Airbnb New York Data Analysis

Data Science Project Fall 2018

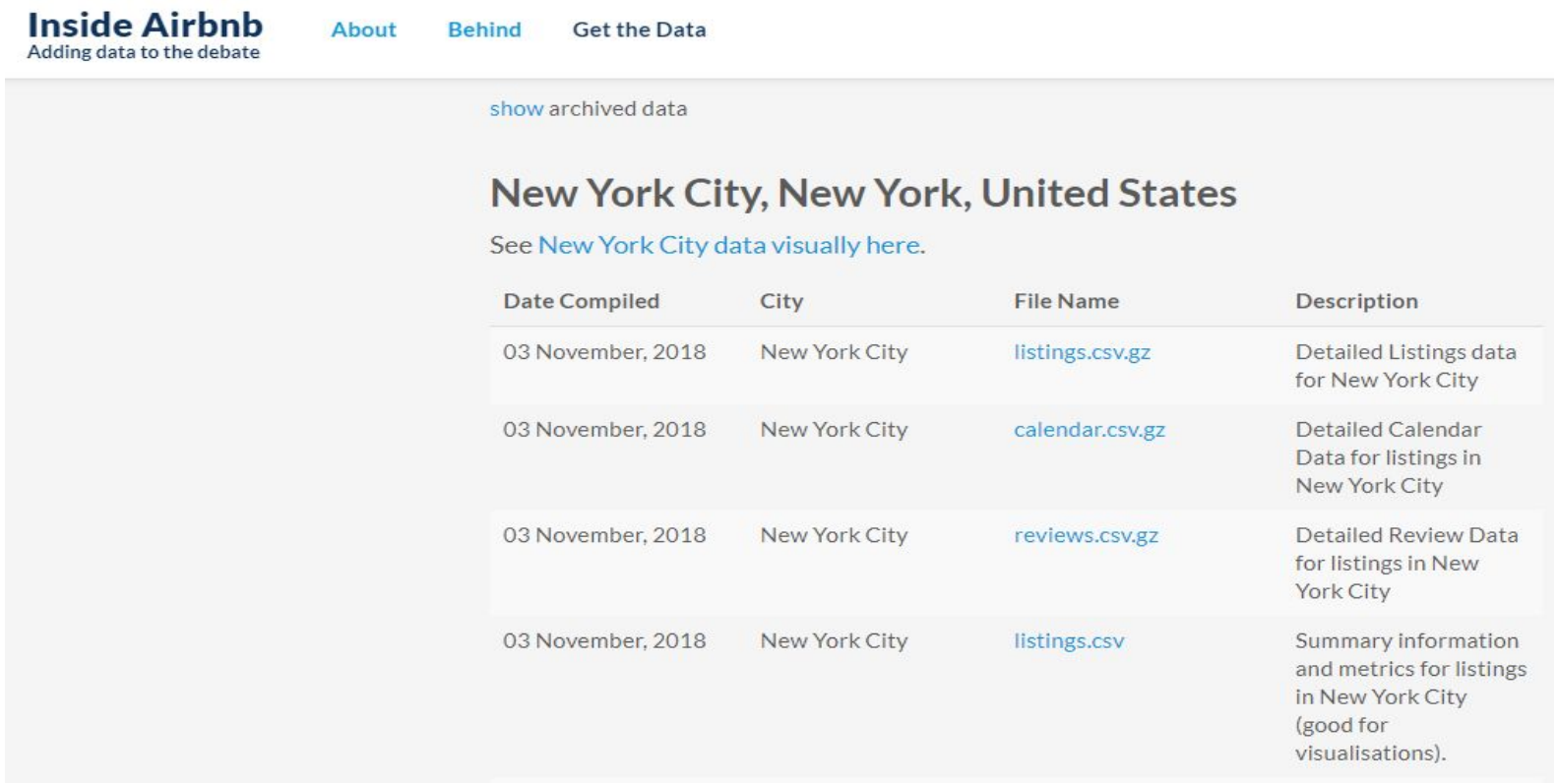
Divya Natolana Ganapathy
Prajwal Gadad

Objectives

- **Customer Point of view:**
  - What is it that makes the price of one listing different from another

