



ACADGILD

SESSION 7: Basic Statistics

Assignment 1

Table of Contents

1. Introduction 3

2. Objective 3

3. Prerequisites 3

4. Associated Data Files 3

5. Problem Statement..... 3

7. Approximate Time to Complete Task 18

6.Expected Output	3
-------------------------	---

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)
```

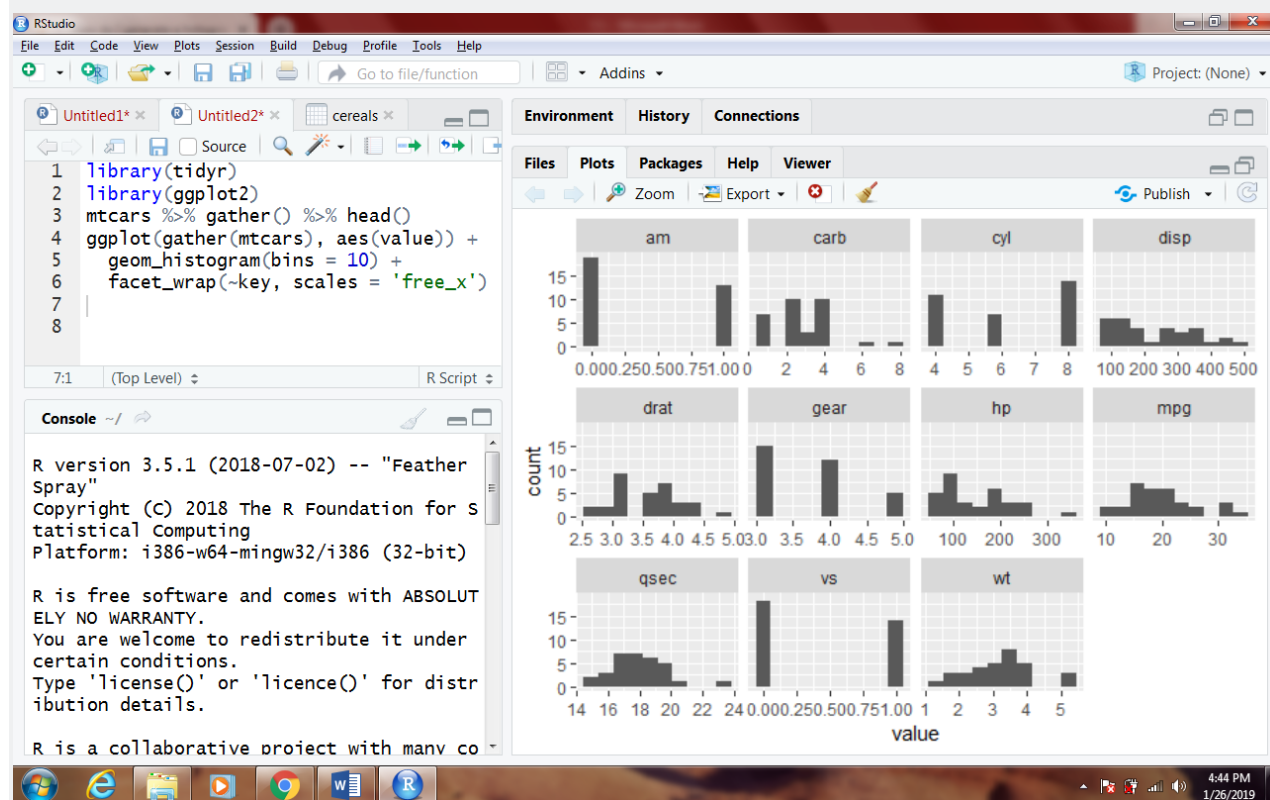
```
mtcars %>% gather() %>% head()
```

Data Analytics

```
#> 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')
```

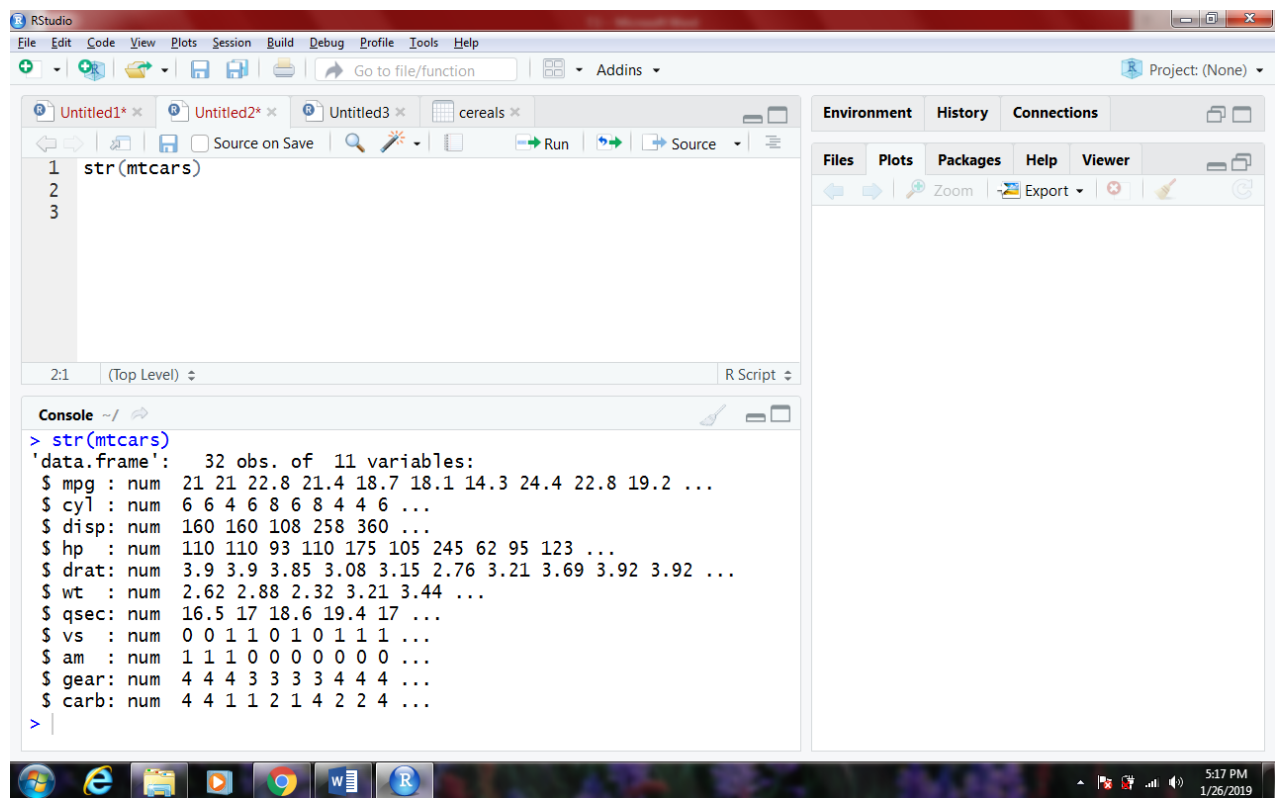


2. Check the probability distribution of all variables in **mtcars**.

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

`Str(mtcars)`

Data Analytics

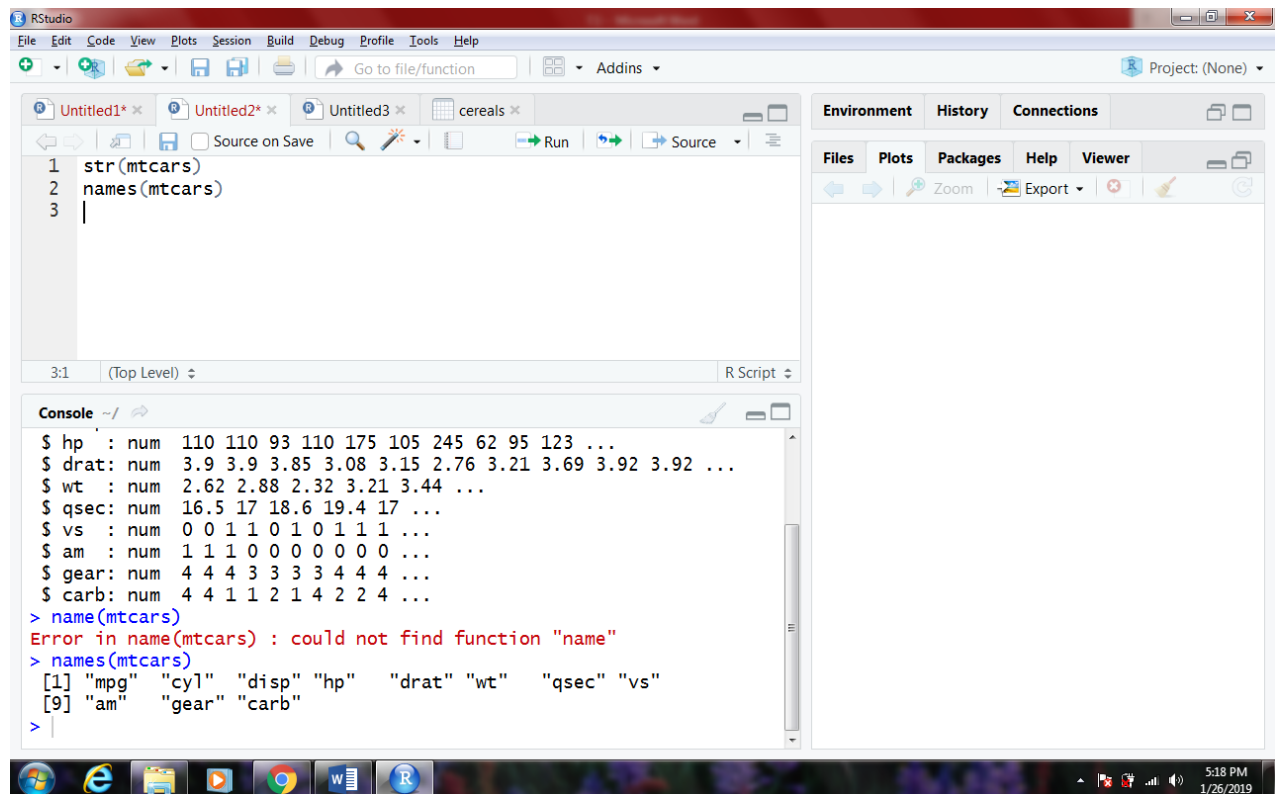


The screenshot shows the RStudio interface. The source editor on the left contains the command `str(mtcars)`. The console on the right displays the output of this command, which is a detailed summary of the `mtcars` data frame, including the number of observations (32) and the data types of the 11 variables.

