Airbnb Rental Analysis

Wenying Yang, Yixiao Wang, Qingyan Wang

2017, May 8

BU.520.710.51.SP17 Special Topics in Risk Management: Big Data Machine Learning

Aunt **(Money Holder)**

Los Angeles

It sounds a good investment, but I need to see detailed numbers...Call your cousin and ask her

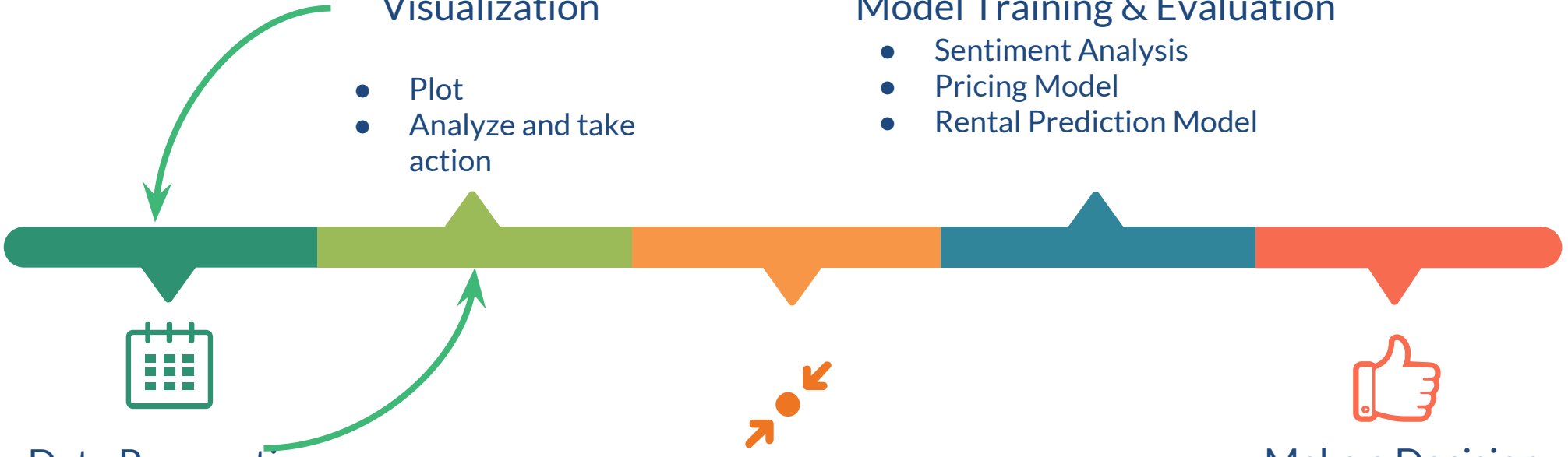
Mia

Boston

I will stay in Boston for 3 years, but the monthly rent is very high there. I'm thinking about buying an apartment and lease the rooms on Airbnb. I was wondering if my mom would be interested in this investment opportunity?

Buy or Not Buy?

Process



Visualization

- Plot
- Analyze and take action

Model Training & Evaluation

- Sentiment Analysis
- Pricing Model
- Rental Prediction Model

Data Preparation

- Collect Data
- Fill Missing Value
- Handle Categorical Columns
- Merge datasets
- Filter Data

Data Transform

- Feature Selection
- Add New Features
- Polynomial Features Expansion

Make a Decision

Findings and Recommendations

Datasets - Listings

Listings.csv includes detailed data of listings in Boston scraped on 2016 September,
Key: Listing ID, 94 Features, 2437 observations

Business Type	Columns	New Columns
Geo Locations	Neighbourhood, Zip Code, Street, Latitude, Longitude...	
Host	Host Since, Host Acceptance Rate, Host Response Rate	Host_weeks: how many weeks till today
Listing Properties	Property Type, Room Type, Bed Type # of Bathroom, # of bed, # of accom, # of guest included Amenities	Split Amenities and convert into dummy variables (0,1)
Review	Total # of Review, Ratings (accuracy, location, etc..)	
Fee	Price, Cleaning fee, Fee for Extra People	Price_Per_Bed: price/number of bed
Reservation	Instant Bookable, Minimum Nights, Maximum Nights	

