SC1015

MINI PROJECT

FOO JIN RUI
LOH RUI JIE JOEL
TAN YU

PROBLEM INTRODUCTION:
CAN WE PREDICT HAPPINESS?

Happiness is Fleeting

Pollution Levels

Climate change Is getting worse[1]

War & Political Instability

Civil unrest with Russia, Sub Saharan Africa[2], Myanmar

COVID-19 Pandemic

The global pandemic has Instated multitudinous lockdowns across the globe, and Inhibiting people from doing things and meeting people they love[3]

High Inflation

US Inflation Rate Hits New 40-Year High of 7.9%[4]

1. Mulvaney, K. (2022, February 28). Climate change already worse than expected, says new UN report. National Geographic. https://www.nationalgeographic.com/environment/article/climate-change-already-worse-than-expected-un-report
2. Africa will continue to suffer coups and civil wars in 2022. (2021, November 8). The Economist. https://www.economist.com/the-world-ahead/2021/11/08/africa-will-continue-to-suffer-coups-and-civil-wars-in-2022
3. How COVID-19 affects mental health. (2020, May 18). Singapore Institute for Clinical Sciences (SICS). https://www.a-star.edu.sg/sics/news-views/blog/blog/covid-19/angst-and-anger-why-does-the-covid-19-pandemic-make-us-so-upset
4. TRADING ECONOMICS. (2022, March 10). United States Inflation Rate - March 2022 Data - 1914–2021 Historical - April Forecast. https://tradingeconomics.com/united-states/inflation-cpi#:%7E:text=US%20Inflation%20Rate%20Hits%20New,of%201982%2C%20matching%20market%20expectations.

PROBLEM STATEMENT

HAPPINESS IS IN JEOPARDY

INTRODUCING:

CANTRIL LADDER

Overview of Cantril Ladder

Gallup World Poll

Questionnaire based on answers to the main life evaluation questions. It asks respondents to rank their current lives based on a scale of 0 to 10.

Compared to Dystopia

A dystopia is a hypothetical country which evaluation question values equal to the world's lowest national averages for each of the six factors.

6 Factors

01 **Levels of GDP**
GPD Per Person

02 **Life Expectancy**
Healthy Life Expectancy of a Country

03 **Generosity**

04 **Social Support**

05 **Freedom**
Perceived Freedom to Make Life Choices

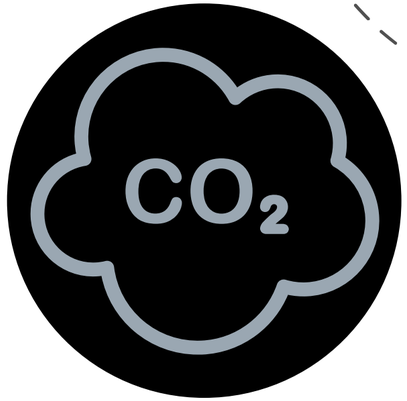
06 **Corruption**
Perception of Corruption

Are There Other Factors We Can Use to Predict the Cantril Ladder Score of a Country?

Governmental or economi

Two Other Factors

CO2 Emission Levels



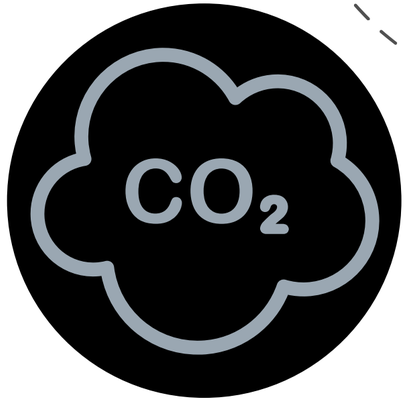
To track environmental Implication of a country

Number of Internet Users

To track social media penetration rates of a country

Hypothesis

CO2 Emission Levels



Lower levels will make people happier because the Climate Change is getting better!

Number of Internet Users

High internet usage will make people happier as they have more access to entertainment and freedom of speech.

DATA CLEANING
ISSUES

Each Year, the Cantril Ladder Only Adds Approximately 150 Countries into the List. And, There Are Only So Many Countries in the World.

In Other Words, Our Data Set Was Too Small!

So, What Can We Do?

We Decided to Compile the Cantril Ladder Scores over the Last 15 Years to Capture More Data Points

DATA CLEANING
PROCESS

Data Wrangling Process

01 Access Raw Data

02 Rename Columns

03 Drop Unnecessary Columns

04 Reshape the Data

05 Type Cast Objects

06 Merge and Drop NA

01 Access Raw Data

- Abstracted Raw Data from Multiple Sources like:
  - Cantril Ladder Score: https://ourworldindata.org/happiness-and-life-satisfaction
  - CO2 Emissions: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions
  - Percentage of Internet Users: https://data.worldbank.org/indicator/IT.NET.USER.ZS

