Mark Hageman

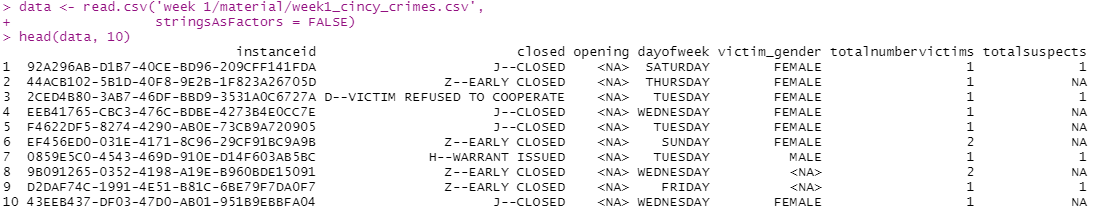
BANA 7025-001

Homework #1

October 18, 2020

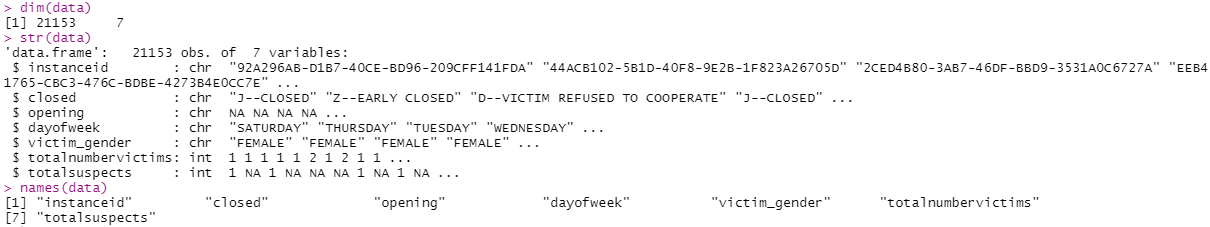
## Acquainting yourself with the data

1. *Import the data set into RStudio. Show the code to import the data and then display the first 10rows of data in the console.*

The data are imported with the read.csv() function. The stringsAsFactors parameter of this function is set to “FALSE” to prevent R from converting text data into the Factor data type. The imported data is assigned to the variable “data” and the first ten rows are output with the head() function.  


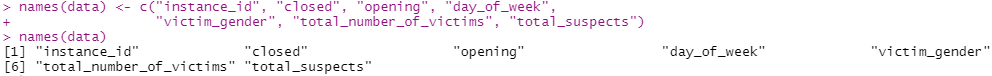
2. *Examine the structure of the data set.*

The structure of the data set is examined with three commands: dim(data) returns the number of rows and columns; str(data) returns summary information about the data including field names and data types; names(data) returns the name of each of the fields.



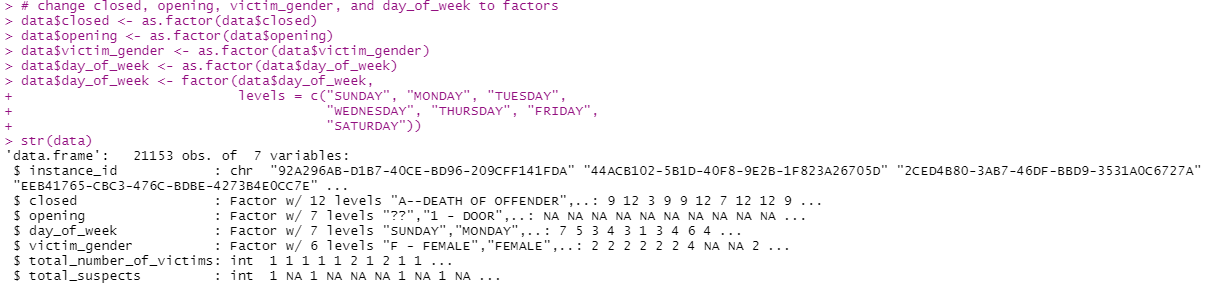
3. *Do the variable names need changed/edited? If so, how would you change them?*

The field names need to be changed to “snake case” for better readability. This is accomplished using by setting names(data) to the appropriate values.



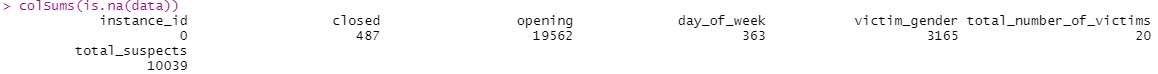
4. *Do any variable types need changed? Explain why or why not, and change any variable types as you see fit.*

Let’s change the “closed”, “opening”, “victim\_gender”, and “day\_of\_week” fields to factors. This is done with the as.factor() function. We will also re-order the day\_of\_week factor so it makes sense.



