***Storm and Climate Data Record (SCDR)***

Implementation Plan

Date of record:

Jan 1, 2017 to Dec 1, 2019

Principal Investigator: Camilo Hoyos

## I. Introduction

### 1.1 Overview and purpose

The city of Miami has requested an investigation into storm and crime data reported by the police department. Specifically, they are looking to see if the cost of crimes done during a storm are different than the crimes done while there is no storm. The purpose is to then identify if there is a correlation to loss and crimes during storms and crimes with the absence of storms.

### 1.2 Define why you need data analysis

Data analysis is needed to help with supporting decisions. One can simply ask “is option A or option B better?” and wonder what the true value is to their question. Once a question is asked, data analysis is the process of utilizing data to drive decisions. (Import.io, 2019) What is important is to take the bias out of an individual’s gut check and instead derive the decision based on the data they analyze.

## II. Data Preparations

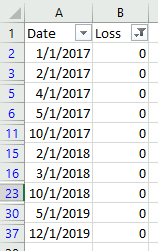
### 2.1 Name data sources

There are two data sources. First, is the dataset of crime during storms which is labeled as “crimestormQ.csv”. The second source file is a dataset of crime while no storms are present. Both of the data sets are from the Miami police department. Both datasets contain two attributes; date, and loss which is in an undefined currency but presumed to be in thousands of US dollars.

### 2.2 Filter through unnecessary data

There is a tough call to make with unnecessary or incomplete data. In the ‘crimenostormQ.csv’ file we do have entries where there is ‘0’ for loss. The issue is one cannot discern whether this means there was genuinely no loss, or if these are incomplete records. **An assumption will be made for this project that it means there is genuinely no loss**. Otherwise, the attribute “Date” is consistently in a date format of mm/dd/yyyy, and “Loss” is consistently a float value. None of the values are also outliers as they all fall within the expected parameters, such as the date being between January 1, 2017 and December 31, 2019.

*Loss entries at ‘0’ that will be assumed as genuine “no loss”*



### 2.3 Define your parameters

Since the objective is to identify if storms cause an increase in loss the major parameters to identify is if crime occurred during a storm, what is its loss, and if crime occurred without a storm, what is its loss. What’s distinct about this scenario is they are already segregated by file, so the file itself is a parameter. In addition, we have the date attribute which can be used to execute a time analysis regardless of the initial analysis request. In summary, the parameters are “storm” as a boolean, “loss” as a float, and “date” in mm/dd/yyyy.

### 2.4 Identify measurement priorities

In the previous section we touch on this briefly. In the end the priority of measurement for this analysis is if a crime occurred we capture whether a storm was present. This is currently a boolean as we have no other information on the type of storm. The other major priority is the loss to determine if the value of loss is higher or lower during a storm. Since we have the two major attributes we can identify if storms cause an increase in loss.

### 2.5 Ensure collected data fits the need

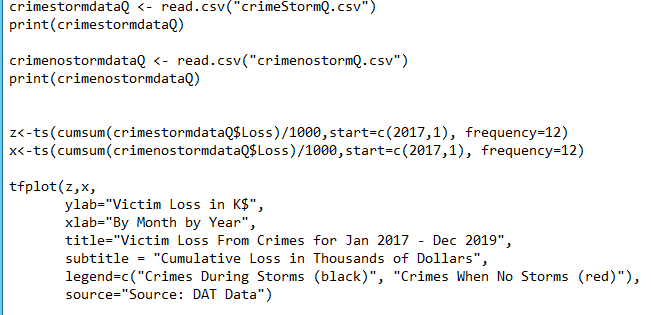
The data collected does fit the need, but at a very rudimentary level. For instance, classification of the storm could create a more meaningful analysis. It is possible that the type of storm might drive the loss more than another. Without this information it cannot be derived. We also need to know more details about the crime committed. A burglary at a home residence would be a lower loss than at a major bank. Likewise, how is loss calculated for life? How does one measure murder versus the theft of a bike in terms of loss? While the data fits the initial ask, the questions are very surface level and a more critical set of questions would cause this data set to be incapable of assisting.

