Keenin Vye

Fundamentals of Data Analytics — C740

IVP Task 1

Section A

I removed the following rows:

1. Row 226 because it did not have any Sector data associated with it.

I removed the following columns:

* 1. Event Clearance Code:
     1. 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.
  2. Event Clearance Description
     1. 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.
  3. Event Clearance Subgroup
     1. 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.
  4. Zone/Beat
     1. 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.
  5. Latitude and Longitude
     1. I removed these two columns because this information is present in the `location` column and there is no reason to have duplicate data.
  6. Initial Type Description
     1. 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.
  7. Initial Type Subgroup
     1. 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.
  8. At Scene Time
     1. I removed this because it was not present in all rows of the data.
  9. Event Clearance Date/time
     1. 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

|  |  |  |
| --- | --- | --- |
| SUMMARY OUTPUT | | |
|  | |  |
| *Regression Statistics* | | |
| Multiple R | 0.922070287 | |
| R Square | 0.850213614 | |
| Adjusted R Square | 0.840851964 | |
| Standard Error | 13.38325702 | |
| Observations | 18 | |

|  |  |  |  |
| --- | --- | --- | --- |
| RESIDUAL OUTPUT | |  |  |
|  |  |  |  |
| *Observation* | *Predicted Y* | *Residuals* | *Standard Residuals* |
| 1 | 150.7673045 | -6.964404479 | -0.536397367 |
| 2 | 156.7187382 | 44.46876185 | 3.424977229 |
| 3 | 89.5525581 | -8.599458099 | -0.662328946 |
| 4 | 127.8117746 | -9.968574586 | -0.767778088 |
| 5 | 119.3097265 | -6.931726478 | -0.533880512 |
| 6 | 77.64969075 | -4.894390747 | -0.376965226 |
| 7 | 98.05460621 | 3.392993793 | 0.261327862 |
| 8 | 121.8603409 | -9.48234091 | -0.730328445 |
| 9 | 86.15173886 | 4.365461145 | 0.336227149 |
| 10 | 81.9007148 | 4.517585199 | 0.34794372 |
| 11 | 118.4595217 | -3.348921667 | -0.257933434 |
| 12 | 77.64969075 | 4.669709253 | 0.359660292 |
| 13 | 81.05050999 | 9.46669001 | 0.729123015 |
| 14 | 166.0709911 | -11.33769107 | -0.873227231 |
| 15 | 115.9089072 | -13.09500723 | -1.008575452 |
| 16 | 74.2488715 | 3.971628496 | 0.305894219 |
| 17 | 81.05050999 | 2.63519001 | 0.202961931 |
| 18 | 150.7673045 | -2.865504479 | -0.220700716 |

The fit of the regression is fairly strong and I would have a large amount of confidence in it. What tells me this, is the R-square value. The R-square value of 0.8502 tells us that our data strongly fits the line of regression, but could be improved. Along with that, I have created an ANOVA table to view the Standard Error value. With a Standard Error of 13.383257, we have more assurance that the fit of our model is strong.

Section D

The outliers (Zones H) 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:

1. 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 |25|).

Once I plotted the new, clean data, I observed the R-square value was 0. 0.959, as opposed to 0.85 in the initial model with the outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| RESIDUAL OUTPUT | |  |  |
|  |  |  |  |
| *Observation* | *Predicted Y* | *Residuals* | *Standard Residuals* |
| 1 | 141.1909629 | 2.61193714 | 0.509543982 |
| 2 | 89.70189147 | -8.748791467 | -1.706738641 |
| 3 | 121.8825611 | -4.039361087 | -0.788009827 |
| 4 | 114.7313012 | -2.353301172 | -0.459088556 |
| 5 | 79.69012758 | -6.934827585 | -1.352865507 |
| 6 | 96.85315138 | 4.594448618 | 0.896297851 |
| 7 | 116.8766791 | -4.498679146 | -0.877614875 |
| 8 | 86.8413875 | 3.6758125 | 0.717087755 |
| 9 | 83.26575754 | 3.152542458 | 0.615006776 |
| 10 | 114.0161752 | 1.09442482 | 0.213503446 |
| 11 | 79.69012758 | 2.629272415 | 0.512925796 |
| 12 | 82.55063155 | 7.966568449 | 1.554140393 |
| 13 | 154.0632307 | 0.670069292 | 0.130718986 |
| 14 | 111.8707972 | -9.056897205 | -1.766844768 |
| 15 | 76.82962362 | 1.390876382 | 0.271336044 |
| 16 | 82.55063155 | 1.135068449 | 0.221432319 |
| 17 | 141.1909629 | 6.71083714 | 1.309168827 |

|  |  |
| --- | --- |
| SUMMARY OUTPUT | |
|  |  |
| *Regression Statistics* | |
| Multiple R | 0.979326566 |
| R Square | 0.959080523 |
| Adjusted R Square | 0.956352558 |
| Standard Error | 5.294139803 |
| Observations | 17 |

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 most often way to fix this issue is to use a transformation function such as Log. “The most common way to improve a model is to transform one or more variables, usually using a ‘log’ transform… After transforming a variable, note how its distribution changes, the r-squared of the regression changes, and the patterns of the residual plot changes. If those improve (particularly the r-squared and the residuals), it’s probably best to keep the transformation.” (Statwing, n.d.). After transforming my model with the Log function, I observed a more condensed model. One that does seem more clustered together along the x and y axis, yet has a slightly worse R-Square value of 0.950.

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 ~1.889952 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. “Thousands of values must be drawn for each input variable in a Monte Carlo simulation. More numerously drawn variables leads to smaller error in the predictions made by the model. Once the simulation is completed, the statistical analysis of the output can be analyzed.” (Fundamentals Of Data Analytics, n.d.). To generate these inputs, I used the Normal Inverse distribution with a mean of 1.889952 and a standard deviation of 1.189974 (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.

My recommendations to the precinct when handling this data are:

1. Encrypt the data so that if a 3rd party gets a hold of it, they cannot decipher the details of the incidents.
2. The precinct should hold this data in a password protected database so no unauthorized users can access it.
3. When the precinct talks about this data, they should reference every incident/event by its ID so that no contextual data about the infractions is talked about.
4. I would advise the officers at the precinct to not talk about the location or number of officers at the scene in order to not let that vital information to fall into the wrong hands.

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.