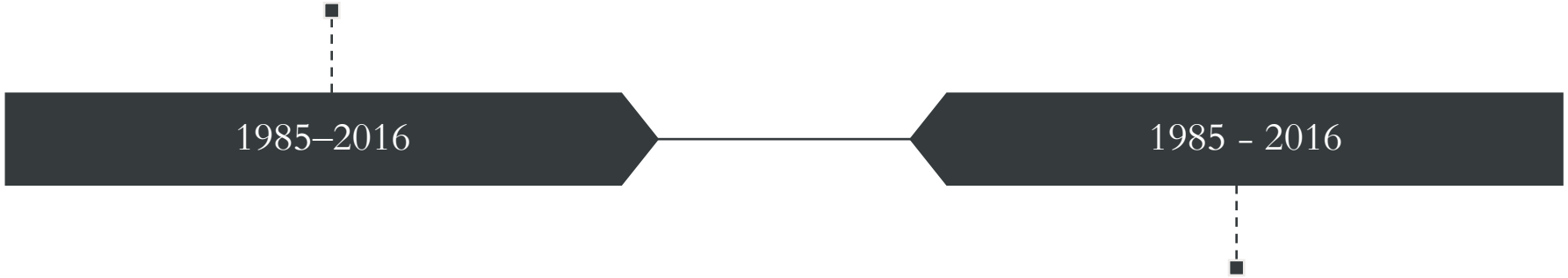
SILENT EPIDEMIC OF HUMAN MIND

- Analysis and presentation
- Joyita Roy

DATA SOURCES:

World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from: http://databank.worldbank.org/data/source/world-development-indicators



Kaggle: ttps://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

PATHWAY:

Asking questions

Data wrangling

EDA

Draw conclusions

Communicate the results

DATA WRANGLING:

I used Jupyter Notebook – Read, Asses and Clean Data

While Analyzing the main parameters:

Country

Year

Suicide no

Population

Gender

Unique categorical values are found that there are:

- 101 countries
- 7 regions
- 3 income groups
- 6 Generations
- 6 age groups

EXPLORATORY DATA ANALYSIS(EDA)

