

**UNIVERSITI TUNKU ABDUL RAHMAN**

**Faculty of Information and Communication Technology (FICT)**

**UCCB3224 Data Mining Techniques**

**Assignment Jun 2023**

**Team Name: Group 20**

**Assignment Marksheet**

By signing below, we confirmed that the work produced was original and purely based on our own sentence construction. Should there be any plagiarism detected, we agreed mark penalization on the part(s) detected.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | **Name** | **ID** | **Practical**  **Group** | **Program (IA/IB/DE)** | **Signature** |
| 1 | Alvin Tan Yong Chuan | 2003184 | P4 | IB | *Alvin* |
| 2 | Carmen Lee | 2002631 | P5 | IB | *Carmen* |
| 3 | Pang Tse Ki | 2002067 | P3 | IB |  |
| 4 | Tan Ping Cheun | 2103476 | P5 | IB | *Cheun* |
| 5 | Wong Chee Yee | 2003002 | P4 | IB | A close-up of a signature  Description automatically generated |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Criteria** | **Weightage** | **Member**  **1** | **Member**  **2** | **Member**  **3** | **Member**  **4** | **Member**  **5** |
| **Individual:** |  |  |  |  |  |  |
| Final Deliverables (Quality of Work) | 4%  (Scale\*4 / 5) |  |  |  |  |  |
| **Team:** |  |  |  |  |  |  |
| Report and  Documentation | 3%  (Scale\*3 / 5) |  |  |  |  |  |
| Effort and Technical Capability (Overall as a Team) | 3%  (Scale\*3 / 5) |  |  |  |  |  |
| Final Deliverables (Quality of System) | 3%  (Scale\*3 / 5) |  |  |  |  |  |
| Innovativeness | 2%  (Scale\*2 / 5) |  |  |  |  |  |
|  | **15%** | % | % | % | % | % |

**Marking Scheme**

|  |  |
| --- | --- |
| **Scale (0-5)** | **Description** |
| 5 | Excellent work produced. Evidence of in-depth study **and** critical thought |
| 4 | Good work produced. Evidence of in-depth study **or** critical thought |
| 3 | Average work produced. Evidence of adequate study **or** thought/idea, although not extensively covered |
| 2 | Below average work produced. Evidence of some study **or** thought/idea, but not quite adequate. |
| 1 | Poor work performance **or** work **not** supported by any study/basis |
| 0 | Not attempted |

## **Executive Summary**

The goal of this data mining project is to solve the significant issues of fair pricing and transparency in the taxi service industry where the public lost trust on. Our descriptive analytics have revealed information on peak travel hours, well-travelled routes, and fare distribution, while our predictive algorithms have demonstrated great accuracy in projecting taxi fares. We have done the data-driven analysis to analyse the actual price. Moreover, there are several factors involved which includes distance travelled, duration of the trip and the number of passengers. Therefore, these factors can be used to further analyse. But firstly, must be processed then only performed the Exploratory Data Analysis (EDA). There are several visualisation tools had been used such as scatter plot, histogram, boxplot and many more to determine the data. For instance, according to the scatter plot that we have created, we acknowledge the longer the distance, the higher the travel fee. This is because the scatter plot formed an upward-sloping pattern. In addition, we also get to know the peak hour for occupancy of the taxi which is around 8 to 14 and 18 to 22 hours by using the histogram plot. On top of that, there are five data modelling being created such as the gradient boosting regressor, ridge regression, random forest and decision tree. This project establishes the foundation for data-driven decision making in the transportation sector, which has the potential to transform pricing strategies and customer satisfaction in taxi services for the benefit of both customers and service providers.

**Table Of Contents**

Contents

[**Executive Summary** 4](#_Toc144779777)

[**1.0** **Business Understanding** 6](#_Toc144779778)

[**1.2 Problem Framing** 7](#_Toc144779779)

[**1.3 Business Objectives** 8](#_Toc144779780)

[**1.4 Project Goals** 8](#_Toc144779781)

[**1.5 Project Motivation** 9](#_Toc144779782)

[**1.6 Current Solutions** 9](#_Toc144779783)

[**1.7 Gantt Chart** 10](#_Toc144779784)

[**1.8 Development Tools** 10](#_Toc144779785)

[**2.0** **Data Understanding** 12](#_Toc144779786)

[**3.0** **Data Preparation** 28](#_Toc144779787)

[**4.0** **Data Modelling** 29](#_Toc144779788)

[**5.0** **Evaluation** 34](#_Toc144779789)

[**Conclusion** 35](#_Toc144779790)

[**Reference** 36](#_Toc144779791)

## **Business Understanding**

**1.1 Business Information**

Nowadays, the transportation industry plays a vital role in this society. The transportation industry can conveniently people from one place to another, especially taxi services. This service has become a vital transportation mode or transportation tool for the part of the people due to it helps people to solve the problems such as time constraints or inconvenience of not having a vehicle. Hence, fair pricing and transparency have become significant aspects that will affect people's satisfaction and even trust when they enjoy this taxi service.

Fair pricing is about the reasonable price paid by the customer. The price paid must be reasonable, justifiable, and overly high than the service that the taxi provided to them. A reasonable price can increase customer satisfaction and trust making them feel it is worth it to enjoy this kind of service. Transparency is also vital same with the fair price. This is because transparency means that customers are clear about the prices. Clear about how the fare is calculated about the distance, such as how much customers need to pay for every mile [1]. Additional charges about some situations or even the situations are unexpected such as specific areas, peak hours, or the traffic condition.

Therefore, the challenges faced by the Taxi service are, how to reasonably calculate the fair price and the transparency, especially the special conditions for the additional charge. This is challenging due to the fair price and transparency must be reasonable and not minimize the customer's satisfaction and trust in the taxi service. To solve this problem, a data mining methodology is needed. Data mining has valuable insight from the large dataset to make wisely decision-making. Data mining can provide the analyse ride taxi data and provide conclusion about fair price and transparency they are reasonable.

## **1.2 Problem Framing**

**Unfair Taxi Fare Pricing**

There are some problems occur on taxi fare price. This is because the main problem is the local taxi is too expensive for customers to ride. It will occur some unfair situations which some ethical consideration on taxi drivers which they will not run the meter as others do. [2] In this case they will charge the price higher which is unfair for the customer. Taxi fares are calculated according to the meter so if the taxi driver does not do so his has the intend to scam customers. Besides, there are additional charged RM1 for the third passenger with maximum 4 sits or charged for the baggage store in boot. [3] Normally, these services should be included in the price according to the meter calculated. For e-hailing there had include these services and it just count for the distance for journey.

**Customer loyalty and pricing**

Price is a key factor that has a significant impact on customer loyalty. It's important to recognize that each customer has unique affordability thresholds that can heavily influence their decisions when purchasing or using services. [4] In this scenario, customers can choose from three different taxi fare levels: high, low, and moderate. By offering different pricing level, this allows them to select for a fare that fits their budget. Customers who prioritize premium service may select for a higher fare, while customers looking for a budget-friendly option may select for a moderate and lower fare.

## **1.3 Business Objectives**

To accurately predicting the fare prices according to several factors which includes the distance travelled, duration of the trip and additional charges. Hence the company can ensure the customers are charged appropriate fares that equalise the value of the service given.

## **1.4 Project Goals**

1. Build Accurate Predictive Models: To precisely anticipate taxi prices, use predictive models such as gradient boosting regressors, decision trees, random forests, and linear regression. These models directly contribute to the goal of determining fair pricing and making accurate fare predictions.
2. Model Comparison and Selection: Evaluate the performance of various predictive models, including linear regression, decision trees, random forests, and gradient boosting regressors, to determine which one provides the best accuracy and reliability for fare predictions.
3. Feature Importance Analysis: To determine which factors such as distance travelled, duration of the trip, and any additional factors that affect the cost. The goal of pricing optimisation is supported by this information, which influences pricing methods.
4. Hyperparameter Tuning: Improve the prediction accuracy and resilience of the models by fine-tuning their parameters, such as tree depth in decision trees, the number of estimators in random forests, and gradient boosting regressors.
5. Model Monitoring and Maintenance: Keep an eye on the performance of predictive models, such as gradient boosting regressors, decision trees, random forests, and linear regression, to make sure they continue to be correct over time. Regular model maintenance helps to ensure fair pricing and the satisfaction of the customers.
6. 6. RMSE and R2 goals: The goal of final RMSE to achieve is maximum 5.00 and the goals of R2 (percentage accuracy) need to be achieved is minimum 75%.

