# Data Wrangling Homework 1 – By Esha Singh

# DATA PROFILING

## CrimeDate
Categorical

HIGH CARDINALITY

| | |
|---|---|
| Distinct | 2300 |
| Distinct (%) | 0.8% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 2.2 MiB |

| | |
|---|---|
| 04/27/2015 | 421 |
| 06/05/2016 | 255 |
| 12/20/2018 | 212 |
| 01/20/2017 | 205 |
| 10/25/2017 | 202 |
| Other values (2295) | 291466 |

## CrimeCode
Categorical

HIGH CARDINALITY
HIGH CORRELATION

| | |
|---|---|
| Distinct | 81 |
| Distinct (%) | < 0.1% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 2.2 MiB |

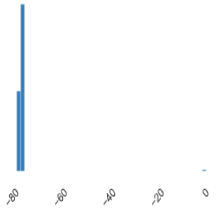
| | |
|---|---|
| 4E | 48061 |
| 6D | 38977 |
| 5A | 2535 |
| 7A | 2439 |
| 6J | 162 |
| Other values (76) | 139738 |

## Longitude
Real number (ℝ)

HIGH CORRELATION

| | |
|---|---|
| Distinct | 97441 |
| Distinct (%) | 33.3% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Infinite | 0 |
| Infinite (%) | 0.0% |
| Mean | -76.18956175 |

| | |
|---|---|
| Minimum | -81.5291885 |
| Maximum | 0 |
| Zeros | 1635 |
| Zeros (%) | 0.6% |
| Negative | 291126 |
| Negative (%) | 99.4% |
| Memory size | 2.2 MiB |

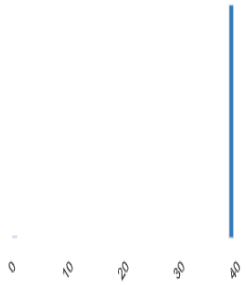



## Latitude
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

| | |
|---|---|
| Distinct | 95904 |
| Distinct (%) | 32.8% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Infinite | 0 |
| Infinite (%) | 0.0% |
| Mean | 39.08722695 |

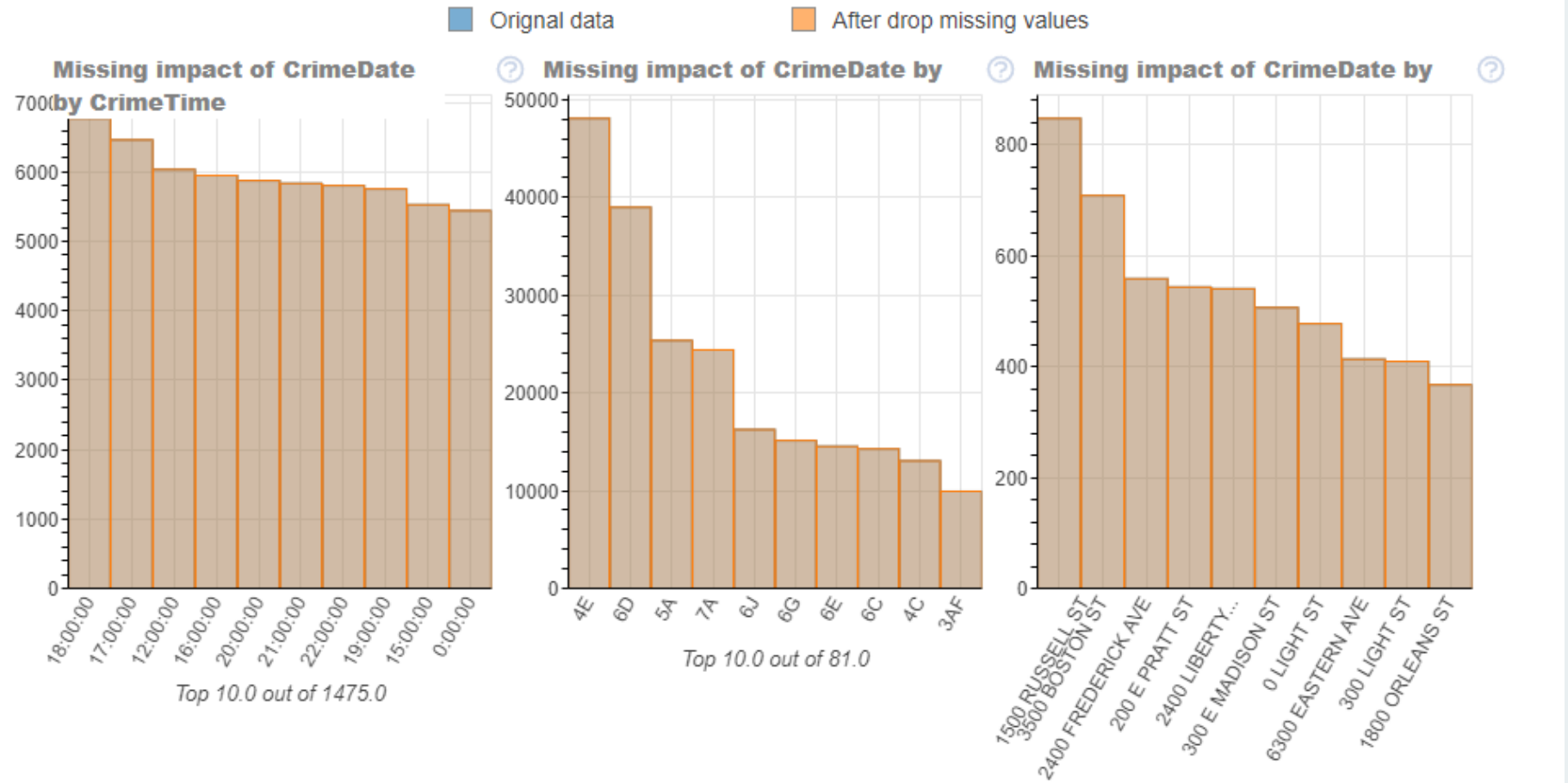
| | |
|---|---|
| Minimum | 0 |
| Maximum | 39.66133158 |
| Zeros | 1635 |
| Zeros (%) | 0.6% |
| Negative | 0 |
| Negative (%) | 0.0% |
| Memory size | 2.2 MiB |



