

# PREDICTING UEFA CHAMPIONS LEAGUE WINNER 2022

---

By Andrea Fernandez

MSDS 696 Data Science Practicum II



# UEFA CHAMPIONS LEAGUE



- This league consists of 32 teams broken up into 8 groups
- These teams must qualify each year in order to participate which creates some diversity of the teams.
- The best teams in the European leagues all compete throughout the year starting from June to the next May

# RESEARCH QUESTION

- The winner is going to be determined by who can qualify for this years Champions League
- The winner will also depend on past performances and winners
- Question: **Is it possible to predict the winner of the Champions League based on past winners and performances?**
- I will produce results, but the winner will not be know for a while!



# DATA

- First dataset labeled champs:
  - All matches from 1955-2015 seasons
  - Includes group, round, team1, team2, half-time score, full-time score, aggregate, and date
  - Dimensions are 23x6554
- Second dataset labeled ucldata:
  - consists of the winner and runner up from 1955-2019
  - Includes team, country, coach, formation, mvp, position, and season
  - Dimensions are 7x127



# DATA CLEANING

## Ucldata

- No missing values
- Missing values were filled in with 'unknown'
- No data changes needed

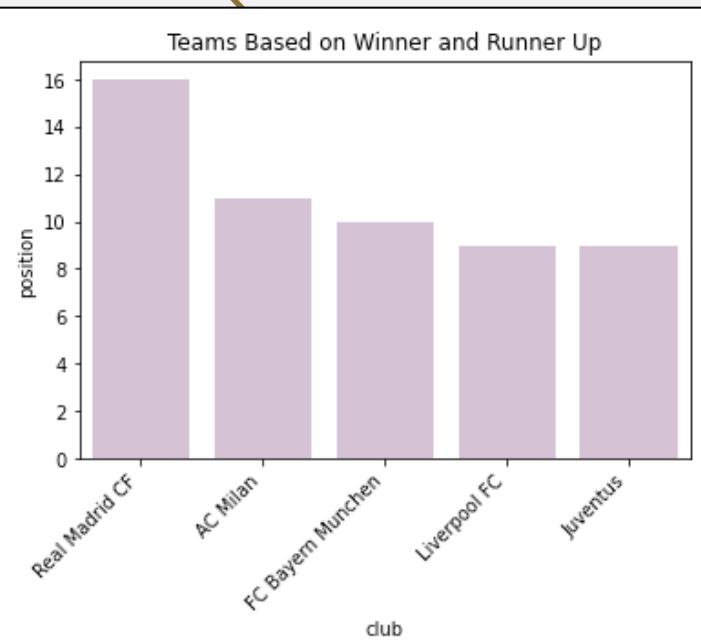
## Champs

- Contained missing values for multiple columns
- Replaced missing values with “no penalties” or “no info”
- Removed foreign characters

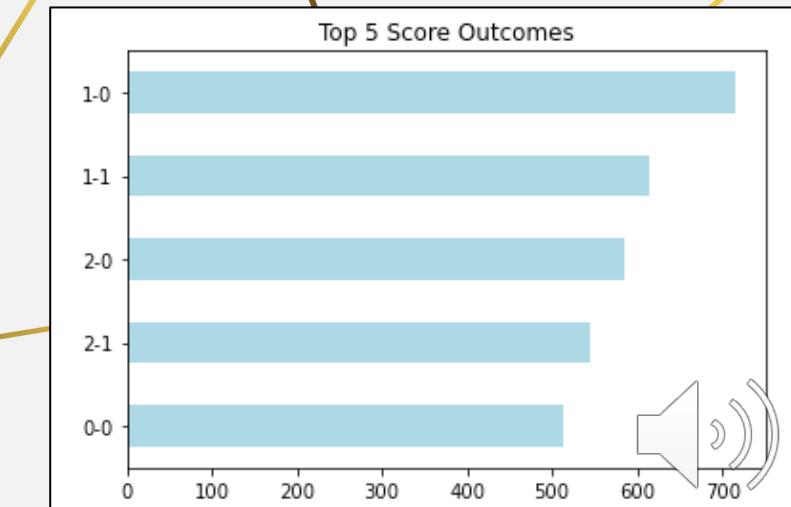
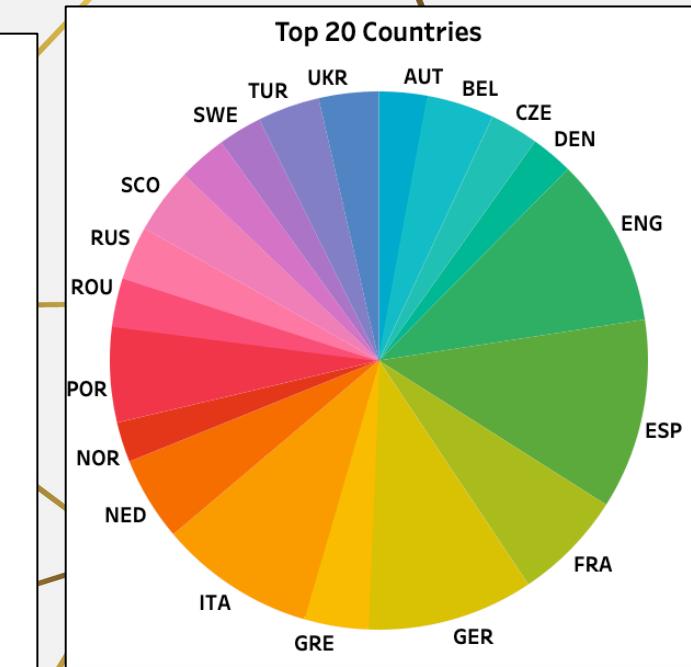


