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

Implementation Plan

Date of record:

Jan 1, 2017 to Dec 1, 2019

Principal Investigator: [Brandon Hobbs]

* 1. **Overview and Purpose**

Previous analysis of the Storm and Crime Data Report (SCDR) data resulted in a report, titled *Project One*. The police department has requested a new analysis with new data. They are now investigating the crimes during storms and examining the rising cost of crimes when storms are occurring. This report will provide the detectives with information on the next possible string of crimes. The data analyzed consists of crimes that occurred in the city of Miami between Jan 1, 2017 and Dec 1, 2019 during storm and non-storm events.

* 1. **Data Preparation**

The data sources delivered by the Detectives consists of two CSV files (crimeStormQ.csv and crimeNoStormQ.csv) that contain the monthly cumulative loss valuation for storm and non-storm periods. This data was combined into a single dataset within Excel. Validation was preformed to ensure that each CSV contained the same number of months. No invalid or unnecessary data was seen within the 36 rows of data (3 years times 12 months).

As with *Project One*, Excel, was chosen due to ease of use, its powerful pivoting capabilities, and speed at which new hypotheses can be tested. Moreover, Excel has a myriad of available functions and plot-types that can aid in analysis.

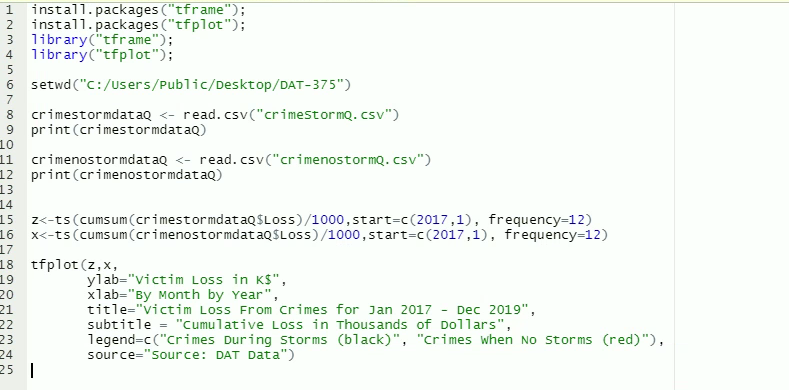
In addition to the *Date* and *Loss* attributes in each CSV file, three new attributes were created: *Year*, *Quarter*, *Season* (all extracted from the *Date* field). In the newly created dataset loss appears twice, once for no storms and once for storms. *Season* and *Quarter* attributes were added to see if any new patterns may be discovered by adding new categorical attributes.

These values were further analyzed by adding two cumulative sums: cumulative sum by month and cumulative sum by quarter.

* 1. **Validation of Raw Data and Union Data**

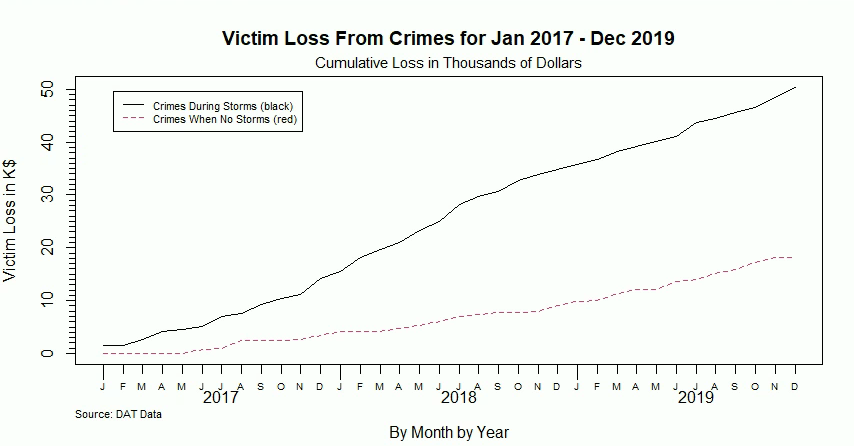
As the data was merged an R-script was used to plot the data from the raw CSV files. This R-generated plot was then compared to the plot generated from the Union created in Excel.

The script used in RStudio is shown in Figure 1.



**Figure 1: RStudio Script used for Validation**

The plot generated from RStudio using the raw data directly is shown in Figure 2.



**Figure 2: Plot Generated by RStudio from the Raw Data**

This RStudio plot was then compared to the plot of the data in Excel, Figure 3.

**Figure 3: Plot Generated from the Union Created in Excel**

The plots are the same and therefore the Excel data is considered to be un-adulterated.

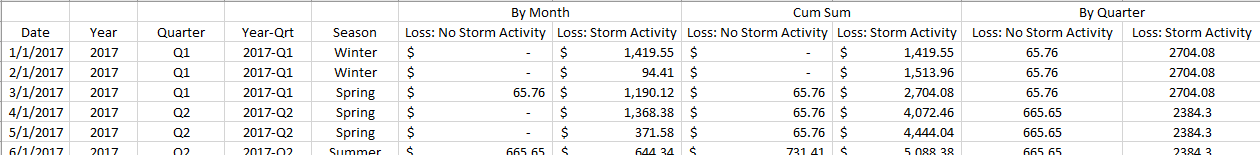
* 1. **Measurement Priorities**

As the data was being read and understood a few hypotheses were developed. The rest of this SCDR will be spent trying to prove or disprove these:

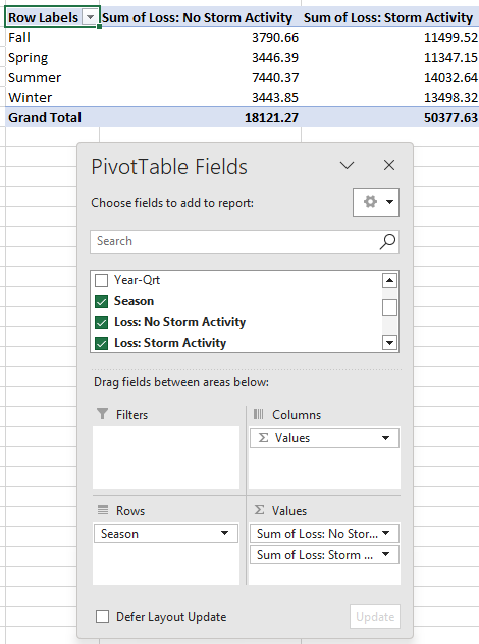
* Storm activity increases the loss valuation
* Loss, by season, will not fluctuate greatly year-on-year
  1. **Data Analysis**

As the data was analyzed in Excel no scripts were needed. However, as mentioned previously the union of the two CSV was validated. Validation was also performed on the monthly, season-based, and quarter-based cumulative sums – across all three summation techniques no-storm activity summed to $18,121.27 and storm activity summed to $50,377.63.

To facilitate the analysis and answer the hypotheses the data was labeled with headers and then a pivot table was created, Figures 4 and 5.



**Figure 4: Raw Data with Additional Attributed and Headers Labeled**



**Figure 5: Example Pivot Table Used to Generate More Granular Views and Slices of the Data**

* 1. **Drawing Conclusions**

To begin to prove or disprove the hypotheses the data was pivoted and then plotted to aid in drawing conclusions. The next sub-sections will cover the methods used to analyze the data and any plots generated during the exploration.

* + 1. **Storm activity increases the loss valuation**

To understand if storm activity brings greater loss the timeseries data was plotted. If storm activity does cause greater loss, then the curve generated from the storm data, regardless of shape, will be above the curve generated from no-storm activity, Figure 6. Note: Figure 6 was generated from the quarter cumulative sum but the curves should be the same for monthly- or season-based cumulative sums.

**Figure 6: Cumulative Sum of Loss by Storm and No-Storm Activity Categories**

Figure 6 potentially proves the hypothesis that the presence of storms increases the loss across all quarters; however, cumulative sums can hide relative changes. To further validate this hypothesis the loss per month and year are plotted by category, see Figures 7a and 7b.

**Figure 7a (left):** **Month by Month Loss by Storm and No-Storm Activity Categories**

**Figure 7b (right):** **Year by Year Loss by Storm and No-Storm Activity Categories**

Figure 7a proves the hypothesis that the presence of storms increases the loss. Except for August 2017 and a few months in 2019 the losses in the Storm Activity category are all above the No Storm activity category. Figure 7b has no periods where the No Storm activity lies above Storm Activity again confirming the hypothesis.

* + 1. **Loss, by season, will not fluctuate greatly year-on-year**

To analyze the seasonality of losses three plots were generated. The first plots, Figures 8 and 9, were created to explore if the seasonality varies across the three years – no variation would have the three lines grouped tightly and appear linear.

**Figure 8: Loss Seasonality per Year of No-Storm Activity**

**Figure 9: Loss Seasonality per Year of Storm Activity**

Loss as a function of season within the No Storm activity category shows great variation, Figure 8. All three years are non-linear and have a large spread – Fall of 2019 was six time greater than 2017 or 2018. Storm activity curves are much more linear and all generally trend upward.

Figures 8 and 9 showed that the seasonality, over year, did vary greatly, but suggest, subtly, that the summer months have the greatest loss. To investigate this idea the year dimension was removed and the loss per category was summed by the season, Figure 10.

**Figure 10: Loss per Season Regardless of Year**

Figure 10 clearly shows that loss during the summer months does indeed spike. Moreover, the ratio of loss during no storm activity to storm activity actually increases in the summer – going from ~30% in all other seasons to >50%. This suggests that loss, regardless of the presence of storms, is increasing in the summer months.

To strengthen this observation Figure 11 was created.

**Figure 11: Seasonal Loss as a Percentage of the Total**

Figure 11 further confirms that the summer months do carry the largest percentage of loss. It also confirms that summers with periods of non-storm activity will have a large increase in loss – growing to over 40%.

Figure 11 also shows that while storms are present seasonality is less important as all season-based losses remain close to an expected mean of 25%.

Figures 8 through 11 disprove the hypothesis. Summer carries the largest percentage across all slices of data – except for Winter 2017 with Storm Activity. Moreover, the seasonality over the three years of data has shifted dramatically, especially for periods of no storm activity (Figure 8).

* 1. **Parting Words**

Based upon this analysis storm activity appears to cause an increase in losses. This is most prevalent in the summer months. Moreover, the summer has the largest amount of loss regardless of storm activity. This data seems to suggest that more vigilance would be warranted in the summer.

The data analyzed for the previous SCDR, *Project One*, had data that detailed the crime type. If this data was added to the loss valuation a more thorough, targeted, analysis could be performed. That is, perhaps what is being targeted changes with the season. Knowing what is being targeted in each season could allow even more vigilance leading to deterrence and prevention.

It would also be good to understand what happened in 2019. That is, loss during periods of no storm activity have increased every year (up 162% YoY in 2019) but were down 75.8% (YoY) in periods with storm activity. This is unexpected because this was a trend reversal as there was a 144.5% increase (YoY) in 2018. Understanding what caused this would be valuable as a continued reduction in loss is desirable.

* 1. **Citations**

Hobbs, B. (2023). *DAT 375 project one summary report*. [Unpublished report]. SNHU.

Larose, D. T. (2015). Data mining and predictive analytics (2nd ed.). Wiley Global Research (STMS). https://mbsdirect.vitalsource.com/books/9781118991121