**NT Region’s Crime Data Statistics Analysis Report**

date: 29th of September, 2021

## Introduction

1Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime (result for assessment 2). With the increasing advent of computerized systems, crime data analysts can help Law enforcement officers speed up the process of solving crimes. 2 Using data analysis, we can extract previously unknown, useful information from unstructured data.

I apply the data provided by the NTG (Northern Territory Government of Australia Open Data). The reason for choosing this data is that there is no secret about the main issue here in NT: the continuously elevated crime rate compared with other states and territories in Australia. At this assessment stage, we focus on the crime data, which contains numerous raw data for us to analyze and figure out some hidden reason behind the annoying continuously elevated data. This set of data is a detailed analysis of each area, each type of crime, and the number of crimes in the Northern Territory in the past 10 years. These datasets are pretty informative, in particular, we live in the city of Darwin, Palmerston, as well as the Northern suburb known as the Great Darwin area,

## 2. Methodology

**2.1 Data Collection**

In data selection step we are collecting data from NTG , the local government open source data website <https://data.nt.gov.au/> . We download the newest data which upgraded couple of days ago and stored them into our work folder waited to be further process.

Also, in order to get the crime rate in the whole Northern Territory and each region, we have chosen the NT population data from another data source “Northern Territory Economy” website <https://nteconomy.nt.gov.au/population> to get the population of each region. This is based on the assumption that the population of each region from 2011 Aug to 2021 Jul is the same as the population as at Jun 2020.

We apply Excel, Vscode, Jupyter, and Python to achieve all the statistic analysis.

**2.2 Data sorting**

The basic datasets we can approach include CSV, xlsx. CSV format is convenient for Python to fetch but not that perceptual for humans to get information, so we use Excel to get a pivot table, in order to reformat our target columns of data.

Our data types include Categorical data and ratio data. Categorical data are mainly about Offence types: Assault, Actual Break-in, Attempted Break-in, Attempted Murder, Break-in, Commercial break-in, Domestic Violence, Driving Causing Death, Homicide and Related Offences, House break-in, Illegal Use of Motor Vehicle, Manslaughter, Motor Vehicle Theft and Related Offences, Non-Assaultive Sexual Offences, Northern Territory Balance, Offence, Theft and Related Offences, Property Damage, Robbery, Sexual Assault, Theft of Motor Vehicle Parts or Contents. Ratio data from the original dataset is the number of offenses recorded by Police in the specified month of the specified type in the specified area. For the purpose of sampling and data profiling, we extract the raw data to create our new datasets, including more ratio data like “the total number of offenses of each month each year in each region”, “the crime rate of each region in each month”, and “the number of offenses per ten thousand people in each region” for our project. 图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Picture 1 - the screenshot of the columns from our datasets

We add up all the crimes as title “all crime” and areas containing Darwin, Palmerston, Alice Spring, Nhulunbuy, Kathrine, Tennant Creek, and the rest of the area besides these main towns we classify into NT Balance.

We select 10 years of data from August 2011 to July 2021, sample 50 samples from 120 months, and use Excel form to calculate the sum, use Python programming skills to get average, variance, maximum, minimum, and IQR values and standard of the number of crimes in each region.

**2.3 Process and Theory**

Statistic packages we used: pandas, mathplotlib, stats models, numpy, packages,

Here we apply two methods to solve it:

1. use Excel filter function to get a sum of different crimes as “all crime” then use Python taped formula then we print out mean, mode, medium, etc.
2. use Python pivot\_table and group by function in the Pandas to extract the target data and use “aggfunc” to apply the sum function and mean function in the Numpy, and according to the pivot\_table, we can easily get mean, sum, of population.

Example: set up Value=” number of offenses” index=” offence type”,columns=[” year”, month number”] the result as follows:

文本

中度可信度描述已自动生成

The formula as follow:

Mean:

(1)

The Z-Statistic:

(2)

is the mean of the sample

is the population mean

is the population standard deviation

n is the sample size

The standard deviation of population/sample；

(3)

(4)

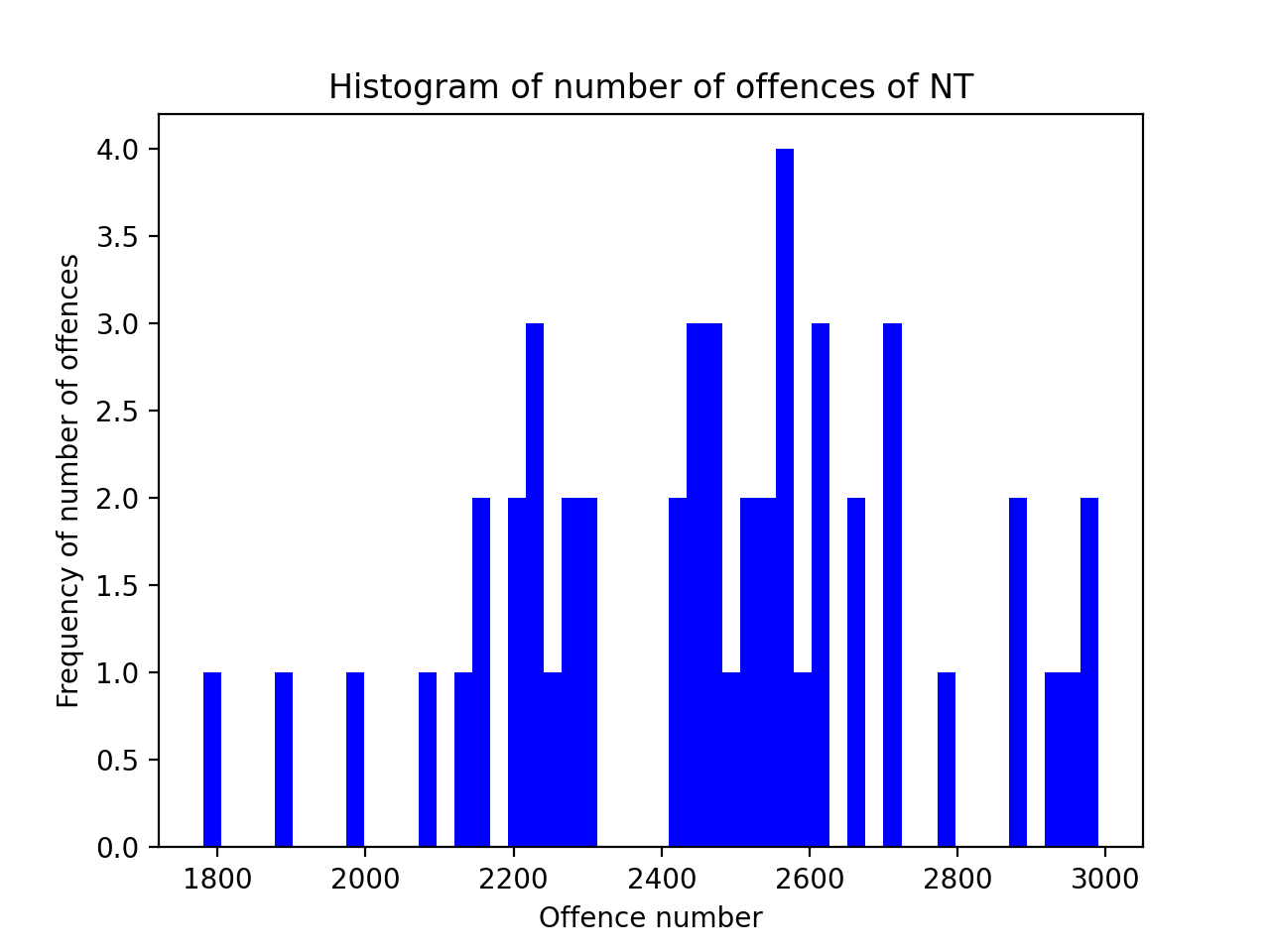
then use a random function to sample (or set up a sample fetch window), Each of us is in charge of two areas of data and gets 4 result forms, we merge 4 Excel/CSV data into one file(using Python) in order to process further data

