Predictive Modeling Business Report

V1.0

Prepared by: Lavanya Sreeram

Prepared for: GL Batch Nov 22

May 14, 2023

PREDICTIVE MODELING

Business Report

Document Information

Author:	Lavanya Sreeram
Document Title	Lavanya_Sreeram_14_05_2023
Issue Date:	14-05-2023
Version:	V 1.0
Subject:	Predictive Modeling Business Report by Lavanya Sreeram

Revision History

Version	Date	Revised By	Reason
1.0	14-05-2023	Lavanya Sreeram	First Issue

Contents

Figures:

Tables:

1 PROBLEM 1: LINEAR REGRESSION

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 DESCRIPTION

Firm_level_data.csv data set is provided. Data Dictionary for Firm_level_data is as below:

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.

1.1 EXPLORATORY DATA ANALYSIS

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. (8 marks)

Answer:

1.1.1 Data Exploration

The Dataset has 759 rows, 10 columns. 'sales' is the Target Variable. The below snap shows the dataset first 5 observations.

	Unnamed: 0	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	0	826.995050	161.603986	10	382.078247	2.306000	no	11.049511	1625.453755	80.27
1	1	407.753973	122.101012	2	0.000000	1.860000	no	0.844187	243.117082	59.02
2	2	8407.845588	6221.144614	138	3296.700439	49.659005	yes	5.205257	25865.233800	47.70
3	3	451.000010	266.899987	1	83.540161	3.071000	no	0.305221	63.024630	26.88
4	4	174.927981	140.124004	2	14.233637	1.947000	no	1.063300	67.406408	49.46

Fig1.1.1 Firm Data top 5 observations – as given

The below snap shows the bottom 5 observations.

	Unnamed: 0	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
754	754	1253.900196	708.299935	32	412.936157	22.100002	yes	0.697454	267.119487	33.50
755	755	171.821025	73.666008	1	0.037735	1.684000	no	NaN	228.475701	46.41
756	756	202.726967	123.926991	13	74.861099	1.460000	no	5.229723	580.430741	42.25
757	757	785.687944	138.780992	6	0.621750	2.900000	yes	1.625398	309.938651	61.39
758	758	22.701999	14.244999	5	18.574360	0.197000	no	2.213070	18.940140	7.50

Fig1.1.2 Firm Data bottom 5 observations – as given

On first look, it can be inferred that the feature 'Unnamed: 0' must be dropped. Additionally, it can be noted that the data is not scaled.

The below snap shows the datatypes of all features of the dataset.

```
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    759 non-null    int64
 1   sales         759 non-null    float64
 2   capital       759 non-null    float64
 3   patents       759 non-null    int64
 4   randd         759 non-null    float64
 5   employment    759 non-null    float64
 6   sp500         759 non-null    object
 7   tobinq        738 non-null    float64
 8   value         759 non-null    float64
 9   institutions  759 non-null    float64
dtypes: float64(7), int64(2), object(1)
```

(Fig1.1.3): Firm Data description - initial

Except 'sp500', all the features in the data are numeric in nature ('int64' or 'float64' type). 'sp500' is object type, and as per data dictionary the feature is binary. 'sp500' shall be encoded.

Drop 'Unnamed:0' feature. Now, the dataset top 5 observations are as shown below:

	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	826.995050	161.603986	10	382.078247	2.306000	no	11.049511	1625.453755	80.27
1	407.753973	122.101012	2	0.000000	1.860000	no	0.844187	243.117082	59.02
2	8407.845588	6221.144614	138	3296.700439	49.659005	yes	5.205257	25865.233800	47.70
3	451.000010	266.899987	1	83.540161	3.071000	no	0.305221	63.024630	26.88
4	174.927981	140.124004	2	14.233637	1.947000	no	1.063300	67.406408	49.46

(Fig 1.1.4): Firm data after dropping 'Unnamed:0' feature

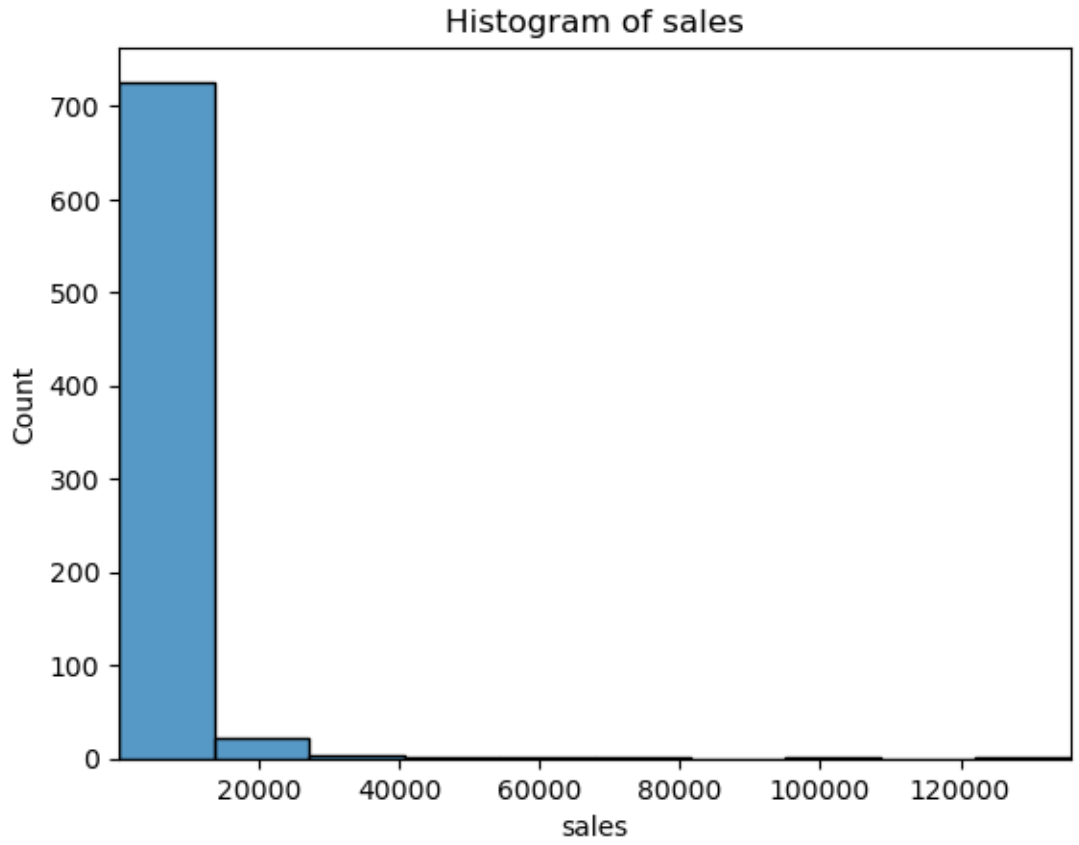
The statistical data descriptions is as shown in below snap.

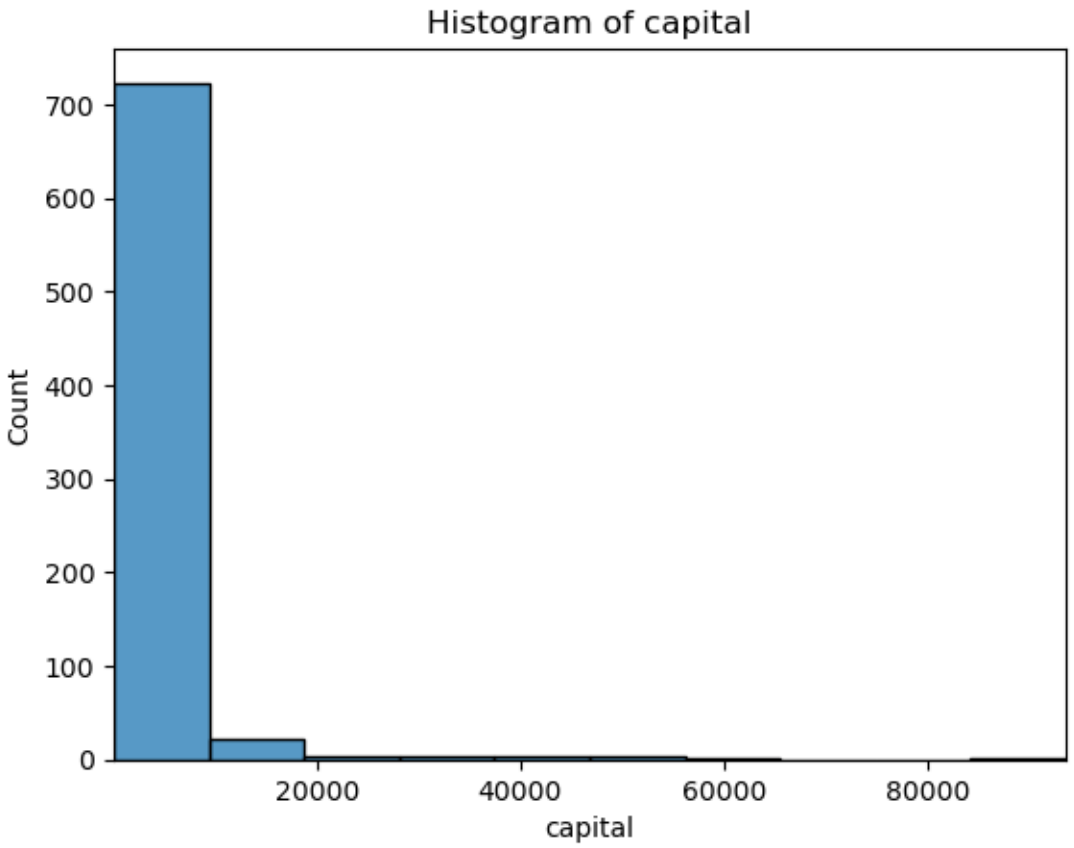
	count	mean	std	min	25%	50%	75%	max
sales	759.0	2689.705158	8722.060124	0.138000	122.920000	448.577082	1822.547366	135696.788200
capital	759.0	1977.747498	6466.704896	0.057000	52.650501	202.179023	1075.790020	93625.200560
patents	759.0	25.831357	97.259577	0.000000	1.000000	3.000000	11.500000	1220.000000
randd	759.0	439.938074	2007.397588	0.000000	4.628262	36.864136	143.253403	30425.255860
employment	759.0	14.164519	43.321443	0.006000	0.927500	2.924000	10.050001	710.799925
tobinq	738.0	2.794910	3.366591	0.119001	1.018783	1.680303	3.139309	20.000000
value	759.0	2732.734750	7071.072362	1.971053	103.593946	410.793529	2054.160386	95191.591160
institutions	759.0	43.020540	21.685586	0.000000	25.395000	44.110000	60.510000	90.150000

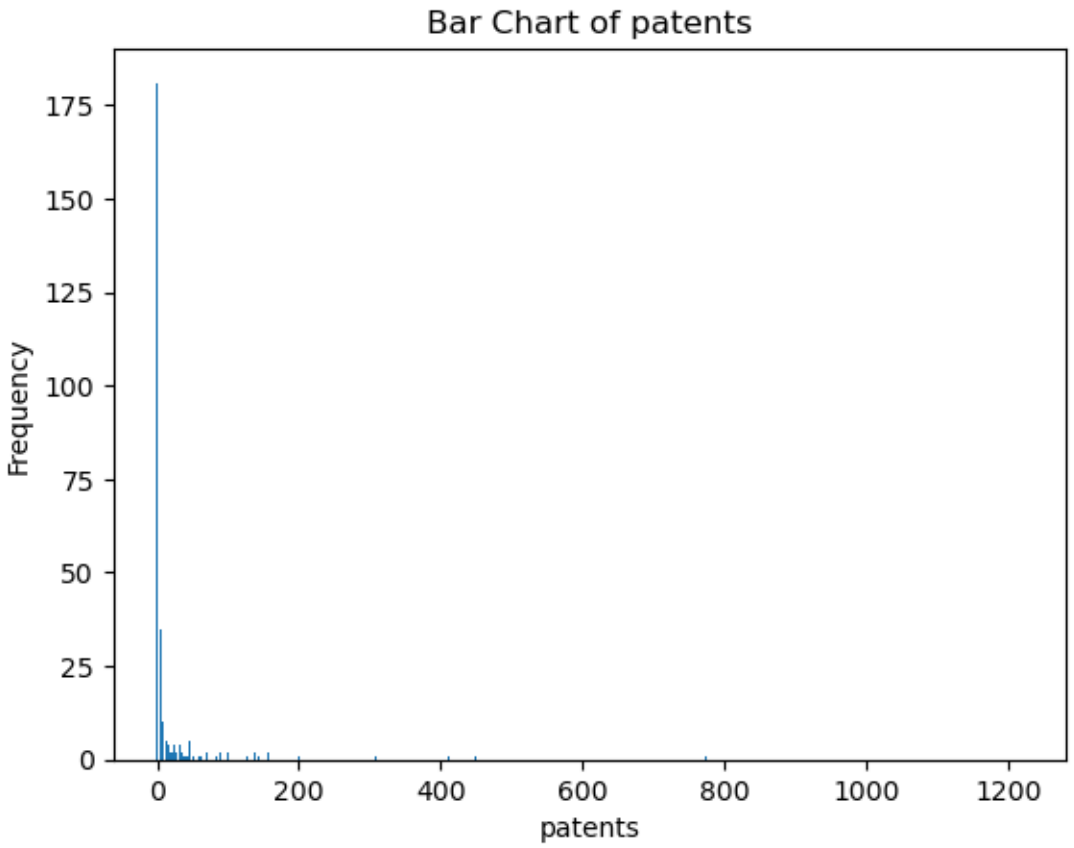
(Fig1.1.5): Firm data statistical description -initial

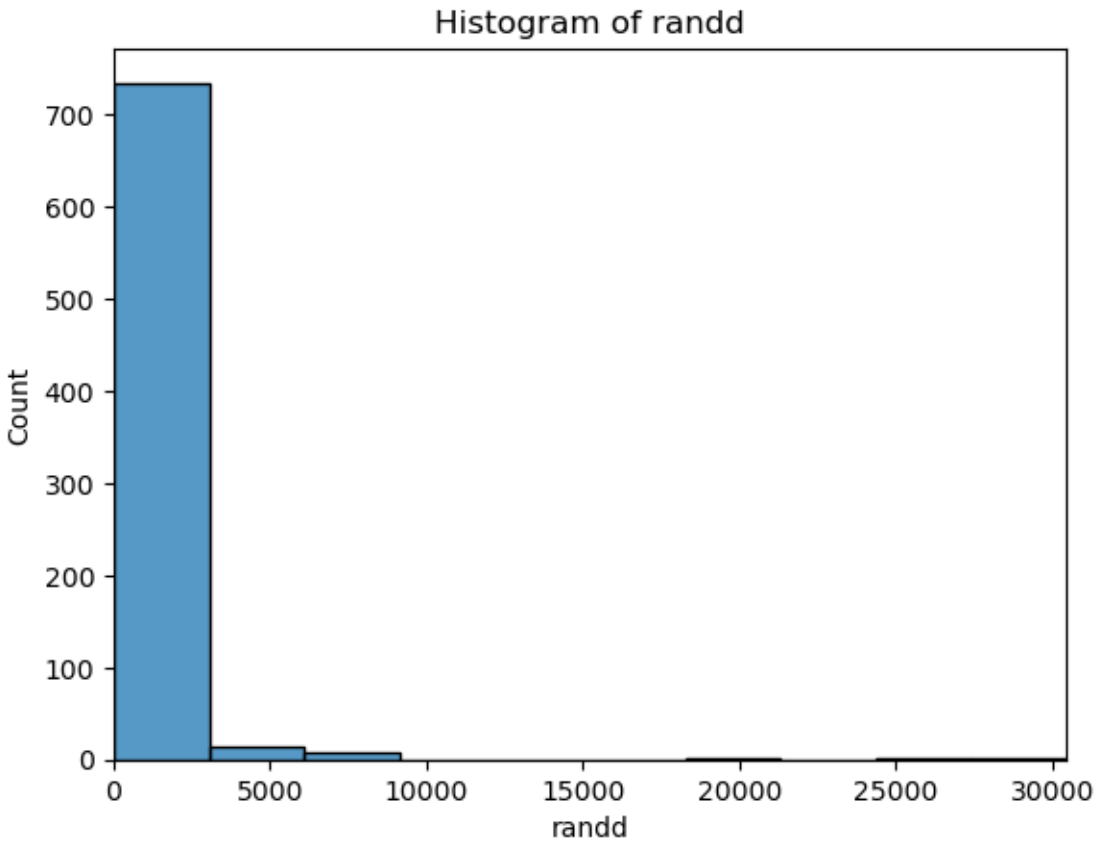
1.1.2 Univariate Analysis:

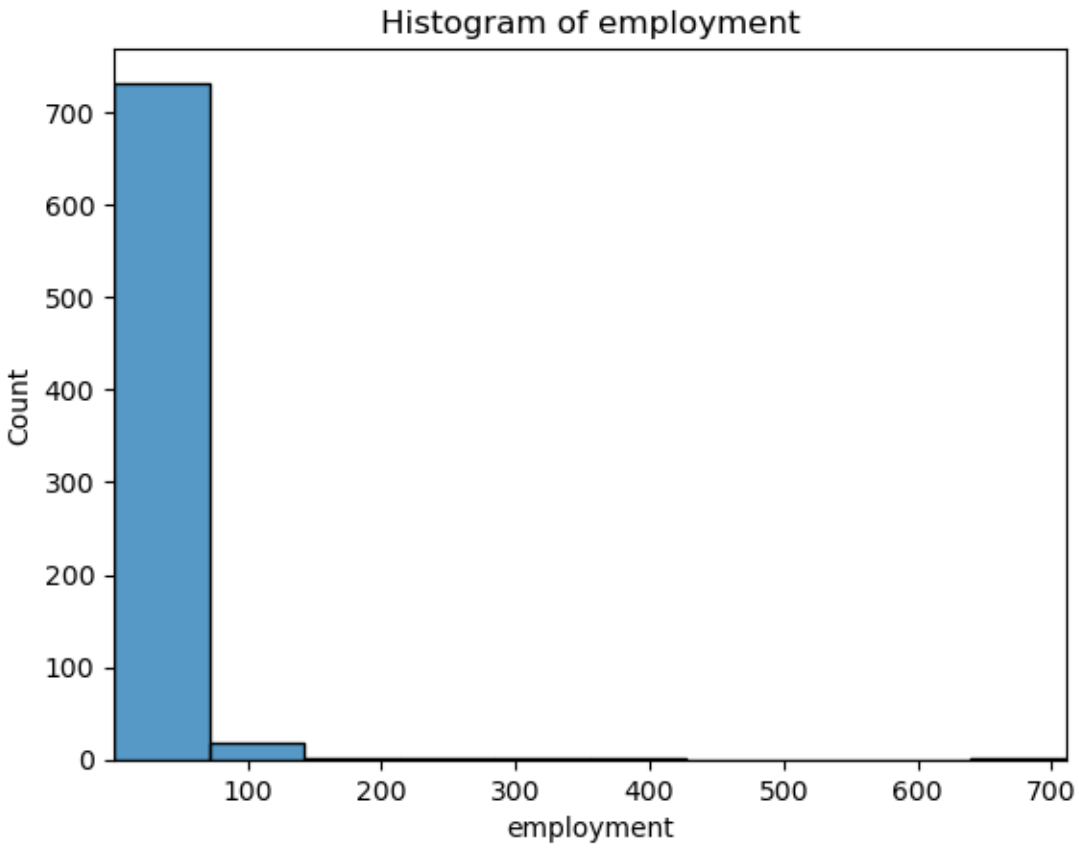
Univariate analysis on the dataset is as shown below using histograms:

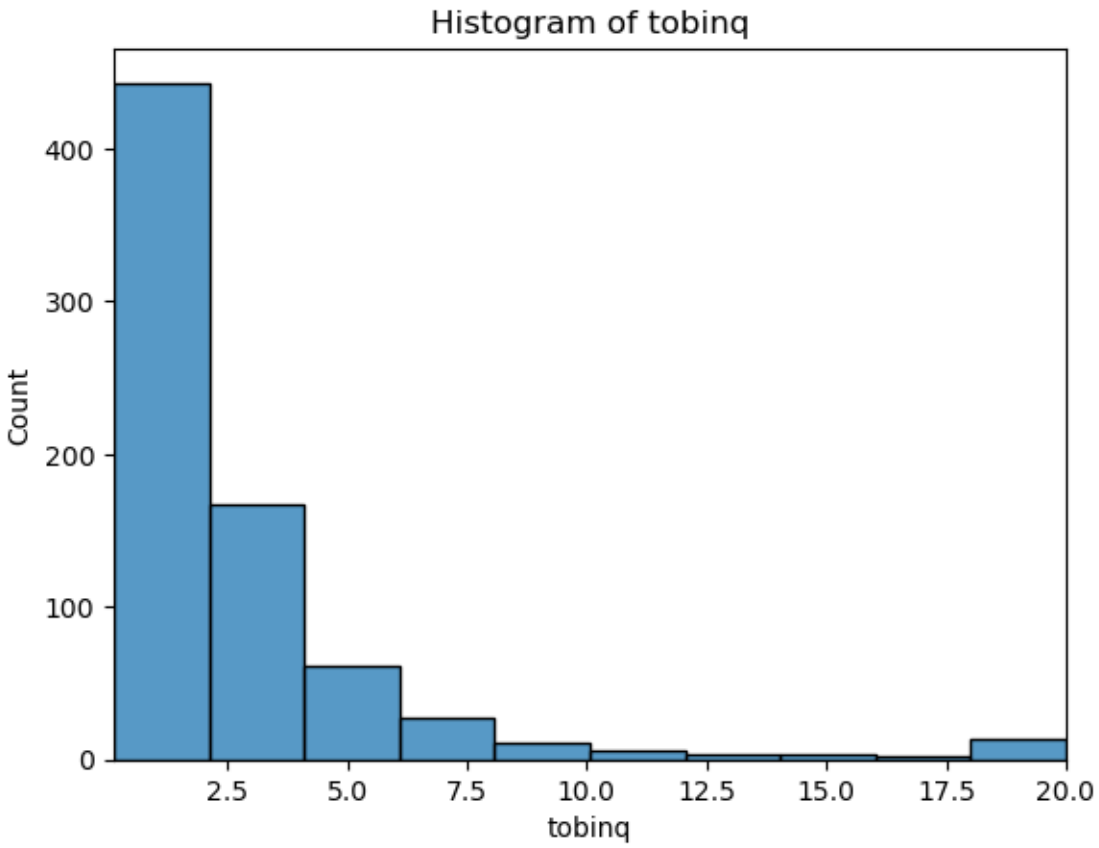


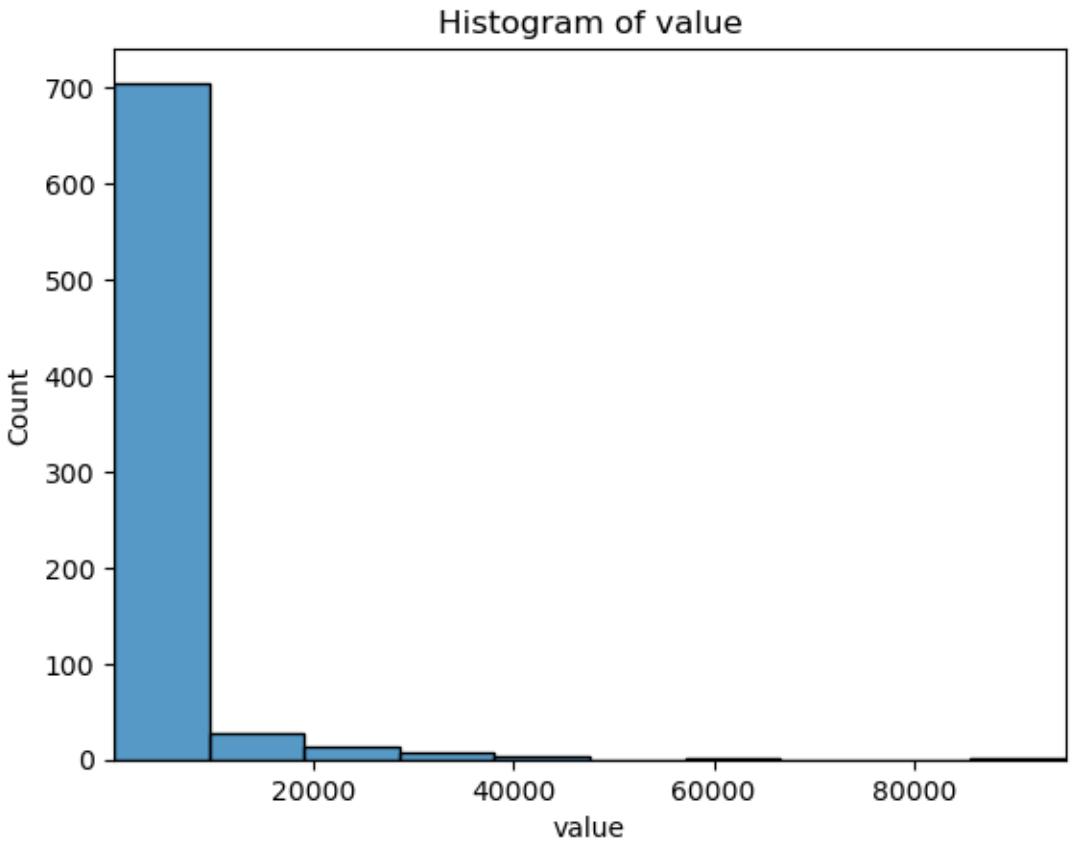

Histogram of capital
Count
700
600
500
400
300
200
100
0
20000
40000
60000
80000
capital


Bar Chart of patents
Frequency
175
150
125
100
75
50
25
0
0
200
400
600
800
1000
1200
patents

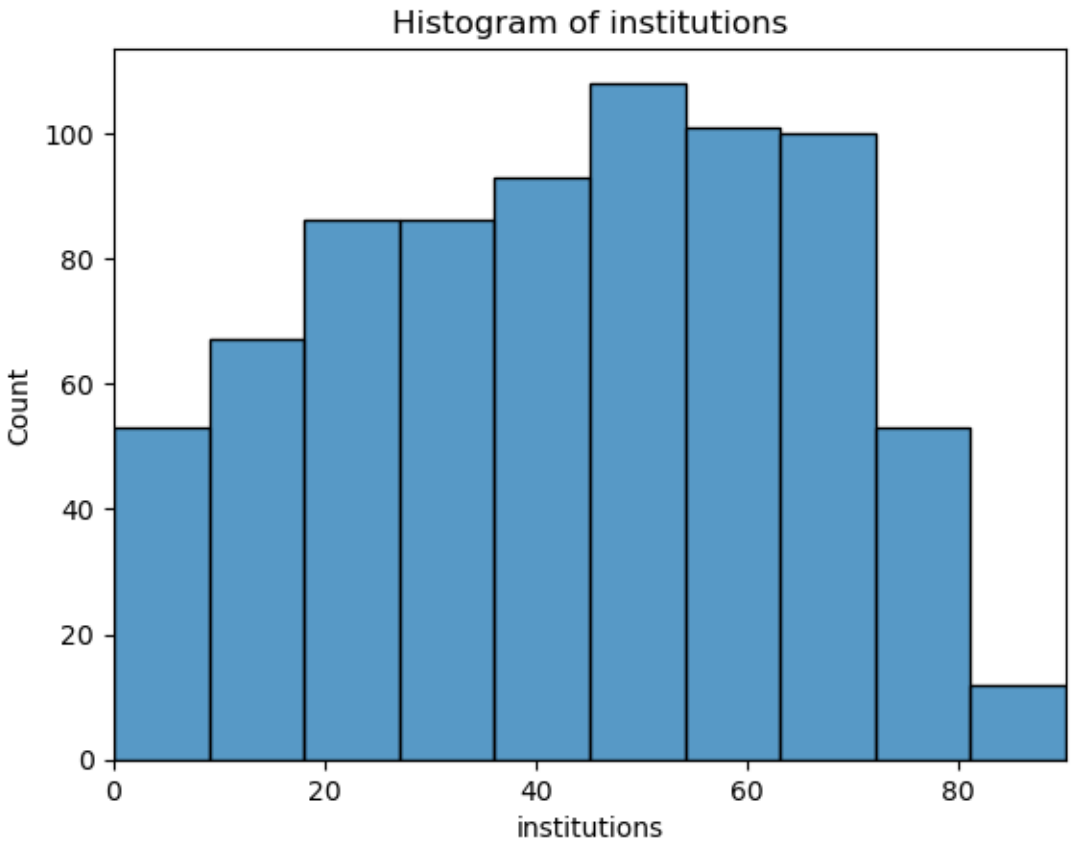

Histogram of randd
Count
700
600
500
400
300
200
100
0
0
5000
10000
15000
20000
25000
30000
randd


Histogram of employment
Count
0
100
200
300
400
500
600
700
100
200
300
400
500
600
700
employment


Histogram of tobinq
Count
400
300
200
100
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
tobinq



All the above histograms show that the data is right skewed or positively skewed meaning that the mean is greater than the median. This is because the presence of a few large values on the right pulls the mean towards higher values. The tail of the distribution extends towards larger values, indicating the presence of outliers or extreme values on the right side. This can happen as few observations in ‘sales’ or ‘value’ are significantly higher in weight/value compared to the majority. This is generally the case with economic observations like stock prices.



