Data Visualization Project

Name: Swastik Shrey & Preetam Bolla
Section: S11
Roll no's: 211EE148 & 211EE132 respectively

***Problem Statement* — This project aims to create an infographic/visual graph to find a correlation between *a family's household income* and their *inclination towards switching to renewable sources of energy*.**

Python libraries used[1]:

1. ***NumPy* – mathematical operations**
2. ***Pandas* – extracting and organizing data**
3. ***Matplotlib* – plotting data**
4. ***Seaborn.* – building over matplotlib graphs**
5. ***Plotly* – for animated World heat map**

Design Plan of the Project

THE STEPS WE BROADLY FOLLOWED ARE:

1. Create a **hypothesis** and target a set of data that we'd need
2. **Find the data** & suitable visualizing plot
3. Conduct Exploratory Data Analysis **(EDA)**
4. Clean and **organize the data** to our requirements
5. Find **outliers** using a swarm plot
6. Creating **3D plots** to and compare the **values** of our 3-variable dataset
7. Creating **animated World heat map** plot to see the **change of these values** over 20 years
8. Finding **Spearman and Pearson** correlation between median GNI and Renewable energy usage **year wise.**
9. Finding correlations between total change in GNI and Renewable energy usage over 20 years **country wise**
10. Finding the same correlation for **rich & poor** countries separately to see if rich countries have a better correlation
11. Finally using all data analysis to find out the particular countries with **the highest contribution to the correlation** and analyzing them to find commonality and hence, the affecting factor.
12. **Plotting correlograms** to visualize all the correlations.

Terminologies used ahead to increase readability:

- GNI: *Gross National Income (per capita)*
- NRG: *Renewable energy usage by countries*
- SPC: *Spearman Correlation*
- PEC: *Pearson Correlation*
- EDA: *Exploratory Data Analysis*

Note: Every python function mentioned in the report or used in the project has been explained by comments. Kindly find the jupyter notebooks (.ipynb) on our GitHub repository.

GitHub Link: S-Shrey-09/DataVizProject: Python Data

Uniqueness of the Project

The project is completely unique and the code writing has been done completely by us. The code writing followed the chronology:
Decide what we want a piece of code to do – **searching the official documentation[1] to find a function that helps us execute** – adapt the function to our needs

I. Create a hypothesis and target a set of data that we'd need

The hypothesis was created with an aim to find a relation that can help maximize the usage of renewable sources of energy in the world. This project is its first step.

The data that we chose were:
1. **GNI** – Household income per capita. We didn't choose GDP because we wanted to see an individual household's inclination towards adoption of renewable energy rather than the whole country's
2. **Renewable energy usage per capita** – Per capita data again chosen to see an individual household's relation.

II. Find the Data & suitable visualizing plot

We found the country wise data on **World Databank's** website. It was raw data with lots of problems and null datapoints that we had to fix.

We **used DataToViz** website's help as well as matplotlib's documentations to choose the appropriate plots for data visualization.

1. Since we had 3 variables – Years, Countries and Values = f (Years, Countries) we chose **a 3D graph.**
2. 3D values were also represented on **a 2D world map by animating** it to show different values for different years.
3. Outliers can be found with multiple types of plots. We went **with swarm plots** for it's easier to understand visuals.
4. Finally, plotting **correlograms** for all correlations that we made throughout the project.

III. Conduct Exploratory data analysis (EDA)

EDA is done in order to identify patterns and analyse the data on surface level.
.describe( ) is a useful pandas function to find the minimum, maximum, mean etc of the used data.
.info( ) is a pandas function that helps find null data points which helped us choose countries which had complete data from 1999-2019.
We arranged the data in descending order for the years 1999, 2004, 2009, 2014 & 2019 to see which countries lead and which countries perform poor to get a basic idea of correlation before going ahead.

[1] All documentations and resources used have been linked to at the end.

IV. Clean and organize the data to our requirements

Cleaning and organizing data comprised of these steps:

1. Choosing **years 1999-2019**
2. Choosing **12 rich countries, 12 poor countries, and World average** with equal representation of continents as well.
3. **Removing outliers** identified by the swarm plot for the World map graph in order to keep the scale understandable.
4. **Removing null value** data points to avoid errors.
5. Creating a separate csv with the above changes using pandas.

