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

Implementation Plan

Date of record:

Jan 1, 2017 to Dec 1, 2019

Principal Investigator: Daniel Gorelkin

Date: June 03, 2025

## I. Introduction

### 1.1 Overview and purpose

This Storm and Crime Data Report (SCDR) will examine the new data that was provided by the police department of the city of Miami and will provide the detectives with information on the next possible string of crimes based on the data collected from Jan 1, 2017, and Dec 1, 2019, during storm events. The report will examine whether there is a rise in the cost of crimes during stormy periods compared to non-stormy crime events and will investigate possible trends in crime events and the timeframe within which such events may occur in the future.

### 1.2 Defining why we need data analysis

Data analysis is the process of converting raw data and obscured insights into meaningful information that can contribute to an informed decision-making process and forecast future outcomes based on facts and evidence, rather than assumptions, while identifying trends by examining historical data. Additionally, data analysis can assist in building prediction models and pinpointing important variables that eventually can be easily presented in the form of plots and visualizations to stakeholders. In order to conduct a successful analysis, we might need to clean our data and manipulate it by some sort of aggregation, or even apply some filters or format conversions to make it efficient without any background noise that can clutter the results.

## II. Data Preparations

### 2.1 Naming data sources

This report will be prepared based on two raw data sources *crimeStormQ.csv*, and *crimenostormQ.csv* provided by the police department. Assuming that these reports originate from a credible source, contain no errors or bias, and that they adhere to the lineage standards, we will first visually inspect and explore the data to clean and prepare it for analysis by using R as our statistical tool of choice. The files provided contain records of the loss throughout three observation years, sorted by month, where the *crimenostormQI* file provides observations where no storm events occurred, while the *crimeStormQ* file provides loss information where the weather in Miami was stormy.

### 2.2 Filtering through unnecessary data

Importing and skimming the data in the files will provide us with the following information:

A screenshot of a computer

AI-generated content may be incorrect.

This output extract shows that our data from the *crimeStormQ.csv* file (and similarly, the *crimenostormQ.csv* file) consists of 36 records, organized into two columns: *Date*, which holds the values in a string format, and *Loss*, which represents the numerical value of the loss for the specified month of the year. Similarly, we can see that our column names are structured with clean and concise naming and should not be modified. However, we can see that our Date field is stored in a character format, which should be converted into a date format in order to be analyzed. Additionally, as our records only show the first day of each month, they provide no meaning to the report, which means the deepest drill we can conduct at our analysis level is a monthly analysis.

### 2.3 Defining our parameters

To start evaluating storm-related crime in Miami, we would like to analyze whether we can indeed identify a loss increase in crime during stormy weather. To do so, as our data contains only two parameter columns, we will first convert the *Date* field to date format and ensure there are no missing or unusual values in the field records. Then, we will define our variables as *Date* (a date type) and Loss (a double type). Our *crimestormdataQ* will represent the records during stormy weather, and *crimenostormdataQ* will represent a crime observation during a clear day.

A close-up of a white background

AI-generated content may be incorrect.

### 2.4 Identifying measurement priorities

To define our measurement priorities and to help detectives prepare for future crimes during storms, we will first examine whether and we can confirm a significant increase in loss, and if such a pattern exists, we will try identify the trend and predict the next timeframe of potential extreme crime loss that could occur by analyzing seasonal observations. To do so, we will differentiate between storm and non-storm periods and aggregate our records by year, comparing each individual quarter and month while measuring the financial impact of the crime. To measure our success, we would expect to see a significant difference in crime loss between stormy and non-stormy periods, as well as identify a distinct increase in a particular quarter or month of the year.

### 2.5 Ensuring collected data fits the need

To evaluate whether the collected data can address the question of whether there is a relationship between storm events and crime loss impact in Miami, we will validate that the data fully covers the time frame in the analysis, is clean, and contains no missing or unexpected values. We will also ensure that all field values are in the desired data type and that the volume of data is sufficient for conducting the analysis. Furthermore, we will verify the lineage of the data, confirming its origin from top to bottom to ensure it has not undergone any transformation, manual correction, or distortion while traveling through the pipeline, which might lead to misleading results.

## 

## III. Data Analysis

### 3.1 The scripts used

To conduct the analysis mentioned above, we will start by plotting the data starting Jan 2017 from *crimestormdataQ* and *crimenostormdataQ* in a single time series object to compare the cumulative loss between crimes during storms and without storms. This will help us to visualize whether loss grows faster during stormy events than in non-storm conditions.  
Next, we could run a one-sided t-test to compare whether the loss is indeed higher during storms compared to a non-stormy loss, statistically confirming the results. Digging further by focusing on the stormy weather report, we can plot all the crime events (losses) from *crimestormdataQ,* and aggregate all our losses by year to discover if there is a specific year with the greatest loss, or if there is any trend we can identify. Next, if any trend were discovered, we could examine the reasons behind it in more detail and make predictions based on it.

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

To begin our analysis, we will first load our CSV files and take a brief look at the data:

A screenshot of a computer

AI-generated content may be incorrect.

Next, we will clean our data by converting the character values into formal date fields and inspect the data for empty fields:

A screenshot of a computer

AI-generated content may be incorrect.

Next, we will prepare our data for plotting and plot the "Victim Loss From Crimes for Jan 2017 - Dec 2019" plot.



Followed by conducting the Welch Two Sample t-test.

A close-up of a white background

AI-generated content may be incorrect.