From the above histogram, it can be noted that the 'institutions' is not as skewed as rest of the features. In fact, the feature is more uniform and has left skewedness in data.

Count plot is used for the binary variable 'sp500' as shown below:

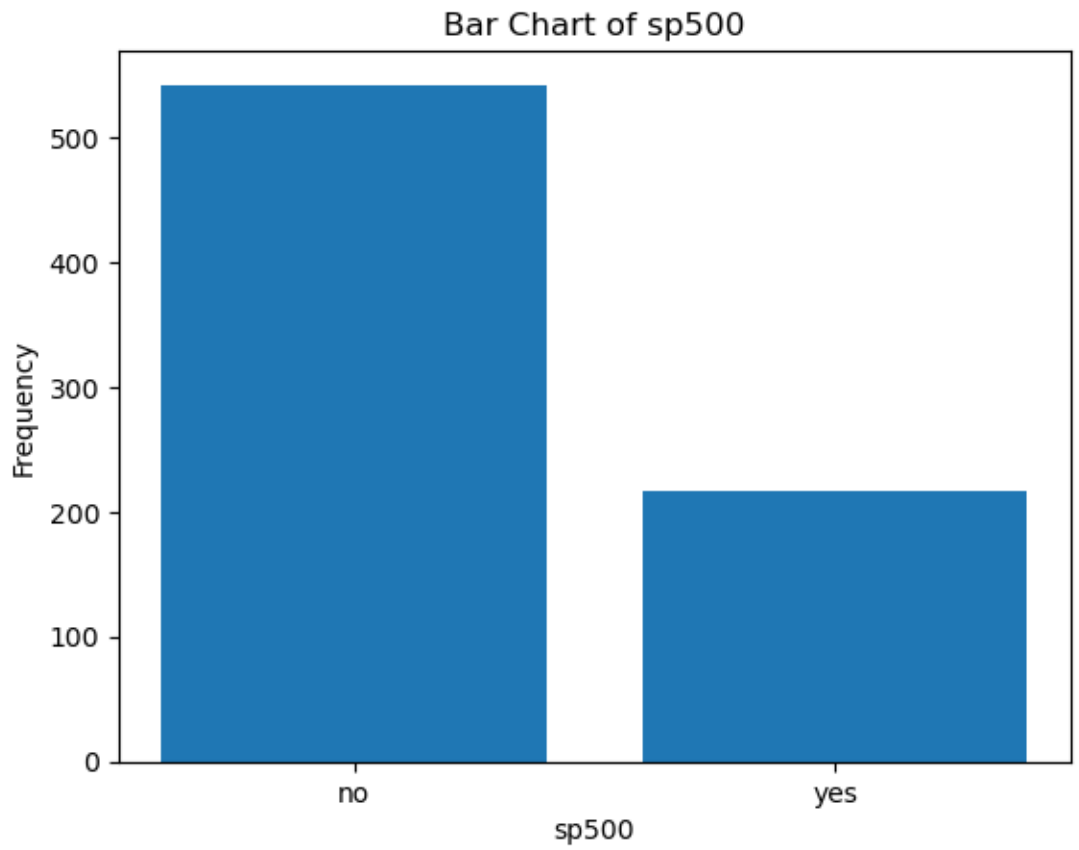
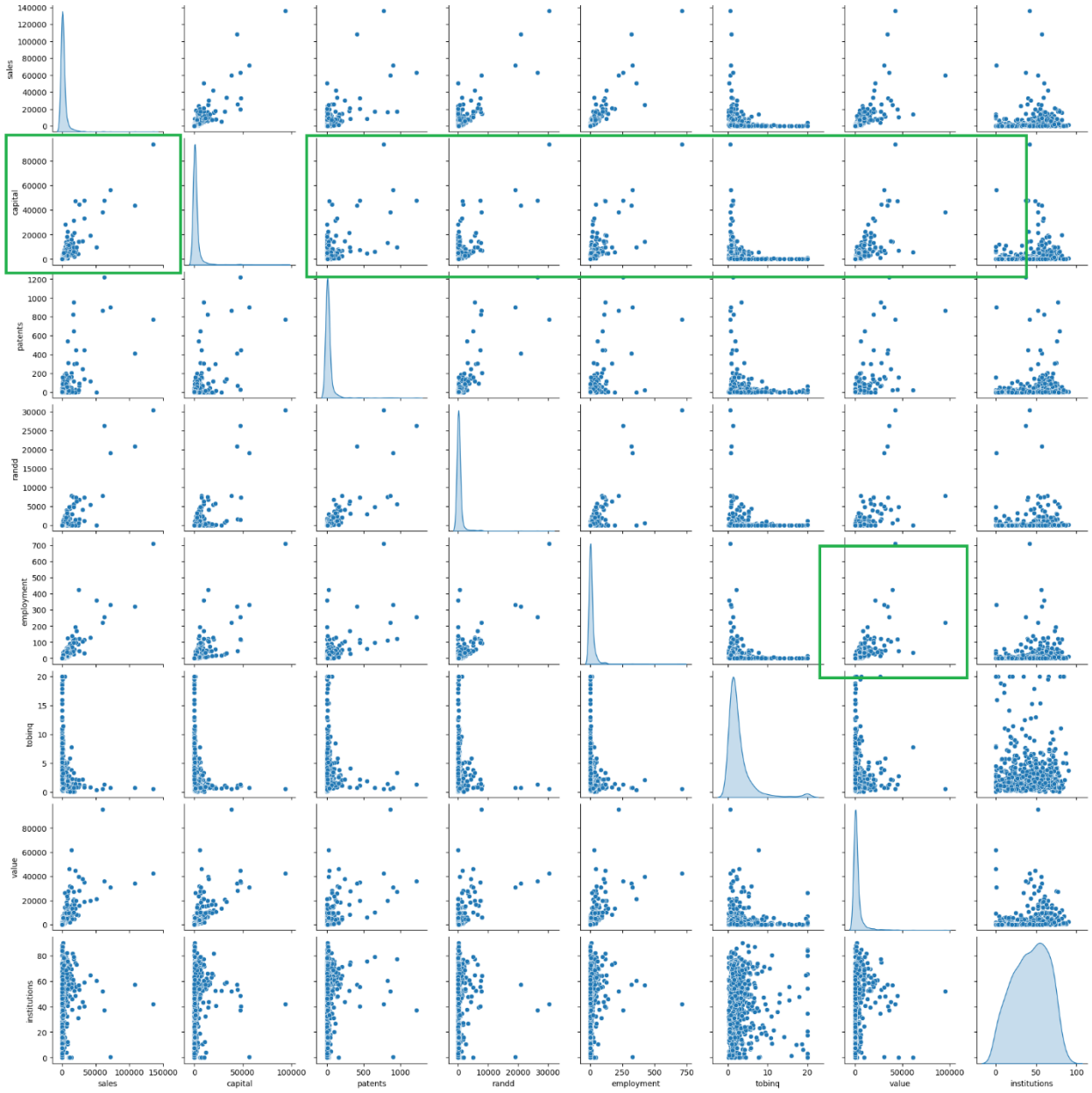


Fig1.1.6 Histogram and countplot of the firm data

The plot shows that the observations with S&P 500 index companies are less in number compared to opposite.

1.1.3 Bivariate Analysis

Bivariate analysis is performed on the data and the pair plot is shown below:
Fix- highlight co-relation features.



(Fig 1.1.7): Firm data – Pairplot

There is lot of positive high correlation & linear relationship for 'capital' variable with multiple variables. Similar insights are drawn from the heatmap.

1.2 DATA PREPARATION

Impute null values if present? Do you think scaling is necessary in this case? (8 marks):

Answer:

1.2.1 Null Values

There are no duplicates in the data. There are 21 null/missing values in 'tobinq' column. Boxplots are generated to visualize the skewness in data as shown below.

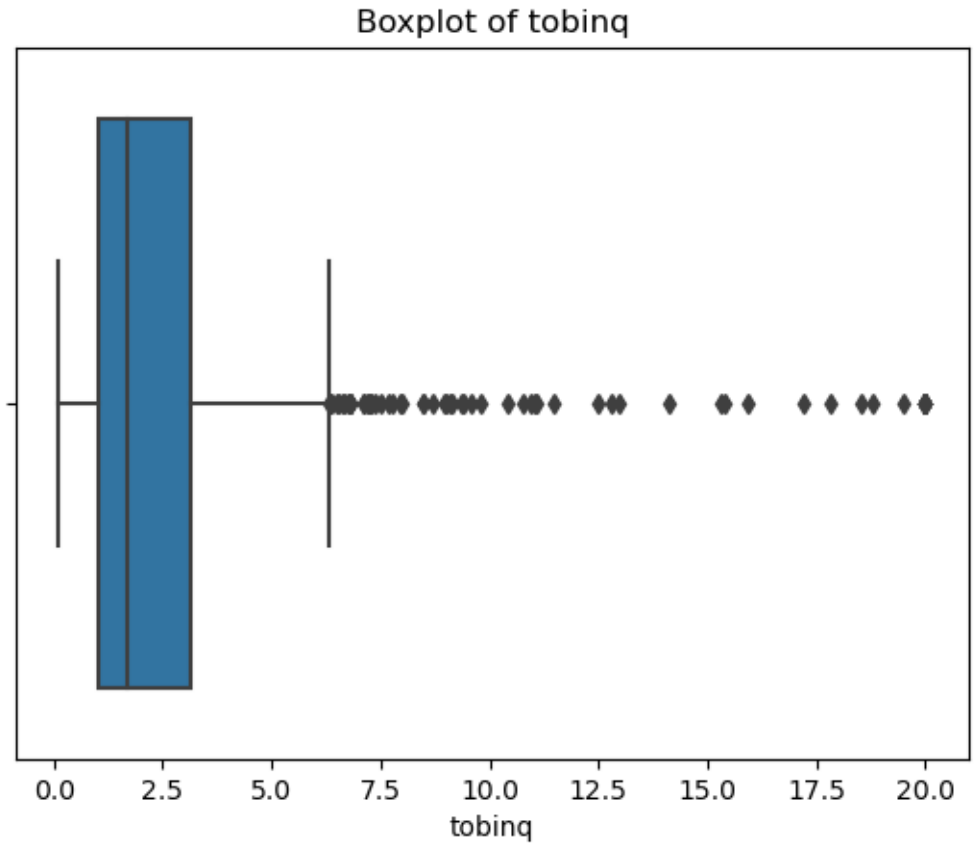


Fig1.2.1 'tobinq' boxplot

The data is right skewed. The missing values in 'tobinq' are treated by median imputation. Median imputation is preferred when the distribution is skewed, as the median is less sensitive to outliers than the mean.

After treating the null values in the data, the data info can be seen as below with no null values.

```
 #   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         759 non-null    float64
 7   value          759 non-null    float64
 8   institutions   759 non-null    float64
dtypes: float64(7), int64(1), object(1)
```

Fig1.2.2 Firm Data Info after null values are treated

1.2.2 Encode Data

The 'sp500' feature is the only categorical feature and contains binary values. After one-hot encoding, the dataset looks like below.

	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	826.995050	161.603986	10	382.078247	2.306000	0	11.049511	1625.453755	80.27
1	407.753973	122.101012	2	0.000000	1.860000	0	0.844187	243.117082	59.02
2	8407.845588	6221.144614	138	3296.700439	49.659005	1	5.205257	25865.233800	47.70
3	451.000010	266.899987	1	83.540161	3.071000	0	0.305221	63.024630	26.88
4	174.927981	140.124004	2	14.233637	1.947000	0	1.063300	67.406408	49.46

Fig1.2.2.1 Firm Data after encoding

The dataset types are as below after the encoding:

```
 #   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    int64
 6   tobinq         759 non-null    float64
 7   value          759 non-null    float64
 8   institutions   759 non-null    float64
dtypes: float64(7), int64(2)
```

Fig1.2.2.2 Firm Data after converting datatype

1.2.3 Outliers

The boxplot is plotted to visualize the outliers in all numerical features i.e. except ‘sp500’ which is essentially binary/categorical feature.

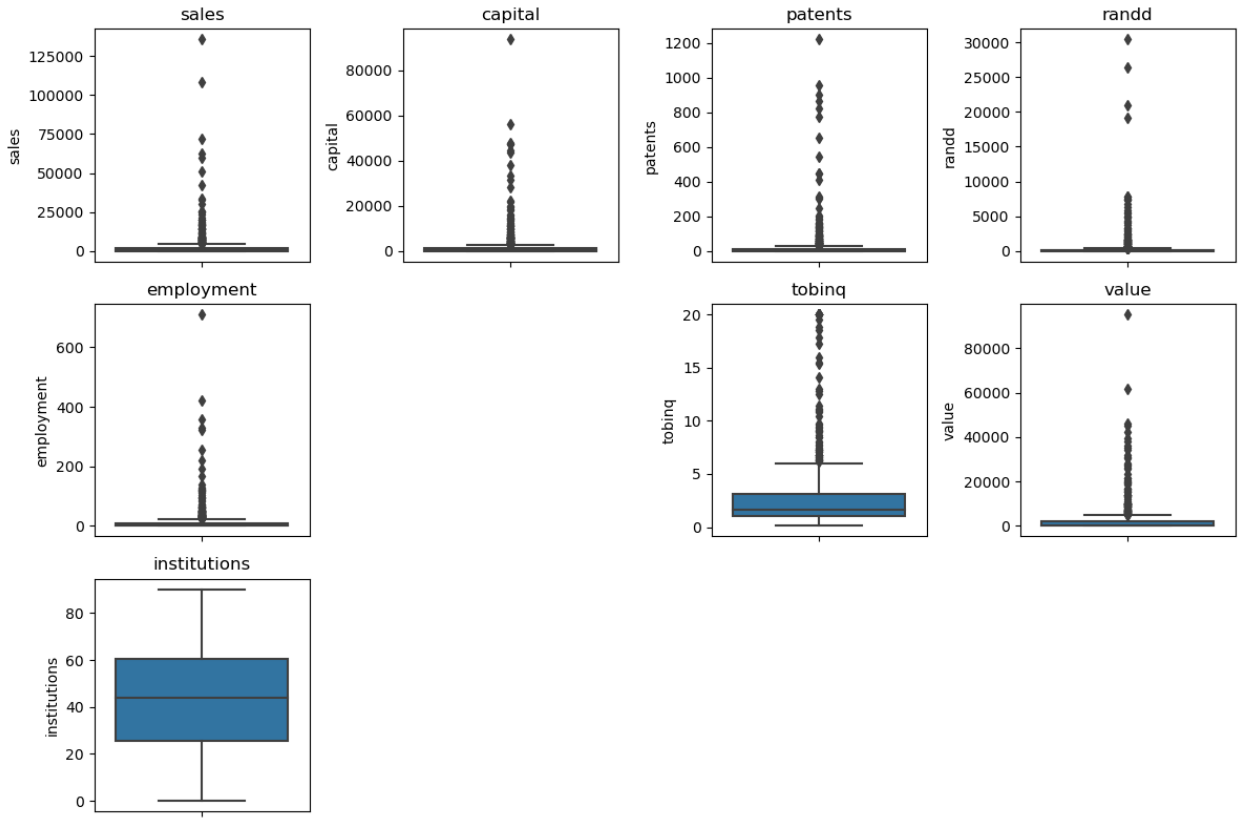


Fig1.2.3.1 Firm Data Boxplot showing outliers

There are multiple outliers in many columns as shown in above plot.

The black dots in the boxplots show that there are multiple outliers in multiple columns. Except 'institutions' remaining features have outliers. Majority of the variables are highly skewed as well and this can be seen in the boxplots as significantly larger tails with large magnitude (Y value).

Treat outliers:

All the outliers are treated by adjusting them to the lower and upper bound values calculated by the IQR value.

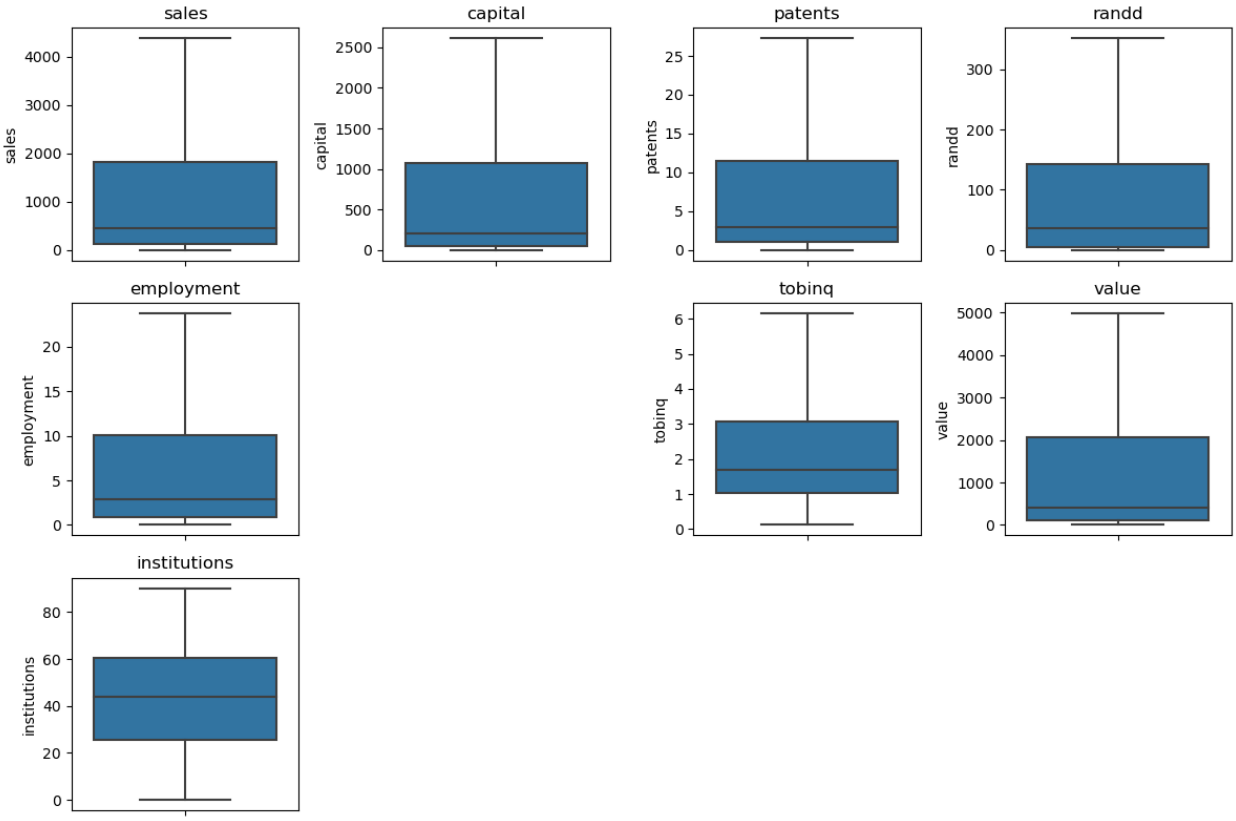


Fig1.2.3.2 Firm Data Boxplot after treating outliers

The above boxplot is after treating the outliers showing no outliers.

1.2.4 Co-relation

4 co-relation heat maps are shown in this section. These are heatmaps of independent features , heatmaps of all features before and after treating outliers.

Before treating outliers:

The below plot shows the correlation between all variables before outliers are treated:

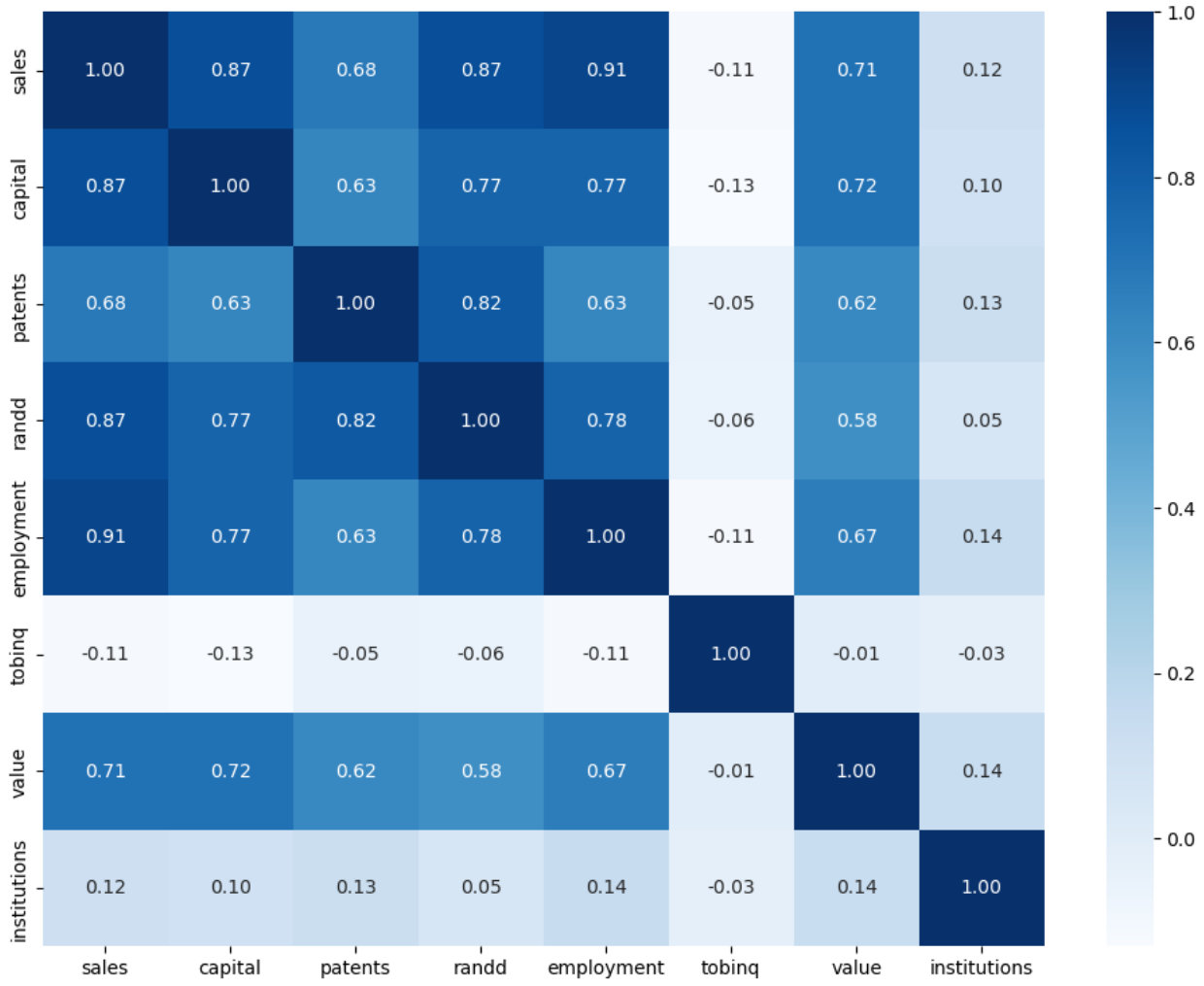


Fig1.2.4.1 Heatmap – of all variables

The below plot shows the correlation between independent variables before outliers are treated.

