2022

PREDICTIVE MODELLING GRADED PROJECT REPORT

DSBA

Girish Chadha
29/12/2022

Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

Problem 1

Problem 1: Linear Regression

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

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. (8 marks)

Solution: Reading the data

	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

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

Table 1

Checking data types

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 759 entries, 0 to 758
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sales         759 non-null    float64
 1   capital       759 non-null    float64
 2   patents       759 non-null    int64  
 3   randd         759 non-null    float64
 4   employment    759 non-null    float64
 5   sp500         759 non-null    object 
 6   tobinq        738 non-null    float64
 7   value         759 non-null    float64
 8   institutions  759 non-null    float64
dtypes: float64(7), int64(1), object(1)
memory usage: 53.5+ KB
```

Shape - 759 Rows and 9 columns

Describing dataset

	sales	capital	patents	randd	employment	tobinq	value	institutions
count	759.000000	759.000000	759.000000	759.000000	759.000000	738.000000	759.000000	759.000000
mean	2689.705158	1977.747498	25.831357	439.938074	14.164519	2.794910	2732.734750	43.020540
std	8722.060124	6466.704896	97.259577	2007.397588	43.321443	3.366591	7071.072362	21.685586
min	0.138000	0.057000	0.000000	0.000000	0.006000	0.119001	1.971053	0.000000
25%	122.920000	52.650501	1.000000	4.628262	0.927500	1.018783	103.593946	25.395000
50%	448.577082	202.179023	3.000000	36.864136	2.924000	1.680303	410.793529	44.110000
75%	1822.547366	1075.790020	11.500000	143.253403	10.050001	3.139309	2054.160386	60.510000
max	135696.788200	93625.200560	1220.000000	30425.255860	710.799925	20.000000	95191.591160	90.150000

Table 2

Checking Null Values

```
sales            0
capital          0
patents          0
randd            0
employment       0
sp500            0
tobinq          21
value            0
institutions     0
dtype: int64
```

There are 21 entries missing in tobinq coloumn.

UNIVARIATE ANALYSIS

```
BoxPlot of sales
----------------------------------------------
```

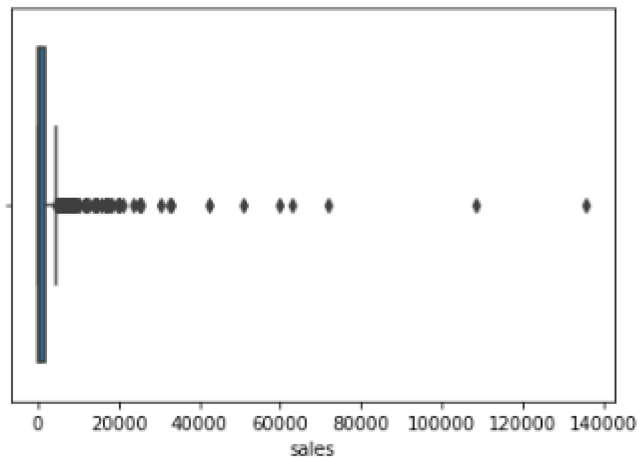


Figure 1

There seem to be many outliers in the data. Maximum number of sales are 135696 million dollars but average sales are 2689 million dollars only.

```
BoxPlot of employment
----------------------------------------------
```

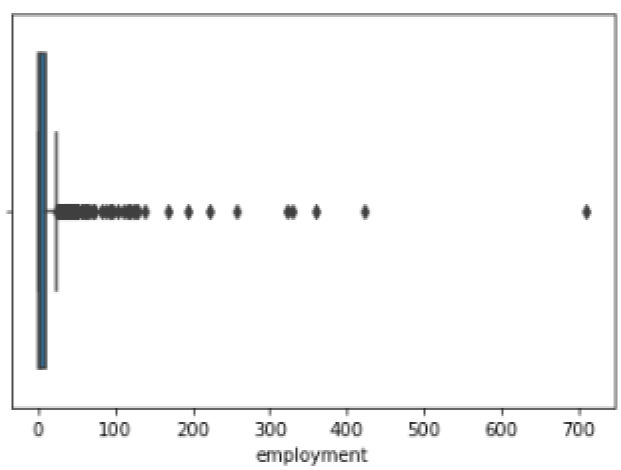
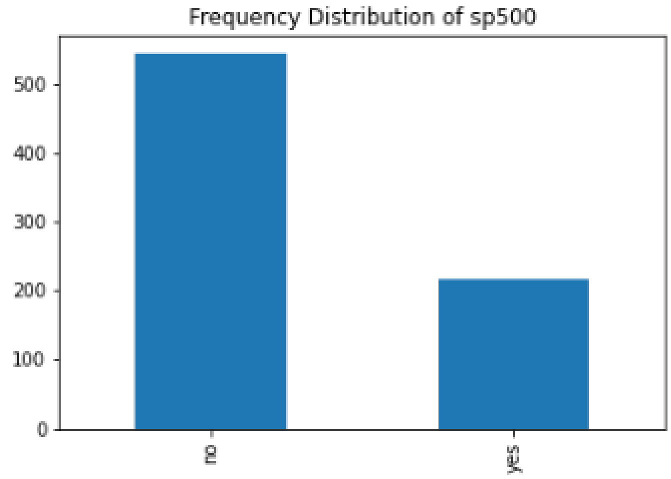


Figure 2

There seem to be many outliers in the data. Maximum number of Employment is 710 but average employment is 14 only.

```
Details of sp500
----------------------------------------------------------------
no      542
yes     217
Name: sp500, dtype: int64
```



217 firms have membership in the S&P 500 index while 542 firms don't have membership in sp500.

Figure 3

BIVARIATE ANALYSIS

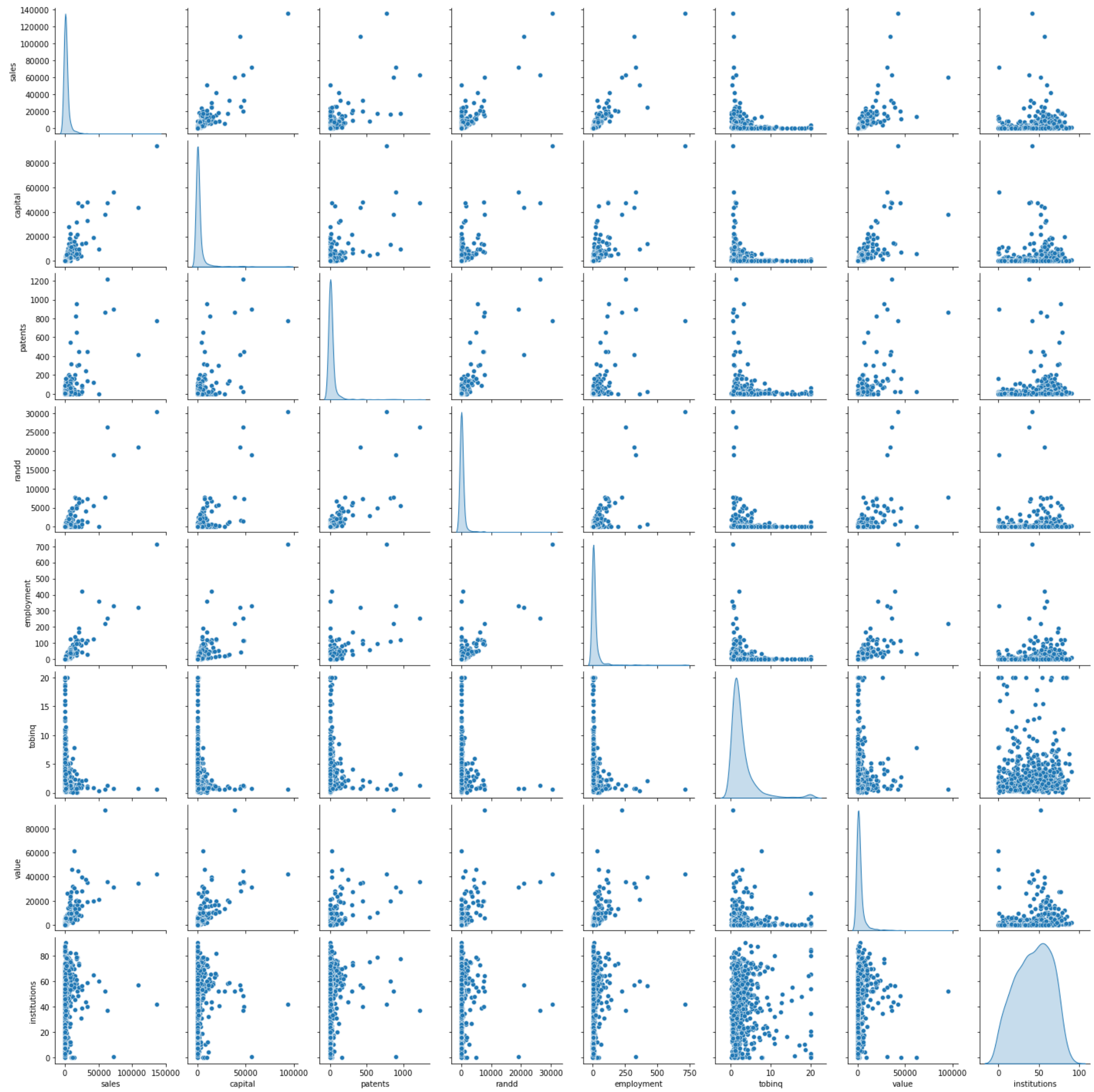


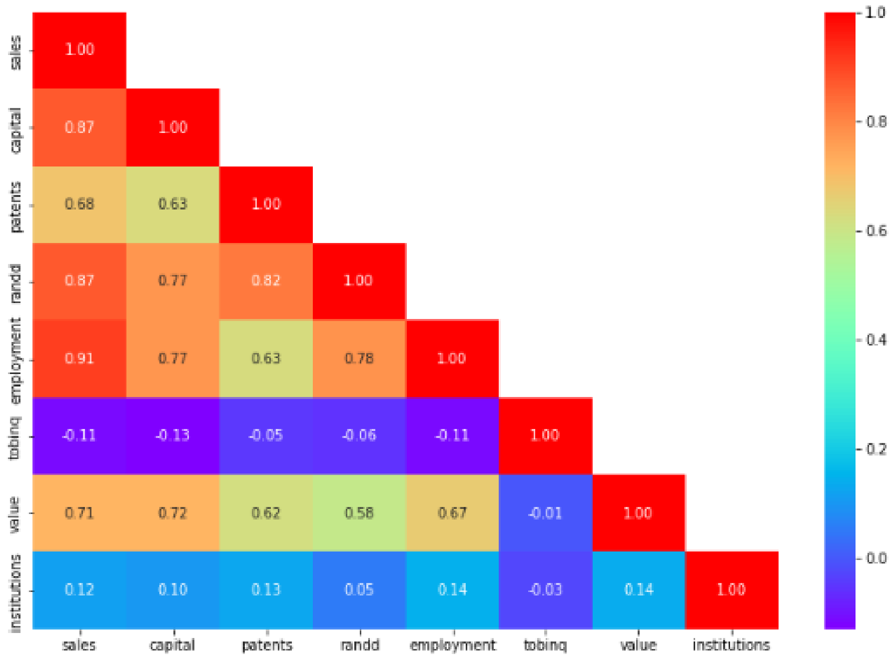
Figure 4

Observe that the relationship between 'Sales' and other attributes is not really linear.

However, the plots also indicate that linearity would still capture quite a bit of useful information/pattern.

Several assumptions of classical linear regression seem to be violated

MULTIVARIATE ANALYSIS



There seems to be a very high correlation between sales & capital coloumn, randd & patents coloumn, employment & sales coloumn, tobinq seems to have a negative correlation with other attributes.

Figure 5

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

```
sales            0
capital          0
patents          0
randd            0
employment       0
sp500            0
tobinq          21
value            0
institutions     0
dtype: int64
```

There are 21 entries missing in tobinq coloumn. These need to be treated. We won't drop the null values instead we will replace them with the median value of the coloumn itself.

Median values of all variables

```
sales           448.577082
capital         202.179023
patents           3.000000
randd            36.864136
employment        2.924000
tobinq            1.680303
value           410.793529
institutions     44.110000
dtype: float64
```

So we create 2 simple true or false columns with titles equivalent to sp500_yes & sp500_No.

We will also be dropping one of those two columns to ensure there is no linear dependency between the two columns.

Creating dummy variables

	sales	capital	patents	randd	employment	tobinq	value	institutions	sp500_yes
0	826.995050	161.603986	10	382.078247	2.306000	11.049511	1625.453755	80.27	0
1	407.753973	122.101012	2	0.000000	1.860000	0.844187	243.117082	59.02	0
2	8407.845588	6221.144614	138	3296.700439	49.659005	5.205257	25865.233800	47.70	1
3	451.000010	266.899987	1	83.540161	3.071000	0.305221	63.024630	26.88	0
4	174.927981	140.124004	2	14.233637	1.947000	1.063300	67.406408	49.46	0

Table 3

Using median filler null value are imputed

```
sales           0
capital         0
patents         0
randd           0
employment      0
tobinq          0
value           0
institutions    0
sp500_yes       0
dtype: int64
```

Checking boxplot of all numeric variables

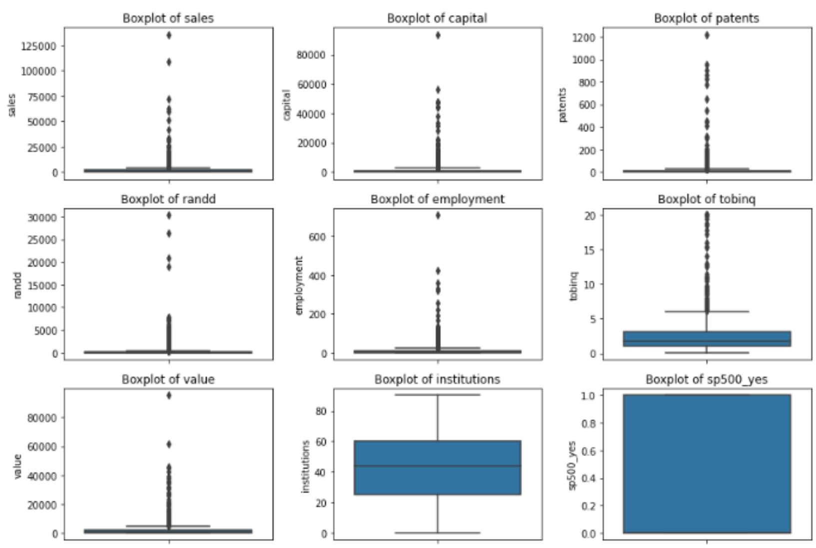


Figure 6

Scaling the data

	sales	capital	patents	randd	employment	tobinq	value	institutions	sp500_yes
0	-0.213704	-0.281030	-0.162882	-0.028842	-0.273914	2.493756	-0.156696	1.718839	-0.632747
1	-0.261802	-0.287143	-0.245190	-0.219303	-0.284216	-0.577847	-0.352317	0.738279	-0.632747
2	0.656027	0.656624	1.154052	1.424056	0.819869	0.734749	3.273585	0.215929	1.580410
3	-0.256841	-0.264737	-0.255479	-0.177659	-0.256243	-0.740066	-0.377803	-0.744789	-0.632747
4	-0.288514	-0.284354	-0.245190	-0.212208	-0.282206	-0.511899	-0.377183	0.297142	-0.632747

There are many outliers in the data. Scaling is necessary in this case as sales and randd are in million of dollars and other variables are having smaller units. To ensure that the model is predicting well we need to scale it before proceeding.

Table 4

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.

Taking X as all independent variable and Y as dependent variable(Sales).

Splitting the data into test and train sets in 30:70 ratio.

Test set -