## **1.5 Project Motivation**

The taxi services have slowly started to wane. This is because most of the people nowadays have the concerns of the taxi services due to the unfair pricing and the prices are lack of transparency. Therefore, by building this data mining models, it can help to predict the fair prices accurately and more transparency.

## **1.6 Current Solutions**

To solve the taxi pricing transparency problem, the data-driven pricing analysis need to be used to analyse and predict the actual price for customer by using data mining technique. This approach aims to enhance the pricing transparency, build the trust to the customers, and make sure the taxi prices they received is reasonable. There have few of key components to solve the current problems.

First component is the data understanding and preprocessing. Data that those related to taxi pricing, distance travelled, duration of the trip, and the total number of occupants need to be collected. The data need to be cleaned and pre-processed to ensure the dataset is unify so that it can be suitable for analysis.

The second component is data mining and analysis. The Exploratory Data Analysis (EDA) need to be performed to explore and analyse the dataset. This can be beneficial to identify the patterns and correlations between attributes by visualization such as using scatter, bar and line chart which can really influence the taxi fare.

The third component need to be done is predictive modelling. This predictive model is developed to predict the fair pricing from taxi. The models need to see the key factors such as distance, duration and occupancy.

The fourth component that needs to be done is evaluation. The evaluation was done by using the test set in the dataset and applying it to the final model which decided on the fine-tuning phase to predict the more accurate price.

By using the data mining technique, this solution may solve the fair and transparent pricing issues in transportation industry. It is not only can make sure the customers charged reasonably, but also build the trust and satisfaction, this can lead to improve customer relationships and business growth.:

## **1.7 Gantt Chart**



The Gantt chart for the project outlines the key milestones and tasks spanning from August 5th to September 11th. During the project initiation phase (August 5th to August 13th), the focus was on understanding the given tasks, conducting a thorough literature review, defining data mining goals and business objectives, articulating project motivation, and analysing the current solution. This foundational work allowed for a smooth transition into the Data Understanding phase (August 12th to August 14th), where the team described the data and performed exploratory data analysis (EDA). Subsequently, the Data Preparation phase (August 14th to August 19th) enabled the preparation of the data for analysis. The Model Selection and Development phase (August 20th to August 25th) was dedicated to developing data mining models. Following this, the models were diligently evaluated from August 25th to August 30th. Finally, deployment planning took place from August 30th to September 2nd. The project concluded with a comprehensive summary and documentation finalization (September 2nd to September 4th) and a project presentation on September 11th. This structured timeline ensured a systematic and efficient progression through the project phases, facilitating successful project completion.

## **1.8 Development Tools**

1. Integrated Development Environment (IDE)

Jupyter Notebook will function as our primary development environment, offering an interactive notebook-based interface that visualizations and explanations that stimulating a well-structure documentation of our analysis. It offers a dynamic, notebook-based user interface that enables seamless blending of code, justifications, visualizations, and analysis. The creation of organized, recorded analysis depends on this integration. We can create and run code in small pieces with Jupyter Notebook, view the results right away, and add rich text explanations right along with our code. This feature improves our work's clarity while also making it easier to collaborate with team members or share our insights with others. The need to transition between tools is decreased since we can carry out data visualization, statistical analysis, and model creation in the same environment.

In addition to Jupyter Notebook, we use Colab for certain aspects of our projects. Google's cloud-based platform Colab provides a supplement to our on-site development environment. It has a number of benefits, including the availability of extra computing power like GPUs and TPUs, which can greatly speed up computationally demanding activities. Colab makes it easier to collaborate by letting numerous users edit the same document at once. This is especially helpful when we need to consult with coworkers or seek advice from specialists.

1. Programming language

The main implementation of the project will rely heavily on the Python programming language. Python's capabilities include efficiently processing large data sets, performing complex mathematical operations, and creating multi-dimensional arrays. Additionally, Python has an extensive ecosystem of data analysis and machine learning libraries that allow us to efficiently preprocess, analyse, and model datasets.

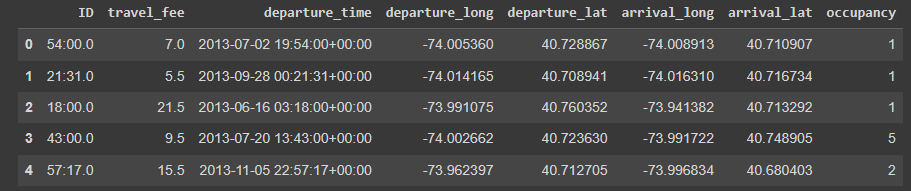
1. Data analysis and manipulation libraries

To undertake data preprocessing and exploratory data analysis, our reliance will be placed on widely used libraries such as Pandas and NumPy. These libraries provide powerful tools for cleaning and transforming datasets to make them suitable for analysis. While Pandas is user-friendly and NumPy has excellent execution speed.

## **Data Understanding**

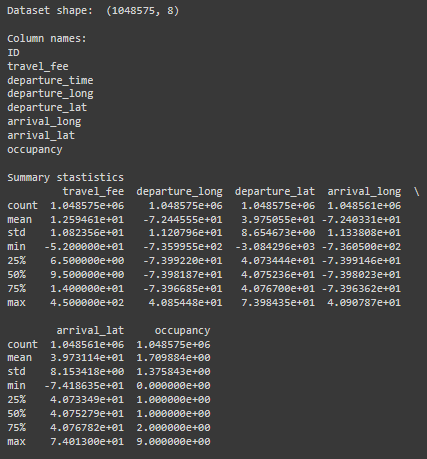
**2.1 Loading Dataset**

1. The dataset.csv is being loaded and read to use in the following sections. The diagram below is the table head of the dataset after being loaded.

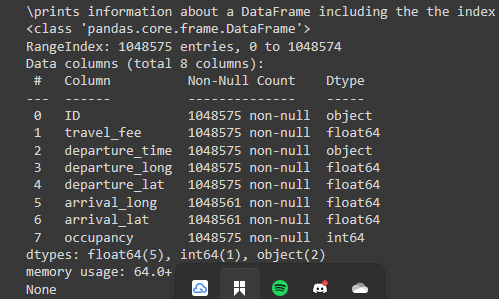


**2.2 Data Description and Visualization**

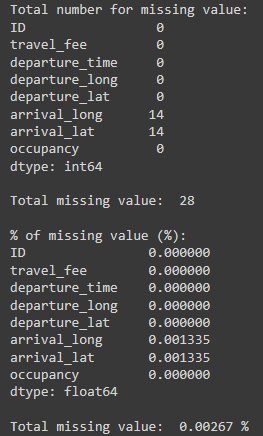
1. Based on the diagram below, the dataset has 1048575 rows of data and 8 attributes in the dataset. The diagram below also has provided the attributes which are ID, travel fee, departure time, departure longitude, departure latitude, arrival longitude, arrival latitude, and occupancy. Furthermore, it also shows the summary statistics in diagram below.



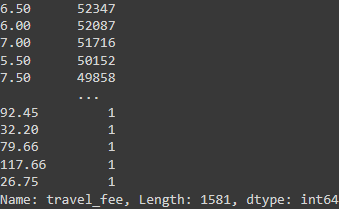
1. The diagram below shows how many non-null and data types used in every attributes.



