**PREDICTIVE MODELLING BUSINESS REPORT**

**BY**

**SUDEEP KUMAR DAS**

**PGPDSBA.O. SEP22.B**

**TABLE OF CONTENT**

**LIST OF FIGURES ------------------------------------------------------------------------ 03**

**LIST OF TABLES -------------------------------------------------------------------------- 05**

**Problem 1**

* 1. **Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis ---------------------------------06**
  2. **Impute null values if present? Do you think scaling is necessary in this case? ----------------------------------------------------------------------17**
  3. **Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (30:70). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE ------------------------------18**
  4. **Inference: Based on these predictions, what are the business insights and recommendations. ----------------------------------------------20**

**Problem 2:**

**2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. -------21**

**2.2) Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis) ---------------------------------------32**

**2.3) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Compare both the models and write inferences, which model is best/optimized. --------------------33**

**2.4) Inference: Based on these predictions, what are the insights and recommendations. -----------------------------------------------------34**

**LIST OF FIGURES**

**Figure 1: Distribution of variable sp500 --------------------------------08**

**Figure 2: Distribution of Sales Variable ---------------------------------08**

**Figure 3: Distribution of Capital Variable ------------------------------ 09**

**Figure 4: Distribution of Patents Variable ----------------------------- 09**

**Figure 5: Distribution of R & D variable -------------------------------- 10**

**Figure 6: Distribution of Employment Variable ---------------------- 10**

**Figure 7: Distribution of Tobinq Variable -------------------------------11**

**Figure 8: Distribution of Value Variable ---------------------------------11**

**Figure 9: Distribution of Institution Variable --------------------------12**

**Figure 10: Scatter plot of Capital and Sales ----------------------------12**

**Figure 11: Scatterplot of Patents and Sales ----------------------------13**

**Figure 12: Scatterplot of R&D and Sales --------------------------------13**

**Figure 13: Scatterplot of Employment and Sales ---------------------14**

**Figure 14: Scatterplot of Tobinq and Sales -----------------------------14**

**Figure 15: Scatterplot of Value and Sales -------------------------------15**

**Figure 16: Scatterplot of Institutions and Sales -----------------------15**

**Figure 17: Multivariate Analysis of the Dataset ----------------------16**

**Figure 18: Top 5 variables which influence the sales ----------------20**

**Figure 19: Distribution of Weight Variable ---------------------------- 24**

**Figure 20: Distribution of Frontal Variable -----------------------------25**

**Figure 21: Distribution of Age of Occupant --------------------------- 25**

**Figure 22: Distribution of Year of accident ---------------------------- 26**

**Figure 23: Distribution of Year of Vehicle ------------------------------ 26**

**Figure 24: Distribution of Airbag Deploy ------------------------------- 27**

**Figure 25: Distribution of Severity of Injury --------------------------- 27**

**Figure 26: Distribution of dvcat -------------------------------------------28**

**Figure 27: Distribution of Survived Variable ---------------------------28**

**Figure 28: Distribution of Airbag Variable ------------------------------29**

**Figure 29: Distribution of Seatbelt ---------------------------------------29**

**Figure 30: Distribution of abcat -------------------------------------------30**

**Figure 31: Distribution of occRole ----------------------------------------30**

**Figure 32: Multivariate Analysis of the dataset -----------------------31**

**LIST OF TABLES**

**Table 1: First and last few rows of the dataset -------------------------06**

**Table 2: Info of the dataset --------------------------------------------------07**

**Table 3: Missing values in the dataset ------------------------------------07**

**Table 4: Description of the dataset ----------------------------------------07**

**Table 5: Correlation Matrix of the dataset ------------------------------16**

**Table 6: Missing Values in the Dataset ----------------------------------- 17**

**Table 7: Scaled dataset --------------------------------------------------------18**

**Table 8: One Hot encoding on sp500 ------------------------------------- 18**

**Table 9: First few rows of the Train & Test Dataset --------------------19**

**Table 10: First and Last few rows of the dataset ------------------------21**

**Table 11: Info of the dataset -------------------------------------------------22**

**Table 12: Description of the dataset ---------------------------------------23**

**Table 13: Missing Value Check -----------------------------------------------23**

**Table 14: Correlation Matrix of the Dataset -----------------------------31**

**Table 15: First few rows after encoding -----------------------------------32**

**Table 16: First and last few rows of train dataset ---------------------- 33**

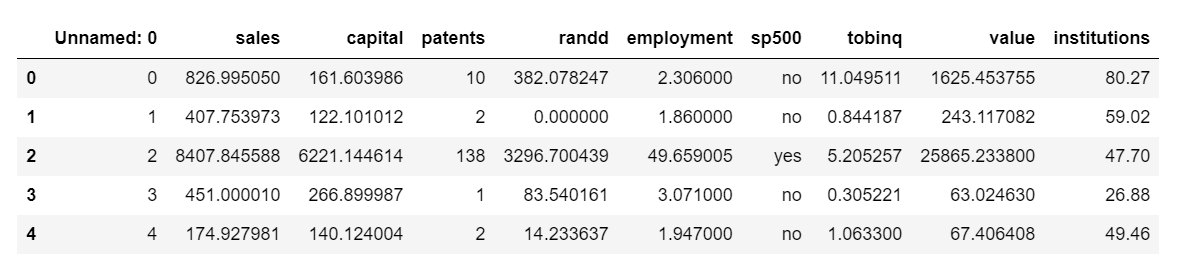
**Problem 1:**

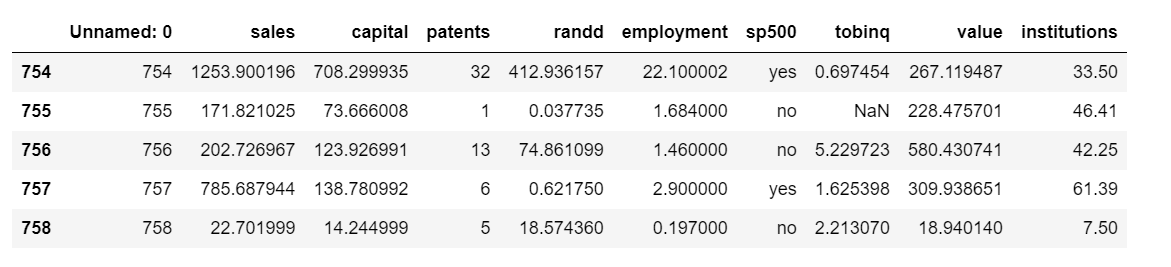
You are a part of an investment firm and your work is to do research about these 759 firms. You are provided with the dataset containing the sales and other attributes of these 759 firms. Predict the sales of these firms on the bases of the details given in the dataset so as to help your company in investing consciously. Also, provide them with 5 attributes that are most important.

* 1. **Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis.**

**Solution:**

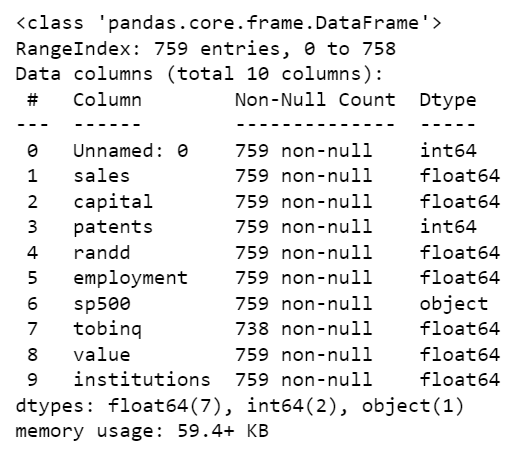
After importing the necessary libraries, we will read the data and perform the exploratory data analysis. The following table shows the first and the last few rows of the dataset:





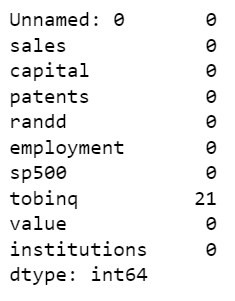
***Table 1: First and last few rows of the dataset***

The basic info of the dataset is as follows:



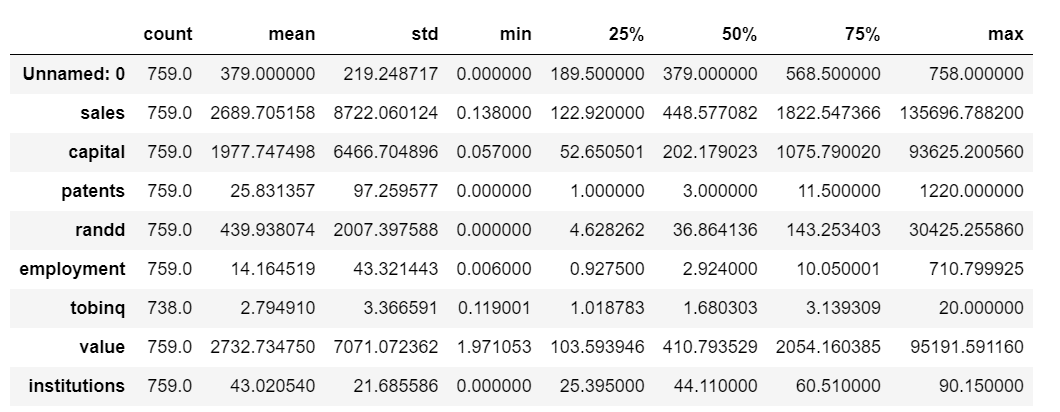
***Table 2: Info of the dataset***

Missing value check reveals the following result:



***Table 3: Missing values in the dataset***

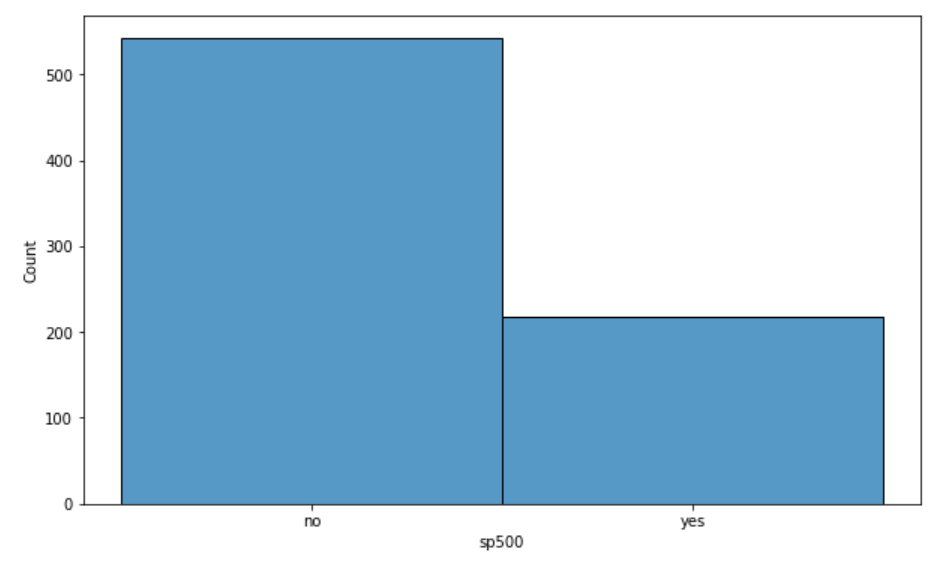
The description of the dataset is as follows:



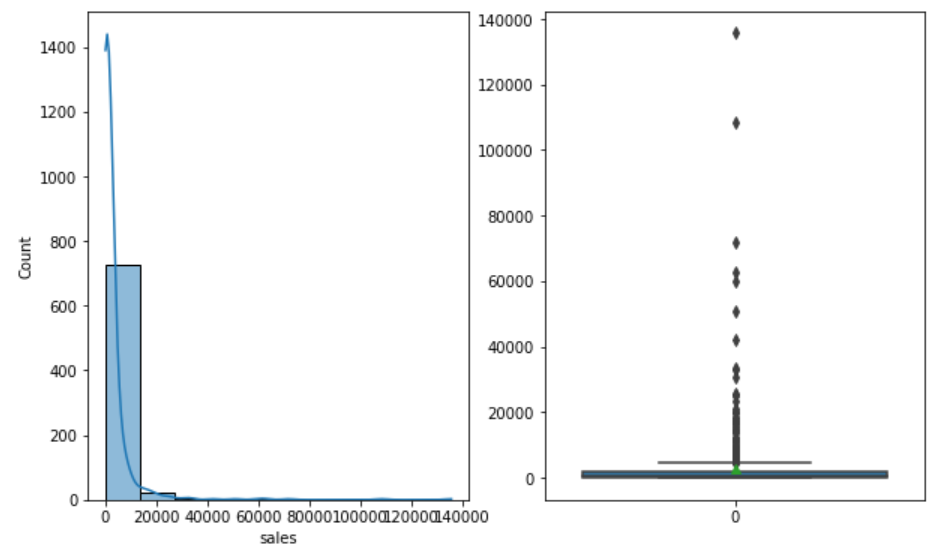
***Table 4: Description of the dataset***

**Exploratory Data Analysis:**

Univariate Data Analysis:

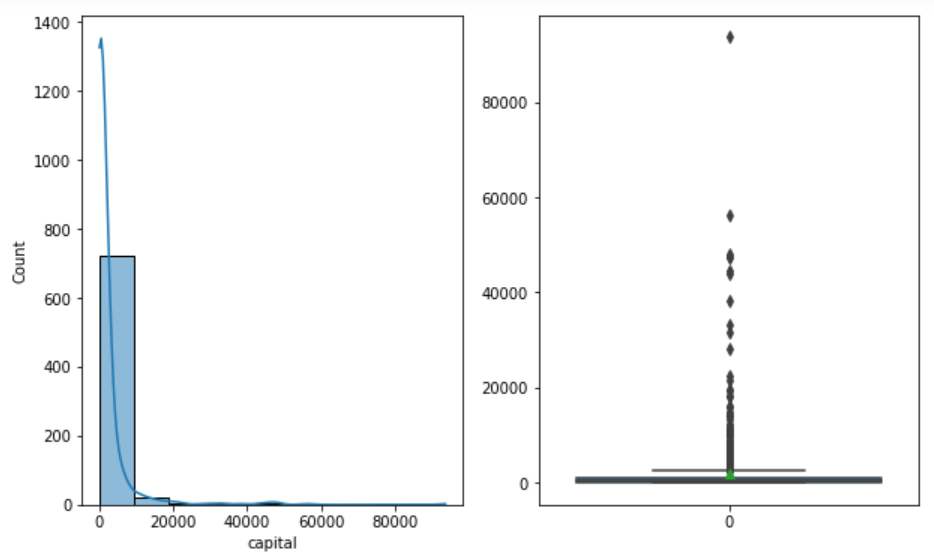


***Figure 1: Distribution of variable sp500***



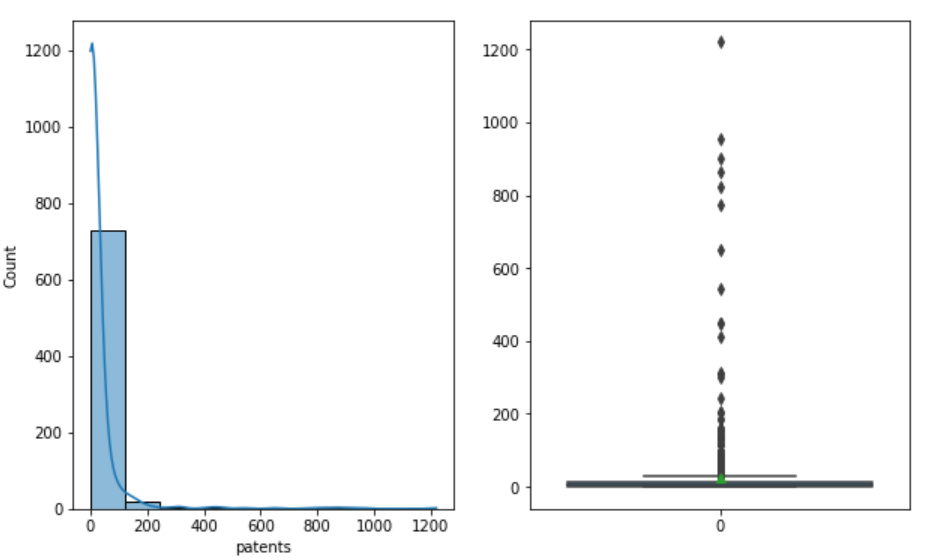
***Figure 2: Distribution of Sales Variable***

The above figure reveals that the data is not normally distributed. There are outliers present in the dataset and the mean sale amount is 2690.



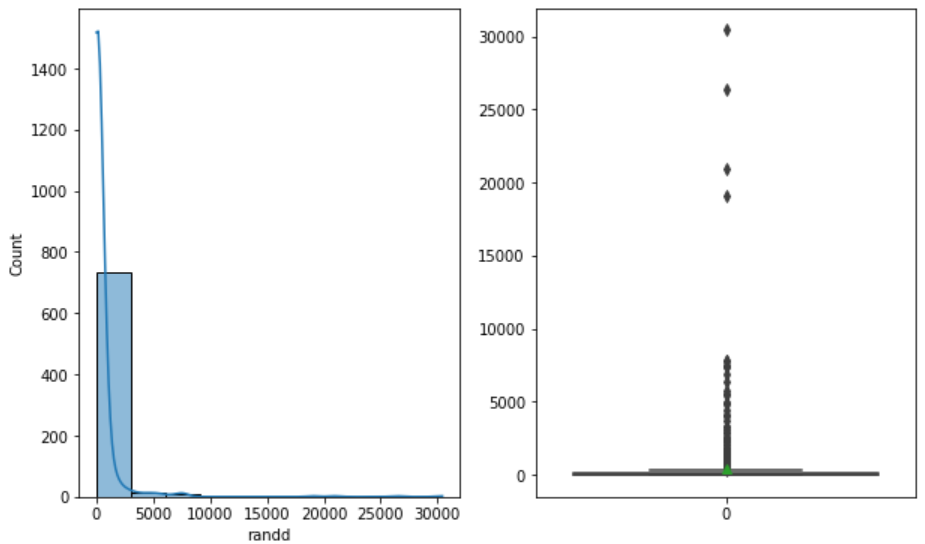
***Figure 3: Distribution of Capital Variable***

The above figure reveals that the data is not normally distributed. There are outliers present in the dataset and the value is 1978.



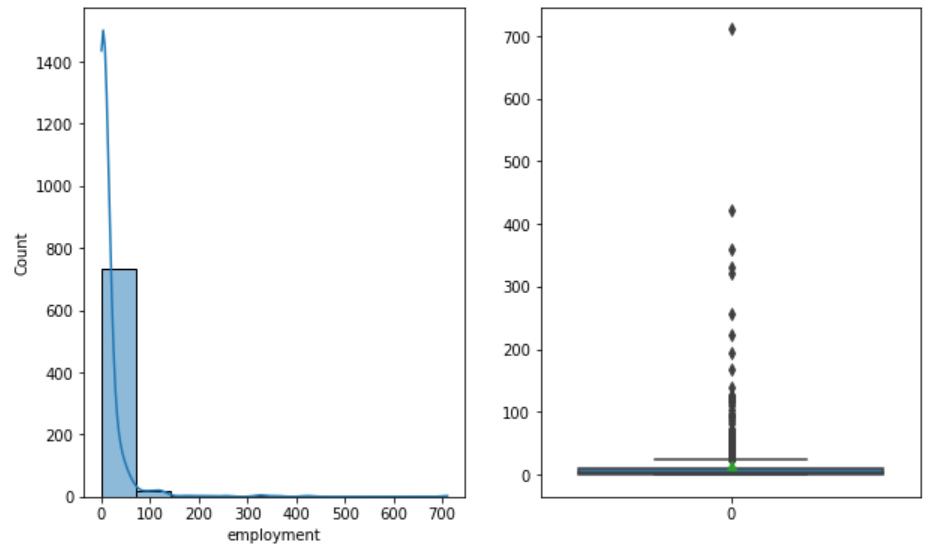
***Figure 4: Distribution of Patents Variable***

The above figure reveals that the data is not normally distributed. There are outliers present in the dataset and the mean amount is 25.8.



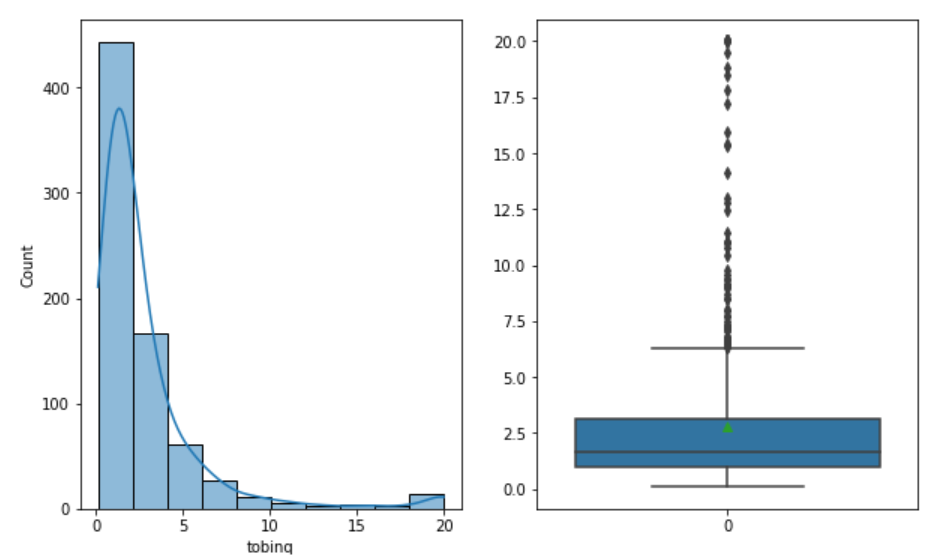
***Figure 5: Distribution of R & D variable***

The above figure reveals that the data is not normally distributed. There are outliers present in the dataset and the mean amount is 440.



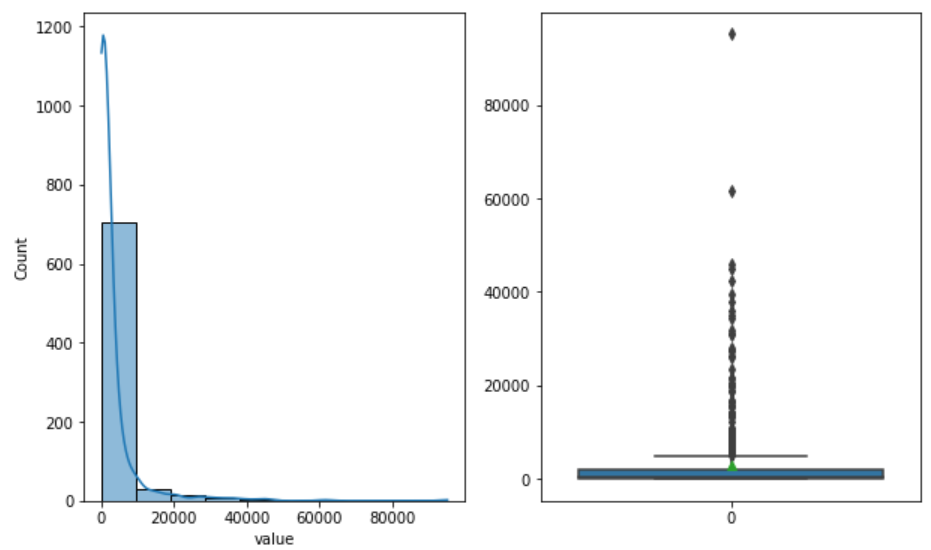
***Figure 6: Distribution of Employment Variable***

The above figure reveals that the data is not normally distributed. There are outliers present in the dataset and the mean amount is 14.16.



