PREDICTIVE MODELLING

Name: Vamsi Krishna Bommu

Batch: PGPDSBA.APRIL23.B

CONTENTS

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 basis 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.

Exploratory Data Analysis:

- The dataset includes various attributes for 759 firms and having sales as the target variable.
- Checking the null values across all the columns presented in the dataset and getting the type of the values presented in each column for eg., Dtype: int64. There were 21 null values in tobinq columns.

```
No.             0
sales           0
capital         0
patents         0
randd           0
employment      0
sp500           0
tobinq         21
value           0
institutions    0
dtype: int64
```

- Getting the shape of the dataset using data.shape. It consists of 759 rows and 10 columns. And different types of data types were presented in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 759 entries, 0 to 758
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   No.           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
None
(759, 10)
```

- To view the data in the descriptive manner, we use data.describe() method to get the statistical summary of numerical columns, offering insights into mean, median, standard deviation, and distribution range of each numeric attribute. This is important for identifying

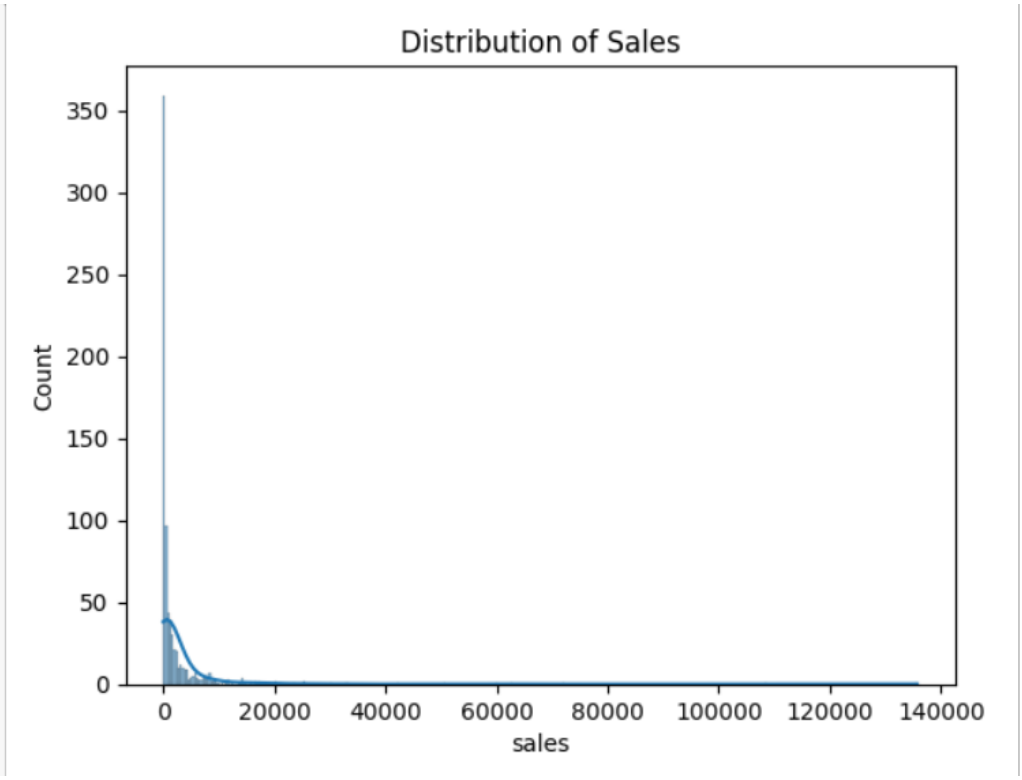
outliers, understanding data spread, and preliminary insights into data normalization needs.

	No.	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

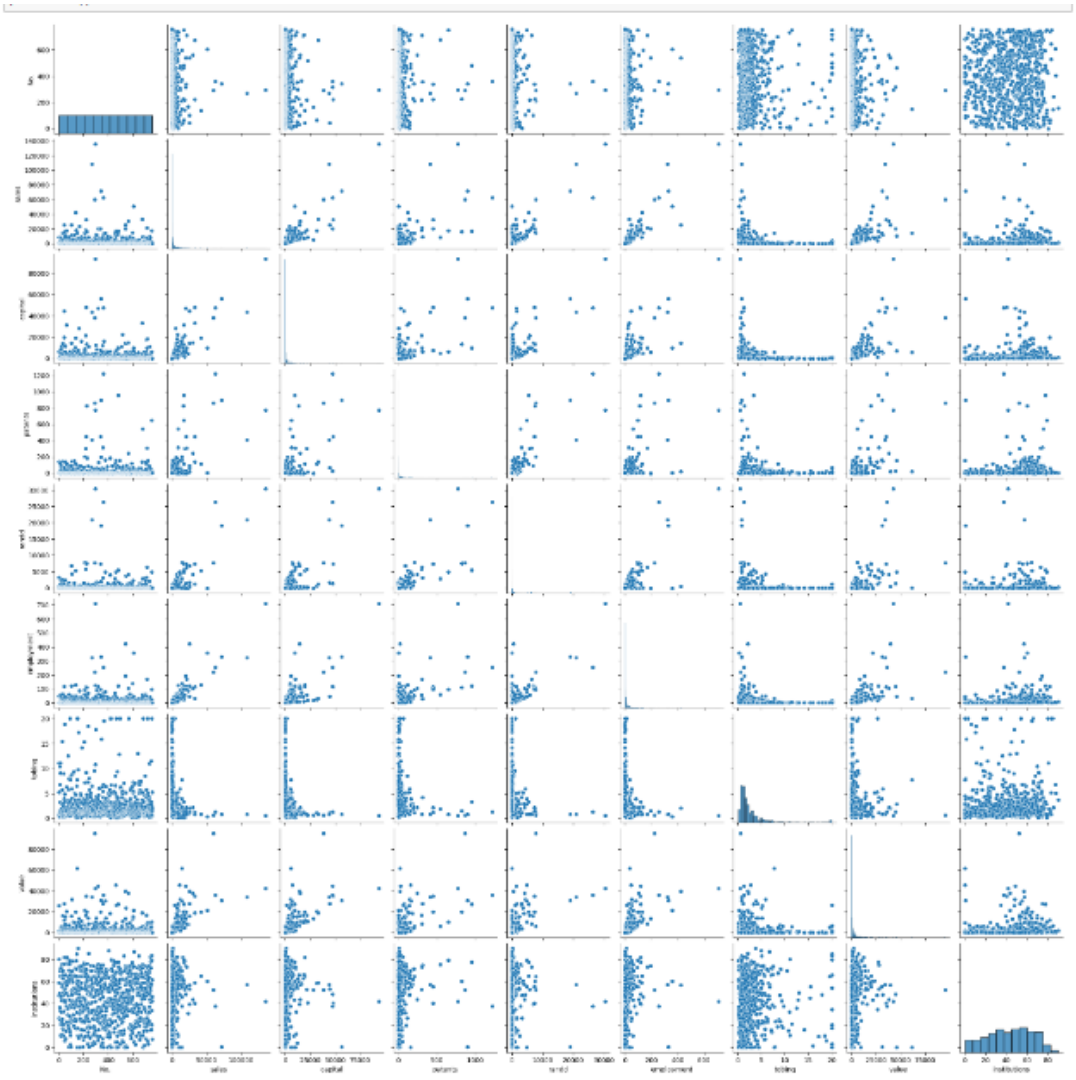
Univariate Analysis:

- A histogram with a kernel density estimate (KDE) for the 'sales' column was plotted, showing the distribution of sales across the