```
     const   capital   patents     randd  employment    tobinq     value  \
626    1.0 -0.102446 -0.111439 -0.182776    0.053299 -0.333052 -0.070880
333    1.0 -0.303676 -0.245190 -0.214691   -0.314105 -0.717031 -0.385324
257    1.0 -0.222758 -0.059996 -0.175741   -0.174105 -0.191825 -0.242456
173    1.0 -0.181128 -0.265767 -0.184957   -0.151469  0.116599 -0.072309
242    1.0 -0.243753 -0.245190 -0.219303   -0.291376 -0.183503 -0.336054

     institutions  sp500_yes
626     -0.532988  -0.632747
333     -0.984276  -0.632747
257      1.228328  -0.632747
173      1.230635   1.580410
242      1.956480  -0.632747
```

Train set -

```
     const   capital   patents     randd  employment    tobinq     value  \
480    1.0 -0.298194 -0.255479 -0.195788   -0.300685 -0.529094 -0.381837
622    1.0 -0.293509 -0.234901 -0.194253   -0.248644 -0.452728 -0.363394
638    1.0 -0.132883 -0.070285 -0.180110    0.111415 -0.259944 -0.087443
389    1.0 -0.295400 -0.234901 -0.196957   -0.299368 -0.150969 -0.375077
748    1.0 -0.258258 -0.245190 -0.179909   -0.251785 -0.307684 -0.311286

     institutions  sp500_yes
480     -0.373791  -0.632747
622     -1.113941  -0.632747
638      0.227003   1.580410
389     -0.847690  -0.632747
748     -1.244528  -0.632747
```

Applying linear regression to train sets fitting OLS model and printing the OLS regression summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     960.3
Date:                Sat, 24 Dec 2022   Prob (F-statistic):         1.37e-306
Time:                        02:05:36   Log-Likelihood:                -13.768
No. Observations:                 531   AIC:                             45.54
Df Residuals:                     522   BIC:                             84.01
Df Model:                           8
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -0.0036      0.011     -0.329      0.742      -0.025       0.018
capital        0.3071      0.020     15.565      0.000       0.268       0.346
patents       -0.0563      0.027     -2.096      0.037      -0.109      -0.004
randd          0.2361      0.029      8.052      0.000       0.179       0.294
employment     0.4170      0.018     23.136      0.000       0.382       0.452
tobinq        -0.0120      0.012     -1.039      0.299      -0.035       0.011
value          0.1027      0.017      5.886      0.000       0.068       0.137
institutions   0.0026      0.012      0.213      0.832      -0.022       0.027
sp500_yes     -0.0052      0.014     -0.375      0.708      -0.032       0.022
==============================================================================
Omnibus:                      231.591   Durbin-Watson:                   1.932
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31508.283
Skew:                           0.809   Prob(JB):                         0.00
Kurtosis:                      40.703   Cond. No.                         7.58
==============================================================================

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

Table 5

How to check for Multicollinearity

There are different ways of detecting (or testing) multicollinearity. One such way is Variation Inflation Factor.

Variance Inflation factor: Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is "inflated" by the existence of correlation among the predictor variables in the model.

General Rule of Thumb:

If VIF is 1, then there is no correlation among the k th predictor and the remaining predictor variables, and hence, the variance of βk is not inflated at all.
If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.

The purpose of the analysis should dictate which threshold to use.

```
VIF values:

const           1.003437
capital         3.454206
patents         5.453357
randd           7.314665
employment      3.171786
tobinq          1.041429
value           2.972769
institutions    1.295102
sp500_yes       1.646031
dtype: float64
```

The VIF values indicate that the features

Multicollinearity affects only the specific independent variables that are correlated. Therefore, in this case, we can trust the p-values of patents and randd variables.

To treat multicollinearity, we will have to drop one or more of the correlated features (patents,randd).

We will drop the variable that has the least impact on the adjusted R-squared of the model.

```
R-squared: 0.936
Adjusted R-squared: 0.935
```

On dropping 'patents', there is no change in adj. R-squared.

```
R-squared: 0.928
Adjusted R-squared: 0.928
```

On dropping 'randd', adj R square dropped by 0.007

```
R-squared: 0.907
Adjusted R-squared: 0.906
```

On dropping 'capital', adjusted Rsquare dropped by 0.029

```
R-squared: 0.871
Adjusted R-squared: 0.869
```

On dropping 'employment', adjusted Rsquare dropped by 0.066 , employment seem to be a significant variable hence can't be dropped.

```
R-squared: 0.936
Adjusted R-squared: 0.935
```

On dropping 'tobinq', adjusted Rsquare didn't drop.

```
R-squared: 0.932
Adjusted R-squared: 0.931
```

On dropping 'value' adj R square dropped by 0.004

```
R-squared: 0.932
Adjusted R-squared: 0.931
```

On dropping 'institutions' adj R square dropped by 0.004

```
R-squared: 0.936
Adjusted R-squared: 0.936
```

On dropping 'sp500_yes' adj R square increased by 0.001

Since there is no effect on adj. R-squared after dropping the 'patents' , 'tobinq' column, we can remove it from the training set.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     1272.
Date:                Sat, 24 Dec 2022   Prob (F-statistic):          1.47e-308
Time:                        02:06:12   Log-Likelihood:                -16.429
No. Observations:                 531   AIC:                             46.86
Df Residuals:                     524   BIC:                             76.78
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.0027      0.011     -0.244      0.807      -0.024       0.019
capital          0.3150      0.019     16.193      0.000       0.277       0.353
randd            0.1870      0.018     10.172      0.000       0.151       0.223
employment       0.4291      0.017     24.881      0.000       0.395       0.463
value            0.0834      0.015      5.476      0.000       0.053       0.113
institutions    -0.0014      0.012     -0.113      0.910      -0.025       0.023
sp500_yes       -0.0043      0.014     -0.306      0.760      -0.032       0.023
==============================================================================
Omnibus:                      225.738   Durbin-Watson:                   1.922
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31707.449
Skew:                           0.744   Prob(JB):                         0.00
Kurtosis:                      40.827   Cond. No.                         4.08
==============================================================================
```

Table 6

```
VIF values:

const           1.002020
capital         3.335841
randd           2.854966
employment      2.886611
value           2.249028
institutions    1.266903
sp500_yes       1.640098
dtype: float64
```

Since there is a very small effect (0.001) on adj. R-squared after dropping the 'sp500_yes' column, we can remove it from the training set.

```
                          OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     1529.
Date:                Sat, 24 Dec 2022   Prob (F-statistic):          3.75e-310
Time:                        02:06:15   Log-Likelihood:                -16.477
No. Observations:                 531   AIC:                             44.95
Df Residuals:                     525   BIC:                             70.60
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.0026      0.011     -0.239      0.811      -0.024       0.019
capital          0.3146      0.019     16.225      0.000       0.276       0.353
randd            0.1873      0.018     10.216      0.000       0.151       0.223
employment       0.4287      0.017     24.950      0.000       0.395       0.463
value            0.0820      0.015      5.645      0.000       0.053       0.111
institutions    -0.0030      0.011     -0.272      0.786      -0.025       0.019
==============================================================================
Omnibus:                      225.660   Durbin-Watson:                   1.922
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31318.780
Skew:                           0.747   Prob(JB):                         0.00
Kurtosis:                      40.594   Cond. No.                         3.92
==============================================================================
```

Table 7

```
VIF values:

const           1.001747
capital         3.319886
randd           2.845405
employment      2.870100
value           2.050564
institutions    1.031616
dtype: float64
```

VIF for all the features is <3.5

Now that we do not have multicollinearity in our data, the p-values of the coefficients have become reliable and we can remove the non-significant predictor variables