```
1 str(mtcars)
2
3
```

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num   6  6  4  6  8  6  8  4  4  6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num   3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num   2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num   0  0  1  1  0  1  0  1  1  1 ...
 $ am  : num   1  1  1  0  0  0  0  0  0 ...
 $ gear: num   4  4  4  3  3  3  4  4  4 ...
 $ carb: num   4  4  1  1  2  1  4  2  2  4 ...
```

names(mtcars)



The screenshot shows the RStudio interface. The source editor on the left contains the command `names(mtcars)`. The console on the right displays the output of this command, which is a character vector containing the names of the variables in the `mtcars` data frame. The output is displayed in two lines.

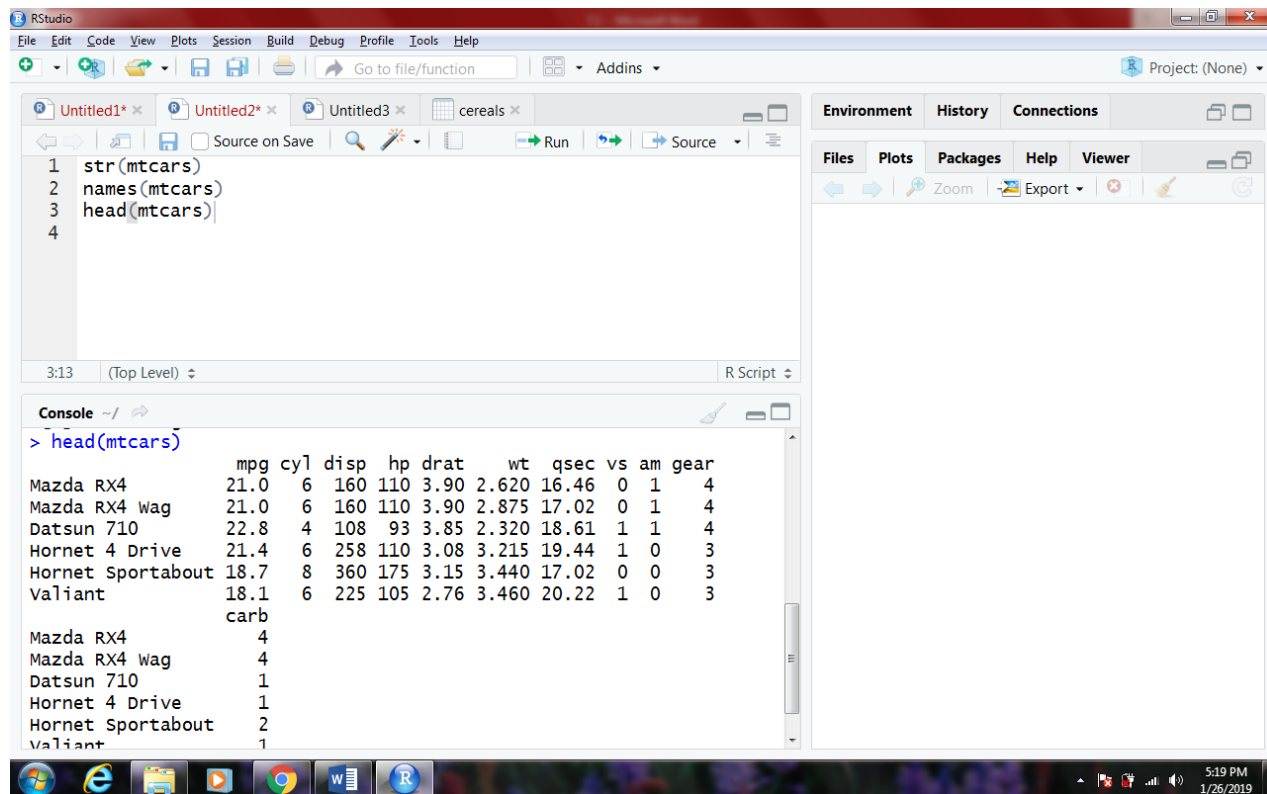
```
1 str(mtcars)
2 names(mtcars)
3
```

```
$ hp : num  110 110 93 110 175 105 245 62 95 123 ...
$ drat: num   3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
$ wt  : num   2.62 2.88 2.32 3.21 3.44 ...
$ qsec: num  16.5 17 18.6 19.4 17 ...
$ vs  : num   0  0  1  1  0  1  0  1  1  1 ...
$ am  : num   1  1  1  0  0  0  0  0  0 ...
$ gear: num   4  4  4  3  3  3  4  4  4 ...
$ carb: num   4  4  1  1  2  1  4  2  2  4 ...

> name(mtcars)
Error in name(mtcars) : could not find function "name"
> names(mtcars)
[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs"
[9] "am" "gear" "carb"
```

