MSDS 422

Assignment 1.

COVID 19 EDA Project

By: Naseer Fhaeem

Summary:

In this project, I performed an exploratory data analysis of the COVID 19 dataset provided by ecdc.europa.eu. The purpose of this project is to apply basic descriptive statistics, visualization, feature engineering, and feature scaling to understand the data and the underlying patterns.

Data Preparation:

I start my data preparation by downloading the dataset from the source, which comes in a CSV format. I then load the dataset into Python using Pandas' *read_csv* function. Once I get the data to my local machine, I display a few columns to see the structure of the data. I then renamed two columns to be more intuitive and easier to work with throughout the data analysis process.

Data Exploration:

Next, I explore the dataset set by running descriptive statistics functions on the dataframe. I also look deep into the data by using Pandas' *scatter_matrix* to see the relationship between the numerical variables of the dataset. Then, I summarized the dataset by grouping all the data by countries and summing the daily cases and deaths.

First, I visualized the two dataframes that I created to determine which countries have the highest number of COVID 19 cases and deaths. After visualizing the top 25 countries in terms of total cases and total deaths, I noticed that the US was on top of the list. However, I know that this was not the right way

to compare countries since they wouldn't be on the same scale as they have different populations. I started feature engineering by converting the populations of each country to millions. Then, I calculated cases per million and deaths per million. In order to narrow down my focus, I selected Brazil, China, India, Italy, Qatar, and the US to include in my visualization. I graphed a daily line chart for the countries that I picked and realized that India's cases were so low that it was not comparable with other countries. I then tried to aggregate the total cases per million to each month to see a better trend of the COVID 19 cases. I still was not able to see India's COVID 19 cases on the same graph as other countries. To fix this issue, I had to scale the total number of COVID 19 cases of the selected countries. I did so by using Scikit-learn's *MinMaxScaler*. The MinMaxScaler, scaled all the values between 0 and 1.

After scaling the total number of cases per million on monthly basis, I was able to see the total monthly COVID 19 cases for all the countries on my list and could compare

Conclusion:

It's always important to inspect a data set both manually and programmatically. Before looking at the data, I thought that the US had the worst case of COVID 19. My first two graphs followed this bias and showed that the US has the highest number of COVID 19 cases and deaths. However, after a few feature engineering and scaling I realized that the US is not even in the top 5 highest COVID 19 cases per million.

For the EDA notebook on github, click here: