Uncovering the Secrets of Airbnb:

What impacts the Airbnb's Spatial Distribution in U.S. cities?

——Data Mining Project by Group BeautifulSoap4

Content

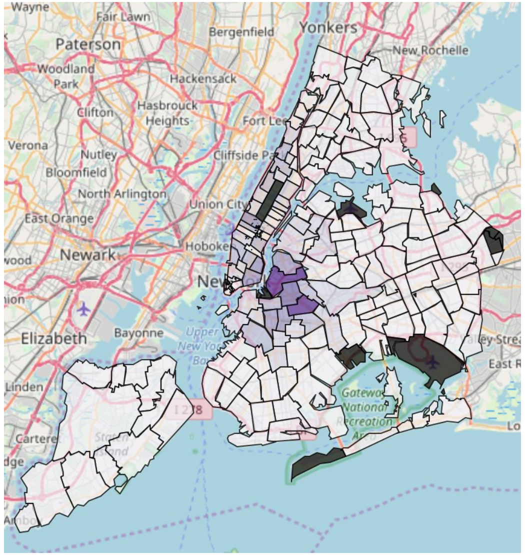
1. Initial Exploration
2. Feature Engineering
3. Modeling Products
4. Summary and Suggestion

Part 1: Initial Exploration

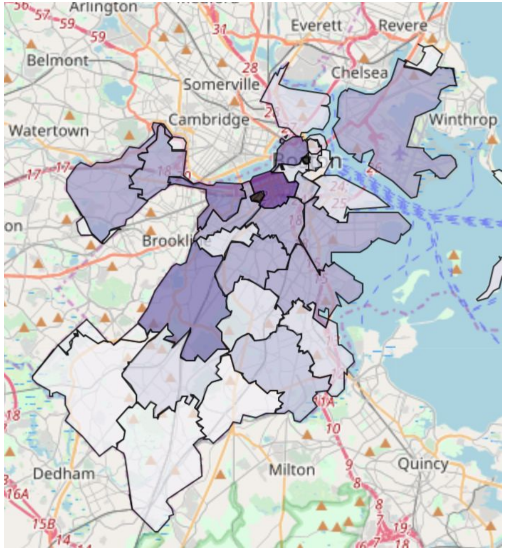
How are airbnb distributed in NY, Boston and DC?

- A first look at the airbnb spatial distribution
- Let’s take the three major U.S. cities on the East Coast for example

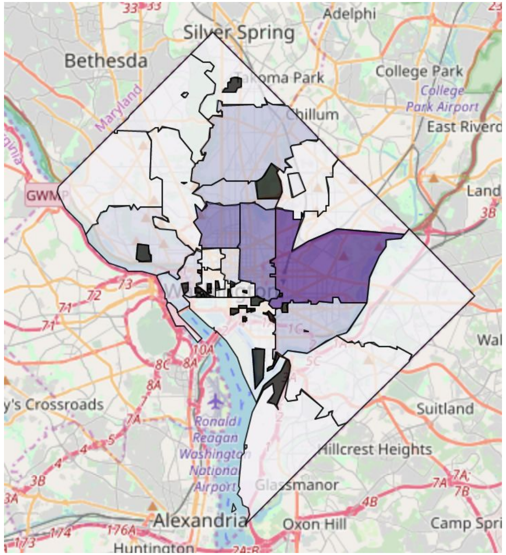
New York



Boston



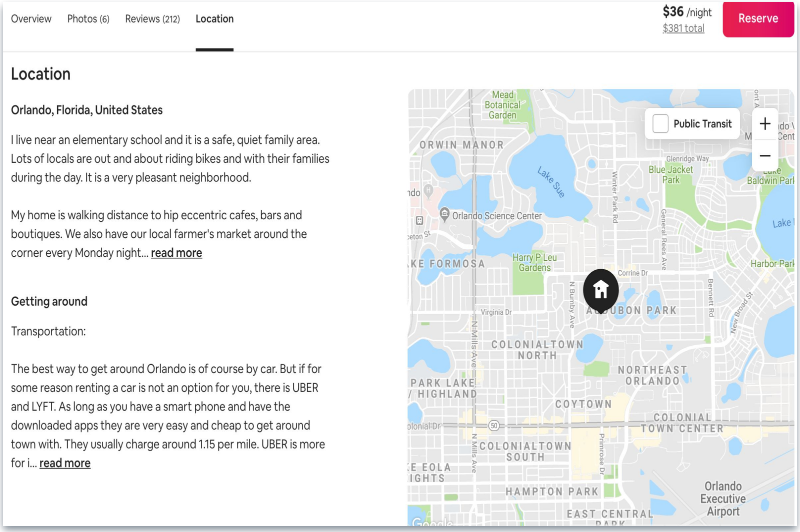
Washington D.C.



