**List of contents:**

1. **Business Question…………………………………………………………………………….3**
   1. **Selection of independent and dependent variables………………….3**
   2. **Relationship Diagram……………………………………………………………….3**
   3. **Design of Hypothesis……………………………………………………………….4**

**2.2 Task 2.2 Data preparation………………………………………………………………….8**

**2.3. Task 2.3 Statistical testing………………………………………………………………….15**

**2.4 Deployment considerations……………………………………………………………….24**

**References………………………………………………………………………………………………26**

**Appendix A- Data Visualizations**

**Appendix B- Sample Python Code**

**1.0 Business Question:**

The main aim of the study is to predict the crime rate in Chicago and to understand the factors which are driving crime rate.

Data set used in this case study was taken from the city of Chicago Open data by referring to the government of Chicago website. This data set records crime data in the city of Chicago between 2001 to January 2022.

For the following case study, we have used victim-based crime dataset to predict the number of crimes and the reasons driving the crime rate.

* 1. **Selection of independent and dependent variables from the collected datasets.**

By analysing the dataset, we have noticed that crime is dependent on various factors. And then we can classify them as either dependent or independent variables. In our case study, we would like to measure the crime rate, or the number of crimes occurred in Chicago USA, which is depending on the different features like location, Arrest status etc., So, we can conclude crime rate as our dependent feature and rest of the features as independent variables. **(Definition and Examples of Independent and Dependent Variables, 2022)**

From the available features, we have considered the below variables for the case study, along with the crime data set, income and population statistics have been considered and integrated with the existing data set to check if the income range and population has any impact for driving the crime rate.

* **Dependent(Target variable) : Crime rate(ID\_COUNT)**

**Table 1 Showing the list of independent variables**

|  |  |
| --- | --- |
| Variable Name | Variable Content |
| Primary Type | Description of the type of crime committed. |
| Arrest | This variable indicates whether the arrest was made. Carrying Boolean values.(TRUE or FALSE) |
| Community Area | Community area can be considered as the important factor in determining the crime rate. By analysing the crime rate across different districts, we can possibly understand the districts with higher crime rate and act accordingly to reduce the crime count for those districts. |
| Income Level | Higher the income for the region will probably reduce the crime as most the people can be considered as educated and financially well balanced thereby people are not motivated to not commit any crimes. |
| Year | The year the crime occurred |
| Population |  |
| House price |  |

**Location** Where the crime occurred.

From the table 1, in the crime\_data.csv dataset, we will use these independent features and understand their impact on driving the crime rate.

All these features have a prominent role in understanding the crime impact in Chicago, so we have considered these features and dropped rest of the features for determining factors which are driving crime rate in Chicago.

**1.2 Diagram showing relationships among the identified variables.**

**Diagram

Description automatically generated**

**Figure 1 Representing the relationship between the chosen independent and dependent variables**

A flow diagram is used to represent the identified dependent and independent variables from the dataset. Raw data is processed, and then dependent and independent features are identified. Using simple linear regression algorithm, we will try to predict the number of crimes and results will be discussed. And for the crime analysis we can use Python seaborn, matplotlib, and other plotting libraries to develop different charts like bar, pie, scatter, and data driven decisions can be made to control the crime rate.

A good understanding of the relationships that may exist between different variables can be achieved by creating a graphical representation that includes these variables. Scatterplot, bubble chart, heatmaps are the different ways to represent a correlation matrix ( Dependencies) between the variables in the dataset.

The statistical relationship between two variables is known as correlation. A positive correlation indicates that both variables move in the same direction, whereas a negative correlation indicates that as the value of one variable rises, the value of the other variable declines. (Correlation, 2022)

Below is the table representing the dependency of crime rate (ID\_COUNT) on different independent variables.

Table

Description automatically generated

We can notice that there is strong positive relationship between the dependent variable ID\_COUNT**(crime rate)** to population and relatively less dependency on House price.

let us consider the independent variables population, house price to predict our target variable crime rate.

Below figure showing the heatmap for the chosen dependent and the independent variables after dropping the other variables which have relatively less dependency on the target variable.

Shape, background pattern, square

Description automatically generated

**Figure 2 Heat MAP Showing the dependencies for the identified variables**

In here, we can notice that there is high dependency of crime rate( ID\_COUNT) on the variable **population**.

