SESSION 7: Basic Statistics

Assignment 1

Table of Contents

1. Introduction

This assignment will help you understand the concepts learnt in the session.

2. Objective

This assignment will test your skills on basic statistics.

3. Prerequisites

Not applicable.

4. Associated Data Files

Not applicable.

5. Problem Statement

1. Histogram for all variables in a dataset **mtcars.**

   Write a program to create histograms for all columns

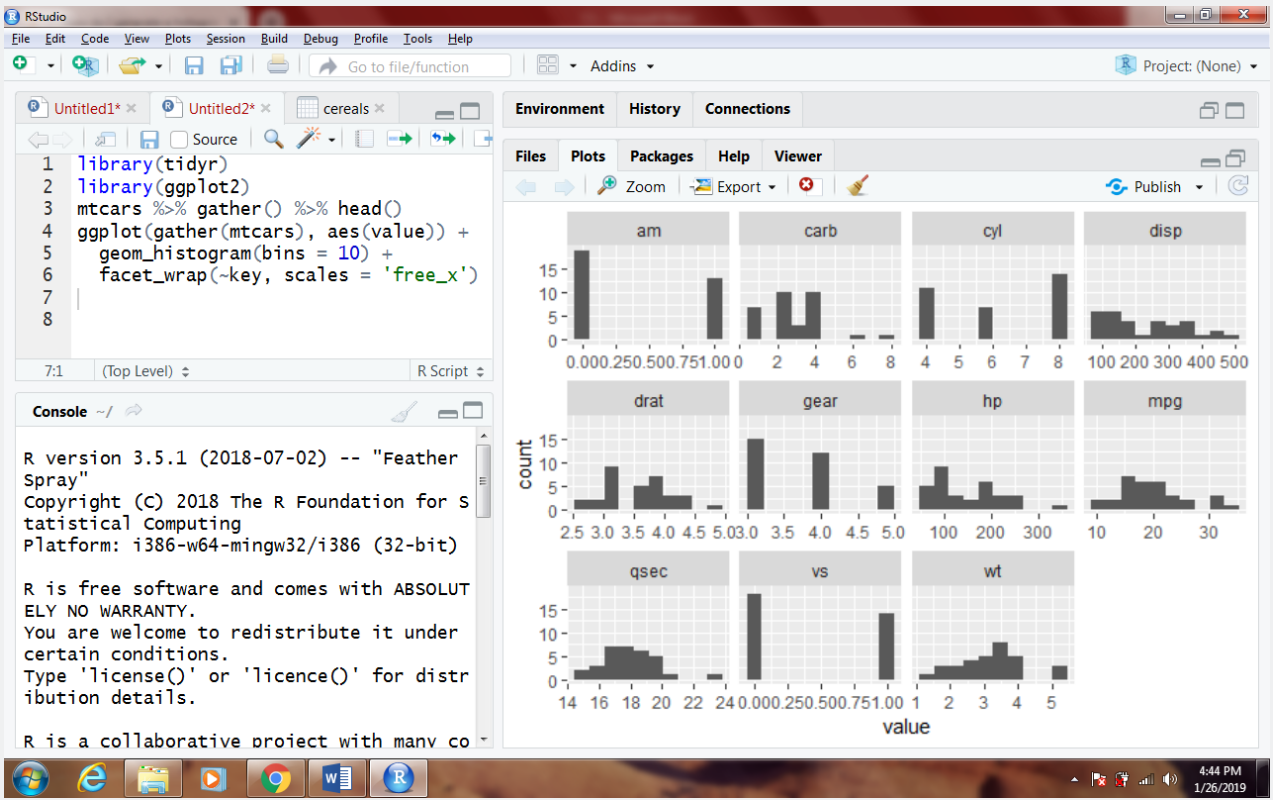
```
Ans : - library(tidyr)
        library(ggplot2)
```

```
mtcars %>% gather() %>% head()
```

```
#>   key value
#> 1 mpg  21.0
#> 2 mpg  21.0
#> 3 mpg  22.8
#> 4 mpg  21.4
#> 5 mpg  18.7
#> 6 mpg  18.1
```

Using this as our data, we can map `value` as our x variable, and use `facet_wrap` to separate by the `key` column:

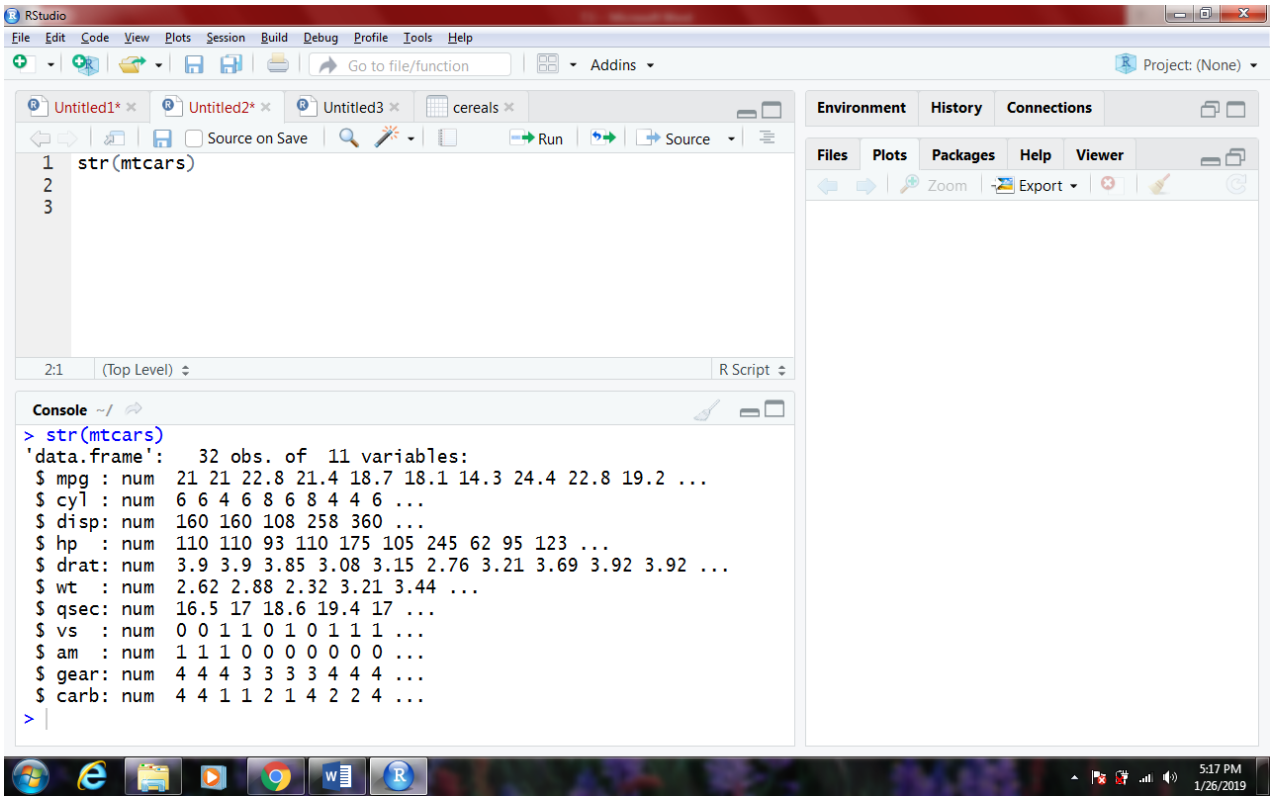
```
ggplot(gather(mtcars), aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x')
```



