COMS4060A Data Visualisation and Exploration - Assignment 1

Lisa Godwin 2437980 **Nihal Ranchod 2427378** **Zach Schwark 2434346**

All plots and jupyter notebook can be found at this Github Link.

1 Data Cleaning

1.1 Date Fields

1. Total entries: 1174870
   Improper date entries: 135126
   Percentage of improper dates: 11.50%
2. Total entries: 1174870
   Improper date entries: 0
   Percentage of improper dates: 0.00%
3. Figure 1

	date_fueled	date_captured
0	2022-04-07	2022-04-07
1	2012-11-07	2016-08-30
2	2012-09-22	2012-09-28
3	2019-05-04	2019-05-04
4	2022-02-15	2022-02-15

Figure 1: Showing Dates are in correct format

4. Answer done in code.
   Removing dates for both date_fueled and date_captured in this range. If one column (date_fueled or date_captured) has a valid date while the other column has an invalid date (in the future or before 2005), it could lead to inconsistencies. For example, analysing time differences between fueling and capturing, such inconsistencies could distort the results. By removing invalid dates from both columns, it simplifies the data cleaning process, ensuring that all remaining records are within the valid date range.
5. The distribution plot 2 shows the valid fuelling dates within the specified range (from 2005 to the current date).

   - Distribution of Fuelling Dates:
     The data seems to be heavily skewed towards the more recent dates. The KDE line peaks significantly towards the right-hand side of the plot, suggesting that most of the refuelling entries occurred within a more condensed recent time period.

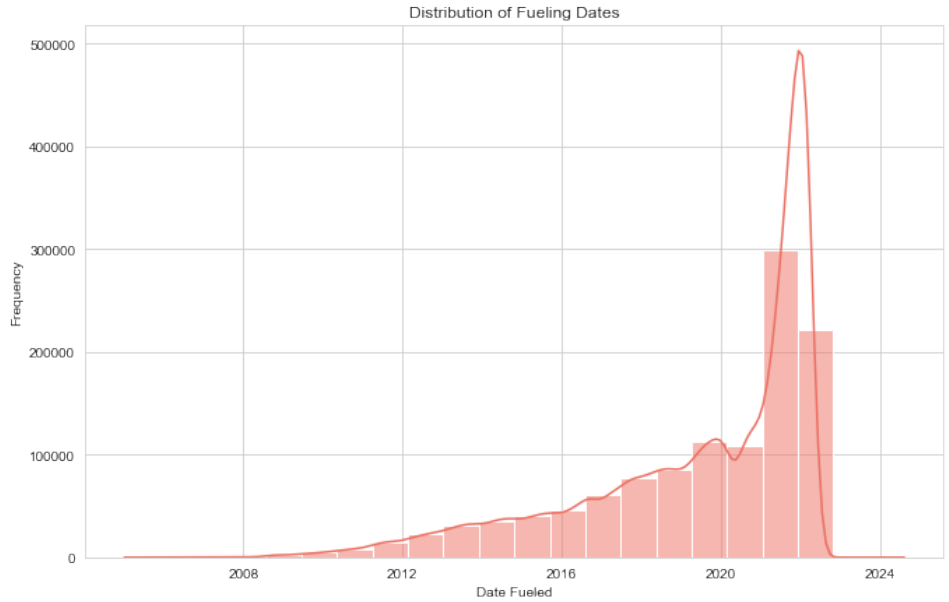

Figure 2: Distribution Plot of fuelling dates

- Historical Refuelling Activity:
  There is a gradual increase in the number of entries over time, with a sharp rise as you approach the more recent dates. This could indicate that the dataset either has more recent entries or that vehicle refuelling was recorded more frequently as time progressed.
- Binning:
  The histogram bins are wide enough to capture general trends, but there might be clustering towards specific time periods, which the KDE line emphasises.

1.2 Numeric Fields

1. Percentage of missing entries in 'gallons': 6.33%
   Percentage of missing entries in 'miles': 87.60%
   Percentage of missing entries in 'odometer': 12.65%
2. Figure 3

$$miles = gallons \times mpg \tag{1}$$

$$gallons = \frac{miles}{mpg} \tag{2}$$

$$mpg = \frac{miles}{gallons} \tag{3}$$

Conversion to float to check for conversion errors.

	gallons	miles	mpg
0	NaN	NaN	NaN
1	12.120	NaN	31.6
2	7.991	NaN	28.5
3	10.575	NaN	46.8
4	11.651	244.4	21.0

After applying missing values function.

	gallons	miles	mpg
0	NaN	NaN	NaN
1	12.120	382.9920	31.6
2	7.991	227.7435	28.5
3	10.575	494.9100	46.8
4	11.651	244.4000	21.0

Figure 3: Values have been calculated for miles, gallons and miles per gallon

3. Figure 4

	odometer	cost_per_gallon_float	total_spent_float
0	73370.0	NaN	NaN
1	11983.0	5.599	67.86
2	98233.0	5.450	43.53
3	163802.0	5.110	54.00
4	NaN	3.029	35.29

Figure 4: Converted odometer, cost_per_gallon and total_spent to be floats

4. We are unable to make meaningful observations about these graphs due to the lack of a clear distribution in Figure 11. This suggests the presence of significant outliers that obscure our data, making it difficult to draw conclusions. Given the diverse range of currencies involved, it was anticipated that interpreting Total Spent and Cost per Gallon would be challenging. So we are going to use a boxplot 18 to help us see the outliers and where they should be removed.

   After examining the boxplots 18, it is evident that there are significant outliers present in the data. These outliers skew the distributions, making it challenging to draw accurate and

