# **An investigation into the effects of transportation and oil consumption throughout the countries in the world using regression and statistical analysis.**

## Executive summary

The purpose of this study is to investigate what the impact of transportation has on oil consumption throughout the world. Throughout this assessment we will be looking at the hypotheses:

- Is there a relationship between the cars, trucks and buses per 1000 and the oil consumption per person datasets?. This should indicate just what impact transportation has on oil consumption.
- Is there a relationship between the three HDI (Human Development Index) categories of oil consumption per person. This indicates whether or not there is a difference in the level of oil consumption in different HDI ranking countries.

The first hypothesis was addressed using linear regression modelling techniques. Here, we ascertained that there was a slight positive correlation present despite outliers using the Kendall Tau test. The second hypothesis was answered using statistical investigations. Here, we were able to deduce that there was a difference in the way that oil is consumed in countries with different HDI score categories using the Kruskal Wallis test. Overall, it can be stated that there does appear to be a link with the amount of transportation per 1000 and the amount of oil consumption per person. But there are clearly many other factors that must influence the amount of oil consumption by looking at the results.

## Introduction.

Oil consumption is known to cause global warming which in turn melts the polar ice caps. As the ice caps melt, less light is reflected away from the planet resulting in a negative feedback loop that causes the planet's temperature to rise above uninhabitable conditions. "*In the next twenty years, average global temperatures will increase by more than 2°C*" (Harari, 2018, p77). Because of this, it would be beneficial to understand what impact transportation has on the amount of oil consumption throughout the world, as well as understanding how HDI has an effect too (ClientEarth, 2022).

## Data

The data used for this investigation was extracted from the Gap minder website and loaded into R studio using the read.csv() function. The three datasets that I have used are listed in Table 1. The str() and summary() functions provided the dimensions of the datasets.

Table 1: Summary statistics of the chosen dataset

| Dataset | Minimum | Median | Mean | Max | Standard deviation | Number of Variables (years) | Number of observations (countries) |
|---|---|---|---|---|---|---|---|
| Oil consumption | 0.026 | 1.265 | 1.650 | 10.50 | 1.738 | 56 | 79 |
| Car, buses and trucks per 1000 | 2.26 | 344.50 | 352.38 | 820.0 | 231.08 | 7 | 161 |
| HDI | 0.4960 | 0.8245 | 0.8055 | 0.936 | 0.1015 | 31 | 189 |

Pre-processing:

For this dataset, I only want to include observations which are present in all three datasets, as the linear regression and statistical investigations should be all encompassing and relevant throughout this assessment. To do this, I will use the dplyr and tidyverse packages to 'left_join()' the datasets together. I want to use the year that is most recent. In the 'Cars, trucks & busses' dataset, the latest data that is available is from 2007. Therefore, I will extract the years 2007 from each of the datasets by using the 'select()' function. I will then use the 'omit_na()' function to remove any observations with no recorded data in all three categories.

Finally, I need to create a categorical variable for the statistical investigation section. To do this I need to find the appropriate level sizes so I ran the 'summary()' function on the filtered HDI dataset:

Table 2: A table showing the summary of the filtered HDI dataset.

| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
|---|---|---|---|---|---|
| 0.4960 | 0.7465 | 0.8245 | 0.8055 | 0.8898 | 0.9360 |

I want my categories to represent this spread of data shown in Table 2. I will be using three categorical levels of 'Low', 'Medium' and 'High' to describe the HDI. The 1st Quartile should be between the 'Low' and 'Medium' category, and the 3rd Quartile should be between the 'Medium' and 'High' categories. The categorical variables were created using the 'cut()' function. When running the 'head()' function, my data looks as follows (**Appendix 1**):

Table 3: A table showing the variables used  for this assessment with the data type

| Country (Nominal data) | Oil per person per year (tonnes) (Ratio data) | Cars, trucks & buses per 1000 persons (ratio data) | HDI (Ratio data) | Categorical description of HDI (Ordinal data) |
|---|---|---|---|---|

Data limitations

The most recent data that was available was from 2007. It would have been better to use data that is more contemporaneous and relevant to what we are trying to understand with environmental concerns now. A lot of data is missing when combining the datasets, especially from lower ranking HDI countries. It would have been better to represent these lower HDI countries. Not much statistical power will be generated from these datasets as the number of observations is relatively low (66 observations and 5 variables).

Linear Regression between two variables

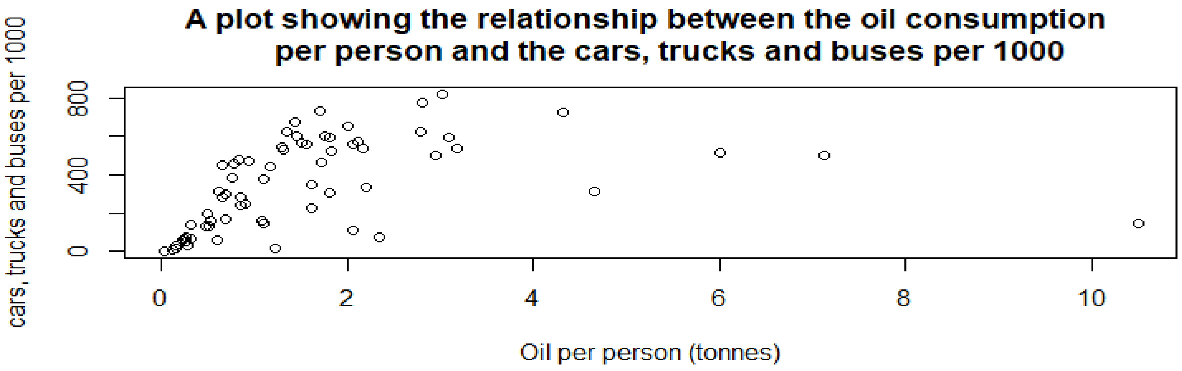
In this assessment, we want to investigate if there is a relationship between the variables 'Oil consumption per year' and ' the Cars, trucks & buses per 1000 persons'. We also want to measure how strong this relationship is.

Hypothesis

H0: Correlation coefficient of the oil consumption per person and the cars, trucks and buses per 1000 is equal to zero

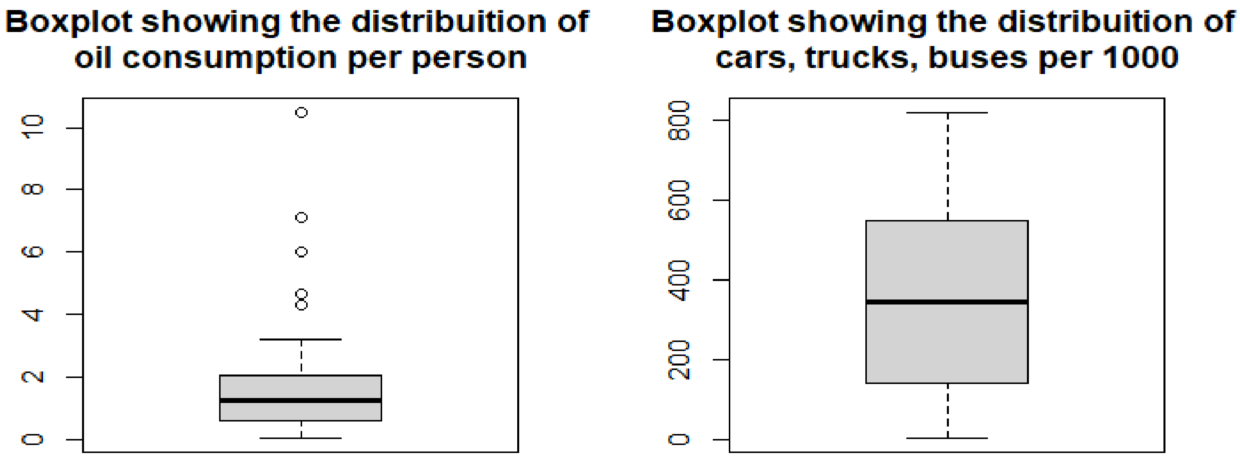
H1: Correlation coefficient of the oil consumption per person and the cars, trucks and buses per 1000 is not equal to zero

