Fundamentals of Data Analytics – C740

Student Name: David Bailey

Student ID: #000658614

Date: 10/12/2019

Student Mentor Name: Kelly Smith

Dataset Preparation

The dataset seemed fairly clean after using conditional formatting to check the event ID number for any duplicate records. I did, however, remove one row and one column. Row 226, with “CAD Event Number” 16000105361, was missing a value in the “District/Sector” feature and needed to be removed to prevent it from being classified in the wrong district, or as a blank district of its own. The column “At Scene Time” was only filled in 400 out of 1047 times and isn’t useful with that many missing values. The remainder of the data was complete. No outliers are present in the target feature of “Officers At Scene.” While all features might not be needed for this regression analysis, they could be helpful in later analysis.

# Data Sheets Observations (See Pivot Tables and Graphs in “Raw Data”)

At first glance, March 27th looks to have more than twice as many events than the 26th or 28th, but if you expand the pivot table you see a different story. The 26th and 28th don’t have a full day’s reporting, which makes the event totals look lower for those days. When you look at the average number of events per hour for the reported hours, March 26th has the most events, at an average of 34.71 per hour, compared to 24.29 and 18.25 for the 27th and 28th, respectively.

For the reported period, three event types (event clearance groups) were much higher than the others. Disturbances had 167 events, traffic-related calls had 164, and suspicious circumstances had 150. The events with the minimum number of calls were weapons calls and harbor calls, with one event each.

The range of events by sector is 94, with the minimum at 31 for sector O and the maximum of 125 for sector H. The mean number of events per district for this period is 58.06.

# Fit (See Table in “Regression” Worksheet)

The linear regression line doesn’t seem incredibly far off. Visually, it appears that two values are errors or outliers and are hurting the fit of our linear regression. Looking at the table of the absolute value of errors from expected and actual values, two values seem much higher than the others, which seems to confirm that these two data points should be inspected to see if we can get the linear regression line to fit better.

# Outliers (See “Regression – No outliers” Worksheet)

The two outliers affected the slope and intercept of our linear regression line equation, and the R2 value quite a bit. The one outlier labeled district “ ” is actually an error that should have been caught in the data scrubbing. The missing value for that record should have been removed or audited so that it could be attributed to the correct sector. Instead, this single record is having much more effect on the regression line that it would if it were correctly included with the right district.

After removing the two outliers, we can see that the linear regression line equation changes. The slope increases and the intercept decreases. It also visually shows a better fit to the remaining data, has a lower sum of absolute error, and a higher R2 value.

# Residuals (See “Residual Plot” Worksheet)

The residual plot shows that most of the data, other than the two outliers, is a fairly constant distance from our expected values. My recommendation again would be to remove the two large residual values that appear to be outliers.

# Qualification (See “Qualification” Worksheet)

The data that has been provided is a very small sample compared to the population. Only about 48 hours of event data is available in the sample, and depending on many factors surrounding those two days, could lead our analysis in a path away from the true population parameters. Based on this limited data set, we assume with a 95% confidence level, the department would not qualify for the additional funding. The threshold amount of officers per event would not be reached from the assumptions we gather from this small sample.

# Precautions

Since the raw data file lists longitude, latitude, and descriptions of possible criminal activity, this data set should be kept confidentially and securely in the precinct. This information can’t be excluded from the gathering process, because it is needed by the department’s officers at the time of a new event and likely for future audits. The analyst should use caution when transferring the files, accessing them on insecure networks, and storing them on local hardware. Adequate password protection and file encryption should be used, and the analyst should only access files in the presence of authorized people.