- **Host Point of view:**
  - How to become a successful host on Airbnb

Data Source



Inside Airbnb
Adding data to the debate

About Behind Get the Data

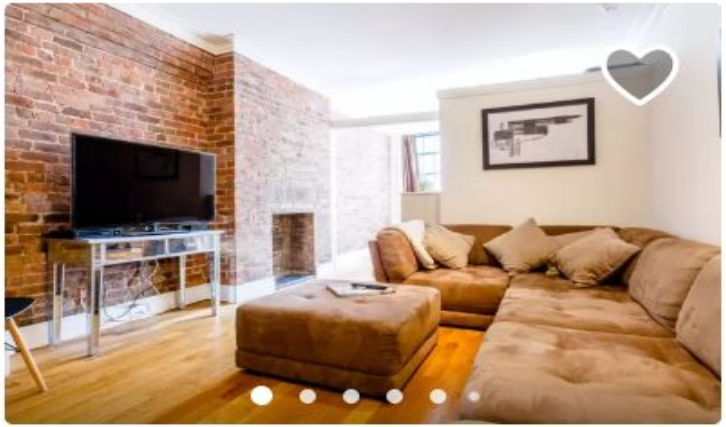
show archived data

New York City, New York, United States

See New York City data visually here.

Date Compiled	City	File Name	Description
03 November, 2018	New York City	listings.csv.gz	Detailed Listings data for New York City
03 November, 2018	New York City	calendar.csv.gz	Detailed Calendar Data for listings in New York City
03 November, 2018	New York City	reviews.csv.gz	Detailed Review Data for listings in New York City
03 November, 2018	New York City	listings.csv	Summary information and metrics for listings in New York City (good for visualisations).

Customer Point of View

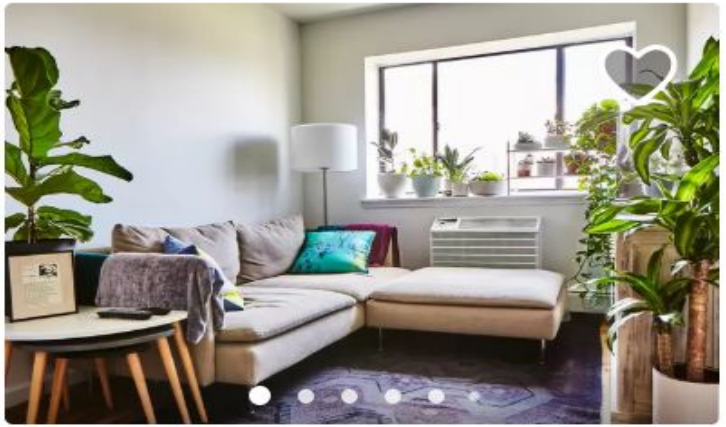


PRIVATE ROOM · 2 BEDS

Soho loft with massive couch for extra sleepers!

$80 per night · Free cancellation

★★★★½ 102



PRIVATE ROOM · 1 BED

Sunny, Modern room in East Village!

