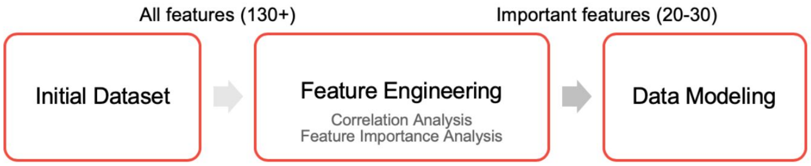
Airbnb Distribution Project

Project Goal

When Airbnb decides to expand to a new city, it is crucial to gain insights into the Airbnb Distribution Pattern to perfect budgeting and advertising strategies in that city. Therefore, our team aims to analyze relevant factors that impact airbnb's spatial distribution and build prediction and classification models. We picked three representative cities on the East Coast, New York, Boston and Washington D.C., for our investigation.

Project Summary

The overall logic of this project is to first construct an initial dataset, then conduct the feature engineering to select a subset of key features as input of models, finally build up prediction and classification models with the best performances by applying various machine learning algorithms.



All features (130+)

Important features (20-30)

Initial Dataset

Feature Engineering

Correlation Analysis

Feature Importance Analysis

Data Modeling

Content	Explanation
Initial Exploration	- Processed the airbnb listing data into "number of airbnb's in each zip-code area" - Plotted the processed data on choropleth map with folium to see the overall distribution - Location WordCloud Analysis to take a first look at potential factors Result: decided to include geographic, demographic, economic and social data that might impact airbnbs' distribution in our dataset, consisting of more than 130 features (described on Page 6 in slides)
Feature Engineering	**Correlation Analysis** - Analyzed correlation between each feature and airbnb's amount. - With scatterplots and correlation heatmap, we found out relevant (positive/negative) factors that impact airbnb distribution, such as young people, new housing, etc - This guides our feature selection - for example - we should include features like the number of young people and new housing, instead of population of all age groups and housing built a century ago. **Feature Importance Analysis** - After a couple of trials of modeling, we found that including too many features would yield models with bad performances. So we performed the feature importance analysis using random forest with 10 times of shuffle split and 500 estimators, to rank the feature importances for feature selection Result: filtered the original 130+ features to 29 features for modelling next step

Modelling Products	With the 29 input features, we built prediction and classification models with supervised learning (regression) and unsupervised learning (clustering) **Density Prediction with Regression** - First tried to train Ordinary Least Squares model, expecting to yield the coefficients of each features on airbnb amount to compare how different features' impacts vary across cities. Yet the overfitting problem is pretty serious, with R-squared being 1 for Washington D.C. - We reflected that it may be caused by inputting too many features into the model (actually outweigh the sample size in DC), and that random forest may be more useful to dealing with a relatively small dataset with a large number of features. Then we used RandomForestRegressor. With the use of GridSearchCV, we evaluated the best parameter combinations for the final model, which gave us better results with higher accurate rate and explanation power. **Clustering for Management** - Used principal components analysis (PCA) to compress 29 features into 10 components. - Then, tried k-means, DBSCAN and other 6 clustering methods to train our model with three cities' data. Through the visualization result of tsne, we found out that k-means clustering with 6 centers has the optimal clustering performance. It successfully differentiated different types of areas with distinctive characters. - We then assigned these clusters with unique labels, 'popular', 'residential', 'luxury' , etc. by analysing their characters. These labels can provide insight for the management of airbnb in diverse areas. For example, the areas of 'residential' type have few airbnbs and a lower tendency to rent despite a large number of restaurants and houses in this area. So when expanding into a new city, Airbnb may as well put less efforts for area in this type

What We Believed We Did Great At

1. We collected and processed a large number of external datasets, such as transportation, demographics and real estate data, which are highly relevant for this topic and significantly contributes to final results.
2. We spent great efforts in analyzing all features and filtering them into a subset of more crucial features as input of modelling, which proves to be effective, since omitting less relevant features greatly improves the model performances.
3. The clustering model successfully identified and differentiated different types of areas with distinctive characters. Besides, we also help make sense of the results by summarizing them with different labels, such popular area, luxury area, residential area, etc, which gives direct implications for audiences.
4. We went beyond from technical analysis to discuss its business insights and real-life application for airbnb management team, which gives the modelling practical meaning.

What We Would Like to Further Explore

It takes a lot of work and time to build up a large dataset and transform them into the input data for models, which limit our scope of analysis (three cities). For future exploration, we would like to include data of more cities and try to differentiate the airbnb distribution patterns in different cities. With the training of more cities' data, these models should work better and better for other regions.