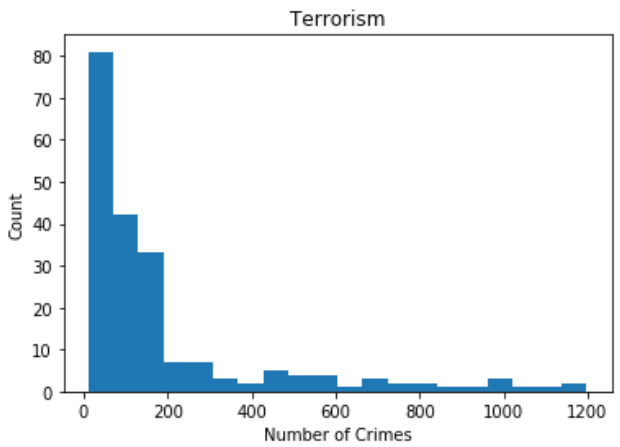
INVESTIGATION OF CRIME DATA IN RUSSIA (2003-2020)

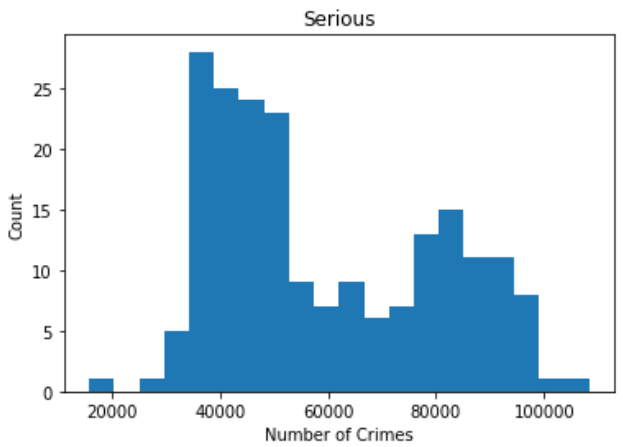
BY: EVAN EDMUNDS

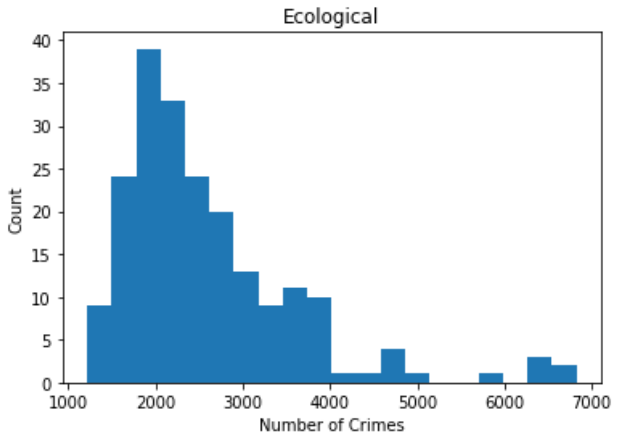
DSC 530

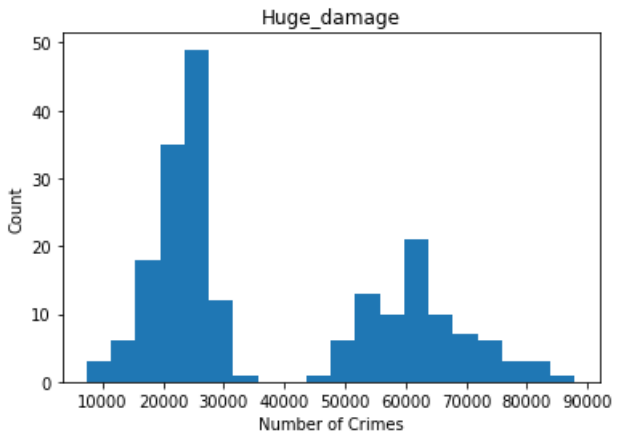
VARIABLES OF INTEREST

- month: The month crime data was collected
- Total_crimes: total number of crimes in a time period
- The rest of the variables are counts of specific crimes that were reported in a given month
  - Serious
  - Huge_damage
  - Ecological
  - Terrorism
  - Extremism
  - Murder
  - Harm_to_health
  - Rape
  - Theft
  - Vehicle_theft
  - Fraud_scam
  - Hooligan
  - Drugs
  - Weapons







