C740 – Performance Assessment of the Fundamentals of Data Analytics

Morrell J. Parrish

Western Governors University

Abstract

The governor of Washington has offered an additional funding incentive for police departments throughout the state. The Seattle police department requested that their incident logs from March 26, 2016 through March 28, 2016 be reviewed and analyzed to determine if they qualify for additional funding. In order to secure additional funding, the Seattle police department needs to meet the minimum standard of having 2.5 qualified officers on the scene per incident (Western Governors University, 2020)

C740 – Performance Assessment of the Fundamentals of Data Analytics

The local police department has requested help in securing additional funding; in order to secure funding, they have provided a raw data file, this file contains incidents ranging from March 26, 2016 through March 28, 2016.

Part 1: Data Analysis

Explain why you removed each column and row from the "Raw Data" spreadsheet, or why you imputed data in empty fields as you prepared the data for analysis.

The data set submitted by the Seattle police department contains info from March 26, 2016 to March 28, 2016; however, not all of the data is useful or will be used. The following columns will be used for this analysis: *event clearance code*, *event clearance date*, *event clearance group, district and sector, zone and beat, and officers at scene*. The below columns were exempt from analysis and not used because they did not contain any useful data and/or was missing data.

- **Columns A** thru **C** are unique incident identifiers and are assigned to only one specific event and are not relevant to the study.
- **Columns E** and **F** are subcategories and descriptive fields that describe the main "**Event Clearance Group**" column G; therefore, they will not be utilized for this analytical review.
- **Column I** is the address of the incident; addresses are tied to districts and zones, which can be represented by columns J and K.
- **Columns L** thru **O** are locations coordinates; since we are using columns J and K - districts and zones, we do not need the coordinates of the incidents.

- **Column Q** is a subcategory of **Column R**. Both of these columns contain redundant data and are best represented by **Column G** – the event clearance group.
- Column S is missing data and will not be used; however, it should be noted because of the vast amount of missing data, the department may want to review their data entry procedures; a review of their process should correct the inaccuracies of the missing data.

Create data sheets using your cleaned dataset, provide each of the following to represent the requested aggregated data.

The following datasheets are located in the attached excel file "**Assessment_Morrell_Parrish**".

- Clean DataSet
- Number of Events by Date
- Incidents by Event Types
- Number of Events by Sectors
- Linear Regression Chart w/ Outliners
    - Regression Output 1
- Liner Regression Chart w/o Outliners
    - Regression Output 2

Summarize your observations from reviewing the data sheets you have created.

The submitted data from the Seattle police department contains incidents covering a three-day period from March 26, 2016 to March 28, 2016. Upon initial review of the datasheets, you can see that the reported incidents on day two more than doubled the reported incidents on day one and day three. The types of incidents varied over the course of three days; disturbances were the most reported type incidents; they accounted for about 16% of the reported incidents.

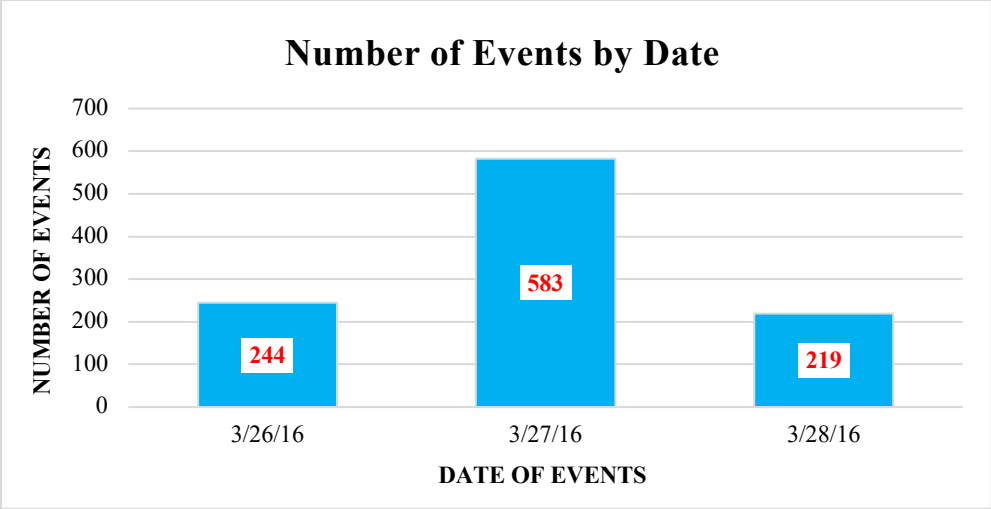

***Figure 1**: Events by Date*

The top ten incidents including disturbances were traffic related calls, suspicious circumstances, false alacad, motor vehicle collision investigations, car prowls, liquor violations, trespassing, mischiefs (nuisance), and behavioral health. I also noticed that the top three reported incidents accounted for about 46% of the total reported incidents.

After reviewing the incidents by district-sectors; there are two outliners in the data set - **district - sector H** and the **unknown** district. After removing or discarding the outliners, district M has the most reported incidents with 91 incidents and 176 officers on scene, which accounts for about 10% of the total incidents reported.

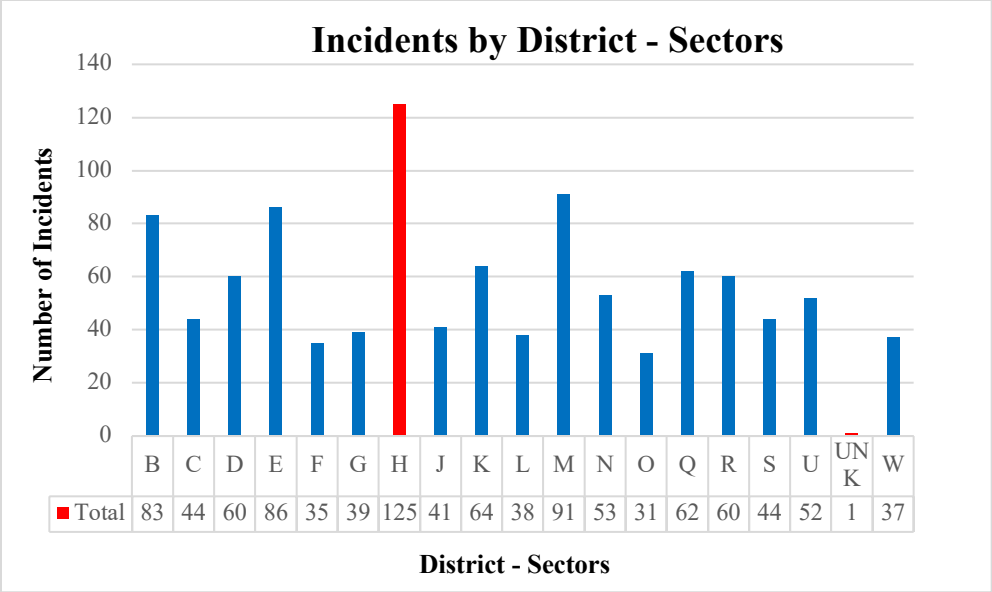

***Figure 2**: Number of Incidents by District - Sectors*

Part 2: Recommendation

Describe the fit of the linear regression line to the data, providing graphical representations or tables as evidence to support your description.

The regression model shows that there are two outliners **district H** (125, 165) and the **unknown district (**1,1). The linear regression line with the outliners has a value of *1.491x + 21.914* and without the outliners the value is *1.8324x + 7.3058*. The linear regression figure without outliners shows that there are less officers on scene when the reported incidents are low; it also shows that there is an average of *1.8* officers on scene per incident; also, the upward trend shows when incidents increase so do officers on scene respectively. In order to determine which linear regression model has the best goodness of fit; we will need to compare the statistical measure of *R-squared*; which measures how close the data are fitted to the regression line, R^2 is always between 0 and 1. In the first linear regression model with outliners the R^2 *is .08502;* this indicates that 85% of the variance in *y* is predictable from *x*. In the second linear regression model without outliners the R^2 is .9591; which, indicates that 95% of the variance in *y* is predictable from *x*; so, the second linear regression model without the outliner is the better fit.

