AirXGB

Tom Raczkowski

Question

Which are the most important factors when pricing an Airbnb?

Why is this question important?

- With the rise of Airbnb, the job of pricing vacation rentals in large cities is shifting from massive hotel chains or agents to people with one or two properties and little domain knowledge.

- This presents us with a classic supply and demand problem.

- Priced too high and they will be left with an empty property.
- Priced too low and they are slashing their margins.

Methodology

Utilized a dataset managed by InsideAirbnb, a group who seeks to keep Airbnb honest by scraping the site and keeping the data for some cities freely available on the internet

Obtained and cleaned the data for 9 large American cities (4500+ listings in each market)

Built a model to predict price bins (<$150, $151-$300, $301-$400, $401-$500) using Random Forest and XGBoost algorithms

Compared the cities using the same model to see how similarly they behaved

Assumptions and Preprocessing

Removed information related to the host because they are the ones setting the price.

-Host description, response time, etc.

Backfilled any columns I was planning on using with assumptions.

EX: If a host was marked as having 0 properties I would fill with 1

id	48852 non-null int64
listing_url	48852 non-null object
scrape_id	48852 non-null int64
last_scraped	48852 non-null object
name	48824 non-null object
summary	47464 non-null object
space	33201 non-null object
description	48817 non-null object
experiences_offered	48852 non-null object
neighborhood_overview	28841 non-null object
notes	18385 non-null object
transit	30391 non-null object
access	28476 non-null object
interaction	27346 non-null object
house_rules	29259 non-null object
thumbnail_url	0 non-null float64
medium_url	0 non-null float64
picture_url	48852 non-null object
xl_picture_url	0 non-null float64
host_id	48852 non-null int64
host_url	48852 non-null object
host_name	48746 non-null object
host_since	48746 non-null object
host_location	48563 non-null object
host_about	29776 non-null object
host_response_time	31322 non-null object
host_response_rate	31322 non-null object
host_acceptance_rate	0 non-null float64
host_is_superhost	48746 non-null object
host_thumbnail_url	48746 non-null object
host_picture_url	48746 non-null object
host_neighbourhood	42031 non-null object
host_listings_count	48746 non-null float64
host_total_listings_count	48746 non-null float64
host_verifications	48852 non-null object
host_has_profile_pic	48746 non-null object
host_identity_verified	48746 non-null object
street	48852 non-null object
neighbourhood	48842 non-null object
neighbourhood_cleansed	48852 non-null object
neighbourhood_group_cleansed	48852 non-null object
city	48791 non-null object
state	48851 non-null object
zipcode	48172 non-null object
market	48696 non-null object
smart_location	48852 non-null object
country_code	48852 non-null object
country	48852 non-null object

latitude	48852 non-null float64
longitude	48852 non-null float64
is_location_exact	48852 non-null object
property_type	48852 non-null object
room_type	48852 non-null object
accommodates	48852 non-null int64
bathrooms	48713 non-null float64
bedrooms	48796 non-null float64
beds	48783 non-null float64
bed_type	48852 non-null object
amenities	48852 non-null object
square_feet	533 non-null float64
price	48852 non-null object
weekly_price	8199 non-null object
monthly_price	6706 non-null object
security_deposit	28101 non-null object
cleaning_fee	35867 non-null object
guests_included	48852 non-null int64
extra_people	48852 non-null object
minimum_nights	48852 non-null int64
maximum_nights	48852 non-null int64
calendar_updated	48852 non-null object
has_availability	48852 non-null object
availability_30	48852 non-null int64
availability_60	48852 non-null int64
availability_90	48852 non-null int64
availability_365	48852 non-null int64
calendar_last_scraped	48852 non-null object
number_of_reviews	48852 non-null int64
first_review	37916 non-null object
last_review	37972 non-null object
review_scores_rating	36946 non-null float64
review_scores_accuracy	36870 non-null float64
review_scores_cleanliness	36893 non-null float64
review_scores_checkin	36835 non-null float64
review_scores_communication	36886 non-null float64
review_scores_location	36816 non-null float64
review_scores_value	36817 non-null float64
requires_license	48852 non-null object
license	0 non-null float64
jurisdiction_names	6 non-null object
instant_bookable	48852 non-null object
is_business_travel_ready	48852 non-null object
cancellation_policy	48852 non-null object
require_guest_profile_picture	48852 non-null object
require_guest_phone_verification	48852 non-null object
calculated_host_listings_count	48852 non-null int64
reviews_per_month	37916 non-null float64

Simplifying Input for the Model

To predict more accurately, some portions of the data needed to be removed…

There were some listings that were outliers. Some were too expensive, and others had data that was just plain wrong.

EX: A shared room in Brooklyn with 1 bedroom and 17 bathrooms for $35.00 a night?

-I trimmed my data (losing around 3% of usable entries for NYC) to under $500.00 per night.

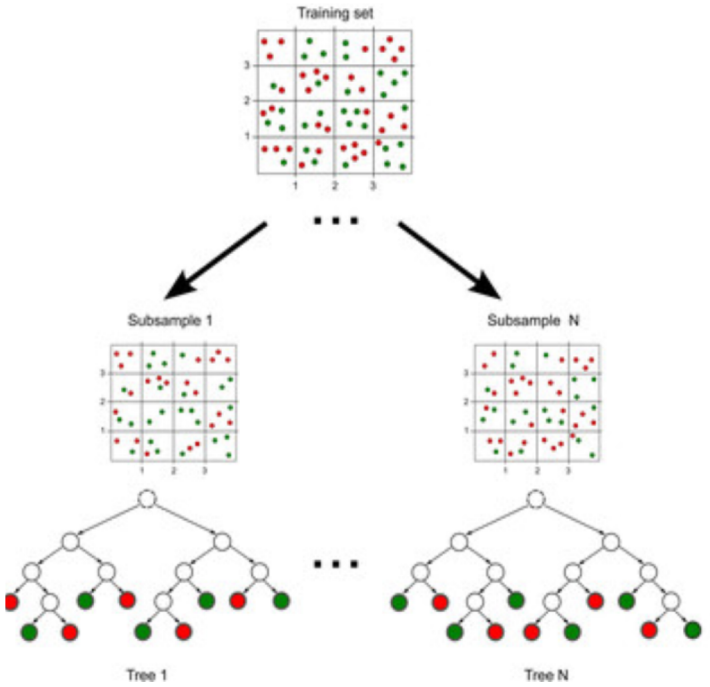
```
nyc[nyc.bathrooms == 17]
```

	property_type	room_type	accommodates	bathrooms	bedrooms	beds	price	cleaning_fee	latitude	longitude
44853	House	Shared room	2	17.0	1.0	2.0	35.0	30.0	40.696762	-73.939594

