**\Visualization of Quantitative Variables with M&Ms**

**As a result of completing this exercise you should be able to:**

* Summarize and examine the data using head, summary, fivenum and descr functions in R
* Recognize *categorical* and *quantitative* variables
* Produce *charts and graphs for quantitative variables* in the R Statistical Software and understand the differences:
  + Boxplot and Histogram
* Create scatter plots using plot function

**Instructions**

1. Download the “ISAT251\_MMs.csv” file on Canvas. There are six columns in total in the dataset.

student id color defect total.number weight

Each column will correspond with ONE and only one variable

* 1. In the student column, record ***recorder’s JMU eID*** (e.g. instructor’s is yang4cx)
  2. In the id column, record the ***number*** of M&M that recorder is observing. The first M&M that student collects data on will be 1, the second M&M you collect data on will be 2, and so on.
  3. In the color column, record the ***color*** of each M&M:
* **R** for Red, **BR** for Brown, **O** for Orange, **Y** for Yellow, **BL** for Blue and **G** for Green
* It’s important to mark these exactly as requested. Recorders should use capital letters, be sure not to include a space before or after the color code and be sure not to confuse BR and BL (or worse, mark B for either brown, or blue, or both - this will cause big problems later).
  1. In the defect column, record ***type of defect*** on the M&M. Use the following coding scheme:
* **N** = No defect found
* **C**= Cracked, chipped or broken shell
* **L** = Letter missing or only partially printed on the shell
* **M** = More than one defect
* Similar to the ***color*** column, recorders should type these codes exactly as what are listed above to prevent problems later which need to be fixed by data cleaning.
  1. In the total.number column, mark the total number of M&Ms each recorder observed (it will be around 50 for a regular sized bag). The value will be the same all the way down each recorder’s column (the same bag).
  2. In the weight column, insert the weight of your M&M (the whole bag) with unit **gram**. The value will be the same all the way down for the same bag

1. Which variables are **categorical**, and which are **quantitative**? Filling the rightmost column with the information you get in step 3 to step 5.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Categorical or Quantitative?** | If categorical, list *all possible values* that variable can take on. If quantitative, describe the *range* of possible values (w/ units). |
| student | categorical | (There will be one value for each student who collected the data before, what are they?)  Grenobwk, schiesjc, bell2ml, edejerrr, overbyjc, golds2gd, Wolfelc, henriqta, suterbs, janigaek, Tangbl, COCKRACJ, sungelee, hooveret, mageema, light2ra, hellerem, burrisss, placeba, lagodkrf, lupatb, robin3mn, jacks2gs, fellenjd, billaj, lottnb, dempsekl, lukerzt, tarterkb, podrassn, moran2tp, tenoriaa, delzernm, belislka, Richa7jl, baxterka, Saclolcr, peter3kr, ellio2ce, brotemar, MCKEEAM, cortinem, knabele, lebedeay, sumieljt, sortorae, henderin, patte2mg, romerojg, obidinnc, emchea, morto2tm, harrinnr, shoem2tc, lovingdg, webbat, ander2cf, quinonrm, Kellyfe, campbens, colli2pl, garnerbp, heshmarr, golandtn, |
| id | categorical | (There will be one value for each M&M recorded by each student, what are the possible id values?)  1-61 |
| color | categorical | (What are the labels or letters you have in the color column?)  **R** for Red, **BR** for Brown, **O** for Orange, **Y** for Yellow, **BL** for Blue and **G** for Green |
| defect | categorical | (What are the labels or letters you have in the defect column?)   * **N** = No defect found * **C**= Cracked, chipped or broken shell * **L** = Letter missing or only partially printed on the shell * **M** = More than one defect |
| total.number | quantitative | 52-61 M&Ms |
| weight | quantitative | 42-52 g |

1. Now it’s time to **bring the data into R.**
2. Open the RStudio Statistical Software. Use file.choose function to navigate and get the full path to your file where your “ISAT251\_MMs.csv”stored in it. Then, store the path to a variable call myfile.

> myfile <- file.choose()

**The full path to *your* file is: /Users/zshindc/Downloads/ISAT251\_MMs.csv**

1. Create a new variable called mms.data which will hold every observation for records everyone collected outside of the class. The argument header=TRUE tells R that the first row of my data set contains variable or column names.

> mms.data <- read.csv(myfile,header=TRUE)

Once you’ve loaded the CSV file using read.csv, you can check to make sure it’s there with the head command, which pulls out the first six observations only.

> head(mms.data)

**Snapshot your R code and the first six lines of your data set and paste them here:**

A close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. Before you jump into your analyses, please use the summary function to get an overview of your data as well as to check whether there are weird data values or missing values.

> summary(mms.data)

**Snapshot your code and the summary of your data set and paste them here. Use the information to fill in the rightmost column of the table in the step 1 (you may need extra information from the next step):**

A screenshot of a cell phone

Description automatically generated

A screenshot of text

Description automatically generated

1. Provide **descriptive statistics** for each variable in the dataset. (Look at the Lab 1 if you forgot what this means.) We’re going to do it a slightly different way in this lab… because in R, there’s always more than one way to accomplish a task.
2. First, be sure you have the descry package installed and load it with the library command in R:

> install.packages("descr")

> library(descr)

> descr(mms.data)

**Snapshot your code and the summary of your data set and paste them here. Use the information the information to fill in the rightmost column of the table in the step 1:**

A screenshot of a cell phone

Description automatically generatedA close up of text on a white background

Description automatically generated

1. Before we continue, we have to get the MM’s data per bag rather than per chocolate, because the variables we are interested in weight and total.number, are at bag level and have many duplicates. If there are 50 chocolates in a bag, there are 50 duplicates of weight values and total.number values for that bag in our data.

To get the values for weight and total.number per bag, we can select the first row of data of each bag (each student’s first record), where the id is 1. The which function can help us get those rows. It checks all the rows and find those satisfy the condition mms.data$id==1. Then, the which function return us the index numbers of those rows (which make condition to be True).

To access subset of data from mms.data, we use squared brackets “[ “ “]”. Within squared brackets, type the row indexes and column indexes. Remember to separate the two arguments by a comma. In the square brackets. the selected row indexes are which(mms.data$id==1). For the column indexes, we want all the column, so we don’t need to type anything after the comma. Without specific column indexes, R will get all the columns in the data. **Using the following statement, we get MM’s data at bag level. Please use the data, mms.data.bags, to answer the following question 4.c., 5 and, 6**.

> mms.data.bags <- mms.data[which(mms.data$id==1),]

Before you go ahead, please check whether you get the data correctly by using head function and summary function. **Please paste the snapshot of your codes and the outputs in R here.**

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated

> head(mms.data.bags)

> summary(mms.data.bags)

1. For quantitative variables, you can create a **five number summary.** Let’s create the five number summary of the **weight variable**.(Five number summary will work on any quantitative variable.)

> fivenum(mms.data.bags$weight)

**Snapshot your code and the output of this function in R and paste them here. In addition, please describe what these numbers mean blow.** (Hint: Look them up in your book & label your output.**)**

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Minimum – smallest value in dataset

First quartile (Q1) – point at which 25% of our observations fall below that value

Median – point at which 50% of our observations fall below that value

Mean – average value of all observations

Third quartile (Q3) – point at which 75% of observations

Maximum – largest value in dataset”

1. Please create **a histogram** and **a box plot** for the **weight** variable in the mms.data.bags. You can find all the code for your plots in the appropriate part of Section 2 in your textbook. **Please paste the snapshot of your codes and your plots below. In addition, for the *box plot*, please explain how to read it, specifically, describe which part of the box plot indicating the minimum, Q1, median, Q3, maximum, and outlier, if any, individually.**

*Thank Dr. Radziwill’s help for the Exercises*

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The white dot at 42g is an outlier. The left line at 46g is the minimum. The gray box that is left of the black line is Q1. The gray box that is right of the black line is Q3. The right line at 52g is the maximum.

1. Please create **a scatter plot** using the weight column and total.number column in the mms.data.bags. **Please paste the snapshot of your codes and your plot below**.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated