Report 1: Analysis and Visualization

Sai Teja Reddy Palam (202106549)
St. Francis Xavier University

1 Analysis

The statistical analysis in the COVID-19 Symptoms Checker data-set can be expressed in tabular form as shown below:

	Tiredness	Dry-Cough	Difficulty-in-Breathing	Sore-Throat
Range	[1, 0]	[1, 0]	[1, 0]	[1, 0]
Mean	0.5	0.5625	0.5	0.3125
Mode	1	1	1	0
Standard Deviation	0.500001	0.496079	0.500001	0.463513
Median	0.5	1.0	0.5	0.0

	None_Sympton	Pains	Nasal-Congestion	Runny-Nose
Range	[1, 0]	[1, 0]	[1, 0]	[1, 0]
Mean	0.0625	0.363636	0.545455	0.545455
Mode	0	0	1	1
Standard Deviation	0.242062	0.481046	0.49793	0.49793
Median	0.0	0.0	1.0	1.0

	Diarrhea	None_Experiencing	...	Gender_Female	Gender_Male
Range	[1, 0]	[1, 0]	...	[1, 0]	[1, 0]
Mean	0.363636	0.090909	...	0.333333	0.333333
Mode	0	0	...	0	0
Standard Deviation	0.481046	0.28748	...	0.471405	0.471405
Median	0.0	0.0	...	0.0	0.0

	Gender_Transgender	Severity_Mild	Severity_Moderate
Range	[1, 0]	[1, 0]	[1, 0]
Mean	0.333333	0.25	0.25
Mode	0	0	0
Standard Deviation	0.471405	0.433013	0.433013
Median	0.0	0.0	0.0

	Severity_None	Severity_Severe	Contact_Dont-Know	Contact_No
Range	[1, 0]	[1, 0]	[1, 0]	[1, 0]
Mean	0.25	0.25	0.333333	0.333333
Mode	0	0	0	0
Standard Deviation	0.433013	0.433013	0.471405	0.471405
Median	0.0	0.0	0.0	0.0

	Contact_Yes
Range	[1, 0]
Mean	0.333333
Mode	0
Standard Deviation	0.471405
Median	0.0

[5 rows x 25 columns]

Figure 1: Analysis

Aside from statistical measurements such as Range, Mean, and Mode, I had chosen two more metrics which are Standard Deviation and Median for analyzing the data set. The metrics are explained as follows:

- Range: A data set's range is the difference between its maximum and minimum values.
- Mean: A data set's mean, or arithmetic mean, is the sum of all values divided by the total number of values.
- Mode: The most frequent number that is the number that occurs the most frequently.
- Standard Deviation: It tells us how much the supplied values depart from the mean value for the attributes present int the data set.
- Median is the value in the middle of a data set which means that 50% of the data points have a value less than or equal to the median, and 50% of the data points have a value greater than or equal to the median.

Since all the attribute values except the Country in the data-set have either 0's or 1's which tells whether a person is effected or not, that is False: "0" or True: "1" against the list of Symptoms and also tells the age under the age category, Gender and Contact, so the range for all of these attributes are [1, 0]. For the provided data set, the difference between the the Standard Deviation and the Mean values for attributes Tiredness, Difficulty-in-Breathing is close to zero. For attributes Severity_Mild, Severity_Moderate, Severity_None and Severity_Severe the different between the Standard Deviation and the Mean values are the same. The most frequent number which is the Mode for some of the attributes like Dry-Cough, Sore-Throat, None_Symptom, Pains are the same as Median.

2 Visualization

2.1 Pie Chart

The Pie Chart below depicts the Severity of the patients present in the data-set. Severity_Mild, Severity_Moderate, and Severity_Severe are the attributes evaluated for this visualization. In this case, the entire pie chart symbolizes 100 percent and the share of each severity is depicted in distinct colors with respective percentages showing

various categorical serveries of the given individuals. From Figure 2, we can infer that there are many Moderate (Severity_Moderate) cases as compared to other categories and least number of Mild (Severity_Mild) cases across the data-set.

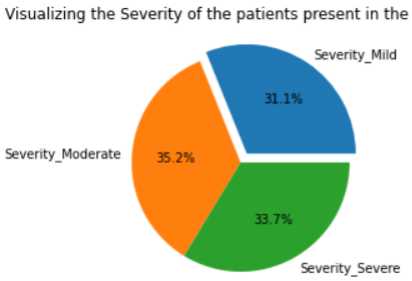

Figure 2: Pie Chart

2.2 Polar Chart

The Polar Chart below depicts the number of Males and Females who are affected by Covid-19 under each age category has been visualized. For this visualization, the following attributes are used which are Age_0-9, Age_10-19, Age_20-24, Age_25-59, Age_60+, Gender_Female, Gender_Male. The entire polar chart symbolizes 100 percent of the Men and Women where each category is depicted in distinct colors, with respective percentage values represented in numbers across all ages. From Figure 3, we can infer that there are more Men who are affected by Covid as compared to Women under age category Age_0-9 and under age category "Age_10-19, it is the opposite where more Women are affected by Covid as compared to Men.

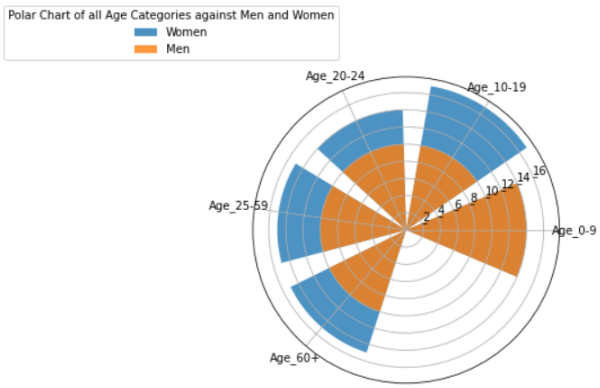

Figure 3: Polar Chart

2.3 Donut Chart

The Donut Chart below depicts number of Women who are affected by Covid-19 under each age category has been visualized. For this visualization, the following attributes are used which are Age_0-9, Age_10-19, Age_20-24, Age_25-59, Age_60+, Gender_Female. The entire Donut Chart consists of percentage of Women who are affected by Covid with respective percentage values across all age categories. From Figure 4, we can infer that the least number of women who are affected by Covid are under age category "Age_0-9" while the women who are above 25 years are largely affected.

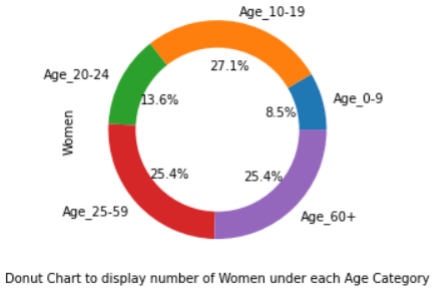

Figure 4: Donut Chart

2.4 Correlation Matrix

This is a correlation matrix which illustrates the relationship between variables or attributes in our case. The correlation between all possible pairs of values are shown in a matrix format. By looking at the Figure 5, we can infer that the relationship is strong if the relationship score is 1, if the relationship is zero then it indicates that the relationship is neural and if the relationship is -1 then it indicates that the relationship is negative or weak.

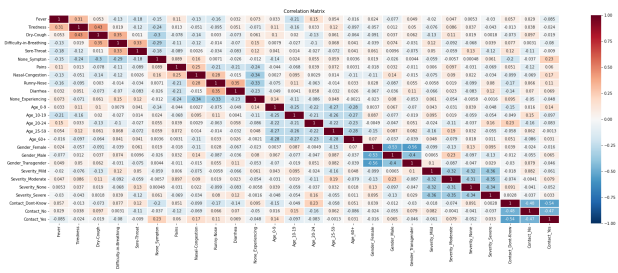

Figure 5: Correlation Matrix

2.5 Bar Chart

The Bar Chart below depicts number of Men and Women who are affected by Covid-19 under each Age Category has been visualized. For this visualization the following attributes are used which are Age_0-9, Age_10-19, Age_20-24, Age_25-59, Age_60+, Gender_Female, Gender_Male. From Figure 6, we can infer that the majority of the Covid cases are seen in Women under age category "Age_10-19", "Age_20-24" compared to Men while the majority of Men who are affected are the elder people (Age_60+).

