A closer look into NY parking violations

Cody Hegwer, Jeff LaPrade, Adam Grabowski

Description

For our semester project, we intend to analyze data regarding parking violations in New York City. Along the way we may discover other interesting questions we wish to answer, but our initial interest will be in questions such as: "If I wish to traverse the city of New York and never pay for parking, can I identify a route to take that minimizes my chance of receiving a ticket?" or "Is there a profile we can establish from the data to predict what cars will receive tickets?".
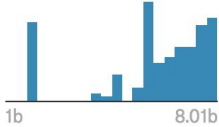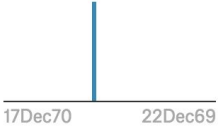
Prior Work

- After choosing our data set, we identified several examples of other students performing similar data mining projects concerning both NY parking violations, and parking violations in other cities. It actually was a more popular research topic than we anticipated, but the other student projects were asking different questions that what will be exploring.
- An example of a professional inquiry done by a college professor: http://iquantny.tumblr.com/post/83696310037/meet-the-fire-hydrant-that-unfairly-nets-nyc
- Another example of a professional inquiry done by a data analyst: https://www.linkedin.com/pulse/toronto-parking-ticket-data-analysis-angela-ju/

Datasets

Currently held on Adam's desktop computer.

- 43 attribute types
- 9.10 million objects
- Contains data ranging from years 1970 all the way to 2014.

	# Summons Number	A Plate ID	A Registration State		A Plate Type		📅 Issue Date	# Vio
	1b — 8.01b	2722215 unique values	NY 77% NJ 10% Other (67) 13%		PAS 72% COM 21% Other (86) 8%		17Dec70 — 22Dec69	0
1	1361929741	FCJ5493	NY		PAS		1970-12-18T00:00:00	
2	1366962000	63540MC	NY		COM		1971-02-02T00:00:00	
3	1342296187	GCY4187	NY		SRF		1971-09-18T00:00:00	
4	1342296199	95V6675	TX		PAS		1971-09-18T00:00:00	
5	1342296217	FYM5117	NY		SRF		1971-09-18T00:00:00	
6	1356906515	GFM1421	NY		PAS		1971-09-18T00:00:00	
7	1337077380	18972BB	NY		999		1971-10-10T00:00:00	
8	1364523796	WNJ4730	VA		PAS		1973-04-05T00:00:00	

Proposed Work

Items to be Cleaned:

- The data appears fairly clean so far. The only cleaning we have currently identified that will be necessary is dealing with missing values. Since there is only a very small number of missing values, we currently plan to simply exclude those rows as it will hardly reduce the number of available entries.

Data Preprocessing:

- Most of the data is already complete and in an ideal format. However, some attributes are irregular formats and will need to be transformed. The only attributes we have identified as of now that require transformation are the various Date and/or Time attributes.

Tools

- Tableau
- Python
- Pandas
- Matplotlib

Evaluation

In our evaluation of the data, we will be primarily concerned with the metrics: accuracy, error rate, and precision.

Due to the nature of our questions, the main events we wish to predict are future events that, for the sake of this study, we would not have access to. Therefore, the precision of our classification will be more readily evaluated since we may predominantly be testing on data already in our dataset as opposed to new, more recent data.