12/30/2023

BUSINESS REPORT

PREDICTIVE MODELLING

KAVIN BHARATHI

Contents

List Of Tables

List Of Figures

Problem Statement 1: Linear Regression

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

Data Dictionary

Column Name	Description	Data Type
sales	Sales (in millions of dollars).	Float
capital	Net stock of property, plant, and equipment.	Float
patents	Granted patents.	Integer
randd	R&D stock (in millions of dollars).	Float
employment	Employment (in 1000s).	Float
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	Object
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.	Float
value	Stock market value.	Float
institutions	Proportion of stock owned by institutions	Float

Table 1: Data Description - Dataset 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.

EDA

The Data is imported and below are the observations:

- The Data has 759 Rows and 9 Columns
    - **Data types:** Float – 7, Integer – 1, Object – 1

	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
count	759.000000	759.000000	759.000000	759.000000	759.000000	759	738.000000	759.000000	759.000000
unique	NaN	NaN	NaN	NaN	NaN	2	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN	no	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN	542	NaN	NaN	NaN
mean	2689.705158	1977.747498	25.831357	439.938074	14.164519	NaN	2.794910	2732.734750	43.020540
std	8722.060124	6466.704896	97.259577	2007.397588	43.321443	NaN	3.366591	7071.072362	21.685586
min	0.138000	0.057000	0.000000	0.000000	0.006000	NaN	0.119001	1.971053	0.000000
25%	122.920000	52.650501	1.000000	4.628262	0.927500	NaN	1.018783	103.593946	25.395000
50%	448.577082	202.179023	3.000000	36.864136	2.924000	NaN	1.680303	410.793529	44.110000
75%	1822.547366	1075.790020	11.500000	143.253403	10.050001	NaN	3.139309	2054.160386	60.510000
max	135696.788200	93625.200560	1220.000000	30425.255860	710.799925	NaN	20.000000	95191.591160	90.150000

Table 2: Data Summary

- The Object Variable – sp500 contains two categories and majority of it belongs to No Category
- There are no duplicates in the dataset.
- 21 missing values found in the Tobinq column and it will be treated later.

Univariate Analysis:

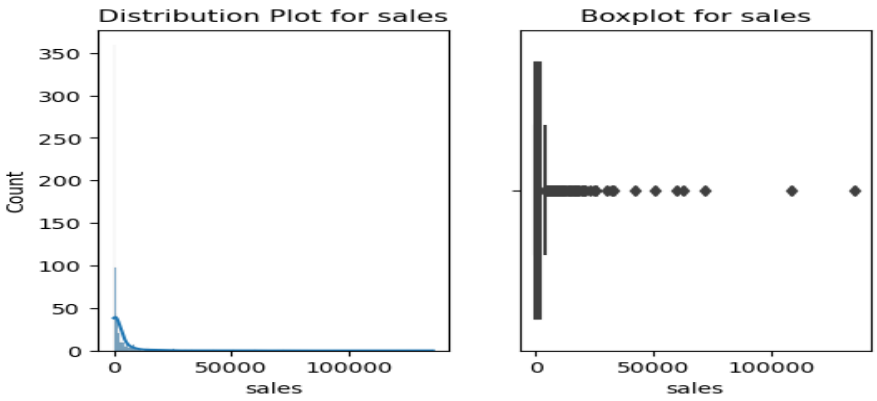

Figure 1 - Distribution & Boxplot of Sales

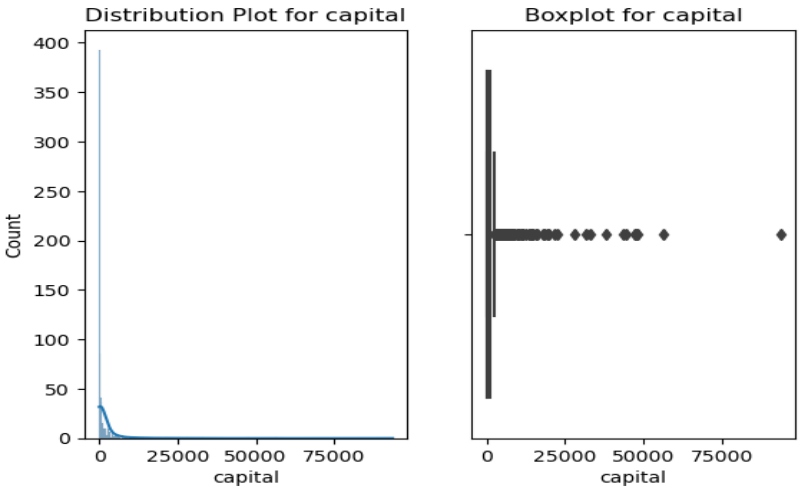

Figure 2 - Distribution & Boxplot of Capital

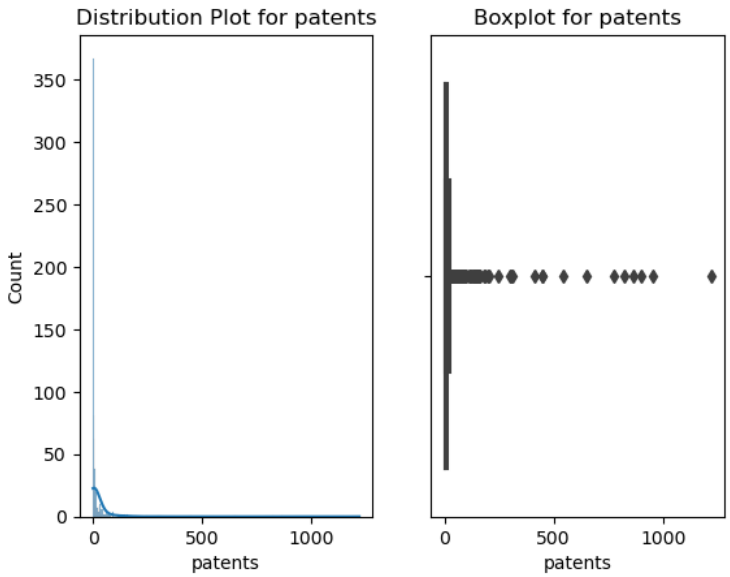

Figure 3 - Distribution & Boxplot of Patents

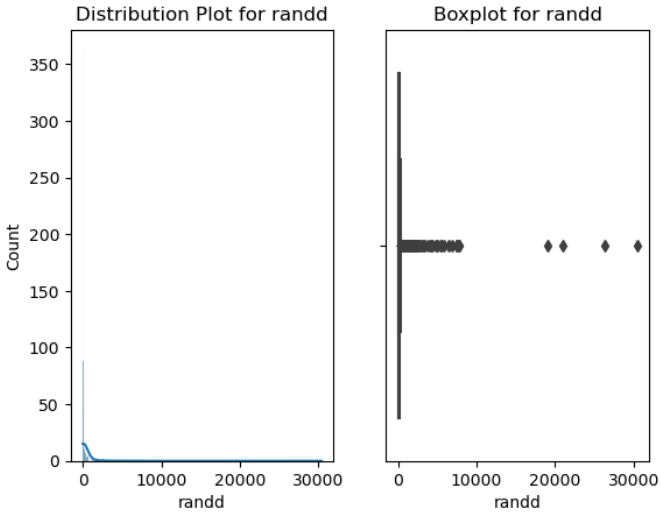

Figure 4 - Distribution & Boxplot of R&D

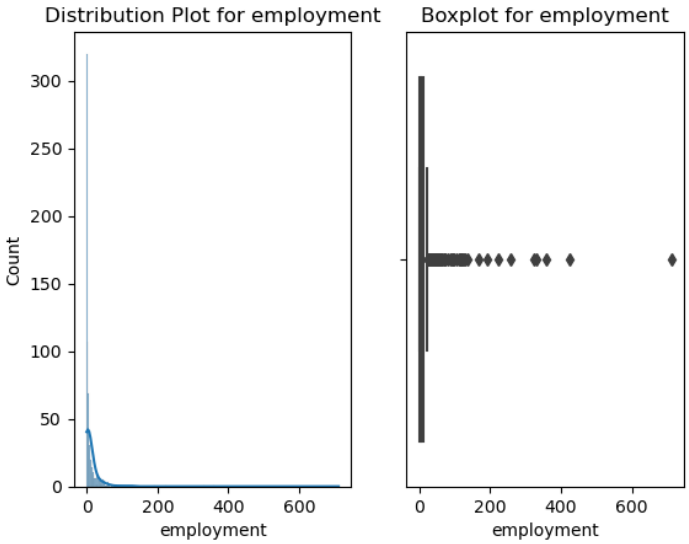

Figure 5 - Distribution & Boxplot of Employment

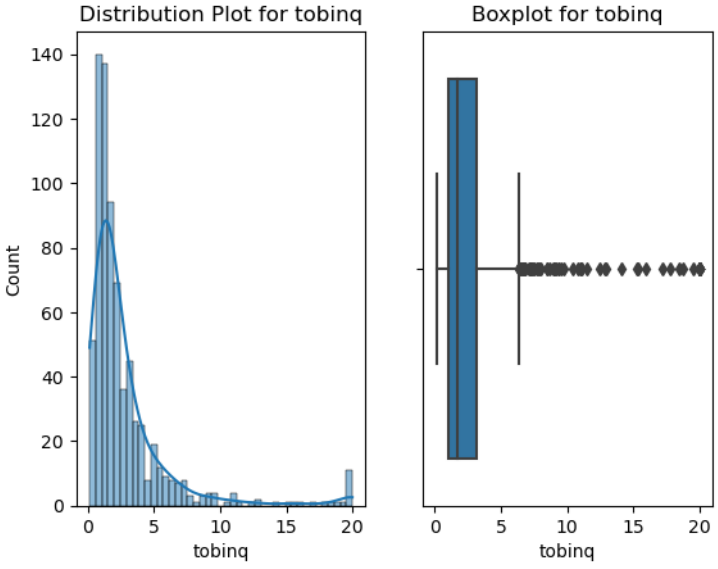

Figure 6 - Distribution & Boxplot of Tobinq

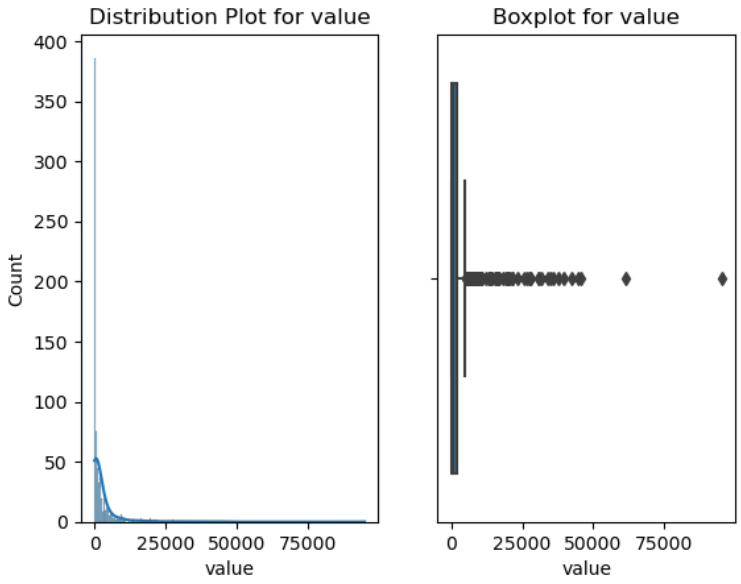

Figure 7 - Distribution & Boxplot of Values

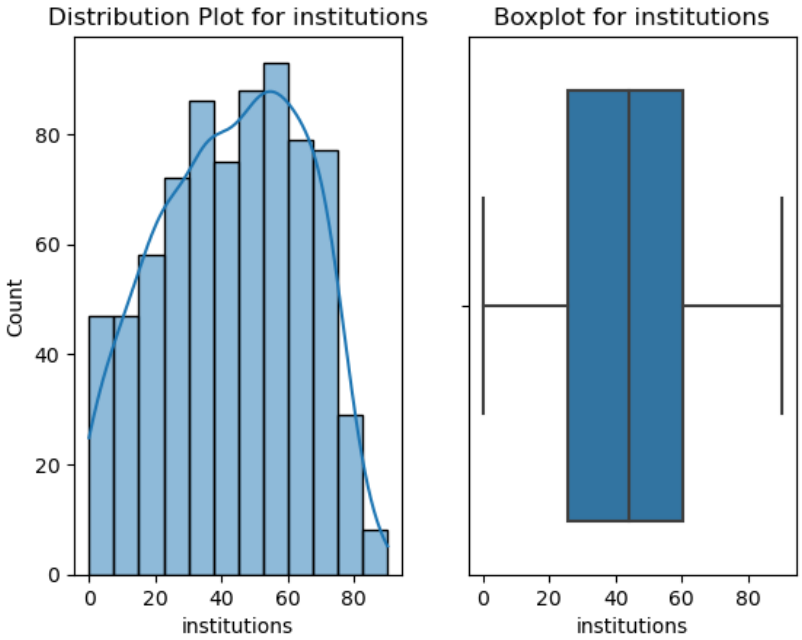

Figure 8 - Distribution & Boxplot of Institutions

- Mostly the sales of the firms is around 10000.
- Most of the firms are having capital less than 25000 and the patents acquired by them were also in small quantities. So, this indicates that most of the firms are smaller or mid-sized ones.
- However, there are certain outlier population indicating the firms belonging to larger sized.

Bivariate Analysis:

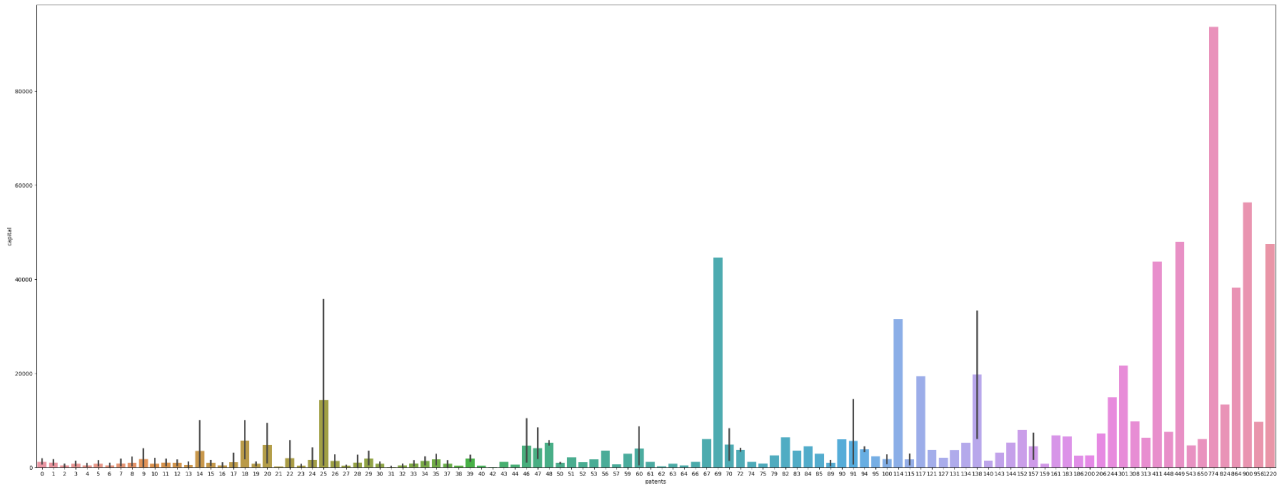

Figure 9 - Distribution of Capital and Patents

- When the capital is High the Patents is also high. This indicates that firms with higher capital is holding much patents.

Multivariate Analysis:

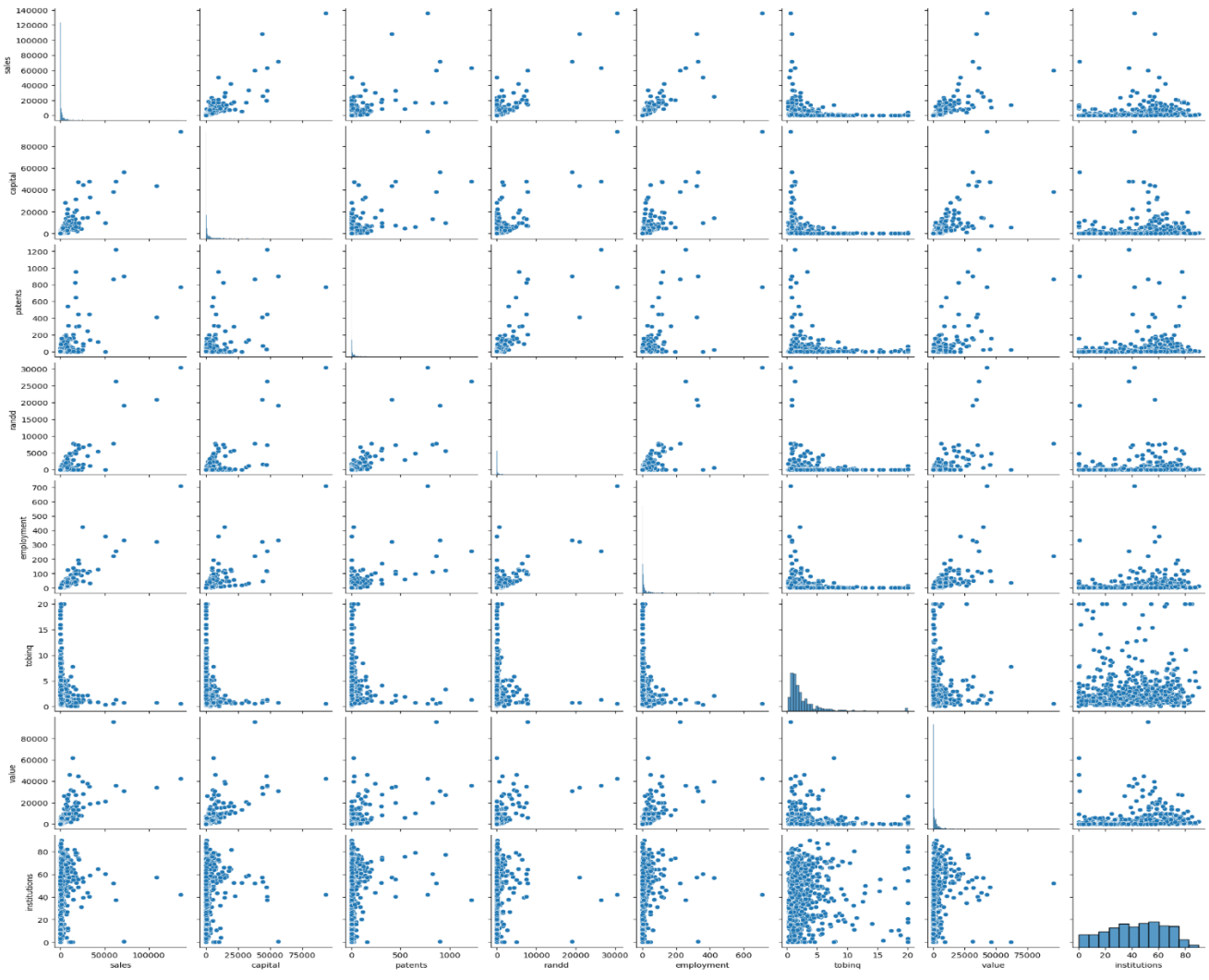

Figure 10 - Pairplot

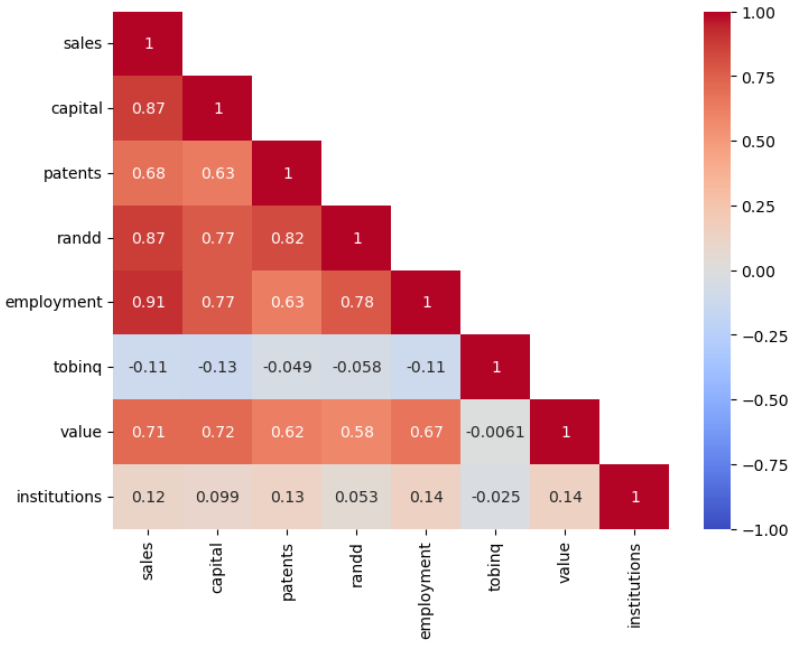

Figure 11 - Heatmap

- From the above pair plot and heatmap we can see that there is a strong correlation between few variables like Employment and sales, R & D stock and Sales, Capital and Sales, R & D stock and Patents.

Outlier Detection & Treatment:

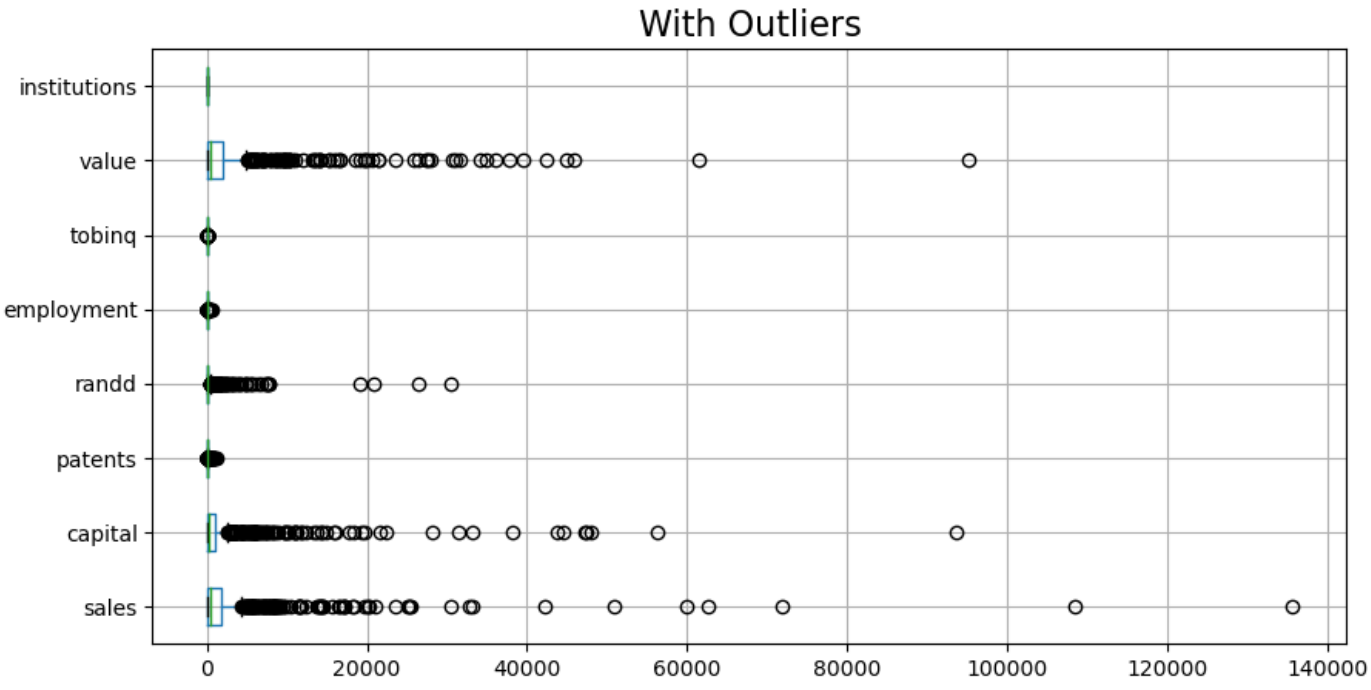

Figure 12 - Before Outlier Treatment

- The black circles represent the outliers and it is present in all the columns except Institutions.
- Majority of the variables are highly skewed towards right.

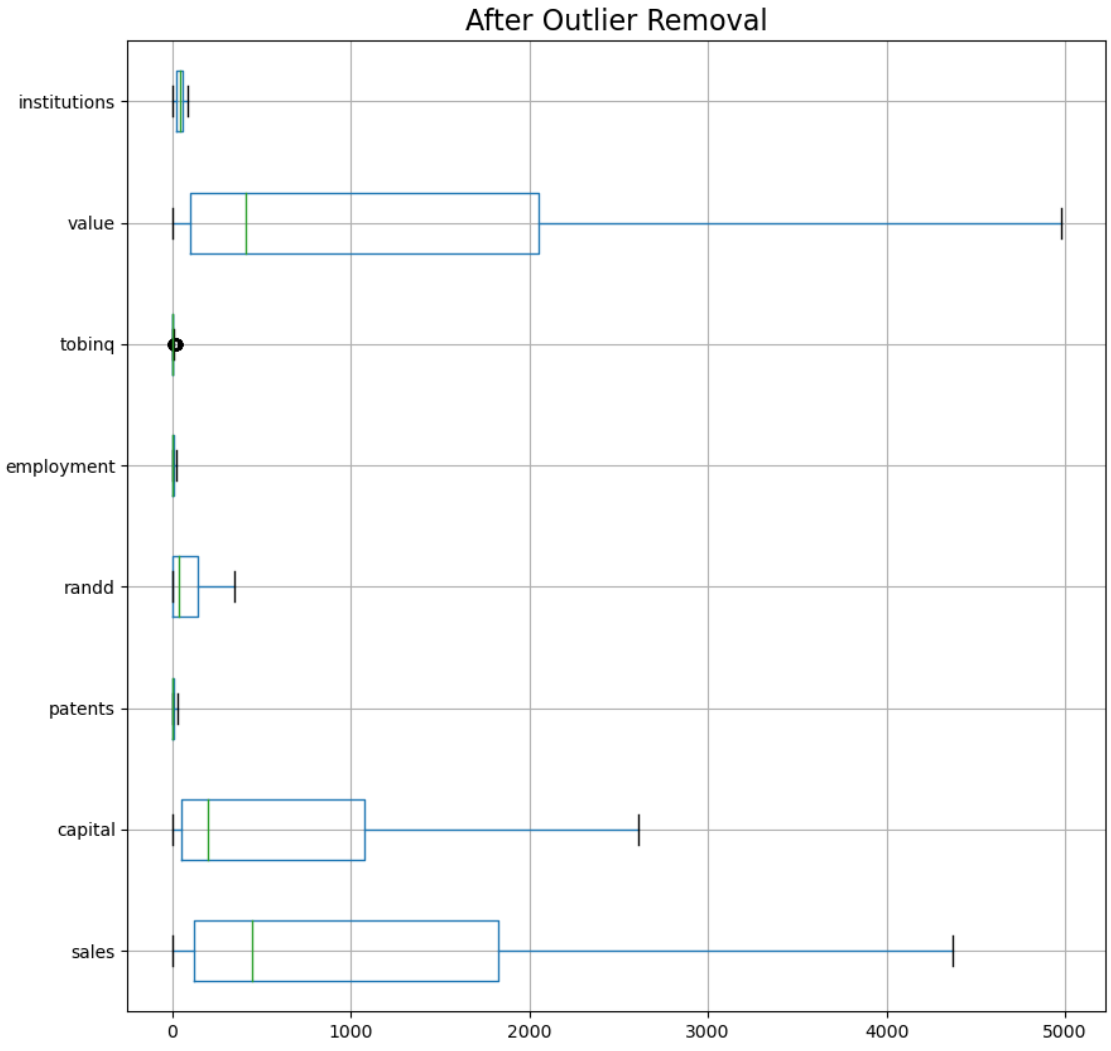

Figure 13 - After Outlier Treatment

- The outliers are treated by IQR value as seen above.

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

- The variable **tobinq** is having 21 null values and it has been replaced using median after finding the skewness value as 3.29.
- Yes, scaling can be done as it enhances the performance of the algorithm.

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.

- One Hot encoding is done for the column **sp500.**

	sales	capital	patents	randd	employment	tobinq	value	institutions	sp500_yes
0	-0.267788	-0.591504	0.221152	1.979986	-0.564800	2.493756	0.142598	1.718839	0
1	-0.542217	-0.632706	-0.583181	-0.782879	-0.619331	-0.577847	-0.645807	0.738279	0
2	2.052715	1.962722	1.955496	1.979986	2.055116	0.734749	2.055843	0.215929	1
3	-0.513909	-0.481679	-0.683723	-0.125658	-0.471265	-0.740066	-0.748521	-0.744789	0
4	-0.694622	-0.613908	-0.583181	-0.670901	-0.608694	-0.511899	-0.746022	0.297142	0

Table 3 – Sample dataset after Encoding

Train and Test Split

- The model for test and train has been split into 70:30 Ratio.
- The Linear Regression Model is built and fitted into the Training dataset.
- The feature importance is derived and we can see the co efficient of the variables below.

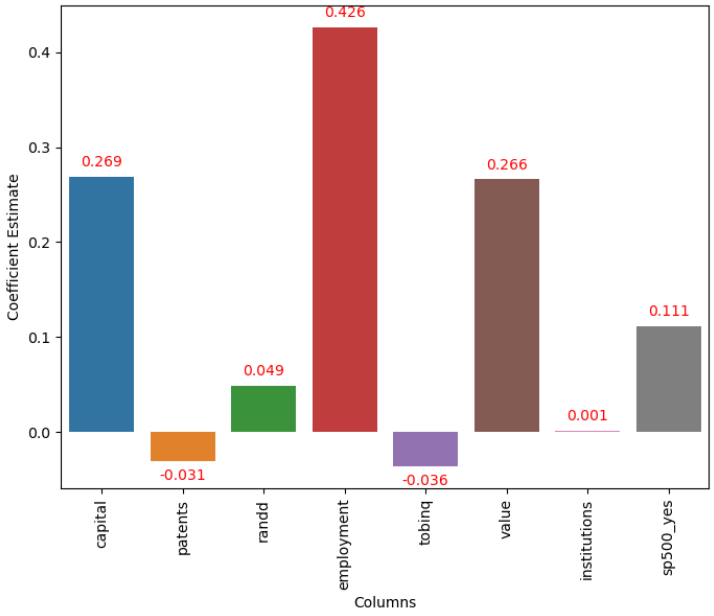

Figure 14 – Top Variables with Co-Efficient

- It clearly shows that Employment plays a crucial role followed by Capital and Value variable.
- These three important variables are the driving factors of the model and hence we can conclude that even if the other variables are not present or close to zero these three variables can be focused.

Model Performance

Model 1:

- To check the model performance, we calculate R Square value.
- **R Square for Train data: 0.935**
- **RMSE for Train Data: 0.259**
- **R Square for Train data: 0.923**
- **RMSE for Train Data: 0.263**

This is a very good model as it shows 92% variance of the testing data was captured by the model.

Model 2:

- The Sklearn model using OLS method gives an similar response as seen below.
- **R Square for Train data: 0.935**

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.935
Model:	OLS	Adj. R-squared:	0.934
Method:	Least Squares	F-statistic:	945.6
Date:	Sat, 30 Dec 2023	Prob (F-statistic):	6.02e-305
Time:	17:13:13	Log-Likelihood:	-36.474
No. Observations:	531	AIC:	90.95
Df Residuals:	522	BIC:	129.4
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0291	0.017	-1.733	0.084	-0.062	0.004
capital	0.2689	0.025	10.547	0.000	0.219	0.319
patents	-0.0306	0.018	-1.677	0.094	-0.067	0.005
randd	0.0494	0.019	2.547	0.011	0.011	0.088
employment	0.4264	0.025	16.763	0.000	0.376	0.476
tobinq	-0.0358	0.013	-2.692	0.007	-0.062	-0.010
value	0.2656	0.028	9.405	0.000	0.210	0.321
institutions	0.0013	0.013	0.099	0.921	-0.024	0.026
sp500_yes	0.1111	0.044	2.544	0.011	0.025	0.197

Omnibus:	183.987	Durbin-Watson:	1.955
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1261.544
Skew:	1.341	Prob(JB):	1.15e-274
Kurtosis:	10.059	Cond. No.	8.46

Figure 15 – OLS Regression Results for Train Data

- **R Square for Test data: 0.932**

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	373.1
Date:	Sat, 30 Dec 2023	Prob (F-statistic):	4.75e-123
Time:	17:41:46	Log-Likelihood:	-5.9422
No. Observations:	228	AIC:	29.88
Df Residuals:	219	BIC:	60.75
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0308	0.024	1.288	0.199	-0.016	0.078
capital	0.2368	0.046	5.126	0.000	0.146	0.328
patents	-0.0182	0.027	-0.670	0.504	-0.072	0.035
randd	-0.0175	0.030	-0.591	0.555	-0.076	0.041
employment	0.4880	0.042	11.610	0.000	0.405	0.571
tobinq	-0.0348	0.018	-1.919	0.056	-0.071	0.001
value	0.3181	0.038	8.436	0.000	0.244	0.392
institutions	0.0290	0.019	1.498	0.136	-0.009	0.067
sp500_yes	-0.1146	0.058	-1.989	0.048	-0.228	-0.001

Omnibus:	133.046	Durbin-Watson:	2.228
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1376.268
Skew:	2.064	Prob(JB):	1.40e-299
Kurtosis:	14.306	Cond. No.	7.11

Figure 16 - OLS Regression Results for Test Data

- Using OLS Method we can see that the model is performing slightly better in the test model.
- So, it is better to go with the SKlearn model for interpretation.

