Terrence Zhu
Gaurav Suri
Psy300 Sec. 01
Sep 7th, 2021

PSY 300, Fall, 2021 Introduction to Data Science. HW 1

This HW is due on Wednesday, Sept 8 at Noon. HWs submitted after this deadline will be deducted 10% / day

Instructions: Please submit your HW in a word file (or google docs. If you do google docs, make sure I have permission to view the document). Please copy paste code and output from R Studio into your HW document. Here is an example question and response.

* * *

Example Question: Assign the value 161 to variable called 'var1'. Find the cube of the contents of 'var1'

Answer:

Code:
```
var1 <- 161
var1^3
```

Output:
```
4173281
```

Comments:
If you have comments you'd put them here

* * *

Some responses may involve graphs as output. Just export those and paste them into your HW.

There are ten questions.

Q1) (1 point each. Total 10 points) Perform the following arithmetic and object creation in R.

a) 8+8+4

Answer:
Code: 8+8+4
Output: [1] 20

b) 8-4

Answer:
Code: 8-4
Output: [1] 4

c) 8/4

Answer:

Code: 8/4

Output: [1] 2

d) 8*4

Answer:

Code: 8*4

Output: [1] 32

e) Create an object named 'a' composed of 8+4

Answer:

Code: a<-8+4

f) Create an object named 'b' composed of 8-4

Answer:

Code: b<-8-4

g) Multiply 'a' and 'b'

Answer:

Code: a*b

Output: [1] 48

h) Divide 'a' by 'b'

Answer:

Code: a/b

Output: [1] 3

i) Compute 'a' raised to the power 'b'

Answer:

Code: a^b

Output: 20736

j) Find a function (Google if you need to) to find the maximum of 'a' and 'b'

Answer:

Code: max(a)

max(b)

Output: [1] 12

[1] 4

"https://www.datamentor.io/r-programming/examples/minimum-maximum/"

Q2) (10 points) (a) Create a vector called 'v1' composed of the numbers 3,4,5,6,7 (b) Create a vector called 'v2' composed of the numbers 23,14,18,16,21 (c) Perform the operation v1 + v2. Describe what R did. (d) Write a statement to access the third element of 'v2'.

Answer:

a) Code: v1<-c(3,4,5,6,7)

b) Code: v2<-c(23,14,18,16,21)

c) Code: v1 + v2

Output: [1] 26 18 23 22 28

R use calculation to sum up each individual of the same position in the list, v1 with the same position of the other list, v2.

d) Code: v2[3]

Output: [1] 18

Q3) (10 points) Examine the dataset 'iris' that comes standard in R. (a) How many variables does it have? (b) which variables are quantitative and which are qualitative?

Answer:

a) Code: head(iris)

Output: There are 5 variables in the dataset 'iris'.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

b) The quantitative variables are sepal length, sepal width, petal length, and petal width. The qualitative variable is species.

Q4) (10 points) (a) Write one line of code to access the Sepal.Width column of 'iris'. (b) Find the mean and standard deviation of Sepal.Width.

Answer:

a) Code: iris[, 2]

Output:

```
 [1] 3.5 3.0 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 3.7 3.4 3.0 3.0 4.0 4.4 3.9 3.5 3.8 3.8
[21] 3.4 3.7 3.6 3.3 3.4 3.0 3.4 3.5 3.4 3.2 3.1 3.4 4.1 4.2 3.1 3.2 3.5 3.6 3.0 3.4
[41] 3.5 2.3 3.2 3.5 3.8 3.0 3.8 3.2 3.7 3.3 3.2 3.2 3.1 2.3 2.8 2.8 3.3 2.4 2.9 2.7
[61] 2.0 3.0 2.2 2.9 2.9 3.1 3.0 2.7 2.2 2.5 3.2 2.8 2.5 2.8 2.9 3.0 2.8 3.0 2.9 2.6
[81] 2.4 2.4 2.7 2.7 3.0 3.4 3.1 2.3 3.0 2.5 2.6 3.0 2.6 2.3 2.7 3.0 2.9 2.9 2.5 2.8
[101] 3.3 2.7 3.0 2.9 3.0 3.0 2.5 2.9 2.5 3.6 3.2 2.7 3.0 2.5 2.8 3.2 3.0 3.8 2.6 2.2
[121] 3.2 2.8 2.8 2.7 3.3 3.2 2.8 3.0 2.8 3.0 2.8 3.8 2.8 2.8 2.6 3.0 3.4 3.1 3.0 3.1
[141] 3.1 3.1 2.7 3.2 3.3 3.0 2.5 3.0 3.4 3.0
```

b) Code: mean(iris[, 2])

sd(iris[,2])

Output: [1] 3.057333

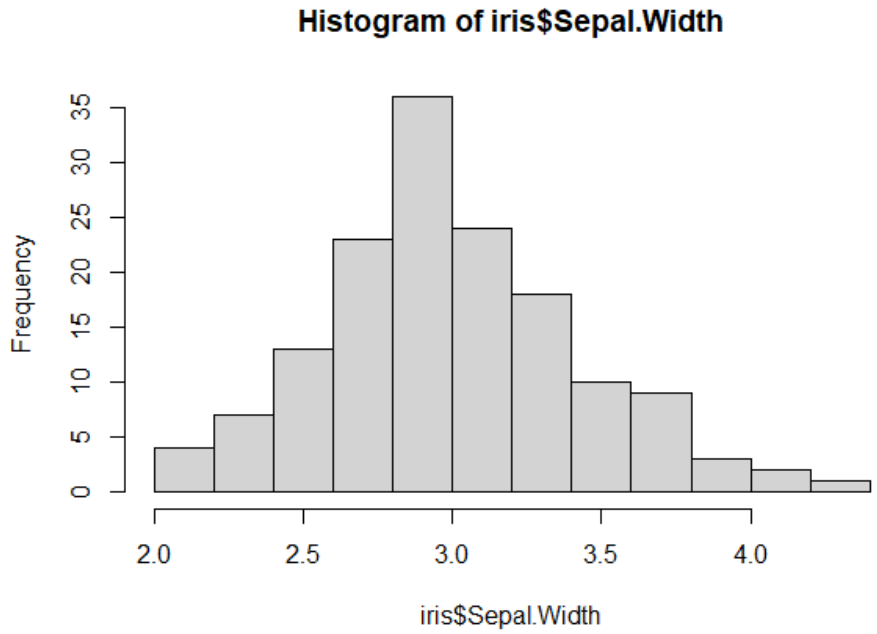
[1] 0.4358663

Q5) (10 points) Draw a histogram of Sepal.Width in 'iris'. If you choose a flower at random, what range (interval) is its sepal width most likely to lie in (In the histogram the width of each interval is 0.2)

Answer:

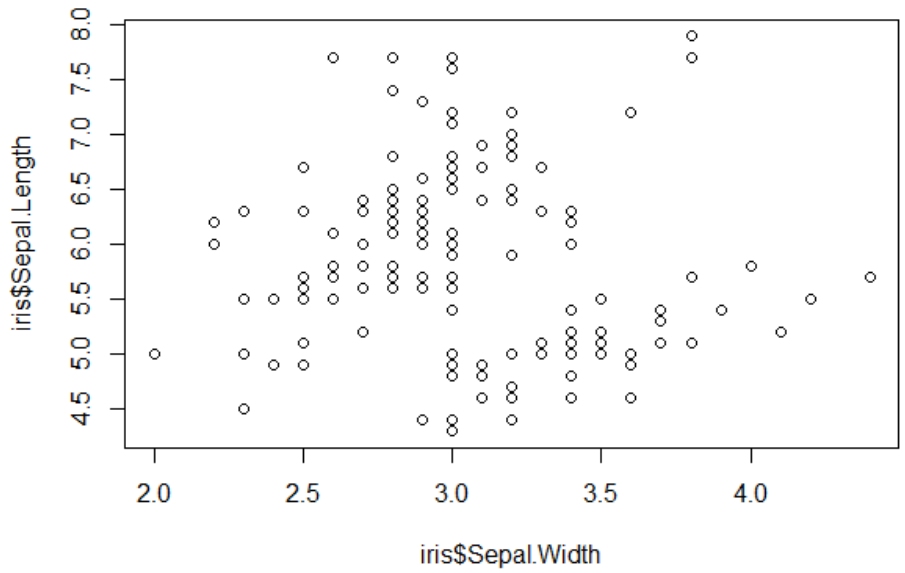
a) Code: hist(iris$Sepal.Width)

