Great Learning

PREDICTIVE MODELING

PROJECT REPORT

By

HARI HARAN

22st June, 2023

Great Learning

PROBLEM 1 : Linear Regression

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

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

Data Dictionary for Firm_level_data:

1. sales: Sales (in millions of dollars).
2. capital: Net stock of property, plant, and equipment.
3. patents: Granted patents.
4. randd: R&D stock (in millions of dollars).
5. employment: Employment (in 1000s).
6. sp500: Membership of firms in the S&P 500 index. S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States
7. tobinq: Tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value.
8. value: Stock market value.
9. institutions: Proportion of stock owned by institutions.

Performing EDA,

For better readability we have rearranged and renamed the columns,

	Instituions	R&D_stocks	S&P500_index	Patents	Employment	Stock_value	Capital_value	Sales
0	80.27	382.078	no	10	2.306	1625.454	161.604	826.995
1	59.02	0.000	no	2	1.860	243.117	122.101	407.754

S&P500_index has been encoded as, No = 0 and Yes = 1, for further analysis.

Sales is our target variable for further analysis for the Linear Regression Model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 759 entries, 0 to 758
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Instituions    759 non-null    float64
 1   R&D_stocks     759 non-null    float64
 2   S&P500_index   759 non-null    int64
 3   Patents        759 non-null    int64
 4   Employment     759 non-null    float64
 5   Stock_value    759 non-null    float64
 6   Capital_value  759 non-null    float64
 7   Sales          759 non-null    float64
dtypes: float64(6), int64(2)
memory usage: 47.6 KB
```

```
Instituions      0
R&D_stocks       0
S&P500_index     0
Patents          0
Employment       0
Stock_value      0
Capital_value    0
Sales            0
dtype: int64
```

Table 1.1 - Data Info and Checking for null values.

There are **no Null values and duplicate values** in the data.

	count	mean	std	min	25%	50%	75%	max
Instituions	759.0	43.021	21.686	0.000	25.395	44.110	60.510	90.150
R&D_stocks	759.0	439.938	2007.398	0.000	4.628	36.864	143.253	30425.256
S&P500_index	759.0	0.286	0.452	0.000	0.000	0.000	1.000	1.000
Patents	759.0	25.831	97.260	0.000	1.000	3.000	11.500	1220.000
Employment	759.0	14.165	43.321	0.006	0.928	2.924	10.050	710.800
Stock_value	759.0	2732.735	7071.072	1.971	103.594	410.794	2054.160	95191.591
Capital_value	759.0	1977.747	6466.705	0.057	52.651	202.179	1075.790	93625.201
Sales	759.0	2689.705	8722.060	0.138	122.920	448.577	1822.547	135696.788

Table 1.2 - Data Description

```
Instituions      -0.168071
R&D_stocks       10.270483
S&P500_index      0.949540
Patents           7.766943
Employment        9.068875
Stock_value       6.075996
Capital_value     7.555091
Sales             9.219023
dtype: float64
```

