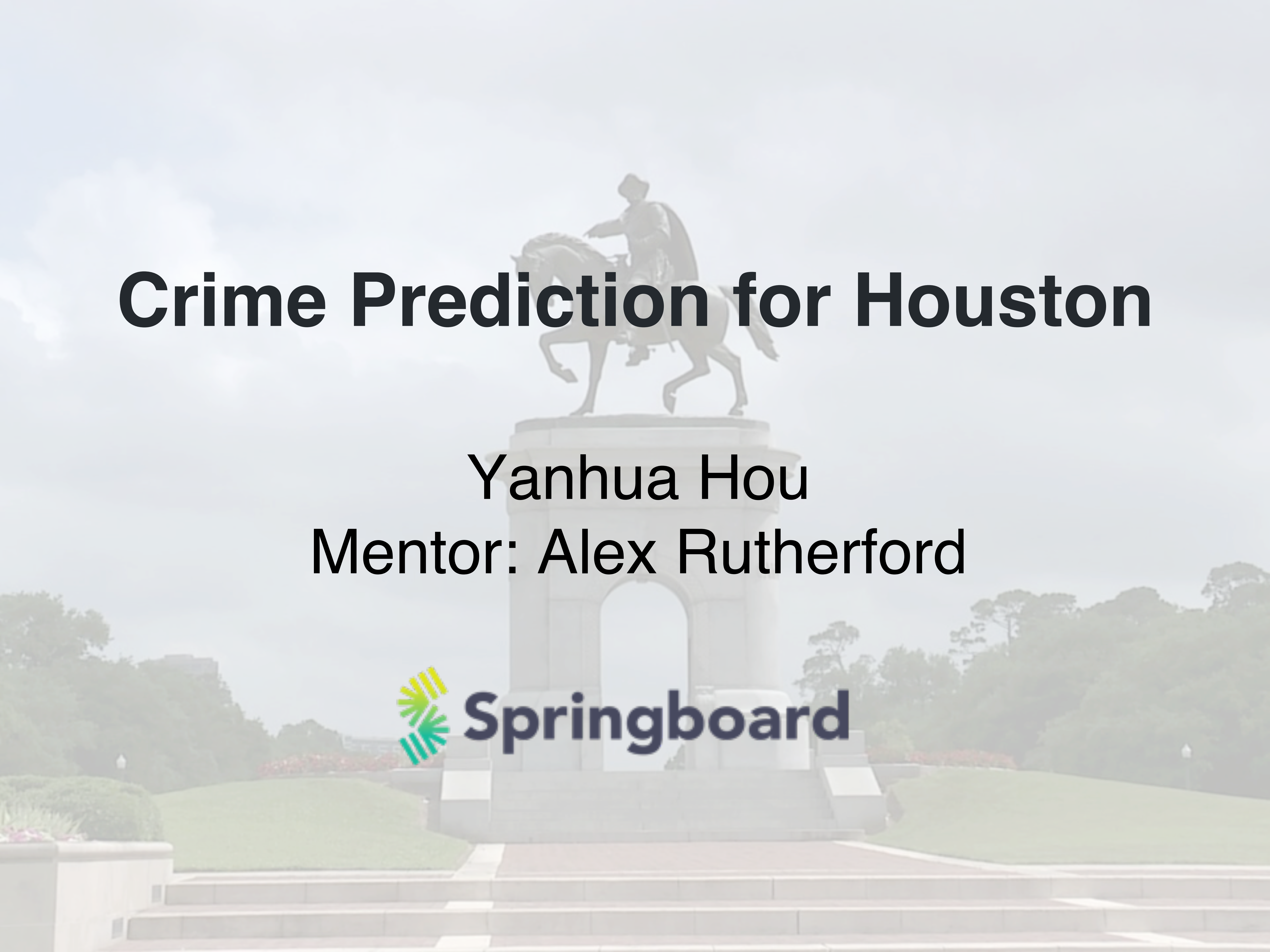


Crime Prediction for Houston

Yanhua Hou
Mentor: Alex Rutherford

Springboard

Outlines

- Problems and Clients
- Data Acquisition and Data Cleaning
- Exploratory Data Analysis
- Machine Learning Models and Predictions
- Conclusions and Future Work

Problems

- How safe is the city we live in regarding crime rates?
- What are the main types of crimes in neighborhoods?
- How has the crime rate changed over past years?
- Which types of crimes have increased and which has decreased and why?

Problems

- Most importantly what message can we take from historical data to reduce crimes?
- In case of crimes how can we be more prepared to minimize losses?

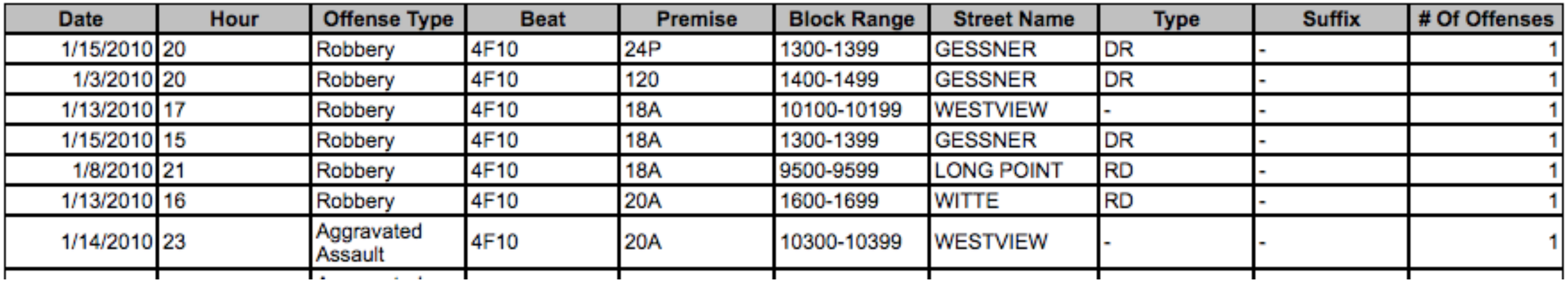
?

Clients

- Police board, government, and general public would be beneficiaries.
- Police officers can use this model to be better deployed.
- The government can take precautions more efficiently.
- Residents can use this model to better protect their lives and properties.

Data Acquisition

The crime data is acquired from HPD



Date	Hour	Offense Type	Beat	Premise	Block Range	Street Name	Type	Suffix	# Of Offenses
1/15/2010	20	Robbery	4F10	24P	1300-1399	GESSNER	DR	-	1
1/3/2010	20	Robbery	4F10	120	1400-1499	GESSNER	DR	-	1
1/13/2010	17	Robbery	4F10	18A	10100-10199	WESTVIEW	-	-	1
1/15/2010	15	Robbery	4F10	18A	1300-1399	GESSNER	DR	-	1
1/8/2010	21	Robbery	4F10	18A	9500-9599	LONG POINT	RD	-	1
1/13/2010	16	Robbery	4F10	20A	1600-1699	WITTE	RD	-	1
1/14/2010	23	Aggravated Assault	4F10	20A	10300-10399	WESTVIEW	-	-	1

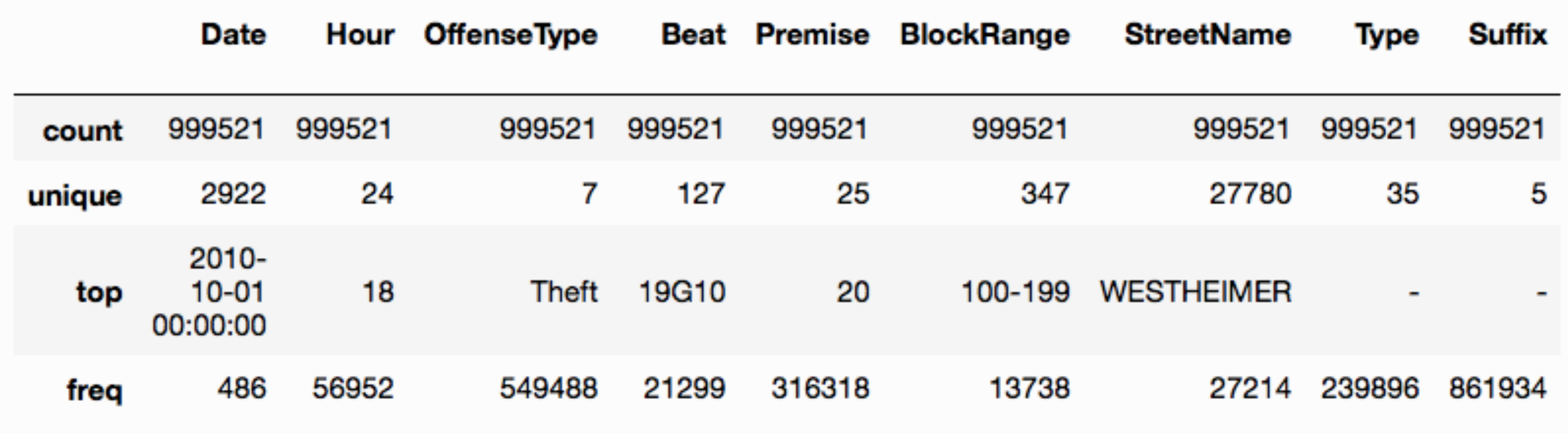
It reports **seven** types of crimes on a monthly basis:

- murder
- rape
- robbery
- aggravated assault
- burglary
- theft
- auto theft

Data Cleaning

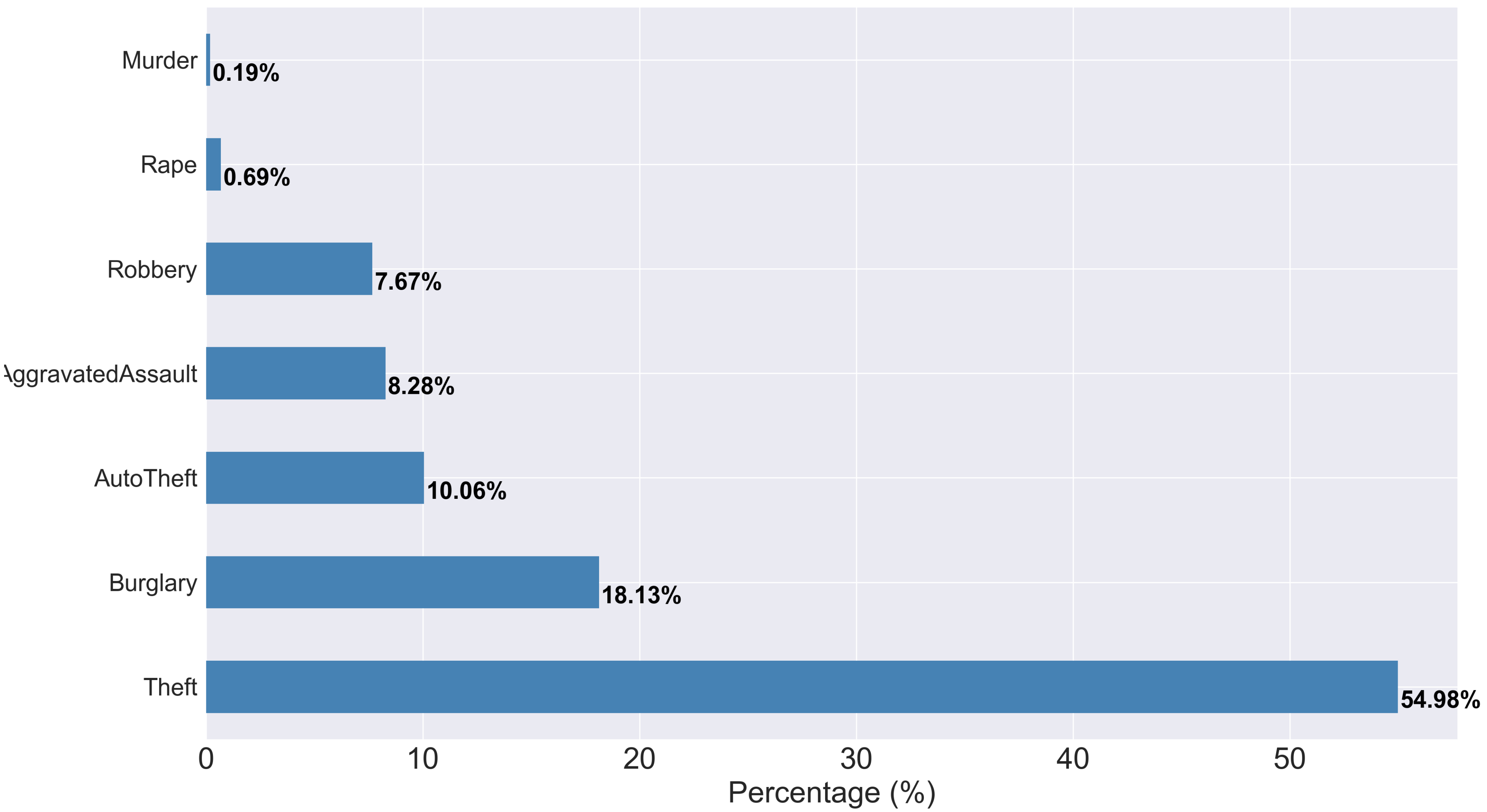
- Drop empty columns, empty rows, and the rows with missing ***‘Date’***.
- Fill columns with missing values by ‘UNK’.
- Select ***Date’*** in the range ‘2010-01-01’ to ‘2017-12-31’.
- Clean column '***Hour***' to contain 24 unique integers 0-23.
- Clean column ‘***OffenseType***’ to contain 7 types of crimes.
- Reduce the number of ‘***Premise***’ from 126 to 25.
- Clean ‘***Beat***’, ‘***Type***’, ‘***Suffix***’, ‘***StreetName***’ by stripping whitespaces.
- Check duplicates and drop duplicated observations.

A summary statistics of ‘object’ features



	Date	Hour	OffenseType	Beat	Premise	BlockRange	StreetName	Type	Suffix
count	999521	999521	999521	999521	999521	999521	999521	999521	999521
unique	2922	24	7	127	25	347	27780	35	5
top	2010-10-01 00:00:00	18	Theft	19G10	20	100-199	WESTHEIMER	-	-
freq	486	56952	549488	21299	316318	13738	27214	239896	861934

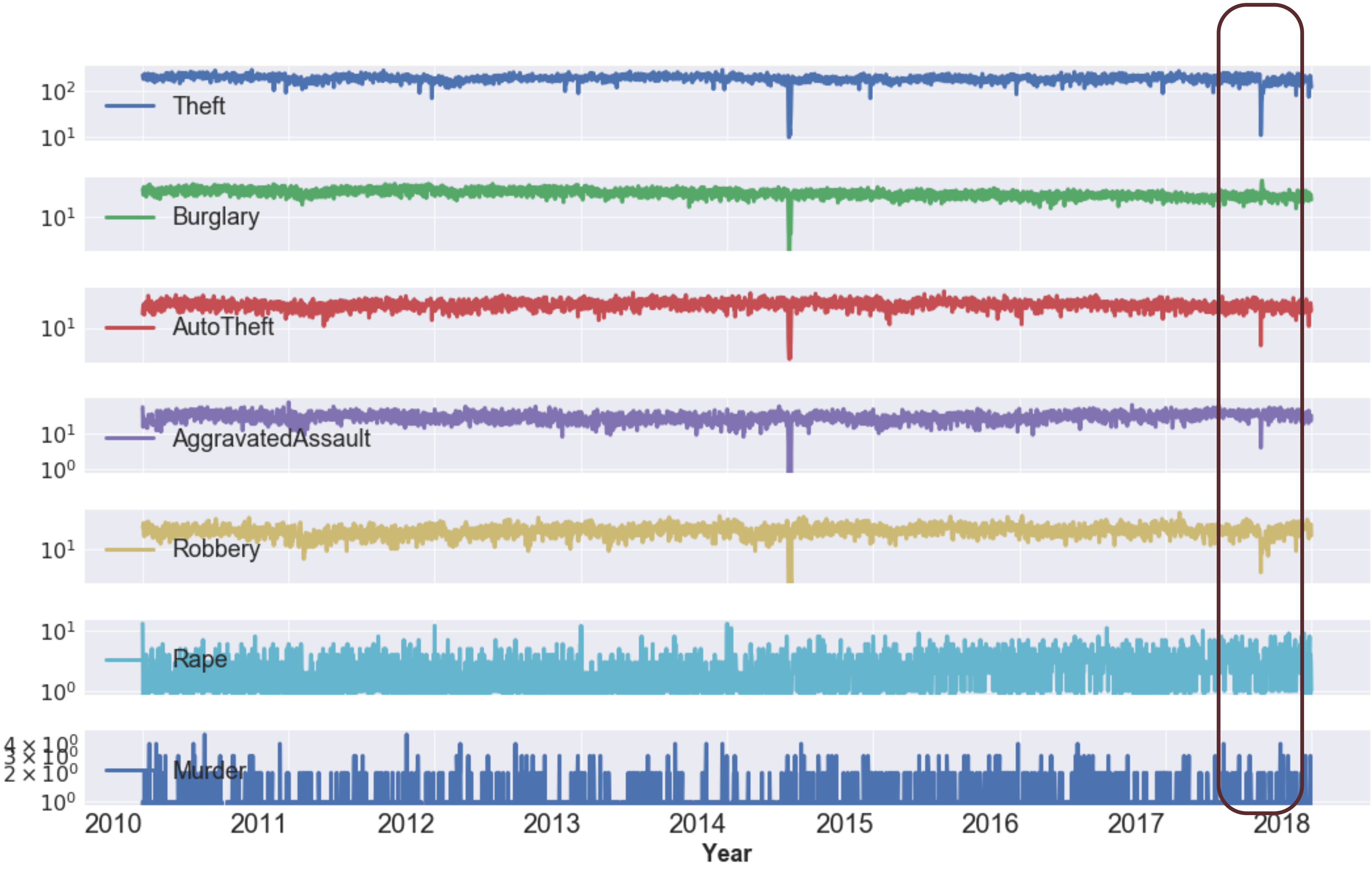
Exploratory Data Analysis



‘Theft’ is the dominant crimes, followed by ‘Burglary’, ‘AutoTheft’, ‘AggravatedAssault’ and ‘Robbery’ while ‘Rape’ and ‘Murder’ only takes a rather small portion.

