Mark Baber - 17076749

Faculty Of Computing and Engineering Sciences

Assessment Cover Sheet and Feedback Form 2020-21

Module Code:	Module Title:	Module Team:
MS4S08	Applied Statistics for Data Science	Filippo Cavallari, Penny Holborn, Angelica Pachon
Assessment Title and Tasks: Set Tasks - not-time constrained 1		Assessment No. 1
Date Set: 11-Oct-2020 12:00	Submission Date: 24-Nov-2020 21:00	Return Date: 23-Dec-2020 21:00

IT IS YOUR RESPONSIBILITY TO KEEP RECORDS OF ALL WORK SUBMITTED

Marking and Assessment
This assignment will be marked out of 100% This assignment contributes to 50% of the total module marks.
Learning Outcomes to be assessed (as specified in the validated module descriptor https://icis.southwales.ac.uk/): 1) Learning Outcome 1: To understand the concepts and theory of statistical analysis, and explain the wider context of their value in Data Science. 2) Learning Outcome 2: Determine and use statistical techniques to assess practical situations and interpret real-world complex data.
Provisional mark only: subject to change and / or confirmation by the Assessment Board

Mark Baber - 17076749

Contents

Mark Baber - 17076749

1 – Introduction

This report will look at the foundation called World Car Free Day (WCFD) which is a movement to encourage people around the world to give up a car for one day (September 22nd) and see how beneficial it could be. The report will also look at additional data such as government datasets on the areas around transport and pollution, to see if small differences can have a big impact.

The report will be broken up into multiple sections from exploring WCFD, looking at datasets and using hypothesis testing and some statistical analysis which will be followed by a conclusion on the matter of WCFD.

2 - World Car Free Day 2020 (WCFD)

World Car Free Day has been running globally since 2000 (World Carfree Day, 2009) and is a movement to help ease up the use of cars for one day a year (September 22nd), this movement has allowed people to see the benefits of having empty roads with room for cycling, skating and walking (Awareness Days, 2020).

The original idea was brought to light back in 1997 (World Carfree Day, 2009) within Lyon, France, by protestors representing 50 groups from 21 countries (World Carfree Day, 2009) under the organisation European Youth For Action (EYFA). These protests were raising awareness around the use of cars within urban areas and ways to improve green city planning.

So why join in with this event you might ask, well to start with it would be beneficial to see less vehicles on the road which could improve the air quality of the area. Showing people an alternative to cars and having the roads be a lot safer depending on the area. This will be explored within the next section.

Very good introduction.

3 – Exploratory Data Analysis (EDA)

After exploring what WCFD is all about in the last section, this section will look to explore some datasets around the air quality, greenhouse gases and vehicles registered by fuel types. This section will use the exploratory data analysis on these datasets to see what stands out.

3.1 - Greenhouse Gases (GGs)

This section will explore a dataset from StatsWales which measured the Emissions of Greenhouse Gases by sector. Before exploring the data, it is important to clean some of this data to make the next steps easier. What is the time frame, how big is the data etc?

To clean this data, it would be best practice to have the year going down the left hand column and the industry going across the columns, this should allow for an easier time importing the data into SAS Studio. After getting the data into SAS Studio, here we can run a summary analysis and the results were as follows:

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
Agriculture	Agriculture	5689.46	422.4118801	5118.56	6391.50	24	0.2676964	-1.2134139
Business	Business	11039.47	2572.30	7891.08	16644.42	24	0.9925980	-0.1192864
Energy Supply	Energy Supply	16790.55	2356.68	11455.05	21229.73	24	-0.3725606	-0.0109979
Industrial Process	Industrial Process	2514.73	542.0361040	1439.72	3319.44	24	-0.4966827	-0.7133737
Land Use Change	Land Use Change	-469.8161689	85.8083550	-607.6379650	-295.0838420	24	0.3331598	-0.2189693
Public	Public	463.5565065	138.7175718	307.6078310	769.1970780	24	0.9772693	0.2198283
Residential	Residential	4592.48	637.4749495	3590.59	5539.11	24	-0.3342250	-1.2030472
Transport	Transport	6440.40	249.6841919	6018.78	6901.35	24	0.0453253	-0.8798184
Waste Management	Waste Management	2321.92	851.5478941	1244.13	3447.68	24	-0.0198377	-1.7479535
Exports	Exports	550.0392023	75.0574433	437.1167500	696.3460560	24	0.4890822	-0.6708075

(Fig 1 – Summary Analysis of Greenhouse Gases.)

