Project 1: Training Guide

DAT-375

Brandon Hobbs

March 22, 2023

To guide new hires or those new to analyzing data a simple data set will be analyzed. Commentary will be provided on the choices and decisions made to effectively analyze the data.

The data set to be analyzed contains 250 rows and 9 columns of data and represents historical data from the city of Miami from 10/1/2019 to 10/31/2019. The data consists of storm and crime data from this time period. The analysis will seek to determine if there is an increase in crimes with storms present.

This guide will provide commentary on the following topics: suitability of analysis type; parameter definition and selection; selection of tools; validation.

# Section 1: Type of analysis

Analysis comes in to two forms: qualitative and quantitative. From Smith (2022):

Qualitative analysis relies on thick description and deep understanding of the subject being researched, obtained from in-depth interviews, observations, and/or close readings of text. This type of research typically looks at case studies and can be used to understand local phenomena.

Quantitative analysis instead relies on the statistical analyses of numerical data obtained from surveys, experiments, or administrative records. From this, inferences can be made and correlations between variables analyzed to understand more generalized phenomena.

The analysis we are being asked to conduct falls in between both. We are being asked to use numerical data that corresponds to categories, e.g., crime occurred or storm activity occurred, and see if there is in fact a correlation between the two phenomena.

Fitting a scatter plot, e.g., crime activity vs storm activity, with a best fit line and then calculating the R2 value is quantitative, but determining the logical categories and then assigning the events to them is qualitative – takes a bit of “gut instinct”.

# Section 2: Define parameters and collect data

Before defining the parameters needed for analysis it is important to set up the hypothesis we will be testing for. Because we are being asked to determine if storm activity corresponds to higher crime rate, the following hypotheses were created:

* More crime occurs when story activity is present
* Less crime occurs when no storm activity is present
* Storm activity will have a higher coefficient of determination to crime activity than a random variable

Based on these hypotheses the data parameters needed will be: crime activity type, storm activity type, categorical crime activity (a Boolean variable), categorical storm activity (a Boolean variable), and a random variable (for example, day of the week).

# Section 3: Tool Selection

Because, correlation is being requested, i.e., does higher crime correlate to storm activity, the ability to create scatter plots, trendlines, calculate statistical values, and create new data types, i.e., the Boolean categorical variables, the tool chosen needs to be flexible.

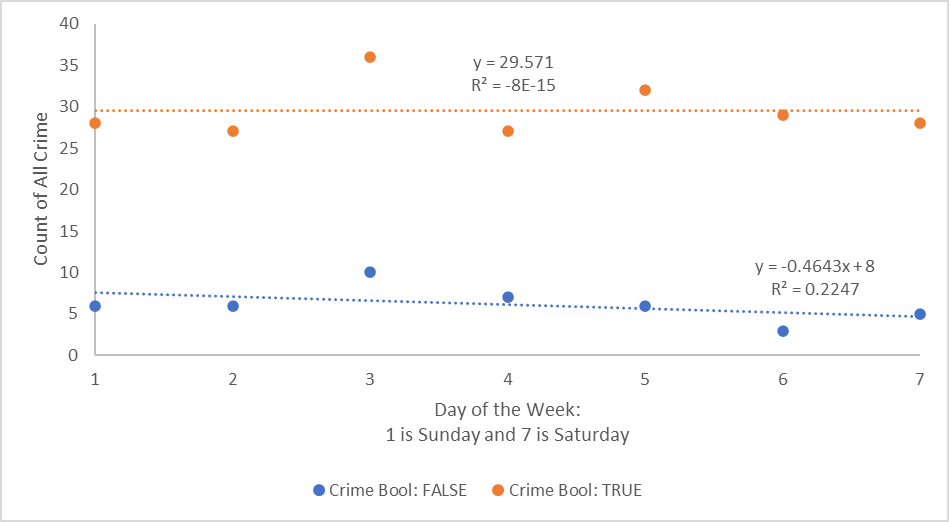
A valuable tool would be Python and its panada library, to help with data blobs, and the Matplotlib library, to create the plots.

R is also capable of doing this type of analysis. R is very effective in analyzing categorical data with its ability to create strip, jitter, and swarm plots. In a deeper analysis involving Classification and Regression trees (CART) R would be invaluable. From the EPA (2023), “CART analysis is used in data exploration to classify systems that differ due to natural causes. CART analysis may be used to determine the relative importance of different variables for identifying homogeneous groups within the data set.”

For example, if the Miami Police want to predict the likeliness of a particular crime subgroup happening during a particular storm activity a CART analysis could help make that determination.