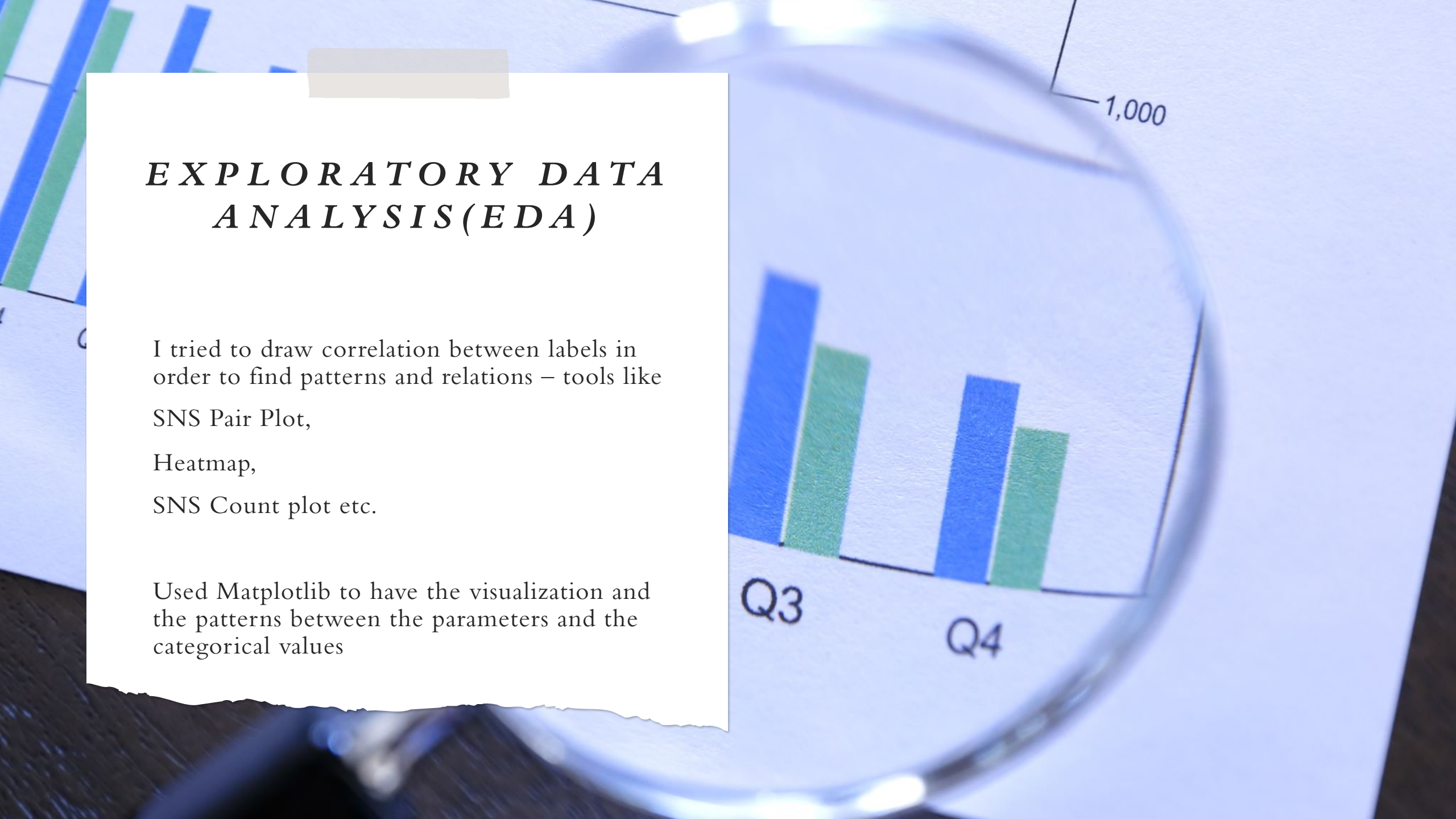
I tried to draw correlation between labels in order to find patterns and relations – tools like

SNS Pair Plot,

Heatmap,

SNS Count plot etc.

Used Matplotlib to have the visualization and the patterns between the parameters and the categorical values



DATA PROCESSING

Feature Engineering:

I added latitude & longitude features which helped to regionize the countries.

Also the dataset is handled with a better proficiency as neighbouring countries have similar circumstances that may affect civils and their suicidal rates.

Feature transformation:

As the dataset had a lot of features and dummy data existence, so in model training I used *binary encoding* to *get dummies* which helped the data to get dimensionally reduced.

Seeing that the variance of data is large that the outliers showing that the data has outliers, I decided:

- There is nothing to be deleted.
- Not to use a ML model that is affected by outliers.

Feature selection:

As I am dealing with many features, I applied Multiple linear regression and used Random Forest Regressor Model to predict.

QUESTIONS I ASKED:

What is the age range on which suicidal actions are the most?

What is the most correlated feature with the increase of suicidal actions?

What is the expected number of suicides given other labels that is related to it?

ANSWERS I PERCEIVED:

Q: *1. What is the age range on which suicidal actions are the most ?*

It is seen that the suicide number increases in the age range(35-54) Which is at mostly the working age group and it is low in age ranges beyond 75 and lower than 14.

Q:2. *What is the most correlated feature with the increase of suicidal actions?*

The correlation function shows that there is a high correlation between population and suicide number so I have gone into regression to see the correlation visualized and to know its type I found that is a positive correlation between suicide number and population.

As I looked into population two different Gender I found that: The suicides among males is more than that among females. The sex label values are nearly equal in the population, so this indicates that the previous result that males commit more suicides than females is true.

There are some historical features those are neglected as it has no purpose in this work. As an example the political era for each age range in each country, and the years that had the most suicidal actions

- As an example, the huge increase of suicidal actions in Russian federations at the years of the cold-war.

Q3. *What is the expected number of suicides given other labels that is related to it?*

See my Model