# EXPLORATORY DATA ANALYSIS

- Ucl data contained all object columns , while champs data had 9 integer
- Real Madrid (ESP) has been in the final the most
- Spain is represented the most throughout 1955-2016
- The most common full-time score in the competition is 1-0

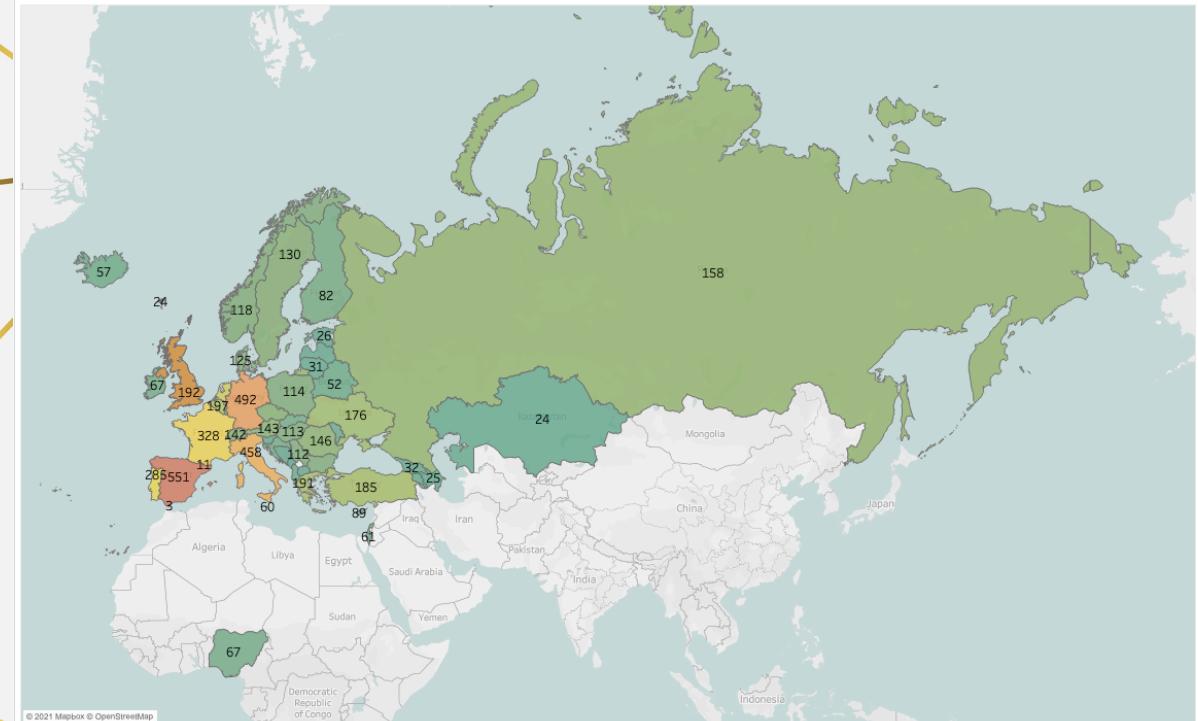


Real Madrid	204
Bayern Munich	156
Barcelona	141
Manchester United	131
AC Milan	127
Juventus	117
SL Benfica	112
Dinamo Kiev	110
FC Porto	108
AFC Ajax	98
Name: home, dtype: int64	

