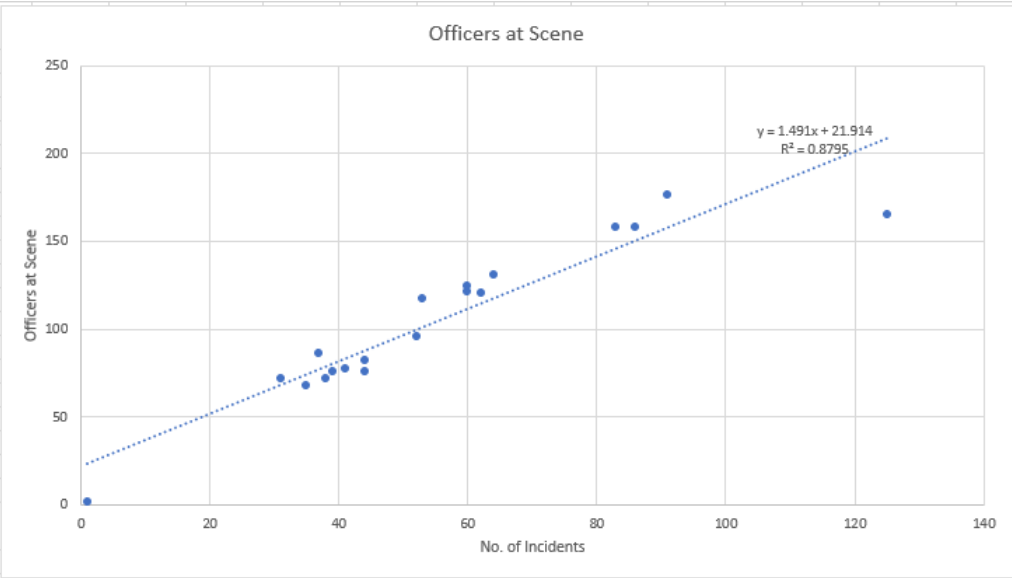
Fundamentals of Data Analytics

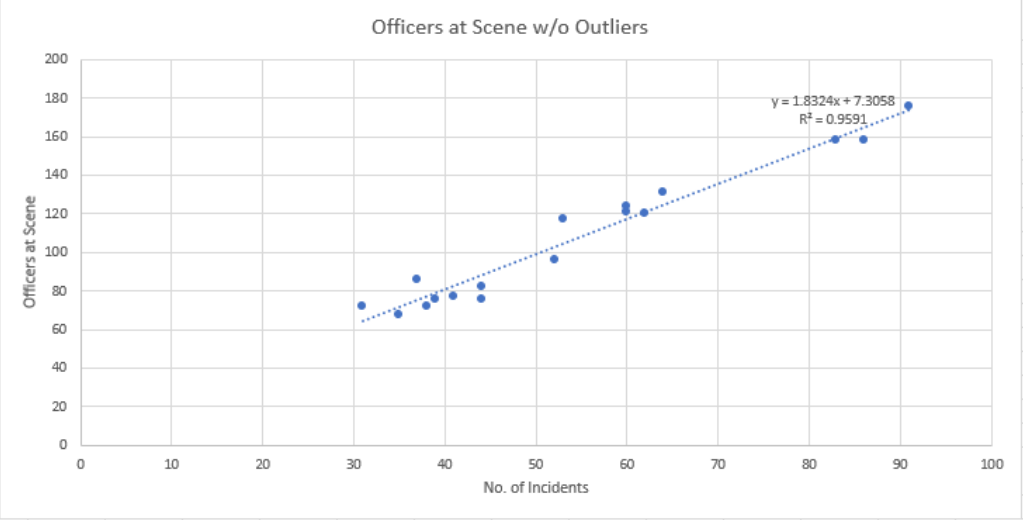
Part 1.