# Results & Charts

```
plot(df_str)

plot_missing(df_str)
plot_missing(df_str, "CrimeDate")
```



# Results & Charts

```
In [9]: skim(df_int)
```

skimpy summary

*Data Summary*

| dataframe | Values |
|---|---|
| Number of rows | 292761 |
| Number of columns | 4 |

*Data Types*

| Column Type | Count |
|---|---|
| float64 | 3 |
| int32 | 1 |

*number*

| column_name | NA | NA % | mean | sd | p0 | p25 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|
| Longitude | 1600 | 0.56 | -77 | 0.044 | -82 | -77 | -77 | -76 | |
| Latitude | 1600 | 0.56 | 39 | 0.03 | 38 | 39 | 39 | 40 | |
| Location 1 | 290000 | 100 | nan | nan | nan | nan | nan | nan | |
| Total Incidents | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | |

End

# Analysis Dashboard

## Overview

Overview | Alerts 25 | Reproduction

### Dataset statistics

| | |
|---|---|
| Number of variables | 16 |
| Number of observations | 292761 |
| Missing cells | 867451 |
| Missing cells (%) | 18.5% |
| Duplicate rows | 8584 |
| Duplicate rows (%) | 2.9% |
| Total size in memory | 35.7 MiB |
| Average record size in memory | 128.0 B |

### Variable types

| | |
|---|---|
| Categorical | 12 |
| Unsupported | 2 |
| Numeric | 2 |

- The dashboard has clickable buttons that can be used to see all the analysis done on the data.
- The data has been brought to a single dashboard with all explanation

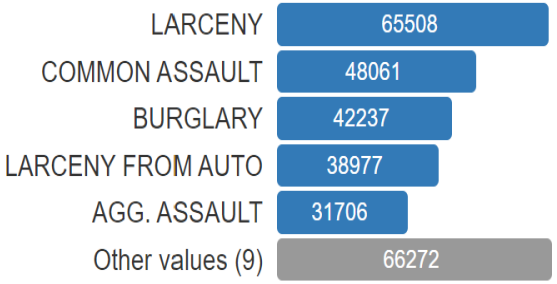
+

# Understanding the Overall Crime distribution

**Description**
Categorical

HIGH CORRELATION

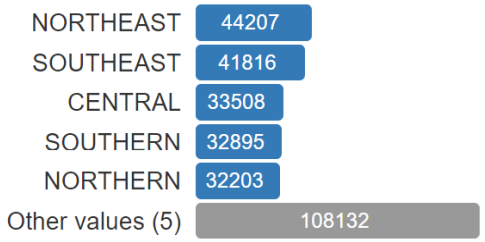
| | |
|---|---|
| Distinct | 14 |
| Distinct (%) | < 0.1% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 2.2 MiB |