$55 per night · Free cancellation

★★★★★ 289 · Superhost

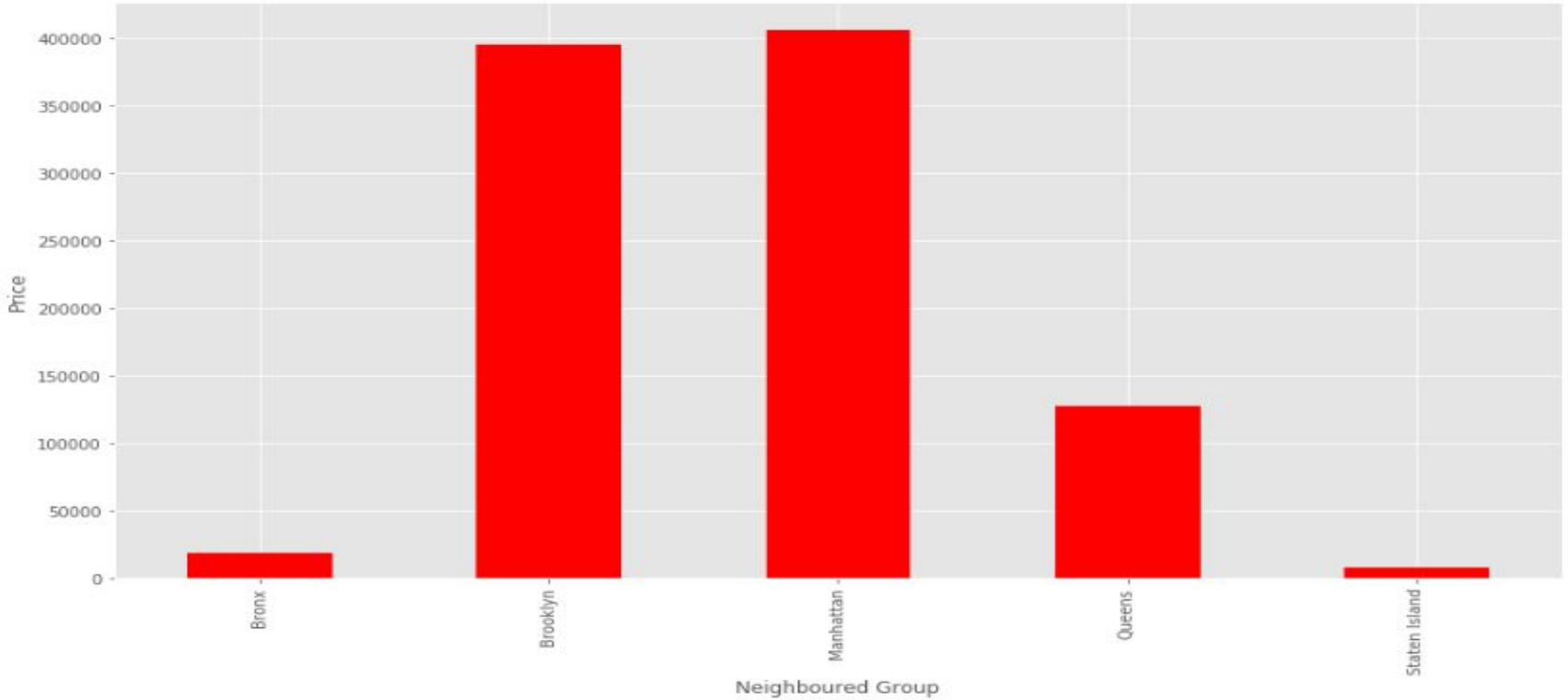
- Listings.csv:

```
In [6]: cleaned_listings.columns

Out[6]: Index(['id', 'name', 'summary', 'space', 'description',
               'neighborhood_overview', 'notes', 'transit', 'access', 'interaction',
               'house_rules', 'host_id', 'host_since', 'host_location', 'host_about',
               'host_response_time', 'host_response_rate', 'host_acceptance_rate',
               'host_is_superhost', 'host_picture_url', 'host_neighbourhood',
               'host_listings_count', 'host_total_listings_count',
               'host_verifications', 'host_has_profile_pic', 'host_identity_verified',
               'street', 'neighbourhood', 'neighbourhood_cleansed',

In [53]: cleaned_listings.shape

Out[53]: (50041, 81)
```