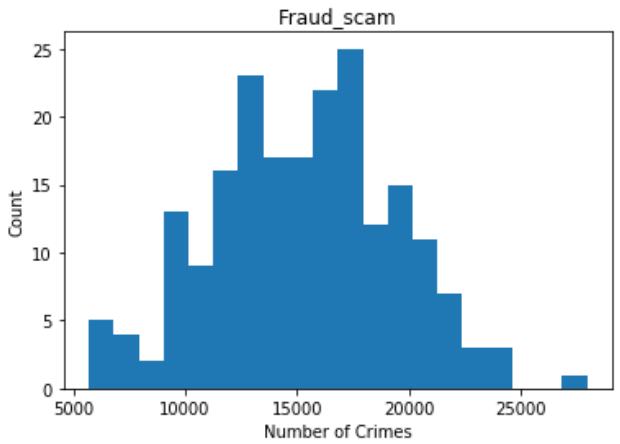


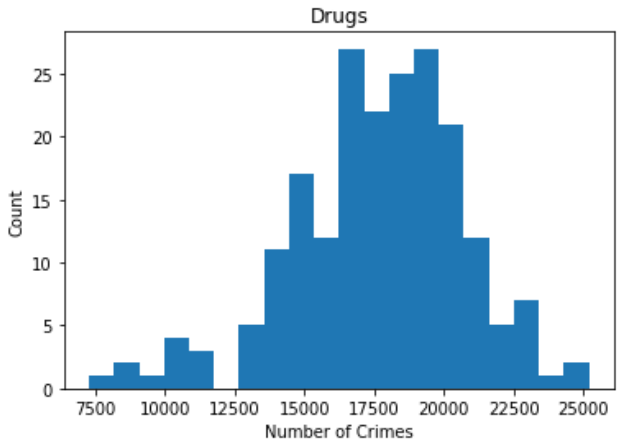
HISTOGRAMS

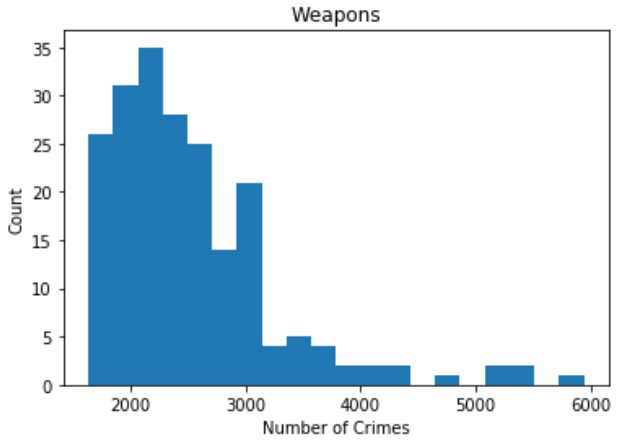
Looking at the plots, it seems that there are some outliers in the data:

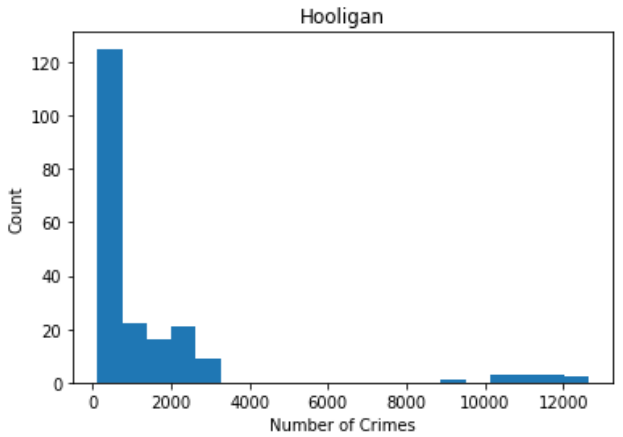
- In the Ecological variable, the right three bins with values.