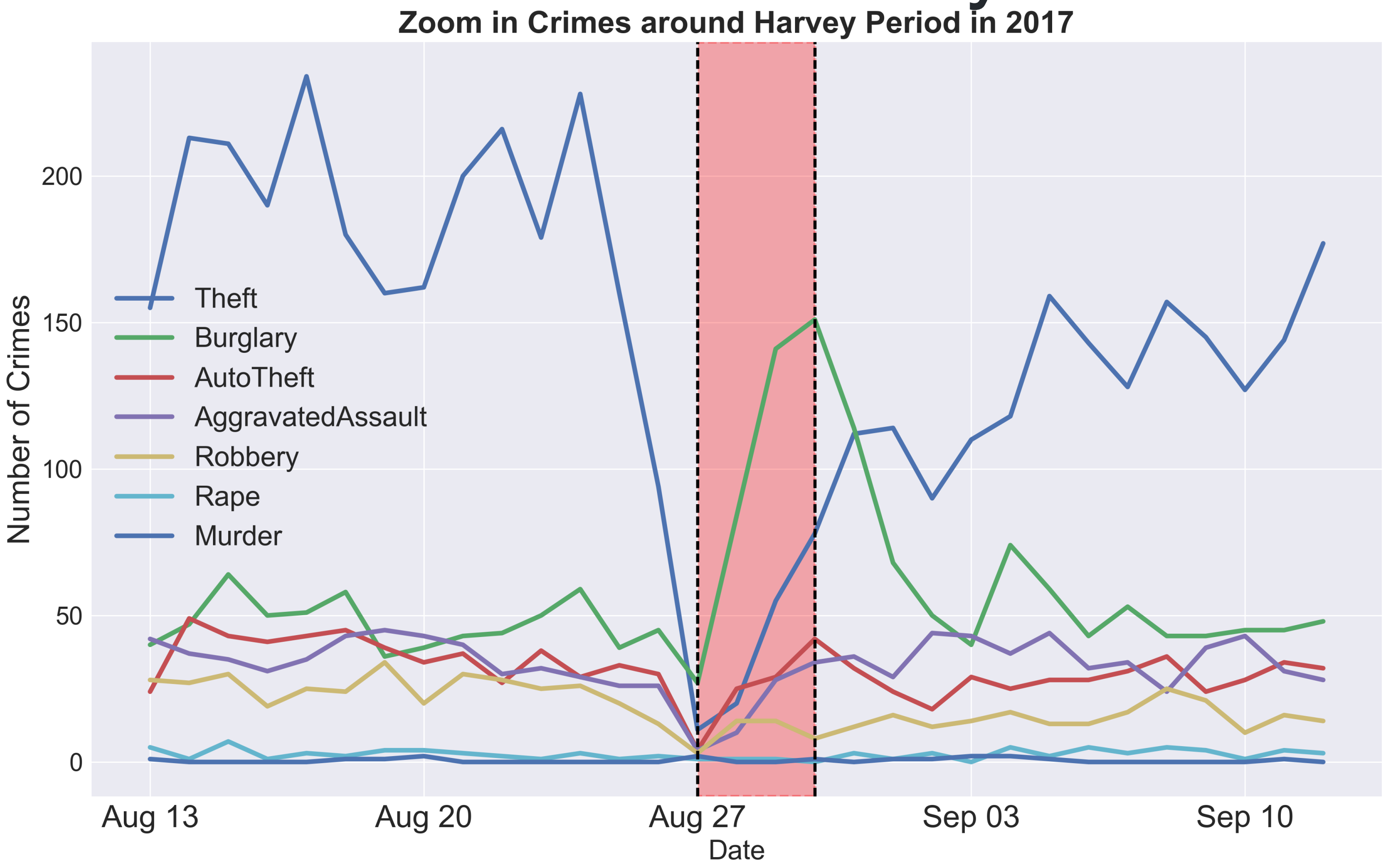
Exploratory Data Analysis

Time series analysis



Exploratory Data Analysis

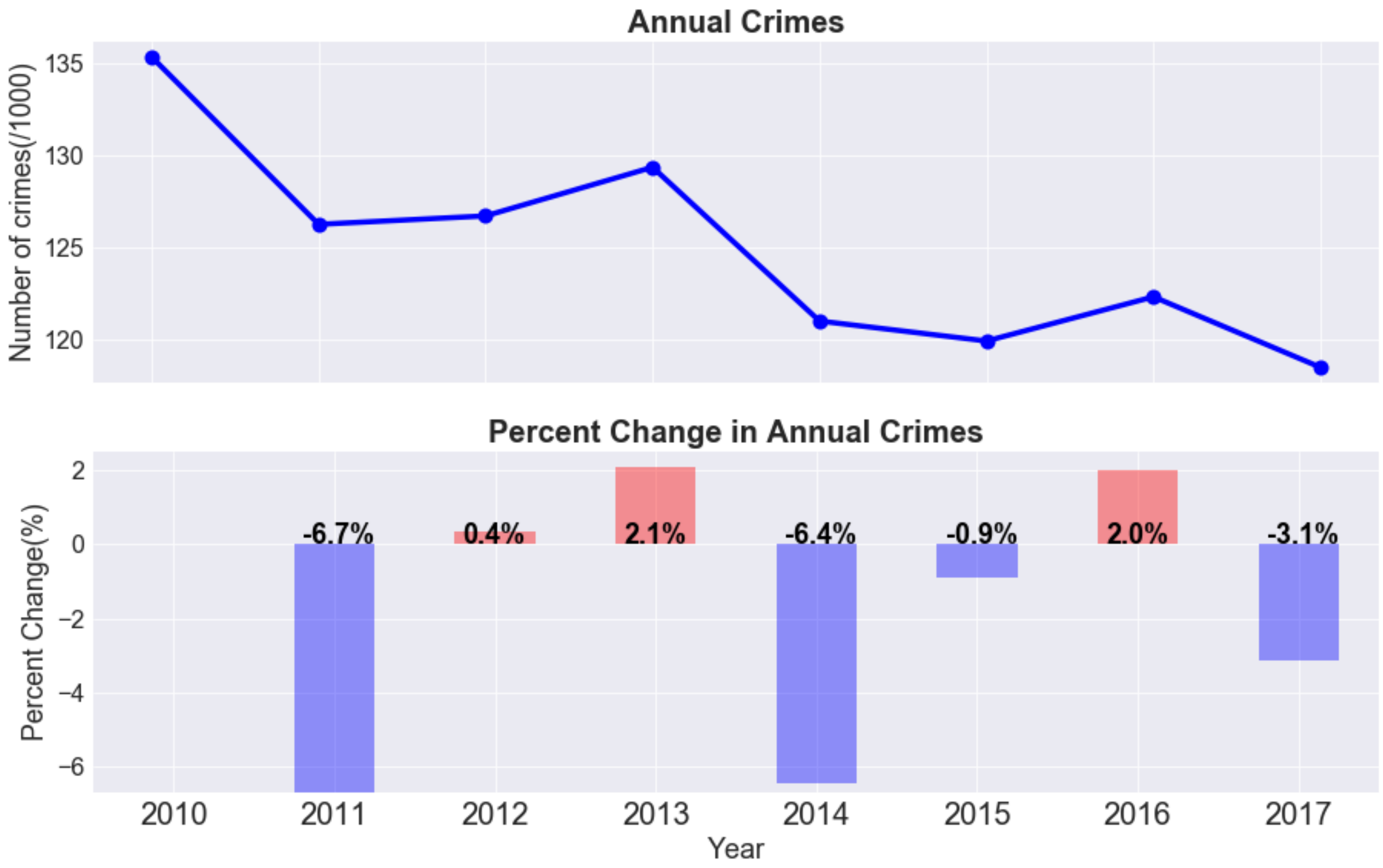
Time series analysis



In the hurricane Harvey Period, crimes dropped overall. However, 'Burglary' quickly increased followed by 'Theft', 'AutoTheft', 'Robbery' and 'Assault'.

Exploratory Data Analysis

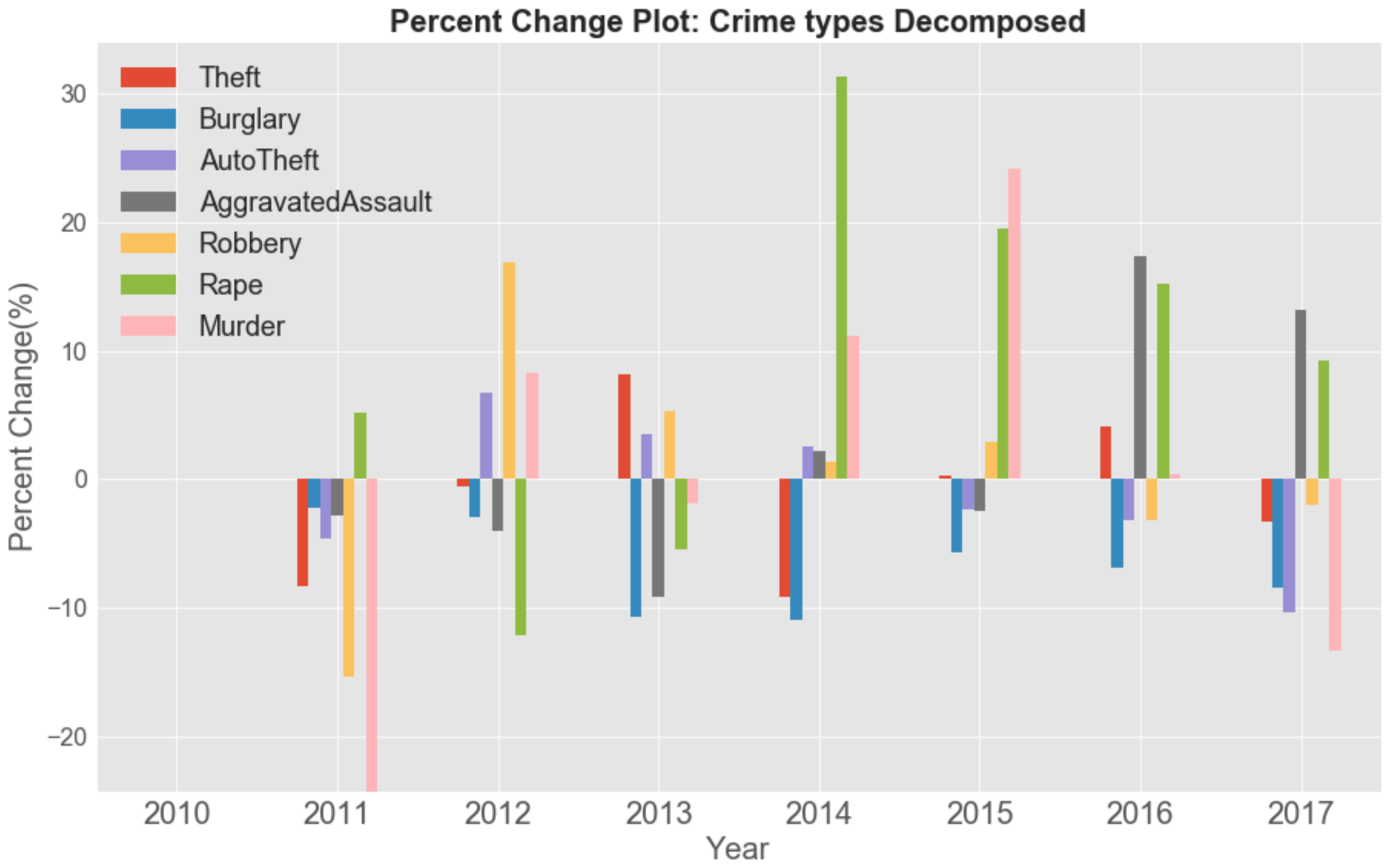
Time series analysis: year



The crime trend: overall crimes in Houston have a decreasing trend in 2017 compared to 2016.

Exploratory Data Analysis

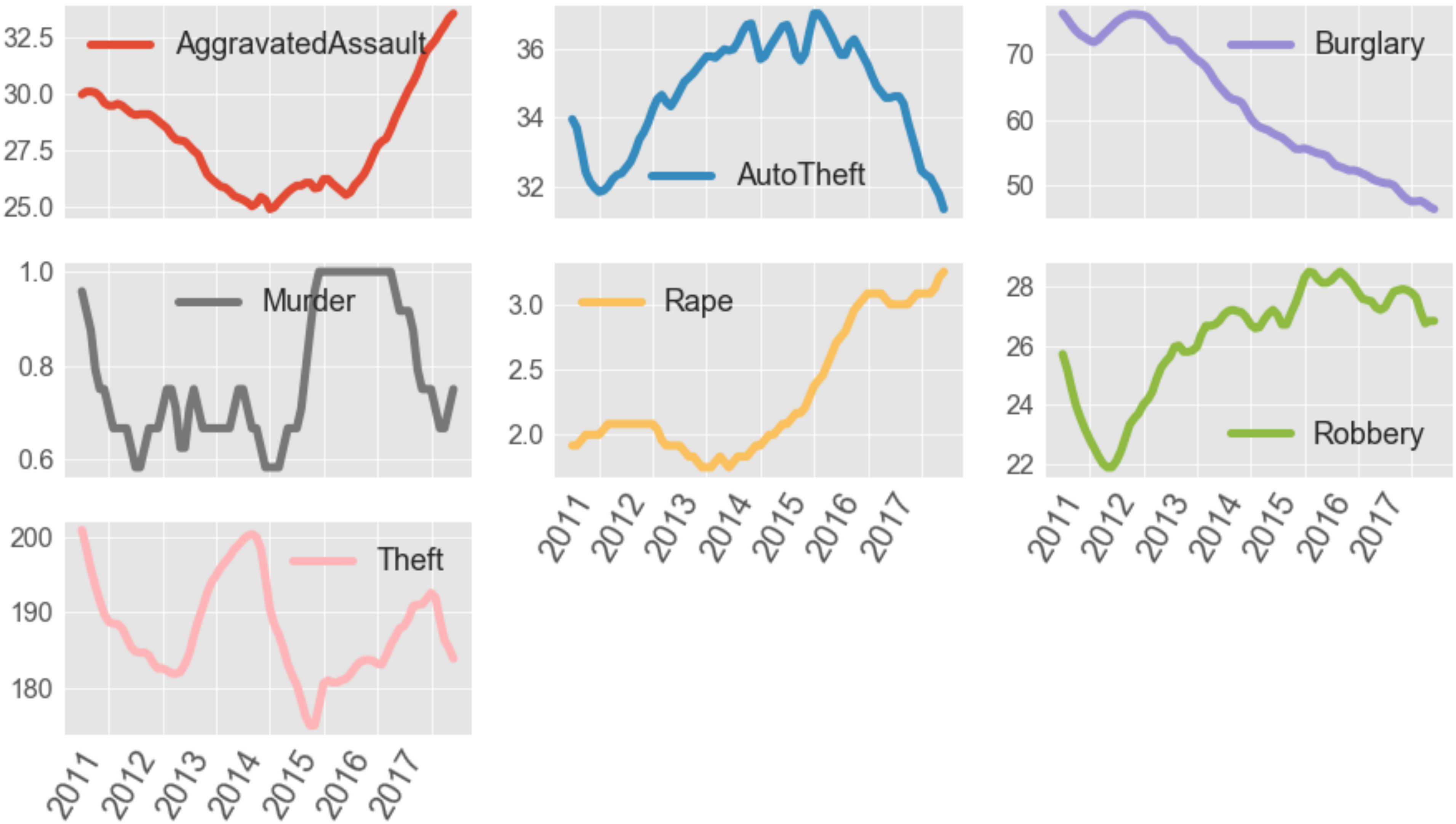
Time series analysis: year



The crime types decomposed: violent crimes like ‘AggravatedAssault’ and ‘Rape’ increased.

