# BUSINESS REPORT

# PREDICTIVE MODELLING

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 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.

### 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.

Firm Dataset is shown Below

| | sales | capital | patents | randd | employment | sp500 | tobinq | value |
|---|---|---|---|---|---|---|---|---|
| 0 | 826.995050 | 161.603986 | 10 | 382.078247 | 2.306000 | no | 11.049511 | 1625.453755 |
| 1 | 407.753973 | 122.101012 | 2 | 0.000000 | 1.860000 | no | 0.844187 | 243.117082 |
| 2 | 8407.845588 | 6221.144614 | 138 | 3296.700439 | 49.659005 | yes | 5.205257 | 25865.233800 |
| 3 | 451.000010 | 266.899987 | 1 | 83.540161 | 3.071000 | no | 0.305221 | 63.024630 |
| 4 | 174.927981 | 140.124004 | 2 | 14.233637 | 1.947000 | no | 1.063300 | 67.406408 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 754 | 1253.900196 | 708.299935 | 32 | 412.936157 | 22.100002 | yes | 0.697454 | 267.119487 |
| 755 | 171.821025 | 73.666008 | 1 | 0.037735 | 1.684000 | no | NaN | 228.475701 |
| 756 | 202.726967 | 123.926991 | 13 | 74.861099 | 1.460000 | no | 5.229723 | 580.430741 |
| 757 | 785.687944 | 138.780992 | 6 | 0.621750 | 2.900000 | yes | 1.625398 | 309.938651 |
| 758 | 22.701999 | 14.244999 | 5 | 18.574360 | 0.197000 | no | 2.213070 | 18.940140 |

**Table 1.1- Firm level data**

**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.

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | sales | 759 non-null | float64 |
| 1 | capital | 759 non-null | float64 |
| 2 | patents | 759 non-null | int64 |
| 3 | randd | 759 non-null | float64 |
| 4 | employment | 759 non-null | float64 |
| 5 | sp500 | 759 non-null | object |
| 6 | tobinq | 738 non-null | float64 |
| 7 | value | 759 non-null | float64 |
| 8 | institutions | 759 non-null | float64 |

**Table 1.2 - Datatypes in the Dataset and Null check**

dtypes: float64(7), int64(1), object(1)

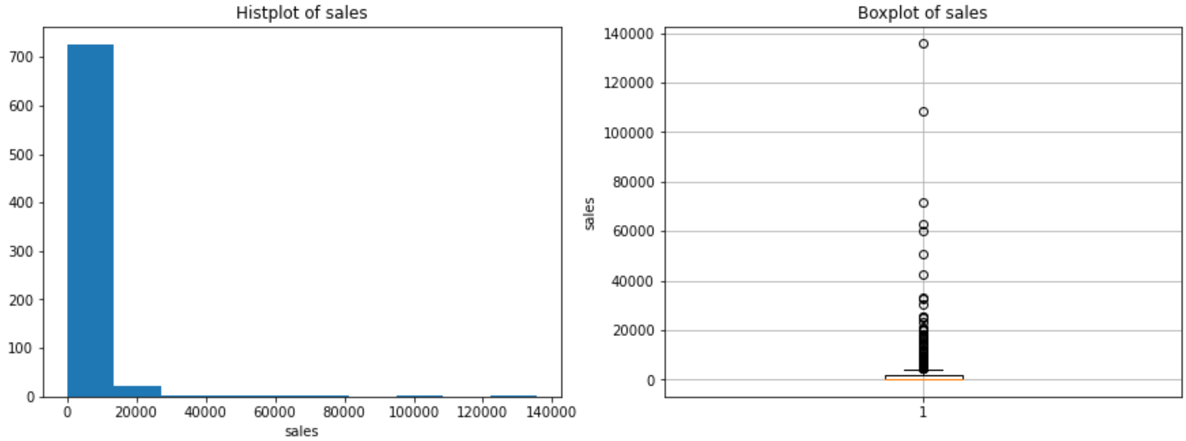
- We observe that tobinq has missing Data which will be imputed later on.
- We shall be predicting Sales which will be our dependent variable and rest will be independent variable

| | sales | capital | patents | randd | employment | tobinq | value |
|---|---|---|---|---|---|---|---|
| count | 759.000000 | 759.000000 | 759.000000 | 759.000000 | 759.000000 | 738.000000 | 759.000000 |
| mean | 2689.705158 | 1977.747498 | 25.831357 | 439.938074 | 14.164519 | 2.794910 | 2732.73475( |
| std | 8722.060124 | 6466.704896 | 97.259577 | 2007.397588 | 43.321443 | 3.366591 | 7071.072362 |
| min | 0.138000 | 0.057000 | 0.000000 | 0.000000 | 0.006000 | 0.119001 | 1.971053 |
| 25% | 122.920000 | 52.650501 | 1.000000 | 4.628262 | 0.927500 | 1.018783 | 103.593946 |
| 50% | 448.577082 | 202.179023 | 3.000000 | 36.864136 | 2.924000 | 1.680303 | 410.793529 |
| 75% | 1822.547366 | 1075.790020 | 11.500000 | 143.253403 | 10.050001 | 3.139309 | 2054.16038! |
| max | 135696.788200 | 93625.200560 | 1220.000000 | 30425.255860 | 710.799925 | 20.000000 | 95191.5911( |

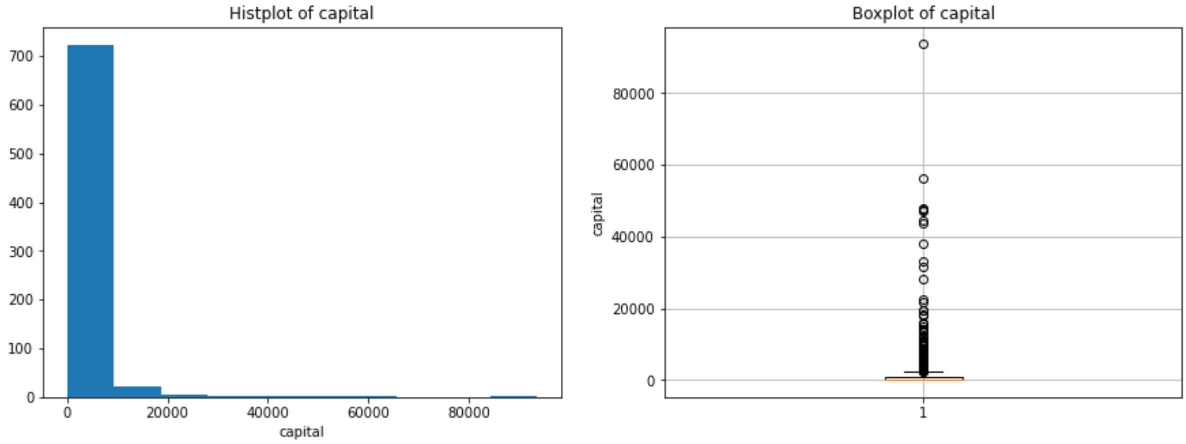
**Table 1.3 - Data Description of Firm Level Data**

# Univariate Analysis

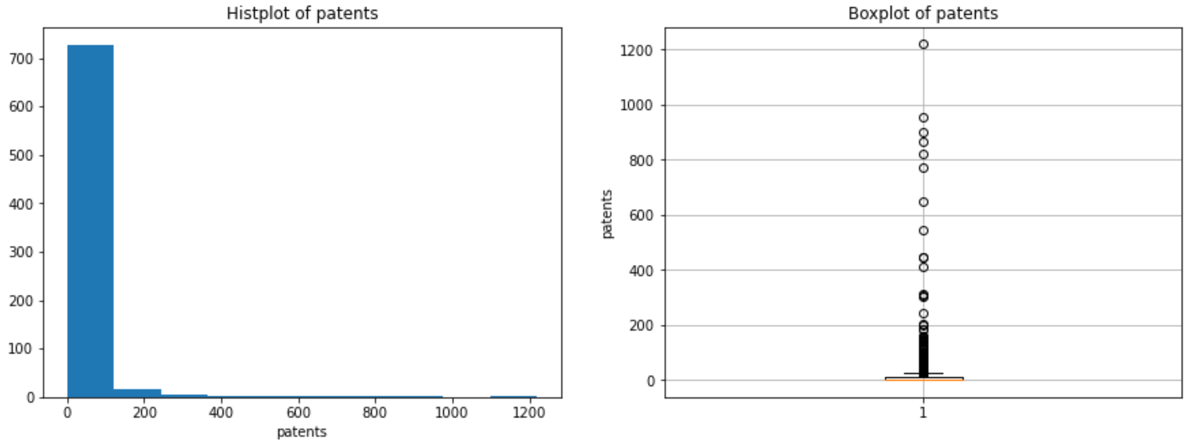
### Plot of Histogram and Boxplot of each attribute is as shown below



