Mark Hageman

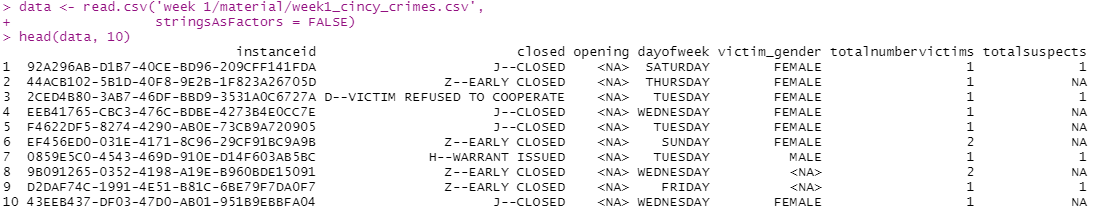
BANA 7025-001

Homework #1

October 18, 2020

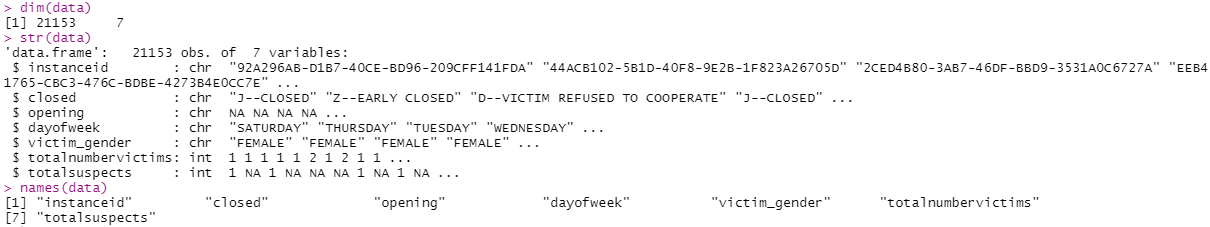
## Acquainting yourself with the data

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.  


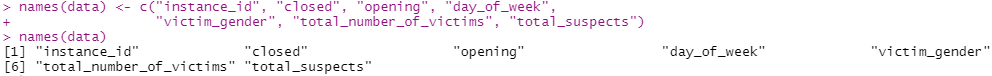
2. *Examine the structure of the data set.*

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.



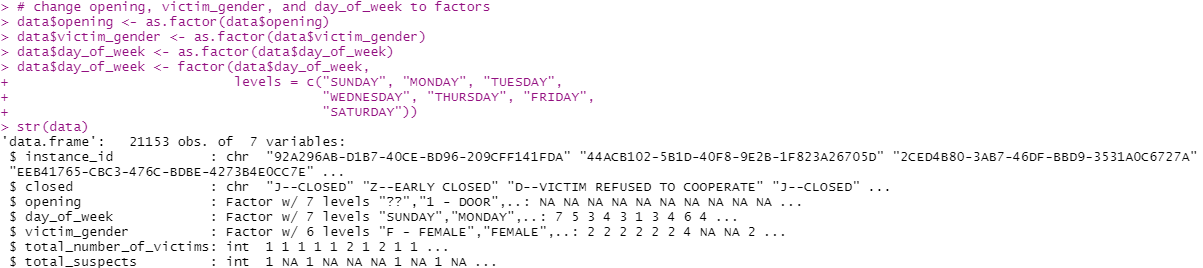
3. *Do the variable names need changed/edited? If so, how would you change them?*

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.



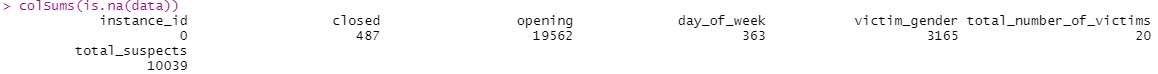
4. *Do any variable types need changed? Explain why or why not, and change any variable types as you see fit.*

Let’s change the 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.



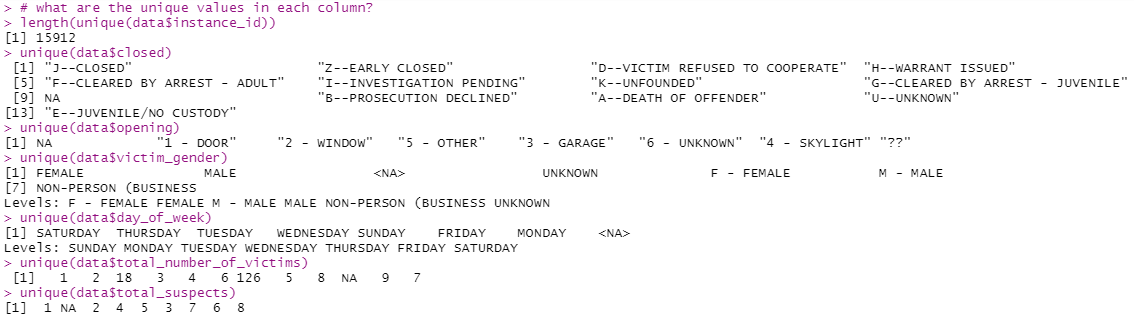
In this data set of 21,153 observations, there are only 470 complete cases. For this reason, I would not remove observations containing missing values, as this would result in a greatly diminished data set.



## Data Cleaning

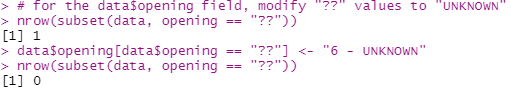
6. *Look at unique values for every column. Do values in a column need combined, relabeled, or removed? (e.g., Are there multiple ways that a column labels missing values or genders? Should any values be removed or recoded?) Show your process for modifying values and your rationale for doing so. You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

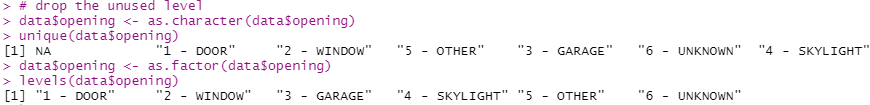


The instance\_id field has only 15,912 unique values. Since there are 21,153 observations and no NA values, that means there are duplicates. Not sure how to deal with this just yet, so we’ll leave it as-is for now.

For the “opening” field there are 19,562 rows with the “??” value. We will change the NA values to “6 – UNKOWN”.



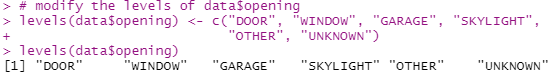
Next, we will drop the now unused level “??” by first converting the field into the character data type, then back into a factor.



Next we will replace the “NA” values with “6 – UNKNOWN”.



Finally, we will re-assign the values of each level by removing the index from the value.



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.

