Deexith Mysore Nagaraj,
Pavan Bharadwaj Holenarasipura.

San Francisco Crime Classification

Problem Selection

- To categorize the different types of crime in San Francisco using the San Francisco Crime Dataset.
  Source: Dataset is taken from Kaggle (https://www.kaggle.com/c/sf-crime)
- **Tasks:**
  - **Task 1:** Build a classifier to predict Crime category.
  - **Task 2**: Build a clustering model using various features to predict the category of crime.

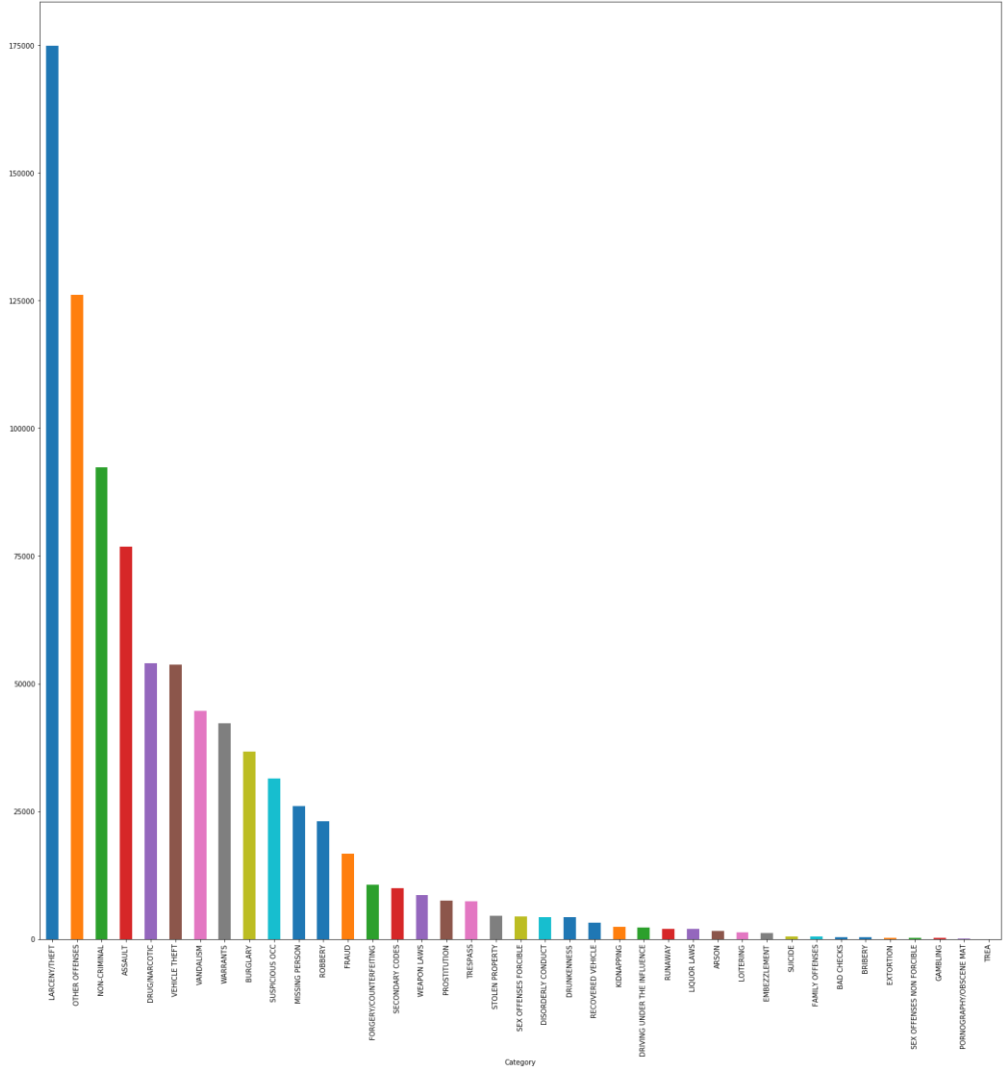
Data-Frame

- **Date**: The timestamp of the crime recorded.
- **Category**: The category of the of the crime recorded.
- **Description**: A short note on the crime.
- **DayOfWeek**: The day on which the crime took place.
- **PdDistrict**: The police department, under which the crime is reported.
- **Resolution**: The status of the crime, resolved or unresolved.
- **Address**: The address of the crime scene:
- **X**: The latitude of the crime scene.
- **Y**: The longitude of the crime scene.
- **Zip-code**: Zip-code of the area where the crime was reported.
- **Weather**: Weather information of each day.

