Individual project

This dataset is called "Infogears". Since, nowadays the world is fighting against coronavirus, I have decided to take it and get new, more interesting staff and compare them with information we get every day. The data is a new one; more and more rows are added daily. And, since the data is new, there is no much statistics done on it.

Hypothesis: Let's understand does the society wear masks, does wearing masks useful and what are the common symptoms of the virus.

Here is my github link: https://github.com/Anahit-Manukyan/Infogears/tree/master

R Markdown

```
## Loading required package: NLP

## Loading required package: RColorBrewer

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##     annotate

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: magrittr
```

Let's look at the structure of the dataset.

```
## [1] "Number of observations: 30088"

## [1] "Number of variables: 33"
```

As we can see, there exist 30088 observations and 33 variables in our data set.
Among which householdHeadcount is numerical, and the symptoms are in an integer type since, as we can see, each person checked he/she has the symptom (1) or not (0).

First of all, let's clean the data and make it more practical.

1. Since the column "ï..age" is hard to find or type, let's just rename it to "age".
   Here are our 33 columns.

2. Also, since the values of age are too long, we can change them

The values in the original dataset

```
## [1] interval_36_45       interval_26_35       interval_46_55
## [4] interval_18_25       interval_66_75       interval_13_17
## [7] interval_75_and_more interval_56_65
## 8 Levels: interval_13_17 interval_18_25 interval_26_35 ... interval_75_and_more
```

Here are the new values of the variable "age"

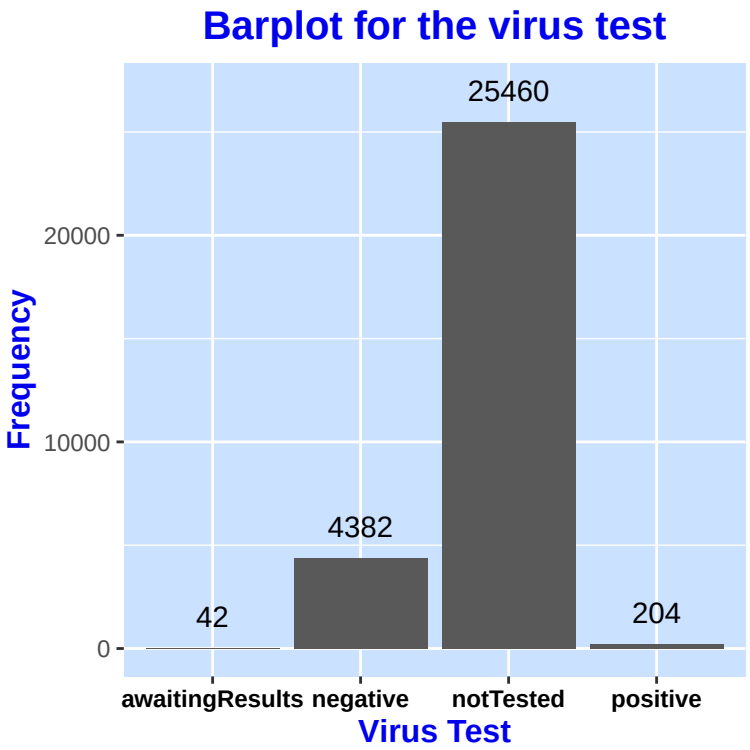
```
## [1] "36-45" "26-36" "46-55" "18-25" "66-75" "13-17" "75+"   "56-65"
```

3. Let's understand which columns we need.

```
##  [1] "age"                 "antibodyTest"        "createdAt"
##  [4] "exposureLevel"       "faceCovering"        "gender"
##  [7] "healthIssues"        "householdHeadcount"  "leftHomeTimes"
## [10] "mentalHealthImpact"  "updatedAt"           "virusTest"
## [13] "zipCode"             "bodyAche"            "diarrhea"
## [16] "difficultyBreathing" "disorientation"      "fatigue"
## [19] "headAche"            "id"                  "irritatedEyes"
## [22] "leftForExercise"     "leftForOther"        "leftForShopping"
## [25] "leftForWork"         "lossOfSmell"         "noSymptoms"
## [28] "persistentCough"     "soreThroat"          "temperature"
```

Since, in my opinion, we do not need the information about user agent, I have deleted the column. As a result, we are left with 30 columns.

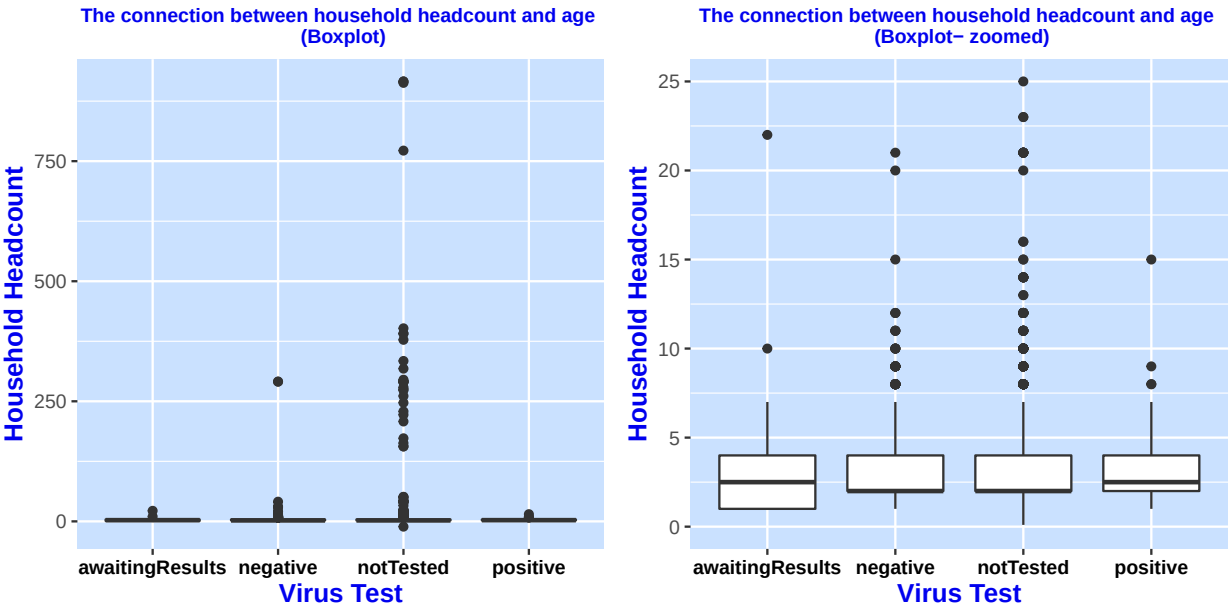
3. Let's make numerical variables factors, in order to do the calculations easily.



As we can see most of the patients are not tested or negatively tested, as a result, we cannot get clear results regarding the infected people.

Lets understand the test results and household headcount.

```
## Warning: Removed 45 rows containing non-finite values (stat_boxplot).
```



From the first graph we can see that all the categories have many outliers however, their number is extremely

dominant for the people who are not tested. Positively tested people have household headcount mostly equal to 3, however, the outliers show that some people have
relationship with more than 7 people. The good point that we can get from the graph is that the people who are waiting for their
test results have shortened their household headcount (or they are just they are shown now).

Let's have a closer look on the virus test results.

```
## [1] "Positive: 204"
```

```
## [1] "Awaiting results: 42"
```

```
## [1] "Negative: 4382"
```

As we can see, majority of the people in this dataset are not tested.
However, from 30088 observations only 204 have the virus.

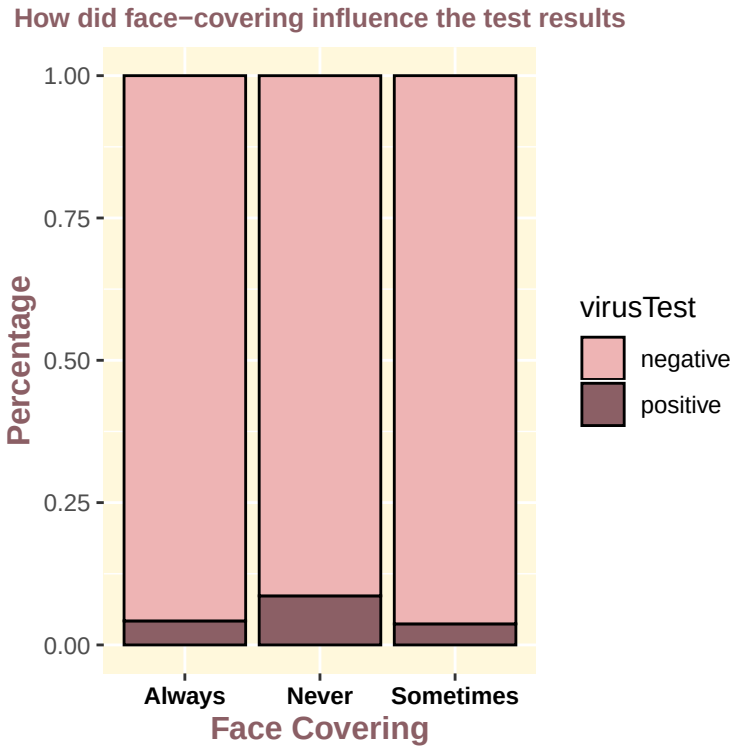
Let's understand, does the society wear masks or not, and how it influences the spreading of the virus.

```
## [1] "Never wore a mask: 693"
```

```
## [1] "Always wore a mask: 16625"
```

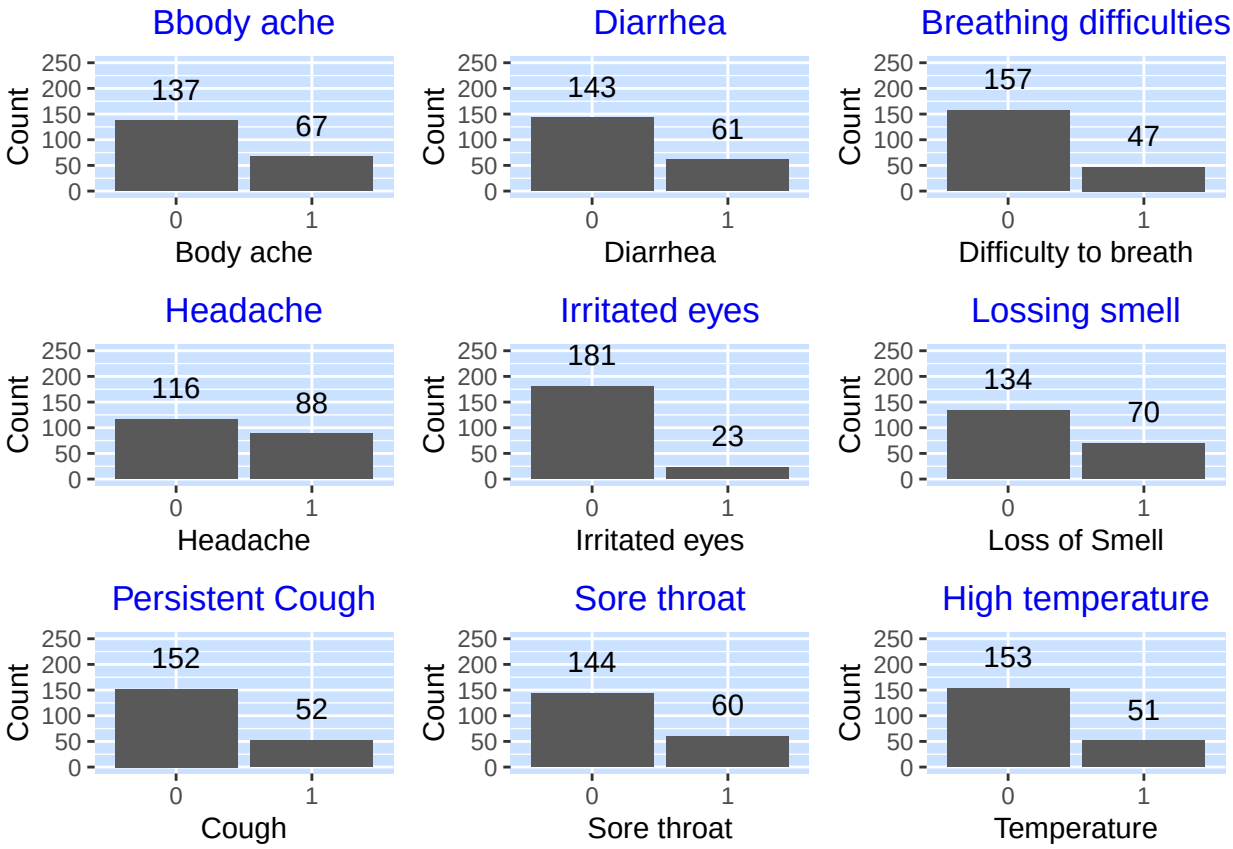
```
## [1] "Sometime wore a mask: 4439"
```

As we can see, the majority of people wore masks.
Let's understand how much wearing masks is dependent to the test being positive.



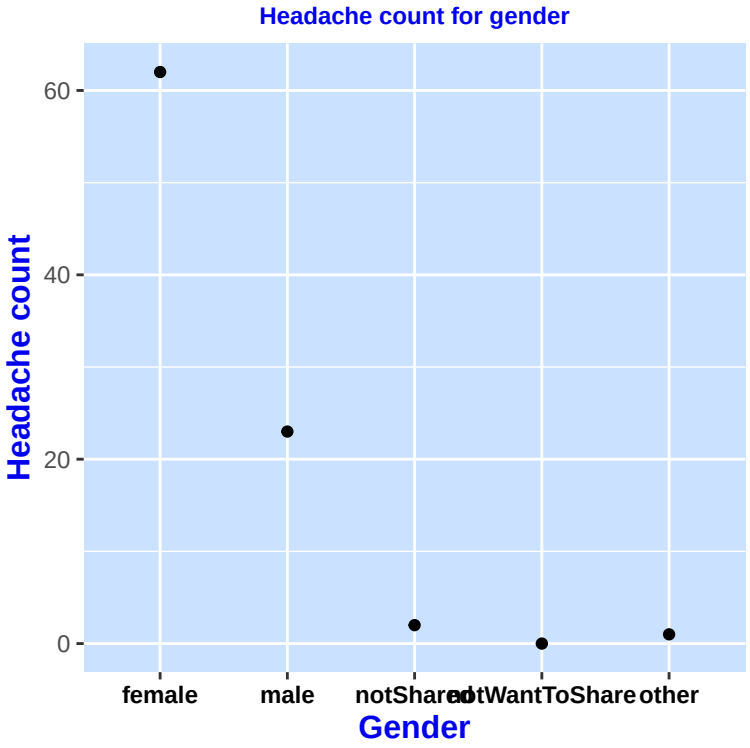
As we can see, the people who never wore a mask had more changes of getting the virus,than the people who always or sometimes wore the masks. However, the results of wearing sometimes or always nearly do not vary.

Here are the major symptoms of the people with positive test results



As we can see, the majority of people do not experience these symptoms. Most commonly, people experience headache and body ache.
Having irritated eyes is not a common symptom.
Although, some symptoms such as cough, breathing difficulties or high temperature are told to be common, the statistics does not show that.

Since headache is the most common symptom, let's understand does it vary from gender to gender



Here are the people who have the virus and experience headache. As we can see, women are more likely to have headaches.

As the thesis statement suggests, we are going to find the people with the virus who shared information about their communication.

