HW1 – IDSC 4444 – FALL 2021

TOTAL POINTS: 5

Instructions:
Your HW submission should consist of one pdf file containing the answers to all the questions included in the homework. For the hands-on part of the HW where you are required to use R, you must also copy and paste the code you used to produce your results, or you will lose points.

PRACTICE ON R

For this part homework, you need to analyze the dataset LaptopSales.csv which contains the details of the sales for some laptops at several stores, during the first 10 days of January 2008.
Into your submission, copy and paste the code uses to answer each question, the output of the code and any additional explanation (if required). Your final submission should be a PDF file where, for each question, you provide the code used to answer that question, the output generated by the code and explanation where required.
Additionally, copy and paste the entire R code at the end.
NOTE: you must provide your R code as part of each answer, or you will lose points regardless of whether the answer in correct. You do not need to submit the R script file. Simply copy and paste the R code associated to each question/answer in your submission.
See the example at the end of this doc on how to structure the answers.

Part 1 (1.4 points):

a) (0.2 points) Import the laptop sales dataset, give it a proper name
b) (0.2 points) Summarize your data, check for missing values; if missing values exist, report for which variable(s) and row number(s) (you do not need to fix the missing data).
c) (0.2 points) What is the average price of a laptop and what is the median price?
d) (0.2 point) What is the average price of a laptop with Integrated Wireless? And what is the average price without Integrated wireless?
e) (0.2 points) What is the configuration type with the highest price?
f) (0.2 points) How many laptops have HD size < 150?
g) (0.2 points) What is the total value, in $, of all the laptops sold in this dataset? (In other words, what is the sum of all the prices)

Part 2 (3.6 points): For each plot you construct, you also need to provide a description of the insights you get from it. All your plots should have a proper title, a proper labeling of (both) axis and proper axis limits (for both). Proper axis limit means that the axis limits should be such that no value is excluded or cut out of the plot. You may be required additional things for individual plots, see the questions below. You are required to use ggplot for the plots.

a) (0.5 points) Plot the distribution of the Customer Store Distance variable and provide a description of the insights you get from it. Your plot should be created in ggplot and include the following options: the fill color should vary depending on the count in each bar. The fill gradient should be purple for low values and darkblue for high values. The alpha should be set to 1.
b) (0.5 points) Plot a boxplot of the variable Retail Price (only) and provide a description of the insights you get from it. Use ggplot. For this plot, you do not need to label the X axis (but you do need a title and label the Y axis). Your plot should display outliers (if any) in orange, with a shape of a triangle and size 3 (HINT: to understand how to change shapes see http://www.sthda.com/english/wiki/ggplot2-point-shapes)
c) (1 points) Plot a boxplot to compare the retail price of laptops divided in groups based on the variable HD.Size..GB.; provide a description of the insights you

get from it. Use ggplot; each individual box in your plot should have a different fill color, based on HD.Size..GB. The outliers should be identified with a shape and color of your choice.

d) Finally, we want to understand what type of relationship exists between the retail price and the battery life hours of a laptop.
    a. (0.5 points) Plot the relationship between battery life hours and the retail price. Use ggplot; the color of the points should change based on battery life hours. What do you see? Provide an explanation.
    b. (1.1 points) Plot a histogram of retail price, divided by categories based on the battery life hours and provide a description of the insights you get from it. Use ggplot. Your histogram should have a different color for each category and an alpha = 0.8.

    HINT: you may need to create a new "categorical" variable for the battery life hours, that is currently a numerical variable.

    Here an example of how to do that:

    Assume you have a dataset called “mydata” and you have a numerical variable “age”. You want to group the variable age into categories: Elder, Middle Aged, Young. You first need to create a new variable that will contain the categories, let us call it “agecat” and then assign the records to the different categories based on some rules you decide. For example, all records for which age > 75, are assigned to the category “Elder”.

    See example code below:

    mydata$agecat[age > 75] <- “Elder”

    mydata$agecat[age > 45 & age <=75] <- “Middle Aged”

    mydata$agecat[age <= 45] <- “Young”

Example of how to structure your answers

NOTE: you can copy and paste your code or take a screenshot and attach it to the doc. Similarly, for the output, graphs, etc.. you can take a screenshot and paste it into the submission.

