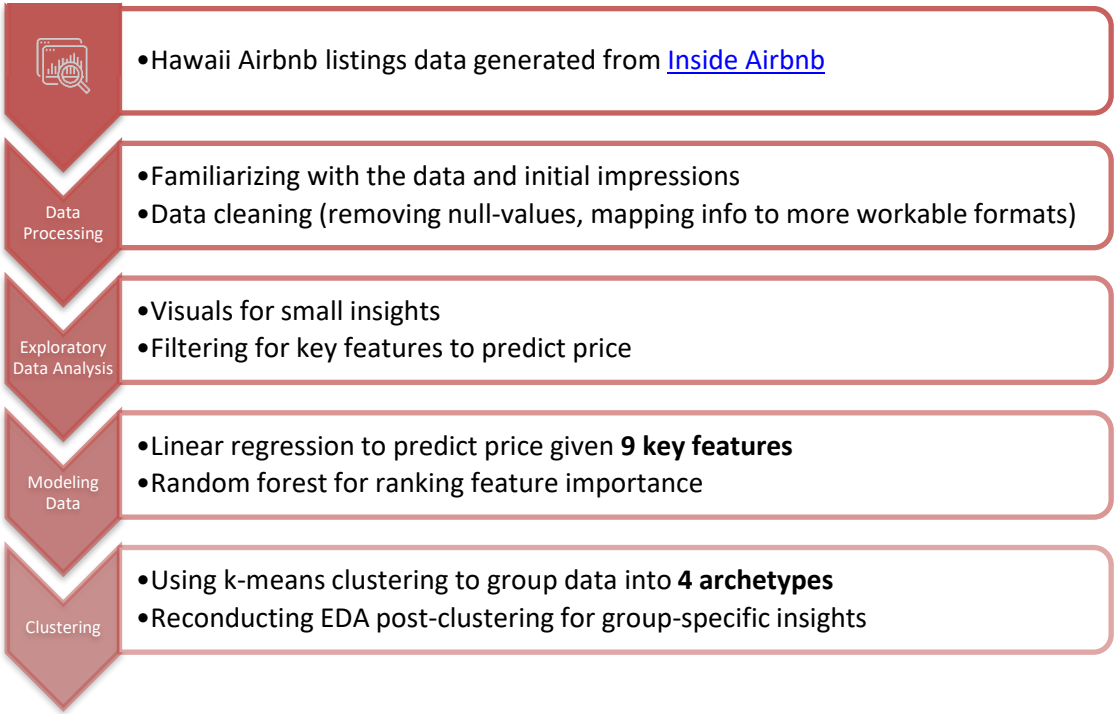
WHARTON ANALYTICS FELLOWS FALL 2022 DATA CHALLENGE

Kyle Liao

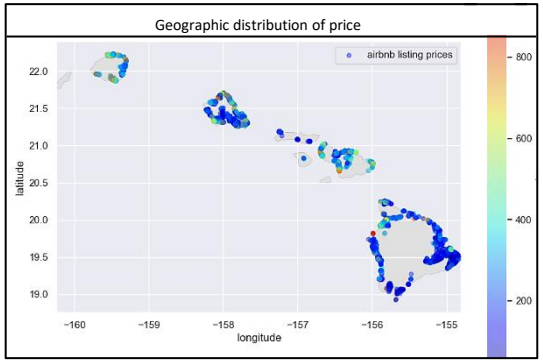
Introduction: Data Analytics Pipeline for Airbnb Data

