# **INTRODUCTION**

The History of the National Football League (NFL) spans 99 years. The league represents 2 conferences, The National Football Conference (NFC) and the American Football Conference (AFC) who represent 32 teams.  These 2 conferences merged in 1966 to form the NFL we know today. One of the leading reasons for the merger was issues with player salaries.

## **Problem Definition**

Provide an optimal solution to the NFL team to minimize the amount spent on the Quarterback position players and help in deciding whether to change an existing player with a better player.

## **Why NFL?**

The NFL Teams spend a considerable amount on maintaining their roster of players. The NFL receives billions of dollars in subsidies from the government. This makes the NFL players’ lofty salaries a controversial subject. The quarterback is the most visible and the most important player on an NFL team. Therefore, we will analyze the quarterbacks’ salaries.

## **How?**

* Understand the data of NFL
* Use the historical data to generate insights about the game
* Find out the factors affecting the problem we are trying to solve
* Perform required analysis on the data
* Use the value of data and solve the problem specific questions

In the project, we attempt to study and model the NFL player’s Statistical data to answer the following questions:

* How much are quarterback salaries?
* How are quarterback salaries expected to increase?
* Should teams replace an expensive quarterback for one with a lower salary and better rating?

# **DATA RESOURCES**

We have taken a dataset from Kaggle which consists the data about all players in NFL teams.

(<https://www.kaggle.com/zynicide/nfl-football-player-stats>)

The other dataset which we used specific Quarterbacks stats is from the website

(<https://www.lineups.com/nfl/player-stats/quarterback-qb-stats>)

The provider of the Team Stats is a company that specializes in providing NFL betting data. They have scrapped the data from sportsline.com.

## **Credibility and Reliability of the Data Resource**

We find the data to be appropriate as the datasets have data from 1900’s to recently updated data and also the points on which we choose the datasets are,

* Both the datasets have a high update frequency
* High usability rating on the Kaggle website
* Provenance of the data is specified for both the datasets

# **ANALYTICS**

## **Data Preprocessing:**

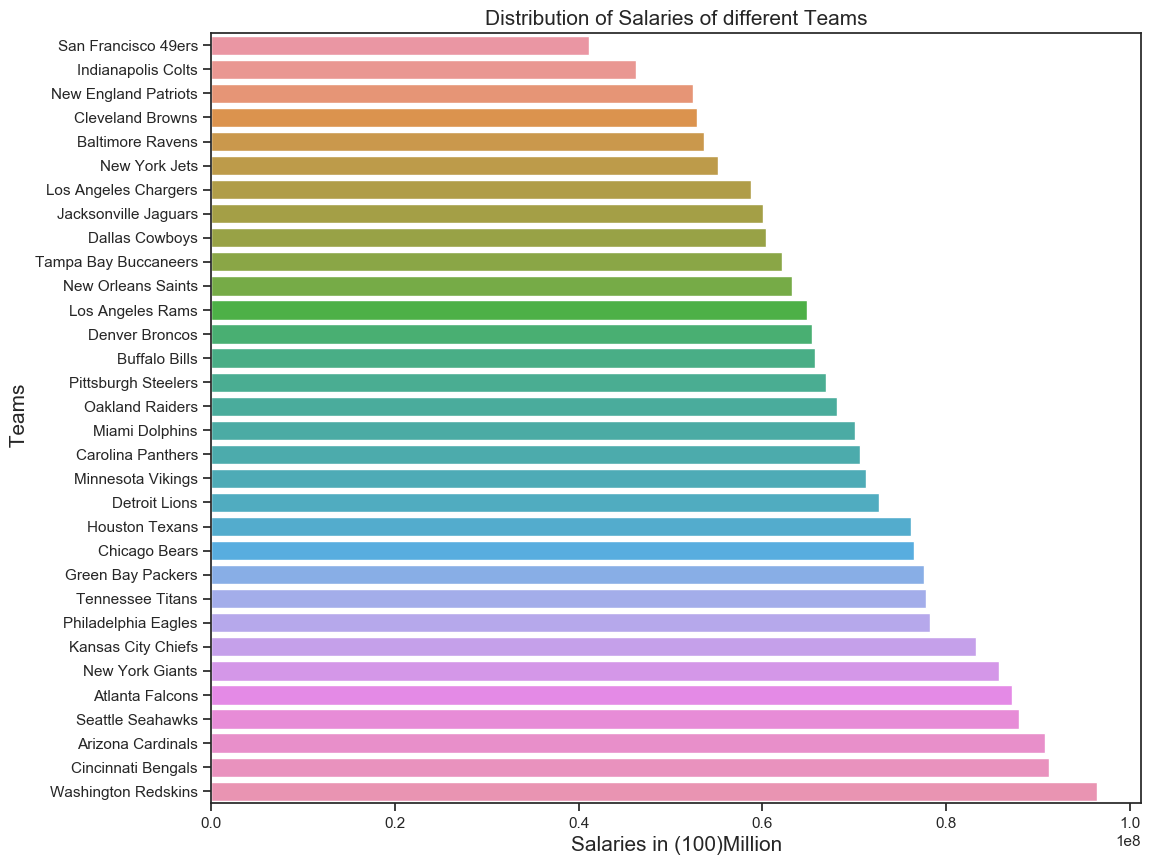
Following steps were performed in order to process our data:

