Statistical Analysis of influencing Factors on the Winner of the Eurovision Song Contest

Bachelorarbeit

eingereicht von: **Daniela Hnátová**
Matrikelnummer: 11716064

im Fachhochschul-Bachelorstudiengang Wirtschaftsinformatik (0470)
der Ferdinand Porsche FernFH

zur Erlangung des akademischen Grades

Bachelor of Arts in Business

Betreuung und Beurteilung: DI Eszter Geresics-Földi, MSc, BSc

Wiener Neustadt, Mai 2022

Ehrenwörtliche Erklärung

Ich versichere hiermit,

1. dass ich die vorliegende Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Inhalte, die direkt oder indirekt aus fremden Quellen entnommen sind, sind durch entsprechende Quellenangaben gekennzeichnet.
2. dass ich diese Bachelorarbeit bisher weder im Inland noch im Ausland in irgendeiner Form als Prüfungsarbeit zur Beurteilung vorgelegt oder veröffentlicht habe.

Daniela Hnátová

Wien, 03.05.2022

Unterschrift

Kurzzusammenfassung:

Statistische Analyse von Einflussfaktoren auf Gewinner des Eurovision Song Contests

Diese Bachelorarbeit analysiert die Einflüsse der Faktoren Genre, Liedsprache, Alter und YouTube Aufrufe auf die erreichte Platzierung von Eurovision Song Contest Kandidatinnen und Kandidaten unter Verwendung statistischer Forschungsmethoden für dessen Umsetzung die Programmiersprache für statistische Berechnungen R zusammen mit R-Studio als integrierte Entwicklungsumgebung verwendet wurde. Nachdem zunächst der Faktor Genre wegen fehlender objektiver Datengrundlage für weitere Untersuchungen verworfen werden musste, wurde die Einflussanalyse eines jeden der drei übrigen Faktoren als Unterproblem in Form einer statistischen Hypothese formuliert. Die Analyse des Effekts der Sprache auf die Platzierung mittels Varianzanalyse zeigte einen geringen, signifikanten positiven Effekt. Während der metrische Faktor Alter unter Anwendung des Pearsonschen Korrelationskoeffizienten keine signifikante Korrelation zeigte, legten die Daten bei der Analyse des Einflusses der metrischen Variable YouTube Aufrufe einen moderaten, signifikanten senkenden Effekt auf die Platzierung nahe. Für die Teilnehmenden des Eurovision Song Contests 2022 wurde ein Vorhersagemodell mit den YouTube Aufrufen als unabhängige Variable erstellt welches Italien, Serbien und die Ukraine unter den top drei Platzierungen sieht.

Zusammenfassend formuliert, legen die Daten nahe, dass zwei der vier Faktoren des ursprünglichen Forschungsproblems einen signifikanten Effekt auf die Platzierung im Eurovision Song Contest ausüben.

Schlagwörter:

Eurovision Song Contest, ESC, Gewinner, Statistik, R, Vorhersagemodell, Regression

Abstract:

Statistical Analysis of influencing Factors on the Winner of the Eurovision Song Contest

This Bachelor Thesis analyses the influencing factors genre, language, age and YouTube views on the ranking for contestants in the Eurovision Song Contest using statistical research methods for which the programming language for statistical calculating R was used along with R-Studio as integrated development environment. After omitting the factor genre from further analysis due to a lack of objective data, each of the remaining three factors has been formulated as subproblems using statistical hypothesis. The effect of the language on the ranking has been approached with a variance analysis, showing a slight significant effect. Whereas the metric factor age could not indicate a significant correlation using Pearson's correlation coefficient, the metric variable YouTube views demonstrated a moderate, significant lowering effect on the rank. A winner prediction for 2022 has been performed applying a linear regression model with views as the independent variable. This prediction model sees Italy, Serbia and Ukraine among the top three.

In conclusion, the data indicated that two out of the four factors from the initial research problem have a significant effect on the winner of the Eurovision Song Contest.

Keywords:

Eurovision Song Contest, ESC, winner, statistics, R, prediction model, regression

Acknowledgement

I would like to take a moment to express my gratitude to many people without who I would not be where I am now.

First of all, I would like to thank my supervisor Mrs DI Eszter Geresics-Földi, MSc, BSc for her support, guidance, valuable feedback and time spent to help me on my journey.

Secondly, I would like to show my appreciation to Mr Prof.(FH) DI Dr. Martin Staudinger for his initial help with my proposal and for providing me with interesting literature sources for my thesis.

I would also like to thank my classmates Michael Kolodziejek, Cornelius Plaiasu, Johannes Resch and Tatjana-Nadine Tavares for their help and mutual support during the past three years.

Further thanks goes to my parents, my family and my sister for supporting and always having time to listen to me and letting me talk their ears off.

A special thank you I'd like to express to my significant other Michael Maschek, without whom I would not be pursuing a bachelor's degree in Business Informatics nor living in Austria. Thank you for encouraging me to keep going and supporting me during times of feeling helpless.

Table of Contents

1. Introduction and Goal

The Eurovision Song Contest (abbreviated as ESC) has a deep tradition in Europe. Raykoff and Tobin describe ESC as „[...] the largest and most-watched international festival of popular music, as well as one of the world's longest-running annual television programs." (Raykoff and Tobin 2007, p. XVII) Some even argue that the Eurovision Song Contest represents the history of modern Europe as mentioned in „The Secret History of Eurovision". (Oliver 2011)

About 40 countries compete every year for the title of the winner of the Eurovision Song Contest. For this, a pre-selected candidate with a self-written song represents their country. The contest consists of two semi-finals and a final. Only those who receive enough of votes can continue to the final (with exception of the Big Five: France, Germany, Italy, Spain and the United Kingdom as well as the hosting country). The placement in the final is decided by televoting and votes of a jury. The winner receives a glass trophy and a fame across multiple countries that lasts at least for the following year in which the winner's country hosts the next ESC. (Eurovision 2022i)

Since 1965 the ESC reoccurs every year (with exception of 2020) in May on the screens of most European and many other countries, depending on its popularity within each country. (Eurovision 2022a) With such a long history, many fans of the ESC try to discover possible influencing winning factors or even predict the winner beforehand. The goal of this thesis is therefore to find out if it is possible to predict the winner of the ESC based on genre, language, views on YouTube and age of the artist or artists. This objective will be thoroughly explained from a statistic's point of view in the chapter „Research Question and Methods" that also holds explicitly formulated research question, scientific methods and hypothesis.

This thesis focuses on the four influencing factors language, age, genre and views on YouTube in the span of the last 10 years, namely 2011-2021. It does not focus on any other possible factors such as voting bias, order effects, host country or other years than the specified time span of 2011-2021. Further out of scope topics are influencing economic factors on hosting countries, influences on tourism, prediction models via Twitter or prediction methods using artificial intelligence as those topics have already been covered in existing research papers as will be outlined in the chapters „Literature Research and Current Situation" as well as „Research Gap".

This paper is relevant to competing artists, future host countries and fans of ESC. The possible future host countries may gain from the findings of this thesis by their application while choosing the year's selected performer, their possibility of winning and therefore the chance of hosting. It must namely be considered that the hosting country has to undergo a tremendous amount of preparation work. Therefore, knowing that there is a high (or low) probability of winning brings a time advantage for those preparations. Fans of ESC, myself included, may find a comfort in knowing, that their assumptions are indeed true and even scientifically proven in this thesis.

2. Personal Motivation

One might ask why this topic is important or relevant at all. That has been proven to me by the amount of confused faces that I have witnessed while explaining what my bachelor thesis is about.

Several reasons of importance will be outlined as follows.

First of all, being able to predict the winner brings an enormous advantage to the winning country. Hosting of the ESC brings great opportunities not only to the music industry but also to tourism and the economy as a whole for the host country. This alone brings motivation to deal with the topic. The better prepared and the earlier a host country knows its winning probability, the better it can prepare for the enormous investments, construction works and marketing arrangements necessary. (Bard 2017) With a background in marketing and coming from a family who does not only talk a lot about economy but also is very well familiar with the hassles involved in arranging and managing events, I find this opportunity to aid countries by publishing relevant research information, exciting.

Secondly, the factor of academic pleasure of prediction has a big importance in my personal motivation of the topic. Especially after trying to predict the winner of the year 2021 without any usage of proper statistic methods by combining views on YouTube with the overall mood on internet platforms and guessing correctly, with reasonably good results, my motivation to investigate this topic on a more scientific level was awakened.

Figure 1: Dana International

And finally, this subject is significant because of what it represents as a whole. The ESC is a pure representation of togetherness, acceptance, diversity and more. In the words of Bohlmann: „Eurovision has provided a distinctive and very public forum for cultural and musical integration.“ (Bohlmann 2004, p. 288) By providing a platform to artists such as Conchita Wurst (Figure 3), Verka Serduchka (Figure 2) or Dana International (Figure 1), ESC demonstrated their stance on the topic of LGBTQ+.

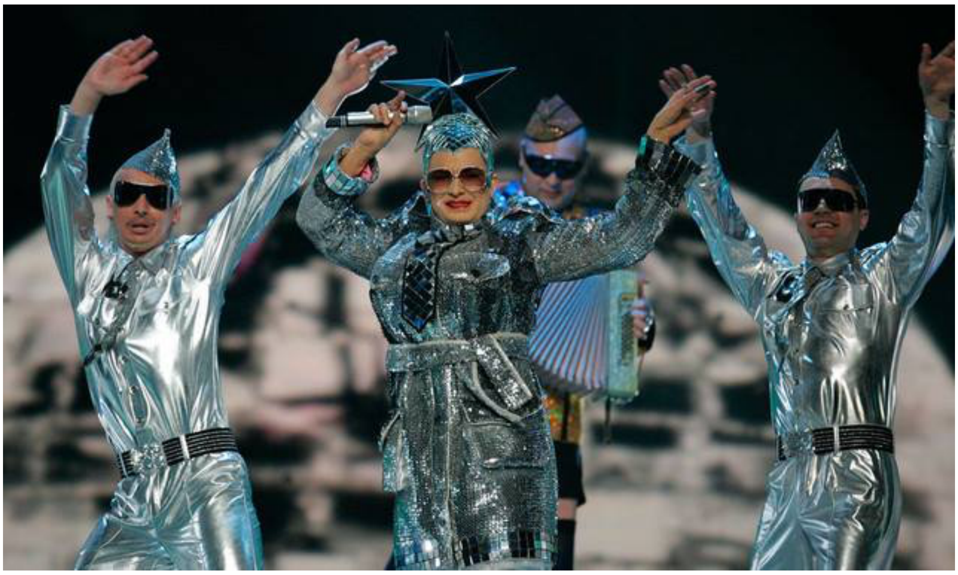


Figure 2: Verka Serduchka

This international music competition shows that there are practically no limits of what is possible when it comes to the performances. Equality, acceptance, tolerance within our heterogeneous society are values that I personally find tremendously important. Therefore, this last point adds to the list of motivating factors to approach this topic for my bachelor’s thesis.

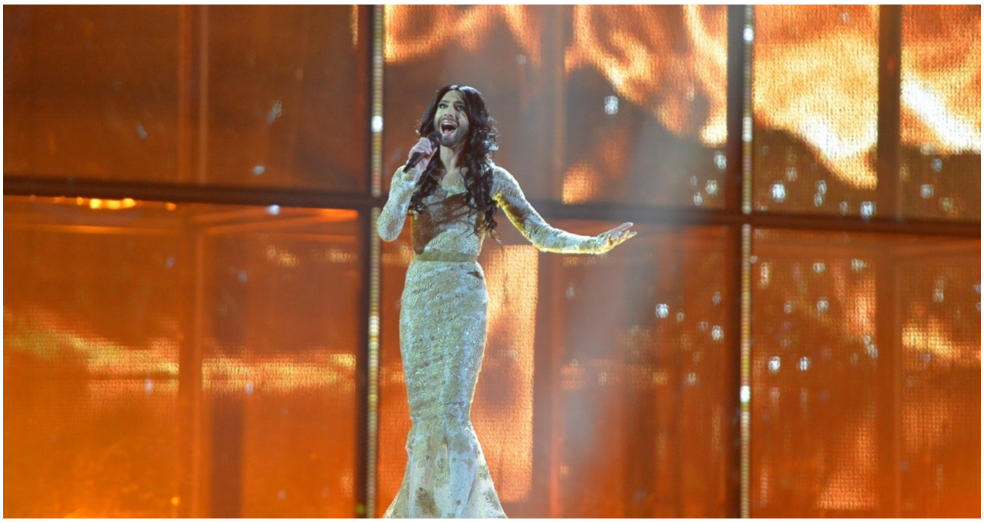


Figure 3: Conchita Wurst

3. Literature Research / Current Situation

The current chapter describes works of other authors that addressed the topic of the Eurovision Song Contest.

At first, there is a regression analysis by Karlsson (2015), analysing the voting patterns among different countries in the jury voting. Karlsson illustrates insights into voting biases and identifies patterns between the countries Greece & Cyprus, Russia & former UDSSR countries as well as Ireland and the United Kingdom. The author visualises those results in a scatter plot and on maps.

Secondly, there are Kumpulainen et al. (2020) who analysed over a million of tweets on Twitter to predict the results of televoting for 2019 with sentiment analysis and correlation coefficients. In this work, the authors analysed of the metric factor number of tweets whereas Kakouris et al. (2016) were using microblogging text from Twitter to identify hidden patterns, classifying tweets by emotions in tweets such as joy, surprise, anger and others in order to predict the winner of 2014. Like this thesis, Kakouris et al. used R for answering their statistical questions.

Thirdly, Budzinski and Pannicke (2016) study whether same hits and same artists are more popular across certain countries and cultures with trend analysis, Gini-Coefficient and the Herfindahl-Hirschman-Index.

Another topic, influencing factors are taken from is “geographical proximity, migration and cultural characteristics.” (Blangiardo and Baio 2014) That includes for instance Blangiardo and Baio (2014) who look among other countries into Serbia and Montenegro, figuring out whether these countries vote for each other because they like the same style of music and share cultural characteristics. In addition, there are Millner et al. (2015) studying a similar topic. However, Blangiardo and Baio use the Bayesian hierarchical model, whereas Millner et al. work with descriptive statistics and regression models. Additionally, Millner et al. also take the effects of the order of appearance of the performers into their calculations. The work of Antipov and Pokryshevskaya (2017) analyses the order of appearance as well, using the Pearson correlation analysis.

The subsequent table 1 gives an overview of the above discussed results of my literature research.

Nr.	Author(s)	Title	Year of publication	Investigated years	Topic	Method	Result	Publication
1	Andreas Karlsson	Eurovision Song Contest: Regression analysis highlighting the voting patterns	2015	1975- 2014	Voting patterns, alliances & biases,	Regression analysis	Scatter plot, visualisations on maps.	Blog
2	Kumpulainen et al.	Predicting Eurovision Song Contest Results Using Sentiment Analysis	2020	2019	Tweets analysis to predict televoting	Sentiment analysis, correlations, spearman correlation coefficients	„Twitter tweets have fairly strong correlation with televoting behaviour.“	National Defence University, Helsinki, Finland
3	Budzinski and Pannicke	Do preferences for pop music converge across countries? Empirical evidence from the Eurovision Song Contest	2016	1975-2016	Popularity of hits and artists across countries and cultures	Herfindahl-Hirschman- Index (HHI) and Gini-Coefficient Trend Analysis	No significant trend.	Ilmenau University of Technology, Ilmenau, Germany
4	Blangiardo and Baio	Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models	2014	1998-2012	Positive or negative bias based on geographical proximity, migration and cultural characteristics	Bayesian hierarchical model	Evidence of mild positive bias, none of negative bias.	University College London, London, UK

5	Millner et al.	Fair oder Foul? Punktevergabe und Platzierung beim Eurovision Song Contest	2015	1999-2014	The impacts of the serial position of a performance, the language of the song and the existence of voting blocs.	Descriptive statistics Regression models	Participants that perform later receive more points on average. Very weak evidence for voting bias.	Ernst-Abbe-Hochschule Jena – University of Applied Sciences, Jena, Germany
6	Antipov and Pokryshevskaya	Order effects in the results of song contests: Evidence from the Eurovision and the New Wave	2017	2009-2012	Order of appearance on their ranking	Pearson correlation coefficients	Weak statistical evidence.	Judgment and Decision Making, Vol. 12, No. 4
7	Kakouris et al.	Detecting Hidden Patterns in European Song Contest—Eurovision 2014	2016	2014	Microblogging text from Twitter to find the winner of the Eurovision 2014	Classifying the tweets by the emotions like: joy, surprise, anger, fear, sadness, disgust in R.	Prediction not correct, can only make estimations.	University of Southern Denmark, Odense, Denmark

Table 1: Literature research overview

4. Research Gap

As the previous chapter showed, biases and voting patterns for jury voting have been studied using different methods. As for the televoting, the factors tweets on Twitter (number of tweets as numeric factor as well as emotions within tweets as categoric factor) and their effects on the ranking have been analysed. However, only Kakouris et al. used R as tool for answering their statistical questions, leaving room for improvement in the analyses among the other authors.

The choice of tool is not the only possible improvement for future research. The scientific community on the topic of influencing factors and their effects on the ESC ranking also lacks a research paper using statistic methods and R as answering tool for the factors language, age, views on YouTube and genre. This thesis will fill this gap.

5. Background and History of Eurovision Song Contest

In this chapter it will be explained how it is possible that countries outside Europe, such as Australia, compete in the Eurovision Song Contest. Furthermore, the history of the ESC, its rules and the program of the semi-finals and the final will be outlined. The chapter finishes with a short insight into politics and highlights of the Eurovision Song Contest.

5.1 Background

Important, reoccurring terms will be explained before the deep dive into the world of the ESC.

5.1.1 EBU

EBU stands for European Broadcasting Union. The EBU was established in the year 1950 with the aim to form an alliance providing a public service media connecting worldwide. (EBU 2022a) Today, the EBU consists of 69 members in 56 countries that are inside the European Broadcasting Union, as well as 20 countries that are outside, such as Australia or the USA. (EBU 2022b)

One of EBU’s companies is the Eurovision, a television network, that is in charge of of the exchange and providing of programs. (Eurovision Services 2022) The Eurovision organises contests such as the Eurovision Song Contest, Junior Eurovision Song Contest, Eurovision Young Dancers, Eurovision Young Musicians or Eurovision Choir of the Year. (Events Eurovision 2022) The members of the EBU can compete in those contests, which answers the question why Australia or Azerbaijan, that are not part of Europe, can join the Eurovision Song Contest.

5.1.2 Big Five

When talking about ESC, one frequently comes across the words „Big Five“. The „Big Five“ refers to five countries that are the highest paying EBU members. The Big Five are every year the same countries, listed as follows: Italy, Germany, Spain, the United Kingdom and France. As a compensation to their generous financial contribution, these countries gain from the advantage of automatically joining the finals. In addition to the Big Five, also the so-called host city (=host) is given the advantage of a guarantee of direct placement in the final.

5.1.3 Host city

The host country is defined by the last year's winner. Usually, the winning country is the host country with only few exceptions. One of these exceptions was when the Netherlands won in 1959. Typically, Netherlands should have been the host of 1960 but that was not possible, because of the high expenses (Netherlands would have to host second year in a row). The ESC in 1960 was therefore hosted in London. (Eurovision 2022b)

The hosting of the ESC is an immense commitment that requires massive financial investment as well as extensive organisation management. According to the EBU website the ESC is financed, among others, by:

"a contribution from the Host Broadcaster, which is generally between €10 and €20 million, depending on local circumstances and available resources. A contribution from the Host City, either financially or 'in kind' (e.g. covering expenses of city branding, side events, security, etc.);" (Eurovision 2022c)

The management of the organisation has always been a challenge but with the recent developments regarding the COVID-19 pandemic, the planning has become even more complex.

The host country must select a host city that fulfils certain criteria of the ESC. For example, the venue must be able to hold at least 10 000 audience members, must be in reach of international airport and of hotel possibilities that are able to accommodate at least 2000 guests. (Jordan and Zwart 2017)

In order to make the ESC 2021 possible given the COVID-19 regulations, multiple scenarios had to be created that reflected the situation at hand. (Eurovision 2022d) Challenges such as the before mentioned regulations, make the organisation especially difficult.

Apart from the pride of being a host city for ones own nation, there are also other benefits. Hosting the ESC is worth the effort due to its "increases in international tourism spending, tax revenue and export". (Bard 2017)

5.2 History

A member of EBU Marcel Bezençon, inspired by the Italian Sanremo Festival, founded the ESC as an intention of unity between European countries after the Second World War. (West 2017, p.8)

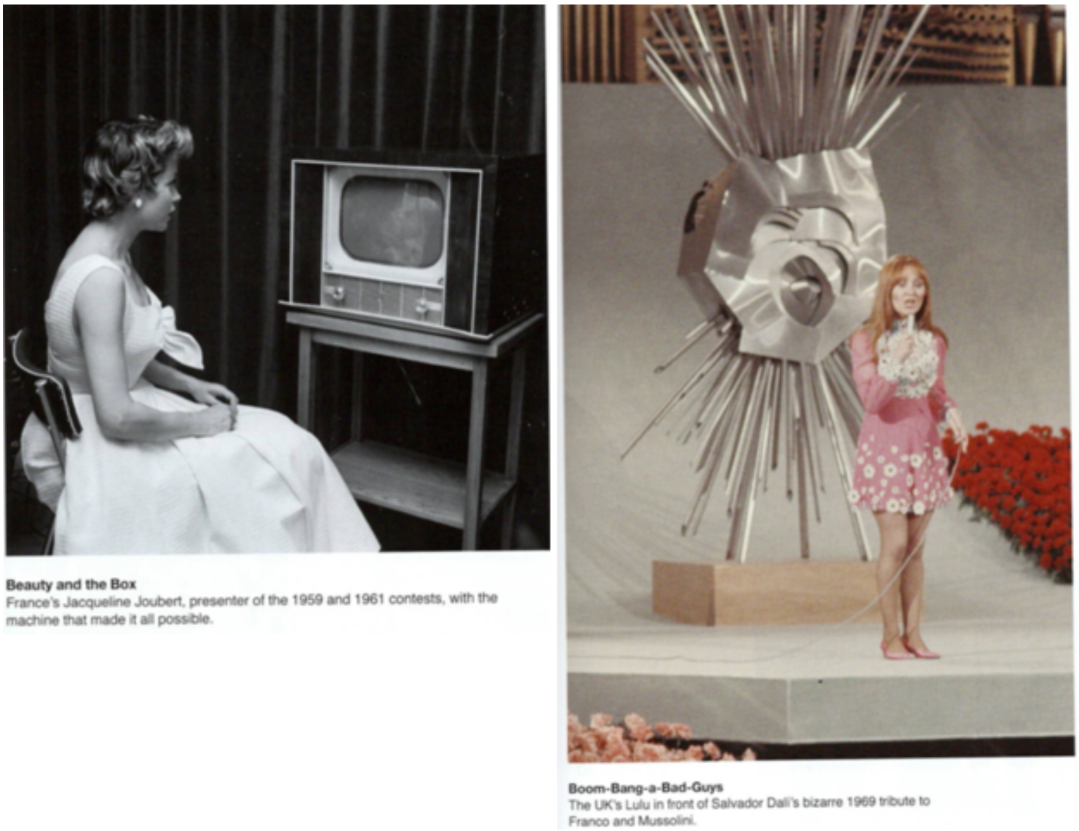
Although, the maximum number of countries that can participate in the ESC is 44, the real number fluctuates at around 40. (Eurovision 2022g) This number changes yearly as it is common for some countries to withdraw, usually because of financial reasons or low interest within the audience. There are two semi-finals that all participants, except for the Big Five and the hosting country, must compete in. (Eurovision 2022i)

The placement in which semi-final and the running order the performers sing, is decided by a random draw to ensure "[..] a more entertaining, better paced show, that allows different styles of song their space to shine." (Eurovision 2022e)

The first ever held ESC was on May 24th 1956 in Lugano, Switzerland with seven competing countries, that were allowed two songs each. Those countries were France, Germany, Belgium, Italy, Luxembourg, the Netherlands and Switzerland. The first contest was very different from how we know it today. The television broadcast was black and white in 1956, a time in which TV sets were very expensive and not wide-spread. Therefore, in order to reach a broader audience, the show was transmitted via radio in addition. Compared to today's ESC version, it was very minimalistic, with simple staging and live orchestrate music. Even the choice of languages in which the first competitors sang, were different from nowadays. The national language was the choice in 1956, whereas it is English that prevails in last years. (Eurovision 2022f)

Since 1956, the ESC has been held annually with only one exception - the year 2020. Due to the COVID-19 pandemic, the contest had to be cancelled for the first time in the history of the Eurovision Song Contest.

To compare the early years of the ESC to the latest ones, images have been scanned from the book "Eurovision! A History of Modern Europe Through the World's Greatest Song Contest" (West 2017) and screenshots of live videos from the official Eurovision YouTube Channel from the year 2021 have been taken.



Beauty and the Box
France's Jacqueline Joubert, presenter of the 1959 and 1961 contests, with the machine that made it all possible.

Boom-Bang-a-Bad-Guys
The UK's Lulu in front of Salvador Dali's bizarre 1969 tribute to Franco and Mussolini.

Figure 4: The beginnings of the Eurovision Song Contest

The contrast between those two figures is visible immediately. Figure 4 shows the beginnings of the ESC. The left image is called “Beauty and the Box”, illustrating Jacqueline Joubert, the presenter from France, posing with a television. The picture on the right, is a portrait of Lulu from the UK, from the year 1969, where colour already made an appearance.

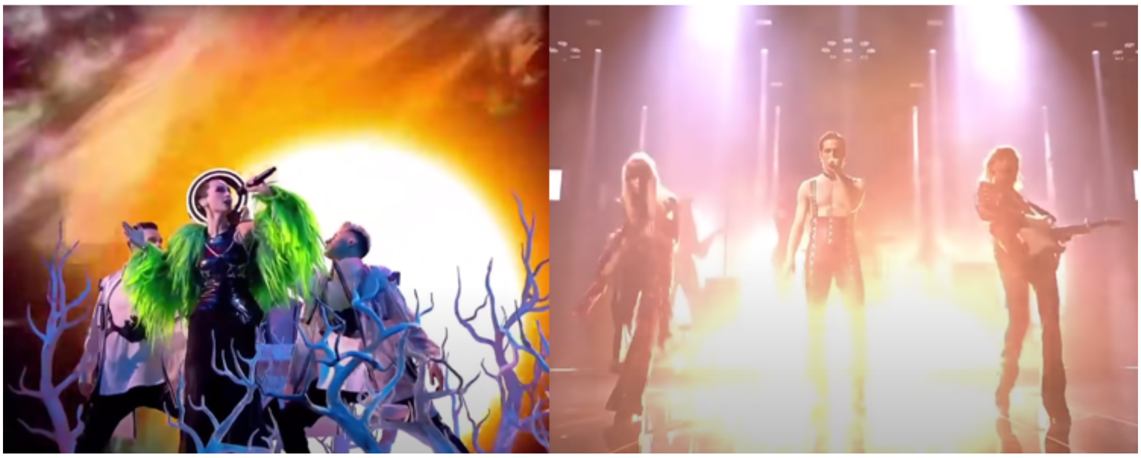


Figure 5: The ESC from the year 2021

The figure 5 shows what the ESC looks like today. The colours are magnificent, the stage and technology has grown. There are practically no limits to the choice of costume, simply, following the slogan “dress to impress”. According to the official ESC website, there were 20 fire fountains, 48 stage flames, 10 fog machines and 8 confetti canons available for the special effects in the ESC of 2019. (Royston 2019)

5.3 Rules

5.3.1 National Selection

The countries‘ method of selecting their representative contestant for the ESC may be chosen freely. Therefore, different nations follow different selection methods. One method, for example, is splitting the vote between expert jury (50%) and online app voting (25% international and 25% national voters). This method is applied by the Czech Republic. Another method is the hosting of the country‘s own music competition in which a mix of televoting and expert jury votes are applied in order to decide for the national performer. This selection method is used by Albania (Festivali i Këngës), Sweden (Melodifestivalen), Italy (Sanremo) or Norway (Melodi Grand Prix). Other countries such as Austria or Bulgaria select their performers through an internal selection, which means that the public has no say in the decision. (Eurovision World 2022)

5.3.2 Language

The rules about the language in which the song is to be performed has changed many times since the beginning of the ESC in 1956.

Until 1965, so the first nine years, there were no language restrictions. At this time, all competitors decided to sing in their national language. After the unexpected decision of Sweden to sing in English in 1965, a rule has been established, allowing performers to sing in the country‘s language only. (West 2017, p. 51) That rule lasted for six years and was then replaced, officially allowing to sing in any language. (West 2017, p. 80) This time, the new rule was in power for five years when the first rule from 1965 was reintroduced and remained active until 1999. (West 2017, p. 101) The latest rule established by EBU from 1999 states that, „Each Participating Broadcaster is free to decide the language in which its Contestant(s) will sing.“ (Eurovision 2022g)

5.3.2.1 Variety

The variety of languages is as diverse as Europe and other EBU member countries themselves. Any language, even imaginary ones, are allowed to be used, which makes the possibilities practically endless.

After the establishment of the most recent language rule, most of performers switched to singing in English. Unsurprisingly, participating countries whose national language is English (Australia, Ireland and the United Kingdom), did not aberrate from this trend. However, some countries, such as Albania, France, Italy or Spain tend to sing, at least partially, in their national languages.

5.3.2.2 Recent Development

Recently, an interesting trend has occurred, in which more and more countries perform either completely in their national language or at least insert a verse or two in the national language in an otherwise English song. This trend became apparent in 2021, when the top three contestants performed in other languages than English: Italian (the winner Måneskin) and French (second place Barbara Pravi, third place Gjon’s Tears).

Many argue that singing in English brings the audience closer to the performer since they understand it better. I personally don’t see much of a significance in that, considering that most of the ESC viewers are from countries in which English is not a national language, meaning the language relationship is not that strong. This argument gets support by the fact that Salvador Sobral won in 2017 while singing in Portuguese. At this point it is to be noted again that it was an Italian song that won in 2021, followed by two French acts placing second and third. These events leave to wonder if the shift of the preferred language will continue in the future and whether English will lose even more of its importance as an influencing factor on the ESC winner.

Conclusively, some of the most popular languages used in songs during ESC performances are listed as follows:

- English
- French
- German
- Italian
- Spanish

5.3.3 Genre

As far as my research goes, there are no regulations on the genre of the song to be performed in the Eurovision Song Contest. There are rules for the song itself, for example the length, number of performers and even a regulation on lyrics. The maximum length of a song accounts for three minutes, whereas the number of artists is limited to six performers, none of which may be an animal. (Eurovision 2022g). The before mentioned regulation on lyrics states the following:

„No lyrics, speeches, gestures of a political, commercial or similar nature shall be permitted during the ESC. No swearing or other unacceptable language shall be allowed in the lyrics or in the performances of the songs". (Eurovision 2022g)

Eurovision comes across as filled with pop music with moving ballads and elegant chansons. The genre range grew over the years as much as the contest itself. Thirteen out of the fourteen songs that were performed in 1956 in Lugano, were either of the genre chanson or ballad. The contestant from Germany served as an exception to that by choosing rock and roll as their genre for their song. (Eurovision 2022f)

The genre mix of the latest ESC in 2021, with a total of 39 competing countries, was respectively bigger, consisting of genres such as rock, pop, ballad, chanson, 80's pop or folktronica.

5.3.4 Age

As stated by the EBU, all performers must be at least 16 years old on the day of the final. (Eurovision 2022g)

5.3.5 Running Order

Before the contest begins, all countries except for the six countries with a guaranteed place in the final, are randomly placed in a pool in either first or second semi-final. Afterwards, they are drawn randomly and given a place in their previously appointed semi-final. (Eurovision 2022e)

5.3.6 Voting

As stated by EBU: "The voting is compulsory in all the countries of the Participating Broadcasters. [..] the televoting and the national jury voting [..] to ensure a central control and verification of the results." (Eurovision 2022g)

The voting consists of two parts and is split in 50% each. The first part is the jury voting. The jury members must be music professionals and judge the vocal capacity, the performance on stage, the composition and originality of the song and the overall impression by the act. There are some rules that come along being a jury member.

Some of the rules are:

- “The jury voting is always monitored by an independent notary in each country
- The jury consists of a variety of members in terms of age, gender, and background
- All jury members must be citizens of the country they are representing
- None of the jury members must be connected to any of the participating songs/artists in such a way that they cannot vote independently.
- Members shall not have been part of a National Jury the preceding two years” (Eurovision 2022h)

The second part is televoting, during which the audience can vote for their favourite song. Voting can be done by telephone, sending of a SMS or by using of the official app “Eurovision”.

There are slight differences in voting between semi-finals and the final, as follows:

5.3.6.1 Semi-finals

Only the countries competing in semi-final 1 can vote for that event and those countries who participate in semi-final 2 can only vote for the second semi-final. In addition, countries of Big Five and the host country are obliged to make voting possible for both semi-finals.

5.3.6.2 Final

All competing countries are obliged to make voting possible during final, even those who did not continue from semi-finals.

5.3.7 Point System

Both, jury vote and televote give 1 to 8, 10 and 12 points to their top 10 songs. The 12 points, also known as “douze points”, are given to the performer they like the most. On the other end of the scale, it is also possible to reach 0 points, also known as “nul points”. That does not happen very often.

Voting in the semi-finals determines 10 qualified countries, that can continue their journey to the grand final. Those 10 countries are announced on the end of the respective semi-finals, without displaying of their ranking. The ranking of the rest of the countries that didn’t qualify is therefore also not mentioned.

As for the final, the jury votes are represented at first by each nation’s representative. After that, the televote results are shared, starting with the country with the lowest points, working up to those with most points received by the jury. This ensures the most exciting and surprising way to finding out the winner. (Eurovision 2022i)

5.4 Program Description

The Eurovision Song Contest, as mentioned before, consists of three events that are broadcasted across multiple countries, during one week in May. The first semi-final always falls on Tuesday, the second semi-final on Thursday and the grand final takes place on Saturday, no matter the time-zone of the host city, the shows begin at 21:00 CEST. Before each of the events start, the Eurovision logo makes an appearance playing Prélude du Te Deum by Charpentier. The contest starts by a video introduction of the host country and short recapitulation of the last year, followed by an opening act from a non-competing artist or the winner of the last year.

5.4.1 Semi-finals

Depending on the total amount of competitors per year, each Semi-final consists of approximately 20 participants. Those participants then get called out by the name of their country as an introduction. Based on the running order announced at the draw, the competitors perform their song. Before each contestant’s performance, there is a short clip featuring the the artist in form of a so-called postcard (West 2017, p. 71). After all of the contestants perform their songs, the voting lines are opened. The contest continues by showing short clips with voting numbers of all the participants in between interval acts and displaying three of the six countries, that are automatically qualified for the final. Afterwards, the voting lines are closed and the 10 countries to continue to the final are announced by the hosts.

5.4.2 Final

A maximum of 26 countries are allowed to compete in the Grand Final. (Eurovision 2022g) The time plan is not very different from the Semi-finals, except for the fact that it lasts about double as long. As in Semi-finals, the postcards are shown, the artists perform and then the voting lines open. The difference between Semi-finals and the Final is especially in the announcing of the placement. The jury vote is announced by each country's representator live, giving all 1 to 8, 10 and 12 points. After that, the first overview of points is visible. Next, the televotes are announced by the hosts, starting at the lowest placements to create excitement until the last minute. Once all points are announced, the winner is revealed. The winner receives a glass microphone trophy, as pictured in figure 6 and performs their song one more time.