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     1529.
Date:                Sat, 24 Dec 2022   Prob (F-statistic):          3.75e-310
Time:                        02:06:17   Log-Likelihood:                -16.477
No. Observations:                 531   AIC:                             44.95
Df Residuals:                     525   BIC:                             70.60
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.0026      0.011     -0.239      0.811      -0.024       0.019
capital          0.3146      0.019     16.225      0.000       0.276       0.353
randd            0.1873      0.018     10.216      0.000       0.151       0.223
employment       0.4287      0.017     24.950      0.000       0.395       0.463
value            0.0820      0.015      5.645      0.000       0.053       0.111
institutions    -0.0030      0.011     -0.272      0.786      -0.025       0.019
==============================================================================
Omnibus:                      225.660   Durbin-Watson:                   1.922
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31318.780
Skew:                           0.747   Prob(JB):                         0.00
Kurtosis:                      40.594   Cond. No.                         3.92
==============================================================================

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

Table 8

As observed in the above model (olsres_11), 'institutions' has a p-value greater than 0.05. So, we can drop it because it is not significant in predicting 'sales'.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     1914.
Date:                Sat, 24 Dec 2022   Prob (F-statistic):          8.36e-312
Time:                        02:06:18   Log-Likelihood:                -16.514
No. Observations:                 531   AIC:                             43.03
Df Residuals:                     526   BIC:                             64.40
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0027      0.011     -0.248      0.804      -0.024       0.019
capital        0.3145      0.019     16.237      0.000       0.276       0.353
randd          0.1877      0.018     10.284      0.000       0.152       0.224
employment     0.4282      0.017     25.124      0.000       0.395       0.462
value          0.0819      0.015      5.644      0.000       0.053       0.110
==============================================================================
Omnibus:                      225.939   Durbin-Watson:                   1.923
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31049.067
Skew:                           0.753   Prob(JB):                         0.00
Kurtosis:                      40.431   Cond. No.                         3.89
==============================================================================

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

After dropping the features causing strong multicollinearity and the statistically insignificant ones, our model performance hasn't dropped. This shows that those variables didn't have predicting power.

Table 9(OLS_RES12)

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.936
Model:	OLS	Adj. R-squared:	0.935
Method:	Least Squares	F-statistic:	1914.
Date:	Sat, 24 Dec 2022	Prob (F-statistic):	8.36e-312
Time:	02:06:31	Log-Likelihood:	-16.514
No. Observations:	531	AIC:	43.03
Df Residuals:	526	BIC:	64.40
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0027	0.011	-0.248	0.804	-0.024	0.019
capital	0.3145	0.019	16.237	0.000	0.276	0.353
randd	0.1877	0.018	10.284	0.000	0.152	0.224
employment	0.4282	0.017	25.124	0.000	0.395	0.462
value	0.0819	0.015	5.644	0.000	0.053	0.110

Omnibus:	225.939	Durbin-Watson:	1.923
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31049.067
Skew:	0.753	Prob(JB):	0.00
Kurtosis:	40.431	Cond. No.	3.89

Observations

R-squared of the model is 0.936 and adjusted R-squared is 0.935, which shows that the model is able to explain ~93.5% variance in the data. This is extremely good.

A unit increase in the capital will result in a 0.3145 unit increase in the sales, all other variables remaining constant.

A unit increase in the R&D will result in a 0.1877 unit increase in the sales, all other variables remaining constant.

A unit increase in the employment will result in a 0.4282 unit increase in the sales, all other variables remaining constant.(MOST IMPORTANT ATTRIBUTE)

A unit increase in the Stock market value will result in a 0.0819 unit increase in the sales, all other variables remaining constant.

R square of training data

93.5% of the variation in the sales is explained by the predictors in the model for train set

Equation of Linear Regression

Sales = -0.0026976549480944656 + 0.3144643716579352 * (capital) + 0.1877071035767714 * (randd) + 0.4281775881727028 * (employment) + 0.08186917601013671 * (value)

We can now use the model for making predictions on the test data.

RMSE on train data

0.24961439816650766

Let us plot the fitted values vs residuals

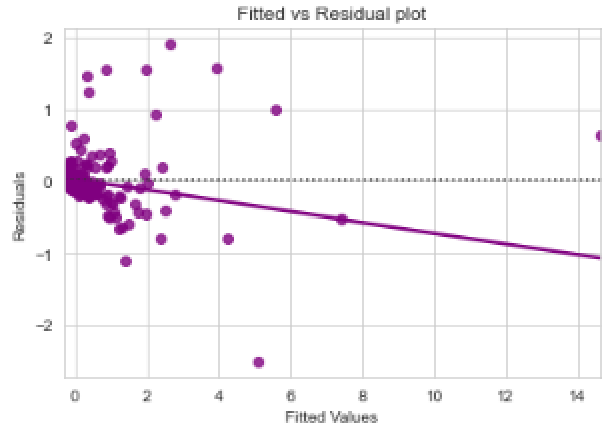


Figure 7

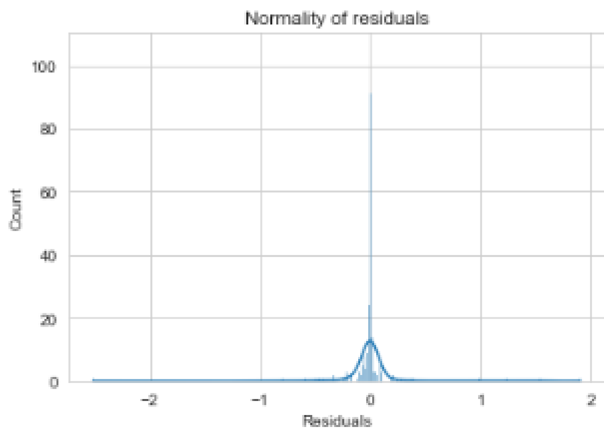


Figure 8

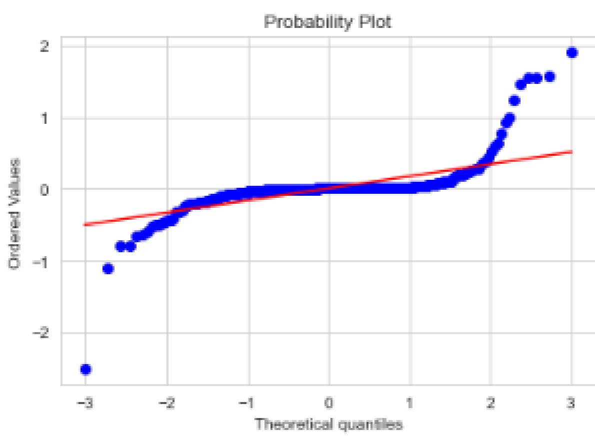


Figure 9(QQ plot)

<u>RMSE on test data -
0.3520590072129984</u>

<u>MAE on train data -</u>

0.08419071412781504

MAE on test data -
0.07218065761152341

We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
MAE indicates that our current model is able to predict sales within a mean error of 0.07 units on the test data.
Hence, we can conclude the model "ols_res12" is excellent for prediction as well as inference purposes.

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

Conclusion

The final Linear Regression equation is

sales = -0.0026976549480944656 + 0.3144643716579352 * (capital) + 0.1877071035767714 * (randd) + 0.4281775881727028 * (employment) + 0.08186917601013671 * (value)

R-squared of the model is 0.936 and adjusted R-squared is 0.935, which shows that the model is able to explain ~93.5% variance in the data. This is extremely good.

A unit increase in the capital will result in a 0.3145 unit increase in the sales, all other variables remaining constant.

A unit increase in the R&D will result in a 0.1877 unit increase in the sales, all other variables remaining constant.

A unit increase in the employment will result in a 0.4282 unit increase in the sales, all other variables remaining constant.(MOST IMPORTANT ATTRIBUTE)

A unit increase in the Stock market value will result in a 0.0819 unit increase in the sales, all other variables remaining constant.

Insights & Recommendations

1)5 most important attributes are "capital","randd","employment","value" & "sp500".
Sales can now be predicted using our final linear regression model equation and we can see how each attribute affects the sales.

