Remove the following columns:

* Column F: Event Clearance SubGroup - same information as column E.
* Column G: Event Clearance Group - same information as column E.
* Column P: Initial Type Description - same information as column E.
* Column Q: Initial Type SubGroup - same information as column E.
* Column R: Initial Type Group - same information as column E.
* Column M: Longitude – information is contained in Column O.
* Column N: Latitude - information is contained in Column O.
* Column S: At Scene Time – too many missing values.

Remove the following row:

* Row: CAD CDW ID 1702543 – too many missing values

**Data sheets**

|  |  |
| --- | --- |
| Date | Number of Events |
| 3/26/16 | 243 |
| 3/27/16 | 583 |
| 3/28/16 | 219 |

Table 1.1: Number of Events by Date

Figure 1.1: Number of Events by Date

|  |  |
| --- | --- |
| Event Clearance Description | Number of Events |
| DISTURBANCE, OTHER | 123 |
| SUSPICIOUS PERSON | 105 |
| PARKING VIOLATION (EXCEPT ABANDONED VEHICLES) | 67 |
| TRAFFIC (MOVING) VIOLATION | 65 |
| MOTOR VEHICLE COLLISION | 62 |
| TRESPASS | 54 |
| THEFT - CAR PROWL | 53 |
| LIQUOR VIOLATION - INTOXICATED PERSON | 50 |
| MISCHIEF, NUISANCE COMPLAINTS | 39 |

Table 1.2: Number of Incident Occurrence by Event Type

Figure 1.2: Number of Incident Occurrence by Event Type

|  |  |
| --- | --- |
| District/Sector | Number of Events |
| H | 125 |
| M | 91 |
| E | 86 |
| B | 83 |
| K | 64 |
| Q | 62 |
| D | 60 |
| R | 60 |
| N | 53 |
| U | 52 |
| C | 44 |
| S | 44 |
| J | 41 |
| G | 39 |
| L | 38 |
| W | 37 |
| F | 35 |
| O | 31 |

Table 1.3: Number of Events by Sector

Figure 1.3: Number of Events by Sector

* From table 1.1 and figure 1.1, the date with the highest number of events is 3/27/16 with 583 incidents.
* From table 1.2 and figure 1.2, type “Disturbance, Other” has the highest occurrences, with 123 incidents.
* From table 1.3 and figure 1.3, sector H has the highest number of events with 125 incidents. Sector L, W, F and O are safest with less than 40 incidents.

Using the excel tool Data Analysis for the data in Linear Regression.xlxs, the summary statistics table is produced:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 21.9141501 | 8.1767407 | 2.68005932 | 0.01581947 |
| No. of Incidents (X) | 1.49104316 | 0.1338378 | 11.140673 | 3.1087E-09 |

Table 2.1: Coefficients and p-value

Figure 2.1: Linear Relationship Between Number of Officers at Scene and Number of Incidents

From figure 2.1, we have the simple linear regression model:

The estimators are and .

Based on the information from the linear regression model, for every incident x increase, an additional 1.49 officers appear at the scene.

The value of coefficient of determination R2 = 0.8759, which means approximately 87.59% of the data points should fall within the regression line. This percentage suggests a fairly decent model.

The p-value from table 2.1 is 0.0158, which is less than 0.05. This suggests a statistically significant finding, and all other assumptions are valid.

While most of the data points are clustered around the regression line, there are two points that lie far away from the main cluster of the data, which are (1,1) and (125,165). These points are outliers.

After removing the outliers, we obtain the new scatter plot:

Figure 2.2: Linear Relationship Between Number of Officers at Scene and Number of Incidents after removing Outliers

The initial fit for the regression line equation was

, with R2 = 0.8759.

After removing the outliers from the dataset, the new equation is:

, with R2 = 0.9591.

The removal of the outliers yields a new value for R2 that is much closer to 1. The new value of R2 = 0. 9591, which means approximately 95.91% of the data points should fall within the regression line. This high percentage suggests a better model than the previous one and is likely to produce much better predictions.

**Improve the linear regression model based on interpretation of the plot.**

Figure 2.2. Officers at Scene Residual Plot

In general, the residual plot above does not show a clear pattern. The positive and negative residuals is scattered around zero line. This random pattern indicated that the linear model provides a decent fit to the data.

According to figure 2.2, the residual error of point (1,1) and (125,165) are -22.405 and

-43.294 respectively, and they also have the furthest distances from the zero line. Therefore, removing them will reduce the degree of scattering, as well as improve our linear regression model.

The initial fit for the regression line equation was:

After removing the outliers from the dataset, the new equation is:

The estimator represents the minimum number of officers at scene when there is no incident. The values of for both models are 21.914 and 7.3058, and they are greater than 2.5. Therefore, the police departments are eligible to receive addition funding from the state governor.

There are some limitations posed by the current data that will affect the result of this study. The sample size is small, which may not represent the current situation of all police departments of the entire state. Also, there are some areas that will have higher crime rate and require more attention than others, and vice versa. The researchers have to take these aspects into consideration before making the final decision.

**Precautions**

* Misuse of statistics: The sample size is 19, which is small. This is called bad data, so the researcher should collect more samples.
* Transparency: The state governor will be funding this research so there is a financial conflict of interest. The researcher may not include the data that doesn’t support the desired conclusion. Therefore, the researchers need to be honest about their data.
* Confidentiality: The information such as actual address that an incident occurs, victims’ names, officers’ names should not be collected by the researcher.
* Privacy: The data that is collected by the researcher should not be seen by everyone. The research should protect the data with some form of protection such as password, and only grant access to authorized users.