# PGA TOUR Analysis and Regression

By Kyle Arbide

# Project Scope

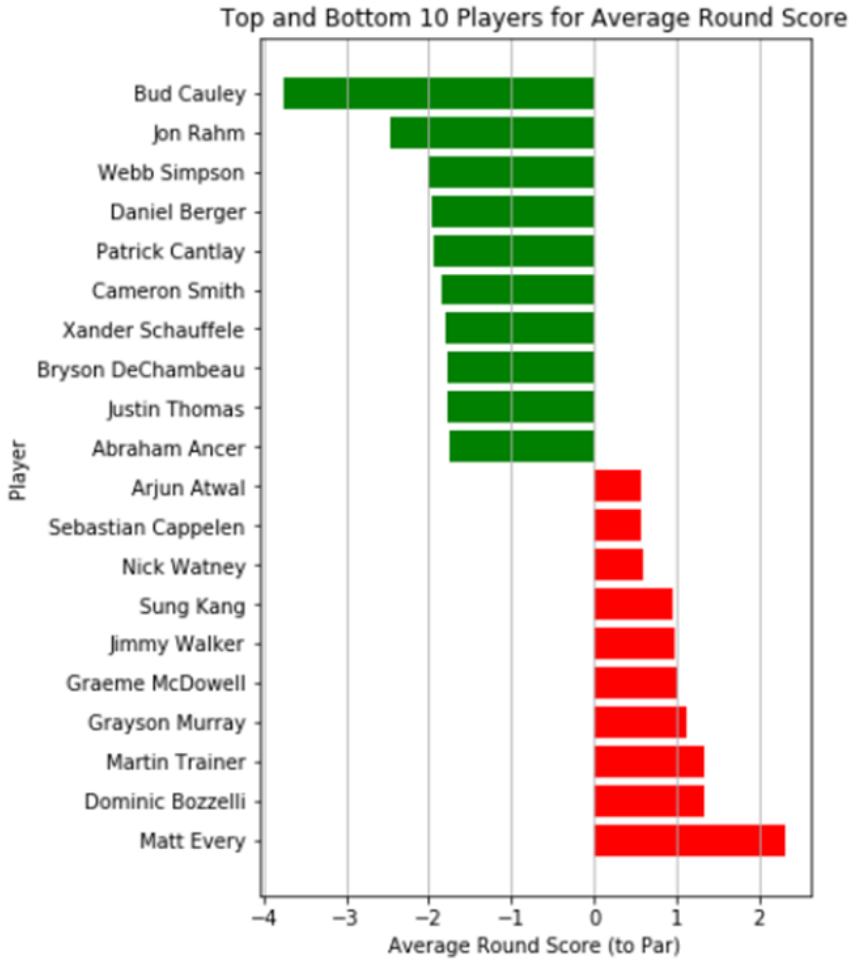
- **Objectives:**
  - Acquire and clean data for PGA Tournament Results and Player Statistics
  - Perform exploratory analysis on player statistic and create visuals that represent player skill
  - Create a predictive model that uses player skill to predict performance based on round/course