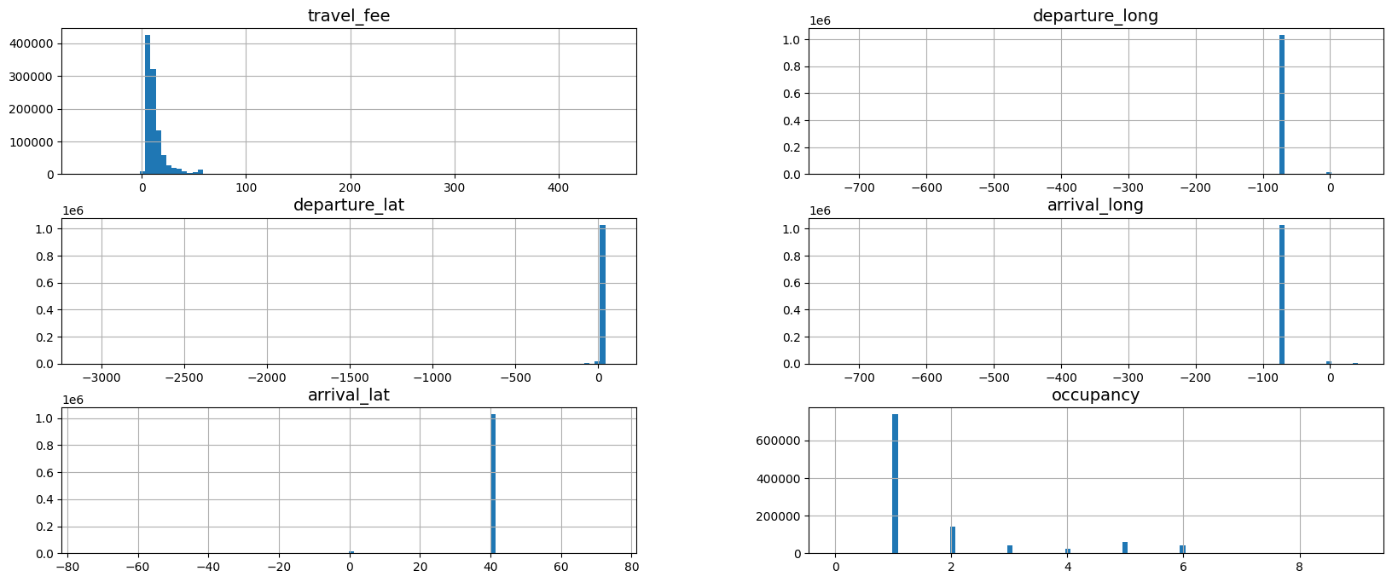
1. The following diagram shows the total number and percentage of missing value for every attributes. From the diagram below, it shows 14 missing values in arrival longitude and latitude attributes.



1. The diagram below shows how many counts for rows of data in every travel fee.



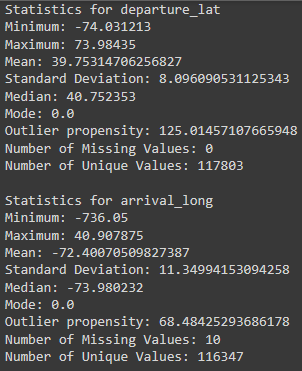
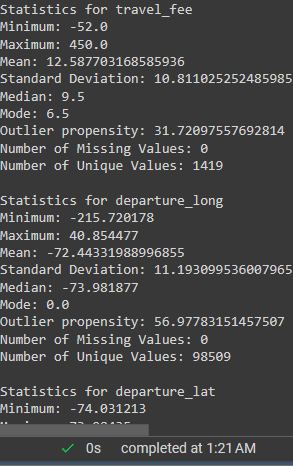
1. The diagram below shows the histogram that used for every attributes.

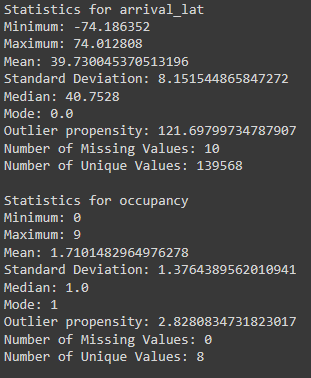


1. The diagram below shows the splitting of the dataset in train set and test set. It shows that there is about 80% of dataset is going to train set and 20% going to test set. After splitting the dataset, there are total of 838860 rows going to train set and 209715 rows going to test set.

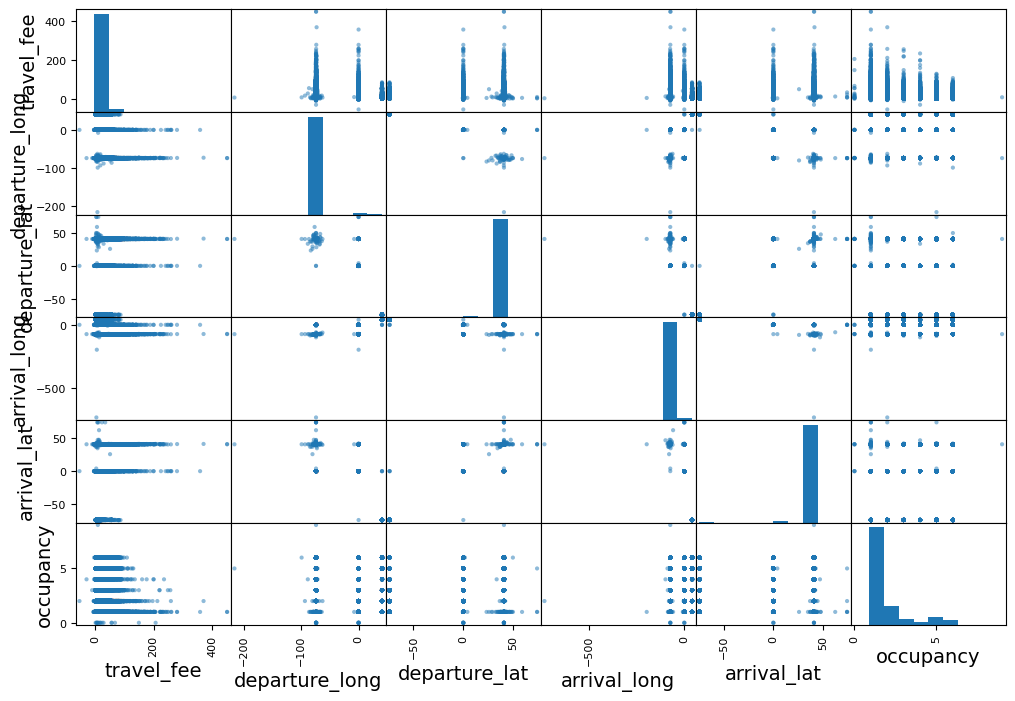


1. The diagrams below show the more in-depth statistics for every attribute including the minimum value, maximum value, mean, standard deviation, median, mode, outlier propensity, number of missing values and number of unique values.

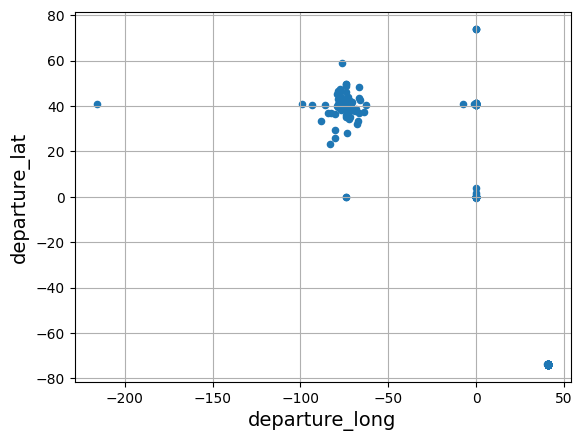




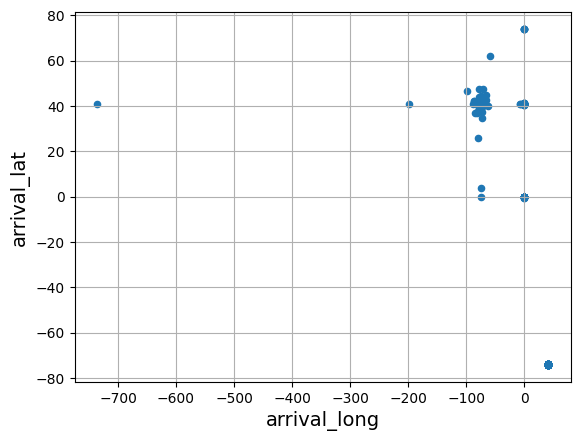
1. The next diagram shown the scatter matrix of every attribute which to analyse the relationship between attributes.



