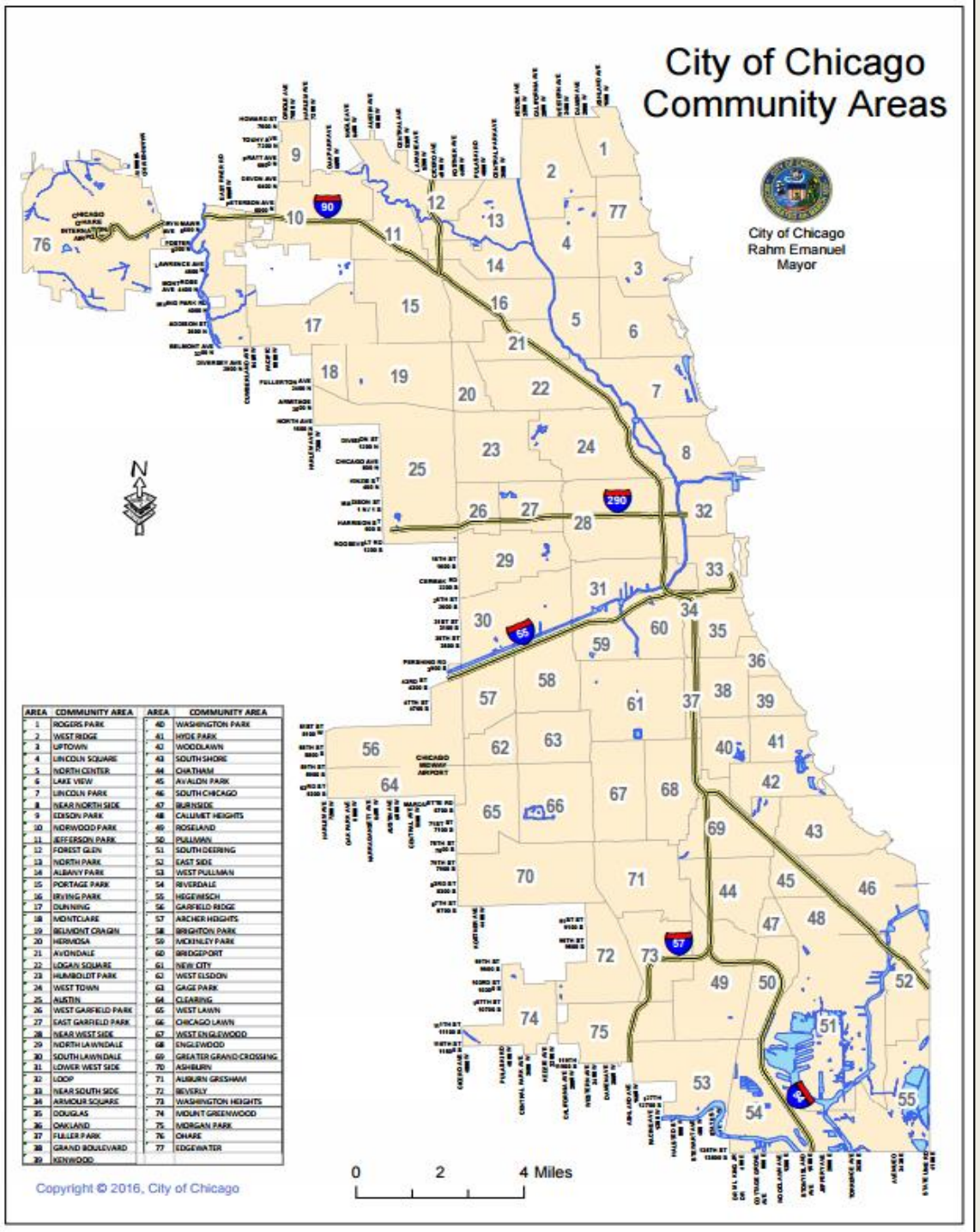
POTHOLE ANALYSIS IN THE CITY OF CHICAGO

Melanie Hanna

July 2017



AREA	COMMUNITY AREA	AREA	COMMUNITY AREA
1	ROGERS PARK	40	WASHINGTON PARK
2	WEST RIDGE	41	HYDE PARK
3	UPTOWN	42	WOODLAWN
4	LINCOLN SQUARE	43	SOUTH SHORE
5	NORTH CENTER	44	CHATHAM
6	LAKE VIEW	45	AVALON PARK
7	LINCOLN PARK	46	SOUTH CHICAGO
8	NEAR NORTH SIDE	47	BURNSIDE
9	EDISON PARK	48	CALUMET HEIGHTS
10	NORWOOD PARK	49	ROSELAND
11	JEFFERSON PARK	50	PULLMAN
12	FOREST GLEN	51	SOUTH DEERING
13	NORTH PARK	52	EAST SIDE
14	ALBANY PARK	53	WEST PULLMAN
15	PORTAGE PARK	54	RIVERDALE
16	IRVING PARK	55	HEGEWISCH
17	DUNNING	56	GARFIELD RIDGE
18	MONTCLARE	57	ARCHER HEIGHTS
19	BELMONT CRAGIN	58	BRIGHTON PARK
20	HERMOSA	59	MCKINLEY PARK
21	AVONDALE	60	BRIDGEPORT
22	LOGAN SQUARE	61	NEW CITY
23	HUMBOLDT PARK	62	WEST ELSDON