firms. This visualization is essential for understanding the distribution's skewness, detecting outliers, and determining the sales performance of firms briefly.



Bivariate Analysis:

- Pairwise relationships between all variables were visualized using sns.pairplot(data), providing insights into the relationships and correlations between sales and other attributes. This is crucial for identifying potential predictors for the regression model.



Handling Null Values:

- Null values were identified and filled with the mean of their respective columns. This approach is used for the continuous numerical variables. However, the approach might vary depending on the variable's nature and distribution.
- The necessity for scaling was not directly addressed. Given that linear regression can be sensitive to the scale of input features, especially when interpreting coefficients, scaling could be beneficial for models involving variables with vastly different scales and units. However, the direct impact of scaling on model performance (R-square, RMSE) was not assessed.

Data Handling:

- Non-numeric values were identified, and a binary encoding was applied to replace 'yes' and 'no' with 1 and 0, respectively. This encoding is suitable for binary categorical variables but might not be adequate for nominal variables with more than two categories if present.

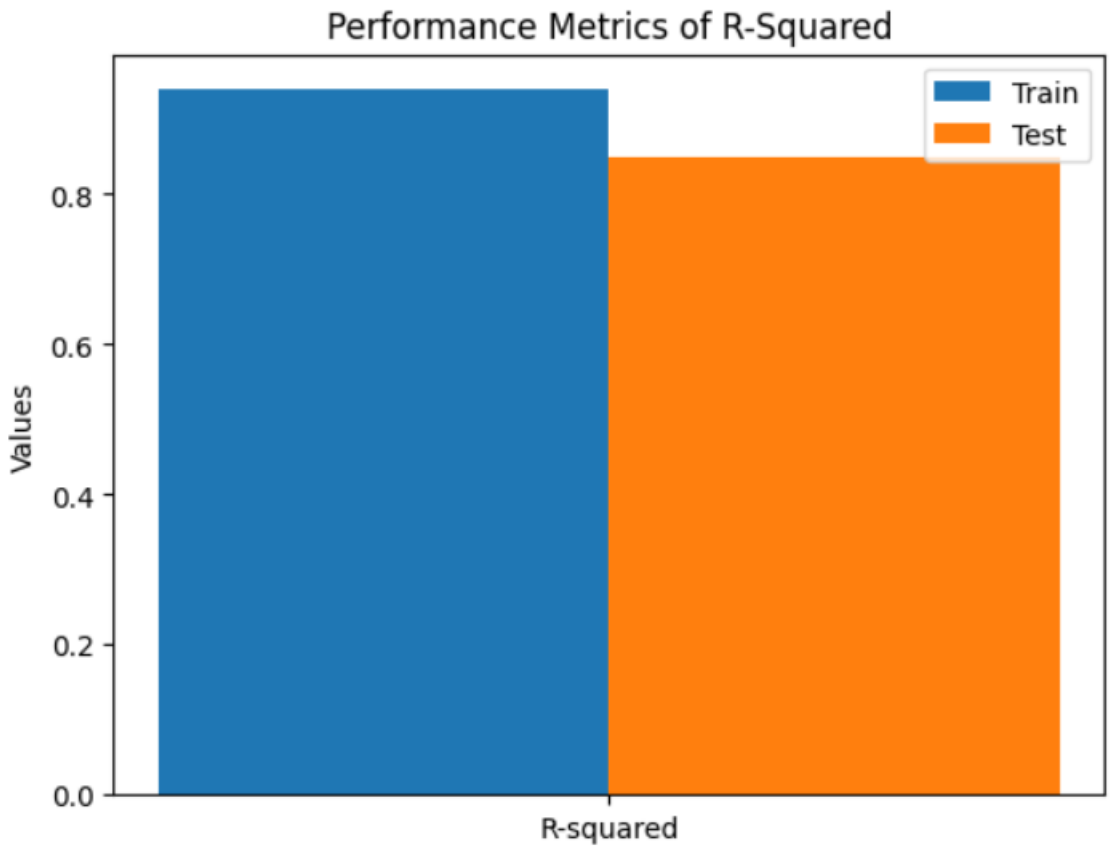
Splitting the Dataset:

- The data was split into a training set (70%) and a testing set (30%), following common practices for model validation.

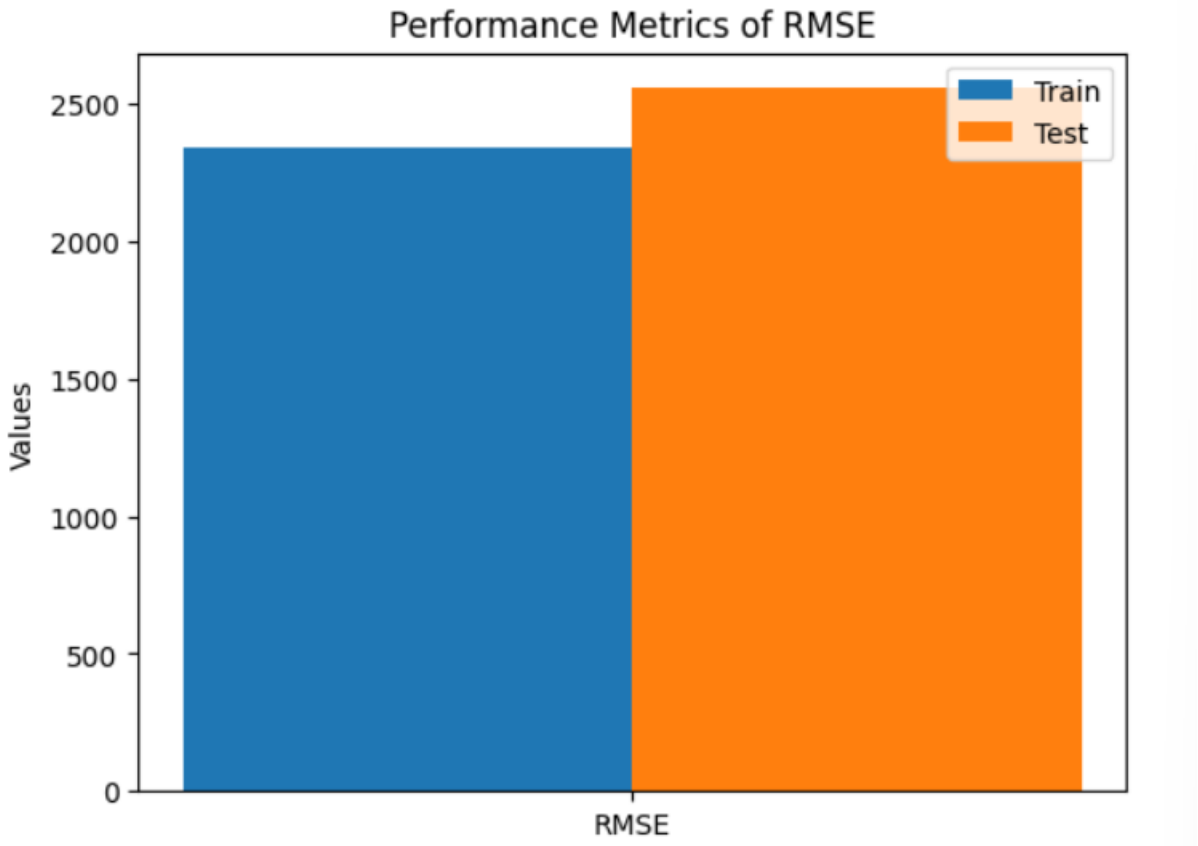
Model Creation:

- A linear regression model was applied and trained, and its performance was evaluated using R-square and RMSE on both training and testing sets.