Datasets - Calendar

Calendar.csv includes detailed calendar data for listings in Boston scraped on 2015 Oct and 2016 Oct

Host can snooze, temporarily or permanently deactivate listings through calendar to pause listing and hide it from search results

Key: Listing ID, 3 Features, 2176812 observations

Columns	Description	New Columns
Date	Calendar Date	
Available	Whether a listing is available for booking on a specific date. Note: it is not relevant to whether the listing has been booked.	ava_2016: Count available days by listing id between Sep 2015 to Sep 2016 ava_2017: Also count available days by listing id between Oct 2016 to Oct 2017. Also add available days during boston Airbnb peak season (July-September)
Price	If the hosting is available on a particular date, a price is required	Average Price per listing

Datasets - Reviews

Reviews.csv includes Detailed Review Data for listings in Boston scraped on 2016 Oct.

Key: Listing ID, 6 Features, 37303

80% of hosts leave a review for their guests.
72% of guests leave a review for hosts.

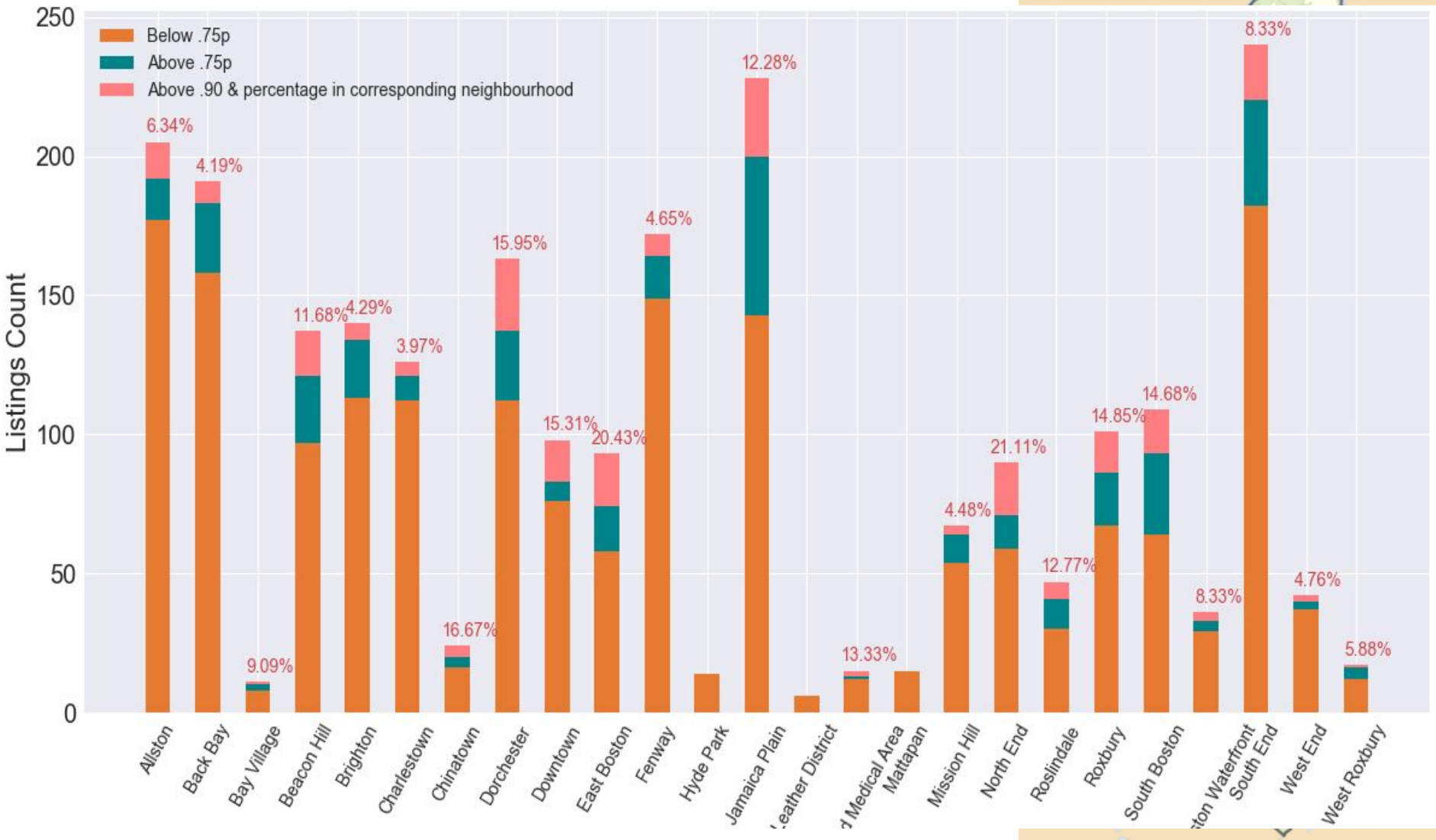
Columns	Description	New Columns
Date	The date guests leave a review	number_reviews_2016: Count number of reviews group by listing id between Sept 2015 and Oct 2016
Comments	Review that guests leave	review_score_range: map comments into three categories based on 'review_scores_rating'.

