THE UNIVERSITY OF HONG KONG

COMP3522 REAL-LIFE DATA SCIENCE

Are Fares Fair?
A Data-driven Analysis into Hong Kong's Green MiniBus Routes

Final Report

By

Jayawardana Wickramasinghe Pathiranage Lakindu Ransika

(3036094631)

Supervised by

Dr. Loretta Choi

School of Computing and Data Science

May 3, 2025

1. Introduction

1.1. Background

Hong Kong’s public transportation system is one of the most efficient and widely used in the world, transporting around 11.7 million passengers daily (Transport Department, 2025c). This system mainly consists of the Mass Transit Railway (MTR), bus, and Green MiniBus (GMB). Given this daily reliance, the affordability of their fares is of significant public interest. In recent years, according to various news articles, there has been growing discontent across different districts regarding the fairness of these prices. For example, according to a lawmaker, Tien, the East Rail Line is about 30% cheaper than the West Rail Line while having lower travel times. This disparity has drawn public criticism (Yè, 2022). While this issue of fare equity is system-wide, the GMB's critical role in connecting underserved communities (Transport Department, 2017, pp. 43–44) makes it a logical starting point for analysis.

1.2. Problem statement

While these criticisms highlight the existence of a fundamental gap, the absence of an objective methodology makes these claims purely anecdotal. Focusing on GMB, the central problem can be broken down into two sub-problems: firstly, modeling the fares for GMB routes across different districts in Hong Kong, and secondly, determining any districts with unreasonably expensive GMB prices.

1.3. Project Significance

The significance of this project lies in its capacity to bridge the gap between subjective public perception and objective reality. The systematic modeling of fares and classification of unfairly priced districts provides tangible evidence against GMB fare inequity. This contributes towards shaping future government policy, thereby resulting in improved standards of living.

1.4. Objectives

To assess the equity of GMB fares, this project accomplished the following objectives:

- Develop a robust and interpretable regression statistical model for GMB fares.
- Identify districts with unreasonably expensive GMB fares by interpreting the model's coefficients.

1.5. Deliverables

This project delivered a GitHub repository, which contained a Jupyter notebook. This notebook presented the data processing, EDA, model training, and model evaluation pipelines for the final model.

1.6. Outline

This report is organized into five main sections. Section 2 highlights the methodology detailing the project setup, data collection, data processing, Exploratory Data Analysis (EDA), and modeling pipelines. Section 3 presents the modeling results and the interpretations of the final model. Section 4 provides a discussion of these results, interpreting their real-world implications while highlighting the project limitations. Finally, section 5 provides a concluding summary of the project's findings and potential future work.

2. Methodology

This section details the methodology employed in developing a statistical model for classifying unfairly priced GMB districts. Primarily, various datasets were collected and processed. An EDA was then conducted to understand the features collected and their relationship to the target variable. Uninformative or redundant features were then removed to refine the feature set. Finally, various models were trained and evaluated to select the optimal model. This section begins by outlining the software and tools used for the project.

2.1. Project Setup

Python was chosen for its extensive ecosystem of data processing, visualization, and machine learning libraries. Google Collab, a cloud-based interactive Jupyter Notebook environment, was selected due to its collaborative nature. Additionally, Supabase, a Postgres development platform, was utilized to store and share datasets. The key Python libraries were NumPy for array-based numerical operations, Pandas for data wrangling, Folium for geospatial visualization, Matplotlib for data visualization, and sci-kit learn for machine learning.

2.2. Data Collection

This section details the collection of datasets that were used for EDA and statistical modeling. Four categories of data were collected: GMB data, geographic data, points of interest data, and demographic and socioeconomic data.

2.2.1. GMB Data

Two GMB datasets were collected. This included the 'Route information of GMB'(Transport Department, 2025b) and 'Green Minibus Route' (Transport Department, 2025a) datasets from DATA.GOV.HK. Both datasets were chosen because they contributed different useful features. The first dataset consisted of full fare, outbound/circular/inbound, journey time, regularity of operation, and stop locations for each route, while the second dataset contained a LineString Object and distance for each GMB route. The LineString Object stored the entire GMB route, allowing for more accurate and fine-grained geospatial data processing.

2.2.2. Geographic Data

Geographic data consisted of the 'District boundary' dataset (Home Affairs Department, 2018) obtained from DATA.GOV.HK. The borders for each district were stored in GeoJSON format, which allowed the use of efficient geospatial data processing libraries such as GeoPandas. Additionally, this dataset was chosen as it allowed for the efficient extraction of district-wise features from other collected datasets.

2.2.3. Points of Interest (POI) Data

Residential, commercial, and industrial areas were considered to be POI due to their tendency to attract specific population segments. In order to travel to these places, public transport is often used, resulting in higher demand. Hence, data from the aforementioned categories were collected.

Data was gathered from various sources. Public residential areas (Hong Kong Housing Authority, 2025) and industrial areas (Planning Department, 2021) were collected from

DATA.GOV.HK. Private residential areas (Wikipedia, 2025) were extracted from Wikipedia. Finally, to construct a more reliable dataset for commercial areas, two different information sources—the Corporate Locations website (Corporate Locations, n.d.) and the Hong Kong office rental (Hong Kong Office Rental, n.d.) website—were merged.

2.2.4. Demographic and Socioeconomic Data

Given the shaping of government policies by various demographic and socioeconomic factors, this project included three such datasets, 'average household size and median monthly household income of households by District Council district' (Census and Statistics Department, 2024c), 'domestic households by District Council district and monthly household income' (Census and Statistics Department, 2024b), and 'land area, mid-year population and population density by District Council district' (Census and Statistics Department, 2024a). District-wise median monthly income, percentage of households in three income brackets, land area, and population density were chosen from the aforementioned datasets. Both median monthly income and income distribution were chosen as median income does not describe the income stratification within a district. Land area was included as a spatial indicator, and population density was selected as a proxy for GMB demand. Additionally, average household size and mid-year population were excluded to avoid multicollinearity.

In summary, data encompassing 4 different categories—GMB, geographic, points of interest, demographic, and socioeconomic—were collected from eleven different sources. The curated data was then processed to enhance its impact on subsequent EDA and model development stages.