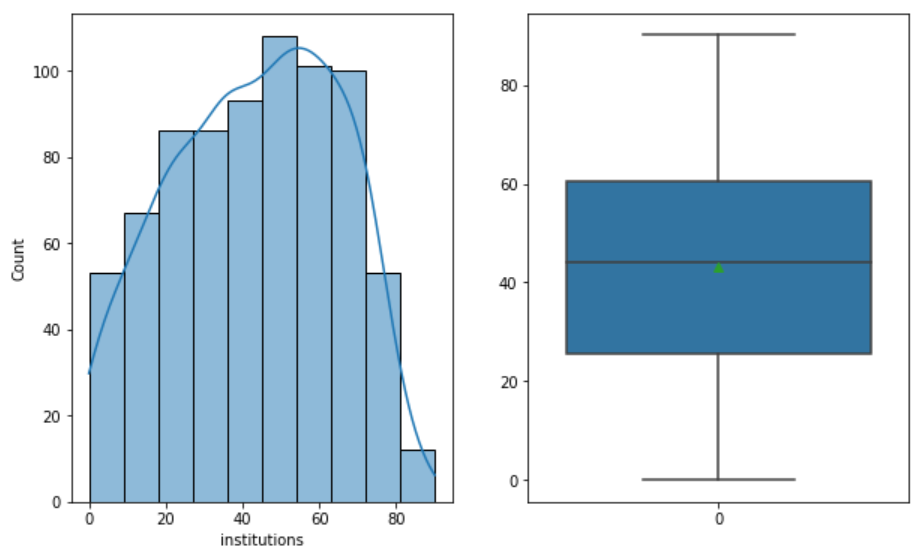
***Figure 7: Distribution of Tobinq Variable***

The above figure reveals that the data is not normally distributed. There are outliers present in the dataset and the mean amount is 2.79.



***Figure 8: Distribution of Value Variable***

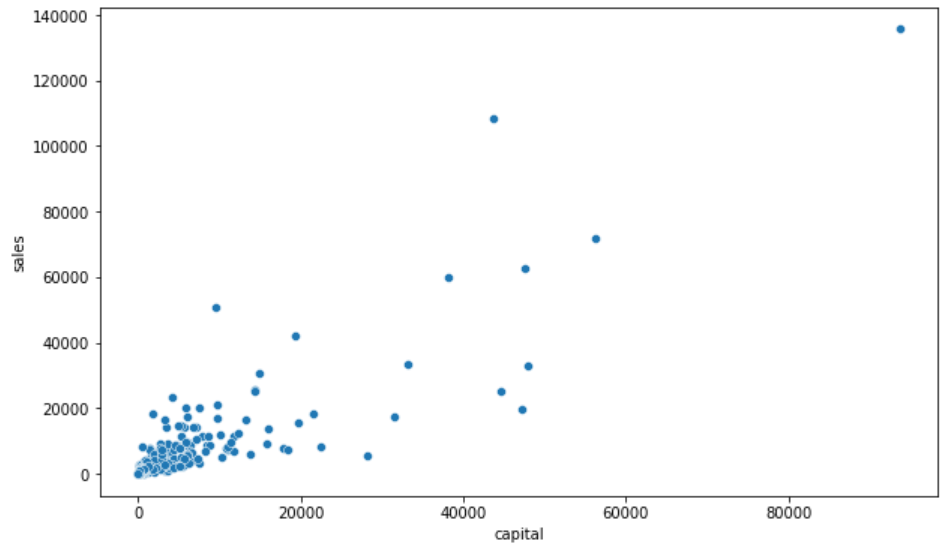
The above figure reveals that the data is not normally distributed. There are outliers present in the dataset and the mean amount is 2733.



***Figure 9: Distribution of Institution Variable***

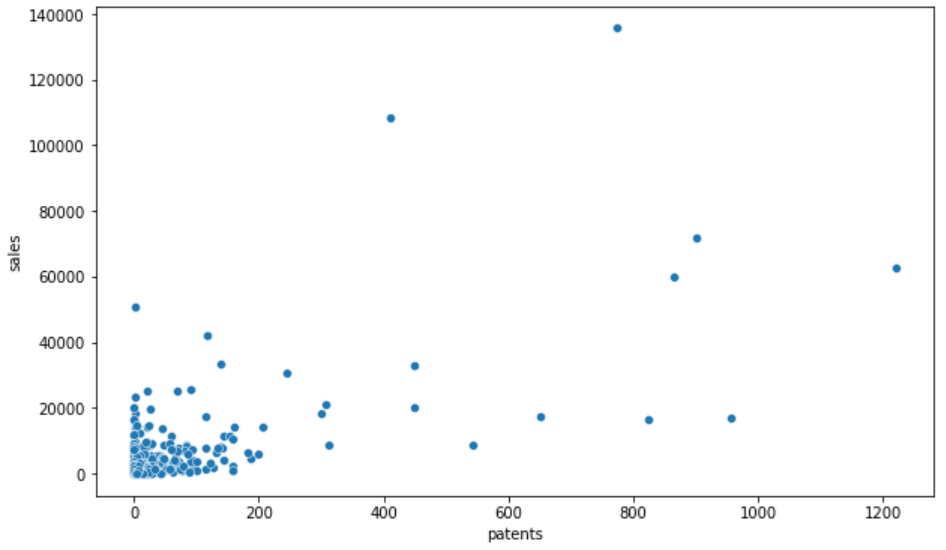
The figure shows that the data is normally distributed. There are no outliers present in the dataset and the mean value is 43.

Bivariate Analysis:



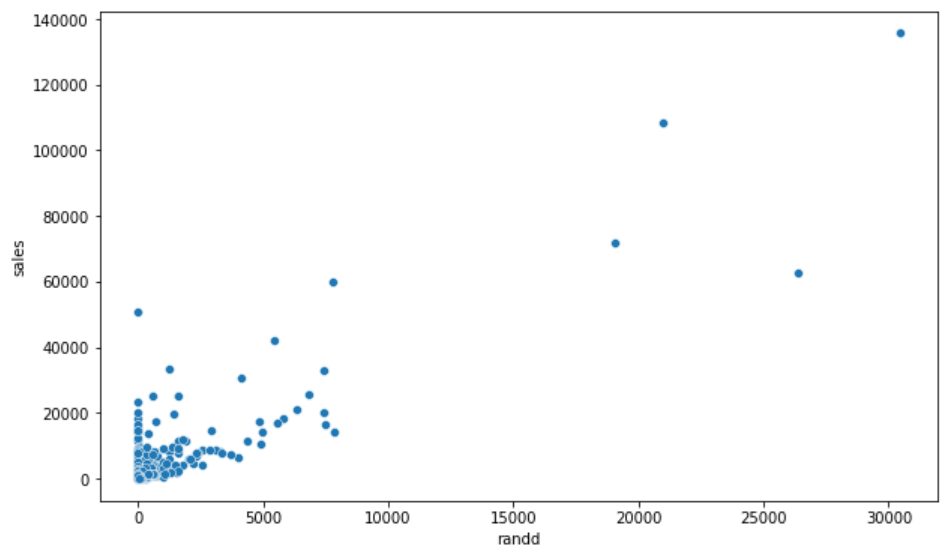
***Figure 10: Scatter plot of Capital and Sales***

The above figure depicts the relationship between the variables capital and sales. We can see that there is no definitive relationship between the variables.



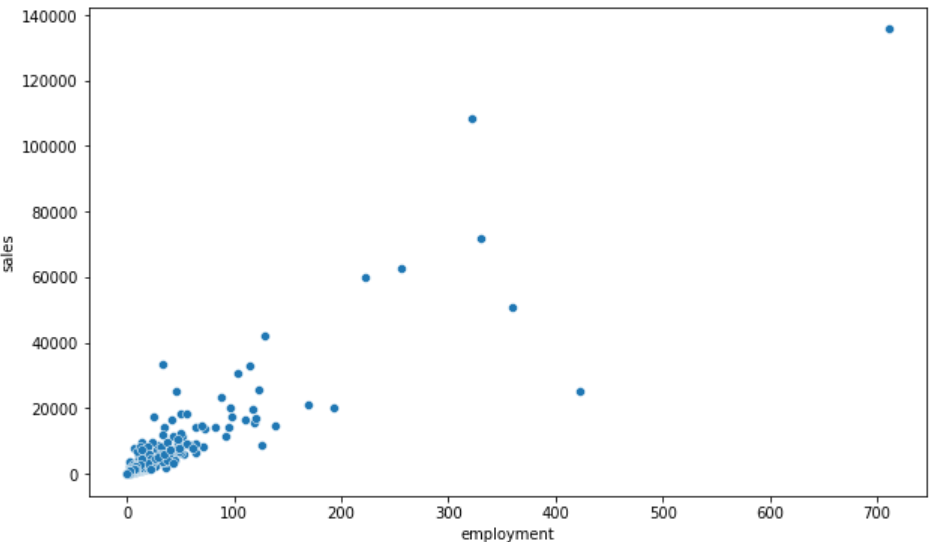
***Figure 11: Scatterplot of Patents and Sales***

The above figure depicts the relationship between the variables patents and sales. We can see that there is no definitive relationship between the variables.



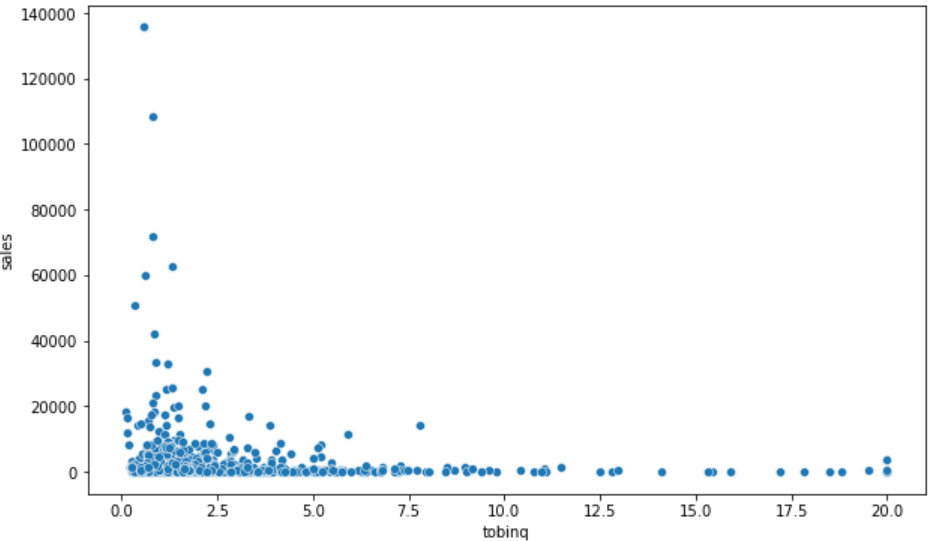
***Figure 12: Scatterplot of R&D and Sales***

The above figure depicts the relationship between the variables R&D and sales. We can see that there is no definitive relationship between the variables.