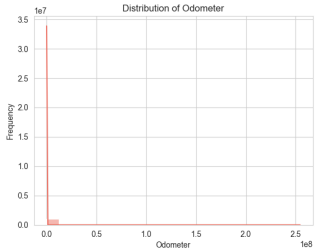

Figure 5: Distribution Plot for Odometer

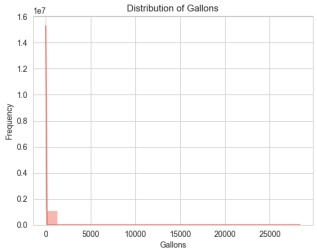

Figure 6: Distribution Plot for Gallons

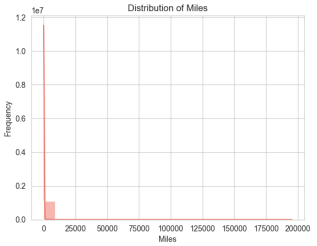

Figure 7: Distribution Plot for Miles

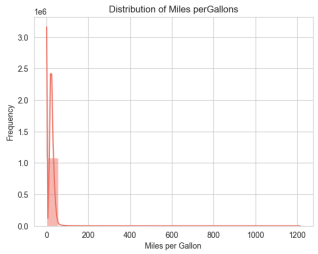

Figure 8: Distribution Plot for Miles per Gallon

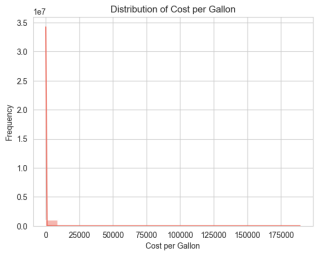

Figure 9: Distribution Plot of Cost per Gallon

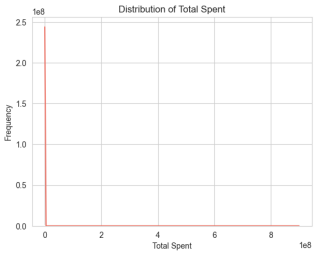

Figure 10: Distribution Plot for Total Spent

Figure 11: Initial Distribution Plots

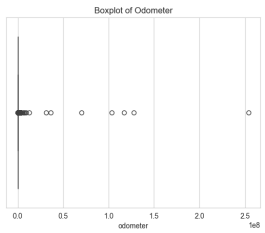

Figure 12: Boxplot for Odometer

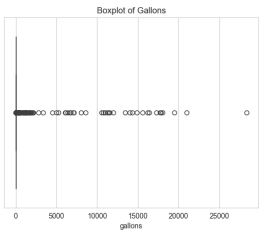

Figure 13: Boxplot for Gallons

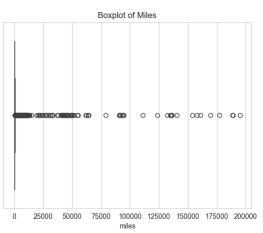

Figure 14: Boxplot for Miles

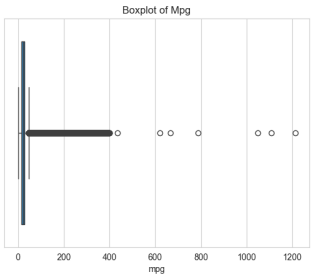

Figure 15: Boxplot for Miles per Gallon

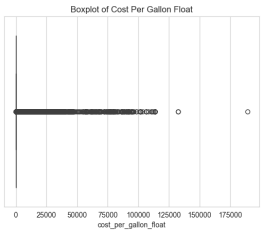

Figure 16: Boxplot of Cost per Gallon

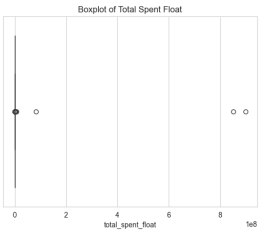

Figure 17: Boxplot for Total Spent

Figure 18: Boxplots to view outliers

informed conclusions. Therefore, we filter out these extreme values using percentiles before conducting further analysis.

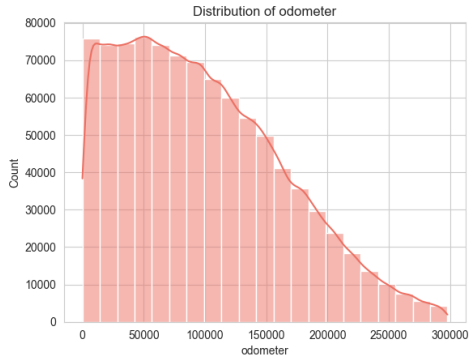

Figure 19: Distribution Plot for Odometer

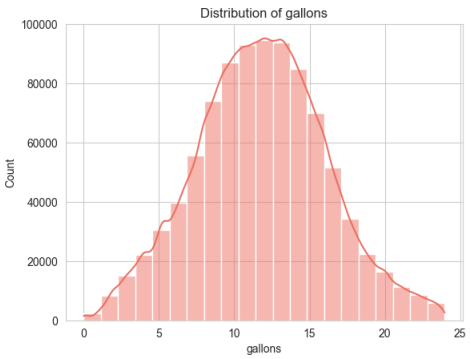

Figure 20: Distribution Plot for Gallons

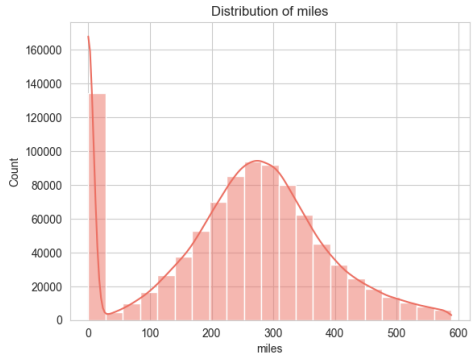

Figure 21: Distribution Plot for Miles

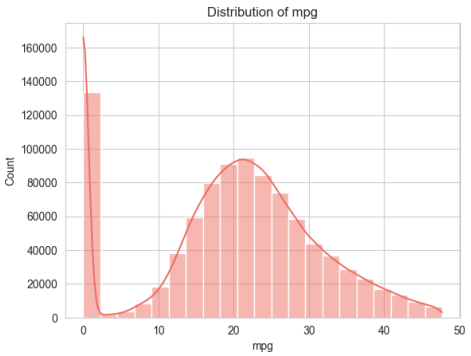

Figure 22: Distribution Plot for Miles per Gallon

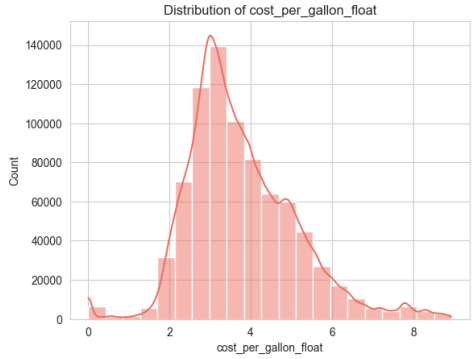

Figure 23: Distribution Plot of Cost per Gallon

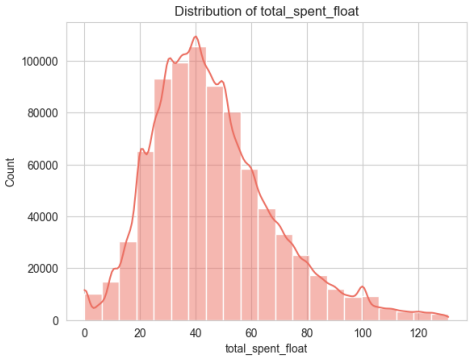

Figure 24: Distribution Plot for Total Spent

Figure 25: Distribution Plots after Outliers have been removed

The odometer distribution in Figure 19 is right-skewed, with most vehicles having low mileage and a steady decline as mileage increases. The peak at lower odometer readings indicates a predominance of newer or less-driven vehicles, while a long tail shows fewer high-mileage vehicles. The gallons distribution in Figure 20 is symmetric and bell-shaped, with the most frequent fuel purchase amount around 12-13 gallons, reflecting consistent consumption patterns. The miles distribution in Figure 21 is right-skewed, showing a high frequency of very short trips near 0-50 miles and another peak around 250-300 miles,

suggesting a mix of short and long journeys. The miles_per_gallon distribution is right-skewed, with a high frequency of very low values between 0-5 and another peak around 20-30, indicating a mix of highly inefficient and moderately efficient fuel consumption. he distributions for cost_per_gallon_float and total_spent_spent exhibit right-skewed patterns. For cost_per_gallon_float in Figure 23, most values cluster around 2 to 4 units. For total_spent_float in Figure 24, the values primarily cluster between 20 and 60 units. These distributions reflect a right-skew due in part to the exclusion of outliers, which were removed to simplify the analysis. This approach tends to highlight lower cost and spending values while higher values are less represented. Consequently, the datasets, which include a higher number of transactions from regions with lower fuel costs, may not fully capture the global variation in fuel prices and total expenditures