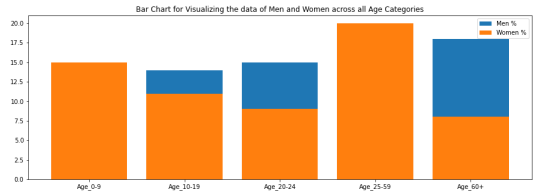

Figure 6: Bar Chart

2.6 Count Plot

The Count Plot is used to display the number of occurrences of the observation in the categorical attributes. The following are Count Plots which are plotted against Country alone, Country versus the people who are experiencing Fever as a symptom and Country versus the people who are not experiencing any kind of symptoms.

2.6.1 Count Plot for Country

This plot displays the number of people who are affected by Covid in each Country present in the data-set. From Figure 7, we can infer that Italy has highest number of cases which has been affected the most while the least affected is the United Arab Emirates (UAE) when taking countries into consideration and also there is "Other" which contains the cases from all over the world.

2.6.2 Count Plot for Country vs Fever Symptom

This Count Plot depicts the number of people in the data-set who are affected by Covid-19 and having Fever as a symptom and the people who do not have it against each country is plotted.

- Fever: 0 is the number of people who are affected by Covid but not having the symptom

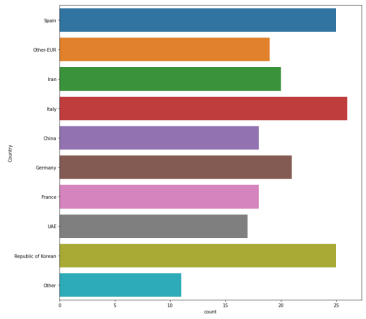

Figure 7: Count Plot for Country

- Fever: 1 is the number of people who are affected by Covid but having the symptom

From Figure 8, we can infer that Italy has the highest number of people who are affected by Covid but not suffering from Fever while the people from Spain and Other-EUR has been experiencing the most when compared to all other countries.

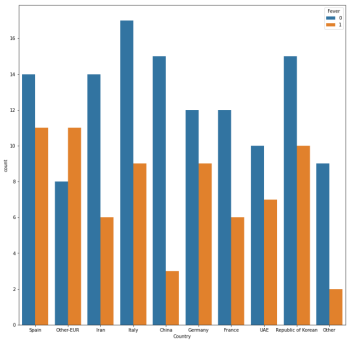

Figure 8: Count Plot for Country vs Fever Symptom

2.6.3 Count Plot for Country vs None_Experiencing Symptom

This Count Plot depicts the number of people in the data-set who are affected by Covid-19 but not experiencing any kind of symptoms against each country is plotted.

- None_Experiencing: 0 is the number of people who are affected by Covid and having the symptoms
- None_Experiencing: 1 is the number of people who are affected by Covid but not experiencing any kind of symptoms

From Figure 9, we can infer that Iran, China and UAE have the least number of people who are af-

fected by Covid but not experiencing any kind of symptoms while it is the highest for Italy.

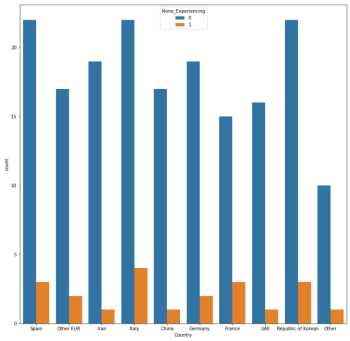

Figure 9: Count Plot for Country vs None_Experiencing Symptom