Predictive Modeling
Business Report

By: Ambrish Verma

Date: 30-Dec-2023

Contents

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 to help your company in investing consciously. Also, provide them with 5 attributes that are most important.

Questions for Problem 1:

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

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

1.3) Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (30:70). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE. (8 marks)

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

Dataset for Problem 1: Firm_Level_data.csv

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. (8 marks)

- The data file provided consists of 759 entries and 10 attributes.
- The attributes: Sales, capital, randd, value seem to be denoting the amount in $ terms whereas the attributes: patents, employment, tobinq and institutions do not have any unit.
- ```
  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
  ```
- The data provided was in random order, no sequence was observed in the dataset.
  ```

- Of all the attributes, all the attributes except sp500 have numeric values. Sp500 has Boolean information (True/False).
- Only attribute: 'tobinq' has null values (21 out of 759). Rest all the attributes have values in the respective rows.
- Attribute 'Unnamed:0' has sequence numbers for each row and there is no value it seems to be adding. Hence, this attribute is removed from the dataset.
- Following is the pairplot across various attributes:

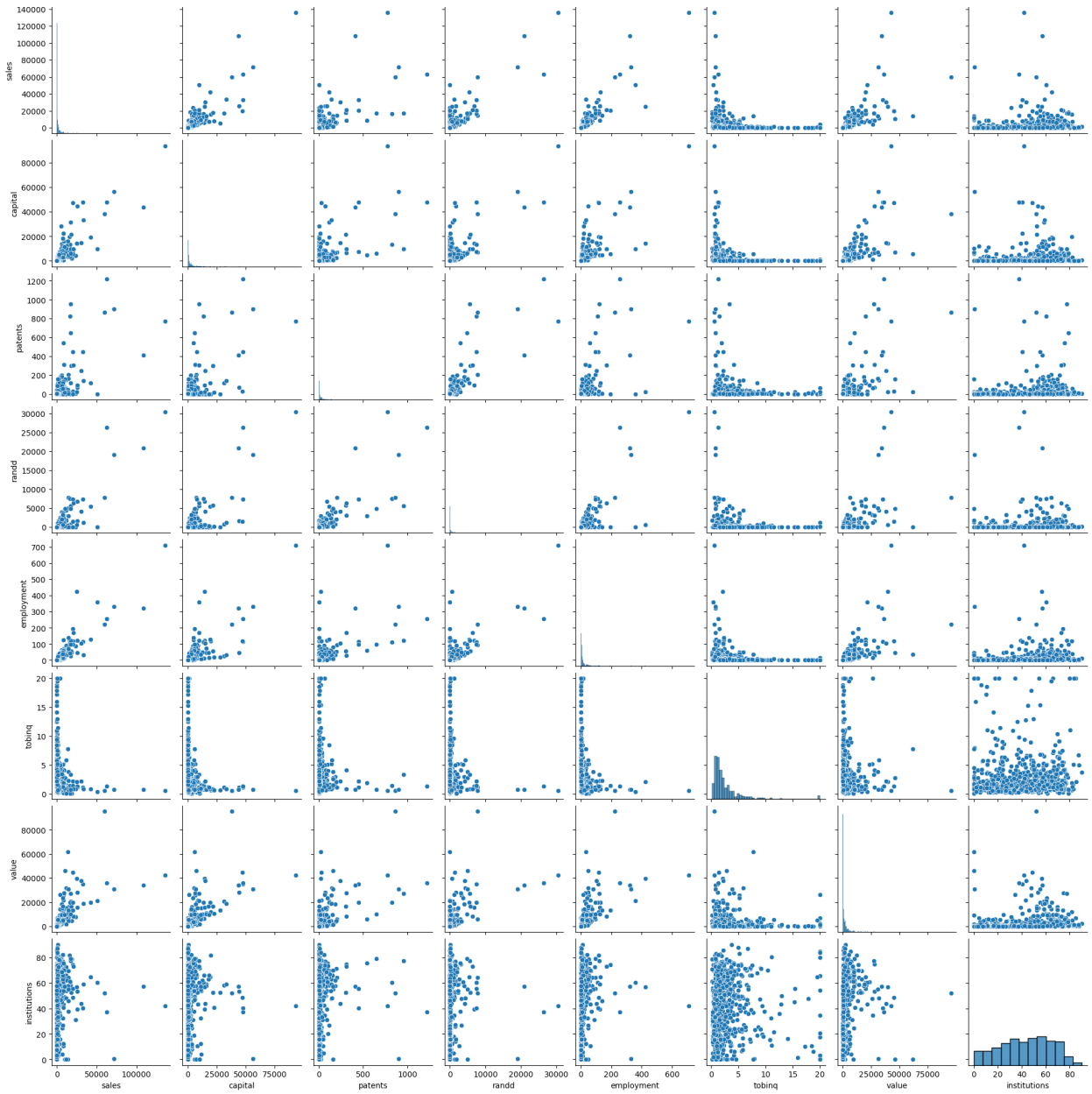

Figure 1 Bivariate analysis - pairplot before MinMax scaling

- On performing univariate analysis, it was observed that the values across attributes was uniformly distributed in the dataset.

- The attribute: ‘sp500’ was encoded with one hot encoding. This was done to perform scaling on the dataset.
- MinMax scaling was applied on the dataset. After scaling, the bivariate analysis did not show any change in the pairplot, indicating the scaling has had no effect on the correlation between the columns.
- Following is the bivariate analysis after MinMax scaling was performed:

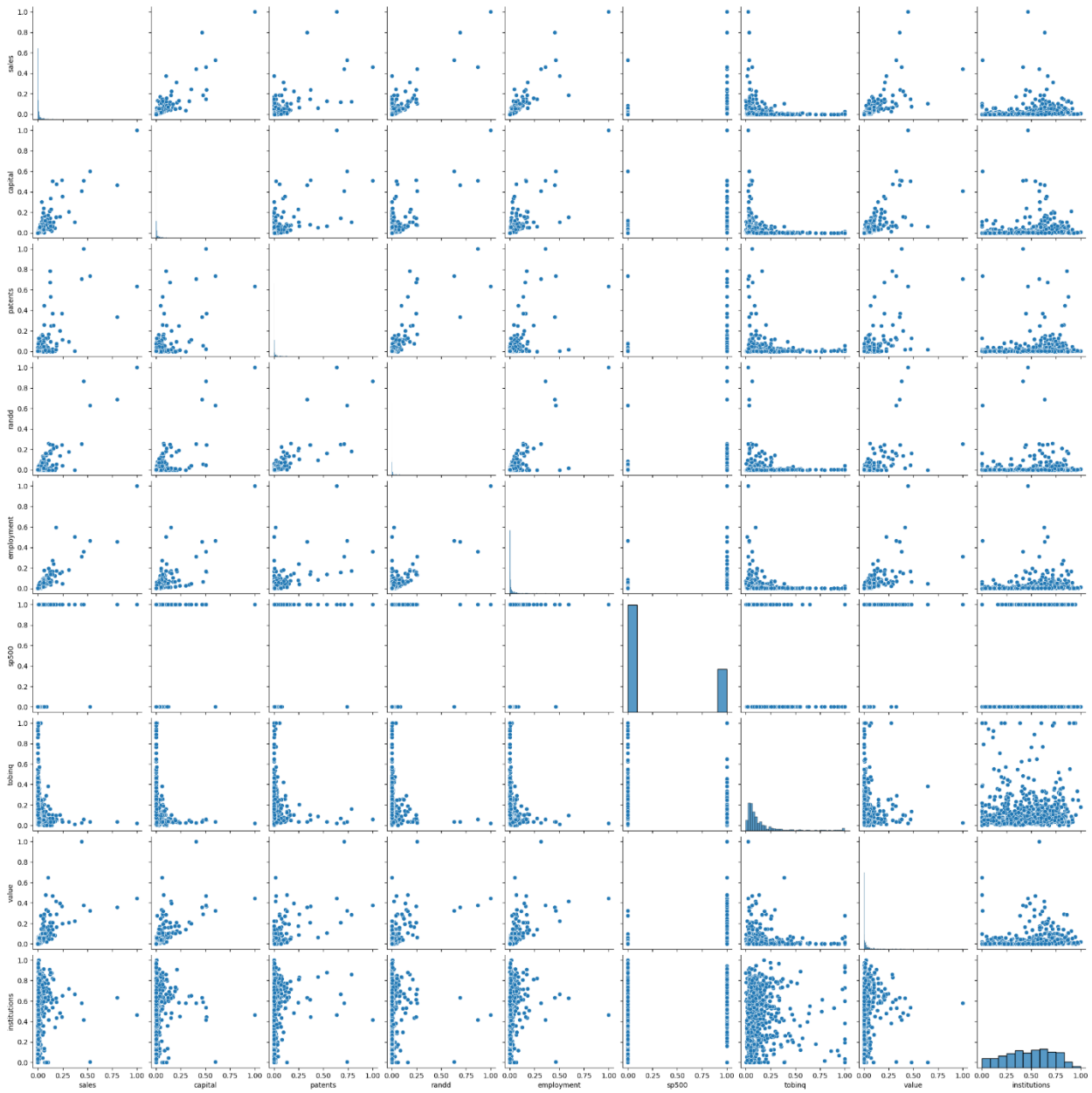

Figure 2 Bivariate analysis - pairplot after MinMax scaling

- Following is the correlation matrix to show the correlation between any 2 columns in the dataset

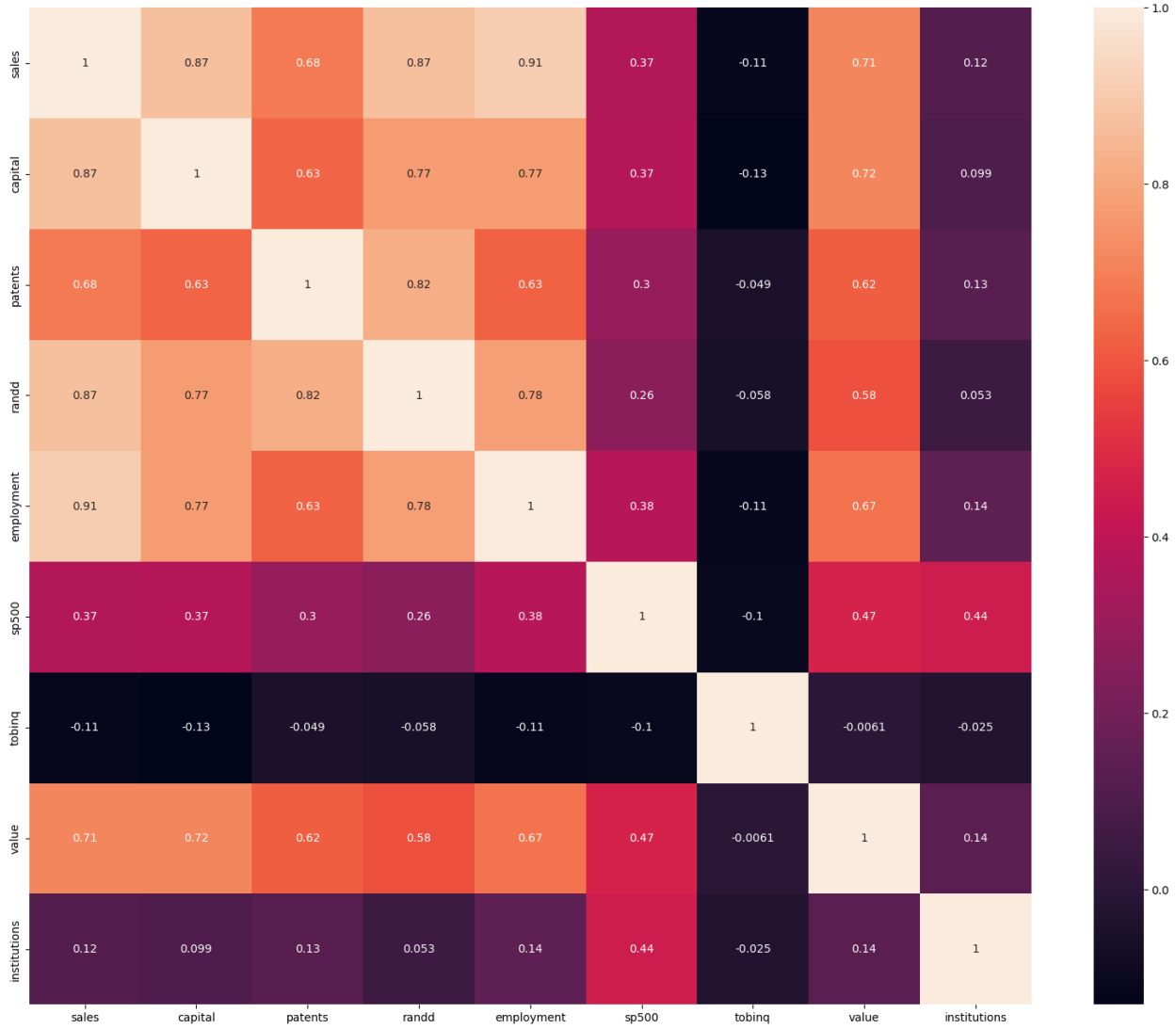

Figure 3 Correlation Matrix

- The below inferences can be drawn from the correlation matrix:
  - The attributes: sales, employment, capital, value and patents have a high degree of correlation among each other.
  - The attributes 'tobinq' and 'institutions' do not have any significant correlation with any other attribute.
- The resultant dependant attribute: sales seems to be the most dependant on the independent attributes: employment, capital, randd, value and patents. Hence, these are the 5 most important attributes for sales.

- Following is the univariate analysis of the attributes:

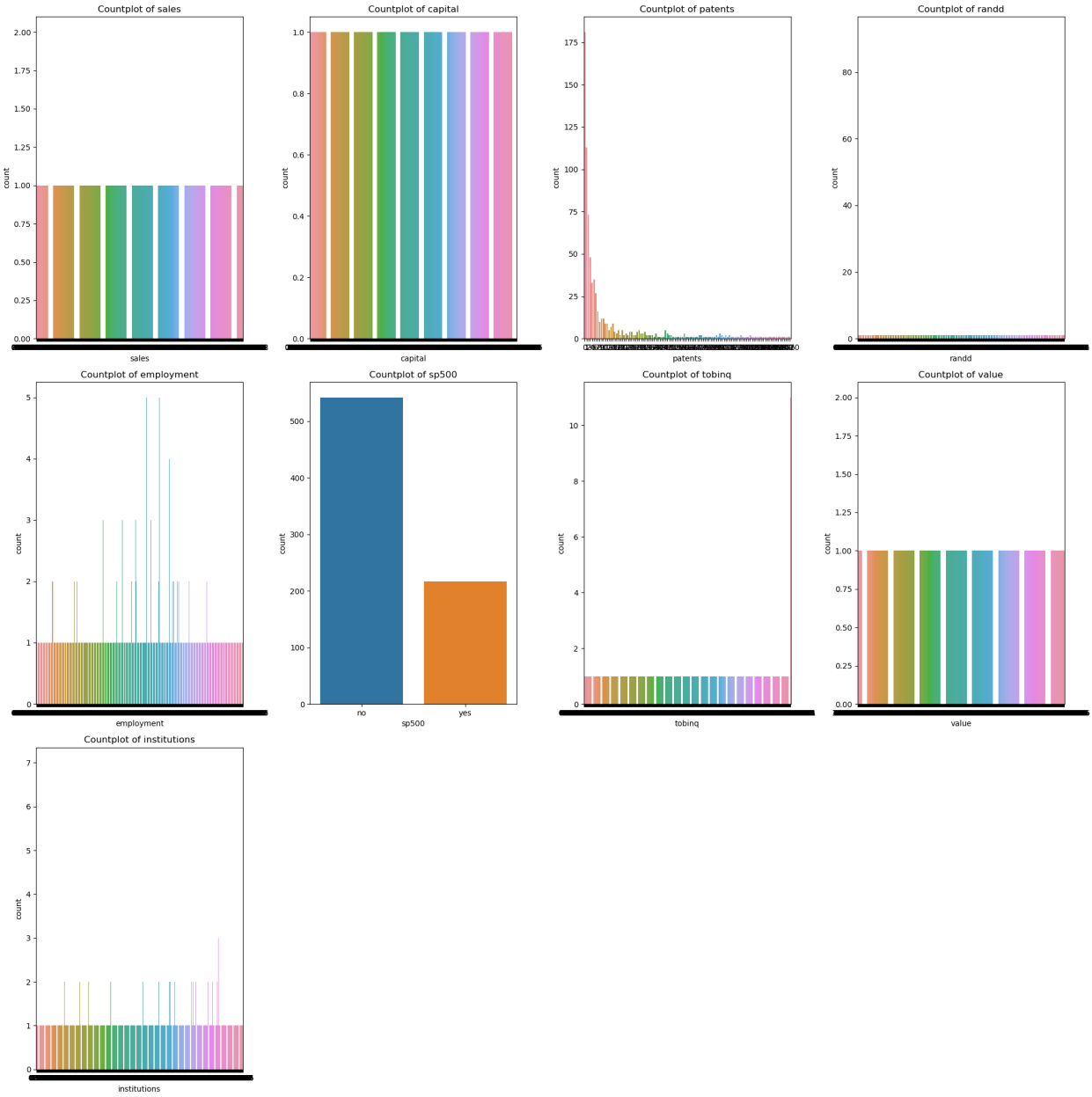

Figure 4 univariate analysis – countplot

- Box-plot of the attributes – univariate analysis:

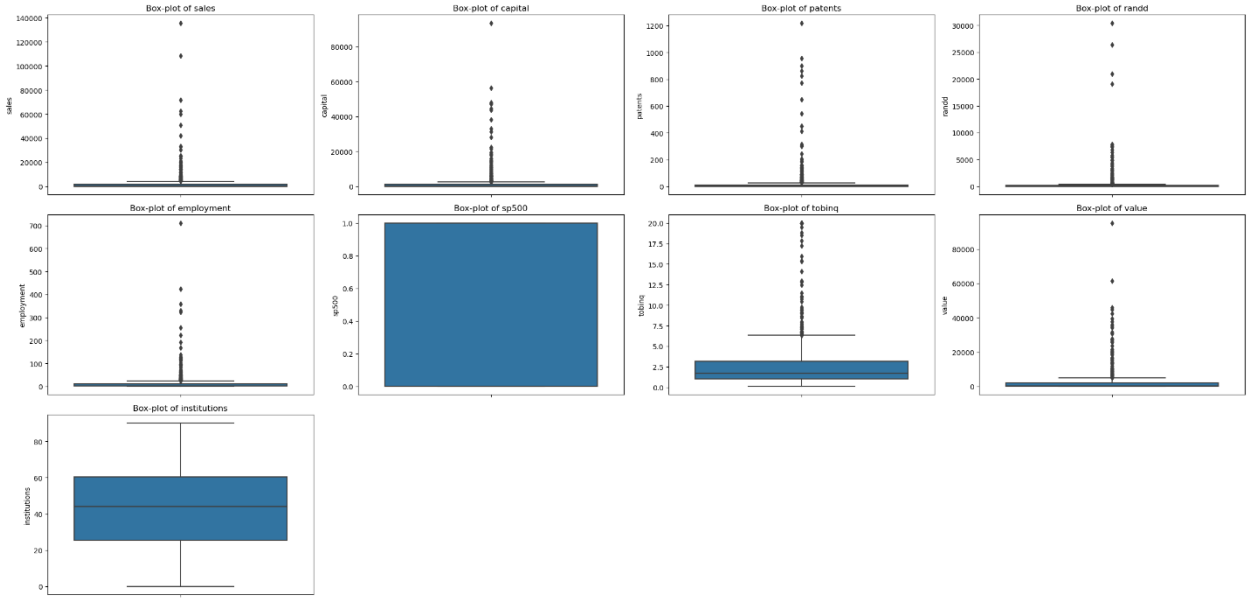

Figure 5 Box-plot - indicating presence of outliers

- Since there is significant presence of outliers and linear regression technique is sensitive to outliers, it is feasible to treat the outliers.
- This is done by moving the outliers to the upper and lower limits of IQR.
- After outlier treatment, following are the box-plots of the attributes:

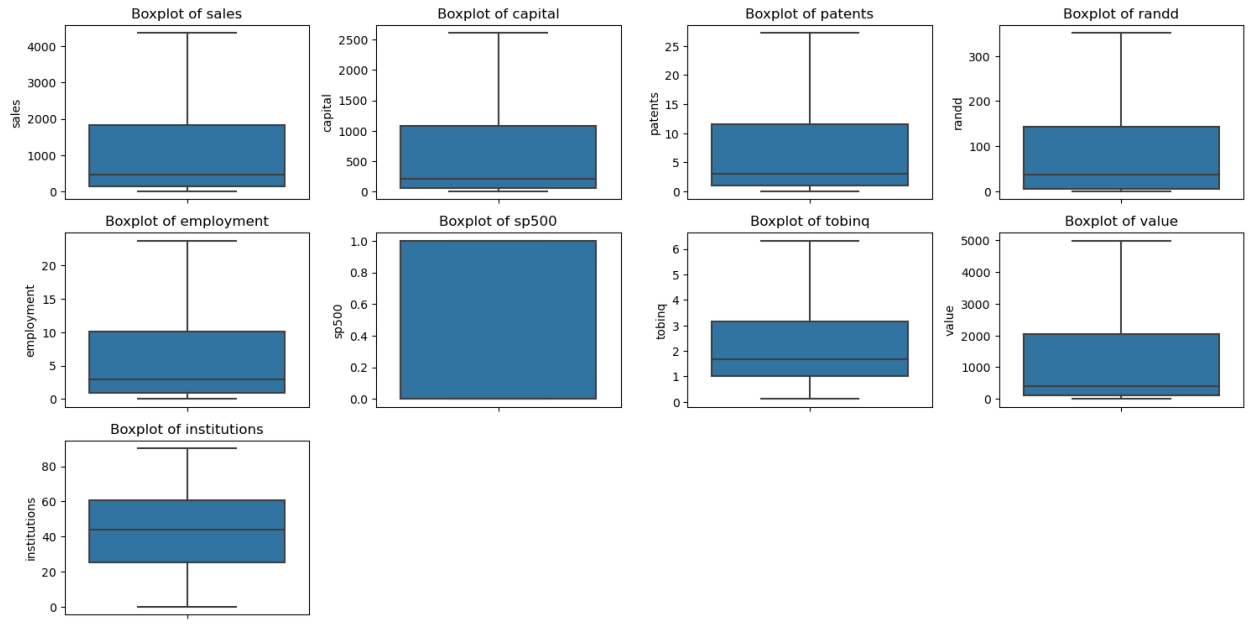

Figure 6 Box-plot after outlier treatment

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

- of all the attributes available in the dataset, 'tobinq' only has null values. As is illustrated above, this attribute does not have any significant correlation with any other.
- Imputing null vales with the median values is expected not to have any significant impact on the predictions that will be performed. Hence, the null values are replaced with the median values for the attribute: tobinq
- Scaling is beneficial in the context of the dataset that has been presented and the modelling technique they will be applied, even though it may not be necessary. Moreover, there is a risk if we do not scale the data. In the dataset that has been presented, the attributes: capital, randd, value seem to be denoting the amount in $ terms and the scale is in million $ whereas the attributes: patents, employment, tobinq and institutions do not have any unit. Furthermore, the attribute: employment has values in units of 1000s and the attribute: institutions seems to be storing data in percentages. Hence, scaling this is the logical choice.
- Scaling has been performed on the dataset provided. Box-plot, pair-plot and correlation matrices above are showing the univariate and bivariate analyses after scaling was performed.