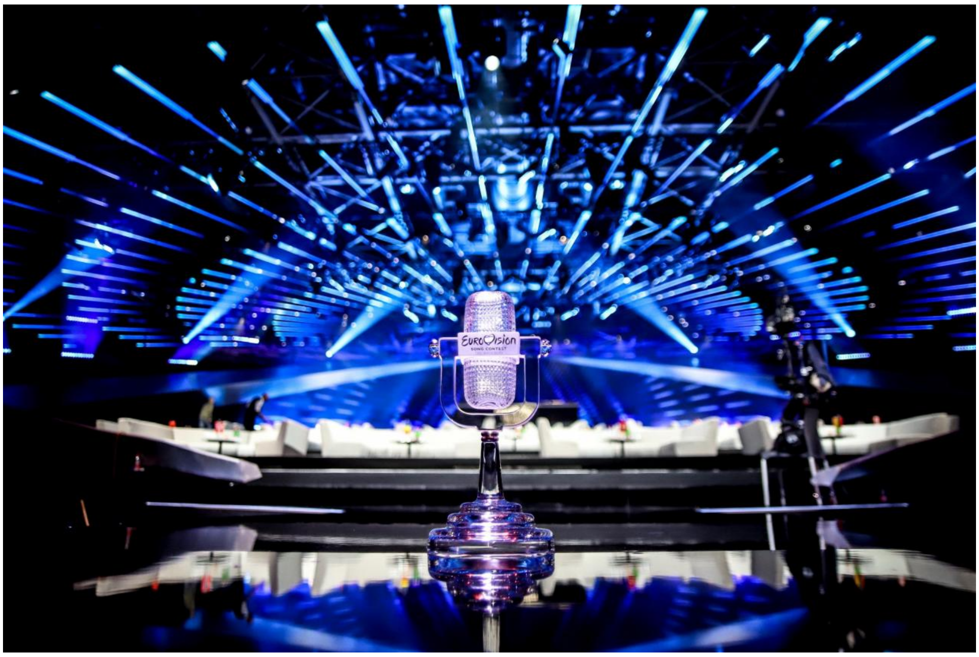


Figure 6: Eurovision 2019 trophy (© Thomas Hanses)

5.5 Highlights

Unfortunately, many of the ESC winners are remembered only by the Eurovision community after their victory. However, for some participants, the ESC was just a beginning of their stardom. Some of them are:

- ABBA, winner of 1974 with their song Waterloo
- Céline Dion, winner of 1988 with her song Ne Partez Pas Sans Moi
- Conchita Wurst, winner of 2014 with their song Rise Like A Phoenix
- Lordi, winner of 2006 with their song Hard Rock Hallelujah
- Måneskin, winner of 2021 with their song Zitti e Buoni

Another interesting fact is the outcome of 1969 in which a total of four countries won. A movie from 2020 is also worth mentioning, namely the Eurovision Song Contest: The Story of Fire Saga by Will Ferrell. The movie takes place in a small town in Iceland, portraying a couple of friends whose biggest dream is to compete in the ESC. They take us with them on their journey, starting with the national selection all they way to the actual contest. The production from Netflix also featured some of the actual ESC participants from last years such as Conchita Wurst (winner 2014), Alexander Rybak (winner 2009, 15th place 2019), Loreen (winner 2012), Netta (winner 2018), Salvador Sobral (winner 2017) or Jamala (winner 2016). The movie was especially popular as it was published in June 2020 and acted as a type of replacement of the actual contest that has been cancelled.

As a result of the popularity of the Eurovision Song Contest, the USA have also decided to make their own version of the show called American Song Contest and will take place for the first time between February and March 2022. (Eurovision 2022j)

5.6 Politics

The ESC claims not to be a political event. That proves to be difficult when so many countries and cultures join on one stage. West (2017), presents interesting insights on the connection of historical events and the Eurovision Song Contest. One of them, that is close to my heart, is from 1968, in which the Czech singer Karel Gott was asked by Austria to represent them. Czechoslovakia wasn’t part of the EBU and therefore could not take place in the contest. But that is not the reason why Austria decided to do this move. 1968 was a very eventful year in the history of Czechoslovakia. It started with hope for the folk given by the new political party in power but the hope disappeared as soon as August came and with it the invasion of Czechoslovakia by troops from Russia. (West 2017, p. 62-63)

Ukraine competed in 2016 with their song 1944, surrounded by many controversies. The song is about the deportation of Crimean Tatars by Stalin. Many people were wondering about the non-politicalness of such a song. However, the EBU has decided that the song did not match current events and was therefore allowed to be performed. The song won. (West 2017, p. 296-297)

6. Research Question and Methods

The overall research question and its derived overall hypotheses of this thesis are:

ID: Q1	Overall Research Question
Is it possible to predict the winner of the Eurovision Song Contest using the influencing factors language, age of the main artist, views on YouTube and genre?	
ID: H1	Overall Hypotheses
It is possible to use the influencing factors language, age of the main artist, views on YouTube and genre in a prediction model in order to determine the winner of the Eurovision Song Contest.	

Table 2: Overall Research Question

The following figure 7 created by me in the visual collaboration tool Creately illustrates the procedure that is to be followed in order to answer the overall research question Q1. In this model, rectangles represent activities whereas the ellipses contain artefacts serving as input/output for the activities.

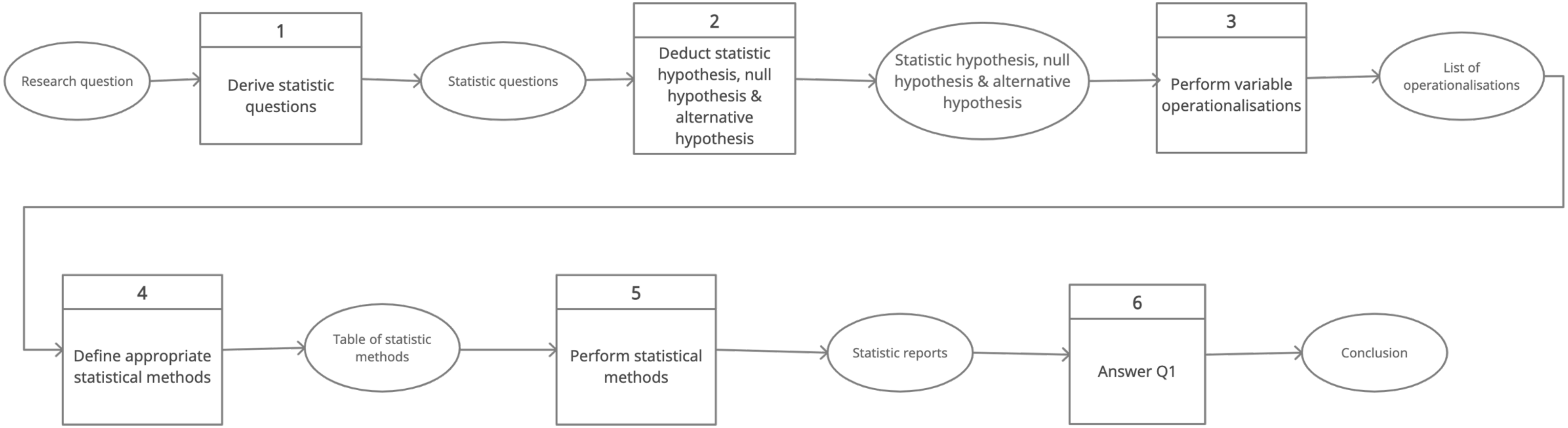

Research question
1
Derive statistic questions
Statistic questions
2
Deduct statistic hypothesis, null hypothesis & alternative hypothesis
Statistic hypothesis, null hypothesis & alternative hypothesis
3
Perform variable operationalisations
List of operationalisations
4
Define appropriate statistical methods
Table of statistic methods
5
Perform statistical methods
Statistic reports
6
Answer Q1
Conclusion

Figure 7: Research Procedure (by Daniela Hnátová)

The following table gives an overview in which chapter each of the in figure 7 listed artefacts can be found.

Artefact	Chapter
0 - Research Question	6
1 – Statistic Questions	6
2 – Statistic hypotheses, null hypotheses and alternative hypotheses	8.1
3 – List of operationalisations	8.2
4 – Table of statistic methods	8.3
5 – Statistic Reports	9
6 - Conclusion	10

Table 3: Overview of artefacts

The statistical methods used in activity 5 will be illustrated and explained in more detail in chapter 8.3. The following list gives an overview of some of the statistical methods that will be used.

- Histogram
- Scatterplot
- Barplot
- Boxplot
- Parallel boxplot
- Residual plot
- Q-Q plot
- Chi-square Test
- Regression analysis
- t-test
- Wilcoxon test
- ANOVA table

These statistic methods including the creation of statistic reports (output of activity 5) will be performed using the programming language for statistic computing R. R is a language that will create results according to accepted statistic methods. During the input throughout lectures from the Ferdinand Porsche Fernfachhochschule, its powerful functionality has been demonstrated suitable multiple times for the purpose of answering statistic research questions. R will be executed in the open-source environment RStudio that also serves as my IDE. RStudio does not only execute R seamlessly and comes with third party functionalities but also provides the for this thesis needed automated reporting (knitting) option.

Following figure 7 (Research Procedure), activity 1 (Derive statistic questions) has been performed, using the input of the overall research question Q1.

The following table shows the output artefact of activity 1:

ID: Q1.L	**Statistic Question Language**
Does the language of a competing song have an effect on the ranking in the Eurovision Song Contest?	
ID: Q1.A	**Statistic Question Age**
Does the age of a contestant have an effect on their ranking in the Eurovision Song Contest?	
ID: Q1.V	**Statistic Question Views on YouTube**
Do the YouTube views of a competing song have an effect on the ranking in the Eurovision Song Contest?	
ID: Q1.G	**Statistic Question Genre**
Does the genre of a competing song have an effect on the ranking in the Eurovision Song Contest?	

Table 4: Statistic Questions

Before proceeding with activity 2 (Deduct statistic hypotheses, null & alternative hypotheses), which is to be found in chapter 8.1, the data collection procedure and methods are outlined in the proceeding chapter „Data Collection“.

7. Data Collection

This chapter describes the empirico-statistical elicitation for the sample data needed. One unit of observation consists of the tuples (language, ranking), (age, ranking), (views on YouTube, ranking) and (genre, ranking) respectively. The total population holds 1596 units of observations, comprising of all historic ESC participants, beginning with the very first contest in 1956 and ending with 2021. With one Eurovision Song Contest being carried out per year (excluding 2020 because of COVID-19 regulations), a total of 65 song contests have taken place until today (January 2022). The following table gives an overview of the number of participants per year.

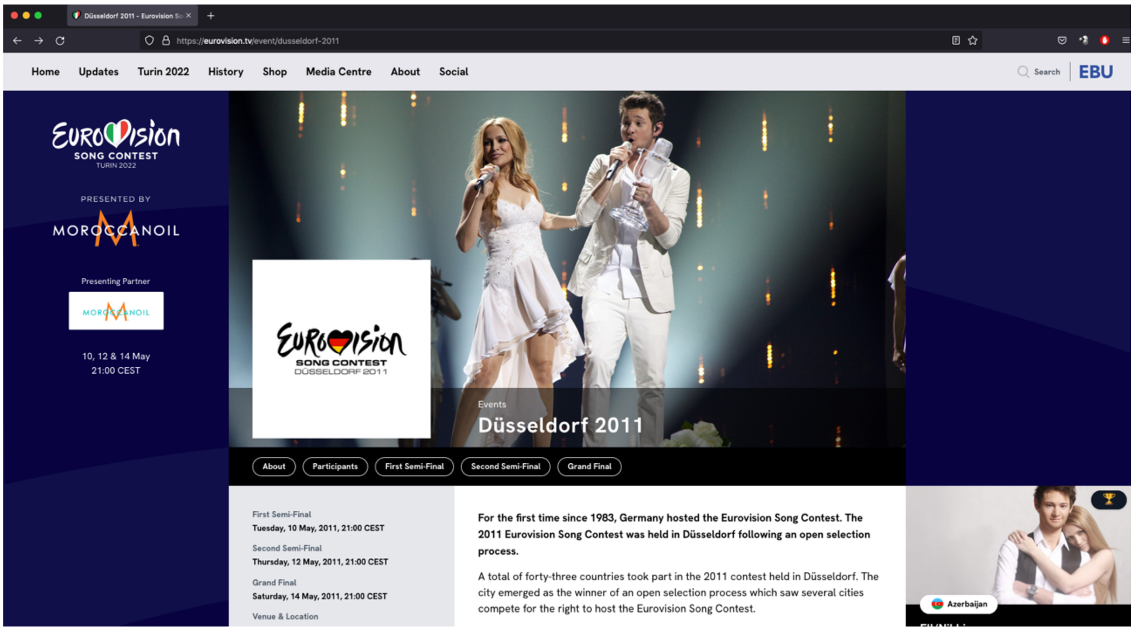
Year	Nr. of Particip.	Year	Nr. of Particip.	Year	Nr. of Particip.	Year	Nr. of Particip.
1956	7	1973	17	1990	22	2007	42
1957	10	1974	17	1991	22	2008	43
1958	10	1975	19	1992	23	2009	42
1959	11	1976	18	1993	25	2010	39
1960	13	1977	18	1994	25	2011	43
1961	16	1978	20	1995	23	2012	42
1962	16	1979	19	1996	23	2013	39
1963	16	1980	19	1997	25	2014	37
1964	16	1981	20	1998	25	2015	40
1965	18	1982	18	1999	23	2016	42
1966	18	1983	20	2000	24	2017	42
1967	17	1984	19	2001	23	2018	43
1968	17	1985	19	2002	24	2019	41
1969	16	1986	20	2003	26	2020	0
1970	12	1987	22	2004	36	2021	39
1971	18	1988	21	2005	39		
1972	18	1989	22	2006	37		

Table 5: Number of participants per year 1956-2021

For Q1.L, Q1.A and Q1.G, the collection of the attributes language, age, genre and ranking is required for every participant. For this, the sample size comprises of all ESC participants from 2011 until 2021, accounting for a total of 408. That is 25,56% of the total population.

The above-mentioned time frame was decided as it is most representative for the current events. Rules remained relatively unchanged throughout the past ten years, which is definitely not true for the time before as regulations regarding the language, for example, changed regularly (cf. chapter Background and History of Eurovision Song Contest). In addition, the number of participants varied a lot throughout the years before (only 7 in the year 1956 and fluctuating now at about 40). Therefore, the time frame of the past ten years seems the most stable, making it perfect for an objective sample for the statistical methods.

It has been decided to sample in a systematic way, using a secondary source, namely the official ESC website (https://eurovision.tv) as it is impossible to collect primary data from the past. Having access to the most trustworthy source of data, the website of the organiser themselves, this is not an issue. The following figure 8 shows the screen of the secondary source, eurovision.tv.



First Semi-Final
Tuesday, 10 May, 2011, 21:00 CEST

Second Semi-Final
Thursday, 12 May, 2011, 21:00 CEST

Grand Final
Saturday, 14 May, 2011, 21:00 CEST

Venue & Location

For the first time since 1983, Germany hosted the Eurovision Song Contest. The 2011 Eurovision Song Contest was held in Düsseldorf following an open selection process.

A total of forty-three countries took part in the 2011 contest held in Düsseldorf. The city emerged as the winner of an open selection process which saw several cities compete for the right to host the Eurovision Song Contest.

Figure 8: Secondary data source: eurovision.tv (eurovision.tv)

As for Q1.L, the collection of the data will proceed as follows:

1. Selecting of the desired year: Menu -> History -> History by year -> Year
2. Clicking on “Participants” and going through each participant’s page
3. Gathering of the data from “Lyrics” about the language based on following:
   a. Language is English -> English
   b. Language other than English and only one language -> Non-English
   c. Language mix of more languages -> Mix

Following figures clarify the steps of the process.

In Figure 9, the navigation highlighting step 1 is shown.

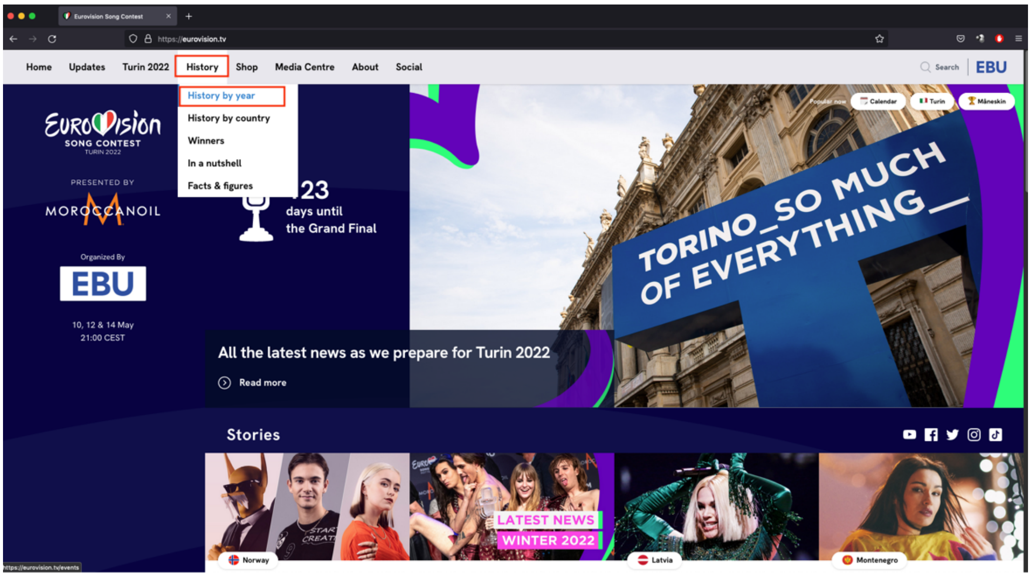


Figure 9: The Navigation of eurovision.tv (eurovision.tv)

Figure 10 illustrates the overview of all events by year.

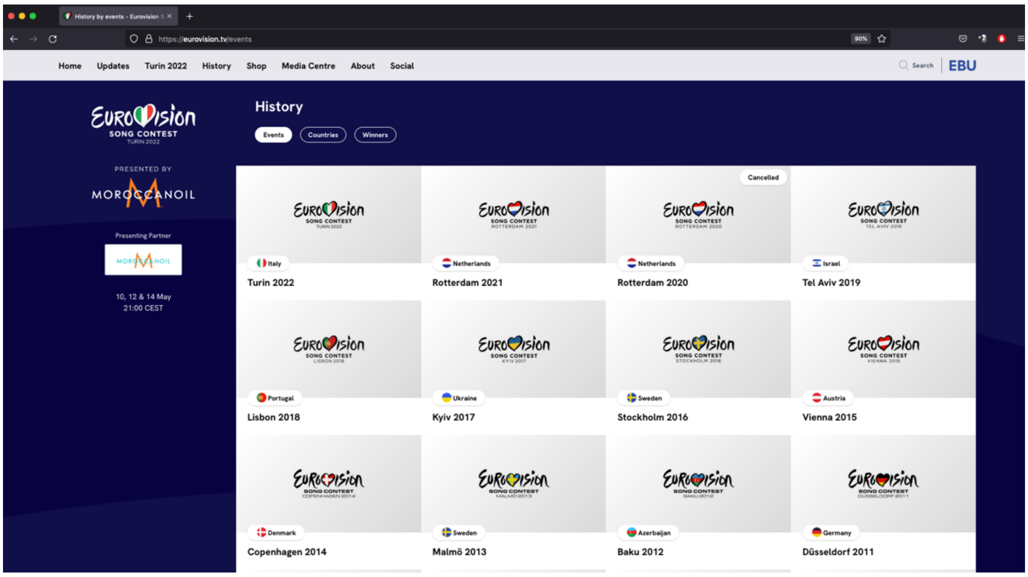


Figure 10: Overview of all events by year (eurovision.tv)

Afterwards, Figure 11, displays an overview of a specific year, in this example, the year 2011 with a highlight of participants, in red square.

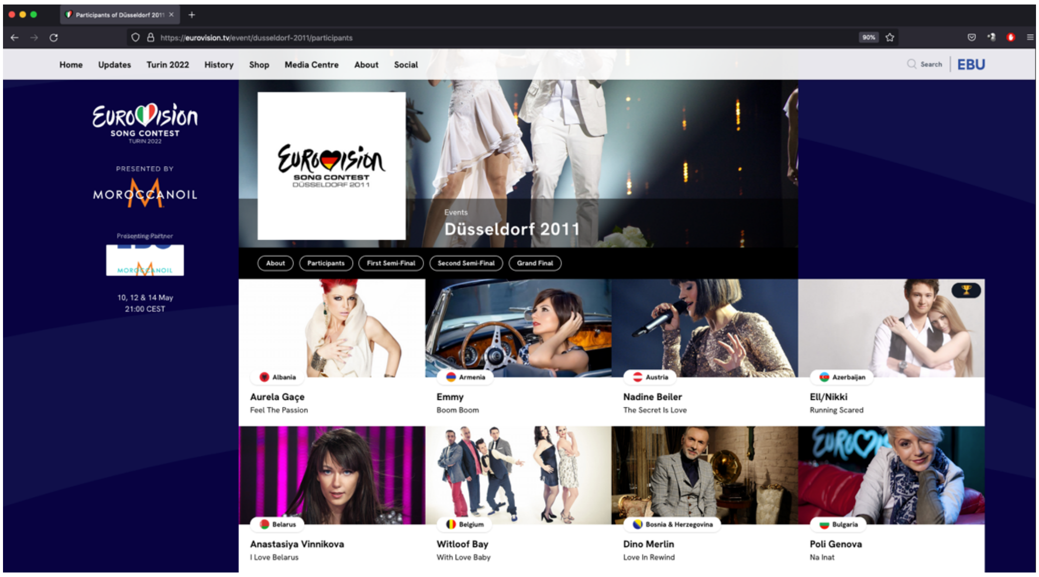
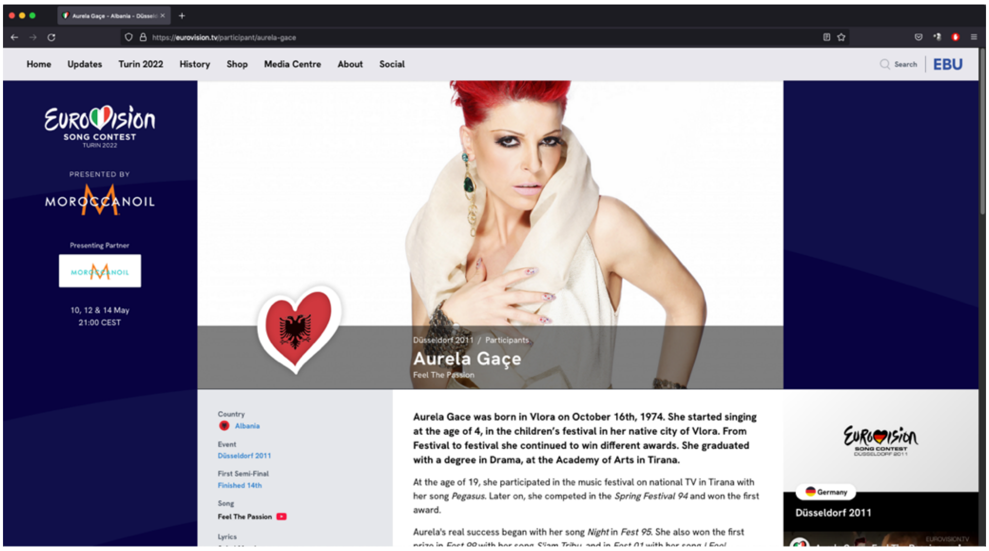


Figure 11: Overview of the year 2011 (eurovision.tv)

As per step 2, Figure 12 features a specific participant, in this case the singer Aurela Gaçe from Albania.



Aurela Gaçe

Feel The Passion

Country: Albania

Event: Düsseldorf 2011

First Semi-Final: Finished 14th

Song: Feel The Passion

Lyrics

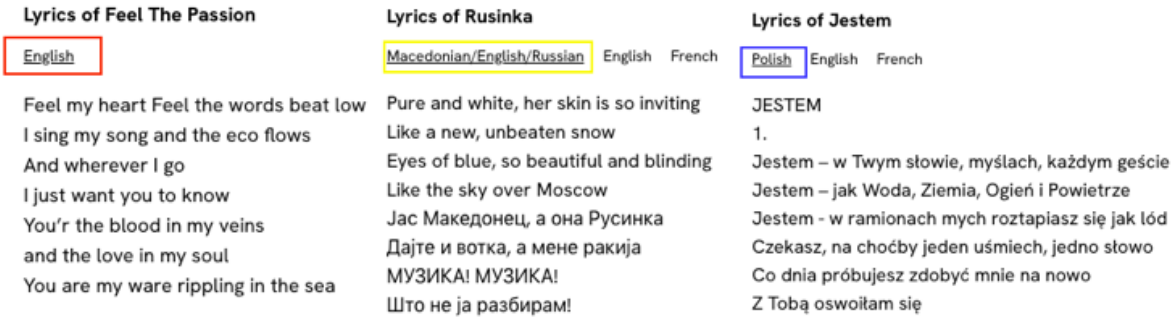
Aurela Gace was born in Vlora on October 16th, 1974. She started singing at the age of 4, in the children's festival in her native city of Vlora. From Festival to festival she continued to win different awards. She graduated with a degree in Drama, at the Academy of Arts in Tirana.

At the age of 19, she participated in the music festival on national TV in Tirana with her song *Pegasus*. Later on, she competed in the *Spring Festival 94* and won the first award.

Aurela's real success began with her song *Night* in *Fest 95*. She also won the first prize in *Fest 99* with her song *S'iam Tribu*, and in *Fest 01* with her song *I Feel*

Figure 12: Overview of a participant (eurovision.tv)

The data for Q.1L are gathered and split based on the type as stated in step 3. Figure 13 shows example for type "English" highlighted in red, the type "Non-English" in blue and the type "Mix" in yellow.



Lyrics of Feel The Passion

English

Feel my heart Feel the words beat low
I sing my song and the eco flows
And wherever I go
I just want you to know
You'r the blood in my veins
and the love in my soul
You are my ware rippling in the sea

Lyrics of Rusinka

Macedonian/English/Russian English French

Pure and white, her skin is so inviting
Like a new, unbeaten snow
Eyes of blue, so beautiful and blinding
Like the sky over Moscow
Јас Македонец, а она Русинка
Дајте и вотка, а мене ракија
МУЗИКА! МУЗИКА!
Што не ја разбирам!

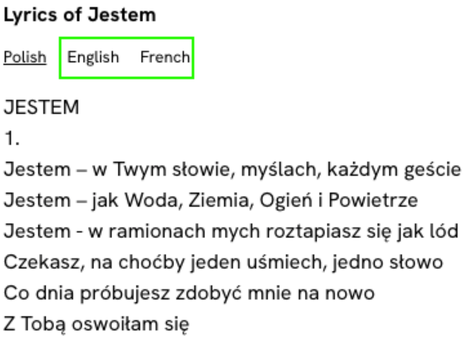
Lyrics of Jestem

Polish English French

JESTEM
1.
Jestem – w Twym słowie, myślach, każdym geście
Jestem – jak Woda, Ziemia, Ogień i Powietrze
Jestem - w ramionach mych roztapiasz się jak lód
Czekasz, na choćby jeden uśmiech, jedno słowo
Co dnia próbujesz zdobyć mnie na nowo
Z Tobą oswoiłam się

Figure 13: Language types English, Non-English and Mix (eurovision.tv)

For both Non-English and Mix, there is a translation in English (sometimes also French) added next to the lyrics, shown in Figure 14 with a green square.



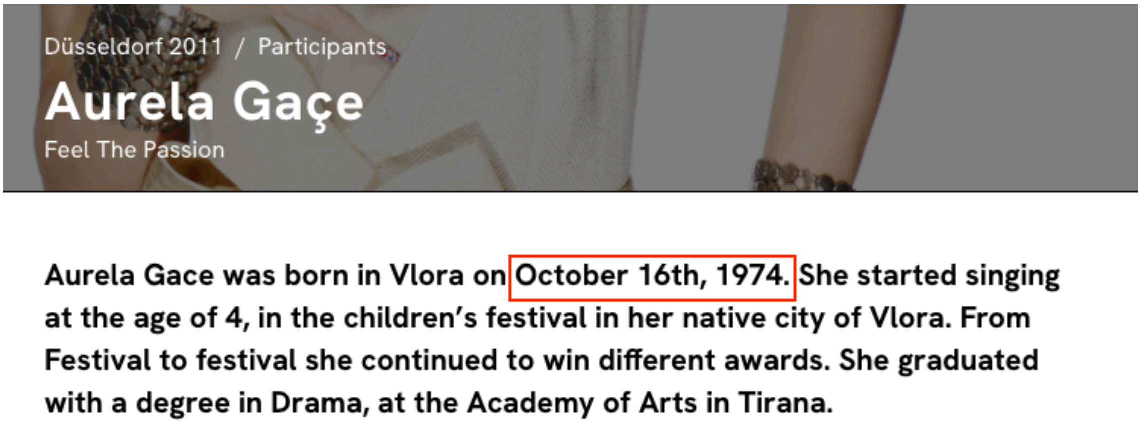
Lyrics of Jestem

Polish English French

JESTEM
1.
Jestem – w Twym słowie, myślach, każdym geście
Jestem – jak Woda, Ziemia, Ogień i Powietrze
Jestem - w ramionach mych roztapiasz się jak lód
Czekasz, na choćby jeden uśmiech, jedno słowo
Co dnia próbujesz zdobyć mnie na nowo
Z Tobą oswoiłam się

Figure 14: Translation (eurovision.tv)

The collection of data for Q1.A is identical with Q1.L in steps 1 and 2. Step 3, as visible on Figure 15, comprises of reading the first paragraph of the participant's page from eurovision.tv, where the age of the artist is stated. In cases in which the official ESC website does not state the age, the book series "The Complete & Independent Guide to the Eurovision Song Contest" by Simon Barclay will be used as a data source.



Düsseldorf 2011 / Participants

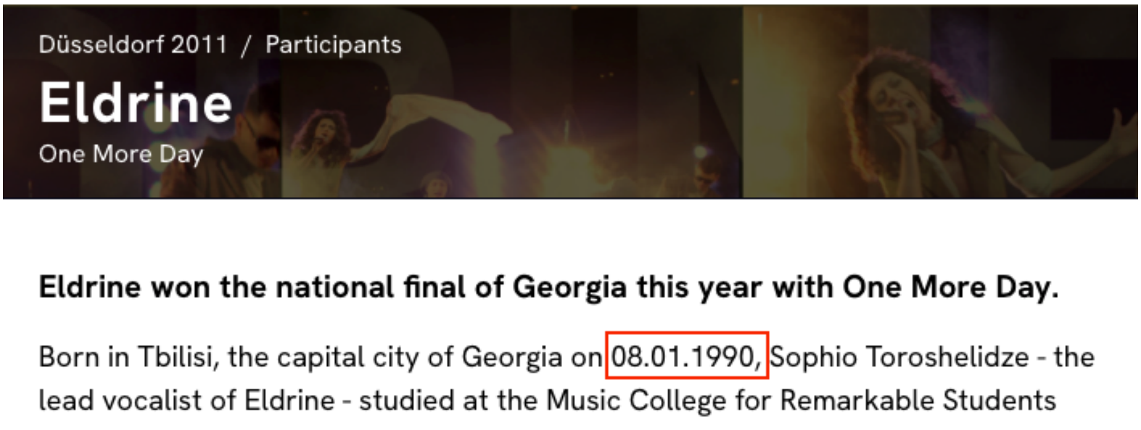
Aurela Gaçe

Feel The Passion

Aurela Gace was born in Vlora on October 16th, 1974. She started singing at the age of 4, in the children's festival in her native city of Vlora. From Festival to festival she continued to win different awards. She graduated with a degree in Drama, at the Academy of Arts in Tirana.

Figure 15: Age of participant (eurovision.tv)

In case of more performers, in form of a band or a duet, it has been decided to take the age of the lead performer. For example, as shown in Figure 16, in the year 2011 the Georgian band Eldrine's lead vocalist's age (in red square) will be taken.



Düsseldorf 2011 / Participants

Eldrine

One More Day

Eldrine won the national final of Georgia this year with One More Day.

Born in Tbilisi, the capital city of Georgia on 08.01.1990, Sophio Toroshelidze - the lead vocalist of Eldrine - studied at the Music College for Remarkable Students

Figure 16: Age - example multiple performers (eurovision.tv)

The collection of necessary information about the attribute of Genre has proven to be more difficult than initially expected. The problem is namely that the genre is not explicitly stated on the official website. That means, that in order to gather all of the data required, listening to each song is needed. Not only is that very demanding when it comes to timely resources but it also means, that the data cannot be objective, as my impressions of each song vary from person to person, making a reproduction by others for scientific purposes impossible. Additionally, while describing the topic of genre, Meijer wrote in his Master Thesis that "Some songs could fit into multiple genres, but I have decided to place them into one single category which is emphasized most in the performance." (Meijer: 2013, p. 62), proving that the genre is not a suitable attribute as neither objectivity nor an unambiguous categorisation can be ensured. Therefore, Q1.G will not be further discussed in this thesis.

For illustration purposes, there are 2 units of observation for each of the statistic questions Q1.L and Q1.A in tables 6 and 7:

Q1.L	
Language	**Ranking**
English	13
Non-English	5

Table 6: Example: Unit of Observation – Language

Q1.A	
Age	**Ranking**
27	23
18	10

Table 7: Example: Unit of Observation – Age

For Q1.V, the collection of the information on views on YouTube and ranking was made by me, making it a primary source. This year (2021), I was gathering data in form of views on YouTube since January until May for all of the 39 contestants. For each contestant, I noted the amount on views from the applicable YouTube channel, one day (17.05.2021) before the beginning of the Eurovision Song Contest. Then I created a list of possible rankings based on the amount of views, which I compared later during the final on 22.05.2021 and noted the differences.

For illustration purposes, there are 2 units of observation for the statistic question Q1.V in table 8:

Q1.V	
Views on YouTube	**Ranking**
980 000	21
3 100 000	2

Table 8: Example: Unit of Observation – Views on YouTube

The following table summarises this chapter.

Research Question ID	Type of Sampling	Sample Size	Relative Sample Size
Q1.L	Secondary	408	25,56 %
Q1.A	Secondary	408	25,56 %
Q1.V	Primary	39	2,44 %
Q1.G	Sampling not objectively possible		

Table 9: Summary of Chapter Data Collection

8. Statistical Reports Preparation

In order to perform the statistical reports, at first the appropriate statistical hypotheses (null and alternative hypotheses), tables of operationalisations and the therefore to be conducted statistical research methods are to be defined. The results of those activities (2, 3 and 4) are shown in this chapter.

8.1 Statistical Hypotheses

As a next step in the research procedure, activity 2 must be executed: Deduct alternative and null hypotheses. The result of this activity is visible in the following tables 10: Statistic Question Language, 11: Statistic Question Age and 12: Statistic Question Views on YouTube.