Exploratory Data Analysis

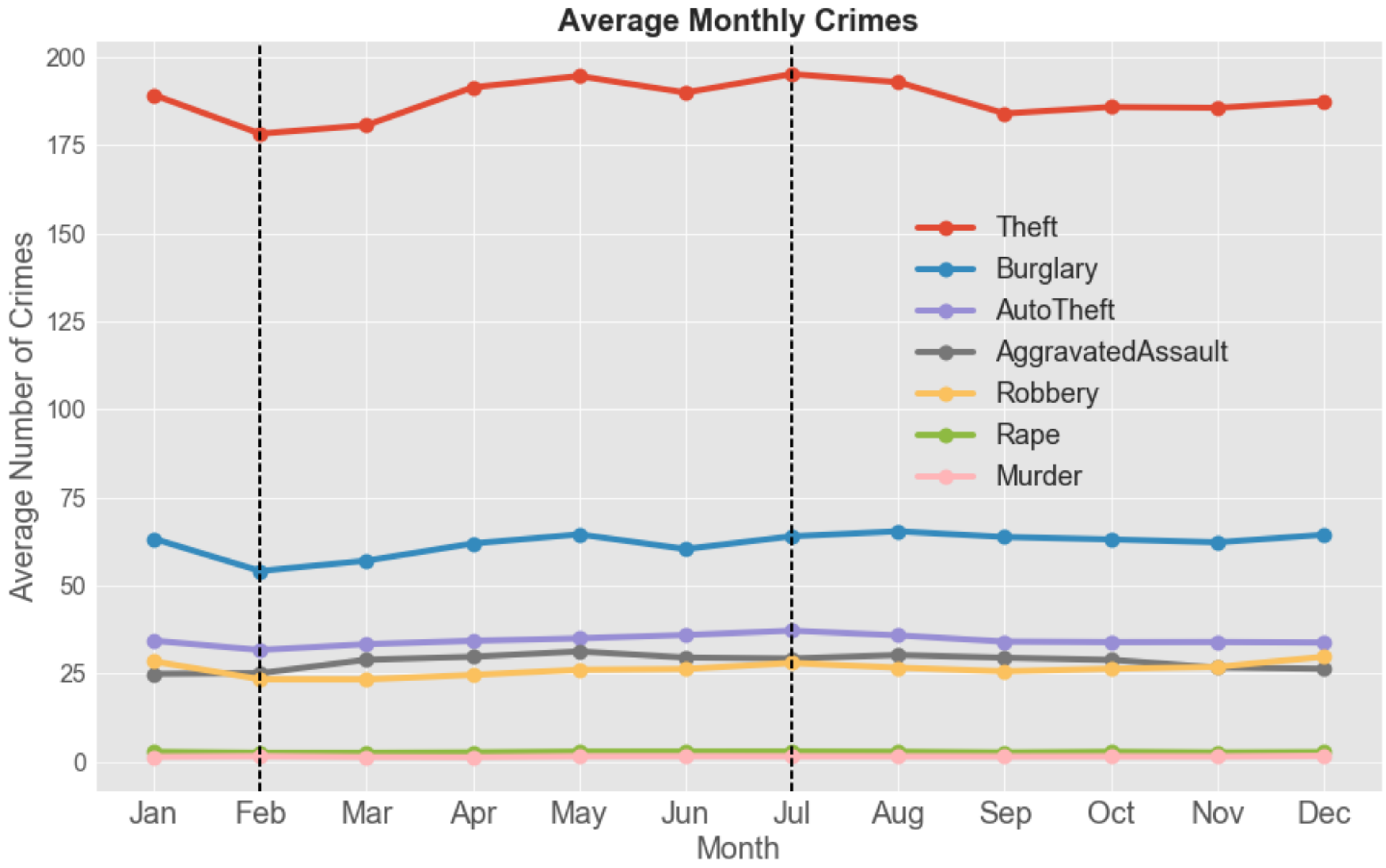
Time series analysis: trends



We apply the moving average model to smooth out short-term fluctuations and highlight long-term trends.