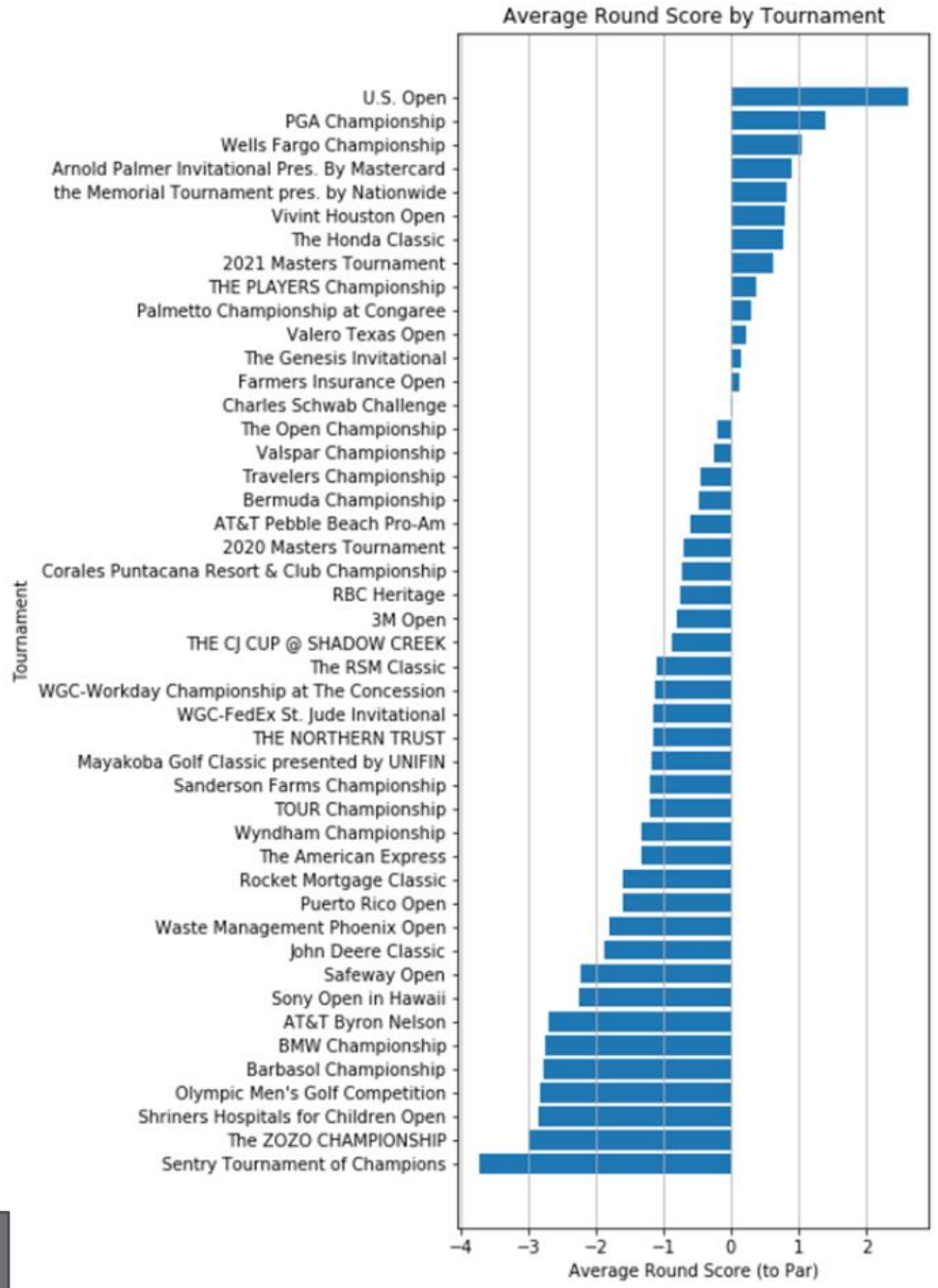
# Data Description

- Player Statistics were pulled from Kaggle Datasets
- Player Results were pulled from SportData.io API
- Data was transformed from the following format shown on the left into more comprehensive round scores
- API pull was built to merge the data from all players and tournaments, and to identify edge cases.
- Data taken from the 2020 season and first few tournaments of 2021

| | DoubleEagle | Eagle | Birdie | IsPar | Bogey | DoubleBogey | WorseThanDoubleBogey | Number | Par |
|---|---|---|---|---|---|---|---|---|---|
| 0 | False | False | False | True | False | False | False | 10 | 4 |
| 1 | False | False | False | True | False | False | False | 11 | 4 |
| 2 | False | False | False | True | False | False | False | 12 | 4 |
| 3 | False | False | False | True | False | False | False | 13 | 5 |
| 4 | False | False | True | False | False | False | False | 14 | 3 |
| 5 | False | False | False | True | False | False | False | 15 | 4 |
| 6 | False | False | False | False | True | False | False | 16 | 5 |
| 7 | False | False | False | True | False | False | False | 17 | 3 |
| 8 | False | False | False | True | False | False | False | 18 | 4 |

