Data Analysis and Visualisation Principles

Assignment Report

Student Name: Vivian Omo-Ojugo
Student Code: S4100903
Module Title: CT7202 - Data Analysis and Visualisation Principles
Module Tutor: Dr Thiago Viana
School of Computing and Technology
University of Gloucestershire 4/05/2021

Data Analysis of Twenty-Four Month Period of Crown Prosecution Service

Introduction:

Dataset

For this report, I will make use of a twenty-four months data set of Crown Prosecution Service Case Outcomes by Principal Offence Category available to be downloaded at data.gov.uk site and Moodle. The Dataset, if gotten from Moodle, is in zip format and has datasets for various years, of which, we are expected to download a twenty four months period of data that is, two years data set for analysis. Some years have incomplete months and in order to carry out proper analysis on this dataset, these missing months need to be filled in using either of these methods:

i) Extrapolation

ii) Using the mean based on different years or months

iii) Filling with the next month available in the data set.

For this analysis, I uploaded two years namely: 2014 and 2015 with the former having no missing month and the latter missing the month of November. In order to make up for the missing month of November in 2015, I added the January 2016 month. An advantage to this method is that I get to carry out my analysis, with real data but then, there would be no seasonality compared to using the Mean whereby I get to pick up trends for the missing months, but I miss out the trend of the missing month which happens to be November 2014 in this case. For the extrapolation method, if properly predicted, it would be the most accurate method but it might also give room for bias because for the missing month, a different situation might occur which the prediction is unlikely to pick up and this might just skew the dataset. Consider a situation where we are looking out for unique crimes, which would be the outlier(s) in this dataset but, a prediction model wont pick up the outliers.

Each month's file consists of 44 rows and 51 columns. 43 are the different states and regions of UK while 1 row contains the national data that is sum of all the 43 rows. There are 12 categories of every crime with 4 columns each containing Convicted, Percentage

Convicted, Unsuccessful and Percentage Unsuccessful which makes up a total of 48 columns. Of these 51 columns, there is 1 unique category named “Number of Admin Finalized” which was 100% unsuccessful thereby making it 2 columns namely, “Number of Admin Finalized Unsuccessful” and “Percentage of L Motoring Offences Unsuccessful”.

Problem Analysis:

In this section I will try to identify and analyze the problems that we may have to face and solve while dealing with this data. This data is in 24 files, so we can either combine these files into 1 single file, which is also a common practice, or work on the 24 files individually but working on files individually will expand the work unnecessarily, make it more complex and limit the analysis that may be extracted from the data.

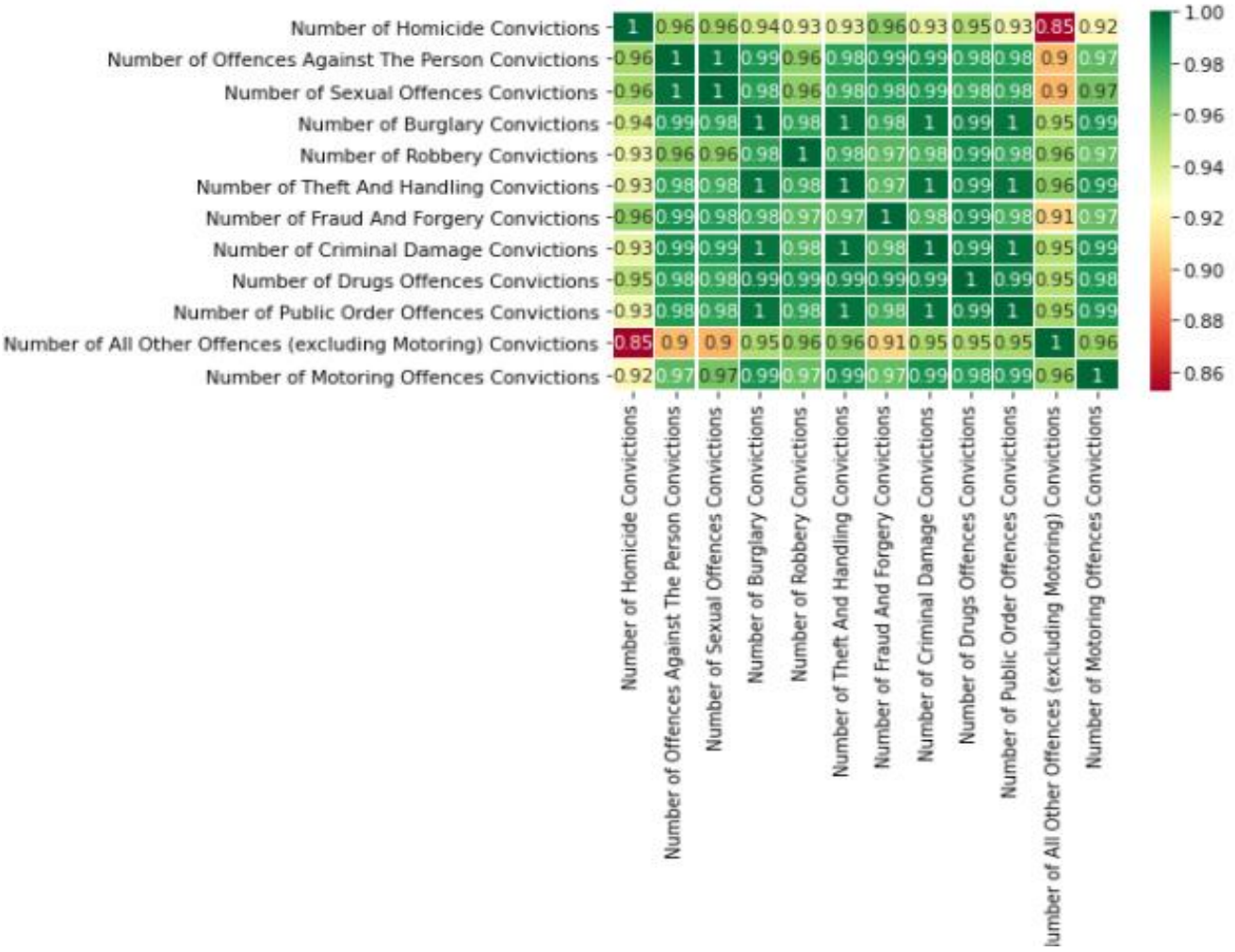
The other apparent issue with this data is that there is no unique identifier in the data by de fault and data in all files appear identical. In order to get a year and month column to disting uish between files, I go ahead to merge all twenty four months columns into one file. Also, data cleaning needs to be done. For this, I search for missing values using np.nan and replac e them with the number ‘0’. I also use regular expressions to eliminate the ‘%’, ‘,’ symbols a nd to remove empty spaces, after this, I use the isnull.sum function to ensure there are no missing values in the dsataset. I then drop some columns leaving a total of fourteen column s which are:

('Date', 'County', 'Number of Homicide Convictions’, 'Number of Offences Against The Perso n Convictions', 'Number of Sexual Offences Convictions', 'Number of Burglary Convictions', ' Number of Robbery Convictions', 'Number of Theft And Handling Convictions', 'Number of F raud And Forgery Convictions', 'Number of Criminal Damage Convictions', 'Number of Drug s Offences Convictions', 'Number of Public Order Offences Convictions', 'Number of All Othe r Offences (excluding Motoring) Convictions', 'Number of Motoring Offences Convictions').