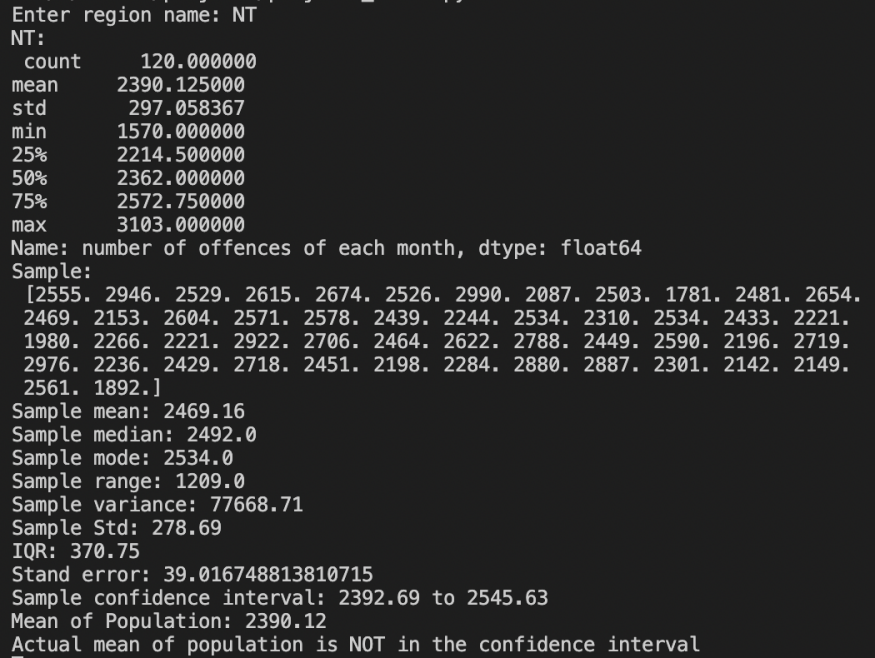
## Result

The total number of offences of of each month in each region is ratio data. All the result are based on the condition of randomly choosing 50 total offenses of each month in NT and the main region as samples to compute the statistics and plot the histogram. The main region includes Darwin, Alice Springs, Katherine, Nhulunbuy, Palmerston, Tennant Creek, and NT Balance.

**3.1 Histogram and python execute result analysis**

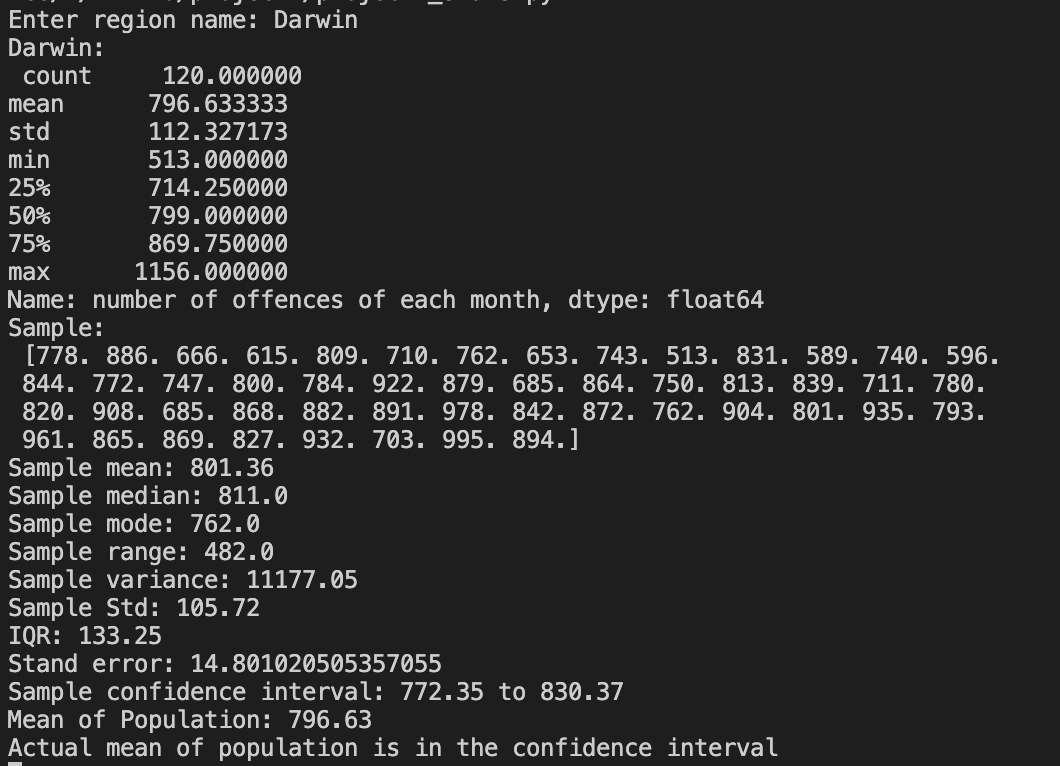
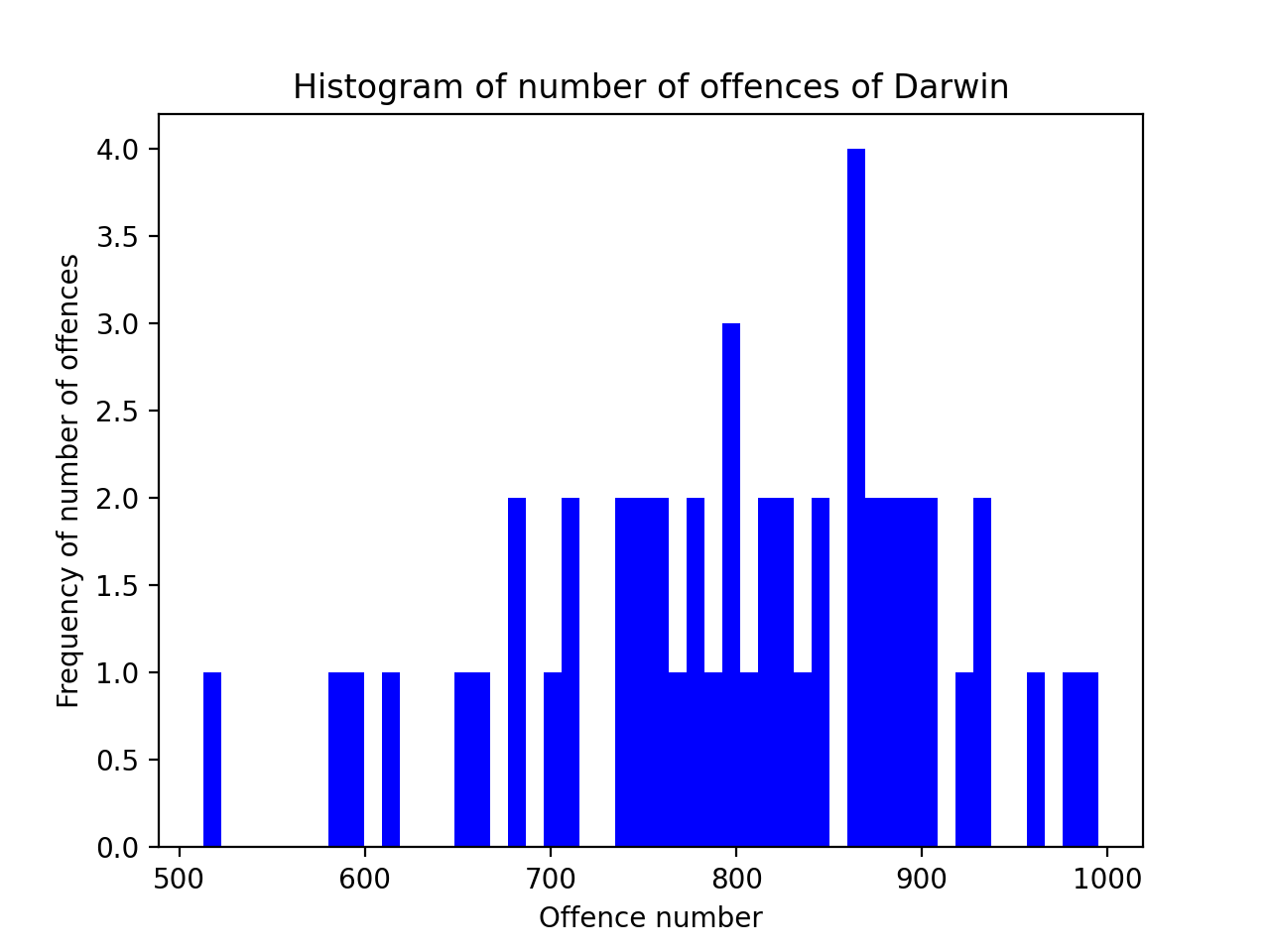
Figure 1. Histogram of number of offences of each month in NT