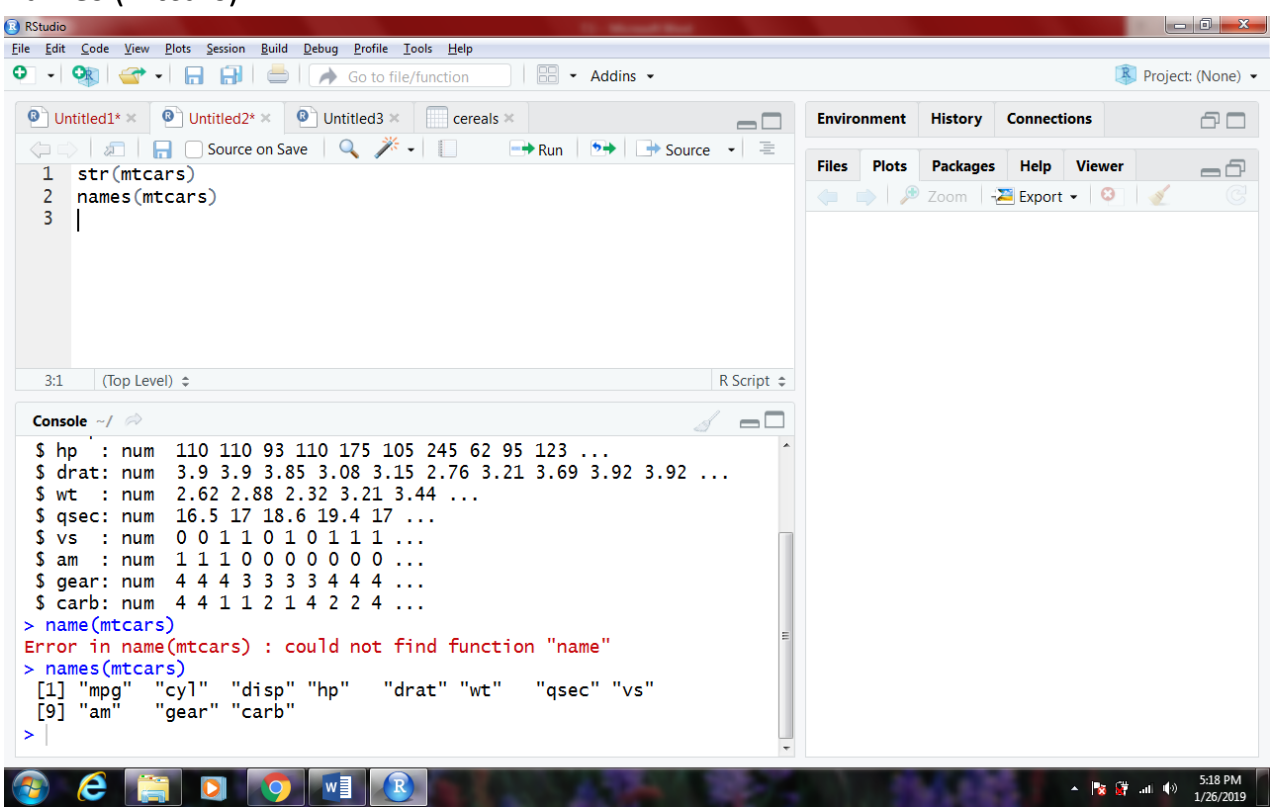
2. Check the probability distribution of all variables in **mtcars**.

Ans- First we look at the structure of the data set

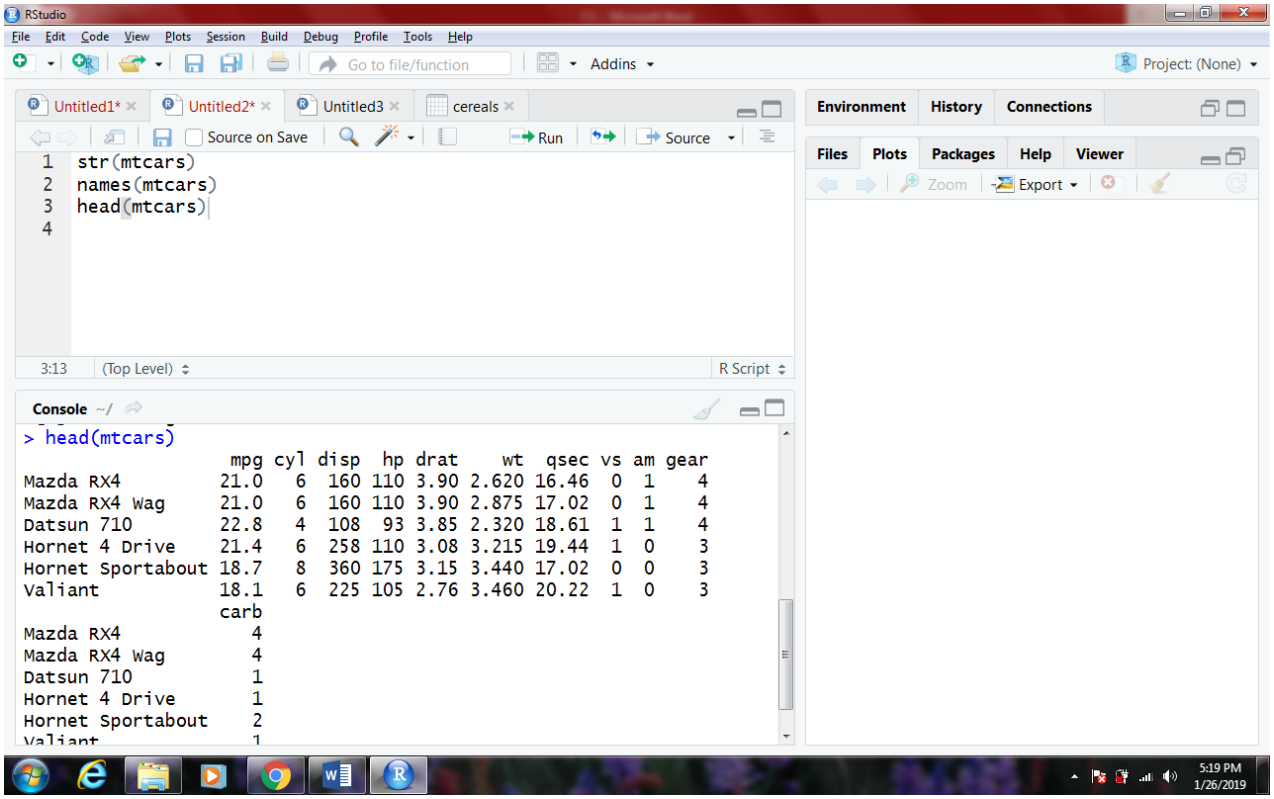
Str(mtcars)