Fig1.2.4.2 Heatmap – of independent variables - before outliers

Below chart shows there is lots of multi-collinearity and co-relation between variables after treating the outliers.

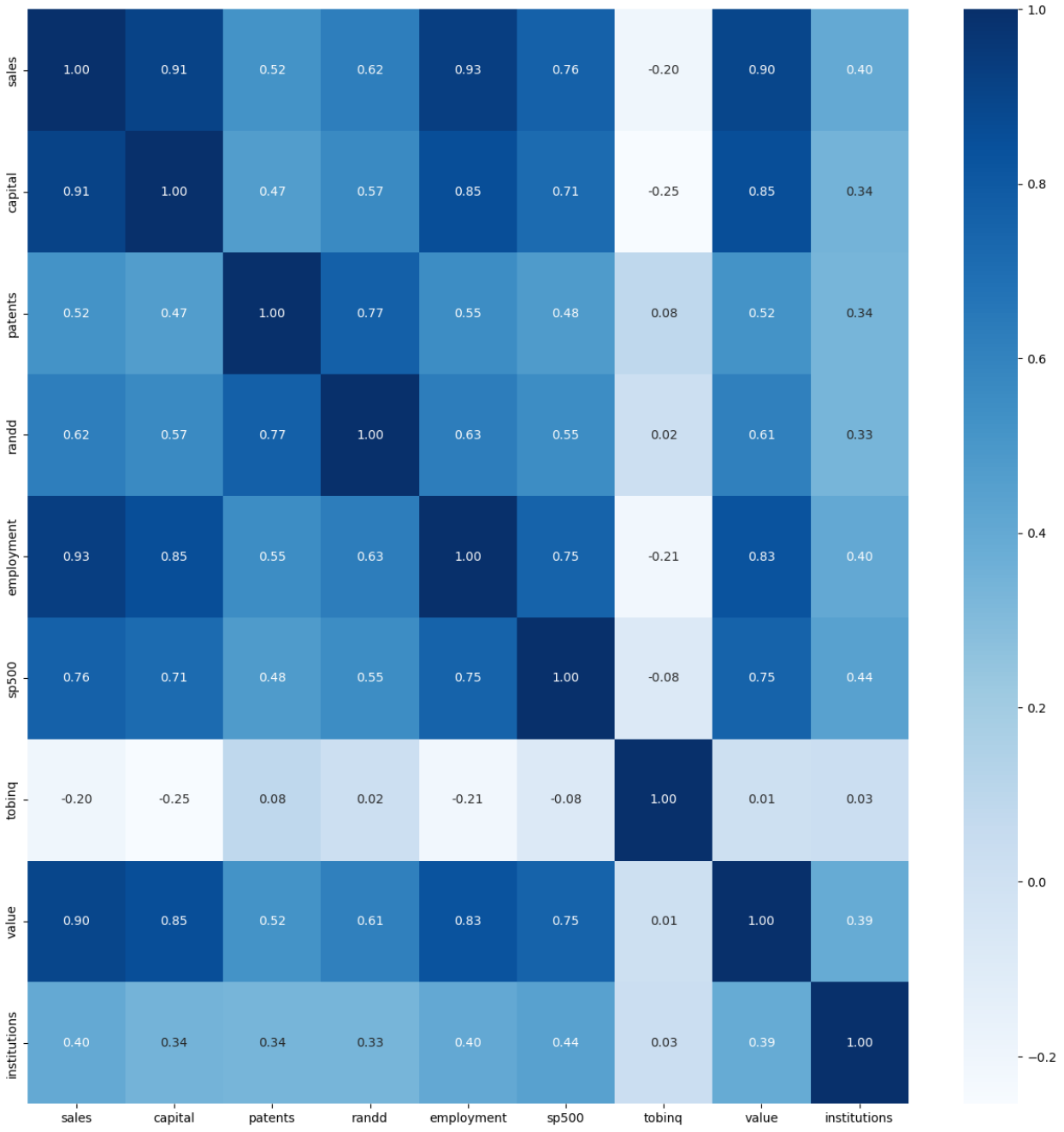


Fig1.2.4.3 Heatmap – of all variables - after treating outliers

The below heatmap shows the co-relation matrix of independent variables:

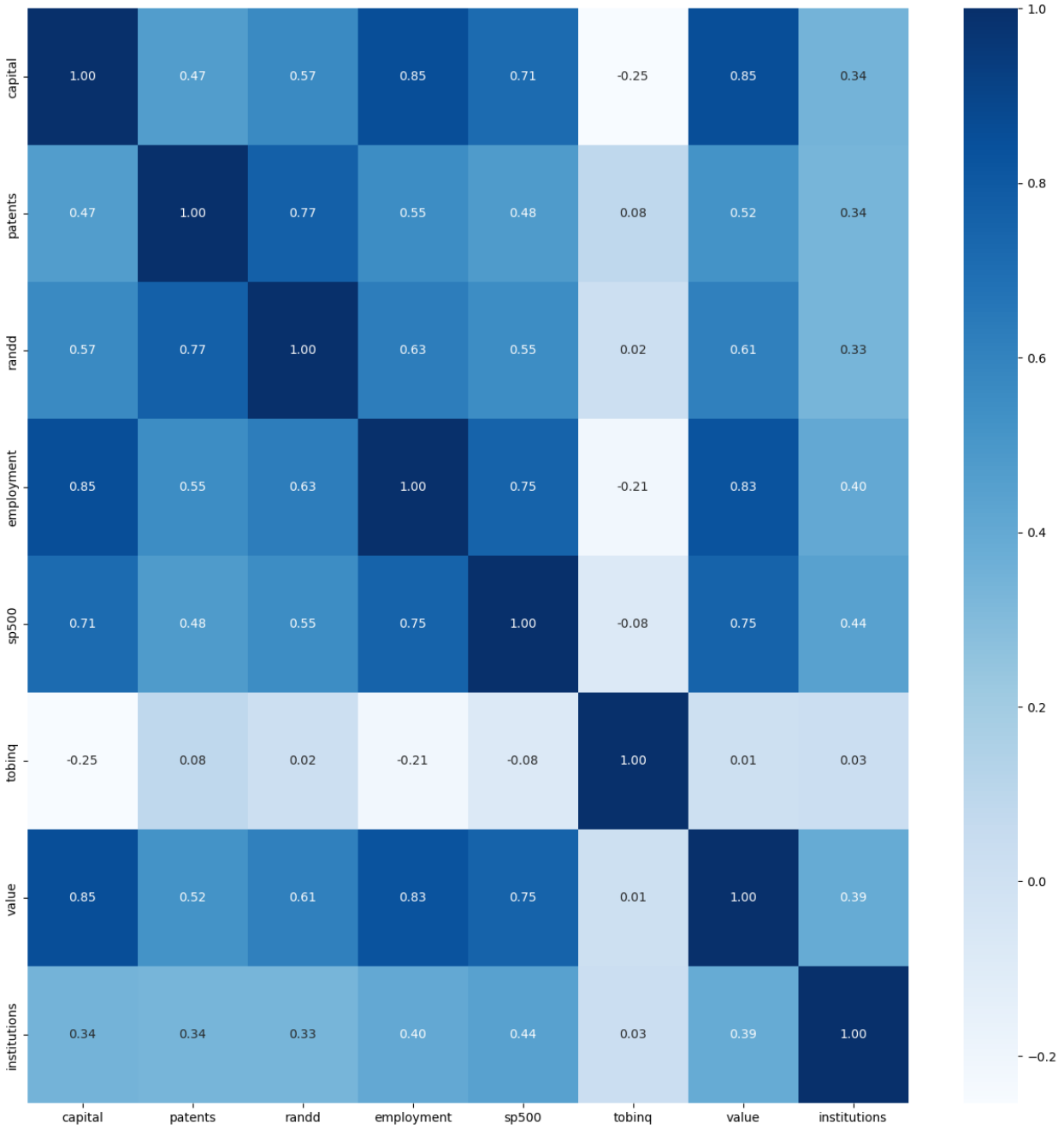


Fig1.2.4.4 Heatmap – of independent variables - after treating outliers

Based on the above correlation matrix plot:

There are several strong and moderate positive correlations, indicating that certain variables tend to move together in a positive direction. For example, ‘capital’ shows strong positive correlations with ‘patents’, ‘randd’, and ‘employment’.

'tobinq', 'institutions' are weakly co-related. 'patents', 'randd', 'sp500' are moderately co-realted with values around .6/.7. 'capital', 'employment', 'value' are highly co-related with correlation values over .91.

'value' shows moderate positive correlations with multiple variables, indicating potential associations with those variables. The variable 'institutions' generally has weak positive correlations with other variables.

The variable 'tobinq' has weak negative correlations with several other variables, suggesting a weak inverse relationship.

There is high co-relation between the independent variables. This needs to be addressed. It's important to note that correlation alone does not imply causation. Further analysis and domain knowledge are needed to determine the underlying relationships and causality between variables.

1.2.5 Scaling

The below snap shows the data description after the data is prepared.

	count	mean	std	min	25%	50%	75%	max
sales	759.0	1236.090089	1528.690552	0.138000	122.920000	448.577082	1822.547366	4371.988416
capital	759.0	728.715785	959.394531	0.057000	52.650501	202.179023	1075.790020	2610.499299
patents	759.0	7.800395	9.952684	0.000000	1.000000	3.000000	11.500000	27.250000
randd	759.0	99.512662	127.195056	0.000000	4.628262	36.864136	143.253403	351.191114
employment	759.0	6.925381	8.184188	0.006000	0.927500	2.924000	10.050001	23.733752
sp500	759.0	0.285903	0.452141	0.000000	0.000000	0.000000	1.000000	1.000000
tobinq	759.0	2.300408	1.723035	0.119001	1.036000	1.680303	3.082979	6.153448
value	759.0	1375.431494	1754.489690	1.971053	103.593946	410.793529	2054.160386	4980.010044
institutions	759.0	43.020540	21.685586	0.000000	25.395000	44.110000	60.510000	90.150000

Fig1.2.5.1 Firm Data after scaling

Scaling - Magnitude Differences:

The variables in the dataset have significantly different scales. For example, 'sales' and 'value', 'capital' have much larger magnitudes compared to 'patents', 'employment', and 'institutions'. Scaling can help to ensure that all features contribute equally to the model.

That being said, it is not necessary to perform scaling on the data. In the later sections, ols regression model and linear regression model are fit on the original and scaled data. The models show no major difference in terms of performance metrics.

1.3 CREATE MODELS

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. (8 marks)

Answer:

1.3.1 Split Data

The 'sales' feature is the target variable, y. Rest of the data is X. Using sklearn train_test_split feature, the data X and y is split into train and test sets in a 70:30 ratio.

Train Test Data for the Stats Model:

The Train dataset top 5 rows are shown below.

	const	capital	patents	randd	employment	sp500	tobinq	value	institutions
626	1.0	1315.696256	15.0	73.275818	16.472000	0	1.657513	2231.870118	31.47
333	1.0	15.258002	2.0	9.252643	0.566000	0	0.381755	9.877838	21.69
257	1.0	538.188036	20.0	87.388641	6.627000	0	2.126738	1019.443780	69.64
173	1.0	807.215091	0.0	68.900185	7.607001	1	3.151469	2221.768944	69.69
242	1.0	402.508010	2.0	0.000000	1.550000	0	2.154388	358.040202	85.42

The test dataset top 5 rows are shown below.

	const	capital	patents	randd	employment	sp500	tobinq	value	institutions
480	1.0	50.688001	1.0	47.173386	1.147000	0	1.006168	34.516077	34.92
622	1.0	80.960002	3.0	50.251263	3.400000	0	1.259892	164.840772	18.88
638	1.0	1119.000008	19.0	78.623947	18.988003	1	1.900413	2114.826950	47.94
389	1.0	68.742010	3.0	44.827785	1.204000	0	2.262480	82.287341	24.65
748	1.0	308.770949	2.0	79.026939	3.264000	0	1.741800	533.056000	16.05

The training and test dataset sample for stats model **after scaling** is shown below:

	const	capital	patents	randd	employment	sp500	tobinq	value	institutions
626	1.0	0.612227	0.723860	-0.206409	1.167240	0	-0.373364	0.488463	-0.532988
333	1.0	-0.744145	-0.583181	-0.710087	-0.777545	0	-1.114266	-0.778833	-0.984276
257	1.0	-0.198723	1.226569	-0.095381	-0.036482	0	-0.100860	-0.203035	1.228328
173	1.0	0.081876	-0.784265	-0.240832	0.083340	1	0.494257	0.482702	1.230635
242	1.0	-0.340238	-0.583181	-0.782879	-0.657234	0	-0.084802	-0.580261	1.956480

	const	capital	patents	randd	employment	sp500	tobinq	value	institutions
480	1.0	-0.707191	-0.683723	-0.411760	-0.706508	0	-0.751635	-0.764781	-0.373791
622	1.0	-0.675617	-0.482640	-0.387546	-0.431039	0	-0.604284	-0.690451	-1.113941
638	1.0	0.407071	1.126027	-0.164334	1.474865	1	-0.232299	0.421708	0.227003
389	1.0	-0.688360	-0.482640	-0.430213	-0.699538	0	-0.022027	-0.737535	-0.847690
748	1.0	-0.438007	-0.583181	-0.161164	-0.447668	0	-0.324414	-0.480442	-1.244528

1.3.1.1 VIF (Variance Inflation Factor)

The below snap the sorted features by Importance of the sm ols original train dataset.

	feature	VIF
5	tobinq	2.656490
7	institutions	3.846556
4	sp500	4.193898
1	patents	4.285710
2	randd	4.690289
0	capital	8.251460
3	employment	8.859125
6	value	10.207006

Fig1.3.1.1.1 Variance Inflation Factor on ols train dataset

Below snap shows the sorted features by VIF importance of the sms ols scaled train dataset.

	feature	VIF
7	institutions	1.272333
5	tobinq	1.426445
4	sp500	1.965792
1	patents	2.654913
2	randd	2.937165
3	employment	5.130612
0	capital	5.667341
6	value	6.382272

Fig1.3.1.1.2 Variance Inflation Factor on scaled ols train dataset

In both cases, 'employment', 'capital', 'value' have high multicollinearity with VIF value > 5. However, dropping variables blindly based on VIF value may not always result in better models.

Using VIF, features with high co-linearity can be identified. For reference, the range of VIF:

VIF	Starts at 1
VIF = 1	no correlation between this independent variable and the other variables
1 < VIF < 5	Indicates moderate collinearity
VIF > 5	exceeding 5 or 10 indicates high multicollinearity between this independent variable and the other.

Table1 VIF Range

1.3.2 Model 1 – OLS Initial

OLS Stats Model is fit on the train data. The performance metrics on the initial model are shown as below.

Initial OLS Model Training Set Metrics:
RMSE: 394.33716619028996
R-squared: 0.9359702538559448

Initial OLS Model Test Set Metrics:
RMSE: 400.0021149372728
R-squared: 0.9240311293641787

The summary of the results is as shown below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     953.8
Date:                Sun, 14 May 2023   Prob (F-statistic):          7.27e-306
Time:                        04:39:53   Log-Likelihood:                -3927.4
No. Observations:                 531   AIC:                             7873.
Df Residuals:                     522   BIC:                             7911.
Df Model:                           8
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           73.7353     48.629      1.516      0.130     -21.798     169.269
capital          0.4051      0.042      9.650      0.000       0.323       0.488
patents         -4.6622      2.786     -1.674      0.095     -10.134       0.810
randd            0.6385      0.232      2.749      0.006       0.182       1.095
employment      78.5650      4.756     16.518      0.000      69.221      87.909
sp500          167.8727     66.445      2.527      0.012      37.341     298.404
tobinq         -40.9833     12.057     -3.399      0.001     -64.670     -17.296
value            0.2455      0.025      9.649      0.000       0.196       0.296
institutions     0.2129      0.901      0.236      0.813      -1.558       1.984
==============================================================================
Omnibus:                      184.867   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1273.725
Skew:                           1.347   Prob(JB):                    2.59e-277
Kurtosis:                      10.093   Cond. No.                     9.83e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.83e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
```

Fig1.3.2 Model 1 – OLS Initial Model

1.3.3 Model 2 – OLS Dropped p>.05

All features with p values above .05 are dropped and the new model regression performance met rics are:

Training Set Metrics:
RMSE: 395.3975671797366
R-squared: 0.9356254296687838

Test Set Metrics:
RMSE: 398.4877494359508
R-squared: 0.9246052606459508

And the summary of result is as shown below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     1269.
Date:                Sun, 14 May 2023   Prob (F-statistic):          2.32e-308
Time:                        04:39:53   Log-Likelihood:                -3928.8
No. Observations:                 531   AIC:                             7872.
Df Residuals:                     524   BIC:                             7901.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         80.1957     37.396      2.144      0.032       6.731     153.660
capital        0.4091      0.042      9.777      0.000       0.327       0.491
randd          0.4002      0.183      2.192      0.029       0.042       0.759
employment    77.7994      4.713     16.509      0.000      68.541      87.057
sp500        161.0654     65.150      2.472      0.014      33.078     289.053
tobinq       -43.3778     11.967     -3.625      0.000     -66.888     -19.868
value          0.2449      0.025      9.623      0.000       0.195       0.295
==============================================================================
Omnibus:                      183.370   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1228.445
Skew:                           1.344   Prob(JB):                    1.76e-267
Kurtosis:                       9.950   Cond. No.                     9.51e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.51e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fig1.3.3 Model 2 – OLS Model dropped p>.05

The model assumptions, such as normality and absence of autocorrelation, should be carefully evaluated.

1.3.4 Model 3 – OLS Scaled Data

The performance metrics on the OLS Scaled data are as below:

Scaled Model Training Set Metrics:
RMSE: 394.33716619028996
R-squared: 0.9359702538559448

Scaled Model Test Set Metrics:
RMSE: 400.002114937273
R-squared: 0.9240311293641786

regression result on scaled data are as shown below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     953.8
Date:                Sun, 14 May 2023   Prob (F-statistic):          7.27e-306
Time:                        04:43:20   Log-Likelihood:                -3927.4
No. Observations:                 531   AIC:                             7873.
Df Residuals:                     522   BIC:                             7911.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1192.7832     25.570     46.648      0.000    1142.551    1243.016
capital      388.3593     40.246      9.650      0.000     309.294     467.424
patents      -46.3704     27.705     -1.674      0.095    -100.798       8.057
randd         81.1654     29.520      2.749      0.006      23.172     139.158
employment   642.5668     38.901     16.518      0.000     566.146     718.988
sp500        167.8727     66.445      2.527      0.012      37.341     298.404
tobinq       -70.5691     20.762     -3.399      0.001    -111.356     -29.782
value        430.5140     44.615      9.649      0.000     342.866     518.161
institutions   4.6133     19.536      0.236      0.813     -33.765      42.991
==============================================================================
Omnibus:                      184.867   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1273.725
Skew:                           1.347   Prob(JB):                    2.59e-277
Kurtosis:                      10.093   Cond. No.                         8.47
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig1.3.4 Model 3 – OLS Initial Model on scaled data

1.3.5 Model 4 – OLS Scaled and refined

The scaled and new model after dropping all features with p>0.05 has below metrics:
Scaled Improved Model Training Set Metrics:
RMSE: 395.3975671797366
R-squared: 0.9356254296687838

Scaled Improved Model Test Set Metrics:
RMSE: 398.48774943595146
R-squared: 0.9246052606459505

The complete summary is as shown below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     1269.
Date:                Sun, 14 May 2023   Prob (F-statistic):          2.32e-308
Time:                        04:43:20   Log-Likelihood:                -3928.8
No. Observations:                 531   AIC:                             7872.
Df Residuals:                     524   BIC:                             7901.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1194.0317     25.405     46.999      0.000    1144.123    1243.940
capital      392.2491     40.119      9.777      0.000     313.435     471.063
randd         50.8673     23.203      2.192      0.029       5.285      96.449
employment   636.3050     38.544     16.509      0.000     560.586     712.024
sp500        161.0654     65.150      2.472      0.014      33.078     289.053
tobinq       -74.6922     20.607     -3.625      0.000    -115.174     -34.210
value        429.4345     44.626      9.623      0.000     341.767     517.102
==============================================================================
Omnibus:                      183.370   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1228.445
Skew:                           1.344   Prob(JB):                    1.76e-267
Kurtosis:                       9.950   Cond. No.                         7.59
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig1.3.4 Model 4 – OLS Model on scaled data dropped p>.05

1.3.6 Model 5 – OLS Data – dropping variables

Blindly dropping all features above p>0.05 is not always correct as the dropped features may have multi-coliinearity but still carry important information about the target variable. So instead, drop each variable based on p-value, lower rmse, higher r-squared, Higher F-Statistic. This is iterative process and in the final improved model 'sales','patents','institutions','randd' are dropped.

The OLS Stats improved regression model equation can be given as below:

y = (81.7013 * const) + (0.4118 * capital) + (80.6845 * employment) + (176.4655 * sp500) - (39.6549 * tobinq) + (0.2466 * value)

1. When all other predictors are zero, the const is 81.703 which is not bad.
2. for a one-unit increase in the "sp500" variable, holding all other variables constant, the sales price increases by 176.4655
3. For one-unit increase in ‘tobinq’ the sales decreases by -39.6549 as the ‘-‘ sign indicates an inverse relationship.

The performance metrics and the summary of this improved ols model are as below:

Training Set Metrics:
RMSE: 397.20671726067104

R-squared: 0.935034987496708

Test Set Metrics:
RMSE: 394.38620894692014
R-squared: 0.9261493138182268

Performance metrics on test data-

1. An R-squared value of 0.926 suggests that the model explains approximately 92.6% of the variance in the sales data.
2. The RMSE for the test set indicates the average deviation of the predicted sales values from the actual sales values in the test set. A lower RMSE is desirable, indicating better predictive performance. The RMSE of 394.38 suggests an average deviation of approximately 398.488 which is pretty good considering the target variable range.
3. The target variable has a mean of 2689.705158 and a standard deviation of 8722.060124. In this context, an RMSE of 394 is relatively small compared to the magnitude of the target variable. It indicates that, on average, the model's predictions have an error of approximately 394 units, which is relatively low compared to the overall scale of the target variable. Hence, this is an acceptable value of RMSE.

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.935
Model:	OLS	Adj. R-squared:	0.934
Method:	Least Squares	F-statistic:	1511.
Date:	Sun, 14 May 2023	Prob (F-statistic):	6.21e-309
Time:	07:38:21	Log-Likelihood:	-3931.2
No. Observations:	531	AIC:	7874.
Df Residuals:	525	BIC:	7900.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.7013	37.525	2.177	0.030	7.983	155.419
capital	0.4118	0.042	9.811	0.000	0.329	0.494
employment	80.6845	4.542	17.766	0.000	71.763	89.606
sp500	176.4655	65.005	2.715	0.007	48.764	304.167
tobinq	-39.6549	11.889	-3.335	0.001	-63.011	-16.299
value	0.2466	0.026	9.659	0.000	0.196	0.297

Omnibus:	183.570	Durbin-Watson:	1.965
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1224.353
Skew:	1.347	Prob(JB):	1.36e-266
Kurtosis:	9.934	Cond. No.	9.45e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.45e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Fig1.3.6 Model 5 - OLS Improved model

1.3.7 Model 6 – Simple Linear Regression

Below are the Lienar regression results:

Coefficients: [[0.40506319 -4.66215536 0.63853807 78.56497068 167.87266108
-40.98329451 0.24554026 0.21287606]]
Intercept: [73.73534908]

Training set performance:
RMSE: 394.337
R-squared: 0.936
MSE 155501.80063898838

Test set performance:
RMSE: 400.002
R-squared: 0.924
MSE 160001.69195429183

1.3.7.1 Actual v predicted scatter plot

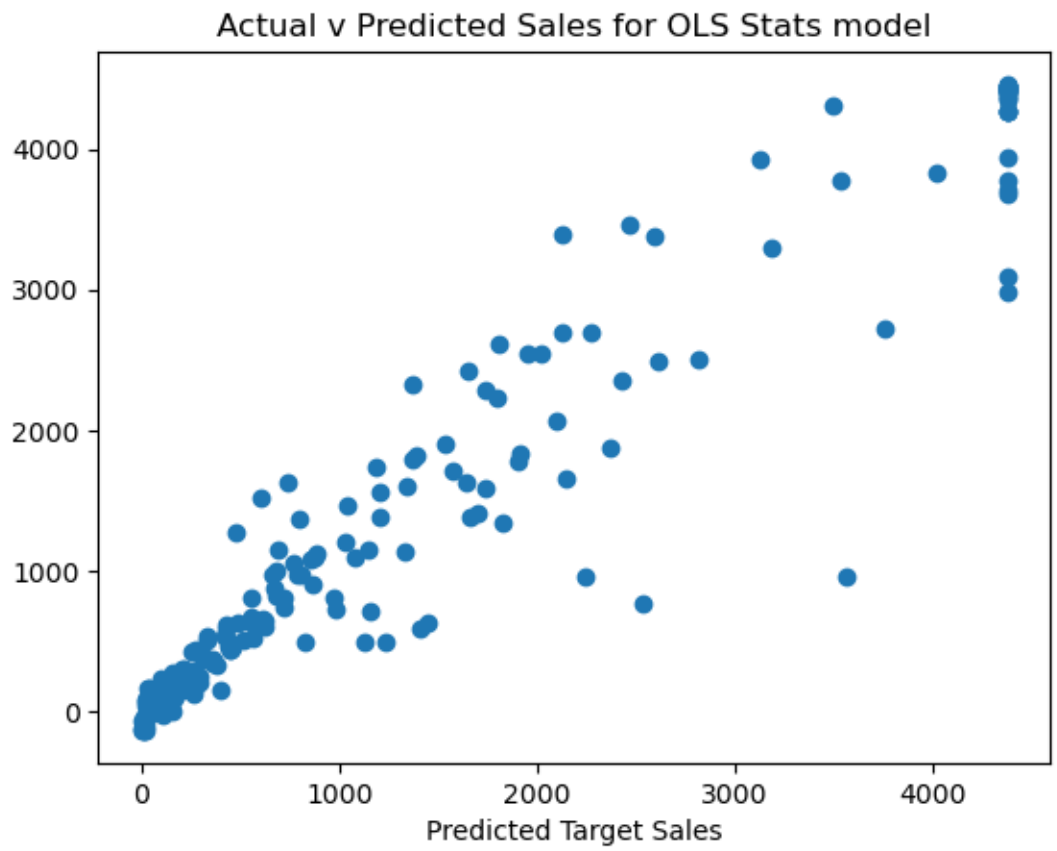


Fig1.3.7.1 Scatter Plot- Actual v Predicted Sales for OLS Stats model.

The above scatter plot shows between actual and predicted values. Inferences are:

1. There are 3 areas, one where the sales are below '1000', between '1000-2000' & higher.
2. The model correctly predicts level 1, whereas the variance increases as the sales value inc reases.
3. The variance of the error is not constant across various levels of your dependent variable.

1.3.7.2 Residual plot

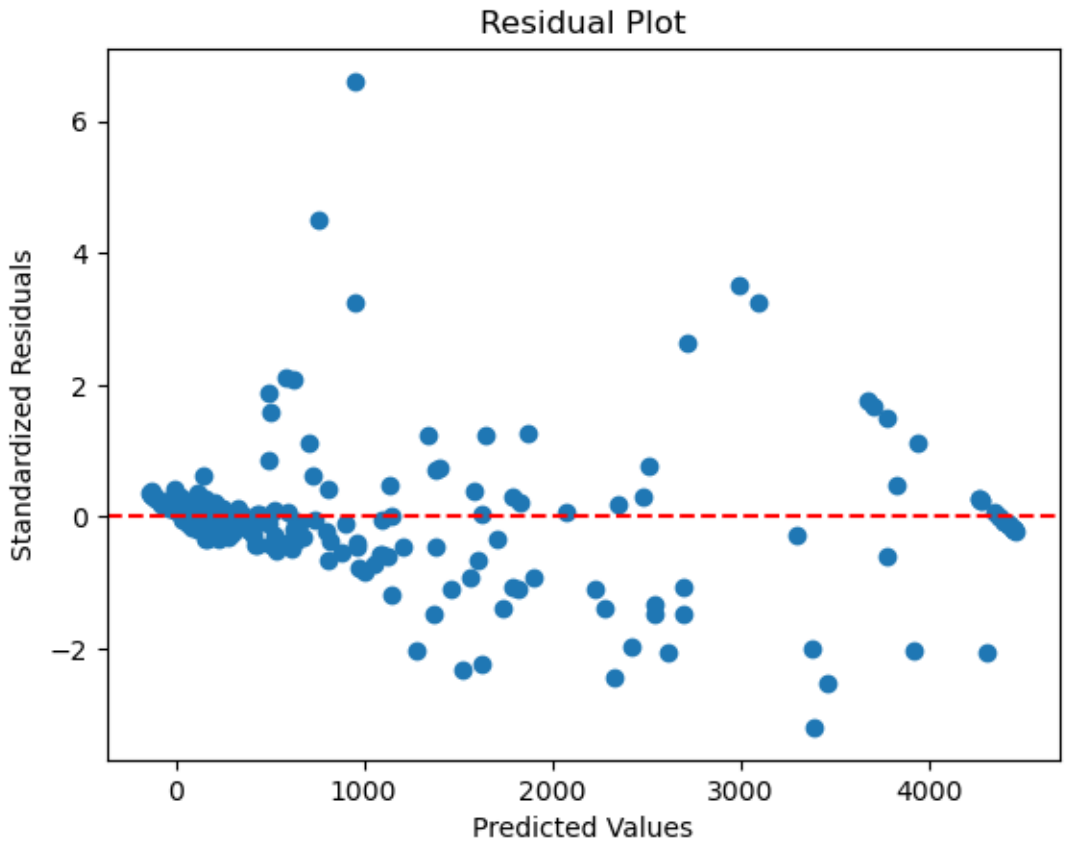


Fig1.3.7.2 Residual Plot OLS Stats Model

1. The residual points are symmetrically distributed, tending to cluster towards the middle of the plot.
2. They're clustered around the lower single digits of the y-axis.
3. In general, there aren't any clear patterns.
4. This is an ideal residual plot and is acceptable.

Note- Scatter plots and residual plots for all models have been plotted and all show similar result.

1.3.8 Model 7 – Linear Regression on Scaled Data

Model performance results are:

Coefficients: [[388.35931937 -46.37038134 81.16536349 642.56680013 167.87266108
-70.56912892 430.5139761 4.61330005]]
Intercept: [1192.78318419]

Training set performance:
RMSE: 394.337
R-squared: 0.936
MSE 155501.8006389884

Test set performance:
RMSE: 400.002
R-squared: 0.924
MSE 160001.69195429125

1.3.9 Model 8 – Linear Regresion on Scaled data- dropping features over VIF 5

Model performance results are:

Coefficients: [[21.6288429 429.0632531 2050.9331315 -246.32957 54.98948947]]

Intercept: [669.99885843]

Training set performance:
RMSE: 854.696

R-squared: 0.699

MSE 730504.8254082266

Test set performance:

RMSE: 948.881

R-squared: 0.573

MSE 900375.8143274568

1.3.10 Scikit learn Linear Models Discussion & Equation

Below is the discussion on Scikit learn's linear models:

1. As discussed above, and based on the Model 6,7,8 results, scaling does not necessary give drastic results in this dataset. The Model 6, 7 have similar or same performance result upto or above 5 decimal points.
2. The best linear model is Model 6 – Simple Linear regression without scaling.
3. Linear regression can be given by below equation (Model 6 – Simple Linear Regression):

   y = 0.40506319 * capital - 4.66215536 * patents + 0.63853807 * randd + 78.56497068 * employment + 167.87266108 * sp500 - 40.98329451 * tobinq + 0.24554026 * value + 0.21287606 * institutions + 73.73534908
4. Dropping features, blindly based on VIF value may not give better model. Infact, as seen in model 8, the model may show decrease in predicting performance.

1.3.11 Compare Scikit learn and OLS model – Model 5, Model 6

1. Scikit learn gives Test set performance of RMSE: 400.002, R-squared: 0.924
2. Stats model gives test set performance of RMSE: 394.38620894692014, R-squared: 0.92 61493138182268
3. Both models are strong models for regression. However, Stats model is slightly a better m odel compared to scikitlearn.

1.4 INFERENCE

Based on these predictions, what are the business insights and recommendations. (6 marks)

1. There might still be multicollinearity which needs to be addressed with business team.
2. Stats model is used for model prediction given equation y = (81.7013 * const) + (0.4118 * capital) + (80.6845 * employment) + (176.4655 * sp500) - (39.6549 * tobinq) + (0.2466 * value)
3. 'capital', 'employment', 'tobinq', 'value', 'sp500' are the best predictors of the target variable.
4. 'institutions', 'patents', 'raandd' do not show a significance value in predicting the target variable. However, this does not mean they are poor predictors.
5. Focus on Capital Investment - as this variable has a high positive impact on 'sales'
6. Monitor S&P 500 Performance: The coefficient of the "sp500" variable is 176.4655. This indicates that the performance of the S&P 500 index has a positive impact on the response variable. It is recommended to closely monitor the trends and movements of the S&P 500 index as it can potentially influence the 'sales' outcome.
7. 'tobinq' has an inverse relationship and the company should evaluate this variable carefully as this has a negative effect on the 'sales'.

2 PROBLEM 2: LOGISTIC REGRESSION, LDA, CART

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.

DATA DESCRIPTION

Data Dictionary for Car_Crash

Dataset for Problem 2: Car_Crash.csv

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 populations sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

2.1 EXPLORATORY DATA ANALYSIS

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. (8 marks)

Answer:
The data shows 11217 rows, 16 columns. 'Survived' is the target or feature of interest with 'survived' having higher value count and hence the positive class.

2.1.1 Sample Data

The Data provided has these top rows as shown below.

	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
0	0	55+	27.078	Not_Survived	none	none	1	m	32	1997	1987.0	unavail	driver	0	4.0	2:13:2
1	1	25-39	89.627	Not_Survived	airbag	belted	0	f	54	1997	1994.0	nodeploy	driver	0	4.0	2:17:1
2	2	55+	27.078	Not_Survived	none	belted	1	m	67	1997	1992.0	unavail	driver	0	4.0	2:79:1
3	3	55+	27.078	Not_Survived	none	belted	1	f	64	1997	1992.0	unavail	pass	0	4.0	2:79:1
4	4	55+	13.374	Not_Survived	none	none	1	m	23	1997	1986.0	unavail	driver	0	4.0	4:58:1

Fig2.1.1.1 Sample Crash Data top 5

The bottom five rows from the dataset are shown below.

	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
11212	11212	25-39	3179.688	survived	none	belted	1	m	17	2002	1985.0	unavail	driver	0	0.0	82:107:1
11213	11213	10-24	71.228	survived	airbag	belted	1	m	54	2002	2002.0	nodeploy	driver	0	2.0	82:108:2
11214	11214	10-24	10.474	survived	airbag	belted	1	f	27	2002	1990.0	deploy	driver	1	3.0	82:110:1
11215	11215	25-39	10.474	survived	airbag	belted	1	f	18	2002	1999.0	deploy	driver	1	0.0	82:110:2
11216	11216	25-39	10.474	survived	airbag	belted	1	m	17	2002	1999.0	deploy	pass	1	0.0	82:110:2

Fig2.1.1.2 Sample Crash Data bottom 5

The data contains 'Unnamed: 0', 'caseid' features which are dropped. The dataset top 5 observations are as shown below:

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity
0	55+	27.078	Not_Survived	none	none	1	m	32	1997	1987.0	unavail	driver	0	4.0
1	25-39	89.627	Not_Survived	airbag	belted	0	f	54	1997	1994.0	nodeploy	driver	0	4.0
2	55+	27.078	Not_Survived	none	belted	1	m	67	1997	1992.0	unavail	driver	0	4.0
3	55+	27.078	Not_Survived	none	belted	1	f	64	1997	1992.0	unavail	pass	0	4.0
4	55+	13.374	Not_Survived	none	none	1	m	23	1997	1986.0	unavail	driver	0	4.0

Fig2.1.1.3 Sample Crash Data top 5-(dropped unnamed, caseid)

The data contains many numerical and categorical columns. Further exploratory data analysis is as below.

2.1.2 Data Statistical Description

The data description of categorical and numerical is shown below.

	count	mean	std	min	25%	50%	75%	max
weight	11217.0	431.405309	1406.202941	0.0	28.292	82.195	324.056	31694.04
frontal	11217.0	0.644022	0.478830	0.0	0.000	1.000	1.000	1.00
ageOFocc	11217.0	37.427654	18.192429	16.0	22.000	33.000	48.000	97.00
yearacc	11217.0	2001.103236	1.056805	1997.0	2001.000	2001.000	2002.000	2002.00
yearVeh	11217.0	1994.177944	5.658704	1953.0	1991.000	1995.000	1999.000	2003.00
deploy	11217.0	0.389141	0.487577	0.0	0.000	0.000	1.000	1.00
injSeverity	11140.0	1.825583	1.378535	0.0	1.000	2.000	3.000	5.00

Fig2.1.2.1 Crash Data Description 1

	count	unique	top	freq
dvcat	11217	5	10-24	5414
Survived	11217	2	survived	10037
airbag	11217	2	airbag	7064
seatbelt	11217	2	belted	7849
sex	11217	2	m	6048
abcat	11217	3	deploy	4365
occRole	11217	2	driver	8786

Fig2.1.2.2 Crash Data Description 2

2.1.3 Missing Values

Missing values:

Below information shows the missing values in the data.