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

The following are the observations for the model

	Actual Values	Fitted Values	Residuals
0	-0.751485	-0.697459	-0.054026
1	-0.606816	-0.560103	-0.046712
2	0.678812	0.864440	-0.185627
3	-0.728405	-0.711163	-0.017242
4	-0.592357	-0.456111	-0.136246

Table 4 – Actual, Fitted & Residual

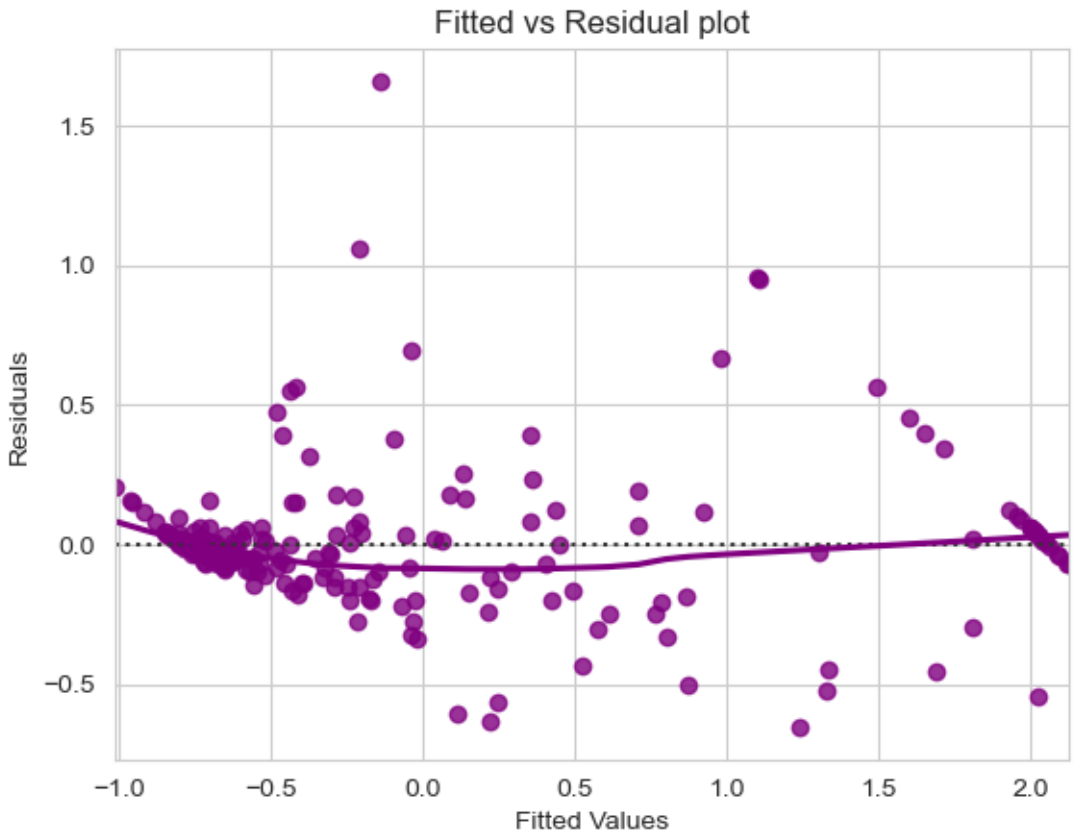

Figure 17 – Fitted vs Residual

- The key variables to considered by the firm are Capital, Employment and value.
- These three factors are influencing the business.
- Hence, the firm needs to have a closure look into these three items.
- Also, it is advisable to have a look at the research and development stocks of the firms as it is closely related with almost all the factors.

Problem Statement 2: Logistic Regression and Linear Discriminant Analysis

You are hired by the Government to do an analysis of car crashes. You are provided details of car crashes, among which some people survived and some didn't. You have to help the government in predicting whether a person will survive or not on the basis of the information given in the data set so as to provide insights that will help the government to make stronger laws for car manufacturers to ensure safety measures. Also, find out the important factors on the basis of which you made your predictions.

Data Dictionary

Column Name	Description	Data Type
dvcat	factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+	Object
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)	Float
Survived	factor with levels Survived or not_survived	Object
airbag	a factor with levels none or airbag	Object
seatbelt	a factor with levels none or belted	Object
frontal	a numeric vector; 0 = non-frontal, 1=frontal impact	Integer
sex	a factor with levels f: Female or m: Male	Object
ageOFocc	age of occupant in years	Integer
yearacc	year of accident	Integer
yearVeh	Year of model of vehicle; a numeric vector	Float
abcat	Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy, nodeploy and unavail	Object
occRole	a factor with levels driver or pass	Object
deploy	a numeric vector	Integer
injSeverity	a numeric vector; 0: none, 1: possible injury, 2: no incapacity, 3: incapacity, 4: killed; 5: unknown, 6: prior death	Float
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.	Object

Table 5 - Data Dictionary – Dataset 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.

EDA

The Data is imported and below are the observations:

- The Data has 11217 Rows and 16 Columns
  - **Data types:** Float – 3, Integer – 5, Object – 8
- Column injSeverity is having 77 Null values and it is replaced with Median.

	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	11
count	11217.000000	11217	11217.000000	11217	11217	11217	11217.000000	11217	11217.000000	11217.000000	11217.000000	11217	11217	
unique	NaN	5	NaN	2	2	2	NaN	2	NaN	NaN	NaN	3	2	
top	NaN	10-24	NaN	survived	airbag	belted	NaN	m	NaN	NaN	NaN	deploy	driver	
freq	NaN	5414	NaN	10037	7064	7849	NaN	6048	NaN	NaN	NaN	4365	8786	
mean	5608.000000	NaN	431.405309	NaN	NaN	NaN	0.644022	NaN	37.427654	2001.103236	1994.177944	NaN	NaN	
std	3238.213319	NaN	1406.202941	NaN	NaN	NaN	0.478830	NaN	18.192429	1.056805	5.658704	NaN	NaN	
min	0.000000	NaN	0.000000	NaN	NaN	NaN	0.000000	NaN	16.000000	1997.000000	1953.000000	NaN	NaN	
25%	2804.000000	NaN	28.292000	NaN	NaN	NaN	0.000000	NaN	22.000000	2001.000000	1991.000000	NaN	NaN	
50%	5608.000000	NaN	82.195000	NaN	NaN	NaN	1.000000	NaN	33.000000	2001.000000	1995.000000	NaN	NaN	
75%	8412.000000	NaN	324.056000	NaN	NaN	NaN	1.000000	NaN	48.000000	2002.000000	1999.000000	NaN	NaN	
max	11216.000000	NaN	31694.040000	NaN	NaN	NaN	1.000000	NaN	97.000000	2002.000000	2003.000000	NaN	NaN	

Table 6 – 5 Point Summary

- Using the above 5 point Summary we can see that most of the persons survived in the accidents.

Univariate Analysis

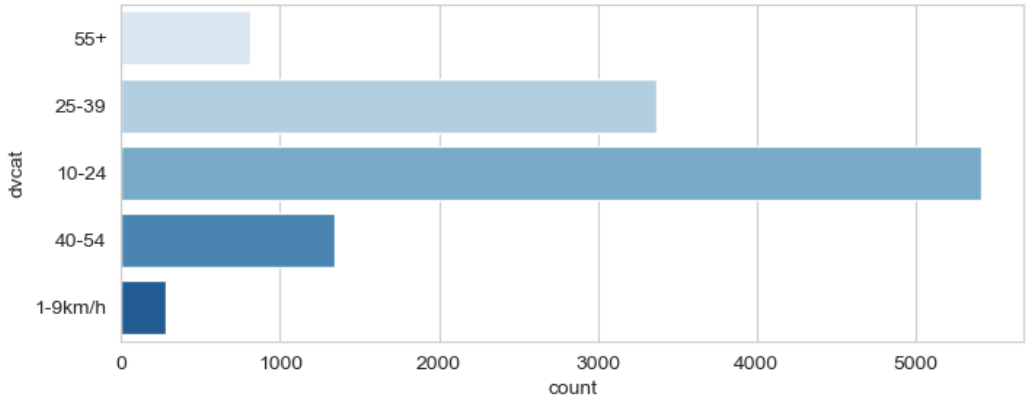

Figure 18 – dvcat

- From the above charts we can see that most of the crashes were happened while the speed was between 10 – 24 followed by 25-39 and 40-54.

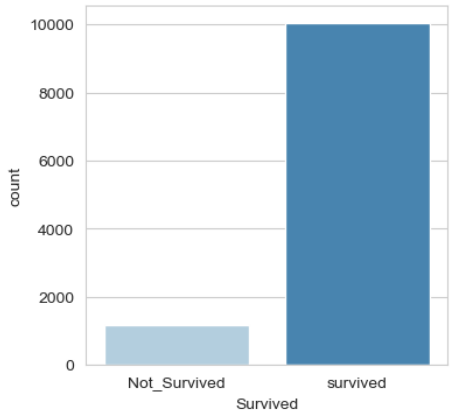

Figure 19 – Survived Split

- There is a huge difference between Survived (10037) and Not survived (1180).

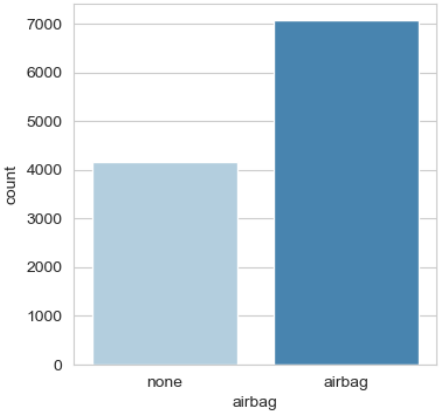

Figure 20 – Airbag Split

- In the 11217 crashes 7064 cars were equipped with airbag.

