IDVE Assignment One

Muhammed Muaaz Dawood (2425639)
Mujammil Mohsin Gulam Sakhidas (2436109)
Sayfullah Jumoorty (2430888)

1 Question 1

1.1 Data Fields

1. Percentage of date fueled entries that are not proper dates: 11.66%
2. Found in Jupyter Notebook.
3. Found in Jupyter Notebook
4. Found in Jupyter Notebook
5. There were very few fuel-ups recorded in the early years, but the numbers started growing quickly after 2012. The biggest jump happened in the last few years, with way more fuel-ups recorded between 2020 and 2024 than ever before. This could be because more people started using the service that tracks these fuel-ups, or because of changes in how often people need to fill their tanks.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

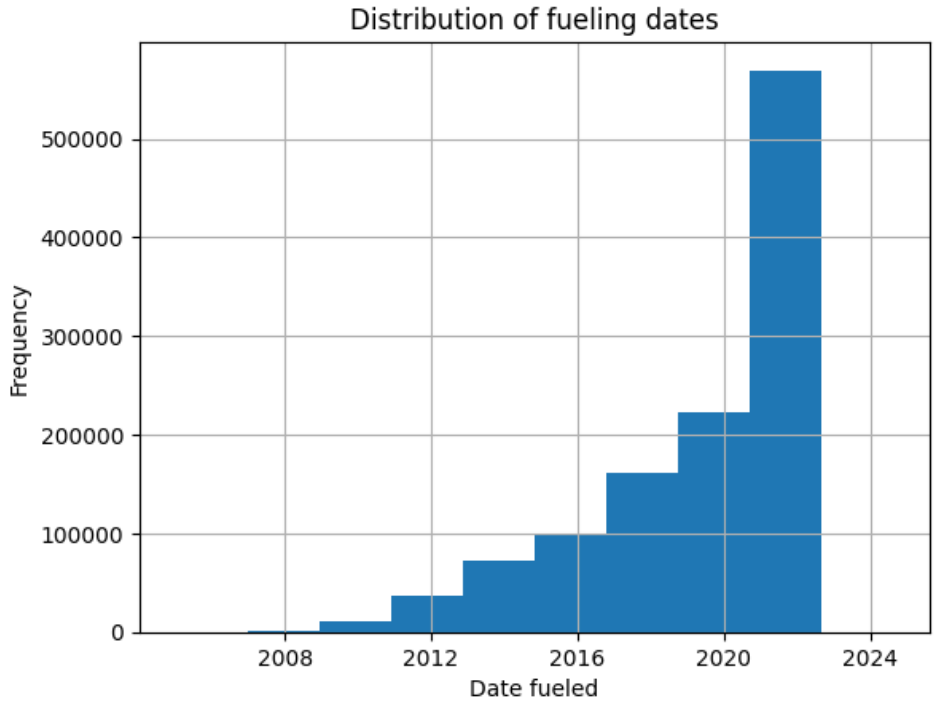

Figure 1: Distribution of fueling dates

1.2 Numeric fields

1.
   - Percentage of missing gallons entries: 6.32%
   - Percentage of missing miles entries: 87.56%
   - Percentage of missing odometer entries: 12.68%

2. Found in Jupyter Notebook

3. Found in Jupyter Notebook

4.
   - Distribution of miles: Normal distribution with slight right skew, peaking at 250-300 miles. Range: 0-700 miles. Reflects typical driving patterns, with most trips falling in a moderate range and fewer long-distance journeys.
   - Distribution of gallons: Right-skewed normal distribution, peaking at 10-15 gallons. Range: 0-30 gallons. Indicates common refueling habits and variety of fuel tank sizes across different vehicle types.
   - Distribution of MPG: The MPG distribution is unimodal, peaking around 20-22 MPG and ranging from 0-60 MPG. It shows a right skew, with a long tail towards higher efficiencies. This pattern reveals a diverse vehicle fleet centered on moderate fuel efficiency, with some highly efficient outliers. The spread likely represents a mix of vehicle types, from larger or less efficient models to more fuel-efficient compact cars or hybrids.

