Airbnb Pricing
Sankarhan Acharya, Paul de Fusco, Parinaz Azdavari
Report 2 – Data Pipelines and Process

1. Raw Data Sources

- Summaries of each dataset description from report 1
- Table for Dataset name, source location, destination in your data pipeline, data movement and processing scripts and notebooks, and data size

Dataset Name	**Source**	**Destination**	**Acquisition Notebooks, Code, Documents**	**Data Size**	**Other Notes, e.g., Confidentiality, Notes from data provider. Etc.**
Affordability_Wide_2017Q3_Public.csv	www.kaggle.com	AWS	Acquired through download	2.6 MB	N/A
MarketHealthIndex_Zip.csv	www.kaggle.com	AWS	Acquired through download	2.7 MB	N/A
Zip_MedianRentalPrice_1Bedroom.csv	www.kaggle.com	AWS	Acquired through download	393 KB	N/A
Zip_MedianRentalPricePerSqft_1Bedroom.csv	www.kaggle.com	AWS	Acquired through download	247 KB	N/A
Zip_PriceToRentRatio_AllHomes.csv	www.kaggle.com	AWS	Acquired through download	8 MB	N/A

Zip_Zhvi_AllHomes.csv	www.kaggle.com	AWS	Acquired through download	25.9 MB	N/A
ZriForecast_Public.csv	www.kaggle.com	AWS	Acquired through download	4 KB	N/A
population_by_zip_2000.csv	www.kaggle.com	AWS	Acquired through download	58.2 MB	N/A
population_by_zip_2010.csv	www.kaggle.com	AWS	Acquired through download	59.6 MB	N/A
price.csv	www.kaggle.com	AWS	Acquired through download	5.1 MB	N/A
pricepersqft.csv	www.kaggle.com	AWS	Acquired through download	5.6 MB	N/A
kaggle_income.csv	www.kaggle.com	AWS	Acquired through download	5 MB	N/A
Air_Quality_Measures_on_the_National_Environmental_Health_Tracking_Network.csv	www.kaggle.com	AWS	Acquired through download	49.6 MB	N/A
Crime_Data_LA_from_2010_to_Present.csv	www.kaggle.com	AWS	Acquired through download	385.3 MB	N/A

2016_General_-_Election_Results_by_precinct__complete_eCanvass_dataset_.csv	www.kaggle.com	AWS	Acquired through download	47.9 MB	N/A
calendar.csv	www.insideairbnb.com	AWS	Acquired through download	62.9 MB	
listings.csv	www.insideairbnb.com	AWS	Acquired through download	29 MB	
reviews.csv	www.insideairbnb.com	AWS	Acquired through download	38.9MB	
data_hotel .zip	www.kaggle.com			434.6 MB	
Weather data TBD	WIP				
Consumer Datasets	WIP				
Addl Datasets TBD					

2. Data Exploration, Cleaning, Wrangling and Engineering

- Data Exploration Summary

Initial discussions on modeling the prediction have been focused on producing a regression model to forecast the ideal price for a home. As a consequence, the initial efforts in data exploration have been centered around identifying variables that have an effect on the price variable.

- Data Preprocessing Approach

So far, we have focused on the datasets provided by insideairbnb.com. These datasets consist in the calendar.csv, the listings.csv and the reviews.csv. We joined the calendar and the listings to produce a larger file named ‘Master’ containing listing information for each calendar day a particular listing was offered. The ultimate goal is to predict the price for each day and produce market aggregate measures.

- Approach for storing processed and/or integrated data

So far the data has been stored on ipython notebooks. In the meantime, Parinaz has been working on setting up an AWS instance which will eventually host all or most of our data.

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

Each dataset is being processed to augment the Master dataset. So far the master dataset has the following columns:

```
id                              int64
listing_url                       object
scrape_id                         int64
last_scraped                      object
name                             object
summary                           object
space                            object
description                       object
experiences_offered                object
neighborhood_overview               object
notes                            object
transit                          object
access                           object
interaction                       object
house_rules                        object
thumbnail_url                      object
medium_url                         object
picture_url                       object
xl_picture_url                     object
```

```
host_id                            int64
host_url                          object
host_name                         object
host_since                        object
host_location                     object
host_about                        object
host_response_time                object
host_response_rate                object
host_acceptance_rate              object
host_is_superhost                 object
host_thumbnail_url                object
maximum_nights                     int64
calendar_updated                  object
has_availability                 float64
availability_30                    int64
availability_60                    int64
availability_90                    int64
availability_365                   int64
calendar_last_scraped             object
number_of_reviews                  int64
first_review                      object
last_review                       object
review_scores_rating             float64
review_scores_accuracy           float64
review_scores_cleanliness        float64
review_scores_checkin            float64
review_scores_communication      float64
review_scores_location           float64
review_scores_value              float64
requires_license                  object
license                          float64
jurisdiction_names                object
instant_bookable                  object
cancellation_policy               object
require_guest_profile_picture     object
require_guest_phone_verification  object
calculated_host_listings_count     int64
reviews_per_month                float64
listing_id                         int64
date                              object
price_y                          float64
```

In summary, this dataset produces price curves for each listing as a function of many variables.

