***Crime Rate Predictive Analysis and Smart Police Patrolling***

Utkarsh Sharma 1 (**Corresponding Author**)

Yash Garg 2

1, 2 ASET, Amity University, Uttar Pradesh, India

1 [skhiearth@gmail.com](mailto:skhiearth@gmail.com) , 2 [ygarg704@gmail.com](mailto:ygarg704@gmail.com)

1 C-273, Millenium Apartments, Sector-18, Rohini, New Delhi-110089, India

2 118, 2nd floor, Deepali Enclave, Gate No.1, Pitampura, New Delhi-110034, India

***Abstract – teko likhna hai***

1. **INTRODUCTION**

**Teko likhna hai**

1. **RELATED WORK**

In paper [1], the authors are using the open crime dataset of the LA County Sheriff to research the issue of police patrol preparation. They suggest a new approach to building a cluster network to effectively distribute patrols based on information entropy. This minimises the time-to-arrival of the police and decreases the total number of police on patrol.

ROTA-Analytics, a web-based framework that aims to provide outputs for the prediction of crime occurrence, is discussed in paper [2]. This prediction of crime events allows patrol managers to draw up a list of predefined locations (points) and the time at which each police vehicle needs to patrol. ROTA-Analytics promotes the collection of many machine and statistical learning approaches to construct a crime prediction environment in different areas of the region.

In the paper[3], the authors suggest a solution to promote the efficient distribution of police forces in a city between multiple districts, based on the degree to which each district is vulnerable to crime for a given month at a given time, on a given day. In this work, they used the Chicago Crime Dataset.

A structure for the implementation of police patrol strategies based on an urban street network is suggested in the thesis [4], which essentially includes crime hotspots and the rest of the territory. This system consists of three strategies, including a district model, a repeated coverage patrol routing strategy, and an infrequent coverage patrol routing strategy.

In paper [5], the framework is web-based with the aim of showing the areas of crime (hotspots). The framework contains three kinds of users: Informer, Patrolman and Decision Maker. The first one, filling up the diagram, will report crimes. Consequently, through the diagram, the Decision Maker will see hotspots and he / she will disperse the Patrolmen in the city better.

1. **PROPOSED TECHNIQUE**

Primarily, the dataset which was taken from [San francisco’s government website](https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry/data) was prepared for Experimental Data Analysis by data cleaning. The data was analysed and visualized for further understanding and to answer questions like which district has the highest crime rate. After EDA the dataset was grouped into 4 quartiles which are ‘Low’, ‘Medium’, ‘High’, and ‘Extreme’. For grouping, whisker plot was used to determine how many qurtiles should the data be divided into, for more accurate results, pandas quartile based discretion function was used to determine number of quartiles to formed. After grouping, the remaining rows were 20160 and 5 columns. We used grouping to make our function for classification. Many classification algorithms like Logistic Regression, K nearest neighbour, Decision Tree, Random Forest, XG Boost and Voting classifier were used along with ANN (Artificial Neural Network). The research methodology is shown in Fig(1). Our Best model is the ANN Model with test accuracy of 79.81%.

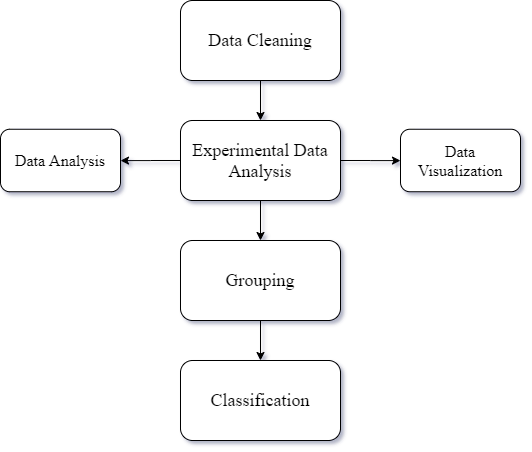


Figure 1. Application Framework.