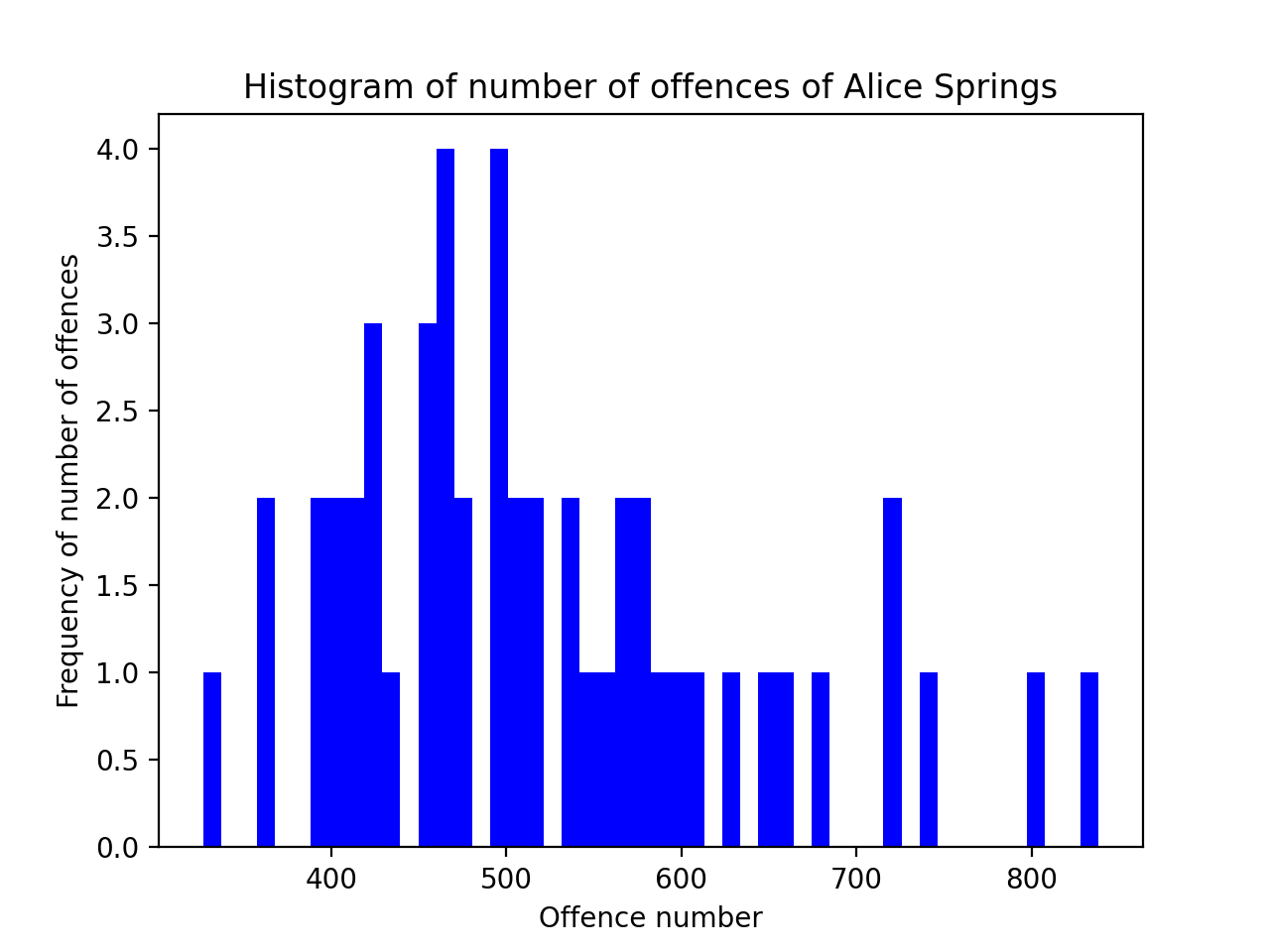
The figure of the NT crime is ratio data, and the frequency of several offenses is Discrete data due to the crime can’t happen 2.3, we can only tell 3 or 4. From the executed output, the spread of the data includes range 1209, IQR 370.75, variance 77668.71, and std 278.69. the population means of the 120-month offenses number is 2390.12, which is in the confidence interval between 2350.56 to 2500.32 based on the sample mean of 2425.44. From the histogram, the total number of offenses each month in NT represents a negative skew due to the sample size being only 50.

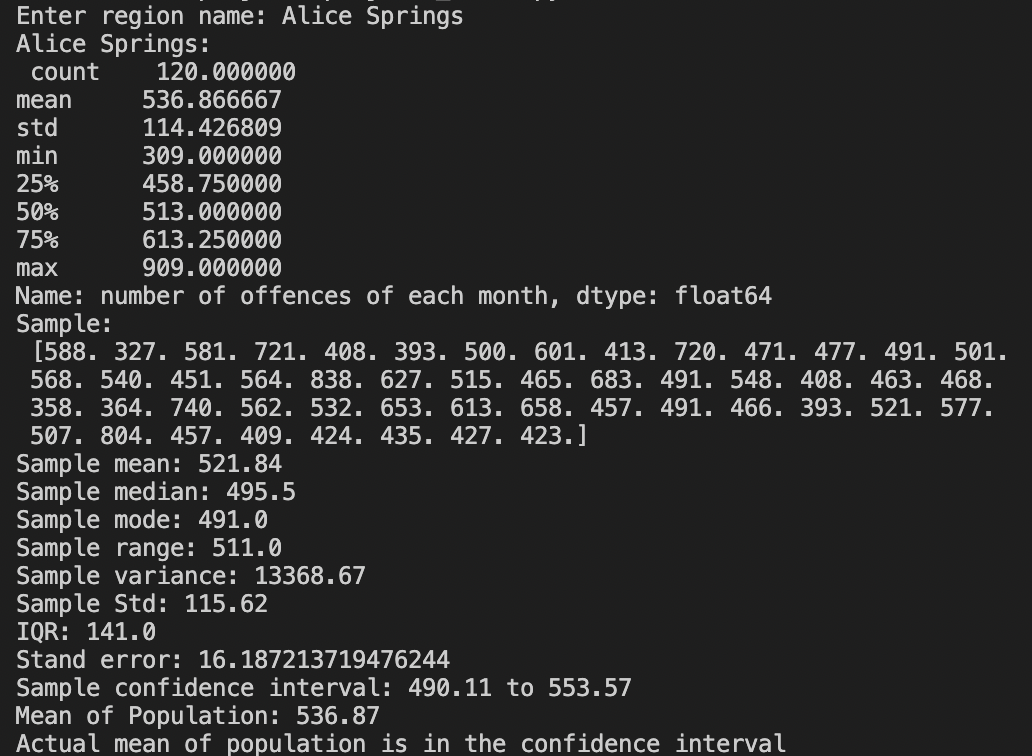
Figure 2. Histogram of the total number of offences each month in Darwin



From the Python executed output the population mean of 796.63 of 120 months offenses number is between the confidence interval between 772.35 to 830.37 based on the sample mean of 801.36. The centrality is better measured by the mean of the sample. The executed output shows the spread of the data includes range 482, IQR 133.25, variance 11177.05, and standard deviation of the sample is 105.72.

