COVID-19 Data Analysis

Erik Nuroyan

The problem/data description

Introduction

As the virus COVID-19 spreads all over the globe, taking lives of a lot of people and giving more and more data day by day which needs processing, this project concentrates on analyzing the data gathered from distinct regions of the United States and to provide comprehensive descriptions of findings which may be used by various medical and political experts to determine further steps to prevent the increase in the number of infected people.

Data Information

The data comes from the results of the questionnaire answered by people anonymously on the website https://infogears.org/ [1]. There is not much processing done on the data. Some work is done to understand correlation between the family size and mental stress level or correlation between the frequency of leaving home and different age groups.
The data consists of 16768 observations of 33 variables. There are 12 variables of character type, 18 of numeric type, 1 of logical type,2 of POSIXct type.

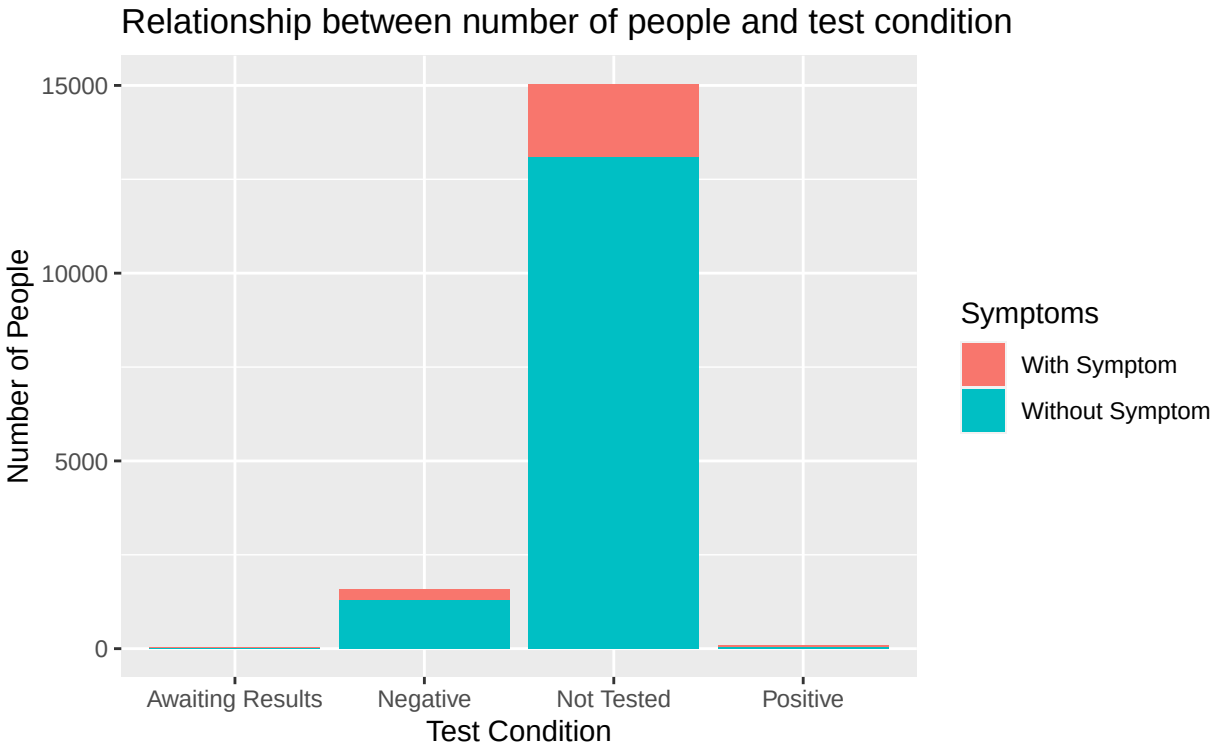
Main hypotheses

1. Those who don't have symptoms are not likely to be tested for COVID-19.
2. People who are tested positive are mostly asymptomatic.
3. The users of this website are mostly of young age.
4. Those who are tested positive and are symptomatic are mostly aged (>65).
5. People having no symptoms are likely to take antibody test.
6. Those who are tested positive have lost sense of smell more than those with negative results.
7. The number of positive tests increased by the time.

Note: The number of the hypothesis in this section corresponds to that of in the Analysis section following this one.