- 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	

3. Approach for Feature Engineering and Data Modeling

- Summary of feature sets
- Table for feature set including links to input datasets, feature engineering scripts and notebooks, and provisional data size

Feature	Input Datasets (Dependencies)	Destination	Related Notebooks, Code, Documents	Other Notes
The features are listed section 2	listings, reviews, calendar	listings_explo .ipynb	listings_explo .ipynb	

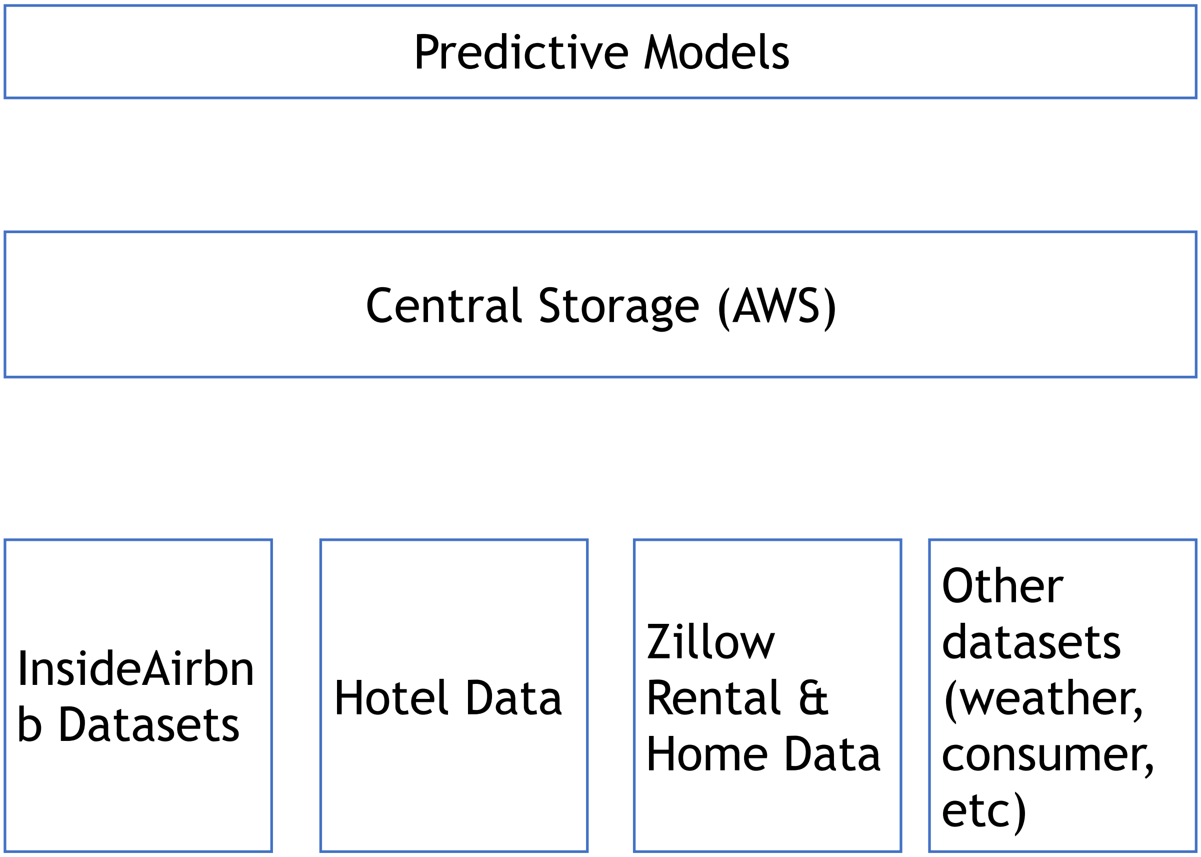
4. Approach for Data Access

- Initial design for data querying interfaces
- Justification for manual vs. programmatic access

The data querying interfaces will ultimately provide access to the central logical schema. At this point there hasn't been a need for designing them as the datasets used are all imported manually into the ipython notebook.
Ideally we would like to opt for programmatic access

5. Data Pipeline

- Description of the needs, approach, and data access and refresh frequency
- Logical diagram showing major data pipeline components for data sources and sinks



Predictive Models

Central Storage (AWS)

InsideAirbnb Datasets

Hotel Data

Zillow Rental & Home Data

Other datasets (weather, consumer, etc)

Refresh Frequency:

Inside Airbnb: Depends on scraping abilities. Ideally once a day, maybe once a week
Hotel Data: No refresh planned at this time
Zillow Data: No refresh planned at this time
Other datasets: depends on the nature of the dataset. Weather could be daily, consumer more rare

6. Set up for your data environment

- Cloud vs. local, database vs. flat files, etc. (see lecture slides for further suggestions)

7. Bullets for each team member’s individual contributions in Step 2

Paul: produced code to continue data exploration and build predictive model

Sankarshan: produced code to continue data exploration and build predictive model

Parinaz: investigated AWS account opening

8. Any major updates to Step 1 as a result of data pipeline step