Risto B. Rushford
GSCM 451-Spr 2019
HW01

Short-Answer Problems

These concepts can appear on the optional short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1. What is a data table and how are the data values organized?

A data table is a computer file containing data organized in a rectangular table with the variable names listed in the first row of each column, and the measurements/values of those variables listed below.

2. What is a csv file? What are its properties, its primary advantage and its primary disadvantage?

CSV stands for comma separated value, it is a file which contains variable names and values separated by values. CSV files are easy to use by a variety of software to store data in tabular form.

3. What is the distinction between categorical and continuous variables? Provide an example variable of each along with some sample values.

A categorical variable is one that is either of one or some other type, such as either male or female. A continuous variable is one that is measured, such as the temperature read on a thermometer or the length of a board.

4. Why do bar charts but not histograms have gaps between the bars?

Bar charts are used to plot categorical data, and the order of the bars don't affect the meaning of the data. Histograms are used to plot quantitative data. There are no gaps between the bars of histograms because each bar represents a "bin" of data and there are no gaps between the bins. The bars of a histogram cannot be rearranged.

5. Compare the mean to the median. Indicate when they differ.

The mean is the average of a collection of numerical data. The median is the data point located at the middle of a sequentially ordered list of values. The difference is in how they are derived:

Given the dataset: 3, 5, 43

- Example of a mean: (3+5+43) / (3) = 17
- Example of a median: the middle data point: 5

6. What is reproducibility and how do we attain it?

Reproducibility is when work such as a lab experiment or data analysis activity is performed and documented such that anyone given the same or similar tools may perform the exact same work and produce similar if not identical results.

7. What is histogram arbitrariness based on bin width and bin starting point

The frequency of occurrence of values for continuous variables affects how a histogram can be effectively used for a given set of data. Bin width is used to capture the frequency of occurrence within a range of measurement values. The starting bin should be chosen keeping in mind the lowest values of the data.

7.1. histogram under-smoothing

When the bin width for a histogram plot is too small. This can result in having too many bins that provide little or no value to a decision maker.

7.2. histogram over-smoothing

When the bin width of a histogram plot is too large, clumping together data makes it difficult for decision-makers to see exactly where in a dataset significant measurements are being taken.

8. What is an outlier and how does it affect data analysis?

Outliers affect data analysis by skewing averages and by shifting attention toward sometimes statistically insignificant or anomalous occurrences. These can obscure more significant patterns in the data and lead to false conclusions or misunderstood representations and relationships in the data.

Worked Problems

1. Create an Excel data table

Consider the data in Figure 1, randomly selected from a data file of the body measurements of thousands of motorcyclists.

	A	B	C
1	Gender	Weight	Height
2	F	150	66
3	F	138	66
4	M	240	
5	M	178	71
6	F	130	64
7	M	200	74
8	F	140	70
9	M	220	77

Figure 1: Gender, Weight and Height of eight motorcyclists.

a. List each of the variable names in Figure 1 and classify each as continuous or categorical.

b. Manually enter each data value from Figure 1 into a worksheet (such as Excel). Copy and paste the data table from Excel (or whatever) into your homework document. A screen pic (not of the entire screen but just the worksheet) also works.

Gender	Weight	Height
F	150	66
F	138	66
M	240	
M	178	71
F	130	64
M	200	74
F	140	70
M	220	77

c. Every data table, whether in R or Excel, has a name. Excel stores a data table in a worksheet, and every worksheet has a name. What is the name of that worksheet (located on a tab toward the bottom left corner) on which you entered your data? Compare the name of the data table to the names of the variables defined in that worksheet. (This is the same distinction in any analysis system, such Excel or R.)

Read these data directly from the Excel file you created on your computer system into an R data table named **d**. Copy to your homework document the listing from the lessR **Read** function you used to read the data that displays the variable names, variable types, and first and last data values.

```
d <- Read("/home/risto86/Drive/1 Classes by Quarter/1 Spring 2019/GSCM 451 - Forecasting/Week 1_4-1-2019/HW01_451.xlsx")
```

d. Verify that the data values for the variables in the data table are stored within the analysis system in the intended format. That is, from the output of **Read** describe how the variable types defined by R correspond to your description of categorical and continuous variables.

As shown below, the Read() function identified two data types: character and integer. The character data types correspond to categorical variables (in this case, gender). The integer data types correspond to continuous, or measured, variables (in this case, weight and height).

```
Data Types
-------------------------------------------------------
character: Non-numeric data values
integer: Numeric data values, integers only
-------------------------------------------------------

 Variable                 Missing  Unique
     Name     Type  Values  Values  Values   First and last values
-----------------------------------------------------------------------------------
1    Gender character     8       0       2   F  F  M ... M  F  M
2    Weight   integer     8       0       8   150  138  240 ... 200  140  220
3    Height   integer     7       1       6   66  66  NA ... 74  70  77
-----------------------------------------------------------------------------------
```

e. Display the data from within R with the **`print(d)`** function call and then copy to your homework document. Note that simply entering the name of the R object, here called **`d`**, is an abbreviation to invoke the **`print`** function.

```
print(d)
  Gender Weight Height
1      F    150     66
2      F    138     66
3      M    240     NA
4      M    178     71
5      F    130     64
6      M    200     74
7      F    140     70
8      M    220     77
```

f. From the previous answers, compare the data stored in Excel and then compare to the representation of the data stored in R. What does **`NA`** refer to in the R data table?

Whereas the .xlsx file had an empty cell for the third "Height" value, R outputs the missing data as "NA". This simply means that there is missing information.

2. Bar Chart of Frequencies (from data)

A motorcycle clothing company makes jackets of three styles: Lite, Medium and Thick. The company needs guidance as to how many different jackets of each type to bring to a gathering of motorcyclists of a specific brand, here BMW or Honda motorcycles. The data are recorded from past sales of motorcycle jackets to owners of BMW or Honda motorcyclists.

For now, just analyze the type of jackets sold, and ignore the motorcycle type.
Consider the data on the web at:

http://web.pdx.edu/~gerbing/data/Jackets.csv

Data file

a. Verify that there is an actual data file called Jackets.csv at the specified web address (URL) by pointing your browser at that URL. Copy and paste the first several lines of the data file and paste into your homework document. Display in a monospaced font such as Courier New to maintain the positioning of the columns.

```
Bike,Jacket
BMW,Lite
Honda,Lite
Honda,Lite
Honda,Med
```

---------->

Bike	Jacket
BMW	Lite
Honda	Lite
Honda	Lite
Honda	Med

b. What is the format of that document?

The document format is of .csv, which stands for comma-separated-values. Opening the document in Excel (or in my case, Google Sheets) translates it into a tabular table format as displayed above.

c. Just by looking at the data directly, without any R or Excel analysis, how many variables are in the data file? What are their names? What is the relevant variable for this analysis?

There are two variables: **Bike**, and **Jacket**. The relevant variable for this analysis is **Jacket**, since the company is determining how many of each type of jacket to the motorcyclist gathering.

