
Boston Airbnb Fair Pricing Tool and Recommender

— Dan Rossetti, Data Scientist and
Founder of Cheap Stays, LLC —

Agenda

- Introduction / Problem Statement / Data Sources / Architecture
- Exploratory Data Analysis
- Feature Engineering
- Feature Selection
- Modeling
- Recommender
- Conclusions / Next Steps

Intro | Problem Statement | Data Sources | Architecture

Introduction

Airbnb General:

- Hosts allow guests to stay at their property for a fee
- Alternative to traditional hotel / hostel
- Passive (or main) income for hosts

Airbnb Listing Prices:

- Pricing very subjective
- Vary wildly
- Finding best deals - time consuming
- Setting listing price - time consuming

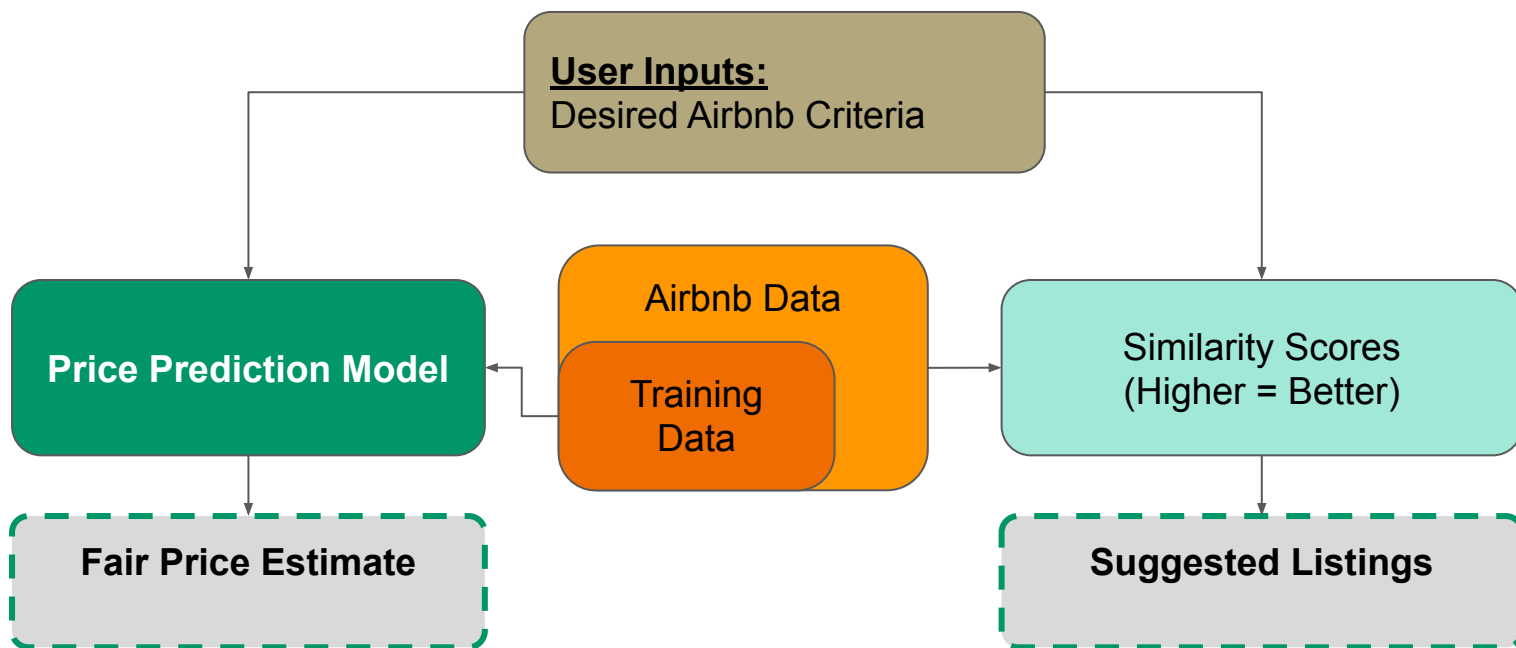
Problem Statement at a Glance

- Be the ***Kelly Blue Book*** of Airbnb's
- Estimate fair prices for
 - Guests
 - Hosts
 - Developers
- Goal: RMSE of \$20
- Recommend criteria-based listings
 - Criteria in \Rightarrow Fair Price / Recommended Listings out
- Fast, easy, quick

Data Sources

- Airbnb Data - *Inside Airbnb* [1]
 - Includes: 75 numerical and categorical variables
- Geographic Data: Listing locations to Subway Stations
 - *MBTA Stations* [2] - 132 total Stations
 - Google Geocoding API [3]
 - T-stop addresses \Rightarrow latitude / longitude