In order to visualise the strength of the relationships between the remaining variables in each column, I plot a correlation heatmap using seaborn. With the correlation varying between -1 and +1, where +1 means, there is a perfect positive linear correlation, 0 means no correlation whatsoever and -1 indicates a perfect negative linear correlation.

Let us take a look at the figure figure below.

Fig 1. Correlation Heatmap



From the Heat map above, we can see that the number of offences against the person and t he Number of sexual offences conviction has a perfect positive correlation value of 1. We ca n also see that number of all other offences(excluding motoring) Convictions, has a negativ e correlation of -0.85 with Number of Homicide convictions.

To further narrow it down, I plot a scatter matrix of select variables using seaborn library ag ain. For this purpose, I temporarily rename columns from V1 through to V14, visualise, and t hen drop more columns, leaving five variables with about three of the five having various co rrelation levels. The five columns for further analysis are:

1: 'Date' == V1

2: 'County' == V2

3: 'Number of Offences Against The Person Convictions' == V4

4: 'Number of Sexual Offences Convictions' == V5

5: 'Number of Robbery Convictions' == V7

Fig 2i. Scatter matrix with five variables

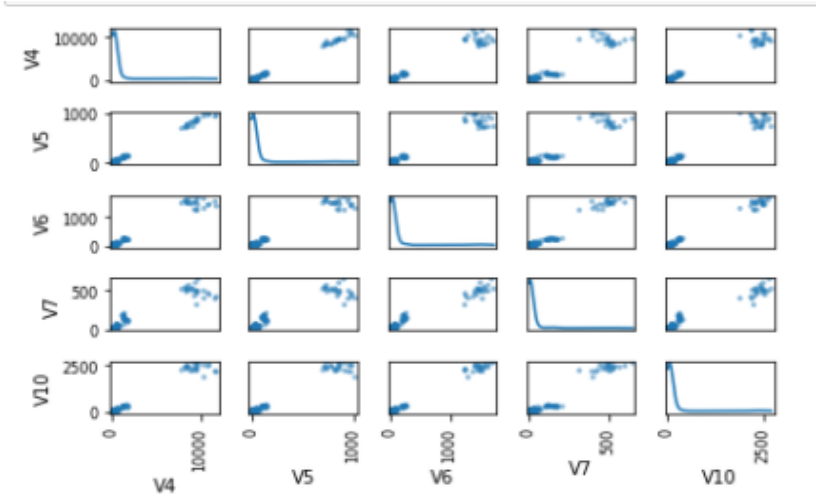
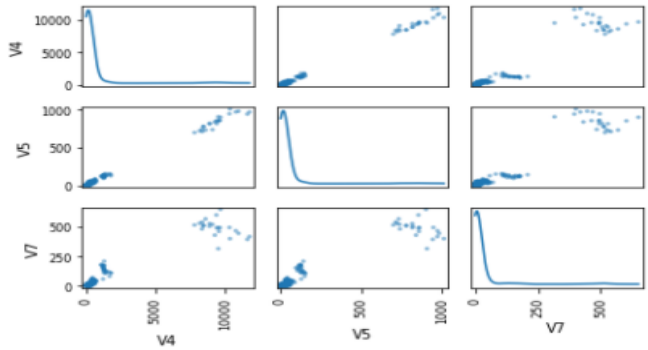


Fig 2ii. Scatter matrix of narrowed down variables



In the figure above there appears to be a strong positive correlation between variables' V4(Number of Offences Against The Person Convictions) and V5(Number of Sexual Offences Co nvictions)This means that as the number of V4 offence increases, so does the number of V5

offence and vice versa. Again there seems to be some sort of relationship between, V7(Number of Robbery Convictions) and V4(Number of Offences Against The Person Convictions) /V5(Number of Sexual Offences Convictions). This discovery leads me to analyse these variables further.

For ease in referencing, Variables are renamed in this manner:
Number of Offences Against the Person Convictions changed to 'Person'
Number of Sexual Offences Convictions changed to 'Sexual'
Number of Robbery Convictions changed to ' Robbery'.

After this I then got ahead to get the Descriptive statistics of my new dataset. Find statistics below:

Fig 3.

	Person	Sexual	Robbery
count	1032.000000	1032.000000	1032.000000
mean	439.207364	39.205426	22.852713
std	1415.404559	126.673383	76.440271
min	29.000000	0.000000	0.000000
25%	109.000000	8.000000	3.000000
50%	169.500000	14.000000	6.000000
75%	267.250000	27.000000	11.000000
max	11741.000000	1011.000000	650.000000

From the statistics above, all variables have different averages/means with the mean of the Person crime been the largest. This tells us that the Person crimes, is the highest crime followed by Sexual crimes and then Robbery crimes, which has the smallest mean value. Also, all variables, have standard deviations which are way higher than their corresponding means, this indicates a rightly skewed dataset and it is an indication that the standard deviation is spread over a wide range thus making the dataset unreliable. A dataset with less skew, and standard deviation less than half of the mean will be a more reliable dataset.

Also, from the table above, we can see that the 50% (50^{th} percentile) which are the median values are significantly lower than their corresponding means, for every column. This again indicates that the dataset, is rightly skewed. Also, due to the difference in value between the mean, median and mode, we can rightly say that the data distribution is asymmetric. In a situation, where a dataset is skewed or asymmetric as in the case of the dataset being analysed here, the median will be a better measure of central tendency because it is less susceptible to the influence of outliers unlike the mean, which is prone to the influence of outliers.

Data Visualisation:

In this session, I carry out univariate, bivariate and multivariate analyses by visualising box plots and scatterplots.

Fig 4i.

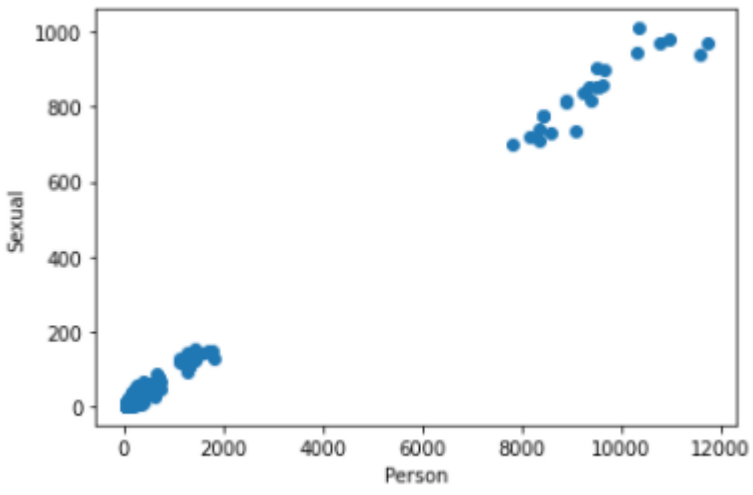


Fig 4ii.