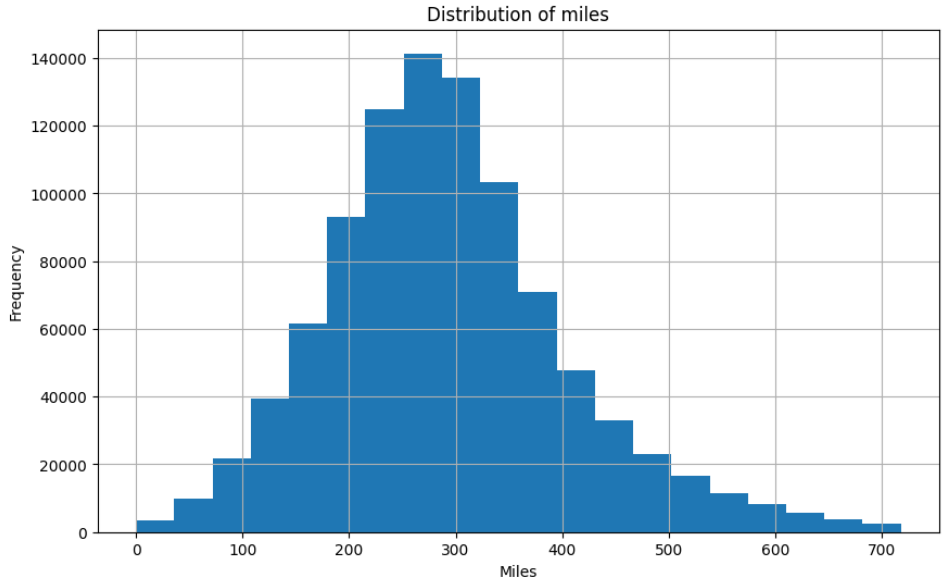

Figure 2: Distribution of Miles

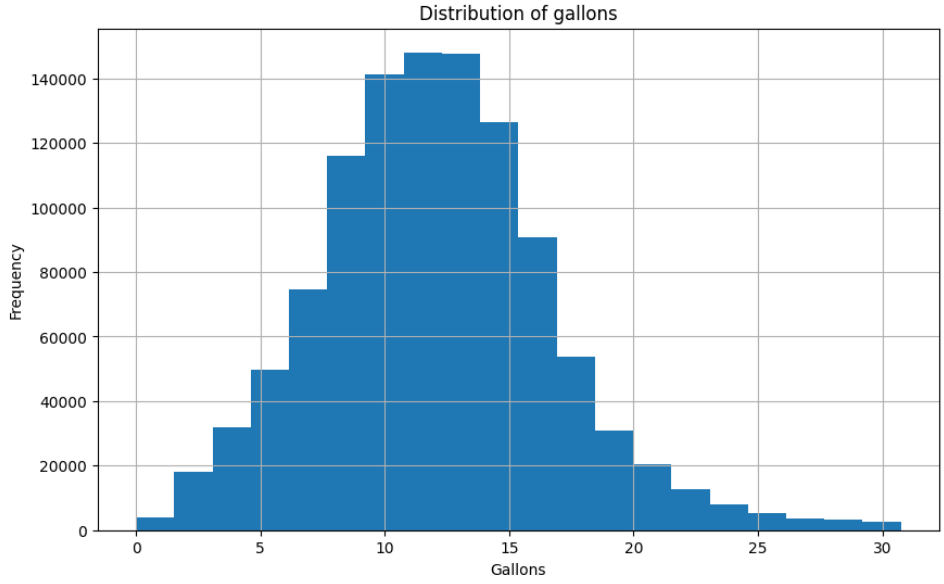

Figure 3: Distribution of Gallons

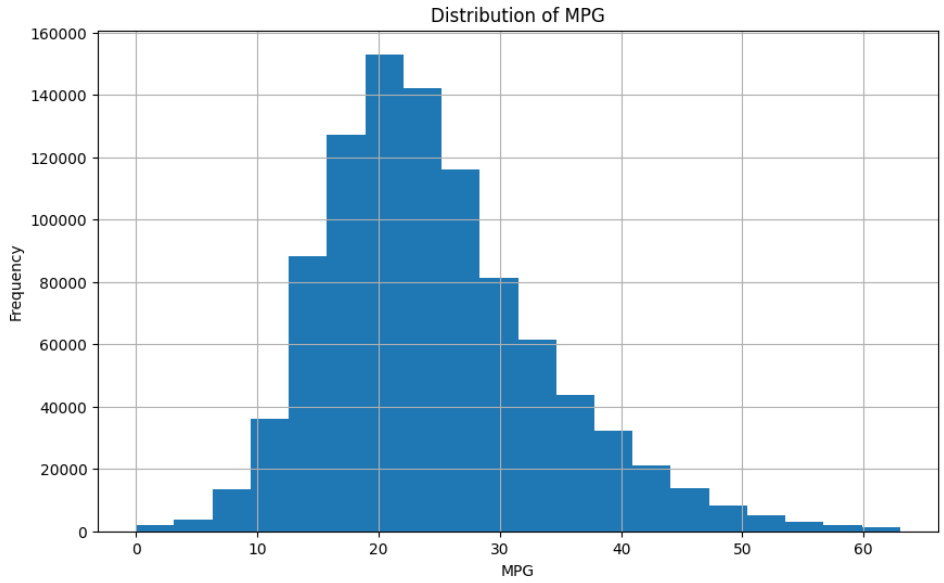

Figure 4: Distribution of miles per gallons

5. The statistical descriptions generally align with expectations for fuel usage data. The date range, mean gallons, and typical MPG values appear sensible. However, there are notable outliers that raise concerns about data quality. Extreme maximum values for gallons (984.71), MPG (165,900), and miles (23,238.4) are unrealistic and likely errors. The mean values and quartiles for gallons, MPG, and miles driven between fill-ups fall within reasonable ranges. The most frequent values seem plausible, except for the oddly low odometer reading of 1.