With the figure 1 above, we can see a few things which stand out.

- The means are very different for most variables.
- All variables have the same sample size.
- 4 variables have a negative skewness.
- Only 1 variable has a positive kurtosis.

3.2 - Energy Consumption

UK/Year?

This section will explore a dataset from the Office for National Statistics (ONS) which measured the energy consumption of each industry. This data will be explored and see what parts of the data stand out and if any of it would need to be cleaned further.

First of all the data would need to be cleaned a little, but just to rearrange the rows and columns so it would be easily imported into SAS Studio and also to change the variable names as the industry sectors had really long names. These have been changed and can be seen in appendix 6.2.

After doing a summary analysis within SAS, the output was as follows:

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
AFF	AFF	2.2448276	0.2063046	1.9000000	2.6000000	29	0.0267419	-1.3146976
M&Q	M&Q	7.5586207	1.2451383	5.7000000	9.5000000	29	0.1502842	-1.4097375
M	M	36.4413793	6.8601602	24.9000000	44.6000000	29	-0.4840110	-1.4025994
EGSA	EGSA	50.7241379	9.3110018	26.6000000	61.5000000	29	-1.3407337	1.2586235
WSWMR	WSWMR	0.8862069	0.1846312	0.7000000	1.6000000	29	2.0818206	7.0614881
C	C	4.0620690	0.4850727	3.2000000	4.8000000	29	-0.3440610	-0.9133193
WRTRMV	WRTRMV	4.5827586	0.1605103	4.3000000	4.9000000	29	0.3011164	-0.7198300
T&S	T&S	27.7758621	3.0259184	22.2000000	33.7000000	29	-0.2906443	-0.3457649
A&F	A&F	1.4586207	0.1118585	1.2000000	1.7000000	29	-0.2317092	0.0750590
I&C	I&C	0.4034483	0.0823007	0.3000000	0.5000000	29	-0.0662758	-1.5176605
FI	FI	0.1000000	0	0.1000000	0.1000000	29	.	.
R	R	0.3655172	0.0483725	0.3000000	0.4000000	29	-0.6890964	-1.6437247
PST	PST	0.8862069	0.1407230	0.7000000	1.1000000	29	0.0140434	-1.2937451
ASSA	ASSA	1.1413793	0.0627765	1.0000000	1.2000000	29	-0.5818704	-0.4934028
PAD	PAD	3.0379310	0.7692361	1.7000000	3.9000000	29	-0.6440221	-1.1829709
E	E	1.7068966	0.4208249	1.1000000	2.5000000	29	0.3002874	-1.1170625
HHSW	HHSW	2.0310345	0.2406887	1.4000000	2.5000000	29	-0.6902199	0.5408158
AER	AER	0.5379310	0.1612757	0.3000000	0.8000000	29	0.3707680	-1.2305383
O	O	0.4482759	0.0508548	0.4000000	0.5000000	29	0.0728290	-2.1481481
AOH	AOH	0	0	0	0	29	.	.
CE	CE	56.9137931	3.6822668	49.6000000	62.5000000	29	-0.4798321	-0.9083952

(Fig 2 – Summary Analysis of Energy Consumption by Industry.)

Firstly, this is a much bigger table to explore, but will mainly focus on Transport whilst also looking at the bigger energy consumers within the same dataset.

- 2 datasets do not have a Std Dev, Skewness and Kurtosis but will be omitted.
- The means for the datasets are wildly different (which shows how much more energy consumption some industries use).
- There is a mix of positive and negative skewness on the data with the same for the kurtosis.

3.3 - Car Fuel Types

This section will explore a dataset around the number of cars registered in the UK, by their fuel type which was sourced from the Vehicle Licensing Statistics. Here are the results from running a summary analysis.

car-type=diesel

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
car-year	car-year	2006.50	7.6485293	1994.00	2019.00	26	0	-1.2000000
no-registered	no-registered	6710.02	3730.49	1576.20	12397.64	26	0.2280508	-1.3930830

car-type=gas

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
car-year	car-year	2006.50	7.6485293	1994.00	2019.00	26	0	-1.2000000
no-registered	no-registered	31.1115000	16.3595627	1.7880000	50.9500000	26	-0.5435983	-0.9801010

car-type=other

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
car-year	car-year	2006.50	7.6485293	1994.00	2019.00	26	0	-1.2000000
no-registered	no-registered	159.2811538	205.4060497	2.1100000	783.7500000	26	1.7965748	2.8270756

car-type=petrol

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
car-year	car-year	2006.50	7.6485293	1994.00	2019.00	26	0	-1.2000000
no-registered	no-registered	20174.92	1255.68	18348.09	21976.37	26	-0.0537793	-1.4655568

(Fig 3 – Summary Analysis of cars registered by fuel type.)