5. The summary statistics in Figure 26 reveal several significant data quality issues, which initially made it challenging to draw accurate conclusions. Extreme standard deviations and unrealistic maximum values across multiple columns suggest the presence of severe outliers and potential data entry errors. For example, the odometer reading had an implausible maximum of 254 million miles, and the gallons, mpg, miles, cost per gallon, and total spent columns also showed exaggerated maximum values. Minimum values of zero in several columns further indicated possible missing or miss entered data. The monetary columns, such as cost per gallon and total spent, were particularly difficult to analyse due to the presence of over 100 different currencies in the dataset, contributing to the observed extreme variability. After addressing these outliers, the data became more consistent and reliable, improving the accuracy and interpretability of the results.

Summary statistics for data without outlier removal:

	odometer	gallons	mpg	miles	cost_per_gallon_float	total_spent_float
count	1024016.000	1098145.000	1098145.000	1098145.000	1089593.000	1096068.000
mean	104001.680	12.798	22.159	269.432	107.894	2751.337
std	340794.908	74.550	15.744	726.395	1682.883	1184992.917
min	0.000	0.000	0.000	0.000	0.000	0.000
25%	45926.000	8.988	15.500	181.338	2.999	32.480
50%	91882.000	11.953	21.800	267.019	3.859	47.010
75%	146925.000	14.937	28.500	342.741	5.310	70.810
max	254362100.000	28380.000	1214.300	195321.200	189270.590	899647469.000

Summary statistics for data with outlier removal:

	odometer	gallons	mpg	miles	cost_per_gallon_float	total_spent_float
count	932494.000	919953.000	922574.000	924049.000	801898.000	807771.000
mean	98608.355	11.805	20.314	244.014	3.773	45.978
std	66345.408	4.334	11.313	138.315	1.331	22.076
min	0.000	0.000	0.000	0.000	0.000	0.000
25%	44288.000	8.881	14.700	169.201	2.879	30.400
50%	89440.000	11.812	21.100	259.769	3.519	42.700
75%	143385.750	14.683	27.400	332.572	4.540	57.940
max	297731.000	23.961	47.700	588.588	8.940	130.700

Figure 26: Statistics for Numerical Data

2 Feature Engineering

1. Figure 27

	currency
0	NaN
1	$
2	£
3	£
4	$

Figure 27: New currency column created

2. Already done. Can be seen in Figure 4
3. Figure 28

	make	model	year	user_id
0	suzuki	swift	2015	674857
1	bmw	x3	2009	461150
2	mercedes-benz	e300	1998	133501
3	bmw	320d	2010	247233
4	honda	passport	2019	1038865

Figure 28: Car make, model, year, user ID columns created from user_url

4. Determined the predominant currency type in the dataset which was US gallons and set the conversion factor. Using US gallon to litre conversion factor: 3.78541. Can be found in Figure 29

$$litres\_filled = gallons \times 3.78541 \tag{4}$$

	currency_type	litres_filled
0	Other	NaN
1	US	45.879
2	UK	30.249
3	UK	40.031
4	US	44.104

Figure 29: Converted gallons to litres

5. Conversion factor from miles to kilometres: 1.60934. Can be found in 30

$$km\_driven = miles \times 1.6093474 \tag{5}$$

	miles	km_driven
0	NaN	NaN
1	382.992	616.364
2	227.743	366.517
3	494.910	796.478
4	244.400	393.323

Figure 30: Converted miles to kilometres

6. Figure 31

$$litres\_per\_100km = \frac{litres\_filled \times 100}{km_driven} \tag{6}$$

	litres_filled	km_driven	litres_per_100km
0	NaN	NaN	NaN
1	45.879	616.364	7.444
2	30.249	366.517	8.253
3	40.031	796.478	5.026
4	44.104	393.323	11.213

Figure 31: Calculated litres per 100km

3 Vehicle Exploration

1. Figure 32

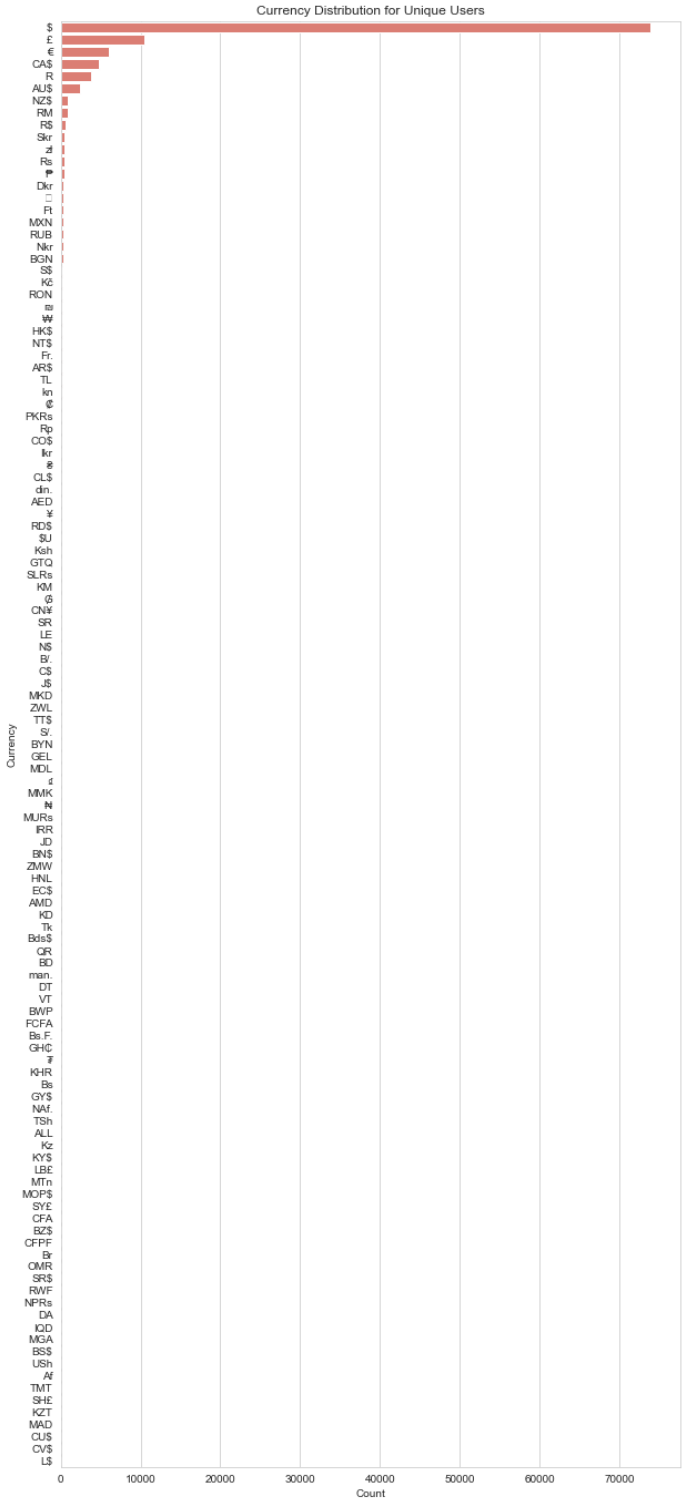

Figure 32: New currency column created

2. The app is most popular during the beginning of the month, specifically with in the first 10 days of the month which can be seen in Figure33