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.*

We can find the total number of missing values for each column with the colSums(is.na(data)) command.



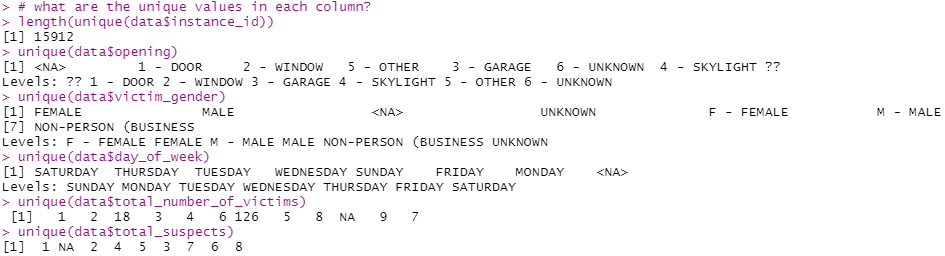
In this data set of 21,153 observations, there are only 470 complete cases. For this reason, I would not remove observations containing missing values, as this would result in a greatly diminished data set.



## Data Cleaning

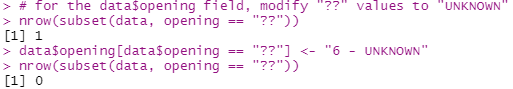
6. *Look at unique values for every column. Do values in a column need combined, relabeled, or removed? (e.g., Are there multiple ways that a column labels missing values or genders? Should any values be removed or recoded?) Show your process for modifying values and your rationale for doing so. You will definitely spend a couple hours on this step.*

We can list the unique values in each column with the unique() command, using each field as the input parameter in turn. For “instance\_id”, we will also determine the number of unique values with the command length(unique(data$instance\_id)id)).

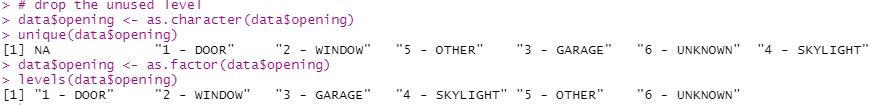


The “instance\_id” field has only 15,912 unique values. Since there are 21,153 observations and no NA values, that means there are duplicates. Not sure how to deal with this just yet, so we’ll leave it as-is for now.

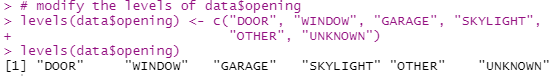
For the “opening” field there is one row with the “??” value. We will change these values to “6 – UNKOWN”.



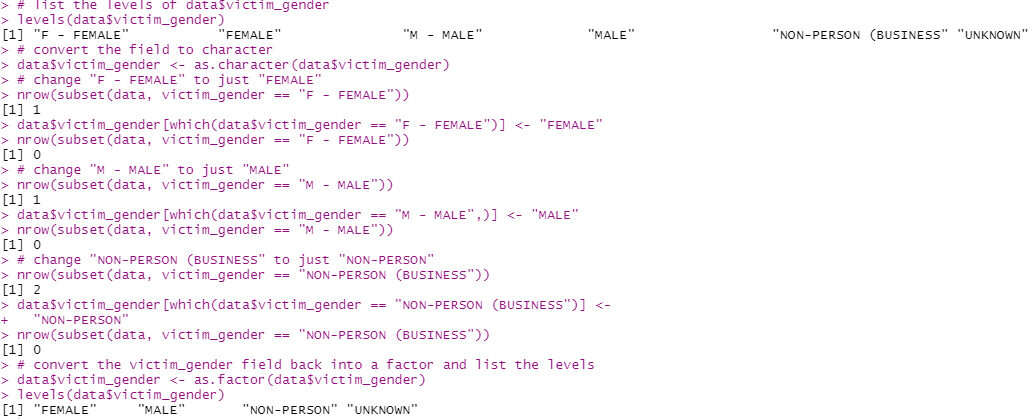
Next, we will drop the now unused level “??” by first converting the field into the character data type, then back into a factor.



Then we will modify the values of each level by removing the indexes.



For the “victim\_gender” field, there is one row each with the values “F – FEMALE” and “M – MALE”, and two rows with the value “NON-PERSON (BUSINESS”. Let’s convert the “F – FEMALE” and “M – MALE” values to “FEMALE” and “MALE” to match the other observations. We will also change the value of “NON-PERSON (BUSINESS” to just “NON-PERSON”. This will be accomplished by first converting the field into the character data type, then changing the values, then converting the field back into a factor.

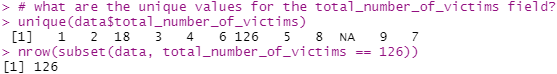


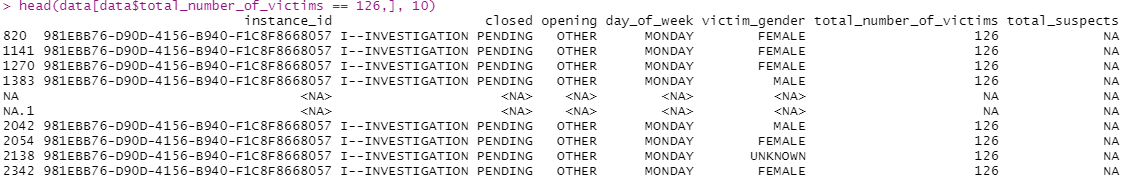
7. *Are there any outliers or aberrant values in the numeric columns? How do you know? Do you remove or recode them? Show your process for modifying values and your rationale for doing so. (You should leverage information from other analytics/statistics/quantitative courses you’ve taken either in the Business Analytics program or elsewhere throughout your education.)*

There do not appear to be any outliers in the “total\_suspects” field.

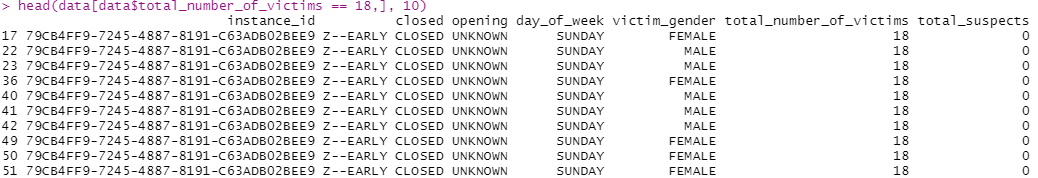


The “total\_number\_of\_victims” field does have what appears to be an outlier value of 126. However, there are 126 rows with this value, which implies that there may be one row for each victim, which seems logical given the fact that the “victim\_gender” field assumes one victim. This possibility is made stronger by the fact that each of the 126 rows have the same value for “instance\_id”. This must have been a crime involving a crowd of people.



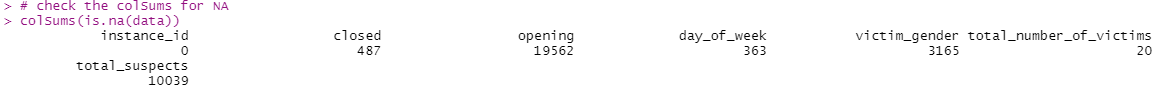


The same is true for the value of 18; all records with the value of 18 for “total\_number\_of\_victims” have the same “instance\_id” value. We will leave this field as-is.

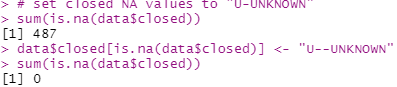


8. *Take care of any missing values. Do you keep them in the data set, remove observations, impute missing values, or use some other procedure? Show your processes and any rationale for doing so.*

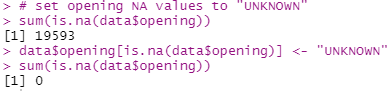
We can use the colSums(is.na(data)) command to see how many NAs remain in each column.



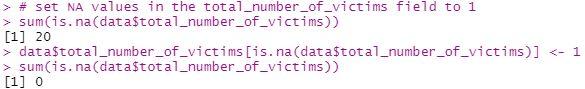
There are 487 rows with NA in the “closed” field. That field also has a value “U—UNKNOWN”, so let’s set the NAs to “U-UNKNOWN”.



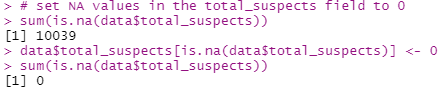
We can also set the rows with NA in the “opening” field to the level “UNKNOWN”.



It seems logical to assume that the crime would not exist without at least one victim, so we will replace the NAs in the “total\_number\_of\_victims” field with the value 1.



With the “total\_suspects” field, we will set the NAs to 0. It should be safe to assume that if no suspects have been recorded, then there probably aren’t any yet.



It is not advisable to guess at the day of the week given that we have no other data about when the crime occurred (such as a date), so we will leave the NAs in this field.

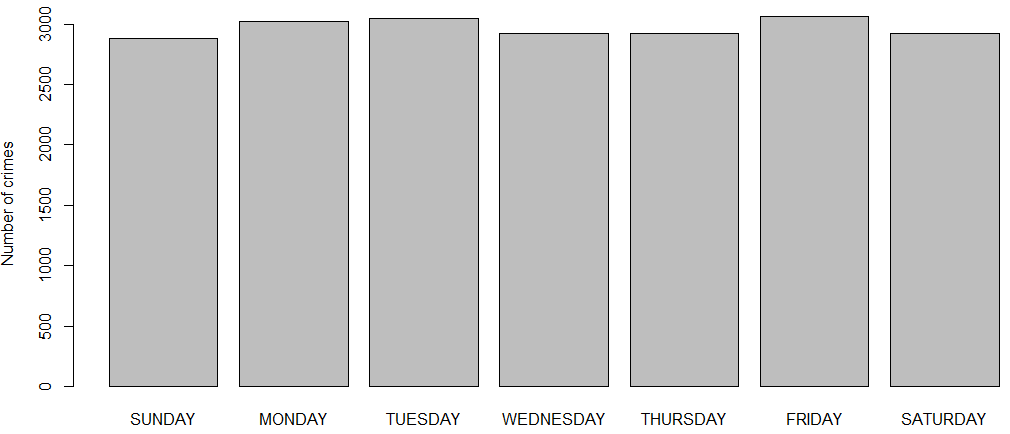
## EDA (Exploratory Data Analysis)

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

**Variable: day\_of\_week**

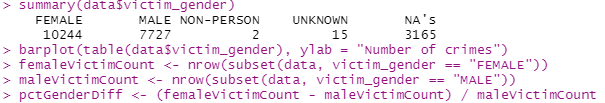
Examining the “day\_of\_week” variable does not reveal any significant insights. Crimes appear to be committed at similar frequencies for each day of the week.



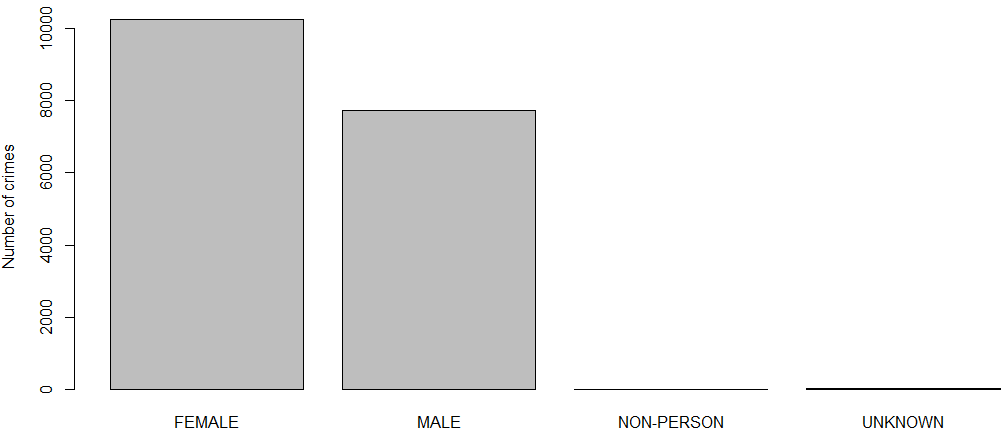


**Variable: victim\_gender**

In this data set, females are the victim in 32.6% more crimes than males.