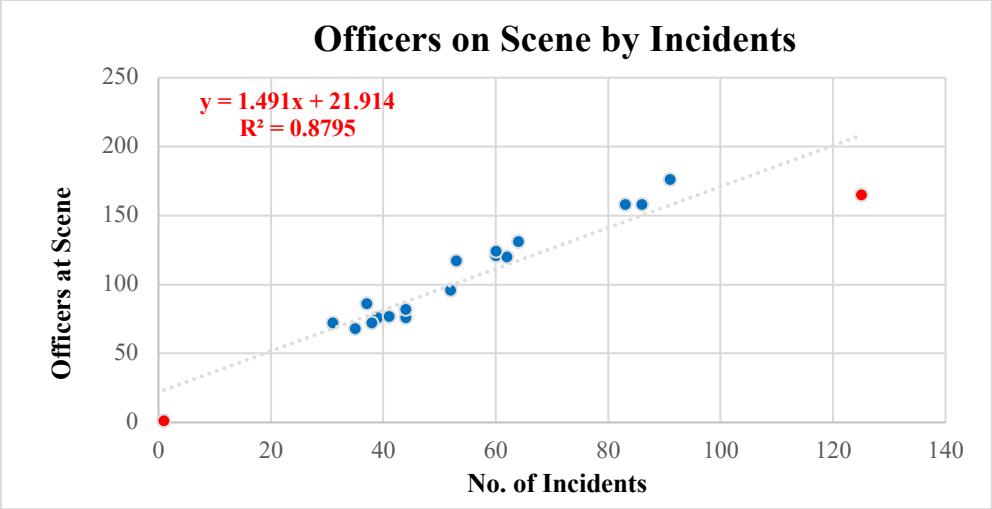

Figure 3: *Linear Regression Chart with Outliners*

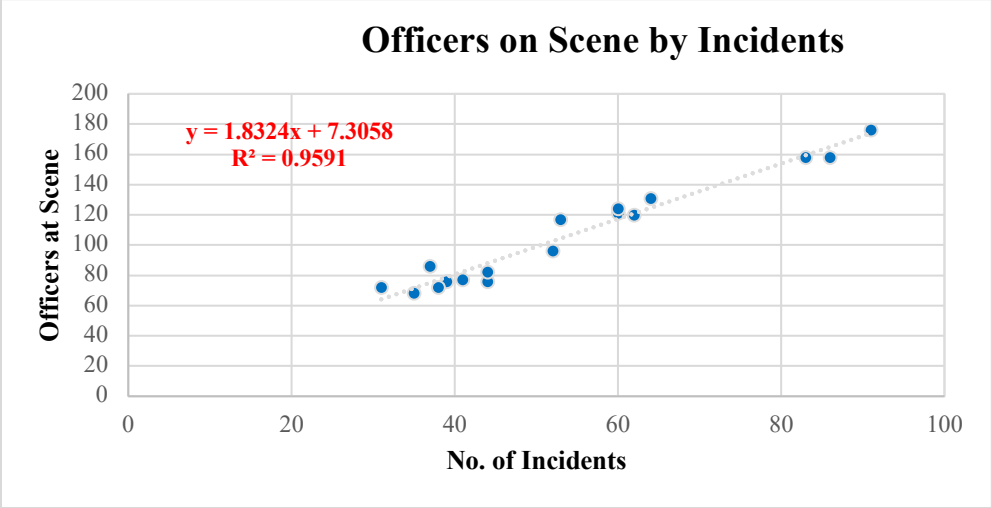

Figure 4*: Linear Regression Chart without Outliners*

Describe the impact of the outliers on the regression model, providing graphical representations or tables as evidence to support your description.

District H and the **unknown district** are both outliners; these outliners have the highest difference between the actual value and the expected value given in the linear regression; additionally, these two outliners collectively cause the overall variance to be relatively low at *.85;* removing these outliners the overall variance increase to *.95*, thus moving the variance closer to 1. ***Figure 1*** shows the linear regression without the outliners.

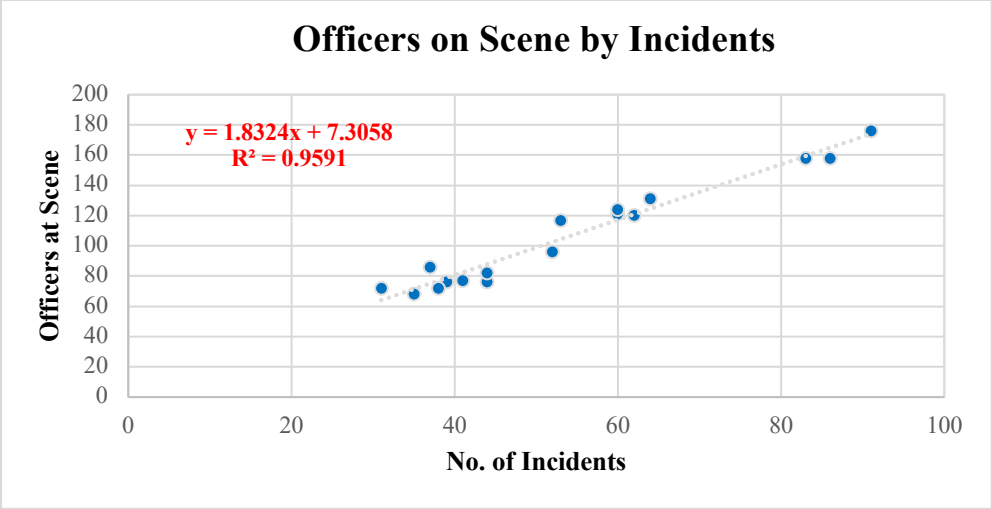

Figure 5: *Linear Regression Chart without Outliners*

Create a residual plot and explain how to improve the linear regression model based on your interpretation of the plot.

