OPIM 5604

Spring Semester

2016

Group Project

Team-3

Instructor:

Iva Stricevic

Team Members:
Ashish Gupta
Guanwei Tao
Nachiket Garge
Sindhushree S M
Ziou Zhang

EXECUTIVE SUMMARY

American Football is the most popular sport in America. The organization which organizes the biggest American Football tournament is the National Football League (NFL). This professional American football league consists of 32 football teams, divided equally between the National Football Conference (NFC) and the American Football Conference (AFC).

Our Project team found an interesting area in this field which is what led to the start of this project. Every football player is paid a salary as part of a contract. When a player's contract ends, the player is addressed as a 'Free Agent', who is free to be picked by any team for contract. Deciding salaries for such players has in it, the scope of building a predictive model.

Objective:

- As a team, learn predictive modelling concepts through applying SEMMA approach effectively. Research and learn about NFL as an example to build a predictive model. Based on team experiential learning process with this project, to better understand the data exploration, data visualization, pattern discovery, modelling and better interpret results obtained from the model.
- Build a predictive model with enough capacity to predict the salaries of future free agents based on their previous performance and salary information.

"The work contained and presented here is the work of Spring 2016 class Team#3 and Spring 2016 class Team#3 alone"

TABLE OF CONTENTS

I. Data Collection and Consolidation

The data collection was done from the below sources –

http://www.spotrac.com/nfl/freeagents

http://www.nfl.com

This data is available for free on the mentioned pages without any licensing requirement.

The data collected was spread across multiple files. The files had data representing the salaries of every player from 2012 to 2016 and their performance statistics from the year 2011 to 2015. We then segregated the data according to our business objective's requirements and created a final JMP file. This file has a total of 289 rows.

The JMP file that we created has 24 variables. The variables and their description is given in the Data Dictionary.

The Data dictionary was available in the same sources from where we collected the data and is a part of the Appendix.

Assumptions:

Based on the data collected, we made a few assumptions for our modeling:

1. NFL decides the salary of a player based on their previous year's performance.
2. NFL decides the salary of a player based on their previous average salary.

II. Data pre-processing

Sampling:

In our dataset, we had population of 1023 football players. Every player belongs to one of the following field positions as an Offensive player (See Appendix for definition):

Quarterback
Center
Running back
Fullback
Wide receiver
Tight End
Left guard and right guard
Left tackle and right tackle

Every player belongs to one of the following field positions as a Defensive player (See Appendix for definition):

Defensive tackle
Defensive end
Linebacker
Safety
Cornerback

One of the major issues in considering all 1023 players in a single JMP file was that certain player positions have their set of performance metrics while the others will not share the same performance metrics.

E.g. A player in Centre position will majorly participate in 'passing' the ball to Quarterback, rather than 'receiving' it. This means that we will have data for performance variable of 'receiving' for Quarterback but will not have it for Centre player.

Therefore, the approach we took as part of Sampling was to include only the 'Quarterback' position players for modeling. This sample data has 289 rows (289 Quarterbacks).

Missing Values:

	Count	Number of columns missing	Patterns	Player	Team	Pos	average
1	286	0	00000000000000000000000	0	0	0	
2	2	1	00000000000000000001000000	0	0	0	
3	1	2	00000000000000000011000000	0	0	0	

Fig 2.1

Fig 2.1 shows the missing values that we found in our JMP file. The 3 rows that had the missing values were deleted for our Modeling. This was because we did not get data for the players in order to perform missing value imputation.

Correlation:
We noticed there were some variables strongly correlated and few others which were not. There were few variables which were positively correlated and few others that were negatively correlated. (See Appendix for correlation charts)
Few of the major correlations we identified are as follows:

Variable1 vs Variable2	Correlation	Comments	
Rank vs Average salary	-0.7	As the value of Rank increases, the Average salary decreases. i.e. A player of rank 10 is paid lesser than the player of rank 1.	**Correlations** average salary: 1.0000, Rank: -0.7019 Rank: -0.7019, 1.0000
Attempted pass vs Completed pass	0.99	When a player attempts a pass, it is very likely he completes it. Therefore, the high correlation is justifiable.	**Correlations** Completed pass: 1.0000, Attempted pass: 0.9949 Attempted pass: 0.9949, 1.0000
Year vs Completed pass	-0.03	The column Year shows which year's file the row was fetched from and therefore has no correlation with any of the performance or salary variables.	**Correlations** Completed pass: 1.0000, Year: -0.0322 Year: -0.0322, 1.0000

Treating highly correlated variables:

We have learnt that if we run models with highly correlated variables, we may get model with high R-square which may look perfect, but in fact, it would a failure. Metaphorically we can say that it is like comparing a person with himself; the model will have no practical meaning. Therefore, it is essential to eliminate such variables as per the business sense.

Performing principal component analysis on highly correlated variables, gives the mathematical dependencies of variables but wouldn't give more of the business sense. Therefore, the approach we took towards elimination of highly correlated variables was using the method of feature engineering (See Appendix for definition) instead of performing Principal Component analysis.

i.e. We created a variable 'QB Rating' based on the formula fetched from the website http://www.nfl.com/help/quarterbackratingformula

The formula of QB Rating used in the world of NFL Rankings -
(25 attempted_pass + 1000 completed_pass + 50 yards + 4000 touchdown - 5000intercept)/12attempted pass

Fig 2.2 below shows the correlation between the above mentioned variables.

Multivariate

Correlations

	Completed pass	Attempted pass	Yards	Touch Down	Intercepts
Completed pass	1.0000	0.9949	0.9933	0.9368	0.8527
Attempted pass	0.9949	1.0000	0.9896	0.9191	0.8762
Yards	0.9933	0.9896	1.0000	0.9501	0.8459
Touch Down	0.9368	0.9191	0.9501	1.0000	0.7203
Intercepts	0.8527	0.8762	0.8459	0.7203	1.0000

Scatterplot Matrix

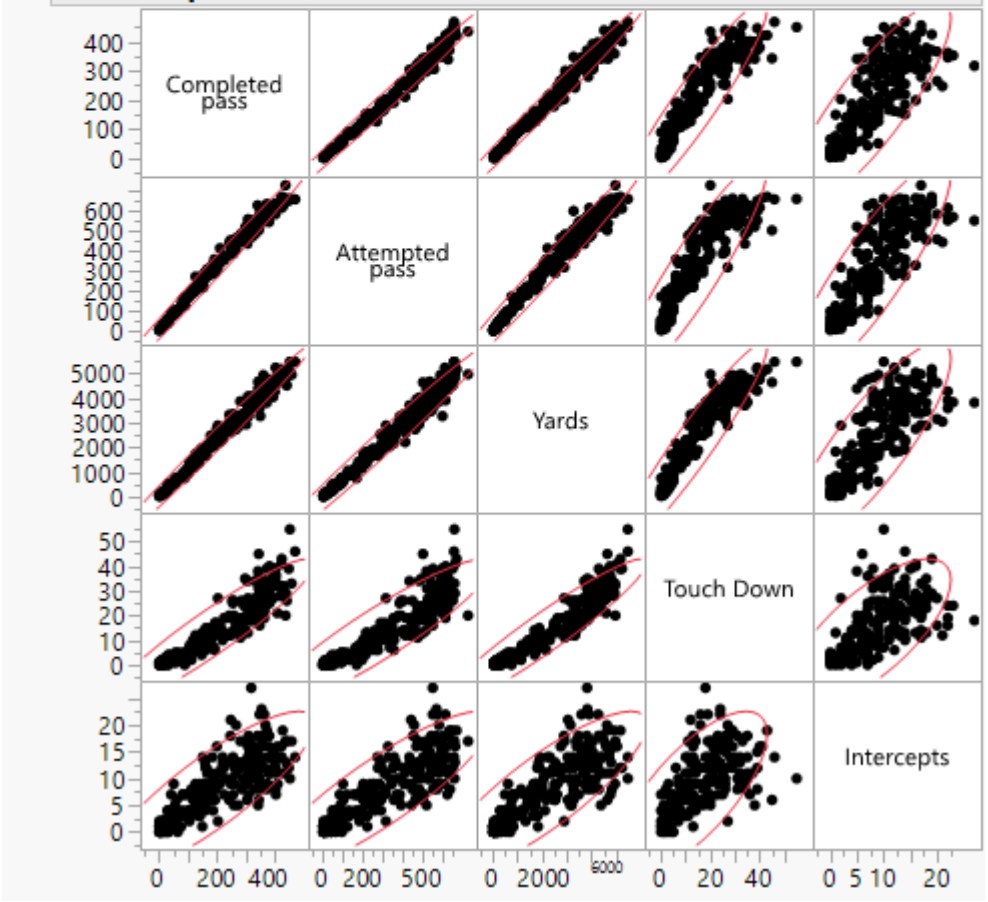


Fig 2.2

Distributions:

From the distributions of variables (See Appendix for distributions), we can see that the target variable 'average salary' is not normally distributed. Normal distribution of average salary becomes essential since it is a requirement in regression models to have a normally distributed target variable in order to get a good model.

If the target variable is right skewed Log transformation is preferred. Our target Variable, average salary is right skewed, and therefore, we performed Log transformation to normalize the target variable.

Advantages of using Log transformation on average salary are:

- To narrow down the range as average salaries had huge variations, ranging

from $450,000 to $22,133,333.

- Easy reconversion to interpret the final results.

Distribution of average salary before and after transformation

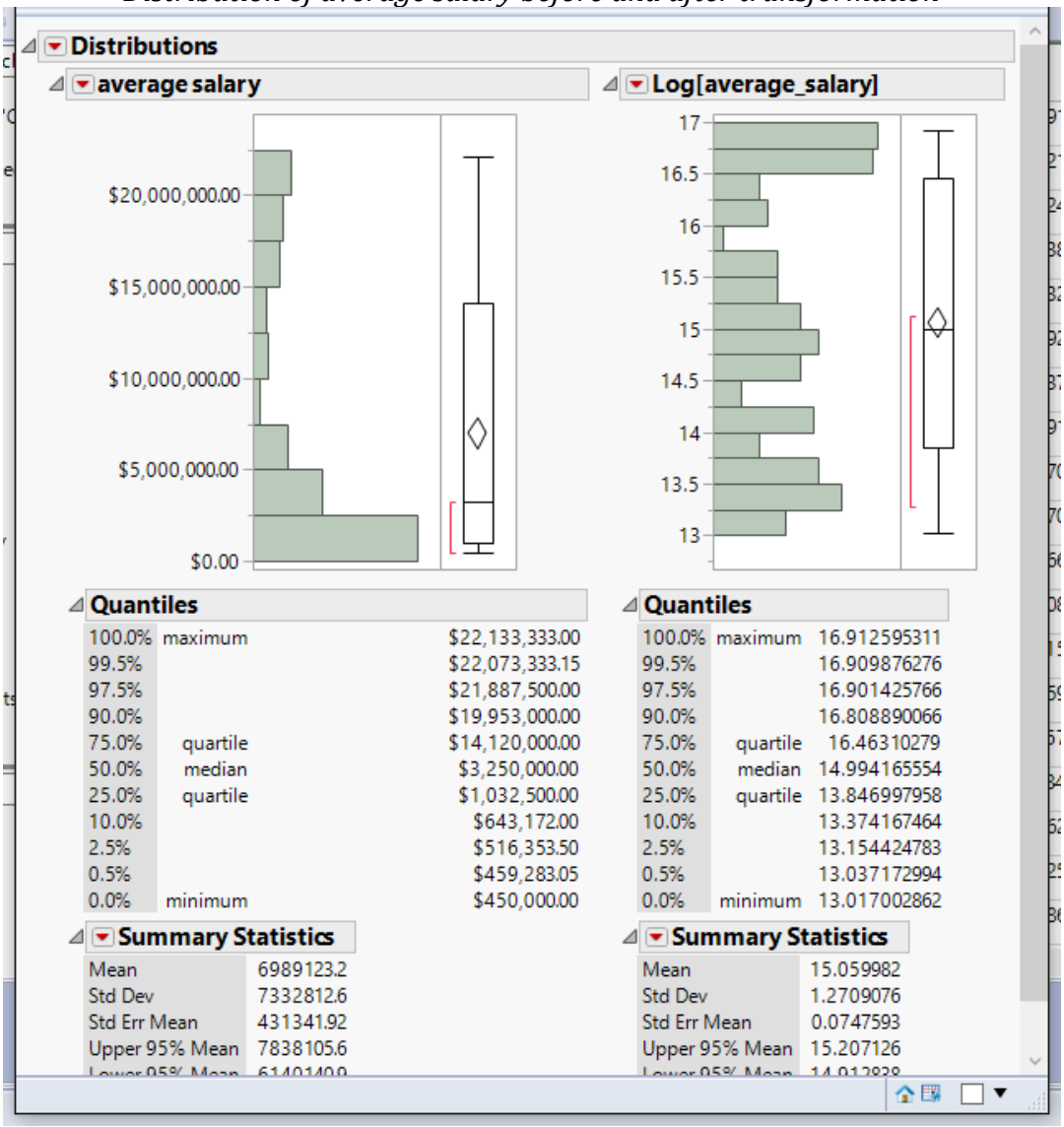


Fig 2.3

Outliers:

Outlier analysis was performed for all the variables. Fig 2.4(a) – Fig. 2.4(d) show the distribution of various important variables.

Before identifying the important attributes, distribution of all variables were conducted and are shown in the following charts.

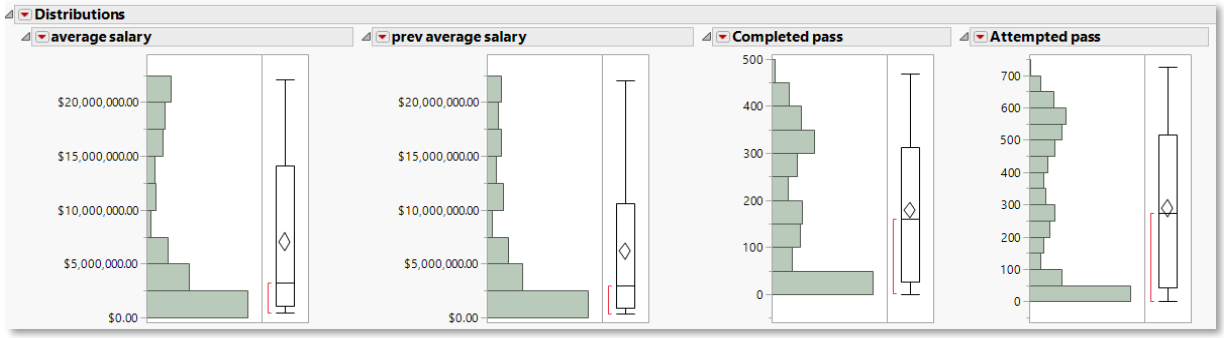


Fig 2.4(a)

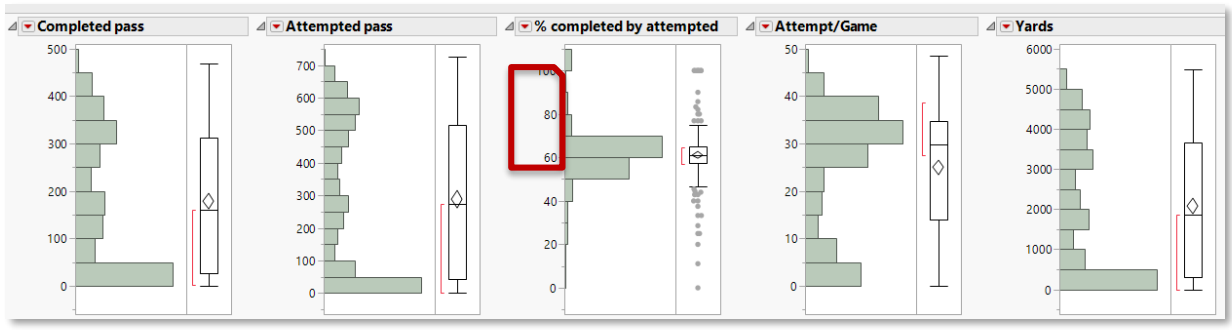


Fig 2.4(b)

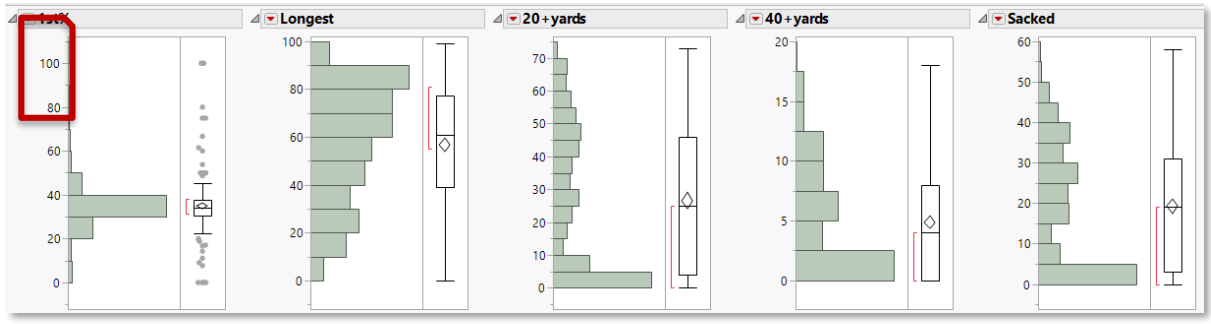


Fig 2.4(c)

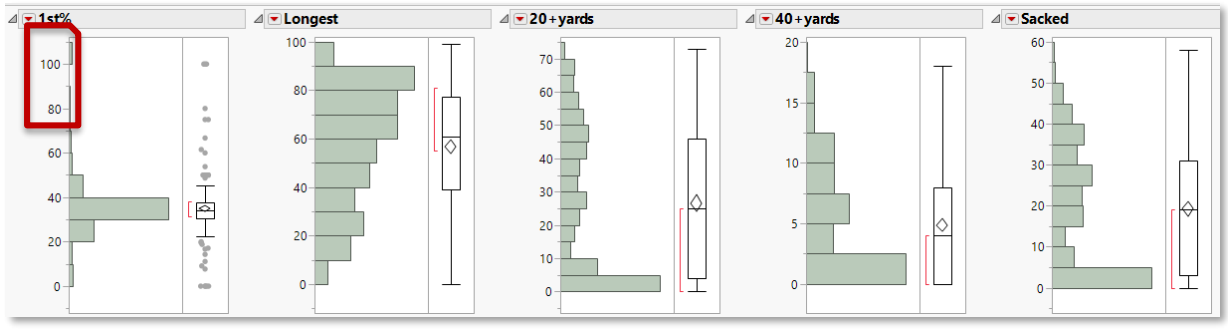


Fig 2.4(d)

By identifying distributions, "%completed by attempted", "1st%", " QB_Rating", "yards" have significant outliers. However, more analysis evidence was required to support the decision of keeping or discarding those outliers.

According to results from Prediction profiler run by "fit the module"(shown in Fig 2.5(a) and Fig 2.5(b)), we concluded that variables of "%completed by attempted"," 1st%", "QB_Rating", "yards attempted", are not significantly impacting the salary. Therefore, the outliers from those four variables were retained in the data.

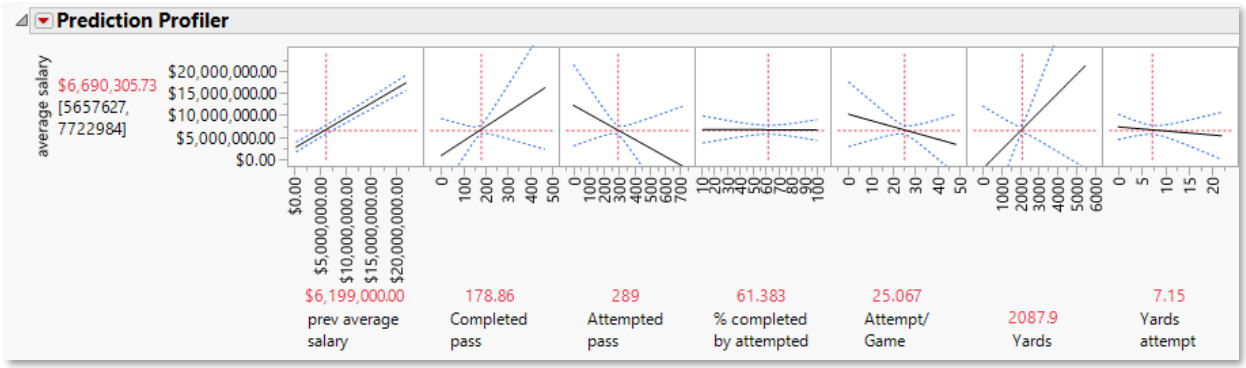


Fig 2.5(a)

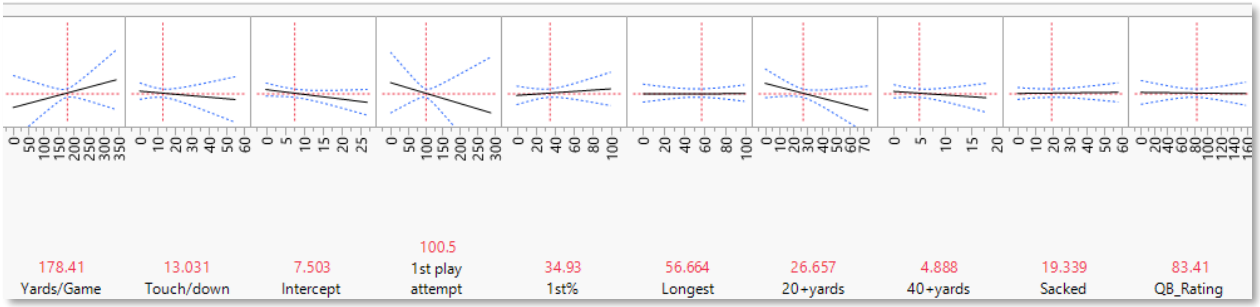


Fig 2.5(b)

Further, in our data set we have only 289 rows. Therefore, we did not decide to remove outliers because even if there are only a few rows for a particular performance variable as outliers, they'll contribute a significant percentage to the total data. In fact, the amount of data available is too less to state with confidence that the identified rows are outliers.

However, we removed a few rows based upon the following approach:

We computed the differences in current salary and previous year's salary of each player. Fig 2.6 shows that there are few rows with huge differences in the current and previous year's salaries.

Distribution of difference in salary of previous year and current year

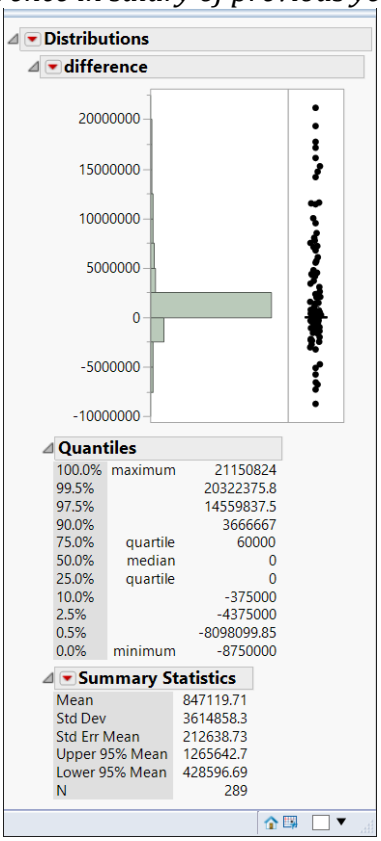


Fig. 2.6

To be able to predict for these rows, the variations in performance variables would also have to be high. However, the data of performance variables varies in a particular range and such high differences in performance variables are not possible. Therefore, we interpreted that there are other factors contributing to such high raise in salaries and excluded the rows (accepting the limitation of the data) that have very high difference in current year's and last year' average salary from the prediction. Let us illustrate this with an example-

Football player Brock Outlier who started his career in 2012 and improved his performance statistics over the years, has a salary of $18,000,000 currently.
Fig 2.7 shows his average salary of 2016 with performance statistics of 2015 and his average salary of 2015 with performance statistics of 2014.
His average salary is a whopping $17,120,830 more in 2016 when compared to last year (2015). To justify this high raise in salary, his performance statistics is not sufficient enough. Thus, pointing out that there are other factors (out of scope with respect to data available) for this difference in salary.

Player	Rank	Year	average salary	prev average salary	Attempt/Game	Yards/attempt	Yards/Game	1st_Play attempts	1st%	Longest	20+ yards	40+ yards
Brock Osweiler	31	2016	$18,000,000.00	$879,170.00	34.4	7.2	245.9	91	33.1	72	17	4
Brock Osweiler	65	2015	$879,170.00	$879,170.00	2.5	5.2	13	2	20	38	1	0

Fig 2.7

We identified 8 such players and excluded those rows for our models.

III. Data visualization and pattern discovery

After preprocessing of the data, we discovered some visual visualizations and patterns to give us a hint and to help us know more about relationships between the variables.

