# Modeling Social Data
# Final Project

Zach Gleicher, Matt Piccolella, and Edo Roth

# Stop and Frisk

Police Activity in New York City

# The Problem

- “In 2012, New Yorkers were stopped by the police 532,911 times.” (New York Civil Liberties Union)
- 473,544 (89%) were totally innocent
- Can this performance be improved?

# Our Objectives

- Finding patterns in the data -- who is being stopped? When? Where?

- Can we predict when arrests are being made?

# The Data

- Public Records from the NYPD
- ~500,000 data points with 118 variables
- Mostly Boolean (Did suspect have a weapon? Was suspect frisked? Was physical force used by officer?)
- Other Factor Indicators (Age, Race, Sex, Location of Stop, etc.)

# Filtering and Cleaning Up

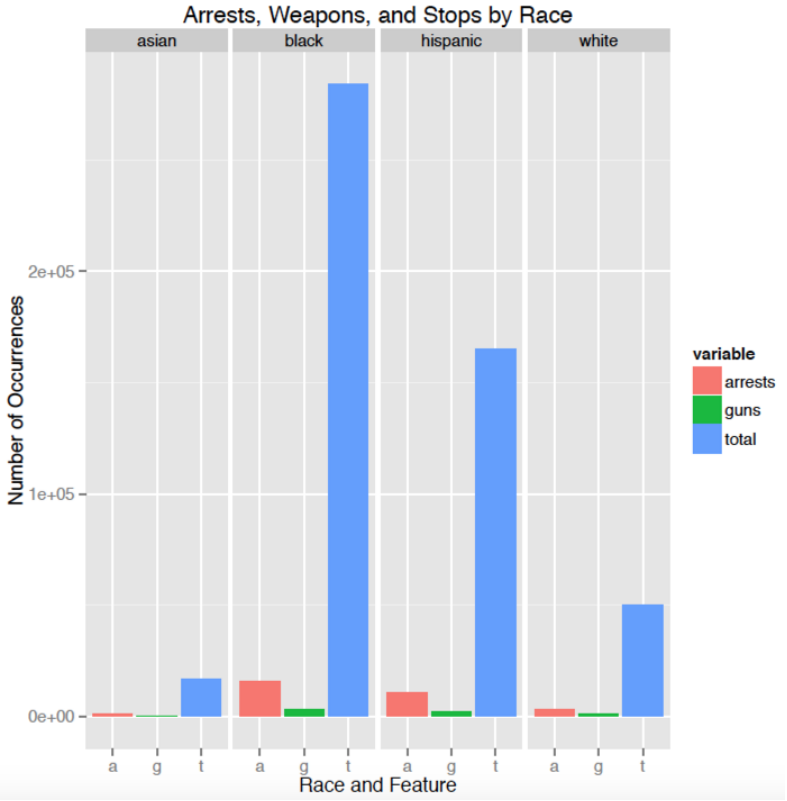
- Command line utils to:
  - Separate the columns we use
  - Remove malformed data
  - Parse integers