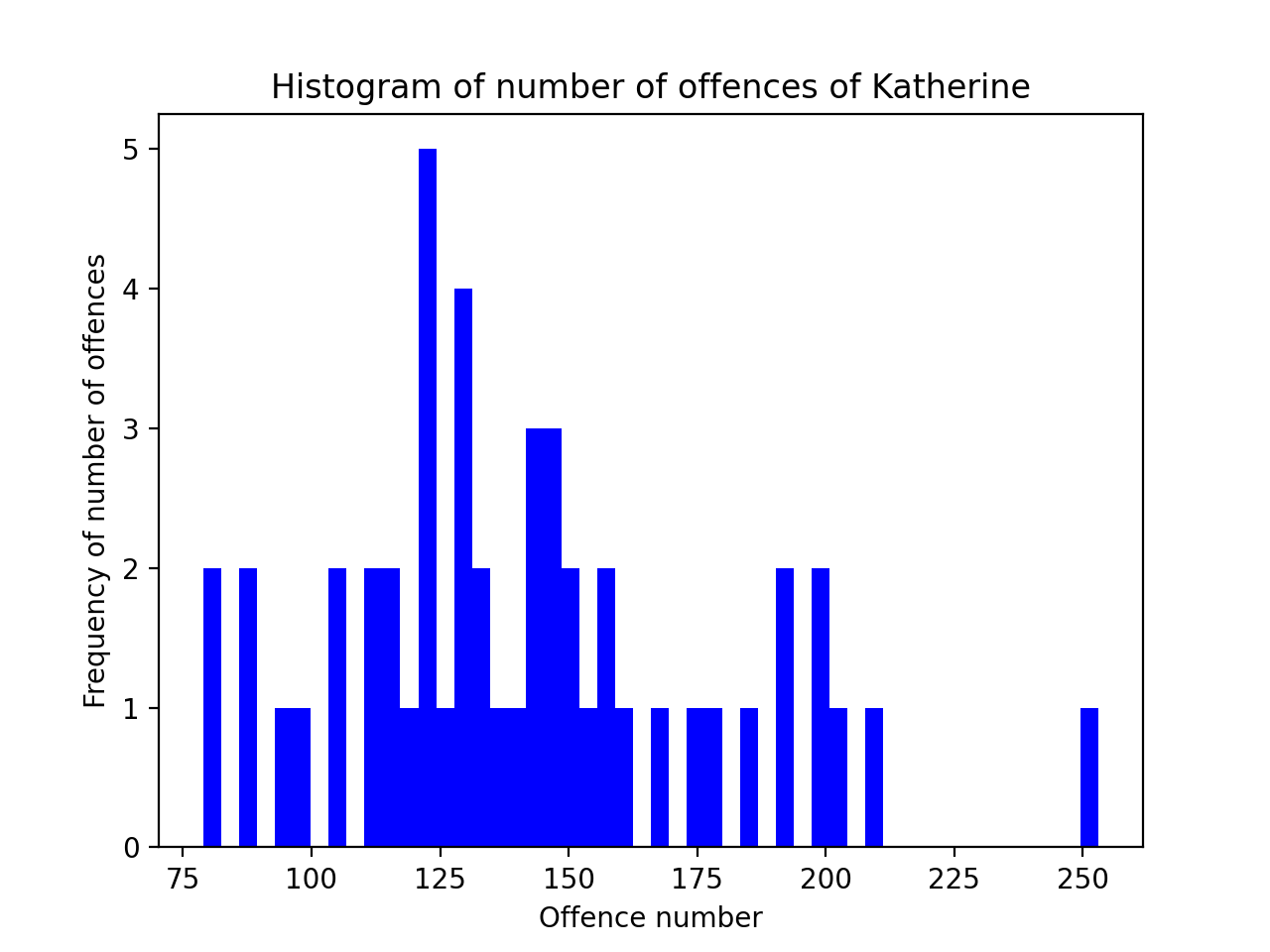
Figure 3. Histogram of total number of offences of each month in Alice Springs

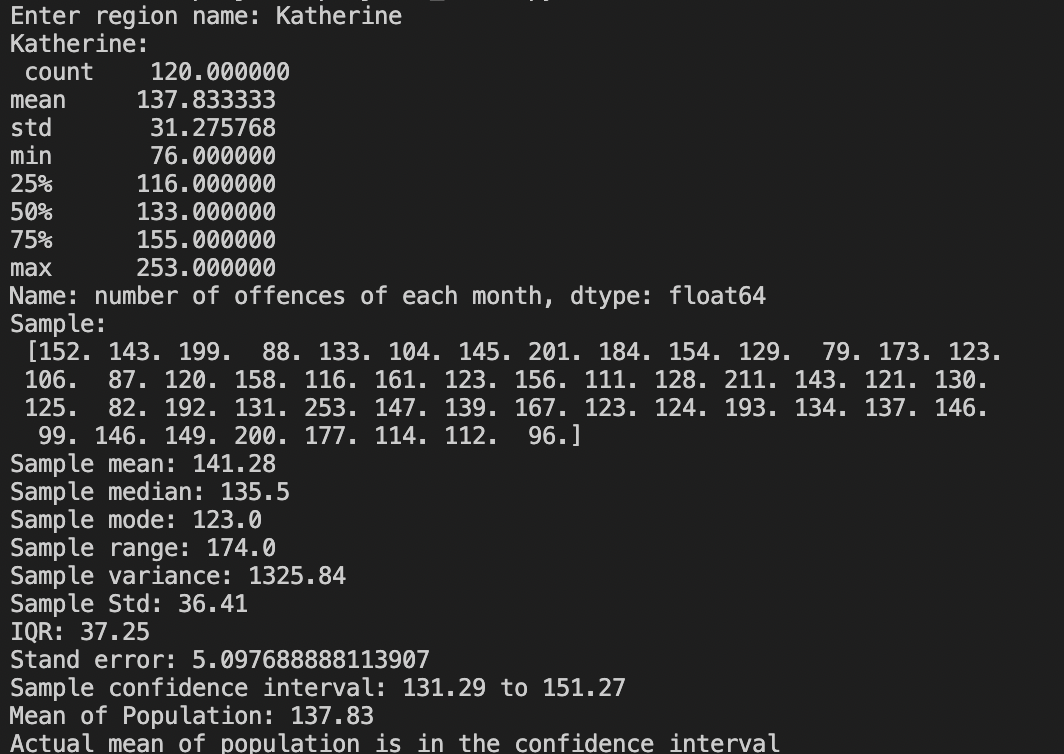




From the Python executed output, the population mean of 536.87 of 120 months offenses number is between the confidence interval between 490.11 to 553.57 based on the sample mean of 536.87. From the executed output, the spread of the data includes range 511, IQR 141, variance 13368.67, and standard deviation of sample 115.62. From the histogram, the total number of offenses each month in Alice Springs represents a Positive Skew distribution due to the sample size being only 50.