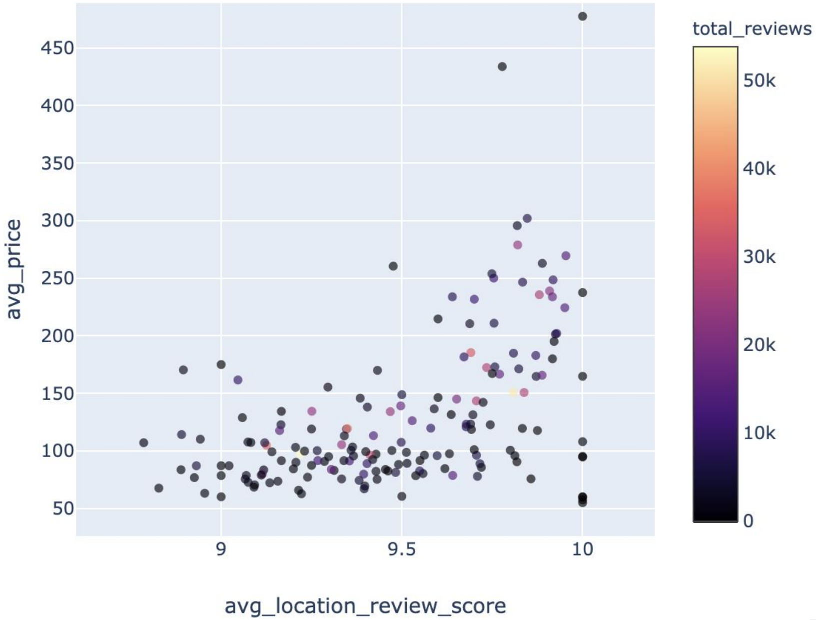
Part 1: Initial Exploration

For Airbnb, the location is crucial!

- Airbnb hosts usually emphasize a lot about their location features of their house on their first pages



- Price and popularity of an Airbnb house are positively correlated to the average location review score



Part 1: Initial Exploration

What impacts airbnb distribution?

- What do people care about location that might potentially impact airbnb distribution?
- What keywords are often mentioned?

New York

Boston

Washington D.C.

Part 1: Initial Exploration

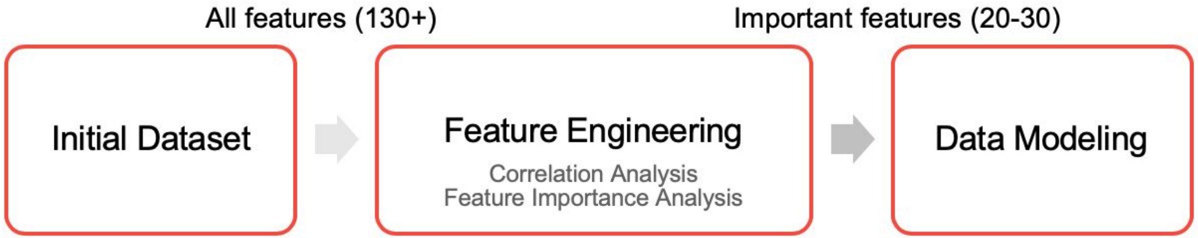
Dataset Building

- Airbnb listing data, as well as geographic, economic, social, real-estate data are included in our dataset for analysis of factors that impact airbnb distribution
- Note: All the data are processed into zipcode-level data for zipcode level analysis

No.	Data	Description	Range	Source
1	Airbnb listing data	- locations, room types, prices, ratings, reviews	as of 2019	Inside Airbnb
2	transportation	- locations of subway and bus stations	as of 2019	City Opendata
3	c_distance	- the great-circle distance between centroid of zip code and the city hall of each city	as of 2016	US zipcode package
4	venues	- tourist attractions, restaurant, market	as of 2019	Google Places
5	populatition	- total population, population by age/race	as of 2019	US zipcode package
6	household income	- average household income	as of 2019	US zipcode package
7	education	- number of people of each education level	as of 2019	US zipcode package
8	employment	- number of full-time, part-time, unemployed people	as of 2019	US zipcode package
9	crime	- crime data of NYC, BOS and DC	2018.8 - 2019.1	CIty Opendata
10	real estate	- building-year, rental & home values, room amount, house vacancy and occupancy	as of 2019	US zipcode package, Zillow Housing Data

What impacts airbnb distribution?

- Next, we conduct the feature engineering, to select the important features as preparation for modeling in part 3
  - **Correlation Analysis** (Scatter Plot & Correlation Heatmap)
  - **Feature Importance Analysis** (Random Forest)

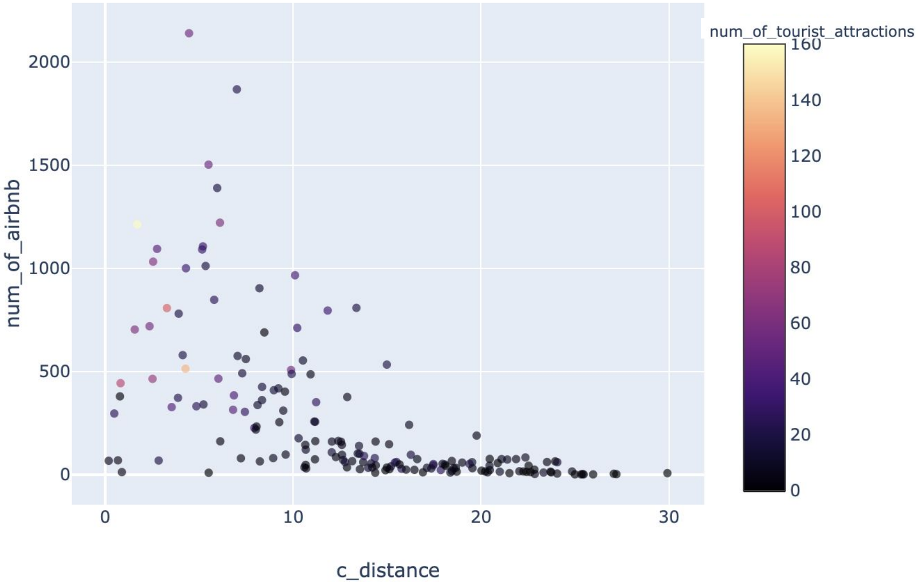


Part 2: Feature Engineering

A first look at the potential factors

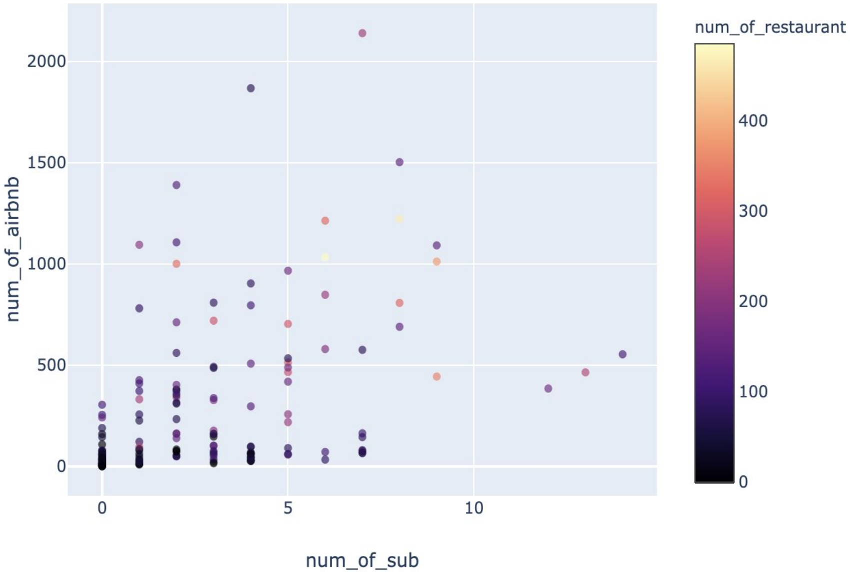
Airbnb distribution is highly tourism-related.

Airbnb tends to have a positive correlation with the number of **tourist attractions** in the neighborhood, and a negative correlation with its **distance from city center**.



The more convenient the neighborhood is, the more airbnbs there are.

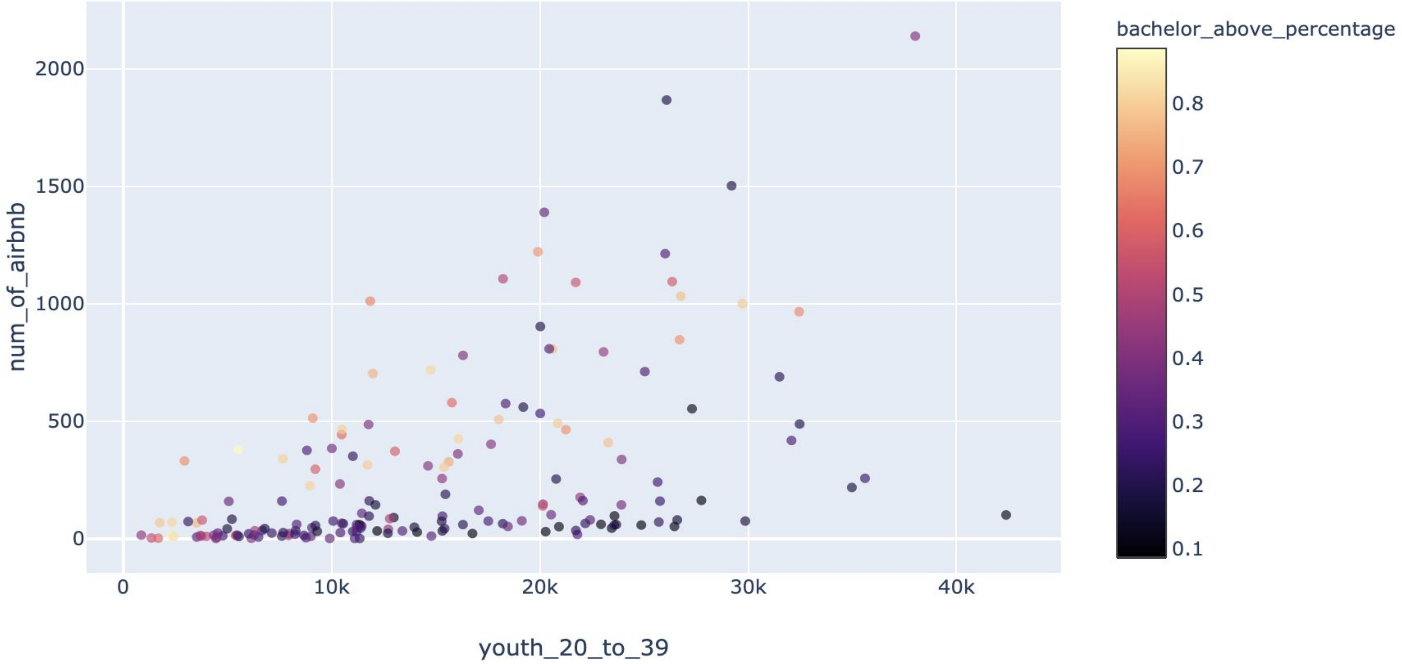
Airbnb distribution is positively impacted by **transportation** and **venues** (restaurant) in the neighborhood.



