IDVE Assignment 1

Ilona Smith - 2471308
Luca von Mayer - 2427051
Tumi Jourdan - 2180153
Tao Yuan - 2332155

[1]*School of Computer Science and Applied Mathematics, University of the Witwatersrand*

Abstract

This report details the process of applying basic cleaning techniques on the logbook data given for IDVE Assignment 1 and exploring the dataset. Observations are provided in plots, descriptive statistics and explanations for relevant sections.

1 Data Cleaning

1.1 Date Fields

1.1.1 Percentage of date fueled entries that are not proper dates

To find how many dates were incorrect, each date field was checked by first seeing if the field was empty, or if it failed a conversion to a DateTime datatype. In these cases it was marked as and non-proper date.

Percentage of invalid 'datefueled' entries: **11.68%**

1.1.5 Distribution of fueling dates

Day of the Month

Most days are relatively consistent with each other. There are four days of exception where the count is significantly lower, the 25th, 29th, 30th and 31st.

25th – Possibly attributed to the day of Christmas which may result in lower traffic on the roads and filling stations being closed 29th to 31st – likely attributed to February not having these days and hence an overall lower contribution.

Month of the Year

Tthe most notable difference is the (relatively) high spike in March. One possible explanation is due to the offset of days in February resulting in those who would have filled up at the end of February instead filling up in March, and the early refilling resulting in another fill in March instead of April.

Some seasonality can also be observed in the middle of the year when compared to the two ends of the year. There is more refuels in December – February when compared to April – September. This is likely due to the holiday season where travel is more prominent as well as the winter months requiring heating in cars, thereby consuming more fuel.

An incremental increase can also be observed in June to August which may be attributed to the summer months in USA (as previously discussed, a large majority of the data is from USA) and the increasing use of air-con in cars during this time.

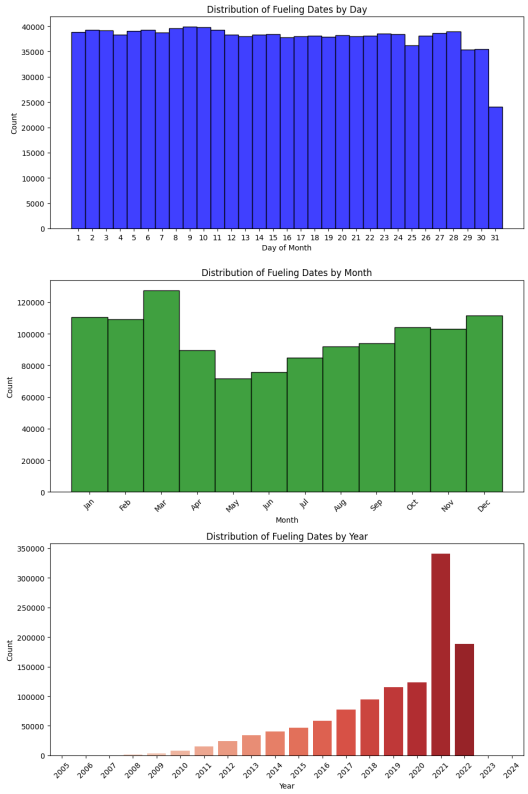

Figure 1: Distribution of fueling dates

Year

There is an increasing trend of the count of fueling records per year until 2021 where it reached its peak. This can be attributed to the increasing usage of the app. Sharp increase in 2021 could be attributed to the service reaching peak popularity and decreases in subsequent years being a decrease in popularity with the last two years only receiving a few hundred logs.

1.2 Numeric Fields

1.2.1 Percentage of gallons, miles, and odometer entries that are missing.

The results were found by getting the mean of the subset of NaN values using the .isna() function.

Percentage of missing values in 'gallons': **6.32%**
Percentage of missing values in 'miles': **87.55%**
Percentage of missing values in 'odometer': **12.70%**

1.2.2 Consolidating miles, gallons and mpg.

The values were calculated with the knowledge that **mpg = miles/gallons**. By manipulating this equation any one of the missing 3 values could be calculated with the other 2 values.

1.2.3 Converting values in Pandas

The commas where removed from the numeric fields, following this they were converted to floats if they were strings.

1.2.4 Distributions of odometer readings, gallons, mpg and miles driven

The distribution is extremely skewed to the right due to the outliers and the large range of values compacting the bins. This results in most of the data being concentrated in a narrow range at the lower end of each scale, while a small number of extreme values stretch the distribution far to the right. The compacted bins at the lower end make it difficult to discern the finer details of the distribution where most of the data points lie.

