2018

Seattle Police Department 911 Call Log Analysis

With Linear regression analysis to determine additional funding

Provided by: Kristopher Chesney

Contents

[Part 1: Emergency 911 Call Log Analysis 2](#_Toc527831253)

[A. Dataset Preparation 2](#_Toc527831254)

[B. Data Cleaning Explanation 2](#_Toc527831255)

[I. Rows Deleted 2](#_Toc527831256)

[II. Columns Deleted 2](#_Toc527831257)

[III. Columns Kept 3](#_Toc527831258)

[IV. Other Considerations 3](#_Toc527831259)

[C. Data Sheets 4](#_Toc527831260)

[I. Date and Number of Events 4](#_Toc527831261)

[II. Number of Incident Occurrences by Event Type 5](#_Toc527831262)

[III. Sectors (and Beats) and Number of Events 8](#_Toc527831263)

[D. Summary of Data Sheet Observations 9](#_Toc527831264)

[I. Date and Number of Events 9](#_Toc527831265)

[II. Number of Incident Occurrences by Event Type 10](#_Toc527831266)

[III. Sectors (and Beats) and Number of Events 10](#_Toc527831267)

[Part 2: Linear Regression Analysis 11](#_Toc527831268)

[E. Linear Regression Line Fit 11](#_Toc527831269)

[F. Impact of Outliers 12](#_Toc527831270)

[G. Residual Plot 13](#_Toc527831271)

[H. State Funding Qualification Analysis 14](#_Toc527831272)

[I. Sensitive Data Precautions 14](#_Toc527831273)

[J. References 15](#_Toc527831274)

# Part 1: Emergency 911 Call Log Analysis

## A. Dataset Preparation

The provided “Raw Data.xlsx” file has been cleaned to remove any major errors or outliers. No duplicate records were found. The attributes of the dataset were then reduced further to support a more efficient analysis of the requested information. The finalized data set has been attached to this report, labeled as “Cleaned Raw Data.xlsx”.

## B. Data Cleaning Explanation

### I. Rows Deleted

1. No rows were deleted in this dataset.

### II. Columns Deleted

1. **At Scene Time** was deleted for the following reasons:

* 61.7% of the fields in this column are blank.
* Without further expert knowledge, derivation of this field from the “Event Clearance Date” field would be inaccurate and unreliable.

1. **Incident Location** was deleted for the following reasons:
   * This data can be derived from **Longitude** and **Latitude** columns.
   * This data was not needed to complete the task of this analysis.
2. **Event Clearance Code, Event Clearance SubGroup, Event Clearance Group,** and **District/Sector** were deleted for the following reason:
   * Relevant data in these columns could be derived from, or were better explained by, other columns that were retained.
3. **CAD CDW ID, General Offense Number, Hundred Block Location, Census Tract, Longitude, Latitude, Incident Location, Initial Type Description, Initial Type Subgroup, Initial Type Group, At Scene Time,** and **OFFICERS\_AT\_SCENE** were deleted for the following reason:
   * This data was not needed to complete the task of this analysis.

### III. Columns Kept

1. **Event Clearance Date** was kept for the following reason:
   * Further analysis of this attribute was requested to determine the date and number of events.
2. **CAD Event Number** was kept for the following reason:
   * This data provided a unique number for each event.
3. **Event Clearance Description** was kept for the following reasons:
   * This data was more granular and informational than both the **Event Clearance Group** and **Event Clearance Subgroup** columns.
   * This data has a direct correlation with the **Event Clearance Code** column, but the descriptions make this data more informative and therefore, more presentable.
4. **Zone/Beat** was kept for the following reasons:
   * This data contained the information requested in regards to **Sector.**
     1. In all instances of this log data, **Sector** was equal to the first letter in **Zone/Beat.**
   * This data contained additional information that could be provided without diminishing the accuracy of the requested information.

### IV. Other Considerations

In order to provide the appropriate tables and bar graphs, **Zone/Beat** was split into two new columns, labeled **Sector** and **Beat**, respectively. This change also allowed **Row 226** (**CAD Event Number**: 16000105361) to be maintained. This row was missing data for the **Sector** column, but as **Zone/Beat** has a direct correlation with **Sector**, this replacement was sufficient and accurate.

## C. Data Sheets

The following information includes tables and bar graphs displaying an analysis of the requested information.

### I. Date and Number of Events

1. Table displaying the date and number of related events.

|  |  |
| --- | --- |
| Date | Count of CAD Event Number |
| 26-Mar | 244 |
| 27-Mar | 583 |
| 28-Mar | 219 |
| **Grand Total** | **1046** |

1. Bar graph displaying the date and number of related events.

### II. Number of Incident Occurrences by Event Type

1. Table displaying the number of incident occurrences by event type.

|  |  |
| --- | --- |
| Event Types | Count of CAD Event Number |
| JUVENILE DISTURBANCE | 1 |
| HARBOR - DEBRIS, NAVIGATIONAL HAZARDS | 1 |
| PERSON WITH A WEAPON (NOT GUN) | 1 |
| RECKLESS ENDANGERMENT, LITTERING, PARKS CODE VIOLATIONS | 1 |
| CRISIS COMPLAINT - PICK-UP OR TRANSPORT | 1 |
| NARCOTICS ACTIVITY REPORT | 1 |
| ANIMALS - INJURED, DEAD, DANGEROUS | 1 |
| NARCOTICS FOUND, RECOVERED | 1 |
| FELONY WARRANT SERVICE | 1 |
| PARKS EXCLUSION | 1 |
| LEWD CONDUCT | 2 |
| MISDEMEANOR WARRANT SERVICE | 2 |
| FORGERY, BAD CHECKS | 2 |
| BICYCLE THEFT | 2 |
| NARCOTICS, DRUG TRAFFIC LOITERING | 2 |
| PROPERTY - MISSING | 2 |
| STRONG ARM ROBBERY | 3 |
| MOTORIST ASSIST | 3 |
| ANIMAL NOISE, STRAYS, BITES | 3 |
| MARIJUANA PUBLIC USE (NOT DISPENSARY) | 3 |
| GANG GRAFFITI | 3 |
| MISSING PERSON | 3 |
| PEDESTRIAN VIOLATION | 3 |
| AUTO RECOVERY | 3 |
| BURGLARY - RESIDENTIAL, OCCUPIED | 3 |
| PROWLER | 3 |
| ALACAD - COMMERCIAL PANIC (FALSE) | 3 |
| BLOCKING VEHICLE | 4 |
| ABANDONED VEHICLE | 4 |
| FRAUD (INCLUDING IDENTITY THEFT) | 4 |
| BURGLARY - UNOCCUPIED STRUCTURE ON RESIDENTIAL PROPERTY | 4 |
| LICENSE PLATE THEFT OR LOSS | 5 |
| CASUALTY (NON CRIMINAL/TRAFFIC) - MAN DOWN, SICK PERSONS, INJURED, DOA) | 5 |
| ARMED ROBBERY | 5 |
| LIQUOR VIOLATION - ADULT | 5 |
| NOISE DISTURBANCE, RESIDENTIAL | 6 |
| HARASSMENT, THREATS | 6 |
| NARCOTICS, OTHER | 7 |
| ALACAD - RESIDENTIAL PANIC (FALSE) | 7 |
| BURGLARY - COMMERCIAL | 9 |
| BURGLARY - RESIDENTIAL, UNOCCUPIED | 9 |
| PROPERTY - FOUND (FOLLOW UP TO SPD CASE) | 10 |
| SUSPICIOUS CIRCUMSTANCES - BUILDING (OPEN DOOR, ETC.) | 10 |
| FIGHT DISTURBANCE | 13 |
| HAZARDS | 13 |
| PROPERTY DESTRUCTION | 14 |
| AUTO THEFT | 15 |
| THEFT - MISCELLANEOUS | 16 |
| DRIVING WHILE UNDER INFLUENCE (DUI) | 18 |
| SHOPLIFT | 21 |
| ALACAD - RESIDENTIAL BURGLARY (FALSE) | 22 |
| ASSAULTS, OTHER | 24 |
| NOISE DISTURBANCE | 24 |
| CRISIS COMPLAINT - GENERAL | 25 |
| ALACAD - COMMERCIAL BURGLARY (FALSE) | 32 |
| SUSPICIOUS VEHICLE | 35 |
| MISCHIEF, NUISANCE COMPLAINTS | 39 |
| LIQUOR VIOLATION - INTOXICATED PERSON | 50 |
| THEFT - CAR PROWL | 53 |
| TRESPASS | 54 |
| MOTOR VEHICLE COLLISION | 62 |
| TRAFFIC (MOVING) VIOLATION | 66 |
| PARKING VIOLATION (EXCEPT ABANDONED VEHICLES) | 67 |
| SUSPICIOUS PERSON | 105 |
| DISTURBANCE, OTHER | 123 |
| **Grand Total** | **1046** |

