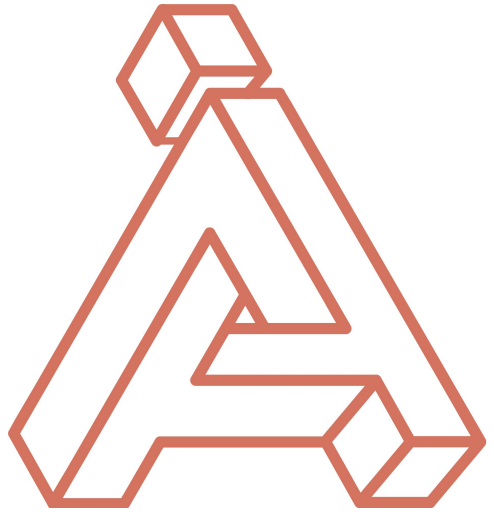


AI Hack

The AirBnB Challenge

Una, Shawn, Dimitri, Jaume

February - March 2020


AIRBNB
Investment Strategy
AIHACK20 – TEAM 10

Airbnb Project Overview

Business Problem

Researches show current **rental market** in top tier cities e.g. New York, London has a **very low yield** of ~2%. Airbnb could be a better investment

Project Aim

- Help investors to estimate Airbnb **Monthly Revenue**;
- Identify and recommend the **most profitable** neighbourhood in New York.

Key Factors for Predict

Room Properties
- No. Accommodates
- Room Type (private/shared)
- Amenities
- Cancellation Policy etc.

Geographical Features
- Neighbourhood (bar/restaurant/entertainment)
- Competitors Pricing
- Transport etc.

Demographic Features
- Age
- Household Income etc.

Result: Top 6 Factors Affecting Monthly Revenue

1. No. Accommodates
2. Room Type (Entire home/apt)
3. Location
4. Convenience / Accessibility
5. Neighbourhood (Entertainment)
6. Cleanliness

Technical Analysis - Pre-Processing

Data Preprocessing

Created listings.csv copy with state of New York listings

Split into training and test set using Scikit-Learns train_test_split(test_size = 0.2). Proceeded with training set

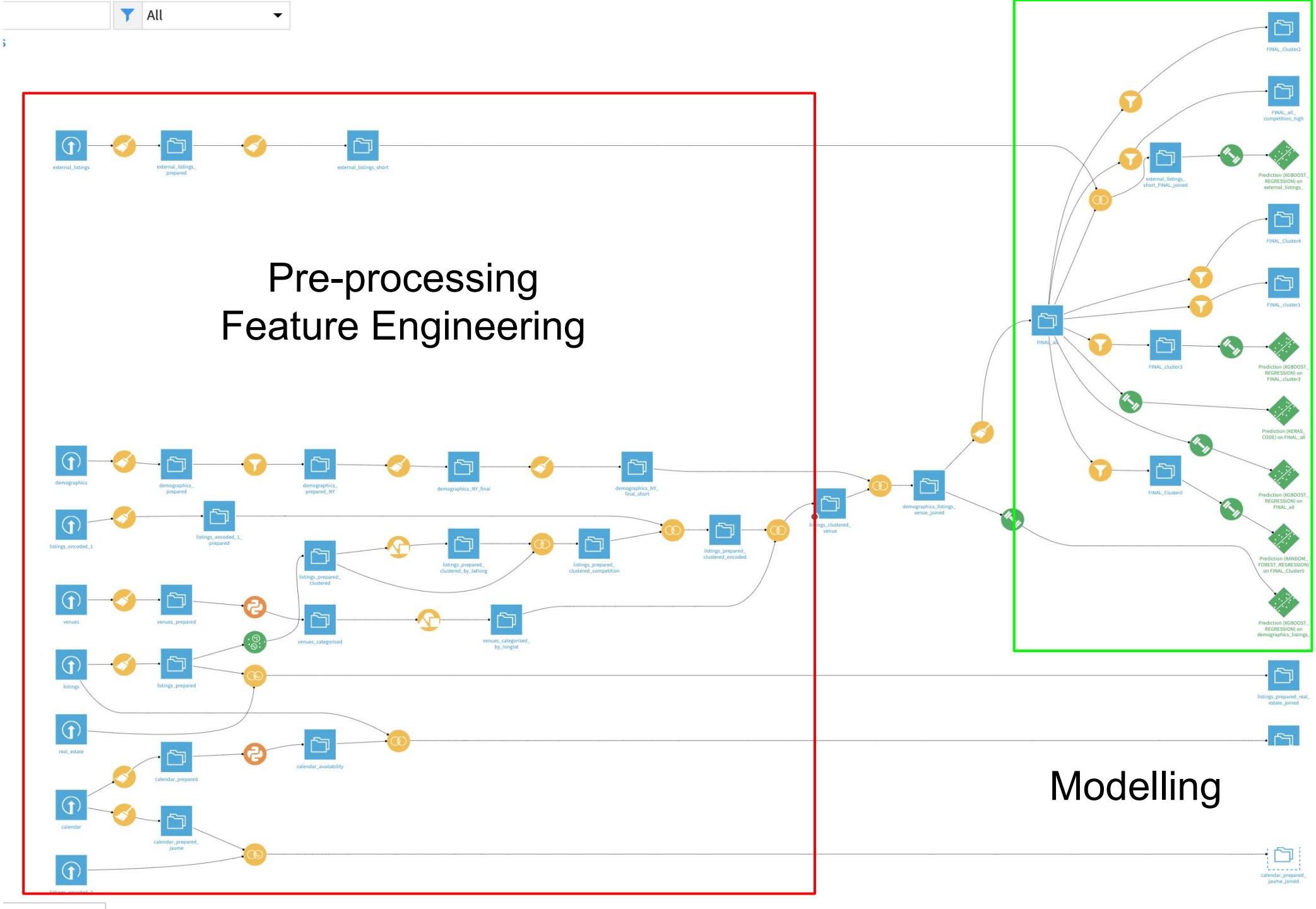
Used Scikit-Learns SimpleImputer(strategy = ‘median’) to correct NaN entires

