Problem Set 2 Solutions

Harris Goodman

2023-11-02

Github Link:

https://github.com/Harris-Go/EC349/blob/ac90b30a5040d84735a9059b4d77c930439d9011/ProblemSet2.Rmd

---

1. Write your name and surname in bold font, with your Name ONLY in italics.

Answer: ***Harris*** **Goodman**

2. Print your student number and what year you are in on different lines.

Answer: 2101274

3

3. Briefly state, in italics, why you are doing this Data Science Module

Answer: *I really enjoy using computational methods to solve problems using data and I am keen to learn and utilise the basic techniques of data science.*

4. Make an unordered list of at least 3 things you have learned in this module.

Answer:

- Data Science Methodologies
- Integrating R and Github
- Data types in R

5. Create a table showing the 1st 5 stages of the John Rollins General Data Science Methodology and the quetions associate with them.

Answer:

Stage	Questions
Business/Problem Understanding	What problem/question are you trying to solve/answer?
Analytic Approach	How can you use data to answer this question?
Data Requirements	What data is needed to answer the question?
Data Collection	What is the source of the data? How will you collect and receive it?
Data Understanding	Is the data indicative of the problem to be solved?

6. Create an ordered list of at least 3 other Data Science methodologies.

Answer:

1. CRISP-DM (Cross-Industry Standard Process for Data Mining)
2. TDSP (Team Data Science Process)
3. OSEMN (Obtain. Scrub. Explore. Model. Interpret)

7. Provide and ordered list the weaknesses of the CRISP-DM methodology and how to address them

Answer:

1. Outdated
   - Add phases (if needed)
2. Documentation heavy
   - Document enough... but not too much
3. Not a project management approach
   - Define team roles
   - Combine with a team coordination process
4. Can ignore stakeholders
   - Ensure actionable insight
5. Slow start
   - Iterate quickly

8. Download and include a University of Warwick Logo in your document.

Answer:

9. Write and execute a code to randomly generate 10 numbers from a normal distribution with mean 12 and standard deviation 3.

```
rnorm(10, mean = 12, sd = 3)
```

Answer:

```
##  [1]  8.061447 10.675538 14.208037 11.196428 12.666556 15.562815 10.556907
##  [8] 10.686777 13.545766 11.724101
```

10. Write and execute a code to calculate the mean of 12 random numbers generated from the intervalof integes from 5 to 25.

```
a <- sample(5:25, 12)

mean(a)
```

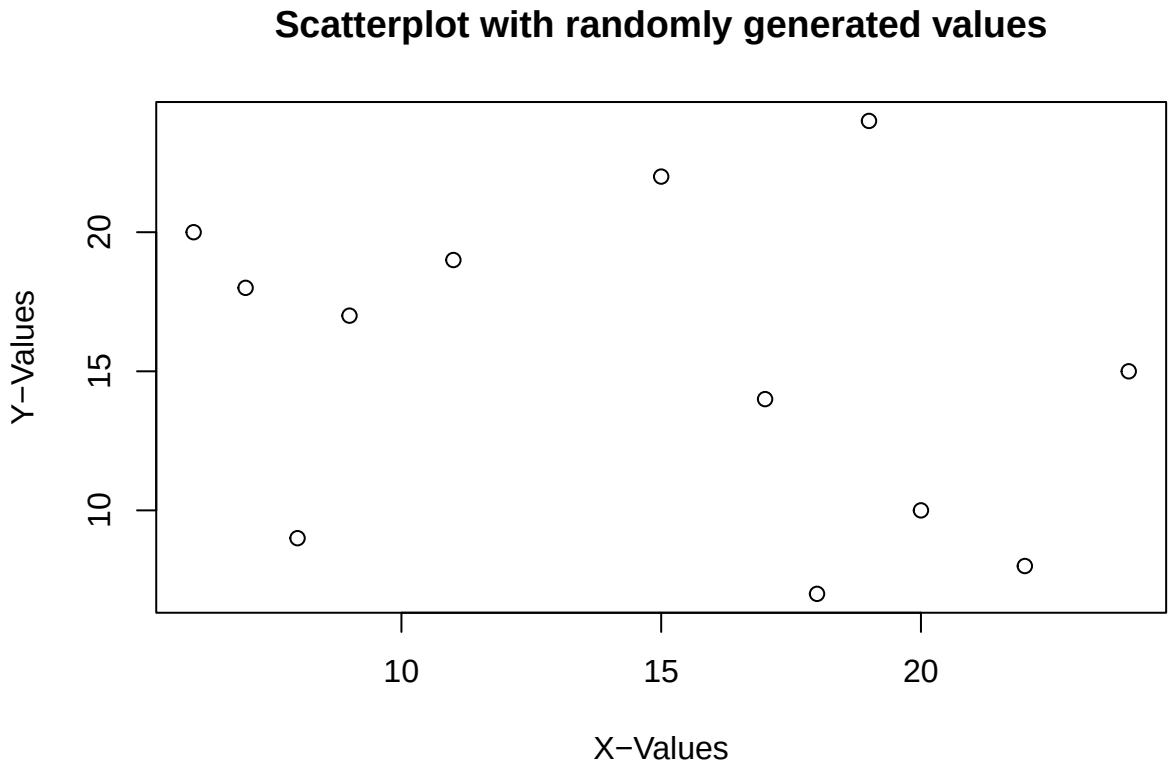
Answer:

```
## [1] 16.16667
```

11. Write and execute a code to create a scatter plot containing 12 points whose coordinates were randomly geenrated in the interval 5 to 25. Provide a title for the figure, label the x and y axis.

```
x <- sample(5:25, 12)
y <- sample(5:25, 12)
```

```r
plot(x, y,
     main = "Scatterplot with randomly generated values",
     xlab = "X-Values",
     ylab = "Y-Values")
```



Answer:

12. Write and execute a code to create a function to calculate standard error. Then, calculate the standard error of a set of 15 randomly generated numbers from a normal distribution with mean 12 and standard deviation 3.

```r
standard_error <- function(numbs){

  sd(numbs)/sqrt(length(numbs))

}

b <- rnorm(15, mean = 12, sd = 3)

standard_error(b)
```

Answer:

```
## [1] 0.7706472
```

```r
library(plotrix)
std.error(b)
```

```
## [1] 0.7706472
```

13. Given that BMI=kg/m2, where kg is the person's weight in kilograms and m2is height in meters squared, write and execute a code to create a function that calculates the BMI of who weighs 85kg and is 1.9m tall.

```r
BMI <- function(kg,m){

  kg/(m^2)

}

BMI(85,1.9)
```

Answer:

```
## [1] 23.54571
```