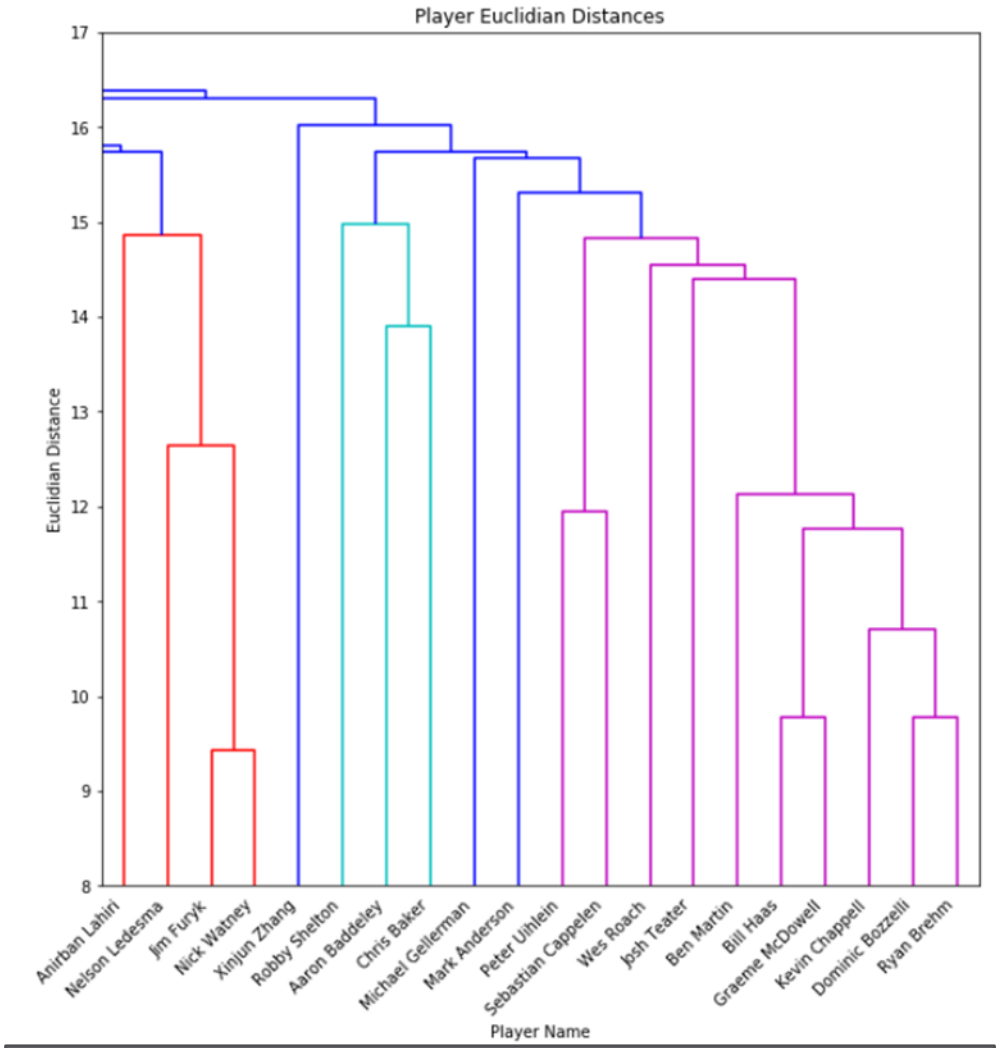
# Exploratory Analysis



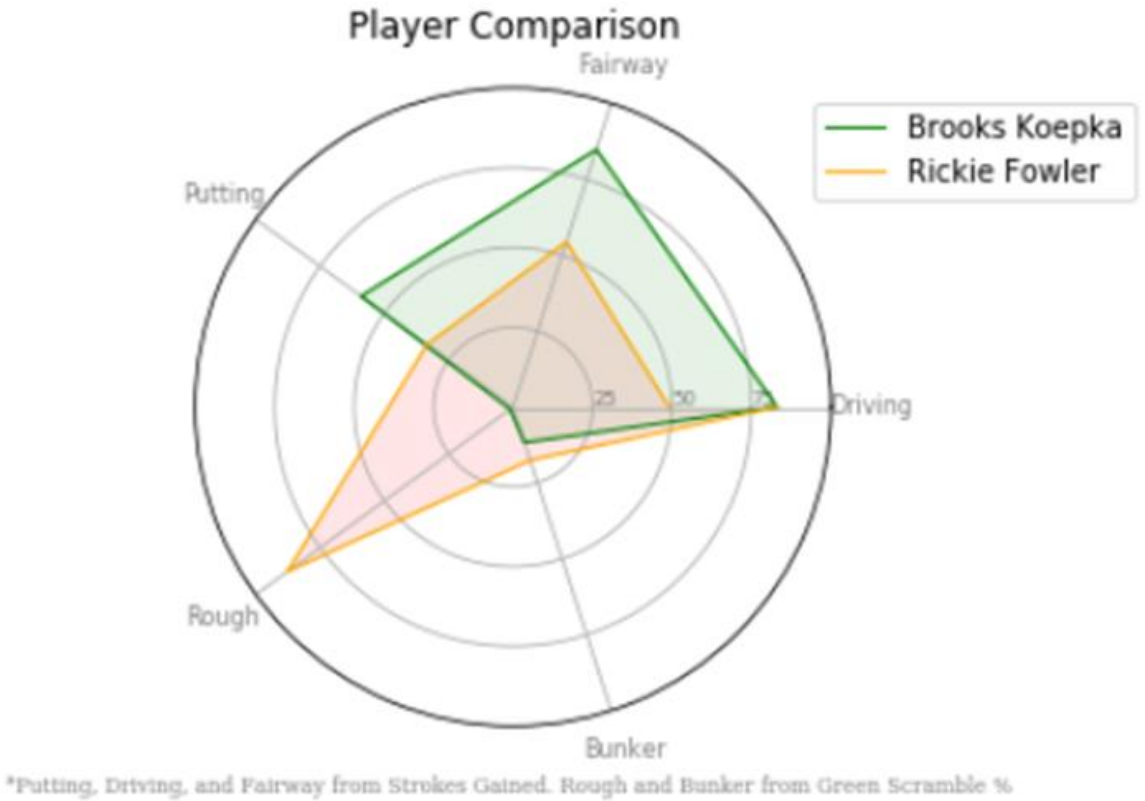




Summaries of the average round scores by player and tournament

# Player Comparison Visuals




Dendrogram created using Euclidian distances from Tournament Results




Spider Plot comparing player statistics in different parts of the course

# Regression Model

- Regression models using all the tournament results, or all the round results failed
- Instead, one tournament was selected, and only results from that tournament were used
- Small sample size limited the number of independent variables used in the model. This prevented the model from finding specific traits but instead used more general stats
- Min-max scaler used so weights could be compared

| Variable | Coef |
|---|---|
| Intercept | 6.07691841 |
| Par3Average | -15.1449274 |
| Par4Average | 10.4459427 |
| Par5Average | 5.67832195 |
| AverageSG | -3.32189072 |
| **R-squared** | 0.43257 |
| **n** | 132 |

Results from the regression model run for The Players Championship

# Results/Conclusion

- Exploratory analysis and player comparisons give some interesting, easy to read insights
- The regression model was not as accurate as hoped, but sports predictor models generally score around .6-.7 for R-squared
- Perhaps with a larger sample size, or data from across multiple years the model would improve in accuracy or be able to take more skill-specific variables
- Full report is available on my website: https://kylearbide.github.io/ -WIP

LinkedIn: https://www.linkedin.com/in/kyle-arbide-96b6491a1/