Question1：

Code explanation:

I first install “readxl” and “xlsx” packages, and then use library to call them. library(readxl); library(xlsx)

And then I read the death\_cases dataset, (note that here it is lowercase d), and I read the excel file saved in my computer, location of the excel file is: C:/Users/USER/Downloads/COVID-19 in Alabama.xlsx.

Code: death\_cases<-read\_xlsx("C:/Users/USER/Downloads/COVID-19 in Alabama.xlsx"). now in the enviroment area (top right of the R studio), you can see the death\_cases appears at here. If you click the drop down arrow, you can see what variables in this dataset named “death\_cases”:

Graphical user interface, text, application

Description automatically generated

Now I want to see the data type of each variable: I use code: str(death\_cases). Run it I get:

Text

Description automatically generated with medium confidence

and then I use table(is.na(death\_cases)) to find the missing value, result is 271 False and 1 True, meaning we have 271 good values and only one missing values,

A picture containing text

Description automatically generated

so I use is.na(death\_cases)to see who is that N/A value. If you run this code, you can see row 68th is N/A, because county in the code counts from 1 and excel counts from 2, so at the end, the 69th line of excel is blank, and the 68th line of the code is a missing value.

Background pattern

Description automatically generated with low confidence

And then I want to omit the only one missing value, and name the new dataset as “Death\_cases”(capital D). Death\_case<-na.omit(death\_cases). You can see in the environment area, the Death\_cases appears there.

A screenshot of a computer

Description automatically generated with low confidence

from now on, the dataset name changes from “death\_cases” to “Death\_case”

Since I want to calculate Death\_rate, I need to convert the datatype,

Death\_case$Cases<-as.numeric(Death\_case$Cases)

Death\_case$`Total Tested By County`<-as.numeric(Death\_case$`Total Tested By County`)

Death\_case$Deaths<-as.numeric(Death\_case$Deaths)

str(Death\_case)

Run them, I get the result:

Text

Description automatically generated with medium confidence

As you can see, the data type has been changed.

And then I use Death\_case$Death\_rate<-Death\_case$Deaths/Death\_case$Cases to calculate Death\_rate, and put it to the new vector named: Death\_case$Death\_rate. Note that the Death\_rate is in the dataset named”Death\_case”. Now the Death\_case increase a variable from 4 variables to 5 variables.

Text, application

Description automatically generated

We have one variable added, which is the last one: $Death\_rate.

Now you can see the Death\_rate table, you click the right side of R studio, which is Environment section, you click the “Death\_case”, (capital D). you will get the table as follows:

Table

Description automatically generated

But, As of here, our code is not finished, it just shows us the result of death\_rate, I need to sort them by decreasing order, but don’t hurry, I want to plot the line chart for the death\_rate first. So I use plot(Death\_case$Death\_rate,type="o",xlab = "Counties",ylab = "Death\_rate") to plot the line chart, in this code, $ means the Death\_rate is one variable in the dataset Death\_case, and “o”means I want the line chart, and xlab is the name of the independent variable, ylab is the name of the dependent variable. And then you can run it and get the line chart as follows:

Chart

Description automatically generated

You may wonder what is the 0 - 68 means, each of them means one county, for example, 2 means Autauga, 3 means Baldwin, 4 means Barbour, etc. Furthermore, from the graph you can see, the death\_rate is from 0.010 to 0.4380, each county match their own death\_rate.

Ok, now I want to sort the death\_rate by decreasing order. The code is : Death\_case<-Death\_case[order(Death\_case$Death\_rate, decreasing= T),] Finally, run this code, click the Death\_case(remember it is a name with Capital D) in the environment (top right side of R studio) area. You will see the table as follows, the death\_rate has been sorted by decreasing order :

Table

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidenceTable

Description automatically generated

Table

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidence

Then I want to do the summary statistics: summary(Death\_case), the result is:

A picture containing text, receipt, screenshot

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cases | Total Tested by County | Deaths | Death\_rate |
| min | 621 | 2975 | 25 | 0.009015 |
| 1st Qu | 1874 | 9323 | 52.5 | 0.020106 |
| Median | 3728 | 16848 | 100.0 | 0.024481 |
| Mean | 8189 | 41835 | 168.7 | 0.025075 |
| 3rd Qu | 9778 | 43156 | 195.5 | 0.029479 |
| Max | 81003 | 457504 | 1566.0 | 0.043804 |

As you can see, the Russell county has the lowest death\_rate, 0.9%

Dallas county has the highest death\_rate, 4.3%.

Pike county has 2.4% death\_rate, under the mean death\_rate of all alabama state.

There are 36 counties are under the mean, which is 2.5% death\_rate.