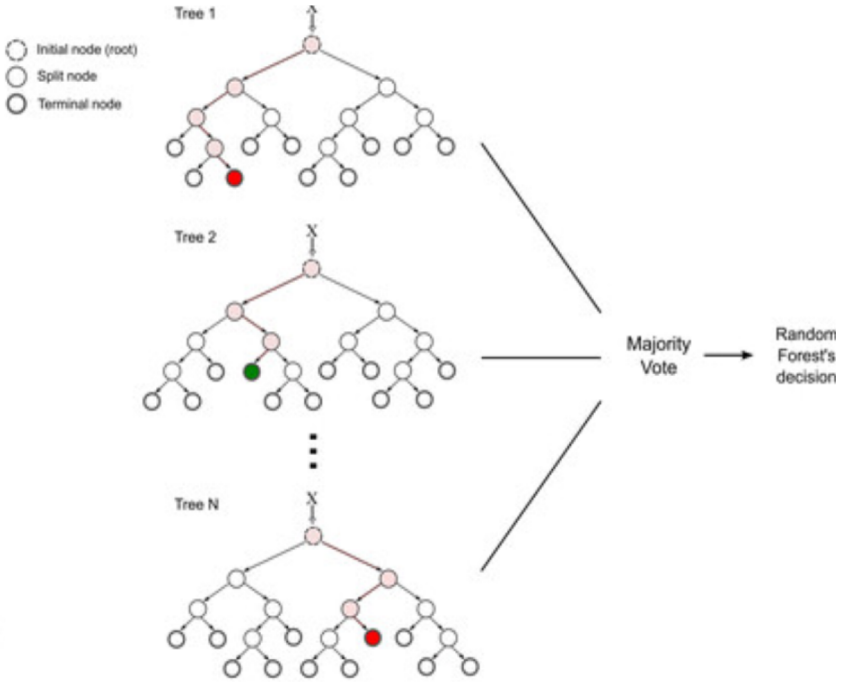
```
nyc = nyc[nyc['bedrooms'] <= 5]
nyc = nyc[nyc['beds'] <= 6]
nyc = nyc[nyc['bathrooms'] <= 4]
nyc = nyc[nyc['accommodates'] <= 10]
nyc = nyc[nyc['price'] <= 500]
```

Models Used

- Random Forest – This model builds decision trees, making splits at random and then aggregates them to lead to a less biased model in the end.



- XGBoost – Also based on decision trees. It is one of the strongest predictive algorithms currently available and is based on gradient boosting.



Most Important Features

As the saying goes:

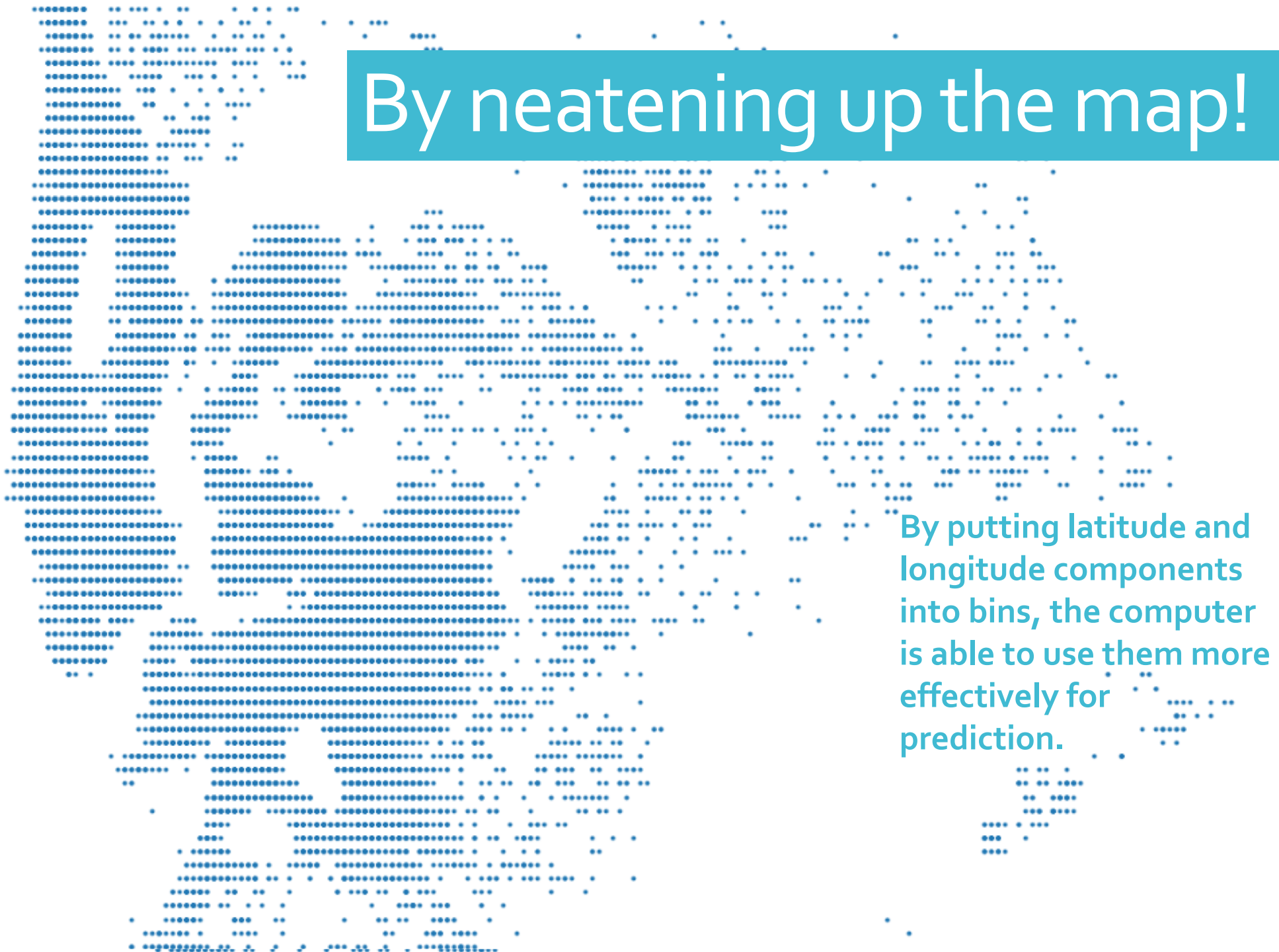
"Location, Location, Location!"

Latitude and Longitude were consistently at the top of the feature importance when it came to the models of all cities.

Both ~30%

How can we use location?

By neatening up the map!



By putting latitude and longitude components into bins, the computer is able to use them more effectively for prediction.

Other Notable Features

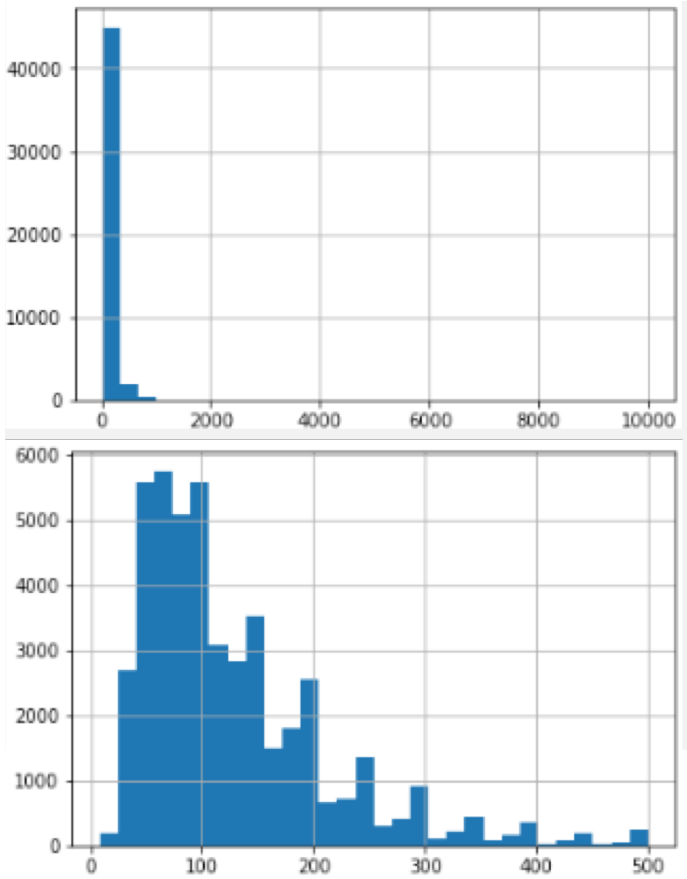
- The number of bedrooms, bathrooms, and beds included in the rental.
- Amenities that were included with the rental were also considered as part of the final model.
  - Common things like Wifi, TV, a hair dryer, etc.
  - Things denoting luxury such as a doorman, a pool, fireplace, etc.
  - Features which may be important for smaller subsets of people or families.
    - Disabled Parking spot
    - Ground floor access
    - Changing table
    - Outlet covers

New York City

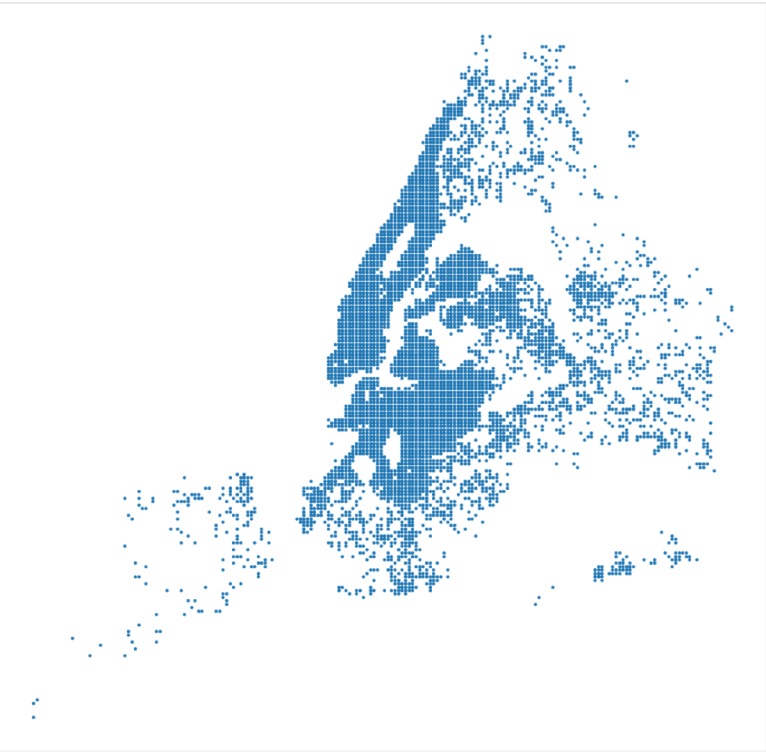
Best Random Forest Accuracy: 71.11%
Best Xgboost Accuracy: 71.38%

Top 3 Amenities (Phase 2):

Family/Kid Friendly:	6.49%
TV:	5.25%
Lock on Bedroom Door:	3.35%



	importance
longs	0.349399
lats	0.279368
accommodates	0.122990
bedrooms	0.089245
bathrooms	0.080988
beds	0.078010