1. **DATA CLEANING, ANALYSIS AND VISUALIZATION**

The dataset was taken from San Francisco’s Government website. It contains police department incident reports from year 2003 to 2018 (May). Initially it had 2,160,953 rows and 35 columns. In Data Cleaning, initially we checked for null values which we found many, so those rows were removed and then, out of 35 columns only 8 were of use which were day, date, time, district, result and X and Y coordinates and location. After the wanted columns were selected time and date column were merged into one using pandas library ‘DatetimeIndex’ function. Furthermore, the dataset was analyzed and visualized for more clarity.

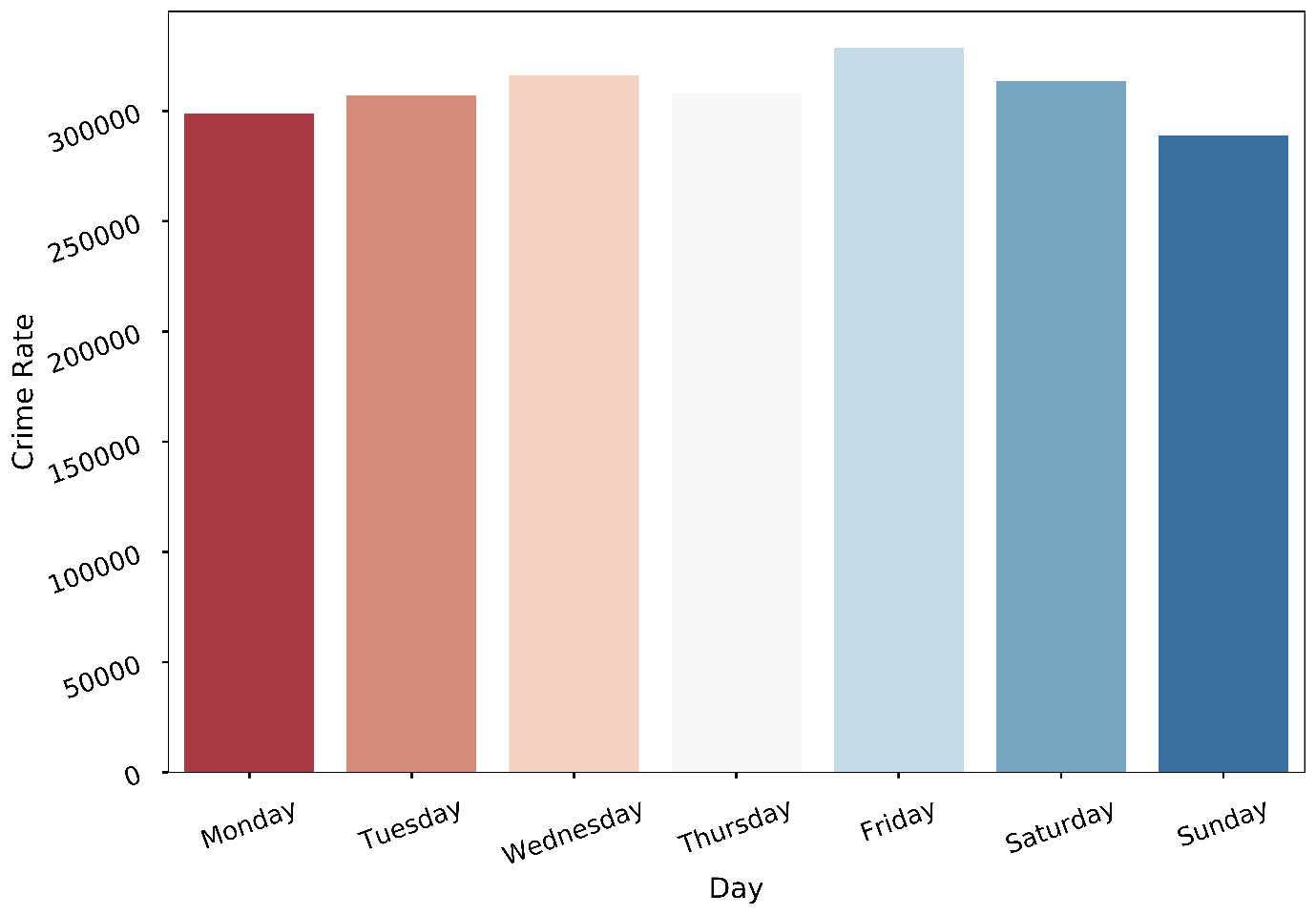


Figure 2. Crime Rate per Day.