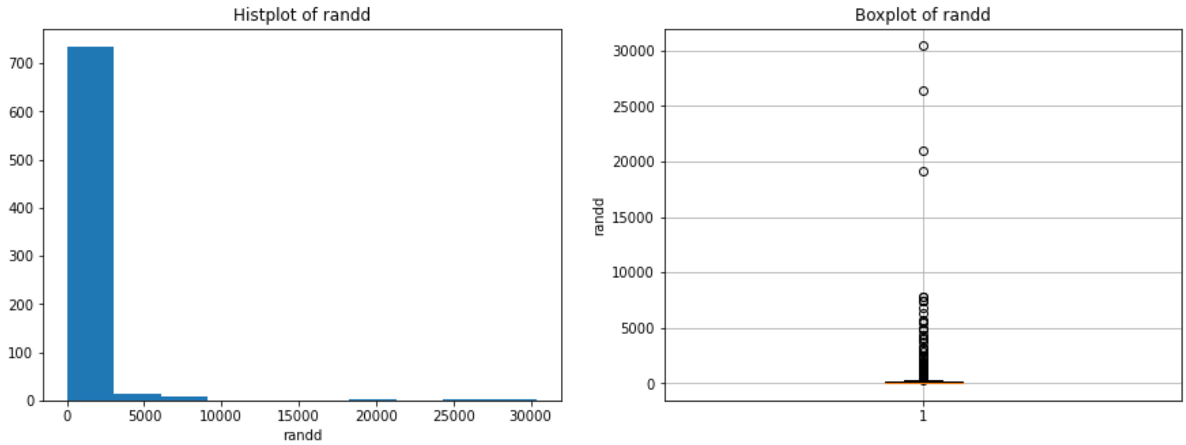

## Fig 1.1 - Boxplot and Histplot of Sales




## Fig 1.2 - Boxplot and Histplot of Capital




## Fig 1.3 - Boxplot and Histplot of Patents




## Fig 1.4 - Boxplot and Histplot of R&D Stock

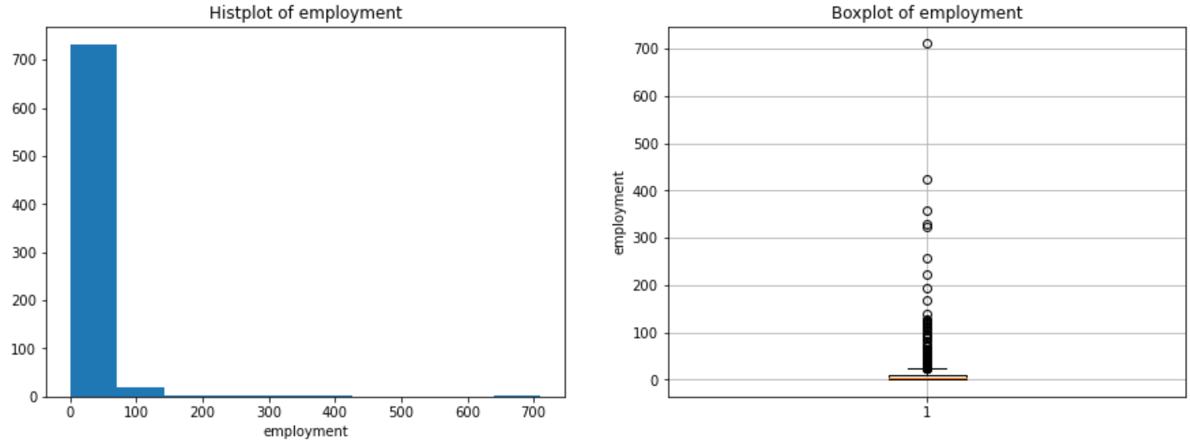



**Fig 1.5 - Boxplot and Histplot of Employment**

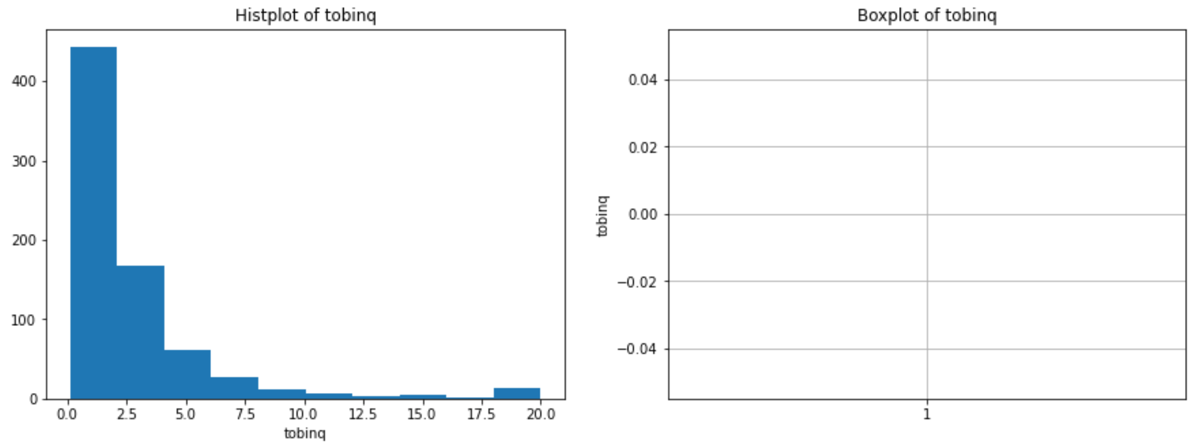



**Fig 1.6 - Boxplot and Histplot of Tobin's q**

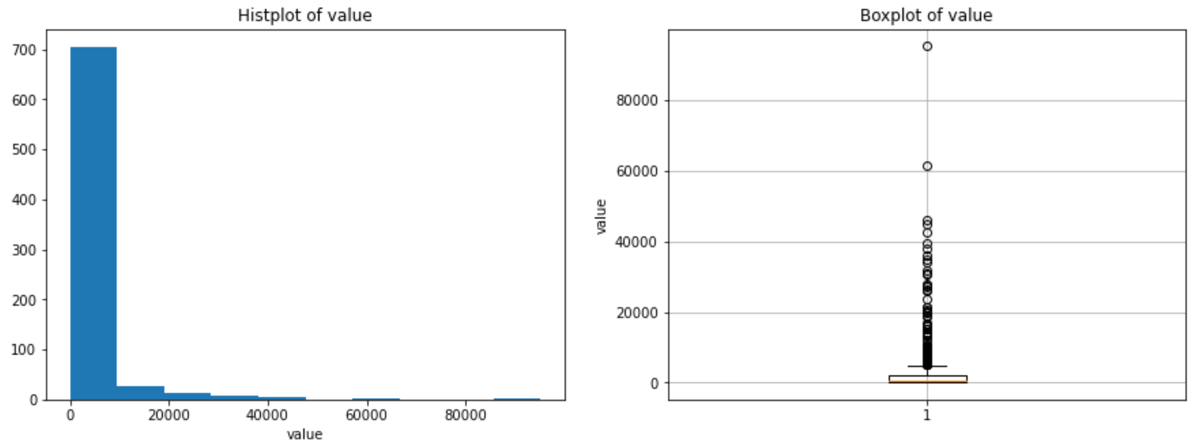



**Fig 1.7 - Boxplot and Histplot of Stock market value**

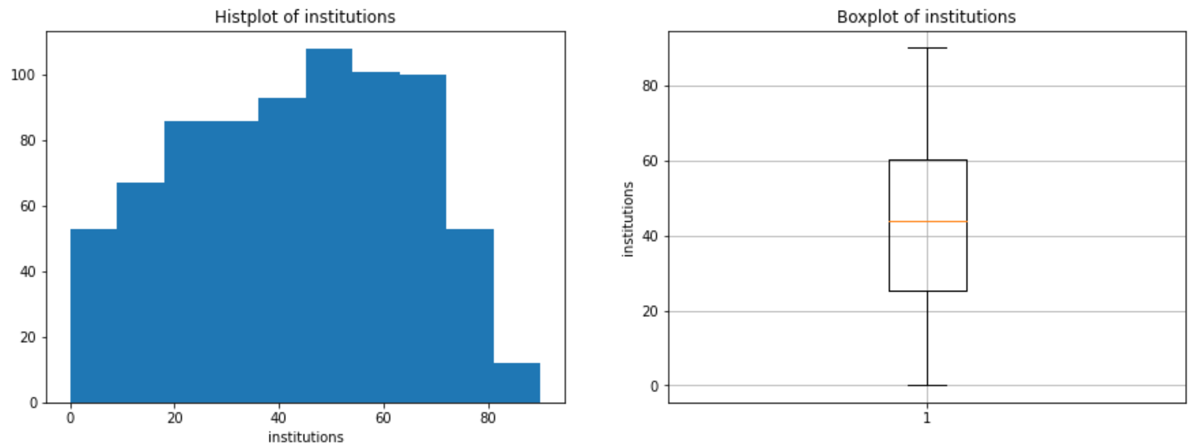



**Fig 1.8 - Boxplot and Histplot of Proportion of stock owned by institutions**

- We see that all the numeric columns except Proportion of stock owned by institutions are all Right Skewed
- Proportion of stock owned by institutions is slightly Left skewed but has no Outliers