Used Scikit-Learns OneHotEncoder on categorical features such as ‘room_type, cancellation_policy’

Created numeric Pipeline with Imputer and Standard Scalar

Created full Pipeline using ColumnTransformer on the numeric Pipeline and categorical OneHotEncoding

Process resulted in 40,731 x 238 dataframe ready for training and validation

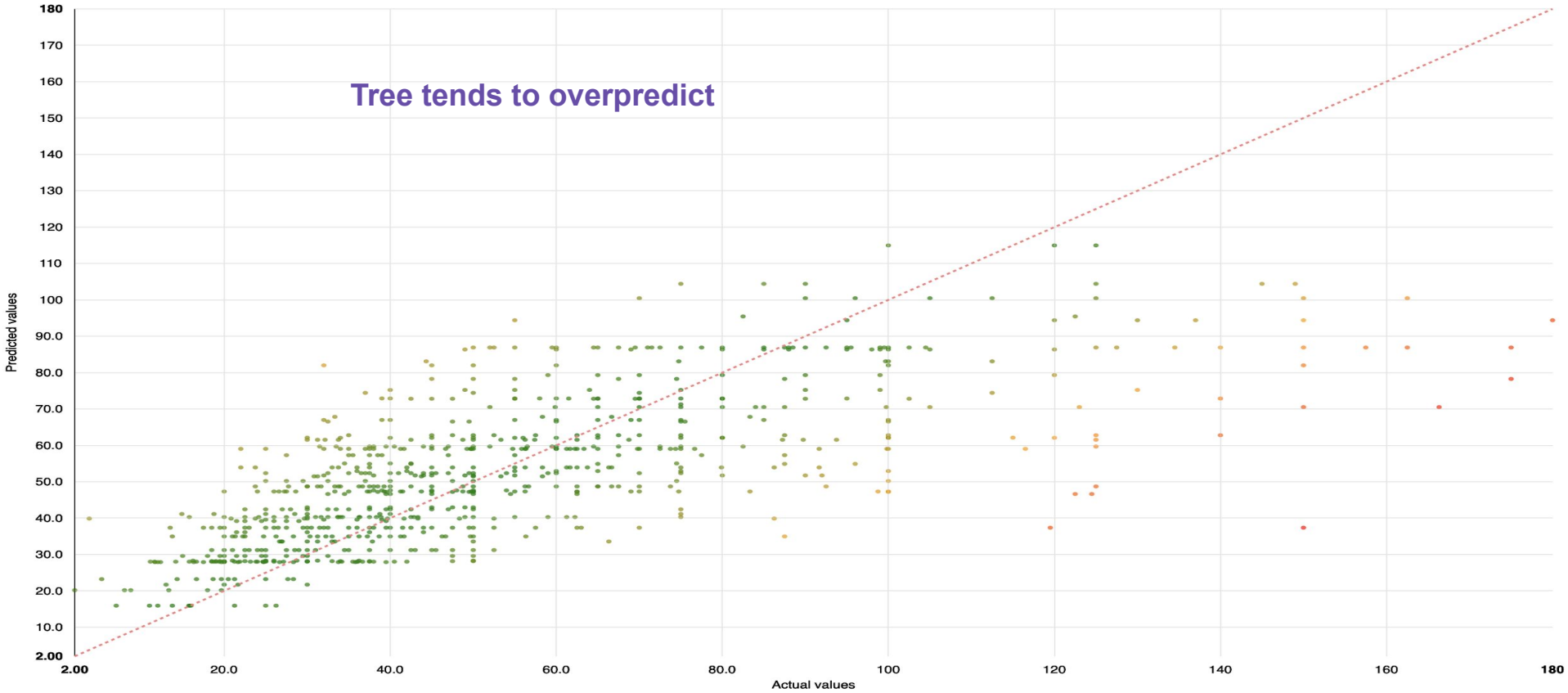

All
Pre-processing
Feature Engineering
Modelling

Technical Analysis - Modelling

Decision Tree

- Used for its interpretability and ranking of factor importance.
- Residual Analysis - large deviation
- **No significant** difference when relevant features selected by us rather than automatically selected
- Being close to many **museums**, **bars** and other **entertainment** options, along with # of people accommodated and having entire home most important.