* 1. **Design of hypothesis.**

The fact that the Increase in crime rate is related to the growth in population is well established. However, this relationship is less clear. A hypothesis can be formulated to prove or disprove our assumption in relation to our business question to predict the crime rate.

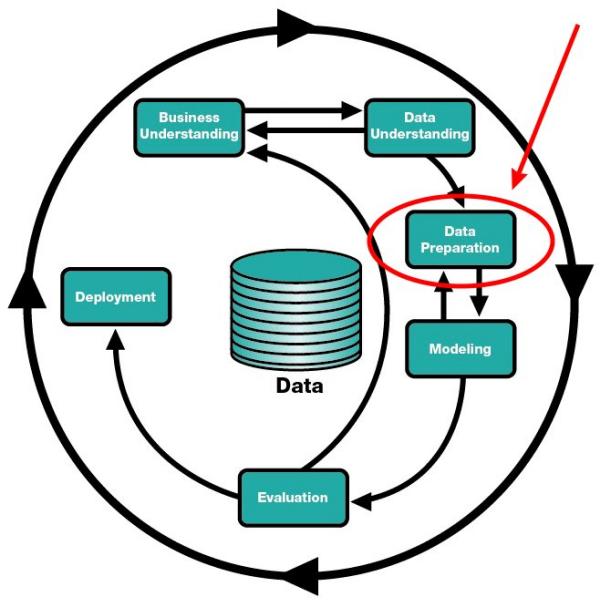
**Null hypothesis H0:**  The rate of crime and the number of people in a city are independent.

**Alternate hypothesis H1:** The rate of crime and population growth are related.

Based on the significance level **p** we can either accept or reject the null hypothesis.

If P value is less than 0.05 it is statistically significant. Indicating strong evidence against null hypothesis, as the probability is less than 5 % for the null to be correct. Further, detailed discussion of test results is made in section 2.3

**Task 2.2 Data preparation**

**2.2.1 Data Preparation:** Before data can be analyzed, it must first be organized. Data preparation is an iterative process of transforming unstructured and noisy raw data into a more organised and usable form that is ready for further analysis. (Abdallah and Webb, 2022) In here, we will be discussing the data preparation process by using Python (Pandas- Data Manipulation and Analysis library) 

**Figure 3 Showing data preparation process in crisp DM lifecycle (Source: Author: (Mayo, 2022))**

**2.2.2 Data Cleaning:**

As soon as we load the data. The first step and probably the biggest data cleaning task is to check on the missing values. Most of the datasets have missing values and the probability increases with the size of the data set. In here, the dataset considered for the analysis has over 6 million rows and 22 columns.

Let’s draw a bar plot to check the missing values for each column.

A picture containing icon

Description automatically generated

**Figure 4 Showing bar Plot for number of missing values**

From the above bar chart, we can notice that there are lot of missing values in the location-based information, ward, and community area. Let’s understand it more accurately by calculating the missing value percentage of all the features.

Below are the steps followed in data cleaning:

1. Calculated the missing value percentage for all the columns in the dataset.

Graphical user interface, text, application

Description automatically generated

We have noticed that that nearly 90 % of the location-based data is missing.

1. Dropped variables(Features) which are not useful for the analysis. Using drop method. As we can notice from bar chart and from the percentage of missing values. As, there are lot of missing values present in location relevant data like location, X and Y coordinates. Since these features are not direct numeric values, we can’t use summary Statistical functions to fill in the missing values. Hence, we are dropping these features from the data set. Graphical user interface, text, application

   Description automatically generated
2. Label encoder is used to convert category characteristics to numeric values. The date attribute is splitted into new attributes like month, hour, day, day of the week, weekday, and day period. (Categorical encoding using Label-Encoding and One-Hot-Encoder, 2022)

Text

Description automatically generated with medium confidence

1. By analysing the type of missing data. Missing values are treated using **fillna** function. A forward filling approach (method=**ffill**) is used to fill the missing values.

Graphical user interface, text, application

Description automatically generated

Inconsistent data issues like **Capitalization** whichisobserved especially in the categorical variables is removed as python is case sensitive.

**2.2.3 Data Filtration:**

Data is filtered from the year 2001 to 2015 as we have noticed that in the integrated new population data set consists of data from 2001 to 2015. Hence crime data is filtered to these specific years to balance both the datasets and to avoid missing values.

