Fundamental of Data Analytics- C740 Task 1

by

Serena Zatara

The local police department of Raccoon City has recruited my analytic skill set for consulting purposes as their data analyst. The RPD is interested in discovering the behavior, trends, and needs of their department and from observations what my recommendation would be for increasing their funding. The Chief of Police has informed me that to meet the county's funding requirement; they must have an average of 2.5 officers at their incident scene.

For my consulting, I will take the requests from the chief of police and present my analysis in two separate parts.

Part 1:

Analyze the logs from 911 calls from within Raccoon City and summarize the data. I will prepare and clean the raw data from and records that are outliers of not necessary. Once the data has been cleaned, I will then create the following tables and bar graphs:

- table: date and number of events
- bar graph: date and number of events
- table: number of incident occurrences by event type
- bar graph: number of incident occurrences by event type
- table: sectors and number of events
- bar graph: sectors and number of events

For reporting purposes, I will only be including my observations found in my analysis and present the bar graphs to support my findings with visualizations.

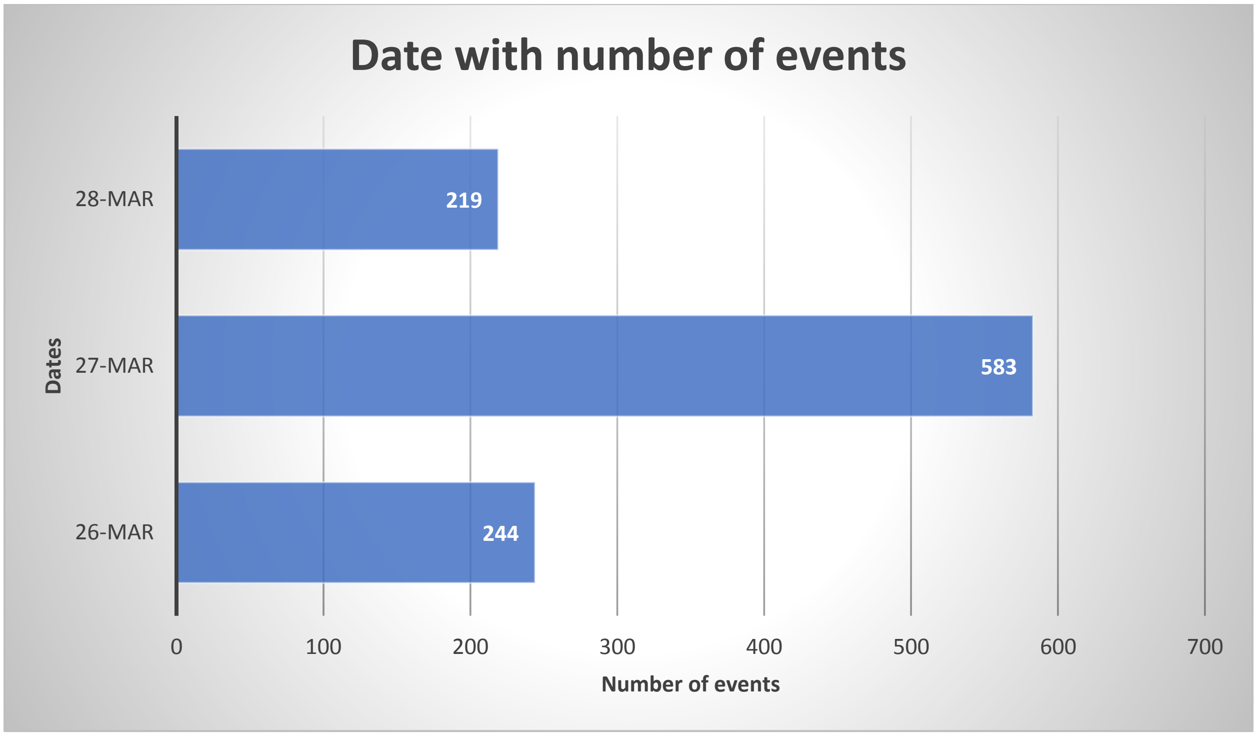
Part 2:

As part of my requested consulting duties, my finding will help the chief of police with my recommendations to see if the RPD meets the minimum of 2.5 officers onsite per incident to receive additional funding from the state governor. Included in this analysis will be an existing table that was constructed. Also how I will take this existing information and refine this data with a new Linear Regression Model, "*Simple linear regression is a way to model the linear relationship between two quantitative variables, using a line drawn through those variables' data points, known as a regression line is common use of a regression line to make predictions*"(zyBooks).

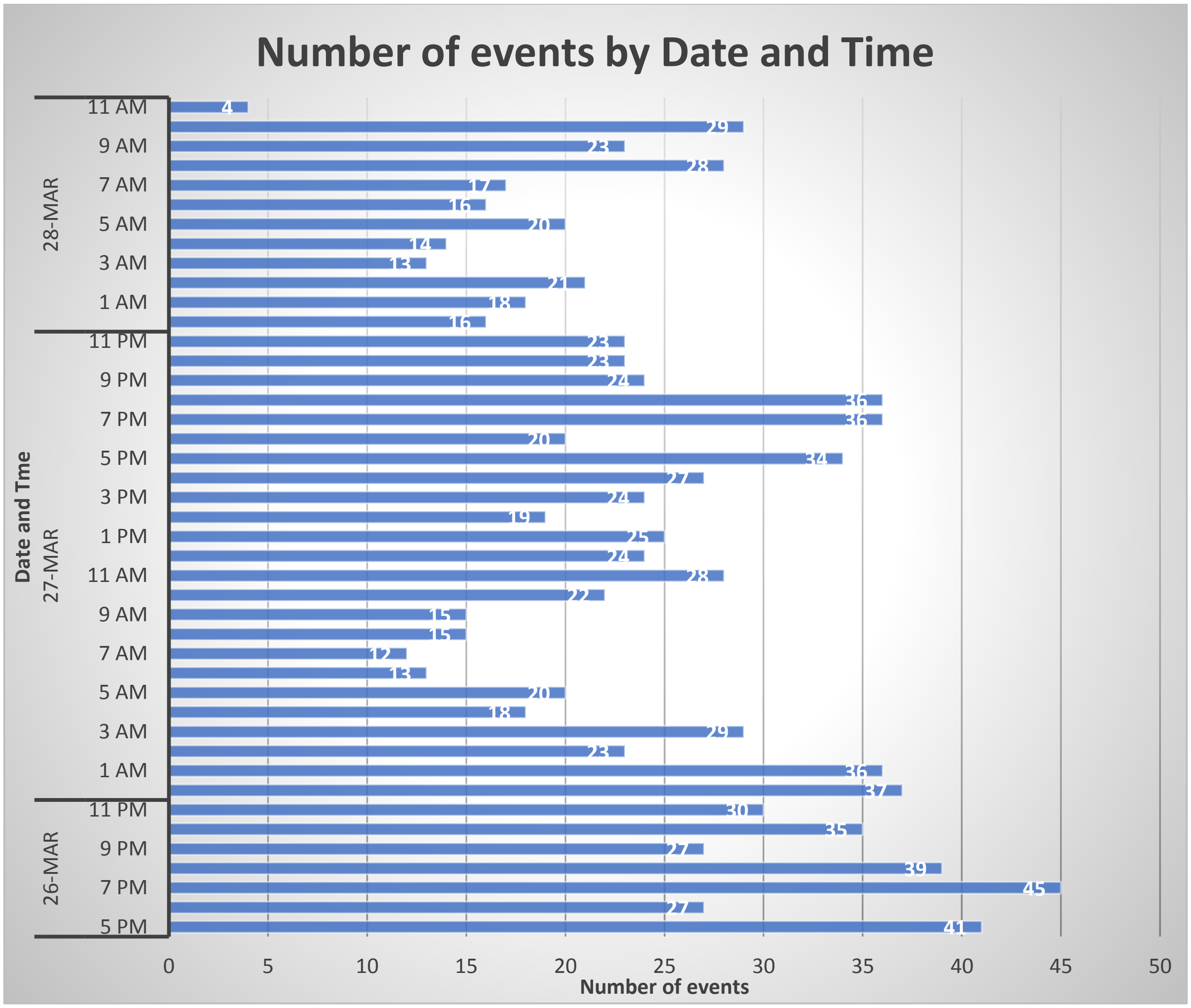
Part 1

For the beginning of my analysis, I received the raw data form Chief Irons provided me from the existing 911 call from March 26th through March 28th. From this data, I found that some of the data was not necessary for my analysis. Thus I imputed the following data columns Longitude, Latitude, Incident location, Census tract, CAD CDW ID, Hundred block locator, Event Clearance Code, and Initial Date due to lack of data in this column. I then set the excel sheet to search and locate in event number to identify and duplicate data. This did not yield in and duplicate results. Upon my findings, I discovered that event 226 was missing the sector information. For this, I chose to use the zone of the sector and filled the value in with F.

