# MAIN GOAL:

Connection between economic health of population and the ‘happiness’ of the population.

**World Happiness**
https://www.kaggle.com/mathurinache/world-happiness-report?select=2020.csv
**Human Freedom Index**
https://www.kaggle.com/gsutters/the-human-freedom-index?select=hfi_cc_2019.csv
^2019 sheet has data 2011-2017??
**GNI per capita by Purchasing Power Parity**
https://data.worldbank.org/indicator/NY.GNP.PCAP.PP.KD

**Hf_score (0-10)**
**The Human Freedom Index (HFI) is the most comprehensive freedom index so far created for a globally meaningful set of countries. ... On a scale of 0 to 10, where 10 represents more freedom, the non-weighted average rating for 159 countries in 2014 was 6.93. https://www.cato.org/human-freedom-index-new**
**Hf_rank (1-159)**

**Happiness.Rank [NEED TO FIND SOURCE]**

**Happiness.Score (0-10)**
**Happiness score or subjective well-being (variable name ladder ): The survey measure of SWB is from the Dec 23, 2016 release of the Gallup World Poll (GWP), which covers the years from 2005 to 2016. Unless stated otherwise, it is the national average response to the question of life evaluations. The English wording of the question is “Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. https://s3.amazonaws.com/happiness-report/2017/StatisticalAppendixWHR2017.pdf**

**GNI_capita**
**https://datahelpdesk.worldbank.org/knowledgebase/articles/378832-what-is-the-world-bank-atlas-method**

**Useful Notebook for how to Analyse the Connection**
https://www.kaggle.com/rblcoder/mental-health-happiness-economics-human-freedom/data?

**Useful Notebook for how to Use World Happiness Data**
https://www.kaggle.com/charlievbc/human-freedom-index-interactive-visualization

**International Country Codes**
https://unstats.un.org/unsd/methodology/m49/ OR from WIIL dataset from Lab 3

https://www.kaggle.com/juanumusic/countries-iso-codes
^csv with codes

**NOTES FROM LECTURE:**

Need to write primary key for data in metadata i.e. what columns uniquely identify each rows value

Need to state the structure of the data e.g. dictionary of dictionary, is key the columns, what is the file key stored as, wide or long format?

# Summary of the work to do:

- Obtain a suitable data set
    - Ensure that you have data which has good quality and is clean from serious errors
    - Get rid of country rows without happiness index from the human freedom index
    - Append Human Freedom Index 2017 with Happiness Index 2017
    - Misspelling Countries/Country Codes
    - Remove no values, negative numbers
    - Filter out only 2017 data Human Freedom Index
- Produce a few summaries (aggregates) of some attributes
    - Ranking countries in terms of happiness & economics
        - Middle Country, Mean Values, Highest Ranking, Lowest Ranking (Could also involve sub-indicators of happiness & economics)

Write a report and submit it

The report should have a three-section structure that corresponds to the marking scheme:
1.a section that describes the data source(s), the format/contents of the data, the rights associated with the data, and some comment on any strengths or limitations of the dataset;
2.a section that describes the initial transformation and cleaning that you did (include here the parts of Python code that you used, or a description that is detailed enough to be followed);

this section should end with a brief explanation of where (in the submitted material) is found the metadata for the resulting dataset(s) which you will use in the rest of your project;
3.a section that describes and explains some simple analysis that you have done (again, show the code and also the output of the analysis).

25th September - Clean up the Data

If n/a's >30% or some other threshold, remove column (grok module 6 stuff)
Should we cut rows when there is an n/a or just the value

2 October - Aggregates & Report
9 October
16 October - Finish this
Sunday, 25th October 2020

# Report Draft

## Instructions

1. a section that describes the data source(s), the format/contents of the data, the rights associated with the data, and some comment on any strengths or limitations of the dataset;
2. a section that describes the initial transformation and cleaning that you did (include here the parts of Python code that you used, or a description that is detailed enough to be followed); this section should end with a brief explanation of where (in the submitted material) is found the metadata for the resulting dataset(s) which you will use in the rest of your project;
3. a section that describes and explains some simple analysis that you have done (again, show the code and also the output of the analysis).

# Project Stage 1

SIDS: 500445930, 500486852

## Data Sources

### World Happiness

Source & Metadata:

- Data collected from: https://www.kaggle.com/mathurinache/world-happiness-report?select=2020.csv
- Source: https://worldhappiness.report/ed/2017/
- Metadata: https://s3.amazonaws.com/happiness-report/2017/StatisticalAppendixWHR2017.pdf

This dataset originating from the World Happiness Report contains life-evaluation indicators by country which reflect how happy citizens perceive themselves to be (Helliwell, J., Layard, R., & Sachs, J. 2017). The data acquired from Kaggle was in csv format and was 155 by 12. It contained 11 indicator columns with 155 unique country entries. The data is licensed by CC0 1.0 i.e. Public Domain Dedication. The data can be downloaded here, https://www.kaggle.com/mathurinache/world-happiness-report?select=2020.csv, the complete written report with metadata can be found here, https://worldhappiness.report/ed/2017/.

In assessing the happiness levels of each country, they took the national averages of one question in the World Gallup Poll (GWP), which has been polling individuals across the world. The English translation can be found below:

“Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?” (Sourced from Helliwell, J., Layard, R., & Sachs, J. (2017). World Happiness Report 2017, New York: Sustainable Development Solutions Network.)

Around 1000 people are surveyed from each country each year, but this value slightly fluctuates, as shown below:

Table 1: Number of ladder (WP16) observations for WP5-years - Part 1

| Country/territory (wp5 ID) | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| United States (1) | | 1001 | 1225 | 1004 | 1003 | 1005 | 1008 | 2094 | 1005 | 2048 | 1019 | 1032 |
| Egypt (2) | 999 | | 1024 | 1105 | 2112 | 2053 | 5296 | 4186 | 1149 | 1000 | 1000 | 1000 |
| Morocco (3) | | | | | | 1006 | 1001 | 3000 | 1007 | | 2050 | 1008 |
| Lebanon (4) | 996 | 1000 | | 1000 | 2010 | 2027 | 2007 | 2013 | 1000 | 1000 | 1000 | 1000 |
| Saudi Arabia (5) | 1004 | | 1006 | 1150 | 2052 | 2038 | 2022 | 1077 | 2036 | 2035 | 1012 | 1000 |
| Jordan (6) | 1000 | | 1016 | 1007 | 2016 | 2000 | 2000 | 2000 | 1000 | 1000 | 1000 | 1000 |
| Syria (7) | | | | 1209 | 2100 | 2035 | 2041 | 2043 | 1022 | | 1002 | |
| Turkey (8) | 995 | | 1001 | 1004 | 999 | 1000 | 1001 | 2000 | 1000 | 2003 | 1002 | 1001 |

The following is a brief summary of what other key variables in this dataset mean and their source:

| Column Name | Economy..GDP.per.Capita. | Health..Life.Expectancy. | Freedom | Generosity |
|---|---|---|---|---|
| Values Type | Numeric | Numeric | Numeric | Numeric |
| Format | float | float | float | float |
| Missing Value | 0 | 0 | 0 | 0 |
| Meaning | GDP per Capita covers the Gross Domestic Product across countries, converted to international dollars (US Dollar) using Purchasing Power Parity rates. It is sourced from the August 10, 2016 release of the World Development Indicators (WDI) | Healthy Life Expectancy (HLE) is the average national lifespan of individuals, and was sourced from the World Health Organization (WHO), the World Development Indicators (WDI), and statistics published in journal articles. | Freedom to make life choices is the national averages of this question in the GWP: “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?” | Generosity is the national averages of this question in the GWP: “Have you donated money to a charity in the past month?” |

The documentation for the data is comprehensive; in particular it explains why its data collection method of life evaluation is a better reflection of happiness because it circumvents temporal, day-to-day fluctuations in mood. The accompanying report clearly outlines the way that data

within each column e.g. “Corruption, Trust and Good Governance” was sourced (either from the Gallup Poll or World Bank data), and justifies its methodology to procuring their values.

One limitation is that the number of people surveyed from each country varies both between countries and by year. This could mean that the degree to which the sample accurately represents the overall population varies. Additionally, while attempts have been made to avoid selection bias, due to the nature of polls, there may be an underlying bias on who agrees to complete the poll, such as those who do not work late nights, or are at home most of the day. While most of the data was procured through the Gallup Poll, as some data was produced from World Bank data, that may be subjected to governmental bias and how individual national governments choose to represent their own country.

