# DAT 375 Project One

# Data Analysis Process Job Aid

**David France 5/25/23**

### Who should use this job aid?

This job aid is intended for new hires so they can see how we address various data analysis problems and techniques.

### Introduction

The Miami Police Department has identified a link between crime and storm activity. They have asked us to provide a report of storm activity and crime to help test this hypothesis and anticipate future crime. This includes a file containing all crime activity that occurred at the same time as storm activity. It also includes a count of the number of crimes grouped by the type of storm activity.

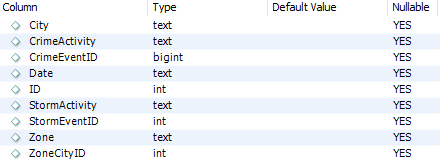
### Section 1: Type of analysis

There are two main types of analysis when presented with a new data set – exploratory analysis and hypothesis testing. Exploratory analysis is appropriate when we don’t have any thoughts or ideas about relationships within the data. This might be used when presented with a new data set and simply asked for any insights we might find. Hypothesis testing is appropriate when we have some idea or direction of relationships within the data and want to see if we are correct (Larose & Larose, 2015). Because the Miami PD has defined the relationship for us, we will use hypothesis testing in this example.

### Section 2: Define Parameters and collect data

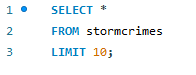
Defining the parameters means choosing the correct variables to include in the analysis. To do this, we can work backwards from what data is needed and choose our variables from there. The data we need for the report is a CSV file containing all records for crimes that occurred at the same time as a storm activity and a count of crimes grouped by storm activity.

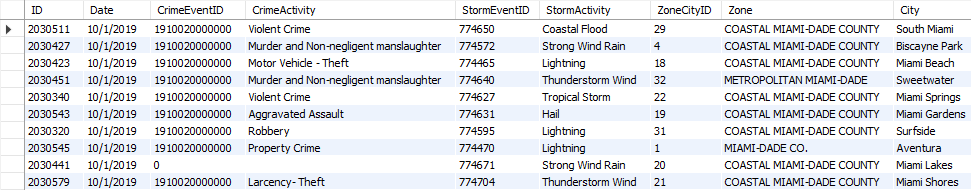
To decide which variables to include, we should look at the schema for the data:



Because we need a file with only records that have both crime and storm activity, we will have to include either CrimeActivity or CrimeEventID, as well as StormActivity or StormEventID. It is good practice to use ID columns, as they will contain unique values, so we will use CrimeEventID and StormEventID to filter our data. Notice how both columns are nullable, we will have to write a query that ensures we only take records where both columns have a value.

We need to include a count of crimes grouped by storm activity in our report, so it looks like we will need to include the StormActivity parameter in our analysis. Let’s double check to make sure that makes sense. We can select the first 10 rows of data to see what each column’s data looks like:





The StormActivity column contains activity types, such as ‘Coastal Flood’ and ‘Hail’. This matches what we are looking for so we can use this column for our table. We can use the CrimeEventID from our first query to count the number of crimes and group them by StormActivity.

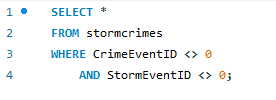
### Section 3: Tool Selection

There are many tools available for data analysis, including R, MySQL, and Excel. For this analysis MySQL makes sense because it is fast, easy to use, and our analysis is relatively simple. If we needed to go more in depth, perhaps creating models to predict when a crime would occur, R is a much more powerful tool and would be a good choice. Excel could work for this data set, since it is only 250 rows, but it struggles with larger sets. It also takes a few extra steps to filter out the data we are looking for. With MySQL we only need to do two queries.

We need to provide a visualization of the count of crimes by storm activity. The standard for this type of information is a simple bar graph. It conveys information quickly and makes it easy to compare different categories at a glance. We will use Excel to create our graph because it is easy to quickly customize a graph for simple data, such as what we have. R has the ability to customize any graph and would also be an acceptable choice.

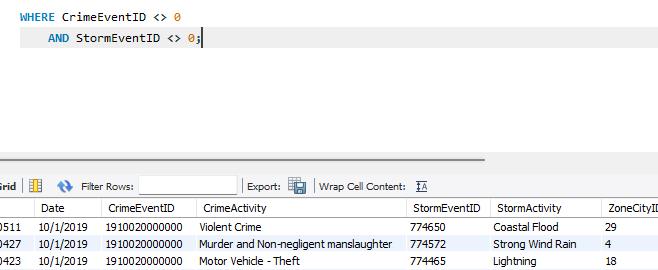
### Section 4: Validation

Now it’s time to run our analysis in MySQL. We will start by filtering out rows that have either CrimeEventsID = 0 or StormEventsID = 0. To do this we will select all the columns where both of those variables are not 0:

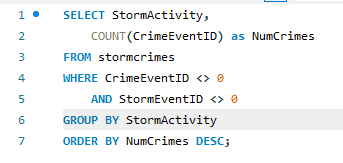


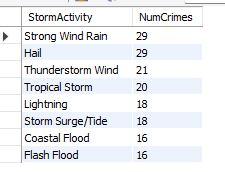


Now that we have our data, we can export it to a CSV. This process is simple on MySQL:



If we wanted to do extended analysis in MySQL using this result set, we could load this data into a new table and work with that. We could also open it in Excel and work with it there. For instance, we could use an Excel pivot table to count the number of crimes, grouped by storm activity. Because we are already in MySQL, we can simply do that here:





Notice how we select the StormActivity parameter and a COUNT() of CrimeEventID, which is aliased as ‘NumCrimes’, then GROUP BY StormActivity. I also chose to ORDER BY NumCrimes to make a clean table. From here, I used Excel to create the graph. I simply copy/pasted the table into an Excel sheet and turned it into a bar graph:

Looking at the graph, it’s easy to see that ‘Strong Wind Rain’ and ‘Hail’ contain the highest number of crimes. The Miami PD will likely take that information and increase police presence when those conditions are predicted. Conversely, the three flood-related categories have the fewest number of crimes. The Miami PD shouldn’t be as concerned in hurricane conditions compared to typical storms. We should conclude our analysis by following up with the department to make sure our analysis is reasonable and that we’ve provided the information they are looking for.

# References

Larose, D., & Larose, C. (2015). *Data Mining and Predictive Analysis.* Hoboken: John Wiley & Sons, Inc.