Exploratory Data Analysis

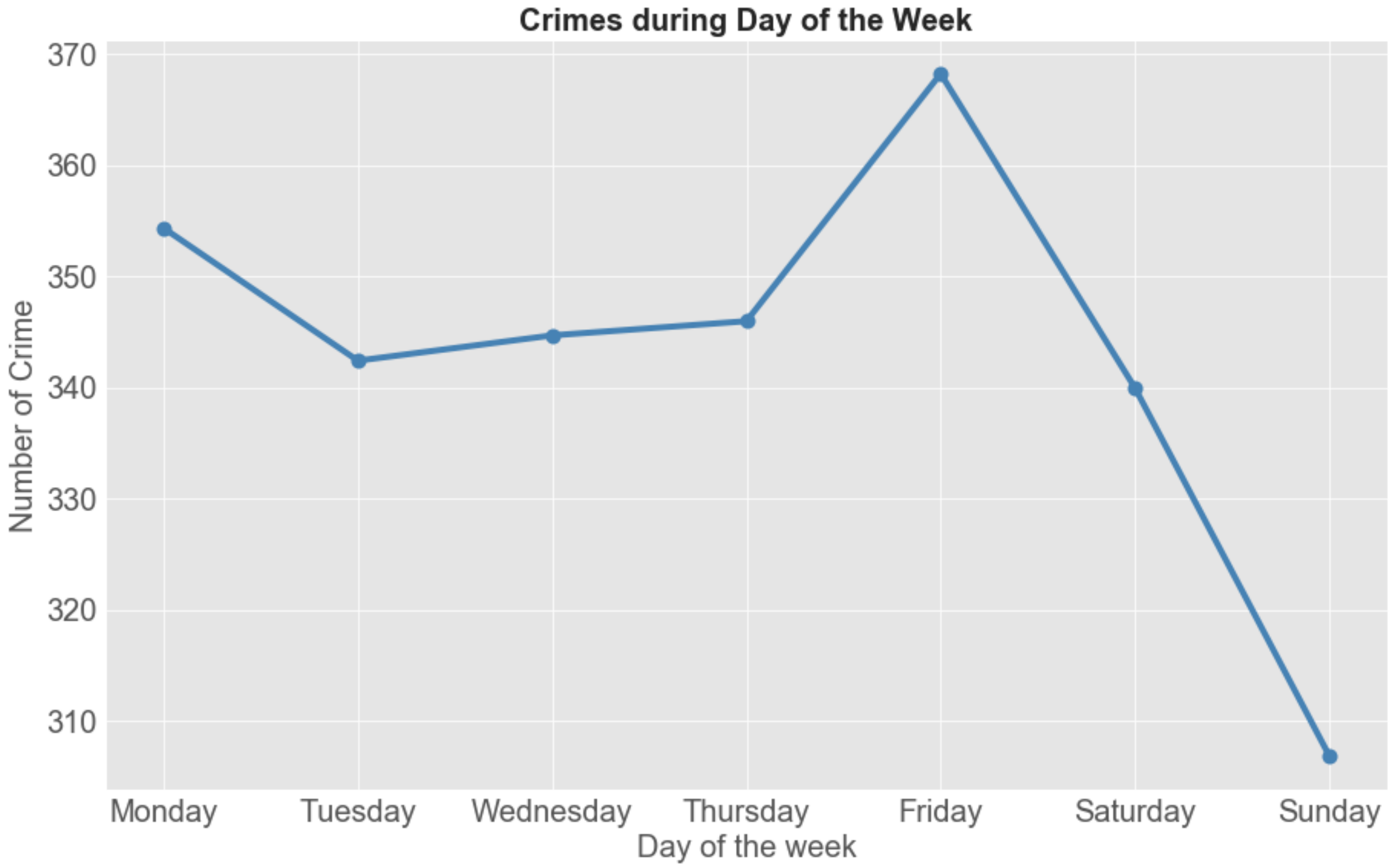
Time series analysis: month



Month tends: summer months have more occurrences than winter months

Exploratory Data Analysis

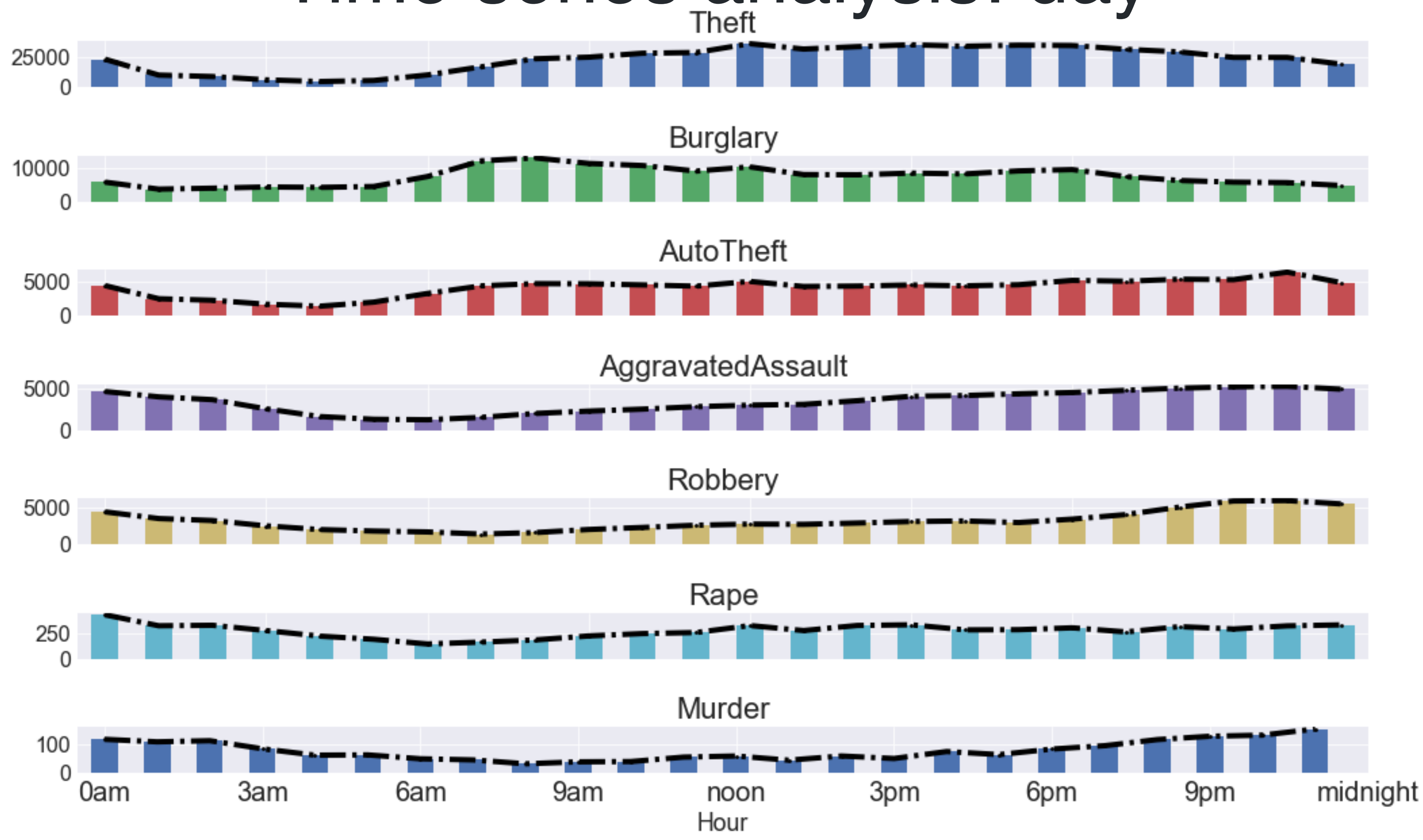
Time series analysis: week



Day tends: there are most crimes on Fridays and least crimes on Sundays on a weekly basis.

Exploratory Data Analysis

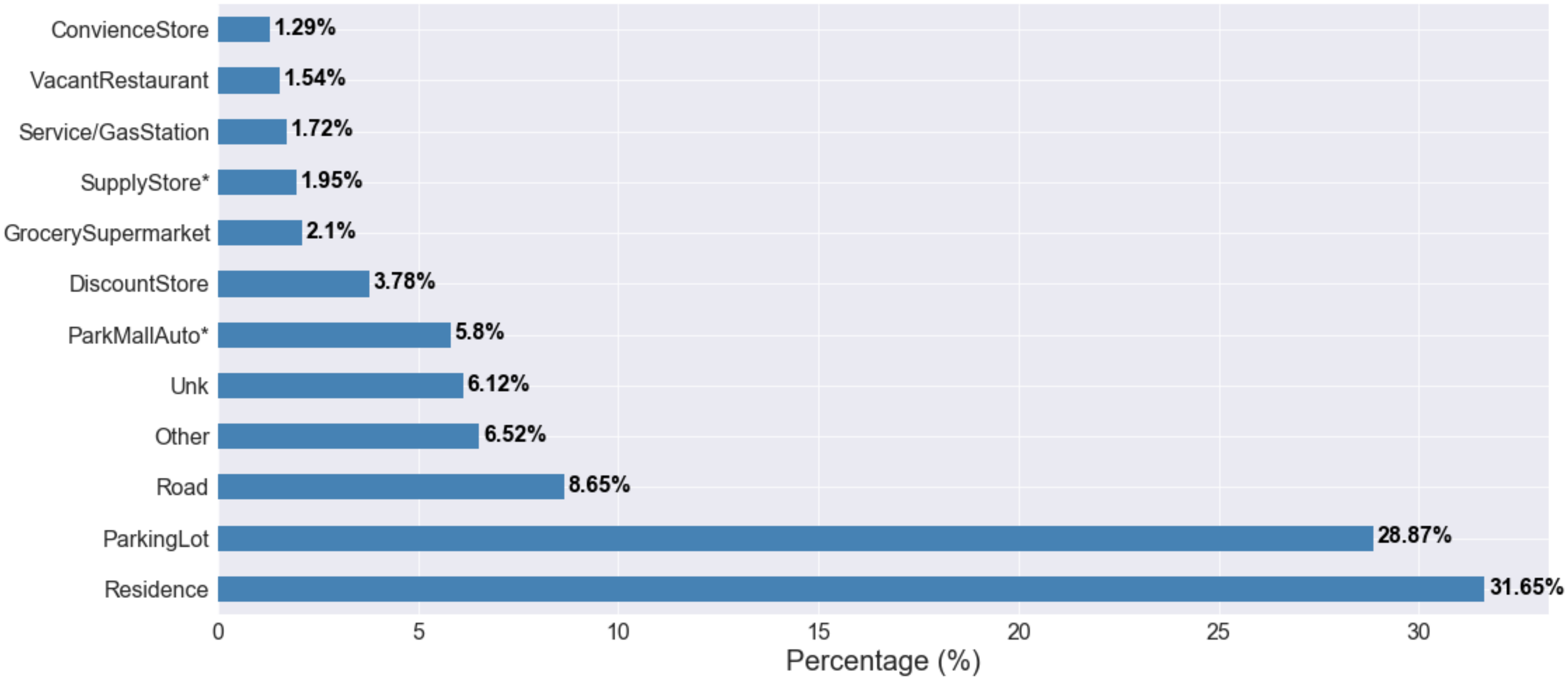
Time series analysis: day



‘Theft’ hits a peak in the middle of day; ‘Burglary’ peaks in the early morning; ‘Auto theft’, ‘Aggravated Assault’, ‘Robbery’ and ‘Murder’ peaks around late or middle night.

Exploratory Data Analysis

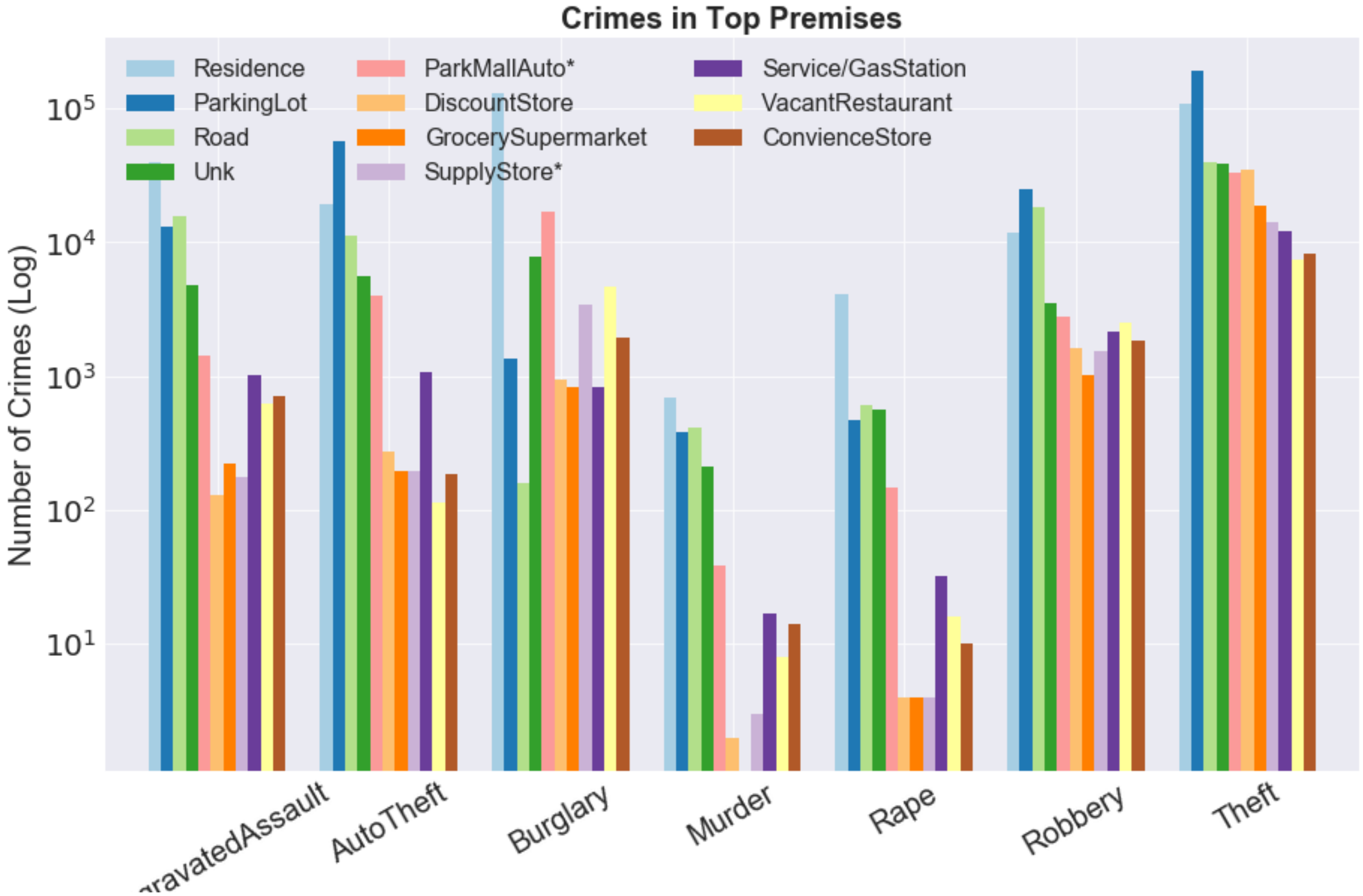
Geographical aspects analysis



The residential place and parking lot had the most frequent crimes.