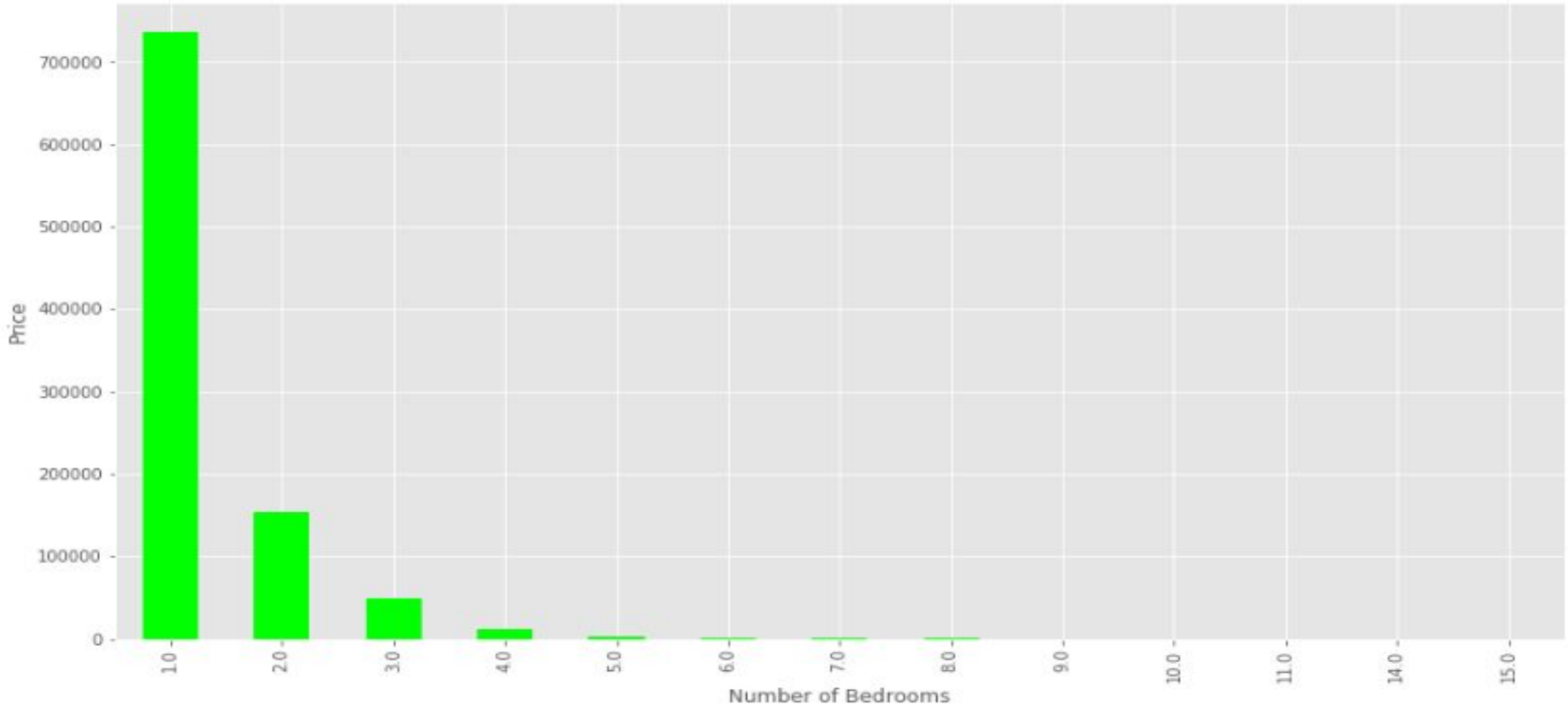
- Reviews.csv:

```
In [55]: print(reviews_detailed.columns)
         print(reviews_detailed.shape)

         Index(['id', 'date', 'reviewer_id', 'reviewer_name', 'comments'], dtype='object')
         (1051974, 5)
```

