Data Obtaining and Machine Learning Final Project

Natalie Aflalo
Linoy Yarkoni

https://github.com/natalieaflalo/Project-Eurovision

Main Steps Of the Project

Data Acquisition → Data Handling → EDA → Machine Learning

Research Questions:

- ♡ **Which song has the highest odds for wining an Eurovision contest giving a list of songs.**
- ♡ **How many points will the winning song get.**

EUROVISION
SONG CONTEST
ITALY 2022

Data Sources:

- **Crawling:**
  - https://en.wikipedia.org/wiki/List_of_Eurovision_Song_Contest_entries_(1956%E2%80%932003)
  - https://en.wikipedia.org/wiki/List_of_Eurovision_Song_Contest_entries

  Used crawling for each year in the links above.
- **Selenium:**
  - https://eurovision.tv/events
  - https://en.wikipedia.org/wiki/List_of_Eurovision_Song_Contest_entries_(1956%E2%80%932003)
  - https://en.wikipedia.org/wiki/List_of_Eurovision_Song_Contest_entries
- **Other:**
  - **CSV file of artist names and genders.**

Data Sources-Crawling

- **We used functions from ‘pandas’, ‘BeautifulSoup’ and ‘Request’ to crawl the tables from Wikipedia and save the records in data frames: data frame named: ‘df_finals_1956_2003’ for years 1956-2003 (final), data frames named: ‘df_finals_2004_present’, ‘df_semi_finals_1_2004_present’ and ‘df_semi_finals_2_2004_present’ (final, semi final 1, semi final 2). In the end, we combined the data frames above in to ‘result’ data frame.**

Data Sources-Selenium

- **We used Selenium to fill the names of hosting country and the country that won the previous contest for years 1956-2021. In addition we mine the links for each contest using selenium.**

Data Handling:

♡ **Outliers:**

- **We deleted the data about year 1956 because there was no score, only winner had placing and every artist preformed two songs (instead of one).**

♡ **Cleaning values from unwanted characters:**

- **The data from Wikipedia had characters such as '[a],[b],[ab]' and other comments added to the songs language, song name and artist columns so we removes this characters.**

"Od nas zavisi" (Од нас зависи)
"Light a Candle"
"Dans le jardin de mon âme"
"Never Let It Go"
"Addicted to You"
"Tell Me Who You Are"
"Na jastuku za dvoje" (На јастуку за двоје)

11	Israel	Shiri Maimon	"Hasheket Shenish'ar" (השקט שנשאר)	Hebrew, English	4	154
12	Serbia and Montenegro	No Name	"Zauvijek moja" (Заувијек моја)	Montenegrin	7	137
13	Denmark	Jakob Sveistrup	"Talking to You"	English	9	125
14	Sweden	Martin Stenmarck	"Las Vegas"	English	19	30
15	Macedonia	Martin Vučić	"Make My Day"	English	17	52
16	Ukraine	GreenJolly	"Razom nas bahato" (Разом нас багато)	Ukrainian, English[a]	19	30

Data Handling:

♡ **Conversions:**

- **We converted the column ‘type of contest’ into two categorical columns: the first one is: ‘was in final’ if the value is 1 the country was in final, else 0. The second column is: ‘which semi final’: the value 0 means final, the value 1 means semi final 1 and the value 2 means semi final 2.**
- **We also converted the column ‘gender’ to categorical while 0 is for ‘male’, 1 is for ‘female’ and 2 is for ‘group’.**
- **We combined the columns of ‘Macedonia’ and ‘North- Macedonia’.**

Data Handling:

♡ **Duplicates:**

Songs that were in semi final and final (since 2004), had two records each, so after splitting ‘type of contest’ column into two categorical columns we combined between the records from the final and the semi final and deleted the duplicated records.

EDA-Visualization:

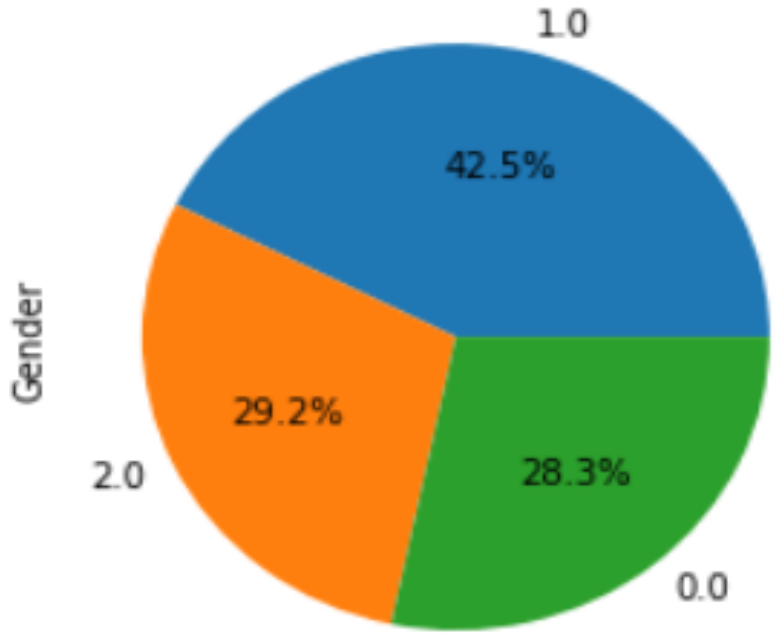
- ♡ **This pie chart is showing the ratio of the participants' gender in all contests:**

28.3% are male.

42.5% are female.

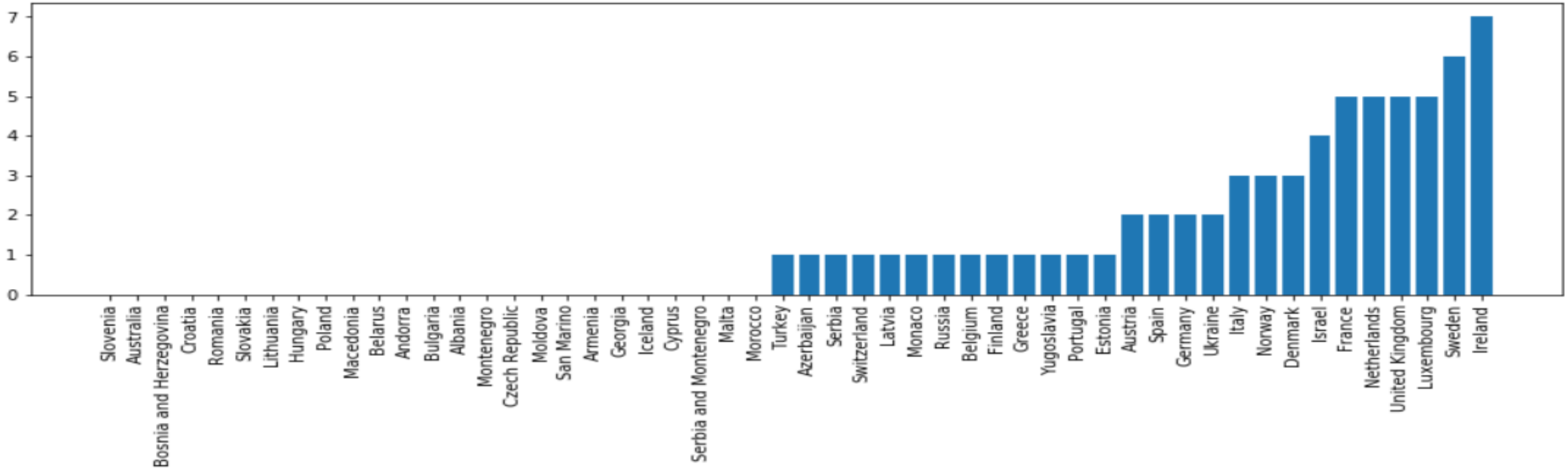
29.2% are groups.

