# San Francisco Crime Research

# Analyzing and predicting crime in the city is very important for people

- The idea of this study is to help people planning their journey in the San Francisco. It will help them to choose the right route where the crime is low.
- This project can help people investigate crime rate in their cities.

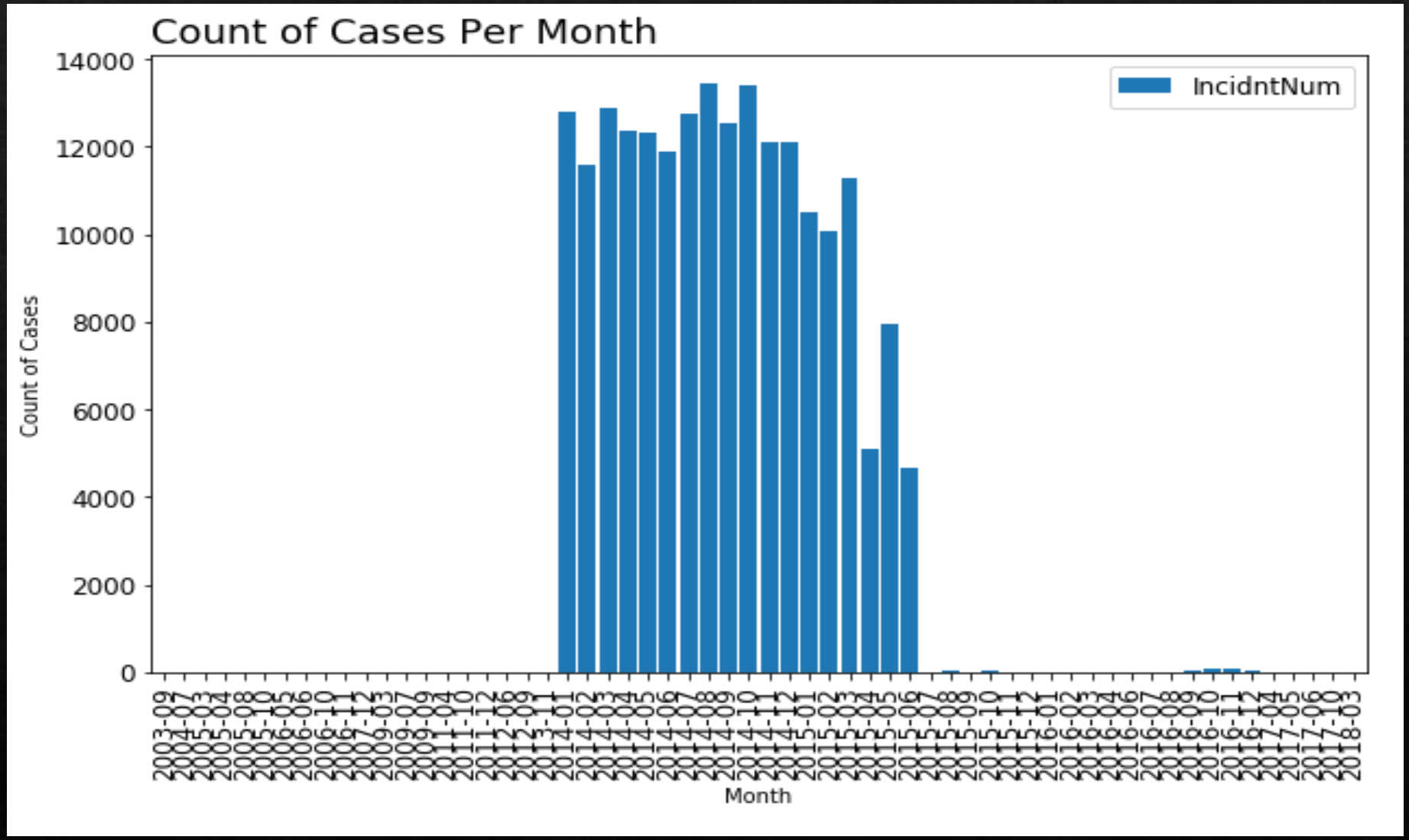
# Data analysis

After preparing the data we decided to analyze the data- we visualized a different charts