```
Instituions      -0.168071
R&D_stocks        1.162978
S&P500_index      0.949540
Patents           1.162219
Employment        1.186553
Stock_value       1.195849
Capital_value     1.190265
Sales             1.189942
dtype: float64
```

Table 1.3 - Skewness Before (left) and After (right) Outlier treatment in the data

<u>Observations</u> -

- There are no null values in the data. Here we have dropped "Unnamed: 0 " column.
- There are 6 float types and 2 integer type variables.
- The data has **759** rows and **8** columns.
- The skewness in the data in not symmetrical and R&D stocks has the highest value.
- The R&D stock , patents, employment, stock_value and capital_value are important factors.
- The skewness in the data is almost symmetrical after outlier treatment.
- R&D has good investment considering all intuitions.

Univariate Analysis,

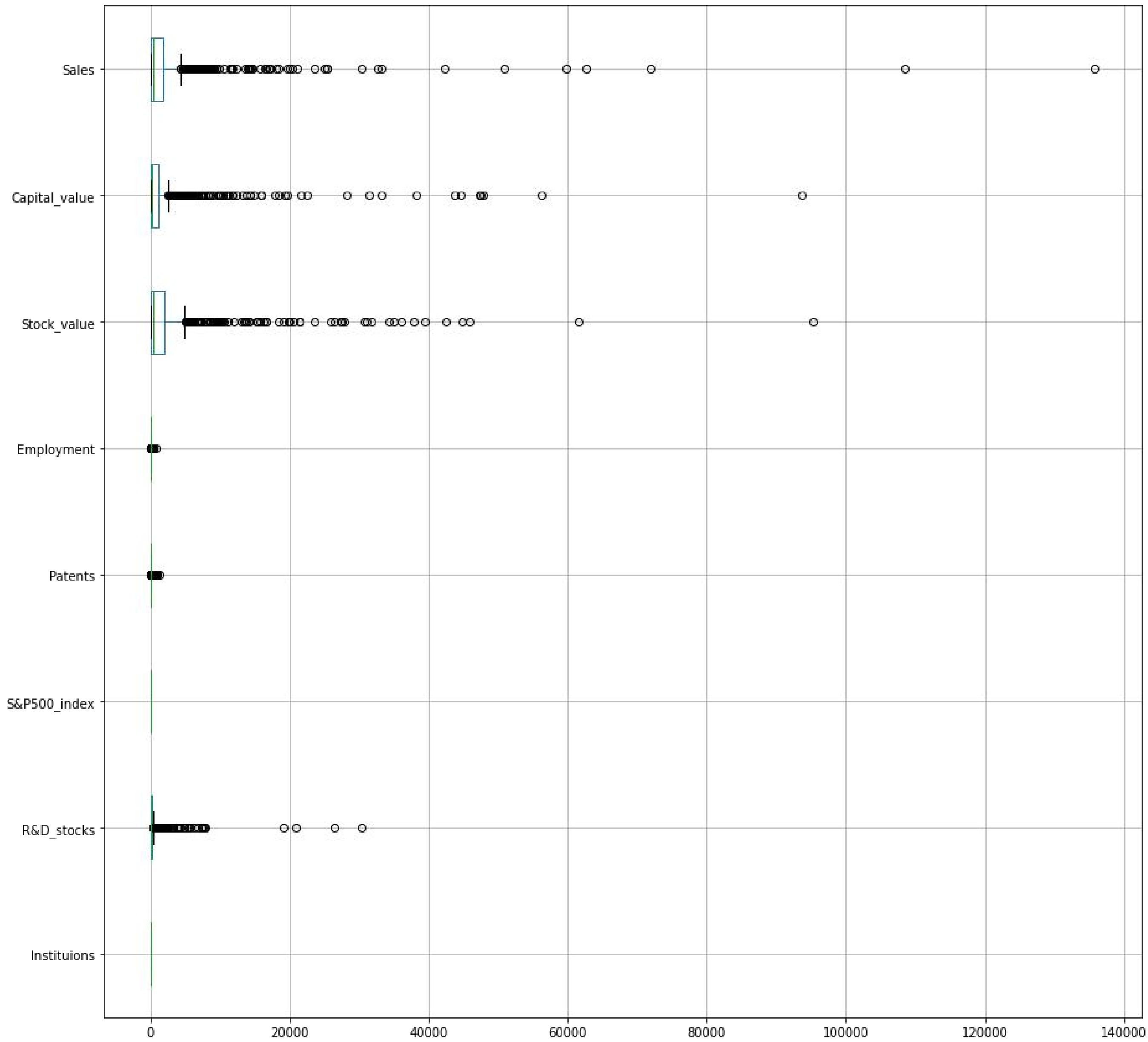

Fig1.1 - Box Plot to check Outliers

We are going to treat outliers for further analysis to get better model and increase accuracy. The treatment depends on the analyst choice whether to analyse with the given data or cap the values for better analysis. Although, its recommended to treat outliers before modelling for data with large amounts of variables results because with more variables the skewness may increase and the results may become biased.

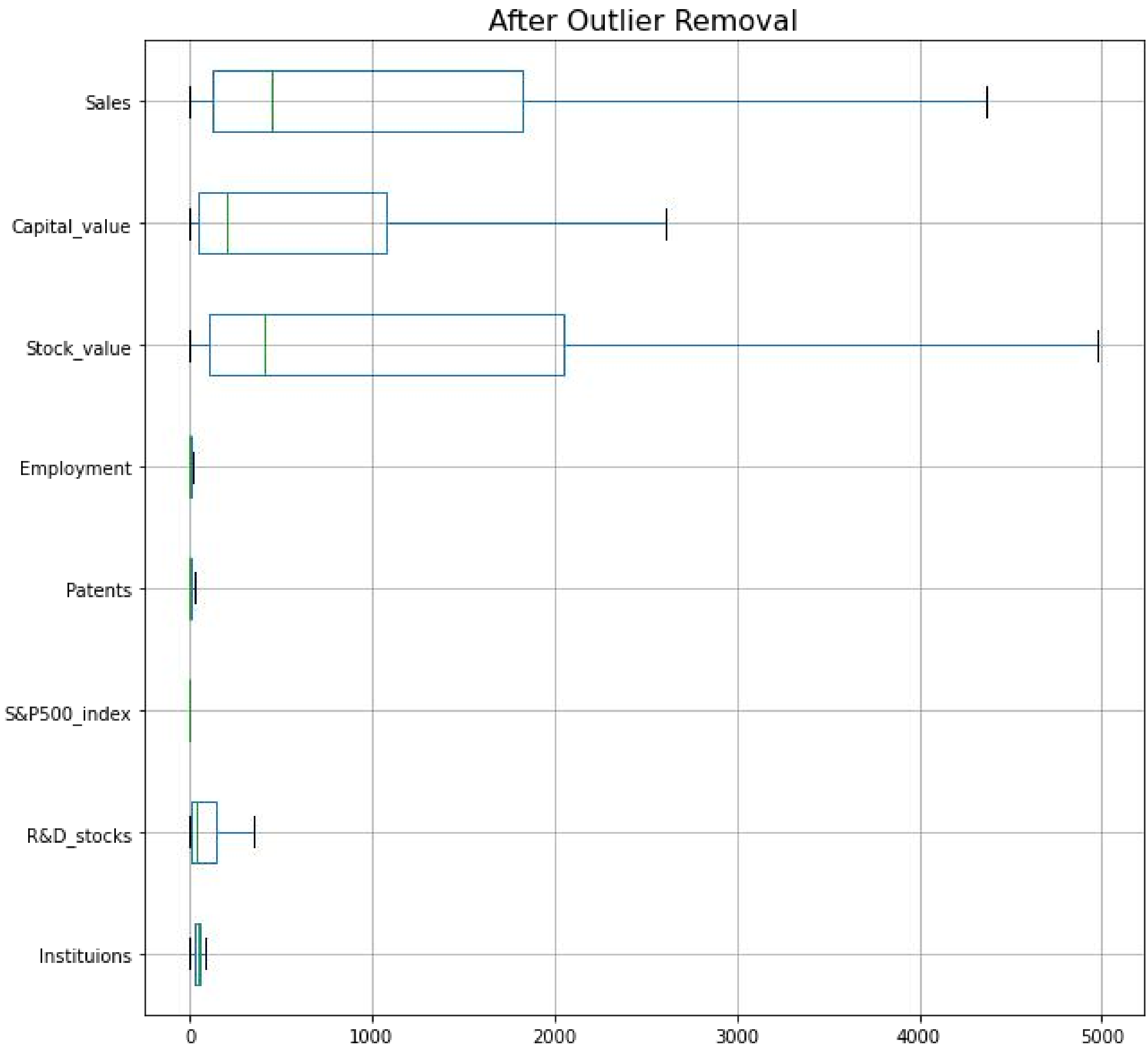

Fig1.2 - Box Plot : After Outlier Treatment

All the outliers have been treated and the data looks good for further analysis.

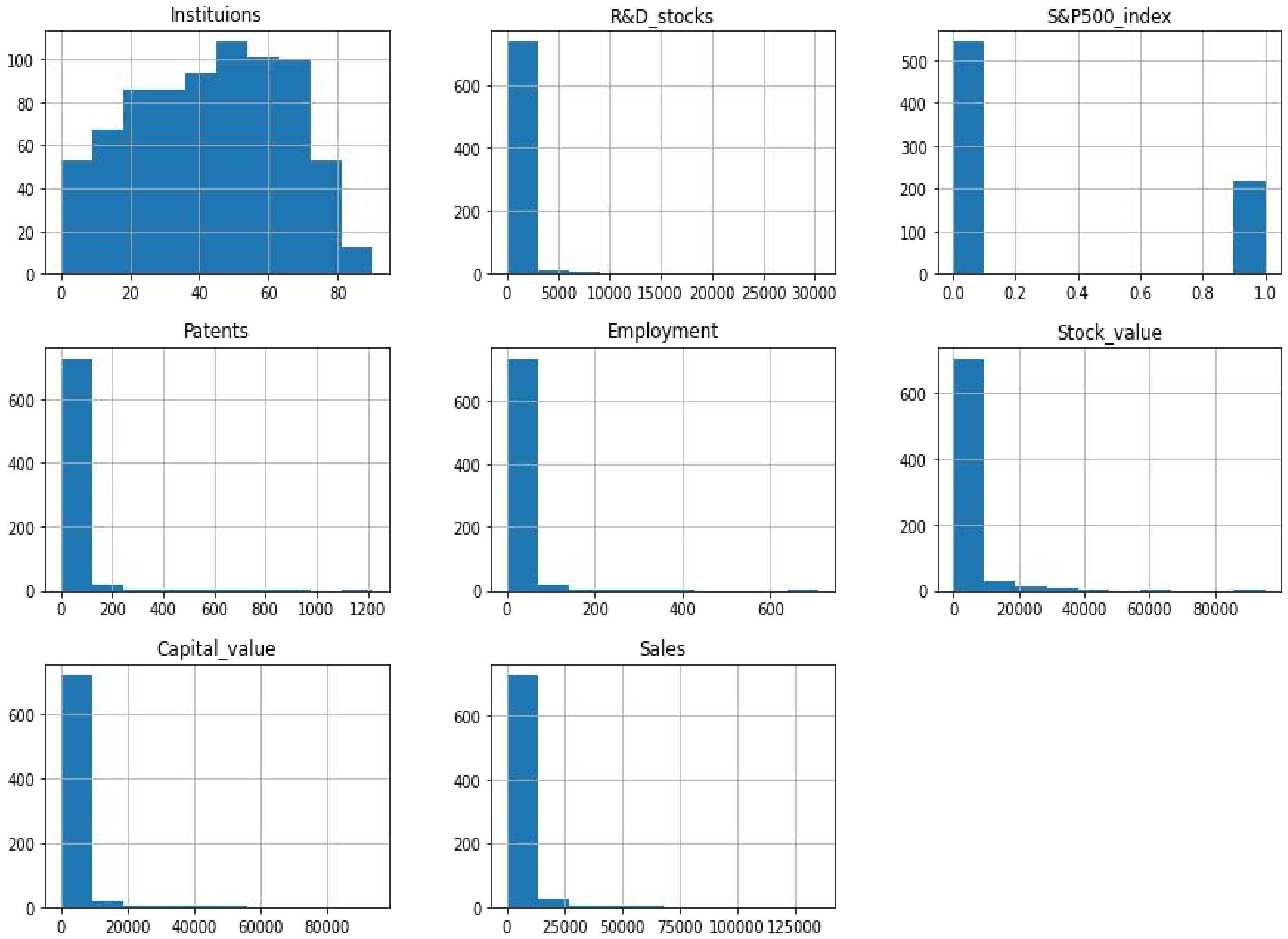

Fig 1.3 - Histogram Plot.

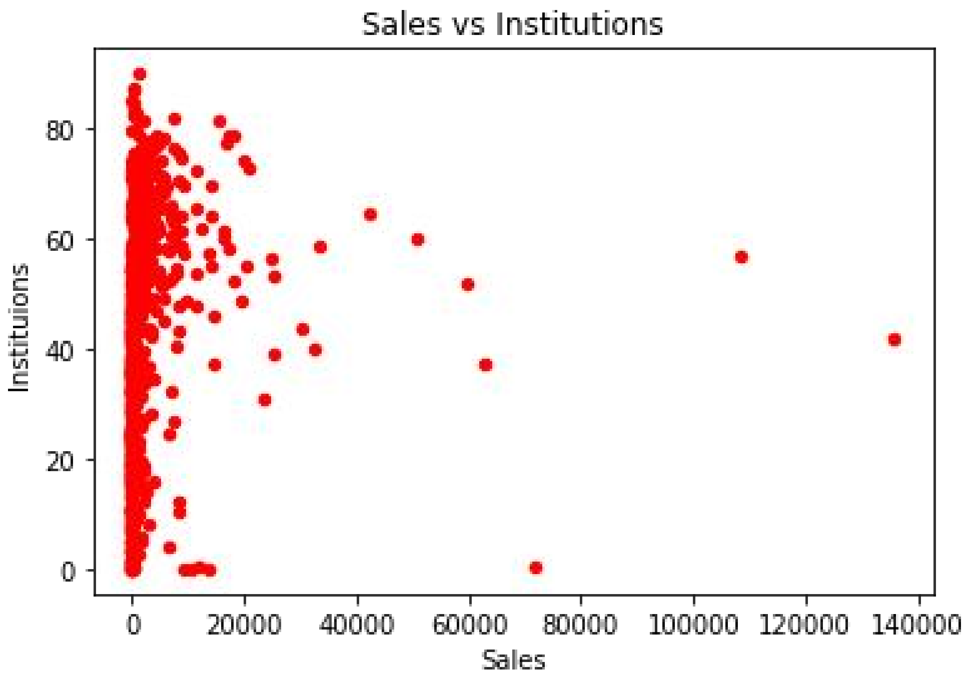

Fig 1.4 - Scatter Plot: Sales Vs Institutions.

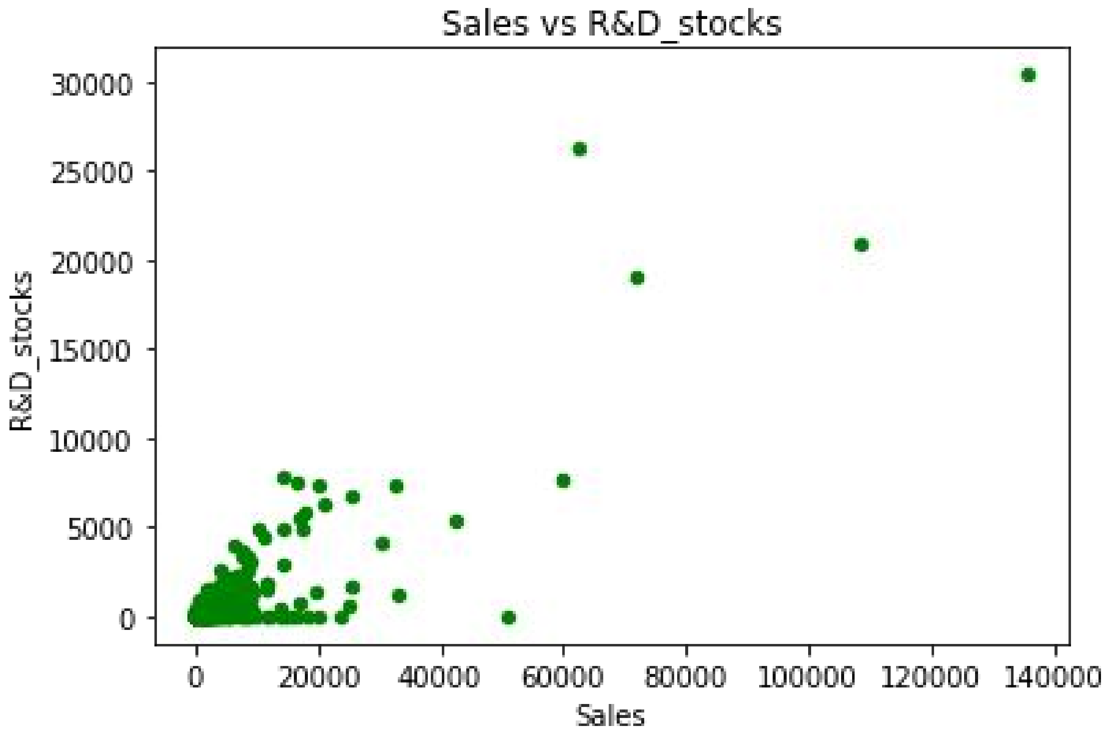

Fig 1.5 - Scatter Plot: Sales Vs R&D Stocks.

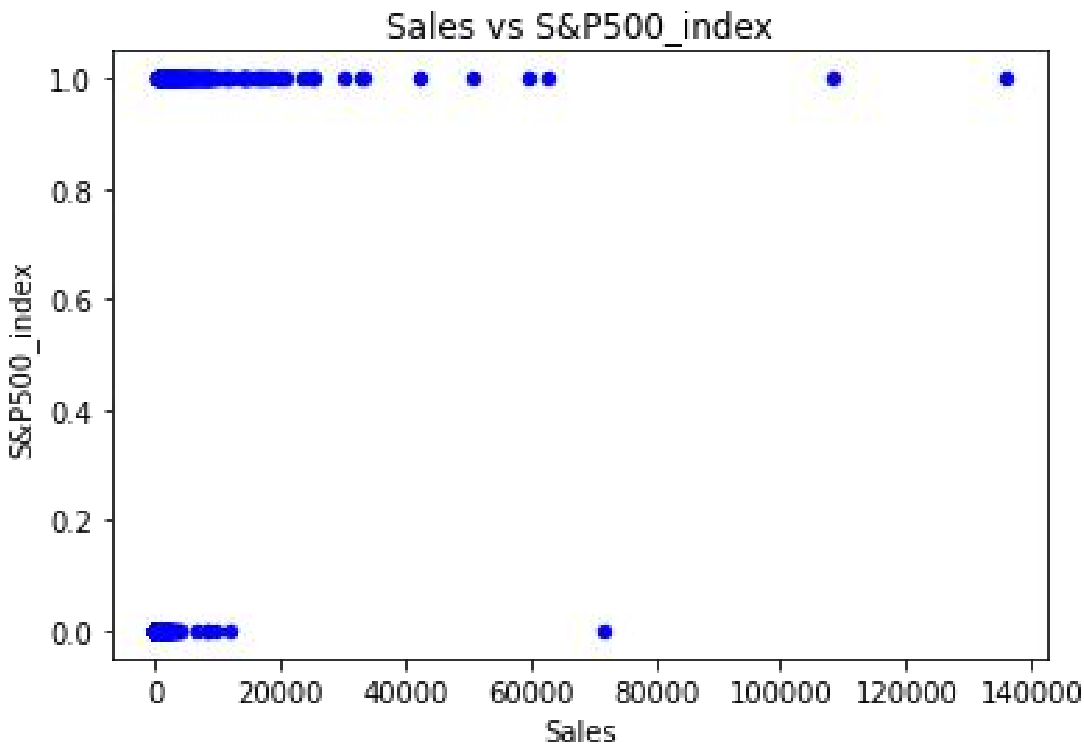

Fig 1.6 - Scatter Plot: Sales Vs S&P Index.

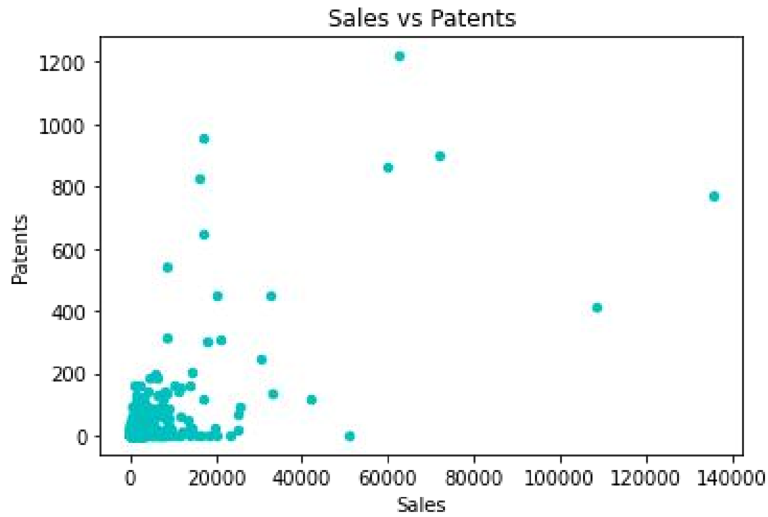

Fig 1.7 - Scatter Plot: Sales Vs Patents.

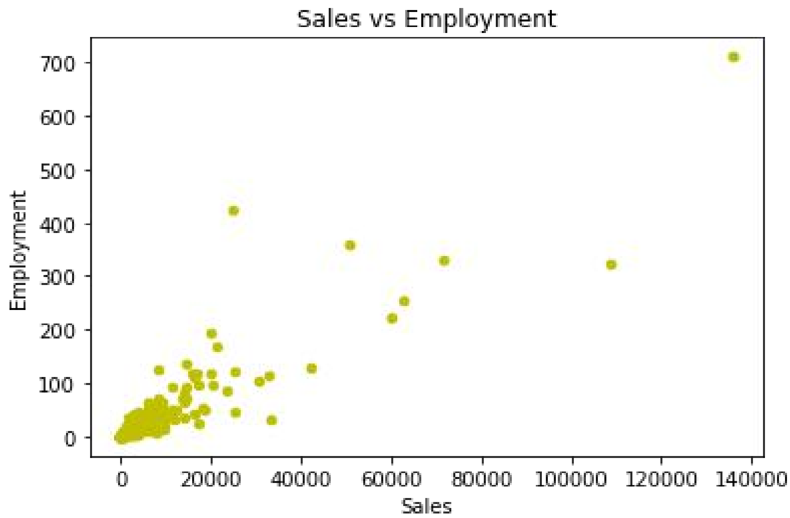

Fig 1.8 - Scatter Plot: Sales Vs Employment.

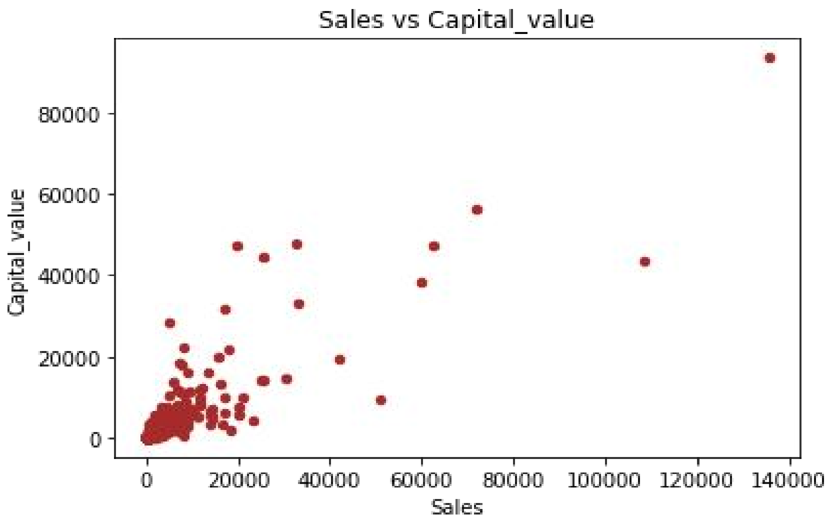

Fig 1.9 - Scatter Plot: Sales Vs Capital value.

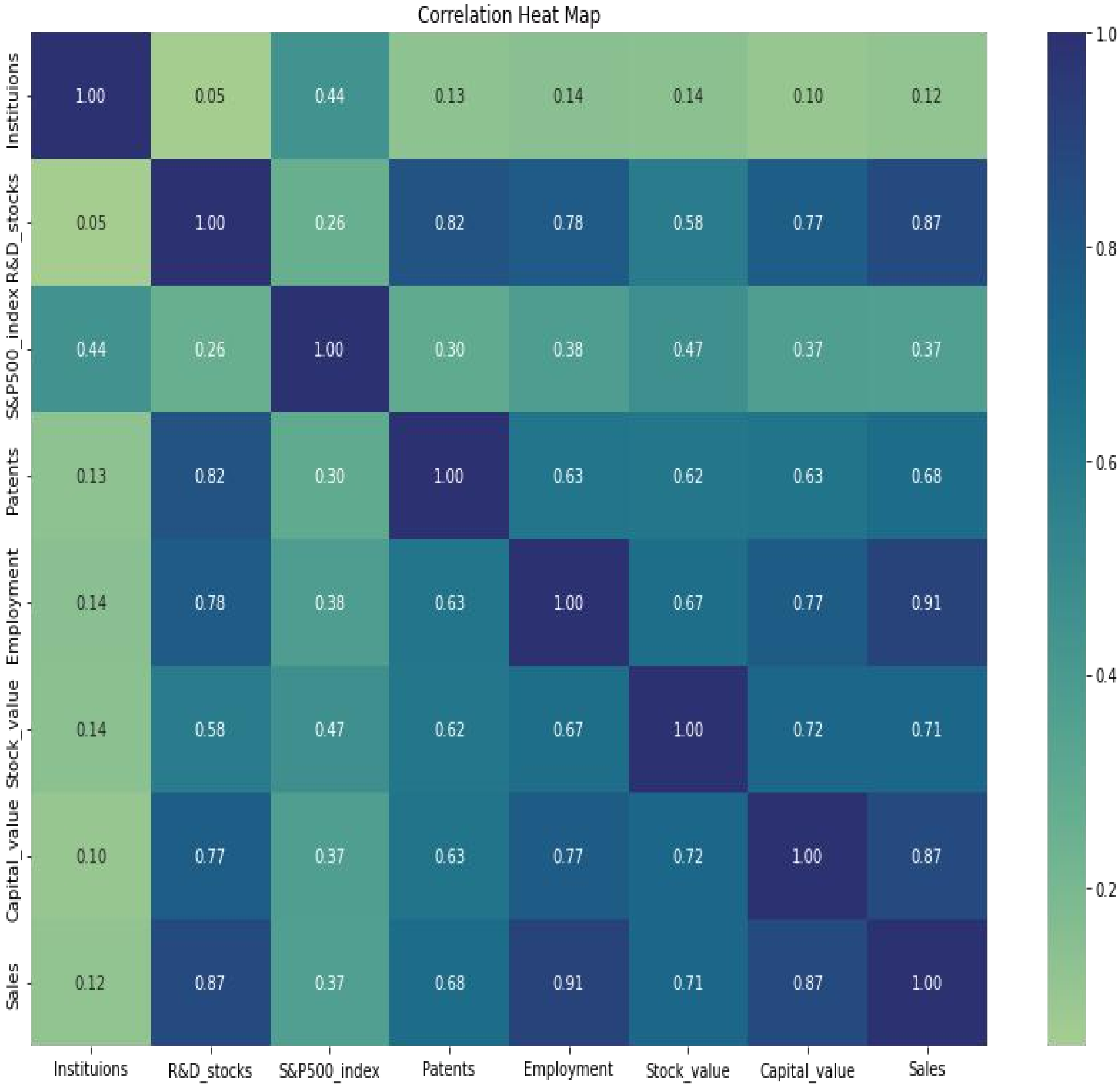

Fig 1.10 - Correlation Heat Map

From the above plot,

The R&D stock , patents, employment, stock_value and capital_value are important factors. These variables are highly correlated with sales. Employment has the highest correlation with sales.

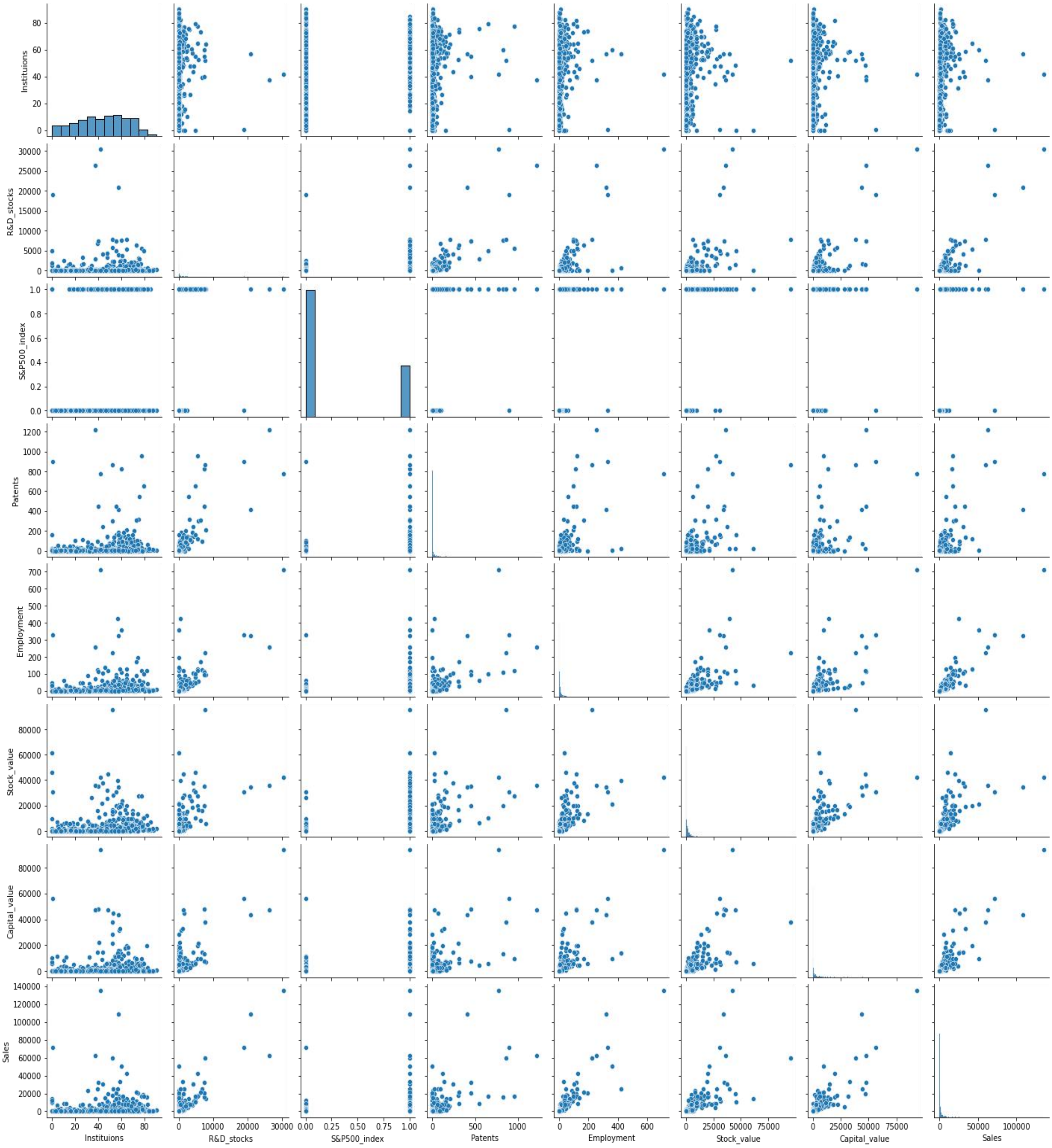

Fig 1.11 - Pair Plot

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

There are null values in the data set.

S&P500_index has been encoded as, No = 0 and Yes = 1, for further analysis.

