Market the Market

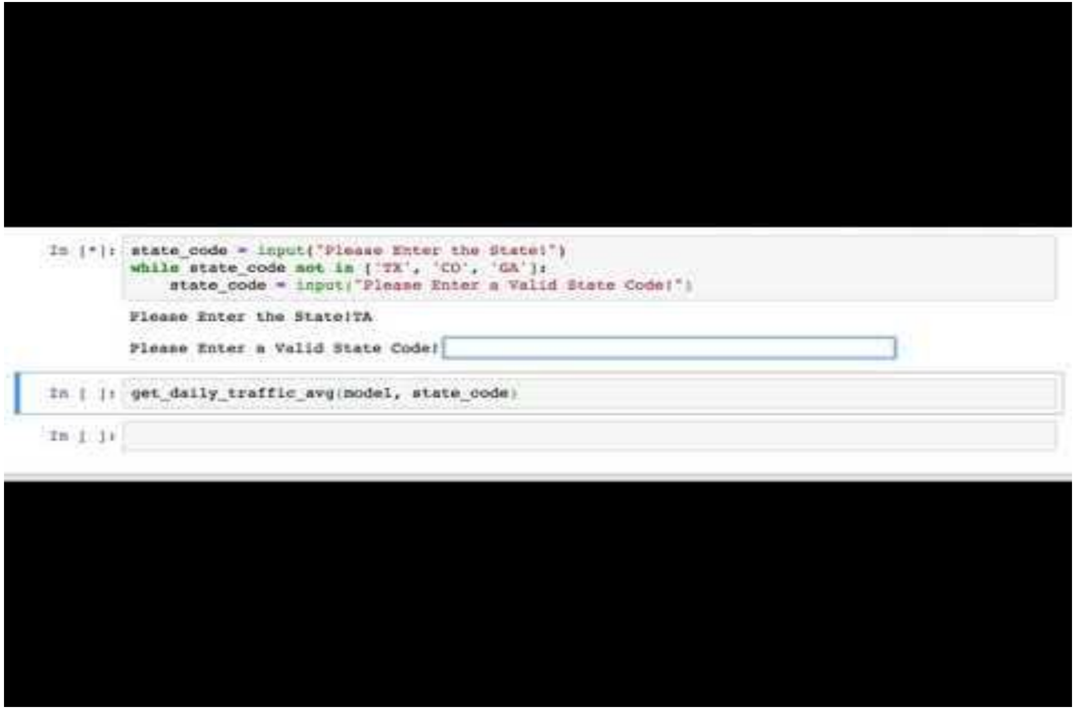
The Nightwatch

Yiqing Liang, Lingrui Luo, Yinhe Lu, Xinyuan He

+

MARKET

Demo



```
In [*]: state_code = input("Please Enter the State!")
        while state_code not in ['TX', 'CO', 'GA']:
            state_code = input("Please Enter a Valid State Code!")

        Please Enter the State!TA

        Please Enter a Valid State Code!

In [ ]: get_daily_traffic_avg(model, state_code)

In [ ]:
```

Overview

To find the most ideal place to start a new supermarket, we first use data visualization tools including carto, to visualize the patterns of the dataset, and use different statistical model to predict the traffic, as an indicator of profitability.

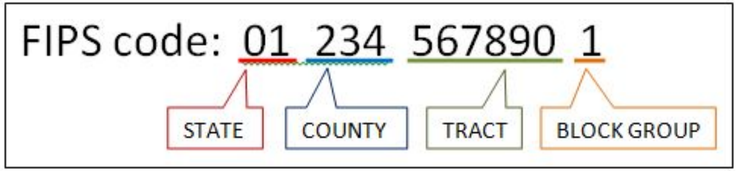
Question & Hypothesis

Can different factors such as store location, community population, competitor count, customer income… predict customer traffic?

Outsourced Datasets

- Location dataset
  - fipsCode → longitude and latitude
  - https://www.quora.com/Where-are-latitude-longitude-coordinates-for-all-census-block-FIPS-codes-available
- Income dataset
  - fipsCode → zipcode
  - Zip code → income
  - https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

Outsourced Datasets



FIPS code: 01 234 567890 1

STATE

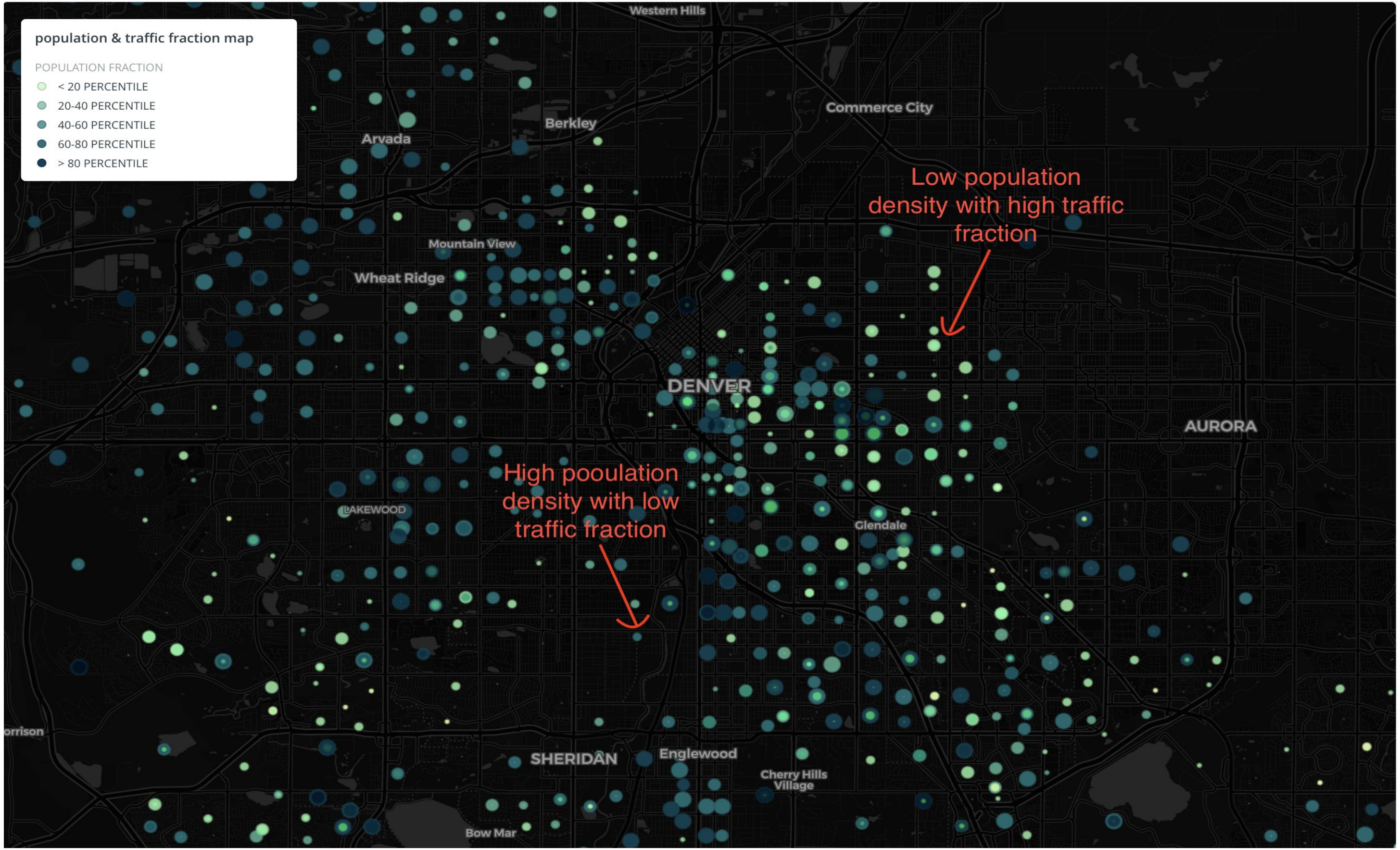
COUNTY

TRACT

BLOCK GROUP

- Location dataset
  - fipsCode → longitude and latitude
  - https://www.quora.com/Where-are-latitude-longitude-coordinates-for-all-census-block-FIPS-codes-available
- Income dataset
  - fipsCode → zipcode
  - Zip code → income
  - https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

