# DAT 375 Project One Template

# Data Analysis Process Job Aid

### Who should use this job aid?

This job aid is for those who will be preparing data for analysis.

### Introduction

Preparing data to execute analysis is one of the key parts of the process of analyzing data, if not the most important. To do this we must identify the necessary data, which variables are relevant to the analysis, and clean the data before analytics can be run. Once the data is cleansed, the analysis can be executed to identify relationships between variables.

### Section 1: Type of analysis

The current project is looking to identify any patterns between crimes and storms in the Miami area. The data set used is historical data from the month of October in 2019 from the city of Miami that indicates the crimes and the type of storm for each date. This data has 250 rows of information, including ID, Date, CrimeEventID, CrimeActivity, StormEventID, StormActivity, ZoneCityID, Zone, and City.

The data must be cleaned before it can be used, so the identification of potential data errors will be important. The first step is to identify if there is any missing data. Sorting the data in the csv in Excel, it can be seen that values for CrimeActivity are null when CrimeEventID is equal to zero, and vice versa. This is the same for StormActivity and StormEventID. So, there is no obviously missing data.

Also within Excel, it can be seen that there are no errors with the information contained within CrimeActivity or StormActivity, as there are no duplicates identified, or misspelled entries, when it is filtered (see Appendix A). There are no apparent errors in the data set that would impact the queries that are to be performed. The data appears to pass the most common data cleansing techniques using Excel (Digital Vidya, 2020), and can now be analyzed with respect to the variables of interest with an expectation of accurate information used for the results.

### Section 2: Define Parameters and collect data

The client for this task, the Miami Police Department, wants to identify relationships between crimes and storm activity. For the storm information, the StormEventID does not provide any valuable information. It is enumerating the storm events uniquely without repeating and is not a valuable key for data. The StormActivity column provides a description of the type of storm that was occurring on a given day and would be the valuable input for this analysis.

The crime data has two inputs, the CrimeEventID and CrimeActivity. The CrimeEventID does not have useful information, as all values are either 0 or 1910020000000, where 0 means there is no crime and the other value is the same for all crimes, regardless of the description. The CrimeActivity variable will be the useful input for this analysis.

Since the client is seeking this information in order to anticipate timeframes for crimes in the future based on the storm activity, the valuable inputs identified for storm and crime, StormActivity and CrimeActivity, will be the most important variables to include. Additionally, this information can be consolidated by creating a count of the number of a given crime during a specific storm (see Appendix B). This would allow a more accessible presentation of a table of data for the client.

### Section 3: Tool Selection

SQL would be the preferred way to aggregate the desired data as it makes the sorting and creation of tables of data remarkably simple. Should additional data or information wish to be added to the analyses, the creation of additional tables, linking those tables, and having an overarching schema of interrelated data will be much easier. As well, SQL is designed to handle much greater volumes of information quickly compared to Excel. It also allows for multiple individuals to work on the same project simultaneously without locking out others, as Excel does (Chen, 2016). Additionally, SQL queries allow for data to be exported to a multitude of other environments for visualization or incorporation of other programming tasks.