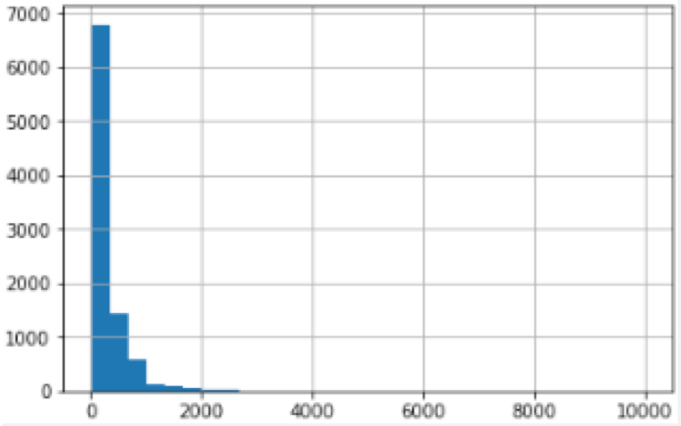


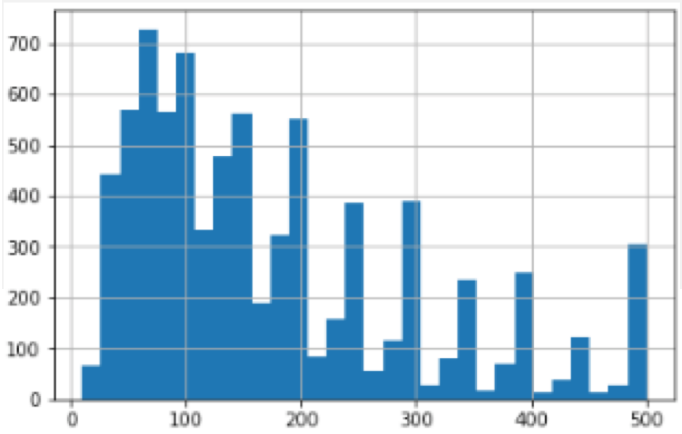
Austin

Best Random Forest Accuracy: 59.43%
Best Xgboost Accuracy: 58.58%

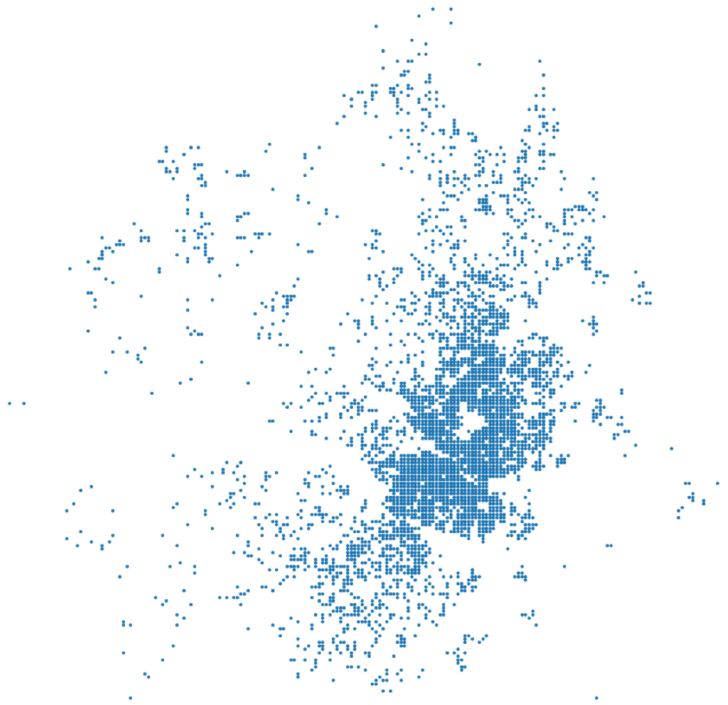
Top 3 Amenities (Phase 2):

Cable TV:	4.19%
Pets Live on The Property	2.49%
Lock on Bedroom Door:	2.11%





	importance
longs	0.380143
lats	0.269449
accommodates	0.113901
bathrooms	0.092954
beds	0.080168
bedrooms	0.063385

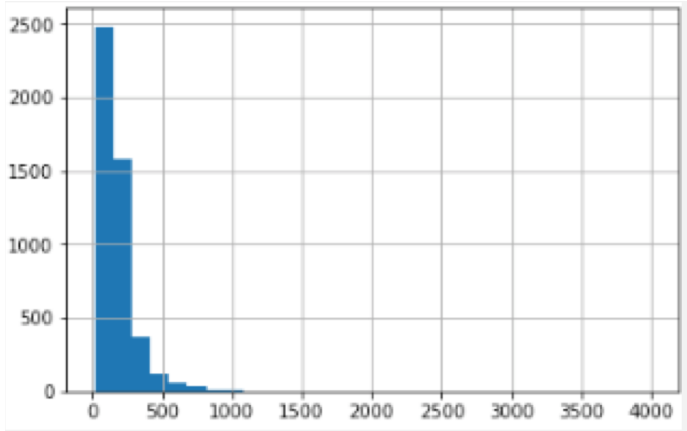


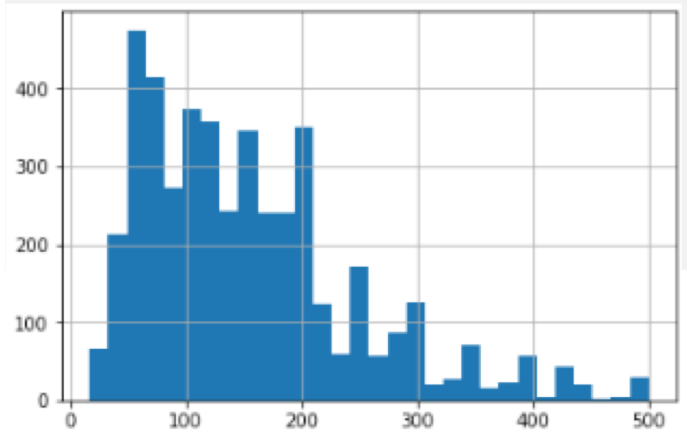
Boston

Best Random Forest Accuracy: 70.43%
Best Xgboost Accuracy: 69.97%

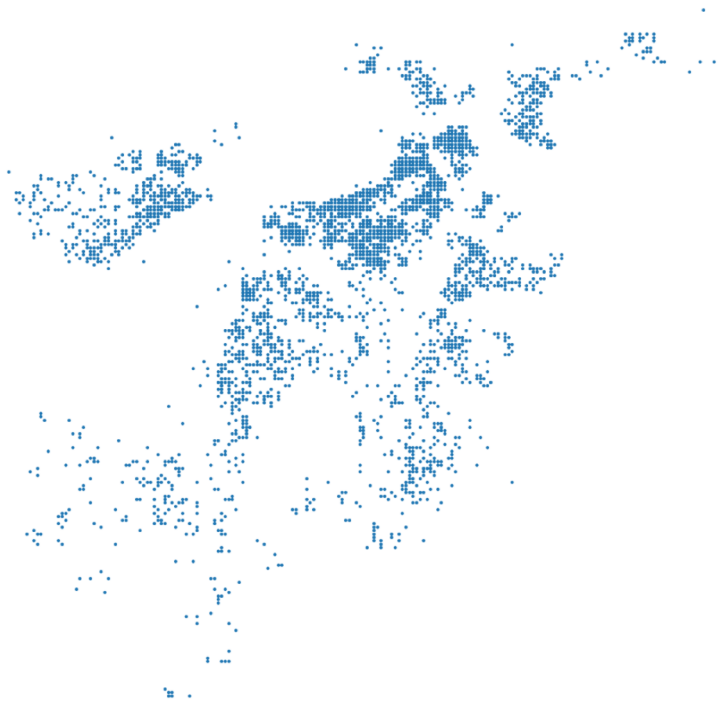
Top 3 Amenities (Phase 2):

TV:	15.0%
Family/Kid Friendly:	5.17%
Lock on Bedroom Door:	4.36%





	importance
longs	0.306106
lats	0.287602
accommodates	0.163451
beds	0.085911
bathrooms	0.079038
bedrooms	0.077892

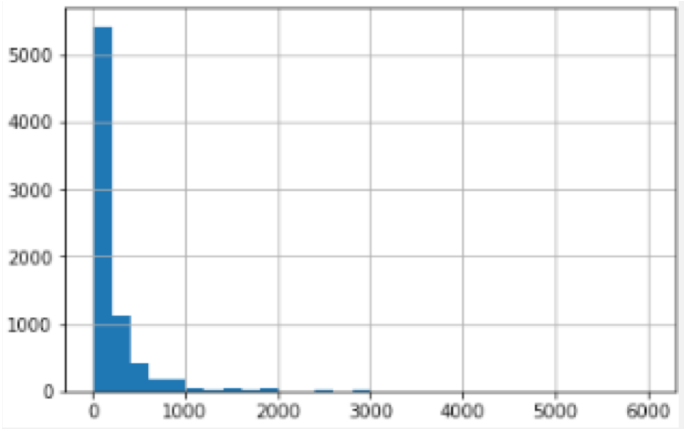


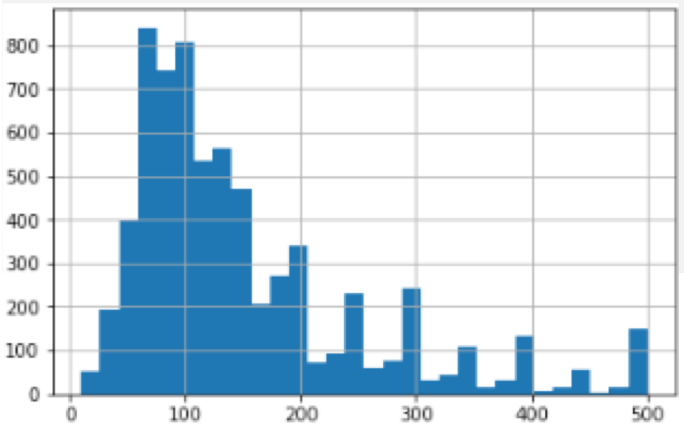
Washington DC

Best Random Forest Accuracy: 64.10%
Best Xgboost Accuracy: 63.19%

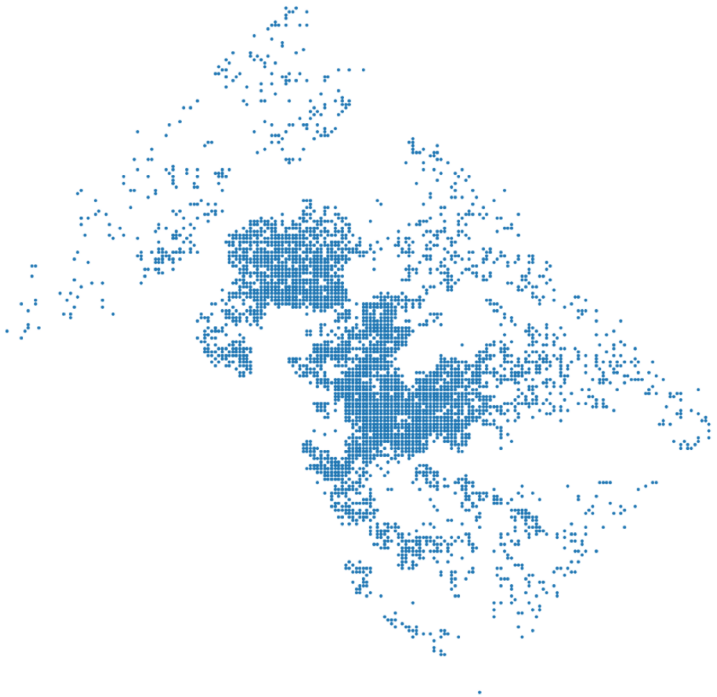
Top 3 Amenities (Phase 2):

TV:	6.12%
Family/Kid Friendly	4.48%
Gym:	3.15%





	importance
longs	0.324132
lats	0.267740
accommodates	0.141307
bathrooms	0.092090
beds	0.088197
bedrooms	0.086534

