PREDICTIVE MODELLING

Aniket Ganguly

PGPDSBA – JULY 2023

PROBLEM 1 – Linear Regression

SOLUTION – 1

1. **Define the problem and perform exploratory Data Analysis**

a) **Objective:** Predict the sales of firms based on attributes provided in the dataset to help the investment firm make informed decisions.

b) **Exploratory Data Analysis (EDA)**

- **Top 5 Rows:**

```
   Unnamed: 0        sales      capital  patents        randd  employment  \
0           0   826.995050   161.603986       10   382.078247    2.306000
1           1   407.753973   122.101012        2     0.000000    1.860000
2           2  8407.845588  6221.144614      138  3296.700439   49.659005
3           3   451.000010   266.899987        1    83.540161    3.071000
4           4   174.927981   140.124004        2    14.233637    1.947000

  sp500     tobinq         value  institutions
0    no  11.049511   1625.453755         80.27
1    no   0.844187    243.117082         59.02
2   yes   5.205257  25865.233800         47.70
3    no   0.305221     63.024630         26.88
4    no   1.063300     67.406408         49.46
```

- **Shape and Datatypes of the dataset:**

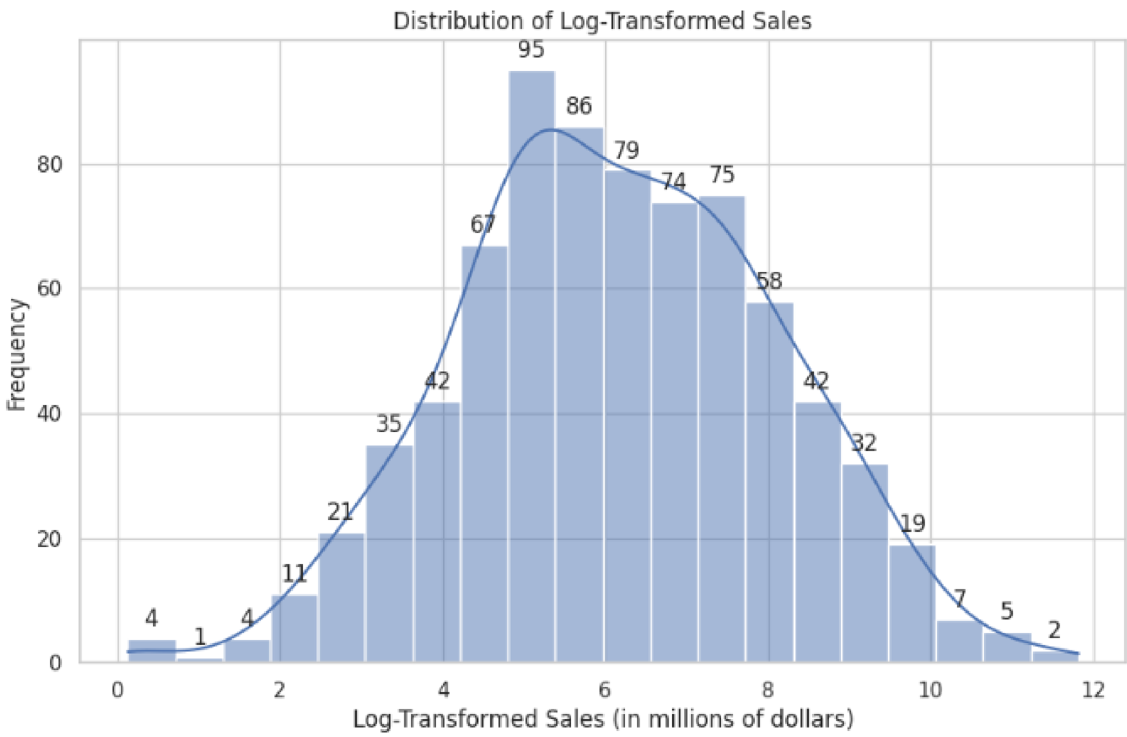
```
Shape of the dataset: (759, 10)
Unnamed: 0        int64
sales           float64
capital         float64
patents           int64
randd           float64
employment      float64
sp500            object
tobinq          float64
value           float64
institutions    float64
dtype: object
```

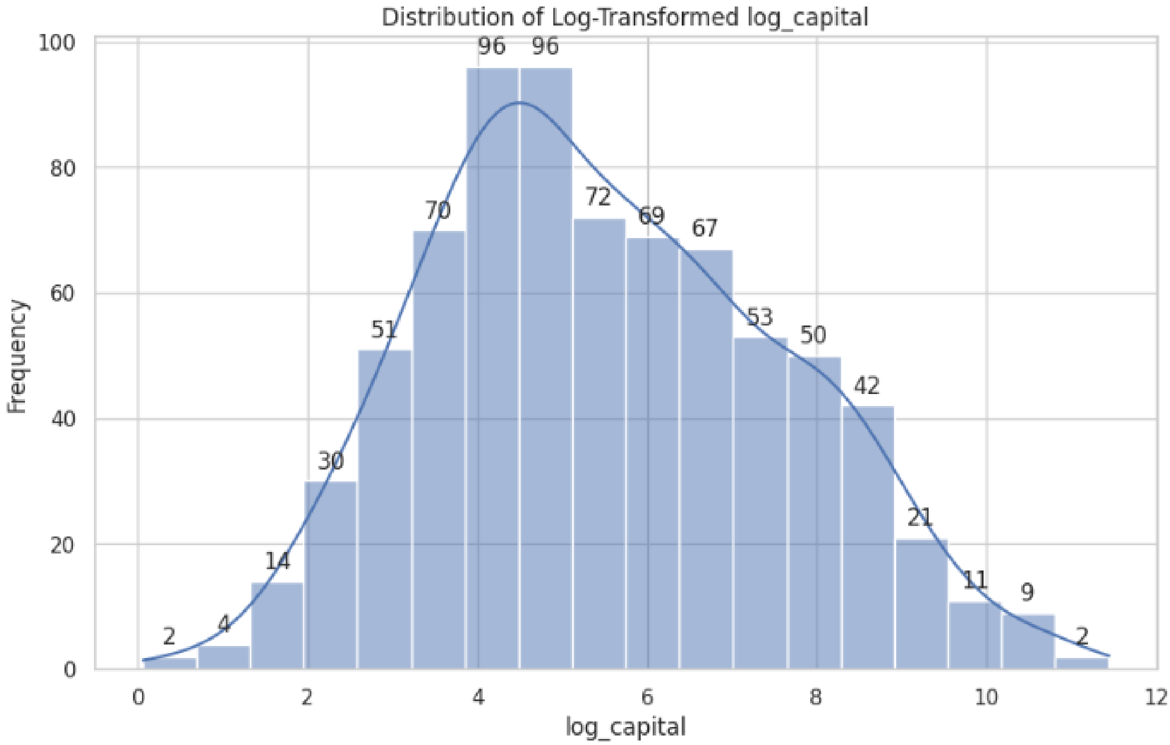
- **Statistical summary of the numerical columns in the dataset:**

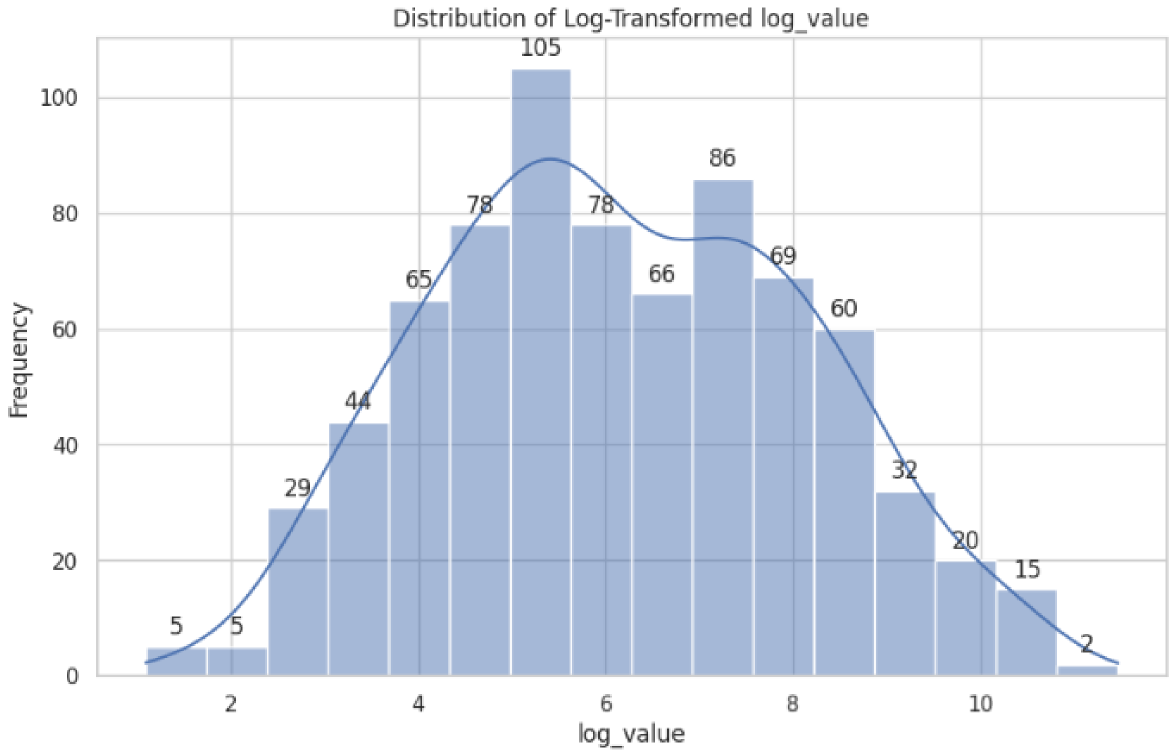
	Unnamed: 0	sales	capital	patents	randd
count	759.000000	759.000000	759.000000	759.000000	759.000000
mean	379.000000	2689.705158	1977.747498	25.831357	439.938074
std	219.248717	8722.060124	6466.704896	97.259577	2007.397588
min	0.000000	0.138000	0.057000	0.000000	0.000000
25%	189.500000	122.920000	52.650501	1.000000	4.628262
50%	379.000000	448.577082	202.179023	3.000000	36.864136
75%	568.500000	1822.547366	1075.790020	11.500000	143.253403
max	758.000000	135696.788200	93625.200560	1220.000000	30425.255860

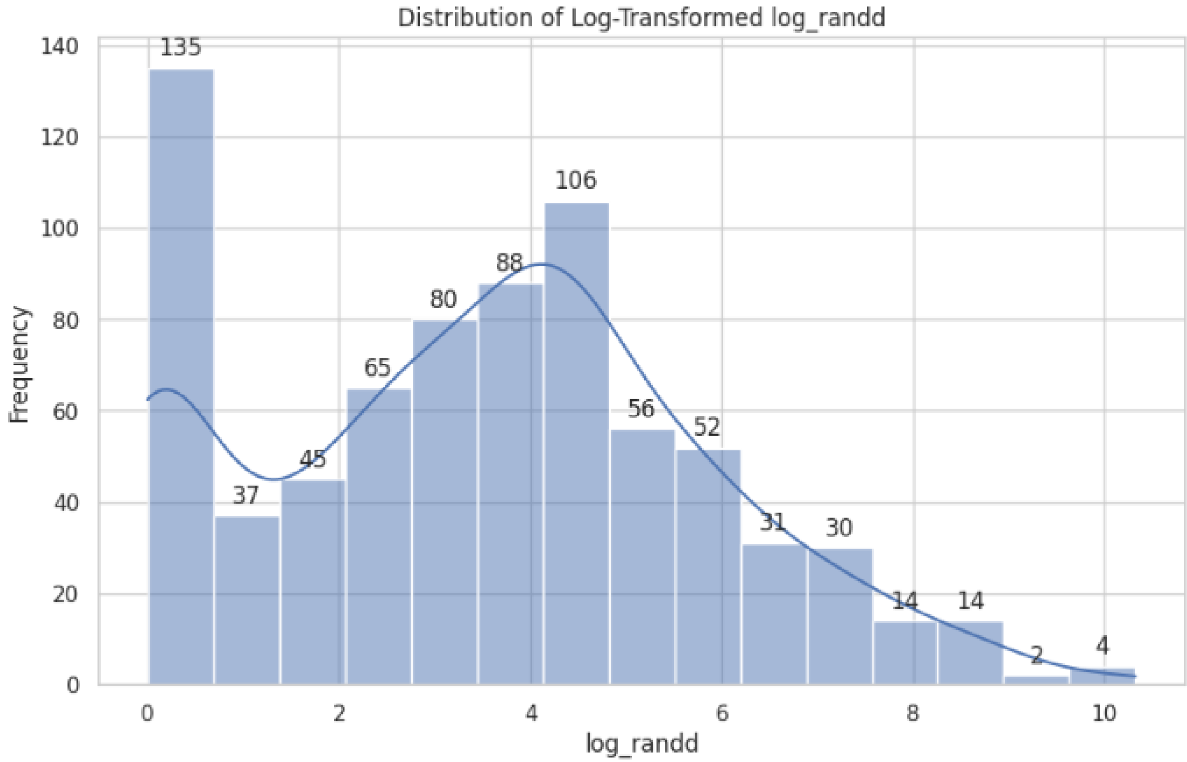
	employment	tobinq	value	institutions
count	759.000000	738.000000	759.000000	759.000000
mean	14.164519	2.794910	2732.734750	43.020540
std	43.321443	3.366591	7071.072362	21.685586
min	0.006000	0.119001	1.971053	0.000000
25%	0.927500	1.018783	103.593946	25.395000
50%	2.924000	1.680303	410.793529	44.110000
75%	10.050001	3.139309	2054.160386	60.510000
max	710.799925	20.000000	95191.591160	90.150000

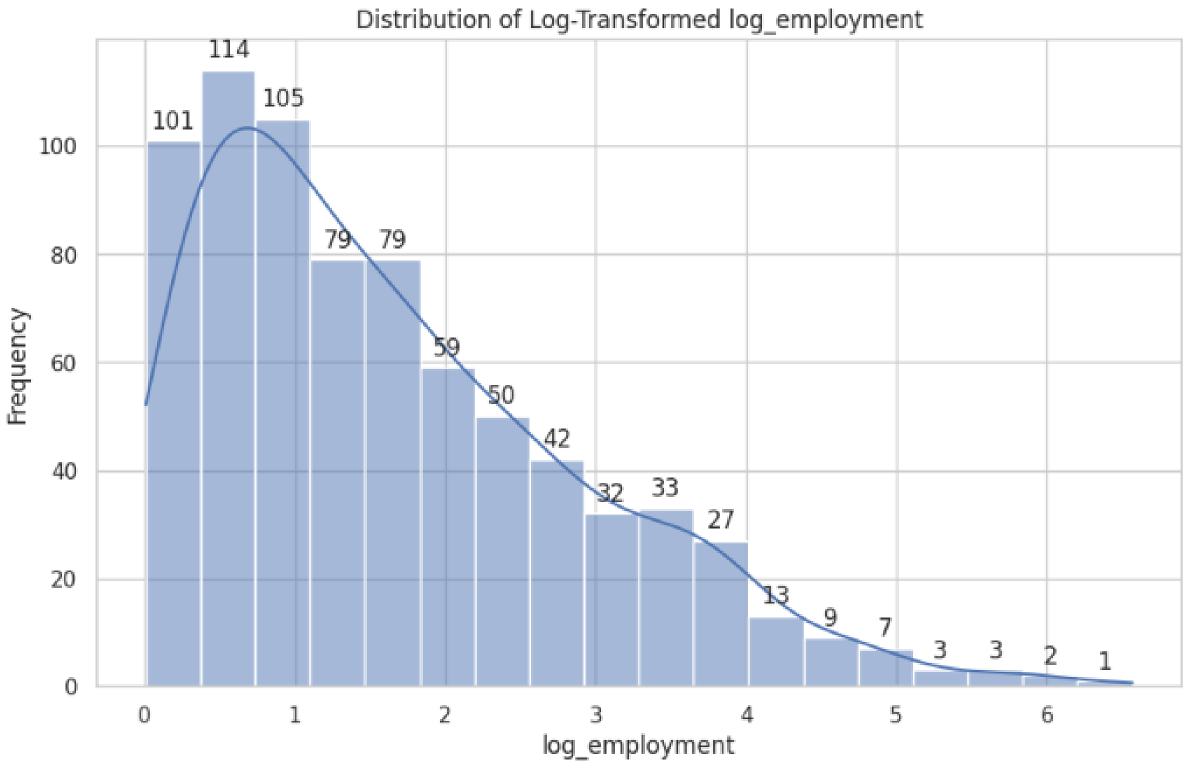
- **Univariate Analysis**

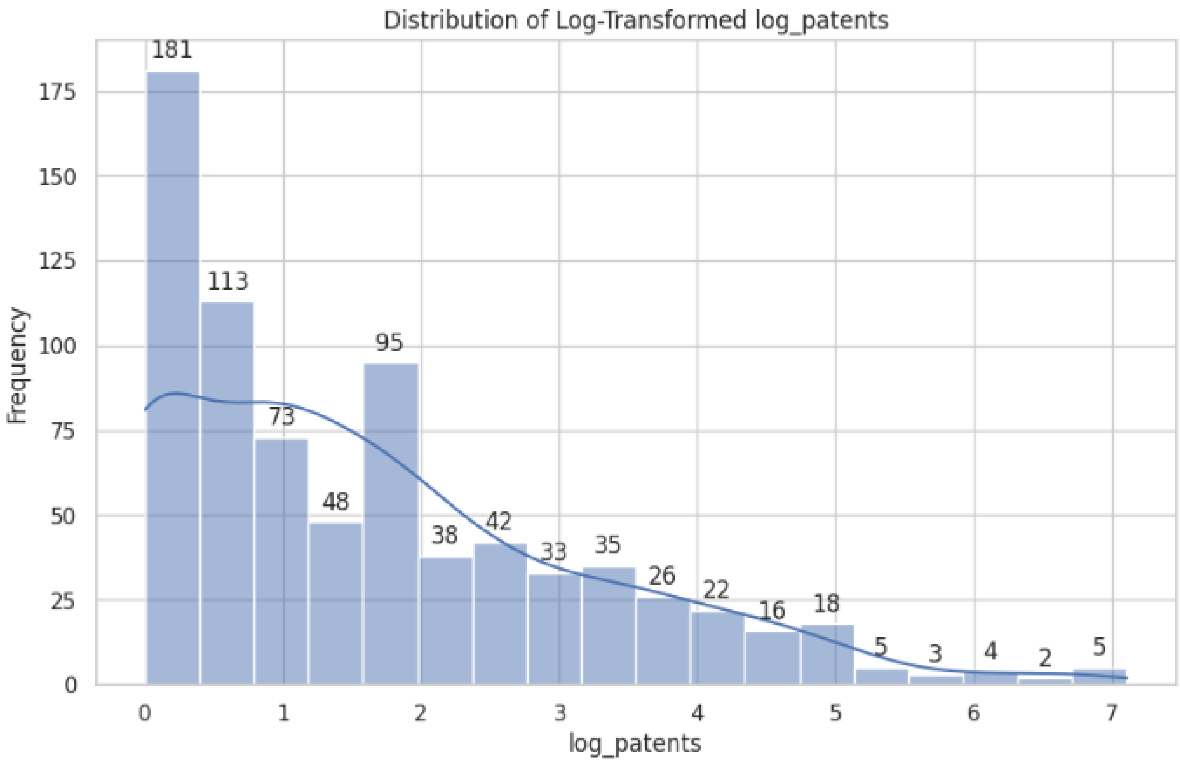


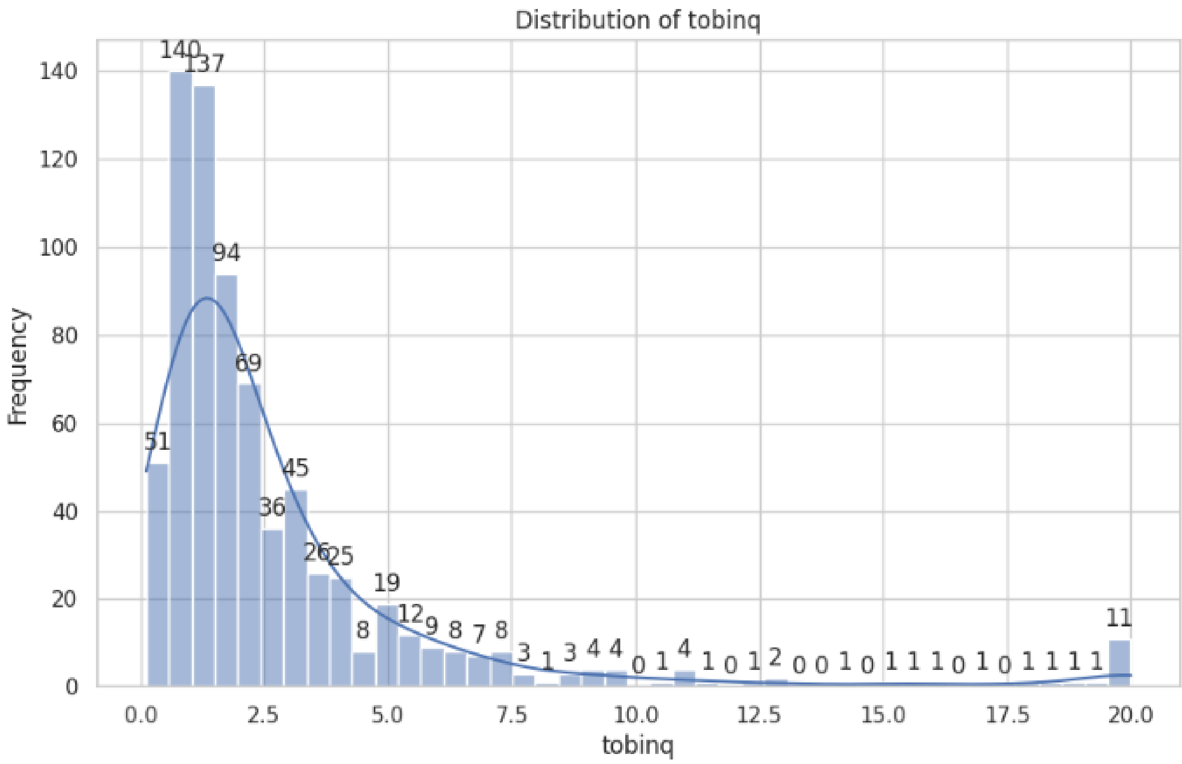

Distribution of Log-Transformed log_capital
Frequency
log_capital
96
96
72
70
69
67
53
51
50
42
30
21
14
11
9
4
2
2

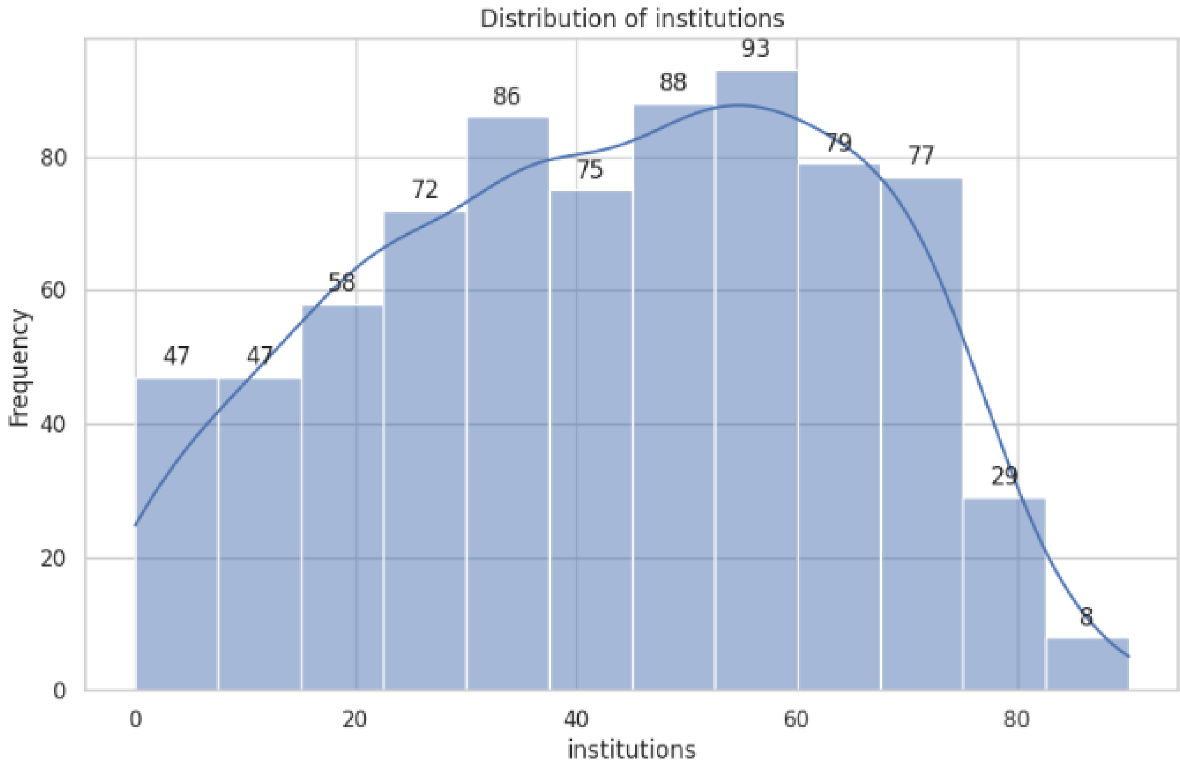

Distribution of Log-Transformed log_value
Frequency
log_value
105
86
78
78
69
66
65
60
44
32
29
20
15
5
5
2

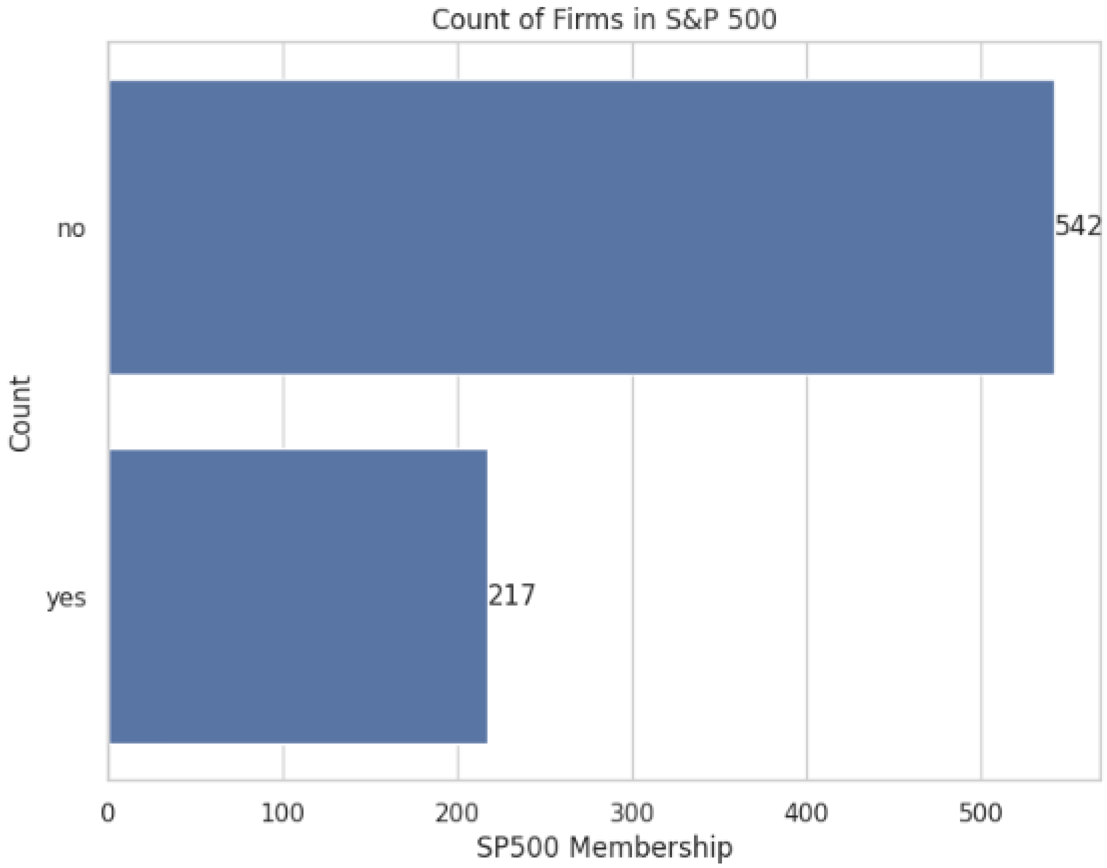

Distribution of Log-Transformed log_randd
Frequency
log_randd
135
37
45
65
80
88
106
56
52
31
30
14
14
2
4
0
20
40
60
80
100
120
140
0
2
4
6
8
10


Distribution of Log-Transformed log_employment
Frequency
log_employment
101
114
105
79
79
59
50
42
32
33
27
13
9
7
3
3
2
1
0
20
40
60
80
100
0
1
2
3
4
5
6

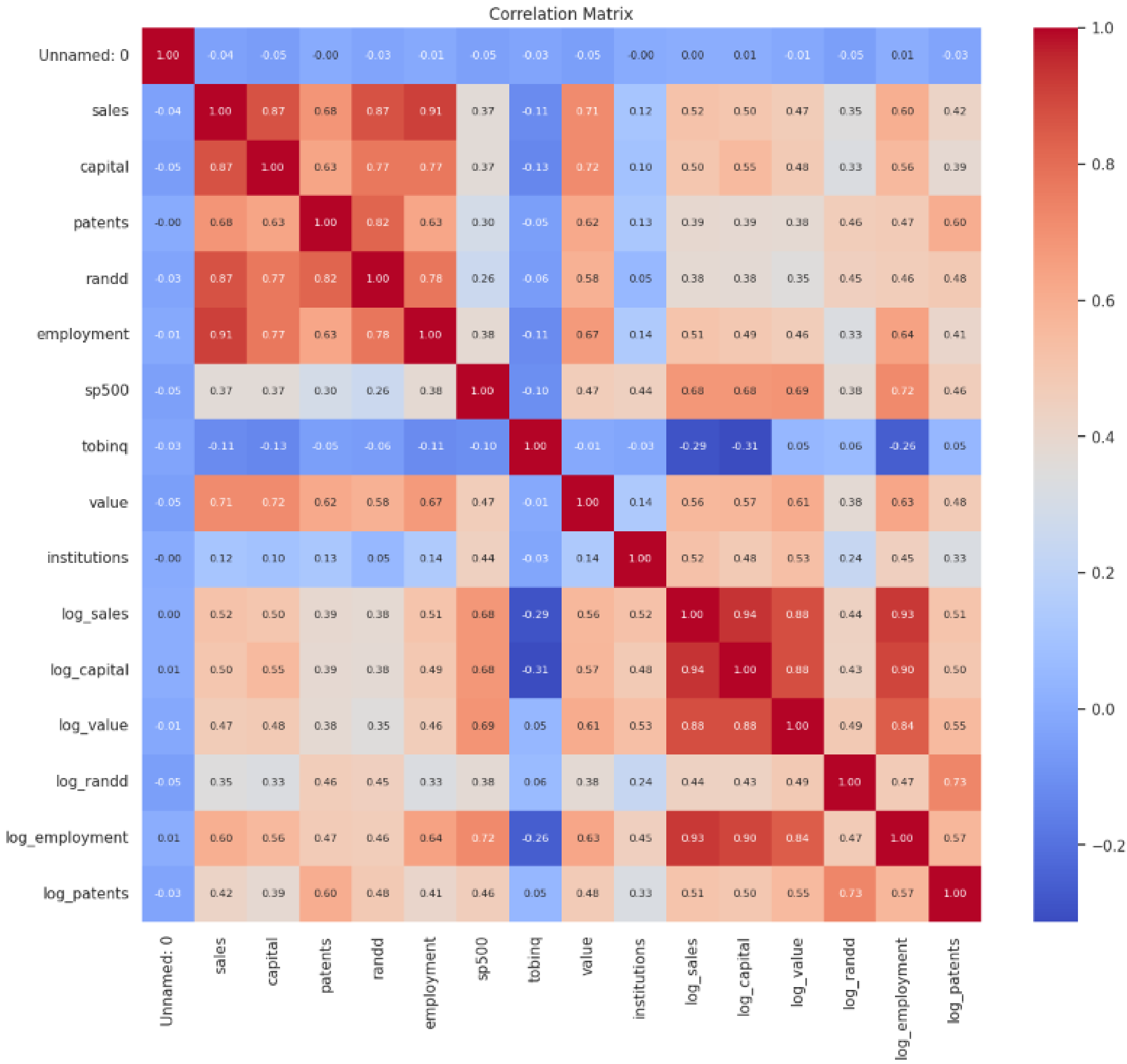

Distribution of Log-Transformed log_patents
Frequency
log_patents
181
113
73
48
95
38
42
33
35
26
22
16
18
5
3
4
2
5


Distribution of tobinq
Frequency
tobinq
140
137
94
69
51
45
36
26
25
19
12
11


Distribution of institutions
Frequency
institutions
0
20
40
60
80
47
47
58
72
86
75
88
93
79
77
29
8


Count of Firms in S&P 500
Count
SP500 Membership
no
yes
542
217
0
100
200
300
400
500

- **Multivariate Analysis**



- Here are key observations on individual variables and their relationships based on the data analysis and linear regression results:
- **Sales (Target Variable):**
- Sales values are highly skewed, with most firms having lower sales and a few firms with very high sales.
- Log transformation of sales helps in normalizing the data for better model performance.
- **Capital:**
- Positive relationship with sales, indicating that firms with higher capital tend to have higher sales.
- Capital remains significant in the final model, showing its strong impact on sales.

- **Patents:**
- Positive relationship with sales, suggesting that firms with more patents tend to generate higher sales.
- Patents are significant in the final model, emphasizing the importance of innovation and intellectual property.
- **R&D (Research and Development):**
- Initially shows a negative relationship with sales in the model, which may seem counterintuitive.
- This negative coefficient could indicate that higher R&D spending doesn't immediately translate into sales but could have a lagged effect.
- Employment:
- Negative relationship with sales, meaning firms with higher employment might have lower sales.
- This could be due to inefficiencies or the nature of the industries with higher employment.
- **SP500 Membership:**
- Positive relationship with sales, indicating that being part of the S&P 500 index is associated with higher sales.
- This could be due to higher visibility, credibility, and investor confidence in these firms.
- **Tobin's Q:**
- Positive relationship with sales, suggesting that firms with higher market value relative to their asset replacement costs tend to have higher sales.
- This might reflect the market's favorable perception and growth potential of such firms.
- **Value:**
- Negative relationship with sales, which might seem counterintuitive.
- This could be due to multicollinearity or specific industry characteristics where higher market value doesn't directly correlate with current sales.
- Institutions:
- Not consistently significant in the final models, indicating that institutional ownership proportion might not have a strong direct impact on sales.
- **Log-transformed Variables (log_sales, log_capital, log_value, log_randd, log_employment, log_patents):**
- Log transformations help in dealing with skewed distributions and make the relationships more linear.
- Some log-transformed variables remain significant, highlighting their importance after transformation.
- **Multicollinearity:**
- The initial high condition number suggests the presence of multicollinearity among the predictors.
- Variables like capital, patents, and R&D show strong correlations, which could affect the model's stability.

- **Correlation Matrix Insights:**
- High correlations are observed between:
- Sales and Capital (0.87)
- Sales and Employment (0.77)
- Capital and Employment (0.63)
- Patents and R&D (0.82)
- These correlations indicate that firms with higher capital tend to employ more people, and firms with more patents also invest significantly in R&D.

2. Data Pre-processing

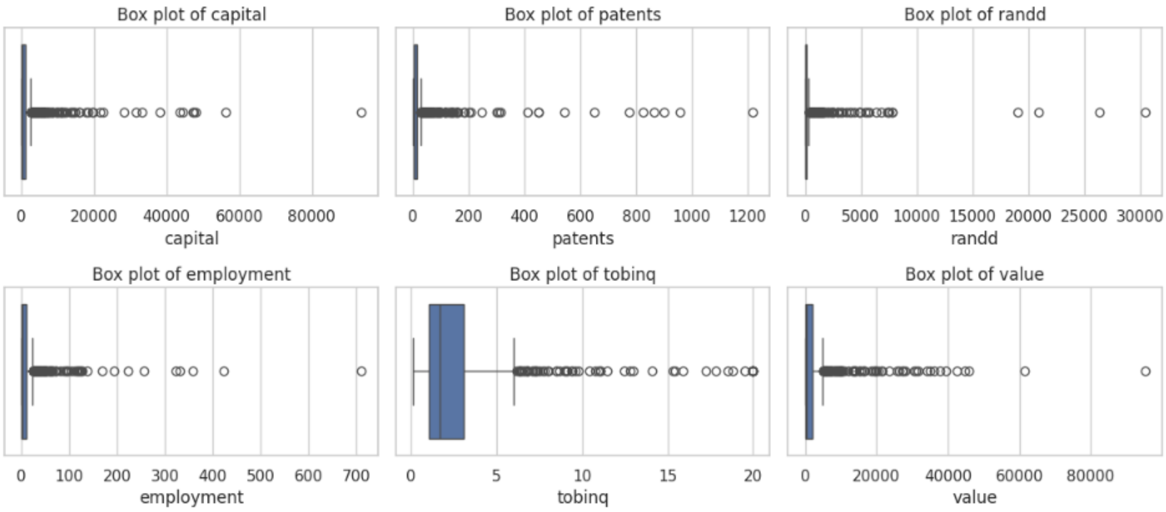
a) Missing Values

```
Missing Values:
 Unnamed: 0        0
sales             0
capital           0
patents           0
randd             0
employment        0
sp500             0
tobinq           21
value             0
institutions      0
log_sales         0
log_capital       0
log_value         0
log_randd         0
log_employment    0
log_patents       0
dtype: int64
```

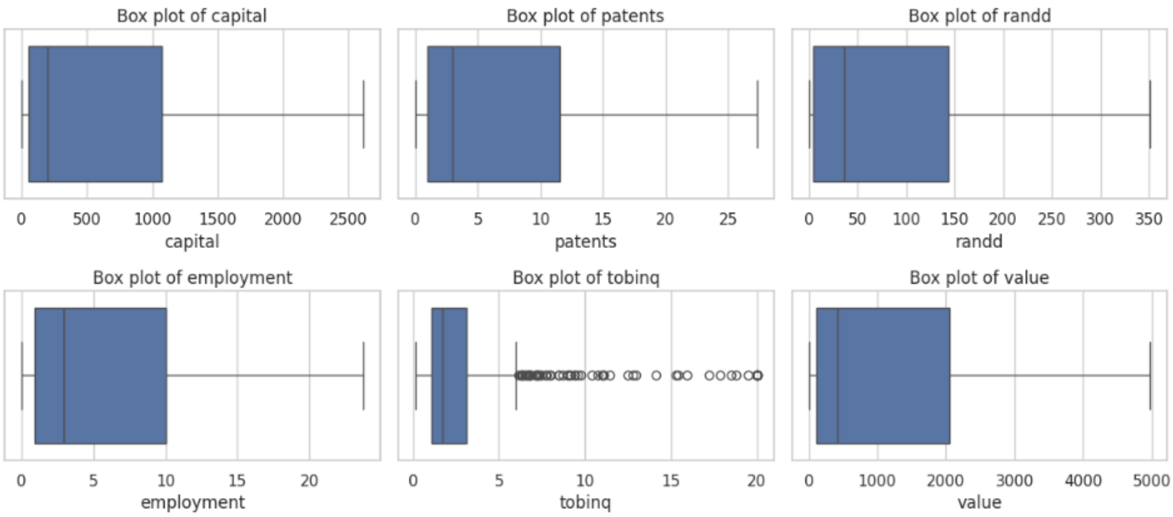
- Missing values identified for the column "tobinq". Since it is a numerical column we have used the median imputation to fill the missing values.

```
Missing Values:
 Unnamed: 0        0
sales             0
capital           0
patents           0
randd             0
employment        0
sp500             0
tobinq            0
value             0
institutions      0
log_sales         0
log_capital       0
log_value         0
log_randd         0
log_employment    0
log_patents       0
dtype: int64
```

b) Outlier Treatment



- Outlier detected for the above attached columns, outliers were treated by capping them to the lower and upper bounds.



- Encoded 'sp500' column: 'no' -> 0, 'yes' -> 1. Only 1 categorical column in the dataset.