Example: Cantril Ladder Score

```
df
```

	Entity	Code	Year	Ladder Score
0	Afghanistan	AFG	2008	3.724
1	Afghanistan	AFG	2009	4.402
2	Afghanistan	AFG	2010	4.758
3	Afghanistan	AFG	2011	3.832
4	Afghanistan	AFG	2012	3.783
...	...	...	...	...
1944	Zimbabwe	ZWE	2016	3.735
1945	Zimbabwe	ZWE	2017	3.638
1946	Zimbabwe	ZWE	2018	3.616
1947	Zimbabwe	ZWE	2019	2.694
1948	Zimbabwe	ZWE	2020	3.160

1949 rows × 4 columns

Example: CO2 Emissions

```
dfCO2
```

	Code	country	Year	co2	co2_per_capita	trade_co2	cement_co2	cement_co2_per_capita	coal_co2	coal_co2_per_capita	...
0	AFG	Afghanistan	1949	0.015	0.002	NaN	NaN	NaN	0.015	0.002	...
1	AFG	Afghanistan	1950	0.084	0.011	NaN	NaN	NaN	0.021	0.003	...
2	AFG	Afghanistan	1951	0.092	0.012	NaN	NaN	NaN	0.026	0.003	...
3	AFG	Afghanistan	1952	0.092	0.012	NaN	NaN	NaN	0.032	0.004	...
4	AFG	Afghanistan	1953	0.106	0.013	NaN	NaN	NaN	0.038	0.005	...
...	...	...	...	...	...	...	...	...	...	...	...
25186	ZWE	Zimbabwe	2016	10.738	0.765	1.415	0.639	0.046	6.959	0.496	...
25187	ZWE	Zimbabwe	2017	9.582	0.673	1.666	0.678	0.048	5.665	0.398	...
25188	ZWE	Zimbabwe	2018	11.854	0.821	1.308	0.697	0.048	7.101	0.492	...
25189	ZWE	Zimbabwe	2019	10.949	0.748	1.473	0.697	0.048	6.020	0.411	...
25190	ZWE	Zimbabwe	2020	10.531	0.709	NaN	0.697	0.047	6.257	0.421	...

25191 rows × 60 columns

02 Renaming Columns

- Using pandas rename() function
  - Example: For CO2 Emissions DataFrame, renamed:
    - "iso_code" to "Code" so It matches the column name In "Cantril Ladder" DataFrame
    - This makes merging the DataFrames easier
  - The same method is also used for Internet Users Dataset

Example: CO2 Emissions

```
dfCO2. rename(columns = {'co2_per_capita' : 'CO2 Per Capita'}, inplace = True)
```

iso_code	country
AFG	Afghanistan
AFG	Afghanistan

	Code	co
0	AFG	Afgha

03 Drop Unnecessary Columns

- Using pandas drop() function
  - Example: For Internet Users DataFrame, dropped unnecessary columns like "Indicator Name" and "Indicator Code:
  - This is to make the DataFrame easier to read and saves memory
  - The same method is also used for CO2 Emissions

Example: Internet Users

```
internet_df = internet_df.drop(["Indicator Name", "Indicator Code"], axis=1)
```

Country Name	Country Code	Indicator Name	Indicator Code	1960	:
Individuals using the					

Country Name	Country Code	1960
Aruba	ABW	NaN

04 Reshape the Data

- Using pandas melt() function
  - Example: For Internet Users DataFrame, the "Year" data are columns for each year. Instead, we want each row to map to each year with each country to make Exploratory Data Analysis easier.
  - This was only done for Internet Users DataFrame
  - Also renamed some columns to improve readability

Example: Internet Users

```
internet_df = internet_df.melt(["Country Name", "Code"])
internet_df.rename(columns = {'variable' : 'Year', 'value' : 'Internet Users Percentage'}, inplace = True)
```

Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964	1965	...	2011
Aruba	ABW	Individuals using the Internet (% of population)	IT.NET.USER.ZS	NaN	NaN	NaN	NaN	NaN	NaN	...	69.000000

	Country Name	Code	Year	Internet Users Percentage
10906	Aruba	ABW	2001	17.1
11172	Aruba	ABW	2002	18.8
11438	Aruba	ABW	2003	20.8
11704	Aruba	ABW	2004	23.0
11970	Aruba	ABW	2005	25.4

05 Type Cast Objects

- Using pandas astype() function
  - Example: For Internet Users DataFrame, the "Year" data are by default object types. Using the astype(int) function, we are able to type cast them successfully to the right format.
  - This is necessary to merge Internet Users DataFrame with the Cantril Ladder DataFrame.

Example: Internet Users

```
internet_df['Year']=internet_df['Year'].astype(int)
```

```
internet_df.dtypes
```

```
Country Name                    object
Code                            object
Year                            object
Internet Users Percentage      float64
dtype: object
```

```
internet_df['Year']=internet_df['Year'].astype(int)
internet_df.dtypes
```

```
Country Name                    object
Code                            object
Year                             int64
Internet Users Percentage      float64
dtype: object
```

06 Merge and Drop NA

- Using pandas merge() and dropna() function
  - Example: For Internet Users and Ladder Scores DataFrames, we first merged them with the merge() function.
  - Then, we removed rows which contains NA values with dropna() function.
  - We also dropped some more unnecessary columns so the final data set contains only "Internet Users Percentage" and "Ladder Score"
  - The same method is also used for CO2 Emissions Dataset

Example: Internet Users

```
concat_df = pd.merge(happiness_index_df, internet_df, on=["Year", 'Code'], how='left')
concat_df = concat_df.drop(["Entity", "Code", "Year", "Country Name"], axis=1)
concat_df.dropna(inplace= True)
```

	Ladder Score	Internet Users Percentage
0	3.724	1.840000
1	4.402	3.550000
2	4.758	4.000000
3	3.832	5.000000
4	3.783	5.454545
...	...	...