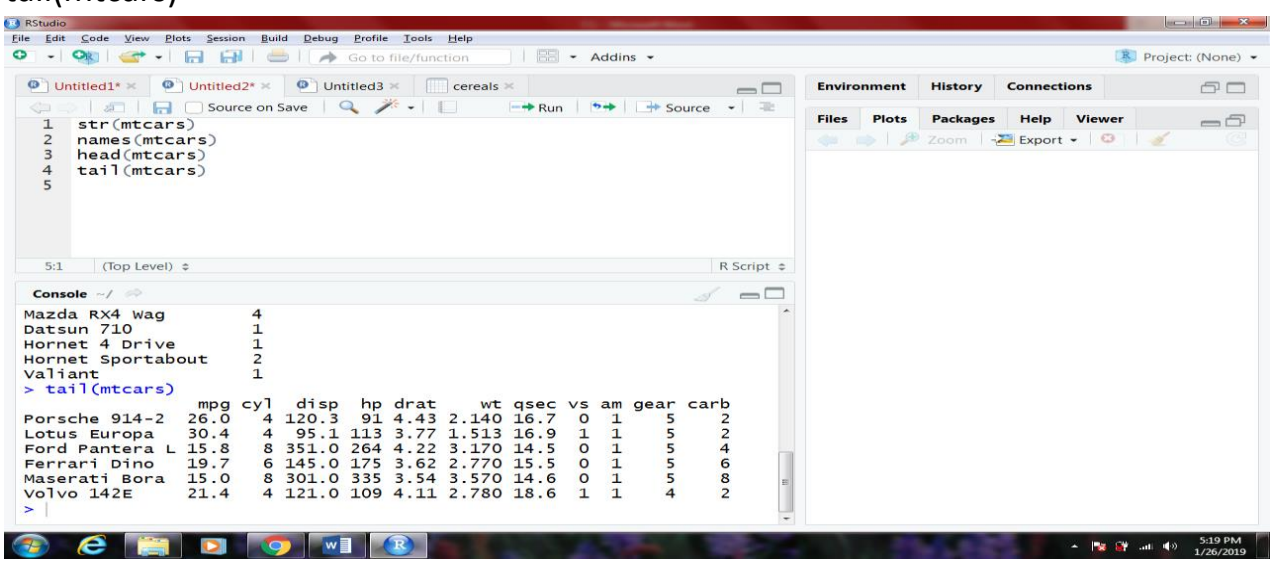
names (mtcars)



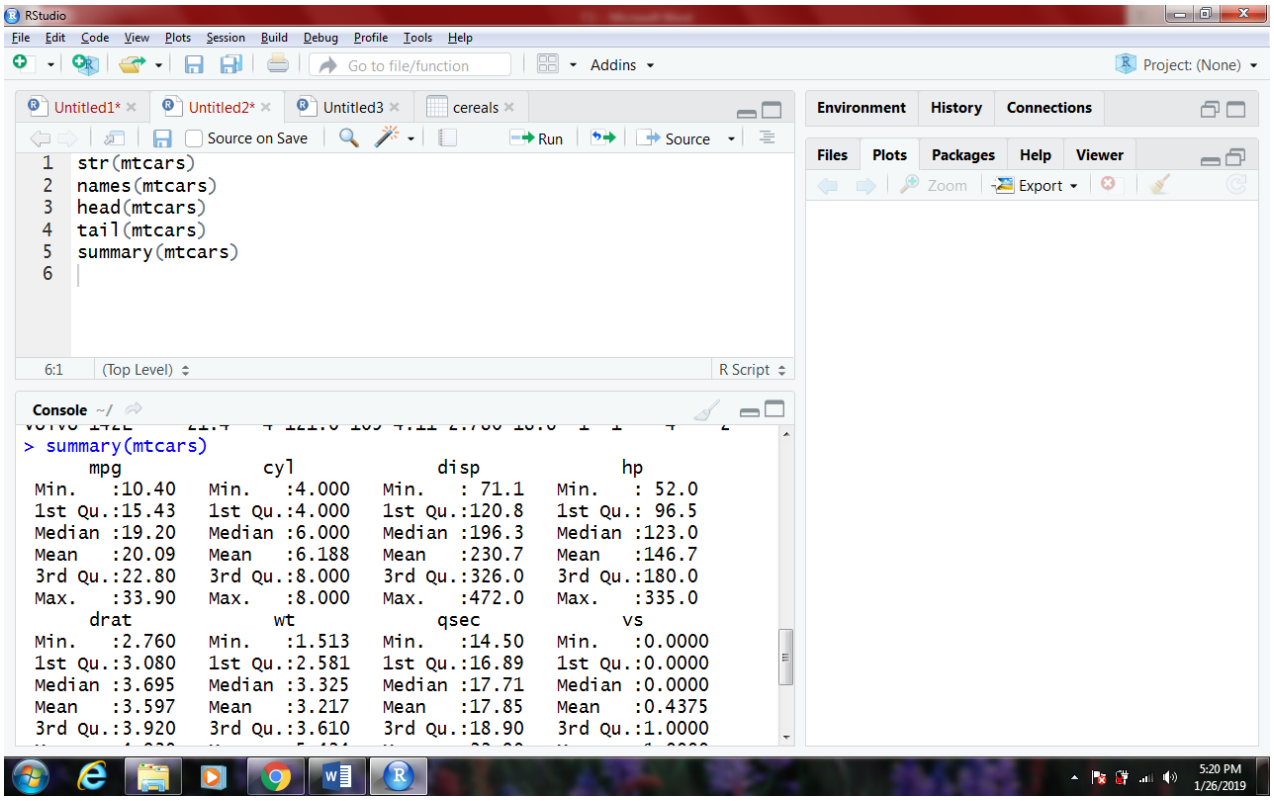
head(mtcars)



tail(mtcars)



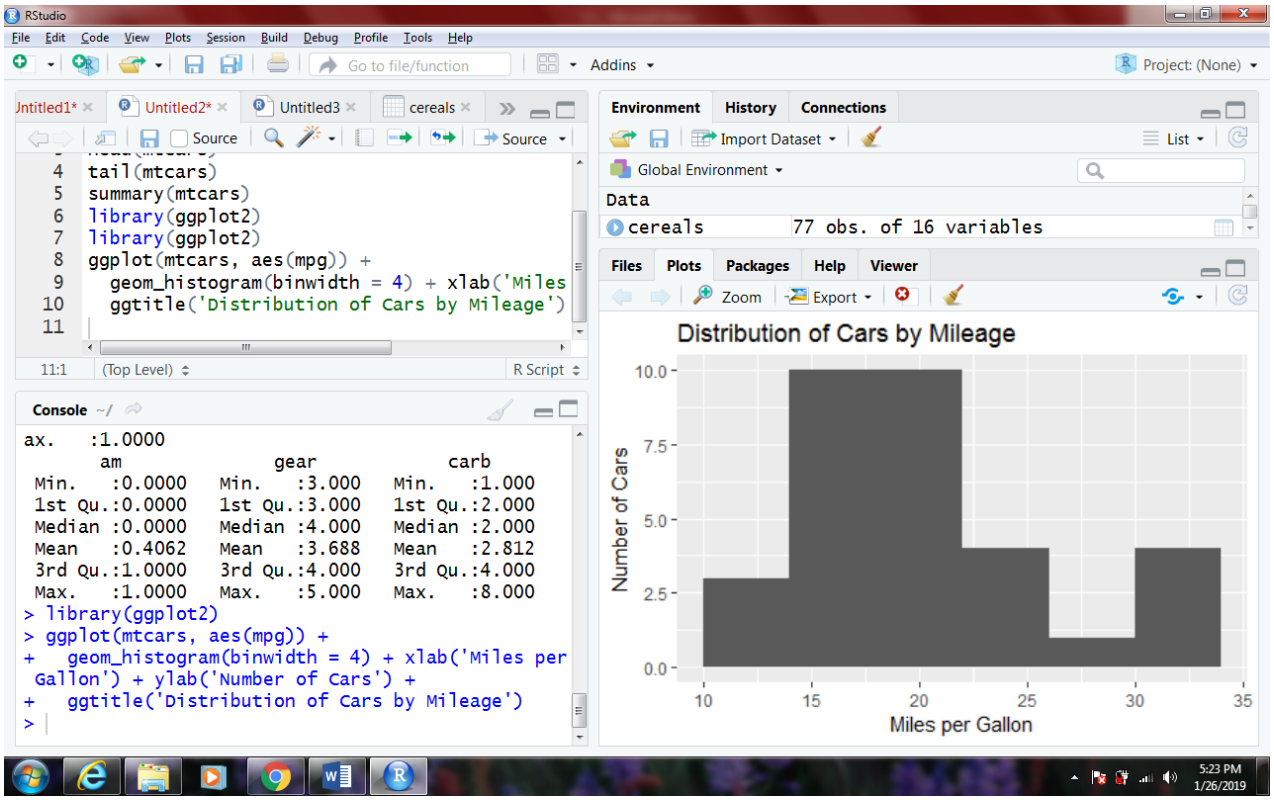
summary(mtcars)