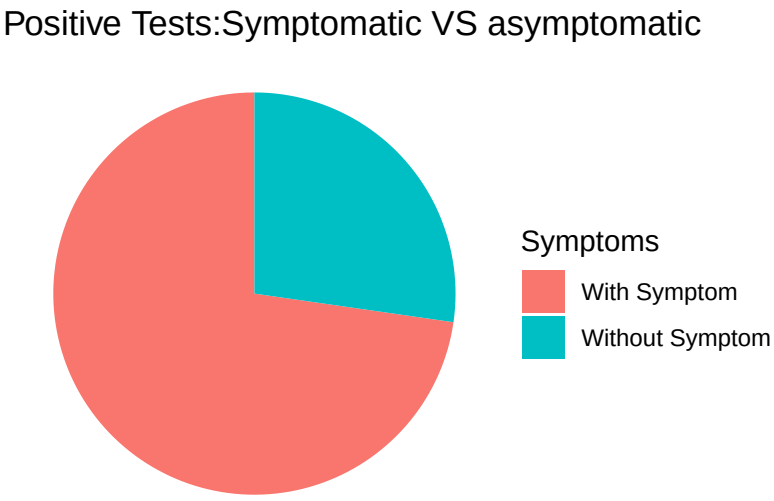
Analysis

1. It is logical to think that those who don't have symptoms are not likely to be tested for the Coronavirus, since they may feel OK and, thus, not find any need of testing. Now, let's look at what we have, when we analyze the data.



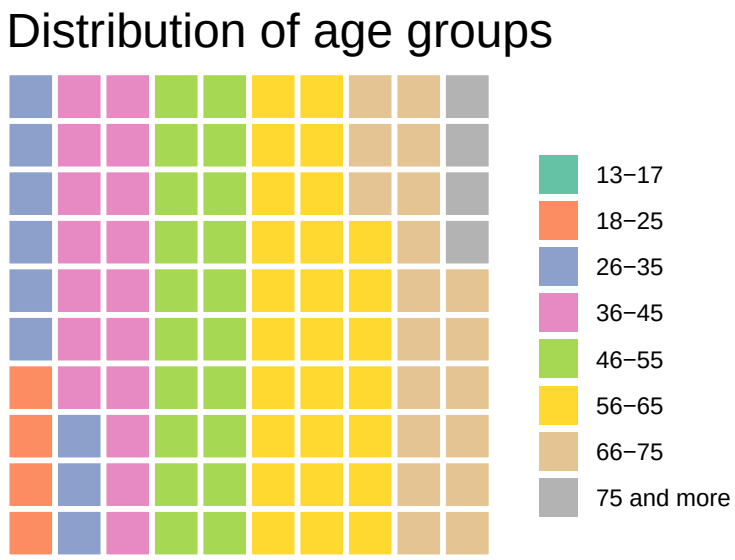
As the graph shows, the number of people who have not been tested for Coronavirus mostly comprised those, who had no symptoms, and this proves the statement in our hypothesis.
If we look at the number of people who had negative test results, most of them had no symptoms,as well.

2. This hypothesis refers to those, who were tested positive and we are interested in the number of people who were without symptoms.Below is the graph showing the proportion of people with symptoms and without symptoms among positive tested ones.



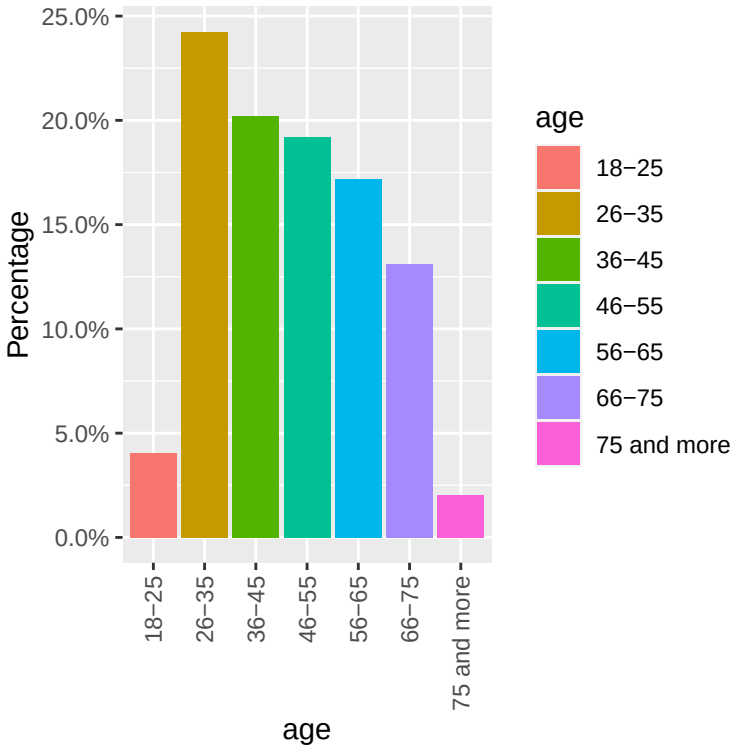
The graph shows that nearly 3/4 of people diognised with COVID-19 were with symptoms and only about 1/4 of them were without symptoms.However, this doesn't actually mean, that those who are asymptomatic may not be infected, since from **1.** we see that most of the people who have no symptoms are not likely to be tested for the virus.