EXPLORATORY
DATA ANALYSIS

Removal of Anomalies

- Using sklearn.ensemble IsolationForest() to remove anomalous points in the multi variate setting.
    - CO2 Per Capita against Ladder Score
    - Internet Users Percentage against Ladder Score
- Using a contamination value of 0.01
- Rows with anomaly value as -1 are dropped

Example

index	Ladder Score	Internet Users Percentage	Anomalies_scores	Anomaly
200	4.031	36.74474744	0.06934851548473075	1
201	3.762	37.31205037	0.031044666890072237	1
202	3.499	39.36299738	0.005161830643152676	1
203	3.505	41.41379464	0.008894334829710804	1
204	3.461	58.0	0.0005652145331268565	1
205	3.471	61.0	-0.0009194564183481191	-1

Removal of Anomalies

CO2 Per Capita against Ladder Score

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1924 entries, 0 to 1923
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Ladder Score    1924 non-null   float64
 1   CO2 Per Capita  1924 non-null   float64
dtypes: float64(2)
memory usage: 30.2 KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1904 entries, 0 to 1903
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Ladder Score    1904 non-null   float64
 1   CO2 Per Capita  1904 non-null   float64
dtypes: float64(2)
memory usage: 29.9 KB
```

Removal of Anomalies

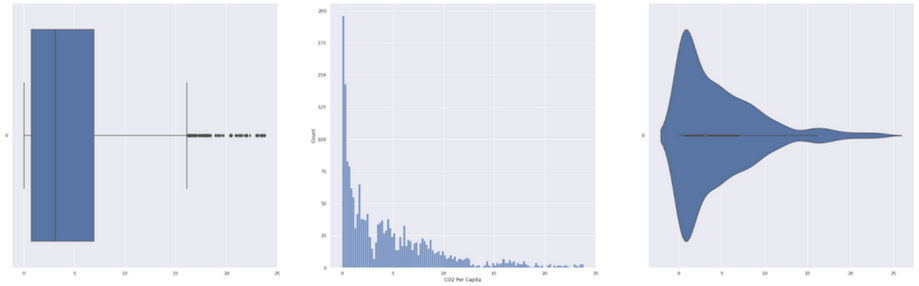
Internet Users Percentage against Ladder Score

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1768 entries, 0 to 1767
Data columns (total 2 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Ladder Score               1768 non-null   float64
 1   Internet Users Percentage  1768 non-null   float64
dtypes: float64(2)
memory usage: 27.8 KB
```