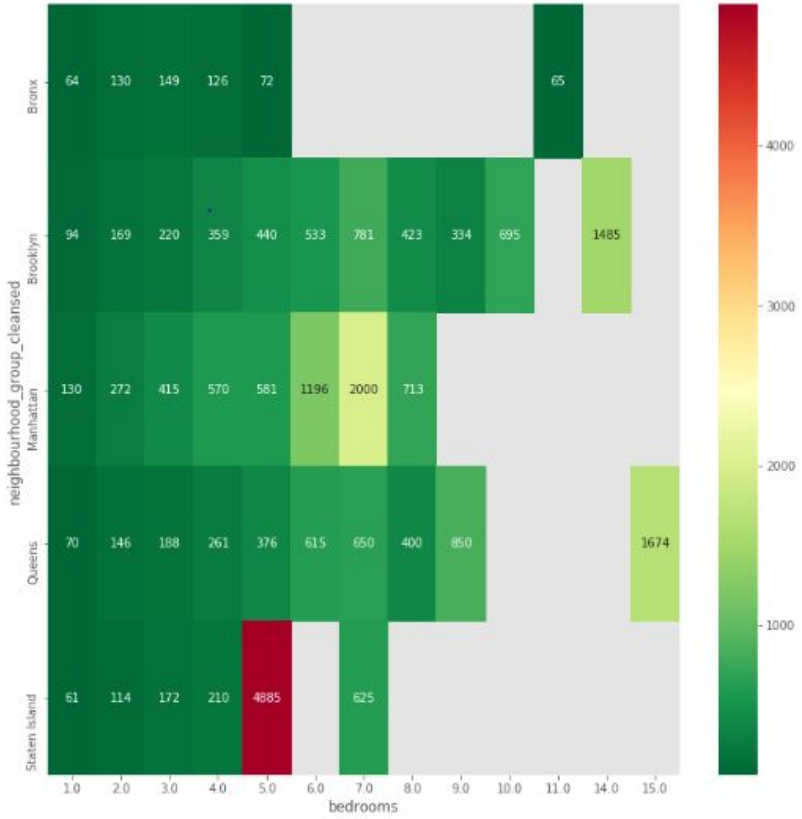
Price vs Neighbourhood Group

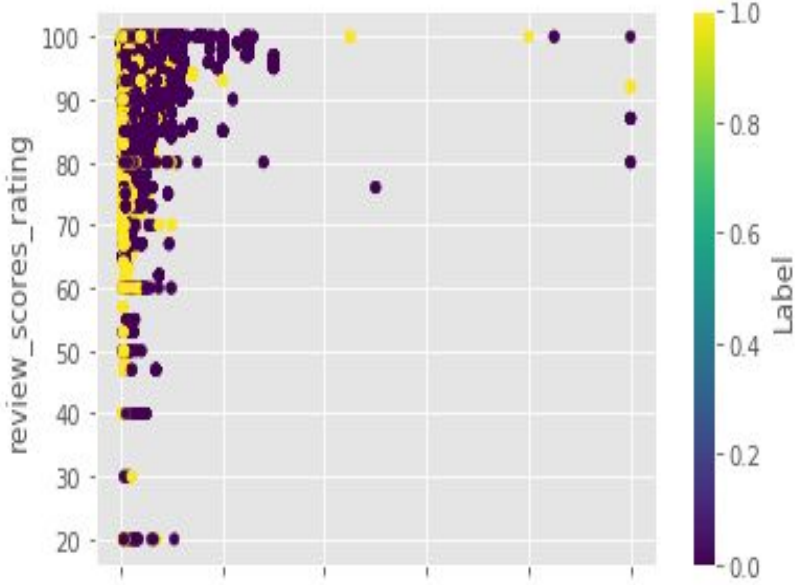


Price vs Number of Bedrooms



Neighbourhood Group vs Bedrooms





Words that affect price

Walking distance

Bedroom

Bed Size

Coffee maker

Queen Size

Hudson River

Clean

Easy


stay
place
great
apartment
location
host
amazing
restaurant
nice
erin
clean
home
check
jessica
experience
time
recommend
room
back
great location
neighborhood
subway
always
comfortable
picture
everything
close
beautiful
responsive
communication
easy
amenity
n't
airbnb
good
exactly
loved
town
thank
city
better
day
view
walk
night
modern
amy
area
described
shop
next
house
bed
needed
way
definitely
fantastic
helpful
friendly
really
space
coffee
shoot
safe
photo
nyc
need
quiet
well
ilo
lot
home
enough
bit
family
even
best
enjoyed
west
nearby
pool
village
plenty
located
convenient
building
question
floor
wonderful
food
feel
though
spacious
sure
walking
bar
coming
felt
highly
made
away
central
staying
perfect
much
accommodating
bathroom
new york
although
subway station
sandra
late
arrival
quick
heart
manhattan
quite
touch
around
respond
chinatown
brooklyn
inside
key
weekend
totally
chinese
henry
complex
come
first
came
listing
flexible
found
drink
looking
pretty
expected
broadway
see
store
comfy
living
problem
decorated
kind
missed
huge
park
seen
every
available
super
overall
look
think
fun
towel
morning
quickly
person
thing
museum
rooftop
cafe
retreat
confused
meet
center
answered
took
booking
sick
point
rum
going
est
paul
near
visit
far
within
happy
couple
attentive
apt
art
rather
distance
seemed
people
contemporary
still