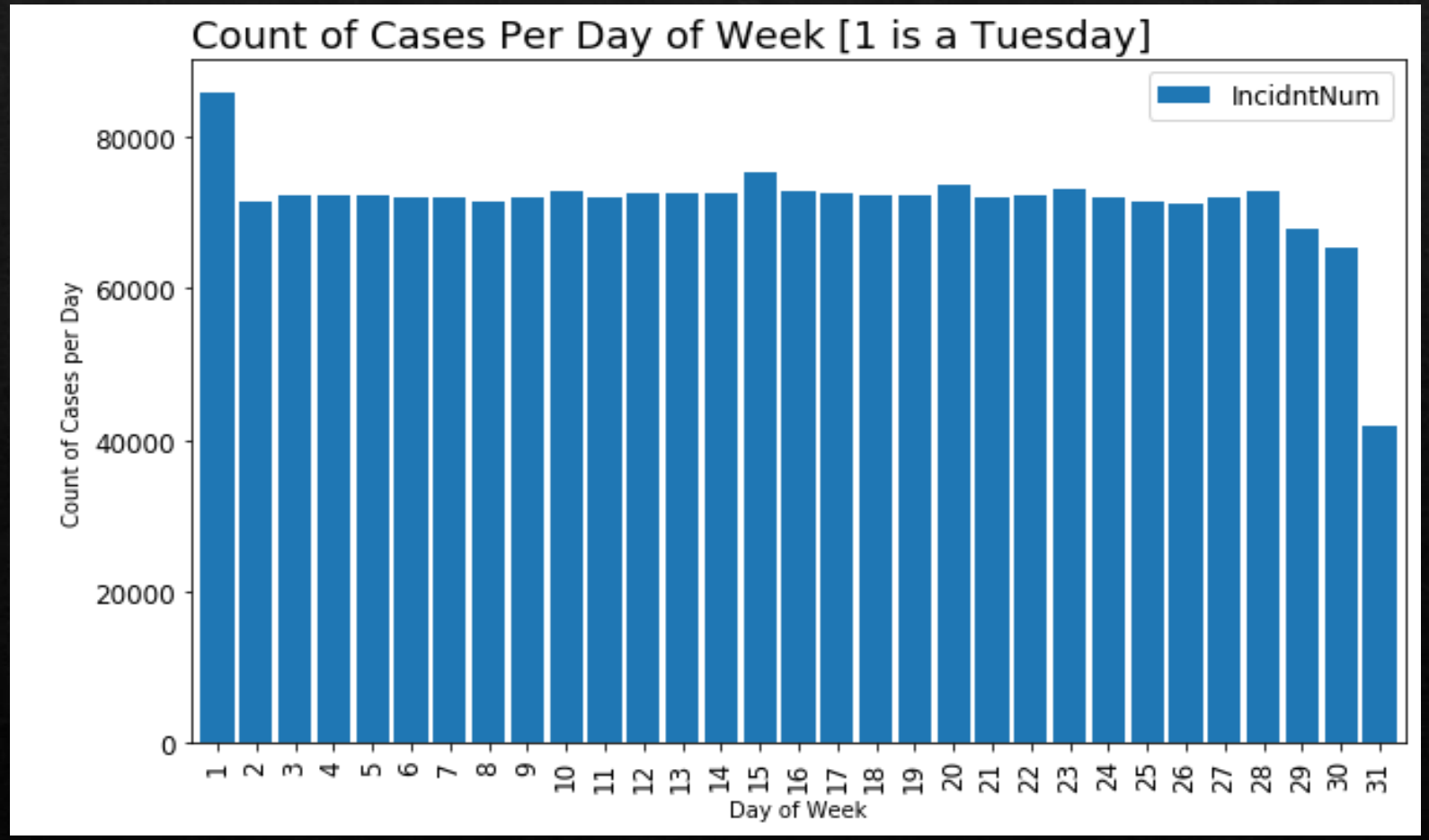
We will analyze:

- **Number of crimes per month**
- **Number of crimes occurring on each day**
- **Number of  crimes  occurring on each hour**

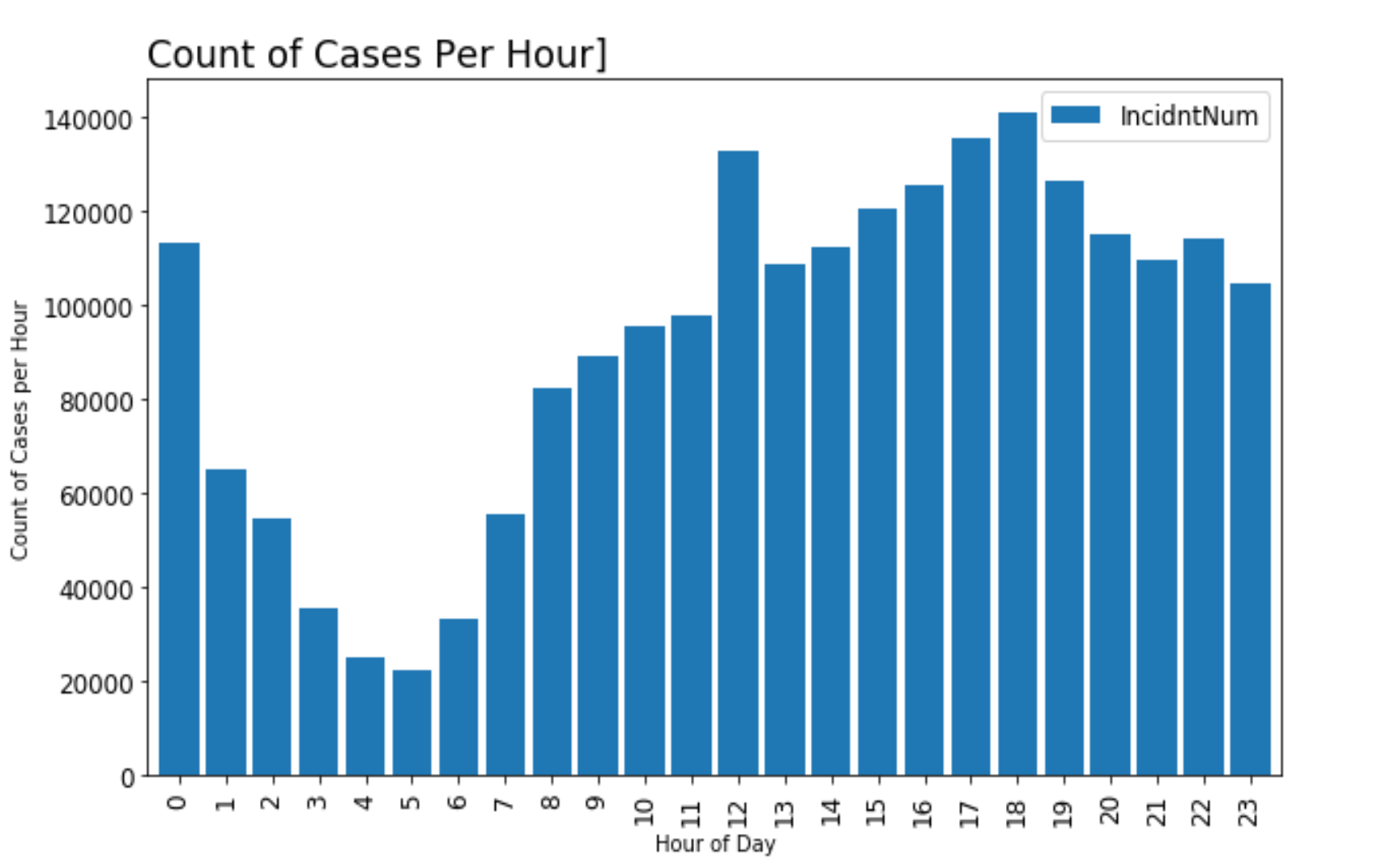
# Number of crimes per month



# Number of crimes occurring on each day



# Number of crimes occurring on each hour



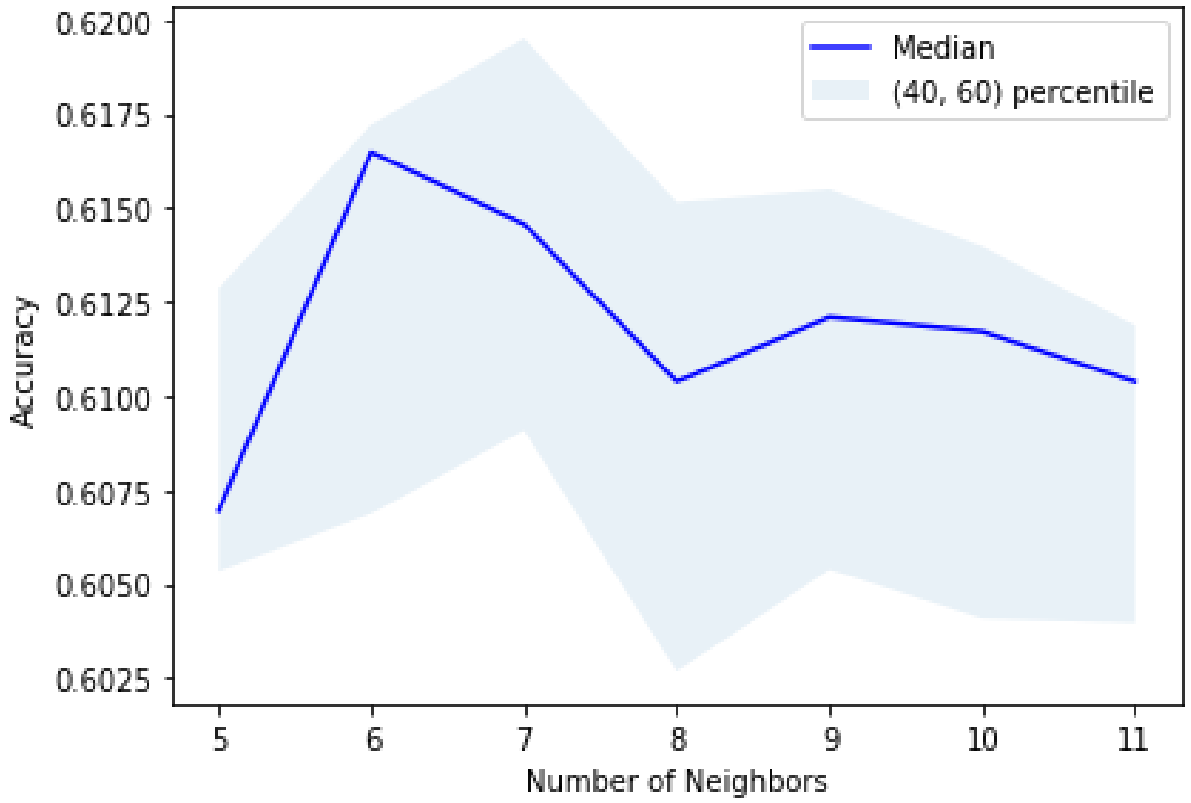
# Predictive Modeling

There are of 4 different models to predict new crimes:

- **Logistic Regression**
- **Random forest**
- **kNN ( k Nearest Neighbor)**
- **Decision Tree**

Firstly I selected the best parameters for models (e.x. the number of neighbors for kNN), and then used these parameters. I used the accuracy metric to control the models.

# k Nearest Neighbor

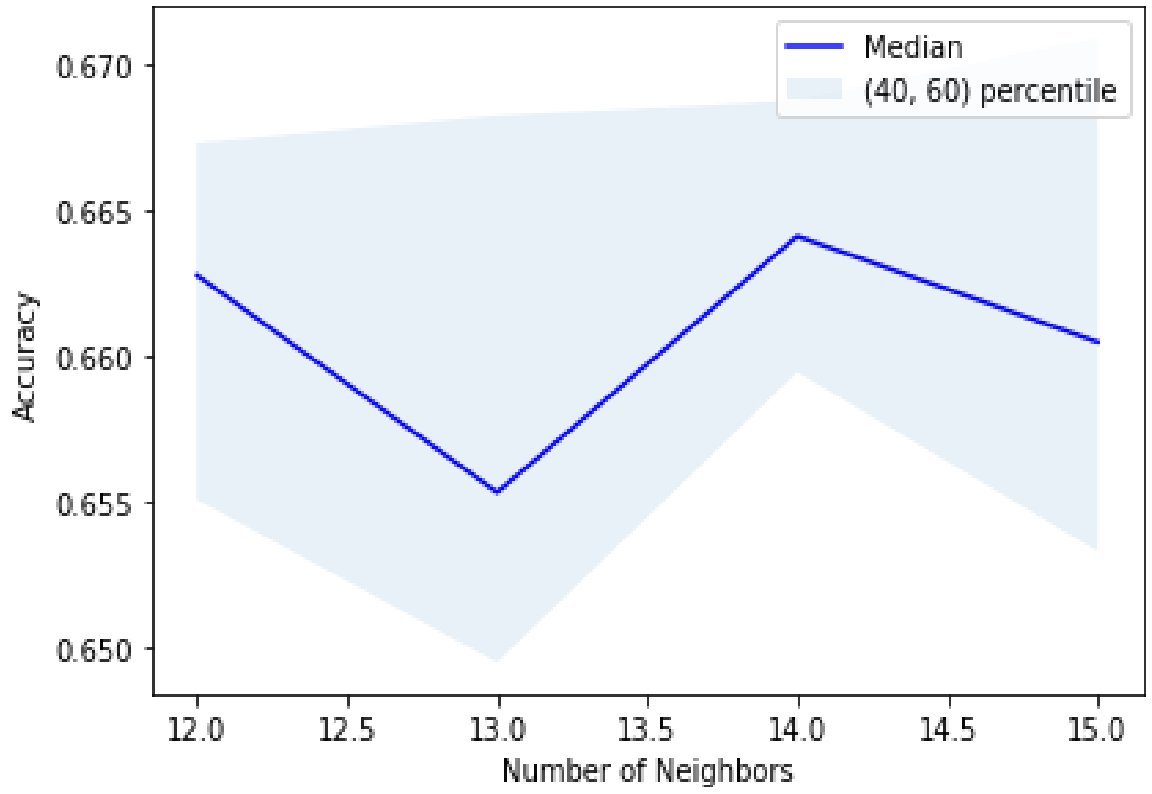



```python
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
#K Nearest Neighbor(KNN)
model = KNeighborsClassifier(n_neighbors = 6)
model.fit(X_train, y_train)
y_pred=model.predict(X_test)
accuracy_score(y_test, y_pred)
```

```
0.51
```

Result: The accuracy score is equal 0.51. It is low, we will not use this model

# Random forest

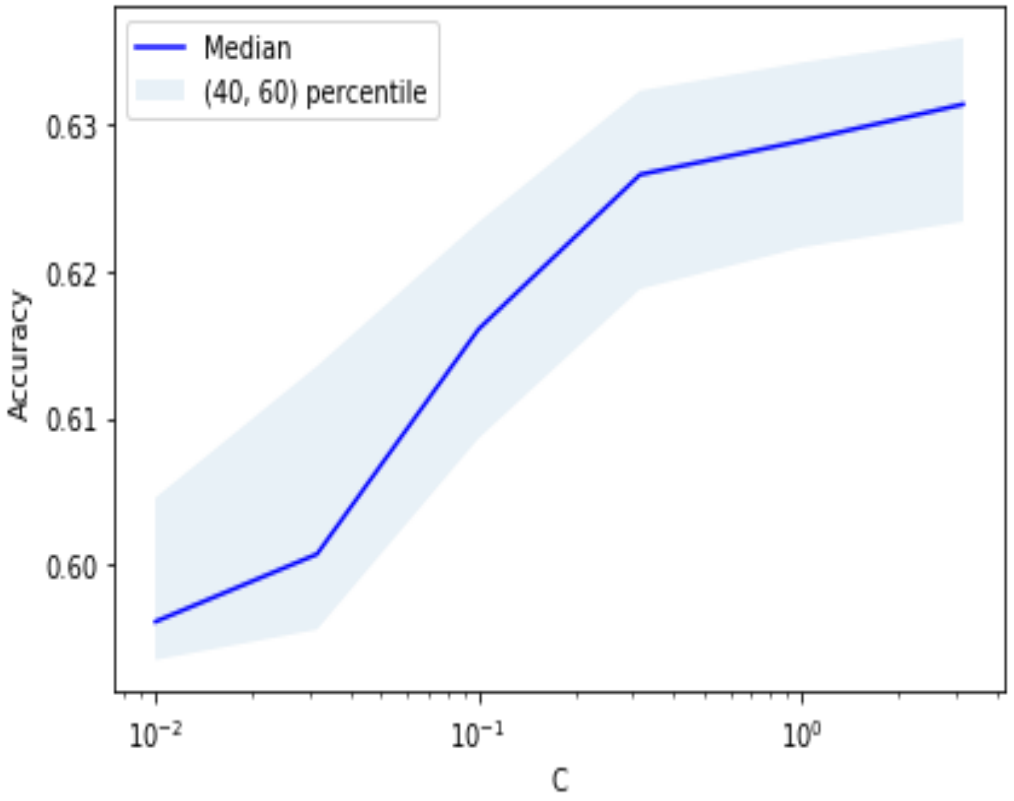



```
#Random Forest
model = RandomForestClassifier(n_estimators = 14, max_features = 'sqrt')
model.fit(X_train, y_train)
y_pred=model.predict(X_test)
accuracy_score(y_test, y_pred)
```

```
0.89
```

Result: The accuracy score is equal 0.89. It is the highest score. This model is convenient to us.