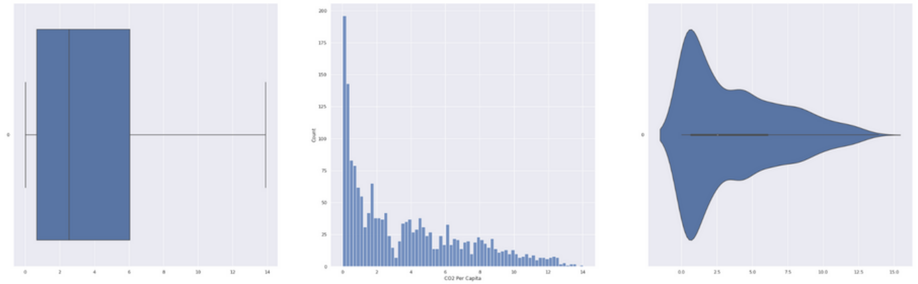
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1750 entries, 0 to 1749
Data columns (total 2 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Ladder Score               1750 non-null   float64
 1   Internet Users Percentage  1750 non-null   float64
dtypes: float64(2)
memory usage: 27.5 KB
```

Uni-Variate Data Visualisation

CO2 Per Capita

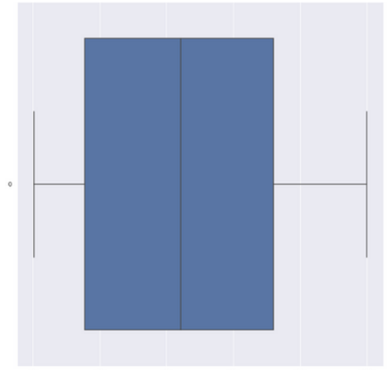


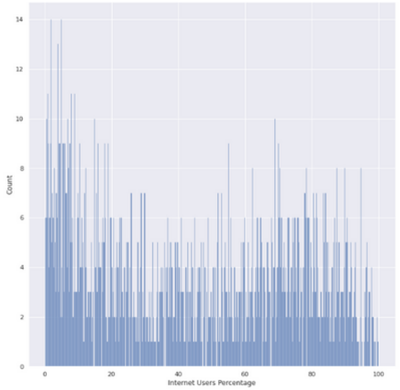
(outlier removal)

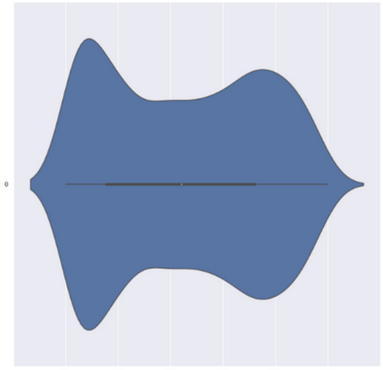


Uni-Variate Data Visualisation

Internet Users Percentage