Figure 1:



From Figure 1, the plot appears to have a few outliers that could affect the correlation between the two variables. There does seem to be some weak positive correlation if we were to remove the outliers present towards the bottom right of the plot. However, when doing research into the outlier, I can find evidence that the data is not anomalous. The outlier is Singapore and is known to use a high amount of oil per person. "*Singapore relies on fossil fuels more than any other country"* (Power engineering international. 2021). Also there appears to have been restrictions put in place *"Singapore has announced that it will freeze the number of vehicles allowed in the country"* Fansided. (2017). For these reasons, I don't think it would be appropriate to discard the outlier in this dataset.

Figure 2:



Boxplot showing the distribuition of oil consumption per person

Boxplot showing the distribuition of cars, trucks, buses per 1000

Figure 2 shows that there are 5 outliers present in the 'oil consumption' dataset. The whiskers for the 'cars, trucks and buses' dataset are uneven suggesting the data may be skewed here. To be able to use a parametric test, some data transformation will most likely be required. I performed the summary() function for the linear model in R Studio. The R squared value is quite low at 0.1229. *"The larger the R-squared is, the more variability is explained by the linear regression model."*(Mathworks. (N.D.). This low R squared value makes sense as in figure 3, we can see visually that the residuals are quite far from the linear regression line.

The population slope and population R squared is significantly different to zero as the p value (0.00391) is less than the significance level of 0.05. This means that there is a good chance that the slope from the linear model is statistically significant. Here we are rejecting the null hypothesis that the slope is zero.

Table 4: The summary of the residuals indicate that they are not really symmetrical about zero:

| Minimum | 1st quartile | Median | 3rd quartile | Maximum |
|---|---|---|---|---|
| -615.81 | -175.97 | -18.33 | 185.36 | 403.77 |

The Y intercept for the linear model for this data is 275.49, we are also given the slope value of 46.6 which states that for every unit of oil consumption, the number of cars, buses and Trucks per 1000 increases by 46.6.

Figure 3:

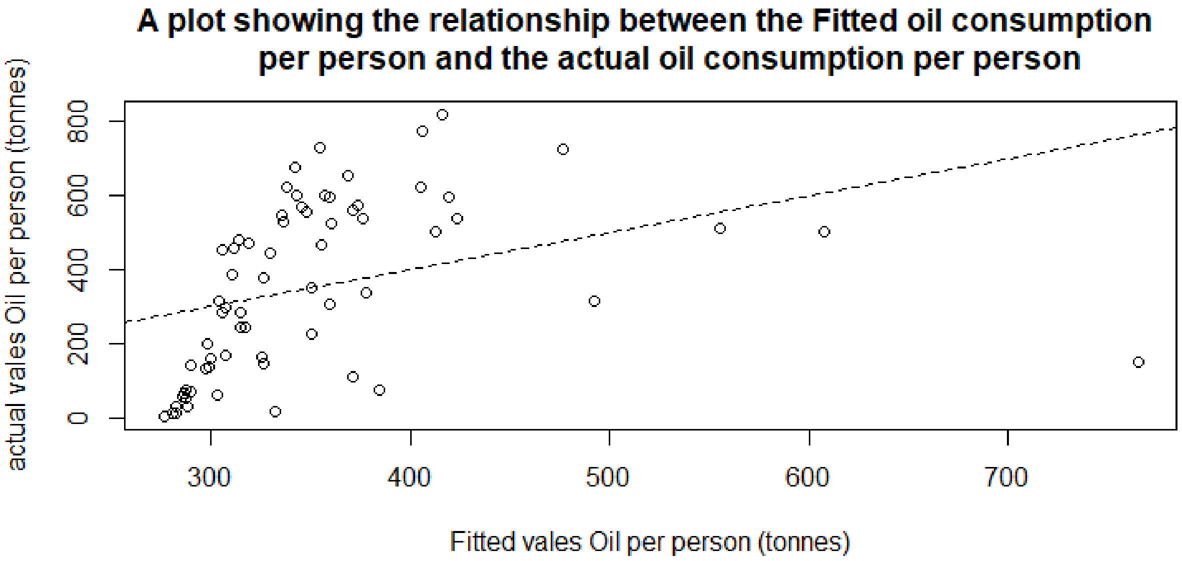
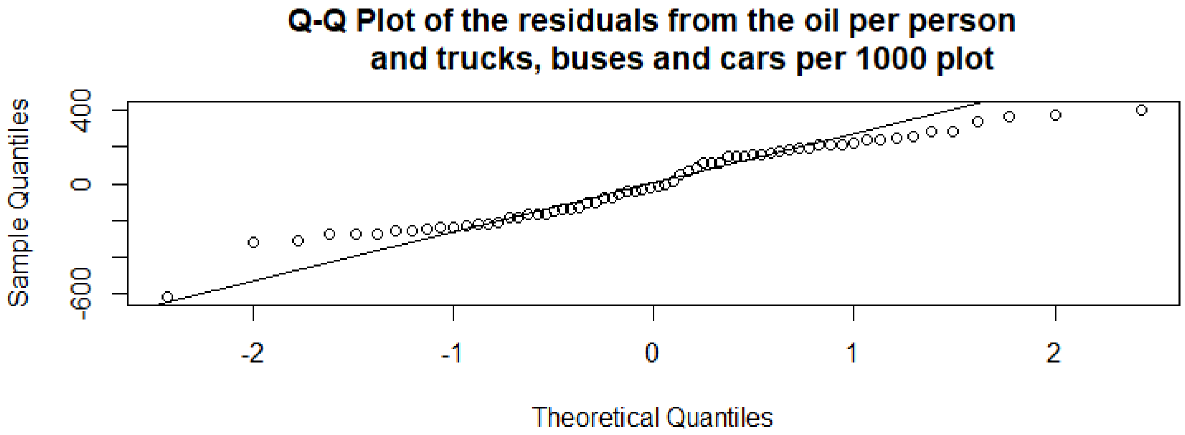


Figure 3 shows a plot of the relationship between the two variables. I used the lm() function in r studio. This was then converted to fitted values by using the predict() function. The residuals of the above linear model were obtained using the resid() function. Even if we could remove the outliers, the plot still seems to contain points which show heteroscedasticity. The variance seems somewhat unequal.

In order to satisfy the trend of the linear model, the residuals should:

(1) be normally distributed

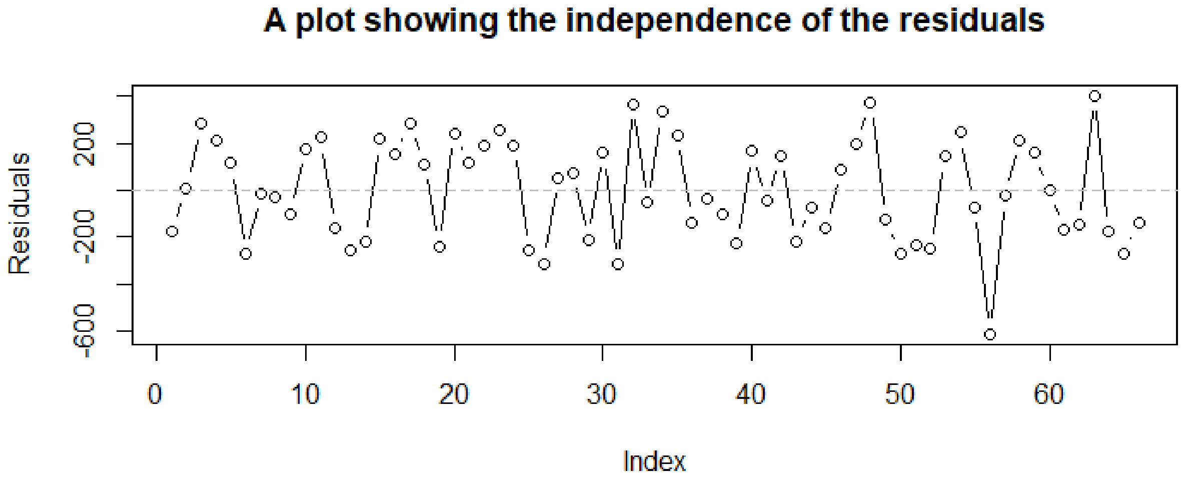
Figure 4:



The above Q-Q Plot indicates that the points are aligned along the line of normality but there are some signs of tailing.