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

- The attribute: 'sp500' was encoded with one hot encoding. This was done to perform scaling on the dataset. The univariate and bivariate analyses were performed after one-hot encoding was performed.
- First the dependent column: sales is removed from the dataframe and stored in a separate variable: y. The rest of the independent columns are kept in the dataframe and stored in a separate variable: X. This dataset is the split into test and train in the ratio of 30:70.
- Following is the summary statistic for Linear regression using statmodels package without performing any step to remove multi-collinearity:

OLS Regression Results

Dep. Variable:	sales	R-squared (uncentered):	0.961
Model:	OLS	Adj. R-squared (uncentered):	0.961
Method:	Least Squares	F-statistic:	1623.
Date:	Sat, 06 Jan 2024	Prob (F-statistic):	0.00
Time:	03:30:37	Log-Likelihood:	2347.0
No. Observations:	531	AIC:	-4678.
Df Residuals:	523	BIC:	-4644.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
capital	0.2913	0.028	10.466	0.000	0.237	0.346
patents	-0.0432	0.025	-1.726	0.085	-0.092	0.006
randd	0.1432	0.052	2.746	0.006	0.041	0.246
employment	0.4169	0.025	16.911	0.000	0.368	0.465
sp500	0.0011	0.000	2.324	0.021	0.000	0.002
tobinq	-0.0045	0.001	-3.117	0.002	-0.007	-0.002
value	0.1660	0.017	9.576	0.000	0.132	0.200
institutions	0.0007	0.000	1.555	0.121	-0.000	0.002

Omnibus:	180.219	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1255.197
Skew:	1.303	Prob(JB):	2.74e-273
Kurtosis:	10.067	Cond. No.	301.

Figure 7 Linear Regression summary stat: no column dropped.

- A constant was added to the training data.
- Variance Inflation Factor (VIF) was calculated and VIF values obtained were:

VIF values:

const	7.634492
capital	5.657160
patents	2.658440
randd	2.943306
employment	5.265370
sp500	3.051603
tobinq	1.424990
value	6.702470
institutions	1.286556

- Because of the high VIF, the column: ‘value’ was dropped from the training data.
- Following is the OLS summary stat on the linear regression with the column: ‘value’ dropped.

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.925
Model:	OLS	Adj. R-squared:	0.924
Method:	Least Squares	F-statistic:	915.5
Date:	Sat, 06 Jan 2024	Prob (F-statistic):	9.04e-289
Time:	19:55:13	Log-Likelihood:	-3970.9
No. Observations:	531	AIC:	7958.
Df Residuals:	523	BIC:	7992.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-44.0201	50.899	-0.865	0.388	-144.012	55.972
capital	0.6433	0.037	17.479	0.000	0.571	0.716
patents	-4.2275	3.021	-1.399	0.162	-10.162	1.707
randd	0.6666	0.252	2.647	0.008	0.172	1.161
employment	92.2936	4.923	18.749	0.000	82.623	101.964
sp500	344.7045	69.265	4.977	0.000	208.633	480.776
tobinq	9.4955	11.553	0.822	0.411	-13.200	32.191
institutions	0.4480	0.977	0.458	0.647	-1.472	2.368

Omnibus:	211.263	Durbin-Watson:	1.898
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2196.540
Skew:	1.439	Prob(JB):	0.00
Kurtosis:	12.539	Cond. No.	4.73e+03

Figure 8 OLS summary stats after dropping 'value' column

- After dropping 'value' column from the training data, following is the list of VIF values obtained:

VIF values:

const	7.394310
capital	3.702183
patents	2.657769
randd	2.942804
employment	4.793522
sp500	2.819500
tobinq	1.154716
institutions	1.285582

- Since all the columns are now less than a VIF value of 5, it can be safely assumed that the multicollinearity is under check.
- When calculating VIF after dropping 'value', the column 'employment' has a high degree of collinearity. But, after dropping this column, it is observed that the r-squared drops significantly.

Moreover, the F-statistic also drops significantly. Despite this, the column is dropped as the r-squared is still acceptable(0.874) as shown in the OLS summary below:

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.874
Model:	OLS	Adj. R-squared:	0.872
Method:	Least Squares	F-statistic:	604.9
Date:	Sat, 06 Jan 2024	Prob (F-statistic):	7.44e-232
Time:	19:46:30	Log-Likelihood:	-4107.4
No. Observations:	531	AIC:	8229.
Df Residuals:	524	BIC:	8259.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	40.6375	65.496	0.620	0.535	-88.030	169.305
capital	1.0434	0.039	26.937	0.000	0.967	1.119
patents	2.1205	3.878	0.547	0.585	-5.498	9.739
randd	1.4212	0.321	4.425	0.000	0.790	2.052
sp500	787.0852	84.130	9.356	0.000	621.812	952.358
tobinq	-18.2349	14.802	-1.232	0.219	-47.313	10.844
institutions	2.2952	1.256	1.827	0.068	-0.172	4.763

Omnibus:	67.182	Durbin-Watson:	2.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	572.843
Skew:	-0.045	Prob(JB):	4.06e-125
Kurtosis:	8.088	Cond. No.	4.46e+03

Figure 9 OLS summary stats after dropping 'employment' column

- The VIF data after dropping ‘employment’ column:
  VIF values:

```
const           7.336120
capital         2.457350
patents         2.624381
randd           2.867641
sp500           2.492323
tobinq          1.135790
institutions    1.272515
```

- The remaining columns have VIF in check. Proceeding to removing the columns have significant P value. The columns: institutions and tobinq have bee removed. After all the columns removal, following is the final OLS result summary:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.873
Model:                            OLS   Adj. R-squared:                  0.872
Method:                 Least Squares   F-statistic:                     907.5
Date:                Sat, 06 Jan 2024   Prob (F-statistic):          1.88e-234
Time:                        20:14:29   Log-Likelihood:                -4108.3
No. Observations:                 531   AIC:                             8227.
Df Residuals:                     526   BIC:                             8248.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -0.7825     55.759     -0.014      0.989    -110.320     108.755
capital         1.0574      0.037     28.708      0.000       0.985       1.130
randd           1.4945      0.242      6.168      0.000       1.019       1.970
sp500         783.4404     83.039      9.435      0.000     620.312     946.568
institutions    2.2910      1.249      1.835      0.067      -0.162       4.744
==============================================================================
Omnibus:                       67.610   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              582.370
Skew:                          -0.046   Prob(JB):                    3.47e-127
Kurtosis:                       8.130   Cond. No.                     4.33e+03
==============================================================================
```