Fig(2) answers the question, ‘which day has the highest crime rate in San Francisco?’ and we can clearly see from the plot above that Friday has the highest crime rate out of all days and Sunday has the lowest crime rate (lower than 300,000) in San Francisco in all the years from 2003 to 2018.

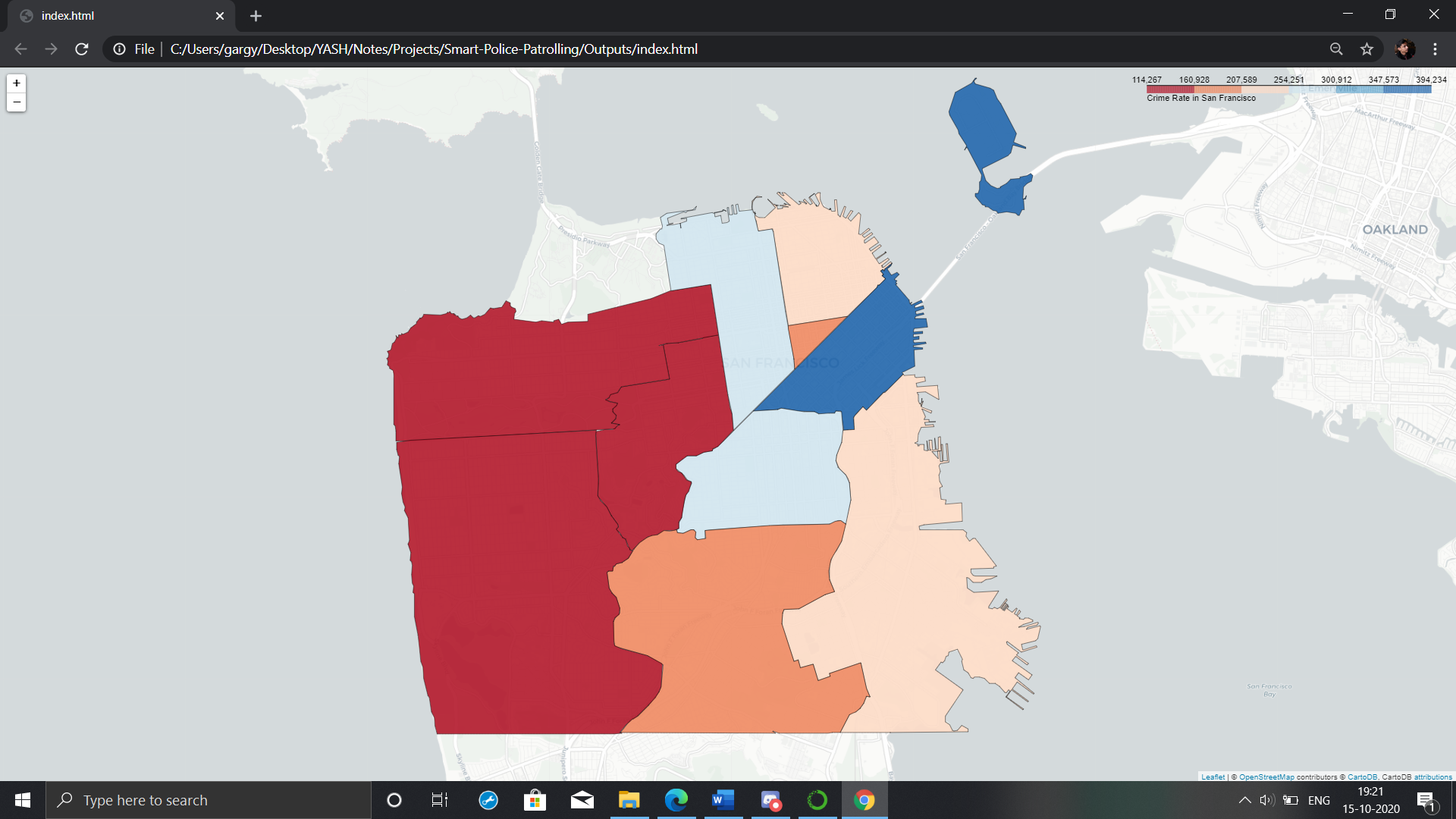


Figure 3. Density Map of San Francisco.

