Predicting COVID-19 Infection

Team Information: Nikhila Chiluka, Anamika Chatterjee

1 Introduction

Since 2019, we have been experiencing the Covid-19 pandemic. The cause of this pandemic has had a significant impact on the lives of citizens as well as the global economy. One of the most important tasks in this pandemic has been determining whether a person is infected with Severe Acute Respiratory Syndrome Coronavirus 2. (SARS-CoV-2).

To accomplish this, we are using a dataset that records all possible symptoms, physiological, and geographical parameters to predict infections.

The primary goal is to identify the best model for predicting COVID-19 infection in a person.

2 Dataset

The dataset used in the project is a covid-19 symptom dataset.(HUNGUND, 2019) It is a dataset used to detect the infection of Covid-19 in an individual. The dataset was created based on the WHO guidelines and Ministry of Health and Family Welfare, India. (WHO, 2019)(MoHFW, 2019)

2.1 Dataset Description

The shape of the dataset is 316800x27.

2.2 Samples

The dataset consists of 316800 records.

2.3 Measurements

Measurements: The dataset consists of 27 features, out of which 23 features describe the symptoms, age, gender, contact and country. The remaining four features describe the severity of infection.

3 Description of Measurements

Fever: It determines an individual's fever symptom. This field accepts integers.

Tiredness: It determines an individual's tiredness symptom. This field accepts integers.

Dry-Cough: This field accepts integers and determines an individual's Dry Cough symptom.

Difficulty-in-Breathing: It determines the difficulty in breathing symptom in an individual. This field takes integers.

Sore-Throat: It determines the sore throat symptom in an individual. This field takes integers.

Pains: It determines the Pains symptom in an individual. This field takes integers.

Nasal-Congestion: It determines the nasal congestion symptom in an individual. This field takes integers.

Runny-Nose: It determines the runny nose symptom in an individual. This field takes integers.

Diarrhea: It determines the diarrhea symptom in an individual. This field takes integers.

Gender_Female: It determines the gender of an individual. It takes integers.

Gender_Male: It determines the gender of an individual. It takes integers.

Gender_Transgender: It determines the gender of an individual. It takes integers.

Severity_Mild: It determines the severity of infection in an individual. It takes integers.

Severity_Moderate: It determines the severity of infection in an individual. It takes integers.

Contact_Dont_Know: This feature tells if the infected person has contact. It takes integers.

Contact_No: This feature tells if the infected person has contact. It takes integers.

Contact_Yes: This feature tells if the infected person has contact. It takes integers.

Country: This feature explains the country of the individual.

No_Common_Symptoms: It determines if the individual has common symptoms or not. This field takes integers.

Age_0-9: It determines the age group of an indi-

vidual. It takes integers.
Age_10-19: It determines the age group of an individual. It takes integers.
Age_20-24: It determines the age group of an individual. It takes integers.
Age_25-59: It determines the age group of an individual. It takes integers.
Age_60+: It determines the age group of an individual. It takes integers.
Severity_None: It determines the severity of infection in an individual. It takes integers.
Severity_Severe: It determines the severity of infection in an individual. It takes integers.
No_Other_Symptoms: It determines if the individual has other symptoms or not. This field takes integers.

4 Data Cleaning

4.1 Data Sampling

The shape of the original dataset is 316800x27. A subset of the total data is taken into account during analysis and visualisation. Therefore, a sample of 75% of the data was taken and used for the observations.

4.2 Data Cleaning

The dataset has 27 features as discussed in the section

Taking into account the characteristics that are associated with the Covid 19 Infection's severity. The dataset contains four features(Columns) that include information on the severity as follows: Severity_None, Severity_Mild, Severity_Moderate, and Severity_Severe. The above Columns are merged into a single feature as Condition. Then the categories were converted to numeric data for convenience in analysis.

The features Gender_male, Gender_female, and Gender_Transgender that include information on an individual's gender were combined into one feature called Gender, and categorical data was then converted.

The features that contained the data indicating whether the person had contacts were also combined into one feature called Contact. The merged columns were Contact_yes, Contact_No, Contact_DontKnow. The number of features are now reduced to 20.

5 Analysis

The statistical metrics are performed to identify the trends in the dataset. The Covid-19 dataset was subjected to a variety of calculations, including mean, mode, range, total, standard deviation, correlation, and variance. In Figure 1, the calculated values are shown.

5.1 Range

Range always gives the interval of minimum and maximum value in a feature or dataset. As mentioned, the features are either binarized or categorical so most of the values are 0 and 1 or numerical values that represent distinct categories. Due to this nature of the dataset the range metric does not yield much insight into the dataset as for most of the columns except for the columns which are merged it only helps to identify the extreme values in a dataset which is in case of binary data it is already known to be 0 and 1.

5.2 Mean

Mean is the most frequently used measure of central tendency it will give the average value of an attribute in a dataset. Mean of a binary attribute can help us in identifying whether that attribute has more or less presence in our dataset, a mean value ¡0.5 signifies less presence of the binary attribute whereas the mean value ¿0.5 signifies more presence. For other categorical attributes having numerical values from 0-2 or 0-3 a mean value which is closer to 0 will imply more presence of categories 0 and 1 in the attribute while a mean value closer to 2 or 3 will signify more presence of categories 2 or 3.

5.3 Mode

:Mode helps to identify the most frequently occurring value in an attribute of the dataset. It can give a measure of the central tendency of the dataset. For our dataset mode is quite a significant metric. The mode helps us to identify which attribute is more prevalent in the binary dataset. For example, the mode of 1 for a symptom attribute will imply that symptom is more prevalent in our dataset since 1 signifies the presence of an attribute in binary data.

5.4 Sum

Sum will give the total count of values in each attribute.

5.5 Variance

Variance is metric that will give us the spread of the attributes in the dataset with respect to their means. It gives an idea of the distribution in our dataset.

5.6 Standard deviation

Standard deviation is always zero or more than zero. It helps to know if the features are closer to or farther from the mean. This metric can help us identify the type of distribution our dataset has.

5.7 Correlation

Correlation factor is calculated to find the similarity between the features. The correlation values are between -1 to +1. If the correlation value is between -1 to 0 then the features are considered to be negatively correlated where as a correlation value of 0 to 1 signifies positive correlation between features. If the correlation value of two features is very close to zero then it implies that the two features are dissimilar. Figure 2 describes the correlated values.

6 Visualizations

6.1 Word Cloud

The first visualization is a word cloud, the word cloud is done with the country feature. It represents the frequency of each country in the feature. The country with highest frequency is appeared with the maximum size like china. Figure 3 represents the word cloud which is done for country vs sum

6.2 Heatmap

The figure 4 represents a Heatmap. It represents the correlation of one feature with the other. The correlation matrix values which are in Figure 2 are used to plot the heatmap.

6.3 Donut chart

The visualization is performed on Condition feature with value counts. In the Donut chart each condition is differentiated by different colours. The dataset is equally distributed so the difference in the infection levels is very minimal. Figure 6 represents the value counts for each infection level.

6.4 Horizontal Bar Chart

The horizontal bar graph is plotted against the different symptoms in an individual in the dataset and the frequency of occurence the symptoms. It helps us to identify which symptom is mostly seen in people. Figure 5 represents the bar chart. From the figure we can deduce that dry-cough is the most prevelant symptom in the dataset whereas people having no symptoms are the least common.

6.5 Line Graph

Figure 7 represents the Line graph. It is a plot of age group and the count of people in different age groups with the age columns.It gives us a clear insight into the distribution of people in different age groups.

7 Distribution Curve

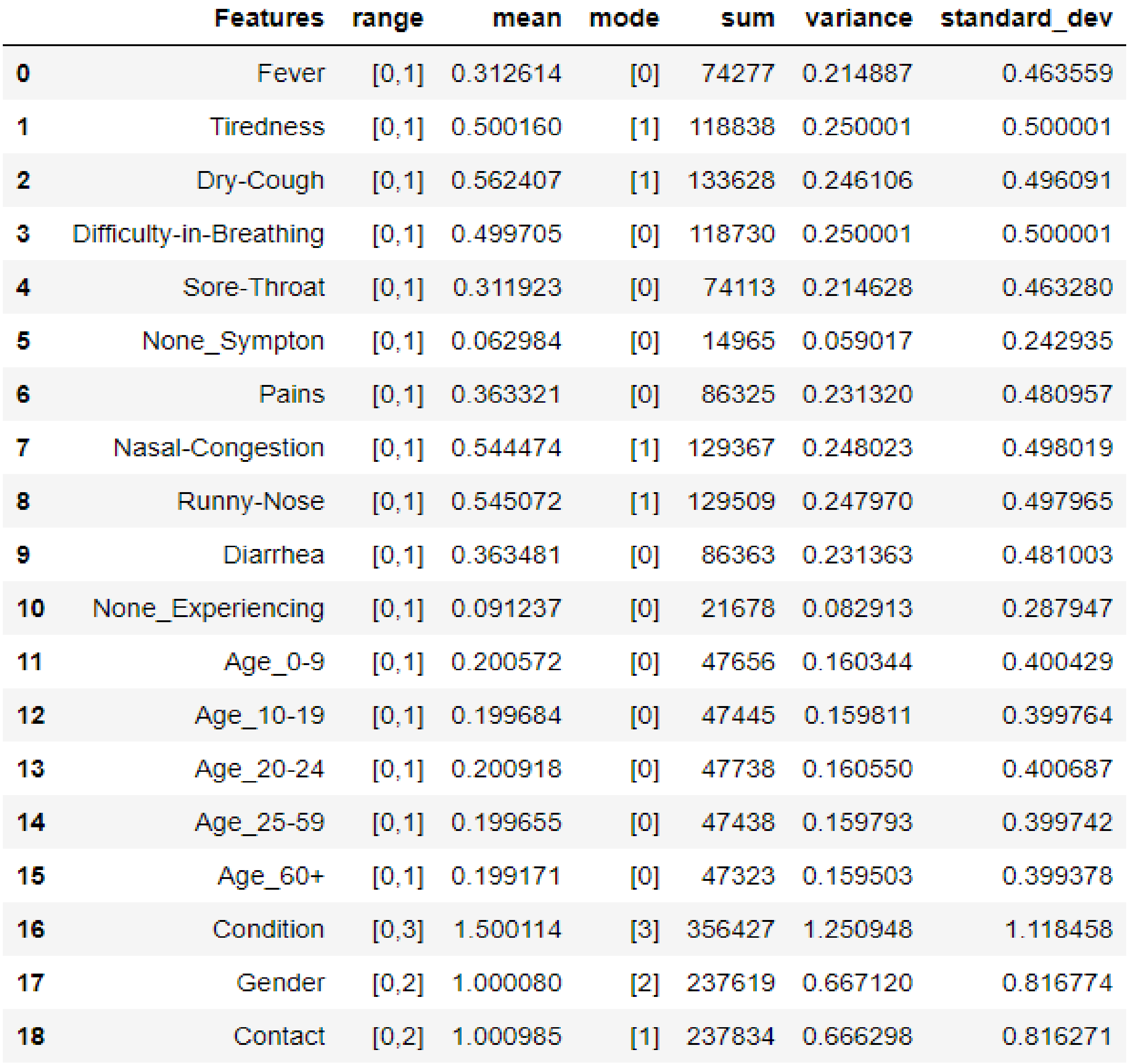
Figure 8 represents the distribution curve of the mean of the features. We observe that the distribution of the means of the dataset gives a curve very close to a normal distribution curve.

References

BILAL HUNGUND. 2019. Covid-19 symptom checker dataset. Kaggle, Dataset.

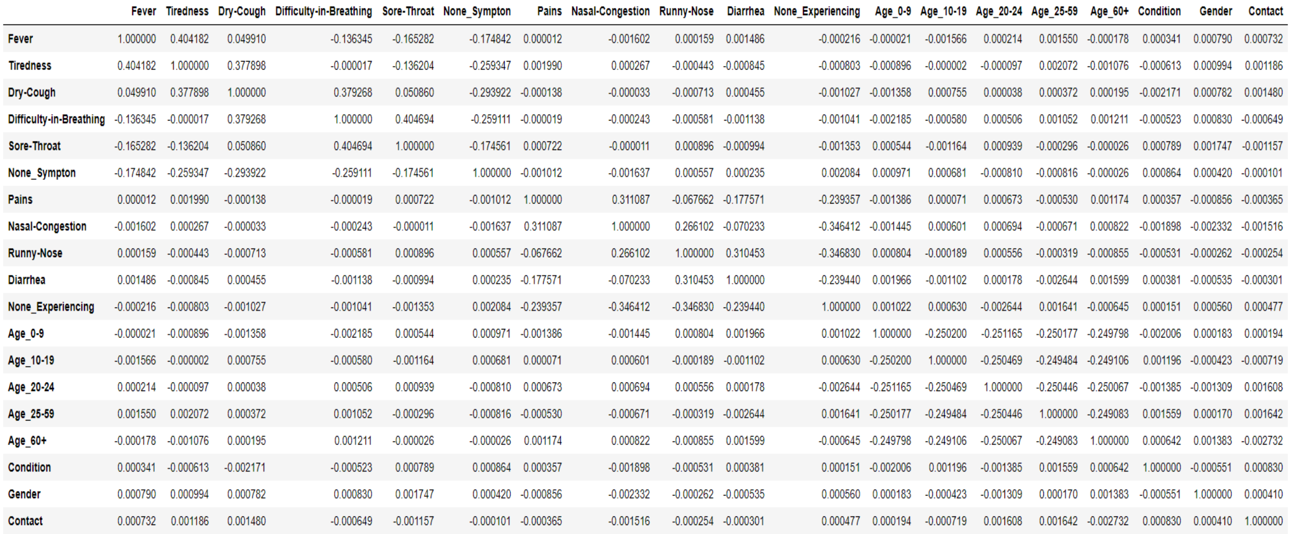
MoHFW. 2019. Ministry of health and family welfare india. MoHFW, MoHFW-Guidelines.

WHO. 2019. World health organisation guidelines on the symptoms of covid-19. WHO, WHO guidelines.



	Features	range	mean	mode	sum	variance	standard_dev
0	Fever	[0,1]	0.312614	[0]	74277	0.214887	0.463559
1	Tiredness	[0,1]	0.500160	[1]	118838	0.250001	0.500001
2	Dry-Cough	[0,1]	0.562407	[1]	133628	0.246106	0.496091
3	Difficulty-in-Breathing	[0,1]	0.499705	[0]	118730	0.250001	0.500001
4	Sore-Throat	[0,1]	0.311923	[0]	74113	0.214628	0.463280
5	None_Sympton	[0,1]	0.062984	[0]	14965	0.059017	0.242935
6	Pains	[0,1]	0.363321	[0]	86325	0.231320	0.480957
7	Nasal-Congestion	[0,1]	0.544474	[1]	129367	0.248023	0.498019
8	Runny-Nose	[0,1]	0.545072	[1]	129509	0.247970	0.497965
9	Diarrhea	[0,1]	0.363481	[0]	86363	0.231363	0.481003
10	None_Experiencing	[0,1]	0.091237	[0]	21678	0.082913	0.287947
11	Age_0-9	[0,1]	0.200572	[0]	47656	0.160344	0.400429
12	Age_10-19	[0,1]	0.199684	[0]	47445	0.159811	0.399764
13	Age_20-24	[0,1]	0.200918	[0]	47738	0.160550	0.400687
14	Age_25-59	[0,1]	0.199655	[0]	47438	0.159793	0.399742
15	Age_60+	[0,1]	0.199171	[0]	47323	0.159503	0.399378
16	Condition	[0,3]	1.500114	[3]	356427	1.250948	1.118458
17	Gender	[0,2]	1.000080	[2]	237619	0.667120	0.816774
18	Contact	[0,2]	1.000985	[1]	237834	0.666298	0.816271

Figure 1: Statistical Metrics



	Fever	Tiredness	Dry-Cough	Difficulty-in-Breathing	Sore-Throat	None_Sympton	Pains	Nasal-Congestion	Runny-Nose	Diarrhea	None_Experiencing	Age_0-9	Age_10-19	Age_20-24	Age_25-59	Age_60+	Condition	Gender	Contact
Fever	1.000000	0.404182	0.049910	-0.136345	-0.165282	-0.174842	0.000012	-0.001602	0.000159	0.001486	-0.000216	-0.000021	-0.001566	0.000214	0.001550	-0.000178	0.000341	0.000790	0.000732
Tiredness	0.404182	1.000000	0.377898	-0.000017	-0.136204	-0.259347	0.001990	0.000267	-0.000443	-0.000845	-0.000803	-0.000896	-0.000002	-0.000097	0.002072	-0.001076	-0.000613	0.000994	0.001186
Dry-Cough	0.049910	0.377898	1.000000	0.379268	0.050860	-0.293922	-0.000138	-0.000033	-0.000713	0.000455	-0.001027	-0.001358	0.000755	0.000038	0.000372	0.000195	-0.002171	0.000782	0.001480
Difficulty-in-Breathing	-0.136345	-0.000017	0.379268	1.000000	0.404694	-0.259111	-0.000019	-0.000243	-0.000581	-0.001138	-0.001041	-0.002185	-0.000580	0.000506	0.001052	0.001211	-0.000523	0.000830	-0.000649
Sore-Throat	-0.165282	-0.136204	0.050860	0.404694	1.000000	-0.174561	0.000722	-0.000011	0.000896	-0.000994	-0.001353	0.000544	-0.001164	0.000939	-0.000296	-0.000026	0.000789	0.001747	-0.001157
None_Sympton	-0.174842	-0.259347	-0.293922	-0.259111	-0.174561	1.000000	-0.001012	-0.001637	0.000557	0.000235	0.002084	0.000971	0.000681	-0.000810	-0.000816	-0.000026	0.000864	0.000420	-0.000101
Pains	0.000012	0.001990	-0.000138	-0.000019	0.000722	-0.001012	1.000000	0.311087	-0.067662	-0.177571	-0.239357	-0.001386	0.000071	0.000673	-0.000530	0.001174	0.000357	-0.000856	-0.000365
Nasal-Congestion	-0.001602	0.000267	-0.000033	-0.000243	-0.000011	-0.001637	0.311087	1.000000	0.266102	-0.070233	-0.346412	-0.001445	0.000601	0.000694	-0.000671	0.000822	-0.001898	-0.002332	-0.001516
Runny-Nose	0.000159	-0.000443	-0.000713	-0.000581	0.000896	0.000557	-0.067662	0.266102	1.000000	0.310453	-0.346830	0.000804	-0.000189	0.000556	-0.000319	-0.000855	-0.000531	-0.000262	-0.000254
Diarrhea	0.001486	-0.000845	0.000455	-0.001138	-0.000994	0.000235	-0.177571	-0.070233	0.310453	1.000000	-0.239440	0.001966	-0.001102	0.000178	-0.002644	0.001599	0.000381	-0.000535	-0.000301
None_Experiencing	-0.000216	-0.000803	-0.001027	-0.001041	-0.001353	0.002084	-0.239357	-0.346412	-0.346830	-0.239440	1.000000	0.001022	0.000630	-0.002644	0.001641	-0.000645	0.000151	0.000560	0.000477
Age_0-9	-0.000021	-0.000896	-0.001358	-0.002185	0.000544	0.000971	-0.001386	-0.001445	0.000804	0.001966	0.001022	1.000000	-0.250200	-0.251165	-0.250177	-0.249798	-0.002006	0.000183	0.000194
Age_10-19	-0.001566	-0.000002	0.000755	-0.000580	-0.001164	0.000681	0.000071	0.000601	-0.000189	-0.001102	0.000630	-0.250200	1.000000	-0.250469	-0.249484	-0.249106	0.001196	-0.000423	-0.000719
Age_20-24	0.000214	-0.000097	0.000038	0.000506	0.000939	-0.000810	0.000673	0.000694	0.000556	0.000178	-0.002644	-0.251165	-0.250469	1.000000	-0.250446	-0.250067	-0.001385	-0.001309	0.001608
Age_25-59	0.001550	0.002072	0.000372	0.001052	-0.000296	-0.000816	-0.000530	-0.000671	-0.000319	-0.002644	0.001641	-0.250177	-0.249484	-0.250446	1.000000	-0.249083	0.001559	0.000170	0.001642
Age_60+	-0.000178	-0.001076	0.000195	0.001211	-0.000026	-0.000026	0.001174	0.000822	-0.000855	0.001599	-0.000645	-0.249798	-0.249106	-0.250067	-0.249083	1.000000	0.000642	0.001383	-0.002732
Condition	0.000341	-0.000613	-0.002171	-0.000523	0.000789	0.000864	0.000357	-0.001898	-0.000531	0.000381	0.000151	-0.002006	0.001196	-0.001385	0.001559	0.000642	1.000000	-0.000551	0.000830
Gender	0.000790	0.000994	0.000782	0.000830	0.001747	0.000420	-0.000856	-0.002332	-0.000262	-0.000535	0.000560	0.000183	-0.000423	-0.001309	0.000170	0.001383	-0.000551	1.000000	0.000410
Contact	0.000732	0.001186	0.001480	-0.000649	-0.001157	-0.000101	-0.000365	-0.001516	-0.000254	-0.000301	0.000477	0.000194	-0.000719	0.001608	0.001642	-0.002732	0.000830	0.000410	1.000000

Figure 2: Correlations

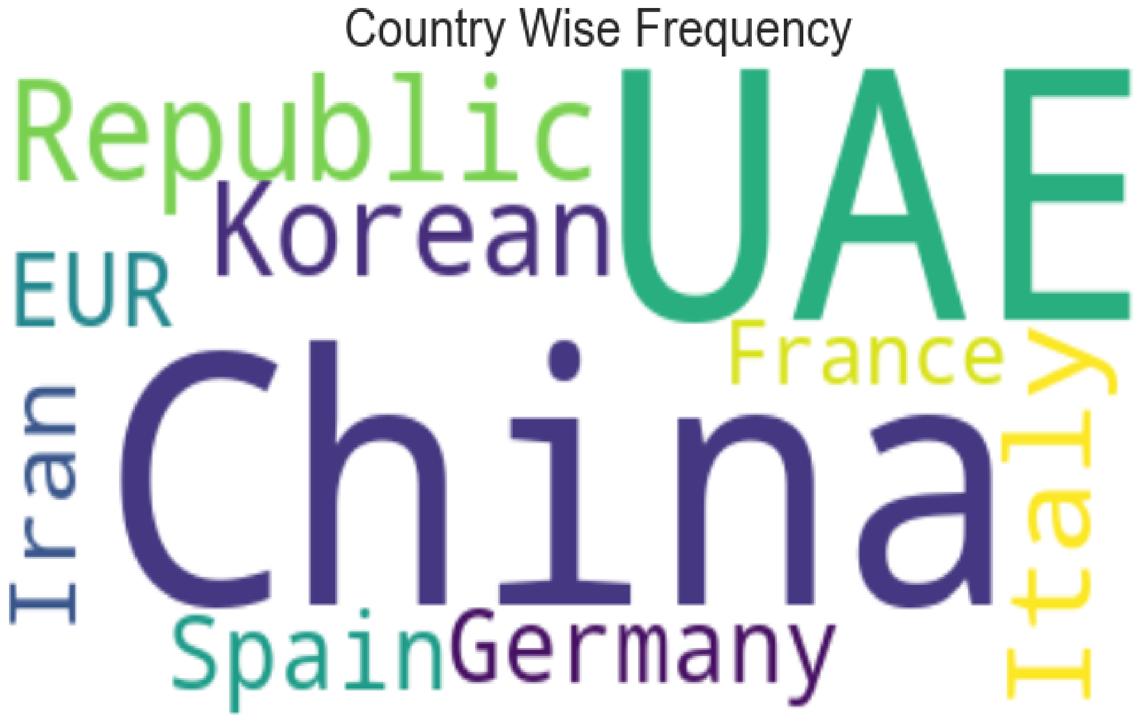


Figure 3: Word Cloud

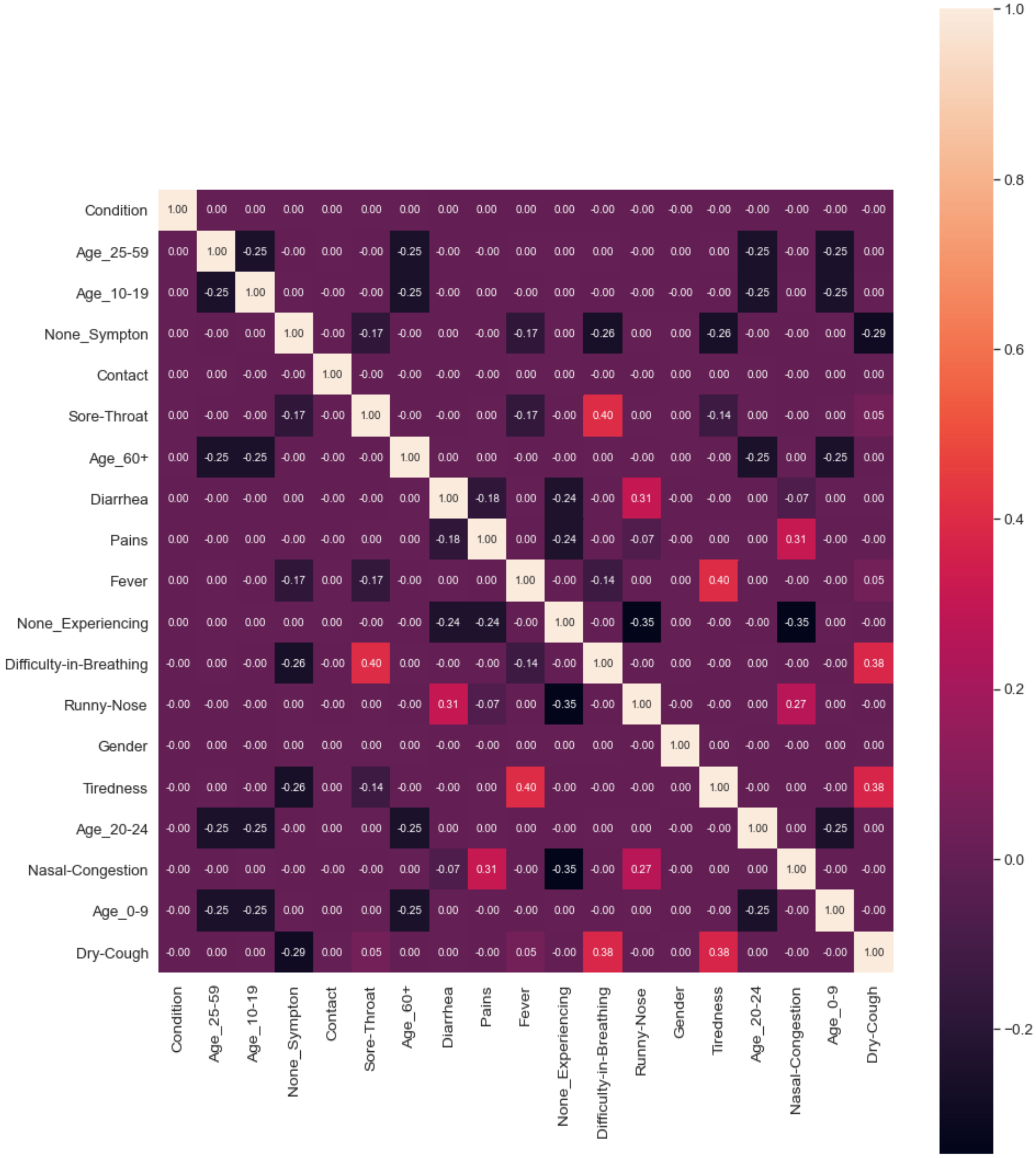


