AKM1

Michael Segaline

C740

10/31/2020

Abstract

The following data analysis is uncovering trends for the Seattle Police Department along a three day time frame. The data is well structured with few null values and all date -time groups are in a consistent format. The dataset contains 1047 rows and 20 columns. The column labeled “At Scene Time” has many null values and will be deleted. One row, number 460 has three null values for location and will be deleted as well. The ultimate goal of this analysis is to help the Seattle Police Department determine eligibility for the new funding standards set forth by the Governor. This paper will detail the data preparation and mining process, along with the linear regression to ultimately determine: the Seattle Police Department does not meet the eligibility requirements of 2.5 Officers per incident as per the Governor’s requirements.

Part 1

A.  Prepare a dataset from the data provided in the “Raw Data” spreadsheet. Remove any potential errors or outliers, duplicate records, or data that are not necessary.

B. The following columns where deleted with an explanation as to why. “At Scene Time”, because it has too many missing values and does not relate to the purpose of the analysis. “Initial Type Group”, “Initial Type Subgroup”, and “Initial Type Description” where deleted because they reflect the initial but not necessary verified incident descriptions. “Incident Location”, “Latitude”, and “Longitude” columns where deleted because they are redundant to the indicated sectors and unnecessary for management to utilize. The “Census Tract” column could be used later as a metric for police officer number proportioning, however, not required for the task.

“Zone /Beat” and “Hundred Block Location” are redundant because the bigger picture i/e “District/ Sector” column is what is need for this analysis. The columns, “Event Clearance Description” and “Event Clearance Subgroup” are redundant and too finely drilled down for this exploration; the “Event Clearance Group” column appears to be the top of the hierarchy and implies the final result as it received a code as compared to the “Initial Type Group”. The columns: “Event Clearance Code” was deleted due to brevity requiring a dictionary type look up. Plus, the “Event Clearance Group” descriptions will be less visually confusing on the following bar graphs. Lastly, the “Cad Event Number” and “General Offense Number” columns where deleted due to brevity and would cause visual confusion to graphs.

A total of four columns remain for analysis: “CAD CDW ID”, “Event Clearance Group”, “Event Clearance Date”, “District/Sector” and “OFFICERS\_AT\_SCEENE”. Upon checking for duplicate values, there are no apparent duplicate values as it pertains to the CAD CDW ID; all CAD CDW ID numbers are unique. Upon searching for null values, one was located with a blank sector bearing the CAD CDW ID of 1702543, the row was deleted. Up to this point the data appears to be clean, with unique CAD CDW ID numbers and no erroneous record are evident. The current data size is 1046 rows and five columns. Furthermore, the “Event Clearance Date” column, which has a date / time format, was split to sperate the date from the times. This was done so there would be less confusion on the visualization, since every date time stamp for the three days is unique and counter to analysis. Therefore, only the date is remaining in the “Event Clearance Date”.

The “Event to Clearance Date” column is now going to by split to isolate only the date and not the times. If the times are included, the bar graphs will be unreadable because each date -time value will count as a unique entity; we only want the date, not the time as well. That is done in Excel by inserting a new column and naming it date and calling a function “left”, indicating the rows to start and adding the argument numeral “5” to isolate only the date. After highlighting the date and formatting the converted number to date, the result is a the date without a timestamp. The new column has been named “Date”.

C. An “Events by Date” sheet has been added to the Excel document in which the two features have been copied and pasted from the “Clean Data” sheet: Event Clearance Group and Date. On said sheet, a pivot table was used on the “Date” column to get a count of events by date and visualize them via a bar graph. The bar graph is labeled “Events by date”. The next spreadsheet, “Incidents by event type”, had one column copied to it: “Event Clearance Group”. On said column a pivot table was used to get a tally of unique events. From the pivot table result, a bar graph named “Incidents by Event type” was produced and labeled on the same sheet; with Number of Incidents on the Y axis and Incident Event Types on the X axis. The final sheet, “Sector by Events”, had the “District/Sector” column analyzed by a pivot table; resulting in the unique sectors and totals per sector. Of which, a bar graph labeled “Events by Sector” was produced, with the Number of Incidents on the Y axis and sectors on the X axis.

D. Summarizing the observations of the graphs of each created spreadsheet, starting with “Incident by Event Type”. It appears that the date 3/27/16 had a significant spike in events, totaling 583. More than double that of the previous day (243) and following day (219). Next, examining the frequency of the “Incidents by Event Type”, there are 32 unique incident event types. Of these event types, it appears that “Disturbances” where the most commonly reported with 167. A close second was “Traffic Related Calls” totaling 164 incidents and in third was “Suspicious Circumstances” totaling 150. When analyzing the incidents by sector, there are 18 total sectors with incidents reported. Sector H is distinctly in the lead with 125 incidents. Sectors: M, E, and B are the nearest comparable, between 86 to 91 incidents.