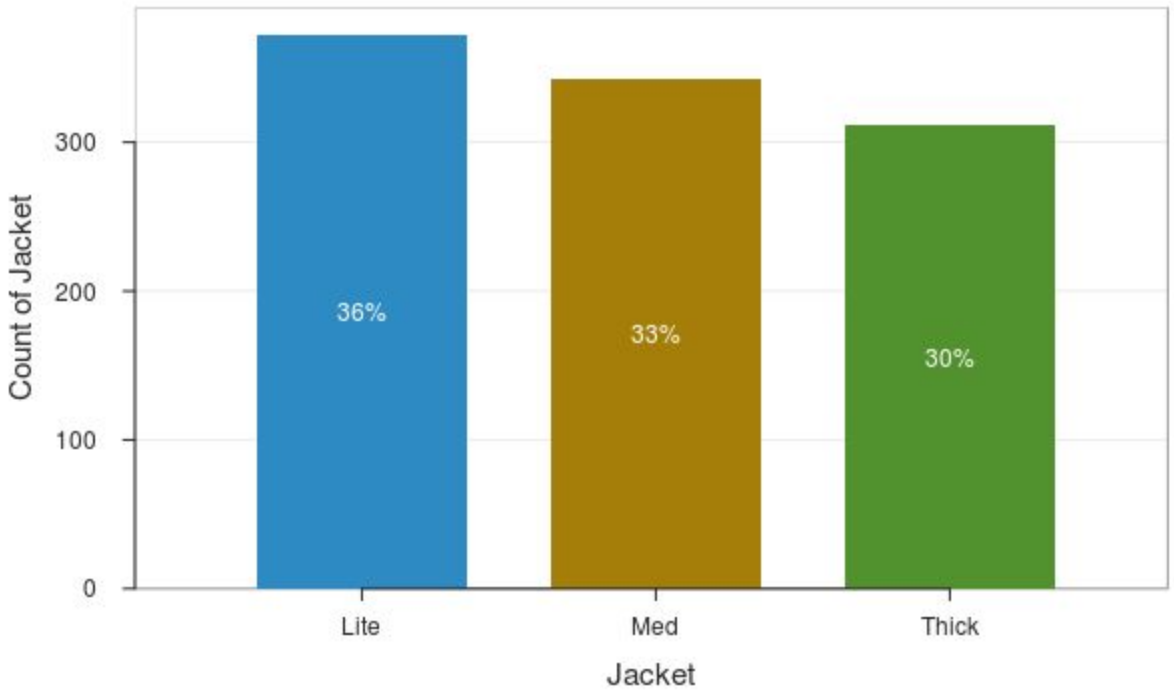
Analysis

Read the data into R directly from the web and then present the equivalent results (such as from the R BarChart function) for the distribution of Jacket Types with:

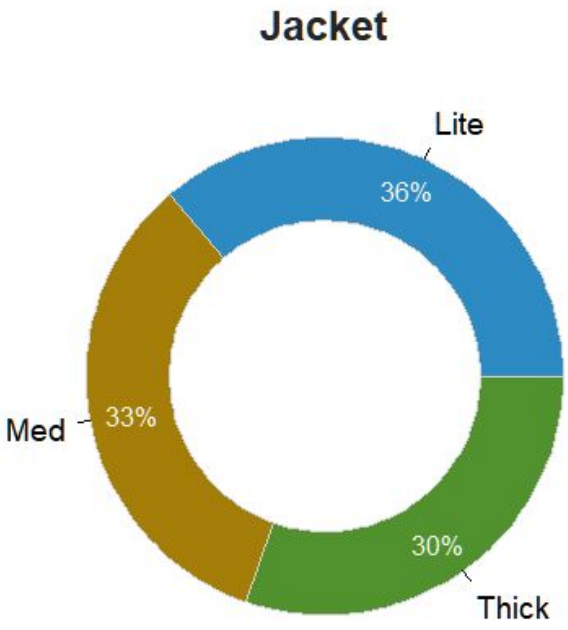
d. frequency table

	Lite	Med	Thick	Total
Frequencies:	372	342	311	1025
Proportions:	0.363	0.334	0.303	1.000

e. bar chart



f. pie (or ring) chart



Interpret

g. interpretation, what do these results mean?

These results show the proportion of each jacket type to be ordered for the motorcyclist gathering, with roughly a third of the total order being from each type, but slightly more for the Lite, and slightly less for Thick.

3. Histogram

The data for this exercise are included with lessR, obtained when the package was downloaded, so do not need an Internet connection to access. To read these built-in data sets, add the `in.lessR=TRUE` option to the Read statement.