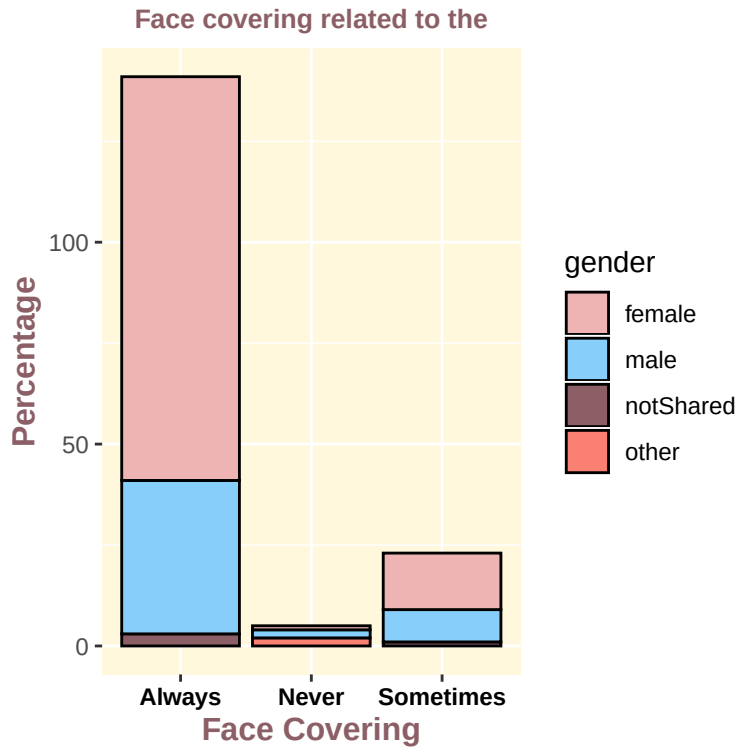
We are going to take a close look at a random zip code and find out the problems at that neighborhood

```
## [1] "As a result, there exists 5903 zip codes in our dataframe, among which in 167 unique zip codes
```

By checking the length of unique values of Zip Code, I wanted to understand if there exist people from the same neighborhood of from the same family. Since, from 30088 observations only 5903 are unique, we can understand that there exist many people having some contact to each other.

```
## [1] "52 people, who have the virus had direct contact with others."
```

Let's understand, did they wore the masks or not.



As we can see from the graph the majority of the people who had the virus wore masks. However, if we look at Never-Face-Covering people, we can notice Men are the majority among not wearing the masks. Additionally, the graphs suggest that Women tend to cover the face.

Lets try to find a neighborhood, where the virus is more spread.

```
## # A tibble: 1 x 2
##   zipCode  Freq
##   <fct>   <int>
## 1 10111       3
```

The results show that there are not many people from the same neightborhood who have the virus. Now let's find out how could possibly the virus spread, although frequency=2 is not a big number.

```
## Error: Unknown column `guid`
```

As we can see from the results, the woman, who wears face covering sometimes could have infected the person who always wears it. Additionally, both if them had direct contact. We can say that, they are not from the same family, since their ip_addresses are
different. While they had no health issues, both of the people experience body ache, diarrhea, difficulty of breathing, disorientation, fatigue, headache loss of smell, have persistent cough and high temperature. Finally, they are both in the risking group, since their age interval is from 66 to 75.

```
## Error in table(dat$Marital_status, dat$approval_status): object 'dat' not found
```

```
## Error in eval(expr, envir, enclos): object 'two_way' not found
```

```
## Error in prop.table(two_way): object 'two_way' not found
```

```
## Error in sweep(x, margin, margin.table(x, margin), "/", check.margin = FALSE): object 'two_way' not
```

```
## Error in sweep(x, margin, margin.table(x, margin), "/", check.margin = FALSE): object 'two_way' not
```

Creating a frequency table is a simple but effective way of finding distribution between the two categorical variables. The table() function can be used to create the two way table between two variables.

```
##
##                          Always Never Sometimes
##   noImpact             0   1965   183       611
##   significantImpact    0   4746   117       879
##   someImpact           0   9914   393      2949
```

Cell percentages

```
##
##                                     Always       Never   Sometimes
##   noImpact          0.000000000 0.090315760 0.008411086 0.028082916
##   significantImpact 0.000000000 0.218136692 0.005377580 0.040400791
##   someImpact        0.000000000 0.455669440 0.018063152 0.135542584
```

Row percentages

```
##
##                                   Always      Never  Sometimes
##   noImpact          0.00000000 0.71221457 0.06632838 0.22145705
##   significantImpact 0.00000000 0.82654127 0.02037618 0.15308255
##   someImpact        0.00000000 0.74788775 0.02964695 0.22246530
```

Column percentages

```
##
##                         Always     Never Sometimes
##   noImpact           0.1181955 0.2640693 0.1376436
##   significantImpact  0.2854737 0.1688312 0.1980176
##   someImpact         0.5963308 0.5670996 0.6643388
```

As we can get from the graph, wearing masks and mental health impact have some connection or correlation. Majority of the people have some mental impact.

As we see masks are actually useful. The only thing that I would suggest is that people, especially the ones who are tested positively, sould wear masks to protect lives.