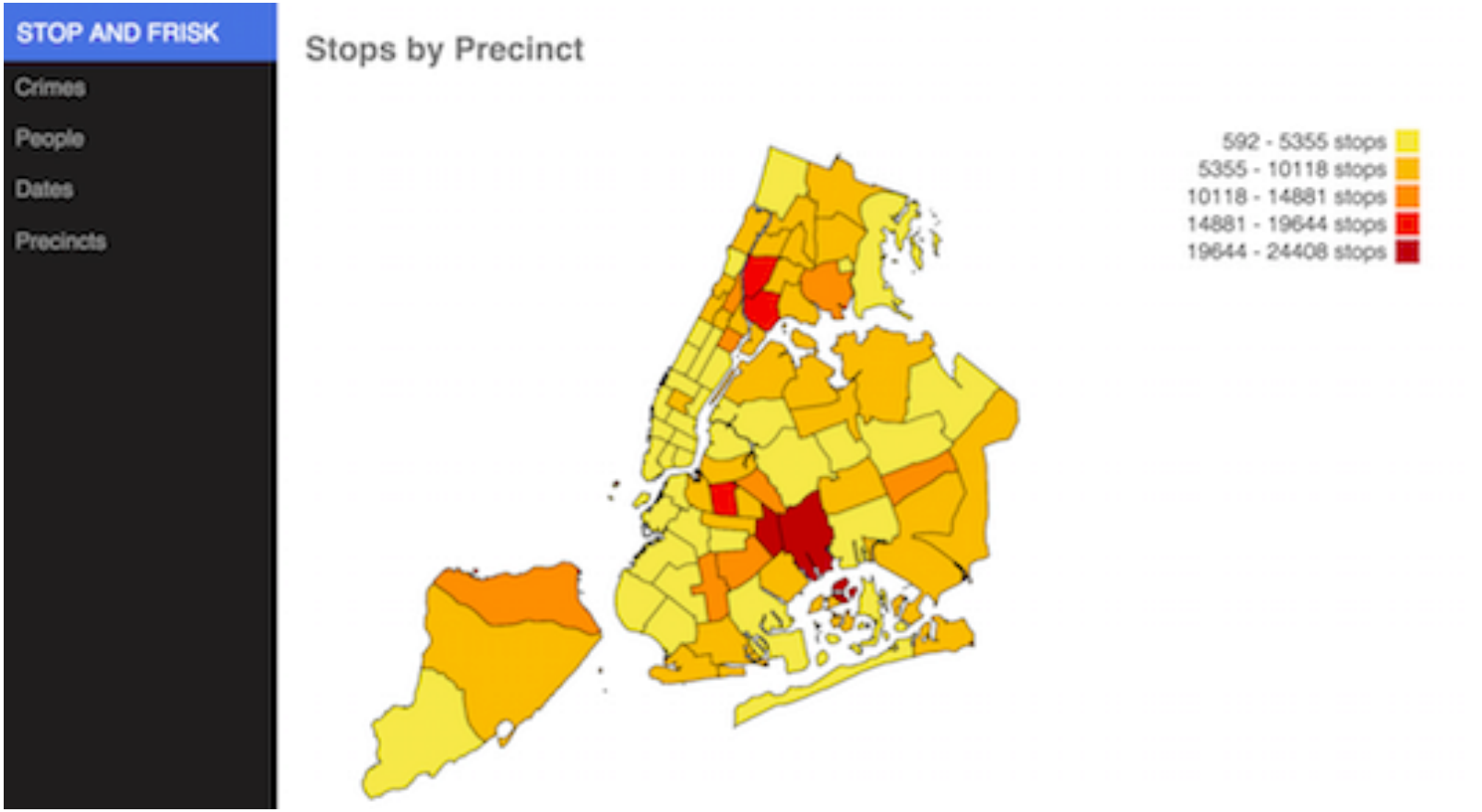
- R code to remove outliers for gender, race, etc.

