Analyzing Public Transportation Fare Discrepancies in Hong Kong

Group Number: A1



Group Members:

Lie Warren Leander (3036094306)

Winiera Sutanto (3035964671)

Huang Da Fang (3036052205)

Clara Natalie Surjadjajadi (3036029686)

## 

## 

## **Table of Contents**

1. Objective

2. Highlights

3. Background

4. Methodology

4.1 Project Focus and Approach

4.2 Conceptual Framework

4.3 Initial Methodological Approach

4.4 Methodological Evolution

4.5 Definition of "Unreasonably Expensive"

5. Bus Dataset

5.1 Data Acquisition

5.2 Data Cleaning

5.3 Exploratory Data Analysis

5.4 Model and evaluation

5.4.1 OLS Regression

5.4.2 Random Forest

5.4.3 XGBoost

5.5 Interpretations and Communications

5.5.1 Key Insights from Bus Data

5.5.2 Overall Conclusions for Bus Fares

5.5.3 Reflection on Methodological Change

5.5.4 Limitations and Future Work for Bus Analysis

6. MTR Dataset

6.1 Data Acquisition

6.2 Data Cleaning

6.3 Exploratory Data Analysis

6.4 Data Modelling and Evaluation

6.5 Interpretations and Communications

7. References

## **1. Objective**

The primary objective of this data science project is to analyze public transportation costs across Hong Kong districts, with a particular focus on bus and MTR fares, to determine whether claims of "unreasonably expensive" transportation in certain areas are supported by data. Specifically, the project aims to:

1. Conduct comprehensive analysis of bus and MTR fare structures across Hong Kong's 18 districts to identify patterns, disparities, and potential inequities.
2. Develop robust predictive models that accurately capture the factors influencing public transportation fares, including distance, geographic location, service providers, and route characteristics.
3. Establish a data-driven methodology to objectively compare transportation costs between different areas of Hong Kong, accounting for relevant socioeconomic factors such as population density and median income.
4. Identify districts with potentially "unreasonably expensive" public transportation by establishing statistical thresholds and evaluating outliers against model predictions.
5. Provide evidence-based insights to inform public discourse on transportation equity and accessibility in Hong Kong.

## **2. Project Highlights**

* **Identified Key Fare Determinants:** Advanced modeling techniques (XGBoost, Random Forest) revealed that MTR fares are primarily influenced by travel distance (R² = 0.96), while bus fares are significantly affected by both specific route groups (particularly harbor crossings) and journey distance (R² = 0.88), providing evidence-based insights into Hong Kong's public transportation pricing structure.
* **Discovered Geographic Disparities:** Data analysis identified specific districts with disproportionately expensive transportation: Tuen Mun and Yuen Long having the highest percentage of "unreasonably expensive" bus routes (21.0% and 18.5% respectively), while MTR routes showed minimal pricing anomalies (0-2% in most districts) except for specific cross-island routes in Tsuen Wan district and Racecourse station.
* **Methodological Innovation**: The project demonstrated significant improvement in model accuracy by pivoting from traditional "fare per kilometer" metrics to raw fare analysis with distance as an independent variable, increasing explanatory power from approximately 30% to over 88% (R²), highlighting the importance of methodological flexibility in transportation pricing research.

## **3. Background**

Public transportation constitutes a critical infrastructure component in Hong Kong, facilitating approximately 11.5 million passenger journeys daily across diverse transport modalities. Buses independently account for 4 million daily passenger trips, while the Mass Transit Railway (MTR) serves as the primary rapid transit system. Despite the widely acknowledged efficiency and comprehensive coverage of Hong Kong's public transportation network, persistent questions regarding fare equity and affordability across different geographical regions have emerged in public discourse.

Recent fare adjustments have precipitated discussions regarding transportation equity, with empirical evidence from media reports indicating significant resident dissatisfaction in specific areas. In Tung Chung, survey data suggests that more than 80% of residents express concerns regarding transportation expenditure. Analogous discussions have emerged regarding differential fare structures between rail corridors, particularly comparisons between the East Rail Line and Tuen Ma Line, prompting academic and public policy debates regarding fare determination mechanisms and equity considerations.

These observed disparities in transportation expenditure can have substantial implications for household financial management, particularly affecting lower and middle-income households residing in peripheral districts who typically encounter extended commuting distances and potentially higher cumulative fares. The economic burden is further intensified when fare increases exceed inflation metrics, raising fundamental questions regarding the equitable distribution of transportation costs across Hong Kong's socioeconomically diverse landscape.

Prior research and journalistic investigations have suggested that residents in particular geographical areas, including Tuen Mun, Yuen Long, and the Islands District, may experience disproportionately elevated transportation costs relative to more centrally situated districts. However, these assertions frequently lack comprehensive statistical analysis incorporating the multifaceted factors that influence transportation pricing structures.

This research aims to address this analytical deficiency by applying rigorous data science methodologies to objectively evaluate transportation costs across Hong Kong's administrative districts. Through the systematic analysis of governmental datasets encompassing route specifications, fare structures, and socioeconomic indicators, this study endeavors to develop multivariate models capable of identifying statistically significant fare anomalies and elucidating the primary determinants of any observed geographical disparities in transportation costs.

## 

## **4. Methodology**

### **4.1 Project Focus and Approach**

This research employs a comprehensive, multi-modal approach to investigating public transportation costs across Hong Kong, focusing specifically on both bus and MTR fare structures. The central research questions guiding this investigation are: (1) What factors significantly influence public transportation fares in Hong Kong? (2) Do certain districts bear disproportionately higher transportation costs? and (3) Can objective criteria be established to identify "unreasonably expensive" transportation in specific areas?

The investigation adopts a data-driven perspective that transcends anecdotal evidence and media narratives about transportation inequities. By leveraging quantitative methodologies and statistical modeling, we aim to provide empirical evidence regarding fare structures and potential disparities. This approach allows us to move beyond subjective perceptions of "expensive" transportation and establish objective parameters for evaluation.

Our dual focus on both bus and MTR systems provides a holistic view of Hong Kong's transportation ecosystem, recognizing that residents often make modal choices based on comparative costs, journey times, and accessibility. This multi-modal analysis enables us to identify whether certain districts face disadvantages across multiple transportation options or whether alternative modes might mitigate high costs in certain areas.

### **4.2 Conceptual Framework**

The investigation employs a systematic framework for transportation cost analysis that incorporates multiple variables affecting fare structures across different transportation modes. Our analysis recognizes that complex interactions between distance-based factors, geographic considerations, service provider variables, and infrastructure requirements determine transportation costs.

The conceptual framework acknowledges that fare structures in public transportation systems reflect operational realities (service provision costs) and policy objectives (affordability, equity, and access). This creates a complex analytical challenge requiring multi-factorial analysis rather than simplistic comparisons. Our framework therefore, incorporates:

1. **Distance Metrics**: Recognizing the fundamental relationship between journey length and operational costs
2. **Geographic Determinants**: Accounting for topographical challenges, tunnels, bridges, and other infrastructure requirements specific to Hong Kong's unique geography
3. **Provider Characteristics**: Including company-specific operational models, fleet composition, and service standards
4. **Service Classifications**: Differentiating between route types (e.g., urban, suburban, cross-harbour, express)
5. **Socioeconomic Context**: Incorporating district-level demographic and economic indicators to contextualize fare structures

This integrated framework enables us to distinguish between fare variations that reflect legitimate operational differences versus those that might indicate inequitable pricing structures.

In our initial project proposal, we formulated specific transportation cost models for both bus and MTR services:

* For buses, we proposed a model incorporating bus fare, stop density, number of unique routes, and median income, expressed as f(Bus Fare, Stop Density, Number of Unique Routes, Median Income).
* For MTR, we proposed a model incorporating fare, car capacity, train frequency, population density, and median income, expressed as f(Fare, Distance, Car Capacity, Train Frequency, Median Income).

These initial frameworks guided our early data collection efforts but were subsequently refined as we gained deeper insights into the transportation data.

### **4.3 Initial Methodological Approach**

Our original methodology, as outlined in the project proposal, focused heavily on calculating and comparing average fare per kilometre metrics across different districts. The proposed calculations for bus fares included:

1. **Average Bus Fare per Kilometre for a bus**:
2. **Average Bus Fare per Kilometre for a stop**:
3. **Average Bus Fare per Kilometre for a district**:

The initial methodology also specified excluding stops less than 1 kilometer apart, based on research suggesting passengers are willing to walk between 600-1300 meters before opting to take a bus (Burke & Brown, 2007). We further planned to exclude airport bus routes due to their distinct fare structures and passenger demographics.

For MTR analysis, our initial approach involved calculating average fares per kilometer between stations and aggregating these to district-level comparisons, while incorporating data on car capacity and train frequency as potential explanatory variables.

### **4.4 Methodological Evolution**

A significant methodological refinement for bus analysis occurred during the research process as we encountered limitations in our original approach. Initially, the study employed average fare per kilometer (HKD/km) as the primary metric for comparison. However, rigorous exploratory data analysis revealed that this approach inadequately captured the complexity of Hong Kong's transportation fare structures, particularly for flat-fare routes and cross-harbor services.

This methodological pivot represented a crucial advancement in our research approach. The initial fare/km metric, while intuitively appealing for its simplicity and apparent comparability, proved problematic for several reasons:

1. **Distortion of Flat-Fare Routes**: Many short-distance services in Hong Kong employ flat-fare structures that appear artificially expensive when converted to per-kilometre metrics
2. **Inadequate Representation of Premium Services**: Cross-harbour routes with tunnel tolls and premium express services incorporate costs not directly related to distance
3. **Masking of Starting Fare Components**: Many routes employ graduated fare structures with higher initial per-kilometre costs that decrease over distance
4. **Statistical Limitations**: The fare/km distribution exhibited problematic statistical properties, including heteroscedasticity and non-normality, that complicated modelling efforts

By transitioning to raw fare analysis with distance as an independent variable rather than a denominator, we achieved several methodological improvements:

* More accurate representation of how fares are actually determined
* Better statistical properties for modelling
* Improved ability to identify the independent contribution of distance versus other factors
* More nuanced insights into fare structures across different route types and geographic areas

A similar refinement to the one above was done for the MTR analysis after the weaknesses of using fare/km were found in analysing the bus data. In addition, the idea of aggregating the data based on districts was abandoned after the initial EDA, as with the numerous available routes, aggregation might not give us the insight we desired and potentially hide important results from us. We also decided not to include car capacity and train frequency in our modelling/analysis because we could not find sufficient data regarding these factors.

### **4.5 Definition of "Unreasonably Expensive"**

To establish objective criteria for identifying "unreasonably expensive" transportation, we employed a statistical outlier approach defining unreasonable expenses as those exceeding 1.5 times the interquartile range (IQR) above the upper quartile (Q3 + 1.5\*IQR). This approach was selected for its statistical robustness, interpretability, and alignment with established outlier detection methodologies in data science literature.

The determination of what constitutes "unreasonably expensive" transportation presents both technical and philosophical challenges. From a technical perspective, we needed a method that would:

1. Identify statistical anomalies rather than subjective judgments
2. Account for the natural variability in transportation costs
3. Be applicable across different transportation modes and districts
4. Maintain statistical validity despite non-normal distributions

From a philosophical perspective, we recognized that "unreasonable" costs must be defined relative to expected costs given relevant factors rather than in absolute terms. A higher fare for a longer journey does not necessarily indicate an unreasonable cost structure.

The selected IQR-based approach offers several advantages:

* **Statistical Robustness**: Unlike mean-based methods, IQR is resistant to extreme outliers and does not assume normality
* **Contextual Sensitivity**: By calculating thresholds within relevant subgroups (e.g., within similar route types or districts), we account for legitimate variations
* **Interpretability**: The concept of statistical outliers is well-established and communicable to non-technical audiences
* **Alignment with Practice**: The 1.5×IQR threshold has widespread acceptance in statistical analysis as a benchmark for outlier detection

This approach represents an evolution from our initial proposal, which suggested using both 90% confidence prediction intervals and 75th percentile thresholds to define "unreasonably expensive" transportation. After evaluating these methods during our exploratory analysis, we determined that the IQR-based approach provided superior robustness and interpretability for our specific research context.

This approach enables us to identify not just routes with high absolute fares, but those with fares that are disproportionately high relative to comparable services, providing a more meaningful basis for evaluating potential inequities in transportation costs across Hong Kong's diverse districts.

## **5. Bus**

### **5.1 Data Acquisition**

1. Overview of Main Datasets

The primary dataset for this analysis comprises a comprehensive list of bus stops, routes, bus numbers, fares, and additional relevant information sourced from the "Routes and Fares of Bus" dataset available on Data.GOV.HK. This dataset serves as the foundation for understanding public transportation in Hong Kong.

1. Integration of Route Grouping

To enhance the dataset, we incorporate a route grouping list from Wikipedia, which categorizes routes into specific types. This categorization aids in structuring our analysis, allowing for differentiated insights based on route type.

1. Socioeconomic Data

We also examine socioeconomic data from Data.GOV.HK to assess its potential influence on bus fares. Understanding the relationship between fare structures and socioeconomic factors is essential for determining equitable pricing.

1. Airport Bus Services

To gather information on airport bus services, we scrape data from the Hong Kong International Airport website. This data provides insights into transportation options available to passengers traveling to and from the airport.

1. Coordinate Conversion

We utilize the geodetic data from the Hong Kong Geodetic Survey website to convert coordinates into a standardized format, which is crucial for spatial analysis.

1. District Boundary Data

Additionally, we incorporate the hksar\_18\_district\_boundary.json dataset, which contains geo-referenced data of district boundaries, including names in both Chinese and English along with their corresponding coordinates. This allows us to create a new column in our dataset that identifies the district in which each bus stop is located, adding critical geographic context.

### **5.2 Data Cleaning**

**Routes Data**

1. Data Type Conversion

In the data cleaning phase, we first handle data type conversion. The original MDB format cannot be directly converted to CSV using an online converter, so we employ Python to convert this data into a pandas DataFrame. Initially, all data types are set as strings, necessitating careful conversion to their appropriate types for analysis.

1. String Manipulation

We perform string manipulation on the ROUTE\_NAMEE column, which includes extraneous characters such as "<br". This cleanup ensures consistency across the dataset, with all route names converted to uppercase.

1. Handling NA Values

To maintain data integrity, we drop rows with NA values, particularly in the fare dataset where some entries are marked as zero. After investigating these zero values, we conclude that they often occur when passengers cannot board a bus from the "Shing Mun Tunnel Bus Interchange." Therefore, we remove these entries to avoid skewing our fare analysis.

1. Final Dataset Construction

The final dataset is constructed by merging the stop dataset with the stop coordinates to obtain latitude and longitude for each bus stop. This combined DataFrame is then enriched with fare data to ensure we have fare information for both ON\_SEQ (boarding) and OFF\_SEQ (alighting) stops. We further integrate route details and company information from their respective DataFrames. This comprehensive merging process results in a dataset that includes fare information, geographic coordinates, route names, and company descriptions, enabling a detailed analysis of public transportation routes.