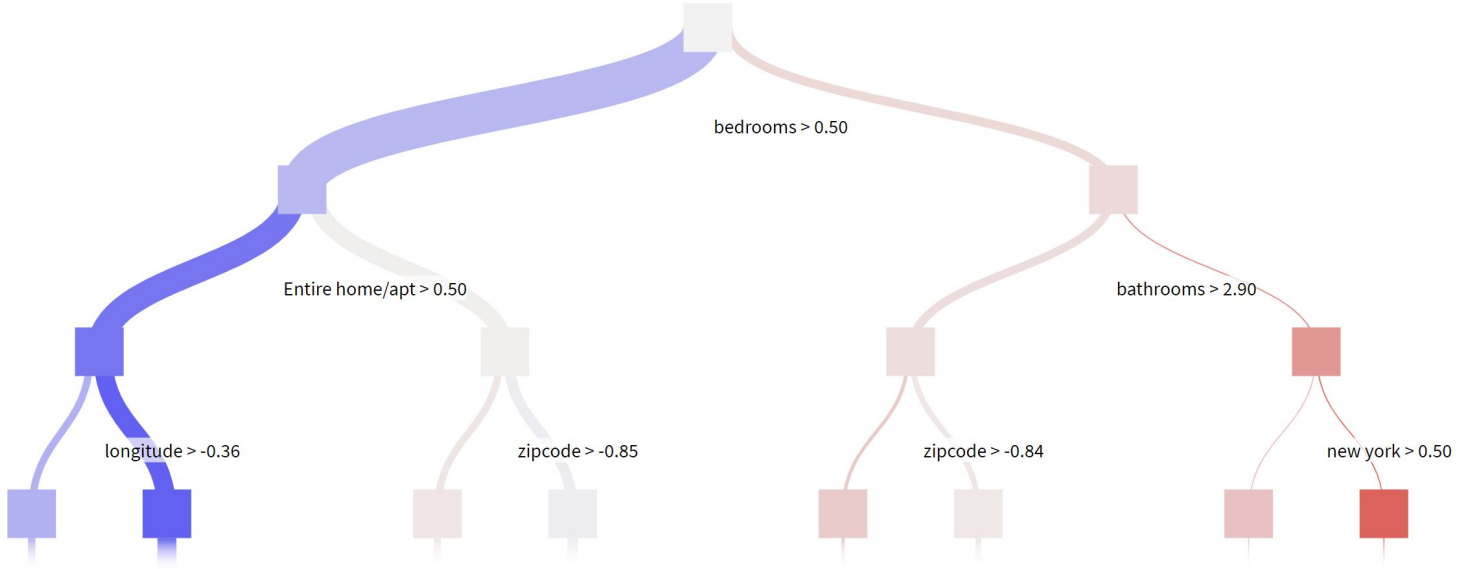
Minimum	25th perc.	Median	75th perc.	90th perc.	Maximum
-37.321	-12.923	-4.0218	8.4320	23.378	62.280
	Average	-0.79236		Standard deviation	19.458



Random Forest

Key Features

feature_name	importance
Entire home/apt	0.260380194
bathrooms	0.254297223
bedrooms	0.089700301
zipcode	0.085265961
longitude	0.074653639
latitude	0.067644101
accommodates	0.041737028
new york	0.015034849
Apartment	0.011890113
Suitable for events	0.011760542



XGBoost

- Selected to better fit the significant residuals , especially many overpredictions.
- Maximum depth of 3 helps prevent overfitting.
- **Best performance** - a few features seem to explain a lot of the residual clusters. Suggests there are clusters of neighbourhoods, # of people accommodated.
- **Most important features** - # of people accommodated, latitude & longitude, availability in the next 30 days, having entire home.