In order to improve the linear regression model; the police department would need to investigate the findings for the outliners, specifically **district H.** We should also note that the officers on scene increases as incidents increase respectively. ***Figure 1*** shows the linear regression with the outliners and ***Figure 2*** shows the linear regression without the outliners.

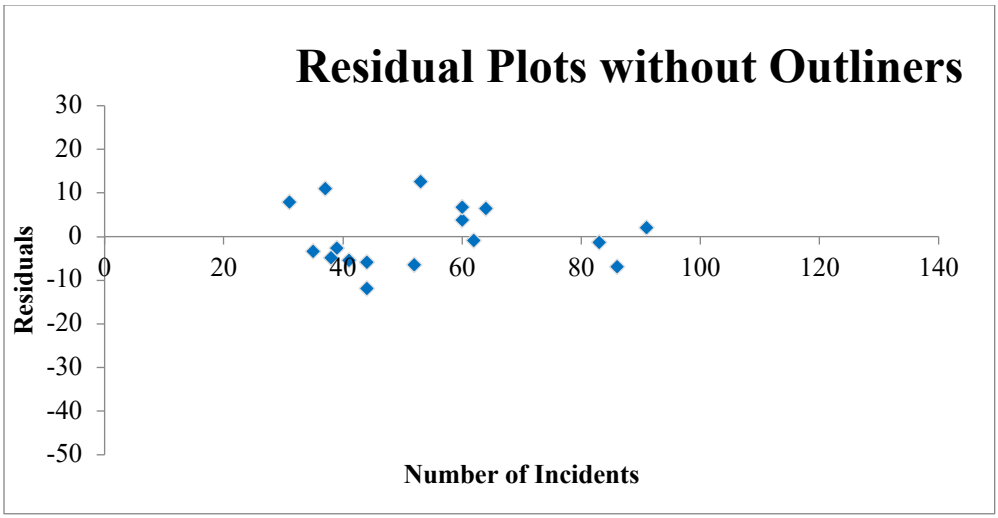

Figure 6: *Residual Plots without Outliners*

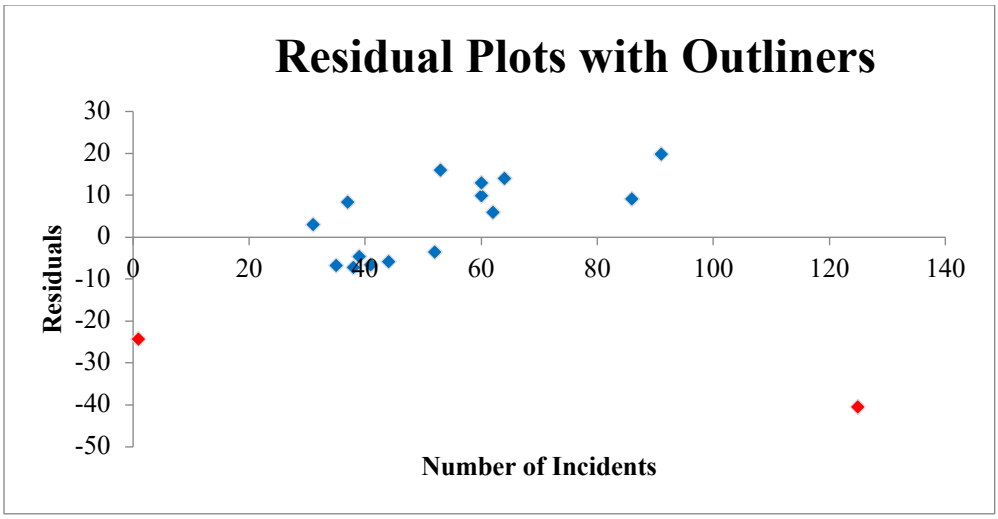

Figure 7: *Residual Plots with Outliners*

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.

