Midterm 2

Parker DeBruyne - V00837207

21/03/2022

Question 1: Use the built-in data set UCBAdmissions to answer this question.

```
head(UCBAdmissions)
```

```
## , , Dept = A
##
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
##
## , , Dept = B
##
##           Gender
## Admit      Male Female
##   Admitted  353     17
##   Rejected  207      8
##
## , , Dept = C
##
##           Gender
## Admit      Male Female
##   Admitted  120    202
##   Rejected  205    391
##
## , , Dept = D
##
##           Gender
## Admit      Male Female
##   Admitted  138    131
##   Rejected  279    244
##
## , , Dept = E
##
##           Gender
## Admit      Male Female
##   Admitted   53     94
##   Rejected  138    299
##
## , , Dept = F
##
##           Gender
## Admit      Male Female
```

```
##   Admitted   22     24
##   Rejected  351    317
```

(a) Create a single table called status_dept_totals which summarizes the total number of applicants who are accepted and rejected for each department.

(b) Print out the status_dept_totals table.

(c) Create a grouped bar plot which displays the information from the status_dept_totals table. Your plot should include the following:

- a main title
- titles for the x-axis and y-axis
- colours to help differentiate the bars
- a legend to identify what each colour represents

(d) Create and print out a vector called percent_dept which contains the percent of applicants who applied to each department (rounded to 2 decimal places). Show any additional code needed to create this vector.

(e) Create a pie chart displaying the information in the percent_dept vector. Your graph should include:

- a main title
- labels for each wedge
- a different colour for each wedge
- the percentages displaying next to each wedge.

```
# (a) Answer below:


# (b) Answer below:


# (c) Answer below:


# (d) Answer below:


# (e) Answer below:
```

Question 2: Use the built-in data set LakeHuron to answer this question.

(a) What is the variable being measured in the data set?

(b) What is the most appropriate type of graph to visualize the distribution of this variable?

(c) Graph the distribution of the variable (using the type of graph that you identified in part (c)). Your graph should include:

- a main title.
- x-axis title.
- scales on the x and y-axis which fully extend from atleast the min value to at least the max value.

(d) Are there any overall trends or seasonal variations that you can see from the graph?

```
# (a) Answer below:

# (b) Answer below:

# (c) Answer below:
```

```
# (d) Answer below:
```

Question 3: Use the built-in data set rock to answer this question.

(a) Create a histogram displaying the distribution of the area variable.

(b) Describe the shape of the distribution (that is, symmetric, left-skewed, right-skewed).

(c) What is an appropriate measure for the center of the area distribution?

(d) Compute the observed value of this statistic.

(e) What is an appropriate measure for the spread of the area distribution?

(f) Compute the observed value of this statistic.

```
# (a) Answer below:

# (b) Answer below:

# (c) Answer below:

# (d) Answer below:

# (e) Answer below:

# (f) Answer below:
```

Question 4: Download and save the new_NHL_data.csv data set that is posted in Brightspace and use it to answer Question 4 and 5.

(a) Read the data set into R and save it as nhl_data.

(b) Suppose you are looking to explore the relationship between the number of goals that a player scores (G) and the number of assists that a player gets (A). What direction do you expect the relationship to have, why?

(c) Create a graph which visualizes the relationship between these two variables. Set the number of assists as the explanatory variable.

(d) What issues do you see in the graph? Can you identify direction and/or form from this graph?

(e) Create a vector which contains the number of assists for players on the Vancouver Canucks (VAN) and call this vector van_assists.

(f) Create a vector which contains the number of goals for players on the Vancouver Canucks (VAN) and call this vector van_goals.

(g) Create a graph which visualizes the relationship between the number of assists and number of goals for players on the Vancouver Canucks. Your graph should include:

- a main title.
- a title for both the x-axis and the y-axis
- the scale should not be in scientific notation.

(h) Describe the direction and form of the relationship.

Ignore this question (i) Compute the correlation between the two variables and describe what this implies for the linearity and strength of the relationship.

```
# (a) Answer below:
```

```
# (b) Answer below:


# (c) Answer below:


# (d) Answer below:


# (e) Answer below:


# (f) Answer below:


# (g) Answer below:


# (h) Answer below:


# (i) Answer below:
```

Question 5: Consider again the new_NHL_data.csv data set that was used in Q4. For this question, we will be focussing on the column associated with the number of penatly minutes (PIM).

(a) Bootstrap 10000 sample third quartile values (Q3) for the number of penalty minutes in the NHL (PIM). Save your bootstrapped Q3's to a vector called boot_Q3.

(b) Plot the sampling distribution for the Q3 statistic. Be sure to include the following in your graph.

- a main title
- a title for the x-axis

(c) Would it be appropriate to use a critical value from a normal distribution in order to find a confidence interval for the true Q3 value?

(d) What is your estimate for Q3?

(e) Compute a 70% confidence interval for the true value of Q3.

(f) Determine how many players in the NHL have had more than your estimate for Q3 minutes of penalty time. Show your code.

(g) Compute the percentage of players in the NHL who have had more than your estimate for Q3 minutes of penalty time. Is this value close to what you might expect? Explain.

```
# (a) Answer below:


# (b) Answer below:


# (c) Answer below:


# (d) Answer below:
```

```
# (e) Answer below:


# (f) Answer below:


# (g) Answer below:
```