



ACADGILD

SESSION 8: Exploratory Data Analytics

Assignment 2

Table of Contents

1.Introduction 3

2.Objective 3

3.Prerequisites 3

4.Associated Data Files 3

5.Problem Statement 3

6.Expected Output 3

7.Approximate Time to Complete Task 3

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

```
library(RcmdrPlugin.IPSUR)  
data(RcmdrTestDrive)
```

Perform the below operations: -

- a. Compute the measures of central tendency for salary and reduction which variable has highest center?

Measures of Central Tendency are Mean, Mode and Median

```
tapply(RcmdrTestDrive$salary,RcmdrTestDrive$reduction,mean)
```

Data Analytics

The screenshot shows the RStudio interface with the following components:

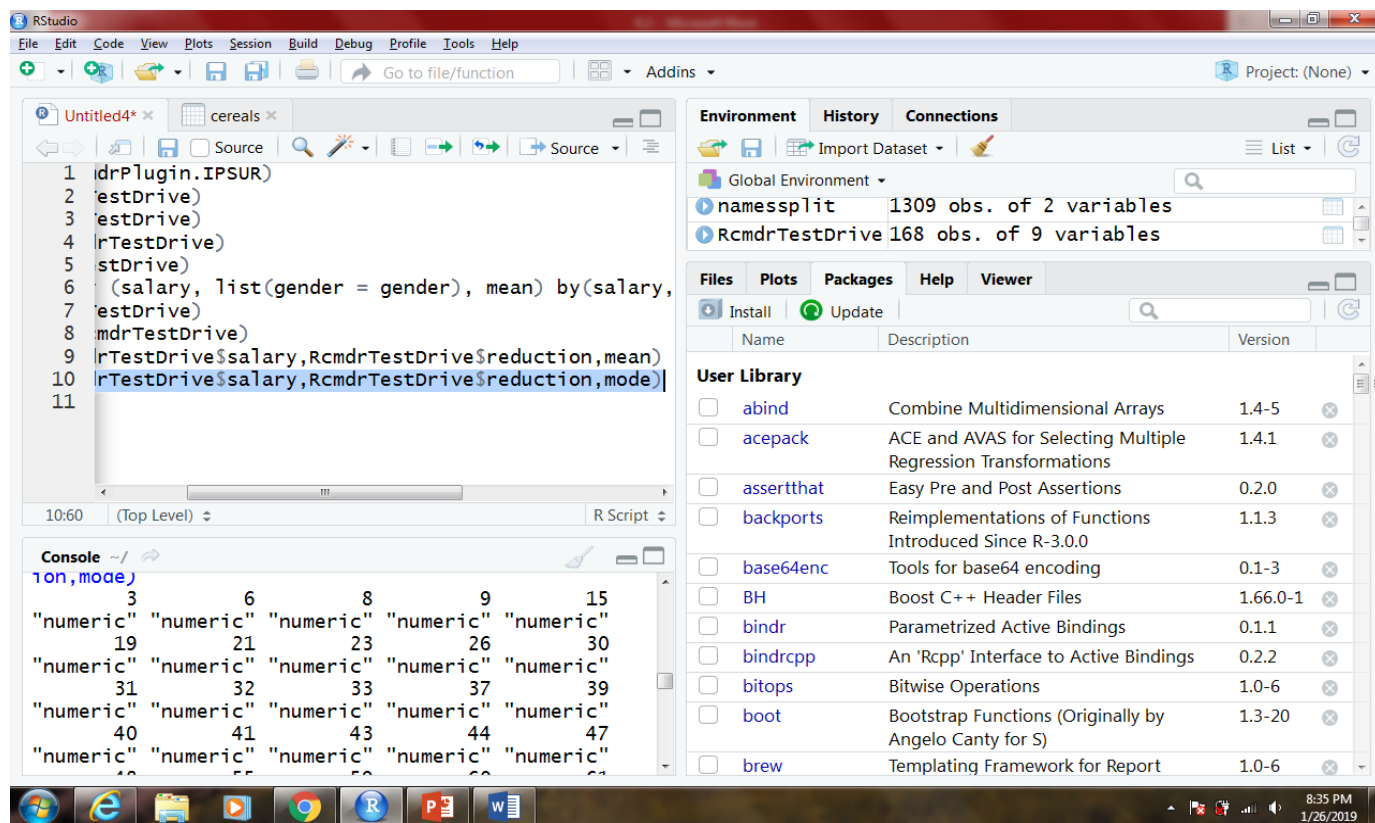
- Source Editor:** Contains an R script with the following code:

```
1 library(RcmdrPlugin.IPSUR)
2 data(RcmdrTestDrive)
3 head(RcmdrTestDrive)
4 attach(RcmdrTestDrive)
5 str(RcmdrTestDrive)
6 x <- tapply(salary, list(gender = gender), mean)
7 head(RcmdrTestDrive)
8 colnames(RcmdrTestDrive)
9 tapply(RcmdrTestDrive$salary, RcmdrTestDrive$reduc
10
```
- Console:** Shows the execution of the script, resulting in a matrix of mean salaries by gender and reduction level:

```
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$reduction, mean)
      3      6      8      9     15
606.1200 584.0800 808.6500 846.0750 651.7300
      19      21      23      26      30
550.2400 808.6300 619.2900 1027.3600 616.1600
      31      32      33      37      39
702.6700 690.3000 790.3300 888.0000 699.6300
      40      41      43      44      47
```
- Environment:** Shows the Global Environment with variables:
 - `namesplit`: 1309 obs. of 2 variables
 - `RcmdrTestDrive`: 168 obs. of 9 variables
- Packages:** Lists installed user libraries with their descriptions and versions.