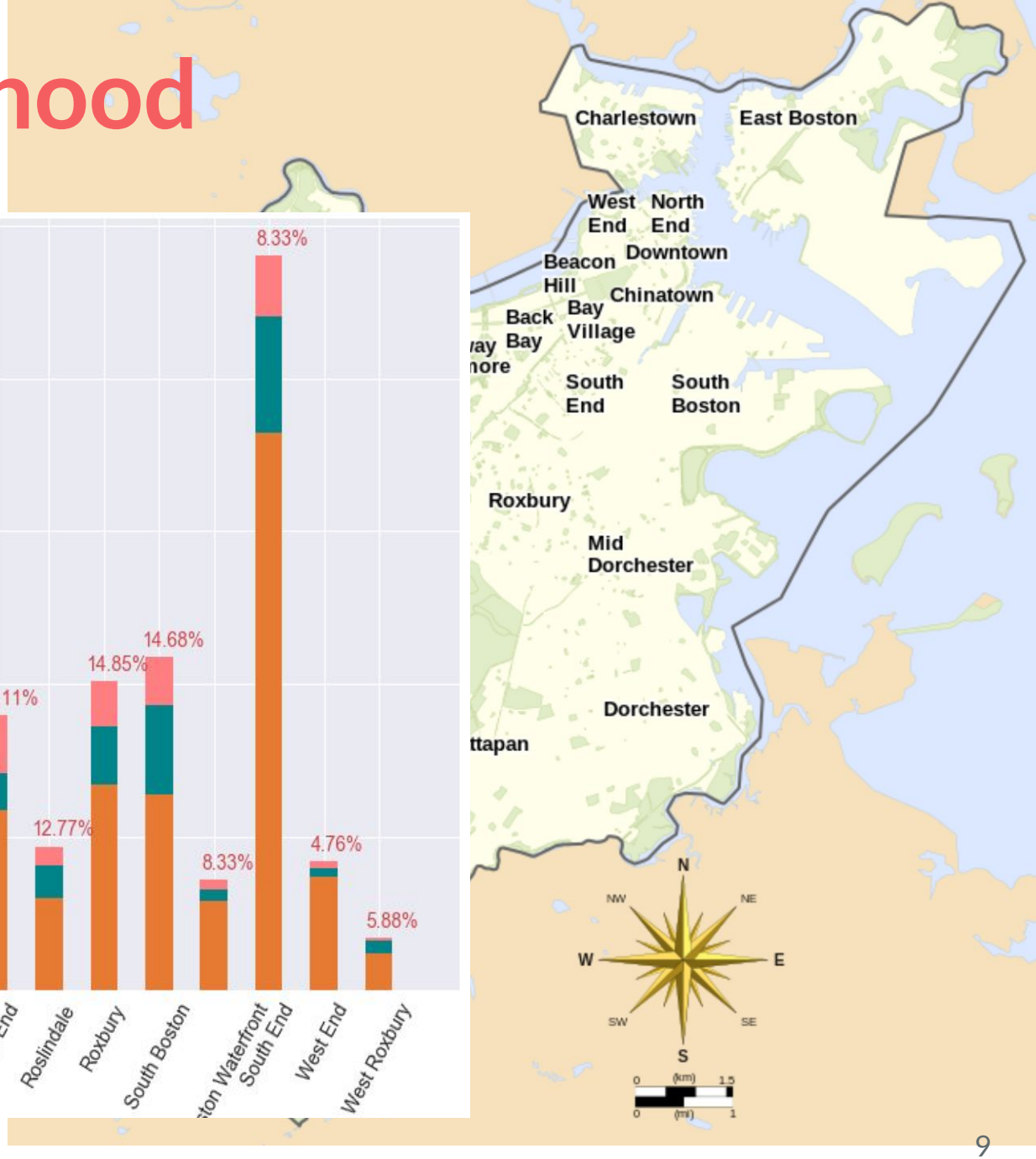
Data Combination

Input Variables

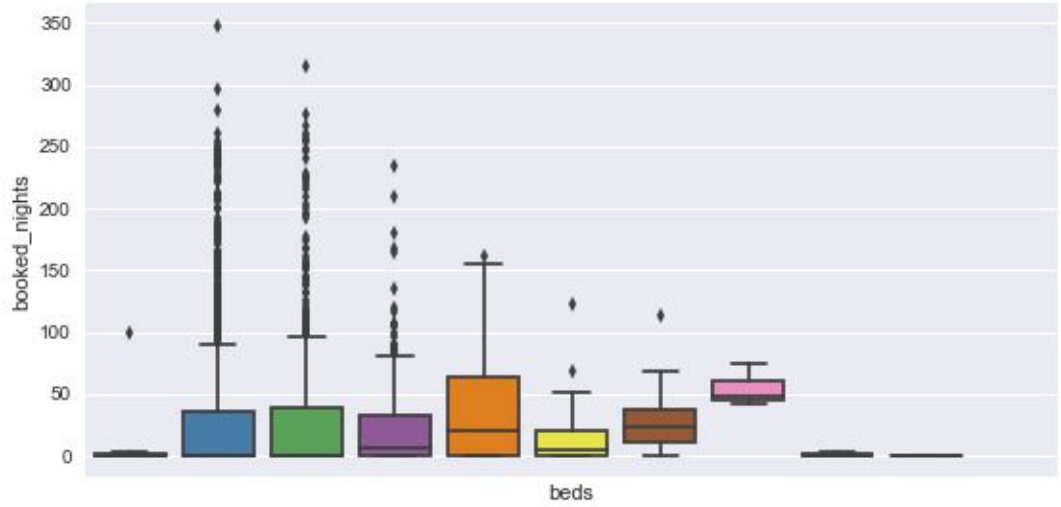
Business Type	Columns
Geo Locations	Neighbourhood
Host	Host weeks, Host Acceptance Rate, Host Response Rate
Listing Properties	Property Type, Room Type, Bed Type # of Bathroom, # of bed, # of accom, # of guest included
Review	Review number in 2016, Review Scores (review number should be excluded for new hosts prediction)
Fee	Price, Cleaning fee, Fee for Extra People
Reservation	Instant Bookable, Minimum Nights, Maximum Nights
Availability	Number of Available in 30 days, 60 days, 90 days, in peak season (July - Sept)