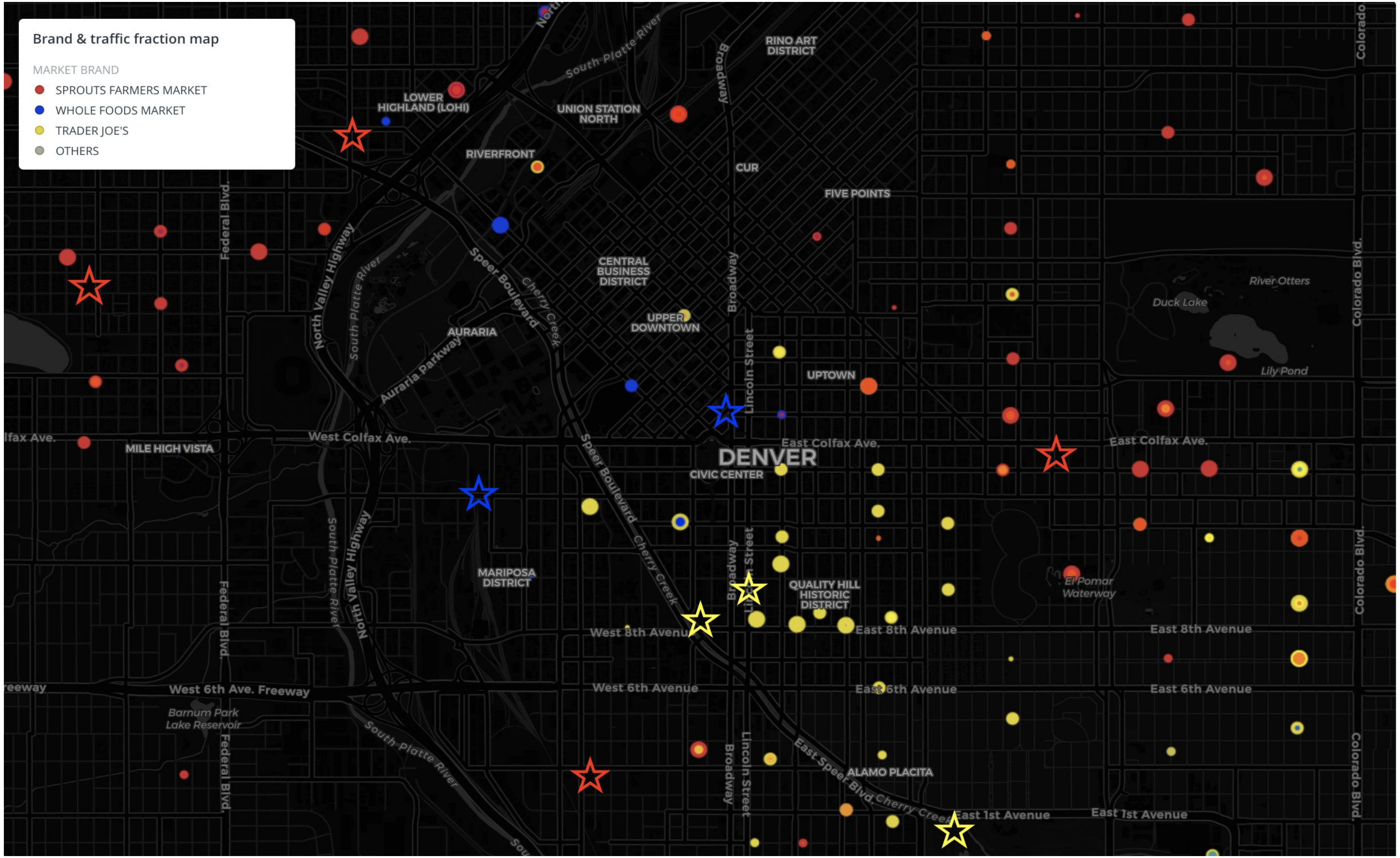
Population density & Traffic fraction



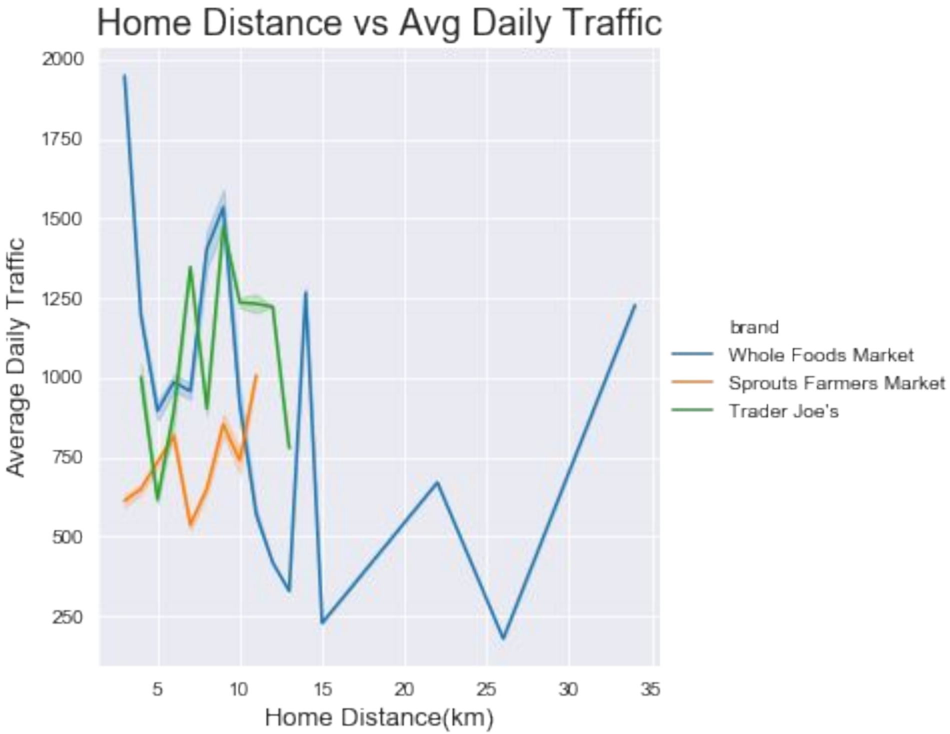
Low population density with high traffic fraction

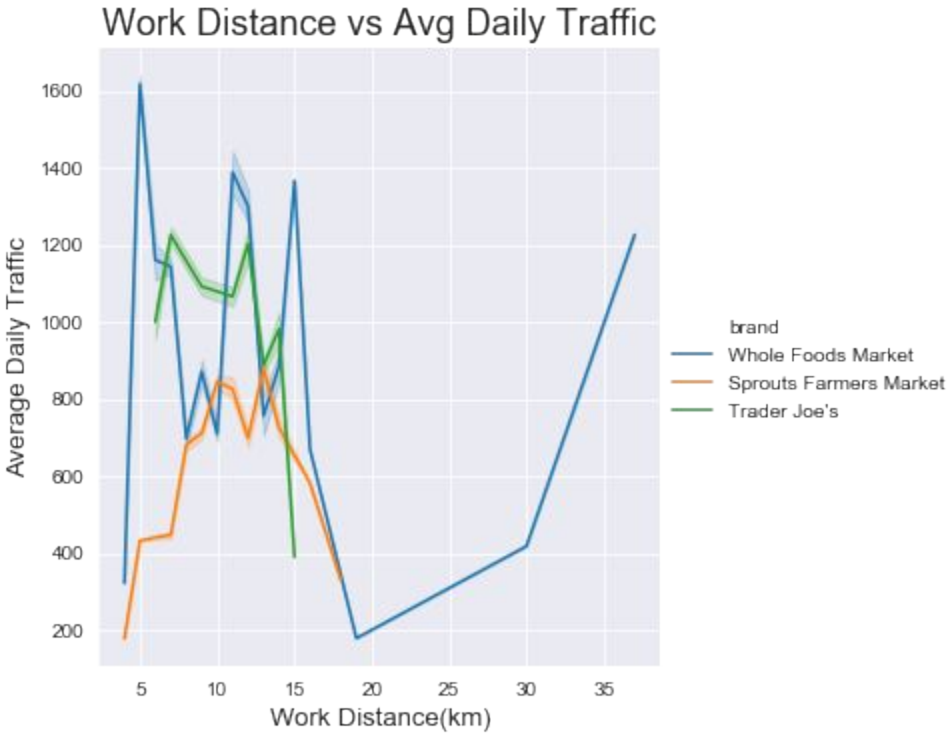
High pooulation density with low traffic fraction