# Logistic Regression



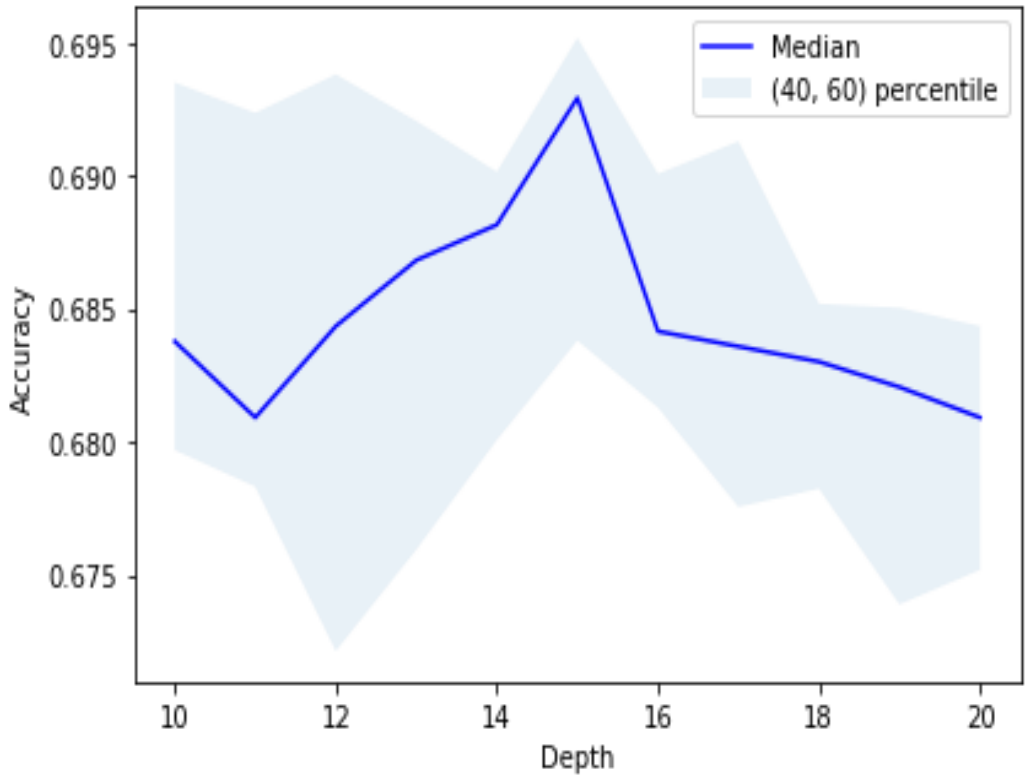

```
#Logistic Regression
model = LogisticRegression(C = 3.1622776601683795  , solver = 'liblinear')
model.fit(X_train, y_train)
y_pred=model.predict(X_test)
accuracy_score(y_test, y_pred)
```

```
C:\Users\slava\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:469:
nged to 'auto' in 0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
```

```
0.3389090909090909
```

Result: The accuracy score is equal 0.34. It is low, we will not use this model.

# Decision Tree




```python
#Decision Tree
model = DecisionTreeClassifier(criterion = "entropy", max_depth = 100)
model.fit(X_train, y_train)
y_pred=model.predict(X_test)
accuracy_score(y_test, y_pred)
```

```
0.6321
```

Result: The accuracy score is equal 0.6321. It is low, we will not use this model.

# Conclusions

In this study, I analyzed the history of crimes in Chicago. I found dependencies between crimes and hours or days. After that I build a few models to predict a new crimes.

I hope my this research will help people planning their journey in the San Francisco. If I visit San Francisco, firstly, I will use this my research. And some people can use this program to analyze and predict the crimes in their city ( as for me I can use it to analyze cities in my country).