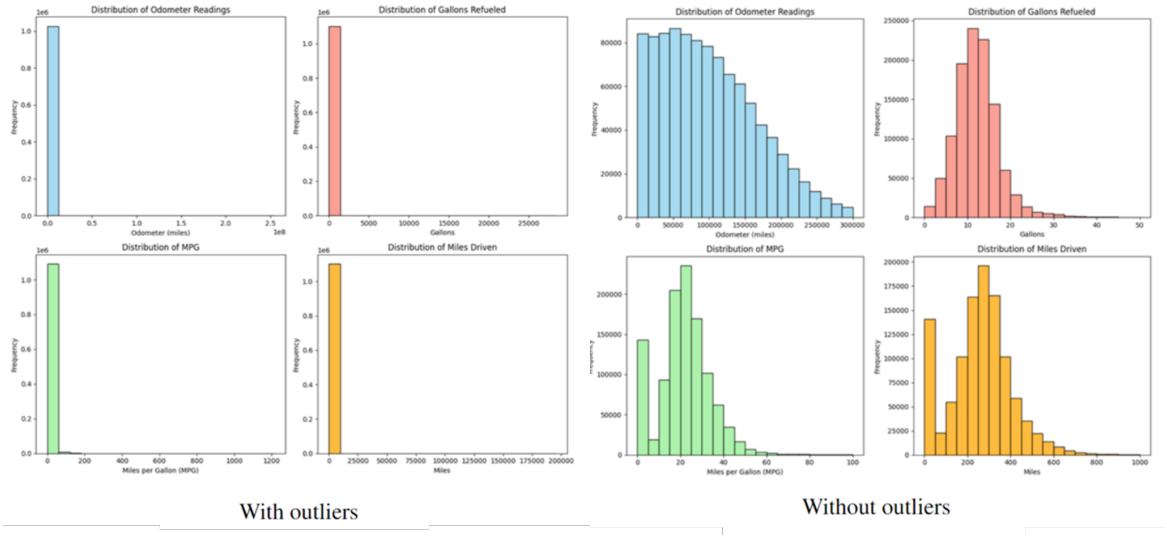

Figure 2: Comparison of data with and without outliers

However when graphed with basic outlier removal.

- Odometer: miles above 300 000 classified as outlier
- Gallons: Above 50 gallons classified as outlier
- MPG: above 100 MPG classified as outlier
- Miles driven: above 1000 miles classified as outlier.

This produces graphs from which more meaningful information can be discerned, although still with a slight right skew. It can be observed that 0 miles per gallon and 0 miles driven has a high frequency which can be attributed to missing data from the user, without which the respective graphs look closer to a standard distribution. Similarly, gallons filled displays a skewed standard distribution.

Distribution of Odometer Readings: The gradual decline in frequency for higher mileage readings is expected, as fewer cars reach very high mileages.

Distribution of Gallons Refueled: The peak is around 10-15 gallons, which is typical capacity for most passenger vehicle fuel tanks.

Distribution of MPG (Miles Per Gallon): This distribution is dependant on and aligns with the distribution of data for gallons filled and miles driven.

Distribution of Miles Driven: The peak is around 225-250 miles, which correspond to the expected mileage per tank of fuel for most vehicles.

1.2.5 Statistical description of the columns for odometer, gallons, mpg and miles

	odometer	gallons	mpg	miles
Mean	103,996	12.8	22.16	269.5
std	340,000	74.47	15.74	725.71
min	0	0	0	0
25%	45,900	8.99	15.6	181.4
50%	91,800	11.95	21.8	267.06
75%	146,900	14.94	28.5	342.77
max	254,300,000	28,380	1,214.3	195,321.2
mode	1	10.57	0	0

Figure 3: mean, standard deviation, max, min, most frequent, and quartiles

- **Odometer:** Mean makes sense as according to `https://www.nytimes.com/2012/03/18/automobiles/as-cars-are-kept-longer-200000-is-new-100000.html?_r=2&ref=business&pagewanted=all&` Cars can be expected to last for 200 000 miles and `https://www.junkcarmedics.com/blog/what-is-the-lifespan-of-a-vehicle-in-the-usa/` suggests that the average mileage of vehicles at end of life is around 150 000 miles. Considering the US form a significant potion of the data and the quality standards of vehicle manufacturers, it should be expected that other countries follow a similar trend. Standard deviation being 340 000 miles does not make sense considering it is over 3 times the mean and almost two times the expected lifetime mileage. Minimum also is not quite as expected because although it is possible for vehicles to be taken to refuel soon after it is purchased, it is likely to have at least some mileage in it between travelling from dealership to refuel station. 25, 50 and 75% are reasonable given the mean. The max of 254,300,000 is particularly concerning as it far exceeds the expected lifetime usage. The mode of 1 is also unexpected and can indicate a default data entry similar to 0 being the minimum.

- **Gallons:** Mean looks to be normal. Standard Deviation looks to be unnaturally high, likely due to the abnormally high max of 28,380 which far exceeds any vehicle capacity. Minimum being 0 is unlikely as it implies that the vehicle does not require fuel (ie: electric car). The likely explanation is misentered data. 25, 50 and 75% all make sense, as it falls within the range for a car's maximum fuel capacity. Mode makes sense when considering not all fill ups are done from empty to max capacity.

- **Mpg:** Mean makes sense according to `https://afdc.energy.gov/data` . Standard deviation is higher than expected, likely affected by the max of 1,214 mpg. Minimum of 0 does not make sense as it translates to not travelling any miles for any number of gallons and is likely due to human error or due to a default value. 25, 50 and 75% are reasonable given the mean. Mode of 0 also does not make sense and is likely to a default value.