Store location & Traffic fraction

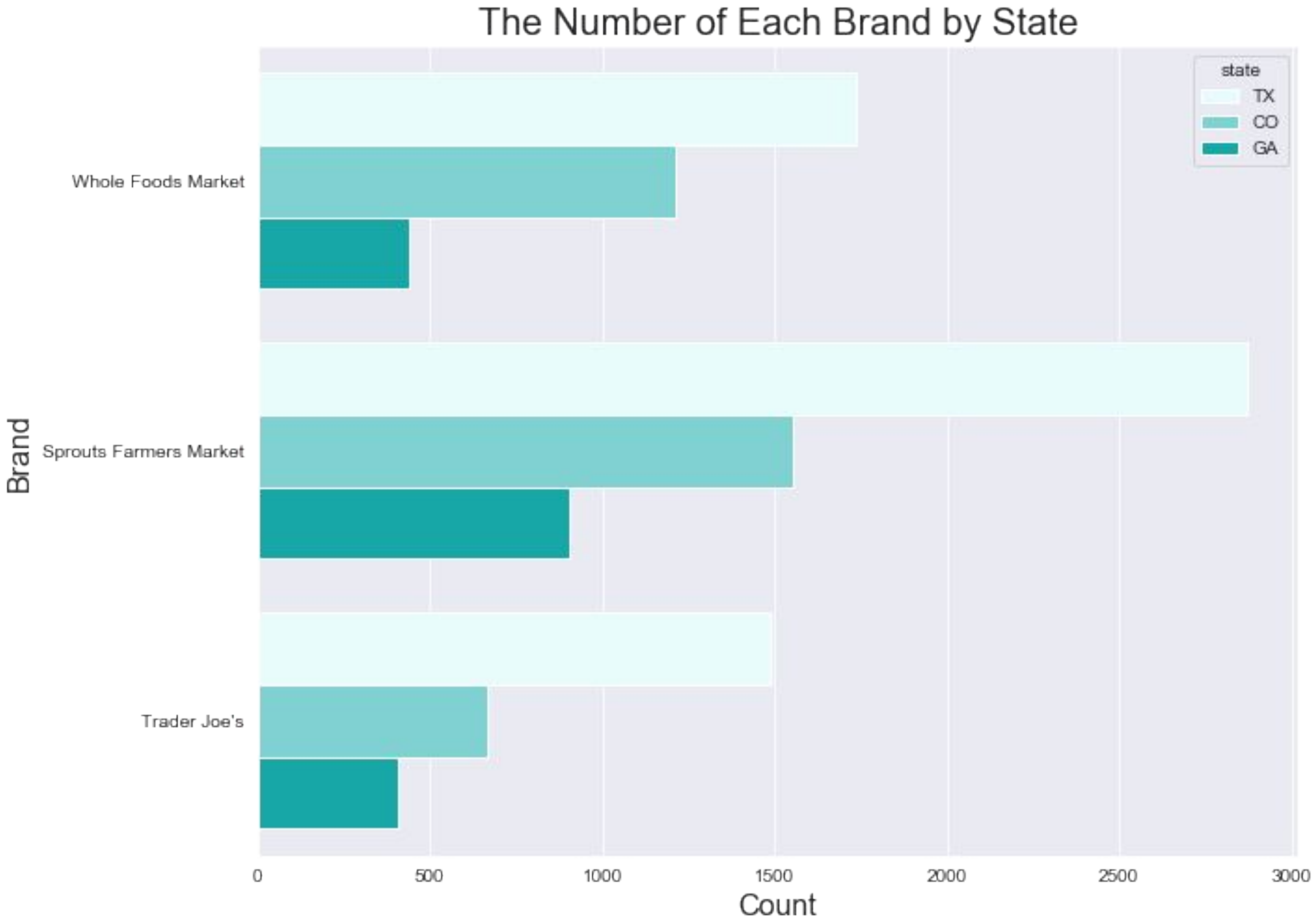


Average Daily Traffic vs Distance

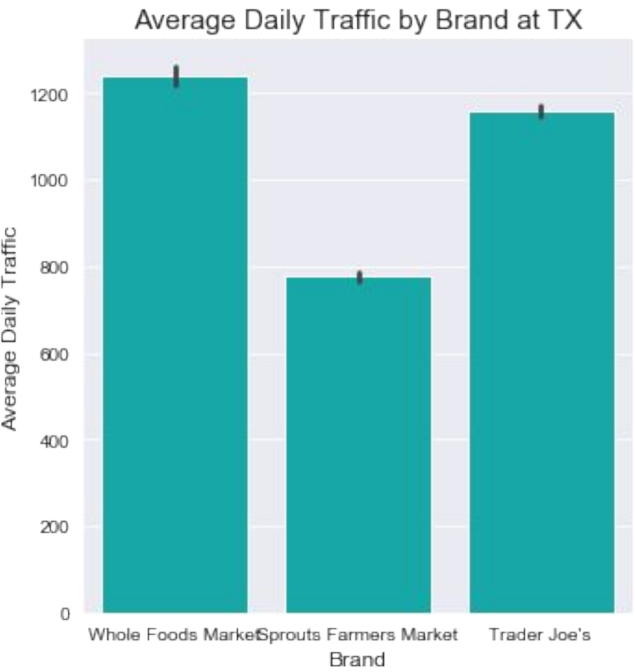


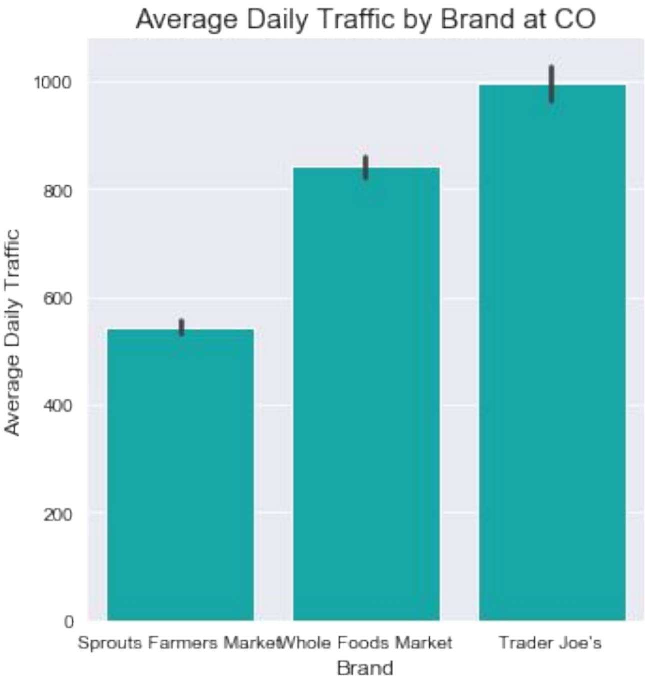


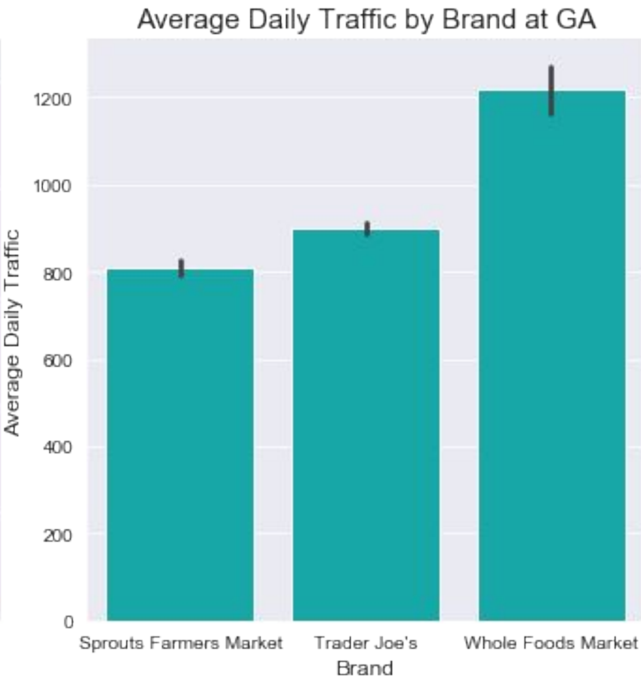
Number of Market



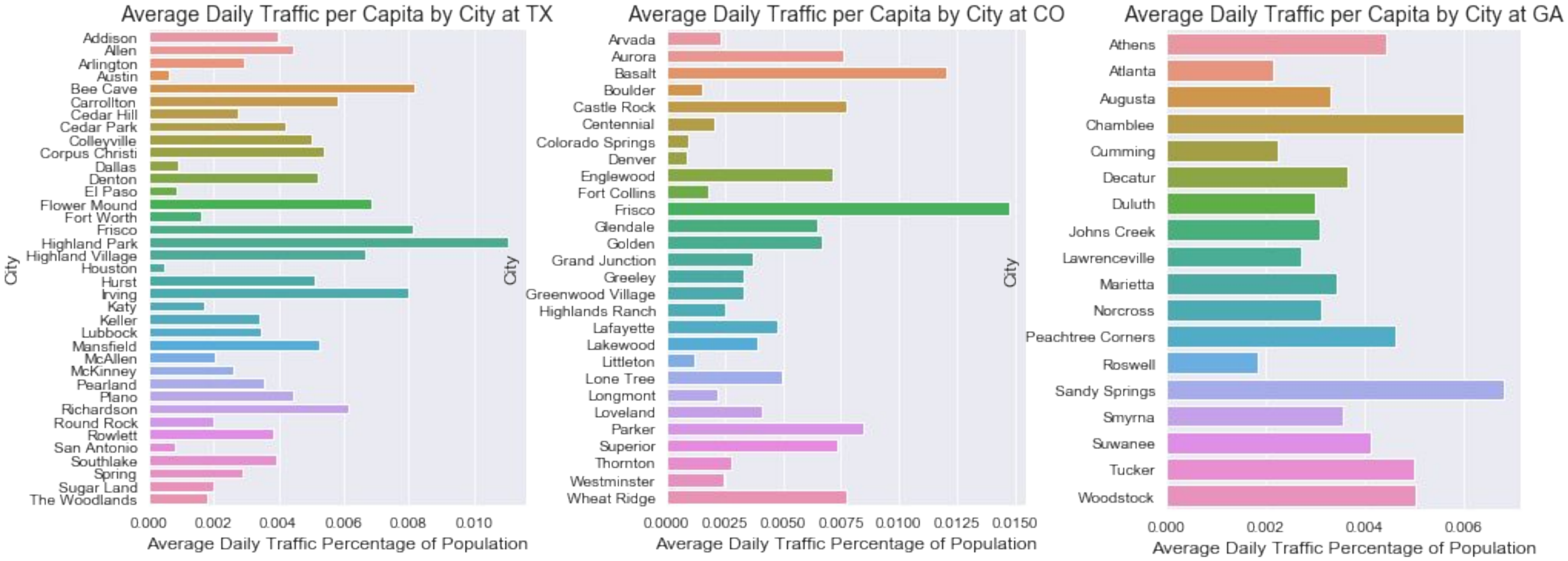
Average Daily Traffic by Brand







Average Daily Traffic in Percentage of Population



]

Machine Learning Model

Model Setting: Regressor!

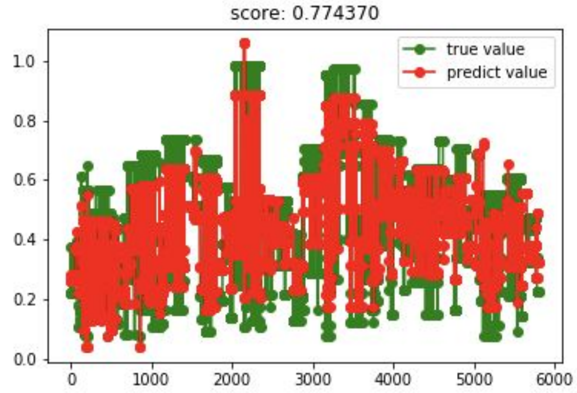
- Input (11)
  - State (3)
  - Home-market distance median
  - Home-market distance 25th percentile
  - Home-market distance 75th percentile
  - Workplace-market distance median
  - Workplace-market distance 25th percentile
  - Workplace-market distance 75th percentile
  - Community household annual income median
  - Community household annual income mean
- Output
  - Average traffic per day

