SESSION 8: Exploratory Data Analytics

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 Variables & Distributions in R.

3. Prerequisites

Not applicable.

4. Associated Data Files

Not applicable.

5. Problem Statement

1. Use the package -**RcmdrPlugin.IPSUR**.
   data(RcmdrTestDrive)

   and perform the below operations:

   a. Calculate the average salary by gender and smoking status.

ibrary(RcmdrPlugin.IPSUR)
data(RcmdrTestDrive)

head(RcmdrTestDrive)

attach(RcmdrTestDrive)

str(RcmdrTestDrive)

x <- tapply (salary, list(gender = gender), mean) by(salary, gender, mean, na.rm = true)

head(RcmdrTestDrive)

colnames(RcmdrTestDrive)

tapply(RcmdrTestDrive$salary,RcmdrTestDrive$gender,mean)

tapply(RcmdrTestDrive$salary,RcmdrTestDrive$smoking,mean)
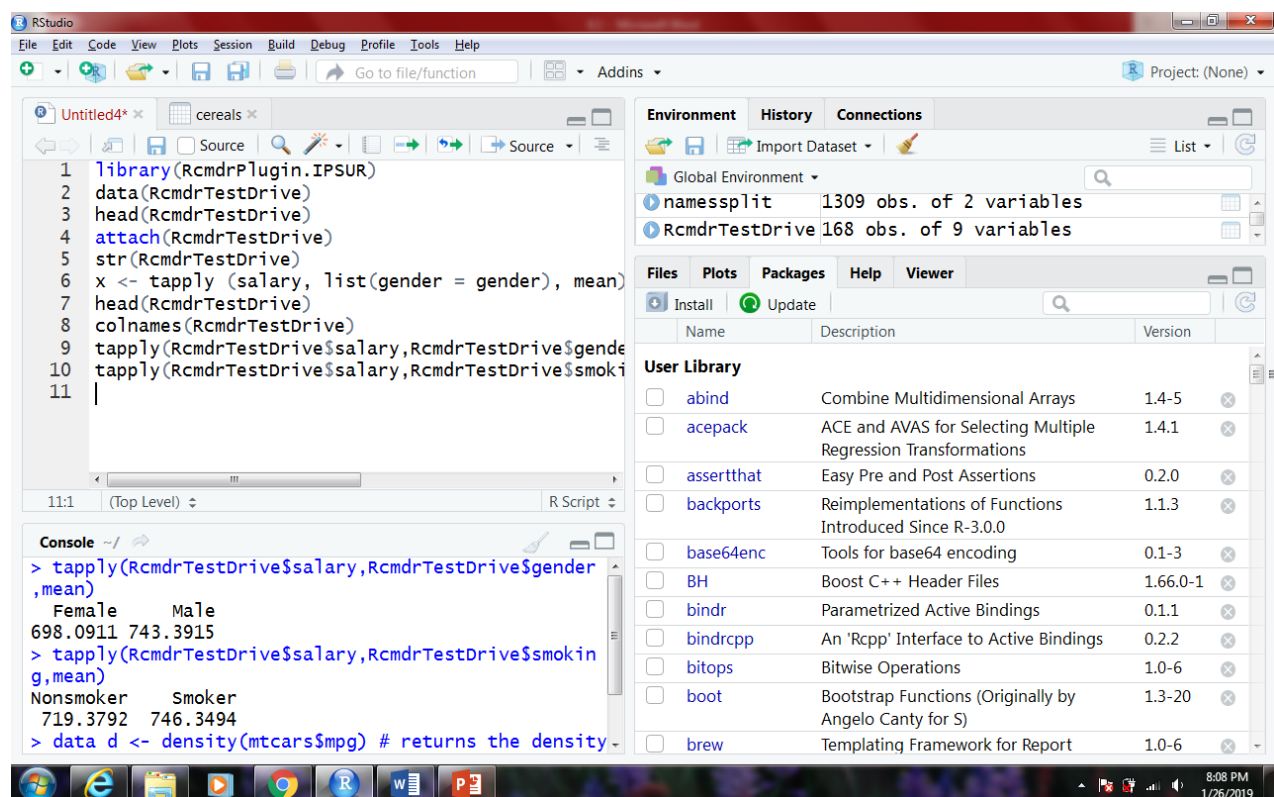
b. Which gender has the highest mean salary?
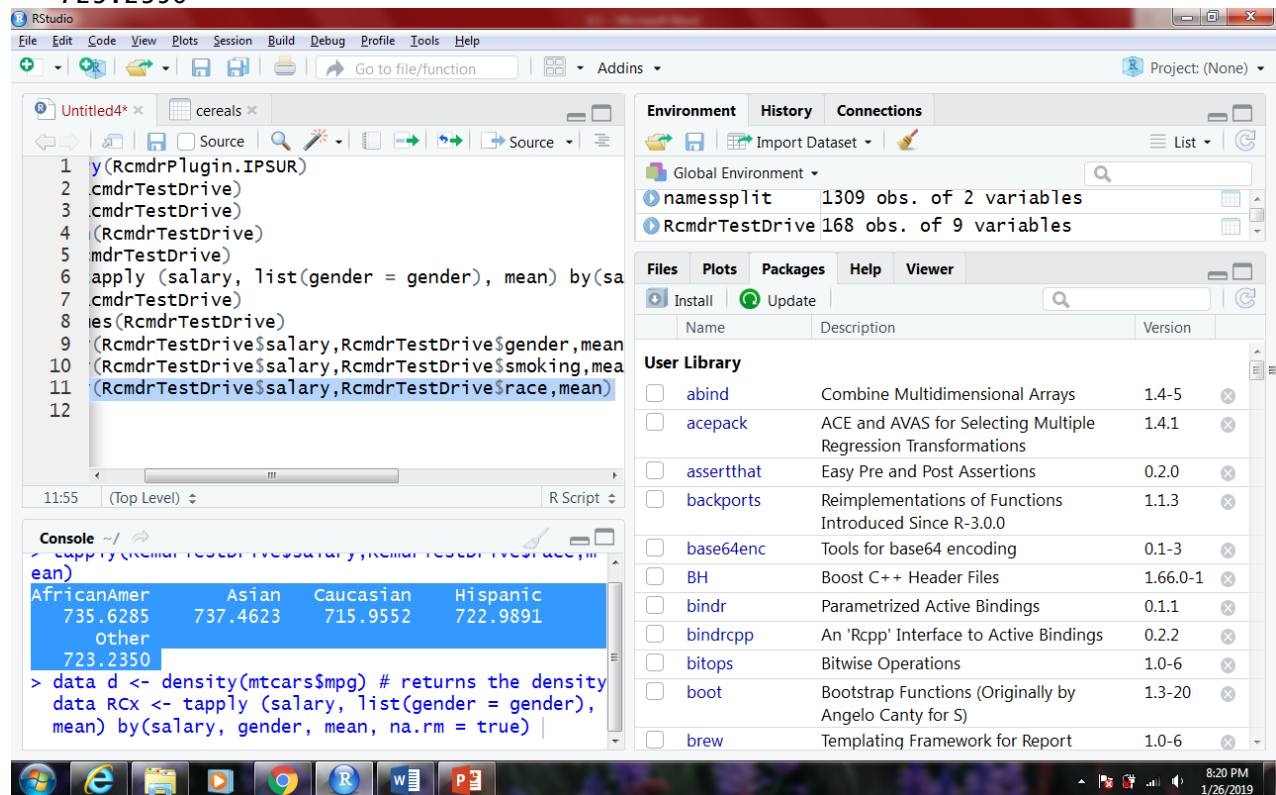
    C. `Female     Male`

    d. `698.0911  743.391`

Male has highest mean salary.

e. Report the highest mean salary.