24	WEST TOWN	63	GAGE PARK
25	AUSTIN	64	CLEARING
26	WEST GARFIELD PARK	65	WEST LAWN
27	EAST GARFIELD PARK	66	CHICAGO LAWN
28	NEAR WEST SIDE	67	WEST ENGLEWOOD
29	NORTH LAWNDALE	68	ENGLEWOOD
30	SOUTH LAWNDALE	69	GREATER GRAND CROSSING
31	LOWER WEST SIDE	70	ASHBURN
32	LOOP	71	AUBURN GRESHAM
33	NEAR SOUTH SIDE	72	BEVERLY
34	ARMOUR SQUARE	73	WASHINGTON HEIGHTS
35	DOUGLAS	74	MOUNT GREENWOOD
36	OAKLAND	75	MORGAN PARK
37	FULLER PARK	76	OHARE
38	GRAND BOULEVARD	77	EDGEWATER
39	KENWOOD		

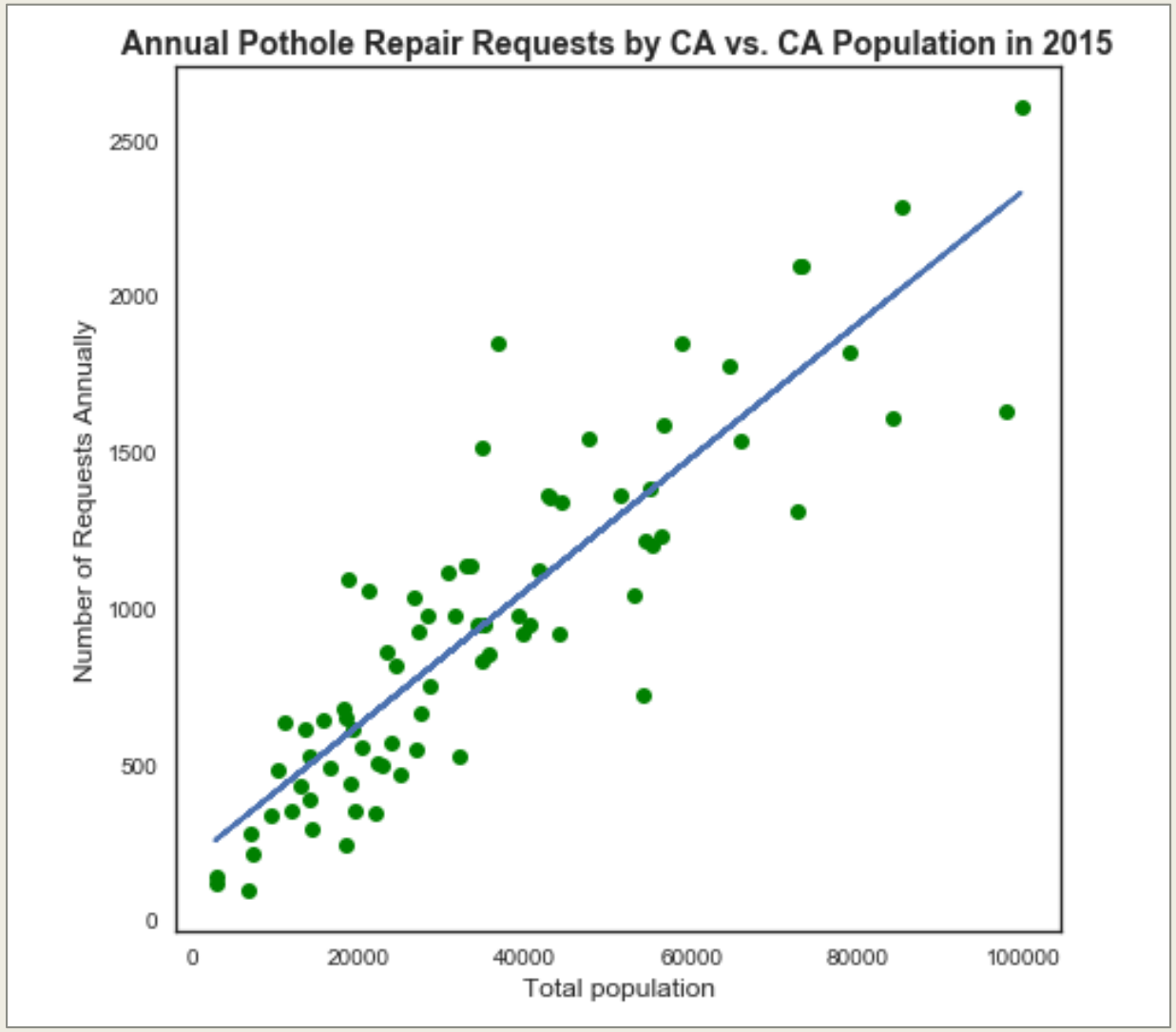
Introduction

- We set out to:
  - *Analyze pothole data across all 77 community areas in the city of Chicago*
  - *Compare to census data to discover patterns in demographic/economic variables*
  - *Create predictive models to forecast pothole metrics for a given geographic area*