```
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
dtype: int64
```

Fig2.1.3.1 Crash Data info-Missing values treated

There are 77 missing values in 'injSeverity' column. Below is boxplot of 'injSeverity':

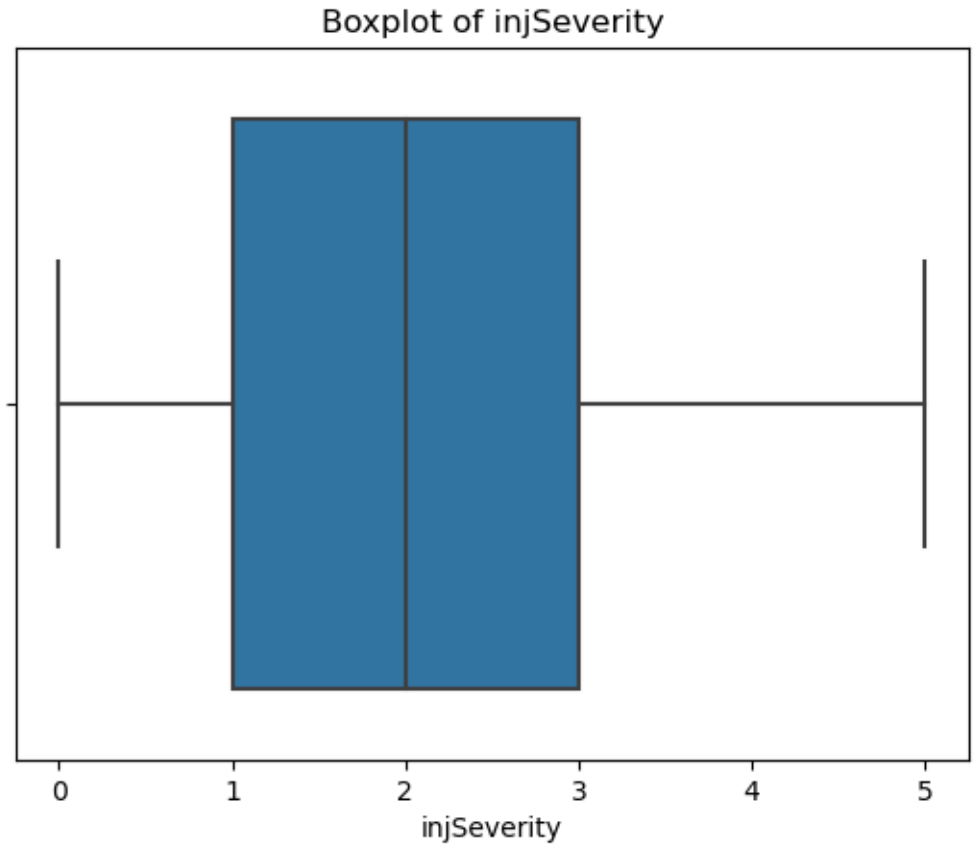
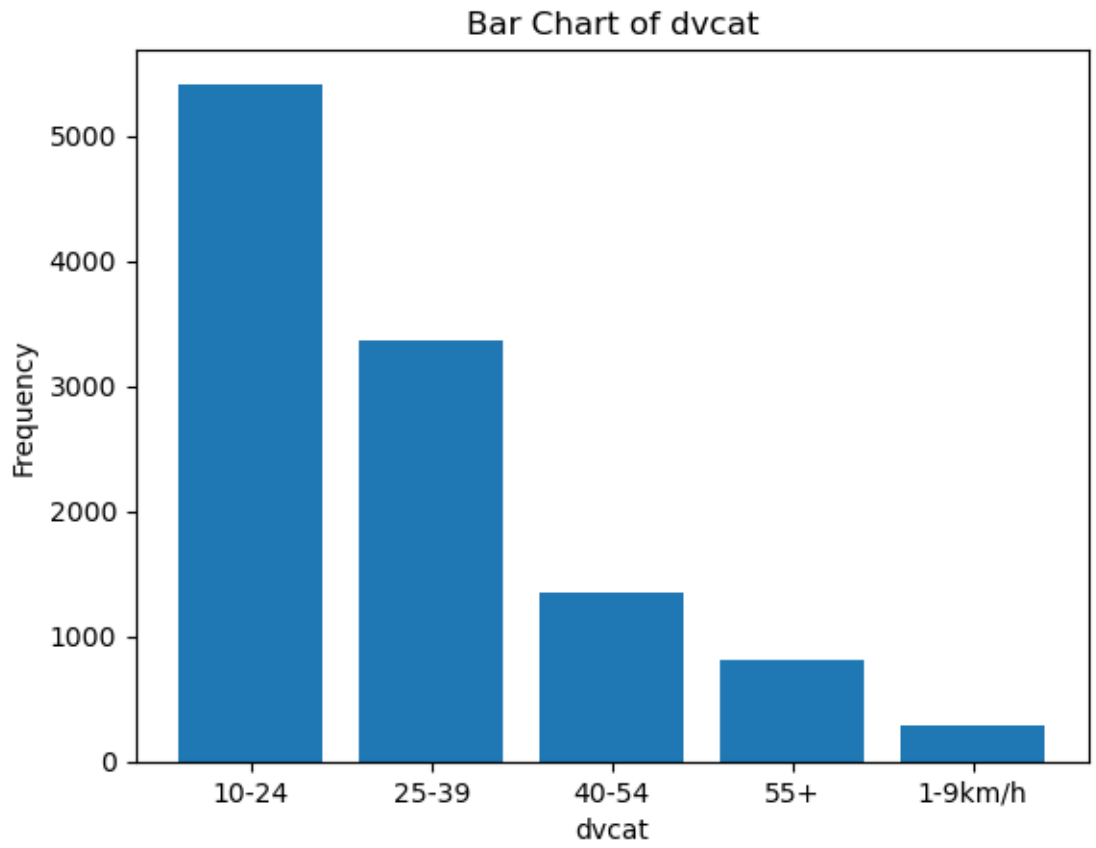


Fig2.1.3.2 Boxplot showing 'injSeverity' skewedness.

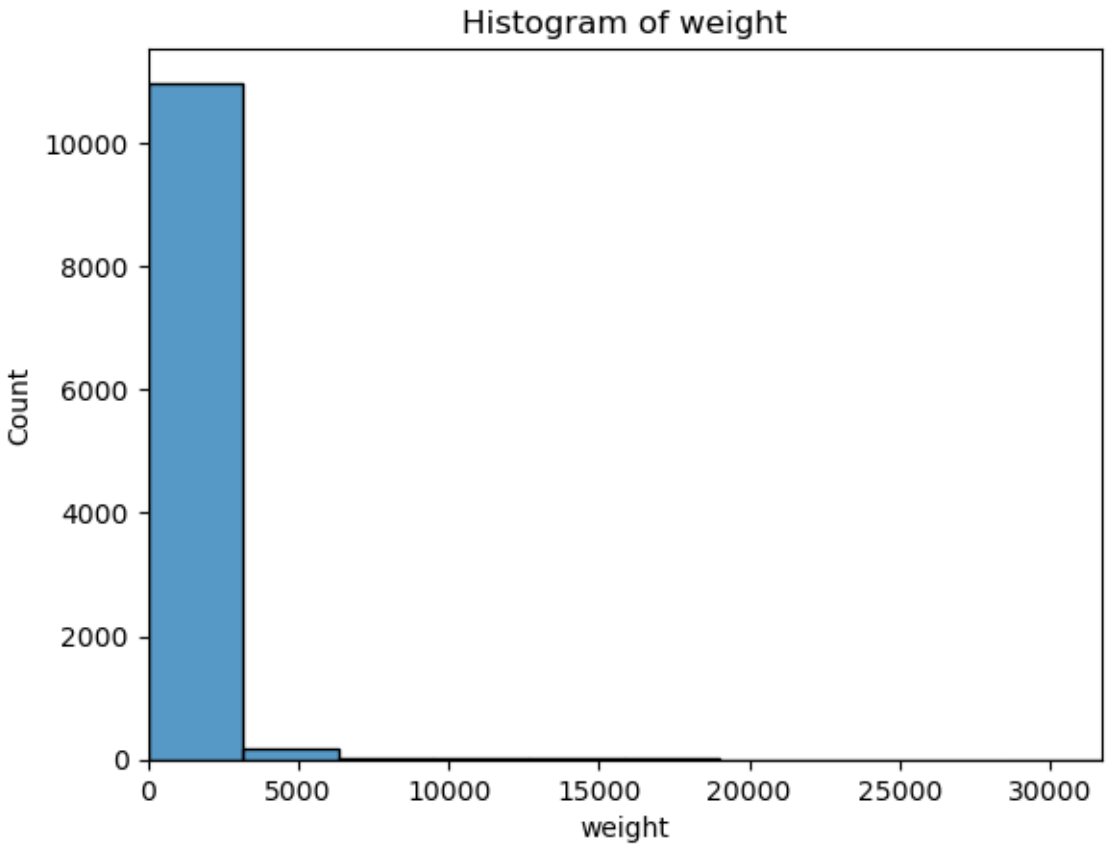
the 'injSeverity' does not have outliers. Data is very lightly right skewed. Hence, the missing values can be imputed based on median. Note- Mean can also be used as there is no high skewedness in data.

2.1.4 Univariate Analysis

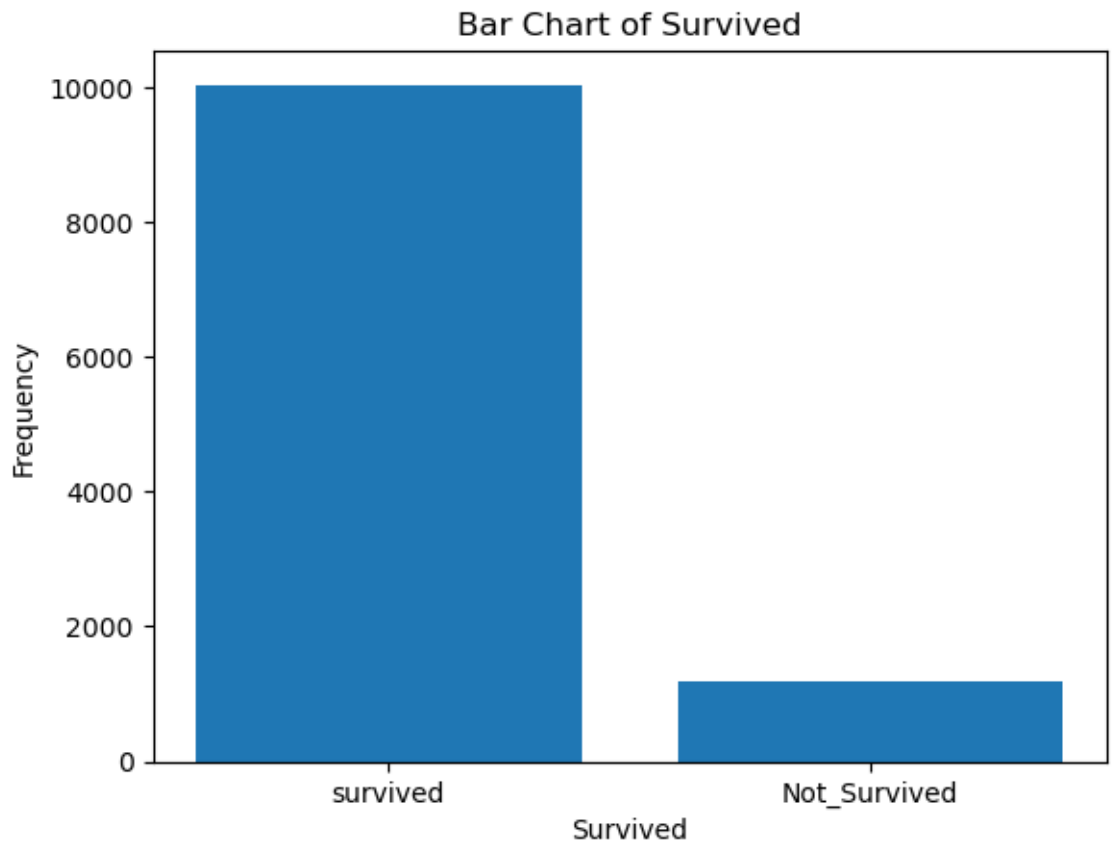
The univariate analysis is as shown below using histogram plots of each features:



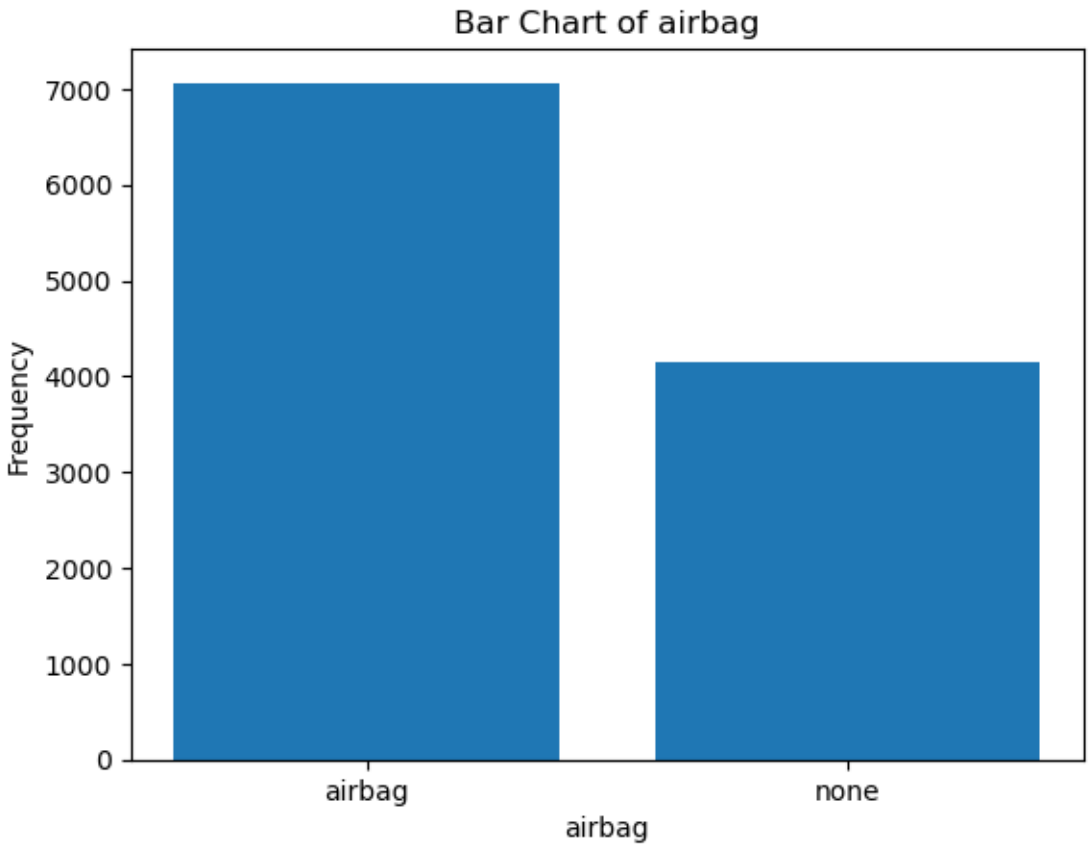
The dvcat is highly right skewed indicating outliers presence.

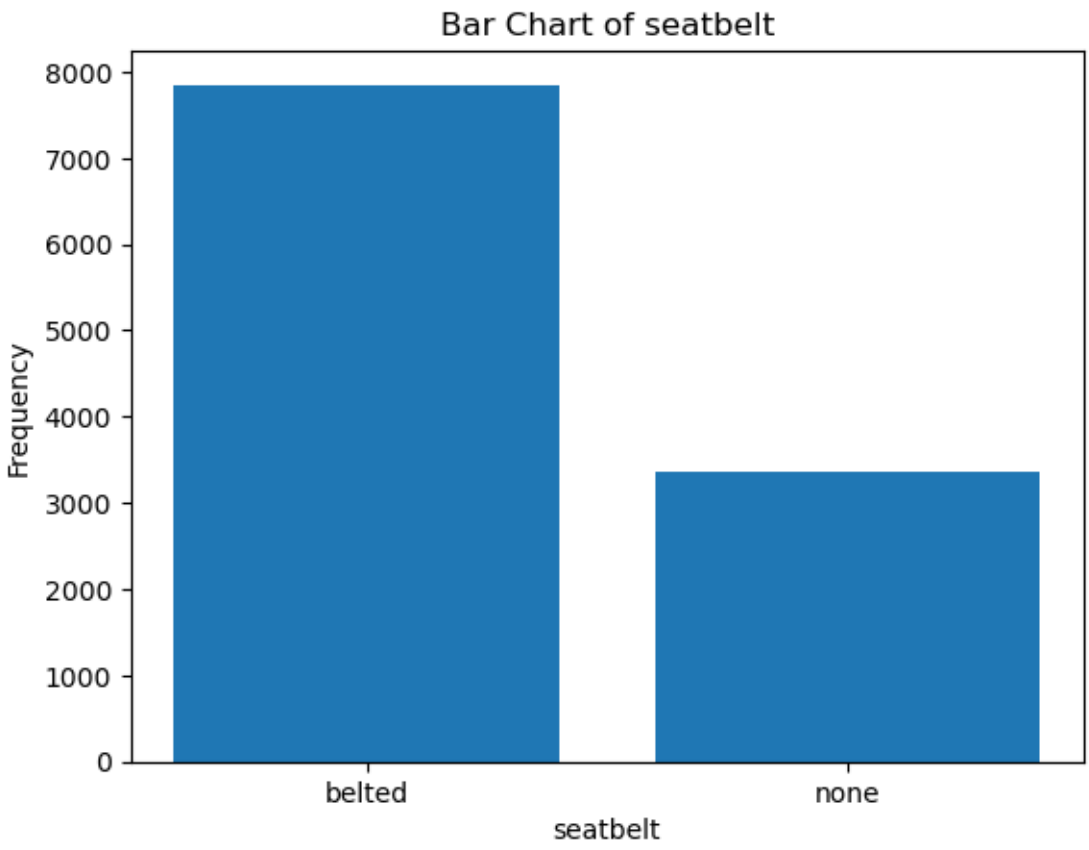


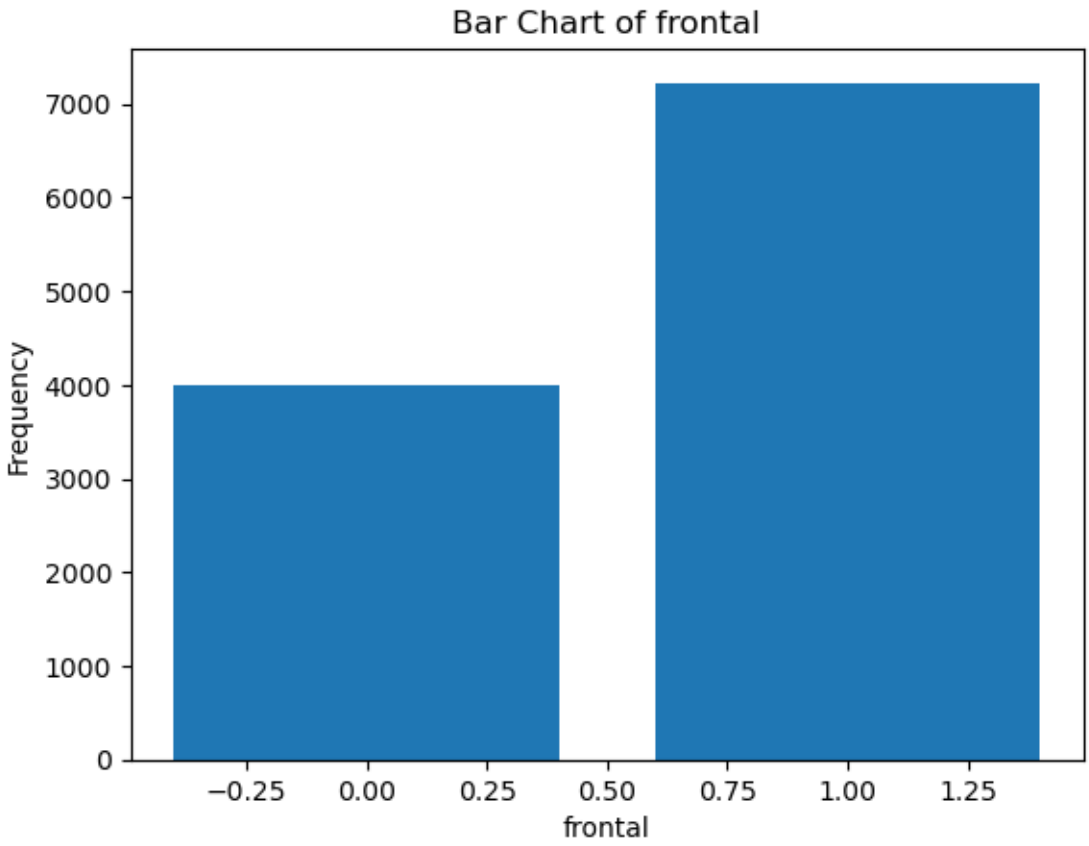
The ‘weight’ is highly right skewed indicating outliers presence.

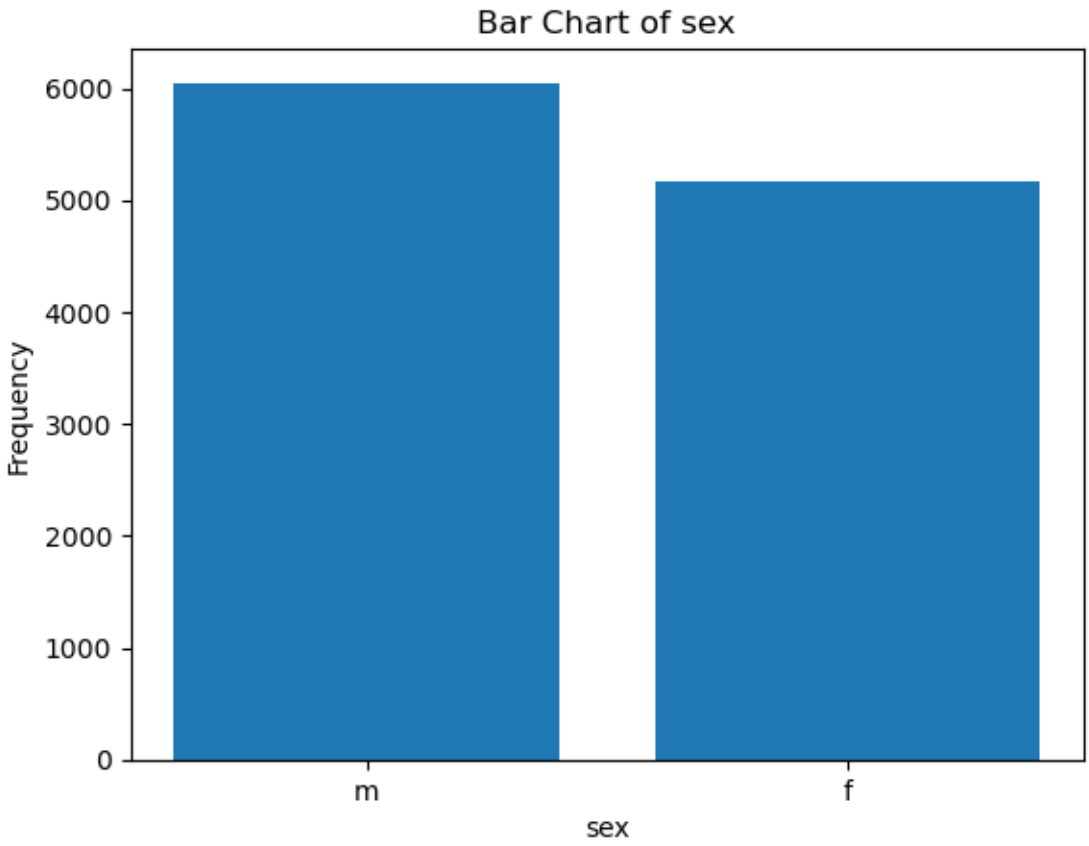


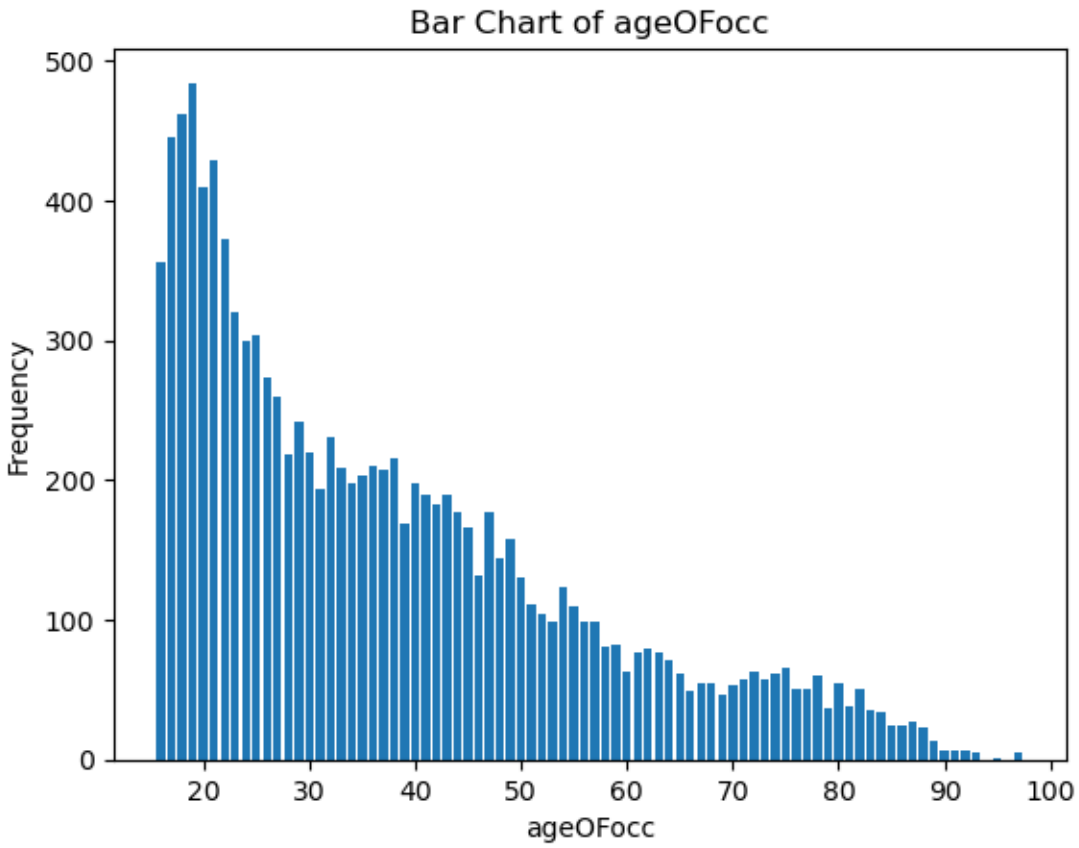
'Survived' i.e. the target variable has much higher counts for 'Survived' than the opposite and while splitting data stratify parameter shall be used.


Bar Chart of airbag
7000
6000
5000
4000
3000
2000
1000
0
Frequency
airbag
none
airbag

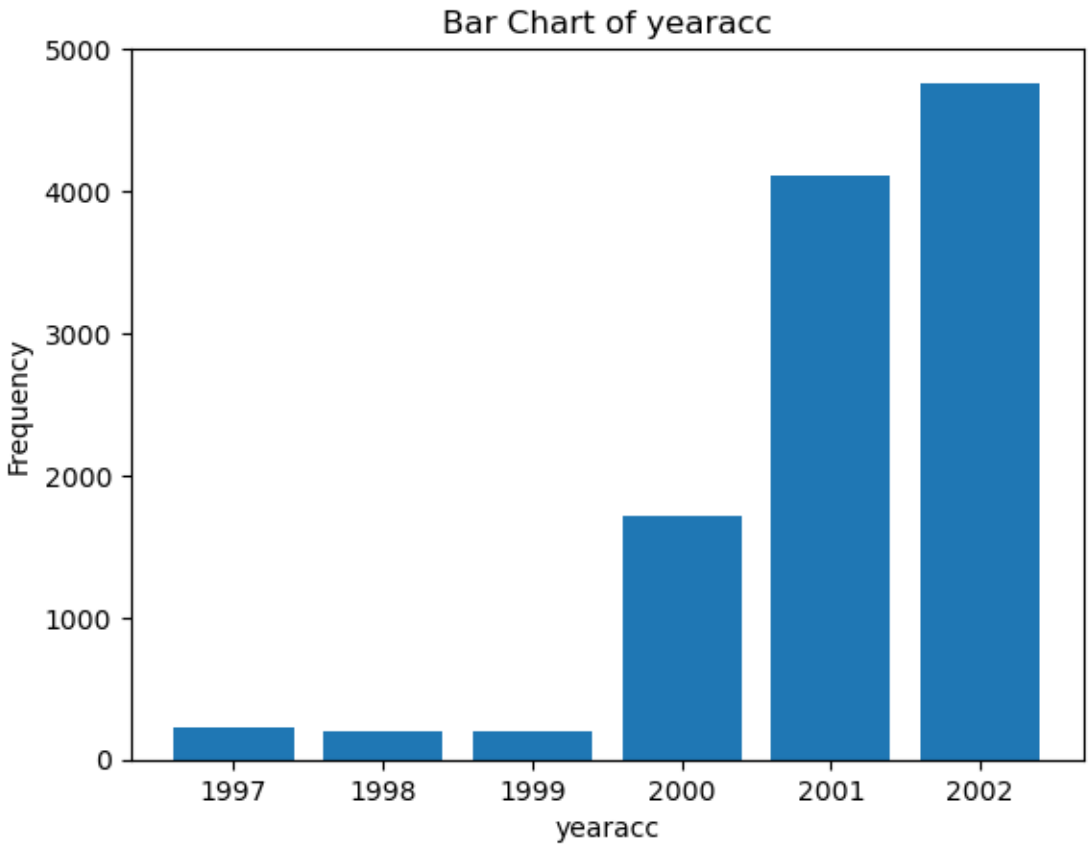

Bar Chart of seatbelt
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency
belted
none
seatbelt


Bar Chart of frontal
Frequency
7000
6000
5000
4000
3000
2000
1000
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
frontal

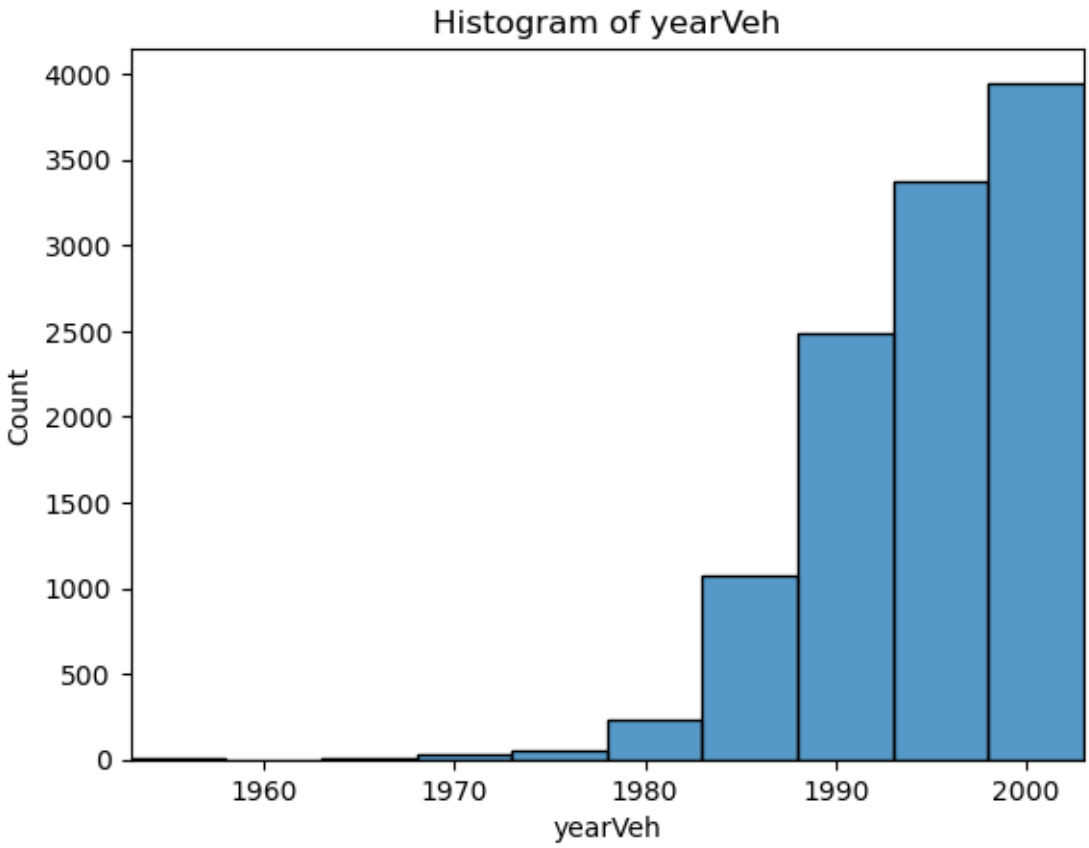

Bar Chart of sex
Frequency
6000
5000
4000
3000
2000
1000
0
m
f
sex



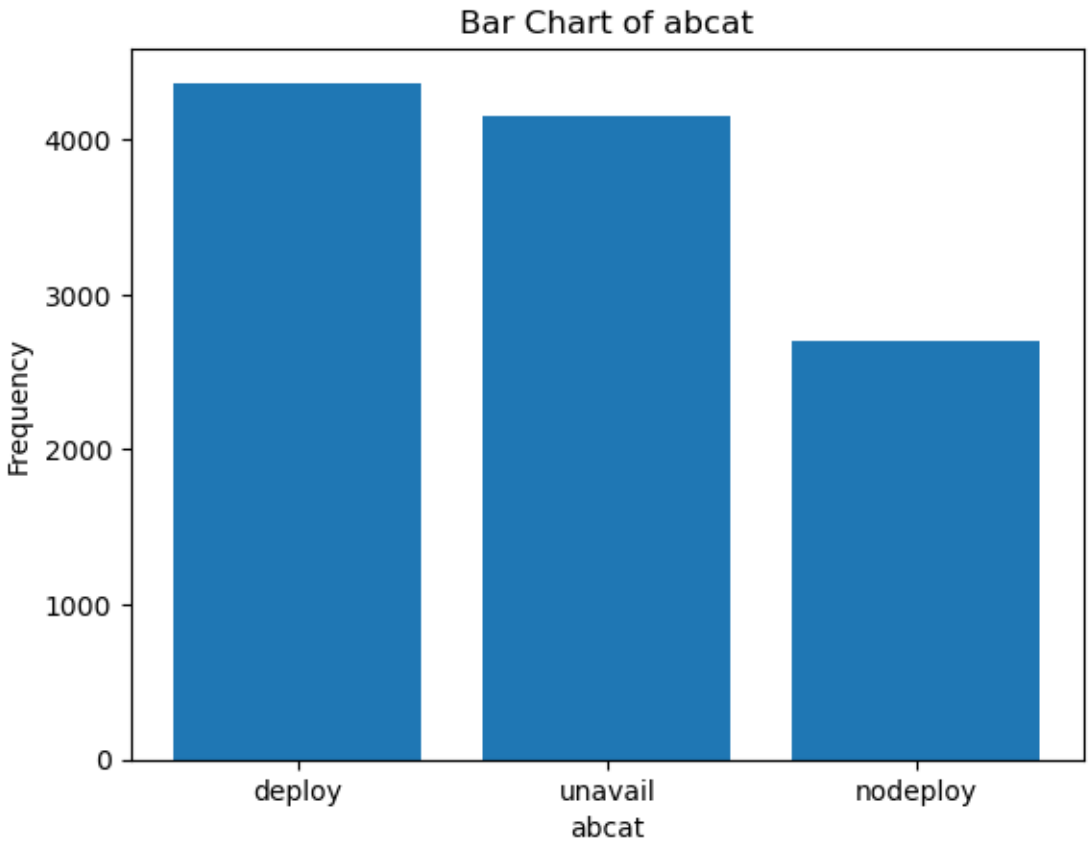
The ‘ageOfOcc’ has a peak in ’10-24’ age group. The data is also highly right skewed indicating the presence of ‘outliers’. Binning of age groups can help in this feature.

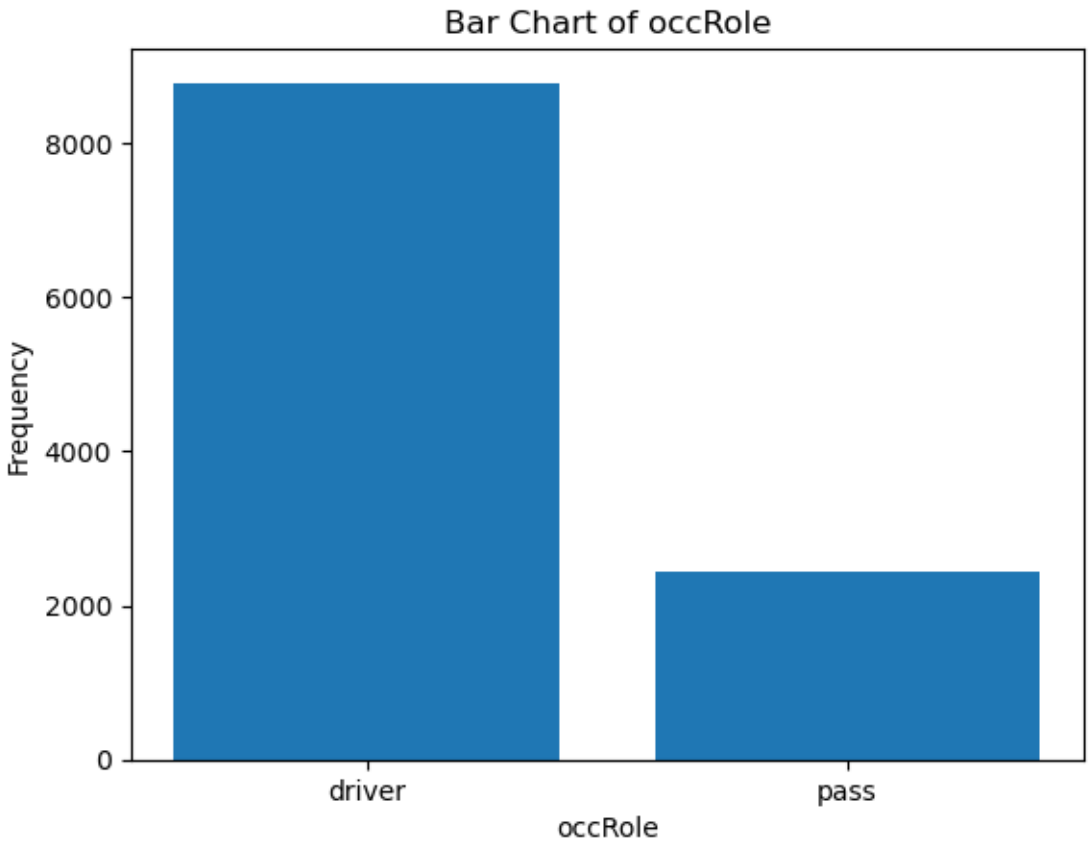


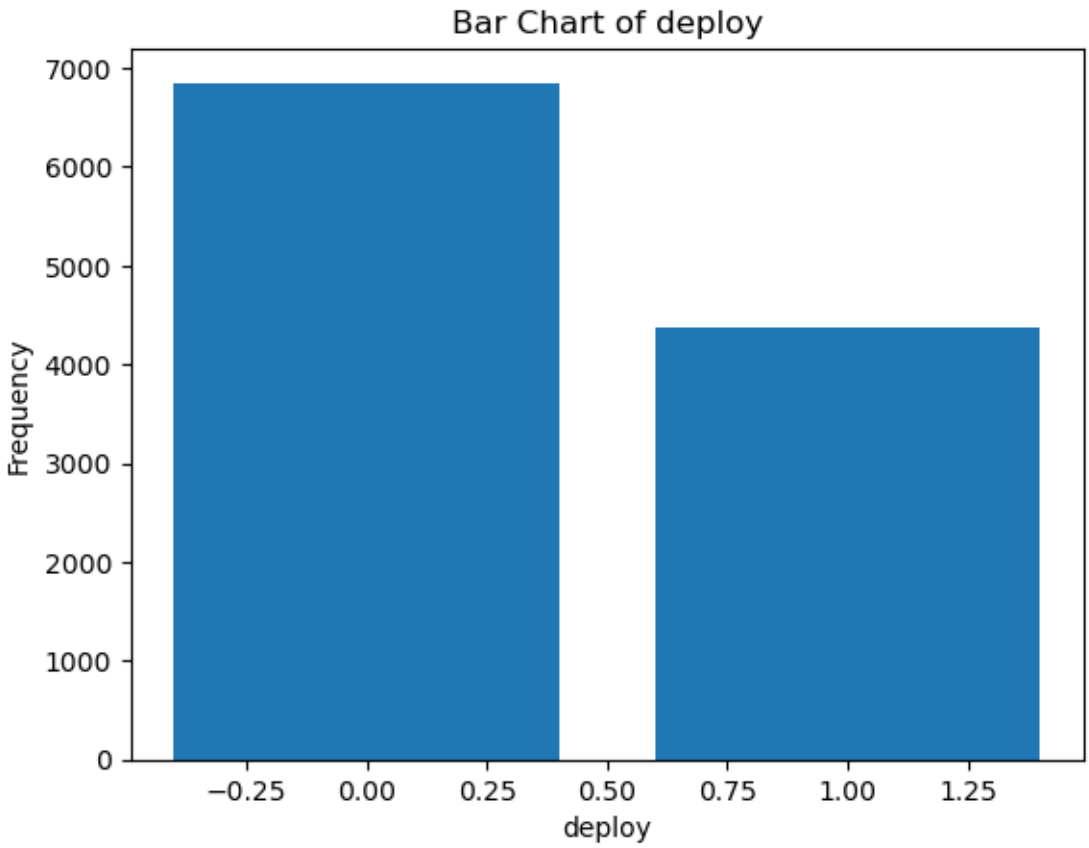