HISTOGRAMS





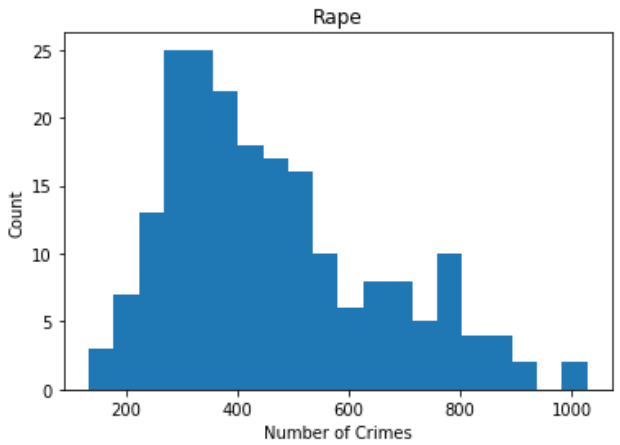


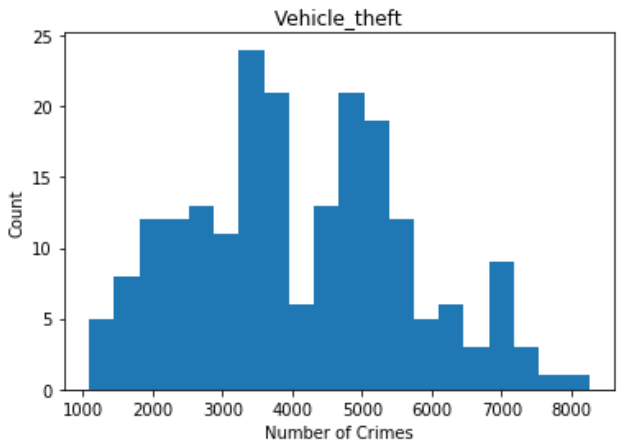


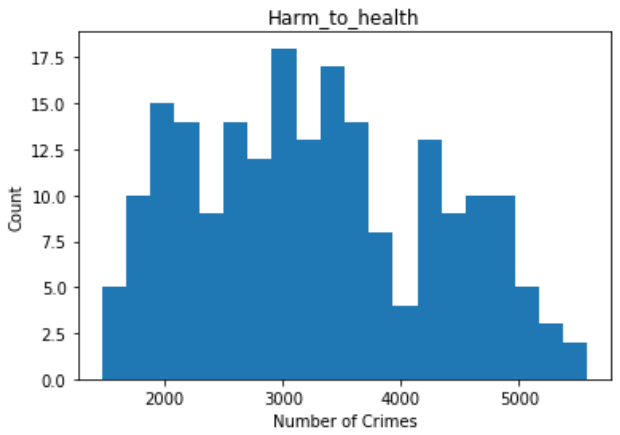
Looking at the plots, it seems that there are some outliers in the data:

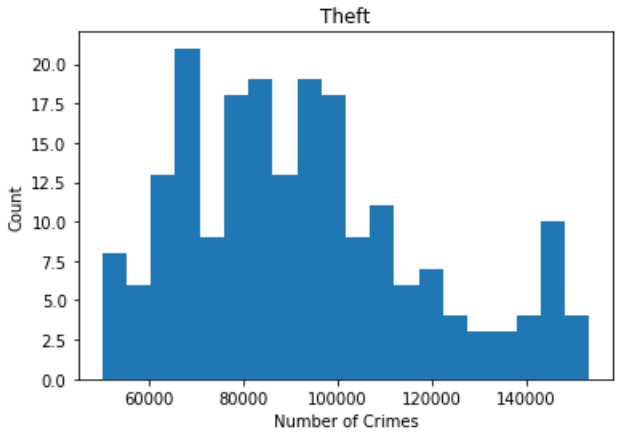
- In the Fraud_scam variable, the far right value
- For the Weapons and Drugs plots, I'm inclined to say that the smaller count bins on the extreme edges are more likely to be tails, than outliers.

HISTOGRAMS

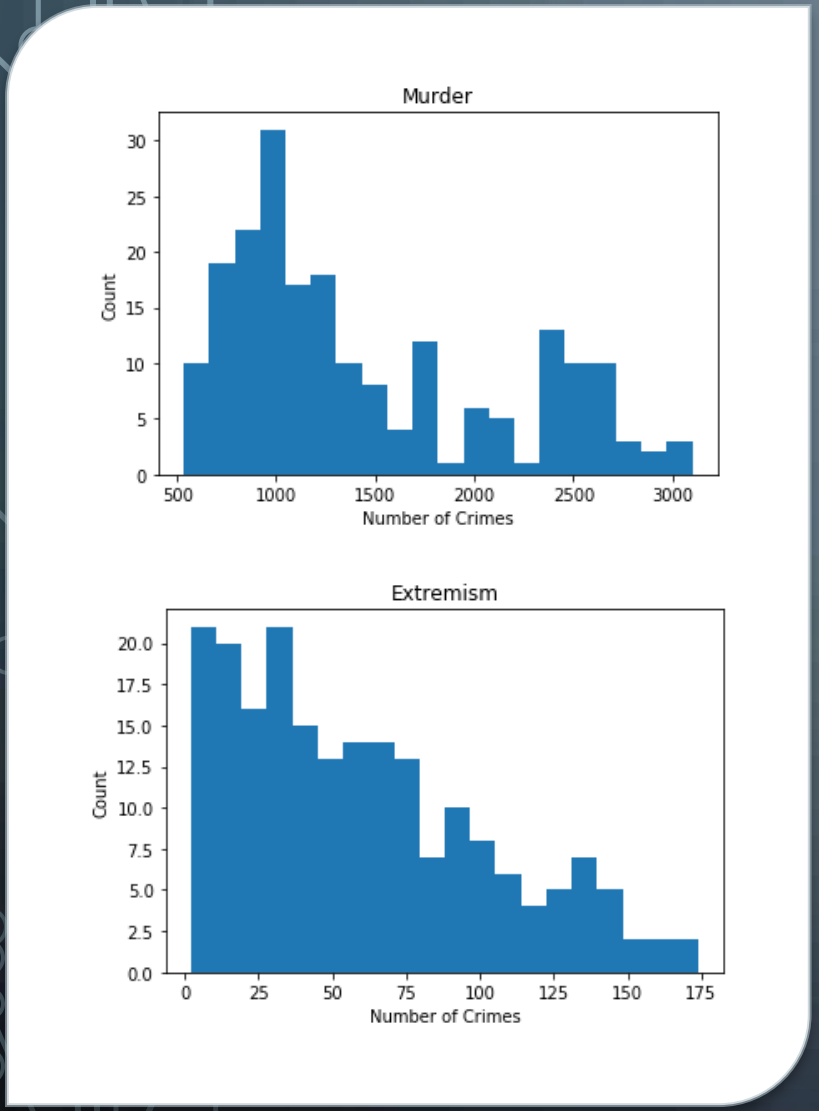








Looking at the plots, I can say that it looks like there are no outliers in the plots, and the values furthest from the modes are the tails of the distribution, not outliers.



HISTOGRAMS

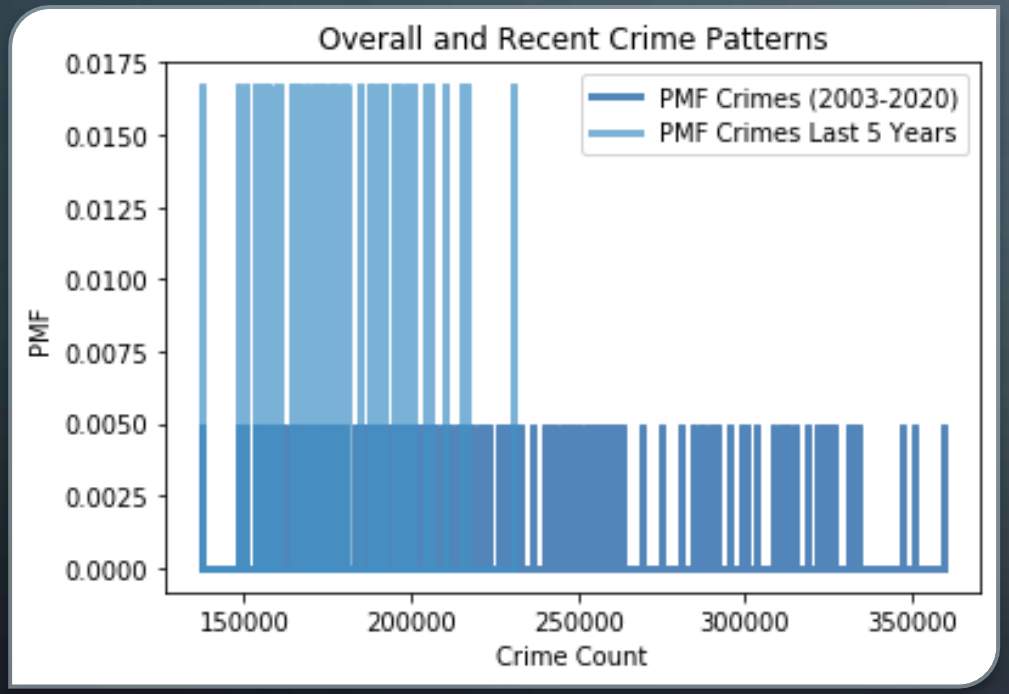
Looking at the plots, I can say that it looks like there are no outliers in the plots, and the values furthest from the modes are the tails of the distribution, not outliers.

Since all of these variables represent counts of real life crime rates, I'm going to leave all data in, even though there do seem to be some outliers present. Since I am doing all my correlation analyses using the spearman method, I'm not concerned with outliers inhibiting that part of my hypothesis testing. I also don't think the outliers are far enough from the mode to have a significant impact on the regression analysis.