Data sets

- City of Chicago's 311 Service Request database from January 2011 to June 2017
  - *More than 477,000 pothole reports*
  - *Includes creation and completion date, number of potholes found on block when crew arrives, physical address of request*
- Chicago Metropolitan Agency for Planning (CMAP)'s Community Snapshot data:
  - *77 community areas*
  - *Sourced from U.S. Census Bureau's 2010-14 American Community Survey, Longitudinal Employment-Household Dynamics data for 2014, and 2014/2015 data from the Illinois Department of Employment Security and the Illinois Department of Revenue*
  - *155 metrics in total (median income, percent vacant housing, etc.)*

DATA EXPLORATION

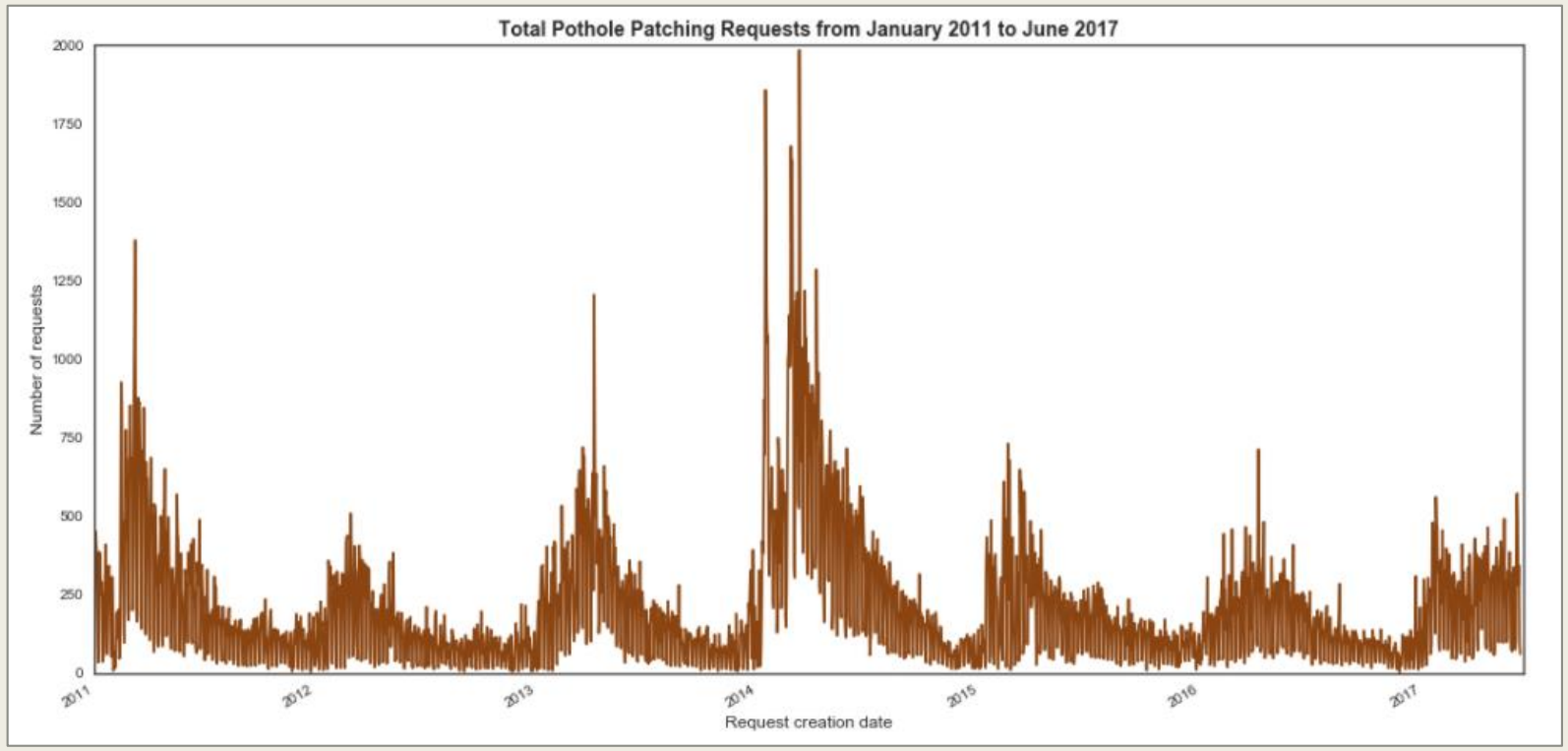


Service requests by community area

- Linear relationship between total population and the annual number of pothole repair requests
- Slope of regression line indicates 2.1 requests are submitted per 100 residents, regardless of community area