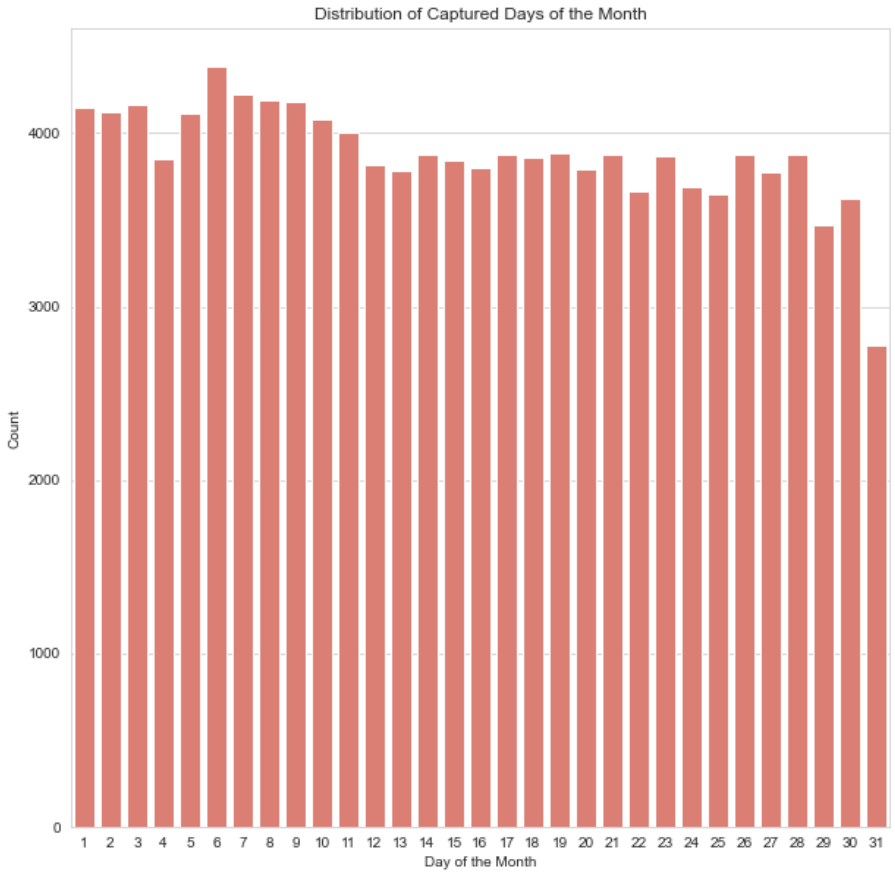

Figure 33: New currency column created

3. There are over a 100 distribution plots for all the different currencies. So we are just going to look at the top 6 currencies.
   In the more developed countries, such as the United States, Europe, United Kingdom, Australia and New Zealand. The distribution of the age of the cars as seen in Figure 40 is generally lower, as they have more cars that are younger. These younger cars have ages that are less than 10 years old. This is also surprisingly true for South Africa as well, even though South Africa is not as developed as the other countries mentioned. Most of the cars for all countries have an age that is less than 20 years old.
4. Ford, Toyota, BMW, Nissan and Volkswagen are the most popular car makes as seen in Figure 41.
   The most popular car models are the mustang, F-150, Land Cruiser, Civic and Corolla as seen in Figure 42.

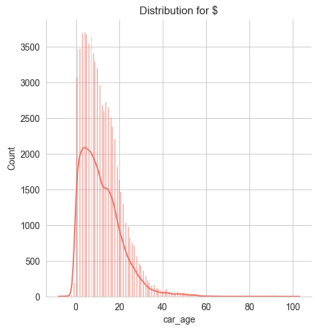

Figure 34: Distribution Plot for United States ($)

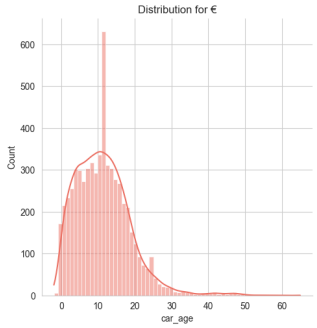

Figure 35: Distribution Plot for Europe (€)

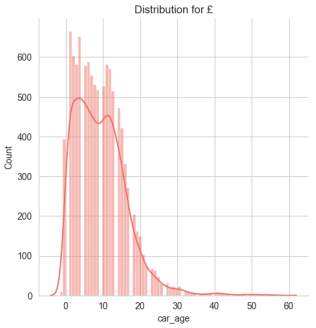

Figure 36: Distribution Plot for United Kingdom (£)

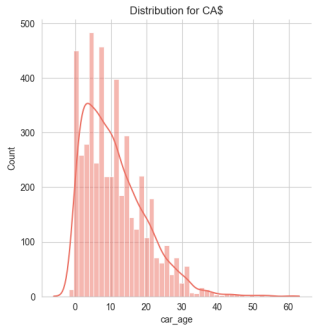

Figure 37: Distribution Plot for Canadian Dollar (CA$)

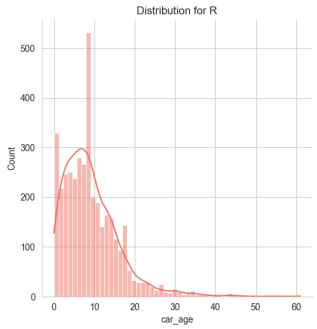

Figure 38: Distribution Plot of South Africa (R)

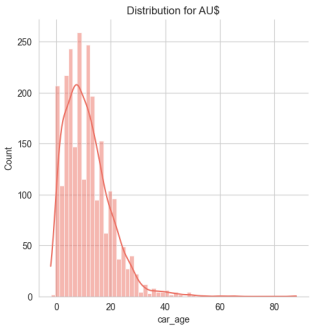

Figure 39: Distribution Plot for Australia (AU$)