```
<matplotlib.legend.Legend at 0x20209e17ed0>
```



```
<matplotlib.legend.Legend at 0x20209e20e50>
```



Inference: Business Insights and Recommendations:

- The model demonstrated reasonable predictive ability, with R-square values indicating a decent level of variance explanation in both training and testing sets. However, a slight overfitting was noted, as indicated by a higher R-square on the training set.
- RMSE values provided insights into the average prediction error, essential for assessing model accuracy and reliability.
- Top 5 Attributes: Identifying the five most important attributes based on their coefficients suggests areas where the firm could focus to improve sales or areas of due diligence for investment decisions.

```
Top 5 Important Attributes:
        Feature  Coefficient
3    employment   108.688744
4         sp500    48.088493
7  institutions     4.649914
2         randd     1.695019
0       capital     0.253374
```

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.

Data Ingestion and Exploratory Data Analysis:

- The dataset includes a range of variables associated with car crash incidents.

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

- Describing the dataset using df.describe() to provide the insights into the basic statistics and the structure of the data.

	sno	weight	frontal	ageOFocc	yearacc	yearVeh	deploy	injSeverity
count	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11217.000000	11140.000000
mean	5608.000000	431.405309	0.644022	37.427654	2001.103236	1994.177944	0.389141	1.825583
std	3238.213319	1406.202941	0.478830	18.192429	1.056805	5.658704	0.487577	1.378535
min	0.000000	0.000000	0.000000	16.000000	1997.000000	1953.000000	0.000000	0.000000
25%	2804.000000	28.292000	0.000000	22.000000	2001.000000	1991.000000	0.000000	1.000000
50%	5608.000000	82.195000	1.000000	33.000000	2001.000000	1995.000000	0.000000	2.000000
75%	8412.000000	324.056000	1.000000	48.000000	2002.000000	1999.000000	1.000000	3.000000
max	11216.000000	31694.040000	1.000000	97.000000	2002.000000	2003.000000	1.000000	5.000000

Check Null Values:

- Check the null values using isnull().sum(). The null value check indicated the presence and absence of missing values across all the features, and filling or removing those values to make the dataset clean for further process. In this dataset there is 77 null values presented in the injSeverity column.

```
sno            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
dtype: int64
```

Descriptive Statistics:

- The descriptive statistics revealed the central tendency, dispersion, and shape of the dataset's numerical features.
- This dataset acquires 11217 rows and 16 columns.

