**Project: Predictive Modeling**

Pavan Kalyan B | PGP DSBA Oct\_C’21 | Jun 2022

**Table of contents**

**Problem-1**

**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.  
**1.2)**  Impute null values if present? Do you think scaling is necessary in this case.  
**1.3)**  Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE.  
**1.4)**  Inference: Based on these predictions, what are the business insights and recommendations.

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

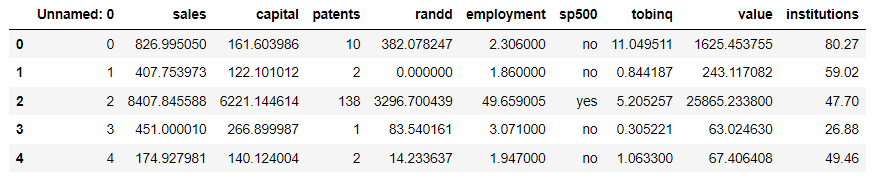
**Problem 1: Linear Regression**

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.

Data Dictionary for Firm\_level\_data:

1. sales: Sales (in millions of dollars).  
2. capital: Net stock of property, plant, and equipment.  
3. patents: Granted patents.  
4. randd: R&D stock (in millions of dollars).  
5. employment: Employment (in 1000s).  
6. sp500: Membership of firms in the S&P 500 index. S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States  
7. tobinq: Tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value.  
8. value: Stock market value.  
9. institutions: Proportion of stock owned by institutions.

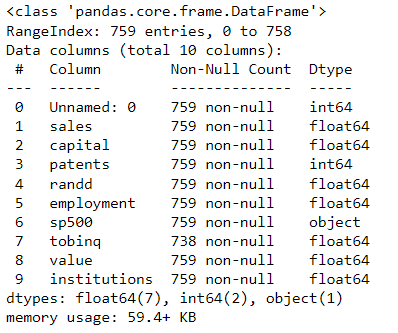
* 1. **Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis.**

****

**Fig 1.1.1 Firm data – first 5 rows**

We can observe that:

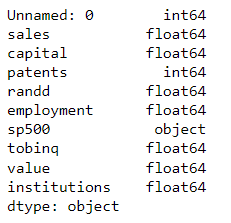
1. sales: Sales (in millions of dollars).  
2. capital: Net stock of property, plant, and equipment.  
3. patents: Granted patents.  
4. randd: R&D stock (in millions of dollars).  
5. employment: Employment (in 1000s).  
6. sp500: Membership of firms in the S&P 500 index. S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States  
7. tobinq: Tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value.  
8. value: Stock market value.  
9. institutions: Proportion of stock owned by institutions.



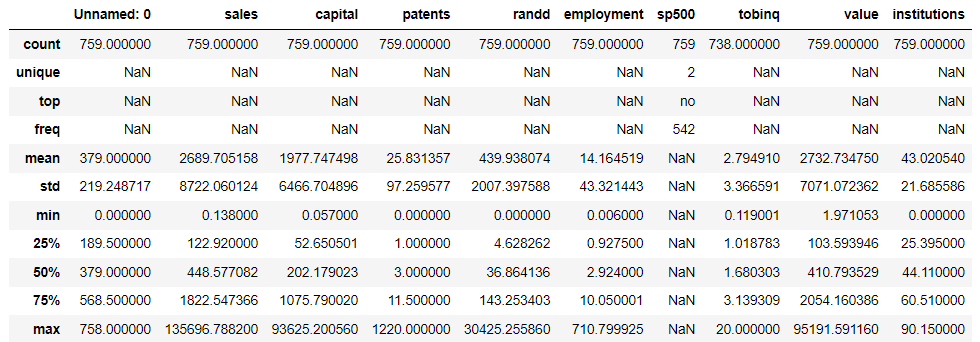
**Fig 1.1.2 Information of table**

We can observe that:

* Sp500 is an object column and rest all are numbers
* We have 759 rows and 10 columns including ‘unnamed: 0’
* The memory used here is 59.4+ kb

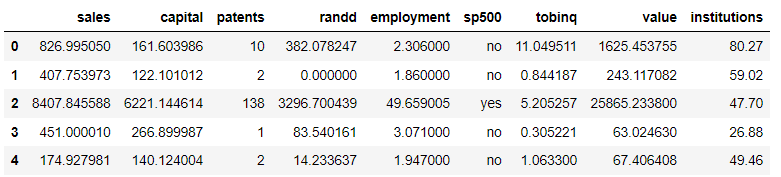


**Fig 1.1.3 Data types of each column**

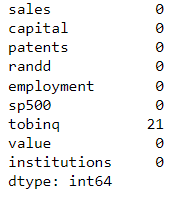
****

**Fig 1.1.4 Description of each column**

* Sales has the 0 as min and 135696.788200 as maxi
* Capital is 0.05 as min as 93625.200 as maxi
* Patents has 0 as min and 1220 as maxi
* Randd has 0 as min and 30425 as maxi
* Employment has 0.006 as min and 710 as maxi
* Tobinq has 0.11 as min and 20 as max

****

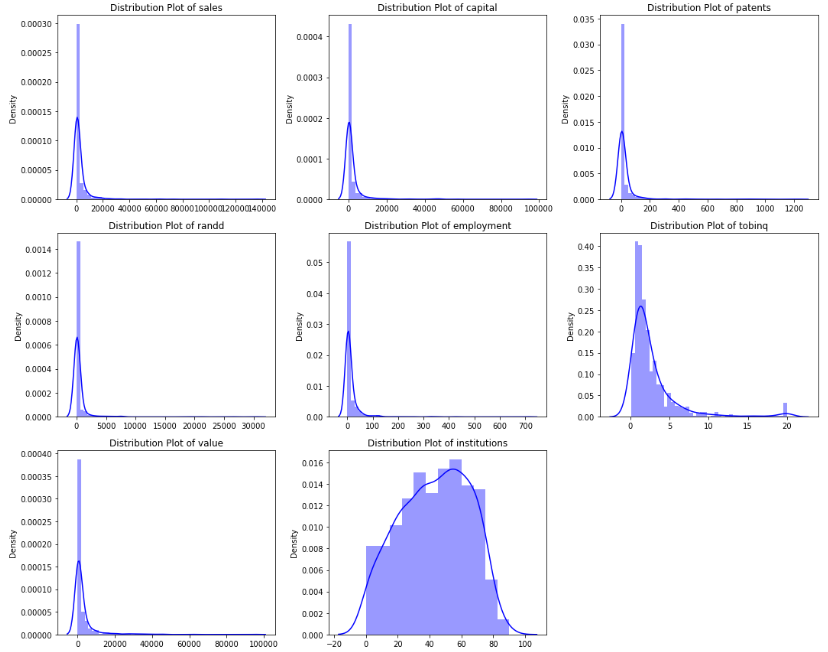
**Fig 1.1.5 After removing the unnamed column**

****

**Fig 1.1.6 Missing values**

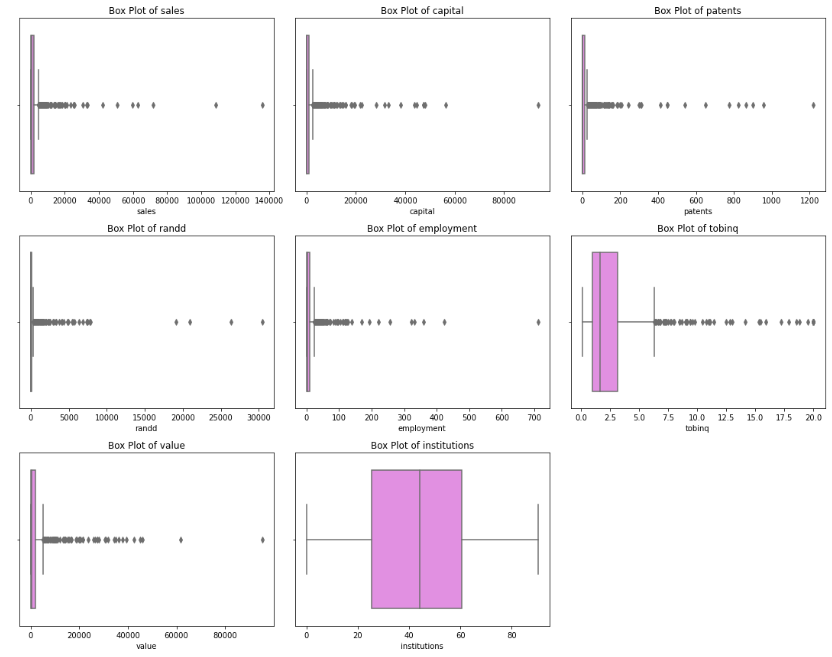
* We can observe that column tobinq has 21 missing values we have to take some actions to avoid this.