Host Point of View

- Calculated the total revenue of each listings:

  Linear Regression-->0.33 (r2-score)

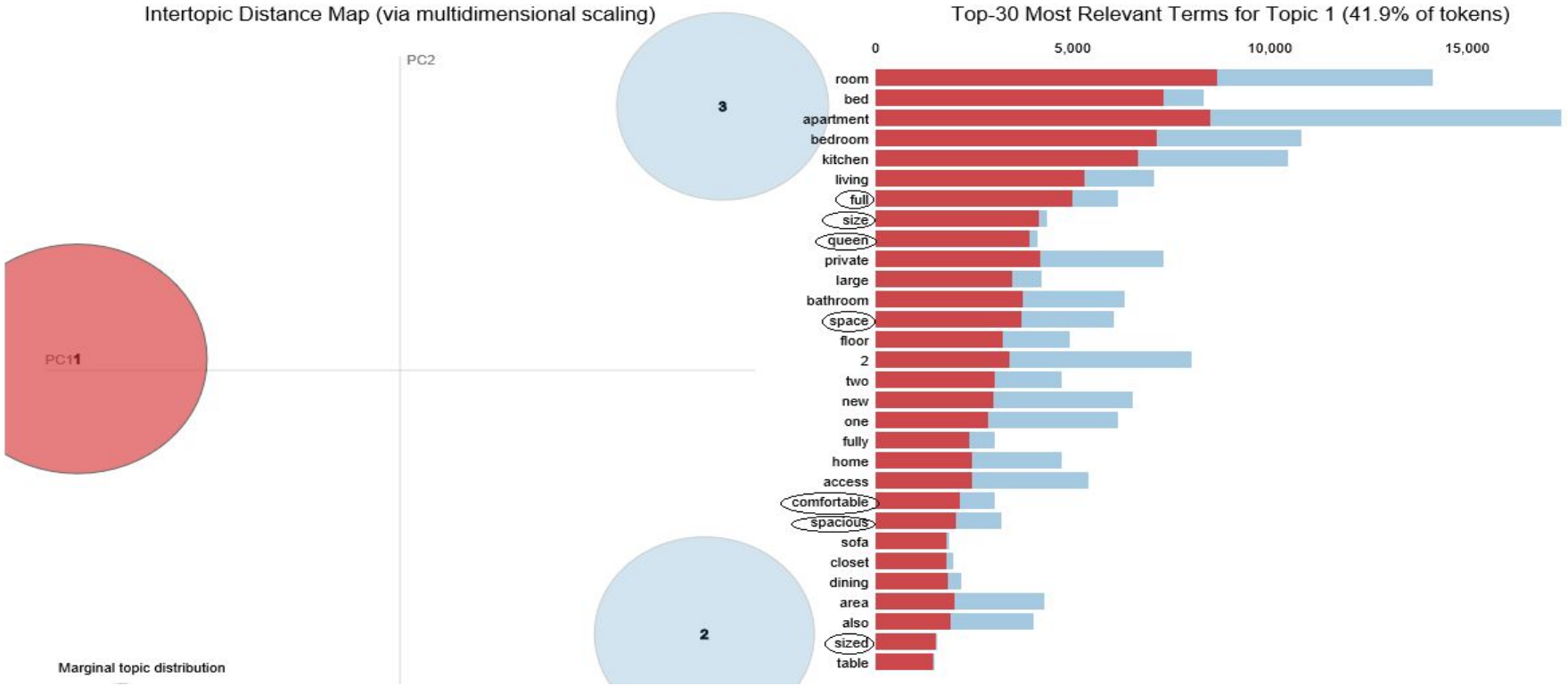
  Random Forest Regressor-->0.49 (r2-score)

```
In [186]: y_test.values[0]
Out[186]: 29836.800000000003

In [185]: regr_y_pred[0]
Out[185]: 34764.59447911568
```