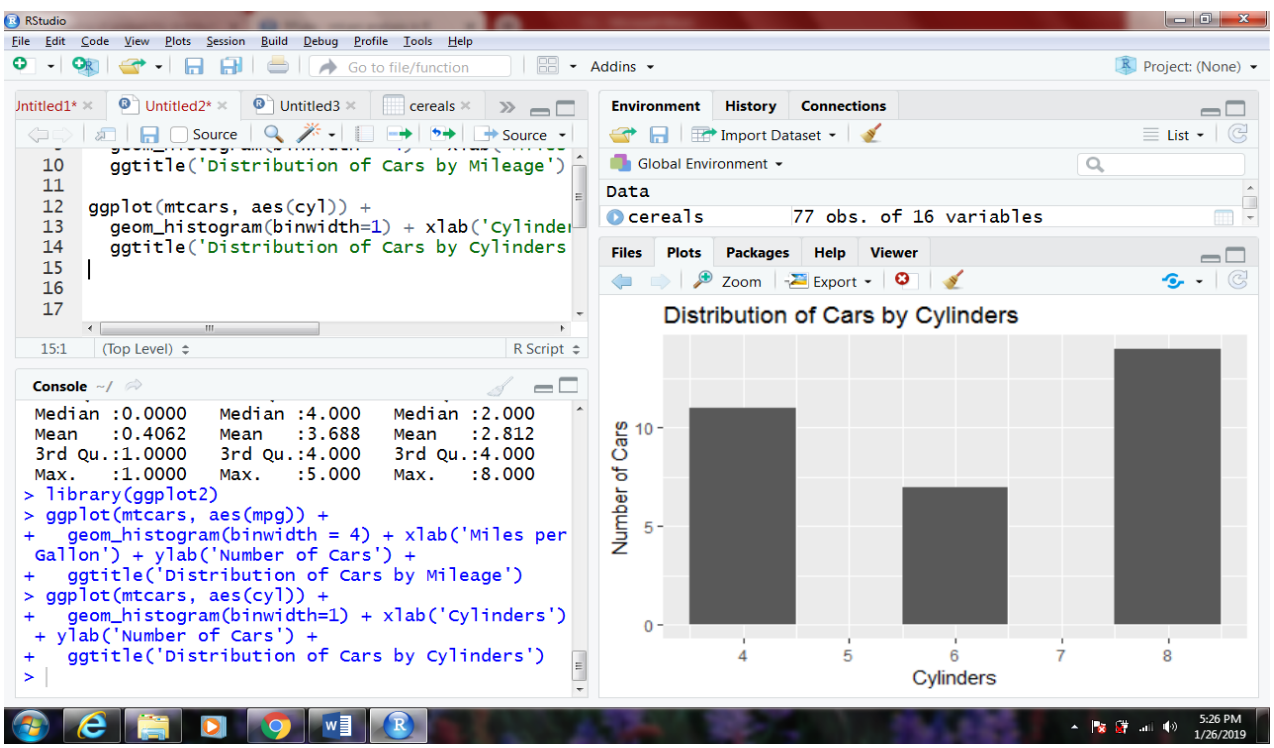
To get a feel for the distribution of some of the data to be analyzed, we plot some histograms, the first against mpg, the second against the number of cylinders, and the third, hp.

library(ggplot2)

ggplot(mtcars, aes(mpg)) +
 geom_histogram(binwidth = 4) + xlab('Miles per Gallon') + ylab('Number of Cars') +
 ggtitle('Distribution of Cars by Mileage')

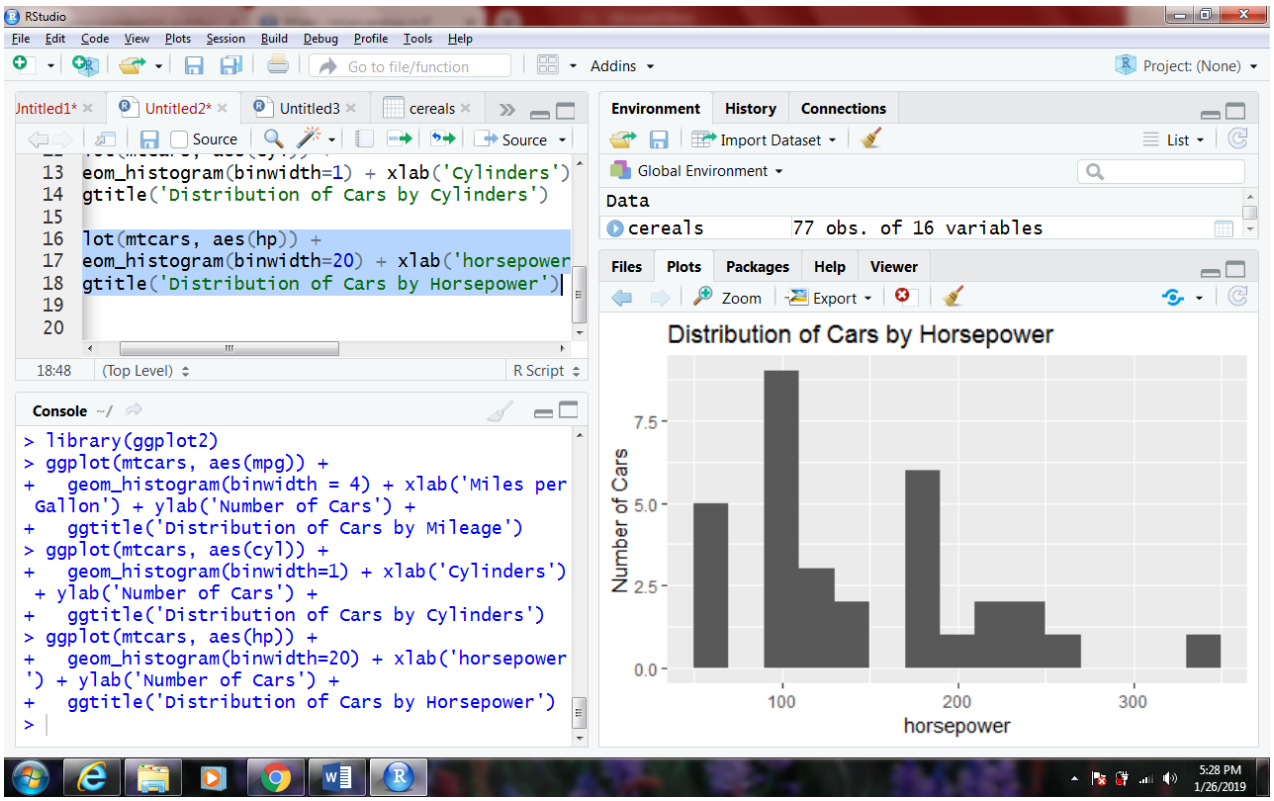


Now we show the histogram for number of cylinders:



Finally, we show the histogram for horsepower:

```
ggplot(mtcars, aes(hp)) +
  geom_histogram(binwidth=20) + xlab('horsepower') + ylab('Number of Cars') +
  ggtitle('Distribution of Cars by Horsepower')
```



We see a good distribution of data across both a wide range of mpg as well as across the different quantity of cylinders, 4, 6, 8, and across a range of horsepower.

Now we look at correlation of hp and mpg.

```
cor(mtcars$mpg, mtcars$hp)
```

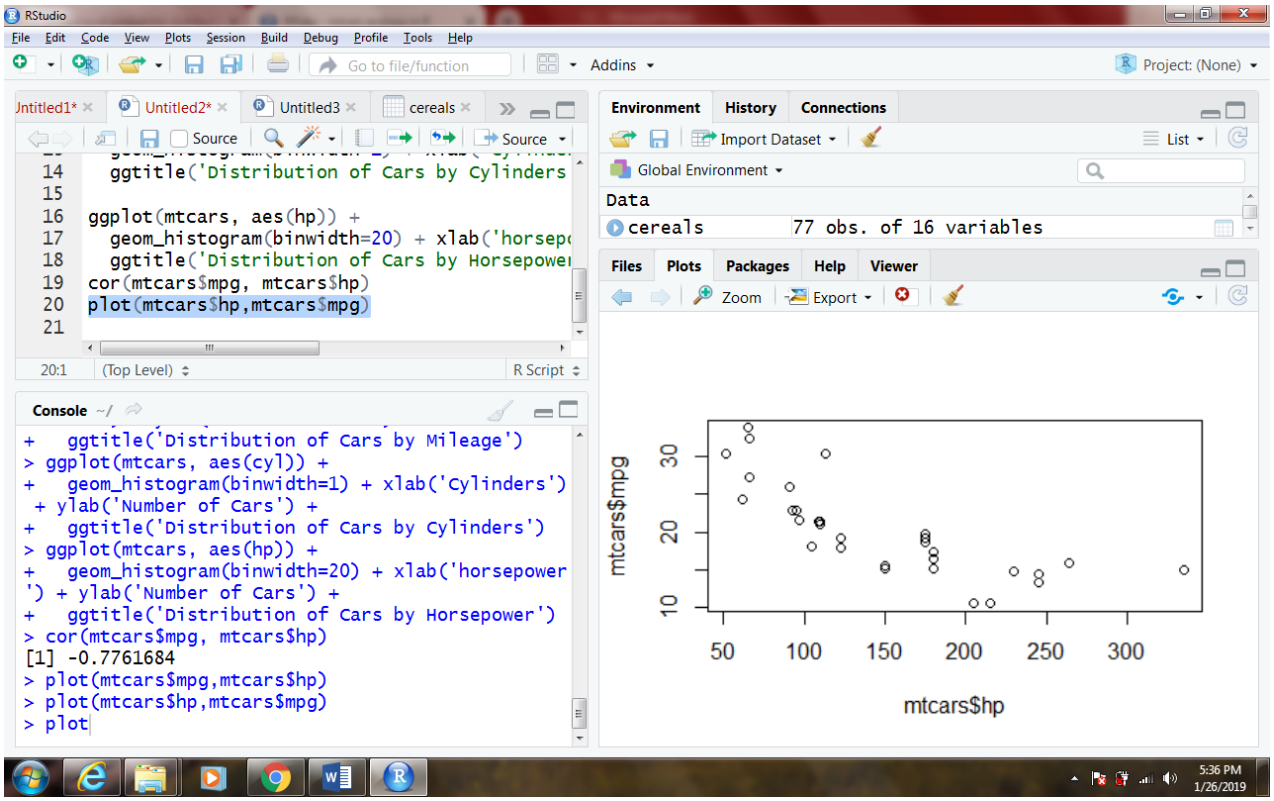
```
1] -0.7761684
```

We find a fairly strong negative correlation.

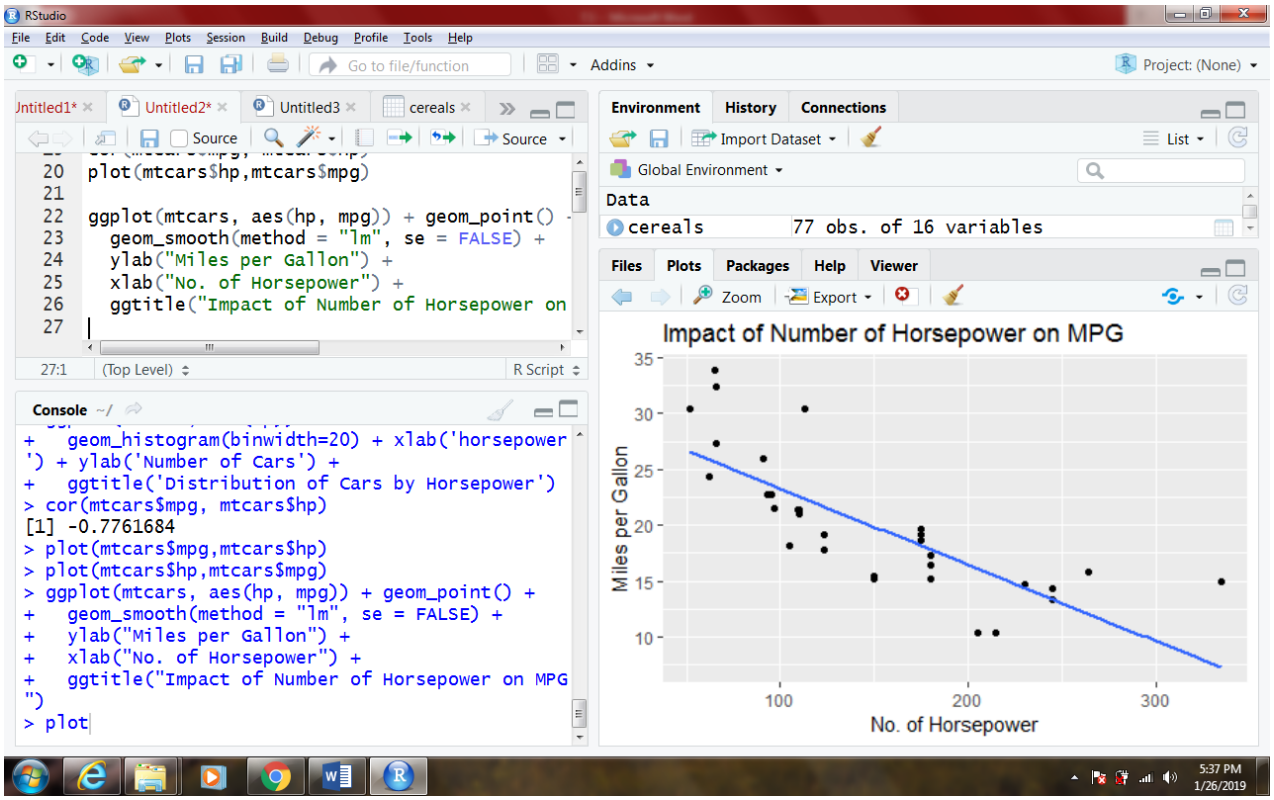
Now we plot the data – HP vs MPG

Below is the effect that number of horsepower has on mpg. We have also shown transmission type (manual = 1, auto = 0) as a point of reference, but it is not a primary part of our analysis.