- **Miles:** Considering the average gallons and miles per gallon, it is expected that the average total miles between refills is 283 miles. This is consistent with the documented average miles. Standard deviation is unnaturally high, likely due to the abnormally high max of 195,321. Minimum of 0 does not make sense as it translates to not travelling between refills. 25, 50 and 75% all make sense. 75 percentile being 342 miles is high for average car consumption but falls into the range that hybrid cars would allow. Mode of 0 also does not make sense and is likely to a default value. In conclusion, Mode, Min and Max are not reliable, and as a result, neither is standard deviation. Only the percentiles are reliable indicators.

2 Feature Engineering

We first created a 'currency' column by extracting the currency symbol from the 'total_spent' field using regular expressions. Next, we converted the 'total_spent' and 'cost_per_gallon' fields to float values by utilising the function created in Numeric Fields.

We then extracted car make, model, year, and user ID from the URL using regex patterns. Cases in which one or more fields were missing from the URL were accounted for.

To convert the imperial units to metric, we created 'litres_filled' by converting gallons to liters (using UK gallon conversion), 'km_driven' by converting miles to kilometers, and finally calculated 'litres_per_100km' as a measure of fuel efficiency.

3 Vehicle Exploration

3.1 Number of unique users per country

A unique user is defined by their user_id found in the url field. This had been previously split up into atomic fields during Feature Engineering.

Using the currency attribute created in the same section we are able to define unique users in each country by grouping by currency and counting unique user_id's. This is shown in4

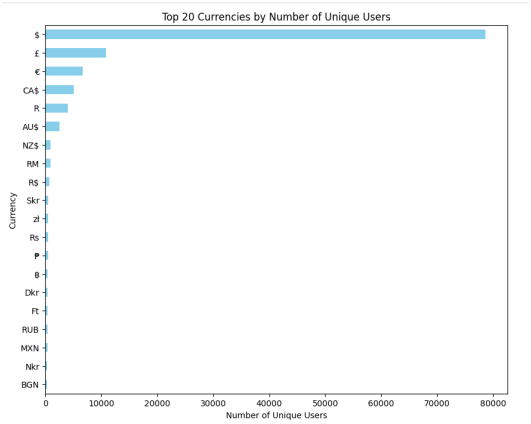

Figure 4: Unique users per country (proxy by currency)

3.2 Popularity of the app

Using a similar approach we can count the number of users per day by grouping by the day of the month and counting unique useridś. This is shown in 5

3.3 Distribution of age of vehicles per country

The age of the vehicle is calculated using the last known year where it refueled and the manufactured year of the vehicle. For the sake of readability only the top 20 is displayed here but the full data calculates for all countries. This is shown in 6

3.4 Popular makes and models

Similarly only the top 20 most common makes and models of vehicles are displayed here for readability. This is shown in 7

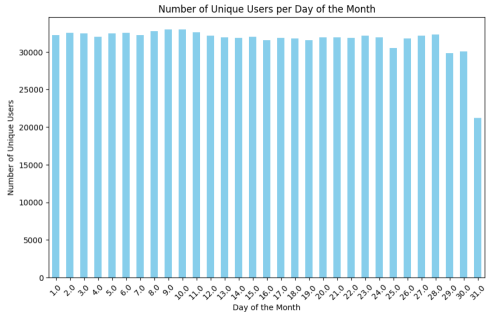

Figure 5: Popularity per month

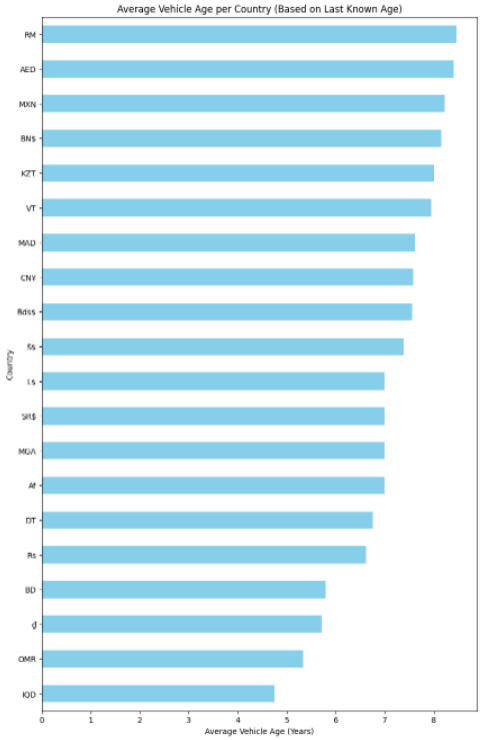

Figure 6: Distribution of top 20 age of vehicles

4 Fuel Usage

4.1 Outlier Removal

We explored ABOD and Isolation Forest for outlier removal due to the high dimensionality of our dataset. ABOD struggled with the large dataset, requiring us to chunk it into segments of 10,000 rows. In comparison, Isolation Forest provided a more rigorous outlier detection. When using the Isolation Forest, using an automatic contamination threshold was too aggressive when removing the outliers. Through iteration we found a contamination threshold of 0.2 alongside IQR to be best. Figure 8 are the stats of that final combination.

