**Part 1**

1. In the workbook, the first sheet is the original raw data (originally from Taskstream). The second sheet is the cleaned dataset.
2. Below is a list of each column that was removed and why it was removed:
   * CAD CDW ID and CAD Event Number – Both appear to be unique identifiers for each offense and are not needed for comparing each incident. General Offense Number will be used if uniquely identifying each incident is needed.
   * Event Clearance Code – The Event Clearance Code matches with the Event Clearance Group and thus is duplicate information. The Event Clearance Group is more descripted when it comes to visualization.
   * Event Clearance Description and Event Clearance SubGroup – Although these features provide a more detailed look at each incident, the Event Clearance Group acts as the event type for the needs of comparison.
   * Hundred Block Location, Longitude, Latitude, Incident Location – Although location information may be useful for more detailed analysis, it is not needed for summarizing the incidents.
   * Census Tract – According to Weessies, a census tract “is an area roughly equivalent to a neighborhood established by the Bureau of Census for analyzing populations.” This information is similar to location or sector information and not needed for summarizing the incidents.
   * Initial Type Description, Initial Type Subgroup, and Initial Type Group – Although these were features initially used to describe and categorize each incident, there is Event Clearance Group to define the final event type better.
   * At Scene Time – As stated in zyBooks, “Discarding features (columns) with too many missing values…” was something to consider for this particular feature. Over half of the fields are blank.

Below are the rows that were modified or removed and why:

* Row 226 (General Offense Number 2016105361) – After removing all of the unneeded columns, this row had an invalid Zone of FS and a missing Sector value. At first glance, one could assume the Sector should be F, but on comparison of the Longitude and Latitude locations, it more closely matched the O Sector. Since this was the only row with incomplete data (accounting for far less than 1% of the entire dataset), I chose to omit it.

1. Tables and bar graphs are the next three sheets in the workbook.
2. When observing the number of events by date, one notices that there were significantly more events on 3/27/2016 than the other two dates. Further analysis on the types of events and other detailed information on that date may show insight as to the reason for the extreme difference in the amount of events.

When observing the number of events by event type, one will see that over half of the event types occurred less than 20 times over the course of the three days. Only three events occurred over 140 times. Further analysis on the specific dates or sectors with the event type may provide other insight.

When observing the number of events by sector, well over half of the sectors had 60 events or less occur over the course of the three days. Four sectors had over 80, with sector H standing out with over 120 events occurring. Further analysis that would break out the sectors by date and/or event type may provide more insight.

**Part 2**

1. The linear regression line for this data model is considered the line of best fit. This is derived from the method of least squares.

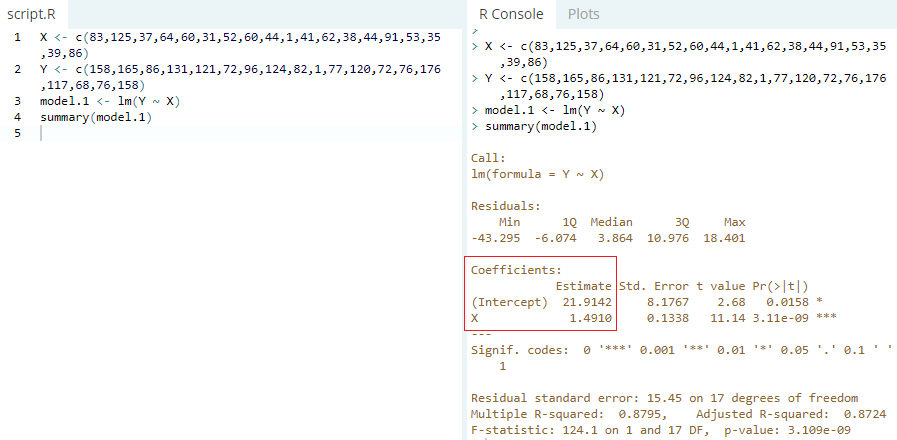


Figure 1 - R code confirming best fit

1. There are two outliers in this model – the undefined sector with one officer and one incident and sector H, which had over 30 more incidents than the second highest sector, with less officers on scene then that same sector. Both affect the slope of the linear regression line slightly. If removed, the slope changes and the coefficient of determination is much higher. More points in the dataset sit closer to, if not on top of, the regression line.