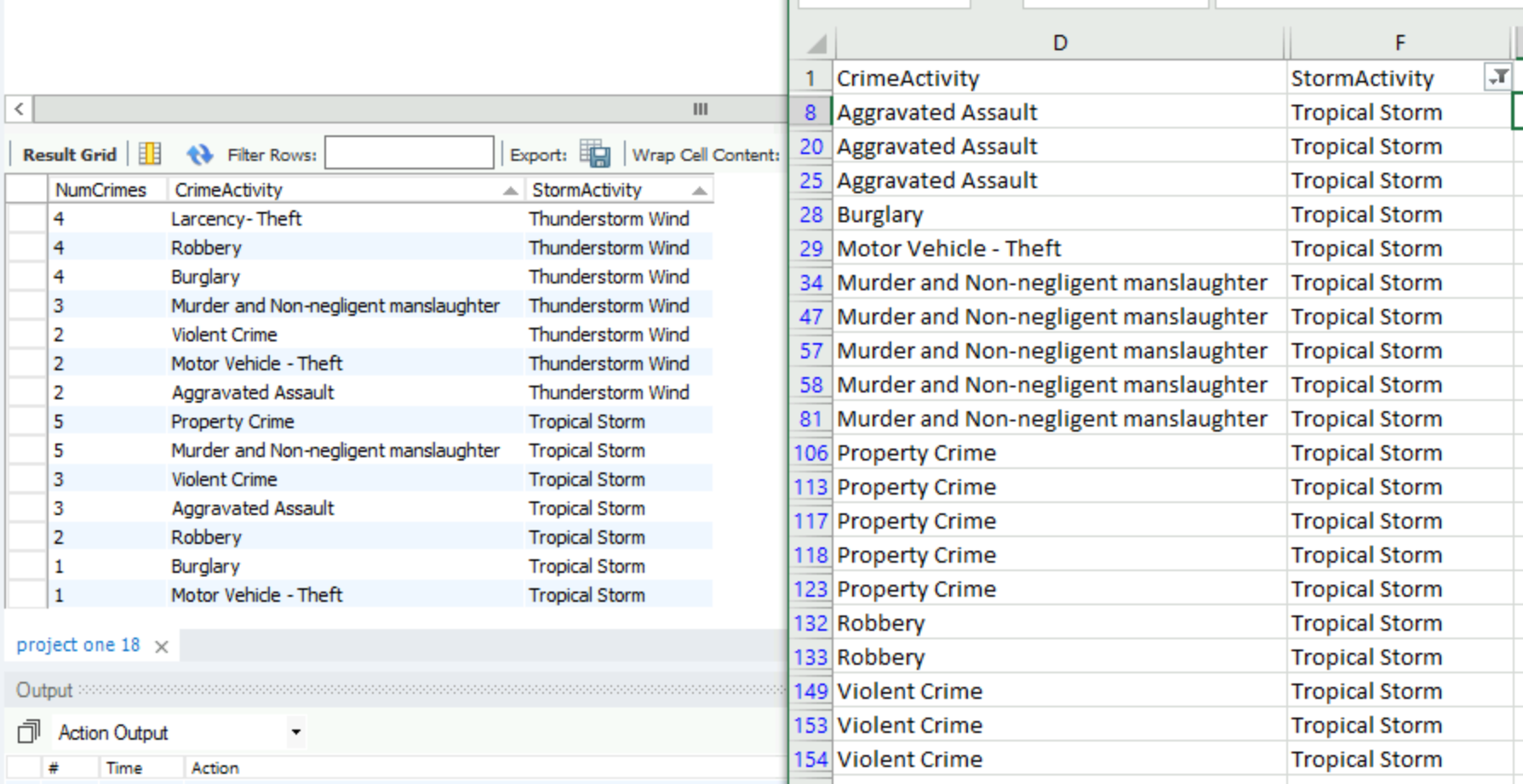
Once the data is aggregated with the parameters we have identified, this data can be visualized using any of several methods. The easiest way to have a highly effective visualization is to use Excel. This allows a choice of a multitude of visualizations to represent the data we have queried (Sharma, 2020). Since we have been working with Excel already through the process of cleaning and validating the data, creating a pivot table in Excel utilizing the query data exported as a csv file will be an easy transition. As well, the facility of creating these tables to express whichever parts of the data makes it a less cumbersome tool (Microsoft, n.d.). A pivot table of this query are available to review in Appendix C, which shows the breakdown of crimes by the weather event as well as the overall number of crimes during each weather event, utilizing the same csv from one query in MySQL.

### Section 4: Validation

The scripts to filter the data from the larger data set will be a select query, choosing the CrimeActivity, StormActivity, and a count of the number of each CrimeActivity. Refer again to Appendix B for elaboration on this script and the output. This allows for the number of each crime type to be summed and sorted with regards to the type of weather experienced at the time. The other variables are not relevant to the request by the client and as such that data, such as the date or city, are not included.

The data that is pulled does not include crime incidents when there is no weather event, or when there is a weather event with no crimes reported. As the client wishes to use this tool to predict the possibility of future crime events during specific types of weather, only the relationship between those variables is identified for the query. Should other variables be requested, such as the city or the date, additional queries on the same larger data set can be executed.

Sampling a random variable in the larger data set by filtering will show that the queries we have performed match the data set. As can be seen in the side-by-side comparison below, the query indicates there were 5 property crimes, 3 violent crimes, and 1 burglary during tropical storms. Filtering the csv file of the larger data set to show the crimes by the filter of Tropical Storm, it can be seen there are 5 property crimes, 3 violent crimes, and 1 burglary in the data set. This confirms that the query is accurate.