Another limitation is that the data source uses the full names of countries as opposed to standardised ISO 3166-1 alpha-3 codes. This may cause problems when merging as the way a country’s name is expressed may differ across datasets. At first glance, the dataset also had a number of anomalous zero values in certain cells which suggested missing data. This will require cleaning to create a fairer spread across the countries and whether this involves removing those countries from the data or creating a model to determine those values will be decided after evaluating the strength of the dataset after removing them.

Helliwell, J., Layard, R., & Sachs, J. (2017). *World Happiness Report 2017*. New York: Sustainable Development Solutions Network.

## Human Freedom Index

Source & Metadata:
HFI Report:
https://www.cato.org/human-freedom-inedex-new
Data Source:
https://www.kaggle.com/gsutters/the-human-freedom-index?select=hfi_cc_2019.csv

This data source was originated from the Human Freedom Index Report and aims to capture a broad but rational accurate picture of overall freedom in the world (Ian Vásquez and Tanja Porčnik, The Human Freedom Index 2018: A Global Measurement of Personal, Civil, and Economic Freedom. Washington: Cato Institute, Fraser Institute, and the Friedrich Naumann Foundation for Freedom, 2018). More specifically the index seeks to provide an estimate to which the negative rights of individuals are respected in their country. With negative rights being defined as freedom from interference, namely from the government, and provided these individual freedoms don't impose on other people's freedom to do the same. It works to capture the freedom enjoyed by people in their respective countries in broad areas such as civil liberties, freedom of speech, religion, association, and assembly.

The data was collected by using only third party data to avoid bias from the authors, and by using 79 data streams from a multitude of sources. This was to ensure that the index has a broader perspective and is not a single view. The dataset actively avoided subjectivity by fulfilling the following criteria taken directly from the 2017 Index which included: ”the data comes from credible external sources and, for the sake of objectivity, are not generated by us; the index is transparent on methodology and sources; and the report covers as large a number of countries over as long a time period as is possible given the data available.” The data seeks to examine 12 different areas that involve both economic and personal freedom to provide the body of the index. These are: rule of law; security and safety; freedom of movement; freedom of religion; association, assembly, and civil society; expression and information; identity and relationships; size of government; legal system and property rights; sound money; freedom to trade; and regulation.
The following is a table with key columns from the data and a description:

| Column Name | pf_movement_domestic | pf_expression_cable | pf_ss_homicide | pf_religion_restirctions |
|---|---|---|---|---|
| Values Type | Numeric | Numeric | Numeric | Numeric |
| Format | float | float | float | float |
| Missing Value | 0 | 0 | 0 | 0 |
| Meaning | Freedom to move domestically means the ability that individuals have to travel within their own country | Access to cable and satellite means the ability of individuals to access these technologies | A value for the amount of homicides that occur in each respective country | A gauge on the religious restrictions imposed on individuals or groups in regards to all different religions |

The dataset we are using includes data from 2011 to 2017 and has 120 columns by 1621 rows. The first four columns deal with the year, country, iso code and region whilst the rest all contain their respective freedoms that make up the overall freedom index for each country, such as homicide, fatalities, criminality etc. The dataset was sourced  from Kaggle and is licensed under Open Data Commons Open Database License which means it is free to use.

As data that describes global human freedom, the human freedom index appears to be the most comprehensive - taking averages across a very broad range of different categories as documented in p.g. 17 and p.g. 19 of the HFI report linked above. It contains a far larger amount of data and different categories than other freedom indices such as Freedom in the World, Worldwide Press freedom index etc, because it includes personal, civil and economic freedom and is created by multiple countries (USA, Canada and Germany) rather than one sole country.

Some limitations with the data include some countries having no values for certain columns which can affect their HFI. Furthermore the data from each country could be skewed as governments may be corrupt in their reporting of certain values for political reasons. However there is a large amount of data from 162 countries which provides a relatively comprehensive amount of information that can be used to give a general idea of freedom in the world. Moreover, the HFI report did mention on p.g. 19 that indicators quantifying drug and alcohol prohibition were not included because insufficient reliable data was available.

Vásquez, I., & Porc̆nik, T. (2019). *The Human Freedom Index: A Global Measurement of Personal, Civil, and Economic Freedom*. Fraser Institute.