Clustering by Hierarchical method:

We performed Hierarchical clustering with all the performance variables. We divided the data in many different no. of clusters and analyzed the variation of all these clusters with current average salary. The one-way analysis plots of different numbered clusters with current average salary are shown in Fig 3.2 and Fig 3.3. Also Fig 3.1 shows the general dendogram structure of hirerchial clustering.

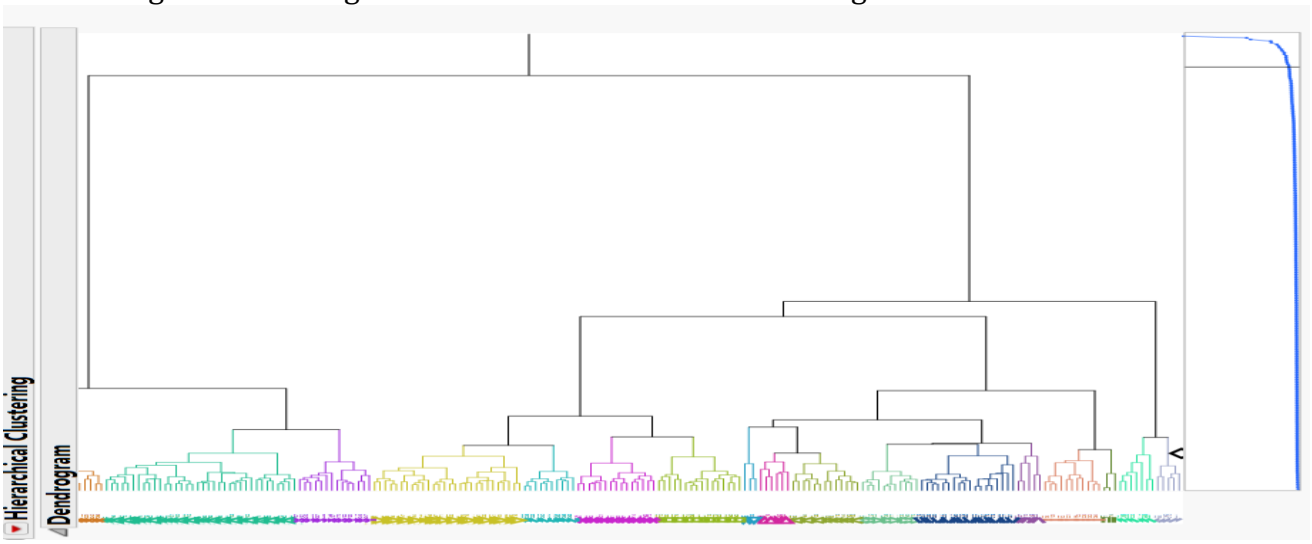


Fig. 3.1

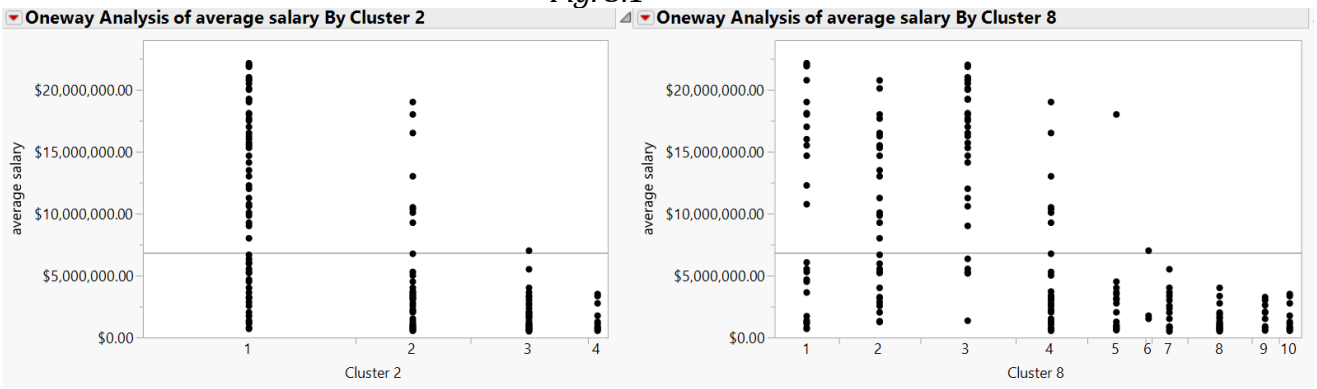


Fig 3.2

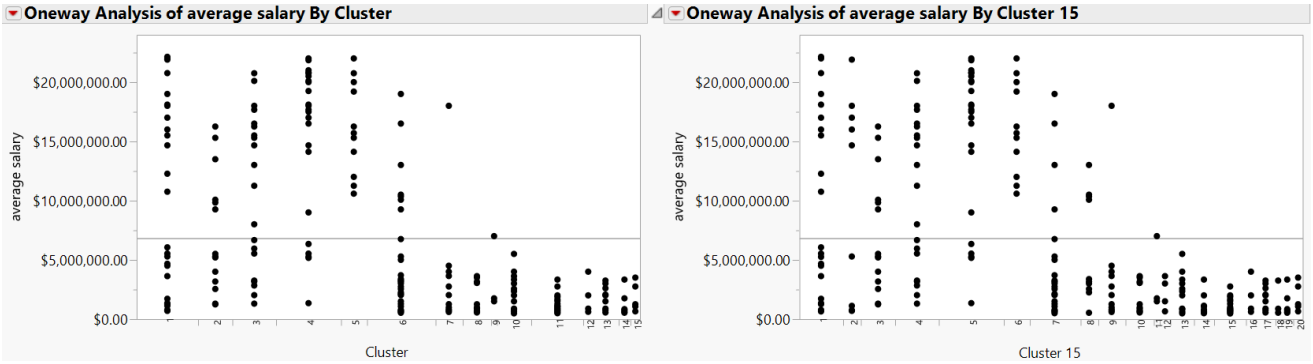


Fig 3.3

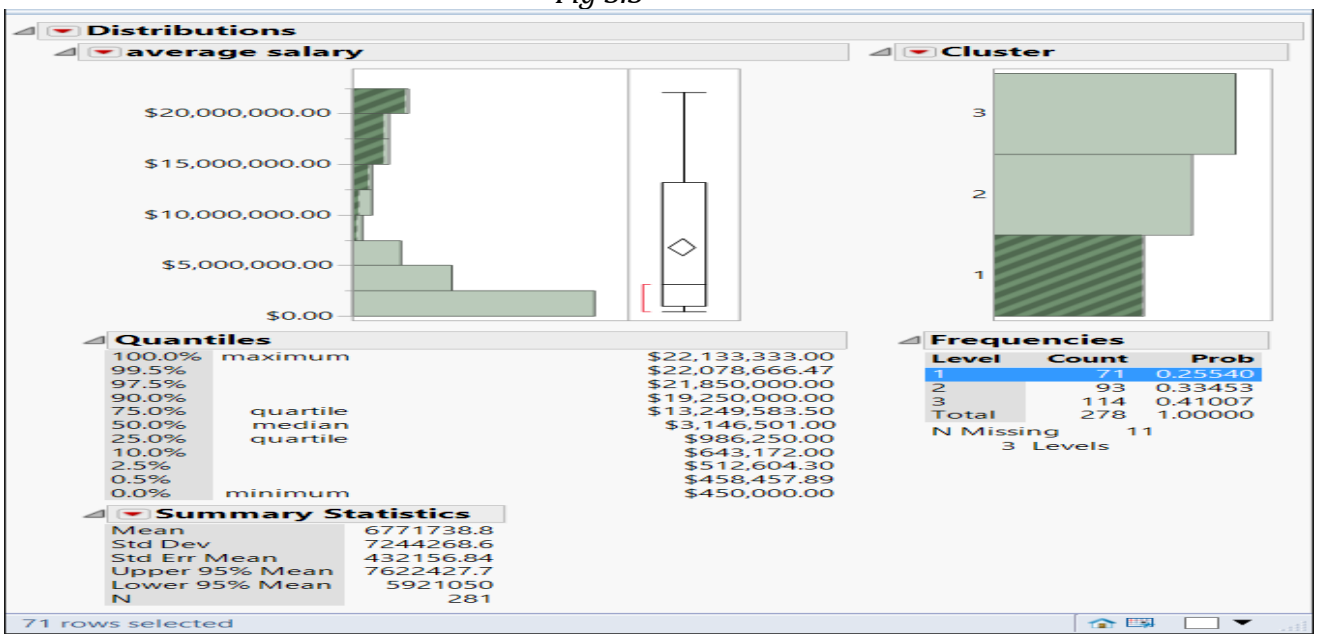


Fig. 3.4

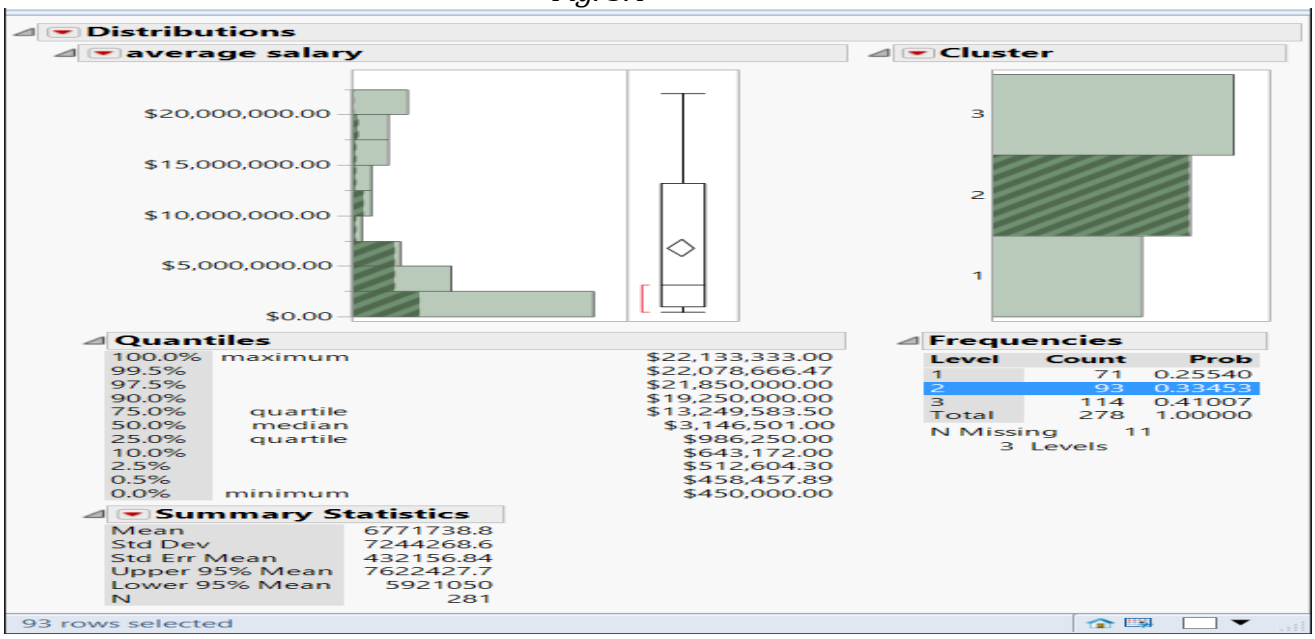


Fig. 3.5

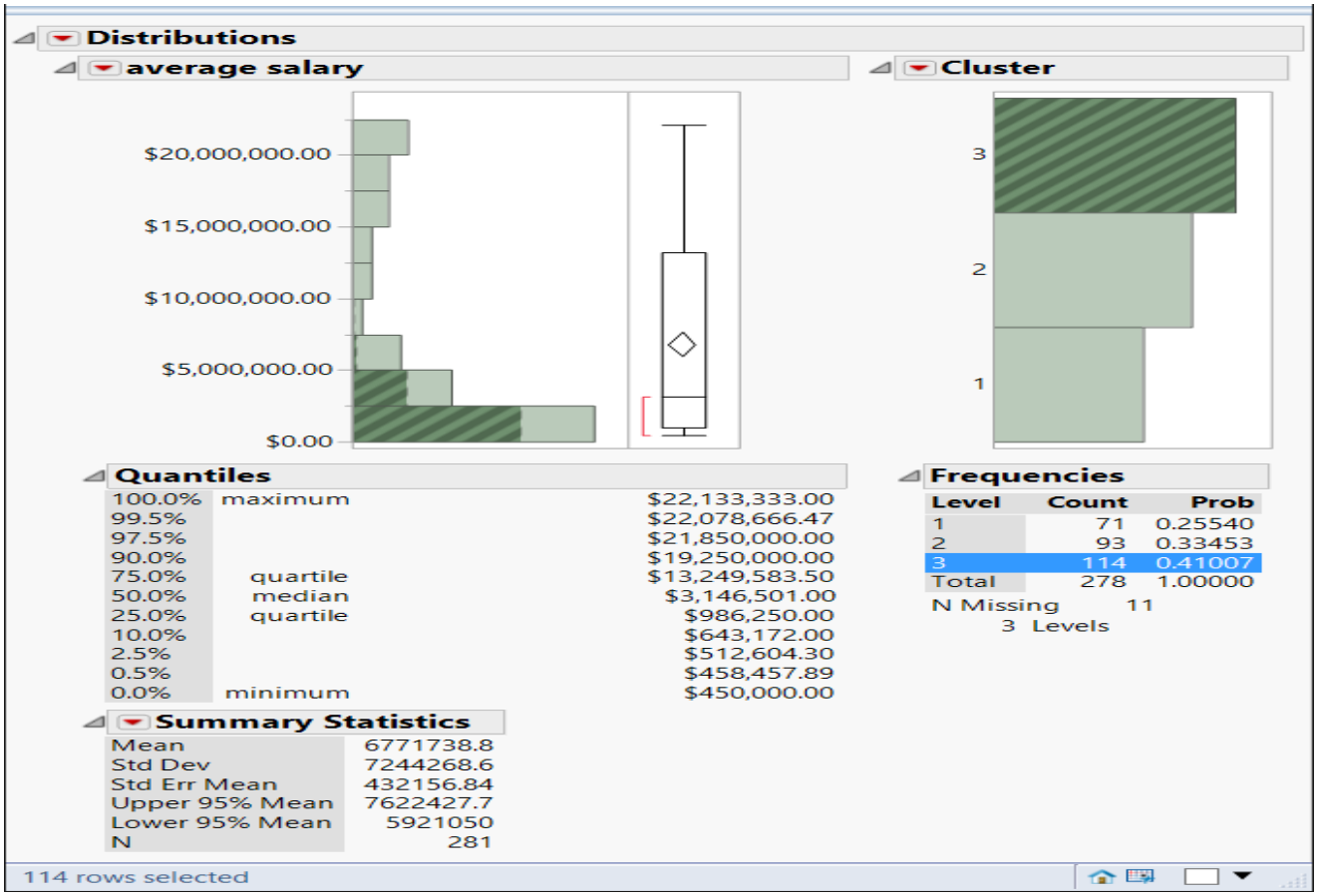


Fig. 3.6

From Fig 3.2 – Fig 3.6, we can infer that there is an underlying pattern of the performance variables that shows dependencies on the current average salary. There are some clusters that fall in lower salary region and some clusters that covers the total range of salary.

From Fig 3.1 -Fig 3.4, it can be seen that there is a conspicuous distinction between clusters belonging to lower and higher range of salaries as the decrease in current average salary is in one discrete step.

From Fig 3.4 – Fig 3.6, We also learnt that –

Cluster 1 has those players who are paid high salaries (Fig. 3.4)

Cluster 2 has those players who are averagely paid (Fig. 3.5)

Cluster 3 has those players who are paid below average salaries (Fig. 3.6)

Hypothesis Testing:

We performed 2-Sample t-test that calculates a confidence interval and does a hypothesis test of the difference between two population means when standard deviations are unknown and samples are drawn independently from each other. (samples are independent).

We performed the following tests:

- Tests whether the average of two different populations are significantly different
- Checked group variances are equal or not ~ ANALYZE -> FIT Y by X -> generate

graph seeing the target variable w.r.t two groups -> Normal Quantile Plot ->Plot Actual by Quantile

If variances are equal, slopes are parallel, if not equal then slopes converge and meet. We can do the unequal variances test for a more formal test of equal variances.

- Group variances are not equal: perform Generated Graph > t Test, check p values

Here our null hypothesis and acceptance hypothesis are,

H0 -> Pre Avg. Salary = Current Avg. Salary

HA -> Pre Avg. Salary ≠ Current Avg. Salary

Our Test showed,

Test Mean

Hypothesized Value	6142004
