**Preliminary Analysis:**

Before implementing the regression analysis on the dataset, we have performed the below initial data checks:

1. Presence of Missing values:

Command in R:

*sum(is.na(Final\_Dataset))*

*0*

where , Final\_Dataset is the name of our dataset file

Since the result is 0, it implies that there are no NULL values in our dataset.

1. Skewness of Data:

Using the skewness() in R, we find that the factors Score, Rushing Yards, Penalties, Touchdowns, Field Goals are skewed to the right i.e. are positively skewed while Total Offensive Plays, Passing Yards are negatively skewed.

Below is the table depicting the result values:

|  |  |
| --- | --- |
| **Field Name** | **Skewness Value** |
| Score | 0.5410502 |
| Total Offensive Plays | -0.2503103 |
| Rushing Yards | 0.3427709 |
| Passing Yards | -0.0479893 |
| Penalties | 0.027623 |
| Touchdowns | 0.7522769 |
| Field Goals | 0.4083343 |

**Table 3.1 Skewness values of the Dataset**

1. Check for Outliers:

Dataset was checked for outliers by plotting histograms against their frequency of occurrences. However no outliers were found.

1. Statistical Summary of dataset:

Below table represents the statistical mean, median and standard deviation for each of the numerical factors in our dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Factor Name** | **Mean** | **Standard Deviation** | **Median** |
| Year | 1991 | 14.29 | 1991 |
| Score | 30.27 | 9.78 | 30 |
| Total Offensive Plays | 66.78 | 7.37 | 68 |
| Rushing Yards | 168.9 | 48.7 | 167 |
| Passing Yards | 232.8 | 85.71 | 241 |
| Penalties | 3.12 | 2.59 | 2 |
| Touchdowns | 2.82 | 1.59 | 3 |
| Field Goals | 1.55 | 1.04 | 1 |

**Table 3.1 Summary of Dataset**

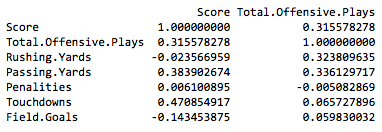
**Statistical Analysis:**

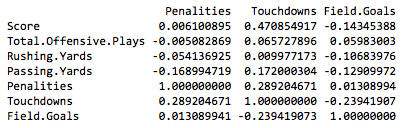
The dataset is analysed using Multiple Regression Analysis method with the Score field as the dependent variable and below as the independent variables:

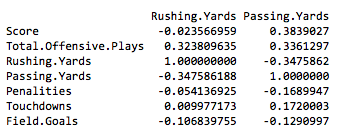
* Total Offensive Plays
* Total Rushing Yards
* Total Passing Yards
* Penalties
* Touchdown
* Field Goals

To remove any redundant or highly correlated factors, we create a correlation matrix. The coefficients from the matrix help in understanding the degree of linear relationship existing between these factors. If these coefficient values are lesser than -0.7, the factors are considered as redundant and highly correlated.

Below results were obtained from calculating using the Pearson’s correlation cofficient on dataset:







Since we do not find any highly correlated values, we go ahead with our regression analysis using all the above factors.

To build a linear model for our analysis, we use the lm() in R with the Score as dependent variable against the rest of factors as independent variable.

The alpha value which we have considered for our analysis is 0.05. Hence factors having p values lesser than the alpha value are the most significant fields in determining the dependent variable which is the Score field as per our analysis.

Upon building the model, we get the below p values for each of the independent variable against the dependent variable.

|  |  |
| --- | --- |
| **Field Names** | **P values** |
| Total Offensive Plays | 0.0102 |
| Rushing Yards | 0.5034 |
| Passing Yards | 0.3091 |
| Penalties | 0.0359 |
| Touchdowns | 0.0214 |
| Field Goals | 0.0484 |

**Table 4.1 p values of Dataset**

Evaluating the p values from our result, we find that the below fields have values less than our alpha value:

* Total Offensive Plays
* Penalties
* Touchdown
* Field Goals

Hence, from our analysis we can say that the above factors have a strong relationship in determining the final Score field in a NFL Superbowl match.

**Conclusion:**

We began our analysis with 6 factors evaluating whether each of these factors are responsible in affecting the final Score of the NFL superbowl match. Upon regression analysis, from the p value analysis only the below four factors had very low values :

* Total Offensive Plays
* Penalties
* Touchdown
* Field Goals

Thus indicating that among the total of 6 factors, the above four factors have a very strong relationship with our dependent variable which is the Score field and hence we can conclude that these factors have a strong relationship in determining and predicting the Score field.