Figure 40: Distribution Plots of age of the vehicles per country

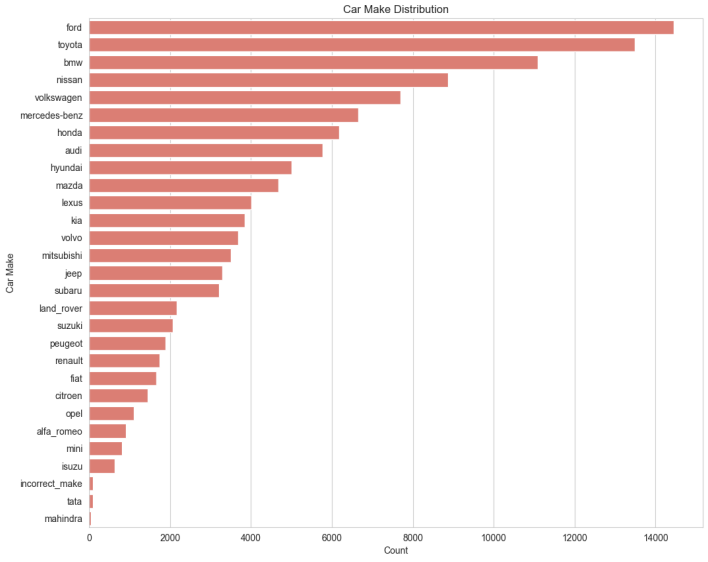

Figure 41: Car make distribution

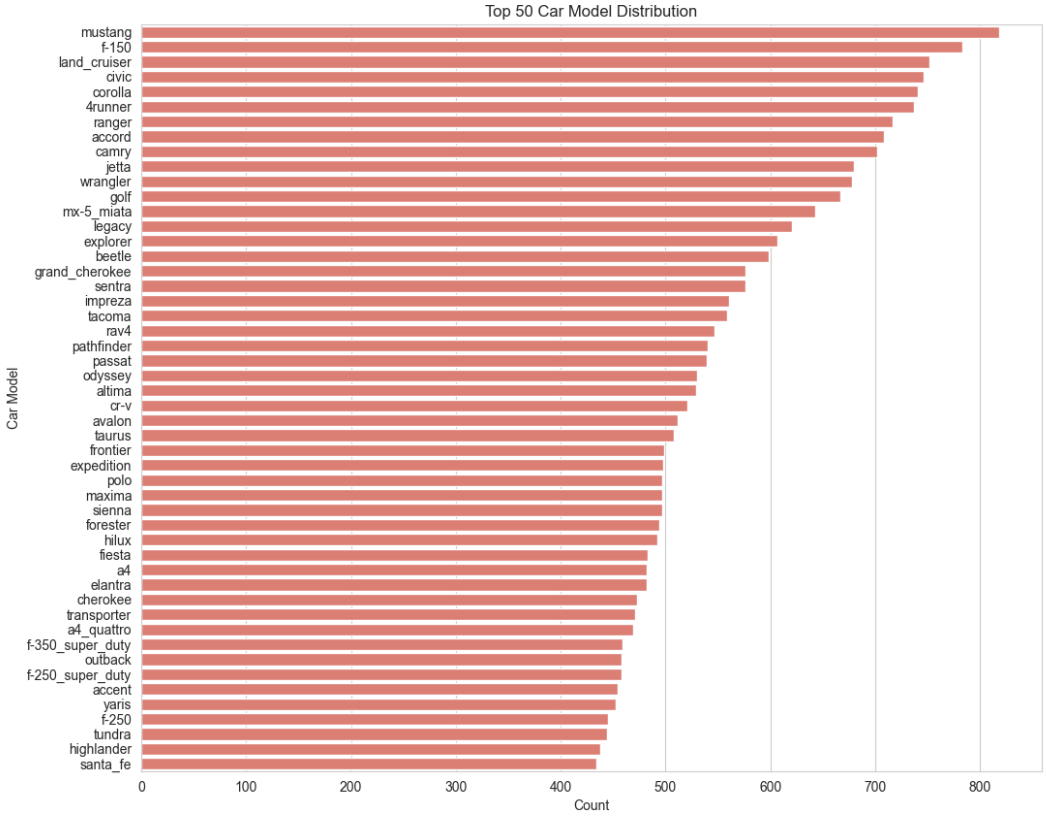

Figure 42: Top 50 car models distribution

4 Fuel Usage

4.1 Outlier Removal

1. The top 5 currencies are:

$	738381
£	86147
€	58441
CA$	46234
R	36034

2. First create cost_per_litre column, see Figure 43

   To remove outliers, first establish reasonable thresholds for each currency and field by performing statistical analysis, including calculating measures like mean, median, quartiles, and the interquartile range (IQR) using pandas. Define thresholds based on these analyses, setting boundaries at 1.5 times the IQR below the first quartile and above the third quartile. Apply these thresholds to filter the dataset, targeting metrics such as total_spent_float, gallons, and cost_per_gallon_float for USD and CAD, and equivalent metrics for GBP, Euro, and Rand.

   General Approach:

	cost_per_gallon_float	cost_per_litre	currency
1	5.599	1.479	$
2	5.450	1.440	£
3	5.110	1.350	£
4	3.029	0.800	$
5	3.739	0.988	$

Figure 43: New currency column created

- **Total Spend (`total_spent_float`):** Outliers are values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$.
- **Gallons (`gallons`) / Litres (`litres_filled`):** Outliers are values significantly outside the typical range.
- **Cost per Gallon / Cost per Litre:** Extreme values may indicate incorrect data, with outliers identified using the IQR method.

Suggested Thresholds:

Currency	Metric	Range	Reasoning
USD ($)	Total Spend	1 to 130	Q1 = 30, Q3 = 56, IQR = 25.75. Outliers below 1 or above 130.
	Gallons	1 to 30	Q1 = 9.47, Q3 = 15.31, IQR = 5.83. Outliers below 1 or above 30.
	Cost per Gallon	1 to 5	Q1 = 2.78, Q3 = 3.99, IQR = 1.22. Outliers below 1 or above 5.
CAD ($)	Total Spend	1 to 130	Q1 = 41.86, Q3 = 72.25, IQR = 30.39. Similar range to USD.
	Gallons	1 to 30	Q1 = 8.99, Q3 = 14.36, IQR = 5.37. Outliers below 1 or above 30.
	Cost per Gallon	3 to 7	Q1 = 4.31, Q3 = 5.55, IQR = 1.24. Values below 3 or above 7.
GBP (£)	Total Spend	1 to 120	Q1 = 37.96, Q3 = 67.59, IQR = 29.63. Outliers below 1 or above 120.
	Litres Filled	1 to 85	Q1 = 31.26, Q3 = 53.13, IQR = 21.87. Outliers below 1 or above 85.
	Cost per Litre	0.5 to 2	Q1 = 1.16, Q3 = 1.36, IQR = 0.2. Outliers below 0.5 or above 2.
Euro (€)	Total Spend	1 to 125	Q1 = 38.2, Q3 = 70.22, IQR = 32.02. Outliers below 1 or above 125.
	Litres Filled	1 to 85	Q1 = 29.07, Q3 = 51.0, IQR = 21.93. Values below 1 or above 85.
	Cost per Litre	0.5 to 2	Q1 = 1.23, Q3 = 1.55, IQR = 0.32. Outliers below 0.5 or above 2.
Rand (R)	Total Spend	100 to 1500	Q1 = 469.32, Q3 = 902.52, IQR = 433.2. Values below 100 or above 1500.
	Litres Filled	1 to 90	Q1 = 33.5, Q3 = 61.0, IQR = 27.5. Values below 1 or above 90.
	Cost per Litre	10 to 25	Q1 = 12.79, Q3 = 16.48, IQR = 3.69. Outliers below 10 or above 25.