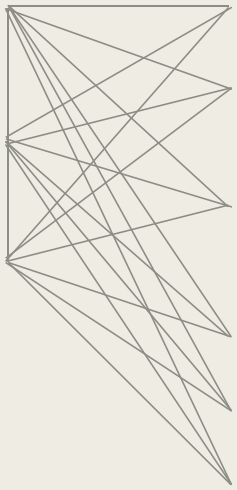
Service requests over time

- Spike in requests in early 2014 aligns with winter vortex
- Regular increase in springtime months
  - *43.5% of all requests occur in spring*



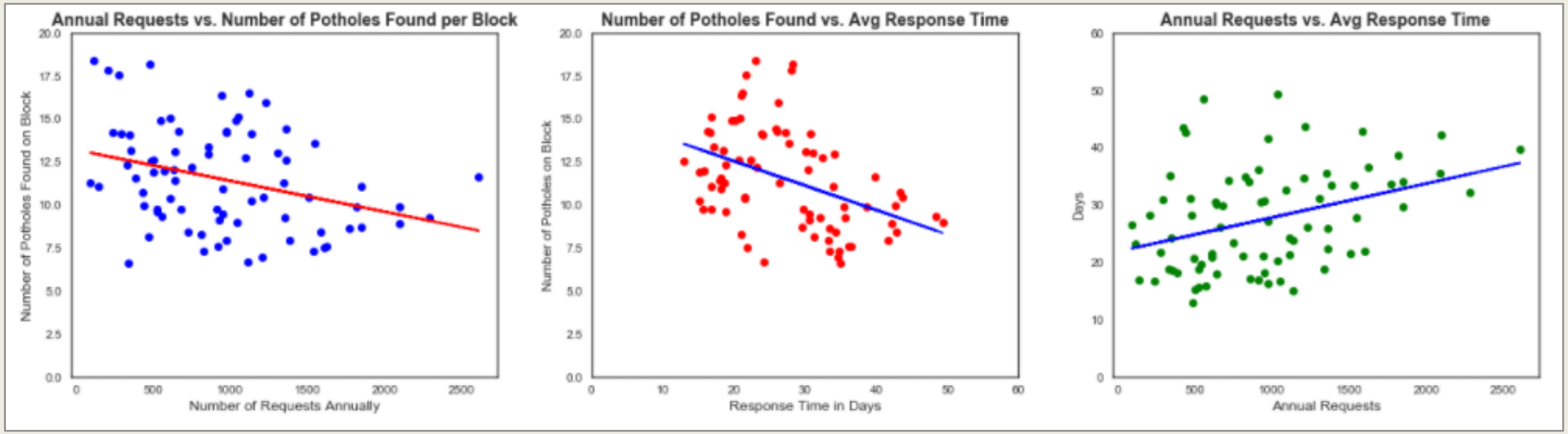
Pothole metrics vs. census data

- Analyzed relationships between the following:
  - *Pothole metrics*
    - Number of potholes found by the DoT per block
    - Number of annual pothole repair requests
    - Response time to repair request
  - *Census data*
    - Median income
    - Percent of land defined as roadways
    - Commuting transportation method (public transit, carpool, etc.)
    - Percent of land as commercial
    - Percent of land as residential
    - Percent of land as industrial



Pothole metrics vs. census data

- Correlations found between all three pothole metrics:
  - *Number of potholes found is negatively correlated with both response time and annual requests*
    - Fewer potholes are discovered with more frequent and prompt service
  - *Response time is positively correlated with annual requests*
    - DoT does not respond as quickly to CA's filing more repair requests OR CA's file more repair requests when the DoT responds slowly