Data Analytics

head(mtcars)



The screenshot shows the RStudio interface. The source editor contains the following R code:

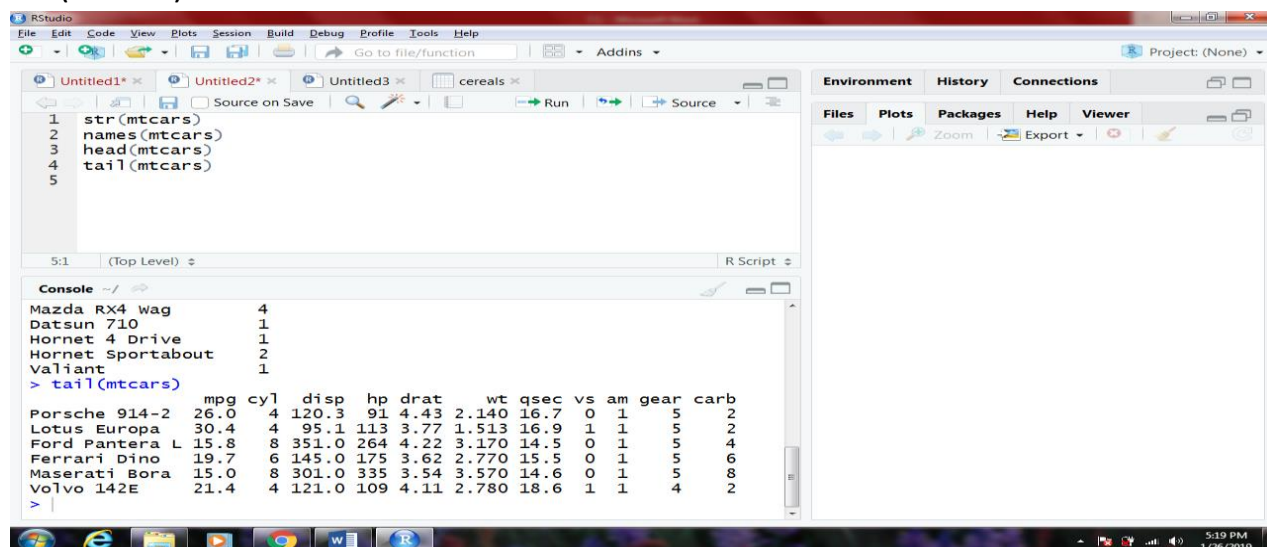
```
1 str(mtcars)
2 names(mtcars)
3 head(mtcars)
4
```

The console displays the output of the `head(mtcars)` command, showing the first six rows of the `mtcars` dataset. The output is as follows:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3

The console also shows the output of `str(mtcars)` and `names(mtcars)`, which are partially visible at the top of the console output.

tail(mtcars)



The screenshot shows the RStudio interface. The source editor contains the following R code:

```
1 str(mtcars)
2 names(mtcars)
3 head(mtcars)
4 tail(mtcars)
5
```

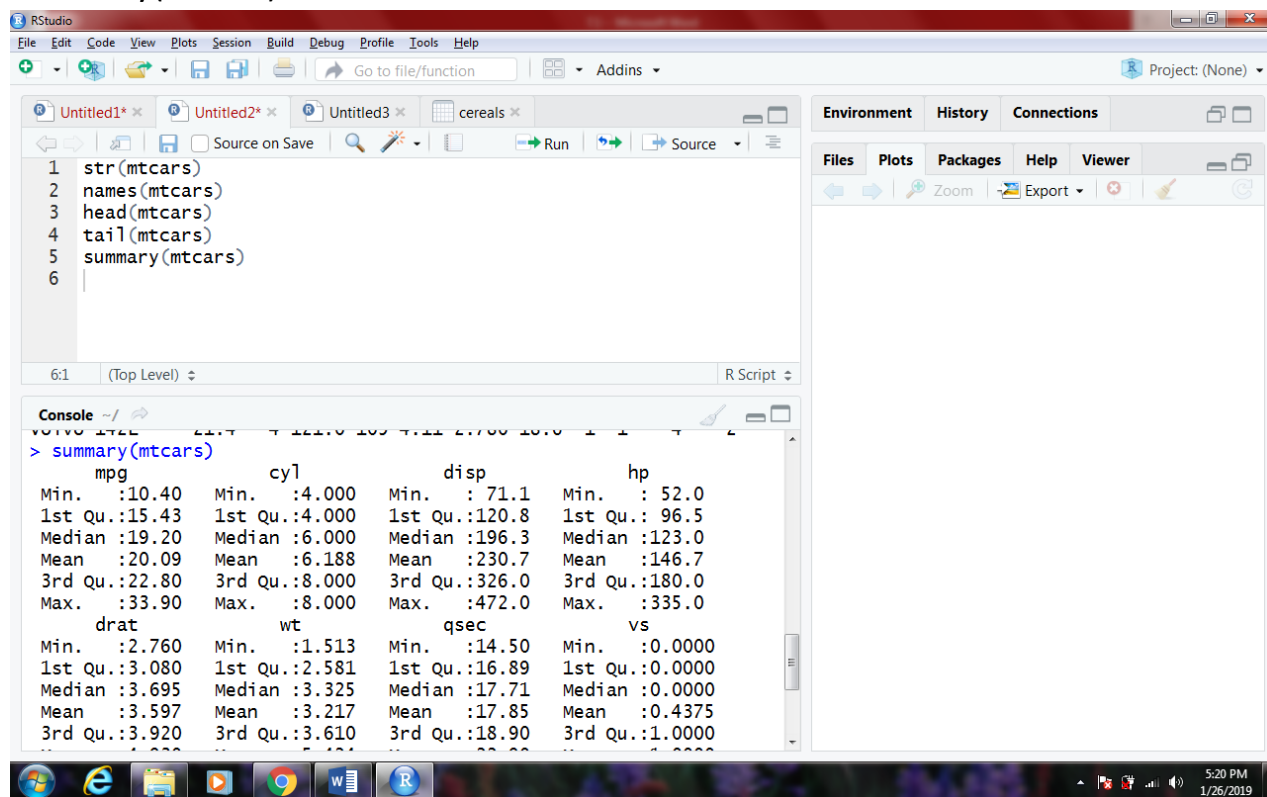
The console displays the output of the `tail(mtcars)` command, showing the last six rows of the `mtcars` dataset. The output is as follows:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.5	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.5	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.6	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.6	1	1	4	2

The console also shows the output of `str(mtcars)` and `names(mtcars)`, which are partially visible at the top of the console output.