```
df.shape
```

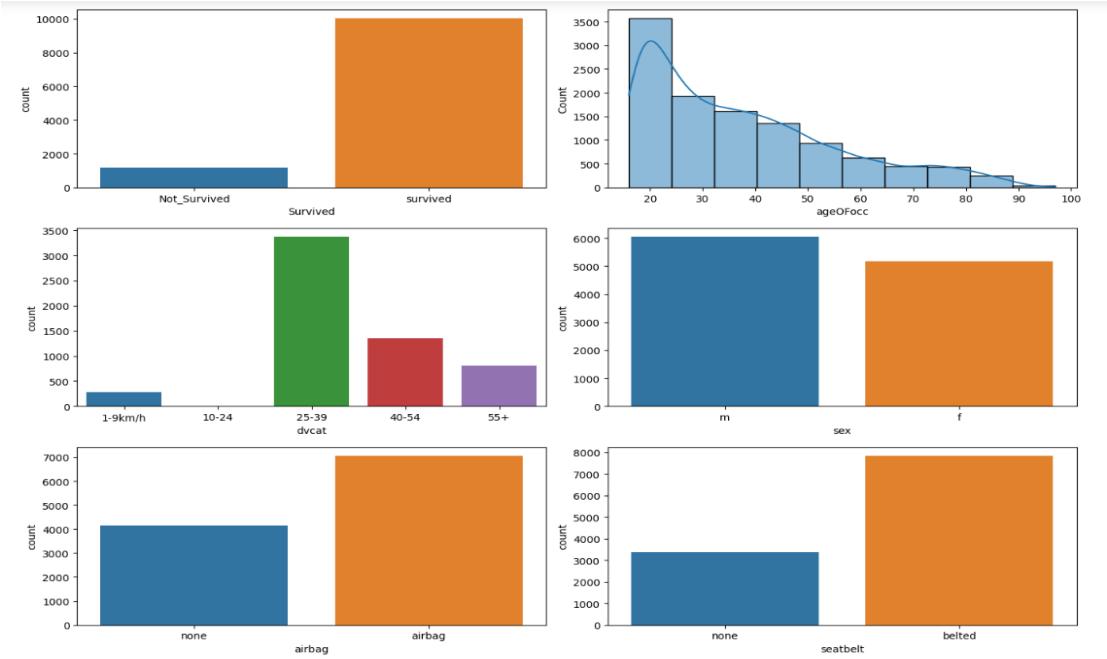
```
(11217, 16)
```

Inference:

- If any null values are presented in the data, then, it must be addressed appropriately to ensure the integrity of the modeling process. The choice of imputation could be based on the nature of missing values and the distribution of the data.

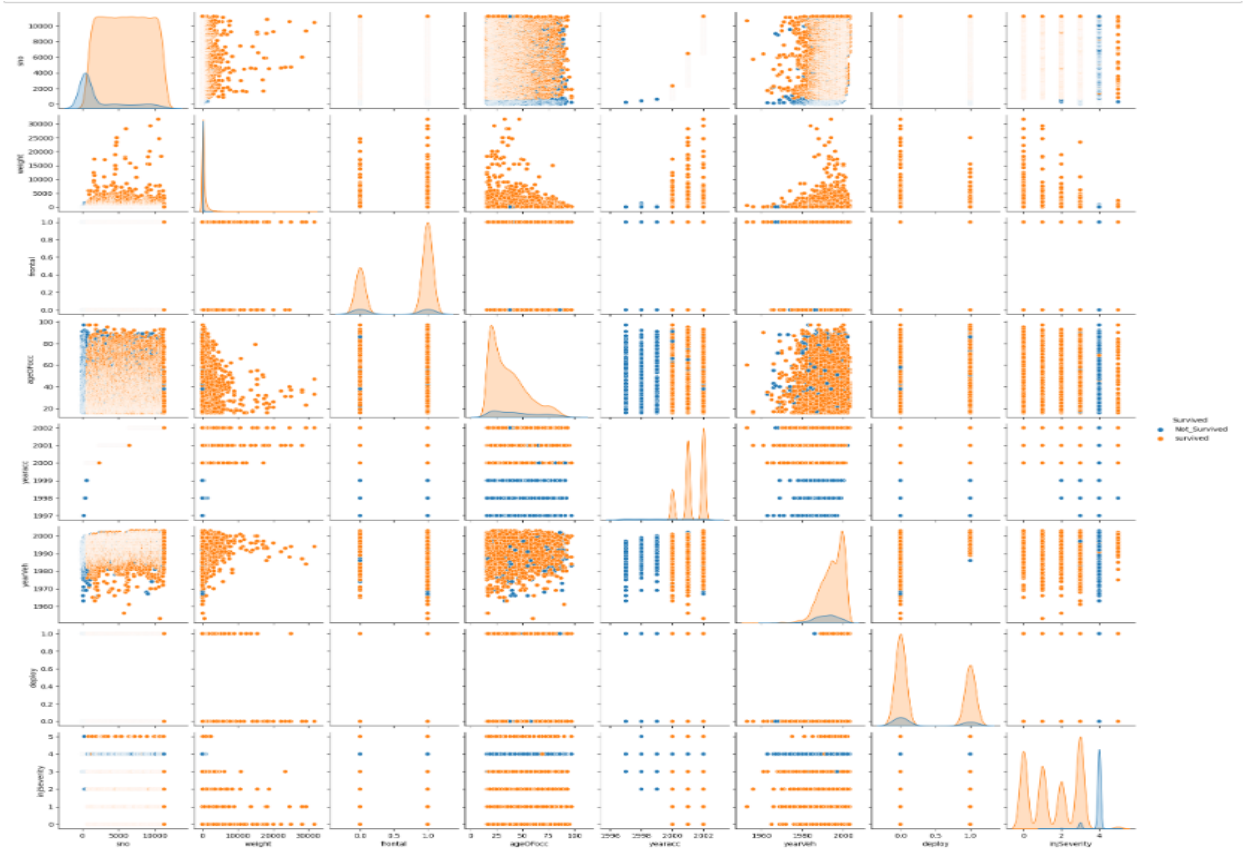
Univariate Analysis:

- In Univariate Analysis, we plotted histograms and count plots provided insights into the distribution of individual variables such as 'age0Focc', 'dvcat', 'sex', 'airbag', and 'seatbelt'. This helped in understanding the skewness, kurtosis, and categorical balance in the dataset.



Bivariate Analysis:

- The pairplot with hue='Survived' allowed us to observe the relationships between pairs of variables and how they correlate with the survival outcome.



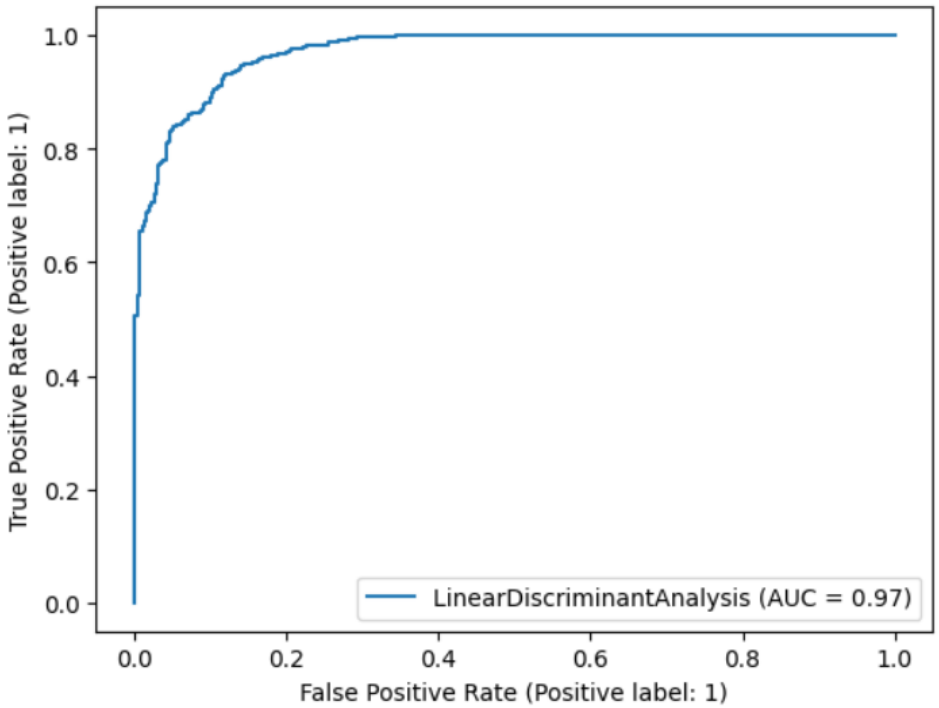
Data Encoding and Splitting of the Data:

- Categorical variables were encoded into numerical formats using 'Label Encoder', facilitating their use in machine learning models.