1. First, we explored the datasets to get an idea of all the information present in datasets. Profiles Dataset has 17 columns and 25043 rows with int, float and object datatypes. The Quarterback data has 19 columns that shows various playing characteristics of a player in the field.
2. Dropping details of the players who are not alive were deleted because the salary values for the dead players were missing in the data.
3. Data cleaning for integer values were performed and then we calculated the player experience and players age based on the draft year and birth date columns respectively.
4. Finally we merged the profiles dataset with Quarterback dataset to get the complete details of the Quarterback which included the salary and all the playing traits of the players.
5. Finally the dataset was filtered for only Quarterback position because we are interested in the Quarterback only. The null values in the data were also removed.
6. Final dataset used for the for the predictive analysis.

## 

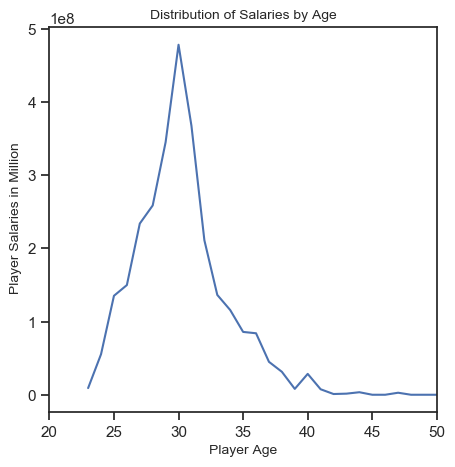
## **Data Visualization**

From the graph below, we can observe that the Washington Redskins spend the most on their team’s total salary while Cleveland Browns spend the least on their team’s total salaries.