**District**
Categorical

HIGH CORRELATION

| | |
|---|---|
| Distinct | 10 |
| Distinct (%) | < 0.1% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 2.2 MiB |




- As we can see larceny and common assault are most in number and we can see other types of crime aggregate to 66272.
- We can see that Northeast and southeast have the highest number of cromes.

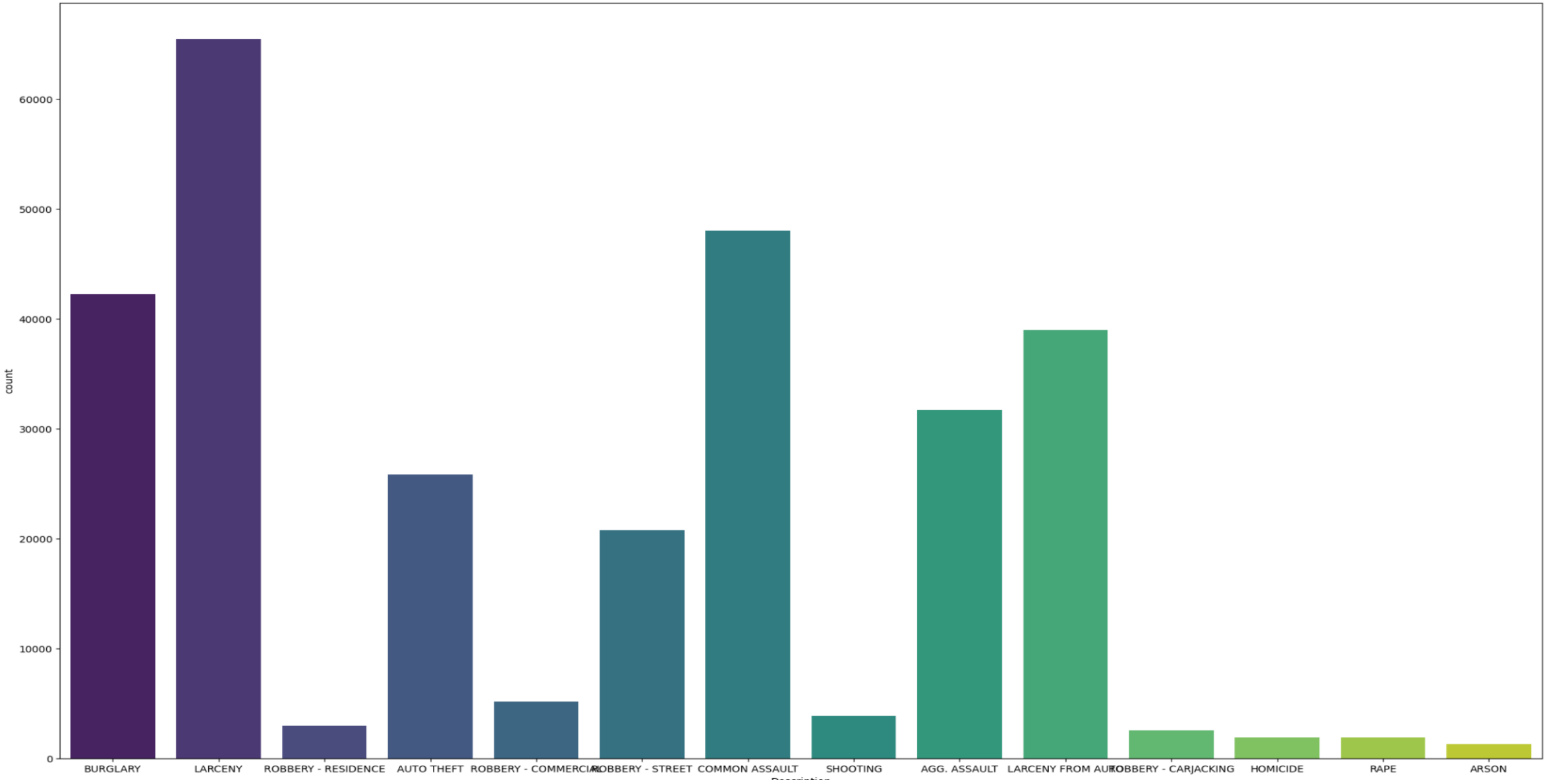
# CORRELATION

This picture shows the correlation between each variable.

