Programming in Python:

Web Scraping Project

Bo J. Taylor

Western Governors University

Abstract

The key topic of this project was the analysis of raw data generated emergency 911 calls. The datasets created from the raw data were used to provide a graphical summary of events by type, date, and location. Before datasets were created, outliers, errors, and duplicates were removed from and missing data was imputed to the raw data. A clean dataset was then generated by omitting columns of data unnecessary for creating summaries of events by type, date, or location. From that clean dataset, columns were extracted to new spreadsheets. A separate spreadsheet was used to analyze 911 calls by type, date, and location. Additional data was imputed as appropriate from those columns in order to generate the bar graphs. Additionally, the data was analyzed to determine if the city was eligible for additional funding based on a minimum requirement for officers onsite per incident using linear regression.

Programming in Python: Web Scraping Project

**Part I**

The first step I took towards creating “clean data” was determining which columns were necessary to chart events by date, type, and location. This allowed me to find any errors in the data by sorting each column in ascending order individually. This moved blank values to the top and also showed me an organized ranged of values. The only record with data that was missing and relevant to my analysis was the record with CAD CDW ID of 1702543. This record was missing the District/Sector value. I was able to impute this value by referencing the following column (Zone/Beat). The contents of this column contained two values: The District/Sector value and a numerical value. Removing the integer from the Zone/Beat value produced the District/Sector value.

**Outlier Data**

I was unable to find any outliers in the columns used in the clean data: Event Clearance Group, Event Clearance Date, District/Sector, and OFFICERS\_AT\_SCENE. Event Clearance Group has a range of values from “animal complaints” to “weapons calls”, and no values appeared to be an outlier. Event Clearance Date only had a range of timestamps for three days, and no values deviated from this range. The District/Sector column contained a total of 18 character values, without much of a potential for deviations or outliers. And finally, the OFFICERS\_AT\_SCENE column contained a range of integers from zero through four. I disregarded the zero values as outliers since those could be the method used by the office when closing out an event that ultimately didn’t require an officer (e.g. a caller resolved the initial complaint and no longer requires an officer’s assistance).

**Duplicate, Redundant, and Unnecessary Data**

The raw dataset did not appear to have duplicate records. I determined this by checking the first three columns for duplicate values: CAD CDW ID, CAD Event Number, and General Offense Number. Using Excel’s conditional formatting, I selected each column and highlighted duplicate values. Excel allows columns to be sorted by color. These columns did not have cells to organize by color, implying there are no duplicate values.

I was unable to discover any duplicate records since each record has unique CAD CDW ID values. However, the table did contain duplicate column data. Column M (Longitude) and Column N (Latitude) contained duplicate information from Column O (Incident Location). Column O contained the information in both Column M and N. If the longitude and latitude was relevant to the analysis, then either Columns M and N or Column O could be removed, depending on which format would be more optimal. Since the data is entirely unnecessary, all three columns were omitted.

After searching the raw data for duplicate records and columns, as well as missing values, I then evaluated the columns of information that were redundant or unnecessary. The following is a breakdown of each column removed with the reasoning for being omitted:

* Column A (CAD CDW ID): This data was useful initially to determine if there are duplicates of a single case. After checking for duplicate records, this data is unnecessary to generate the events tables.
* Column B (CAD Event Number): This could be used to determine duplicate values; however, Column A already provides the data potentially needed for that function.
* Column C (General Offense Number): This data does not serve a purpose in generating tables for events by date, type, or sector.
* Column D (Event Clearance Code): The event clearance group was the best source of data for organizing events by type. The code related to those events was unnecessary.
* Column E (Even Clearance Description): This column provides data that is too granular to chart events by type. Column G will be used since it provides data with event groups.
* Column F (Event Clearance SubGroup): This data resembles column G with increased detail by subgrouping events. This is unnecessary since column G provides groups for event types.
* Column I (Hundred Block Location): Column J provides necessary data to chart events by sector. Data pertaining to locations within sectors is more specific than what is required for this analysis.
* Column K (Zone/Beat): Column J already provides data necessary to chart events by sector. This data is unnecessary.
* Column L (Census Tract): This provides information for statistical subdivisions of a county. This is unnecessary for charting events by date, type, or sector.
* Column M (Longitude): This data is duplicated by Column O (Incident Location). Also, this data is not necessary to chart events by sector.
* Column N (Latitude): This data is duplicated by Column O (Incident Location). Also, this data is not necessary to chart events by sector.
* Column O (Incident Location): The specific latitude and longitude coordinates for events is too granular and unnecessary since the District/Sector column provides the relevant data.
* Column P (Initial Type Description): When considering between the initial even groups and clearance event groups, I debated which would provide more reliable data to organize the events by type. I chose the event clearance group since events may have been initially created under potentially incorrect or missing data. Completed events would provide a better indication of what the event truly reflected.
* Column Q (Initial Type Subgroup): Unnecessary since the clearance type data will be used instead of the initial type data.
* Column R (Initial Type Group): Unnecessary since the clearance type data will be used instead of the initial type data.
* Column S (At Scene Time): In order to chart evens by date, type, or sector, this data is not necessary. This data also is unnecessary to determine the number of officers for each event.

**Events by Date**

I created a histogram of the events by date, using individual dates as bins (Figure 1). The values for 3/26/2016 and 3/28/2016 are considerably lower, at first glance. After reviewing the records for those bins and comparing them to the bin for 3/27/2016, these values do not suggest they are outliers. 3/27/2016 consists of a full 24-hour’s worth of records, while 3/26/2016 and 3/28/2016 are for roughly half of a day each. Cutting the values for the 3/27/2016 bin in half would result in relatively the same bin size as the other two. From the statistics generated by this sample, the only inference I can make is the number of events was consistent for the three days being analyzed, given that the sample data did not contain three full days of event records.

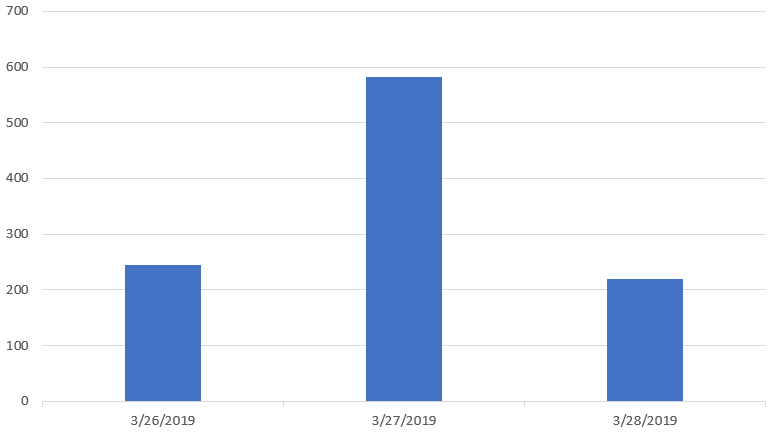


Figure 1. Events by Date. This figure illustrates the number of events per day.

**Events by Type**

The next histogram I generated was the events by type (Figure 2). This chart demonstrates the large difference in the number of event types. The top 50 percentile consists of only four events: false ALACAD (0.06), suspicious circumstances (0.14), traffic-related calls (0.16), and disturbances (0.16). The lowest seven event types amount to one percent of the events. Considering the high probabilities for only a small number of variables, these events may have the most impact when considering the number of officers per event.

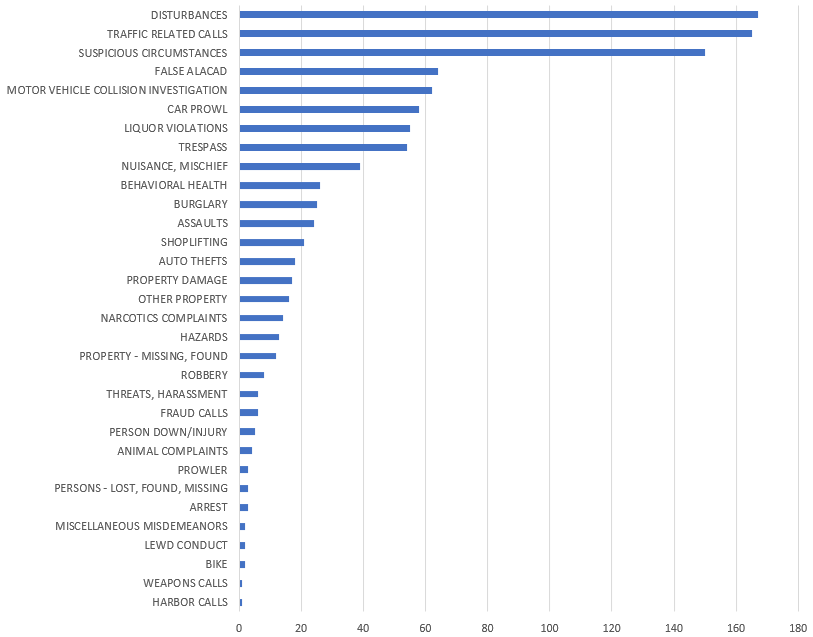


Figure 2. Events by Type. This bar graph illustrates variance in frequencies of each event.

**Events by Sector**

The final histogram I generated was the total number of events separated by sector (Figure 3). Most sectors fall within two standard deviations. However, sector H falls outside of this range (-2.43 to 109.48). Sector H has a Z-Score of 2.57, suggesting it may be considered an outlier (Anonymous, n.d.). After reviewing this histogram, I concluded that sector H would be omitted in future tables that excluded outliers.

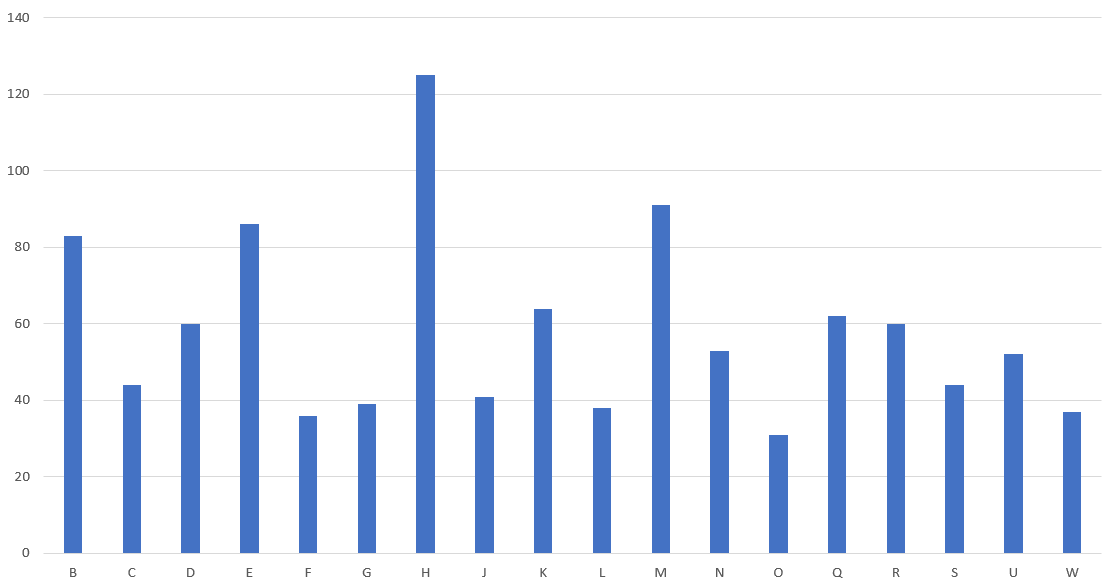


Figure 3. Events by Sector. This histogram depicts the difference in number of events experienced by each district or sector.

**Part II**

The main objective of Part II was to determine if the department qualifies for additional funding based on the average number of officers on scene. The minimum requirement is 2.5 officers on scene. In order to provide a confident response to this question, I identified potential outliers. I then reviewed the data in linear regression and residual plot models. While reviewing the data and models, I discovered potential data privacy concerns. And in conclusion, I proposed whether or not the department should receive additional funding according to the minimum threshold of officers on scene.

**Linear Regression Analysis**

The linear regression chart models the relation between the officers on scene and number of incidents; both of these values are separated by district. In this model, the number of officers on scene is the response variable, which is being predicted according to the number of incidents. Figure 4 illustrates the simple linear regression line with the outliers included in the plot. This gives a coefficient of determination of 87.95%, resulting in a 12.05% error rate. This indicates a good fit, however there are outliers that could be removed to reduce the error rate.

Prior to charting the linear regression models, I determined if any outliers existed. First, I found the mean for both the number of incidents per sector and officers on scene. Second, I determined the standard deviation value for each. Using those two values, I created the “normal” range of values within two units of standard deviation. Number of incidents outside of the range -2.43 to 109.48 are considered outliers, and number of officers outside of the range of 17.46 to 190.54 are also considered outliers. I determined Sector H to be an outlier since the number of incidents was 125, exceeding the 109.48 threshold. I also determined the record with no sector value to be an outlier since the number of officers at scene was 1, which was well below the threshold of 17.46

Removing the two outliers results in the chart in Figure 5. The coefficient of determination of this model is 95.91%, and an error rate of only 4.09%. This cuts the error rate by almost one third. Not only did removing outliers create a reduced error rate, it also altered the regression line slope. Removing outliers increased the predicted officer per event to increase from 1.49 to 1.83.

References

Anonymous. (n.d.). C 740. *Fundamentals/Statistics for Data Analytics V5.* Retrieved from https://learn.zybooks.com/zybook/WGUC740V52018/chapter/1/section/1?modal\_name=about-zybook

Kumashiro, Suzanne S. (2019). AFI 33-332. *Air Force Privacy and Civil Liberties Program*. Retrieved from <https://static.e-publishing.af.mil/production/1/saf_cio_a6/publication/afi33-332/afi33-332.pdf>

McCarron, Daniel [Daniel McCarron]. (2011, December 8). Residuals and Residual plots on Excel [Video file]. Retrieved from <https://www.youtube.com/watch?v=6HcyewrPNEw>