3. Here we'll take into our consideration the age groups and try to figure out, whether those belong to younger age groups use the website more in contrast to the ones who are of older age.



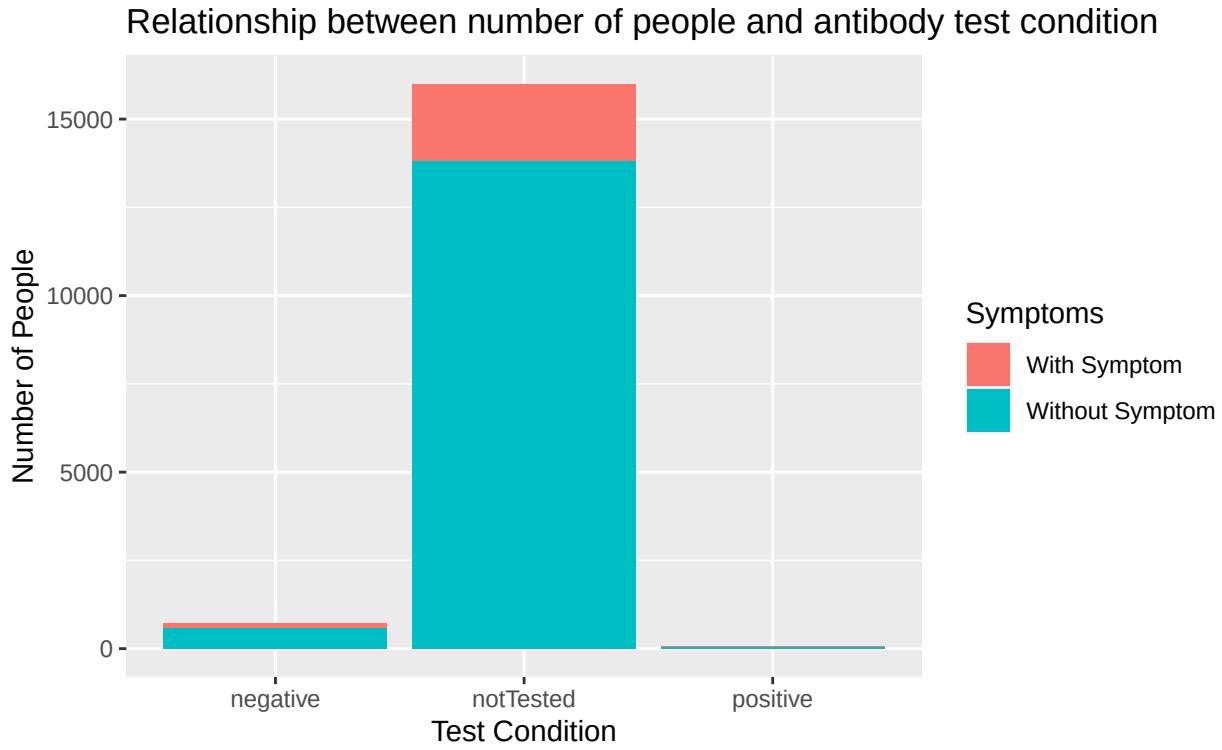
As we can see most of those who took part in the questionnaire are of young age. Nearly 23% of the participants were older than 66. This proves our hypothesis that mostly young people take part in the survey.The reason behind this may be that those who are aged do not use such devices as computers or smartphones, and may even not be aware of such a questionnaire at all.