Product Architecture



Exploratory Data Analysis (EDA)

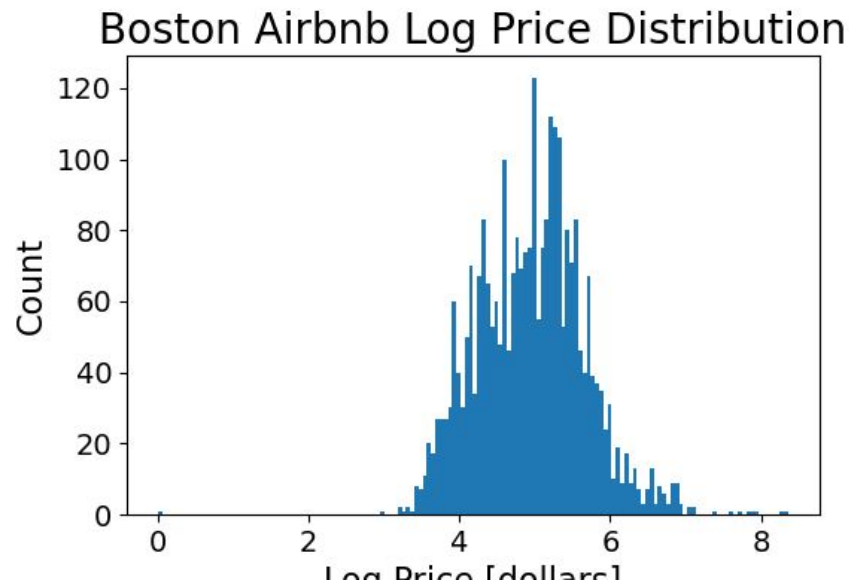
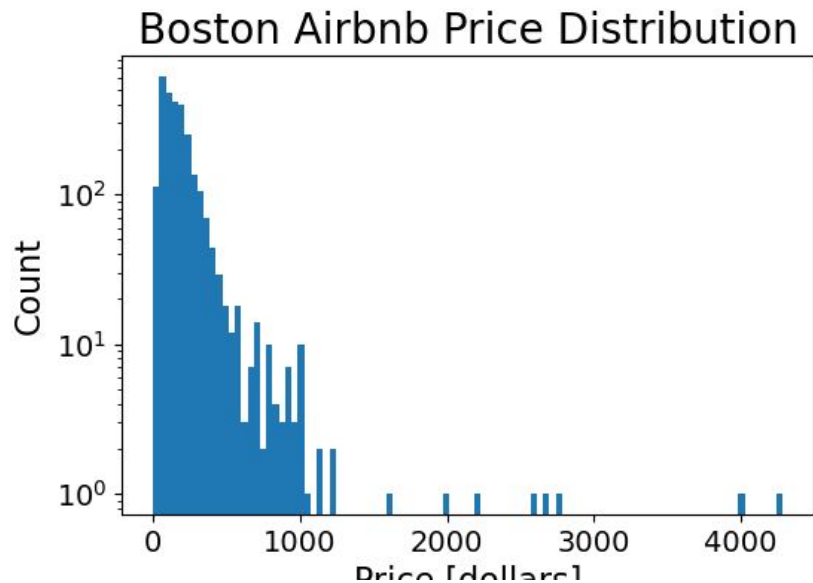
NOTE: All Subsequent EDA Performed on Training Data

Exploratory Data Analysis – Basic Stats

	<u>Mean</u>	<u>Min</u>	<u>25th Pcntl</u>	<u>50th Pcntl</u>	<u>75th Pcntl</u>	<u>Max</u>
Guests Accommodated	3.17	1	2	2	4	16
No. Bedrooms	1.75	1	1	1	2	13
No. Beds	1.79	1	1	1	2	22
Price	\$188.39	\$0	\$83	\$146	\$225	\$4283
Number of Reviews	41.09	0	0	7	44	821

Most Airbnb's are not extravagant

Exploratory Data Analysis – Price



Log transformations help normalize price data \Rightarrow better modeling

Feature Engineering

Feature Engineering - Descriptive Columns (Tokenizing)

Airbnb Data:

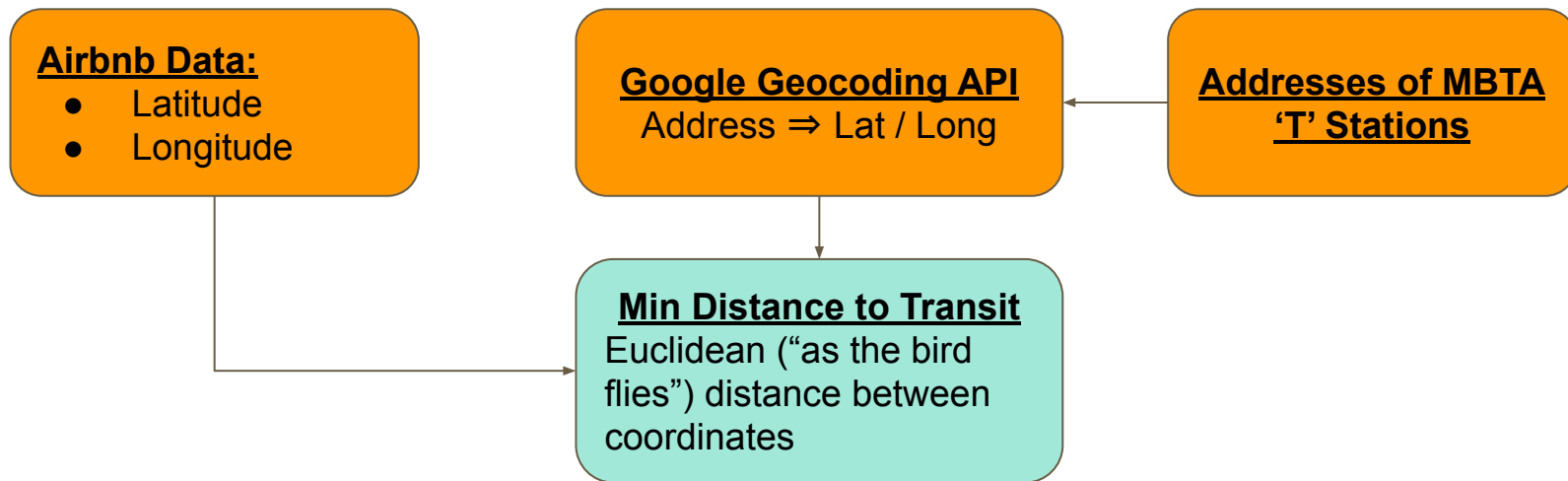
- Amenities
- Name
- Description
- Neighborhood Description
- Host About Information

For Each Category Find:

- Common Words
- Common phrases up to 4 word strings (n-grams)
- In at least 10% of listings

	Amn. word_1	Amn. word_2	...	Name word_1	...	Descr. word_1	...	Nbhd word_1	...	Hst. Abt. word_1	etc...
Listing 1	yes	no	yes	...	no	...	yes	...
Listing 2	yes	yes	no	...	no	...	yes	...
etc...

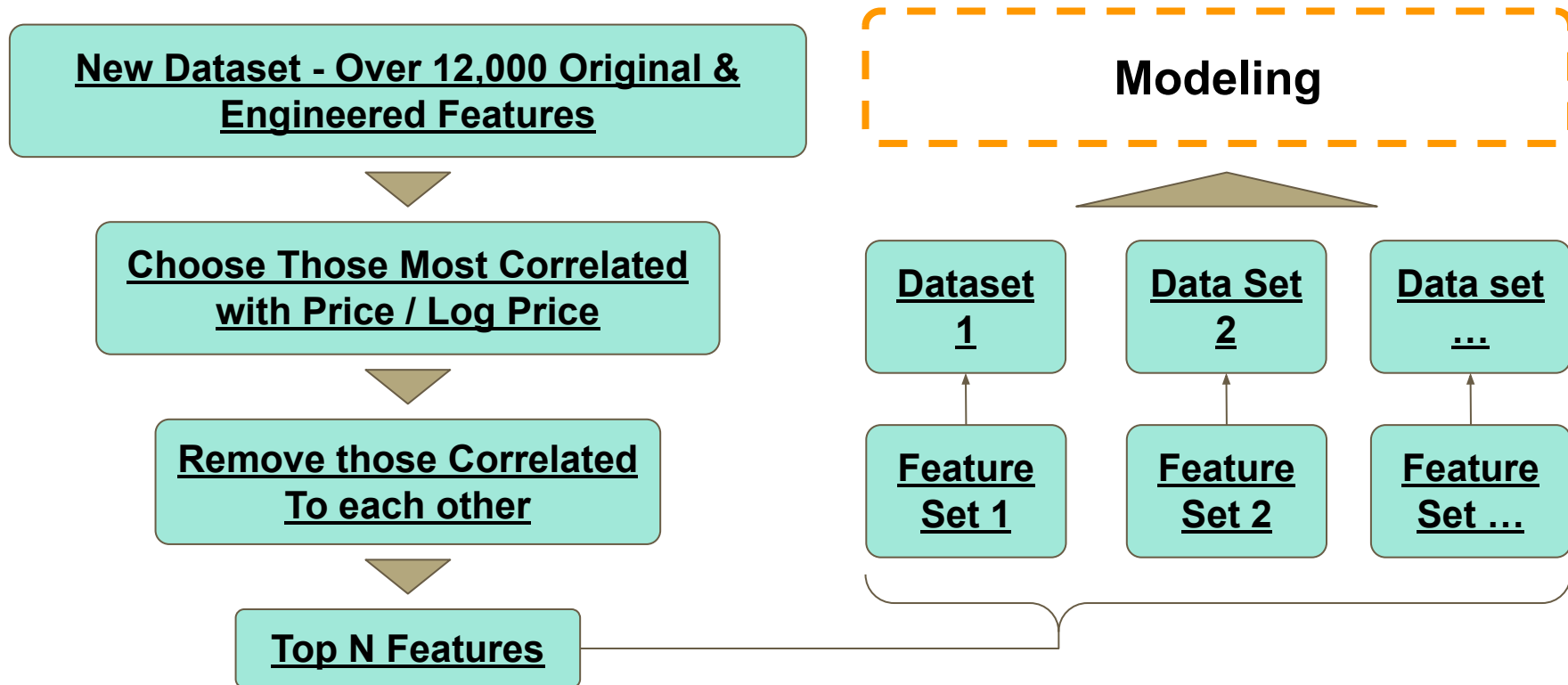
Feature Engineering – Proximity to Public Transit



