Name: Amy Nguyen

TASK 1: FUNDAMENTAL DATA ANALYTICS

Scenario: A local police department is interested in discovering the behavior, trends, and needs of the department based on data that has been collected. As a data analyst, you have been recruited to do consulting work for the department.

Part 1: The police chief asks you to analyze the logs from emergency 911 calls in the city and then provide a summary of that data.

A.) Cleaning the Raw Data Spread

- The cleaned data has been attached as "CleanData.xls".

B.) Data Cleaning Change List

- The column "OFFICERS_AT_SCENE" was changed to "Officers at Scene" to ensure consistent column header formatting within the table.
- No row duplicates were removed as the "Remove Duplicates" tool in Microsoft Excel did not find any duplicates based on "CAD CDW ID", "CAD Event Number", and "General Offense Number". "CAD CDW ID", "CAD Event Number", and "General Offense Number" were used as columns for uniqueness as these numbers should be unique from one incident to the next.
- The columns "CAD Event Number" and "General Offense Number" were removed. The information requested for the graphs were; Date, Number of Events, Number of Incident Occurrences by Event Type, Sectors, and Number of Events. Thus, neither of the two columns are necessary to draw information from. Furthermore, while the two columns do provide a unique identifier for the incident row, only 1 unique identifier is necessary which is why "CAD CDW ID" was not removed.
- The row where "CAD CDW ID" = 1702543 was removed since column "District/Sector" was empty and "Zone/Beat" was filled in with a value that does not correspond with the other values in the column.
    - "Zone/Beat" was filled in as "FS" whereas it should be filled in as "Alphabet, Number" without any commas or spaces (i.e. F1).
- The columns "Hundred Block Location", "Longitude", "Latitude", and "Incident Location" were each removed from the dataset. The previously listed columns in addition to columns "District/Sector" and "Zone/Beat" all provide information on incident location. Since columns "Hundred Block Location", "Longitude", "Latitude", and "Incident Location" all provide redundant information about the incident location and are not necessary for graphing reasons later on, they were removed.

- The column “At Scene Time” was removed from the dataset. Of the 1006 rows of data that were sampled 605 rows were empty for the column. Also, while “At Scene Time” does include a date which is required for a table and graph later on, the column “Event Clearance Date” also provides information on the incident date. Thus, since approximately 60% of the data is missing for the column “At Scene Time” and the critically information needed from “At Scene Time” can be found from another column, the column “At Scene Time” was removed.
- The column “Event Clearance Date” has been adjusted so that only date can be seen since the time is not necessary for the requested tables and graphs.
- The column “Census Tract” was removed since it does not contain any relevant data for the tables and graphs that are to be created.
- The column “Zone/Beat” has been removed since it contains redundant data that is generalized in column “District/Sector”.
- The columns “Initial Type Description”, “Initial Type Subgroup”, and “Initial Type Group” have all been removed as the current data of interest is for the confirmed event type groups and not the initial event type groups.
- The column “Event Clearance Code” has been removed as the column “Event Clearance Group” provides a better understanding over the same information
    - The “Event Clearance Code” gives a numerical classification whereas “Event Clearance Group” gives a text descriptor which is easier for a person to understand and look at.
- The column “Event Clearance SubGroup” has been removed since it is redundant to column “Event Clearance Description” which does a better job of further breaking down/giving further classification to “Event Clearance Group”.
- The column “Officers at Scene” has been left in the data table since it is relevant information for Part 2.

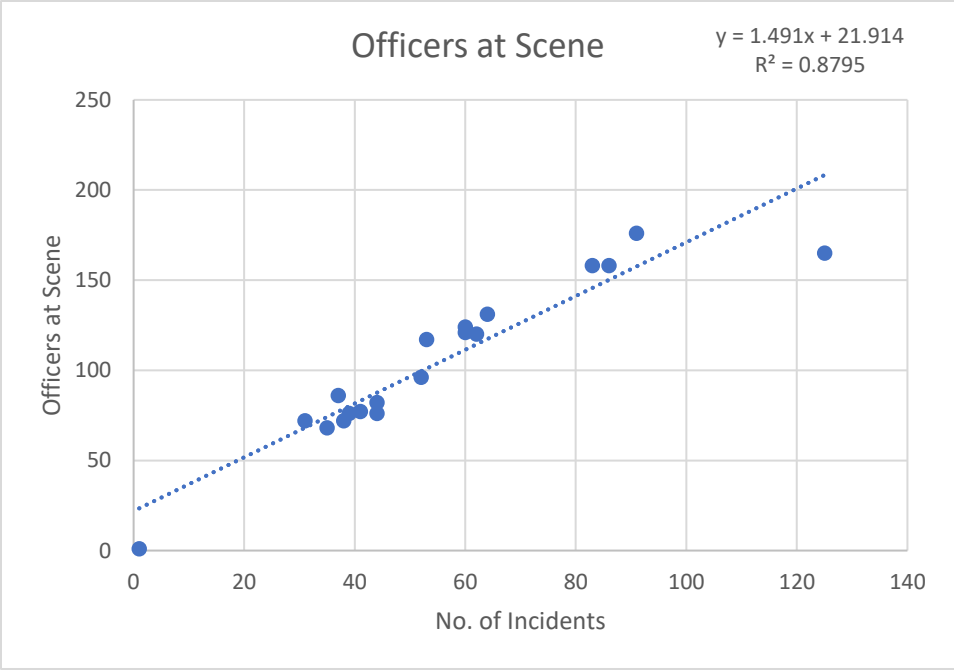
C.) <u>Tables & Graphs</u>

- The table and bar graph for date and number of events can be found in sheet “TG_Date” of “CleanData.xls”.
- The table and bar graph for number of incident occurrences by event type can be found in sheet “TG_Event_Type” of “CleanData.xls”.
- The table and bar graph for sectors and number of events can be found in sheet “TG_Sector” of “CleanData.xls”.

D.) Summary

- From the data tables and graphs created from the cleaned data we can observe the following points:
    - The majority of the 911 calls sampled occurred on Sunday, March 27th of 2016. The specific breakdown for the sample is that approximately 23.2% of the calls occurred on March 26th, 55.8% occurred on March 27th, and 21% occurred on March 28th.
    - The reasons for the 911 calls can be broken into 32 different categories. Of the 32 categories, 29 of the categories have less than 75 incidents reported and of those 29, 13 of the categories have less than 10 incidents reported. The lowest number of reports with 1 report each are "HARBOR CALLS" and "WEAPON CALLS". The top three categories for the calls that made up roughly 46% of the data were "DISTURBANCES" at 150, "TRAFFIC RELATED CALLS" at 164, and "SUSPICIOUS CALLS" at 167. The median number of calls made per category is 15 and the mean number of calls made per category is approximately 33.
    - 18 sectors were involved in the sampled 911 calls. The lowest number of calls was for sector O at 31 and the highest number of calls was for sector H at 125. The median number of calls per sector is approximately 53 and the mean number of calls per sector is approximately 58.

Part 2: The state governor has offered an additional funding incentive for police departments that are able to meet the standard of having a minimum of 2.5 officers onsite per incident. The police department has asked you to analyze their data to determine if the department will be eligible for additional funding, using the attached linear regression.



District Sector	No. of Incidents	Officers at Scene	Fitted Officers at Scene	Residuals
B	83	158	145.667	12.333
H	125	165	208.289	-43.289
W	37	86	77.081	8.919
K	64	131	117.338	13.662
D	60	121	111.374	9.626
O	31	72	68.135	3.865
U	52	96	99.446	-3.446
R	60	124	111.374	12.626
S	44	82	87.518	-5.518
	1	1	23.405	-22.405
J	41	77	83.045	-6.045
Q	62	120	114.356	5.644
L	38	72	78.572	-6.572
C	44	76	87.518	-11.518
M	91	176	157.595	18.405
N	53	117	100.937	16.063
F	35	68	74.099	-6.099
G	39	76	80.063	-4.063
E	86	158	150.14	7.86

The above graph was provided by the given material "Linear Regression.xlsx". The table was provided by the given material "Linear Regression.xlsx", but was modified to include the fitted number of officers at the scene based on the linear regression equation and the resulting residuals.

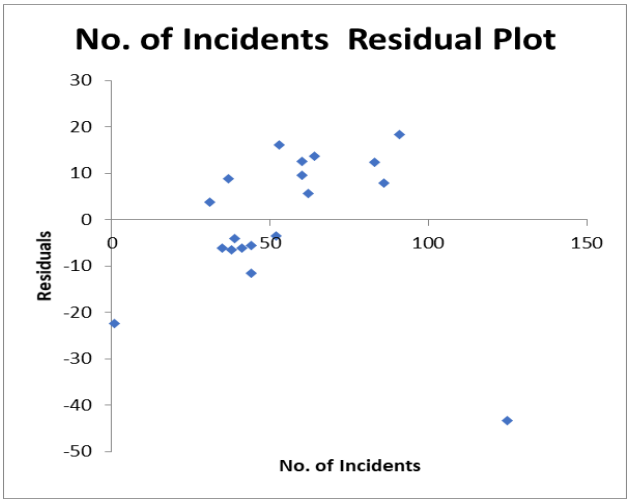
E.) Describe the fit of the linear regression line to the data.

- The fit of the regression line to the data can be described by the R-squared value which is known as the coefficient of determination. The R-squared value is often used as a measure of goodness of fit since it measures the percentage of the variance in the data explained by the model. For the given linear regression model y = 1.491x + 21.914 the R-squared value is 0.8795 or 87.95% which is considered on the high side since 100% is the highest value. That being said typically one would want an R-squared value greater than 90% so the fit of the linear regression data can be considered acceptable, but not the best. This standpoint can be further reiterated if one looks at the ANOVA table generated for the data on sheet "LR_Outlier" of "Cleaned_Data.xlsx". The Residual Sum of Square or SSE for the data is measured to be approximately 4060. This can be possibly considered to be high since one should aim for an SSE value as close to 0 as possible since SSE measures the error or discrepancy between the data and the linear regression model.