1. Exclusion of Airport Bus Routes

We also exclude airport bus routes from our analysis, as these services primarily cater to tourists rather than daily commuters. This exclusion is significant because the pricing structures and characteristics of airport buses differ markedly from regular commuting routes.

1. Distance Calculation

Once we have standardized the coordinates into longitude and latitude, we calculate distances using the geodesic method. While this method may not match the precision of Google Maps API, it provides a reasonable approximation for our analysis. We utilize the district boundary data to classify each stop within its respective district, enhancing our spatial understanding of the bus network.

1. Exclusion of Stops Based on Distance

In our analysis, we exclude any stops that are within 1 kilometer of the boarding location. This decision is grounded in the assumption that passengers are willing to walk a certain distance—between 600 to 1300 meters, as noted by Burke & Brown (2007)—before opting for bus transport. By filtering out these nearby stops, we focus on meaningful travel distances that better reflect the behavior of passengers who are likely to walk before using public transport.

1. Off-Boarding Points and Travel Behavior

We emphasize the furthest off-boarding points for each boarding location to gain insights into passenger travel behavior and its implications for pricing. This methodology ensures that each entry captures the maximum distance passengers are willing to travel after boarding. By analyzing these distances, we can better understand how they influence fare calculations, ultimately informing equitable pricing structures and guiding transport policy and planning.

1. Circular Routes and Mid-Point Analysis

Additionally, we identify circular routes to analyze how far passengers might travel based on their boarding location. By considering the midpoint and allowing for variations of approximately ±5 stops, we better estimate the furthest stop a passenger might reasonably reach. This approach provides a nuanced understanding of passenger behavior and aids in developing a more effective fare strategy.

1. Route grouping Column

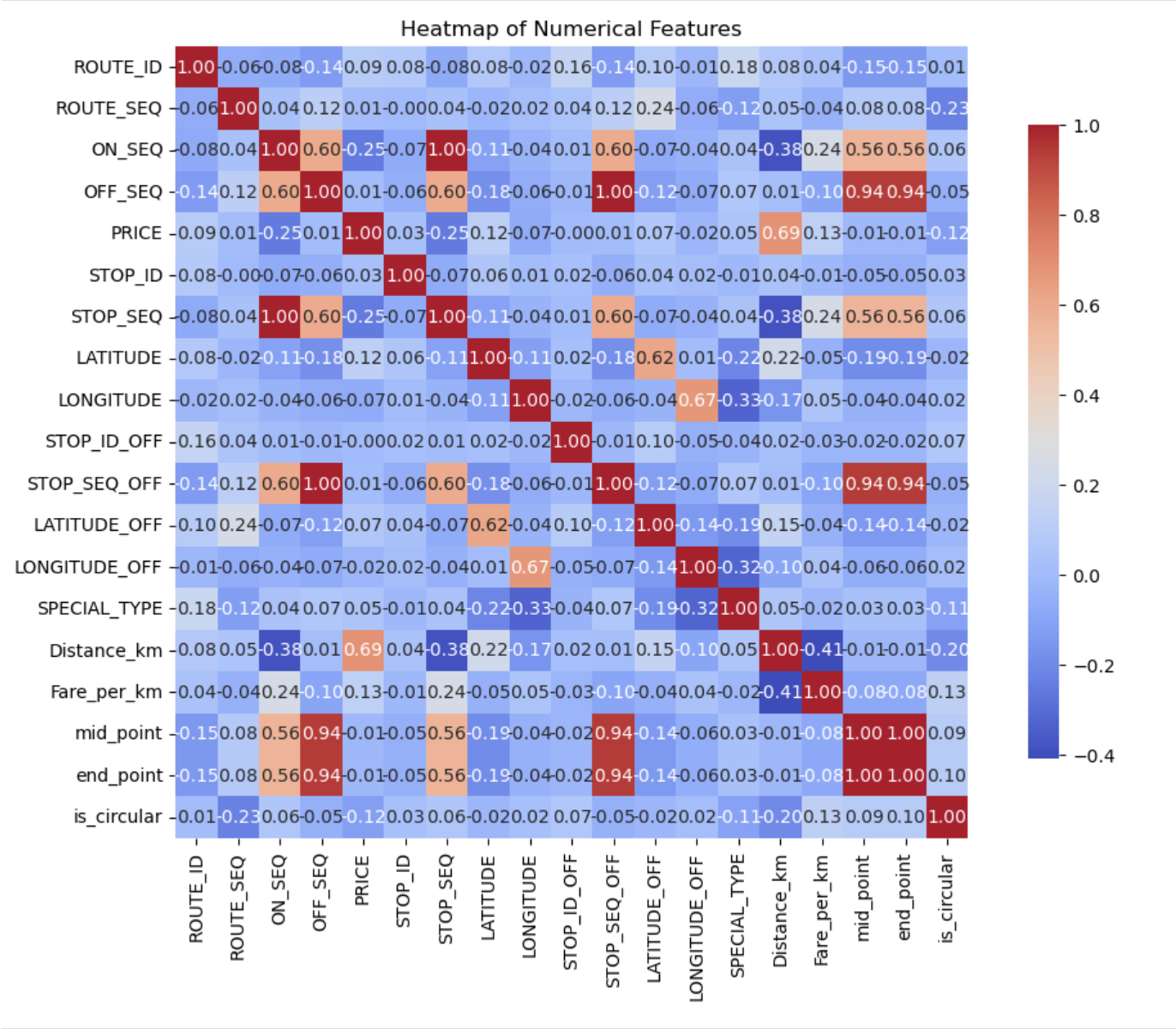
We conducted web scraping from Wikipedia to further categorize the bus routes based on their route types. This categorization is essential, as it allows us to analyze how different route types may influence bus fares. As highlighted in recent news, the fare structure can vary significantly depending on the route type, affecting passenger choices and travel behavior. By integrating this information into our dataset, we enhance our analysis of fare implications and develop a more comprehensive understanding of public transportation dynamics in Hong Kong.

**Socioeconomics data**

To prepare our datasets for analysis, we first created a DataFrame from the economics data. We specified the columns to retain, which included year, district, and monthly median income. After filtering the DataFrame, we separated the data by year, dynamically creating individual DataFrames for each year while setting districts as the index. We specifically chose the year 2022 because it is the only overlapping year of population data and socioeconomics data. We further sorted the corresponding DataFrame by the 'ma\_hh' column in descending order and renamed this column to “income median” for clarity.

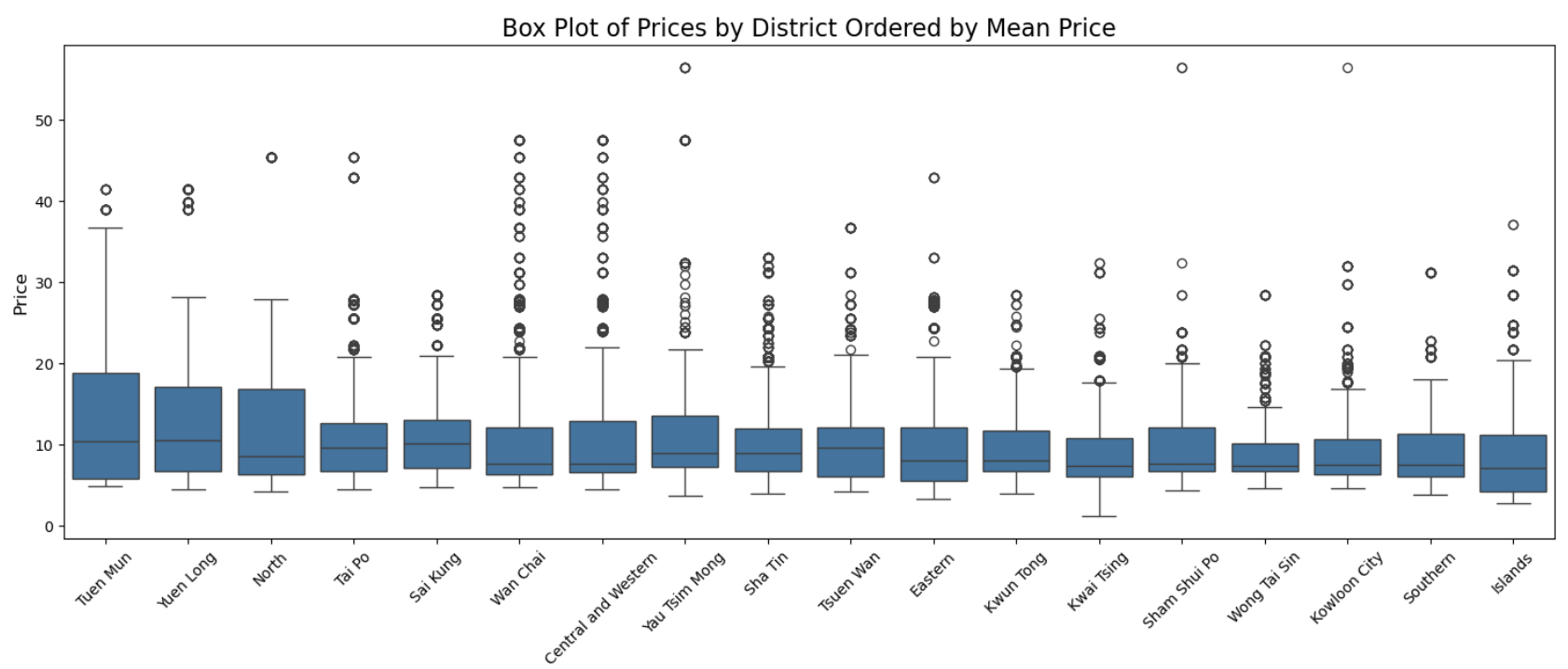
Additionally, we imported a population density dataset from an Excel file, filtering it to keep only the district and population density columns while removing the last row, which was likely unnecessary. We then sorted this by population density in descending order and renamed “Population Density” for better understanding.

**5.3 Exploratory Data Analysis**



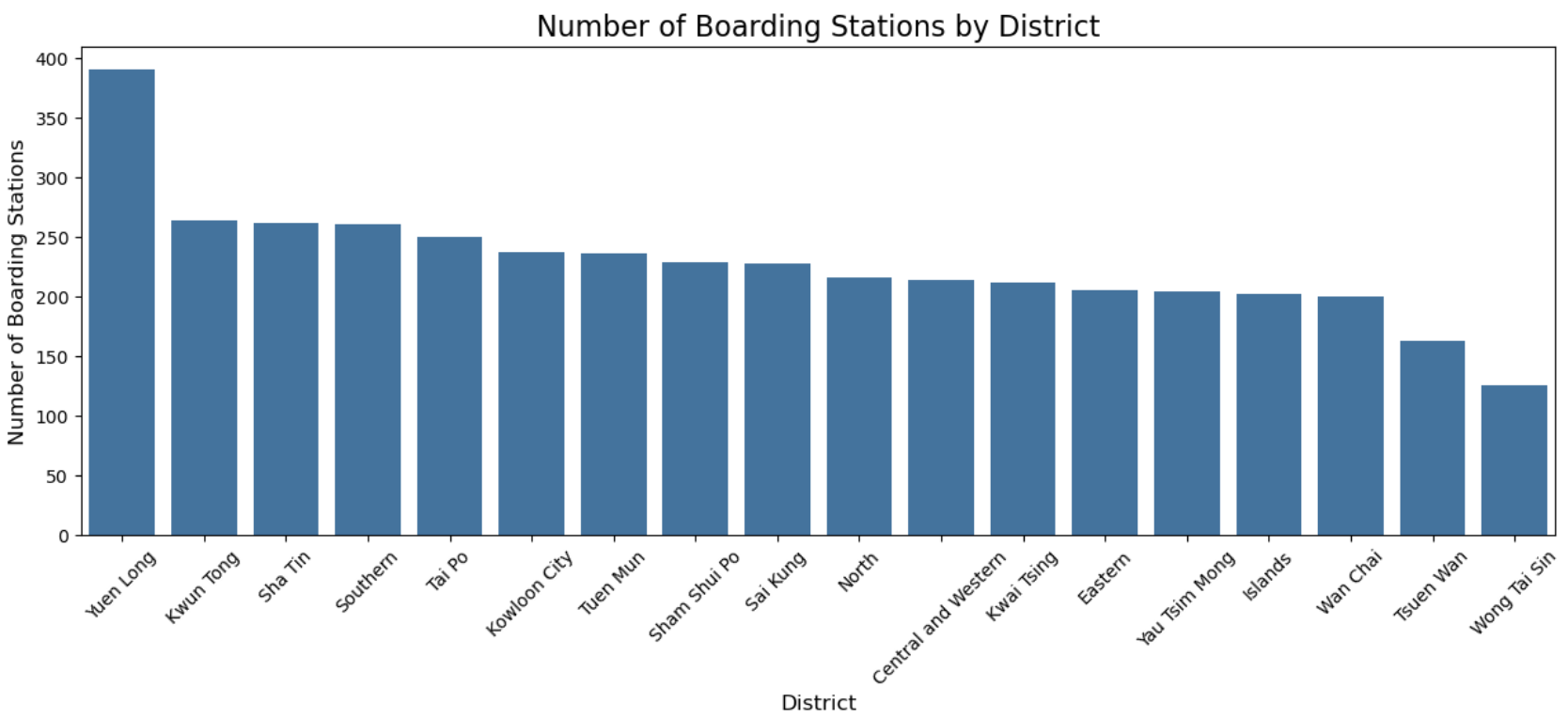
*Figure 1. Heatmap of Numerical Features*

Based on the heatmap, there is a strong positive correlation between transportation prices and distances traveled (Distance\_km). This implies that as the distance of a journey increases, the fare is likely to increase correspondingly.



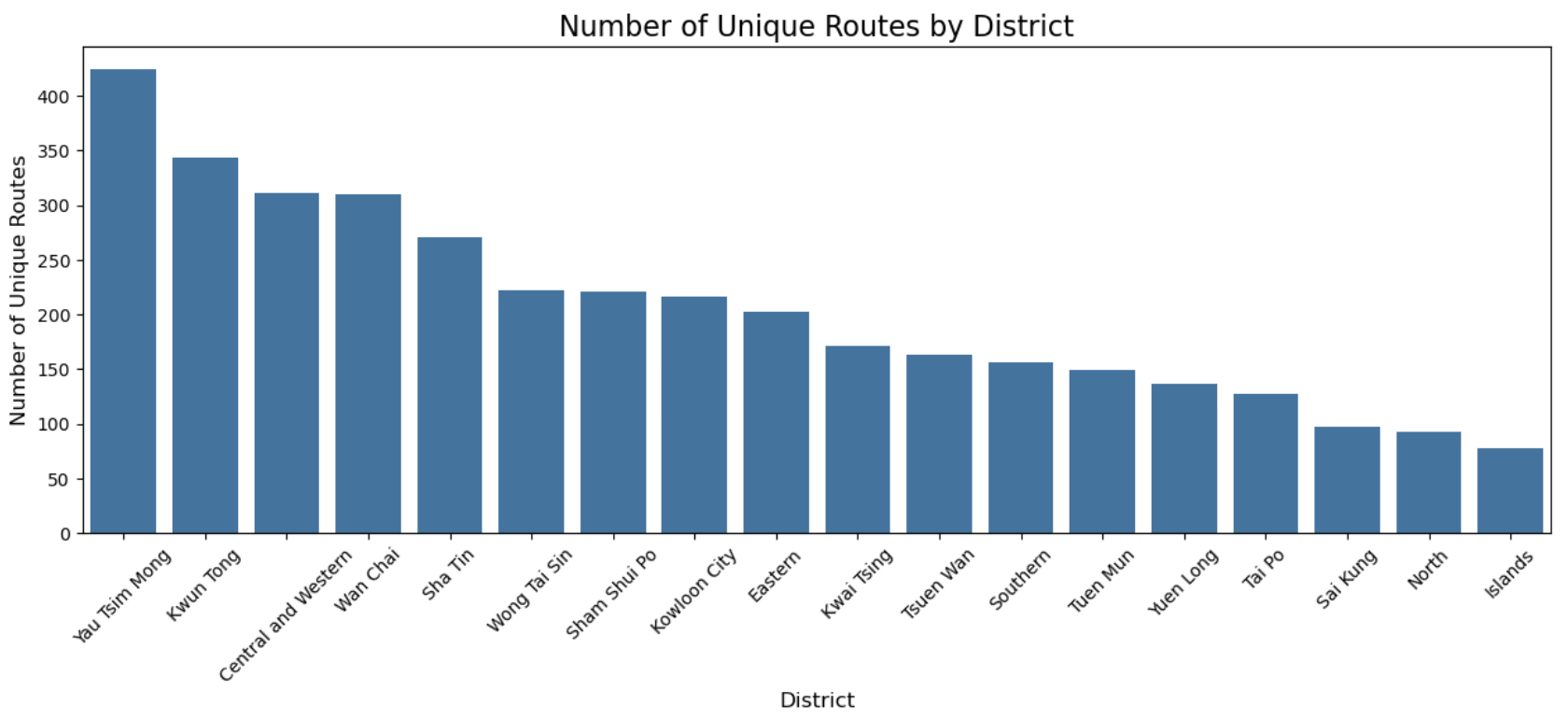
*Figure 2. Box Plot of Prices by District Ordered by Mean Price*

From the boxplot analysis, we observe high prices in **Tuen Mun** and **Yuen Long**. The higher median prices in these districts may correlate with their extensive transit infrastructure, particularly the high number of boarding stations in Yuen Long. A well-developed transit system often comes with higher operational costs, which can be passed on to consumers in the form of elevated fare structures.



*Figure 3. Number of Boarding Stations by Districts*

The presence of numerous boarding stations in Yuen Long indicates a high demand for public transit services, allowing operators to set higher fares, as commuters are often willing to pay for greater access to public transit.



*Figure 4. Number of Unique Routes by Districts*

The districts with the highest number of unique routes, such as **Yuen Tsim Mong**, **Kwun Tong**, and **Central and Western**, serve as major transportation hubs. The high route diversity in these areas indicates robust public transport options, which can attract more commuters.

### **5.4 Model and evaluation**

**Previous Methodology**

The previous approach involved two primary analyses: ANOVA and OLS Regression, both examining the relationship between average fares per kilometer and various factors. The key findings are summarized below:

| **Test Type** | **Hypothesis** | **Key Statistics** | **Conclusion** |
| --- | --- | --- | --- |
| **ANOVA** | Significant differences in average fares per kilometer across districts. | F-statistic: 9.353, p-value: 1.42e-24 | Reject the null hypothesis; significant differences exist across districts. |
| **OLS Regression** | Higher station density relates to higher average fares. | R-squared: 0.003, F-statistic: 107.4, p-value: 3.88e-25 | Significant but weak relationship; station density explains very little of the variance in fares. |
| **OLS Regression** | Higher unique routes relate to higher average fares. | R-squared: 0.005, F-statistic: 181.8, p-value: 2.44e-41 | Significant but weak relationship; unique routes explain very little of the variance in fares. |
| **OLS Regression** | Higher population density relates to higher average fares. | R-squared: 0.001, F-statistic: 26.64, p-value: 2.46e-07 | Significant but weak relationship; population density explains very little of the variance in fares. |
| **OLS Regression** | Higher income median relates to higher average fares. | R-squared: 0.005, F-statistic: 186.0, p-value: 2.91e-42 | Significant but weak relationship; income median explains very little of the variance in fares. |

**Limitations of Previous Methodology**

1. **Narrow Focus on Average Fares per Kilometer**
   * The reliance on average fares per kilometer as the target variable may overlook significant factors influencing pricing, such as service quality, demand fluctuations, and operational costs.
   * Average fares per kilometer might fail to capture regional differences in pricing strategies or consumer behavior.
2. **Low R-squared Values**
   * The regression models reported very low **R-squared values**, indicating that the selected independent variables (e.g., station density, unique routes, population density, income median) explain only a small portion of the variability in fares.
   * This suggests that additional factors or more complex models are necessary to better account for fare variations.
3. **Oversimplified Assumptions**
   * The analysis assumed linear relationships between district characteristics and average fares, which may not fully capture the dynamics of fare pricing.
   * Many hypotheses regarding the influence of district characteristics on fare pricing were weakly supported, highlighting the limitations of the simplistic modeling approach.

**Revised Approach**

Given the identified limitations, we propose transitioning our analysis from average fares per kilometer to raw fare. This change is expected to enhance the comprehensiveness of our analysis and better capture the complexities of fare pricing.

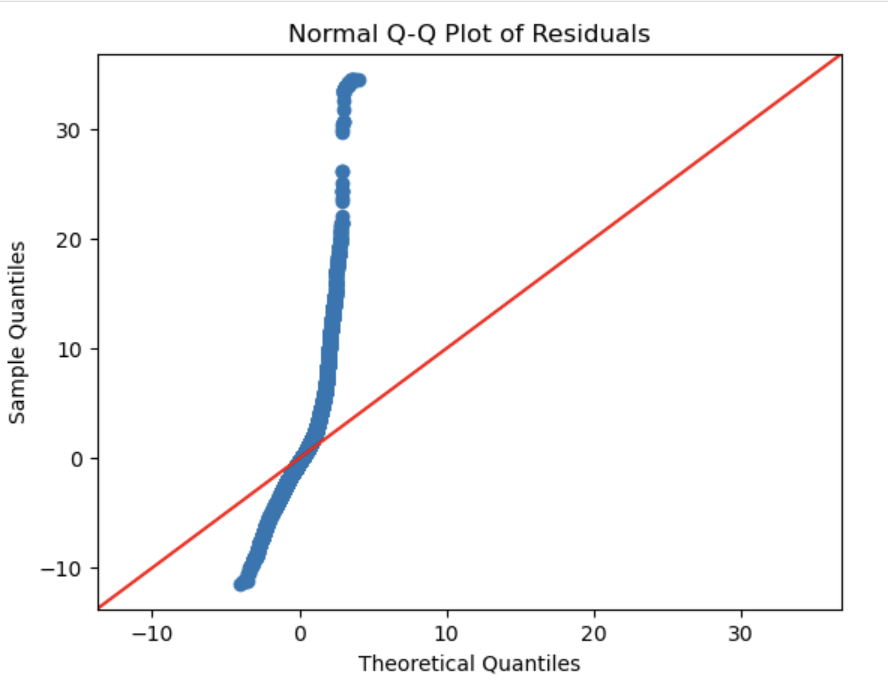
**Advantages of Using Raw Fare**

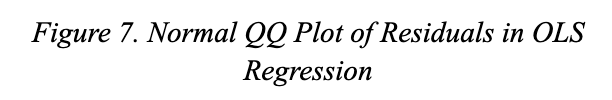
* Holistic Measurement: Raw fare better reflects consumer payments and incorporates all influencing factors, providing a more accurate representation of pricing dynamics.
* Greater Variability: Shifting to raw fare may reveal significant differences in pricing strategies that were previously overlooked in the average fare analysis.

All models, including OLS Regression, Random Forest, and XGBoost, will utilize raw fare as the target variable. This unified approach enhances the comparability of results across different methodologies.

#### **5.4.1 OLS Regression**

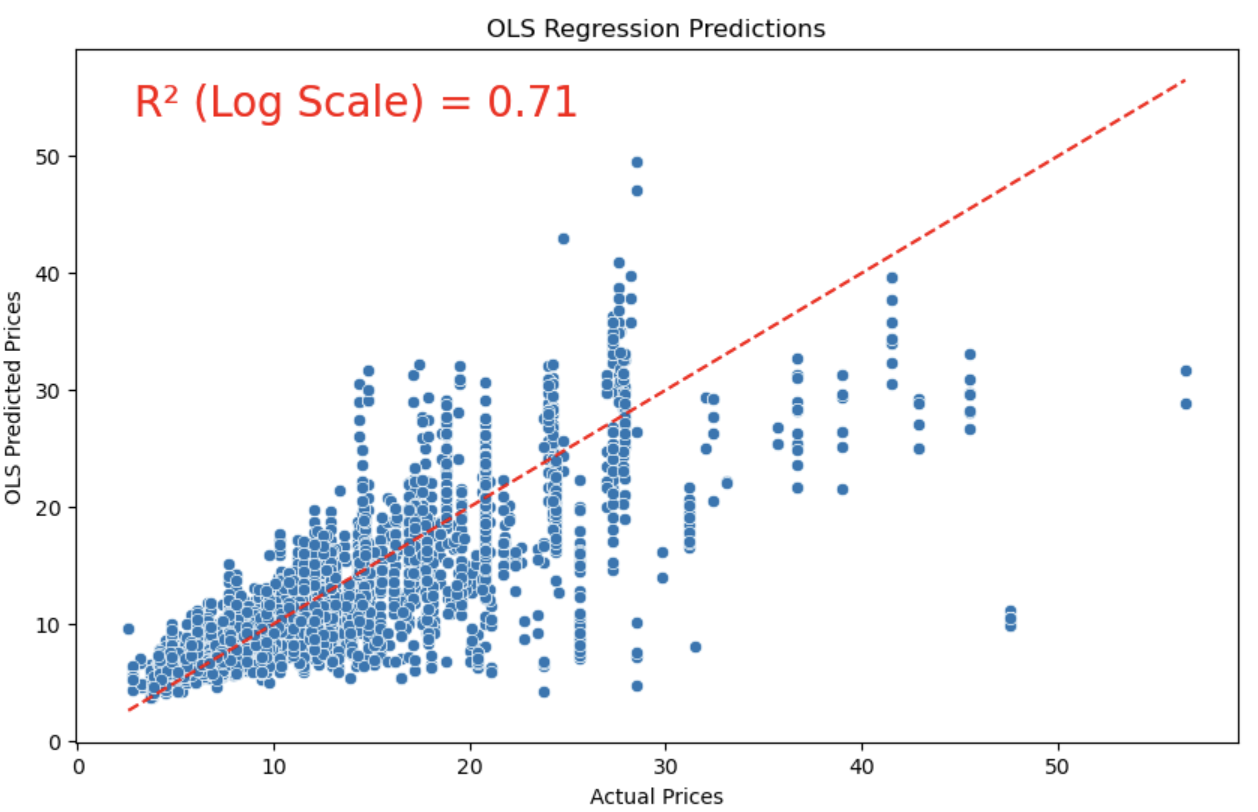
#### 

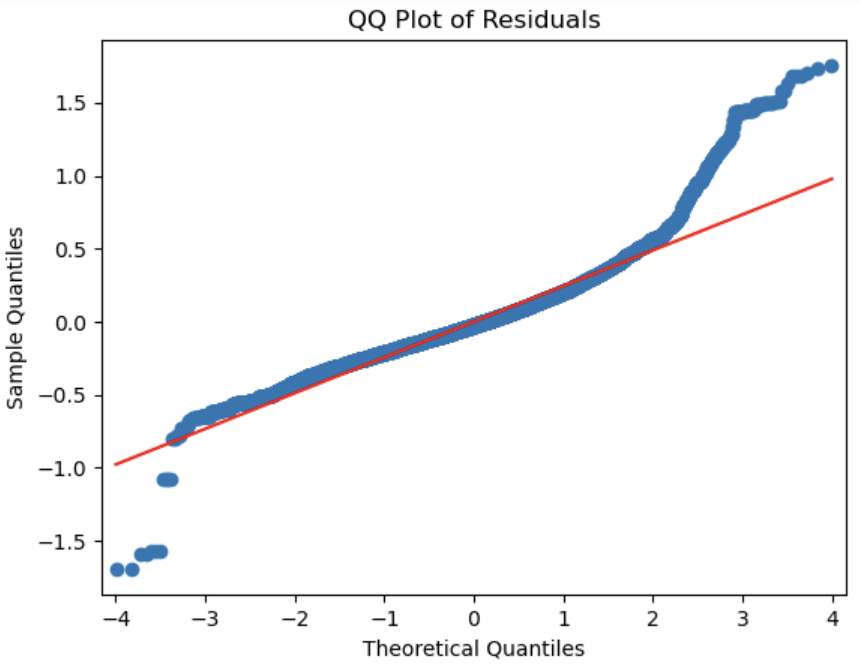
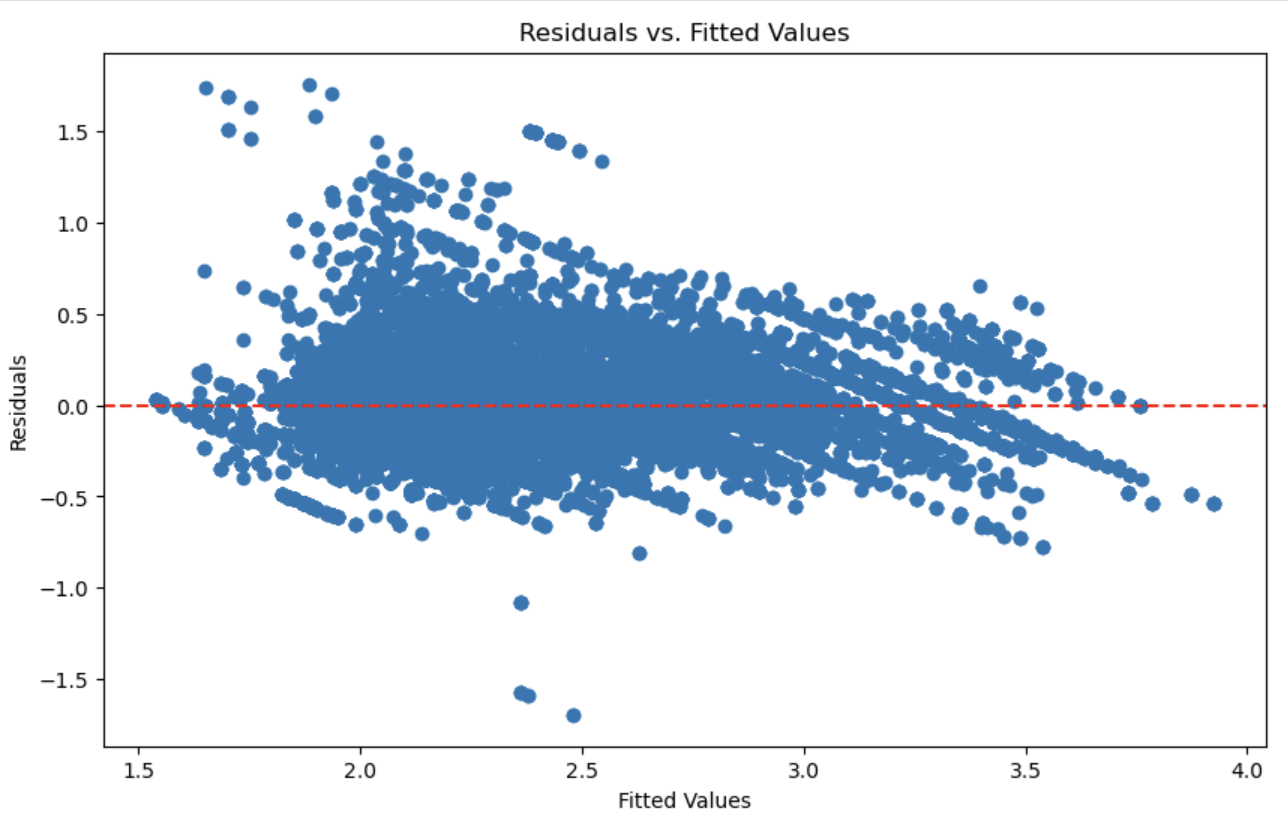
*Figure 5. OLS Regression Predictions*



*Figure 6. Residuals vs. Fitted Values in OLS Regression*

Before conducting the OLS Regression, we performed a Variance Inflation Factor (VIF) analysis to identify and drop features that were highly correlated, helping to mitigate multicollinearity issues. Our initial OLS Regression model achieved an R² of 0.69, indicating it explains a modest percentage of the variation in bus fares. However, we encountered several issues: the residuals fan out with increasing fitted values, indicating heteroscedasticity, and the Q-Q plot showed significant deviations from normality, suggesting that our residuals are not normally distributed and may contain outliers.



*Figure 8. OLS Regression Predictions (after log transformation)*

*Figure 9. Residuals vs. Fitted Values in OLS Regression (after log transformation)*

*Figure 10. QQ Plot of Residuals in OLS Regression (after log transformation)*

To address these issues, we applied a log transformation. Post-transformation, we achieved an R² of 0.71, indicating that 71% of the variance in bus fares is now explained by our model. The cross-validated mean absolute error stands at 0.18, demonstrating a solid performance in predicting actual prices.

While log transformation stabilizes variance, it doesn’t fully resolve all underlying issues. Persistent outliers and heteroscedasticity still pose challenges, as reflected in our residual plots. Consequently, the OLS Regression method may not capture all relationships effectively, necessitating further exploration for a more accurate model.

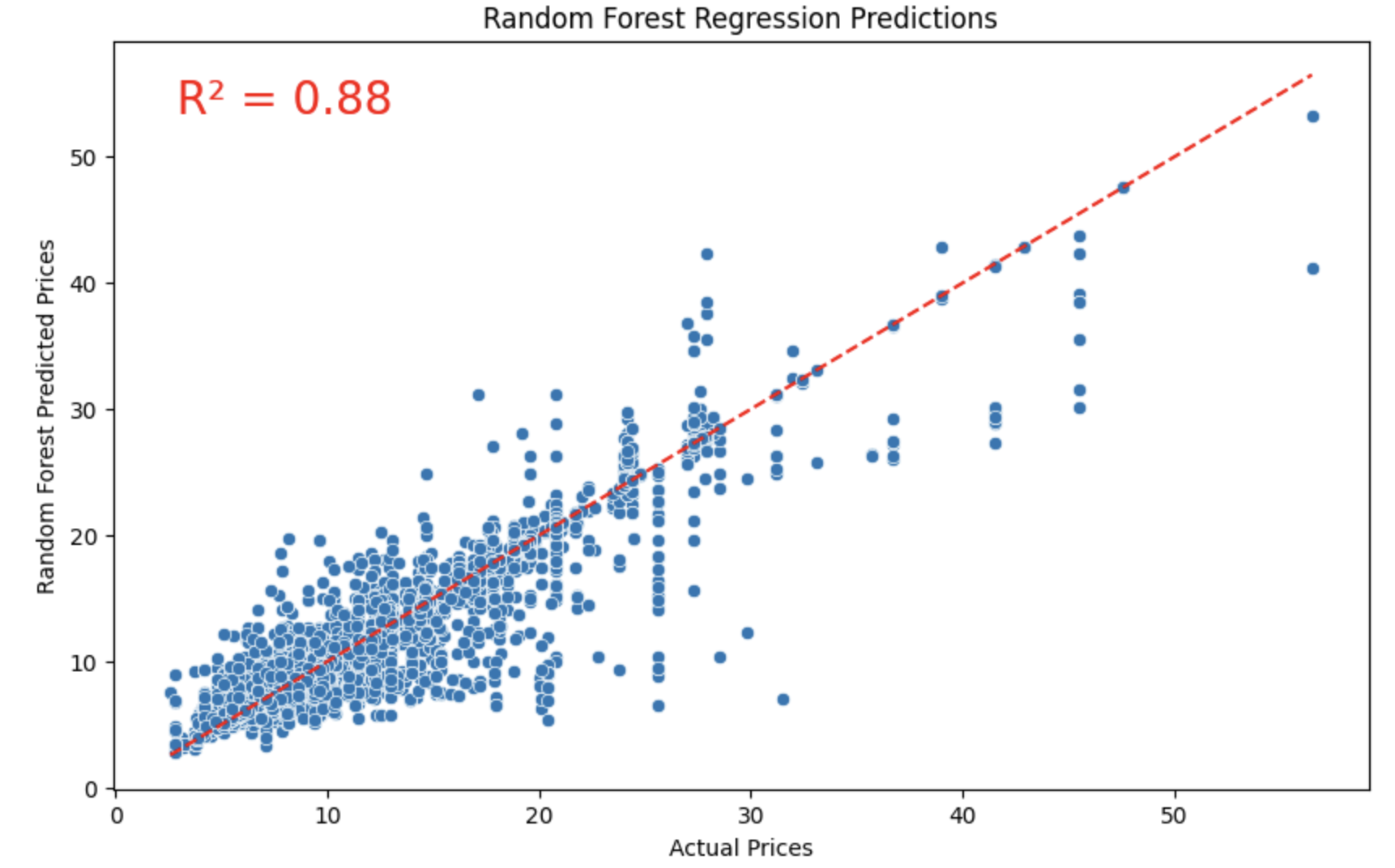
#### **5.4.2 Random Forest**

One of the models we chose to implement is Random Forest due to its robustness in handling complex relationships between features, particularly in our analysis of distance and fare. This ensemble method constructs multiple decision trees and aggregates their predictions, which helps improve accuracy when identifying unreasonably expensive routes.

The graph of Random Forest Regression Prediction illustrates the relationship between actual prices and the predicted prices generated by the Random Forest model. Each point represents an observation, with actual prices plotted on the x-axis and predicted prices on the y-axis. The red dashed line signifies the best-fit line, showcasing the alignment of predictions with actual values. With an R² value of 0.88, the model demonstrates a strong correlation, indicating that it explains 88% of the variance in the actual prices. This high R² suggests that the Random Forest model effectively captures the underlying patterns in the data. The model achieved an RMSE of 2.18, representing an average prediction error of about HK$2.18, and an MAE of 1.22, showing that our predictions typically deviate by around HK$1.22 from actual fares.

The graph of Residuals of Random Forest Model presents a histogram of the residuals, which are the differences between the predicted and actual prices. Ideally, a well-functioning model should yield residuals that are normally distributed around zero. In this case, the distribution shows a peak at zero, indicating that most predictions are close to the actual values, though some skewness suggests the presence of outliers or variability not fully accounted for by the model. The red dashed line represents the theoretical normal distribution, and deviations from this line highlight areas where the model may have limitations in its predictive performance.

The analysis of residuals revealed a peak at zero, suggesting that many predictions are quite close to the actual values. The residuals exhibited a symmetric distribution around the zero line, indicating that the model's predictions are unbiased and that errors are evenly spread. Approximately 88% of predictions fell within ±10 HKD of actual values. The model performed well for average fares (10-30 HKD) but faced some difficulties with extreme prices (very low or very high routes). The primary factors influencing price were identified as distance, the Western Harbour Crossing, COMPANY\_CODE\_KMB, the Cross-Harbour Tunnel, and the Eastern Harbour Crossing.



*Figure 11. Residuals of Random Forest Model*

*Figure 12. Random Forest Regression Predictions*

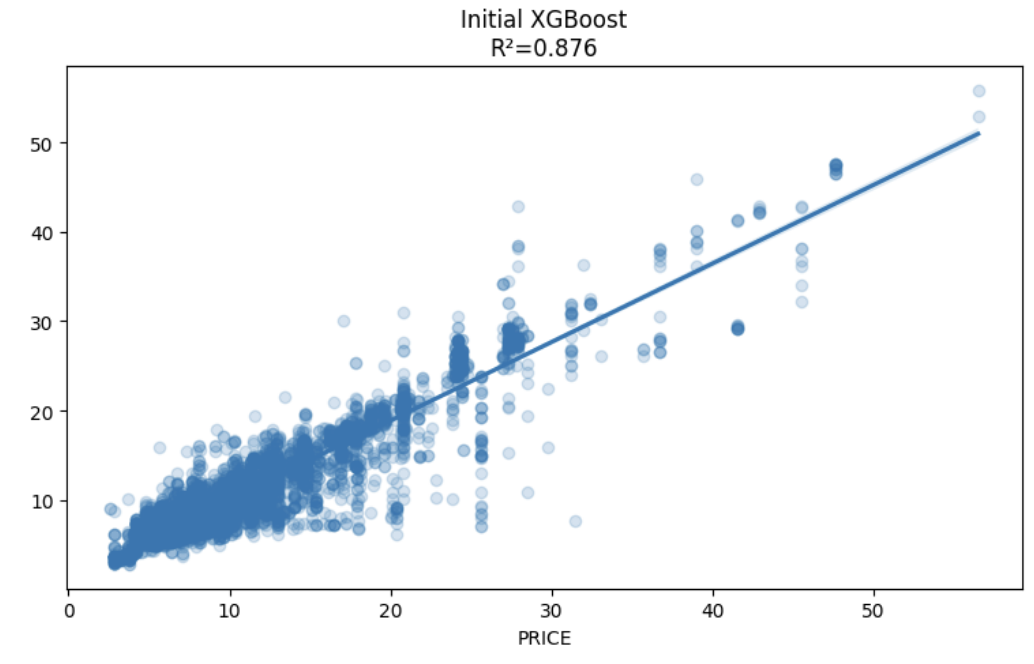
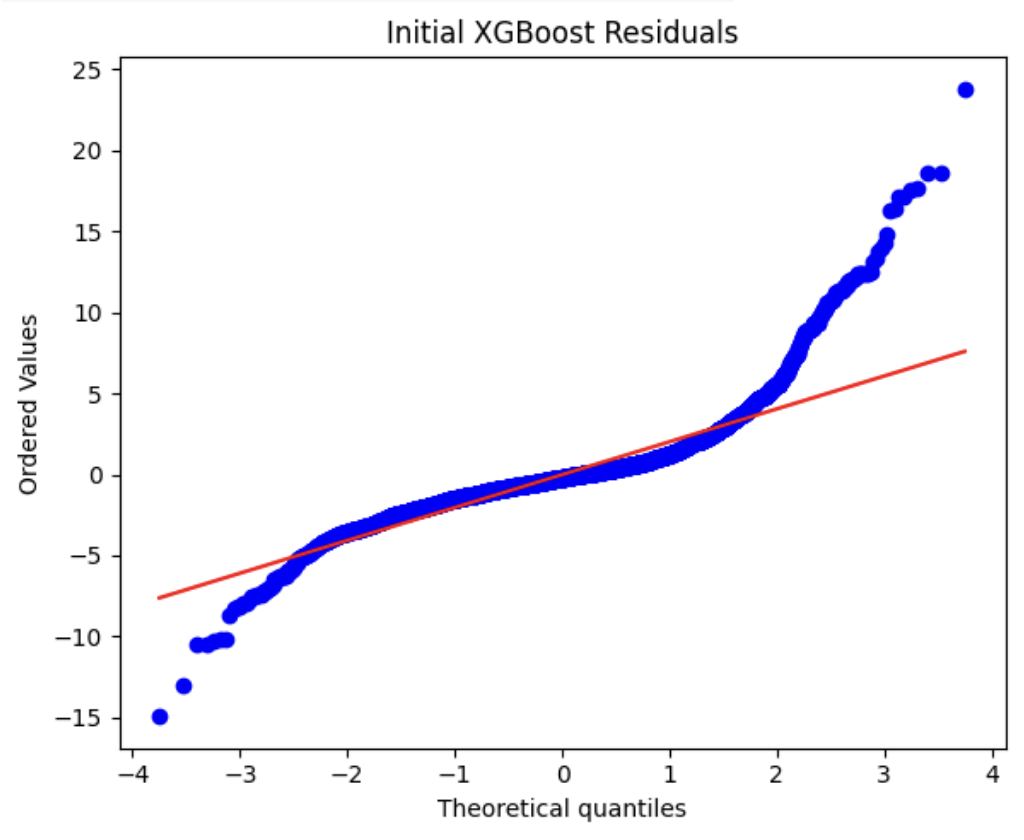
#### 