A first look at the potential factors

Airbnb distribution also has something to do with demographics of the area.

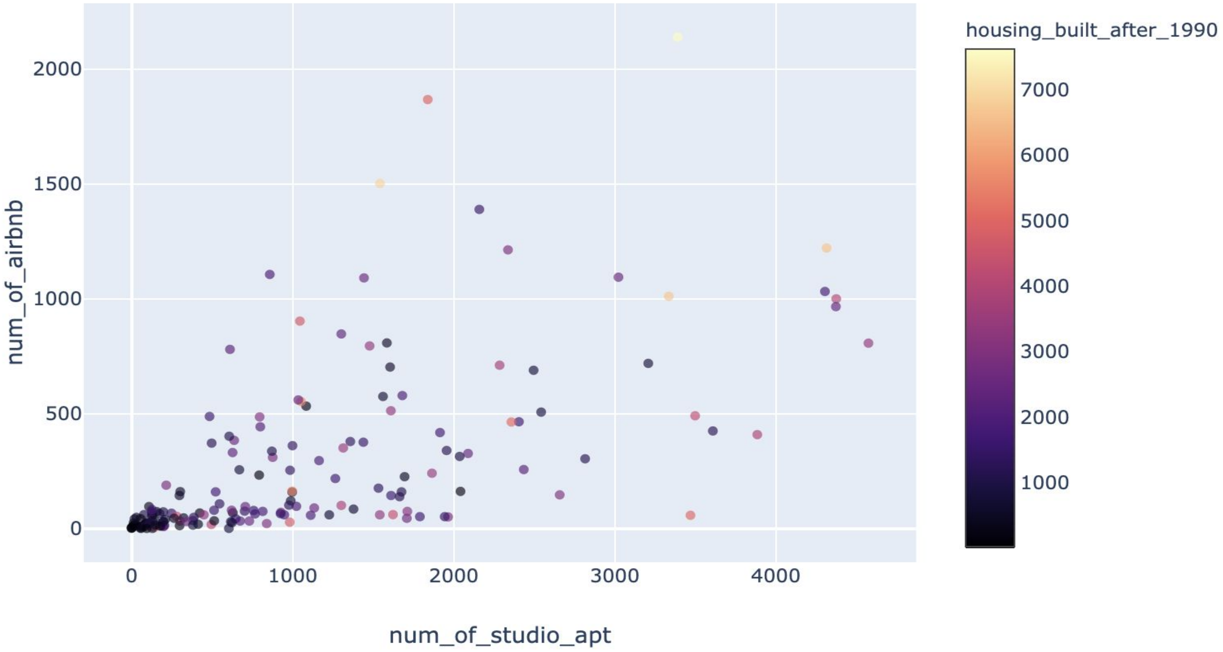
- Regions with more **young and well-educated people** ("creative class") are more likely to have a denser airbnb distribution



A first look at the potential factors

Airbnb distribution seem to have positive correlation with the amount of studios and new housing.

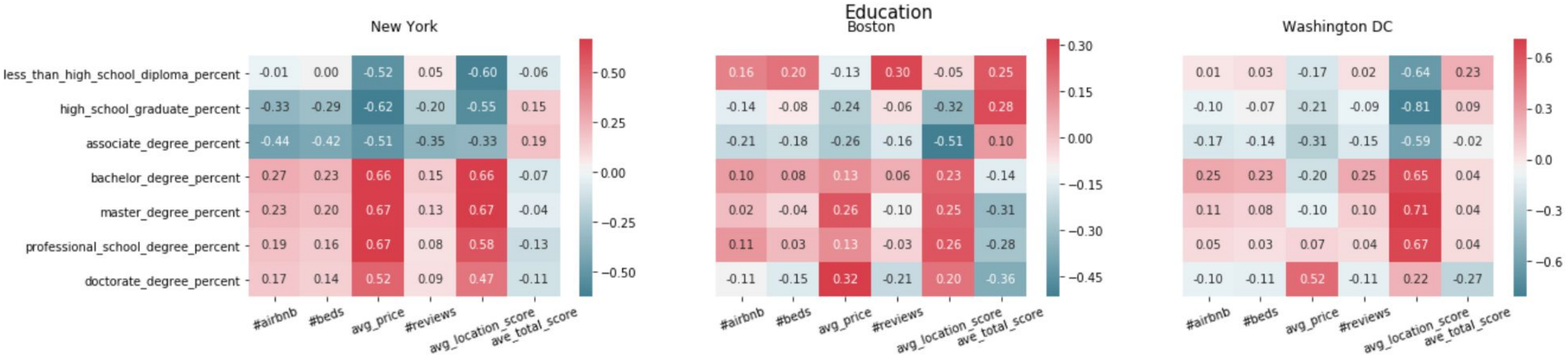
- In terms of the real estate aspect, airbnbs are more likely to locate in regions with more **studios** and **new housing** (built after 1990)



Part 2: Feature Engineering

What impacts airbnb distribution?

Airbnb distribution v.s. Education



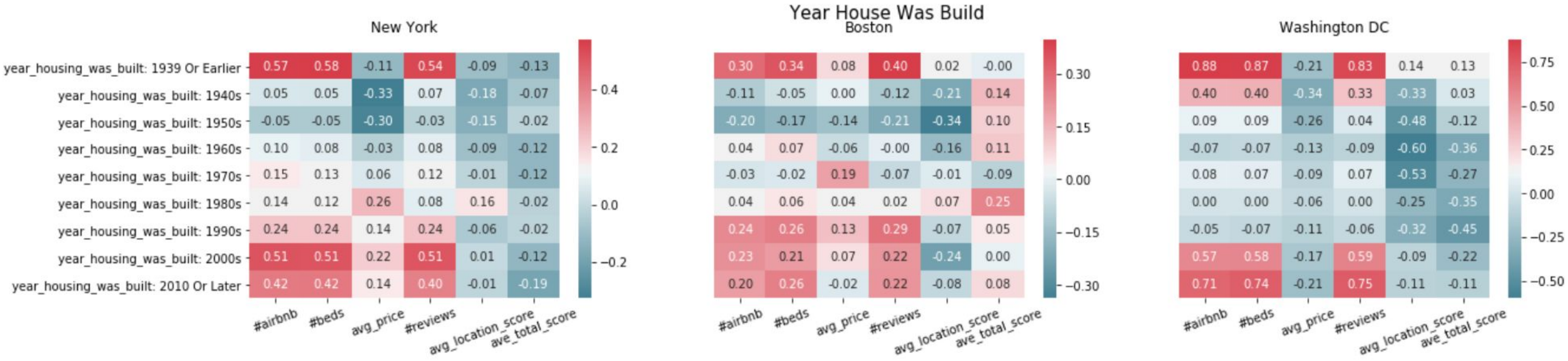
Education rate impacts airbnb distribution:

- Generally, the percentage of people with higher education is highly correlated with number of airbnb
- Select the feature - the percentage of people with bachelor degree and above - for data modelling

Part 2: Feature Engineering

What impacts airbnb distribution?

Airbnb distribution v.s. Year Housing was Built

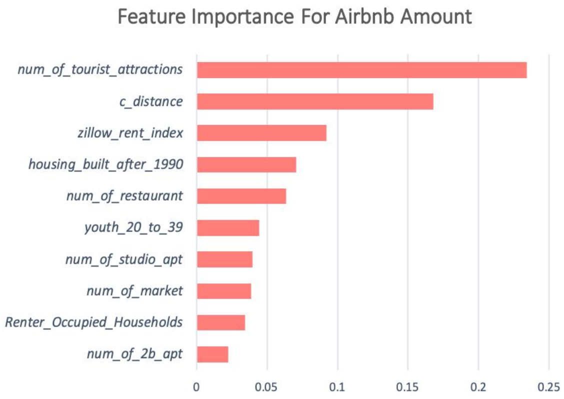


Number of newly-built housing impacts number of airbnbs:

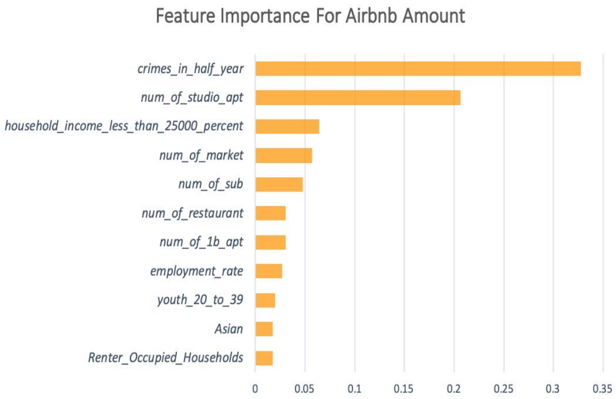
- Generally people tend to choose newly built housing for airbnb
- Select the feature - num of housing built after 1990s - for data modeling

Feature Importance with Random Forest

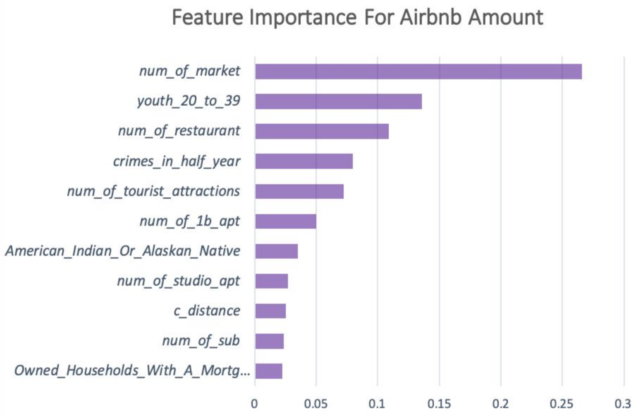
Top 10 Features that Impact Airbnb Distribution



