IDVE Discussion questions

# 1.1.5 Plot the distribution of fueling dates and comment on the results.

On a scale of 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

27th to 31st – likely attributed to February not having these days and hence an overall lower contribution.

A graph showing the amount of fuel in the fueling months

Description automatically generated with medium confidence

On a scale of months of the year

The 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 aircon in cars during this time.

A graph showing the amount of fuel in the month

Description automatically generated

On a scale years 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.

A graph of a number of fuel prices

Description automatically generated

# 1.2.4 Plot the distributions and comment on the distributions

The distribution is extremely skewed to the right due to the outliers and the space availability compacting the bins.

A group of graphs with numbers

Description automatically generated

Graphing 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.

From here the data is less skewed to the right with gallons refuelled closer to a standard distribution. It can be observed that 0 miles per gallon and 0 miles driven has a high frequency which can be attributed to a type of outlier, without which the respective graphs also look closer to a standard distribution.

A group of graphs showing different types of data

Description automatically generated

# 1.2.5 Compute the statistical description of the columns: mean, standard deviation, max, min, most frequent, and quartiles. Do these results make sense?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 |

**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 milage 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 milage.

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 milage 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 mis entered 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.

# 4.1.2 Discuss what was done with the outliers

Pending answers

# 4.2.1 Look at the difference in cost per litre per country for January 2022 - use the average currency conversion rate to Rands (quote your values and source). Are there any notable differences? Discuss reasons why this may/may not be the case.

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/>

A graph with blue bars

Description automatically generated

# 4.2.2 Looking at the odometer readings, find examples of where users have missed logging a fill-up. Give a basic rule for identifying this, and estimate how many there are in the dataset.

A basic approach can be used by estimating how far a user is expected to travel on a full tank of fuel. Based on the average total miles between refills being 283 miles calculated earlier, and the 75th percentile users travelling 342 miles in electric cars, it is safe to assume that 400 miles is an upper bound on how far a user can travel on a full tank of fuel.

With this in mind, it is possible to estimate whether a user has missed logging a fill up by comparing logging dates and the odometer readings for each unique user id. If the odometer reading differs by more than 400 miles it is relatively safe to assume they missed a log. Further granularity could be done by looking at how much they fuelled but it is not possible to know whether it was a small top up or a fuel from empty.

# 4.2.3 Plot the average distance (in km) per tank per country. Which country has the largest average distance? Provide some explanations for why this might be the case.

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

A graph with blue bars

Description automatically generated

# 4.2.4 Take the top 5 most popular vehicles in SA (ie, those with currency set to R). Compute their fuel efficiency and discuss whether these values are realistic.

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%207.3l%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>

A screenshot of a computer

Description automatically generated

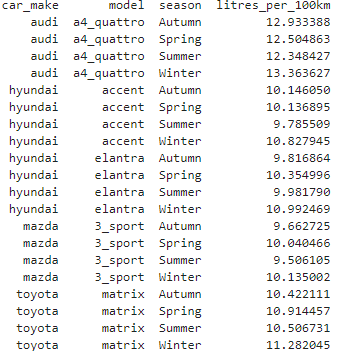
# 4.2.6 Which vehicles are the most fuel efficient in each country?

A screenshot of a computer screen

Description automatically generated

# 4.2.7 Plot the difference in fuel efficiency for the top 5 Canadian vehicles between seasons. Would you expect to see big differences, and do you see them?

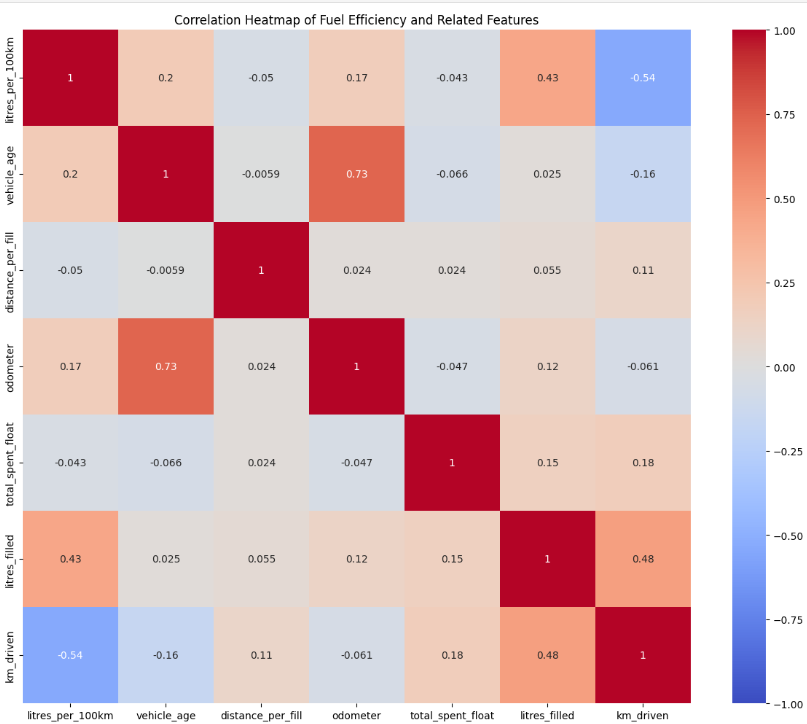
Expecting small difference between the seasons, in particularly between summer and winter and this is shown as winter consumption is increased compared to the same vehicle in summer for all vehicles.



A graph of different colored bars

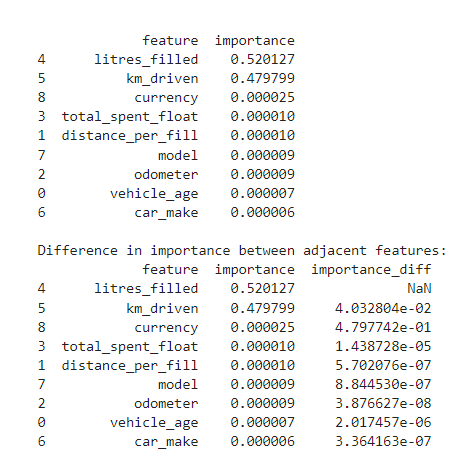
Description automatically generated

# 4.2.8 Show the correlations between fuel efficiency and other features.



# 4.2.9 Use a random forest to get a list of the most important variables. How different are they from each other, and how do these relate to the variables from the correlations above?

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.



A graph with text on it

Description automatically generated