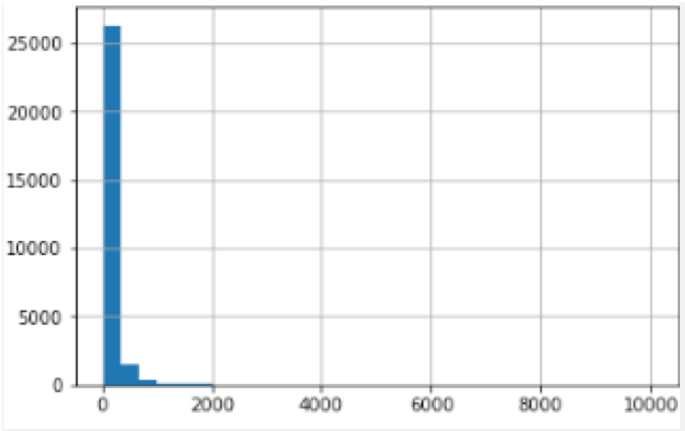


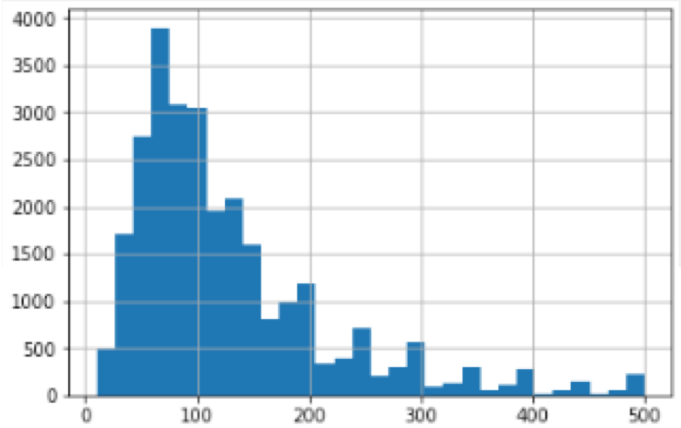
Los Angeles

Best Random Forest Accuracy: 69.36%
Best Xgboost Accuracy: 70.49%

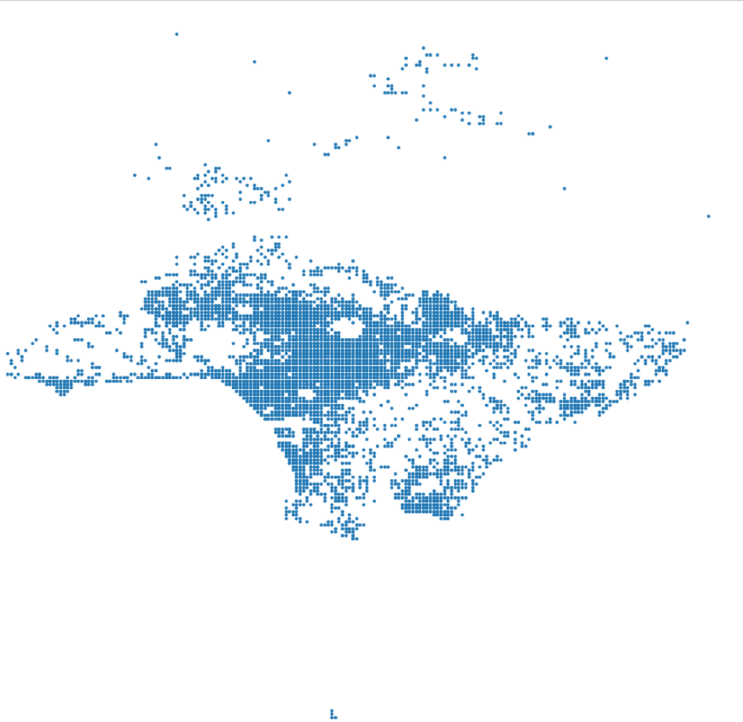
Top 3 Amenities (Phase 2):

TV	4.92%
Family/Kid Friendly:	4.45%
Lock on Bedroom Door:	2.34%





	importance
longs	0.390756
lats	0.297963
accommodates	0.117806
bathrooms	0.069883
beds	0.065443
bedrooms	0.058150

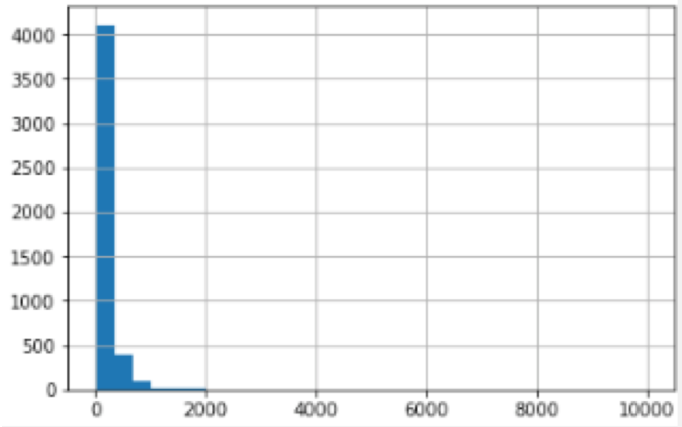


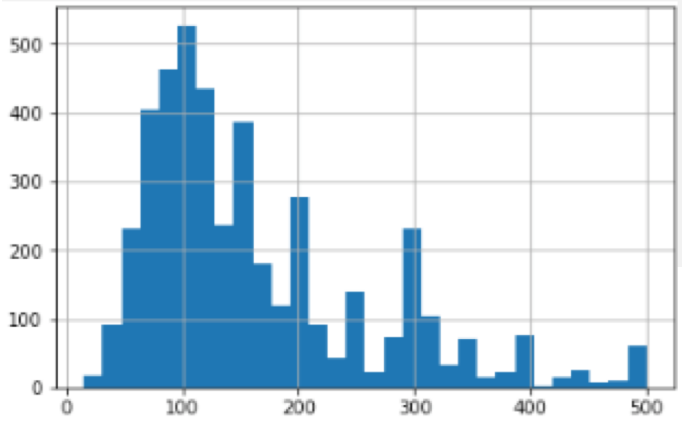
New Orleans

Best Random Forest Accuracy: 57.28%
Best Xgboost: 61.10%

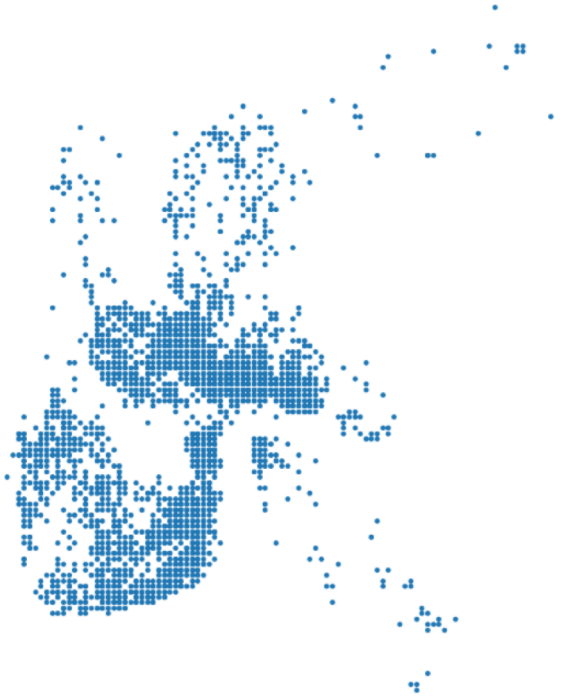
Top 3 Amenities (Phase 2):

Elevator	6.54%
Pets live on this property:	4.43%
Cable TV:	3.91%





	importance
longs	0.318255
lats	0.238131
accommodates	0.138072
beds	0.130174
bathrooms	0.089669
bedrooms	0.085699

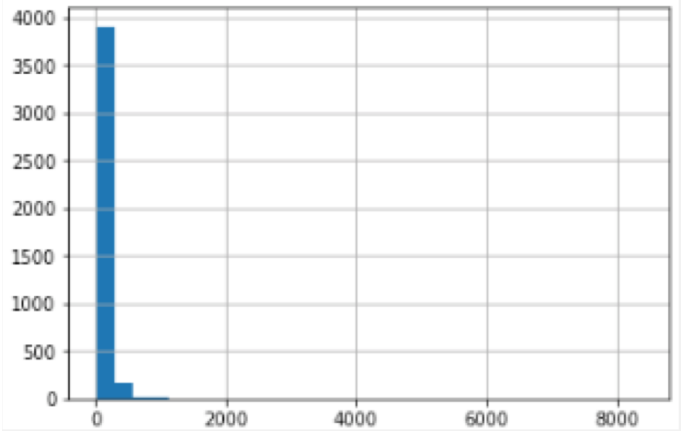


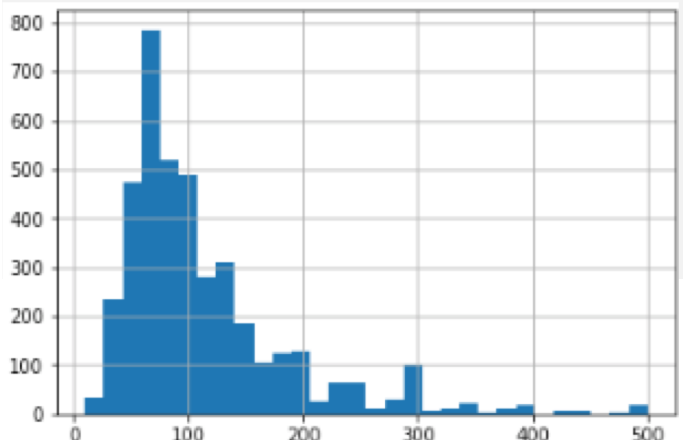
Portland

Best Random Forest Accuracy: 72.76%
Best Xgboost: 71.69%

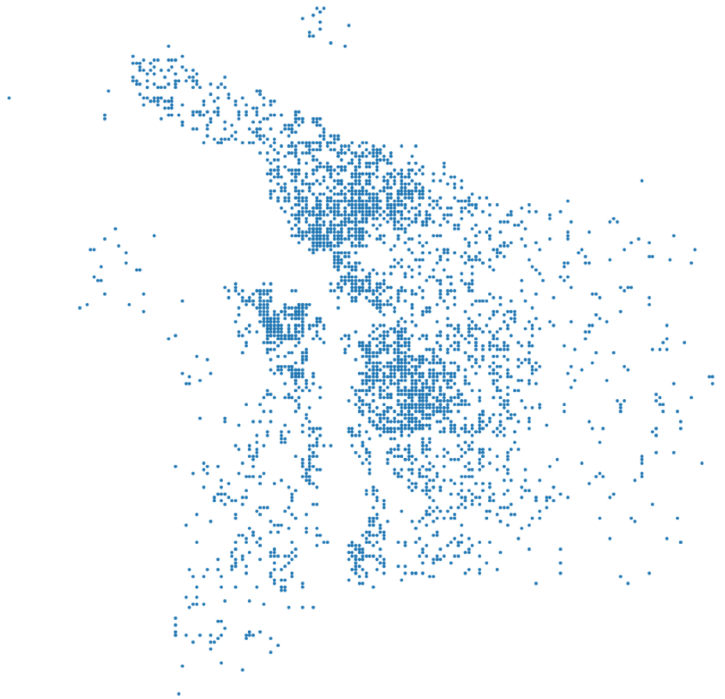
Top 3 Amenities (Phase 2):

Elevator:	**6.17%**
TV:	**4.55%**
Family/Kid Friendly:	**4.31%**





	importance
longs	0.320857
lats	0.254800
accommodates	0.142996
bathrooms	0.103354
beds	0.102602
bedrooms	0.075392