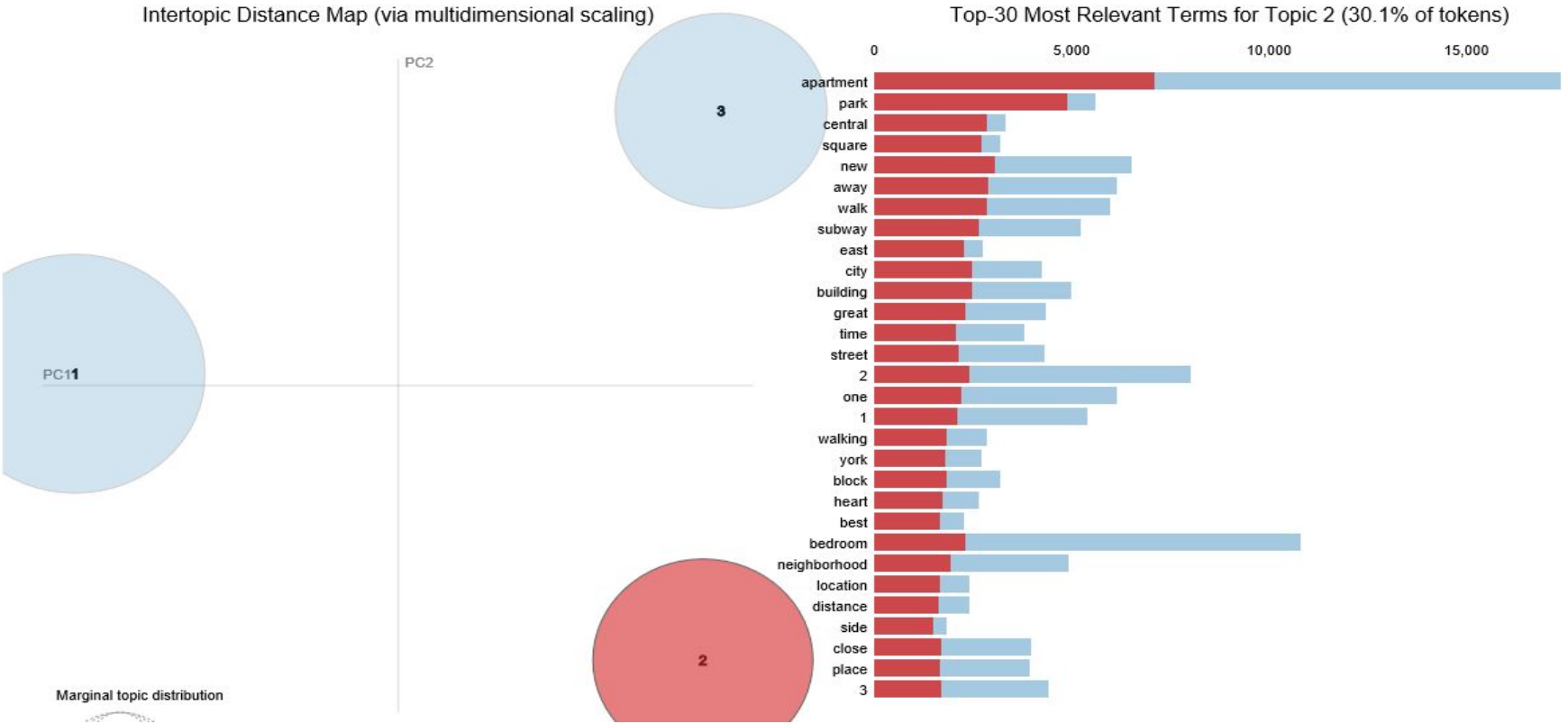
```
In [190]: y_test.values[2]
Out[190]: 19968.0

In [189]: regr_y_pred[2]
Out[189]: 20738.201244561
```

- LDA (Latent Dirichlet Allocation) on Description of Listings: Top 3 topics (Luxury, Location,Budget)