# Bivariate Analysis

### Heat Map and Pair Plot of the Data is as shown below

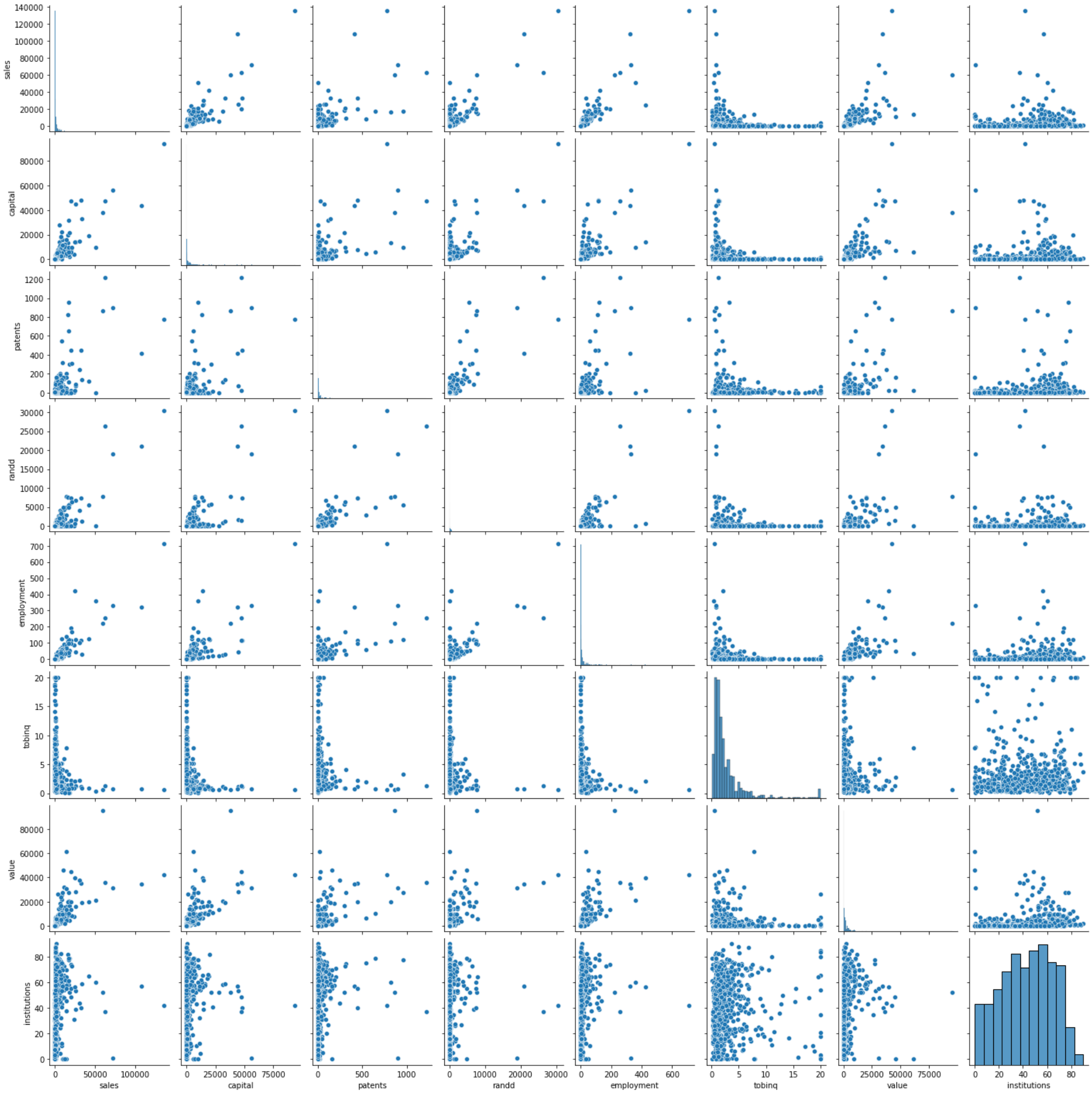



**Fig 1.9 - Pairplot of Firm Dataset**

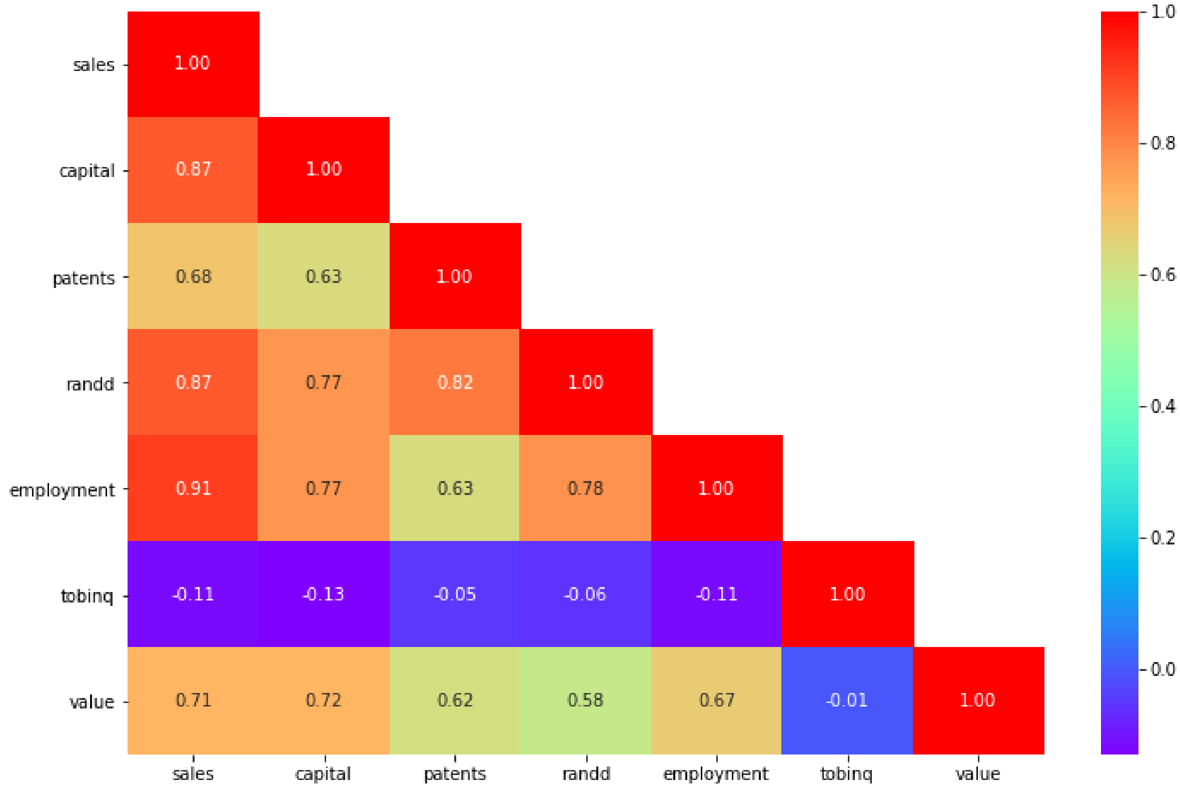



**Fig 1.10 - Heatmap of Firm Dataset showing correlation**

- We see that Tobin q has no very minute correlation with any of the attributes
- Sales which is the dependant variable is highly correlated with Capital, R&D Stock, Employment And Moderate-to-High Correlated with Patents and Stock Market Value

## 1.2) Impute null values if present? Do you think scaling is necessary in this case?

Null Value Check

| Columns | Null Value present |
|---|---|
| sales | 0 |
| capital | 0 |
| patents | 0 |
| randd | 0 |
| employment | 0 |
| sp500 | 0 |
| tobinq | 21 |
| value | 0 |
| institutions | 0 |

**Table 1.4 - Null values present in data set of Investment Firm**

**Scaling**

- Feature Scaling is not required in Linear Regression models but is used depending on the training Algorithm that is being considered

- If Normal Equation is being implemented then there is No stepwise Optimization process hence feature scaling is not necessary.
- However when Gradient descent Algorithm is used, Scaling is recommended, otherwise the algorithm might take much longer to converge.

## 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.

Only one Categorical Variable - sp500 (Membership of firms in the S&P 500 index)

| sales | capital | patents | randd | employment | tobinq | value | institutions | sp500 | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 826.995050 | 161.603986 | 10.0 | 382.078247 | 2.306000 | 11.049511 | 1625.453755 | 80.27 | 0 |
| 1 | 407.753973 | 122.101012 | 2.0 | 0.000000 | 1.860000 | 0.844187 | 243.117082 | 59.02 | 0 |
| 2 | 8407.845588 | 6221.144614 | 138.0 | 3296.700439 | 49.659005 | 5.205257 | 25865.233800 | 47.70 | 1 |
| 3 | 451.000010 | 266.899987 | 1.0 | 83.540161 | 3.071000 | 0.305221 | 63.024630 | 26.88 | 0 |
| 4 | 174.927981 | 140.124004 | 2.0 | 14.233637 | 1.947000 | 1.063300 | 67.406408 | 49.46 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 754 | 1253.900196 | 708.299935 | 32.0 | 412.936157 | 22.100002 | 0.697454 | 267.119487 | 33.50 | 1 |
| 755 | 171.821025 | 73.666008 | 1.0 | 0.037735 | 1.684000 | 1.680303 | 228.475701 | 46.41 | 0 |
| 756 | 202.726967 | 123.926991 | 13.0 | 74.861099 | 1.460000 | 5.229723 | 580.430741 | 42.25 | 0 |
| 757 | 785.687944 | 138.780992 | 6.0 | 0.621750 | 2.900000 | 1.625398 | 309.938651 | 61.39 | 1 |
| 758 | 22.701999 | 14.244999 | 5.0 | 18.574360 | 0.197000 | 2.213070 | 18.940140 | 7.50 | 0 |

**Table 1.5 - encoded Data set of Firm**

#### Spliting Data into 70:30

### Performing Simple Linear Regression using OLS model

| | | | |
|---|---|---|---|
| Dep. Variable: | sales | R-squared: | 0.936 |
| Model: | OLS | Adj. R-squared: | 0.935 |
| Method: | Least Squares | F-statistic: | 960.3 |
| Date: | Tue, 22 Feb 2022 | Prob (F-statistic): | 1.37e-306 |
| Time: | 01:57:37 | Log-Likelihood: | -4831.5 |
| No. Observations: | 531 | AIC: | 9681. |
| Df Residuals: | 522 | BIC: | 9719. |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |
| Omnibus: | 231.591 | Durbin-Watson: | 1.932 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 31508.283 |
| Skew: | 0.809 | Prob(JB): | 0.00 |
| Kurtosis: | 40.703 | Cond. No. | 2.90e+04 |

| | coef | std err | t | P>(t) | 0.025 | 0.975 |
|---|---|---|---|---|---|---|
| const | 52.8005 | 233.075 | 0.227 | 0.821 | -405.080 | 510.681 |
| capital | 0.4142 | 0.027 | 15.565 | 0.000 | 0.362 | 0.467 |
| patents | -5.0452 | 2.407 | -2.096 | 0.037 | -9.774 | -0.317 |
| randd | 1.0261 | 0.127 | 8.052 | 0.000 | 0.776 | 1.276 |
| employment | 83.9581 | 3.629 | 23.136 | 0.000 | 76.829 | 91.087 |
| tobinq | -31.4063 | 30.227 | -1.039 | 0.299 | -90.787 | 27.975 |
| value | 0.1267 | 0.022 | 5.886 | 0.000 | 0.084 | 0.169 |
| institutions | 1.0555 | 4.964 | 0.213 | 0.832 | -8.697 | 10.808 |
| sp500 | -100.4375 | 267.857 | -0.375 | 0.708 | -626.648 | 425.773 |

**Table 1.6 - OLS Table**

**Below are some of the Observations of Linear Refression Model**

1. The variation in the independent variable which is explained by the dependent variable is 93.6378 %
2. The Root Mean Square Error (RMSE) of the model is for the training set is 2164.4938172647676
3. The Root Mean Square Error (RMSE) of the model is for testing set is 2953.569036057085
4. The coefficient of determination R^2 of the prediction on Train set 0.9363784533904187
5. The coefficient of determination R^2 of the prediction on Test set 0.892768228595857
6. The Root Mean Square Error (RMSE) of the model is for testing set is 2953.5690360571057
7. VIF Scores

- capital ---> 4.119837852562865
- patents ---> 3.8852813007814486
- randd ---> 6.229053691145338
- employment ---> 3.888358497441767
- tobinq ---> 1.5665491224775512
- v-alue ---> 3.241413503399095
- institutions ---> 2.4474660981670664
- sp500 ---> 2.220617734070768

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

- The RMSE value tells us that the average deviation between the predicted sales made by the model and the actual sales is 2953.56 (Test Data)
- We can Observe from the above predictions and calculations that RMSE for Training and testing data sets are moderately low. which indicates that Model Predictions are Quiet Good when compared to Actual Observations.
- Coefficient of Determination R^2 on Test data is 0.8927 which tells us that the predictor variables explain about 89% of the variance in the response variable.
- F statistic has a very low p value (practically low) Meaning that the model fit is statistically significant, and the explained variance isn't purely by chance.