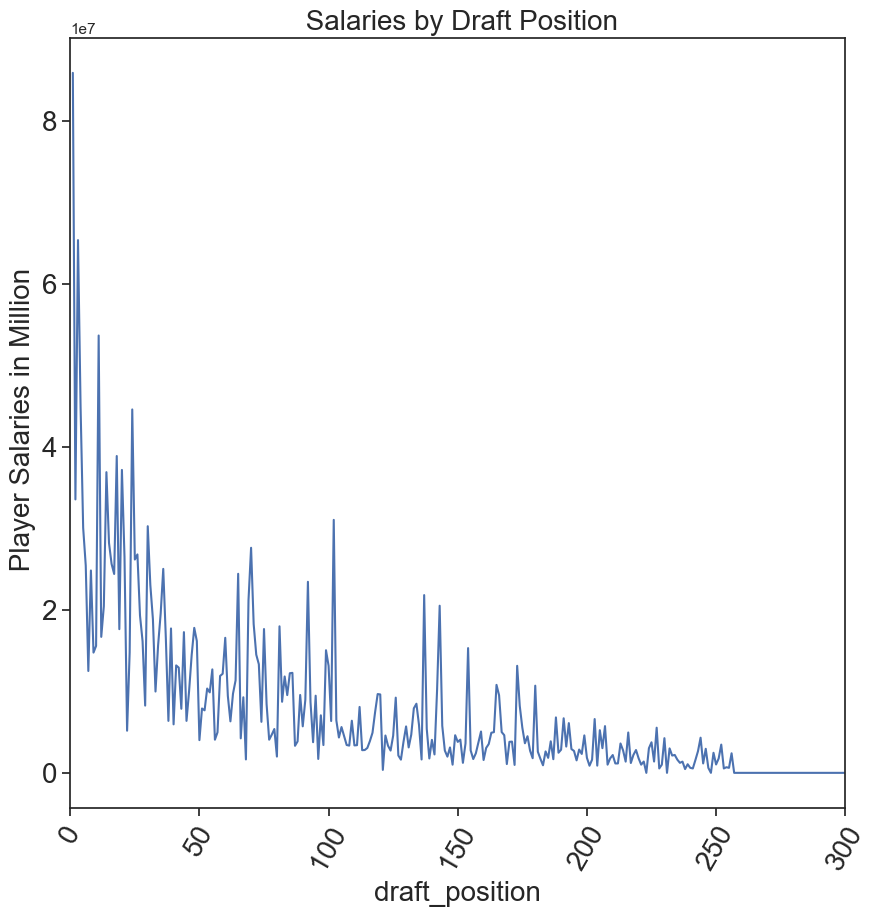


*Figure 1: Distribution of Salaries in Different Teams*

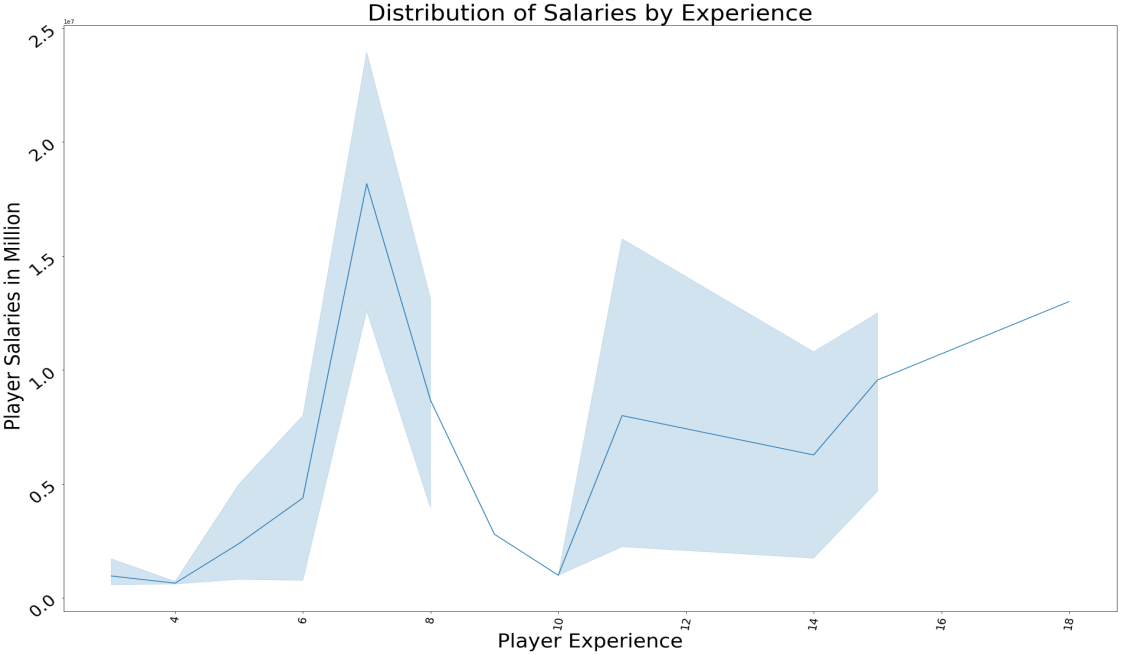
From the graph below, we can observe that the annual salary of NFL player peaks between at the age of 27-30 years.