Despite the effectiveness of Isolation Forest, some outliers persisted. To address this, we applied IQR outlier removal on the data already cleaned by Isolation Forest.

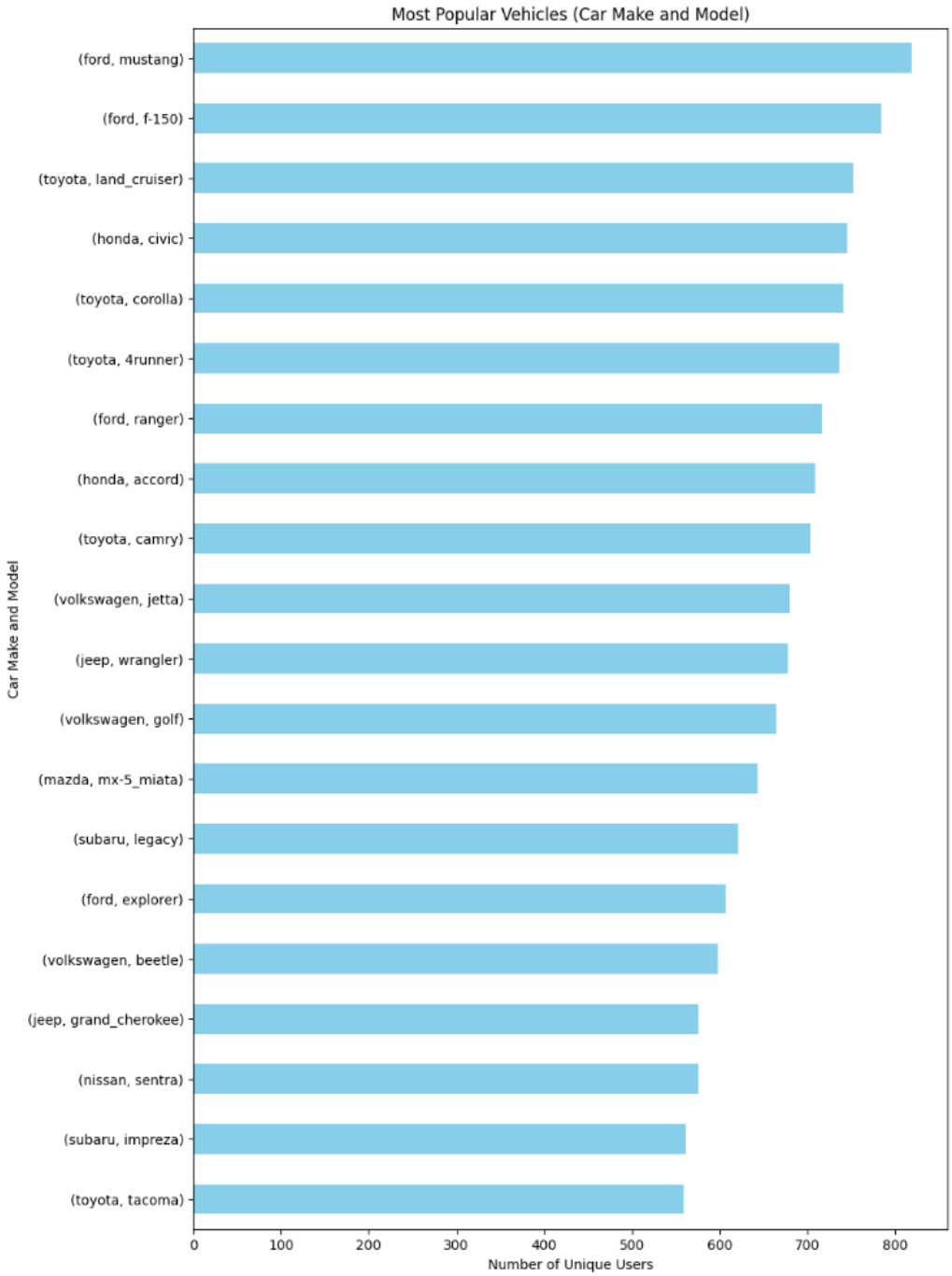

Figure 7: Top 20 most popular makes and models

Additionally, to catch potential outliers caused by incorrect currency entries—especially if the cost per gallon didn't align—we implemented an arithmetic cleaning process. This ensured that the cost per gallon, total gallons, and total spent correlated correctly, removing 8,000 invalid entries.

Overall, the dataset from the top five countries initially had 835,228 elements. After outlier removal, it was reduced to 530,238 elements, resulting in 304,990 rows being removed.