After ensuring that the results are significant and the police department's assumption is confirmed, we will group our losses by year and plot the results in the "Monthly Loss Over Time" plot.



A computer code with text

AI-generated content may be incorrect.

Then, after identifying some spikes in specific months, we will break down the losses into a monthly or yearly view (or quarterly) to find the pattern and plot the "Total Monthly Crime Loss Over Time" graph.

A screenshot of a computer program

AI-generated content may be incorrect.

Lastly, by grouping all the months and summing the yearly loss for each month, we will output the "Total Crime Loss by Month (Sum of All Years)" plot, which predicts the potential next highest crime losses.

A screenshot of a computer program

AI-generated content may be incorrect.

## IV. Drawing Conclusions

### 4.1 Presenting the results of the analysis

As shown in the scripts presented above, a clear trend in crime increase during storms is evident, leading to a rising cost of crimes when storms occur, compared to the same timeframe without storm events in Miami, as suspected by the police department. From the "Victim Loss From Crimes for Jan 2017 - Dec 2019" plot, we can see that the rising cost of crimes is in a constant and steady increase, while the cumulative victim loss during storms doubles itself compared to the same year without storms. Hence, remaining reluctant to the issue and not flattening the curb could even triple the loss in the coming years.

A graph of a loss

AI-generated content may be incorrect.

We will also test this assumption of rising loss during storms by conducting a Welch Two-Sample t-test to prove the rise statistically. In this test, we will be comparing whether the mean loss during storms is significantly larger than the mean loss when no storms occurred.

We will set our **H₀ (null) hypothesis** as: The mean loss during storms ≤ The mean loss without storms, and run the test.

A screenshot of a computer

AI-generated content may be incorrect.

From the test output, we can see that since our p-value is much less than 0.05, there is enough evidence to reject H₀, and conclude with 95% assurance that losses are significantly higher during storms, and the police assumption is correct.

Next, we will sample our data to try to identify some trends and patterns in crime activity.

First, we will inspect the volume of loss aggregated by years and plot a trendline on a scatterplot to visualize the volume of loss.

A graph of a graph of a crime loss

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

A graph with a red line and blue dots

AI-generated content may be incorrect.

From these graphs, we can see that there is an evident spike during the middle of the 2018th year, and despite that, 2019 saw less crime loss compared to 2018, we can still see some seasonal behavior with an increase in loss during the second half of each year. Hence, we will examine the loss for each month individually, as shown in the "*Total Monthly Crime Loss Over Time*" time series graph.

A graph of a number of months

AI-generated content may be incorrect.

This plot gives us insights that the majority of the loss comes from the months of December, February, and July. Therefore, we will next aggregate our yearly data from the set into a monthly view by summing up the loss for each respective month:

A graph with a line and a red dot

AI-generated content may be incorrect.

Here, we can clearly see the pattern, for example, in which months we can expect the most valuable crimes to occur, namely July and December.

To conclude, indeed, the assumption of the Miami police department is true: there is a clear rise in crime loss during stormy weather compared to the regular crime level, and the trend is growing rapidly. Particularly, we can name the months of July and December to bring the greatest loss. Specifically, in these months we can expect the loss to be doubled compared to the rest of the months in a year, as can be observed from the following loss by month graph:

A graph of different colored bars

AI-generated content may be incorrect.

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

With the results of the analysis in mind, we can see that the provided dataset, which is assumed to be authentic and accurate, was sufficient to prove that losses are significantly higher during storms, and the most common months for high loss crime are as mentioned above. However, to provide more to-the-point predictions, or even identify the type of storm and the expected time of a future event, we would need to delve into data that provides hourly crime information and align it with the type of storm and the location where the crime occurred. By having that data, we could build a prediction model that would reveal with high accuracy the circumstances and the date and time of the next “big job”.

### 4.3 Reporting potential new findings and recommendations

Our recommendations, based on the analysis of the provided dataset, are to focus most police resources during the hurricane season on the month of July, which stands out as having the most significant loss compared to all the other months. Similarly, we would recommend enforcing the streets during December each year. Despite this month being excluded from the hurricane season, there is a clear increase in crime during this period, most likely due to the overwhelming number of sales advertisements before the holidays (could be tested).

**References:**

*Data Visualization: What it is and why we use it.* (n.d.). MicroStrategy. <https://web.archive.org/web/20191107194208/https://www.microstrategy.com/us/resources/introductory-guides/data-visualization-what-it-is-and-why-we-use-it>

*Gaur, C. (2024, November 21). Data Lineage | Tools and best practices. XenonStack.* [*https://www.xenonstack.com/insights/data-lineage*](https://www.xenonstack.com/insights/data-lineage)

*Knight, M. (2023, January 5). Data lineage demystified. DATAVERSITY.* [*https://www.dataversity.net/data-lineage-demystified/*](https://www.dataversity.net/data-lineage-demystified/)

Larose, D. T. (2015-04-13). Data Mining and Predictive Analytics, 2nd Edition. [[VitalSource Bookshelf version]]. Retrieved from vbk://9781118991121

*Power BI - Data Visualization | Microsoft Power Platform. (n.d.).* [*https://www.microsoft.com/en-us/power-platform/products/power-bi*](https://www.microsoft.com/en-us/power-platform/products/power-bi)

*Tidyverse packages. (n.d.).* [*https://www.tidyverse.org/packages/*](https://www.tidyverse.org/packages/)

**Storm\_and\_Climate\_Data\_Record\_RScript.R:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**