Other features were engineered but won't be discussed

Feature Selection

Correlations - Quantitative Feature Selection Process



Modeling

Modeling

Primary Metric: Root Mean Squared Error

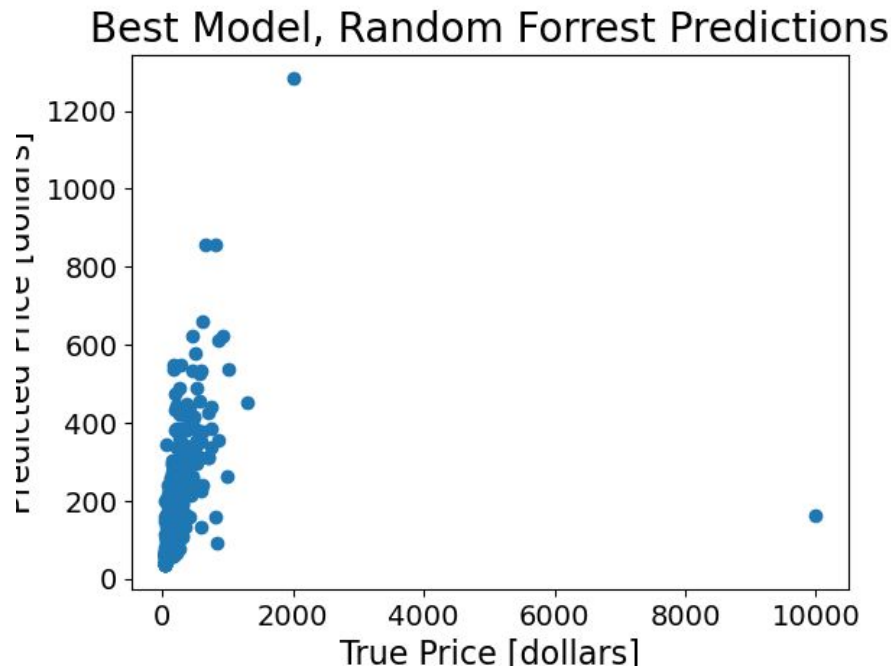
- Standard Deviation of Errors
 - ~67% of values within +/- RMSE
- R-squared - secondary metric

Modeling Steps Taken

- Seven different regression model varieties
- Ran models on different feature sets
- Optimized models
- Dimensionality reduction