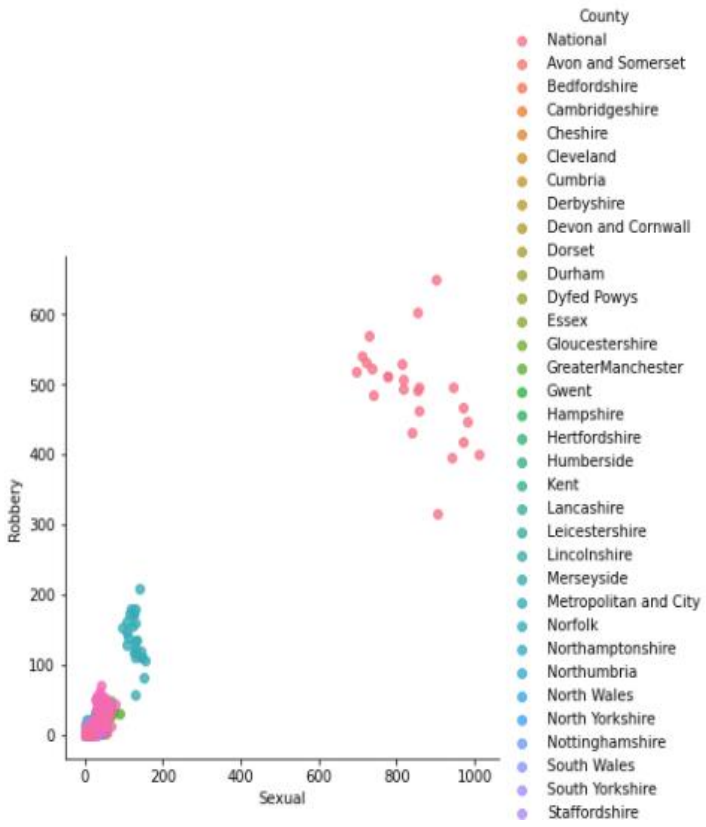


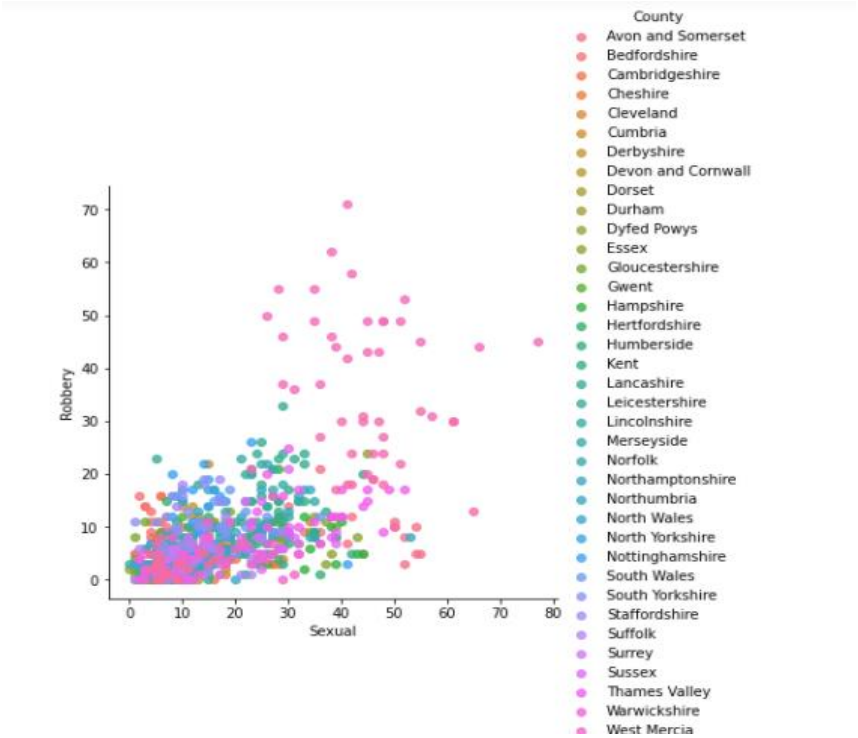
Figure 4i) above, is a scatterplot showing a possible relationship between Number of sexual offence convictions and Number of robbery convictions. But there appears to be two groups of this crime as a cluster is seen in the lower left corner of the graph with another cluster at the upper right corner of the same graph and a large gap in between both groups. It could be as a result of data collection methodology errors or some other reason.

Figure 4ii) above is a multivariate analysis graph. It is plotted using three variables, namely Sexual, Person and Counties as the hue. In this graph, we can see grouped clusters at the lower left and upper part of the graph which seem to show slight correlation, between the variables, with some counties having very high Sexual and Robbery crime convictions. Clusters and gaps can affect the result of my analyses so I would go ahead to eliminate them.

To get a better view of these scatter plots, i would go ahead to remove certain counties. I would then drop certain rows, which are the counties that are highly populated. Namely:

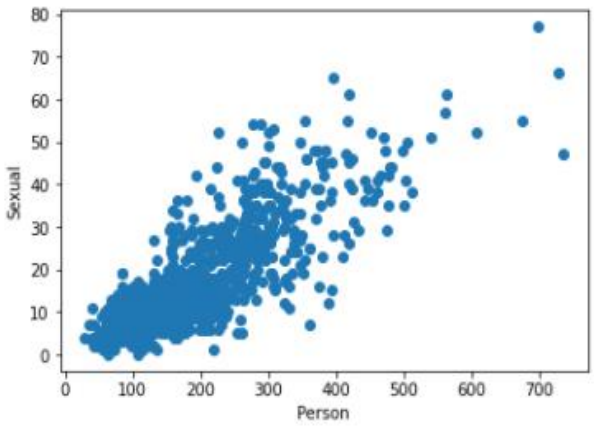
National, Metropolitan and Greater Manchester. After dropping these rows, column numbers drop from 1032 to 960. I would then visualize these scatter plots again to get the figures below:

Fig 5i.



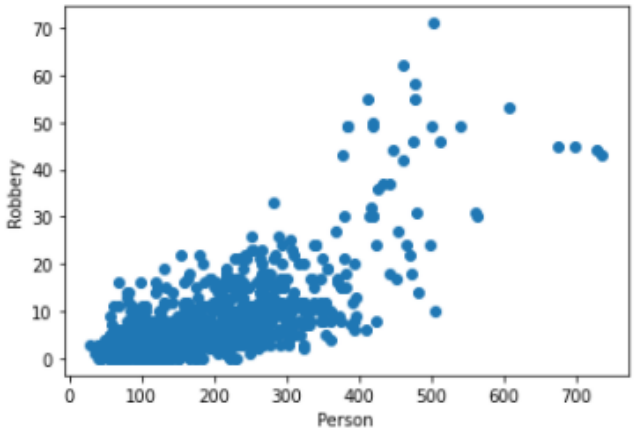
After taking out the highly populated cities, I visualize the three variable scatter plot again and fig 5i is produced without visible gaps in the plot and reduced clusters and a positive correlation.

Fig 5ii.



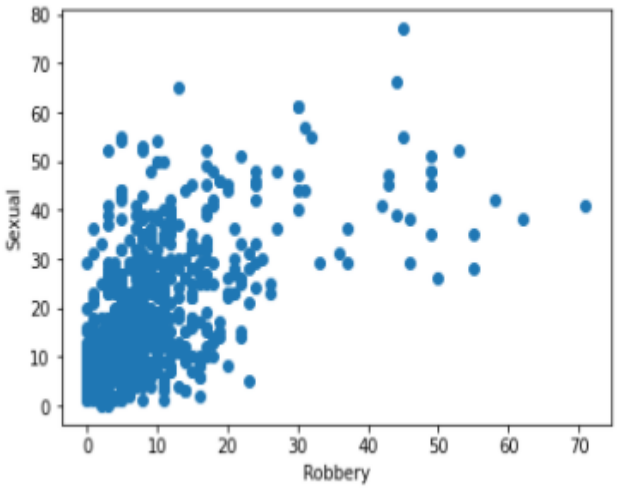
The scatter plot above is one which plots Sexual crimes against Person offence and it shows a certain level of positive correlation, with few outlier observations.

Fig 5.iii.



