Analyzing the Relationship between Fuel Prices and Public Train Usage
Assignment 3 | AR5959D

Chan Junhao(A0204255N)

Chen Xinyang(A0283777J)

Wen Ge(A0283782R)

Xu Binyao(A0289644L)

Contents

1. Introduction

As a densely populated and highly urbanized country, Singapore's robust public transport system plays a pivotal role in maintaining the city-state's infrastructure. Trains, as the foremost mode of rail transit, are instrumental in reducing urban traffic congestion and enhancing the overall efficiency of daily commutes. Data from 2020 to 2024 indicates that petrol prices in Singapore have been both volatile and on an upward trajectory. Although this data is not directly linked to the usage of public transport, there is a reasonable assumption that the rising costs of fuel may increase the operational costs of owning a private vehicle, thus making public transport a more appealing option for daily commuters. This project sets out to investigate these assumptions further and aims to rigorously analyze the correlation between fuel prices and the adoption of public transportation through the development of comprehensive statistical models.

2. Literature Review

Extensive scholarly work has focused on the dynamics between fuel price changes and public transportation preferences. Notably, a detailed study conducted in France by Delsaut in 2014 assessed how fuel price fluctuations influence the demand for both road and rail transport. This study highlighted that a 10% increase in fuel prices typically results in a short-term 1.4% decrease in road traffic, which potentially expands to a 2.8% reduction in the longer term, consequently enhancing rail traffic volumes. Similarly, a 2021 study by Chen in China investigated the reaction of domestic gasoline price variations on consumer behavior from the perspective of behavioral economics. This research indicated that an increase in gasoline prices generally compels car owners to lessen their driving frequency and shift towards more economical and sustainable modes of transport, such as subways and buses. Collectively, these studies underscore that fluctuations in fuel prices significantly affect public transportation choices, though more research is needed to specifically quantify this impact within the context of Singapore.

3. Research Question

The central inquiry of this research is: **"How do changes in fuel prices influence public train ridership in Singapore?"** Evidence from prior research indicates that shifts in fuel prices can meaningfully influence the utilization rates of public transport systems. This study posits that individuals might prefer using cars for convenience and long-distance travel under normal circumstances but may shift towards more economical public transport options like trains when fuel costs become prohibitive. This research specifically concentrates on examining trains as the preferred public transport mode and endeavors to develop a sophisticated predictive model that can explore and establish the relationship between fuel price fluctuations and train ridership. The goal is to generate a model that not only aids in understanding current usage patterns but also serves as a strategic forecasting tool for transport operators such as SMRT Corporation in Singapore, potentially guiding future infrastructure developments and service enhancements.

4. Data Sourcing

In order to thoroughly investigate the posed research question, it was crucial to obtain extensive, accurate data on public train usage and fuel prices within Singapore. This was accomplished by accessing detailed datasets from the Land Transport Authority's DataMall API and Singstat.gov, which are authoritative sources that provide up-to-date and relevant data for such analyses. These sources were meticulously selected to ensure the reliability and applicability of the data for this project, with references included in the documentation for verification and further exploration.

5. Data Cleaning

The data analysis process began with a meticulous cleaning of the dataset to remove any discrepancies such as missing values or duplicates. This crucial phase ensures the data's accuracy and reliability, which is essential for conducting effective statistical analysis and modeling. Effective data cleaning not only enhances the quality and usability of the data but also significantly boosts the precision of the findings derived from the analysis. By establishing a solid data foundation, this step aids in paving the way for more detailed and reliable analysis and modeling efforts, which are crucial for drawing accurate conclusions and making informed decisions based on the study's outcomes.