Scaling is the process of standardization of data to transform the data in such a way that it will have a **mean 0** and standard **deviation 1**. Scaling helps us to balance the impact of different variables ,present in the data, on the distance between them and in-turn helps to improve the quality and performance of the model.

Here we are applying the **StandardScaler()** method.

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.

Applying Linear Regression Model,

Splitting the data into Train and Test data, - 70:30 split. Considering, the “Sales” as the target variable.

X _train shape : (531, 7) ; y_train shape : (531, 1); X_Test Shape : (228, 7); y_test shape : (228, 1)

X shape - (759, 7)
Y shape - (759, 1)

The coefficients for each of the independent attributes :
- The coefficient for **Institutions** is 0.11609135363276048
- The coefficient for **R&D_stocks** is 0.6120568947187326
- The coefficient for **S&P500_index** is 172.12798009965283
- The coefficient for **Patents** is -5.808956240929047
- The coefficient for **Employment** is 82.43808341024011
- The coefficient for **Stock_value** is 0.20783028524978664
- The coefficient for **Capital_value** is 0.45978979902852246

The coefficient of determination R^2 of the prediction on Train set 0.935.
The coefficient of determination R^2 of the prediction on Test set 0.922.

Applying Linear Regression Model Linear Regression using Statsmodel (OLS),

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.935
Model:	OLS	**Adj. R-squared:**	0.934
Method:	Least Squares	**F-statistic:**	1067.
Date:	Sat, 17 Jun 2023	**Prob (F-statistic):**	6.47e-305
Time:	16:02:08	**Log-Likelihood:**	-3933.2
No. Observations:	531	**AIC:**	7882.
Df Residuals:	523	**BIC:**	7917.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-20.9972	40.250	-0.522	0.602	-100.069	58.075
Instituions	0.1161	0.910	0.128	0.899	-1.672	1.904
R&D_stocks	0.6121	0.234	2.611	0.009	0.152	1.073
S&P500_index	172.1280	67.100	2.565	0.011	40.310	303.946
Patents	-5.8090	2.793	-2.080	0.038	-11.295	-0.323
Employment	82.4381	4.664	17.675	0.000	73.275	91.601
Stock_value	0.2078	0.023	8.985	0.000	0.162	0.253
Capital_value	0.4598	0.039	11.742	0.000	0.383	0.537

Omnibus:	189.932	**Durbin-Watson:**	1.939
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1369.342
Skew:	1.376	**Prob(JB):**	4.48e-298
Kurtosis:	10.370	**Cond. No.**	9.81e+03

Table 1.4 - OLS Regression Result

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.81e+03. This might indicate that there are strong multicollinearity or other numerical problems.

The final Linear Regression equation is

(-21.0) * const + (0.12) * Institutions + (0.61) * R&D_stocks + (172.13) * S&P500_index + (-5. 81) * Patents + (82.44) * Employment + (0.21) * Stock_value + (0.46) * Capital_value +

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

Business Insights and Recommendations:

- The five most of important attributes are The R&D stock , patents, employment, stock_value and capital_value
- Employment has the highest correlation with sales and is important of all the five above attributes mention above.
- The inquisitions should maintain the current employment numbers or increase for better sales. It shows that they increase their sales and marketing people to improve the sales.
- The business have good patents investment and can invest more in their R&D stocks, as all businesses should to improve their service and product qualities.
- The current Tobin's q ratio is not looking good in terms of their values. The recommendation is to improve and asses their existing physical assets and change them before it looses further market value.
- The Capital value and stock values are looking good for now but the stock values may decline subjecting to future market scenarios.
- Improving their R & D stocks may increase their product and in turn their stock values.
- The firm can collect data from these institutions, to research and collect data from their consumers or customers.
- The Capital value looks good and it may increase steadily, provide it maintains the current in the sales.

Great Learning

Problem 2: Logistic Regression and Linear Discriminant Analysis

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

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

Data Dictionary :

1. dvcat: factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+
2. weight: Observation weights, albeit of uncertain accuracy, designed to account for varying sampling probabilities. (The inverse probability weighting estimator can be used to demonstrate causality when the researcher cannot conduct a controlled experiment but has observed data to model)
3. Survived: factor with levels Survived or not_survived
4. airbag: a factor with levels none or airbag
5. seatbelt: a factor with levels none or belted
6. frontal: a numeric vector; 0 = non-frontal, 1=frontal impact
7. sex: a factor with levels f: Female or m: Male
8. ageOFocc: age of occupant in years
9. yearacc: year of accident
10. yearVeh: Year of model of vehicle; a numeric vector
11. abcat: Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy, nodeploy and unavail
12. occRole: a factor with levels driver or pass: passenger
13. deploy: a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed.
14. injSeverity: a numeric vector; 0: none, 1: possible injury, 2: no incapacity, 3: incapacity, 4: killed; 5: unknown, 6: prior death
15. caseid: character, created by pasting together the populations sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

Performing EDA,

	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	1987.0	unavail	driver	0	4.0	2:13:2
1	25-39	89.627	Not_Survived	airbag	belted	0	f	54	1997	1994.0	nodeploy	driver	0	4.0	2:17:1

Table 2.1 - Data head