2.3. Data Preprocessing

This section outlines the data cleaning, feature engineering, and feature transformation to prepare the curated dataset for EDA.

2.3.1. Data Cleaning

This stage involved identifying missing, erroneous, or inconsistent data. A manual and programmatic investigation concluded that the datasets were of very high quality. The only issue identified was the inconsistent naming for the 'Central and Western' district, which used both 'and' and '&'. This was fixed by merging the related keys.

2.3.2. Feature Engineering

The district boundary data was used to engineer various district-wise features, given the project's objective to identify district-level patterns. In conjunction with the Google Maps API, the residential, commercial, and industrial districts were extracted. A district is residential/commercial/industrial if at least one such area exists in that specific district. A relaxed binary approach was chosen as it is more robust to incomplete data in the POI sources. Afterwards, the stops for each route were utilized to engineer the district-wise stop count, a measure of accessibility. Finally, since each route passes through many districts, to create route-level features, each district-wise feature was averaged across all the districts passed by a given route. For instance, if a route passed through Southern and Western, the route-level median income would be calculated by averaging across the median incomes for Southern and Western districts.

2.3.3. Feature Transformation

To transform the feature set into a statistically interpretable form, textual features such as route sequence and 'regularity of operation' were one-hot encoded. Afterwards, the last column was dropped to prevent the dummy variable trap.

Due to the sensitivity of gradient descent-based machine learning models to scaling, standardization was used. Standardization was avoided on binary features, as breaking the binary distribution of the data would result in loss of information.

2.4. Exploratory Data Analysis (EDA)

To better examine the dataset before modeling, an EDA was conducted. A comprehensive summary of the processed dataset, histograms for each feature, a covariance heatmap, various geospatial maps, and a feature importance plot were constructed. Pandas handled intermediate processing, matplotlib plotted visualizations, and Folium constructed geospatial plots.

2.4.1. Data Overview

The dataset consisted of 1143 rows with 16 features each. Table 1 highlights the data type, mean, standard deviation (std), minimum, and maximum values for the primary features of interest. The complete data overview is available in Appendix A.

Feature Name	Description	Data Type	Mean	Std	Min	Max
full_fare	Full fare in HKD for a single journey	float	7.8	3.7	0.0	28.6
distance_m	Distance in metres for the entire route	float	7,961	6,275	435	45,443
journey_time	Journey time in minutes for the entire route	int	17.2	10.9	1.0	78.0
route_sequence	Flag if route is inbound	boo1	0.67	0.47	0	1
service_R	GMB that operates regularly	bool	0.56	0.50	0	1

Table 1: Descriptive statistics for key Features

The summary statistics in Table 1 reveal a strong right-skew in the distributions of the target variable, distance_m, and journey_time. This skew is evidenced by their maximum values, which are more than 5 standard deviations above their respective means. Given that distance_m and journey_time exhibit similar skewness to full_fare, it was hypothesized that they are strong predictors for full_fare.

2.4.2. Univariate Analysis and Data Distribution Comparison

To visually validate this hypothesis, histograms with a Kernel Density Estimate (KDE) curve were plotted (as shown in Figure 1).

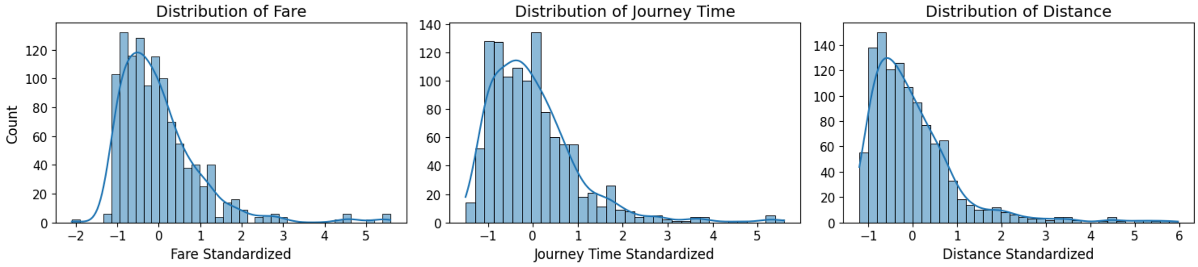


Figure 1: Distributions of fare, journey time, and distance

The standardized distributions of fare, journey time, and distance have similar shapes and right skewness. This observation further reinforces the hypothesis.

2.4.3. Bivariate Analysis and Multicollinearity Assessment

The similar distributions observed between journey time and distance in Figure 1 additionally highlight potential multicollinearity among features. Given the adverse effects of multicollinearity, an assessment is required. A correlation matrix for the key features was plotted as shown in Figure 2. The correlation matrix for the entire feature set is available in Appendix B.

Most significantly, journey time and distance have a very high correlation of 0.91. The analysis also reveals strong correlations between the four income-related features and among the one-hot encoded service_R and service_T. These high correlations confirm the presence of multicollinearity, highlighting the information redundancy within the feature set, thereby establishing a need for feature selection.