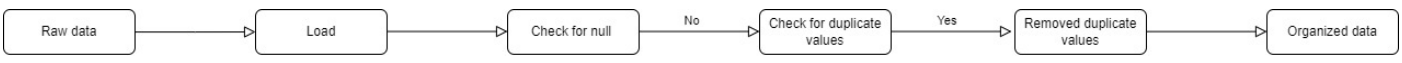

Fig.1 Data Cleaning Process Illustrated

6. Exploratory Data Analysis (EDA)
6.1 Petrol Prices Trend Analysis

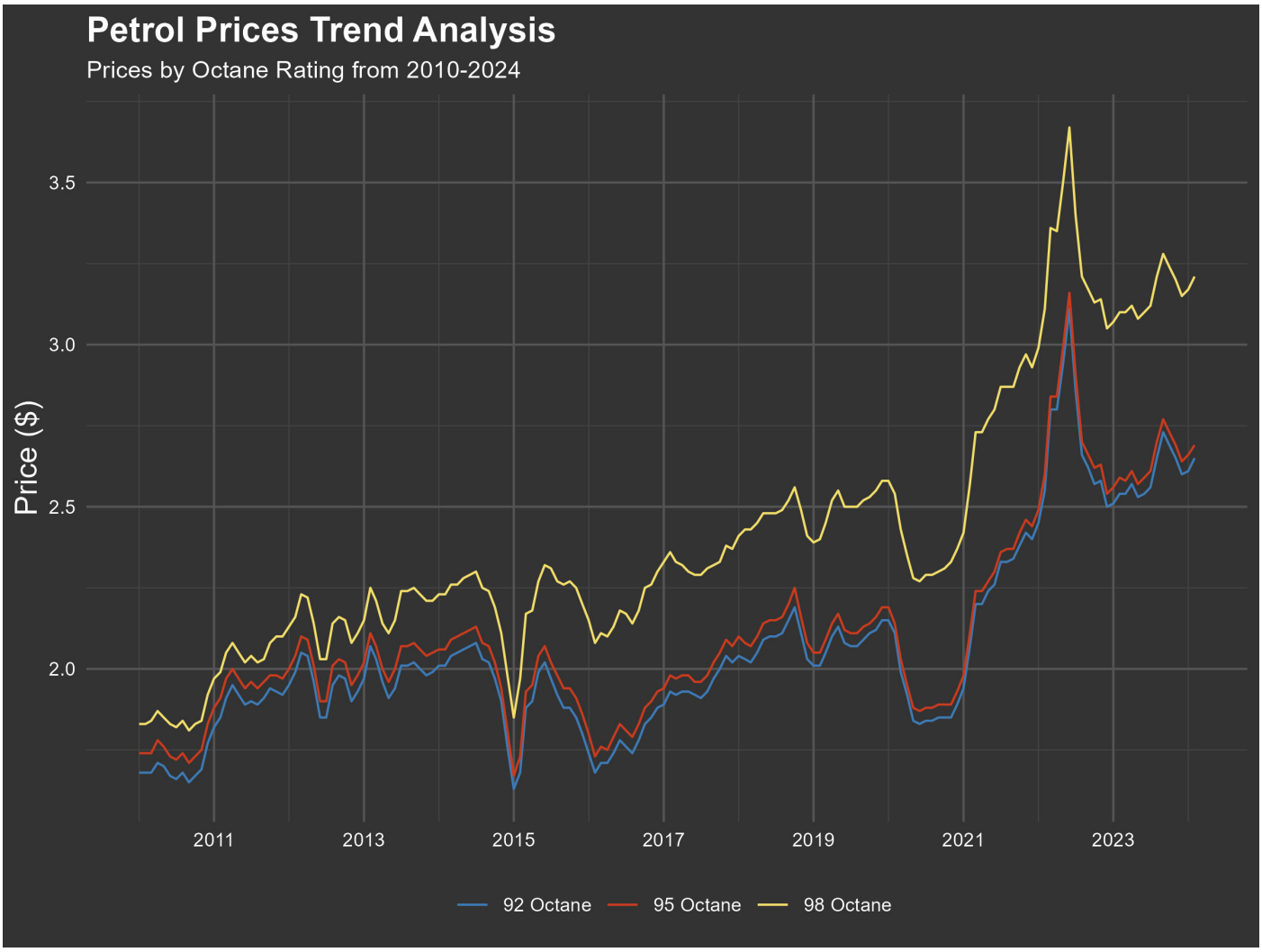

Fig.2 Trend line of monthly fuel price for Petrol (92, 95 and 98 Octane)

Observation:
1. The relative price difference between the 3 fuel types generally remains the same, with 92 being the cheapest and 98 being the most expensive.

2. The prices of the three types of petrol show an increasing overall trend, with a sudden drop in 2015.

3. The price peak was in September 2023, and has declined since then.

Conclusion:
The prices of petrol of 92, 95 and 98 octanes show an increasing trend between 2010 to 2024.

6.2 Time Series Decomposition Analysis of Petrol Prices

To gain insights into the underlying patterns of petrol prices for 92, 95, and 98 octanes, a time series decomposition analysis was conducted, breaking the data into four main components: data, trend, seasonal, and remainder. The data component represents the original recorded data, serving as the baseline for further analysis. The trend component smooths out short-term fluctuations to reveal the long-term movements in petrol prices, while the seasonal component captures predictable annual patterns related to changes in demand during different seasons. The remainder component accounts for irregularities not explained by trend or seasonal effects, often due to unexpected market events. This comprehensive analysis aids in isolating the various factors influencing petrol prices, enabling more accurate forecasting and strategic planning.

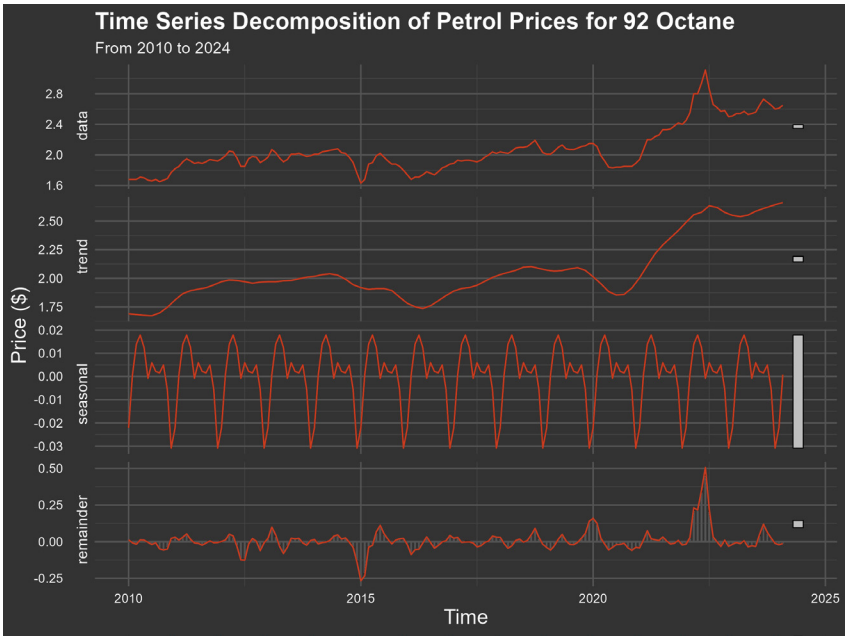

Fig.3 Seasonal decompositions of petrol price of 92 octane

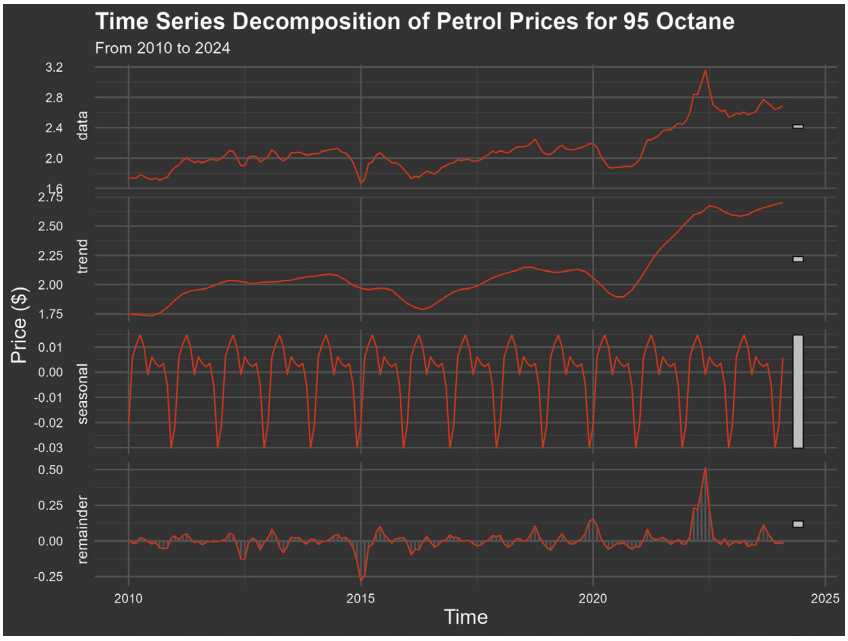

Fig.4 Seasonal decompositions of petrol price of 95 octane

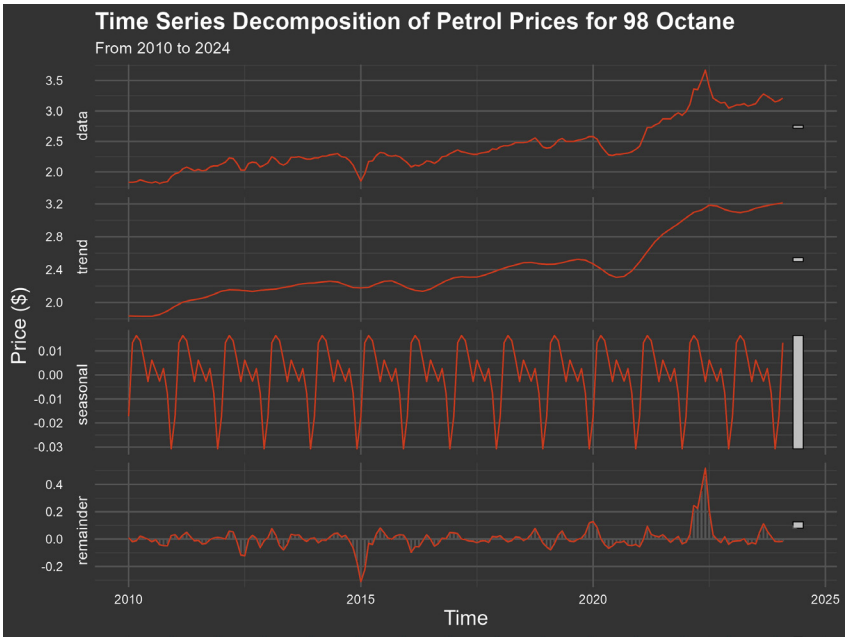

Fig.5 Seasonal decompositions of petrol price of 98 octane

Observation:
1.All fuel types show an increasing trend in prices from 2010 – 2024.
2.All fuel types show a consistent seasonal yearly cycle, where the price peaks in January and decreases towards the end of the year.
3.2 sharp residuals are identified in 2015 and 2023 for all fuel types, perhaps due to a strong external factor which influenced the change in fuel prices.

Conclusion:
The identified trend line and consistent seasonal cycle could be used for possible feature engineering or time series forecasting.

6.3 Passenger Traffic Analysis

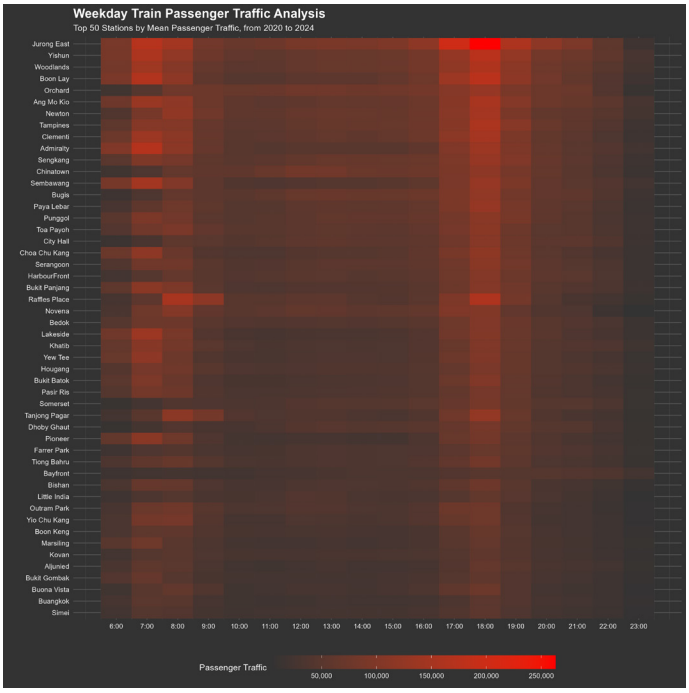

Fig.6 Analysis plot of train usage of top 50 train stations on weekdays in Singapore, from 2020-2024