(no outliers)

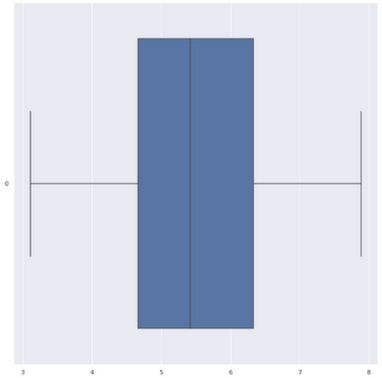


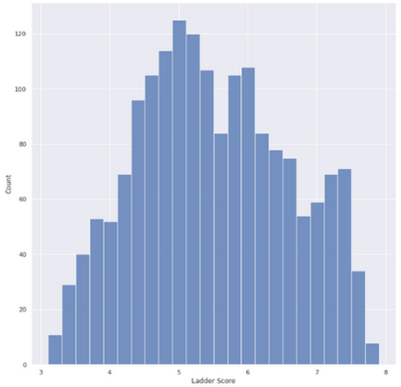


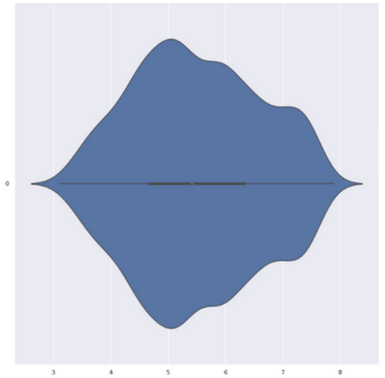


Uni-Variate Data Visualisation

Ladder Score
(no outliers)

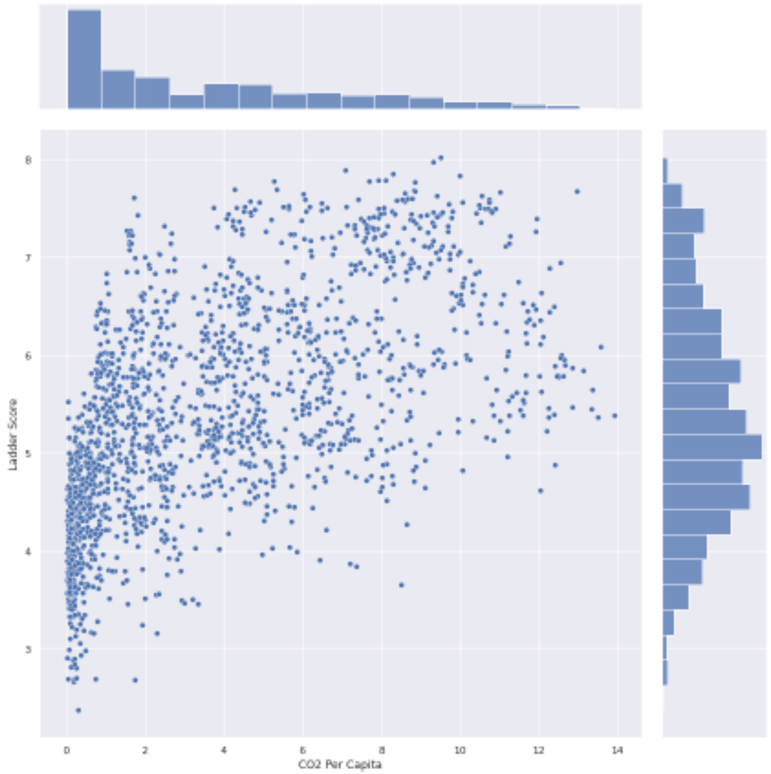


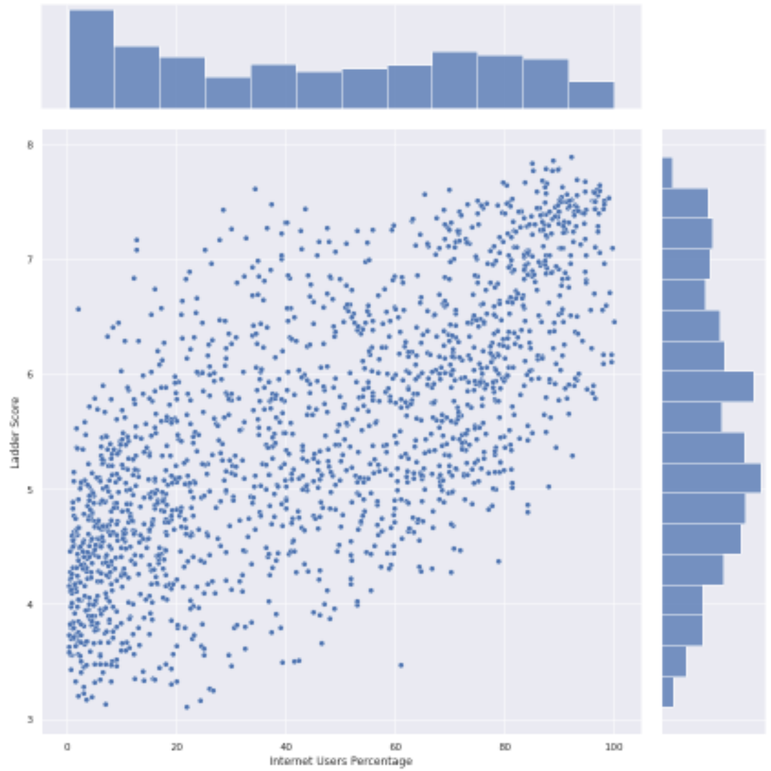




Bivariate Visualisation

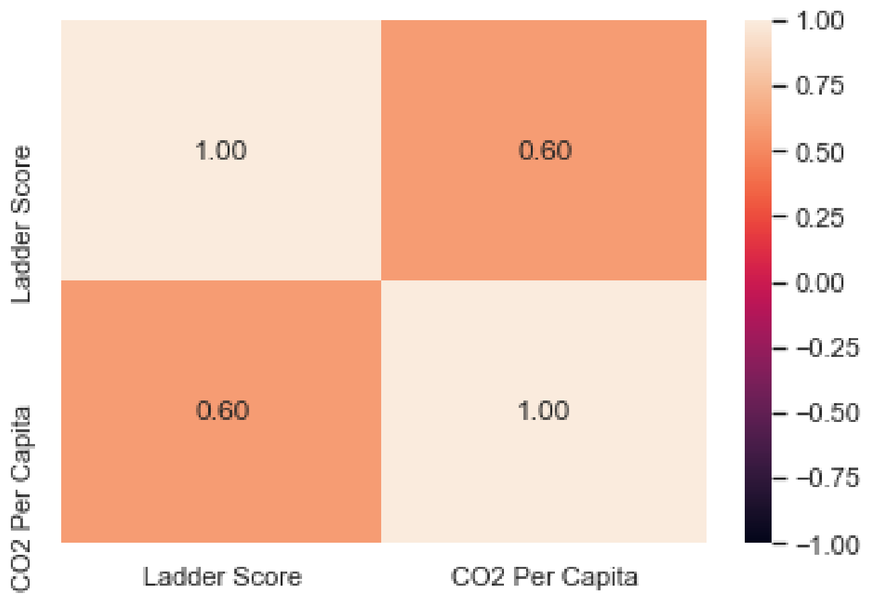
Join plots reveals positive correlation for (CO2 Per Capita against Ladder Score) and (Internet Users Percentage against Ladder Score).



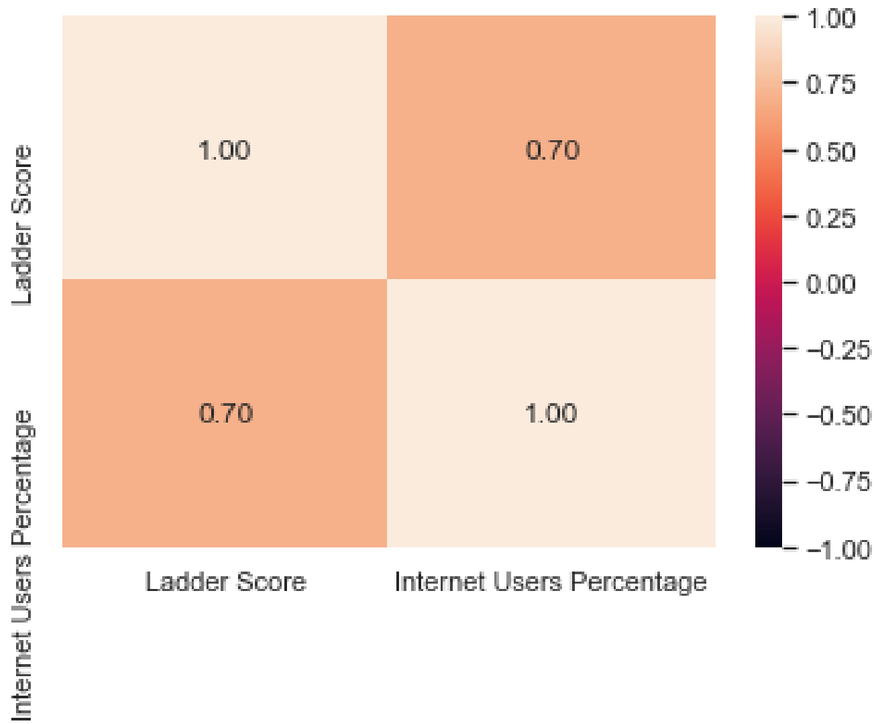


Correlation Matrix

Relatively high correlation coefficient of 0.6