Excel, however, is chosen due to ease of use, especially its powerful pivoting capabilities, and speed at which new hypotheses can be tested. Moreover, Excel has a myriad of available functions that can aid in analysis.

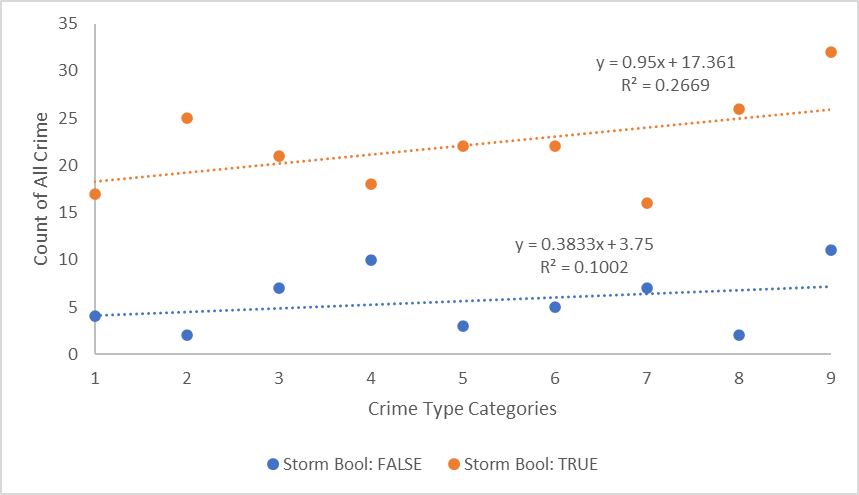
For example, the day of the week is determined from the *Date* records using the Excel function WEEKDAY(). The categorical crime activity variable (which is TRUE when a crime activity record is not null and FALSE when it is null) is then plotted against the day of week.



**Figure 1: Scatterplot and regression lines for categorical crime activity versus day of the week**

Figure 1 showed the correlation of crime versus an arbitrary variable, day of the week. The coefficients of determination show very weak correlation (R2 << 1.0). Figure 1 also shows that crime is insensitive to the day of the week, i.e., orange trendline has a slope of zero.

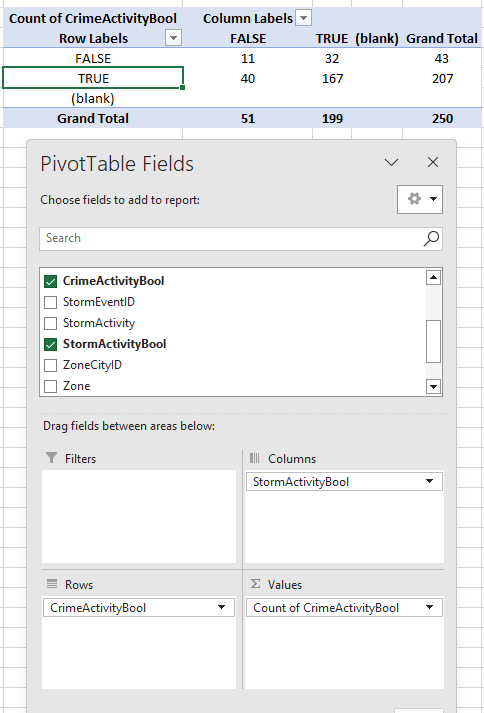
A similar scatterplot may be created with the type of crime, see Figure 2.



**Figure 2: Scatterplot and regression lines for categorical crime activity versus crime types**

Figure 2 shows that crime is higher for all subcategories when a storm is present, i.e., orange line is always above the blue line. It also suggests that the presence of a storm does correlate more strongly than the day of the week (R2 = 0.2669 >> R2 = -8E-15).

Another example, is pivoting the data into useful quantitative tables to look for patterns. One of the hypothesis is that higher crime should occur with storm activity and less crime with no storms. Using Excel’s pivot table capabilities, the following table is generated:



**Figure 3: Pivot table of Storm Activity Boolean and Crime Activity Boolean**

Within Figure 3 the intersection of TRUE/TRUE (crime activity present and storm activity noted) is the largest occurrence (167 occurrences). Figure 3 also shows that the rate of crime when a storm is active (167) falls when no storm is active (40).