**UNIVARIATE ANALYSIS:**

****

**Fig: 1.1.7 Distribution plot for whole data**

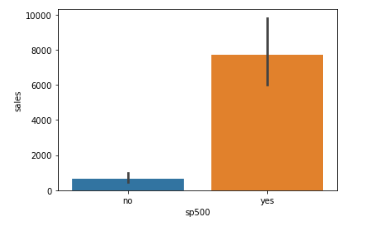
* We can observe that columns ‘sales’, ‘capital’, ‘patents’, ‘randd’, ‘employment’ are normally distributed
* ‘tobinq’ and ‘value’ are right skewed

****

**Fig 1.1.8 box plot for whole data**

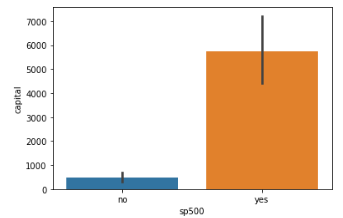
* All the columns except ‘institutions’ have outliers

**BIVARIATE ANALYSIS:**

****

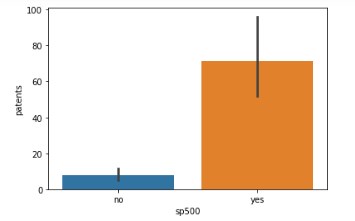
**Fig 1.1.9 Bar plot sales with sp500**

* Sales with yes is higher than no in sp500

****

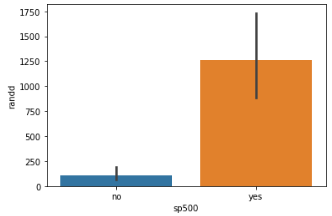
**Fig 1.1.10 Bar plot capital with sp500**

* Capital with yes is higher than no in sp500

****

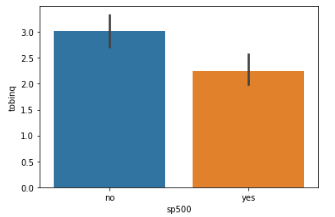
**Fig 1.1.11 Bar plot patents with sp500**

* Patents with yes is higher than no in sp500

****

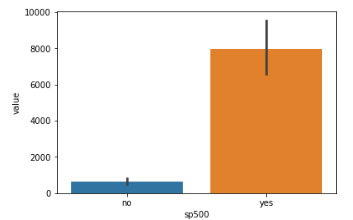
**Fig 1.1.12 Bar plot randd with sp500**

* Randd with yes is higher than no in sp500

****

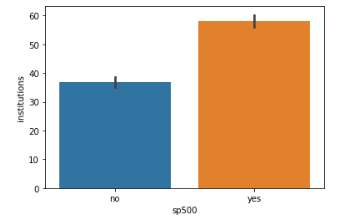
**Fig 1.1.13 Bar plot tobinq with sp500**

* Tobinq with no is higher than yes in sp500

****

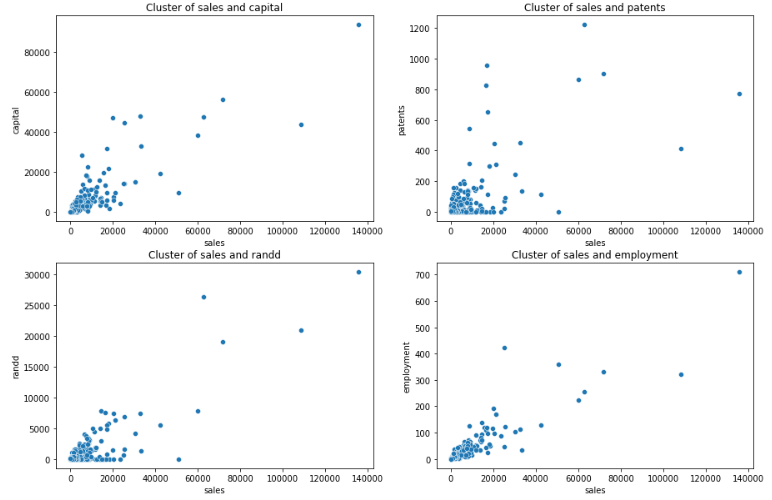
**Fig 1.1.14 Bar plot value with sp500**

* Value with yes is higher than no in sp500

****

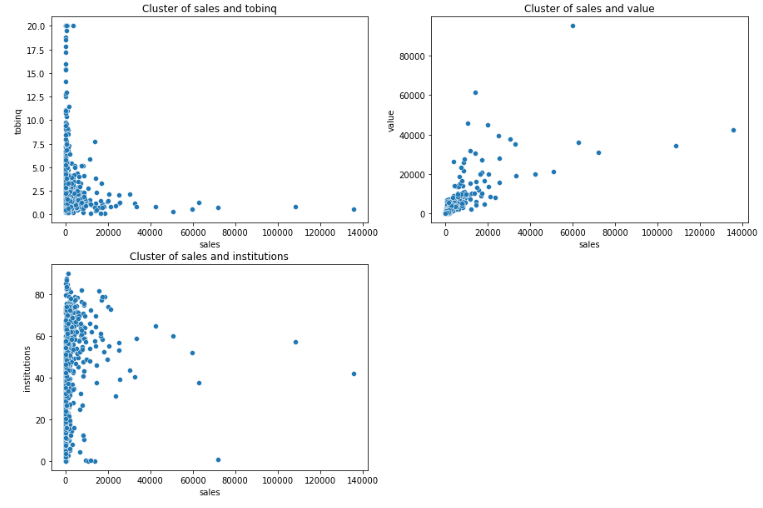
**Fig 1.1.15 Bar plot institutions with sp500**

* institutions with yes is higher than no in sp500

****

**Fig 1.1.16 Scatterplot**

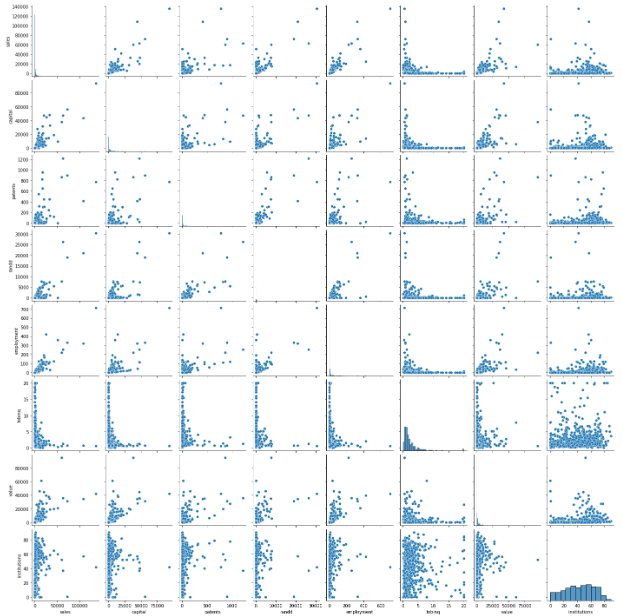
* We can observe that there is moderately positive correlation in the scatter

****

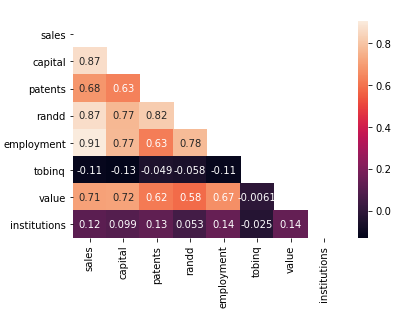
**Fig 1.1.17 Scatterplot**

* We can observe that there is no proper pattern in the plots

**PAIR PLOT:**

****

**Fig 1.1.18 Pair plot**

****

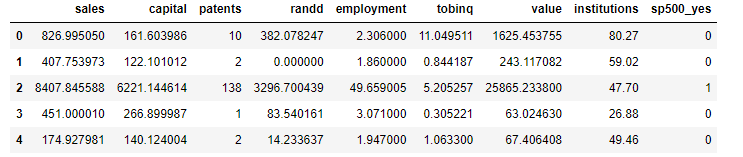
**Fig 1.1.19 Heat map plot for correlation**

* 1. **Impute null values if present? Do you think scaling is necessary in this case**

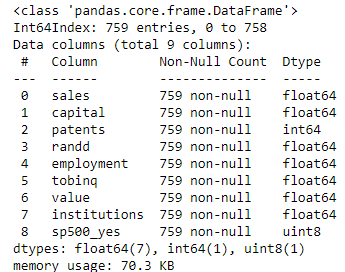
****

**Fig 1.2.1 Null value check**

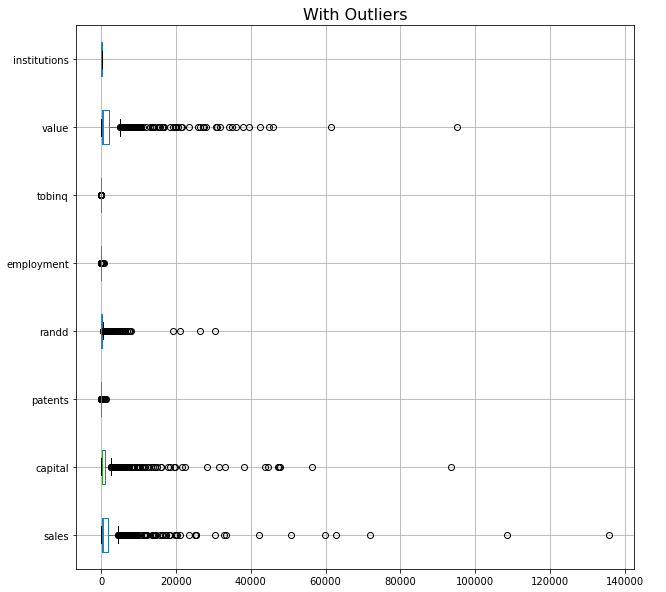
* We can observe that there is no null values as we have replaced with mean values

****

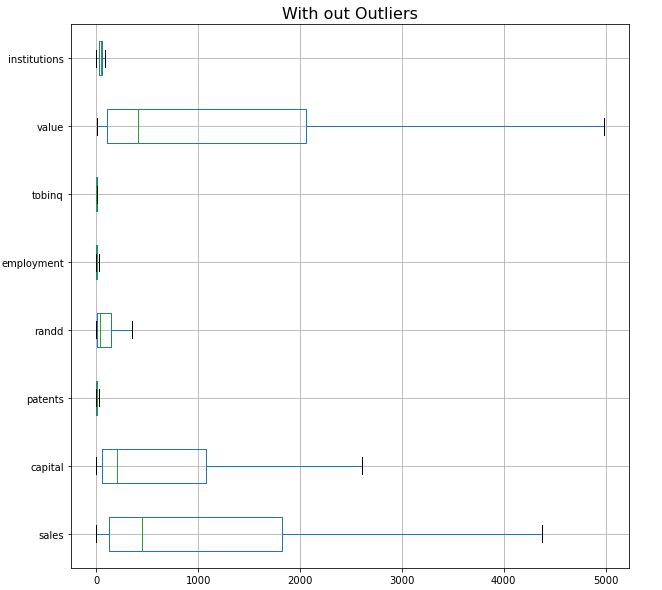
**Fig 1.2.2 Converting the sp500 column to dummies**

****

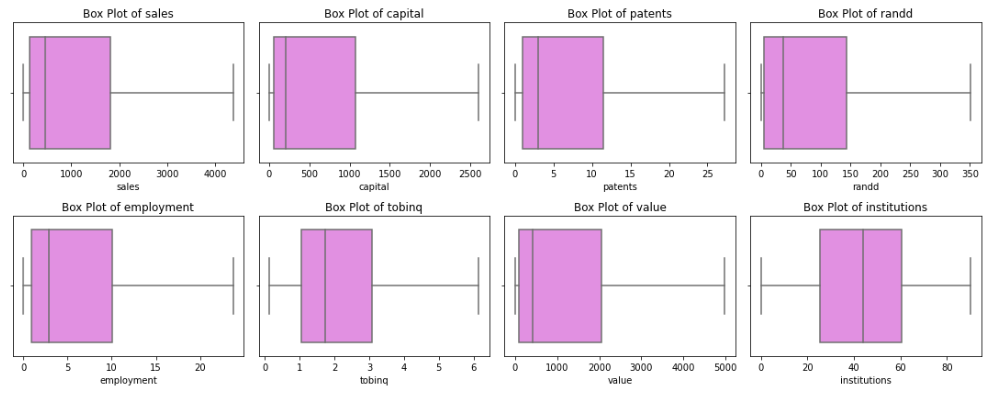
**Fig 1.2.3 Information of columns**

****

**Fig 1.2.4 Before outlier treatment**

****

**Fig 1.2.5 After outlier treatment**

****

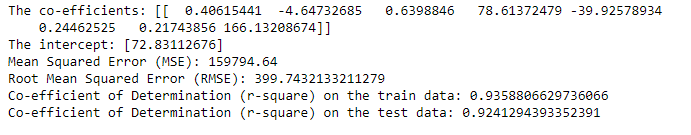
**Fig 1.2.6 Box plots for all columns after outlier treatment**

* We can observe that there are no outliers in the data now as it has been treated

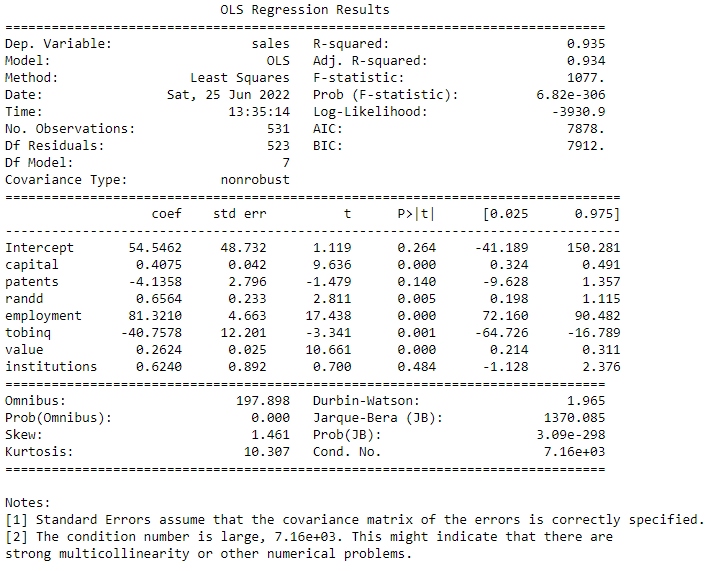
**1.3)** **Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE.**

