BANA 7025
Data Wrangling with R
Homework #1

Deadline to submit via Canvas: 9:00AM, Monday, Oct 25th, 2020 (only one person per group sends to me)

Submission expectations

You will submit an .R script containing all code and a Word document containing code and explanations for each question.

NOTE!

You should use Base R for all questions on this assignment! In other words, do not load any packages! I will deduct 5 out of 10 total points on this assignment if you load any packages (e.g., tidyverse, dplyr, plyr, rio)!
I'm more than aware that we'll use the Tidyverse for the remainder of the course. However, this assignment is meant to assess your Base R skills that we learned this week.

Example of how to format your Word document using question #1

We import the data set into RStudio, showing the first 10 rows after importing, with code and output as shown below.

```
data <- read.csv("C:/Users/Justin/Documents/week1_cincy_crimes.csv",
         stringsAsFactors = FALSE)
head(data, 10)
```

	instanceid	closed	opening	dayofweek	victim_gender	totalnumbervictims	totalsuspects
1	92A296AB-D1B7-40CE-BD96-209CFF141FDA	J--CLOSED	<NA>	SATURDAY	FEMALE	1	1
2	44ACB102-5B1D-40F8-9E2B-1F823A26705D	Z--EARLY CLOSED	<NA>	THURSDAY	FEMALE	1	NA
3	2CED4B80-3AB7-46DF-BBD9-3531A0C6727A	D--VICTIM REFUSED TO COOPERATE	<NA>	TUESDAY	FEMALE	1	1
4	EEB41765-CBC3-476C-BDBE-4273B4E0CC7E	J--CLOSED	<NA>	WEDNESDAY	FEMALE	1	NA
5	F4622DF5-8274-4290-AB0E-73CB9A720905	J--CLOSED	<NA>	TUESDAY	FEMALE	1	NA
6	EF456ED0-031E-4171-8C96-29CF91BC9A9B	Z--EARLY CLOSED	<NA>	SUNDAY	FEMALE	2	NA
7	0859E5C0-4543-469D-910E-D14F603AB5BC	H--WARRANT ISSUED	<NA>	TUESDAY	MALE	1	1
8	9B091265-0352-4198-A19E-B960BDE15091	Z--EARLY CLOSED	<NA>	WEDNESDAY	<NA>	2	NA
9	D2DAF74C-1991-4E51-B81C-6BE79F7DA0F7	Z--EARLY CLOSED	<NA>	FRIDAY	<NA>	1	1
10	43EEB437-DF03-47D0-AB01-951B9EBBFA04	J--CLOSED	<NA>	WEDNESDAY	FEMALE	1	NA

NOTE: Code or output submitted without any explanations or rationale will receive zero points. You must show code and explain what your code/output mean. However, you do not need an endless number of sentences for each question. Your responses should answer each question—no more, no less.

Homework Questions to Answer

You will import into RStudio the *week1_cincy_crimes.csv* file from the Week 1 folder you downloaded for today's class, clean the data set, and perform some elementary exploratory data analysis. Data is taken from the City of Cincinnati Open Data Portal website, which you may need to read to place context in your answers.

Acquainting yourself with the data

1. Import the data set into RStudio. Show the code to import the data and then display the first 10 rows of data in the console.
2. Examine the structure of the data set.
3. Do the variable names need changed/edited? If so, how would you change them?
4. Do any variable types need changed? Explain why or why not, and change any variable types as you see fit.
5. How many missing values are present per column? Would you remove an entire observation if it contained a missing value? Why or why not? Give a good rationale for your answer.

Data cleaning

6. Look at unique values for <u>every</u> column. Do values in a column need combined, relabeled, or removed? (e.g., Are there multiple ways that a column labels missing values or genders? Should any values be removed or recoded?) Show your process for modifying values and your rationale for doing so. You will definitely spend a couple hours on this step.
7. Are there any outliers or aberrant values in the numeric columns? How do you know? Do you remove or recode them? Show your process for modifying values and your rationale for doing so. (You should leverage information from other analytics/statistics/quantitative courses you've taken either in the Business Analytics program or elsewhere throughout your education.)
8. Take care of any missing values. Do you keep them in the data set, remove observations, impute missing values, or use some other procedure? Show your processes and any rationale for doing so.

EDA (Exploratory Data Analysis)

9. Show appropriate visualizations or summaries for all *character* variables. Do any insights appear as a result of these?
10. Show appropriate visualizations or summaries for all *numeric* variables. Do any insights appear as a result of these?