Table 1: Suggested thresholds for outlier detection based on IQR method.

3. Number of values removed after accounting for outliers: 110491

4.2 Fuel Efficiency

1. USD ($) to Rands (R) - $1 = R15.518617
   CAD (CA$) to Rands (R) - CA$ = R12.289890
   GBP (£) to Rands (R) - £1 = R21.037880
   Euro (€) to Rands (R) - €1 = R17.575064 [1]

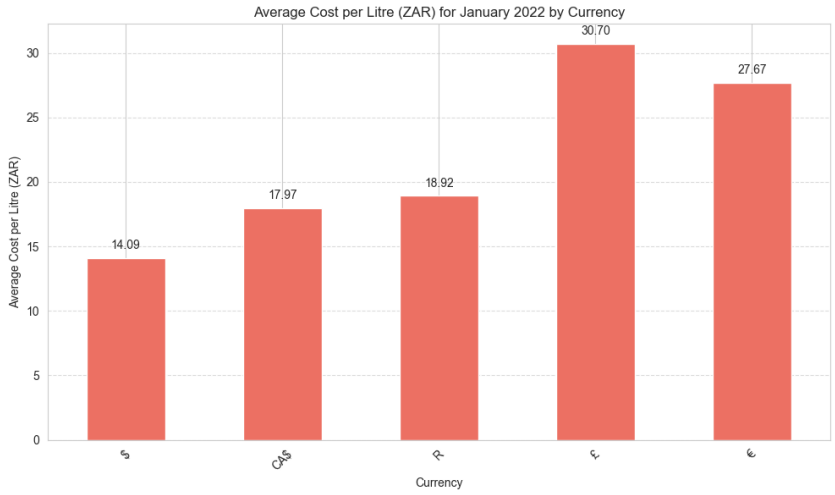

Figure 44: Average cost per litre for January 2022 in Rands

The exchange rate is a dominant factor in the differences in fuel costs per litre across countries when converted to Rands. Stronger currencies like the GBP and Euro (Figure 44) result in higher apparent costs when converted to a weaker currency like the Rands, while the USD and CAD, although strong, show more moderate increases.

2. A basic rule for identifying missed fill-ups is to calculate the difference in odometer readings between consecutive fill-ups. If the difference exceeds 500 miles, it can be assumed that a fill-up was missed. This rule was applied by sorting the data by user_id and date_fueled to accurately calculate the odometer difference for each user's consecutive entries as seen in Figure 45.
   Estimated number of potential missed fill-ups: 327175

	user_id	date_fueled	odometer	odometer_diff	potential_missed_fill_up
515881	1000007	2021-02-07	104824.000	NaN	False
836141	100002	2011-12-19	73864.000	NaN	False
826075	100002	2012-01-03	74605.000	741.000	True
826908	100002	2012-01-05	75000.000	395.000	False
845790	100002	2012-01-19	76322.000	1322.000	True
850442	100002	2012-01-23	76733.000	411.000	False
844515	100002	2012-02-12	78909.000	2176.000	True
853110	100002	2012-02-13	79370.000	461.000	False
827734	100002	2012-03-08	82317.000	2947.000	True

Figure 45: Potential missed fill ups of a user

3. Country with the largest average distance per tank: € with 522.5636636344653 km.
   he higher average distance travelled per tank in Europe compared to the USA, Canada, and South Africa can be attributed to several factors. European vehicles are generally more fuel-efficient due to stricter emissions regulations and higher fuel prices, leading to greater distances per tank despite often having smaller fuel tanks. Additionally, European cars are typically smaller and more efficient than the larger vehicles prevalent in North America, Figure 46.

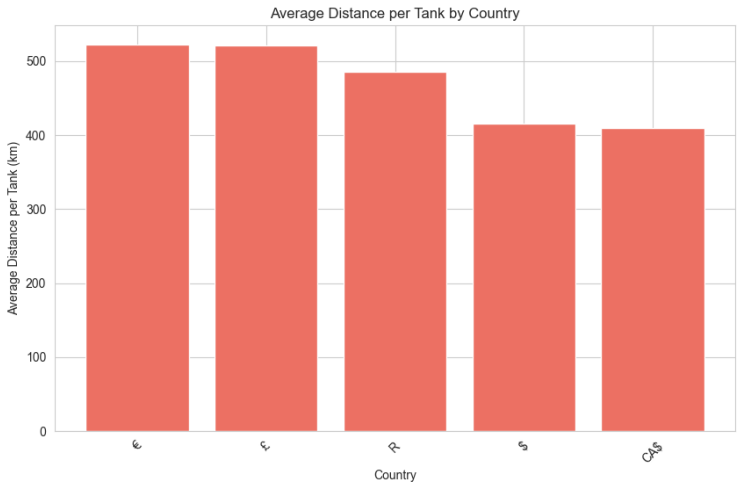

Figure 46: Average distance per tank by country

4. The graph suggests that newer vehicles generally tend to drive further distances between fill-ups, particularly from the 1980s onwards. However, there are significant outliers in earlier years, which may reflect specific models with unique characteristics rather than a broad trend. The consistent increase in average distance per tank in more recent years aligns with advancements in vehicle technology, such as improved fuel efficiency and possibly larger fuel tanks, Figure 47.

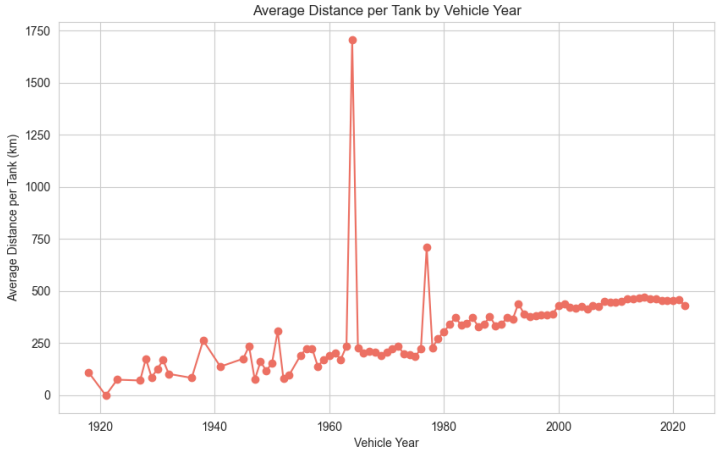

Figure 47: Average distance per tank by Vehicle Year

5. The average fuel efficiency values for these top 5 vehicles in South Africa appear realistic. They reflect expected consumption ranges for various vehicle types, from standard to performance-oriented models. While individual models may vary, these figures are consistent with typical fuel efficiency for the vehicles described, Figure 48.

