Post Graduate Program in Data Science

Module 4 – R for Data Science – Graded Assignment 2

1. Use the dataset named ‘Salary Data.xlsx’: It contains the data of 30 sample salary of women and men in an organization.
2. Form the null and alternate hypothesis.
3. Perform an appropriate hypothesis testing to identify if there is a difference between the salary of men and women.
4. What is the business conclusion based on the p-value.
5. The dataset named ‘Prime TV Shows Data set’ contains the details of tv shows in Amazon Prime OTT. Using this data:
6. Use a bar chart the plot the number of shows in each ‘Age of viewers’ category
7. Plot a histogram for the ‘IMDb Rating’.
8. Plot a bar chart for the top 3 shows based on ‘IMDb Rating’.
9. Plot a pie chart between the number of shows released and the years 2018,2019,2020.
10. Use the dataset named as ‘Question 3.xlsx’. With the help of **ggplot**:
11. Use the sheet named Stock price and plot a line chart between Date and Close columns. Give the title, axis labels and colors
12. Use the sheet named Sales vs Profit and plot a scatter plot between sales and profit columns. Use the region column to give colors in the scatter plots.
13. Use the sheet named Product Category Sales and plot a bar chart between Product Category and Sales. Add at least 5 layers to make the chart attractive.
14. Use the dataset named as advertising.csv.
15. Build a linear regression model with Sales as the target column
16. What is the R square-value
17. What would be the sales, if the TV spending is 150, The radio spending is 40, Newspaper spending is 70.

**Instructions:**

1. This is a graded assignment.
2. Once the assignment is done submit the R file in the LMS as a zip file.
3. You can add a word document for the explanation part in the zip. However this is not mandatory.