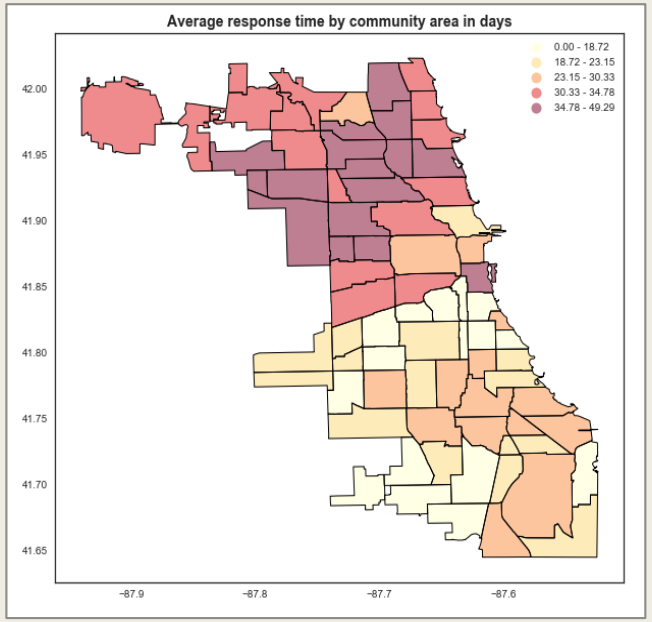


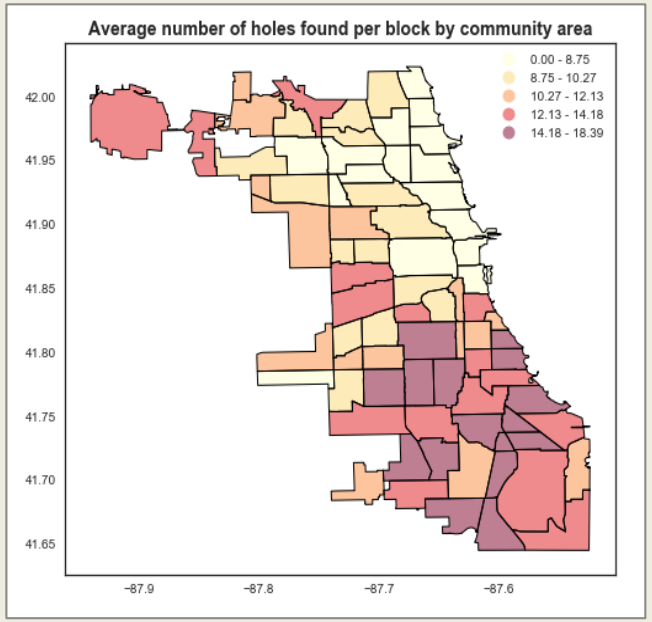
Other relationships found

- As median income ⇧, annual repair requests ⇧
- As median income ⇧, number of potholes found ⇩
- As percent of land zoned as residential ⇧, annual requests ⇧
  - *Tied to correlation with total population*
- As percent of land zoned as commercial ⇧, annual requests ⇧
- As percent of land zoned as commercial ⇧, response time ⇧
- As percent of land zoned as commercial ⇧, number of potholes found ⇩

North vs. South

- Southern community areas typically see more potholes per block than northern areas
  - *Southern average: 13.1 holes*
  - *Northern average: 9.0 holes*
- Northern areas wait longer on average for pothole repair than southern community areas
  - *Southern average: 21.1 days*
  - *Northern average: 36.4 days*