*Figure 13. Top 20 Features by Importance in Random Forest*

*Figure 14. QQ Plot of Residuals in Random Forest*

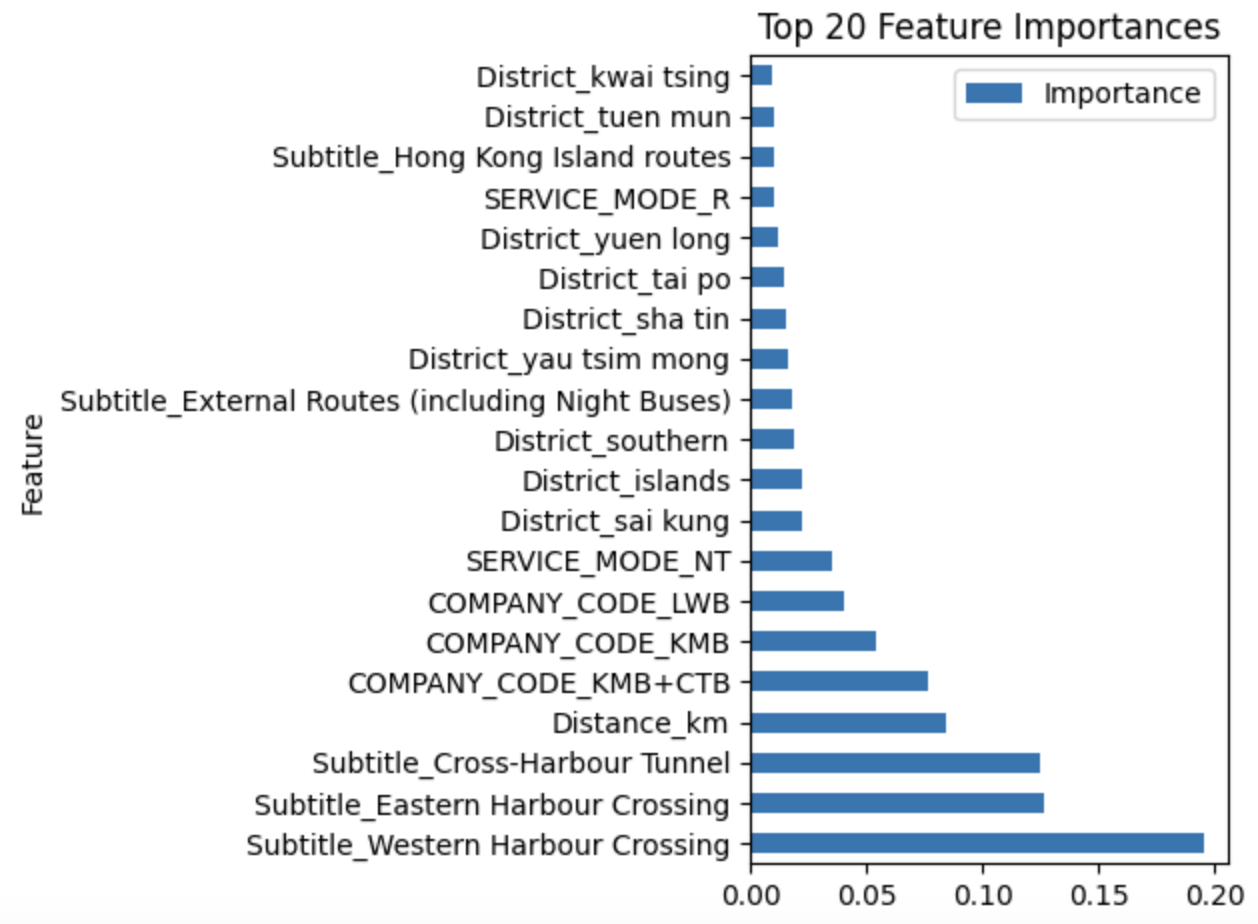
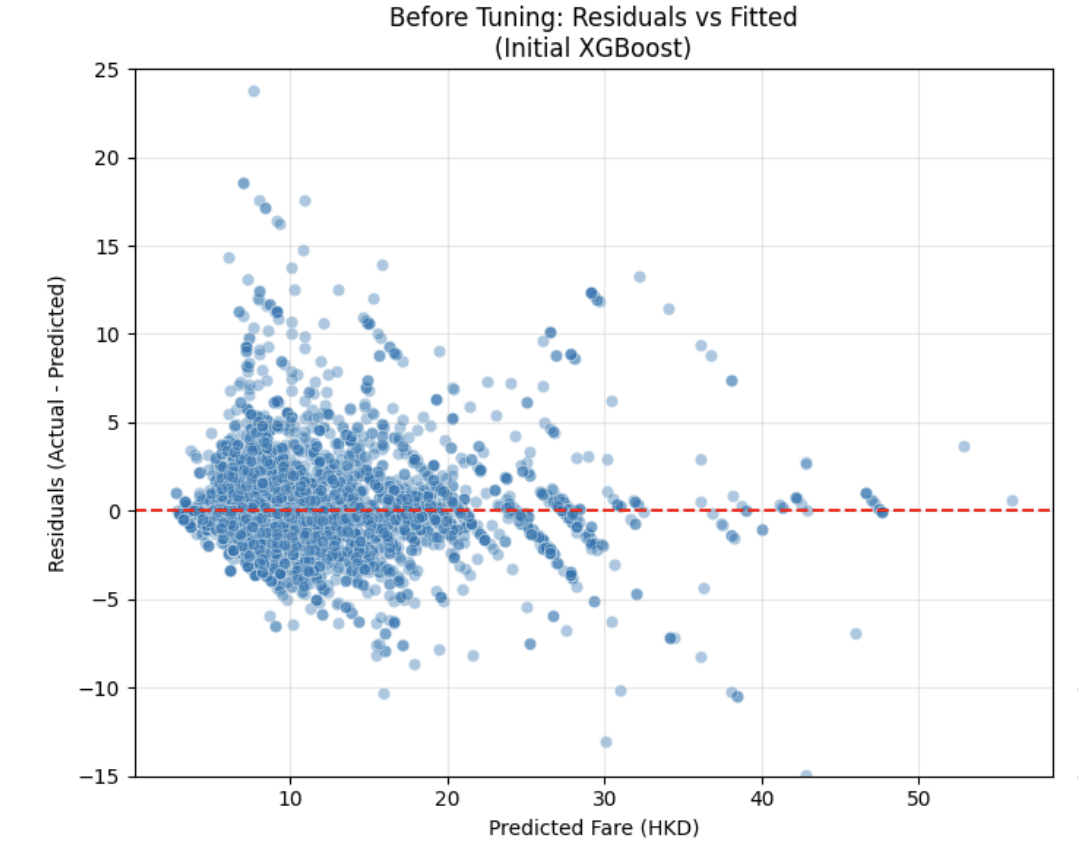
#### **5.4.3 XGBoost**

After exploring initial modelling approaches, we implemented XGBoost for its ability to handle the complex, non-linear relationships between distance and fare revealed in our exploratory analysis. This ensemble method builds sequential decision trees where each tree corrects errors from previous ones, providing the precise predictions needed to identify "unreasonably expensive" routes. XGBoost offers robust feature importance metrics that resist multicollinearity issues present in our geographic and service variables. Additionally, its built-in regularization capabilities were essential for preventing overfitting while maintaining high predictive power, particularly important given our dataset's high dimensionality after one-hot encoding categorical variables.



*Figure 15. XGBoost Regression Prediction (before tuning)*

*Figure 16. XGBoost Residuals (before tuning)*

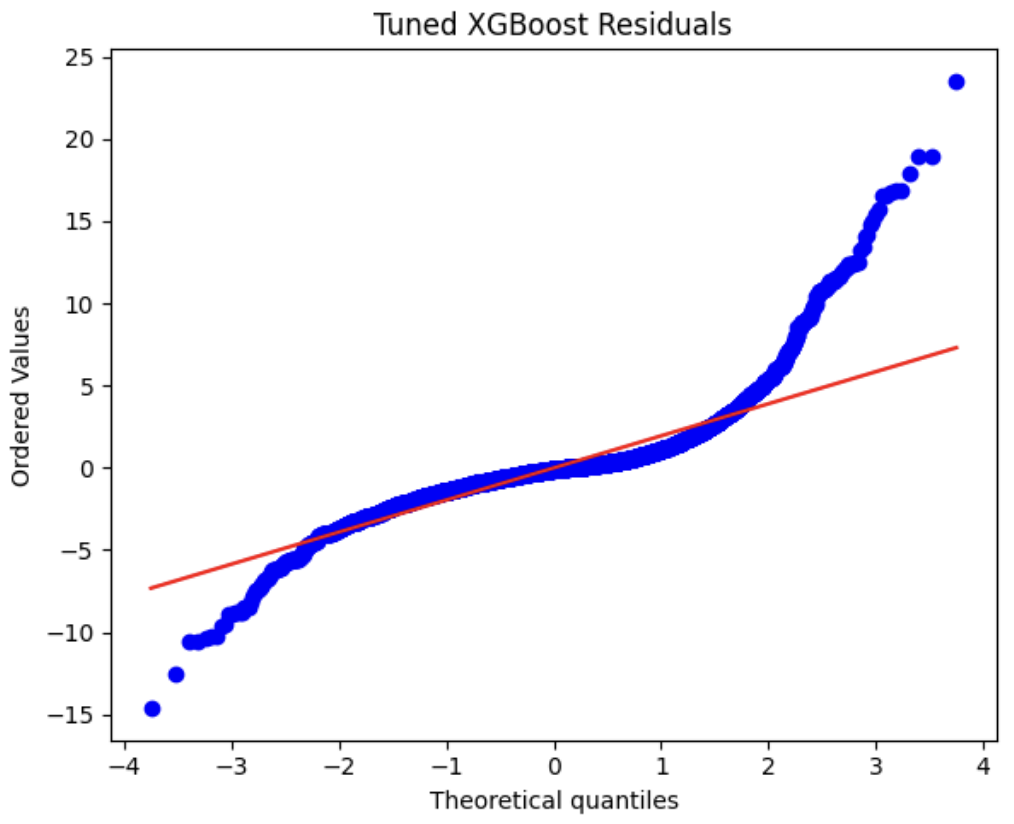
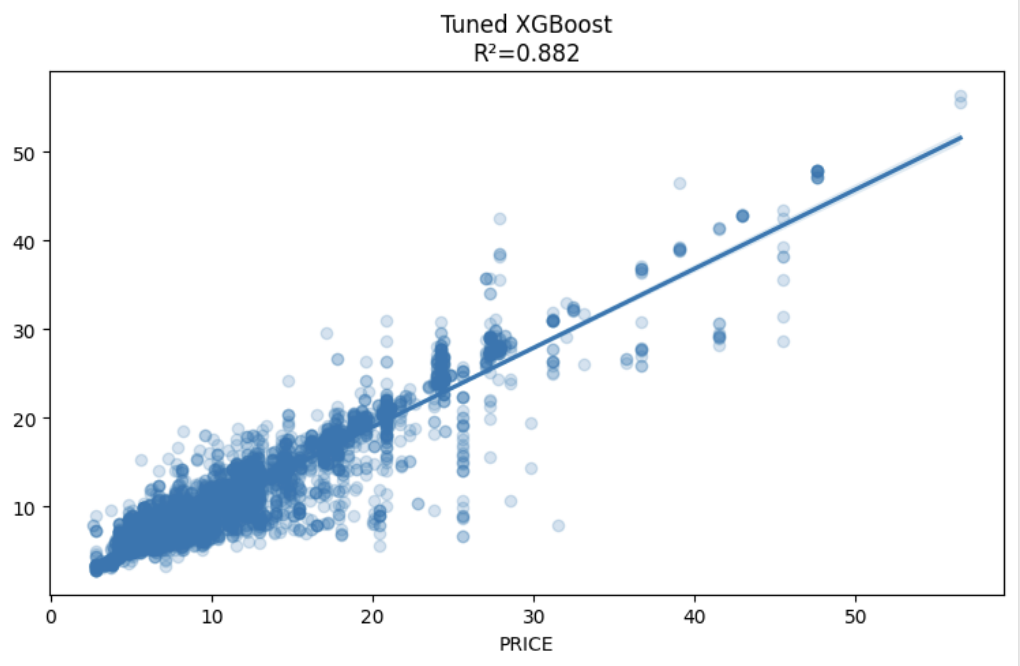


*Figure 17. Residuals vs Fitted Values in XGBoost (before tuning)*

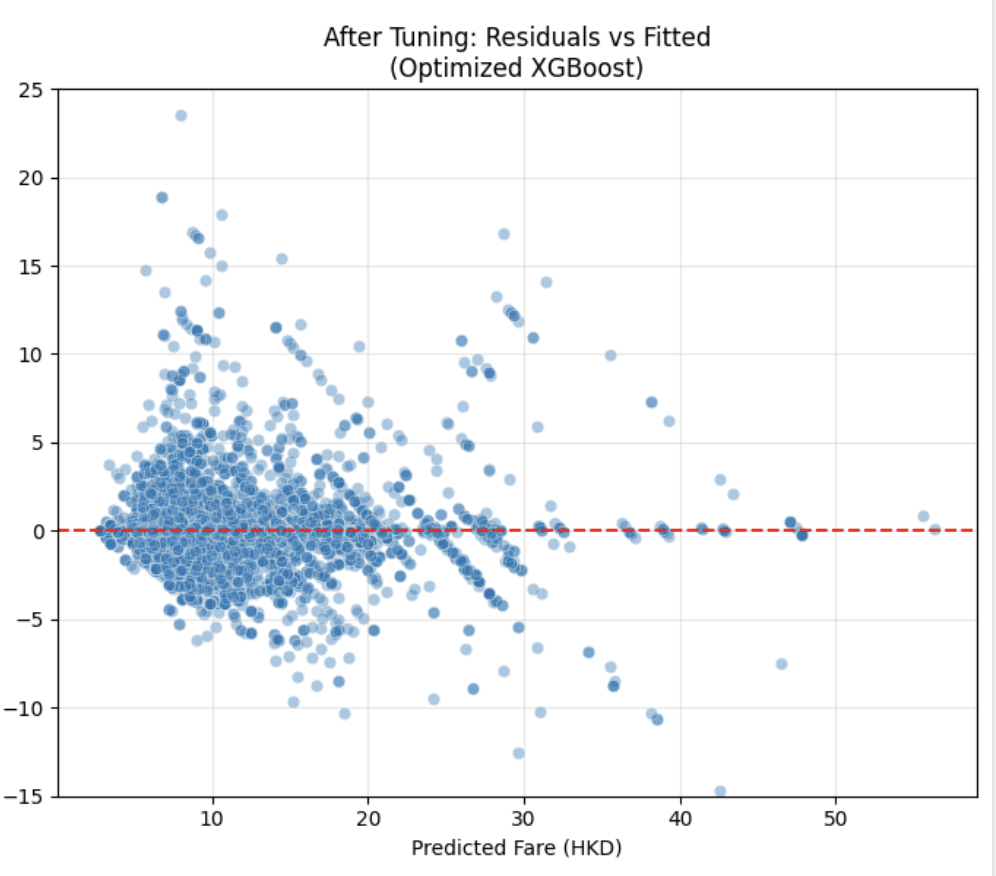
*Figure 18. Top 20 Feature Importances in XGBoos*

Our initial XGBoost model showed promising results with an R² of 0.876, indicating it explains approximately 88% of the variance in bus fares across Hong Kong. The model achieved an RMSE of 2.23, representing an average prediction error of about HK$2.23, and an MAE of 1.38, showing that on average, our predictions deviate by about HK$1.38 from actual fares.

