**Author: Neal Sherman**

**Project: Chapter 2 Answer Key for ‘Data Science Using Python and R’**

**Working With the Data**

**For the following exercises, work with the bank\_marketing\_training data set. Use R to solve each problem.**

**11) Download the program and open the compiler. What is contained in the bottom‐right window? The top‐left (for R)?**

The top left window of RStudio contains the script, the bottom left contains the console, the bottom right contains plots, and the top right contains the environment.

**12) Type a comment stating that you are working on Chapter 2 exercises.**

> #This comment states that i am working on Chapter 2 Exercises

**13) Locate the “Run” button and note whether there is a keyboard shortcut.**

The keyboard shortcut for “Run” is Ctrl+Enter.

**14) Execute the comment from the previous exercise. What is the output? Explain your answer.**

> #This comment states that i am working on Chapter 2 Exercises

The output is the comment itself

**15) Import the following packages: For R, import the ggplot2 package. Make sure you both install and open the package.**

> #Install the ggplot2 package

> install.packages("ggplot2")

Error in install.packages : Updating loaded packages

Restarting R session...

> install.packages("ggplot2")

Installing package into ‘C:/Users/Neal/Documents/R/win-library/3.5’

(as ‘lib’ is unspecified)

also installing the dependency ‘isoband’

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/isoband\_0.2.1.zip'

Content type 'application/zip' length 3303519 bytes (3.2 MB)

downloaded 3.2 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/ggplot2\_3.3.0.zip'

Content type 'application/zip' length 4007170 bytes (3.8 MB)

downloaded 3.8 MB

package ‘isoband’ successfully unpacked and MD5 sums checked

package ‘ggplot2’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Neal\AppData\Local\Temp\RtmpK6qk5P\downloaded\_packages

> #This comment states that i am working on Chapter 2 Exercises

> #Import the ggplot2 package

> library(ggplot2)

**16) Import the bank\_marketing\_training data set and name it bank\_train.**

> #Rename bank\_marketing\_training as bank\_train

> bank\_train <- bank\_marketing\_training

**17) Create a contingency table of the variables response and previous\_outcome from the bank\_train data set. Do not save the output from the code.**

> table(bank\_train$response, bank\_train$previous\_outcome)

failure nonexistent success

no 2390 21176 320

yes 385 2034 569

**18) Rerun the code from the previous exercise, this time saving the output as t1 (for R code).**

> #Save the previously created table and rename it

> t1 <- table(bank\_train$response, bank\_train$previous\_outcome)

**19) After saving the output in the previous exercise, display the output using the name of the saved output.**

> #Print t1

> t1

failure nonexistent success

no 2390 21176 320

yes 385 2034 569

**20) Save the contingency table under a different name. This time, use your last name and favorite number as the name; for example, larose42.**

> #Rename t1

> sherman24 <- t1

**21) Save the first nine records of the bank\_train data set as their own data frame.**

> #Create a dataframe of rows 1-9

> df1 <- bank\_train[1:9,]

**22) Save the age and marital records of the bank\_train data set as their own data frame.**

> #Create a dataframe of columns 1 and 3

> df2 <- bank\_train[,c(1,3)]

**23) Save the first three records of the age and marital variables as their own data frame.**

> #Create a dataframe of columns 1 and 3 of the first 3 rows

> df3 <- bank\_train[1:3,c(1,3)]

**HANDS‐ON ANALYSIS**

**24) Import the adult\_ch3\_training data set using the “Heading: Yes” setting. Rename the data set adult once it is imported.**

> #rename adult\_ch3\_training as adult

> adult <- adult\_ch3\_training

**25) Write a comment explaining the change in the data set name.**

> #the data set name was changed from adult\_ch3\_training to adult

**26) Import the following packages: For R, import the rpart package. Make sure you both install and open the package.**

> #install rpart package

> install.packages("rpart")

Installing package into ‘C:/Users/Neal/Documents/R/win-library/3.5’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/rpart\_4.1-15.zip'

Content type 'application/zip' length 767500 bytes (749 KB)

downloaded 749 KB

package ‘rpart’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Neal\AppData\Local\Temp\Rtmp8uIgSc\downloaded\_packages