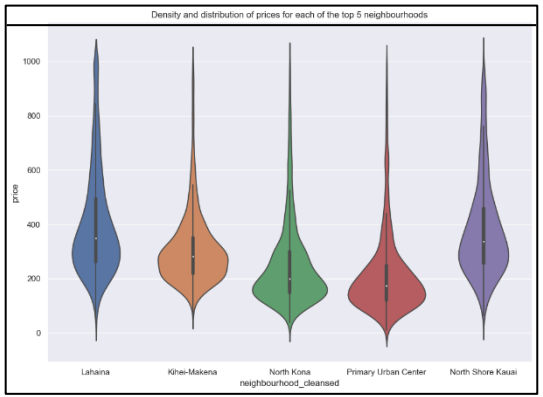


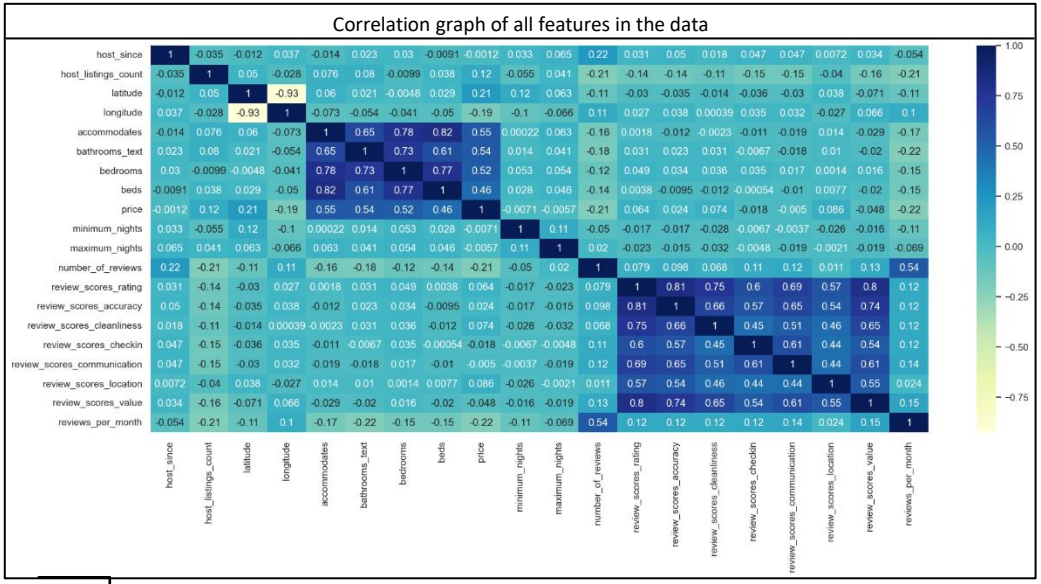
airbnb

Wharton
University of Pennsylvania

Part 1. Exploratory Data Analysis







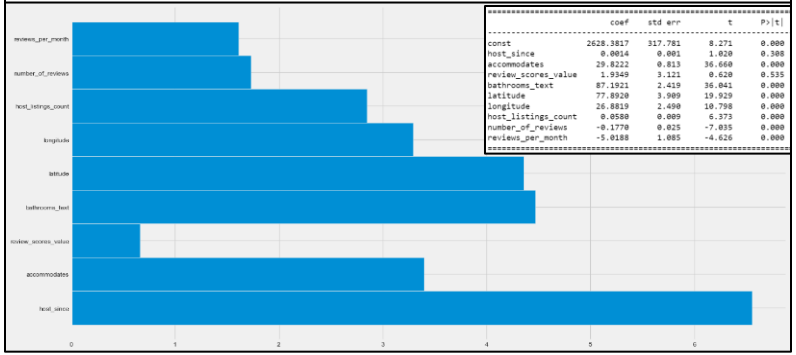
9 key features selected: host experience, # of people accommodated, # of bathrooms, review score, latitude, longitude, host listing count, reviews per month, and number of reviews.

airbnb

Wharton
University of Pennsylvania

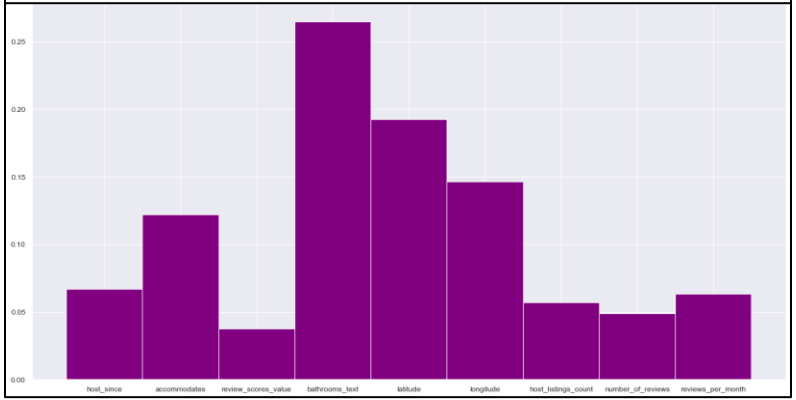
Part 2. Modeling Data

Linear regression coefficients corresponding to features (log scale)



	coef	std err	t	P>|t|
const	2628.3817	317.781	8.271	0.000
host_since	0.0014	0.001	1.020	0.308
accommodates	29.8222	0.813	36.660	0.000
review_scores_value	1.9349	3.121	0.620	0.535
bathrooms_text	87.1921	2.419	36.041	0.000
latitude	77.8920	3.909	19.929	0.000
longitude	26.8819	2.490	10.798	0.000
host_listings_count	0.0580	0.009	6.373	0.000
number_of_reviews	-0.1770	0.025	-7.035	0.000
reviews_per_month	-5.0188	1.085	-4.626	0.000

