Fossil fuels and HDI: Carbon emissions as an indicator of human human standards of living.

Laura Vodden

Executive Summary

Carbon dioxide (CO_2) emissions are produced by burning fossil fuels, primarily as a source of energy (Olivier et al., 2012). It has been demonstrated that countries with the highest CO_2 emissions have the wealthiest people and greates economic growth (Hickel, 2020; Holtz-Eakin & Selden, 1995; Jorgenson et al., 2017), and that there is global inequality in the production of carbon emissions, reflecting this (Chancel & Piketty, 2015; Oxfam, 2015).

While the wealth of a country is often expressed in terms of economic value, this is not the only measure of a country's wealth (Hickel, 2020). The Human Development Index (HDI) is a measure of average human wellbeing in terms of health (measured by life expectancy), education (measured by mean years of schooling) and standard of living (measured by gross national income per capita). The HDI is the mean of the normalized indices for each of the three dimensions (UNDP, 2020).

This report aims to investigate the possibility that economic growth and human wellbeing go hand in hand, by comparing CO_2 emissions with HDI and population data. To this end, the report will test:

1. Whether there is a difference between total CO_2 emissions in high and low HDI countries;
2. Whether CO_2 emissions per person increase with HDI; and
3. Whether a country's population can be used to its total CO_2 emissions.

In order to conduct this investigation, data was cleaned and analysed in RStudio, using the following methods to answer these three questions, respectively:

1. A two-sample Wilcoxon test
2. A two-sample T-test
3. A linear regression analysis

The report finds that there is a difference between CO_2 emissions in high and low HDI countries, CO_2 emissions do generally increase with HDI, and that population is not a good predictor for CO_2 emissions and therefore HDI.

The conclusion drawn from this report is that a CO_2 emissions (indicative of a country's economic wealth) do tend to increase with the wellbeing of its people, as indicated by HDI. This suggests that economic wealth and human development go hand in hand, since the countries with the highest HDIs also had the highest CO_2 emissions, however the countries producing the highest emissions were not necessarily the most populous, indicating that population is not a significant contributor to either HDI or CO_2 emissions.

Introduction

The Human Development Index (HDI) is an indicator of a country's level of development, expressed as a unitless average of three dimensions: health, education and income (Wolff et al., 2011). A country's development index can be categorised as low (0.0 to 0.5), medium (0.5 to 0.8) and high (0.8 to 1.0).

The HDI can be used to demonstrate global inequality in wealth, living standards and even carbon emissions (Chancel & Piketty, 2015; Oxfam, 2015). Motivated by an interest in how carbon emissions can highlight the uneven distribution of wealth, this report aims to use global environmental data from 2014 (the most recent Gapminder data available) to test whether there is a difference between total CO_2 emissions in high and low HDI countries, whether CO_2 emissions per person increase with HDI, and whether a country's population can be used to its total CO_2 emissions.

Data

The software used in the analysis is R version 1.2.5033 (Orange Blossom).

The data used in the following analyses can be accessed from the Gapminder.org website. Specifically, environmental data comprises Total CO_2 emissions (measured in tonnes), and CO_2 emissions per person (measured in tonnes per person), These datasets provided numeric variables to be compared against Human Development Index (HDI) data, which can be accessed from the 'Society' section of the Gapminder website.

Initially, the data comprised a population size of up to 192 countries (depending on the data accessed). Prior to commencing the analyses, however, some data cleaning was required. Firstly, only data from the year 2014 were retained, using the **select()** function. 2014 data were chosen because this was the most recent year with data available for all variables of interest and therefore the most representative of the current global situation. Secondly, in order to more easily measure the numeric variables against HDI in a meaningful way, the HDI data were grouped into three categories: 'low', 'medium' and 'high', in accordance with those categories set by the UN (Hickel, 2020; Wolff et al., 2011).The following code achieved this:

```
HDI_Data$HDI[HDI_Data$HDI < 0.5] = "Low"
HDI_Data$HDI[HDI_Data$HDI < 0.8] = "Medium"
HDI_Data$HDI[HDI_Data$HDI < 1.0] = "High"
```

