Exercise 1：

Check your current directory：

dir()

Set your directory to Where you download the csv file

setwd("~/desktop/….")

Now, go find a good data set from data.gov and download it to your working directory. Let's say we got something like this:

```
id,name,salary,start_date,dept
1,Rick,623.3,2012-01-01,IT
2,Dan,515.2,2013-09-23,Operations
3,Michelle,611,2014-11-15,IT
4,Ryan,729,2014-05-11,HR
5,Gary,843.25,2015-03-27,Finance
6,Nina,578,2013-05-21,IT
7,Simon,632.8,2013-07-30,Operations
8,Guru,722.5,2014-06-17,Finance
```

Let's make a csv using the data provided above.

Read the data created and name it data

data <- read.csv("data.csv")

print(data)

Let's see if it is actually a dataframe

print(is.data.frame(data))

We do the normal stuff regarding knowing the columns and the rows

ncol(data)

nrow(data)

We wanna extract the maximum salary from the csv file and name it max

max(data$salary)

Let's get the row of the of that person who got the most salary

subset(data, data$salary == max(data$salary))

We want to extract data for the IT department

subset(data, data$dept == "IT")

Get the persons whose id is bigger than 3 and salary is more than 500
subset(data, data$salary >500 & data$id > 3)

Get people who joined after 2019
subset(data.as.Date(start_date) > as.Date (“2019-01-01”))

Let’s take out all NA in the data set
na.omit(data)

Write a new csv file that contains only the persons whose id is bigger than 3 and print new data
write.csv(subset(data$id >3), ”new.csv”)

read.csv(“new.csv”)

What if I do not want the column with the row names?

write.csv(data,”output.csv”, row.names = FALSE)

Exercise two

Now go to the built in data set and find an interesting dataset make that dataset a csv file

Find the standard deviation of all of the columns in that data set and export that information in a csv file

Let’s look at the internal structure of the csv files that you saved with str()