Data Analytics

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)

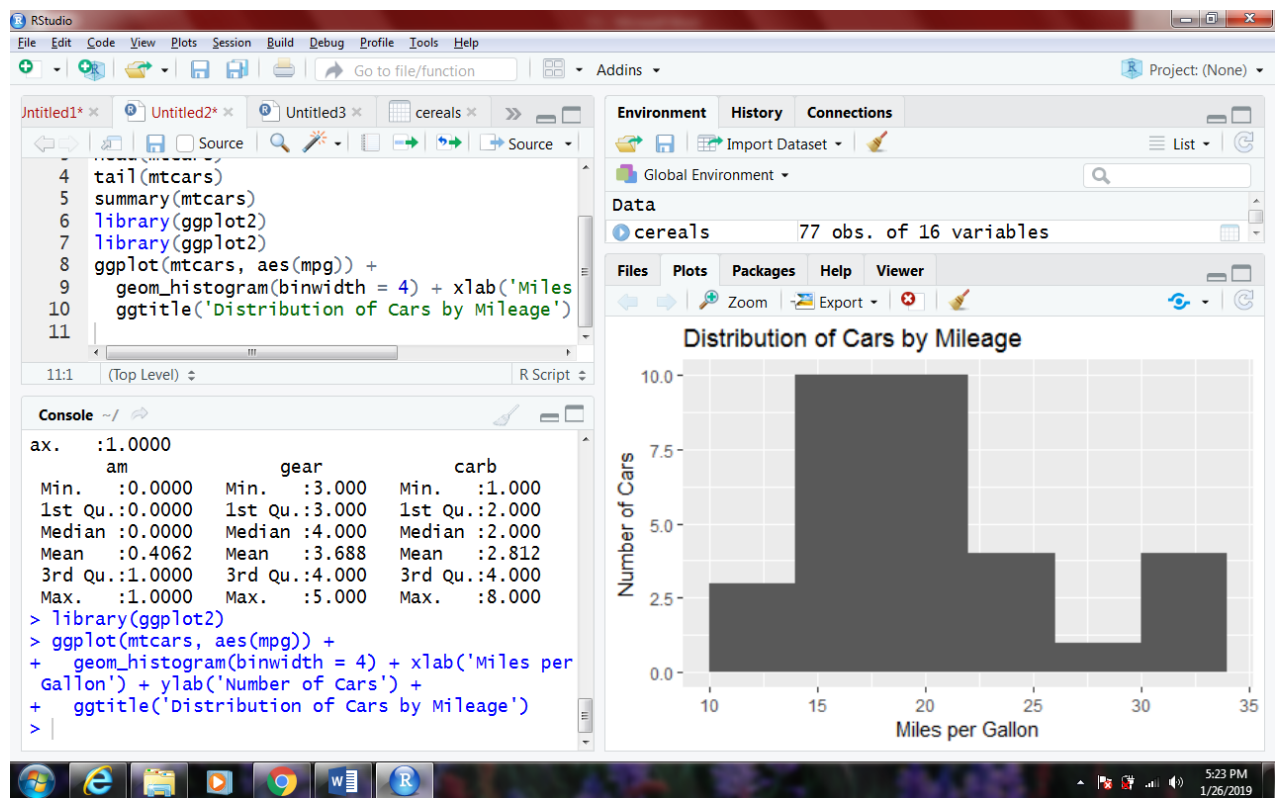
```
ggplot(mtcars, aes(mpg)) +
```

```
  geom_histogram(binwidth = 4) + xlab('Miles per Gallon') + ylab('Number of Cars')
```

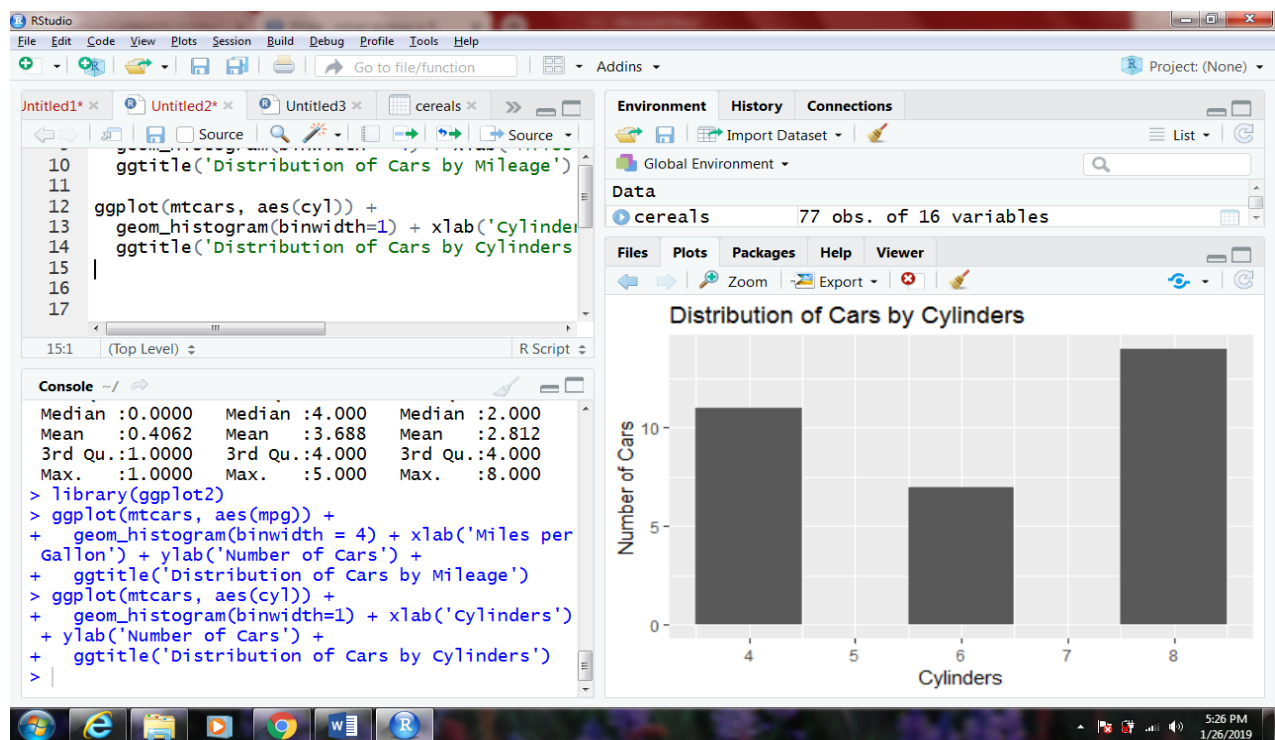
```
+
```

```
  ggtitle('Distribution of Cars by Mileage')
```