2 Feature Engineering

All the answers relating to this section can be found in the Jupyter Notebook.

3 Vehicle Exploration

1. Plot showing number of unique users per country

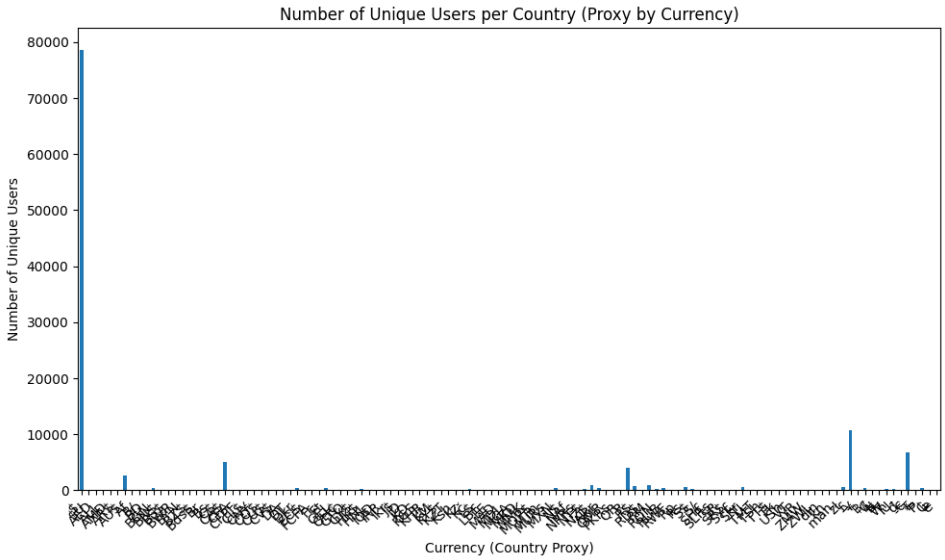

Figure 5: Plot of number of unique users per country

2. Plot of number of unique users per day

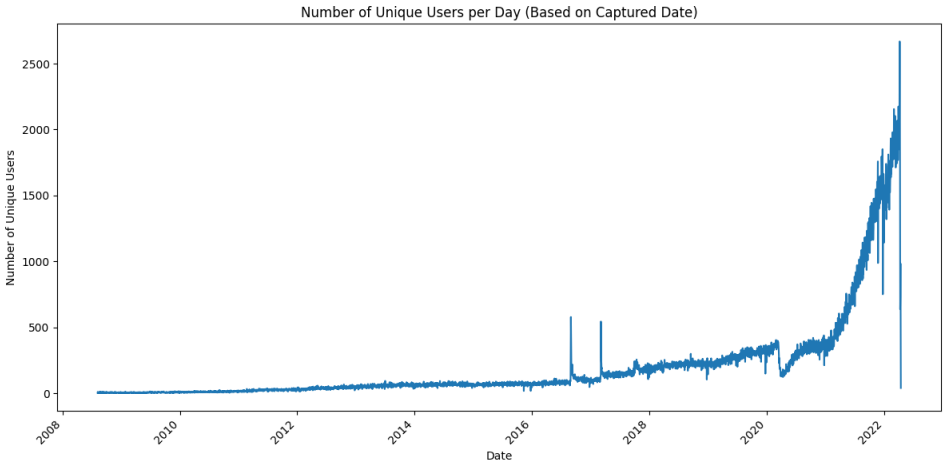

Figure 6: Plot of number of unique users per day

3. Found in Jupyter Notebook

4. The table 4 shows the the top 10 car makes and models.

Car Make	Car Model	Counts
Honda	Civic	8001
Toyota	4Runner	7758
Toyota	Corolla	7694
Ford	F-150	7644
Honda	Accord	7583
Ford	Mustang	7506
Ford	Ranger	7405
Toyota	Land Cruiser	7372
Toyota	Camry	7311
Jeep	Wrangler	7022

4 Fuel Usage

4.1 Outlier Removal

1. Top 5 Currencies by Number of Transactions:
   $: 735419
   £ : 86125
   € : 58243
   CA$: 46108
   R : 35974

2. Our outlier removal approach for the top 5 currencies uses the Isolation Forest algorithm, applied separately to each currency to account for regional differences. This method detects multivariate outliers by considering multiple features like volume, distance, and cost, effectively identifying anomalies such as currency misclassifications and data entry errors. The algorithm's sensitivity is adjustable via the contamination rate. This comprehensive approach aims to improve data quality while preserving legitimate variations in fuel purchase behavior across different currencies and regions.