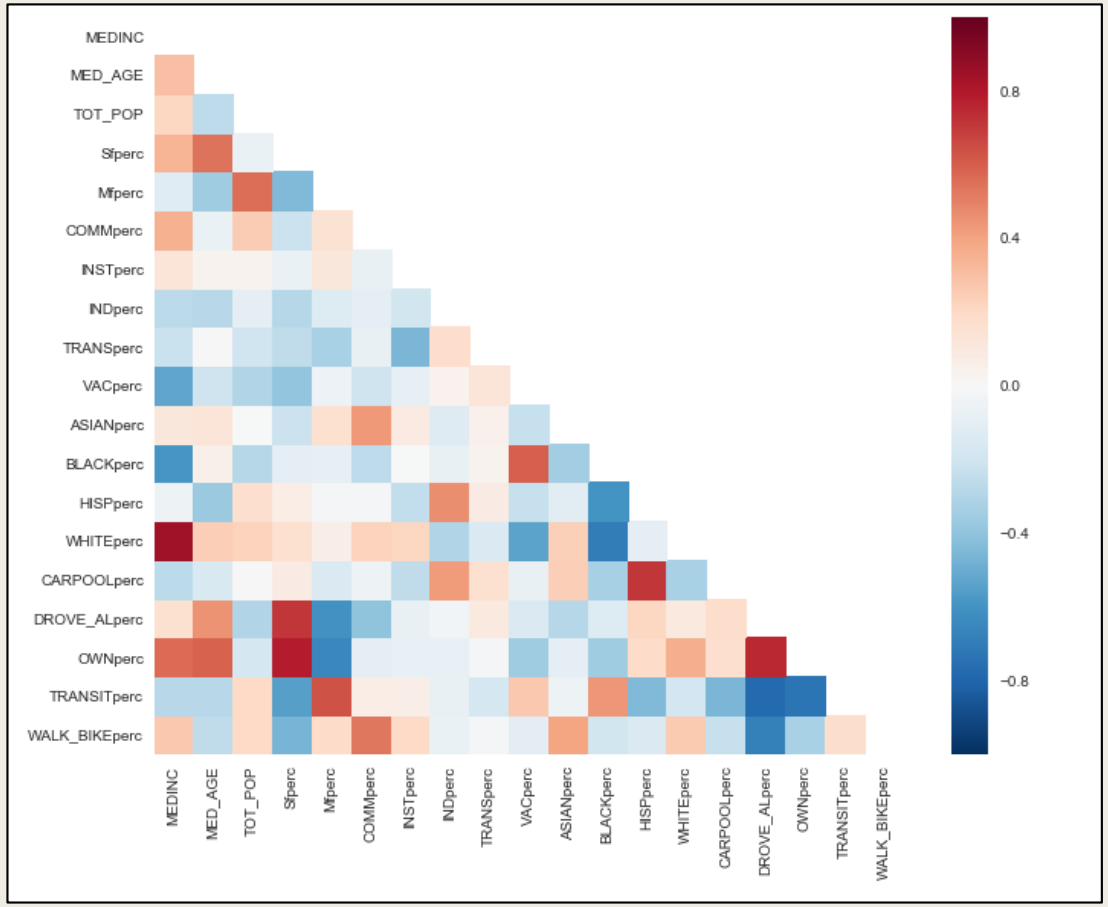


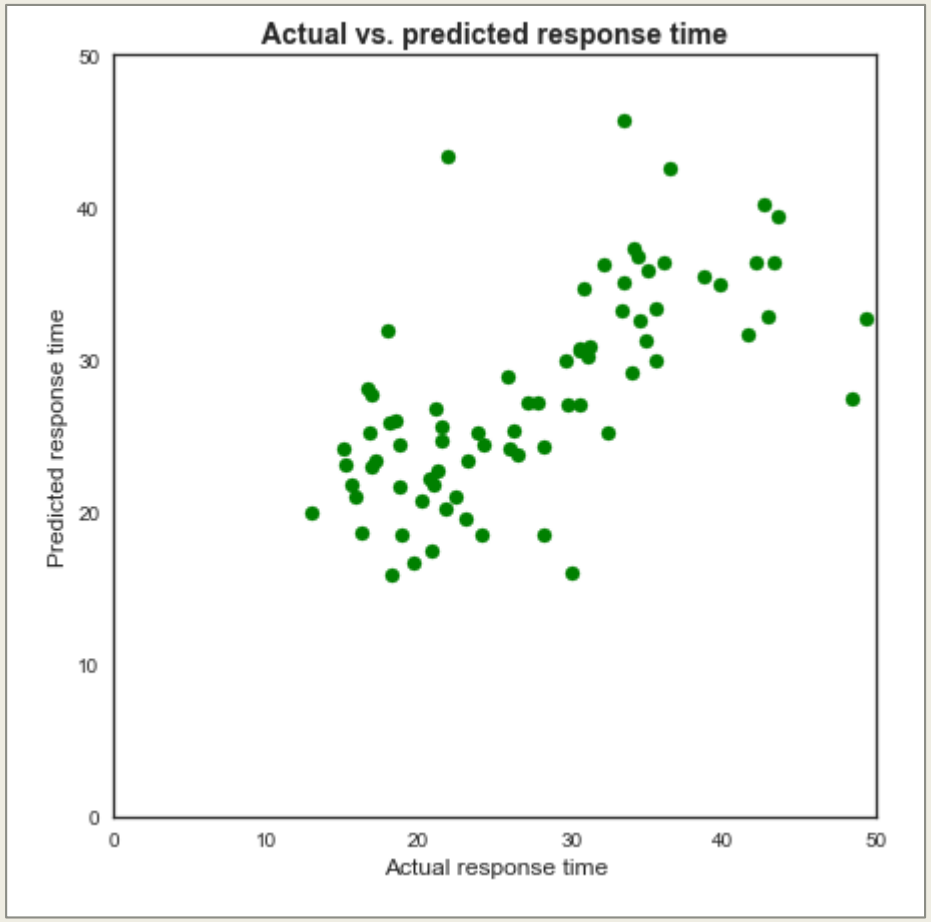


PREDICTIVE MODELS

Linear regression

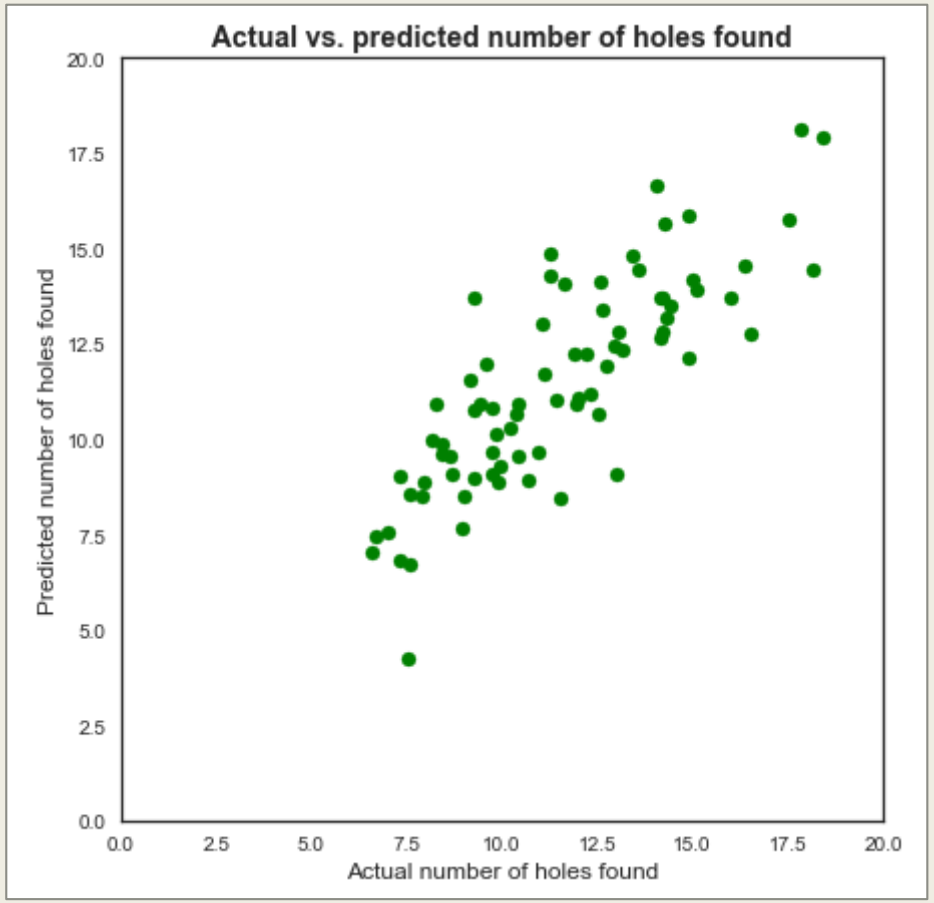
- Selected 19 features from the census data covering:
  - *Median age*
  - *Median income*
  - *Race*
  - *Mode of transport for commuting*
  - *Percent of land used as industrial, commercial, residential, & transportation*
  - *Percent of owner-occupied housing*
- Correlation matrix shows some relationships but none strong enough to skew the model





Linear regression on response time