Actual Estimate	6989123
DF	288
Std Dev	7332813

Fig 3.7

It is evident that hypothesized mean is less than actual estimate.

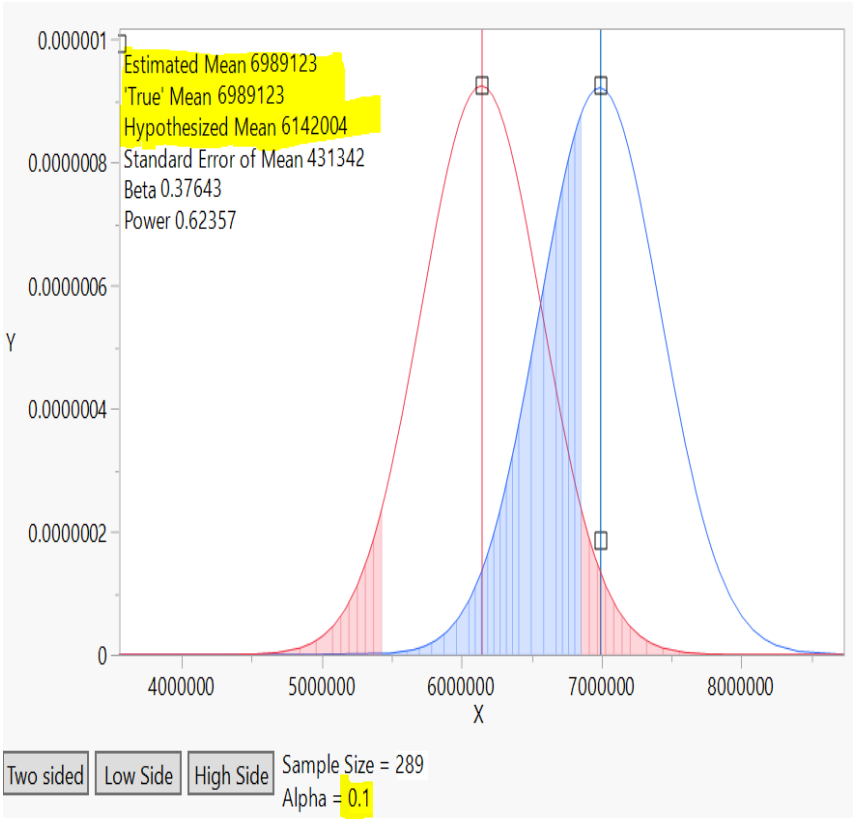
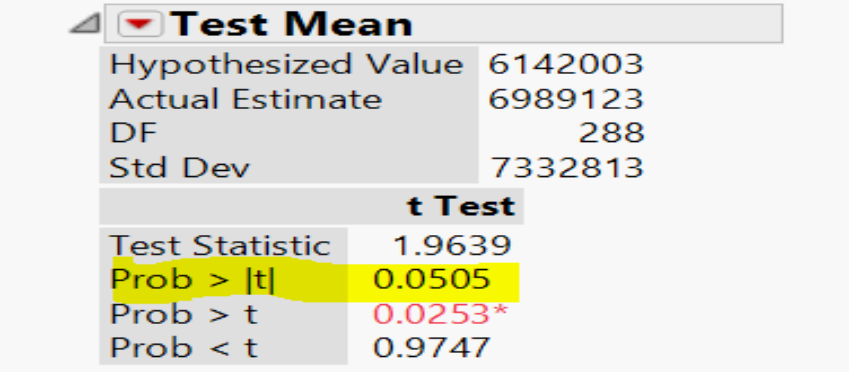


Fig 3.8



Test Mean

Hypothesized Value	6142003
Actual Estimate	6989123
DF	288
Std Dev	7332813

t Test

Test Statistic	1.9639
Prob > |t|	0.0505
Prob > t	0.0253*
Prob < t	0.9747

Fig 3.9

The results are evident from the two tailed T test that, P value is 0.0505, whereas our alpha value is 0.01 at 90% confidence interval.
As p value is less than alpha value we can reject our null hypothesis that average salary and previous average salary means are equal.

IV. Predictive Modeling

We split the Quarterback players' data into 2 sets – Training and Validation. 25% of the rows are taken as Validation data and 75% is Training data.
Our aim is to predict the salary of Quarterback players based on their performance statistics of previous year. The approach would be to first find out what variables in performance can be most significant factor to the predict the amount of salary.
We have considered three main models to predict.

1. Regression modeling
2. Decision tree
3. Neural modeling

Regression modeling:

Since our target variable, which is average salary is a continuous variable we built a linear regression model

Summary of linear regression model

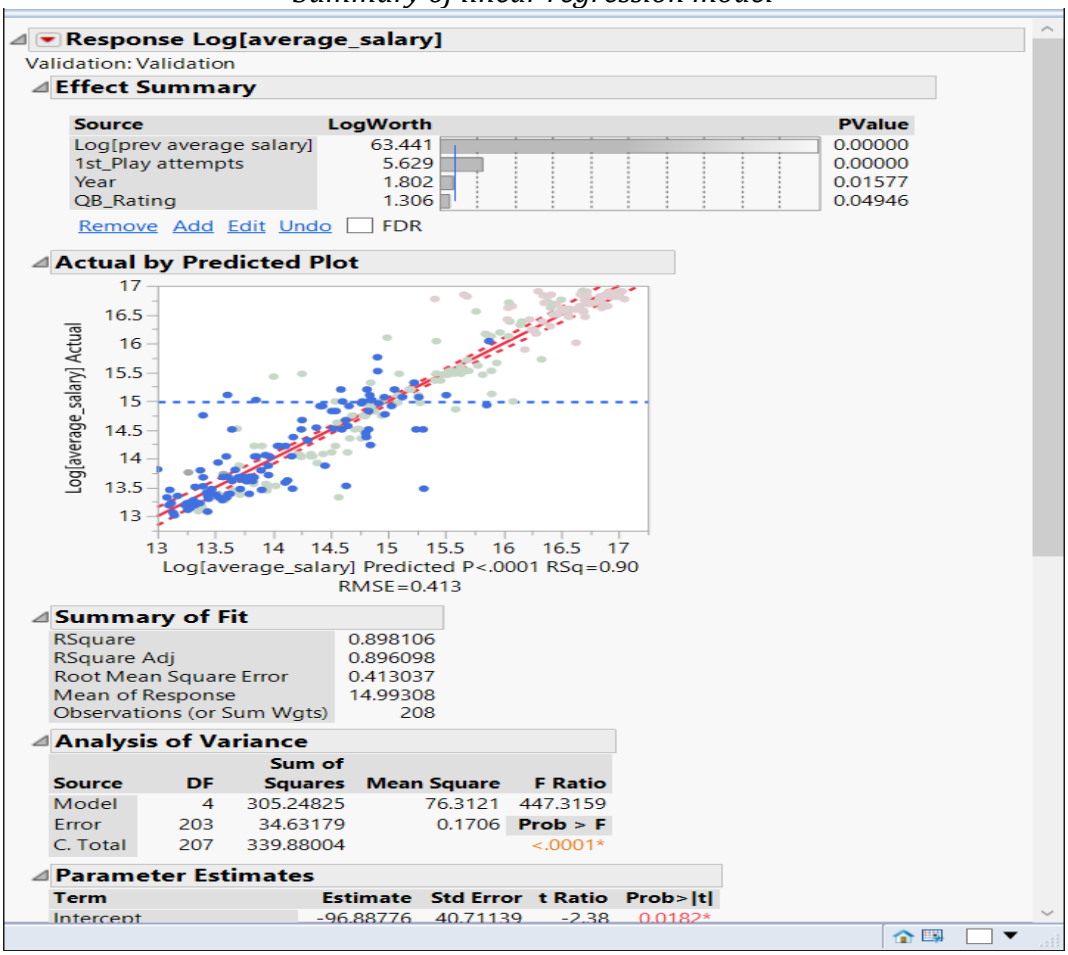


Fig. 4.1

Regression model shows that log(prev average salary), 1st_play attempts, year, and QB_rating are the influencing variables in predicting the current year's average salary. The R-square and adj R-square obtained are 0.89. This indicates that the selected variables explain 89% of variation in predicting next year's salary. The P-values of these variables are less than 0.05 (Fig 4.1) which states that the hypothesis that the intercept of these variables is zero in the prediction formula of the current salary can be safely rejected with a 95% confidence. Overall R-square for the model is 0.898 which is good, since the model fit is better as R-square value approaches 1.
Also the value of 'Prob>F' is less than 0.001 (Fig 4.1). This value measures the probability of obtaining a large F Ratio (See Appendix for definition) of 447.3. The obtained low value of 'Prob > F' indicates that the observed F Ratio is unlikely. Thus, it is evidence that there is at least one significant effect in the model

Regression Assumptions:

1) Linearity: Fig.4.2 shows the strong linear relationship.