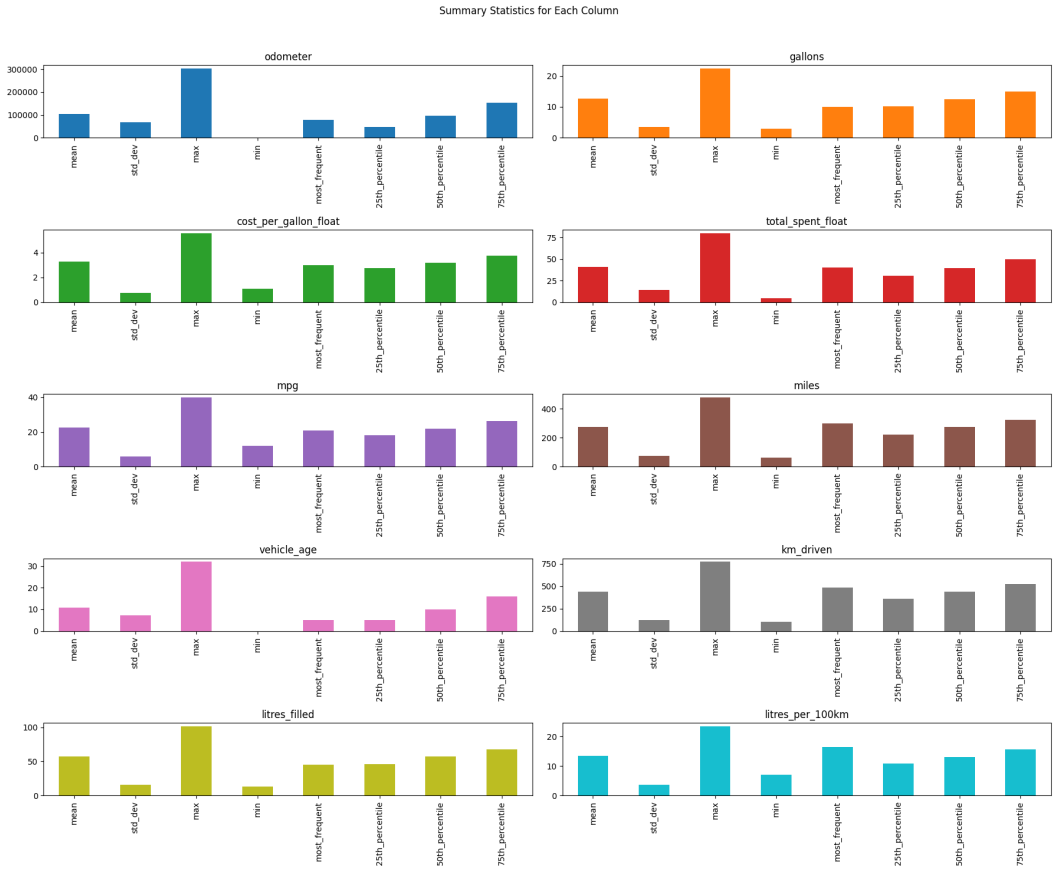

Figure 8: Statistics for the USD after outlier removal

4.2 Fuel Efficiency

4.2.1 Difference in cost per litre per country for January 2022

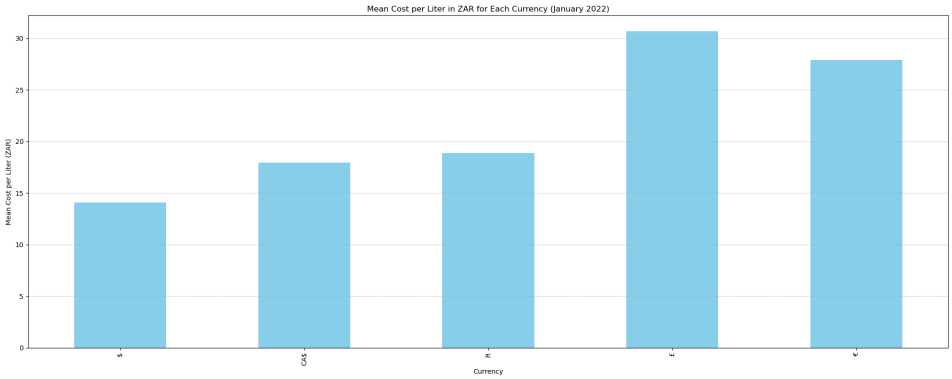

Figure 9: Mean Cost per Liter in ZAR for Each Currency (January 2022)

Source for conversion: `https://www.xe.com/currencytables/?from=ZAR&date=2022-01-15#table-section`

The average price of fuel in USA is just under R15, this increases to around R17 in Canada and R18 in South Africa. European prices are relatively higher at R31 in England and R27 in the rest of

Europe. This can be attributed to the higher tax laws in Europe. https://taxfoundation.org/data/all/eu/gas-taxes-in-europe-2022/

4.2.2 Missed logging of fill ups

To find whether a fill up has been missed we get the the odometer differences between fill-ups for each users. This will find how much they have driven between fill-ups (without using the miles value). We then get this average distance for each user i.e the average amount a user drives between fill-ups. The general rule is then defined as when a users drives more than twice this average distance without recording a fill-up then we can assume they likely missed a fill-up. So for e.g. if a user usually fills up on average every 200 miles and there is an instance of them going 500 miles without a fill-up (> 2x the average) then we can conclude they missed a fill-up.