The scatter plot above shows the relationship between Robbery and Offence against the person crimes. Its shows a positive correlation that does not appear so strong or linear per say.

Fig 5.iv.



The relationship depicted between Sexual and Robbery crimes in the above visualized image is neither linear nor correlated as the dots are scattered all over the graph. There are also some outliers observed in the graph.

For the scatter plots visualized above, the mere fact that i noticed certain levels of correlations amongst some variables, does not necessarily indicate that a change in one variable, is the reason for a change in the other variable. In statistics, it is stated that 'correlation is not causation'. It could be that the relationship is as a result of some other factor(s) or it could be coincidence.

I will investigate these relationships further with other forms of visualization. Worthy of note is that after dropping certain rows, that is taking out the counties, there are certain adjustments, in the data statistics. Find the new descriptive statistics below:

Person	Sexual	Robbery	
count	960.000000	960.000000	960.000000
mean	186.822917	16.613542	7.960417
std	105.616053	12.439117	8.803645
min	29.000000	0.000000	0.000000

	Person	Sexual	Robbery
25%	106.000000	7.000000	3.000000
50%	161.000000	13.000000	6.000000
75%	248.250000	24.000000	10.000000
max	736.000000	77.000000	71.000000

Although the standard deviations of the dataset are still high, which still leaves it right skewed with high variance, the mean and the median values are closer now.

I would go ahead to do a univariate analysis on Person, Sexual and Robbery variables using boxplots and histograms visualizations. The box plot will enable me show the results in a single graph as well as summarise the data. Another useful attribute of the boxplot, is its ability to compare different categorical data more easily and its usefulness in detecting ouliers. These boxplot will show the five number summary of the dataset and they are: minimum score, first quartile, median, upper quartile and maximum score.

Fig 6.i.

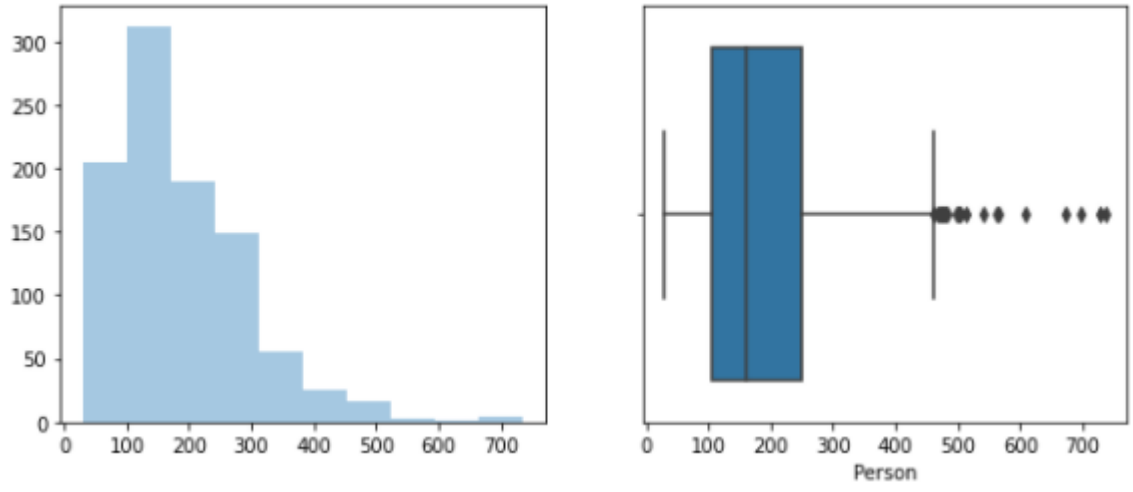


Fig 6.i. above shows a histogram and a boxplot of the offence against the person crime. From the image of the histogram , we can see that the majority of the data values are concentrated on the right side with one mode as one peak can be sighted. For the box plot, we can also see that the median is closer to the left/bottom part of the box which also gives an indication that the distribution is positively skewed. Again, there are values located outside of the whiskers of the boxplot .These values are outliers.

Fig 6.ii.

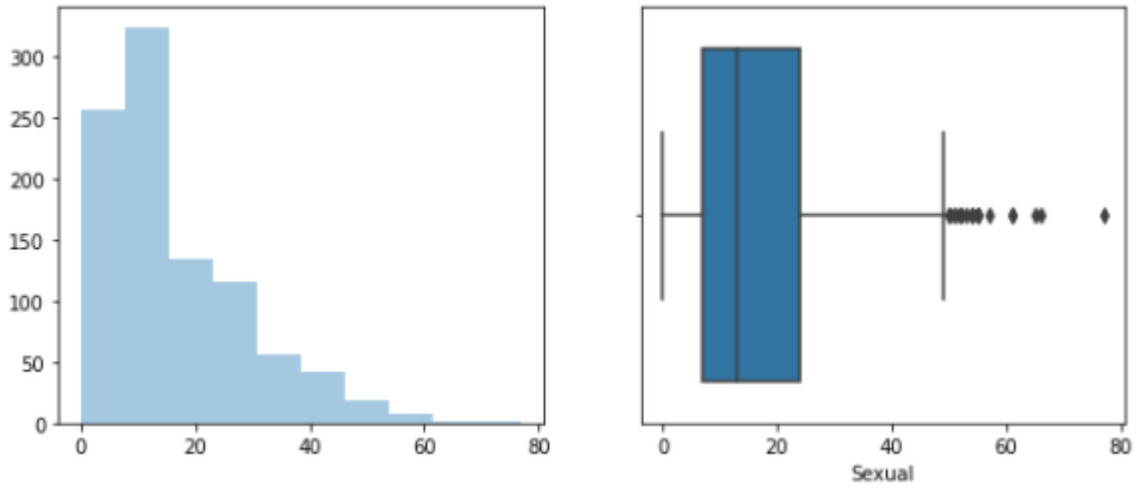


Fig 6.ii. above shows a histogram and a boxplot of the Sexual crime. From the image of the histogram , we can see that the majority of the data values are concentrated on the right side with one mode as it has only one peak. For the box plot, we can also see that the median is closer to the left/bottom part of the box which also gives an indication that the distribution is positively skewed. Again, there are values located outside of the whiskers of the boxplot .These values are outliers.

Fig 6.iii.

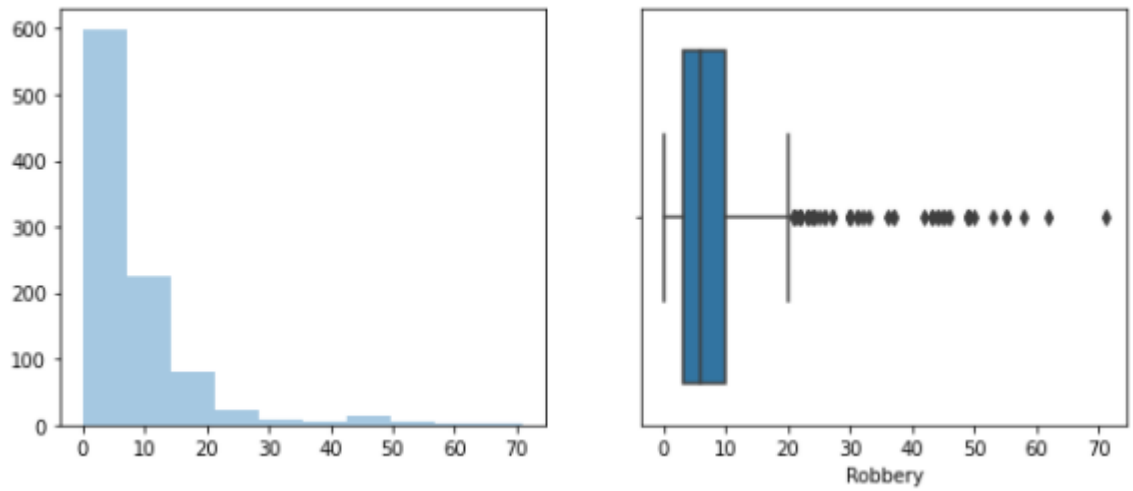


Fig 6.iii. above shows a histogram and a boxplot of the Robbery crime. From the image Of the histogram , we can see that the majority of the data values are concentrated on the right side with one mode as one peak can be sighted. For the box plot, we can also see that the median is closer to the left/bottom part of the box which also gives an indication that the distribution is positively skewed. Again, there are values located outside of the whiskers of the boxplot .These values are outliers.