A new dataframe was compiled, using **left_join()**, comprising the 2014 data for the HDI data as well as the environmental data relating to CO_2 emissions. Missing data were removed using **na.omit()** and a new variable (population) was generated by dividing Total_CO_2 by CO_2 emissions per person:

```
Data_2014$Population <- (Data_2014$Total_CO2)*1000 /
Data_2014$CO2_pp
```

Finally, it must be noted that data available are not a random sample. The data include a population of all countries for which Gapminder data are available. In order to satisfy the requirements for any significance test, and to try to avoid any bias inherent in the original population data, a random sample of 100 countries was generated, using the **sample()** function. A selection of the resulting dataframe is shown in *Table 1*. Each of the three objectives covered in this report will refer to this random sample. *Table 2* provides further detail about the variables used. The full dataframe can be viewed in *Appendix 1*.

Table 1: A selection of the random sample used in this analysis, derived from 2014 Gapminder society and environment data.

country	HDI	CO2_pp	Total_CO2	Population
Malaysia	Medium	8.1300	2.43e+05	2.988930e+07
Montenegro	High	3.5300	2.21e+03	6.260623e+05
South Africa	Medium	8.9800	4.90e+05	5.456570e+07
Andorra	High	5.8300	4.62e+02	7.924528e+04
Eritrea	Low	0.2100	6.97e+02	3.319048e+06
Guinea-Bissau	Low	0.1600	2.71e+02	1.693750e+06
Lithuania	High	4.3200	1.28e+04	2.962963e+06
Portugal	High	4.3200	4.51e+04	1.043981e+07
Argentina	High	4.7900	2.04e+05	4.258873e+07
Croatia	High	3.9600	1.68e+04	4.242424e+06

Table 2: Summary table of variables used in this report.

Variable	Type	Sample Size	Mean	Median
Country	Categorical	100	N/A	N/A
HDI	Categorical	100	N/A	N/A
CO2_pp	Numeric	100		
Total_CO2	Numeric	100		
Population	Numeric	100		

Objective 1: Emission inequality: Test whether there is a difference between total CO_2 emissions in high and low HDI countries.

Using data from the categorical HDI and numeric Total_CO_2 variables, the first objective tests whether high HDI countries and low HDI countries have different total

CO_2 emissions, and therefore whether carbon emission inequality can be demonstrated.

Methods

Exploratory data analysis

A boxplot provided the first look at the difference between the distribution of data for high and low HDI countries. This was achieved using the **geom_boxplot()** function from the **ggplot2** package to plot HDI on the x axis and Total_CO_2 on the y axis. *Figure 1* shows that the range of total CO_2 emissions for high HDI countries is much greater than that for low HDI countries. The y axis limit has been set to 60000 to prevent data from being drowned out by outliers in the high HDI category.

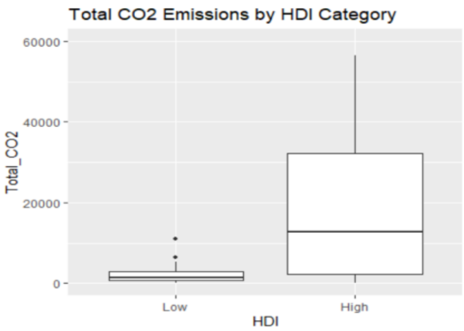

Figure 1: Exploratory boxplot showing range of total CO_2 emissions differs between low and high HDI countries.

Initially, the intention was to perform a two-sample t-test on these data. However, removal of the outliers resulted in removal of specifically high HDI countries from the dataset. This meant that any mean taken for the high HDI countries would not be representative of reality and introduced unwanted bias to the analysis. It was therefore decided that using a nonparametric test would provide a more accurate result, since for nonparametric tests, outliers and deviations from normality do not affect the ability of the test to be performed (Ramdas et al., 2017).

In this case, a two-sample Wilcoxon test was deemed appropriate, as it is the nonparametric equivalent of the two sample t-test, and tests for the median of the sample (Verzani, 2014), which was likely to be more descriptive of the sample than the mean in this case.