- We can also observe that VIF(Variance inflation factor) is low which indicates that multicollinearity cease to exist.
- Important Attributes to sales are Employment, Capital & Patents.

# 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.

### 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.

**Data Set of Car Crash is as shown below**

| | dvcat | weight | Survived | airbag | seatbelt | frontal | sex | ageOFocc | yearacc | yearVeh | a |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 55+ | 27.078 | Not_Survived | none | none | 1 | m | 32 | 1997 | 1987.0 | un |
| 1 | 25-39 | 89.627 | Not_Survived | airbag | belted | 0 | f | 54 | 1997 | 1994.0 | nc |
| 2 | 55+ | 27.078 | Not_Survived | none | belted | 1 | m | 67 | 1997 | 1992.0 | un |
| 3 | 55+ | 27.078 | Not_Survived | none | belted | 1 | f | 64 | 1997 | 1992.0 | un |
| 4 | 55+ | 13.374 | Not_Survived | none | none | 1 | m | 23 | 1997 | 1986.0 | un |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 11212 | 25-39 | 3179.688 | survived | none | belted | 1 | m | 17 | 2002 | 1985.0 | un |
| 11213 | 10-24 | 71.228 | survived | airbag | belted | 1 | m | 54 | 2002 | 2002.0 | nc |
| 11214 | 10-24 | 10.474 | survived | airbag | belted | 1 | f | 27 | 2002 | 1990.0 | de |
| 11215 | 25-39 | 10.474 | survived | airbag | belted | 1 | f | 18 | 2002 | 1999.0 | de |
| 11216 | 25-39 | 10.474 | survived | airbag | belted | 1 | m | 17 | 2002 | 1999.0 | de |

**Table 2.1 - Car Crash Dataset**

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | dvcat | 11217 non-null | object |
| 1 | weight | 11217 non-null | float64 |
| 2 | Survived | 11217 non-null | object |
| 3 | airbag | 11217 non-null | object |
| 4 | seatbelt | 11217 non-null | object |
| 5 | frontal | 11217 non-null | int64 |
| 6 | sex | 11217 non-null | object |
| 7 | ageOFocc | 11217 non-null | int64 |

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 8 | yearacc | 11217 non-null | int64 |
| 9 | yearVeh | 11217 non-null | float64 |
| 10 | abcat | 11217 non-null | object |
| 11 | occRole | 11217 non-null | object |
| 12 | deploy | 11217 non-null | int64 |
| 13 | injSeverity | 11140 non-null | float64 |
| 14 | caseid | 11217 non-null | object |

### Table 2.2 - Datatypes of each columns

| | weight | frontal | ageOFocc | yearacc | yearVeh | deploy | injSever |
|---|---|---|---|---|---|---|---|
| count | 11217.000000 | 11217.000000 | 11217.000000 | 11217.000000 | 11217.000000 | 11217.000000 | 11140.00( |
| mean | 431.405309 | 0.644022 | 37.427654 | 2001.103236 | 1994.177944 | 0.389141 | 1.825583 |
| std | 1406.202941 | 0.478830 | 18.192429 | 1.056805 | 5.658704 | 0.487577 | 1.378535 |
| min | 0.000000 | 0.000000 | 16.000000 | 1997.000000 | 1953.000000 | 0.000000 | 0.000000 |
| 25% | 28.292000 | 0.000000 | 22.000000 | 2001.000000 | 1991.000000 | 0.000000 | 1.000000 |
| 50% | 82.195000 | 1.000000 | 33.000000 | 2001.000000 | 1995.000000 | 0.000000 | 2.000000 |
| 75% | 324.056000 | 1.000000 | 48.000000 | 2002.000000 | 1999.000000 | 1.000000 | 3.000000 |
| max | 31694.040000 | 1.000000 | 97.000000 | 2002.000000 | 2003.000000 | 1.000000 | 5.000000 |

### Table 2.3 - Car Crash Data Description

| Columns | Null Check |
|---|---|
| dvcat | 0 |
| weight | 0 |
| Survived | 0 |
| airbag | 0 |
| seatbelt | 0 |
| frontal | 0 |
| sex | 0 |
| ageOFocc | 0 |
| yearacc | 0 |
| yearVeh | 0 |
| abcat | 0 |
| occRole | 0 |
| deploy | 0 |
| injSeverity | 77 |
| caseid | 0 |

**Table 2.4 - Null Check on Car Crash Dataset**

**From the above table we can see that injSeverity has missing values. 'injSeverity' columns contains 5 levels so we impute the NaNs with their respective Modal Values.**

| Columns | Null Check |
|---|---|
| dvcat | 0 |
| weight | 0 |
| Survived | 0 |
| airbag | 0 |
| seatbelt | 0 |
| frontal | 0 |
| sex | 0 |
| ageOFocc | 0 |
| yearacc | 0 |
| yearVeh | 0 |
| abcat | 0 |
| occRole | 0 |
| deploy | 0 |
| injSeverity | 0 |
| caseid | 0 |

**Table 2.5 - Null Check on Car Crash Dataset after iumputing**

**Some Observations from the Dataset**

1. Since there are some columns that have different data set we convert these columns into 'Object' datatype. Some of them are

- 'injSeverity' - a numeric vector; 0: none, 1: possible injury, 2: no incapacity, 3:incapacity, 4: killed; 5: unknown, 6: prior death
- 'deploy' - has 2 categories ; 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed,
- 'frontal' - a numeric vector; 0 = non-frontal, 1=frontal impact
- 'yearacc' - year of Accident ; 1997, 1998, 1999, 2000, 2001, 2002

1. There are no Duplicate rows.
2. From the Dataset, Survived is 89.48% and Not Survived are 10.52%
3. Once the Columns are converted, we have only 3 columns with numeric datatype

Univariant Analysis

**Histogram plot of Weight, Age of Occupation and Year of model of vehicle**

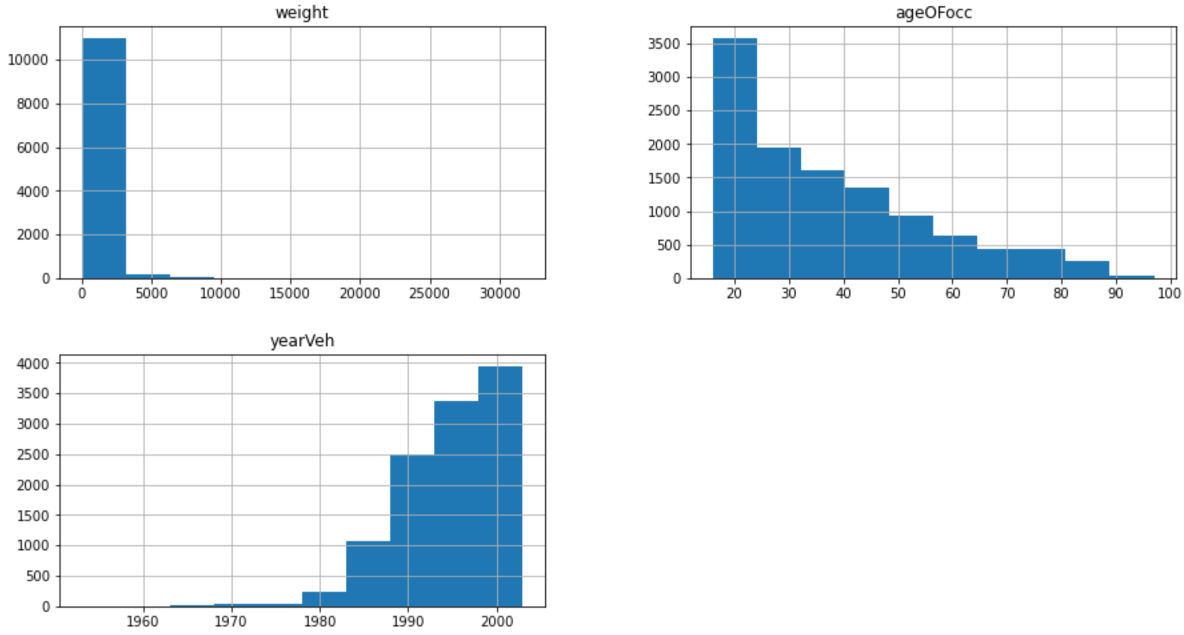



**Fig 2.1 - Histogram of Weight, Age of Occupation and Year of model of vehicle**

- We can observe that Most of the Car weigh between 0 and 5000
- From the Age of occupant graph we see that most of the Occupant involved are between the age 20 to 40 and gradually the count decresses
- From the Year of the Model of Vehcile involved in Car Crash, we can observe that most of the Vehicles are between 1990 & 2000 year Model