****

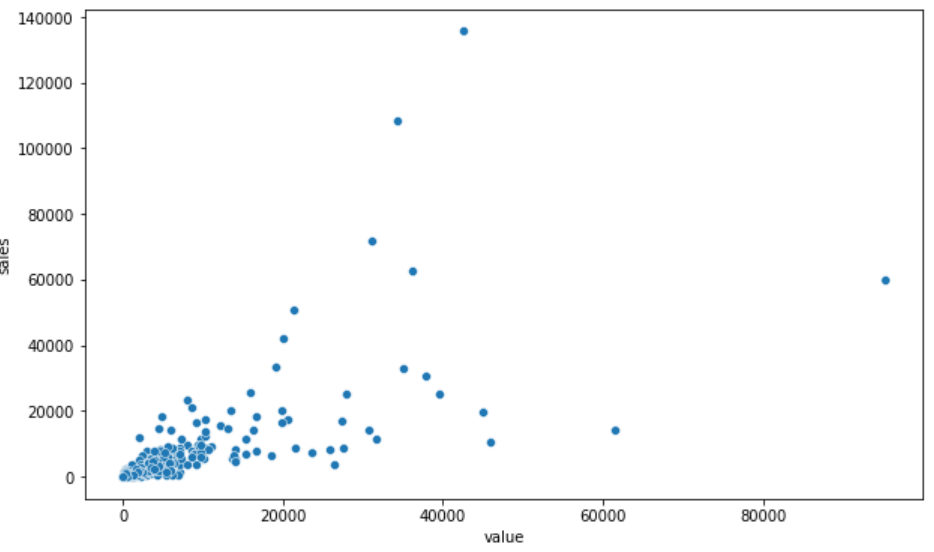
***Figure 13: Scatterplot of Employment and Sales***

The above figure depicts the relationship between the variables employment and sales. We can see that there is no definitive relationship between the variables.

****

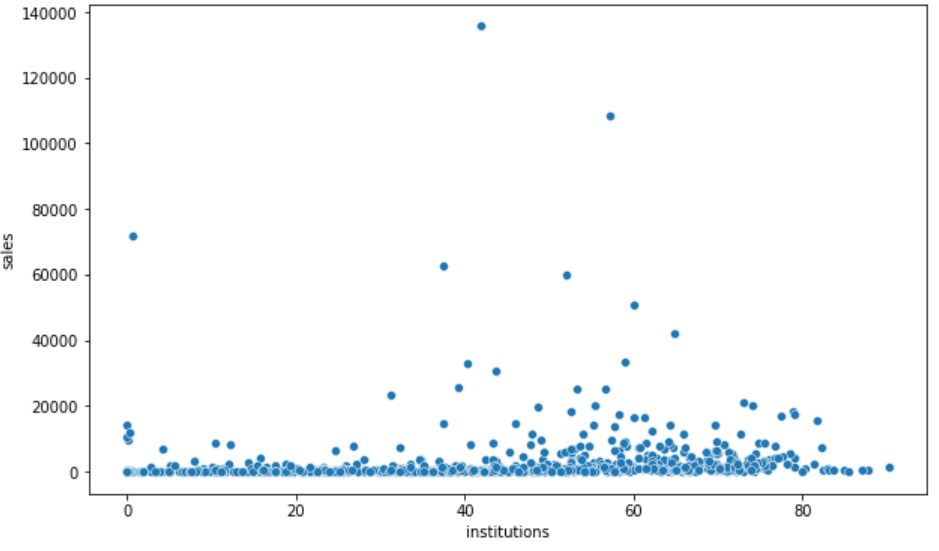
***Figure 14: Scatterplot of Tobinq and Sales***

The above figure depicts the relationship between the variables tobinq and sales. We can see that there is no definitive relationship between the variables.



***Figure 15: Scatterplot of Value and Sales***

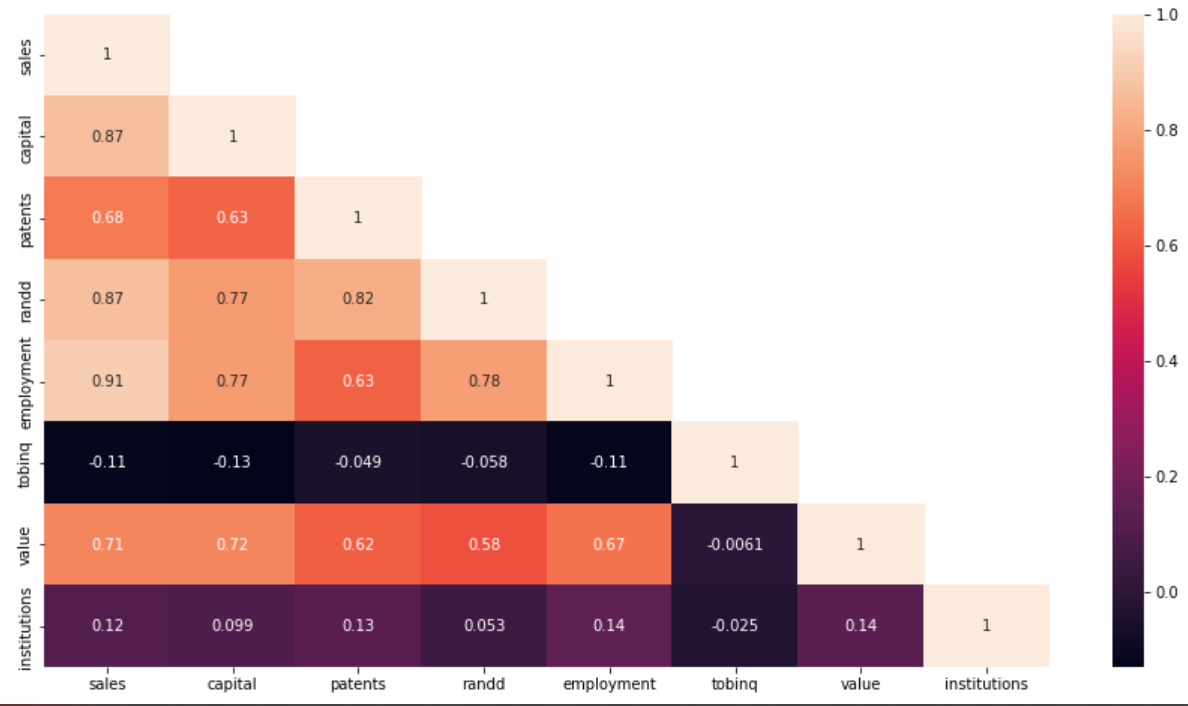
The above figure depicts the relationship between the variables Value and sales. We can see that there is no definitive relationship between the variables.



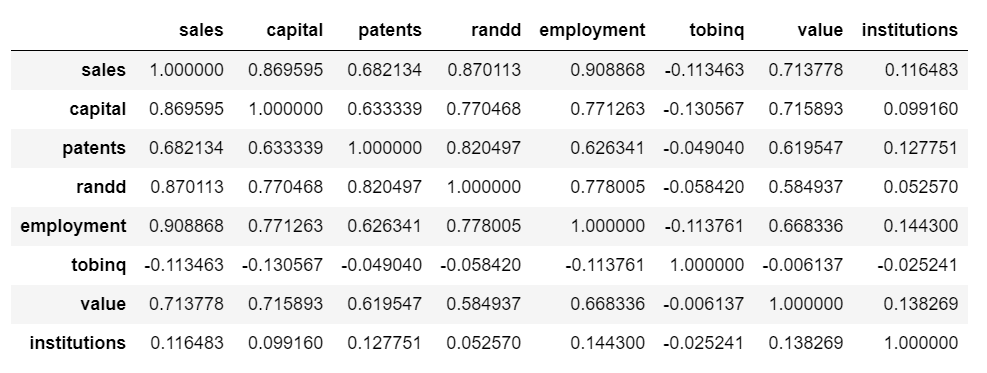
***Figure 16: Scatterplot of Institutions and Sales***

The above figure depicts the relationship between the variables institutions and sales. We can see that there is no definitive relationship between the variables.

Multivariate Analysis:



***Figure 17: Multivariate Analysis of the Dataset***



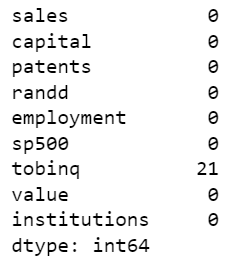
***Table 5: Correlation Matrix of the dataset***

Observations of EDA:

* There are 759 rows and 10 columns in the dataset.
* There are some missing values in the variable tobinq, which we will take care of in the later stages. There are no duplicate values in the dataset
* Almost all the variables have outliers which we need to take care of to proceed with our modelling.
* Almost all the variables are either float or int data type except sp500 which is object data type.
* There are no clearly visible relationship between the dependent and the independent variables. Although some data points suggest a correlation but we need to look further to establish a clear relationship.
* The heatmap plot of the dataset shows that there is a relationship between R&D and sales, Capital and sales, patents & sales, Value and Sales and Employment and sales.
* The heat map also reveals some other correlations between the variables in the dataset.
  1. **Impute null values if present? Do you think scaling is necessary in this case?**

**Solution:**

The following are the variables which contains null values:



***Table 6: Missing Values in the Dataset***

Only the variable Tobinq contains 21 null values. We will impute these values with median.

Scaling:

Yes, scaling is necessary since the dataset is not normally distributed and as we can see all the columns have their own range of distribution and because of that the algorithm might not work properly is why scaling of the dataset is required.

The scaled data is as follows:



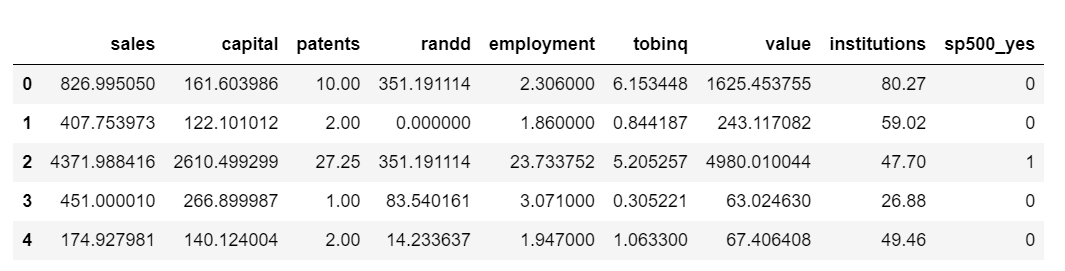
***Table 7: Scaled dataset***

* 1. **Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (30:70). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE.**

**Solution:**

Encoding:

The variable sp500 is the only variable that is object type. We will perform one hot encoding on this variable. The following is the output after the encoding process:



***Table 8: One Hot encoding on sp500***

Data Split:

For the data split we will first split the dependent and the independent variables and then split these into train and test dataset in the ration70:30.

The following is the first few rows of the train and test dataset.





***Table 9: First few rows of the Train & Test Dataset***

Linear Regression:

After the EDA and missing value treatment and the outlier treatment, we performed one hot encoding and scaling to the dataset. This made our dataset ready for Linear Regression. Through Linear Regression we wish to find out the variables which helps us in predicting the sales based on the variables given. It will give us the relationship between the dependent and the independent variables.

We perform the linear regression by using “.fit” method and find out the result.

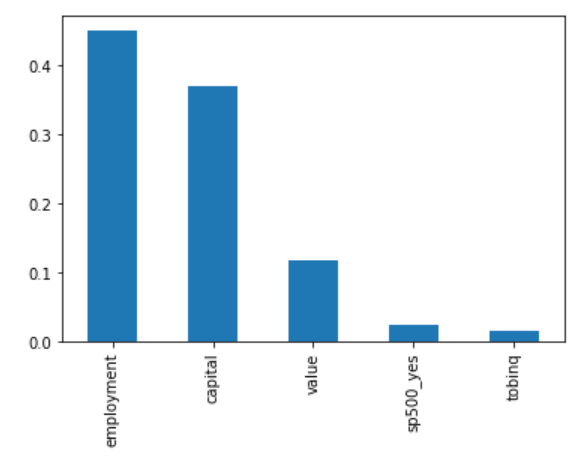
Performance Matrix:

To measure the effectiveness of the Linear regression model we have to measure the output. For this we will find out the r-squared and RMSE value of the train and test dataset.

The r-square value is 93.4% and RMSE is 389.6 for the train data.