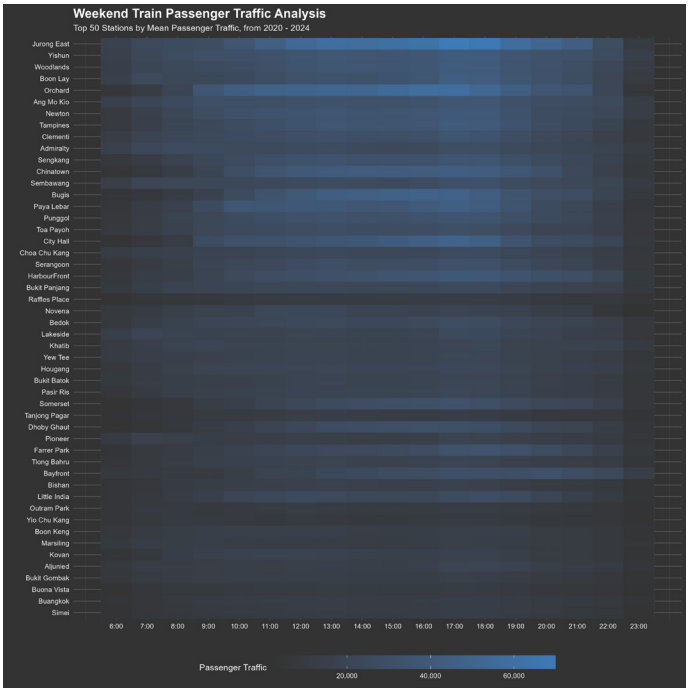

Fig.7 Analysis plot of train usage of top 50 train stations on weekends in Singapore, from 2020-2024

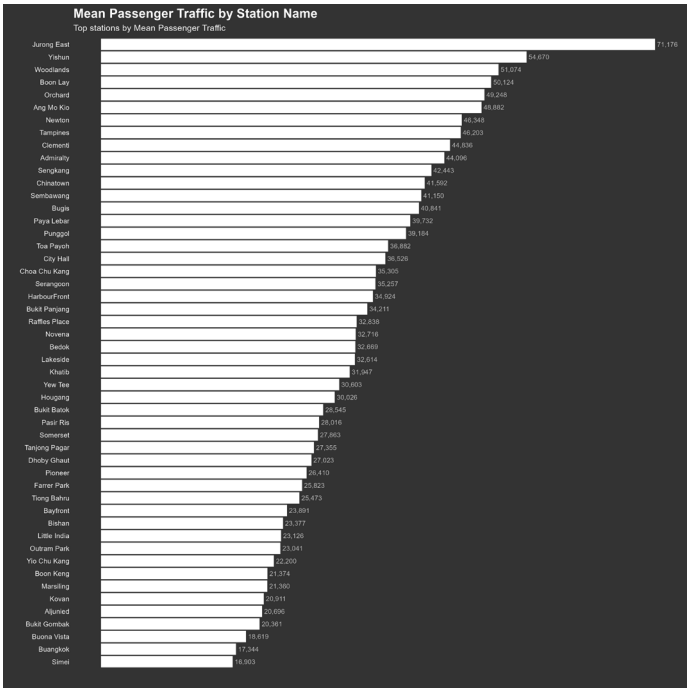

Fig.8 Analysis plot of mean passenger traffic for each station, from 2020-2024

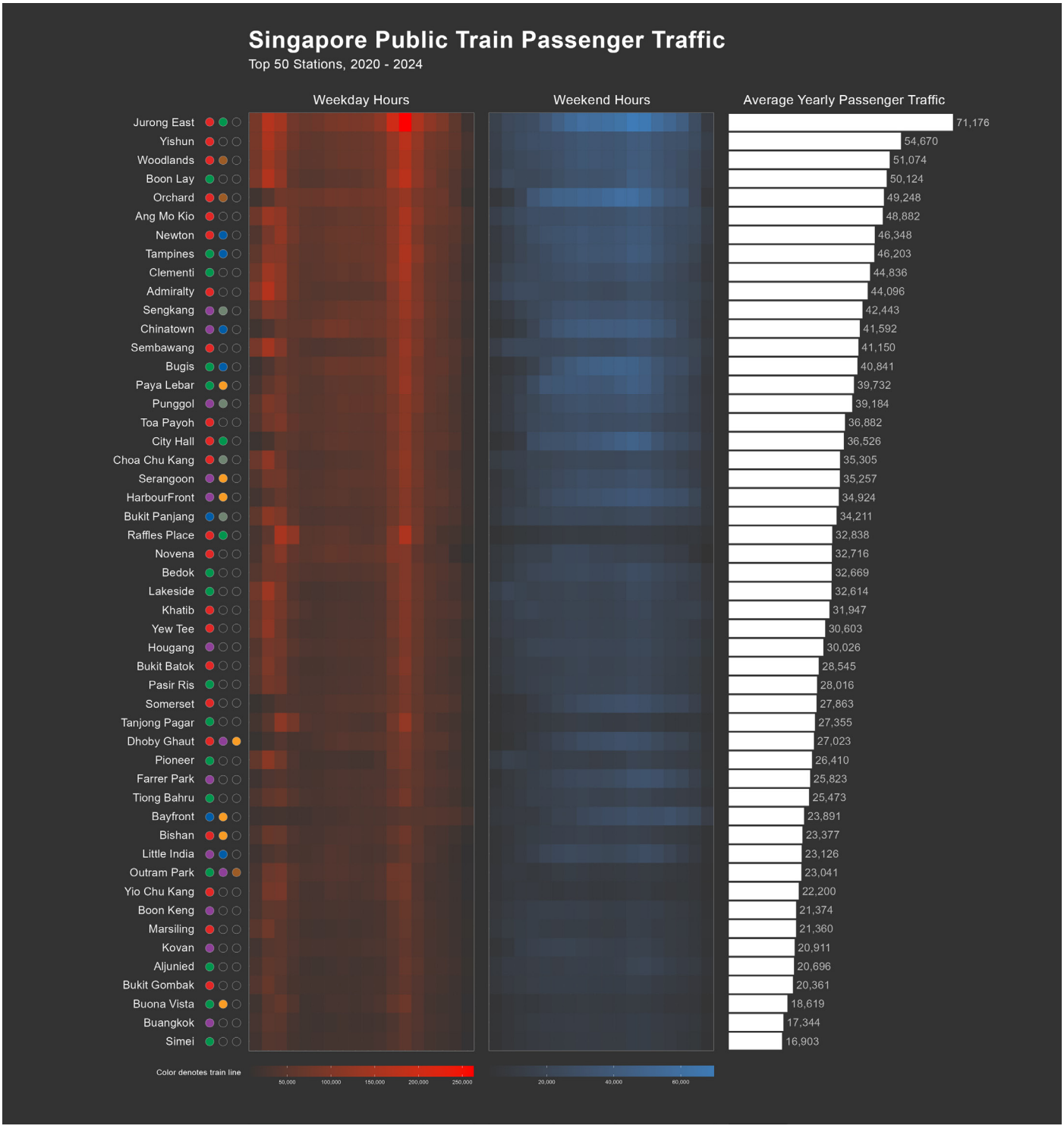

Fig.9 Combined analyis plot of train station usage

Observation:

1.Jurong East, Yishun and Woodlands are the top three stations with the highest mean passenger traffic. Jurong East has 30.2% more than its 2nd place rival Yishun.