Feature importance scores generated by random forest decision model



- Linear regression coefficients corresponding to features graphed on a logarithmic scale
- Larger magnitude implies a larger influence on resulting prediction
- Statistics were calculated for each feature; ‘host_since’ and ‘review_scores_value’ insignificant

- Feature importance scores as determined by random forest decision model
- Higher scores (larger magnitude) means that the specific feature will have a larger effect on the price model
- # of bathrooms is the most important feature, followed by latitude, longitude, and # of people accommodated

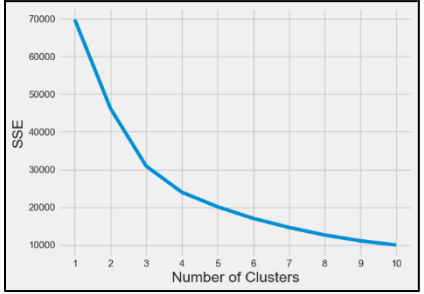
airbnb

Wharton
University of Pennsylvania

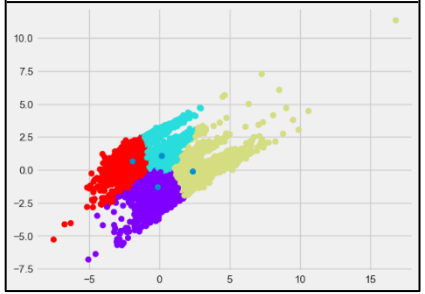
Part 3. Clustering and Post-EDA

Scale features → Use principal component analysis (PCA) for dimension reduction → Use K-means algorithm for clustering → Find 2D vector coordinates → Graph clusters → Conduct post-clustering EDA

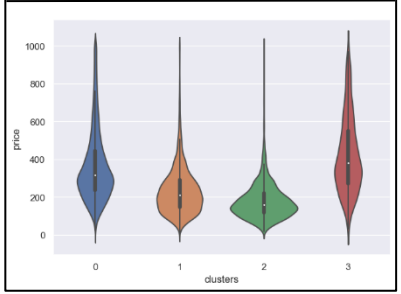
Elbow method graph for cluster count



Graphing coordinates for 2-D post-PCA vectors



Density and distribution of price for 4 clusters



Centroids for four main archetypes of Hawaii Airbnb listings

Cluster	Days hosting	Host listing count	Latitude	Longitude	Guests	Bathrooms	Bedrooms	Beds	Price	Review Score	Reviews	Reviews/month
1	2259.383	111.5177	20.33584	-156.211	5.463832	1.92021	1.975775	2.988989	363.1719	4.647645	18.05997	0.713817
2	2264.835	35.77911	21.30692	-157.692	3.626732	1.059996	1.195245	1.923624	232.4216	4.73904	42.06496	1.411348
3	2547.212	7.241441	19.95914	-155.826	3.102354	1.085949	1.217546	1.691869	177.5453	4.818006	101.4458	2.615278
4	2182.065	103.6904	21.7785	-158.813	5.822527	1.99726	2.169863	3.218874	423.1081	4.529105	13.3726	0.611513

airbnb

Wharton
University of Pennsylvania

Conclusions and Extensions

Step 1. EDA

- Cleaned and processed data
- Created a geographic distribution for price that displays association between latitude and price
- Found neighborhoods with most listings and graphed price density and distribution
- Created correlation graph to isolate 9 key features for regression model

Step 2. Modeling

- Trained linear regression model
- Calculated coefficients and corresponding significance values (p-values) for 9 key features identified in EDA
- Trained a random decision forest to rank feature importance

Step 3. Clustering

- Scaled data down for use in clustering algorithm
- Used PCA to reduce dimension of data from 9 to 2 (for graphing)
- Used K-means clustering algorithm to group data
- Found 4 main archetypes of Hawaii Airbnb listings, as well as centroid for each

Conclusions

- According to the linear regression model, the top 3 features for predicting price are # bathrooms, # of people accommodated, and the latitude
- According to the random decision forest model, the top 3 most important features are # bathrooms, latitude, and longitude, followed by the # of people accommodated

Extension and Improvement

- If I had more time, I would do the following:
  - See how a gradient-boosted tree would predict prices based on features
  - Rigorously check for collinearity before selecting features for models
  - Use Tableau for EDA visualizations (ran into pandas export error)
  - Learn how to implement NLP on text descriptions description

airbnb

Wharton
University of Pennsylvania

Thank you for your time and consideration!