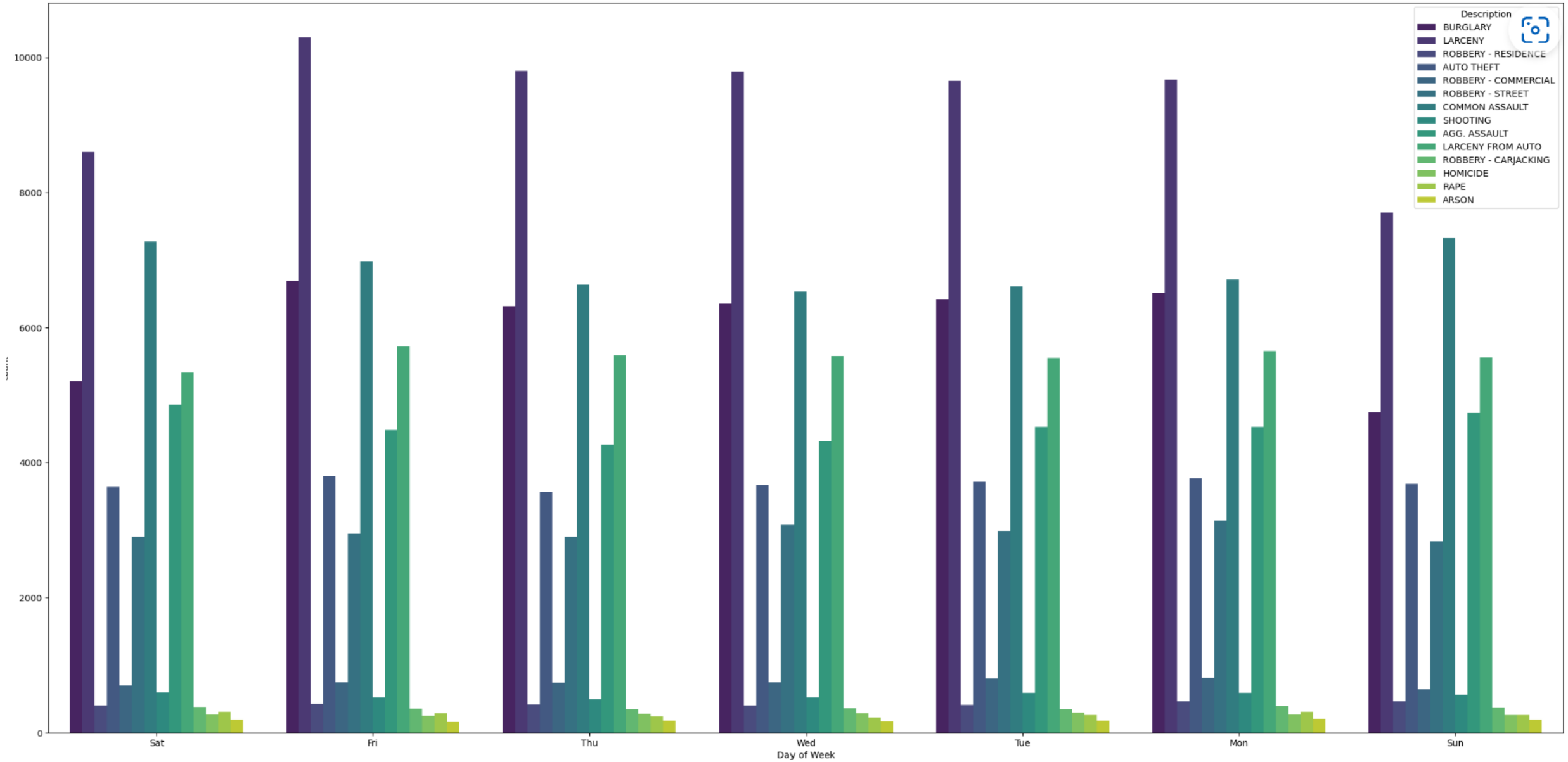
## Alerts

| | |
|---|---|
| `Location 1` has constant value "0.0" | Constant |
| `Total Incidents` has constant value "1" | Constant |
| Dataset has 11311 (3.9%) duplicate rows | Duplicates |
| `CrimeDate` has a high cardinality: 2300 distinct values | High cardinality |
| `CrimeCode` has a high cardinality: 81 distinct values | High cardinality |
| `Longitude` is highly correlated with `Latitude` | High correlation |
| `Latitude` is highly correlated with `Longitude` | High correlation |
| `CrimeCode` is highly correlated with `Description` | High correlation |
| `Location 1` is highly correlated with `CrimeCode` and 3 other fields | High correlation |
| `Total Incidents` is highly correlated with `CrimeCode` and 3 other fields | High correlation |
| `Description` is highly correlated with `CrimeCode` | High correlation |
| `District` is highly correlated with `Location 1` and 1 other fields | High correlation |
| `CrimeTime` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |
| `Location` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |
| `Inside/Outside` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |
| `Weapon` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |
| `Post` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |
| `Neighborhood` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |
| `Premise` is an unsupported type, check if it needs cleaning or further analysis | Unsupported |

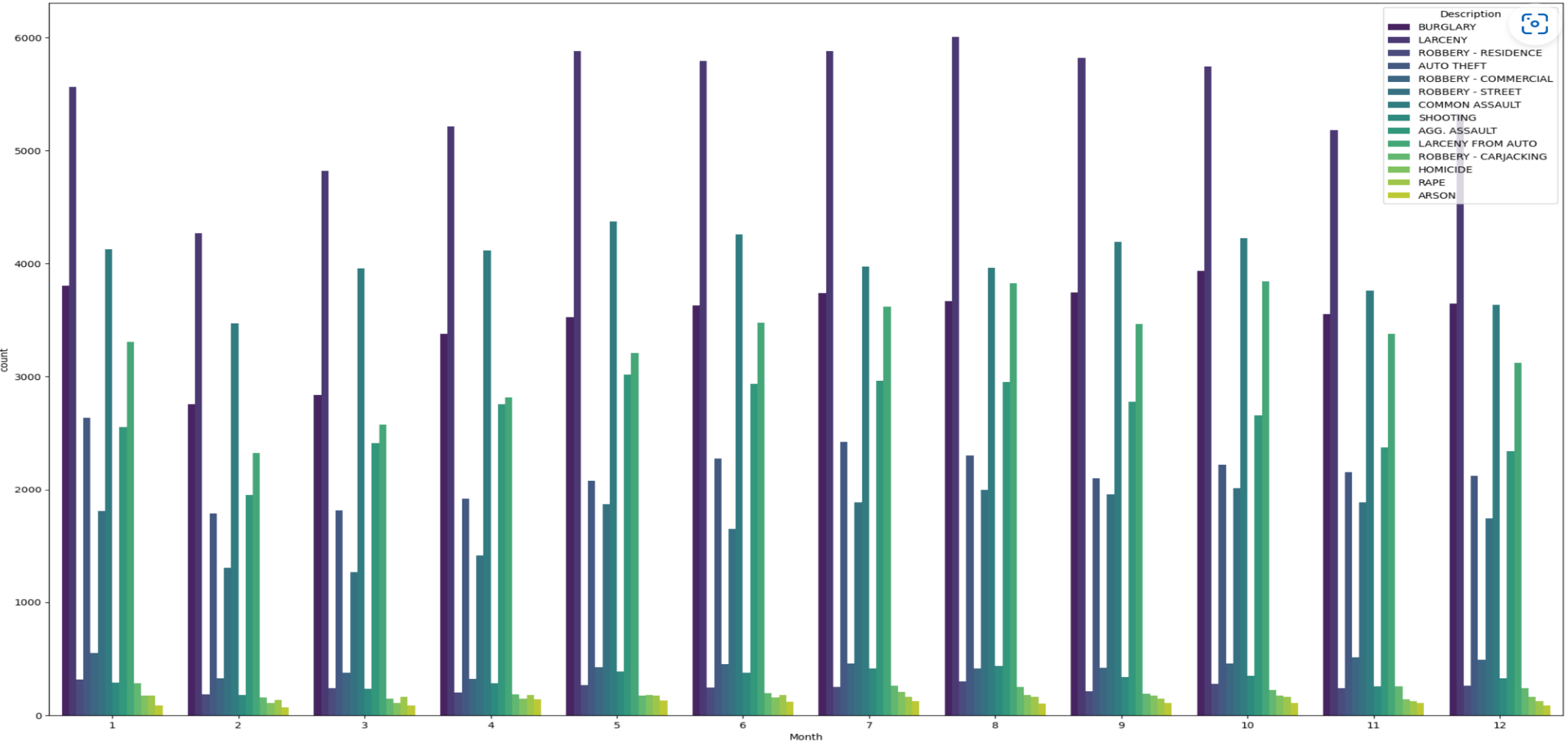
# Overall Crime Distribution



