Practice R Coding Assignment

Devin Cline

2023-02-18

This assignment is worth 100 points. There are 10 questions and each question is worth 10 points each. If a question has multiple parts, the points for each part is provided. The assignment is due Tuesday February 21 by 11:59pm CST. If the assignment is turned in late, points will be deducted as explained in the class syllabus.

Make sure you answer any concept questions asked. There are some questions that do not require you to provide code. Text responses should be placed outside the code chunks. Your code should support the answers you provide for any concept questions.

Note: if you use the View() function, comment out the function before you knit the file.

Submission Instructions: Once you have completed the assignment, knit the RMarkdown document to a pdf. Save the pdf as 'Practice_R_Assignment_your_last_name.' Upload the pdf to the Practice R Coding Assignment in Module 2 on Canvas.

Part 1

Data Set Scenario: Orion is a fictitious retail company. The data you will be analyzing will be employee information from this company.

1. The first task is to import the Excel spreadsheet into R. In the data folder of this project, there is an Excel sheet entitled 'employee_payroll' that contains information about employees at Orion.

Import the data into R and create a data frame named employee_payroll.

```
library(readxl)
employee_payroll <- read_excel("/cloud/project/data/employee_payroll.xlsx")
```

2. Answer some basic questions about the data using the code chunks below. Type any responses outside the code chunks. The code should support your response.

a. How many rows and columns are in the data? 2.5 points There are 424 rows and 8 columns

```
#View(employee_payroll)
str(employee_payroll)
```

```
## tibble [424 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Employee_ID       : num [1:424] 120101 120102 120103 120104 120105 ...
##  $ Employee_Gender   : chr [1:424] "M" "M" "M" "F" ...
##  $ Salary            : num [1:424] 163040 108255 87975 46230 27110 ...
##  $ Birth_Date        : num [1:424] 7535 4971 -2535 -600 6929 ...
##  $ Employee_Hire_Date: num [1:424] 17348 12205 6575 9132 15826 ...
##  $ Employee_Term_Date: num [1:424] NA NA NA NA NA NA NA NA NA NA ...
##  $ Marital_Status    : chr [1:424] "S" "O" "M" "M" ...
##  $ Dependents        : num [1:424] 0 2 1 1 0 2 2 0 3 1 ...
```

b. How many observations are in the data? 2.5 points There are 424 observations

```
str(employee_payroll)
```

```
## tibble [424 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Employee_ID       : num [1:424] 120101 120102 120103 120104 120105 ...
##  $ Employee_Gender   : chr [1:424] "M" "M" "M" "F" ...
##  $ Salary            : num [1:424] 163040 108255 87975 46230 27110 ...
##  $ Birth_Date        : num [1:424] 7535 4971 -2535 -600 6929 ...
##  $ Employee_Hire_Date: num [1:424] 17348 12205 6575 9132 15826 ...
##  $ Employee_Term_Date: num [1:424] NA NA NA NA NA NA NA NA NA NA ...
##  $ Marital_Status    : chr [1:424] "S" "O" "M" "M" ...
##  $ Dependents        : num [1:424] 0 2 1 1 0 2 2 0 3 1 ...
```

c. Are any of the variables character? If yes, list them. 2.5 points Employee_Gender and Marital_Status

```
str(employee_payroll)
```

```
## tibble [424 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Employee_ID       : num [1:424] 120101 120102 120103 120104 120105 ...
##  $ Employee_Gender   : chr [1:424] "M" "M" "M" "F" ...
##  $ Salary            : num [1:424] 163040 108255 87975 46230 27110 ...
##  $ Birth_Date        : num [1:424] 7535 4971 -2535 -600 6929 ...
##  $ Employee_Hire_Date: num [1:424] 17348 12205 6575 9132 15826 ...
##  $ Employee_Term_Date: num [1:424] NA NA NA NA NA NA NA NA NA NA ...
##  $ Marital_Status    : chr [1:424] "S" "O" "M" "M" ...
##  $ Dependents        : num [1:424] 0 2 1 1 0 2 2 0 3 1 ...
```

d. Examine the data frame you imported into R. What looks ‘odd’ to you about the column Employee hire and birth date? 2.5 points

You do not have to do any work outside this assignment. Just simply state, in your own words, what you think is ‘odd’ about the date variables.

They are listed as integer values (presumably because dates are stored as integers in excel)

3. Are there any variables that have missing values? If yes, list them.

Employee_Term_Date

```
colSums(is.na(employee_payroll))
```

```
##        Employee_ID    Employee_Gender             Salary         Birth_Date
##                  0                  0                  0                  0
## Employee_Hire_Date Employee_Term_Date     Marital_Status         Dependents
##                  0                308                  0                  0
```

4. Calculate the mean and standard deviation of Salary.

```
mean_salary <- mean(employee_payroll$Salary)
mean_salary
```

```
## [1] 38041.51
```

```
sd_salary <- sd(employee_payroll$Salary)
sd_salary
```

```
## [1] 31741.14
```

5. Find the first and third quartile of the employee salaries. Interpret both the first and third quartile in the context of the data.

The first quartile is 26742.50 and the third is 36386.25, meaning 25% are paid less than 26742.50 and 75% are paid less than 36386.25.

```
quantile(employee_payroll$Salary)
```

```
##        0%       25%       50%       75%      100% 
##  22710.00  26742.50  28685.00  36386.25 433800.00
```

6. Are there any outlier values in the employee salaries? Yes or No. Use the code chunk below to support your Yes or No answer. Use the same methodology provided in class to detect outliers.

Yes

```
salary_q1 <- quantile(employee_payroll$Salary, 0.25)
salary_q3 <- quantile(employee_payroll$Salary, 0.75)
salary_IQR <- IQR(employee_payroll$Salary)

salary_lower_thresh <- salary_q1 - 1.5*salary_IQR
salary_upper_thresh <- salary_q3 + 1.5*salary_IQR

print(which(employee_payroll$Salary < salary_lower_thresh))
```

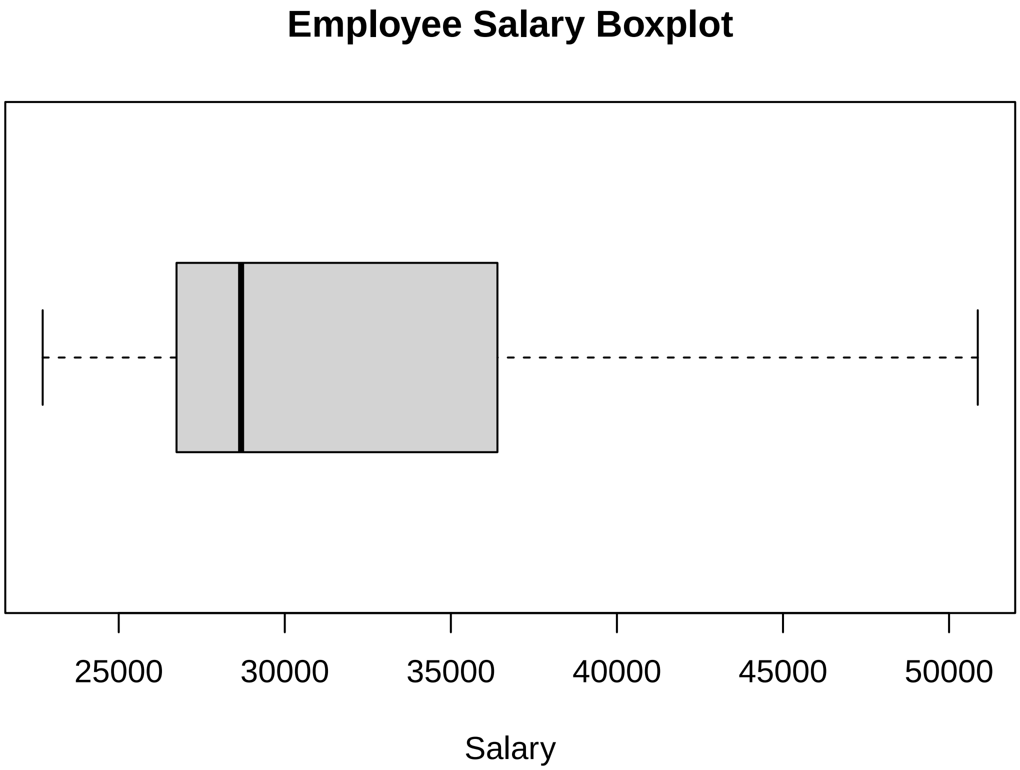
```
## integer(0)
```

```
print(which(employee_payroll$Salary > salary_upper_thresh))
```

```
##  [1]   1   2   3  99 100 101 102 105 108 109 124 125 126 128 130 131 135 137 139
## [20] 161 162 163 165 167 170 175 186 187 188 202 207 211 216 217 231 233 241 242
## [39] 243 245 249 251 253 255 256 261 264 265 417 418 419 420 421 424
```

7. Create a boxplot of the employee's salaries. You can create either a horizontal or vertical boxplot. Label the appropriate axis of the boxplot. Provide a descriptive title for the boxplot (need to provide something more than 'boxplot'). The boxplot should not contain any outlier values.

```
boxplot(employee_payroll$Salary, main="Employee Salary Boxplot", xlab="Salary", outline=FALSE, horizonta
```

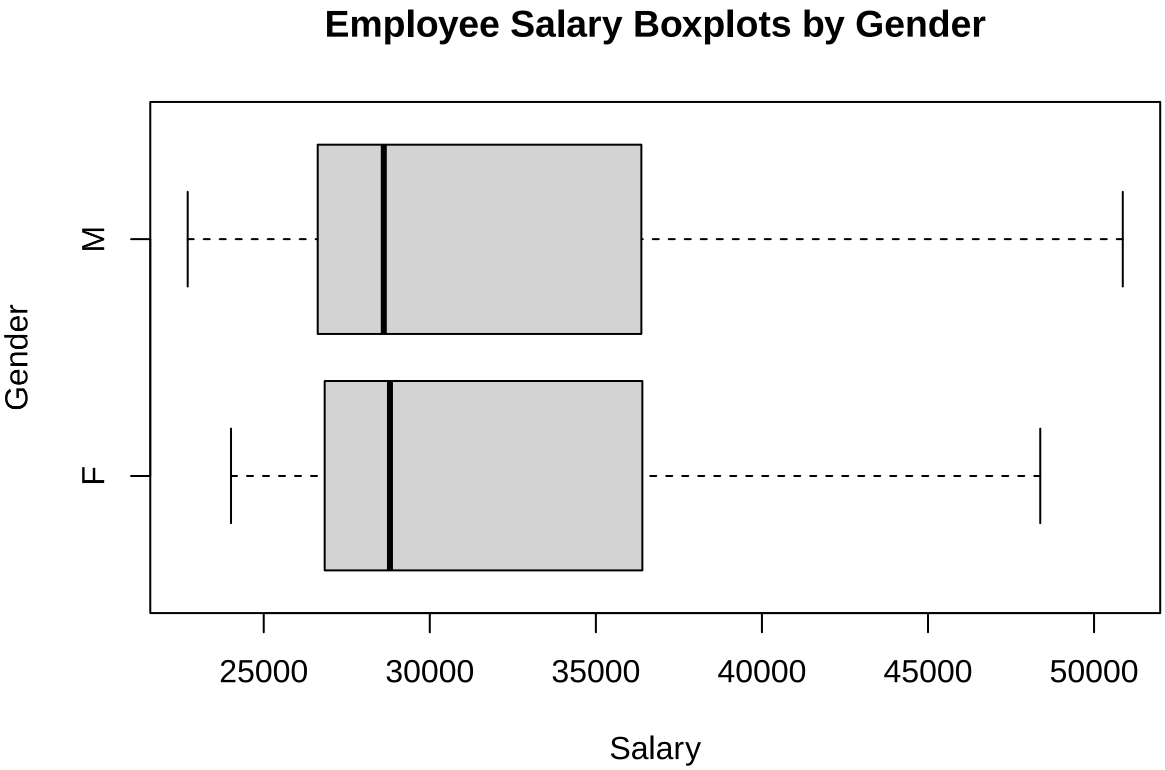


8. Using the boxplot you created in #7, what is the median salary (estimate from the boxplot-you do not have to run any code). Interpret the median salary in the context of the data.

I would estimate the median salary to be 28,000, meaning 50% earn less and 50% earn more than that amount.

9a. Construct a grouped boxplot (group by employee gender) for salary. You may construct a horizontal or vertical boxplot. Provide a descriptive title for the boxplot (need to provide something more than 'boxplot'). Label each axis of the boxplot appropriately. The boxplot should not contain any outlier values. 5 points

```
boxplot(employee_payroll$Salary ~ employee_payroll$Employee_Gender, main="Employee Salary Boxplots by Ge
```



9b. Examine the grouped boxplot in #9a. Answer the two questions below. 5 points

Is there one gender that is making significantly more than another?

Based on the boxplot, for the majority of workers I would say no. The median incomes are similar across genders and the bulk of the salaries are comparable across genders. However, the highest earning men do earn more than the highest earning women and the lowest earning men earn less than the lowest earning women.

Is there one gender that has particularly more variability in their salaries versus another?

Men have more variability in their salaries as seen in the 'whiskers.' There is a larger range for men than women.

Part 2

10. For this part of the assignment, you will be using the Chipolte data scenario that has been discussed in class. There is a data dictionary posted in Module 2 in Canvas to assist you with a description of the columns.

```
#import the data
chipotle_df <- read_excel("/cloud/project/data/Chipolte_clean_top1_labels.xlsx")
```

Scenario: You work as a data analyst for Chipolte. You report directly to the Director of Analytics. The Director has asked you to use the survey data collected to provide some insight about the participants interest in healthy options.

This is all the information your Director has provided. It is up to you to provide some descriptive statistics that will give Chipolte insight into the survey participants interests in healthy food options.

Task: Use some descriptive statistics techniques that you have learned in class to provide to your Director. You need to strictly use the descriptive statistics shown in class. There are three requirements:

1. Need to use, at a minimum, two variables in the Chipolte data frame.

2. Develop a code chunk that is well-commented that describes what descriptive statistics that are being computed.You may use visuals as well.
3. Write a short but concise paragraph that uses the descriptive statistics to tell the Director of Analytics something useful about the survey participants interest in healthy food options. The Director of Analytics will be sharing your insights to the Chief Marketing Officer so ensure that your paragraph is written in non-technical terms. State what the descriptive statistics tells the Director of Marketing in the context of the problem using no analytical or statistical terms.

Having healthy options is an important factor for customers in choosing a restaurant and Chipotle is largely perceived as good at providing these options. The most important factor for a customer deciding on a restaurant by far is taste with an average importance of 4.93 on a scale of 1-5. Price is the second most important factor, at 4.67 and having healthy options is the third most important factor of six, coming in at 4.53. In terms of performane, Chipotle's average rating for providing healthy menu options is 4.42, which is between 'good' and 'very good.' After visualizing the data, we can see that the majority of responses fall between 4 and 5. To reiterate, having healthy options is important for customers and Chipotle is seen as doing a good job at providing them.

(note: the boxplot for ranking chipotle as having healthy options includes a value of 6, indicating an error in the data that I would fix before doing a more formal analysis. I don't believe significant data cleaning was an intended part of the question as written and because we have not covered it much in class.)

```
# identify restaurants with sufficient and clean data to compare
table(chipotle_df$top1)
```

```
## 
##   Chick-Fil-A      Chipotle El Pollo Loco      In-n-out        Panera 
##            33            42            57            76            91 
##        Subway 
##            40 
```

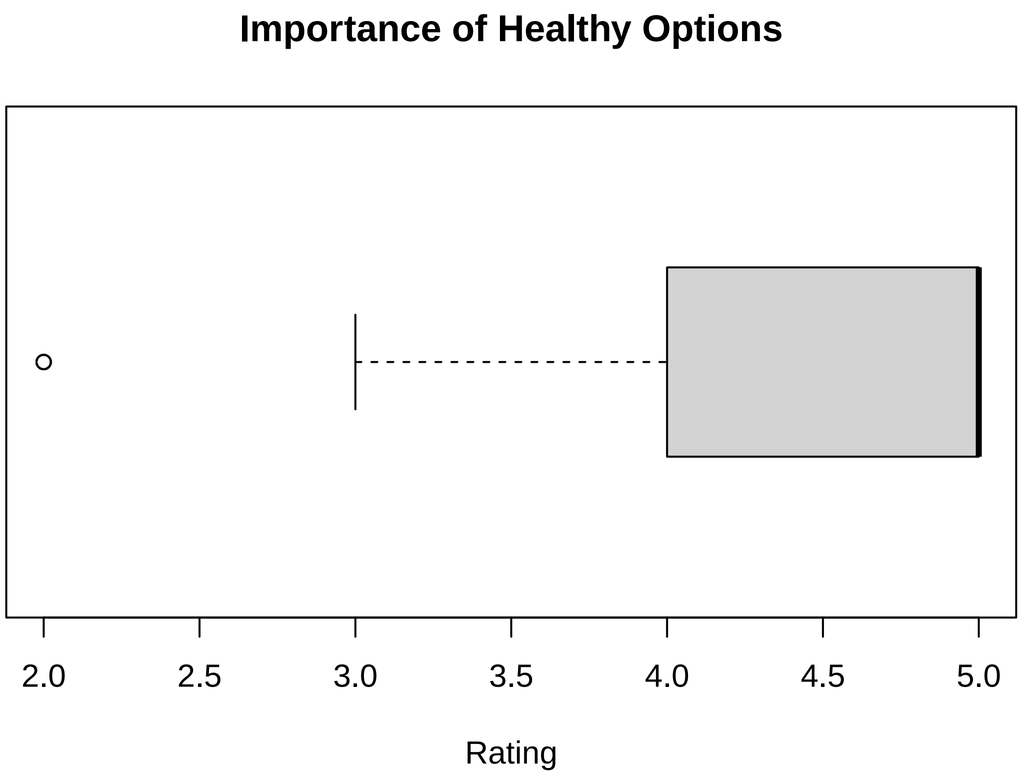
```
# FIRST TOPIC: How important is having healthy options compared to other factors

# create dataframe of important factors for deciding on restaurant
important_factors_df <- data.frame(chipotle_df$importantconvenience, chipotle_df$importantvariety, chipo

# find means of importance of different factors on visiting a restaurant
important_factors_means <- colMeans(important_factors_df, na.rm=TRUE)
important_factors_means
```

```
## chipotle_df.importantconvenience     chipotle_df.importantvariety 
##                         4.514019                         4.310127 
##       chipotle_df.importantprice     chipotle_df.importanthealthy 
##                         4.665615                         4.534591 
##       chipotle_df.importanttaste    chipotle_df.importantambience 
##                         4.926984                         4.242138 
```

```
# create boxplot of importance of healthy options
boxplot(chipotle_df$importanthealthy, horizontal=TRUE, main="Importance of Healthy Options", xlab="Rati
```