Fig (3) shows the density map of San Francisco and we can clearly see that Park, Richmond and Taraval boroughs have low crime rate (< 160,928) while Tenderloin and Ingleside have crime rate between 160,928 and 207,589, Central and Bayview have crime rate between 207,589 and 254,251 and Northern and Mission have less than 300,000 and greater than 254,000 and finally Southern borough has the highest crime rate in all of San Francisco.

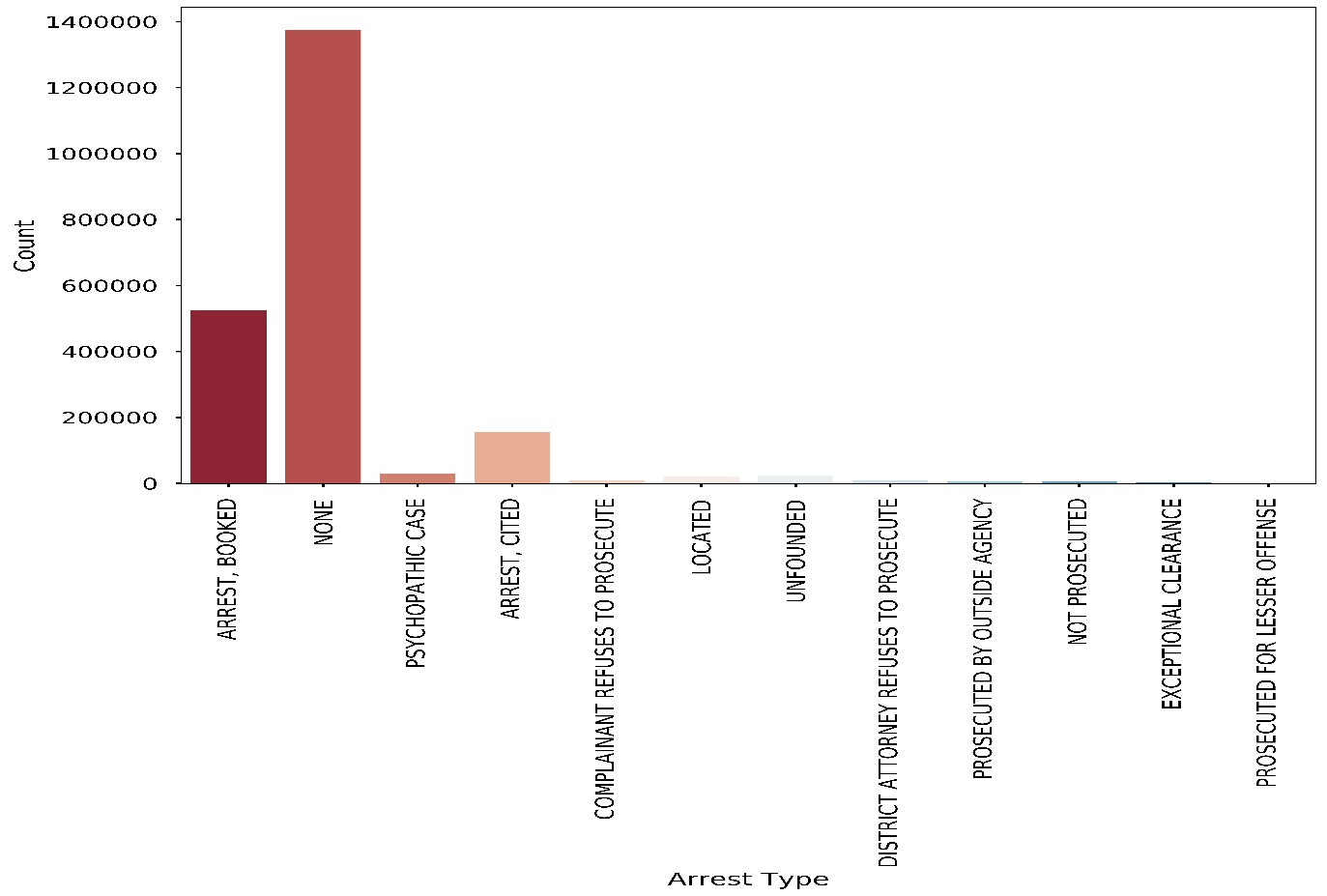


Figure 4. Arrest Count.

Fig(4) represents the types of arrests in our dataset. As we can see, almost 1400000 cases have not been solved from 2003 to 2018 whereas, in around 500000 cases, the crime has been solved and the suspect has been arrested and booked, and for other cases where the suspect has noot been prosecuted or located the numbers are less than 200000.

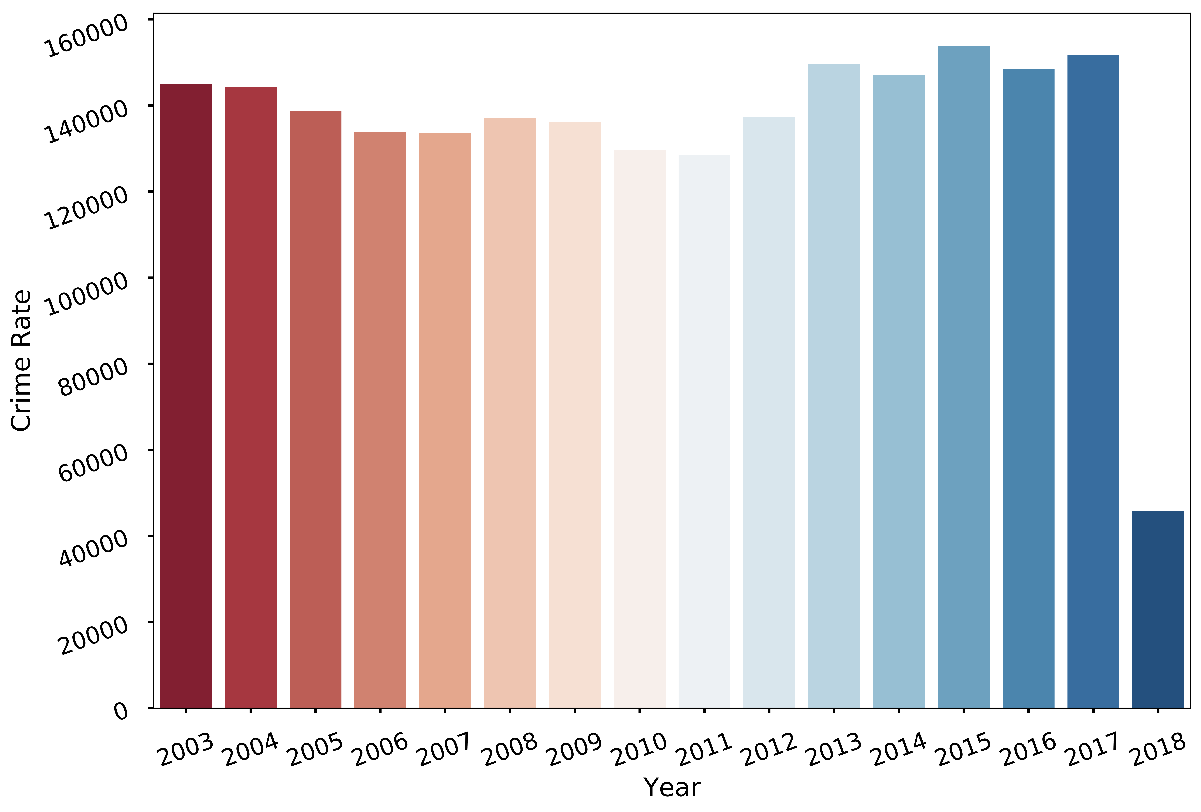


Figure 5. Crime Rate per year.