2.The peak hours for weekdays are 7-8am and 5-7pm.

3.The peak hours for weekends are generally in the middle of the day. The range is wide from 10.30am to 6.30pm.

4.There are 22 MRT interchangeable stations in the top 50 high passenger traffic stations, especially at stations on the North South Line (red line).

Conclusion:

Jurong East station has the highest passenger traffic.

6.4 Correlation Analysis Between Petrol Price And Passenger Traffic

A Pearson's correlation test was conducted to ascertain whether a correlation exists between petrol prices and passenger traffic.

Petrol Type	Correlation Coefficient	P-value
98 Octane Petrol	0.35	0.01267
95 Octane Petrol	0.36	0.01032
92 Octane Petrol	0.36	0.01007

Table.1 Correlation coefficient between petrol price and passenger traffic

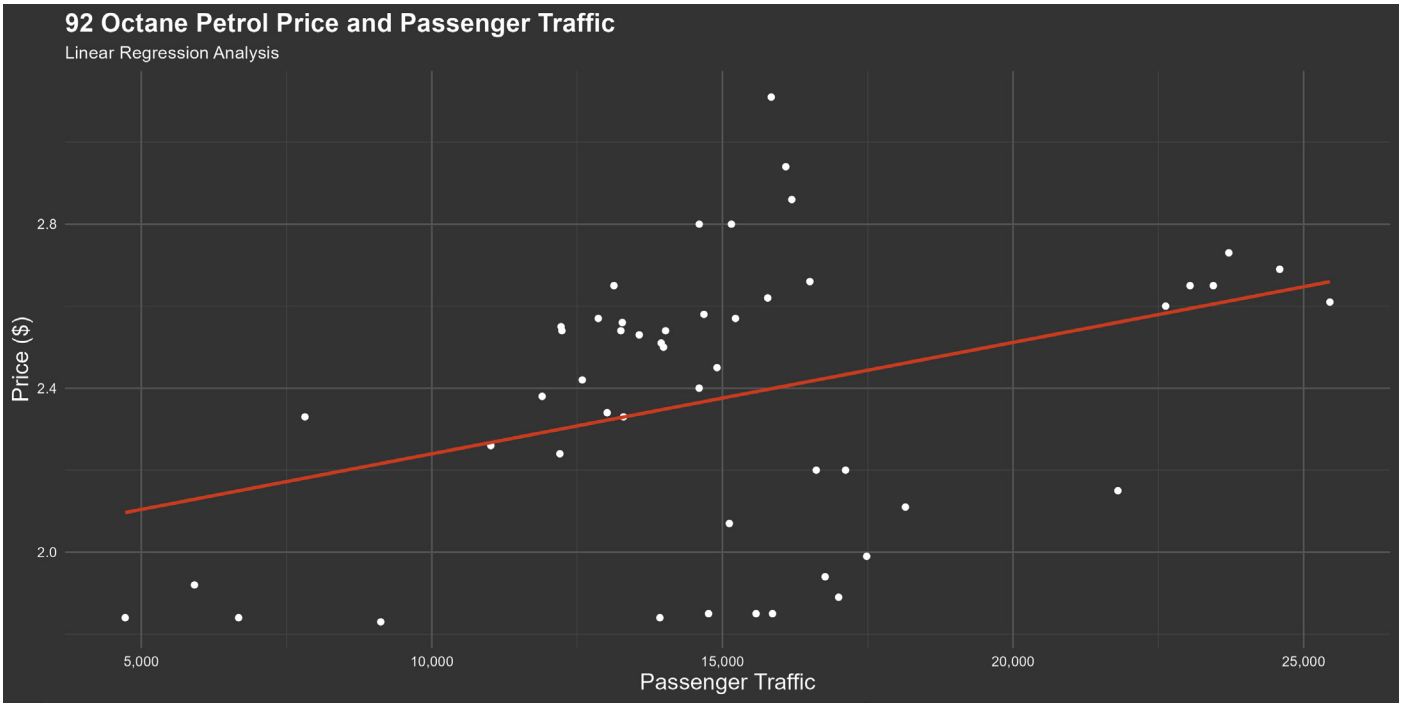

Fig.10 Correlation analysis between petrol price of 92 octane and passenger traffic

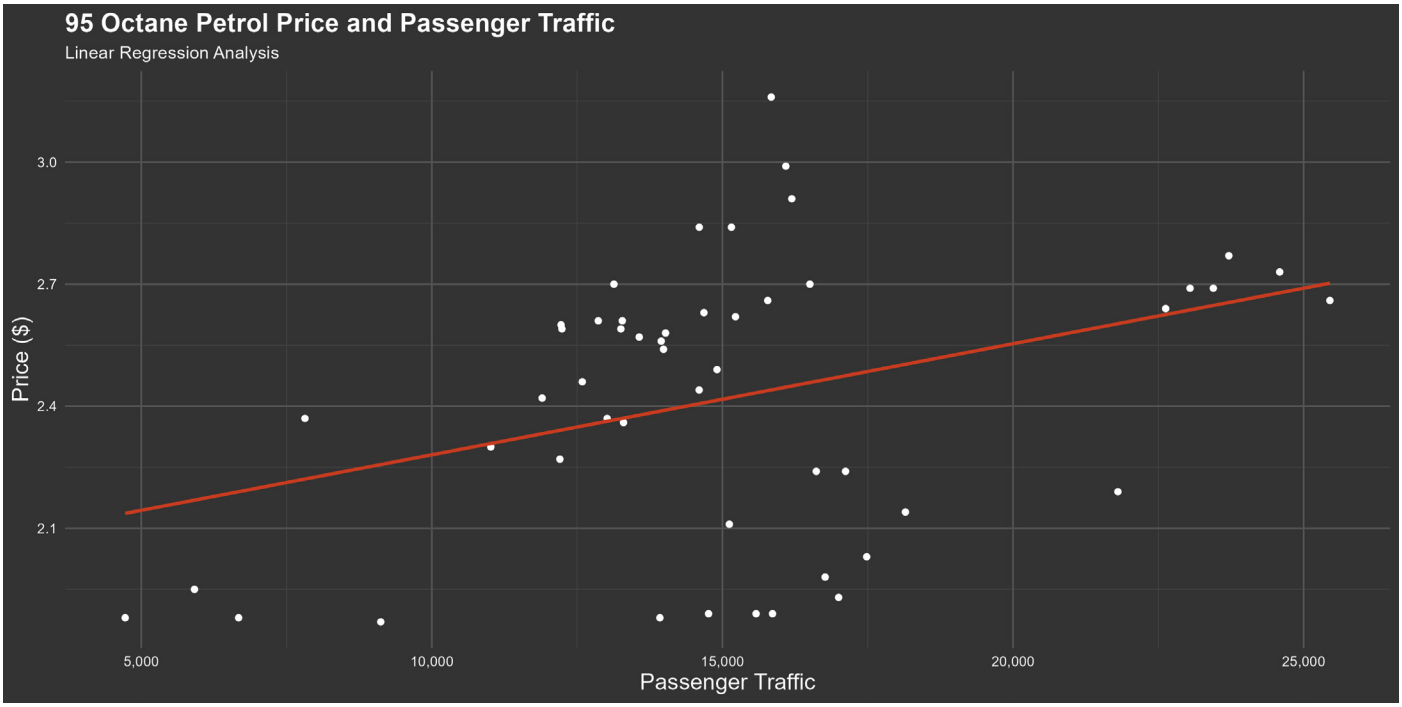

Fig.11 Correlation analysis between petrol price of 92 octane and passenger traffic

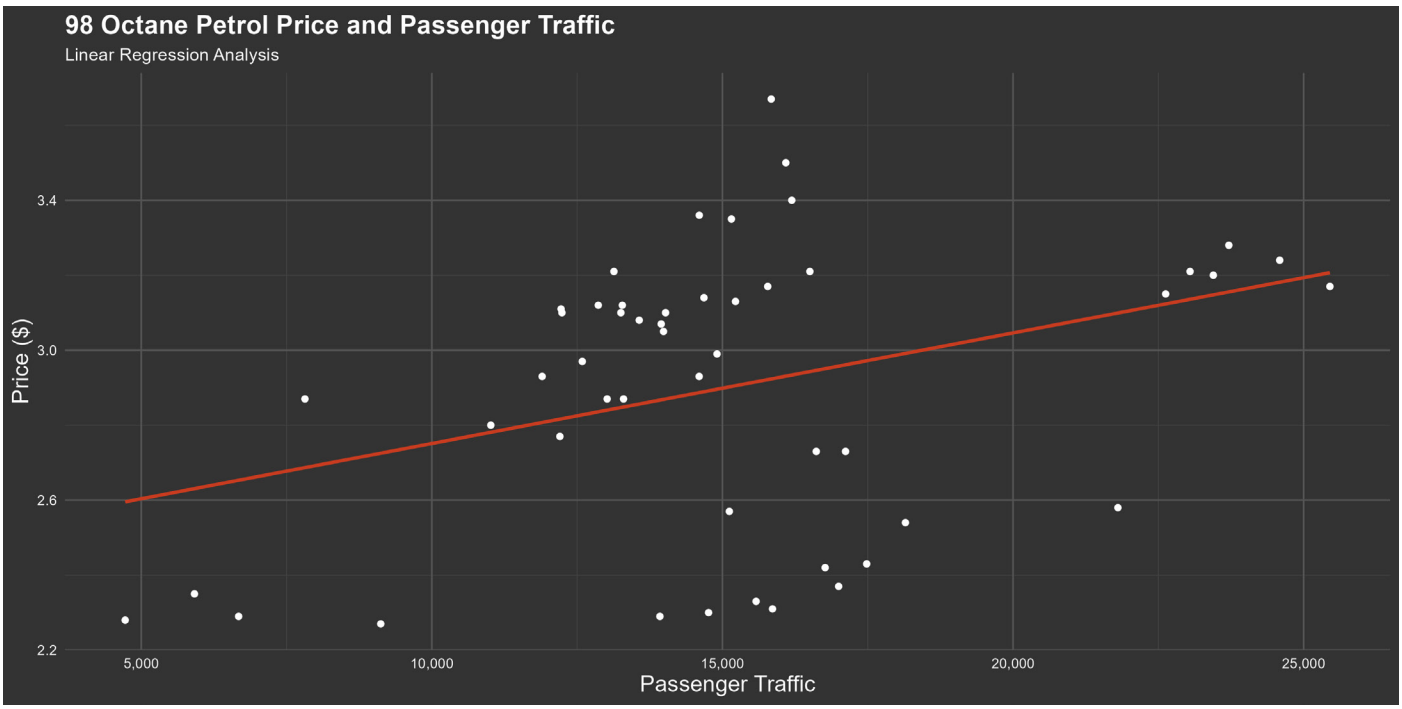

Fig.12 Correlation analysis between petrol price of 92 octane and passenger traffic

Conclusion:

1. The analysis has demonstrated a significant positive correlation between petrol prices and train passenger traffic in Singapore. This finding suggests that as petrol prices increase, more individuals opt to use the train system, likely due to the rising costs of operating private vehicles. However, it is crucial to note that while the correlation is significant, it does not inherently imply causation. Additional factors may influence train usage rates, and these must be considered to fully understand the dynamics at play.

