Question3：

Code explanation:

First I install the package “readxl”, then recall it using library(readxl)

Then I want to read the excel file and assign it to the new vector named “vaccine”: vaccine<-read\_xlsx("C:/Users/USER/Downloads/COVID-19 County Vaccine Data.xlsx").

After you run this code, at the environment area of R studio, you can see this data named”vaccine”, 69 observations of 6 variables:

Table

Description automatically generated

If you click the “vaccine”button, you can see the table has already imported into R studio:

Table

Description automatically generated

Then I want to see how many N/A values in this dataset, I use code: table(is.na(vaccine)) to see how many N/A values in vaccine dataset. After you run this code, you can see the result:



False means no N/A values, and 414 False means there are 414 values in the dataset. Since no true, we don’t have missing values.

Since there is a -1 in the “poplation>16”group, I think it is not right, cause our data should greater than 16, so I name the numbers less than 16 as N/A, using code: vaccine$`Population > 16`[which(vaccine$`Population > 16`<=16)]=NA;

And then I want to ask R studio, is that value N/A? so I use code: is.na(vaccine), then I found that value who less than 16 is N/A.

Chart

Description automatically generated

Now I want to delete/omit that value, cause it is a N/A, I use code: Vaccine<-na.omit(vaccine). Note that now it is capital V in vaccine.

Then I want to calculate completed rate, it equals to people completely vaccinated divided by population>16, I use code: Vaccine$completed\_rate<-as.numeric(Vaccine$`People Completely Vaccinated`)/as.numeric(Vaccine$`Population > 16`)

Then I plot the completed rate line chart, I assign county as independent value and completed rate as dependent value, the code is: plot(Vaccine$completed\_rate,type="o",xlab = "Counties",ylab = "completed\_rate"). The line chart is bottom:

Chart

Description automatically generated

Our counties are in the range of 1-69, each number represents one county, and our completed rate are in the range of 0.1544 to 0.3505.

Then I want to sort the completed rate by the decreasing order, I use code: Vaccine<-Vaccine[order(Vaccine$completed\_rate, decreasing= T),] Now, when you click the right side, environment group, second Vaccine button (Capital V), you will get the table automatically, and also get the completed rate in the decreasing order, the table is at bottom:

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generatedAs you can see in the rightmost column of this table, the completed rate is ordered by decreasing order.

Then I want to do the summary statistics, summary(Vaccine), the result is as follows:

A picture containing text, receipt, screenshot

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Population>16 | People receiving at least one dose | People Completely Vaccinted | Completled rate |
| min | 8111 | 19.24 | 2425 | 0.1544 |
| 1st Qu | 17852 | 29.02 | 4906 | 0.2426 |
| Median | 34944 | 32.42 | 8404 | 0.2622 |
| Mean | 144211 | 32.69 | 42207 | 0.2681 |
| 3rd Qu | 85204 | 26839 | 23161 | 0.2941 |
| Max | 4903185 | 46.15 | 1461323 | 0.3506 |

Then I want to see the bar chart of completed rate by county, code:

barplot(Vaccine$completed\_rate, names.arg =Vaccine$County , beside = FALSE,

horiz = FALSE, density = NULL, angle = 45, col = NULL,

border = par("fg"), xlab ="COUNTY", ylab = "Completed rate")

result is:

Chart

Description automatically generated

Due to the fact that we have 68 counties, cannot fit the bar chart, so it only gives us several names at equal intervals.

Madison county has the highest completed rate, which is 35.05%,

Russell county has the lowest completed rate, which is 15.44%.

Mean completed rate is 26.81%, Pike county is little under the mean, has 24.54%.

Clay county is at the median level of completed rate, which is 26.22%.