**Boxplot of Weight, Age of Occupation and Year of model of vehicle**

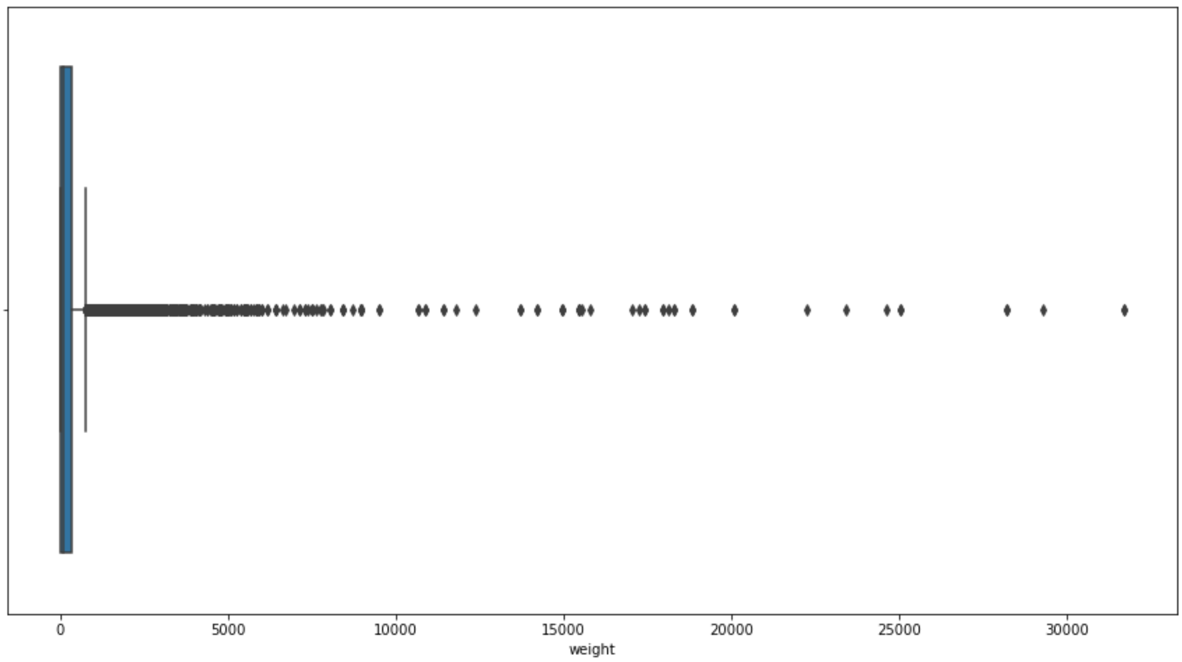



**Fig 2.2 - Boxplot of Weight**

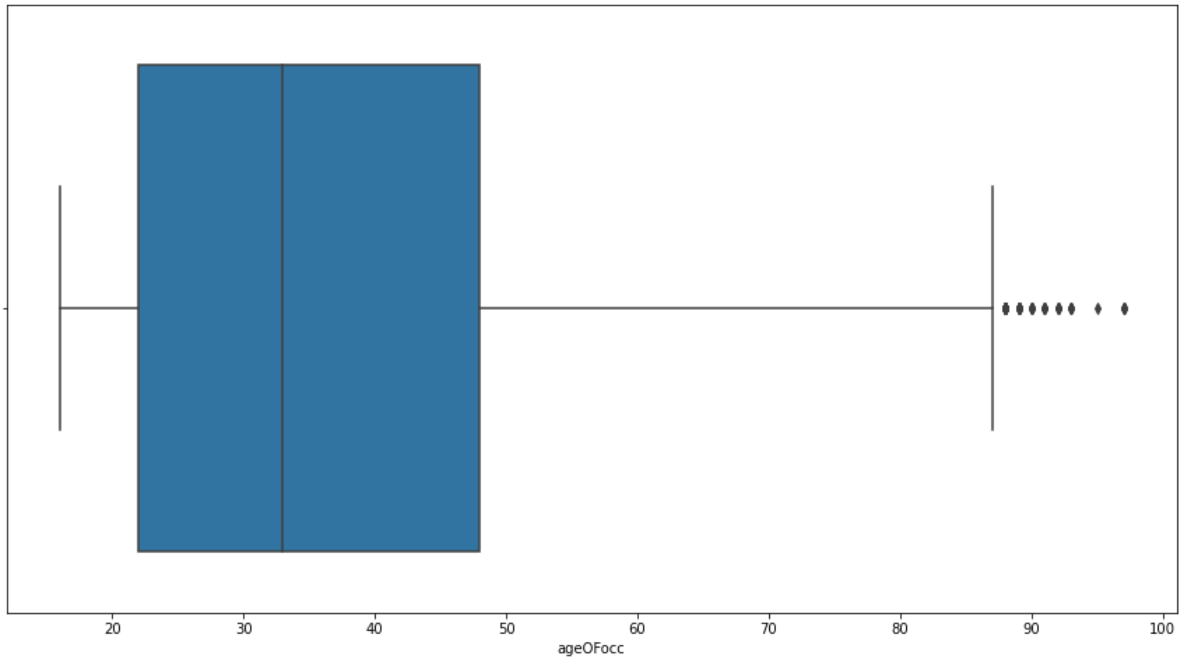



**Fig 2.3 - Boxplot of Age of Occupation**

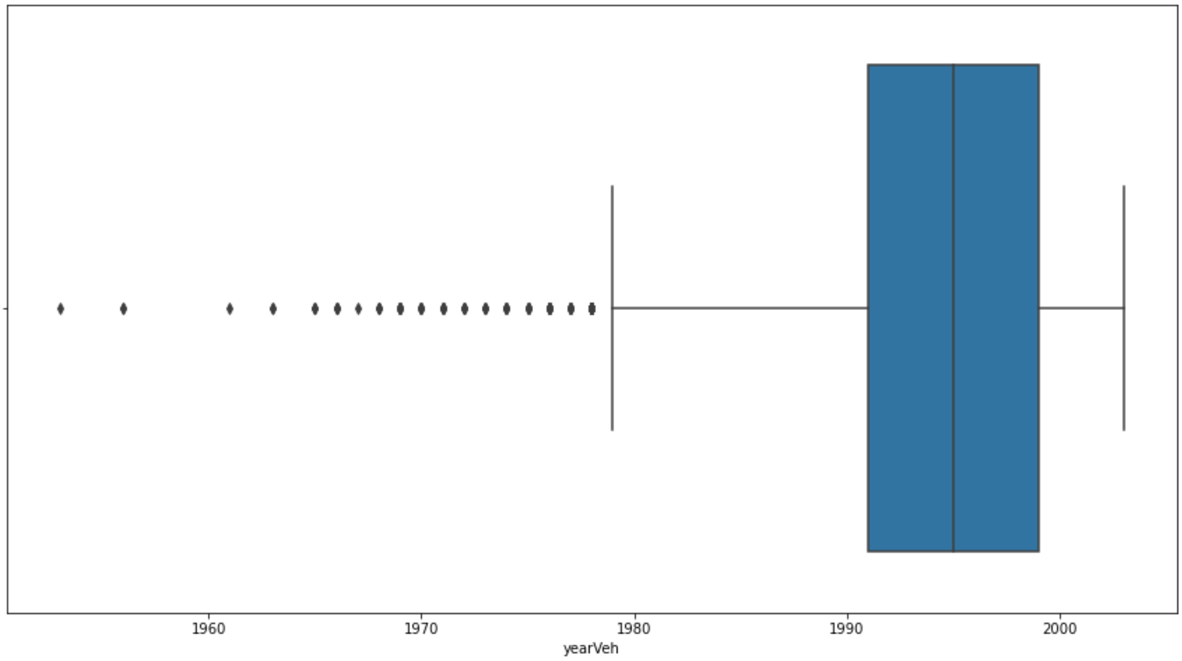



**Fig 2.4 - Boxplot of Year of model of vehicle**

## Bivariant Analysis

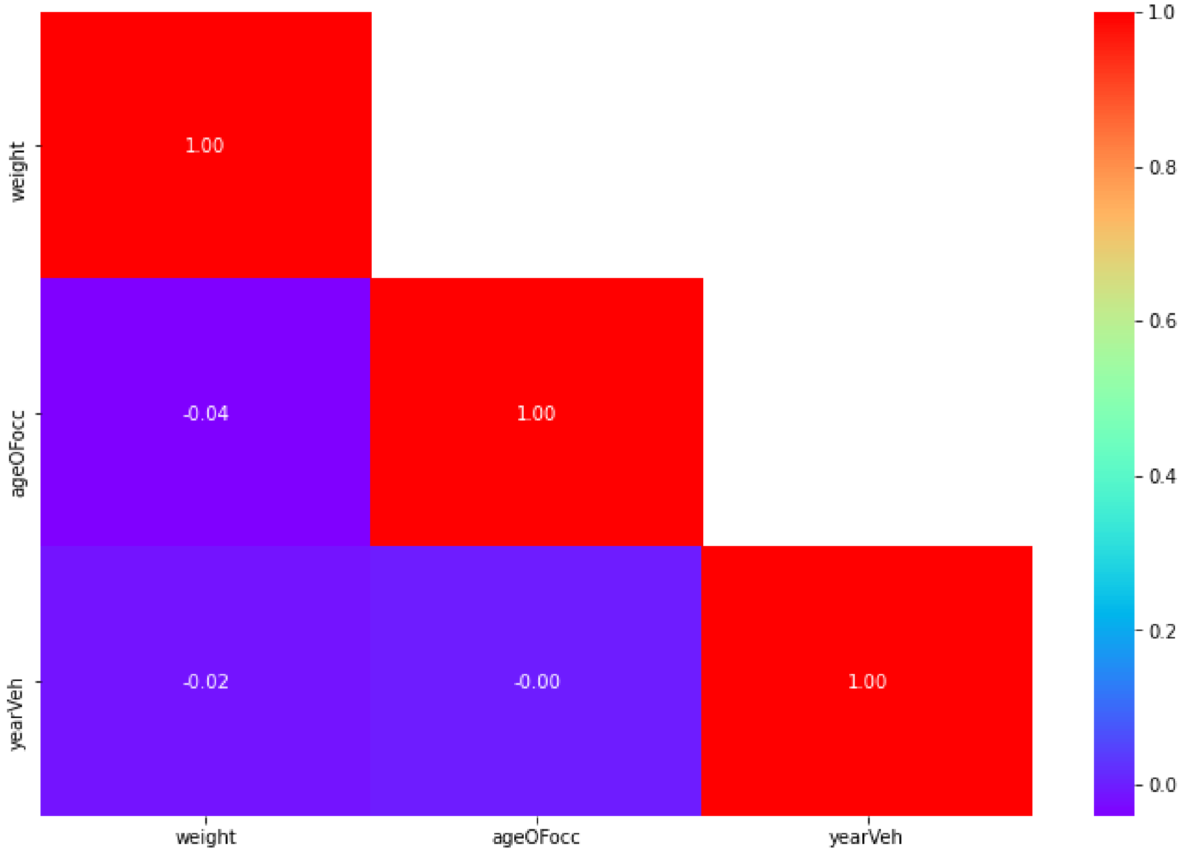



**Fig 2.5 - Heatmap of Car Crash(Weight, Age of Occupation and Year of model of vehicle)**

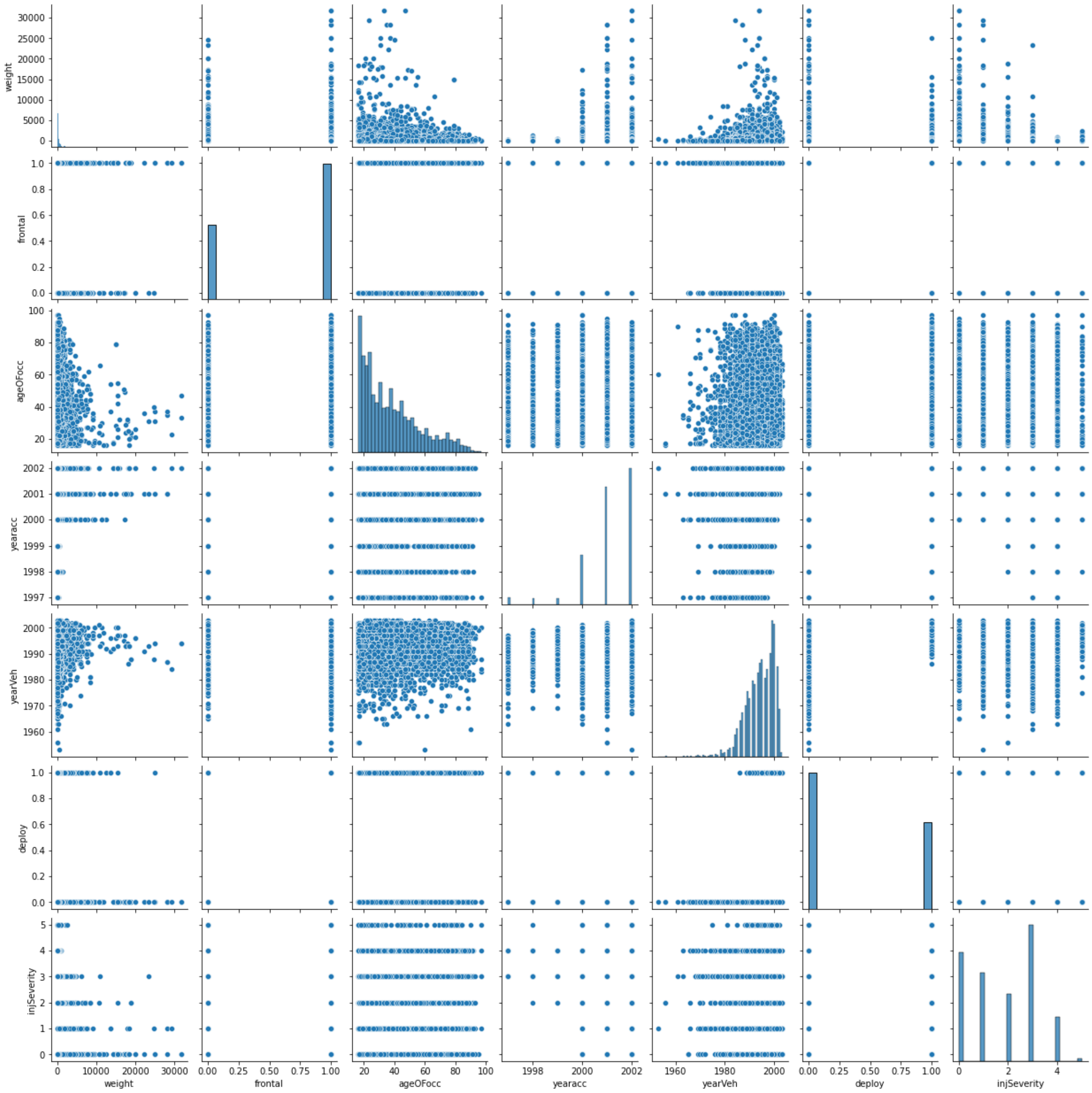


**Fig 2.6 - Pair Plot of Crash Crash Dataset**

## 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).

## &

## 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.

| | dvcat | Survived | airbag | seatbelt | frontal | sex | yearacc | abcat | occRole | deploy | injSe |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 55+ | Not_Survived | none | none | 1 | m | 1997 | unavail | driver | 0 | 4 |
| 1 | 25-39 | Not_Survived | airbag | belted | 0 | f | 1997 | nodeploy | driver | 0 | 4 |
| 2 | 55+ | Not_Survived | none | belted | 1 | m | 1997 | unavail | driver | 0 | 4 |
| 3 | 55+ | Not_Survived | none | belted | 1 | f | 1997 | unavail | pass | 0 | 4 |
| 4 | 55+ | Not_Survived | none | none | 1 | m | 1997 | unavail | driver | 0 | 4 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

| | dvcat | Survived | airbag | seatbelt | frontal | sex | yearacc | abcat | occRole | deploy | injSe |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 11212 | 25-39 | survived | none | belted | 1 | m | 2002 | unavail | driver | 0 | 0 |
| 11213 | 10-24 | survived | airbag | belted | 1 | m | 2002 | nodeploy | driver | 0 | 2 |
| 11214 | 10-24 | survived | airbag | belted | 1 | f | 2002 | deploy | driver | 1 | 3 |
