PREDICTIVE MODELLING PROJECT - EXTENDED

DSBA

Krishnabhamini Sinha

Contents

Data Dictionary for Problem 1: (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.

Data Dictionary for Problem 2: (Car_Crash)

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.

Problem 1.1 - Define the problem and perform Exploratory Data Analysis

- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Key meaningful observations on individual variables and the relationship between variables

Problem Definition:

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.

Solution:

Shape:

The given dataset has 759 rows and 10 columns.

```
data.shape
```

```
(759, 10)
```

Data Types:

The data type for each column is enlisted as below in the figure given below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 759 entries, 0 to 758
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    759 non-null    int64
 1   sales         759 non-null    float64
 2   capital       759 non-null    float64
 3   patents       759 non-null    int64
 4   randd         759 non-null    float64
 5   employment    759 non-null    float64
 6   sp500         759 non-null    object
 7   tobinq        738 non-null    float64
 8   value         759 non-null    float64
 9   institutions  759 non-null    float64
dtypes: float64(7), int64(2), object(1)
memory usage: 59.4+ KB
```

First five rows:

	Unnamed: 0	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	0	826.99505	161.60399	10	382.07825	2.30600	no	11.04951	1625.45376	80.27000
1	1	407.75397	122.10101	2	0.00000	1.86000	no	0.84419	243.11708	59.02000