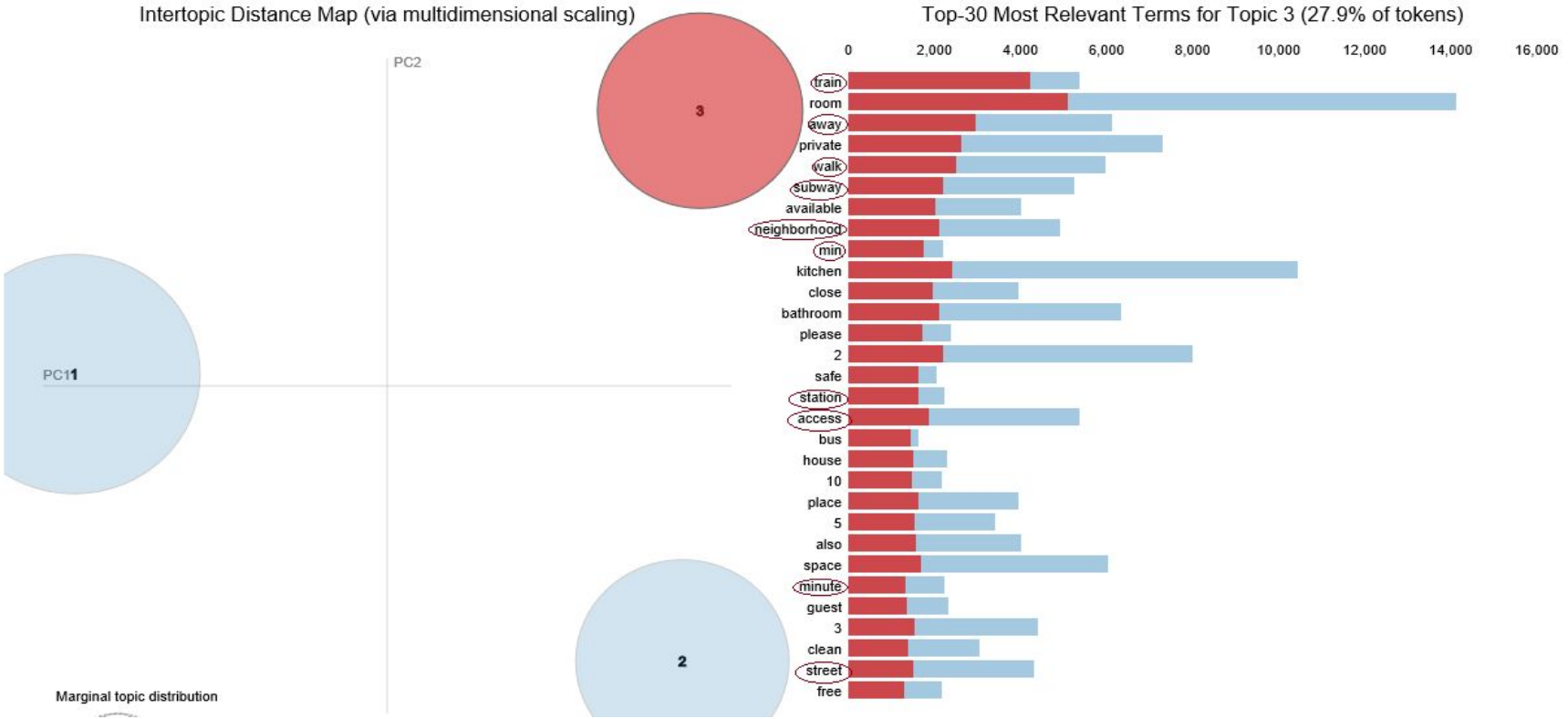
Topic 1 (Luxury) :Appeal to travellers who value Luxury.



Topic 2 (Budget): Appeal to budget travellers.



Topic 3 (Location): Appeal to traveller who value convenience and proximity.



Conclusion

- Price is dependent on the following attributes :
  - Location.
  - Type of rooms.
  - Number of rooms .
  - Keywords used in the description and comments.
- The host could become successful by targeting three types of customer groups.
- We predicted the pricing of a new listing and also were able to suggest keywords that should be used for success in their description.

Failures and Improvements

- Plot of price against the neighbourhood did not give us too much insight.
- The predicted price vs actual price of a listing has some variance.

- To improve the accuracy we need to add more features.
- We can use important keywords extracted from comments and descriptions as feature to the model.

Thank You!