	make	average_fuel_efficiency
0	bmw	10.007
1	ford	11.865
2	nissan	11.280
3	toyota	11.473
4	volkswagen	9.649

Figure 48: Fuel Efficiency of the top 5 makes of cars in South Africa

6. Figure 49

	make	average_fuel_efficiency	currency
0	peugeot	7.399	$
1	citroen	7.622	$
2	renault	7.985	$
3	fiat	8.521	$
4	opel	8.642	$
5	hyundai	7.259	£
6	kia	7.433	£
7	honda	7.787	£
8	toyota	7.937	£
9	mini	7.976	£
10	honda	7.503	€
11	citroen	7.700	€
12	suzuki	7.758	€
13	mini	8.010	€
14	hyundai	8.549	€
15	renault	7.454	CA$
16	fiat	7.587	CA$
17	citroen	8.016	CA$
18	alfa_romeo	8.551	CA$
19	mini	9.118	CA$
20	honda	7.626	R
21	fiat	7.747	R
22	renault	7.760	R
23	citroen	8.099	R
24	hyundai	8.269	R

Figure 49: Top 5 fuel efficiency car makes in the top 5 countries

7. The Figure 50 illustrates the seasonal fuel efficiency of the top 5 Canadian vehicles: Renault, Fiat, Citroen, Alfa Romeo, and Mini. Renault and Mini show improved fuel efficiency in colder seasons, with Renault being most efficient in Winter and Mini in Spring. Fiat's efficiency is consistent but drops in Winter, indicating higher fuel consumption. Citroen is most efficient in Autumn but lacks data for Spring and Winter. Alfa Romeo's fuel efficiency fluctuates, with the best performance in Summer and the worst in Spring. Overall, the graph reveals how different vehicles respond to seasonal changes, with some performing better in colder or warmer conditions.

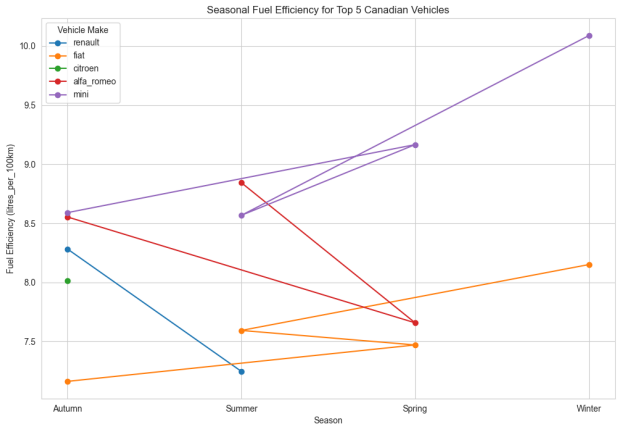

Figure 50: Fuel Efficiency over the seasons for the top 5 Canadian Vehicles

8. The matrix indicates that there is a moderate relationship between the litres_per_100km and km_driven, but the other variables (age_of_vehicle and model) do not show any significant correlations with either litres_per_100km and km_driven. This suggests that fuel efficiency is more likely to be influenced by how much a vehicle is driven rather than its age or model as seen in Figure 51.

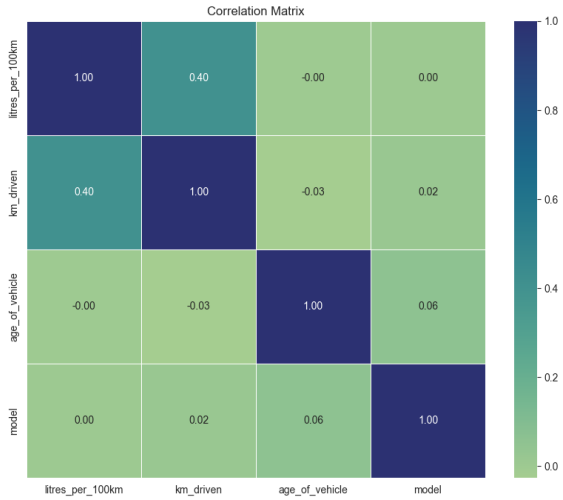

Figure 51: Correlation Matrix

9. The Random Forest model reveals that mpg is the only feature with a significant importance score, indicating it is the primary driver in predicting the target variable. This is consistent with the strong correlation between mpg and litres_per_100km, both of which measure fuel efficiency. All other features show negligible importance, suggesting they do not meaningfully contribute to the model's predictions. Additionally, while some features like miles and km_driven have moderate positive correlations with litres_per_100km, most others exhibit weak or negative correlations, indicating limited relevance to fuel efficiency which can be seen in Figure 52 and 53.