Figure 4: HeatMap

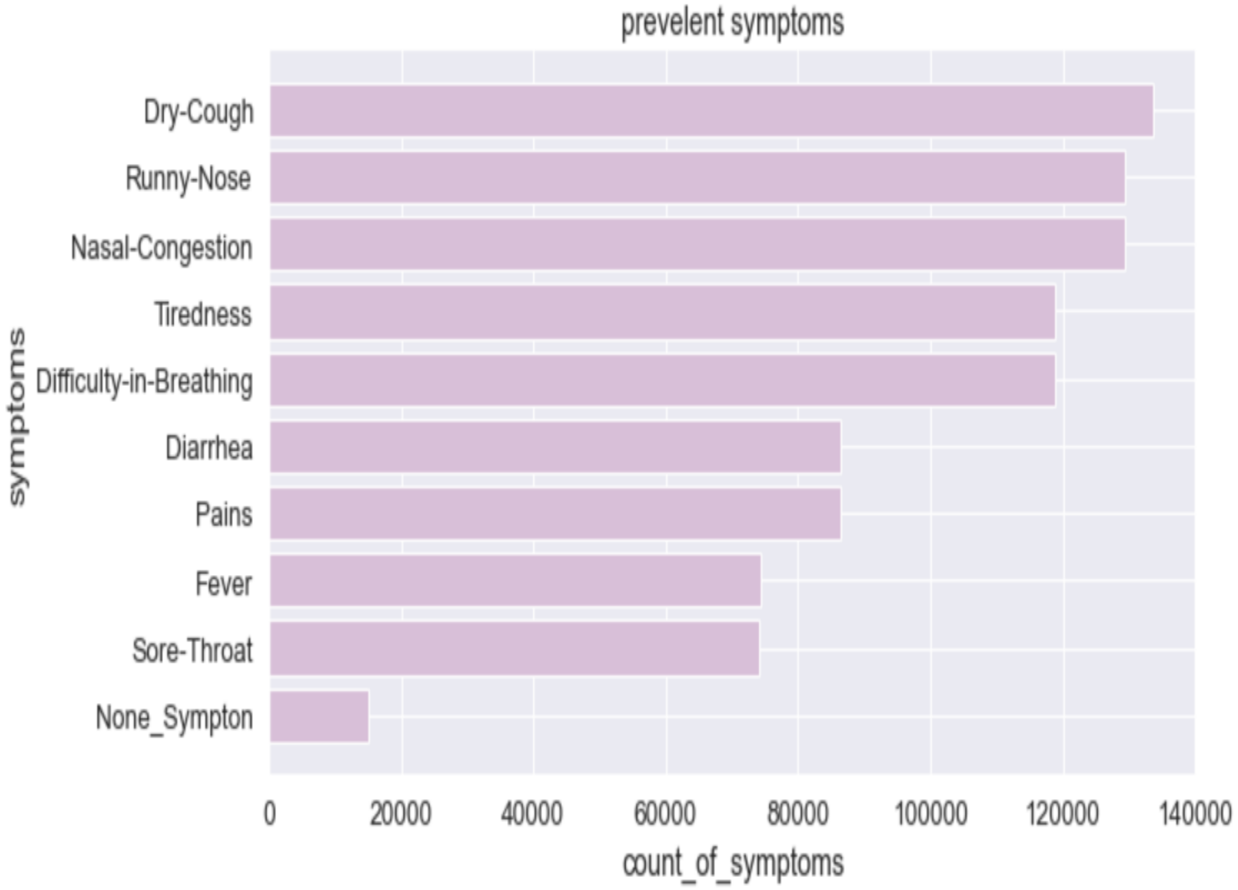


Figure 5: Prevelant Symptoms Bar Graph