2	2	8407.84559	6221.14461	138	3296.70044	49.65900	yes	5.20526	25865.23380	47.70000
3	3	451.00001	266.89999	1	83.54016	3.07100	no	0.30522	63.02463	26.88000
4	4	174.92798	140.12400	2	14.23364	1.94700	no	1.06330	67.40641	49.46000

Last five rows:

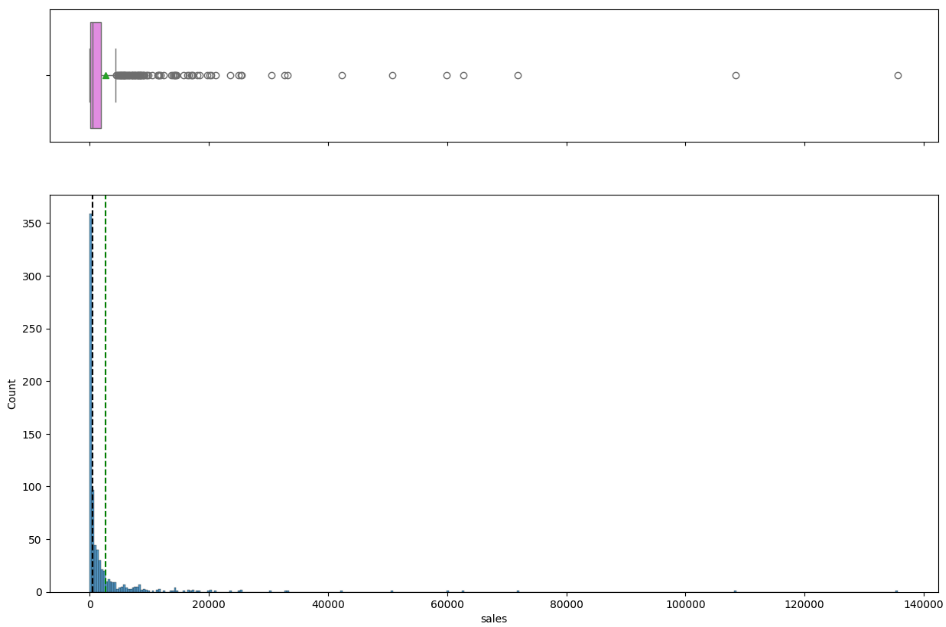
	Unnamed: 0	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
754	754	1253.90020	708.29994	32	412.93616	22.10000	yes	0.69745	267.11949	33.50000
755	755	171.82102	73.66601	1	0.03774	1.68400	no	NaN	228.47570	46.41000
756	756	202.72697	123.92699	13	74.86110	1.46000	no	5.22972	580.43074	42.25000
757	757	785.68794	138.78099	6	0.62175	2.90000	yes	1.62540	309.93865	61.39000
758	758	22.70200	14.24500	5	18.57436	0.19700	no	2.21307	18.94014	7.50000

Statistical Summary:

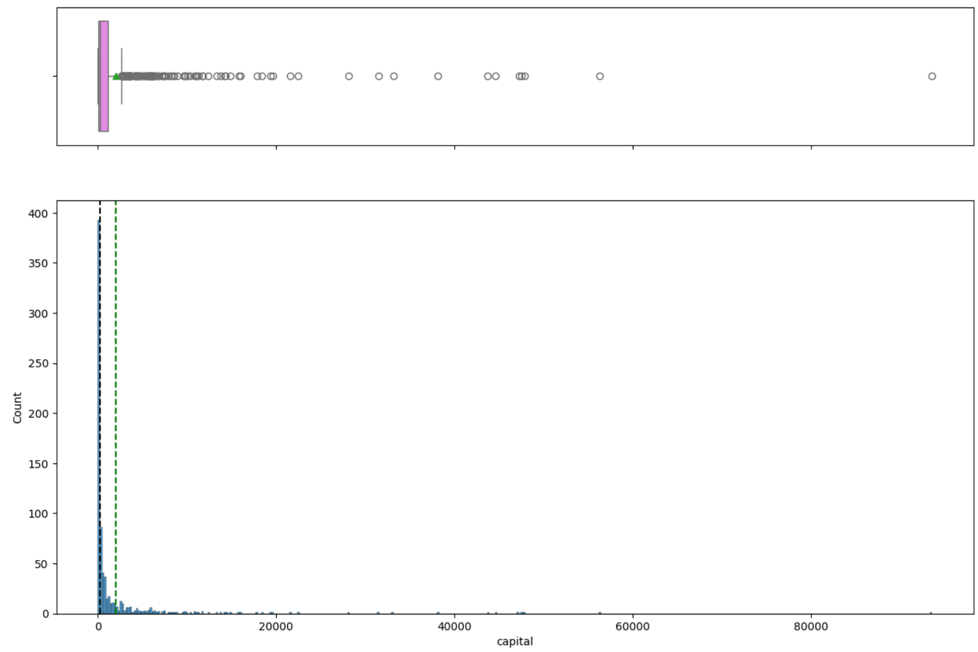
	sales	capital	patents	randd	employment	tobinq	value	institutions
count	759.00000	759.00000	759.00000	759.00000	759.00000	738.00000	759.00000	759.00000
mean	2689.70516	1977.74750	25.83136	439.93807	14.16452	2.79491	2732.73475	43.02054
std	8722.06012	6466.70490	97.25958	2007.39759	43.32144	3.36659	7071.07236	21.68559
min	0.13800	0.05700	0.00000	0.00000	0.00600	0.11900	1.97105	0.00000
25%	122.92000	52.65050	1.00000	4.62826	0.92750	1.01878	103.59395	25.39500
50%	448.57708	202.17902	3.00000	36.86414	2.92400	1.68030	410.79353	44.11000
75%	1822.54737	1075.79002	11.50000	143.25340	10.05000	3.13931	2054.16039	60.51000
max	135696.78820	93625.20056	1220.00000	30425.25586	710.79993	20.00000	95191.59116	90.15000

Univariate Analysis:

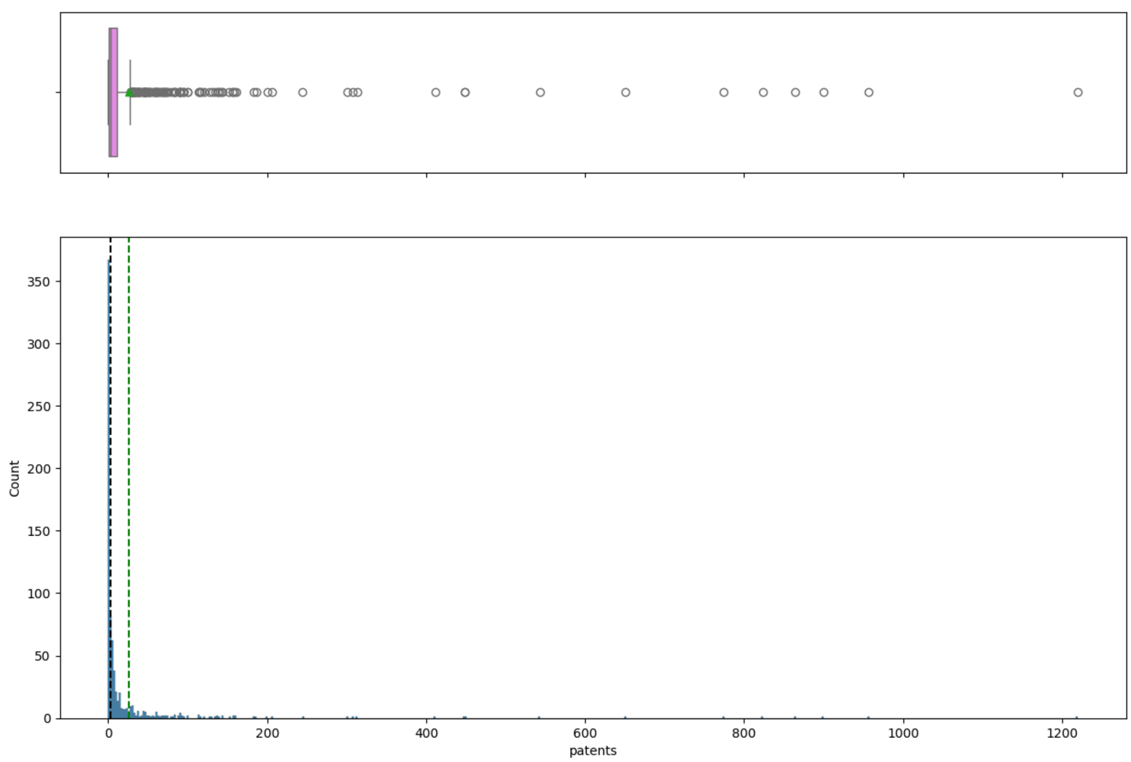
Observation on Sales:



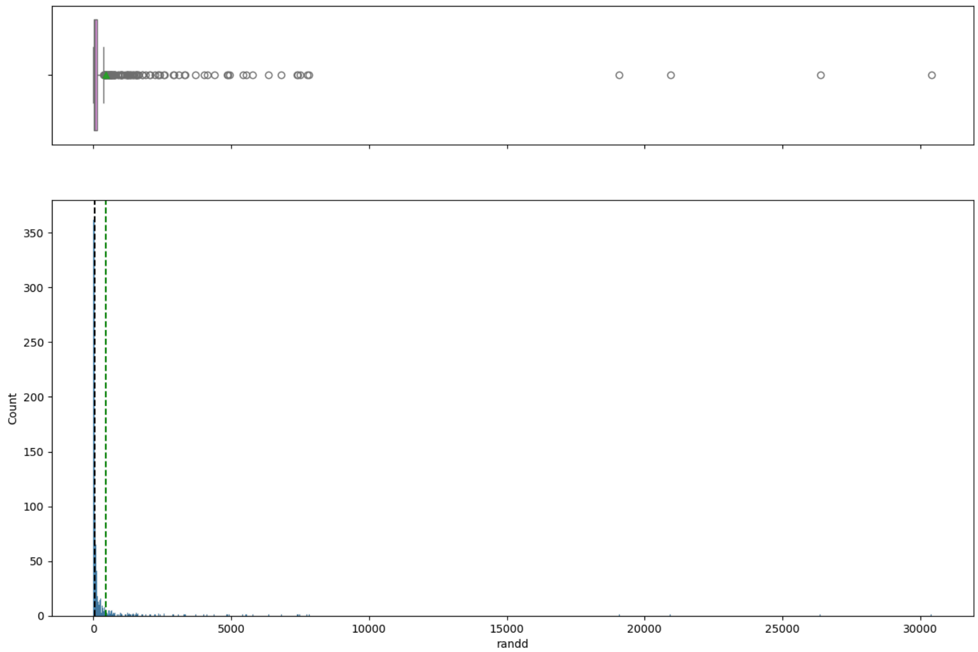
Observation on capital:



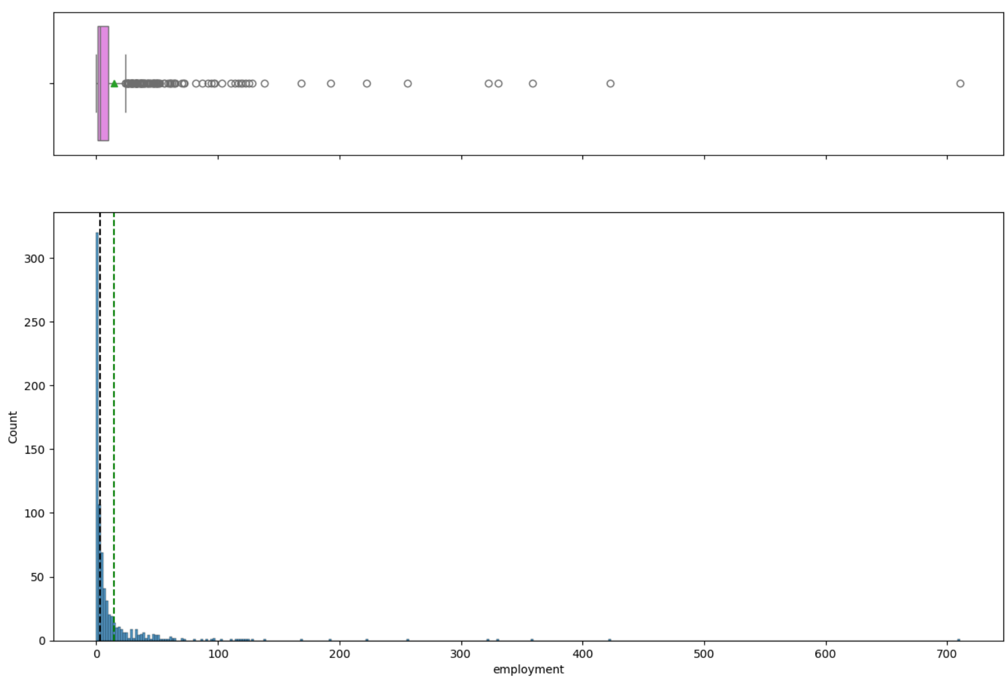
Observation on number of patents:



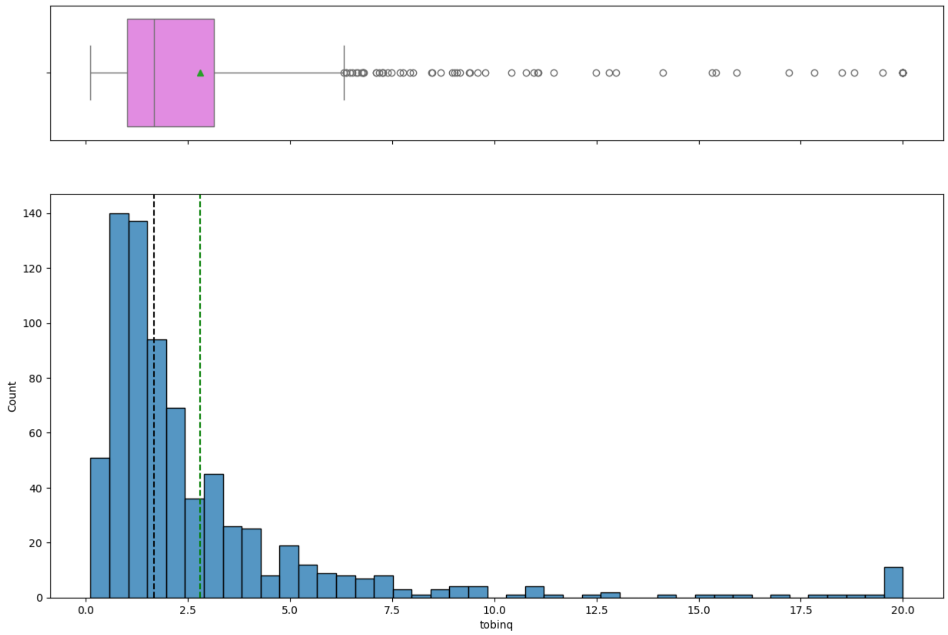
Observations on number of R&D:



Observations on number of employment:

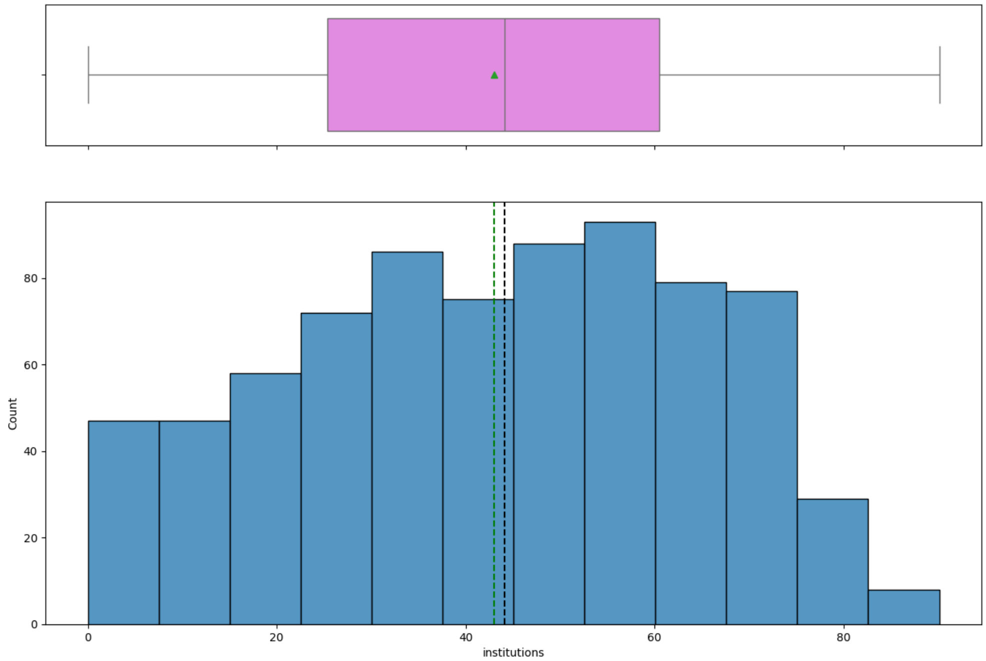


Observations on tobinq:

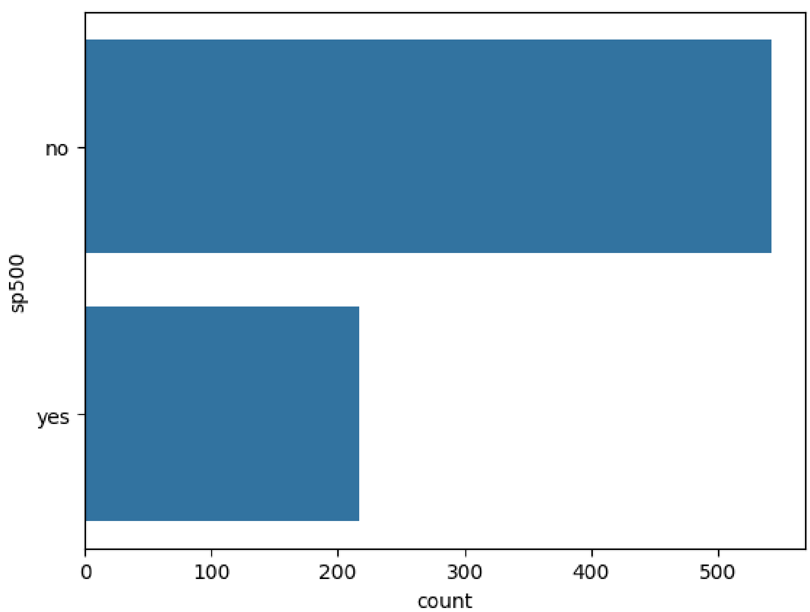


Observations on stock market value:

Observations on institutions:

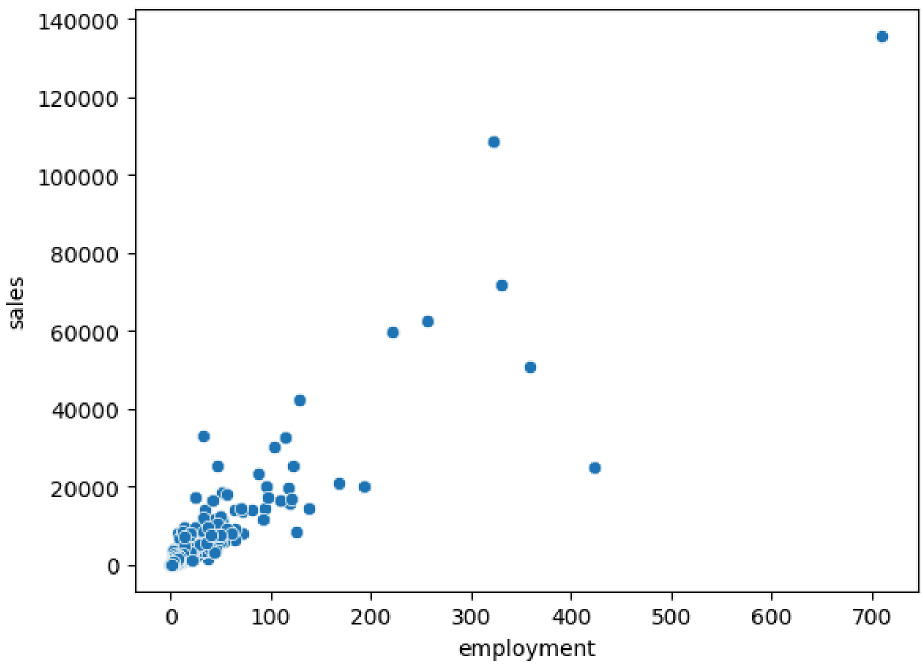


Observations on sp500:

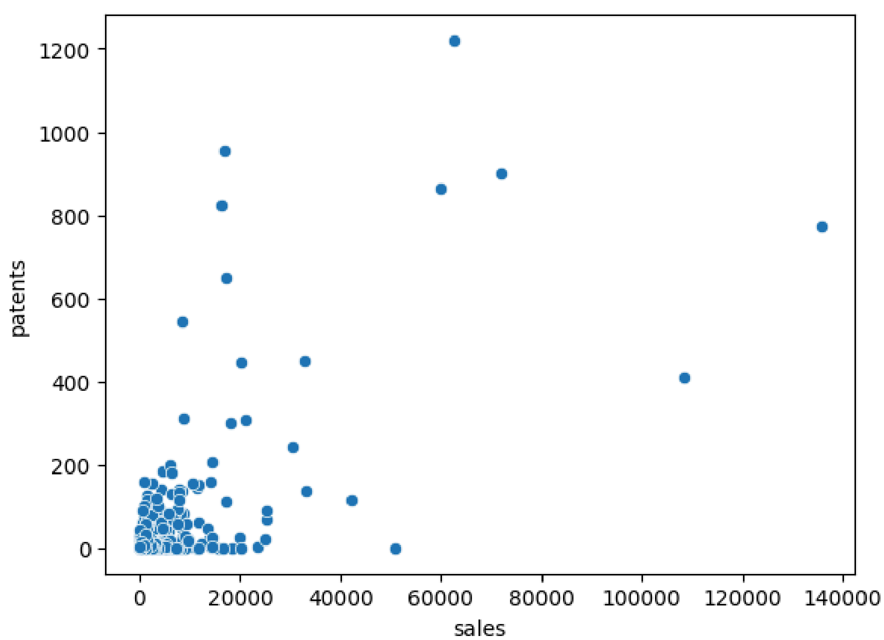


Multivariate analysis

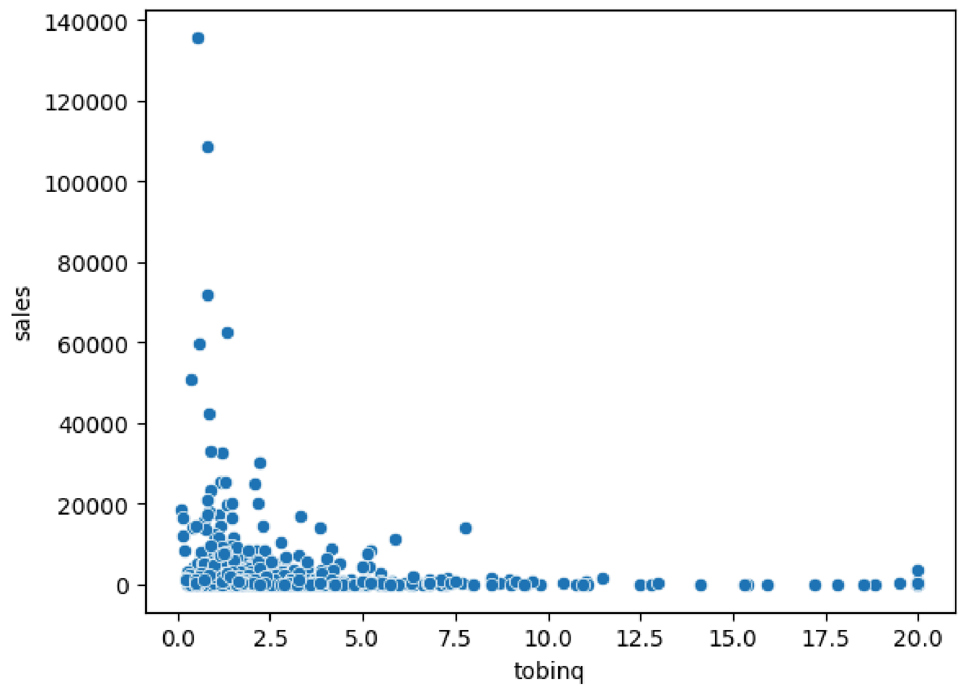
Observation of employment on sales:



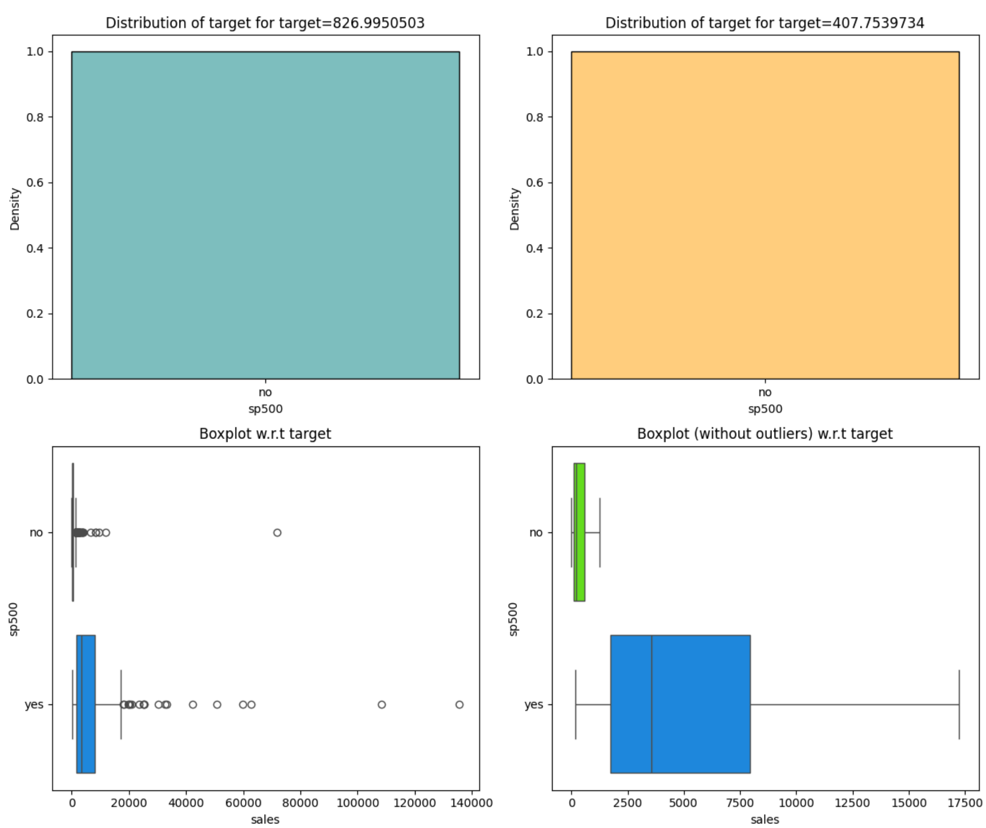
Observation of Patents on sales:



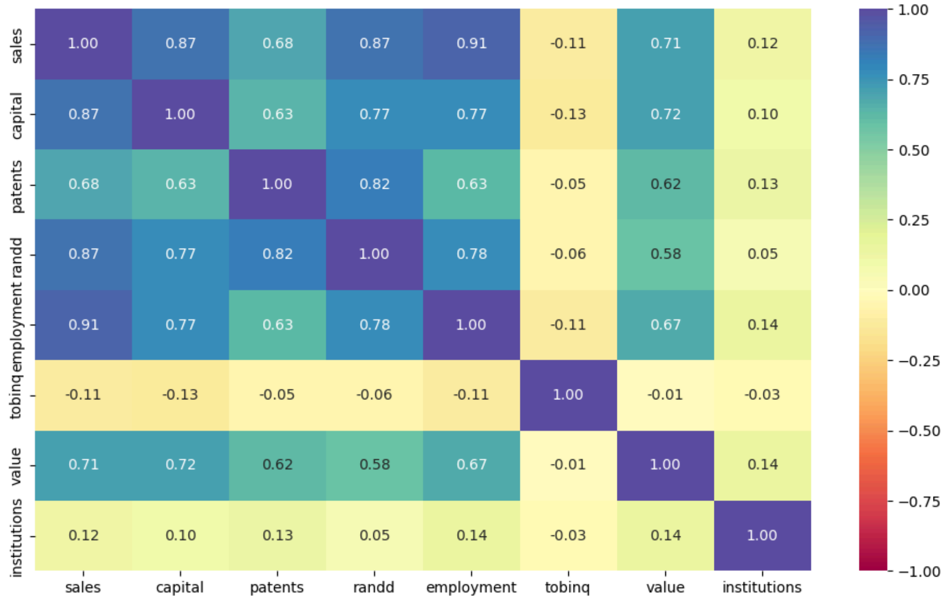
Observation of tobinq on sales:

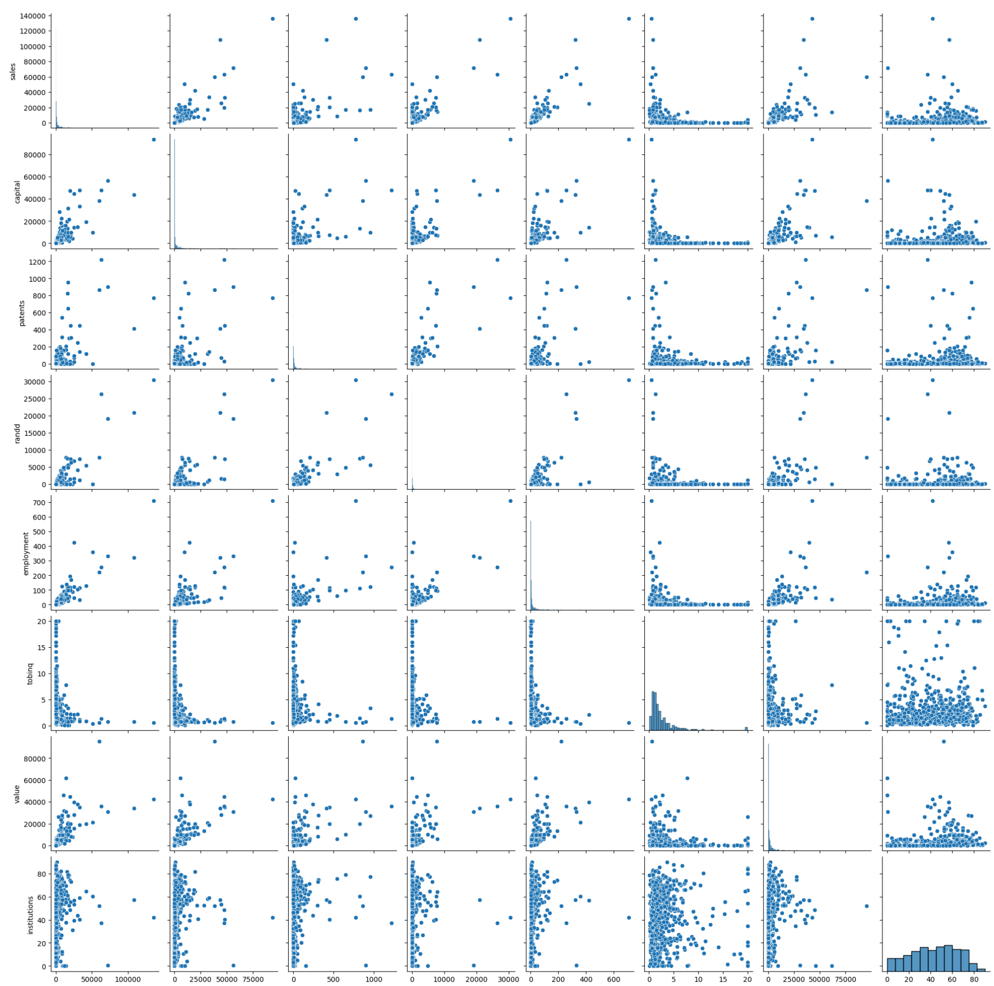


Distribution of sales on sp500:



Correlation between Numerical Variables:





Key meaningful observations on individual variables and the relationship between variables:

- R&D stocks are much lesser in values than other attributes like sales and capital.
- Most firms have around 14,000 employees.
- Most firms have tobinq values <=2.5.
- Most of the firms do not have membership in the S&P 500 index.
- Firms that have membership in the S&P 500 index have a higher number of sales.

Problem 1.2 - Data Preprocessing

Prepare the data for modelling: - Missing value Treatment (if needed) - Outlier Detection(treat, if needed) - Encode the data - Data split

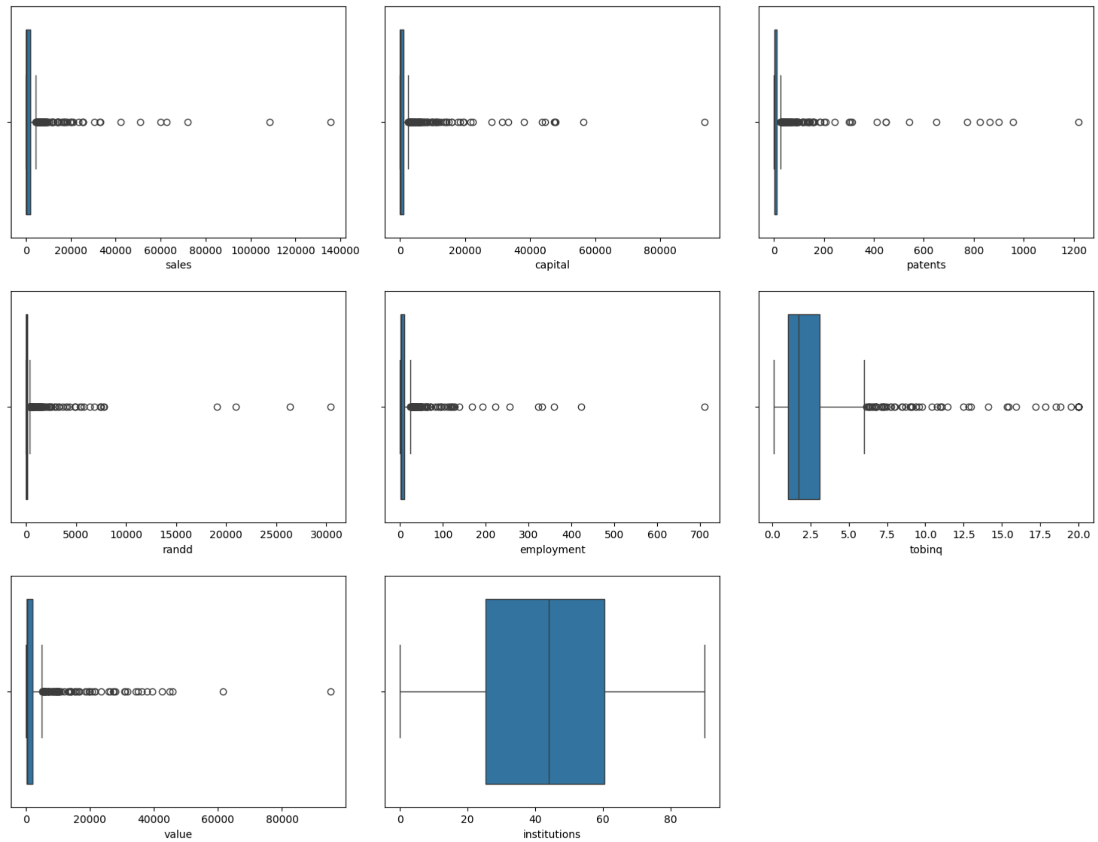
Solution:

```
data['tobinq'] = data['tobinq'].fillna(data["tobinq"].mean())
```

21 tobinq null values are imputed with its mean value. No null values exist in the dataset now.

sales	0
capital	0
patents	0
randd	0
employment	0
sp500	0
tobinq	0
value	0
institutions	0

Outlier treatment:



There are too many outliers in the data and treating those might alter the nature of findings and impact our intended results. Hence, we will not treat the outliers.

Encode the data

	const	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	1.00000	161.60399	10	382.07825	2.30600	0	11.04951	1625.45376	80.27000
1	1.00000	122.10101	2	0.00000	1.86000	0	0.84419	243.11708	59.02000
2	1.00000	6221.14461	138	3296.70044	49.65900	1	5.20526	25865.23380	47.70000
3	1.00000	266.89999	1	83.54016	3.07100	0	0.30522	63.02463	26.88000
4	1.00000	140.12400	2	14.23364	1.94700	0	1.06330	67.40641	49.46000

The column sp 500 has been label encoded using LabelEncoder from sklearn.preprocessing.

Data split

Data has been split into training and test sets in a 70:30 ratio.

```
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=1)
```

```
print("Number of rows in train data =", x_train.shape[0])
print("Number of rows in test data =", x_test.shape[0])
```

```
Number of rows in train data = 531
Number of rows in test data = 228
```

Problem 1.3 - Model Building - Linear regression

- Apply linear Regression - Using Statsmodels Perform checks for significant variables using appropriate method by building multiple models - Create multiple models by dropping insignificant variables - Check the performance of all models on train and test sets using different performance metrics.

Solution:

Using Statsmodels we create a Linear Regression model with all the original variables.

```
                            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:                Mon, 19 Aug 2024   Prob (F-statistic):          1.38e-306
Time:                        15:45:18   Log-Likelihood:                -4831.5
No. Observations:                 531   AIC:                             9681.
Df Residuals:                     522   BIC:                             9719.
Df Model:                           8
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           53.6940    233.554      0.230      0.818    -405.128     512.516
capital          0.4142      0.027     15.563      0.000       0.362       0.467
patents         -5.0445      2.407     -2.096      0.037      -9.773      -0.316
randd            1.0261      0.127      8.052      0.000       0.776       1.276
employment      83.9582      3.629     23.135      0.000      76.829      91.087
sp500         -102.0063    267.962     -0.381      0.704    -628.423     424.410
tobinq         -31.2964     30.297     -1.033      0.302     -90.816      28.223
value            0.1267      0.022      5.885      0.000       0.084       0.169
institutions     1.0627      4.964      0.214      0.831      -8.690      10.816
==============================================================================
Omnibus:                      231.611   Durbin-Watson:                   1.932
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31505.675
Skew:                           0.809   Prob(JB):                         0.00
Kurtosis:                      40.701   Cond. No.                     2.90e+04
==============================================================================
```

R-squared comes up to be 93.6% while the adjusted R-squared is 93.5. Performance of the model 1 on the training and test data is shown as below:

Training Performance

	R-squared	Adj. R-squared
0	0.93638	0.93528

Test Performance

	R-squared	Adj. R-squared
0	0.89277	0.88834

We notice that there are multiple variables that have p-values > 0.05. We are going to drop those one by one until we reach a reasonably optimal number of independent variables. We create model number 2.

The outcome of this operation is as follows:

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                  sales   R-squared (uncentered):                   0.942
Model:                            OLS   Adj. R-squared (uncentered):              0.942
Method:                 Least Squares   F-statistic:                              1717.
Date:                Mon, 19 Aug 2024   Prob (F-statistic):                   4.94e-323
Time:                        15:57:11   Log-Likelihood:                         -4832.1
No. Observations:                 531   AIC:                                      9674.
Df Residuals:                     526   BIC:                                      9696.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
capital        0.4173      0.026     15.951      0.000       0.366       0.469
patents       -4.9106      2.358     -2.083      0.038      -9.543      -0.279
randd          1.0161      0.124      8.214      0.000       0.773       1.259
employment    84.1633      3.528     23.853      0.000      77.232      91.095
value          0.1217      0.020      5.974      0.000       0.082       0.162
==============================================================================
Omnibus:                      234.378   Durbin-Watson:                   1.926
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31417.831
Skew:                           0.840   Prob(JB):                         0.00
Kurtosis:                      40.646   Cond. No.                         384.
==============================================================================
```

Performance on training and test data is as follows:

Training Performance

	R-squared	Adj. R-squared
0	0.93623	0.93563

Test Performance

	R-squared	Adj. R-squared
0	0.89216	0.88973

Although we can see that there is not much of a difference in the performances of model 1 and model 2, we would like to ensure that the independent variables used in the model show a significant relationship with the dependent variable. So, we now have variables in the model with p-values < 0.05. And we finalize model number 2.

Problem 1.4 - Business Insights & Recommendations

- Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. - Comment on the Linear Regression equation from the final model.

Solution:

Steps involved in this project:

- Data has been studied, cleaned and unnecessary columns such as "Unnamed : 0" have been removed from the data.
- Visualization of variables in the data has been done both individually and with respect to the dependent variable - "sales".
- Categorical variables like "sp500" were labeled encoded to prepare data for building a linear regression model.
- Data has been divided into training and test sets in the ratio of 70:30 respectively.
- Linear Regression model is built first with all the variables included. The R-squared comes out to be 93.6% i.e., the model can explain 93.6% of variance.
- Due to the presence of high p-values of independent variables, we prune the data to include variables with p-values less than 0.05.
- The second model created does not result in a difference in the R-squared value. But the p-values of the independent variables are less than 0.05 and hence, meet our needs.

The Linear Regression Model Equation:

```
sales = 0.41733392846709144 + -4.910625601703053 * ( patents ) +

1.0160573682973648 * ( randd ) +  84.16331797689071 * ( employment )

+  0.12171917365830337 * ( value )
```

Interpretation of the Coefficients:

- ➢ Intercept (0.417):

This is the base level of sales when all other variables (Patents, R&D, Employment, Value) are zero. It serves as a starting point for the equation but is typically less important in a business context unless variables are close to zero.
Patents (-4.911):

- ➢ For each additional patent, sales are expected to decrease by approximately 4.911 units, holding other factors constant.

Business Insight: A negative coefficient suggests that more patents are associated with lower sales. This could indicate that focusing too much on patent development might be diverting resources away from activities that drive sales. The business should consider evaluating whether the types of patents pursued are aligned with market needs or revenue-generating opportunities.
R&D (1.016):

- ➢ For each unit increase in R&D expenditure, sales are expected to increase by approximately 1.016 units, holding other factors constant.

Business Insight: Positive correlation between R&D and sales suggests that investment in research and development is likely leading to innovation that drives sales. The business should continue investing in R&D, ensuring that it is focused on areas that directly impact sales growth.
Employment (84.163):

- ➢ For each additional unit of employees, sales are expected to increase by approximately 84.163 units, holding other factors constant.

Business Insight: A very strong positive impact of employment on sales indicates that having more employees significantly contributes to sales growth. This could be due to increased production capacity, better customer service, or enhanced operational efficiency. The business might consider expanding its workforce or optimizing employee productivity to further boost sales.
Value (0.122):

- ➢ For each unit increase in stock value, sales are expected to increase by approximately 0.122 units, holding other factors constant.

Business Insight: The positive but smaller coefficient for value suggests that while increasing the perceived or actual value of products or services contributes to sales, its impact is less pronounced than other factors. The business should focus on value enhancement strategies like improving quality or brand positioning but recognize that other factors like employment and R&D might have a more significant impact on sales.

Actionable Insights:

- Re-evaluate Patent Strategy: Investigate why patents are negatively correlated with sales. It may be beneficial to focus on patents that have direct commercial potential or streamline the patenting process to reduce costs.

- Sustain or Increase R&D Investment: Continue investing in R&D, as it positively impacts sales. Ensure that R&D activities are closely aligned with market demands and customer needs.

- Optimize Workforce: Given the strong positive impact of employment on sales, consider expanding the workforce strategically or enhancing employee efficiency through training or better tools.

- Maintain stock Value: While improving value does positively affect sales, it may be secondary to other factors. Ensure that growth and steady development of stock values are maintained throughout for good sales profit.

By focusing on these areas, the business can leverage the insights from the regression analysis to drive sales growth effectively.

Problem 2.1 - Define the problem and perform exploratory Data Analysis

- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables.

Problem Definition:

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.