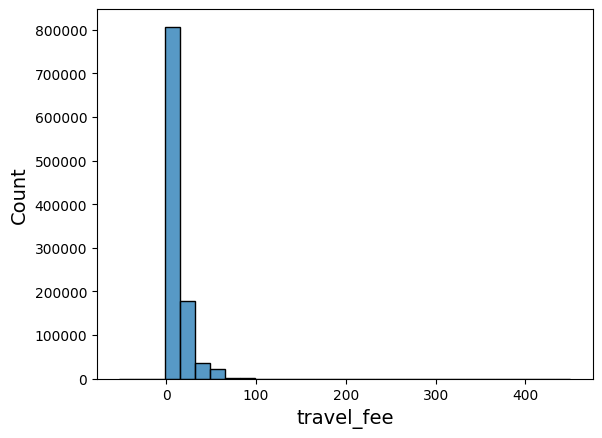
1. The diagram below shows the initial and original scatter graph for the relationship of departure longitude and latitude. This graph may not see any proper shape or any relevant relationship of these two attributes.



1. The diagram below shows the initial and original scatter graph for the relationship of arrival longitude and latitude. This graph is similar with the diagram above which may not see the shape and relationship of these two attributes.



1. The diagram below visualizes about the histogram graph with how many rows of dataset in a range of travel fee.

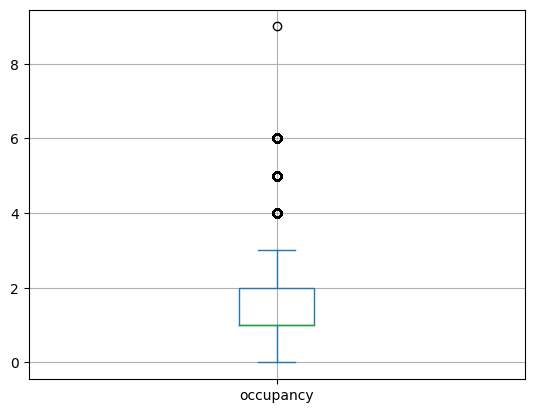


1. The diagram below illustrates the histogram graph which shows how many occupancies according to the range of departure time.

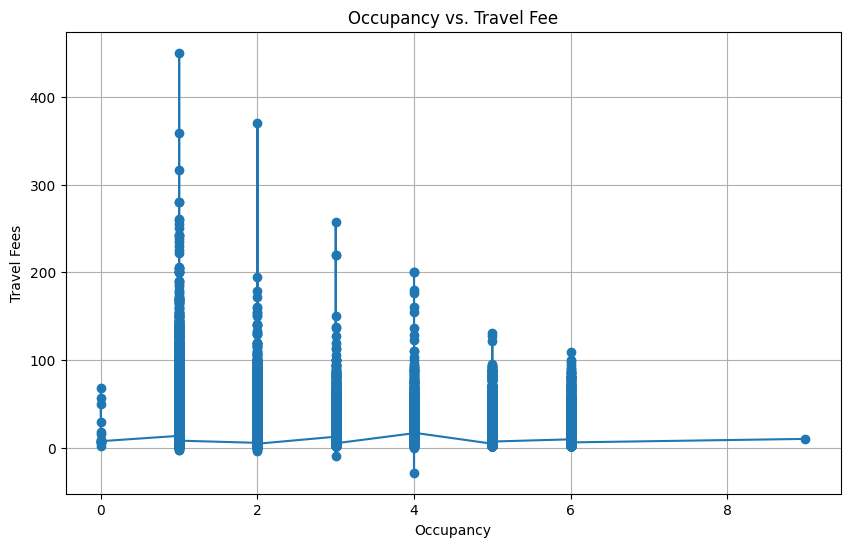
A graph with blue squares

Description automatically generated

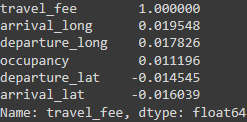
1. The diagram below visualizes about the box-plot graph with the shows the median value of travel fee per person.



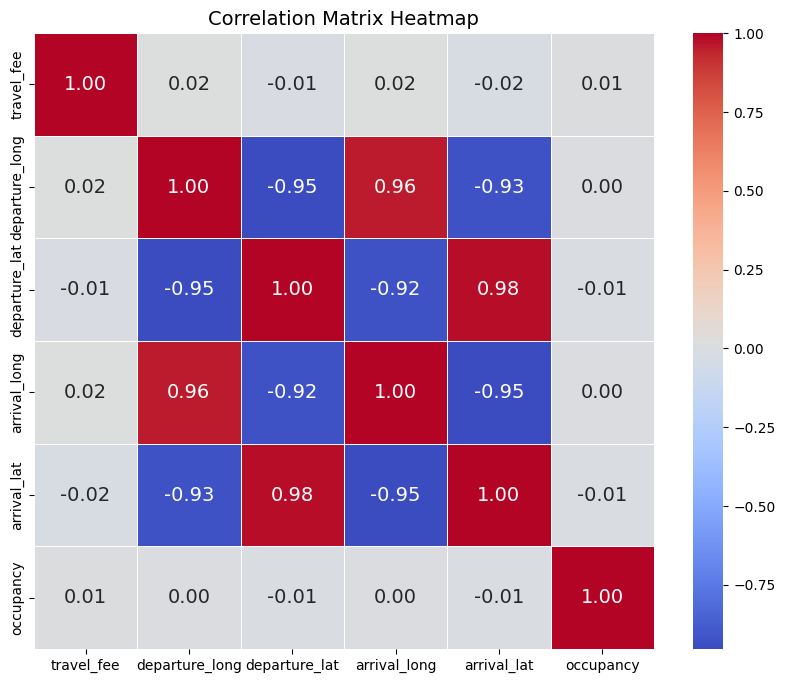
1. The diagram below visualizes about the line graph with the sorted data of occupancy and travel fee.



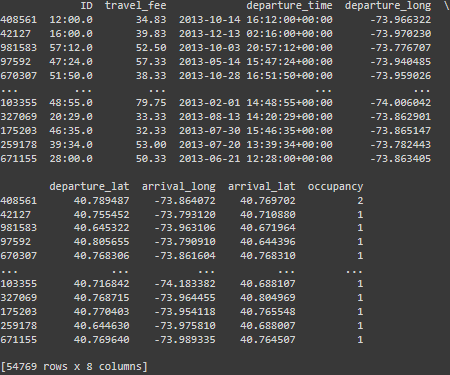
1. The diagram below shows the correlation matrix of travel fee between other attributes.



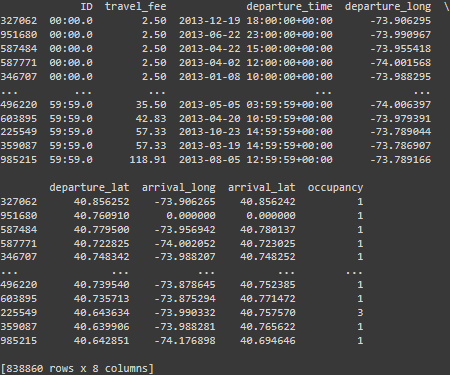
1. The diagram below shows the correlation matrix heatmap which shows the numeric variables of our datasets. The heatmap shows the Pearson correlation coefficients between pairs of numeric variables.



1. The diagram below shows the filtered data in training set of travel fee that more than RM30



1. The diagram below shows the table which sort the travel fee in ascending order which from the lowest price to highest.