**Model 1:**

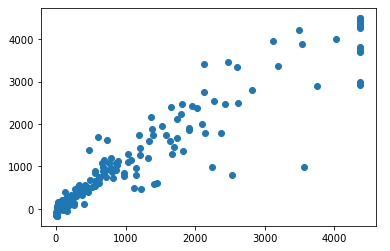
****

****

**Model 2:**

****

**Fig 1.3.1 'sales on capital+patents+randd+employment+tobinq+value+institutions'**

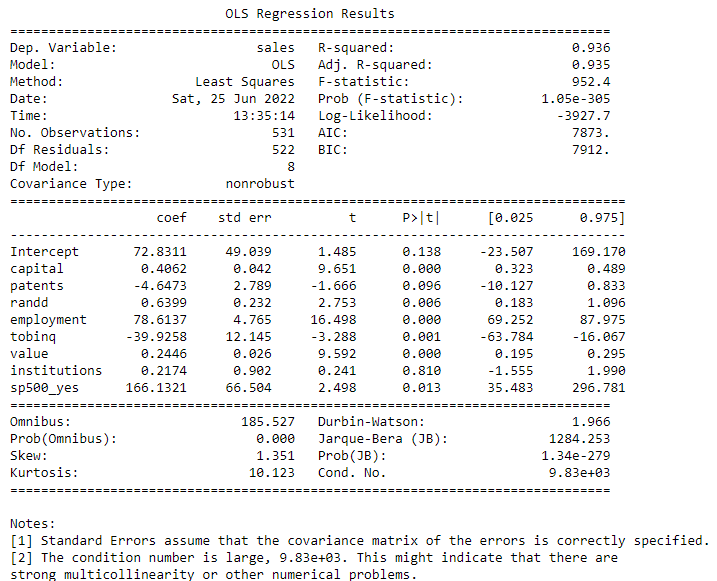
****

**Fig 1.3.2 scatter plot on sales prediction**

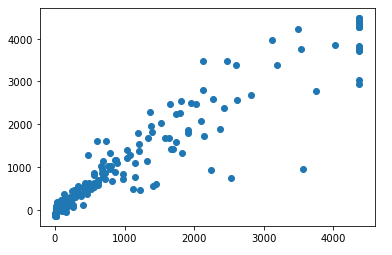
* The plot on sales is positively correlated for the prediction

****

**Model 3:**

****

**Fig 1.3.3 'sales on capital+patents+randd+employment+tobinq+value+institutions+sp500\_yes'**

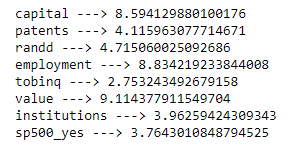
****

**Fig 1.3.2 scatter plot on sales prediction**

* The plot on sales is positively correlated for the prediction

****

**Variance Inflation factor**

****

A variance inflation factor (VIF) detects multicollinearity in regression analysis.

Multicollinearity is when there’s correlation between independent variables, this affects our model results.

VIF explains how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

we can observe that there is a heavy influence on multicollinearity which needs to be treated.

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

* Based on the insights, we can observe that firm level data is to understand which company has given the higher outcome in sales. so based on the outcome we can invest to the companies consciously.
* Based on the vif check we know that, This firm data has some attributes which are collinear 'patents', 'randd', 'tobinq', 'institutions' and 'sp500\_yes'.
* The important attributes which are impacting the sales are 'capital','employment', 'value', and 'tobinq' these play a major role in deciding of sales. After using prediction models we got to know these are the important attribute for the firm investment to decide the where to invest in sales.
* In real time or after using the model we observe that if the capital increases, the sales increases.

**Conclusion:**

Based on the models we used linear regression and stats models:

* linear regression model1 - taking all the variables expect sales we got an accuracy of 93%
* model2 - based on the price how other independent variables are correlated is shown - 93% accuracy
* model3 - based on the price how each and every individual records are correlated is shown - 93% accuracy

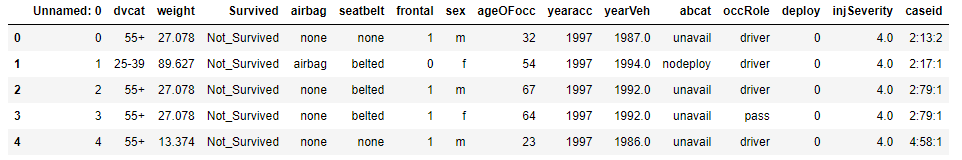
All the three models giving same accuracy so we can go with linear regression itself for futuras it is good for continuous variables.

**Problem 2: Logistic Regression and Linear Discriminant Analysis**

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.

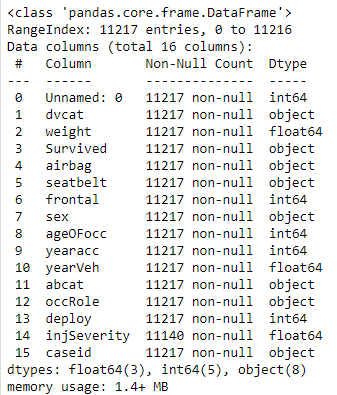
1. dvcat: factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+  
2. weight: Observation weights, albeit of uncertain accuracy, designed to account for varying sampling probabilities. (The inverse probability weighting estimator can be used to demonstrate causality when the researcher cannot conduct a controlled experiment but has observed data to model)  
3. Survived: factor with levels Survived or not\_survived  
4. airbag: a factor with levels none or airbag  
5. seatbelt: a factor with levels none or belted  
6. frontal: a numeric vector; 0 = non-frontal, 1=frontal impact  
7. sex: a factor with levels f: Female or m: Male  
8. ageOFocc: age of occupant in years  
9. yearacc: year of accident  
10. yearVeh: Year of model of vehicle; a numeric vector  
11. abcat: Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy, nodeploy and unavail  
12. occRole: a factor with levels driver or pass: passenger  
13. deploy: a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed.  
14. injSeverity: a numeric vector; 0: none, 1: possible injury, 2: no incapacity, 3: incapacity, 4: killed; 5: unknown, 6: prior death  
15. caseid: character, created by pasting together the population's sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

**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.**

****

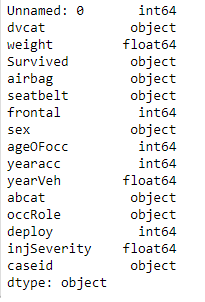
**Fig 2.1.1 Car crash data**

* The data tells use about the car accidents in the past years and shows the major injury and whether that person has survived or not.