There are multiple things which stand out in with these two datasets.

- They all have the same sample size which is good.
- The means are very different for the fuel types, which is to be expected.
- The data is slightly skewed for all 4 fuel types, with 2 to the left and 2 to the right.

Car year is categorical, so does not make snese to incude in summary statistics.

4 - Data Analysis

After carrying out the EDA for the 3 datasets, this section will do a deeper dive into data analysis for the datasets. For each dataset, there will be 3 different types of data analysis which in turn will help to give a better insight into if WCFD would benefit the environment.

4.1 – Greenhouse Gases (Ggs)

This dataset will be used to carry out a correlation analysis between the different types of GGs. The first step would be to state the main hypothesis, then moving onto a few assumptions.

H0: There is no linear relationship between Transport and the other variables.
H1: There is a linear relationship between Transport and the other variables.

in relation to GG emissions?

Here we can run a goodness of fit test for normal distribution:

H0: The data is not normally distributed.
H1: The data is normally distributed.

These are the wrong way around

First is the test for normality, here are the p-values for the goodness of fit tests.

Agriculture	>0.150
Business	<0.010
Energy Supply	>0.150
Industrial Process	0.135
Land Use Change	>0.150
Public	>0.150
Residential	>0.150
Transport	>0.150
Waste Management	0.136
Exports	>0.150

Seeing as there is a mix of p-values from this test, we will go ahead and assume the distribution is not normal. Here we would reject the alternative hypothesis and accept the null hypothesis.

Now that we know it is non-parametric, we can move to a non-parametric correlation analysis. Here let us see if there is a correlation between Transport and the other variables, after running a correlation analysis the results were as follows.

Spearman Correlation Coefficients, N = 24
Prob > |r| under H0: Rho=0

	Agriculture	Business	Energy Supply	Industrial Process	Land Use Change	Waste Management	Exports	Public	Residential
Transport	0.35189	0.54925	-0.07438	0.21526	0.02566	0.60157	-0.15094	0.59635	0.60230
Transport	0.0917	0.0054	0.7298	0.3124	0.9053	0.0019	0.4814	0.0021	0.0018

(Fig 4 – Spearman Correlation between the different types of Greenhouse Gases)

Here we can see a few things:

- Transport has 2 negative correlations (Energy Supply & Exports).
- The strongest correlation is with Residential, followed by Waste Management (WM).

The main results which should be brought to attention are Residential and WM with a very strong significant relationship with a p-value of 0.0018 and 0.0019, which shows there is a significant relationship between them at a 5% level. This would conclude that we would reject the null hypothesis and accept the alternative hypothesis.

This shows that there is a relationship between transport and multiple areas of industry, which shows that by having less vehicles could help to improve air quality of several areas.

Good Scatter plots would be useful here.

4.2 – Energy Consumption

This dataset will be used to carry out a lower one-tailed one sample t-test. This test will be used to see if there has been a 40% reduction within greenhouse gases, whilst this is still smaller than the stated percent of 95% reduction by 2050 (Messenger, 2020), it would be interesting to see if there is any difference so far.

H0: μ = < 40
H1: $\mu \neq$ < 40

When first carrying out a T-Test, the first thing to check is Normality. This was done within SAS and the output was as follows:

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.95742	Pr < W	0.2835
Kolmogorov-Smirnov	D	0.10597	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.078053	Pr > W-Sq	0.2182
Anderson-Darling	A-Sq	0.484945	Pr > A-Sq	0.2184

(Fig 5 – Tests for Normality (Energy Consumption))

As the P-value is > 0.05 this would suggest the data is normally distributed. This would allow us to move on to the next step and see the p-value from the t-test. Here are the results from the t-test with a mu of < 40.

N	Mean	Std Dev	Std Err	Minimum	Maximum
29	27.7759	3.0259	0.5619	22.2000	33.7000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
27.7759	-Infty	28.7317	3.0259	2.4013	4.0924

DF	t Value	Pr < t
28	-21.76	<.0001

(Fig 6 – Tests of Location)

With a p-value of <.0001, this would suggest there is a significant relationship at the 5% level. So, we should reject the null hypothesis and conclude that there is a significant relationship between the levels of energy used by the transport sector around a mu of < 40%.

So what does this mean?