```
<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  float64
 10  abcat        11217 non-null  object
 11  occRole      11217 non-null  object
 12  deploy       11217 non-null  int64
 13  injSeverity  11140 non-null  float64
 14  caseid       11217 non-null  object
dtypes: float64(3), int64(4), object(8)
memory usage: 1.3+ MB
```

```
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
```

- **injSeverity** has null values

Table 2.2 - Data Info and Checking for null values.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
dvcat	11217	5	10-24	5414	NaN	NaN	NaN	NaN	NaN	NaN	NaN
weight	11217.0	NaN	NaN	NaN	431.405	1406.203	0.0	28.292	82.195	324.056	31694.04
Survived	11217	2	survived	10037	NaN	NaN	NaN	NaN	NaN	NaN	NaN
airbag	11217	2	airbag	7064	NaN	NaN	NaN	NaN	NaN	NaN	NaN
seatbelt	11217	2	belted	7849	NaN	NaN	NaN	NaN	NaN	NaN	NaN
frontal	11217.0	NaN	NaN	NaN	0.644	0.479	0.0	0.0	1.0	1.0	1.0
sex	11217	2	m	6048	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ageOFocc	11217.0	NaN	NaN	NaN	37.428	18.192	16.0	22.0	33.0	48.0	97.0
yearacc	11217.0	NaN	NaN	NaN	2001.103	1.057	1997.0	2001.0	2001.0	2002.0	2002.0
yearVeh	11217.0	NaN	NaN	NaN	1994.178	5.659	1953.0	1991.0	1995.0	1999.0	2003.0
abcat	11217	3	deploy	4365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
occRole	11217	2	driver	8786	NaN	NaN	NaN	NaN	NaN	NaN	NaN
deploy	11217.0	NaN	NaN	NaN	0.389	0.488	0.0	0.0	0.0	1.0	1.0
injSeverity	11140.0	NaN	NaN	NaN	1.826	1.379	0.0	1.0	2.0	3.0	5.0
caseid	11217	6488	73:100:2	7	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2.3 - Data Description.

'Unnamed: 0' column dropped as its is not important for further analysis.
caseid will not imputed as it's the target variable.

injSeverrity has null values.

In further steps, following ***object type*** will be replaced to ***integer type*** :

- Survived - Not survived = 0 and Survived = 1
- airbag - none = 0 and airbag = 1
- seatbelt - none = 0 and belted = 1
- sex - m = 0 and f = 1
- occRole - driver = 0, pass = 1

The below table shows that we have replaced the above variables as integer type,