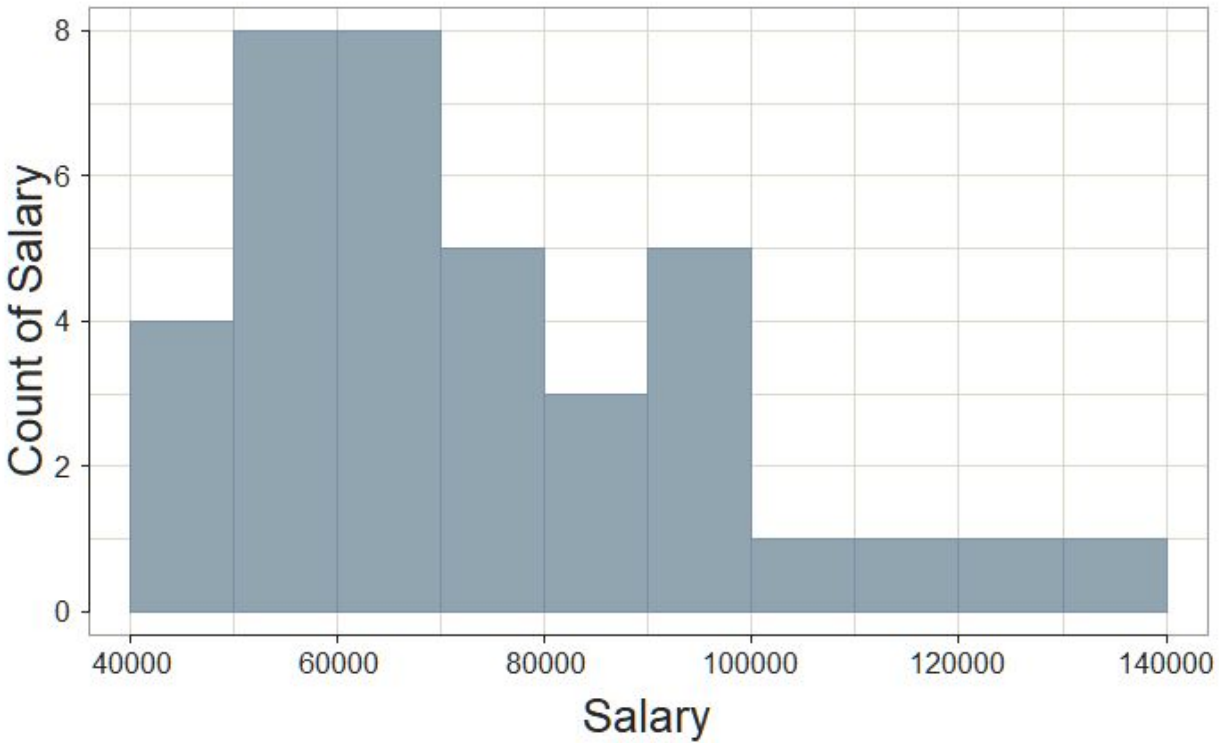
To read the Employee data directly from lessR:

```
d <- Read("Employee", in.lessR=TRUE)
```

or, read from the web at:

```
d <- Read("http://lessRstats.com/data/employee.xlsx")   [Excel file]
```

a. Obtain the default histogram for Salary. [That means display it.]



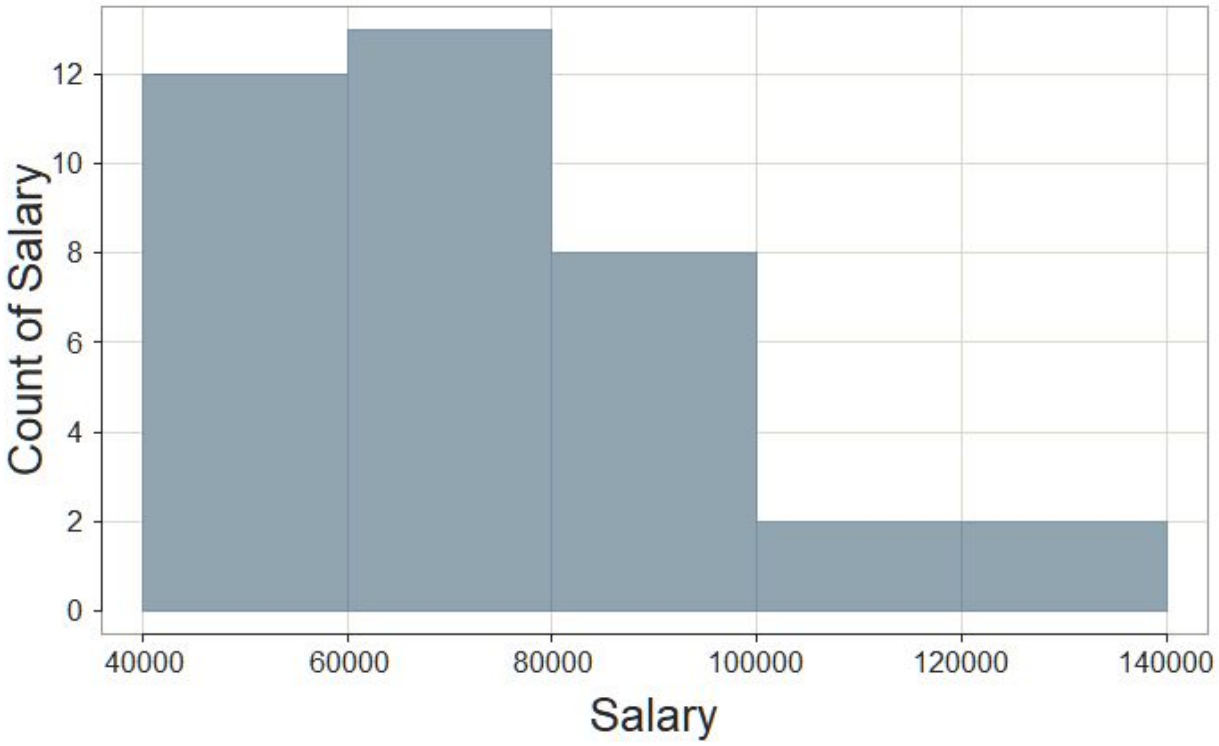
b. What is the default bin width? Starting point of the first bin?

The default bin width is $10,000, with the first bin beginning at $40,000

c. Interpret the histogram.

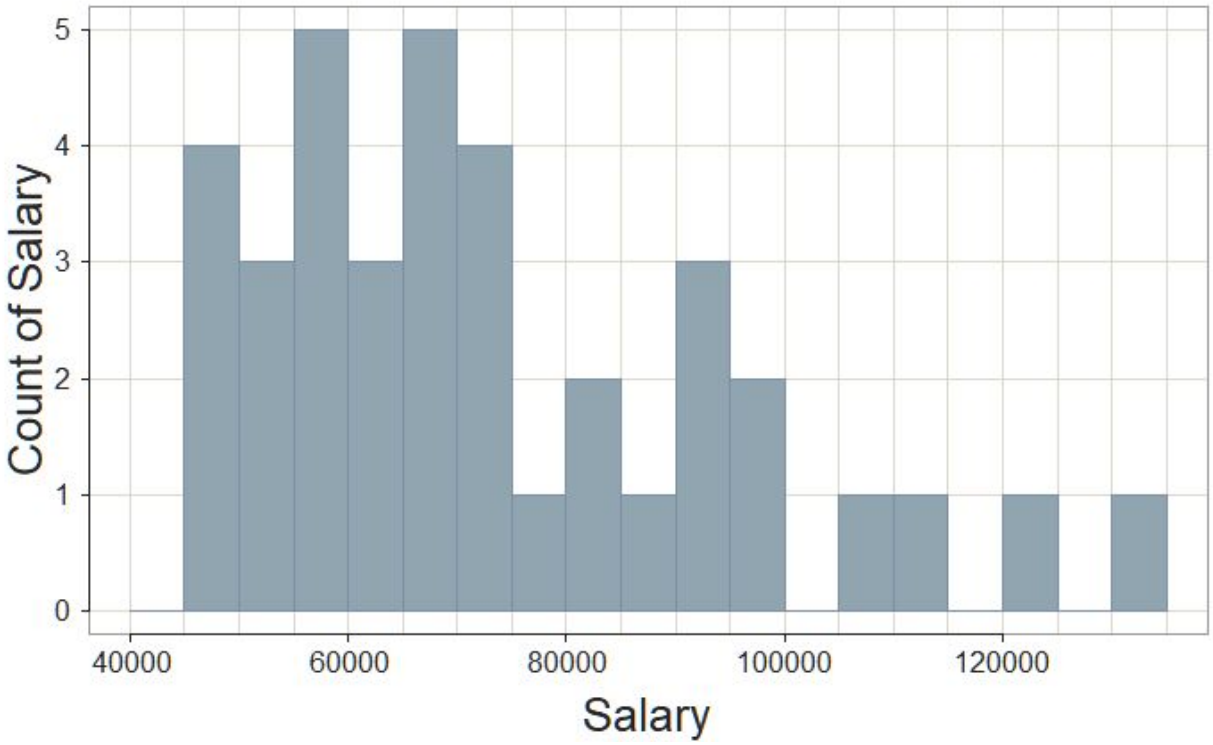
The histogram shows the distribution of wages in each bin, showing that the highest proportion of wages are between $50,000 and $70,000, but with instances going up to $140,000

