Syracuse Crime Analysis

IST 718, Instructor: Prof. Acuna
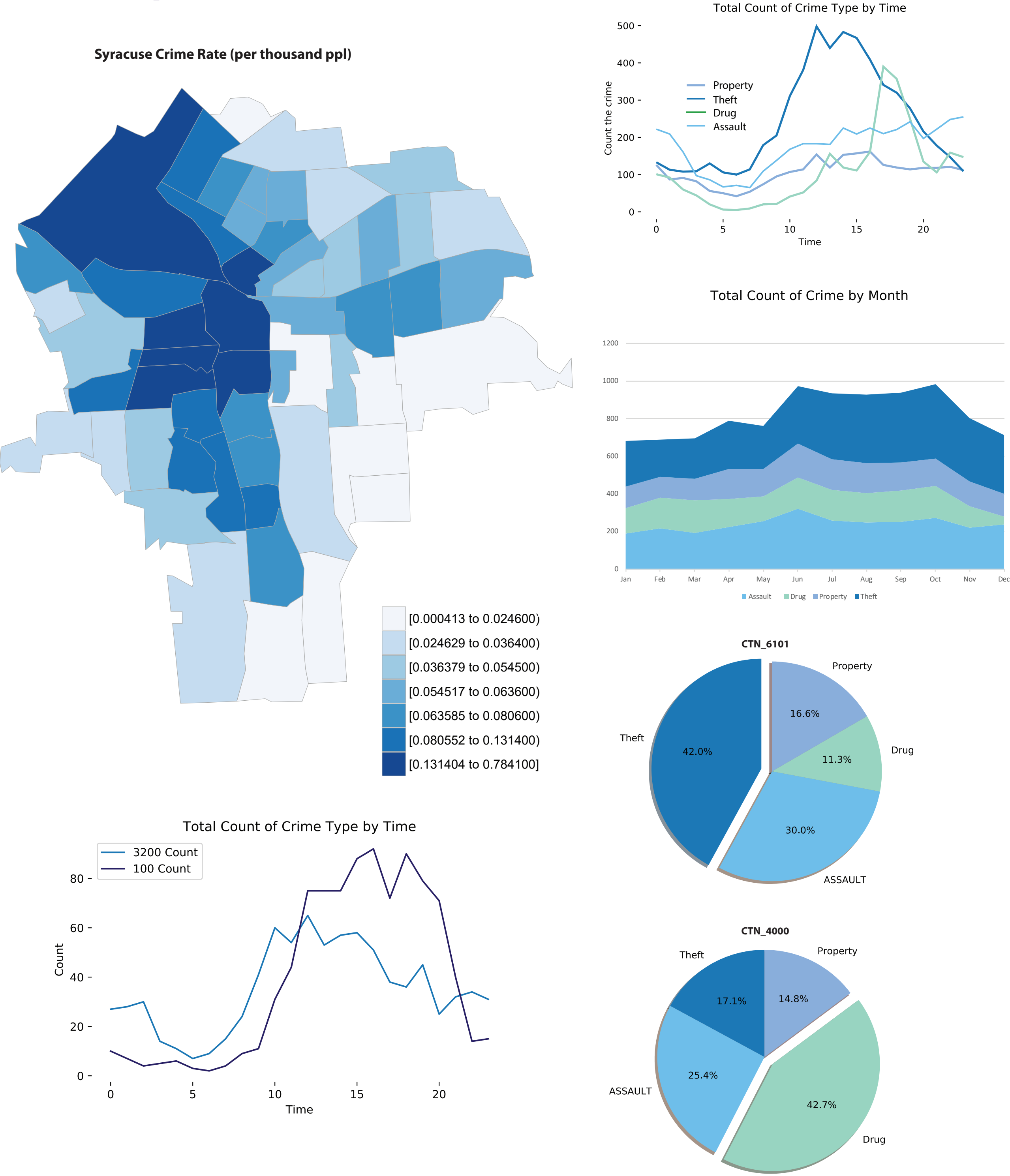Group 5:
Xiaoyi Yan, Yunteng Gu, Yung Hsin Hsu

Problems and Objectives:

Crimes take place in the society all over the world and have strong impact on the community. Syracuse, the place we are living in, has high rate of crimes but low chance of arrest. We want to analyze Syracuse crime data, especially for the types of crime, to find important crime factors, to raise cautious of the public, and to avoid crime more effectively.

Data Description:

Data Duration: January 1st, 2017 – June 3rd, 2018
Variables: Census Tract Code, Date, TimeStart, TimeEnd, Code_Defined
Observations: 15,266
Code_Defined: Assault, Drug, Property, Theft

Data Exploration:

Syracuse Crime Rate (per thousand ppl)

[0.000413 to 0.024600)
[0.024629 to 0.036400)
[0.036379 to 0.054500)
[0.054517 to 0.063600)
[0.063585 to 0.080600)
[0.080552 to 0.131400)
[0.131404 to 0.784100]

Total Count of Crime Type by Time

Total Count of Crime by Month

CTN_6101

CTN_4000

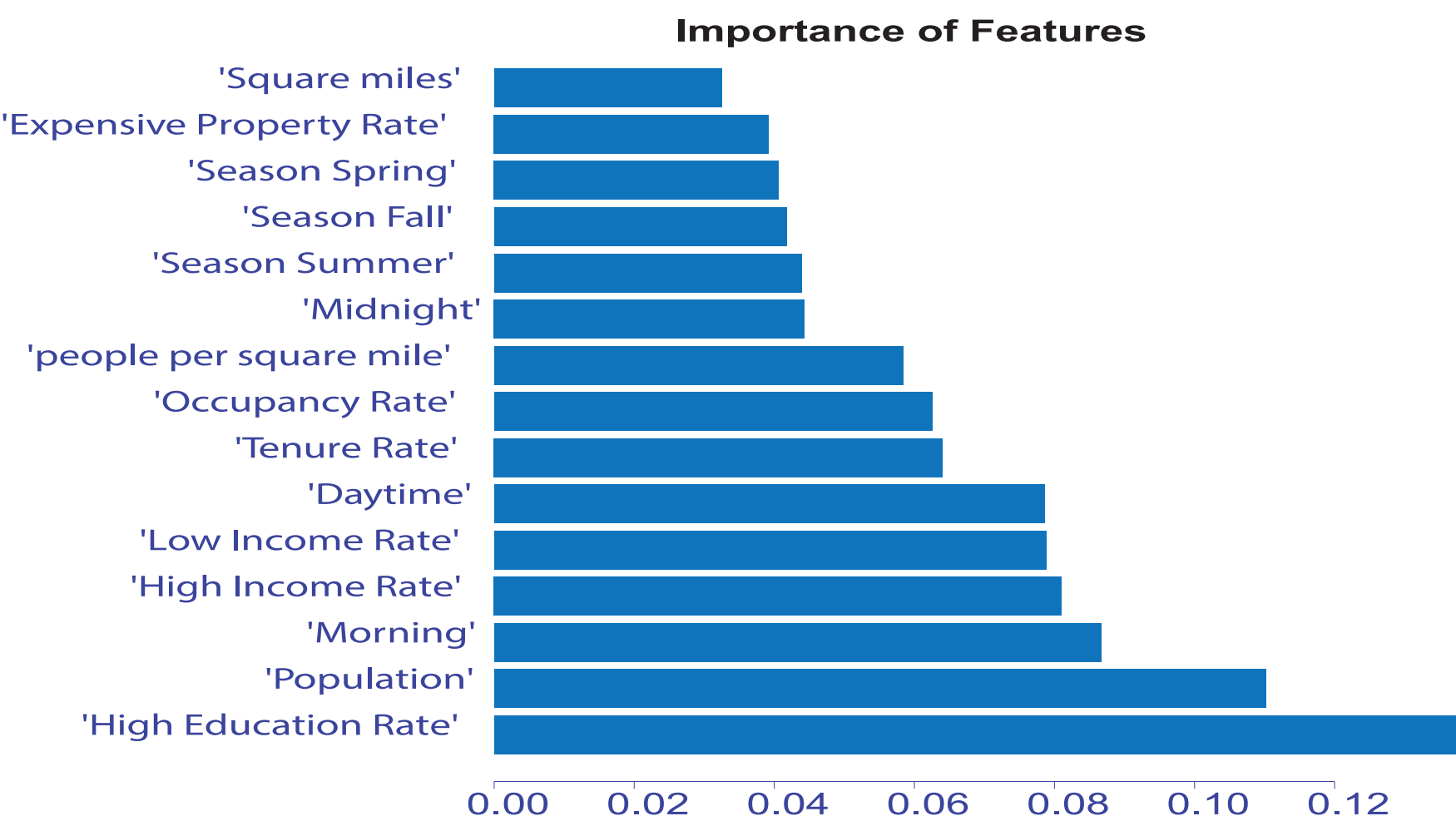**Total Count of Crime Type by Time**

Feature Preparation:

Date: Weekday, Dummy for Month, Season
TimeStart & TimeEnd: Time, Dummy for Day Part
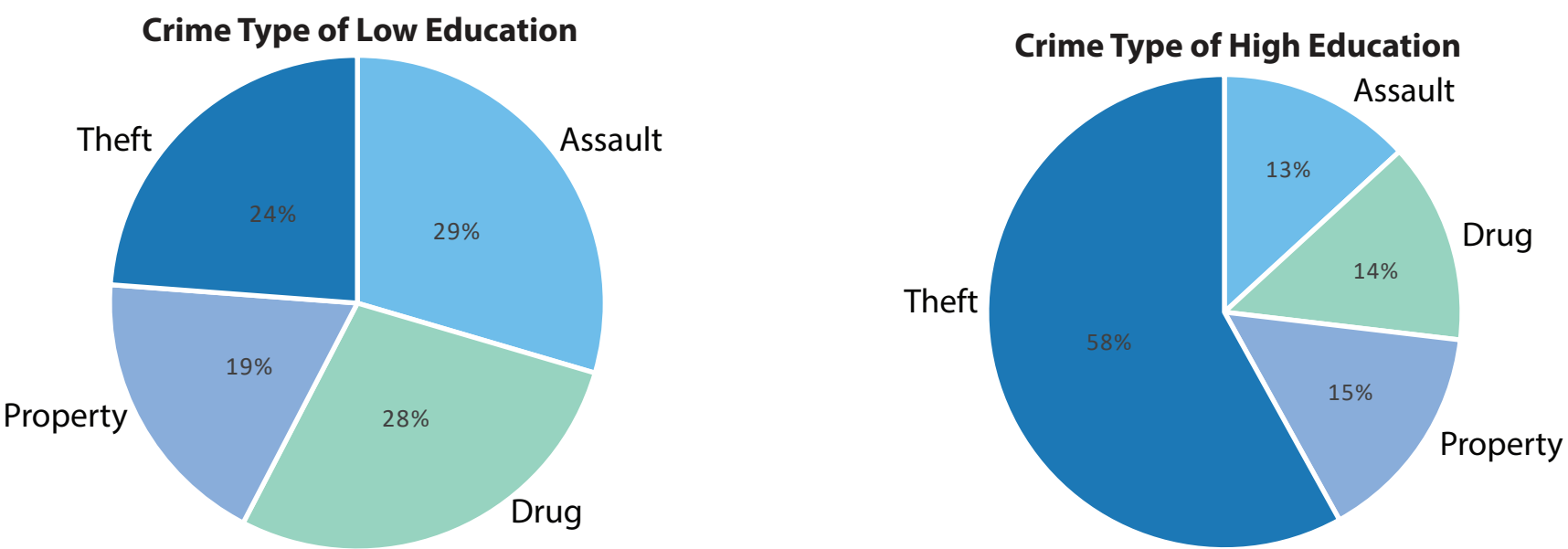Rate of Census Data: Tenure Rate, High Education Rate, etc...

Models Analysis:

	Model 1: Logistic regression classification	Model 2: Random Forest
Split Percentage	Training: 0.6 Validation: 0.3 Testing: 0.1	Training: 0.6 Validation: 0.3 Testing: 0.1
Parameters	Best RegParam: 0.0 Best ElasticNetParam: 0.0	numTrees=5 maxDepth=9
Variables	'Midnight', 'Morning', 'Daytime', 'Population', 'Square miles', 'people per square mile', 'Tenure Rate', 'Occupancy Rate', 'Low Income Rate', 'High Income Rate', 'Expensive Property Rate', 'High Education Rate', 'Season Spring', 'Season Summer', 'Season Fall'	'Midnight', 'Morning', 'Daytime', 'Population', 'Square miles', 'people per square mile', 'Tenure Rate', 'Occupancy Rate', 'Low Income Rate', 'High Income Rate', 'Expensive Property Rate', 'High Education Rate', 'Season Spring', 'Season Summer', 'Season Fall'
Test_AUC	0.3709	0.3891
Majority Baseline	0.3598	0.3598
Accuracy	0.4240	0.4263

Logistic regression Confusion Metrics

Random Forest Confusion Metrics

Results:

Importance of Features

High Education and Population have significant influence on the type of crime.

Crime Type of Low Education

Crime Type of High Education

Low Education Area:
More even distribution of each type of crime

High Education Area:
Theft is in the top freqency crime

Conclusion:

There are various types of crime, and are affected by different factors. We recommend the police to take into account of different factors for different kind of crimes to better assign their resources.
For areas with more educated residences, polices may need to strengthen night patrols.
For area with less educated residences, polices may need to sustain or increase police force all day long.

Improvement:

The accuracy of our result is not appealing, because our data capacity needs to be increased. We need to collect more data to have better insights to improve the accuracy of our models. At the same time, our model is mainly used to make recommendations to the police. If we need stronger predictability, we need to use more predictive models such as SVM.

References:
[1] Syracuse Open Data: http://data.syrgov.net/datasets/weekly-part-1-crime-offenses-2018
[2] United States Census Bureau: https://www.census.gov/