Macrodata Refinement

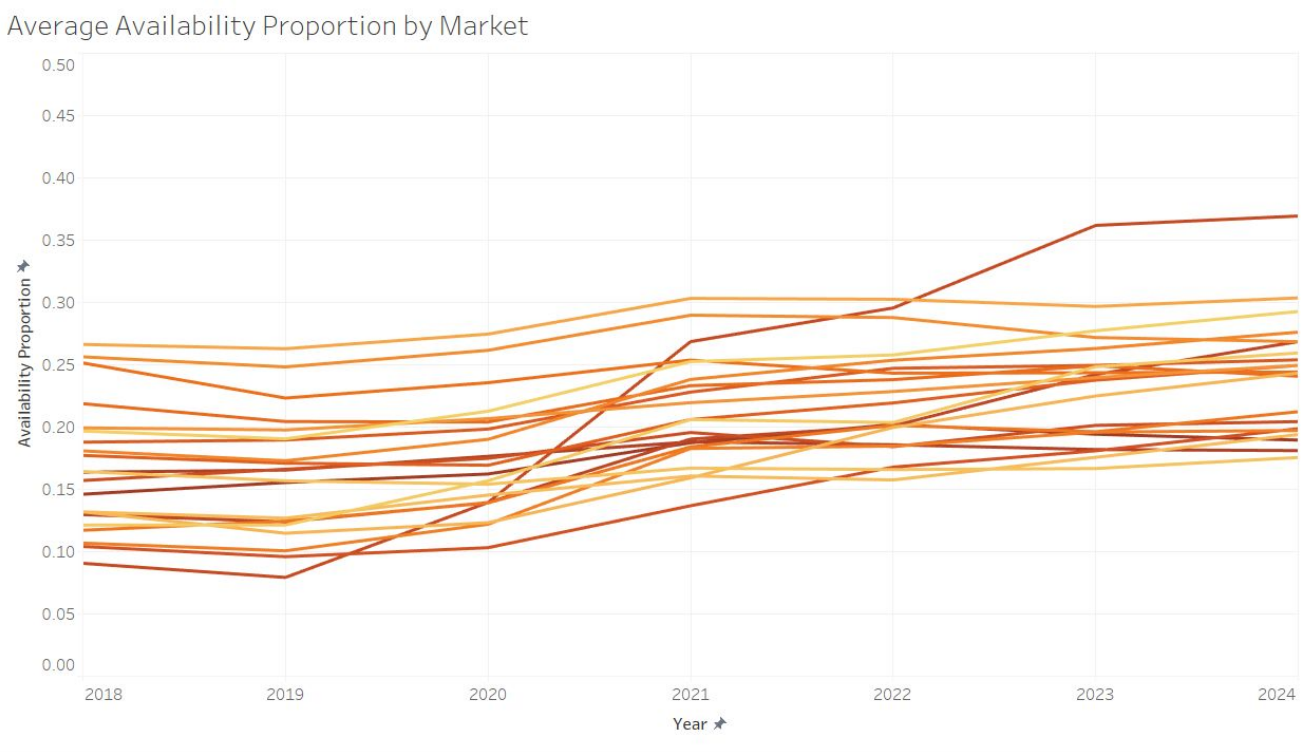
The work is mysterious and important

savills

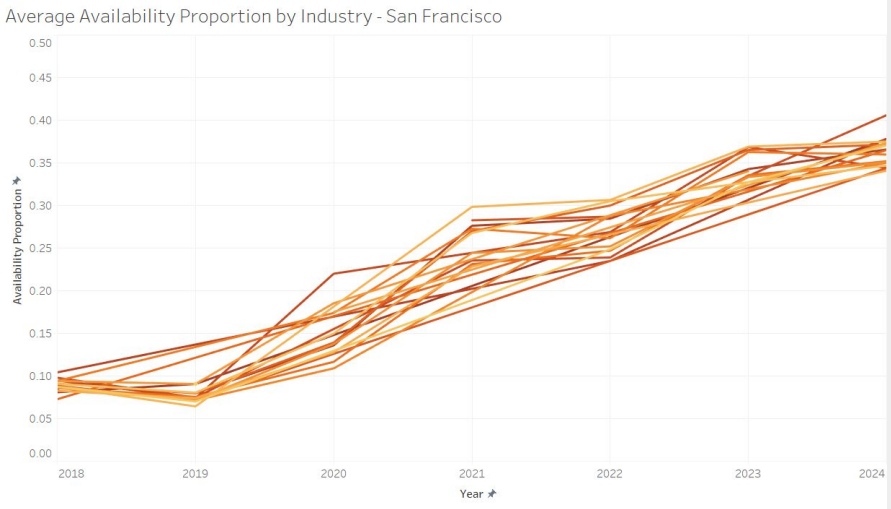
Agenda

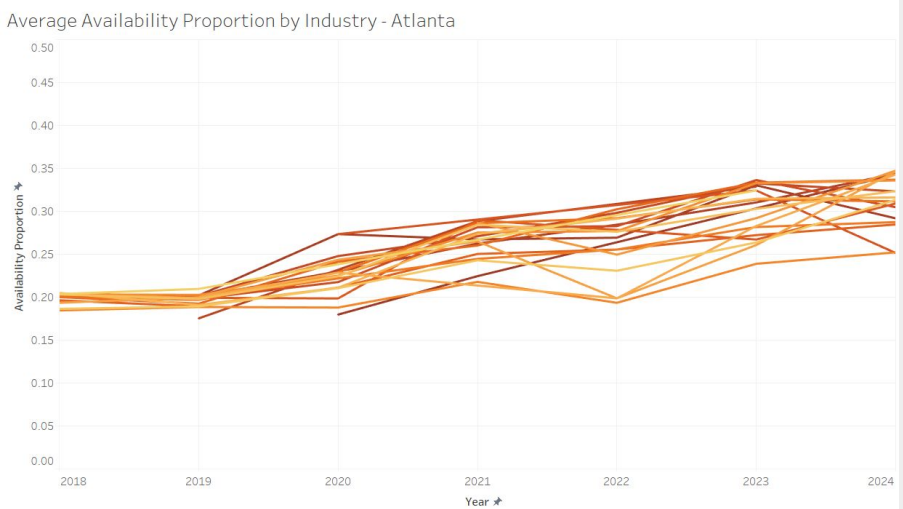
1. Market Trends
2. Problem
3. Our Solution
4. Model in Context

