Airbnb Pricing
Sankarhan Acharya, Paul de Fusco,
Parinaz Azdavari
Report 3

Key Findings through EDA

These key findings are key to the constriction of population segments:

Key finding: we have plotted many variables with respect to price. There is a wide range of values (mean price by neighborhood, property square footage, number of bathrooms, number of rooms, availability, etc.) but in general the market price is very elastic when it comes to smaller properties (defined as lower sq footage and lower bed/bathrooms) to the left of the curve, and it becomes very inelastic when it comes to higher value ranges.

Data Exploration, Cleaning, Wrangling and Engineering

Data Exploration Summary

As discussed with our advisor, we have plotted price with respect to many variables. The key challenge has been the overwhelming amount of data. The approach has been to create population segments.

Data Preprocessing

Most of the data preprocessing was already done and consisted in the construction of a ‘master’ data frame containing all listings and calendar info for San Diego county properties.
Additional data processing to prepare segmentation has consisted in grouping by the listings with respect to many key variables (square footage, bed/bath combinations, etc.).

Storing processed and/or integrated data

Processed dataset description for each processed dataset including why you want to process it that way

The data processing has not really changed since the last update. We have a master data frame which has resulted from merging listings with prices. We would like to add data from the reviews file as well but we are processing it with sentiment analysis first. However, at this time we do not see clear way to use the reviews data (more details in update section).
The airbnb scraper is working but scrapes very basic data. As recommended by Julian we need to first construct a final schema for the master listings file and then aim to scrape those specific fields rather than replicating the entire www.insideairbnb.com datasets.
Additional datasets will be used but we are not evaluating their use as part of data exploration at this time.

Table for processed data sets including processed data set name, input datasets, link to the processing scripts and notebooks, and provisional data size

Dataset Name	Input Datasets (Dependencies)	Destination	Related Notebooks, Code, Documents	Provisional Data Size	Other Notes
Master	Listings, Reviews, Calendar. Eventually more with different structure.	AWS	listings_explo .ipynb	125 MB	

Feature Engineering and Data Modeling

–Summary of feature sets

```
'id', u'listing_url', u'scrape_id', u'last_scraped', u'name',
       u'summary', u'space', u'description', u'experiences_offered',
       u'neighborhood_overview', u'notes', u'transit', u'access',
       u'interaction', u'house_rules', u'thumbnail_url',
u'medium_url',
       u'picture_url', u'xl_picture_url', u'host_id', u'host_url',
       u'host_name', u'host_since', u'host_location', u'host_about',
```

```
       u'host_response_time', u'host_response_rate',
u'host_acceptance_rate',
       u'host_is_superhost', u'host_thumbnail_url',
u'host_picture_url',
       u'host_neighbourhood', u'host_listings_count',
       u'host_total_listings_count', u'host_verifications',
       u'host_has_profile_pic', u'host_identity_verified', u'street',
       u'neighbourhood', u'neighbourhood_cleansed',
       u'neighbourhood_group_cleansed', u'city', u'state', u'zipcode',
       u'market', u'smart_location', u'country_code', u'country',
u'latitude',
       u'longitude', u'is_location_exact', u'property_type',
u'room_type',
       u'accommodates', u'bathrooms', u'bedrooms', u'beds',
u'bed_type',
       u'amenities', u'square_feet', u'price', u'weekly_price',
       u'monthly_price', u'security_deposit', u'cleaning_fee',
       u'guests_included', u'extra_people', u'minimum_nights',
       u'maximum_nights', u'calendar_updated', u'has_availability',
       u'availability_30', u'availability_60', u'availability_90',
       u'availability_365', u'calendar_last_scraped',
u'number_of_reviews',
       u'first_review', u'last_review', u'review_scores_rating',
       u'review_scores_accuracy', u'review_scores_cleanliness',
       u'review_scores_checkin', u'review_scores_communication',
       u'review_scores_location', u'review_scores_value',
u'requires_license',
       u'license', u'jurisdiction_names', u'instant_bookable',
       u'cancellation_policy', u'require_guest_profile_picture',
       u'require_guest_phone_verification',
u'calculated_host_listings_count',
       u'reviews_per_month'
```

–Table for feature set including links to input datasets, feature engineering scripts and notebooks, and provisional data size

We have not created any new features but that is something we plan to do

Data Access Design

–Design for data querying interfaces

–Justification for manual vs. programmatic access

TBD

Bullets for each team member's individual contributions in Step 3

Sankarshan: constructed sentiment analysis notebook. Continued data exploration
Paul: continued data exploration and took AWS courses
Parinaz: continued data exploration and worked on AWS

Any major updates to Steps 1 and 2 as a result of exploratory data analysis

Status Update:

The team has continued to explore the core airbnb data. The main milestone consists in the creation of segments based on listing features in order for more accurate regressions to be executed. The development with respect to all obvious variables has been made but the main concern is that this may not be sufficient to create enough segments. Additionally, while the master file contains nearly 100 features, one challenge has been that only a handful of them can be used to categorize apartments. However, further exploration is needed as additional categorizations for segment creation might become obvious as a result of feature engineering and other research.

Meanwhile, Sankarshan has also created an ElasticSearch model and started sentiment analysis with the Reviews data. While this task is open ended, an initial idea was centered on the possibility of creating a more insightful review score as an alternative to the 1-5 star standard rating applied. As of now, the outcome of Sankarshan's research is that it will be hard to create such alternative score as the majority of reviews at hand are very positive. One workaround might consist in the creation of a model that does not use the star review score. An alternative workaround might be to find data for a different region in the world where a more useful mix of star scores is present.

Finally, the team has worked on expanding their expertise on AWS. Parinaz and Paul have taken AWS courses on oreilly.com and are working on setting up the environment so it can host large datasets and be used to run notebooks.

Next steps:

1) Complete the main milestone (paragraph 1) by creating more segments and run regressions for them
2) Continue to explore sentiment analysis and work on alternative approaches
3) Load data into AWS and make basic preparations for the environment