4.3 – Car Fuel Types

This dataset will be used to explore an ANOVA analysis between the different car fuel types.

H0: There is no significant difference between the car fuel types.
H1: There is a significant difference between the car fuel types.

To start with exploring this hypothesis, there are several assumptions which must be met first. The first would be to explore the data, which was done in the EDA section above.

H0: The samples came from a normally distributed population.
H1: At least one sample did not come from a normally distributed population.

Next there will be a goodness of fit test for normal distribution, the results were as follows.

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.2602756	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.9418482	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	11.4266772	Pr > A-Sq	<0.005

Good

(Fig 7 – Test for Normality – Cars registered.)

Here we can see that the p-value is <0.010 which would suggest, the dataset is not normally distributed. So, we would reject the null hypothesis and accept the alternative. This would allow us to move onto a Nonparametric One-Way ANOVA.

Whilst running the ANOVA analysis, the results were as follows:

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
88.2563	3	<.0001

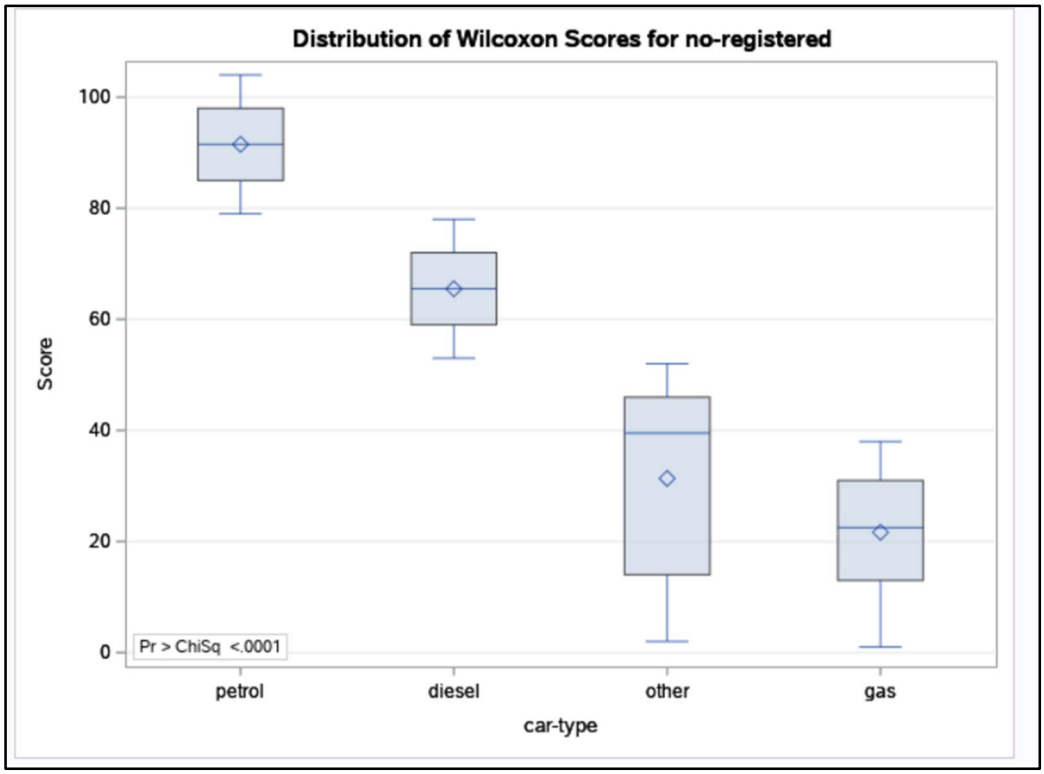
(Fig 8 – Kruskal-Wallis test for car fuel types.)

Here we can see a p-value of <.0001 which is significant at the 1% level, suggesting there is a significant relationship between the fuel types. This can be explored further with the pairwise two-sides multiple comparison analysis.

Pairwise Two-Sided Multiple Comparison Analysis			
Dwass, Steel, Critchlow-Fligner Method			
Variable: no-registered			
car-type	Wilcoxon Z	DSCF Value	Pr > DSCF
petrol vs. diesel	6.1858	8.7480	<.0001
petrol vs. other	6.1858	8.7480	<.0001
petrol vs. gas	6.1858	8.7480	<.0001
diesel vs. other	6.1858	8.7480	<.0001
diesel vs. gas	6.1858	8.7480	<.0001
other vs. gas	2.3059	3.2611	0.0966

(Fig 9 – Pairwise Two-Sides Multiple Comparison Analysis)

Here we can see there is a significant relationship between all groups of fuel types except for other (electric cars) and gas cars. This would suggest there is not enough evidence to accept the null hypothesis and would accept the alternative. This is further shown within the boxplots below.