I will now go ahead to treat these outliers as they are likely to distort my predictive data ana lyses. Before treatment, I first check for the level of skew of my dataset. For the 'Person' var iable I get a skew value of 1.3289061374900129, for the 'Sexual' variable, I get a value of 1.2 730440930691227 and for the 'Robbery' variable, I get a value of 2.9661517095320242. As we can see, all values are above '1' and the level of skew of a dataset should range between -1 and +1 for it to be considered normal. This further emphasizes the fact that the outliers n eed to be treated. For the treatment of my outliers, I make use of the interquartile range me thod (IQR). Other methods such as the trimming method can be used also, but that would m ean removing half of my dataset. So in other to preserve my dataset for correct analyses, I c hose the method mentioned above. After treatment of outliers, I rename the variables and visualise again using boxplots and histograms and there are no outliers this time. Below are the images after outlier treatment.

Fig 7.i.

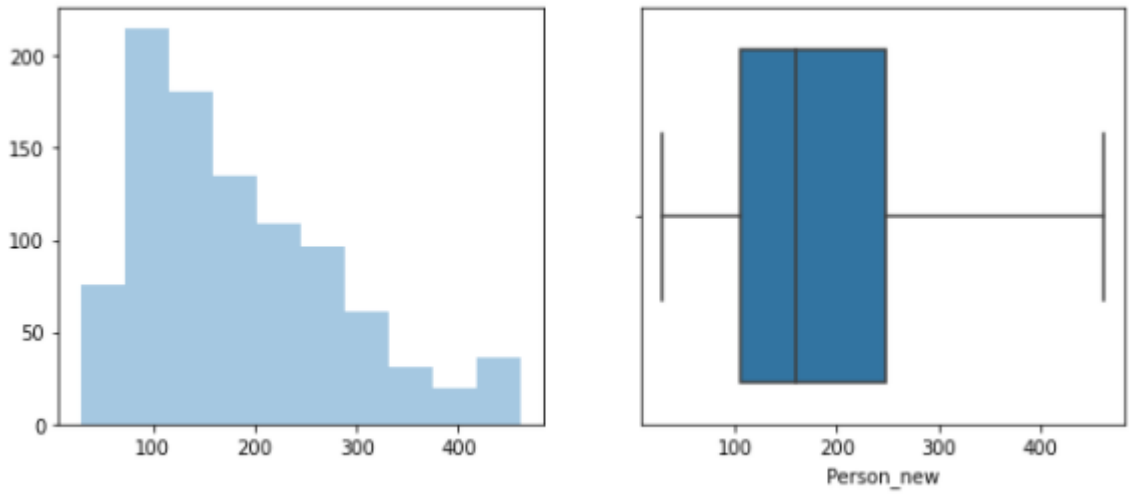


Fig 7.ii.

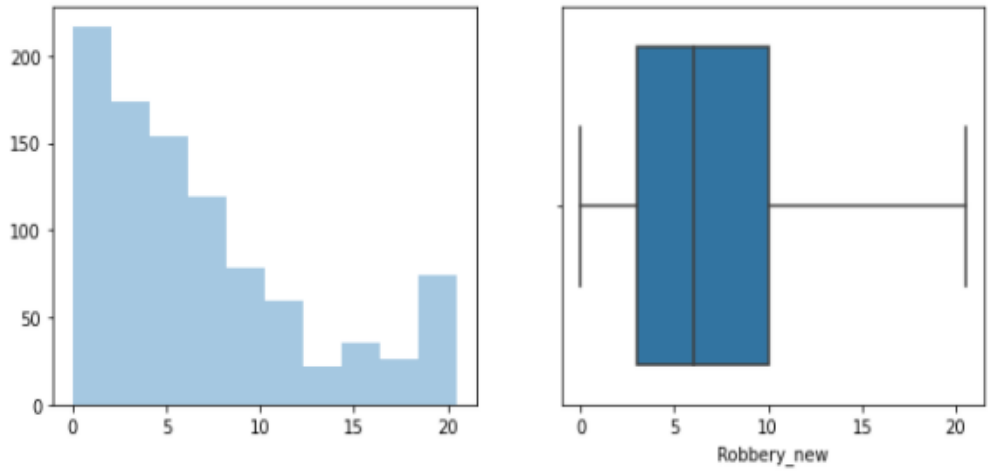
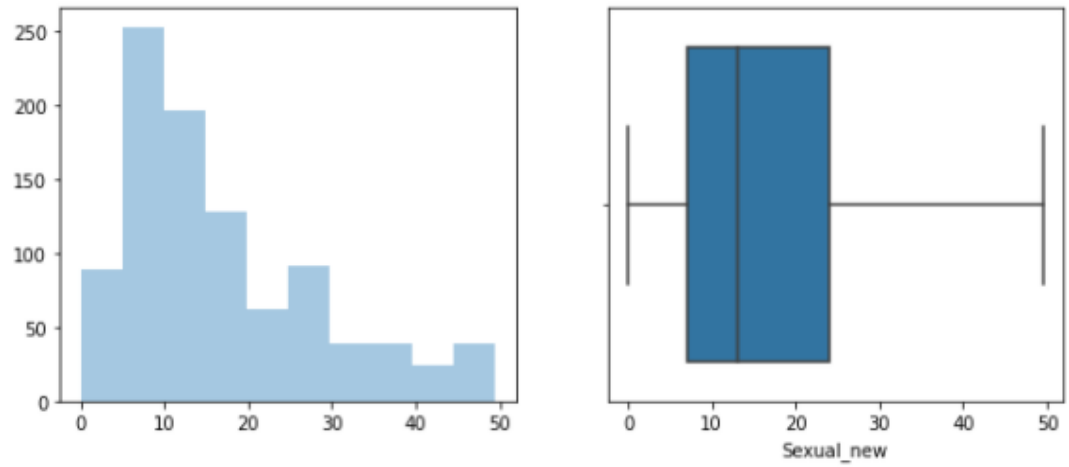


Fig 7.iii.

