Data Analysis for an Investment Firm and the Government Using Predictive Modeling

Name: Aishwariya Hariharan
PGP-DSBA Online September’ 23
Date: 10 April 2024

Contents

List of Figures

List of Tables

Problem 1

Business Context

You are part of an investment firm and your work is to research 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 based on the details given in the dataset to help your company invest consciously. Also, provide them with 5 most important attributes.

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

Data Description

Data Dictionary

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.

Sample of the Dataset

	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

Table 1: Sample dataset

Data Types

```
 #   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
```

Figure 1: Data type

- Total number of rows = 759
- Total number of columns = 10
    - 7 columns are of float type
    - 2 columns are of integer type
    - 1 column is of the object type
- The data was also checked for duplicate rows. There are no duplicate rows in the dataset.

Statistical Summary of the Dataset

	Unnamed: 0	sales	capital	patents	randd	employment	tobinq	value	institutions
count	759.00000	759.00000	759.00000	759.00000	759.00000	759.00000	738.00000	759.00000	759.00000
mean	379.00000	2689.70516	1977.74750	25.83136	439.93807	14.16452	2.79491	2732.73475	43.02054
std	219.24872	8722.06012	6466.70490	97.25958	2007.39759	43.32144	3.36659	7071.07236	21.68559
min	0.00000	0.13800	0.05700	0.00000	0.00000	0.00600	0.11900	1.97105	0.00000
25%	189.50000	122.92000	52.65050	1.00000	4.62826	0.92750	1.01878	103.59395	25.39500
50%	379.00000	448.57708	202.17902	3.00000	36.86414	2.92400	1.68030	410.79353	44.11000
75%	568.50000	1822.54737	1075.79002	11.50000	143.25340	10.05000	3.13931	2054.16039	60.51000
max	758.00000	135696.78820	93625.20056	1220.00000	30425.25586	710.79993	20.00000	95191.59116	90.15000

Table 2: Statistical summary

Exploratory Data Analysis

Univariate Analysis

- Membership of firms in the S&P 500 index

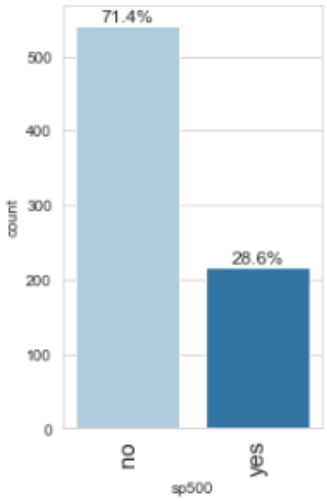


Figure 2: Membership of firms in the S&P 500 index

➔ 71.4% of the firms are not enrolled in the S&P 500 index
➔ Only 28.6% are enrolled.

- Distribution of sales

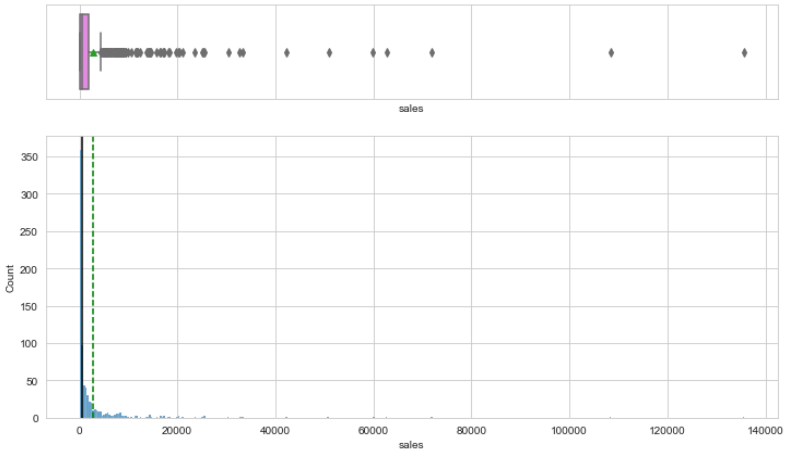


Figure 3: Distribution of sales

➔ Most of the firms have sales value below $20000 million.

- Capital assets' worth: Net stock of property, plant, and equipment.

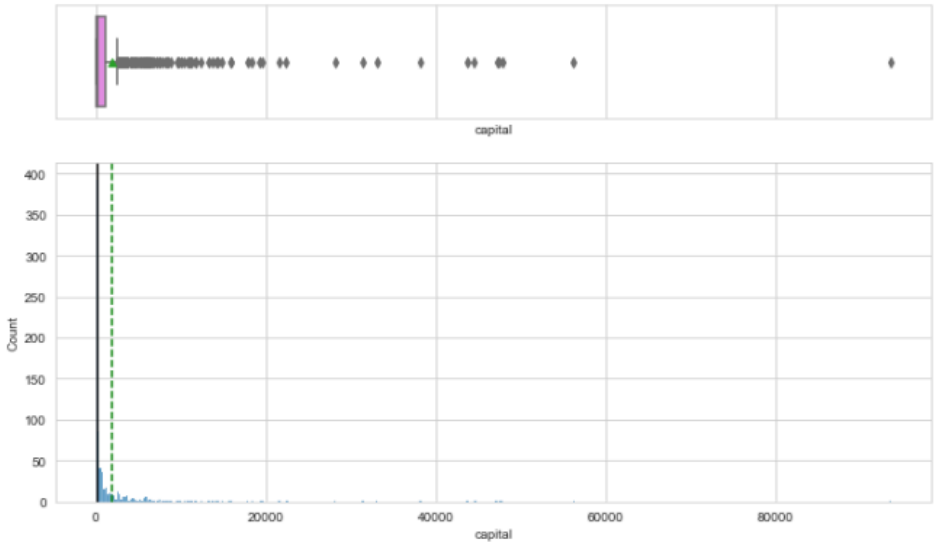


Figure 4: Net stock

➔ The net worth of all stocks put together - property, plant, and equipment - of most firms is below $20000 million.

- Patents

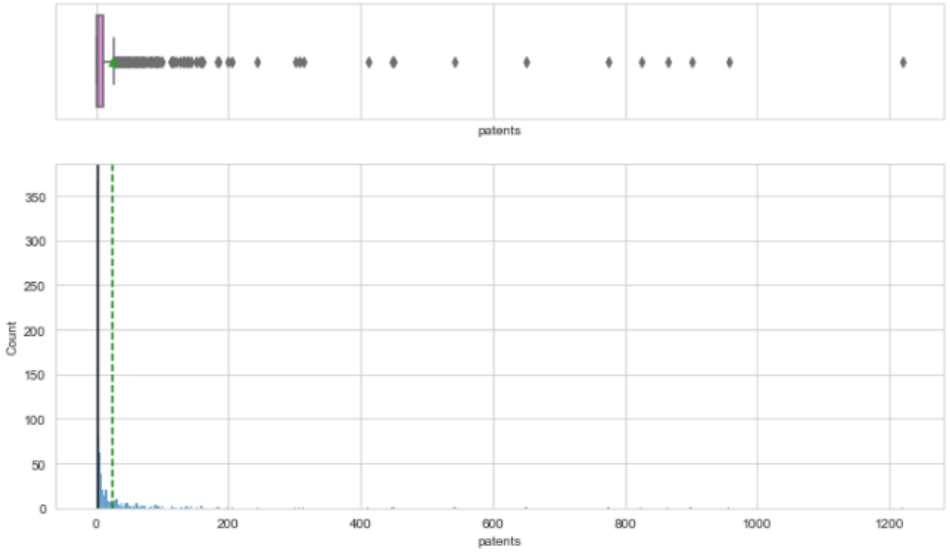


Figure 5: Number of patents

➔ The number of patents owned by most firms is less than 200.

- Randd

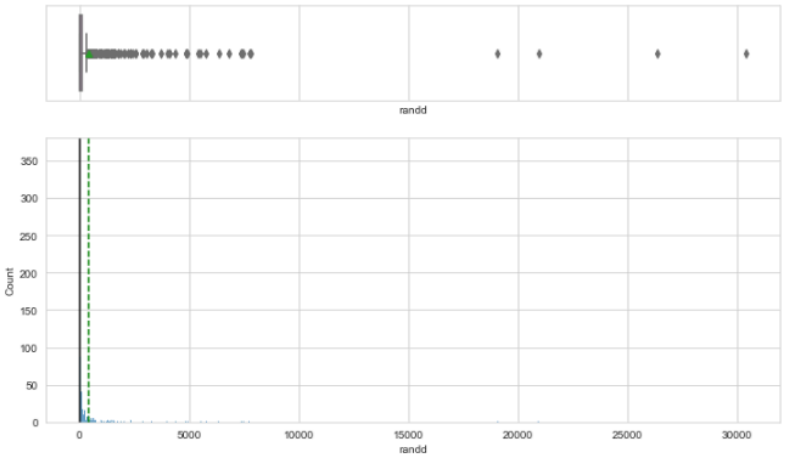


Figure 6: R&D stock worth

➔ Most firms have invested in R&D that's worth way less than $5000 million.

- Employment

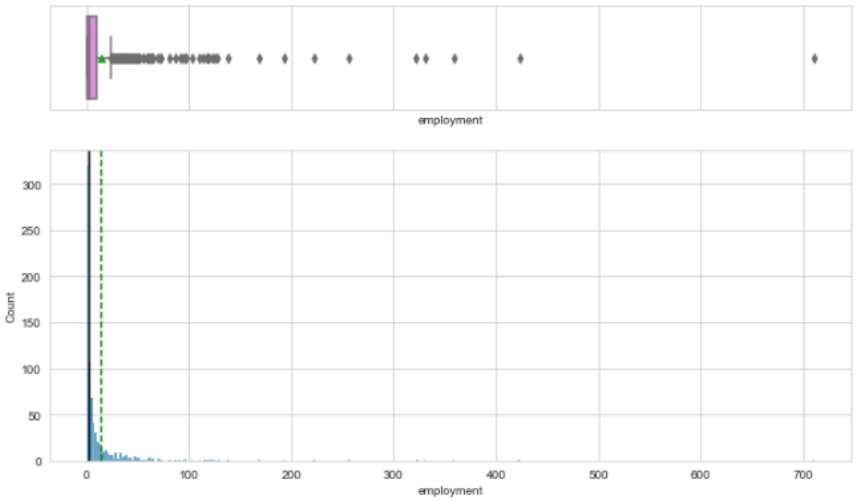


Figure 7: Number of employees

➔ The number of employees in most firms is less than 100 thousand.

- Tobinq

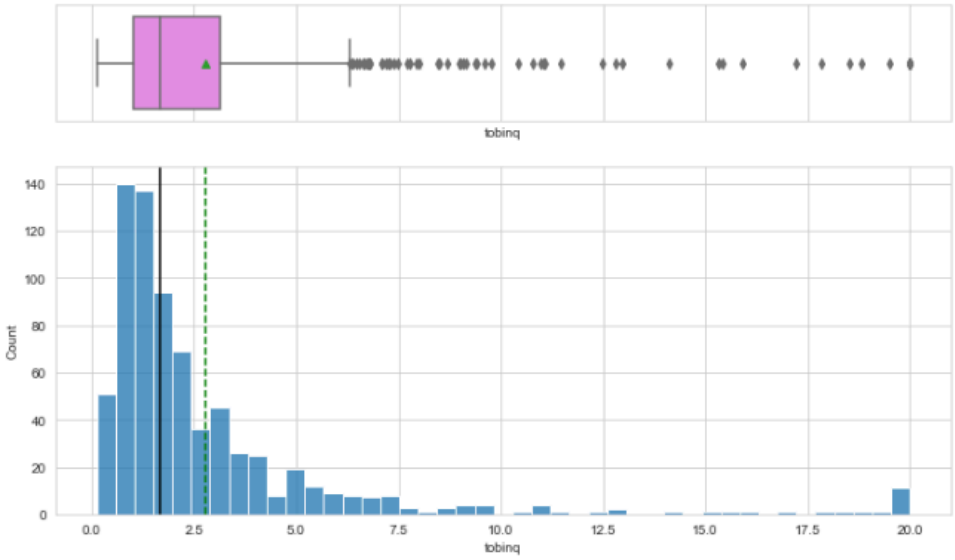


Figure 8: Tobinq

➔ The median of the ratio between a physical asset's market value and its replacement value for the firms surveyed is about 2.5.

➔ This means the physical asset's market value is 2.5 times its replacement value. The existing asset is worth 2.5 times the value of a new asset of a similar kind the firm might have to invest in.

- Value

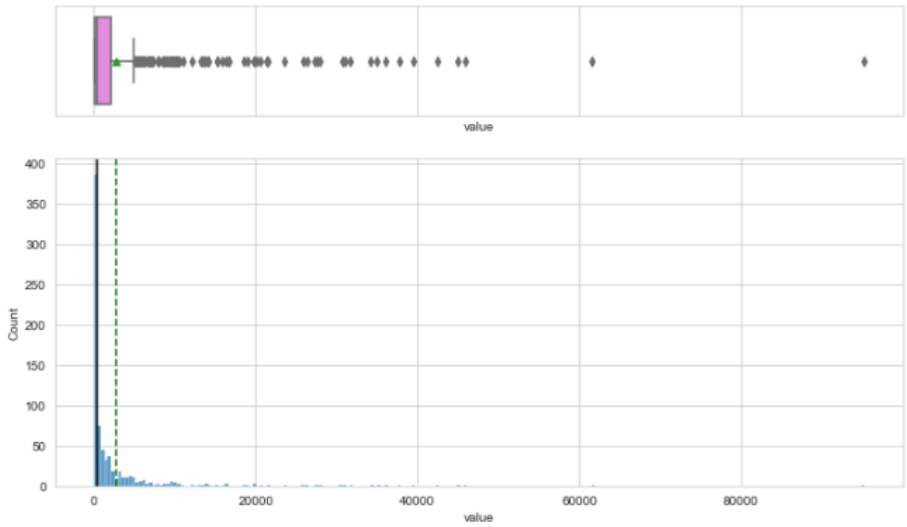


Figure 9: Stock market value of firms

➔ The median of the stock market value of most firms is around $5000 million.

- Institutions

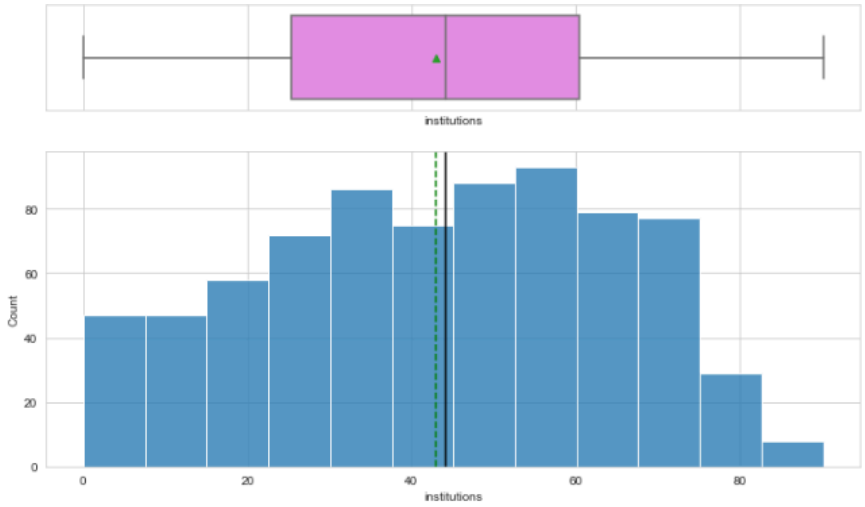


Figure 10: Proportion of stock owned by institutions

➔ The median of proportion of stock owned by institutions is about 40%.