Figure 10 Final OLS summary statistics on training data

- OLS summary on test data:

Root Mean Square Error: 597.0127765793177

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.873
Model:	OLS	Adj. R-squared:	0.872
Method:	Least Squares	F-statistic:	907.5
Date:	Sat, 06 Jan 2024	Prob (F-statistic):	1.88e-234
Time:	20:57:37	Log-Likelihood:	-4108.3
No. Observations:	531	AIC:	8227.
Df Residuals:	526	BIC:	8248.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.7825	55.759	-0.014	0.989	-110.320	108.755
capital	1.0574	0.037	28.708	0.000	0.985	1.130
randd	1.4945	0.242	6.168	0.000	1.019	1.970
sp500	783.4404	83.039	9.435	0.000	620.312	946.568
institutions	2.2910	1.249	1.835	0.067	-0.162	4.744

Omnibus:	67.610	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	582.370
Skew:	-0.046	Prob(JB):	3.47e-127
Kurtosis:	8.130	Cond. No.	4.33e+03

Figure 11 OLS Summary and RMSE on test data

- RMSE on test data = 597.013
- R-Squared on test data = 0.873
- F-statistic = 907.5
- Kurtosis = 8.13

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

- Based on the above Linear regression model, the linear equation that is obtained is: Firms_sales = -0.782525685847105 + 1.0573980764301987 * (capital) + 1.4944849942400398 * (randd) + 783.4404049798086 * (sp500) + 2.2909776483964825 * (institutions)
- There is almost 1:1 correlation between capital and sales.
- The randd and sales are also positively correlated. For Every 1.5 times of R&D stock, there is an impact on the sales.
- If the firm is listed in sp500, then the prospect of sales increases significantly than when it is not listed in sp500.

- Institutions and sales have a positive correlation, in that, for every proportion of stock held by institutions, the sales increase by 2.2 times of the increase in this proportion of stock of those institutions.

Recommendations:

- Investing in firms that are listed in sp500 are expected to generate more sales than the firms that are not listed. Hence, investing in sp500 firms is assured to give better returns.
- The firms that have as much stock owned by institutions have a better chance of sales numbers. So, look for the firms that have as much stock owned by institutions and invest in them.
- The firms that have a higher R&D outlay are expected to show better sales performance. Hence, invest in firms that have higher R&D budget.
- The firms that have better balance sheets (better capital management) are likely to showcase better sales figures. Hence investing in capital rich firms is a safer option.

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 based on 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 based on which you made your predictions.

Questions for Problem 2:

2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. (8 marks)

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). (8 marks)

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. (8 marks)

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

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. (8 marks)

- Observations on the dataset provided: 11217 rows and 16 columns available.
- The columns, their datatypes and the number of non-null rows per column:

```
#   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   11217 non-null  int64
 1   dvcat        11217 non-null  object
 2   weight       11217 non-null  float64
 3   Survived     11217 non-null  object
 4   airbag       11217 non-null  object
 5   seatbelt     11217 non-null  object
 6   frontal      11217 non-null  int64
 7   sex          11217 non-null  object
 8   ageOFocc     11217 non-null  int64
 9   yearacc      11217 non-null  int64
 10  yearVeh      11217 non-null  float64
 11  abcat        11217 non-null  object
 12  occRole      11217 non-null  object
 13  deploy       11217 non-null  int64
 14  injSeverity  11140 non-null  float64
 15  caseid       11217 non-null  object
```

- In all the columns, only the column: 'injSeverity' has 77 null values.
  All the null values in injSeverity was only for recorded crashes where the person had survived. Replacing with the median value is not going to impact the analysis adversely. Hence replaced the nulls with median values.
- All the columns, wherever possible were encoded to ensure proper analysis in subsequent sections. Details are mentioned in the section 2.2.