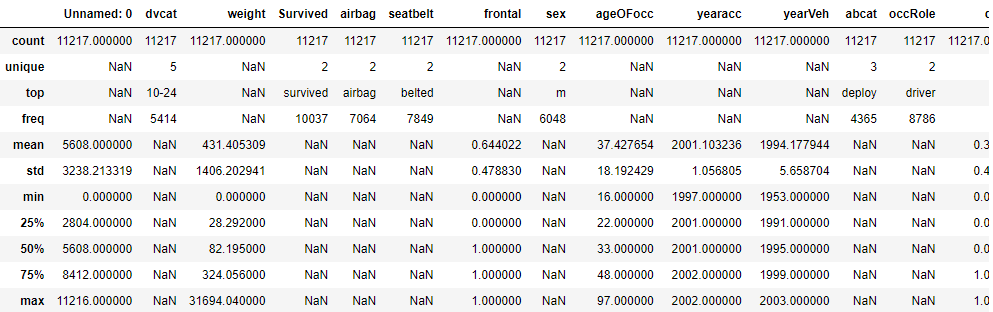
****

**Fig 2.1.2 Information of columns**

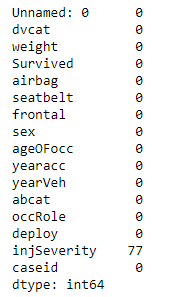
* We can observe that there are many object columns other than weight and ageofocc column which is continuous
* The memory usage used here is 1.4+ mb
* There are 11217 rows and 16 columns

****

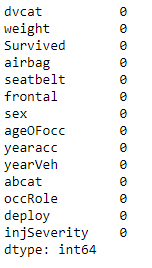
**Fig 2.1.3 Data types of each column**

****

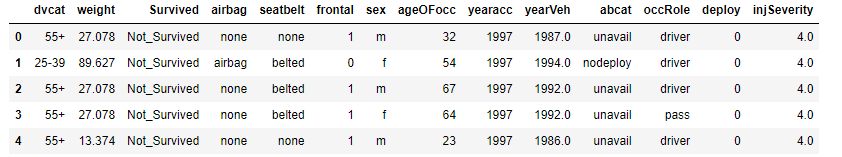
**Fig 2.1.4 Description of each column using describe**

****

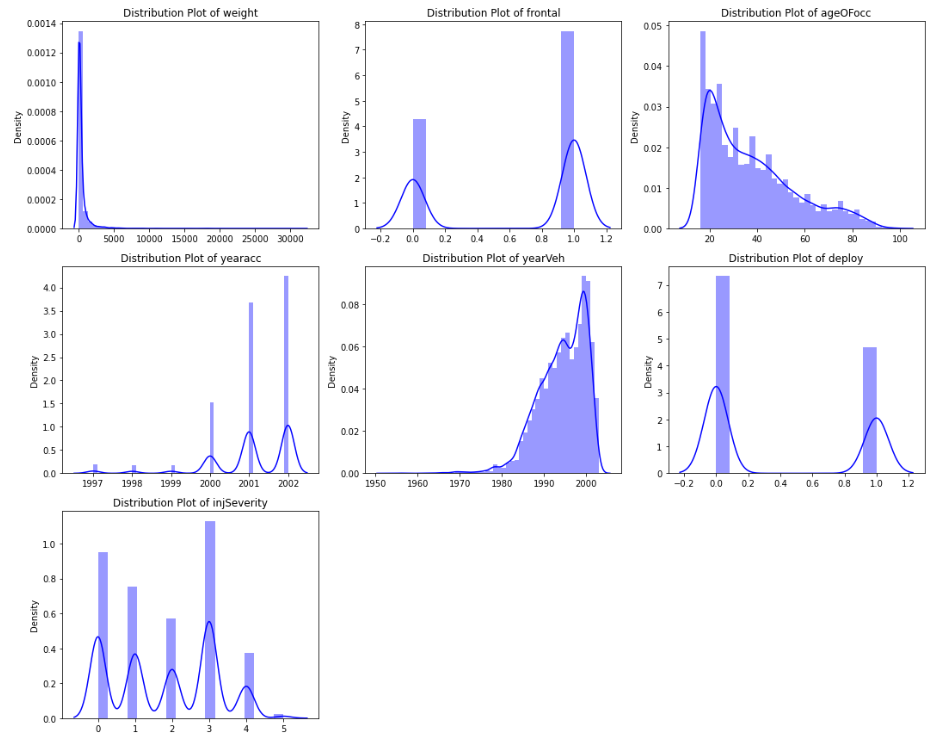
**Fig 2.1.5 Before clearing null values**

****

**Fig 2.1.6 After clearing null values**

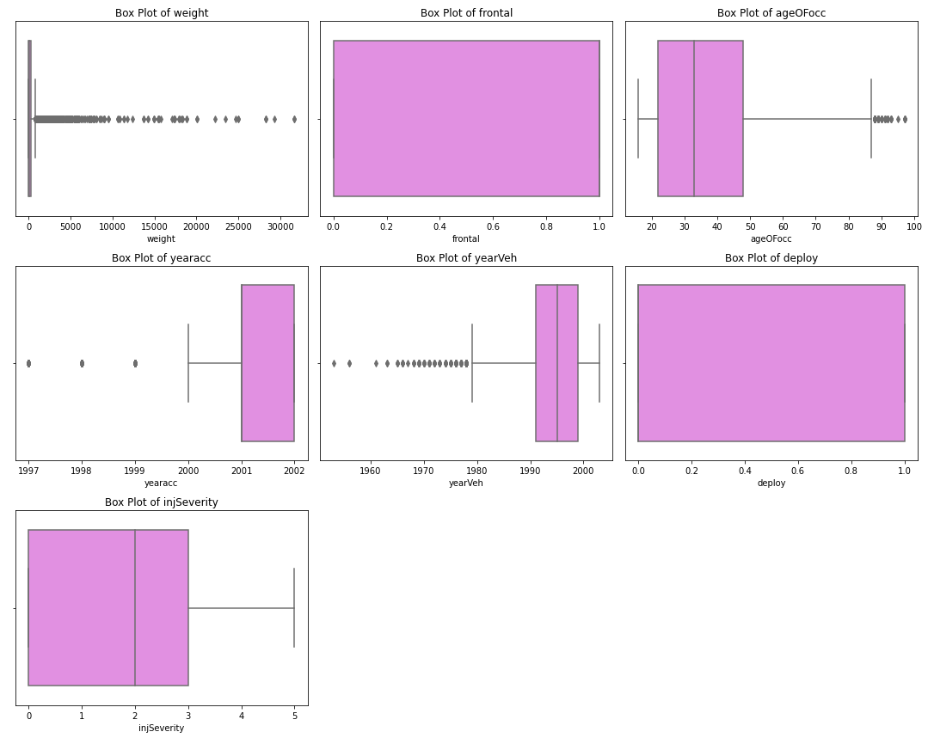
**Fig 2.1.7 After null values treatment**

**UNIVARIATE ANALYSIS:**

****

**Fig 2.1.8 Distribution plots of car data**

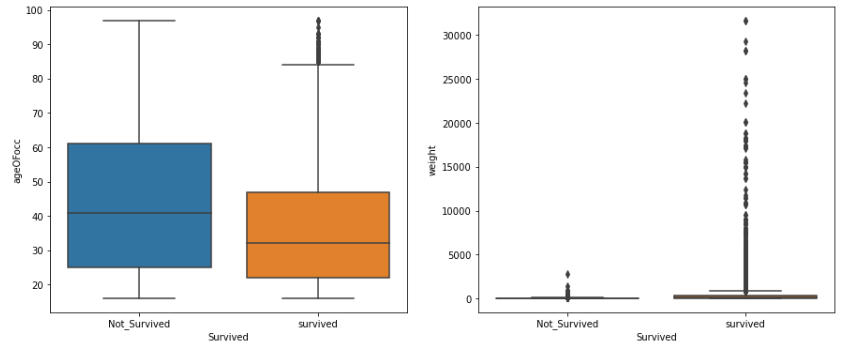
* Weight and ageofocc columns are right skewed positively
* The rest all doesn’t have any pattern