4. Now we are going to look at what the picture looks like, when we take into our consideration the age groups, that have been tested positive for the virus and our assumption is that most of them are aged (>65 years old).



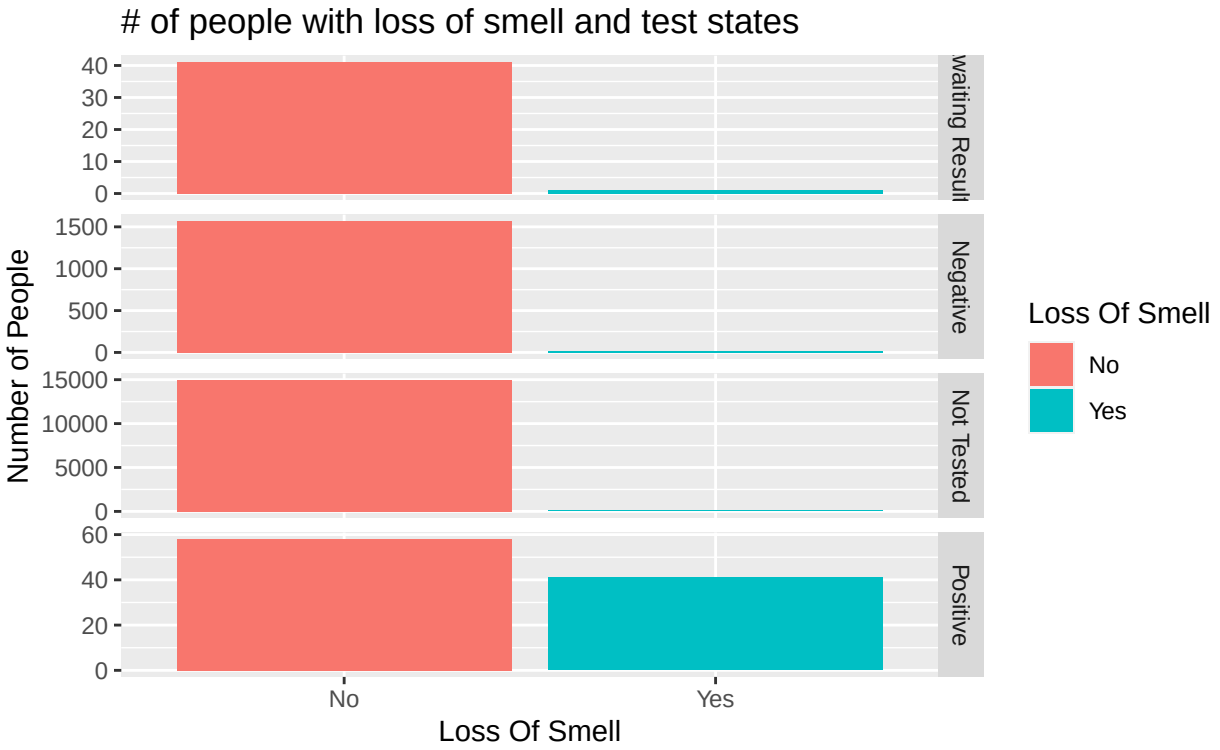
The graph shows that the maximum number of positive tested ones are from the age group 26-36, and the groups, which were in our interest are 66-75 and 75 and more, which have third and first lowest number of positive tests among all the age groups, respectively .It is interesting to note that those, who are aged 75 and more and 18-25 have 2 lowest number of positive tests.This can be interpreted as those who are 18-25 years old may have good immune system and those who are 75 and more do not leave home that much.

5. Here we are going to see what portion of people who have no symptoms take antibody tests. Our initial guess is that people who are asymptomatic are likely to take the antibody test to see whether they had virus or not.