- Model explains 37% of the variance in the data – a moderate value for fit
- Large increase in error from training set to test data, indicating an overfit model that does not generalize well
- Linear regression is not a good choice for this data set

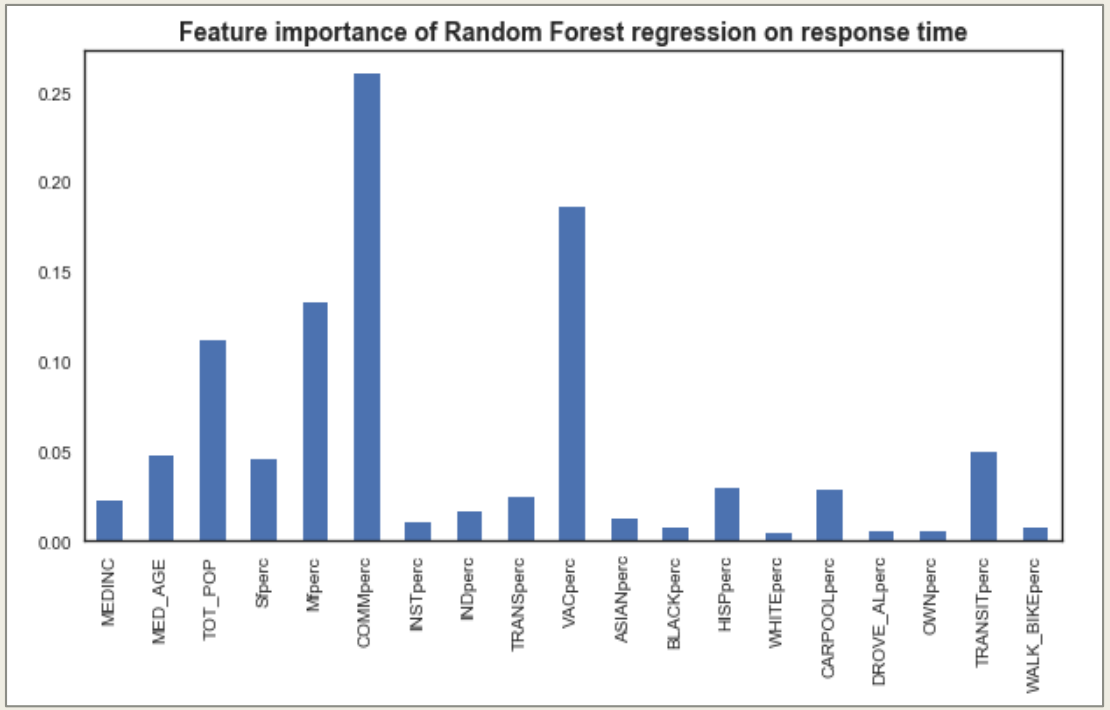


Linear regression on number of holes

- Model explains 65% of the variance in the data – a much better fit than the response model
- Small error on both training and test data
- Linear regression models the number of holes found per block very well