A. See Excel. Represents 1045 data points. One data point has been removed because the “Sector” was missing.
B. See Excel. Represents 1045 data points. One data point has been removed because the “Sector” was missing.
C. Two inferences about least squares
    a. Assuming the simple linear regression model below, one would could infer that the expected number of officers in a District Sector that had 100 incidents would be Y=1.49x+210914 officers where X=100. Thus, we would expect there to be 171 officers needed for 100 incidents.



    b. Similar to the inference above, we could infer that if 171 officers are needed for 100 incidents we could also infer that there would be 1.71 officers at each incident or that there would be 2 or more officers at 58.48% of the 100 incidents.
    c. Referencing the above linear regression, we would expect zero officers to be present until there were 23.404 incidents. We would expect that at 23.404 incidents one officer would be present.
D. Two inferences about the outliers’ impact on the data
    a. When outliers are removed the Y-intercept changes from 21.914 to 7.3058. Likewise, the R-Squared value approaches nearer towards 1 indicating a more reliable source for estimating the expected value of "Y" per a Value of "X". For example, if there were 70 incidents over a 3-day period we would be able to calculate the estimated total number of offices needed to support the 70 incidents as Y=1.8324x+7.3058 with X=70. With an R-Squared value of 0.9591 we would expect our estimation without outliers to be more

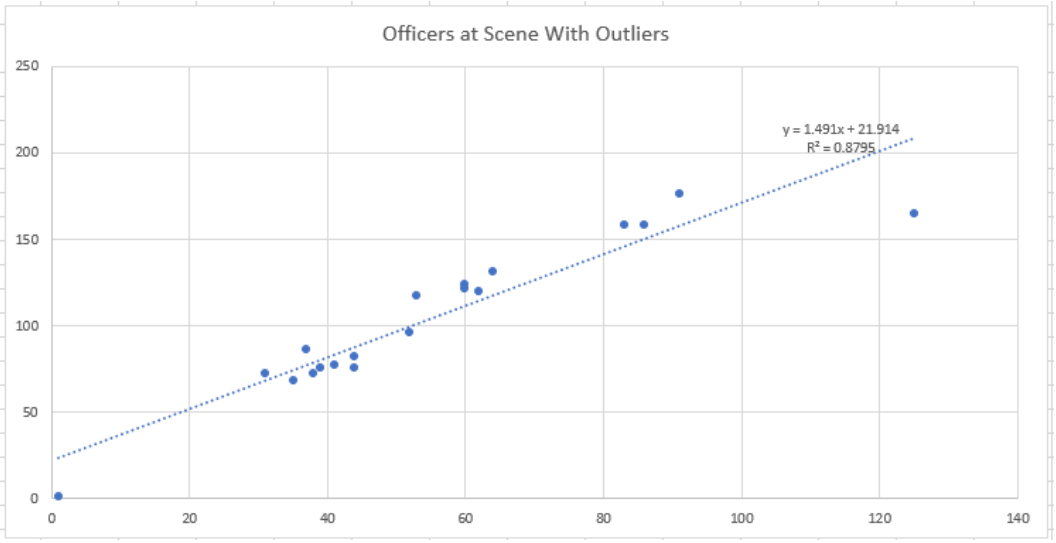
accurate than the linear regression model with outliers.



b. Without outliers, we would expect there to be zero officers until 9.1382 incidents within a 3-day period. At that point, we would expect 1 officer to be needed to cover the incidents. 9.1382 incidents for one officer seems to be a more probable expectation than the 23.404 incidents for one officer. Despite whether it is more probable or not, we can see that the removal of the outliers has a great impact on the y-intercept and the slope of the linear regression line. Accordingly, we could infer that as incidents rise there would be more officers needed per an incident in the model without outliers than the model with the outliers.

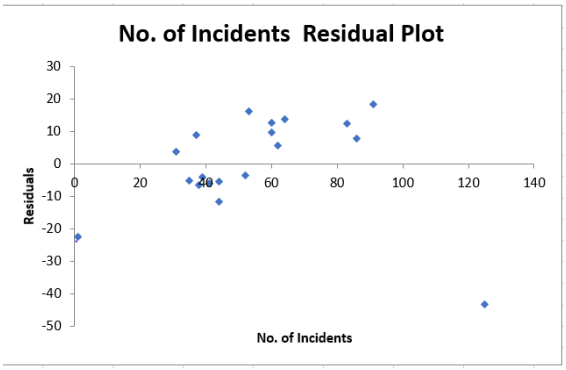
c. Linear regression is in a simple way to estimate or predict, if you will, a variable based on trend line or slope and a selected x value. It is important to note that the slope or the trend line is affected by outliers. Essentially you can have a steeper or less steep slope with different y and x variables. When your data contains outliers, those outliers can greatly affect the slope of your trend line. Thus, your estimation or prediction would give you a less accurate take on reality or future events. Like the examples above when the outliers are removed from the data we can see that the data points fairly close to the trend line which is what we want. With outliers, we have data points that are far from the trend line and indicate a less accurate trend line. This is also shown with an R-Squared value that is further from 1 than when outliers are removed

