Introduction to Statistics and Data Science with M&Ms

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

- Describe the *metadata* associated with a set of data (WHO, WHAT, WHEN, WHERE, HOW & WHY)
- Recognize *categorical* and *quantitative* variables
- Understand what a *level of analysis* is
- Identify the *cases* and *variables* in a data set
- Produce *charts and graphs* in the R Statistical Software and understand the differences:
    - Pie Chart, Bar Chart, Segmented Bar Chart
- Produce *contingency tables* showing the joint, marginal, and conditional distributions of two categorical variables using `table` and `CrossTable`
- Produce *descriptive statistics* for quantitative variables
- Load data and new packages (e.g. `gmodels`) into R

Instructions

0. Download the "ISAT251_MMs.csv" file on Canvas. There are six columns in total in the dataset.

student	id	color	defect	total.number	weight
grenobwk	**1**	**BL**	**N**	**56**	**49**

Each column will correspond with ONE and only one variable

a. In the `student` column, record ***recorder's JMU eID*** (e.g. instructor's is yang4cx)
b. 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.
c. 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).
d. 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.

e. 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).
f. 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. Now it's time to **bring the data into R.**

a. Move the file to a directory on your local machine that you will remember, and write down the full path to the filename here. For example, mine is `D:/R/ISAT251_MMs.csv` because I called the file `ISAT251_MMs.csv` and I stored it on my `D:` drive in the subfolder I called `R`.

The full path to my file is: `D:/R/ISAT251_MMs.csv`

---

b. Open the R Statistical Software. On the lab machines, it's in the Programs area. When you've opened R, it will look like this:

```
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

That caret (>) at the bottom is the R Prompt. R is like a fancy calculator: you type your commands in at the caret, and then R executes your wishes and performs calculations, plots charts and graphs, or sometimes both. **DO NOT TYPE THE CARET**, just the command after the caret. Once you hit enter, you should see the results from typing that command. Now use the `file.choose` function or command to get the path to your csv file and save the path to a variable (called `myfile`). Once you hit enter, a new window will pop up (If not, check the flashing application on the taskbar and click it.)

The file we are looking for is "`ISAT251_MMs.csv`"; once you find it through the pop-up window. The path to the file will be saved in the variable `myfile`. Type the variable name and hit enter. The content stored in the variable will be shown, which the full path to your file. *Copy and paste the path below.*

```
> myfile <- file.choose()
> myfile
```

Please type the full path to *your* file here :

/Users/zshindc/Documents/JMU OneDrive/ISAT 251/R/ISAT251_MMs.csv

I create a new variable called `mms.data` which will hold every observation for every M&M, and use `read.csv` function to load the data in the csv file into R. The argument `header=TRUE` in `read.csv` function tells R that the first row of my data set contains variable names (student, id, color, etc.) Once I've loaded the CSV file using `read.csv`, I can check to make sure it's there with the `head` command or function, which pulls out the first six observations only. You will have more columns in your data and your data will look different from the example.

```
> mms.data <- read.csv(myfile,header=TRUE)
> head(mms.data)
  student id color defect weight total.number
1 allenrj  1    BL      L   50.14           57
2 allenrj  2    BL      N   50.14           57
3 allenrj  3    BL      L   50.14           57
```

```
4 allenrj  4    BL      N   50.14     57
5 allenrj  5    BL      N   50.14     57
6  Pinoja  4    BL      N   48.30     49
```

Copy and paste your R codes and the first six lines of *your data set* here, like the following example:

```
student id color defect total.number
1 daceyij  1    BR      N           57
2 daceyij  2     O      N           57
3 daceyij  3    BL      N           57
4 daceyij  4     G      L           57
5 daceyij  5    BL      N           57
6 daceyij  6     O      N           57


grenobwk      1      BL     N      56
grenobwk      2      BL     N      56
grenobwk      3      BL     L      56
grenobwk      4      BL     N      56
grenobwk      5      BL     N      56
grenobwk      6      BL     L      56
```

Error: unexpected numeric constant in "grenobwk 6" popped up

2. Now it's time to prepare **the tables**, which can display the counts of observations (called a frequency distribution), the percentage of observations in each category (called a relative frequency distribution, because the counts are displayed relative to the total number of observations), or sometimes both.

a. Prepare the **frequency distribution of defects** using this code. Paste the snapshot(s) of your codes and results below.

```
table(mms.data$defect)
```

```
   C    L    M    N
 194 1168  169 2024
```

b. Prepare the **frequency distribution of colors** using this code. Paste the snapshot(s) of your codes and results below.

```
table(mms.data$color)
```

```
  BL   BR    G    O    R    Y
 715  720  648  619  490  363
```

c. Prepare the **relative frequency distribution of defects** using this code. Notice that we are just wrapping the code above in a new command that computes the percentages. Paste the snapshot(s) of your codes and results below.

```
prop.table(table(mms.data$defect))
```

```
          C           L           M           N
0.05457103 0.32855134 0.04753868 0.56933896
```

d. Prepare the **relative frequency distribution of defects and round to two significant digits** using this code. Notice that we are just wrapping the code above in a new command that computes the percentages. Paste the snapshot(s) of your codes and results below.

```
round(prop.table(table(mms.data$defect)),2)
```

```
   C    L    M    N
0.05 0.33 0.05 0.57
```

e. Prepare the **relative frequency distribution of colors** using this code. Notice that we are just wrapping the code above in a new command that computes the percentages. Paste the snapshot(s) of your codes and results below.

```
prop.table(table(mms.data$color))
```

```
        BL         BR          G          O
0.2011252 0.2025316 0.1822785 0.1741210
         R          Y
0.1378340 0.1021097
```

f. Prepare the **relative frequency distribution of colors and round to two significant digits** using this code. Notice that we are just wrapping the code above in a new command that computes the percentages. Paste the snapshot(s) of your codes and results below.

```
round(prop.table(table(mms.data$color)),2)
```

```
  BL   BR    G    O    R    Y
0.20 0.20 0.18 0.17 0.14 0.10
```

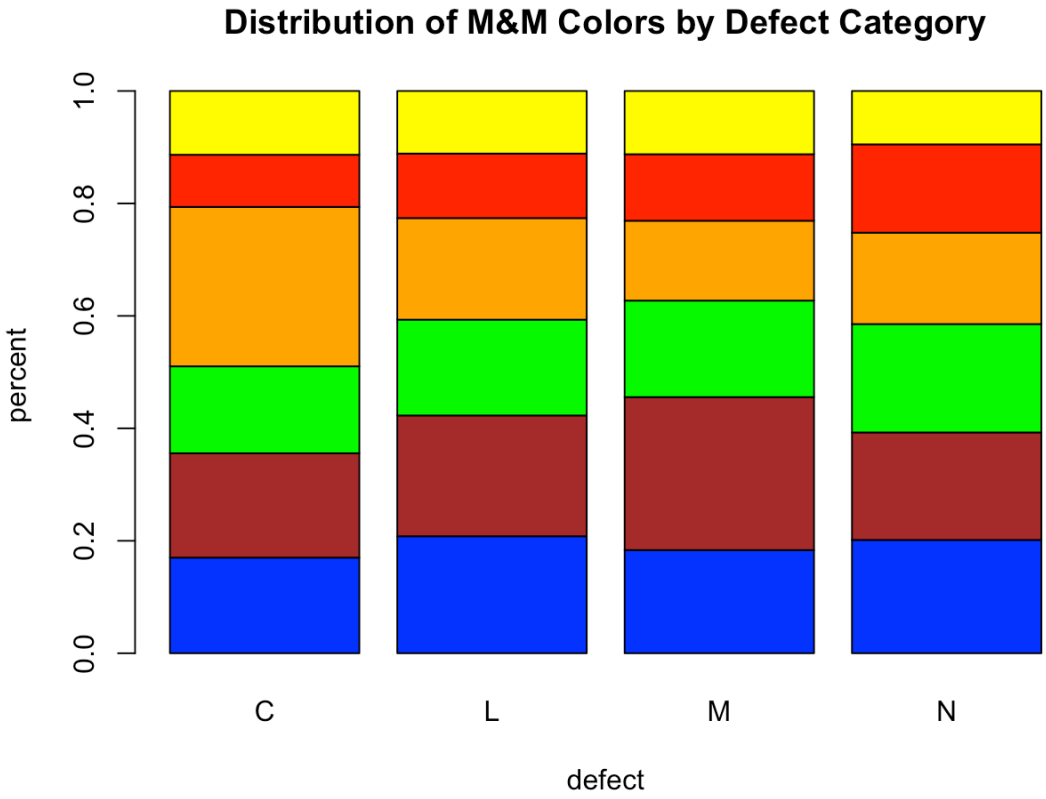
3. Prepare a **pie chart** and a **bar chart** for the frequency distribution of colors using `pie` function and `barplot` function. If needed, you can find all the code in the appropriate part of Section 2 in your book.

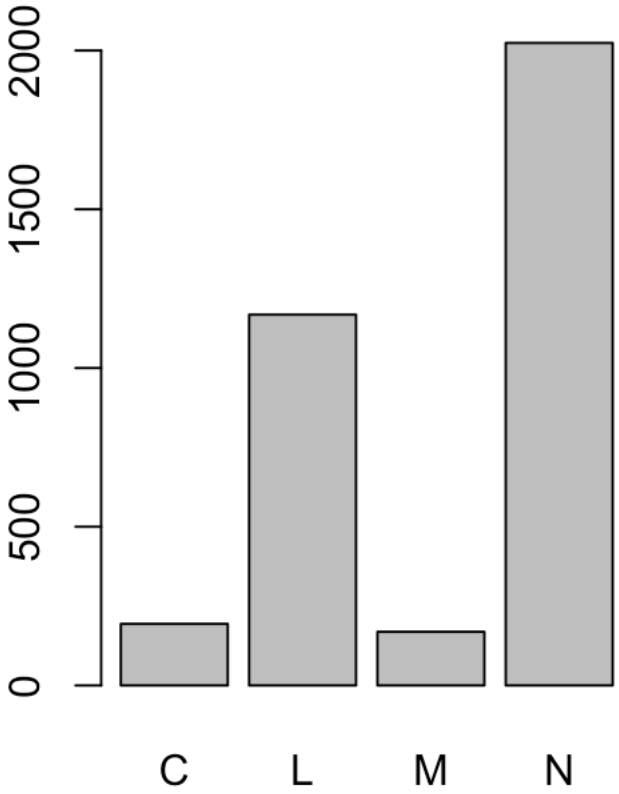
Next, create a **pie chart** and a **bar chart** for the frequency distribution of defects. Finally, create a **segmented bar chart** showing the **joint frequency distribution of color and defect**. Try to label as much as you can.

Paste the snapshot(s) of your codes and plots for all five charts below. Did you use the **area principle** when you produced your plots?

```
table(defect)

defect

C   L   M   N

194 1168  169 2024

barplot(table(defect))

my.defect <- c("blue", "brown", "green", "orange")

barplot(table(color),main="My M&M Defect Distribution",xlab="M&M Defects",ylab-"Number of M&Ms in Bag",col=my.defect)

col=c("red","blue")

barplot(table(defect),main="MY M&M Defect Distribution"

barplot(table(defect),xlab="M&M Defects")

table(mnms$defect)

mm.counts <- as.vector(table(mnms$defect))