ID: Q1.L	**Statistic Question Language**
Does the language of a competing song have an effect on the ranking in the Eurovision Song Contest?	
ID: A1.L	**Alternative Hypothesis Language**
The Language groups of ESC songs do not have the same mean and therefore have an influence on Ranking.	
ID: N1.L	**Null Hypothesis Language**
The Language groups of ESC songs have the same mean and therefore have no influence on Ranking.	

Table 10: Statistic Question Language

ID: Q1.A	**Statistic Question Age**
Does the age of a contestant have an effect on their ranking in the Eurovision Song Contest?	
ID: A1.A	**Alternative Hypothesis Age**
There is a linear correlation between the age of a contestant and their ranking in the Eurovision Song Contest.	
ID: N1.A	**Null Hypothesis Age**
There is no linear correlation between the age of a contestant and their ranking in the Eurovision Song Contest.	

Table 11: Statistic Question Age

ID: Q1.V	Statistic Question Views on YouTube
Do the YouTube views of a competing song have an effect on the ranking in the Eurovision Song Contest?	
ID: A1.V	**Alternative Hypothesis Views on YouTube**
There is a linear correlation between the YouTube views of a competing song and the ranking in the Eurovision Song Contest.	
ID: N1.V	**Null Hypothesis Views on YouTube**
There is no linear correlation between the YouTube views of a competing song and the ranking in the Eurovision Song Contest.	

Table 12: Statistic Question Views on YouTube

8.2 Operationalisations

This subchapter holds the output of activity 3 (perform variable operationalisations). One unit of observation represents an ESC act in a year. Attributes of those units of observations are described in the following tables of operationalisations: 13 (Operationalisation Language), 14 (Operationalisation Age) and 15 (Operationalisation Views on YouTube).

Statistic question	Q1.L	
Variable	Language	Ranking
Variable type	Independent variable	Dependent variable
	Categorical	Metric
	Nominal	Interval scale
		Discrete
Manifestations / Interval	Mix, NEng, Eng	[1;44]

Table 13: Operationalisation Language

Statistic question	Q1.A	
Variable	Age	Ranking
Variable type	Independent variable	Dependent variable
	Metric	Metric
	Ratio scale	Interval scale
	Continuous	Discrete
Interval	[16;122]	[1;44]

Table 14: Operationalisation Age

Statistic question	Q1.V	
Variable	Views on YouTube	Ranking
Variable type	Independent variable	Dependent variable
	Metric	Metric
	Ratio scale	Interval scale
	Discrete	Discrete
Interval	[0;∞]	[1;44]

Table 15: Operationalisation Views on YouTube

8.3 Statistical Research Methods

In this subchapter the result of activity 4 (define appropriate statistical methods) is to be seen. All of the significance tests are compared to a significance level of 0,05.

8.3.1 Analyses of single variables

At first, the distribution, visualisation and population validity test method are defined for each of the three independent variables in the following three tables 16-18: Analysis of single variable Language, Analysis of single variable Age and Analysis of single variable Views.

Language	
Distribution	Contingency table
Visualisation	Bar plot
Population validity	Chi-square

Table 16: Analysis of single variable Language

Age		
Distribution	Central tendency	Mean, min, max
	Statistical dispersion	Spectrum, variance, standard deviation
Visualisation	Histogram	
Population validity	t-test	

Table 17: Analysis of single variable Age

Views on YouTube		
Distribution	Central tendency	Median, 1st quartile, 3rd quartile
	Statistical dispersion	Interquartile range, median absolute deviation (MAD)
Visualisation	Histogram and Boxplot	
Population validity	Wilcoxon signed-rank test	

Table 18: Analysis of single variable Views on YouTube

8.3.2 Analyses of two variables

Secondly, the statistical connection, visualisation and population validity test method are defined for each of the three statistical hypotheses (sets of variables) in the following three tables 19-21: Analysis of influence of Language on Ranking, Analysis of influence of Age on Ranking and Analysis of Views on Ranking. In addition, for the pair Views => Ranking, a regression model is calculated. On this model, its model diagnostics are analysed and a prediction method is defined as illustrated in table 21.

Language => Ranking (1 categoric and 1 metric variable)	
Statistical connection	Aggregate mean, variance analysis
Visualisation	Parallel boxplot
Population validity	ANOVA table

Table 19: Analysis of influence of Language on Ranking

Age => Ranking (2 metric variables)	
Statistical connection	Correlation: Pearson
Visualisation	Scatterplot
Population validity	Pearson's correlation coefficient
NO CORRELATION	

Table 20: Analysis of influence of Age on Ranking

Views => Ranking (2 metric variables)	
Statistical connection	Correlation: Spearman
Visualisation	Scatterplot
Population validity	Spearman’s rank correlation coefficient
CORRELATION EXISTS =>	
Model	Linear regression model
Model diagnostics	Residual plot, Q-Q plot
Prediction	Demo data, sample data 2022, linear regression model

Table 21: Analysis of influence of Views on YouTube on Ranking

9. Statistical Reports

The statistical research methods as outlined in chapter 8.3. were performed and documented in forms of statistical reports in this chapter. The chapter is split into two sub-chapters: 9.1: Analyses of single Factors (Language, Age and Views) and 9.2: Analyses of two Factors (Language and Ranking, Age and Ranking as well as Views and Ranking).

9.1 Analyses of single Factors

Before answering the statical hypothesis (cf. 8.1) for each of variable pairs (cf. 9.2), the involved independent variables and their distribution is analysed with the methods as described in chapter 8.3 as follows.

9.1.1 Language

Data fact sheet

The sample to be analysed consists of 408 observations with 5 variables.

These variables are:

- CountryYear: a String (character vector) identifying each observation
- Year: an integer identifying the year in which the contestant participated
- Country: a three character long String identifying the country the contestant is representing
- Language: a categoric variable with the following manifestations: Mix, Eng, NEng
- Ranking: an integer that informs about the achieved rank of the contestant in the interval [1;44]

In this statistic report, only the variable Language will be analysed.

It is to be analysed whether the three manifestations of Language follow a uniform distribution. Consequently, it will be analysed how well does the sample represent the whole population.

Data management

At first the data file gets imported as data frame and the variable Language will be transformed into a factor.

```
#Data Frame
dfLanguage = read.table(file = "ESCdata_Language.csv", sep = ";", head
er = TRUE)
head(dfLanguage)

##   CountryYear Year Country Language Ranking
## 1       ALB21 2021     ALB     NEng      21
## 2       AUS21 2021     AUS      Eng      34
## 3       AUT21 2021     AUT      Eng      30
## 4       AZE21 2021     AZE      Eng      20
## 5       BEL21 2021     BEL      Eng      19
## 6       BUL21 2021     BUL      Eng      11

str(dfLanguage)

## 'data.frame':    408 obs. of  5 variables:
##  $ CountryYear: chr  "ALB21" "AUS21" "AUT21" "AZE21" ...
##  $ Year       : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2
021 ...
##  $ Country    : chr  "ALB" "AUS" "AUT" "AZE" ...
##  $ Language   : chr  "NEng" "Eng" "Eng" "Eng" ...
##  $ Ranking    : int  21 34 30 20 19 11 27 16 35 28 ...

#All variables have been identified under the correct data type
automatically except for Language.
#Language will be extracted from the data frame as a vector and
factorised:

language = dfLanguage$Language
language = as.factor(language)
str(language)

##  Factor w/ 3 levels "Eng","Mix","NEng": 3 1 1 1 1 1 2 1 2 3 ...

#Looking for missing values (NA):

ok = complete.cases(dfLanguage)
sum(ok)

## [1] 408

#Creating a contingency table with absolute values:

language_abs = table(language)
language_abs

## language
##  Eng  Mix NEng
##  285   48   75
```

```
#If the contingency table is correct, its sum must equal the number of
observations
sum(language_abs)

## [1] 408

#Sorting the contingency table decreasingly:
language_abs = sort(language_abs, decreasing = TRUE)
language_abs

## language
##   Eng NEng  Mix
##   285   75   48
```

Visualisation

```
#install.packages("wesanderson")
library(wesanderson)
barplot(language_abs, main="Distribution of ESC song languages between
2011-2021 (N=408)", col=wes_palette("Moonrise3"), ylim = c(0,300))

#In a uniform distribution, each category manifestation would be
408/3=136 as illustrated in the barplot with a red horizontal line.
abline(h=(408/3), col="red")
```

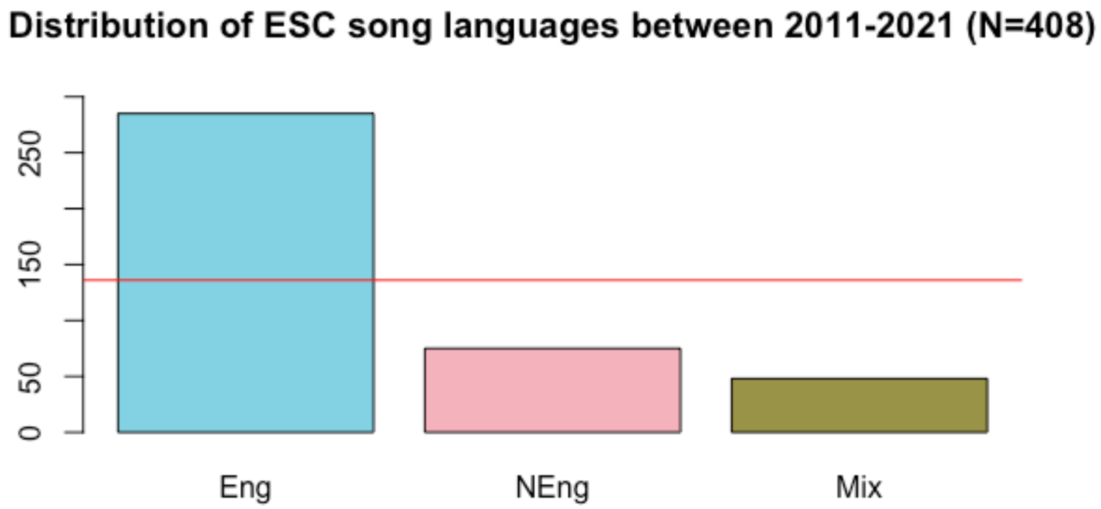


Figure 17: Distribution of ESC song languages between 2011-2021

As visible from figure 17: Distribution of ESC song languages between 2011-2021, more than 2/3 of the songs among the 408 observations, are in English (Eng). Followed by songs in any other language than English (NEng) with only 75. Lastly, songs with a mix of languages account for less than 50. Apparently, English is more popular than expected.

Interim conclusion: The sample does not show a uniform distribution among the three manifestations because English is significantly above the red line (mean) whereas NEng and Mix are under.

Population validity

In order to analyse the validity of the result within the whole population, a chi-squared test on uniform distribution with a significance level of 0.05 is performed as follows.

```
#This R function takes the absolute contingency table as argument.

chisq.test(language_abs)

##
##  Chi-squared test for given probabilities
##
## data:  language_abs
## X-squared = 247.54, df = 2, p-value < 2.2e-16
```

The null hypotheses says: The factor Language is distributed uniformly. (Alternative hypotheses: is not distributed uniformly).

The p-value of approximately 0 (with test statistic of 247.54) is clearly under the defined significance level of 0.05. Therefore the null hypothesis can be dismissed.

Final conclusion: *The data indicates that the factor Language is not distributed uniformly.*

9.1.2 Age

Data fact sheet

The sample to be analysed consists of 408 observations with 5 variables.

These variables are:

- CountryYear: a String (character vector) identifying each observation
- Year: an integer identifying the year in which the contestant participated
- Country: a three character long String identifying the country the contestant is representing
- Age: a metric variable within the following interval: [16;122]. The Age is theoretically a continuous value but it is only measured in whole years for this analysis. The Age has a ratio scale with the natural zero point being the birth.
- Ranking: an integer that informs about the achieved rank of the contestant in the interval [1;44]

In this statistic report, only the variable Age will be analysed.

The analysis comprises of the distribution, the indicators of the central tendency and statistical dispersion. Following, a t-test/Wilcoxon test will be used to investigate a potential deviation of the mean/median among the whole population.

Data management

At first the data file gets imported as data frame and the variable Age must be of data type integer.

```
#Data Frame
dfAge = read.table(file = "ESCdata_Age.csv", sep = ";", header = TRUE)
head(dfAge)

##   CountryYear Year Country Age Ranking
## 1       ALB21 2021     ALB  35      21
## 2       AUS21 2021     AUS  25      34
## 3       AUT21 2021     AUT  35      30
## 4       AZE21 2021     AZE  30      20
## 5       BEL21 2021     BEL  41      19
## 6       BUL21 2021     BUL  23      11

str(dfAge)
```

```
## 'data.frame':    408 obs. of  5 variables:
##  $ CountryYear: chr  "ALB21" "AUS21" "AUT21" "AZE21" ...
##  $ Year       : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2
021 ...
##  $ Country    : chr  "ALB" "AUS" "AUT" "AZE" ...
##  $ Age        : int  35 25 35 30 41 23 22 26 33 32 ...
##  $ Ranking    : int  21 34 30 20 19 11 27 16 35 28 ...

#All variables have been identified under the correct data type
automatically. Age is of type integer as desired.
#For easier handling, age will be extracted from the data frame as a
vector:

age = dfAge$Age
str(age)

##  int [1:408] 35 25 35 30 41 23 22 26 33 32 ...

#Looking for missing values (NA):

ok = complete.cases(dfAge)
sum(ok)

## [1] 407

#There is one observation with a missing value.
grep(FALSE,ok)

## [1] 327

ok[327]

## [1] FALSE

age = age[-327]
#The NA value is now deleted.
```

Visualisation

In order to analyse the distribution of the metric variable Age its density will be visualised in a form of a histogram:

```
hist(age, freq = TRUE, xlim = c(10,80), ylim=c(0,120) ,labels = TRUE,
main="Histogram of ESC participants' Age from 2011-2021 (n=407)", xlab
= "Age", col=rainbow(13))
#The function rug illustrates every single value as a line under the
histogram.
rug(age)
#Now the mean (red) as well as the median (blue) are shown as vertical
lines.
```

```
abline(v=mean(age),col="red")
abline(v=median(age),col="blue")
```

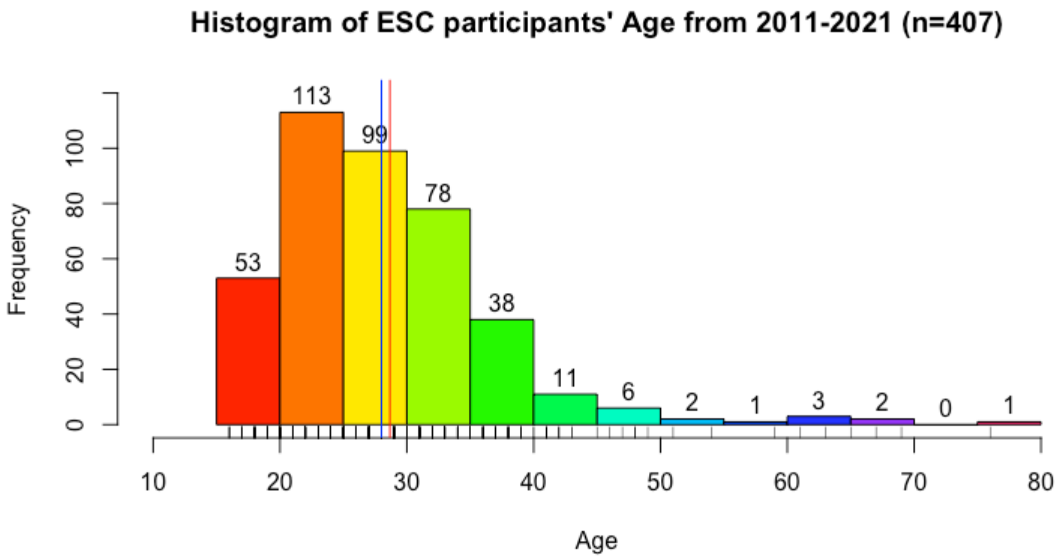


Figure 18: Histogram of ESC participant's age from 2011-2021

Figure 18 (Histogram of ESC participant’s age from 2011-2021) shows: The data distribution is right skewed. The most frequent Age is between 20 and 25 (113 appearances). The second most frequent age is between 25 and 30 (99), followed by 30 to 35 (78). 53 participants are between the minimum age (16) and 20. Not many participants are older than 50. The mean (red) is approximately 29 whereas the median (blue) is ca. 28. These values are close together, indicating robust data.

Central tendency and statistical dispersion

```
#The 5-point summary (minimum, 1st quartile, median, 3rd quartile and
maximum):
summary(age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   16.00   23.00   28.00   28.66   33.00   76.00

#Spectrum:
diff(range(age))

## [1] 60
```

```
#Variance
var(age)

## [1] 71.38149

#Standard deviation
sd(age)

## [1] 8.448757
```

The minimum age is 16 and the maximum age is 76, giving us a spectrum of 60. The first quartile is 23, the median 28 and the third quartile 33. The mean of 28.66 is as seen from the histogram, close to the median (28), making the mean a robust and trustworthy indicator of the data's distribution.

The values spread with a standard deviation of ±8.4 around the mean.

Population validity

Due to the robust data, a t-test will be favourited over the Wilcoxon test for the evaluation. Now it is to be analysed to what extend the mean of 29 (28.66) can be true within the total population.

The null hypothesis is: "The true mean is 29 years." The alternative hypothesis is: "The true mean is not 29 years."

```
t.test(age, mu=28.66)

##
##  One Sample t-test
##
## data:  age
## t = 0.0080963, df = 406, p-value = 0.9935
## alternative hypothesis: true mean is not equal to 28.66
## 95 percent confidence interval:
##  27.84012 29.48666
## sample estimates:
## mean of x
##  28.66339
```

Final conclusion: *The p-value of 0.9935 (test statistic of 0.0080963) is above the significance level of 0.05. Therefore the null hypothesis cannot be dismissed. The data does not contradict that the true mean of the total population might deviate from 29. The 95% confidence interval is between 27.8 and 29.5 and contains the reference value (28.66).*

9.1.3 Views

Data fact sheet

The sample to be analysed consists of 39 observations with 5 variables.

These variables are:

- CountryYear: a String (character vector) identifying each observation
- Year: an integer identifying the year in which the contestant participated
- Country: a three character long String identifying the country the contestant is representing
- Views: a metric variable within the following interval: [0;∞]. Views is a discrete value since YouTube does not count half views. Views has a ratio scale with the natural zero point being 0 views.
- Ranking: an integer that informs about the achieved rank of the contestant in the interval [1;44]

In this statistic report, only the variable Views will be analysed.

The analysis comprises of the distribution, the indicators of the central tendency and statistical dispersion. Following, a t-test/Wilcoxon test will be used to investigate a potential deviation of the mean/median among the whole population.

Data management

At first the data file gets imported as data frame and the variable Views must be of data type integer.

```
#Data Frame
dfViews = read.table(file = "ESCdata_Views.csv", sep = ";", header = T
RUE)
head(dfViews)

##   CountryYear Year Country   Views Ranking
## 1       ALB21 2021     ALB  980000      21
## 2       AUS21 2021     AUS 1100000      34
## 3       AUT21 2021     AUT  706000      30
## 4       AZE21 2021     AZE 4900000      20
## 5       BEL21 2021     BEL 1100000      19
## 6       BUL21 2021     BUL  780000      11

str(dfViews)
```

```
## 'data.frame':    39 obs. of  5 variables:
##  $ CountryYear: chr  "ALB21" "AUS21" "AUT21" "AZE21" ...
##  $ Year       : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2
021 ...
##  $ Country    : chr  "ALB" "AUS" "AUT" "AZE" ...
##  $ Views      : int  980000 1100000 706000 4900000 1100000 780000 2
500000 2400000 827000 620000 ...
##  $ Ranking    : int  21 34 30 20 19 11 27 16 35 28 ...

#All variables have been identified under the correct data type
automatically. Views is of type integer as desired.
#For easier handling, Views will be extracted from the data frame as a
vector:

views = dfViews$Views
str(views)

##  int [1:39] 980000 1100000 706000 4900000 1100000 780000 2500000 24
00000 827000 620000 ...

#Looking for missing values (NA):

ok = complete.cases(dfViews)
sum(ok)

## [1] 39

#There are no missing values.
```

Visualisation

In order to analyse the distribution of the metric variable Views its density will be visualised in a form of a histogram. In order to make the x axis easier to read, the views will be divided by a million.

```
viewsMio=views/1000000
viewsMio=round(viewsMio,2)
head(viewsMio)

## [1] 0.98 1.10 0.71 4.90 1.10 0.78

#install.packages("wesanderson")
library(wesanderson)
hist(viewsMio, freq = TRUE,labels = TRUE, main="Histogram of views of
ESC competing songs on YouTube 2021 (n=39)", xlab = "Views in Mio.", c
ol=wes_palette("Moonrise3"),ylim=c(0,40))
#The function rug illustrates every single value as a line under the
histogram.
rug(viewsMio)
```

```
#Now the mean (red) as well as the median (blue) are shown as vertical lines.
abline(v=mean(viewsMio),col="red")
abline(v=median(viewsMio),col="blue")
```

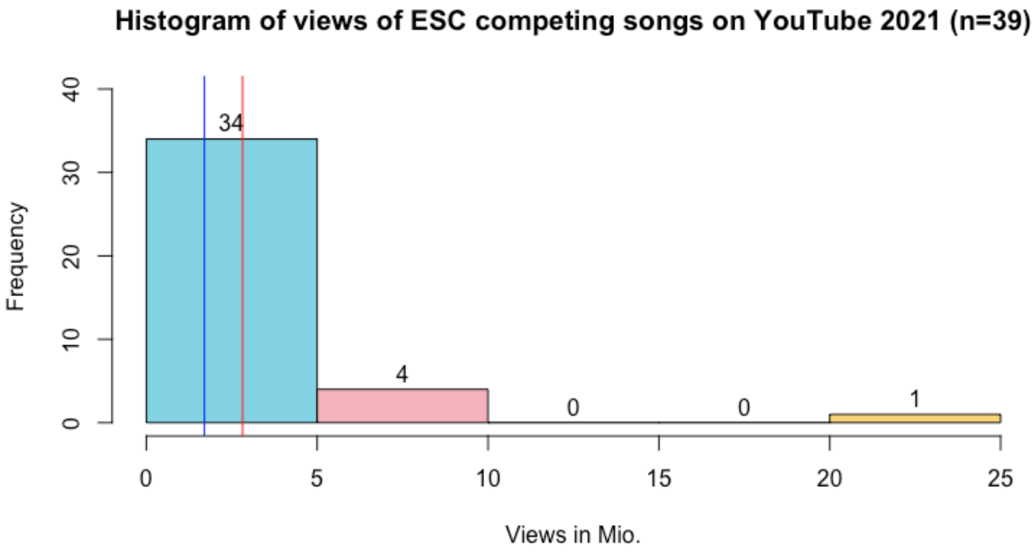


Figure 19: Histogram of views of ESC competing songs on YouTube 2021

Figure 19 (Histogram of views of ESC competing songs on YouTube 2021) shows:
The data distribution is skew right. Out of 39 songs, 34 reached less than 5 Mio views. Four songs reached between 5 and 10 Mio views, leaving one song with more than 20 Mio views. The mean (red) is approximately 3 Mio whereas the median (blue) is approx. 2 Mio. These values are relatively close together. However, it can be clearly seen that the mean is moved to the right due to the one outlier.

Central tendency and statistical dispersion

```
#The 5-point summary (minimum, 1st quartile, median, 3rd quartile and maximum):
summary(viewsMio)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.520   1.050   1.700   2.816   2.450  21.000

#Interquartile range
IQR(viewsMio)

## [1] 1.4
```

```
#Median absolute deviation (MAD)
mad(viewsMio)

## [1] 1.03782
```

The minimum amount of Views is 0.52 Mio and the maximum amount is 21 Mio. The first quartile is 1.05 Mio, the median 1.7 Mio and the third quartile 2.45 Mio. The mean of 2.816 Mio is clearly deviating from the median (1.7 Mio), making the median a more trustworthy indicator than the mean for the data's distribution. Therefore, instead of the spectrum one calculates the interquartile range and instead of the standard deviation (and variance) the median absolute deviation (MAD). The data spreads between the IQR of 1.4 Mio. (2.450-1.050). The values spread with a MAD of ±1.04 around the median.

Due to the non-robust data, a boxplot that behaves less sensitive to outliers than the histogram, follows as visualisation in addition to the histogram.

```
boxplot(viewsMio, horizontal = TRUE)
abline(v=mean(viewsMio),col="red")
```

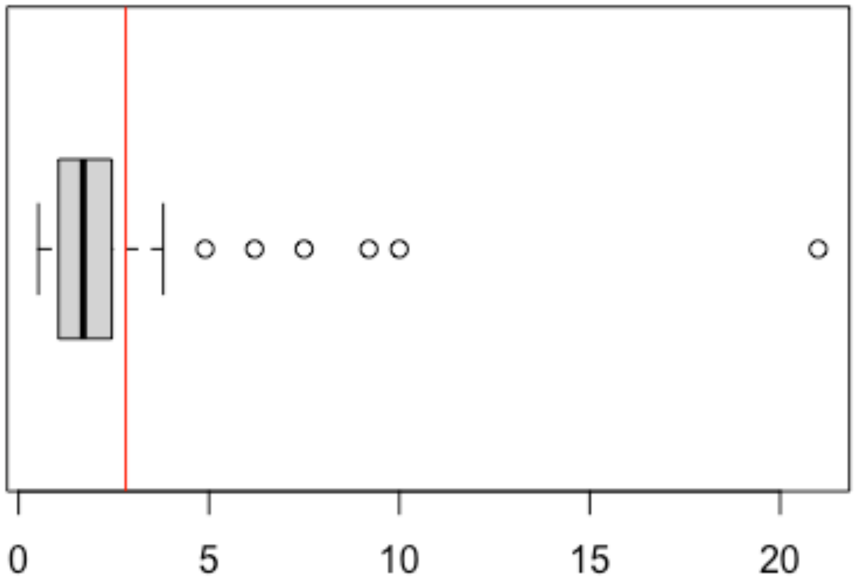


Figure 20: Boxplot of Views in Mio

Figure 20 (Boxplot of Views in Mio) shows:
Also the boxplot shows that the data is right skewed as the median (thick black line) is not perfectly centred but moved slightly to the left. The mean (red) is outside the interquartile range, confirming the assumption of non-robust data. In addition, six outliers are outside the inner fence with one extreme case.

Population validity

Due to the lack of robust data, the Wilcoxon test will be favourited over the t-test for the evaluation. Now it is to be analysed to what extend the median of 1.7 Mio can be true within the whole population.

The null hypothesis is: “The true median is 1.7 Mio views.”

The alternative hypothesis is: “The true median is not 1.7 Mio views.”

```
wilcox.test(viewsMio, mu = 1.7, conf.int = TRUE, conf.level = 0.95)

## Warning in wilcox.test.default(viewsMio, mu = 1.7, conf.int = TRUE,
conf.level =
## 0.95): cannot compute exact p-value with ties

## Warning in wilcox.test.default(viewsMio, mu = 1.7, conf.int = TRUE,
conf.level =
## 0.95): cannot compute exact confidence interval with ties

## Warning in wilcox.test.default(viewsMio, mu = 1.7, conf.int = TRUE,
conf.level =
## 0.95): cannot compute exact p-value with zeroes

## Warning in wilcox.test.default(viewsMio, mu = 1.7, conf.int = TRUE,
conf.level =
## 0.95): cannot compute exact confidence interval with zeroes

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  viewsMio
## V = 399.5, p-value = 0.6791
## alternative hypothesis: true location is not equal to 1.7
## 95 percent confidence interval:
##  1.400026 2.649933
## sample estimates:
## (pseudo)median
##       1.800067
```

Final conclusion: *The p-value of 0.6791 is above the significance level of 0.05. Therefore the null hypothesis cannot be dismissed. The data does not contradict that the true median of the total population might deviate from 1.7 Mio. The 95% confidence interval is between 1.4 and 2.7 Mio and contains the reference value (1.7). For the interpretation of the Wilcoxon test results one is to keep in mind that because ties and zeroes exist, the p-value as well as the confidence interval might be inaccurate.*

9.2 Analyses of two Factors

Now that the single factors have been described with statistical reports, this chapter holds the actual answers of the three hypotheses from chapter 8.1.

9.2.1 Language and Ranking

Data fact sheet

The sample to be analysed consists of 408 observations with 5 variables.

These variables are:

- CountryYear: a String (character vector) identifying each observation
- Year: an integer identifying the year in which the contestant participated
- Country: a three character long String identifying the country the contestant is representing
- Language: a categoric variable with the following manifestations: Mix, Eng, NEng
- Ranking: an integer that informs about the achieved rank of the contestant in the interval [1;44]

In this statistic report, the variables Language (categoric) and Ranking (metric) will be analysed.

It is to be analysed whether the Language has an influence on the Ranking. Given the case that there is an influence, it will be consequently analysed whether this influence is valid only in the sample size or also within the total population.

Data management

At first the data file gets imported as data frame and the variable Language transformed into a factor.

```
#Data Frame
dfQ1L = read.table(file = "ESCdata_Language.csv", sep = ";", header =
TRUE)
head(dfQ1L)

##   CountryYear Year Country Language Ranking
## 1       ALB21 2021     ALB     NEng      21
## 2       AUS21 2021     AUS      Eng      34
## 3       AUT21 2021     AUT      Eng      30
## 4       AZE21 2021     AZE      Eng      20
## 5       BEL21 2021     BEL      Eng      19
## 6       BUL21 2021     BUL      Eng      11
```

```
str(dfQ1L)

## 'data.frame':    408 obs. of  5 variables:
##  $ CountryYear: chr  "ALB21" "AUS21" "AUT21" "AZE21" ...
##  $ Year       : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2
021 ...
##  $ Country    : chr  "ALB" "AUS" "AUT" "AZE" ...
##  $ Language   : chr  "NEng" "Eng" "Eng" "Eng" ...
##  $ Ranking    : int  21 34 30 20 19 11 27 16 35 28 ...

#All variables have been identified under the correct data type
automatically except for Language.
#The variable Language must be factorised.

dfQ1L$Language = as.factor(dfQ1L$Language)
str(dfQ1L)

## 'data.frame':    408 obs. of  5 variables:
##  $ CountryYear: chr  "ALB21" "AUS21" "AUT21" "AZE21" ...
##  $ Year       : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2
021 ...
##  $ Country    : chr  "ALB" "AUS" "AUT" "AZE" ...
##  $ Language   : Factor w/ 3 levels "Eng","Mix","NEng": 3 1 1 1 1 1
2 1 2 3 ...
##  $ Ranking    : int  21 34 30 20 19 11 27 16 35 28 ...

#Looking for missing values (NA):

ok = complete.cases(dfQ1L)
sum(ok)

## [1] 408

#There are no missing values.

#ANOVA table

meanByLanguage = aggregate(Ranking ~ Language, data = dfQ1L, mean)
meanByLanguage = meanByLanguage[order(meanByLanguage$Ranking),]
```

Whereas Eng and NEng lead to a similar ranking (21 and 22 respectively), Mix seems to have a noticeable lowering effect on the ranking (17).

Visualisation

In order to analyse a potential effect of the Language on the Ranking, a parallel boxplot illustrates grouped metric data of Ranking. It shows the distribution of the Ranking grouped by Language types.

```
#library("vcd")
#install.packages("wesanderson")
library(wesanderson)
boxplot(Ranking ~ Language, data = dfQ1L, main="Ranking in dependence
of ESC songs' Language between 2011-2021 (n=408)", col=wes_palette("Mo
onrise3"))
#Adding the median for all three groups:
abline(h = median(dfQ1L$Ranking), col = "red")
```

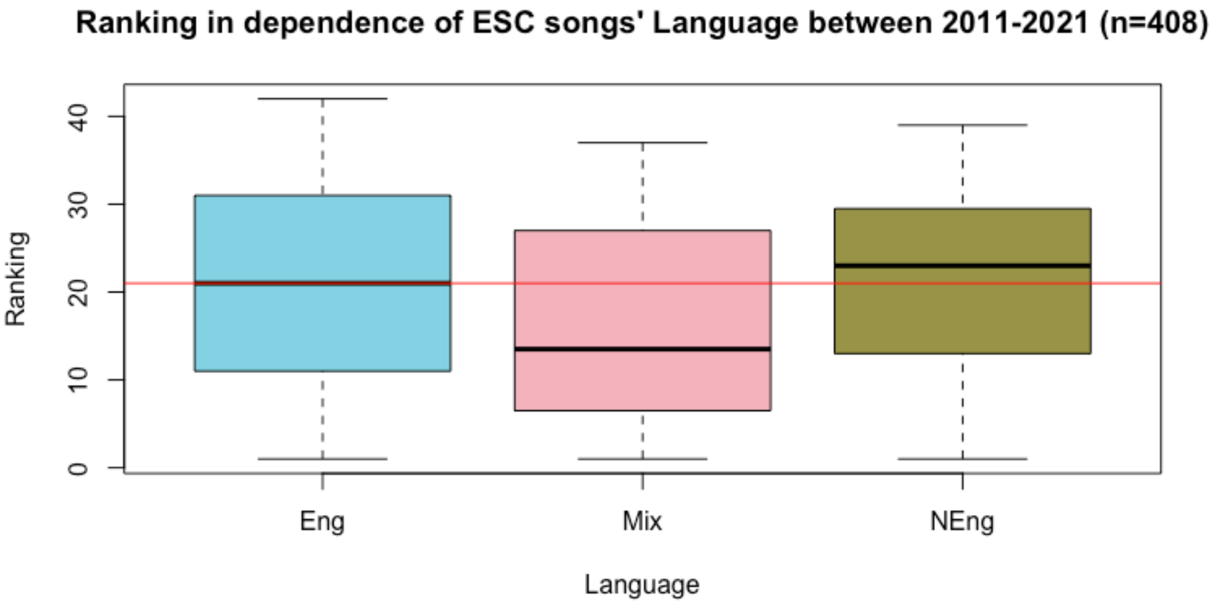


Figure 21: Ranking in dependence of ESC songs' language between 2011-2021

Figure 21 (Ranking in dependence of ESC songs' language between 2011-2021) shows:

The interquartile ranges of Eng and NEng are between ca. 10 and 30 whereas the interquartile range of Mix is slightly below (ca. 8 to 28). Eng is spread in a symmetric matter (median is in the middle of the box), Mix is right skewed (median is moved to the left) and NEng is slightly left skewed (median is moved slightly to the right). Eng does not seem to have an effect on the Ranking as its group median (thick, black) matches the total median (red). NEng seems to have a very small increasing effect on the ranking as its group median is slightly above the total median. Mix as the only one of the three groups seems to have a worth mentioning rank-lowering effect as its group median is noticeably under the total median.

Population validity

Now it is to be analysed whether the effects we saw on the parallel boxplot are also reflected in the variances (F-value) and whether there are just a coincidence in our sample or are also true within the total population. This can be done by performing an Analysis of Variances (ANOVA table).