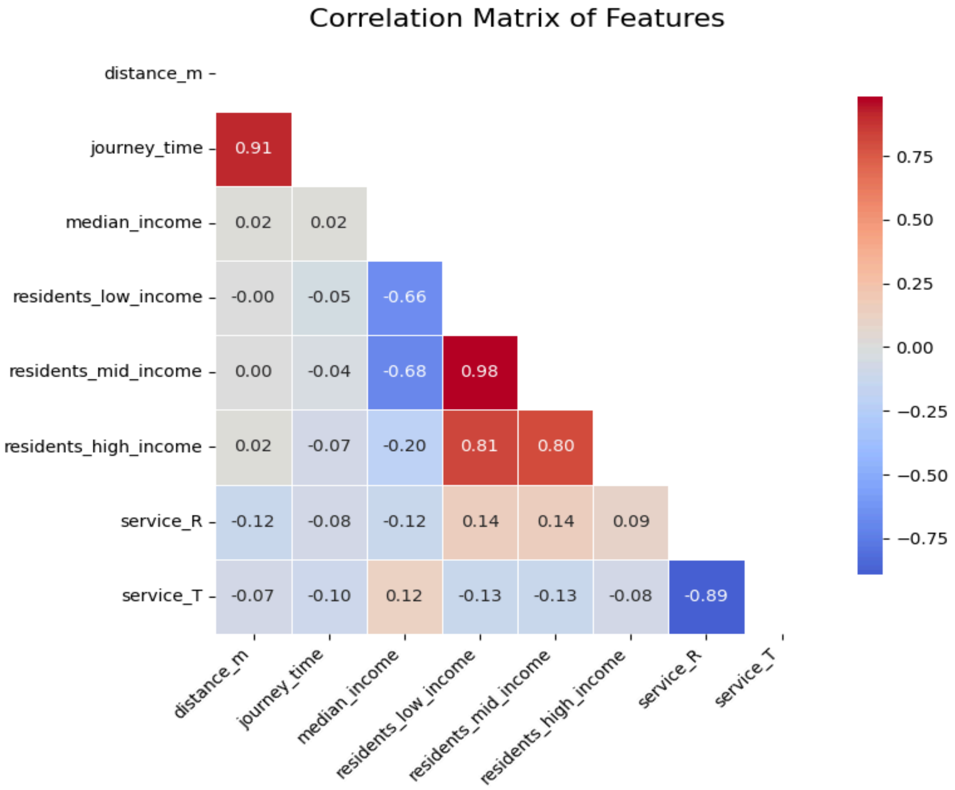


Figure 2: Correlation matrix for key features

2.4.4. Preliminary Feature Importance

To explore the potential predictive power of each feature, while accounting for multicollinearity, a feature importance plot was constructed using Elastic Net regression (as shown in Figure 3). Elastic Net regression is effective at penalizing correlated features while allowing direct interpretation of its penalized coefficients as feature importance.

From Figure 3, it is clear that distance, service_R, and route_sequence are the most important features. The high coefficients of service_R and service_T highlight the predictive value in both features despite the high correlations. The information redundancy of median income to the other income-related features is indicated by its 0 importance. This analysis confirms the need for feature selection for developing the final model.

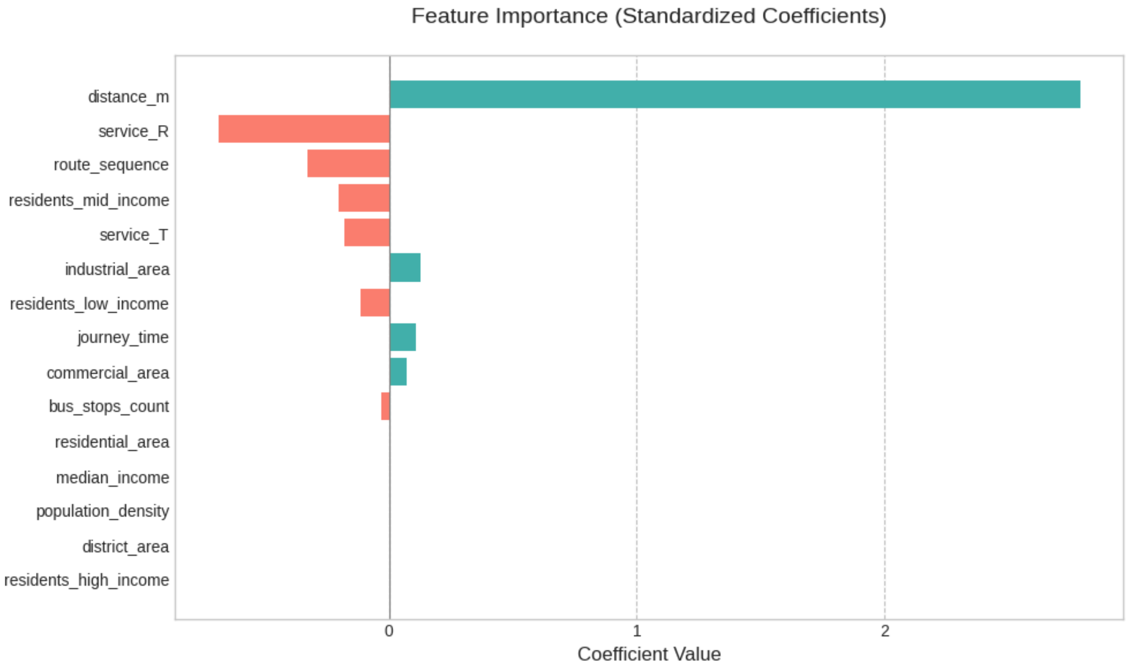


Figure 3: Elastic Net regression feature importance plot

2.4.5. Geospatial Analysis

Before model selection and training, to visualize any districts with unreasonably expensive fares, a bubble map of fares was plotted. Additionally, it was overlayed on an accessibility heatmap to identify any relationships between fares and demand (as shown in Figure 4).

According to Figure 4, Yau Tsim Mong, Wan Chai, Eastern, Kwun Tong, and Tuen Mun had the largest number of large blue circles, highlighting them as potential GMB inequity districts. Yau Tsim Mong, Tai Po, Yuen Long, Kwai Tsing, Tseung Wan, Kwun Tong, Central & Western, and Southern had the brightest red spots, showcasing them as the districts with the greatest accessibility. Additionally, only Yau Tsim Mong overlapped between the two district lists, highlighting the poor relationship between accessibility and fares.