names(mm.counts) <- c("Cracked,chipped or broken shell", "Letter missing or only partially printed on the shell", "More than one defect", "No defect found")

percents<- round(mm.counts/sum(mm.counts)*100,2)

my.labels <-paste(names(mm.counts)," ",percents,"%",sep="")

pie(mm.counts,labels=my.labels,main="My M&M Defects Distribution",col=names(mm.counts))

barplot(prop.table(mm.ct,2))

mm.ct <- table(mnms$color,mnms$defect)

mm.ct

barplot(prop.table(mm.ct,2))

mm.colors <- c("blue","brown","green","orange","red","yellow")
```

```
barplot(prop.table(mm.ct,2),main="Distribution of M&M Colors by Defect
Category",xlab="defect",ylab="percent",col=mm.colors)
```





Unable to set up pie chart for frequency distribution of defects.

4. Now it's time to prepare **fancy contingency tables** using the `CrossTable` method in the gmodels package. At this step, you'll learn how to import a new package to the base R distribution. First, you have to download the new functionality from the internet:

```
install.packages("gmodels")
```

[When the first menu pops up, go all the way to the bottom and select HTTP Mirrors. When the second menu pops up, choose a site close to your geographical location. Even though it's a little farther away, I like to use the California (CA) sites because they rarely give me problems. If any questions pop up, click `"Yes"`.]

Next, wake up the new package so that you can use it:

```
library("gmodels")
```

And prepare the contingency table with this code:

```
CrossTable(mms.data$color,mms.data$defect)
```

Paste the snapshot(s) of your codes and your contingency table below. What do all the numbers mean?

Cell Contents

|-------------------------|

| N |

| Chi-square contribution |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 3555

| mms.data$defect

mms.data$color	C	L	M	N	Row Total
BL	33	243	31	408	715
	0.928	0.278	0.263	0.002	
	0.046	0.340	0.043	0.571	0.201
	0.170	0.208	0.183	0.202	
	0.009	0.068	0.009	0.115	
BR	36	251	46	387	720
	0.276	0.882	4.049	1.282	
	0.050	0.349	0.064	0.537	0.203
	0.186	0.215	0.272	0.191	
	0.010	0.071	0.013	0.109	
G	30	199	29	390	648
	0.813	0.908	0.106	1.203	
	0.046	0.307	0.045	0.602	0.182
	0.155	0.170	0.172	0.193	
	0.008	0.056	0.008	0.110	

O	55	211	24	329	619
	13.331	0.286	1.001	1.556	
	0.089	0.341	0.039	0.532	0.174
	0.284	0.181	0.142	0.163	
	0.015	0.059	0.007	0.093	
R	18	134	20	318	490
	2.857	4.525	0.466	5.459	
	0.037	0.273	0.041	0.649	0.138
	0.093	0.115	0.118	0.157	
	0.005	0.038	0.006	0.089	
Y	22	130	19	192	363
	0.242	0.966	0.176	1.041	
	0.061	0.358	0.052	0.529	0.102
	0.113	0.111	0.112	0.095	
	0.006	0.037	0.005	0.054	
Column Total	194	1168	169	2024	3555
	0.055	0.329	0.048	0.569	

All the numbers represent the probabilities of 2 categorical variables (M&M defects and colors).

5. Produce a simplified contingency table using this code, and paste the snapshot(s) of your codes and results below.