Shape of data:

```
data.shape
```

```
(11217, 16)
```

Data types:

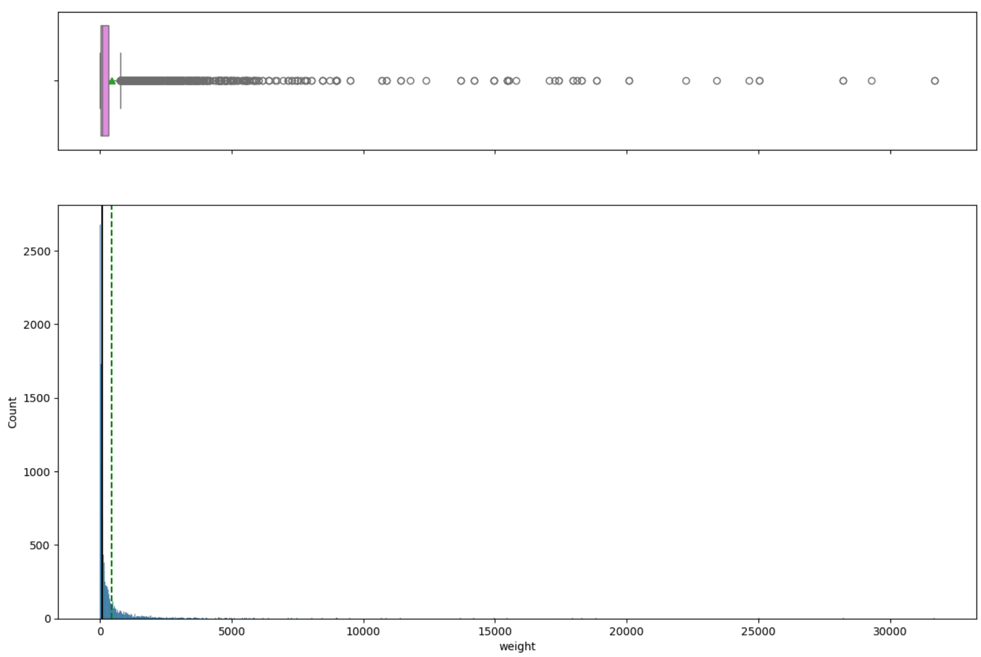
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11217 entries, 0 to 11216
Data columns (total 16 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Unnamed: 0   11217 non-null  int64  
 1   dvcat        11217 non-null  object 
 2   weight       11217 non-null  float64
 3   Survived     11217 non-null  object 
 4   airbag       11217 non-null  object 
 5   seatbelt     11217 non-null  object 
 6   frontal      11217 non-null  int64  
 7   sex          11217 non-null  object 
 8   ageOFocc     11217 non-null  int64  
 9   yearacc      11217 non-null  int64  
 10  yearVeh      11217 non-null  float64
 11  abcat        11217 non-null  object 
 12  occRole      11217 non-null  object 
 13  deploy       11217 non-null  int64  
 14  injSeverity  11140 non-null  float64
 15  caseid       11217 non-null  object 
dtypes: float64(3), int64(5), object(8)
memory usage: 1.4+ MB
```

Statistical Summary:

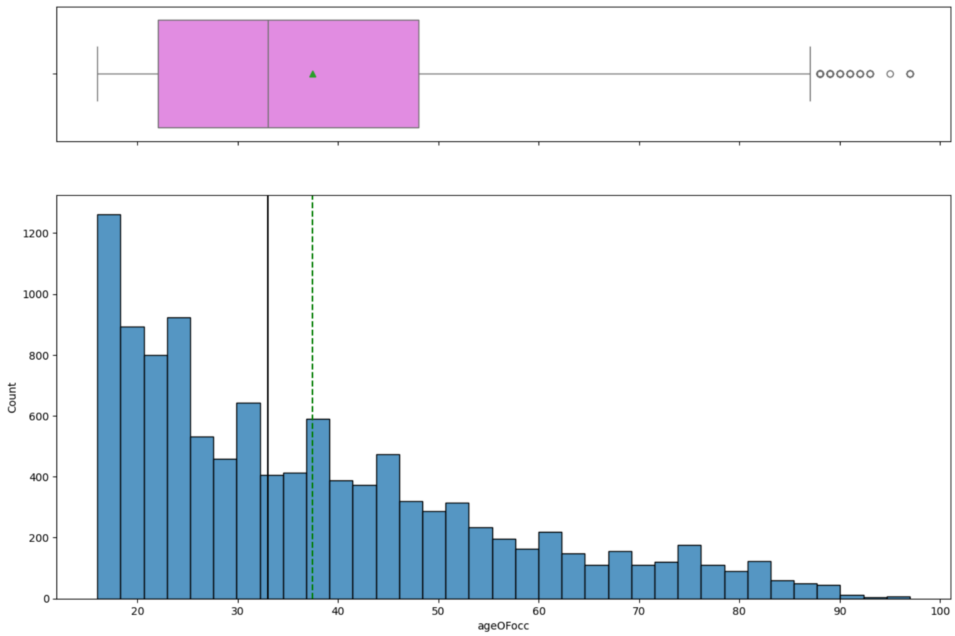
	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
count	11217.00000	11217	11217.00000	11217	11217	11217	11217.00000	11217	11217.00000	11217.00000	11217.00000	11217	11217	11217.00000	11140.00000	11217
unique	NaN	5	NaN	2	2	2	NaN	2	NaN	NaN	NaN	3	2	NaN	NaN	6488
top	NaN	10-24	NaN	survived	airbag	belted	NaN	m	NaN	NaN	NaN	deploy	driver	NaN	NaN	73:100:2
freq	NaN	5414	NaN	10037	7064	7849	NaN	6048	NaN	NaN	NaN	4365	8786	NaN	NaN	7