New York

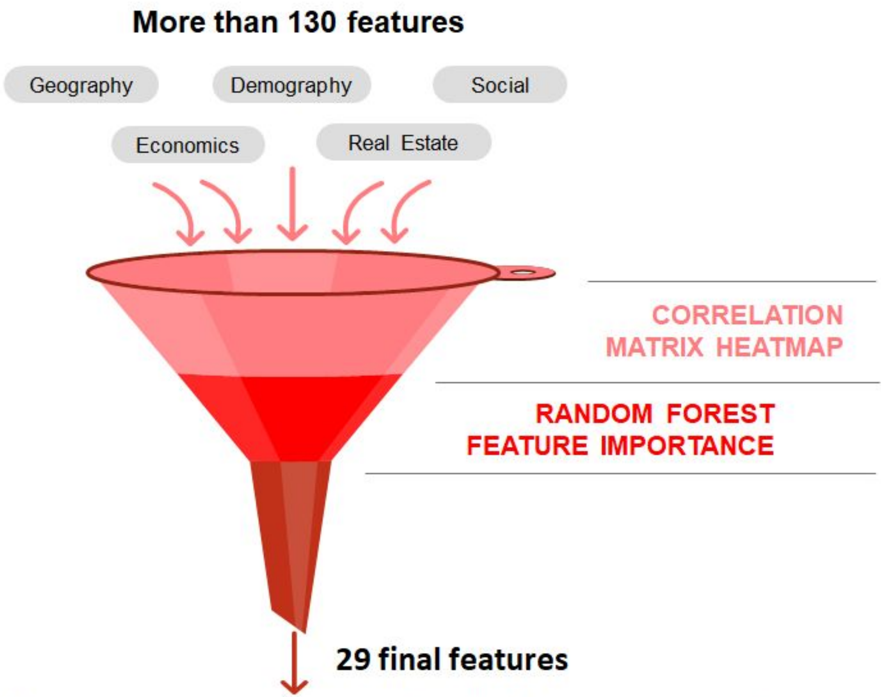


Boston



Washington D.C.

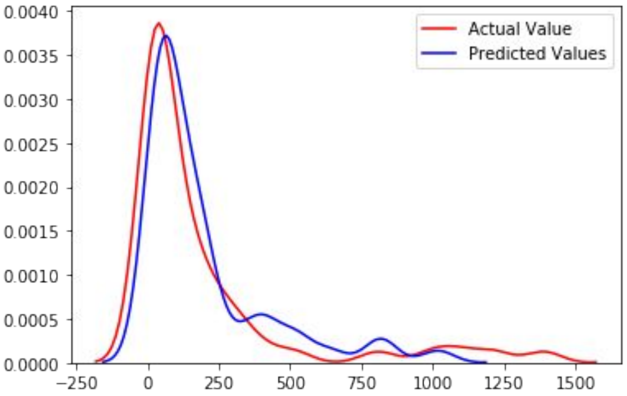
What impacts airbnb distribution?



bus, subway, C-distance, restaurants, markets, tourist attractions travel agency, population, children, youth, race, household income, education, employment, crimes, new housing, home values, room amount, house vacancy and occupancy, rental

Part 3: Modeling Products

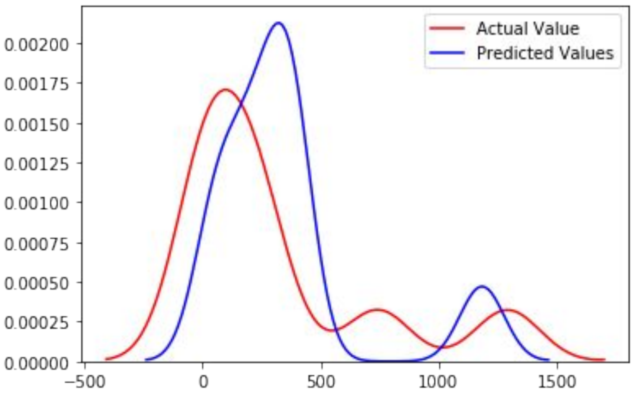
Model 1: Density Prediction with Random Forest



New York

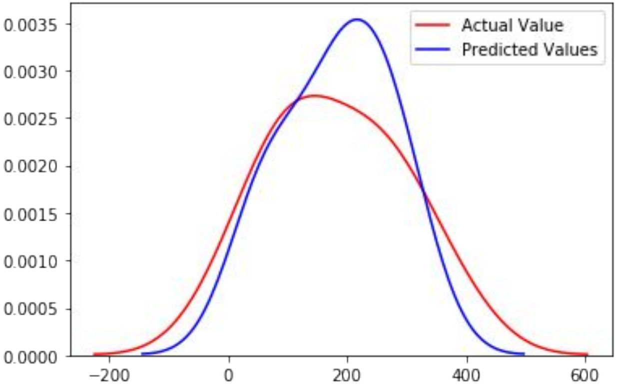
(larger dataset -> higher accuracy)

Model score: 0.7550
Mean Absolute Error: 27.36 degrees.
Accuracy: 90.22 %



Washington D.C.