Data Analytics



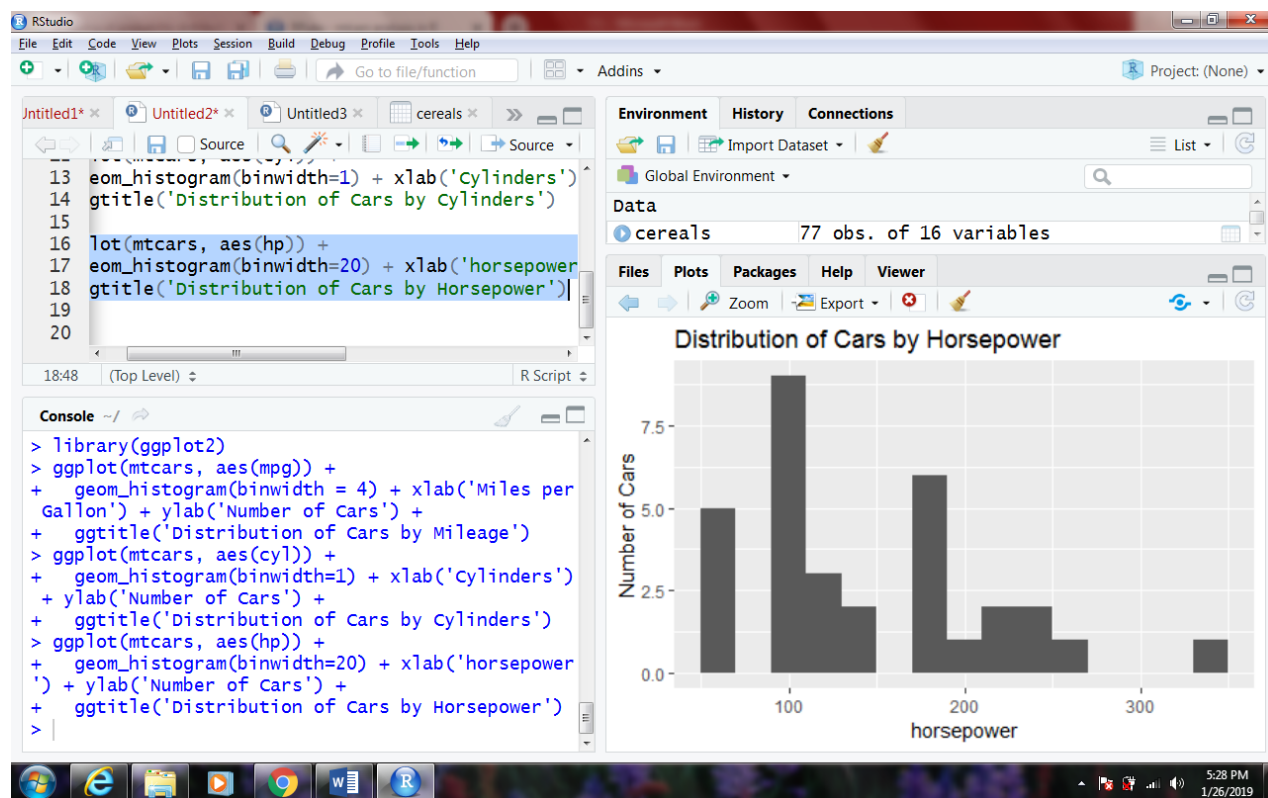
Now we show the histogram for number of cylinders:



Data Analytics

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
```

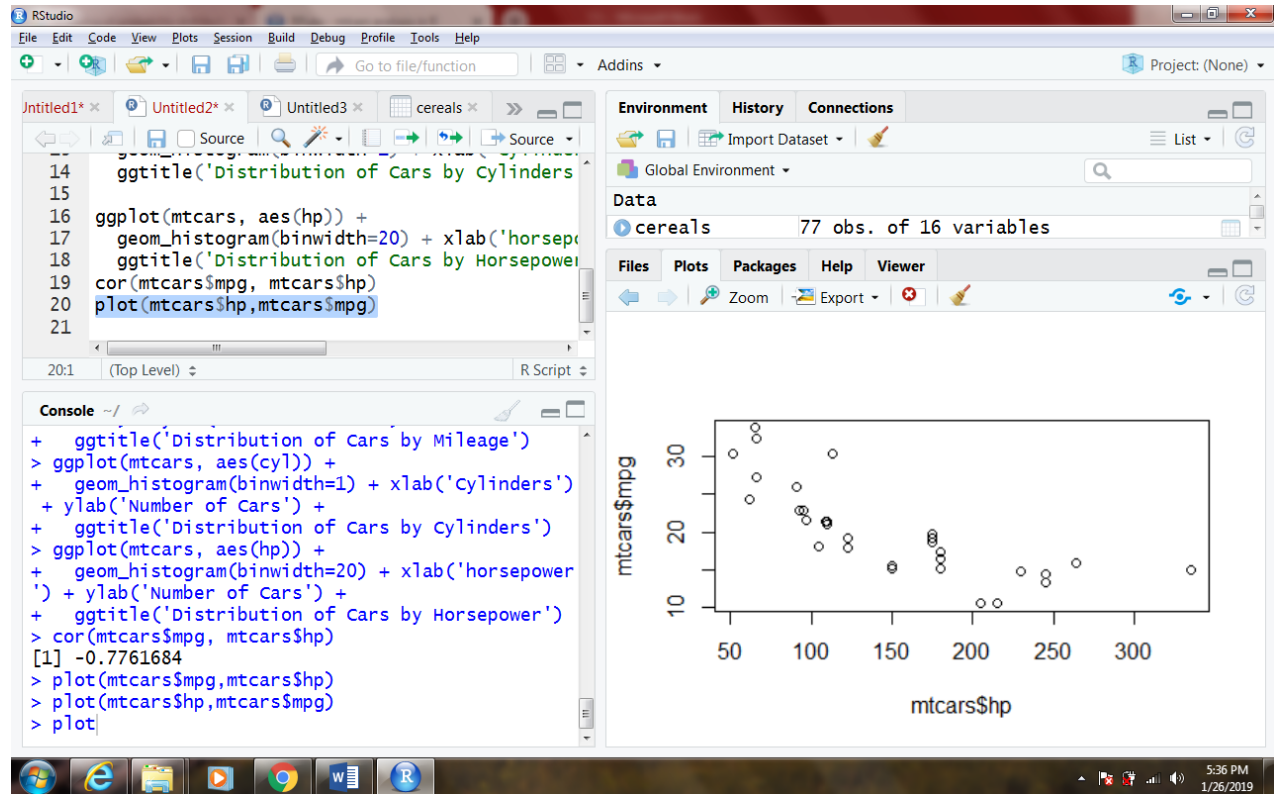
We find a fairly strong negative correlation.

Now we plot the data – HP vs MPG

Data Analytics

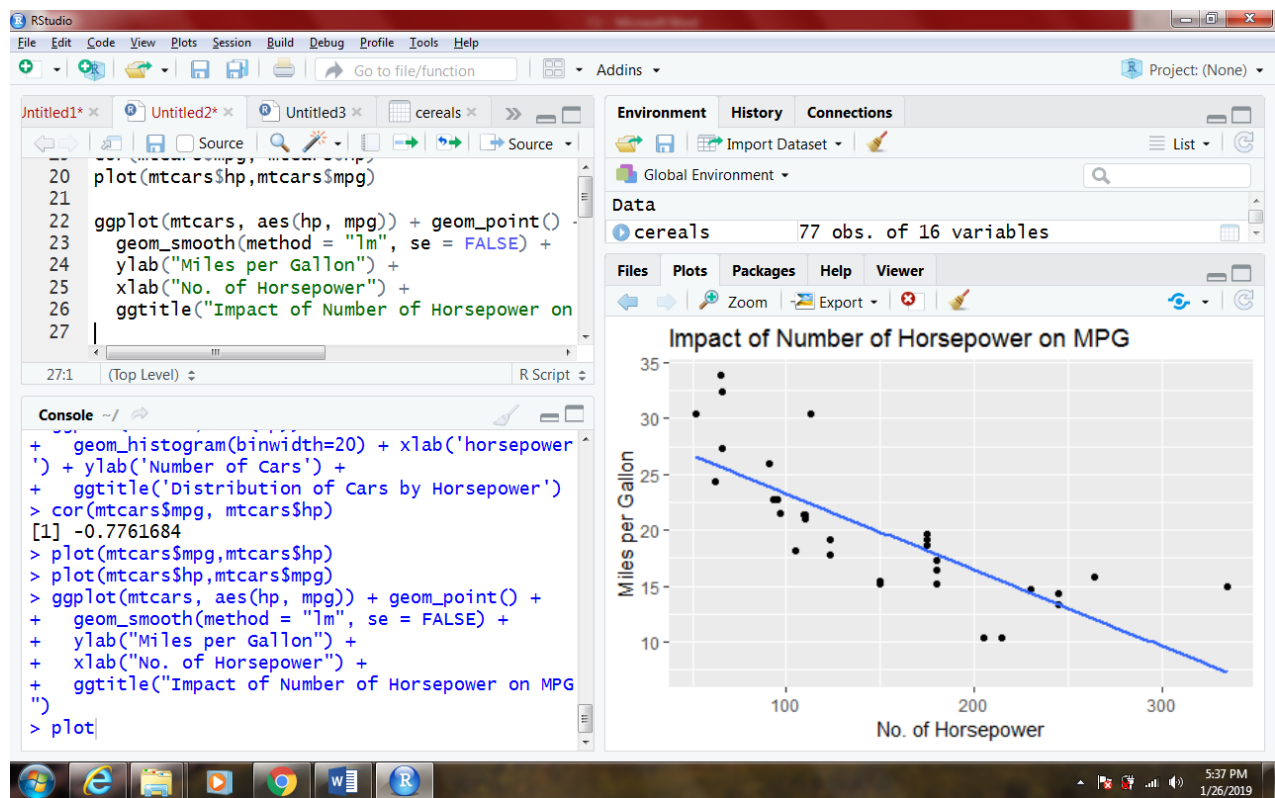
Below is the effect that number of horsepower has on mpg. We have also shown transmission type (manual = 1, auto = 0) as a point of reference, but it is not a primary part of our analysis.

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



HPv/s MPG

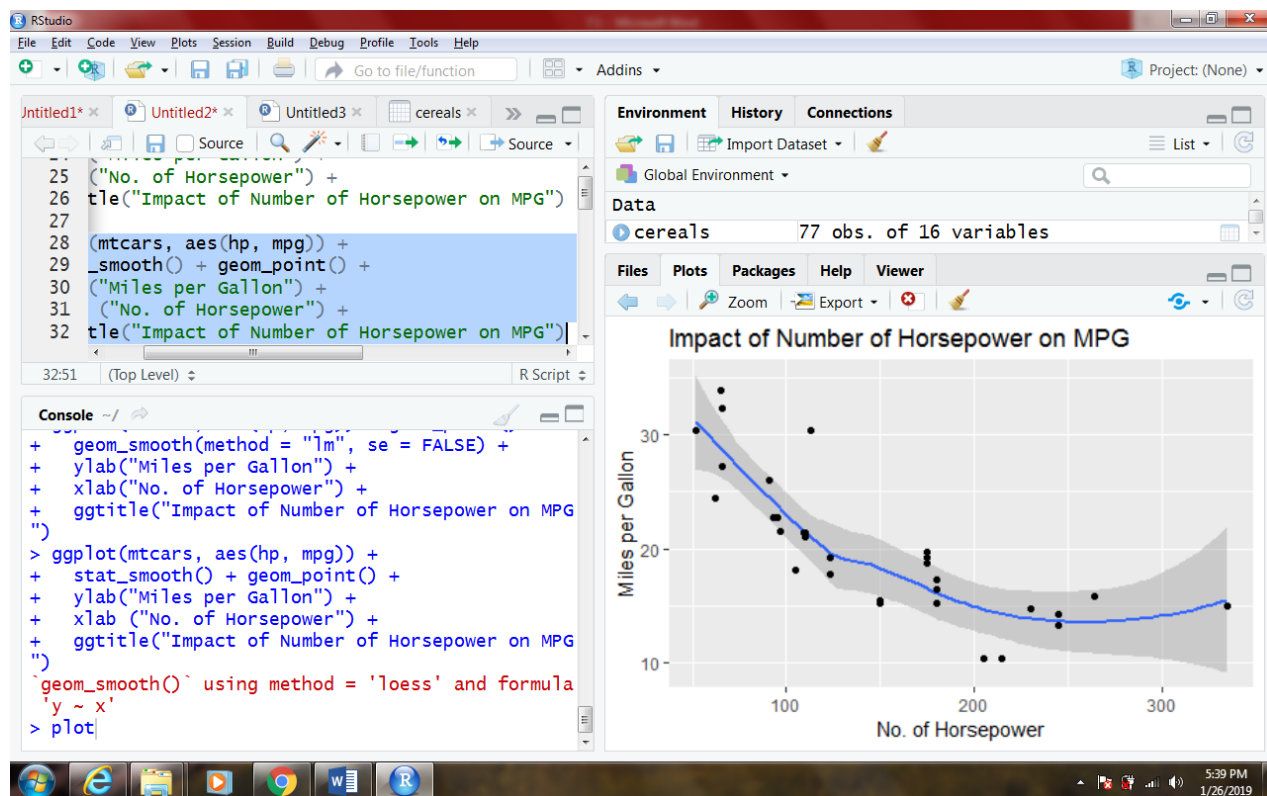
Data Analytics



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")
```

Data Analytics



Effect of number of cylinders on mpg

The correlation of mpg and cyl is shown below.