- ♡ **We can see that the majority of the participants are women.**

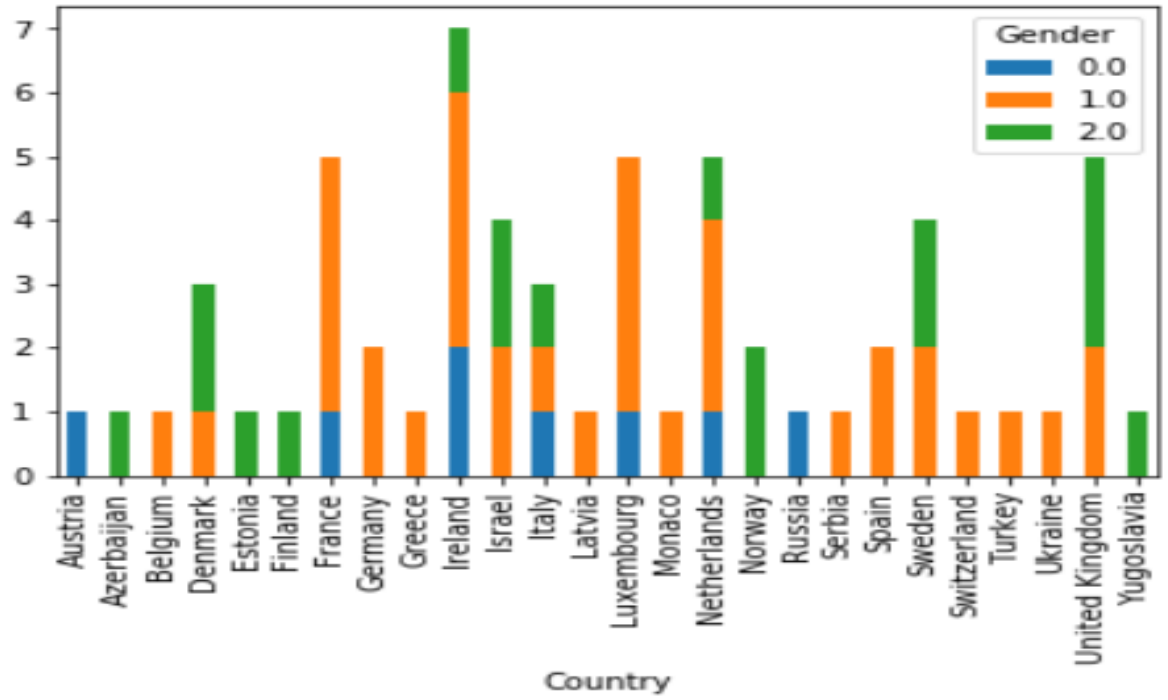


EDA-Visualization:

♡ **This bar plot is showing the number of winnings for each country (x-axis is the countries, y-axis is the amount of winnings):**



EDA-Visualization:

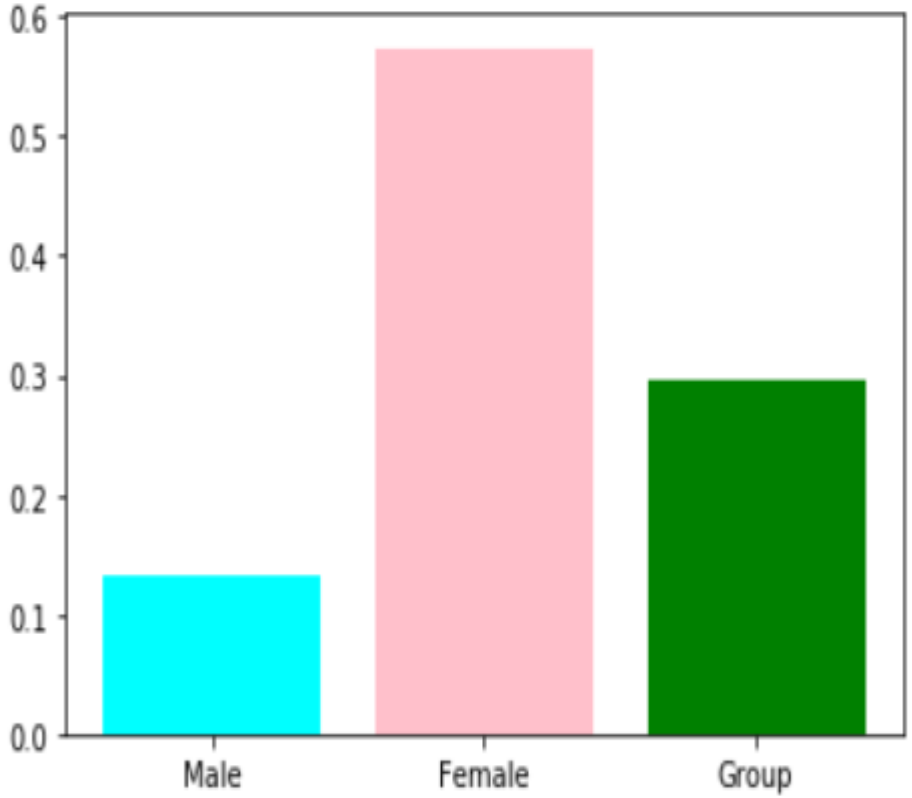


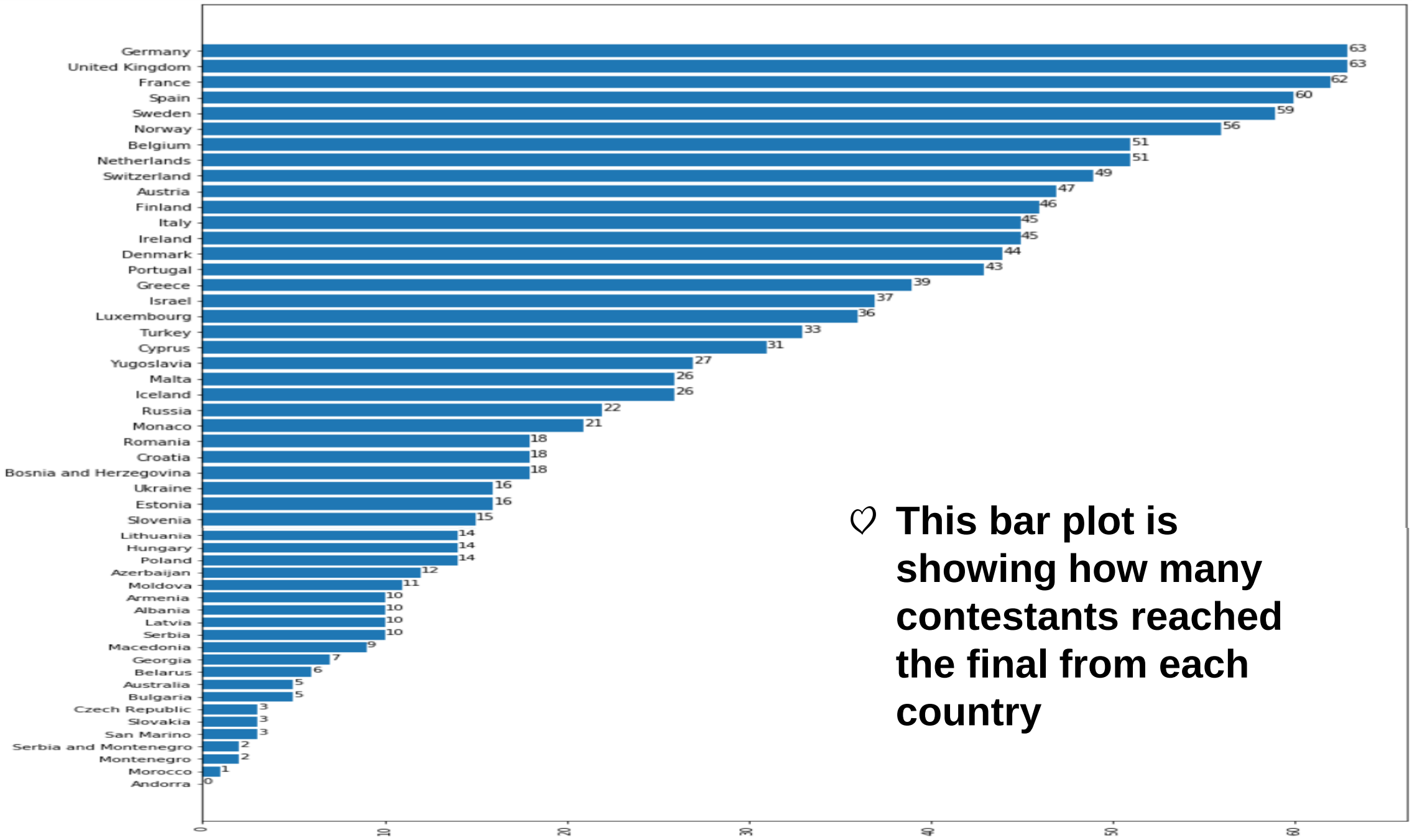
- This bar plot is showing the number of winnings for each country (x-axis is the countries, y-axis is the amount of winnings) considering the gender (0-male, 1-female, 2-group):
- We can infer that the country that won the most represented by women.