Hypotheses

The null hypothesis for this significance test was that the sample median for total CO_2 emissions for high HDI countries would be different from the sample median for total CO_2 emissions for low HDI countries. The alternative hypothesis was that there would be difference. This is expressed as:

$$H_0: M = 1310$$
$$H_A: M \neq 1310$$

Significance level of the test

The significance level (or the probability of committing a type-1 error) of this test was 5%. Therefore, $\alpha = 0.05$.

Statistical test and assumptions

A two-sample Wilcoxon test assumes that both datasets simple random samples, and that each has at least 11 cases, both of which were satisfied in this scenario. The test was run using the following code:

```
wilcox.test(Total_CO2 ~ HDI,
data=High_Low_HDI,alternative="two.sided",mu=0)
```

Results and Discussion

The results of the two sample Wilcoxon test showed that there is sufficient evidence to suggest there is a difference between the CO_2 emissions for high and low HDI countries (**p-value = 0.0001254**). In effect, carbon emission inequality has be demonstrated.

This result agrees with the preliminary analysis shown in *Figure 1*. The boxplot shows that there is a much different distribution of CO_2 data for high HDI countries and low HDI countries. This result agrees with existing studies showing that show there is global inequality in carbon emission production (Chancel & Piketty, 2015; Oxfam, 2015).

Objective 2: Test whether CO_2 emissions per person increase with HDI.

Objective 2 seeks to determine the *nature* of the difference between CO_2 emissions in high and low HDI countries, by testing if CO_2 emissions per person increase with HDI.

Methods

Exploratory data analysis

The first step in analysing the CO_2_pp data was to test for normality. The **qqnorm()** and **qqline()** functions were used, to visualise the distribution of CO_2_pp data, before and after removing outliers (*Figure 2*).

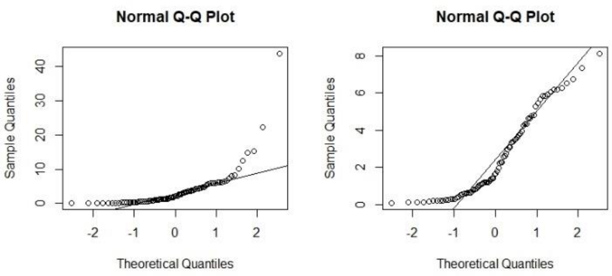

Figure 2: Exploratory QQplots showing approximately normal distribution of data before and after outlier removal.

The qqplot indicates that the data were already approximately normal, however, there were some outliers in the data, which, when removed, resulted in the normal distribution in *Figure 2* (right).

Hypotheses

The null hypothesis was that there would be no difference in CO_2 emissions between the two HDI categories. That is, the mean of CO_2 emissions is the same for both HDI categories. The alternative hypothesis was that the mean CO_2 emissions would be higher in the high HDI category compared to the low HDI category.

$$H_0: \mu Low = \mu High$$

$$H_A: \mu High > \mu Low$$

Significance level of the test

As for objective 1, the significance level is 5% ($\alpha = 0.05$).

Statistical test and assumptions

Solving this objective required a two-sample t-test. This test assumes that the data are a simple random sample and that observations come from a population that is normally distributed. Both of these requirements were satisfied; the data were from a random sample of 100 counties, and the qqplots confirmed the normality of the distribution.

In order to perform this test, it was necessary to subset the Data_2014 data using the **subset()** function, so that there were two datasets to compare: one for high HDI countries and one for low HDI countries. The following code achieved this:

```
Low_HDI <- subset(Data_2014_OM, HDI == "Low")
High_HDI <- subset(Data_2014_OM, HDI == "High")
```

Figure 3 shows that the individual datasets are both approximately normally distributed, high HDI more so than low HDI.

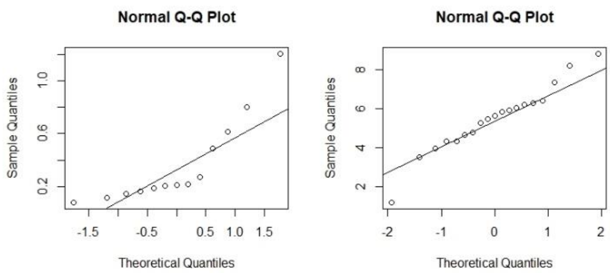

Figure 3: QQplots showing normal distribution of low HDI (left) and high HDI (right) countries.