## GNI per capita by Purchasing Power Parity

Source & Metadata:
Data Source:
https://data.worldbank.org/indicator/NY.GNP.PCAP.PP.KD
Source Metadata:
https://datahelpdesk.worldbank.org/knowledgebase/articles/906531-methodologies#:~:text=National%20accounts%20and%20balance%20of,those%20reported%20in%20official%20sources.

This dataset covers the Gross National Income (GNI) per capita in 2017 and originates from current reports gathered by the Bank's country management units and official sources. It has been converted to international dollars using Purchasing Power Parity (PPP) rates, which compares the world's income against the US dollar. GNI works to calculate the total income earned by a country's people and business, and includes both investment income and money received from abroad.

"GNI_capita" is a measure of a country's income calculated using the World Bank Atlas Method. Low-income countries had a GNI per capita of $1,025, middle income countries sat between $1,026 and $4,035, upper middle income countries between $4,036 and $12,475 and high income countries had a GNI per capita of $12,476 or more.
https://blogs.worldbank.org/opendata/new-country-classifications-2016

Sourced from (International Comparison Program, World Bank | World Development Indicators database, World Bank | Eurostat-OECD PPP Programme), the .csv file was procured from Kaggle and contained 264 regional entries and 64 columns denoting countries and years. The

data is licenced by CC-BY 4.0, which allows users to copy, modify and distribute data in any format for any purpose, including commercial use. The data is sourced by the World bank and is drawn from the statistical systems of the countries that are a part of the world bank. The data can be downloaded here: https://data.worldbank.org/indicator/NY.GNP.PCAP.PP.KD

Some of the column variables are described below:

| Column Name | Country Name | Country Code | Indicator Name | 2017 |
|---|---|---|---|---|
| Values Type | Alphabetic | Alphabetic | Alphabetic and Special characters | Numeric |
| Format | string | string | string | float |
| Missing Value | n/a | n/a | n/a | Empty cell |
| Meaning | Standard Name of Country in English | Standardised ISO 3166-1 alpha-3 codes of each country and region. | Description of GNI per Capita value shown for each year, all cell values are “GNI per capita, PPP (current international $)” | The GNI per Capita for each country and region for the year of 2017. There are additional columns for all years between 1960 and 2019 (inclusive). |

Some of the main strengths of this dataset is their use of standardised ISO 3166-1 alpha-3 codes, which allows for direct comparisons to be made with data that also employ these. The metadata and documentation for this information is readily available in .csv format, which assists with understanding the values more. Additionally, as the years are divided up by columns, specific data for a year can be easily indexed. Furthermore, in comparison to GDP per capita, GNI per capita is a better indication of the economic health of the country’s own residents and businesses. Furthermore, the purchasing power parity calculation allows a fair comparison of purchasing capacity between different countries.
https://www.thebalance.com/gross-national-income-4020738
https://www.investopedia.com/updates/purchasing-power-parity-ppp/#:~:text=Purchasing%20power%20parity%20(PPP)%20is,standards%20of%20living%20between%20countries.

On the other hand, there are some missing values, which could act as confounding factors to our findings depending on which regions are primarily affected by these. Thus, when analysing the GNI and producing aggregate values of regions, it is important to be aware of what regions

are over and underrepresented, particularly post-data cleaning. As these data values are procured from respective countries' governments by the World Bank, these values are strongly subjected to governmental bias, and are susceptible to being altered or misrepresented as these values can be representative of a country's economic power.

## Utility Dataset: ISO 3166-2 Country Codes

Data Source:
https://www.kaggle.com/juanumusic/countries-iso-codes/metadata
Metadata:
https://www.iso.org/iso-3166-country-codes.html

This ISO Codes dataset contains a universal international standard code system for identifying unique countries and originated from the International Organisation for Standardisation. The dataset is under CC0 1.0 Universal (CC0 1.0) Public Domain Dedication which allows us to freely copy and modify without permission.

| Column Name | English short name lower case | Alpha-2 code | Alpha-3 code | Numeric code | ISO 3166-2 |
|---|---|---|---|---|---|
| Values Type | Alphabetical | Alphabetical | Alphabetical | Numeric | Alpha-numeric |
| Format | String | String | String | Integer | String |
| Missing Value | NA | NA | NA | NA | NA |
| Meaning | Common name of country | 2 Letter Identification Code for Country | 3 Letter Identification Code for Country | 2 or 3 Digit Identification for Country | ISO 3166-2: followed by Alpha-2 Code |

The dataset contains 5 columns with 246 entries. This dataset is reliable because it originates from a source that is not motivated by an alternative economic motive apart from global standardisation. It has no missing values which means that all officially recognised countries are listed in this dataset. This dataset is very relevant to our project because it allows us to merge all our other dataset together.