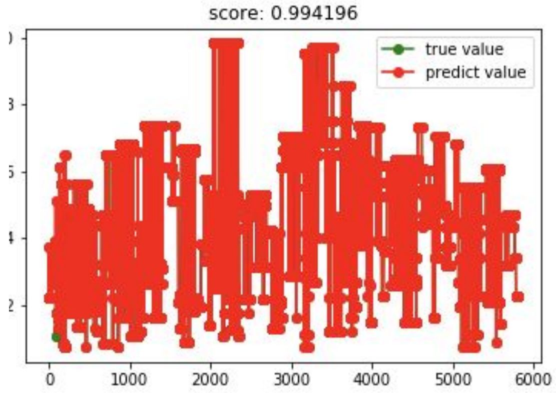
Data Preprocessing

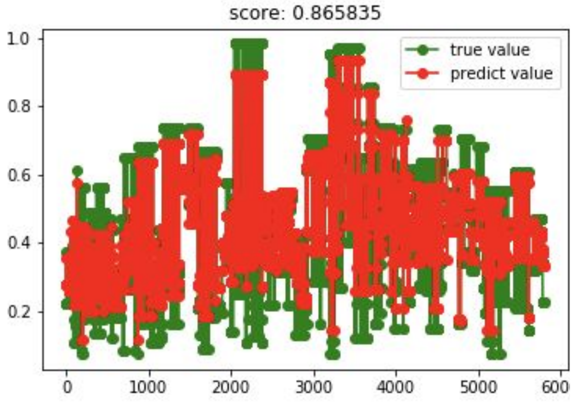
daily_traffic_avg	/2000.
state	one-hot
Distance (any)	/10.
Market count by city	/10.
Income (any)	/1000.

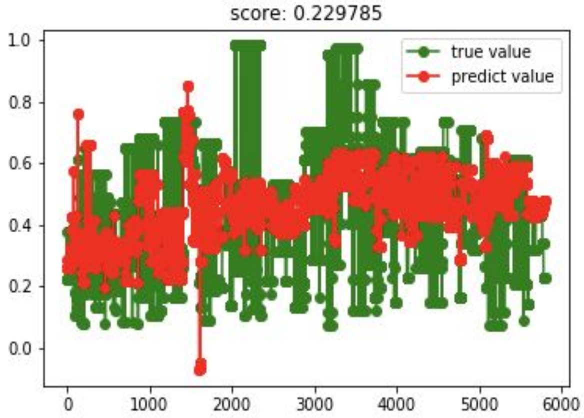
Different Model?

- **SVR** 0.774370
- **Decision Tree Regressor** 0.994196
- **Gradient boosting regressor** 0.865835
- **Linear Regression** 0.229785


score: 0.774370

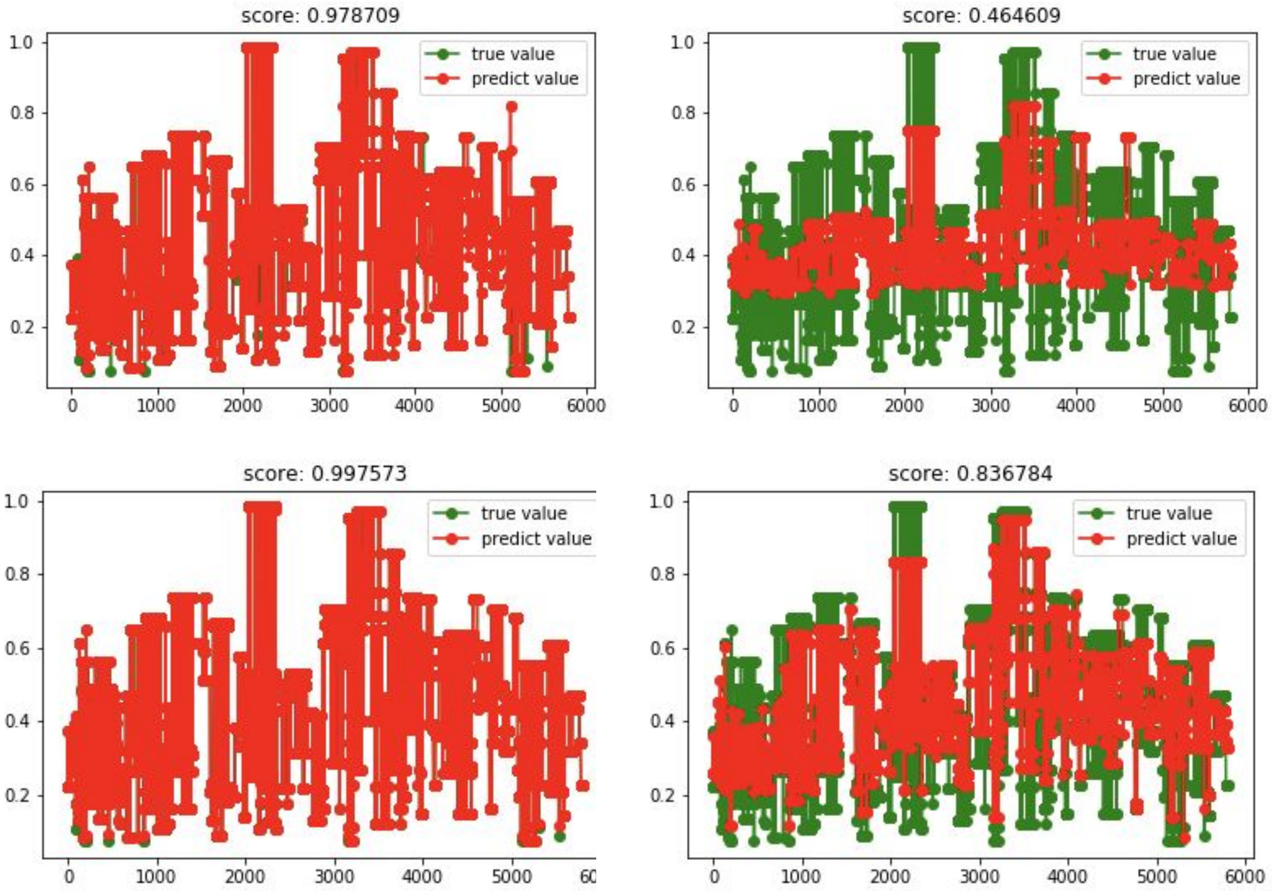

score: 0.994196


score: 0.865835


score: 0.229785

Different Model?