EDA-Visualization:

- **This bar plot is showing the normalized value of winnings by gender (x-axis is the gender, y-axis is the normalized value)**

Amount of winners from specific gender divides by amount of winners in total.

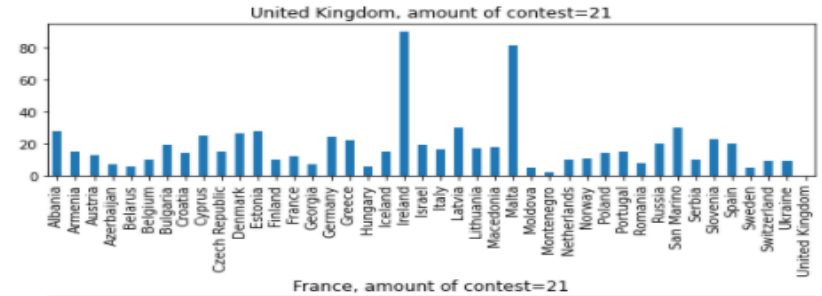


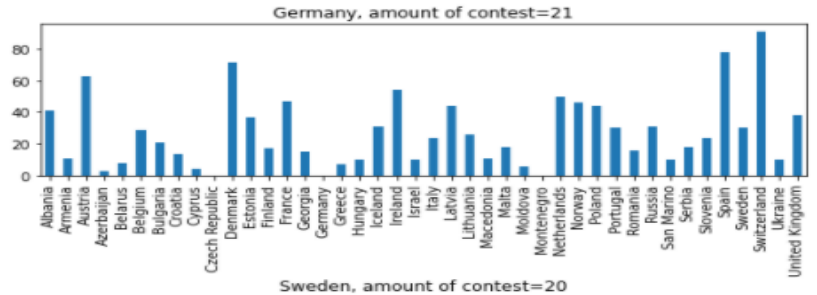


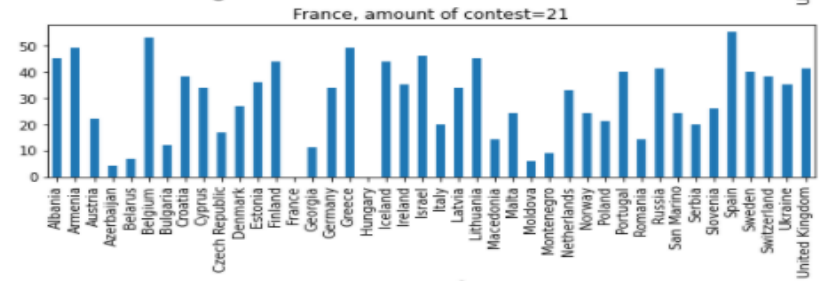
♡ **This bar plot is showing how many contestants reached the final from each country**

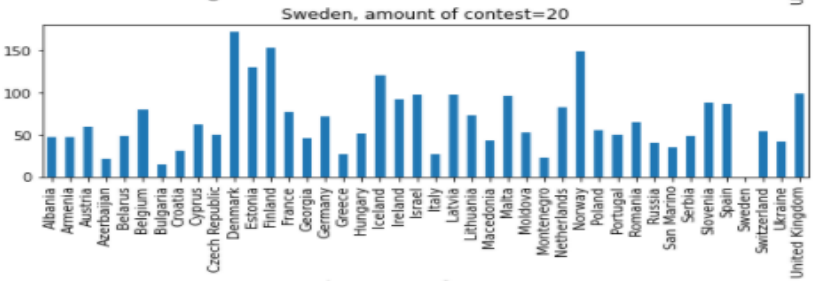
EDA-Visualization:

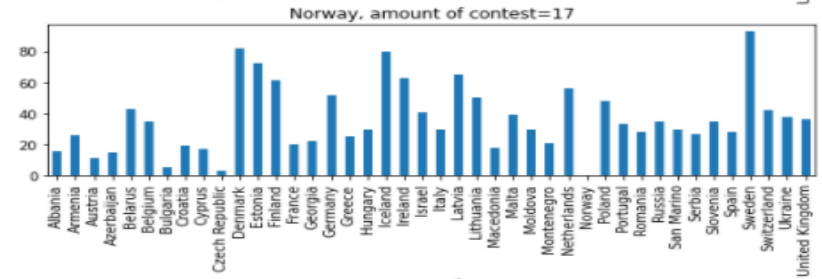
- These individual bar sub-plots are showing the sum of final points from each jury to each finalist country.

