# DAT 375 Project One Template

# Data Analysis Process Job Aid

### Who should use this job aid?

This job aid can be useful for many audiences, but the particular audience this job aid is aimed at is newly hired data analysts at our consulting firm. This job aid is to be used as a reference to these new hires using my current project as a scenario for the data analysis processes that our organization uses to work with clients. The purpose of this job aid is to improve job performance and employee productivity.

### Introduction

The first step of an analysis project is to determine the problem statement. You must ask questions to ensure you know exactly what is being asked and to be able to determine what next steps must be taken. After determining the problem statement and deciding the type of analysis that will be done, you must define the parameters and then collect your data. Choosing what tool to use is also an important step. Many tools are available for an analysis project, so you must determine which tool will best suit your data set and parameters. After completing your analysis, you must also validate your conclusions. This can be done by sampling the data set and then validating with calculators or graphs. You must determine which validation tool will be the most beneficial for the data set and type of analysis.

### Section 1: Type of analysis

The Miami police department has requested a Storm and Crime Data Report (SCDR) using historical data from the city of Miami from 10/1/2019 to 10/31/2019. They believe they have made a connection linking an increase in crimes with storms, so they are looking for a predictive analysis to anticipate future timeframes for these crimes.

A predictive analysis uses historical and current data to predict future probabilities. This SCDR may prove ineffective due to having such a small sample of data (only one month), but the Miami police department hopes to use it to predict future crimes in October. This predictive analysis would be better suited with a larger data set, to include October data over many years, or even the full calendar year over many years.

### Section 2: Define Parameters and collect data

After determining the type of analysis to be completed, you need to define the parameters. The parameters in this analysis include ID, Date, CrimeEventID, CrimeActivity, StormEventID, StormActivity, ZoneCityID, Zone, and City. You must narrow down the data you need to work with only the data that will assist in your analysis. For this predictive analysis, we would want to collect and analyze the data for Date, CrimeActivity, StormActivity, and ZoneCityID. This will allow us to analyze what crime occurred during what storm activity on what date at what location.

We know that this data set is only for the dates 10/1/2019 to 10/31/2019. The different values of CrimeActivity include Violent Crime, Murder and Non-negligent manslaughter, Mother Vehicle – Theft, Aggravated Assault, Robbery, Property Crime, Larceny – Theft, and Burglary. The different values of StormActivity include Coastal Flood, Strong Wind Rain, Lightning, Thunderstorm Wind, Tropical Storm, Hail, Storm Surge/Tide, and Flash Flood. We will use this data to analyze what crime activity occurred during which storm activity to predict this for future timeframes.

### Section 3: Tool Selection

Selecting the appropriate tool to use for your analysis is extremely important. We need to ensure that the chosen tool is capable of processing the size of the data set we are working with and that it offers all commands and/or functions needed to perform our analysis. I chose to use MySQL for this analysis and Excel for visualization. Using MySQL, I can narrow this down using SELECT FROM WHERE commands to only rows that contain a crime activity and a storm activity on the same date which leaves me with 167 rows to analyze. I would export this data set to Excel to create pivot tables showing the analysis results.

Using Excel to create pivot tables can be a very useful tool for visualization. You can choose to create bar charts, line graphs, pie charts, and more. For this analysis I chose to use bar chart. This bar chart I made as an example simply shows the count of storm activities grouped by crime activity. You can take this even further and add the dates and city zones also.

Chart

Description automatically generated

### Section 4: Validation

Using MySQL, you can write scripts to extract the data you are looking for. You can use SELECT FROM WHERE commands to pull exactly what you are looking for, and you can use GROUP BY and ORDER BY commands to organize the data in a format that makes it easier to view. The following is an example of a script you can use to view the count of storms and crimes happening on the same day by zone.

Graphical user interface, text, application, email

Description automatically generated

Another script that can be used to filter the needed data is to use a SELECT FROM WHERE command to pull the data that shows the crime activity during a certain storm activity grouped by Zone. The example below will allow you to see exactly what crimes happened in what zone while there was a coastal flood. This script can be changed to any storm activity, or you can reverse it and pull the data based on crime activity instead of storm activity, which is also shown below.

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Another way to validate your data is to write a script that shows you the count of crime activities during each storm activity per zone. Here is an example of that.

Graphical user interface, text, application, email

Description automatically generated

With this script you can see for example that during a Costal Flood in Coastal Miami-Dade County, there were 12 crime activities. You can also see that during a Flash Flood in Inland Miami-Dade there were 2 crime activities. You can also take this further and add in the Date column to view the date that these storms and crimes occurred.

Graphical user interface, text, application, email

Description automatically generated

When validating your data and ensuring that you have the correct parameters, you always want to double check your data sampling with your stakeholders. This will help determine if your parameters are in the scope of your project and will help clear up any confusion or questions you may have.