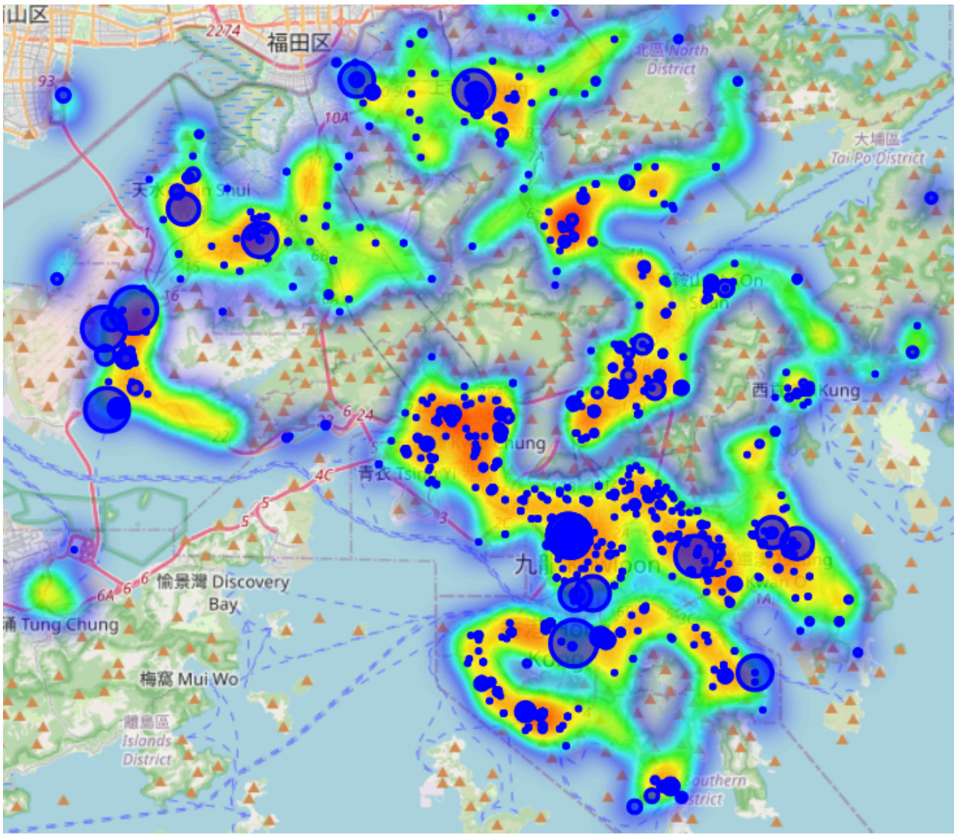


Figure 4: Heat map overlayed on a bubble map

Considering the results of the EDA, it was hypothesized that distance is the most important feature. Additionally, a conservative hypothesis that no district is unreasonably expensive for GMB was established.

2.5. Model Selection

This stage involved engineering more complex interaction features and feature selection. To address the presence of highly correlated variables in Figure 2, interaction features were introduced. Given the large number of features present, feature selection was chosen to improve interpretability and to mitigate the risk of overfitting. Statistical t-tests alongside Elastic Net linear regression were chosen as the primary methods of feature selection. To ensure no potentially strong models were discarded, each model was evaluated using the optimal hyperparameters found via k-fold cross-validated grid search.

2.6. Model Evaluation

Six models were evaluated on the final feature set. This included four linear regression models—Ordinary Least Squares (OLS), Ridge, Lasso, Elastic Net—and two non-linear tree-based models—XGBoost and Random Forest. These six models were selected given their built-in feature importance metric. Linear models were employed due to their high interpretability, and non-linear models were utilized as they can capture non-linear patterns. Additionally, four distinct linear models were chosen due to their differing regularisation terms, which incorporate different soft constraints, and two distinct non-linear models were selected due to the differences in how important features are determined.

To accurately account for train-test split randomness and overfitting, a k-fold cross-validated grid search was utilized to find the optimal hyperparameters. The optimal hyperparameters were selected by choosing the hyperparameter combination that resulted in the lowest standard deviation for MSE among the top 50 hyperparameter combinations with the lowest MSE. This method ensures that the chosen model is both highly performant and stable.

The performance across each CV fold was used to construct two bar plots. The first bar plot compared the mean R^2 across models, while the second compared the standard deviation of R^2. Mean measures the overall performance, while instability in standard deviation is an indicator of overfitting. Afterwards, the final model was chosen, and interpretations were extracted using the feature importance plot. Finally, to determine the unreasonably expensive districts, an OLS model was trained. OLS was selected as it is one of the most interpretable models, and it was trained using the most influential feature and the districts passed by each route as features. A district was considered to be unreasonably expensive if its standardized coefficient had a z-score greater than 3, which conservatively indicates the presence of outliers.

3. Results

This section outlines the chronological model selection process, followed by model evaluation and final model interpretation. Primarily, an iterative feature selection process for the final model was conducted, where the trade-off between performance and interpretability was considered to determine the final feature set. Following this, the best model for the final feature set was found. This section concludes with a detailed interpretation of the final model and the identification of any unreasonably expensive districts.

3.1. Model Selection

The process of model development was iterative and required balancing predictive power with interpretability. This subsection details the progression from a complex model to a final OLS model with two core predictors (as summarized in Table 2).

Model No	Model Description	R^2	Objective	Observations
1	Elastic Net with full interaction (153)	0.834	Maximize predictive power when using interaction terms	Most important features were interaction terms making the model uninterpretable
2-3	Elastic Net with recursive feature selection (88)	0.833	Improve interpretability by removing weak predictors	The model remained uninterpretable. Interaction term approach was abandoned.
4	OLS with multicollinearity based feature selection (9)	0.779	Establish OLS baseline	Strong performance and interpretability
5	OLS with t-test selection (5)	0.764	Improve interpretability	Substantial gain in interpretability
6	OLS with Geographic Features (18)	0.797	Evaluate geographic features for performance	Highest R^2 without interaction terms. Uses only two core predictors, 17 districts and distance.

Table 2: Chronological model selection summary

3.1.1. Initial Exploration with Interaction Terms (Models 1-3)

As indicated in Table 2, the initial approach was to maximize the predictive power of interaction regression. This expanded the feature set to 153 features, resulting in a high R^2 of 0.834. However, as noted in the observation columns, Model 1 was uninterpretable as the feature importance plot was dominated by interaction features.

To address this, Models 2-3 were constructed by removing weak predictors with less than 0.001 feature importance. Despite this reduction to 88 features, the interaction features still dominated the feature importance plot, thereby rendering it uninterpretable.

3.1.2. Pivot to More Interpretable OLS Models (Models 4-5)

Given the difficulties in interpreting interaction regression, interaction features were abandoned. As mentioned in Table 2, OLS was selected as the model. This was to use statistical t-tests for feature selection. Given OLS's sensitivity to multicollinearity, features with high correlation and high VIF index were removed. The resulting feature set consisted of distance_m, journey_time, route_sequence, district_area, population_density, residents_low_income, bus_stops_count, service_R and service_T. Model 4 achieved a high R^2 of 0.779 and was very interpretable.

To further increase the interpretability, a t-test was conducted to remove less important features. This reduced the feature set to distance_m, route_sequence, residents_low_income, service_R, and service_T. Despite the drop in R^2 by 0.015, the reduction of 5 features greatly boosted interpretability.

3.1.3. Experimenting with Geographic Features (Model 6)

To explore further feature set simplifications, the OLS model suggested for classifying unreasonably expensive districts was also trained. While having only two core predictors, distance and districts passed, as shown in Table 2, model 6 managed to achieve the highest R^2 among models with non-interaction features. Additionally, its range of VIF values was between 1.04 and 2.01, suggesting moderate but acceptable multicollinearity. For this model, the Islands District was removed as it is served by only a single GMB route.

3.1.4. Final Feature Set Selection

Given the difficulty in interpreting interaction regression models, they were dropped. By conducting multicollinearity and t-test-based feature selection, Model 5 managed to achieve a high R^2 while having interpretability. However, the feature set of Model 6 was chosen due to Model 6 having the highest R^2 among the non-interaction term regression models while having the highest interpretability, and acceptable multicollinearity.

3.2. Model Evaluation

The mean and standard deviation of R^2 averaged across each fold for the Elastic Net, Lasso, OLS, Random Forest, Ridge, and XGBoost models can be found in Figure 5.

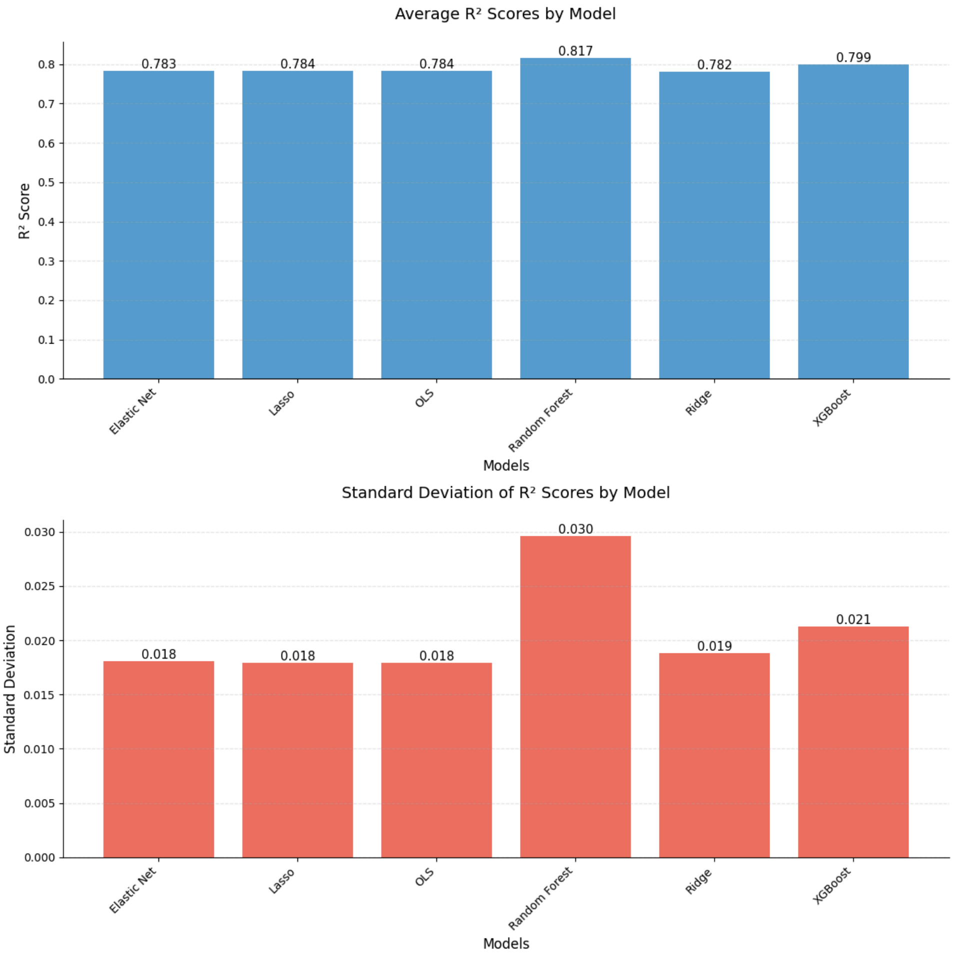


Figure 5: Model evaluation results

From Figure 5, it can be observed that all the linear models have slightly lower mean R^2 compared to the more complex non-linear models, while having much lower standard deviations of R^2. This highlights that while being competitive in terms of model performance, the linear regression models are much less prone to overfitting than the non-linear models. Given this stability, the linear models are preferred for constructing the final model.

Since all four linear models have similar means and standard deviations of R^2, and to make the final model more robust, the final model would be constructed by taking the average across each model coefficient. For example, if distance had coefficients a, b, c, and d, the final model would have the coefficient (a+b+c+d)/4 for distance.

3.3. Final Model Interpretation

A feature importance plot of the final model, "Mean", alongside its constituent models, is observed in Figure 6. The red error bars indicate one standard deviation from the mean coefficient, highlighting the level of agreement between models.

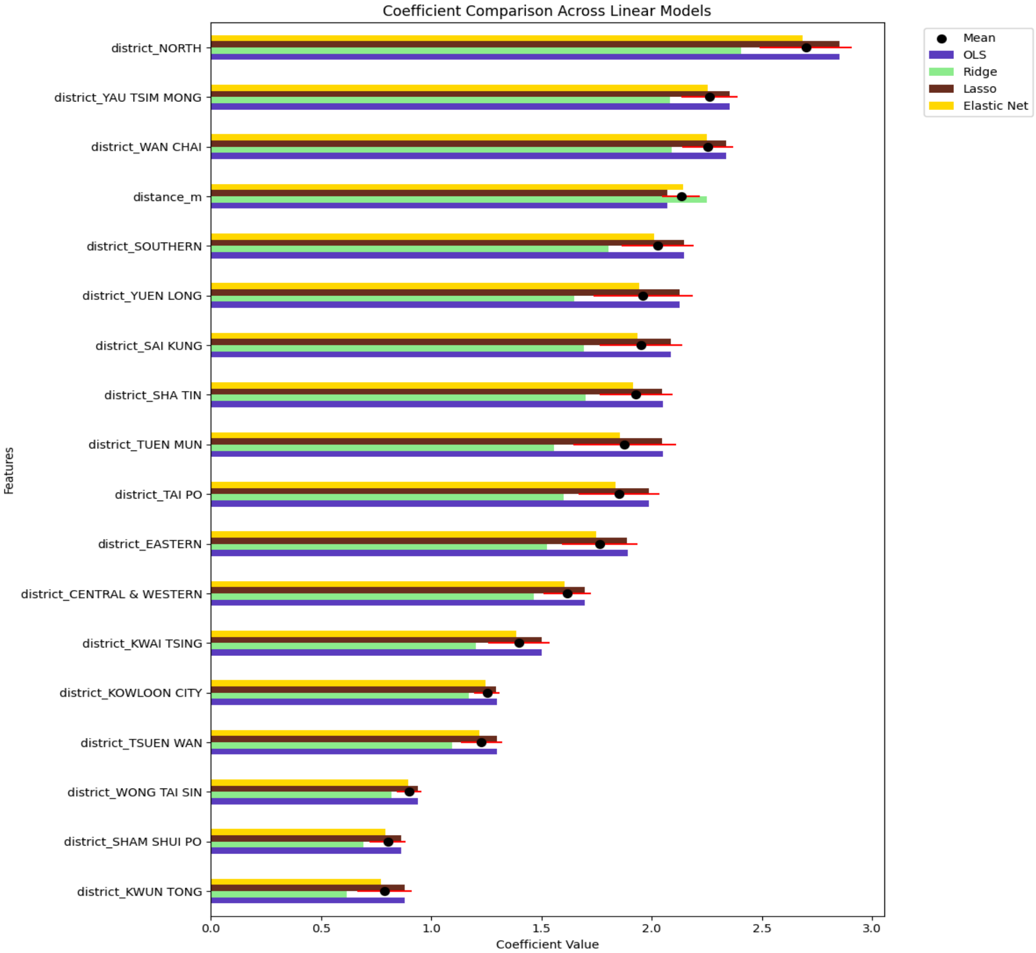


Figure 6: Feature importance plot of final model and its constituent models

According to Figure 6, North, Yau Tsim Mong, and Wan Chai have the highest coefficients. Given the nature of multi-label encoding, the coefficient of each district corresponds to the average cost induced when a GMB passes through that specific district. Therefore, North, Yau Tsim Mong, and Wan Chai are the most expensive districts and contribute much more to fares than distance.

Distance is ranked the fourth most important feature, as shown in Figure 6, highlighting it as a crucial predictor for GMB fare. The model classifies this relationship with high precision: on average, for every 1 km traveled on a GMB, HKD 0.34±0.01 is paid.

In Figure 6, it can be observed that all models, excluding Ridge, have coefficients within one standard deviation of the mean coefficient for all the features. Additionally, this is observed despite the mean model incorporating information from the Ridge model. These observations highlight the robustness and reliability of the mean model.

Figure 7 depicts the standardized coefficient plot of the final model. It is observed that all districts have a z-score less than 2. Since a district is classified as unreasonably expensive if it has a z-score greater than 3, no district is unreasonably expensive for GMB.

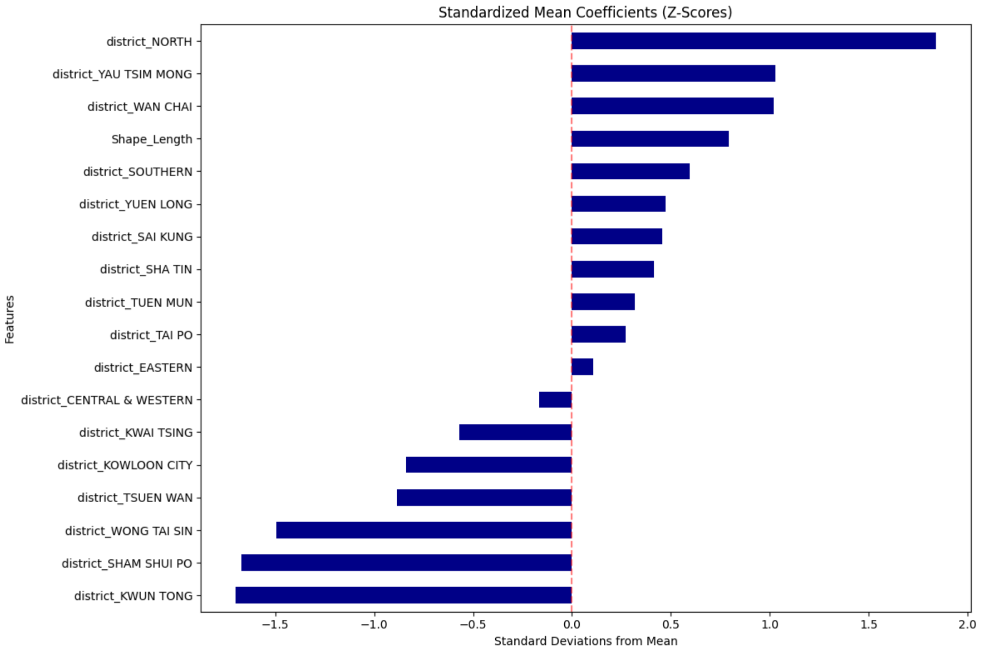


Figure 7: Z-scores of final model coefficients

4. Discussion

4.1. Classifying the Districts with Unreasonably Expensive GMB Fares

A primary objective of this project was to identify districts with GMB fare inequity. The analysis found no districts that met the predefined criteria for being classified as unreasonably expensive. This finding offers quantitative evidence to policymakers suggesting that district-level fare adjustments may not be immediately necessary. This is further supported by the model indicating distance traveled as a primary driving factor for GMB fares.

4.2. Limitations

This project mainly focused on assessing GMB fare inequity across districts. However, there may be more granular areas where fares are unreasonably expensive. Additionally, defining route-level features as the simple district-wise average may result in noise. For example, a district that contains only 1% of a route provides equal contribution to the route-level feature as the majority contributing district. The statistical threshold for unreasonably expensive might be too conservative, potentially overlooking districts that, while not extreme outliers, could still be considered to have fare inequity. Finally, this analysis was conducted over GMBs, but fare inequity may be present in other modes of public transportation, such as MTR and bus.