Output:



When choosing a flower at random, the most likely to be picked would be 3.0 within range of sepal width.

Q6) (10 points) Use the 'plot' function to graph the relationship between Sepal.Width (X axis) and Sepal.Length (Y axis) in the 'iris' data set. Describe (in 2-3 sentences) the relationship that exists between these two variables.
Answer:
Code: plot(iris$Sepal.Width, iris$Sepal.Length)
Output:



In this graph, it shows no relationship between the iris's sepal length and iris's sepal width. The variables in the graph are plotted in a linear horizonal line; thus, there is relationship within the graph.

Q7) (10 points) (a) Run ggplot(data = mpg). What do you see? How do you explain this? (b) How many rows are in mpg? How many columns? (c) What does the 'drv' variable describe? [Hint: 'help(mpg) might be useful].

Answer:

a) After running ggplot(data = mpg), the bottom right of R studio switch to plot and shows a gray screen.

b)

Code: mpg

Outcome:

```
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4           2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4           2    2008     4 auto(av)   f        21    30 p     compact
 5 audi         a4           2.8  1999     6 auto(l5)   f        16    26 p     compact
 6 audi         a4           2.8  1999     6 manual(m5) f        18    26 p     compact
 7 audi         a4           3.1  2008     6 auto(av)   f        18    27 p     compact
 8 audi         a4 quattro   1.8  1999     4 manual(m5) 4        18    26 p     compact
 9 audi         a4 quattro   1.8  1999     4 auto(l5)   4        16    25 p     compact
10 audi         a4 quattro   2    2008     4 manual(m6) 4        20    28 p     compact
# ... with 224 more rows
```

There are 234 rows and 11 columns in mpg.

c) The variable ‘drv’ describe the type of drive train: f = front-wheel drive, r = rear-wheel drive, 4 = 4wd

