C740: Fundamentals of Data Analysis Project

Elizabeth Sweet

Western Governors University

C740: Fundamentals of Data Analysis Project

A local police department requests an analysis of data collected from 911 calls placed between March 26, 2016, and March 28, 2016. The department also wants to see if they qualify for additional funding offered by the Governor for a department that has a response rate of at least 2.5 officers per incident. (Taskstream, n.d.)

Part 1

The raw data provided contains twenty-one fields and 1046 records that include information about the time, date, location, and types of incidents that occurred between March 26, 2016, and March 28, 2016.

A: Dataset Preparation

See accompanying Excel worksheet Clean Data.

B: Data Preparation Explanation

For this analysis, the focus is on the fields of CAD CDW ID, Event Clearance Group, Event Clearance Date, District/Sector, and Officers_At_Scene. All other fields were deleted. Several of these fields are redundant, while others infringe on the privacy of the citizens involved. Also, CAD CDW ID number 1702543 was removed due to missing data in the District/Sector field. One could try to use the Zone/Beat and the given coordinates to fill in the missing value, but given the size of the dataset, the loss of one record will not distort the resulting analysis.

C: Data Sheets

See Excel worksheets Events by Date, Occurrences by Type, and Events by Sector.

D: Data Sheets Observations

The first inquiry was the number of events by date. March 27 has significantly more events occur than on the other dates. Additional investigations into the types of events, if a higher rate of incidents on Sundays is regular, and other information may provide more in-depth insight into why the count of events is so much higher.

The next inquiry was occurrences by types. Seventy-five percent of the event types occurred less than 50 times. Disturbances, traffic-related calls, and suspicious circumstances were the three most common occurrences with 150 occurrences or more. Weapon calls and harbor calls occurred the least with only one each. Exploring the sectors, dates, and times may provide more understanding as to why these three types happen at a much higher rate.

The final inquiry was events by sector. Sector H had the highest number of events with 125, followed by sectors M, E, and B. Sector O had the least with 31. The majority of sectors have less than 60 occurrences over the given period. Additional study could help to understand why sector H has such a high rate compared to other sectors.

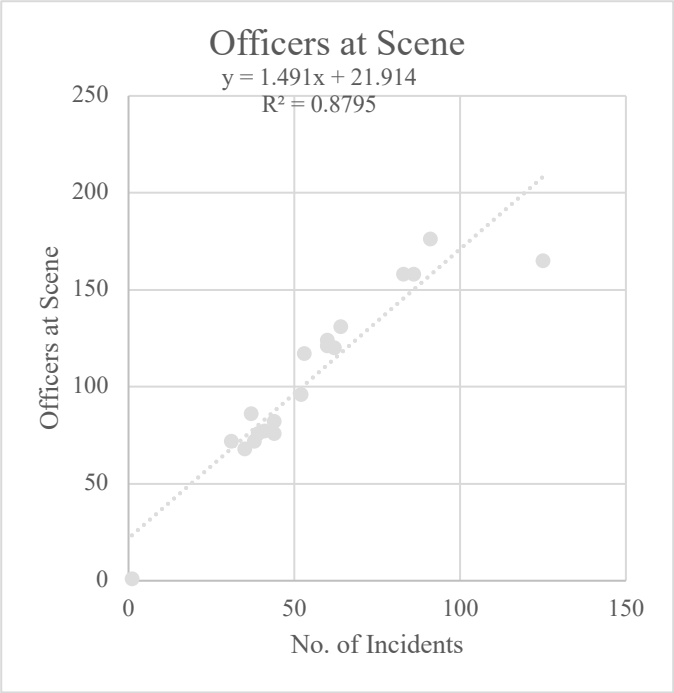
PART 2

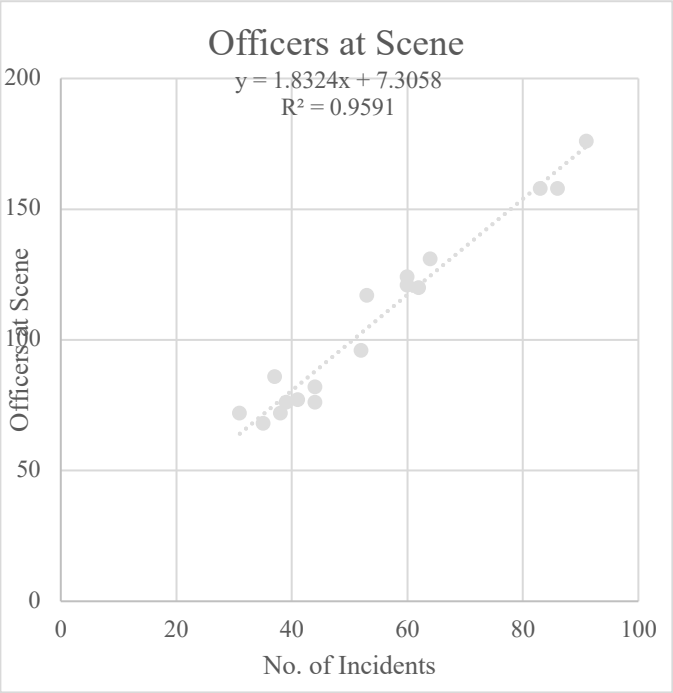
E: Fit

The linear regression line, as seen on the accompanying Excel worksheet labeled Linear Regression, is trending in a positive direction, meaning as the number of incidents increases, so too does the number of officers at the scene. The R squared value is high at 87.95%, leaving 12.05% of the variations in predictions due to factors other than the number of incidents.

