Keenin Vye
Fundamentals of Data Analytics — C740
IVP Task 1

Section A

I removed the following columns:

a. Event Clearance Code:
    i. I removed this because it is the same information as Event Clearance Description, but in the form of a number. This information means nothing to everybody except police officers, the text description is much more beneficial.
b. Event Clearance Description
    i. I removed this because it was a more detailed version of the Event Clearance Group. Event Clearance Group is more general and much more intuitive to understand when grouping incidents.
c. Event Clearance Subgroup
    i. I removed this because, again, it was a more detailed version of the Event Clearance Group. Once I chose to use the Event Clearance Group as my incident type, I removed all other superfluous type information. Event Clearance Group holds enough information to accurately interpret the data based on incident type.
d. Zone/Beat
    i. I removed this because it was more detailed information than just the zone of the incident. For the questions asked of us, the 'beat' information is not important, so we can ignore this column.
e. Latitude and Longitude
    i. I removed these two columns because this information is present in the `location` column and there is no reason to have duplicate data.
f. Initial Type Description
    i. I removed this because it was a more detailed version of the Initial Type Group. Initial Type Group is more general and much more intuitive to understand when grouping.
g. Initial Type Subgroup
    i. I removed this because, again, it was a more detailed version of the Initial Type Group. Once I chose to use the Initial Type Group as my incident type, I removed all other superfluous type information. Initial Type Group holds enough information to accurately interpret the data based on incident type.
h. At Scene Time
    i. I removed this because it was not present in all rows of the data.
i. Event Clearance Date/time
    i. I added Event Clearance Date/time because I wanted to group my events by day and not by minute. Grouping the number of events by day tells a much better story thinly a minutely aggregate.

Section B

See supporting worksheet in Excel Doc (B. Events per Day, B. Events per Incident Type, B. Events per Sector).

Section C

The fit of the regression is fairly strong and I would have a large amount of confidence in it. The reason this is, is because of the R-squared value. The R-squared value of 0.87953 tells us that our data strongly fits the line of regression. Along with that, I have created an ANOVA table to view the Standard Error value. With a Standard Error of 6.2117, we have more assurance that the fit of our model is strong.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.934348495
R Square	0.87300711
Adjusted R Square	0.865070055
Standard Error	9.963968094
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10920.00944	10920.00944	109.9913056	1.41051E-08
Residual	16	1588.490563	99.28066018		
Total	17	12508.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-6.811748955	6.21179943	-1.096582243	0.289054975
158	0.597146029	0.05693789	10.48767399	1.41051E-08

Section D

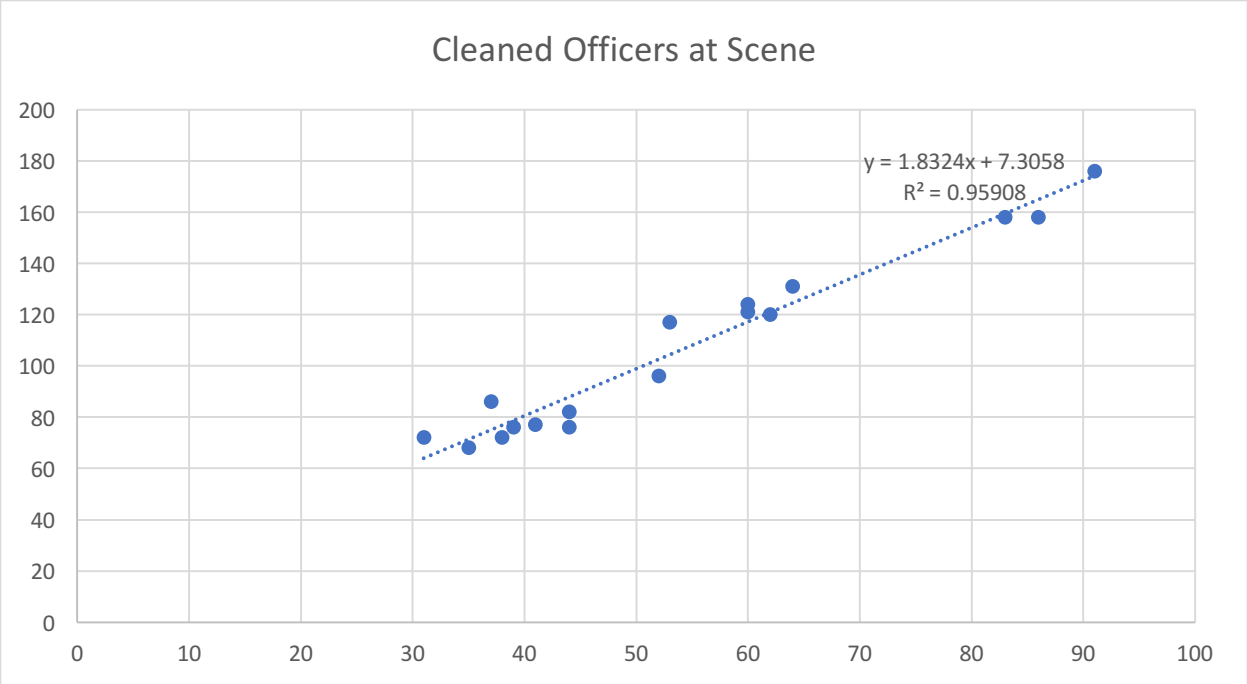
The outliers (Zones H and NULL) in the data, skew the R-squared value to be lower and cause the fit and strength of the model to be decreased. The reason I deemed these as outliers is because:

B. The NULL zone data is missing the Zone it belongs to, so should be removed from the sample set. Also, it had the second largest residual value, making it a clear outlier.
C. The Zone H data was removed because it had the largest residual and that residual was much larger than could be allowed (the determining threshold set was |1.25|).

Once I plotted the new, clean data. I could easily see there was an improvement, right off the bat the R-squared value was 0.959, as opposed to 0.879 in the initial model with the outliers.

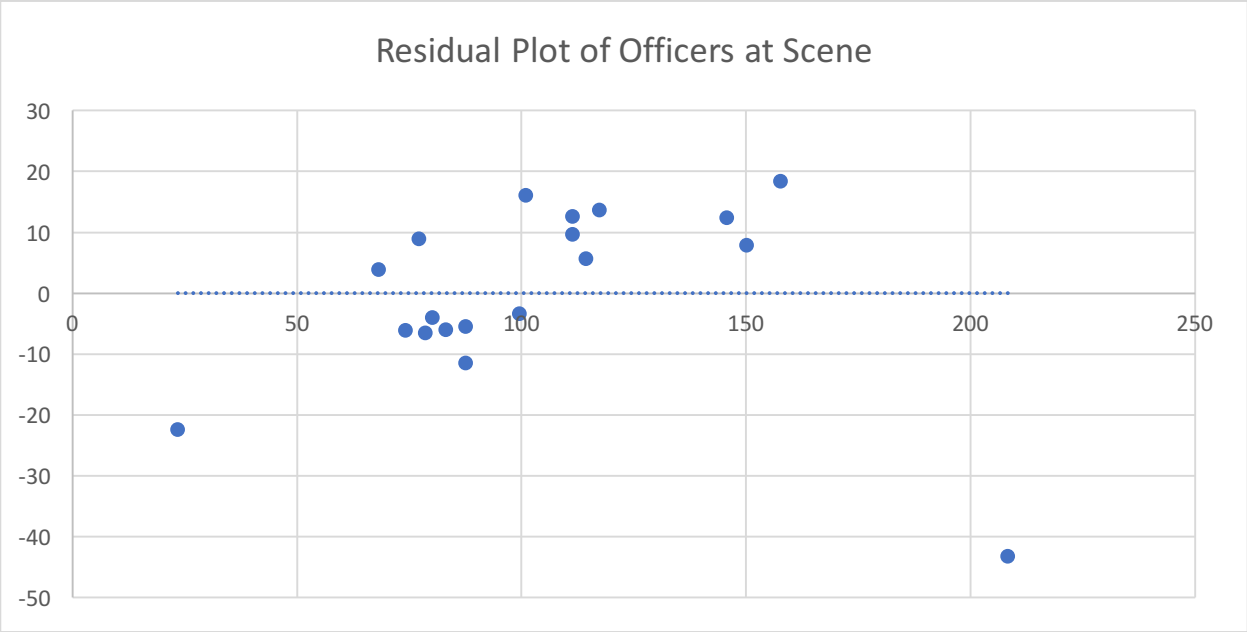
RESIDUAL OUTPUT

Observation	*Predicted Officers at Scene*	*Residuals*	*Standard Residuals*
1	145.6707326	12.32926743	0.82093202
2	208.2945454	-43.29454537	-2.882724282
3	77.08274713	8.917252868	0.593746421
4	117.3409125	13.6590875	0.909476768
5	111.3767399	9.623260147	0.640755214
6	68.13648816	3.863511839	0.257248097
7	99.44839456	-3.448394559	-0.229607925
8	111.3767399	12.62326015	0.840507233
9	87.52004926	-5.520049265	-0.367546995
10	83.04691978	-6.046919779	-0.402628145
11	23.40519331	-22.40519331	-1.491827532
12	114.3588262	5.641173823	0.375611953
13	78.57379029	-6.573790294	-0.437709294
14	87.52004926	-11.52004926	-0.767051033
15	157.5990779	18.40092213	1.225207115
16	100.9394377	16.06056228	1.06937658
17	74.10066081	-6.100660809	-0.406206438
18	80.06483346	-4.064833456	-0.270652896
19	150.1438621	7.856137941	0.523093138



See supporting worksheet in Excel Doc (D. Clean Regression).

Section E



This first residual plot is from the data of the given regression model. I used the equation to determine the predicted values and then generate my residuals. I was most pleased with this graph, but to me it seemed to have minor heteroscedasticity properties. The two often ways to fix this are to use a transformation function such as LOG, or that there is a variable missing. I initially tried to transform the function with log to see if that would help my residual model. It gave me a more condensed model, that does seem more clustered together along the x and y axis, which is what we are looking for.

On the note of adding another variable, if we were able to plot time of day or day of week, that may give us a better plot. I can imagine the residual plot clustering on two ends, one for peak night time crime and one for day time crime.

Section F

The police office does not currently qualify for the funding offered by the governor. This is because they have a minimum standard of ~1.889101 police officers per incident, which is below the required 2.5 set by the governor.

I then generated 5,000 pieces of sample data for a Monte Carlo simulation. To generate these inputs, I used the Normal Inverse distribution with a mean of 1.889101 and a standard deviation of 1.189723 (derived from the cleaned data). This resulted in 1562 scenarios where the police department would meet the funding requirements and 3437 scenarios in which they would not.

See supporting worksheet in Excel Doc (F. Monte Carlo).

Section G

In the future, the police department will have a 29.0%-32.0% chance of receiving funding in the future from the governor's new policy, and a 68.0%-71.0% chance of not receiving funding. This was found by running the Monte Carlo simulation in Section F and calculating the probability for those 5,000 scenarios. While that gave me a good starting point, running these simulations more times (I ran them a total of 100 times), gives me the bounds of the actual percentages of receiving funding. Those bounds being 29.0%-32.0% chance of receiving funding in the future from the governor's new policy, and a 68.0%-71.0% chance of not receiving funding.

Section H

There are numerous precautions I would have with handling this data in the real world.

1. You would want to remove the location of the incidents. You want to do this because if criminals got a hold of this data, they could see which locations are best patrolled/serviced by the police and which places are not.
2. You would also want to remove the time of the incidents, for the same reasons at stated for removing the locations.
3. Most importantly I would encrypt this data whenever transferring it to a new location of giving it to another entity.

Section I

Statwing. "Interpreting Residual Plots to Improve Your Regression." *Interpreting Residual Plots to Improve Regression*, Statwing, docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/.

Western Governors University. *Fundamentals Of Data Analytics*. WGU, learn.zybooks.com/zybook/WGUFundamentalsOfDataAnalytics.