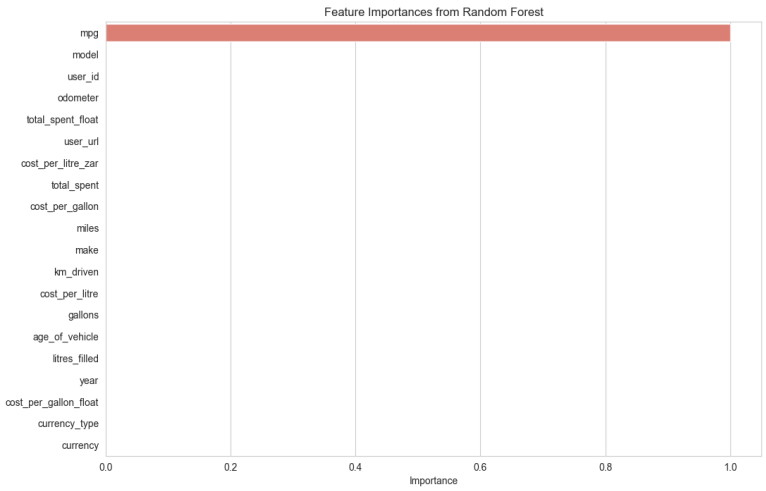

Figure 52: Feature Importance's from Random Forest

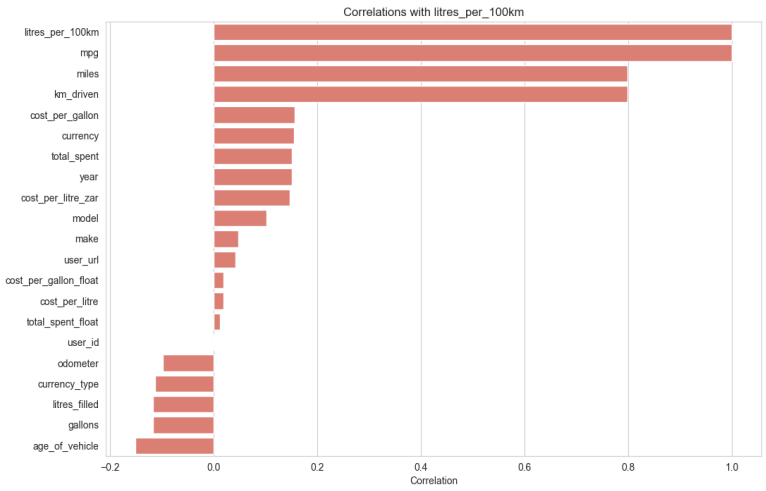

Figure 53: Correlation with respect to 100kms per litres

4.3 Fuel Usage in SA

1. Done in code which can be found in jupyter Notebook
2. Figure 54

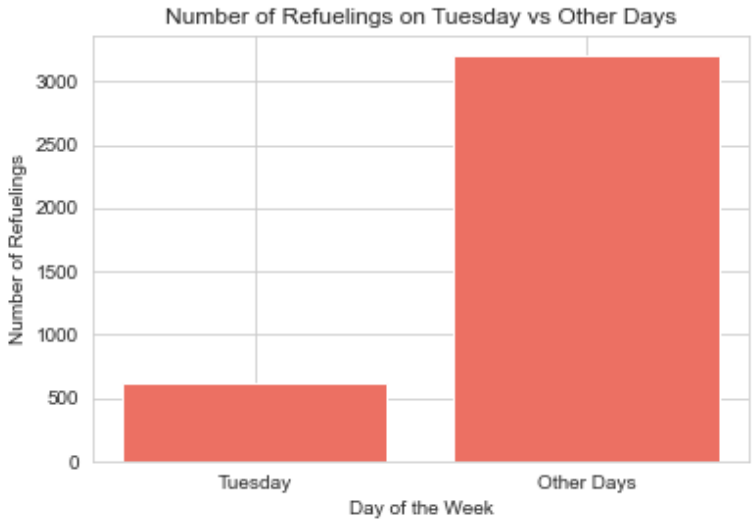

Figure 54: Cost per Litre for South African Drivers

3. Refuelings on Tuesday: 617
   Refuelings on Monday: 599
   Refuelings on Wednesday: 558
   Refuelings on Thursday: 549
   Refuelings on Friday: 546
   Refuelings on Saturday: 430
   Refuelings on Sunday: 526
   More people refueled on a Tuesday compared to any other day.

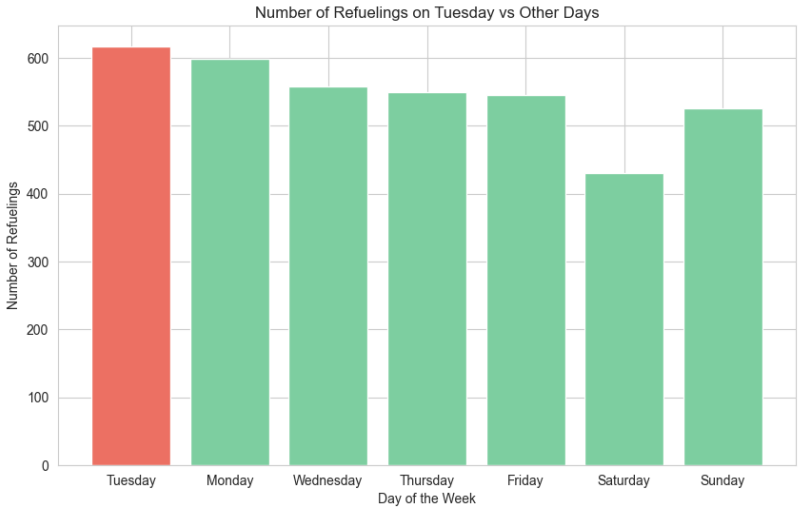

Figure 55: Refuelling of a car on Tuesday vs other days of the week

4. Done in code which can be found in jupyter notebook
5. Figure 56

	date_fueled	cost_per_litre	price_change_indicator
0	2011-10-05	10.260	
1	2011-11-01	10.239	down
2	2011-11-01	10.239	no change
3	2011-11-02	10.599	up
4	2011-11-02	10.260	down
...	...	...	...
271	2022-04-05	19.509	down
272	2022-04-05	21.599	up
273	2022-04-05	19.329	down
274	2022-04-05	20.841	up
275	2022-04-06	21.411	up

Figure 56: Indicator showing whether price goes up or down

6. No, more people do not refuel on the first Wednesday of the month when the price goes down (Figure 57).

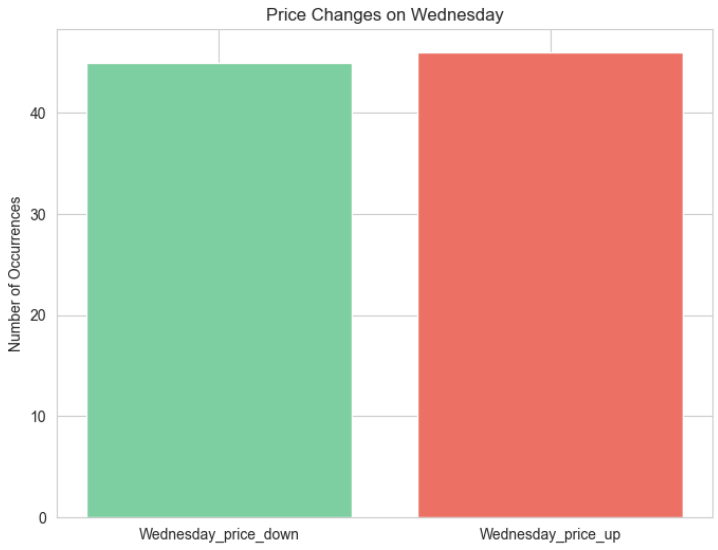

Figure 57: Price Change on Wednesday

7. Yes, more people refuel on a Tuesday when the price goes up (Figure 58).

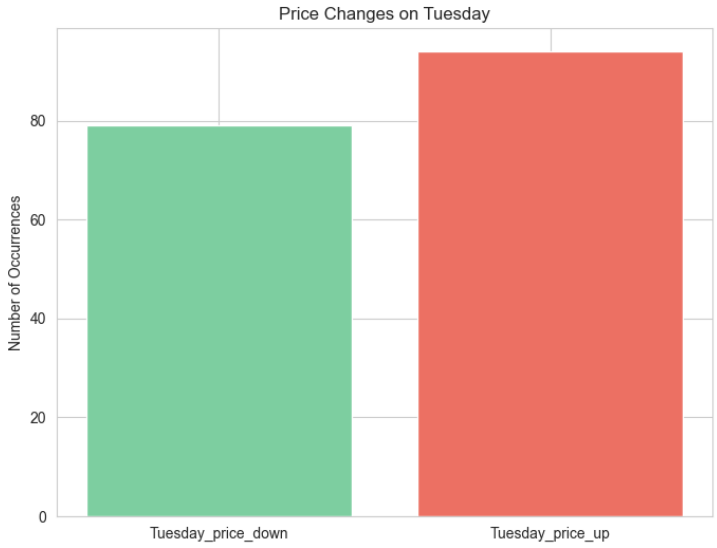

Figure 58: Price Changes on Tuesday

References

[1] X Rates. 2024. URL: https://www.x-rates.com/average/?from=USD&to=ZAR&amount=1&year=2022#google_vignette (visited on 08/18/2024).