2) Among all attributes employment seems to be greatly affecting sales.

3) More Capital, More R&D, More Employment & More stock value will greatly affect sales of firms and we can use our linear regression model to predict sales.

Problem 2

Problem 2 :Logistic Regression, LDA

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.

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.

Reading the dataset

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
0	55+	27.078	Not_Survived	none	none	1	m	32	1997	1987	unavail	driver	0	4.0	2:13:02
1	25-39	89.627	Not_Survived	airbag	belted	0	f	54	1997	1994	nodeploy	driver	0	4.0	2:17:01
2	55+	27.078	Not_Survived	none	belted	1	m	67	1997	1992	unavail	driver	0	4.0	0.138206019
3	55+	27.078	Not_Survived	none	belted	1	f	64	1997	1992	unavail	pass	0	4.0	0.138206019
4	55+	13.374	Not_Survived	none	none	1	m	23	1997	1986	unavail	driver	0	4.0	4:58:01

Table 10

Data Types

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11217 entries, 0 to 11216
Data columns (total 15 columns):
 #   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  int64  
 10  abcat        11217 non-null  object 
 11  occRole      11217 non-null  object 
 12  deploy       11217 non-null  int64  
 13  injSeverity  11140 non-null  float64
 14  caseid       11217 non-null  object 
dtypes: float64(2), int64(5), object(8)
memory usage: 1.3+ MB
```

Describing the data

	weight	frontal	ageOFocc	yearacc	yearVeh	deploy	injSeverity
count	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000
mean	431.405309	0.644022	37.427654	2001.103236	1994.177944	0.389141	1.826781
std	1406.202941	0.478830	18.192429	1.056805	5.658704	0.487577	1.373871
min	0.000000	0.000000	16.000000	1997.000000	1953.000000	0.000000	0.000000
25%	28.292000	0.000000	22.000000	2001.000000	1991.000000	0.000000	1.000000
50%	82.195000	1.000000	33.000000	2001.000000	1995.000000	0.000000	2.000000
75%	324.056000	1.000000	48.000000	2002.000000	1999.000000	1.000000	3.000000
max	31694.040000	1.000000	97.000000	2002.000000	2003.000000	1.000000	5.000000

Table 11

Imputing the null values

```
dvcat          0
weight         0
Survived       0
airbag         0
seatbelt       0
frontal        0
sex            0
ageOFocc       0
yearacc        0
yearVeh        0
abcat          0
occRole        0
deploy         0
injSeverity   77
caseid         0
dtype: int64
```

There are 77 null values in 'injSeverity' Column

11217 Rows & 15 columns

Filling the null values by median of respective column

```
dvcat          0
weight         0
Survived       0
airbag         0
seatbelt       0
frontal        0
sex            0
ageOFocc       0
yearacc        0
yearVeh        0
abcat          0
occRole        0
deploy         0
injSeverity    0
caseid         0
dtype: int64
```

Null values are replaced by median values of the coloumn. 'caseid' coloumn is dropped.

UNIVARIATE ANALYSIS

```
dvcat
24-Oct     5414
25-39      3368
40-54      1344
55+         809
1-9km/h     282
Name: dvcat, dtype: int64
```

```
Survived
survived        10037
Not_Survived     1180
Name: Survived, dtype: int64
```

```
airbag
airbag    7064
none      4153
Name: airbag, dtype: int64
```

```
seatbelt
belted    7849
none      3368
Name: seatbelt, dtype: int64
```

```
sex
m    6048
f    5169
Name: sex, dtype: int64
```

'dvcat' has many entries as 24-Oct. As 10-24 group is missing we will have to change them to 10-24.

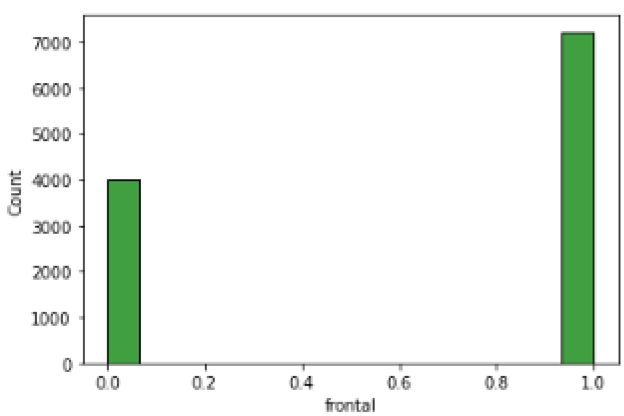
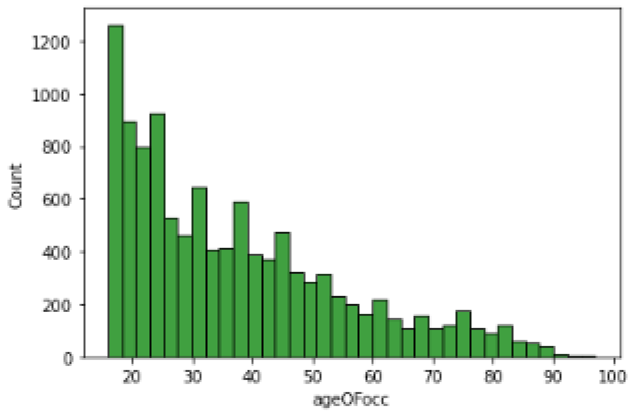


Figure 11



BoxPlot of ageOFocc

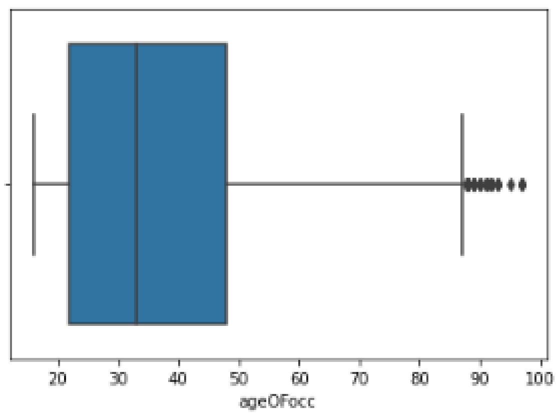


Figure 12

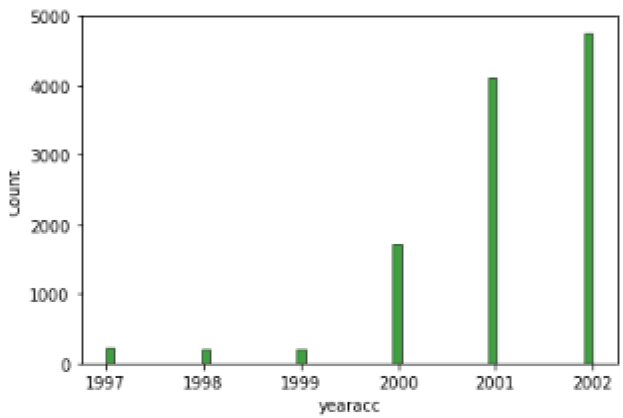
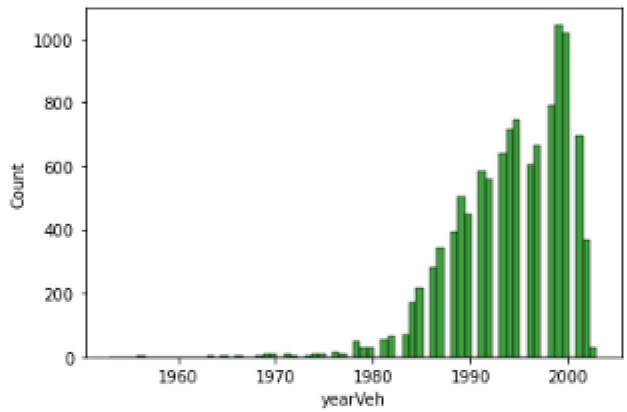


Figure 13




```
BoxPlot of yearVeh
----------------------------------------------------
```

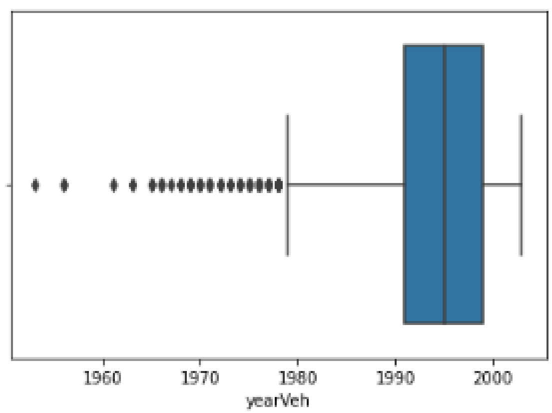


Figure 14

There are many outliers in the data.

Frontal impacts are more as compared to non frontal impacts.

The average age of occupants of car is 37 years, minimum age is 16 years and maximum 97 years.

Count of accidents has increased drastically from around 200 (year 1997) to around 4800 (year 2002).

The model year of vehicles are maximum from the year 1998-2000.