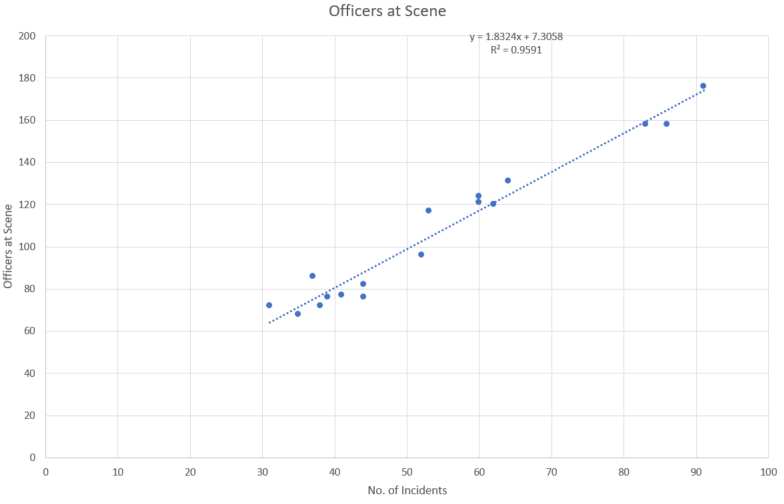


Figure 2 - Linear Regression after removing outliers

1. The residual plot below shows a slight positive trend while sharply going negative with the last value. This puts doubt on both the ‘mean of zero’ and ‘constant variance’ assumptions. This would most likely be improved by removing the outliers.

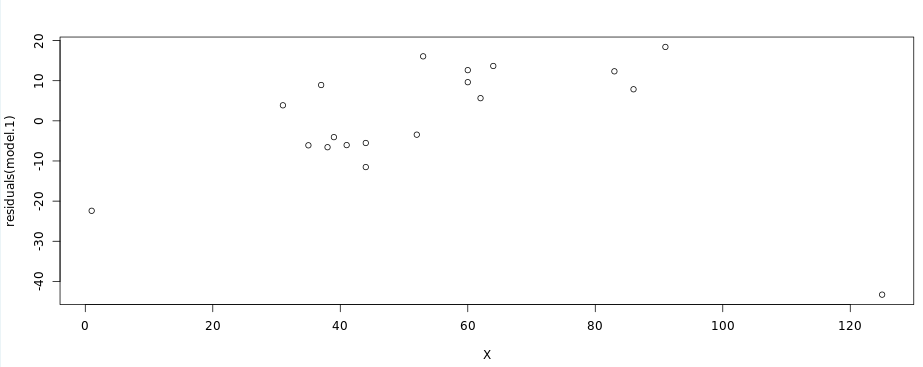


Figure 3 - Residual plot for the dataset

1. The police department does not qualify for additional state funding. Although the regression model shows significance, as shown in Figure 1 with the p-value, none of the sectors had an average of at least 2.5 officers on scene per incident. Removing the outliers only improves this average slightly and not enough to meet the minimum requirements for funding. Some of the limitations in analyzing this dataset are the size and scope of the dataset. Although the full dataset is over 1000 rows, the linear regression model is based off of only the 19 sectors. The data is only a three day period and may not be a good representation of the rest of the year. It might be more beneficial to analyze the average number of officers on scene each day over an extended period of time.
2. When handling sensitive data, precaution needs to be taken to ensure it stays protected. The zyBooks resource discusses many ways to keep the data safe, such as locking down the computer and files that contain the data. The amount of people that have access to the data also needs to be taken into consideration. Having the least amount of people with access minimizes the likelihood of a data breach.

**References**

zyBooks. *C 740: Fundamentals/Statistics for Data Analytics V5.* (2018) Retrieved from: <https://learn.zybooks.com/zybook/WGUC740V52018>

*R Code Runner Powered by DataCamp.* (2018) Retrieved from: <http://rcoderunner.zybooks.com/index.html>

Taskstream (2018) Raw Data [Raw Data.xlsx], Linear Regression [Linear Regression.xlsx]. Retrieved from: [https://tasks.wgu.edu/student/000554641/course/13800005/task/8f332b46-675e-444c-b65c- 7ac544b37d37/overview](https://tasks.wgu.edu/student/000554641/course/13800005/task/8f332b46-675e-444c-b65c-%20%097ac544b37d37/overview)

Weessies, Kathleen. *Finding Census Tract Data: About Census Tracts* (July 17th, 2018). Retrieved from: <https://libguides.lib.msu.edu/tracts>