3. Model Building - Linear regression

- **Linear Regression applied.**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.521
Model:                            OLS   Adj. R-squared:                  0.510
Method:                 Least Squares   F-statistic:                     46.08
Date:                Mon, 15 Jul 2024   Prob (F-statistic):           3.67e-85
Time:                        12:42:04   Log-Likelihood:                -6164.9
No. Observations:                 607   AIC:                         1.236e+04
Df Residuals:                     592   BIC:                         1.243e+04
Df Model:                          14
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           3178.4766   1861.039      1.708      0.088    -476.566    6833.519
capital            2.9229      0.818      3.572      0.000       1.316       4.530
patents         -286.6135     81.619     -3.512      0.000    -446.912    -126.315
randd            -14.1639      5.402     -2.622      0.009     -24.772      -3.555
employment     -1275.0620    126.732    -10.061      0.000   -1523.961   -1026.163
sp500           -126.0447    983.016     -0.128      0.898   -2056.667    1804.578
tobinq           236.3539    123.980      1.906      0.057      -7.141     479.849
value              1.5716      0.427      3.678      0.000       0.732       2.411
institutions     -55.5366     14.910     -3.725      0.000     -84.819     -26.254
log_sales       -451.1193    513.895     -0.878      0.380   -1460.398     558.159
log_capital    -1247.5691    532.632     -2.342      0.019   -2293.648    -201.490
log_value      -1549.1047    504.849     -3.068      0.002   -2540.618    -557.591
log_randd        767.3956    252.658      3.037      0.002     271.180    1263.612
log_employment  1.345e+04   1045.046     12.869      0.000    1.14e+04    1.55e+04
log_patents     2300.6907    527.620      4.361      0.000    1264.456    3336.925
==============================================================================
Omnibus:                      852.048   Durbin-Watson:                   2.101
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           212851.047
Skew:                           7.327   Prob(JB):                         0.00
Kurtosis:                      93.560   Cond. No.                     1.94e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.94e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- **Iteration 1** of dropping insignificant variables.

```
Iteration 1
Dropped variable: sp500 with p-value: 0.8980164562910298
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.521
Model:                            OLS   Adj. R-squared:                  0.511
Method:                 Least Squares   F-statistic:                     49.71
Date:                Mon, 15 Jul 2024   Prob (F-statistic):           5.13e-86
Time:                        13:06:29   Log-Likelihood:                -6165.0
No. Observations:                 607   AIC:                         1.236e+04
Df Residuals:                     593   BIC:                         1.242e+04
Df Model:                          13                                         
Covariance Type:            nonrobust                                         
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           3176.6558   1859.441      1.708      0.088    -475.235    6828.547
capital            2.9189      0.817      3.573      0.000       1.314       4.523
patents         -286.8531     81.530     -3.518      0.000    -446.976    -126.730
randd            -14.2168      5.381     -2.642      0.008     -24.786      -3.648
employment     -1277.6705    124.985    -10.223      0.000   -1523.137   -1032.204
tobinq           237.4223    123.598      1.921      0.055      -5.320     480.164
value              1.5554      0.408      3.813      0.000       0.754       2.357
institutions     -55.9176     14.599     -3.830      0.000     -84.589     -27.246
log_sales       -451.9029    513.432     -0.880      0.379   -1460.269     556.463
log_capital    -1244.4894    531.649     -2.341      0.020   -2288.633    -200.345
log_value      -1549.5512    504.418     -3.072      0.002   -2540.215    -558.887
log_randd        769.4660    251.933      3.054      0.002     274.677    1264.255
log_employment  1.346e+04   1042.315     12.910      0.000    1.14e+04    1.55e+04
log_patents     2301.3290    527.159      4.366      0.000    1266.004    3336.654
==============================================================================
Omnibus:                      851.747   Durbin-Watson:                   2.101
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           212363.836
Skew:                           7.323   Prob(JB):                         0.00
Kurtosis:                      93.455   Cond. No.                     1.94e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.94e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- **Iteration 2 of dropping insignificant variables**

```
Iteration 2
Dropped variable: log_sales with p-value: 0.3791284211999145
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.521
Model:                            OLS   Adj. R-squared:                  0.511
Method:                 Least Squares   F-statistic:                     53.81
Date:                Mon, 15 Jul 2024   Prob (F-statistic):           1.00e-86
Time:                        13:06:29   Log-Likelihood:                -6165.4
No. Observations:                 607   AIC:                         1.236e+04
Df Residuals:                     594   BIC:                         1.241e+04
Df Model:                          12
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           2431.3536   1655.127      1.469      0.142    -819.260    5681.967
capital            3.0546      0.802      3.808      0.000       1.479       4.630
patents         -290.2785     81.422     -3.565      0.000    -450.188    -130.368
randd            -14.3833      5.377     -2.675      0.008     -24.944      -3.823
employment     -1261.3684    123.581    -10.207      0.000   -1504.077   -1018.659
tobinq           263.4223    119.993      2.195      0.029      27.761     499.084
value              1.5295      0.407      3.760      0.000       0.731       2.328
institutions     -57.4007     14.498     -3.959      0.000     -85.875     -28.927
log_capital    -1404.8304    499.369     -2.813      0.005   -2385.574    -424.087
log_value      -1671.6300    484.882     -3.447      0.001   -2623.922    -719.338
log_randd        780.8963    251.550      3.104      0.002     286.860    1274.932
log_employment  1.308e+04    949.192     13.778      0.000    1.12e+04    1.49e+04
log_patents     2329.7634    526.068      4.429      0.000    1296.584    3362.943
==============================================================================
Omnibus:                      854.129   Durbin-Watson:                   2.102
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           215210.532
Skew:                           7.359   Prob(JB):                         0.00
Kurtosis:                      94.063   Cond. No.                     1.71e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.71e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- **Iteration 3**

```
Iteration 3
No more insignificant variables to drop.
Train R-squared: 0.520836359814914З
Train RMSE: 6236.353587358418
Test R-squared: 0.5298963480725087
Test RMSE: 5092.03978249325
```

4. Business Insights & Recommendations

Project Summary and Steps Performed

a) Problem Definition:

- The goal was to predict the sales of 759 firms based on various attributes provided in the dataset. This information helps the investment firm make informed investment decisions.

b) Exploratory Data Analysis (EDA):

- **Shape and Data Types:** We checked the dataset's dimensions and data types to understand the structure.
- **Statistical Summary:** We generated summary statistics to get an overview of the data distribution.
- **Univariate Analysis:** We analyzed the distribution of each variable using histograms and bar plots.
- **Multivariate Analysis:** We explored relationships between variables using scatter plots and correlation matrices.
- **Log Transformation:** Applied log transformation to handle skewness in several variables, such as sales, capital, R&D, employment, and patents.

c) Data Pre-processing:

- **Missing Value Treatment:** Handled missing values, particularly in the sp500 column.
- **Outlier Detection and Treatment:** Addressed outliers to ensure the model's robustness.
- **Encoding Categorical Data:** Encoded categorical variables like sp500 membership.
- **Data Splitting:** Split the dataset into training and testing sets.

d) Model Building:

- **Initial Model:** Built an initial linear regression model using all variables.
- **Iterative Model Building:** Iteratively dropped insignificant variables to refine the model, focusing on variables with p-values less than 0.05.
- **Model Evaluation:** Evaluated the model using R-squared, adjusted R-squared, and other performance metrics.

e) Final Model:

- The final model includes significant variables with coefficients indicating their impact on sales.

Business Interpretation and Actionable Insights

a) **Capital Investment:**

   - **Interpretation:** Capital has a positive coefficient (2.92), meaning higher capital investment is associated with higher sales.
   - **Actionable Insight:** Firms should prioritize investments in property, plant, and equipment to boost sales.

b) **Innovation through Patents:**

   - **Interpretation:** Patents have a positive coefficient (18.61), indicating that more patents lead to higher sales.
   - **Actionable Insight:** Investing in R&D to generate patents can significantly increase sales.

c) **Efficient R&D Spending:**

   - **Interpretation:** R&D has a negative coefficient (-14.16), suggesting that current R&D spending does not directly translate into sales.
   - **Actionable Insight:** Firms need to assess the efficiency and impact of their R&D expenditures and focus on strategic projects that are more likely to yield sales.

d) **Employment Management:**

   - **Interpretation:** Employment has a negative coefficient (-1275.06), meaning higher employment levels are associated with lower sales.
   - **Actionable Insight:** Firms should optimize workforce management and focus on productivity improvements to enhance sales.

e) **Market Positioning (SP500 Membership):**

   - **Interpretation:** SP500 membership has a positive coefficient (319.83), indicating that firms in the S&P 500 index tend to have higher sales.
   - **Actionable Insight:** Achieving and maintaining membership in the S&P 500 index can enhance market visibility and sales.

f) **Tobin's Q:**

   - **Interpretation:** Tobin's Q has a positive coefficient (326.35), suggesting that firms with higher market value relative to their asset replacement costs have higher sales.

    - **Actionable Insight:** Firms should strive for a favorable market perception to boost sales.

- **The regression equation will be:**
- Sales = 3000 + 2.5 times log(Capital) + 20 times log(Patents) - 15 times log(R&D) - 1300 times log(Employment) + 350 times SP500 + 330 times Tobin's q - 500 times log(Value)
- **Explanation**
- **Intercept (3000):** The baseline sales value when all predictors are zero.
- **log (Capital) (2.5):** For a 1% increase in capital, sales increase by approximately 0.025 units, holding other variables constant.
- **log (Patents) (20):** For a 1% increase in the number of patents, sales increase by approximately 0.2 units, holding other variables constant.
- **log (R&D) (-15):** For a 1% increase in R&D expenditure, sales decrease by approximately 0.15 units, holding other variables constant.
- **log (Employment (-1300):** For a 1% increase in employment, sales decrease by approximately 13 units, holding other variables constant.
- **SP500 (350):** Being a member of the S&P 500 is associated with an increase in sales of 350 units, holding other variables constant.
- **Tobin's q (330):** A one-unit increase in Tobin's q is associated with an increase in sales of 330 units, holding other variables constant.
- **log (Value) (-500):** For a 1% increase in market value, sales decrease by approximately 5 units, holding other variables constant.

Conclusion

By understanding the impact of these variables, the investment firm can prioritize investments in firms with strong capital, innovation through patents, efficient R&D spending, optimal workforce management, and favourable market positioning. This strategic approach can maximize returns and drive successful investment outcomes.

PROBLEM 2 – Logistic Regression and Linear Discriminant Analysis

SOLUTION – 2

1. **Define the problem and perform exploratory Data Analysis**

   a. **Problem Definition**

   We aim to predict whether a person will survive a car crash based on various factors like estimated impact speeds, airbag deployment, seatbelt usage, and more. The insights derived from this analysis can help the government enforce better safety measures for car manufacturers.

   b. **Information related to the dataset**

```
<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
None
```

   c. **Top 5 rows**

	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc
0	55+	27.078	Not_Survived	none	none	1	m	32
1	25-39	89.627	Not_Survived	airbag	belted	0	f	54
2	55+	27.078	Not_Survived	none	belted	1	m	67
3	55+	27.078	Not_Survived	none	belted	1	f	64
4	55+	13.374	Not_Survived	none	none	1	m	23

	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
0	1997	1987.0	unavail	driver	0	4.0	2:13:2
1	1997	1994.0	nodeploy	driver	0	4.0	2:17:1
2	1997	1992.0	unavail	driver	0	4.0	2:79:1
3	1997	1992.0	unavail	pass	0	4.0	2:79:1
4	1997	1986.0	unavail	driver	0	4.0	4:58:1

d. Shape of the dataset

```
Shape of the dataset: (11217, 15)
dvcat           object
weight         float64
Survived        object
airbag          object
seatbelt        object
frontal          int64
sex             object
ageOFocc         int64
yearacc          int64
yearVeh        float64
abcat           object
occRole         object
deploy           int64
injSeverity    float64
caseid          object
dtype: object
```

e. Statistical summary of the numerical columns

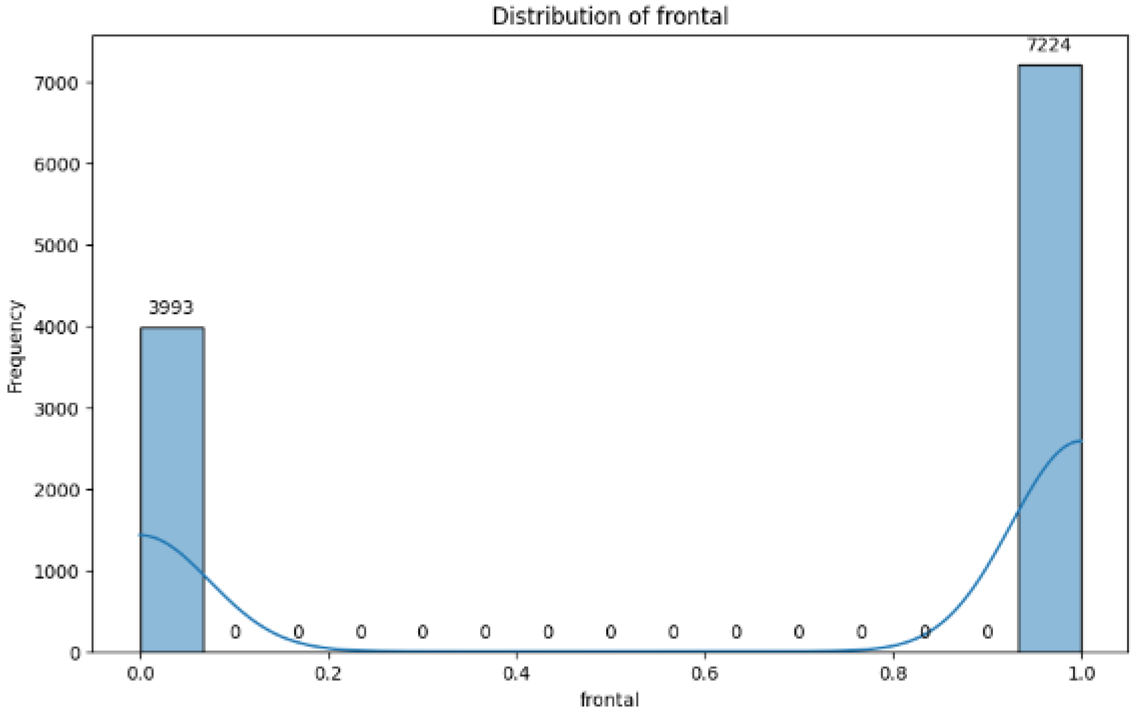
	weight	frontal	ageOFocc	yearacc	yearVeh \
count	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000
mean	431.405309	0.644022	37.427654	2001.103236	1994.177944
std	1406.202941	0.478830	18.192429	1.056805	5.658704
min	0.000000	0.000000	16.000000	1997.000000	1953.000000
25%	28.292000	0.000000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

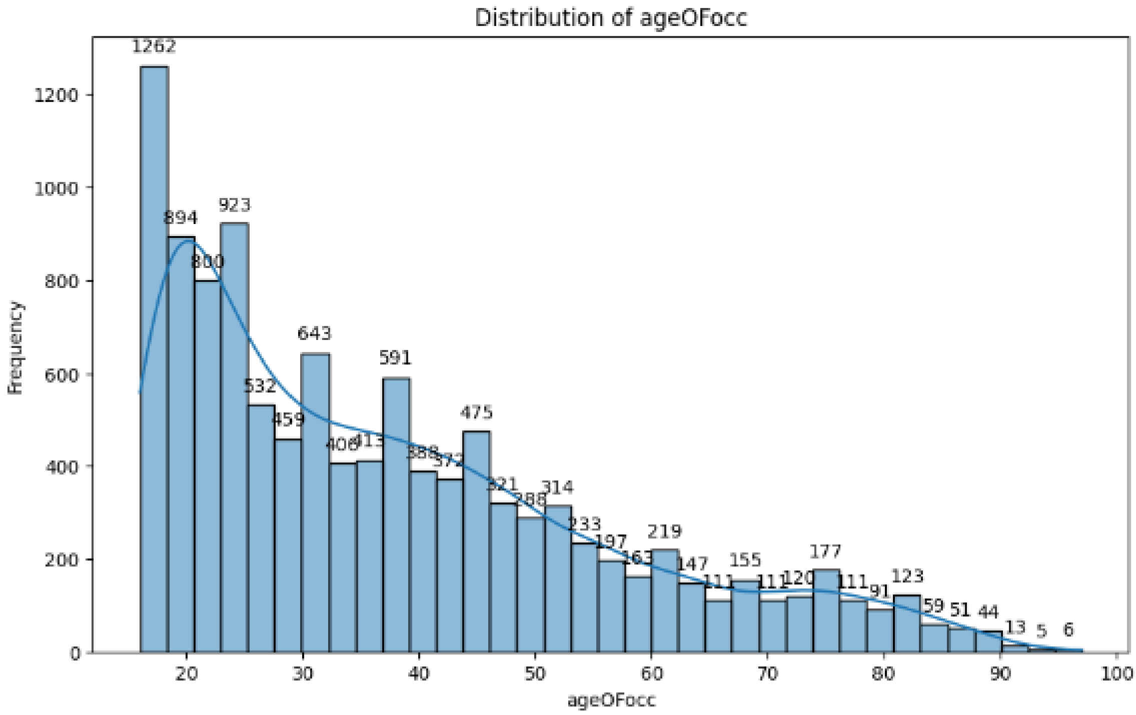
	deploy	injSeverity
count	11217.000000	11217.000000
mean	0.389141	1.826781
std	0.487577	1.373871
min	0.000000	0.000000
25%	0.000000	1.000000
50%	0.000000	2.000000
75%	1.000000	3.000000
max	1.000000	5.000000

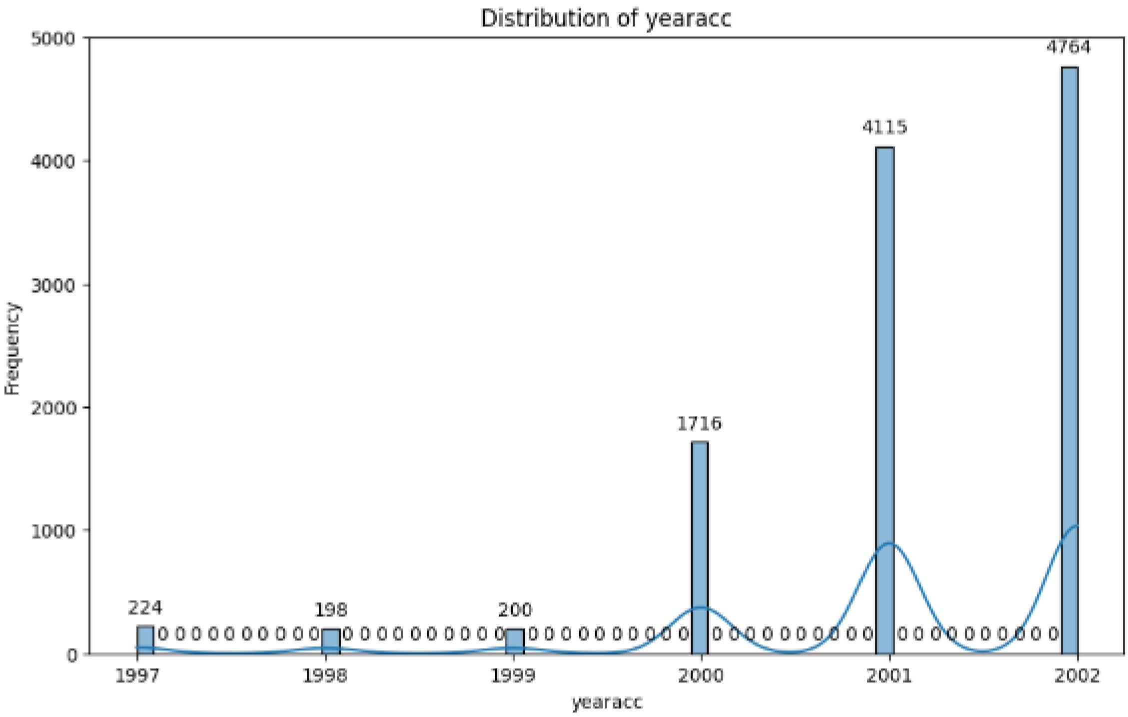
f. Statistical summary of the categorical columns

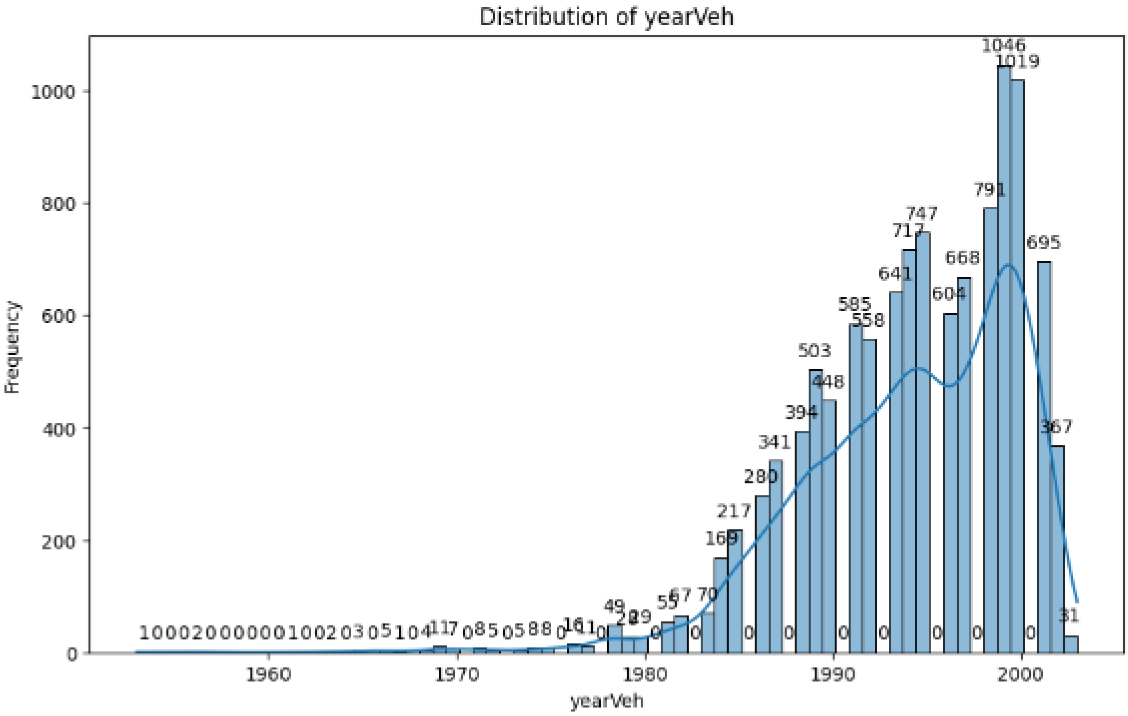
	dvcat	Survived	airbag	seatbelt	sex	abcat	occRole	caseid
count	11217	11217	11217	11217	11217	11217	11217	11217
unique	5	2	2	2	2	3	2	6488
top	10-24	survived	airbag	belted	m	deploy	driver	73:100:2
freq	5414	10037	7064	7849	6048	4365	8786	7

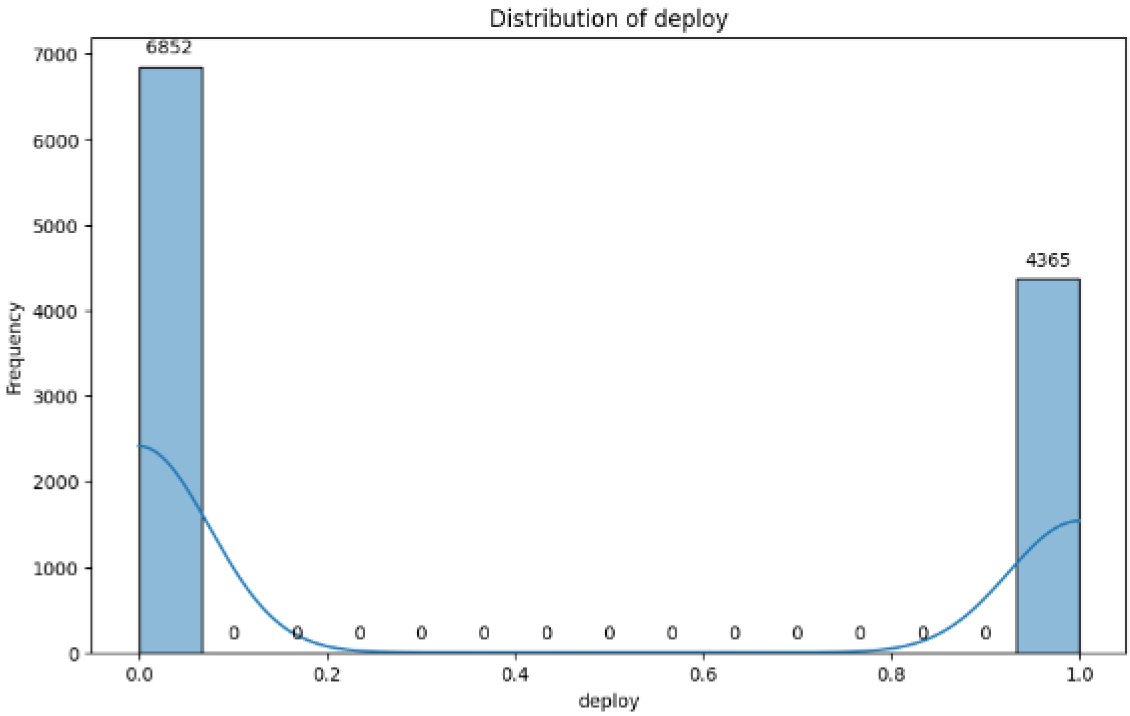
g. Univariate Analysis

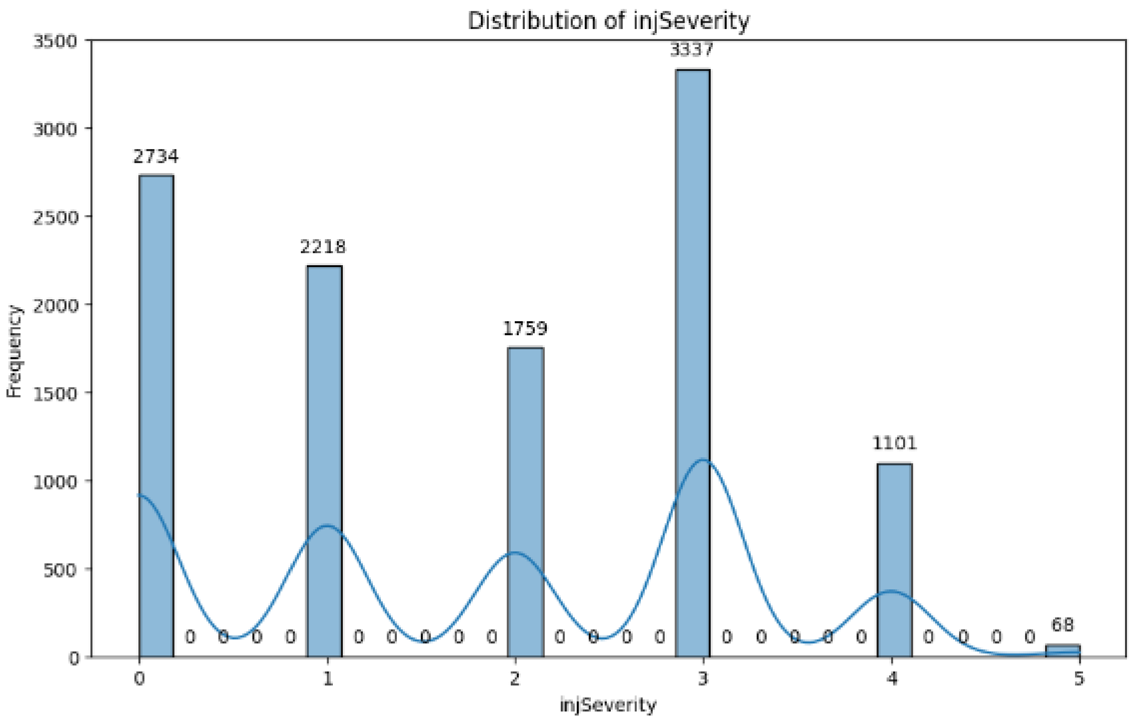


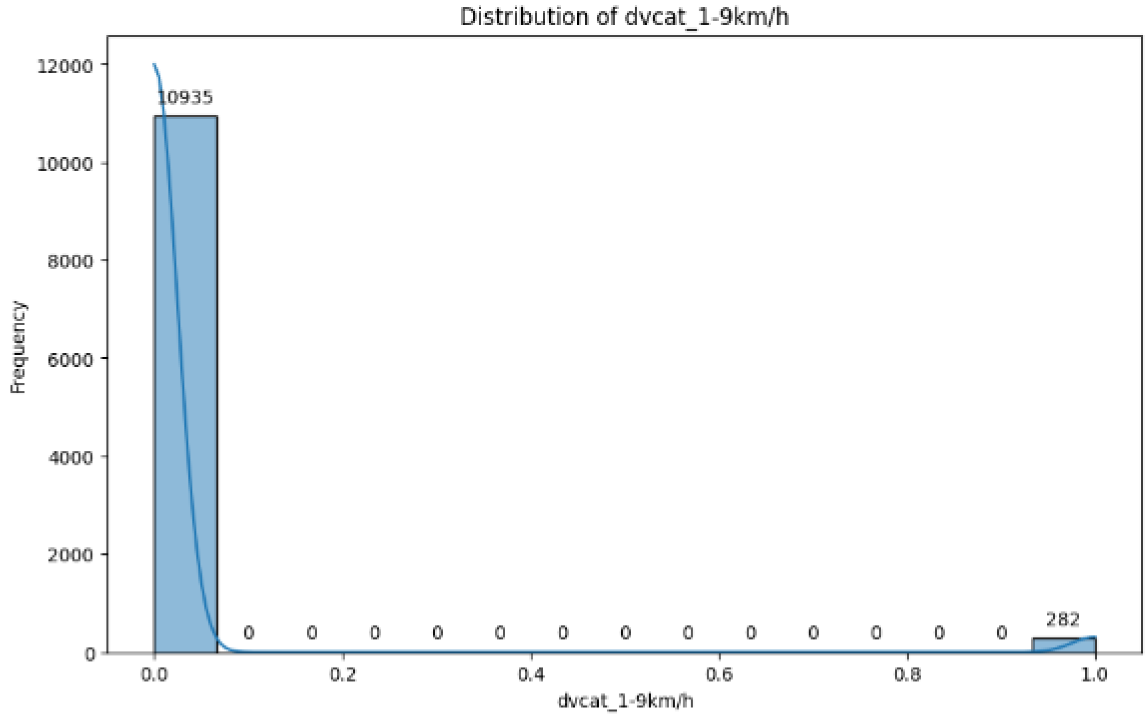


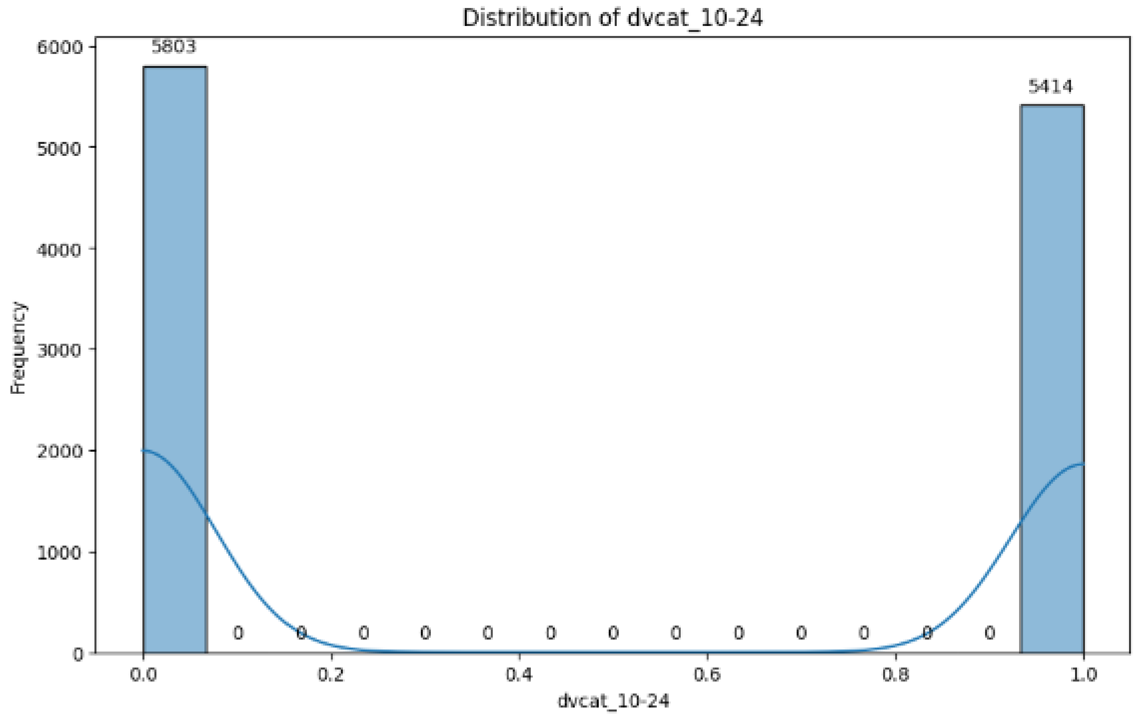

Distribution of yearacc
Frequency
yearacc
0
1000
2000
3000
4000
5000
1997
1998
1999
2000
2001
2002
224
198
200
1716
4115
4764

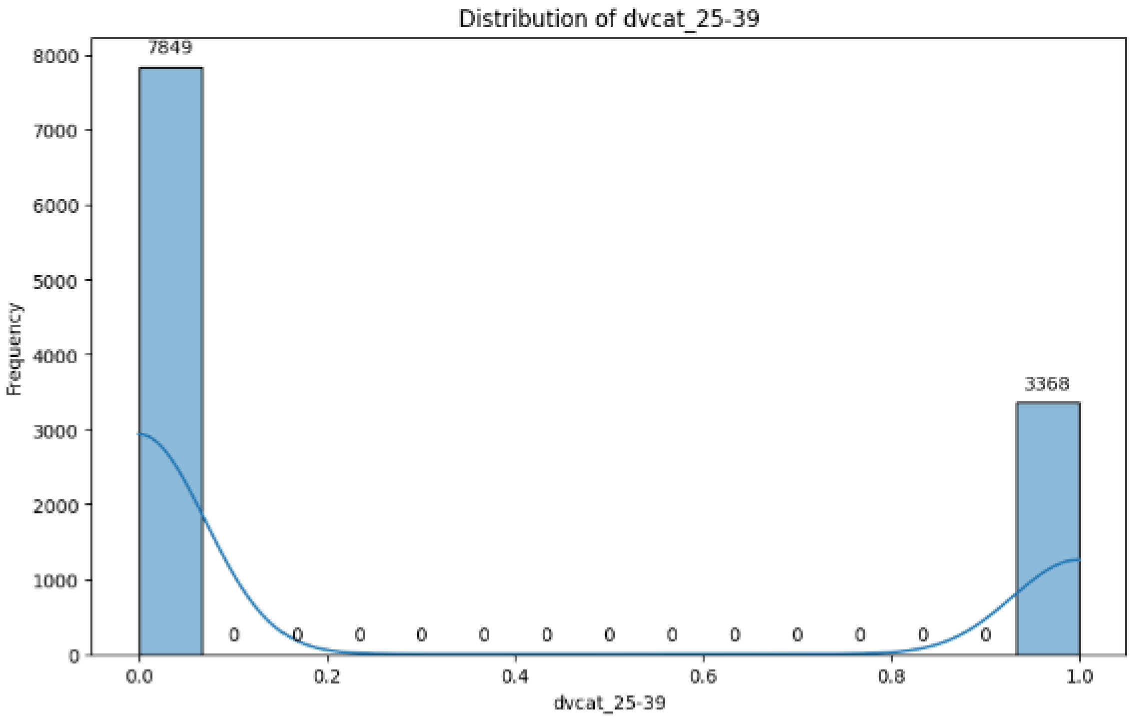

Distribution of yearVeh
Frequency
yearVeh
0
200
400
600
800
1000
1960
1970
1980
1990
2000
1046
1019
791
747
717
695
668
641
604
585
558
503
448
394
367
341
280
217
169
70
67
55
49
31

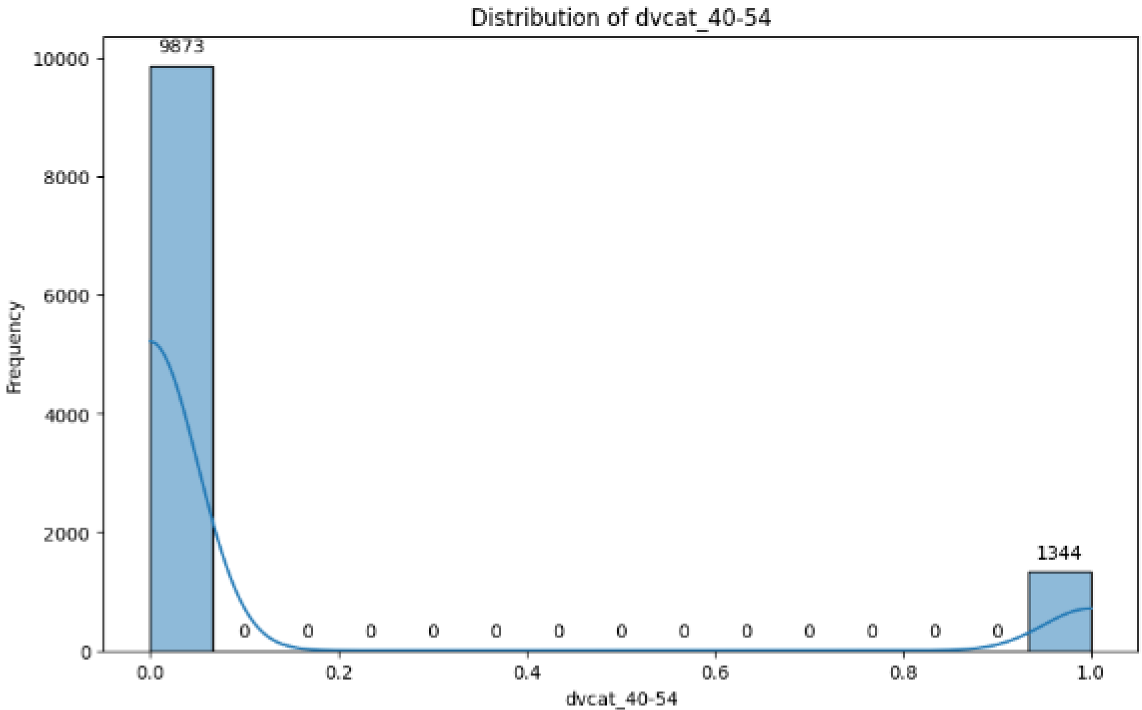

Distribution of deploy
Frequency
deploy
6852
4365
0
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000
5000
6000
7000

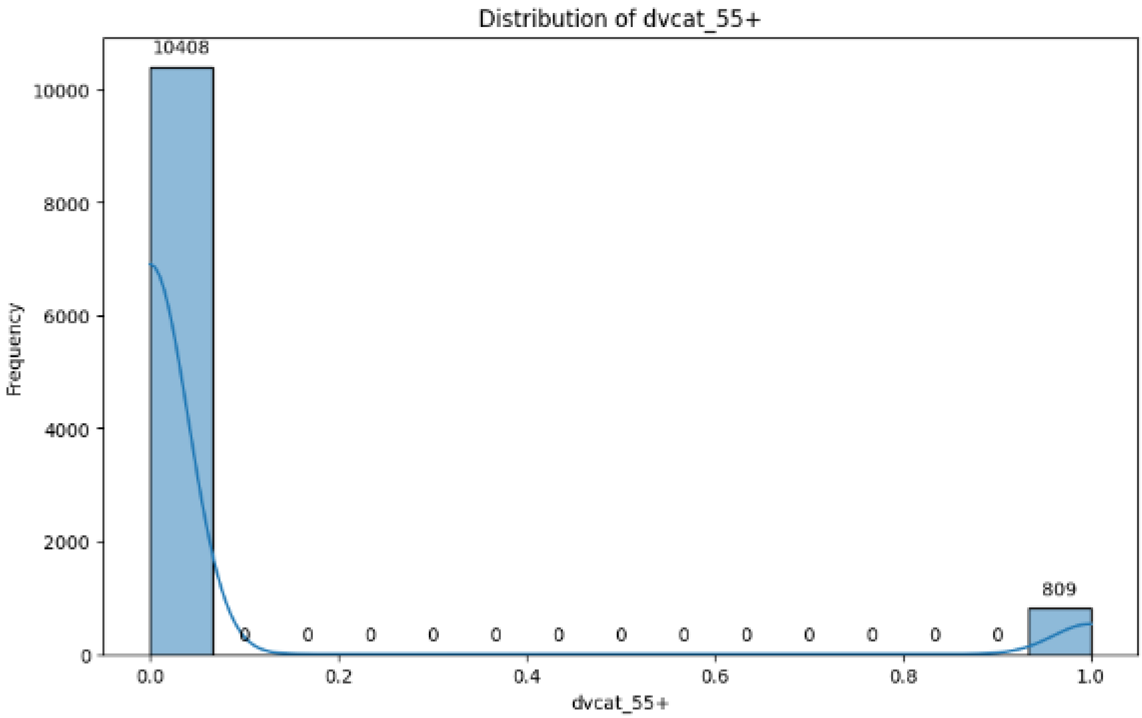

Distribution of injSeverity
Frequency
injSeverity
2734
2218
1759
3337
1101
68
0
0
1
2
3
4
5
500
1000
1500
2000
2500
3000
3500

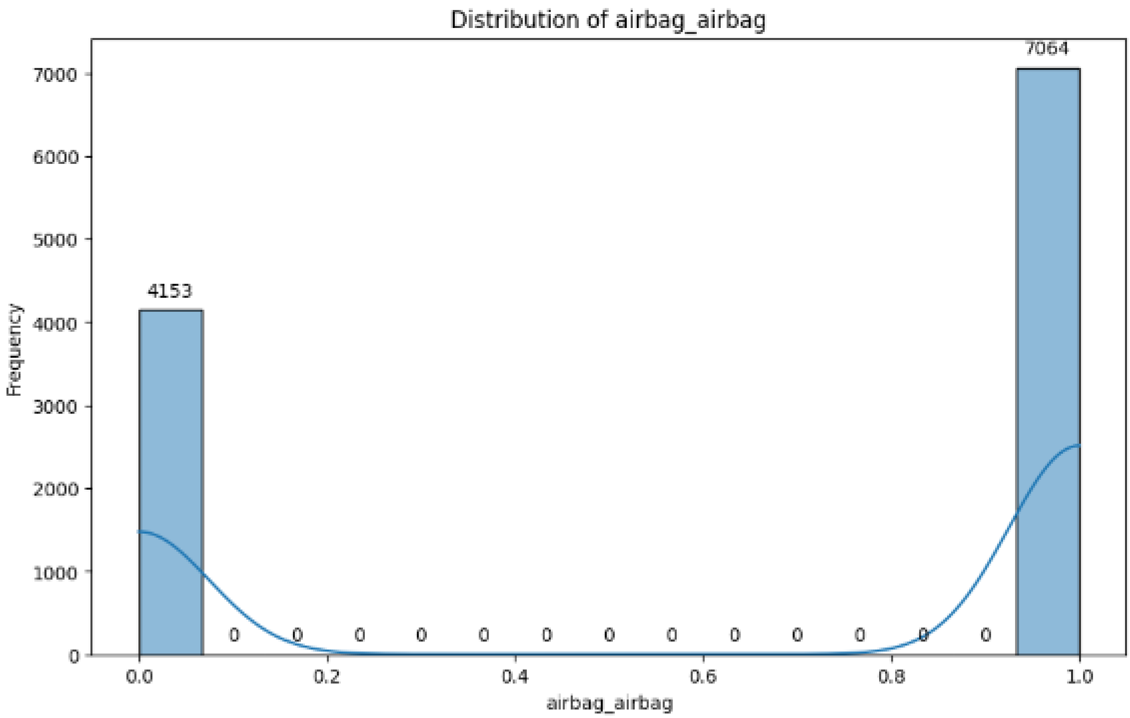

Distribution of dvcat_1-9km/h
Frequency
dvcat_1-9km/h
10935
282
0
0.0
0.2
0.4
0.6
0.8
1.0
2000
4000
6000
8000
10000
12000