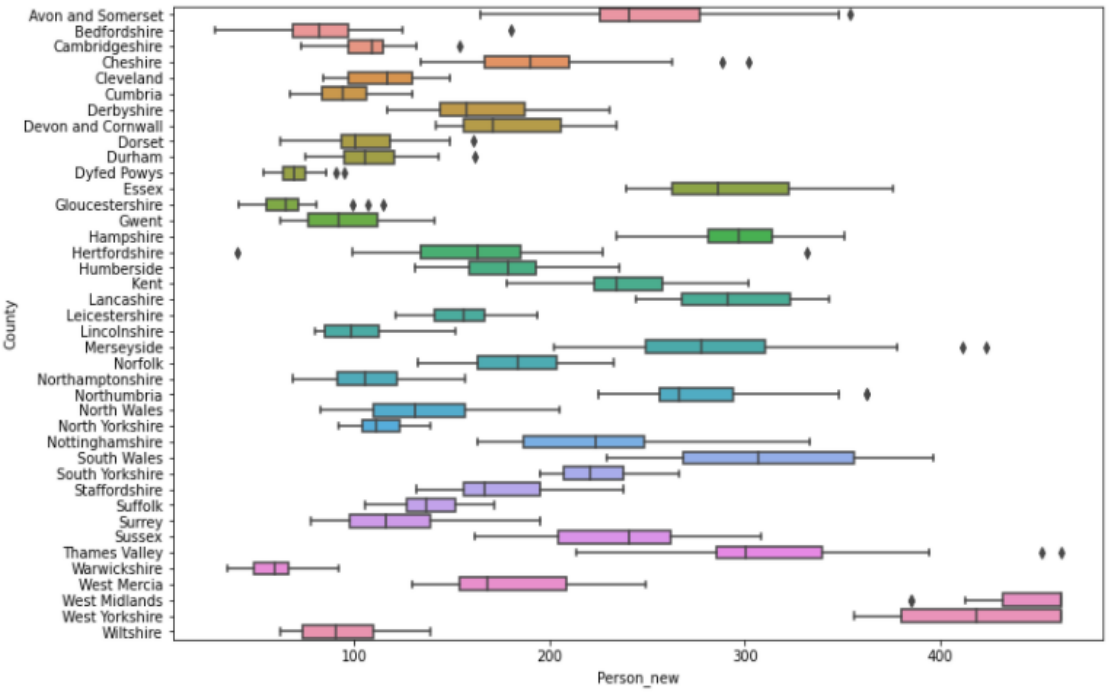


I also get the descriptive statistics for the new dataframe and the values are different as adj ustments have been made. See figures below:

	Person_new	Sexual_new	Robbery_new
count	960.000000	960.000000	960.000000
mean	185.023698	16.474479	7.092708
std	99.218821	11.988225	5.735684
min	29.000000	0.000000	0.000000
25%	106.000000	7.000000	3.000000
50%	161.000000	13.000000	6.000000
75%	248.250000	24.000000	10.000000
max	461.625000	49.500000	20.500000

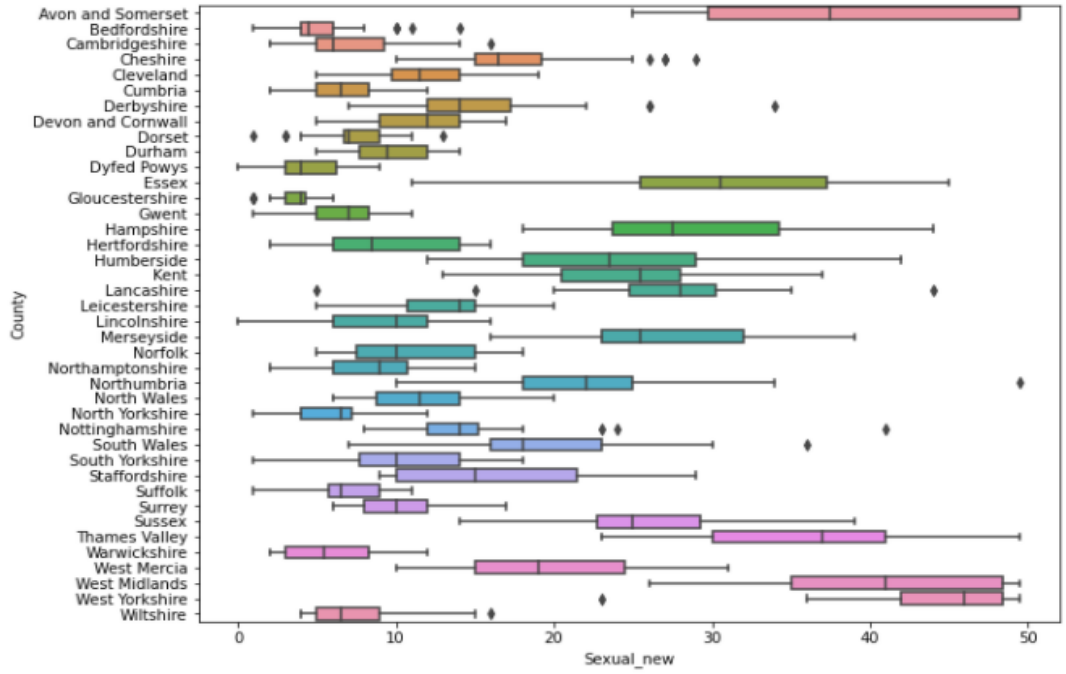
In comparison to the first two statistics, the standard deviations of the variables are all less than the mean, and the mean, though still higher than the median, has its values way closer to the mean this time.The dataset is also less skewed. All these, make the dataset much more reliable for predictive analysis.

Fig 8.i. Boxplot of Person Crimes Across Different Counties



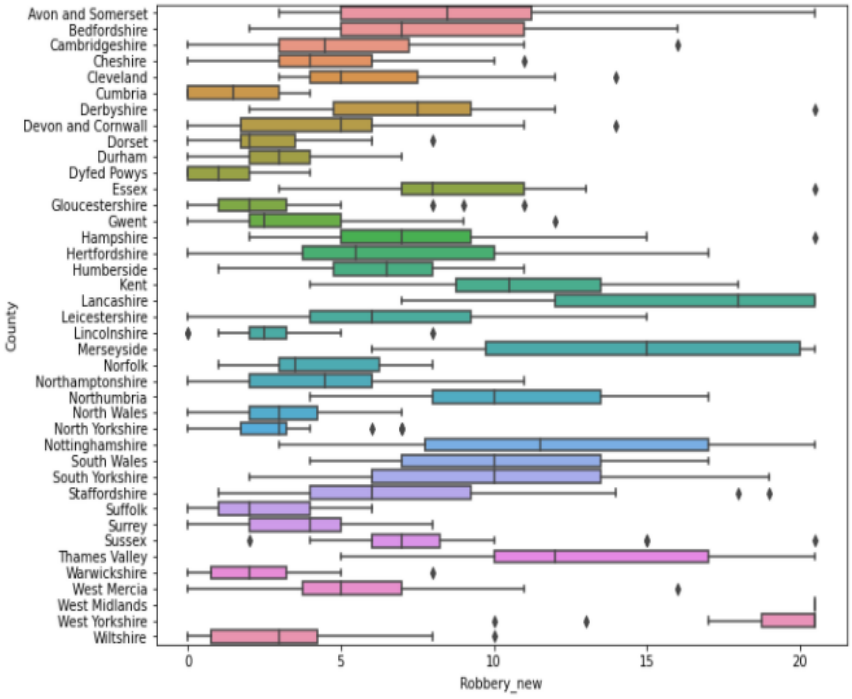
In the figure above, we can see that West Midlands, has the highest level of offence against the person crime and Warwickshire has the lowest level of offence against the person crime.

Fig 8.ii. Boxplot of Sexual Crimes Across Different Counties



The figure above, shows that Avon and Somerset, Thames Valley, West Midlands and West Yorkshire have very high levels of Sexual offence crimes with Gloucestershire and Bedfordshire having the lowest rates.

Fig 8.iii. Boxplot of Robbery Crimes Across Different Counties



From Fig 8.iii, we can see that Robbery rates are highest in Avon and Somerset, West Midlands and West Yorkshire and lowest in Dyfed Powys.

Hypothesis Testing:

Pearson correlation coefficient

This is a hypothesis test, used to check the linear relationship between two variables.

Null Hypothesis(H_0) states that there is no relationship between sexual crime and offence against the person crime.

Alternate Hypothesis(H_A) states that there is a relationship between sexual crimes and offence against the person crimes.

I carried out pearson correlation test and below is an image of the result obtained

```
]: from scipy.stats import pearsonr
   df = pd.read_csv('data2.csv')
   pearsonr(df['Sexual_new'], df['Person_new'])

]: (0.8192994359744694, 1.1890195386959991e-233)
```

What does this mean?

The first value 0.8192994359744694 shows that there is a positive correlation between Sexual and Person crimes. The second value 1.1890195386959991e-233 which is the P-value, is way less than the 0.05. Therefore, with a 99.9% confidence interval, Null Hypothesis can be rejected in favour of the alternate hypothesis.

Independent T Test

Null Hypothesis: Crime changes statistically enough between January to November in the same year.

Alternate Hypothesis : Crime does not change statistically enough between January to November in the same year.

I use the Homicide column in January 2014 to compare the Homicide column in November 2014 , After I run the hypothesis test, we see that both columns have a high level of similarity . therefore, with a 99.9% confidence interval, we can reject Null Hypothesis because the p-value is less than 0.05 and it states that the samples are similar.

```
from scipy.stats import ttest_ind
import numpy as np

Jan = pd.read_csv('Jan.csv')
Nov = pd.read_csv('Nov.csv')

data1 = Jan.iloc[:,1]
data2 = Nov.iloc[:,1]

stat, p = ttest_ind(data1, data2)

print(p)

if p>0.05:
    print('Similar')
else:
    print('Different')
```

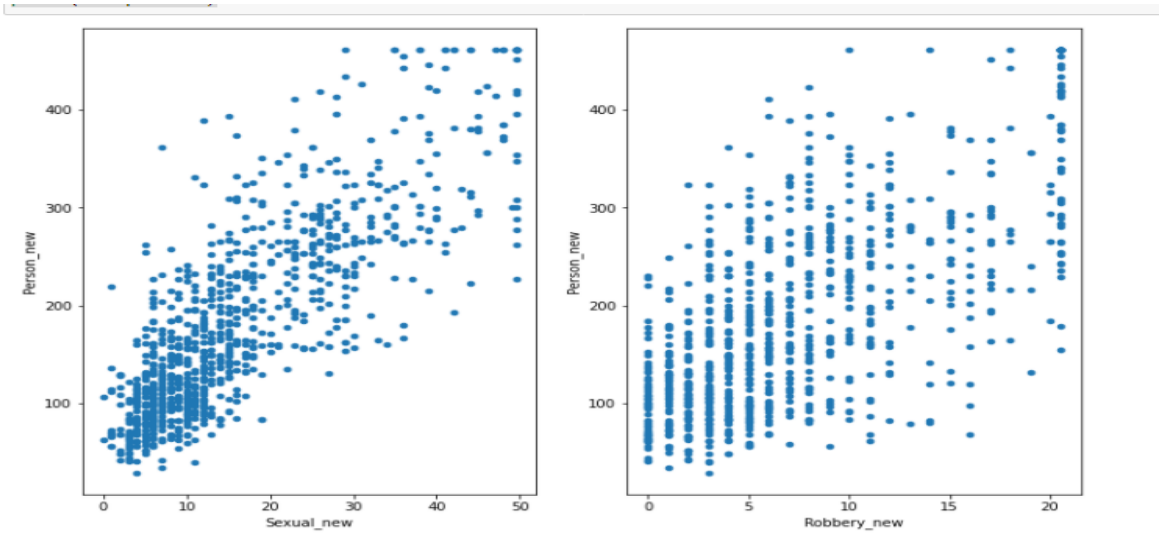
```
0.7906469982848555
Similar
```

REGRESSION

Linear Regression model:

For my Regression model, I made use of Linear regression. Firstly, I visualized using a scatter plot, to see the relationships between the independent variable X (Robbery) and dependent variable Y(Person) and another between Person(Y) and Sexual(X).

Fig 9



From the Figures above, there appears to be a positive relationship between Sexual and Person crimes. Meanwhile, there is no relationship between Robbery and Person crimes. I would then go ahead to generate the model. Given that there seems to be a correlation between sexual and person crimes, I would then go ahead to create a linear relationship, where my B1 is the sexual crime and from my model. I get the following results:

Intercept: 73.313321

Sexual_new: 6.780814

Where B*o* is 73.313321 and B1 is 6.780814.

This means that for every single increase in sexual crimes, the number of crimes against the person goes up by 6.780814. Therefore our linear regression equation becomes:

Yperson = 73.313321 + 6.780814Xsexual

This is important information on its own, I would then proceed to test the model by creating a new value. I check to see what the rate of Person crimes would be if sexual crimes is incre ased by 50 and this is what I get:

Sexual_new

50

412.354015

This means Person crimes would be 412.354015 for every 50 count increase in sexual crimes . This is prediction.

Advantages of Linear Regression Models:

1. It is easy to implement
2. It is efficient to train

Limitations of Linear Regression Model

1. It assumes that the relationship between the dependent and independent variables are linear.
2. It is susceptible to outliers
3. It is prone to noise

CLUSTERING

K-means Clustering:

For my k-means clustering Algorithm, I change my dataset to numbers, as the k-means algorithm works best with numerical values. I use numbers 1 to 40 to represent the forty counties in the dataset. I used the df.replace code.

For my Clustering I came up my own classification by visualising in order to have separate clusters , that is clusters that do not overlap.

Fig 10.i.

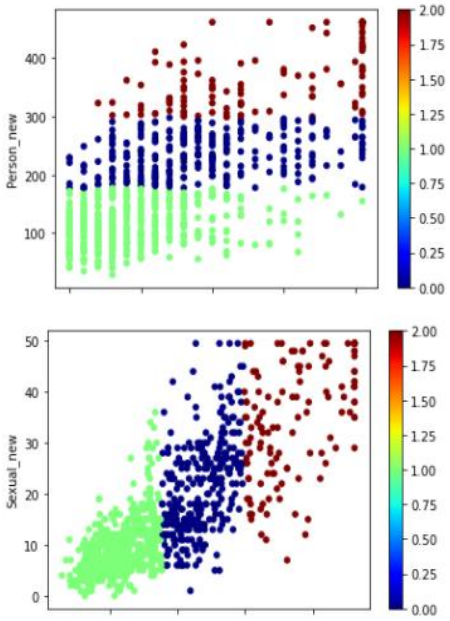
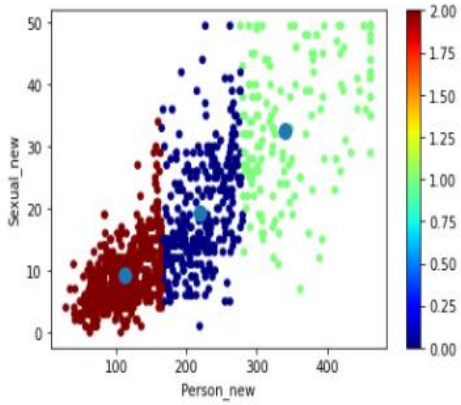
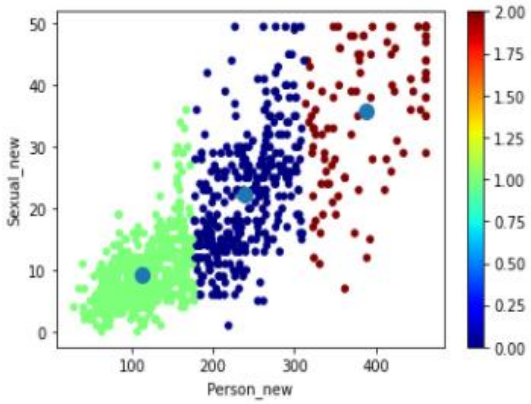


Fig 10.ii.