Estimated number of missed fill-ups: 49103

4.2.3 Average distance (in km) per tank per country.

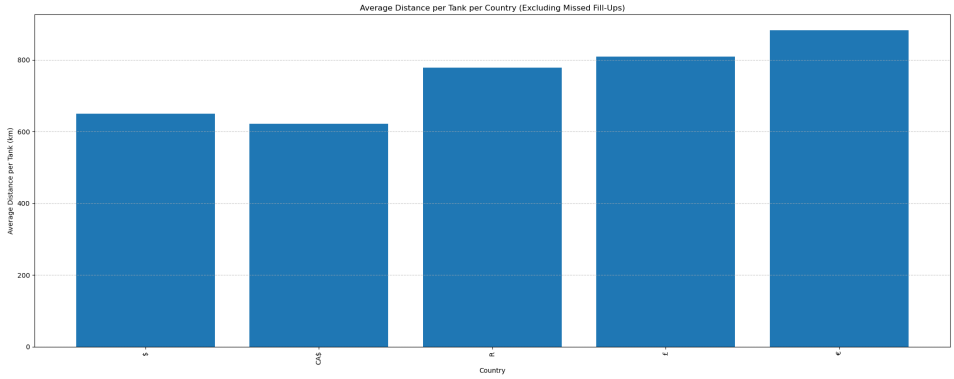

Figure 10: Average Distance per Tank per Country (Including Missed Fill-Ups)

European countries tend to have a higher average distance per tank. This may be due to the following reasons.

- Higher fuel prices push the users to prefer vehicles with higher fuel efficiency.
- Europe has a dense network of efficient, high-speed roads (like the Autobahn) where vehicles can maintain optimal speeds for better fuel economy.
- Manual transmission cars are more common in Europe than in the U.S., and they generally offer better fuel efficiency compared to automatic transmissions.

4.2.4 Correlation of fill ups between distance travelled to the age of vehicle

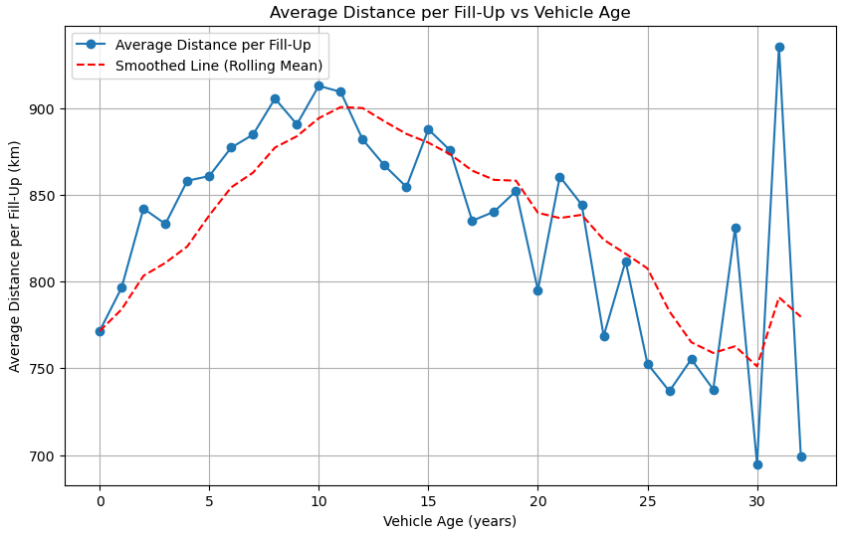

Figure 11: Average Distance per Fill-Up vs Vehicle Age

For the most part newer vehicles do drive further distances between fill-ups, although it seems that vehicles from about 10 years ago drive the most between fill-ups.

4.2.5 Fuel efficiency for the top 5 most popular vehicles in SA

These values are realistic as according to the manufacturer they are expected to consume less litres per 100km. Combined with suboptimal driving conditions it is normal for it to consume higher than quoted.

- **Mitsubishi** - `https://www.group1mitsubishi.co.za/mitsubishi-pajero-swb/`
- **Hilux** - `https://www.cars.co.za/motoring-news/new-toyota-hilux-48v-mild-hybrid-fuel-economy-revealed/210741/`
- **Fortuner** - `https://topauto.co.za/features/101031/comparing-the-new-toyota-fortuner-hybrids-efficiency-vs-other-r900000-suvs/#:~:text=Toyota%20says%20that%20the%2048V,average%20of%207.31%2F100km.`
- **Jimny** - `https://www.cars.co.za/motoring-news/suzuki-jimny-what-its-like-to-live-with/113307/`
- **Polo** - `https://www.autotrader.co.za/cars/news-and-advice/buying-a-car/which-volkswagen-polo-is-better-petrol-or-diesel/6841`

	car_make	model	litres_per_100km
0	mitsubishi	pajero	14.347408
3	toyota	hilux	13.300572
2	toyota	fortuner	13.193800
1	suzuki	jimny	11.187171
4	volkswagen	polo	9.248023