Graphical user interface, text, application, email

Description automatically generated

**2.2.4 Data Transformation:**

Data Transformation has been achieved by using the groupBy function and then using merge to recombine it to the original dataset.

Graphical user interface, text, application, email

Description automatically generated

In the above screenshot we have used groupby on feature **year** and **ID\_COUNT**(crime count for the year) is later stored to a data frame Crime rate and finally we have merged it to the Population dataset as shown below.

Graphical user interface

Description automatically generated

In the final dataset –(**Merged\_crime\_population)**obtained we are left with the features Year, Population, House price and ID\_COUNT

**2.2.5 Data Integration:**

After the cleaning, transformation the next step in analysis is combing of two different datasets.

Different datasets are integrated to predict the crime rate for the following study. **Income** and **population** are the important features in predicting the crime rate as these features tend to have a lot of impact on driving the crime rate.

2 different data frames have been pre-processed, identical features are identified and merged using merge function by the following steps

1. Population data set is loaded by using the head method first 5 rows are printed.

Table

Description automatically generated

1. The next step it to load the dataset crime\_rate and displayed the top 5 rows as shown below. Other features from this data set have been dropped as we haven’t noticed any strong relationship with the other variables.

Table

Description automatically generated

As now, we can see that **Year** is the common feature in both the datasets. Data integration has been performed on both the Data Frames with respect to the **Year column** using the **pd.merge()** function, and then we have printed the first few rows of the data frame.

Graphical user interface, text, table

Description automatically generated with medium confidence

In the above screenshot we can see that data of the population\_data set is merged with the data of the crime\_rate dataset.

**Task 2.3.0 Data analysis**

* + 1. **Identification and justification of a statistical test.**

Statistical tests can be used to test our hypothesis and we can determine whether a predictor(population) variable has a statistically significant relationship with an outcome variable(Crime\_Rate)

Below is the hypothesis which we have formulated for our assumption that crime rate is related to population.

**Null hypothesis H0:**  **Population** and **Crime rate** are not related to each other.

**Alternate hypothesis H1:** **Crime rate** and **population** are related to each other.

And then we can either accept or reject our null hypothesis based on the significance level **α.**

A simple flow chart can be made to understand the type of statistical testing which can be used to prove or disprove our hypothesis.

In here, both dependent and independent variable are quantitative variables and, we have only one predictor variable. As we are considering only the effect of population in driving the crime rate, we can use simple regression for our model to predict the outcomes as explained in the figure5.

Diagram

Description automatically generated

**Figure 5 Flow chart showing steps to choose correct statistical test(Source: How to identify correct statistical test)**

To justify the statistical test for our case study, the best way to construct a scatter plot and observe the data distribution. Using **plotly** library a scatter plot has been generated to identify the relation between independent variable **Population** and Dependent variable **ID\_COUNT**.

Chart, scatter chart

Description automatically generated

Figure 6 Scatter plot showing distribution of Population and Crime count

**Observations from the Scatter graph:**

* Dots represent the true data and line represents the regression model(best fit line).
* A clear linear relationship can be noticed between population and ID\_COUNT(Crimerate)
* As the data is exhibiting a linear trend. Linear regression would be apt to explain the general impact of population on ID\_COUNT(Crimerate).
  + 1. **Application of this test to the integrated dataset**

We are using simple linear regression algorithm to test if the population has any effect on crime rate.

**P** value is nearly equal to **zero**, so we can reject of null hypothesis and accept our alternate hypothesis. (McLeod, 2022) Thereby, we can conclude that crime rate is dependent on the population. Further detailed explanation of the test out comes is done in the section **2.3.3**

Table

Description automatically generated with low confidence

Figure 7 linear regression results for target variable population

Now, Let consider the independent variable House\_price and target variable Id\_count.

A picture containing text, indoor, screenshot

Description automatically generated

Figure 8 Scatter plot showing distribution of House price and Crime Count

We can see the inconsistent distribution of data in the above plot. Let us consider the variable house price along with the population and try to check if it makes any difference to our model performance. In here we have two independent variable, Population, and house price for our dependent variable Crime rate. So, we will be using **multiple linear regression.**

Table

Description automatically generated

**Figure 9 Multiple linear regression results**

We notice the p value(the significance level **α** )for house price is very high **(0.683)** much greater than 0.05. So, the variable house price is no where related to the crime rate. Lets discuss the test results in detail in the following section and evaluate our simple linear regression model performance.