****

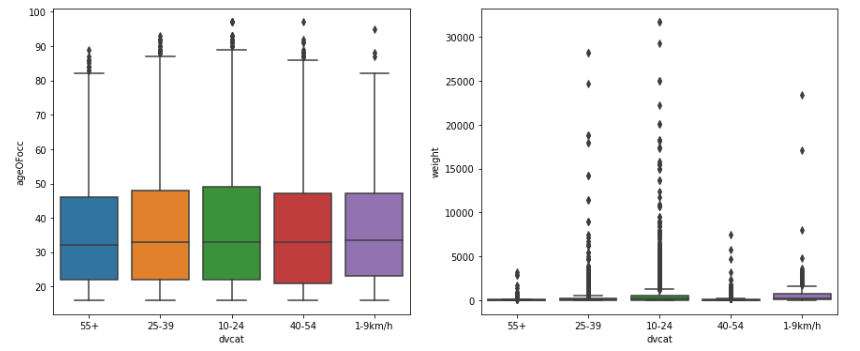
**Fig 2.1.9 Box plots of car data**

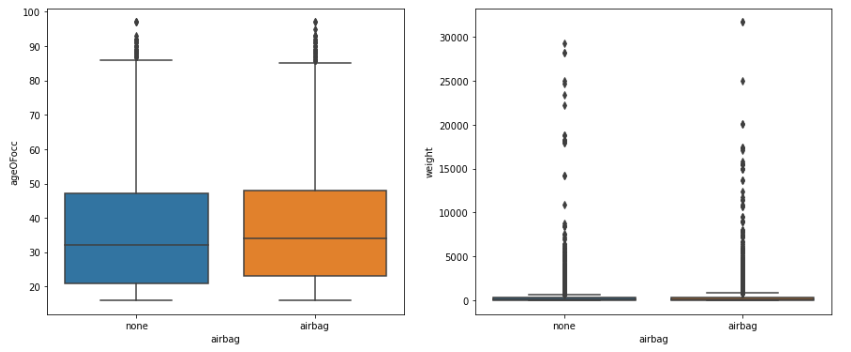
* Weight and ageofocc has an outlier we need to treat it
* The category columns don’t need any of the treatment for outliers

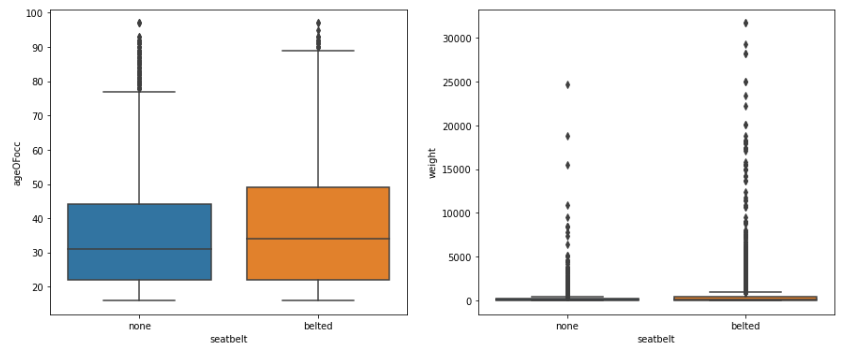
**BIVARIATE ANALYSIS:**

****

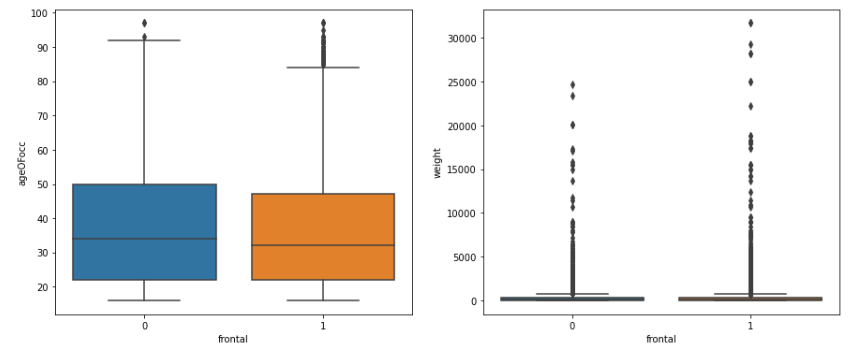
**Fig 2.1.10 Boxplots of survived vs ageofocc**

**Fig 2.1.11 Boxplots of dvcat vs ageofocc**

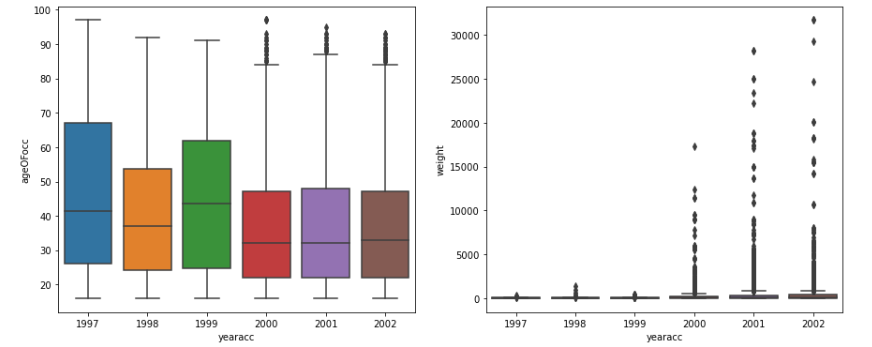
****

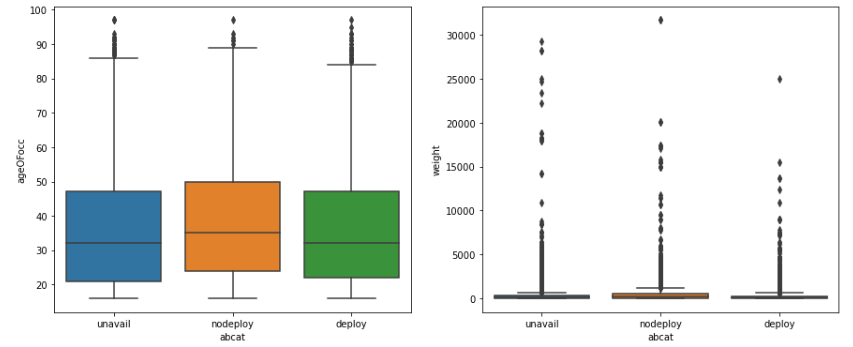
**Fig 2.1.12 Boxplots of airbag vs ageofocc**

**Fig 2.1.13 Boxplots of seatbelt vs ageofocc**

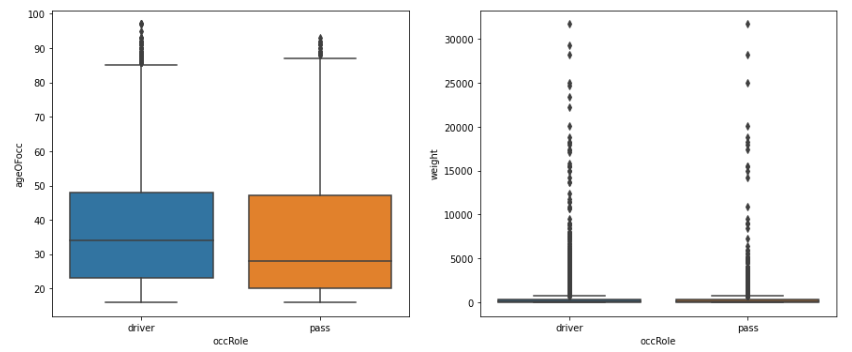
****

**Fig 2.1.14 Boxplots of frontal vs ageofocc**

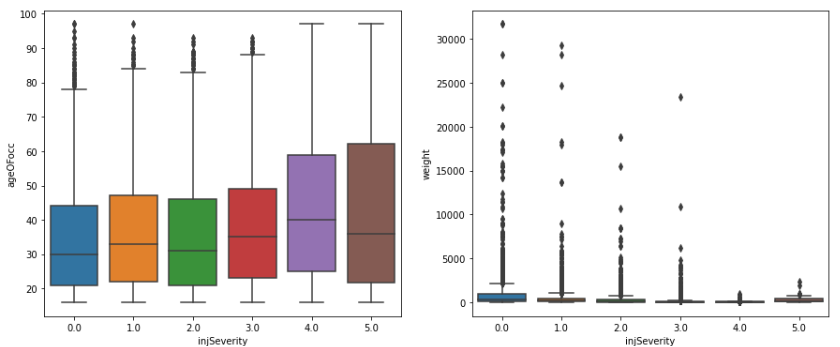
****

**Fig 2.1.15 Boxplots of yearacc vs ageofocc**

**Fig 2.1.16 Boxplots of abcat vs ageofocc**

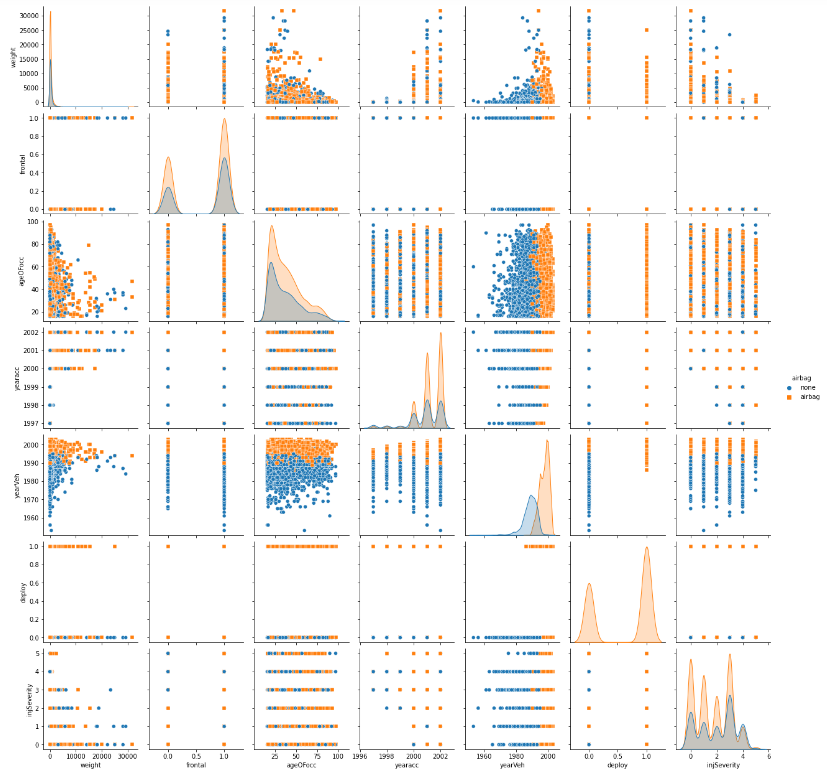
****

**Fig 2.1.17 Boxplots of driver vs ageofocc**

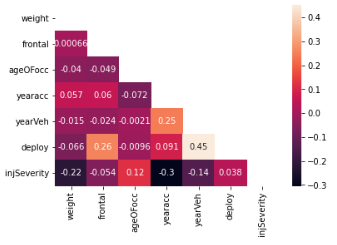
****

**Fig 2.1.18 Boxplots of injseverity vs ageofocc**

**PAIR PLOT:**

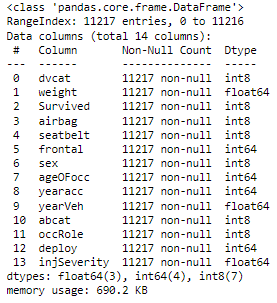
****

**Fig 2.1.19 Pair plot of survived as hue**

****

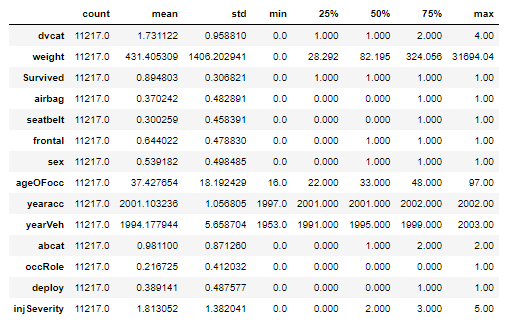
**Fig 2.1.20 Heat map plots for correlations**

**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).**

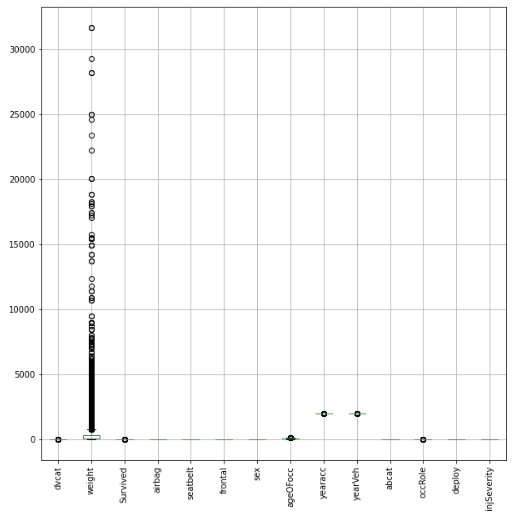
****

**Fig 2.2.1 Information of columns**

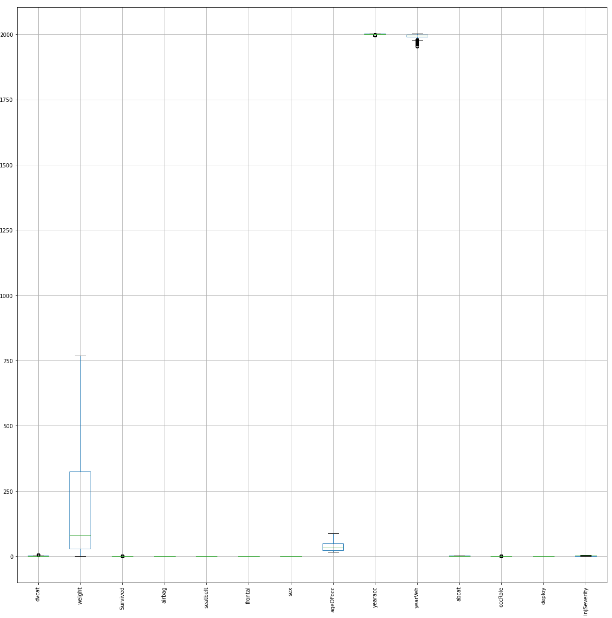
* We can observe that the categorical columns with objects are converted into integer
* The memory has also been minimized from 1mb to 690kb

****

**Fig 2.2.2 Description of each column**

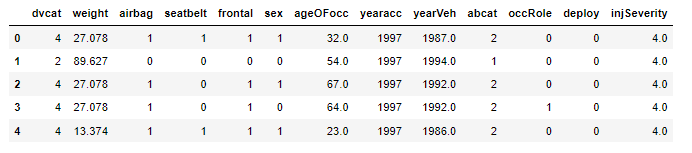
****

**Fig 2.2.3 Before outlier treatment**

****

**Fig 2.2.4 After outlier treatment**

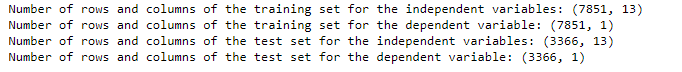
The data is splitted into 70:30 ratio and given to X and y