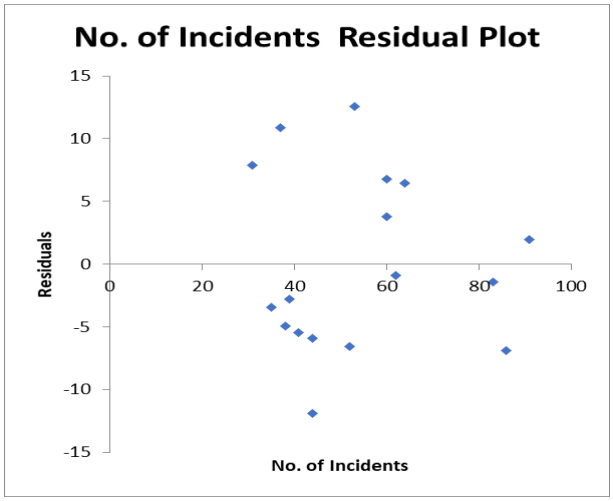
F.) Describe the impact of the outliers on the regression model.

- The values that can greatly affect the accuracy and fit of a regression model are outliers. Outliers are any values that are significantly then the other values in a set of data. These values can greatly impact a regression model because they often result in higher residual values.
- In the given data I consider two values to be outliers. The first outlier is the data point for the unknown district sector. It has 1 incident report with 1 officer at the scene. The data point for the unknown district can be considered an outlier since it is from an incomplete data point that does not have a listed district sector. It would be unwise to keep the data point in the model since there is no way of knowing if it should be classified as its own datapoint or if the data should actually be included in a different district. Furthermore, it has a higher residual value in comparison to most of the other datapoints. The second outlier is district H. District H is a clear outlier that is negatively affecting the accuracy of the linear regression model as it has an absolute residual value of approximately 43 which is much higher than the other datapoints. One can also see from looking at the Linear Regression graph that the data point for district H also lies a significant distance away from the regression line in comparison to each other the other datapoints.

G.) Create a residual plot and explain how to improve the linear regression model.



G1. Residual Plot with Outliers



G2. Residual Plot without Outlier

- The best way to improve the linear regression model is to remove the outliers from the data. From the residual plots above, the residual plot that includes outliers in the data or G1 has a large residual range of -50 to 30. The residual plot without the outliers or G2 on the other hand has a much smaller range of -15 to 15. Furthermore, there is a more even spread of residuals on plot G2 over plot G1.

- One can continue to see that removing the outliers from the data greatly benefits the linear regression model by comparing the R-squared value and SSE values for the two regression models as well. The R-squared value of the model with the outliers is computed to be 0.8795 or 87.95% while the R-squared value of the model without the outliers is computed to be 0.9591 or 95.91%. The model without the outliers also has an improved SSE value of approximately 788 which is much closer to 0 then the original 4060.
    - Please refer to sheets "LR_Outlier" and "LR_No_Outlier" for the graphs and ANOVA table with this information.

H.) Using the linear regression analysis, explain if the department qualifies for additional state funding, including any limitations posed by the available data to the assessment of the department's current funding eligibility.

- If one were to look at the mean number of officers per incident of the sample data, that is if one were to take the total number of officers and divide it by the total number of incidents, the mean number of officers per incident for the sample data is approximately 1.9. Therefore, based on this information, the department does not qualify for funding.

$$y = \frac{Total\ Officers\ at\ Scene}{Total\ \#\ of\ Incidents} = \frac{1976}{1046} = 1.889$$

- Although it was found that the department does not qualify for additional funding it is important to keep in mind that there is one large limitation to the data. The limitation to the data is that it is sampled from only three days' worth of incidents. Three days is a very small amount of time and does not allow for an accurate or thorough understanding of the average number of officers per incident. It would be better to use a larger dataset that covers a larger range of dates in order to create a more accurate model.

I.) Describe the precautions or behaviors that should be exercised when working with and communicating about the sensitive data in this scenario.

- In this scenario there are several behaviors that one should exhibit when working with the data. The first is to always make sure one is analyzing data with integrity and accountability. In the current scenario, the data will be used to determine additional funding for the police departments that are able to meet the standard minimum of 2.5 officers per incident. It is important that the analyst does not feel swayed by the idea that there is a reward if the results point towards a certain result. This is so that the analyst does not do anything to intentionally or unintentionally manipulate the data towards a more favorable outcome which in this case could be manipulating the data to meet the 2.5 minimum.

- Another precaution that an analyst must take is to make sure that the analyst is as transparent and truthfully as possible. The methods for analysis that are used and the data used should be clearly presented in a manner that does not hide any information that could significantly impact the findings of the data.
- Lastly any private information should remain confidential. In this particular scenario the private data of addresses were censored so that one couldn't actually find those locations.
- In conclusion, it is important that the analyst working on the data does so in an ethical and professional manner that is unbiased, transparent, and without hidden motives. Furthermore, the analyst must make sure to protect the confidentiality of any bits of data that could be considered sensitive and/or private.