```
<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  int64
 3   airbag       11217 non-null  int64
 4   seatbelt     11217 non-null  int64
 5   frontal      11217 non-null  int64
 6   sex          11217 non-null  int64
 7   ageOFocc     11217 non-null  int64
 8   yearacc      11217 non-null  int64
 9   yearVeh      11217 non-null  float64
 10  abcat        11217 non-null  object
 11  occRole      11217 non-null  int64
 12  deploy       11217 non-null  int64
 13  injSeverity  11217 non-null  float64
 14  caseid       11217 non-null  object
dtypes: float64(3), int64(9), object(3)
memory usage: 1.3+ MB
```

Table 2.4 - Data Info - Changing Object type to Int Type

Univariate Analysis,

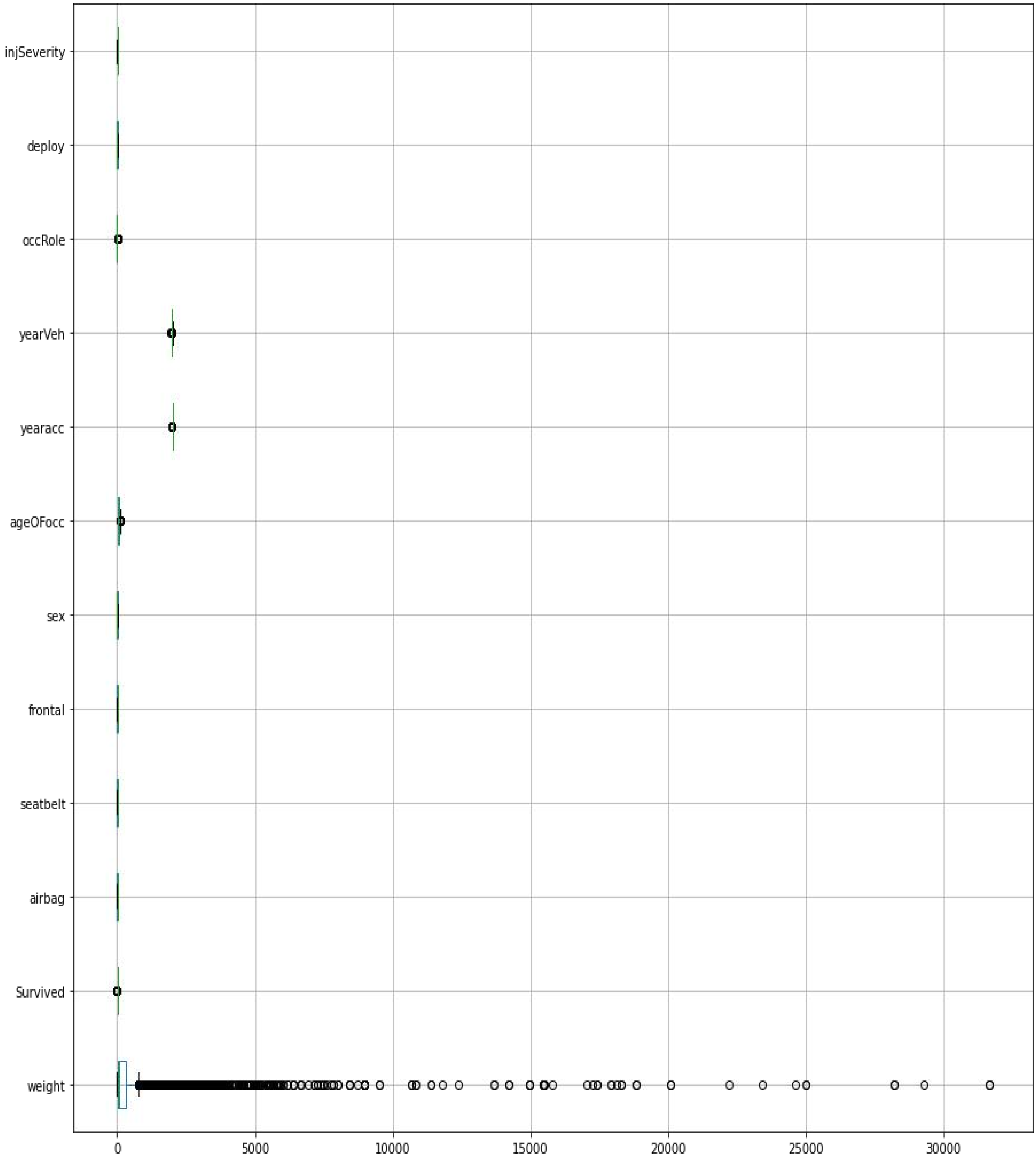

Fig.2.1 - Boxplot

The above plot shows that "weight" has lot of outliers and its need to be treatment. For correct model prediction, outlier treatment maybe done to get more accuracy .

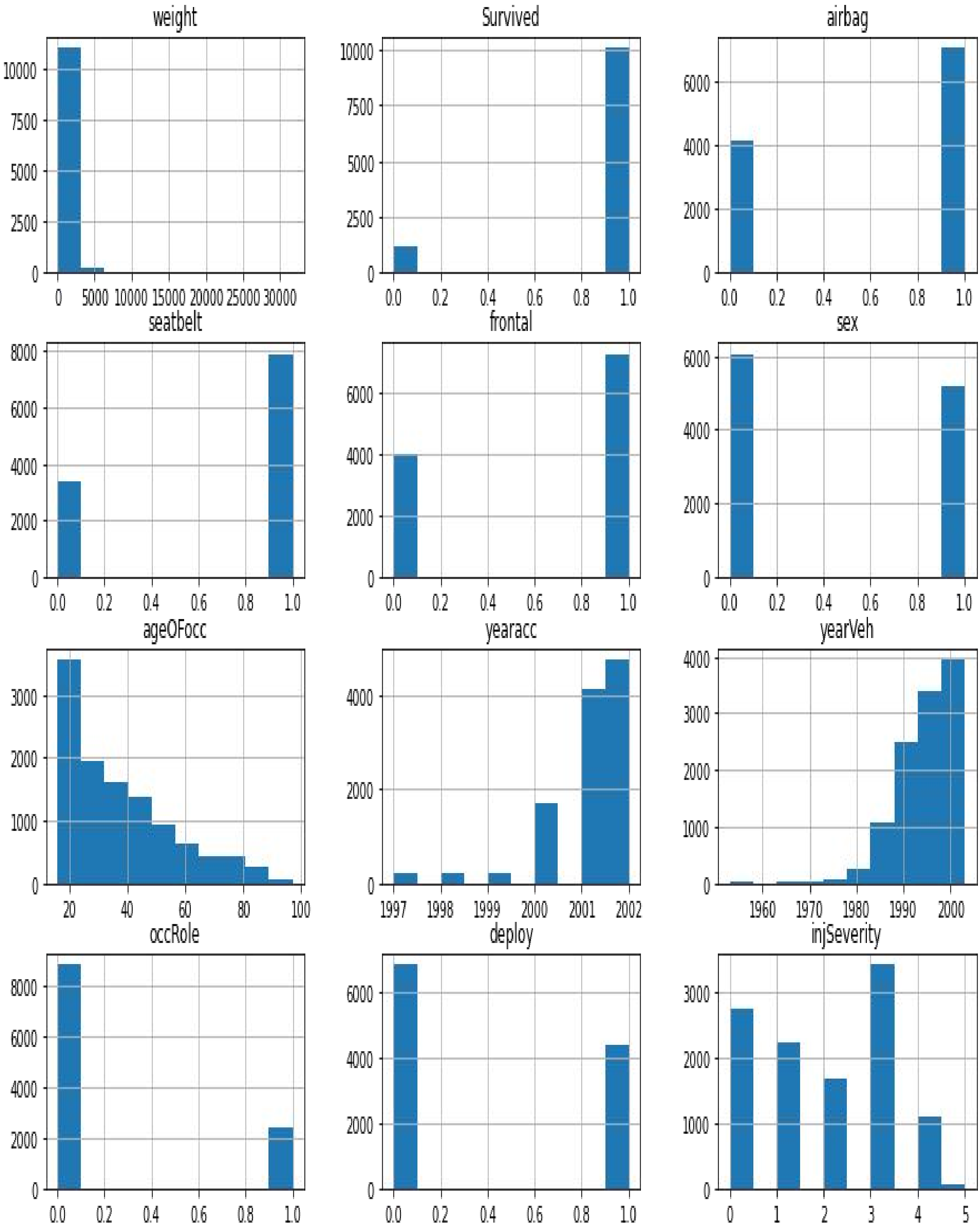

Fig.2.2 - Histgram Plot

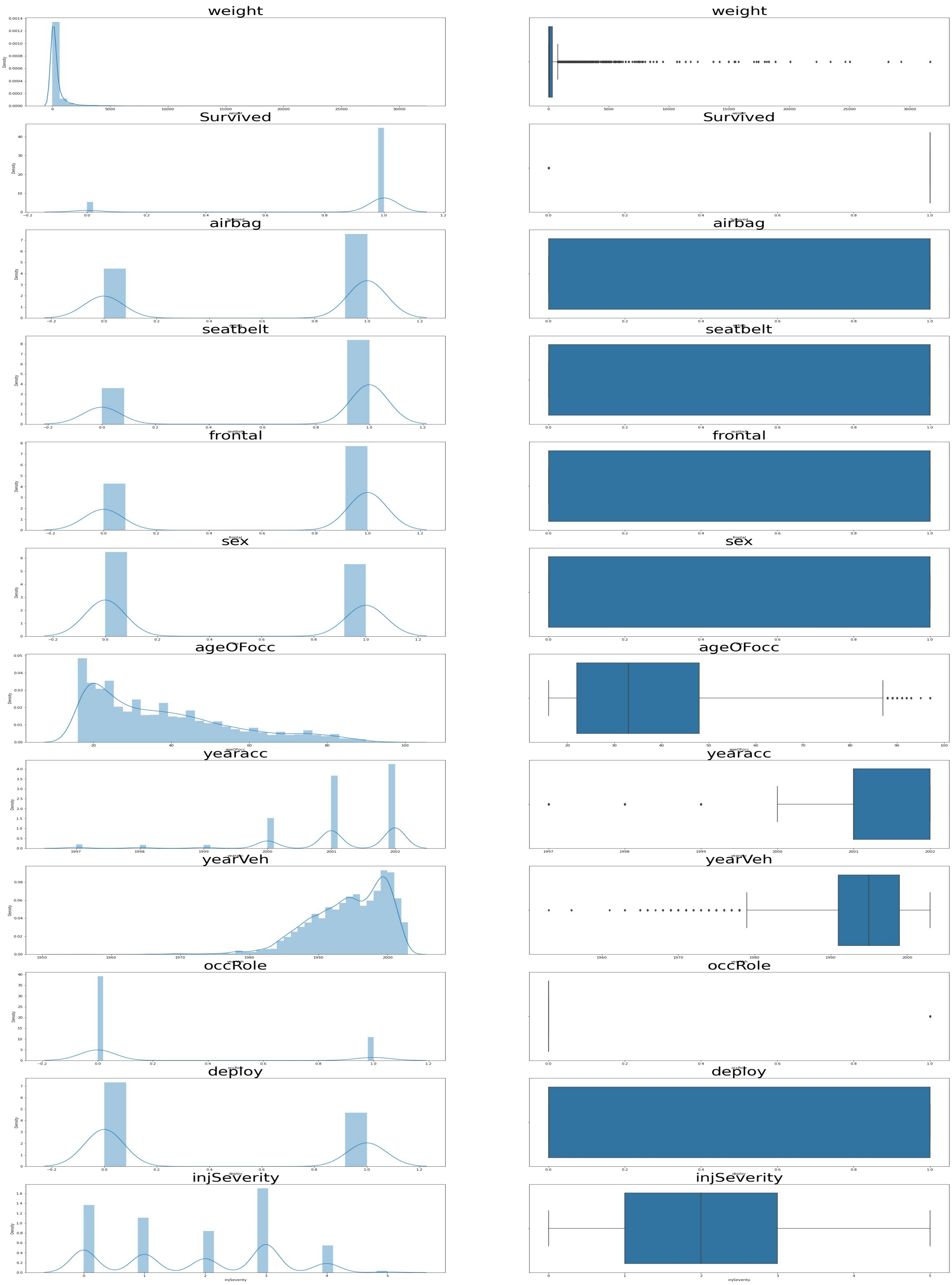

Fig.2.3 - Distribution Plot and Boxplot

	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	occRole	deploy	injSeverity
weight	1.000000	0.091640	-0.003574	0.078739	0.000659	0.006471	-0.040111	0.056892	-0.015226	-0.000219	-0.065783	-0.220659
Survived	0.091640	1.000000	0.139679	0.206467	0.107990	0.046499	-0.135473	0.549885	0.165096	-0.023460	0.054346	-0.517637
airbag	-0.003574	0.139679	1.000000	0.157501	-0.050272	0.092886	0.025109	0.181478	0.766181	-0.086011	0.611983	-0.124394
seatbelt	0.078739	0.206467	0.157501	1.000000	-0.066590	0.117071	0.066066	0.149208	0.180534	-0.047712	0.044132	-0.283063
frontal	0.000659	0.107990	-0.050272	-0.066590	1.000000	-0.055639	-0.048856	0.059768	-0.024267	-0.033721	0.260388	-0.053709
sex	0.006471	0.046499	0.092886	0.117071	-0.055639	1.000000	0.063575	0.025957	0.097390	0.116228	0.036143	0.021284
ageOFocc	-0.040111	-0.135473	0.025109	0.066066	-0.048856	0.063575	1.000000	-0.072271	-0.002070	-0.052485	-0.009556	0.123495
yearacc	0.056892	0.549885	0.181478	0.149208	0.059768	0.025957	-0.072271	1.000000	0.247743	-0.018217	0.091252	-0.300495
yearVeh	-0.015226	0.165096	0.766181	0.180534	-0.024267	0.097390	-0.002070	0.247743	1.000000	-0.018416	0.452448	-0.138475
occRole	-0.000219	-0.023460	-0.086011	-0.047712	-0.033721	0.116228	-0.052485	-0.018217	-0.018416	1.000000	-0.084323	0.018918
deploy	-0.065783	0.054346	0.611983	0.044132	0.260388	0.036143	-0.009556	0.091252	0.452448	-0.084323	1.000000	0.036133
injSeverity	-0.220659	-0.517637	-0.124394	-0.283063	-0.053709	0.021284	0.123495	-0.300495	-0.138475	0.018918	0.036133	1.000000

Table 2.5 - Correlation Table

	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	occRole	deploy	injSeverity
weight	1.977407e+06	39.538440	-2.427227	50.754708	0.443452	4.535851	-1026.128991	84.546094	-121.156698	-0.126992	-45.103025	-427.336576
Survived	3.953844e+01	0.094139	0.020695	0.029038	0.015865	0.007112	-0.756185	0.178301	0.286642	-0.002966	0.008130	-0.218732
airbag	-2.427227e+00	0.020695	0.233184	0.034863	-0.011624	0.022359	0.220582	0.092612	2.093616	-0.017113	0.144089	-0.082727
seatbelt	5.075471e+01	0.029038	0.034863	0.210122	-0.014616	0.026751	0.550940	0.072281	0.468288	-0.009011	0.009863	-0.178699
frontal	4.434517e-01	0.015865	-0.011624	-0.014616	0.229278	-0.013280	-0.425586	0.030244	-0.065752	-0.006653	0.060792	-0.035419
sex	4.535851e+00	0.007112	0.022359	0.026751	-0.013280	0.248487	0.576538	0.013674	0.274715	0.023872	0.008785	0.014612
ageOFocc	-1.026129e+03	-0.756185	0.220582	0.550940	-0.425586	0.576538	330.964474	-1.389464	-0.213142	-0.393423	-0.084764	3.094151
yearacc	8.454609e+01	0.178301	0.092612	0.072281	0.030244	0.013674	-1.389464	1.116838	1.481539	-0.007932	0.047020	-0.437354
yearVeh	-1.211567e+02	0.286642	2.093616	0.468288	-0.065752	0.274715	-0.213142	1.481539	32.020936	-0.042937	1.248330	-1.079169
occRole	-1.269921e-01	-0.002966	-0.017113	-0.009011	-0.006653	0.023872	-0.393423	-0.007932	-0.042937	0.169770	-0.016940	0.010735
deploy	-4.510303e+01	0.008130	0.144089	0.009863	0.060792	0.008785	-0.084764	0.047020	1.248330	-0.016940	0.237732	0.024263
injSeverity	-4.273366e+02	-0.218732	-0.082727	-0.178699	-0.035419	0.014612	3.094151	-0.437354	-1.079169	0.010735	0.024263	1.896717

Table 2.6 - Covariance Table

Fig.2.4 - Correlation Heatmap

The above tables and plot shows ,

- A positive and strong correlation between ***yearVeh*** , ***airbag*** and ***deploy.*** This means that older models of vehicles don't have airbags and it lead to accident.
- While the ***yearVeh*** and ***airbag*** have negative correlation with ***injSeverrity*** and positive correlation with ***deploy,*** which means the new have vehicles have airbags and their deployment in-time have prevented any major injury or severe condition.
- **Drivers** are mostly injured than the **passengers** unless the passengers are not wearing **seatbelt.**
- **Seatbelt** in marginally are positive with most variables except with ***frontal*** and ***injSeverrity*** and frontal have caused sever accidents unless the airbags have deployed.

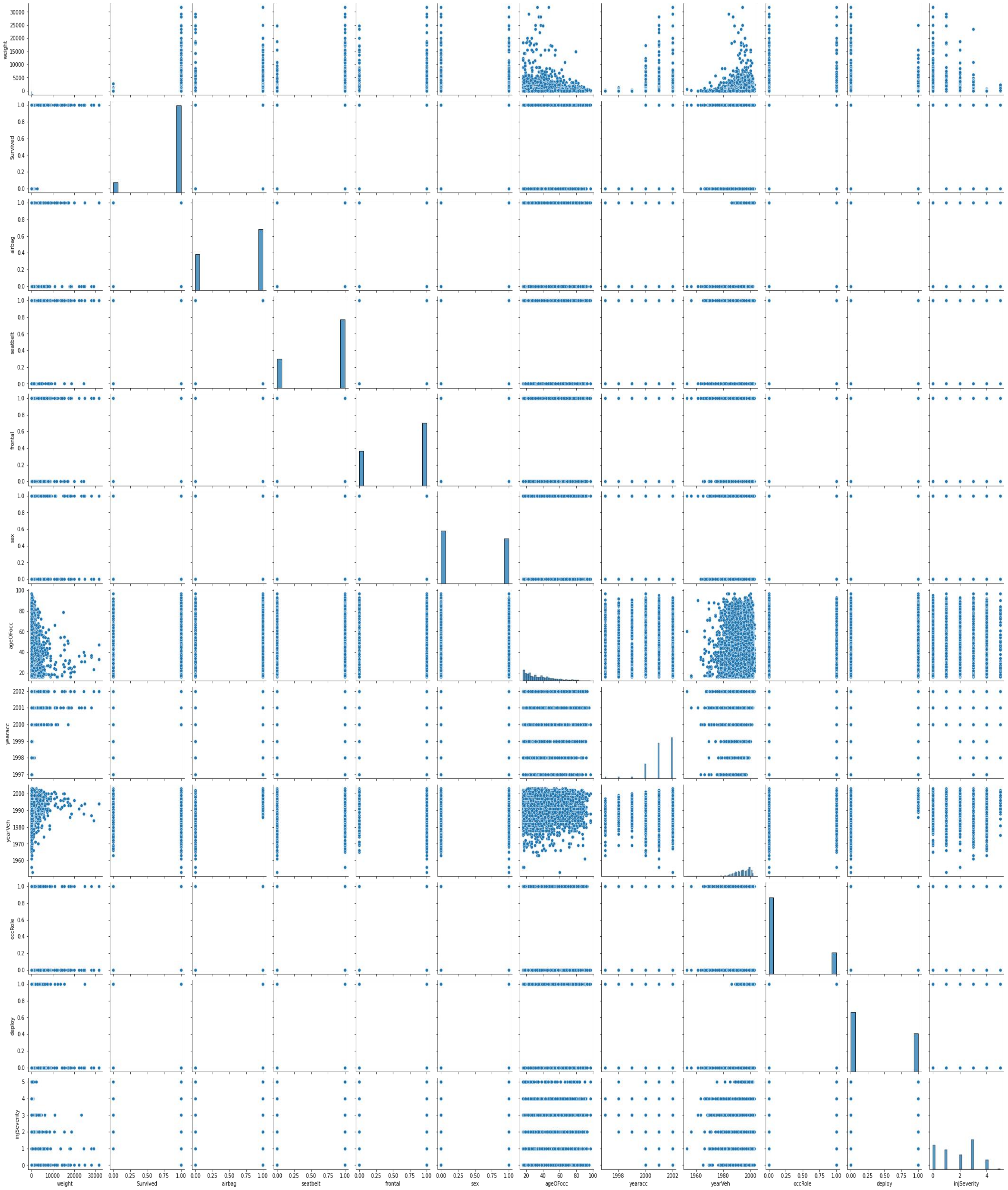

Fig.2.5 - Pair Plot

Outlier treatment : Most machine learning algorithms (models) do not work accurately in the presence of outlier. It helps us to normalise the data distribution.

weight	11.115386
Survived	-2.573960
airbag	-0.537519
seatbelt	-0.871645
frontal	-0.601667
sex	0.157231
ageOFocc	0.911059
yearacc	-1.671687
yearVeh	-1.026743
occRole	1.375262
deploy	0.454813
injSeverity	0.021729
dtype: float64	

Table 2.7 - Skewness in Data

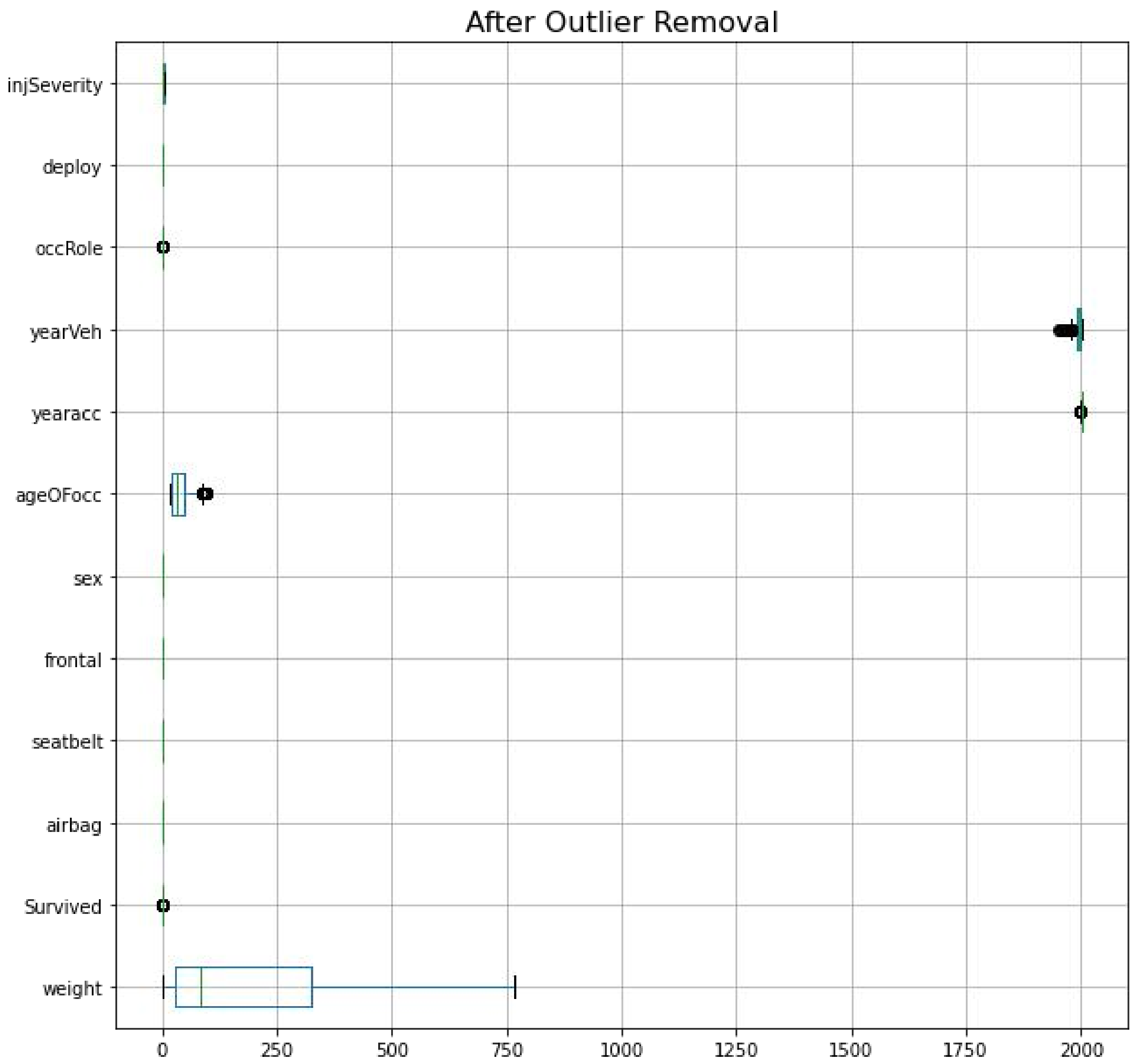

Fig.2.6 - Boxplot After Outlier Treatment.

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).

Label Encoding,

We will apply label encoding to the following using from sklearn LabelEncoder:

- ✓ **dvcat**
- ✓ **abcat**

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
0	4	27.078	0	0	0	1	0	32	1997	1987.0	2	0	0	4.0	2:13:2
1	2	89.627	0	1	1	0	1	54	1997	1994.0	1	0	0	4.0	2:17:1

Table 2.8 - Data head - After Label Encoding

```
<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  int32
 1   weight       11217 non-null  float64
 2   Survived     11217 non-null  int64
 3   airbag       11217 non-null  int64
 4   seatbelt     11217 non-null  int64
 5   frontal      11217 non-null  int64
 6   sex          11217 non-null  int64
 7   ageOFocc     11217 non-null  int64
 8   yearacc      11217 non-null  int64
 9   yearVeh      11217 non-null  float64
 10  abcat        11217 non-null  int32
 11  occRole      11217 non-null  int64
 12  deploy       11217 non-null  int64
 13  injSeverity  11217 non-null  float64
 14  caseid       11217 non-null  object
dtypes: float64(3), int32(2), int64(9), object(1)
memory usage: 1.2+ MB
```

Table 2.9 - Data info - After Label Encoding

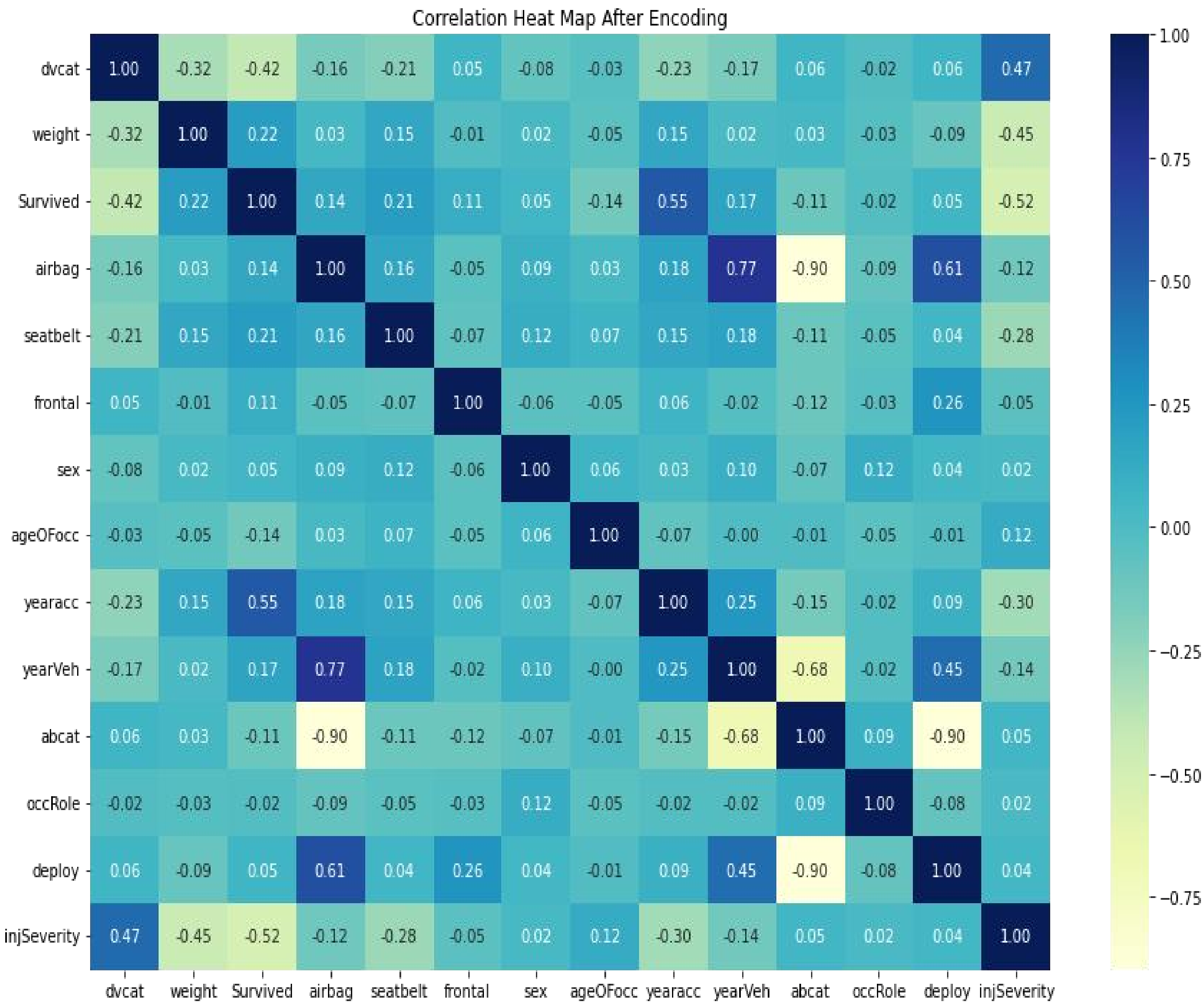

Fig.2.7 - Correlation Heatmap - After Label Encoding

Creating Train_test data and Splitting the data - 70:30 :

Train- test shape,
X _train shape : (7851, 13)
y_train shape : (7851,)
X_Test Shape :(3366, 13)
y_test shape : (3366,)

Now we will apply Logistic Regression and LDA (Linear Discriminant Analysis).

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

Logistic Regression Model :

For Train,