d. Using the bin.width option deliberately over-smooth the histogram (bins too wide). Why is the resulting display not optimal?



Using a bin width of $20,000 results in undersmoothing. This is non-optimal because it clumps all of the measurements into too few brackets. This makes it seem like, even though there's a big difference between incomes of $40k and $60k, like they are categorically the same.

e. Using the bin.width option deliberately under-smooth the histogram (bins too narrow). Why is the resulting display not optimal?



Using a bin width of only $5,000 results in oversmoothing. This is non-optimal because there are now bins with no data, resulting in wasted space in the chart.

4. R transformations and Write Data

a. Within R create two new variables. Convert Height to cm and Weight to kg. Call the new variables Ht_cm and Wt_kg. Show the R code to accomplish these transformations.

I'm assuming the use of the table that we created at the beginning of this exercise:

```
Ht_cm <- d$Height * 2.54
167.64 167.64     NA 180.34 162.56 187.96 177.80 195.58
```

```
Wt_kg <- d$Weight * 0.45359237
68.03886  62.59575 108.86217  80.73944  58.96701  90.71847  63.50293
99.79032
```

b. Carrying extra decimal digits beyond the precision of measurement is misleading. Do another transformation by rounding the newly created values to two decimal digits with the R function **round()**. The form of the function is the same as for Excel: the first parameter value is the definition of the variable, here the variable name, and the second parameter value is the number of decimal digits for which to round). Example: round(x, 4) rounds the values of the variable x to 4 decimal digits.

```
round(Ht_cm, 2)
167.64 167.64     NA 180.34 162.56 187.96 177.80 195.58
```

```
round(Wt_kg, 2)
68.04  62.60 108.86  80.74  58.97  90.72  63.50  99.79
```

Note there was no change in the values for Ht_cm. This is because the conversion from one inch is exactly 2.54 centimeters, which is already at the level of precision specified for this activity.

c. Display the revised data frame in R and copy to your homework document.

	Gender	Weight	Height
1	F	68.03886	167.64
2	F	62.59575	167.64
3	M	108.86217	NA
4	M	80.73944	180.34
5	F	58.96701	162.56
6	M	90.71847	187.96
7	F	63.50293	177.80
8	M	99.79032	195.58

d. Write the data frame to an Excel file. Copy and paste the Excel file into your homework document, either directly or as a screen pic of that document (not the whole screen).

Gender	Weight	Height
F	68.04	167.64
F	62.6	167.64
M	108.86	
M	80.74	180.34
F	58.97	162.56
M	90.72	187.96
F	63.5	177.8
M	99.79	195.58