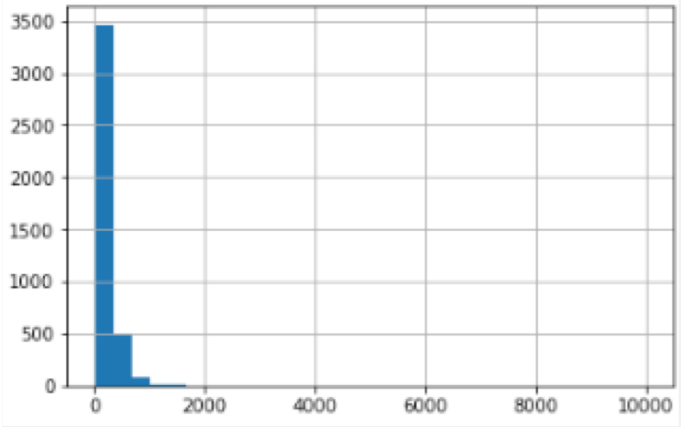


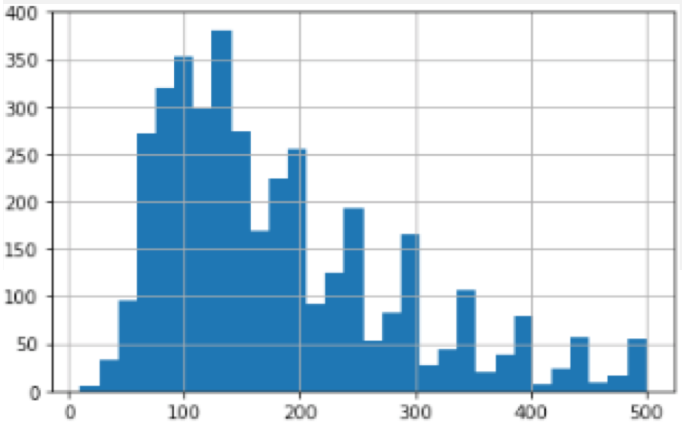
San Francisco

Best Random Forest Accuracy: 69.94%
Best Xgboost Accuracy: 69.59%

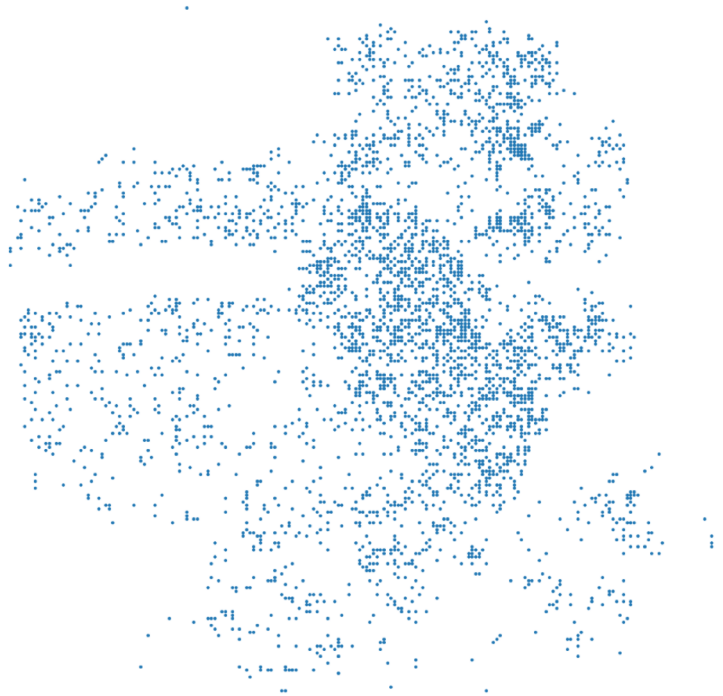
Top 3 Amenities (Phase 2):

Elevator in Building:	7.32%
Cable TV:	5.69%
Family/Kid Friendly	3.87%





	importance
longs	0.366317
lats	0.228682
accommodates	0.142388
bathrooms	0.107168
beds	0.087627
bedrooms	0.067818

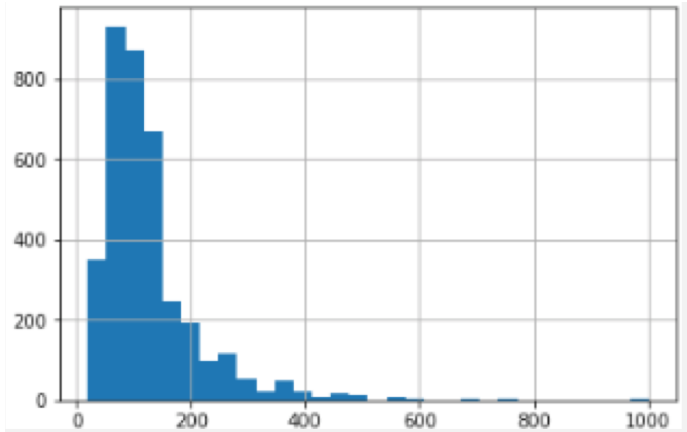


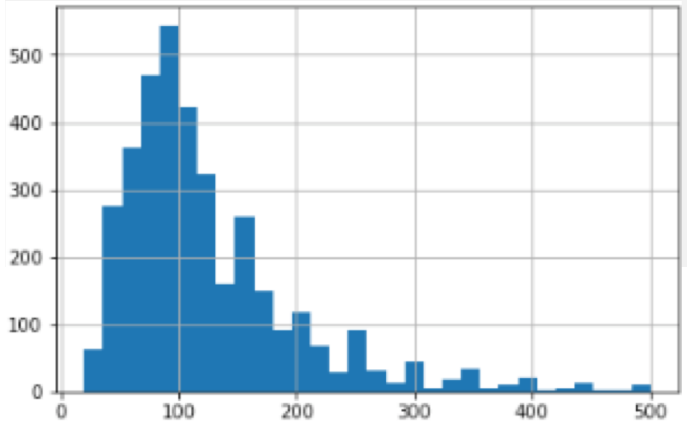
Seattle

Best Random Forest Accuracy: 56.72%
Best Xgboost Accuracy: 58.89%

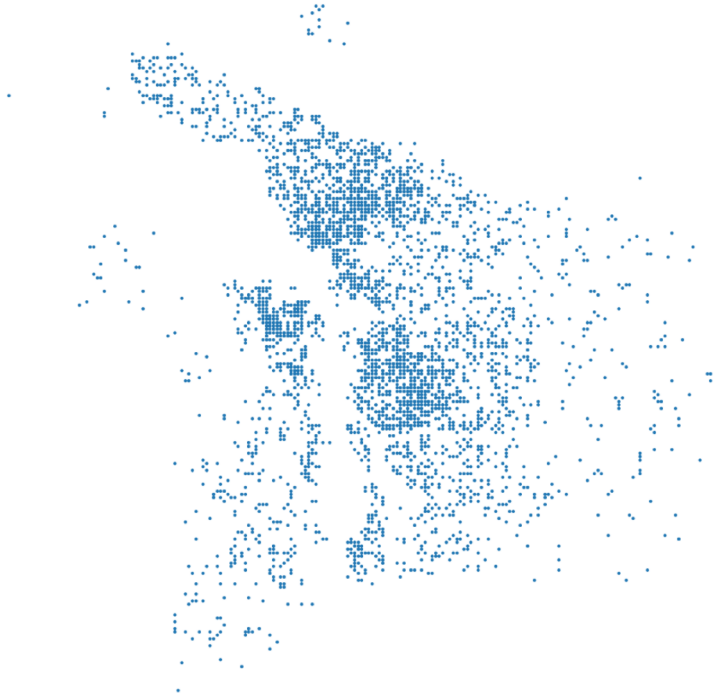
Top 3 Amenities (Phase 2):

TV:	7.01%
Family/Kid Friendly:	3.77%
Lock on Bedroom Door:	2.43%





	importance
longs	0.311484
lats	0.223275
accommodates	0.159433
beds	0.108736
bathrooms	0.105322
bedrooms	0.091751



Upcoming Possibilities

Some possibilities for the near future would be continuing to tune the models to bring better results with less bias, as well as a usable webapp for hosts to enter their own listing data and get a second opinion.

Beyond that, understanding how the shape of a city can affect pricing of Airbnb units could provide more insight

- Distance to crucial points (transportation, attractions, shopping, etc.) can almost definitely help with a model for pricing.

Thanks for taking a look!

- A special thanks to InsideAirbnb for keeping their data open!
  - http://insideairbnb.com