3. Number of values removed after accounting for outliers: 591111

4.2 Fuel Efficiency

1. Breakdown of the average cost per litre in ZAR for January 2022:

   (a) **United States (USD)**: 14.21 ZAR
   (b) **Canada (CAD)**: 17.81 ZAR
   (c) **South Africa (ZAR)**: 18.67 ZAR
   (d) **United Kingdom (GBP)**: 30.71 ZAR
   (e) **Eurozone (EUR)**: 28.00 ZAR

   Source : https://www.oanda.com/currency-converter/

   Notable Differences and Potential Reasons

   United States (USD: 14.21 ZAR)

   - **Lowest Fuel Prices**: The average cost per litre in the US remains the lowest among the top 5 currencies. The US has generally lower fuel prices due to:
     - **Lower Taxes**: Significantly lower fuel taxes compared to European countries and even Canada.
     - **Domestic Oil Production**: Substantial domestic oil production, reducing import dependency.
     - **Efficient Distribution**: Large-scale, efficient distribution networks across the country.

Canada (CAD: 17.81 ZAR)

- **Moderate Fuel Prices**: Canada's average cost per litre is higher than the US but still moderate compared to UK and Eurozone.
  - **Provincial Variations**: Fuel prices in Canada can vary significantly between provinces due to different taxation and environmental policies.
  - **Domestic Production**: Significant domestic oil production helps moderate prices, but not to the extent seen in the US.

South Africa (ZAR: 18.67 ZAR)

- **Comparable to Canada**: South Africa's average cost per litre is similar to Canada's. This is expected due to several factors:
  - **Import Dependency**: South Africa's reliance on oil imports makes it vulnerable to international price fluctuations and exchange rates.
  - **Government Levies**: South Africa has significant government taxes and levies that are added to the base fuel price, contributing to the relatively high price.
  - **Exchange Rate Volatility** :The strength of the Rand against the US Dollar can heavily influence fuel prices.

United Kingdom (GBP: 30.71 ZAR)

- **Highest Fuel Prices**: The UK maintains the highest average cost per litre, consistent with known factors:
  - **High Fuel Taxes**: Substantial taxes on fuel significantly elevate prices.
  - **Environmental Policies**: Stringent policies aimed at reducing carbon emissions contribute to higher costs.
  - **Import Dependency**: Reliance on fuel imports exposes the UK to global price fluctuations.

Eurozone (EUR: 28.00 ZAR)

- **Second Highest Prices**: The Eurozone experiences high fuel prices, though slightly lower than the UK:
  - **High Taxation**: Many European countries impose significant fuel taxes to fund infrastructure and incentivize reduced consumption.
  - **Import Reliance**: Most Eurozone countries depend heavily on oil imports, affecting prices.
  - **Green Initiatives**: Policies promoting a shift towards renewable energy sources often involve higher fuel costs.

Summary of Differences

- **The United States** stands out with significantly lower fuel prices, followed by Canada.
- **South Africa** occupies a middle ground.
- **The UK and Eurozone** maintain substantially higher fuel prices, more than double those in the US.

Conclusion

The notable differences in fuel prices across these countries are largely driven by government taxation policies, domestic production capacities, and environmental regulations. Countries that impose higher taxes on fuel, such as the UK and those in the Eurozone, see significantly higher prices compared to countries like the US and Canada, where taxes are lower and domestic production is higher.

2. Estimated number of missed fill-ups in the dataset: 87 and the general rule for identifying missed fill-ups is to compare the change in odometer readings between consecutive fuel purchases with the reported miles driven. If the odometer difference significantly exceeds

the miles driven (typically by a predefined threshold, such as 50% higher), it suggests a potential missed fill-up. This method assumes that a large discrepancy between the odometer change and reported miles driven likely indicates an unreported fuel purchase occurred between the two recorded fill-ups. The approach requires sorting data by user and date, calculating odometer differences, and flagging instances where this difference surpasses the expected range based on reported miles driven.

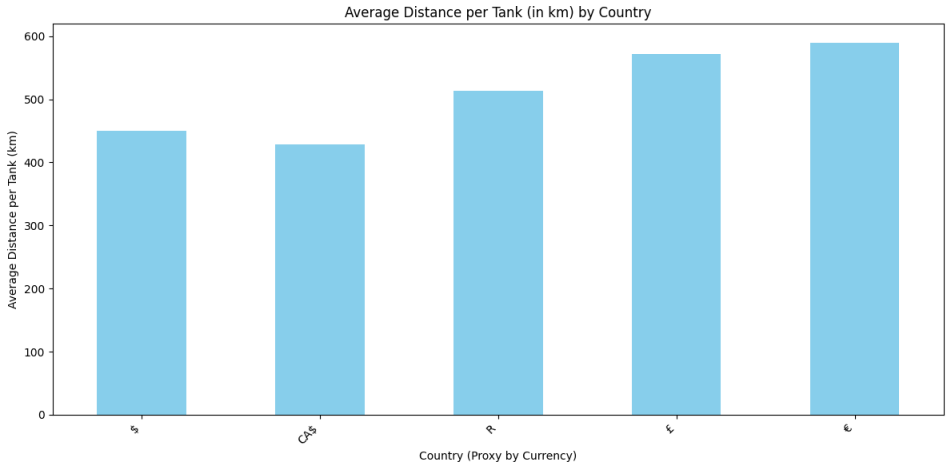

Figure 7: Plot of number of unique users per country

3. The figure 7 shows that countries using the Euro (€) have the highest average distance per tank at 589.93 km. There are several potential explanations for this:

    - **Fuel efficiency**: European countries often have stricter fuel efficiency standards and a higher proportion of smaller, more efficient vehicles. This leads to longer distances traveled on a single tank.
    - **Driving habits**: Europeans may have different driving patterns, possibly with more highway driving or longer commutes, which typically result in better fuel economy.
    - **Fuel prices**: Higher fuel prices in many European countries encourage more efficient driving habits and the use of more fuel-efficient vehicles.
    - **Infrastructure**: Well-developed highway systems in many European countries contribute to more efficient long-distance travel.
    - **Diesel popularity**: Diesel engines, which are more common in Europe, generally offer better fuel economy for long distances.
    - **Cultural factors**: There might be cultural differences in how often people refuel, with Europeans possibly waiting longer between fill-ups.
    - **Vehicle tank sizes**: European cars might have larger fuel tanks on average, allowing for longer distances between refuels.

4. If we focus on the data within the first 250 years (ignoring the outliers beyond that range), the trend suggests that newer vehicles generally drive further distances between fill-ups. The data points in the graph 8 cluster around 300 to 400 km for most newer vehicles, indicating consistent fuel efficiency. While there is some variability, it does appear that newer vehicles tend to achieve higher average distances per tank compared to older ones. Therefore, within the relevant range, the data does support the notion that newer vehicles tend to perform better in terms of distance per tank.

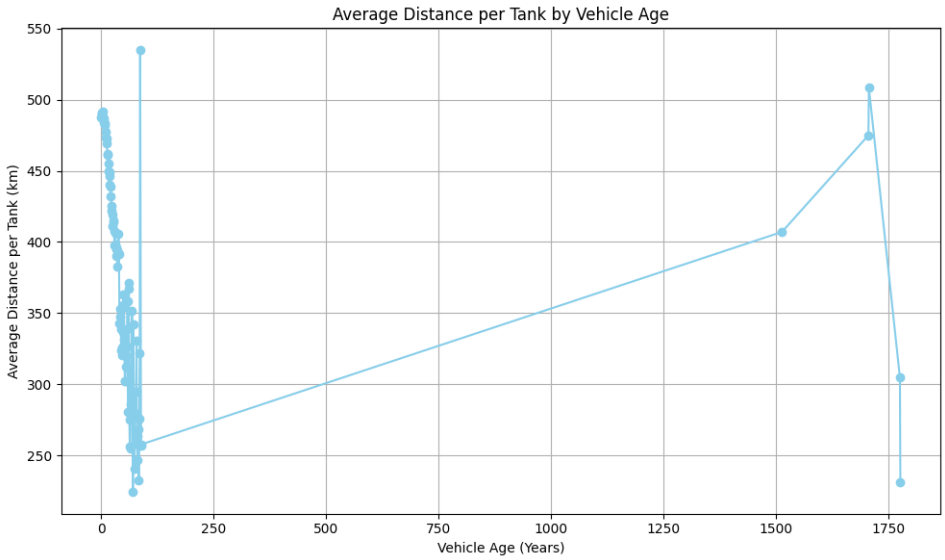

Figure 8: Average Distance per Tank by vehicle age

5.
   - **Mitsubishi Pajero**: 11.81 L/100km - This is within the expected range for an SUV, which typically consumes more fuel due to its larger engine and weight.
   - **Suzuki Jimny**: 9.10 L/100km - This is a realistic figure for a compact off-road vehicle, which balances fuel efficiency with off-road capability.
   - **Toyota Fortuner**: 10.82 L/100km - This is expected for a large SUV that prioritizes power and capability.
   - **Toyota Hilux**: 11.19 L/100km - . The Hilux, a popular pickup, has a fuel consumption that aligns with similar vehicles in its class.
   - **Volkswagen Polo**: 8.01 L/100km - The Polo, being a small and efficient hatchback, shows good fuel efficiency, as expected.

6.
   - Country: $ Vehicle: Opel Tigra, Fuel Efficiency: 5.73 L/100km
   - Country: CA$ Vehicle: Volkswagen Caddy, Fuel Efficiency: 5.43 L/100km
   - Country: R, Vehicle: Suzuki Baleno, Fuel Efficiency: 5.46 L/100km
   - Country: £, Vehicle: Mini Countryman, Fuel Efficiency: 4.64 L/100km
   - Country: €, Vehicle: Fiat Punto Evo, Fuel Efficiency: 4.72 L/100km

7. **Analysis of the Fuel Efficiency Plot for Top 5 Canadian Vehicles**

   Plot Overview:

   The plot 9 depicts the fuel efficiency (in liters per 100 kilometers) for the top 5 Canadian vehicles across different seasons: Fall, Spring, Summer, and Winter. The vehicles included in the plot are:

   (a) **Mazda 3 Sport**
   (b) **Toyota Matrix**
   (c) **Hyundai Accent**
   (d) **Honda Civic**
   (e) **Kia Rondo**

   Expectations of Seasonal Differences

   (a) **Temperature Impact**: In colder months, vehicles typically have lower fuel efficiency due to the extra energy required to warm up the engine and increased rolling resistance on snow or ice.