# Looking at the Data

- Used dplyr and ggplot (Split/Apply/Reduce) to make visualize the data

- Created interactive visualizations with D3.js





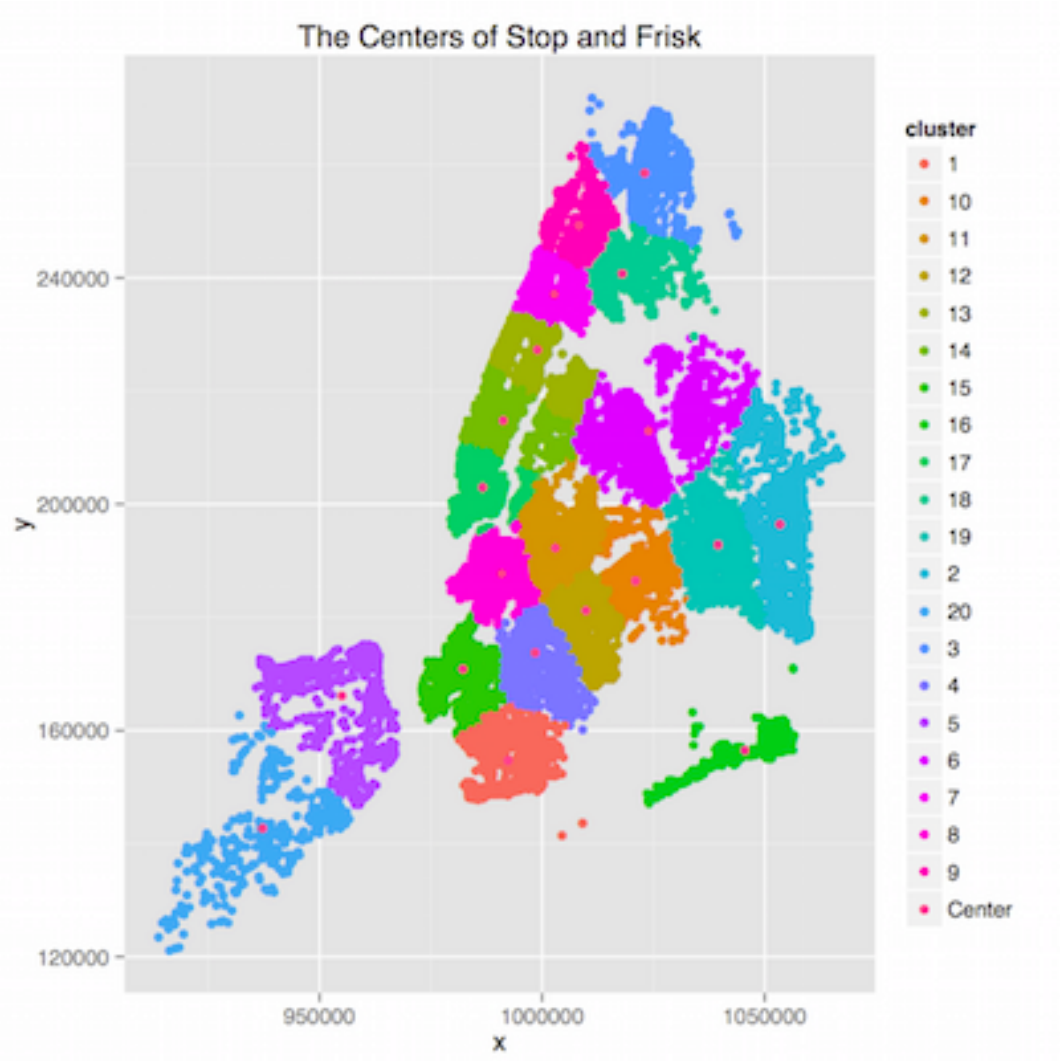

[Stop and Frisk Data]