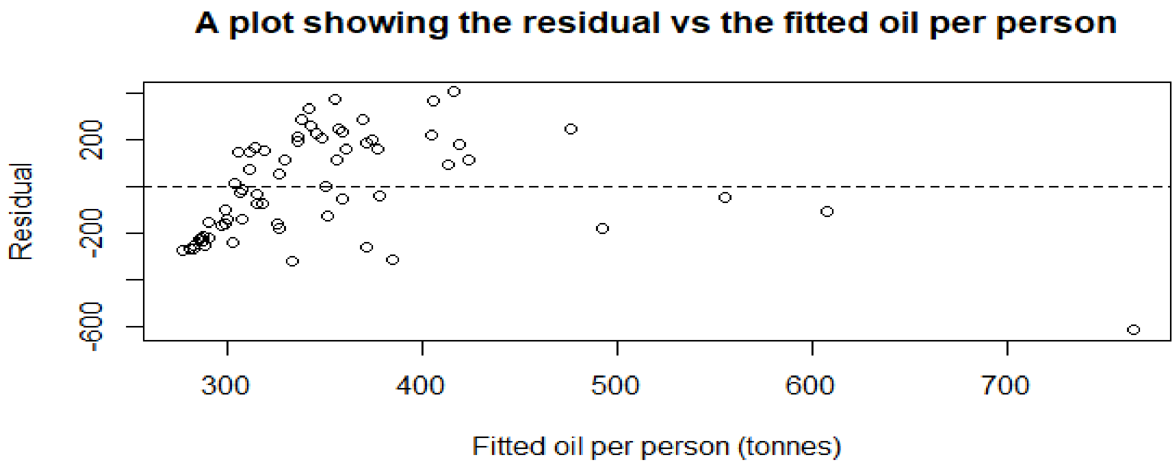
Shapiro-wilks (Your data teacher. 2022) (Towards Data Science, 2019) test shows that the residuals are normally distributed with a p value of 0.03426, which means we can reject the null hypothesis that there is no normal distribution in the data. Less than 0.05 significance level indicates normal distribution.

(2) The residuals should be independent:
Figure 5:



There appears to be little pattern in the sequence of residuals when we see Figure 5. The Durbin-Watson test (Investopedia. 2021) was performed to detect for any autocorrelation. The DW result was 1.7934 which states there is no first order autocorrelation, as the test statistic lies between 1.5 and 2.5. The p value is 0.2029 which means we can accept the null hypothesis that there is no relationship between the residuals (significance level, 0.05).

Figure 6: (3) Have constant variance



It is clear from Figure 6, that the third assumption has been violated as the residuals become negative as the oil per person increases. Because of this, the data was transformed using square root and log transformation (square root was most successful). However, the data still would not comply with this third assumption. The Bartlett test was performed to see if the variances were equal. The null hypothesis states that the variances between the samples are equal. However, the null can be rejected as the p value on the transformed data (square root) is 2.2e-16 (less than 0.05 significance level). This means the variances are unequal confirming the third assumption has been violated. Because of this, I will be using a non-parametric statistical test to determine the correlation coefficient. It would have been preferable to use a parametric test as there is more statistical power. However, I will be using the Kendall Tau test. Kendall Tau result = 0.5156323.

As the result is above 0.5 we can reject the null hypothesis. This means that there is a weak positive relationship between the oil per person consumption and the cars, trucks and buses per 1000 variables despite numerous outliers.