Market Trends

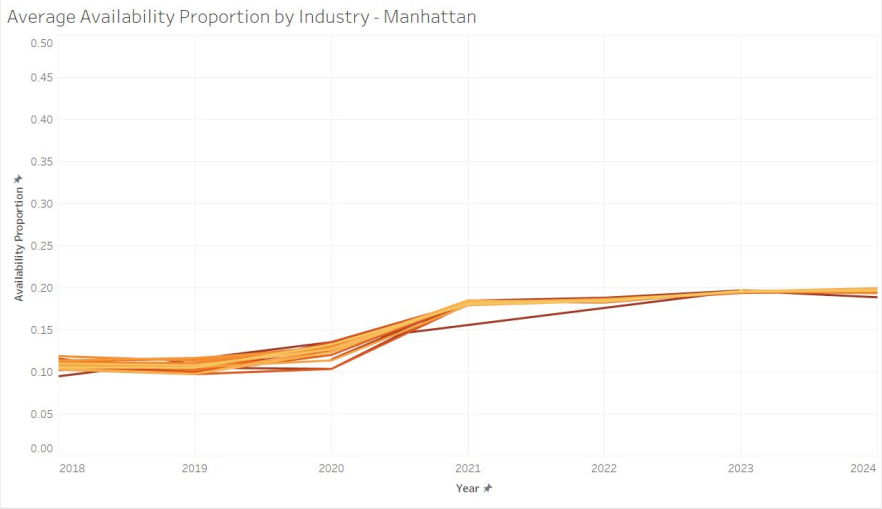


Market Trends

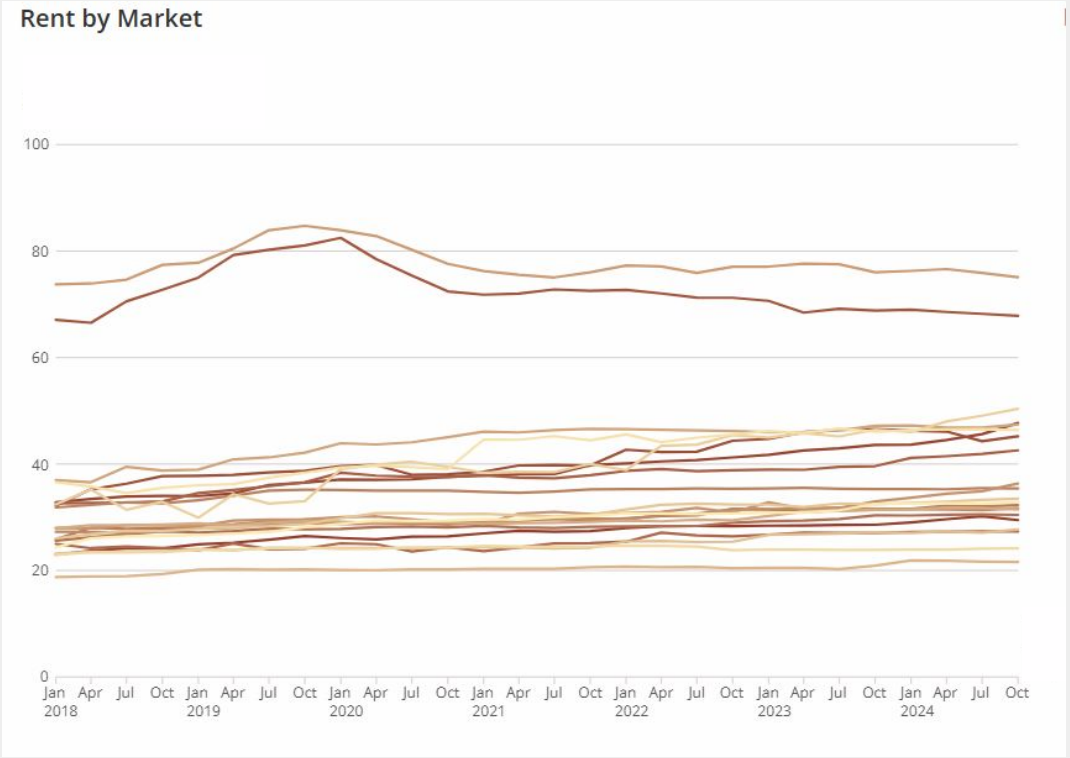




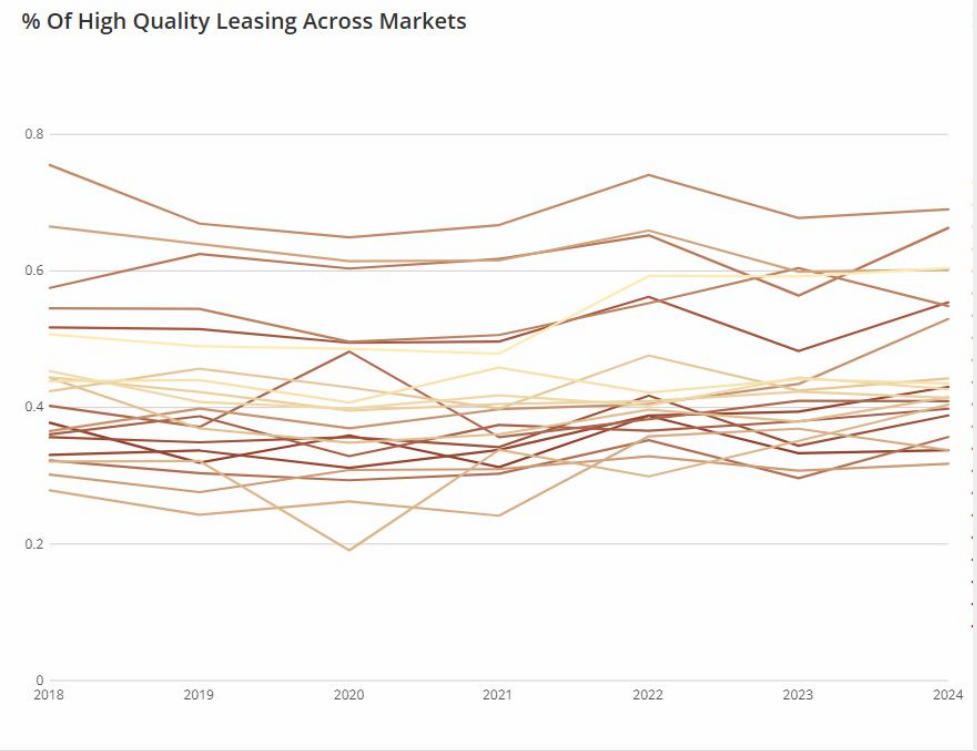
Market Trends



Market Trends



Market Trends



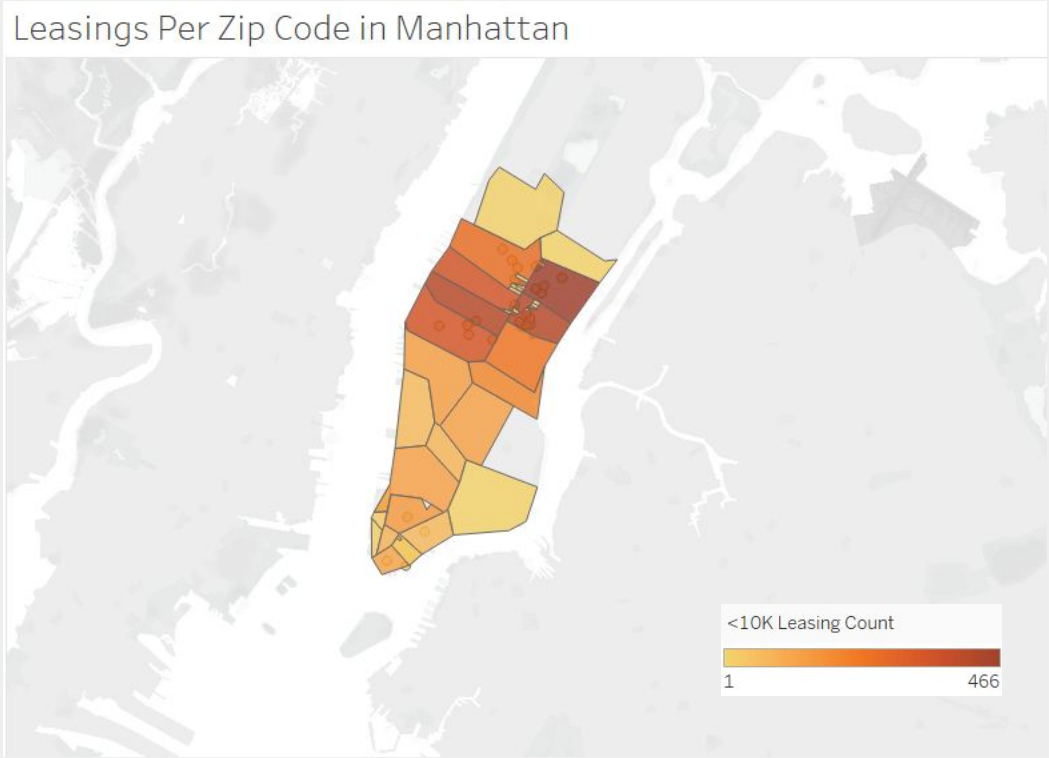
Our Question

Given some company needs & wants, **where in the market** is best?

- Most companies **may already have a market** in mind.
- Provide our answer as a **Zip Code.**
- Each market is different - focus on **Manhattan**, but easily scalable.

Quantify company demands...

Manhattan



Features

Market-level specifics: (leases.csv)

- Available Space, Rent Prices, Square Footage
- Give a sense of **commercial real-estate market performance over time**

Connectivity: (leases.csv)

- Number of New Leases Per Industry (Per Quarter), Population, Businesses Per Capita
- Quantifies **how connected** to Zip Code is in its industry

Features

Crime: (Historical New York City Crime Data - NYPD)

- Felonies, Misdemeanors, Violations
- How **safe** is this Zip Code?

Job Listings: (New York City Job Postings - City of New York)

- New York City Job Postings, broken by part/full time, level of experience (Expert, Entry-Level, Student, ...)
- How **available is the job market**?

Features

Education: (US Colleges and Universities - NYC OpenData)

- Total Enrollment & Employment for university students in the area
- How **available are college students**?

Consolidating the data:

- Not all data is perfectly Zip Code exact
- Impute missing crime and job statistics **based on average of 3 nearest complete values**, grouped by industry, quarter
- Impute remaining small proportion **by median** across industry, time, zip code

Model Architecture

28 features (3 categorical)
Train/Test 80/20

Built on CatBoostClassifier:

- Multicollinearity would be an issue in regression **Decision trees gives advantage**
- Industry is critical, want to **handle categorical variables well**
- CatBoost is **faster** and **prevent overfitting better** than alternatives, especially in a short amount of time
  - Further split 80/20 Train/Validation to accomplish this