## Summary of our Cleaning

Initially we had to merge our three data sources of human freedom index, world happiness and GNI per capita. As we were only seeking to compare the data for the year of 2017, we began by extracting only 2017 data from the human freedom index as this contained data up to 2019. Which was then followed by the same process for GNI per capita.

```
#Extract only 2017 Data for Freedom Index
index_not_2017 = freedom_index[freedom_index["year"] != 2017].index
freedom_index.drop(index_not_2017, inplace=True)

#Extract only 2017 Data from GNI/capita PPP
gni_capita_2017 = gni_capita.iloc[:,[0,1,61]]
print(gni_capita_2017)
```

Now that the data had been extracted and established we sought to merge the three sources under one file.  Due to each data set having slight differences in the name of their “country” column, e.g “countries” or “country” we uniformly named them “countries” and the ISO code column as “ISO codes.”

However, the same countries often had different names between different datasets. Therefore, we tried to combine based on standard ISO codes. However, whilst both GNI per capita and the Human freedom index contained ISO codes, the world happiness report did not. Therefore by importing a list of ISO codes we were able to create a new column in world happiness that contained the ISO codes matched based on country name. Note that in doing so, countries which did not match the official common name had their entire row removed as it was unrealistic to manually search and change country names.

```
#Extract only ISO code & Countries from iso_code
iso_code = iso_code.iloc[:,[0,2]]

#Change Headings of all datasets to Match Freedom Index Dataset
world_happiness = world_happiness.rename(columns={"Country":"countries"})
gni_capita_2017 = gni_capita_2017.rename(columns={"Country Code":"ISO_code",
                                                  "Country Name":"countries",
                                                  "2017":"GNI_capita"})
iso_code = iso_code.rename(columns={"English short name lower case":"countries",
                                    "Alpha-3 code":"ISO_code"})

#Add ISO Code to World Happiness (It doesn't originally have ISO code)
world_happiness = pd.merge(world_happiness, iso_code, on=["countries"])
```

We then deleted country columns from 2 of the 3 datasets that would cause repeated columns when merged. We changed all the empty values in the datasets from their default value to NaN so that we had a uniform standard in our final merged dataset. When we did this, we carefully ensured that the expected range of columns did not include the empty values that we changed. We then successfully merged all the data sets based on the ISO codes column into one file named merged_data. This step also implicitly removed any country's data that was not present in all three data sets.

```
#Drops duplicate country columns from Freedom Index, GNI/capita
del gni_capita_2017['countries']
del world_happiness['countries']

#Changes all empty values in all datasets to NaN (i.e. empty)
gni_capita_2017 = gni_capita_2017.replace('', np.nan)
world_happiness = world_happiness.replace(0, np.nan)
freedom_index = freedom_index.replace('-', np.nan)

#Merges data based on matching ISO Codes to avoid errors in different spelling of countries
#Implicitly removes country data that is not present in all three sources
merged_data = pd.merge(freedom_index, world_happiness, on=["ISO_code"])
merged_data = pd.merge(merged_data, gni_capita_2017, on=["ISO_code"])
print(merged_data)
```

The next step involved cleaning the merged dataset. Firstly to select the data that was relevant to our analysis we selected columns containing: "year", "ISO_code", "countries", "region", "hf_score", "hf_rank", "Happiness.Rank", "Happiness.Score", "GNI_capita ". "Year", "ISO code", countries and regions all serve to pinpoint where the country is in the world geographically and for the year we are collecting the data for. "Hf_score" and "Happiness.Score" are all values from 0-10. "hf_rank" and "Happiness.rank" order the countries from the highest to lowest in each respective score. Whilst GNI_capita" is a measure of a country's income calculated using the World Bank Atlas Method.

```
#Selects columns that are relevant to our analysis
merged_data = merged_data[["year", "ISO_code", "countries", "region", "hf_score",
                           "hf_rank", "Happiness.Rank", "Happiness.Score", "GNI_capita"]]
```