* + 1. **Discussion of Test Outcomes and implication for answering business question.**

1. **Linear Regression**

Constant Intercept value **(C)** = **-1,826e+06**

Depending on the dependent(ID\_COUNT)variable, A negative constant/intercept value should not be a reason for concern. When all independent/predictor variables are set to 0, the anticipated value for the dependent variable will be less than 0 (Gassoumis, 2022)

Typically, in a linear regression model, it is usually the overall relationships between the variables, not the value of the constant, that are most important.

Independent variable intercept value**(X) = 0.7919**

The fitted regression model is ID\_COUNT(Crime\_rate) = -**1,826e+06+0.791\*Population**

We have an extremely low p-value when we look at both coefficients (although it is probably not exactly 0). This indicates that these coefficients and the target have a significant relationship.

In a regression model, R square is a statistical measure that indicates the proportion of a dependent variable's variation that is explained by an independent variable. (How to Interpret R-)squared and Goodness-of-Fit in Regression Analysis - DataScienceCentral.com, 2022)

Then, looking at the R² value, we have **0.79**. Therefore, **about 79 % of the variability of crime rate is explained by the variable population**.

Here the p-value is less than **0.05**, so we can reject our null hypothesis in our case and accept the alternate hypothesis making our assumption true in this case. i.e., population have much effect on driving the crime rate.

1. **Multiple Linear Regression**

R square value is 0.76 which is relatively less when compared to the simple linear regression.

Observing p-values we can notice that for the second variable house price it is relatively very high when compared to other independent variable population. Therefore, we can conclude that rise in house prices is not statistically significant to crime rate.

We can then derive the equation of the best fit line as follows:

**Crime\_rate** **=** -**1,826e+06+0.7759\*Population+177.7889\*house price**

Sample out put showing the Predicted crime rate Vs Actual crime rate for the given population.

Graphical user interface, application

Description automatically generated

Lets check the Mean Absolute Percentage Error(MAPE) for our linear regression model. By using this metric we can meaure the forecasting accuracy of our model to predict the number of crimes.

Graphical user interface, text, application

Description automatically generated

We have noticed the MAPE value as 5.85 %. So, we can conclude that on an average the difference between the Aactual crime rate and the predicted crime rate is 5.85 percentage. Which makes the oveall model performance as 94 %.

Finally, we can evaluate our model performance of crime prediction using linear regression algorithm is as follows:

Overall Accuray : 94 %, Mean Absolute Percentage Error: 5.85 %, R square: 0.797, F Statistic : 51

* + 1. **Acknowledgement of limitations and assumptions of the presented data model.**
* Decision to use linear regression for this model is mostly on the assumption of a linear relation between the independent variable **population** and **dependent** variable crime\_rate.
* To check if out assumption is met or not. We have created a scatter plot**(Figure:6)** and observed that most of the data points in the plot followed the trend line. Indicating that existence of a linear relationship between our variables.
* Earlier, in the analysis process different features like income\_level, population, Arrest rate, primary type, location description were considered to check their impact on the growth in crime rate. As, we have defined our hypothesis on population feature, rest of the features have been dropped from the data set by considering only the population variable as we have noticed a strong linear relation between the variable population and crime rate. This might have probably increased the model performance but impacting its applications in real world.
* Present model works for only Chicago, and it is only confined to population feature. Adding more features might increase or decrease model performance based on the feature importance with respect to crime rate.
  + 1. **Potential enhancement for more accurate analysis**

The present model is only applicable to check the effect of crime rate with only limited features. With the inclusion of additional features like income range, Location description, location, Arrest rate which aren’t included in this study, overall model performance can be increased or decreased based on the feature importance. Along with that the present model can be further enhanced to predict the crime type(Type of crime which can occur at a given location and time) by using like primary type and location features. We can identify the hotspots where a crime is likely to occur in the future, thus we can enable police units to increase a surveillance in these areas. This can be achieved by implementing several classification algorithms like Random Forest, Naïve bayes classifier, Support vector, K-nearest neighbours, and Logistic regression (Kumar and Sahni, 2022). Model performance can be evaluated on all these algorithms and the best model can be chosen for more accurate results.