F: Outliers

Two outliers are affecting the regression line for this data set. The first is at (1,1), which is an unnamed sector. The second is (125,165). This point has a higher number of incidents with fewer officers at the scene. After removing these two points, there is a visible improvement to the regression line. The line passes through more of the data in the graph on the right then on the left.



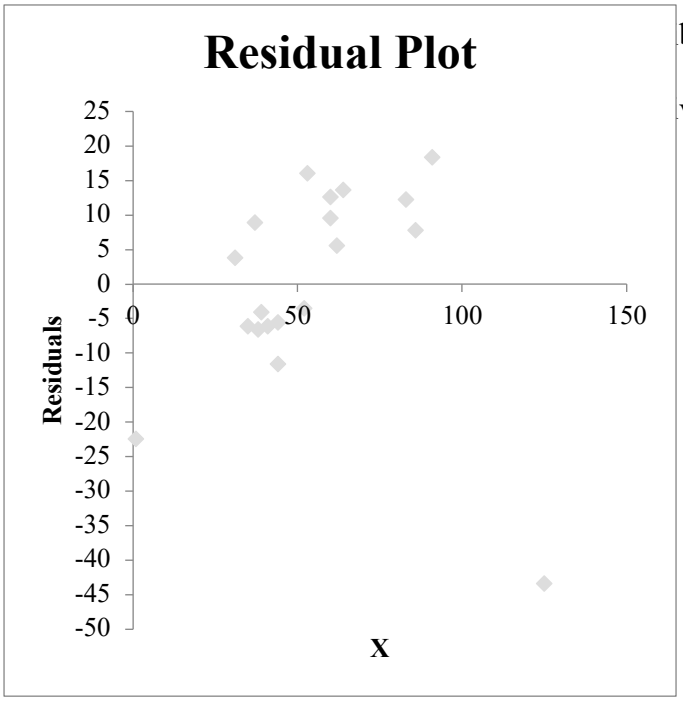


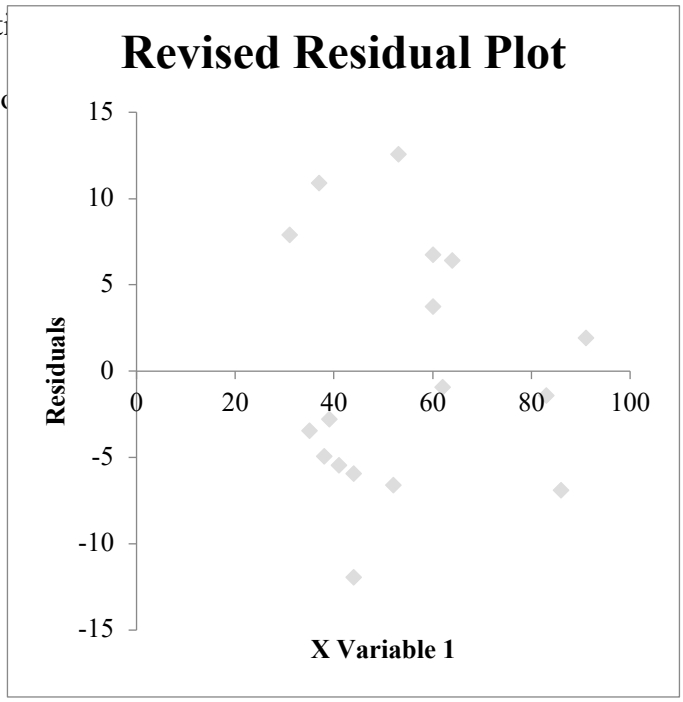
The R squared value has improved by almost 8%, indicating that the new model accounts for more variation in its predictions (Winston, n.d.). A larger percentage of the variation in officers at the scene is explained by the number of incidents and not from an outside. Removing the outliers results in a closer fitting regression line.

G: Residuals

The residual plot, shown below on the left, has a positive trend until the last data point where it falls sharply. From a visual inspection, the graph does not seem to adhere to some of the linear regression assumptions. The points on the graph are not evenly distributed from left to

right. This distribution calls into question the constant variance assumption. Having a mean of zero is also not met. Two points fall well below the others with no balancing point on the positive side. These two points are the outliers for the previous section. After removing these points, the





H. Current Qualifications

This department does not meet the Governor's standard of 2.5 officers per incident to receive extra funding. The number of officers per incident in this department is 1.89. While removing the outliers does increase the mean, it does not raise it enough to meet the requirement. There are some limitations to this analysis. The data collection was over three days. A more extended period might have a different result. There could have been fewer officers on duty that weekend compared to a more normal weekend.

I. Precautions or Behaviors

This data set contains the location of the incidents reported listed in several different ways. It gives the block location, the census tract number, longitude, and latitude. Given this, a person could easily find the location of the incidents. They could use this information to see where there is less of a police presence and use this information to commit further crimes. According to Zybooks, keeping computers locked when not in use and encrypting the sensitive data can help to protect privacy. Also, limiting the number of people who have access to the data will help to curb accidental breaches.

References

Taskstream. (n.d.). *AKM1 TASK 1: Fundamental Analytics*. Retrieved 2020, from https://tasks.wgu.edu/student/001000431/course/13800005/task/424/overview

Winston, W. (n.d.). *Interpreting the R-squared value*. Retrieved 2020, from LinkedIn Learning: https://www.linkedin.com/learning/excel-data-analysis-forecasting/interpreting-the-r-squared-value?u=2045532

Zybooks. (n.d.). *WGU: Fundamental of data analytics*. Retrieved 2020, from https://learn.zybooks.com/zybooks/WGUC740V52018