Bivariate Analysis

- Correlation

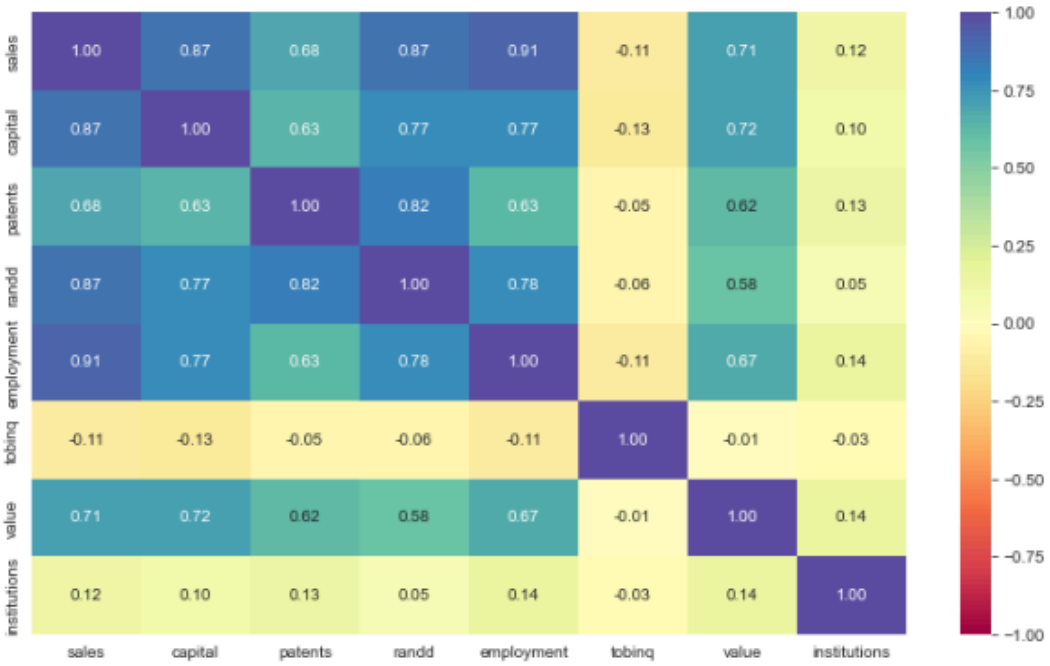


Figure 11: Correlation plot

➔ There's a high correlation between the following:

- ◆ Capital and sales

- Randd and sales
- Employment and sales
- Patents and randd

➔ Therefore, we can say that firms that have a high capital value, good investment in R&D, and greater number of employees have better sales.

➔ More investment in R&D also results in more number of patents granted to the firm and consequently more sales.

- Pairplot

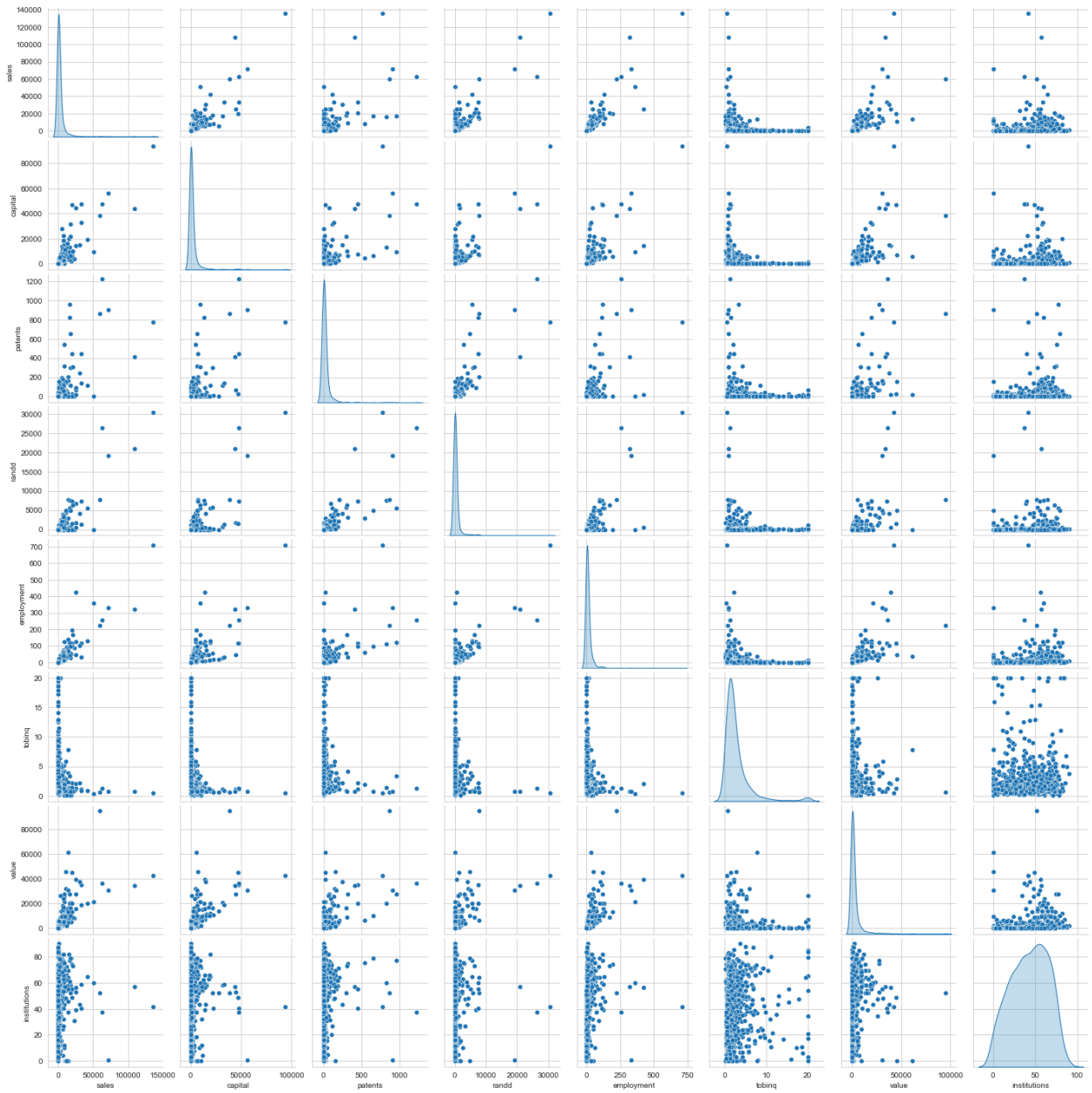


Figure 12: Pairplot

- sp500 vs sales

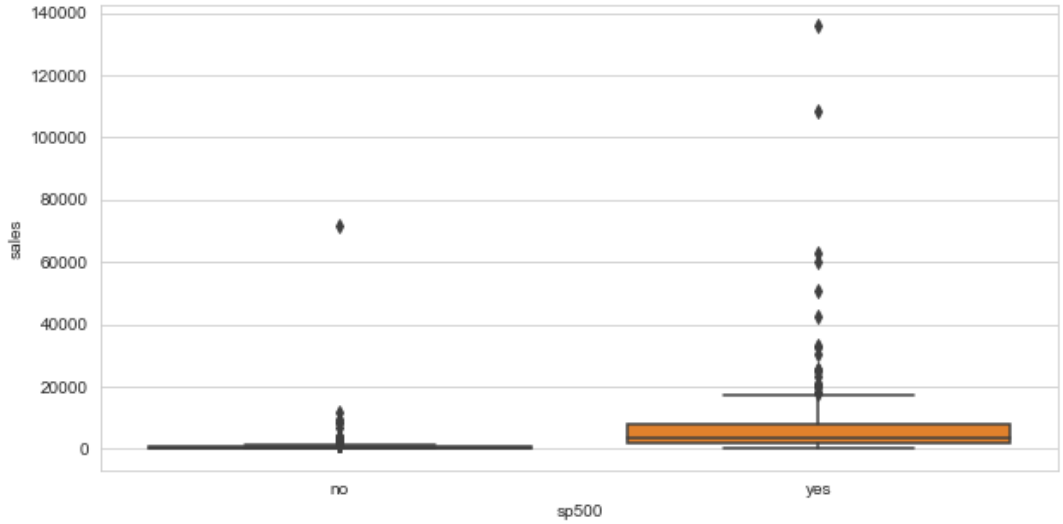


Figure 13: Relationship between sales and enrollment with S&P500

➔ Firms enrolled with the S&P500 have better sales than those not enrolled.

- sp500 vs capital

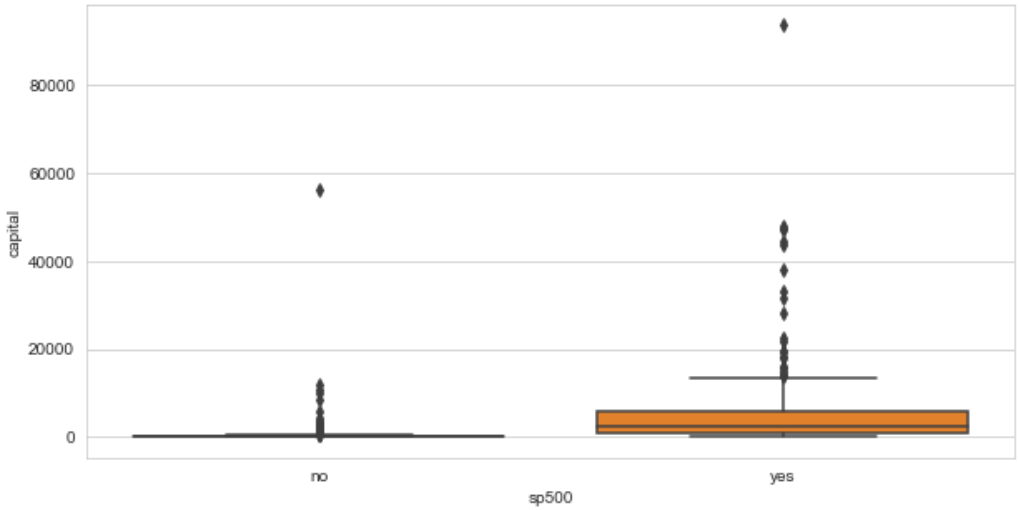


Figure 14: sp500 vs capital

➔ Firms enrolled with the S&P500 are the ones with a high capital value.

- sp500 vs patents

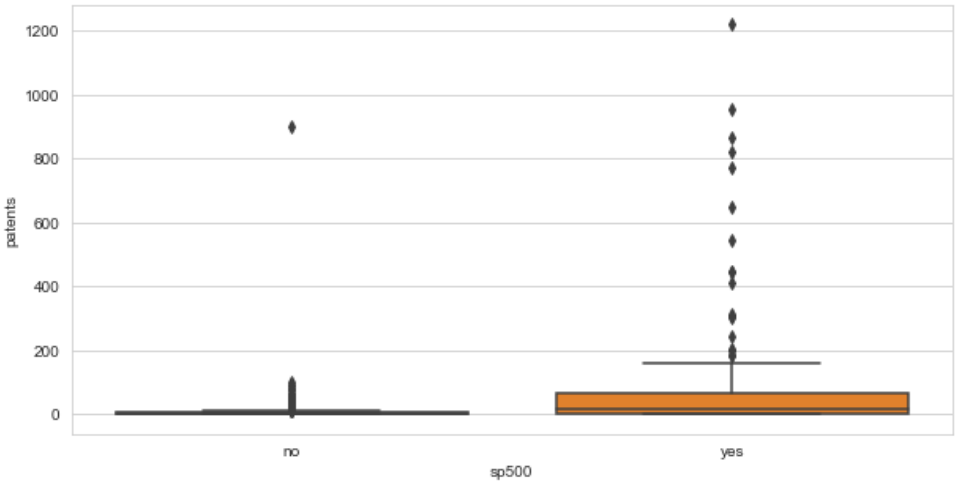


Figure 15: sp500 vs patents

➔ Firms enrolled with the S&P500 have more number of patents than those that are not enrolled.

- sp500 vs randd

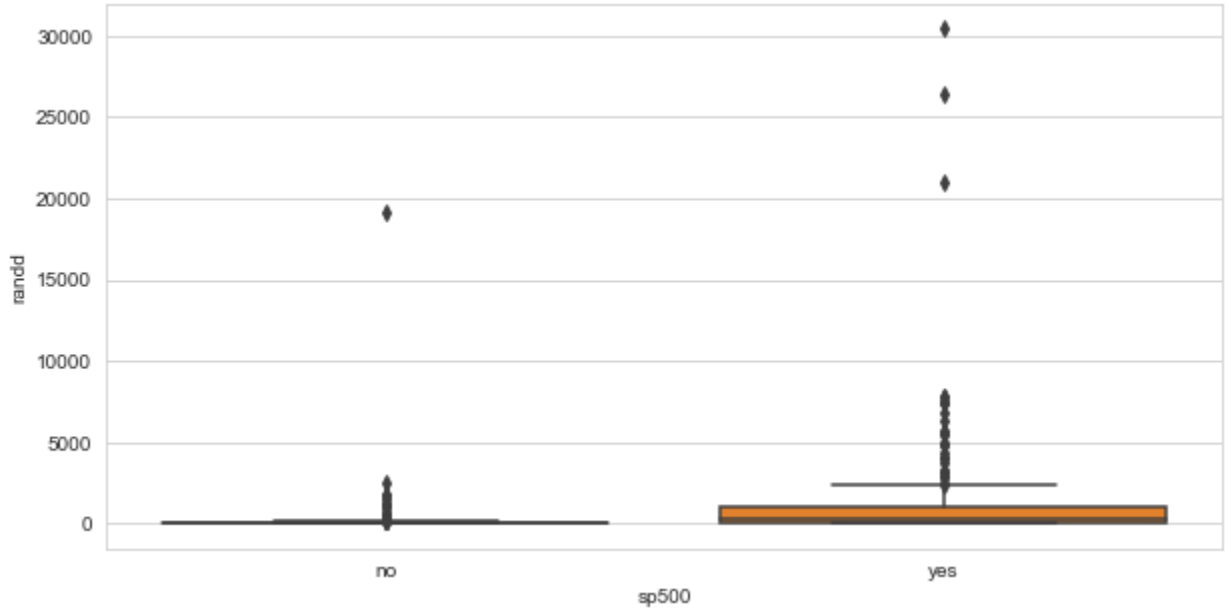
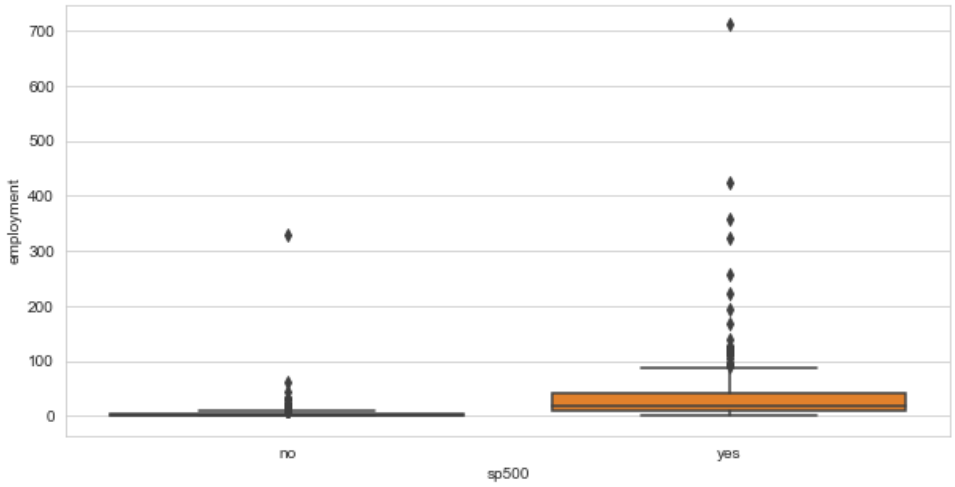


Figure 16: sp500 vs randd

➔ Firms enrolled with the S&P500 are the ones that have higher investment in R&D.