plot(mtcars$hp,mtcars$mpg)

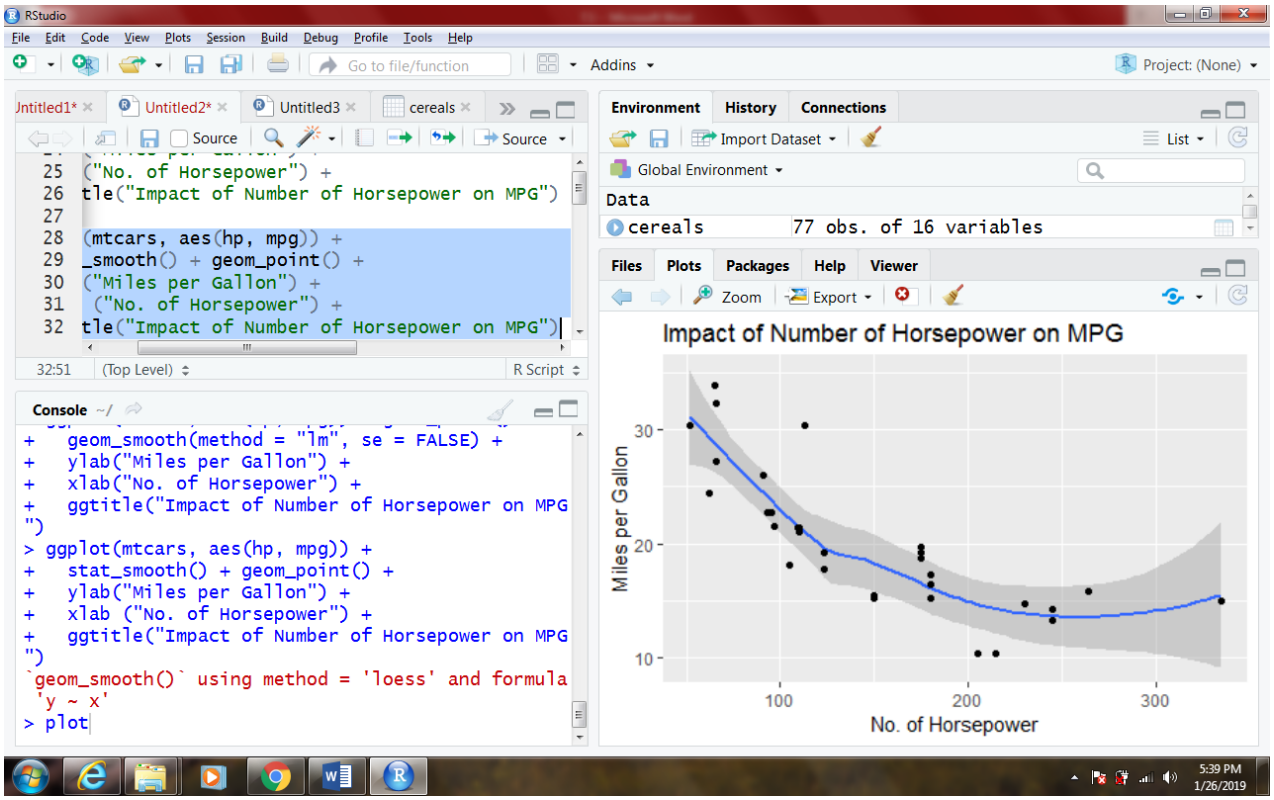


HPv/s MPG



Since the mpg is unlikely to hit zero as the hp increases, we would expect a more asymptotic line. So let's apply stat_smooth to get a better fit.

```
ggplot(mtcars, aes(hp, mpg)) +
  stat_smooth() + geom_point() +
  ylab("Miles per Gallon") +
  xlab ("No. of Horsepower") +
  ggtitle("Impact of Number of Horsepower on MPG")
```



Effect of number of cylinders on mpg

The correlation of mpg and cyl is shown below.

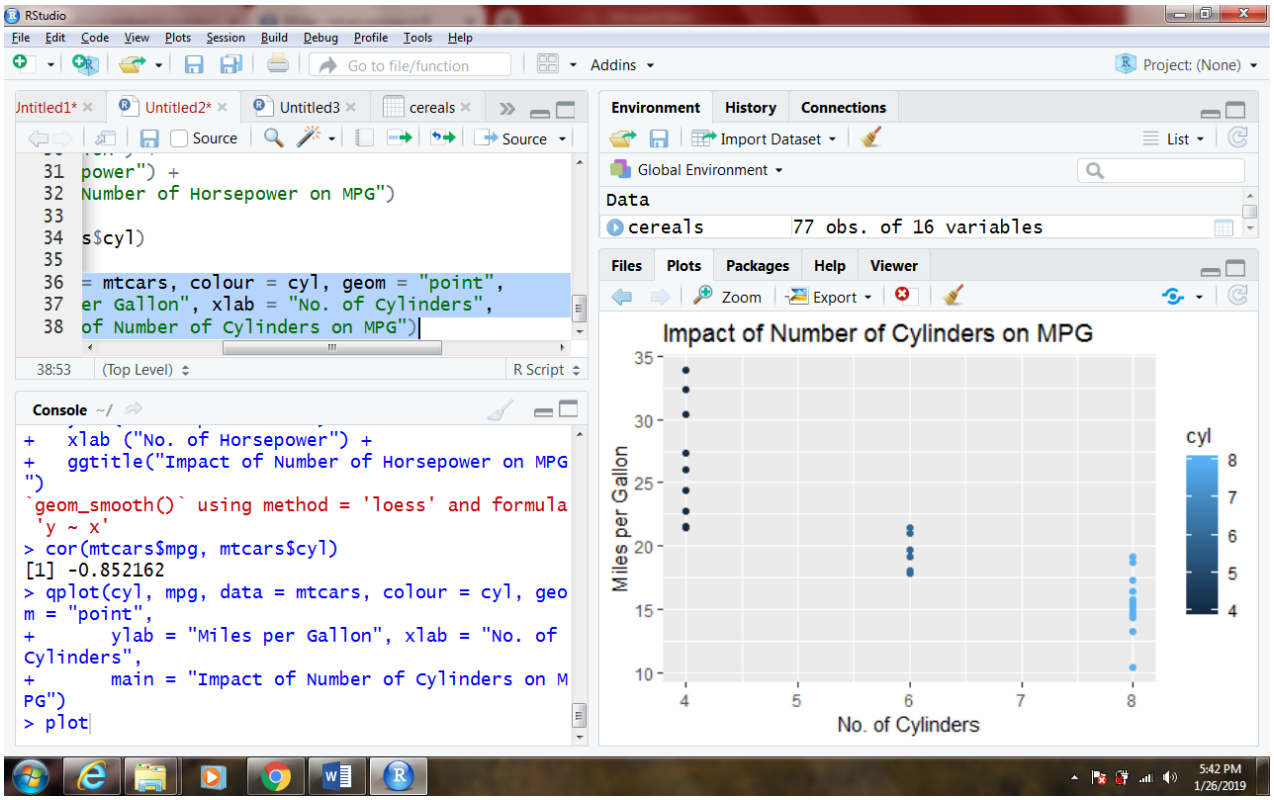
cor(mtcars$mpg, mtcars$cyl)

```
[1] -0.852162
```