| 11215 | 25-39 | survived | airbag | belted | 1 | f | 2002 | deploy | driver | 1 | 0 |
| 11216 | 25-39 | survived | airbag | belted | 1 | m | 2002 | deploy | pass | 1 | 0 |

**Table 2.6 - Categorical columns of Car Crash Dataset**

- From the above table which is a subset of dataframe crash, we have obtained colunns of object type.
- survived cane be one hot encoded, seatbelt can be onehot encoded, abcat can be one hot encoded, occRule can be one hot encoded, sex can be on hot encoded.
- caseid column will be drop as it is unique for each rows and doesn't help much in predictions

| | weight | ageOFocc | yearVeh | dvcat | Survived | frontal | yearacc | deploy | injSeverity | airbag_n |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 27.078 | 32 | 1987.0 | 4 | 0 | 1 | 0 | 0 | 4 | 1 |
| 1 | 89.627 | 54 | 1994.0 | 2 | 0 | 0 | 0 | 0 | 4 | 0 |
| 2 | 27.078 | 67 | 1992.0 | 4 | 0 | 1 | 0 | 0 | 4 | 1 |
| 3 | 27.078 | 64 | 1992.0 | 4 | 0 | 1 | 0 | 0 | 4 | 1 |
| 4 | 13.374 | 23 | 1986.0 | 4 | 0 | 1 | 0 | 0 | 4 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 11212 | 3179.688 | 17 | 1985.0 | 2 | 1 | 1 | 5 | 0 | 0 | 1 |
| 11213 | 71.228 | 54 | 2002.0 | 1 | 1 | 1 | 5 | 0 | 2 | 0 |
| 11214 | 10.474 | 27 | 1990.0 | 1 | 1 | 1 | 5 | 1 | 3 | 0 |
| 11215 | 10.474 | 18 | 1999.0 | 2 | 1 | 1 | 5 | 1 | 0 | 0 |
| 11216 | 10.474 | 17 | 1999.0 | 2 | 1 | 1 | 5 | 1 | 0 | 0 |

**Table 2.7 - Encoded Car Crash Dataset ready for Model Building**

**Data is Split in Training and Testing in the Ratio of 70:30**

**Logistic Regression is now performed on Dataset**

| 0 | 1 | |
|---|---|---|
| 0 | 0.020437 | 0.979563 |
| 1 | 0.001929 | 0.998071 |
| 2 | 0.002533 | 0.997467 |
| 3 | 0.001698 | 0.998302 |
| 4 | 0.013546 | 0.986454 |