Modeling Results

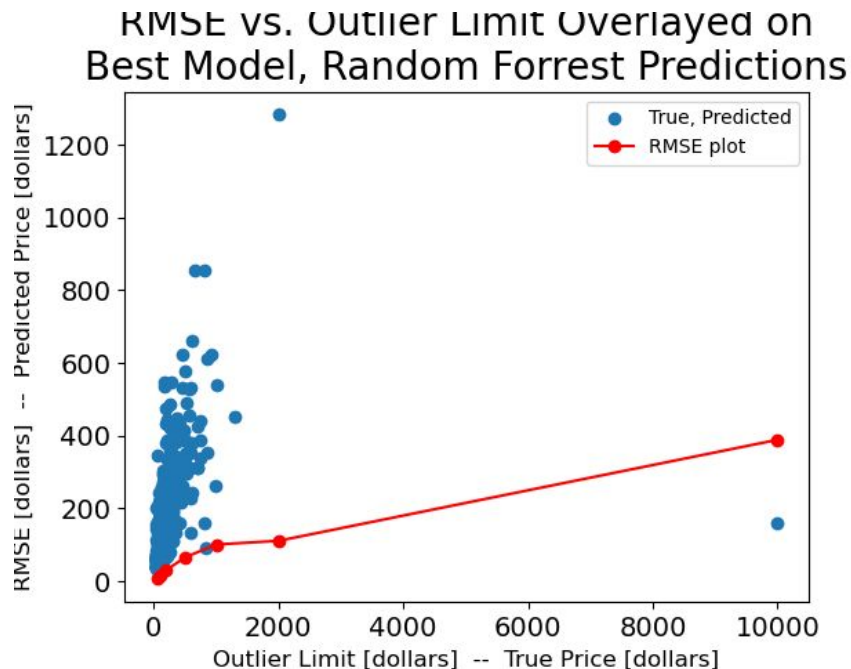
- All results: poor
- RMSEs High \$300s - High \$400s
- Best model:
 - Random Forest Regressor
 - Default Parameters
 - Feature set - 10 Features
- Metrics:
 - RMSE Training Data: \$138.88
 - **RMSE Validation Data: \$388.90**
 - R2 Training Data: 0.56
 - R2 Validation Data: 0.09



Outliers

- Data start to become very scattered after \$500
- Need better features need for higher-priced listings
- \$10,000 listing more than doubles RMSE

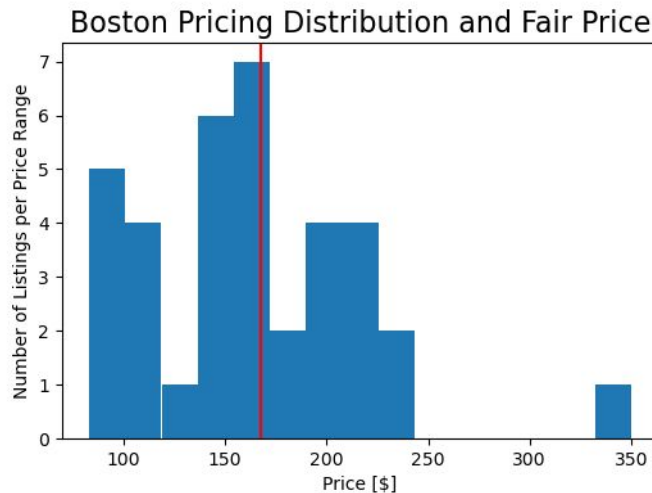
Including higher-priced listings in training data increases RMSE



Recommender

Recommender Tool Outputs

- Provides listing prices as distribution
- Fair price - red line
- Suggested listings: similarity \Rightarrow price



Suggested Listings			

Conclusions & Next Steps

Conclusions

- Poorly performing Fair price predictor created
 - RMSE way above \$20
 - Appropriate for recommender pricing distribution
- Outliers affected model
- Data somewhat unpredictable
- Additional features or data may improve performance
- Recommender performed as envisioned

Next Steps

- In-depth exploration of outliers
- Create additional features not tackled in this phase
- Attempt clustering \Rightarrow new features
- Refine dimensionality reduction methods
- Train a neural network
- Streamlit app