Results and Discussion

The resulting t-statistic for the two-sample t-test was **10.52**, while the tstar value returned was **1.697**.

Because the tstatistic fell within the rejection region, the null hypothesis was rejected. Therefore, there is sufficient evidence to suggest that CO_2 emissions per person increases with HDI.

This is a logical result. Countries with access to greater amounts of power naturally produce more CO_2 emissions (Hickel, 2020). Since power is costly, countries that use the most power should be those that are wealthier (Chancel and Piketty, 2015) with good health, education and employment prospects, so countries with higher carbon emissions should generally have a higher HDI.

Objective 3: Test whether a country's population can be used to predict that country's CO_2 emissions.

The final objective this report aims to address is whether population can be used to predict a country's CO_2 emissions.

Methods

Exploratory data analysis

For this objective, population data and Total_CO_2 data were used. *Figure 4* shows the preliminary box and scatter plots for both variables, showing that there was a small number of extreme outliers.

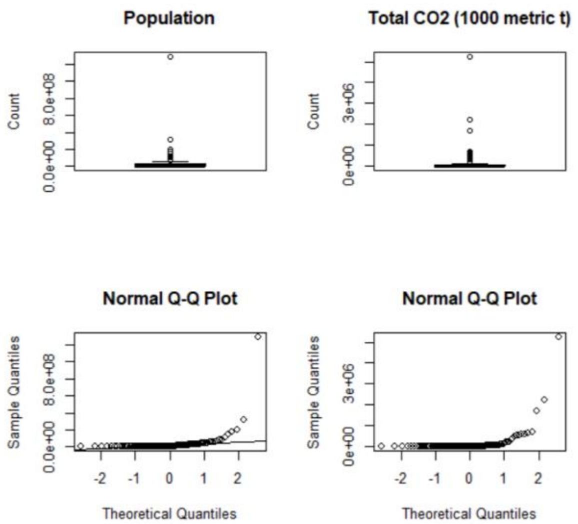

Figure 4: Top: Exploratory box plots for population and total CO_2 data; Bottom: Exploratory QQ plots corresponding to the boxplots.

The most severely outlying data (countries with populations higher than one billion) were removed from the sample, resulting in the plots shown in *Figure 5*.

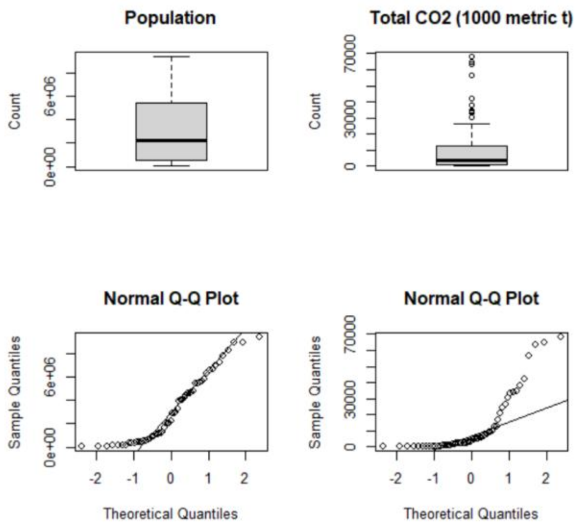

Figure 5: Top: Box plots for population and total CO_2 data; Bottom: QQ plots corresponding to the boxplots. Outliers have been removed.

Hypotheses

The hypothesis being tested is that population data can be used to predict a country's CO_2 emissions.

Estimate the parameters

Base R functions were used in the linear regression analysis. To find the regression coefficients, a linear model was created using **lm()** and output coefficients of the model were found using **summary()**, which can be viewed in *Appendix 4*.

The Rsquared value shows that only 43% of the variability in CO_2 emissions can be explained by this model.

Compute fitted values of statistical model

In order to compute the fitted values of the model, the **predict()** function was used, and the observed values were plotted against the fitted values. In *Figure 6*, the fitted values do not fall on the line, so are not identical to the actual values; in fact, there was little correspondence between the fitted values and the observed values of Total_CO_2.