5. Conclusion

This project was initiated to address the growing public discontent that public transport fares are priced unreasonably. To do so, this project focused on GMB due to its critical role in connecting underserved communities. The central goal was to provide objective evidence regarding these subjective claims.

Through a rigorous process of data collection, processing, and iterative modeling, a final OLS regression model was developed. This model had strong predictive power ($R^2 = 0.797$) and high interpretability. Analysis yielded two key findings. Firstly, no district met the conservative statistical criteria for being unreasonably expensive. Secondly, the model indicates that distance traveled is a primary driver of fares. Furthermore, the analysis suggests that certain districts, such as North, Yau Tsim Mong, and Wan Chai, are associated with a greater contribution to fares.

The significance of these findings is that they provide tangible evidence that does not support the anecdotal claims that GMB, which is a part of Hong Kong's public transportation system, is priced unfairly at the district level. This suggests to policymakers that fare adjustments to the GMB system may not be necessary at this time.

Despite the robustness of the model, its limitations highlight several potential avenues for future work. Analysis could be conducted over more granular areas than the district level, which might allow for a more fine-grained assessment of fare inequity. Additionally, when calculating the district-wise average, a more sophisticated weighted average could be used. The analysis could also be repeated with a less conservative statistical threshold for what is deemed unreasonably expensive. Finally, this analysis could be expanded to other modes of transportation.

References

Census and Statistics Department. (2024a, March 12). *Table 110-02001 : Land area, mid-year population and population density by District Council district*. Census and Statistics Department. https://www.censtatd.gov.hk/en/web_table.html?id=110-02001

Census and Statistics Department. (2024b, April 5). *Table 130-06804 : Domestic households by District Council district and monthly household income*. Census and Statistics Department. https://www.censtatd.gov.hk/en/web_table.html?id=130-06804

Census and Statistics Department. (2024c, April 5). *Table 130-06806 : Average household size and median monthly household income of households by District Council district*. Census and Statistics Department. https://www.censtatd.gov.hk/en/web_table.html?id=130-06806#

Corporate Locations. (n.d.). *Hong kong business locations office district guide*. Corporate Locations. Retrieved October 10, 2025, from https://www.corporatelocations.com.hk/hong-kong-district-guide.php

Home Affairs Department . (2018, March 27). *District boundary (English / Traditional Chinese) (JSON)*. DATA.GOV.HK. https://data.gov.hk/en-data/dataset/hk-had-json1-hong-kong-administrative-boundaries/resource/706ed666-8f6c-4869-8c18-b74f863a5d22

Hong Kong Housing Authority. (2025, July 1). *Location and profile of public housing estates - Public housing estates*. DATA.GOV.HK. https://data.gov.hk/en-data/dataset/hk-housing-eslocator-eslocator/resource/160124ca-0993-4074-a0c2-849db912c455

Hong Kong Office Rental. (n.d.). *Business districts of Hong Kong*. Hong Kong Office Rental. Retrieved October 10, 2025, from https://www.hongkongofficerental.com/business-districts

Planning Department. (2021, December 30). *Table 1: Existing industrial stocks*. DATA.GOV.HK. https://data.gov.hk/en-data/dataset/hk-pland-pland1-2020-area-assessments-of-industrial-land-in-the-territory/resource/2e329437-5953-46a1-a38c-cd544a9266f7

Transport Department. (2017). *Public Transport Strategy Study* (pp. 43–44). https://www.td.gov.hk/filemanager/en/publication/ptss_final_report_eng.pdf

Transport Department. (2025a, February). *Green Minibus Route*. Common Spatial Data Infrastructure (CSDI) Portal. https://portal.csdi.gov.hk/geoportal/?lang=en&datasetId=td_rcd_1697082463580_57453

Transport Department. (2025b, February). *Routes and fares of public transport (GeoJSON) - Route information of GMB (English / Traditional Chinese / Simplified Chinese)*. DATA.GOV.HK. https://data.gov.hk/en-data/dataset/hk-td-tis_23-routes-fares-geojson/resource/ede023b4-f470-4edf-b1a0-dc6713e6021e

Transport Department. (2025c, July). *Fact Sheet on Transport*. Transport Department. https://www.td.gov.hk/en/publications_and_press_releases/publications/free_publications/fact_sheet_on_transport/index.html

Wikipedia. (2025, June 5). *Private housing estates in Hong Kong*. Wikipedia. https://en.wikipedia.org/wiki/Private_housing_estates_in_Hong_Kong

Yè, Z. (2022, May 4). *dōng tiě guò hǎi duàn tōng chē xiàn「dōng píng xī guì」tián běi chén pī zhèng fǔ pāo bō gǎng tiě shì bù fù zé rèn [The opening of the East Rail Cross-Harbour section has led to a rise in prices in the east and west. Tin Pei-chun* *criticised the government*. Epoch Times HK.