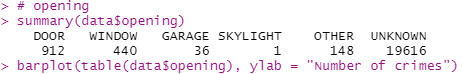


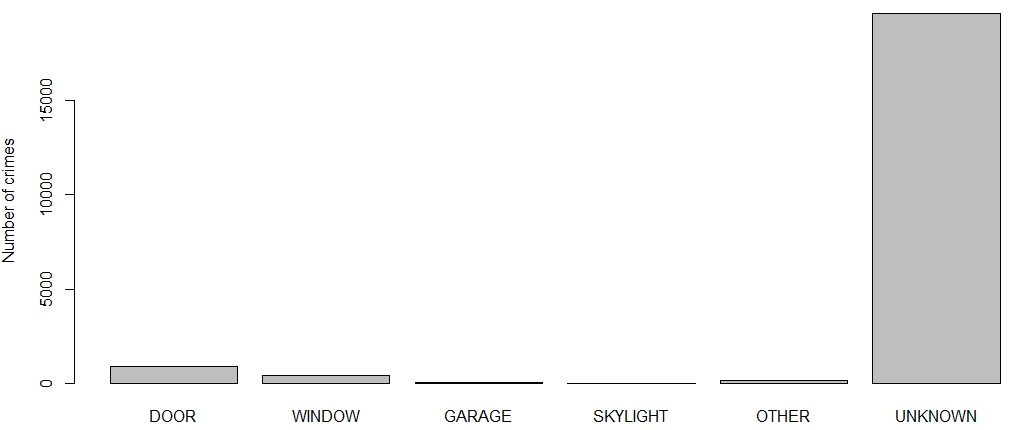




**Variable: opening**

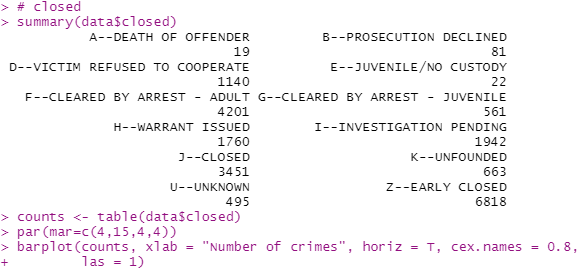
It appears that in most cases, the “opening” aspect of the crime is unknown. In hindsight, it may also be the case that the “NA” values meant “not applicable” as opposed to “not available” if the crime did not take place in a building, or if the crime did not involve unlawful entry.

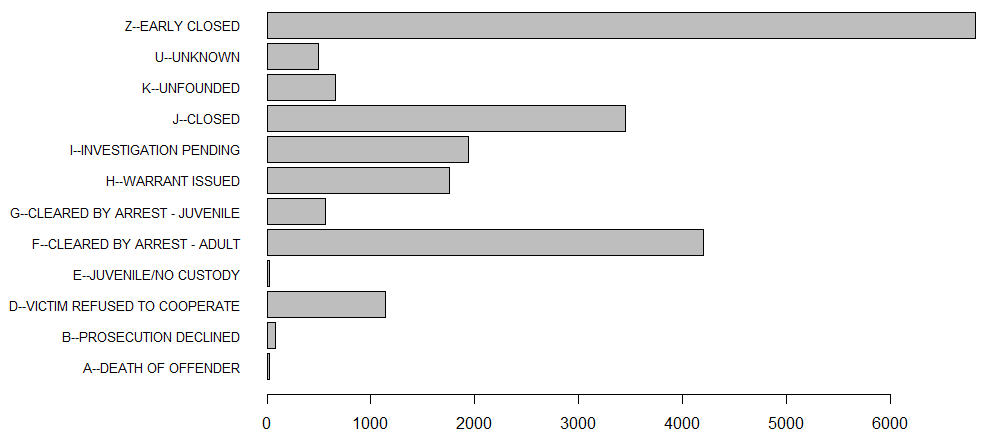




**Variable: closed**

The “closed” variable indicates that more cases fall into the “Z-EARLY CLOSED” category than any other.

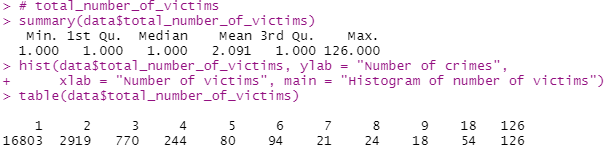




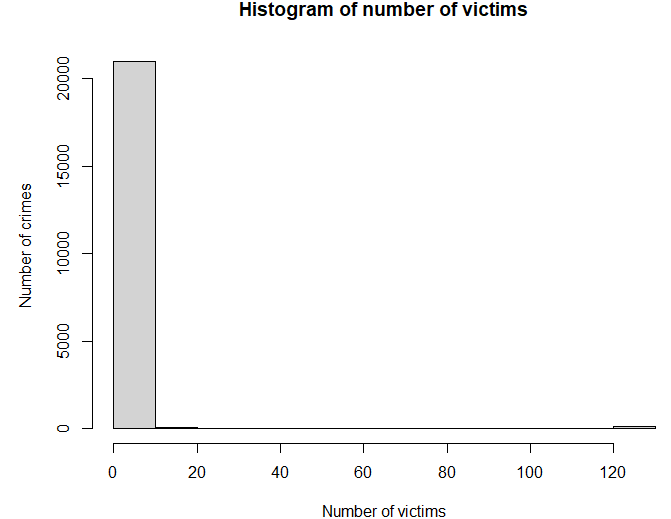
10. *Show appropriate visualizations or summaries for all numeric variables. Do any insights appear as a result of these?*

**Variable: total\_number\_of\_victims**

Most crimes have only a single victim.



It appears that each record represents an individual victim. Because a single crime may have more than one victim, the number of observations does not reflect the number of crimes, but rather the number of victims. It follows that the number of *unique values* in the “instance\_id” field represents the number of crimes.



**Variable: total\_suspects**

A little less than half of the crimes committed do not have any suspects.