Exploratory Data Analysis

Geographical aspects analysis

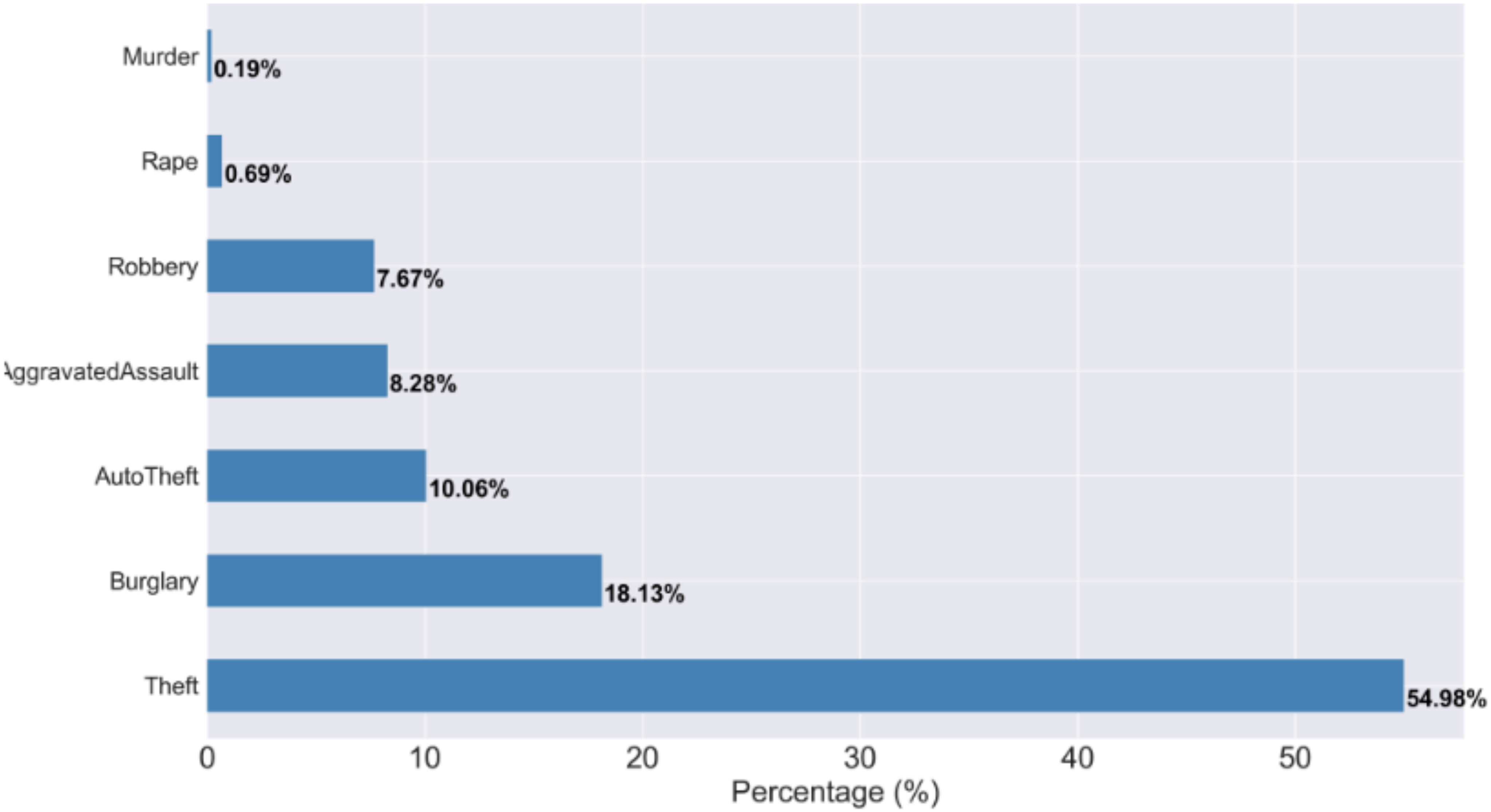


In ‘residence’, dominant crimes include ‘Burglary’, ‘Theft’; Note that violent crimes like ‘Aggravated Assault’ and ‘Rape’ tend to occur most often in ‘residence’ than other premises.

Machine Learning Models

Challenge

I**mbalanced multi-classification** problem with mostly **categorical features**.

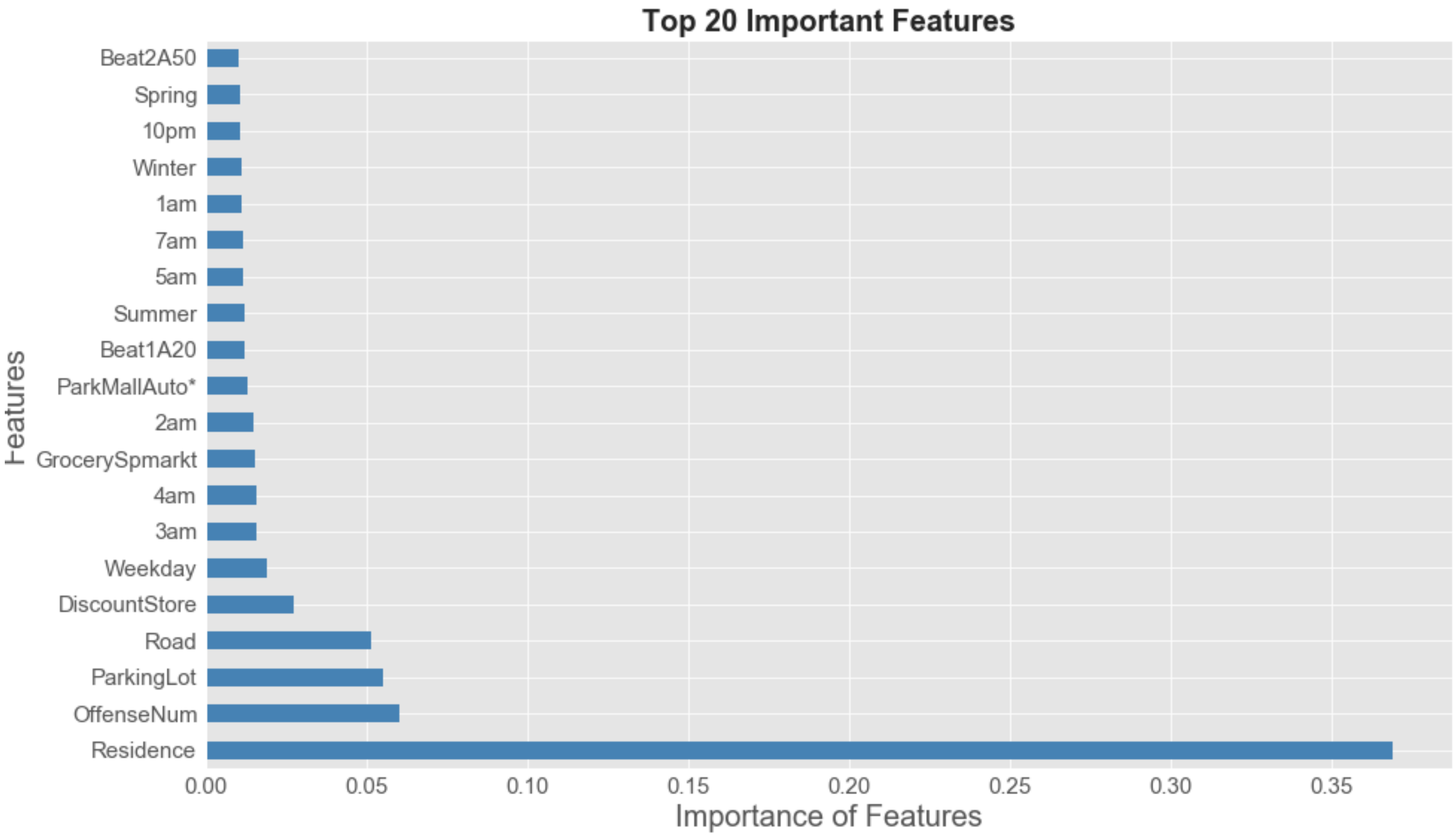


Machine Learning Models

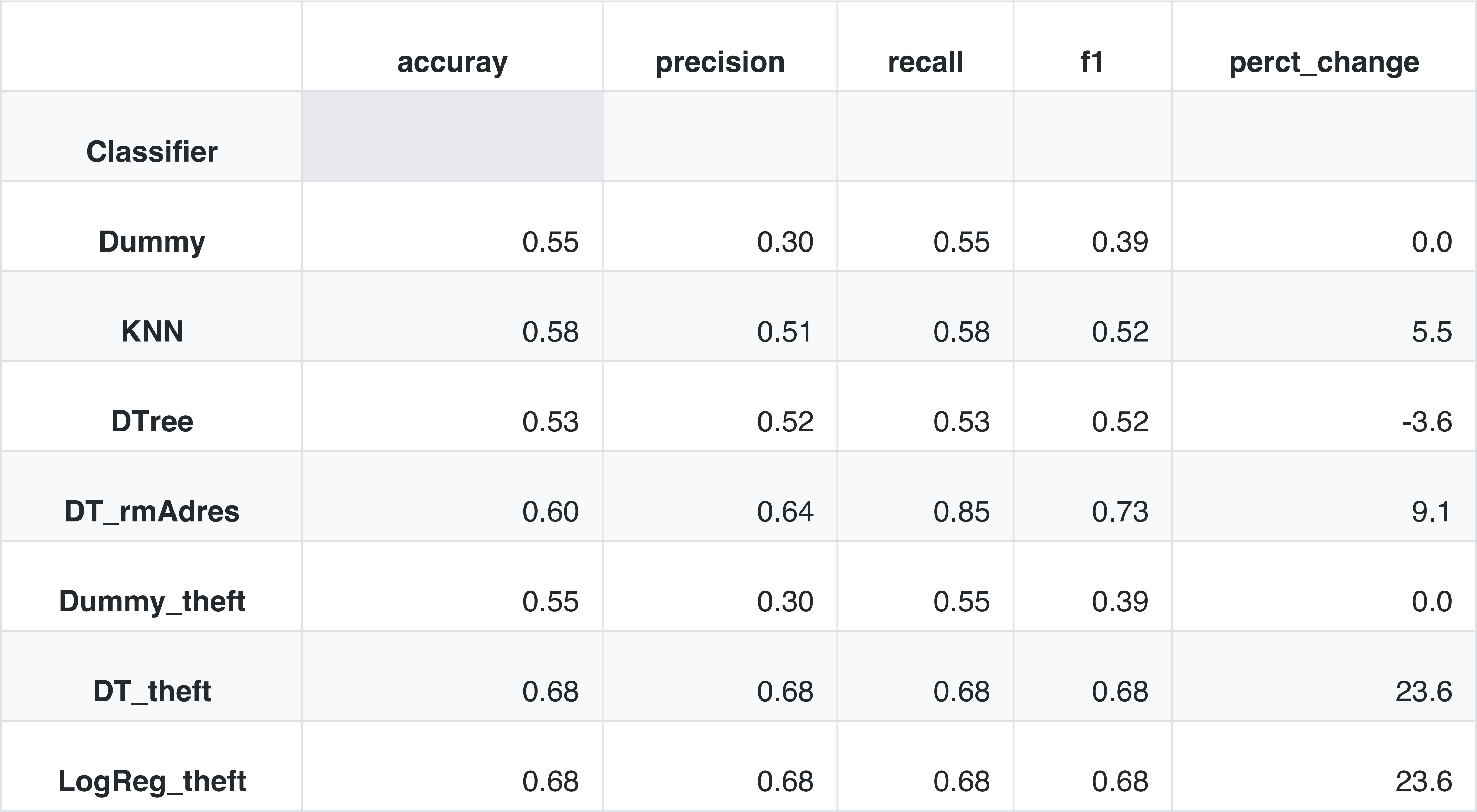
Feature Engineering

- Concatenate ‘***StreetName***’ and ‘***Type***’ columns to ‘***Address***’.
- Drop ‘***StreetName***’, ‘***Type***’ and ‘***Suffix***’.
- Truncate ‘***Beat***’ to keep first 60%, rename others as ‘Other’.
- Similar processing for ‘***Premise***’.
- Add columns ‘***Season***’, ‘***WeekDay***’.
- One-hot/label encode categorical variables.