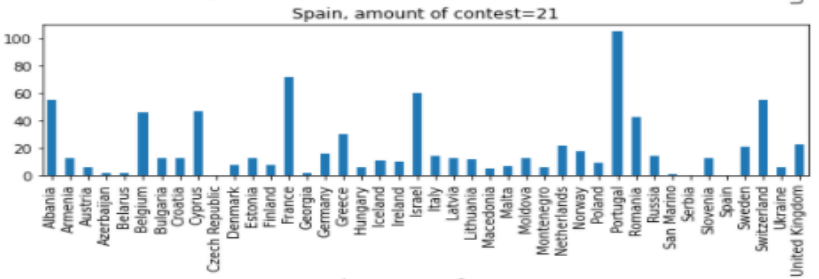


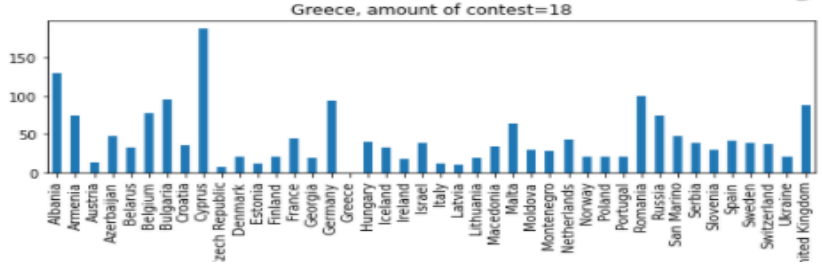


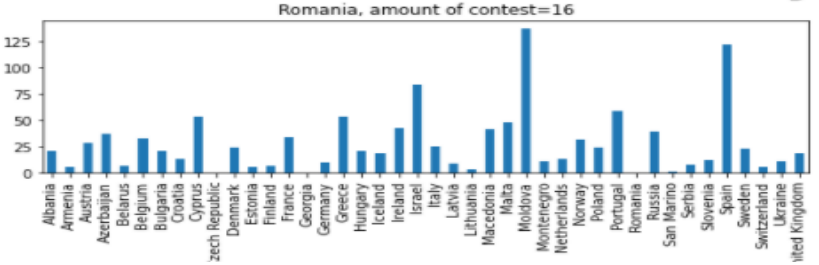






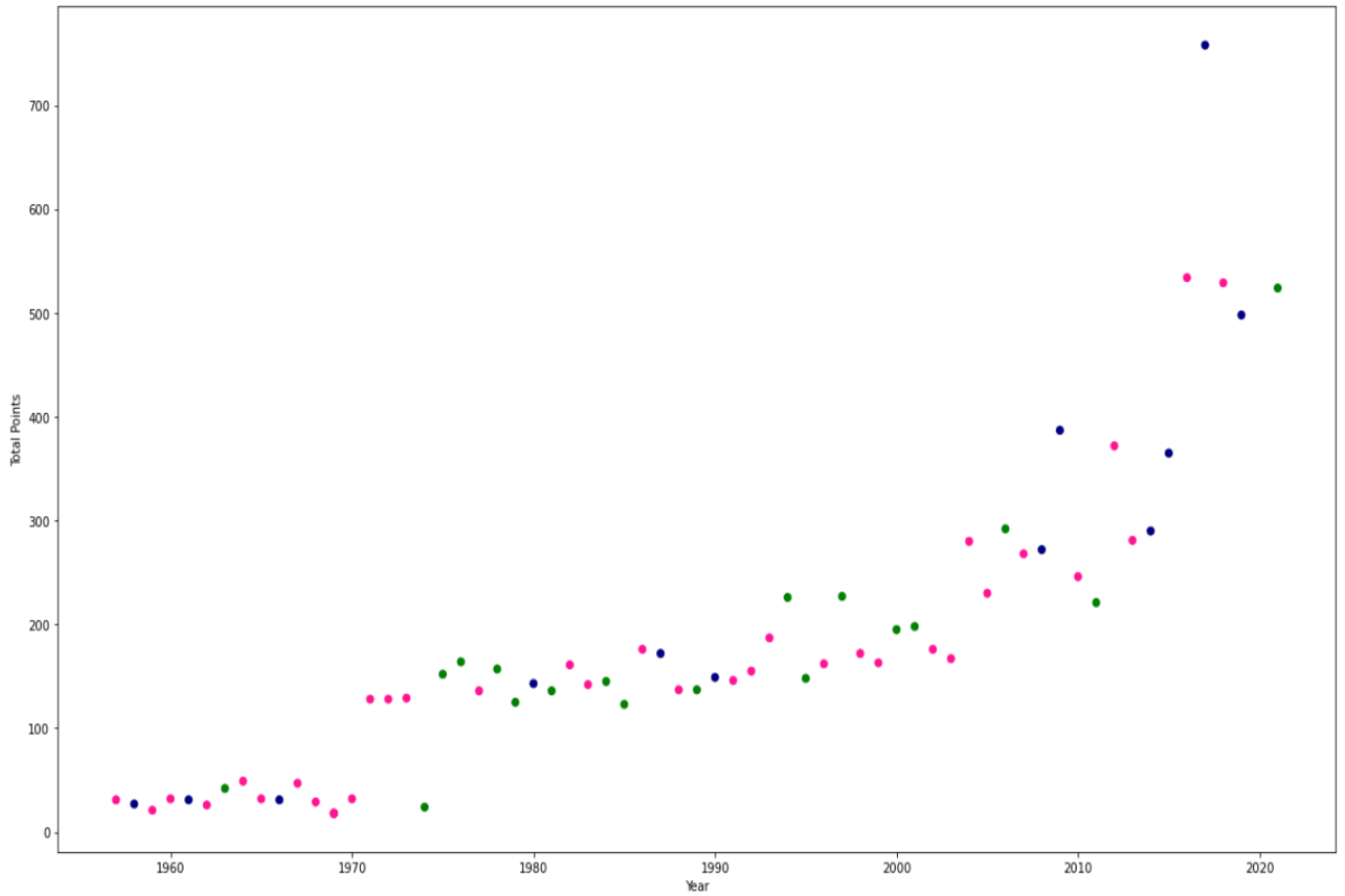






EDA-Visualization:

- **This scatter plot is showing winners' total points of all years (1957-2021) considering the gender.**



'Male: Nave Blue, Female: Deep Pink, Group: Green'

Machine Learning-logistic regression

- ♡ **We trained a model that predicts if a song will reach the final contest or not using logistic regression.**
- ♡ **According to year and gender columns.**
- ♡ **The correction ratio is 81.13% - 70% from the data trained and 30% tested.**

- ♡ **The data we predicted contains songs from 2020 contest which was canceled and songs that were withdrawn.**

```
Initial amount of samples: #1589
Number of training samples: #1112
Number of test samples: #477
Min Value:  [0. 0.]
Max Value:  [1. 1.]
correct: 387
total: 477
correct %: 0.8113207547169812
```

```
In [44]: metrics.confusion_matrix(y_test, y_pred)

Out[44]: array([[ 15,  68],
                [ 22, 372]], dtype=int64)
```

	Actual	Predicted	correct
1431	1	1	1
1519	1	1	1
1447	1	1	1
349	1	1	1
88	1	1	1
436	1	1	1
924	0	1	0
1171	1	1	1
785	1	1	1
998	0	1	0
243	1	1	1
1256	1	1	1

Machine Learning-linear regression

- ♡ **We trained a model that predicts the final points for each song using linear regression.**
- ♡ **According to year and gender columns.**
- ♡ **The R-Squared value is 0.74 (> 0.5) 70% from the data trained and 30% tested.**

```
R2: 0.747022322839864
R2: 0.747022322839864
```

Out[6]:

	Year	Gender	Predicted
0	1968	0	71.133669
1	1974	1	88.400559
2	1976	0	124.970218
3	1979	2	98.937880
4	1982	0	165.347630
5	1986	1	169.155382
6	1988	0	205.725041
7	1990	2	172.963134