There seem to a lot more accidents in years after 2000. This can be due to many reasons that there are higher sales in cars/a boom in teenage car ownership/ records being maintained after these years. More analysis is needed to bring out more insights in this feature. Additionally, binning can help this feature as well.



'yearVeh' has highly left skewed data with most cars being from years after 1990. It can be due to increase of availability of these cars in the market from manufacturing point of view. More analysis is needed to bring out more insights in this feature. Additionally, binning can help this feature as well.


Bar Chart of abcat
Frequency
4000
3000
2000
1000
0
deploy
unavail
nodeploy
abcat


Bar Chart of occRole
8000
6000
4000
2000
0
Frequency
driver
pass
occRole


Bar Chart of deploy
Frequency
7000
6000
5000
4000
3000
2000
1000
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
deploy

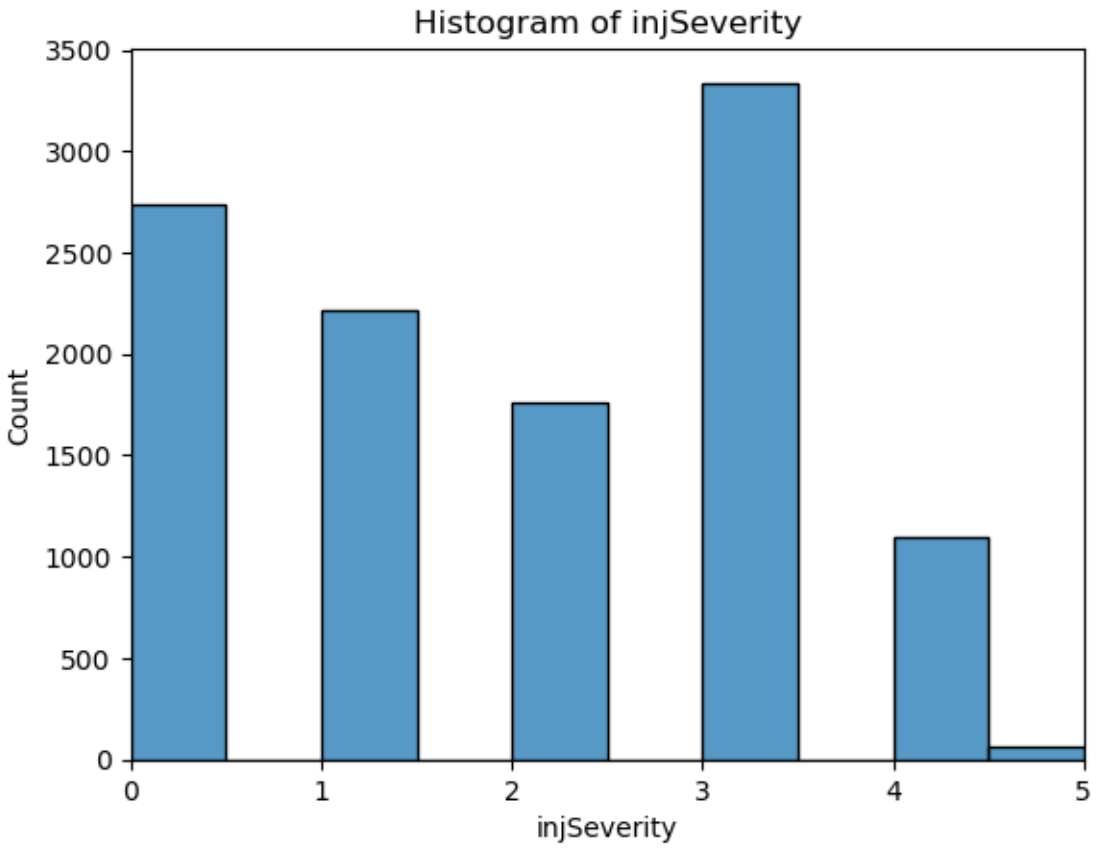


Fig2.1.4 histograms & Countplots of Crash data

The 'injSeverity' of 5 is very less and that can explain the high 'NotSurvived' count. More insights need to be drawn.

2.1.5 Bivariate Analysis

Using calculation, we can find that The vehicles from 1990-2000 are involved in 61% and vehicles after 2000 account for 18% of the crashes.

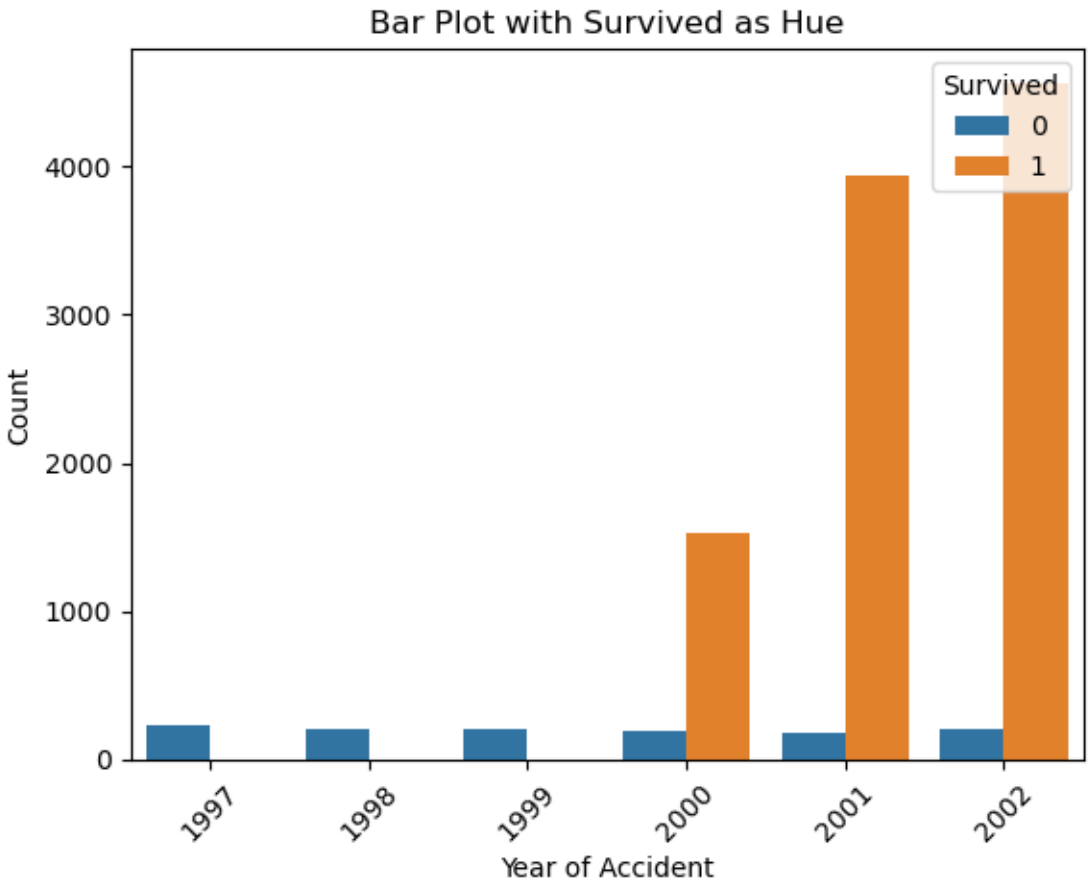


Fig2.1.5.1 Year Of Accident v Survival

The above plot shows a clear distinction that accidents before 2000s resulted only in 'NotSurvived'. Based on this information, the year of accident is encoded into binary variable i.e. before 2000s as 0, after 2000s as 1

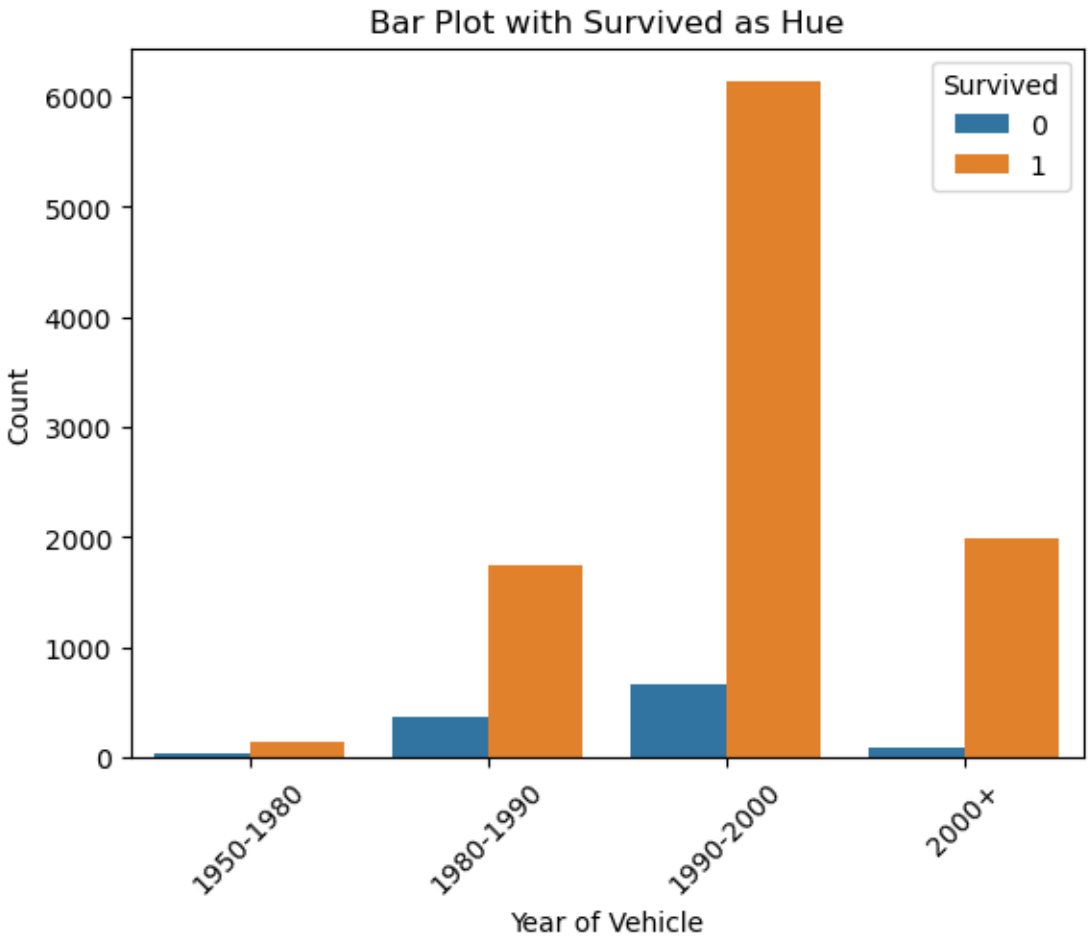


Fig2.1.5.2 Year of Vehicle on barplot with Survived as hue

Above plot shows that Compared to older version cars, passengers/drivers using latest version cars have drastically higher chance of survival. This can be attributed to the airbag facilities in latest cars

Bivariate analysis using pairplot is shown below:

Fig 2.1.5.3 Pairplot of Crash data

It can be seen from the pairplot as well that accidents before 2000s resulted only in 'NotSurvived'.The classes are not well separated as peaks are engulfed by other class.

2.1.6 Encode data

The datatype of each feature is shown below:

```
 #   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
 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
dtypes: float64(3), int64(4), object(7)
```

Fig 2.1.6.1 Crash data info after Encoding.

Inference: Out of 14 columns 7 are type object, rest are either of type float or integer. Most of the existing *numerical* columns are of type float64, int64 already.

'frontal', 'deploy' are already encoded into nominal. dvcat', 'Survived', 'airbag', 'seatbelt', 'sex', 'abcat', 'occRole', 'injSeverity' are the categorical columns which should be encoded to nominal and are encoded to nominal by using data dictionary for few variables and by assigning a value to each category.

'yearVeh', 'injSeverity' are converted to type 'int64'.

The top 5 observations after encoding are shown below.

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity
0	5	27.078	0	0	0	1	0	32	1997	1987	0	1	0	4
1	3	89.627	0	1	1	0	1	54	1997	1994	0	1	0	4
2	5	27.078	0	0	1	1	0	67	1997	1992	0	1	0	4
3	5	27.078	0	0	1	1	1	64	1997	1992	0	0	0	4
4	5	13.374	0	0	0	1	0	23	1997	1986	0	1	0	4

Fig2.1.6.2 Crash data sample observation after Encoding.

Data description is shown below:

	count	mean	std	min	25%	50%	75%	max
dvcat	11217.0	2.731122	0.958810	1.0	2.000	2.000	3.000	5.00
weight	11217.0	431.405309	1406.202941	0.0	28.292	82.195	324.056	31694.04
Survived	11217.0	0.894803	0.306821	0.0	1.000	1.000	1.000	1.00
airbag	11217.0	0.629758	0.482891	0.0	0.000	1.000	1.000	1.00
seatbelt	11217.0	0.699741	0.458391	0.0	0.000	1.000	1.000	1.00
frontal	11217.0	0.644022	0.478830	0.0	0.000	1.000	1.000	1.00
sex	11217.0	0.460818	0.498485	0.0	0.000	0.000	1.000	1.00
ageOFocc	11217.0	37.427654	18.192429	16.0	22.000	33.000	48.000	97.00
yearacc	11217.0	2001.103236	1.056805	1997.0	2001.000	2001.000	2002.000	2002.00
yearVeh	11217.0	1994.177944	5.658704	1953.0	1991.000	1995.000	1999.000	2003.00
abcat	11217.0	0.389141	0.487577	0.0	0.000	0.000	1.000	1.00
deploy	11217.0	0.389141	0.487577	0.0	0.000	0.000	1.000	1.00
injSeverity	11217.0	1.826781	1.373871	0.0	1.000	2.000	3.000	5.00

Fig 2.1.6.3 Crash data statistical description after Encoding

At first look, Standard of living is high, Contraceptive used is Yes. Need to perform regression to understand more insights from the dataset.

2.1.7 Outliers

Below are the outliers after encoding the data.

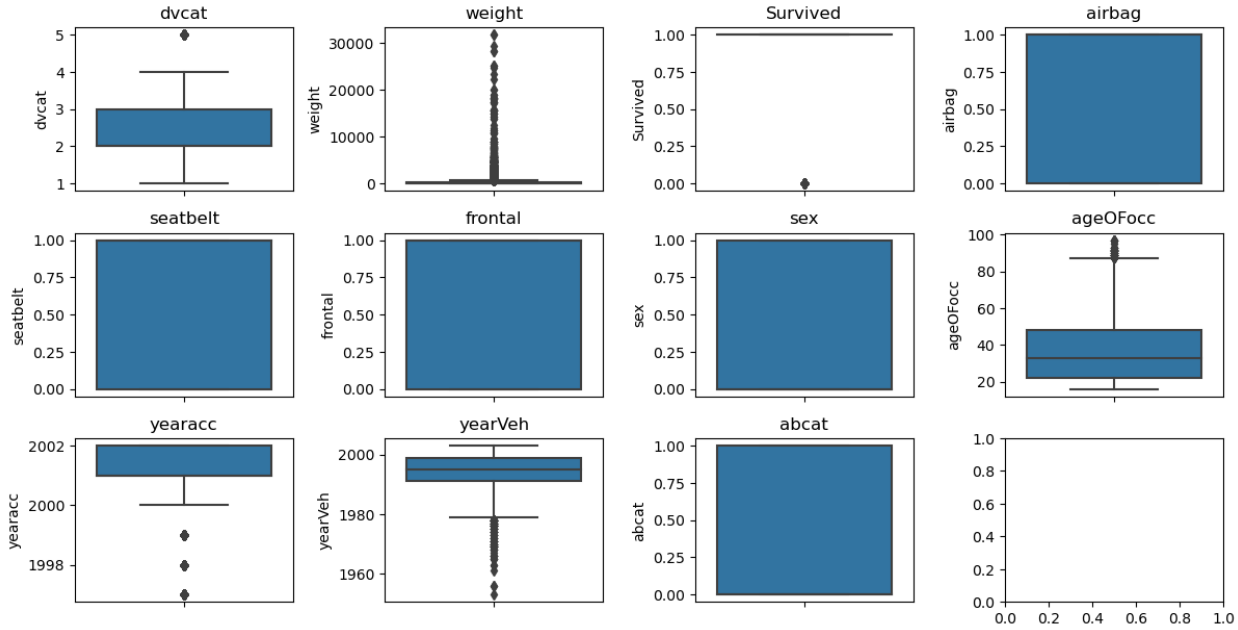


Fig2.1.7.1 Boxplot showing Crash Data outliers

Highest outliers are likely present in 'weight'. The outliers in 'ageOFocc', 'yearacc', 'yearVeh' shall be ignored for now as these are very small number of outliers and also outliers shown in 'Survived' column are not outliers.

Outliers in the 'dvcat', 'weight' features are treated by adjusting them to the lower and upper bound values calculated by the IQR value. Below plot shows the final outliers.

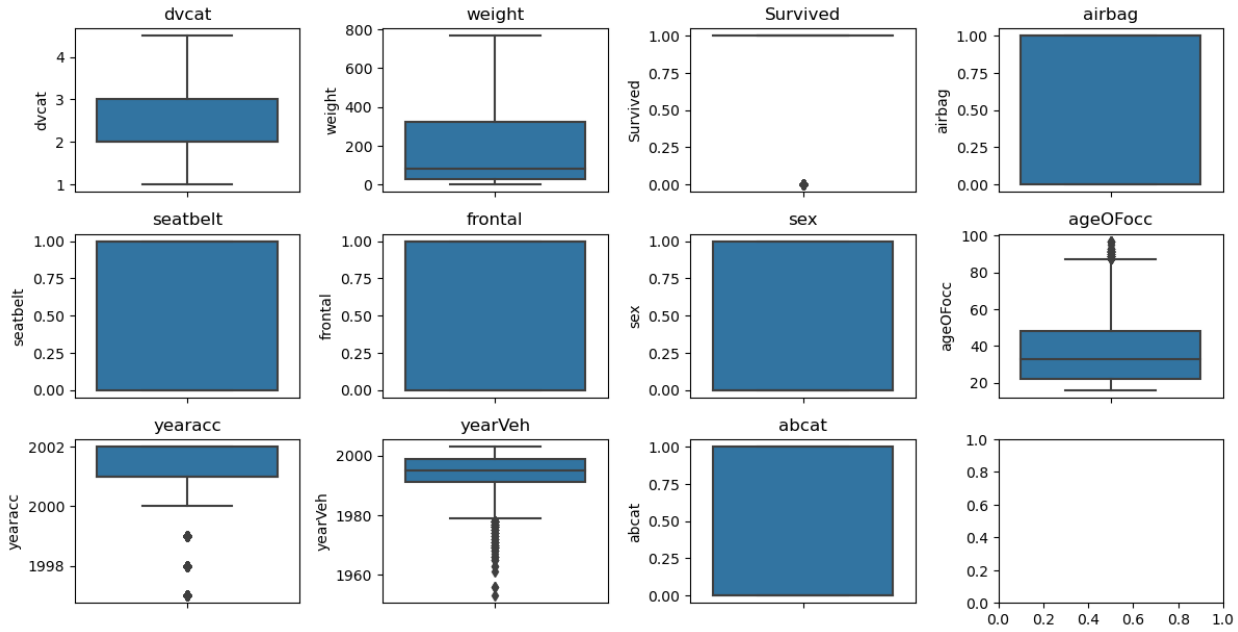


Fig2.1.7.2 Boxplot showing Crash Data after treating outliers

2.2 PREPARE DATA & CREATE MODELS

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). (8 marks)

Answer:

2.2.1 Encode data

Encoding is performed in above section. The top 5 observations after encoding are shown below.

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity
0	5	27.078	0	0	0	1	0	32	1997	1987	0	1	0	4
1	3	89.627	0	1	1	0	1	54	1997	1994	0	1	0	4
2	5	27.078	0	0	1	1	0	67	1997	1992	0	1	0	4
3	5	27.078	0	0	1	1	1	64	1997	1992	0	0	0	4
4	5	13.374	0	0	0	1	0	23	1997	1986	0	1	0	4

Fig 2.2.1.1 Crash data info after Encoding.

2.2.2 Co-relation

The heatmap is plotted to show the correlation between variables.

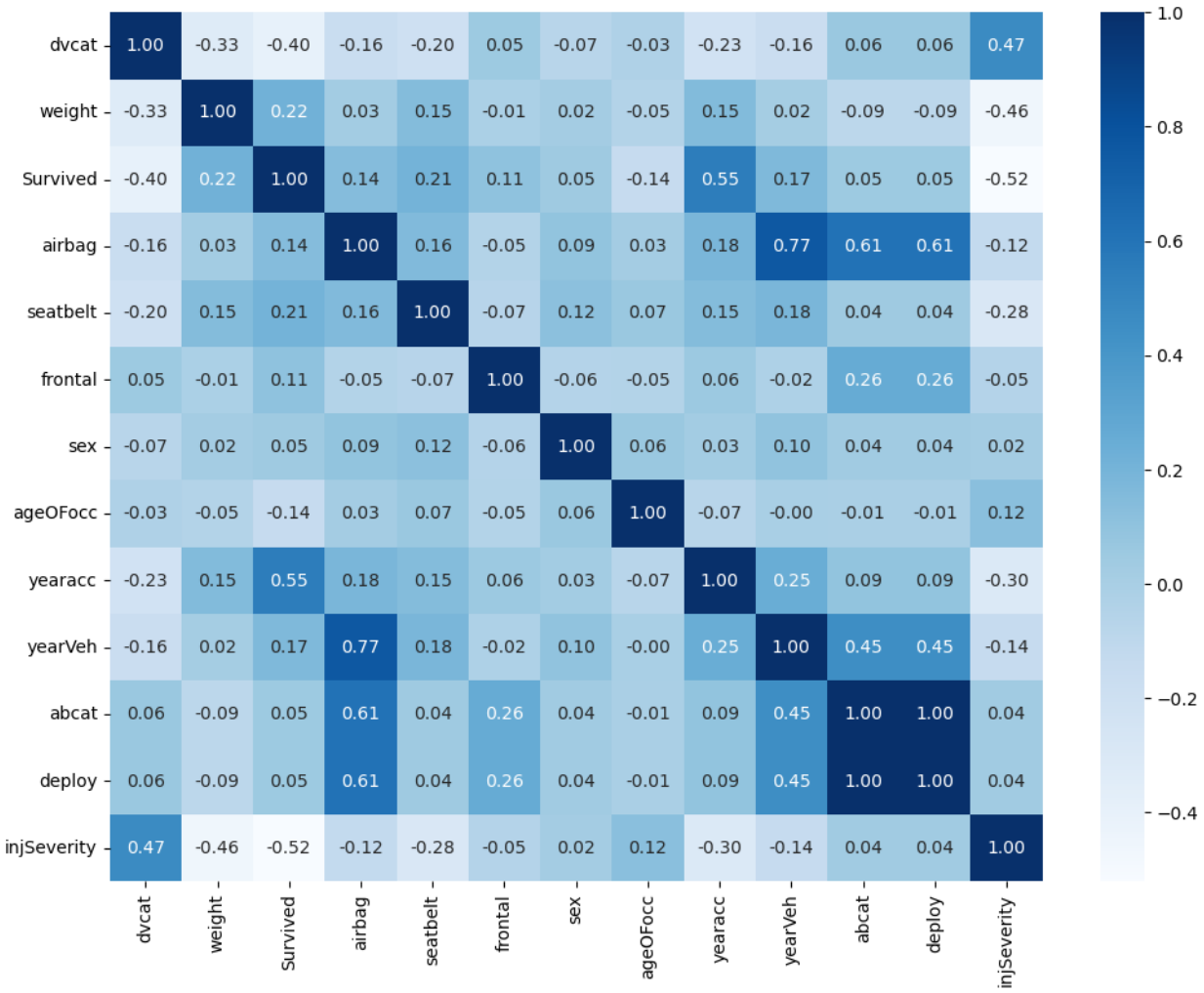


Fig2.2.2.1 HeatMap showing Crash data correlation.

1. 'Survived' shows a negative correlation with 'dvcat' (-0.398861), indicating that higher levels of impact speed category are associated with a lower likelihood of survival.

2. 'Survived' has a positive correlation with 'weight' (0.217553), suggesting that heavier vehicles are more likely to have passengers who survived.

3. 'Survived' also has a positive correlation with 'yearacc' (0.549885), indicating that accidents in more recent years are associated with a higher likelihood of survival.

4. 'Survived' shows a positive correlation with 'abcat' (0.054346) and 'deploy' (0.054346), suggesting that the presence of an airbag and its deployment are associated with a slightly higher likelihood of survival.

5. 'Survived' exhibits a negative correlation with 'injSeverity' (-0.520610), implying that higher injury severity is associated with a lower likelihood of survival.

6. 'yearVeh' shows a relatively strong positive correlation with 'airbag' (0.766181) and 'deploy' (0.611983), indicating that newer vehicles are more likely to have airbags and their deployment.

7. 'weight' has a negative correlation with 'injSeverity' (-0.455644), implying that heavier vehicles are associated with lower injury severity.

8. 'injSeverity' shows a positive correlation with 'dvcat' (0.470624), indicating that higher levels of impact speed category are associated with higher injury severity.

These insights provide a glimpse into the relationships between different variables in the dataset. However, it's important to note that correlation does not imply causation, and further analysis and domain knowledge are necessary to draw meaningful conclusions.

2.2.3 Split Data

Using sklearn train test split method, the dataset is split into 70 30 for training and testing using random state 1 to replicate results, stratify on 'Survived' variable. This will ensure a balance in the data and will not cause biasness while Training or Testing the mode.

The sample train dataset is shown below:

	dvcat	weight	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity
7778	2.0	207.690	1	1	1	1	36	2002	1999	1	0	1	1
9606	3.0	155.956	1	0	1	1	20	2002	2001	1	1	1	1
1085	4.0	9.780	0	1	1	1	22	2000	1991	0	1	0	3
9197	2.0	65.896	1	1	0	0	18	2002	2000	0	1	0	0
5698	3.0	9.285	0	0	1	0	59	2001	1991	0	1	0	3

Fig2.2.3.1 Crash data- Train Data sample

The sample test dataset is shown below:

	dvcat	weight	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity
4985	2.0	49.959	0	1	1	1	75	2001	1994	0	0	0	3
2442	4.5	69.484	0	1	1	0	43	2001	1988	0	1	0	2
7358	2.0	19.866	1	0	1	1	55	2002	2001	0	1	0	3
8396	1.0	53.227	1	1	1	1	80	2002	1998	0	0	0	2
3601	2.0	31.469	1	1	0	0	31	2001	1995	1	1	1	3

Fig2.2.3.2 Crash data- Test Data sample

	dvcat	weight	airbag	seatbelt	frontal	sex	abcat	occRole	deploy	injSeverity	yearAcc_Bin	yearVeh_bin	ageOFocc_bin
7778	2.0	207.690	1	1	1	1	1	0	1	1	1	3	3
9606	3.0	155.956	1	0	1	1	1	1	1	1	1	4	1
1085	4.0	9.780	0	1	1	1	0	1	0	3	1	3	1
9197	2.0	65.896	1	1	0	0	0	1	0	0	1	4	1
5698	3.0	9.285	0	0	1	0	0	1	0	3	1	3	4

Fig2.2.3.3 Crash data train sample after binning

	dvcat	weight	airbag	seatbelt	frontal	sex	abcat	occRole	deploy	injSeverity	yearAcc_Bin	yearVeh_bin	ageOFocc_bin
4985	2.0	49.959	0	1	1	1	0	0	0	3	1	3	4
2442	4.5	69.484	0	1	1	0	0	1	0	2	1	2	3
7358	2.0	19.866	1	0	1	1	0	1	0	3	1	4	4
8396	1.0	53.227	1	1	1	1	0	0	0	2	1	3	4
3601	2.0	31.469	1	1	0	0	1	1	1	3	1	3	2

Fig2.2.3.4 Crash data test sample after binning

However, note that the dataset has some degree of corelation & multi collinearity as shown in heatmaps in above section.