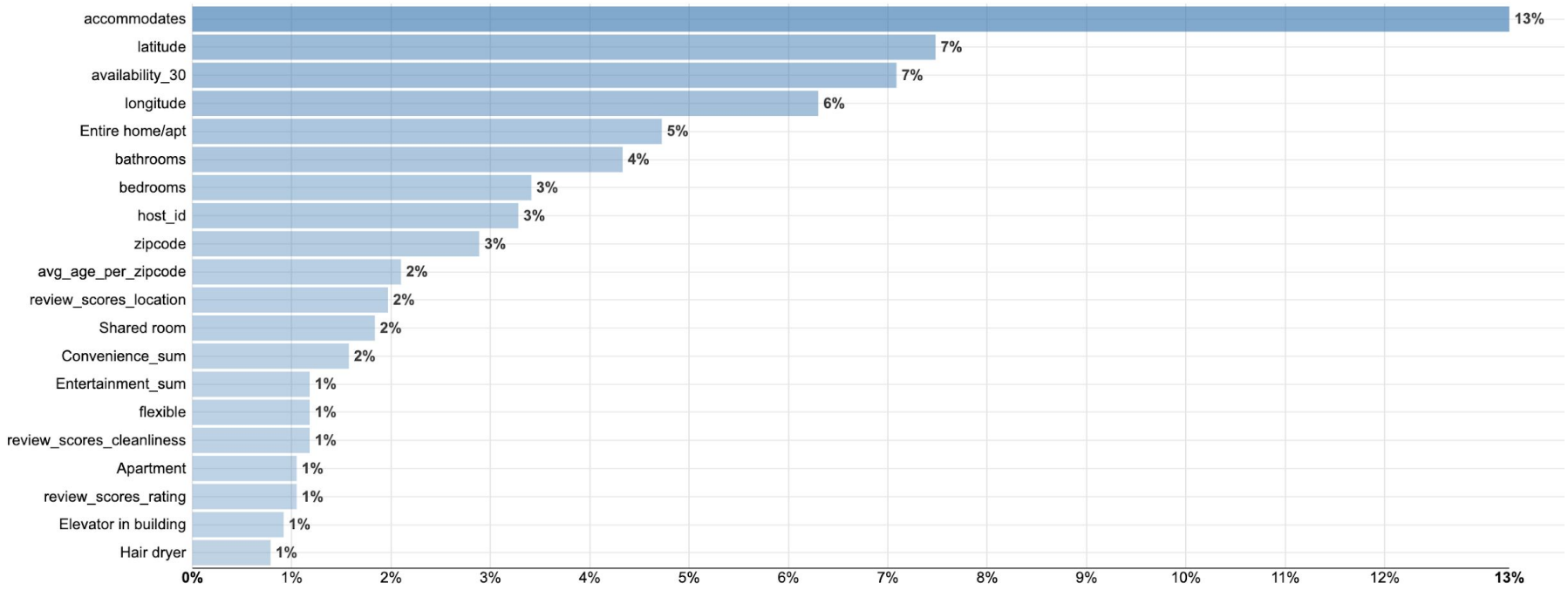
XGBoost

XGBoost | 0.581 | ✔ Done 10 hours ago (2020-03-01 00:40:45)

Trees 111
Max depth 3

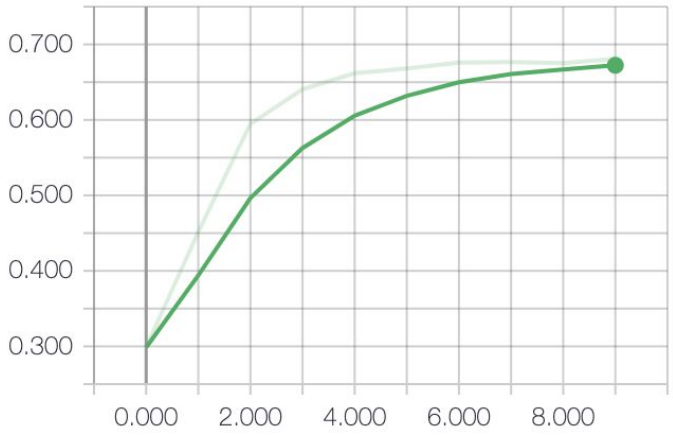
Most important variables
accommodates
latitude
availability_30
longitude
Entire home/apt
bathrooms

Train set 31882 rows
Test set 7944 rows
Train time about 58 seconds
MAKE ACTIVE

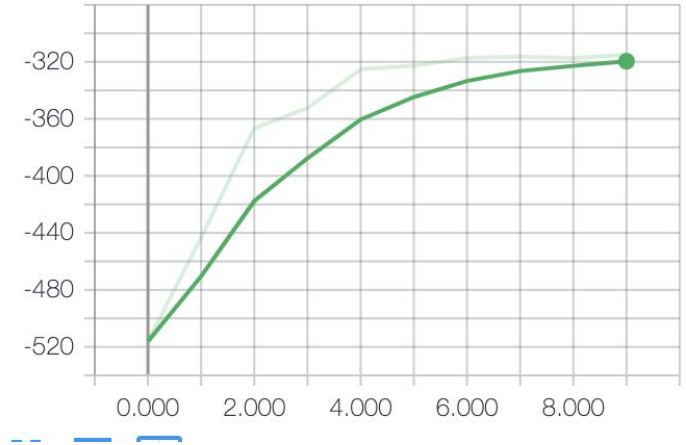


Deep Neural Network

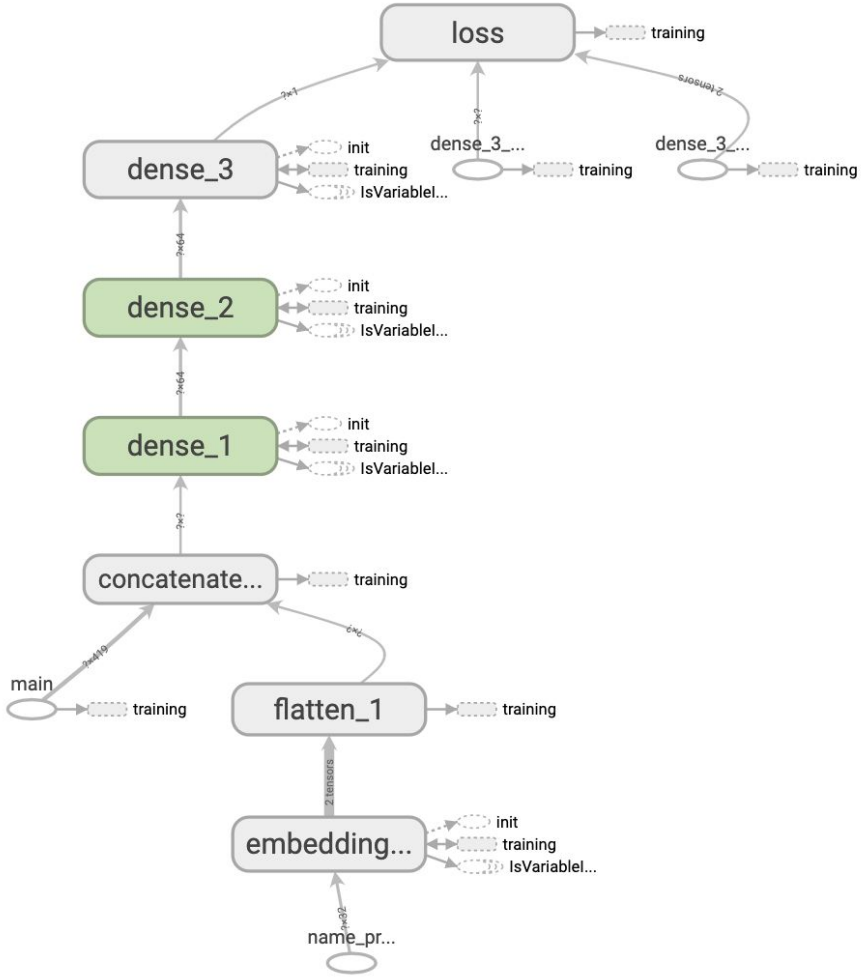
Test EVS



Test MAE



Deep Neural Network

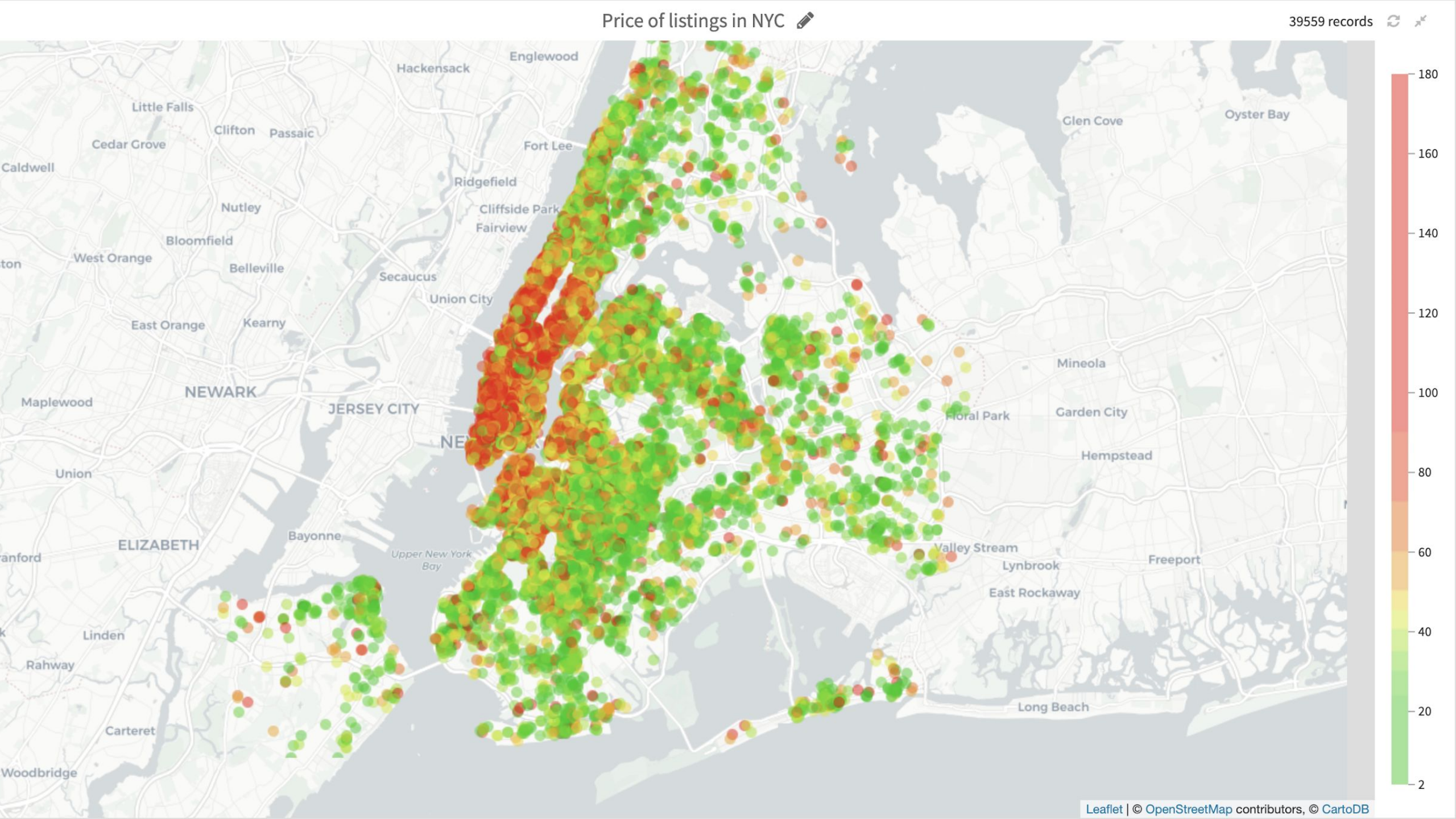


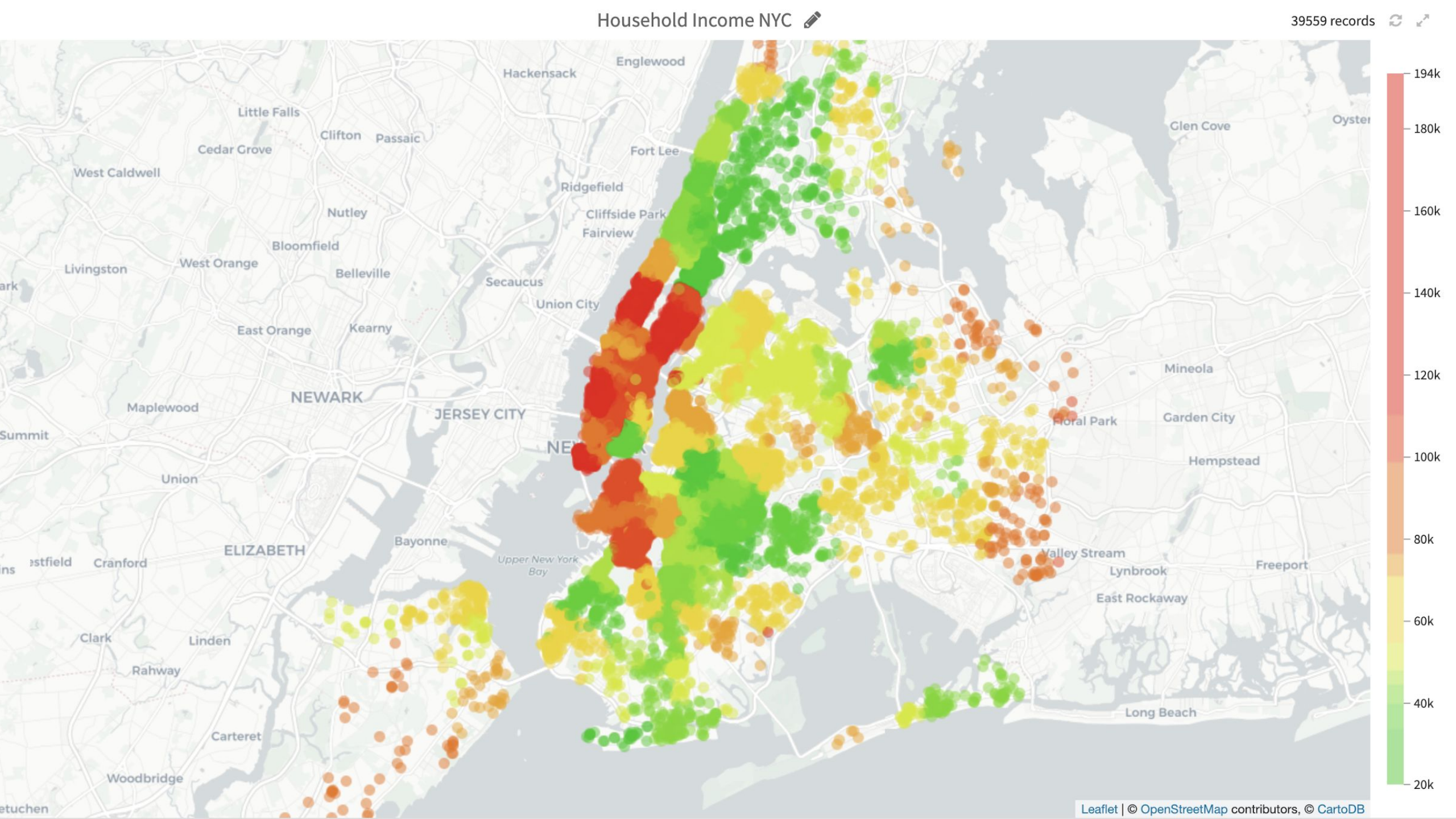

```
    # This input will receive preprocessed text from 'name' column
    input_name_preprocessed = Input(shape=(32,), name="name_preprocessed")
    x_name_preprocessed = Embedding(output_dim=512, input_dim=10000, input_length=32)(input_name_preprocessed)
    x_name_preprocessed = Flatten()(x_name_preprocessed)

    x = concatenate([input_main, x_name_preprocessed])

    x = Dense(64, activation='relu')(x)
    x = Dense(64, activation='relu')(x)

    predictions = Dense(1)(x)

    # The 'inputs' parameter of your model must contain the
    # full list of inputs used in the architecture
    model = Model(inputs=[input_main, input_name_preprocessed], outputs=predictions)

    return model

# Compile your model and return it
#   model   - model defined in 'build_model'
def compile_model(model):

    # The loss function depends on the type of problem you solve.
    # 'mse' is appropriate for a regression.
    model.compile(optimizer='rmsprop',
                  loss='mse')
```

