STAT 306 Project Proposal

Group members: Sicily Xie, Soren Rajani, Yulong Peng

1.The source of the data being used for the project.

The source of the data being used for the project is Ultimate Tennis Statistics (https://www.ultimatetennisstatistics.com/). The dataset used is “Player Stats for Top 500 Players”. Our dataset is available for download at: (https://drive.google.com/uc?export=download&id=1_MECmUXZuuILYeEOfonSGqodW6qVdhsS)

2.A brief description of the variables measured (including when, where, how, in what units, plus any other important information).

There will be 4 explanatory variables being used from the original dataset which are: Age, Current Rank, Best Rank, and Seasons. Age is how old the player is in 2019. Current Rank is the player’s rank in 2019. Best Rank is the best rank the player has ever achieved by 2019. Seasons is the number of seasons the player has played by 2019.

3. The research question, or other motivation, behind the analysis of the data.

Our goal is to predict tennis player prize money earned using player rank, seasons played, and age. Due to the high concentration of prize money at the top of tennis tournaments, we believe that there may be a quadratic relationship between prize money and player rank, i.e. that a small number of players win most large payouts. Our main hypothesis is therefore that there is a quadratic relationship between prize money and player rank.

A secondary hypothesis is that there may be some interaction between age and rank. It may be the case that very successful younger players could attend and win more tournaments and therefore accrue a larger prize pool than rank or age would predict alone.

4. An overview of who will do what on your project across your team members.

Sicily Xie: Keep track of the project progress and keep updated on any changes throughout the project.

Soren Rajani: Analysis of the data and errors to identify patterns from the project's data as well as participating in informing conclusions and supporting decision-making.

Yulong Peng: Cleansing, transforming, and modeling data to discover useful information and writing R codes.

All members will participate in discussions on project design, modification, and analysis throughout the project.

To Do

Graph Y: Prize Money X: Player Rank

Fit Model: Money = Age + Rank + Rank^2 + Best Rank + Season

the model= Age + Rank + Best Rank + Season

the model= Best Rank + Season + Age*Rank

To explore impact of interaction, as well as for other models

the model= Age + Rank + Best Rank + Season + Country

Qqplot normal plot regsubsets residual yhat against ei

AIC(model) BIC