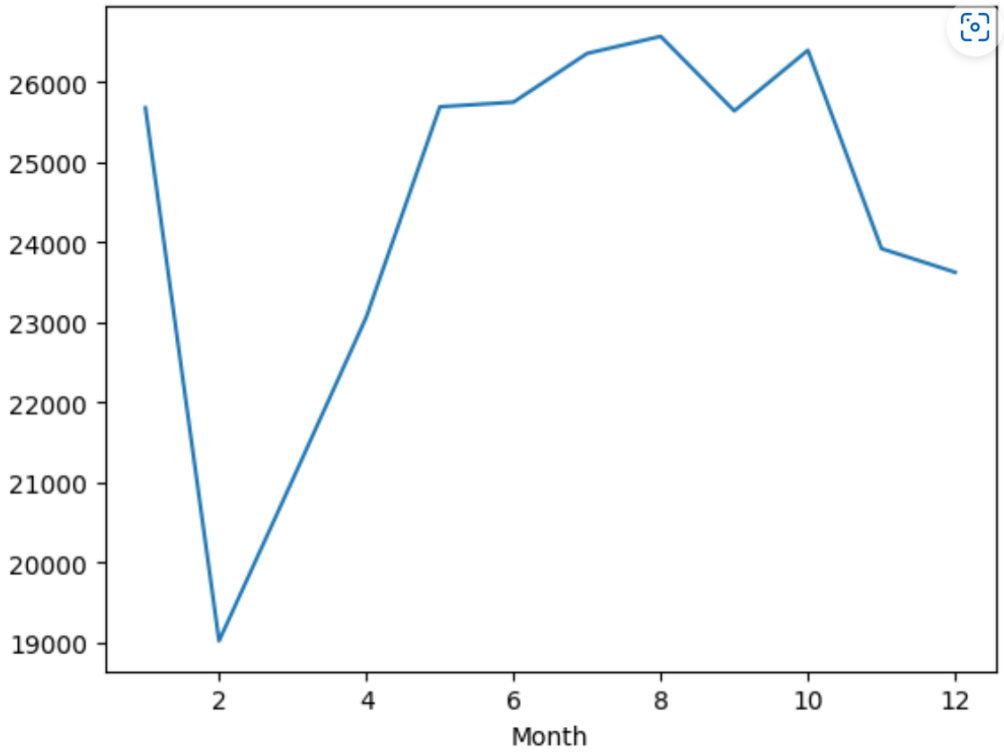
# Overall Crime Distribution by day-of-week