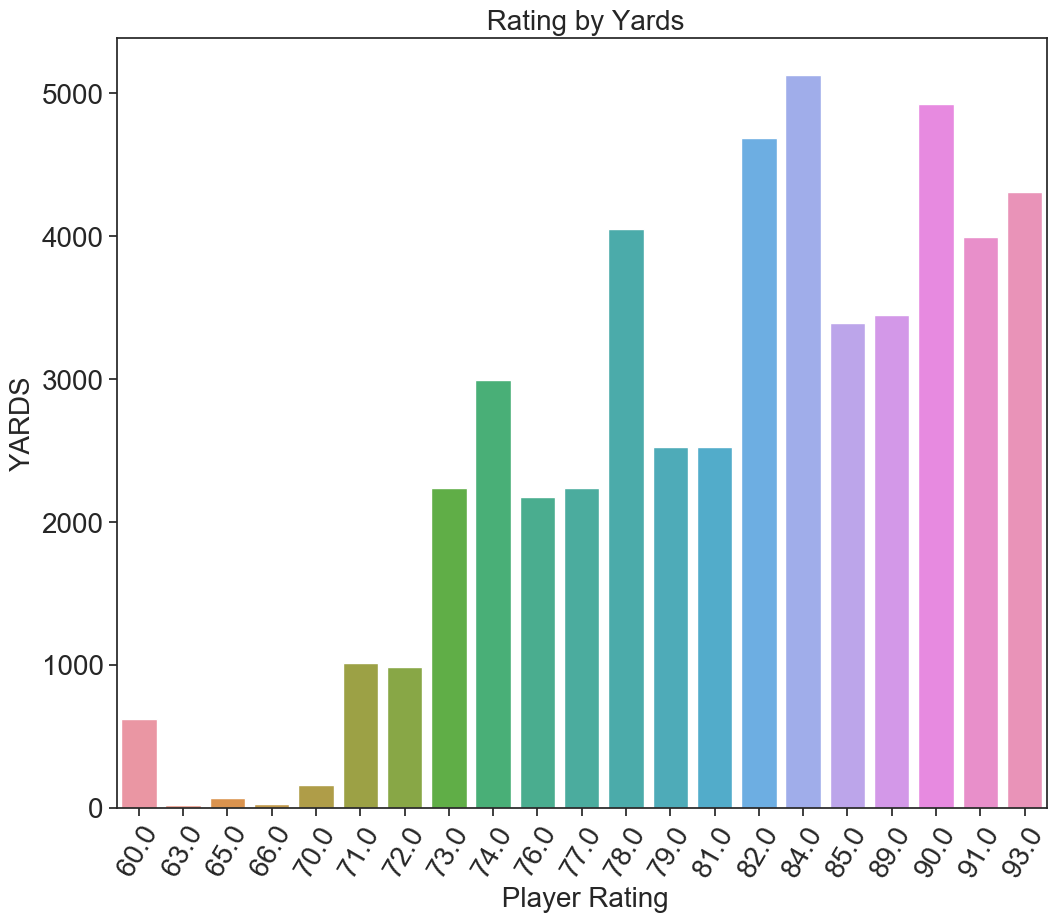
The draft position has a major role in NFL player’s salary. The early drafted players tend to have higher salary. It may also be because better players are chosen early in the draft process.