mean	5608.00000	NaN	431.40531	NaN	NaN	NaN	0.64402	NaN	37.42765	2001.10324	1994.17794	NaN	NaN	0.38914	1.82558	NaN
std	3238.21332	NaN	1406.20294	NaN	NaN	NaN	0.47883	NaN	18.19243	1.05681	5.65870	NaN	NaN	0.48758	1.37854	NaN
min	0.00000	NaN	0.00000	NaN	NaN	NaN	0.00000	NaN	16.00000	1997.00000	1953.00000	NaN	NaN	0.00000	0.00000	NaN
25%	2804.00000	NaN	28.29200	NaN	NaN	NaN	0.00000	NaN	22.00000	2001.00000	1991.00000	NaN	NaN	0.00000	1.00000	NaN
50%	5608.00000	NaN	82.19500	NaN	NaN	NaN	1.00000	NaN	33.00000	2001.00000	1995.00000	NaN	NaN	0.00000	2.00000	NaN
75%	8412.00000	NaN	324.05600	NaN	NaN	NaN	1.00000	NaN	48.00000	2002.00000	1999.00000	NaN	NaN	1.00000	3.00000	NaN
max	11216.00000	NaN	31694.04000	NaN	NaN	NaN	1.00000	NaN	97.00000	2002.00000	2003.00000	NaN	NaN	1.00000	5.00000	NaN

Univariate Analysis:

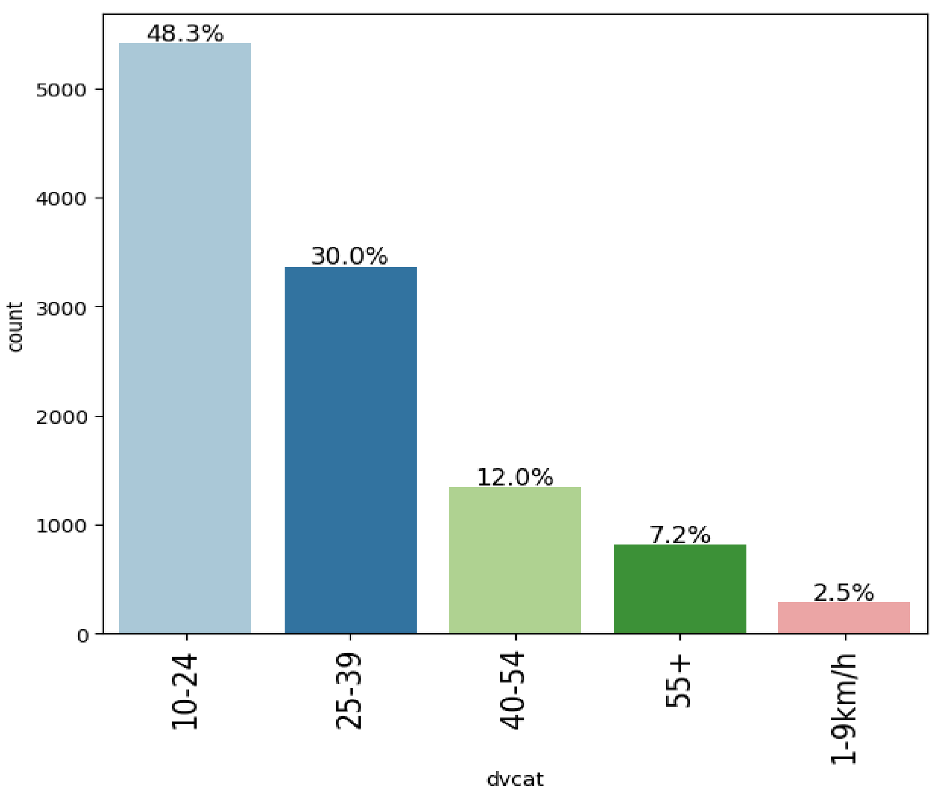
Observations on weight:



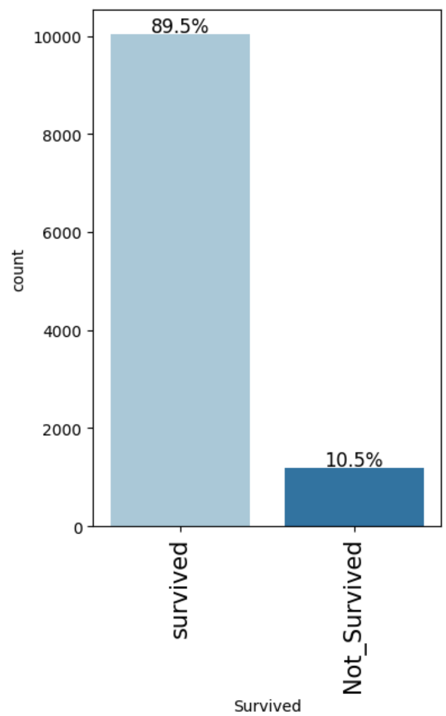
Observations on ageOFocc:



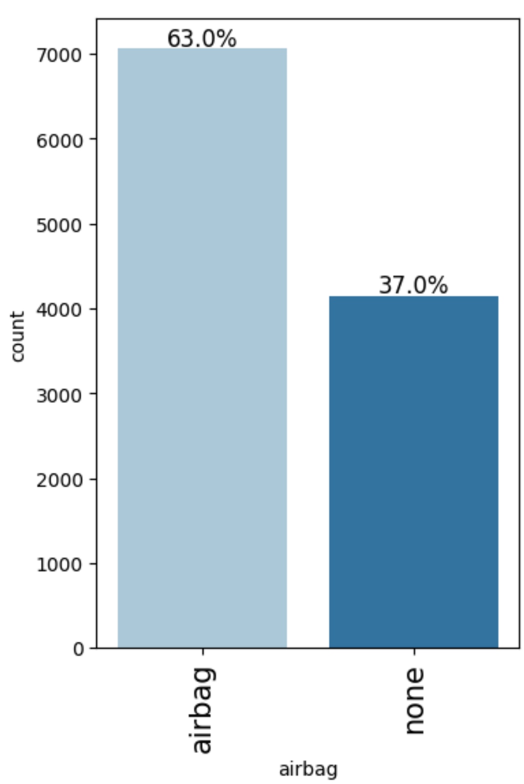
Observations on dvcat:



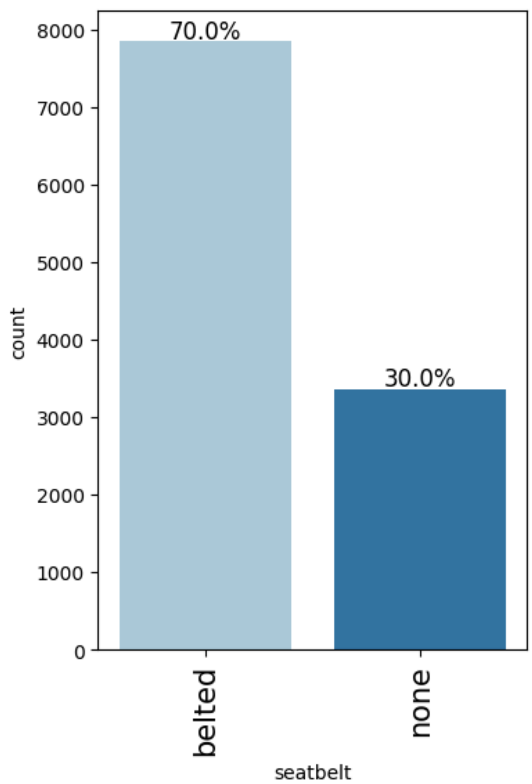
Observations on Survived:



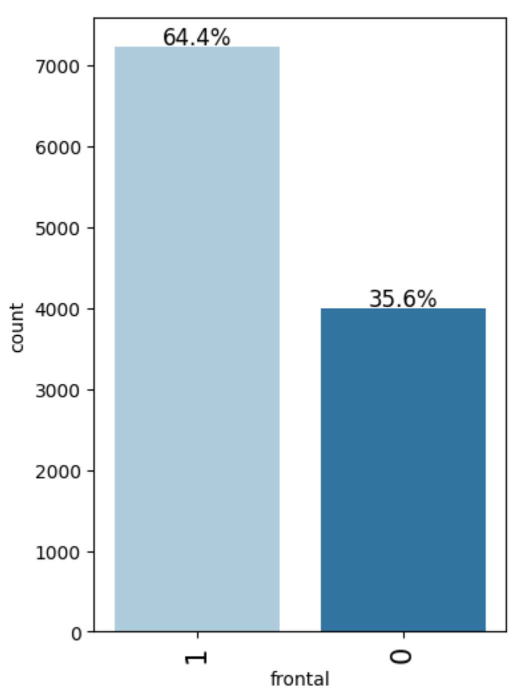
Observations on airbag:



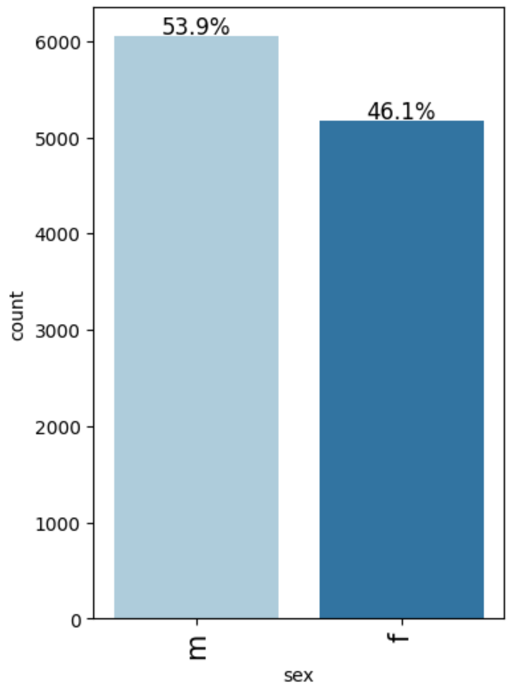
Observations on seatbelt:



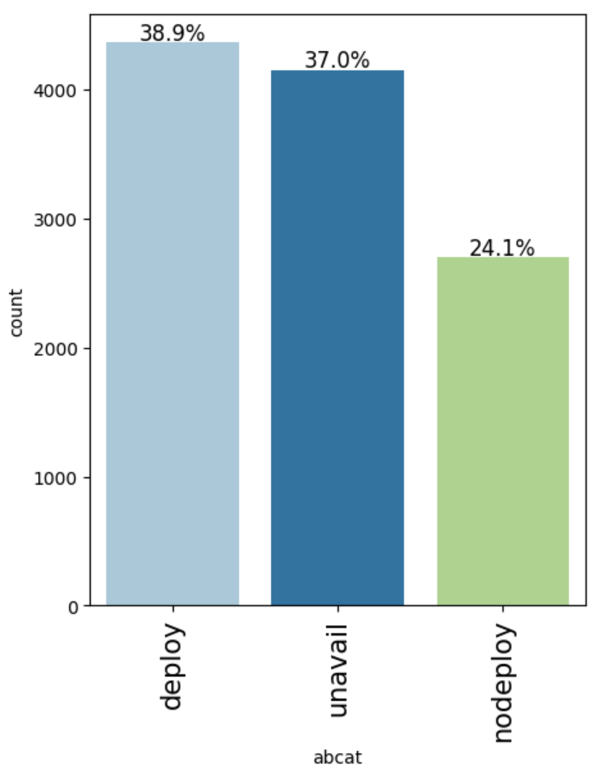
Observations on frontal:



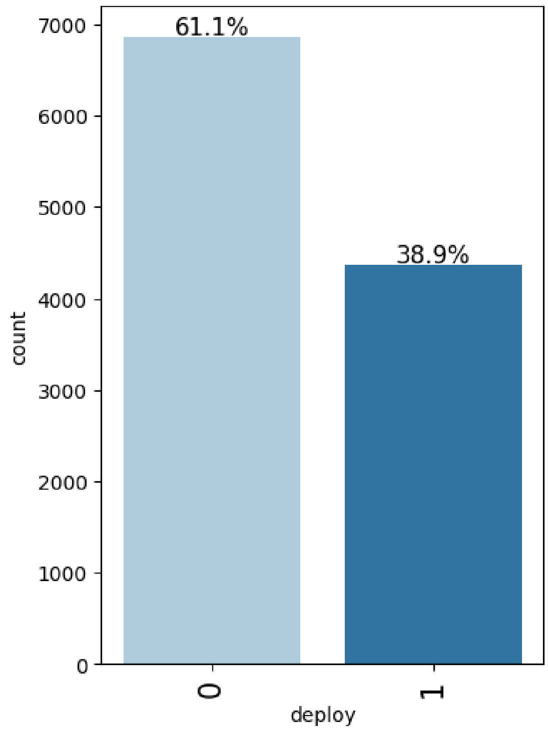
Observations on sex:



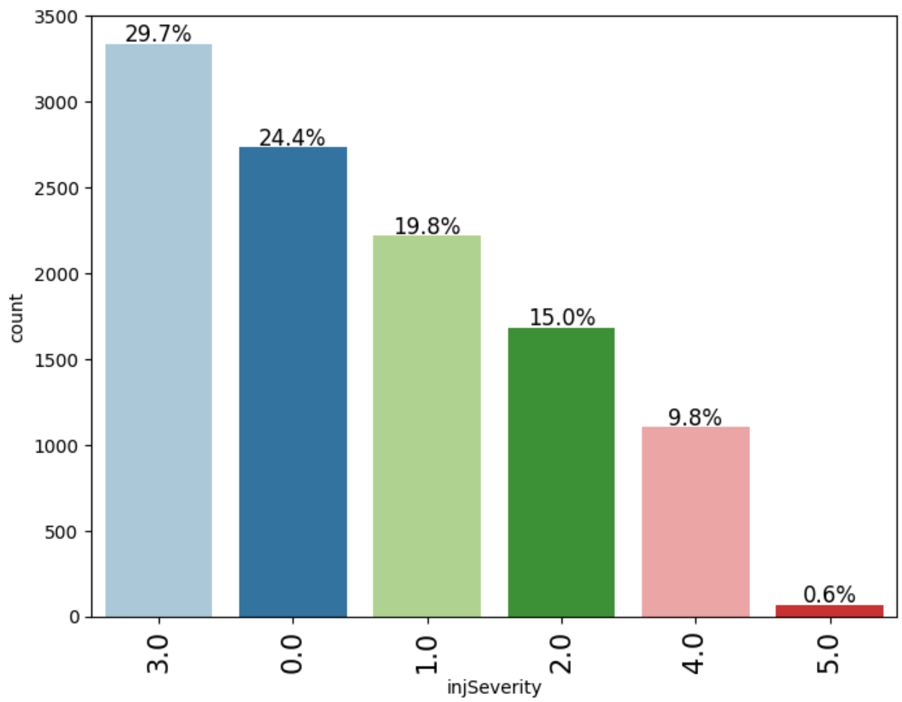
Observations on abcat:



Observations on deploy:

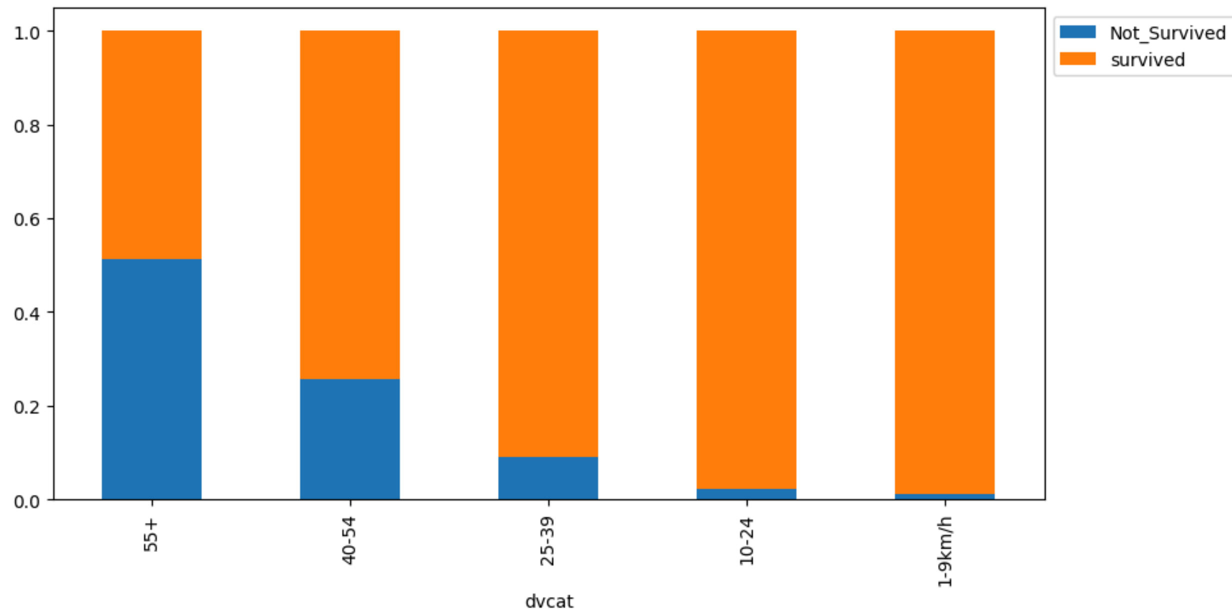


Observations on injSeverity:

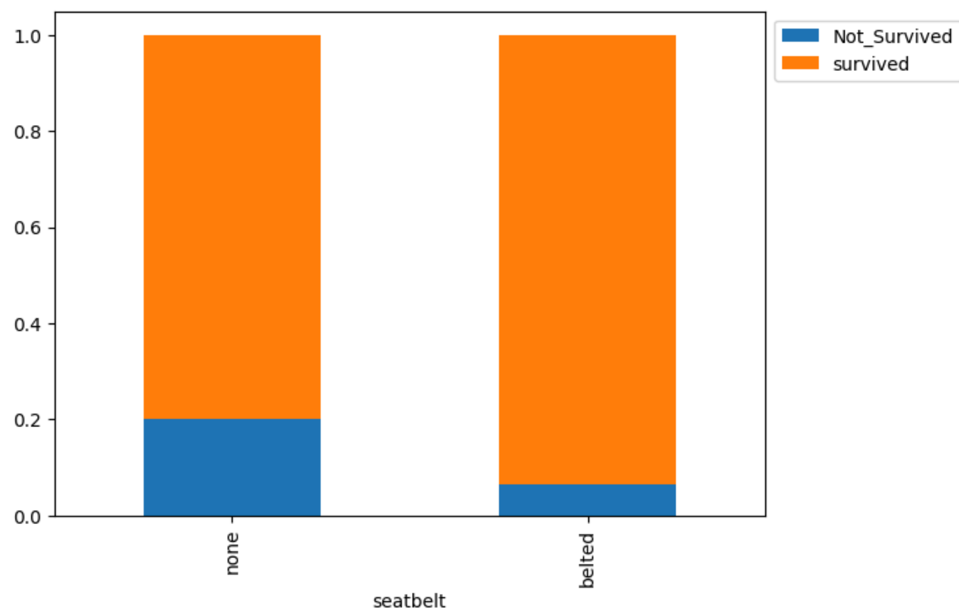