The r-squared value is 93.2% and RMSE is 405 for the test data. The model is working fine and there is no overfitting and underfitting of the data to the model.

The following are the 5 variables which greatly effect the sales.



***Figure 18: Top 5 variables which influence the sales***

* 1. **Inference: Based on these predictions, what are the business insights and recommendations.**

**Solution:**

Insights & Recommendations.

* The dataset consists of 759 rows and 10 columns.
* There are outliers present in almost all the variables. Either this could be due to poor data collection or the presence of extreme data.
* There are some missing values present in the tobinq column of the dataset. This was then imputed with the median values.
* The sale vale is affected by the employment rate. We have observed that when the employment number increases, the sale also increases. The organisation should look into hiring skilled workforce.
* The sale value is also affected by the capital invested. The sale value tends to go up as the capital invested amount goes up.
* The sale value also affected by the value, sp500 and tobinq variable.
* The R&D and the patens features also affects the sale value. The innovation in the products and new patents seem to be affecting the sales a lot.

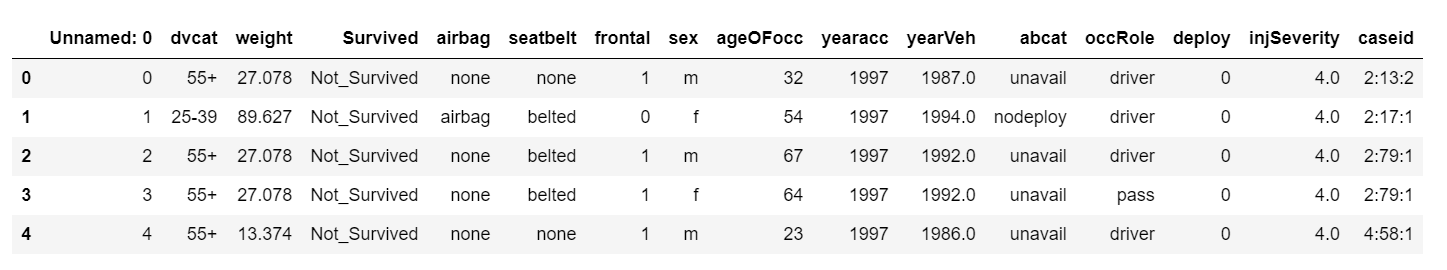
**Problem 2:**

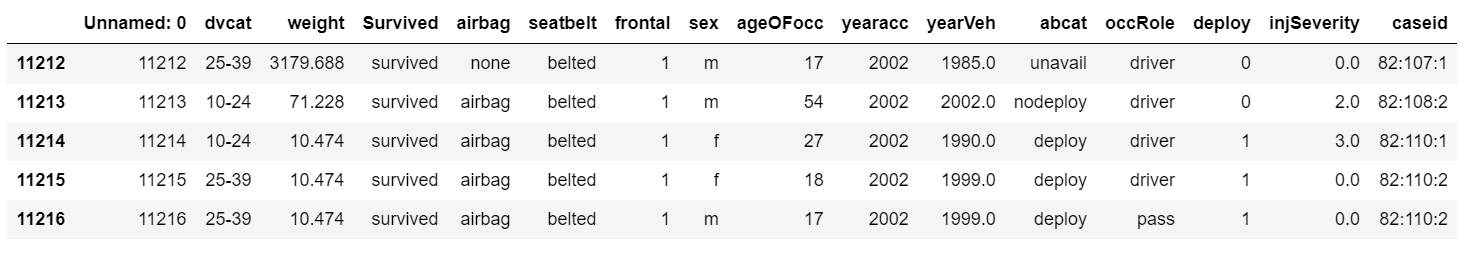
You are hired by the Government to do an analysis of car crashes. You are provided details of car crashes, among which some people survived and some didn't. You have to help the government in predicting whether a person will survive or not on the basis of the information given in the data set so as to provide insights that will help the government to make stronger laws for car manufacturers to ensure safety measures. Also, find out the important factors on the basis of which you made your predictions.

**2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.**

**Solution:**

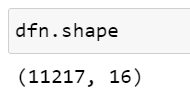
The dataset consists of data related to car crashes provided to us by the government. We will first read the dataset and see what is provided to us.



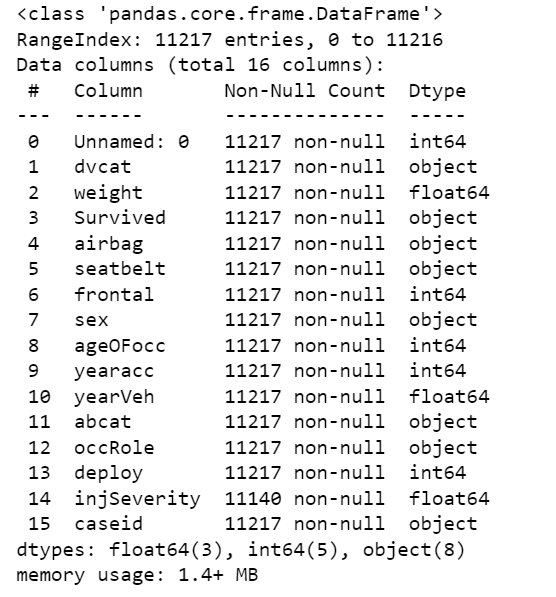


***Table 10: First and Last few rows of the dataset***

The shape of the dataset is as follows. There are 11217 rows and 16 columns in the dataset.

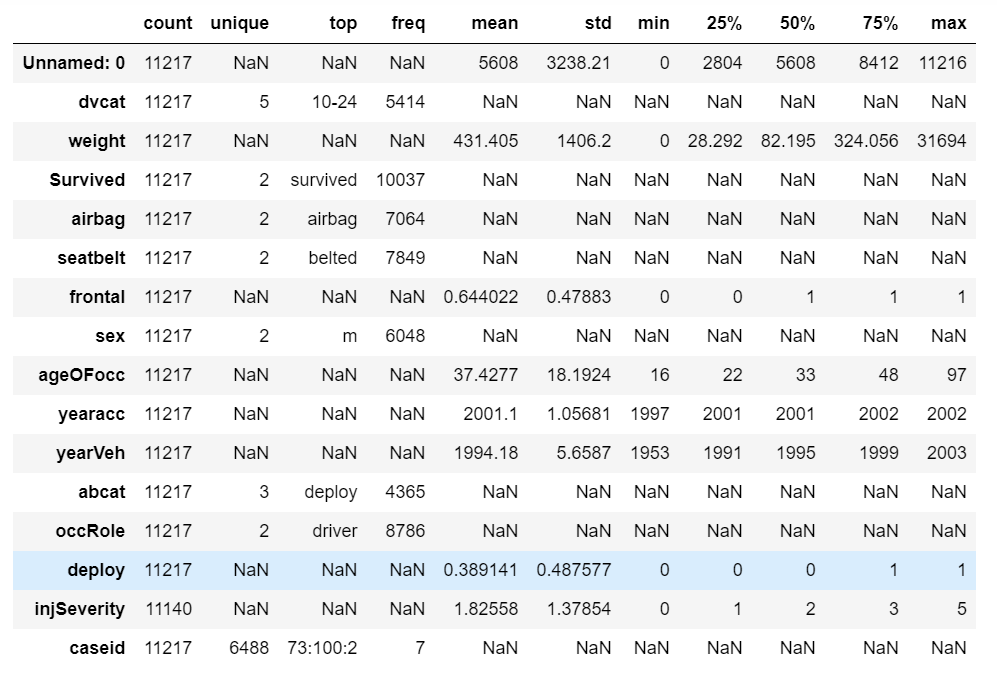


The info of the dataset is as follows:



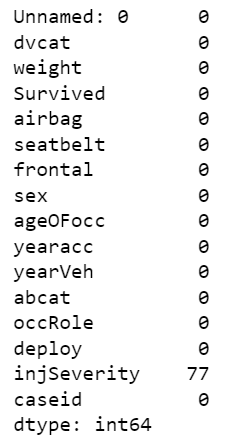
***Table 11: Info of the dataset***

The description of the dataset is as follows:



***Table 12: Description of the dataset***

Missing Value check:

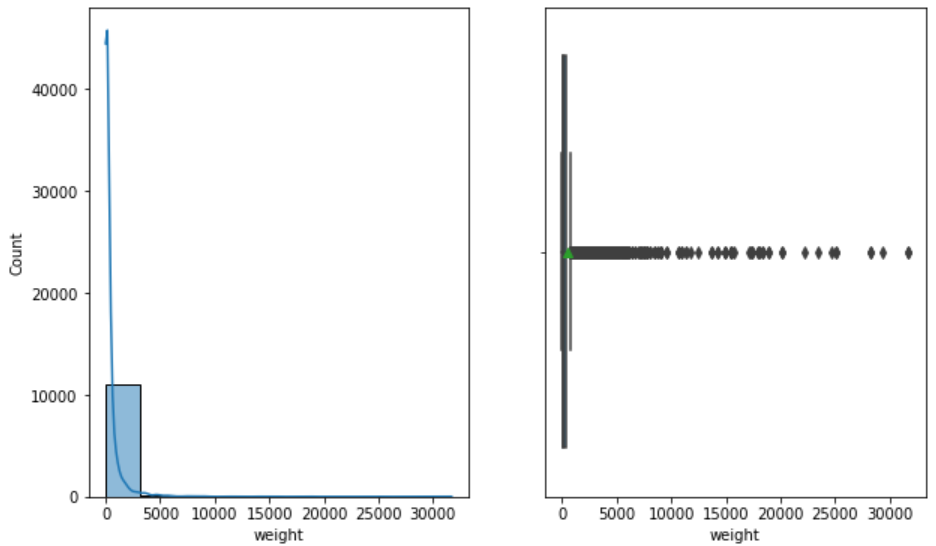


***Table 13: Missing Value Check***

Observations:

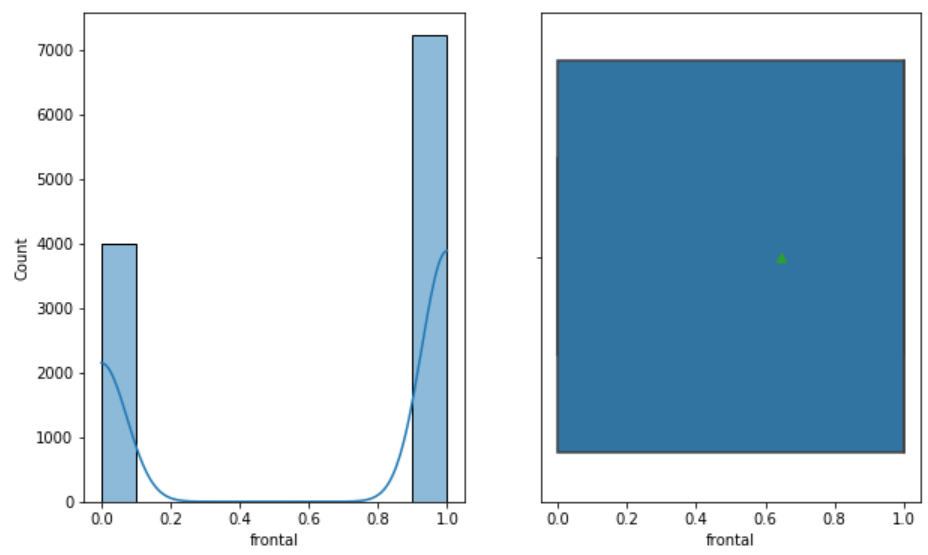
* The dataset contains data regarding the car crash and the observations during these car crashes.
* There are 11217 rows and 16 columns to the dataset.
* There are null values present only in the column InjSeverity which was replaced by mode of the variable.
* Almost 54% of the car crashes happened to male drivers and 46% are female drivers.
* In almost 70% of the car crashes drivers were wearing seat belts.
* In almost 89.4% of the car crashes the driver survived.