Sources

[1] - Inside Airbnb: <http://insideairbnb.com/get-the-data/>

[2] - MBTA Stations: <https://www.mbta.com/stops/subway>

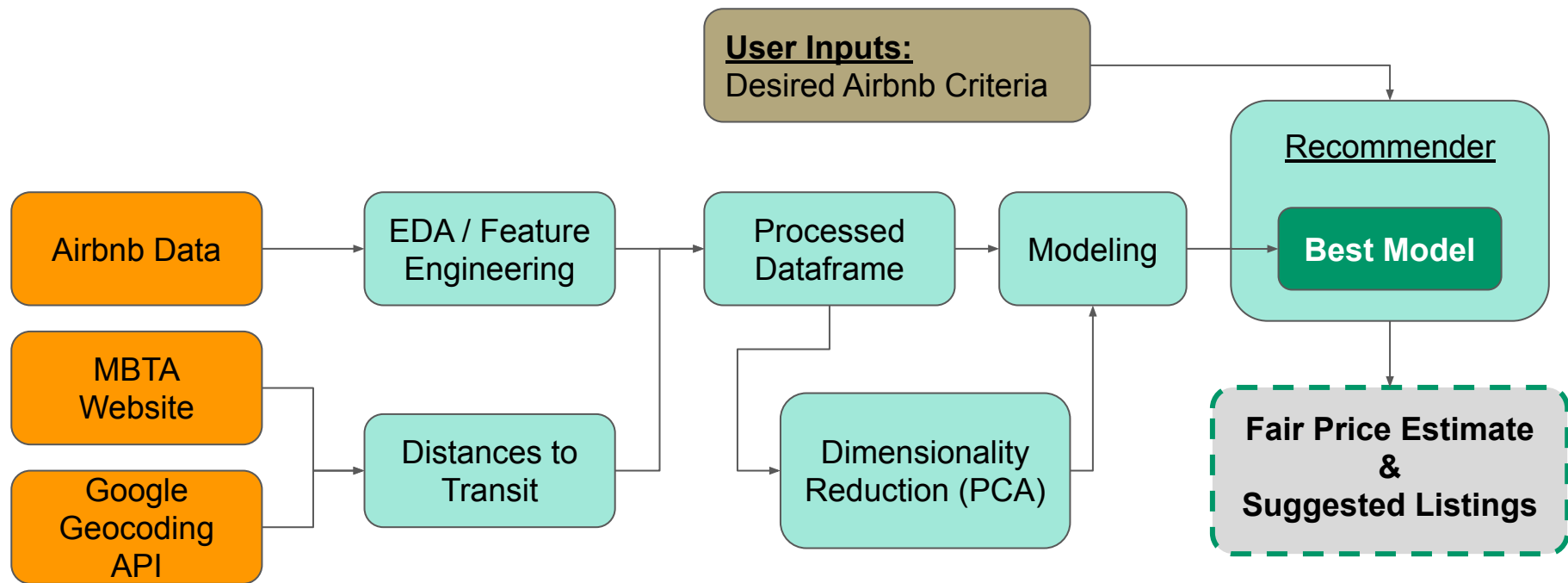
[3] - Google Geocoding API:
<https://developers.google.com/maps/documentation/geocoding>

Other:

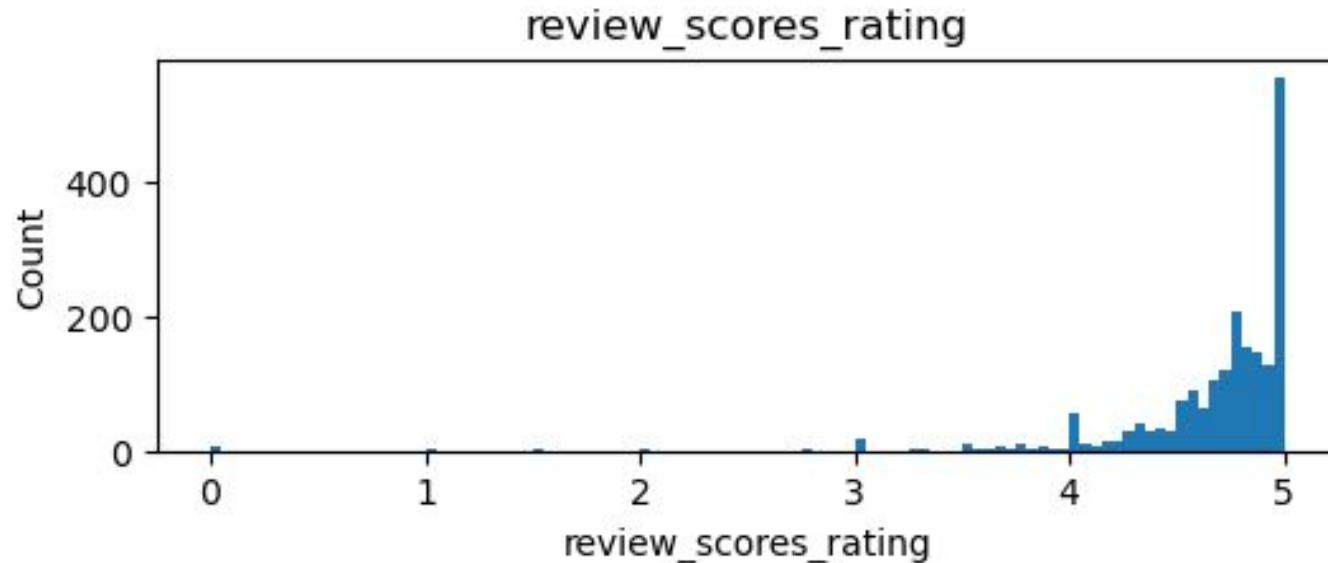
Inside Airbnb Data Dictionary:
<https://docs.google.com/spreadsheets/d/1iWCNjcSutYqpULSQHINyGlnUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596>