Model Score: 0.75500
Mean Absolute Error: 41.64 degrees.
Accuracy: 25.89 %



Boston

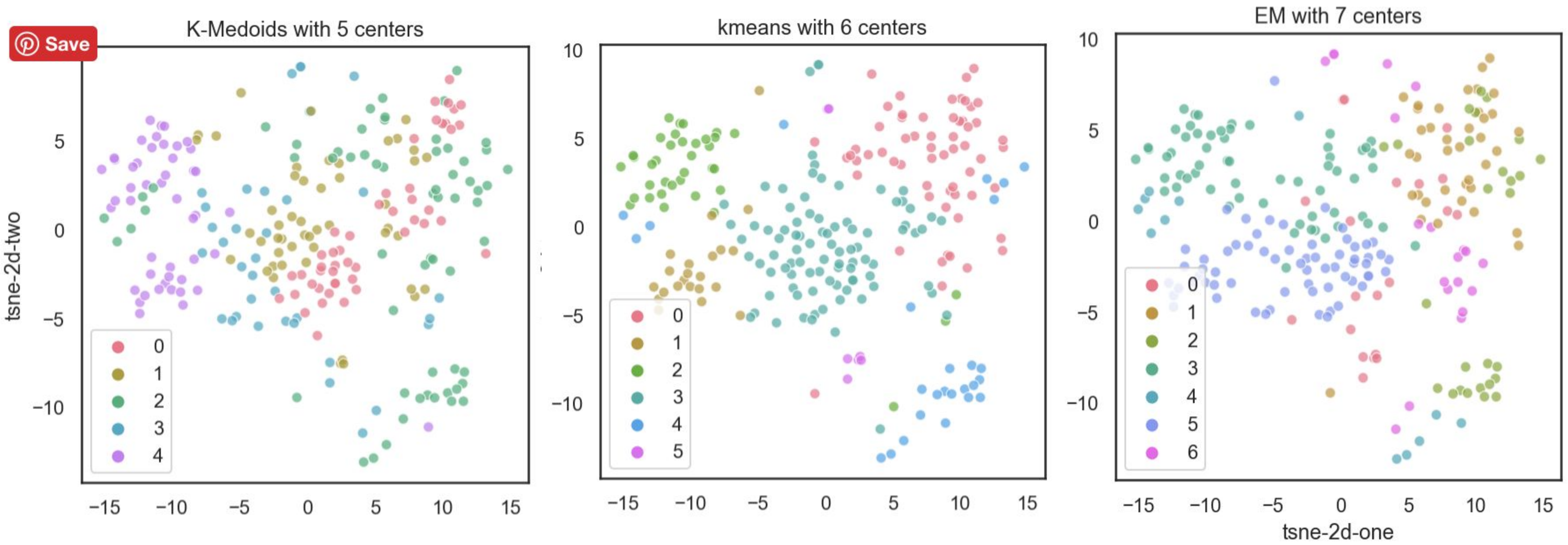
Model Score: 0.6162
Mean Absolute Error: 12.71 degrees.
Accuracy: 78.75 %

- RandomForestRegressor（relatively accurate prediction）
- These features are important factors that influence the distribution of airbnb

Part 3: Modeling Products

Model 2 : Clustering for Management

- With 34 features selected, further use PCA to compress data
- Try K-Means, K-Medoids, Mean-Shift, DBSCAN, EM, etc. clustering methods

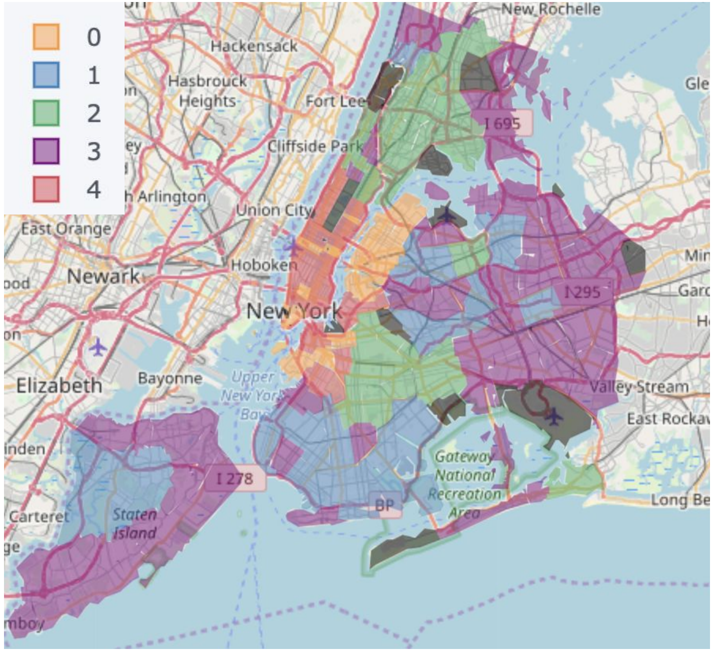


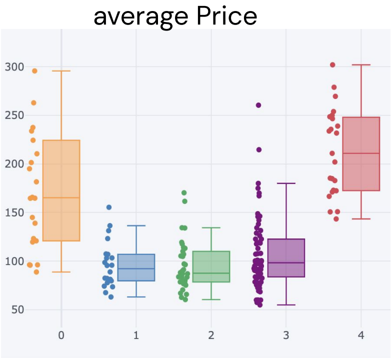
K-Means with 6 centers has the best performance