Bivariate Analysis:

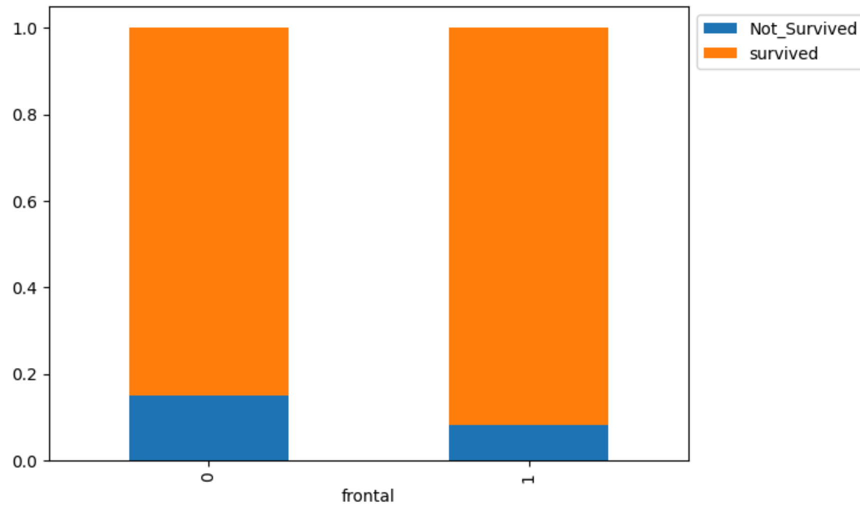
Observation on dvcat and Survived:



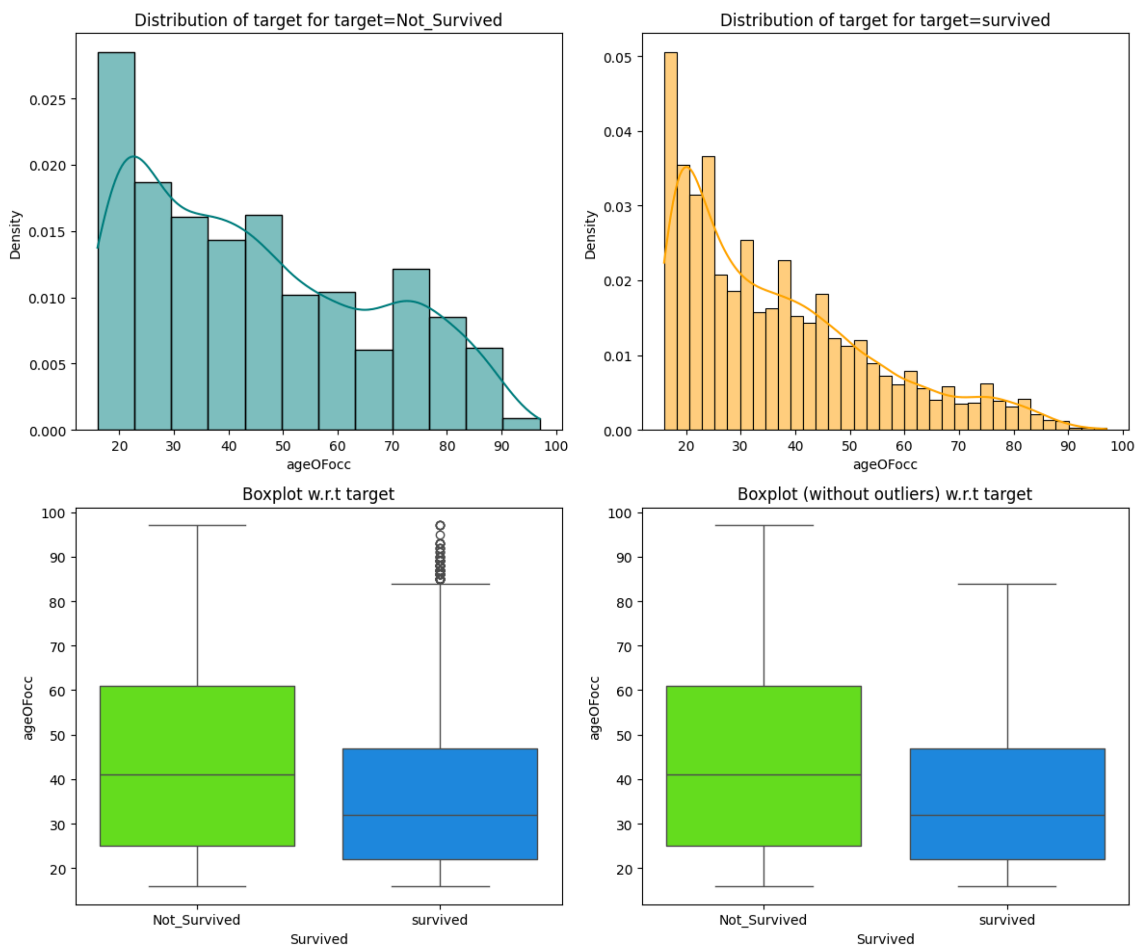
Observation on seatbelt and Survived:



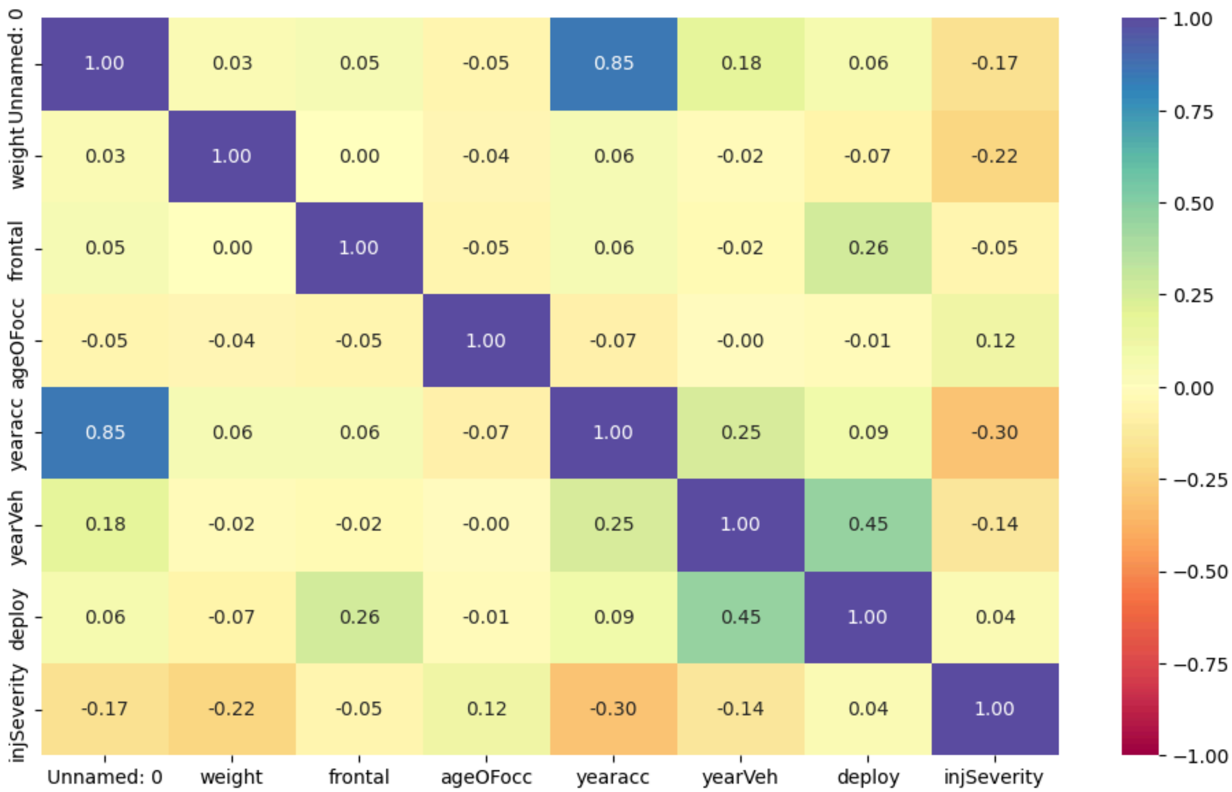
Observation on frontal and Survived:



Distribution of ageOGocc on Survived:



Correlation between numerical variables:



Key meaningful observations on individual variables and the relationship between variables:

- Most vehicle owners are between the ages of 20-30.
- Estimated impact speeds are mostly between 10-24 km/hr.
- 89.5% of people who met with accidents have survived.
- In most cases of accidents airbags were present.
- Frontal impact has been found to be more compared to non-frontal impact.
- 37% airbags were unavailable and 24.1% airbags did not deploy.
- Incapacity is the highest outcome of accidents - 29.7%.
- People who did not survive mostly belong to speed impacts of 40-54 and 55+.
- People who used seatbelts survived more compared to those who did not.
- People with frontal impact survived more than those with non-frontal impact.
- People generally between 20 and 50 years of age survived more.

Problem 2.2 - Data Pre-processing

Prepare the data for modelling: - Missing value Treatment (if needed) - Outlier Detection(treat, if needed) - Drop redundant features (if needed) - Encode the data - Data split

Unnamed: 0	0
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

Values missing in injSeverity. Missing values treated with mode of the variable.