Technical Analysis - Visualisations

Created Geospatial Visualisations on NYC:

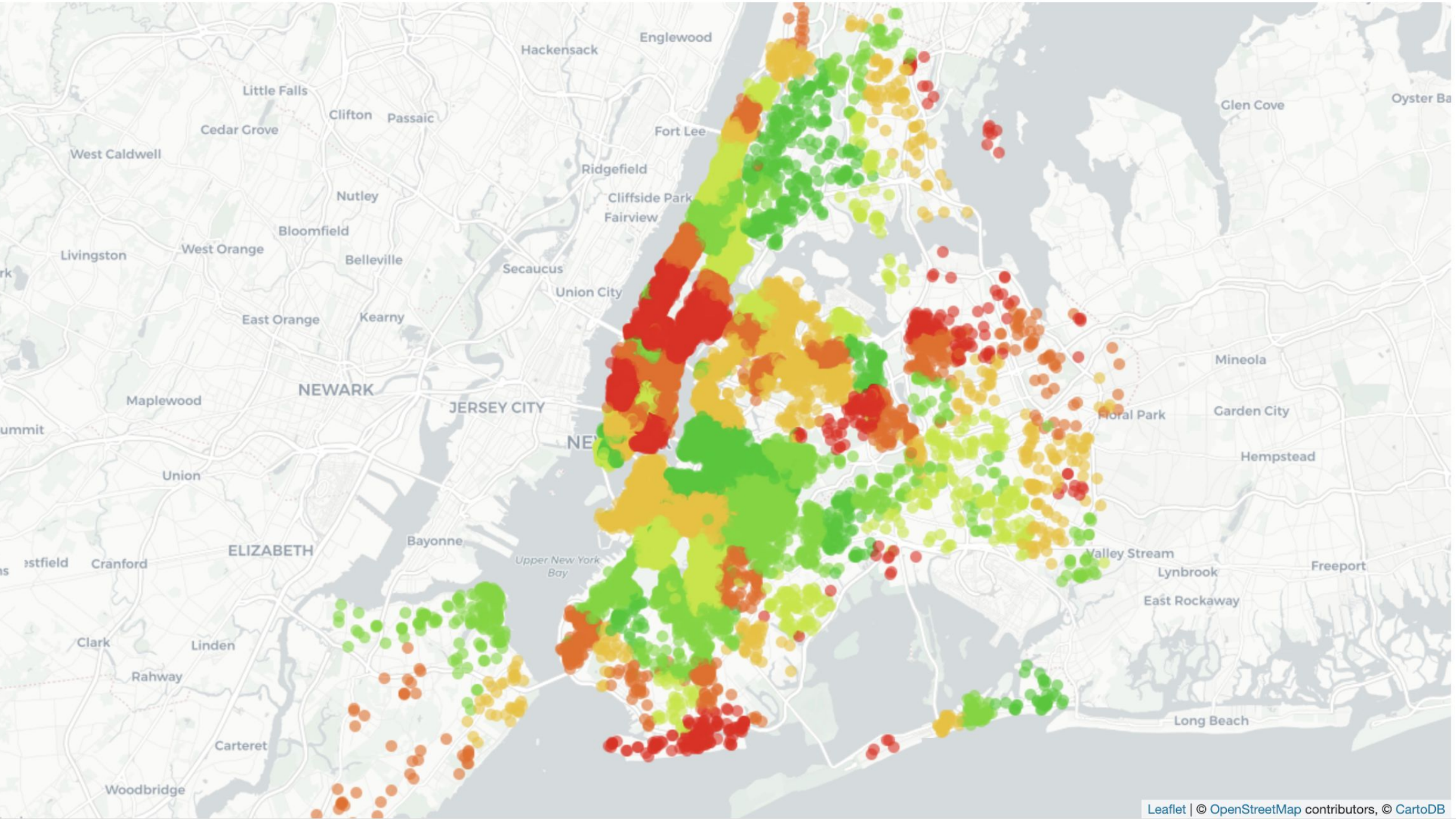
- Listing Price
- Median Household Income
- Mean Age


Price of listings in NYC
39559 records
Englewood
Hackensack
Little Falls
Clifton
Passaic
Cedar Grove
Caldwell
Fort Lee
Ridgefield
Nutley
Cliffside Park
Fairview
Bloomfield
West Orange
Belleville
Secaucus
Union City
East Orange
Kearny
NEWARK
JERSEY CITY
Maplewood
Union
ELIZABETH
Bayonne
Upper New York Bay
Linden
Rahway
Carteret
Woodbridge
Glen Cove
Oyster Bay
Mineola
Floral Park
Garden City
Hempstead
Valley Stream
Lynbrook
Freeport
East Rockaway
Long Beach
180
160
140
120
100
80
60
40
20
2
Leaflet | © OpenStreetMap contributors, © CartoDB

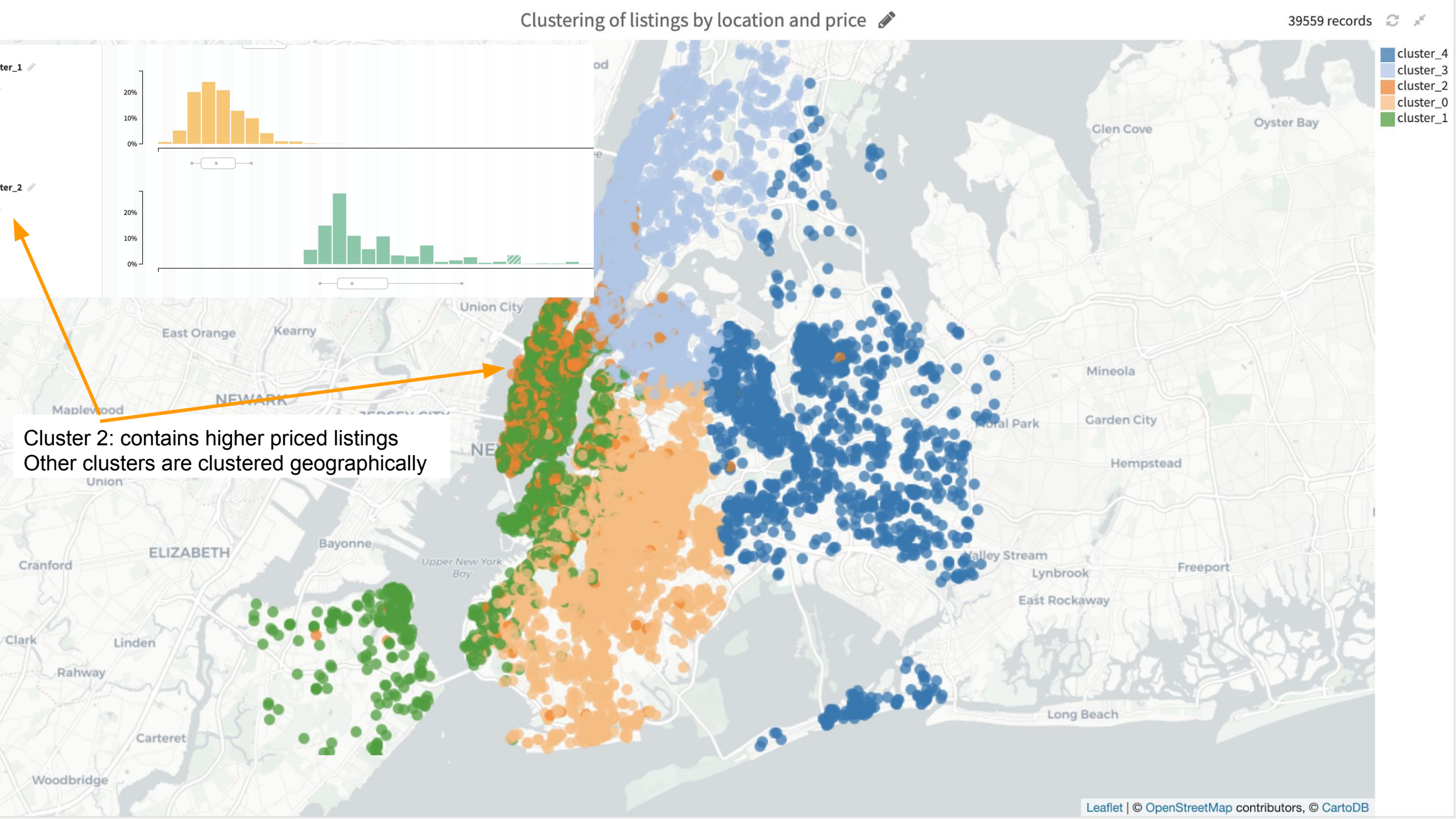

Household Income NYC
39559 records
194k
180k
160k
140k
120k
100k
80k
60k
40k
20k
Englewood
Hackensack
Little Falls
Clifton
Passaic
Cedar Grove
West Caldwell
Fort Lee
Ridgefield
Cliffside Park
Fairview
Nutley
Bloomfield
Livingston
West Orange
Belleville
Secaucus
Union City
East Orange
Kearny
NEWARK
JERSEY CITY
Maplewood
Summit
Union
ELIZABETH
Bayonne
Upper New York Bay
Cranford
Clark
Linden
Rahway
Carteret
Woodbridge
Glen Cove
Mineola
Floral Park
Garden City
Hempstead
Valley Stream
Lynbrook
Freeport
East Rockaway
Long Beach
Leaflet | © OpenStreetMap contributors, © CartoDB

Age distribution NYC

39559 records



We clustered on location and listing price (per person)


Clustering of listings by location and price
39559 records
cluster_4
cluster_3
cluster_2
cluster_0
cluster_1
20%
10%
0%
Cluster 2: contains higher priced listings
Other clusters are clustered geographically
Union City
East Orange
Kearny
NEWARK
Maplewood
Union
ELIZABETH
Bayonne
Cranford
Upper New York Bay
Clark
Linden
Rahway
Carteret
Woodbridge
Glen Cove
Oyster Bay
Mineola
Garden City
Hempstead
Valley Stream
Lynbrook
Freeport
East Rockaway
Long Beach
Leaflet | © OpenStreetMap contributors, © CartoDB

Next Steps

- Airbnb rental investment recommender system
- Automatic marketing

Thank You