HIgh correlation coefficient of 0.7

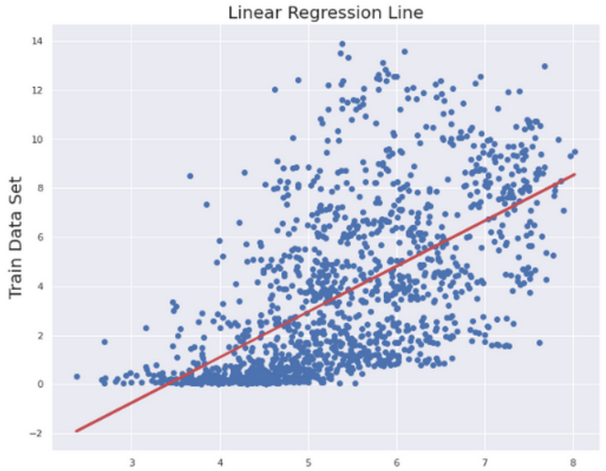


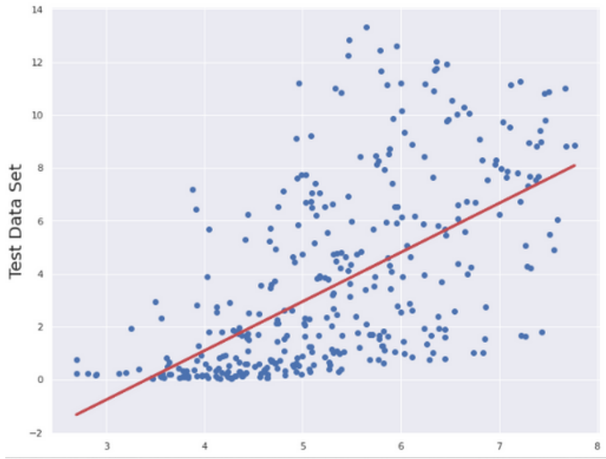
MACHINE LEARING
LINEAR REGRESSION

Linear Regression

- Using sklearn.linear_model Linear Regression().
- Linear Regression is performed for CO2 Per Capita against Ladder Score and for Internet Users Percentage against Ladder Score.
- Datasets are split into train and test randomly using train_test_split() from sklearn.model_selection.

Model Using CO2 Per Capita as Predictor

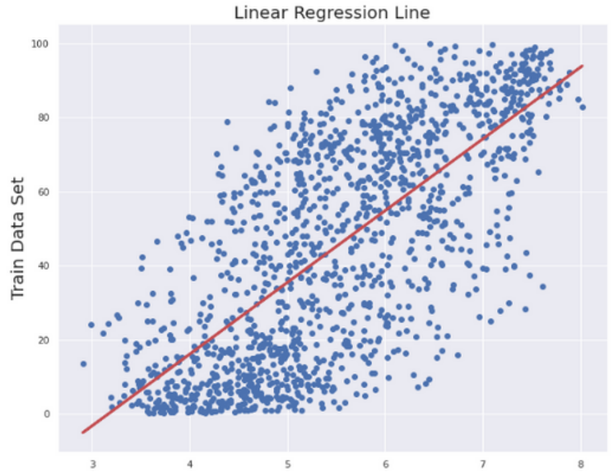


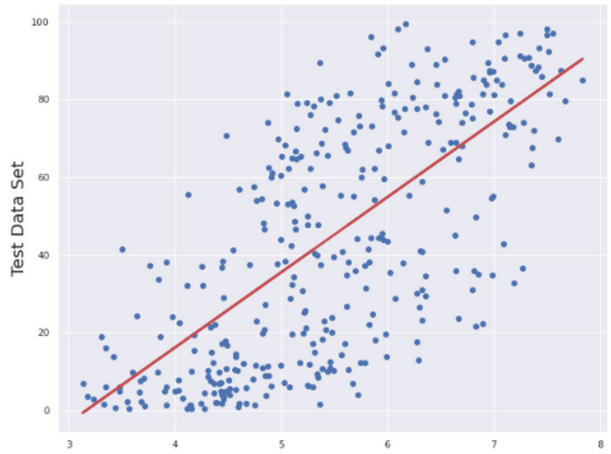