Fig(5) tells us, which year has the highest crime rate in San Francisco and we can clearly see from the plot above that 2015 has the highest crime rate out of all years followed by 2017 and 2013 and then 2014 and 2016.

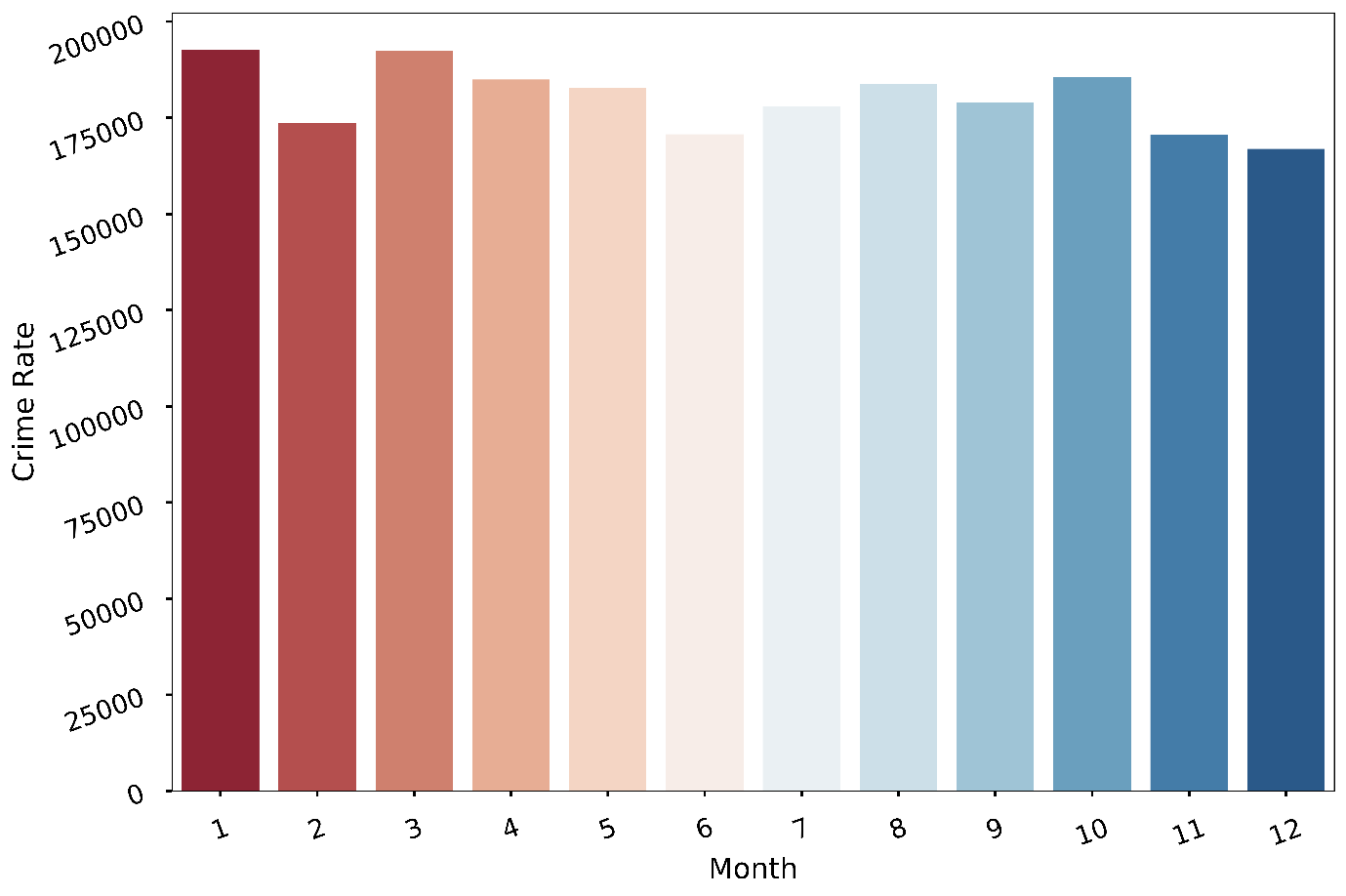


Figure 6. Crime Rate per month.