Additional Materials

Project Architecture



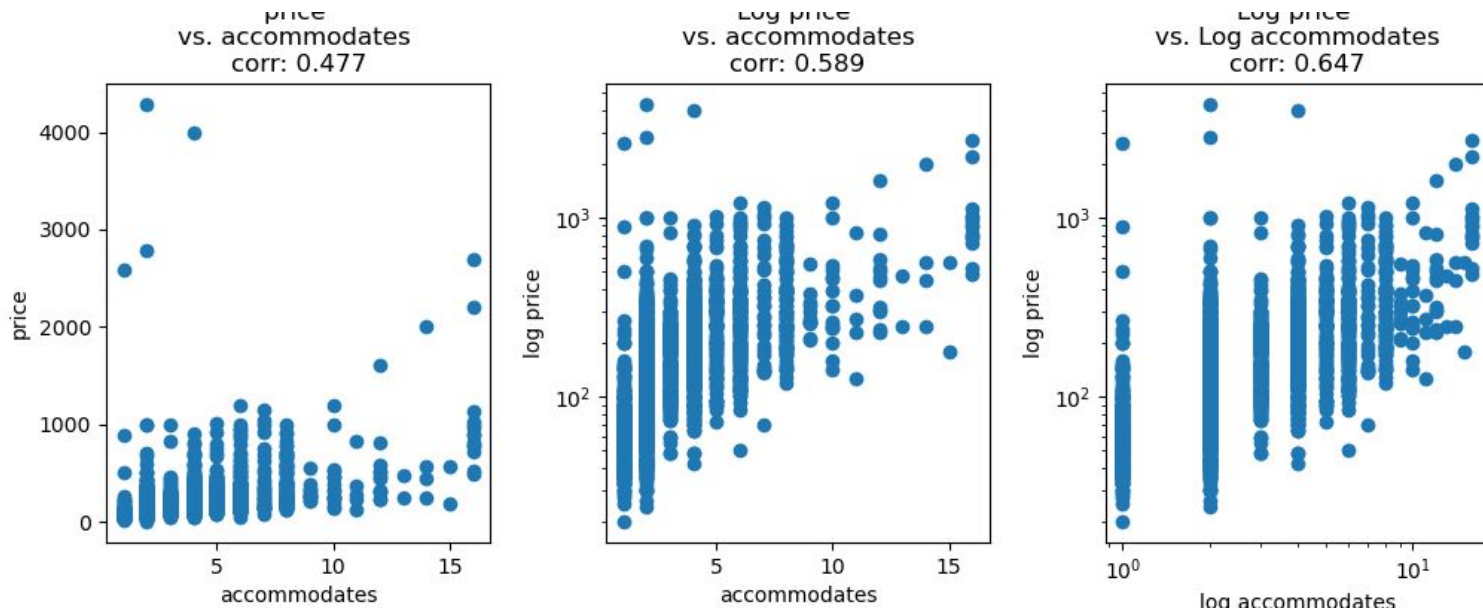
Exploratory Data Analysis – Review Scores



Exponential Distribution

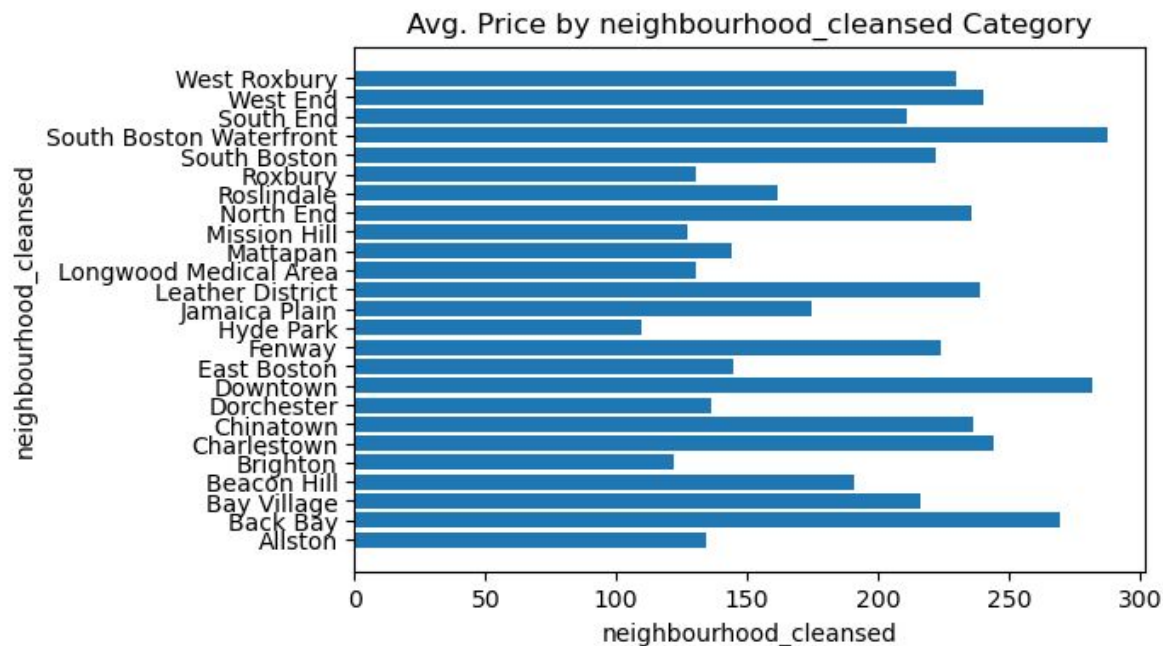
Very few listings score below 4.5 (average of all ratings)