(b) **Air Conditioning in Summer**: Higher temperatures can also reduce fuel efficiency due to the increased use of air conditioning.

(c) **Snow in Winter**: In Canada, winter conditions include snow, which can further exacerbate the reduction in fuel efficiency due to increased rolling resistance and the additional energy required for driving in snowy conditions.

Observed Differences

- **Winter** : The plot shows an increase in fuel consumption (lower fuel efficiency) during the winter for all vehicles. This aligns with the expectation that colder weather increases fuel consumption. In Canada, the presence of snow during winter contributes to this increased fuel consumption.
- **Summer**: Fuel efficiency seems to improve for most vehicles, except for a slight decrease in one or two cases. This could be attributed to milder temperatures and fewer demands on the engine compared to winter.
- **Spring and Fall**: These seasons seem to offer more stable fuel efficiency, with values generally lying between winter and summer, though the exact trends vary by vehicle.

Variability by Vehicle

- **Kia Rondo**: This vehicle shows the largest fluctuation in fuel efficiency, with significant drops in the summer and winter.
- **Honda Civic and Hyundai Accent**: These models exhibit less fluctuation, with relatively stable performance across all seasons.
- **Toyota Matrix**: It shows a noticeable peak in fuel consumption in the spring, which is unusual compared to the others.

Conclusion

While seasonal variations in fuel efficiency are expected, the degree of fluctuation varies by vehicle model. The differences, especially in winter, are quite noticeable for all models, with some showing more pronounced changes than others. The plot confirms that fuel efficiency does change with the seasons, though the extent depends on the specific vehicle.

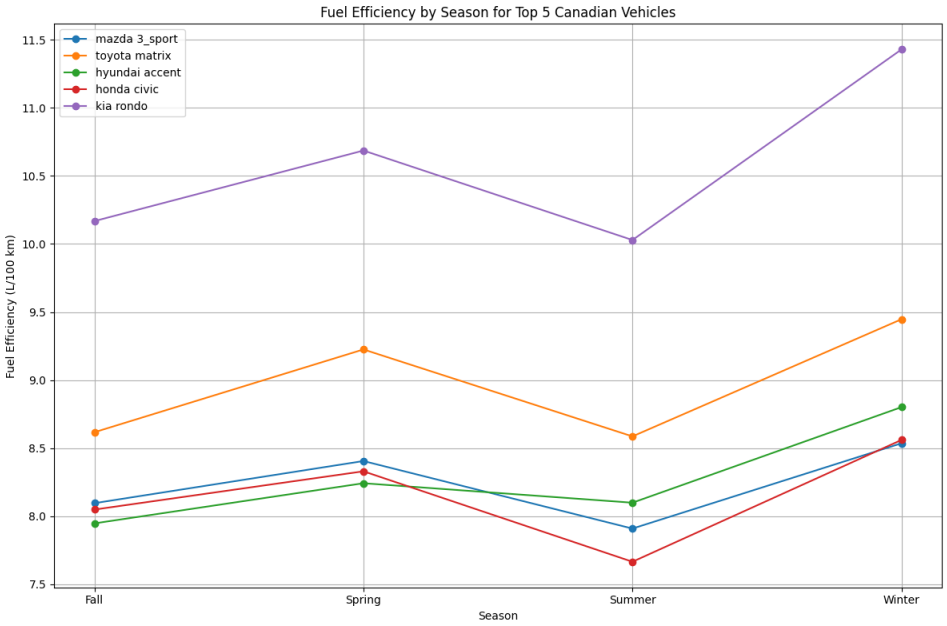

Figure 9: Difference in fuel efficiency for the top 5 Canadian vehicles between season