Fig(6) tells us, which month has the highest crime rate in San Francisco and we can clearly see from the plot above that January and March almost have the highest crime rate out of all months combined from year 2003 to 2018.

1. **GROUPING AND CLASSIFICATION**
2. Grouping

Grouping was done to make the classification column. Formerly, the time column was converted to just hour using pandas DatetimeIndex. After that, crime occurences for every district was calculated on a given timestamp and grouped to the original dataset making a whole new column named Crime Occurences. These occurences were distributed from 0 to 400 across all districts.

Now, the crime occurrence data was transfigured into quartiles. A quartile is a type of quantile which divides the number of data points into four more or less equal parts, or quarters. The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set. In our case, the data was divided into 4 quartiles which were then named low, medium, high and extreme. These quartile were not randomly divided, pandas qcut function was used to divide the data (Crime occurences) into quartiles. Each quartile has approximately 5000 values and this data was then used for classification.

1. Classification

After Analyzing the whole dataset, we performed some classification algorithms and ANN on the dataset to predict whether quartile of crime, that is, low, medium, high and extreme. The data was then encoded using LabelEncoder and pandas get\_dummies function. The test train split size was 0.1 and after grouping the data the total number of rows are 20160 and 7 columns. First we ran some classification algorithms on our dataset which were Logistic Regression, K Nearest Neighbour, Decision Tree, Random Forest with GridSearchCV, XG Boost and Voting Classifier. The highest accuracy in classification algorithms was achieved by voting classifier (**78%**) followed by Random forest (**77%**).

Then, we made our artificial neural network with Input layer of 512 units with activation = ‘relu’ and then a hidden layer also with 512 layers and activation = ‘relu’ and then an output layer with 4 units because we are determining 4 classes and activation = ‘softmax’. The model was then compiled with adam optimizer. We tried many optimizers like SGD, RMSPROP, ADAGRAD, ADADELTA, and ADAM and the best optimizer for our model was ADAM with 100 epochs and batch size of 128. The test model was then evaluated and we achieved the highest test accuracy of (**79.86%**).

1. Tools Used

NumPy, pandas and matplotlib, and seaborn packages were used for the classification of the model, cleaning of the dataset, grouping and experimental data analysis. Sklearn module was used for classification and keras was used for deep learning (ANN).

Python language was used for implementation which was done in jupyter notebook kernels. Python was used because of high availability of libraries and packages.

1. **RESULTS**

(Teko likhna hai) Confusion matrix bana liyo, and result teko hi likhna padega coz mere paas thodi km values ka dataset hai esliye accuracy 2% km aa rhi hai, ANN.ipnyb kholiyo usme confusion matrix bana diyo, tables classififcation\_report ki edhar daal diyo.

1. **CONCLUSION AND FUTURE WORK**

Teko likhna hai

**REFERENCES**

1. Hochstetler, Jacob, Lauren Hochstetler, and Song Fu. "An optimal police patrol planning strategy for smart city safety." 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2016.
2. Junior, Adelson Araujo, et al. "A predictive policing application to support patrol planning in smart cities." 2017 International Smart Cities Conference (ISC2). IEEE, 2017.
3. Yadhunath, Ramshankar & Srikanth, Srivenkata & Sudheer, Arvind & Palaniswamy, Suja. (2019). Identification of Criminal Activity Hotspots using Machine Learning to Aid in Effective Utilization of Police Patrolling in Cities with High Crime Rates. 1-6. 10.1109/CSITSS47250.2019.9031057.
4. Chen, Huanfa. (2019). Developing Police Patrol Strategies Based on the Urban Street Network.
5. Almeida da Costa Júnior, Ademir & Silva, Gilton. (2018). A Decision Support System for Police Patrolling.