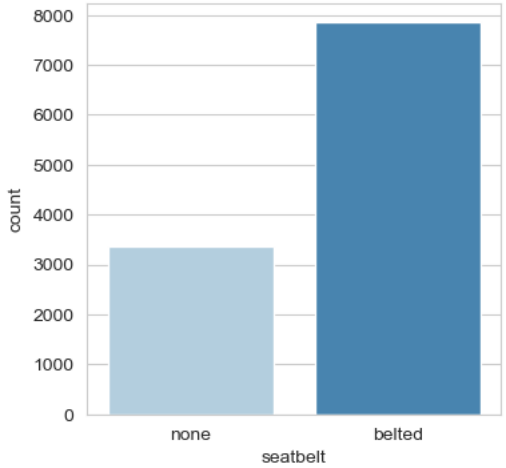

Figure 21 – Seatbelt Split

- In the 11217 crashes 7849 cars were equipped with seatbelt.

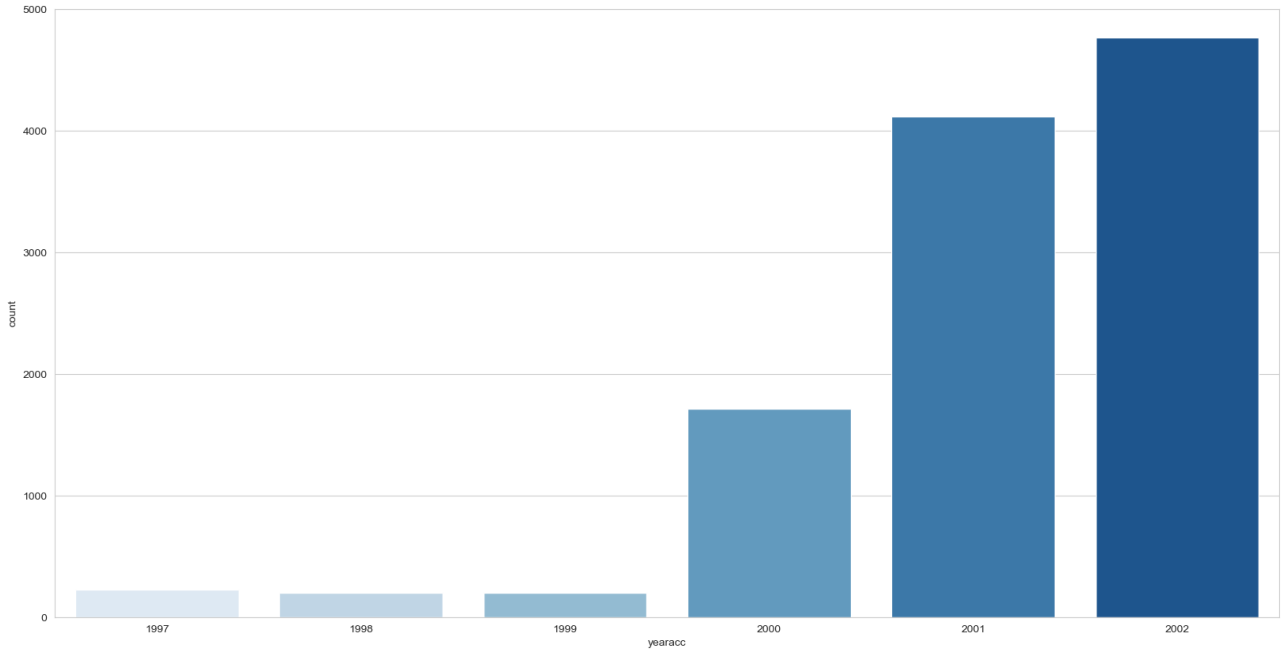

Figure 22 – Car Crash Trend

- From the above chart we can see the increase in the crash trend over the years.

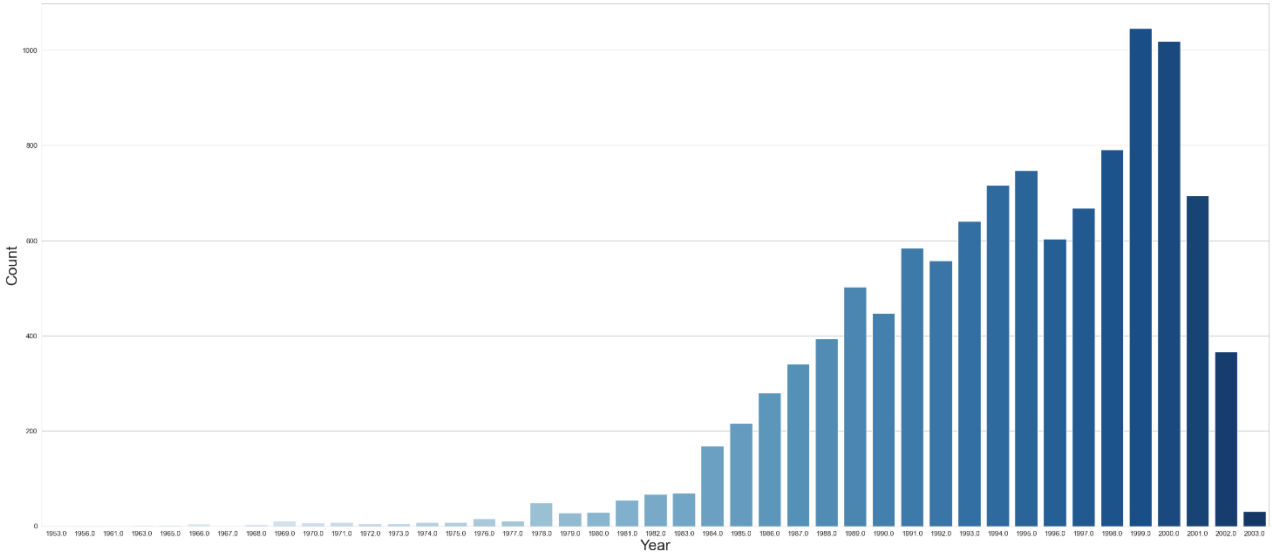

Figure 23 – Car Created Year Trend

- As per the above insights we can clearly see that cars created from later 1980s to 2000 were the ones crashed more.

Bivariate Analysis

Figure 24 - Pairplot

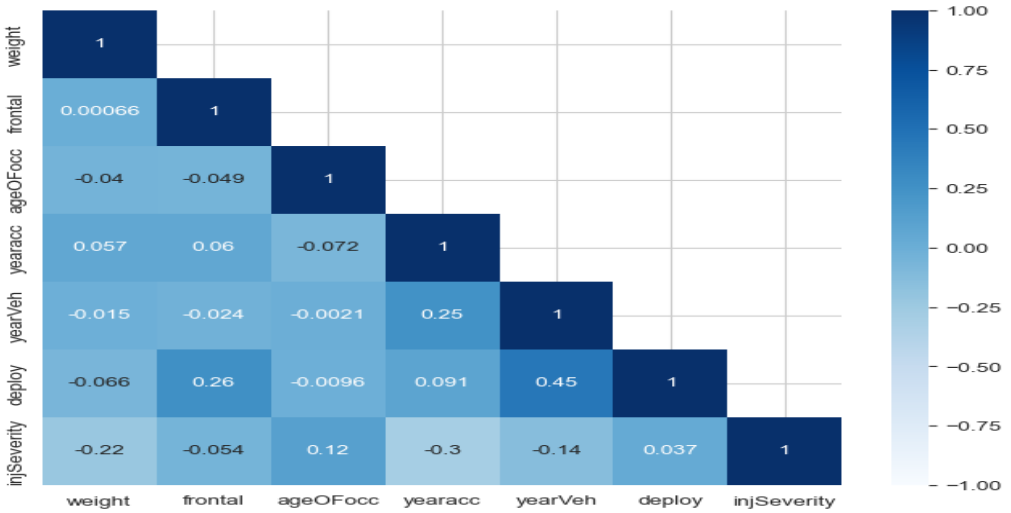

Figure 25 - Heatmap

- From the above heatmap and pairplot we can see that most of the variables are not closely corelated.
- Some of the variables available in the pairplot, do not have the classes well separated. They will not be considered as good predictors.

Outliers

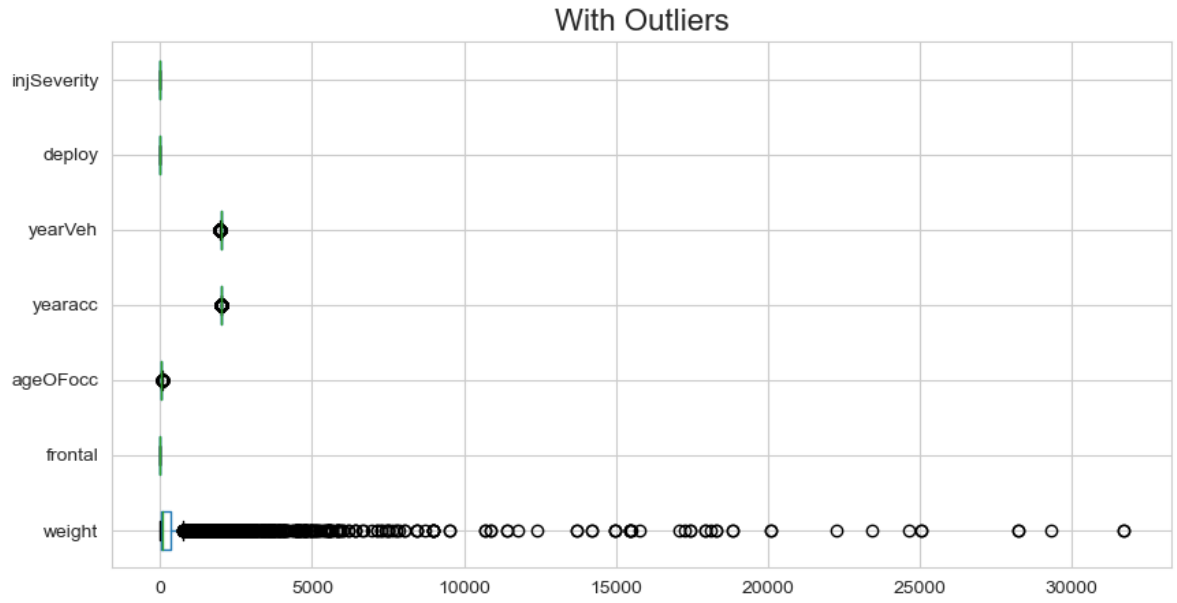

Figure 26 – Outlier Before Treatment

- The variable weight contains more outliers. Hence it is treated as seen below.

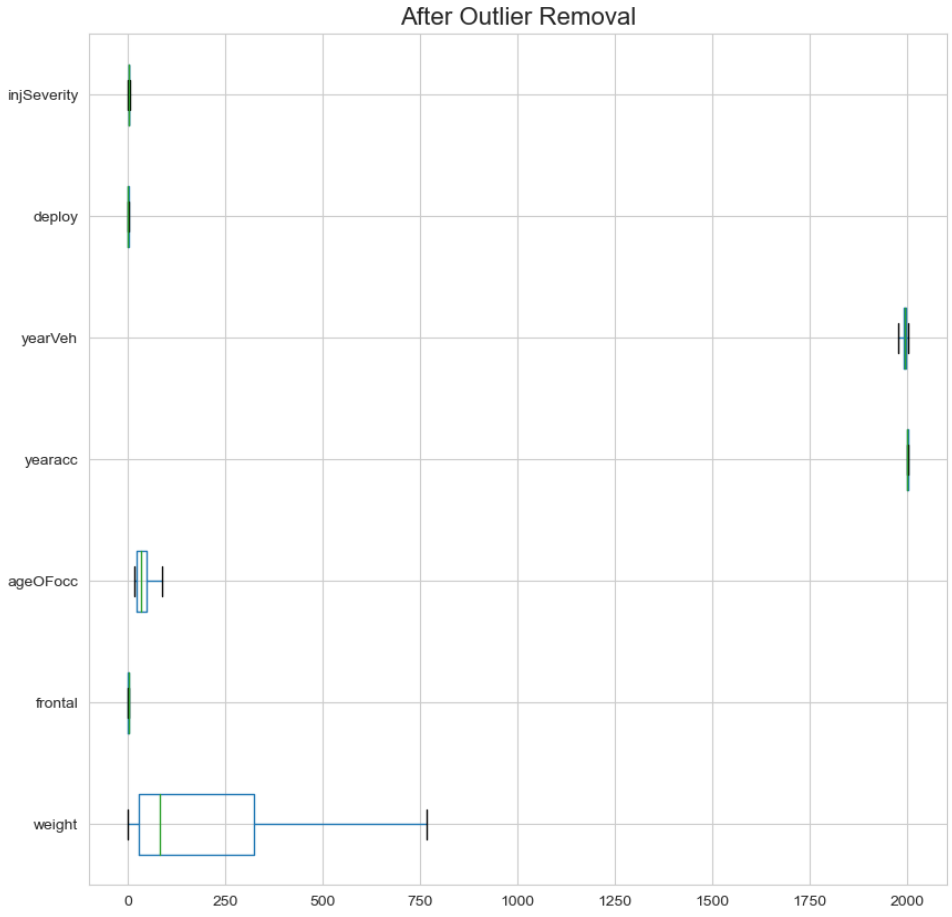

Figure 27 – Outlier After Treatment

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).

Encoding

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	abcat	occRole	deploy	injSeverity
0	1	27.078	0	0	0	1.0	0	32.0	0	0	0.0	4.0
1	3	89.627	0	1	1	0.0	1	54.0	1	0	0.0	4.0
2	1	27.078	0	0	1	1.0	0	67.0	0	0	0.0	4.0
3	1	27.078	0	0	1	1.0	1	64.0	0	1	0.0	4.0
4	1	13.374	0	0	0	1.0	0	23.0	0	0	0.0	4.0

Table 7 - Encoding

- The object datatype columns were encoded with 0s and 1s as seen in the above sample so that the machine learning models can understand the data.
- The target variable “Survived” is replaced by 1 and “Not Survived” by 0.

Train Test Split

- The data has been split into 70:30 Ratio.
- As the dependent variable contains 0s and 1s, to reduce biases stratify method is used.
- Both Logistic regression and LDA is applied.

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

Model Building

Logistic Regression

The model for Logistic Regression is built with solver “Newton – cg”

Performance Metrics

- Classification Report & Confusion Matrix – Train Data

	precision	recall	f1-score	support
0	0.92	0.89	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

Figure 28 – Performance Metrics for Train

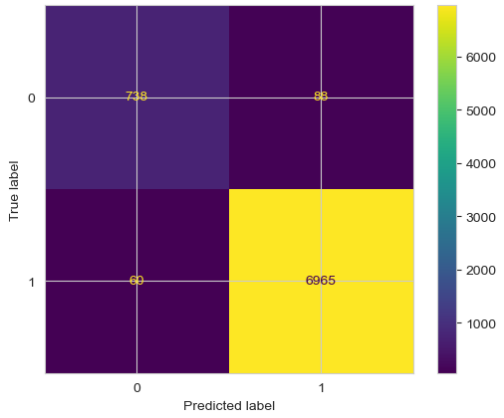

Figure 29 – Confusion Matrix for Train

- Classification Report & Confusion Matrix – 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