- **KNeighborsRegressor** 0.978709
- **AdaBoostRegressor** 0.464609
- **RandomForestRegressor** 0.997573
- **GradientBoostingRegressor** 0.836784



What we learn by models?

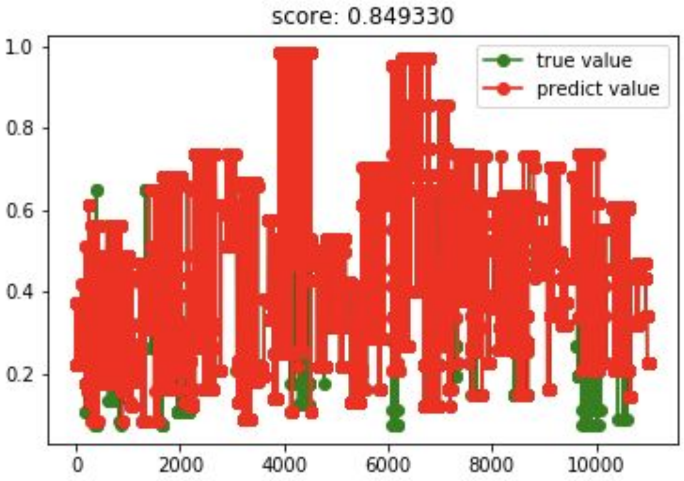
- Our problem is more nonlinear than linear:

  Linear Regression & Adaboost are miserable.

- Decision Tree family is good at this:

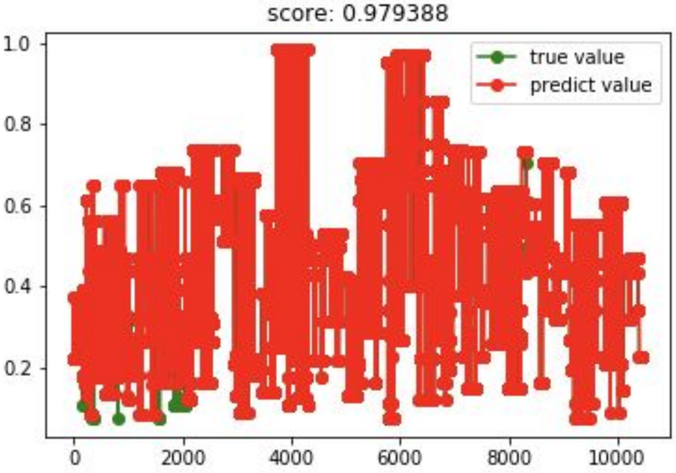
  They learn to form a knowledge map of states and cities, and easily leverage it by guessing where the incoming data point should fall

Data Split Comparison



0.849330
5%training

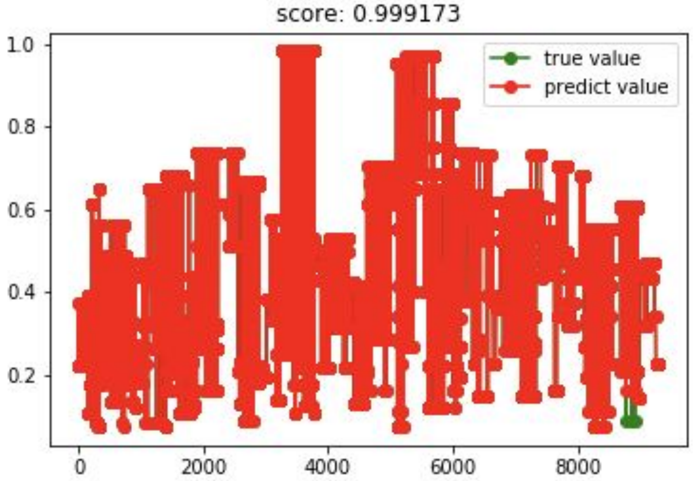
Data Split Comparison

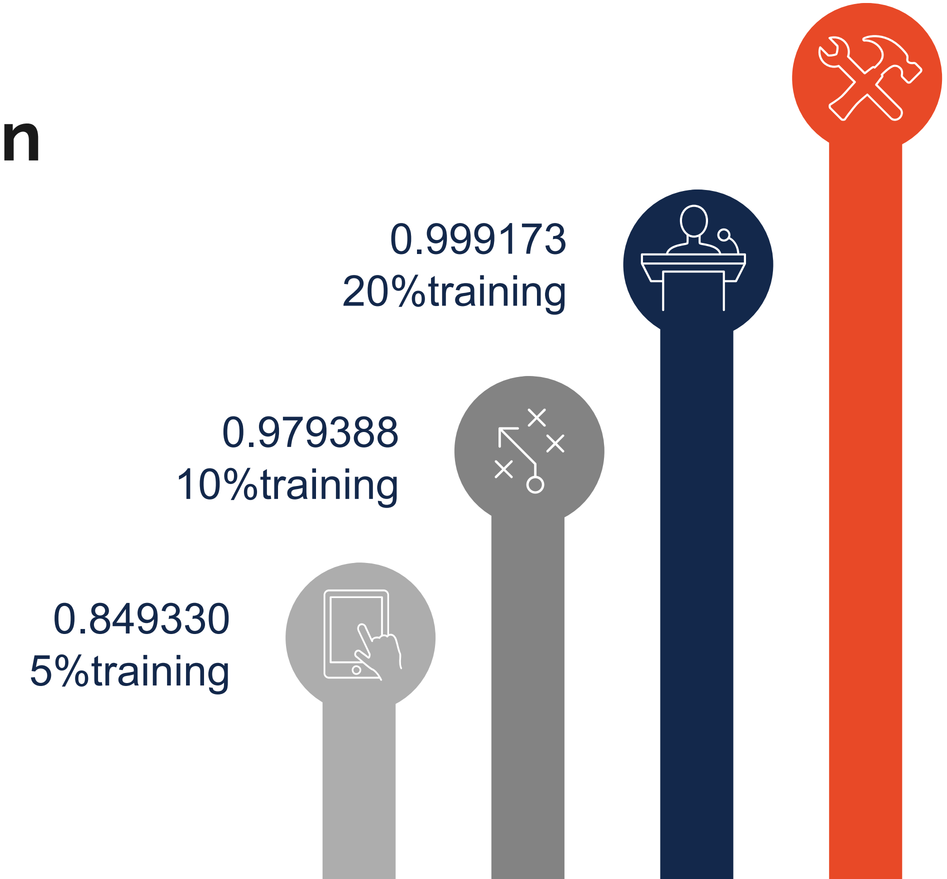


0.979388
10%training

0.849330
5%training

Data Split Comparison



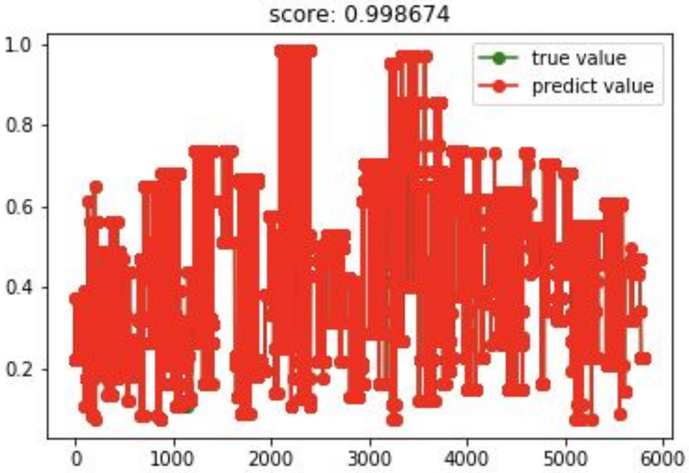


0.999173
20%training

0.979388
10%training

0.849330
5%training

Data Split Comparison



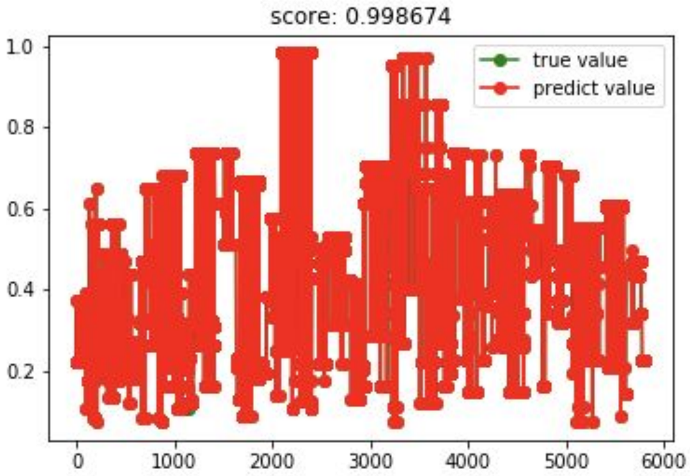
0.998674
50%training

0.999173
20%training

0.979388
10%training

0.849330
5%training

Random Seed?



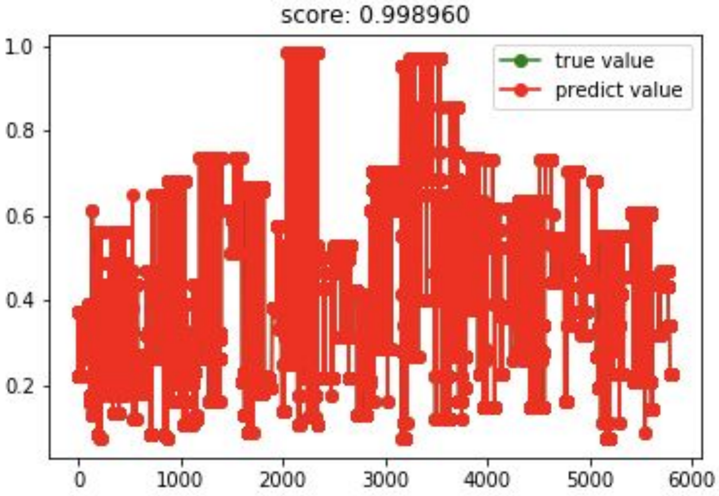
0.998674
50%training

0.999173
20%training

0.979388
10%training

0.849330
5%training

Random Seed?



0.998960
50%training
Seed = 100

0.998674
50%training

0.999173
20%training

0.979388
10%training

0.849330
5%training

What we know so far?

- Our model is soooo coollll!

  Predict well + **Generalize well!! -> caught some key ideas**

- Why they look similar even given different random seed?

  Did random sampling, but did not shuffle!

- The model is good, just because it is good! Nothing to do with seed initialization.

PCA Analysis - Variance

State: 9.56477545e-01 3.72834481e-02 4.93318813e-03 **(high)**

Distance : 4.69759238e-04 2.89794858e-04 2.63631342e-04 1.32852884e-04 6.97724107e-05 4.10084916e-05

Count of competitors in one city: 2.37527445e-05

Income: 1.18478925e-05 3.39851046e-06 **(low)**

Go into the future

- Combine Google Map, user simply needs to provide store location to get the predicted customer traffic.
- Expand the database to other industries such as real estate, restaurant.
- Obtain more potential features such as local GDP and local tax to make the model more powerful.

Thank you.