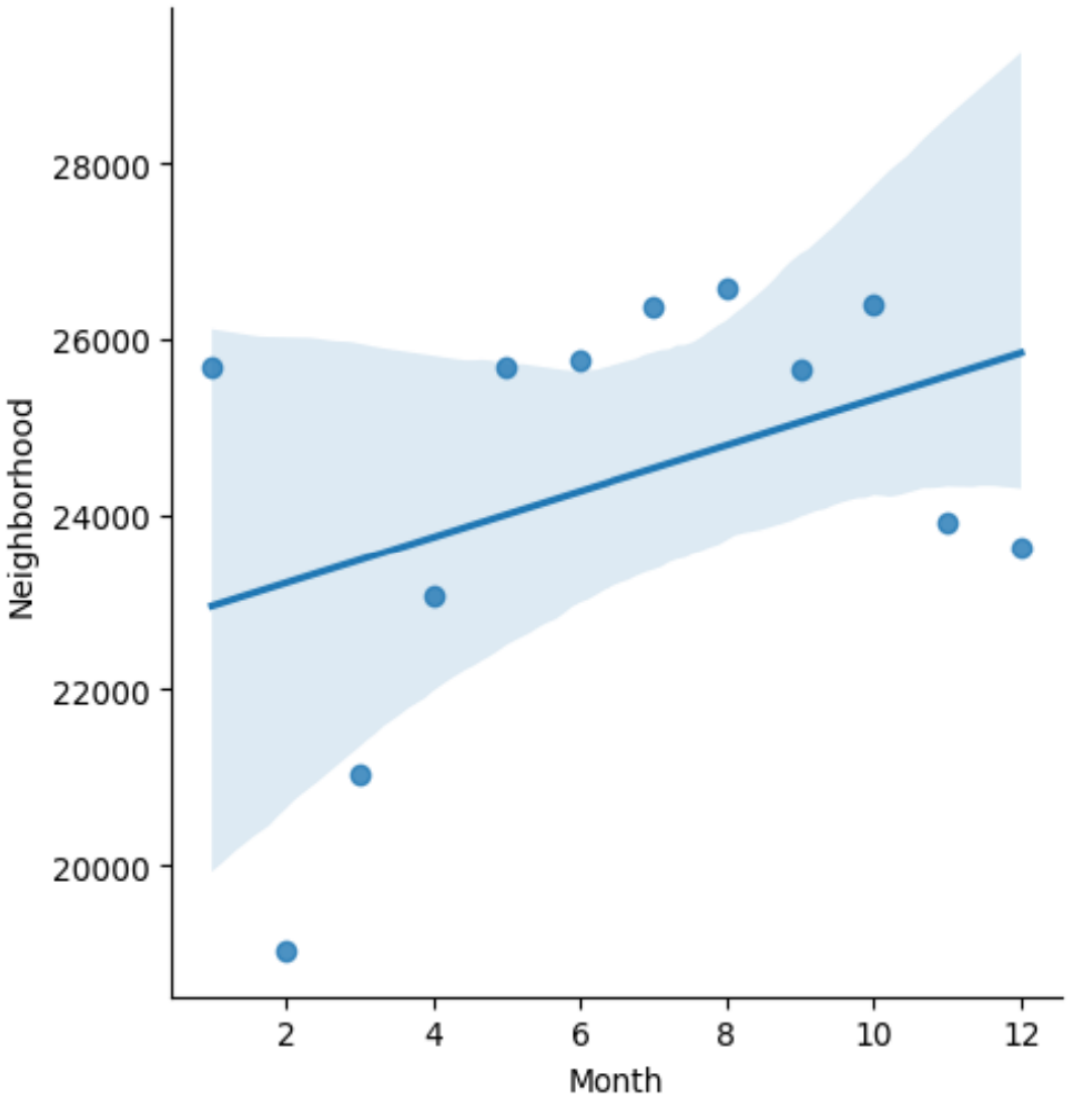


# Overall Crime Distribution By Month



# Crime over months in Neighborhood



# Conclusion

- As the amount of raw data rises, so does the amount of data that is not intrinsically usable; this increases the amount of effort required cleaning and organizing data before it can be evaluated, which is where data wrangling comes into play. The outcome of data wrangling can give crucial metadata statistics for deeper insights into the data; nevertheless, it is critical to guarantee metadata consistency otherwise it might present obstacles.
- Data wrangling enables analysts to evaluate more complicated data faster, generate more accurate findings, and make better judgments as a consequence. Because of its success, many firms have shifted to data wrangling.

# References


- As coder is for code, X is for data Archived 2021-04-15 at the Wayback Machine, Open Knowledge Foundation blog post
- Parsons, M. A.; Brodzik, M. J.; Rutter, N. J. (2004). "Data management for the Cold Land Processes Experiment: improving hydrological science". Hydrological Processes. **18** (18): 3637–3653. Bibcode:2004HyPr...18.3637P. doi:10.1002/hyp.5801.
- "What Is Data Wrangling? What are the steps in data wrangling?". Express Analytics. 2020-04-22. Archived from the original on 2020-11-01. Retrieved 2020-12-06.
- Wickham, Hadley; Grolemund, Garrett (2016). "Chapter 9: Data Wrangling Introduction". R for data science : import, tidy, transform, visualize, and model data (First ed.). Sebastopol, CA. ISBN 978-1491910399. Archived from the original on 2021-10-11. Retrieved 2022-01-12.
- Kandel, Sean; Paepcke, Andreas (May 2011). "Wrangler: Interactive Visual Specification of Data Transformation Scripts". SIGCHI. doi:10.1145/1978942.1979444. S2CID 11133756.