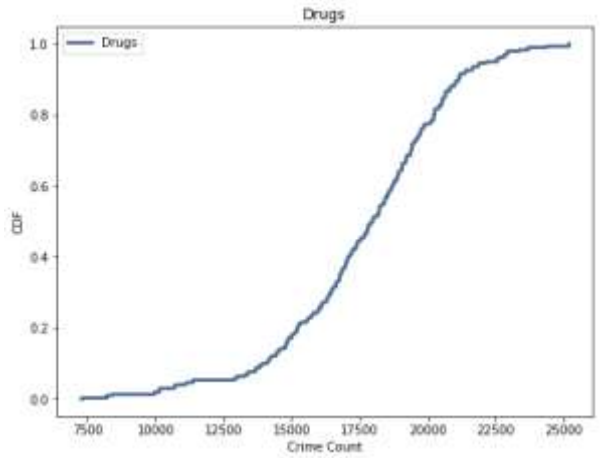
SUMMARY STATS OF ALL VARIABLES

- Serious Crime Stats: mean = 59068.49, mode = 65170.00, spread = 418261854.34, tails = -1.09.
- Huge_damage Crime Stats: mean = 38402.04, mode = 87941.00, spread = 428680561.49, tails = -1.22.
- Ecological Crime Stats: mean = 2592.46, mode = 2308.00, spread = 1051836.74, tails = 3.94.
- Terrorism Crime Stats: mean = 197.84, mode = 63.00, spread = 59958.00, tails = 4.63.
- Extremism Crime Stats: mean = 59.47, mode = 35.00, spread = 1761.29, tails = -0.37.
- Murder Crime Stats: mean = 1451.40, mode = 1287.00, spread = 470643.19, tails = -0.76.
- Harm_to_health Crime Stats: mean = 3276.02, mode = 5021.00, spread = 1022901.57, tails = -0.90.
- Rape Crime Stats: mean = 463.58, mode = 515.00, spread = 35326.43, tails = -0.10.
- Theft Crime Stats: mean = 92647.56, mode = 153394.00, spread = 640817694.90, tails = -0.25.
- Vehicle_theft Crime Stats: mean = 4101.41, mode = 7047.00, spread = 2439404.54, tails = -0.59.
- Fraud_scam Crime Stats: mean = 15277.83, mode = 16912.00, spread = 16846321.88, tails = -0.22.
- Hooligan Crime Stats: mean = 1457.65, mode = 194.00, spread = 6526162.47, tails = 10.17.
- Drugs Crime Stats: mean = 17644.36, mode = 20176.00, spread = 9367464.14, tails = 0.84.
- Weapons Crime Stats: mean = 2525.02, mode = 2089.00, spread = 534915.95, tails = 4.79.