- Splitting the data into 70 and 30 percent for training and testing purposes. This is performed using train_test_split method from sklearn library.
- This splitting method ensures a balanced approach for training and validation.

Model Creation:

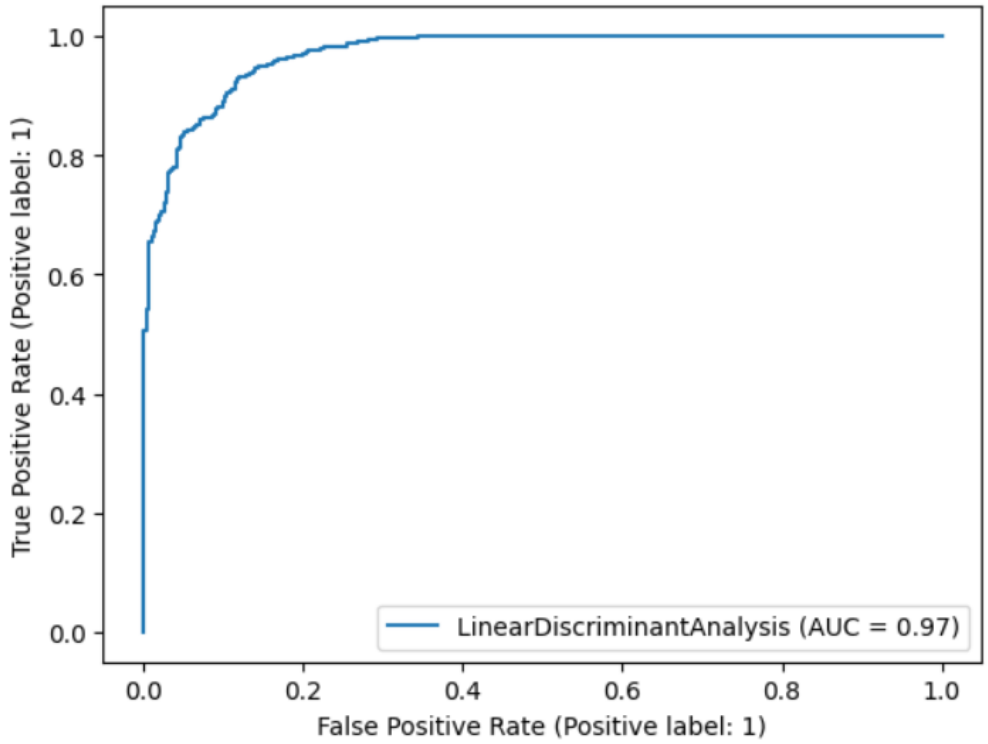
- Both Logistic Regression and Linear Discriminant Analysis were applied to the dataset after handling missing values through mean imputation.
- Logistic Regression and LDA performance were evaluated based on Accuracy, Confusion Matrix, ROC Curve, and ROC_AUC score.
- Both models exhibited comparable accuracy levels on training and test datasets, with the confusion matrix providing insights into true positives, false positives, true negatives, and false negatives.
- ROC Curve and ROC_AUC score further helped in assessing the models' ability to distinguish between the classes.

```
Logistic Regression Performance:
Accuracy on Train Set: 0.9811488982295249
Accuracy on Test Set: 0.9783125371360666
Confusion Matrix on Test Set:
 [[ 332   51]
 [  22 2961]]
```




```
ROC AUC Score on Test Set: 0.989039719419618
```

```
LDA Performance:
Accuracy on Train Set: 0.9574576487071711
Accuracy on Test Set: 0.9581105169340464
Confusion Matrix on Test Set:
 [[ 244  139]
 [   2 2981]]
```




```
ROC AUC Score on Test Set: 0.9706124085220952
```

Inference and Recommendation:

- Safety features such as airbag and seatbelt usage significantly impact survival chances in car crashes.
- The dvcat variable, representing speed categories, shows a strong correlation with survival outcomes, suggesting higher speeds are associated with lower survival rates.
- Enhancing vehicle safety features and promoting their use can significantly improve survival rates.
- Public awareness campaigns highlighting the dangers of high-speed driving could be beneficial.
- Further research could explore more sophisticated models or feature engineering to improve prediction accuracy.