Part 2

E. The Liner Regression analysis spreadsheet has been simplified to 18 defined sectors and one blank sector, by the Seattle PD, tallying the Number of Incidents per sector and Officers at Scene per sector visually enhanced by a scatterplot. A linear regression chart has been previously constructed comparing two tally’s: “Officers on Scene” on the Y axis and “Number of Incidents” on the X axis. There is a regression line of best fit bisecting the scattered plots in a positive direction indicating the correlation between the number of incidents and Police Officers at scene. Noted on the graph is the coefficient of determination scoring 0.8795 with a total equation of Y = 1.491X + 21.914. According to the course book, the “R squared” or “Coefficient of Determination” is the yielded from the Analysis of Variance Table (ANOVA). The ANOVA table produces five metrics that yield the R squared. The Coefficient of Determination is used to detect if a strong linear relationship between Y and X variables. The relationship is measured between 0 and 1 as a percentage (out of 100) of overall accuracy of X predicting Y. Therefore, as the model stands now, there is an 87.6 % chance of the model predicting the amount of officers that respond to an incident. However, there are two glaring outliers: one located at 1,1 and one located 125, 165. These two outliers will have to be removed to improve the line of best fit and coefficient of determination score.

Furthermore, Data Analysis tool was utilized to get the baseline scores of the model as well as visually represent the X Variable Residual Plots and X Variable Line of Fit Plot. The “Number of Officers” and “Number of Incidents” columns where pasted onto a separate spreadsheet labeled “No Outliers”. Of which, the two stated outliers where removed and a scatterplot with a regression line was created. The Excel Data Analysis tool was utilized producing the Summary Output, ANOVA, and Residual outputs. A noticeable difference in the coefficient of determination, now up to 0.9591, the model is approximately nine percent more accurate. Since removing the outliers and the regression line is now the residual plots are grouped closer to the regression line, indicating a much better fit.

F. With the outliers removed, the degrees of freedom for the residual plots have decreased from 17 df to 15 df. The formula of the new regression line is Y =1.8324X + 7.3058. As stated previously the line of best fit has been improved. Furthermore, with the outliers removed the residual standard error has been improved from 9.72 deviations from the regression line to 3.87 deviations. The F statistic has shrunk dramatically from 3.11 x 10^9 to 8.01 x 10^12.

G. A separate spreadsheet was added named “Residual Added”, in which an additional sector named “Z” was created with a notional 75 Incidents and 140 Officers per Scene. A scatterplot graph was created named “Officers per Scene (with notional point)”. A trendline was added through the points with a similar positive slope. This notional point was not created at random, but in trying to predict if the new plot would touch the regression line or come close to it. The graphs with Summary output, and ANOVA prove that the notional plot is touching the line. The Coefficient of Determination has been improved from 0.959 to 0.96. The increase would suggest that the model becomes more accurate with more substantial data points being targeted near the line of best fit. Moreover, the true accuracy of the model can be improved with much more data points. This analysis could be case of “bad data” as indicated per the course textbook, because the sample size is small, less than 20. Unfortunately, this dataset is a generalization from a sample, only giving a small three day window of one year of police operations and should not be considered reflective of the average number of Officers responding to an incident.

H. Located on the spreadsheet labeled “No Outliers”, a tallying of both “Number of Incidents” and “Officers on Scene” has been noted. Additionally, a formula created in the cell below the tally’s named “Officers per Incident”, was created by dividing the “Officers on Scene” tally by the “Number of Incidents” tally. The result is 1.97 Officers per Scene, below the threshold necessary for the state funding set forth by the Governor. Unfortunately, the Seattle Police Department does not qualify for additional funding based off of the given dataset.

I. Cautions and behaviors when working with sensitive data in this scenario. As per chapter 8 of the coursebook: names of officers, names of supervisors and addresses of officers should be kept omitted or anonymized. The spreadsheet could be password protected as well as the device used to analyze the data (unless the data comes from a public repository). If leaving a computer, ensure that it has been properly logged out of. Now knowing that less than two officers respond to any given event during the analyzed days, a criminal could possibly use this information to their advantage. Furthermore, other metrics such as Police Officer response times to incidents, could be exploited by criminals as well. It would be ethically incumbent on the analyst to explain to the Seattle Police Department management, how one years-worth of data could yield a more accurate result and possibly allow them to qualify for the Governor’s monetary incentive.

Work Cited

*Fundamentals of Statistics for Data Analytics*. (n.d.). Retrieved October 21, 2020, from https://learn.zybooks.com/zybook/WGUC740V52018/chapter/8/section/5