Taqdeer ML Project

Team Members:

- **Saad BinOnayq**
- **Mohammed Alkathiri**
- **Mashhor Al-Bakrnn**
- **Nawaf Alrubayyi**
- **Sami Asiri**
- **Waleed Almaslokhi**

تقدير TAQDEER
تقييم TAQEEM
تحت إشراف الهيئة السعودية للمقيمين المعتمدين

Content

1. Introduction

- Machine learning project uses past vehicle damage assessments to predict future vehicle damage assessments according to car information.

About Taqdder:

- Taqdeer is an integrated system for managing, operating, and organizing vehicle damage assessments electronically and professionally.

- Taqdeer system works on the governance of procedures and technical touch with all parties concerned with vehicle damage assessment.

More information: taqeem1. (2021, September 19). تقدير - تقدير ولا أسهل. Retrieved November 15, 2022, from https://www.taqdeer.sa/

How our project related to Saudi's 2030 vision:

Our project will assist in digital transformation which is in line with Saudi Arabia's vision in 2030 and its goals in digital transformation that serves organizations and the country in development and growth. Digital transformation is one of the main targets of the 2030 vision. Our project aims to achieve digital transformation in the field of vehicle damage assessment by automating the assessment process based on machine learning without the need for assessment experts

Challenges With the Dataset:

- More Than 1M Records
- Mixed Inputs
- Arabic Spelling Issues
- Parts Standardization
- Informal Arabic

Problem Statement: A large number of spere parts of cars have a different estimated cost compared to the same car, type and year of manufacture and also the assessment cost has a significant variation, even though they are doing the same job which causes many problems such as the following:

- Non-standardization of the spere cost and assessment cost
- Manipulation and fraud with spare parts prices
- One of the most serious problems is that it may lead to wrong decisions in the future

Hence, we took the data from Taqdeer to understand what factors we should focus on, to facilitate and simplify procedures for those affected. In other words, we want to know what changes they should make to their workplace, in order to avoid the problems and to get the right estimated of the spere cost and assessment cost. We also need to identify which of these factors is most crucial and demands immediate attention.

The goal of the case study We will build a model that can estimate the cost of repairing a damaged car, which includes the price of spare parts and the assessment cost using the regression model which will help on reducing the time and increase the procedures for Taqdeer and will fix the

gab variation. Moreover, the model will standardize the total cost for that cars that have the same damage, brand, type, and the year of manufacture of the vehicle. Our objective for this project is Automation, Productivity, and Accuracy.

2. Data Review

Car repair procedures Analytics, the Dataset which used in this case study has taken from Taqdeer. We had 18 variables and more than 1 million observations on the dataset which only in Riyadh city only and the dataset collected in 2018. The dataset was written in Arabic, At the starting of the project we had to translate all columns and rows into English. Then we had to clustering and groping categories. After that, we combined rows that related to the same cars. We spent a lot of time on the preprocessing stage, and we know that the preprocessing stage is the important stage because all the next stages is relied on this stage so we work carefully, and we had to make sure that our work will not affect the data. After we clean the data, translate, and merge rows, we created new columns that could help us on the project. At the end, the data that we had has 29 variables and 253396 observations.

The dataset from: taqeem1. (2021, September 19). تقدير – تقدير ولا أسهل. Retrieved November 15, 2022, from https://www.taqdeer.sa/

Variables and Explanation:

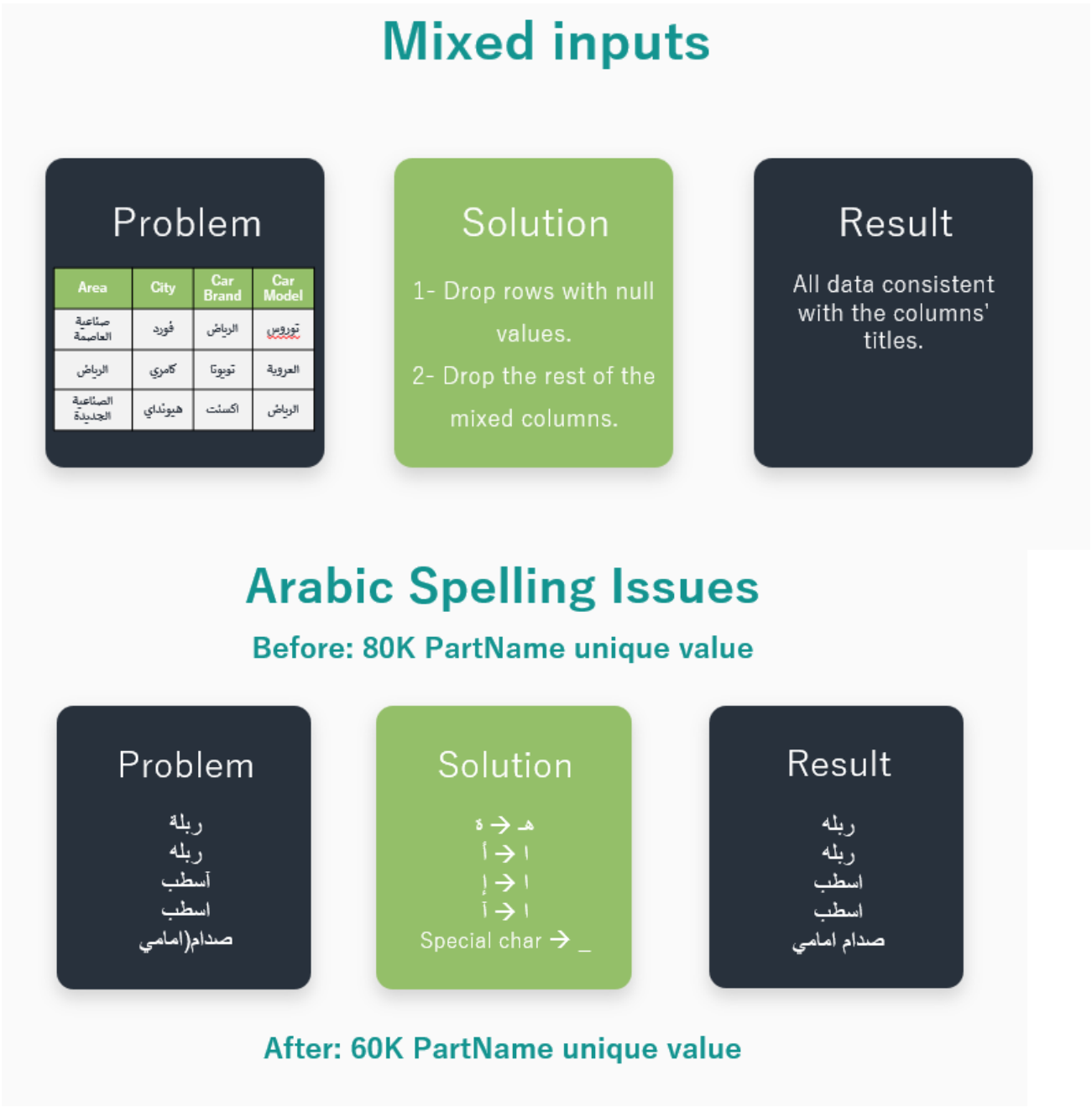
Variables	Explanation
c_id	The car id which defined the car
Area	The place of the branch for Taqdeer
RegistrationTime	The time when they register the car in Taqdeer
CloseTime	When they finish all procedures on the car
CarBrand	The name of the company that the car related to
CarModel	The name of the car
ManufactureYear	The year of manufacture of the vehicle
CarColor	The color of the car
AssessmentCost	The price for people who fix the car
SparePartCost	The price for the spere part of the car
TotalCost	The price of the spere part and the assessment
PaymentType	The payment method how, the price will be pay
DurationTime	The waiting time for finishing all processes
Hour	The time that car inter the Taqdeer
Month	Which month the car finishes the assessment
Day	The number of the day that the car finishes the assessment
WeekDay	The name of the day that the car finishes the assessment
PartsList	The names of the spere part of the car
PositionList	The position of the damaged part

PartStateList	The spere part state wither it is used or new
CarMade	The country that the brand related to
CarClass	Wither it is an expensive car or normal
CarType	Wither the brand has a car, trunk, or both
PartsNumber	How many parts need to be fix
SparePart_Differace%	The assessment cost minus the price of the spare part divided by the price of the spare part
AssessmentEvaluation	Wither the assessment is very high, high, good, low, or very low
TimeEvaluation	Wither the time for assessment is fast, acceptable, or delay
PartOfDay	Either Morning, Afternoon, Evening, or Night
RushHour	Wither the assessment made in the rush hour or not
TotalCostEvaluation	Wither the total cost is very high, high, acceptable, low, or very low

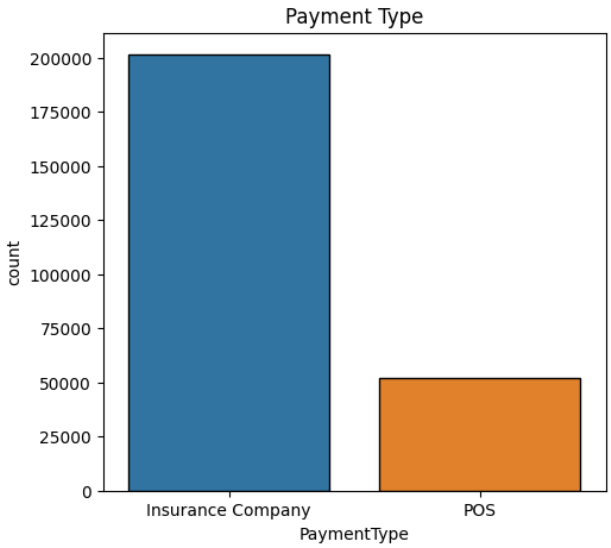
3. Data Preprocessing

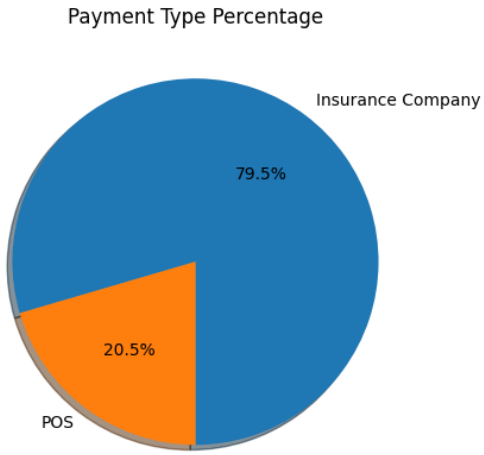
In the preprocessing we must fix the mixed inputs by dropping the rows with null values and the rest of the mixed rows. After that, we had to deal with the Arabic spelling Issues many people wrote the name of the spere part in different spelling, so we had to standardize the spelling. Moreover, we had to translate all columns and rows from Arabic to English.

Example for the preprocessing that have been done:



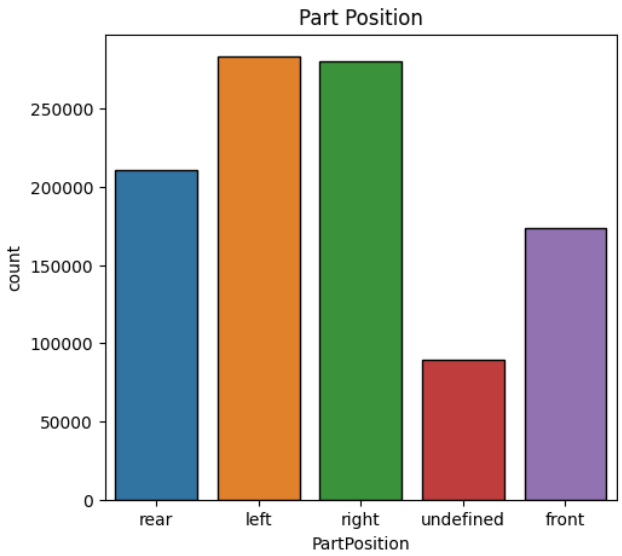
4. Data Exploration

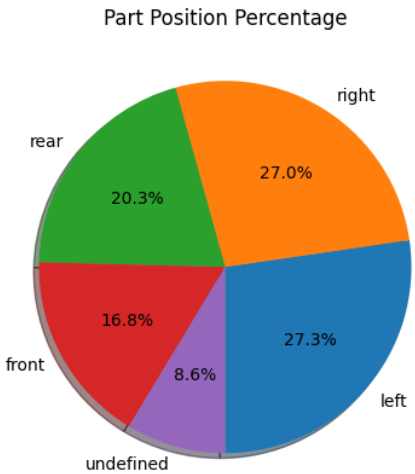




Insights:

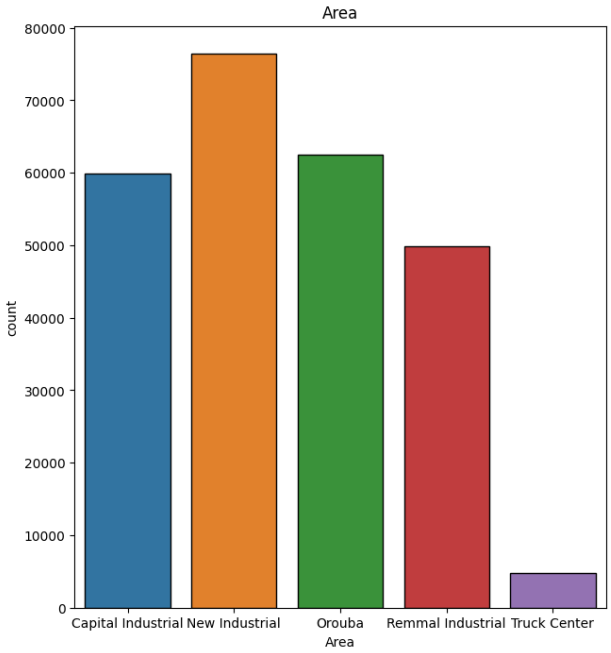
- There's a huge difference between the two payment types
- Insurance Company covered expenses for most of the people
- By looking at the given information, we understand that most of the people take benefit of Car Insurance Companies, whenever an accident occurs

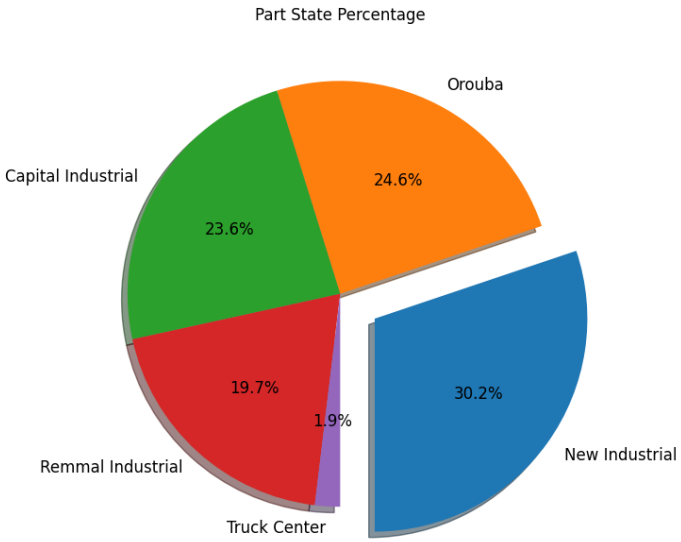




Insights:

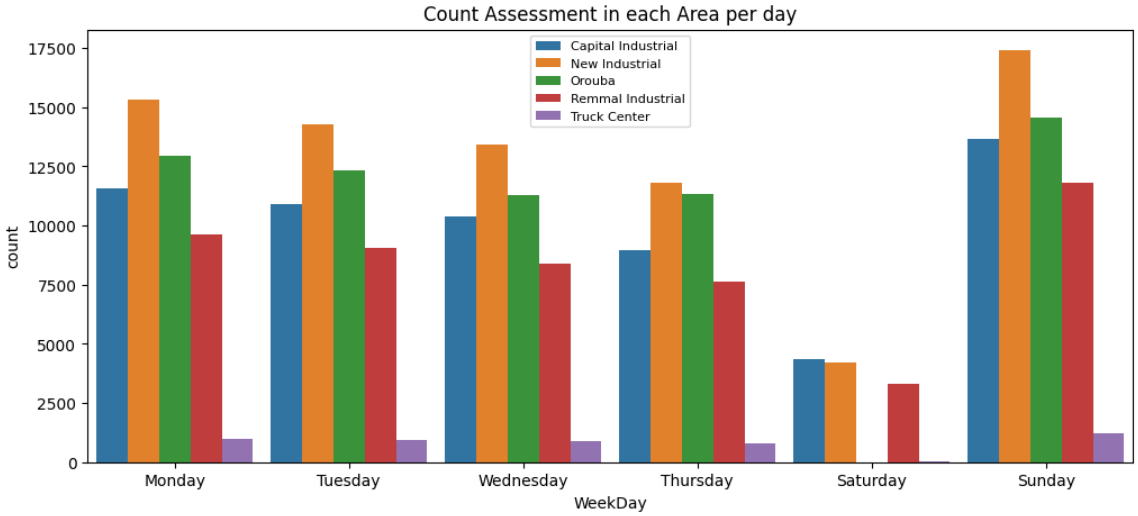
- Left and right parts are the most vulnerable positions of the car whenever an accident takes place
- Whereas the front part of the car is the least likely position to get damaged





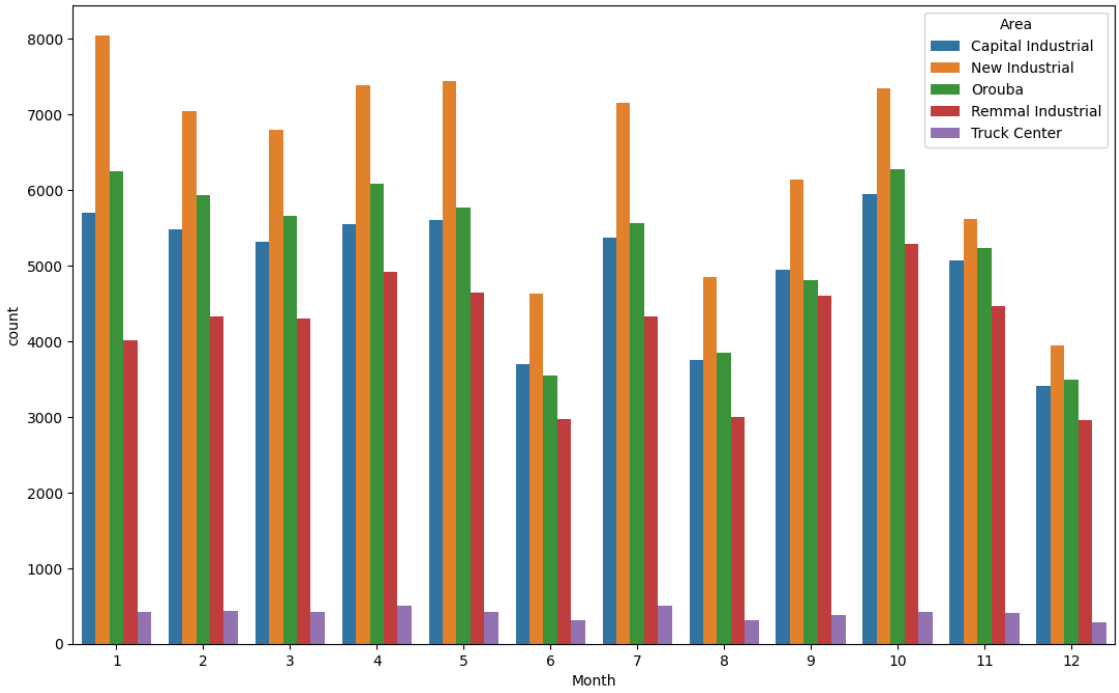
Insights:

- New Industrial is the most commonly visited branch center in Riyadh
- Truck Location is a less visited center since it only caters to trucks.
- The Capital Industrial Area and Orouba have almost the same percentage of visitors in the Riyadh city.



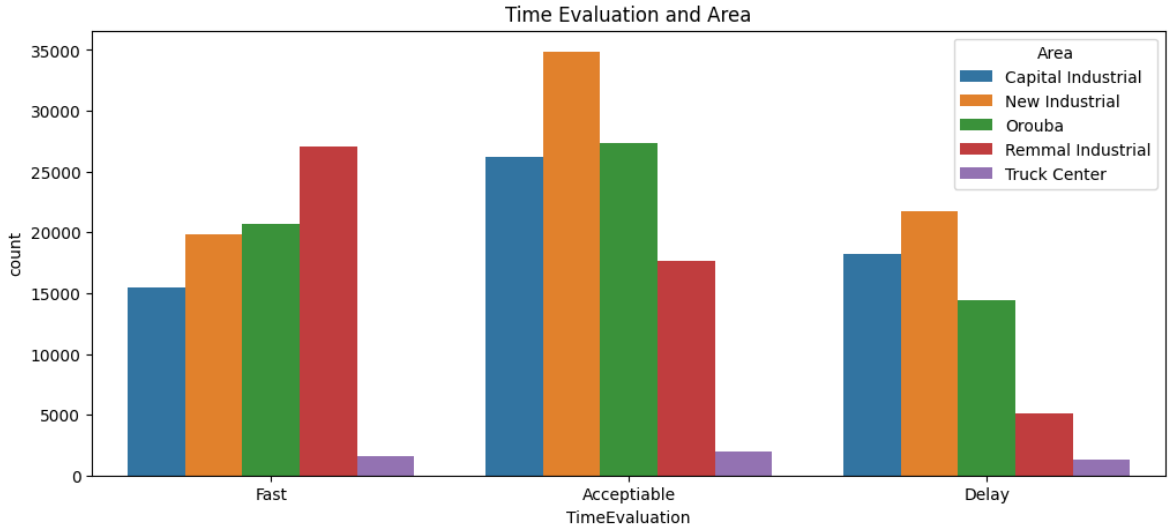
Insights:

- We can observe that Saturday is the least crowded, and it can be noted that the Orouba Center does not operate on Saturday.
- We can observe that Sunday is the busiest day.
- We can observe that the New Industrial center is the busiest on most days.
- There is a clear difference between the number of accidents that occur on weekdays and weekend. Weekdays are more dangerous for driving as compared to on the weekend