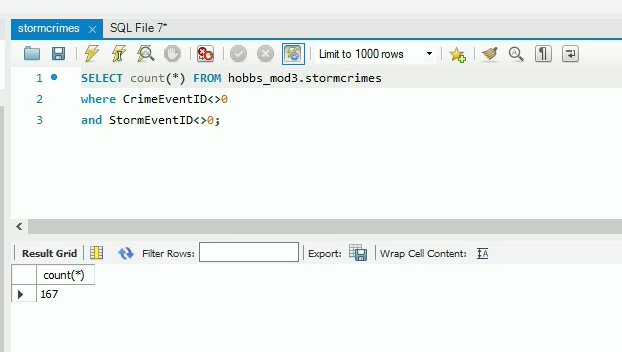
The ability to create these categorical variables and then pivot the data, as shown in Figure 3, within Excel was quick and very visual. Moreover, other quantitative tasks can be conducted easily, e.g., Cramer’s V. Cramer’s V as per Wikipedia (2023), “…is a measure of association between two nominal variables, giving a value between 0 and +1 (inclusive)”. For Figure 3 Cramer’s V was found to be 0.06 and with 1 degree of freedom this suggests no association between crime and storm activity.

# Section 4: Validation

Because the data was extracted and transformed, i.e., new variables added (Storm Activity Boolean and Crime Activity Boolean), it is imperative to validate that the data has not been corrupted by this transformation.

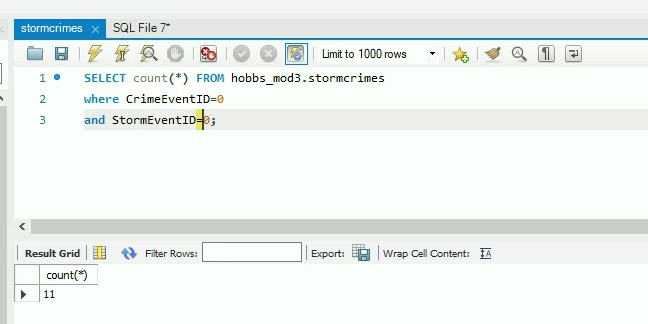
To validate the data transformed in Excel, MySQL is opened and queries are conducted to check each quadrant of the pivot table shown in Figure 3.

Note, in MySQL not equal to zero (< > 0) is the same as TRUE in the pivot table. Thus, the intersection of TRUE/TRUE is found with this query:



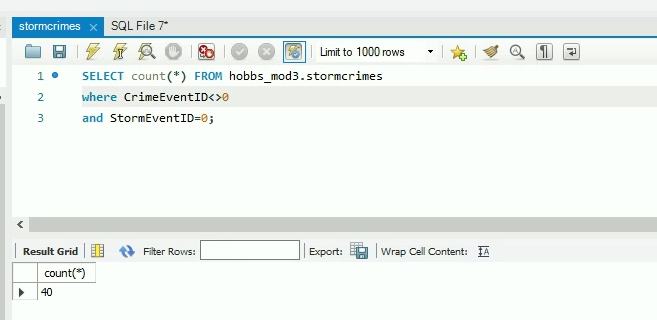
The count of 167 is indeed returned for this pair.

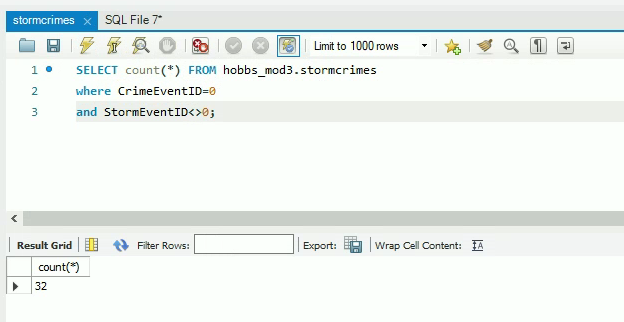
The FALSE/FALSE intersection is validated with this query:



Again, the value returned (11) validates the Excel transformation.

The remaining two intersection pairs are shown below:





The returned counts of 40 and 32, again, validate the transformation and the correctness of the pivot table.

Through this step-by-step commentary the types of analyses, hypotheses we are needing to prove, the tools available and why Excel was chosen, and then finally validation steps were discussed.

This simple analysis did not actually determine if storms and crime actually correlate (different analysis suggest different answers), but how to start breaking the analysis into smaller steps was successfully conducted.

# Citations:

Cramér's V. (2023, March 19) *In Wikipedia*. <https://en.wikipedia.org/w/index.php?title=Cram%C3%A9r%27s_V&oldid=1145529912>

EPA. (2023, February 2). *Additional information on classification and regression tree (CART) analysis.* United States Environmental Protection Agency. https://www.epa.gov/caddis-vol4/basic-analyses-1

Larose, D. T. (2015). Data mining and predictive analytics (2nd ed.). Wiley Global Research (STMS). https://mbsdirect.vitalsource.com/books/9781118991121

Smith, T. (2022, August 2). *Qualitative analysis*. Investopedia. https://www.investopedia.com/terms/q/qualitativeanalysis.asp#toc-qualitative-analysis-vs-quantitative-analysis