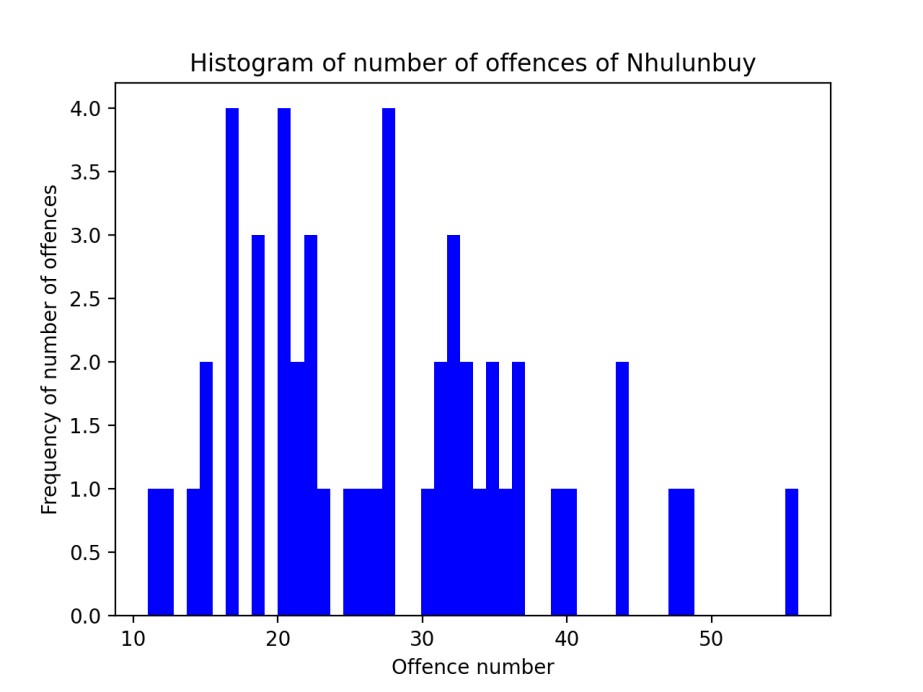
Figure 4. Histogram of the total number of offenses each month in Katherine

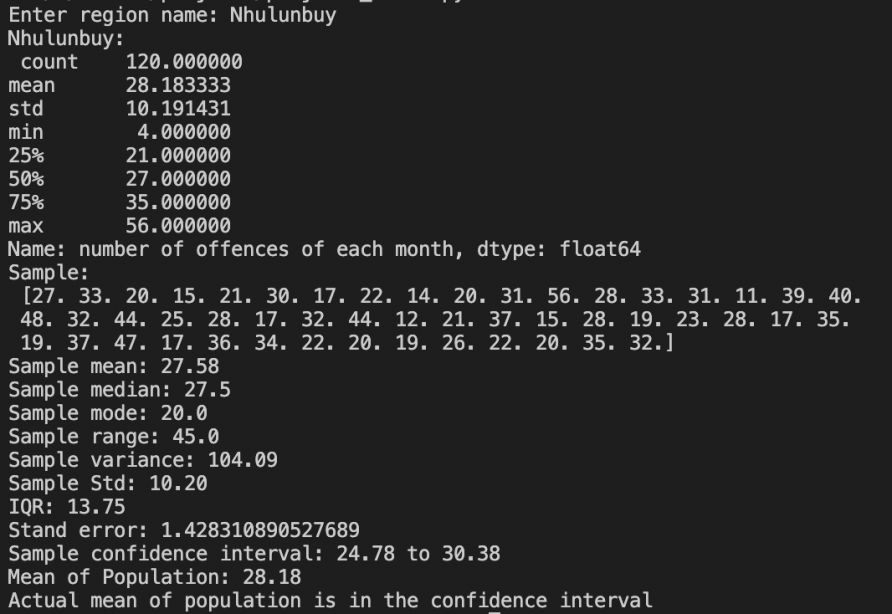




From the Python executed output the population mean of 137.83 of 120 months offences number is between the confidence interval between 131.29 to 151.27 based on the sample mean of 141.28. The executed output shows the spread of the sample includes range 174, IQR37.25, variance 1325.84, and std 36.41. From the histogram, the total number of offences each month in Katherine represents positive skew distribution due to the sample size being only 50.

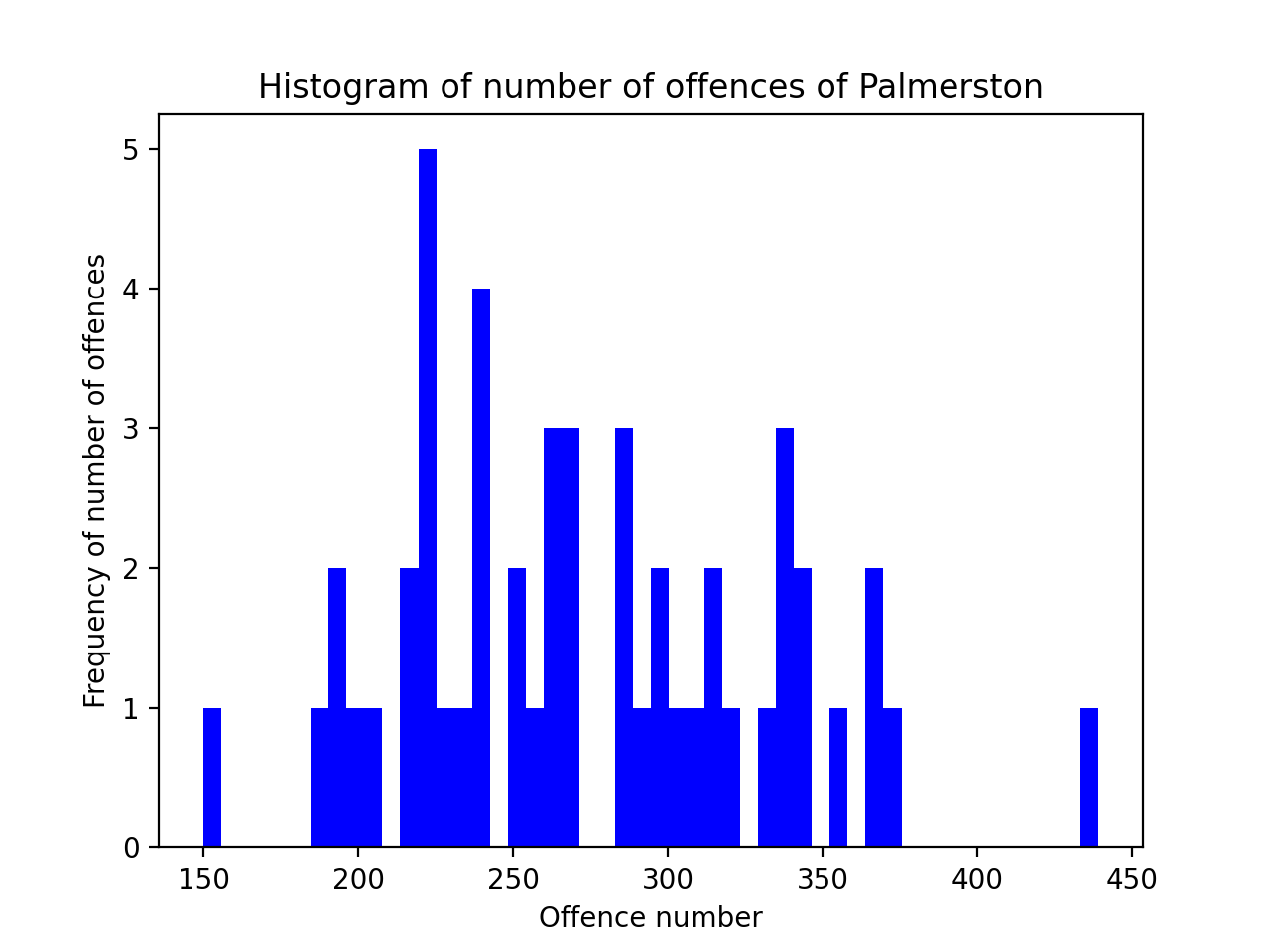
Figure 5. Histogram of the total number of offenses each month in Nhulunbuy