Machine Learning Models

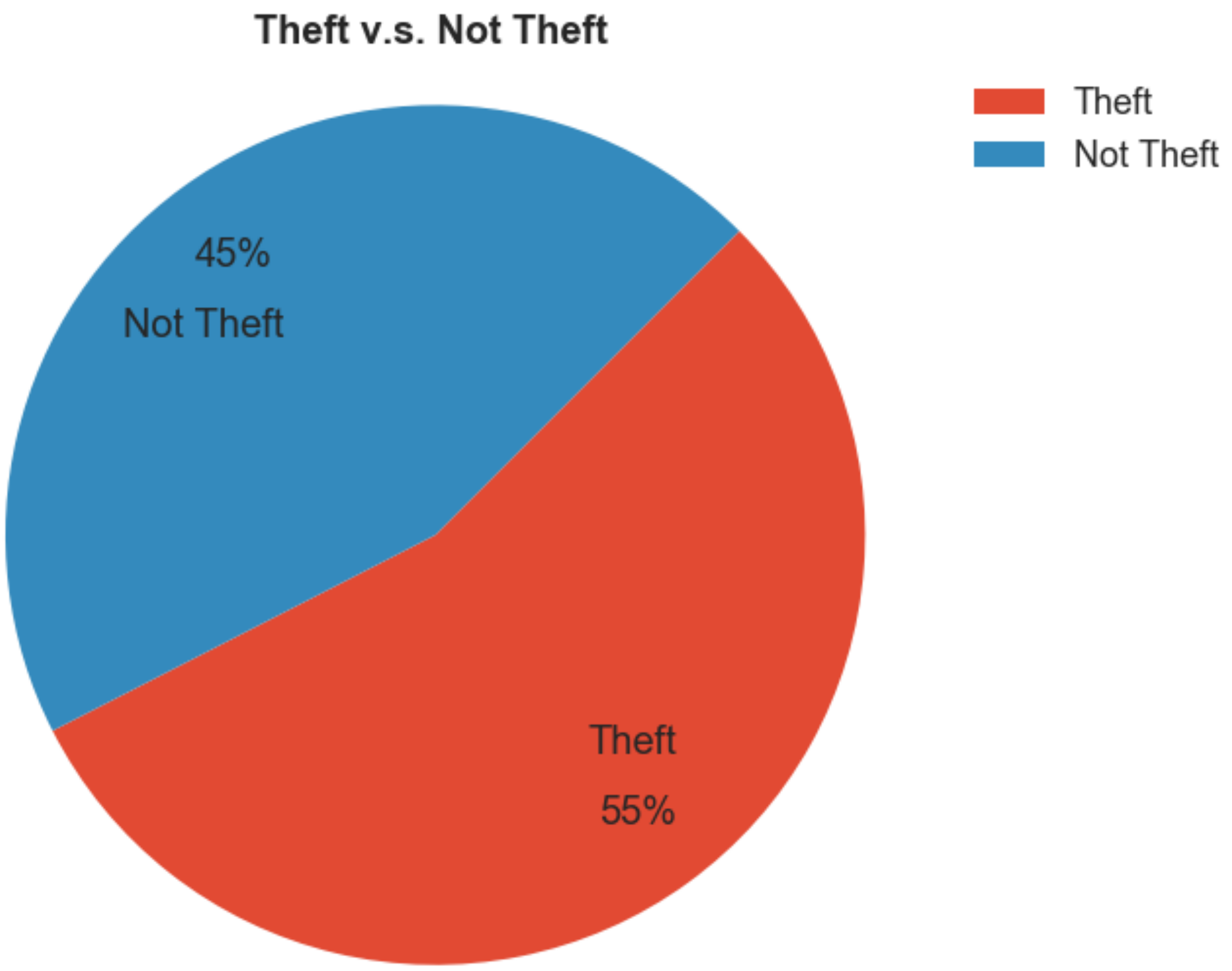


Machine Learning Models



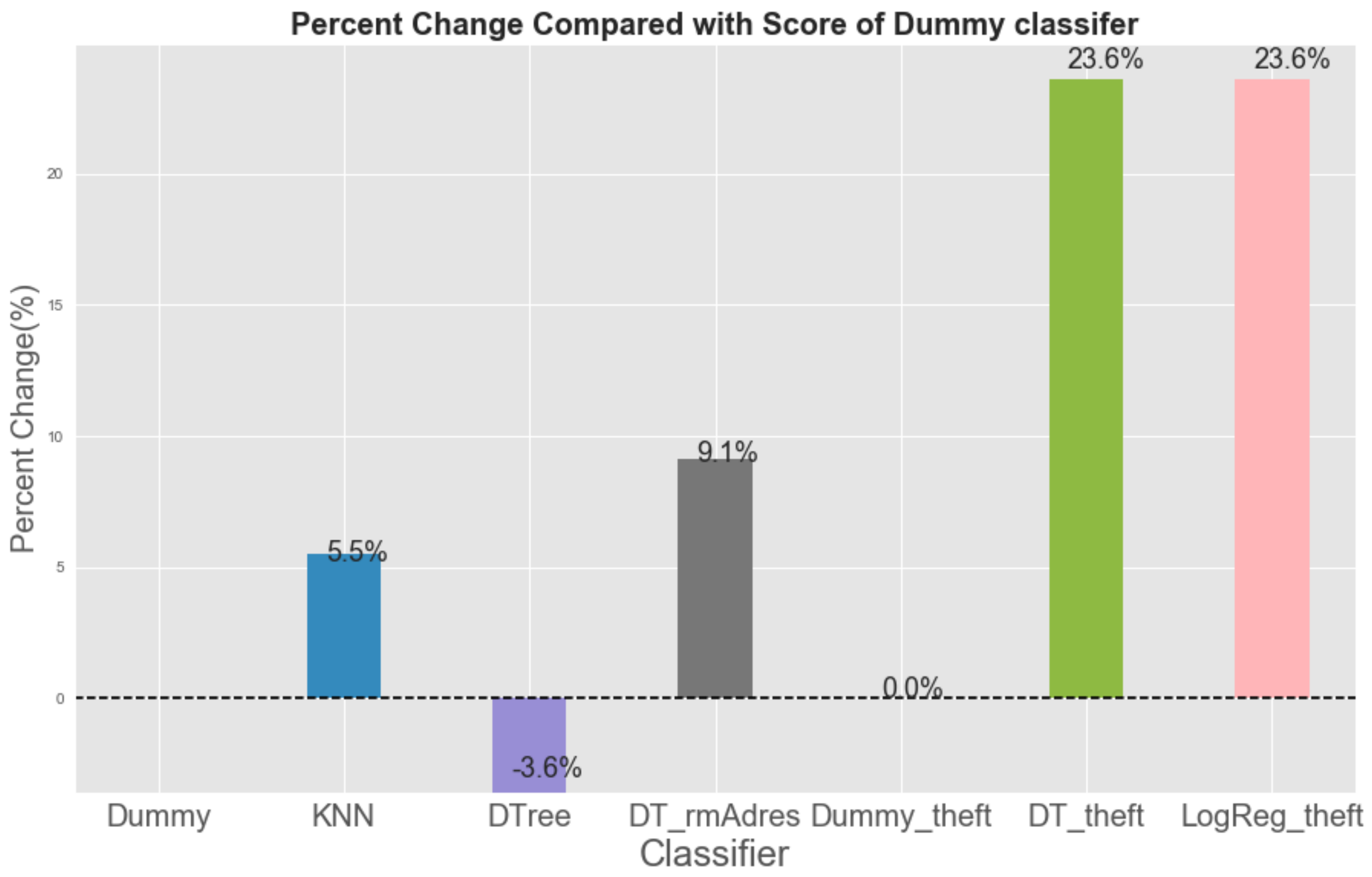
	accuray	precision	recall	f1	perct_change
Classifier					
Dummy	0.55	0.30	0.55	0.39	0.0
KNN	0.58	0.51	0.58	0.52	5.5
DTree	0.53	0.52	0.53	0.52	-3.6
DT_rmAdres	0.60	0.64	0.85	0.73	9.1
Dummy_theft	0.55	0.30	0.55	0.39	0.0
DT_theft	0.68	0.68	0.68	0.68	23.6
LogReg_theft	0.68	0.68	0.68	0.68	23.6

Machine Learning Models

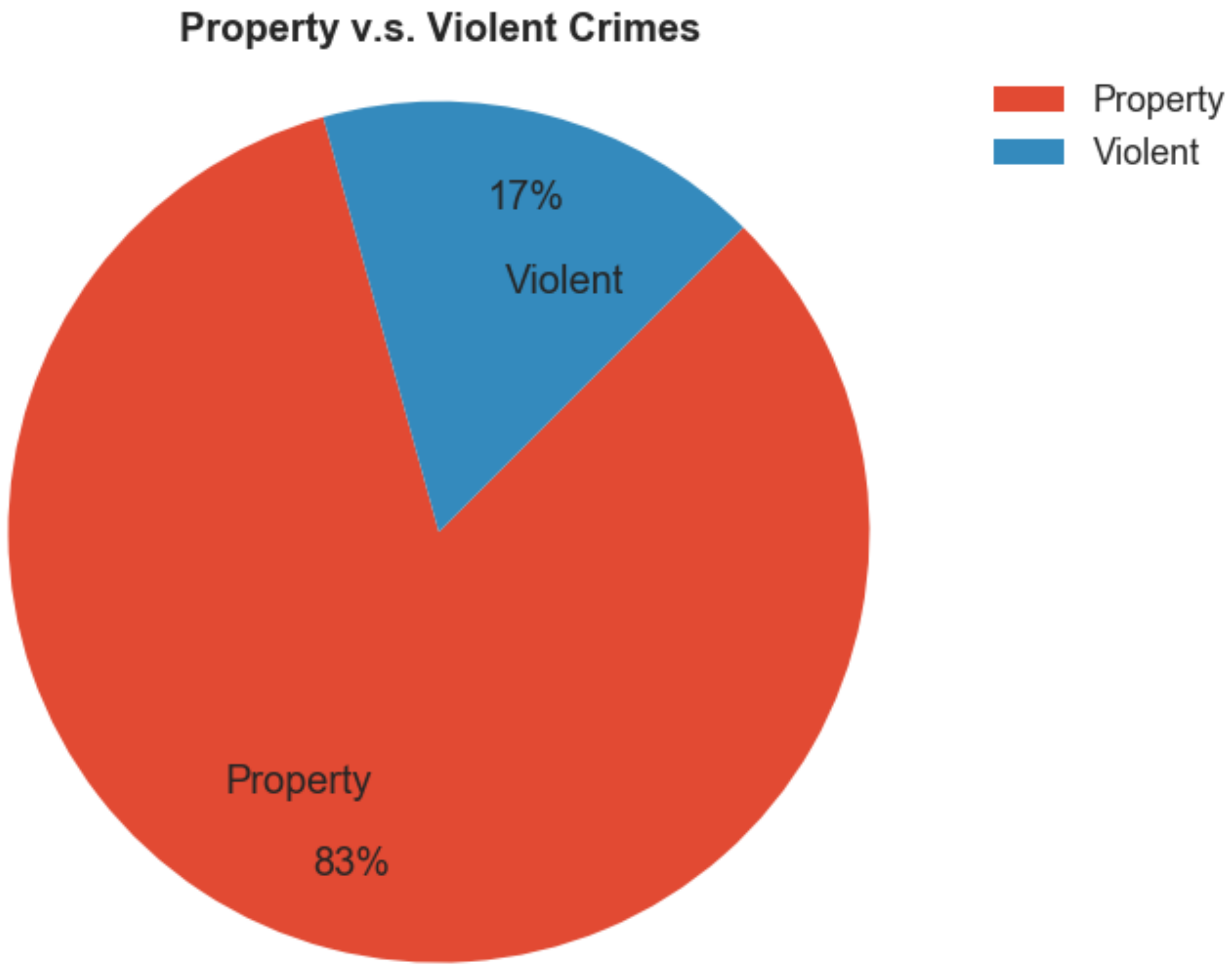


55% is ‘Theft’ and 45% is ‘Not Theft’. We still use the ‘most_frequent’ strategy in dummy classifier. It gives an accuracy of 55%.