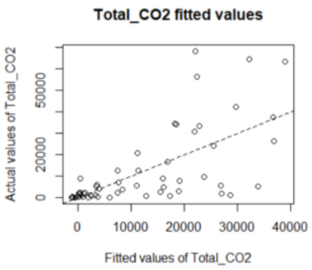

Figure 6: Fitted values (x axis) plotted against actual values for Total_CO_2.

Residuals

The residuals are the difference between the prediction and the observed values, using the **resid()** function. Residuals are assumed to have a normal distribution, constant variance and be approximately independent. The first of these is satisfied, as shown in *Figure 7*.

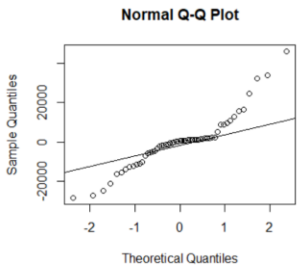

Figure 7: QQplot showing normal distribution of residuals for the majority of the data.

Figure 8 shows that there is no pattern to the residuals. It is reasonable, therefore, to assume that the assumption of constant variance is met. The residuals were plotted in the sequence in which they appeared in the data. Once again, there is no pattern to the data, so the final assumption is satisfied.

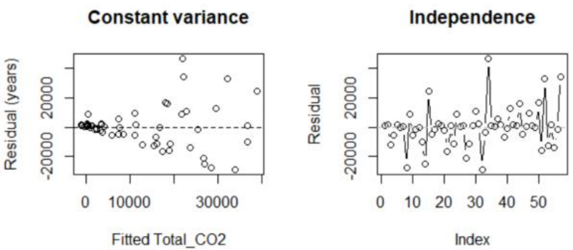

Figure 8: Left: Plot showing constant variance of the residuals; and right: Plot showing independence of the residuals.

Test the significance

The **summary()** function gave a t-statistic of **6.518**, an F-statistic of **42.18**, and a p-value of **2.318e-08**, with **55** degrees of freedom. The very low p-value indicates that there is some trend between population and CO_2 emissions.

Results and Discussion

Figure 9 shows the data plotted with confidence and prediction intervals.

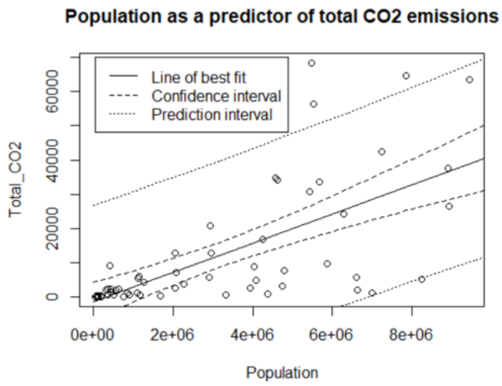

Figure 9: Line of best fit and confidence and prediction intervals show that while there is some correlation between population and total CO_2 emissions, population is not a good predictor for CO_2 emissions

The confidence bands show the uncertainty about the line of best fit, and the prediction bands reflect the uncertainty in the prediction of future observations. The prediction lines are quite wide, reducing confidence in the predictions derived from this linear model. Therefore, a country's population cannot be used as a very reliable predictor of its CO_2 emissions; however, because there is some correlation between the two variables, further investigation into this may be warranted.

Concluding Remarks

This analysis has demonstrated that there is global inequality in carbon emissions, linked to wealth and wellbeing. The study has also demonstrated that it is possible to link carbon emissions to HDI; there is sufficient evidence to suggest that as a country's carbon emissions increase, so too does its citizens' standard of living.

Reference List

Chancel, L., & Piketty, T. (2015). Carbon and inequality: From Kyoto to Paris. Trends in the global inequality of carbon emissions (1998-2013) & prospects for an equitable adaptation fund World Inequality Lab.

Gapminder. (2020). Retrieved August 1, 2020, from https://www.gapminder.org/data/

Hickel, J. (2020). The sustainable development index: Measuring the ecological efficiency of human development in the anthropocene. *Ecological Economics*, *167*, 106-331.

Holtz-Eakin, D., & Selden, T. M. (1995). Stoking the fires? CO2 emissions and economic growth. *Journal of Public Economics*, *57*(1), 85-101.