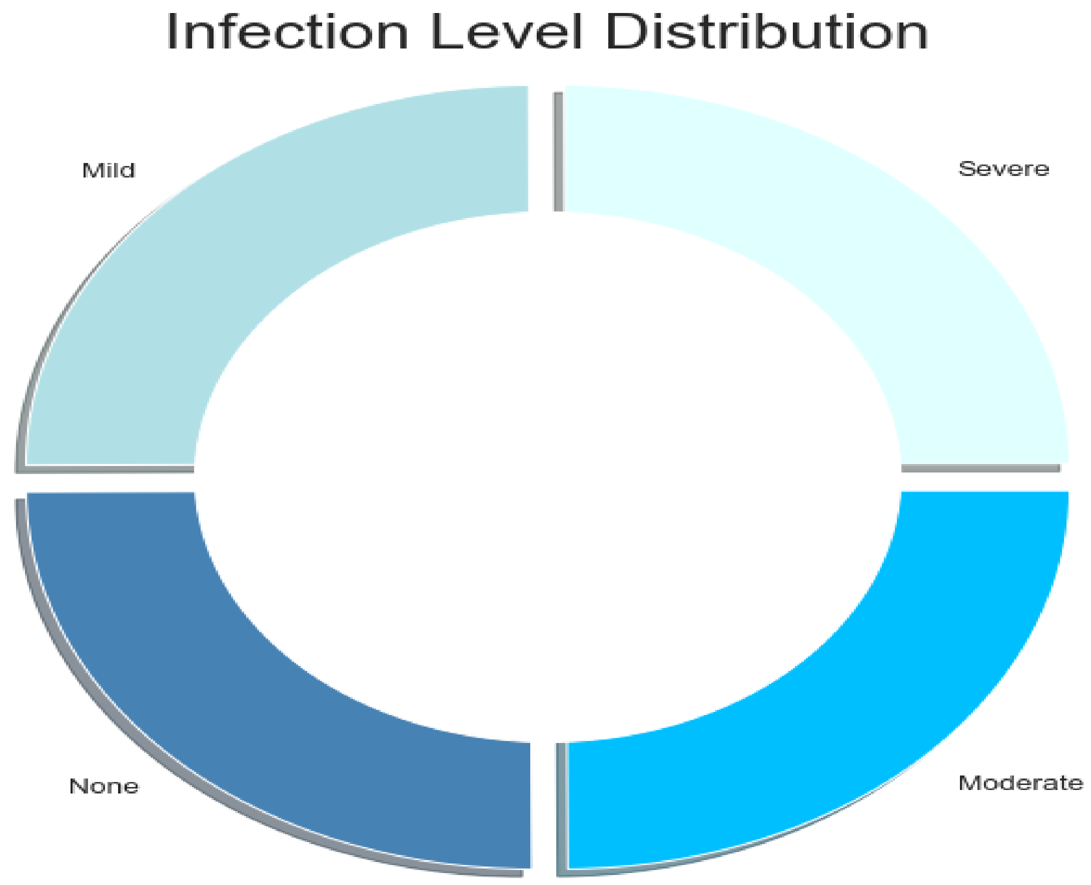


Figure 6: Infection Level in a Donut chart

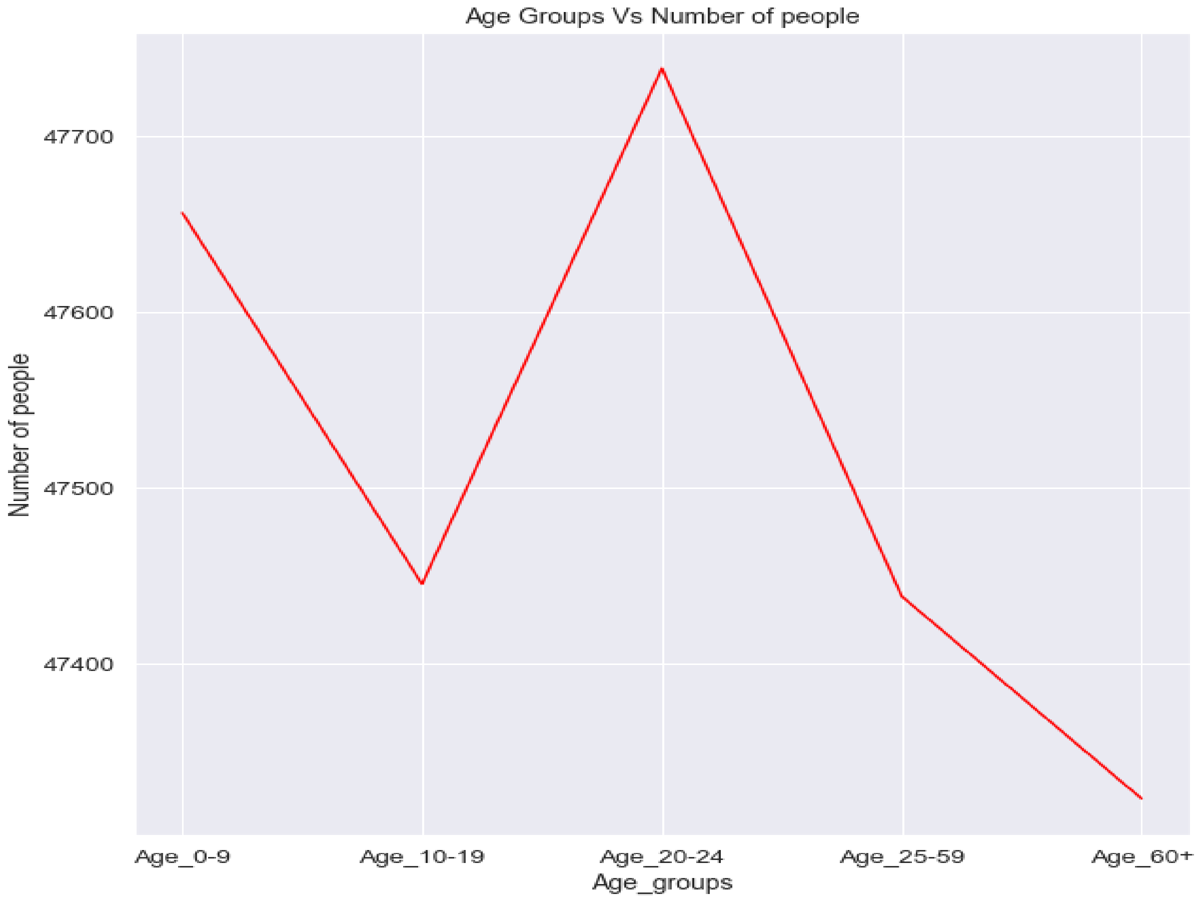