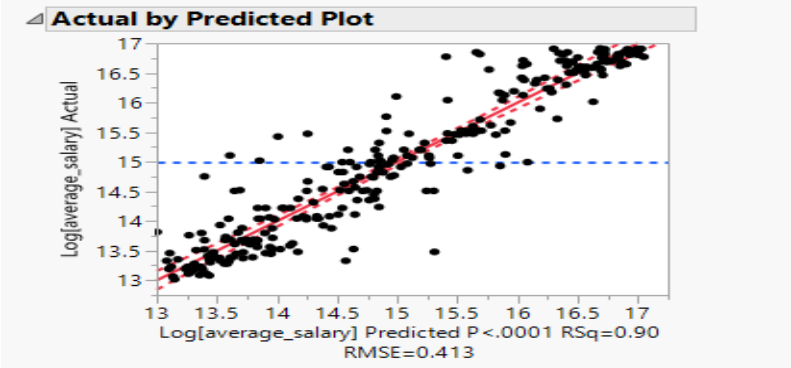


Fig 4.2

2) Homoscedasticity: Fig 4.3 shows that the probability distribution of errors has constanr variance.

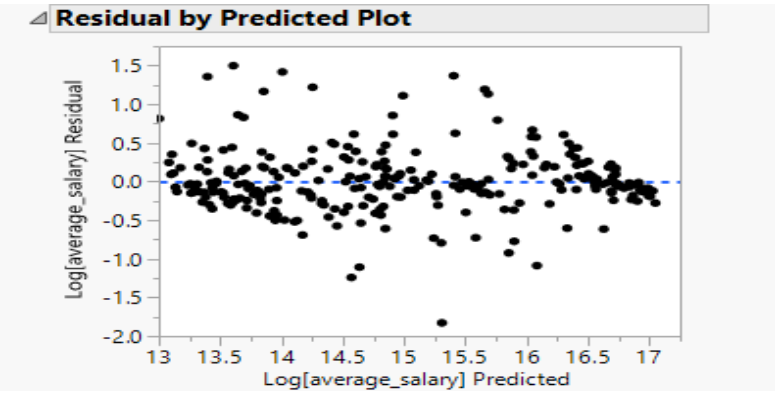


Fig 4.3

3) Normal Distribution of Error: Fig 4.4 shows that error values are normally distributed with predictor variables.

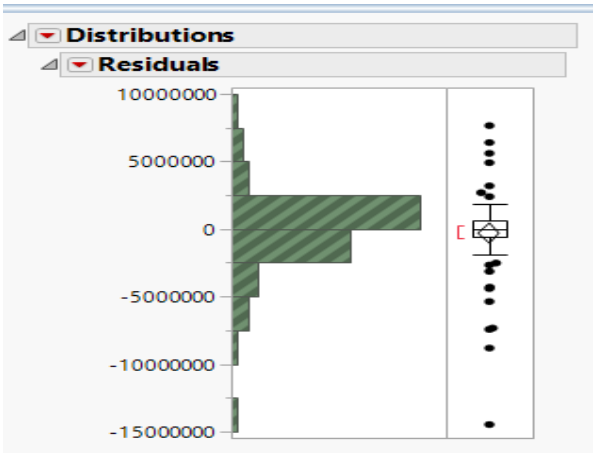


Fig 4.4

4) Independence of Residuals: No or little correlation of residuals with predictor variables.

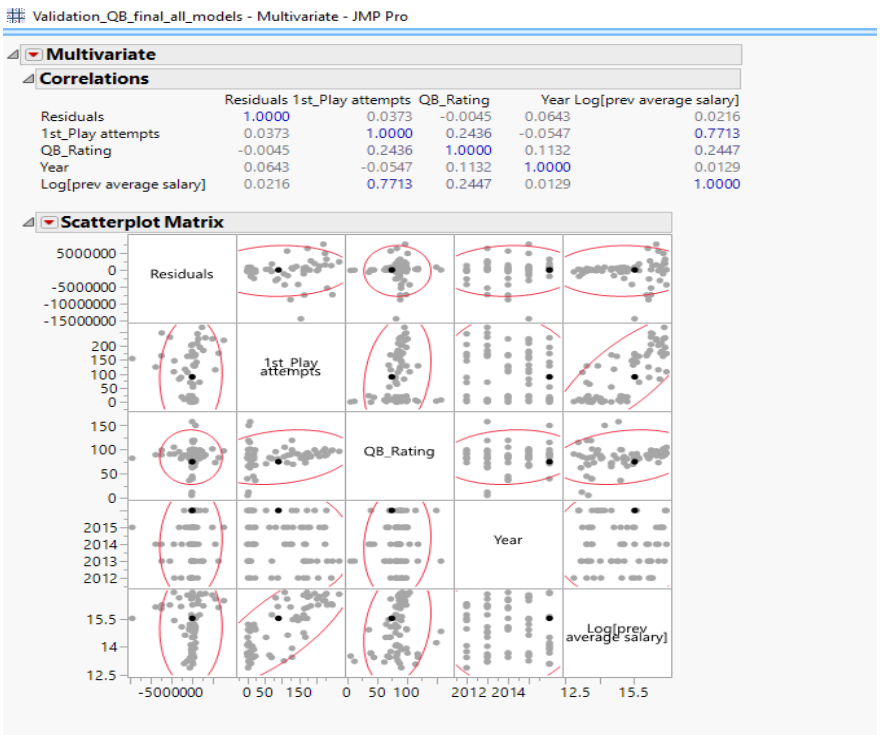


Fig 4.5

5) Check for Multicollinearity:

From Fig 4.5, we can state that the influencing variables in the model are not highly correlated as they have little or no correlation.

6) Error Perfomance:

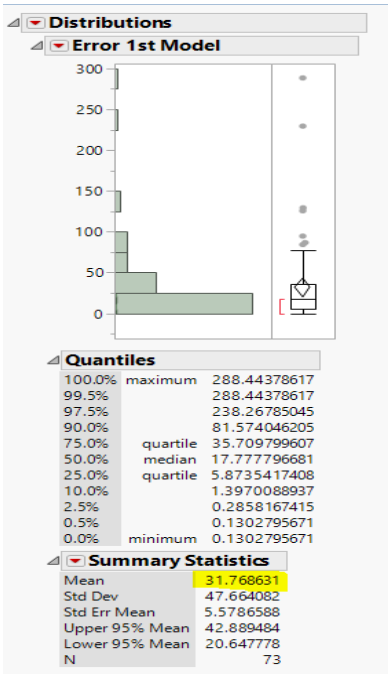


Fig 4.6

The error of regression model for the validation set is 31.7% (Fig 4.6). The lesser the error on validation set, the better the predictive power of the model. However, the error performance on validation set should also not differ much with the error performance on the test set as it would otherwise indicate a overfit or underfit model. From Fig 4.6, clearly that is not the case.

Decision tree

The second continuous model we built was decision tree. Decision tree model provides the best fit training model but it seems slightly over fit.

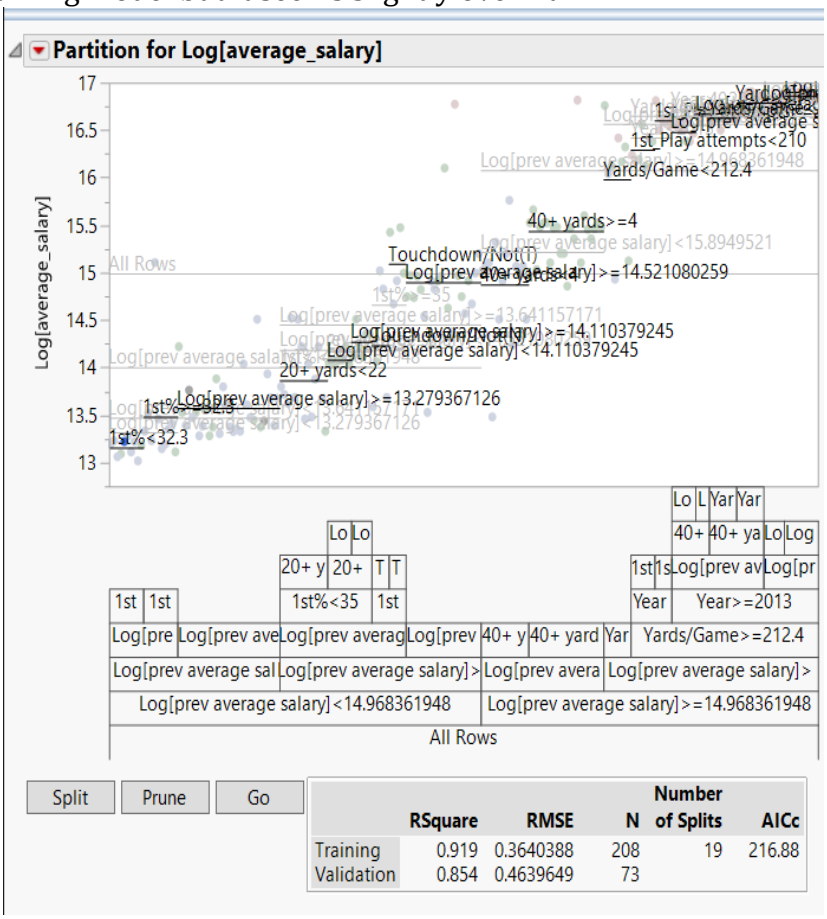


Fig. 4.7(a)

R-square of Training set is 0.919 whereas R-square of Validation is 0.854 (Fig 4.7(a)). However, being close to 1 suggests that the model is good. Further, Decision Tree Model also provided supportive evidence to Regression model as splits in decision tree are based on same variables that were significant in regression model.

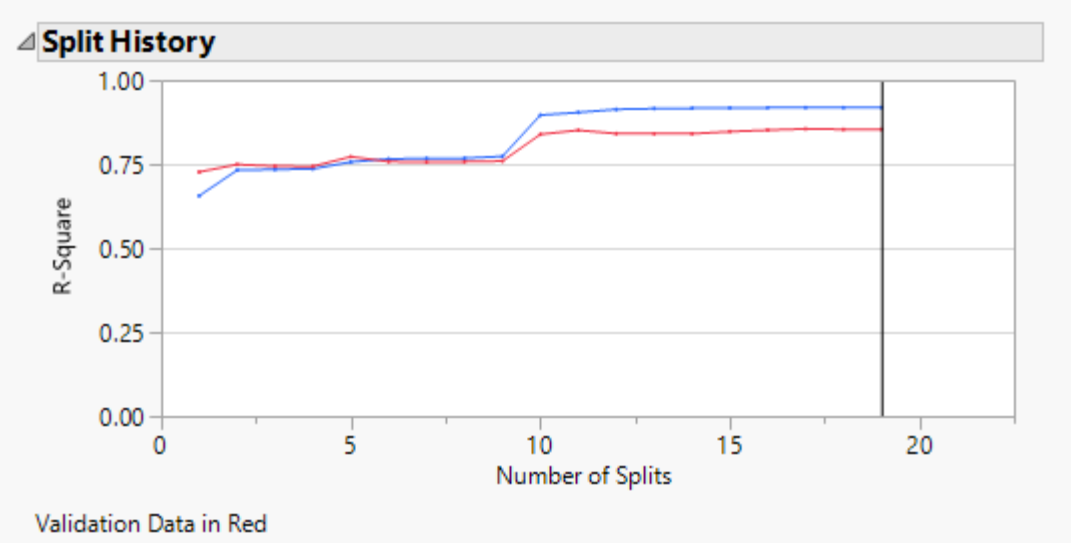


Fig 4.7(b)

The split history shows a plot of R-Square versus the number of splits. The R-Square curve is blue for the training set and red for the validation set. This plot shows that after a certain number of splits, the R-square of training and validation datasets did not vary a lot. This is when we stopped the splitting.
The mean error on validation is estimated to be 37% and is shown in Fig 4.7(c). The error is more when compared to the performance of regression model.