Statistical investigation on a hypothesis:

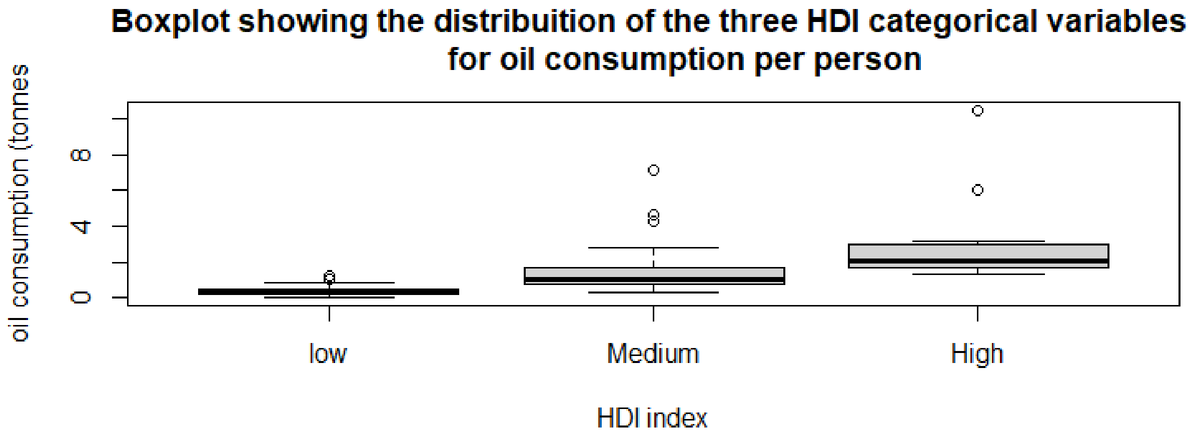
For this section of the assessment I will be looking at the relationship of the oil consumption per person and the Human development Index. I created categorical levels of the human development index in three groups (low is <0.75, medium is 75<>0.88, high is 0.88<>1.0). For this section, I would like to compare the above relationship between these three groups of countries. Because of this, it would be preferable to run an analysis of variance (ANOVA) test. This single test compares the variances between each group while only accounting for a 5% error level.

Null hypothesis: There is no relationship between the three HDI levels of ‘low’, ‘medium’ and ‘high’ and the amount of oil consumption per person.
Alternate hypothesis: There is a relationship between the three HDI levels of ‘low’, ‘medium’ and ‘high’ and the amount of oil consumption per person.

Assumptions for an ANOVA test:

1) Each group of data is a simple random sample from its respective population. We assume this to be the case.
2) Each group of the data ‘low’, ‘medium’ and ‘high’ are normally distributed.
3) All populations have the same standard deviation.

Figure 7:



From the box plot in figure 7, we can see that the spread of the data is generally inconsistent which would mean the data does not meet the assumption for ANOVA testing. We can also perform the Bartlett test of homogeneity of variances to confirm this. When performing this test in r studio, we find the p value (3.1262e-09) is below 0.05 significance level. This means that we can reject the null hypothesis that the variance is equal across all samples.

As this is the case we are unable to use ANOVA testing for our statistical test. Another form of statistical test that could be appropriate to use is the Welchs ANOVA. “*It is an alternative to the Classic ANOVA and can be used even if your data violates the assumption of homogeneity of variances*.” (Statistics How To, N.D.) (Dalgaard, 2008, p134). Here we can test the other assumptions to check if we are able to continue with this type of statistical test.