~ **95.77%** accuracy, **96%** recall, **96%** precision

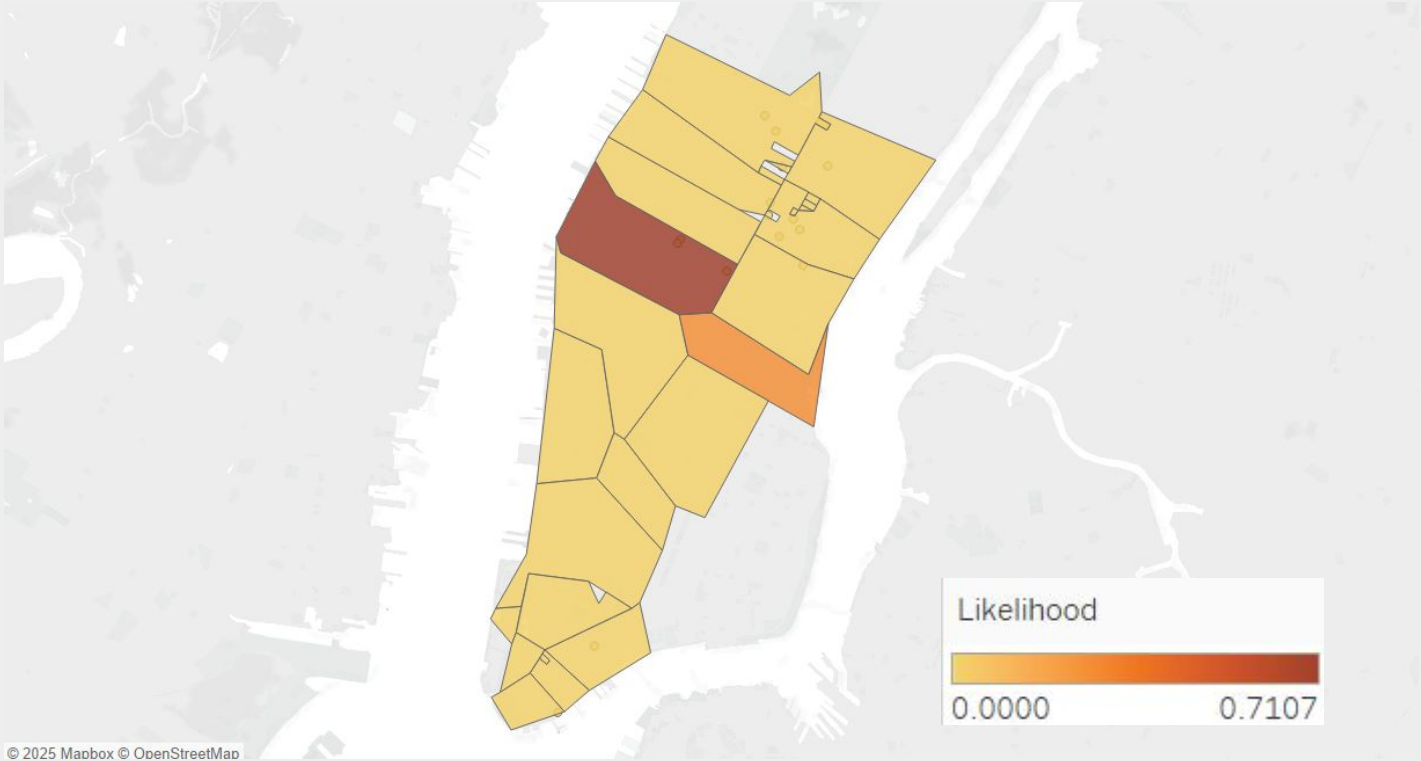
Model in Context

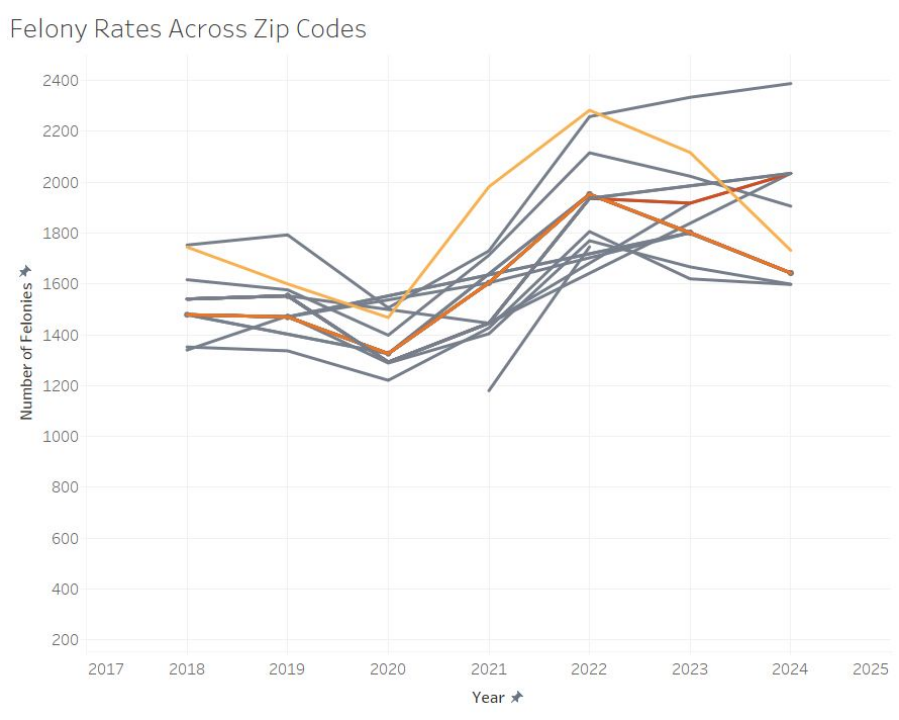
Krish Tech 

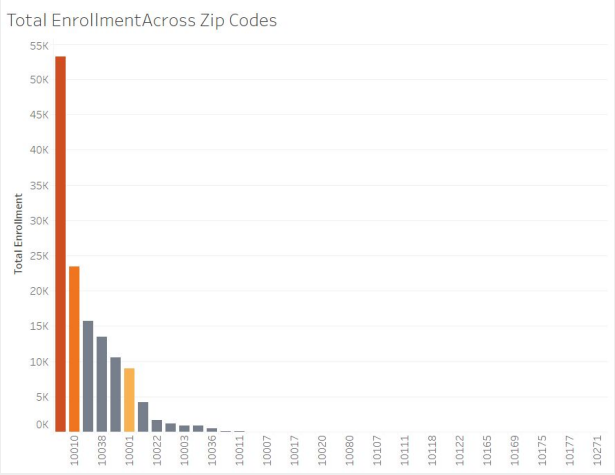
- Tech Company
- Average Sq Ft (~13k), Rent
- Wants Connections
- Wants to be Safe
- Needs Students

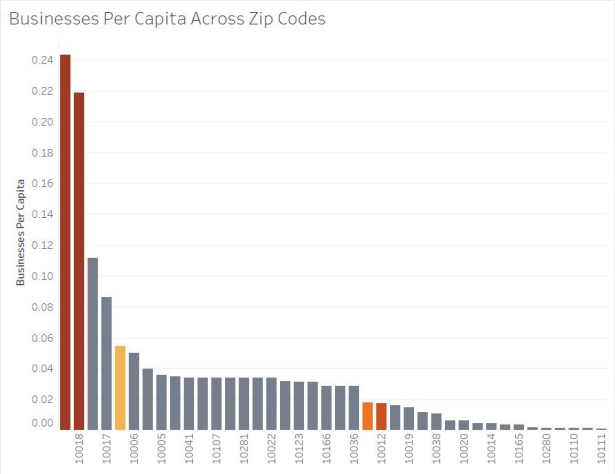
- Median Lease Stats
- 75th in Businesses per Capita, Pop.
- 10th in Crime
- 90th in Student Job Listings, Enroll

Model in Context




Felony Rates Across Zip Codes
Number of Felonies
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
2017
2018
2019
2020
2021
2022
2023
2024
2025
Year


Total EnrollmentAcross Zip Codes
Total Enrollment
55K
50K
45K
40K
35K
30K
25K
20K
15K
10K
5K
0K
10010
10038
10001
10022
10003
10036
10011
10007
10017
10020
10080
10107
10111
10118
10122
10165
10169
10175
10177
10271


Businesses Per Capita Across Zip Codes
Businesses Per Capita
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10018
10017
10006
10005
10041
10107
10281
10022
10123
10166
10036
10012
10019
10038
10020
10014
10165
10280
10110
10111

Limitations

- Our model is trained on specific markets, but is **highly scalable**
  - Can utilize pipeline for other markets, as all data used is widely available for multiple markets

- Assumes businesses have a market in mind
  - Could expand model to consider **multiple markets** simultaneously

- Zip codes between data sets did not match perfectly
  - Solved by **imputing missing values** using geographically proximate data

Summary

- Rather than analyzing all markets at once, **focus on a single market**
- Use leasing data, business & population data, crime rates, job and education data to formulate **quantifiable demands** of a company
- Use these features to **predict what zip code is most desirable**
- **Highly accurate** and **picks up trends** otherwise hard to notice

Thank you!