Once the data has been cleaned and prepped to be analyzed, I started with the first goal of identifying the number of events and their date of occurrence. To count the number of events, I created a new column that I would later use throughout my analysis. This column would be labeled as Number of Events, and to fill this column with data, I counted the Event Clearance Date. The result was a total of 1046 events that occurred between March 26th through March 28th. Upon this finding, I could now identify how many events occurred each day. My findings identified March 26th had 244 events of incidents; March 27th resulted in 583, and March 28th had 219. Clearly making March 27th the top date of incidents.

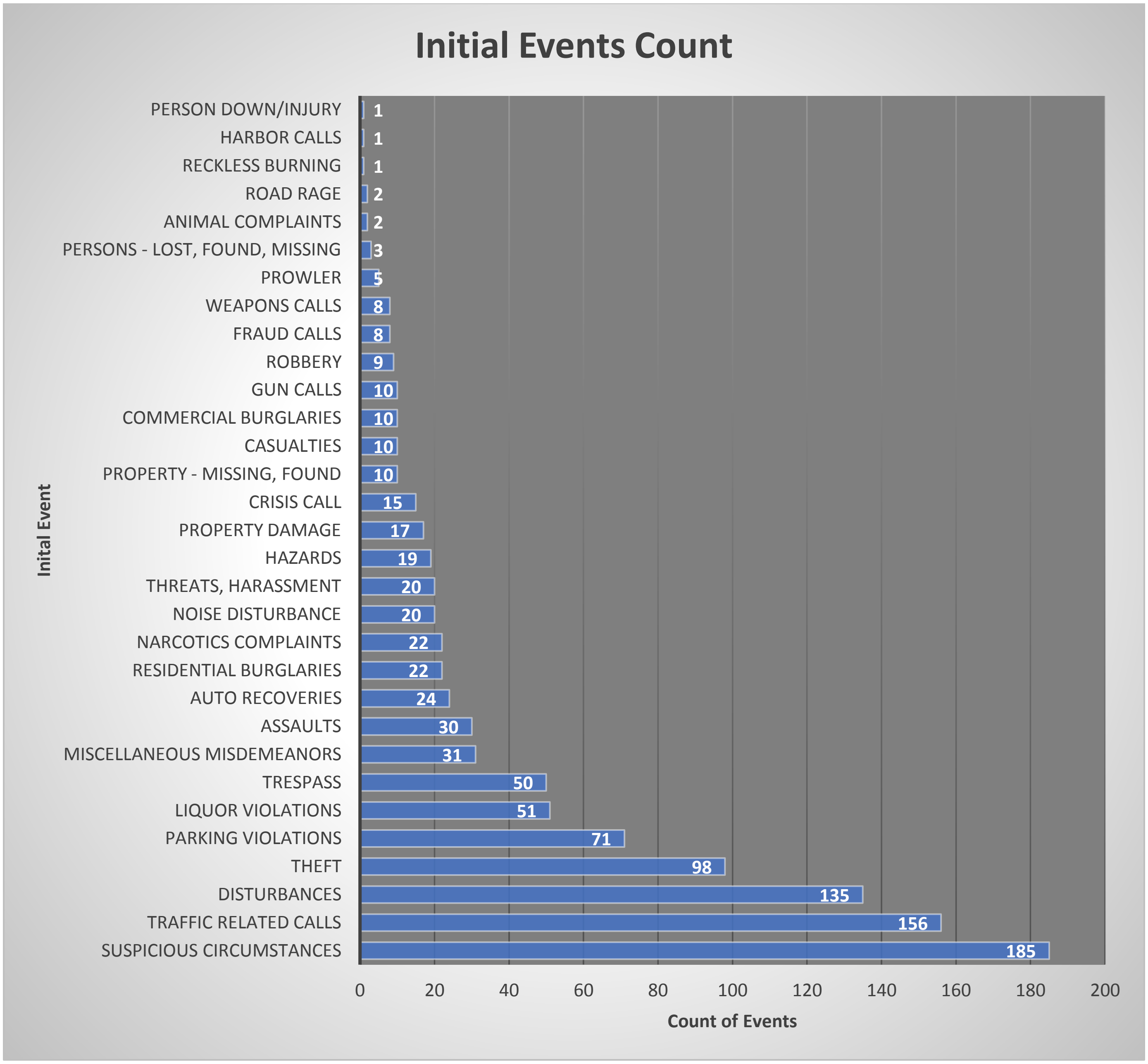


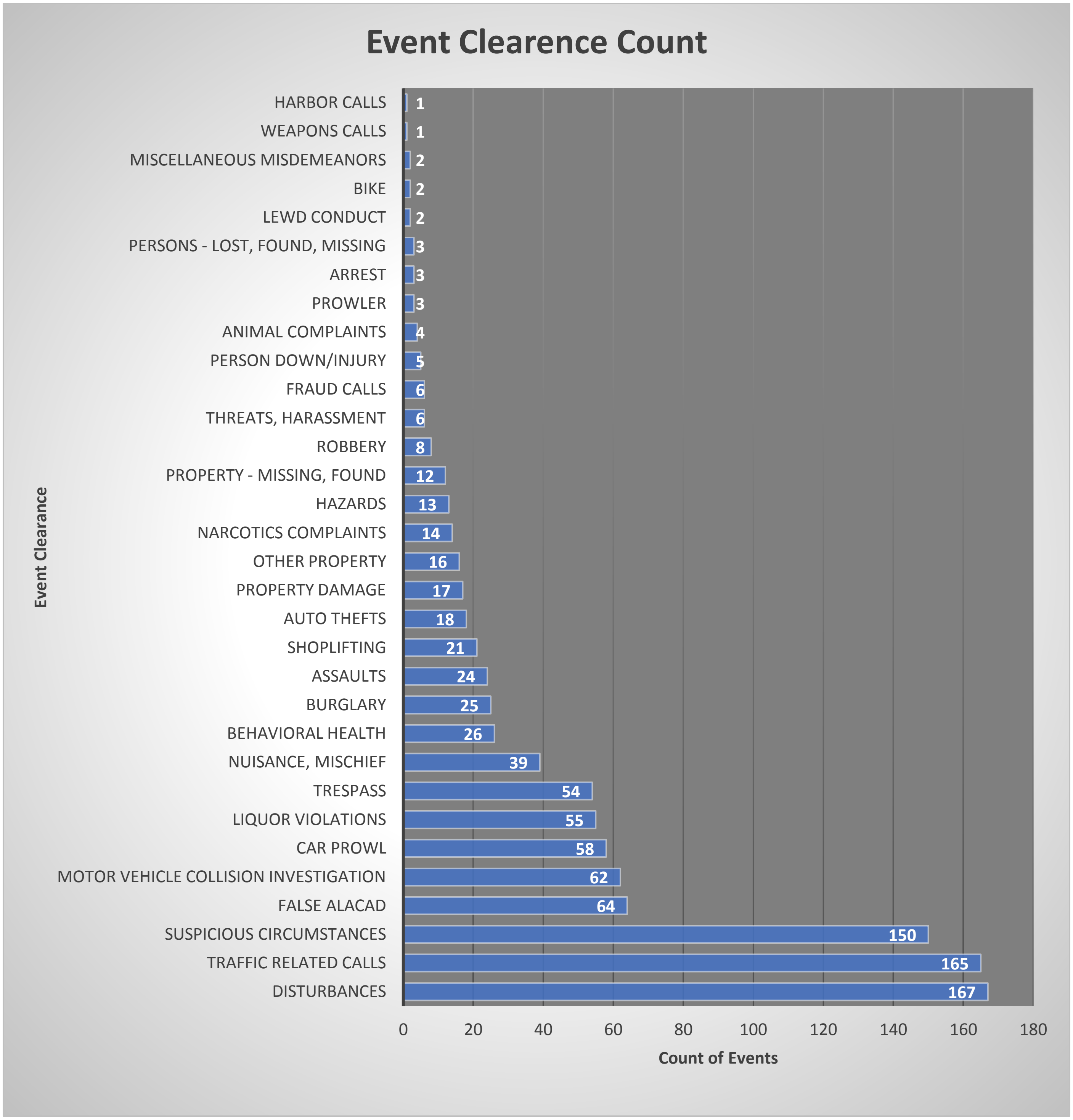
My analysis of the dates took another step as I expanded my data to include the times of day to each day.



This new introduction of time of day provides the RPD and in-depth look into times of data and coordinating the appropriate officers during these times. The data show a higher activity of incidents after 5 pm for Raccoon City.

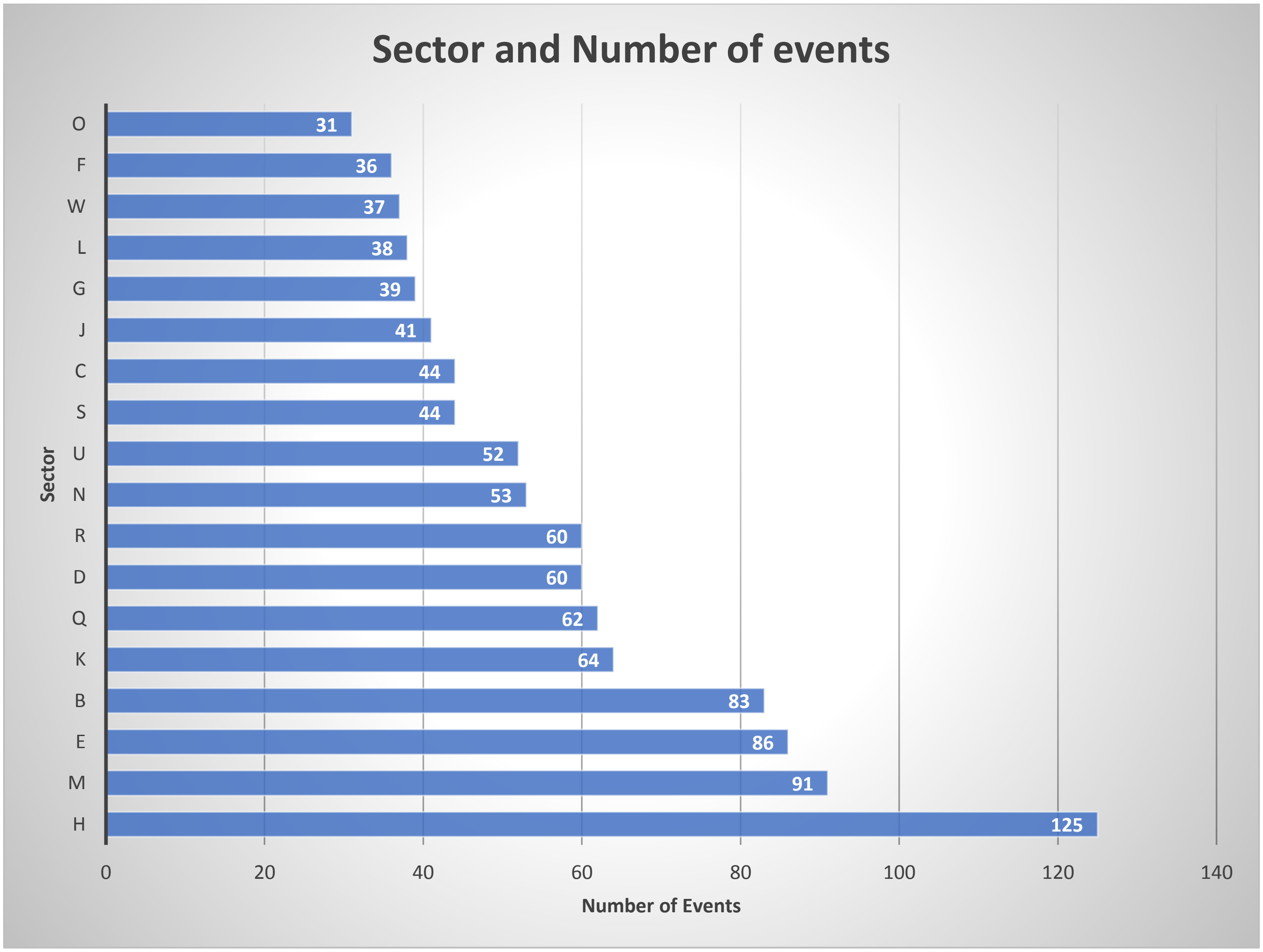
The next step of the report takes a closer analysis of the incidents that occurred and the type category of the incident. I chose to do this in two separate sections by looking at the initial events that the RPD receives and then the actual events cleared in the officers' reports. Utilizing the number of events discovered in the prior bar-graphs, I then combined them to the initial events and events cleared data.


Initial Events Count
Inital Event
PERSON DOWN/INJURY 1
HARBOR CALLS 1
RECKLESS BURNING 1
ROAD RAGE 2
ANIMAL COMPLAINTS 2
PERSONS - LOST, FOUND, MISSING 3
PROWLER 5
WEAPONS CALLS 8
FRAUD CALLS 8
ROBBERY 9
GUN CALLS 10
COMMERCIAL BURGLARIES 10
CASUALTIES 10
PROPERTY - MISSING, FOUND 10
CRISIS CALL 15
PROPERTY DAMAGE 17
HAZARDS 19
THREATS, HARASSMENT 20
NOISE DISTURBANCE 20
NARCOTICS COMPLAINTS 22
RESIDENTIAL BURGLARIES 22
AUTO RECOVERIES 24
ASSAULTS 30
MISCELLANEOUS MISDEMEANORS 31
TRESPASS 50
LIQUOR VIOLATIONS 51
PARKING VIOLATIONS 71
THEFT 98
DISTURBANCES 135
TRAFFIC RELATED CALLS 156
SUSPICIOUS CIRCUMSTANCES 185
0
20
40
60
80
100
120
140
160
180
200
Count of Events