CRIME PATTERNS OVERALL AGAINST RECENT

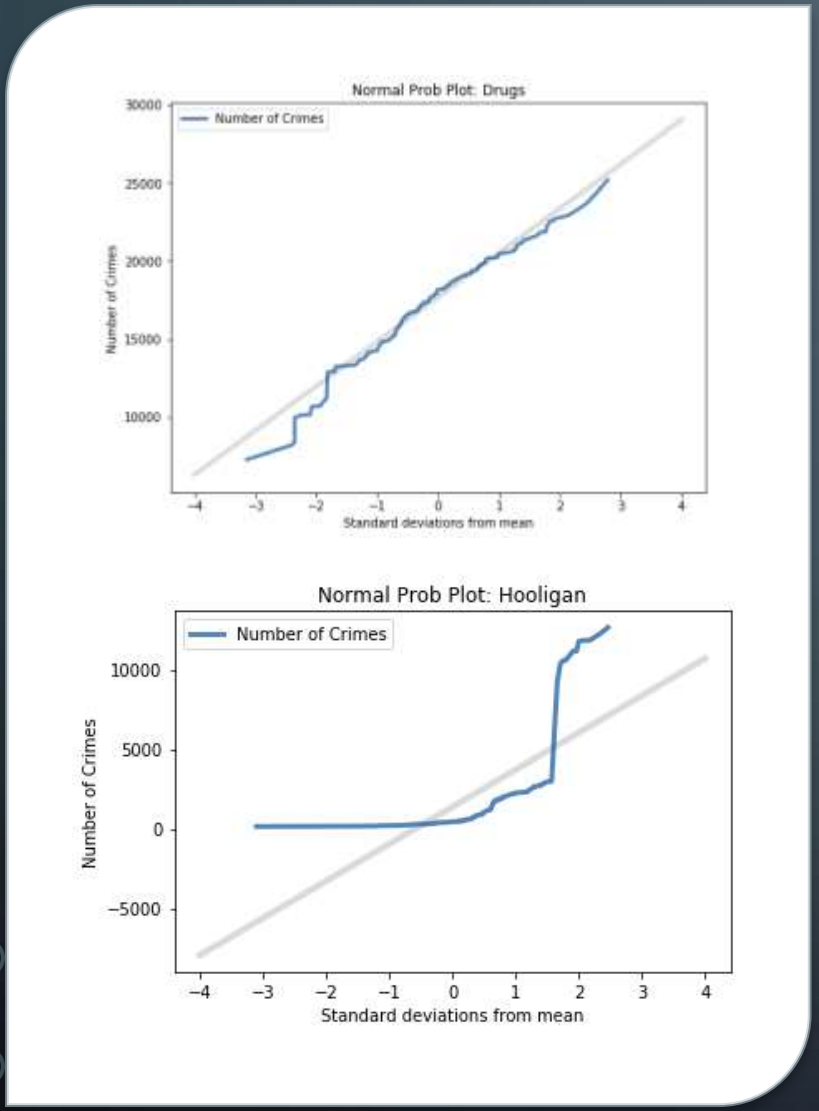


- Created using the Total_crimes variables.
- The plot shows values are more tightly packed than the overall data and gives the impression of a lesser crime rate than the full time length data.



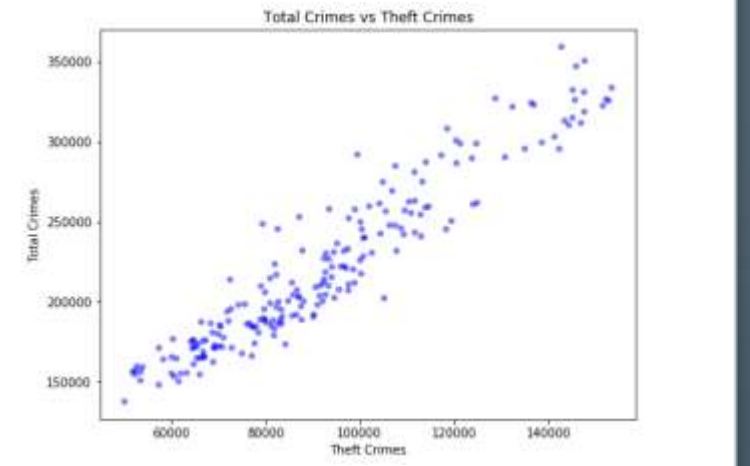
CDF ANALYSIS

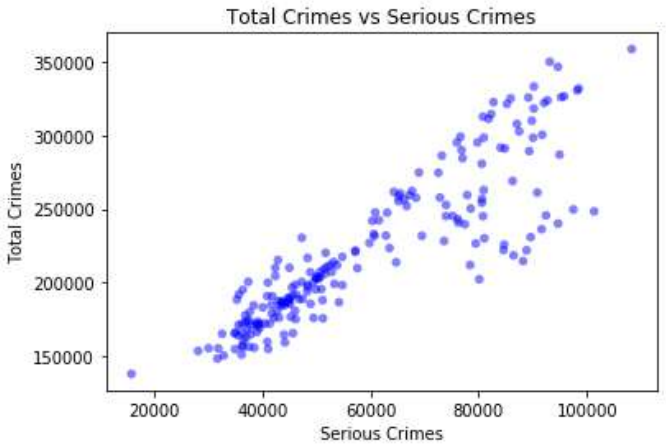
- A CDF of the Drugs and Hooligan variables
- The Drugs CDF bears a somewhat sigmoidal shape, indictating a normal distribution, while the Hooligan CDF lacks this pattern and more resembles a Pareto distribution.
- Illustrates that there is a variance in the normality between different categories of crime. Therefore, linear regression may not be the best predictor for all variables.



ANALYTIC DISTRIBUTION

- A further investigation of the CDFs of the Hooligan and Drugs variables
- These normal probability plots confirm that Hooligan is not normally distributed.
- It also provides more insight into Drugs, showing that while it is normal to about 1.5 standard deviations from the mean, the tails veer drastically from normally greater than 1.5 standard deviations.