> #import rpart package

> library(rpart)

**27) Create a contingency table of workclass and sex. Save the output as table01.**

> #create and store a contingency table of workclass vs. sex

> table01 <- table(adult$workclass, adult$sex)

> #print table01

> table01

Female Male

? 377 452

Federal-gov 149 305

Local-gov 377 592

Never-worked 1 4

Private 3574 6707

Self-emp-inc 54 444

Self-emp-not-inc 178 992

State-gov 201 385

Without-pay 1 4

**28) Create a contingency table of sex and marital status. Save the output as table02.**

> #create and store a contingency table of sex vs. marital status

> table02 <- table(adult$sex, adult$marital.status)

> #print table02

> table02

Divorced Married-AF-spouse Married-civ-spouse

Female 1219 7 761

Male 795 4 6010

Married-spouse-absent Never-married Separated Widowed

Female 95 2160 290 380

Male 104 2717 182 73

**29) Display the sex and workclass values of the person in the first record. What cell of table01 do they belong to? How many other records in the data set have the same sex and workclass values?**

> #print the 2nd and 6th column of the 1st row

> adult[1,c(2,6)]

workclass sex

1 Self-emp-not-inc Male

> #print the value in the 7th row and 2nd column of table01

> table01[7,2]

[1] 992

The person in the first record belongs to table01[7,2]. 992 records have the same sex and workclass values (so 991 other values).

**30) Display the sex and marital status values of the people in records 6–10. Which cells of table02 do they belong to? How many other records in the data set have the same combinations of sex and marital status values?**

> #print the 4th and 6th columns of the 6th through 10th rows

> adult[6:10,c(4,6)]

marital.status sex

6 Married-civ-spouse Male

7 Married-civ-spouse Male

8 Married-civ-spouse Male

9 Married-civ-spouse Male

10 Divorced Male

> #print the values in the 2nd row and the 3rd and 1st columns of table02, respectively

> table02[2,3]

[1] 6010

> table02[2,1]

[1] 795

The people in records 6-9 belong to table02[2,3] (shared by 6010 values), and the person in record 10 belongs to table02[2,1] (shared by 795 values).

**31) Create a new data set that has only records whose marital status is “Married‐civ‐spouse” and name the data set adultMarried.**

> #store the subset of married adults in adultMarried

> adultMarried <- adult[adult$marital.status=="Married-civ-spouse",]

**32) Recreate the contingency table of sex and workclass using the adultMarried data set. What differences do you notice between the sexes?**

> #create, store and print a contingency table of workclass vs. sex for married adults

> table03 <- table(adultMarried$workclass, adultMarried$sex)

> table03

Female Male

? 67 224

Federal-gov 20 203

Local-gov 68 411

Never-worked 1 0

Private 491 3883

Self-emp-inc 24 347

Self-emp-not-inc 57 703

State-gov 33 237

Without-pay 0 2

The major difference is the number of values for each sex (761 females vs. 6010 males), but a greater proportion of married females work in local or state government while a greater proportion married males were self-employed.

**33) Create a new data set that has only records whose age value is greater than 40. Name the new data set adultOver40.**

> #store the subset of adults over 40 in adultOver40

> adultOver40 <- adult[adult$age>40,]

**34) Recreate the contingency table of sex and marital status using the adultOver40 data set. What differences do you notice?**

> #create, store and print a contingency table of sex vs. marital status for adults over 40

> table04 <- table(adultOver40$sex, adultOver40$marital.status)

> table04

Divorced Married-AF-spouse Married-civ-spouse

Female 720 0 305

Male 448 0 3383

Married-spouse-absent Never-married Separated Widowed

Female 38 256 133 354

Male 47 315 74 71

Despite having more total values (4338 for males vs. 1806 for females), men over 40 had fewer values for divorced, separated, or widowed, and a significantly greater proportion of values for married.

**Larose, Chantal D.. Data Science Using Python and R (Wiley Series on Methods and Applications in Data Mining) (pp. 26-27). Wiley. Kindle Edition.**