**Table 2.8 - 1st 5 Rows of Predictions on test set**

- Model Score=0.9802572920647051

AUC and ROC for Training data for Logistic Regression

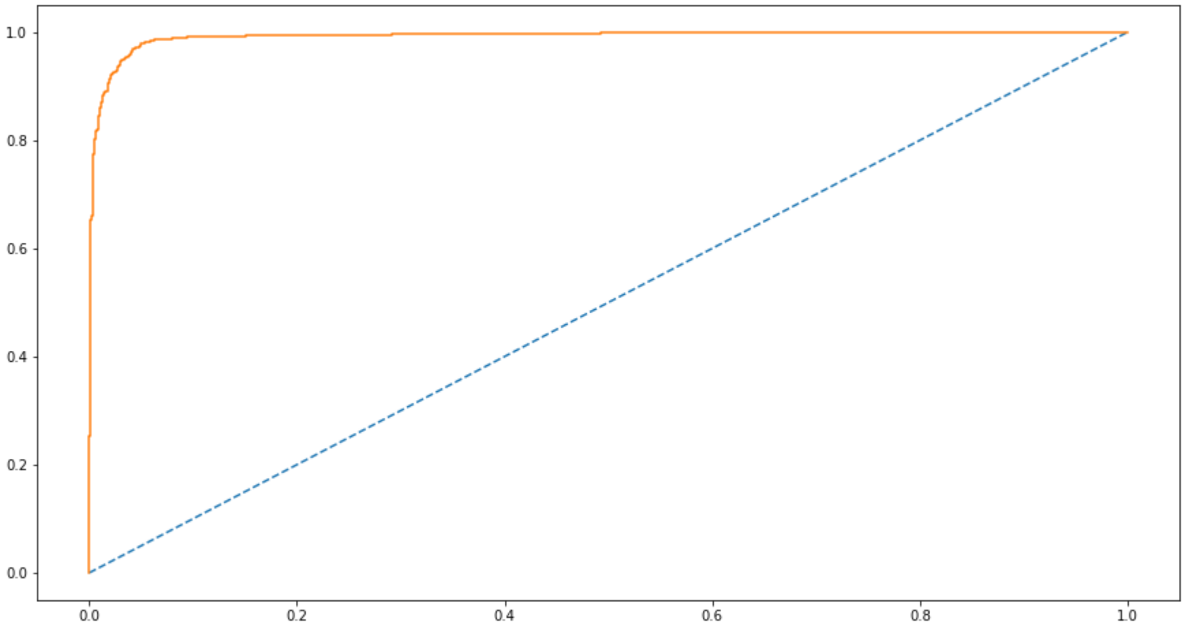



**Fig 2.7 - ROC Curve for Training data for Logistic Regression**

**ROC_AUC Score = 0.991**

AUC and ROC for Test data for Logistic Regression

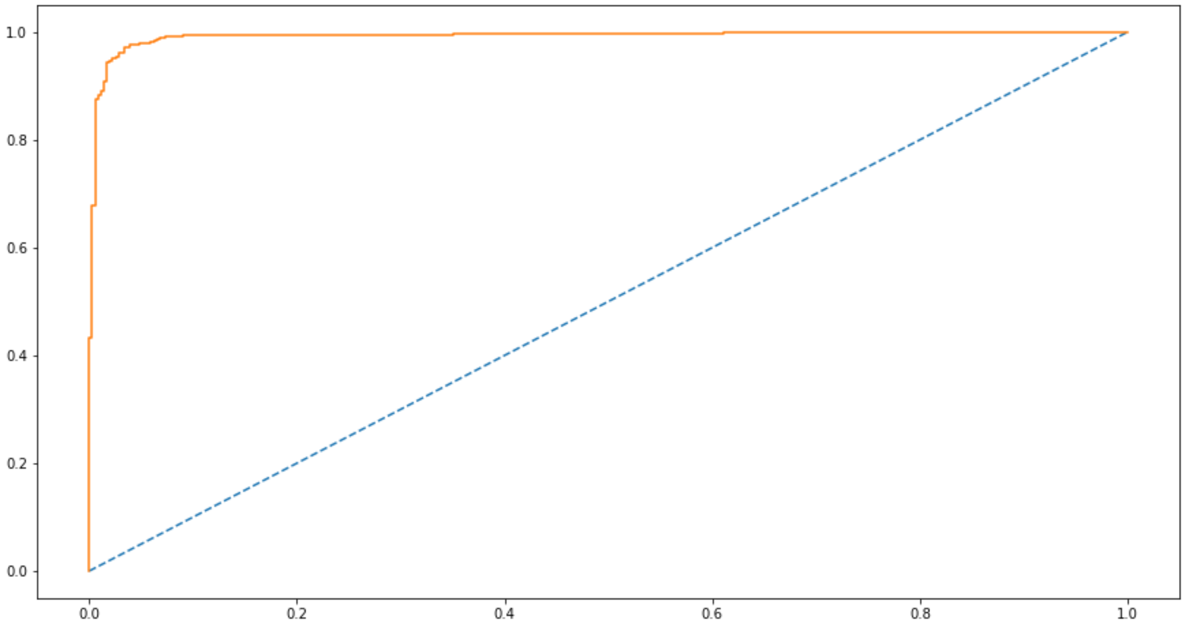



**Fig 2.7 - ROC Curve for Testing data for Logistic Regression**

**ROC_AUC Score = 0.991**

**Confusion Matrix of Training Data is as shown Below**

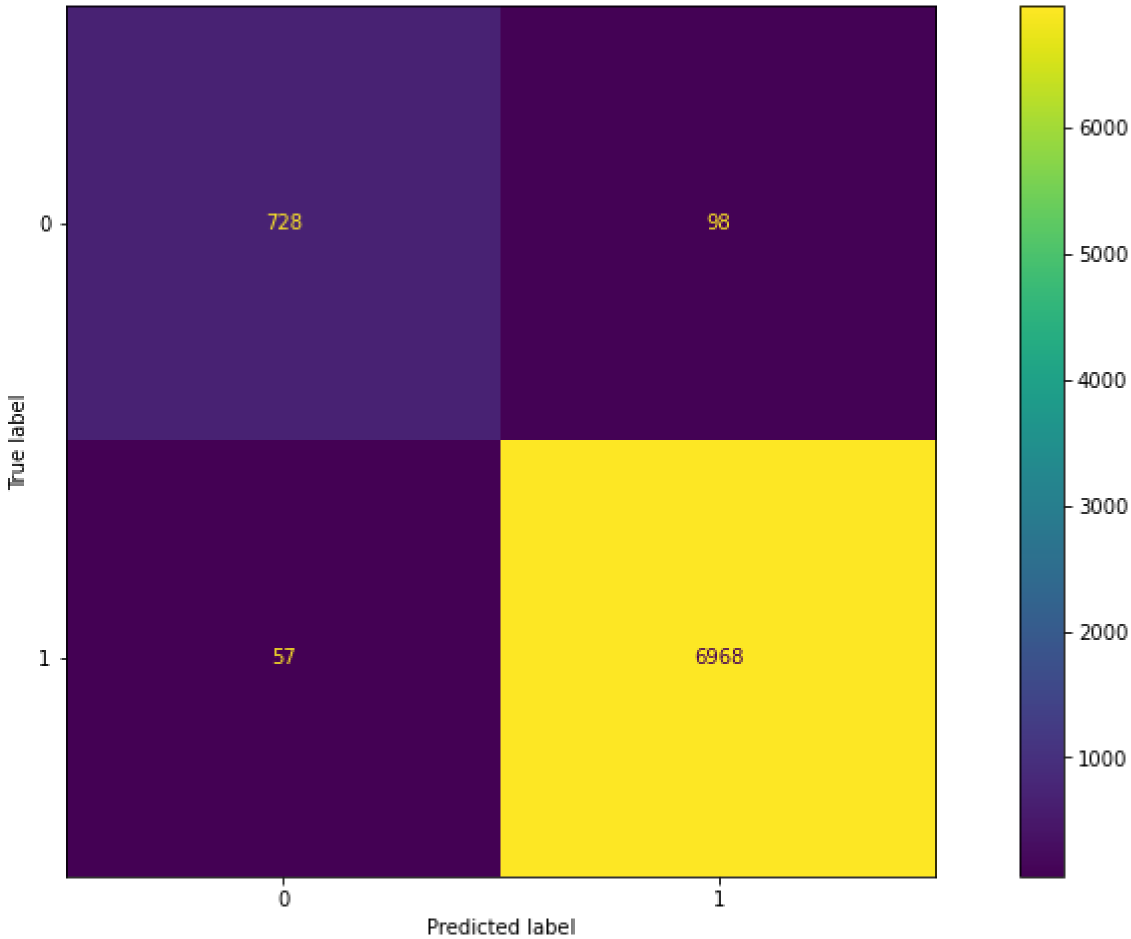



**Fig 2.8 - Confusion Matrix of Training Data for Logistic Regression**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.88 | 0.90 | 826 |
| 1 | 0.99 | 0.99 | 0.99 | 7025 |
| | | | | |
| accuracy | | | 0.98 | 7851 |
| macro avg | 0.96 | 0.94 | 0.95 | 7851 |
| weighted avg | 0.98 | 0.98 | 0.98 | 7851 |

**Table 2.9 - Classification Table for Logistic Regression on Training Data**

- We can see that Accuracy is 98% and F1 score close to 1

**Confusion Matrix of Testing Data is as shown Below**

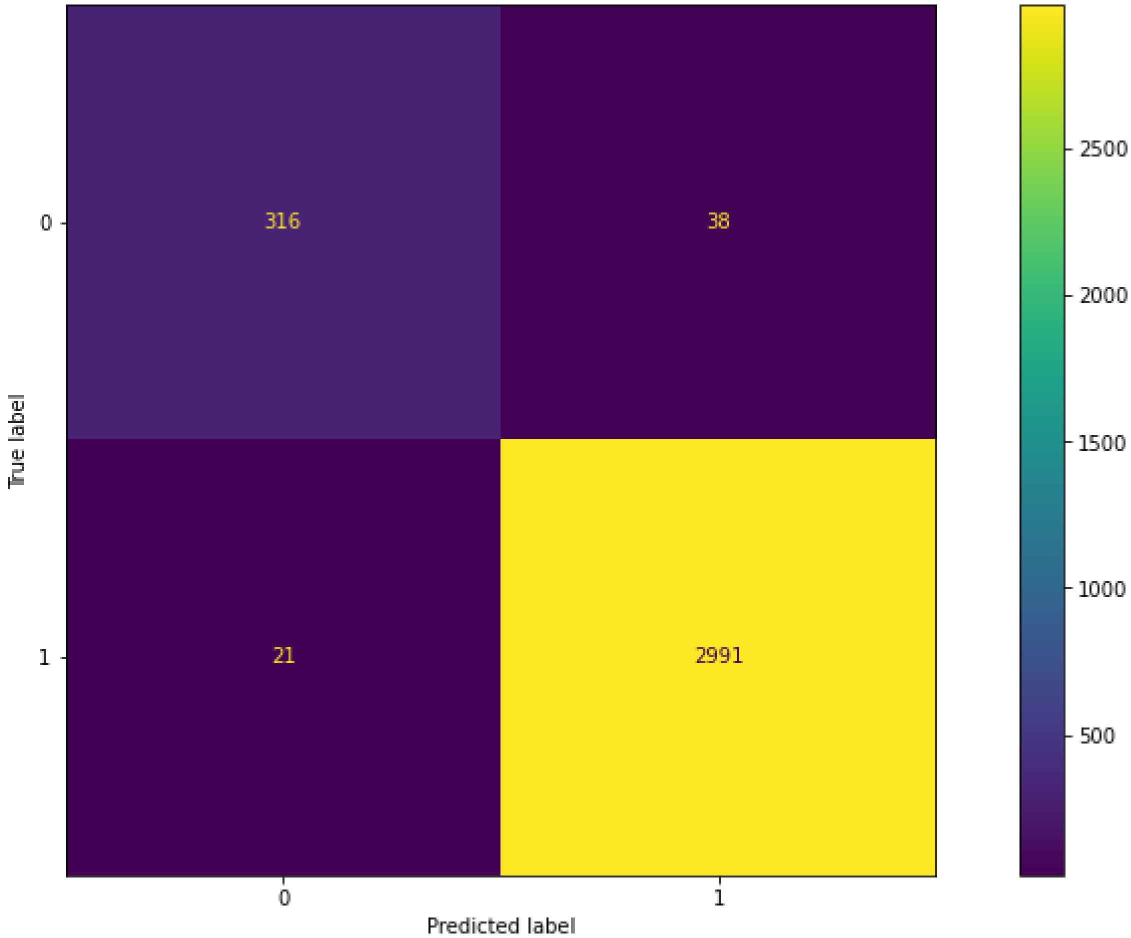



**Fig 2.9 - Confusion Matrix of testing Data for Logistic Regression**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.89 | 0.91 | 354 |
| 1 | 0.99 | 0.99 | 0.99 | 3012 |
| accuracy | | | 0.98 | 3366 |
| macro avg | 0.96 | 0.94 | 0.95 | 3366 |
| weighted avg | 0.98 | 0.98 | 0.98 | 3366 |