Jorgenson, A., Schor, J., Huang, X. (2017). Income inequality and carbon emissions in the United States: A state-level analysis, 1997–2012. *Ecological economics* 134, 40–48.

Olivier, J. G., Peters, J. A., & Janssens-Maenhout, G. (2012). Trends in global CO2 emissions 2012 report. Netherlands Environmental Assessment Agency. doi:10.2788/33777.

Oxfam. (2015). *Extreme carbon inequality*. https://oi-files-d8-prod.s3.eu-west-2.amazonaws.com/s3fs-public/file_attachments/mb-extreme-carbon-inequality-021215-en.pdf?cid=aff_affwd_donate_id78888&awc=5991_1597727317_d526157dc94d5dcf6b9fad7aeb4a1ff1

Ramdas, A., Trillos, N. G., & Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, *19*(2), 47.

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

United Nations Development Programme. (2020). *Human Development Index (HDI)*. Retrieved August 4, 2020, from http://hdr.undp.org/en/content/human-development-index-hdi

Verzani, J. (2014). Using r for introductory statistics. CRC Press.

Wolff, H., Chong, H., & Auffhammer, M. (2011). *Classification, detection and consequences of data error: Evidence from the Human Development Index.* Cornell University, School of Hospitality Administration. http://scholarship.sha.cornell.edu/articles/338