The null hypothesis is: “The Language groups have the same mean and therefore have no influence on Ranking.”

```
summary(aov(Ranking ~ Language, data = dfQ1L))

##              Df Sum Sq Mean Sq F value Pr(>F)
## Language      2    880   440.2   3.413 0.0339 *
## Residuals   405  52232   129.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-value is not 1, showing a slight but noticeable influence of 3.413 of Language on the Ranking. This influence is significant on the 0.05 level (0.0339). The null hypothesis may be dismissed. The data indicates that Language groups can have an effect on the Ranking.

9.2.2 Age and Ranking

Data fact sheet

The sample to be analysed consists of 408 observations with 5 variables.

These variables are:

- CountryYear: a String (character vector) identifying each observation
- Year: an integer identifying the year in which the contestant participated
- Country: a three character long String identifying the country the contestant is representing
- Age: a metric variable within the following interval: [16;122]. The Age is theoretically a continuous value but it is only measured in whole years for this analysis. The Age has a ratio scale with the natural zero point being the birth.
- Ranking: an integer that informs about the achieved rank of the contestant in the interval [1;44]

In this statistical report, the variables Age (independent variable) and Ranking (dependent variable) are relevant. It is to be analysed whether there is a correlation between those two variables. If there is a significant correlation (on significance level 0.05), regression and prediction will be calculated in addition.

Data management

At first the data file gets imported as data frame and the variables Age and Ranking must be of data type integer.

```
#Data Frame
dfQ1A = read.table(file = "ESCdata_Age.csv", sep = ";", header = TRUE)
head(dfQ1A)
```

```
##   CountryYear Year Country Age Ranking
## 1       ALB21 2021     ALB  35      21
## 2       AUS21 2021     AUS  25      34
## 3       AUT21 2021     AUT  35      30
## 4       AZE21 2021     AZE  30      20
## 5       BEL21 2021     BEL  41      19
## 6       BUL21 2021     BUL  23      11
```

```
str(dfQ1A)
```

```
## 'data.frame':    408 obs. of  5 variables:
##  $ CountryYear: chr  "ALB21" "AUS21" "AUT21" "AZE21" ...
##  $ Year       : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2
021 ...
```

```
##  $ Country    : chr  "ALB" "AUS" "AUT" "AZE" ...
##  $ Age        : int  35 25 35 30 41 23 22 26 33 32 ...
##  $ Ranking    : int  21 34 30 20 19 11 27 16 35 28 ...

#All variables have been identified under the correct data type
automatically. Age and Ranking are of type integer as desired.

#Looking for missing values (NA):

ok = complete.cases(dfQ1A)
sum(ok)

## [1] 407

#There is one observation with a missing value.
grep(FALSE,ok)

## [1] 327

ok[327]

## [1] FALSE

dfQ1A = dfQ1A[-327,]
#The NA value is now deleted.
```

Visualisation

In order to analyse a possible correlation between the two metric variables a visualisation of a scatterplot that shows the Age (independent, x axis) in dependence of Ranking (dependent, y axis) will be made.

```
#library("vcd")
plot(Ranking ~ Age, data = dfQ1A, main="Age in dependence of Ranking (
n=407)")
```

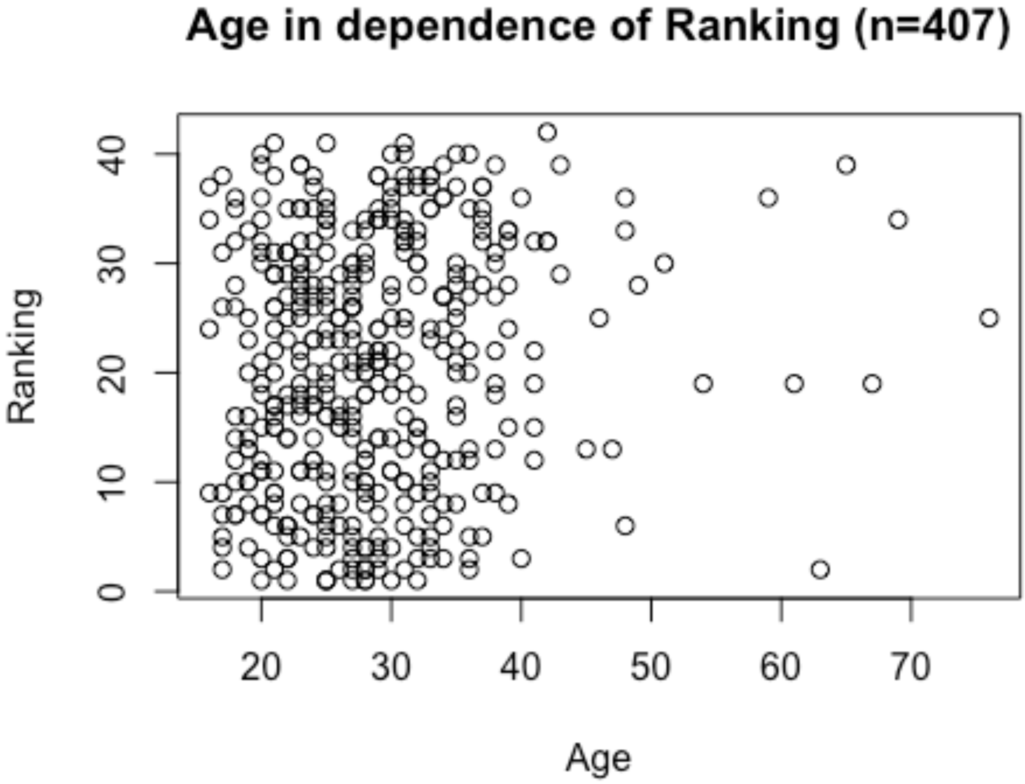


Figure 22: Age in dependence of ranking

According to the scatterplot from figure 22 one can assume that there is no correlation. It is clearly visible that most of the data points are under the age of 40, reaching values of the whole range of Ranking.

Correlation and hypothesis test

In order to gain more certainty whether there is a correlation or not and how relevant it is, the correlation coefficient is calculated as follows.

```
#Since the metric variables are robust, the Pearson’s correlation coefficient is calculated

with(dfQ1A, cor(Ranking, Age))

## [1] 0.1343454
```

The calculated correlation coefficient indicates that there is no correlation. Therefore, regression and prediction cannot be computed in a meaningful way.

9.2.3 Views and Ranking

Data fact sheet

The sample to be analysed consists of 39 observations with 5 variables.

- These variables are: CountryYear: a String (character vector) identifying each observation
- Year: an integer identifying the year in which the contestant participated
- Country: a three character long String identifying the country the contestant is representing
- Views: a metric variable within the following interval: [0;∞]. Views is a discrete value since YouTube does not count half views. Views has a ratio scale with the natural zero point being 0 views.
- Ranking: an integer that informs about the achieved rank of the contestant in the interval [1;44]

In this statistical report, the variables Views (independent variable) and Ranking (dependent variable) are relevant. It is to be analysed whether there is a correlation between those two variables. If there is a significant correlation (on significance level 0.05), regression and prediction are to be calculated in addition.

Data management

At first the data file gets imported as data frame and the variables Views and Ranking must be of data type integer.

```
#Data Frame
dfQ1V = read.table(file = "ESCdata_Views.csv", sep = ";", header = TRU
E)
head(dfQ1V)

##   CountryYear Year Country   Views Ranking
## 1       ALB21 2021     ALB  980000      21
## 2       AUS21 2021     AUS 1100000      34
## 3       AUT21 2021     AUT  706000      30
## 4       AZE21 2021     AZE 4900000      20
## 5       BEL21 2021     BEL 1100000      19
## 6       BUL21 2021     BUL  780000      11

str(dfQ1V)

## 'data.frame':    39 obs. of  5 variables:
##  $ CountryYear: chr  "ALB21" "AUS21" "AUT21" "AZE21" ...
##  $ Year       : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2
021 ...
##  $ Country    : chr  "ALB" "AUS" "AUT" "AZE" ...
##  $ Views      : int  980000 1100000 706000 4900000 1100000 780000 2
500000 2400000 827000 620000 ...
##  $ Ranking    : int  21 34 30 20 19 11 27 16 35 28 ...

#All variables have been identified under the correct data type
automatically. Views and Ranking are of type integer as desired.

#Looking for missing values (NA):

ok = complete.cases(dfQ1V)
sum(ok)

## [1] 39

#There are no missing values.
```

Correlation

Visualisation

In order to analyse a possible correlation between the two metric variables one visualises both together in a scatterplot that shows the Age (independent, x axis) in dependence of Ranking (dependent, y axis).

```
#library("vcd")
#Before creating the scatterplot, the variable Views will be divided
by a million for easier reading.
dfQ1VInMio = dfQ1V
dfQ1VInMio$Views = dfQ1VInMio$Views/1000000
dfQ1VInMio$Views = round(dfQ1VInMio$Views,2)
head(dfQ1VInMio)

##   CountryYear Year Country Views Ranking
## 1       ALB21 2021     ALB  0.98      21
## 2       AUS21 2021     AUS  1.10      34
## 3       AUT21 2021     AUT  0.71      30
## 4       AZE21 2021     AZE  4.90      20
## 5       BEL21 2021     BEL  1.10      19
## 6       BUL21 2021     BUL  0.78      11

plot(Ranking ~ Views, data = dfQ1VInMio, main="Views in Mio in depende
nce of Ranking (n=39)")
```

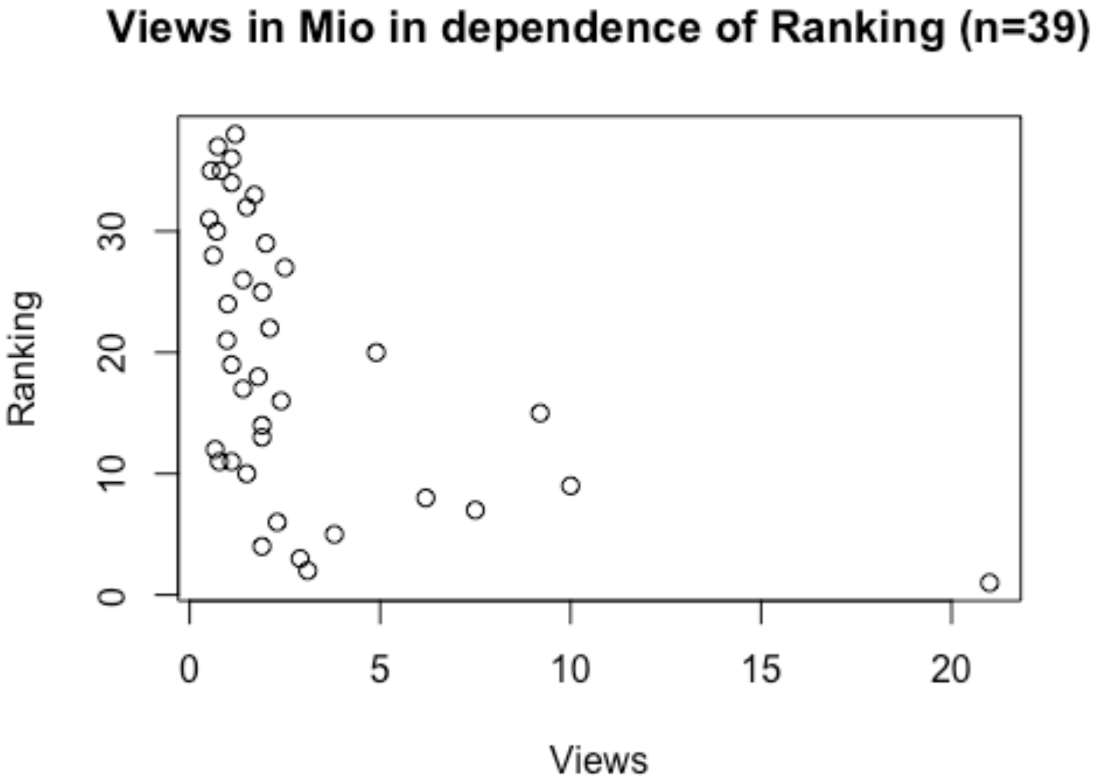


Figure 23: Views in Mio in dependence of ranking

According to the scatterplot in figure 23 one can assume that there is a weak negative correlation. Further, it is visible that the majority of the songs have less than 5 Mio Views, whereas the remainder spreads between 5 and 10 Mio Views with one outlier above 20 Mio Views.

Correlation and hypothesis test

In order to gain more certainty whether there is a correlation or not and how relevant it is, the correlation coefficient is calculated as follows.

```
#Since the metric variable are not robust, the Spearman’s rank
correlation coefficient is calculated:

with(dfQ1V, cor(Ranking,Views, method="spearman"))

## [1] -0.6086849
```

The calculated correlation coefficient “rho” indicates that there is a moderate negative correlation of -0.61.

Population validity

In order to analyse the validity of the correlation coefficient within the total population the null hypothesis “There is no linear correlation.” will be tested with p-value of the method of Spearman.

```
cor.test(~ Ranking + Views, data = dfQ1V, method = "spearman")

## Warning in cor.test.default(x = c(21L, 34L, 30L, 20L, 19L, 11L, 27L
, 16L, :
## Cannot compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  Ranking and Views
## S = 15894, p-value = 3.933e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.6086849
```

The p-value of almost zero dismisses the null hypothesis on the significance level of 0.05. Therefore the data indicates that the alternative hypothesis “There is a linear correlation.” is true also within the total population. The test result makes one aware of the fact that there are ties, meaning that the p-value might be inaccurate.

Regression

Estimation of linear model

```
#The R-function lm creates a linear model in the form: y = a + bx

model = lm(Ranking ~ Views, data = dfQ1V)
model

##
## Call:
## lm(formula = Ranking ~ Views, data = dfQ1V)
##
## Coefficients:
## (Intercept)        Views
##   2.369e+01   -1.457e-06

round(coef(model), 9)

##  (Intercept)        Views
## 23.692108445 -0.000001457
```

The coefficients are:

- Intercept: 23.692
- Coefficient of Views: -0.000001457

Therefore, the linear regression model is:

Ranking = 23.692 - 0.000001457 * Views

This means that participants reach the 24th rank on average. Considering an amount of total participants of about 40 per year, that makes sense. Every one million Views decreases the rank by about 1.5.

```
#The model overview looks as follows:

summary(model)

##
## Call:
## lm(formula = Ranking ~ Views, data = dfQ1V)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.175  -7.423   1.368   7.988  16.057
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.369e+01  2.021e+00  11.722 5.09e-14 ***
## Views       -1.457e-06  4.346e-07  -3.353  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.05 on 37 degrees of freedom
## Multiple R-squared:  0.233,  Adjusted R-squared:  0.2123
## F-statistic: 11.24 on 1 and 37 DF,  p-value: 0.001854
```

The intercept as well as the coefficient of Views are significant on the 0.05 significance level (p values of 5.09e-14 and 0.00185). The coefficient of determination (multiple R-squared) is 0.233, meaning that the model explains 23.3% of the variance of Ranking. The p-value of the F-statistic on the whole linear model is 0.001854 which is significant on the 0.05 level.

```
# Following, the 95% confidence intervals for the parameters are
calculated:

round(confint(model), 9)

##                    2.5 %       97.5 %
## (Intercept) 19.596889948 27.787326943
## Views       -0.000002338 -0.000000577
```

The confidence intervals are:

- for the Ranking (intercept): [19.6;27.8]
- for the Views: [-0.000002338;-0.000000577]

It is to be noted that the Views confidence interval covers only negative numbers, meaning that there is in 95% of data always a positive effect on a participants rank by YouTube views.

Visualisation: Improved Scatterplot

In order to visualise the confidence intervals, at first, demo-data must be created and saved in a data frame.

```
tmp = seq(from = min(dfQ1V$Views), to = max(dfQ1V$Views), length = 100
)
head(tmp)

## [1]  521000.0  727858.6  934717.2 1141575.8 1348434.3 1555292.9

demodata = data.frame(Views = tmp)
head(demodata)

##       Views
## 1  521000.0
## 2  727858.6
## 3  934717.2
## 4 1141575.8
## 5 1348434.3
## 6 1555292.9

# The demo data is created
```

Now the improved scatterplot including regression line (red) and confidence bands (green) may be created.

```
conf = predict(model, demodata, interval = "confidence")

plot(Ranking ~ Views, data = dfQ1V, main="Views in dependence of Ranki
ng (n=39)")


#Adding regression line
abline(model, col = "red")

lines(tmp, conf[,"lwr"], col = "#336633") #color green
lines(tmp, conf[,"upr"], col = "#336633")
```

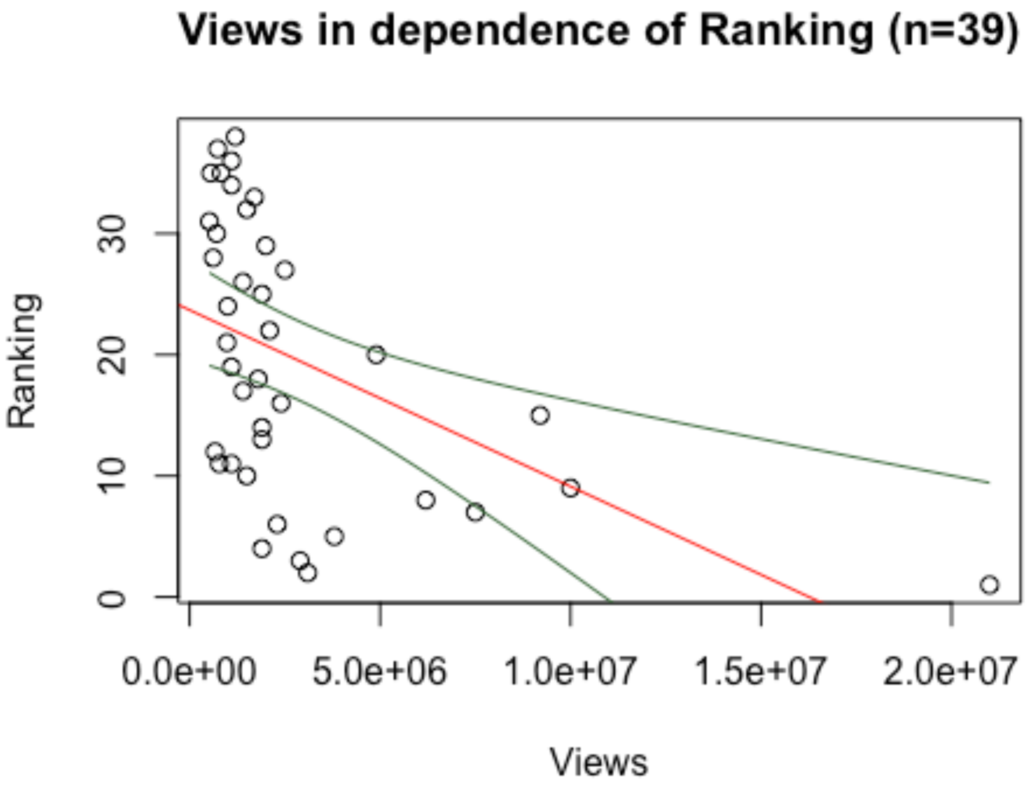


Figure 24: Views in dependence of ranking

The scatterplot from figure 24 shows the slightly falling regression line (red). It can also be seen that a participant needs at least 16 Mio Views to reach the first rank according to the regression model.

Prediction (general)

Model diagnostics

Before predicting, the model diagnostics must be analysed with a residual plot and a Q-Q plot.