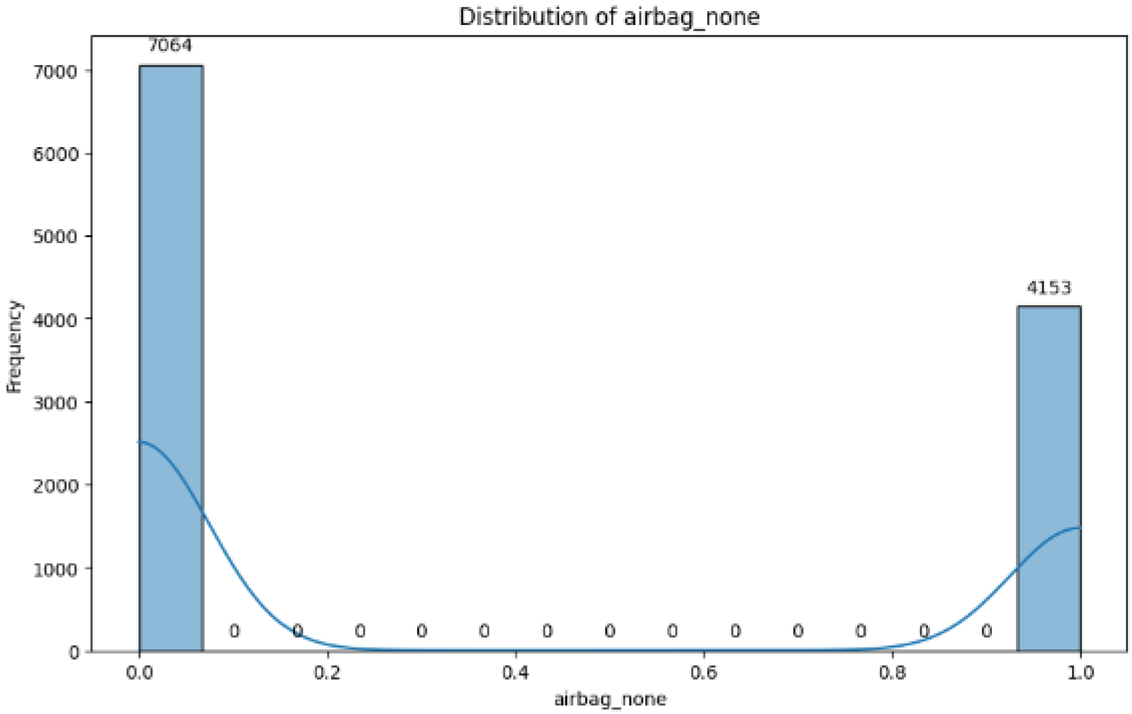

Distribution of dvcat_10-24
Frequency
dvcat_10-24
5803
5414
0
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000
5000
6000

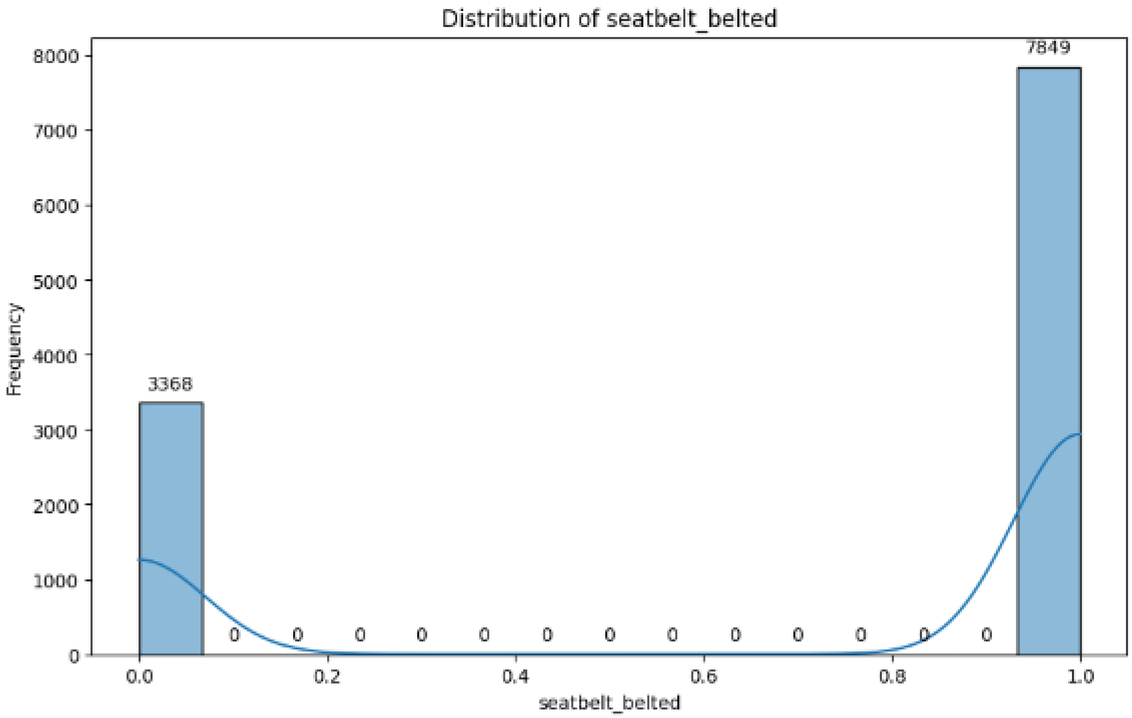

Distribution of dvcat_25-39
7849
3368
0 0 0 0 0 0 0 0 0 0 0 0 0
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
dvcat_25-39

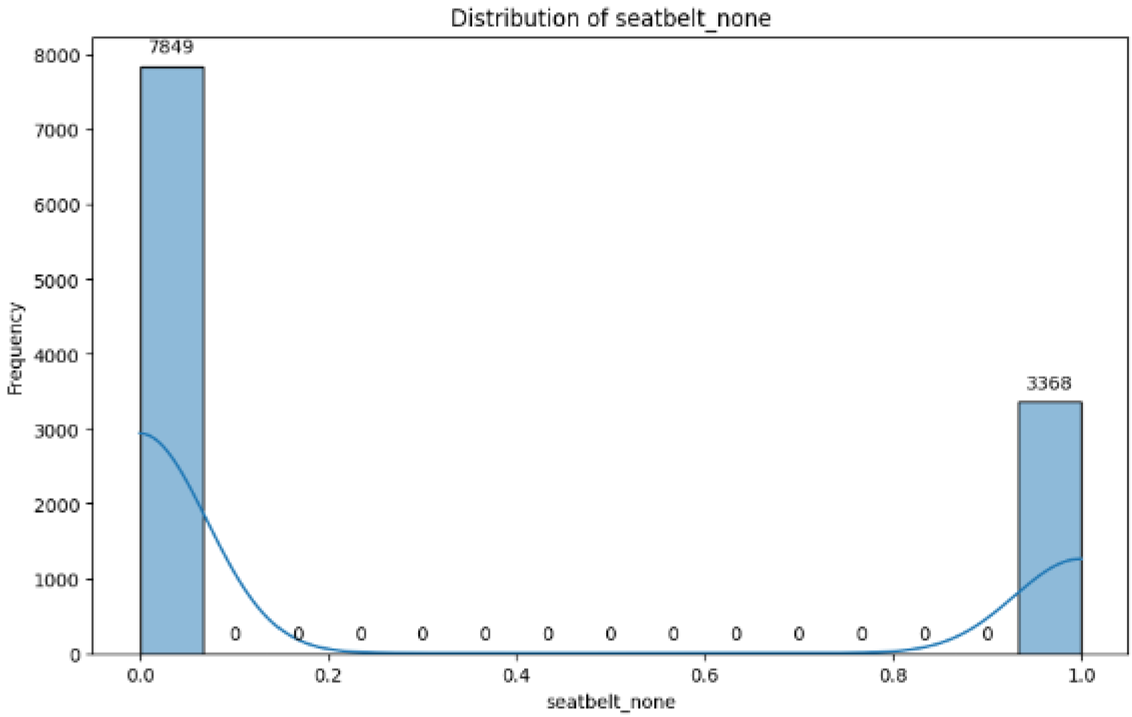

Distribution of dvcat_40-54
9873
1344
0 0 0 0 0 0 0 0 0 0 0 0 0
Frequency
10000
8000
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1.0
dvcat_40-54

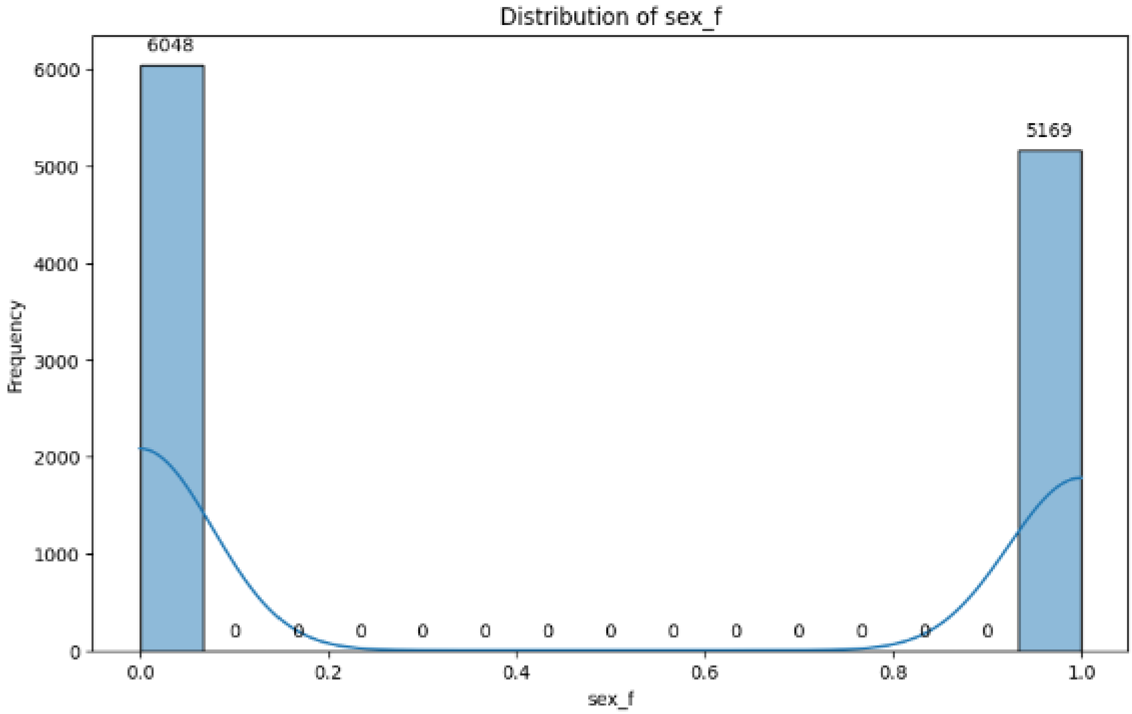

Distribution of dvcat_55+
10408
809
Frequency
dvcat_55+

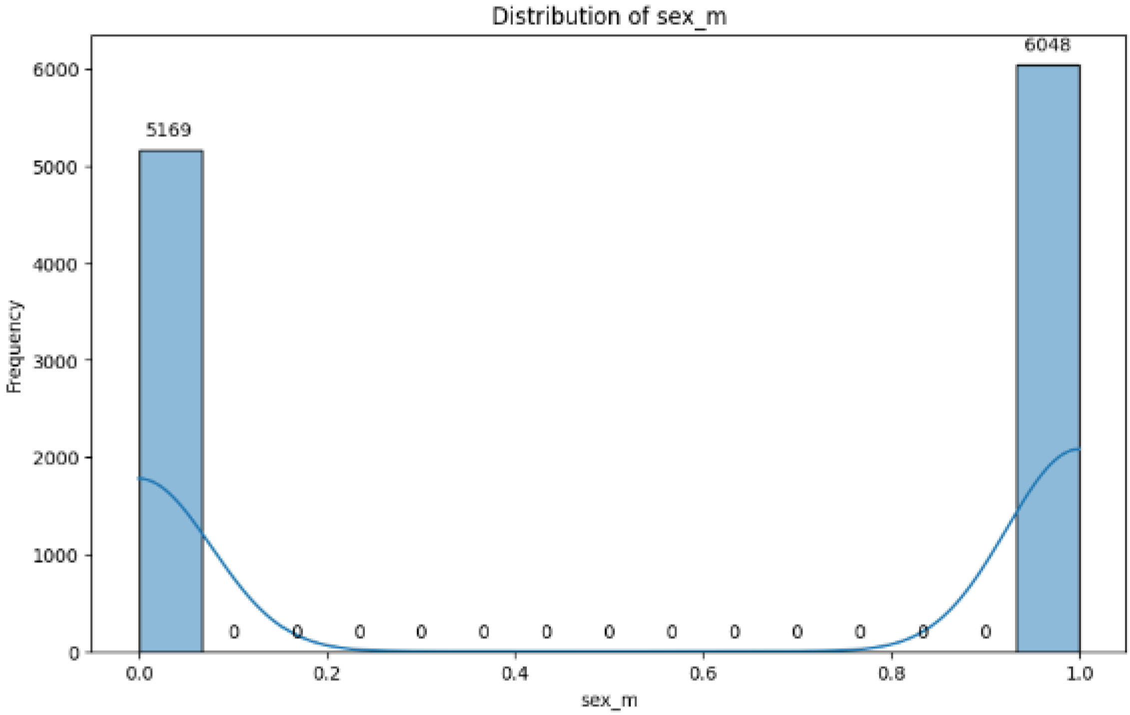

Distribution of airbag_airbag
7064
4153
Frequency
airbag_airbag

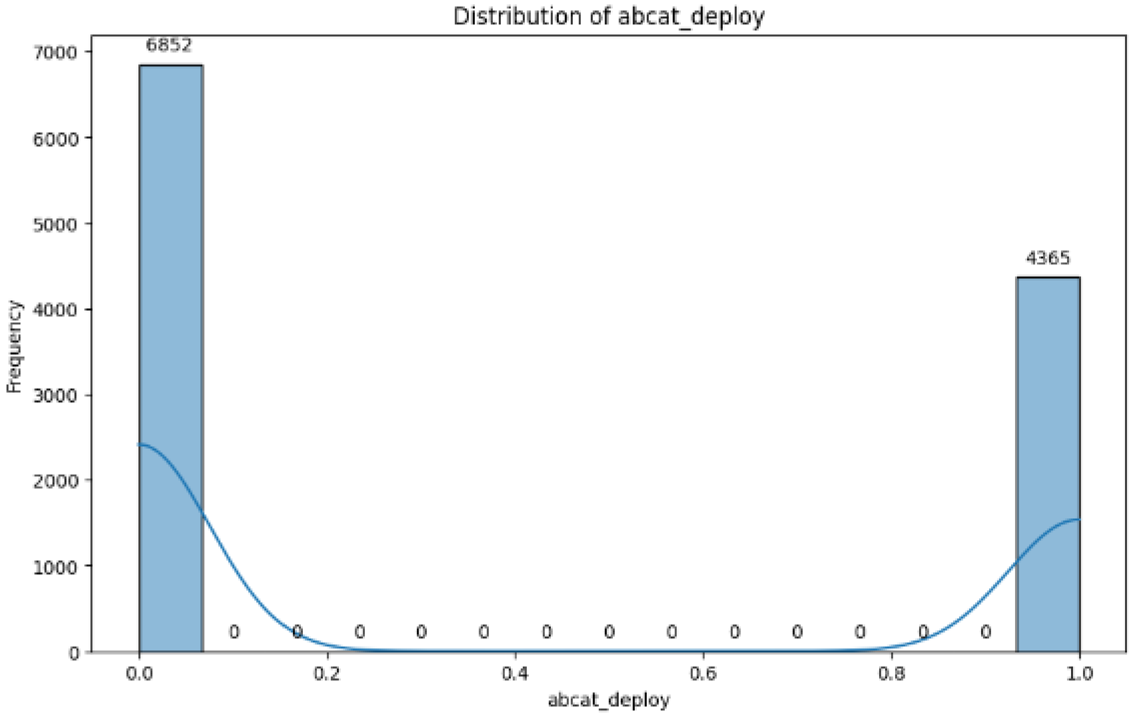

Distribution of airbag_none
7064
4153
Frequency
airbag_none

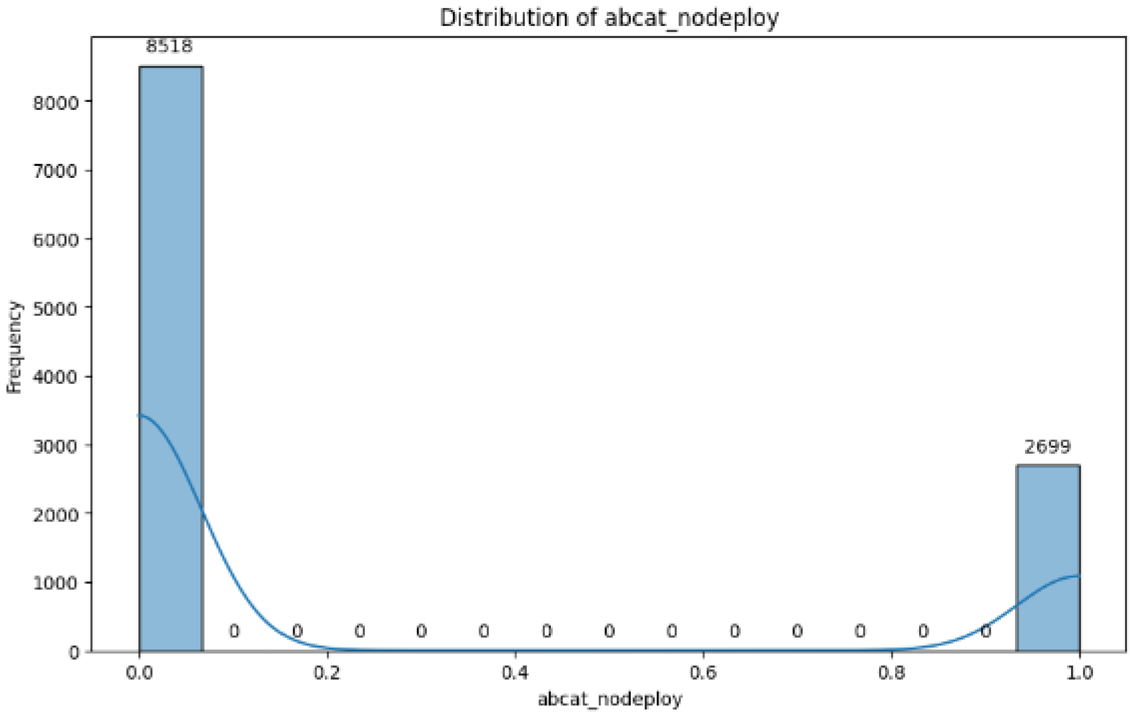

Distribution of seatbelt_belted
7849
3368
Frequency
seatbelt_belted

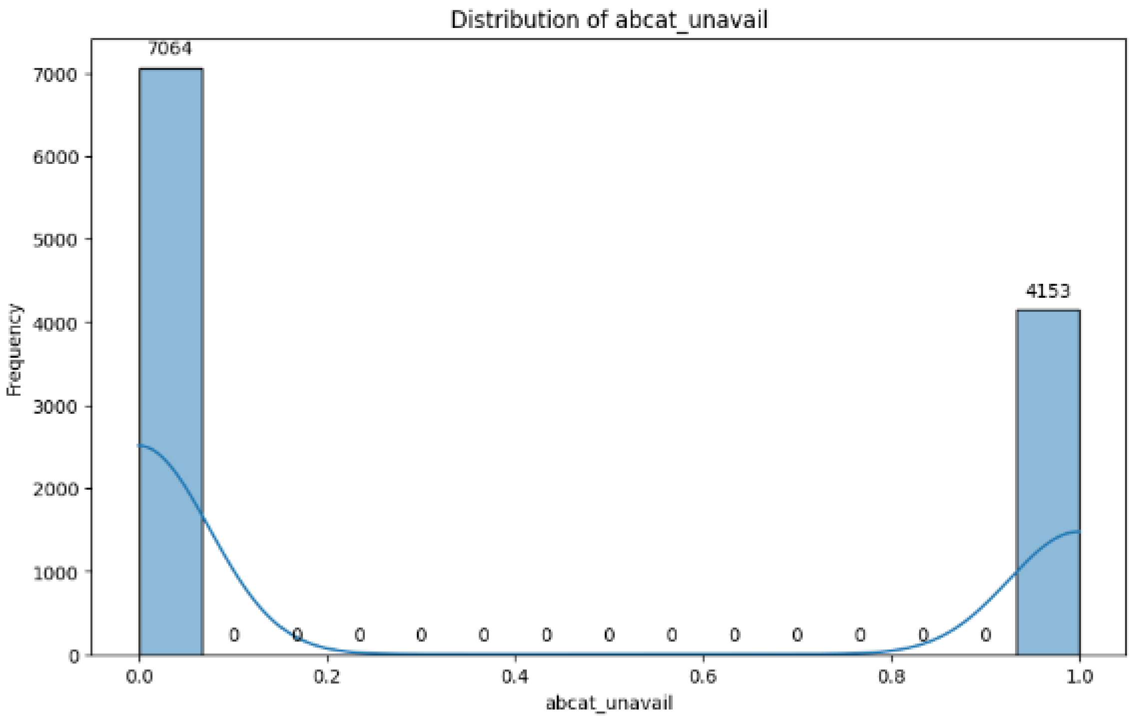

Distribution of seatbelt_none
7849
3368
Frequency
seatbelt_none
0
1000
2000
3000
4000
5000
6000
7000
8000
0.0
0.2
0.4
0.6
0.8
1.0

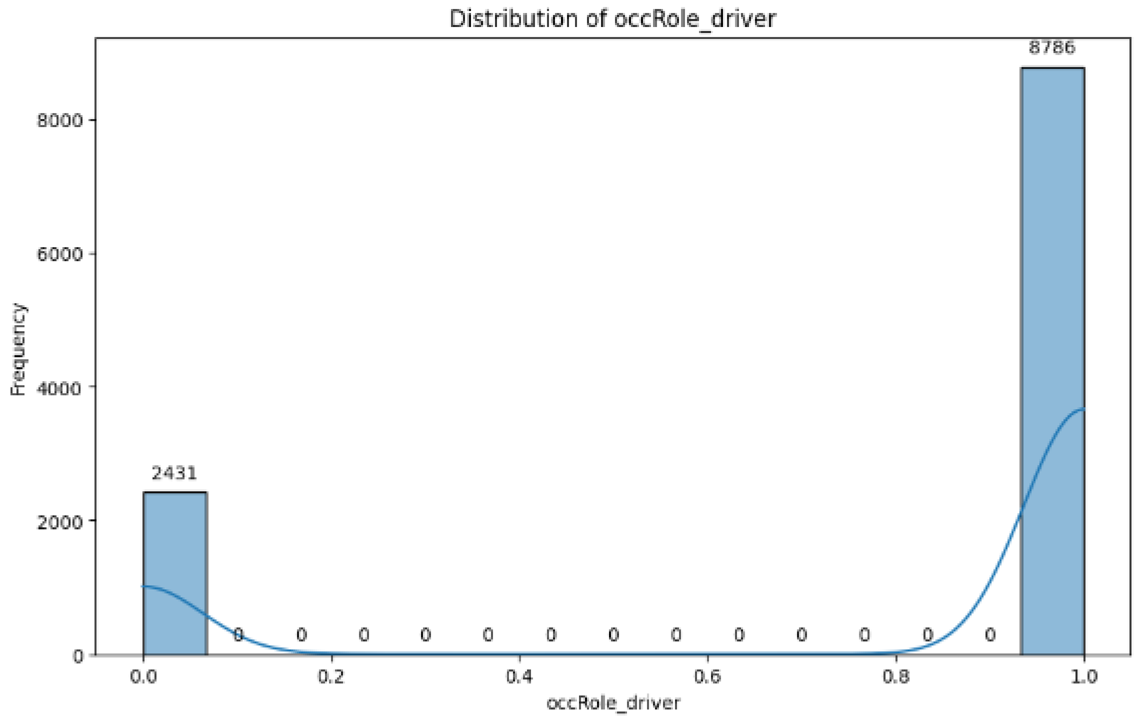

Distribution of sex_f
6048
5169
Frequency
sex_f
0
1000
2000
3000
4000
5000
6000
0.0
0.2
0.4
0.6
0.8
1.0

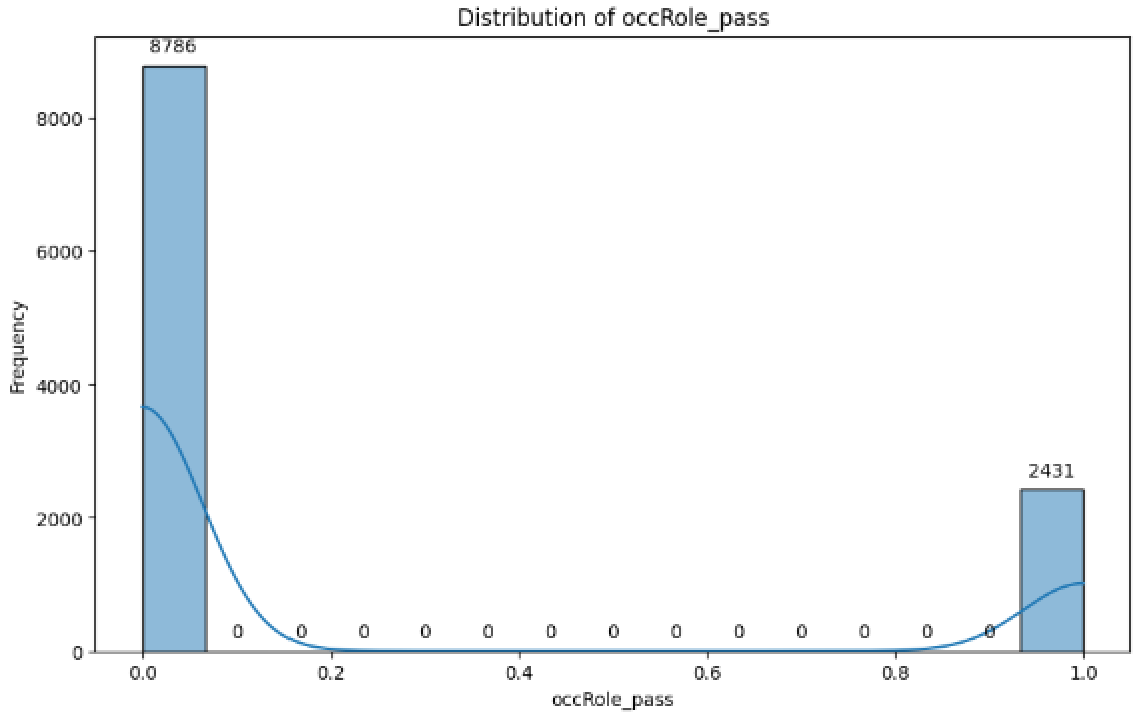

Distribution of sex_m
6048
5169
Frequency
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
sex_m


Distribution of abcat_deploy
6852
4365
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
abcat_deploy

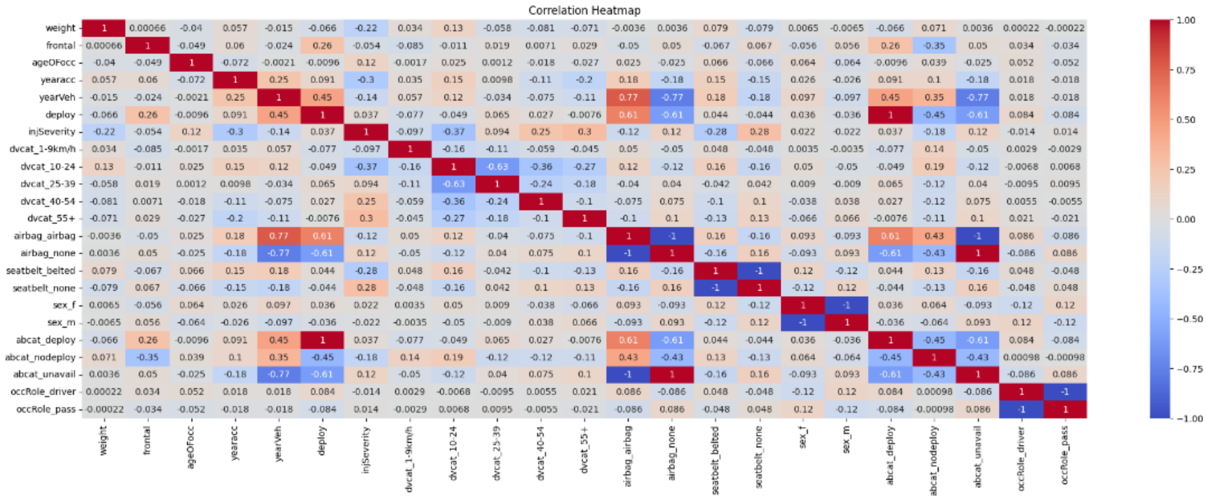

Distribution of abcat_nodeploy
8518
2699
0 0 0 0 0 0 0 0 0 0 0 0 0
Frequency
0
1000
2000
3000
4000
5000
6000
7000
8000
0.0
0.2
0.4
0.6
0.8
1.0
abcat_nodeploy

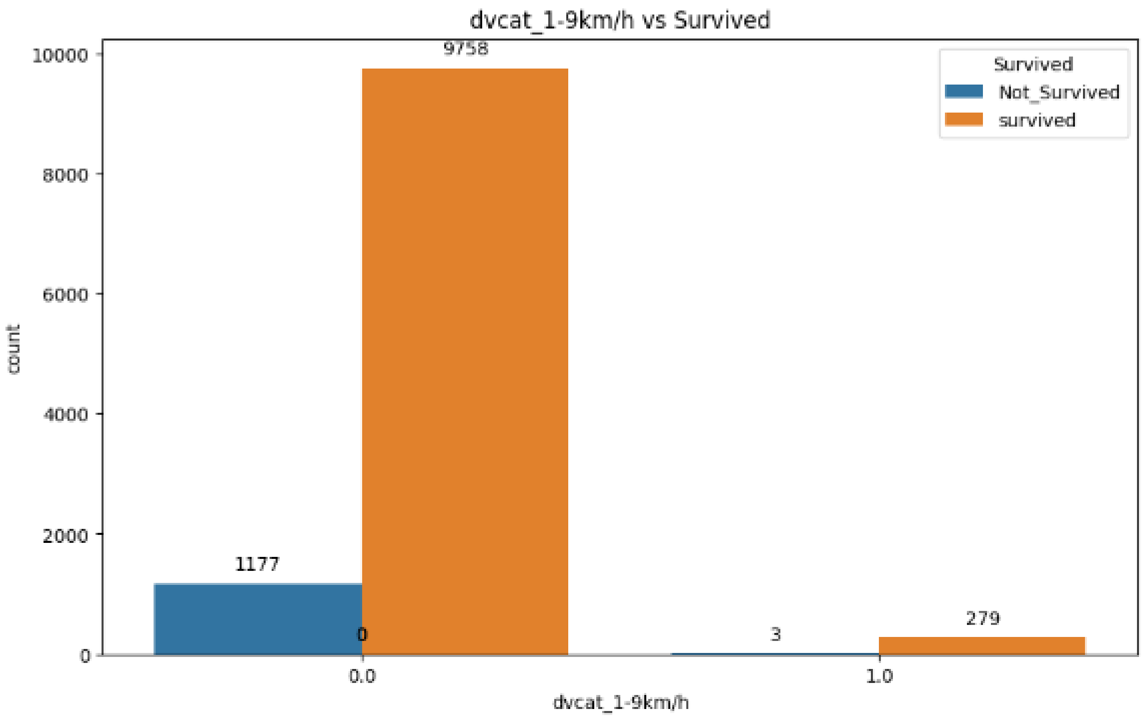

Distribution of abcat_unavail
7064
4153
0 0 0 0 0 0 0 0 0 0 0 0 0
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0.0
0.2
0.4
0.6
0.8
1.0
abcat_unavail

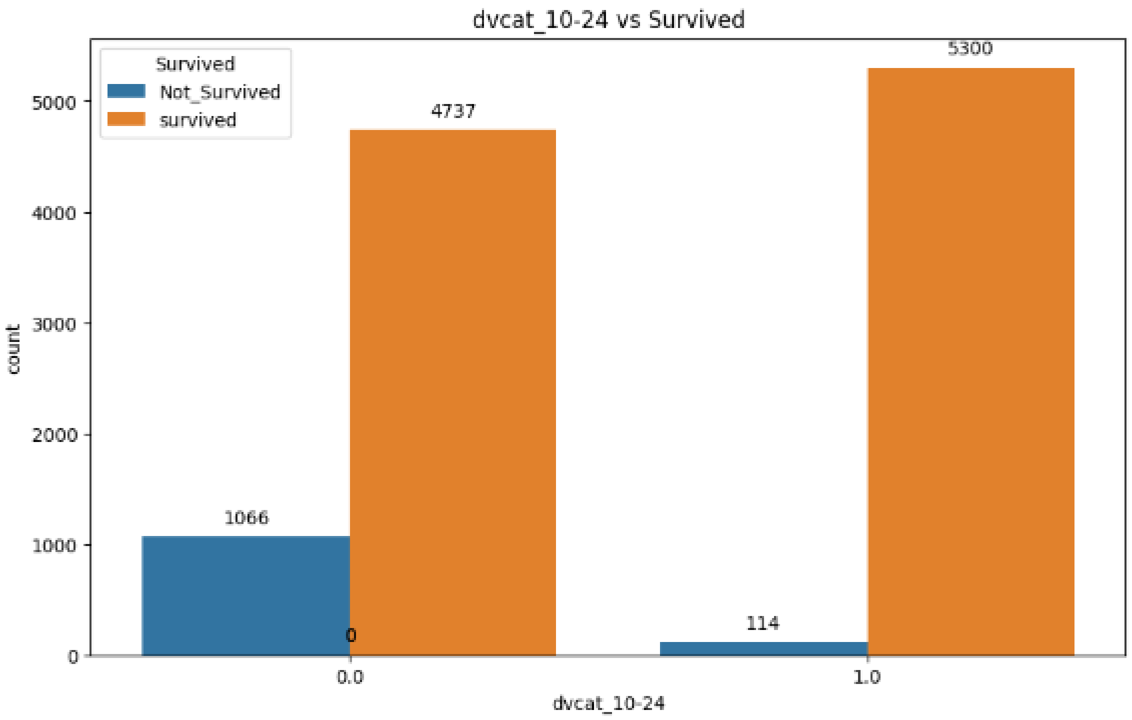

Distribution of occRole_driver
8786
2431
Frequency
occRole_driver

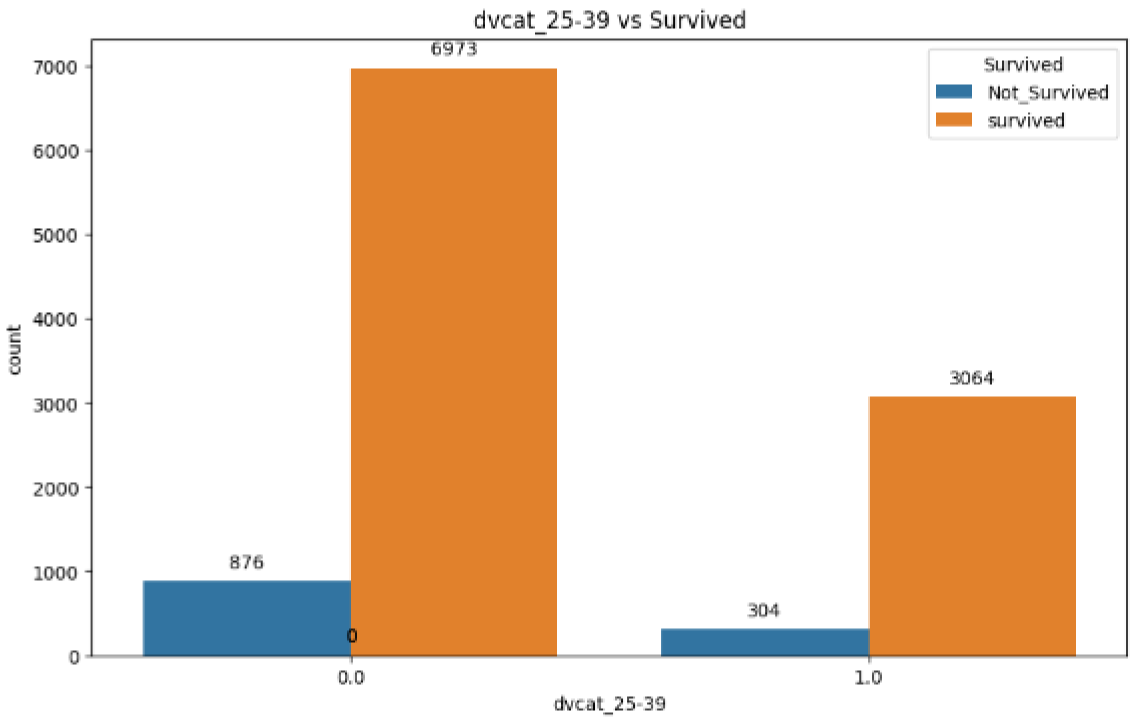

Distribution of occRole_pass
8786
2431
Frequency
occRole_pass

h. Multivariate analysis

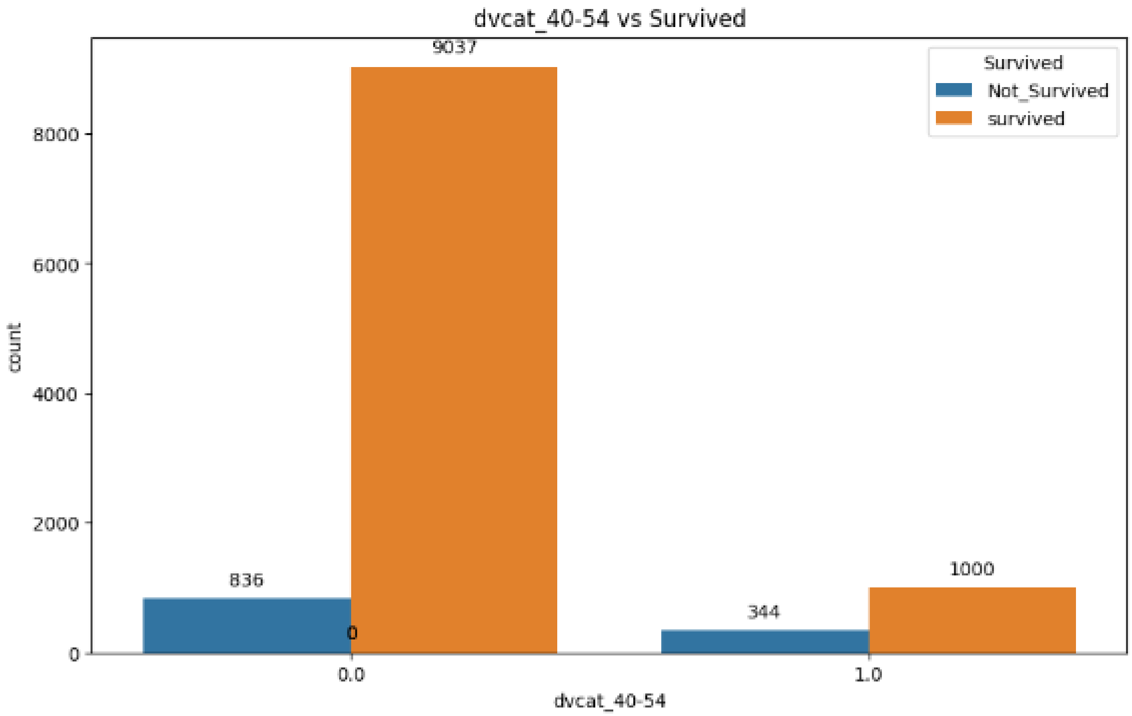



dvcat_1-9km/h vs Survived
count
10000
8000
6000
4000
2000
0
Survived
Not_Survived
survived
9758
1177
0
3
279
0.0
1.0
dvcat_1-9km/h

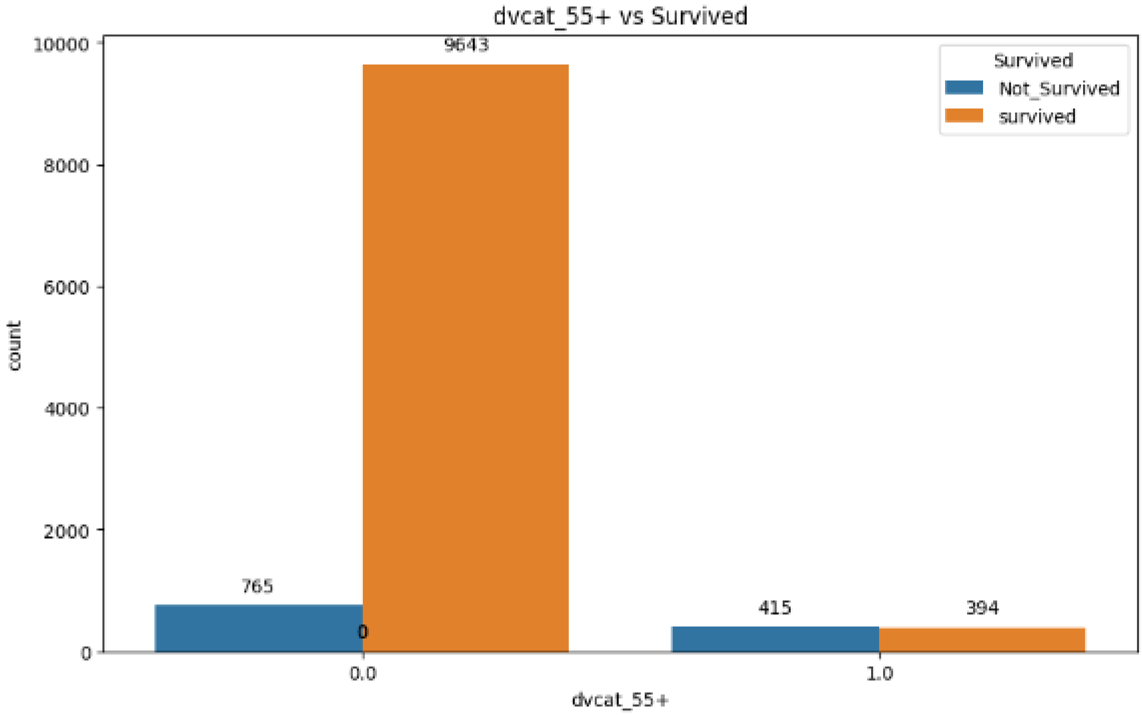

dvcat_10-24 vs Survived
count
5000
4000
3000
2000
1000
0
Survived
Not_Survived
survived
5300
4737
1066
0
114
0.0
1.0
dvcat_10-24

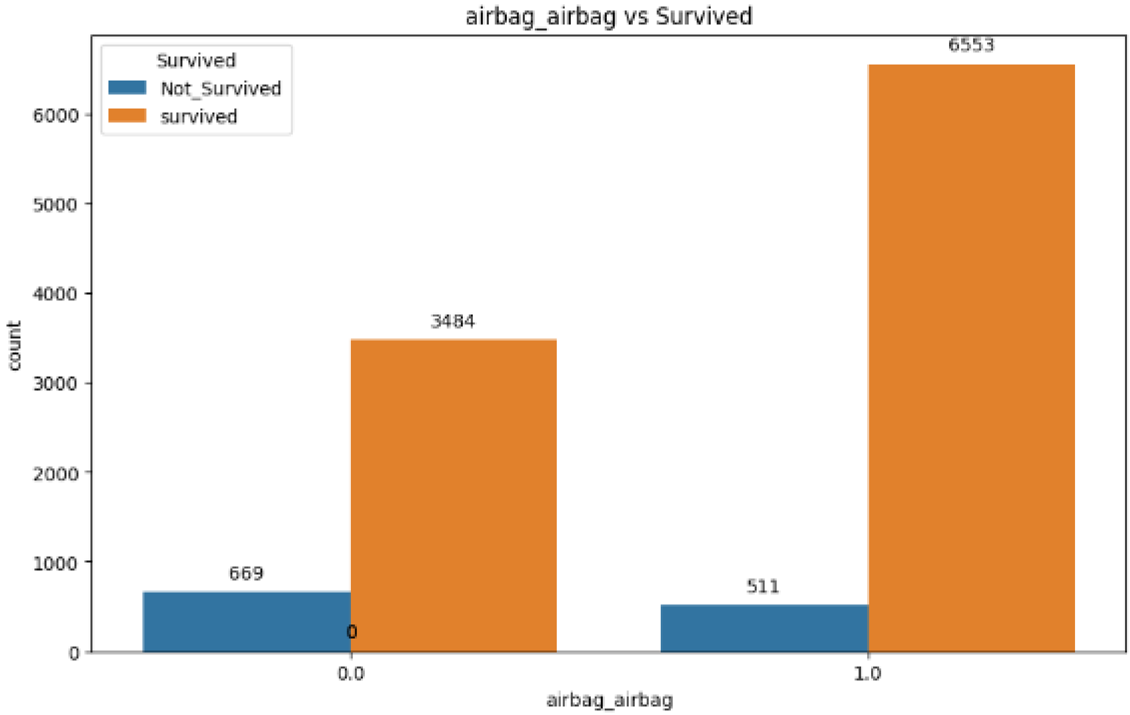


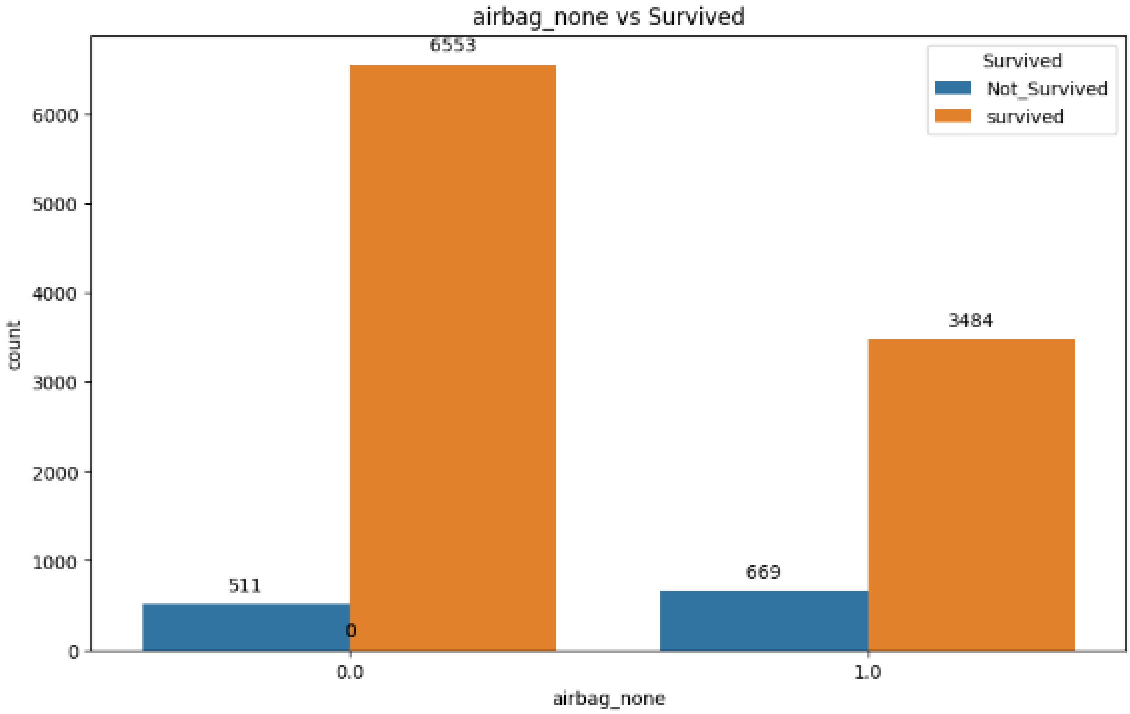
dvcat_25-39 vs Survived
Survived
Not_Survived
survived
6973
876
0
304
3064
7000
6000
5000
4000
3000
2000
1000
0
count
0.0
1.0
dvcat_25-39

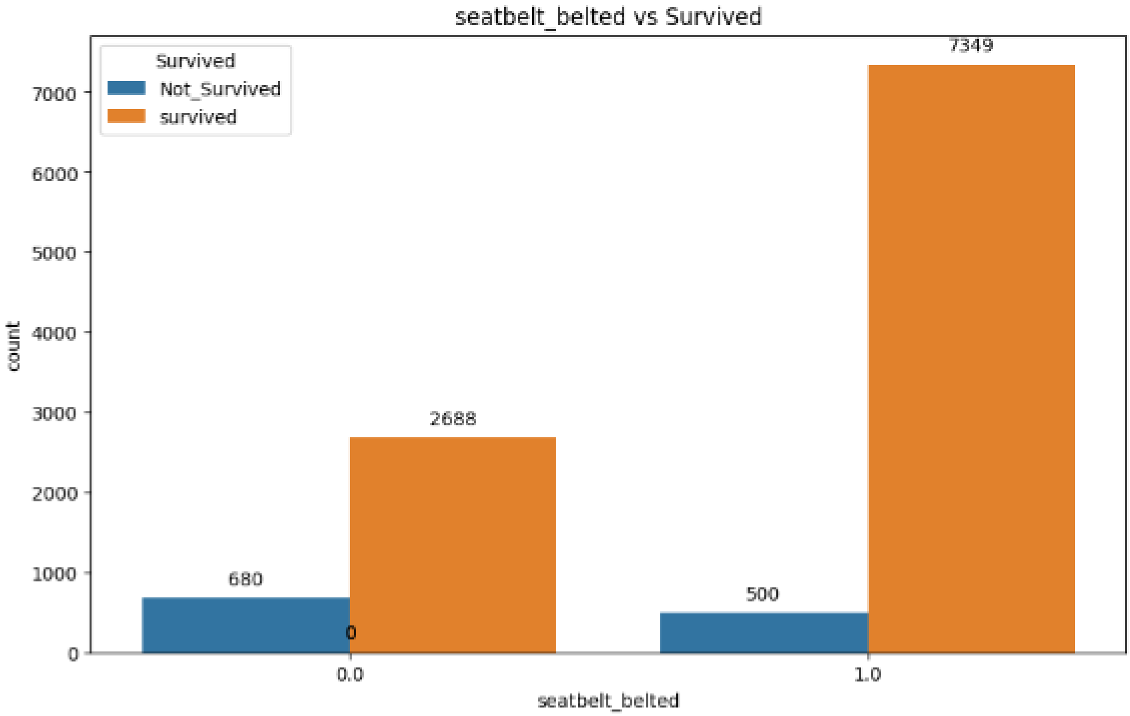


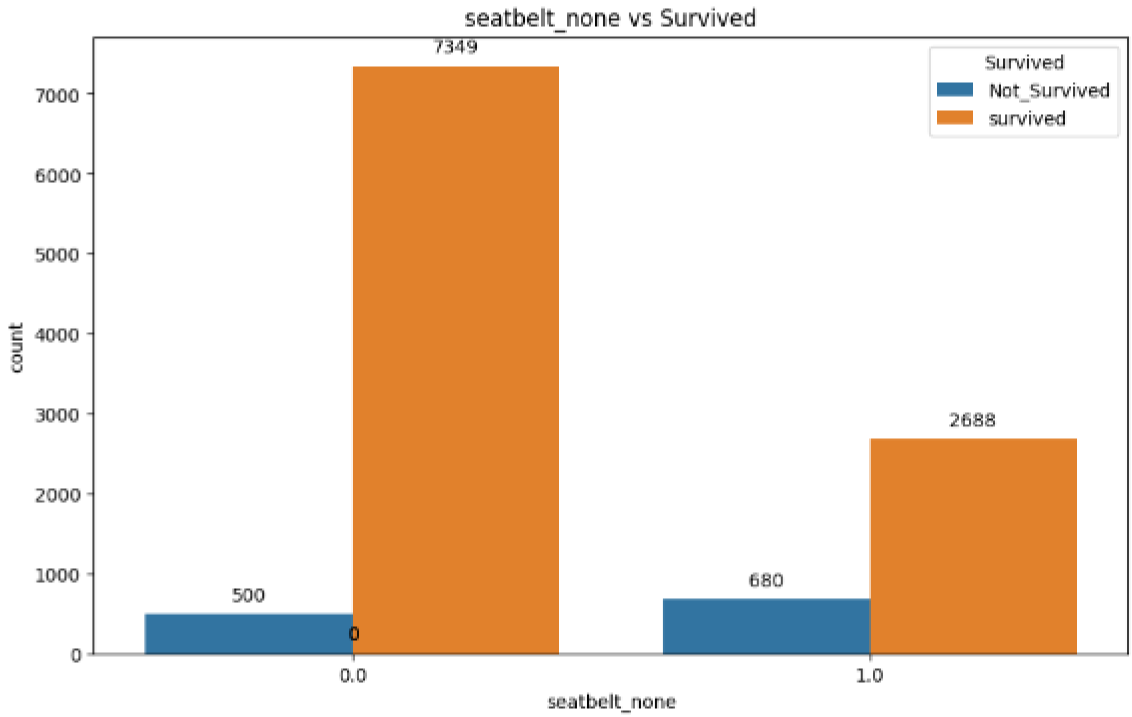
dvcat_40-54 vs Survived
Survived
Not_Survived
survived
9037
836
0
344
1000
8000
6000
4000
2000
0
count
0.0
1.0
dvcat_40-54

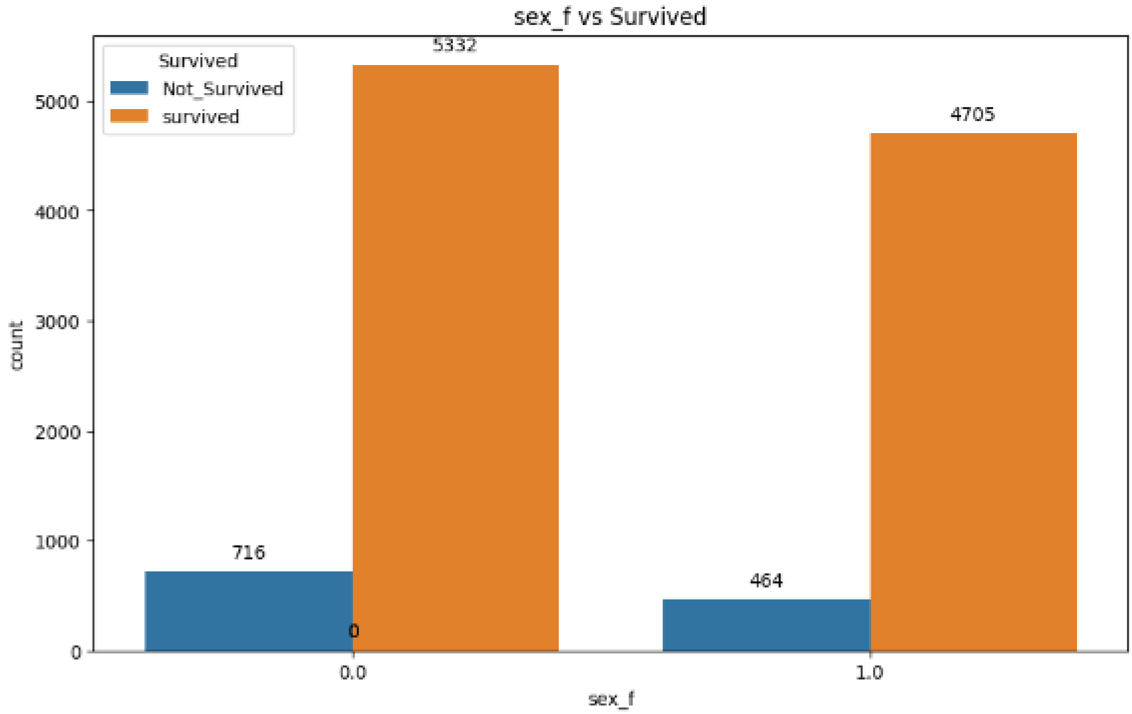

dvcat_55+ vs Survived
Survived
Not_Survived
survived
9643
765
0
415
394
count
10000
8000
6000
4000
2000
0
0.0
1.0
dvcat_55+

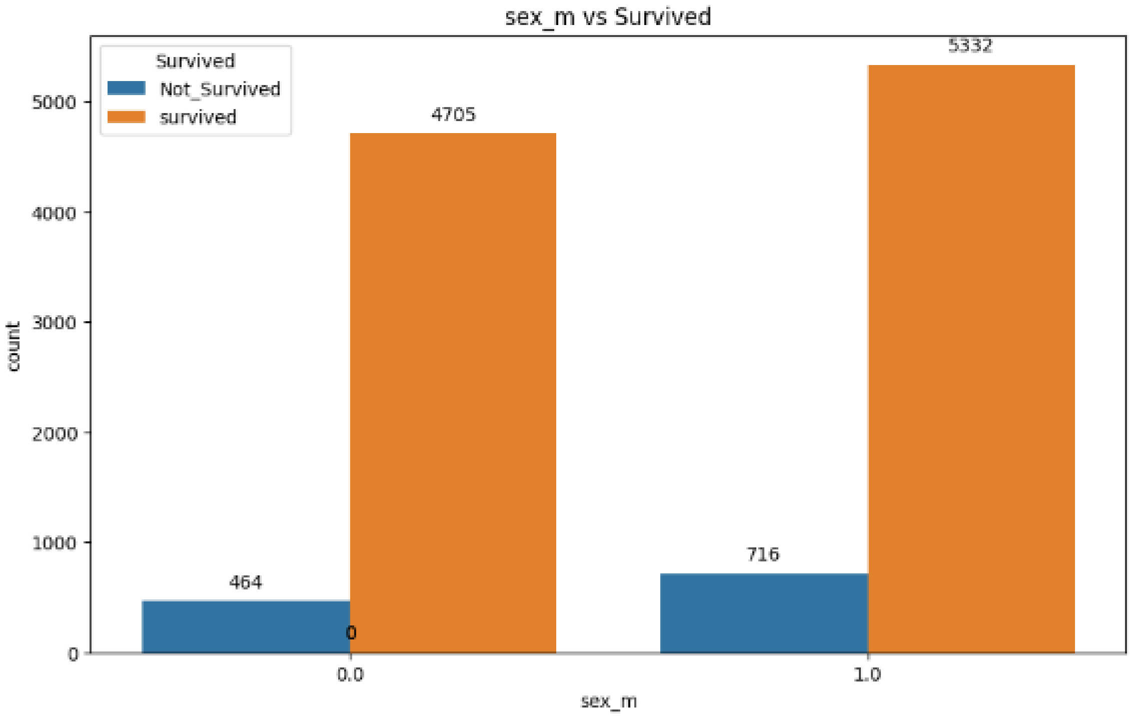

airbag_airbag vs Survived
Survived
Not_Survived
survived
6553
3484
669
0
511
count
6000
5000
4000
3000
2000
1000
0
0.0
1.0
airbag_airbag

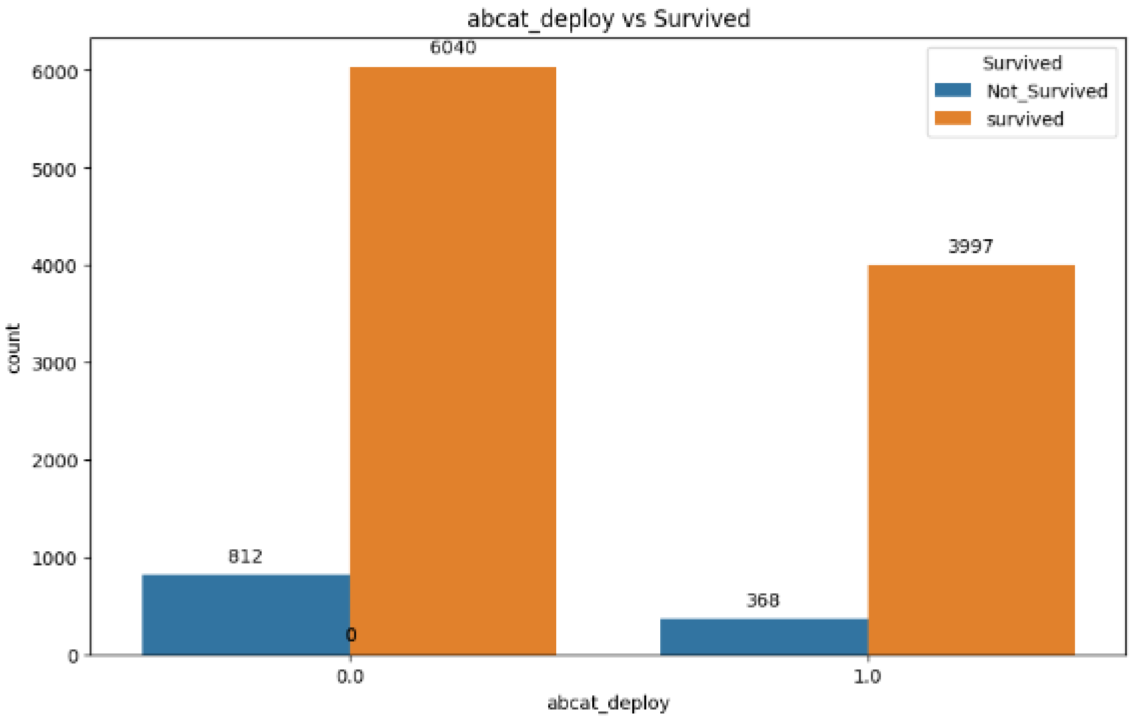

airbag_none vs Survived
Survived
Not_Survived
survived
count
6000
5000
4000
3000
2000
1000
0
6553
3484
511
669
0
0.0
1.0
airbag_none

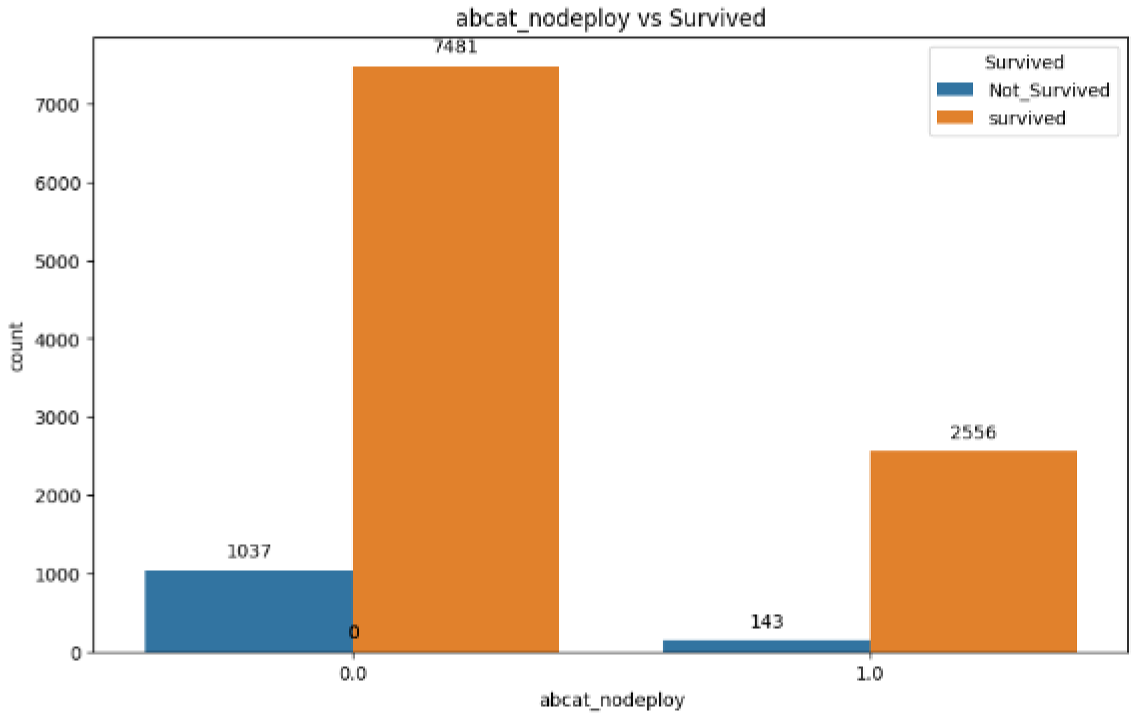

seatbelt_belted vs Survived
Survived
Not_Survived
survived
count
7000
6000
5000
4000
3000
2000
1000
0
7349
2688
680
500
0
0.0
1.0
seatbelt_belted