# Analysis/Prediction

- K-Means Clustering
- Naive Bayes
- Logistic Regression

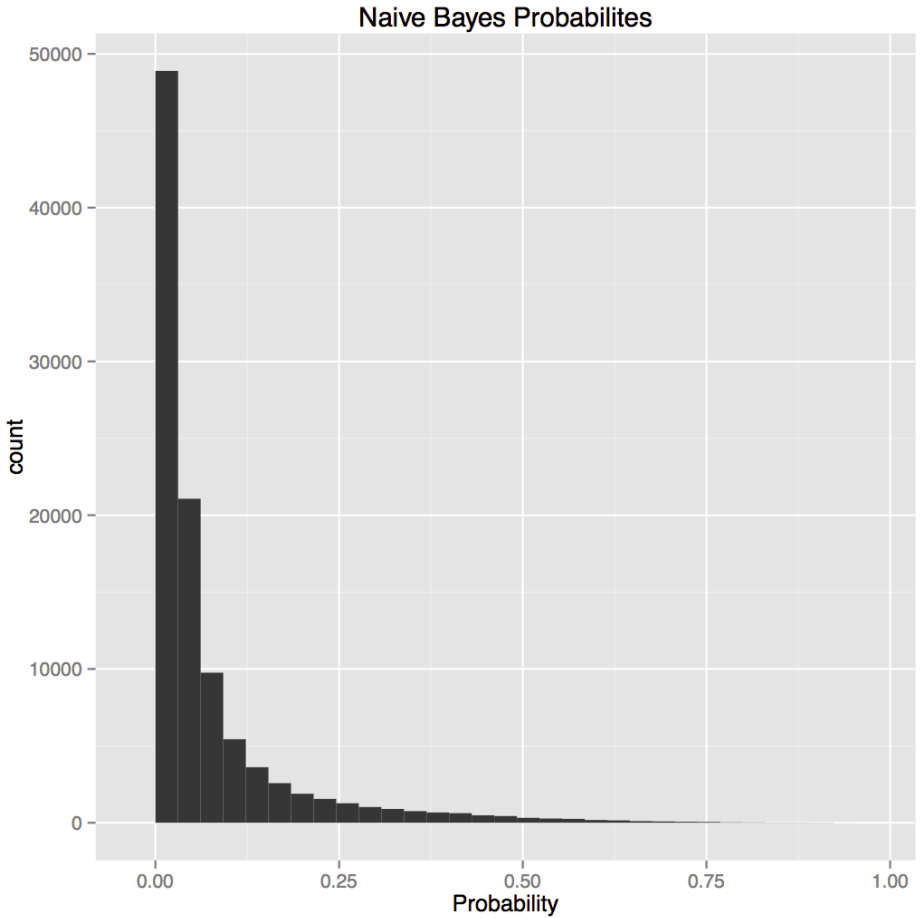
# K-Means Clustering

- Thought clustering stops geographically could help us find centers of Stop and Frisk in the city
- Don’t just “Dive In”
- R-Packages are superior


The Centers of Stop and Frisk
240000
200000
160000
120000
y
950000
1000000
1050000
x
cluster
1
10
11
12
13
14
15
16
17
18
19
2
20
3
4
5
6
7
8
9
Center

# Naive Bayes

- High probability for no arrest
- ROC Curve to measure performance
  - AUC = 0.7199524
- Balancing Data

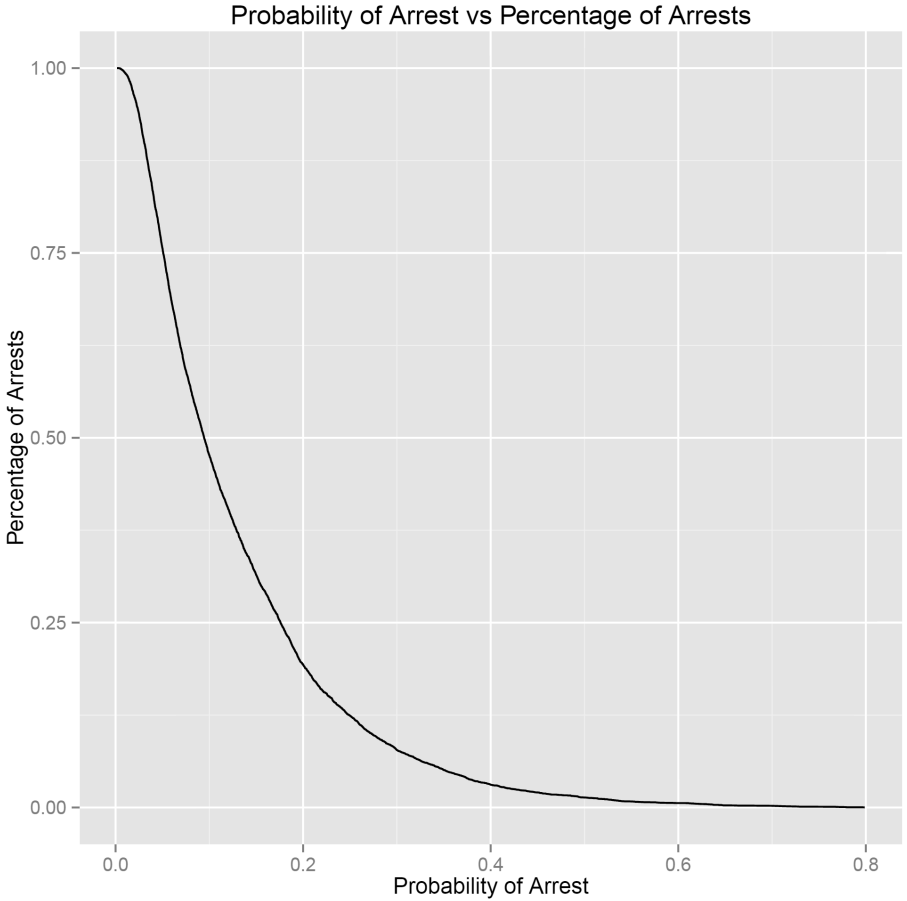


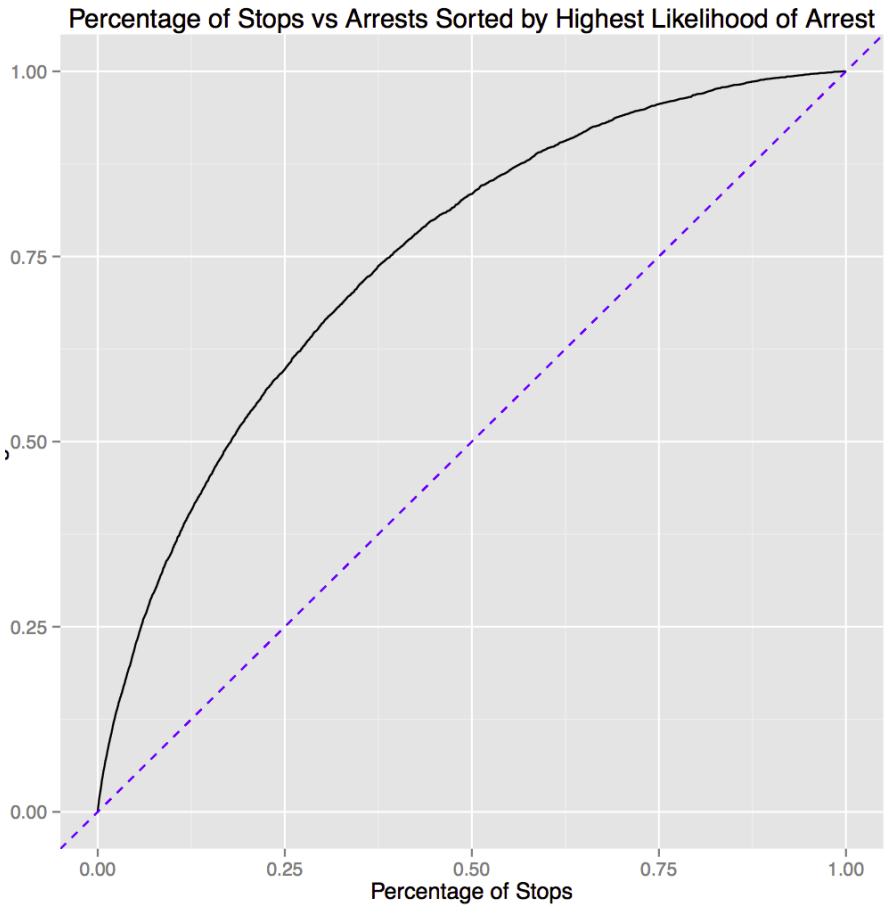
# Logistic Regression

- AUC = 0.7392117 (always slightly better than Naive Bayes)

```
"Get 10 best predictors for Arrest"
    pct42     pct25     pct19     pct44    pct105    pct102 cs_drgtr1      pct9  ac_rept1 cs_objcs1
0.5008073 0.5639895 0.6069064 0.6222683 0.6400148 0.7004087 0.8372146 0.8652883 0.9516741 1.0960916
```

- Attempted to use regularization with Lasso, didn’t improve results


Probability of Arrest vs Percentage of Arrests
Percentage of Arrests
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
Probability of Arrest


Percentage of Stops vs Arrests Sorted by Highest Likelihood of Arrest
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Percentage of Stops

# What's Next?

- Feature Conjunction for classification

- Providing each precinct with most indicative features for arrest