# DAT 375 Project One Template

# Data Analysis Process Job Aid

### Who should use this job aid?

The aim of this job aid is to inform and model a report for newly hired analysts in the creation of future reports to fulfill the analytical goals of the Miami police department in reference to the requested Storm and Crime Data Report (SCDR).

### Introduction

The Miami Police Department has requested the creation of a “Storm and Crime Data Report (SCDR)” that will leverage historical data collected from the city of Miami from the period of 10/1/2019 through 10/31/2019. The Miami Police Department believes strongly that there is a link between storm conditions and increase in crime and hopes to use the results of the SCDR to anticipate and minimize crimes that fall within storm time frames.

### Section 1: Type of analysis

The Miami Police Department strongly believes that crime rates in the city *increase* with storm conditions and would therefore like for our firm to create a report that will be named the Storm and Crime Data Report (SCDR). This report aims to be used as a predictive tool for crime enforcement and planning around and during times of increased storm activity. There may be some potential barriers to producing an accurate analysis such as the quality of the data and sample size. If there aren’t enough observations, the predictions made based on the data may be anecdotal at best. How was the data collected, assembled, stored, transported, etc. and was it done by professionals or supplied from “amateur” sources? Is the data stored in proper form (text, integer, date, etc.) or does it need to be transformed?

The goal of this report, or analysis, is predictive in nature. The Miami Police Department wants to use storm and crime data in order to anticipate the time frames of the increases in crime prior to or during storm conditions. For this rate increase, a linear or multiple regression analysis model would likely be the most simple and effective way of predicting this rate increase.

Logistic regression would be more appropriate if the predicted outcome was more qualitative. Decision trees and random forests would be appropriate if there were many different predictors and could be tailored to predict either qualitative or quantitative results. K-nearest neighbor, deep learning, and other such advanced methods are beyond the scope of this analysis and would likely need a larger data set than the one we have been presented. For the quantitative nature of the outcome the police department wishes, and for simplicities sake, a linear or multiple regression analysis would hopefully achieve both goals.

### Section 2: Define Parameters and collect data

Before beginning any analysis, the data must first be cleaned and prepared. We need to take a good first look at the data in its uncleaned, natural state. The data was sent to us in the form of an Excel file, or .CSV formatted document. There are 9 variables in each observation including: ID, Date, crime event ID (CrimeEventID), crime activity (CrimeActivity), storm event ID (StormEventID), storm activity (StormActivity), zone city ID (ZoneCityID), Zone, and City.

The relevant weather data variables are storm activity, and storm event ID. We are given storm activity ID, which is just a numeric key for each instance of a storm activity, not a “code” for the different types of storm activity, so it may be of use for referencing a particular “storm activity,” however for referencing a particular observation (row), the ID variable can be used which may render storm activity ID irrelevant.

Relevant criminal data, at initial inspection, may only include the crime activity variable. Crime event ID may seem to be relevant based on the title alone, but after inspecting the raw data, all values (in fields that are complete) appear to be the same and would not be useful for identification purposes. Again, for referencing a specific observation or row, the ID variable would likely be best suited.

In summary, the relevant variables for storm and criminal data appear to be “crime activity (CrimeActivity)” and “storm activity (StormActivity)” alone. To really delve deeper into the “analysis” side of things and determine what other data will be needed for our report, we first need a little context. The ID variable will be the best for referencing a specific observation (row), so we may want to keep that. Date is also likely to be of interest for purposes of establishing timelines *or* for grouping crimes together by something other than the type of crime; its usefulness is yet to be seen but it can’t be eliminated without more context. The Zone and City variables may be useful for “hotspot” determination or determining which areas of the city/county have the highest rates of crime during these times (if that is of interest to the Miami police department, and it should be). The zone city ID variable is just a numeric “code” for the city variable, so it may be more useful to use it instead and eliminate the city variable. While doing some quick research however, one can see that the Miami police department may not be interested in many of the areas in question as their jurisdiction does not apply there. Apparently, the Miami Police Department (MPD) only has jurisdiction over a few municipalities in the “city” of Miami, whereas the Miami-Dade Police Department has jurisdiction over the whole Miami-Dade County, minus the MPD areas (Wikipedia contributors, 2022). For this report, all municipalities will be included in the analysis.

### Section 3: Tool Selection

For the analysis tool of choice, I will use R-studio, however, Excel could likely also produce excellent results. R is very flexible and has many of the functions needed to perform these analyses included and is easily automatable, especially if the data is already cleaned and formatted prior to importing. R also has many powerful tools for visualization including bar charts, line graphs, scatterplots, and others that may end up being used here.

For visual aid representation, several tools in the R library could be useful, as mentioned above. For showing the distribution of crime per municipality, a pie chart or perhaps a bar chart could be useful. With that said, the number of municipalities in the Miami-Dade area may make them congested. For a visual representation of the rate of increase in crime I would do a simple line graph or column chart contrasting the number of crimes *per date* against the actual dates. For *types of crimes* and their rates of change over time/dates and amounts, a dual axis chart may be best recommended. For the types of crime per area against type of storm activity a dual axis chart may also be best (Oetting, 2022).

### Section 4: Validation

To clean and prepare the data for analysis with R later, I will use MySQL Workbench to import and filter out irrelevant data. I will start by dropping the CrimeEventID column with the following command: “alter table dat375.stormcrimes drop column CrimeEventID” which removes all of those identical values and then removing observations that are incomplete by selecting *all* records in the data set via “select \* from dat375.stormcrimes” and then using “where StormEventID <>0;” scripts to filter out all records that do not have empty or placeholder ID values, which returns only 167 rows of the original 250 rows of data.

From here, I could probably remove the City column as well, leaving only the ZoneCityID column to identify the different municipalities by, but for readability I have left them in and will just omit them from the analysis to be performed later. The data in its cleaned form contains no empty cells, still has the ID column to identify individual rows, dates, the type of crime committed (CrimeActivity), the type of storm activity (StormActivity, and its identifier StormEventID), the Zone, and ZoneCityID (with city left of course).

Wikipedia contributors. (2022, November 18). *Miami-Dade County, Florida*. Wikipedia. <https://en.wikipedia.org/wiki/Miami-Dade_County,_Florida>

Oetting, J. (2022, August 11). *14 Best Types of Charts and Graphs for Data Visualization [+ Guide]*. https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization