**Intermediate Report**

**Topic: Sports Analysis**

**Abstract**

This project is to develop a predict framework for a match winner. I analysis [Premier League](http://football-data.co.uk/mmz4281/1516/SC0.csv) football data and predict the winner, for which six teams competed against each other. There are four steps for the whole process. Data collection, data preprocessing, data analysis and model building and evaluation. The Linear regression algorithm I would like to use as a model.

**Data Collection**

I collected data from the football-data website. The file which I download have already include the home team, away team, home shot, away shot, and etc. which is from 2012-2019.

**Data Preprocess**

Firstly, I extract ten features for each team by calculating the features mean for the first ten games. Then, I use that data to make the dataset data frame. I have already had dataset for each Premier League season. The rest thing I need to do is that I have to combine datasets to be one.

**Data Analysis**

I plan to use linear regression algorithm to analysis data. After that I can order the team to win from the most possibility to the lowest possibility. The bin graphic I would like to show the result.

**Potential Analysis**

There is a certain amount of error in the data of different seasons to combine them. I guess the accuracy will be between 0.5-0.75.

**List of data sources (URL)**

http://football-data.co.uk/scotlandm.php

**Timeline for this project**

* 2.3 Collect data
* 2.8 1-page project proposal
* 2.28 Recollect data – finished
* 3.24 2-page project intermediate report
* 3.27 Preprocess data - combine five datasets to one need to be done
* 4.1 Analysis data
* 4.10 Build model
* 4.21 Final project report