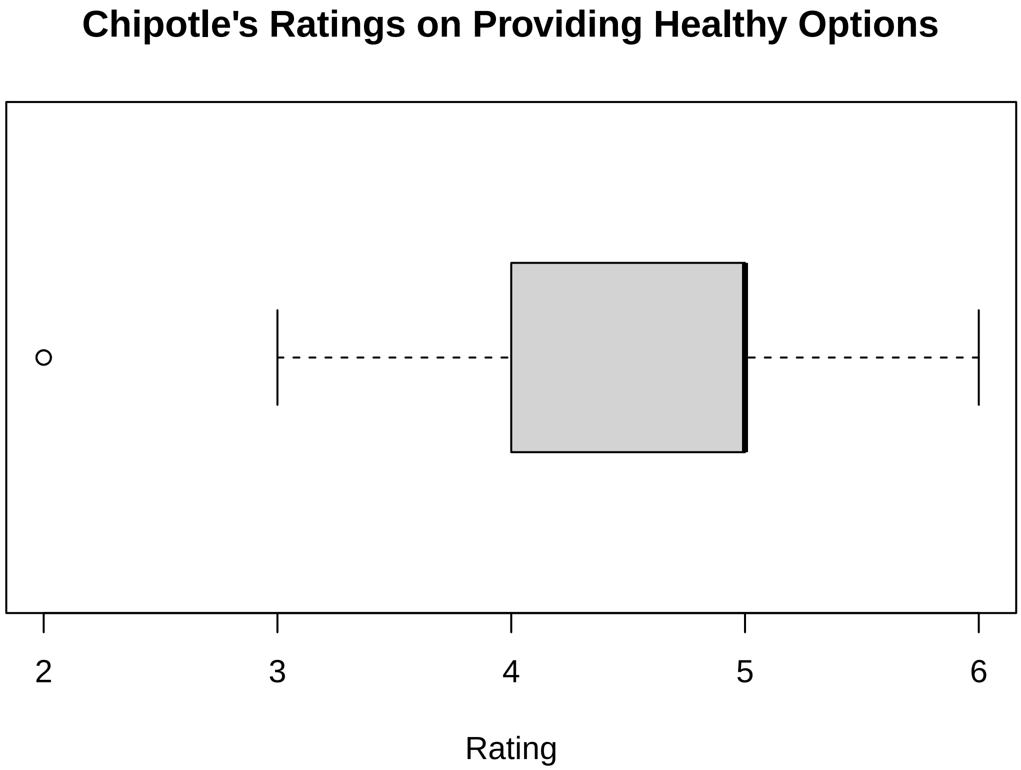


```
# SECOND TOPIC: How well is Chipotle scored in having healthy options

# find mean of rating on Chipotle having healthy options
mean(chipotle_df$chipotlehealthy, na.rm=TRUE)
```

```
## [1] 4.42492
```

```
# create boxplot of chipotle's rating on providing healthy options
boxplot(chipotle_df$chipotlehealthy, horizontal=TRUE, main="Chipotle's Ratings on Providing Healthy Opti
```


Chipotle's Ratings on Providing Healthy Options
2
3
4
5
6
Rating