Appendix 1

Dataframe (100 case random sample

	country	HDI	CO2_pp	Total_CO2	Population
2	Albania	Medium	1.97	5720	2903553.299
3	Algeria	Medium	3.74	145000	38770053.48
4	Andorra	High	5.83	462	79245.28302
5	Angola	Medium	1.29	34800	26976744.19
7	Argentina	High	4.79	204000	42588726.51
9	Australia	High	15.3	361000	23594771.24
12	Bahamas	Medium	6.52	2420	371165.6442
16	Belarus	Medium	6.73	63500	9435364.042
18	Belize	Medium	1.4	495	353571.4286
19	Benin	Low	0.614	6320	10293159.61
23	Botswana	Medium	3.37	7030	2086053.412
24	Brazil	Medium	2.61	530000	203065134.1
25	Brunei	High	22.2	9110	410360.3604
26	Bulgaria	Medium	5.85	42400	7247863.248
31	Canada	High	15.1	537000	35562913.91
32	Cape Verde	Medium	0.948	491	517932.4895
35	Chile	High	4.65	82600	17763440.86
37	Colombia	Medium	1.79	84100	46983240.22
38	Comoros	Low	0.203	154	758620.6897
40	Congo, Rep.	Medium	0.653	3090	4732006.126
41	Costa Rica	Medium	1.62	7760	4790123.457
42	Cote d'Ivoire	Low	0.488	11000	22540983.61
43	Croatia	High	3.96	16800	4242424.242
44	Cuba	Medium	3.08	34800	11298701.3
45	Cyprus	High	5.26	6060	1152091.255
47	Denmark	High	5.91	33500	5668358.714
48	Djibouti	Low	0.804	722	898009.9502
51	Ecuador	Medium	2.75	43900	15963636.36
54	Equatorial Guinea	Medium	4.76	5350	1123949.58
55	Eritrea	Low	0.21	697	3319047.619
58	Fiji	Medium	1.35	1170	866666.6667
63	Georgia	Medium	2.23	8990	4031390.135
64	Germany	High	8.84	720000	81447963.8
66	Greece	High	6.29	67300	10699523.05
68	Guatemala	Medium	1.15	18300	15913043.48
70	Guinea-Bissau	Low	0.16	271	1693750
72	Haiti	Low	0.271	2860	10553505.54
75	Iceland	High	6.04	1980	327814.5695
76	India	Medium	1.73	2240000	1294797688
78	Iran	Medium	8.38	649000	77446300.72
80	Ireland	High	7.36	34100	4633152.174
81	Israel	High	8.23	64600	7849331.713
85	Jordan	Medium	2.97	26500	8922558.923
87	Kenya	Medium	0.306	14300	46732026.14
88	Kiribati	Medium	0.57	62.3	109298.2456
90	Kyrgyz Republic	Medium	1.64	9610	5859756.098
91	Lao	Medium	0.294	1950	6632653.061
93	Lebanon	Medium	3.84	24100	6276041.667
94	Lesotho	Low	1.21	2470	2041322.314
95	Liberia	Low	0.214	935	4369158.879
97	Liechtenstein	High	1.18	44	37288.13559
98	Lithuania	High	4.32	12800	2962962.963
102	Malaysia	Medium	8.13	243000	29889298.89
103	Maldives	Medium	3.07	1330	433224.7557
104	Mali	Low	0.0834	1410	16906474.82
105	Malta	High	5.46	2350	430402.9304
106	Mauritania	Medium	0.689	2710	3933236.575
107	Mauritius	Medium	3.36	4230	1258928.571
110	Moldova	Medium	1.21	4930	4074380.165
111	Mongolia	Medium	7.09	20800	2933709.45
112	Montenegro	High	3.53	2210	626062.3229
113	Morocco	Medium	1.75	59900	34228571.43
114	Mozambique	Low	0.321	8430	26261682.24
115	Myanmar	Medium	0.414	21600	52173913.04
116	Namibia	Medium	1.65	3760	2278787.879
119	New Zealand	High	7.59	34700	4571805.007
121	Niger	Low	0.111	2130	19189189.19
122	Nigeria	Medium	0.546	96300	176373626.4
131	Paraguay	Medium	0.864	5700	6597222.222
132	Peru	Medium	2.05	61700	30097560.98
133	Philippines	Medium	1.05	106000	100952381
135	Portugal	High	4.32	45100	10439814.81
137	Romania	Medium	3.49	70000	20057306.59
138	Russia	High	11.8	1710000	144915254.2
140	Samoa	Medium	1.03	198	192233.0097
141	Sao Tome and Principe	Medium	0.581	114	196213.4251
142	Saudi Arabia	High	19.4	601000	30979381.44
144	Serbia	Medium	4.23	37700	8912529.551
146	Sierra Leone	Low	0.187	1310	7005347.594
147	Singapore	High	10.2	56400	5529411.765
148	Slovak Republic	High	5.65	30700	5433628.319
149	Slovenia	High	6.2	12800	2064516.129
151	South Africa	Medium	8.98	490000	54565701.56
152	South Korea	High	11.6	587000	50603448.28
154	Spain	High	5	234000	46800000
158	St. Vincent and the Grenadines	Medium	1.92	209	108854.1667
160	Suriname	Medium	3.6	1990	552777.7778
161	Swaziland	Medium	1.1	1200	1090909.091
164	Syria	Medium	1.64	30700	18719512.2
165	Tajikistan	Medium	0.629	5190	8251192.369
166	Tanzania	Medium	0.231	11600	50216450.22
167	Thailand	Medium	4.62	316000	68398268.4
168	Timor-Leste	Medium	0.4	469	1172500
170	Tonga	Medium	1.2	121	100833.3333
172	Tunisia	Medium	2.61	28800	11034482.76
174	Turkmenistan	Medium	12.5	68400	5472000
175	Uganda	Low	0.142	5230	36830985.92
179	United States	High	16.5	5250000	318181818.2
183	Venezuela	Medium	6.17	185000	29983792.54
185	Yemen	Low	0.879	22700	25824800.91

Appendix 2

Objective 1 Wilcoxon test output

```
	Wilcoxon rank sum test with continuity correction

data:  Total_CO2 by HDI
W = 59, p-value = 0.0001254
alternative hypothesis: true location shift is not equal to 0
```

Appendix 3

Objective 2 t-test output

```
> degrees.freedom
[1] 32
> tstatistic
[1] 11.78699
> tstar
[1] 1.693889
```

Appendix 4

Objective 3 summary

```
Call:
lm(formula = Total_CO2 ~ Population, data = Data_2014_OM3)

Residuals:
   Min     1Q Median     3Q    Max 
-28724  -5392    546   1851  46336 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.269e+03  2.746e+03  -0.462    0.646    
Population   4.264e-03  6.542e-04   6.518 2.32e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13670 on 55 degrees of freedom
Multiple R-squared:  0.4358,	Adjusted R-squared:  0.4255 
F-statistic: 42.48 on 1 and 55 DF,  p-value: 2.318e-08
```