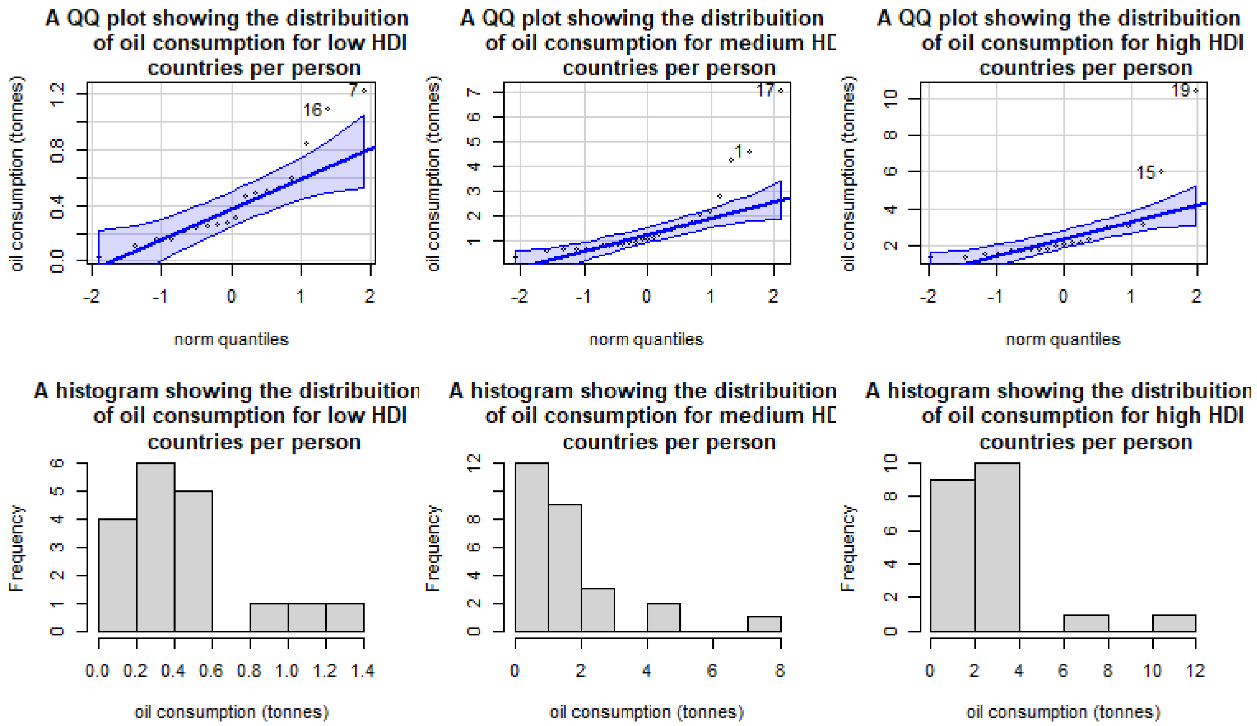
Our data should conform to the next assumption to perform our Welchs ANOVA test. Here I will look for normal distribution within our datasets. I will show this visually using histograms from the base r package and qqPlot() function from the ‘car’ package. When exploring the datasets in

Figure 8, we can see that there is not much normal distribution in the data. I will now see if it would be possible to use transformations to provide normalisations for my data. For this I will be using the Shapiro Wilkes test  (Your data teacher, 2022) (Towards Data Science, 2019) which tests for normality in the datasets.

Table 5: Table summarising the Shapiro-Wilkes test results before and after transformations:

| HDI category | Original | Square root transform | Log transform |
|---|---|---|---|
| high | 1.71e-06 | 5.329e-05 | 0.003418 |
| Medium | 3.7e-06 | 0.0007796 | 0.2168 |
| low | 0.01908 | 0.6751 | 0.2206 |

Figure 8: QQ Plots and Histograms showing the distribution of oil consumption per person by the three categorical HDI variables



Although table 5 does show that the log transformations fall above the 0.05 significance level test for the ‘medium’ and ‘low’ categories, it does not for the ‘high’ category. Because of this, we are unable to reject the null hypothesis that the data is not normally distributed once transformed. Because we are unable to conform to the normalisation assumption, we must use another type of statistical test.

We can use a type of non parametric testing in this case called the Kruskal Wallis test (Verzani, J. 2014). There is less statistical power here than in a parametric test, but this test could still provide some insights into our dataset.