```
CrossTable(mms.data$color,mms.data$defect,prop.r=TRUE,
prop.c=TRUE,prop.chisq=FALSE)

  Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

Total Observations in Table: 3555

mms.data$color	mms.data$defect				Row Total
	C	L	M	N	
BL	33	243	31	408	715
	0.046	0.340	0.043	0.571	0.201
	0.170	0.208	0.183	0.202	
	0.009	0.068	0.009	0.115	
BR	36	251	46	387	720
	0.050	0.349	0.064	0.537	0.203
	0.186	0.215	0.272	0.191	
	0.010	0.071	0.013	0.109	
G	30	199	29	390	648
	0.046	0.307	0.045	0.602	0.182
	0.155	0.170	0.172	0.193	
	0.008	0.056	0.008	0.110	
O	55	211	24	329	619
	0.089	0.341	0.039	0.532	0.174
	0.284	0.181	0.142	0.163	
	0.015	0.059	0.007	0.093	

```
          R |        18 |       134 |        20 |       318
|       490 |
            |     0.037 |     0.273 |     0.041 |
0.649 |     0.138 |
            |     0.093 |     0.115 |     0.118 |     0.157 |
|
            |     0.005 |     0.038 |     0.006 |     0.089 |
|
--------------|-----------|-----------|-----------|-----------|-
----------|
            Y |        22 |       130 |        19 |       192 |
363 |
            |     0.061 |     0.358 |     0.052 |     0.529 |
0.102 |
            |     0.113 |     0.111 |     0.112 |     0.095 |
|
            |     0.006 |     0.037 |     0.005 |     0.054 |
|
--------------|-----------|-----------|-----------|-----------|-
----------|
  Column Total |       194 |      1168 |       169 |      2024 |
3555 |
            |     0.055 |     0.329 |     0.048 |     0.569 |
|
--------------|-----------|-----------|-----------|-----------|-
----------|
```

- **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).

g. 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.

Based on the above contingency table, answer the following questions:

Question	What did you examine in the table? (circle or highlight one)	Your answer
a. Which combination of color and defect type was most common in the bag you examined?	● Each cell's count as a % of the table ● Marginal distribution of colors ● Marginal distribution of defects ● Conditional dist of defects for a given color ● Conditional dist of colors for a given defect	No defect and blue
b. Which color was most common in the bag you examined?	● Each cell's count as a % of the table ● Marginal distribution of colors ● Marginal distribution of defects	Brown

	● Conditional dist of defects for a given color ● Conditional dist of colors for a given defect	
c. Which defect was most common?	● Each cell's count as a % of the table ● Marginal distribution of colors ● Marginal distribution of defects ● Conditional dist of defects for a given color ● Conditional dist of colors for a given defect	No defect
d. Of the cracked M&M's, which color was most common?	● Each cell's count as a % of the table ● Marginal distribution of colors ● Marginal distribution of defects ● Conditional dist of defects for a given color ● Conditional dist of colors for a given defect	orange
e. Of the yellow M&M's, what percentage was more than one defect?	● Each cell's count as a % of the table ● Marginal distribution of colors ● Marginal distribution of defects ● Conditional dist of defects for a given color ● Conditional dist of colors for a given defect	No defect
f. Based your bag, which color was the most likely to be perfect (no defect)?	● Each cell's count as a % of the table ● Marginal distribution of colors ● Marginal distribution of defects ● Conditional dist of defects for a given color ● Conditional dist of colors for a given defect	red

6. Now, we'll use methods to generate even more **descriptive statistics**. These supplement charts and graphs to provide an overview of your data for your audience. Paste the snapshot(s) of your codes and results below. What do these numbers mean? [Note: There are lots of great utilities for descriptive statistics and pretty charts based on contingency tables in the `descr` package. We won't go into them here, but you might want to explore them.]

```
summary(mms.data)
```

```
student             id
 Length:3555       Min.   : 1.00
 Class :character   1st Qu.:14.00
 Mode  :character   Median :28.00
                    Mean   :28.35
                    3rd Qu.:42.00
                    Max.   :61.00
   color             defect
 Length:3555       Length:3555
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```
 total.number      weight
 Min.   :52.00   Min.   :42.00
 1st Qu.:55.00   1st Qu.:48.19
 Median :56.00   Median :50.00
 Mean   :55.71   Mean   :49.43
 3rd Qu.:57.00   3rd Qu.:50.80
```

Max. :61.00 Max. :52.00