- sp500 vs employment



.

Figure 17: sp500 vs employment

➔ Firms enrolled with the S&P500 have more number of employees.

- sp500 vs value

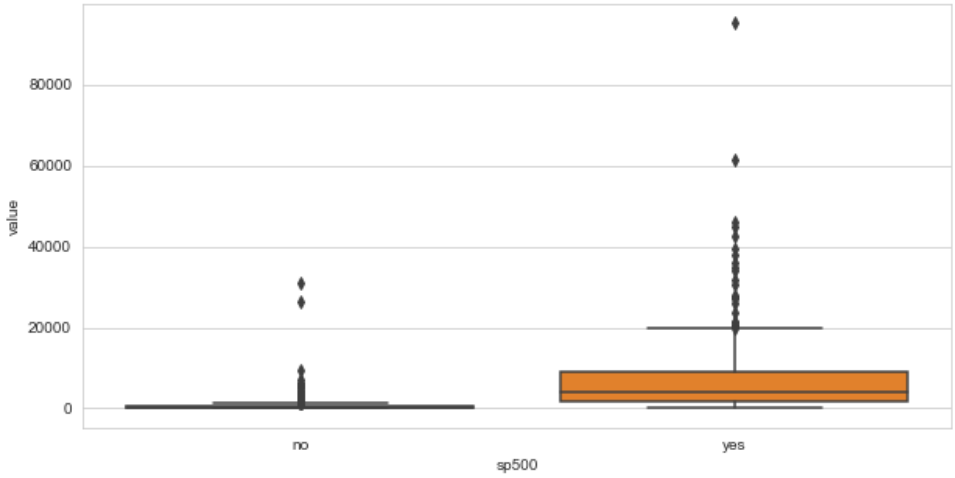


Figure 18: sp500 vs value

➔ Firms enrolled with the S&P500 have a higher stock market value

- sp500 vs institution

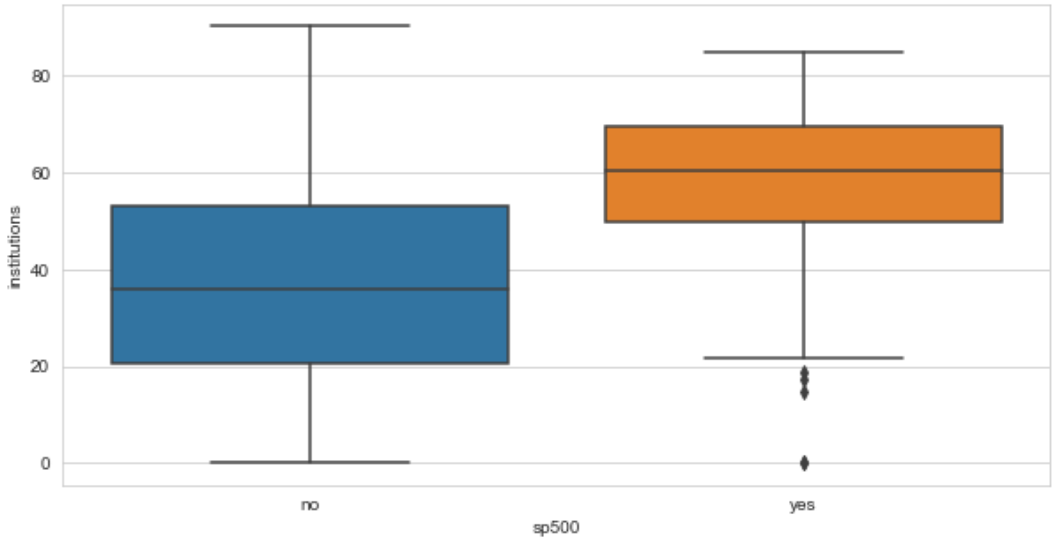


Figure 19: sp500 vs institution

➔ The median proportion of stocks owned by firms enrolled with the S&P500 is 60%, whereas the ones that aren't enrolled have about 35% of stocks ownership.

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

```
Missing Values:
sales            0
capital          0
patents          0
randd            0
employment       0
sp500            0
tobinq          21
value            0
institutions     0
dtype: int64
```

- The tobinq column of the dataset has null values. Hence, there's a need for imputation.
- Scaling is necessary because:

1) The features have vastly different ranges, so scaling is necessary. Features with larger ranges can dominate those with smaller ranges during model training, potentially leading to biased or inefficient models.

2) Algorithms that involve gradient descent, such as linear regression, logistic regression etc., benefit from feature scaling.

- We treat the outliers before scaling

Before treating:

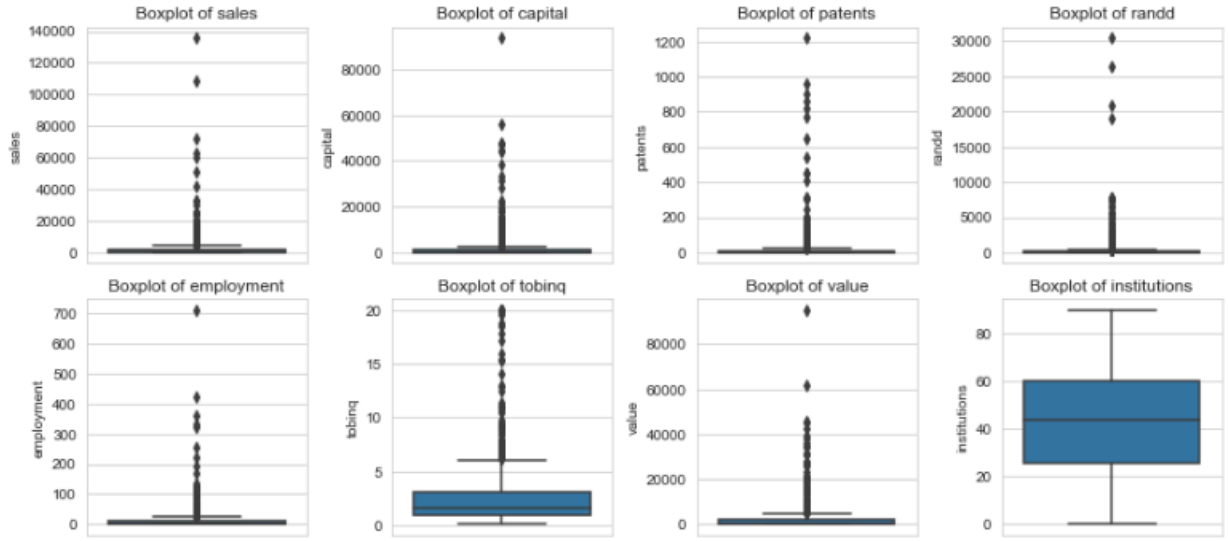


Figure 20: Before treating outliers

After treating:

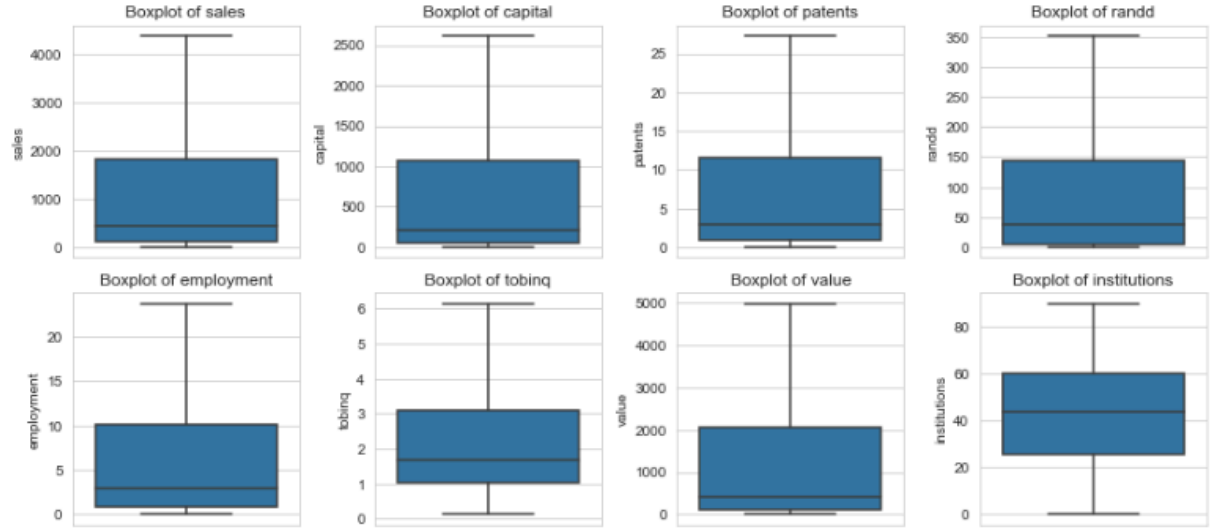


Figure 21: After treating outliers

Scaling using the z-score method

Scaled data

	sales	capital	patents	randd	employment	tobinq	value	institutions
0	-0.26779	-0.59150	0.22115	1.97999	-0.56480	2.23767	0.14260	1.71884
1	-0.54222	-0.63271	-0.58318	-0.78288	-0.61933	-0.84571	-0.64581	0.73828
2	2.05272	1.96272	1.95550	1.97999	2.05512	1.68700	2.05584	0.21593
3	-0.51391	-0.48168	-0.68372	-0.12566	-0.47127	-1.15871	-0.74852	-0.74479
4	-0.69462	-0.61391	-0.58318	-0.67090	-0.60869	-0.71846	-0.74602	0.29714

Table 3: Scaled data

Statistical summary of data after scaling

	count	mean	std	min	25%	50%	75%	max
sales	759.00000	0.00000	1.00066	-0.80904	-0.72867	-0.51549	0.38389	2.05272
capital	759.00000	0.00000	1.00066	-0.76000	-0.70514	-0.54918	0.36200	1.96272
patents	759.00000	0.00000	1.00066	-0.78426	-0.68372	-0.48264	0.37196	1.95550
randd	759.00000	-0.00000	1.00066	-0.78288	-0.74647	-0.49286	0.34411	1.97999
employment	759.00000	0.00000	1.00066	-0.84601	-0.73335	-0.48924	0.38204	2.05512
tobinq	759.00000	0.00000	1.00066	-1.26686	-0.73431	-0.36013	0.45448	2.23767
value	759.00000	0.00000	1.00066	-0.78334	-0.72538	-0.55017	0.38711	2.05584
institutions	759.00000	0.00000	1.00066	-1.98514	-0.81331	0.05027	0.80703	2.17474

Table 4: Statistical summary of data after scaling

1.3) Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE.

Data Preparation for Modelling

We encode categorical feature sp500 and split the data into train and test (70:30) to evaluate the model we build on the train data.

The variable of interest is 'sales'.

Linear Regression using statsmodel(OLS)

Regression summary:

OLS Regression Results

Dep. Variable:	sales	R-squared (uncentered):	0.936
Model:	OLS	Adj. R-squared (uncentered):	0.935
Method:	Least Squares	F-statistic:	950.4
Date:	Tue, 02 Apr 2024	Prob (F-statistic):	7.03e-306
Time:	06:53:16	Log-Likelihood:	-35.777
No. Observations:	531	AIC:	87.55
Df Residuals:	523	BIC:	121.8
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
capital	0.2540	0.026	9.625	0.000	0.202	0.306
patents	-0.0292	0.018	-1.611	0.108	-0.065	0.006
randd	0.0547	0.019	2.829	0.005	0.017	0.093
employment	0.4275	0.025	16.978	0.000	0.378	0.477
tobinq	-0.0466	0.014	-3.420	0.001	-0.073	-0.020
value	0.2927	0.029	10.255	0.000	0.237	0.349
institutions	0.0054	0.013	0.426	0.670	-0.020	0.030
sp500_encoded	0.0556	0.029	1.888	0.060	-0.002	0.114

Omnibus:	192.253	Durbin-Watson:	1.957
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1326.633
Skew:	1.412	Prob(JB):	8.42e-289
Kurtosis:	10.210	Cond. No.	6.47

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 22: Regression summary

- ➔ The R^2 value is 0.936, i.e. 93% of the variance in the sales is explained by the independent variables.
- ➔ The adjusted R^2 value is not very different from the R^2 value since there aren't too many predictors.

- ➔ Except for patents, institutions, and sp500-encoded, all predictors have low p-values, $p < 0.05$. Hence, these 3 variables are not very statistically significant.
- ➔ The standard errors for the variables are low. This indicates more precise estimates.

VIF (Variance Inflation Factor) values

```
const           2.19480
capital         5.66489
patents         2.65796
randd           2.94361
employment      5.26334
tobinq          1.42652
value           6.70601
institutions    1.28667
sp500_encoded   3.05160
dtype: float64
```

Figure 23: VIF values

Patents, randd, tobinq, institutions, and sp500_encoded have VIF values below 5, indicating low multicollinearity.

Capital, employment, and value have VIF values between 5 and 10, suggesting moderate multicollinearity.

Original $R^2 = 0.936$

Original adjusted $R^2 = 0.935$

Dropped variable	R^2	Adjusted R^2
Capital	0.925	0.924
Employment	0.903	0.901
Value	0.925	0.924

If the dropped variable did not contribute much unique information to the model or was conceptually redundant with other variables, its removal may have a minimal impact on the R-squared value.

After dropping the features causing strong multicollinearity and the statistically insignificant ones, our model performance (R-square values, Cond No., etc) hasn't dropped sharply.

This shows that these variables did not have much predictive power.

Considering the moderate decrease in both R^2 and adjusted R^2 values, and the potential loss of interpretability associated with dropping important variables, it may not be advisable to drop any variable from the model based solely on these considerations.

Testing the Assumptions of Linear Regression

For Linear Regression, we need to check if the following assumptions hold:

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality of error terms
5. No strong Multicollinearity

Figure 24: Fitted vs Residual plot for Linear Regression

- The graph shows no pattern in the data. Thus the assumption of linearity and independence of predictors are satisfied.

- From the Shapiro-Wilk test, we get p-value=8.61×10^{-25}
    - Since p-value < 0.05, the residuals are not normal.

- From the test for Homoscedasticity, we get p = 0.6899
    - Since p-value > 0.05 we can say that the residuals are homoscedastic.

Most of the assumptions are satisfied, hence we can fetch the linear regression equation to model the data using a linear regression equation.

Linear regression equation

sales = -0.028786883868058888 + 0.25385092961544264 * (capital) + -0.03000236237465888 * (patents) + 0.05315127066829384 * (randd) + 0.4211458601045237 * (employment) + -0.04609218846465864 * (tobinq) + 0.2819938049088843 * (value) + 0.11173693955527303 * (sp500_encoded)

Observations

- 1 unit increase in the capital leads to a 0.25 times increase in the sales value.
- 1 unit increase in the no.of patents leads to a 0.03 times decrease in the sales value.
- 1 unit increase in investment in R&D stock leads to 0.053 times increase in sales value.
- 1 unit increase in employees (in 1000s) leads to 0.42 times increase in sales value.

RMSE Values:

- RMSE value on train data = 0.258
- RMSE value on test data = 0.262

The Root Mean Squared Errot value indicates the average magnitude of prediction errors in a regression model. Lower RMSE values indicate better predictive accuracy. For both test and train data, the RMSE values are low.

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

Actionable Insights & Recommendations

- Due to the strong correlation between variables such as capital, randd, employment and the variable of interest i.e. sales, firms need to increase their investment in capital assets, research and development, and hire more employees as they expand their business.
- Investment in R&D, patents, and so are other variables and their link to the enrollment of the firm to the S&P500 proves that companies that are doing well with respect to their sales are companies that are enrolled with the S&P500 index that rate companies globally.
- Since the R and R^2 values are nearly 0.9 and odd, 90% of the variance in sales is explained by the existing predictors, which is a fairly good score, we need not drop any variable. It also indicates a good fit for the model.
- The p-values, RMSE values, standard deviations, etc, all indicate that the model is a good fit for the given data.
- Attributes that are the most important are - capital, employment, value, and sp500_encoded.

Problem 2

Business Context

You are hired by the Government to analyze 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 based on the information given in the data set to provide insights that will help the government make stronger laws for car manufacturers to ensure safety measures. Also, find out the important factors based on which you made your predictions.

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

Data Description

Data Dictionary

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

Sample of the Dataset

	Unnamed: 0	dvcat	weight	Survived	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
0	0	55+	27.07800	Not_Survived	none	none	1	m	32	1997	1987.00000	unavail	driver	0	4.00000	2:13:2
1	1	25-39	89.62700	Not_Survived	airbag	belted	0	f	54	1997	1994.00000	nodeploy	driver	0	4.00000	2:17:1
2	2	55+	27.07800	Not_Survived	none	belted	1	m	67	1997	1992.00000	unavail	driver	0	4.00000	2:79:1
3	3	55+	27.07800	Not_Survived	none	belted	1	f	64	1997	1992.00000	unavail	pass	0	4.00000	2:79:1
4	4	55+	13.37400	Not_Survived	none	none	1	m	23	1997	1986.00000	unavail	driver	0	4.00000	4:58:1

Table 5: Sample of dataset (Problem 2)

Data Types

```
 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
```

Figure 25: Data types (Problem 2)

- Total number of rows = 11217
- Total number of columns = 16
    - 3 columns are of float type
    - 5 columns are of integer type
    - 8 column is of the object type
- The data was also checked for duplicate rows. There are no duplicate rows in the dataset.

Statistical Summary of the Dataset

	Unnamed: 0	weight	frontal	ageOFocc	yearacc	yearVeh	deploy	injSeverity
count	11217.00000	11217.00000	11217.00000	11217.00000	11217.00000	11217.00000	11217.00000	11140.00000
mean	5608.00000	431.40531	0.64402	37.42765	2001.10324	1994.17794	0.38914	1.82558
std	3238.21332	1406.20294	0.47883	18.19243	1.05681	5.65870	0.48758	1.37854
min	0.00000	0.00000	0.00000	16.00000	1997.00000	1953.00000	0.00000	0.00000
25%	2804.00000	28.29200	0.00000	22.00000	2001.00000	1991.00000	0.00000	1.00000
50%	5608.00000	82.19500	1.00000	33.00000	2001.00000	1995.00000	0.00000	2.00000
75%	8412.00000	324.05600	1.00000	48.00000	2002.00000	1999.00000	1.00000	3.00000
max	11216.00000	31694.04000	1.00000	97.00000	2002.00000	2003.00000	1.00000	5.00000

Table 6: Statistical summary of dataset (Problem 2)

Checking for null/missing values

The injSeverity column has missing values. So we impute them with the median value.

```
Missing Values:
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
dtype: int64
```

Exploratory Data Analysis

Univariate Analysis

- Impact of different levels of speed (count)

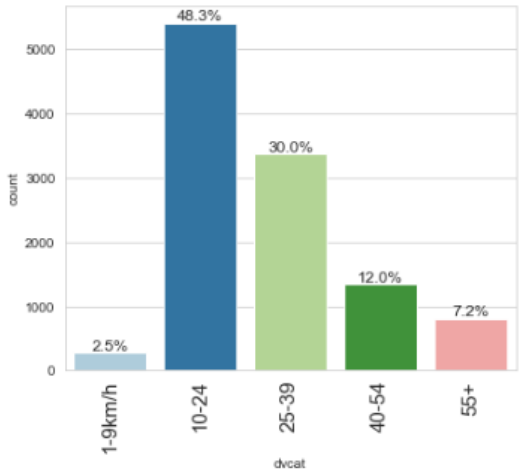


Figure 26: Impact of different levels of speed

➔ Most people drive at a speed of range 10-24 km/h.

- Observation on survival count

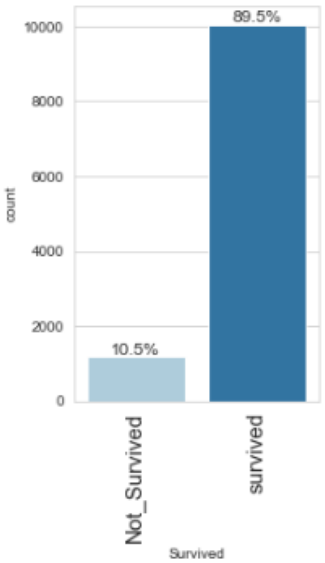


Figure 27: Survival count

➔ 89.5% of people surveyed survived the car crash.

- Observation on the presence of airbags

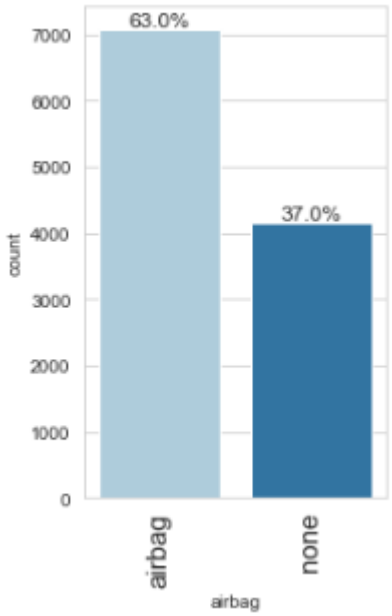


Figure 28: Presence of airbags

➔ 63% of the cars that were involved in the crash had airbags installed in them.

- Observation on the usage of seatbelts

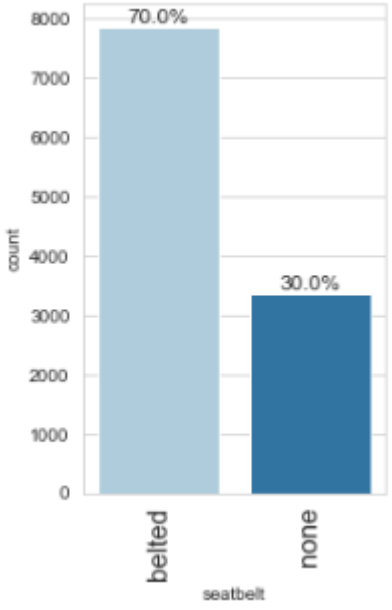


Figure 29: Usage of seatbelts

➔ 70% of the cases had their seat belts on.

- Observation (count) on the sex of the occupant

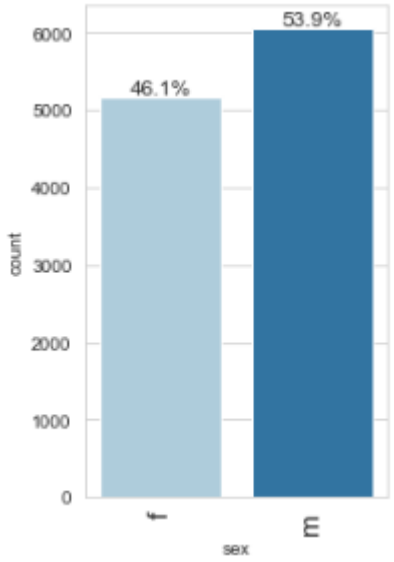


Figure 30: Sex of the occupant

➔ 54% of the occupants were male and 46% female.

- Observation on the deployment of airbags by the occupant

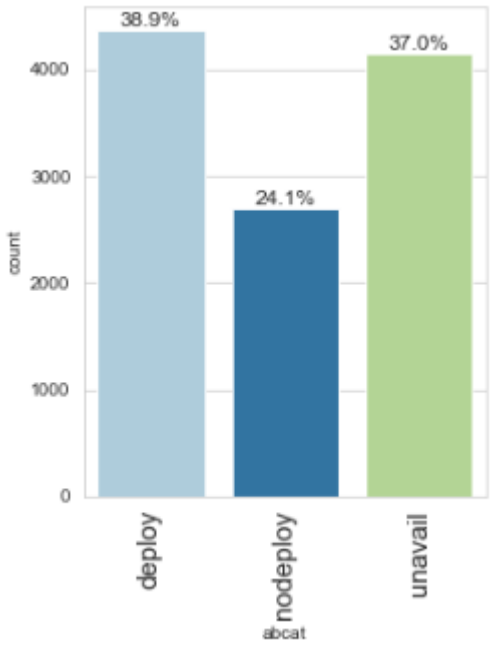


Figure 31: Deployment of airbags by occupant

➔ About 40% of the occupants deployed airbags after the crash, 24% did not, and for the remaining cases, airbags were not available.

- Observation on whether the airbags were deployed or not

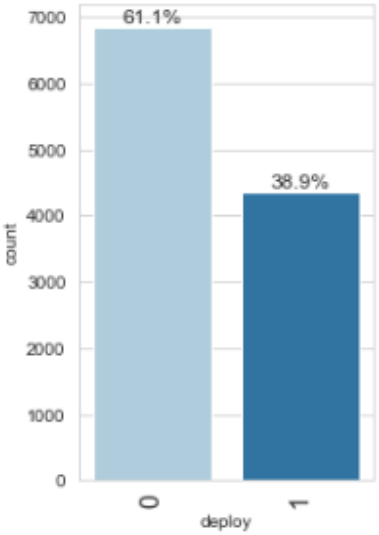


Figure 32: Deployment of airbags

➔ After the occupants deployed airbags, only 40% of the airbags deployed properly.

➔ 61% did not, and that must be a concern.

- Observation on whether the person involved in the car crash was the driver or a passenger

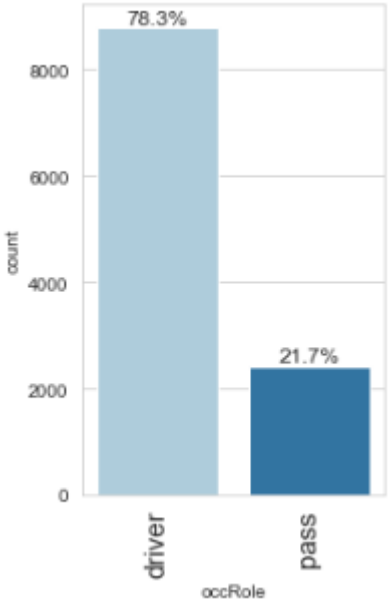


Figure 33: Occupant type

➔ 78% of occupants involved in the carcrash casualty were the drivers and 21% were the passengers.

- Yearly records of the occurrences of car crashes

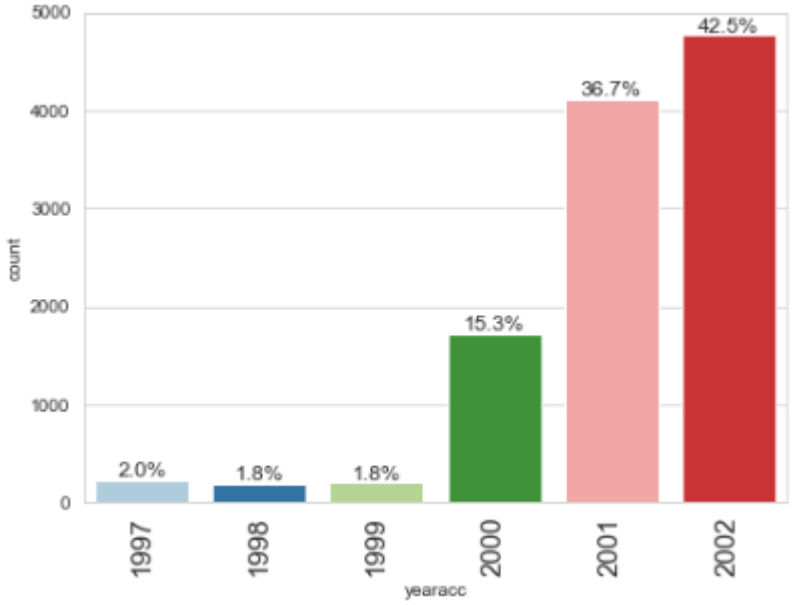


Figure 34: Yearly records of car crashes

➔ 2002 recorded the maximum car crash incidents.

➔ Between 1997 to 1999 there aren't many records of car crash, this could be because the incidents were not kept track of.

- Observations on whether the impact was frontal or non-frontal

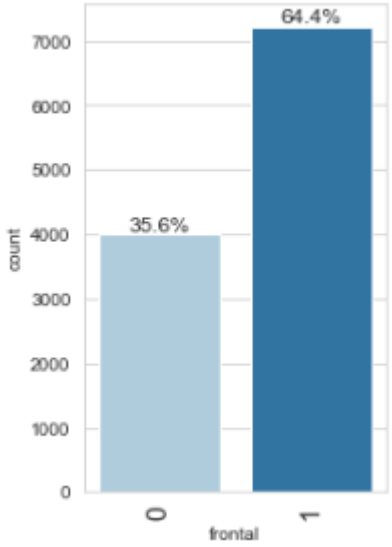


Figure 35: Frontal and non-frontal impact

➔ 64% of the car crash were due to the frontal impact of accident.

- Observation on the severity of injury

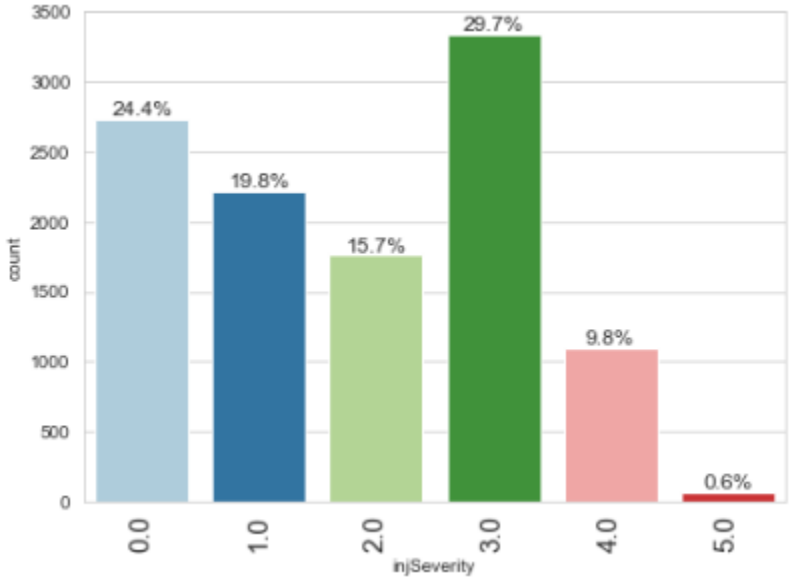


Figure 36: Severity of injury

➔ On a scale of 1 to 5, most severity is ranked 3, which indicates moderate severity.