`tapply(RcmdrTestDrive$salary, RcmdrTestDrive$reduction, mode)`

Data Analytics



`tapply(RcmdrTestDrive$salary,RcmdrTestDrive$reduction,median)`

Data Analytics

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for loading data and performing a tapply operation.
- Environment:** Shows the Global Environment with variables `namesplit` (1309 obs. of 2 variables) and `RcmdrTestDrive` (168 obs. of 9 variables).
- Console:** Displays the output of the `tapply` function, showing a matrix of values.
- Package List:** A list of installed and available packages.

```
1 library(RcmdrPlugin.IPSUR)
2 data(RcmdrTestDrive)
3 head(RcmdrTestDrive)
4 attach(RcmdrTestDrive)
5 str(RcmdrTestDrive)
6 x <- tapply(salary, list(gender = gender), mean)
7 head(RcmdrTestDrive)
8 colnames(RcmdrTestDrive)
9 tapply(RcmdrTestDrive$salary, RcmdrTestDrive$reduc
10 tapply(RcmdrTestDrive$salary, RcmdrTestDrive$reduc
11 tapply(RcmdrTestDrive$salary, RcmdrTestDrive$reduc
12
```

Console Output:

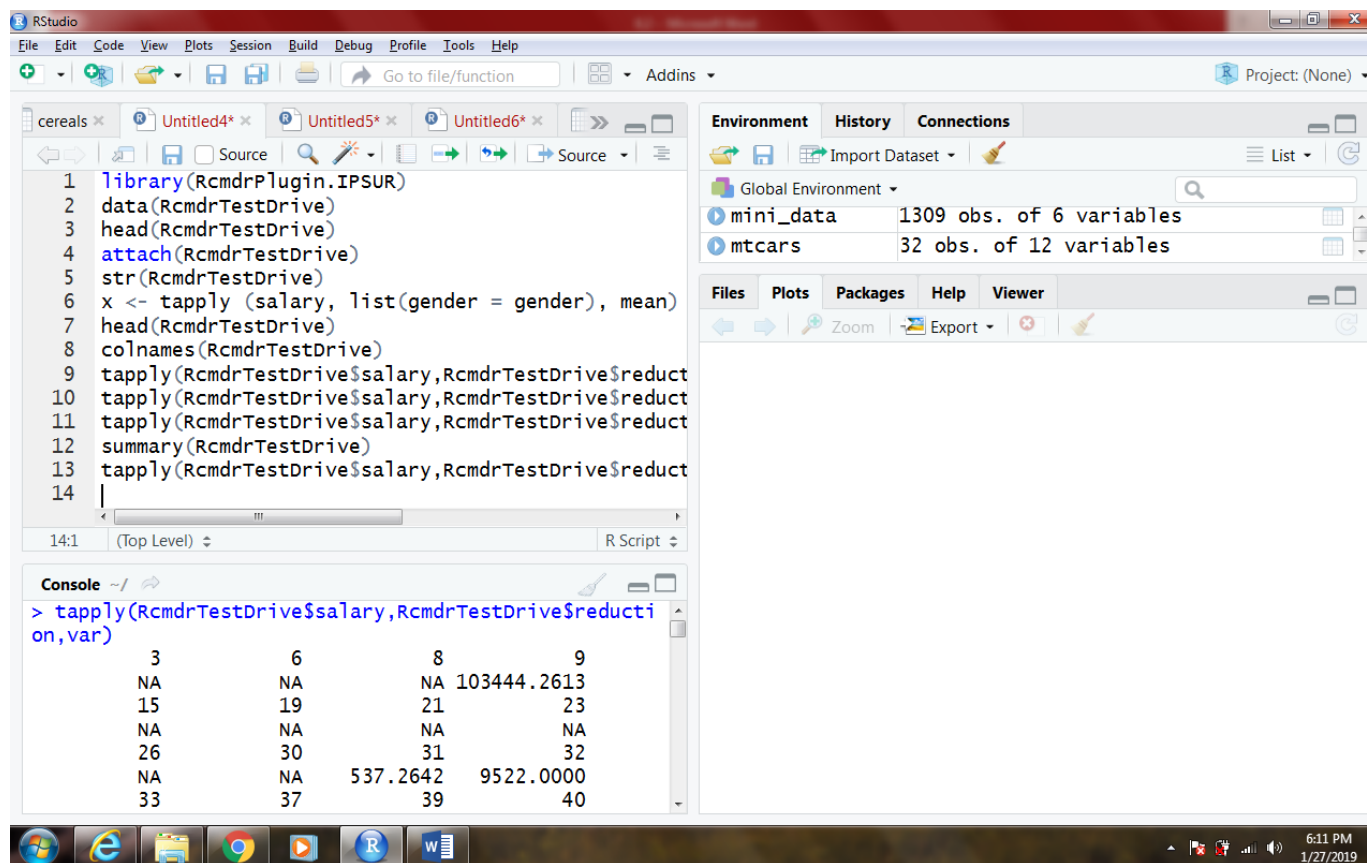
```
ton, median)
      3      6      8      9     15
606.120 584.080 808.650 846.075 651.730
 19    21    23    26    30
550.240 808.630 619.290 1027.360 616.160
 31    32    33    37    39
702.670 690.300 790.330 888.000 699.630
 40    41    43    44    47
703.790 809.260 692.090 849.250 790.820
```

Name	Description	Version
abind	Combine Multidimensional Arrays	1.4-5
acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1
assertthat	Easy Pre and Post Assertions	0.2.0
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.3
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.66.0-1
bindr	Parametrized Active Bindings	0.1.1
bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2.2
bitops	Bitwise Operations	1.0-6
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-20
brew	Templating Framework for Report	1.0-6

Variance:

```
tapply(RcmdrTestDrive$salary, RcmdrTestDrive$reduction, var)
```

Data Analytics



b. Which measure of center is more appropriate for before and after?

As data is continuous , mean is the appropriate measure of central tendency.

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.