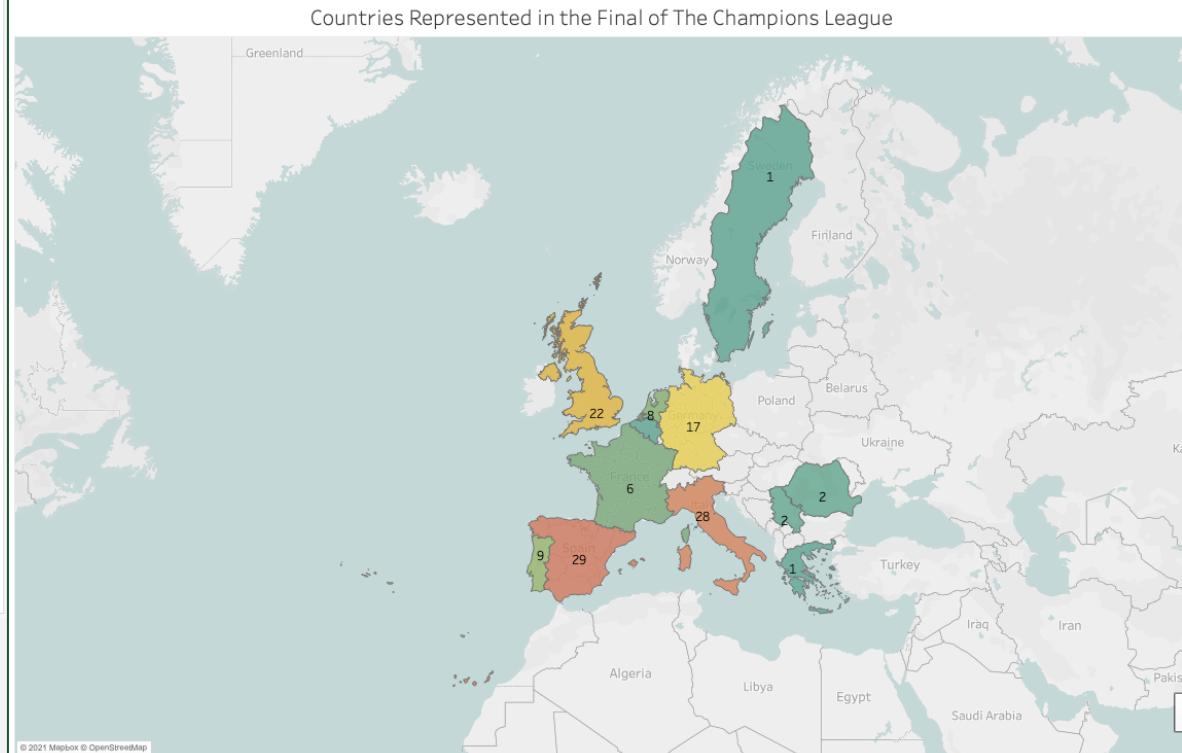


# EXPLORATORY DATA ANALYSIS

Countries Represented in The Champions League



Countries Represented in the Final of The Champions League



## BUILDING THE MODELS

- Three classification models
  - Logistic Regression
  - Support Vector Classification
  - Decision Tree Classification
- Separated the data into feature data and target data
- Standardized, split, and shuffled the data into training and testing sets
- Ran all 3 models and produced a F1 score and accuracy score



# RESULTS

- The Logistic Regression model produced an F1 score of 0.9461 and an accuracy score of 0.9461 as well
- The SVC model produced an F1 score of 0.9611 and an accuracy score of 0.9611
- The XGBClassifier model produced an F1 score of 0.9981 and an accuracy score of 0.9981
- Adjusted XGBClassifier parameters:
  - 1<sup>st</sup> adjustment produced F1 score of 0.9800 and an accuracy of 0.9800
  - 2<sup>nd</sup> adjustment produced F1 score of 0.9557 and an accuracy of 0.9557



## CONCLUSION

- The models proved to be able to predict the outcomes of games based on all the history of previous matches
  - It was very accurate with 99% and can be trusted to be used for every aspect of the tournament
  - Predictions can be made for betting purposes as well as strategies for the teams