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

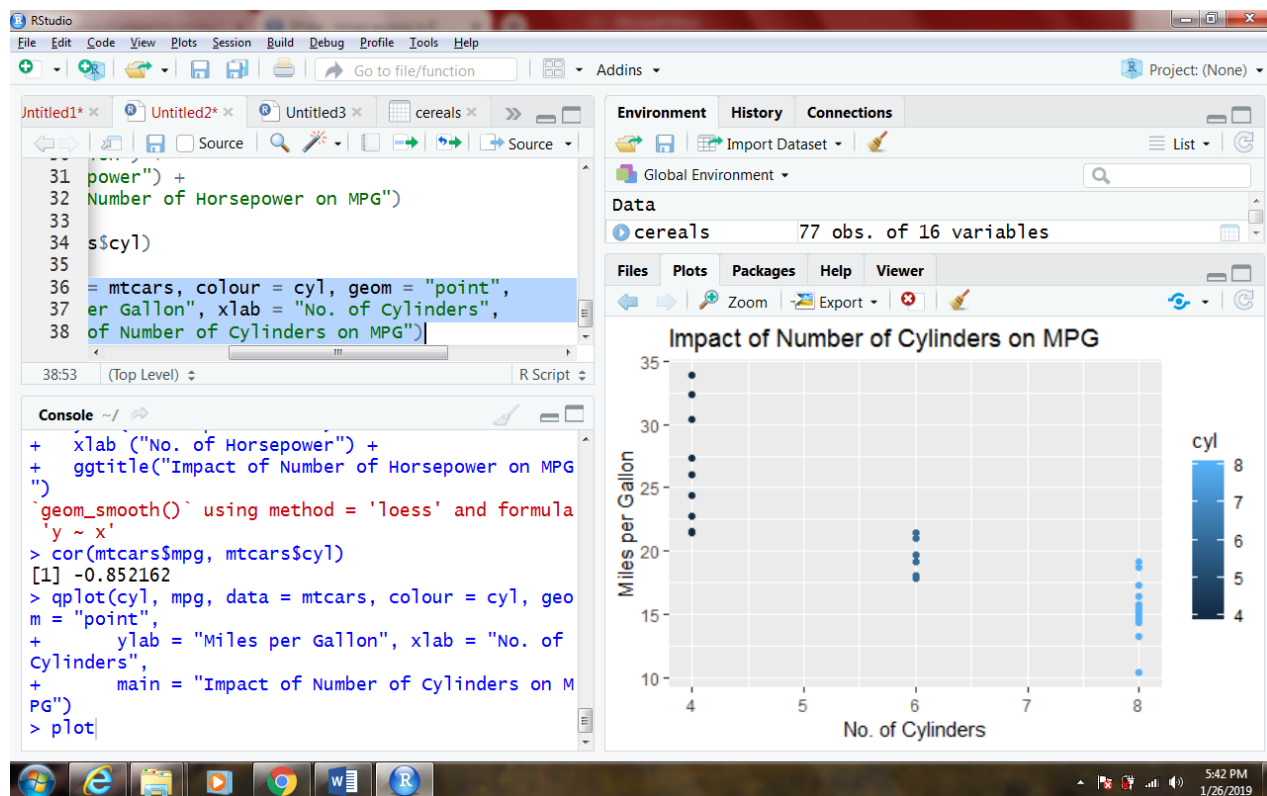
```
[1] -0.852162
```

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")
```

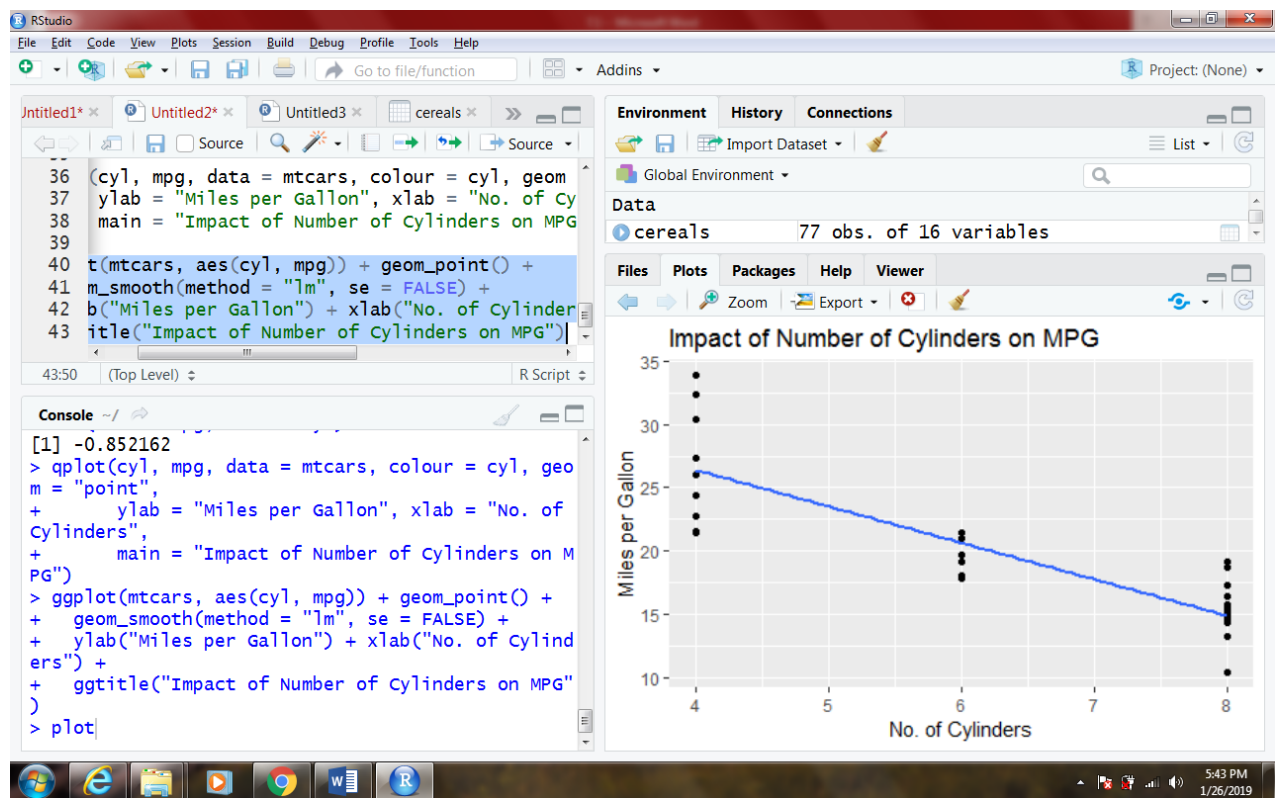
Data Analytics



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")
```

Data Analytics



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.

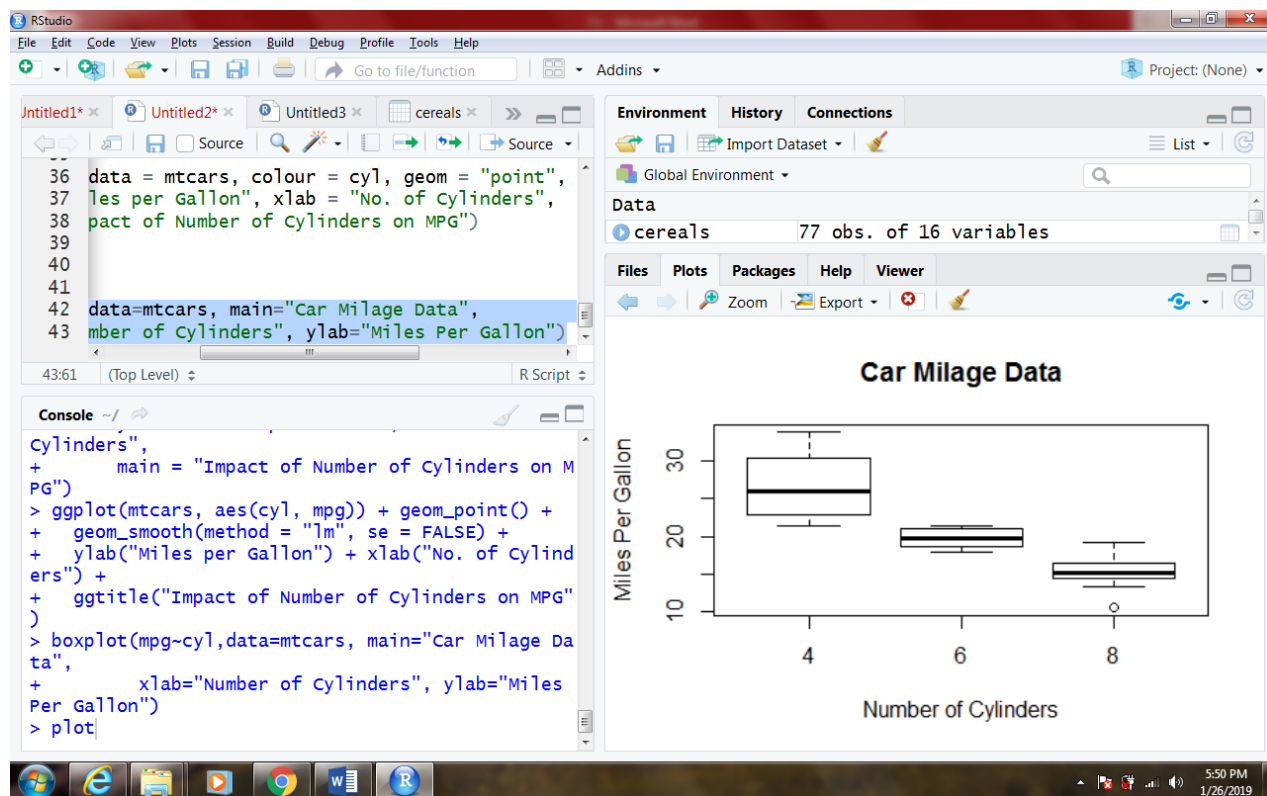
As horsepower or cylinders increase, we see miles per gallon decreasing. While both have a strong negative correlation, we find that the impact of having more cylinders in a car has a greater negative impact on miles per gallon achieved.