****

**Fig 2.2.5 X data**

****

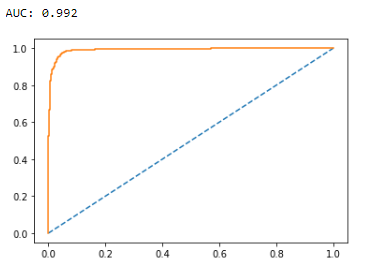
**Fig 2.2.6 Y data**

****

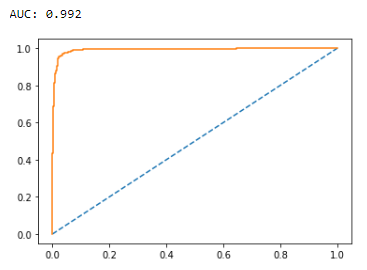
**Fig 2.2.7 splitting of data into train and test**

**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.**

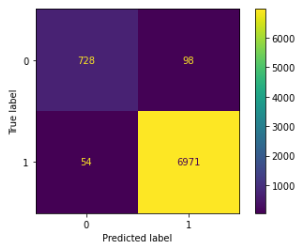
**Model 1 – logistic regression:**

****

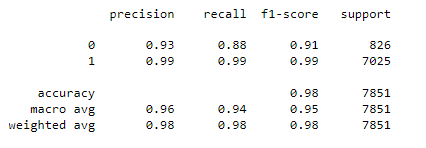
**Fig 2.3.1 Training AUC**

****

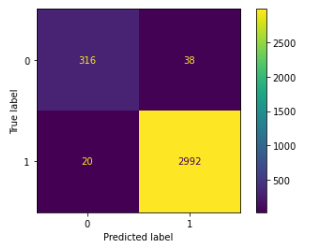
**Fig 2.3.2 Testing AUC**

****

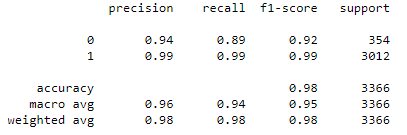
**Fig 2.3.3 Training Confusion matrix**

****

**Fig 2.3.4 Training Classification metrics**

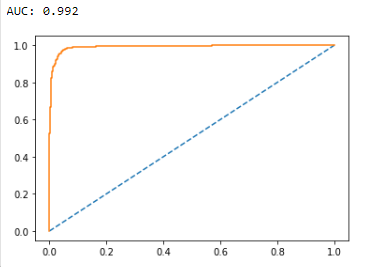
****

**Fig 2.3.5 Testing Confusion matrix**

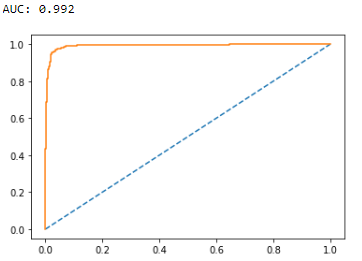
****

**Fig 2.3.6 Training Classification metrics**

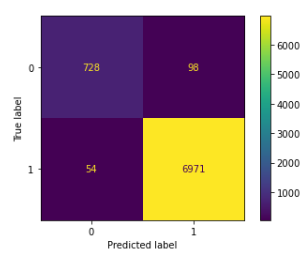
**Model 2 – Grid search:**

****

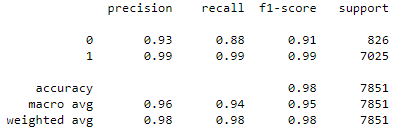
**Fig 2.3.7 Training AUC**

****

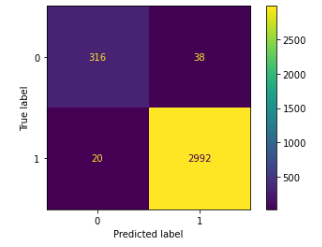
**Fig 2.3.8 Testing AUC**

****

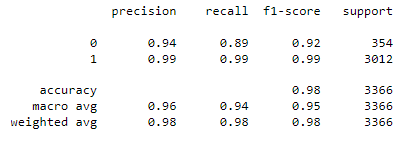
**Fig 2.3.9 Training Confusion matrix**

****

**Fig 2.3.10 Training Classification metrics**

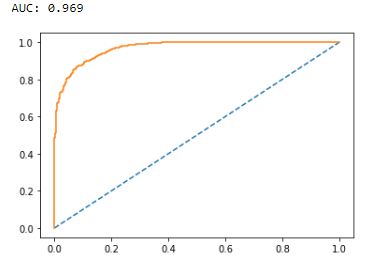
****

**Fig 2.3.11 Testing Confusion matrix**

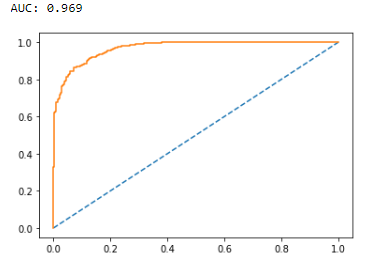
****

**Fig 2.3.12 Testing Classification metrics**

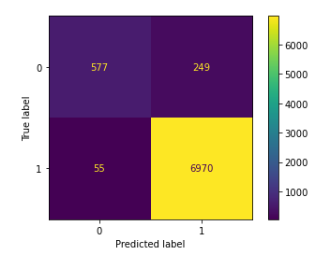
**Model 3 – Linear discriminant analysis:**

****

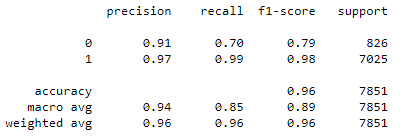
**Fig 2.3.13 Training AUC**

****

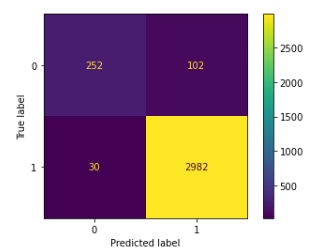
**Fig 2.3.14 Testing AUC**

****

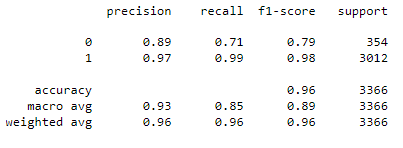
**Fig 2.3.15 Training Confusion matrix**

****

**Fig 2.3.16 Training Classification metrics**

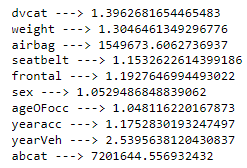
****

**Fig 2.3.17 Testing Confusion matrix**

****

**Fig 2.3.18 Testing Classification metrics**

**Variance Inflation factor:**

****

**Fig 2.3.19 vif for whole data**

* we can observe that there is some collinearity but there is not much by looking into results of vif we got that except 'airbag' and 'abcat' rest other columns are most important and collinearity.

1. logistic regression - training accuracy - 98%, testing accuracy - 98%
2. grid search - training accuracy - 98%, testing accuracy - 98%
3. Linear discriminant analysis - training accuracy - 96%, testing accuracy - 96%

**Inference**

\* logistic and grid search models performs at same level

\* The have same AUC for train and test 99.2% and 99.2%

\* LDA is less quite than other few models

\* AUC for this model is 96.9% and 96.9%

\* All the outputs are similar or closest for all the three models

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

All the 3 models performing good and same at this situation, we can use logistic and grid search.

**we have performed**

--> logistic regression and interpreted the results with 98% accuracy on the training data an 98% on testing data

--> grid search on logistic and interpreted the results with 98% accuracy on the training data an 98% on testing data,

--> Linear discriminant analysis and interpreted the results with 96% accuracy on the training data an 96% on testing data

**🡨-------------------------------------------------------------------------------------------------------------------------------🡪**