## 

## III. Data Analysis

### 3.1 Identify scripts used

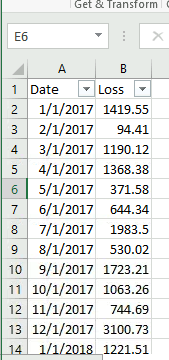
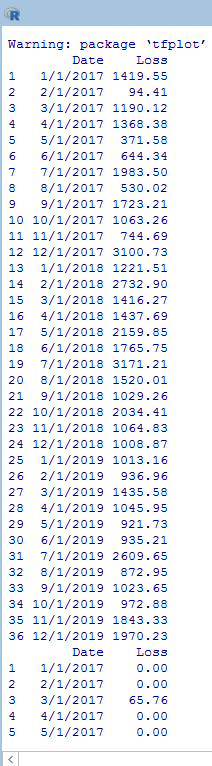
The script being utilized is executing a time series analysis on the two files. These two files then get plotted individually on a time plot. The execution of the script is as follows:



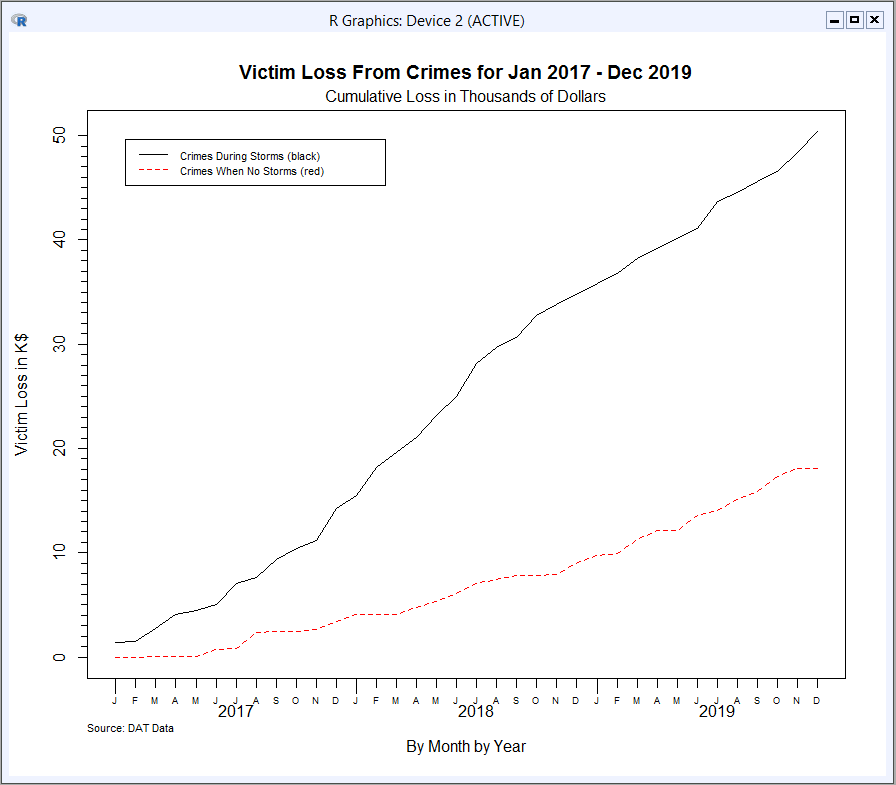
1. The first pair is responsible for reading the data set of crimes during storms and printing this data set.
2. The second pair is responsible for reading the data set of crimes without storms and printing this data set.
3. The third pair is responsible for setting two variables to hold an (x,y) map of the loss and time.
4. The final section is responsible for plotting ‘z’, and ‘x’, as lines on a time plot. You will see that the black line is crimes during storms, and the red line is crimes without storms.

### 3.2 Run the scripts to analyze the data and validate the output

With the script executed it does have two distinct outputs. First in the console there is a general print command to show the data set. We can see that the same data provided in the print matches from the original data set. This data set is small enough to do a visual review.

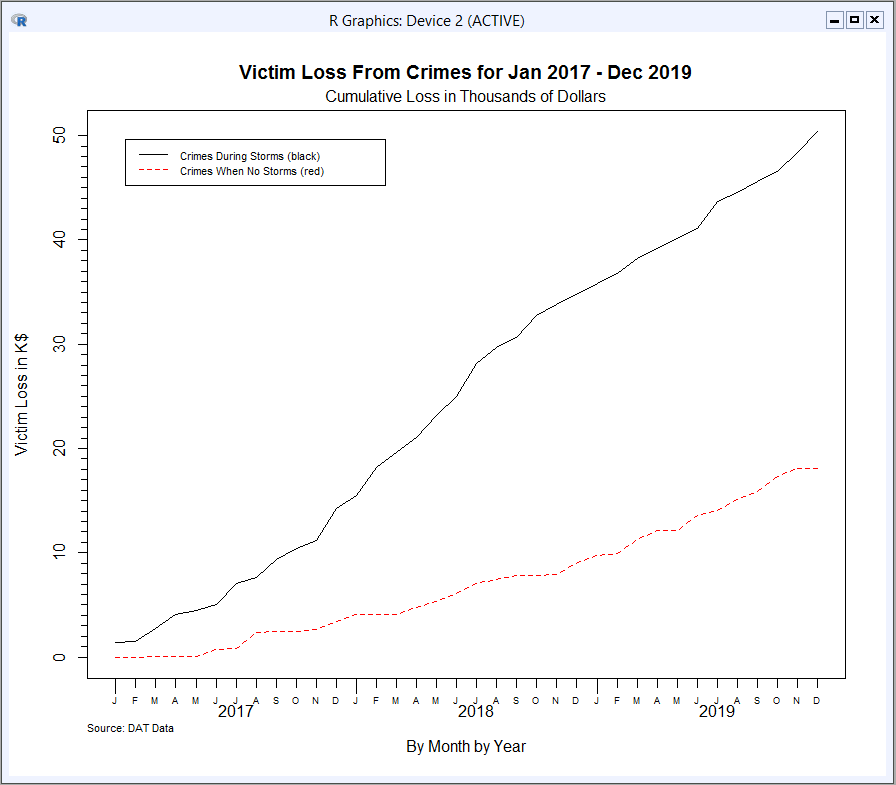


There is also the time plot provided. Upon visual inspection there are no immediate issues such as outliers or improper axis definition. The lines are easy to compare and identify the differences between storms and no storms.



## IV. Drawing Conclusions

### 4.1 Present the results of the analysis to stakeholders



### 4.2 Determine whether the problem was addressed, including any challenges and limitations

The problem was addressed by this data set. From the lines provided in the time analysis, it is clear that there is a relationship between a storm and an increase in loss. As the lines grow further apart, loss during storms increases. There are no immediate challenges with one exception; that is it is hard to discern if the 0 entries are genuine or incomplete records.

Regarding limitations, the data set does not have many attributes. It is not possible to continue to draw more insights from this dataset than what already exists. If the dataset was joined with the original StormCrimes.csv data set and showed the individual loss per crime event, then we could draw some more insights.

### 4.3 Report potential new findings

If the data set could be enriched or joined with the data set that included the storm types, and the types of crimes then this analysis could become more fine-tuned. One might be able to find a relationship between storm types and loss in dollars, a relationship between crime type and loss in dollars, or even a combination of storm types and crime types using interaction terms.

References

Import.io. (2019, October 01). Data analysis: What, how, and why to do data analysis for your organization. Retrieved February 18, 2021, from https://www.import.io/post/business-data-analysis-what-how-why/