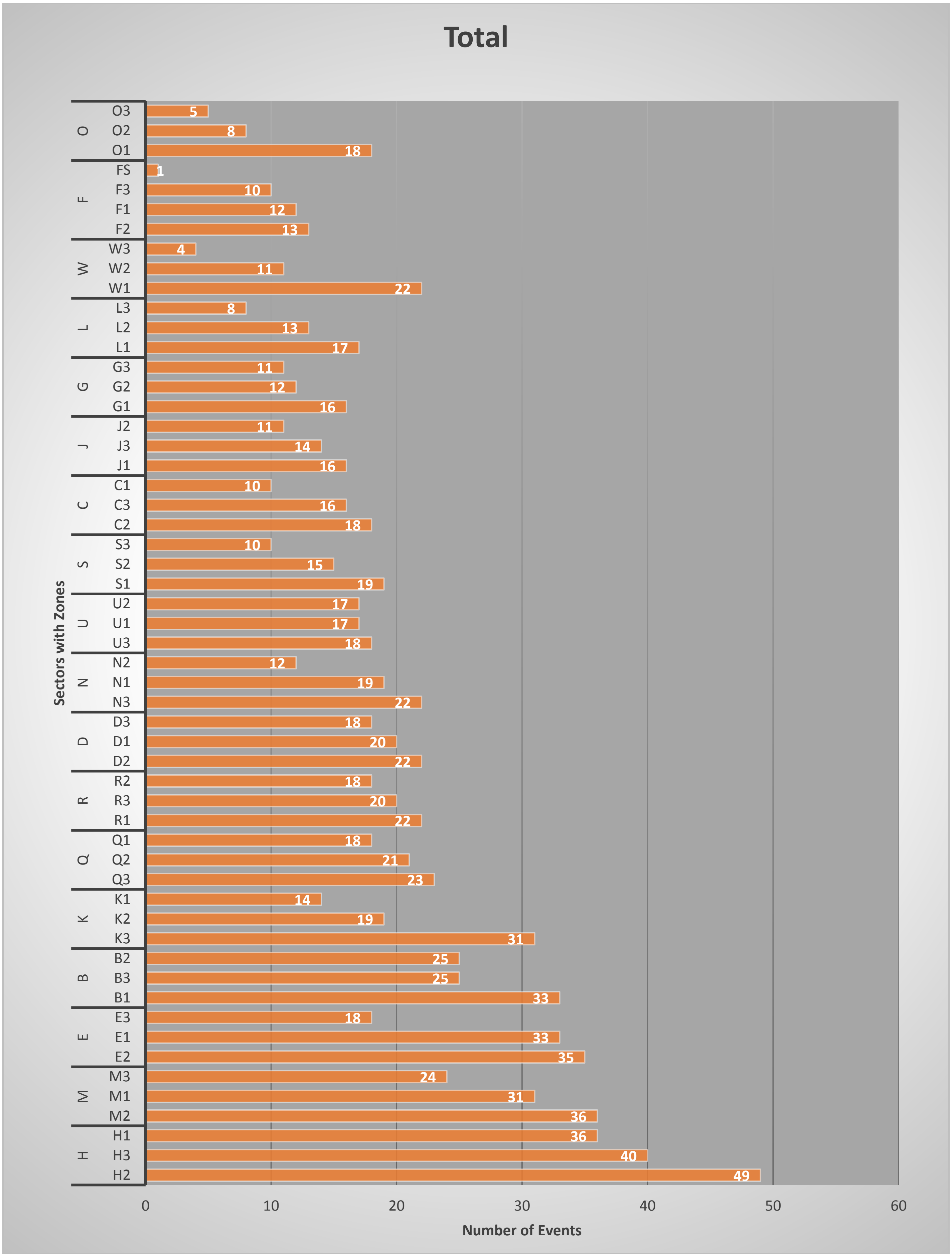
The findings showed for both the top three incidents of events were Suspicious circumstances Disturbances, Traffic-related calls for the initial calls the RPD received. After officers were sent to the scene and filed their reports, the data showed Disturbances, Traffic-related calls, Suspicious circumstances to be the highest of incidents. Along with the graphs provided, you can visually see initial calls aren’t always classified correctly until officers are sent and file their reports.

The final step in my analysis of 911 call chief Irons provided is to analyze the number of incidents and from which sector of the city they occur.



With the visual above, the data support Sector Has the highest sector with most events that occur at 125 incidents. Followed by Sector M (91 incidents), E (86 incidents), and B (83 incidents). To identify a closer analysis for the RPD to map the city and station their officers. My next step was including the zones for each sector.

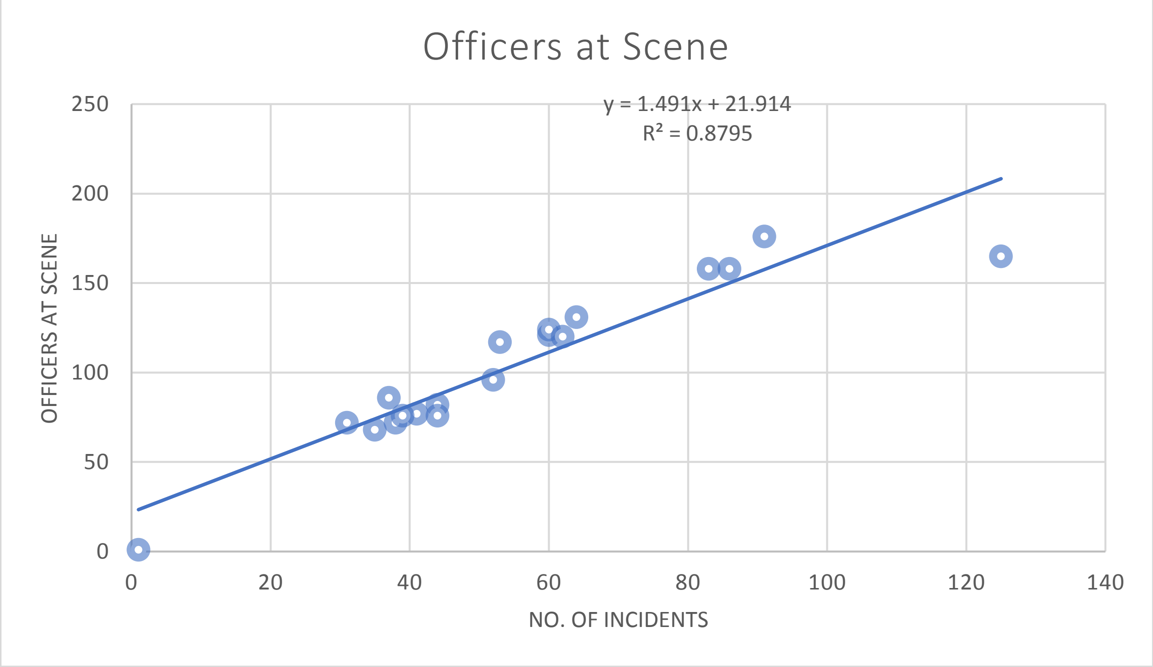
The graph below displays each zone within the sectors. This shows Sector H zone H2 requires higher priority in this sector; M2, E2, and B1 respectively display the highest results zone with incidents. The zone within each sector carefully shows slight decreases for incidents, but with this data visualization, the RPD can now effectively patrol these sectors.