2. When examining the correlation coefficients between the petrol prices of three different octane ratings (92, 95, and 98) and train passenger traffic, it is observed that the coefficients are strikingly similar across all types. This consistency supports the reliability of the findings and indicates a uniform response in passenger behavior regardless of the slight variations in fuel price changes among the different petrol types.

3. To further investigate whether fuel prices can directly influence passenger traffic, a more robust analytical approach will be employed. We plan to utilize machine learning (ML) techniques to test this hypothesis. By applying ML models, we can analyze complex datasets more deeply and identify patterns and relationships that are not immediately apparent through traditional statistical methods. This future work aims to substantiate or challenge our preliminary conclusions and provide a clearer understanding of the factors driving changes in public transportation usage..

7. EDA Summary

In the table below, it outlines the insights from the EDA and its subsequent effect on the model development.

SN	Analysis	Insight	Effect on Model Development
1	Correlation Scatterplot	Significant positive correlation between passenger traffic and fuel prices for all fuel types	Suggests a relationship can be modeled and a prediction model developed
2	Time Series Decomposition	For all petrol types: 1. Increasing trend line observed 2. Consistent seasonal patterns	Could be used for possible feature engineering
3	Heatmap Analysis of Passenger Traffic by Station	Jurong East station has the highest footfall	Develop a model for Jurong East station to test hypothesis

Table.2 EDA Summary

8. Model Selection and Feature Engineering

8.1 Model Selection

The model selection was between a regression model (Random Forest) and a forecasting model (ARIMA). Random forest regression was selected as the most appropriate model due the following reasons:

Does not need differencing.

The dataset was evaluated for stationarity and due to the high trend line and seasonal as seen in the time series decomposition, 2 levels of differencing would have been needed to remove the trends and seasonality. Thus, this would have lost valuable information in the training data.

Able to calculate feature importance

Random forest allows us to calculate feature importance, and thus we can evaluate if certain fuel types or all types affect train usage.

8.2 Feature Engineering

The table below outlines the feature engineering done from raw data.

Feature	Reasoning	Parameters	Raw Data used
Rolling mean for each fuel type	Reduces noises, focuses on the trend line observed in time series	k = 3	Last 3 months prices of each fuel type
Lag	Incorporate effect of past prices on future prices	Lag month = 1	Fuel prices
Seasonality	Seasonality trends exists as seen in the time series	Month and Year	Original Data column

Table.3 Feature Engineering Summary

In summary the model will be trained with 11 features (3 existing and 8 engineered) and 1 target variable.

9. Model Training

The model was trained on 52 training data (Jan 2020 – 2024) with one test data withheld (Feb 2024). Due to the limited number of training data, cross validation was used to split the evaluate the training data, allowing for a more robust estimation of the model's performance and training on the entire dataset. The number of folds (k) was set to 5 and number of trees used was 500.

10. Model Evaluation
10.1 R2 and MAE

mtry	R2	Mean Absolute Error (MAE)
2	0.35	474,087
6	0.37	464,309
11	0.37	455,684

Table.4 Model Tuning Summary

The table above shows the tuning summary evaluation metrics of the model of which mtry 11 was used as the final model. The R2 value of 0.37 tells us the model generally performs decently, but there is room for improvement. As for MAE, no conclusion can be drawn without comparing it to the range / mean of the target variable.