For any row that had NaN i.e. missing value in any of the columns we decided to remove the entire entry. This is because the columns we have are crucial for what we want to compare and analyse. Whilst the likelihood of duplicate entries were low, we tested for this and removed any duplicate rows which would give us inaccurate summaries. There were 246 unique countries identified by the ISO code dataset and due to limited data available in each of the three remaining datasets the final number of unique entries in our merged data is 133.

```
#Remove row entries that have empty cells in any of the columns
merged_data.dropna(inplace=True)

#Remove duplicates
merged_data.drop_duplicates()
```

After this, we then proceeded to check the remaining data values. We first tested if all the values were within their expected range, and found all values to be appropriate. These ranges were acquired from each of the datasets' metadata. For example, all the columns for scores would require their values to be from 0 to 10, thus if any were not, the console would return false.

```
#Checks if all data values are within their expected ranges: returns True if data is within ranges
print((merged_data["hf_score"].astype(float) >= 0).all() and (merged_data["hf_score"].astype(float) <= 10).all())
#Between 0 and 10 inclusive
print((merged_data["Happiness.Score"].astype(float) >= 0).all() and (merged_data["Happiness.Score"].astype(float) <=
10).all()) #Between 0 and 10 inclusive
print((merged_data["hf_rank"].astype(float) >= 1).all() and (merged_data["hf_rank"].astype(float) <= 159).all())
#Between 1 and 159 inclusive
print((merged_data["Happiness.Rank"].astype(float) >= 1).all() and (merged_data["Happiness.Rank"].astype(float) <=
155).all()) #Between 1 and 155 inclusive
print((merged_data["GNI_capita"].astype(float) >= 0).all()) #Ensuring non-negative values
```

Finally, the data was scanned for any special characters, which ensured that all values were in the expected data type i.e. numeric/alphanumeric.alphabetical.

```
#This function check every item of a given column in a given dataframe for special characters that shouldn't be
present and returns a bool
def check_special(column, dataframe):
    special_char = True
    string_check= re.compile('[@_!#$%^&*()<>?/\|}{~:]')
    for item in dataframe[column]:
        if string_check.search(str(item)) == None:
            special_char = False
        else:
            special_char = True
    return special_char

#The following for loop checks all code to ensure the correct types of characters are present for each column i.e.
alphabetic, numeric
for column_name in merged_data:
    if check_special(column_name, merged_data) == True:
        print(column_name, ": some entries have special characters that are invalid" )
    else:
        if pd.to_numeric(merged_data[column_name], errors='coerce').notnull().all() == True:
            print(column_name, ": is entirely numeric")
        elif not merged_data[column_name].str.isnumeric().all() == True:
            print(column_name, " is entirely alphabetical letters or normal punctuation")
        else:
            print(column_name, ": column contains inappropriately mixed data")
```

# Simple Analysis

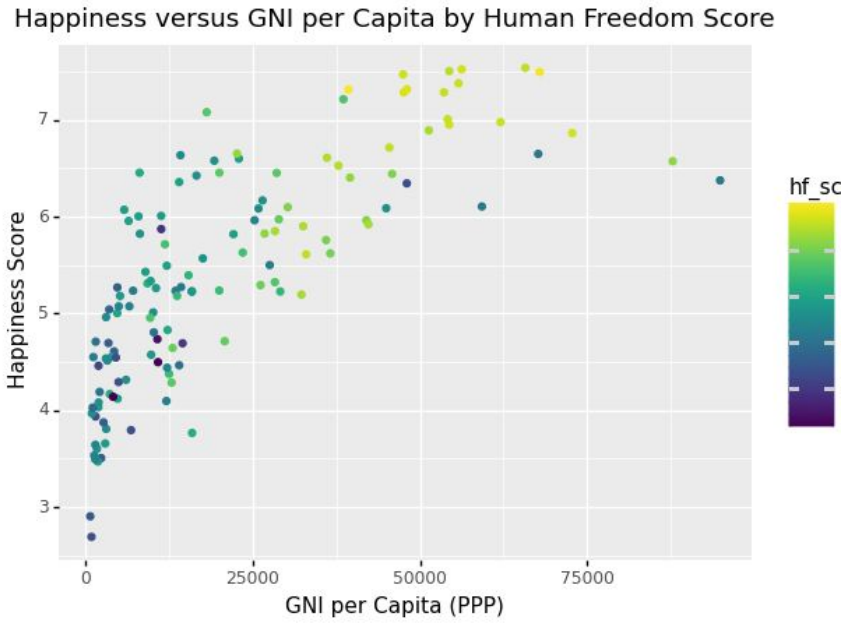
## Simple Aggregates

For each non-ranked variable (Happiness score, Human Freedom Score, as well as GNI per capita), we conducted a data summary to determine their count, mean, standard deviation, minimum, maximum and quartile values. From there, we printed the most extreme countries in each variable (e.g. the country with the largest and smallest GNI per capita).