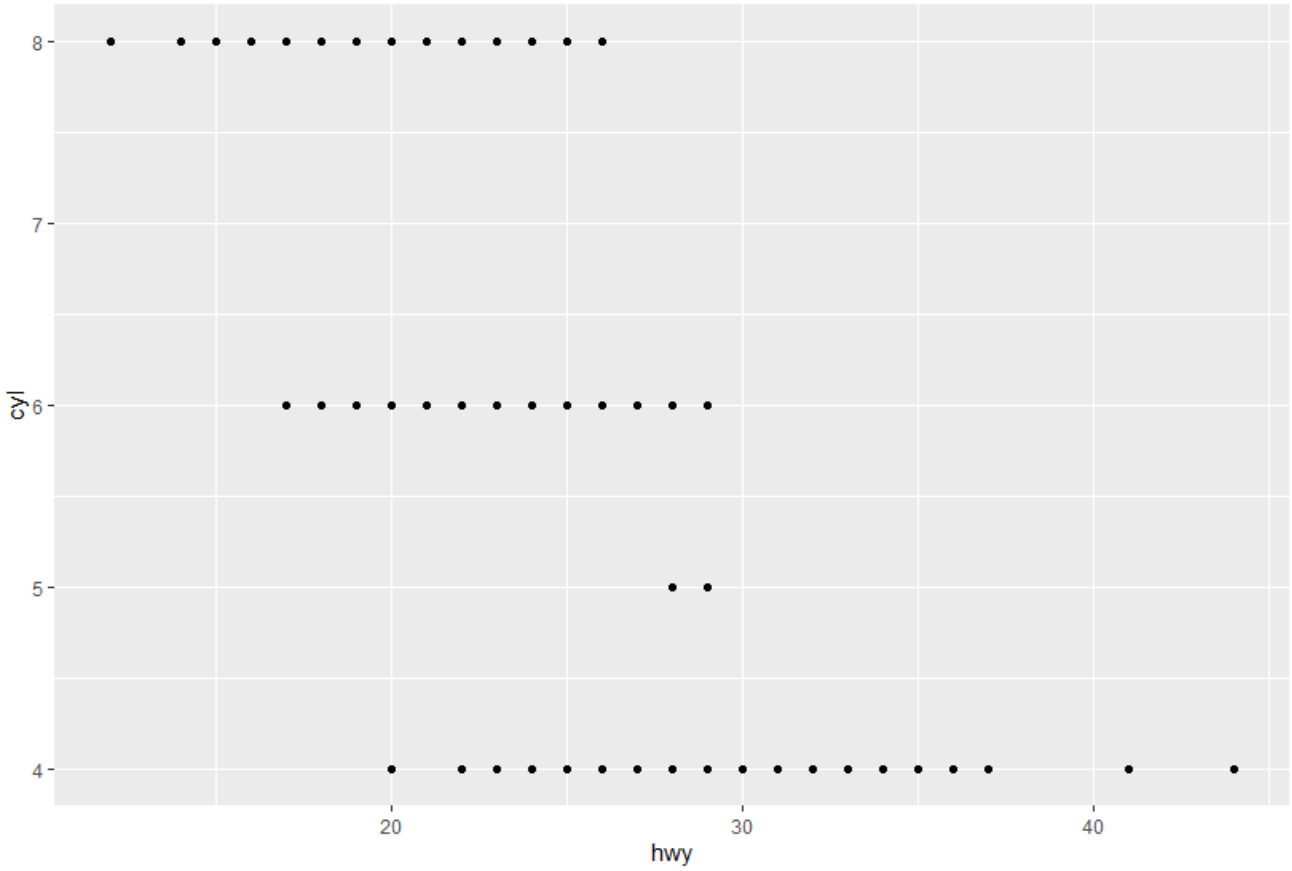
Q8) (10 points) (a) Use ggplot to make a scatterplot of ‘hwy’ vs ‘cyl’. Describe a conclusion you can derive from this data (b) What happens if you make a scatterplot of class vs ‘drv’? Why is this plot not useful?

Answer:

a)

Code: ggplot(data = mpg) + geom_point(mapping = aes(x = hwy, y = cyl))

Output:

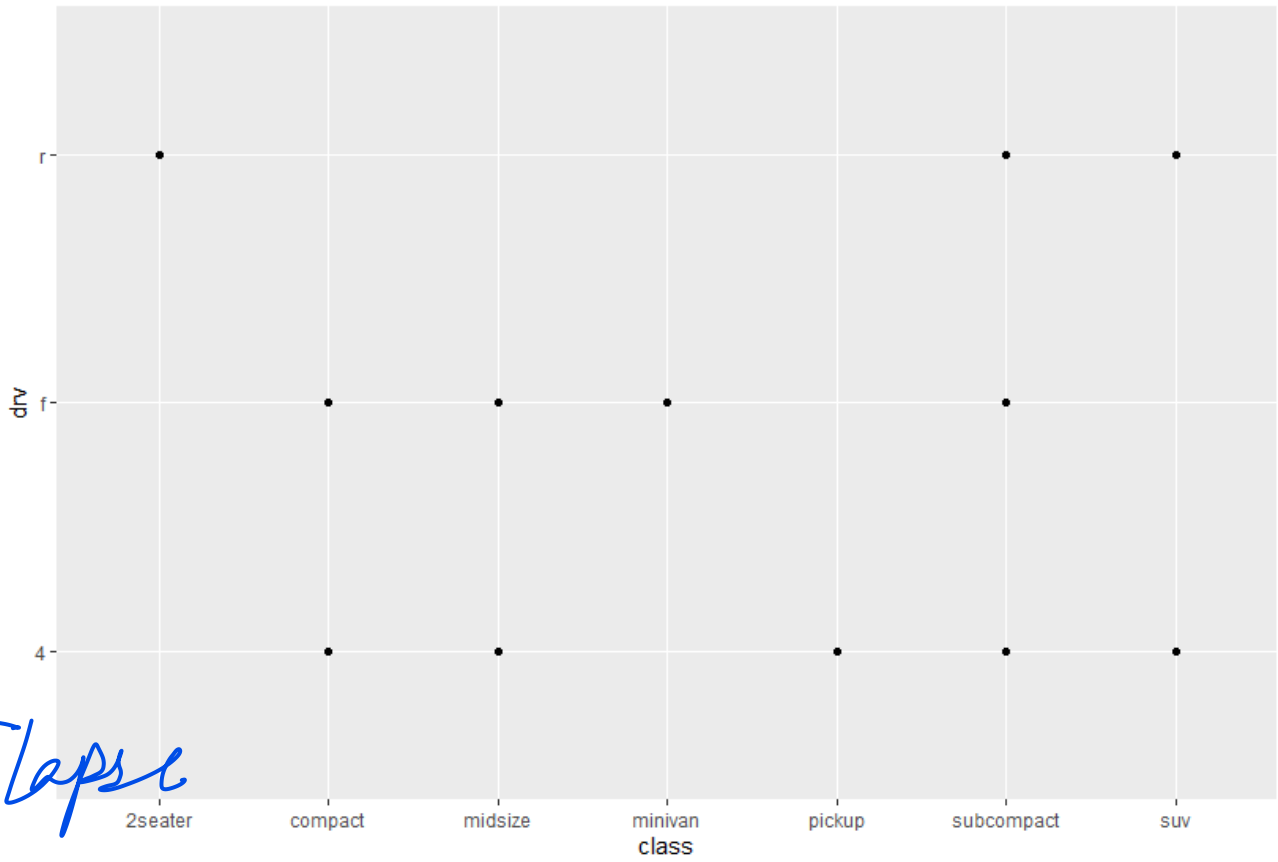


The graph seem to show a negative correlation between ‘hwy’ and ‘cyl’. As ‘hwy’ increases, ‘cyl’ decreases.

b)

Code: ggplot(data = mpg) + geom_point(mapping = aes(x = class, y = drv))

Output:



When R studio created the scatterplot of 'class' vs 'drv', the plot does not look like it is scattered everywhere. Because the values for both 'class' and 'drv' are qualitive, it does not fit well with scatterplot; but would work better if it was created with a bar graph.

Q9) (10 points) Describe in your own words (without copying and pasting) (a) what is data science (b) what data scientists do (c) some subjects that are related to data science
Answer:
a) In my own words, data science is using all the data that being collected and use it for analysis and research.
b) For data scientists, they use the data that has been collected to create a solution and conclusion to help improve anything which varies from daily life, a product or a company.

Q10) (10 points) Research a cool story that illustrates the power of data science. Describe it in your own words (2-3 paragraphs). Do not copy and paste.

As for LinkedIn, they are a leading social media platform for job and social connections. They continue to use big data to help users find connections in job searching and social opportunities to connect with others. Whereas skills endorsement, people you have connected with, and job experience you have listed, LinkedIn uses these data to help improve their mission statements on providing their users to find jobs opportunities and connect with people.

An example of LinkedIn using big data is when getting spam emails from LinkedIn about job suggestions and companies that are seeking employees' help. For instance, I was spammed with job suggestions in my email from my old job about a barista opportunity and other places such as Blue Bottle Coffee to work for them. Because I have inputted that I had worked for Target, Starbucks as a barista in the past and have listed one of my skills for being a barista, LinkedIn has used this data to send to companies to reach out to me about job opportunities. Thus, with the usage of big data, LinkedIn has used my data to send out to companies to reach out to me and using my data to companies that need help with employee seeking.

Source: https://towardsdatascience.com/apple-linkedin-netflixs-big-data-statement-fe937b7e96db