#Questions:

#a) Load the dataset "iris", show the first three rows

```
iris <- data("iris")
head(iris, n = 3)
```

#Output

```
  sepal_length sepal_width petal_length petal_width       class
1          5.1         3.5          1.4         0.2 Iris-setosa
2          4.9         3.0          1.4         0.2 Iris-setosa
3          4.7         3.2          1.3         0.2 Iris-setosa
```

We see from the first three rows, that the dataset has a total of 5 variables (columns).

#b) Summarize the data

```
summarize(iris)
```

```
  sepal_length    sepal_width     petal_length    petal_width                class
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Iris-setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Iris-versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   Iris-virginica :50
 Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

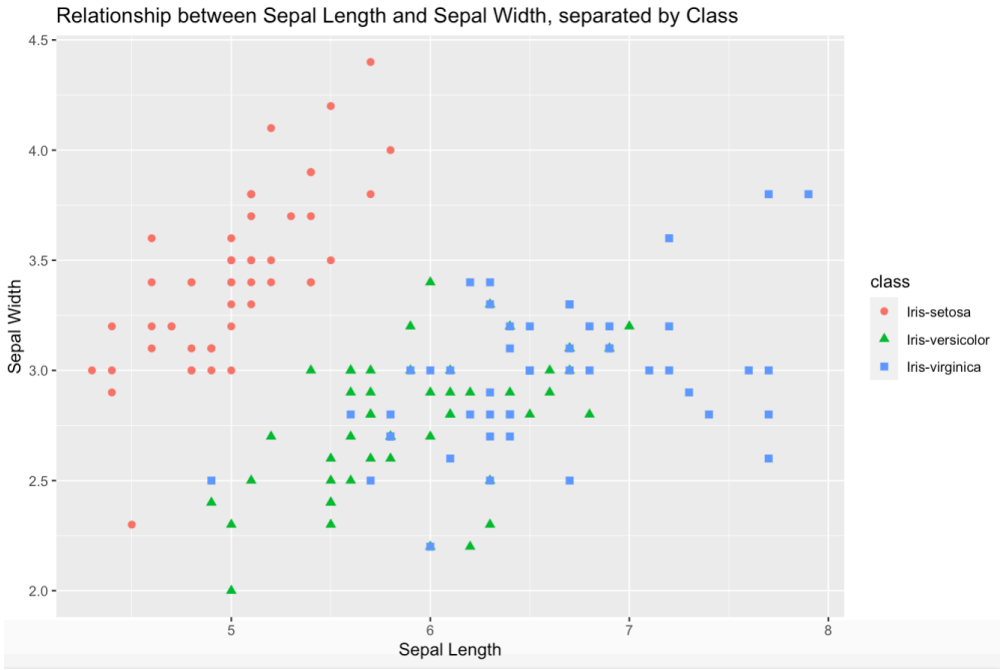
From the summary table, we observe that there are three classes of iris, setosa, versicolor and virginica. The average sepal length is about 5.8, while the average sepal width is about 3. The average petal length is about 3.75 and the average petal

width is about 1.1. The petal length seems to vary more than the sepal length: the petal length goes from 1 to almost 7; while the sepal length goes from 4.3 to 7.9.

#c) Plot the relationship between sepal length and sepal width, with a different color and shape for each class of iris. Your graph should have a proper title, proper axis labels and proper axis limits.

library(ggplot2)

```
ggplot(data = iris, aes(x=Sepal.Length, y=Sepal.Width, color=class, shape = class)) +
  geom_point(size=2) +
  ggtitle("Relationship between Sepal Length and Sepal Width, separated by Class") +
  labs(x = "Sepal Length", y = "Sepal Width")
```



From the graph, we observe that the Iris setosa (in red circles) seems to have an average Sepal length lower than the other classes, but an average sepal width higher than the other classes. Also, there seem to be a very steep, positive relationship between width and length: the width seems to increase exponentially with the length. The values for Iris versicolor (green triangles) and iris virginica (blue squares) seems to overlap, for the great part. While the average for the sepal width may be close for the two classes, the Iris versicolor seems to have a slightly lower average for the sepal

length. For these two classes, the relationship between length and width seems also positive, but less steep, compared to the Iris setosa.