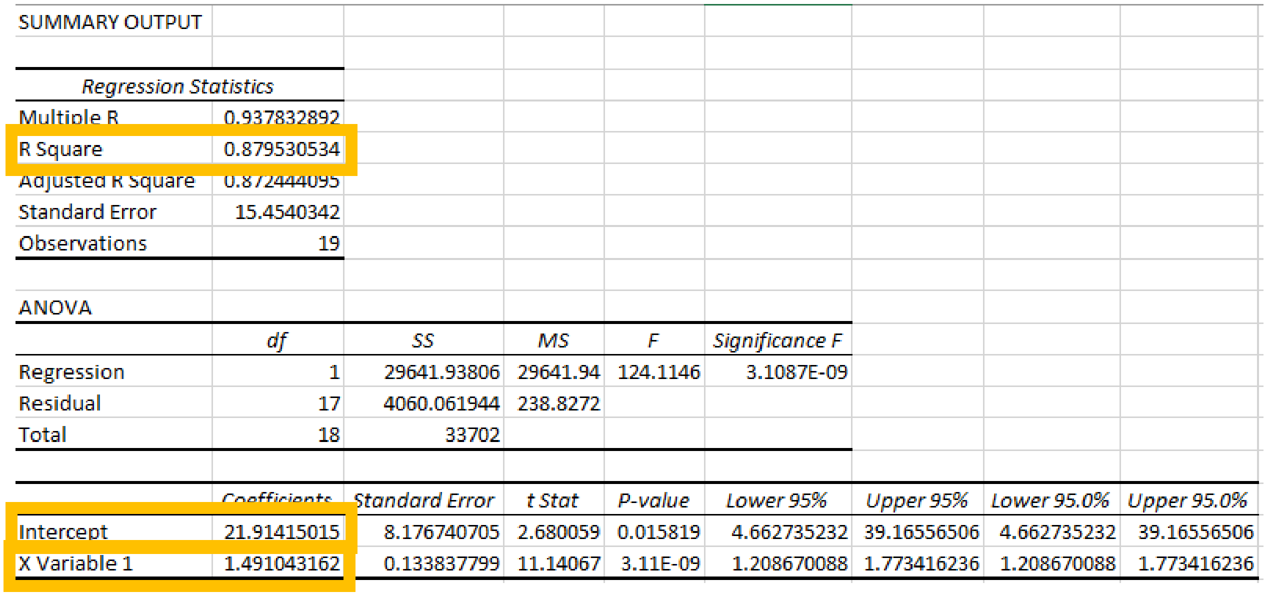

Total
Sectors with Zones
O
O3 5
O2 8
O1 18
F
FS 1
F3 10
F1 12
F2 13
W
W3 4
W2 11
W1 22
L
L3 8
L2 13
L1 17
G
G3 11
G2 12
G1 16
J
J2 11
J3 14
J1 16
C
C1 10
C3 16
C2 18
S
S3 10
S2 15
S1 19
U
U2 17
U1 17
U3 18
N
N2 12
N1 19
N3 22
D
D3 18
D1 20
D2 22
R
R2 18
R3 20
R1 22
Q
Q1 18
Q2 21
Q3 23
K
K1 14
K2 19
K3 31
B
B2 25
B3 25
B1 33
E
E3 18
E1 33
E2 35
M
M3 24
M1 31
M2 36
H
H1 36
H3 40
H2 49
0
10
20
30
40
50
60
Number of Events

Part 2

For the 2nd part of my consulting analysis report, I will help Chief Irons identify if he meets the minimum requirements set by the state's governor, which would be a minimum of 2.5 officers onsite per incident. The start of my analysis was using the already existing data:

District Sector	No. of Incidents	Officers at Scene
B	83	158
H	125	165
W	37	86
K	64	131
D	60	121
O	31	72
U	52	96
R	60	124
S	44	82
	1	1
J	41	77
Q	62	120
L	38	72
C	44	76
M	91	176
N	53	117
F	35	68
G	39	76
E	86	158





SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.937832892
R Square	0.879530534
Adjusted R Square	0.872444095
Standard Error	15.4540342
Observations	19

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29641.93806	29641.94	124.1146	3.1087E-09
Residual	17	4060.061944	238.8272		
Total	18	33702			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	21.91415015	8.176740705	2.680059	0.015819	4.662735232	39.16556506	4.662735232	39.16556506
X Variable 1	1.491043162	0.133837799	11.14067	3.11E-09	1.208670088	1.773416236	1.208670088	1.773416236

For the assessment of the data, there will be two Hypotheses, "*A proposed explanation of a phenomenon, usually as a starting point for further analysis*" (zyBooks). We will have a Null hypothesis and an alternative hypothesis that the conclusion will identify as the outcome from the analysis.

> Null Hypothesis- "Is a statement about a population parameter such as the mean. Null means "amounting to nothing," which typically involves values for the population to indicate no change or no effect." (zyBooks)
>
> Alternative Hypothesis- "Is an alternative statement about a population parameter such as the mean. The alternative is often the research hypothesis of interest in the analysis."

The Null Hypothesis that will be used in this report will be, enough officers arrive to each incident. Alternative Hypothesis, not enough officers arrive at an incident.

Now that the two hypotheses have been stated, let's continue the analysis report and revisit the data provided. Referring to the table above, it displays a simple linear regression of officers at the scene. When looking at The Regression Summary output the current data with outliers include had an intercept of 21.914, X variable of 1.94 and R^2 of 0.879 which is written out as follows:

$$y = 1.491x + 21.914$$

$$R^2 = 0.8795$$

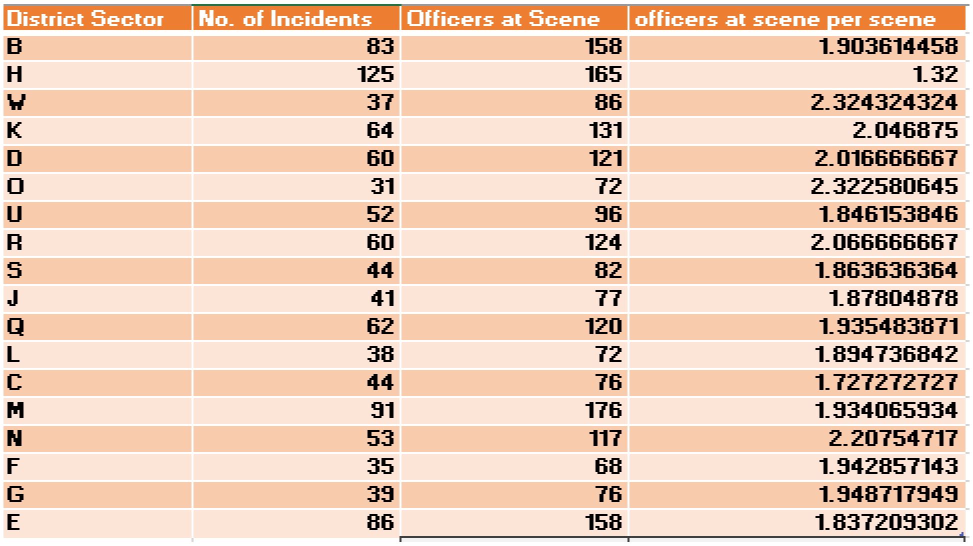
Taking R^2 of 0.8795 giving you 0.7735, then multiplying this number by 100 results in 77.35%. This gives us a representation of the "goodness of fit" or in other words the correlation represented between 0.00 and 1.0, which can also be called the coefficient of determination. The closer this number is to 1.00 or 100% then it will be a perfect fit.

Looking at the plotted graphs I've identified some outliers at the Y-axis, which would be the number of officers being plotted by 50 and so on to 250. This will misrepresent the actual number of officers at each incident to support if RPD reached the 2.5 officer minimum. As the number of incidents increased, the more officers visually responded, which would lead one to accept the Null Hypothesis that enough officers are at each scene, which would be 2.5 or higher. This type of representation of the data is manipulating the visualization and the viewer to believe as the number of incidents increases, so does the number of officers respond to each scene.

However, in addition to this outlier, I will point out and explain another and how removing each of these outliers and data manipulation will provide a more accurate representation of the hypothesis. The fit of the linear regression in this first model shows a simple linear regression assuming that as the number of incidents (X) increases, the variability of officers at the scene (Y) plots the values that remain constant around the regression line. The outlier with the visual presented is how the lowest point represents an unknown sector with one incident and one officer and then jumps to the second plot at 68 officers for 35 incidents. It would be best to remove this strange incident representation of 1 from the data and see how by removing this outlier, changes the regression summary below and the coefficient of determination. Now with the data representing a more significant number of officers, this would be more suitable in a bar graph to depict a visualization of the accumulative amount of officers responding.

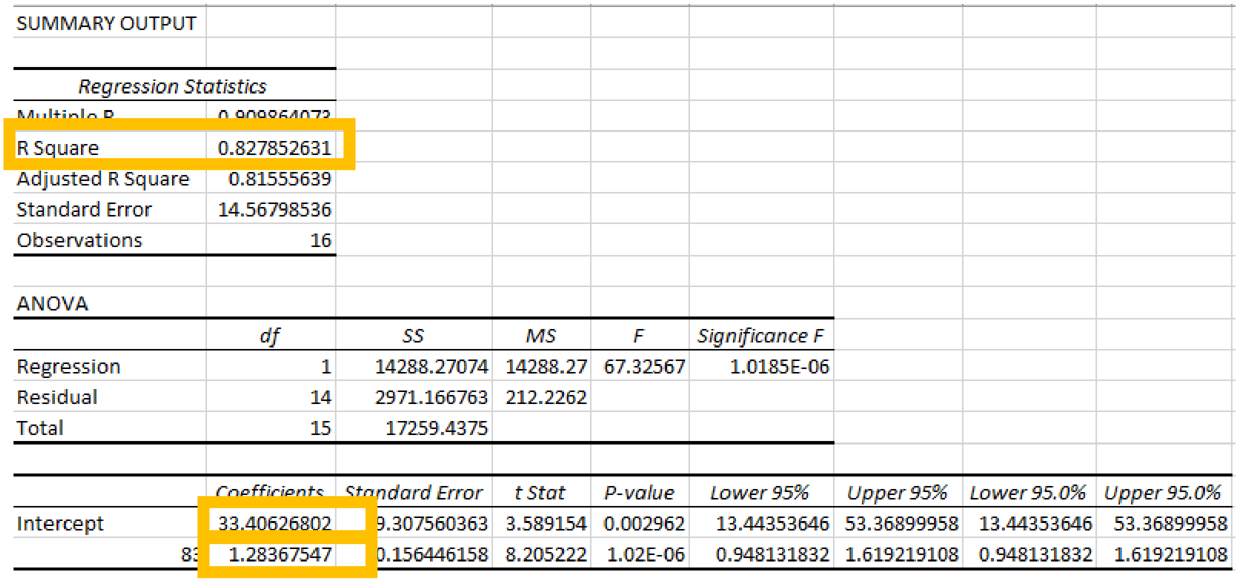
Since this is not the objective of showing the 2.5 officers responding to incidents in hopes of receiving additional funding from the governor, to improve upon this data and graph, I took the officers at a scene divided by the number of incidents.