The average officer on scene is currently *1.8*; however, with the outliners the average officer seems to decrease to *1.49*, moving further away from the *2.5* minimum requirement; so the department does not currently qualify for additional funding; however, the department needs to investigate the reported incidents that reported zero officers on scene to see if this could move them towards the minimum requirement of *2.5* officers on scene per incident. In addition to investigating the incidents that reported zero officers on scene; the department may want to increase the length of the study and look at more data over a longer time period.

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

The safeguard of data is everyone's responsibility; the data within this scenario contains info regarding crimes, their locations, and the police departments response to each one. Just because someone wants to know doesn't necessarily mean they need to know. If this info were to fall into the wrong hands; criminals could then target the area where there are little to no

police response; this could ultimately lead to an increase in crime. In order to maintain data and analytical integrity, the analyst should make sure that the data is secured in one spot and only those people who are involved with the research have access to it. Info should not be shared in any open public forums or over unsecured networks or emails.

References

AKM1 TASK 1: FUNDAMENTAL ANALYTICS. (2020, August 1). Retrieved from Western Governors University: https://tasks.wgu.edu/student/000194226/course/13800005/task/424/overview

Fundamentals/Statistics for Data Analytics. (n.d.). Retrieved August 2020, from Learn Zybooks: https://learn.zybooks.com/zybook/WGUC740V52018

*Optional Footnotes

[1][Add footnotes, if any, on their own page following references. For APA formatting requirements, it's easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. *(Note: If you delete this sample footnote, don't forget to delete its in-text reference as well. That's at the end of the sample Heading 2 paragraph on the first page of body content in this template.)*]

*Optional Tables

Table 1

[Table Title]

Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

*Optional Figures

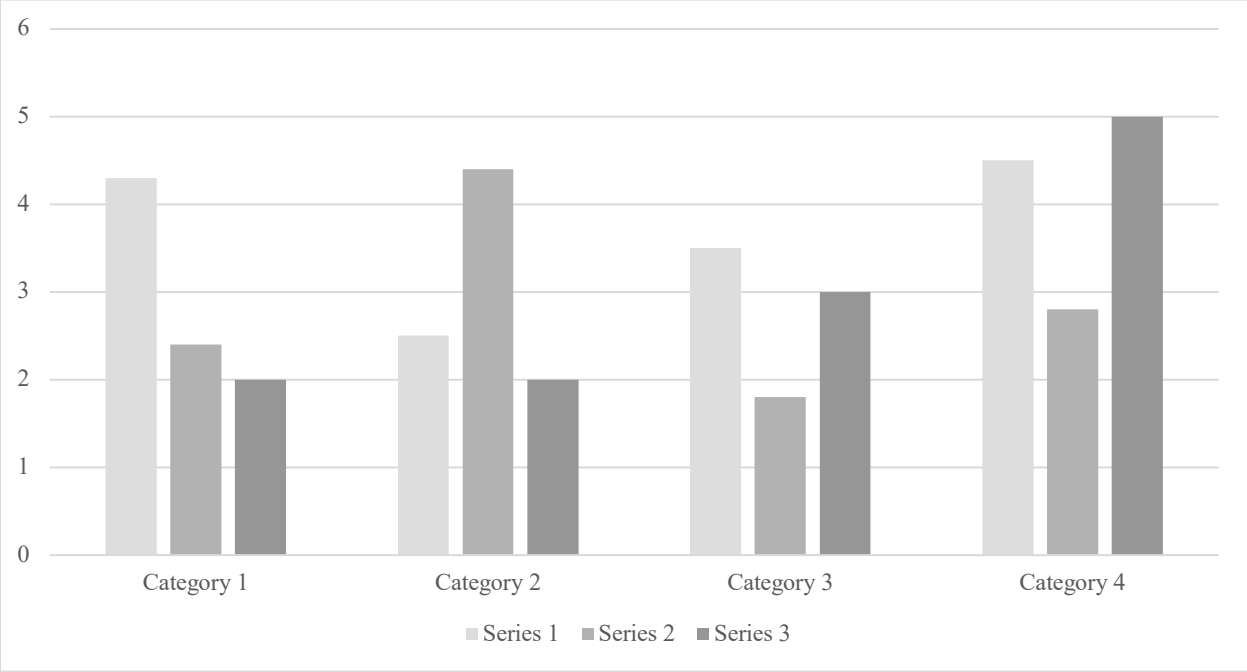

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the *APA Style Manual, 6th Edition*.