1. Bar graph displaying the number of incident occurrences by event type.

### III. Sectors (and Beats) and Number of Events

1. Table displaying the sectors (and beats) and number of events.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Beats | |  |  |  |
| Sectors | 1 | 2 | 3 | | S | Grand Total |
| B | 33 | 25 | 25 | |  | 83 |
| C | 10 | 18 | 16 | |  | 44 |
| D | 20 | 22 | 18 | |  | 60 |
| E | 33 | 35 | 18 | |  | 86 |
| F | 12 | 13 | 10 | | 1 | 36 |
| G | 16 | 12 | 11 | |  | 39 |
| H | 36 | 49 | 40 | |  | 125 |
| J | 16 | 11 | 14 | |  | 41 |
| K | 14 | 19 | 31 | |  | 64 |
| L | 17 | 13 | 8 | |  | 38 |
| M | 31 | 36 | 24 | |  | 91 |
| N | 19 | 12 | 22 | |  | 53 |
| O | 18 | 8 | 5 | |  | 31 |
| Q | 18 | 21 | 23 | |  | 62 |
| R | 22 | 18 | 20 | |  | 60 |
| S | 19 | 15 | 10 | |  | 44 |
| U | 17 | 17 | 18 | |  | 52 |
| W | 22 | 11 | 4 | |  | 37 |
| **Grand Total** | **373** | **355** | **317** | | **1** | **1046** |

1. Bar graph displaying the sectors (and beats) and number of events.

## D. Summary of Data Sheet Observations

### I. Date and Number of Events

There are significantly more events occurring on March 27th. The number of events on March 27th are more than both the other days combined. The other two days, March 26th and March 28th have a similar number of events. After reviewing the data, this is most likely attributed to the first event time showing as 17:14, which is more than halfway through March 26th. Likewise, the data for March 28th ends at 11:08 of that day, just under hallway. Further investigation shows that March 26th is a Saturday, March 27th a Sunday, and March 28th a Monday. In addition, Sunday March 27th 2016 is the US holiday, Easter.

### II. Number of Incident Occurrences by Event Type

This table and graph allow observations of the most common and least common incidents that occurred over the three days of information contained in this dataset. The most common incident is described as **Disturbance, Other**, followed by **Suspicious Person**. The top ten events make up 62.5% of all incident occurrences, while the bottom ten events make up less than 1% (0.956%) of all incident occurrences.

### III. Sectors (and Beats) and Number of Events

Sector **H** had the most events (12% of events) by a significant margin, as the next highest sector had thirty-four less events. Sector **O** had the least number of events (3% of events), but the next lowest sector only had fire more events. While the beats in some sectors could vary significantly, overall, the total of events per beat were relatively close to each other.

# Part 2: Linear Regression Analysis

## E. Linear Regression Line Fit

A linear regression line is a line that best fits all points of the data. One way to achieve this is by using the sum of squared errors, and this has mathematical advantages so is usually the preferred method (zyBooks, 2018). This is called the method of least squares. As can be seen in the provided Linear Regression Plot, the line shows an attempt to most accurately represent the trend of the data.