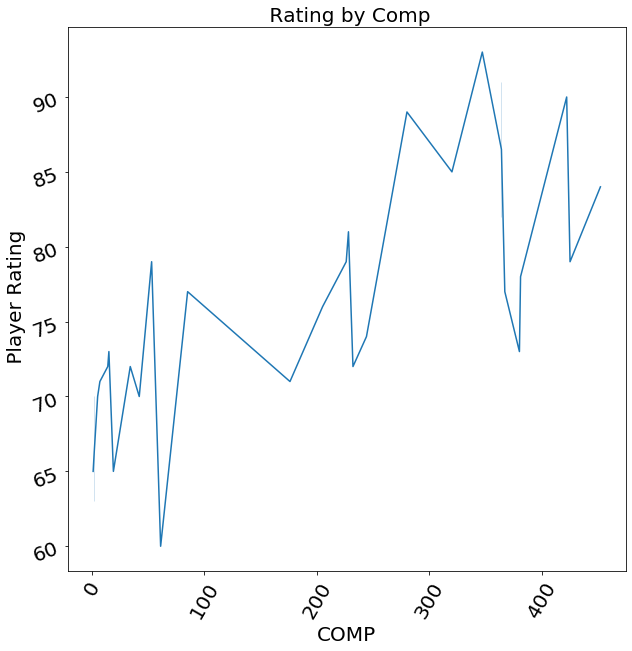
It was observed from the data that players with the experience of 6-8 years has the highest salary value. The salary tends to decrease linearly after 7 years. One major factor that plays vital role is that the productivity of NFL player tends to decrease sharply after they turn 28 years or older.



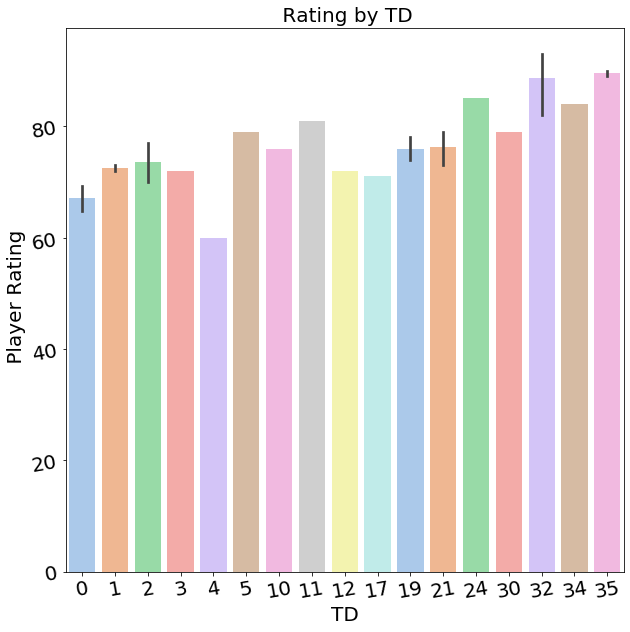
It is observed that the higher rated players have more yards per game value. Higher rated players seems to have more active role in moving the ball in order to achieve touchdown which are primarily focused in modern times to score more points.



The completion percentage (COMP) also has an effect on the player’s rating. It was observed that higher rated players have higher completion percentage. Completion percentage is



Touchdowns are the main focus of teams. Touchdowns are worth 6 points towards the team’s total. Quarterbacks tends to score more touchdowns that other players. It was observed that higher rated quarterbacks score more touchdowns that lower rated quarterbacks.