```
Details of dvcat
------------------------------------------------------
10-24       5414
25-39       3368
40-54       1344
55+          809
1-9km/h      282
Name: dvcat, dtype: int64
```

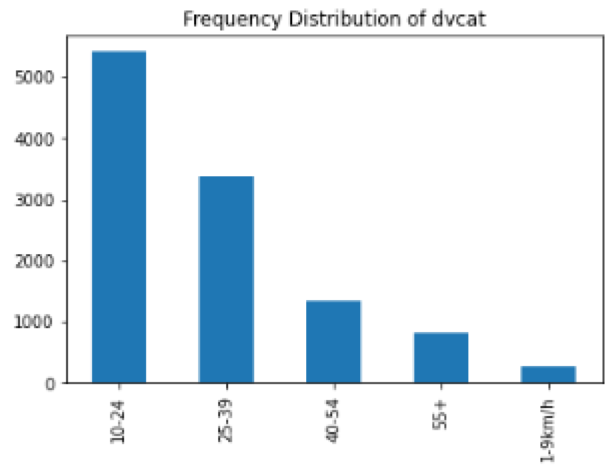


Figure 15

```
Details of Survived
---------------------------------------------------
survived        10037
Not_Survived     1180
Name: Survived, dtype: int64
```

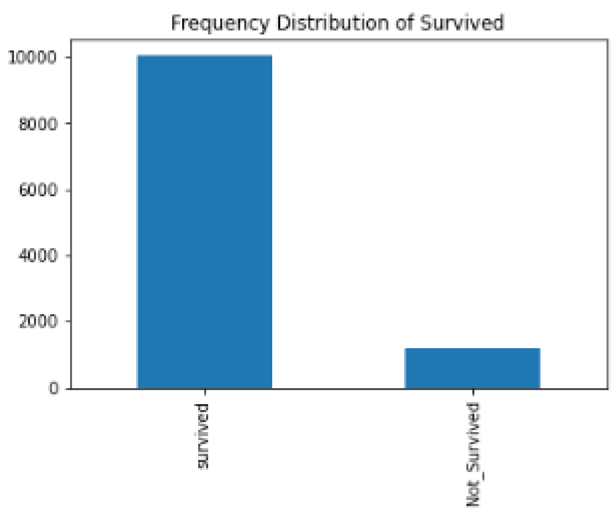


Figure 16

```
Details of airbag
--------------------------------------------------
airbag    7064
none      4153
Name: airbag, dtype: int64
```

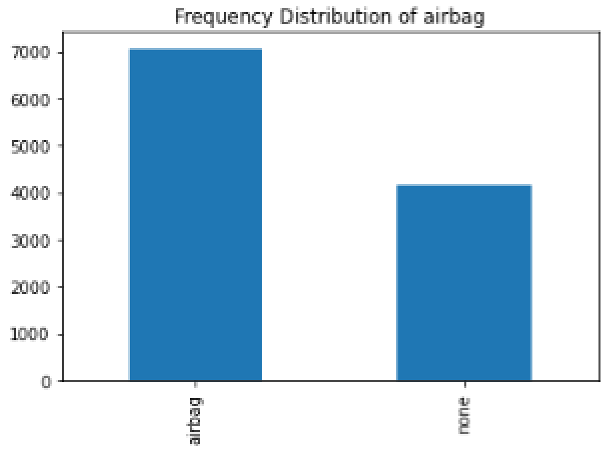


Figure 17