Exploratory Data Analysis – Accommodates vs. Price



No. of People Accommodated Moderately Correlates to Price / Log Price

Exploratory Data Analysis – Neighborhoods

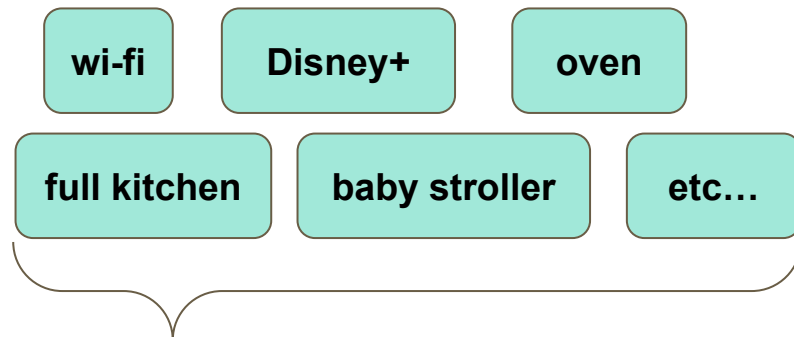


Average Prices Change by Boston Neighborhood

Feature Engineering – Amenities (Tokenizing)

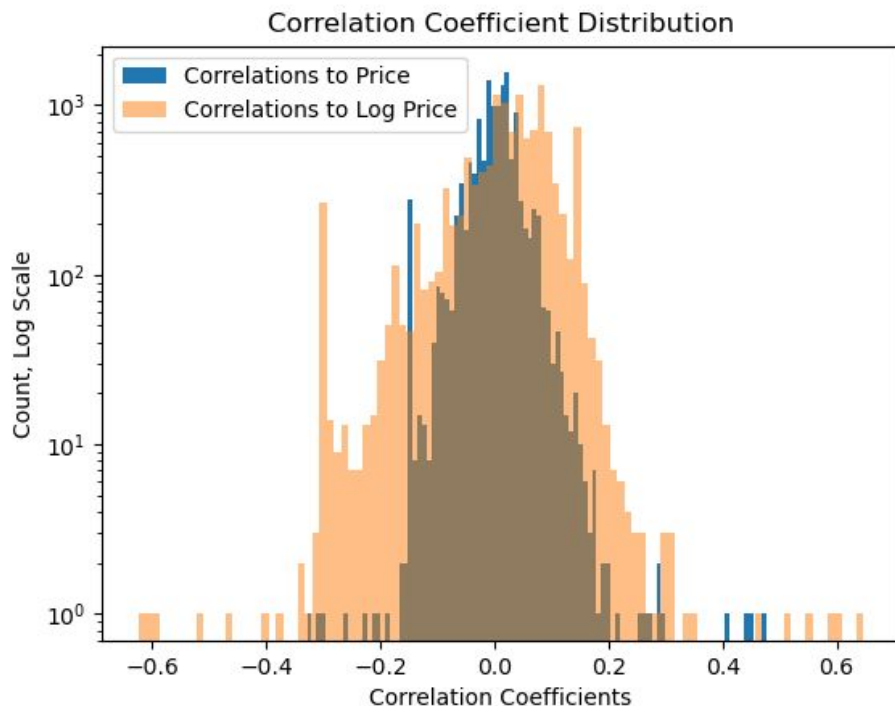
Airbnb Data - Amenities:

- ['wi-fi', 'oven', 'full kitchen' ...]
- ['wi-fi', 'Disney+', 'baby stroller' ...]
- ...



	wi-fi	Disney+	full kitchen	oven	baby stroller	etc...
Listing 1	yes	no	yes	yes	no	...
Listing 2	yes	yes	no	no	yes	...
etc...

Correlations – Very Few Highly Correlated Features



- Ideally: magnitude over 0.6 - 0.7
- Correlations to log price stronger

Poorly correlated features usually not helpful when modeling