Figure 7: Age Group in a Line chart

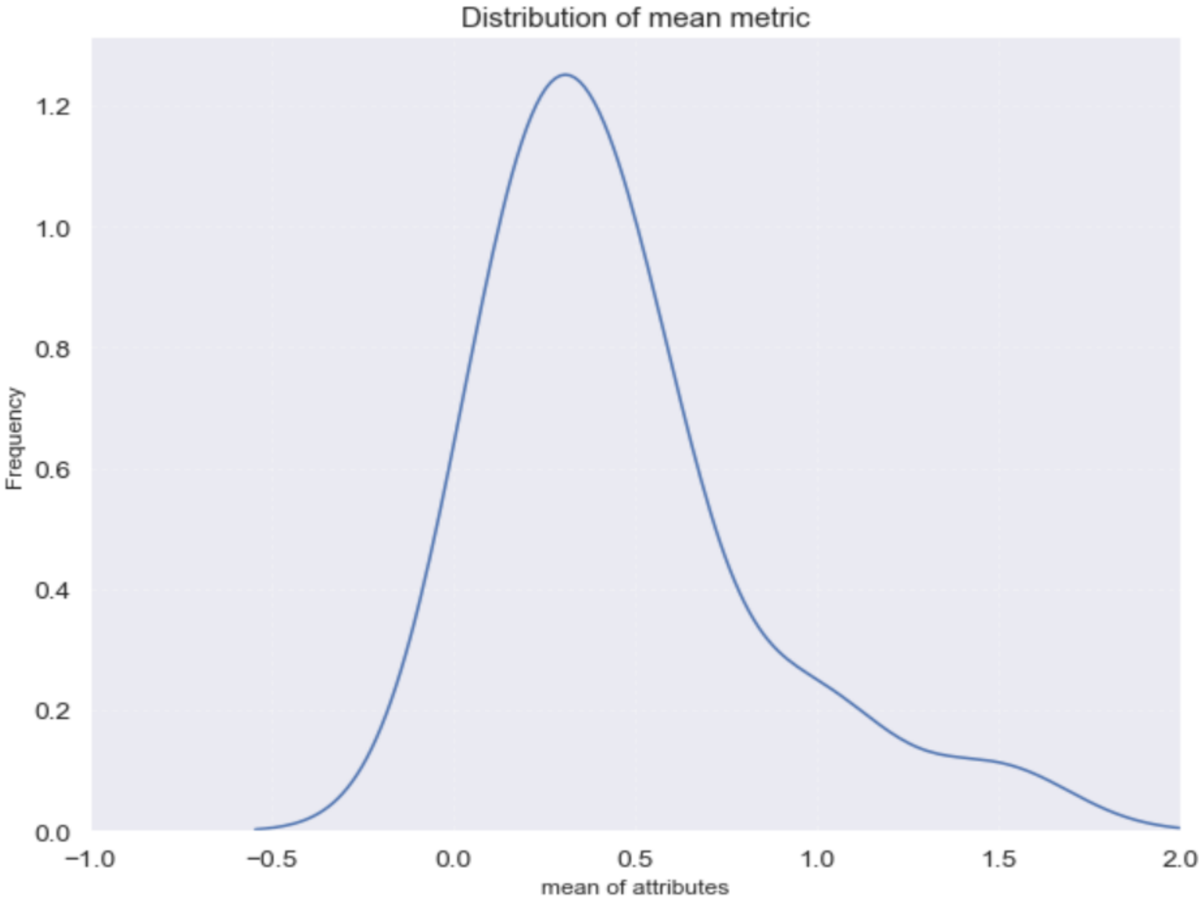


Figure 8: Distribution Curve