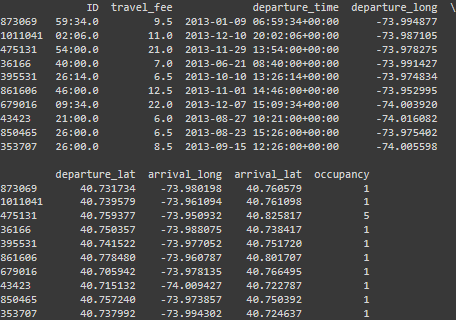


1. The diagram below shows the column selected in dataset.

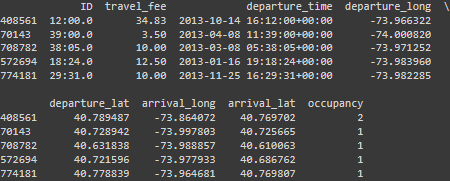
A black screen with white numbers

Description automatically generated

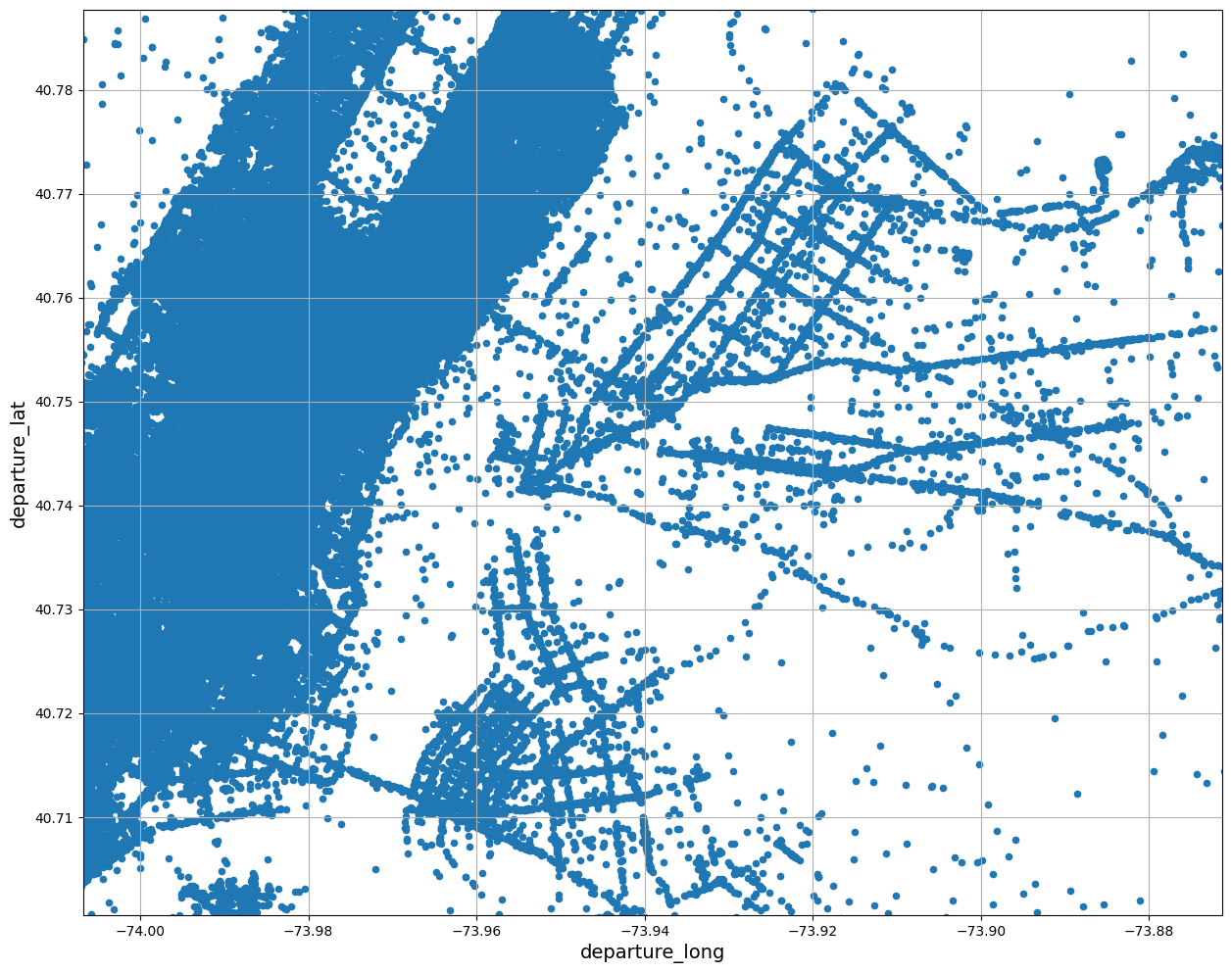
1. The diagram below shows the sorted column data of ID and travel fee into ascending order.



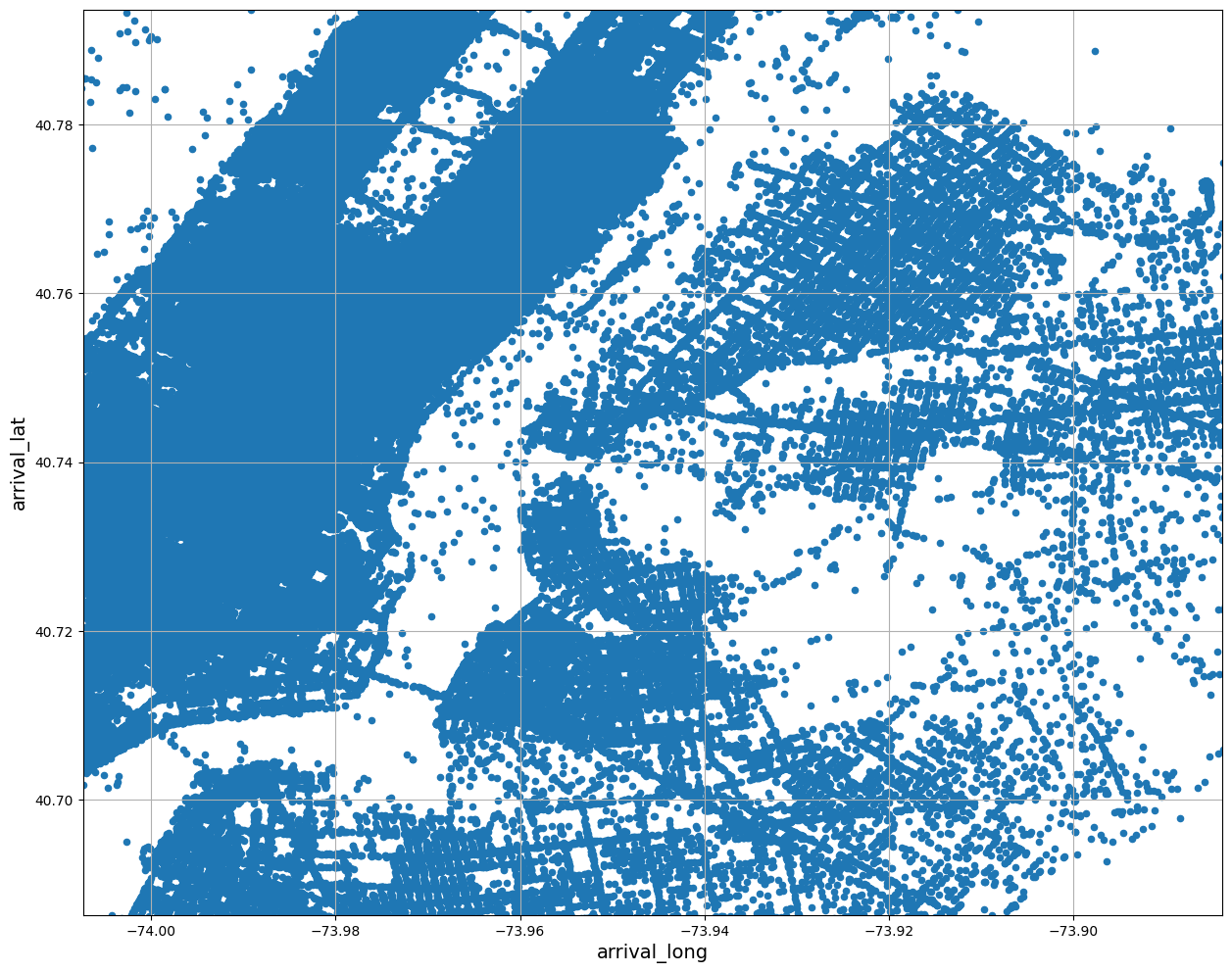
1. The diagram below shows the selected fist five rows.



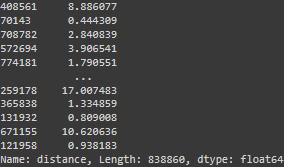
1. The diagram below shows the scatter graph of departure longitude and latitude after quantile. By making the quantile in 0.05 and 0.95, the graph becomes more meaningful and interpretable. The graph shows the location of that rows.



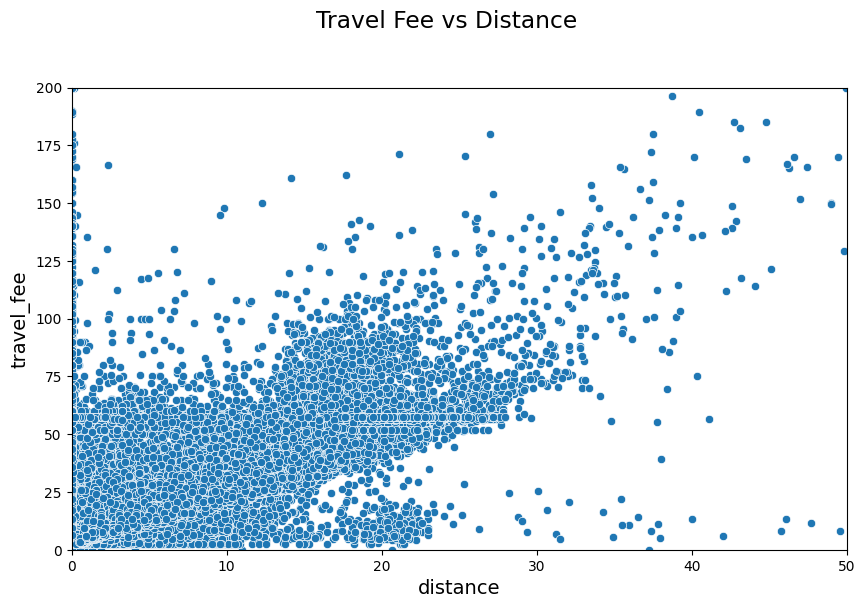
1. The diagram below shows the scatter graph of arrival longitude and latitude after quantile. By making the quantile in 0.05 and 0.95, the graph becomes more meaningful and interpretable. The graph shows the location of that rows.



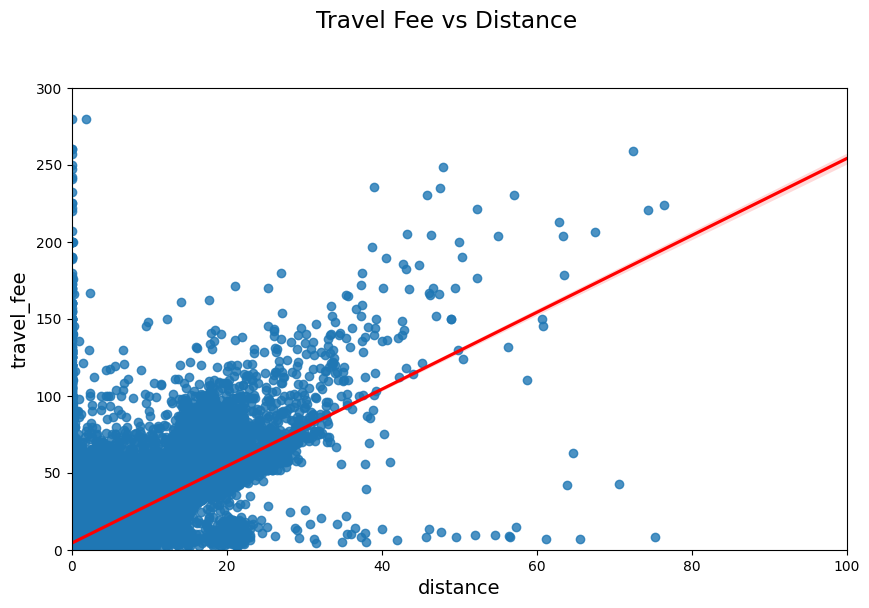
1. The diagram below shows the table of ID and distance, which is the new attribute created by departure longitude, departure latitude, arrival longitude and arrival latitude by using haversine formula.



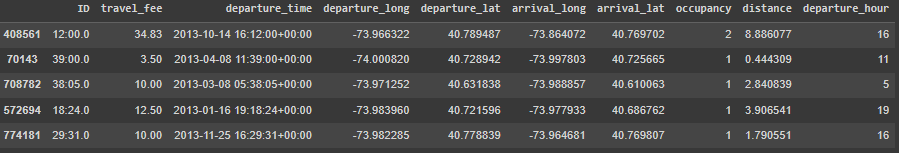
1. The diagram below illustrates the scatter graph between the travel fee and the distance in the axis of 0 to 200.



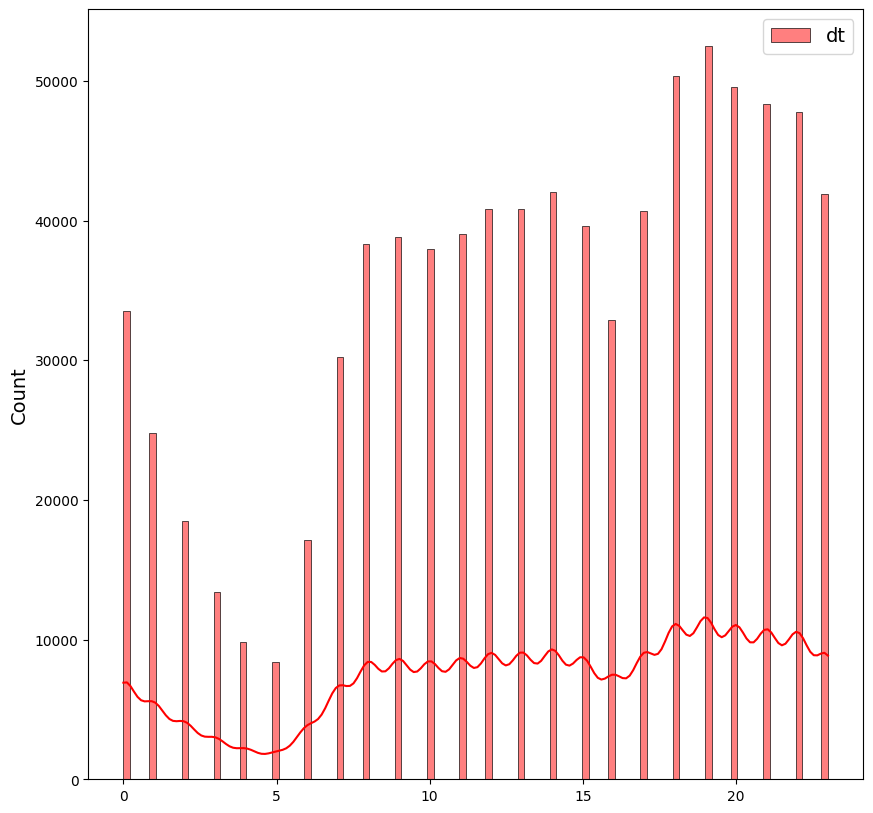
1. The diagram above shows the same scatter graph as the diagram above where it has the relationship between travel fee and distance. However, at this graph, the data points show the travel fee/distance ratio is below a threshold (0.1). Hence, it is removed to ensure the outliers are being removed as well so that it can create more accurate model. In this graph, the axis may become from 0 to 300, and the line, kws included. Based on this scatter graph, it shows the relationship between these attributes due to the higher the travel fees the longer the distance.



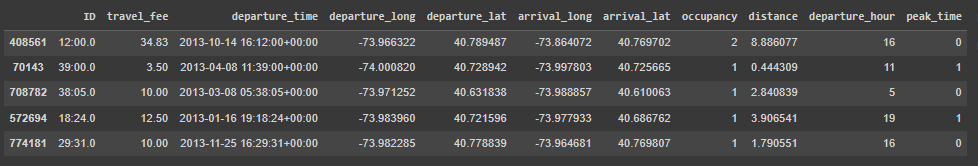
1. After filtering, the outliers remove about .
2. The diagram below shows the table. In this table, the new attribute, departure hour is created based on the departure time attribute.