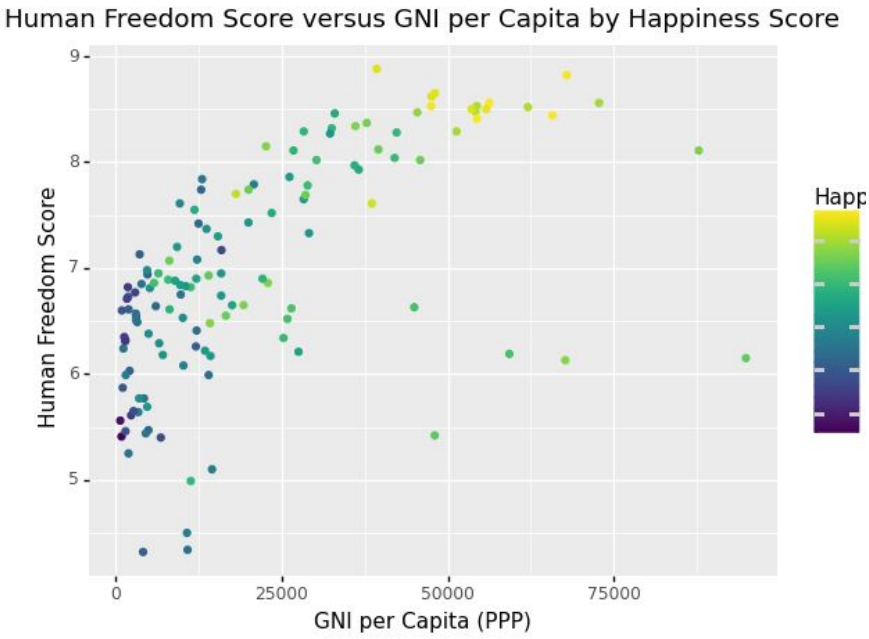
```
Happiness Data Summary
count    133.000000
mean       5.413624
std        1.138588
min        2.693000
25%        4.545000
50%        5.311000
75%        6.357000
max        7.537000
Name: Happiness.Score, dtype: float64
Maximum Happiness in: Norway
Minimum Happiness in: Central Afr. Rep.
Human Freedom Data Summary
count    133.000000
mean       6.969549
std        1.045091
min        4.320000
25%        6.290000
50%        6.850000
75%        7.840000
max        8.880000
Name: hf_score, dtype: float64
Maximum Human Freedom in: New Zealand
Minimum Human Freedom in: Sudan
GNI/capita Data Summary
count      133.000000
mean     21340.322618
std      20316.657271
min        772.401290
25%       5004.577847
50%      13738.524683
75%      32309.450253
max      94819.578441
Name: GNI_capita, dtype: float64
Maximum GNI/capita in: Qatar
Minimum GNI/capita in: Burundi
```

## Graphical Summaries

We also generated a scatterplot using ggplot2. The first graph represents the relationships between GNI per capita and human freedom, with the third variable, happiness, represented by the colour of the plots.




The second graph instead observes GNI per capita against happiness, with human freedom in colour.



## Grouped Aggregates

Additionally, we further analysed our data by creating grouped aggregates and determining data summaries for each one.

1) We first grouped our countries by quartiles of Happiness (1 being lowest quartile,  and then determined the average GNI for each one.

```
Happiness_Quartile
1                          4872.610913
2                         11144.093012
3                         24537.851999
4                         45305.756116
```

2) From our human freedom index dataset, countries were additionally described as a part of either ‘Caucasus and Central Asia’, ‘East Asia’, ‘Eastern Europe’, ‘Latin America and the Caribbean’, ‘Middle East and North Africa’, ‘North America’, ‘Oceania’, ‘South Asia’, ‘Sub-Saharan Africa’ and ‘Western Europe’. For our second aggregation, we used these additional parameters to group the countries and then determined the max happiness of each region.

```
Caucasus & Central Asia          5.819
East Asia                        5.920
Eastern Europe                   6.609
Latin America & the Caribbean    7.079
Middle East & North Africa       7.213
North America                    7.316
Oceania                          7.314
South Asia                       6.572
Sub-Saharan Africa               5.629
Western Europe                   7.537
```

3) As stated previously, low-income countries had a GNI per capita of $1,025, middle income countries sat between $1,026 and $4,035, upper middle income countries between $4,036 and $12,475 and high income countries had a GNI per capita of $12,476 or more. Our third aggregation and data summary involved determining the standard deviation of happiness in each of these defined income brackets.

```
GNI_capita_bins
(0.0, 1025.0]          0.684336
(1025.0, 4035.0]       0.489258
(4035.0, 12475.0]      0.661388
(12475.0, inf]         0.896195
```

4) Finally, we determined the median GNI for each quartile of the Freedom Index. We decided that reporting the median GNI would be a better reflection of each quartile as it would reduce the effect of any potential outliers.

```
Freedom_Quartile
1                      5810.544968
2                      6119.130532
3                     13738.524683
4                     45426.739014
```

## Source Code

```
#Basic Summary
print("Happiness Data Summary")
happiness_summary = merged_data['Happiness.Score'].describe()
print(happiness_summary)
print("Maximum Happiness in:", merged_data.loc[merged_data['Happiness.Score'].idxmax(),'countries'])
print("Minimum Happiness in:", merged_data.loc[merged_data['Happiness.Score'].idxmin(),'countries'])

print("Human Freedom Data Summary")
hf_summary = merged_data['hf_score'].describe()
print(hf_summary)
print("Maximum Human Freedom in:", merged_data.loc[merged_data['hf_score'].idxmax(),'countries'])
print("Minimum Human Freedom in:", merged_data.loc[merged_data['hf_score'].idxmin(),'countries'])

print("GNI/capita Data Summary")
GNI_summary = merged_data['GNI_capita'].describe()
print(GNI_summary)
print("Maximum GNI/capita in:", merged_data.loc[merged_data['GNI_capita'].idxmax(),'countries'])
print("Minimum GNI/capita in:", merged_data.loc[merged_data['GNI_capita'].idxmin(),'countries'])

#Basic Plots
p1 = (ggplot(merged_data, aes(x="GNI_capita", y="hf_score", color = 'Happiness.Score'))+
      geom_point() +
      xlab("GNI per Capita (PPP)") +
      ylab("Human Freedom Score") +
      labs(title = "Human Freedom Score versus GNI per Capita by Happiness Score"))
p2 = (ggplot(merged_data, aes(x="GNI_capita", y="Happiness.Score", color = 'hf_score'))+ \
      geom_point() +
      xlab("GNI per Capita (PPP)") +
      ylab("Happiness Score") +
      labs(title = "Happiness versus GNI per Capita by Human Freedom Score"))
print(p1,p2)

#Grouped Aggregates
#Average GNI for each quartile of Happiness
merged_data["Happiness_Quartile"] = pd.qcut(merged_data["Happiness.Score"], q=4, labels=[1,2,3,4])
print(merged_data.groupby("Happiness_Quartile").agg(
    average_GNI_capita = pd.NamedAgg(column = "GNI_capita", aggfunc=np.mean)
))
#Maximum Happiness grouped by Region
print(merged_data.groupby("region")["Happiness.Score"].max())
#SD of happiness in each Income Brackets
merged_data["GNI_capita_bins"] = pd.cut(merged_data["GNI_capita"], bins=[0,1025,4035,12475,float('inf')]) #These
values from World Bank, Max is Arbitrary
print(merged_data.groupby("GNI_capita_bins").agg(
    sd_happiness = pd.NamedAgg(column="Happiness.Score", aggfunc=np.std)
))
#Median GNI/capita for each quartile of Freedom Index
merged_data["Freedom_Quartile"] = pd.qcut(merged_data["hf_score"], q=4, labels=[1,2,3,4])
print(merged_data.groupby("Freedom_Quartile").agg(
    median_GNI_capita = pd.NamedAgg(column = "GNI_capita", aggfunc=np.median)
))
```