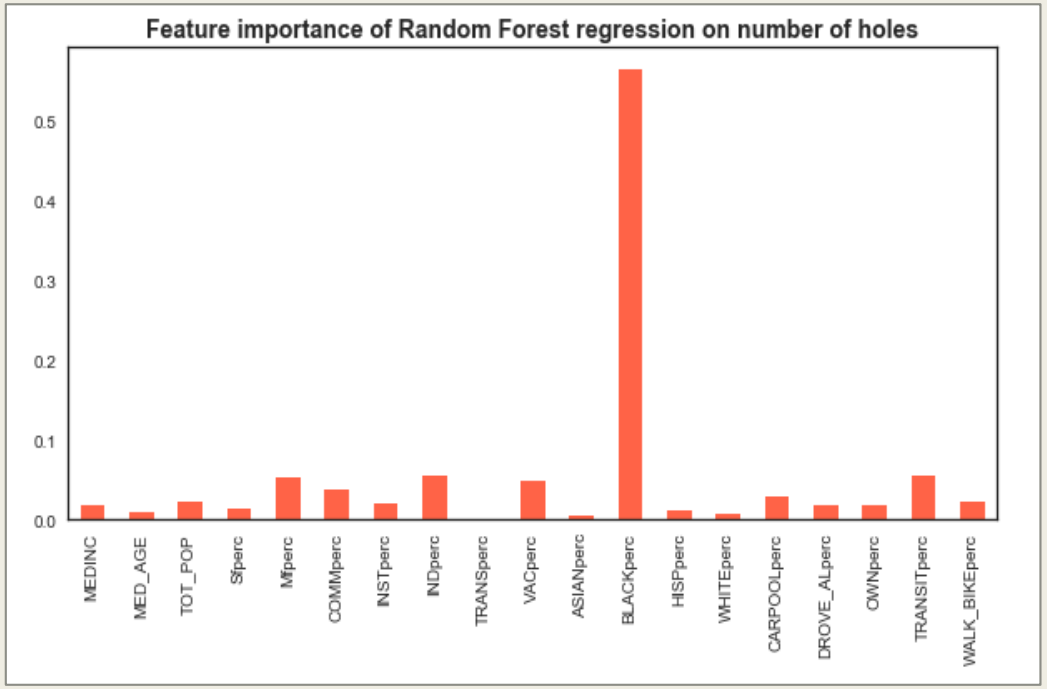
Random tree regression on response time

- Model explains 89% of data's variance but does not generalize well to the test set
- Percentage of land zoned as commercial and percent of vacant land are most important features during training

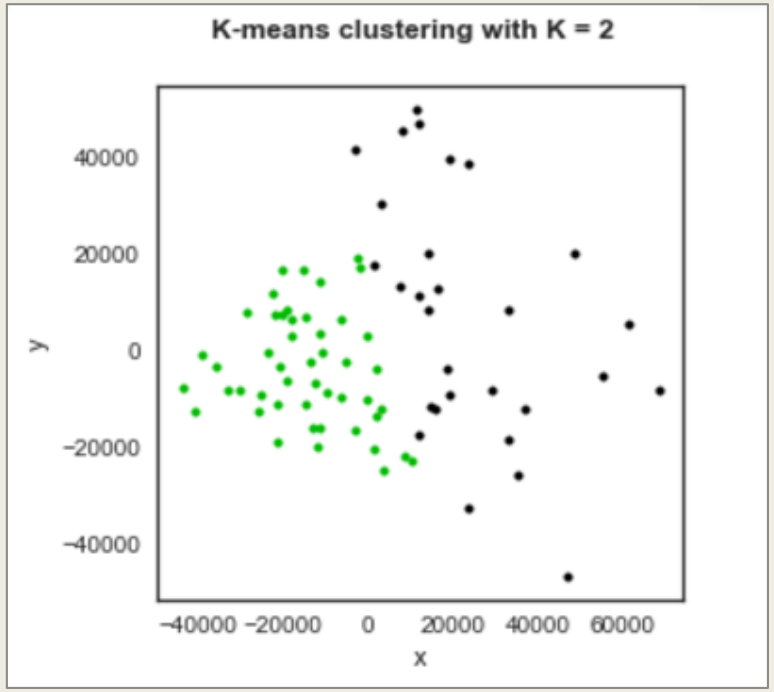


Random tree regression on number of holes

- Model explains 90% of the variance in the data
- Error on the training and test set is remarkably low
- Percentage of black residents is by far the most important feature



Clustering



- Combined census features from regression models and pothole data into one dataset and reduced to 2 dimensions
- Trained clustering model around 2 clusters to explore north-south relationships in community areas
  - *Cluster 1: 12 northern CA's and 2 southern ones*
  - *Cluster 2: 5 northern CA's and 27 southern ones*
- Clustering model then applied to test data
  - *Cluster 1: 3 northern CA's and 0 southern ones*
  - *Cluster 2: 0 northern CA's and 15 southern ones*
- Very accurate generalization

Recommendations and future work

- Use censusreporter.org to gather census data on smaller geographies
- For the City:
  - *Conduct a study to determine the root cause of the north/south divide in 311 service request response time*
- For Chicago residents:
  - *Lobby your alderman if your community area has a longer than average response time for pothole repair requests or if a large number of potholes are expected per block in your neighborhood.*
  - *Plot your daily commute to determine if those streets are pothole-ridden and route a new course on smoother roads.*