References

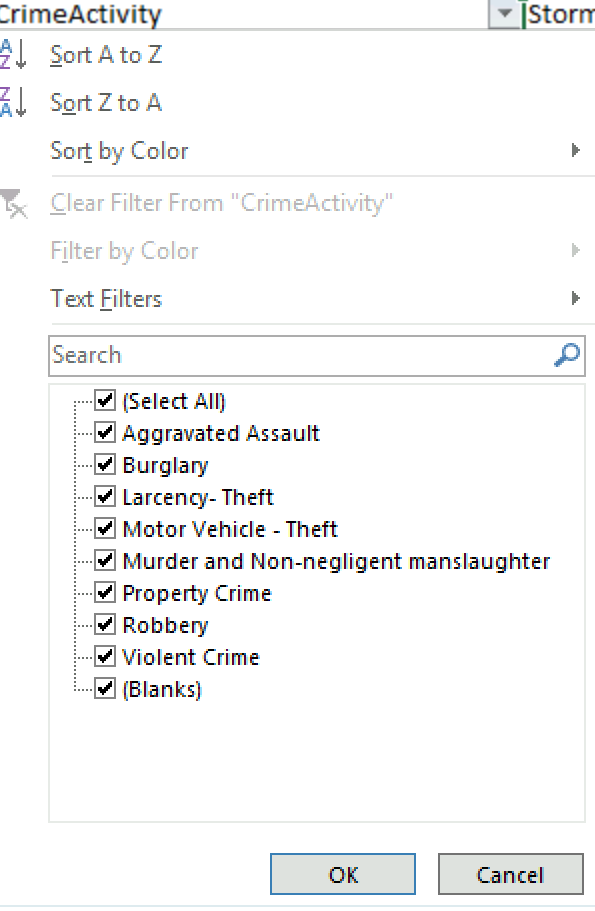
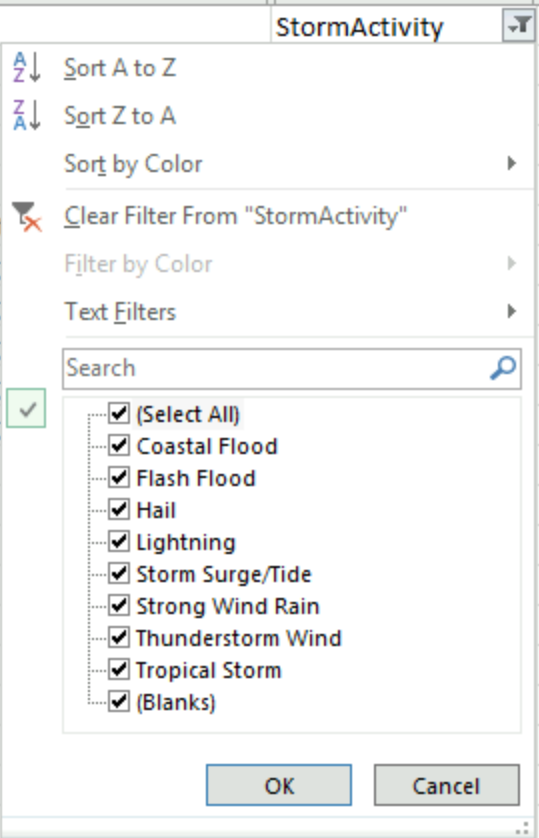
Chen, A. (2016, August 16). *3 reasons why you should prefer SQL server over MS excel for your small business*. Data Recovery Blog. https://www.datanumen.com/blogs/3-reasons-prefer-sql-server-ms-excel-small-business/

Digital Vidya. (2020, March 27). *8 Ways to Clean Data Using Data Cleaning Techniques*. https://www.digitalvidya.com/blog/data-cleaning-techniques/

Microsoft. (n.d.). *Create a PivotTable with an external data source*. Retrieved September 27, 2020, from https://support.microsoft.com/en-us/office/create-a-pivottable-with-an-external-data-source-db50d01d-2e1c-43bd-bfb5-b76a818a927b

Sharma, H. (2020, June 24). *Best Excel Charts Types for Data Analysis, Presentation and Reporting*. Optimize Smart. https://www.optimizesmart.com/how-to-select-best-excel-charts-for-your-data-analysis-reporting/

Appendix A

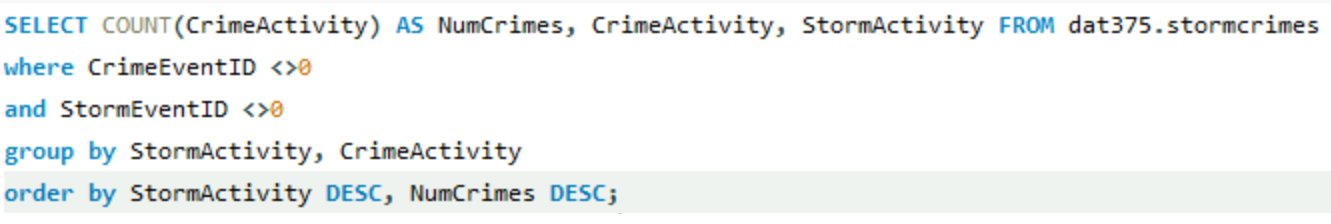
 

Appendix B





Code used:



Appendix C