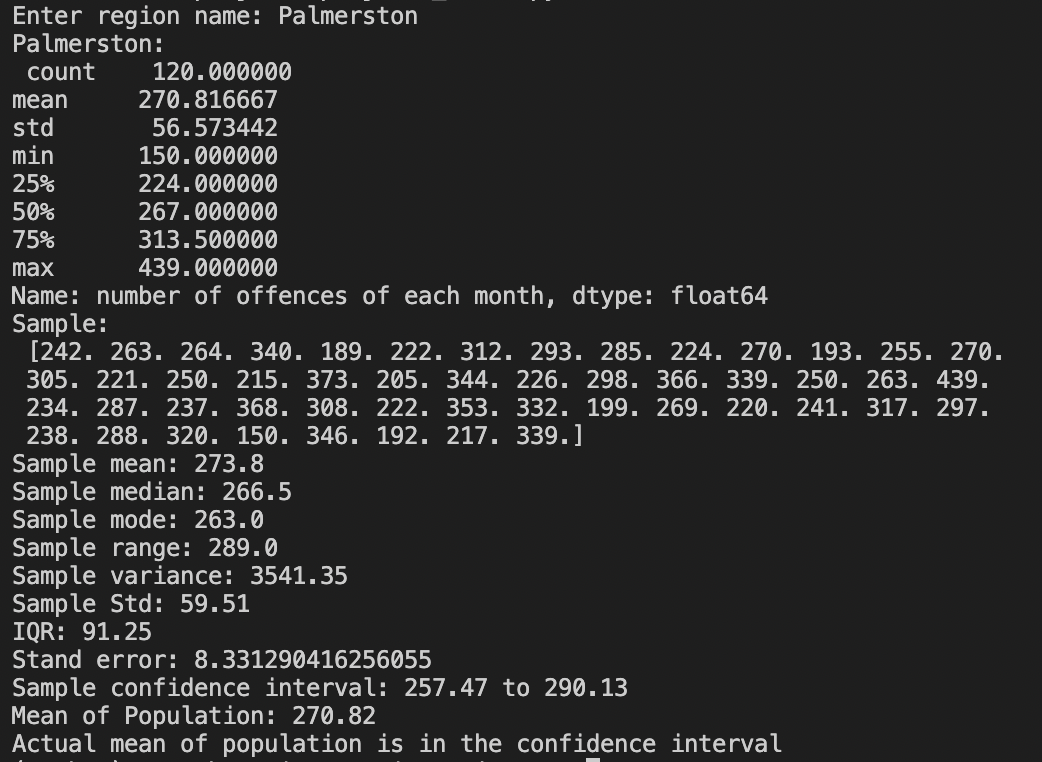




From the Python executed to output, the population mean of 28.18 of 120 months offenses number is between the confidence interval between 24.78 to 30.38 based on the sample mean of 27.58, the executed output shows the spread of the data includes range 45, IQR 13.75, variance 104.09 and standard deviation of sample 10.2. From the histogram, the total number of offenses each month in Nhulunbuy represents a symmetric distribution due to the sample size being only 50.

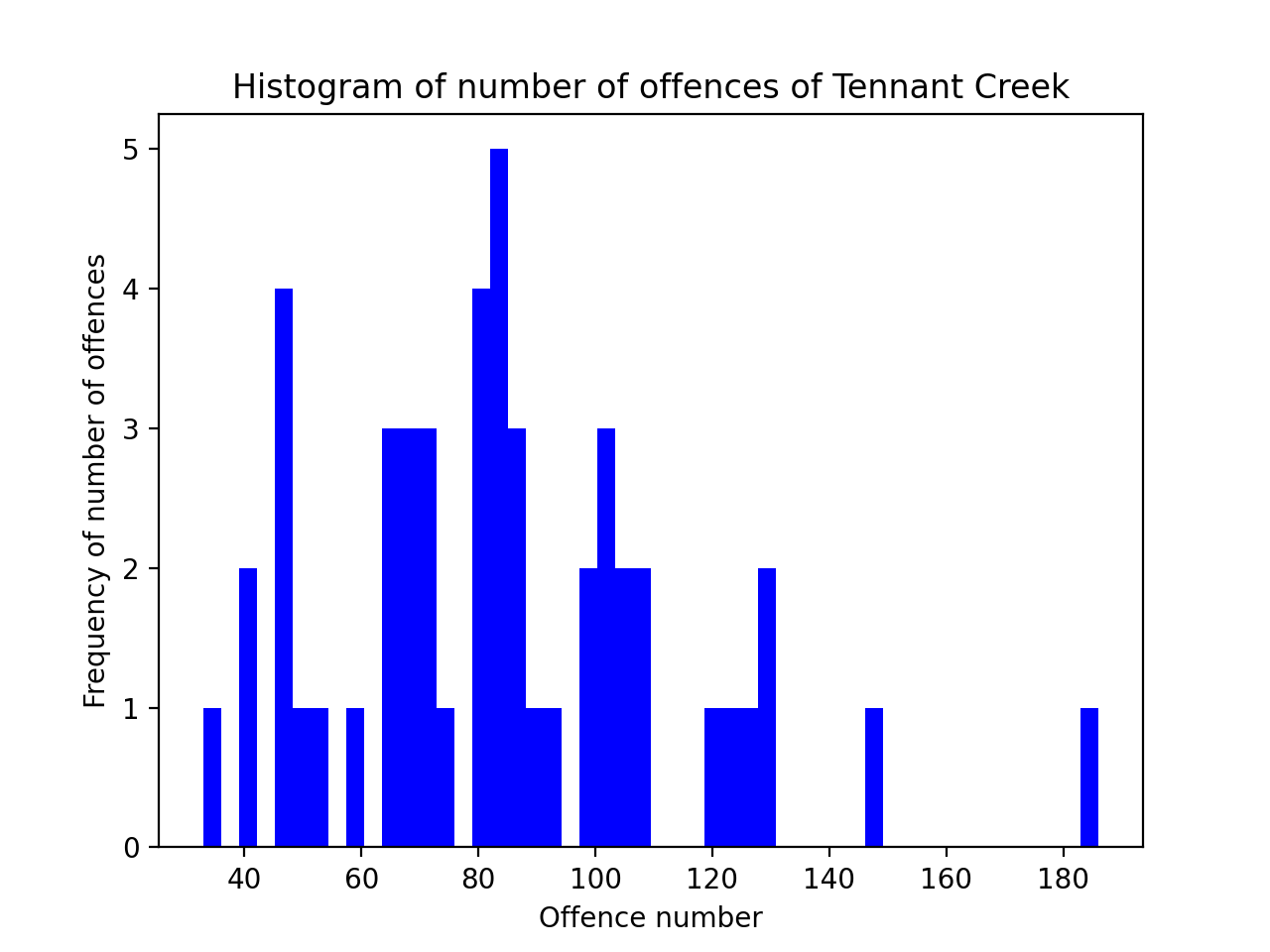
Figure 6. Histogram of the total number of offenses each month in Palmerston