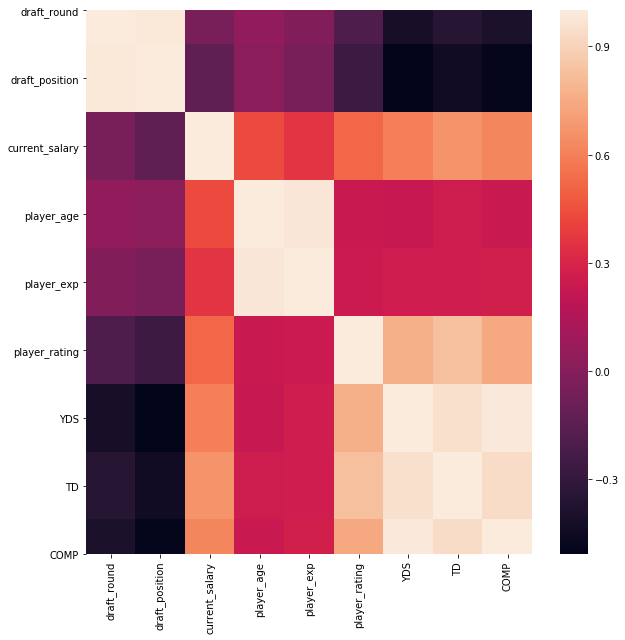


## **Predictive Analytics**

Predictive analytics is a statistical technique that help us make prediction about the future or unknown events by analysing patter and trends in the historical and current data. The likelihood of future outcome can be predicted using predictive analytics. Although, machine learning algorithms can make successful predictions, it is important to remember that no statistical algorithm can predict with 100% certainty.

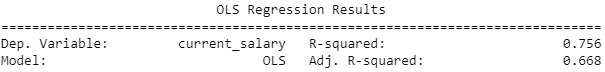
In our data we are using the linear multiple regression model to predict the salaries and ratings of the players.

Before we start predicting let us look into the correlation using heat map so that we can clearly understand which values are highly correlated with each other.



## **Prediction of Salaries**

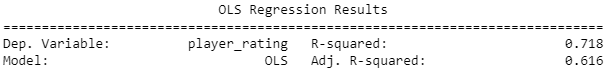
We are predicting the salaries of quarter back players using linear multiple regression model as given below,

As multiple linear regression attempts to model the relationship between the one or more independent variables and a dependent variable by fitting an equation to the observed data, in our model the dependent variable is considered to be the current salary and the independent variable are player age, player experience, player rating, draft round, draft position, yards, touchdown points, completion percentage values. 

We can observe that the R-squared value is 0.756 which indicates that the current salary is about 75.6% dependent on player age, player experience, player rating, draft round, draft position, yards, touchdown points, completion percentage values.

## **Prediction of Ratings**

Similarly, we are predicting the ratings of quarter back players using linear multiple regression model by taking the dependent variable as player rating and the independent variables are player age, player experience, current salary, draft round, draft position, yards, touchdown points, completion percentage values.



We can observe that the R-squared value is 0.718 which indicates that the player rating is about 71.8% dependent on player age, player experience, current salary, draft round, draft\_position, yards, touchdown points, COMP values.

Since the R-squared value is greater than 70% in both the predictions we predicted new salaries and new ratings for the players and created a dataframe from those values in order to use them for our decision making process.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Predicted Data** | **Existing Data** | |  |  | |  |
|  |  |

## **Prescriptive Analytics**

From the above preliminary analysis, we are considering below problem statement for optimization.

Problem statement:

To minimize the amount of salary spent on QB’s in the team. The salary amount is calculated by predicting whether the replacement of existing player with a new player is needed or not.

Considering the predicted salaries and ratings,

In New York Giants team , there is one QB player whome the teams manager wants to replace with an identified potential QB player from Seattle Seahawks..

Now, let us consider the new player being taken into the team as P1(player1) then (1-P1) would be the player in the team whom we are going to replace.

So our objective would be ,

Objective =minimize ( Player1\*Player1\_Predicted\_Salary + (1-Player1)\*Existing\_Player\_Predicted\_Salary)

Let us consider a constraint , when a new player is being taken into the team he is expected to have a rating above than the existing player.

So our contraint would be,

Player1\*Predicted\_Rating+(1-Player1) Existing\_Player\_Predicted\_Rating >Existing\_Player\_Predicted\_Rating

When we try to optimize this problem we are making a decision to either pick a new player or not which could be easily implimented with a Gurobi Optimizer(Mixed Integer Programming model).

Which gives us a binary output either 1 or 0 saying 1 to choose a new player or 0 to keep the same old player.

## **Conclusion :**

From the output of Gurobi Optimizer in order to minimize the amount of salary spent there is no requirement of replacing the existing player with a new player.

## **Shortcommings**

If we had the historical salaries for the entire 99 years, then we can have a more precise model.