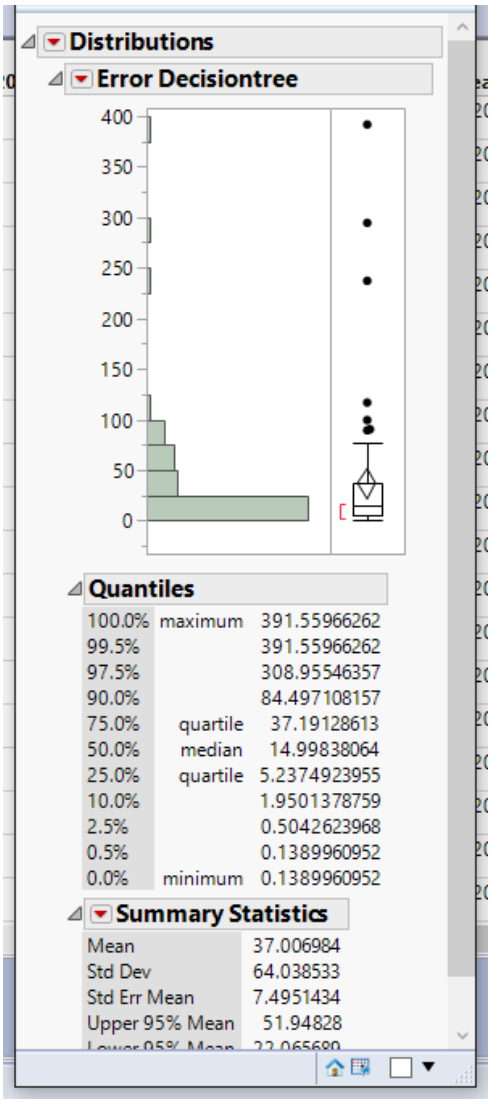
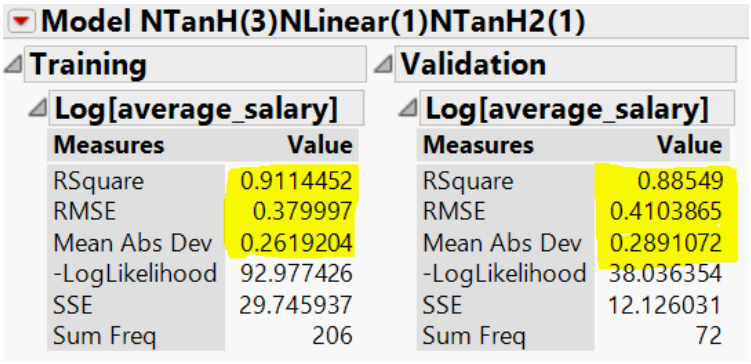


Fig 4.7(c)

Neural network

Neural Network predicts a better fit model as compared with linear regression (R-square 91.1% vs 89%) and almost the same as in Decision tree model (R-square 91.1% vs 91.9%). This can be seen in Fig 4.8(a). However, according to Fig 4.8(b) which shows the distribution of the error on validation dataset, the mean error is slightly higher than linear regression (32.3% vs 31.7%).



Model NTanH(3)NLinear(1)NTanH2(1)

Training		Validation	
Log[average_salary]		Log[average_salary]	
Measures	**Value**	**Measures**	**Value**
RSquare	0.9114452	RSquare	0.88549
RMSE	0.379997	RMSE	0.4103865
Mean Abs Dev	0.2619204	Mean Abs Dev	0.2891072
-LogLikelihood	92.977426	-LogLikelihood	38.036354
SSE	29.745937	SSE	12.126031
Sum Freq	206	Sum Freq	72

Fig. 4.8 (a)

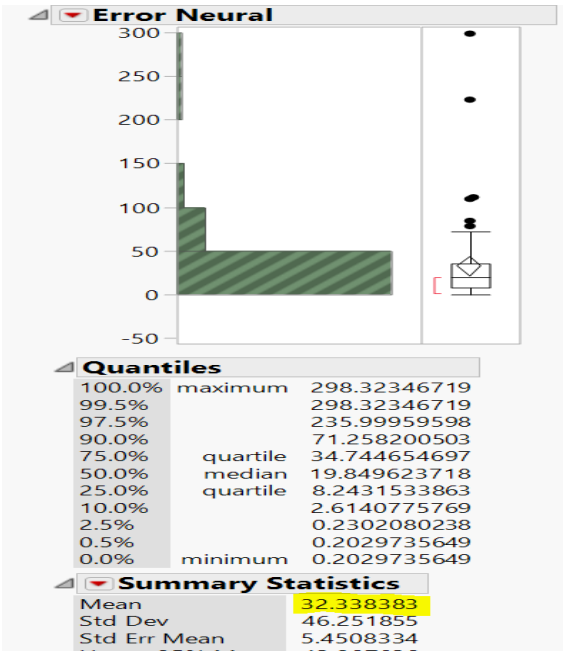


Fig 4.8 (b)

Validation:
The RMSE value is indicative of the error that the model may have during prediction.

Model	RMSE
Regression	0.41 (Fig. 4.1)
Neural	0.36 (Fig. 4.7(a))
Decision	0.39 (Fig. 4.8(a))

But we take the approach of Mean Absolute Error method for error analysis of these 3 models in order to conclude which is our best model.
The models result in a Log of predicted average salary. Comparison of the actual average salary cannot be done against Log values. We therefore find the exponential of Predicted Average Salary.
For Mean Absolut Error values, we use the Formula of **[Abs(Predicted Salary-Average Salary)]/[Predicted Salary]** on all the 3 models.

The comparison of the errors of 3 models with the baseline error is shown below:

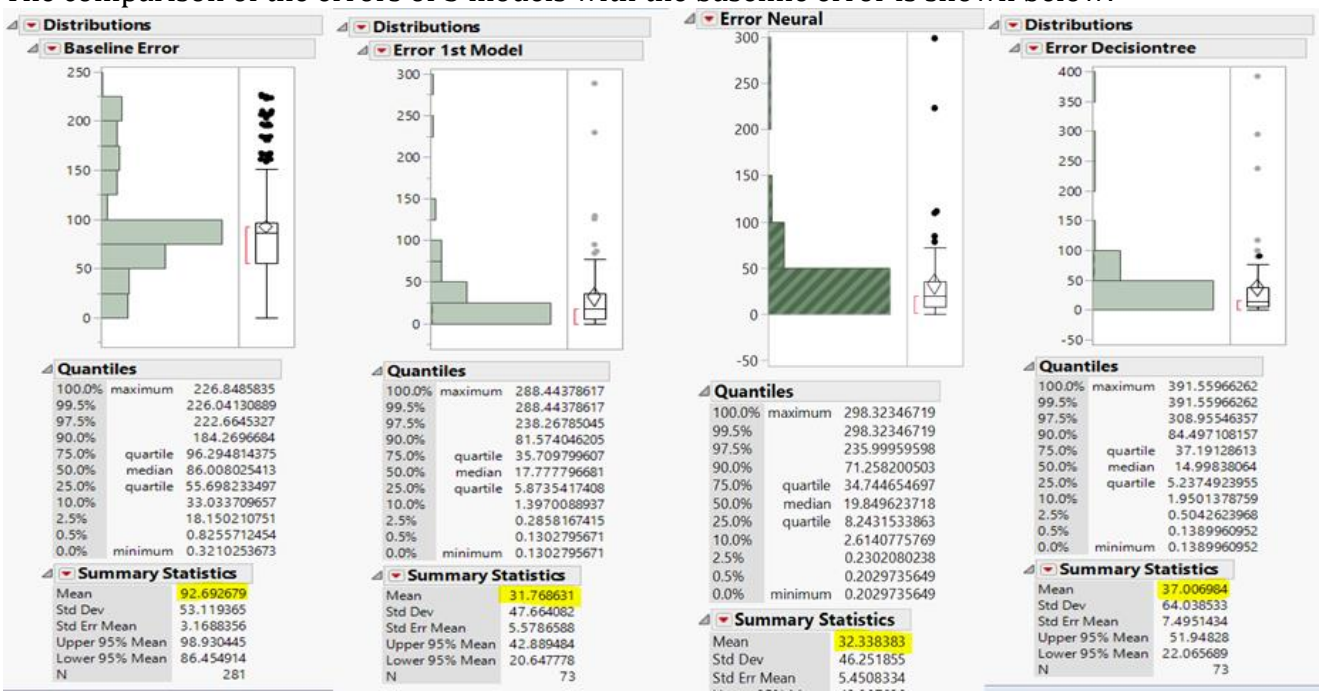


Fig. 4.9

Comparison of models:

Model	R-square	Mean error % on Validation dataset	Baseline error %
Regression	89.8%	31%	92%
Neural	91.1%	32%	
Decision	91.9%	37%	

Conclusion:
Considering the R-square values and mean error performances we decided to choose the Linear regression model to predict the average salaries.

Also, the advantages of choosing the Linear regression model over the others are:

- Predictor Variables are easily interpreted. As per business sense it gave us a clear picture too.
- It has an easy prediction formula
- No hidden issues like – multicollinearity and overfitting
- Since there were not too many missing values and outliers, regression model result can be considered a good fit with clarity.

V. Model Implementation:

In this section we provide some analysis based on our observation of the dataset, as well as the results of all Prediction Models.

The final linear regression formula for our model is as shown below,
Log (avg. salary) = -96.88 + 0.0024 * 1st_PlayAttempts+0.0492*Year+0.8166*Log (prev. average salary) + 0.0027*QB_Rating

The above expression gives us "Predicted log(average salary)".
Here, plus sign indicates positive correlation.
1st Play attempts: The player which passes more in 1st play gets better salary in his next contract.
Year: There is an inflation factor, every year the contract amount tends to increase compared to previous year's contract.
Log (previous average salary): There is strong positive coefficient of previous salary and current salary, which indicates that current salary increases with increase in previous salary.
QB_Rating: This parameter incorporates weighted performance parameters.

QB_Rating = (25 Attempted Pass +1000 completed pass + 50 Yards+ 4000td-5000 intercepts)/12 attempts.
Thus we can say that average salary of a player increases when number of completed passes, yards, touchdown increases, for constant number of attempts.

Insights:
Considering all our model results and analysis, if we were to make some suggestions to the NFL world – provided they find us worth talking to, here's what we have to suggest:

We found that the previous salary which can be considered as a baseline salary of a player, impacts a lot on the future salary of a player. Is this always justified or is it a Bandwagon Effect (Refer appendix for definition) If the management world uses this as an indirect indicator for identifying good players, this can be a systematic bias in the decision making process. The management can take decisions considering only the performance variables, without considering the previous salaries of the players, to actually identify some under-valued players who could perform equally good, when compared to the other high-paid players.

The good news is, we can provide such models which can identify relationships between different player's performance attributes and their salaries. With the help of these models we can identify 'low-valued' and 'high-valued' attributes which can help the management of a franchisee to make trade-offs whether they want the identified "high-valued" performers as team players for a higher-price or if they are satisfied with other "low-valued" performers who may actually be valuable to the team.

VI. Plan for future upgrades

- Incorporate physical characteristics like height and weight of players to available performance data and conduct a comprehensive analysis. We'd invest 55% of our future time of the total time on this project for incorporating the effect of personal characteristics.
- Prepare a model for all the positions following same approach as followed for quarterbacks. We'd invest 35% of our future time on this project for incorporating effects of similar models to make a more comprehensive predictive model. This can help in identifying the most important performance variables affecting the salaries of players and eventually identify 'high-valued' and 'low-valued' attributes. In this way, we can simplify the toughest job of the NFL world!
- Update the model by adding the data of coming years, once it is available. We'd invest only 5% of our time on this.
- Look for other leagues, college level competitions etc. to collect additional performance data. We'd invest only 5% of our time on this.