2.2.4 Model 1 – Decision Tree

Decision tree model is built using criterion='gini', max_depth=5, random_state=0 hyperparameters. Decision trees model is as below:

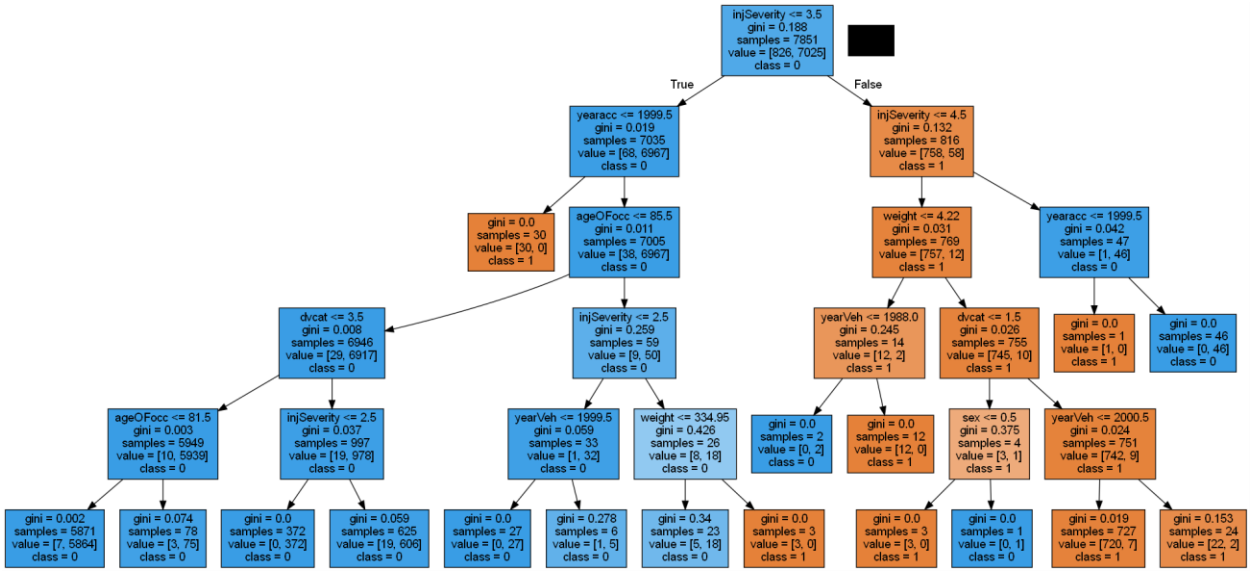


Fig2.2.4.1 Model -1 Decision Tree basic

occRole	0.001082
abcat	0.001437
deploy	0.001581
seatbelt	0.001953
airbag	0.002816
ageOFocc_bin	0.003249
sex	0.003514
yearVeh_bin	0.003794
frontal	0.004831
dvcat	0.005695
weight	0.035250
yearAcc_Bin	0.041527
injSeverity	0.893271

Fig2.2.4.2 Importance of Features

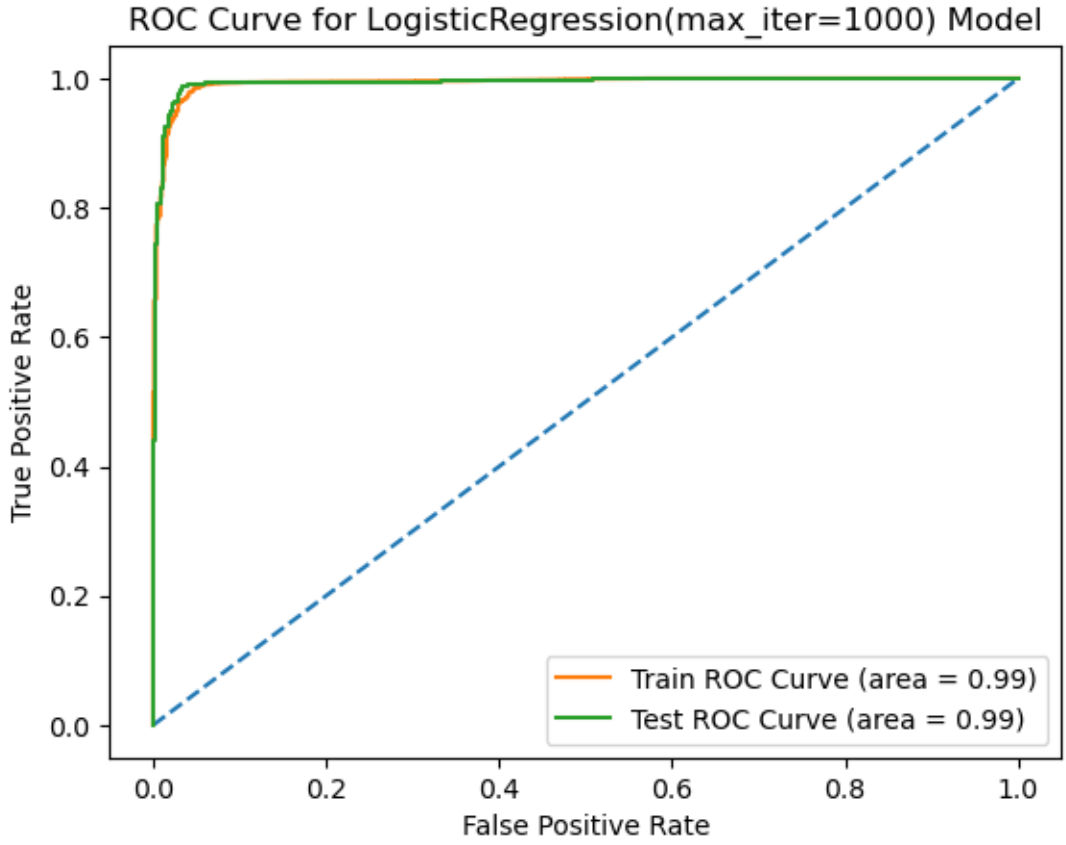
The injSeverity, year of accident, weight, impact speed are the important features in determining the 'Survived' class.

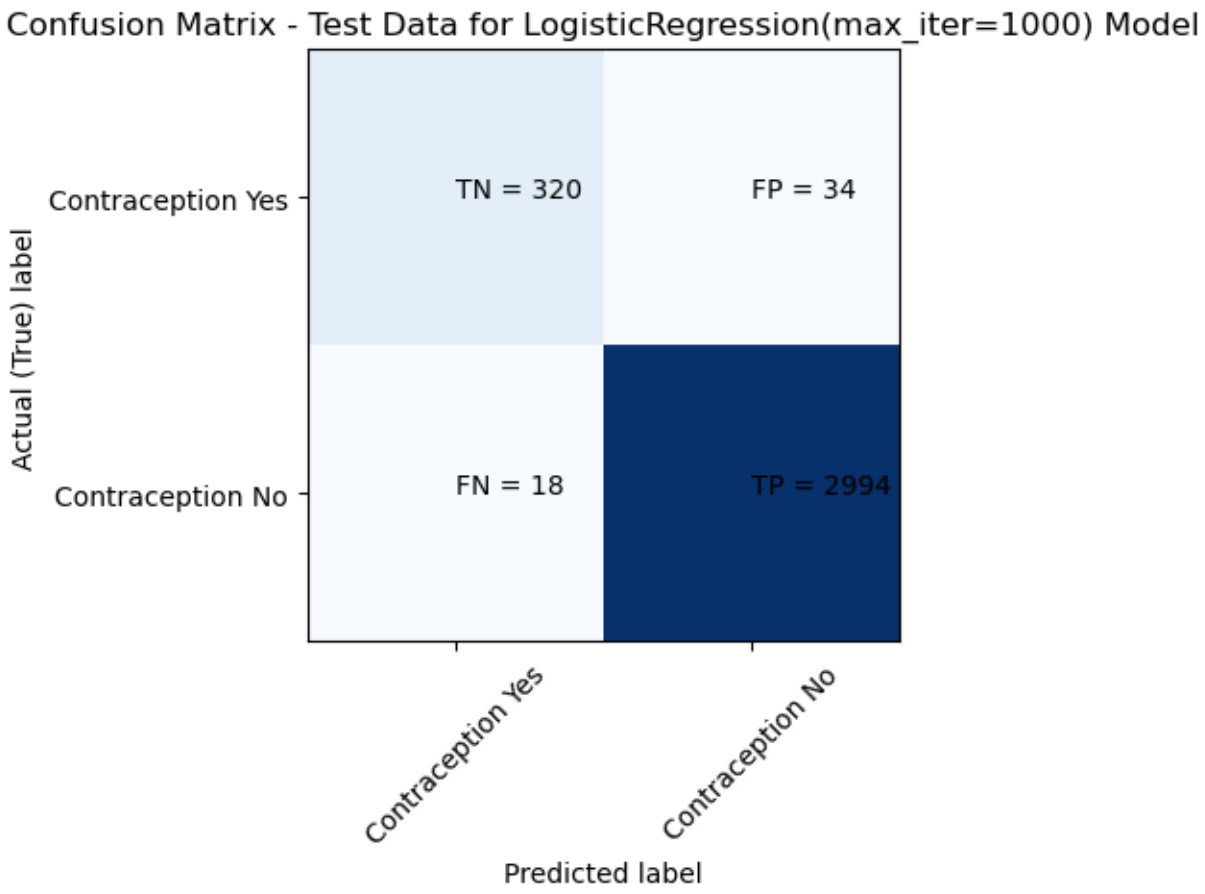
2.2.5 Model 2 – Apply Logistic, LDA, Cart

Logistic Regression, LDA, CART are applied on the dataset using training and performance is checked using the test data.

Hyper parameters used for the models are max_iter=1000 in Logistic regression, n_components=1 in LDA, criterion='gini' in CART.

2.2.6 Model 3 – Apply Logistic, LDA, CART on binned data





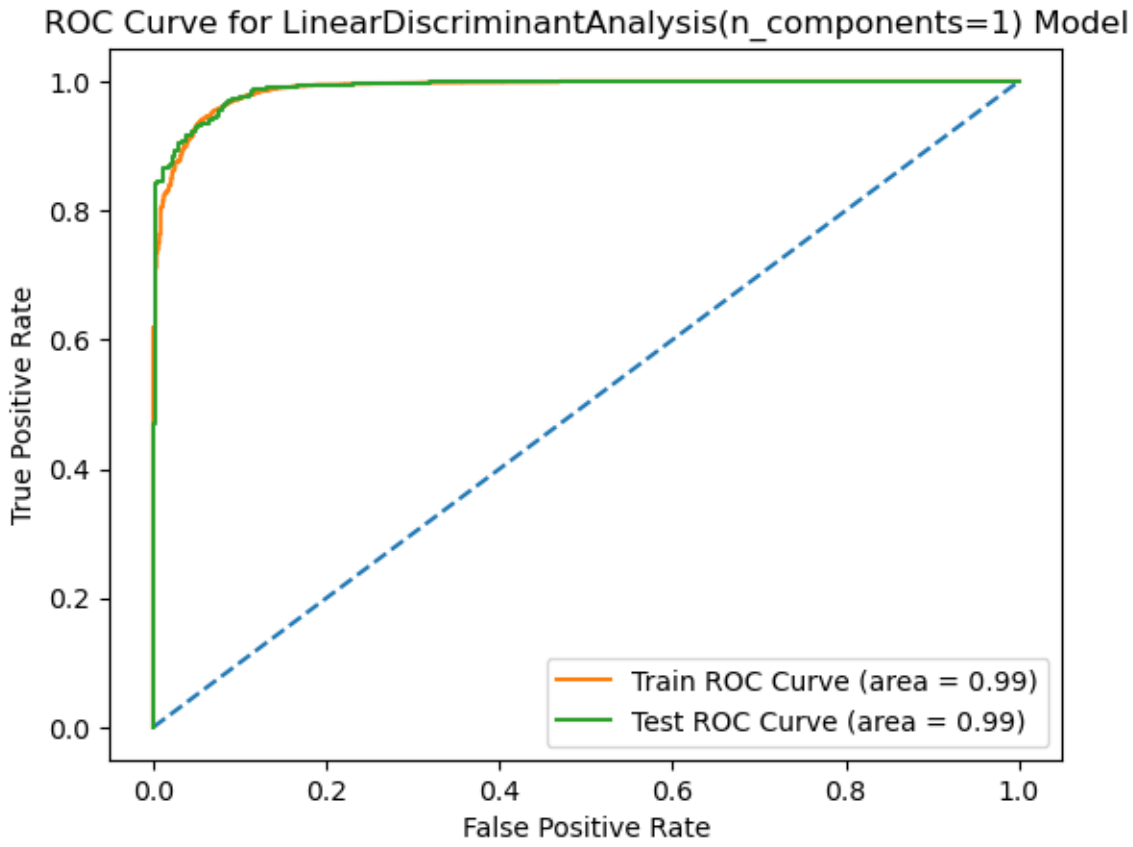
Model: Logistic Regression
Confusion Matrix for Train Data:
[[746 80]
[48 6977]]
Classification Report for Train Data:

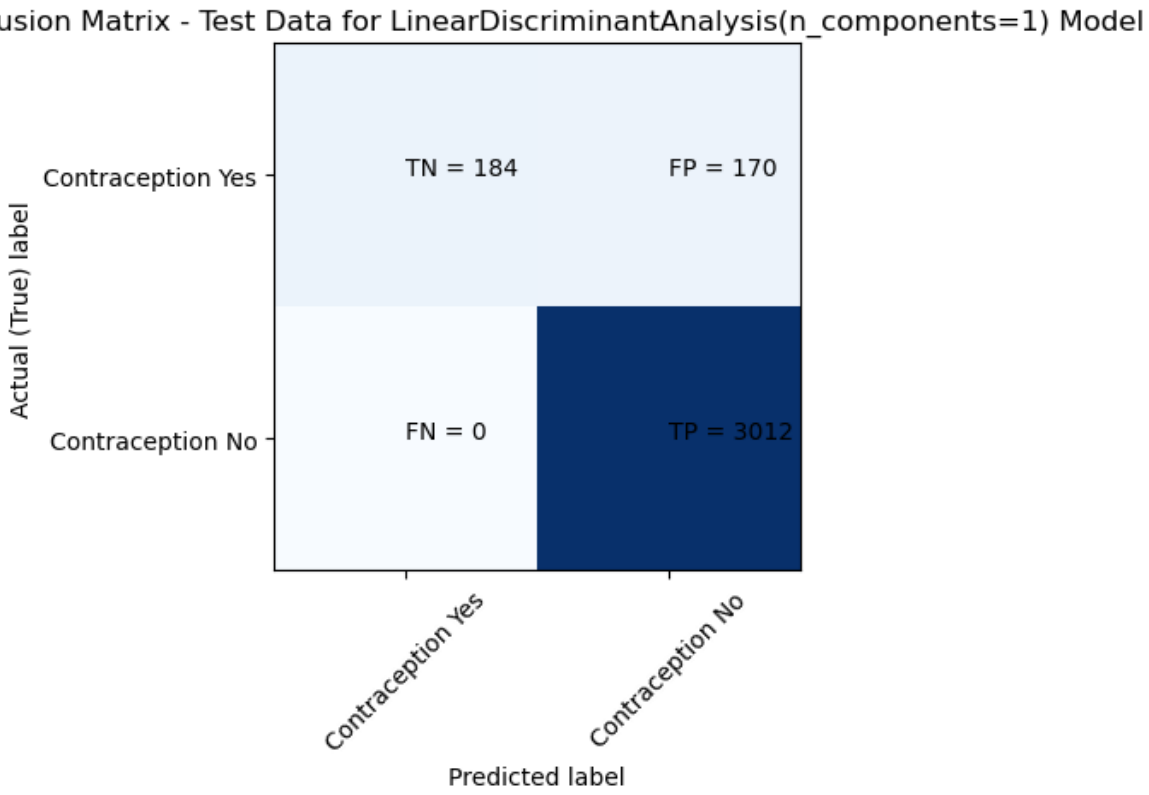
	precision	recall	f1-score	support
0	0.94	0.90	0.92	826
1	0.99	0.99	0.99	7025
accuracy			0.98	7851
macro avg	0.96	0.95	0.96	7851
weighted avg	0.98	0.98	0.98	7851

Model: Logistic Regression
Confusion Matrix for Test Data:
[[320 34]
[18 2994]]

Classification Report for Test Data:

	precision	recall	f1-score	support
0	0.95	0.90	0.92	354
1	0.99	0.99	0.99	3012
accuracy			0.98	3366
macro avg	0.97	0.95	0.96	3366
weighted avg	0.98	0.98	0.98	3366





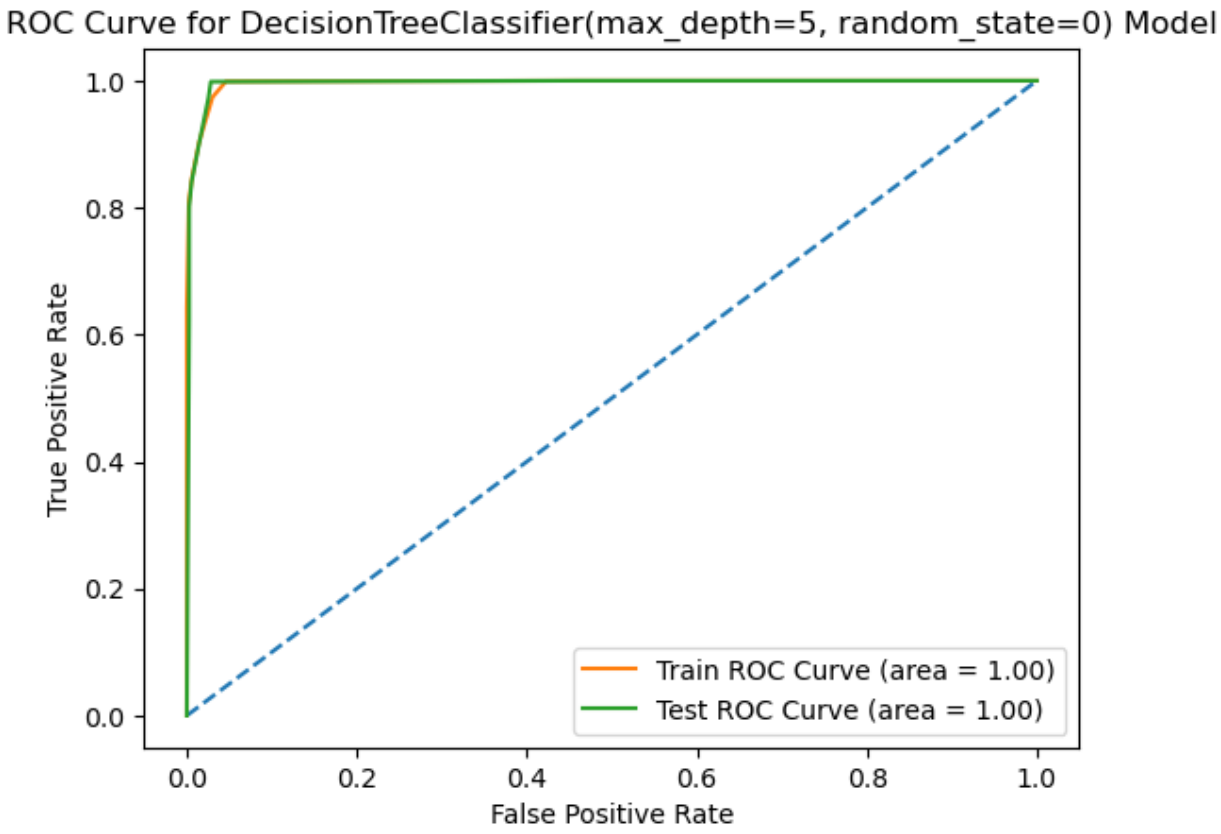
Model: LDA
Confusion Matrix for Train Data:
[[438 388]
[0 7025]]
Classification Report for Train Data:

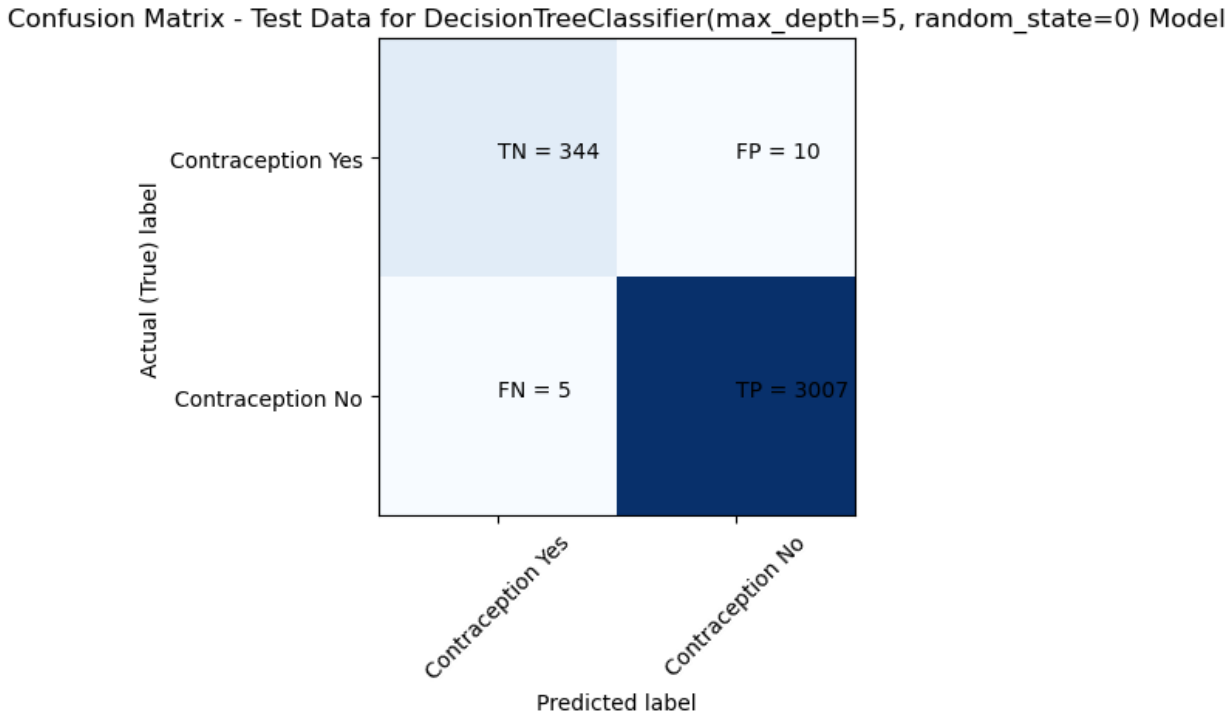
	precision	recall	f1-score	support
0	1.00	0.53	0.69	826
1	0.95	1.00	0.97	7025
accuracy			0.95	7851
macro avg	0.97	0.77	0.83	7851
weighted avg	0.95	0.95	0.94	7851

Model: LDA
Confusion Matrix for Test Data:
[[184 170]
[0 3012]]
Classification Report for Test Data:

	precision	recall	f1-score	support

0	1.00	0.52	0.68	354
1	0.95	1.00	0.97	3012
accuracy			0.95	3366
macro avg	0.97	0.76	0.83	3366
weighted avg	0.95	0.95	0.94	3366





Model: CART
Confusion Matrix for Train Data:
[[788 38]
[9 7016]]
Classification Report for Train Data:

	precision	recall	f1-score	support
0	0.99	0.95	0.97	826
1	0.99	1.00	1.00	7025
accuracy			0.99	7851
macro avg	0.99	0.98	0.98	7851
weighted avg	0.99	0.99	0.99	7851

Model: CART
Confusion Matrix for Test Data:
[[344 10]
[5 3007]]
Classification Report for Test Data:

	precision	recall	f1-score	support
0	0.99	0.97	0.98	354
1	1.00	1.00	1.00	3012

accuracy			1.00	3366
macro avg	0.99	0.99	0.99	3366
weighted avg	1.00	1.00	1.00	3366

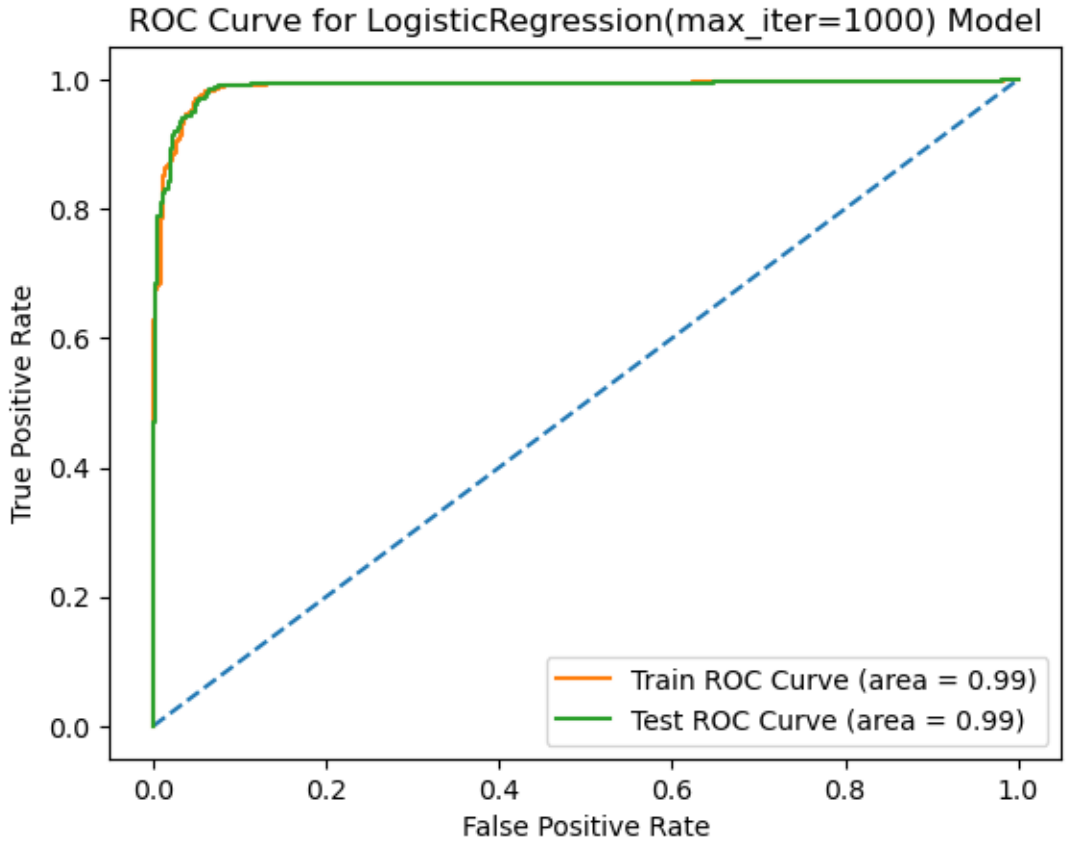
2.3 PERFORMANCE METRICS

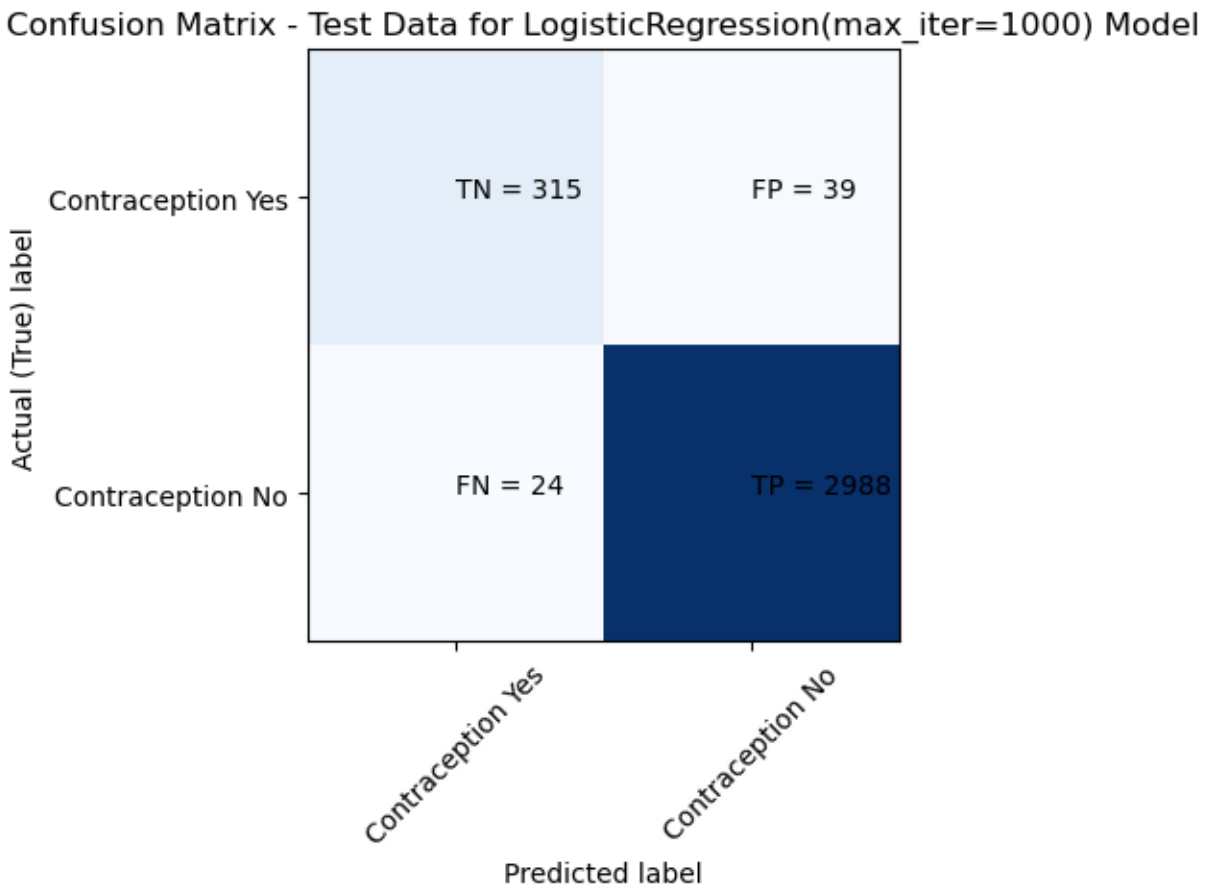
Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Compare both the models and write inferences, which model is best/optimized. (8 marks)

Answer:

2.3.1 ROC Curve, Classification Report, Confusion Matrix

When the model is run without any treatment of Feature Importances or multi -collinearity, below are the classification results.





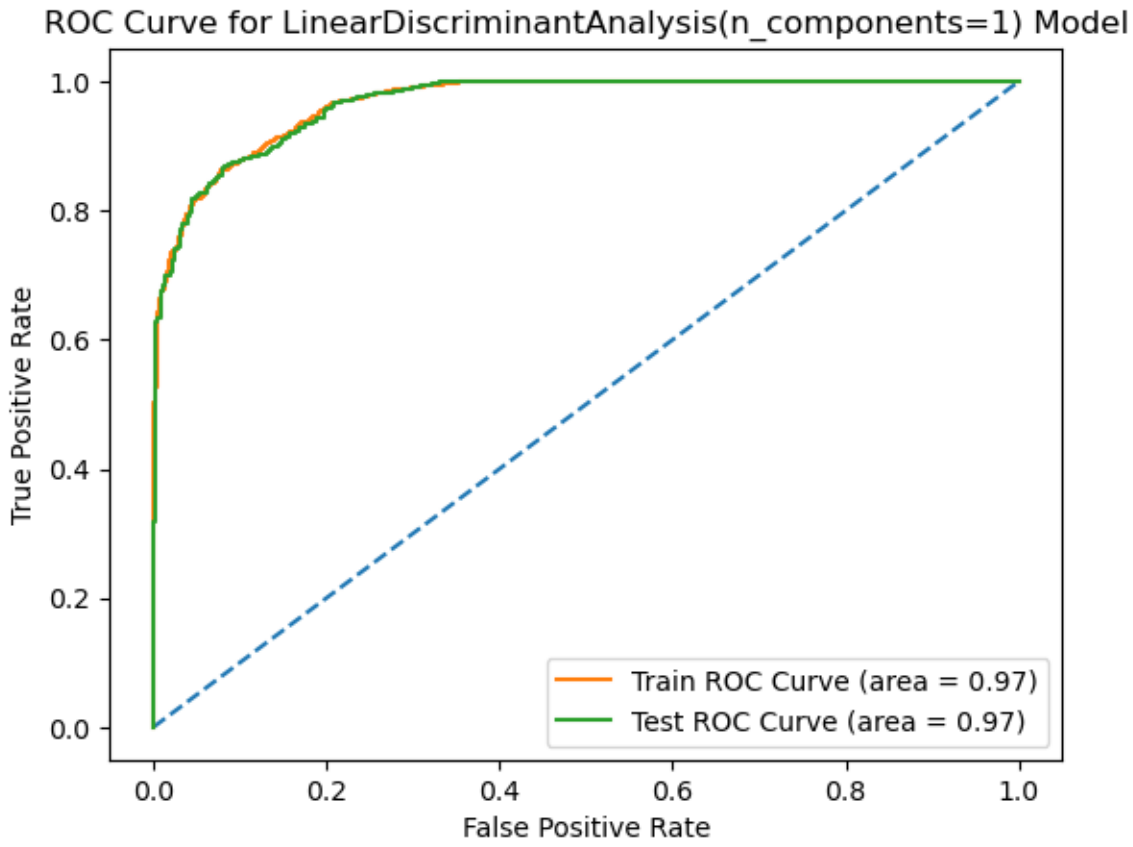
Model: Logistic Regression
Confusion Matrix for Train Data:
[[731 95]
[60 6965]]
Classification Report for Train Data:

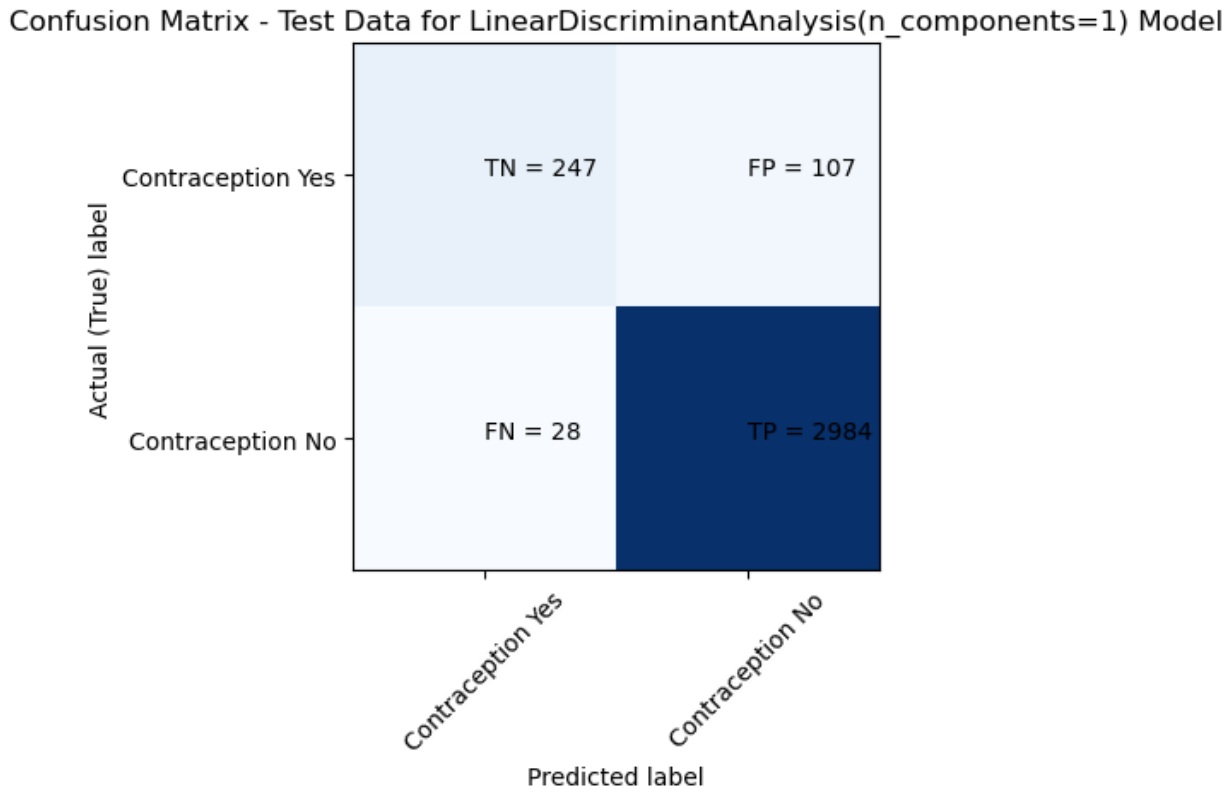
	precision	recall	f1-score	support
0	0.92	0.88	0.90	826
1	0.99	0.99	0.99	7025
accuracy			0.98	7851
macro avg	0.96	0.94	0.95	7851
weighted avg	0.98	0.98	0.98	7851

Model: Logistic Regression
Confusion Matrix for Test Data:
[[315 39]
[24 2988]]

Classification Report for Test Data:

	precision	recall	f1-score	support
0	0.93	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





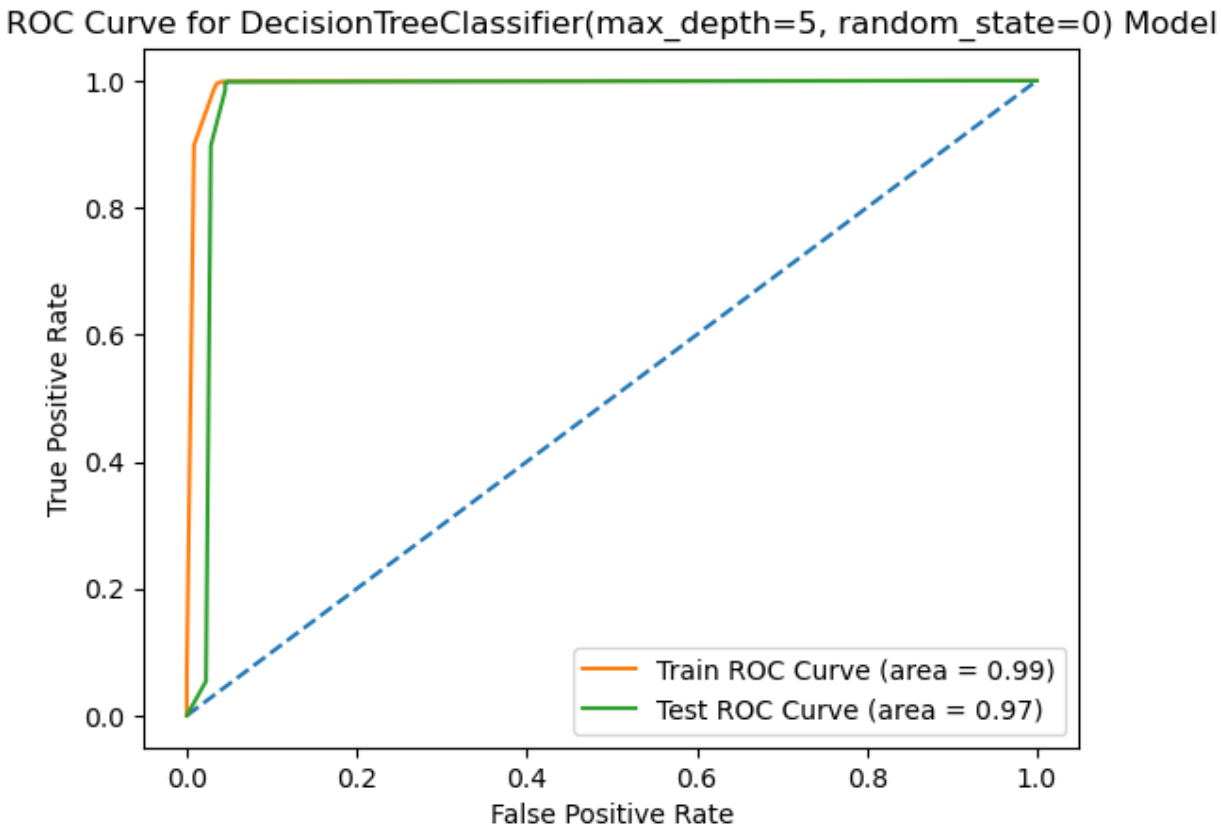
Model: LDA
Confusion Matrix for Train Data:
[[565 261]
[50 6975]]
Classification Report for Train Data:

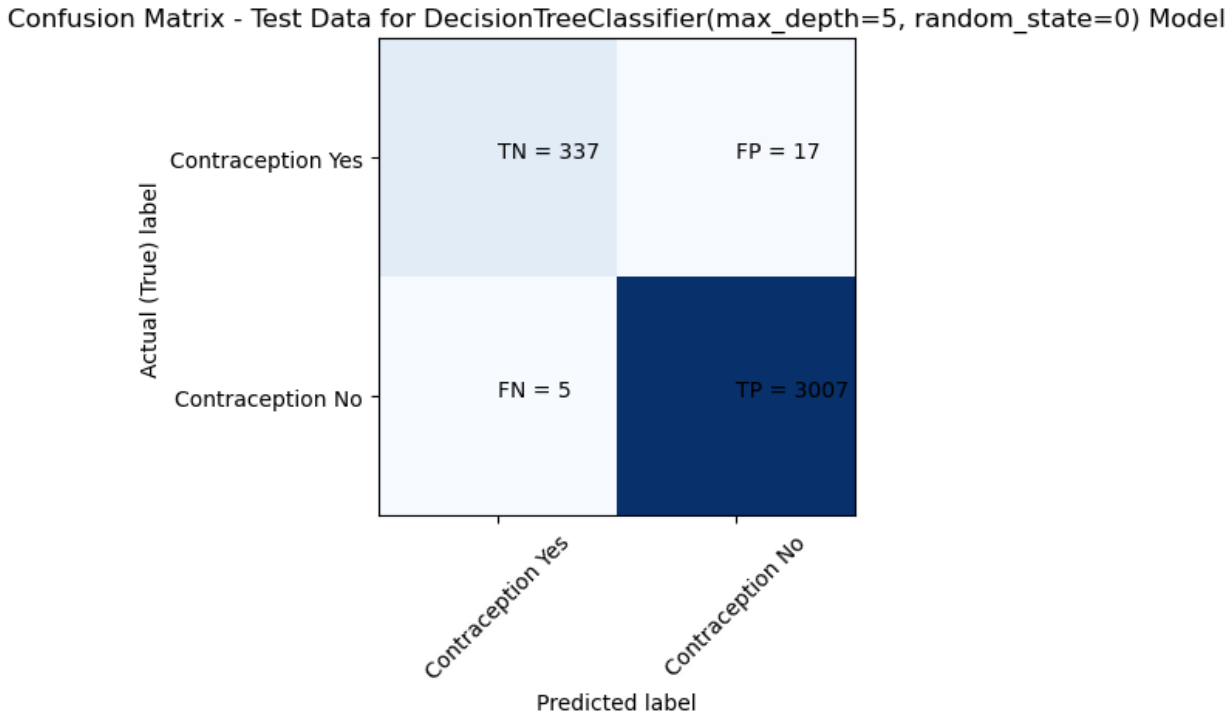
	precision	recall	f1-score	support
0	0.92	0.68	0.78	826
1	0.96	0.99	0.98	7025
accuracy			0.96	7851
macro avg	0.94	0.84	0.88	7851
weighted avg	0.96	0.96	0.96	7851

Model: LDA
Confusion Matrix for Test Data:
[[247 107]
[28 2984]]
Classification Report for Test Data:

	precision	recall	f1-score	support

0	0.90	0.70	0.79	354
1	0.97	0.99	0.98	3012
accuracy			0.96	3366
macro avg	0.93	0.84	0.88	3366
weighted avg	0.96	0.96	0.96	3366





Model: CART
Confusion Matrix for Train Data:
[[791 35]
[9 7016]]
Classification Report for Train Data:

	precision	recall	f1-score	support
0	0.99	0.96	0.97	826
1	1.00	1.00	1.00	7025
accuracy			0.99	7851
macro avg	0.99	0.98	0.98	7851
weighted avg	0.99	0.99	0.99	7851

Model: CART
Confusion Matrix for Test Data:
[[337 17]
[5 3007]]
Classification Report for Test Data:

	precision	recall	f1-score	support
0	0.99	0.95	0.97	354
1	0.99	1.00	1.00	3012

accuracy			0.99	3366
macro avg	0.99	0.98	0.98	3366
weighted avg	0.99	0.99	0.99	3366

2.3.2 Accuracy & ROC

	Model	Accuracy (train)	Accuracy (test)	ROC AUC (train)	ROC AUC (test)
0	Logistic Regression	0.980257	0.981283	0.987644	0.987604
1	LDA	0.960387	0.959893	0.967579	0.966374
2	CART	0.994396	0.993464	0.993739	0.973294

Fig2.3.2 Crash Data Regression Performance Metrics – Accuracy & ROC

CART is the best model here when VIF is not treated.

2.4 INFERENCE

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. (8 marks)

Answer:

2.4.1 Model Discussion

Considering the importance of this classification about 'survive', a model that has best recall and low FP is selected i.e. CART.

1. Both Logistic Regression and CART models perform well in terms of F1-score and accuracy across both iterations. They have F1-scores above 0.90 and high accuracies for both the train and test data.
2. LDA, on the other hand, shows lower F1-scores and accuracies compared to the other models
3. CART consistently give better results than the other models in terms of F1-score and accuracy.
4. Logistic Regression performs slightly better than LDA, achieving higher F1-scores and accuracies.
5. Considering the high performance of CART and Logistic Regression, they can be considered as strong models for predicting the classification task.
6. When multi-collinearity is treated, we shall get a even cleaner and simpler decision tree.

2.4.2 Business Recommendations:

1. The injSeverity, year of accident, weight, impact speed are the important features in determining the 'Survived' class. Conduct safety awareness programs and monitor 'airbags' or similar safety features.

2. Age and car model did not actor is as much contrary to popular belief. We can suggest hence to keep collecting data to understand hotspots, trends. And keep updating the safety policies. As, regular evaluation and updates to road safety policies based on emerging research.

3 REFLECTION REPORT:

Please reflect on all that you learnt and fill this reflection report. You have to copy the link and paste it on the URL bar of your respective browser.
https://docs.google.com/forms/d/e/1FAIpQLScKuVyrmTTM7Pboh0IB4YIBUbJp2NrDZcsY4SCRn3ZUkwmLGg/viewform

<Completed>