== Theme ==

The theme chosen for this project is sustainability and the factors leading to a sustainable society.

In this project, I would like to investigate the relationships between a country’s sustainability index and its factors. For example, if a country is classified as sustainable, which are the most important factors? Do all countries that are classified as sustainable share the same factors that make them sustainable? Also, we want to see if there are relationships between the factors. For example, do countries with high ratings for education automatically have high ratings for gender equality? Vice versa for the countries that are not sustainable.

== Data ==

In order to investigate, I first downloaded data from <http://www.ssfindex.com/> and <http://hdr.undp.org/en/data>

The first set of data contains data prepared by the Sustainable Society foundation. There is a lot of data, and the sustainability scores from many years are provided. For the purpose of this project, we will be looking at the sustainability scores from year 2012.

Here is a snapshot of how the data looks like.

[[File:Terence1.png|800px|center ]]<br>

All the scores in this dataset have been normalized such that the values are numerical values ranging from 1 to 10, with increasing number representing a better indicator for that variable. The dataset provides 33 variables but for the purpose of this project, I have selected the following 7 variables which I think are the most pertinent:

* Safe sanitation
* Clean air
* Gender equality
* Income distribution
* Good governance
* Consumption
* Employment

For the initial analysis, I have used the software High-D to come up with a simple parallel coordinates plot.

Here is how the initial parallel coordinates plot looks like:

[[File:Terence2.png|800px|center ]]<br>

== Initial Analysis ==

For some initial analysis, I have selected certain data ranges of the data to view their relationships. Firstly, I have selected countries with scores between 1 and 2 for safe sanitation. (A) Secondly, I have selected countries with clean air scores of above 8. (B) Lastly, I have selected countries with good governance scores above 8. (C) Here are the graphs that are produced after the above filters.

<<A>>

[[File:Terence3.png|800px|center ]]<br>

<<B>>

[[File:Terence3a.png|800px|center ]]<br>

<<C>>

[[File:Terence3b.png|800px|center ]]<br>

From the graphs above, it can be seen that countries that have low scores for safe sanitation tend to also have low scores for clean air. It can also be seen that these countries tend to have a lower employment rate, although there are 4 countries that show a relatively high employment rate.

Graph B further reinforces the view that safe sanitation and clean air and closely related. In graph B, the countries highlighted are countries with high clean air ratings. These countries also have very high safe sanitation ratings. However, there does not seem to be any relationship between clean air and other variables shown in the graphs. For example, for the variables income distribution and good governance, the countries span almost the entire range of these 2 factors.

In Graph C, the countries shown are countries with high good governance scores. It could be interesting to note that these countries all have maximum scores for safe sanitation and clean air, implying that good governance could indeed affect the level of sanitation and air quality of a country. Also, these countries occupy the higher range of scores for gender equality.

== Clustering ==

Next, I have separated the data into different clusters. For the first form of clustering, I have made use of the Human Development Index created by the United Nations Development Program. The UNDP has already classified countries based on their HDI scores into 4 groups, namely ‘Very high human development’, ‘high human development’, ‘medium human development’, and ‘low human development’.

In order to make use of this information in my visualization, I have added a column to my dataset. This is how the modified dataset looks like:

[[File:Terence4.png|800px|center ]]<br>

Next, to further examine the clusters produced, I will be using Mondrian to come up with parallel box plots.

The purpose of using Mondrian is to make use of Mondrian’s feature where its graphs are all linked. This will make it easier to see which variables are the ones affecting a country’s classification in terms of level of human development.

Firstly, I will examine countries identified as ‘low human development’.

[[File:Terence5.png|800px|center ]]<br>

From the box plots seen here, it can be seen that countries that are classified as low human development are quite far from the average in terms of sanitation when we compare the box plot of the safe sanitation scores of all countries vs the box plot of safe sanitation scores among only countries with low human development. The interquartile range of the countries with low human development is completely below the interquartile range of all countries. The same can also be said about clean air.

All the other variables are also below average except for consumption. Low human development countries actually have higher consumption than the average.

[[File:Terence6.png|800px|center ]]<br>

Countries classified as medium human development fare better with safe sanitation. However, the range is large and there are countries that have very low scores as well as countries with very high scores. Similar to countries with low human development, these countries have lower than average scores for clean air. The ranges for consumption and employment scores span almost the entire possible range, just that the interquartile ranges for both are slightly higher than the average interquartile ranges.

[[File:Terence7.png|800px|center ]]<br>

There is a significant jump in safe sanitation scores among countries with high development. The range of scores is also much smaller and most of the scores are high. The same can be said for the scores for clean air.

[[File:Terence8.png|800px|center ]]<br>

For countries classified as very high human development, the safe sanitation and clean air scores are all clustered at the top, with only a few outliers with still relatively high scores. Gender equality and good governance scores are also higher than the average while consumption is much lower than the average.

**We will now examine the changes in the distributions from countries classified as low human development to countries classified as high development. The box plots will be placed side by side for easy appreciation of how the distributions change.**

[[File:Terence9.png|800px|center ]]<br>

**From this visualization, we can see that as human development increases, safe sanitation scores and gender equality scores increase. The increase in gender equality scores is not as drastic as that of safe sanitation.**

[[File:Terence10.png|800px|center ]]<br>

**Good governance scores generally increase whereas consumption scores generally decrease. However, what’s interesting in these visualizations is that there is a big leap between countries classified as high human development and countries classified as very high human development.**

[[File:Terence11.png|800px|center ]]<br>

**In this visualization, we see variables that do not change much across the different classification of countries. There is an increasing trend for income distribution, but the ranges are all very large for all kinds of countries. The values for employment across all the countries seem very similar.**

**== More in-depth analysis ==**

From all the analysis above, it can be seen that gender inequality is a strong determining factor to see if a country is highly developed or not. Let’s first view the gender equality scores across the countries in a treemap.

[[File:Terence12.png|800px|center ]]<br>

In this treemap, countries are grouped by their continents. The area represents the Sustainable Society Index score. More importantly, the colour represents the gender equality score. A darker shade signifies a lower score whereas a whiter shade signifies a higher score. By looking at the treemap, we can see that countries in Asia tend to have a darker shade, signifying low gender equality scores. On the other hand, in Europe, there are countries that are very white, such as the Scandinavian countries Finland, Norway and Iceland.

But how is the gender inequality score determined? Does a high gender inequality score really mean that there is high gender inequality in the country?

To investigate this question, I have used the other dataset that breaks down gender inequality into smaller variables. (Datatables B, Sheet Table 4).

[[File:Terence13.png|800px|center ]]<br>

For this visualization, I have highlighted only countries that are classified as very high human development and have high gender equality scores. From the visualization above, it can be seen that although gender equality scores are high, the percentage of female adults having at least secondary education still range from low to high values and at the same time, the % of seats in parliament held by women is on the low range. Female labour force participation rate is also lower than male labour force participation rate.

== Interactive Visualization ==

<https://public.tableausoftware.com/views/Assignment2_71/Dashboard1?:embed=y&:display_count=no>

In this interactive dashboard, users will be able to interact with 3 different visualizations. The first one includes bar charts of all countries and their scores. The second one is a parallel coordinate plot of countries against their different dimension scores (similar to above). The third one is another parallel coordinate plot but of broken down gender equality scores.

The first 2 graphs link to the same dataset and there is a single filter for both of them on the right of the first graph. Users will be able to highlight countries that fall under a specific classification. The third visualization has a similar filter but it is not linked to the other 2 visualizations as the data for the third visualization comes from a different dataset.