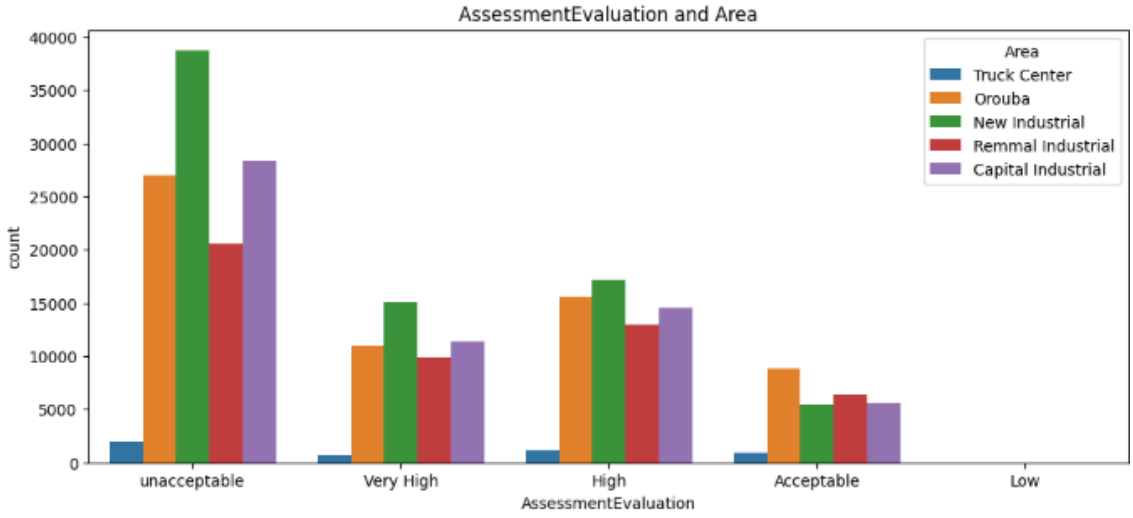
Insights:

- January is the Highest month in car assessment. Whereas December is the lowest month when compared to the other months
- The New Industrial is the most center used for all months



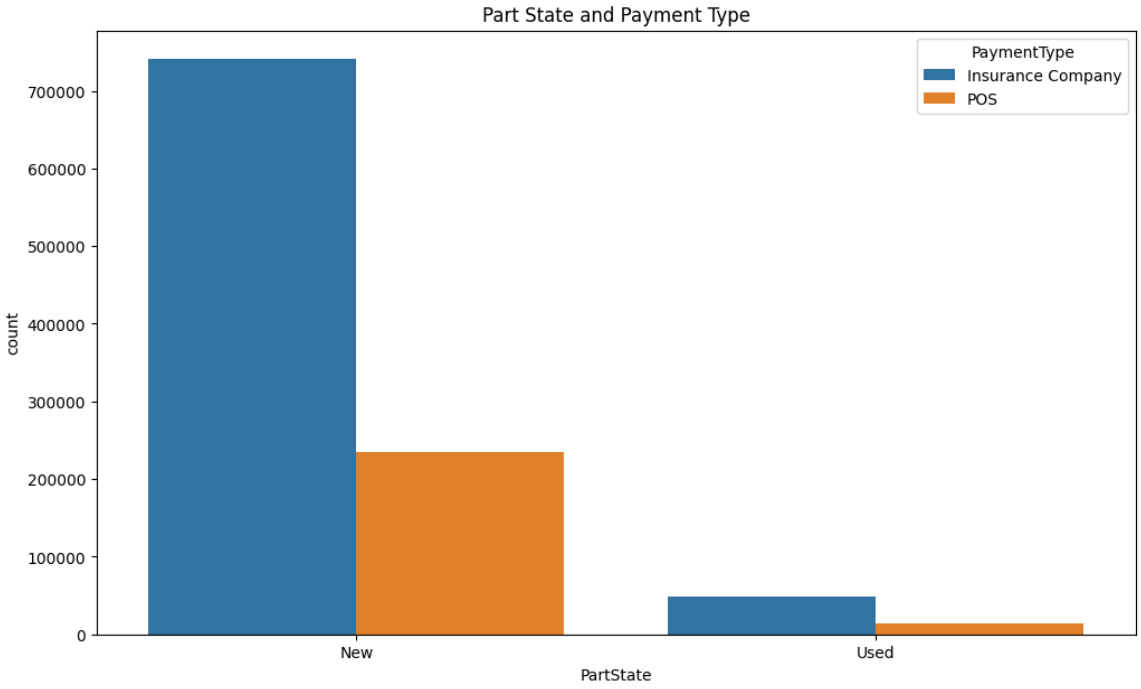
Insights:

- The Remmal Industrial is the Fastest Time Evaluation between the Areas
- The Orouba have good Time Evaluation after the Remmal Industrial
- New Industrial has the highest number of Time Evaluation Delay



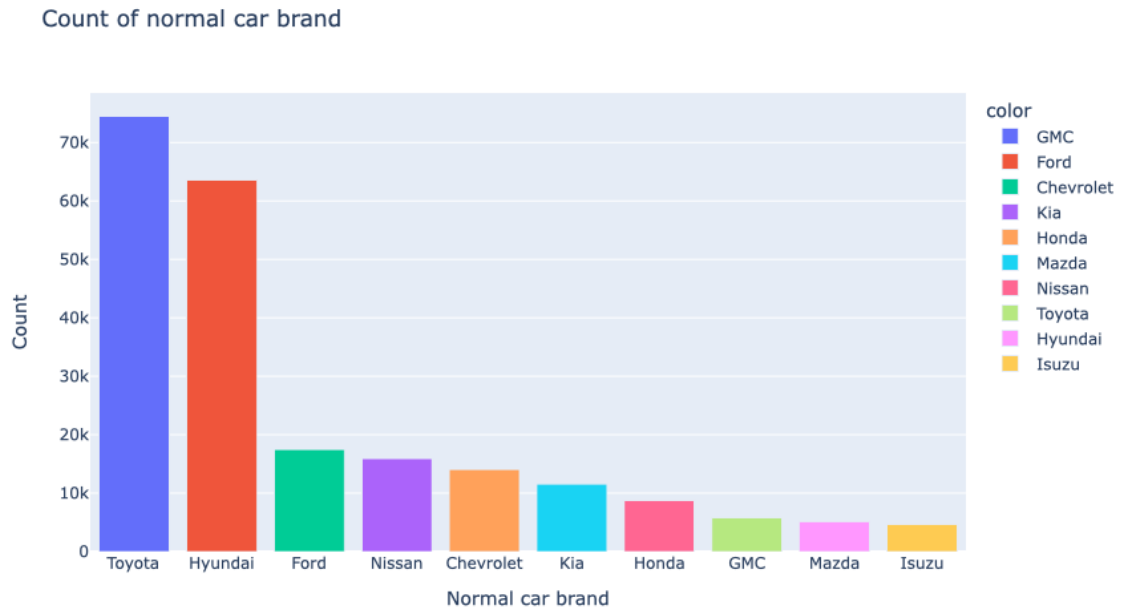
Insights:

- We can observe that the Orouba is the Highest Area have an Acceptable Evaluation
- We can observe that the New Industrial is the Highest Area have an unacceptable Evaluation.
- We can observe that there is no gap between areas in High Assessment Evaluation except Truck Center.



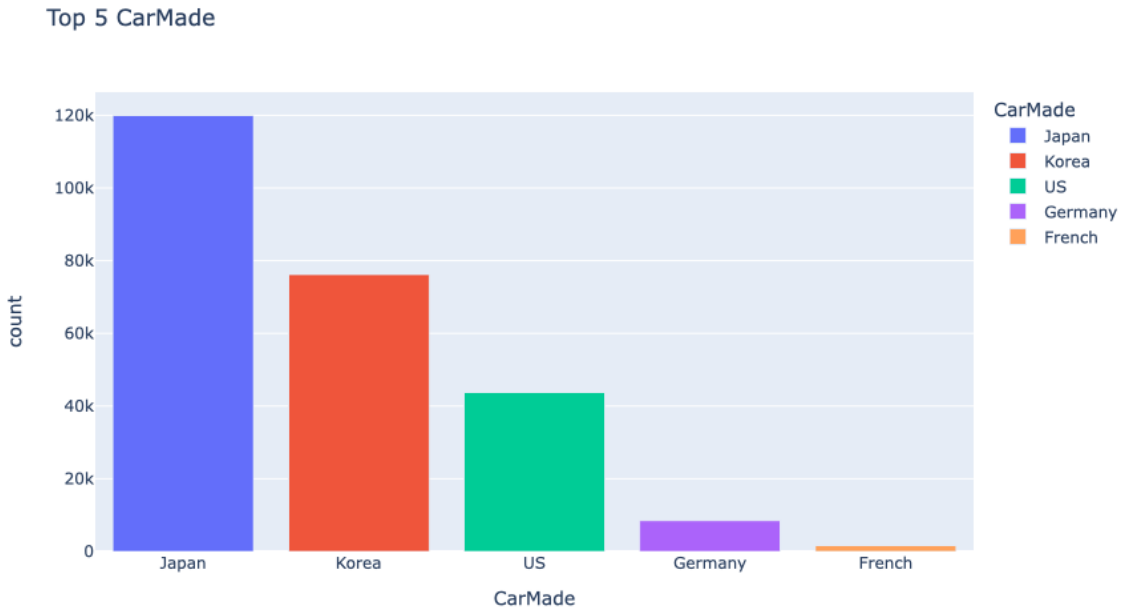
Insights:

- There's a huge difference between payment type when it comes to new parts. Insurance company covers most of them
- A very small difference can be seen when it comes to used parts. Although, the insurance company are more likely to cover the expense of used parts as well but there isn't a big difference when compared to POS
- We can conclude by saying that no matter what the state of the parts is, Insurance Company are more likely to pay for them



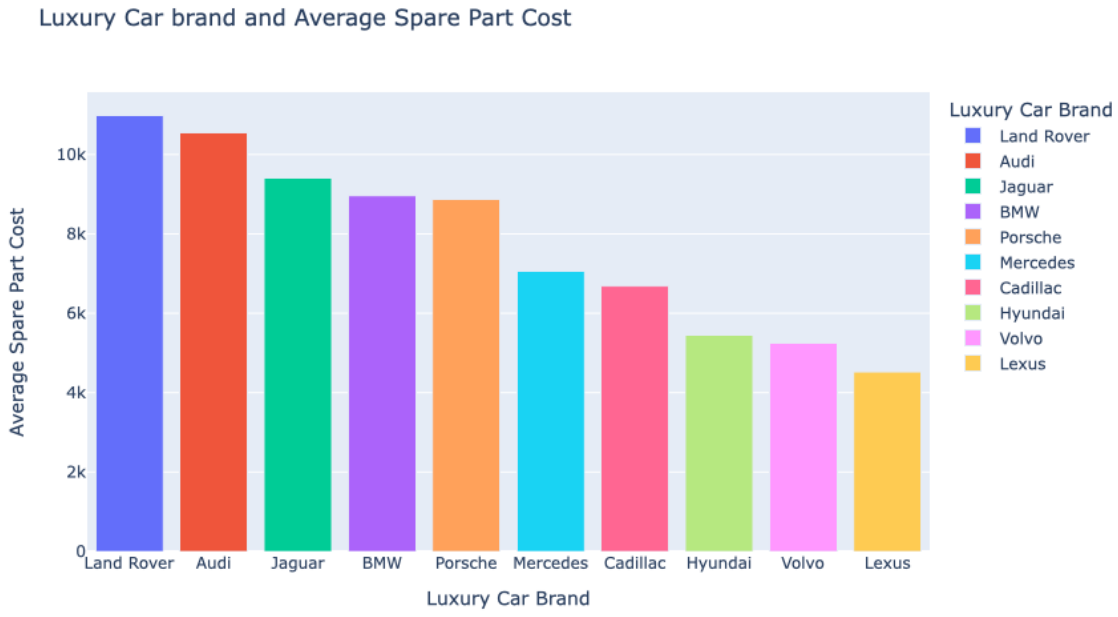
Insights:

- We can observe that Toyota and Hyundai are highly used in Riyadh in the Normal car Brand



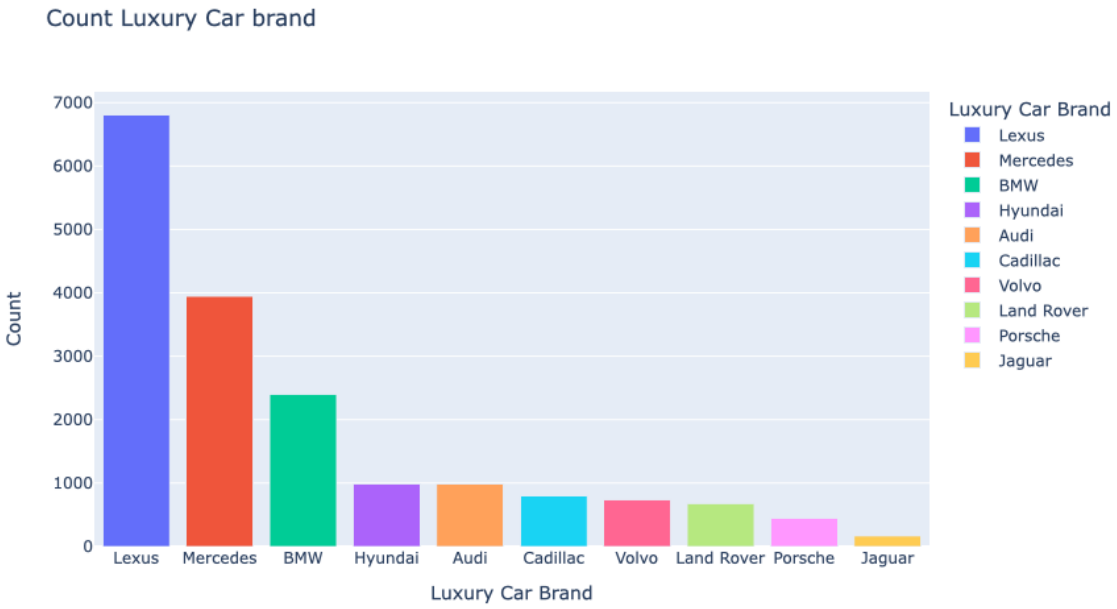
Insights:

- We can observe that Japanese cars is the most cars use in Riyadh
- We can observe that French car is the lowest use in Riyadh



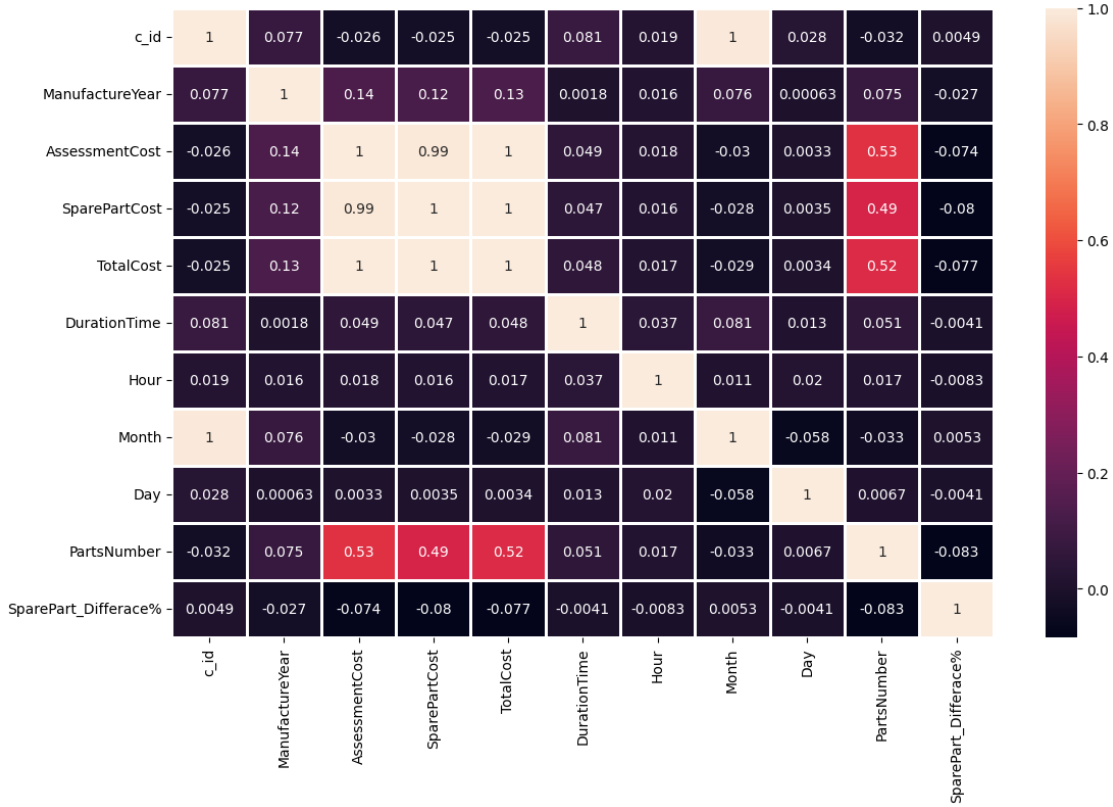
Insights:

- We can observe that Land Rover is the highest Spare part cost, and the Lexus is the lowest one.



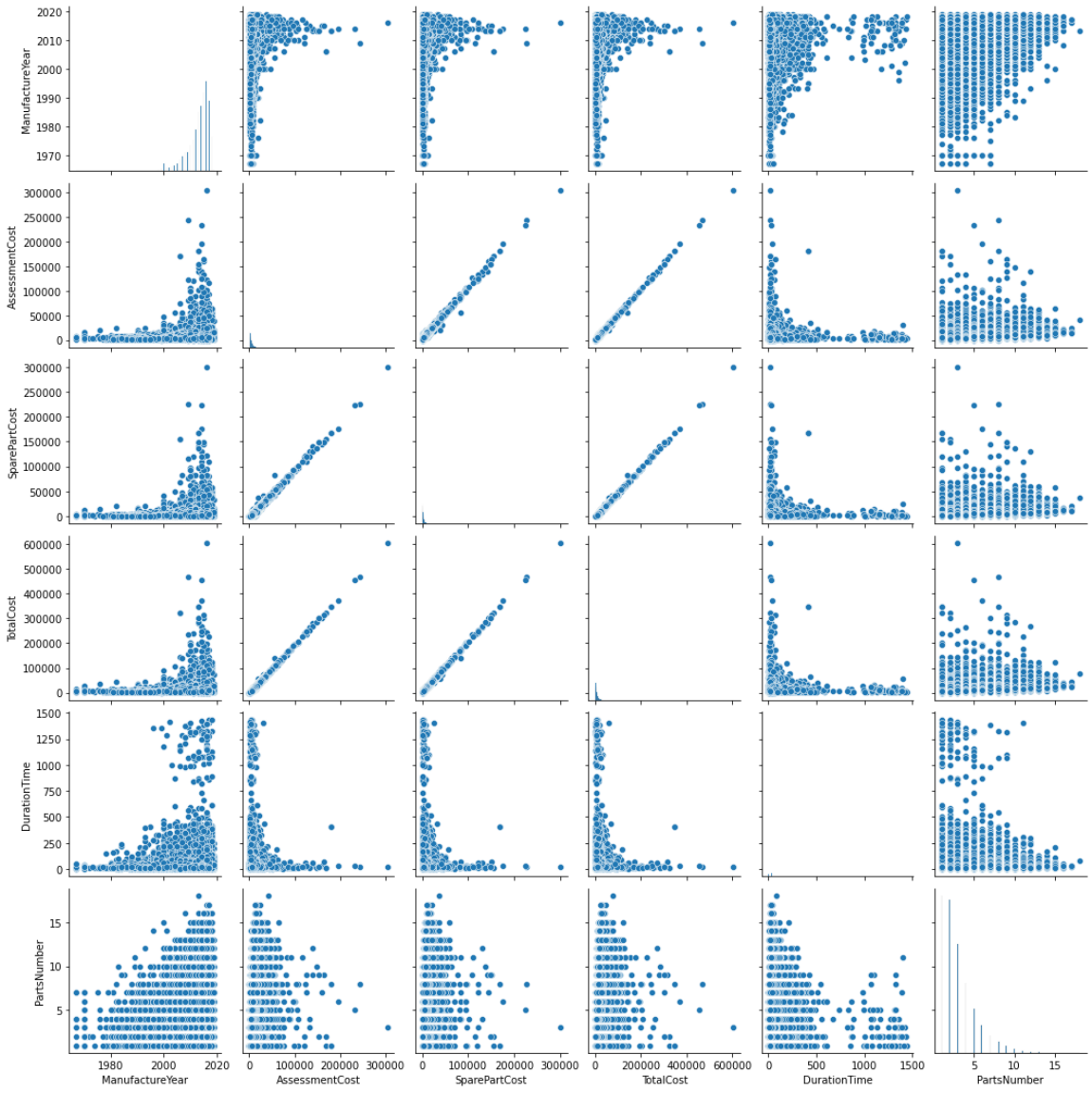
Insights:

- We can observe that Lexus is highly used in Riyadh and the Jaguar is the lowest used in the Luxury car brand.



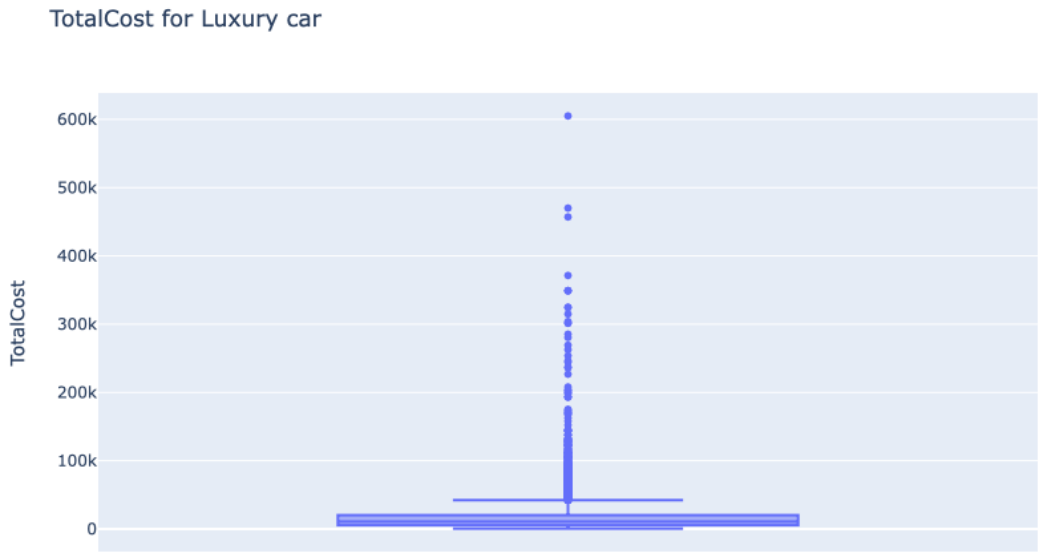
Insights:

- Total Cost it has Data leakage to Assessment Cost and Spare Part Cost
- The day of the week has a very weak realtion with the ManufactureYear
- Assessment Cost and Total Cost have a negative relationship with the month of the year



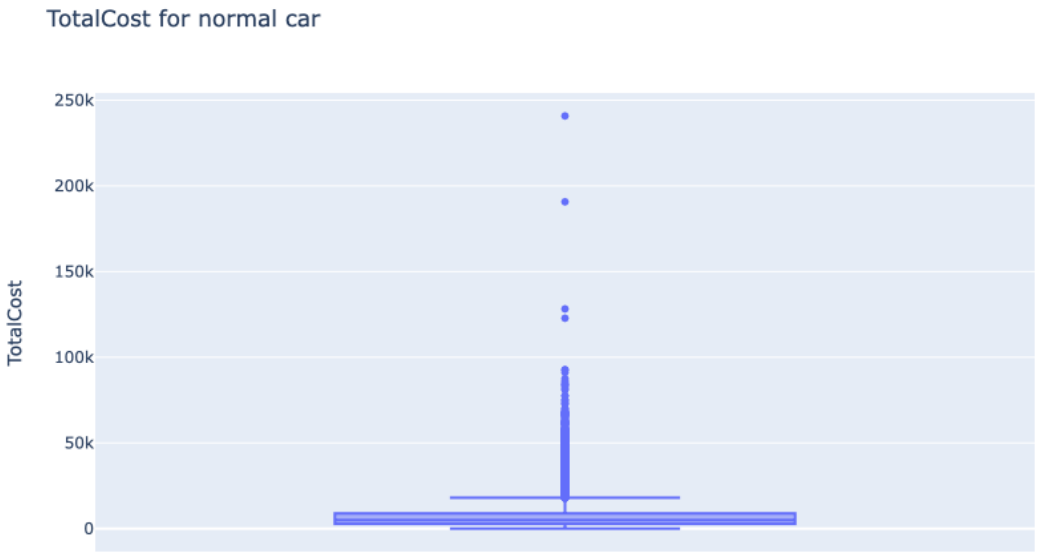
Insights:

- We can observe that when the Manufacture Year is increased the AssessmentCost will increase
- We can see from the graph that Total Cost it has data leakage to Assessment Cost and Spare Part Cost



Insights:

- As we can see, there are many outliers in 'TotalCost' in Luxury cars. we will remove it in the ML models
- The maximum total cost for a luxury car is 605.02K
- Any cost above 42236 is the outlier
- The minimum total cost for a luxury car is 525
- the median is 10.8K



Insights:

- As we can see, there are many outliers in 'TotalCost' in normal cars. we will remove it in the ML models
- The maximum total cost for a normal car is 240.8K
- Any cost above 18.03K is the outlier
- the median is 4972

Initial Conclusions:

1. We understand that most of the people take benefit of Car Insurance Companies, whenever an accident occurs
2. Most people prefer choosing a new part for their vehicles
3. Only a slight number of people go for used parts due to the legitimacy and life expectancy of the used parts

4. Left and right parts are the most vulnerable positions of the car whenever an accident takes place
5. The front part of the car is the least likely position to get damaged
6. New Industrial is the most visited branch center in Riyadh
7. There is a clear difference between the number of accidents that occur on weekdays and weekend. Weekdays are more dangerous for driving as compared to on the weekend
8. January is the Highest month in car assessment. Whereas December is the lowest month when compared to the other months
9. The Remmal Industrial is the Fastest Time Evaluation between the Areas
10. New Industrial has the highest number of Time Evaluation Delay
11. We can observe that the Orouba is the Highest Area have an Acceptable Evaluation
12. There's a huge difference between payment type when it comes to new parts. Insurance company covers most of them
13. A very small difference can be seen when it comes to used parts. Although, the insurance company are more likely to cover the expense of used parts as well but there isn't a big difference when compared to POS
14. We can observe that Toyota and Hyundai are highly used in Riyadh in the Normal car Brand
15. We can observe that Japanese cars is the most cars use in Riyadh
16. We can observe that Lexus is highly used in Riyadh and the Jaguar is the lowest used in the Luxury car brand.

17. We can observe that when the Manufacture Year is increased the AssessmentCost will increase

5. Building Classification Models

Regression Model:

- We choose the TotalCost to be our target in the Regression model and we had to drop the assessment cost and spere cost since it is considered as a data leakage

  XGBoost Regressor:

  We remove the unwanted Columns from a copy from the dataset

```
# Remove unwanted columns from Regression Training
unwanted_cols = ['c_id','RegistrationTime', 'CloseTime', 'Hour','Month', 'Day', 'WeekDay', 'PartsList', 'PositionList', 'PartStateList']
unwanted_cols.append('AssessmentCost')
unwanted_cols.append('SparePartCost')
unwanted_cols.append('SparePart_DifferaceX')
unwanted_cols.append('AssessmentEvaluation')
unwanted_cols.append('TotalCostEvaluation')

unwanted_cols.append('TimeEvaluation')
unwanted_cols.append('DurationTime')

unwanted_cols.append('CarColor')
unwanted_cols.append('CarClass')
# unwanted_cols.append('PartsNumber')



target = 'TotalCost'
```

Here is the Baseline Model for the Regression Models:

```
print(f'>>>>>>>>>>>>>>>>>>>>>>>>Baseline<<<<<<<<<<<<<<<<<<<<<<<<')
reg_baseline = calc_cost_regrs(reg_df[target],baseline=True)
```

```
>>>>>>>>>>>>>>>>>>>>>>>>Baseline<<<<<<<<<<<<<<<<<<<<<<<<
MSE: 72145653.64093144
MAE: 5863.779629443855
RMSE: 8493.859761082205
R2: 0.0
```

The final model evaluation:

```
[ ] # The final model evaluation
    print(f'>>>>>>>>>>>>>>>>>>>>>>>>Difference Between Train & Test Results<<<<<<<<<<<<<<<<<<<<<<<<')
    getAccuracy( X_test, y_test,XGB_R, X_train, y_train)
    cost_diff_reg(XGB_R_train_res, XGB_R_test_res, normalize=True)
```

```
>>>>>>>>>>>>>>>>>>>>>>>>Difference Between Train & Test Results<<<<<<<<<<<<<<<<<<<<<<<<
| Training Accuracy | Testing Accuracy |
|     82.96         |    72.07         |
The differance betewen the train accuracy and test accuracy is: 10.88%
The model in general is (Good)
The Erros is decreesing by:
MSE: -76.984 %
MAE: -8.456 %
RMSE: -33.035 %
R2: 13.119 %
```

The difference between the training accuracy and testing accuracy is: 10%

Classification Model:

We remove the unwanted Columns from a copy from the dataset

```
[ ] cls_df = df.copy()
    # Get dummies for parts regression dataframe
    cls_df = cls_df.join(cls_df['PartsList'].str.join('|').str.get_dummies().add_prefix('part_'))
    cls_df = cls_df.join(cls_df['PositionList'].str.join('|').str.get_dummies().add_prefix('pos_'))
    cls_df = cls_df.join(cls_df['PartStateList'].str.join('|').str.get_dummies().add_prefix('state_'))


    cls_df.head(1)
    # Remove unwanted columns from Regression Training
    unwanted_cols = ['c_id','RegistrationTime', 'CloseTime', 'Hour','Month', 'Day', 'WeekDay', 'PartsList', 'PositionList', 'PartStateList']

    unwanted_cols.append('AssessmentCost')
    unwanted_cols.append('SparePartCost')
    unwanted_cols.append('TotalCost')
    unwanted_cols.append('SparePart_DifferaceX')
    unwanted_cols.append('AssessmentEvaluation')

    unwanted_cols.append('TimeEvaluation')
    unwanted_cols.append('PartOfDay')
    unwanted_cols.append('DurationTime')

    unwanted_cols.append('CarColor')
    # unwanted_cols.append('CarClass')
    # unwanted_cols.append('CarType')
    # unwanted_cols.append('ManufactureYear')

    # unwanted_cols.append('PartsNumber')
    # unwanted_cols.append('PaymentType')

    # unwanted_cols.append('PartStateList')

    target = 'TotalCostEvaluation'

    cls_df.drop(unwanted_cols, axis=1, inplace=True)
```

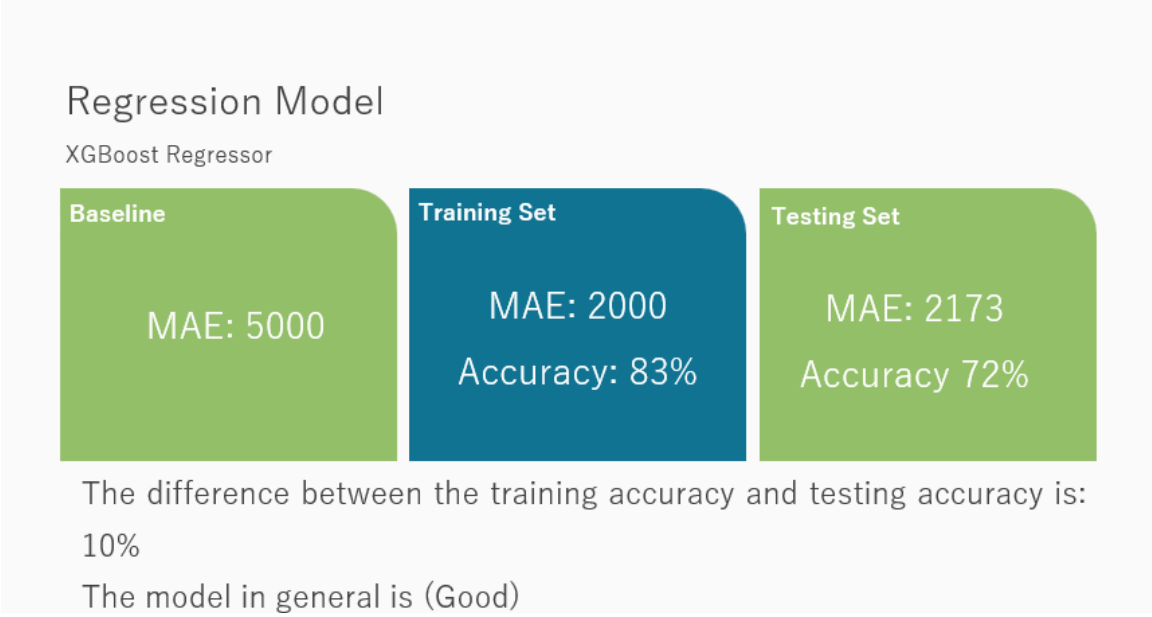
Here is the Baseline Model for the Classification Models:

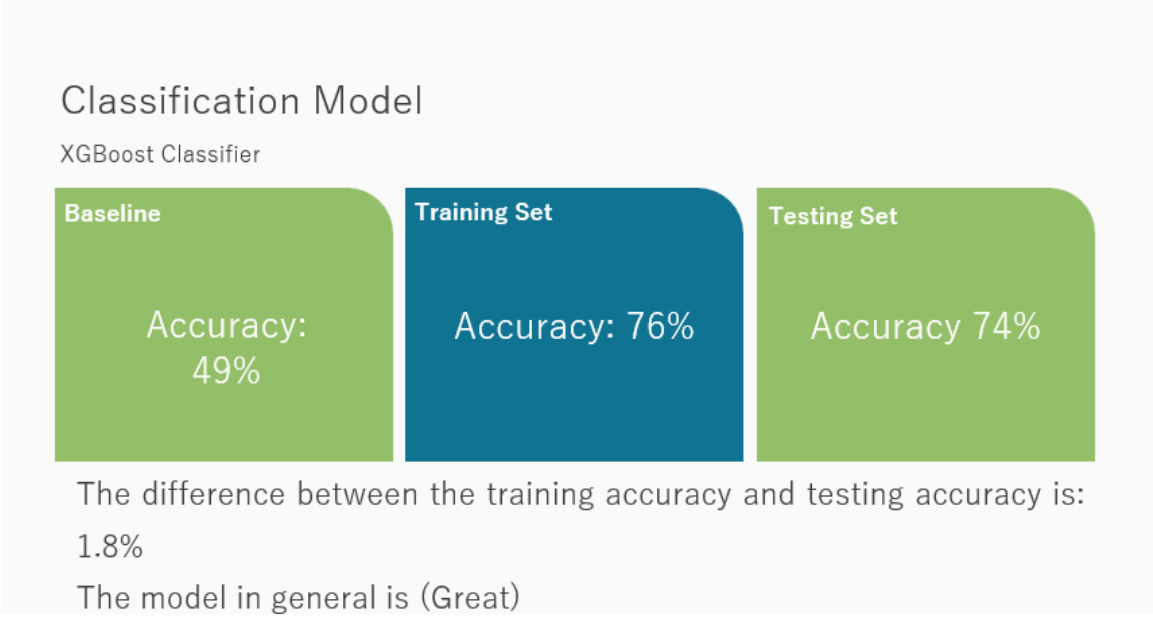
```
[ ] df[target].value_counts(normalize=True)*100

    Acceptable    49.179435
    Low           25.491007
    High          19.507462
    Very High      5.822096
    Name: TotalCostEvaluation, dtype: float64
```

Here is the final Model evaluation:

```
[ ] # The final model evaluation
    cost_diff_class(XGB_C_train_score,XGB_C_test_score)

    | Training Accuracy | Testing Accuracy |
    |     76.24         |    74.41         |
    The differance betewen the train accuracy and test accuracy is: 1.83%
    The model is (Great)
```

6. Results:


Regression Model
XGBoost Regressor
Baseline
MAE: 5000
Training Set
MAE: 2000
Accuracy: 83%
Testing Set
MAE: 2173
Accuracy 72%
The difference between the training accuracy and testing accuracy is:
10%
The model in general is (Good)


Classification Model
XGBoost Classifier
Baseline
Accuracy:
49%
Training Set
Accuracy: 76%
Testing Set
Accuracy 74%
The difference between the training accuracy and testing accuracy is:
1.8%
The model in general is (Great)

ML Codes:

Future work:

- Develop our models
- Contact Taqdeer to present our work
- Testing the models in the real-world life
- Training the models with more new data from Taqdeer.
- Improving the accuracy of the models
- Publish our system to use in Taqdeer centers