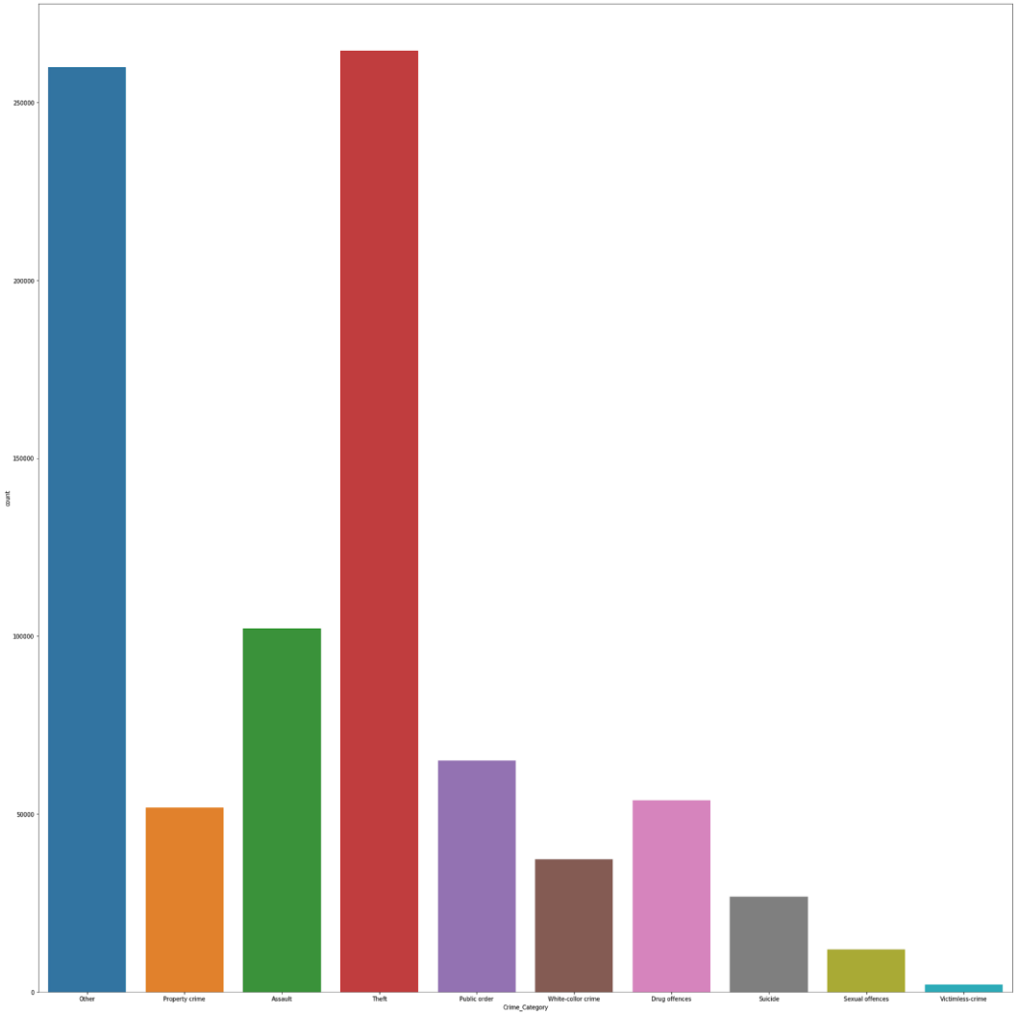
Pre-Processing

- The values are very detailed and doesn't contain any null values. However, it is hard to determine the relationship between the features and the crime classes. Hence additional information is taken from other dataset.

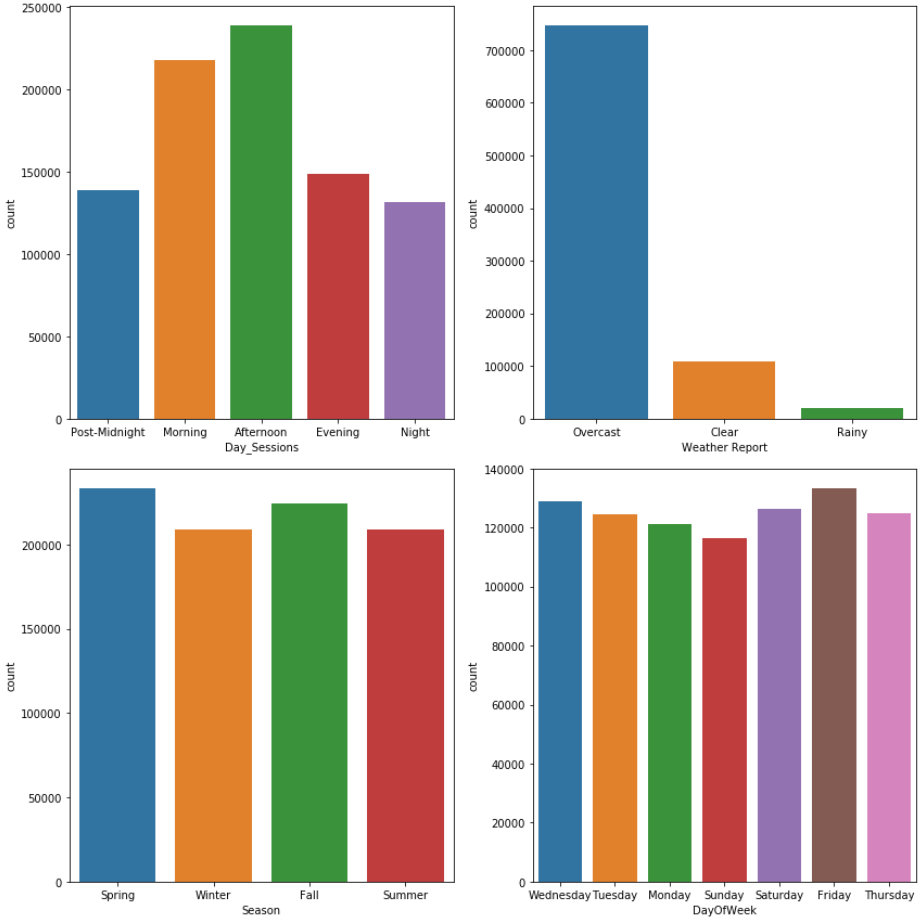
Crime Categories

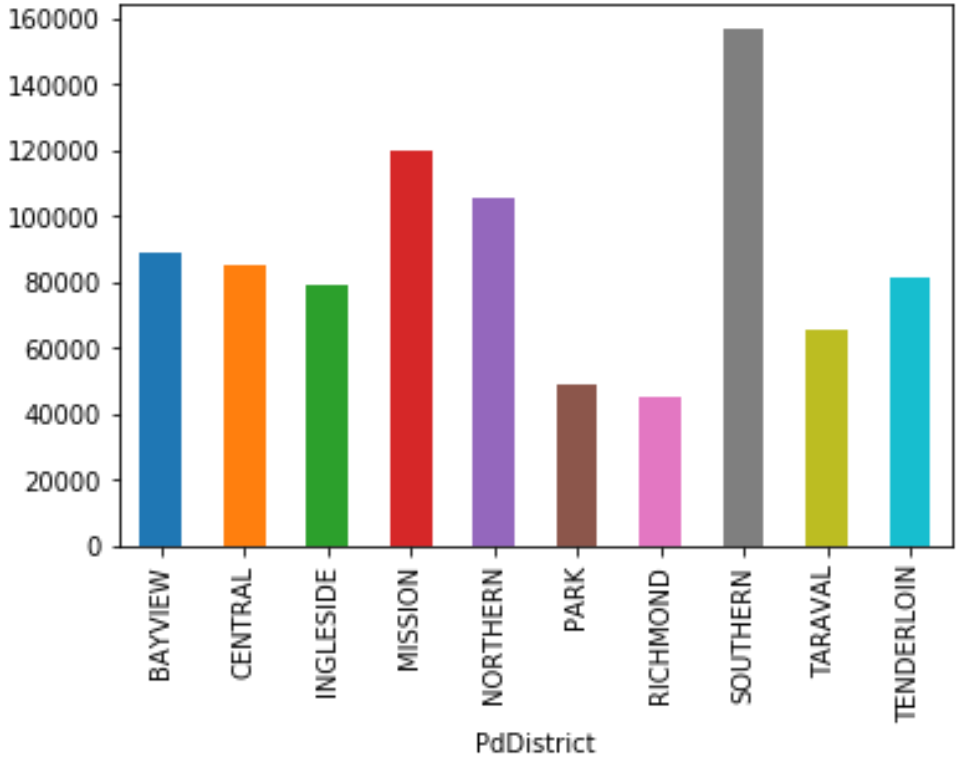


Grouped Crime Categories



Exploratory Data Analysis





Exploratory Data Analysis Contd.

Classification

Classifier	Accuracy, Precision, Recall, F1 Score
KNN	0.297531771632702, 0.25624441256286706, 0.31593356599523625, 0.2749205150139882.
Decision Tree	0.3506780868088218, 0.318143608476642, 0.48890564354700833, 0.3666771009191756.
Random Forest	0.3506171523019914, 0.31642295565768586, 0.4888206903509098, 0.36647755221217937.
Decision Tree using Boosting	0.3506171523019914, 0.31642295565768586, 0.4888206903509098, 0.36647755221217937

Clustering	Adjusted Random Score, Normalized Mutual Score.
K Means	0.671763226297113 4, 0.833943444109823 1
K++	0.518028668805664 1, 0.777924140333750 9

Clustering

Conclusion

- The dataset is highly random, and features were less likely related to the type of crime. However, categorizing the feature and crimes improved the performance of the model slightly.

- Thus, we can conclude from the above results and plots, that the type of crime is less likely dependent of the external factors such as weather, day, time and location.

Insights Gained

- Random Results when there were no relation between the features and classes.
- How the number of classes, affects the model. Higher number of classes affected the accuracy of the model.
- Categorizing the values of features helped to improve the performance of the classification model.

Thank You