```
#The following plot visualises the residuals on the estimated values.
plot(residuals(model) ~ fitted(model))
```

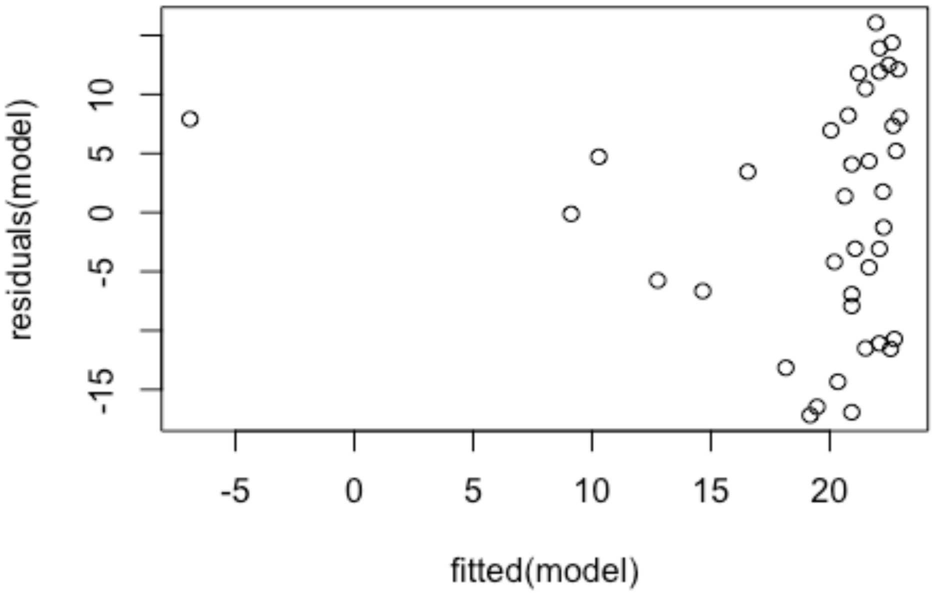


Figure 25: Residual Q-Q Plot

The plot in figure 25 does not show any noticeable problems.

```
#The following plot visualises the uniform distribution of the residuals via Q-Q plot.
qqnorm(residuals(model))
qqline(residuals(model))
```

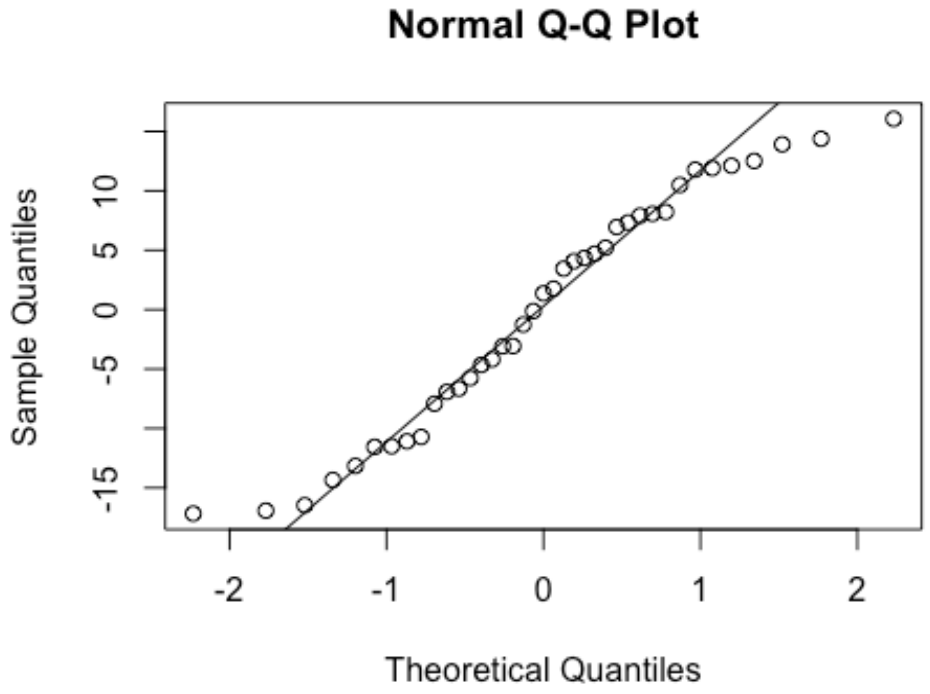


Figure 26: Normal Q-Q Plot

Figure 26 (Normal Q-Q Plot) shows:
It can be seen that the data points follow the line. However, there are exceptions, mostly on the upper edge. The results of the significance tests may be considered but should be interpreted with care.

Summarising, there were no major problems in the model diagnostics to be found. Therefore, we may proceed now with the prediction.

Prediction

Predictions as shown in the following data frame will be made.

```
newdata = data.frame(Views = c(600000,1150000,3500000,17000000))
newdata

##      Views
## 1   600000
## 2  1150000
## 3  3500000
## 4 17000000
```

The prognosis for the expected Rankings for the given Views and their confidence intervals (95%) are calculated as follows.

```
predict(model, newdata, interval = "confidence")

##         fit       lwr      upr
## 1 22.817770  19.01903 26.61651
## 2 22.016294  18.44201 25.59057
## 3 18.591802  15.27684 21.90677
## 4 -1.080807 -13.99007 11.82846
```

The predictions are:

- For 600 000 Views: Rank 23. Confidence interval: [19;27]
- For 1 150 000 Views: Rank 23. Confidence interval: [18;26]
- For 3 500 000 Views: Rank 19. Confidence interval: [15;22]
- For 17 000 000 Views: Rank -1. Confidence interval: [-14;12].

The valid values for the variable Ranking were defined in the beginning and are within the interval: [1;44]. Therefore a more meaningful prediction for 17 Mio Views is: Rank 1 with confidence interval [1;12].

Prediction (2022)

Predictions for the participants of 2022 (data retrieved in April 2022) as shown in the following data frame will be made.

```
dataESC22 = read.table("ESCdata_Views2022.csv", header=TRUE, sep=";")
head(dataESC22)

##   Country   Views
## 1     ALB 2924447
## 2     ARM  977794
## 3     AUS  532362
## 4     AUT 1759357
## 5     AZE  587486
## 6     BEL  769348

#The data consists of 40 observations and is complete. The column
Views is of data type integer as desired.

#Create a vector Ranking that uses the prediction model to predict
Rankings based on Views as can be found in the data frame dataESC22
columns 2: Views (R knows automatically which column to take).


Ranking = predict(model, dataESC22)

#Bind (=add) column Ranking to dataESC22
dataESC22 = cbind(dataESC22, Ranking)

#Sort data frame by column Ranking
dataESC22 = dataESC22[order(dataESC22$Ranking),]

#Round column Ranking to whole numbers
dataESC22$Ranking = round(dataESC22$Ranking,0)
dataESC22

##    Country    Views Ranking
## 21     ITA 47038608     -45
## 34     SRB 12331431       6
## 39     UKR  4838643      17
## 25     MDA  3987846      18
## 36     ESP  3720756      18
## 29     NOR  2930273      19
## 1      ALB  2924447      19
## 30     POL  2844260      20
## 27     NED  2091615      21
## 4      AUT  1759357      21
## 13     FIN  1491262      22
## 40     GBR  1433210      22
## 20     ISR  1179195      22
## 37     SWE  1152061      22
## 23     LTV  1150486      22
## 22     LAT  1135406      22
## 10     CZE  1012613      22
## 14     FRA   995547      22
## 2      ARM   977794      22
## 32     ROU   769670      23
## 6      BEL   769348      23
## 17     GRE   752807      23
## 8      CRO   746855      23
## 26     MNE   665526      23
## 7      BUL   629733      23
## 31     POR   618582      23
## 33     SMR   598478      23
## 5      AZE   587486      23
## 28     MKD   576062      23
## 35     SLO   548119      23
## 3      AUS   532362      23
## 16     GER   491185      23
## 38     SUI   488767      23
## 9      CYP   466854      23
## 12     EST   466248      23
## 24     MLT   384298      23
## 19     IRL   337958      23
## 11     DEN   257512      23
## 18     ISL   241213      23
## 15     GEO   201274      23
```

According to the prediction model, the possible winner of ESC 2022 will be either Italy, Serbia or Ukraine because of their low number in Ranking. The rest of the Rankings are very close to each other in the range 18 to 23, which leaves a lot of room for interpretation.

10. Conclusion

From the original research question

“Is it possible to predict the winner of the Eurovision Song Contest using the influencing factors language, age of the main artist, views on YouTube and genre?”

the following statistical questions and their associated alternative and null hypotheses have been derived:

“Does the genre of a competing song have an effect on the ranking in the Eurovision Song Contest?”

“Does the language of a competing song have an effect on the ranking in the Eurovision Song Contest?”

“Does the age of a contestant have an effect on their ranking in the Eurovision Song Contest?”

“Do the YouTube views of a competing song have an effect on the ranking in the Eurovision Song Contest?”

The first statistical research question about the genre had to be omitted due to a lack of objective categorisation and unavailability of appropriate data.

The second statistical question about the influence of the language of a competing song on the ranking has been answered using variance analysis as method with language being the independent categoric variable and ranking the dependent metric variable. The result of this method shows a slight effect that is significant on the 0.05 level. The result was visualised with a parallel boxplot and the validity within the population was tested with the ANOVA table.

The third statistical question concerning the effect of a contestant’s age on the raking has been approached with the Pearson correlation coefficient as method using age as the independent metric variable and ranking as the dependent metric variable. Neither the Pearson method nor the scatterplot showed a significant correlation.

The fourth statistical question analysing the effect of YouTube views of a competing song on the ranking has undergone the method of correlation coefficient of Spearman with views serving as independent metric variable and ranking as the dependent metric variable. A significant negative effect (on 0,05 level) with a correlation coefficient of -0,61 could be shown. The linear regression model is:

Ranking = 23.692 - 0.000001457 * Views

That means that the effect strength is minus 1,5 ranks per one million views on YouTube. The results were visualised with a scatterplot and the validity within the population was tested with Spearman's rank correlation coefficient. For model diagnostics, residual plot and Q-Q plot were used as statistical methods. Furthermore, the linear regression model was used for a prediction of demo data and sample data for 2022.

The following table 22 visualises the abovementioned results:

<table>
<tr><th colspan="2">Language => Ranking (1 categoric and 1 metric variable)</th></tr>
<tr><td><b>Effect</b></td><td>Yes, slight effect</td></tr>
<tr><td><b>Significant on 0.05 level</b></td><td>Yes (p is 0.0339)</td></tr>
<tr><th colspan="2">Age =&gt; Ranking (2 metric variables)</th></tr>
<tr><td><b>Effect</b></td><td>No correlation could be shown with the Pearson method</td></tr>
<tr><td><b>Significant on 0.05 level</b></td><td>n.a.</td></tr>
<tr><th colspan="2">Views => Ranking (2 metric variables)</th></tr>
<tr><td><b>Effect</b></td><td>Yes, correlation coefficient of -0.61 could be shown with the Spearman method</td></tr>
<tr><td><b>Significant on 0.05 level</b></td><td>Yes (p is 0.001854)</td></tr>
<tr><td><b>Linear regression model</b></td><td>Ranking = 23.692 - 0.000001457 * Views</td></tr>
<tr><td><b>Effect strength</b></td><td>ca. 1,5 ranks down per 1 Mio views</td></tr>
</table>

Table 22: Results summary

As for the overall initial research question, not all of the four factors can be used to predict the winner of the Eurovision Song Contest. Genre could not be analysed, the age does not have a significant effect on the ranking. However, the factors language (slight significant effect) and views (moderate correlation coefficient) can indeed be used for a prediction of the ESC winner.

List of Figures

List of Tables

List of References

Antipov Evgeny A. and Elena B. Pokryshevskaya. 2017. *Order effects in the results of song contests: Evidence from the Eurovision and the New Wave.* Journal of Judgment and Decision Making 4(12): 415-416

Barclay, Simon. 2011-2021.*The Complete & Independent Guide to the Eurovision Song Contest*. Albuquerque*:* Silverthorn Publishing Inc.

Bard, Kendall. 2017. *Does Winning Eurovision Impact a Country's Economy?.* University of Tennessee.

Blangiardo, Marta and Gianluca Baio. 2014. *Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models.* Journal of Applied Statistics 41:10, 2312-2322

Bohlmann, Phillip V.. 2004. *The Music of European Nationalism. Cultural Identity and Modern History*. Santa Barbara: ABC-CLIO, Inc.

Budzinski, Oliver and Julia Pannicke. 2016. *Do preferences for pop music converge across countries? Empirical evidence from the Eurovision Song Contest.* Discussion Paper. Ilmenau University of Technology.

EBU. 2022a. *Our History.* https://www.ebu.ch/about/history. Accessed 12 January 2022.

EBU. 2022b*. Our Members.* https://www.ebu.ch/about/members. Accessed 12 January 2022.

Eurovision. 2022a. *Facts and Figures*. https://eurovision.tv/about/facts-and-figures. Accessed 12 January 2022.

Eurovision. 2022b. *London 1960*. https://eurovision.tv/event/london-1960. Accessed 12 January 2022.

Eurovision. 2022c. *FAQ.* https://eurovision.tv/about/faq. Accessed 12 January 2022.

Eurovision. 2022d. *The 4 ways to make Eurovision 2021 happen.* https://eurovision.tv/about/faq. Accessed 12 January 2022.

Eurovision. 2022e. *Semi-final running orders revealed.* 30.03.2021. https://eurovision.tv/story/semi-final-running-orders. Accessed 12 January 2022.

Eurovision. 2022f. *Lugano 1956.* https://eurovision.tv/event/lugano-1956. Accessed 12 January 2022.

Eurovision. 2022g. *Rules.* eurovision.tv, https://eurovision.tv/about/rules. Accessed 12 January 2022.

Eurovision. 2022h. *Voting.* https://eurovision.tv/about/voting. Accessed 12 January 2022.

Eurovision. 2022i. *How it works.* https://eurovision.tv/about/how-it-works. Accessed 12 January 2022.

Eurovision. 2022j. *'American Song Contest' to launch on Monday 21 February, 2022.* https://eurovision.tv/story/american-song-contest-february-2022. Accessed 13th January 2022.

Eurovision Services. 2022. *Who we are*. https://www.eurovision.net/about/whoweare. Accessed 12 January 2022.

Eurovision World. 2022. *Eurovision: National Selection.* https://eurovisionworld.com/national. Accessed 12 January 2022.

Events Eurovision. 2022. *Live Events.* https://events.eurovision.tv/. Accessed 12 January 2022.

Jordan, Paul and Josianne Zwart. 2017. *What does it take to become a Eurovision host city?.* 30.07.2017. https://eurovision.tv/story/what-does-it-take-to-become-a-eurovision-host-city. Accessed 12 January 2022.

Kakouris, Dionysios, Georgios Theocharis, Prodromos Vlastos and Nasrullah Memon. 2016. *Detecting Hidden Patterns in European Song Contest—Eurovision 2014.* University of Southern Denmark.

Karlsson, Andreas. 2015. *Eurovision Song Contest: Regression analysis highlighting the voting patterns.* 22.05.2015. https://www.datalytyx.com/eurovision-song-contest-regression-analysis-highlights-the-voting-patterns/. Accessed 12 January 2022.

Kumpulainen, Iiro, Eemil Praks, Tenho Korhonen, Anqi Ni, Ville Rissanen, and Jouko Vankka. 2020. *Predicting Eurovision Song Contest Results Using Sentiment Analysis.* National Defence University, Helsinki.

Meijer, Albert. 2013. *Be My Guest: Nation branding and national representation in the Eurovision Song Contest*. Master Thesis. University of Groningen.

Millner, Ralf, Matthias Wolfgang Stoetzer, Christina Fritze and Stephanie Günther. 2015. *Fair oder Foul? Punktevergabe und Platzierung beim Eurovision Song Contest*. Jenaer Beiträge zur Wirtschaftsforschung, No. 2015/2. Ernst- Abbe-Hochschule, Fachbereich Betriebswirtschaft, Jena.

Oliver, Stephen (Director). 2011. *The Secret History of Eurovision* [Film]. Brook Lapping Productions & Electric Pictures

Raykoff, Ivan and Robert D. Tobin. 2007. *A Song for Europe: Popular Music and Politics in the Eurovision Song Contest*. Hampshire: Ashgate Publishing, Ltd.

Royston, Benny. 2019*. Eurovision by numbers: What does it take to put on our show?.* 15.05.2019. https://eurovision.tv/story/eurovision-by-numbers-what-does-it-take-to-put-on-a-show. Accessed 12 January 2022.

West, Chris. 2017*. Eurovision! A History of Modern Europe Through the World's Greatest Song Contest.* London: Melville House UK