This gives an even stronger negative correlation of -0.85

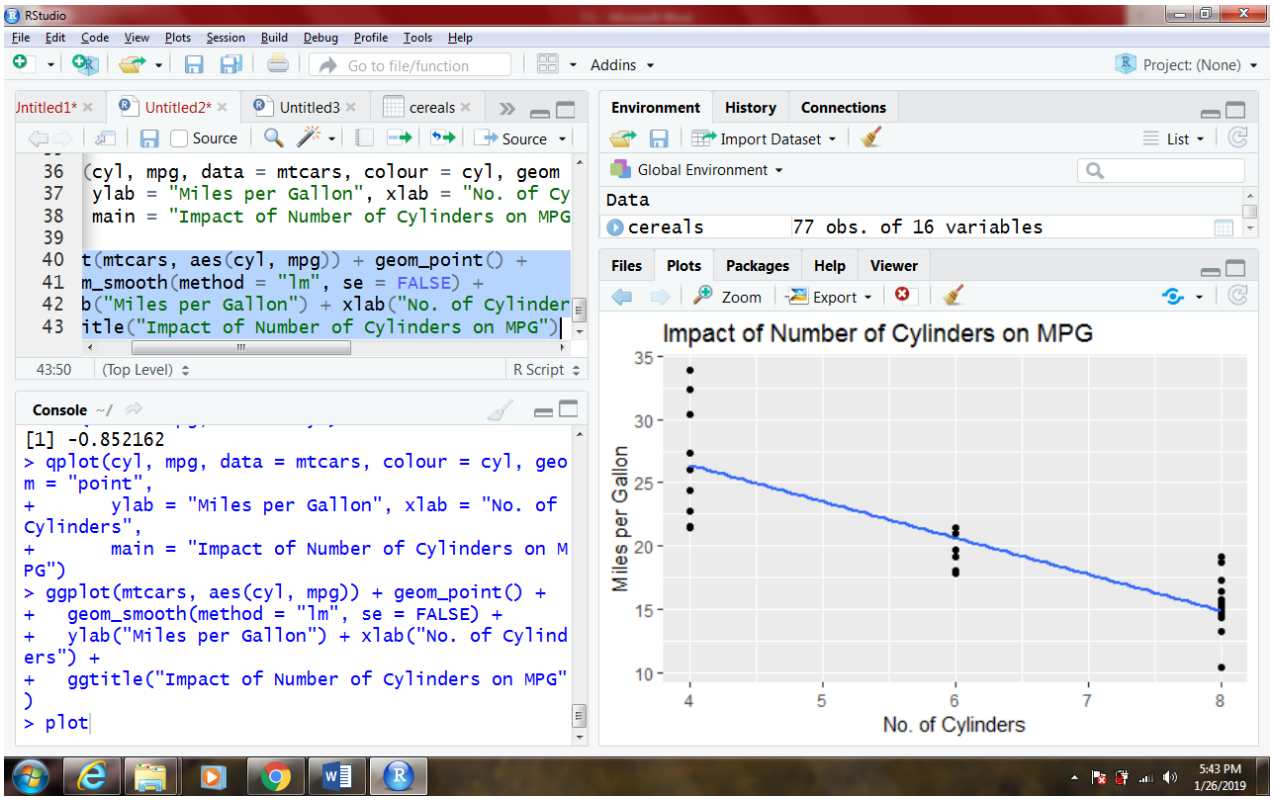
Scatter plot

```
qplot(cyl, mpg, data = mtcars, colour = cyl, geom = "point",
      ylab = "Miles per Gallon", xlab = "No. of Cylinders",
      main = "Impact of Number of Cylinders on MPG")
```



Cyl vrs mpg

```
ggplot(mtcars, aes(cyl, mpg)) + geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ylab("Miles per Gallon") + xlab("No. of Cylinders") +
  ggtitle("Impact of Number of Cylinders on MPG")
```



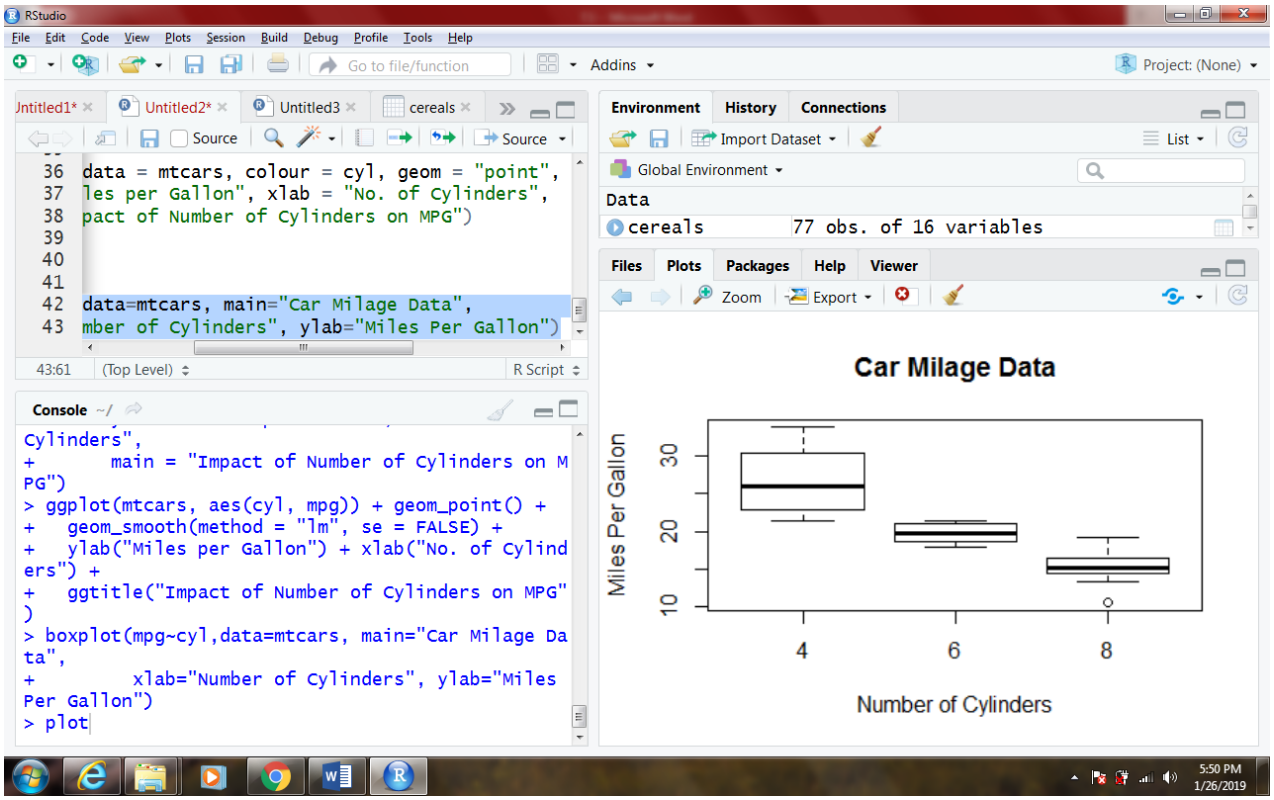
Our analysis shows a strong negative correlation for both number of horsepower (-0.77) as well as number of cylinders (-0.85) on miles per gallon.

As horsepower or cylinders increase, we see miles per gallon decreasing. While both have a strong negative correlation, we find that the impact of having more cylinders in a car has a greater negative impact on miles per gallon achieved.