8. Figure 10 depicts the correlation matrix

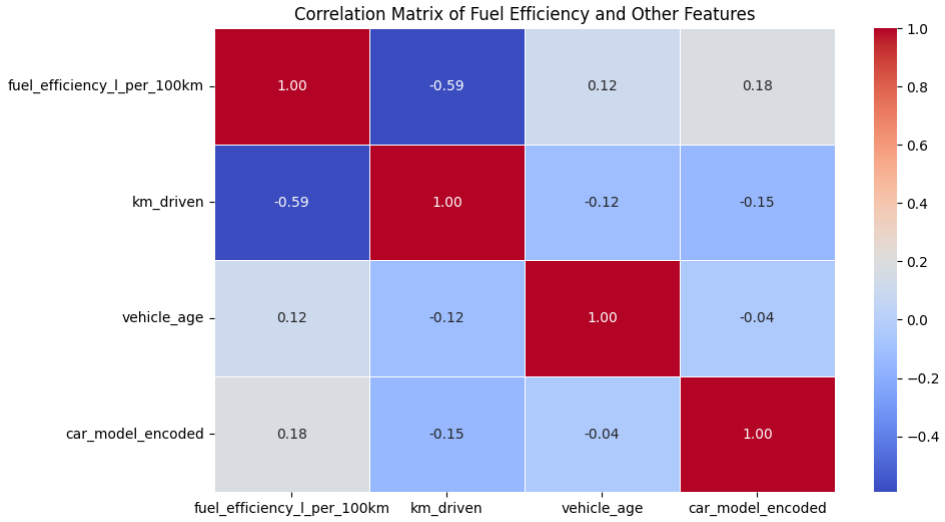

Figure 10: Correlation Matrix

9. **Feature Importances from Random Forest**

In the Random Forest model, `km_driven` stands out as the most important feature, followed by `litres_filled` and `gallons`. Features like `car_model_encoded`, `vehicle_age`, and others have no contribution according to this importance ranking.

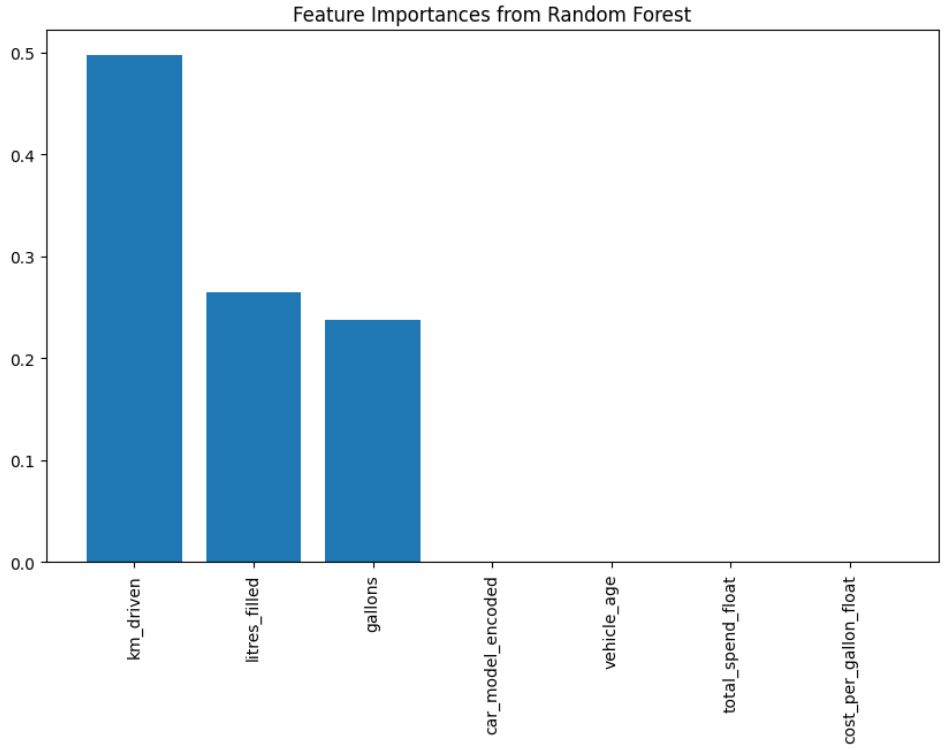

Figure 11: Most important features

Correlation with Fuel Efficiency (L/100km)

From the heatmap, we see that `km_driven` has a significant negative correlation with fuel efficiency (-0.59), indicating that as kilometers driven increase, fuel efficiency (L/100km) tends to decrease. Other variables such as `car_model_encoded` and `vehicle_age` show weaker positive correlations.

Comparison and Analysis

Key Feature: km_driven

- Both the Random Forest feature importances and the correlation matrix indicate that `km_driven` is a critical variable. The Random Forest model assigns it the highest importance (0.4969), and it has the highest absolute correlation with fuel efficiency (-0.591441). This consistency shows that `km_driven` plays a pivotal role in both linear correlation and the more complex decision-making process of the Random Forest.

Other Features

- `litres_filled` and `gallons` have relatively high importance in the Random Forest model, but they don't appear in the correlation matrix provided. These features might be non-linearly related to the target variable, which Random Forest can capture, unlike simple correlation metrics.
- `car_model_encoded` and `vehicle_age` show positive correlations with fuel efficiency but have no importance in the Random Forest model. This discrepancy suggests that while these variables may have some linear relationship with fuel efficiency, they do not significantly influence the model's predictions when all other factors are considered.