In the above K-means algorithm , I classified, using 1 iteration which is not enough/appropri ate as I have to run it multiple times in other to have the best fit for my centroids.

Fig 10.iii



In the above K-means algorithm, I classify using 80 iterations. That is I adjust the centroids 8 0 times unlike the first one.

From the visuals above, I have been able to classify the crimes in different ways, without any centroids, with 1 iteration and with 80 iterations. We can see that the green group, has its s exual crime rate going up when the person crime is increasing. For the blue group, the sexua l crime increases as the person crime increases and stops at a point. But for the Red group, t he rate of increase of the Person crime, does not affect the sexual crimes.

These classifications give an insight which will help with prescriptive analysis.

Advantages of K-means clustering

1. It is Simple
2. It is Efficient

Limitations of K-means clustering

1. I got different errors at different times, and that is due to its randomised approach w hereby it has to be run many times.
2. It does not do well with outliers
3. It is not suitable with all data types. It needs numerical data to be run.

CLASSIFICATION

Decision Tree:

For my decision tree classification, I tried to predict the (Y)county, by giving the values for (X)different crimes namely: Sexual_new, Person_new and Robbery_new. Below is a screensh ot of the classification and predicted result.

CLASSIFICATION

```
']:
import pandas as pd
import numpy as np
from sklearn import metrics
from sklearn.tree import DecisionTreeClassifier

data = pd.read_csv('Col_to_num.csv')

Y = df.iloc[:,1]
X = df.iloc[:,2:5]

dt = DecisionTreeClassifier(min_samples_split=3, random_state=99)
dt.fit(X, Y)

prediction_test = [217,16,7]
predicted_class = dt.predict(np.reshape(prediction_test,[1,-1]))
print("predicted class: ", predicted_class)
```

```
predicted class:  ['Humberside']
```

By Giving values 217, 16 and 7, I got a predicted County ‘Humberside’.

Advantages of Decision Tree Classifier

1. To a large extent, missing values in the data does not affect it.
2. The model is easy to explain

Disadvantages of Decision Tree Classifier

1. It is time intensive , the model requires a considerable amount of time to run.
2. A wrong decision, can result in a significant change in the decision making process.

There are other methods of classification, such as Naïve Bayes, KNN(K-Nearest Neighbours), Neural Networks and others.

Advantages of Naïve Bayes

1. It is not affected by outliers
2. It is highly efficient

Disadvantages of Naïve Bayes

1. It requires large data to run

Advantages of K-Nearest Neighbours

1. It is easy to explain and understand
2. It is highly accurate

Disadvantages of K-Nearest Neighbours

1. It requires a large memory to run
2. It is expensive to implement

Advantages of Neural Networks

1. It is highly tolerant of errors
2. It can perform multiple jobs at once

Disadvantages of Neural Networks

1. It does not explain in detail, its network functionality
2. There is no fixed time frame for the network

References

- Sunil, R 2013-2020, *A comprehensive guide to data exploration*, Analytics Vidhya, viewed 15 March 2021, <https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/>
- Yemulwar, S 2019, *Outlier treatment with python*, Analytics Vidhya, viewed 15 March 2021, <https://medium.com/analytics-vidhya/outlier-treatment-9bbe87384d02>
- Data to fish, 2021, *how to select rows from pandas dataframe,* Data to fish, viewed 1 April 2021, <https://datatofish.com/select-rows-pandas-dataframe/>
- Datacamp Team, 2020, *how to drop columns in pandas,* Datacamp, viewed 18 March 2021, <https://www.datacamp.com/community/tutorials/pandas-drop-column?utm_source=ad>
- Kumara, H 2020, *How to drop rows based on column values using pandas dataframe*, Medium, viewed 1 April 2020, https://medium.com/@harsz89/how-to-drop-rows-based-on-column-values-using-pandas-dataframe-38cf50e4c95a
- Rajabi, R(2019,August 15), *Using standard deviation in python*, Medium, viewed 30 ASpril 2021, <https://towardsdatascience.com/using-standard-deviation-in-python-77872c32ba9b>
- McLeod, S. A. (2019, July 19*). What does a box plot tell you?*, Simply psychology:, viewed 1 May 2021, <https://www.simplypsychology.org/boxplots.html>
- Nelson, D 2021, *Statistical Hypothesis Analysis in Python with ANOVAs, Chi-Square, and Pearson Correlation,* Stack Abuse, viewed 28 April 2021, https://stackabuse.com/statistical-hypothesis-analysis-in-python-with-anovas-chi-square-and-pearson-correlation/
- Minitab Express Support, 2019, *Interpret the key results for Histogram*, Minitab, viewed 29 April 2021, <https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/graphs/histogram/interpret-the-results/key-results/#:~:text=customer%20wait%20times.-,Symmetry,one%20side%20of%20the%20histogram.&text=The%20data%20in%20the%20following%20graph%20are%20left%2Dskewed.,the%20normal%20distribution%2C%20are%20symmetric.>
- Frost, J 2021, *Statistics by Jim*, Jim Frost, viewed 5 May 2021, <https://statisticsbyjim.com/basics/remove-outliers/#:~:text=Removing%20outliers%20is%20legitimate%20only,area%20and%20data%20collection%20process.&text=Outliers%20increase%20the%20variability%20in,results%20to%20become%20statistically%20significant.>
- Waseem, M (2019,December 10), *How To Implement Linear Regression for Machine Learning?,* Brain4ce Education Solutions Pvt, viewed 1 May 2021, <https://www.edureka.co/blog/linear-regression-for-machine-learning/>
- Glen, S (2015, July 9), *Bivariate analysis definition & example,* Statistics How To, viewed 30 April 2021, <https://www.statisticshowto.com/bivariate-analysis/>

- ASQ, 2021, *what is a scatter diagram?,* American Society for Quality, viewed 25 April 2021, https://asq.org/quality
- Excelr, 2021, *Mean Median Mode in data science,* Excelr, viewed 20 April 2021 , https://www.excelr.com/meanmedianmode-in-data-science#:~:text=Mean%20value%20of%20a%20dataset,in%20an%20ordered%20data%20set.
- Singh, D (2019 October 22), *Cleaning up data from outliers*, Plural sight, viewed 15 April 2021, https://www.pluralsight.com/guides/cleaning-up-data-from-outliers
- Dhiraj, K (2019 March 18), *top 5 advantages and disadvantages of decision tree Algorithm,* Medium, viewed 4 May 2021, https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a
- Vinita, 2019 August 3, *advantages and disadvantages of neural networks*, IntelliPaat, viewed 4 May 2021, <https://intellipaat.com/community/21886/advantages-and-disadvantages-of-neural-networks>

-