The residual analysis showed random scatter with no systematic bias and consistent variance across fare ranges (no "fanning"). About 90% of predictions fell within ±10 HKD of actual values. The model worked well for average fares (10-30 HKD) but struggled with extreme pricing (very cheap or very expensive routes). The top price driving factors were identified as harbour crossings, distance, and company & service modes.



*Figure 19. XGBoost Regression Prediction (after tuning)*



*Figure 20. QQ Plot of Residuals in XGBoost (after tuning)*

*Figure 21. Residuals vs Fitted Values in XGBoost (after tuning)*

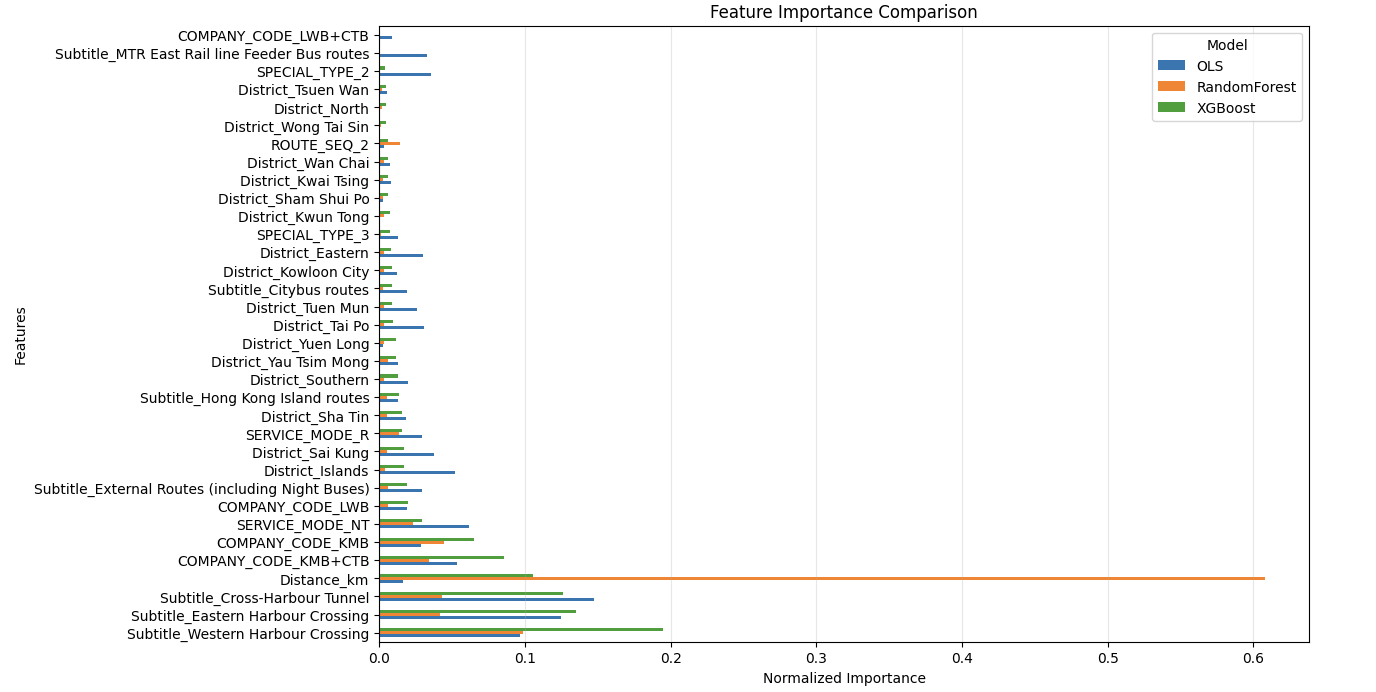
After hyperparameter optimization, our fine-tuned XGBoost model achieved an improved R² of 0.882, indicating it explains a higher percentage of the variance in bus fares. The RMSE improved to 2.17, and the MAE decreased to 1.28, showing better predictive accuracy with predictions deviating by about HK$1.28 from actual fares on average.

The residual analysis confirmed unbiased predictions with a more compact distribution of residuals, especially in the 5-25 HKD range, with 95% of predictions falling within ±10 HKD of actual values. Our analysis showed that tunnel routes average higher fares than non-tunnel routes and exhibit greater variability. Non-tunnel fares cluster tightly around 10 HKD, while tunnel fares cluster around 17 HKD. The model maintained strong performance for average fares but still showed some deviation in predicting extremely cheap or expensive fares.

#### **5.4.4 Feature Importance across all models**

We used multiple models to leverage their unique strengths. OLS Regression provides clear, interpretable results, showing how each factor impacts fares. Random Forest helps us identify complex interactions between features, while XGBoost excels in predictive accuracy, capturing intricate patterns.

By comparing feature importance across XGBoost, Random Forest, and OLS Regression, we validate our findings and ensure a comprehensive understanding of the factors driving bus fares.



*Figure 22. Feature Importance Comparison*

The following table ranks the hypotheses regarding how bus fares vary significantly by different factors:

| **Rank** | **Hypothesis (Bus fares vary significantly by...)** | **Supported or Not** | **Rationale** |
| --- | --- | --- | --- |
| 1 | Journey Distance | Partially Supported | While distance is included in the top factors of all models, it does not show high importance in OLS Regression. |
| 2 | Route Group | Supported | Significant positive coefficients for **Subtitle\_Cross-Harbour Tunnel**, **Subtitle\_Eastern Harbour Crossing**, and **Subtitle\_Western Harbour Crossing** in OLS Regression, supported by Random Forest and XGBoost. |
| 3 | Company Code | Supported | **COMPANY\_CODE\_KMB+CTB** and **COMPANY\_CODE\_KMB** show moderate importance across all models. |
| 4 | Service Modes | Partially Supported | **SERVICE\_MODE\_NT** has moderate importance in OLS Regression and XGBoost but lesser importance in Random Forest. |
| 5 | Geographic District Origin | Partially Supported | **District\_Islands** has a significant negative coefficient in OLS Regression, but minimal importance in Random Forest and XGBoost. |
| 6 | Service Type | Partially Supported | **SPECIAL\_TYPE\_2** shows a positive effect in OLS Regression and minor importance in other models. |

**Key Findings**

1. **Journey Distance**: Although included as a factor, journey distance was not a strong predictor in the OLS Regression model, indicating that other factors may have a more significant impact on fare pricing.
2. **Route Group**: Specific routes, particularly those involving major crossings, showed significant positive coefficients, affirming their importance in fare determination.
3. **Company Code**: The analysis revealed that different bus companies have varying impacts on fare pricing, with certain codes showing moderate importance across all models.
4. **Service Modes**: While service modes contributed to fare variability, their importance varied by model, indicating a need for further exploration.
5. **Geographic District Origin**: The district from which a passenger board can influence fares, though its impact was less pronounced in some models.
6. **Service Type**: Certain service types had a positive effect on fares, but their importance was limited compared to other factors.

### **5.5 Interpretations and Communications**

#### **5.5.1 Key Insights from Bus Data**

From section 5.4.4 (Feature Importance across all models), we identified two major factors affecting bus fares. Firstly, the Specific Route Groups with higher fares, particularly those involving major crossings, exhibited significant positive coefficients, underscoring their importance in fare determination. The notable route groups include “Subtitle\_Cross-Harbour Tunnel,” “Subtitle\_Eastern Harbour Crossing,” and “Subtitle\_Western Harbour Crossing.”

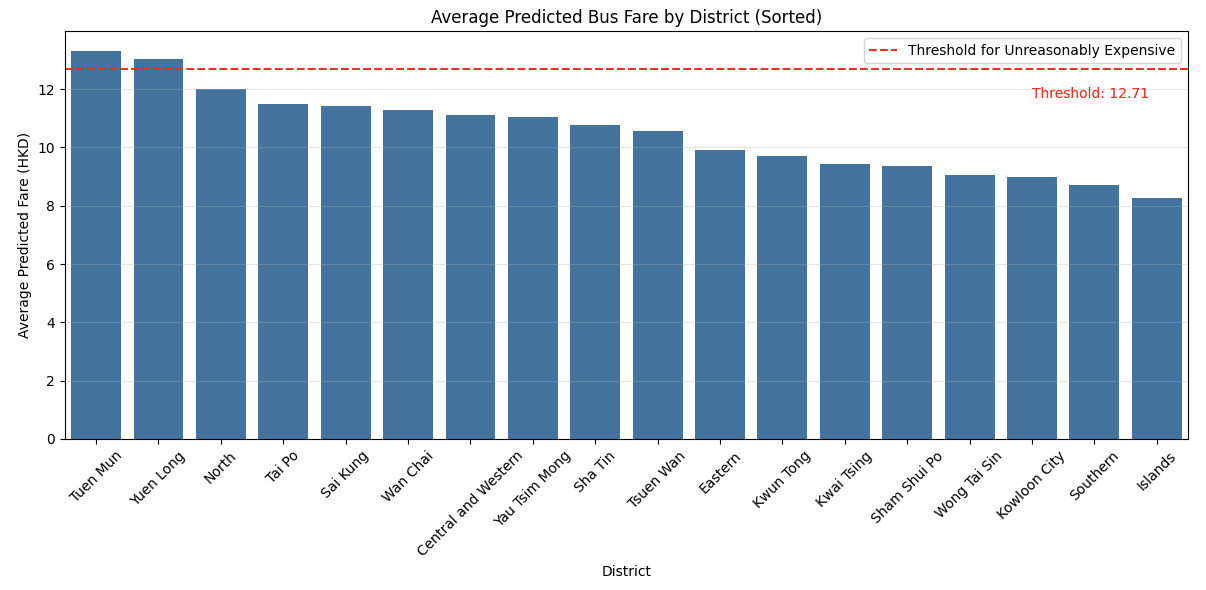
In addition, Journey Distance emerged as another key factor influencing fares. Although the OLS Regression did not support this variable as significant, both the random forest and XGBoost models indicated that Journey Distance possesses high feature importance. This discrepancy highlights the potential limitations of OLS Regression in capturing complex relationships in the data.

Overall, these insights suggest that fare structures may need to be adjusted to reflect the significance of specific route groups and journey distances, ensuring a more equitable pricing strategy for passengers.

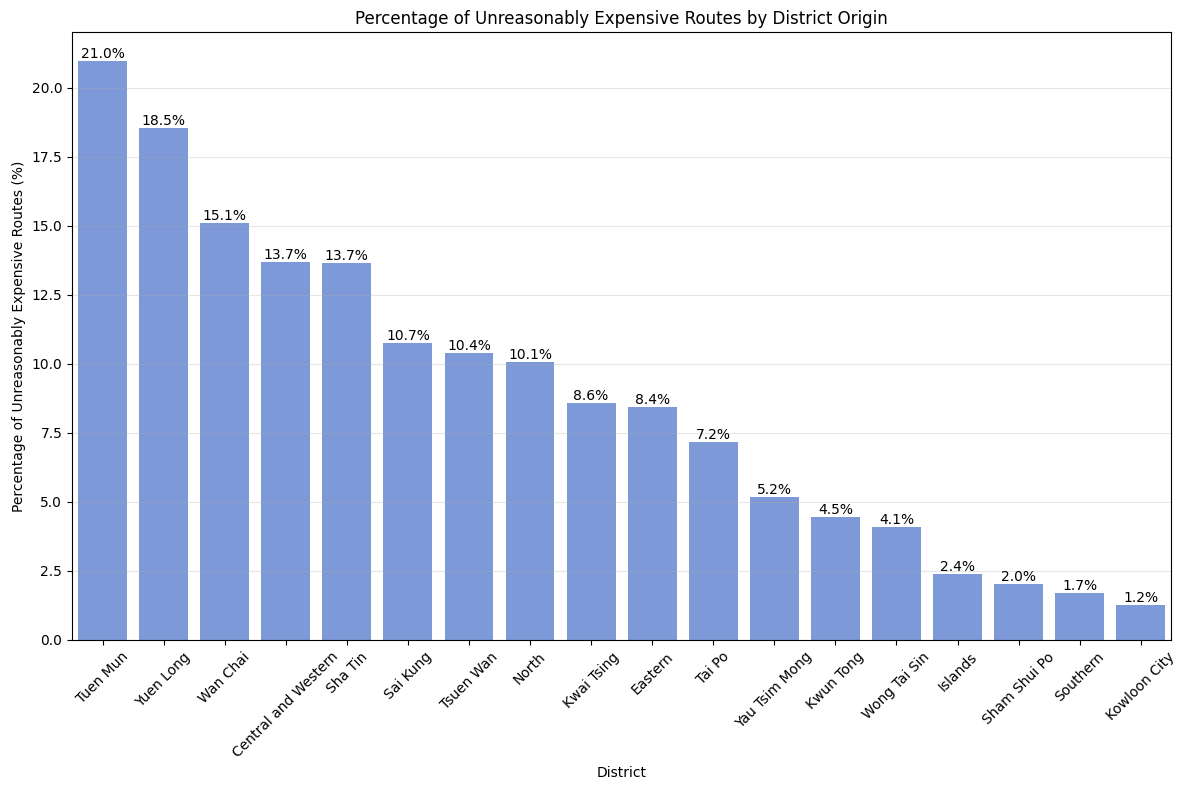
#### **5.5.2 Overall Conclusions for Bus Fares**

The final model we selected is the tuned XGBoost, which yielded an RMSE of 2.17 and an R² value of 0.88. To identify routes that may be considered “unreasonably expensive,” we established a threshold of 12.71 HKD, which is set at 1.5 times the interquartile range (IQR) above the upper quartile. Our analysis revealed that only the Tuen Mun and Yuen Long districts exceeded this threshold, indicating a higher prevalence of potentially unreasonably expensive routes in these areas.

Next, we examined the percentage of unreasonably expensive routes across different districts, further validating our assumption. Both Tuen Mun and Yuen Long emerged as the districts with the highest percentages of unreasonably expensive routes, accounting for 21.0% and 18.5%, respectively. This analysis highlights the need for focused attention on these areas to address fare disparities.



*Figure 23. Average Predicted Bus Fare by District*



*Figure 24. Percentage of Unreasonably Expensive Routes by District Origin*

#### **5.5.3 Reflection on Methodological Change**

As mentioned in Section 4 (Methodology), we shifted our fare calculation from using the average fare per kilometer to the original flat bus fare price after experimenting with various models. While using the average fare per kilometer to identify relationships with other factors affecting bus fare, we did not achieve satisfactory results. Initially, we suspected that our model integration methods were flawed; however, we realized that no matter how we integrated the models, the performance remained below our expectations. This led us to consider the possibility of an incorrect approach from the outset.

Given that bus fares are primarily established at a flat rate—indicating that the cost is uniform regardless of distance and determined instead by the specific boarding stop—integrating distance into our fare calculations may not constitute the most effective strategy. This revelation was unexpectedly disconcerting for our team, as the entire project was predicated on our prior fare calculations. The realization that our initial approach was ineffective was a significant blow to our confidence. Nevertheless, after engaging in discussions with our professors and tutors, we acknowledged that such occurrences are commonplace in data science projects. It is always feasible to revisit earlier stages and amend our approach.

Following the adjustment of our methodology, we ultimately observed substantial improvements in the accuracy of our fare predictions. This experience serves as a reminder that initial hypotheses and models may not always hold true, and that maintaining an openness to reevaluating our strategies can yield more robust and meaningful outcomes.

#### **5.5.4 Limitations and Future Work for Bus Analysis**

While our analysis provided valuable insights into bus fare determination, several limitations must be acknowledged. Firstly, the dataset may not encompass all variables influencing fare prices, such as external factors like traffic conditions, seasonal variations, or special events that can affect bus operations and demand. This lack of comprehensive data may lead to an incomplete understanding of fare dynamics.

Additionally, although the tuned XGBoost model has consistently yielded the best results among the three models we tested, it’s essential to acknowledge that relying solely on this model may introduce inherent biases. To enhance our analysis, we can further work on comparing the feature importances across all three models for validation.

Finally, our threshold for identifying “unreasonably expensive” routes was based on a statistical approach that may not fully capture the perceptions and experiences of passengers. Stakeholder feedback, including insights from regular commuters, could enhance our understanding of fare fairness and help refine our criteria for what constitutes unreasonable pricing.

In summary, while our analysis has yielded important insights into bus fare patterns, addressing these limitations will be essential for future research and for developing more accurate and equitable fare structures.

## **6. MTR Dataset**

### **6.1 Data Acquisition**

The main dataset used in the analysis of MTR fares came from the Hong Kong government’s database website (DATA.GOV.HK) titled “MTR routes, fares and barrier-free facilities.” We chose the csv file titled “MTR Lines (except Airport Express & Light Rail) Fares” which contains the source station, destination station, their respective IDs, and the various types of fares. The most important values in the subsequent analysis are the station name and the Octopus fares.

During the later part of the project, we utilized another dataset found on Kaggle titled “Hong Kong MTR / Subway Network.” This dataset contains all MTR stations in Hong Kong and the next station in the line, with the distance between them. This dataset was utilized to help solve some issues found during modeling with specific routes, which will be discussed in the respective section. Do note that this dataset was not known at the time of the exploratory data analysis and the early parts of modeling.

### **6.2 Data Cleaning**

We cleaned the government’s MTR dataset by keeping only the starting station name, destination station name, and Octopus fare for an adult. Using the names, we scraped Wikipedia to find their respective latitude and longitude. From the latitude and longitude we were able to find the respective districts of the stations. We were also able to use the coordinates and the Geodesic library to find the distance between stations, which the library calculated as finding the direct straight distance between two points (although the weakness to this approach was found later during modeling).

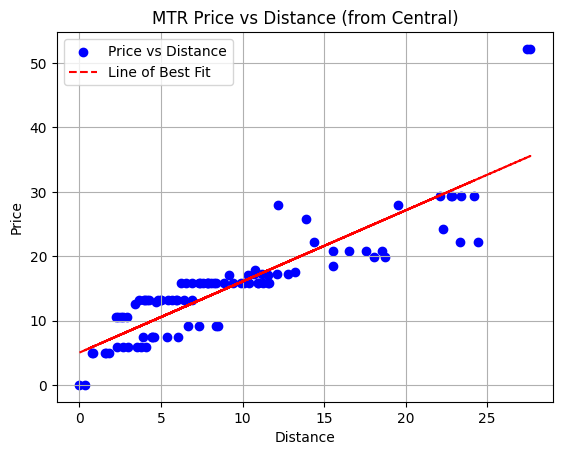
The district labeling was necessary to find insights into the district as a whole and also be able to utilize socioeconomic data. The calculation of the distance between stations was necessary as we presumed it to be an important factor to determine the fares. The distance calculation also helped us to count the fare/km of routes which we also presumed to be important, although later this was abandoned.

The Kaggle dataset was cleaned by transforming the dataset to show all routes possible from every station and their respective distances. As the original dataset contained the distance from one station to the next one, it was similar to nodes and edges. Thus using graph libraries, we were able to find the shortest distance between all possible stations. This was necessary to overcome a limitation in the original distance calculation which will be discussed during the modeling section.

### **6.3 Exploratory Data Analysis**

The goal of our exploratory data analysis (EDA) was to find the relationship between fares and distance and the average MTR fare between districts. Do note that this EDA was conducted while the distance calculation was based on a straight line distance between two stations.

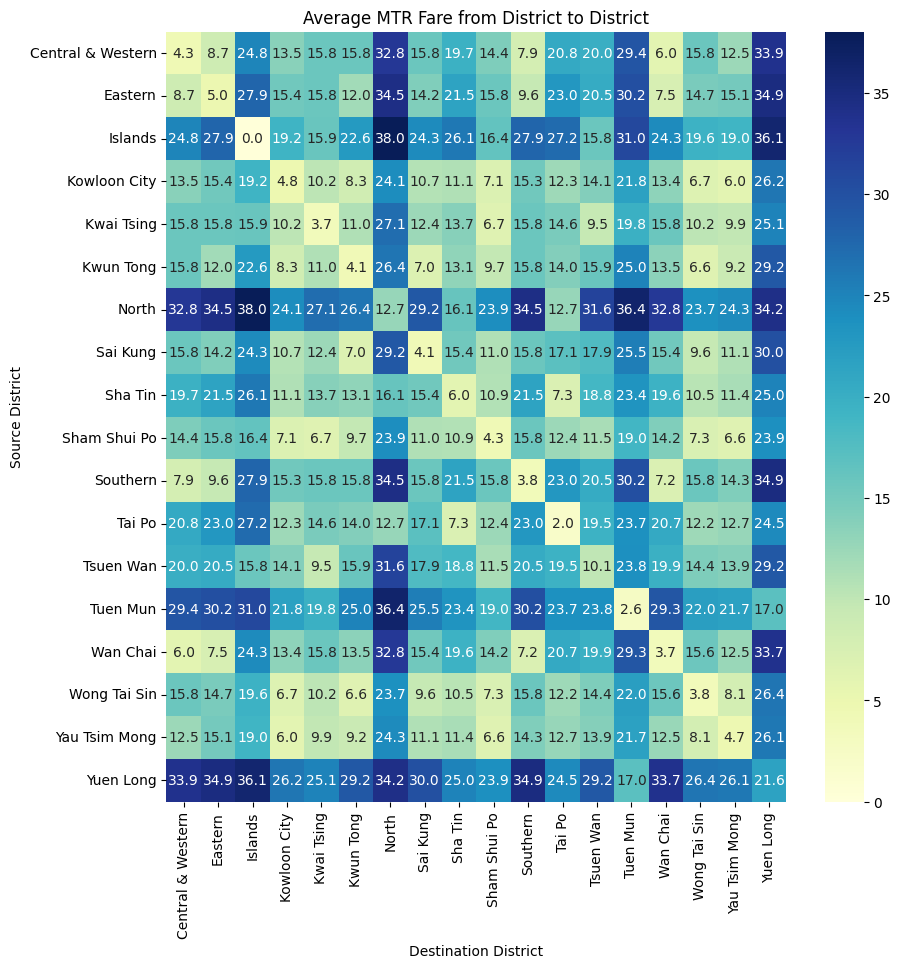
Below is a plot that shows the comparison of MTR price and distance starting from Central Station.



*Figure 25. Graph plotting MTR Price vs Distance from Central*

We chose Central station as it was a relatively central starting point to all the other stations in Hong Kong. From this initial analysis we obtained a positive correlation between prices and distance, with an increase of 1.1 HKD/km travelled. We also found that Lo Wu and Lok Ma Chau station were outliers with extremely high prices.

Below is a correlation matrix based on the average MTR fare from one district to another.



*Figure 26. Matrix showing average MTR fares from district to district*

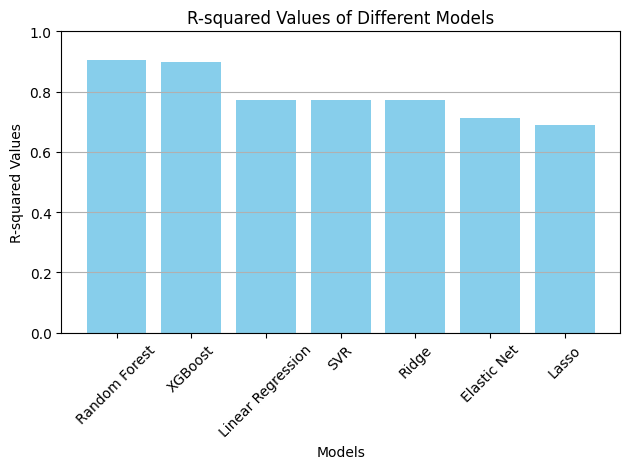
From this matrix, we found that North, Yuen Long, and Tuen Mun districts had higher overall fares. The discrepancy in the North and Yuen Long districts were attributed to Lo Wu and Lok Ma Chau station being located in those districts respectively, coinciding with the previous EDA result. While the higher fare in Tuen Mun could be due to there only being two stations in that district (Siu Hong and Tuen Mun) and it being relatively far from other stations. We also found that the fare of the Islands district to itself was zero.

The most important part of the EDA was that we were able to confirm that fares and distances were correlated and were able to find anomalies in certain stations (Lo Wu and Lok Ma Chau station) and districts (Islands district). After further research, we found that the reason Lo Wu and Lok Ma Chau stations were outliers was due to them being border stations into Shenzhen. As such, going to/from those stations resulted in higher fares which are used to compensate for the rest of the line. For the Islands district, the reason we found the fare going into the district from itself is zero was due to the fact that the Islands district only has one station (Tung Chung). These results allowed us to perform a more thorough data cleaning by removing any routes that go to/come from Lo Wu and Lok Ma Chau station and also remove those that begin and end at the same station.

### **6.4 Data Modelling and Evaluation**

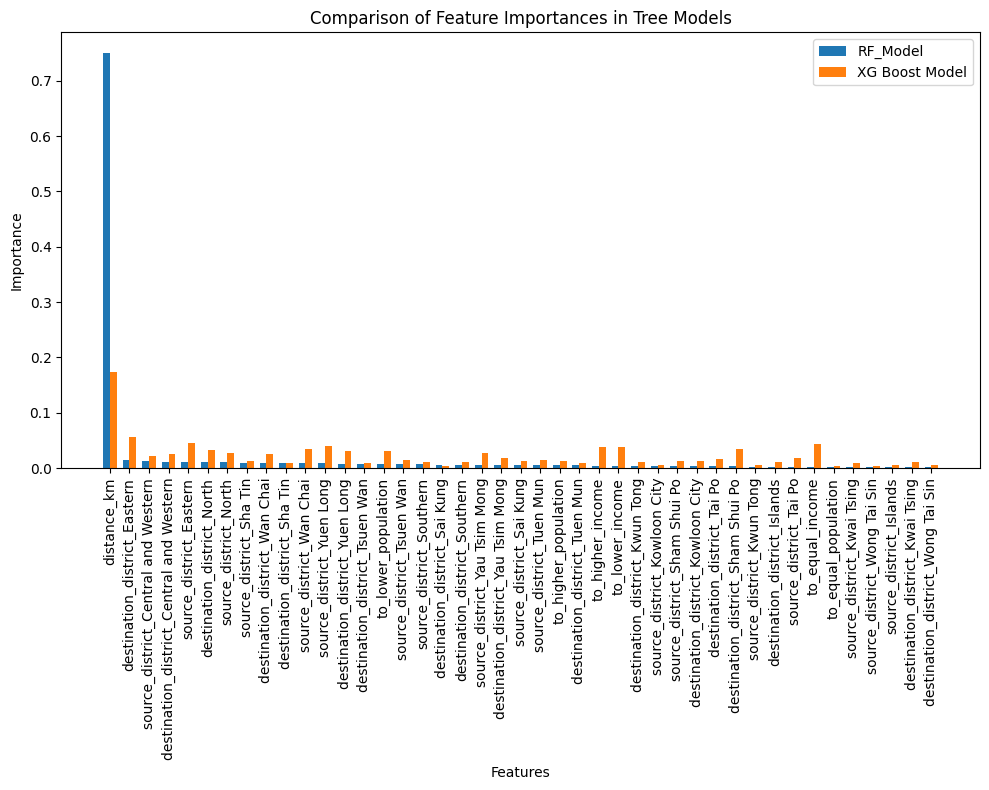
Before we began data modelling, we added the income median by district and population density by district into the dataset. However, we realized that there were only 18 distinct values, which will result in this attribute being more of a label compared to a quantitative value. As a result, we decided to alter the attributes into labels that signify whether we are traveling to a richer/poorer/equally rich district and whether we are traveling to a more/less/equally populated district. We also one-hot encoded the starting and ending district to help facilitate the modelling.

We decided to use various regression models, which were linear regression, linear SVR, ridge regression, lasso regression, and elastic net. We also decided to conduct XGBoost modeling and random forest modeling as our tree modeling. From the various models, we obtained the respective R squared values.



*Figure 27. R Squared values of initial models*

We found that the best models were the random forest model, XGBoost, and linear regression with R squared values of 0.91 and 0.9, and 0.77 respectively. The worst model was Lasso Regression with an R squared value of 0.69



*Figure 28.Comparison of feature importances in tree models*

We then compared the comparison of feature importance from both tree models. We found that both models agree that the most important feature was distance travelled. However, the random forest model gives vastly higher importance to distance, with other factors seemingly having less importance. This is the opposite with the XGBoost model, which gives distance the highest importance, however various other factors are given more importance compared to the random forest model. The XGBoost model seems to put more importance on income median whilst population is a more important factor for the Random Forest model.



*Figure 29. Comparison of coefficients from regression models*

We then decided to check the coefficients from the 5 regression models. Overall, we found that SVR gives coefficients higher values. Linear and ridge regression have similar values for the coefficients. Elastic net and lasso regression also seem to give a lot of variables zero as their coefficients.