## F. Impact of Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population (Prins, McCormack, Michelson, & Horrell, 2018). They can radically effect the

accuracy of any type of predictive analytics, including linear regression lines. The best way to determine outliers is by having expert knowledge of the data to know what doesn’t fit. Outside of that, using visual representation tools can help one identify outliers. In the case of this data, visual representation through a scatter plot was used. The graph below is an example of how outliers can impact a regression line. The blue line seen here is the linear regression line built up on all the data points. The orange line shows the linear regression line using the same data, but excluding the outlier located at (125,165).

## G. Residual Plot

The graph below represents a residual plot, built upon the data in “Linear Regression.xlsx”.

A relatively small residual standard error indicates that the actual future value of Y is likely to be relatively close to the predicted value. Therefore, less residual standard error is better (zyBooks, 2018). Based on this idea, the best way to improve the linear regression model would be to omit the outliers with the largest residual error. In this case, I could remove the data point for the data from District H. In addition, I would perform further investigation to determine which District the data (1,1), without a district, belongs under. Doing this would improve the linear regression model by reducing the overall residual standard error. One way to determine an outlier is by finding the mean of the absolute value for the residual number. Then derive standard deviation of the of that mean. Statistically, 99.73% of data will be within three standard deviations, meaning anything outside of that could be a possible outlier. In the case of this data, this verifies that point (125,165) is a good candidate to be removed as an outlier. In addition, it would be recommended to verify this to be a good choice with those that have expert knowledge of the material.

## H. State Funding Qualification Analysis

The linear regression line for this data can be viewed by the equation, y = 1.491x + 21.914. This is a equation, where equals the slope of the line. This slope is a representation of the direct correlation between the X and Y values, or in the case of this data, the **Number of Incidents** and **Officers at Scene**. This shows us that overall, the Seattle Police Department has just under 1.5 (1.491) officers onsite per incident. In order to be eligible to receive funding, they would need to send just over 1 (1.09) more officers onsite per incident.

## I. Sensitive Data Precautions

Much of the data in this scenario can be called sensitive for various reasons. **Hundred Block Location** and **Incident Location** (Latitude & Longitude) could be used to narrow down common locations for criminal activity. **At Scene Time** could also be used to learn how long it takes officers to typically get to a certain area. These and other pieces of sensitive data in this scenario should be treated with confidentiality. Although there is no personally identifiable information (PII), the data in this scenario should still be treated as sensitive and should not be shared with people not authorized to see it. The data should not be left in a location that others might have access to. Some additional precautions include; storing the data in an encrypted file, ensuring that an unlocked computer is not left unattended, and not discussing details of the data with others. Whenever possible it is important that confidential and sensitive data be discussed in terms of specific instances rather than general concerns. This allows practical solutions to be developed and offered to researchers (Rice & Southall, 2016, p. 123). For this scenario, a very important aspect would be to discuss the data sensitivity with the police department, so that everyone is in agreement about what data can be used to find answers to the questions they have. This allows the project to continue successfully and responsibly.

## J. References

Prins, J., McCormack, D., Michelson, D., & Horrell, K. (2018, 19 October). *What are outliers in the data?* Retrieved from NIST/SEMATECH e-Handbook of Statistical Methods: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

Rice, R., & Southall, J. (2016). *The Data Librarian's Handbook.* London: Facet Publishing. Retrieved October 20, 2018, from https://wgu.idm.oclc.org/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1454580&site=eds-live&scope=site

zyBooks. (2018, October 19). *Creating a regression line: More common method of minimizing sum of squared errors*. Retrieved from C 740: Fundamentals/Statistics for Data Analytics V5: https://learn.zybooks.com/zybook/WGUC740V52018/chapter/5/section/3

zyBooks. (2018, October 20). *Interpreting residual standard error*. Retrieved from C 740: Fundamentals/Statistics for Data Analytics V5: https://learn.zybooks.com/zybook/WGUC740V52018/chapter/5/section/4