To check the extent of linearity of the data. We have also checked the normality of gaussian distribution. It is a symmetric probability distribution about the mean, indicating that data near the mean occur more frequently than data distant from the mean. The normal distribution will show as a bell curve on a graph. **(Improve Linear Regression Using Statistics, 2022)**

Chart, histogram

Description automatically generated

**Figure 10 Showing skewness of the distribution.**

The brown line in the graph above depicts the gaussian distribution we wish to achieve, while the blue line represents the kernel density estimation (KDE) of the provided data after transformation. For more accurate analysis we can apply logarithmic, exponential transformations and skewness can be fixed which can contribute to boost the accuracy of linear model.

Finally, we can include weather, race, unemployment data, census data in the development of new predictive model and implement different machine learning algorithms and check the model performance on each of the algorithms and select one with high model performance to generate more accurate results. Which can be further used to predict the crime in other demographics.

**2.4 Deployment considerations**

* + 1. **Potential benefits**
* law enforcement agencies can use the predicting technology to deploy their resources more efficiently and effectively(Schlehahn and Aichroth, 2022).
* We can discover trends in huge data sets, which may be utilized by police forces to intervene crimes.
* Present model is although limited to Chicago US, but it would be interesting to apply the proposed framework in other countries or regions.

* + 1. **analysis of risks and potential challenges associated with this deployment.**
* The model developed using machine learning algorithm using population as a single independent feature, it is not a ideal to use one single feature to predict class variable, Accuracy achieved with this model would ineffective.
* Deployment of model developed using ml approach is made through building an application programming interface(API) which could be integrated with other software system. Though deployment may not be infective, but the model results would cause inefficient approach to estimate crime rate.
* Initial Model deployed can be retrained by adding additional features as predictors to predict crime rate, challenge identified is no other features available in the crime historical data has correlation to crime rate.
* Deployment of ml model can be made through multiple approaches, API Service, Dashboard, File based results are the ways to implement deployment in various software system. Need to maintain model to be consistent and high throughput to achieve quicker predictions.
  + 1. **The ethical aspects of the deployment of this project**

We are on the verge of shifting from unpredictable crime to preventing crime in many parts of the world. To achieve this goal more academic research is needed, which will be based on accurate predictions of crime. But a basic consideration of ethical principles suggests that these models should be made available only after they have been thoroughly tested and their reliability, sensitivity is certified.

By implementing predictive models for crime prediction, a convict or suspects fate now rests in non-human hands. Considering an all-encompassing approach to human beings, the ethical principle of not inflicting damage, the demands on potential psychological stresses that may exist between police and civilians in the case of inaccurate forecasts. Potential stress assessment should be used to thoroughly educate individuals before they are investigated for controlled interventions and to counsel persons who are at risk. (Siegel, 2022)

Confidentiality should be safeguarded as much as possible, data which is used for the analysis will include the details of individuals(convicts) and other personal information related to the police forces which is highly confidential. Proper measures should be taken to avoid any kind of disturbances in the future.

Confidentiality should be safeguarded as much as possible, in order to

avoid any kind of discrimination of prediabetic individuals, which, in addition

to being unfair to them, might also discourage their access to diagnostic

tests, resulting ultimately of detriment to the progress of research. However

a sound information should include a description of the current possible

limitations in insurance and employment

Confidentiality should be safeguarded as much as possible, in order to

avoid any kind of discrimination of prediabetic individuals, which, in addition

to being unfair to them, might also discourage their access to diagnostic

tests, resulting ultimately of detriment to the progress of research. However

a sound information should include a description of the current possible

limitations in insurance and emplo

**2.4.4Conclusion:**

With the help of this model, it is very easy to find out the relation and patterns among various data sets. Our case study revolves around predicting the crime rate which may happen with the increase in population. using the concept of machine leaning algorithms(Simple linear regression algorithm) a model was built using the training data set that have undergone data cleaning, transformation, and integration. The present model predicts the crime rate with effect to population with an accuracy of 0.94. further many graphs were developed, and interesting statics are found which helped in understanding Chicago crimes which can in turn help in capturing the factors that can keep the society safe. (Schlehahn and Schreiner, 2022)