Zero-Importance Features

- Several features (`car_model_encoded`, `vehicle_age`, `total_spend_float`, `cost_per_gallon_float`) have zero importance in the Random Forest model, implying that the model didn't find them useful for making predictions. Despite `car_model_encoded` and `vehicle_age` having some correlation with the target, their overall impact may be overshadowed by more dominant features like `km_driven`.

Conclusion

- `km_driven` is a strong predictor of fuel efficiency, both in terms of correlation and importance in the Random Forest model.
- **Non-linear Relationships**: The Random Forest model has identified `litres_filled` and `gallons` as important features, which may not show up in simple correlation analyses due to non-linear relationships.
- **Linear but Non-Impactful**: Features like `car_model_encoded` and `vehicle_age` show some linear correlation with fuel efficiency but don't impact the Random Forest model's predictions significantly, suggesting they might not add predictive power when combined with other features.

This analysis shows how the Random Forest model captures complex relationships that simple correlation might miss, particularly for features like `litres_filled` and `gallons`.

4.3 Fuel Usage in SA

1. Found in Jupyter Notebook.
2. Figure 12 shows the SA fuel price over time.

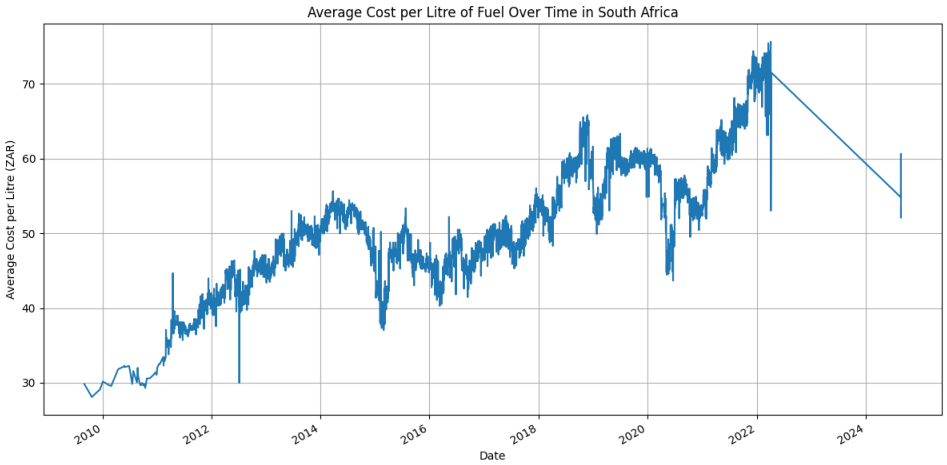

Figure 12: SA fuel price over time

3. Figure 13 shows difference in the number of people refu eling on a Tuesday vs other days.

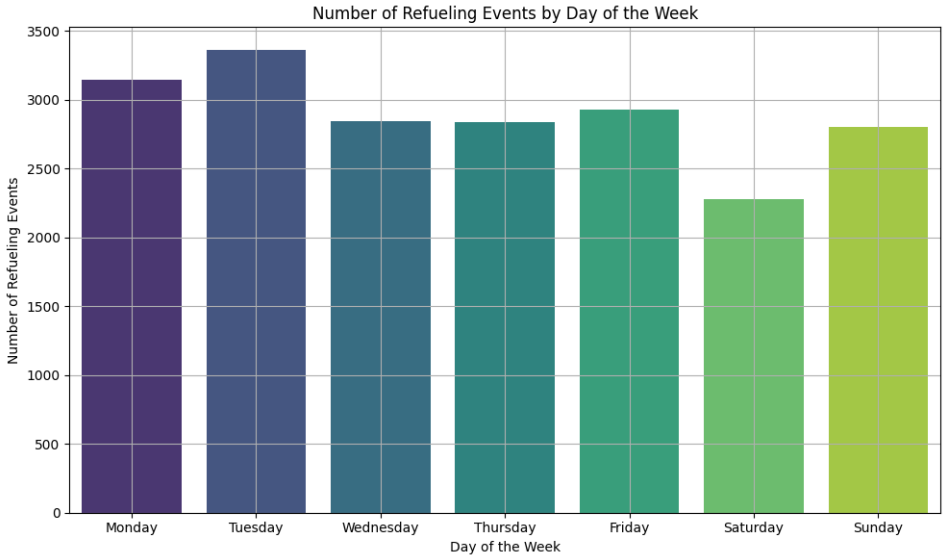

Figure 13: Number of refuels on different days of the week

4. Found in Jupyter Notebook.
5. Found in Jupyter Notebook.
6. Based on the results found, there is no significant difference in refueling events between price decrease (325) and price increase (320) days. The difference of only 5 events is too small to conclude that more people refuel on the first Wednesday of the month when prices go down. The refueling behavior appears to be largely unaffected by these price changes.
7. Based on the results found , more people do not refuel on the first Tuesday of the month when prices go up. In fact, there were slightly fewer refueling events (516) when prices increased compared to when they didn't increase (535). The difference is small (19 events), suggesting that price increases on the first Tuesday of the month do not drive more people to refuel.