Exploratory Data Analysis.



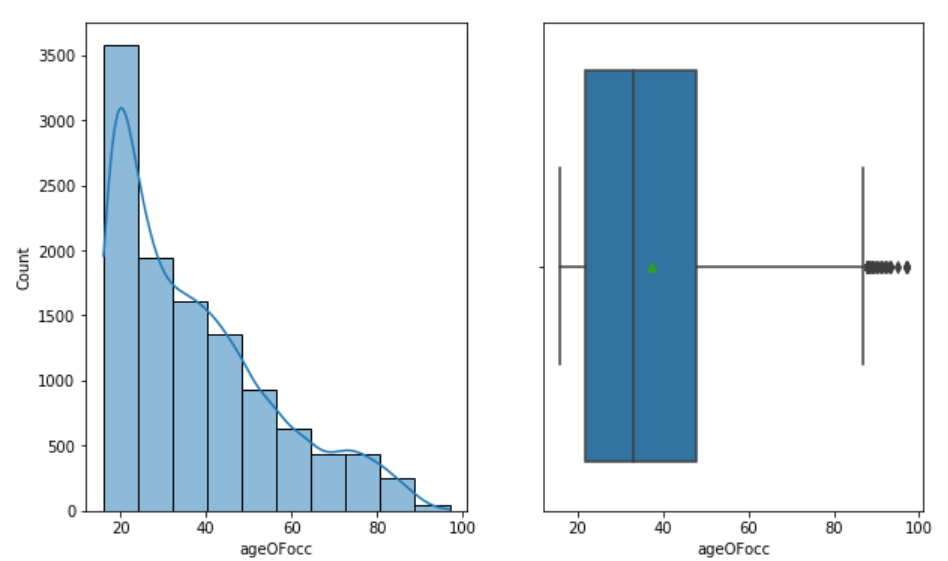
***Figure 19: Distribution of Weight Variable***

The above figure shows the distribution of weight of the vehicles. The average weight is 431.4 with a minimum weight of 0 and a maximum weight of 31694. There are lot of outliers present in the dataset.



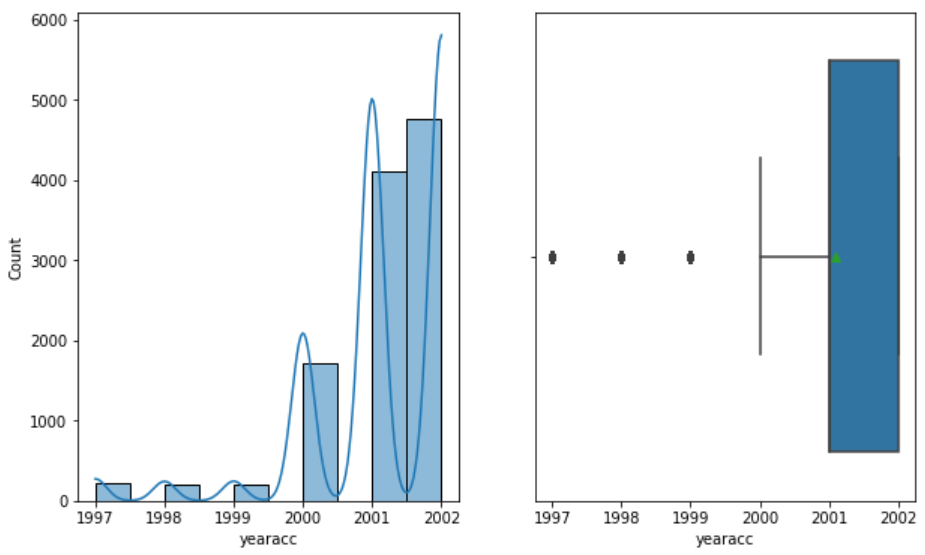
***Figure 20: Distribution of Frontal Variable***

The above figure tells us whether the accident had frontal impact of non-frontal impact. We can see that 64.4% of the accidents had frontal impact and 35.6% had non frontal impact.



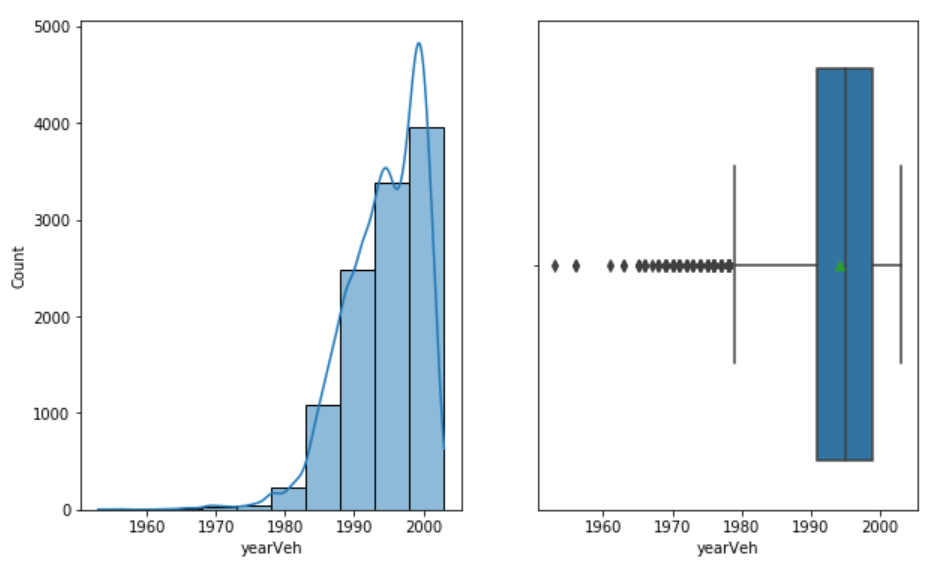
***Figure 21: Distribution of Age of Occupant***

The figure shows the age of occupants in the car crash. The age ranges from 16 years to 97 years and the mean age is 37 years. There are outliers present in the data.



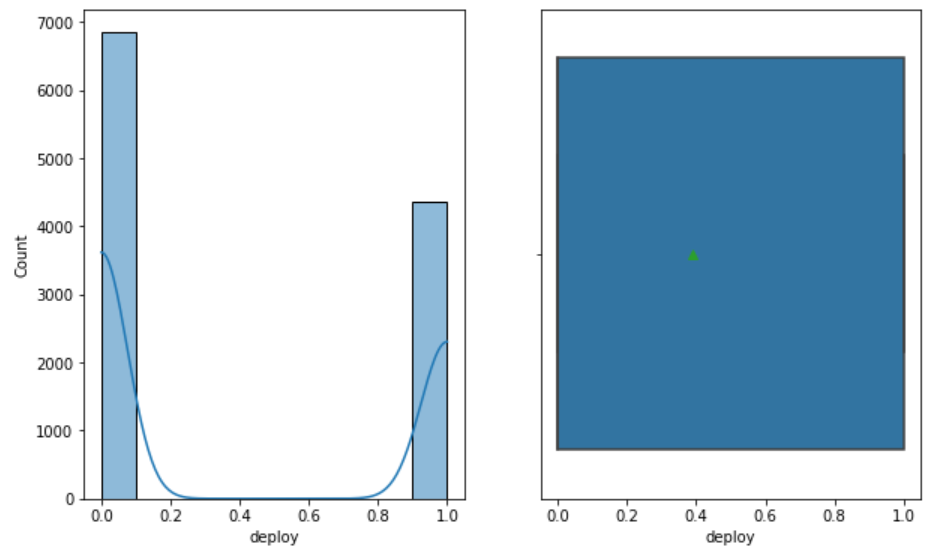
***Figure 22: Distribution of Year of accident***

The dataset shows the year of accident of the car. We can see that the rate of car crash has significantly increased in the year 2001 and 2002. 79% of the accidents took place in the year 2001 and 2002.



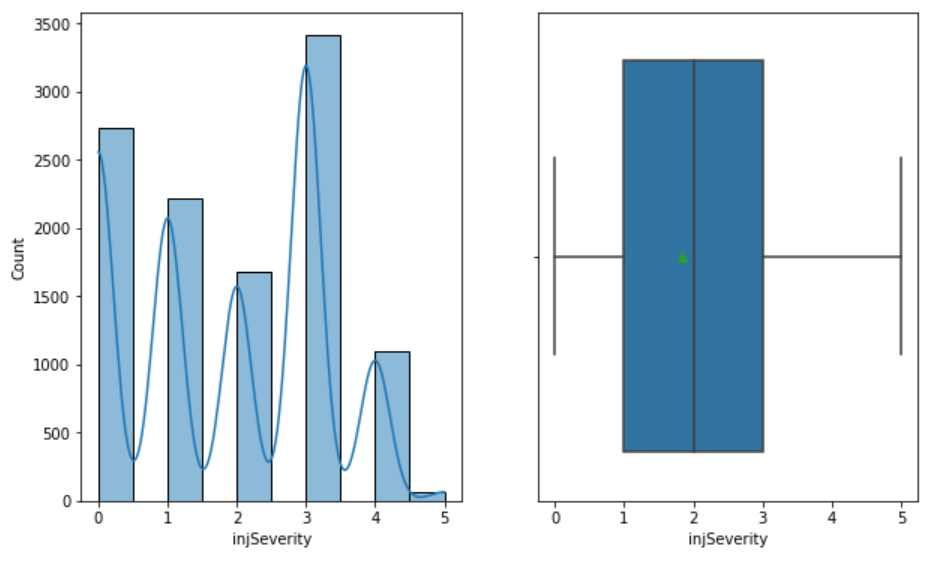
***Figure 23: Distribution of Year of Vehicle***

The above figure shows us the distribution of the vehicle according to the year in which they were manufactured. We can see that there are lot of vehicles in the dataset which were manufactured post 1980 that was involved in a car crash with the highest accident taking place in the vehicle of 1999(1046 crashes).



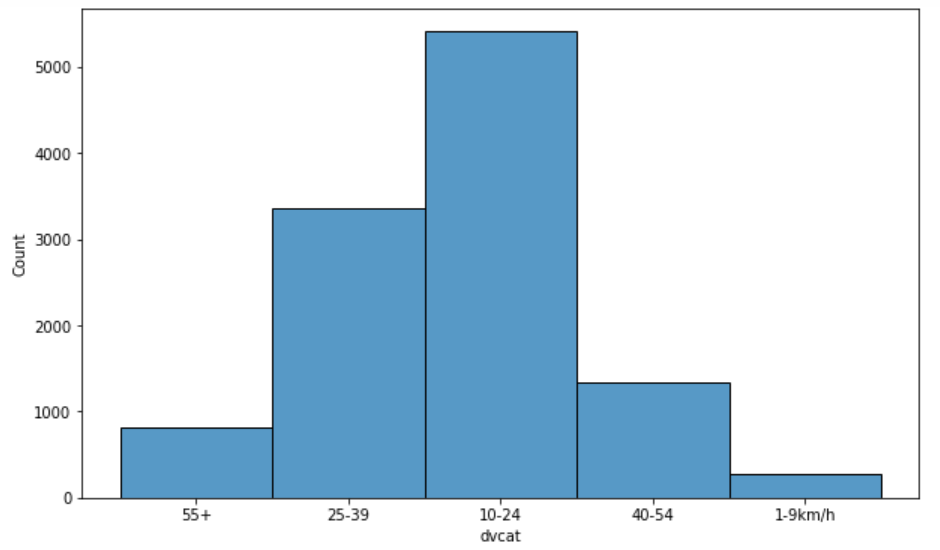
***Figure 24: Distribution of Airbag Deploy***

The above figure displays whether the airbag deployed during the crash or not. We can see that 61% of the times the airbag did not deploy.



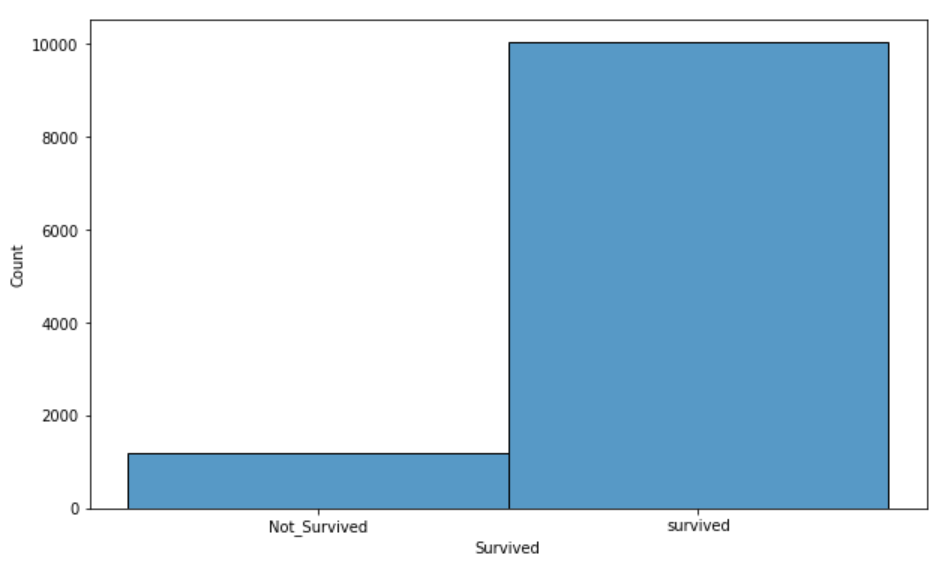
***Figure 25: Distribution of Severity of Injury.***

The above figure describes the severity of the injury of the car crash. We can see that the rate of incapacity is higher than the others. Approx 30% of the car crashes resulted in incapacitation followed be 24% with no injury and 19.7% of possible injury followed by the rest.



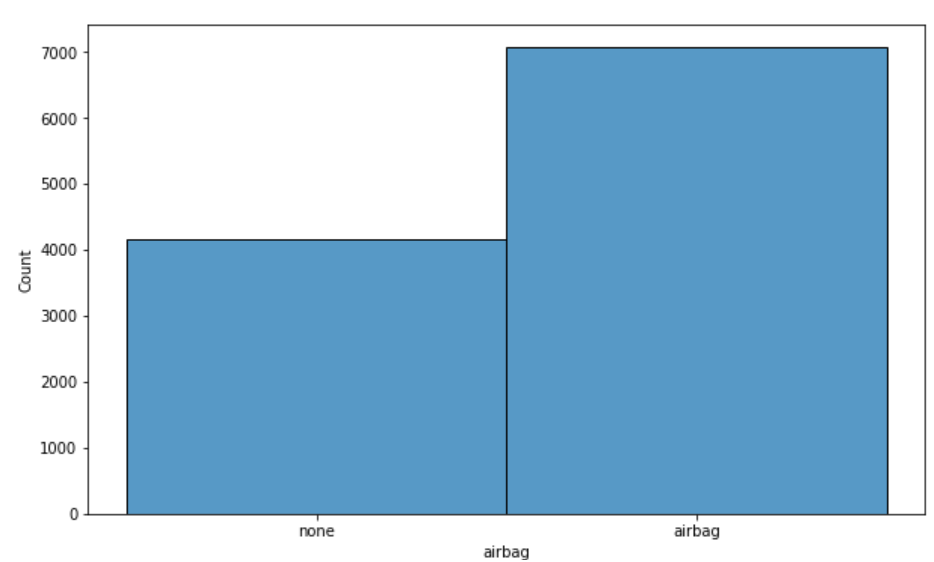
***Figure 26: Distribution of dvcat***

The above figure shows the distribution of the impact speed. The highest is the 10-24 km/hr speed followed by 25-39 km/hr and 40-54km/hr.



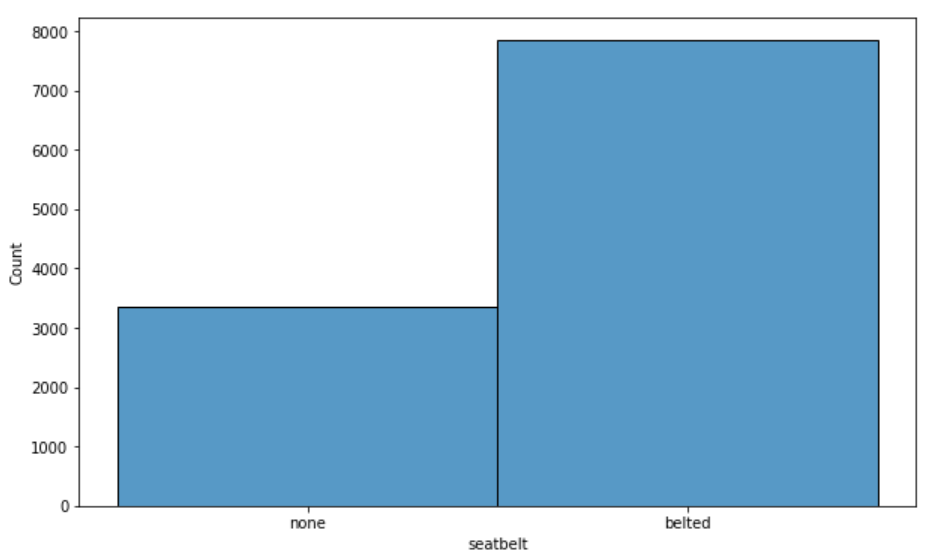
***Figure 27: Distribution of Survived Variable***

The above figure shows the distribution of survival in the car crash. We can see that 89.5% of the times the driver survived the car crash and only 10.5% of the times the driver did not survive.



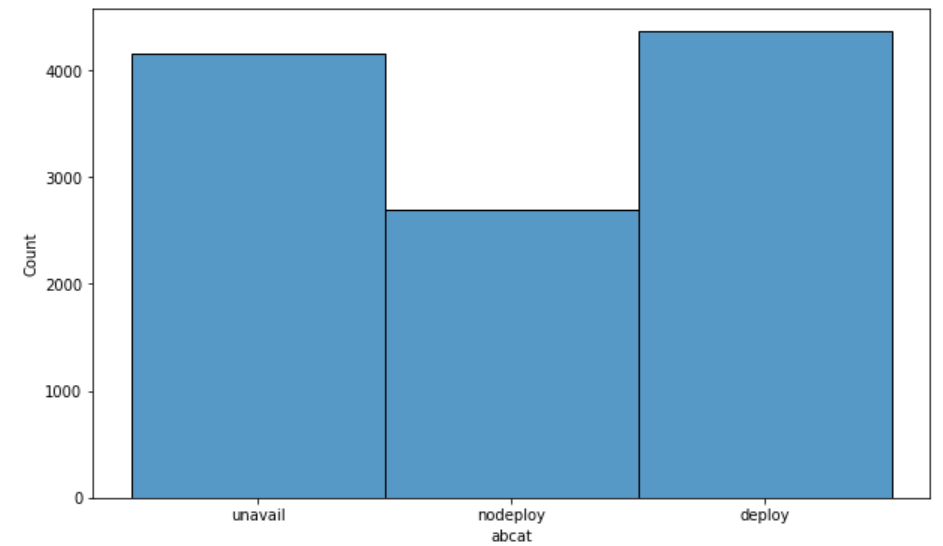
***Figure 28: Distribution of Airbag Variable***

The above figure shows the distribution of airbags in the vehicles. We can see that almost 63% of the vehicles involved in crash had airbags and only 37% of the vehicles did not have airbags.



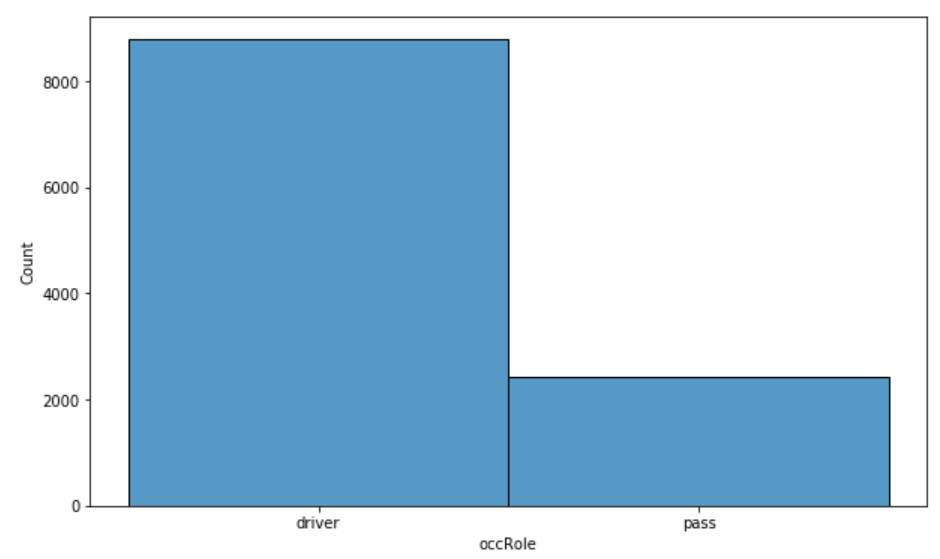
***Figure 29: Distribution of Seatbelt***

The above figure shows whether the driver was wearing seatbelts or not during the crash. We can see that 70% of the drivers were wearing seatbelts and the rest 30% of the car crash drivers were not wearing belts.



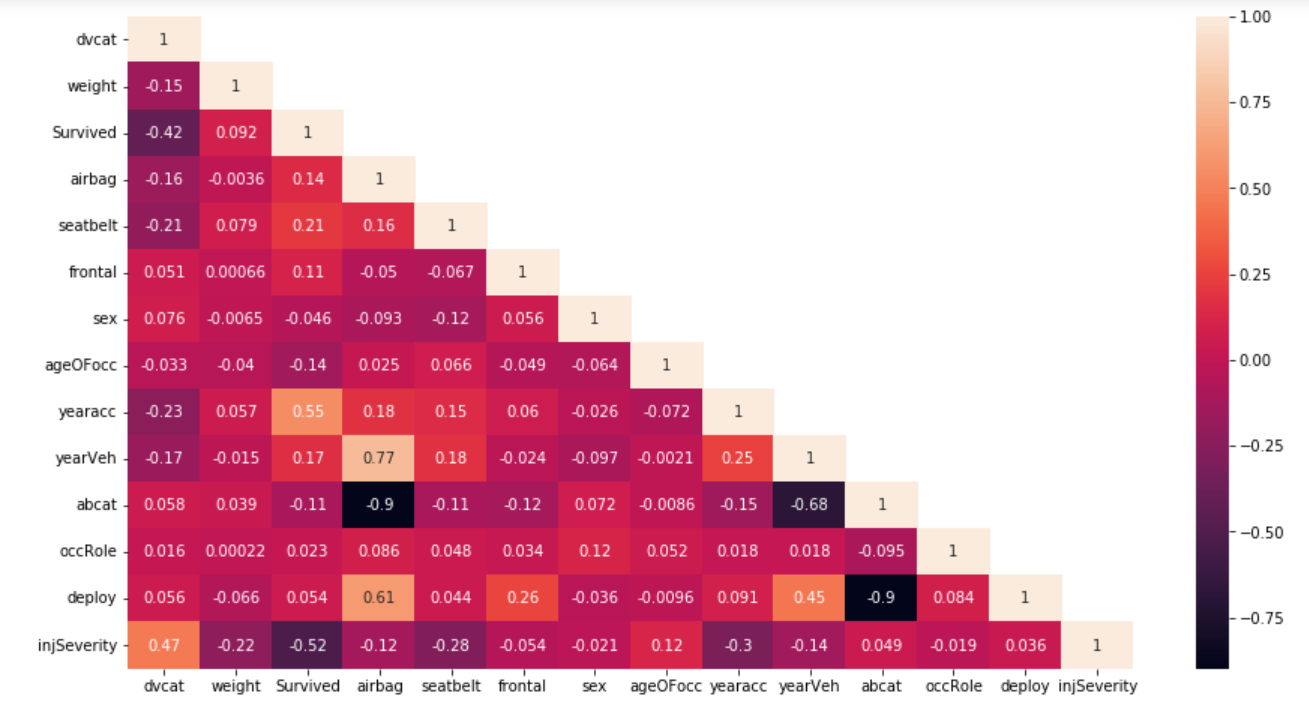
***Figure 30: Distribution of abcat***

The above figure shows the distribution of airbag deploy during a crash. We can see that 39% of the time the airbags deployed, 37% of the time airbag was not available and 24% of the time the airbags did not deploy.