```
0.9916359512333428
[[ 705   92]
 [  59 6995]]
              precision    recall  f1-score   support

           0       0.92      0.88      0.90       797
           1       0.99      0.99      0.99      7054

    accuracy                           0.98      7851
   macro avg       0.95      0.94      0.95      7851
weighted avg       0.98      0.98      0.98      7851
```

Accuracy - 98 %
Precision - 99%
Recall - 99 %
F1 score - 99 %

For Test,

```
0.992289641300704
[[ 339   44]
 [  23 2960]]
              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
```

Accuracy - 98 %
Precision - 99%
Recall - 99 %
F1 score - 99 %

AUC - ROC Plot :

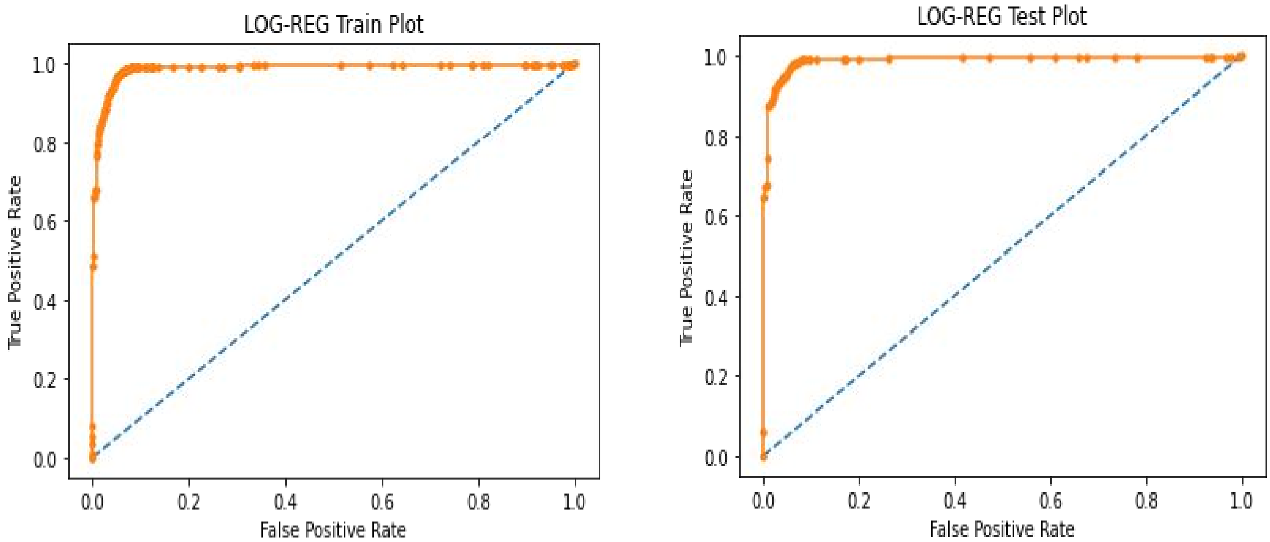

Fig 2.8 - Logistic Regression - AUC Curve Plot- Train and Test

Confusion Matrix :

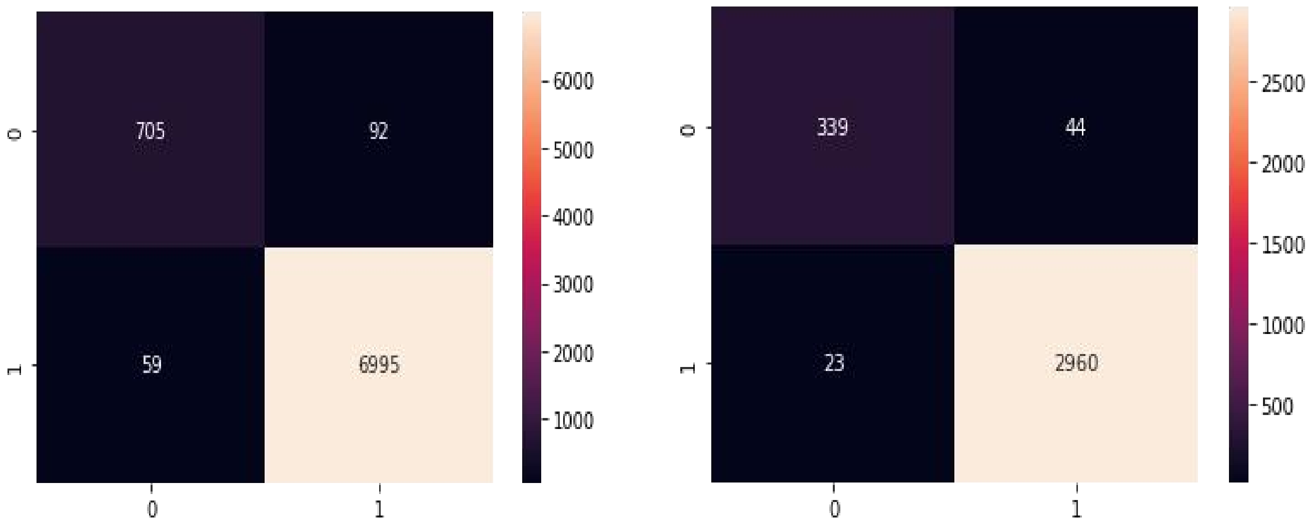

Fig 2.9 - LOGREG -Confusion Matrix- Train (left) and Test (right)

RSME for Train Data = 0.019233218698255
RSME for Test Data = 0.01990493166963755

Observation -

- Based on the above results, the model is neither under fitted nor over-fitted .
- Both train and test data has 98% accuracy. Hence, it's a very good model.

LDA (linear discriminant analysis) Model :

For Train,

```
0.9900765523107457
[[ 547  250]
 [  70 6984]]
              precision    recall  f1-score   support

           0       0.89      0.69      0.77       797
           1       0.97      0.99      0.98      7054

    accuracy                           0.96      7851
   macro avg       0.93      0.84      0.88      7851
weighted avg       0.96      0.96      0.96      7851
```

Accuracy - 96 %
Precision - 97%
Recall - 99 %
F1 score - 98 %

For Test,

```
0.9916191753268522
[[ 276  107]
 [  25 2958]]
              precision    recall  f1-score   support

           0       0.92      0.72      0.81       383
           1       0.97      0.99      0.98      2983

    accuracy                           0.96      3366
   macro avg       0.94      0.86      0.89      3366
weighted avg       0.96      0.96      0.96      3366
```

Accuracy - 96 %
Precision - 97%
Recall - 99 %
F1 score - 98 %

AUC - ROC Plot :

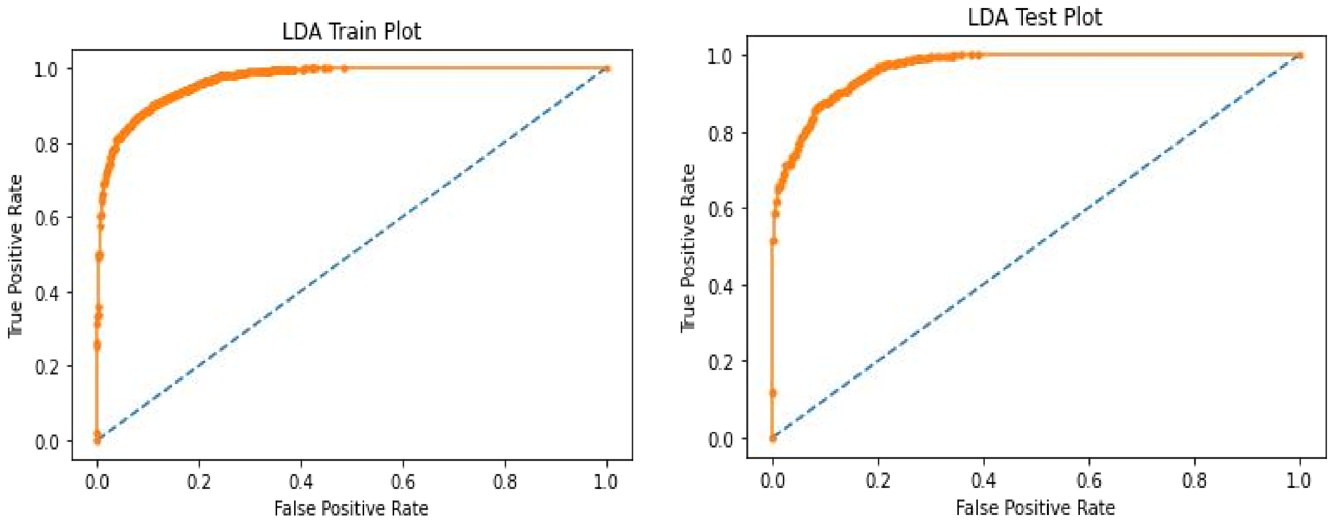

Fig 2.10 - LDA -AUC Curve Plot- Train and Test

Confusion Matrix :

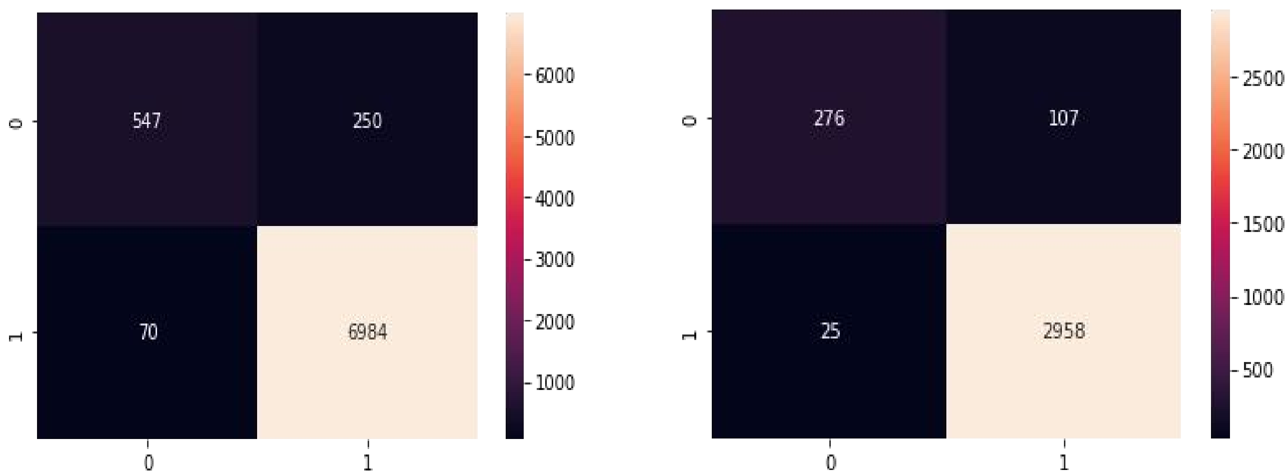

Fig 2.11 - LDA -Confusion Matrix- Train (left) and Test (right)

RSME for Train Data = 0.040759138963189404
RSME for Test Data = 0.0392156862745098

Observation -

- Based on the above results, the model is neither under fitted nor over-fitted .
- Both train and test data has 96% accuracy. Hence, it’s also good model.

	Train Recall	Test Recall
LogReg_Model	0.991636	0.992290
LDA_Model	0.990077	0.991619

Table 2.10 - Table for all Models - Recall values

Observations :

- The Logistic Regression Model has 98%accuracy .
- The LDA Model has 96% accuracy .
- As per the above results both models are very good , but Logistic Regression model looks as the best model when compared.

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

<u>Business Recommendations</u>:

- The above models show a good accuracy and both can be considered as good model.
- The government can spread awareness of the rate of accidents with poor safety standards of older vehicles.
- State sponsored and organised awareness campaigns can be organised through papers and digital media.
- The government may impose fine on older vehicles and give a certain time-period for the owners to dispose of the older vehicles.
- The government can deploy nation wide circulars on road safety rules for seat belts and airbag as mandatory precaution for all vehicles.
- The traffic regulatory or the concerned department can circulate seat belts and airbag as a mandatory equipment for all vehicles.
- If the above rules are applicable to all vehicles and applicable to both the driver and passenger and heavy fine can imposed under non-compliance.
- The government should make seatbelt and airbags as mandatory part for these vehicles to the vehicle manufacturers irrespective of the brand.

New guidelines and threshold can be set out for vehicle to pass quality tests for these vehicles.