Let us first confirm that we meet the statistical assumptions for the Kruskal Wallis test. The first assumption is that there are at least three random variables from each sample. And the next is that there are at least five observations for each sample. Using the count() function we can confirm that our data meets this criteria (low = 18, medium = 27, high = 21). The significance level of this test will be 5%. This is the chance that we will perform a type I error. When we perform the Kruskal-Wallis in R studio we get a p value of 2.709e-09 which is below the 0.05 significance level. Because of this, we can reject the null hypothesis that the medians of each group are the same.

We can break this down further by performing the Games-Howell test (Datanovia, N.D.) which is a non parametric approach to show the statistical significance between our groups.

Table 6: A table showing the results of the Games-Howell test for the oil consumption per person between the three HDI categories.

| HDI Category | P Value |
|---|---|
| Low and medium | 0.001 |
| Low and high | 0.000173 |
| Medium and low | 0.13 |

Concluding remarks

From Table 6 we can see that there is no statistical evidence to support that there is any association between oil consumption per person between the low and medium and low and high HDI categories. However, between the medium and low HDI categories, there is evidence that we could accept the null hypothesis that the medians between these two groups are the same.

It would be expected for there to be an increase in oil consumption in countries with a higher HDI index, as there is more disposable income to use for transport. Table 6 does seem to suggest that this is the case particularly in the ‘high’ HDI ranked countries. There also seems to be slight positive correlation between the transport use and oil consumption, but the result of the Kendall Tau showed only slight positive correlation. It would be better if we were able to obtain more data points to account for the outliers and also enable us to transform the data to use parametric testing to perform tests with more statistical power.

References:

ClientEarth. (2022). *Fossil fuels and Climate Change: The Facts.* https://www.clientearth.org/latest/latest-updates/stories/fossil-fuels-and-climate-change-the-facts/

Dalgaard, P. (2008). *Introductory Statistics with R.* Springer New York.

Datanovia. (N.D.). *Games Howell Post Hoc Tests.* https://rpkgs.datanovia.com/rstatix/reference/games_howell_test.html#:~:text=The%20Games%2DHowell%20method%20is,using%20Welch's%20degree%20of%20freedom.

Fansided. (2017). *Singapore limiting the number of cars on its roads.* https://artofgears.com/2017/10/29/singapore-limiting-number-cars-roads/

Geeks for geeks. (2020). *Bartlett's Test in R Programming.* https://www.geeksforgeeks.org/bartletts-test-in-r-programming/

Harari, Y, N. (2018). *21 Lessons for the 21st Century.* Random House.

Investopedia. (2021). *Durbin Watson Test: What It Is in Statistics with Examples.* https://www.investopedia.com/terms/d/durbin-watson-statistic.asp

Mathworks. (N.D.). *Coefficient of Determination (R-squared).* https://www.mathworks.com/help/stats/coefficient-of-determination-r-squared.html

Power engineering international. (2021). *Singapore relies on fossil fuels more than any country - study* https://www.powerengineeringint.com/coal-fired/singapore-relies-on-fossil-fuels-more-than-any-other-country-study/

Statistics How To. (N.D.). *Welch's ANOVA Definitions, Assumptions.* https://www.statisticshowto.com/welchs-anova/

Towards Data Science. (2019). *6 Ways to Test for Normal Distribution - Which one to use?* https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93

Verzani, J. (2014). *Using R for introductory Statistics.* CRC Press.

Your data teacher. (2022). *A Practical Introduction to the Shapiro-Wilk test for Normality.* https://www.yourdatateacher.com/2022/11/07/a-practical-introduction-to-the-shapiro-wilk-test-for-normality/

Appendices:

Appendix 1:

| Country (Nominal data) | Oil per person per year (Ratio data) | Cars, trucks & buses per 1000 persons (ratio data) | HDI (Ratio data) | Categorical description of HDI (Ordinal data) |
|---|---|---|---|---|
| United Arab Emirates | 4.650 | 313 | 0.819 | Medium |
| Argentina | 0.609 | 314 | 0.817 | Medium |
| Australia | 2.010 | 653 | 0.912 | High |
| Austria | 1.560 | 557 | 0.890 | High |
| Belgium | 3.180 | 539 | 0.907 | High |
| Bangladesh | 0.026 | 2.26 | 0.528 | Low |