```
Details of seatbelt
--------------------------------------------------
belted    7849
none      3368
Name: seatbelt, dtype: int64
```

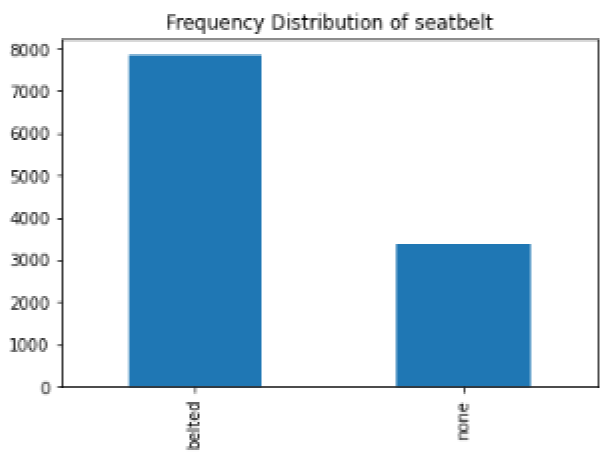


Figure 18

Highest number of count of impact speed was of range 10-24 km per hour.

Only 10037 survived and 1180 didn't survive.

7064 cars did have airbags and 4153 cars didn't have it.

7849 were seatbelted and 3368 weren't seatbelted.

BIVARIATE ANALYSIS
Using Pairplot -

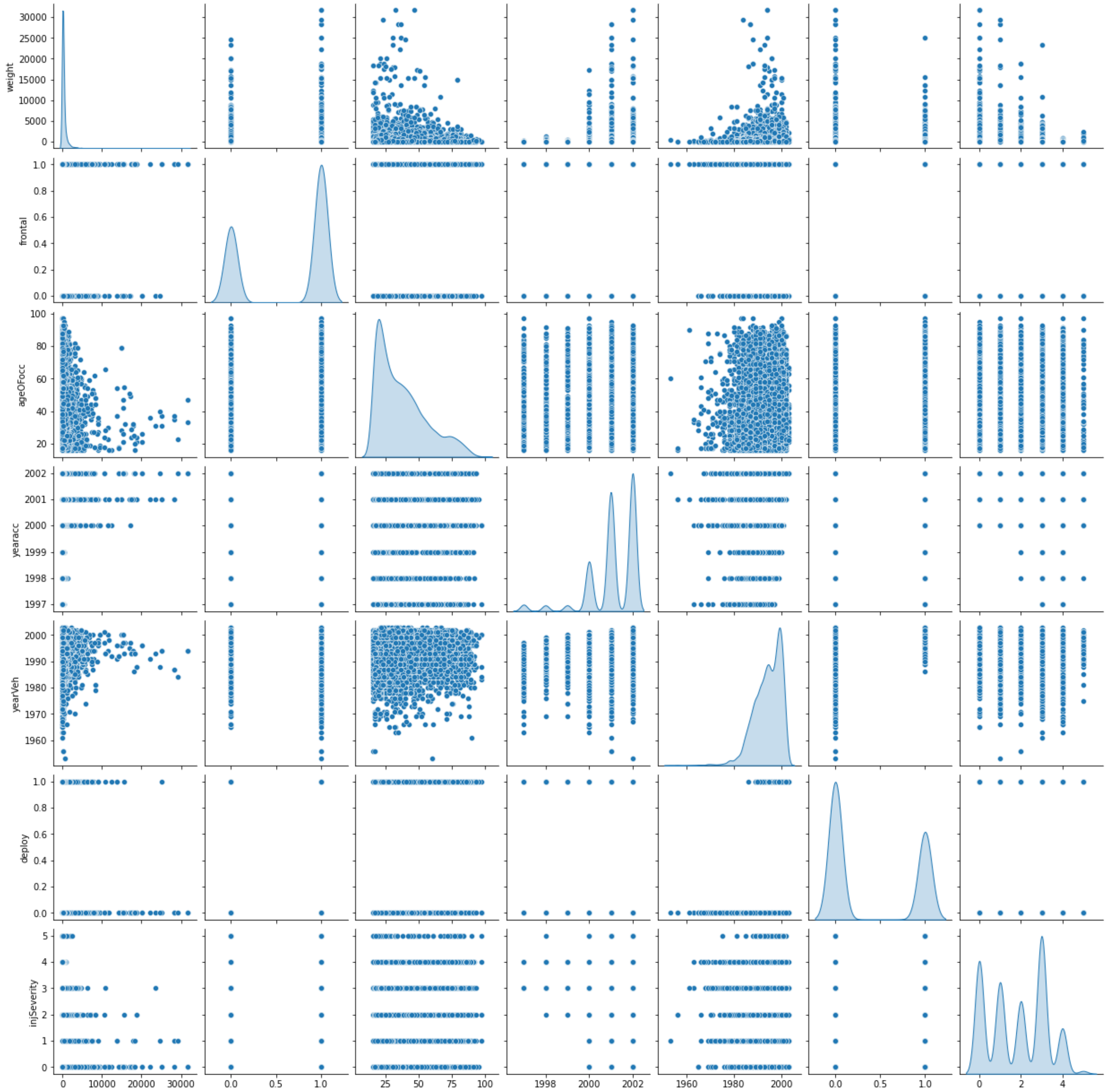
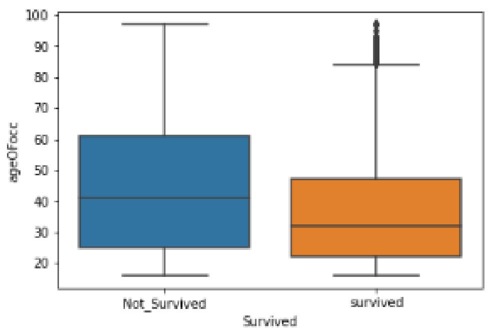
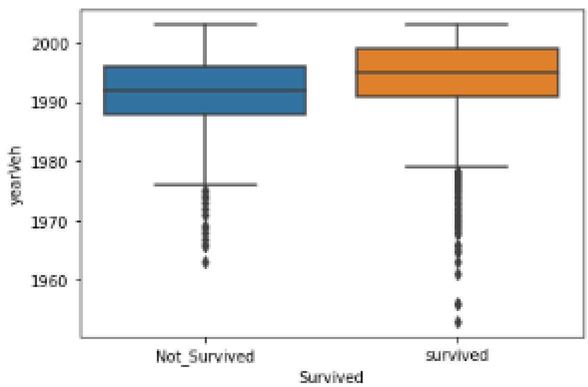


Figure 19

There seems to be non linear relationship between weight and ageoFocc.



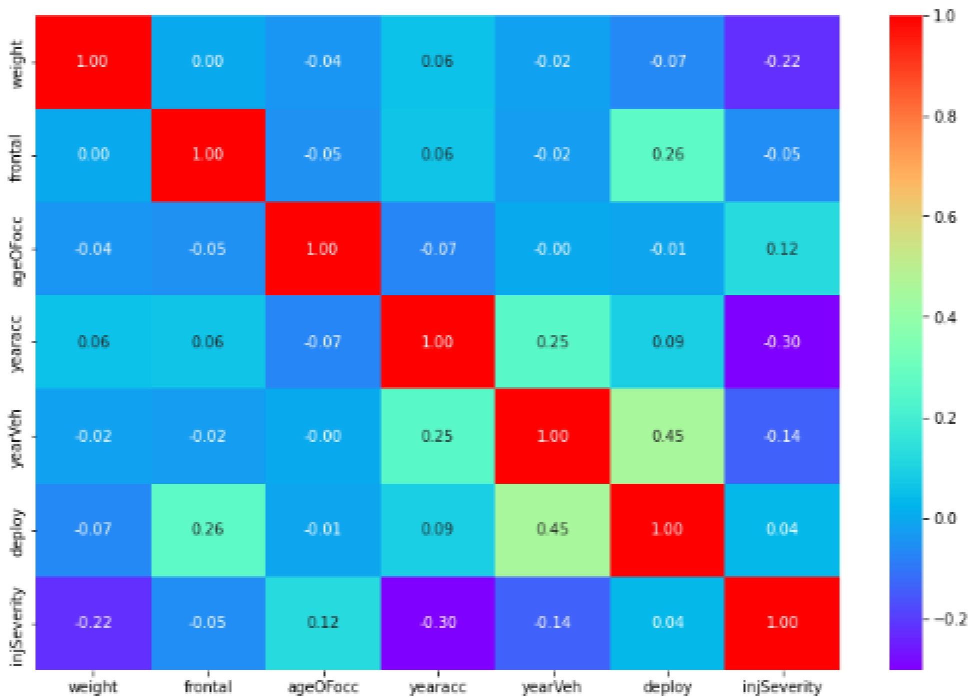
ageOFocc shows some skewness in the distribution between survived and not survived. Distribution is much wider for not survived. Median of Not survived is higher than Survived.



The distribution is almost similar.

Figure 20

MULTIVARIATE ANALYSIS



There seems to be a correlation between frontal & deploy, deploy & yearVeh.

Figure 21

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

Converting all objects to categorical codes

Converting the 'Survived' Variable into numeric by using the LabelEncoder functionality inside sklearn.

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity
0	55+	27.078	0	none	none	1	m	32	1997	1987	unavail	driver	0	4.0
1	25-39	89.627	0	airbag	belted	0	f	54	1997	1994	nodeploy	driver	0	4.0
2	55+	27.078	0	none	belted	1	m	67	1997	1992	unavail	driver	0	4.0
3	55+	27.078	0	none	belted	1	f	64	1997	1992	unavail	pass	0	4.0
4	55+	13.374	0	none	none	1	m	23	1997	1986	unavail	driver	0	4.0

0 means not survived and 1 means survived

Converting the other 'object' type variables as dummy variables

	weight	Survived	frontal	ageOFocc	yearacc	yearVeh	deploy	injSeverity	dvcat_10-24	dvcat_25-39	dvcat_40-54	dvcat_55+	airbag_none	seatbelt_none	sex_m
0	27.078	0	1	32	1997	1987	0	4.0	0	0	0	1	1	1	1
1	89.627	0	0	54	1997	1994	0	4.0	0	1	0	0	0	0	0
2	27.078	0	1	67	1997	1992	0	4.0	0	0	0	1	1	0	1
3	27.078	0	1	64	1997	1992	0	4.0	0	0	0	1	1	0	0
4	13.374	0	1	23	1997	1986	0	4.0	0	0	0	1	1	1	1

Table 12

Train Test Split

Split X and y into training and test set in 70:30 ratio

Build Logistic Regression model

Fit the Logistic Regression model using newton cg as it's a multiclass problem

Model Score of train data(Accuracy) - **0.980639408992485(98%)**

Build LDA(Linear discriminant analysis) Model

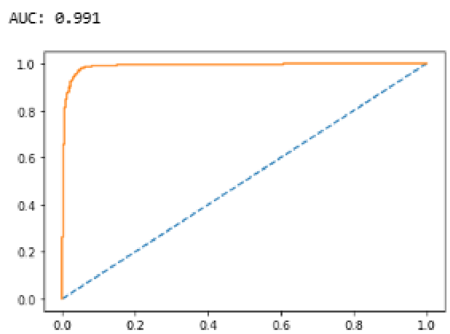
Model score of train data(Accuracy) **- 0.9575850210164311(96%)**

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.

<u>PERFORMANCE OF LOGISTIC REGRESSION MODEL</u>

Model Score(Accuracy of train data) - 0.980639408992485(98%)

ROC Curve & AUC Score(0.991) of train data



Model Score(Accuracy of test data) - 0.9833630421865716(98%)

ROC Curve & AUC Score(0.991) of test data

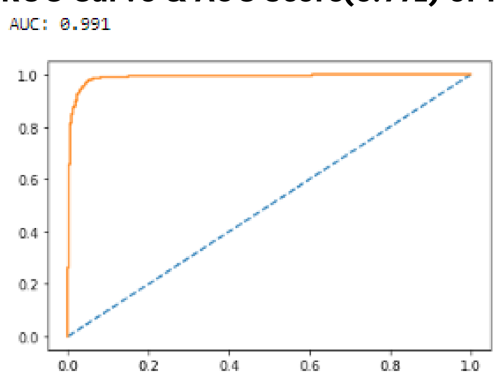
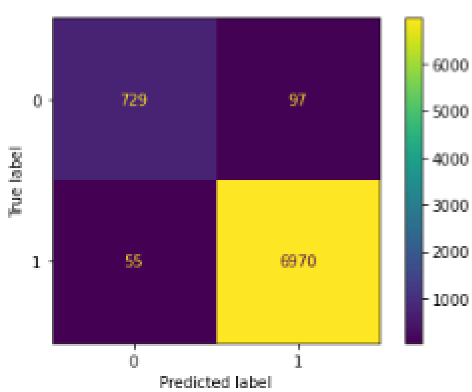
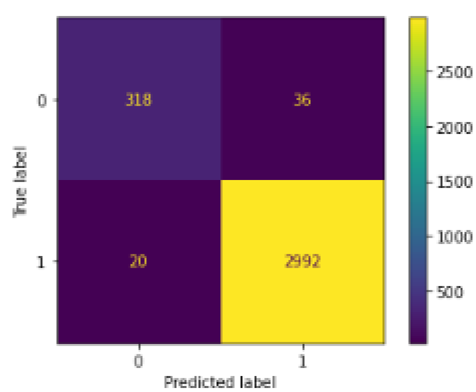


Figure 22

Train & test data confusion matrix





Train data **Test data** **Figure 23**

Train data & test data classification report

Train data