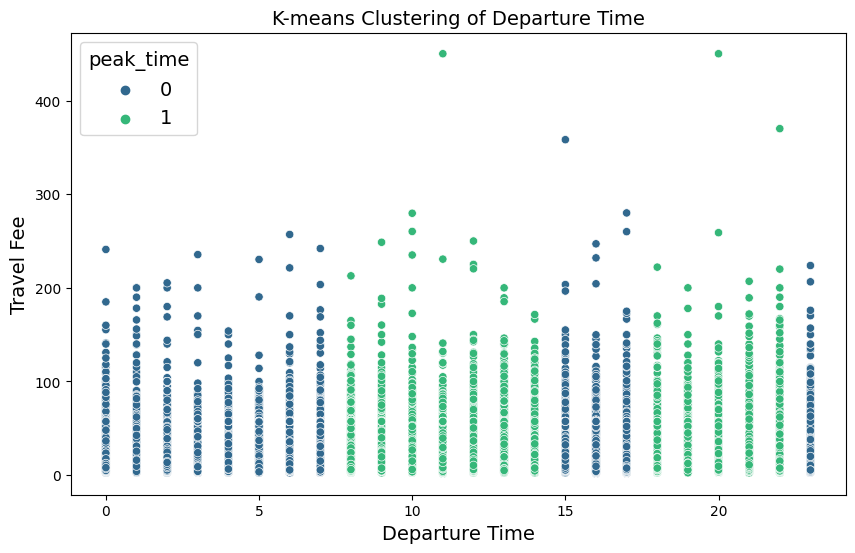
1. The diagram below shows the histogram graph with the red line based on the departure hour in the train set with after filtering. This graph is generated purposely to know which hour is a peak time. Based on the graph, the peak time is between 8 to 14 and 18 to 22 hours.



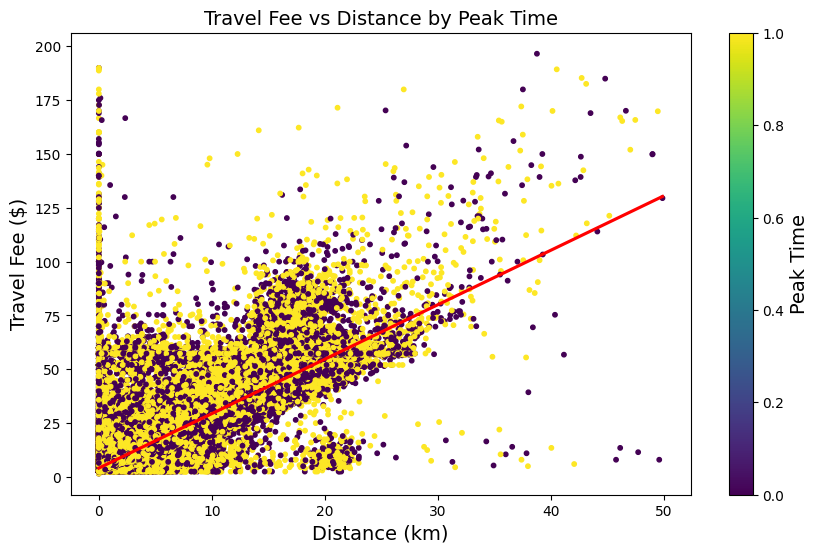
1. The diagram below shows the table which includes the peak time attributes. The classify\_peak\_time function is used to identify which rows is having the peak time and for those who are having the peak time, it will return 1.



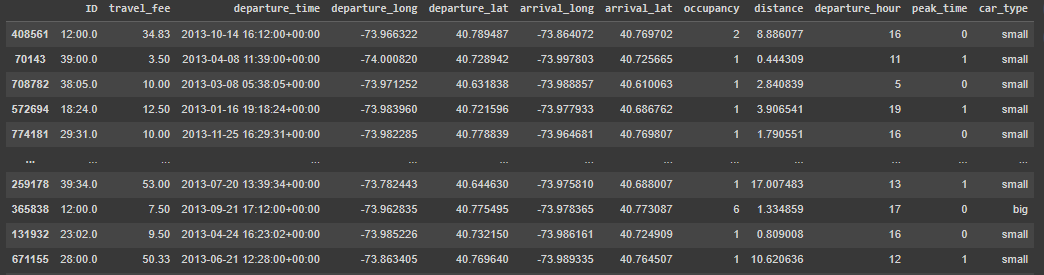
1. The diagram below shows the scatter graph between departure time and travel fee with the peak time plot.



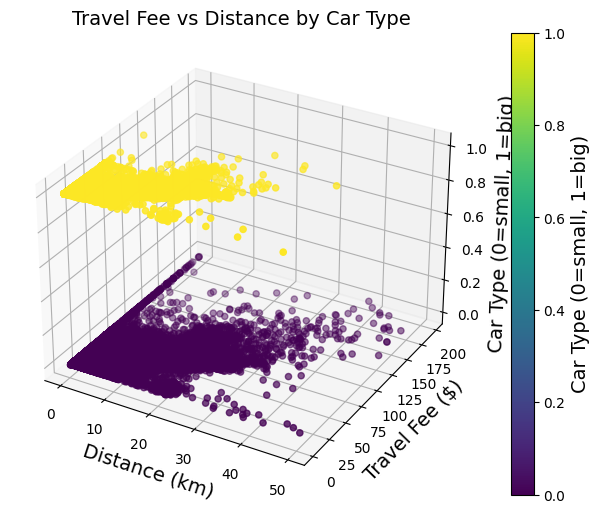
1. The diagram below shows the scatter with reg graph between travel fee and distance by peak time. At this point the train set will undergo another round of filtering, which it may filter out the data rows with travel fee more than $200 and distance more than 50km to remove the outliers and noise.



1. In diagram below shows the table with the new attribute, car type. The car type has classified to small and big car, which means, if it consists of equal or more than 5 occupants, it classifies to the big car, oppositely will be classify to small car.

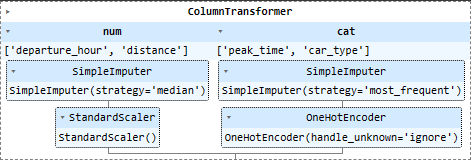


1. The diagram below shows the 3D graph between travel fee and distance by car type. If it is big car, it may categorize to 1, oppositely will categorize to 0 as a small car.

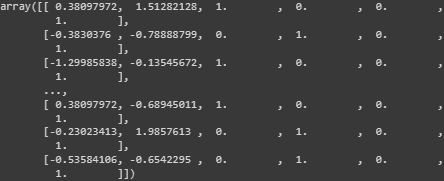


## **Data Preparation**

1. In data preparation, price will serve as a X\_train and price\_labels will be designated as the y\_train. After that, the rows that contain with missing value will be dropped. Furthermore, the simple imputer has been used in it to fill in any missing value, make\_pipeline will be used to create the composite estimator of a sequence of transforms and an estimator. The column transformer will be used in different transformations to different attributes of the dataset. Additionally, the standard scaler will be used to standardize the feature and one hot encoder will be used to convert the categorical variable into indicator. Moreover, there have defining the num\_attr which contain the numerical attributes and cat\_attr which need to include the categorical variable. After that, the pipeline is created for numerical and categorical. Finally, the preprocessing object is created which is a column transformer which can apply the transformation in numerical and categorical pipelines to the appropriate attributes of a dataset.



1. The preprocessing object will be fit and transform with the X\_train (price).



1. Diagram below shows the features name in preprocessing object.

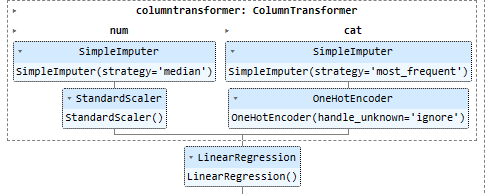


1. The normal dataset preprocessing will also be performed to facilitate a comparison with filtered dataset, helping to determine which method is more suitable to use.

## **Data Modelling**

**4.1 Comparing Normal Dataset and Filtered Dataset**

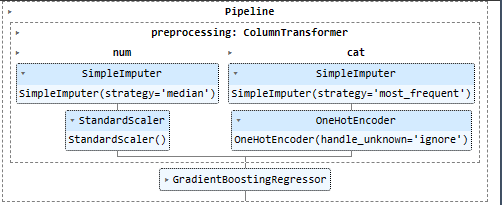
To compare the suitability of the normal dataset and filtered dataset on the modelling, the linear regression method is being used to compare it.



|  |  |
| --- | --- |
| **Dataset Type** | **Root Mean Square Error (RMSE)** |
| Normal Dataset | 5.80 (2d.p) |
| Filtered Dataset | 5.21 (2d.p) |

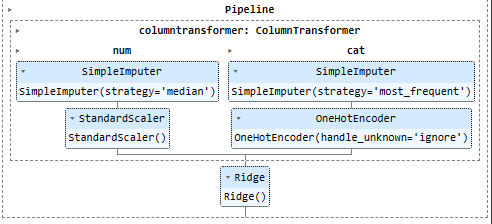
Based on the table above, the RMSE of filtered dataset is smaller than the normal dataset. Thus, filtered dataset is more suitable to use compared to normal dataset.

