Capstone 1 - Statistical Data Analysis

* Identify variables in the data to answer to a project question.

The project question to be answered is; What are the strongest correlators to determine future success of fantasy football players.

The variables in the data that need to be answered for this question include the individual stats associated with positional fantasy football players. There is an abundance of statistics that make up these variables. For example, in the broad scope of the running back position we must consider rushing yards, rushing touchdowns, receptions, receiving yards, and receiving touchdowns. We can then dive deeper to include attempts per game, yards per attempt, and even yards allowed by the defence the player is facing. However to keep this project from being too cumbersome we will stick with the broader scope, only focusing on the individual players statistics.

We also must set up a basis for which we score these statistics. For example, the most common scoring gives one fantasy point per 10 rushing or receiving yards gained by a player, and one point for every 25 passing yards gained. As well, rushing and receiving touchdowns are worth six fantasy points where a passing touchdown is worth four. In most formats players earn points per reception. For this project we will be using the one point per reception format. Players also lose points for turnover, a fumble or interception result in a loss of two points. Turnovers are much too difficult to predict however, so these will not be weighted as heavily.

* Identify strong correlations between pairs of independent variables or between an independent and a dependent variable.

To begin to identify strong correlators between our statistics I used the .corr method. I incorporated this into a plot to better visualize the strongest and weakest correlators between individual statistics and total fantasy points for each position (quarterbacks, running backs, wide receivers, and tight ends).

For Quarterbacks I found that the strongest correlations between the independent variables and the dependent variable, Fantasy Points, were; Touchdowns (0.928), Passing Yards (0,900), Completions (0.835), and Passing Attempts (0.818).

* qb\_corr = qb[['FantPt', 'Cmp', 'Att', 'Yds', 'TD', 'Tgt', 'Att.1', 'Yds.1', 'TD.1', 'Fmb', 'FL', 'Int', 'TD.3']].corr(method='spearman', min\_periods=185)

For Running Backs; Total Touchdowns, rushing and receiving (0.788), Rushing Yards (0.774), Rushing Attempts (0.748), Receptions (0.693), Targets (0.689), Receiving Yards (0.664), and Rushing Touchdowns (0.655).

* rb\_corr = rb[['PPR', 'Att.1', 'Yds.1','Y/A','TD.1', 'Tgt', 'Rec', 'Yds.2', 'Y/R', 'TD.2', 'Fmb', 'FL', 'TD.3']].corr(method='spearman', min\_periods=365)

For Wide Receivers; Receiving Yards (0.949), Receptions (0.932), Targets (0.906), Total Touchdowns (0.749), Receiving Touchdowns (0.741).

* wr\_corr = wr[['PPR', 'Att.1', 'Yds.1','Y/A','TD.1', 'Tgt', 'Rec', 'Yds.2', 'Y/R', 'TD.2', 'Fmb', 'FL', 'TD.3']].corr(method='spearman', min\_periods=90)

For Tight Ends; Receiving Yards (0.956), Receptions (0.946), Targets (0.930), Total Touchdowns (0.656), Receiving Touchdowns (0.643).

* te\_corr = te[['PPR', 'Att.1', 'Yds.1', 'TD.1', 'Tgt', 'Rec', 'Yds.2', 'Y/R', 'TD.2', 'Fmb', 'FL', 'TD.3']].corr(method='spearman', min\_periods=36)
* Practice identifying the most appropriate tests to use to analyze relationships between variables.

After testing the correlations between the independent variables and the dependent variable I tested linear regressions. I ran an OLS regression to confirm that we were getting a p-value that backed up our strongest correlations. For Quarterbacks, I tested Fantasy Points against Touchdowns and the results showed a significantly low p-value (0.00) as well as a small standard error (0.245).

* m = ols('FantPt ~ TD',qb).fit()

I continued running these regressions for the other independent variables and found that the results aligned with the correlations we were receiving from our correlation models.

* qb\_test = qb.rename(columns={'Att.1': 'AttR'})
* m = ols('FantPt ~ AttR',qb\_test).fit()

The OLS results showed a significantly low p-value (0.00) and a small standard error (0.209). The smaller standard error for passing attempts compared to passing touchdowns makes sense in this scenario, passing attempts are much more consistent and common than passing touchdowns.