- Univariate Analysis – Countplot:

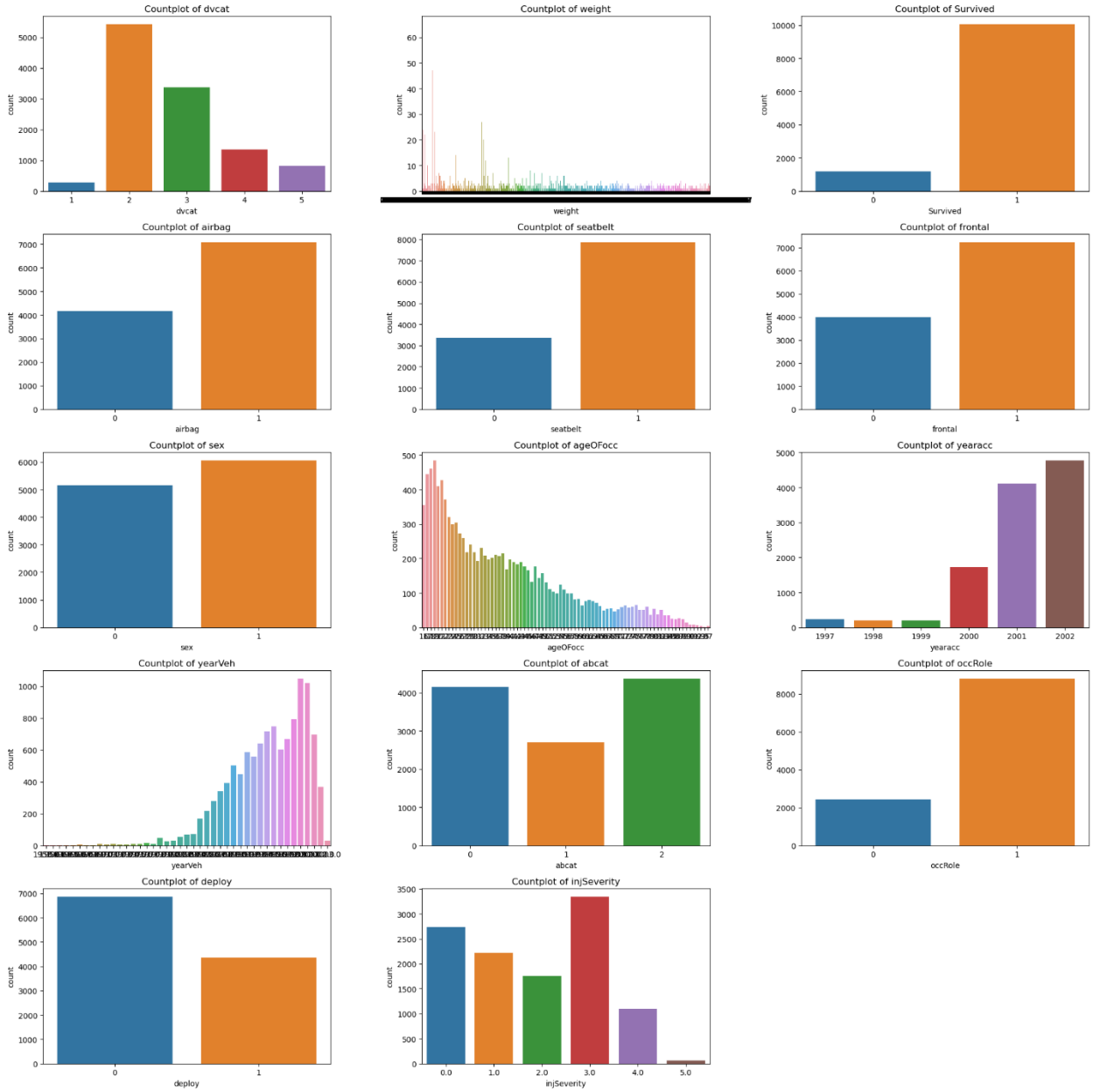

Figure 12 Car Crashes univariate analysis - Countplot

- Univariate analysis – boxplot. It is evident that outliers are present in the attributes: weight, Age of Occupant, Year of accident and Age of Vehicle. Out of these, outliers can have a significant impact on the Vehicle age, Age of occupant and weight of the vehicle.

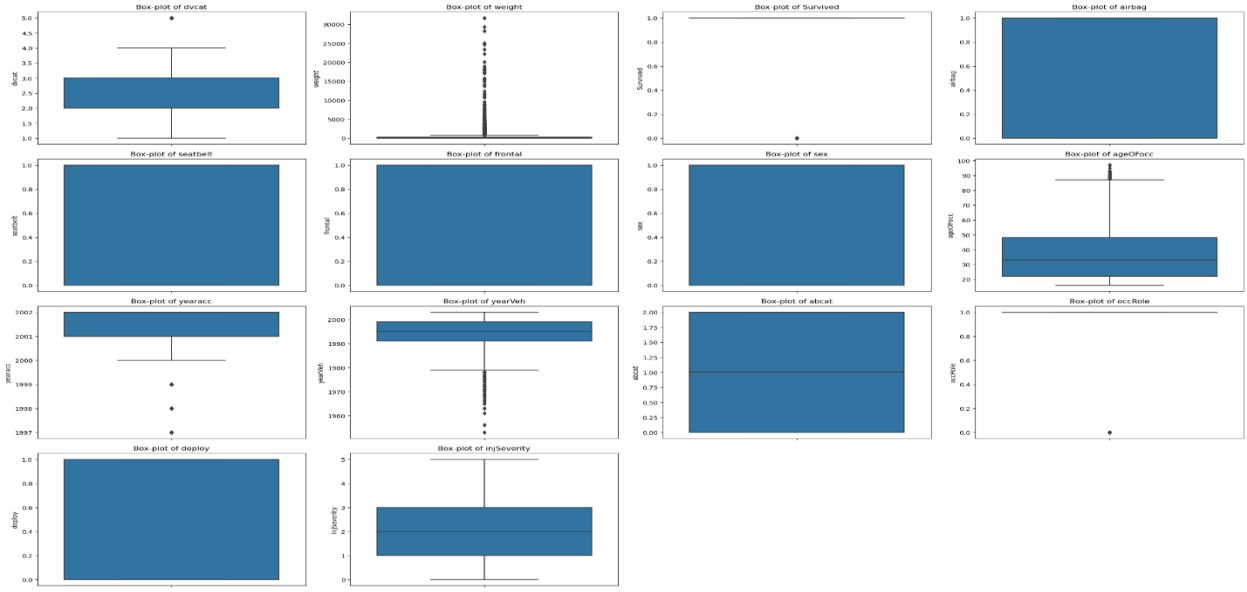

Figure 13 Univariate analysis – boxplot

- Box-plot after treatment of outliers:

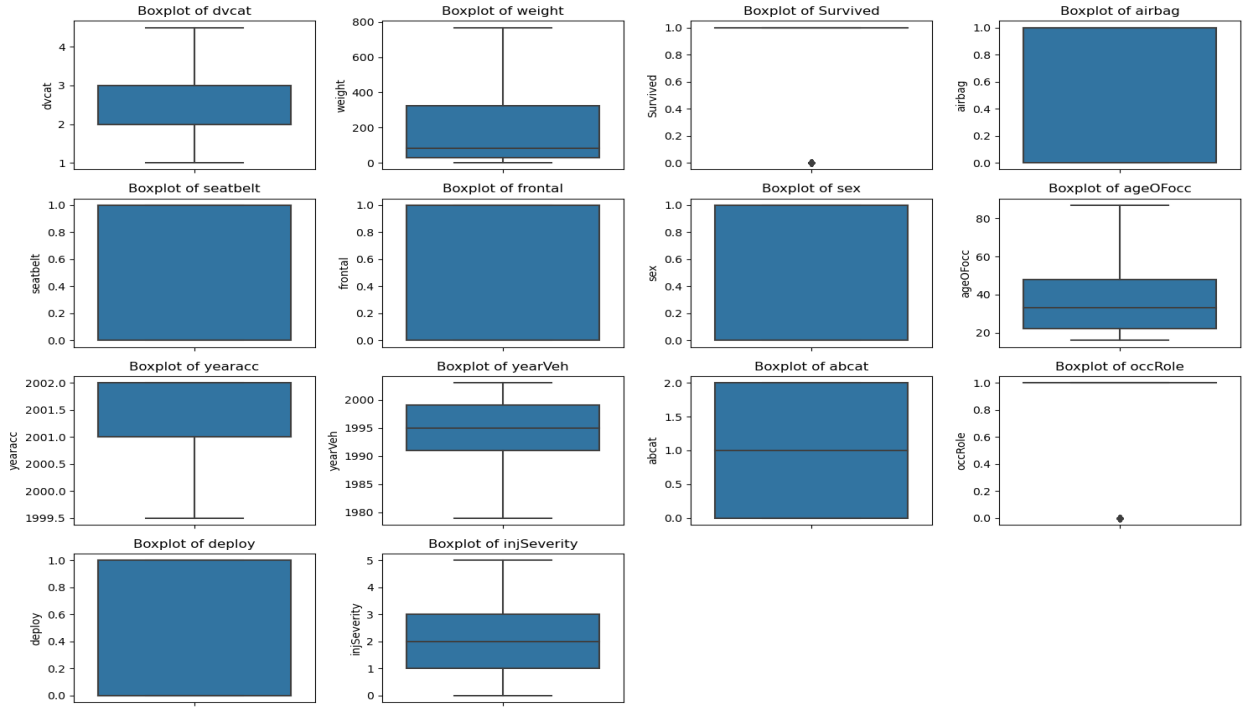

Figure 14 Box-plot of attributes after outlier treatment

- Bi-variate Analysis:

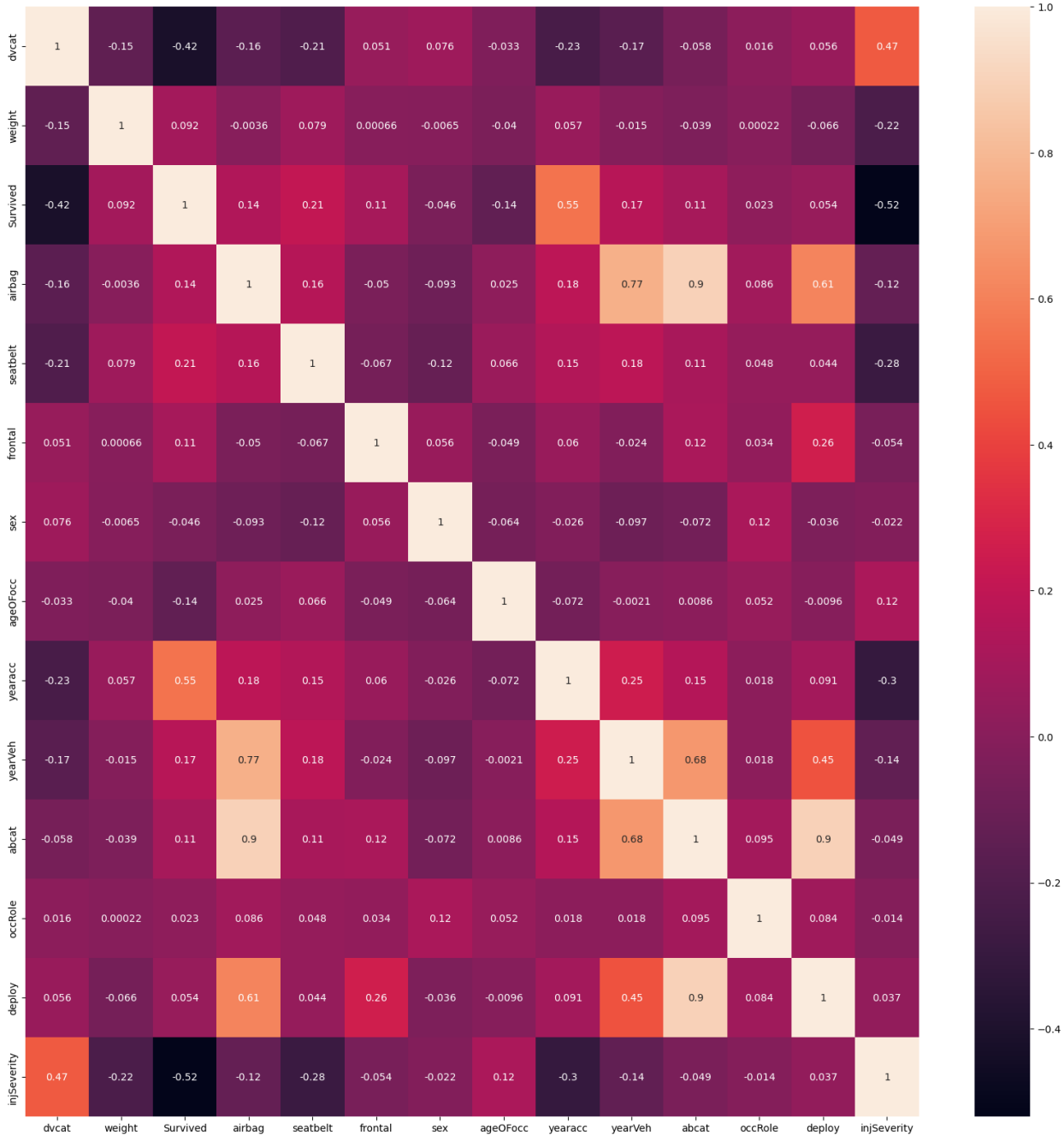

Figure 15 Bivariate analysis - Correlation heatmap

- Bivariate analysis – pairplot:

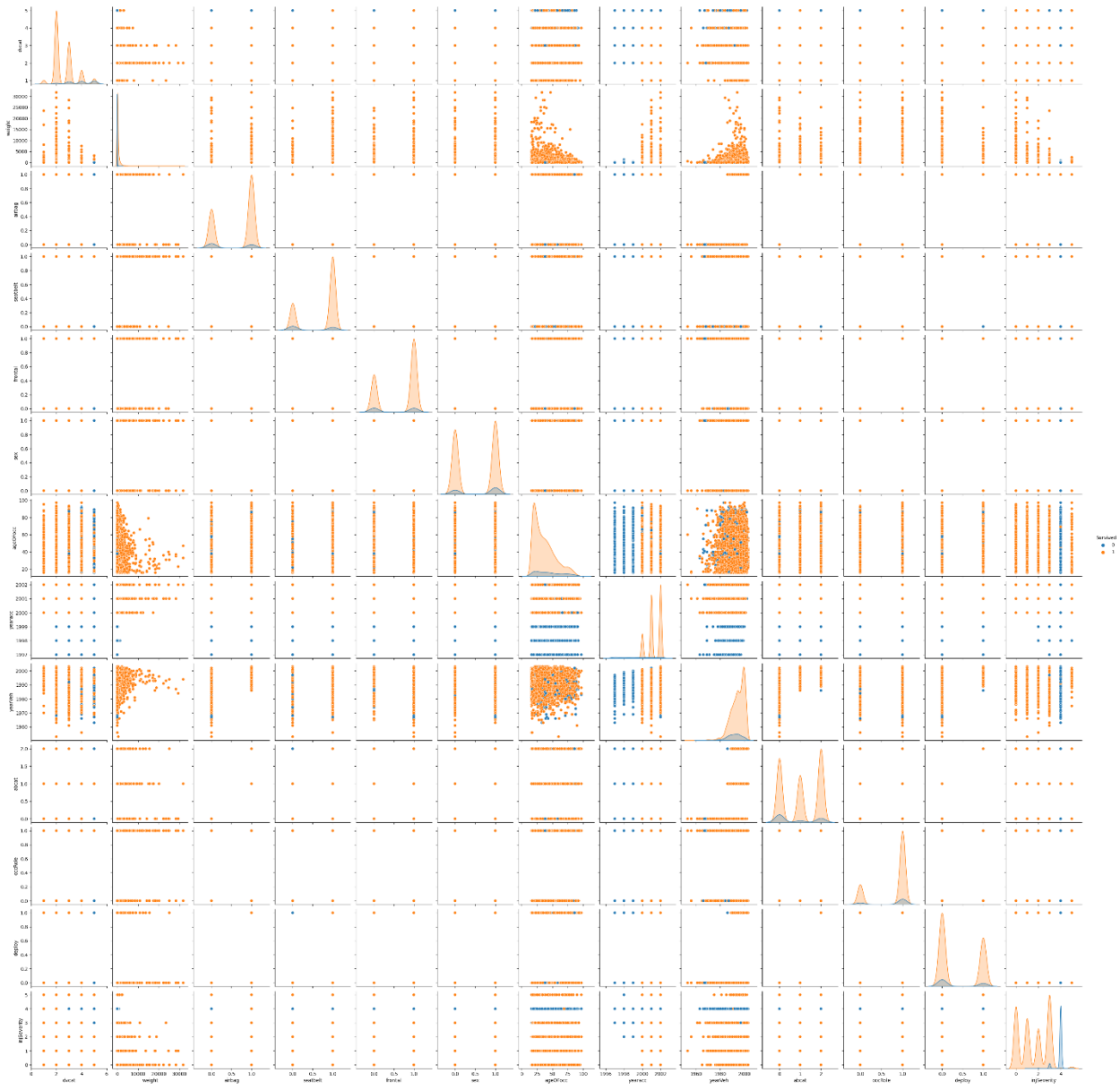

Figure 16 Bivariate analysis - Pairplot

Interpretations:

- In the dataset, very few numbers of non-survivors are present, but this is the class of interest. So, there is no outlier treatment done on Survived column. The column: occRole has the same anomaly in terms of outlier treatment. So, no treatment is done on the column: occRole as well.
- The column: Survived seems to be positively correlated with column: yearacc and significantly negatively correlated with columns: dvcat and injseverity.
- Non-survivors seem to only fall in the category of 55+ km/h(as displayed in the pair-plot for dvcat-Survived column)

- The columns: deploy, occRole, abcat, ageOFocc, airbag, seatbelt, frontal, sex, weight seem to be poor predictors to predict the Survivor in the event of Crash.

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). (8 marks)

- Based on the data analysis, multiple columns were found to be categorical in nature. They were encoded in the following way:
    - the column dvcat can be converted into 5 categories in the following way: 1-9:1, 10-24:2, 25-39:3, 40-54:4, 55+: 5.
    - The column 'Survived' can be encoded into 0 and 1 values, with 0 being not survived and 1 being survived.
    - The column 'airbag' can be converted into binary value, 1 being for airbag and 0 for none.
    - The column 'seatbelt' can be converted into binary value, 1 being for belted and 0 for none.
    - The column 'sex' can be converted into binary value, 1 being for male and 0 for female
    - The column 'yearacc' can be converted into following categories: 1997:0, 1998:1, 1999:2, 2000:3, 2001:4, 2002:5.
    - The column 'abcat' can be converted into following caterogies: unavail=0, nodeploy=1, deploy=1
    - The column 'occRole' can be converted into binary value, 1 being for driver and 0 for passenger.
- The encoding of the above columns were performed as part of the problem 2.1 above.
- The dataset was split into a ratio of 70:30 between training and test data.
- Logistic Regression model and LDA were respectively built on the training data. The models were then deployed onto the test data to assess the accuracy of the models developed.

- Logistic regression model results:

```
Confusion metrics:
 [[ 342   41]
 [  23 2960]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.89      0.91       383
           1       0.99      0.99      0.99      2983

    accuracy                           0.98      3366
   macro avg       0.96      0.94      0.95      3366
weighted avg       0.98      0.98      0.98      3366
```

Figure 17 Logistic Regression - Confusion matrix and Classification report

- LDA model results:

```
Confusion Matrix:
 [[ 275  108]
 [  25 2958]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.72      0.81       383
           1       0.96      0.99      0.98      2983

    accuracy                           0.96      3366
   macro avg       0.94      0.85      0.89      3366
weighted avg       0.96      0.96      0.96      3366
```

Figure 18 LDA - Confusion Matrix and Classification report

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. (8 marks)

- The accuracy scores of logistic regression and LDA on the test set are:
    - Logistic regression accuracy score= 0.9809863339275104
    - LDA accuracy score= 0.9604872251931076

- Logistic Regression - Confusion matrix plot:

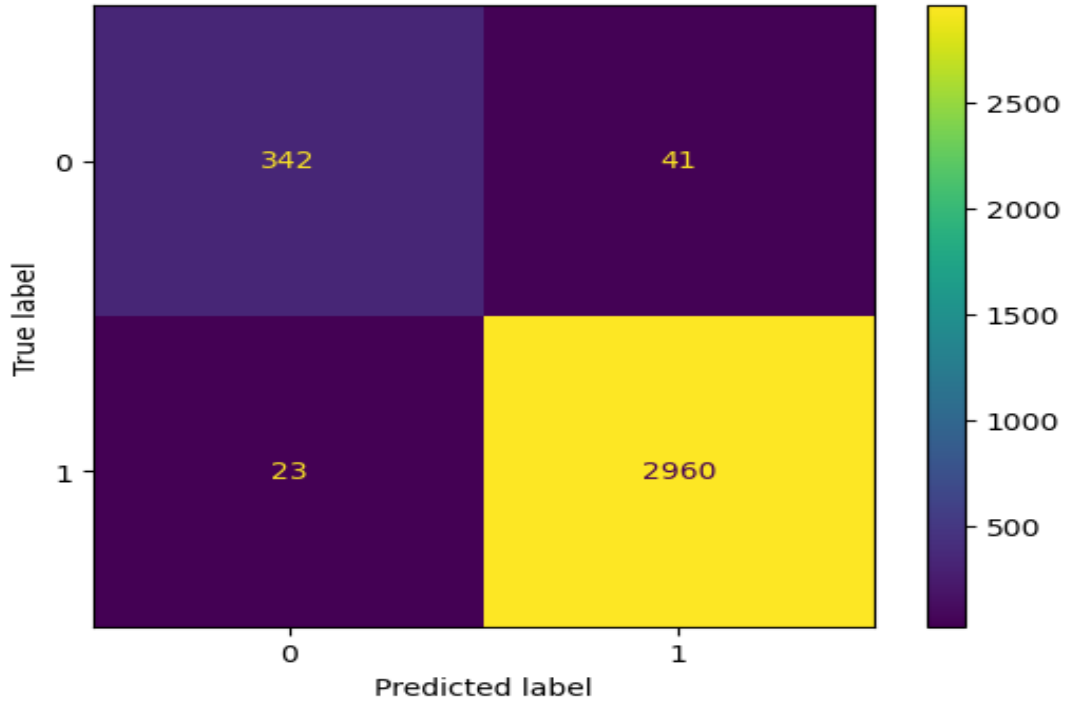

Figure 19 Logistic Regression - Confusion matrix

- LDA – Confusion matrix plot:

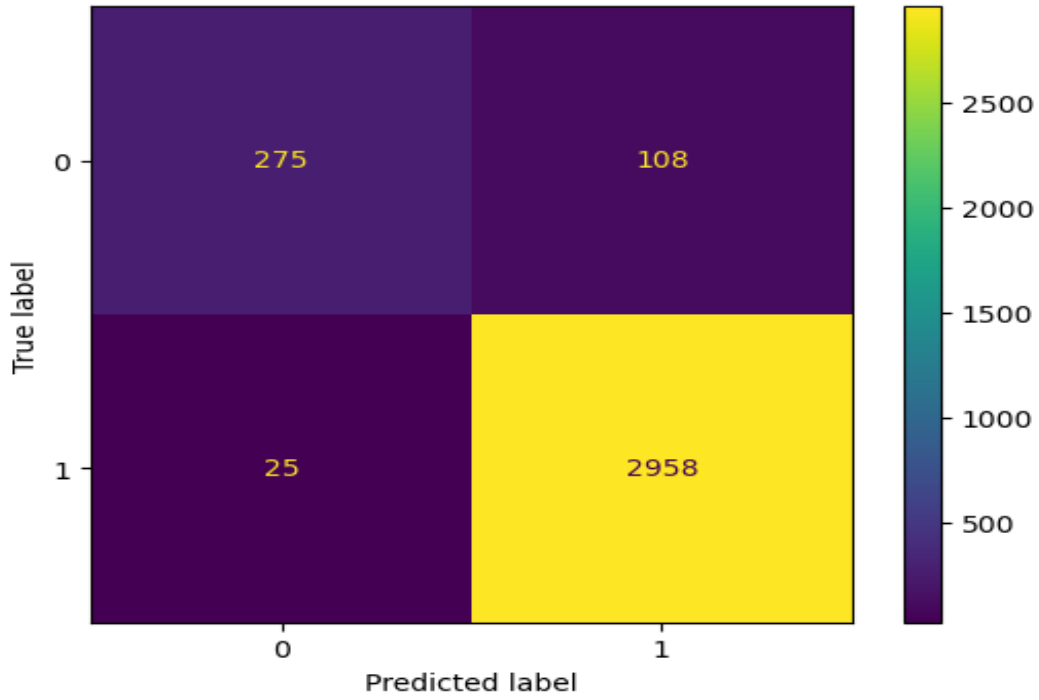

Figure 20 LDA - Confusion matrix plot

- Logistic regression ROC-AUC curve:

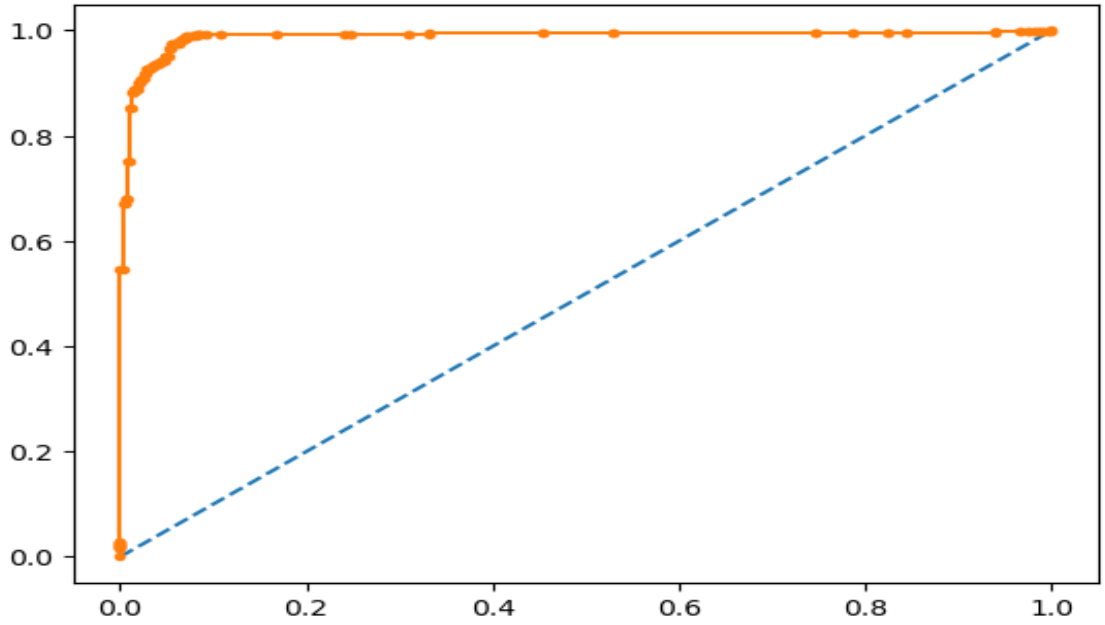

Figure 21 Logistic Regression ROC-AUC curve for Survived = 1

- LDA ROC-AUC curve:

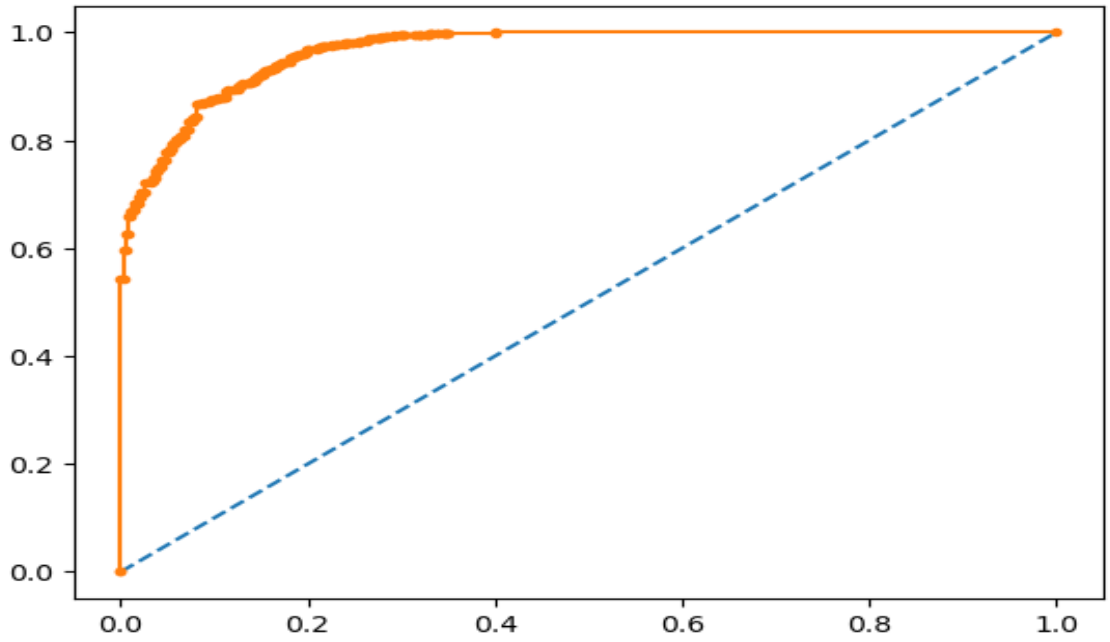

Figure 22 LDA ROC-AUC curve for Survived = 1

- ROC-AUC score:
    - Logistic regression roc-auc score= 0.9426200164728064
    - LDA roc-auc score= 0.8548174205615984

- Between the 2 models, Logistic regression is better optimized. This inference is based on the following data points:
    - The ROC-AUC score for Logistic regression is higher than that of LDA
    - The area under the ROC-AUC curve is higher under Logistic regression than the LDA, indicating the better predictions for survived=1, and by extension, for survived=0
    - F1 score of Logistic regression for Survived=0 is 0.91 and that of LDA is 0.81.
    - Since the class of interest is that of Survived=0(non survivors), the lesser number of non-survivors predicted as survivor is more detrimental to overall utility of the models. In the logistic regression, the non-survivors identified as survivors is 23 whereas, in the LDA, this number is 25. The same is illustrated in the higher recall for Logistic regression than LDA (0.89 for Survived =0 in Logistic regression versus 0.72 for survived=0 in LDA)

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

- To make the recommendations, variables and their coefficients need to be obtained and applied on the logistic regression equation.
- The coefficients for the data set are for probability of Survived=1:

[[-0.81124507 0.00524362 0.2459758 0.68470834 0.78524832 -0.20228853

-0.03103302 0.05702408 -0.04798209 0.34827368 0.10753951 0.10229788

-4.36318589]]

- The constant as part of logistic regression equation: [-0.00336495]
- The equation obtained is:
  P(Y=1) = 1/1+e^-(-0.0034 -
  0.81*(dvcat)+0.005*(weight)+0.24*(airbag)+0.68*(seatbelt)+0.78*(frontal)-0.2*(sex)-
  0.03*(ageOFocc)+0.057*(yearacc)-
  0.048*(yearVeh)+0.348*(abcat)+0.11*(occRole)+0.102*(deploy)-4.36*(Severity))

Recommendations:

- The feature: Severity has the most important predictor of Survivorship. If the Severity of injury is high, the probability of the accident victim's survival is low. If the medical treatment is provided immediately, the chances of survival of the crash victim can be improved. So, action on improving the timely access to medical facility in the event of car crash must be taken. Perhaps a distress call to the nearest medical facility might help.

- The feature: dvcat also negatively impacts the survivorship. So, it can be safely inferred that the more the speed of the car at the time of the crash, the lesser are the chances of survival. So, ensure the speed of the cars is kept under check by any means.
- Out of the frontal and side impact on the car in the event of a crash, chances of survival out of frontal crash are higher than the impact on the sides of the car.
    - Action must be taken to ensure the lane discipline is maintained.
    - Cars crossing to the other side of the road must follow the rules to ensure right-of-way.
- Survivability increases when seatbelt is worn. Hence, ensure the seatbelts are worn whenever the car is moving.
- When the airbags are deployed, survival of the person in the event of crash increases. Hence, rules need to be made stricter to ensure airbags are present and deployed in the event of crash.