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")
```

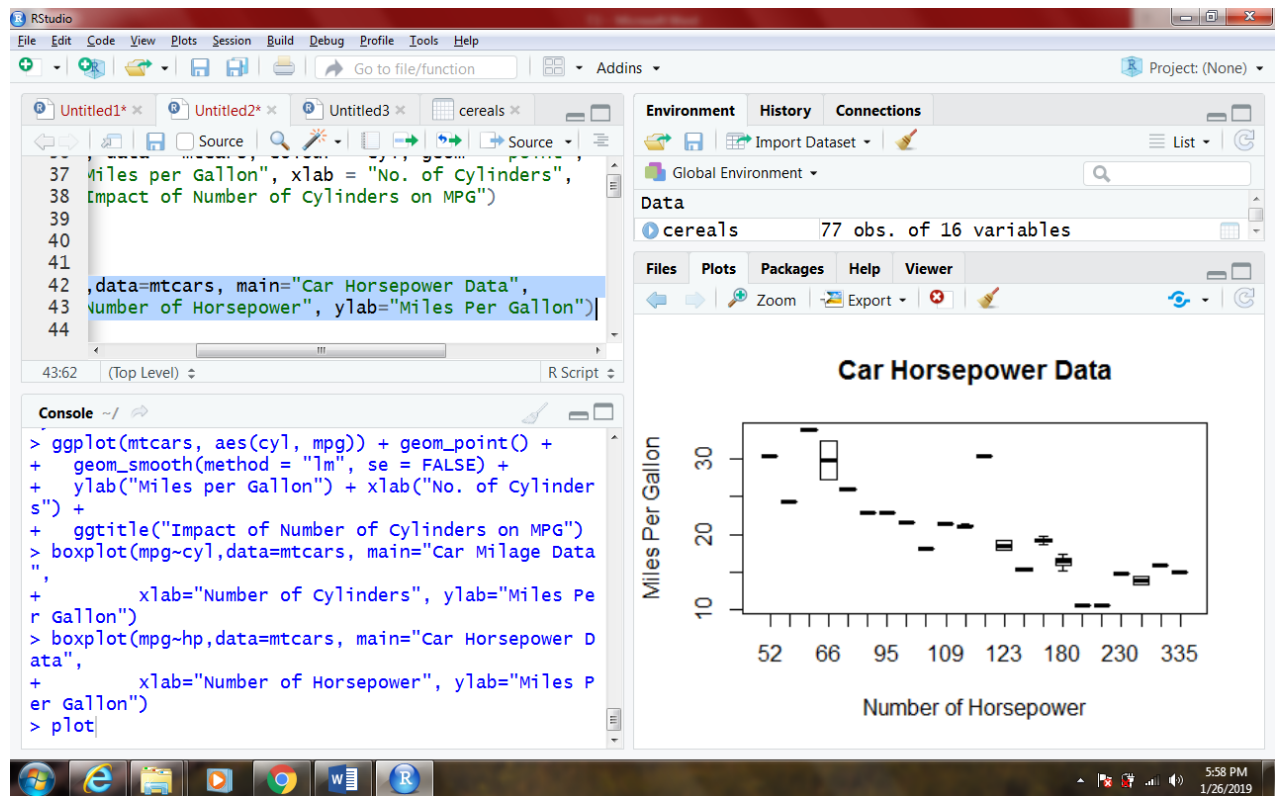
Data Analytics



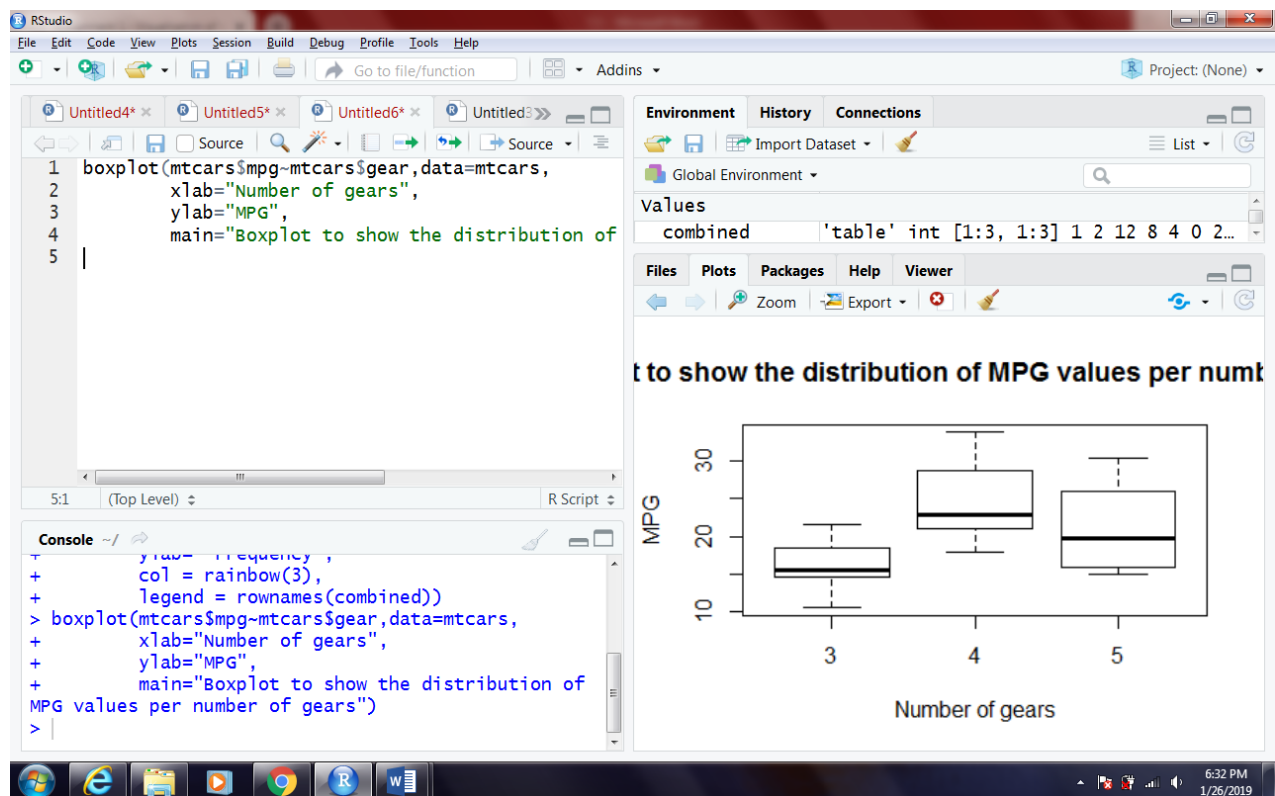
Horsepower mileage data

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


Data Analytics



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.