https://hk.epochtimes.com/news/2022-05-04/1555054

Appendix A

Detailed Data Overview

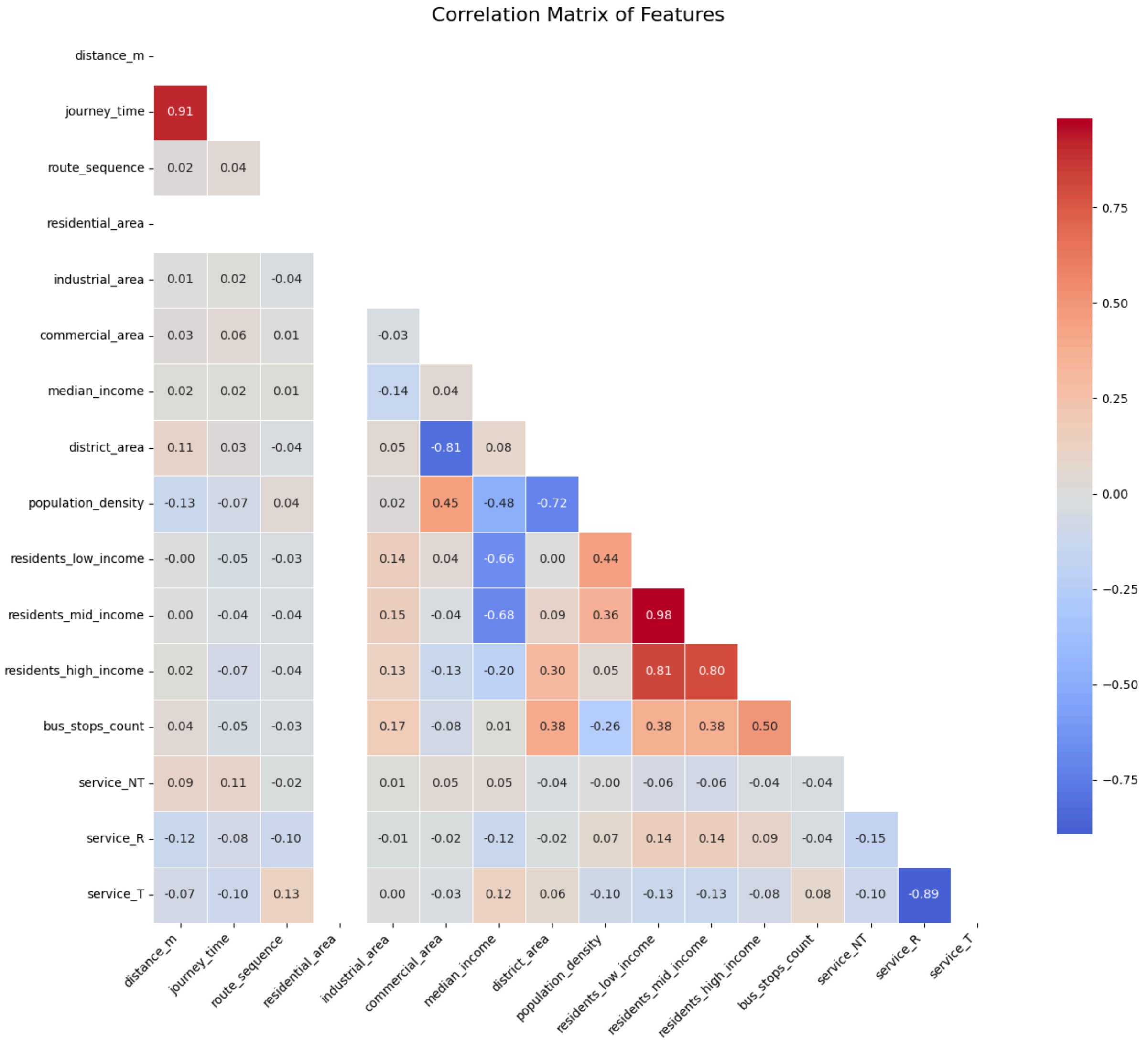
Feature Name	Description	Data Type	Mean	Std	Min	Max
full_fare	Full fare in HKD for a single journey	float	7.8	3.7	0.0	28.6
distance_m	Distance in metres for the entire route	float	7,961	6,275	435	45,443
journey_time	Journey time in minutes for the entire route	int	17.2	10.9	1.0	78.0
route_sequence	Flag if route is inbound (not circular or outbound)	boo1	0.67	0.47	0	1
bus_stops_count	Districtwise average of the number of GMB stops	float	831	245	18	1245
residential_area	Flag if district contains at least 1 residential area	bool	1.00	0.00	1	1
industrial_area	Flag if district contains at least 1 industrial area	bool	0.99	0.08	0	1
commercial_area	Flag if district contains at least 1 commercial area	bool	0.69	0.46	0	1
district_area	Districtwise average of the district land in square km	float	62.54	49.90	0.00	182.28

population_density	Districtwise average of the population density in persons per square km	float	18,875	17,724	0	59,600
residents_low_income	Districtwise average of the number of domestic households with income less than HKD 10,000 in 1000s	float	30,986	11,670	0	55,900
residents_mid_income	Districtwise average of the number of domestic households with income between HKD 10,000 - HKD 30,000 in 1000s	float	50,233	19,852	0	89,600
residents_high_income	Districtwise average of the number of households with income greater than HKD 30,000 in 1000s	float	81,510	27,364	0	127,700
median_income	Districtwise average of the median monthly household income	float	31,048	4,921	24,000	42,600
service_NT	Flag GMBs that operates regularly at a specific time during the night	bool	0.02	0.13	0	1
service_R	Flag GMBs that operates	bool	0.56	0.50	0	1

	regularly					
service_T	Flag GMBs that operates at a specific time	bool	0.38	0.49	0	1

Appendix B

Full Feature Covariance Heatmap


Correlation Matrix of Features
distance_m
journey_time 0.91
route_sequence 0.02 0.04
residential_area
industrial_area 0.01 0.02 -0.04
commercial_area 0.03 0.06 0.01 -0.03
median_income 0.02 0.02 0.01 -0.14 0.04
district_area 0.11 0.03 -0.04 0.05 -0.81 0.08
population_density -0.13 -0.07 0.04 0.02 0.45 -0.48 -0.72
residents_low_income -0.00 -0.05 -0.03 0.14 0.04 -0.66 0.00 0.44
residents_mid_income 0.00 -0.04 -0.04 0.15 -0.04 -0.68 0.09 0.36 0.98
residents_high_income 0.02 -0.07 -0.04 0.13 -0.13 -0.20 0.30 0.05 0.81 0.80
bus_stops_count 0.04 -0.05 -0.03 0.17 -0.08 0.01 0.38 -0.26 0.38 0.38 0.50
service_NT 0.09 0.11 -0.02 0.01 0.05 0.05 -0.04 -0.00 -0.06 -0.06 -0.04 -0.04
service_R -0.12 -0.08 -0.10 -0.01 -0.02 -0.12 -0.02 0.07 0.14 0.14 0.09 -0.04 -0.15
service_T -0.07 -0.10 0.13 0.00 -0.03 0.12 0.06 -0.10 -0.13 -0.13 -0.08 0.08 -0.10 -0.89
distance_m
journey_time
route_sequence
residential_area
industrial_area
commercial_area
median_income
district_area
population_density
residents_low_income
residents_mid_income
residents_high_income
bus_stops_count
service_NT
service_R
service_T