$$\text{Officers at scene per scene} = \frac{\text{Officers at scene}}{Number\ of\ incidents}$$



District Sector	No. of Incidents	Officers at Scene	officers at scene per scene
B	83	158	1.903614458
H	125	165	1.32
W	37	86	2.324324324
K	64	131	2.046875
D	60	121	2.016666667
O	31	72	2.322580645
U	52	96	1.846153846
R	60	124	2.066666667
S	44	82	1.863636364
J	41	77	1.87804878
Q	62	120	1.935483871
L	38	72	1.894736842
C	44	76	1.727272727
M	91	176	1.934065934
N	53	117	2.20754717
F	35	68	1.942857143
G	39	76	1.948717949
E	86	158	1.837209302

As visualized above, it showcases the number of officers per each scene, giving an average of 1.94 officers per scene. This does not meet the 2.5 minimum and to further support that finding The Regression summary of this data provided the following support:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909864073
R Square	0.827852631
Adjusted R Square	0.81555639
Standard Error	14.56798536
Observations	16

ANOVA

	df	SS	MS	F	Significance F
Regression	1	14288.27074	14288.27	67.32567	1.0185E-06
Residual	14	2971.166763	212.2262		
Total	15	17259.4375			

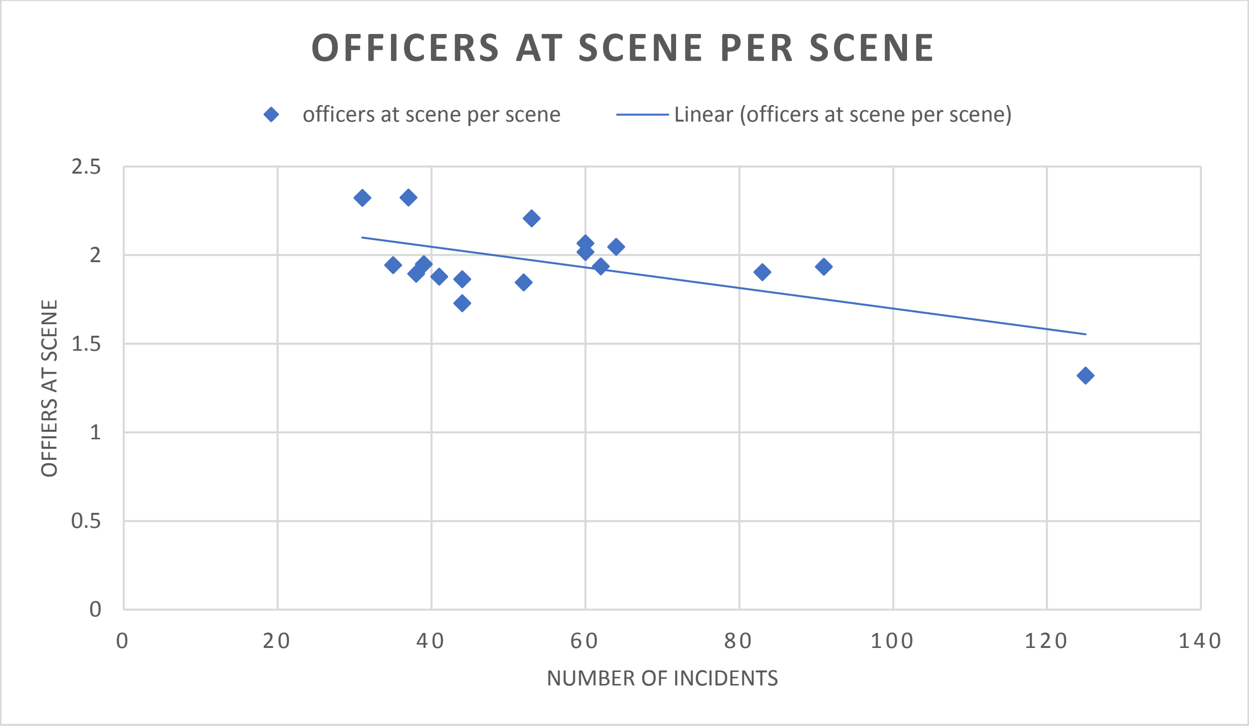
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	33.40626802	9.307560363	3.589154	0.002962	13.44353646	53.36899958	13.44353646	53.36899958
83	1.28367547	0.156446158	8.205222	1.02E-06	0.948131832	1.619219108	0.948131832	1.619219108

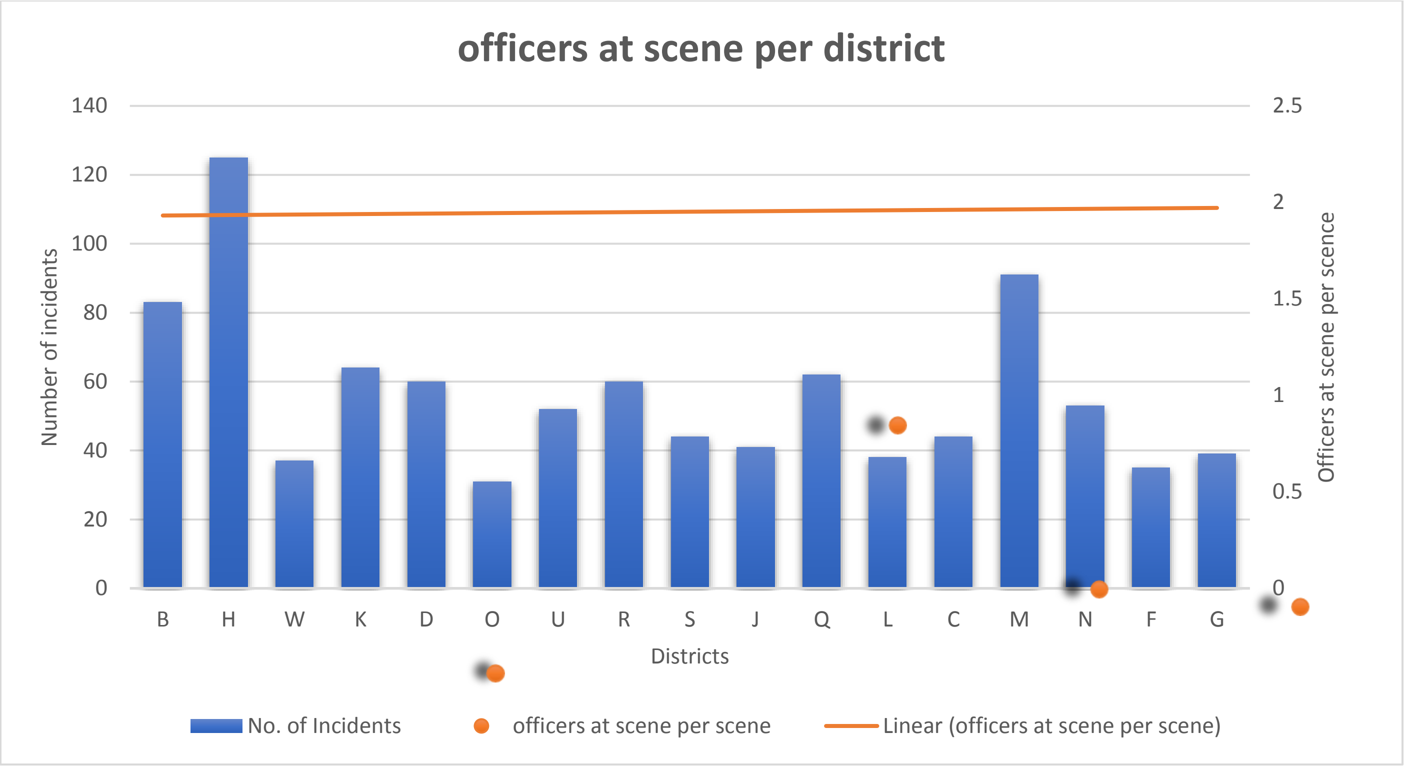
Looking closer at the data provided to identify a specific piece of data that the report will rely on P-Value, “If less than threshold usually 0.01 or 0.05, then one rejects the null hypothesis in favor of the alternative hypothesis” (zyBooks), here the P-Value is 0.0029. With the 1 value outlier taken out the intercept is 33.406, X variable of 1.283 and R^2 of 0.8278, which is written out as follows:

$$y = 1.283x + 33.406$$

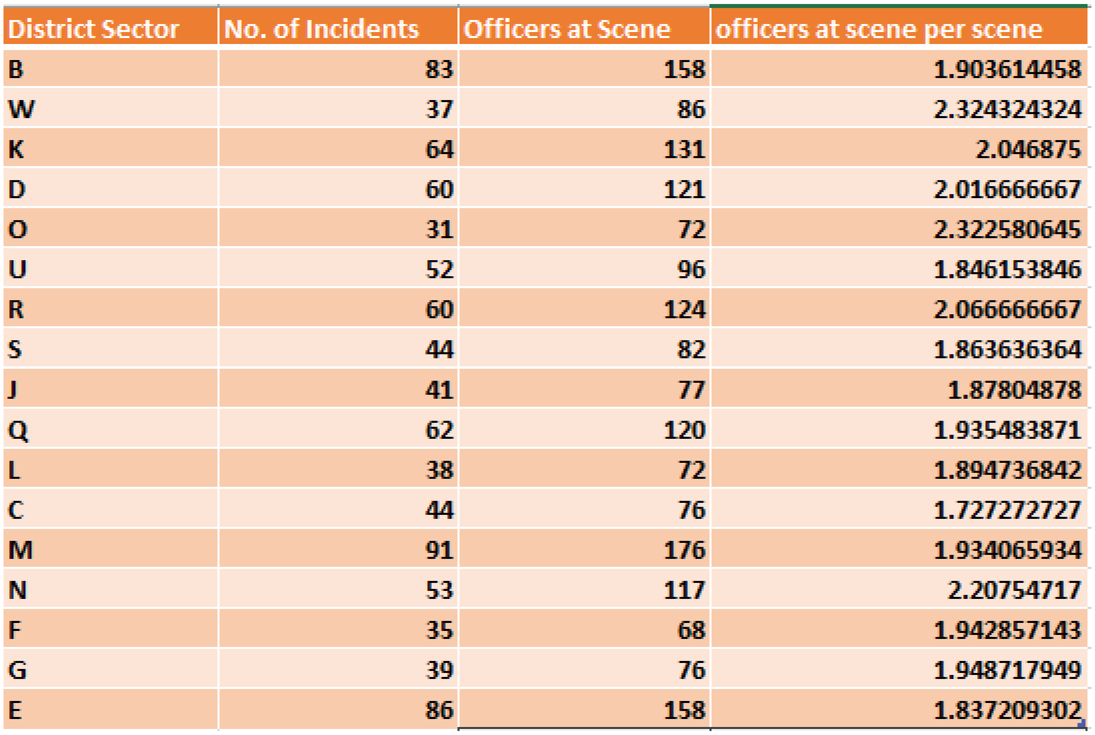
$$R^2 = 0.8278$$

Taking the R^2 of 0.8278 gives you 0.6852, then multiplying this number by 100 results in 68.52%. This is a drop in the coefficient of determination then prior to taking the outlier of 1 out. Let’s graph the data to show visualization support.

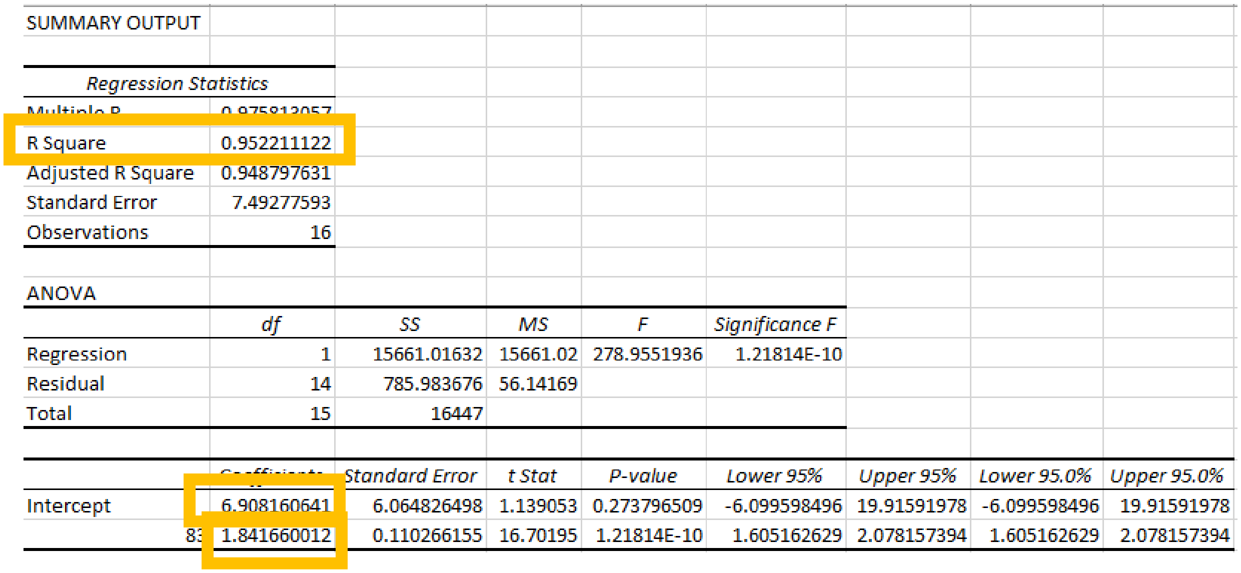




Seeing the points plotted with the number of incidents almost concludes my findings. However, looking, I identified another outlier point found with sector H (125,165) in the first graph, but in the second graph, the line shows a non-zero slop regression when plotting three elements. Seeing that this point now shows as an outlier, I think its best to remove this data/point.