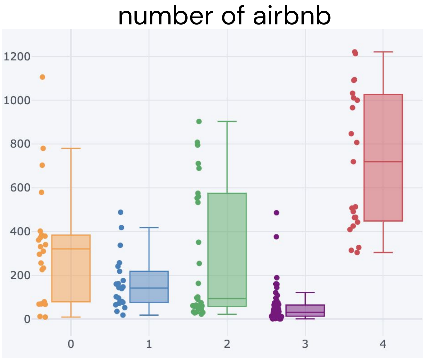
Part 3: Clustering Analysis

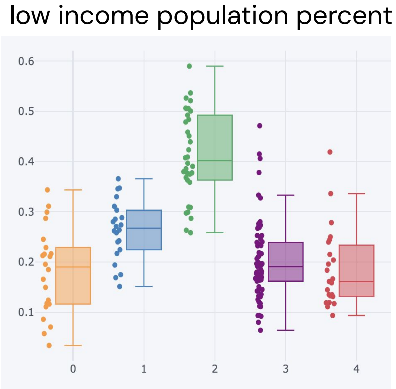
Clustering result with K-Means

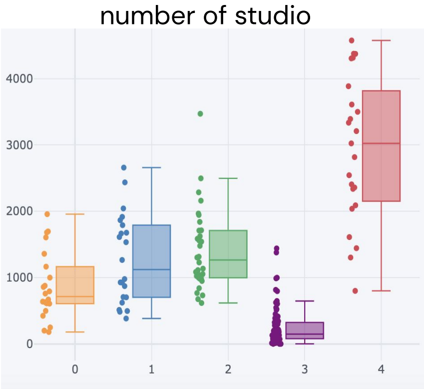
- Our model successfully distinguishes the characteristics of airbnb, population, geography, estates, etc.







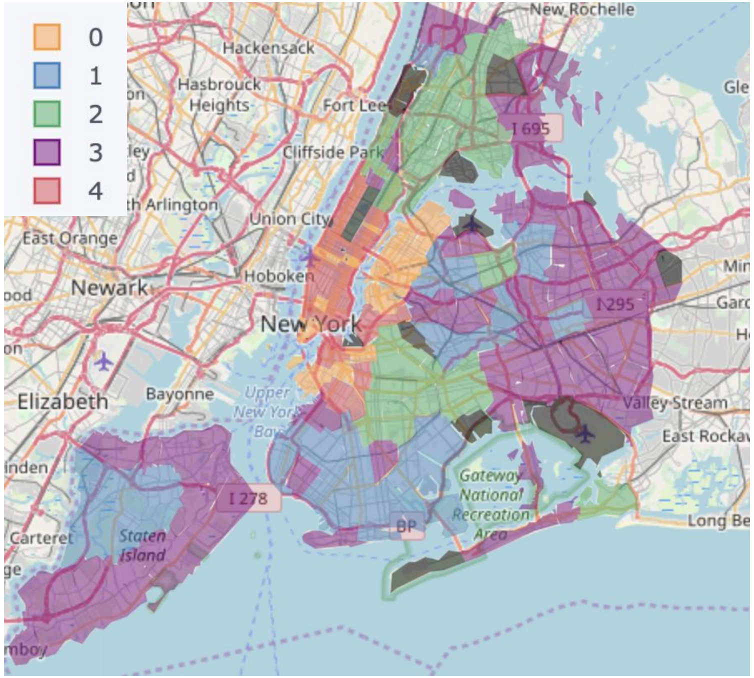




Part 3: Clustering Analysis

Clustering result with K-Means

- Key labels are assigned based on different clusters’ characters.



Popular Area
- Many restaurants
- Relatively high quality
- Asian
- Less youth
- High employ rate
- old buildings

Residential Area
- less airbnb supply
- a lot white
- restaurants
- low rent value

Inexpensive Area
- relative few attractions
- lower quality airbnb
- Black American
- low income population
- low education level
- a lot of old buildings

Remote Area
- few attractions
- few airbnb
- few estates
- few studio
- less youth

Luxury Area
- lots of tourism attractions
- high quality
- large supply of airbnb
- higher education level
- A lot studio

Key Findings and Suggestions

Key findings

- **Convenience** - the airbnb distribution is highly related to convenience and tourism
- **Demographics** - the more “creative class” (educated, creative young workers) in the area, the denser the area is
- **Real Estate** - airbnbs are more likely to locate in regions with studios and new housing

Suggestions for Airbnb Management Team:

- When Airbnb decides to expand into a new city, it can utilize a similar model to predict potential numbers of airbnb in different areas for better budgeting and advertising strategies
- The clustering model can be used to refine the management of airbnb in different types of area. With more cities’s data, this model may work better and better for other U.S cities

References

1. Airbnb distribution in cities may depend on who lives there, not just distance to city centre
   https://www.biomedcentral.com/about/press-centre/science-press-releases/19-09-18
2. US zipcode documentation
   https://pypi.org/project/uszipcode/
3. Airbnb offers World Class service – Thanks to Big data and Machine Learning
   https://www.simplilearn.com/airbnb-uses-big-data-machine-learning-article
4. Use Data Science to find your next Airbnb getaway
   https://towardsdatascience.com/use-data-science-to-find-your-next-airbnb-getaway-3cb9c8333ad1
5. Analyzing and predicting the spatial penetration of Airbnb in U.S. cities
   https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0156-6

Thanks!

Group Member
zd2242@columbia.edu
mj2940@columbia.edu
yz3684@columbia.edu
lm3504@columbia.edu