Mean salary by race

tapply(RcmdrTestDrive$salary,RcmdrTestDrive$race,mean)

```
AfricanAmer        Asian    Caucasian     Hispanic
   735.6285     737.4623     715.9552     722.9891
      Other
   723.2350
```
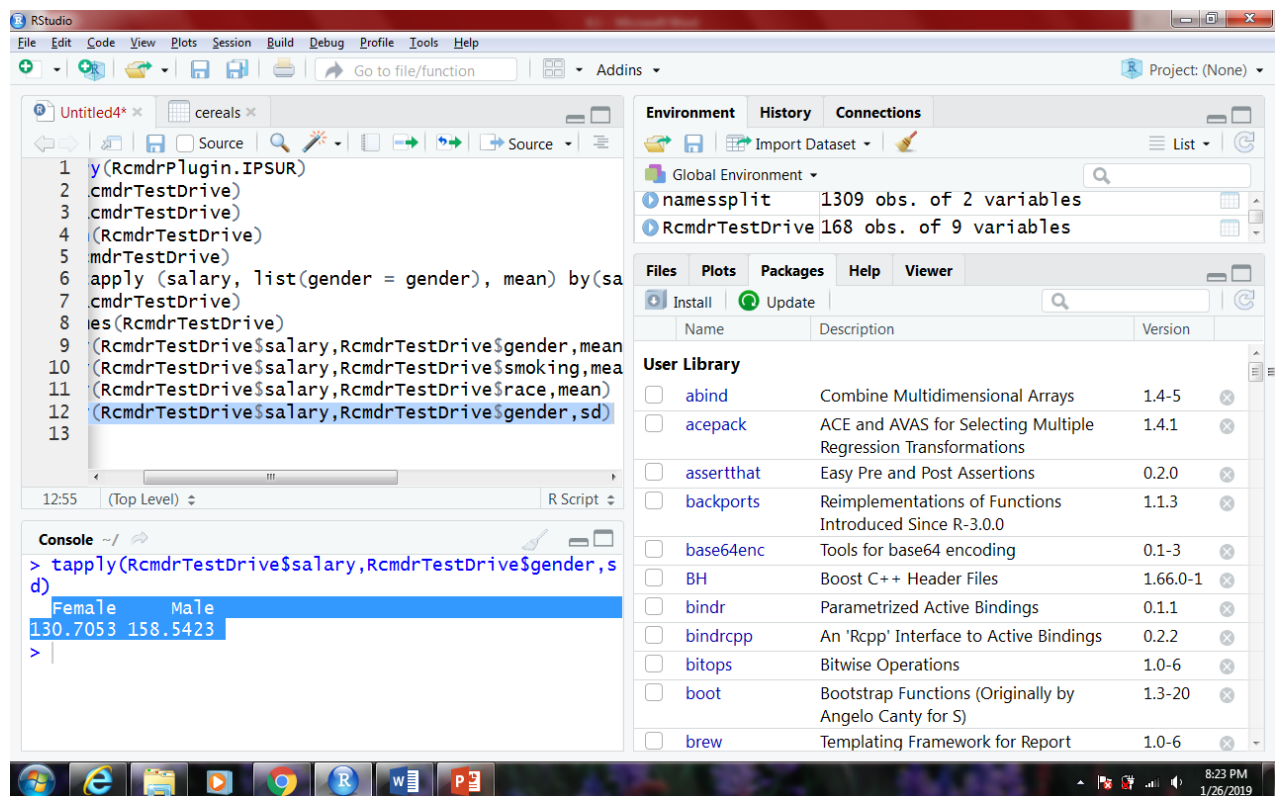

Highest Mean salary is reported by Males -743.391

f. Compare the spreads for the genders by calculating the standard deviation of salary by gender.

tapply(RcmdrTestDrive$salary,RcmdrTestDrive$gender,sd)

```
Female     Male
130.7053 158.5423
```

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.