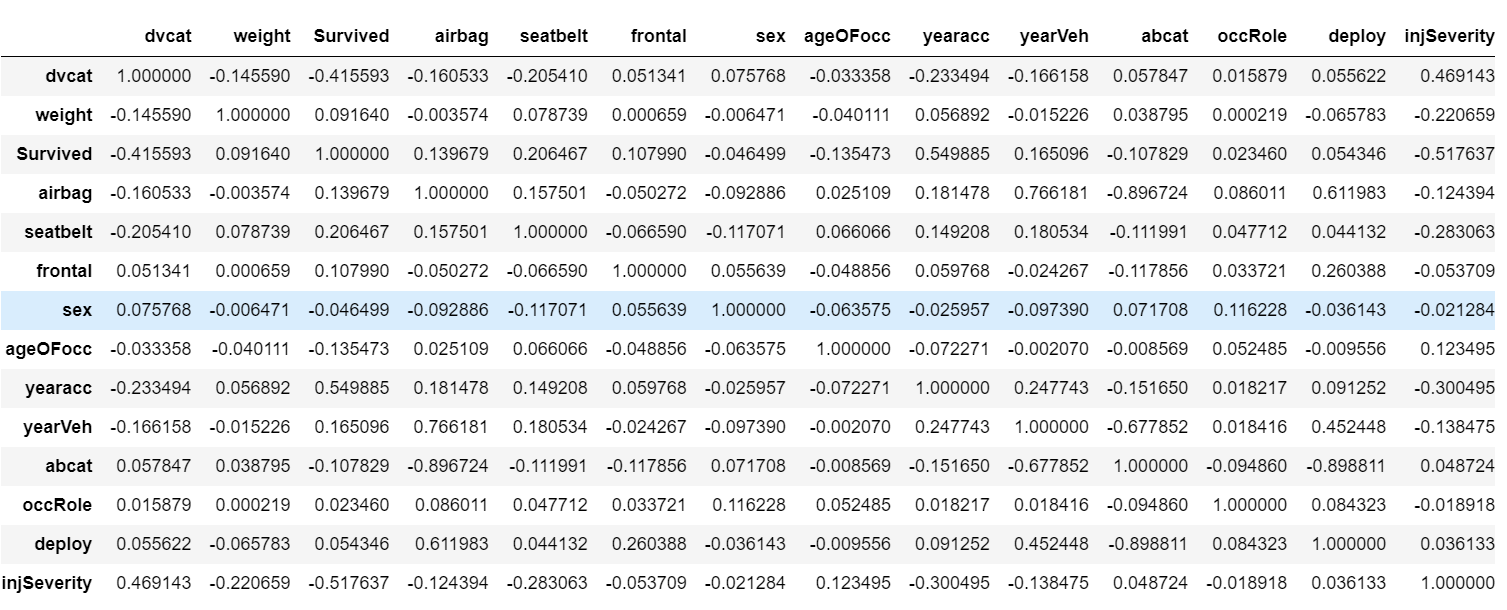
***Figure 31: Distribution of occRole***

The above figure shows the distribution whether the driver or the passenger was injured in the crash. We can see that 78% of the times driver was injured in the crash and only 22% of the time the passenger was injured.



***Figure 32: Multivariate Analysis of the dataset***

The above figure shows the correlation between the different variables in the dataset. The lighter colours depict strong correlation where as dark colours represent low correlation between the variables.



***Table 14: Correlation Matrix of the Dataset***

**2.2) Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis)**

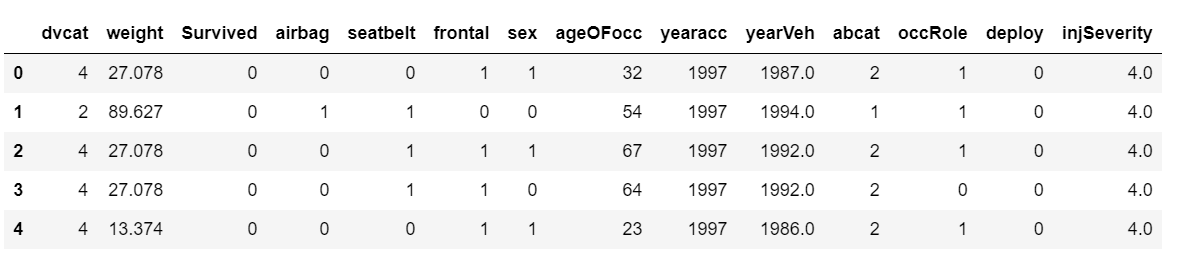
**Solution:**

Data encoding refers to the process of converting data from one representation or format to another, typically with the goal of making it suitable for a specific purpose or analysis. Encoding is a fundamental concept in data preprocessing and manipulation, and it plays a crucial role in data analysis, machine learning, and various other data-related tasks.

Data encoding is a critical step in data preprocessing because it helps ensure that the data is in a suitable format for analysis or modelling. For our dataset, some of our variables needed encoding in order to apply to modelling.

Variables like ‘Survived’, ‘airbag’, ‘seatbelt’, ‘sex’, ‘occRole’ where there were only two inputs were replaced with 1 and 0.

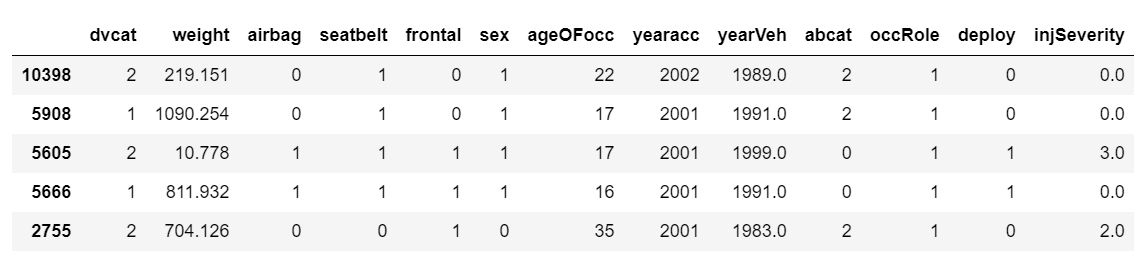
Label encoding is applied to rest of the required variables for the ease of modelling.

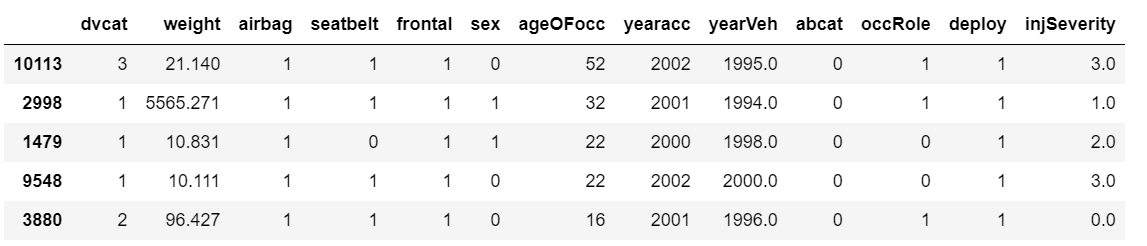


***Table 15: First few rows after encoding***

Train-Test Split.

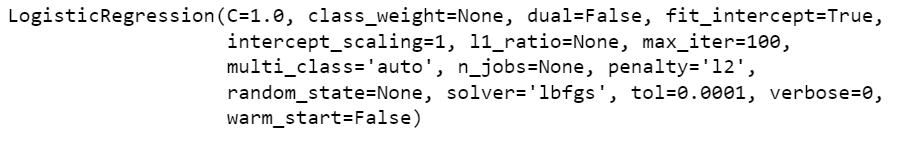
Before going ahead with the modelling step, we need to split the data into train and test. The first and the last few rows of the dataset after the split are as follows.





***Table 16: First and last few rows of train dataset***

Logistic Regression.



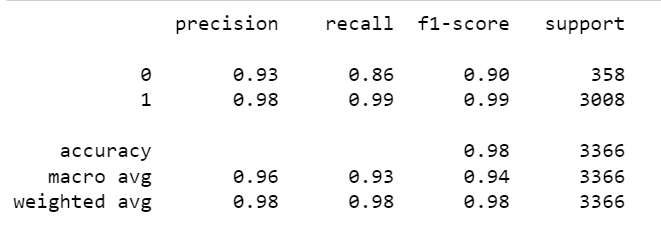
**2.3)** **Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Compare both the models and write inferences, which model is best/optimized.**

**Solution:**

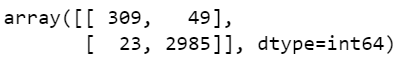
The accuracy score of the logistic regression is



The classification report is as follows:



The confusion matrix is as follows:



**2.4)** **Inference: Based on these predictions, what are the insights and recommendations.**

**Solutions:**

Inferences and Recommendations.

* The dataset contains data regarding the car crash and the observations during these car crashes.
* There are 11217 rows and 16 columns to the dataset.
* There are null values present only in the column InjSeverity which was replaced by mode of the variable.
* Almost 54% of the car crashes happened to male drivers and 46% are female drivers.
* In almost 89.4% of the car crashes the driver survived.
* The average weight of the vehicle is 431.4 with a minimum weight of 0 and a maximum weight of 31694. There are lot of outliers present in the dataset.
* We have observed that 64.4% of the accidents had frontal impact and 35.6% had non frontal impact.
* The age ranges from 16 years to 97 years and the mean age is 37 years. There are outliers present in the data.
* We have observed that the rate of car crash has significantly increased in the year 2001 and 2002. 79% of the accidents took place in the year 2001 and 2002.
* We have observed that there are lot of vehicles in the dataset which were manufactured post 1980 that was involved in a car crash with the highest accident taking place in the vehicle of 1999(1046 crashes).
* We can see that 61% of the times the airbag did not deploy. This needs to be looked into so that safety of the riders can be ensured.
* We can see that the rate of incapacity is higher than the others. Approx 30% of the car crashes resulted in incapacitation followed be 24% with no injury and 19.7% of possible injury followed by the rest.
* The highest is the 10-24 km/hr speed followed by 25-39 km/hr and 40-54km/hr in a car crash. The government should ensure that the drivers are following the speed limit.
* We can see that 89.5% of the times the driver survived the car crash and only 10.5% of the times the driver did not survive.
* We can see that almost 63% of the vehicles involved in crash had airbags and only 37% of the vehicles did not have airbags. The government should take steps to ensure that all vehicles have appropriate airbags available.
* We can see that 70% of the drivers were wearing seatbelts and the rest 30% of the car crash drivers were not wearing belts. The government should take steps to ensure proper traffic rules are observed by the drivers.
* We can see that 39% of the time the airbags deployed, 37% of the time airbag was not available and 24% of the time the airbags did not deploy.
* We can see that 78% of the times driver was injured in the crash and only 22% of the time the passenger was injured. This tells us that the drivers are at a greater risk in any crash and the government should take steps to provide safety of the drivers.