(Fig 10 – Distribution of Wilcoxon Scores for No of Registered Cars by fuel type.)

Very good.

4 – Conclusion

Throughout this report, there have been some research into WCFD and seen what their movement is all about, which was to create safe spaces within cities which would be carless just for one day (Awareness Days, 2020). With this is mind, we carried out some EDA on data around greenhouse gases and vehicles to see if there is any statistical truth to the cause.

Whilst exploring the datasets, we can see there have been a massive burst in the different types of greenhouse gases, the number of registered cars and the energy consumption by industry. This could have been affected by the population growth, even if at a slow rate(Coates, Tanna, & Scott-Allen, 2019).

To conclude on these findings, there are a few things which should be brought up. Whilst the data is mainly local to Great Britain and the rest of the UK, with more accurate data there could be a further analysis concluded. The findings within this report show there are several areas which could be improved upon especially the greenhouse gases sections, whilst travel is a high pollutant it is dwarfed in comparison to the electricity, gas, steam, and air conditioning supply and consumer expenditure.

Whilst the energy consumptions by the industry is increasing, there could also be greener energy being generated with the modern times. This could also be the same for the car fuel types, with so many cars being registered, more and more of those cars are becoming greener with battery powered and a mix of battery and petrol.

Good discussion

5 - References

Awareness Days. (2020, October 26). *World Car Free Day 2020*. Retrieved from Awareness Day: https://www.awarenessdays.com/awareness-days-calendar/world-car-free-day-2020/

Department for Business, Energy & Industrial Strategy. (2020). *2019 UK Greenhouse Gas Emissions, Final figures.* London: National Statistics.

Messenger, S. (2020, September 3). *Climate change: Call to ban new roads as part of challenge*. Retrieved from BBC News: https://www.bbc.co.uk/news/uk-wales-54002555

SAS Institute Inc. (2020, November 16). *SAS Welcome*. Retrieved from SAS OnDemand for Academics Dashboard: https://welcome.oda.sas.com/home

StatsWales. (2020, November 16). *Emissions of Greenhouse Gases by Year*. Retrieved from StatsWales: https://statswales.gov.wales/Catalogue/Environment-and-Countryside/Greenhouse-Gas/emissionsofgreenhousegases-by-year

World Carfree Day. (2009, September 21). *Press Release 2009*. Retrieved from World Carfree Day - Press Release 2009 : https://www.worldcarfree.net/wcfd/wcd-pr2009.php

6 - Appendix

6.1 – Datasets

6.1.1 – Greenhouse Gases

Emissions of Greenhouse Gases by Year - https://statswales.gov.wales/Catalogue/Environment-and-Countryside/Greenhouse-Gas/emissionsofgreenhousegases-by-year

6.1.2 – Energy Consumption

Energy Consumption in Millions of tonnes of oil equivalent (Mtoe): By source and industry section, 1990 to 2018 - https://www.ons.gov.uk/economy/environmentalaccounts/datasets/ukenvironmentalaccountstotalenergyconsumptionbyindustry

6.1.3 – UK CO2 Fuel Type

UK territorial carbon dioxide emissions by fuel type (MtCO2) - https://www.gov.uk/government/publications/uk-greenhouse-gas-emissions-explanatory-notes

6.1.4 – Vehicles Registered

Vehicle Licensing Statistics - VEH0150 - https://www.gov.uk/government/statistical-data-sets/all-vehicles-veh01

6.2 – Variable Names

Agriculture, forestry, and fishing	AFF
Mining and quarrying	M&Q
Manufacturing	M
Electricity, gas, steam, and air conditioning supply	EGSA
Water supply; sewerage, waste management and remediation activities	WSWMR
Construction	C
Wholesale and retail trade; repair of motor vehicles and motorcycles	WRTRMV
Transport and storage	T&S
Accommodation and food services	A&F
Information and communication	I&C
Financial and insurance activities	FI
Real estate activities	R
Professional, scientific, and technical activities	PST
Administrative and support service activities	ASSA
Public administration and defence; compulsory social security	PAD
Education	E
Human health and social work activities	HHSW
Arts, entertainment, and recreation	AER
Other service activities	O
Activities of households as employers; undifferentiated goods and services-producing activities of households for own use	AOH
Consumer expenditure	CE