```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.3609584729246559
Mean Squared Error (MSE)          : 7.3573528795514385
Root Mean Squared Error (RMSE)    : 2.712444078603546

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.356253353450799
Mean Squared Error (MSE)          : 8.003690437204376
Root Mean Squared Error (RMSE)    : 2.8290794328198663
```

Model Using Internet Users Percentage as Predictor



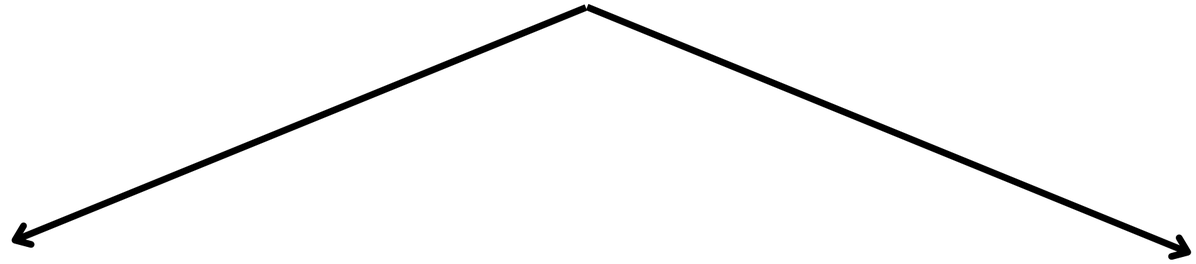