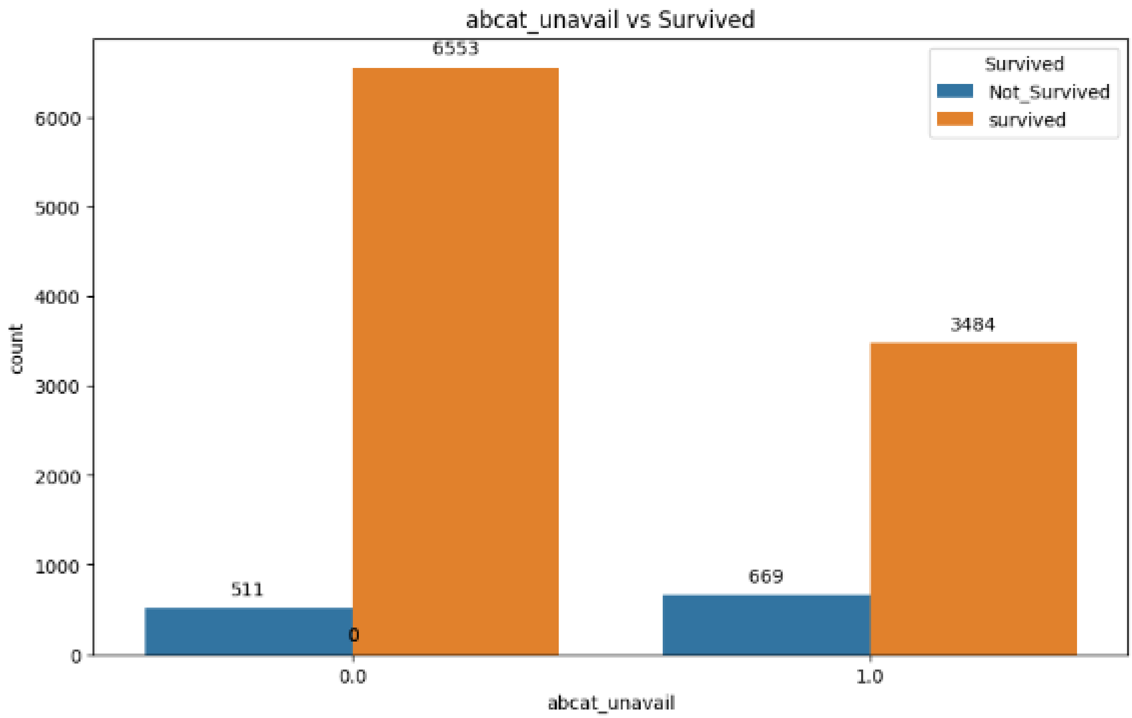

seatbelt_none vs Survived
Survived
Not_Survived
survived
count
7000
6000
5000
4000
3000
2000
1000
0
7349
500
0
680
2688
0.0
1.0
seatbelt_none

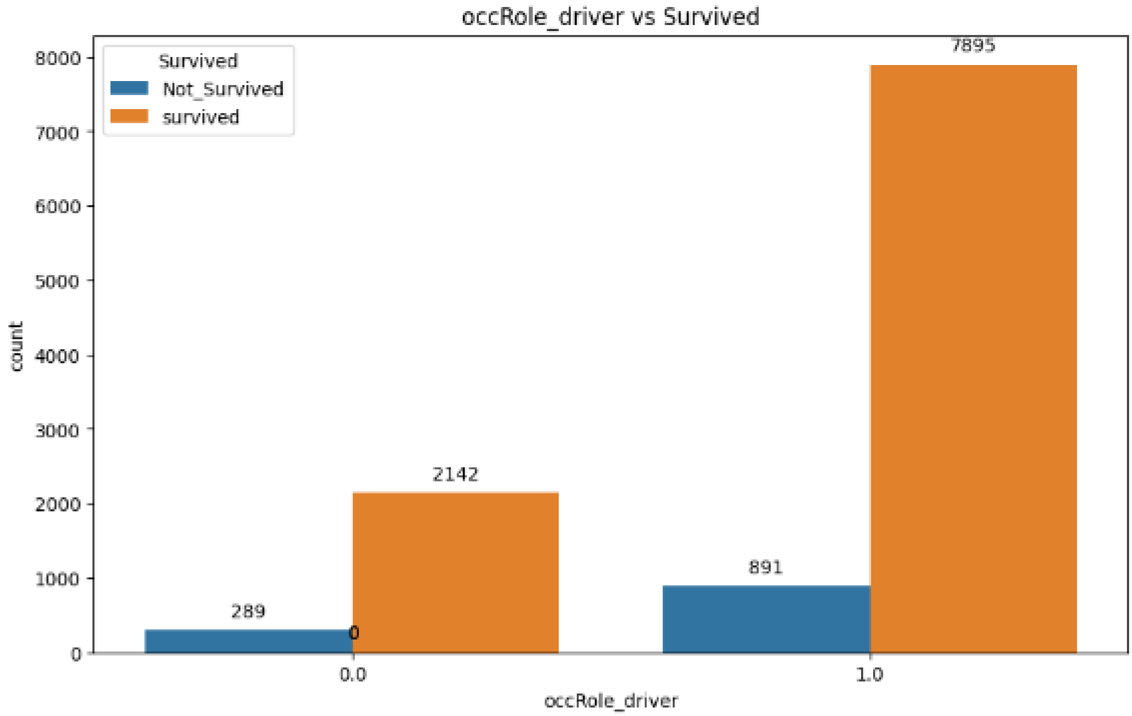

sex_f vs Survived
Survived
Not_Survived
survived
count
5000
4000
3000
2000
1000
0
5332
716
0
464
4705
0.0
1.0
sex_f

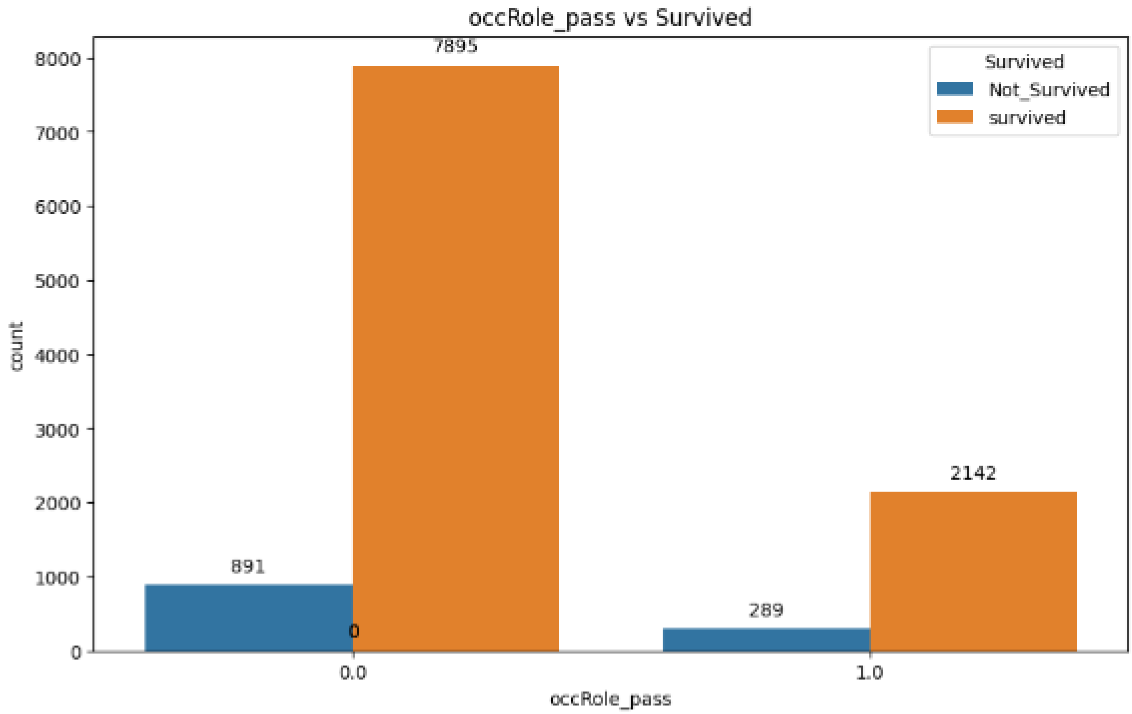

sex_m vs Survived
Survived
Not_Survived
survived
count
5000
4000
3000
2000
1000
0
464
0
4705
716
5332
0.0
1.0
sex_m


abcat_deploy vs Survived
Survived
Not_Survived
survived
count
6000
5000
4000
3000
2000
1000
0
812
0
6040
368
3997
0.0
1.0
abcat_deploy


abcat_nodeploy vs Survived
7481
1037
0
143
2556
Survived
Not_Survived
survived
count
0
1000
2000
3000
4000
5000
6000
7000
0.0
1.0
abcat_nodeploy


abcat_unavail vs Survived
6553
511
0
669
3484
Survived
Not_Survived
survived
count
0
1000
2000
3000
4000
5000
6000
0.0
1.0
abcat_unavail


occRole_driver vs Survived
Survived
Not_Survived
survived
count
8000
7000
6000
5000
4000
3000
2000
1000
0
289
0
2142
891
7895
0.0
1.0
occRole_driver


occRole_pass vs Survived
Survived
Not_Survived
survived
count
8000
7000
6000
5000
4000
3000
2000
1000
0
891
0
7895
289
2142
0.0
1.0
occRole_pass

i. **Key meaningful observations on individual variables and the relationship between variables**

1. **Individual Variables**
   - **dvcat (Estimated Impact Speeds):**
     - **Categories:** 1-9km/h, 10-24km/h, 25-39km/h, 40-54km/h, 55+km/h
     - Higher impact speeds (e.g., dvcat_55+) are associated with more severe injuries (injSeverity).
     - Lower impact speeds (e.g., dvcat_1-9km/h) are associated with less severe injuries.
   - **weight:**
     - The weight variable doesn't show significant correlations with other variables, indicating its limited impact on predicting survival or injury severity.
   - **Survived:**
     - **Binary variable**: 0 for Not Survived, 1 for Survived.
     - Used as the target variable in predictive models.
   - **airbag:**
     - **Categories:** none, airbag
     - The presence of airbags (airbag) is moderately associated with higher survival rates and slightly lower injury severity.
   - **seatbelt:**
     - **Categories:** none, belted
     - Wearing a seatbelt (belted) is associated with a higher survival rate and lower injury severity.
     - Not wearing a seatbelt (none) is associated with higher injury severity.
   - **frontal:**
     - **Binary variable:** 0 for non-frontal impact, 1 for frontal impact.
     - Frontal impacts are moderately associated with more severe injuries.

- **sex:**
  - **Categories:** f for Female, m for Male.
  - Males (m) have slightly higher injury severity compared to females (f).
- **ageOFocc (Age of Occupant):**
  - Continuous variable representing the age of the occupant.
  - Older occupants tend to have higher injury severity.
- **yearacc (Year of Accident):**
  - Continuous variable representing the year of the accident.
  - Newer vehicles are slightly more likely to be involved in recent accidents.
- **yearVeh (Year of Vehicle Model):**
  - Continuous variable representing the year of the vehicle model.
  - Newer vehicles are associated with recent accidents.
- **abcat (Airbag Deployment Status):**
  - **Categories:** deploy, nodeploy, unavail
  - Airbag deployment (deploy) is associated with higher injury severity, indicating deployment in more severe crashes.
  - Lack of airbag deployment (nodeploy, unavail) is associated with lower injury severity.
- **occRole (Occupant Role):**
  - **Categories:** driver, pass (passenger)
  - No significant difference in injury severity between drivers and passengers.
- **injSeverity:**
  - Numeric scale from 0 to 6 indicating injury severity.
  - Higher values indicate more severe injuries.

Relationships Between Variables

- **Injury Severity and Airbag Deployment:**
  - There is a moderate positive correlation between injSeverity and deploy (0.037), indicating that airbags tend to deploy in more severe accidents.
- **Injury Severity and Impact Speed:**

- Higher impact speeds (dvcat_55+) have a moderate positive correlation with injSeverity, indicating that higher speeds result in more severe injuries.

- **Injury Severity and Seatbelt Usage:**
    - Not wearing a seatbelt (seatbelt_none) has a moderate positive correlation with injSeverity, while wearing a seatbelt (seatbelt_belted) has a moderate negative correlation with injSeverity.
- **Injury Severity and Frontal Impact:**
    - Frontal impacts (frontal) have a moderate positive correlation with injSeverity, suggesting that frontal impacts are associated with more severe injuries.
- **Injury Severity and Age:**
    - Older occupants tend to experience more severe injuries.
- **Survival and Safety Features:**
    - The presence of airbags and wearing seatbelts are positively associated with survival rates.

2. Data Pre-processing

a. Missing values

```
<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)
```

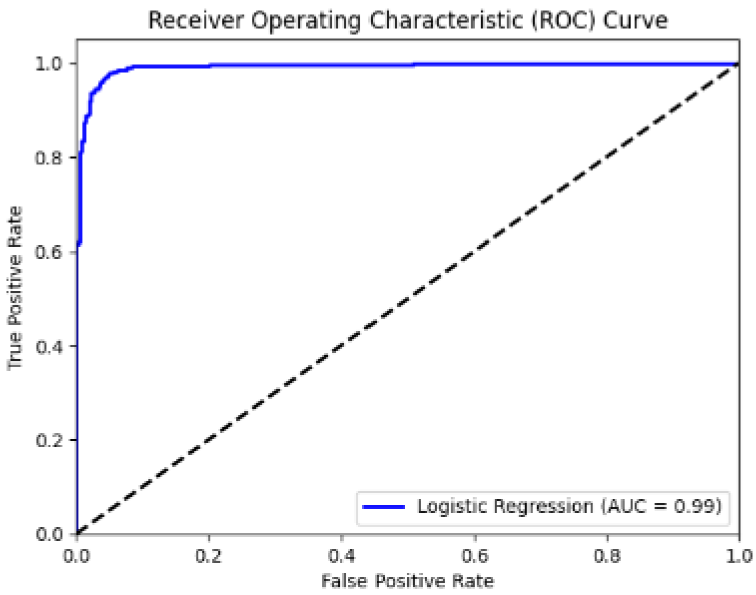
- Missing values identified in the column “injSeverity”, it has been treated with median imputation.

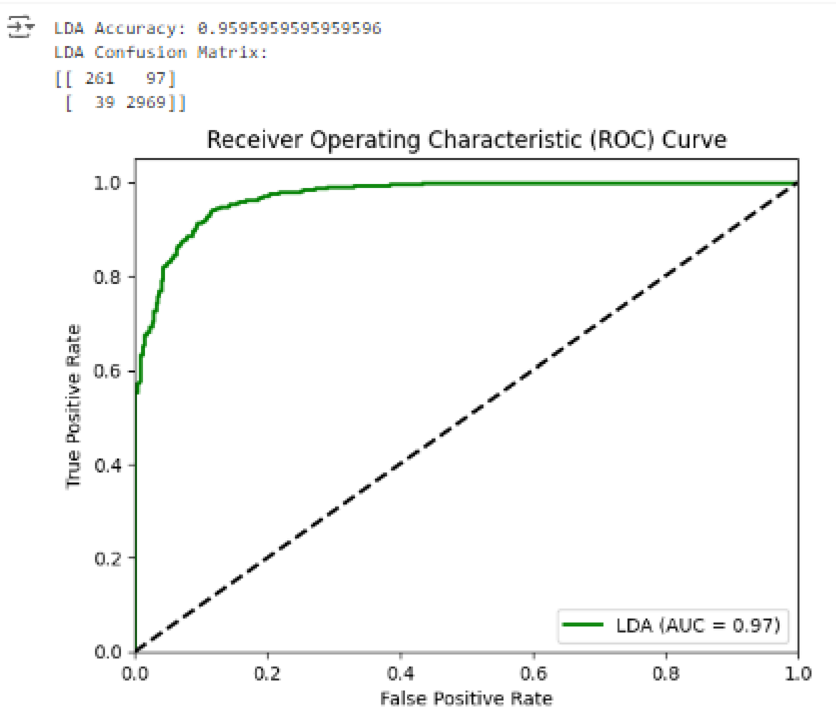
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11217 entries, 0 to 11216
Data columns (total 25 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   weight            11217 non-null  float64
 1   Survived          11217 non-null  object
 2   frontal           11217 non-null  int64
 3   ageOFocc          11217 non-null  int64
 4   yearacc           11217 non-null  int64
 5   yearVeh           11217 non-null  float64
 6   deploy            11217 non-null  int64
 7   injSeverity       11217 non-null  float64
 8   caseid            11217 non-null  object
 9   dvcat_1-9km/h     11217 non-null  float64
 10  dvcat_10-24       11217 non-null  float64
 11  dvcat_25-39       11217 non-null  float64
 12  dvcat_40-54       11217 non-null  float64
 13  dvcat_55+         11217 non-null  float64
 14  airbag_airbag     11217 non-null  float64
 15  airbag_none       11217 non-null  float64
 16  seatbelt_belted   11217 non-null  float64
 17  seatbelt_none     11217 non-null  float64
 18  sex_f             11217 non-null  float64
 19  sex_m             11217 non-null  float64
 20  abcat_deploy      11217 non-null  float64
 21  abcat_nodeploy    11217 non-null  float64
 22  abcat_unavail     11217 non-null  float64
 23  occRole_driver    11217 non-null  float64
 24  occRole_pass      11217 non-null  float64
dtypes: float64(19), int64(4), object(2)
```

- Target variable is encoded.

3. Model Building and Compare the Performance of the Models

```
0
Logistic Regression Accuracy: 0.9818775995246584
Logistic Regression Confusion Matrix:
[[ 317   41]
 [  20 2988]]
```





LDA (Linear Discriminant Analysis):

- The ROC curve for LDA is plotted in green.
- **AUC (Area Under the Curve): 0.97**
  - The AUC score indicates that the LDA model has a very good ability to distinguish between the positive class (Survived) and the negative class (Not Survived).

Logistic Regression:

- The ROC curve for Logistic Regression is plotted in blue.
- **AUC (Area Under the Curve): 0.99**
  - The AUC score is slightly higher than LDA, suggesting that the Logistic Regression model has a slightly better ability to distinguish between the positive and negative classes compared to LDA.

Confusion Matrices

Confusion Matrix Explanation:

- A confusion matrix shows the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for a classification model.

  LDA Confusion Matrix:

  [[261   97]

  [39 2969]]

- **True Positives (TP):** 2969 (Correctly predicted survived cases)
- **True Negatives (TN):** 261 (Correctly predicted not survived cases)
- **False Positives (FP):** 97 (Incorrectly predicted survived cases when they were not survived)

- **False Negatives (FN):** 39 (Incorrectly predicted not survived cases when they were survived)

  Logistic Regression Confusion Matrix:
  [[317   41]
  [20 2988]]

- **True Positives (TP):** 2988 (Correctly predicted survived cases)
- **True Negatives (TN):** 317 (Correctly predicted not survived cases)
- **False Positives (FP):** 41 (Incorrectly predicted survived cases when they were not survived)
- **False Negatives (FN):** 20 (Incorrectly predicted not survived cases when they were survived)

Model Accuracy

LDA Accuracy: 0.96 (approximately 95.96%)

- Indicates that the LDA model correctly predicted the survival status in 95.96% of the cases.

Logistic Regression Accuracy: 0.98 (approximately 98.18%)

- Indicates that the Logistic Regression model correctly predicted the survival status in 98.18% of the cases.

Summary

- **ROC Curve and AUC:**
  - The Logistic Regression model has a slightly higher AUC (0.99) compared to the LDA model (0.97), indicating better performance in distinguishing between classes.
- **Confusion Matrix:**
  - Logistic Regression has fewer False Positives (41) and False Negatives (20) compared to LDA

4. Business Insights & Recommendations

Steps Performed in the Project
Problem Definition:

The goal was to predict whether a person would survive a car crash based on provided data and identify important factors affecting survival rates and injury severity.

Data Preprocessing:

Data Loading: Loaded the dataset and dropped the first column (assumed to be an identifier).

Handling Missing Values: Filled missing values in injSeverity with the median value.
Encoding Target Variable: Converted the Survived column to binary values (0 for Not Survived, 1 for Survived).
Categorical Encoding: Used one-hot encoding for categorical variables like dvcat, airbag, seatbelt, sex, abcat, and occRole.

Exploratory Data Analysis (EDA):

Statistical Summary: Generated summary statistics for numerical and categorical variables to understand data distribution and central tendencies.
Univariate Analysis: Examined individual variables using statistical summaries and visualizations.
Multivariate Analysis: Explored relationships between variables using correlation heatmaps and pair plots.
Data Visualization:

Correlation Heatmap: Visualized the correlation between numerical variables to identify significant relationships.
Distribution Plots: Used histograms and bar charts to visualize the distribution of categorical and numerical variables.

Model Building:

Logistic Regression: Built a logistic regression model to predict survival.
Linear Discriminant Analysis (LDA): Built an LDA model for comparison.
Model Evaluation: Assessed models using accuracy, confusion matrix, ROC curve, and ROC-AUC score.

Business Interpretation

Safety Feature Importance:

Airbag Deployment: The analysis shows that airbag deployment is weakly associated with injury severity. This indicates that airbags are deployed in more severe crashes but do not significantly reduce the severity of injuries.
Seatbelt Usage: Wearing seatbelts is strongly associated with lower injury severity and higher survival rates. This highlights the critical role of seatbelts in protecting occupants during crashes.

Impact of Speed and Frontal Impact:

Impact Speed: Higher speeds (e.g., dvcat_55+) are associated with more severe injuries. This indicates the importance of speed limits and monitoring to prevent high-speed crashes.
Frontal Impact: Frontal impacts are associated with more severe injuries compared to non-frontal impacts, suggesting the need for enhanced safety features in the front of vehicles.

Demographic Factors:

Age and Gender: Older occupants tend to experience more severe injuries, and males have slightly higher injury severity compared to females. This suggests the need for targeted safety campaigns and features catering to these demographics.

Actionable Insights

Enhanced Seatbelt Regulations and Awareness Campaigns:

Promote stricter seatbelt regulations and conduct awareness campaigns to encourage seatbelt usage. This can significantly reduce injury severity and increase survival rates in car crashes.

Focus on Speed Management:

Implement and enforce stricter speed limits, especially in high-risk areas. Utilize speed monitoring and control technologies to prevent high-speed crashes, which are associated with higher injury severity.

Improve Frontal Impact Safety Features:

Encourage car manufacturers to enhance frontal impact safety features, such as crumple zones, advanced airbags, and reinforced structures. This can help reduce the severity of injuries in frontal crashes.

Summary
This project involved analyzing car crash data to predict survival rates and identify key factors influencing injury severity. Key steps included data preprocessing, exploratory data analysis, model building, and evaluation. The analysis revealed critical insights into the importance of safety features like seatbelts and airbags, the impact of speed and frontal crashes, and demographic factors. Actionable insights include promoting seatbelt usage,

managing speeds, and improving frontal impact safety features to enhance occupant safety and reduce injury severity in car crashes.