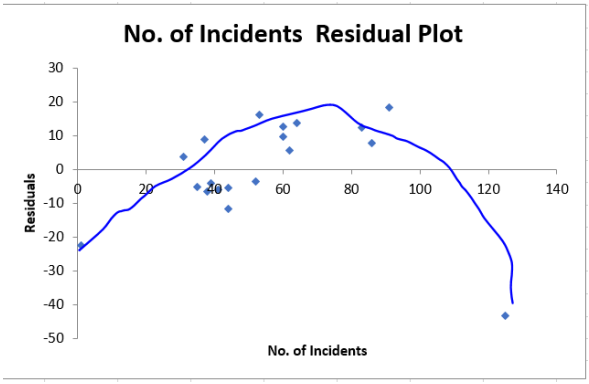
($With\ outliers$: $R^2 = 0.8795$; $Without\ Outliers$: $R^2 = 0.9591$).



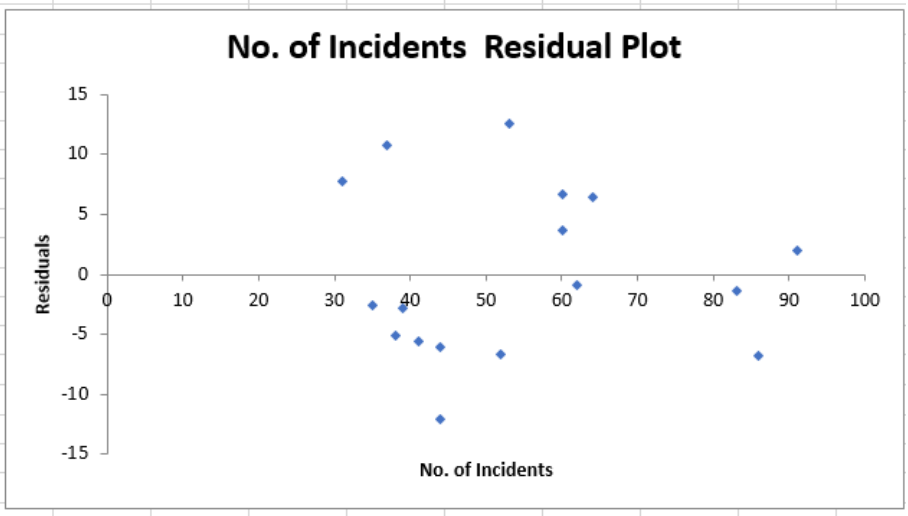
E. Two inferences about residuals represented in the data.

a. In the first linear regression model with outliers it is the SAE is greater than when the outliers are removed. However, with in the linear regression with the outliers it almost looks that at a certain point of incidents the number of officers needed decreases. One could wonder that perhaps at a certain point of incidents less officers are needed per an incident. Reasons could be because the geographical area becomes saturated thus taking officers less time between incidents and quicker response times. When the residuals are represented on a scatter plot it could be determined that perhaps a linear regression line is not the best model for this data set. The slight U-Shape is one indicator of this possibility.





b. When the outliers are removed the residual scatter plot is more random and better fits a linear regression model. Also, the R-Squared value approaches further towards "1" indicating a better fit and predictor of future officer needs based on the number of incidents. See scatter plot below.