**References:**

* Abdallah, Z. and Webb, G., 2022. *Data Preparation*. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/316113863\_Data\_Preparation#:~:text=Data%20preparation%20is%20the%20process,is%20ready%20for%20further%20analysis.> [Accessed 14 February 2022].
* Data Science Central. 2022. *How To Interpret R-squared and Goodness-of-Fit in Regression Analysis - DataScienceCentral.com*. [online] Available at: <https://www.datasciencecentral.com/regression-analysis-how-do-i-interpret-r-squared-and-assess-the/> [Accessed 15 February 2022].
* Gassoumis, Z., 2022. *Negative constant value*. [online] Available at: <https://www.researchgate.net/post/A-negative-coefficient-for-a-constant-in-a-linear-regression#:~:text=Depending%20on%20your%20dependent%2Foutcome,variables%20are%20set%20to%200.> [Accessed 16 February 2022].
* Jmp.com. 2022. *Correlation*. [online] Available at: <https://www.jmp.com/en\_ca/statistics-knowledge-portal/what-is-correlation.html#:~:text=Correlation%20is%20a%20statistical%20measure,statement%20about%20cause%20and%20effect.> [Accessed 16 February 2022].
* Kumar, A. and Sahni, S., 2022. *A Comparative Study of Classification Algorithms for Spam Email Data Analysis*. [online] Available at: <https://www.researchgate.net/publication/267718689\_A\_Comparative\_Study\_of\_Classification\_Algorithms\_for\_Spam\_Email\_Data\_Analysis> [Accessed 16 February 2022].
* Mayo, M., 2022. *7 Steps to Mastering Data Preparation with Python - KDnuggets*. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2017/06/7-steps-mastering-data-preparation-python.html> [Accessed 14 February 2022].
* Medium. 2022. *Improve Linear Regression Using Statistics*. [online] Available at: <https://towardsdatascience.com/statistics-supporting-linear-models-bfc24fb9781f#:~:text=Fixing%20the%20skewness%20of%20data%20can%20provide%20a,has%20a%20long%20tail%20on%20the%20right%20end.> [Accessed 15 February 2022].
* Medium. 2022. *Categorical encoding using Label-Encoding and One-Hot-Encoder*. [online] Available at: <https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd> [Accessed 16 February 2022].
* McLeod, D., 2022. [online] Simplypsychology.org. Available at: <https://www.simplypsychology.org/p-value.html#:~:text=How%20do%20you%20know%20if%20a%20p%2Dvalue%20is%20statistically%20significant%3F&text=The%20level%20of%20statistical%20significance,%E2%89%A4%200.05)%20is%20statistically%20significant.> [Accessed 16 February 2022].
* Schlehahn, E. and Schreiner, R., 2022. *Benefits and Pitfalls of Predictive Policing*. [online] research gate. Available at: <https://www.researchgate.net/publication/304293638\_Benefits\_and\_Pitfalls\_of\_Predictive\_Policing> [Accessed 18 February 2022].
* Siegel, E., 2022. *Six Ethical Quandaries of Predictive Policing - KDnuggets*. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2020/11/six-ethical-quandaries-predictive-policing.html> [Accessed 18 February 2022].
* ThoughtCo. 2022. *Definition and Examples of Independent and Dependent Variables*. [online] Available at: <https://www.thoughtco.com/independent-and-dependent-variable-examples-606828> [Accessed 15 February 2022.

**Appendix A**

**Data Visualizations:**

**A picture containing text, antenna

Description automatically generated**

Figure 11 Line graph represent the number of crimes per year

* Decreasing trend of crimes from 2001 to 2019.

Chart

Description automatically generated

Background pattern

Description automatically generated

Chart, line chart

Description automatically generated

A picture containing window, building, shoji

Description automatically generated

Figure 12 Line graph representing the number of crimes per district.

* Each district is identified with the specific code on the top right from 1.0 to 31.0

**Chart, funnel chart

Description automatically generated**

Figure 13 Showing crimes with location description.

* We can notice that there are high number of crimes on the street and residence.

A picture containing text, building

Description automatically generated

Figure 14 total crimes w.r.t to primary type.

**Appendix B : Sample code**

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Source code link : https://github.com/silentsword03?tab=repositories