```
              precision    recall  f1-score   support

           0       0.93      0.88      0.91       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
```

Test data

```
              precision    recall  f1-score   support

           0       0.94      0.90      0.92       354
           1       0.99      0.99      0.99      3012

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

Model Coefficients

```
array([[ 0.0000e+00,  1.3200e+00, -4.0000e-02,  1.0600e+00, -0.0000e+00,
        -5.9648e+02, -4.0200e+00, -2.1100e+00, -3.3900e+00, -4.2000e+00,
        -5.1600e+00, -2.9830e+02, -7.4000e-01, -4.3000e-01, -5.9602e+02,
        -2.9830e+02, -5.5000e-01]])
```

Model Variables

```
Index(['weight', 'frontal', 'ageOFocc', 'yearacc', 'yearVeh', 'deploy',
       'injSeverity', 'dvcat_10-24', 'dvcat_25-39', 'dvcat_40-54', 'dvcat_55+',
       'airbag_none', 'seatbelt_none', 'sex_m', 'abcat_nodeploy',
       'abcat_unavail', 'occRole_pass'],
      dtype='object')
```

CONCLUSION

Note :

Precision : tells us how many predictions are actually positive

out of all the total positives predicted.

Recall: how many observations of positive class are actually

predicted as positive.

Inferences :

For predicting Not survived (Label 0):

Precision (94%) – 94% of people predicted actually not to survive out of all people predicted to not survive.

Recall (90%) – Out of all the people actually not surviving, 90% of people have been predicted correctly .

For predicting Survived (Label 1):

Precision (99%) – 99% of employees predicted actually survive out of all people predicted to survive.

Recall (99%) – Out of all the people who actually survived , 99% of employees have been predicted correctly .

Overall accuracy of the model – 98 % of total predictions are correct
Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and **overall the model is a good model for classification.**

PERFORMANCE OF LDA MODEL

Linear Discriminant Function
= -4773.68800372 + (-1.20355227e-04xweight)+ (9.10496861e-01xfrontal)+ (-2.87947123e-02xageOFocc)+....+(-4.30119992e-01xoccRole_pass)

Coefficients -

```
array([[-0.  ,  0.91, -0.03,  2.41, -0.02,  0.09, -1.4 ,  0.23,  0.17,
        -1.47, -4.94, -0.07, -0.48, -0.46, -0.03, -0.07, -0.43]])
```

By the above equation and the coefficients it is clear that

- predictor 'yearacc' has the largest magnitude thus this helps in classifying the best
- predictor 'dvcat_55+' has the smallest magnitude thus this helps in classifying the least

Training Data and Test Data Confusion Matrix Comparison

Figure 24

Training Data and Test Data Classification Report Comparison

```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.86      0.72      0.78       826
           1       0.97      0.99      0.98      7025

    accuracy                           0.96      7851
   macro avg       0.91      0.85      0.88      7851
weighted avg       0.96      0.96      0.96      7851


Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.85      0.71      0.78       354
           1       0.97      0.99      0.98      3012

    accuracy                           0.96      3366
   macro avg       0.91      0.85      0.88      3366
weighted avg       0.95      0.96      0.96      3366
```

AUC & ROC FOR TRAINING AND TEST DATA

```
AUC for the Training Data: 0.968
AUC for the Test Data: 0.967
```

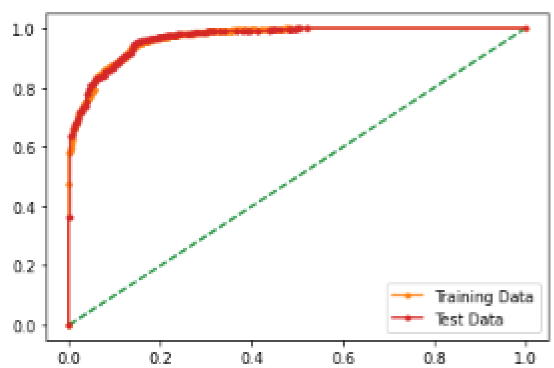


Figure 25

CONCLUSION

Note :

Precision : tells us how many predictions are actually positive out of all the total positives predicted.

Recall : how many observations of positive class are actually predicted as positive.

Inferences :
For predicting Not survived (Label 0):

Precision (85%) – 85% of people predicted actually not to survive out of all people predicted to not survive.

Recall (71%) – Out of all the people actually not surviving, 71% of people have been predicted correctly .

For predicting Survived (Label 1):

Precision (97%) – 97% of employees predicted actually survive out of all people predicted to survive.

Recall (99%) – Out of all the people who actually survived , 99% of employees have been predicted correctly .

Overall accuracy of the model – 96 % of total predictions are correct
Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification.

LOGISTIC REGRESSION MODEL IS BEST/OPTIMISED as the accuracy,AUC,Precision and Recall are greater as compared to the LDA model.

2.4) Inference: Based on these predictions, what are the insights and recommendations?
<u>INSIGHTS AND RECOMMENDATIONS</u>

1. Based on the logistic regression model with 98% accuracy, frontal impact seems to be the most important attribute in determining whether a person has survived or not.

2. To decrease the frontal impact, the government should make strict laws for manufacturers of cars to include good airbags in all cars for the safety of all passengers.

3. Strict laws should be made to ensure passengers always put their seatbelts to lesser the impact of car crash.

4. Good quality airbags with a very fast deploy rate should be installed in all cars to ensure safety of all passengers.

5. Regular airbag testing should be made mandatory by the government to ensure that airbags do deploy in all cars when there is an impact of a car crash.

6. Estimated impact speeds were of range 10-24 km/hr, design should be built to reduce impact speed to decrease damage to passengers.

THANK YOU