**Table 2.10 - Classification Table for Logistic Regression on Testing Data**

- We can see that Accuracy is 98% and F1 score close to 1

## LDA Model

**LDA is now performed on Dataset**

**Confusion Matrix of LDA model is as shown below**

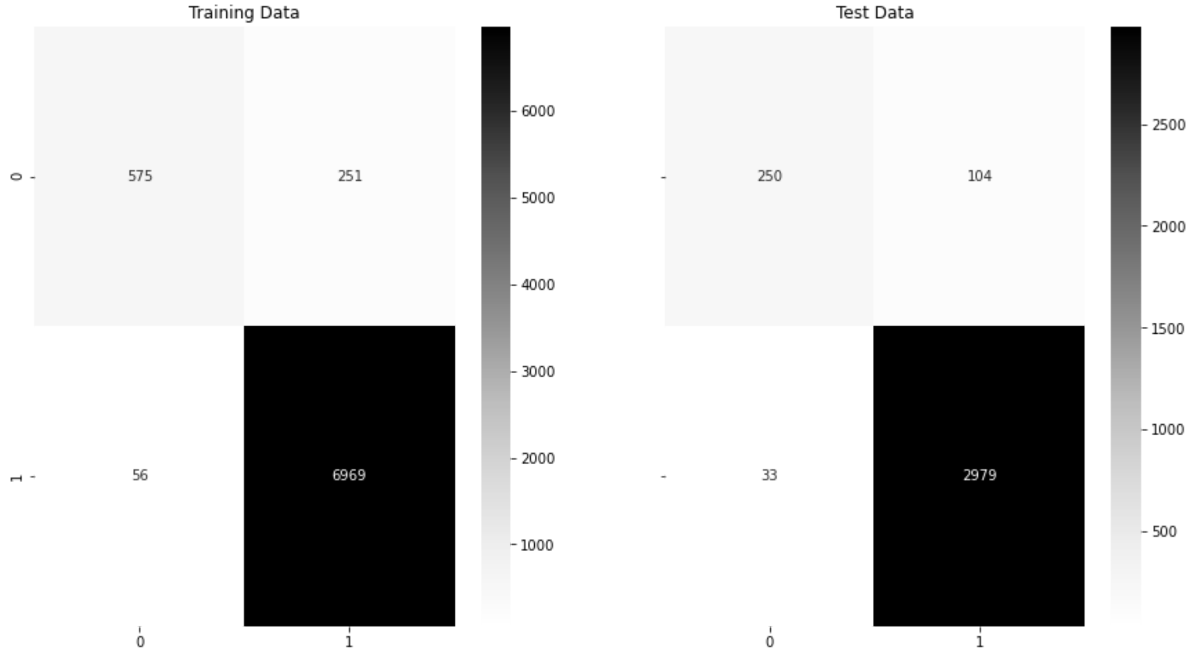



**Fig 2.10 - Confusion Matrix of both Training and testing Data for LDA Model**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.70 | 0.70 | 826 |
| 1 | 0.97 | 0.99 | 0.98 | 7025 |
| accuracy | | | 0.96 | 7851 |
| macro avg | 0.94 | 0.84 | 0.88 | 7851 |
| weighted avg | 0.96 | 0.96 | 0.96 | 7851 |

**Table 2.11 - Classification Table for LDA on Training Data**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.71 | 0.78 | 354 |
| 1 | 0.97 | 0.99 | 0.98 | 3012 |
| accuracy | | | 0.96 | 3366 |
| macro avg | 0.92 | 0.85 | 0.88 | 3366 |
| weighted avg | 0.96 | 0.96 | 0.96 | 3366 |

**Table 2.12 - Classification Table for LDA on Testing Data**

**AUC and ROC for Training data on LDA Model**

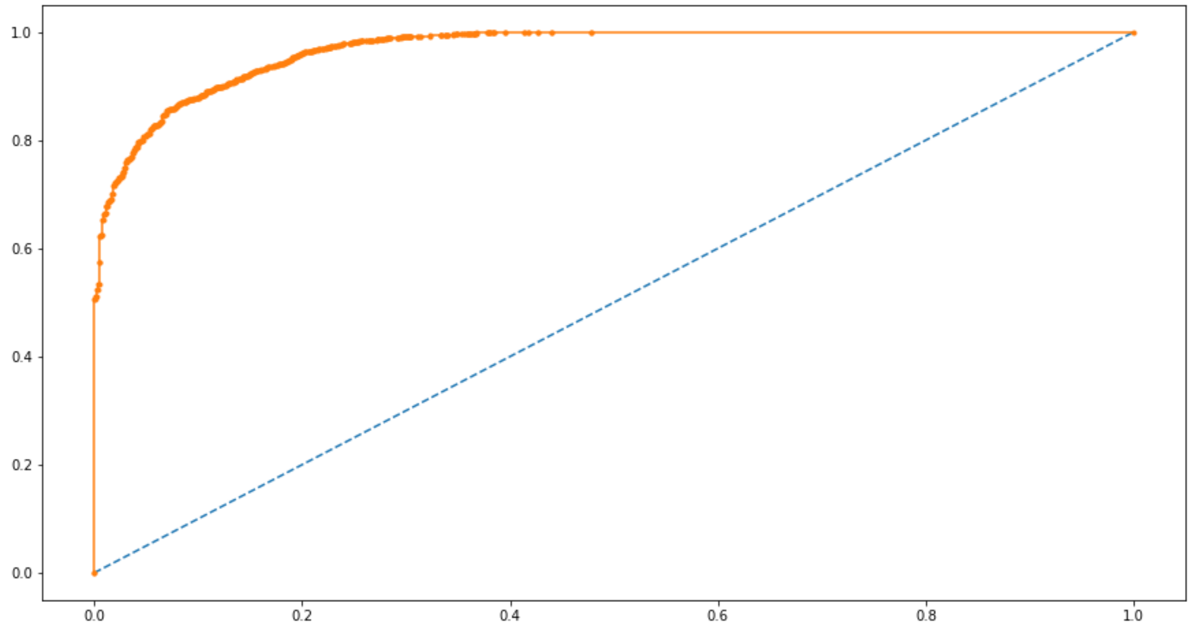



**Fig 2.11 - ROC Curve for Training data for LDA**

**AUC for the Training Data: 0.968**

**AUC and ROC for Testing data on LDA Model**

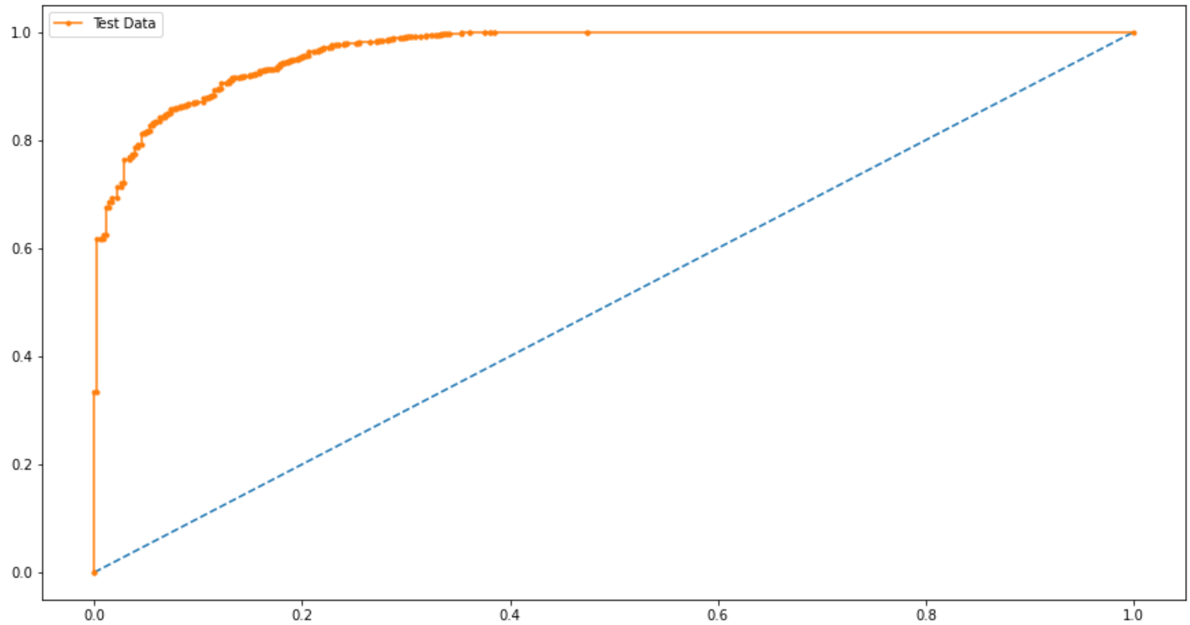



**Fig 2.12 - ROC Curve for Testing data for LDA**

**AUC for the Test Data: 0.967**

## 2.4)Inference: Based on these predictions, what are the insights and recommendations?

**Observations**

1. Model Score of Logistic Regression is 0.9824 where as Model Score of LDA is 0.9592
2. Accuracy on Test Data set for a Logistic Regression Model is 98% where as Accuracy on Test Data set for a LDA Model is 96%
3. ROC_AUC Score for Logistic Regression Model on Test set is 0.991 where as ROC_AUC Score for LDA Model on Test set is 0.967
4. Clearly we can conclude that Logistic Regression Model performs better than LDA model but LDA model also can be considered as the accuracy of it prediction is 96%

**Clearly we can predict if Passenger/ Driver Survived or not with 98% Accuracy Using Logistic Regression Model given the relavent data used for prediction**

**THE END**