

Session 8 – Exploratory

Data Analytics

Assignment - 1



Contents

[Introduction](#page3) [2](#page3)



[Objective](#page3) [2](#page3)



[Prerequisites](#page3) [2](#page3)



[Associated Data Files](#page3) [2](#page3)



[Problem Statement](#page3) [2](#page3)



[Expected Output](#page3) [2](#page3)



|  |  |
| --- | --- |
| Copyrights© 2017, AcadGild. All Rights Reserved | 1 |
|  |  |



**Introduction**



This assignment will help you to understand the key concepts learnt in this session.

**Objective**



This assignment will test your skills on Variables and Distribution in R.

**Prerequisites**



Not Applicable

**Associated Data Files**



Not Applicable

**Problem Statement**



1. Use the package RcmdrPlugin.IPSUR.

data(RcmdrTestDrive)

and perform the below operations:

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

install.packages("RcmdrPlugin.IPSUR")

library(RcmdrPlugin.IPSUR)

head(RcmdrTestDrive)

mean(RcmdrTestDrive$salary)

install.packages("plyr")

library(plyr)

library(reshape2)

library(plyr)

library(ggplot2)

install.packages("RcmdrPlugin.IPSUR")

library(RcmdrPlugin.IPSUR)

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

#of salary

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

#of smoking status

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

1. Which gender has the highest mean salary?

# genders mean salary respectively

#Female Male

#698.0911 743.3915

#so its the gender male which is highest

1. Report the highest mean salary.

#if we are considering the mean of salary then

mean(RcmdrTestDrive$salary)

#724.5164

#its the mean of salary

1. 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

**Expected Output**



Not Applicable

|  |  |
| --- | --- |
| Copyrights© 2017, AcadGild. All Rights Reserved | 2 |
|  |  |