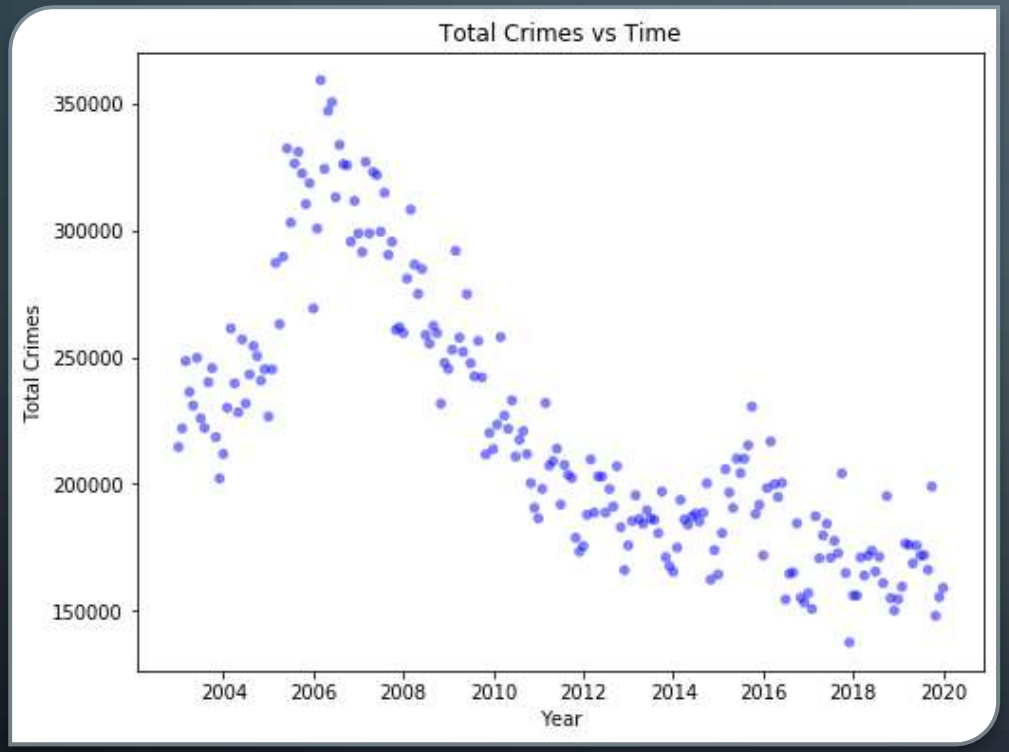




SCATTERPLOTS

- Serious crime stats
  - Spearman's correlation: 0.91
  - Covariance: 938,357,713.01
- Theft crime stats
  - Spearman's correlation: 0.95
  - Covariance: 1,240,667,861.83
- Both of these plots show a possible strong correlation with the Total_crimes variable. The correlation value supports this, as does the covariance value (may be due to large values).
- Either variable should be a good predictor for total crimes.

ADDITIONAL SCATTERPLOT



- This plot shows that the greatest count of crimes happened around 2006 and has decreased since then.
- This answers my question of whether total crime counts have changed over time.

HYPOTHESIS TESTING

It was my hypothesis that there would be strong correlations between crime of different categories. To test this, I calculated the correlation between Serious and Theft using the Spearman method since I have now seen that they are both correlated with Total_crimes. Then I used the CorrelationPermute function on pg. 125 of *ThinkStats* to calculate the p-value. The results are below:

- Spearman Correlation - 0.83
- P-value – 0.0

LINEAR REGRESSION MODELS

Theft regression analysis

Intercept 4.19e+04 (1.07e-18)

Theft 1.94 (1.68e-104)

R^2 0.9023

Std(ys) 5.159e+04

Std(res) 1.616e+04

Serious regression analysis

Intercept 8.88e+04 (1.62e-42)

Serious 2.24 (6.75e-71)

R^2 0.7908

Std(ys) 5.159e+04

Std(res) 2.365e+04

- Two simple regression models were created, Theft predicting Total_crimes and Serious predicting Total_crimes. These two were chosen based on their high correlation with the Total_crimes variable. R-squared values suggest that the Total_crimes vs Theft model is better, as it explains 90.23% of variance between the two variables.