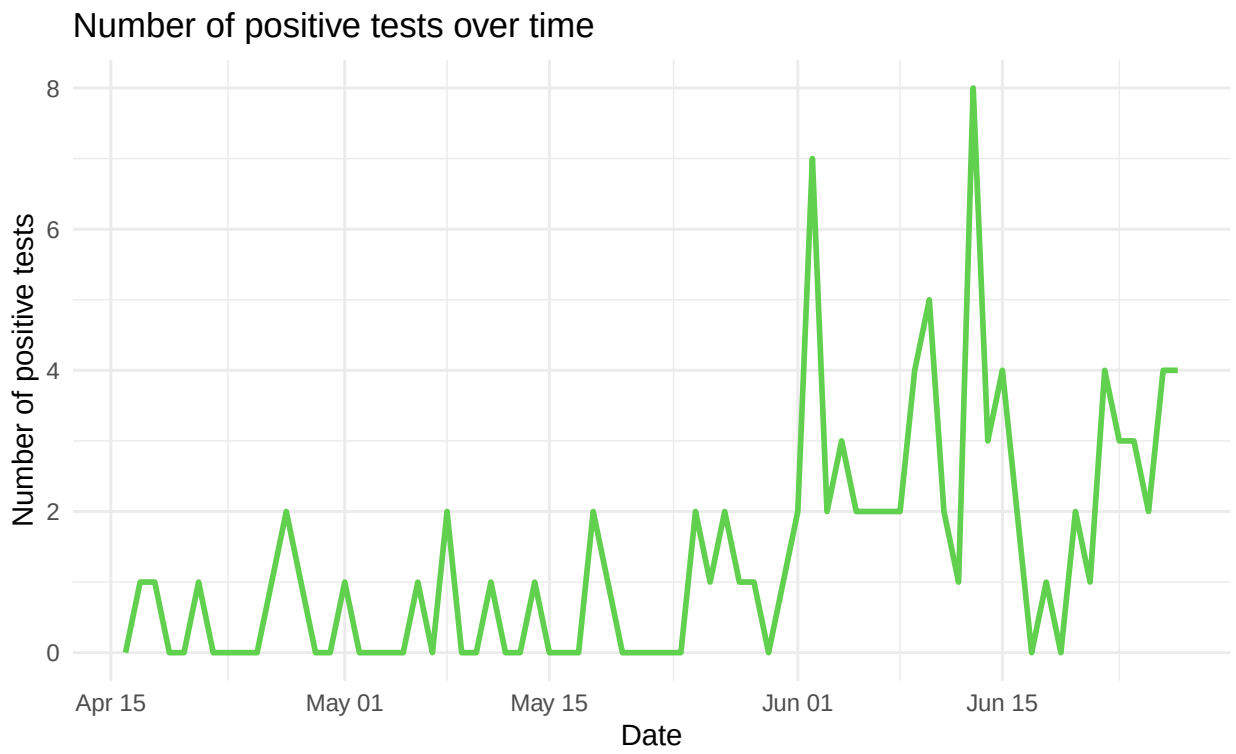
From the graph we see, that most of the people who have no symptoms are not taking any antibody tests and this contradicts our hypothesis. What is interesting in this graph is that those, who have taken the antibody test had mostly negative results than positive and this can be explained as these people may mainly consist the portion who have overcome the virus and are advised by their doctors to take that test.

6. For now we will look at what portion of people who were tested positive and negative have lost smell more and try to understand whether this result proves or disproves our hypothesis.



If we consider the information on the graph, we see that there's a correlation between test result and loss of smell and it's positive as if we look at the number of people with negative results the difference is huge in the number who had loss of smell and who didn't.Yet, the previously mentioned numbers are quite close to each other, when it comes to those, who have been tested positive.

7. For this point we'll take into our consideration the number of positive tests over some amount of time and try to find out, how the number of positive tests changes.



As we cans see from the graph the numbers were relatively low until June. This can be connected with the amount of tests available or low number of people who are tested for the virus. As the virus spreads we see increase in the number of positively tested people in one day, as well.Yet, after June 15 we see an increase but not that much and this may even mean that the situation is gradually getting regularized if we consider small fluctuations after that date.

Summary of findings and recommendations

Summary

To sum up

1. Asymptomatic people are not tested,which is dangerous, since they may be virus carriers and infect people who are in highly risked age groups.
2. The number of tested people is very low.This may be connected with the number of tests available.
3. Symptoms, such as loss of smell may indicate that the person is infected, so it would be better to test such people.
4. Antibody tests are not used that often, which may be again because of the limit in the amount of them.
5. Aged people being considered in the zone of risk do not take the survey that much.

Recommendations

1. It would be better to have data on loss of taste, as well, as this may also be considered as an indicator.
2. In the questionnaire adding a field about chronic illnesses would give opportunity to analyze people who have them, too.
3. It would be better if by some means we could include more aged people who are in the risk zone to better analyze the situation.

References

[1] https://infogears.org/