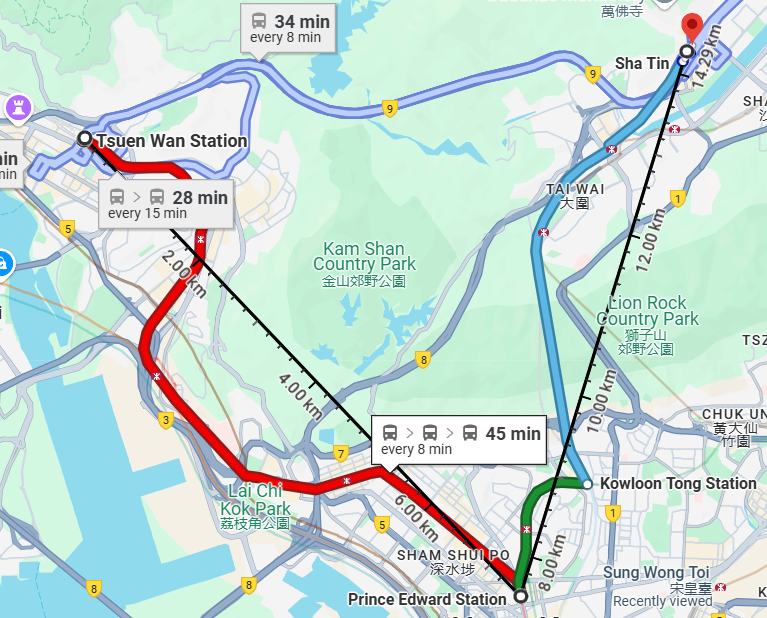
From these models, we decided to choose XGBoost as our chosen model as it has a high R squared value and its intrinsic advantages. We then decided to use the model to predict on all the routes to determine which was unreasonably expensive. The reason we chose to use routes is because it makes more sense than labeling a whole district as expensive, since inside a district there are numerous routes, and aggregating it all into a simple label might give the wrong or hide conclusions. To find the unreasonably expensive routes, we calculated the fare difference (actual fare - predicted fare) from all the routes and added it to the dataframe. We then obtain Q1 and Q3 to calculate the upper bound of Q3 + 1.5 \* Q1, any route that has a fare difference above the upper bound we consider unreasonably expensive.



*Figure 30. Percentage of unreasonably expensive routes by starting district*

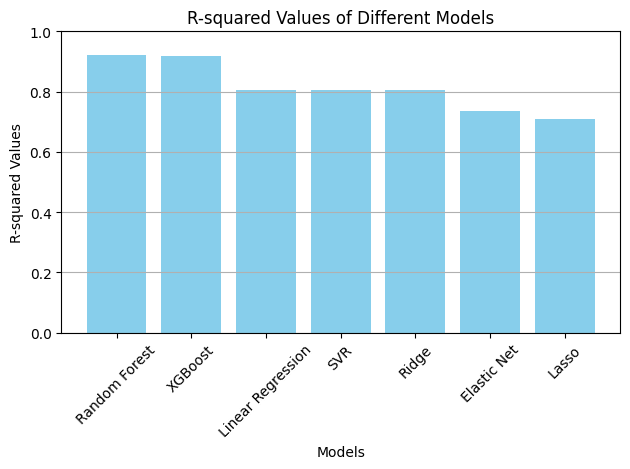
We obtain the result that Tsuen Wan, Islands, and Sha Tin have the most unreasonably expensive routes, with 20.6% of routes in Tsuen Wan considered unreasonably expensive. When we were looking into why this was the case, we discovered the issues lie with how we calculated distances. Since we used the Geodesic library, it calculated the distance as being a straight line between two stations. While this method might work with short distances between stations, it might skew the actual distance travelled. For example, here is the route from Tsuen Wan station to Sha Tin station.





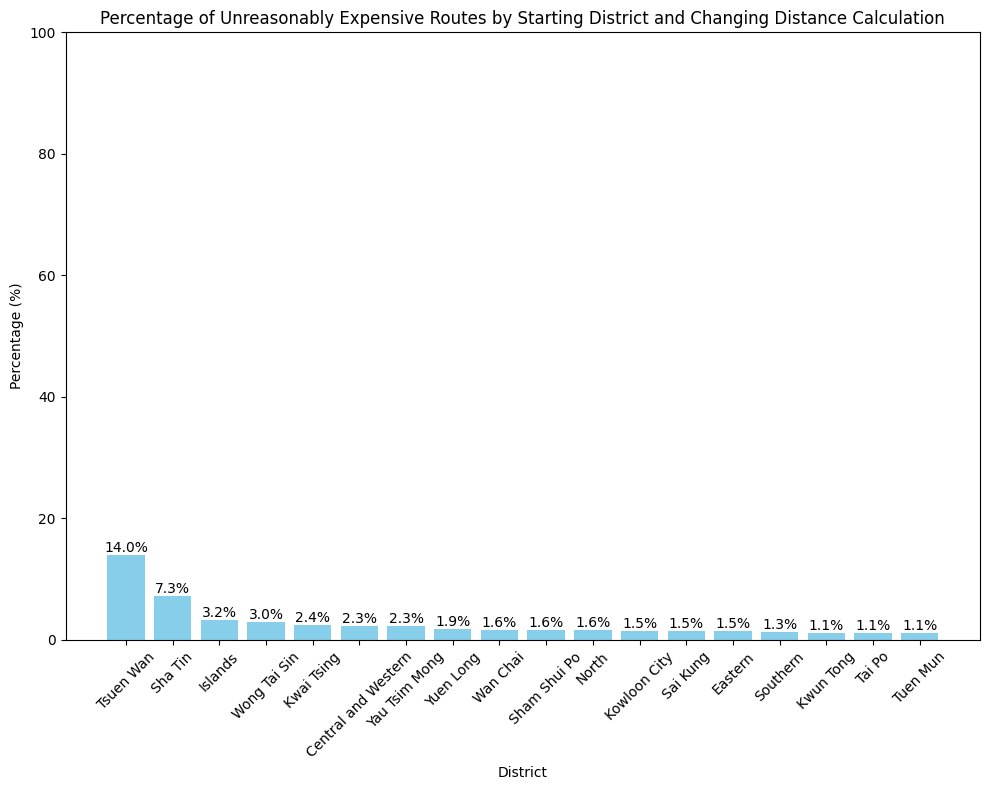
*Figure 31. Map showing different routes from Tsuen Wan station to Sha Tin station*

In the first picture, we can see the direct distance, which is around 7 km and how the Geodesic library calculates it. However the bottom picture shows what it would be like taking the MTR, which means travelling around 14.3 km. As a result, we changed our distance calculation by using the Kaggle dataset mentioned earlier, which makes the stations into nodes and distances between nodes, thus allowing us to use graph search algorithms to find the shortest distance between stations. After changing the distance calculations we ran the models again to obtain the following R squared values.



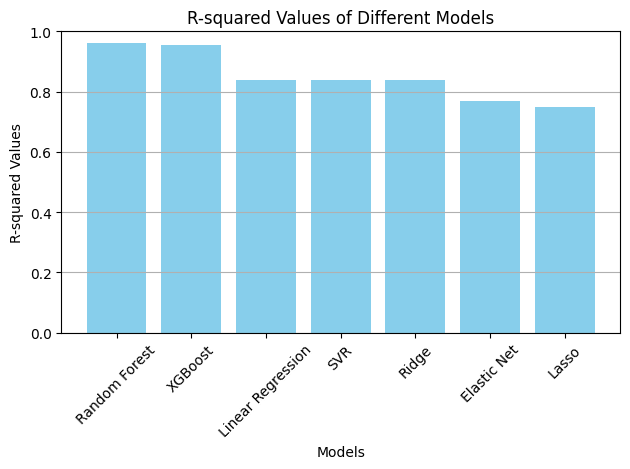
*Figure 32. R Squared values of models after distance recalculation*

After changing the distance calculation, the R squared values of models increased. With random forest, XGBoost, Linear Regression models obtaining values of 0.92, 0.92, and 0.8 respectively. From these models, we again chose the XGBoost model to obtain the following results and redid the analysis.



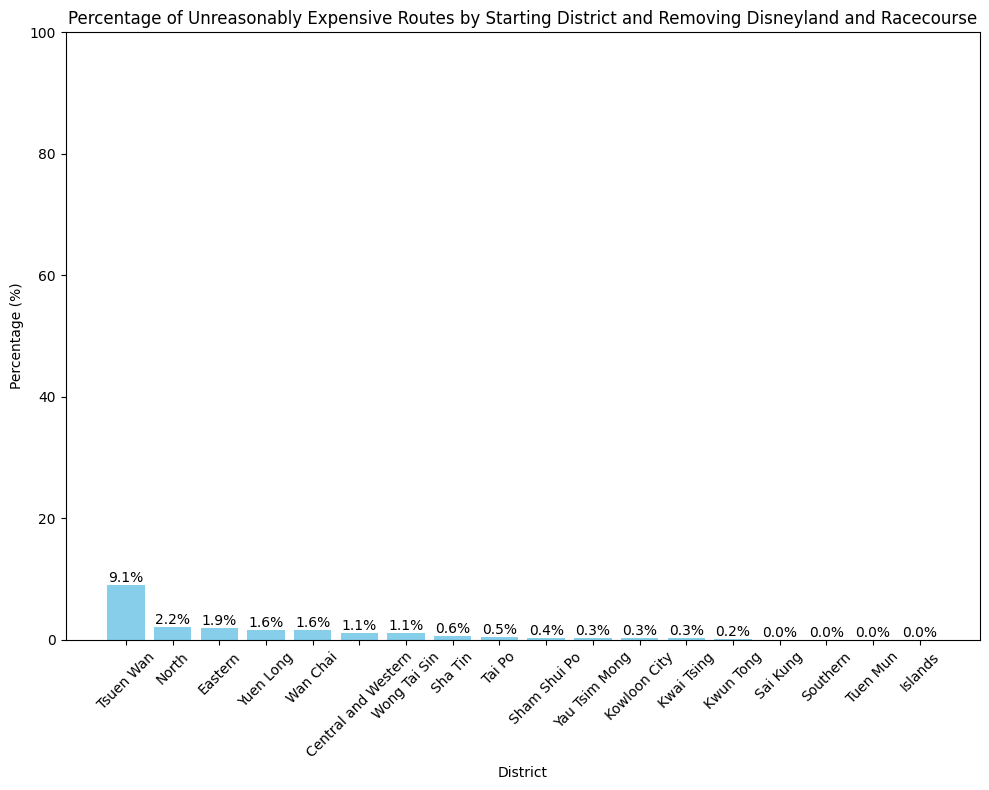
*Figure 33. Percentage of unreasonably expensive routes by starting district after distance recalculation*

By changing the distance calculation, we found that the percentages of expensive routes decreased in general, with Tsuen Wan and Sha Tin still having the most. After further analysis, we found potential explanations for these results. The reason why Tsuen Wan had the most was mainly due to Sunny Bay station and Disneyland Resort station being in the Tsuen Wan district. Due to them being cross island routes, they have higher fares and thus are the more unreasonably expensive routes. The reason why Sha Tin district had a higher number of expensive routes is due to Racecourse station. This station is only open on race days and as such has a higher price, thus explaining why these routes are expensive. We then decided to remove any routes that start at Disneyland Resort station and the Racecourse station. We still kept Sunny Bay station as people live there and pass through to get to the further stations on the Tung Chung line. We then ran the model again, to obtain the R Squared values.



*Figure 34. R Squared values of models after removing Disneyland and Racecourse*

After removing the Disneyland resort district, the R Squared values of all the models increased. With random forest, XGBoost, Linear Regression models obtaining values of 0.96, 0.96, and 0.83 respectively. From these models, we again chose the XGBoost model to obtain the following results and redid the analysis.



*Figure 35. Percentage of unreasonably expensive routes by starting district after removing Disneyland and Racecourse*

As a result of removing the specific stations, Tsuen Wan still has the most unreasonably expensive routes, however this is due to routes to Sunny Bay station being considered expensive by the model. The rest of the districts only have around 0-2% of their routes being unreasonably expensive.

### **6.5 Interpretations and Communications**

From the various modeling done, we obtained the most important factor in determining MTR fares is the distance travelled. We believe this is logical, as traveling longer distances requires more resources by the MTR company, thus charging higher prices. We also obtained that the Tsuen Wan district has the most number of “unreasonably expensive” routes. However this is due to routes starting from Sunny Bay that cross islands to New Territories and Hong Kong Island. As crossing islands requires more resources, it makes sense for these routes to be more expensive. For the other districts, only zero to two percent of routes are considered “unreasonably expensive” by the model. As a result, we believe that MTRs as public transportation in general are fair to Hong Kong residents.

There are some limitations in our analysis. The major limitation being that there are some routes that will probably not be taken due to there being better alternatives. In the earlier Tsuen Wan station to Sha Tin station example, if one were to take a bus they would only need 25 minutes in total, whilst using the MTR would take around 40 minutes and require multiple transfers. In reality, most people would take the bus if they needed to take this route and so that specific MTR route might never be taken. MTR car capacity and frequency are also factors that might be important in determining fares, however we were not able to find enough resources to implement this in a timely manner. Our analysis also does not take into account various subsidy schemes that the MTRs have.

## **7. References**

Hong Kong Transport Department. (n.d.). Routes and fares of public transport. Retrieved from <https://data.gov.hk/en-data/dataset/hk-td-tis_3-routes-and-fares-of-public-transport/resource/707cdb6b-e45c-4e76-aca6-b1622672dd73>

MTR Corporation Limited. (n.d.). MTR Lines (except Airport Express & Light Rail) Fares. Retrieved from <https://data.gov.hk/en-data/dataset/mtr-data-routes-fares-barrier-free-facilities/resource/91e47c22-3e5d-40fe-a779-308be439c41f>

Hong Kong Home Affairs Department. (n.d.). Hong Kong administrative boundaries. Retrieved from <https://data.gov.hk/en-data/dataset/hk-had-json1-hong-kong-administrative-boundaries>

Census and Statistics Department. (2022). *Population by age group and sex*. Retrieved from [<https://www.censtatd.gov.hk/en/wbr.html?ecode=B11303012022AN22>]

OpenStreetMap Community Forum. (2021). How to calculate distance between two coordinates? Retrieved from <https://community.openstreetmap.org/t/how-to-calculate-distance-between-two-coordinates/94431/3>

Transport Department. (2000). Transport Department's statement on bus fares. Retrieved from <https://www.td.gov.hk/en/publications_and_press_releases/press_releases/transport_department/index_id_596.html>

Transport Department. (2020). TD's response on request of reduction in bus fares following Lantau Link toll waiver. Retrieved from <https://www.td.gov.hk/en/publications_and_press_releases/press_releases/transport_department/index_id_3484.html>

Wikipedia. (2022). List of bus routes in Hong Kong. Retrieved from <https://en.wikipedia.org/wiki/List_of_bus_routes_in_Hong_Kong#Hong_Kong_Island_routes>

Wikipedia. (2025). List of MTR stations. Retrieved from <https://en.wikipedia.org/wiki/List_of_MTR_stations>

Citybus. (n.d.). Cityflyer Route. Retrieved from <https://www.citybus.com.hk/routes/airport-bus/route/index.aspx?intLangID=1>

Austroads. (2021). Prioritising Active Transport. Retrieved from <https://austroads.gov.au/__data/assets/pdf_file/0031/653368/AP-R711-24_Prioritising_Active_Transport.pdf>

Kaggle. (2024). Hong Kong MTR / Subway Networkt. Retrieved from <https://www.kaggle.com/datasets/d1om3d3s/hong-kong-mtr-network?select=MTR_Edges_Data.csv>