Figure 12: 5 most popular vehicles in SA

4.2.6 Most fuel efficient vehicles in each country

	currency	car_make	model	litres_per_100km
375	$	ford	sierra	22.873382
735	$	mercedes-benz	cls63_amg_s	22.796597
470	$	isuzu	bighorn	22.780777
1039	$	suzuki	ignis	22.419178
728	$	mercedes-benz	clk63_amg	22.391410
1307	CA$	bmw	535xi	21.370819
1461	CA$	lexus	gs460	21.239221
1452	CA$	land_rover	range_rover_sport	21.137135
1245	CA$	audi	rs6	20.651209
1413	CA$	jeep	commander	20.255758
2112	R	toyota	condor	20.034159
1744	R	alfa_romeo	156	19.377028
2004	R	mercedes-benz	e500	19.255015
2000	R	mercedes-benz	c63_amg	18.753977
2013	R	mercedes-benz	ml63_amg	18.595549
2608	£	mercedes-benz	ml320	15.869755
2771	£	toyota	granvia_	15.697785
2681	£	nissan	stagea	15.563845
2207	£	audi	q7	15.352263
2623	£	mercedes-benz	sprinter_3500	15.260162
3381	€	toyota	caldina	14.295942
2970	€	bmw	760li	14.195057
3256	€	nissan	350z	14.112200
3224	€	mercedes-benz	ml350	13.964030
3198	€	mercedes-benz	cla45_amg	13.916702

Figure 13: 5 most popular vehicles in top 5 countries

4.2.7 Difference in fuel efficiency for the top 5 Canadian vehicles between seasons

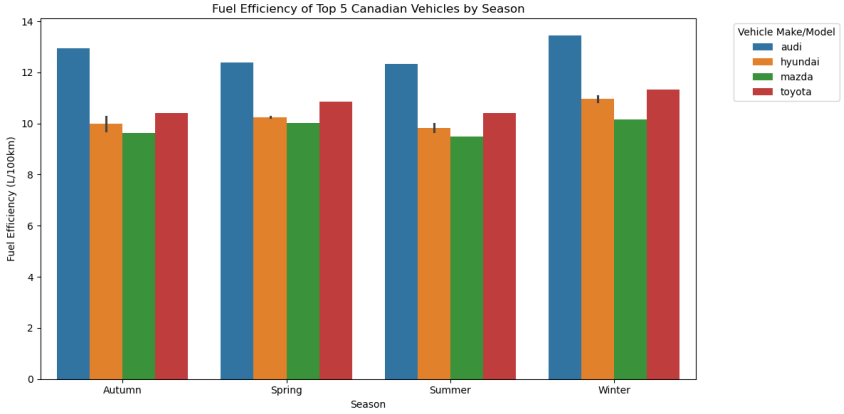

Figure 14: Fuel Efficiency of Top 5 Canadian Vehicles by Season

This makes sense as the winter season is slightly increased, seeing as Canada is a very cold country people would be less likely to walk in winter and thus would drive more in these months for short distances. Furthermore, the weather may effect the running of the car making it less efficient, due to having to constantly use a heater in the car which consumes petrol.

4.2.8 Correlations between fuel efficiency and other features

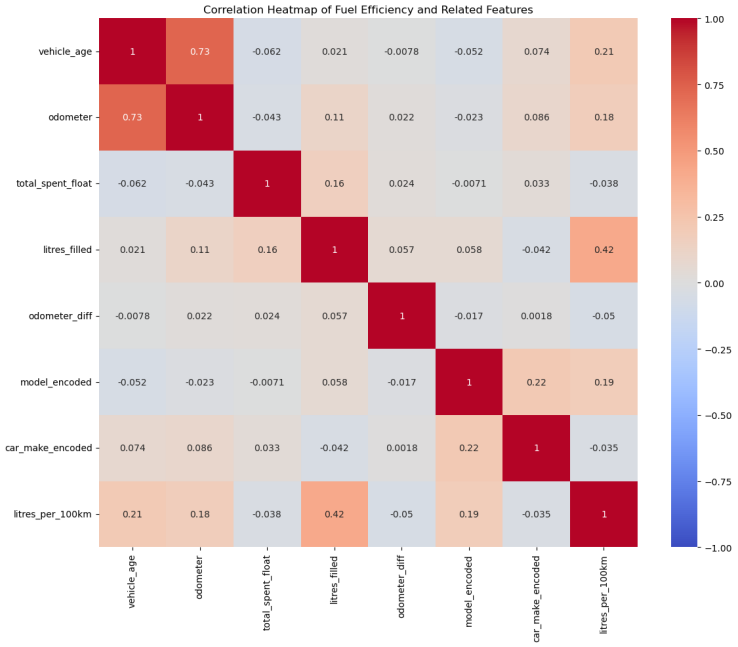

Figure 15: Correlation Heatmap of Fuel Efficiency and Related Features

```
Correlations with litres_per_100km:
litres_per_100km     1.000000
litres_filled        0.417235
vehicle_age          0.205728
model_encoded        0.187408
odometer             0.175322
car_make_encoded    -0.035176
total_spent_float   -0.037832
odometer_diff       -0.050010
```

As predicted the most correlated features with the fuel efficiency are distance travelled (odometer), the age of the vehicle, and the model of vehicle, amongst others such as liters filled. These all intuitively make sense.