Part 2:

F. When working with sensitive data you must be sure to maintain privacy, confidentiality transparency and to protect identity. Privacy is huge. Often as a data analyst you will work with private data that should not be shared with the general public or anyone other than authorized individuals. When sharing or distributing sensitive data you can take precautions to insure the individual you are talking to is in fact the one who should be receiving the data. There are many different ways you can identify and individual. If working within a co-worker in another location you can ask verifying questions before releasing sensitive data. You can also use secure internal communication methods to send a message that must be verified before discussing sensitive data. At my work, we use these methods as well as email. When sending sensitive data via email it is important to encrypt the data. A common way to do this is to use the KEY123 function in outlook or lotus notes. The KEY123 function encrypts the data so only the intended and proper individual can obtain the data sent. Other ways of protecting data is to lock your computer when you step away, collecting data anonymously in surveys, and quickly retrieving papers from the printer with sensitive data.

Equally as important to keeping data private is to be 100 percent transparent in your data analysis. What does this mean, "Transparent"? Have you ever heard the weather man predict the weather to be sunny and it was cloudy or rainy? Wouldn't it be nice if the weatherman would say instead, "We expect the weather to be Sunny skies with a probability of 80% which

was based off a 10-day historical weather sample. Since the sample size is small be on the lookout or possible cloudy skies or rain." OK this would most likely make the weatherman even more of a joke. But the point is we need to be careful how we interpret and transmit data. Giving bad data, omitting statistics or giving bias can lead people to make uninformed and bad decisions. Another example of how data is misused is stating a cause when the data only implies a possible correlation. For example, a data analyst may take a sample individuals that eat 3 or more Snickers a day. The data analyst may conclude that eating 3 or more snickers a day makes an individual overweight. However, true cause may be that individuals that eat 3 or more snickers a day do not exercise thus are overweight. It is important to remember, "Correlation isn't causation" (Section 10.2, ZyBooks, 2016).

There are many ways data can be misrepresented. The p-value, which is highly used by statisticians to find significance at less than 0.05, is a value that if used improperly can easily create misconceptions. It is important to understand that "a p-value of 0.05 does not mean a given hypothesis is 95% likely to be correct. Rather, a p-value of 0.05 just means that if the null hypothesis is true, and all other assumptions are valid, a 5% chance exists of obtaining a result at least as extreme as the observed result" (Section 10.3, Zybooks, 2016). What this means in laymen terms is that one should not rely only in a p-value to determine significance. Rather, a p-value should be only one piece of the puzzle to determine significance.

Last but certainly not least, it is of the upmost importance to be ethical in all processes of data analytics. A data analyst should point out methods used, possible flaws or conflicts, explain the results clearly and ensure confidentiality in all of his/her communications. It is so important as an analyst to have high ethical standards that induces a desire to seek and obtain accurate and truthful statistical results (Section 10.5, Zybooks, 2016).

G. See Excel

H. It could be expected when running a simple Monte Carlo with 1,000 iterations that the precinct in the scenario would have an approximate average of 2.12 officers per an incident. Also, the probability that the precinct would continue to meet the minimum of 2.5 officers per incident is 45.00 percent. It is reasonable to infer that this precinct would not meet the minimum of 2.5 officers per an incident to qualify for the additional funding. It is also a reasonable estimate with a 45.00 percent probability that the precinct would not continue to meet the average of 2.5 officers per an incident which is needed to maintain it eligibility. See graph below for more detail.

Precinct Eligibility	
Average Officers per Incident	2.124797
Incidents w/ Less than 2 Officers	550
Probability to have **"LESS"** than 2 Officers	55.00%
Probability to have **"MORE"** than 2 Officers	45.00%

It is important to note that sector H and one other data point has been removed from the Monte Carlo simulation. These data points skew the data as they are outliers, data points that lay far beyond the normal boundaries of the data. If these data points were left in the Monte Carlo and the linear regression, it would be a less accurate predicter of current and future eligibility.

References

Section 10.2, ZyBooks. (2016, August). *https://my.zybooks.com/#/zybook/WGUFundamentalsOfDataAnalytics/chapter/10/section/2*. Retrieved May 18, 2017

Section 10.3, Zybooks. (2016, August). *https://my.zybooks.com/#/zybook/WGUFundamentalsOfDataAnalytics/chapter/10/section/3*. Retrieved May 20, 2016

Section 10.5, Zybooks. (2016, August). *https://my.zybooks.com/#/zybook/WGUFundamentalsOfDataAnalytics/chapter/10/section/5*. Retrieved May 20, 2016