➔ Only 0.6% of car crashes have ended up with highest severity

- Distribution of weight of car

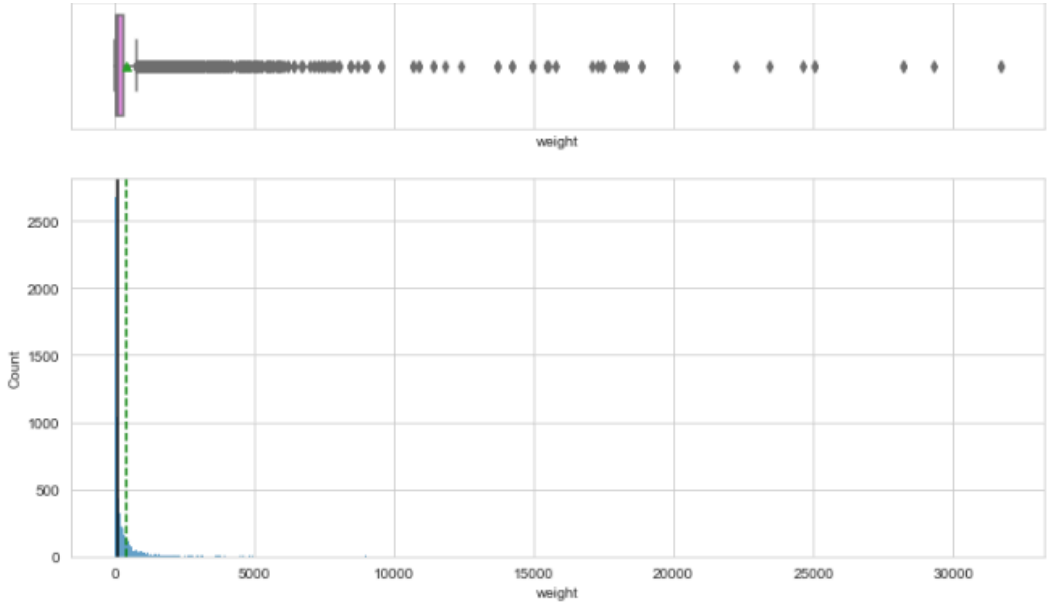


Figure 37: Weight of car

➔ The weight of most cars is below 5000 units.

- Distribution of age of occupant in years

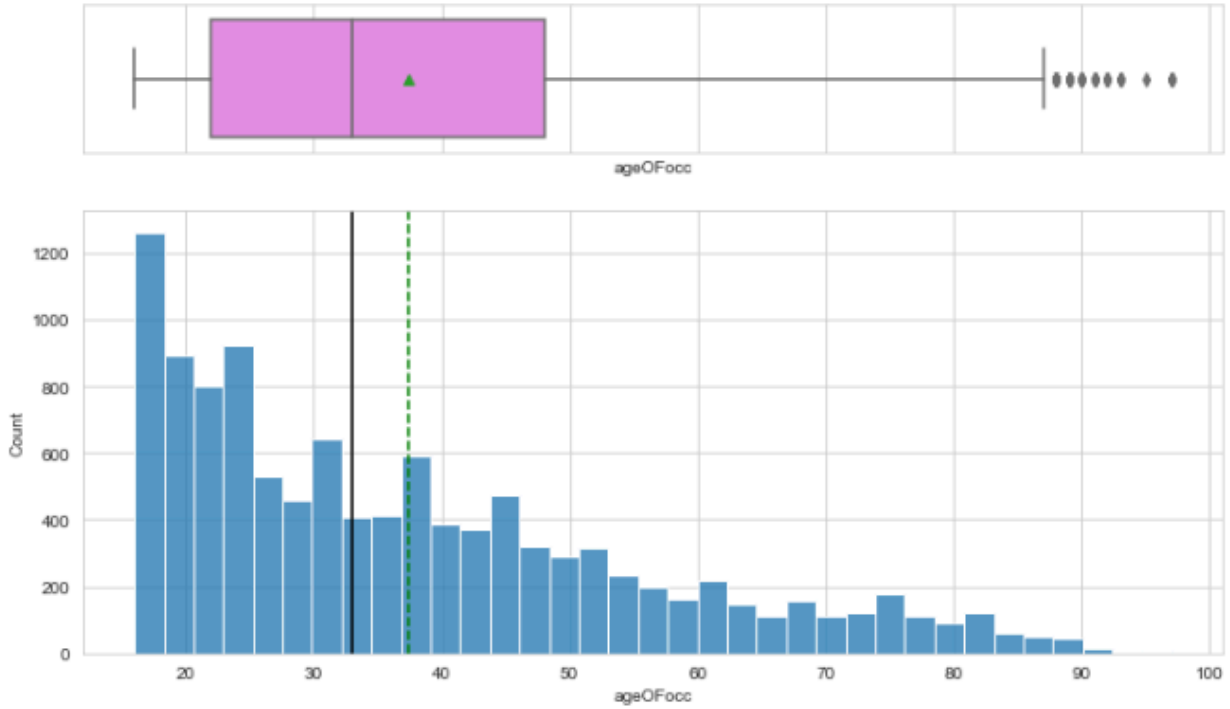


Figure 38: Age of occupants

➔ Te median age of occupants is around 35 years.

➔ The age is right-skewed. Most occupants are aged 20 to 30 years.

- Year of model of vehicle:

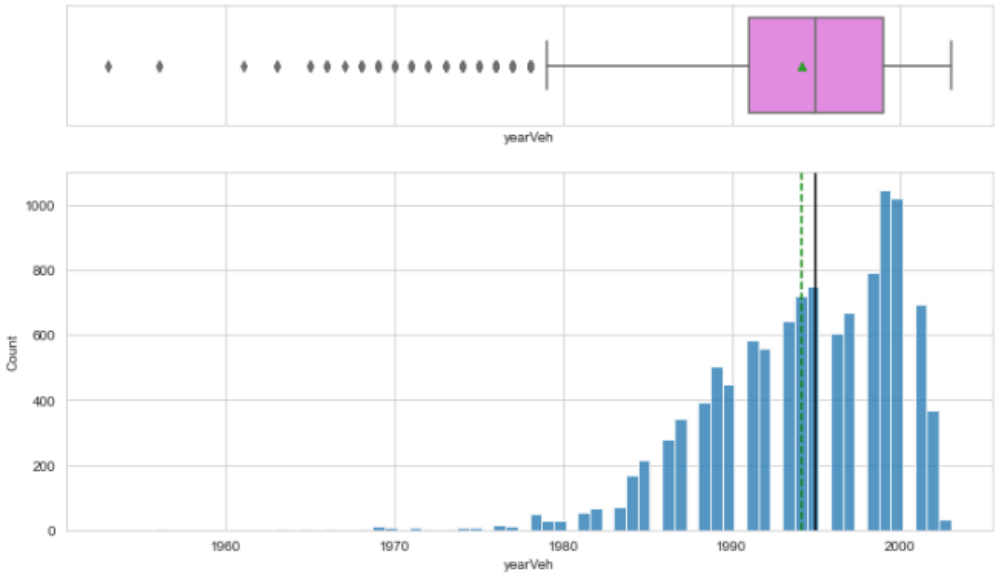


Figure 39: Year model of vehicle

➔ Most vehicles are manufactured after 1990.

➔ Maximum are ones from 2000.

Bivariate Analysis

- Correlation

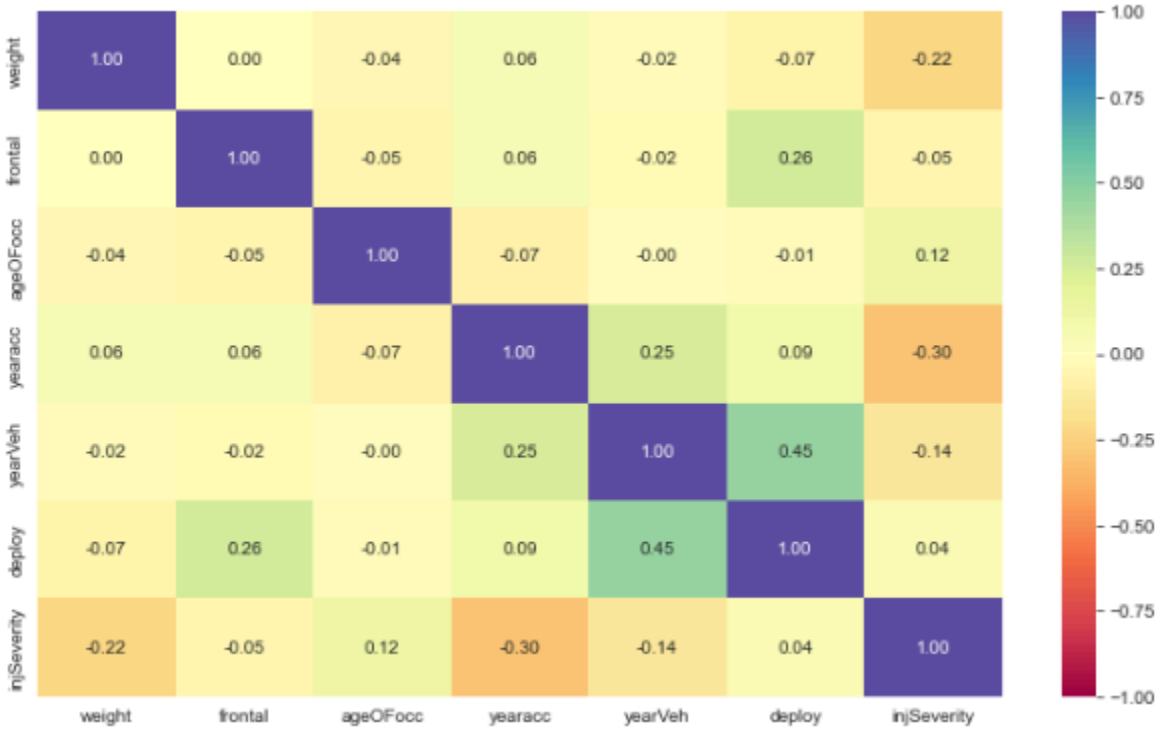


Figure 40: Correlation plot (problem 2)

➔ There is a good correlation between the deployment of airbags year of the vehicle. This could be indicative that new cars had better deployment than old ones.

- Presence of airbag vs Year of model of vehicle

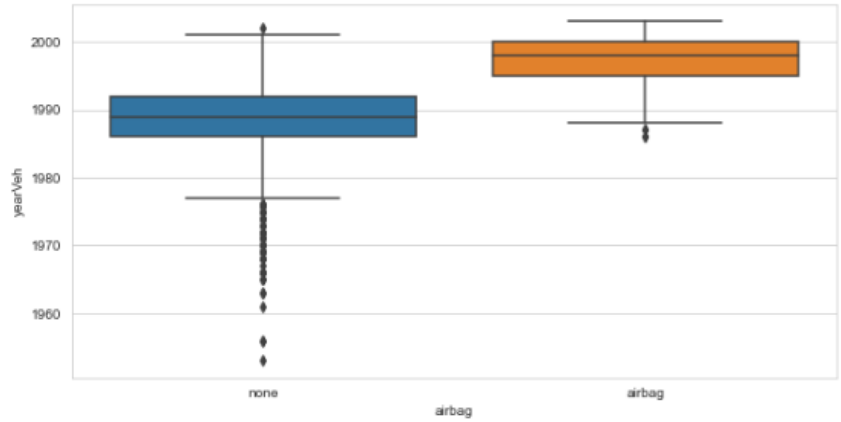


Figure 41: Presence of airbag vs Year of model of vehicle

- Deployment of airbag vs Year of model of vehicle

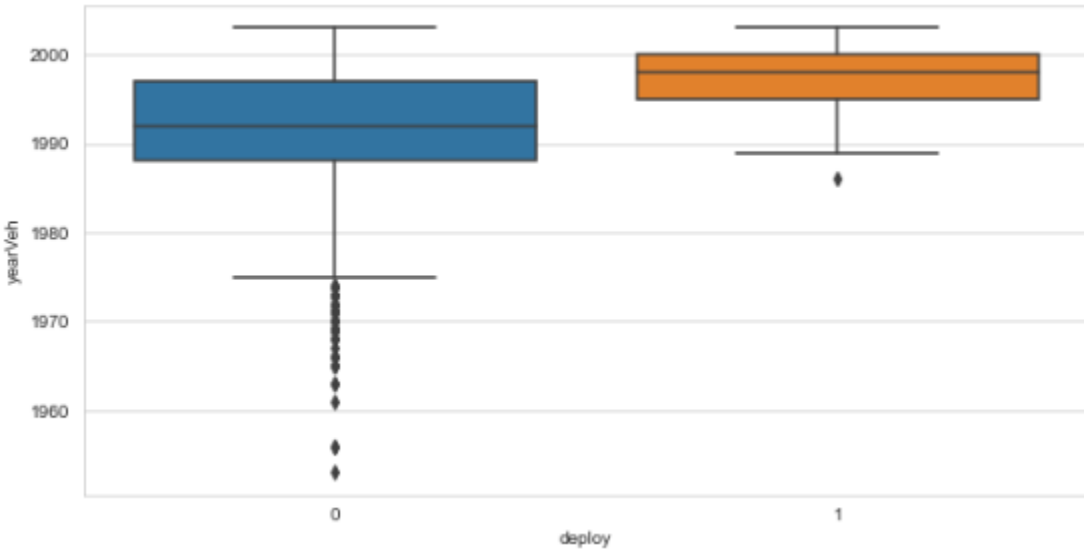


Figure 42: Deployment of airbag vs Year of model of vehicle

- Deployment of airbag of occupant vs Year of model of vehicle

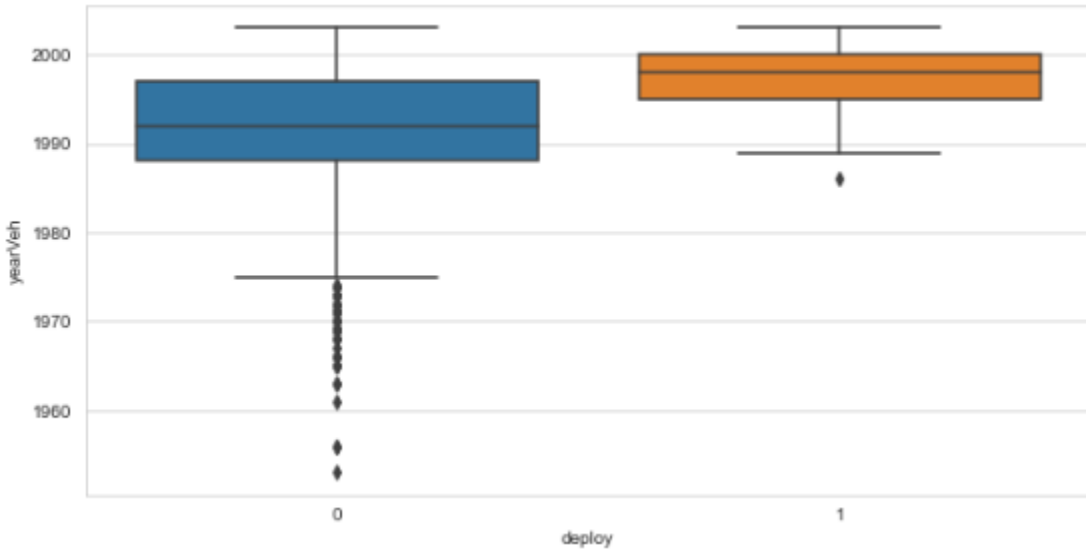


Figure 43: Deployment of airbag of occupant vs Year of model of vehicle

Let's see how the variables have impacted the survival of occupants involved in the crash.