District Sector	No. of Incidents	Officers at Scene	officers at scene per scene
B	83	158	1.903614458
W	37	86	2.324324324
K	64	131	2.046875
D	60	121	2.016666667
O	31	72	2.322580645
U	52	96	1.846153846
R	60	124	2.066666667
S	44	82	1.863636364
J	41	77	1.87804878
Q	62	120	1.935483871
L	38	72	1.894736842
C	44	76	1.727272727
M	91	176	1.934065934
N	53	117	2.20754717
F	35	68	1.942857143
G	39	76	1.948717949
E	86	158	1.837209302



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.975813057
R Square	0.952211122
Adjusted R Square	0.948797631
Standard Error	7.49277593
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	15661.01632	15661.02	278.9551936	1.21814E-10
Residual	14	785.983676	56.14169		
Total	15	16447			

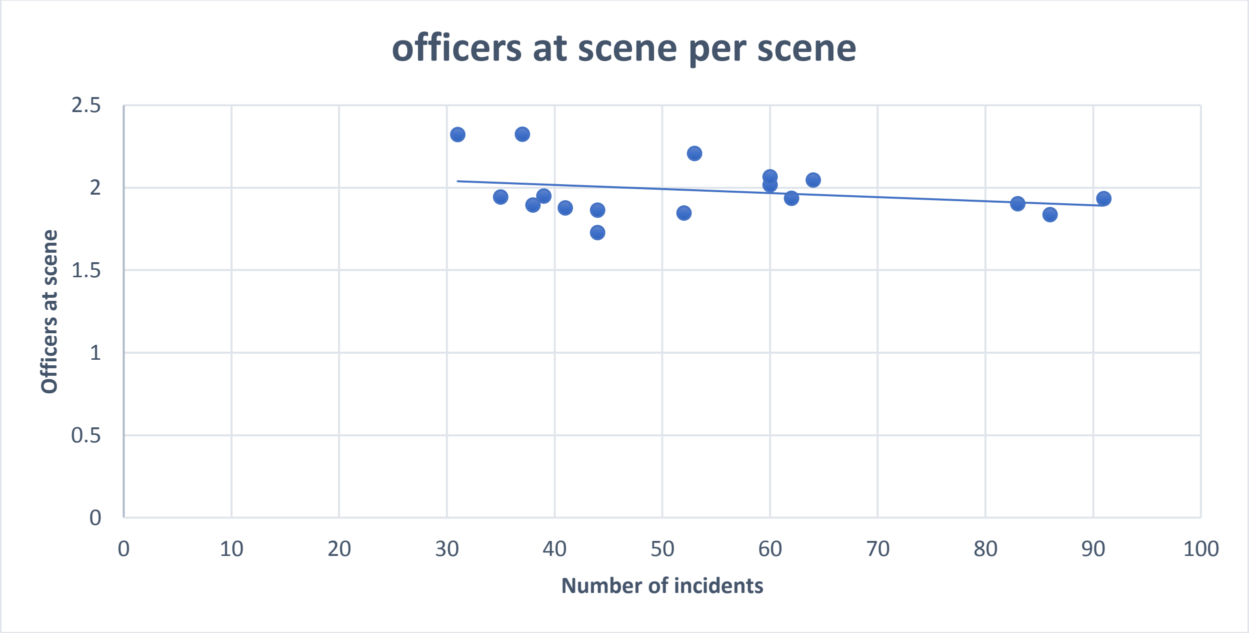
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.908160641	6.064826498	1.139053	0.273796509	-6.099598496	19.91591978	-6.099598496	19.91591978
83	1.841660012	0.110266155	16.70195	1.21814E-10	1.605162629	2.078157394	1.605162629	2.078157394

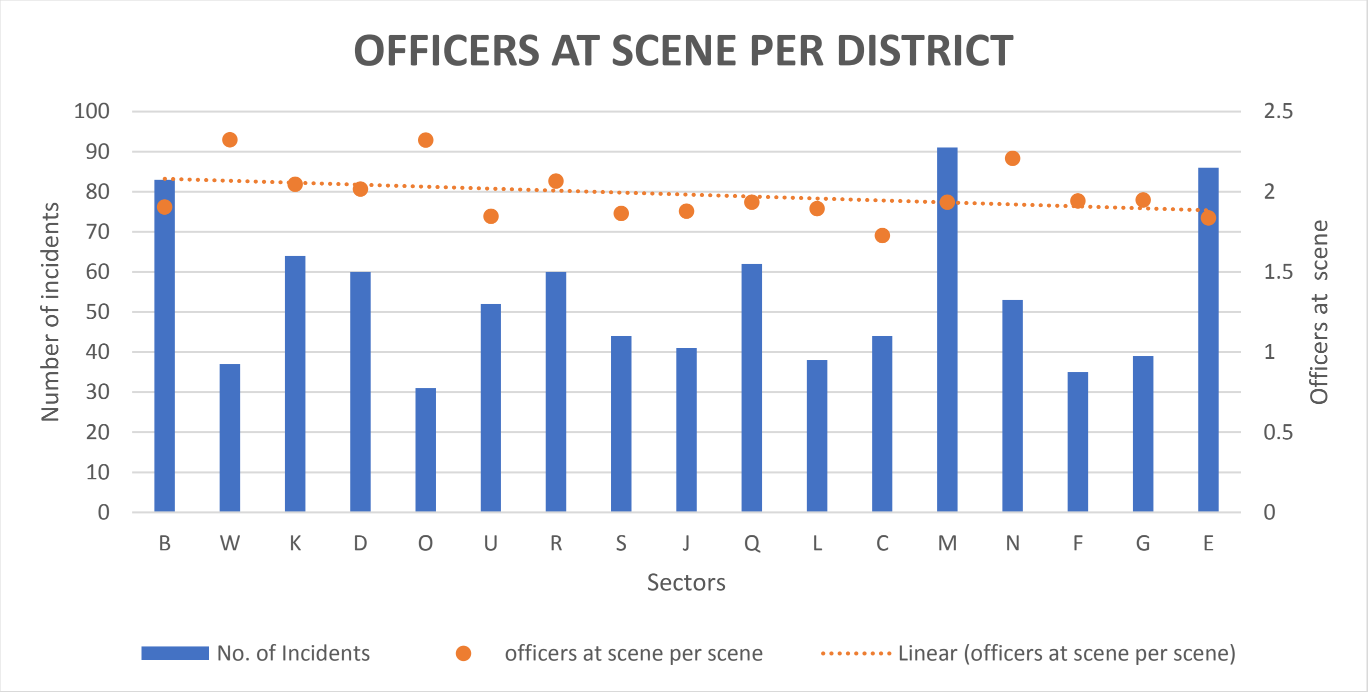
From imputing sector H, the new average of officers per scene is 1.98, increasing slightly; also, the P-value has now increased to 0.2737. With the H point outlier taken out the intercept is 6.908, X variable of 1.841 and R² of 0.9522, which is written out as follows:

$$y = 1.841x + 6.908$$

$$R^2 = 0.9522$$

Taking the R² of 0.9522 gives you 0.9066, then multiplying this number by 100 results in 90.67%. Which is the closest the data has come to 1.0 or 100%. Now that this data provided, it now needs to be showcased visually.





Conclusion

With the new data refined, the number of officers is now able to be visualized and plotted by .5 through 2.5. This provides a more accurate reading of the data per scene. In using all sectors or removing sector H, and the unknown sector with the values of 1, the P-values are either 0.0029 or 0.2737. Which are less than zero, which will result in the rejection of the null hypothesis and conclude with the alternative hypothesis, not enough officers arrive at each incident. The linear regression plotted results both in a non- zero slope. Taking the number of incidents and dividing by officers at the scene results in points showing positive above the line or close to zero on the regression line. In additional support the with the outliers of the strange point of 1,1 and H sector 125,165 taken out the intercept is 6.908, X variable of 1.841 and R^2 of 0.9522:

$$y = 1.841x + 6.908$$

$$R^2 = 0.9522$$

Using R^2 of 0.9522 and gives you 0.9066, then multiplying this number by 100 results in 90.67%. Which is the closest the data has come to 1.0 or 100%, and supports my final conclusion.

Upon my review of the 1046 incidents and isolating the impacted sectors, it is my duty to inform the RPD fails to meet the minimum requirement of additional funding from the state's governor. My recommendation to Chief Irons is to increase current officer efficiency and coordinate with each sector and its zone during high activity time to increase the response rate of officers. Perhaps a special task force to respond in high incident types and relating high activity zones.

The data analyzed is confidential and sensitive. Under the terms of the agreement between myself, chief Irons and Governor Spencer, the results of this report will only be shared with the Governor and Chief

Irons of the RPD. All other data will be appropriately encrypted and stored in a zip file and stored for any further reference or further consulting requests.

Works cited

“Student's t-Test.” *ZyBooks*, learn.zybooks.com/zybook/WGUC740V52018/chapter/4/section/5.

ZyBooks, learn.zybooks.com/zybook/WGUC740V52018/chapter/4/section/2.