3. Write a program to create boxplot for all variables.

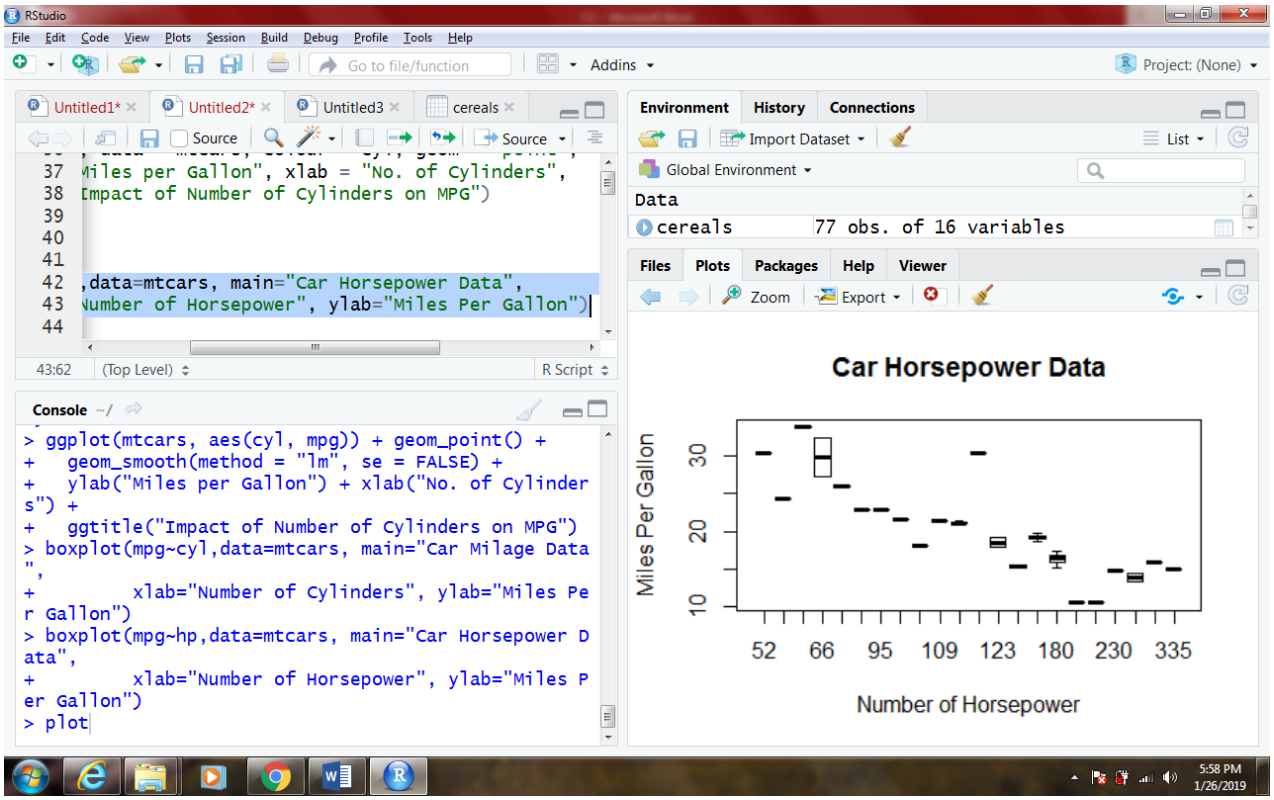
Ans – Box plot of MPG by car cylinders

```
boxplot(mpg~cyl,data=mtcars, main="Car Milage Data",
    xlab="Number of Cylinders", ylab="Miles Per Gallon")
```

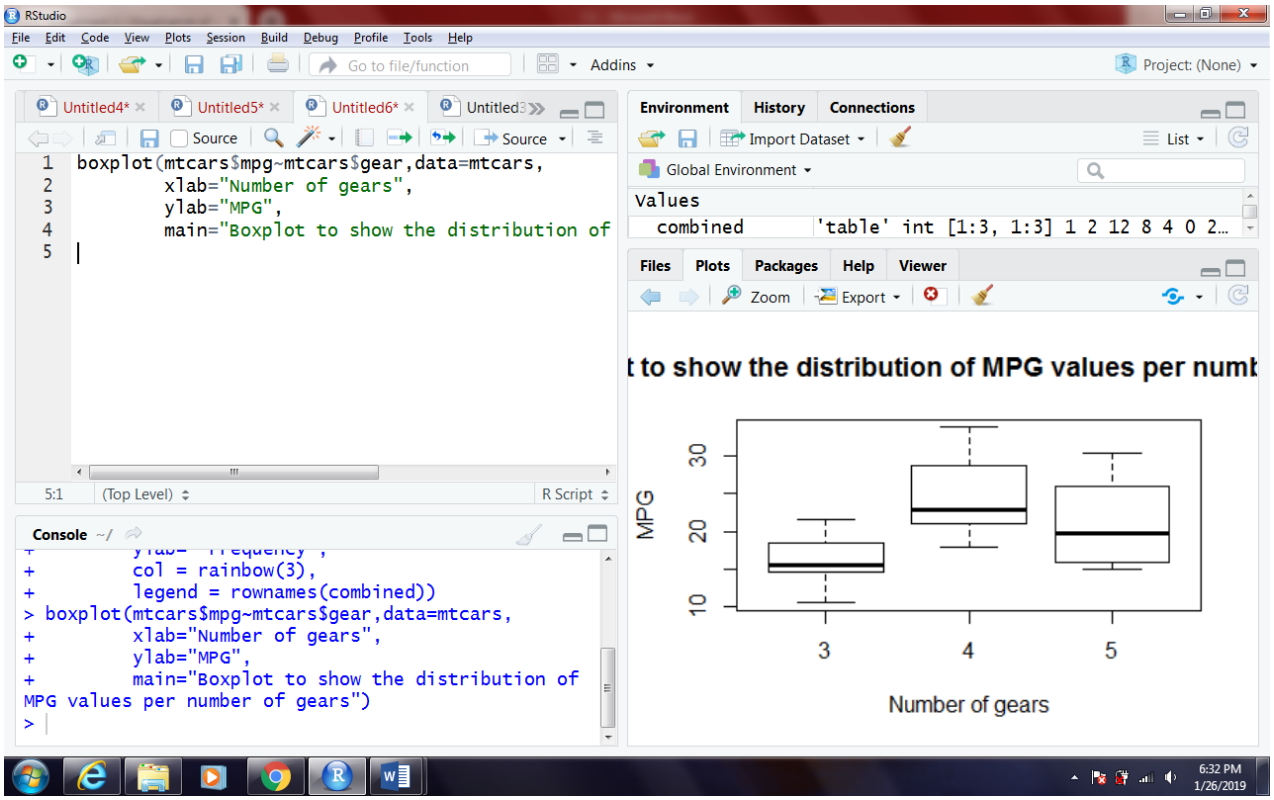


Horsepower mileage data

```
boxplot(mpg~hp,data=mtcars, main="Car Horsepower Data",
      xlab="Number of Horsepower", ylab="Miles Per Gallon")
```



Boxplot to show the distribution of MPG values per number of gears



6. Expected Format

1. R file should be submitted where applicable.
2. R file should be in PDF or in .r format
3. Proper screenshots of the outputs should be submitted as well
4. The r codes, if submitted in any other format, will be subjected to deduction in marks

Note: Your solution will not be entertained if it is any other format, e.g., .zip, .doc, .rtf etc.

7. Approximate Time to Complete Task

30 mins.