```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.5016302374262338
Mean Squared Error (MSE)          : 456.516289247778
Root Mean Squared Error (RMSE)    : 21.366241813846862

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.4855881537973298
Mean Squared Error (MSE)          : 483.25837481388373
Root Mean Squared Error (RMSE)    : 21.983138420477722
```

Statistical Insights

Account for the Correlations between the Ladder Score and each of the factors

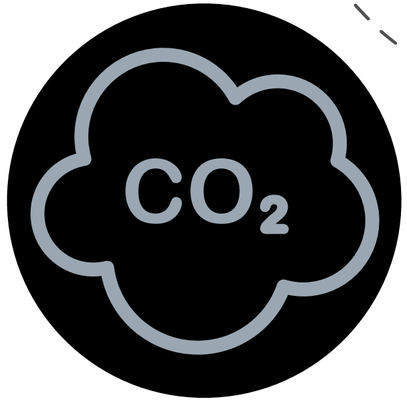


CO2 Emissions

Internet Usage

Back to our Hypothesis...

CO2 Emission Levels

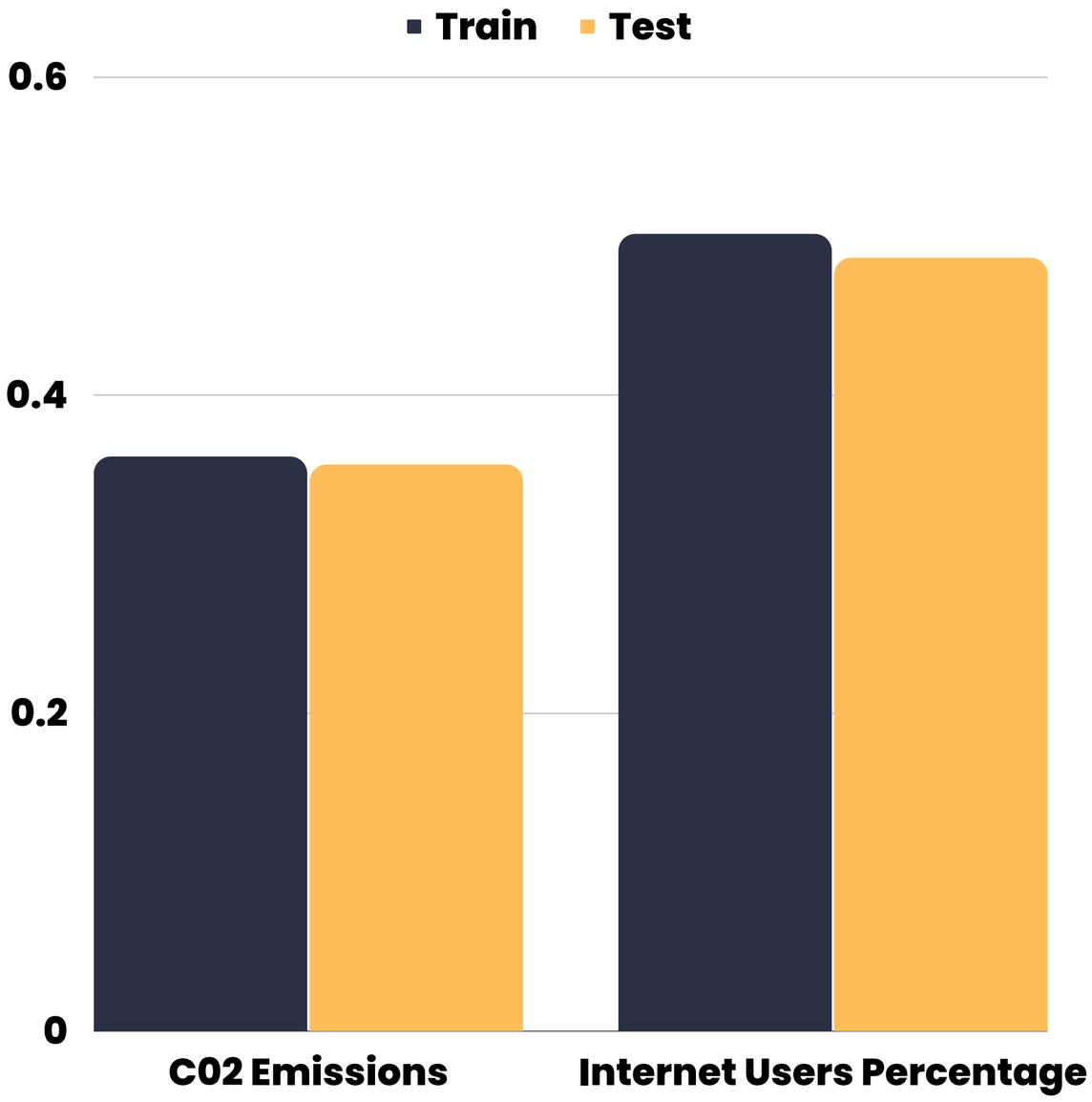


Lower levels will make people happier because the Climate Change is getting better!

Number of Internet Users

High internet usage will make people happier as they have more access to entertainment and freedom of speech.

Explained Variance (R^2)



The higher the R^2, the better the model?

- Both factors (CO2 Emissions and Internet Users Percentage) have **Explained Variance** values close to **0.4**
- Are both of these factors a good predictor of the Cantril Ladder Score (a.k.a Happiness Levels)?
- Statistical Intuitons from Correlation Matrix

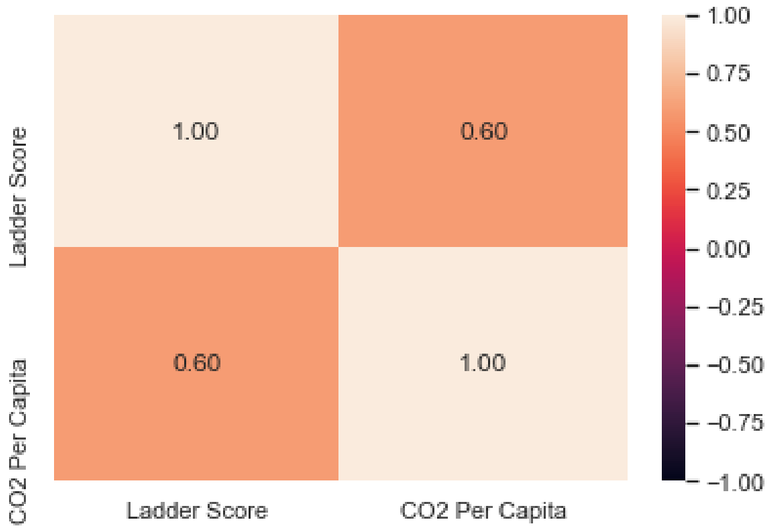
CO2 Emissions

Mean Squared Error

- MSE of 8.00
- However, MSE does not provide us any insight

Explained Variance

- R^2 value of of 0.356
- Does that mean that CO2 emissions per capita is not a good predictor of Cantril Ladder Score?



Correlation Matrix

- Relatively strong Correlation of 0.6
- The higher the CO2 Emissions, the higher the Cantril Ladder Score?
- Or is there more to this relationship than meets the eye?

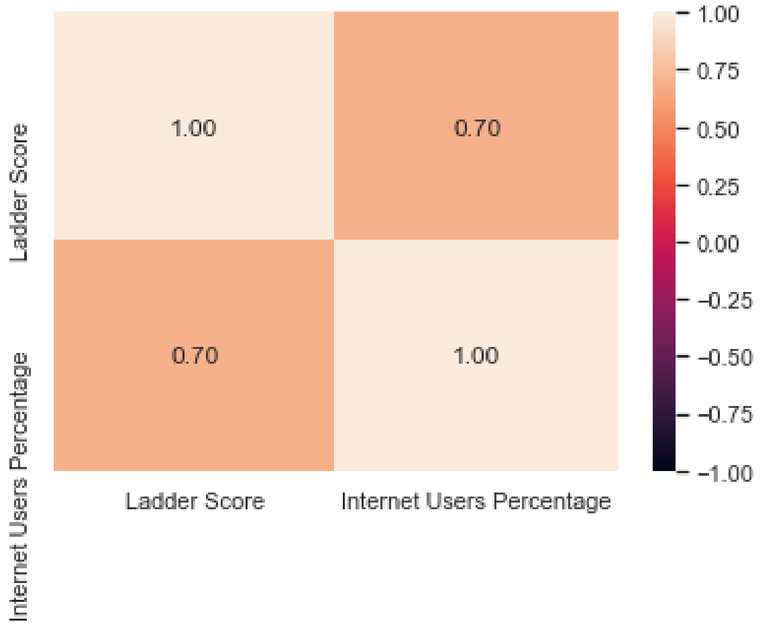
Internet Usage

Mean Squared Error

- MSE of 483
- However, MSE does not provide us any insight

Explained Variance

- R^2 value of of 0.486
- Would this be a better predictor than CO2 emissions per capita in predicting the Cantril Ladder Score?



Correlation Matrix

- Strong Correlation of 0.7
- The higher the Internet Usage per Capita, the higher the Cantril Ladder Score

Conclusions

Low CO2 emissions does not mean Happiness Levels are higher...

- Statistical results have proven otherwise.
- But why?

Internet Usage is a better predictor than CO2 emissions per capita.

- Explained Variance is greater than that of CO2 emissions.
- Stronger Correlation to the Cantril Ladder Score

But higher Internet Usage brought higher levels of Happiness!

- With the advent of cyberbullying, will we see a decrease in Happiness Levels in the future?

CO2 emissions may not necessarily be a bad predictor!

- Still has a relatively strong correlation to the Cantril Ladder Score.
- Low Explained Variance does not mean that it is a bad predictor.

Thank you for your time.

SC1015 Mini-Project

Foo Jin Rui

Tan Yu

Joel Loh