```
data['injSeverity'] = data['injSeverity'].fillna(data['injSeverity'].mode()[0])
```

Outlier Treatment: Outliers have not been treated due to concern of missing important data values.

Feature engineering has been applied. A new column has been created - age of vehicle.

```
data['ageOfVehicle'] = data['yearacc'] - data['yearVeh']
```

Redundant features dropped:

```
data = data.drop(["Unnamed: 0", "caseid"], axis=1)
```

Data split into training and test sets:

```
X = data.drop(["Survived"], axis=1)
y = data["Survived"]
```

Categorical variables label encoded:

```
X['airbag'] = labelencoder.fit_transform(X['airbag'])
X['seatbelt'] = labelencoder.fit_transform(X['seatbelt'])
X['dvcat'] = labelencoder.fit_transform(X['dvcat'])
X['sex'] = labelencoder.fit_transform(X['sex'])
X['abcat'] = labelencoder.fit_transform(X['abcat'])
X['occRole'] = labelencoder.fit_transform(X['occRole'])
```

Problem 2.3 - Model Building and Compare the Performance of the Models

- Build a Logistic Regression model - Build a Linear Discriminant Analysis model - Check Accuracy - Confusion Matrix - Plot ROC curve and get ROC_AUC score - Compare both the models and write inference which model is best/optimized

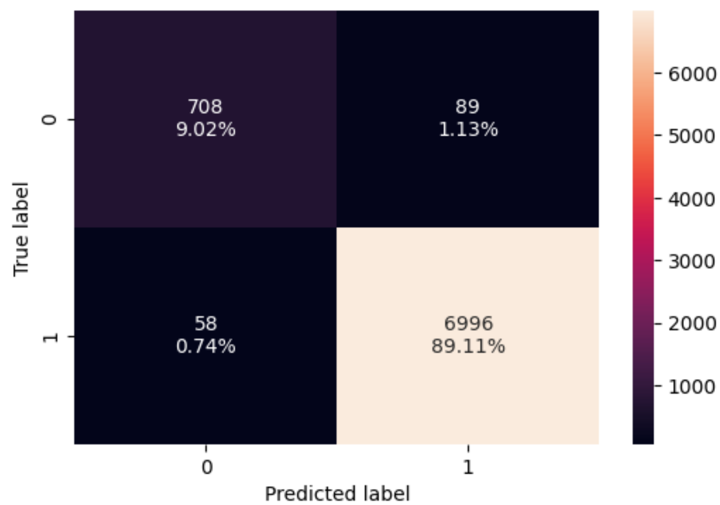
Solution:

Logistic Regression model built with LogisticRegression from sklearn. The performance of the model is as below:

On training set -

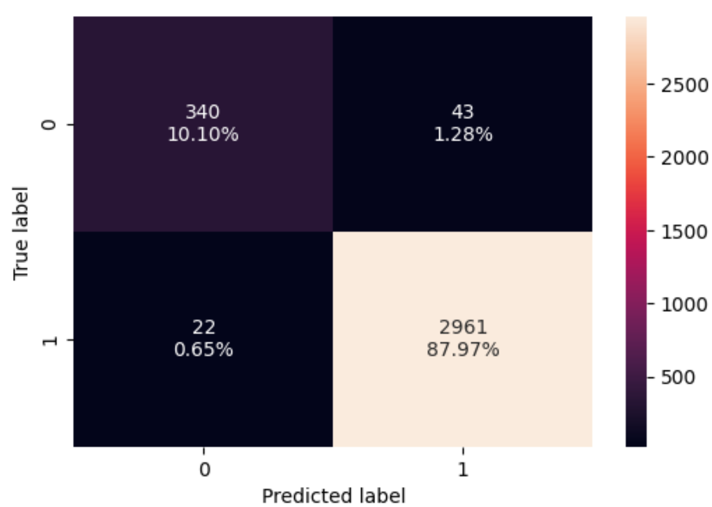
	Accuracy	Recall	Precision	F1
0	0.98128	0.99178	0.98744	0.98960

Confusion matrix:

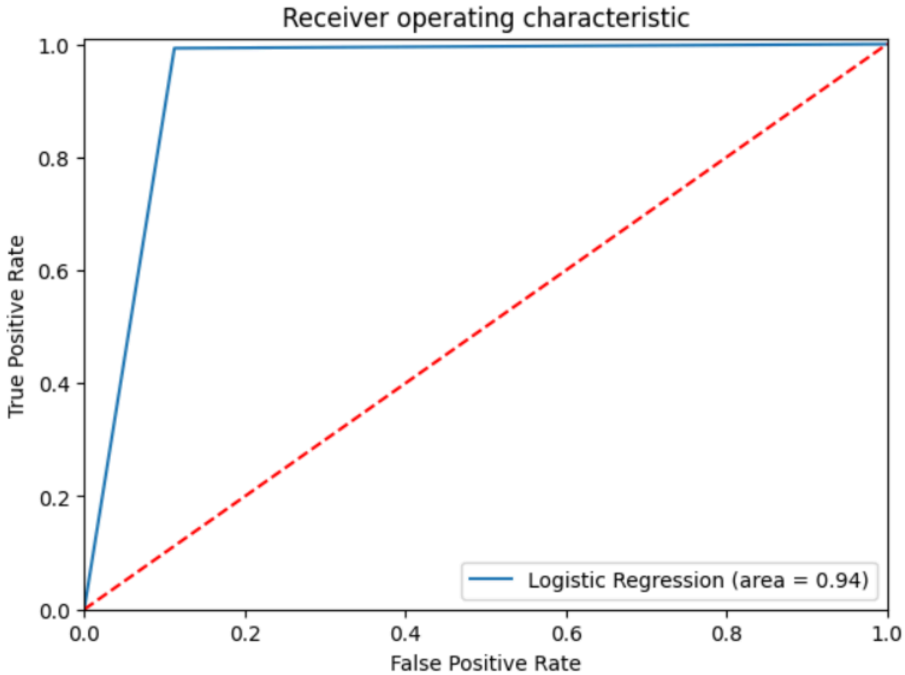


On test set -

	Accuracy	Recall	Precision	F1
0	0.98069	0.99262	0.98569	0.98914



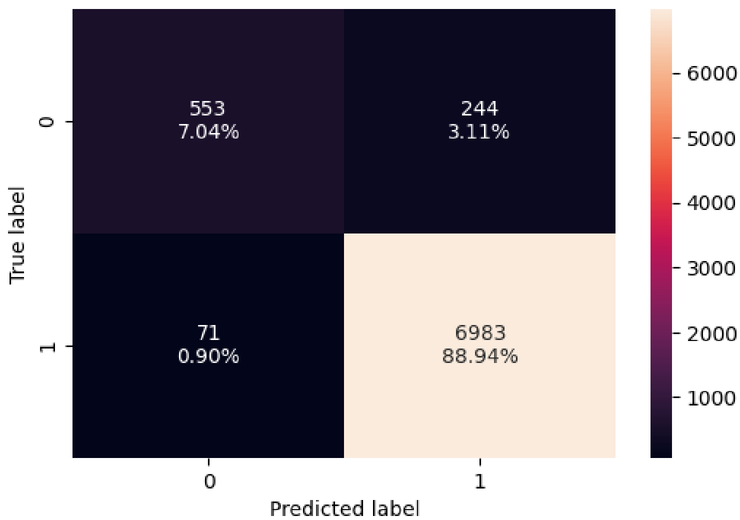
```
Train ROC-AUC score is : 0.9862695698606094

Test ROC-AUC score is : 0.9877976943322868
```



Linear Discriminant analysis is done with LinearDiscriminantAnalysis from sklearn. The performance on training and test sets are as follows:

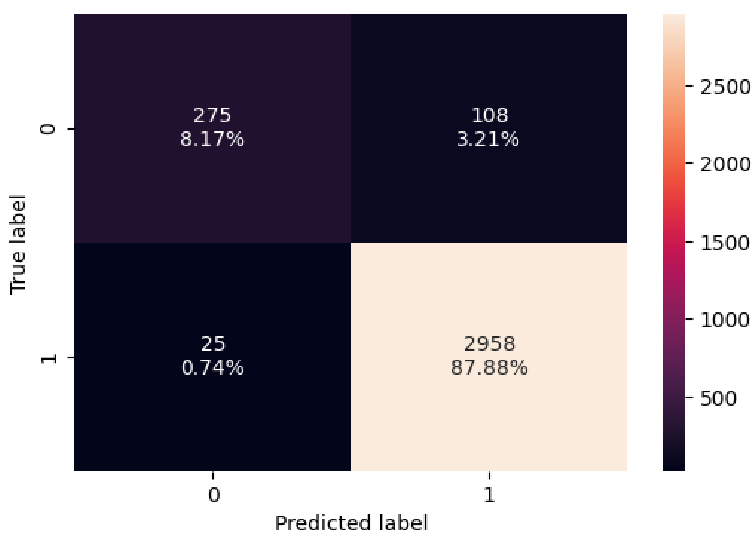
On training set:

	Accuracy	Recall	Precision	F1
0	0.95988	0.98993	0.96624	0.97794

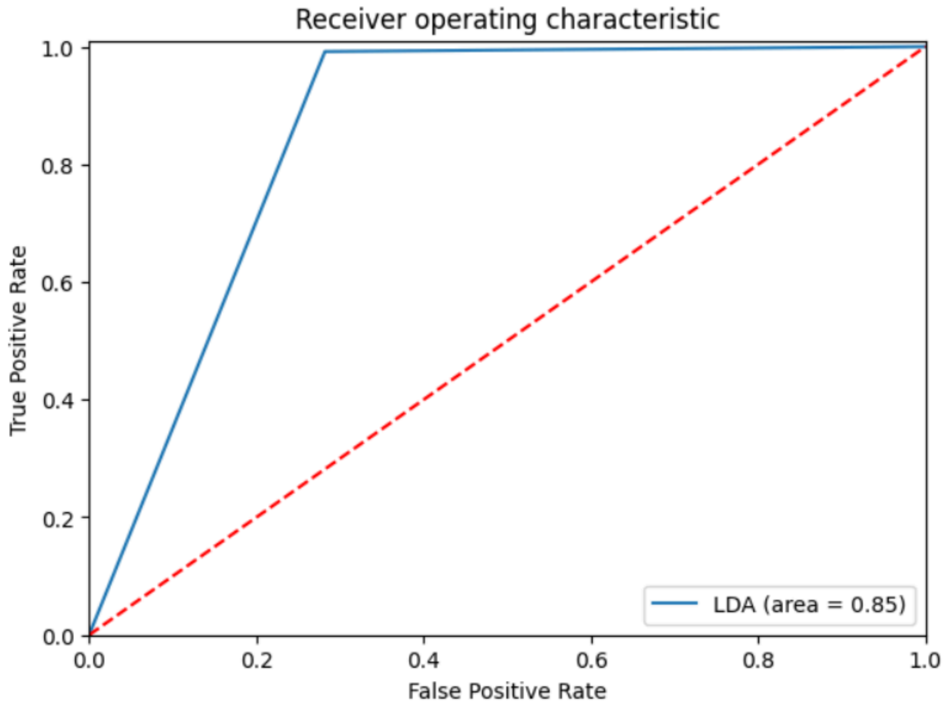


On test set:

	Accuracy	Recall	Precision	F1
0	0.96049	0.99162	0.96477	0.97801



```
Train ROC-AUC score is : 0.9682732489534933

Test ROC-AUC score is : 0.9661668514970385
```



Logistic Regression is the better model based on the provided metrics. It consistently outperforms LDA across all key metrics, including accuracy, precision, F1 score, and ROC-AUC. The higher precision and ROC-AUC scores are particularly significant, as they suggest that Logistic Regression is better at correctly identifying positive cases while minimizing false positives and has a superior overall discriminative ability.

- Higher Precision and F1 Score: These indicate better performance in practical scenarios where the balance between precision and recall is critical.

- Higher ROC-AUC Score: Suggests superior overall performance in distinguishing between the classes across different thresholds, making it more reliable for predictions.

Problem 2.4 - Business Insights & Recommendations

Please explain and summarise the various steps performed in this project. Please provide proper business interpretation (atleast 3) and actionable insights (atleast 3)

Solution:

Summary of the Project Steps:

- Data Collection and Preprocessing - clean the data, handle missing values.
- Feature Selection and Engineering - age of vehicle was created.
- The project involved choosing between different models, specifically Logistic Regression and Linear Discriminant Analysis (LDA). These models were selected based on their suitability for classification tasks.
- Model Training - Both Logistic Regression and LDA models were trained on the training dataset. During this phase, the models learned to predict the target variable based on the features provided.
- Model Evaluation: The performance of the models was evaluated using various metrics such as Accuracy, Recall, Precision, F1 Score, and ROC-AUC scores on both training and test datasets. This step was critical for determining which model performed better and was more reliable for making business predictions.
- Model Selection and Interpretation: Based on the evaluation metrics, Logistic Regression was chosen as the better model due to its superior performance across all key metrics. The model's coefficients and output were interpreted to provide insights into the relationship between features and the target variable.