V. Find outliers using a swarm plot

Outliers disturb the averages of the data as well as the scale of the graphs. Swarm plots for GNI and NRG data show :
GNI outliers: ***Luxembourg***
NRG outliers: ***Norway***

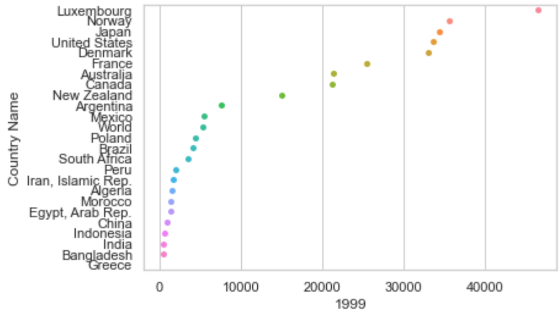


Figure 1: GNI Swarm plot – outlier: Luxembourg

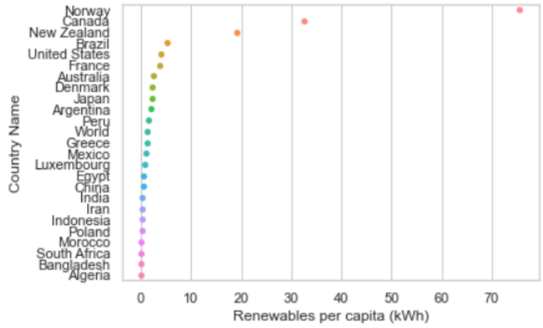


Figure 2: NRG Swarm plot - outlier: Norway

VI. Creating 3D plots[2] to and compare the values of our 3-variable dataset

3D plots are also a part of EDA to see basic correlations.

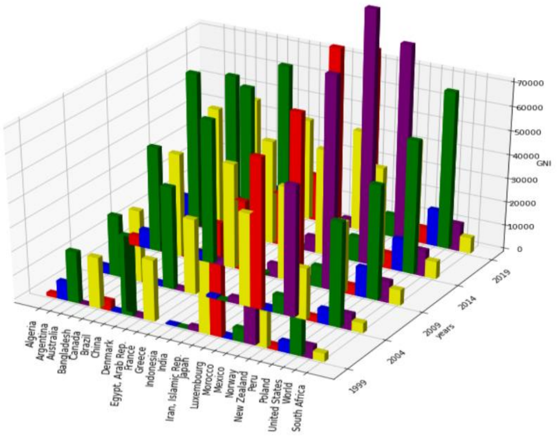


Figure 3: GNI 3D plot

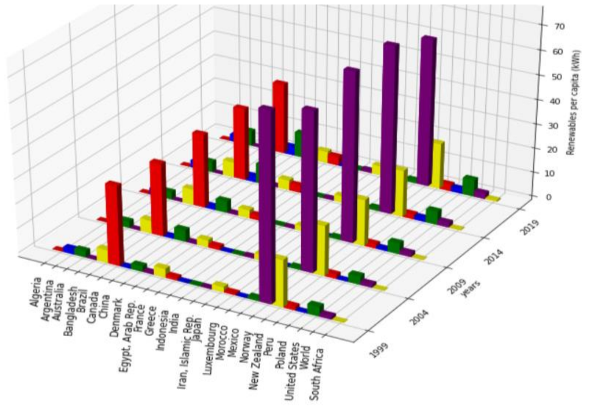


Figure 4: Renewable Energy usage 3D plot

From these 3D graphs we concluded that comparing countries on **the rate of change of values would better** as then they would be in a smaller range of values that can be visualized better.

VII. Creating animated World map plot to See the change of these values over 20 years

The animations across multiple years can't be displayed in the report hence here's a still image. It can be played on any python IDE please do try. This graph was made using Plotly library instead of Matplotlib

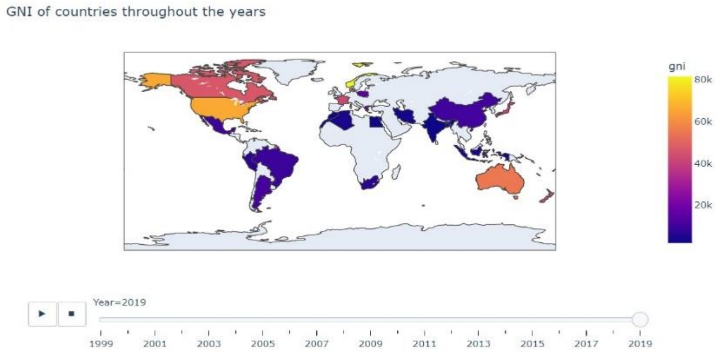


Figure 5: GNI World Heat Map of year 2019

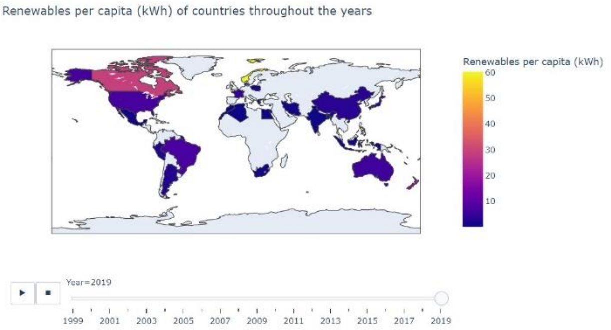

Renewables per capita (kWh) of countries throughout the years
Renewables per capita (kWh)
60
50
40
30
20
10
▶
■
Year=2019
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019

Figure 6: Renewable Energy usage World Heat map of 2019

VIII. Finding Spearman and Pearson correlation between median GNI and Renewable energy usage year wise

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}$$

Spearman correlation coefficient formula

[2] Please refer to out GitHub repo to see the graphs clearly in full size.

Spearman coefficient[3] is basically the covariance of rank variables of the datasets. It's a commonly used correlation coefficient in data analysis along with *Pearson's*, which is the covariance of standard deviation of values of data.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

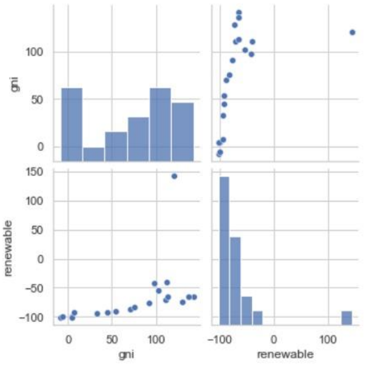
Pearson correlation coefficient formula

Both Spearman and Pearson coefficients can lie between -1 to 1**. A strong correlation is shown by coefficient 1 and -1 (-1 is inverse proportionality) and 0 depicts no correlation**.

We found the median values of GNI and NRG respectively for every year 1999-2019. Then finding a correlation between these values got us the values:

	GNI	NRG
GNI	1	0.876
NRG	0.876	1

Spearman correlation: Median of years



IX. Finding correlations between total change in GNI and Renewable energy usage over 20 years country wise

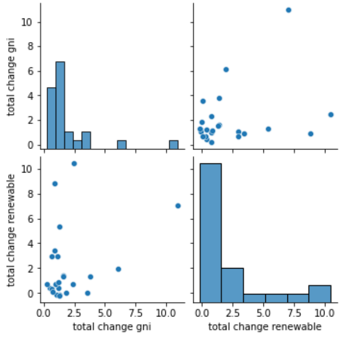
The median values of each year gave us a sense that the correlation might exist for the world collectively so we went ahead with finding correlation country wise. We created a new column using *Pandas* which followed the formula:

$$total.change = \frac{GNI['2019'] - GNI['1999']}{GNI['1999']}$$

And the same ***total_change*** column in NRG data as well whose correlation came out to be:

	GNI	NRG
GNI	1	0.316
NRG	0.316	1

Pearson correlation: Total change of countries



X. Finding the same correlation for rich & poor countries separately to see if rich countries have a better correlation

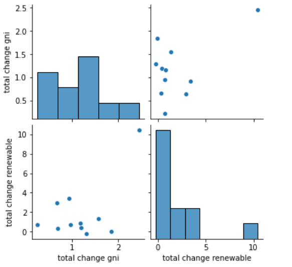
Repeated the exact process for rich and poor countries separately to see if income was the major deciding factor here.

	GNI	NRG
GNI	1	0.551
NRG	0.551	1

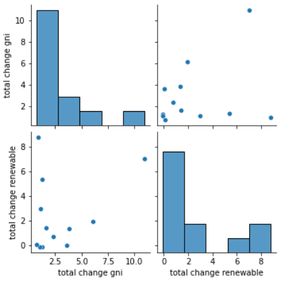
Pearson correlation: Total change of rich countries

	GNI	NRG
GNI	1	0.308
NRG	0.308	1

Pearson correlation: Total change of poor countries



Rich countries



Poor Countries

Rich countries do show a better correlation but it's not too significant.

XI. Finding out the particular countries with the highest contribution to the correlation.

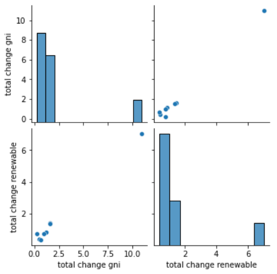
By studying the total increase on GNI and NRG we can choose the indexes of the countries which will give us a good correlation. We also found out the correlation values of the countries that had GNI and NRG changes far off.

	GNI	NRG
GNI	1	0.995
NRG	0.995	1

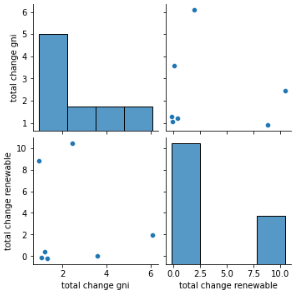
Pearson correlation: Total change of high contributors

	GNI	NRG
GNI	1	-0.091
NRG	-0.091	1

Pearson correlation: Total change of low contributors



Good correlation



Bad correlation

We can conclude that the following list of countries are the ones with best and worst correlations. The correlation wasn't geographical location based. It didn't totally depend on the GNI either and the countries might have a complex commonality that goes beyond the scope of this project but we'll make sure we work on it.

High correlation	Low correlation
Algeria	Bangladesh
Argentina	Canada
Australia	Egypt

[3] More about spearman and pearson correlation here: Spearman's rank correlation coefficient - Wikipedia

Pearson correlation coefficient - Wikipedia

Chine	Indonesia
France	Norway
Japan	Poland
United States	South Africa

List of countries with high and low correlation of NRG and GNI

XII. Plotting Correlograms to Visualize All the Correlations

Correlograms have been plotted and displayed with their respective correlations.

Conclusion

This project has further scope of doing a similar state wise data analysis. Further we can do the same with use of non-renewable energy usage vs GNI to see if the same trend follows. Our current project is just a stepping stone towards the conclusion of the hypothesis; however, we can say that the usage of renewable energy isn’t continent dependent and dependent on the GNI of a country to a limit. We also have a list of countries who showed progress and hence we’ll try to find commonalities in them and reassess our work.

Acknowledgement

We would like to thank Basavaraj sir for teaching us and giving us an amazing opportunity to present what we’ve learnt.

References

[1] Documentations and tutorials used to learn libraries:
a) NumPy - *NumPy user guide — NumPy v1.22 Manual*
Complete Python NumPy Tutorial - YouTube
b) Pandas - *User Guide — pandas 1.4.1 documentation (pydata.org)*
Complete Python Pandas Data Science Tutorial -YouTube
c) Matplotlib - *Users guide — Matplotlib 3.5.1 documentation*
Python Plotting Tutorial w/ Matplotlib & Pandas - YT
d) Seaborn - User guide and tutorial — seaborn 0.11.2 (pydata.org)
Seaborn Tutorial 2021 - YouTube
e) Plotly - Plotly Python Graphing Library

[2] Books used:
1. Python Data Science Handbook – Jake VanderPlas
2. Data visualization with Python – Mario Dobler and Tim Großmann

[3] For choosing data visualizers:
From data to Viz | Find the graphic you need (data-to-viz.com)
44 Types of Graphs & Charts [& How to Choose the Best One]

[4] Data collected from:
Per capita energy consumption from renewables, 2019
India | Data (worldbank.org)
public-apis/public-apis: A collective list of free APIs (github.com)
Wikipedia, the free encyclopedia

[5] Stack overflow for all minor doubts and debugging
Stack Overflow - Where Developers Learn, Share, & Build Careers

[6] 3D axes documentation:
mpl_toolkits.mplot3d.axes3d.Axes3D

[7] Animated World heat map tutorial: Plotly Animation Python | World Development Indicator Data Set - YouTube