**4.2 Gradient Boosting Regressor**



The RMSE of Gradient Boosting Regression after conducting cross validation is 4.83 (2 d.p) and the percentage accuracy (r2) is 79.42%.

**4.3 Ridge Regression**

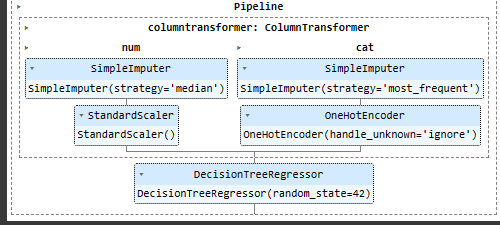


The RMSE of Ridge Regression after conducting cross validation is 5.21 (2 d.p) and the percentage accuracy (r2) is 76.08%.

**4.4 Random Forest Regressor**

The RMSE of Random Forest Regression after conducting cross validation is 5.31 (2 d.p) and the percentage accuracy (r2) is 75.02%.

**4.5 Decision Tree Regressor**



The RMSE of Decision Tree Regression after conducting cross validation is 6.48 (2 d.p) and the percentage accuracy (r2) is 62.86%.

**4.6 Shortlisted Modelling**

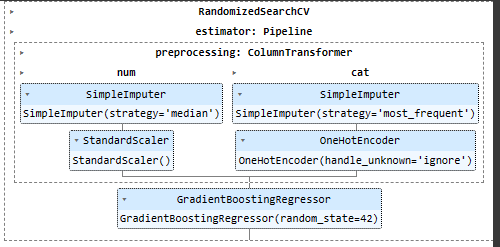
|  |  |  |
| --- | --- | --- |
| **Modelling** | **RMSE** | **Percentage Accuracy (R2)** |
| **Gradient Boosting Regressor** | 4.83 | 79.42% |
| **Ridge Regression** | 5.21 | 76.08% |
| **Random Forest Regressor** | 5.31 | 75.02% |
| **Decision Tree Regressor** | 6.48 | 62.86% |

Based on the table above, the gradient boosting, ridge and random forest regression will be the shortlisted modelling.

**4.7 Fine Tuning**

**4.7.1 Gradient Boost Regressor**

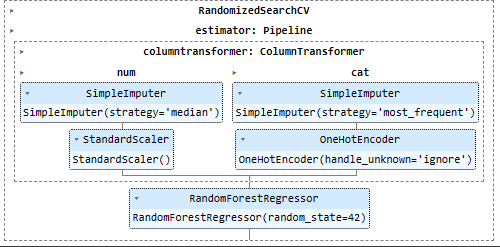
The randomize search cross validation will be applied in this gradient boosting regression. The estimator will range from 1 to 50, the learning rate will be set to 0.01, 0.1 and 0.5 and the max depth will vary from 1 to 5.



From the result, the best parameter for learning rate is 0.5, maximum depth is 4 and estimator is 29. The best RMSE is 4.85 (2d.p) and the percentage accuracy is 79.94%.

**4.7.2 Random Forest Regressor**

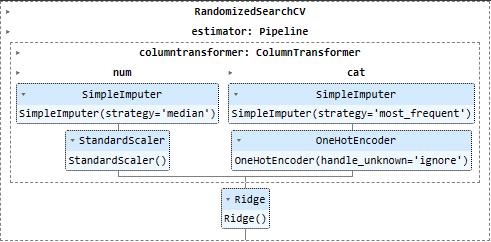
The randomize search cross validation will be applied in this random forest regression. The estimator will range from 1 to 50, max features are auto and sqrt, minimum samples split will vary from 2 to 10, minimum samples leaf will range from 1 to 10 and the max depth will span from 1 to 10.



From the result, the best parameter for max depth is 8, max feature is auto, minimum samples leaf is 4, minimum sample split is 9 and the estimator is 24. The best RMSE is 4.82 (2d.p) and the percentage accuracy is 79.83%.

**4.7.3 Ridge Regression**

The randomize search cross validation will be applied in this ridge regression. The ridge\_alpha will be set as uniform.



From the result, the best RMSE is 5.21 (2d.p) and the percentage accuracy is 76.08%.

**4.7.4 Best Fit Modelling in Evaluation**

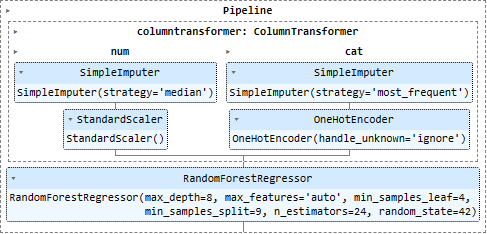
|  |  |  |
| --- | --- | --- |
| **Modelling** | **RMSE** | **Percentage Accuracy (R2)** |
| **Gradient Boosting Regressor** | 4.85 | 79.94% |
| **Random Forest Regressor** | 4.82 | 79.83% |
| **Ridge Regression** | 5.21 | 76.08% |

Based on the table above, Random Forest Regressor has the lowest RMSE which is the best fit of the model to the data. However, Gradient Boosting Regressor got the highest R-squared (R2) which this model is more of the variance in the target variable.

The Random Forest Regressor will be chosen due to Random Forest model might be better due to it is more important to do the precise prediction.

## **Evaluation**

The Random Forest Regressor model will be chosen as test set evaluation model.



Based on the final evaluation, the final RMSE is 4.84 (2d.p) and the final accuracy percentage (R2) is 79.45%.

## **Conclusion**

As a result, this data mining initiative has solved important problems in the transportation sector, particularly those related to taxi services. The issue stems from the high cost of local taxi fares, leading some taxi drivers to adopt the unethical business practices. The main difficulty is that instead of charging according to the meter, some drivers choose to overcharge, which is unethical. Therefore, fair pricing and transparency were the main issues that needed to be addressed. Pricing also has a significant impact on consumer loyalty, with high, low and medium fares. Customers have the freedom to choose a fare that fits their budget thanks to this pricing flexibility. This pricing strategy aims to increase overall consumer loyalty within the business while serving a wide customer base. Offering passengers fair prices that are in line with the value offered while considering variables like distance, travelled time, and additional fees was termed as fair pricing. We developed and identified a real-world business challenge using data mining approaches. Moreover, the weakness of the study might be bias in data. Past pricing practises, such as discriminatory pricing or biased driving behaviour, may have left their mark on historical statistics. Besides, the weakness is model interpretability, it might be difficult to communicate forecasts to customers when using complex models with interpretability issues, such as gradient boosting regressors. In the contrary, additional data modelling will be conducted to identify the best-fitting model in the future. Furthermore, time will be scheduled strategically to allow for more dedicated hours to work on refining the algorithm.

## **Reference**

[1] A. Noulas, V. Salnikov, R. Lambiotte, and C. Mascolo, “Mining open datasets for

transparency in taxi transport in metropolitan environments,” EPJ Data Science, vol. 4,

no. 1, Dec. 2015, doi: https://doi.org/10.1140/epjds/s13688-015-0060-2. ‌

[2] N. Says:, I. R. Says:, and E. Y. says:, “Malaysia Taxi Services Fares,” University, <https://www.university-malaysia.com/useful-information-for-foreign-students/malaysia-taxi-services-fares/> (accessed Aug. 17, 2023).

[3] “Useful tips on using Malaysian taxi services,” Useful tips on using Malaysian taxi services – klia2.info, <https://www.klia2.info/taxis/useful-tips-on-using-malaysian-taxi-services/> (accessed Aug. 17, 2023).

[4] S. B. Assegaff and S. O. Pranoto, “Price determines customer loyalty in ride-hailing services - researchgate,” Price Determines Customer Loyalty in Ride-Hailing Services. , <https://www.researchgate.net/publication/340236752_Price_Determines_Customer_Loyalty_in_Ride-Hailing_Services> (accessed Aug. 19, 2023).