APPENDIX

Data Dictionary

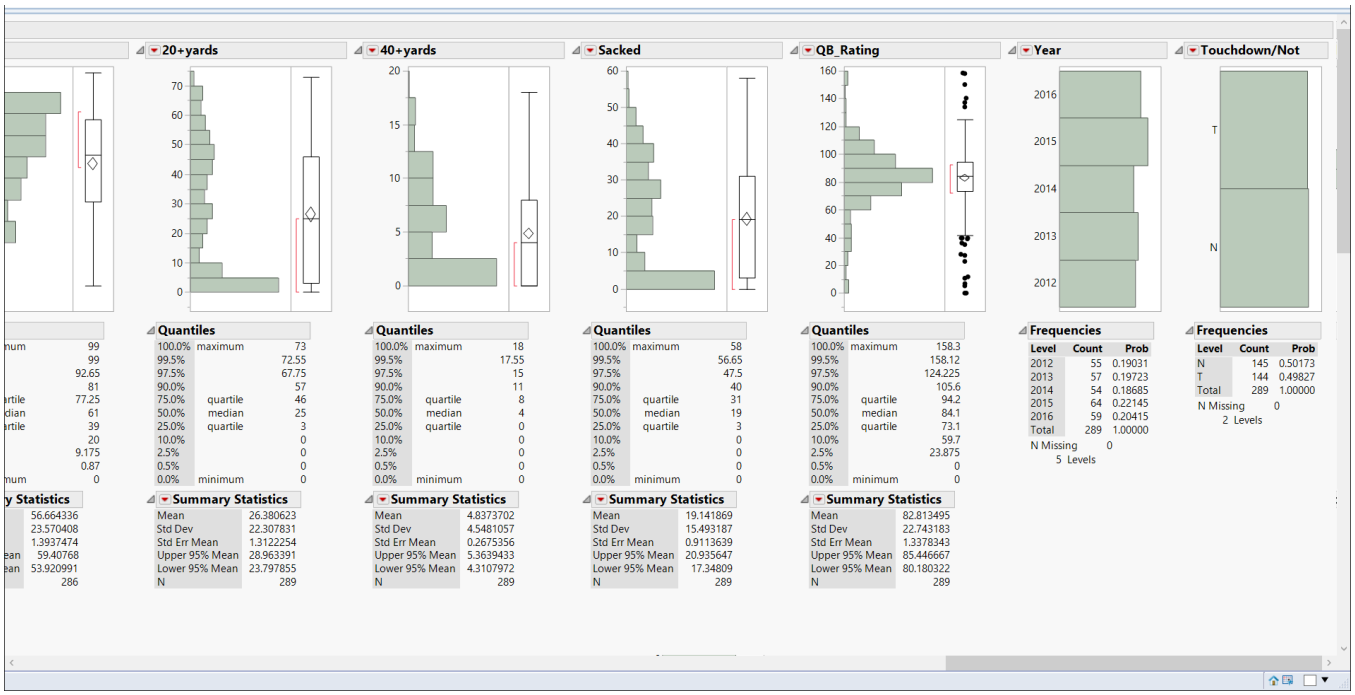
Variable	Description
Player	Player name
Team	The team the player played for
Average salary	Avg salary (per year) a player will earn/earns this year.
Pos	Position of the player
Prev average salary	Avg salary a player earned last year.
Rank	Rank of a player (in all Quaterbacks)
Completed pass	No. of passes completed
Attempted pass	No. of passes attempted
% completed	(Completed pass divided by Attempted pass)*100
Attempt/Game	No of passes attempted/game
Yards	Yards covered in all passes attempted by the quarterback
Yards/attempt	Average yards per attempt
Yards/Game	Yards covered in all passes per game
Touchdown	No. of Touch Downs
Intercepts	No of Intercepts in all passes attempted by the quarterback
1st_Play attempts	No. of attempts in the first play of attack.
1st %	(1st/Att)
Longest	Longest successful pass thrown by a quarterback
20+ yards	No. of successful passes that covered more than 20 yards
40+ yards	No of successful passes that covered more than 40+ yards
Sacked	No of times a quarterback was Sacked by the defense before he could attempt a pass.
QB Rating	Rating of any quarterback calculated by some performance measure technique.
Year	Mentions the year in which a player earned the 'average salary'
Touchdown/Not	It tells whether the longest pass attempted by a quarterback earned touchdown or Not

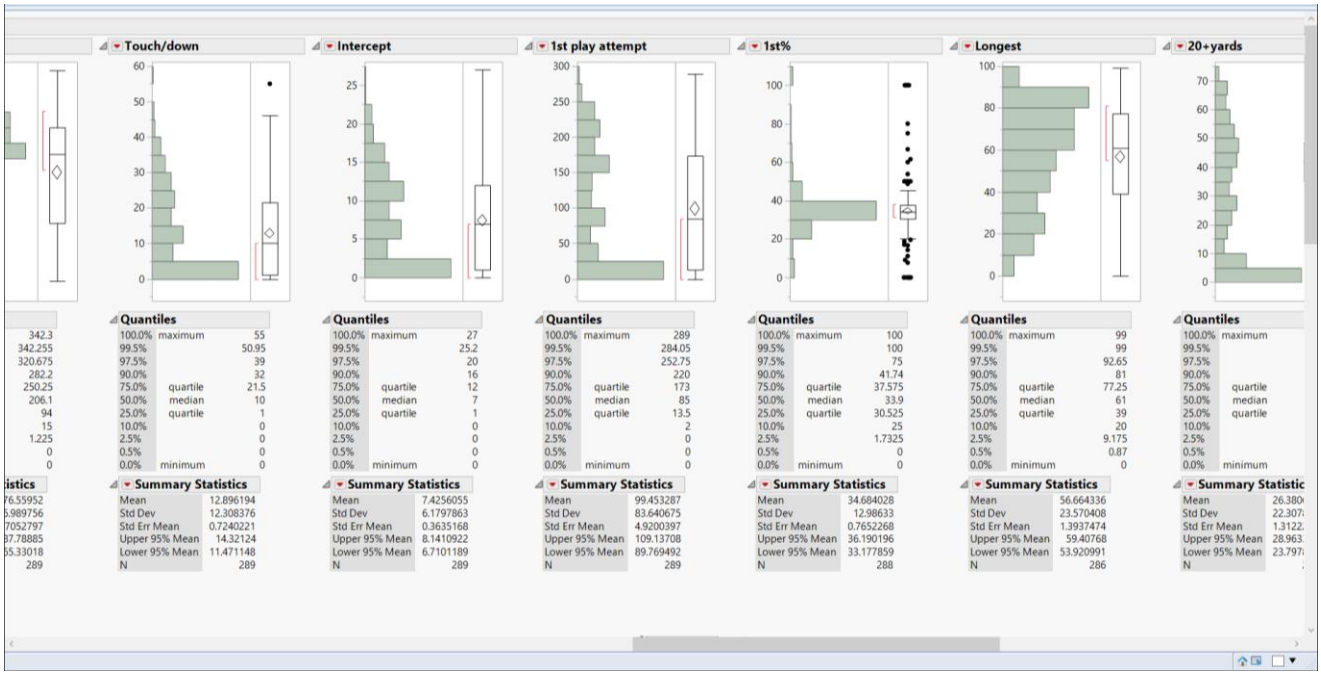
Player positions

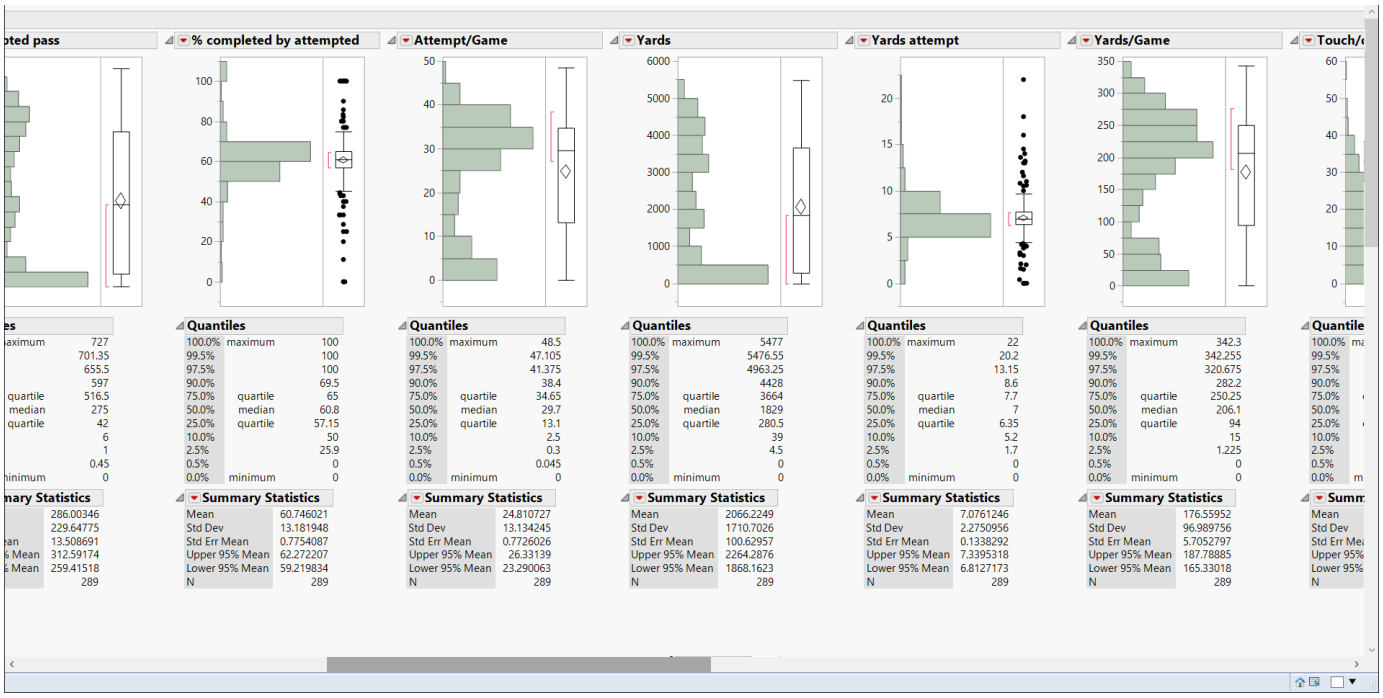
Offense	
Quarterback	The leader of the team. He calls the plays in the huddle, yells the signals at the line of scrimmage, and receives the ball from the center. Then he hands off the ball to a running back, throws it to a receiver, or runs with it
Center	The player who snaps the ball to the quarterback. He handles the ball on every play.
Running back	A player who runs with the football. Running backs are also referred to as tailbacks, halfbacks, and rushers.
Fullback	A player who's responsible for blocking for the running back and also for pass-blocking to protect the quarterback. Fullbacks, who are generally bigger than running backs, are short-yardage runners.
Wide receiver	A player who uses his speed and quickness to elude defenders and catch the football. Teams use as many as two to four wide receivers on every play
Tight End	A player who serves as a receiver and also as a blocker. This player lines up beside the offensive tackle to the right or the left of the quarterback.
Left guard and right guard	The inner two members of the offensive line, whose jobs are to block for and protect the quarterback and ball carriers.
Left tackle and right tackle	The outer two members of the offensive line.

Defense	
Defensive tackle	The inner two members of the defensive line, whose jobs are to maintain their positions in order to stop a running play or run through a gap in the offensive line to pressure the quarterback or disrupt the backfield formation.
Defensive end	The outer two members of the defensive line. Generally, their jobs are to overcome offensive blocking and meet in the backfield, where they combine to tackle the quarterback or ball carrier. On running plays to the outside, they're responsible for forcing the ball carrier either out of bounds or toward (into) the pursuit of their defensive teammates
Linebacker	These players line up behind the defensive linemen and generally are regarded as the team's best tacklers. Depending on the formation, most teams employ either three or four linebackers on every play. Linebackers often have the dual role of defending the run and the pass.
Safety	The players who line up the deepest in the secondary — the last line of defense. There are free safeties and strong safeties, and they must defend the deep pass and the run
Cornerback	The players who line up on the wide parts of the field, generally opposite the offensive receivers

Distributions of variables

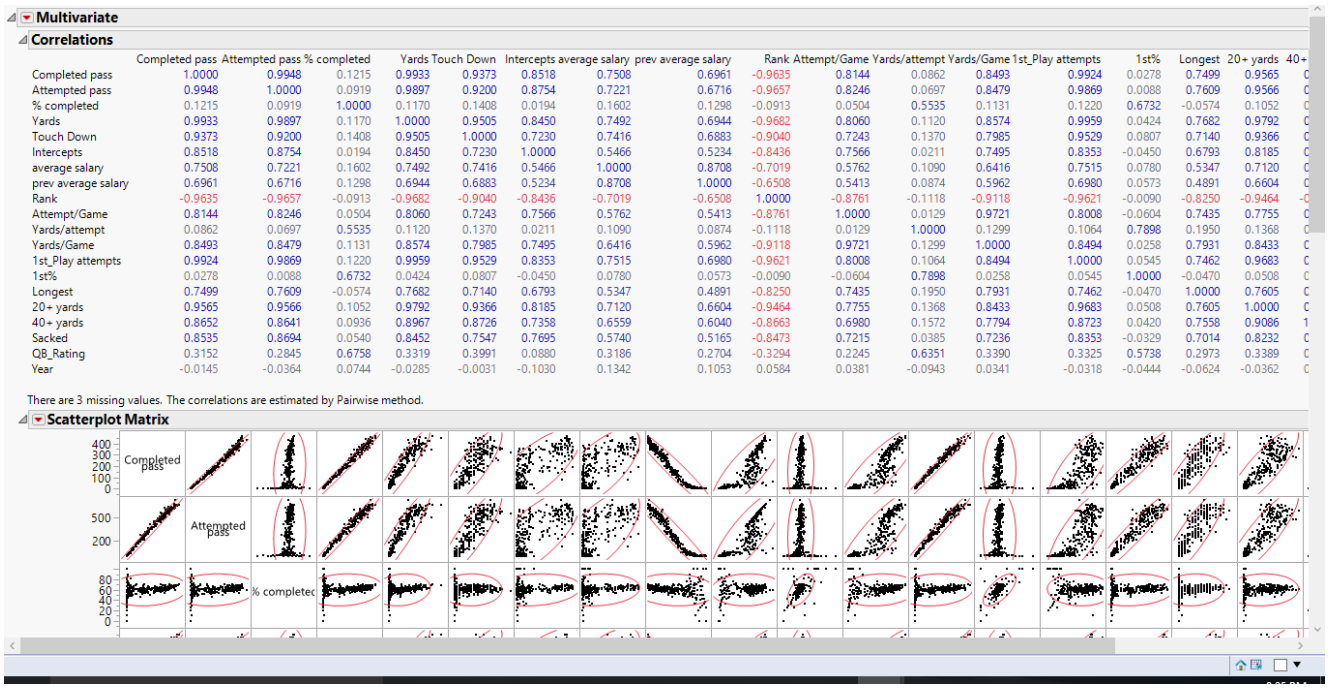





% completed by attempted
Attempt/Game
Yards
Yards attempt
Yards/Game
Quantiles
Summary Statistics


average salary
prev average salary
Rank
Attempt/Game
Yards/attempt
Quantiles
Summary Statistics