10.2 Feature Importance

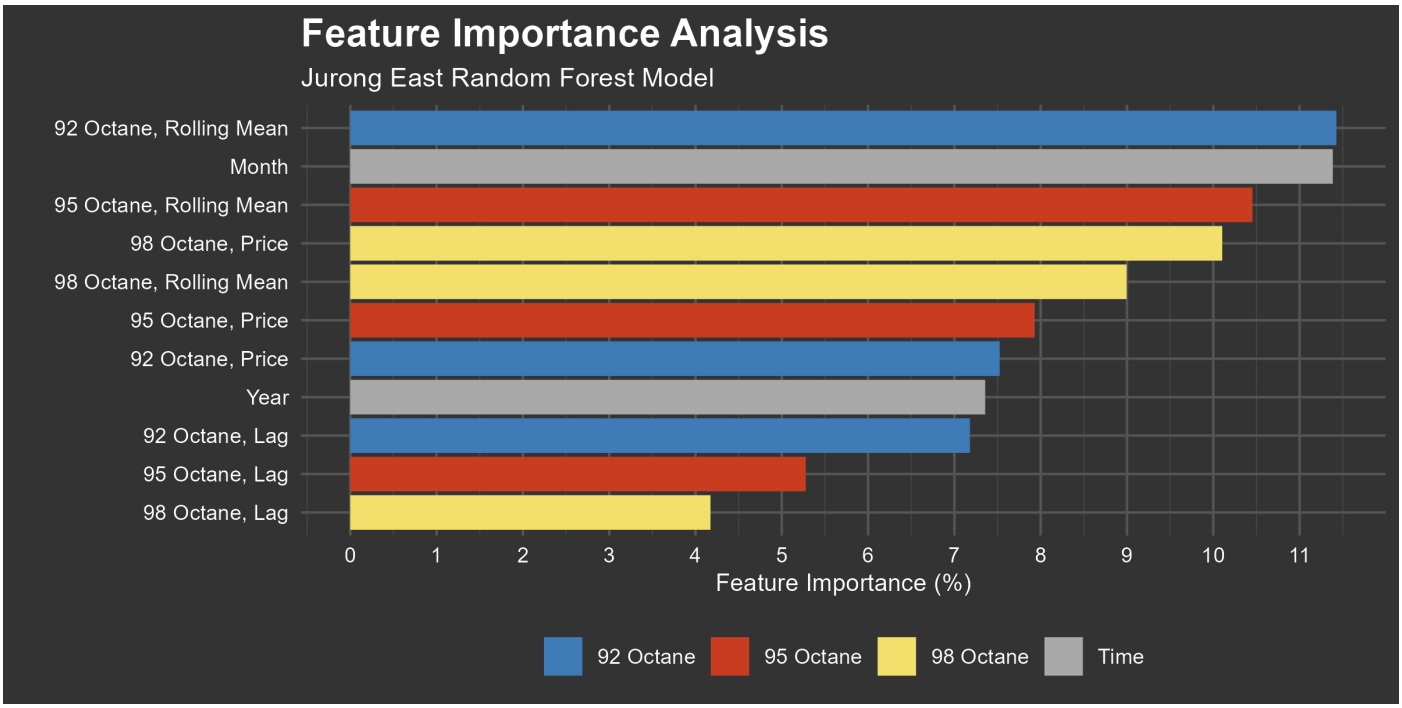

Fig.13 Feature Importance Analysis of the Jurong East Random Forest Model

Observation:

1. No specific fuel type consistently influences train traffic predictions strongly, indicating a nuanced impact of fuel prices on train usage.

2. Rolling mean features outperform simple price and lag features in predictive models, suggesting that averaged data provides a clearer trend.

3. The significant influence of the month on train traffic predictions highlights underlying seasonality, possibly independent of fuel price fluctuations.

Conclusion:

For a more precise model, it would be beneficial to develop separate models for each fuel type. Initial results suggest that a model using 98 Octane might perform better than those using other fuel types. This focused approach could reveal more specific correlations between fuel types and train usage, enhancing the model's robustness and accuracy.

10.3 Test Set

In the final phase of our analysis, the model was applied to a withheld test dataset, specifically the data for February 2024, to assess its predictive accuracy. The model forecasted a train usage value of 3,532,402 for this period. When compared to the actual observed train usage, which was recorded at 3,729,500, the model exhibited a mean absolute percentage error (MAPE) of 5.28%. This level of error demonstrates that the model, despite its simplicity in focusing primarily on fuel price variables, can predict train usage in a station with commendable accuracy, hinting at the effectiveness of integrating fuel price data in predictive models for public transportation usage.

11. Conclusion

This study establishes a clear relationship between fluctuations in fuel prices and the utilization of public train services in Singapore. We began with the hypothesis that an uptick in fuel prices would compel car owners to switch to public transportation within the same month. This is somewhat supported by an observed average correlation coefficient of 0.35 between these variables. While this figure does not suggest an overwhelmingly strong correlation, it is notable because it quantifies the impact of a single, albeit significant, variable—fuel prices—on train usage. This foundational research paves the way for the development of a base predictive model, which can be incrementally refined by incorporating more complex variables such as economic indicators like GDP, as well as environmental factors such as seasonal weather variations. These enhancements could transform the model into a more dynamic and predictive tool, empowering train operators to adjust schedules proactively, thereby minimizing wait times for commuters and enhancing the overall efficiency and satisfaction of public transport users.

The potential benefits of such an advanced predictive model are substantial. Train operators could leverage accurate forecasts to optimize service frequencies and timings, significantly enhancing operational efficiency and customer satisfaction. Furthermore, an effective model would enable better management of resources during peak and off-peak times, ensuring that service provision is closely aligned with passenger demand.

12. Limitations

1. Our findings indicate certain correlations with changes in fuel prices; however, it's important to recognize that train usage is multifaceted and influenced by a variety of factors beyond our primary focus, such as seasonal weather conditions or special events, which can also significantly affect public transportation patterns.

2. The study assumes that the shift from personal cars to public transport is predominantly influenced by changes in fuel prices. This assumption overlooks potential shifts toward other modes of private transportation, such as ridesharing platforms like Grab, which although affected by fuel prices, are not captured within the public transport usage data. This could skew the perceived impact of fuel prices on the broader transportation ecosystem.

3. We have based our analysis on the premise that car ownership in Singapore is primarily for facilitating long-distance travel. Under this assumption, when consumers choose to switch from personal cars to public transport due to increasing fuel prices, they are presumed to opt for trains over buses, given trains' efficiency in covering longer distances quickly. This assumption led us to focus solely on train data, potentially overlooking how buses might also serve as a viable alternative for some consumers affected by rising fuel costs.

13. References

Adhikari, R., & Agrawal, R. K. (2013, February 26). An introductory study on time series modeling and forecasting. arXiv.org. https://arxiv.org/abs/1302.6613

Bates, S., Hastie, T., & Tibshirani, R. (2022, July 18). Cross-validation: What does it estimate and how well does it do it?. arXiv.org. https://arxiv.org/abs/2104.00673

Delsaut, M. (2014). The effect of fuel price on demands for road and rail travel: An application to the French case. Transportation Research Procedia, 1(1), 177–187. https://doi.org/10.1016/j.trpro.2014.07.018

Fuel Data
https://tablebuilder.singstat.gov.sg/table/TS/M212891

Singh, A. (2023, July 28). 6 powerful feature engineering techniques for time series data (using python). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/12/6-powerful-feature-engineering-techniques-time-series/

(PDF) over-differencing and forecasting with non-stationary time series data. (n.d.-a). https://www.researchgate.net/publication/332621968_Over-Differencing_and_Forecasting_with_Non-Stationary_Time_Series_Data

(PDF) time series modelling and decomposition. (n.d.-b). https://www.researchgate.net/publication/307663962_Time_Series_Modelling_and_Decomposition

The study of the impact of domestic gasoline price changes on consumers' behaviors. CNKI. (n.d.). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202201&filename=1021836673.nh

Transport Data
https://datamall.lta.gov.sg/content/datamall/en/dynamic-data.html