Figure 30 – Performance Metrix for Test

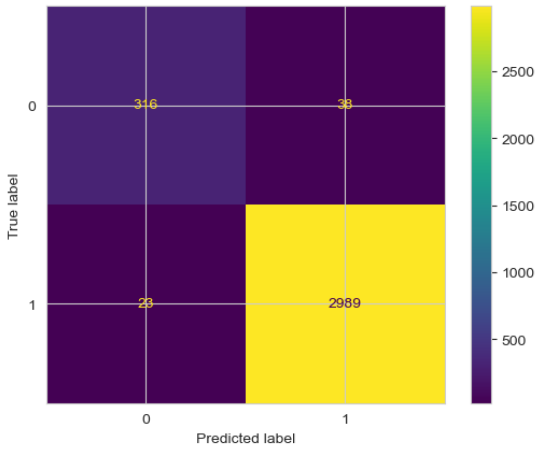

Figure 31 – Confusion Matrix for Test

- AUC and ROC Curve –Train Data
  - **AUC – 0.987**

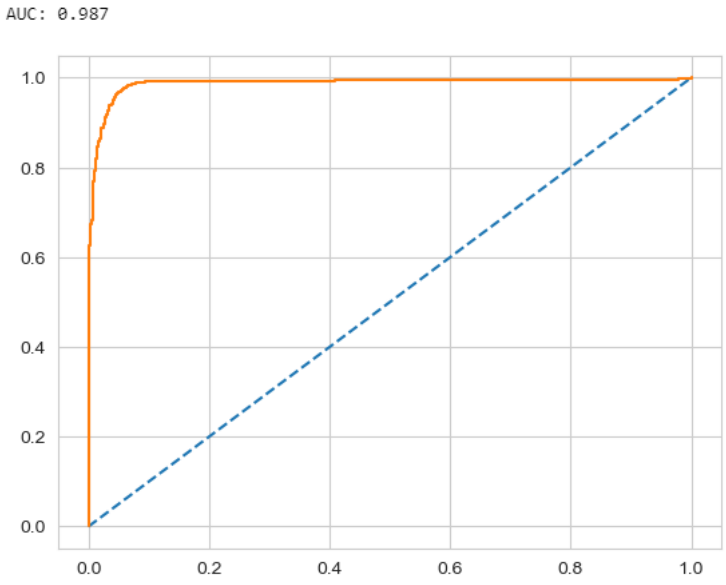

Figure 32 - AUC and ROC Curve –Train Data

- AUC and ROC Curve –Test Data
    - **AUC – 0.987**

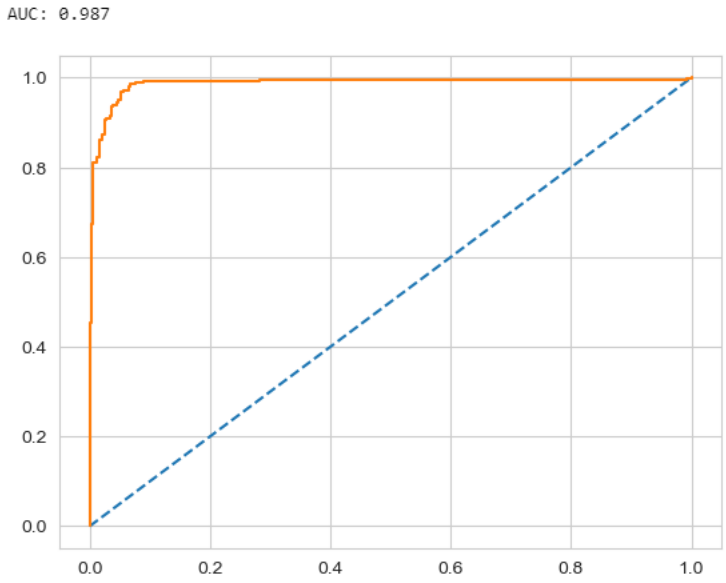

Figure 33 - AUC and ROC Curve –Test Data

Inference:

- The Accuracy and Recall is same for both the test and train data.
- Also, The AUC value matched in both the models.
- This ensures that the model performing well and there is no need of optimization of the model.

Feature Importance:

- Seatbelt and Airbag are the most important variables as they are corelated and impact is high.
- Variables Deploy and Injseverity are also very important.

```
The coefficient for dvcat is 0.6406401640420135
The coefficient for weight is 0.007175080245011356
The coefficient for airbag is 0.8535377824321096
The coefficient for seatbelt is 0.856878902671269
The coefficient for frontal is 1.182136676447145
The coefficient for sex is 0.372492412189811
The coefficient for ageOFocc is -0.03500160338597787
The coefficient for abcat is 0.1851842362839773
The coefficient for occRole is -0.46928689854751626
The coefficient for deploy is -0.6683535461481442
The coefficient for injSeverity is -4.4620811012429
```

Figure 34 – Feature Importance

LDA

- The LDA Model is built with default parameters and evaluated the accuracy score along with the confusion Matrix.
- The ROC-AUC Curve is plotted for both the models.

Performance Metrics

- The Accuracy score is 96 and 97 in the Train and Test data respectively.
- The Recall score is also very similar to each other as seen below.

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.87	0.69	0.76	826
1	0.96	0.99	0.98	7025
accuracy			0.96	7851
macro avg	0.91	0.84	0.87	7851
weighted avg	0.95	0.96	0.95	7851

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.86	0.67	0.75	354
1	0.96	0.99	0.97	3012
accuracy			0.95	3366
macro avg	0.91	0.83	0.86	3366
weighted avg	0.95	0.95	0.95	3366

Figure 35 – Performance Metrics for Train

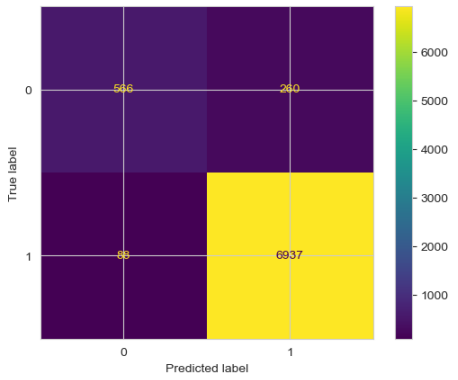

Figure 36 – Confusion Matrix for Train

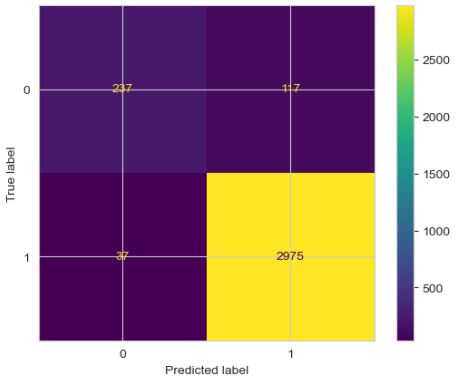

Figure 37 – Confusion Matrix for Test

- AUC and ROC Curve
    - The AUC is also very similar for both the models.

```
AUC for the Training Data: 0.979
AUC for the Test Data: 0.978
```

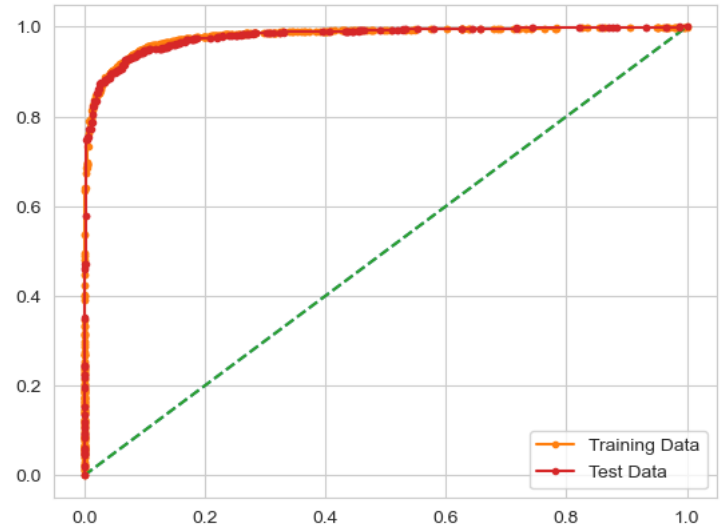

Figure 38 – AUC & ROC Curve

Inference:

- The model has performed very well in both the models as all the scores very similar.
- This implies that the model is very good.

Feature Importance:

- frontal and dvcat are the most important variables as they are corelated and impact is high.
- Variable Injseverity is also very important.

```
The coefficient for dvcat is 0.7981533817183232
The coefficient for weight is -0.0007416459885542775
The coefficient for airbag is 0.47368298157750294
The coefficient for seatbelt is 0.6063892133429107
The coefficient for frontal is 0.8183679580751099
The coefficient for sex is 0.4745402369664335
The coefficient for ageOFocc is -0.02603188304662198
The coefficient for abcat is 0.1205348233537516
The coefficient for occRole is -0.28723479255923134
The coefficient for deploy is -0.07334048867811763
The coefficient for injSeverity is -1.5433558560613099
```

Figure 39 - Coefficient

Model Comparison

	Accuracy		Recall		Precision		AUC	
Model Name	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.98	0.98	0.99	0.99	0.99	0.99	0.987	0.987
LDA Model	0.96	0.95	0.99	0.99	0.96	0.96	0.979	0.978

Table 8 – Model Comparison

- While comparing both the Logistic Regression and LDA model we can conclude that Logistic regression performs better as it gives same results in both train and test.
- So, we can conclude that Logistic Regression performed well using the given Data and well optimized.

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

Business Insights and Recommendations

Important Features:

- As per Logistic Regression Seatbelt, Deploy and Airbag are the most important variables in deciding the survival.
- The LDA model suggests Frontal and dvcat as the most important variable. But it also gives a very good results in Airbag and Seatbelt. So, Seatbelt and Airbag is very important.
- In both the models injseverity is a very important factor. So, it is also very important.
- These features can influence in the future models.

Recommendations:

- Wearing Seatbelt and having the Airbags in the cars can save many lives.
- The government can run an awareness campaign to enforce the people to use seatbelts.

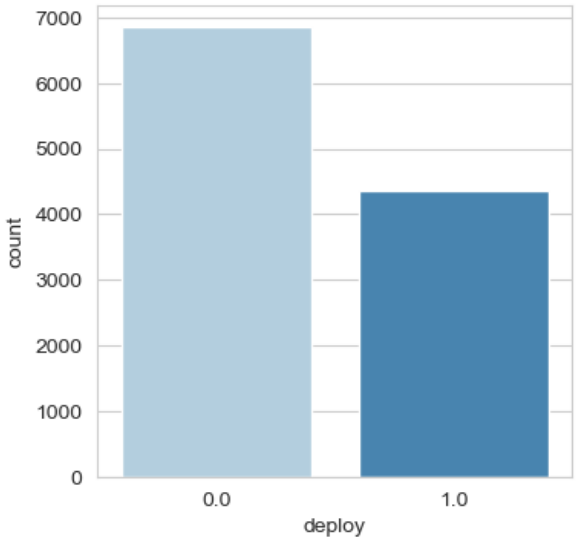

Figure 40 – Deploy Split

- Around 7000 times the Airbag has not been deployed in the car.
- So, it is very essential to deep dive into this issue and mandate the car makers to improve the airbag feature.