Correlations



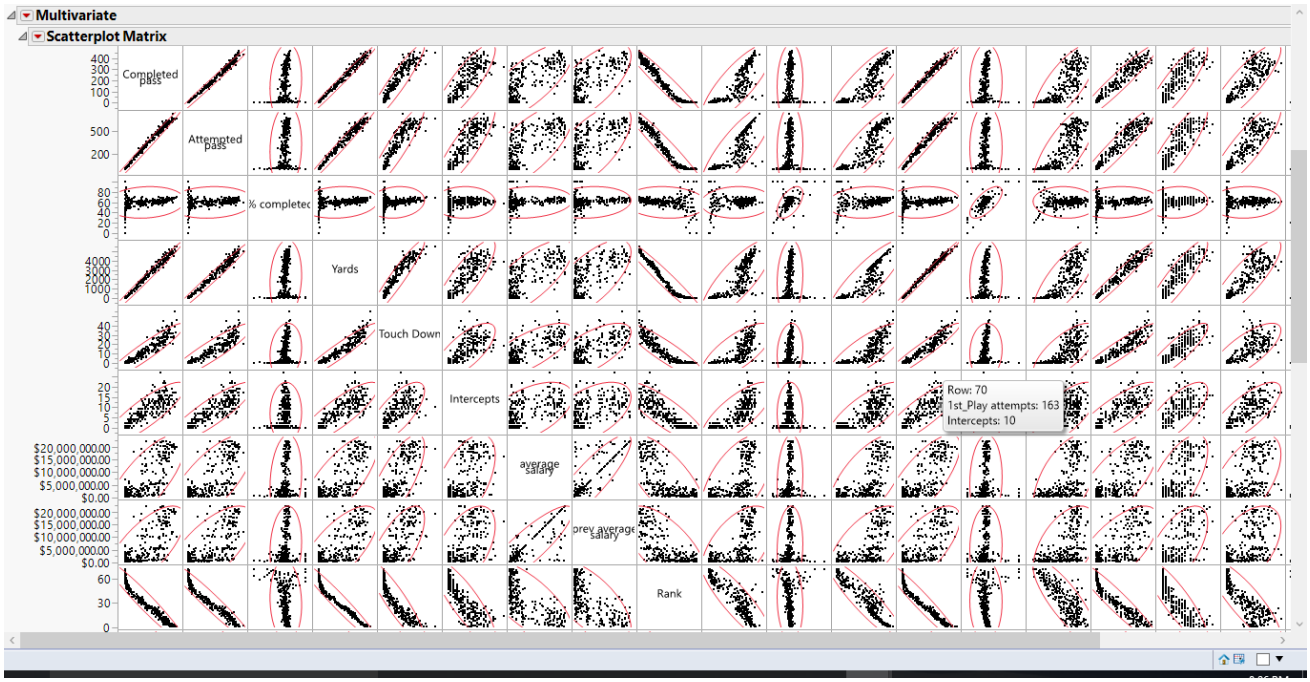
Multivariate

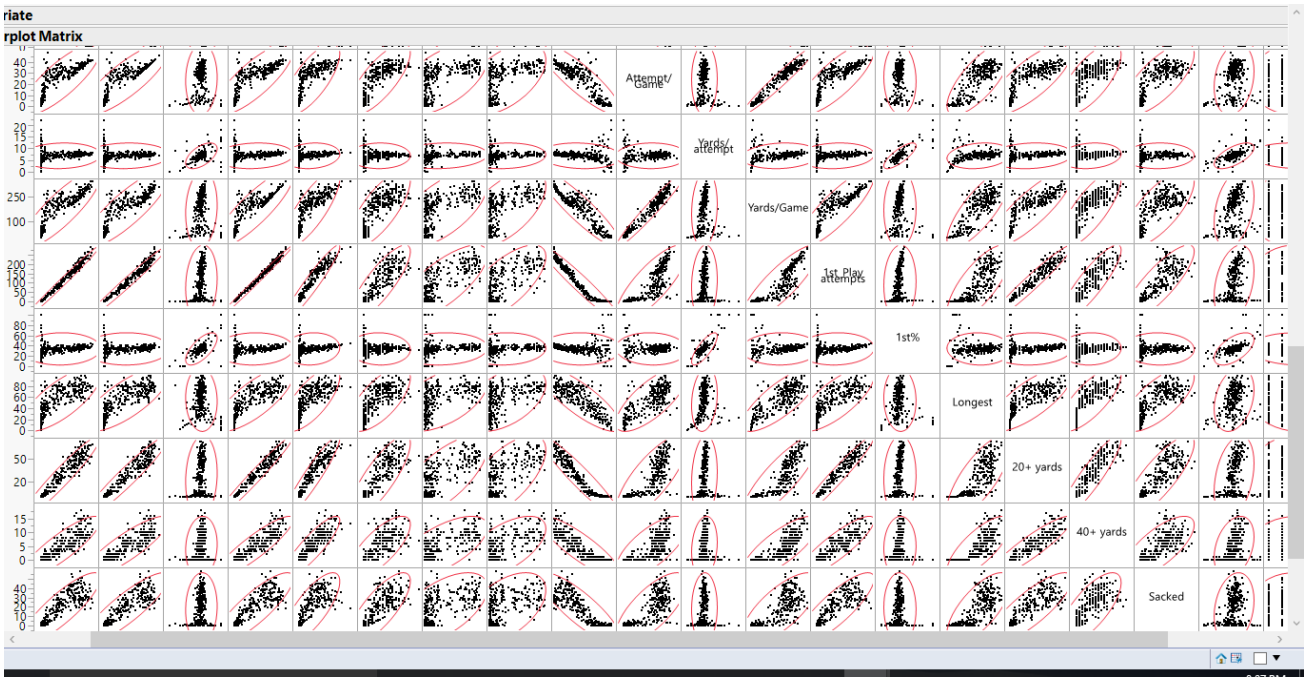
Correlations

	Completed pass	Attempted pass	% completed	Yards	Touch Down	Intercepts	average salary	prev average salary	Rank	Attempt/Game	Yards/attempt	Yards/Game	1st_Play attempts	1st%	Longest	20+ yards	40+
Completed pass	1.0000	0.9948	0.1215	0.9933	0.9373	0.8518	0.7508	0.6961	-0.9635	0.8144	0.0862	0.8493	0.9924	0.0278	0.7499	0.9565	
Attempted pass	0.9948	1.0000	0.0919	0.9897	0.9200	0.8754	0.7221	0.6716	-0.9657	0.8246	0.0697	0.8479	0.9869	0.0088	0.7609	0.9566	
% completed	0.1215	0.0919	1.0000	0.1170	0.1408	0.0194	0.1602	0.1298	-0.0913	0.0504	0.5535	0.1131	0.1220	0.6732	-0.0574	0.1052	
Yards	0.9933	0.9897	0.1170	1.0000	0.9505	0.8450	0.7492	0.6944	-0.9682	0.8060	0.1120	0.8574	0.9959	0.0424	0.7682	0.9792	
Touch Down	0.9373	0.9200	0.1408	0.9505	1.0000	0.7230	0.7416	0.6883	-0.9040	0.7243	0.1370	0.7985	0.9529	0.0807	0.7140	0.9366	
Intercepts	0.8518	0.8754	0.0194	0.8450	0.7230	1.0000	0.5466	0.5234	-0.8436	0.7566	0.0211	0.7495	0.8353	-0.0450	0.6793	0.8185	
average salary	0.7508	0.7221	0.1602	0.7492	0.7416	0.5466	1.0000	0.8708	-0.7019	0.5762	0.1090	0.6416	0.7515	0.0780	0.5347	0.7120	
prev average salary	0.6961	0.6716	0.1298	0.6944	0.6883	0.5234	0.8708	1.0000	-0.6508	0.5413	0.0874	0.5962	0.6980	0.0573	0.4891	0.6604	
Rank	-0.9635	-0.9657	-0.0913	-0.9682	-0.9040	-0.8436	-0.7019	-0.6508	1.0000	-0.8761	-0.1118	-0.9118	-0.9621	-0.0090	-0.8250	-0.9464	
Attempt/Game	0.8144	0.8246	0.0504	0.8060	0.7243	0.7566	0.5762	0.5413	-0.8761	1.0000	0.0129	0.9721	0.8008	-0.0604	0.7435	0.7755	
Yards/attempt	0.0862	0.0697	0.5535	0.1120	0.1370	0.0211	0.1090	0.0874	-0.1118	0.0129	1.0000	0.1299	0.1064	0.7898	0.1950	0.1368	
Yards/Game	0.8493	0.8479	0.1131	0.8574	0.7985	0.7495	0.6416	0.5962	-0.9118	0.9721	0.1299	1.0000	0.8494	0.0258	0.7931	0.8433	
1st_Play attempts	0.9924	0.9869	0.1220	0.9959	0.9529	0.8353	0.7515	0.6980	-0.9621	0.8008	0.1064	0.8494	1.0000	0.0545	0.7462	0.9683	
1st%	0.0278	0.0088	0.6732	0.0424	0.0807	-0.0450	0.0780	0.0573	-0.0090	-0.0604	0.7898	0.0258	0.0545	1.0000	-0.0470	0.0508	
Longest	0.7499	0.7609	-0.0574	0.7682	0.7140	0.6793	0.5347	0.4891	-0.8250	0.7435	0.1950	0.7931	0.7462	-0.0470	1.0000	0.7605	
20+ yards	0.9565	0.9566	0.1052	0.9792	0.9366	0.8185	0.7120	0.6604	-0.9464	0.7755	0.1368	0.8433	0.9683	0.0508	0.7605	1.0000	
40+ yards	0.8652	0.8641	0.0936	0.8967	0.8726	0.7358	0.6559	0.6040	-0.8663	0.6980	0.1572	0.7794	0.8723	0.0420	0.7558	0.9086	1
Sacked	0.8535	0.8694	0.0540	0.8452	0.7547	0.7695	0.5740	0.5165	-0.8473	0.7215	0.0385	0.7236	0.8353	-0.0329	0.7014	0.8232	
QB_Rating	0.3152	0.2845	0.6758	0.3319	0.3991	0.0880	0.3186	0.2704	-0.3294	0.2245	0.6351	0.3390	0.3325	0.5738	0.2973	0.3389	
Year	-0.0145	-0.0364	0.0744	-0.0285	-0.0031	-0.1030	0.1342	0.1053	0.0584	0.0381	-0.0943	0.0341	-0.0318	-0.0444	-0.0624	-0.0362	

There are 3 missing values. The correlations are estimated by Pairwise method.

Scatterplot Matrix





Definitions

F Ratio: Shows the model mean square divided by the error mean square. The F Ratio is the test statistic for a test of whether the model differs significantly from a model where all predicted values are the response mean.

Offensive player: A player who plays in the offense side of the game. The offense is the side which is in possession of the ball. It is their job to advance the ball towards the opponent's end zone to score points.

Defensive player: A player who plays in the defense side of the game. The objective of the defensive team is to prevent the other team from scoring.

Feature Engineering: Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.

Bandwagon effect: The tendency to do (or believe) things because many other people do (or believe) the same. (Herd Mentality)