- Speed:

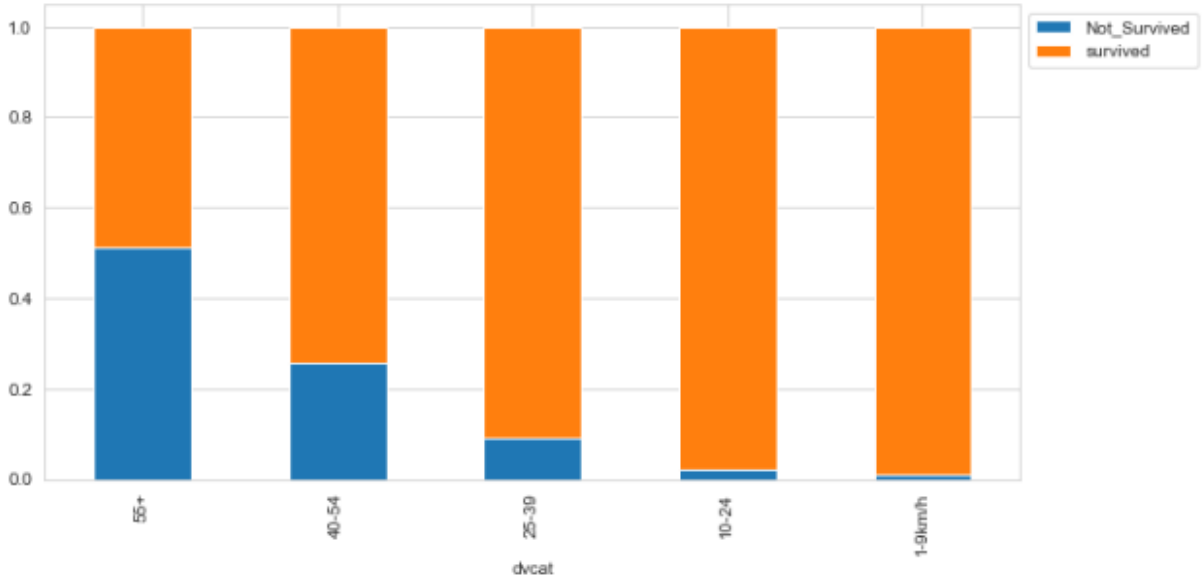


Figure 44: Speed vs Survival

- Presence of airbags

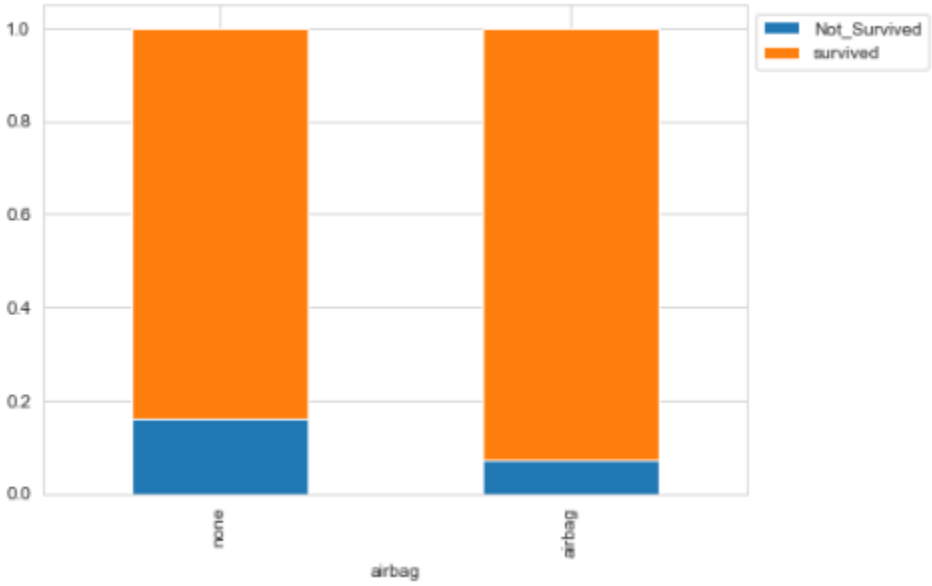


Figure 45: Presence of airbags vs Survival

- Usage of seatbelt

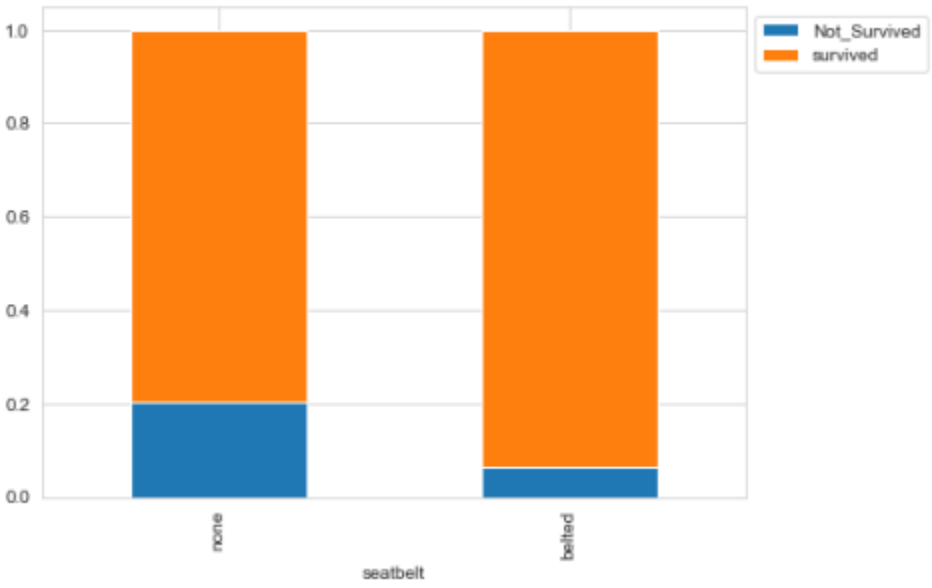


Figure 46: Usage of seatbelt vs Survival

- Impact of crash, whether frontal or non-frontal

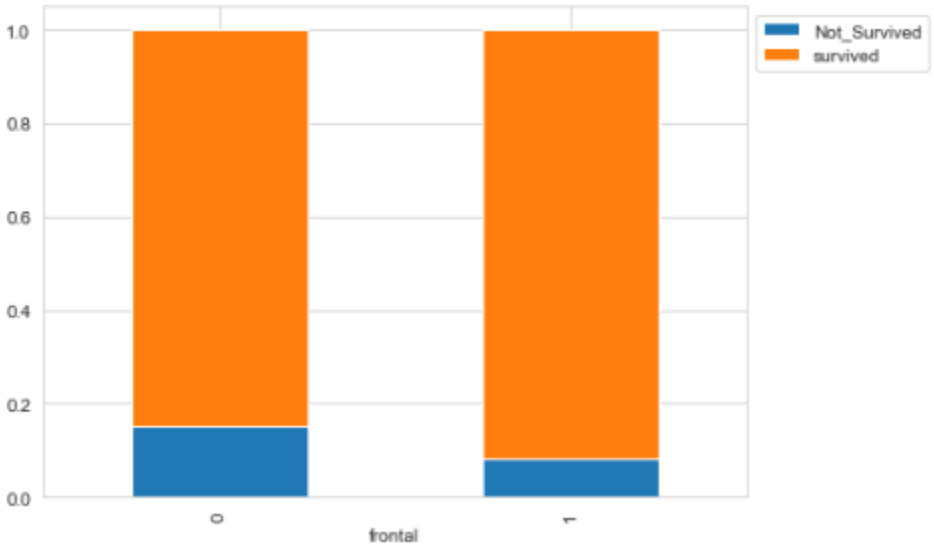


Figure 47: Region of impact of crash vs Survival

- Survival rate based on sex

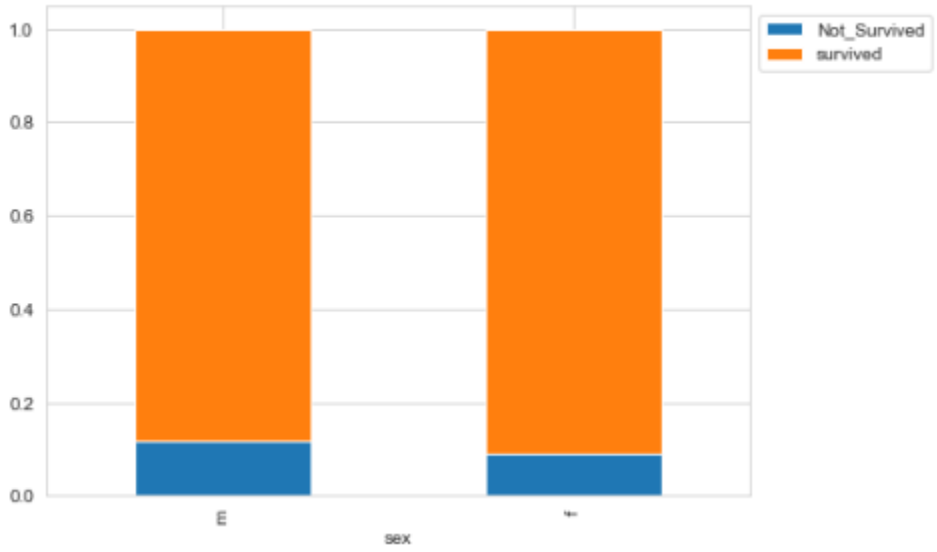


Figure 48: Survival rate based on sex

- Yearly record

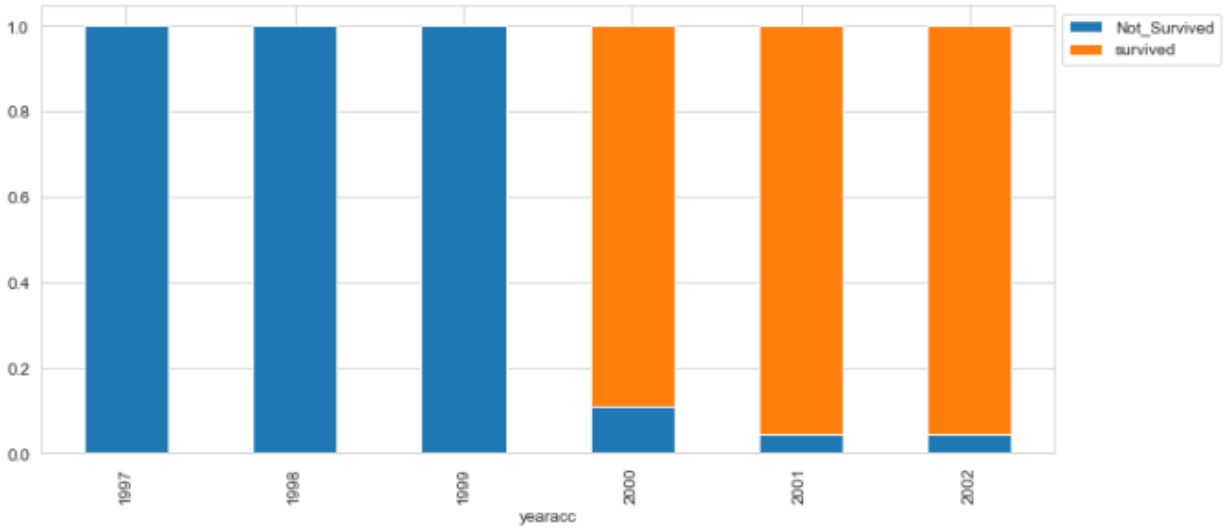


Figure 49: Yearly record vs Survival

- Deployment of airbags by the occupant

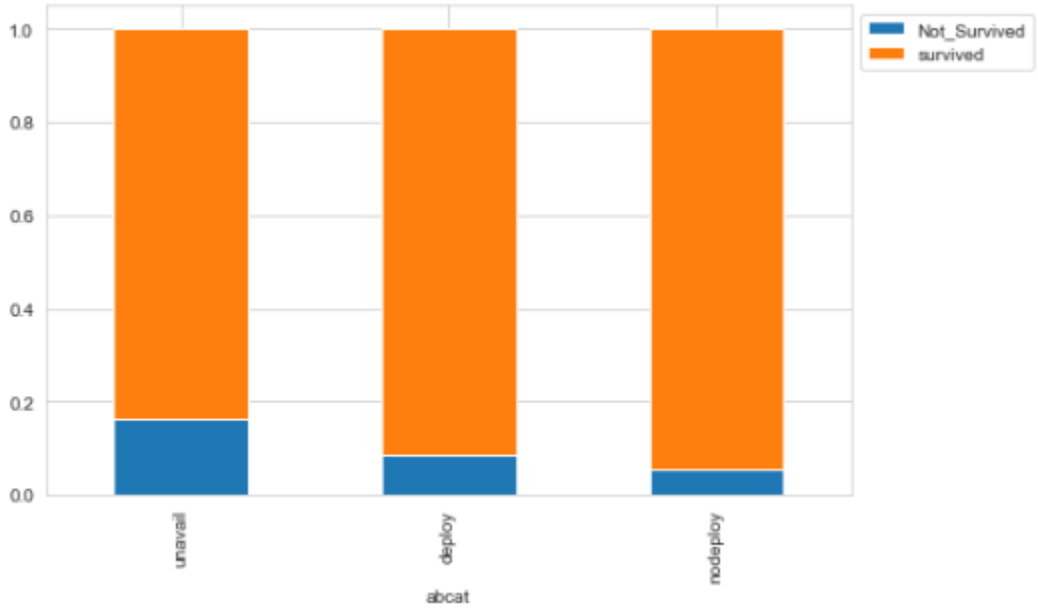


Figure 50: Deployment of airbags by the occupant vs Survival

- Occupant

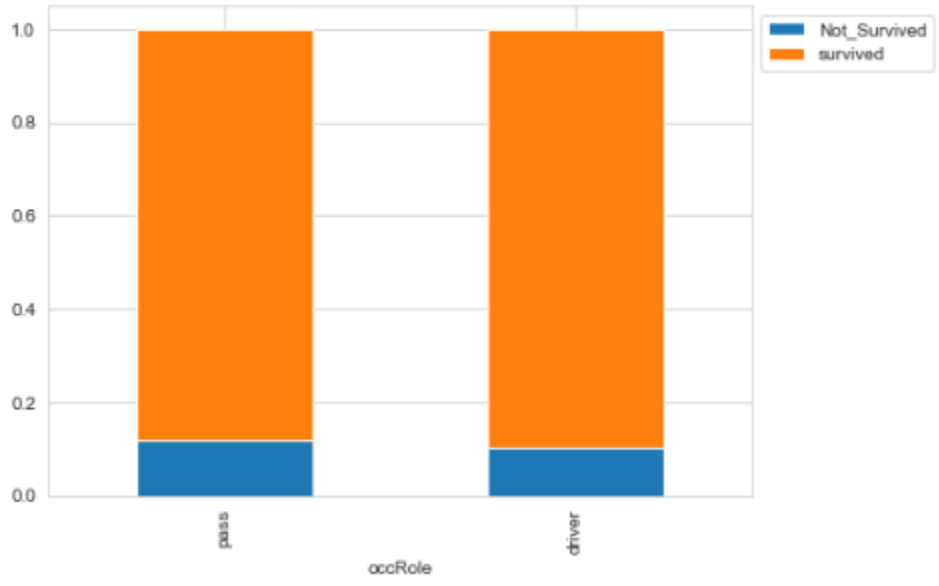


Figure 51: Occupant type vs Survival

- Whether the airbag deployed or not

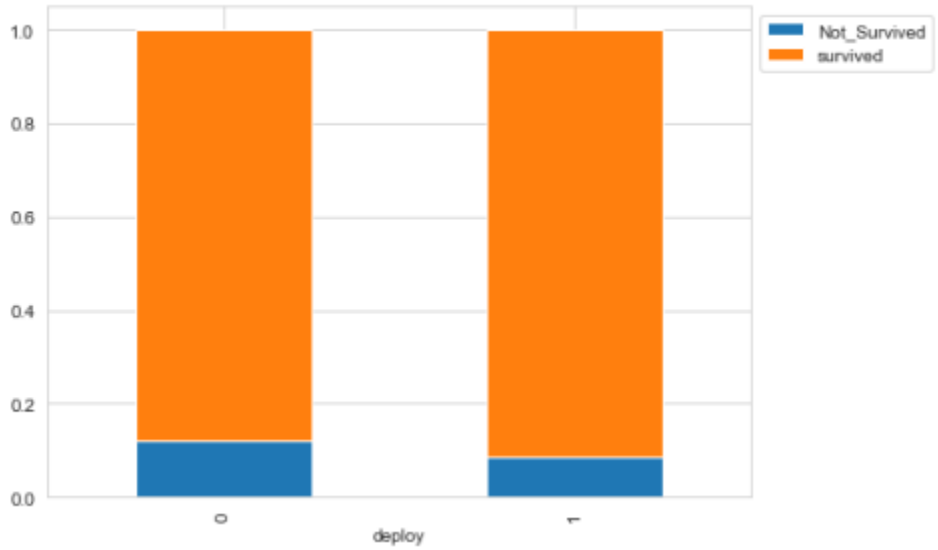


Figure 52: Whether the airbag deployed or not vs Survival

- Severity of injury

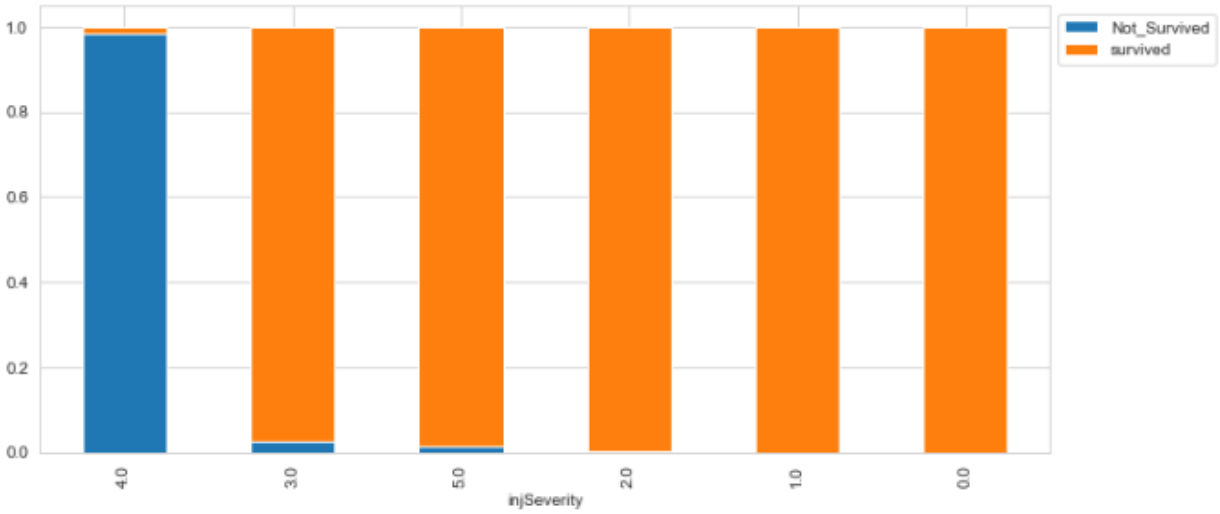


Figure 53: Severity of injury vs Survival

- Age of occupant

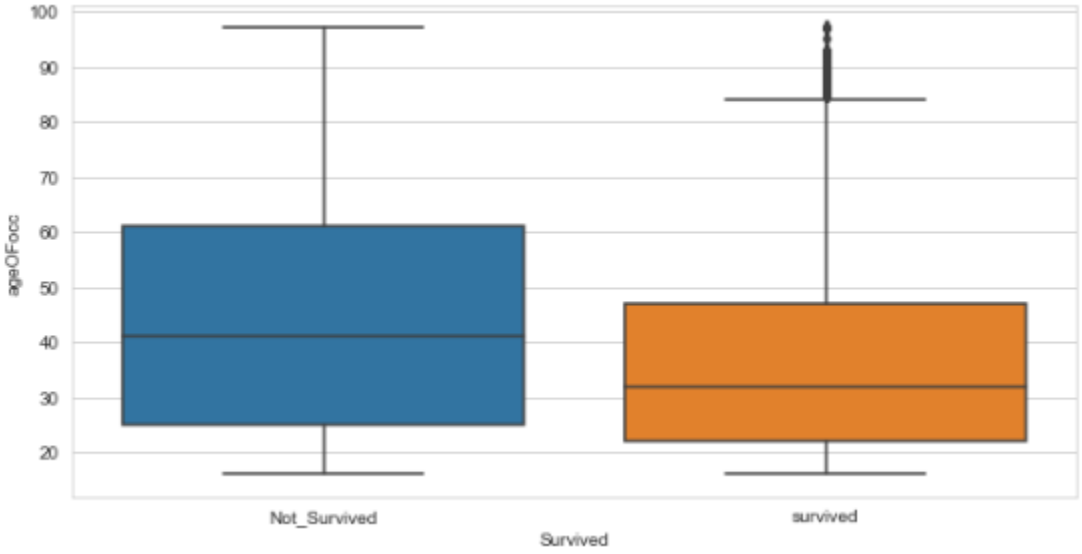


Figure 54: Age of occupant vs Survival

- Impact of speed and age of occupant on survival

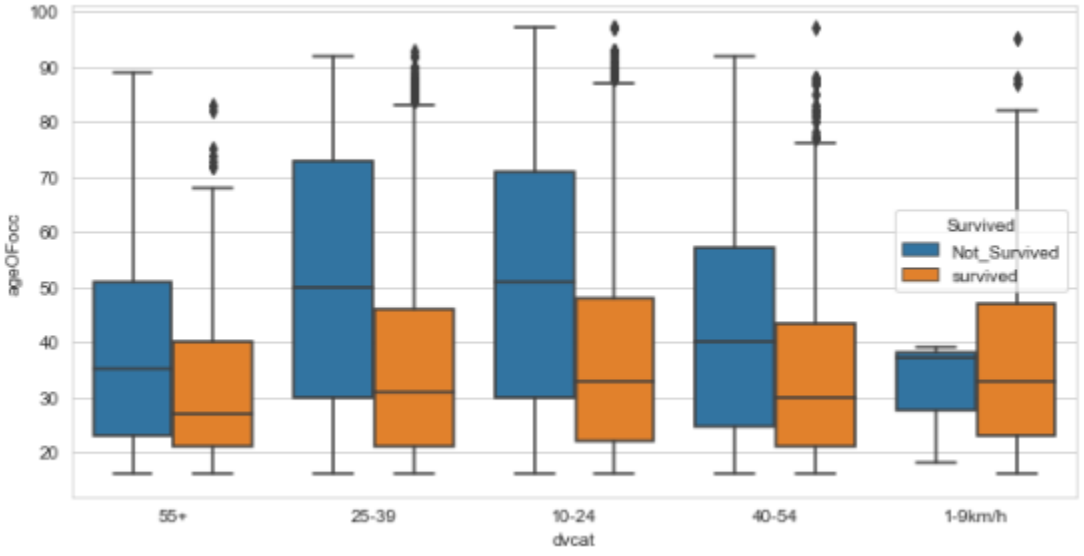


Figure 55: Impact of speed and age of occupant on survival

- Impact of sex and age of occupant on survival

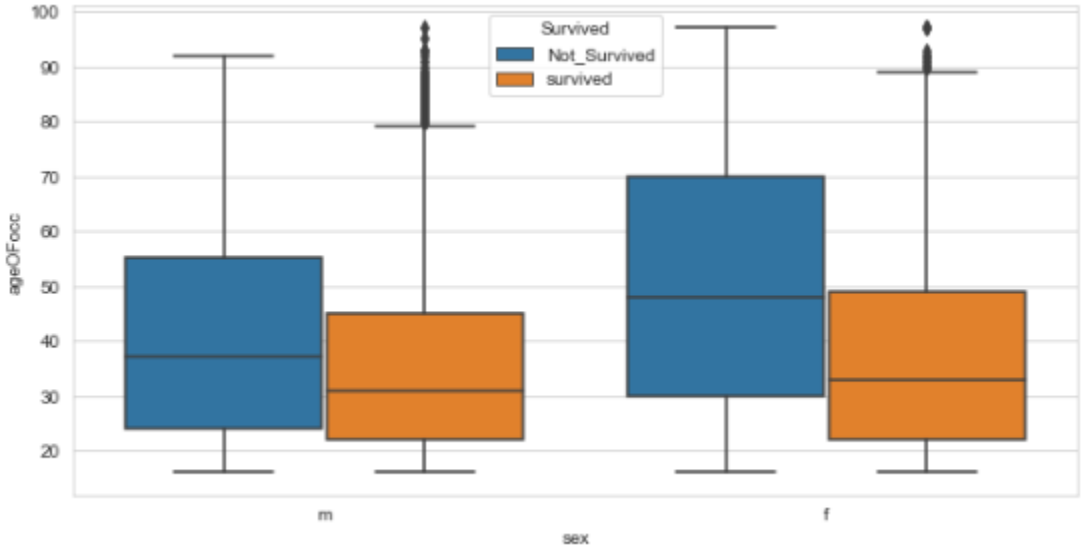


Figure 56: Impact of sex and age of occupant on survival

- Impact of type of occupant and their age on survival

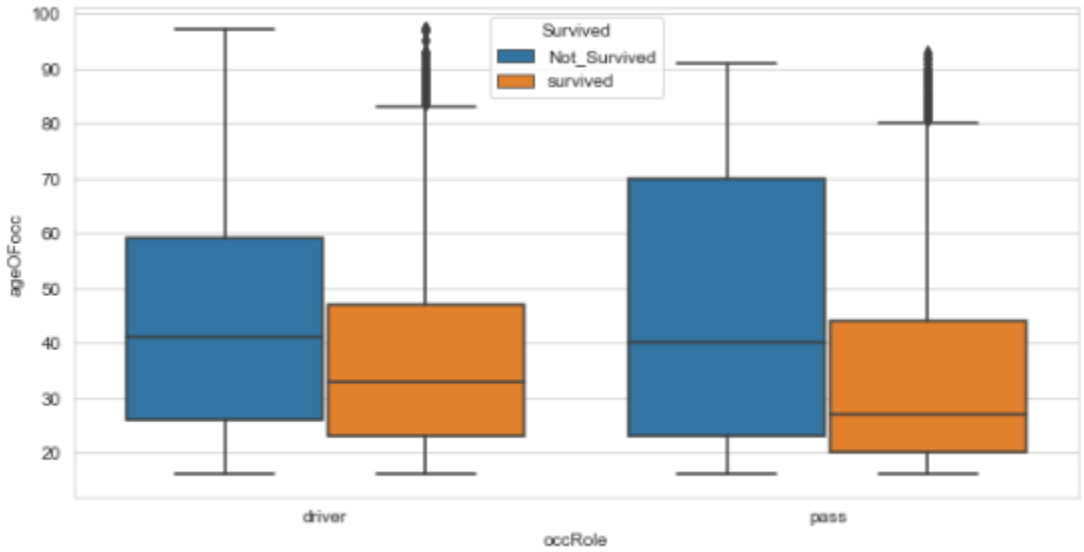


Figure 57: Impact of type of occupant and their age on survival

- Impact of the severity of injury and the age on survival

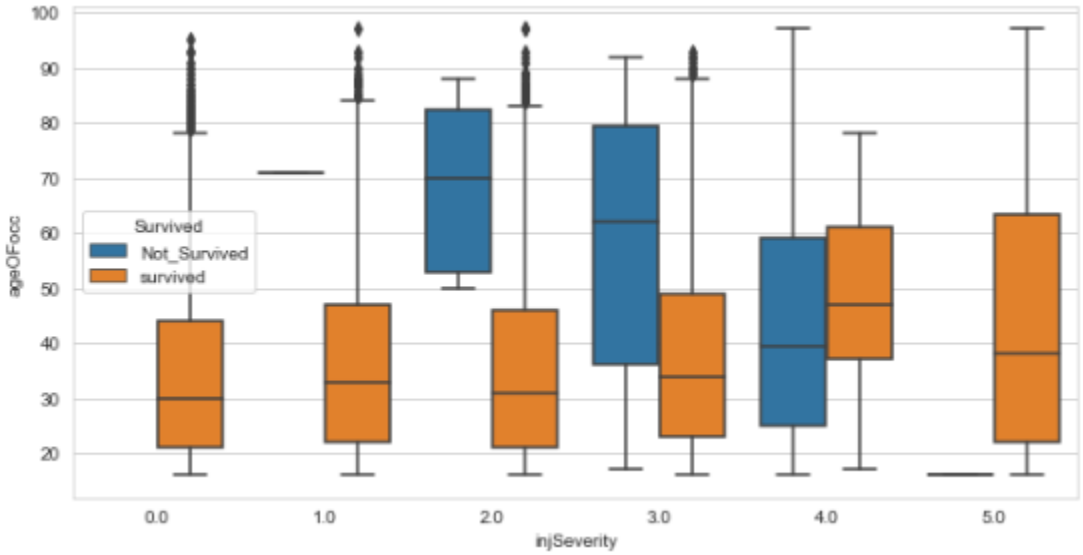


Figure 58: Impact of the severity of injury and the age on survival

- Relationship between the role of the occupant, their age, and their survival

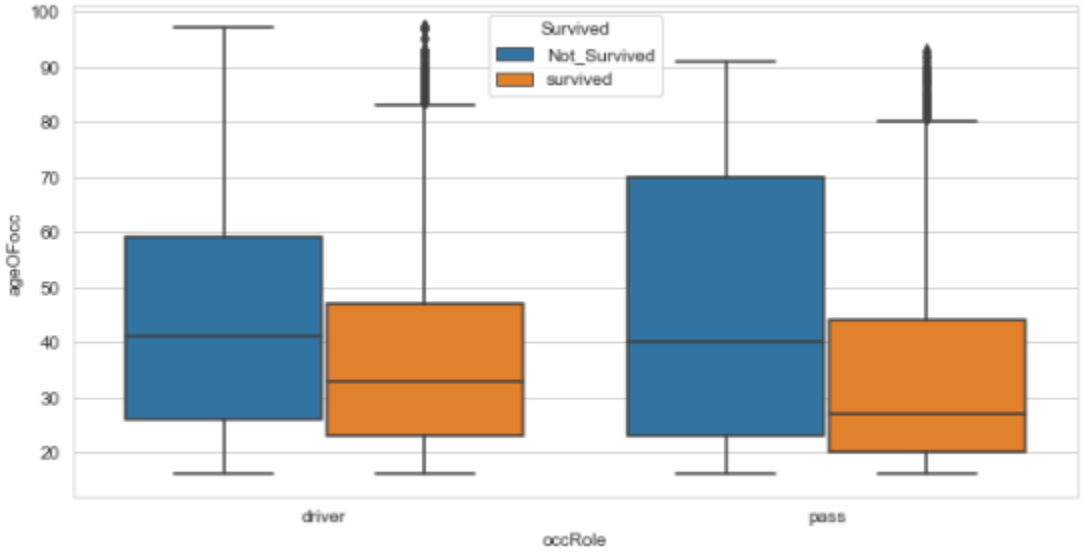


Figure 59: Relationship between the role of the occupant, their age, and their survival

2.2) Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

Outliers & Scaling

Before encoding the data, we treat the outliers.

Before treating outliers

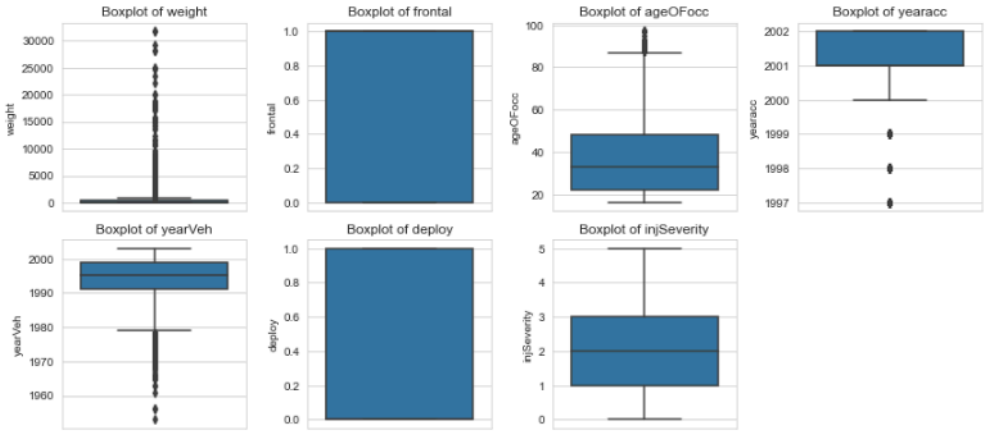


Figure 60: Before treating outlier

After treating outliers

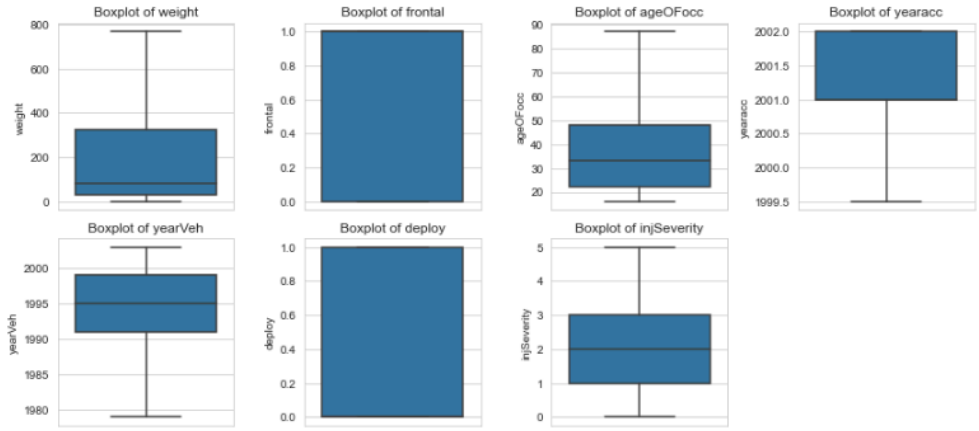


Figure 61: After treating outliers

We scale the data:

Dataset after scaling:

	weight	frontal	ageOFocc	yearacc	yearVeh	deploy	injSeverity
0	-0.73440	0.74347	-0.29821	-2.06767	-1.34089	-0.79815	1.58189
1	-0.49562	-1.34505	0.91487	-2.06767	-0.04576	-0.79815	1.58189
2	-0.73440	0.74347	1.63168	-2.06767	-0.41579	-0.79815	1.58189
3	-0.73440	0.74347	1.46626	-2.06767	-0.41579	-0.79815	1.58189
4	-0.78671	0.74347	-0.79447	-2.06767	-1.52591	-0.79815	1.58189

Table 7: Dataset after scaling (Problem 2)

Summary of scaled data

	count	mean	std	min	25%	50%	75%	max
weight	11217.00000	-0.00000	1.00004	-0.83777	-0.72976	-0.52399	0.39931	2.09293
frontal	11217.00000	-0.00000	1.00004	-1.34505	-1.34505	0.74347	0.74347	0.74347
ageOFocc	11217.00000	-0.00000	1.00004	-1.18044	-0.84960	-0.24307	0.58403	2.73448
yearacc	11217.00000	-0.00000	1.00004	-2.06767	-0.23089	-0.23089	0.99364	0.99364
yearVeh	11217.00000	-0.00000	1.00004	-2.82104	-0.60081	0.13926	0.87934	1.61941
deploy	11217.00000	0.00000	1.00004	-0.79815	-0.79815	-0.79815	1.25290	1.25290
injSeverity	11217.00000	0.00000	1.00004	-1.32972	-0.60182	0.12609	0.85399	2.30980

Table 8: Summary of dataset after scaling (Problem 2)

Data Preparation for Modelling

- We convert dvcat object to categorical codes in an ordinal manner.

10-24' = '1'

'25-39' = '2',

'40-54' = '3'

'55+' = '4'

'1-9km/h' = '5'

- We encode categorical feature sp500 and split the data into train and test (70:30) to evaluate the model we build on the train data.

```
Shape of Training set :  (7851, 17)
Shape of test set :  (3366, 17)
Percentage of classes in training set:
1   0.89479
0   0.10521
Name: Survived_survived, dtype: float64
Percentage of classes in test set:
1   0.89483
0   0.10517
Name: Survived_survived, dtype: float64
```

- The variable of interest is ‘Survived’ which is encoded as 1
- Not survided is encoded as 0.

A model can make wrong predictions as:

- Model predicts which occupant of car crash is likely to survive, but in reality, the outcome could be different. Other factors such as past medical history, financial ability to pay for medical expenses, etc., could influence the survival.
- The reverse case can also occur.

Which case is more important?

- Both cases are important to get an accurate result.

How to reduce the losses?

- In the context of predicting survival rates, we can consider prioritizing metrics like precision, recall, or the F1 score, since the data seems highly imbalanced (**89.5% survived**).

Logistic regression model

Training data

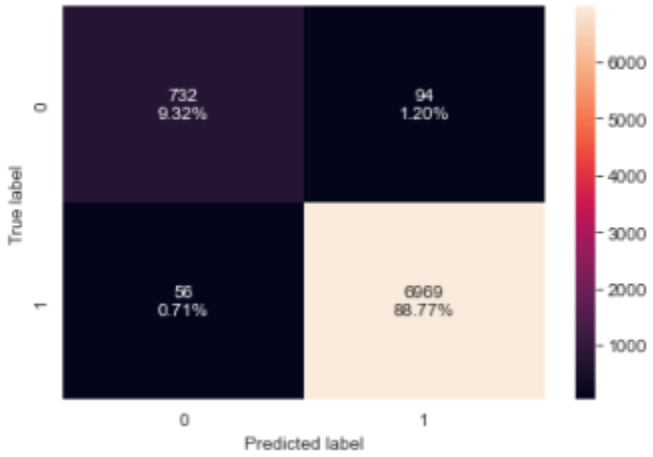


Figure 62: Confusion matrix for training data (logistic regression)

	precision	recall	f1-score	support
0	0.93	0.89	0.91	826
1	0.99	0.99	0.99	7025
accuracy			0.98	7851
macro avg	0.96	0.94	0.95	7851
weighted avg	0.98	0.98	0.98	7851

- Accuracy is 98%. However, it's not a reliable measure for skewed data.
- On predicting survivals:
    - Recall score is 99%, so the model is 99% good at minimizing false negatives, meaning it rarely misses instances of the positive class.
    - F1 score is about 99%, which is not a very high score. The model is fairly good.
    - Precision score is about 99%, i.e. the model is 99% good at making relatively few false positive predictions compared to the total number of positive predictions it has made.
- On predicting cases where occupants do not survive:
    - Recall score is 89%, so the model is 89% good at minimizing false negatives, meaning it rarely misses instances of the positive class.
    - F1 score is about 91%, which is not a very high score. The model is fairly good.
    - Precision score is about 93%, i.e. the model is 93% good at making relatively few false positive predictions compared to the total number of positive predictions it has made.

Testing data

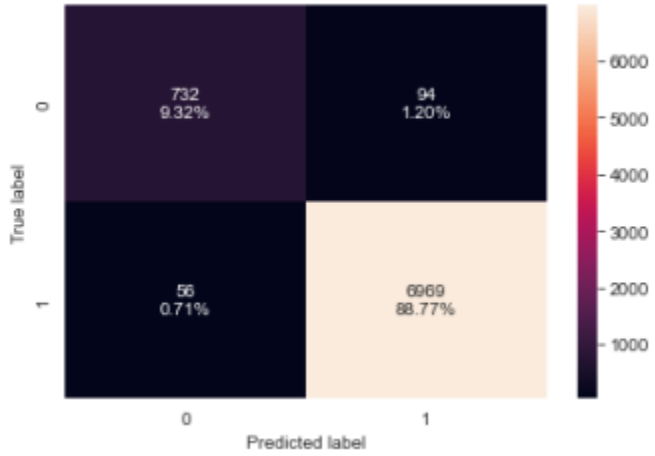


Figure 63: Confusion matrix for testing data (logistic regression)

	precision	recall	f1-score	support
0	0.94	0.89	0.91	354
1	0.99	0.99	0.99	3012
accuracy			0.98	3366
macro avg	0.96	0.94	0.95	3366
weighted avg	0.98	0.98	0.98	3366

- Accuracy is 98%. However, it's not a reliable measure for skewed data.
- On predicting survivals:
    - Recall score is 99%, so the model is 99% good at minimizing false negatives, meaning it rarely misses instances of the positive class.
    - F1 score is about 99%, which is not a very high score. The model is fairly good.
    - Precision score is about 99%, i.e. the model is 99% good at making relatively few false positive predictions compared to the total number of positive predictions it has made.
- On predicting cases where occupants do not survive:
    - Recall score is 89%, so the model is 89% good at minimizing false negatives, meaning it rarely misses instances of the positive class.
    - F1 score is about 91%, which is not a very high score. The model is fairly good.

- Precision score is about 94%, i.e. the model is 94% good at making relatively few false positive predictions compared to the total number of positive predictions it has made.

ROC-AUC curve and score for logistic regression model

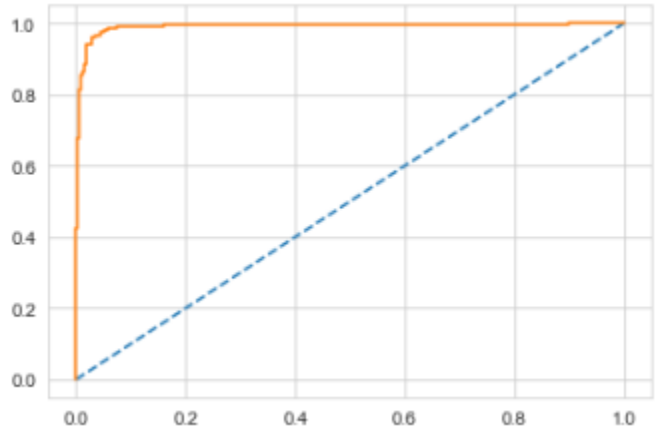


Figure 64: ROC-AUC curve for logistic regression model

AUC score: 0.990
The ROC-AUC score is the same for both test and train data, and it is very good.

Linear Discriminant Analysis Model

Training data

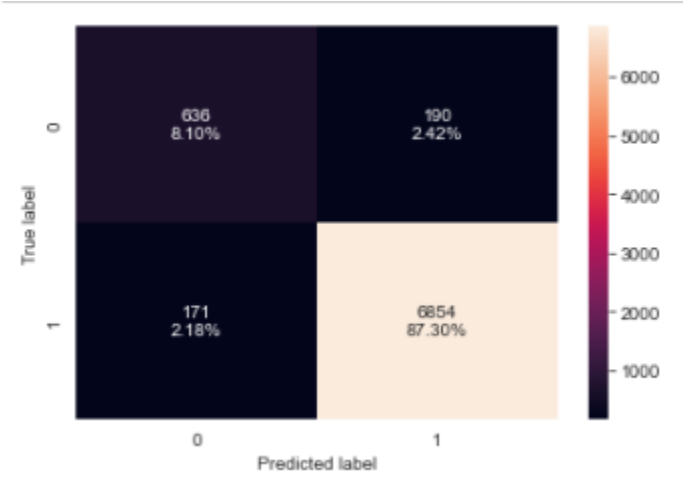


Figure 65: Confusion matrix for training data (linear discriminant analysis)

	Accuracy	Recall	Precision	F1
0	0.95402	0.97566	0.97303	0.97434

- On predicting cases where occupants do not survive:
    - Recall score is 97%, so the model is 97% good at minimizing false negatives, meaning it rarely misses instances of the positive class.
    - F1 score is about 97%, which is not a very high score. The model is fairly good.
    - Precision score is about 97%, i.e. the model is 97% good at making relatively few false positive predictions compared to the total number of positive predictions it has made.

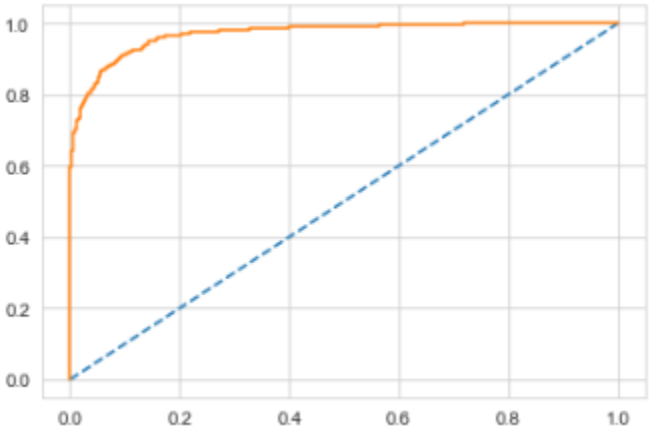


Figure 66: ROC-AUC curve for linear discriminant analysis model

AUC: 0.971
The ROC-AUC score is the same for both test and train data, and it is very good.

Testing data

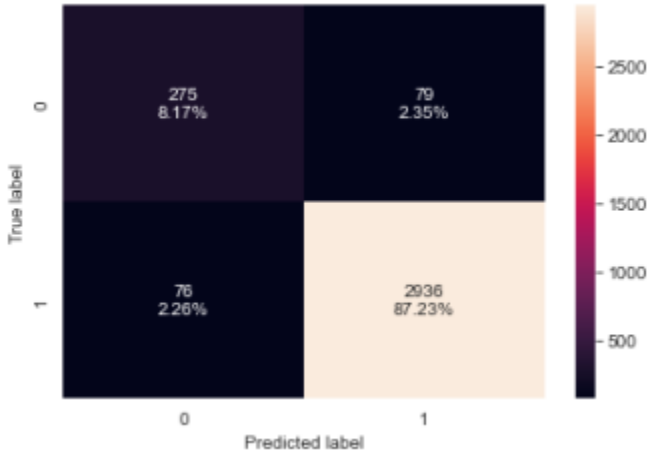


Figure 67: Confusion matrix for testing data for linear discriminant analysis model

	Accuracy	Recall	Precision	F1
0	0.95395	0.97477	0.97380	0.97428

AUC score: 0.971

- On predicting cases where occupants do not survive:
  - Accuracy is 95%.
  - Recall score is 97%, so the model is 97% good at minimizing false negatives, meaning it rarely misses instances of the positive class.
  - F1 score is about 97%, which is not a very high score. The model is fairly good.
  - Precision score is about 97%, i.e. the model is 97% good at making relatively few false positive predictions compared to the total number of positive predictions it has made.

Final Model Selection

Comparing both the models built so far:

- Correctly predicting survivors is more critical (e.g., in medical diagnosis), hence we focus on recall to ensure that we capture as many true survivors as possible, even at the cost of some false positives (misclassified not-survivors).
- For training data and testing data:
  Linear Discriminant Analysis model has a recall score of 97% whereas the logistic regression model has a score of 89%.

  Hence, the Linear Discriminant Analysis model is better than the Logistic Regression model.

Actionable Insights & Recommendations

- After comparing the two models, we can conclude that the Linear Discriminant Analysis model is better than the Logistic Regression model.
- The scores of the model is good and trustable.
- However, the dataset we have analysed is highly skewed. 89% of the cases surveyed have survived the crash. We may not have enough information about people that haven't survived. Hence, we may need to run the models for a larger dataset.
- About 40% of the occupants deployed airbags after the crash, 24% did not, and for the remaining cases, airbags were not available. This issue needs to be addressed as priority. Awareness about availability and how to deploy airbags need to be made rampant.
- After the occupants deployed airbags, only 40% of the airbags deployed properly. 61% did not, and that is a great concern. Such malfunctioning should be penalised and dealt with immediately.
- 64% of the car crash were due to the frontal impact of accident. Therefore, infrastructurally, this has be made more robust