Machine Learning Models

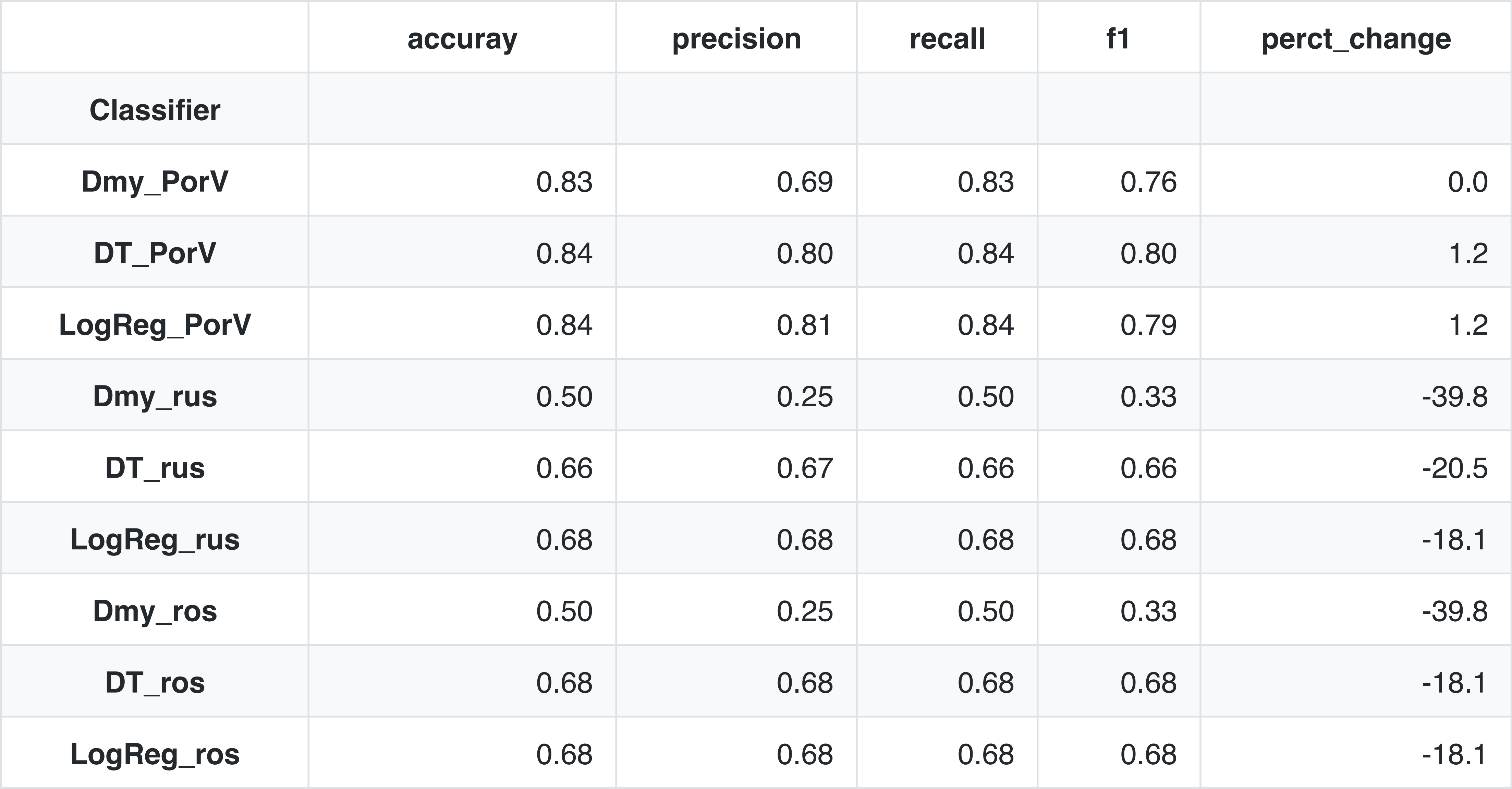


Violent/Property Crime



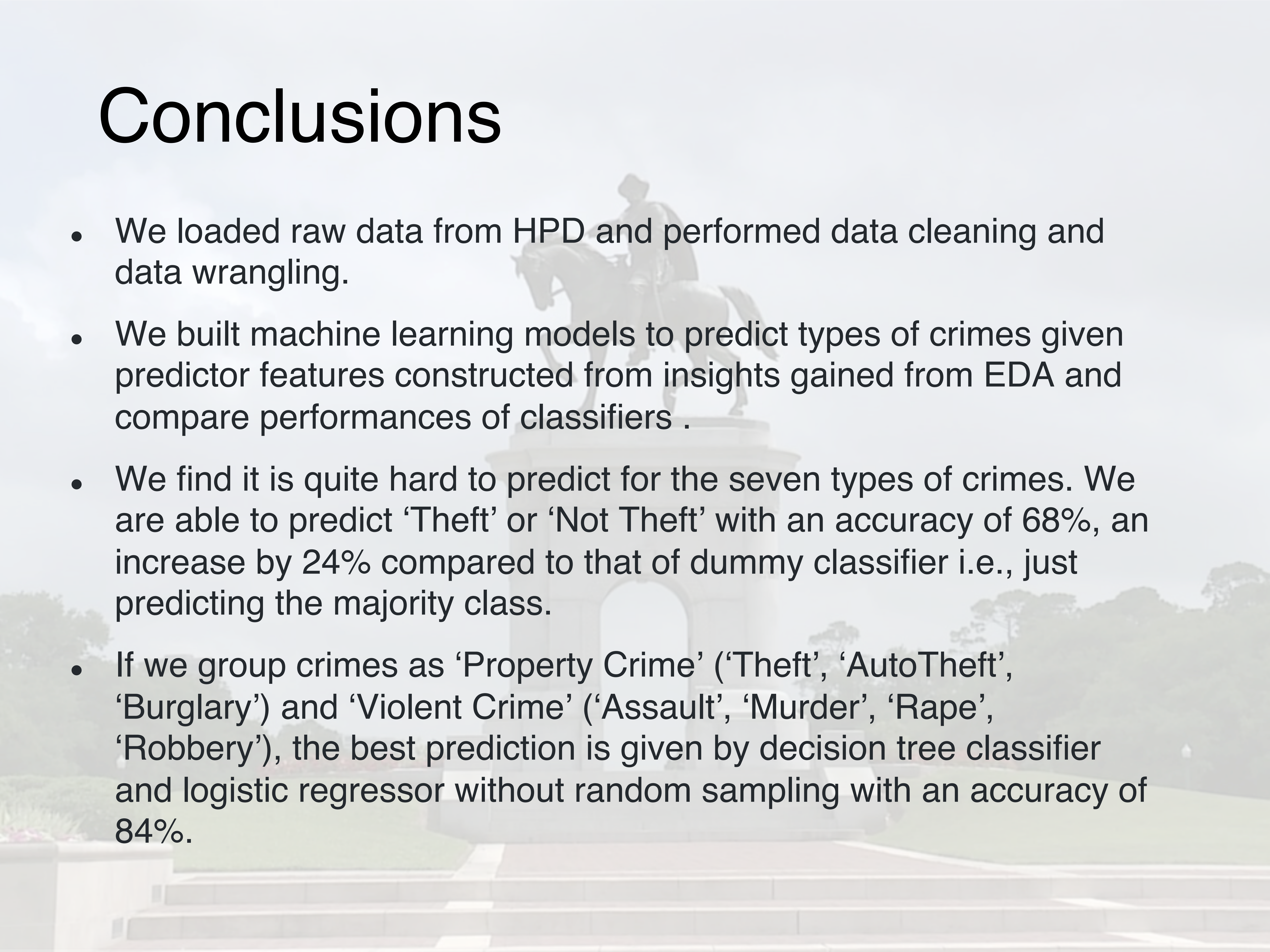
‘Theft’, ‘AutoTheft’, ‘Burglary’ are considered as property crimes, taking a proportion of 83% and ‘Robbery’, ‘AggravatedAssault’, ‘Murder’, ‘Rape’ are considered as violent crimes, taking a proportion of 17%.

Predictions



	accuray	precision	recall	f1	perct_change
Classifier					
Dmy_PorV	0.83	0.69	0.83	0.76	0.0
DT_PorV	0.84	0.80	0.84	0.80	1.2
LogReg_PorV	0.84	0.81	0.84	0.79	1.2
Dmy_rus	0.50	0.25	0.50	0.33	-39.8
DT_rus	0.66	0.67	0.66	0.66	-20.5
LogReg_rus	0.68	0.68	0.68	0.68	-18.1
Dmy_ros	0.50	0.25	0.50	0.33	-39.8
DT_ros	0.68	0.68	0.68	0.68	-18.1
LogReg_ros	0.68	0.68	0.68	0.68	-18.1

‘*Dmy_PorV*’ is the dummy classifier with predicting majority class strategy for property or violent crime; ‘*DT_PorV*’ is the decision tree classifier; ‘*rus*’ stands for random undersampling; ‘*ros*’ stands for random oversampling.



Conclusions

- We loaded raw data from HPD and performed data cleaning and data wrangling.
- We built machine learning models to predict types of crimes given predictor features constructed from insights gained from EDA and compare performances of classifiers .
- We find it is quite hard to predict for the seven types of crimes. We are able to predict ‘Theft’ or ‘Not Theft’ with an accuracy of 68%, an increase by 24% compared to that of dummy classifier i.e., just predicting the majority class.
- If we group crimes as ‘Property Crime’ (‘Theft’, ‘AutoTheft’, ‘Burglary’) and ‘Violent Crime’ (‘Assault’, ‘Murder’, ‘Rape’, ‘Robbery’), the best prediction is given by decision tree classifier and logistic regressor without random sampling with an accuracy of 84%.

Future Work

Improve accuracy:

- Convert address into coordinates and involve that in prediction.
- Get more data on crimes and enrich dataset with more features.

Other interesting questions include:

- How is crime rate correlated with economic status and demographics?
- How is crime rate related with weather?

Other potential datasets:

weather conditions
economic status (unemployment rate)
demographics (population change)