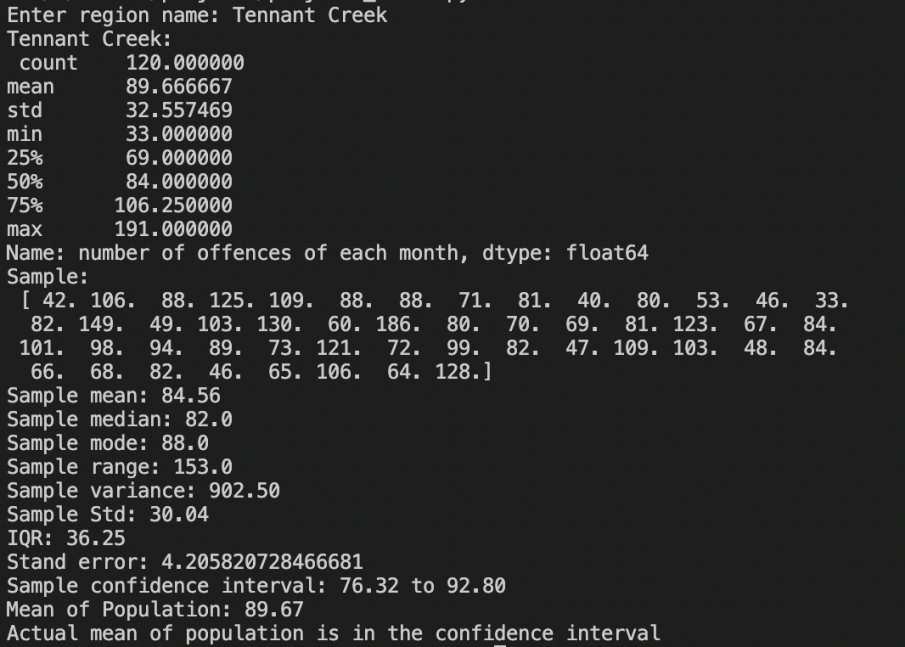




The Python executed output shows that the population means 270.87 of 120 months offenses number is between the confidence interval between 257.47 to 290.13 based on the sample mean of 273.8. The executed output displays the spread of the data, including range 289, IQR 91.25, variance 3541.35, and standard deviation of sample 59.51. From the histogram, the total number of offenses each month in Palmerston represents a positive skew distribution due to the sample size being only 50.

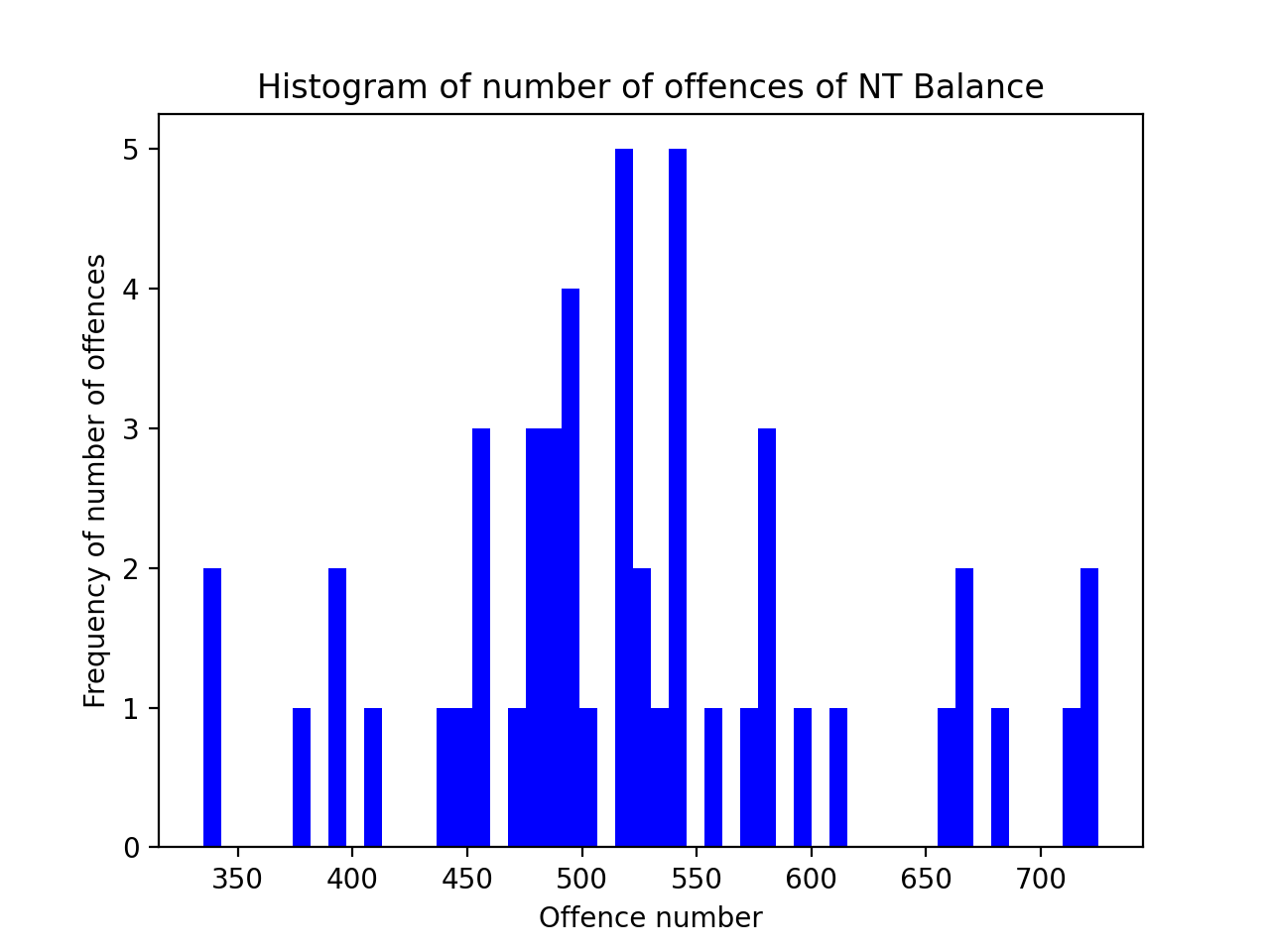
Figure 7. Histogram of the total number of offences each month in Tennant Creek

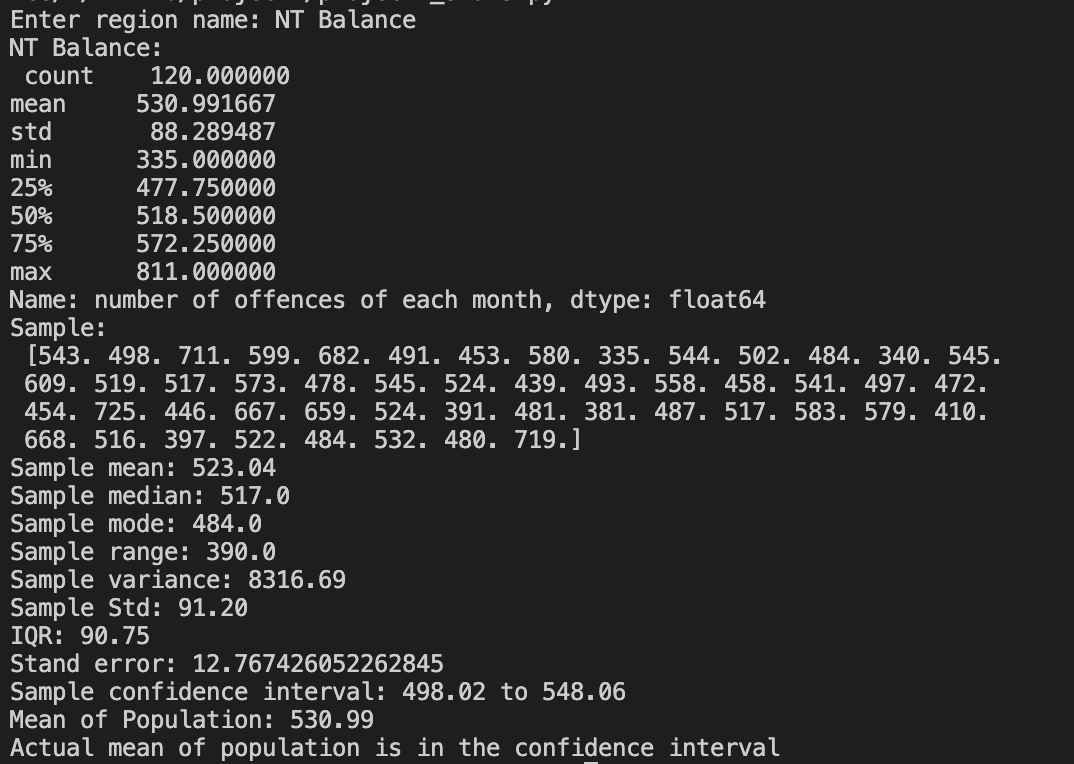




The Python executed output shows that the population means 89.67 of 120 months offenses number is between the confidence interval between 76.32 to 92.80 based on the sample mean of 84.56. From the executed output, the spread of the data includes range 153, IQR 36.25, variance 902.50, and std 30.04. From the histogram, the total number of offenses each month in Palmerston represents a positive skew distribution due to the sample size being only 50.

Figure 8. Histogram of the total number of offences each month in NT Balance





The Python executed output shows that the population means 530.99 of 120 months offenses number is between the confidence interval between 498.02 to 548.06 based on the sample mean of 523.04. The executed output indicates the spread of the data includes range 390, IQR 90.75, variance 390, and std 91.20. From the histogram, the total number of offenses each month in NT Balance represents a symmetric distribution due to the sample size being only 50.

Figure 9. Crime numbers per ten thousand pp in each region

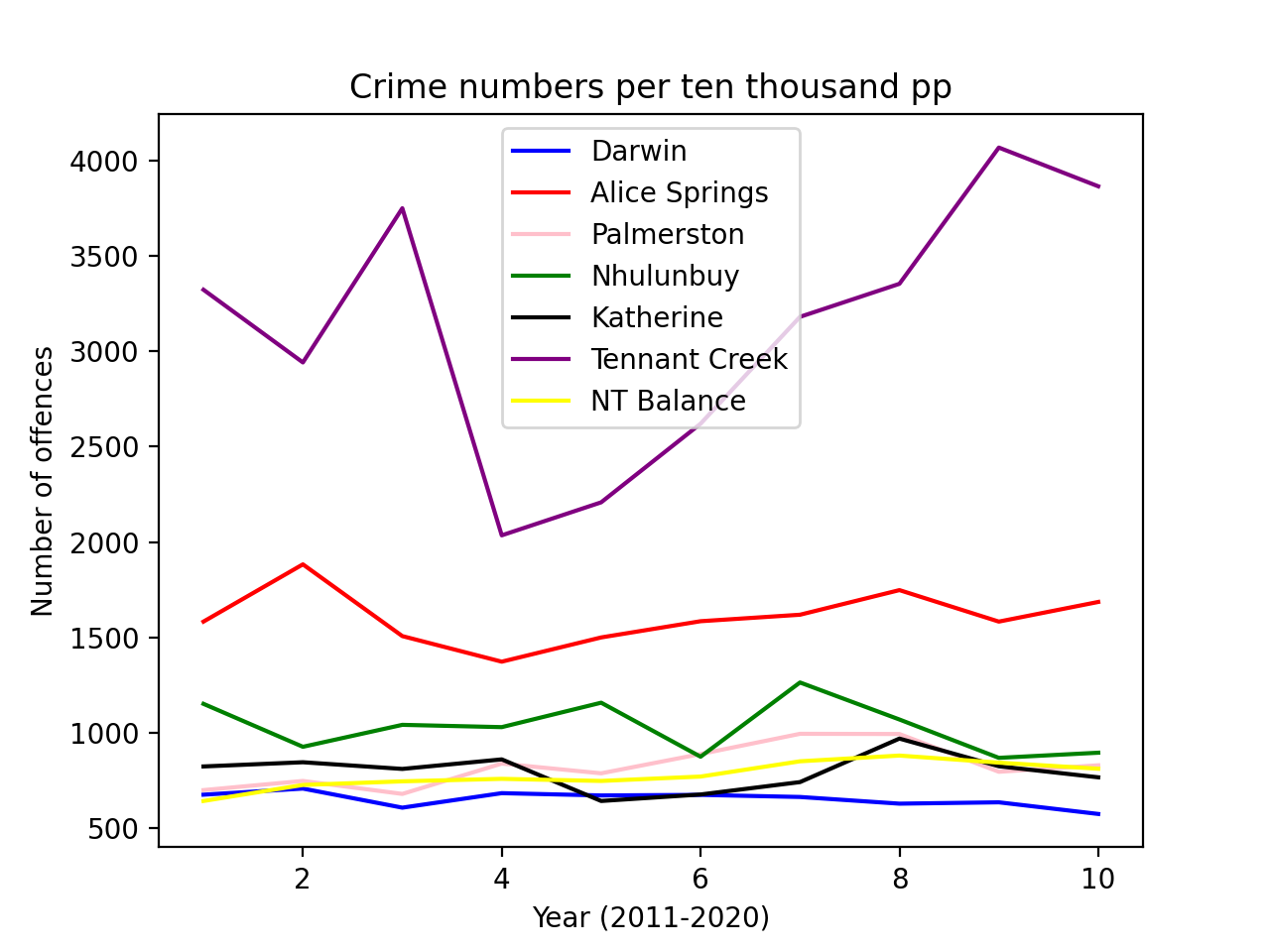


Figure 9 is aimed to show the crime trend in each region. The statistical data in the x-axis is the timeline by years and the y-axis shows the number of offenses of the year, we can see Tennant Creek has an incredibly high offense number than other areas, and then following by Alice Spring, other areas have quite a similar trend compared by the Tennant Creek and Alice Spring.

**3.2 Two sample T-test:**

We got data from 120 months alcohol alcohol-involved offenses and no-alcohol-involved offenses. One of the response variables was offense numbers.

The sample statistics are below:

|  |  |  |  |
| --- | --- | --- | --- |
| Condition | n | x\_bar | s |
| Alcohol involved | 120 | 329.57 | 53.77 |
| No alcohol involved | 120 | 176.03 | 94.54 |

The hypotheses can be stated as follows:

H0:1=μ2 (the average between al involve and no\_al involved no difference, which means alcohol has no relation to offenses numbers)

Ha: μ1≠μ2

We used Python to compute the p-value:

print(st.ttest\_ind\_from\_stats(x1\_bar,s1,120,x2\_bar,s2,120,equal\_var=False, alternative="two-sided"))

>>>>>>

t-statistic=15.4651275, p-value=2.19957204

Two-sample T\_test discussion: The p-value suggests that if the null hypothesis were true, then there is more than 5% chance of observing the difference between the means of the two populations, which means we cannot reject the null hypothesis. In other words, there is NO strong evidence to prove that the number of the average offense with alcohol involved does not differ from non-alcohol involved.

## 4. Conclusion and Evaluation

**Conclusion**

In this project, we mainly analyze the offenses number of each month in a different region, through descriptive and inferential statistics computing, we can tell the central tendency and dispersion of each sample data from each region. Also, we have concluded the crime numbers per ten thousand people in the past 10 years in different regions.

## 5. Question Propose

1. Number of crimes whether is related to alcohol involvement?

2. Based on the number of crimes in the past 10 years, is it possible to predict the number of crimes in each region in the future?

## Reference

[1]Shiju Sathyadevan, Devan M.S. Surya Gangadharan. S. Crime Analysis and Prediction Using Data Mining.Conference Paper. August 2014.

[2]Malathi. A .Dr. S. Santhosh Baboo.An Enhanced Algorithm to Predict a Future Crime using Data Mining. *International Journal of Computer Applications (0975 – 8887)*

*Volume 21– No.1, May 2011*