4.2.9 Using a random forest to get a list of the most important variables

According to the random forest the variables litres filled and km driven are significantly more important than all other features. These are directly correlated as the distance you can travel depends on how much fuel is in the tank.

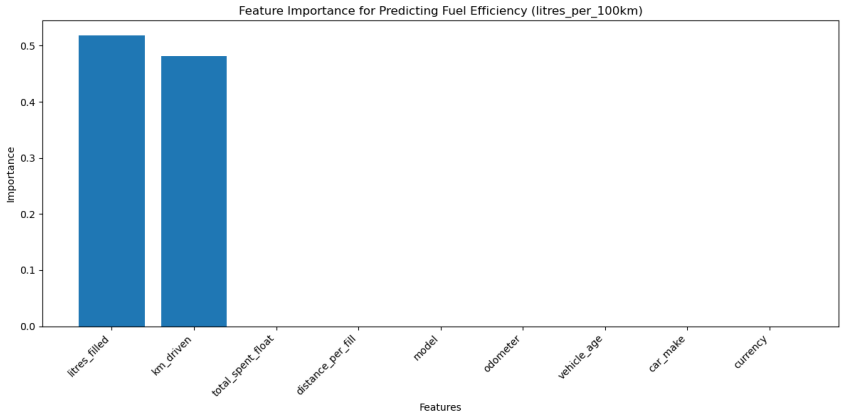

Figure 16: Feature Importance for Predicting Fuel Efficiency (litres_per_100km)

```
              feature  importance
4      litres_filled    0.520127
5          km_driven    0.479799
8           currency    0.000025
3  total_spent_float    0.000010
1  distance_per_fill    0.000010
7              model    0.000009
2           odometer    0.000009
0        vehicle_age    0.000007
6           car_make    0.000006

Difference in importance between adjacent features:
              feature  importance  importance_diff
4      litres_filled    0.520127              NaN
5          km_driven    0.479799     4.032804e-02
8           currency    0.000025     4.797742e-01
3  total_spent_float    0.000010     1.438728e-05
1  distance_per_fill    0.000010     5.702076e-07
7              model    0.000009     8.844530e-07
2           odometer    0.000009     3.876627e-08
0        vehicle_age    0.000007     2.017457e-06
6           car_make    0.000006     3.364163e-07
```

4.3 Fuel Usage in SA

4.3.1 Price over time

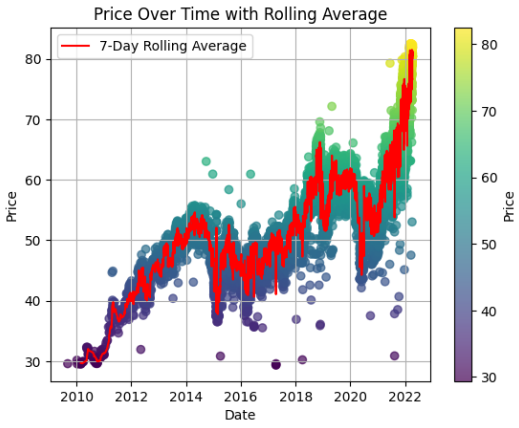

Figure 17: Price over time

Figure 17 shows a general trend in increasing fuel price from 2010 to the highest prices in 2022.

4.3.2 Users Refueling Per Day

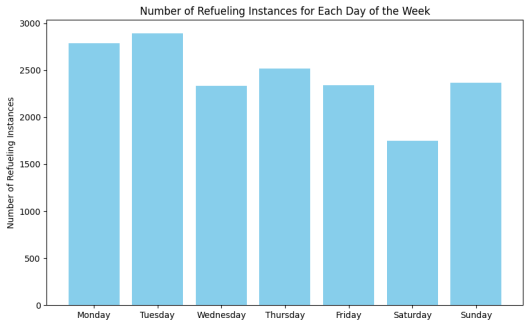

Figure 18: Price over time

The most common days to refuel are Monday and Tuesday. This could be because of people attempting to refuel before the price change on the next Wednesday.

4.3.3 Trends of refueling on Tuesdays and Wednesdays

The results from the data set are :

- Refuelings on first Wednesdays:

Down: 252
No Change: 8
Up: 305
More people do not refuel on the first Wednesday when prices go down.

- Refuelings on first Tuesdays:

Down: 252
No Change: 17
Up: 605
More people refuel on the first Tuesday when prices go up.

The largest difference is in the first Tuesdays where 400 more are refueling on the Tuesday before the price increases, which makes sense as people want to spend less on fuel in the oncoming month.