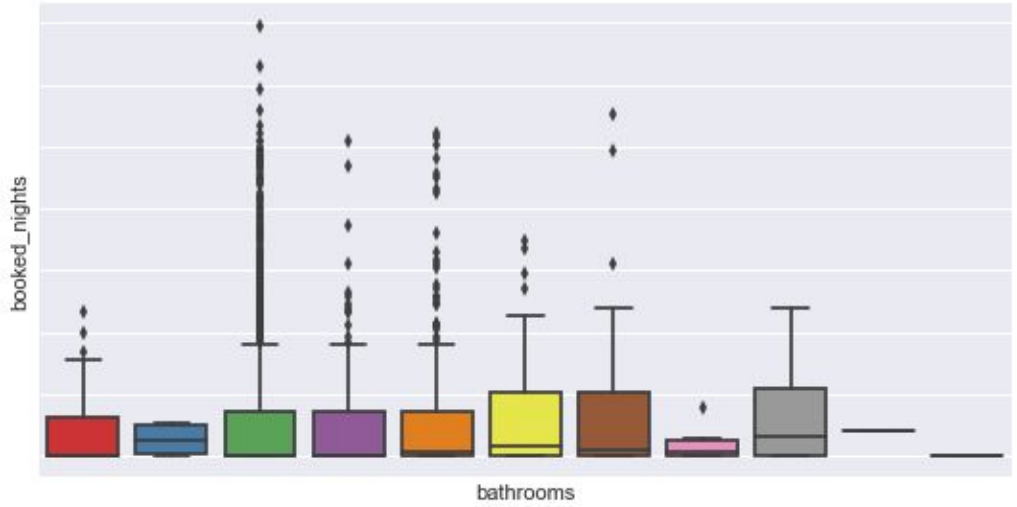
Bookings by Neighbourhood

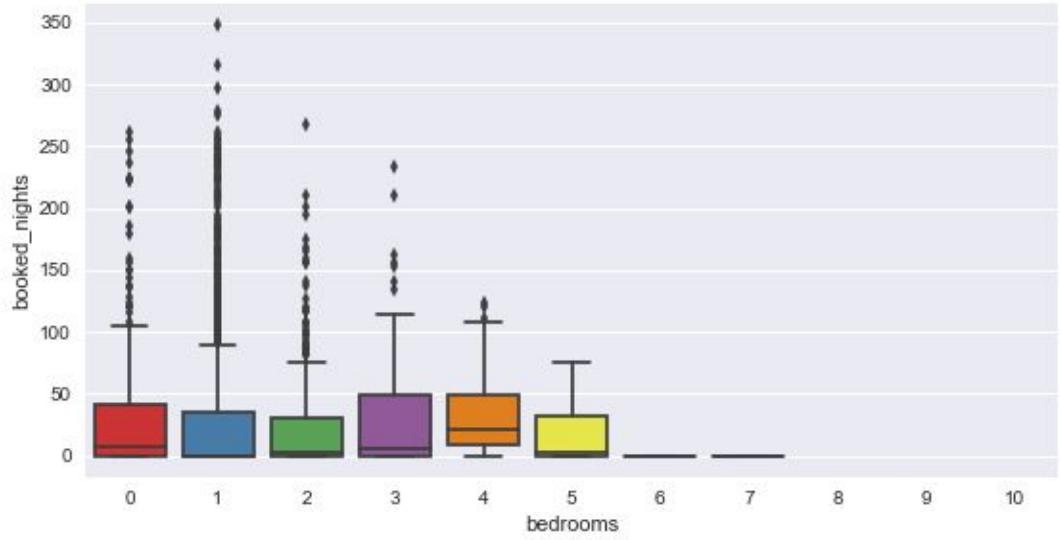


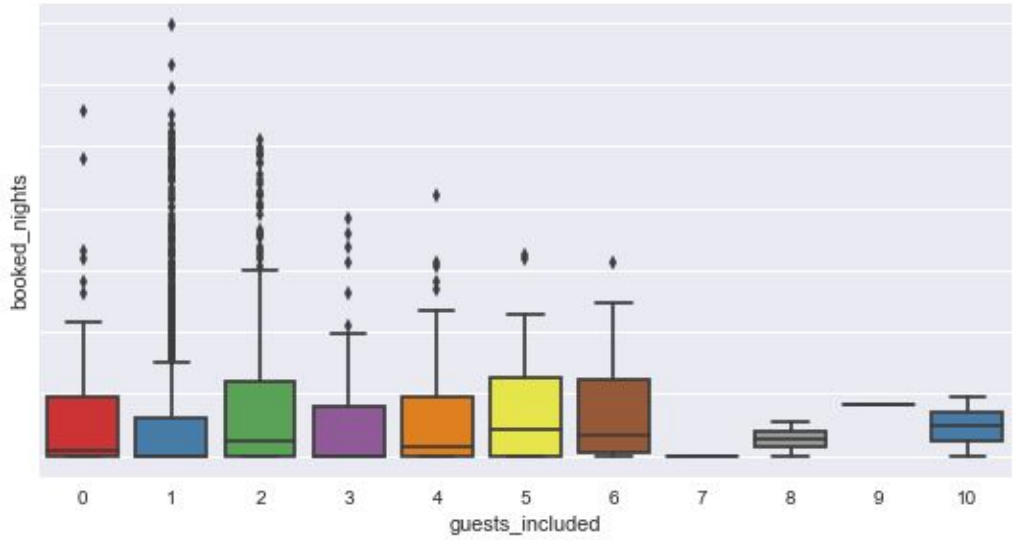


Bookings vs # of beds/bathrooms/bedrooms/guests

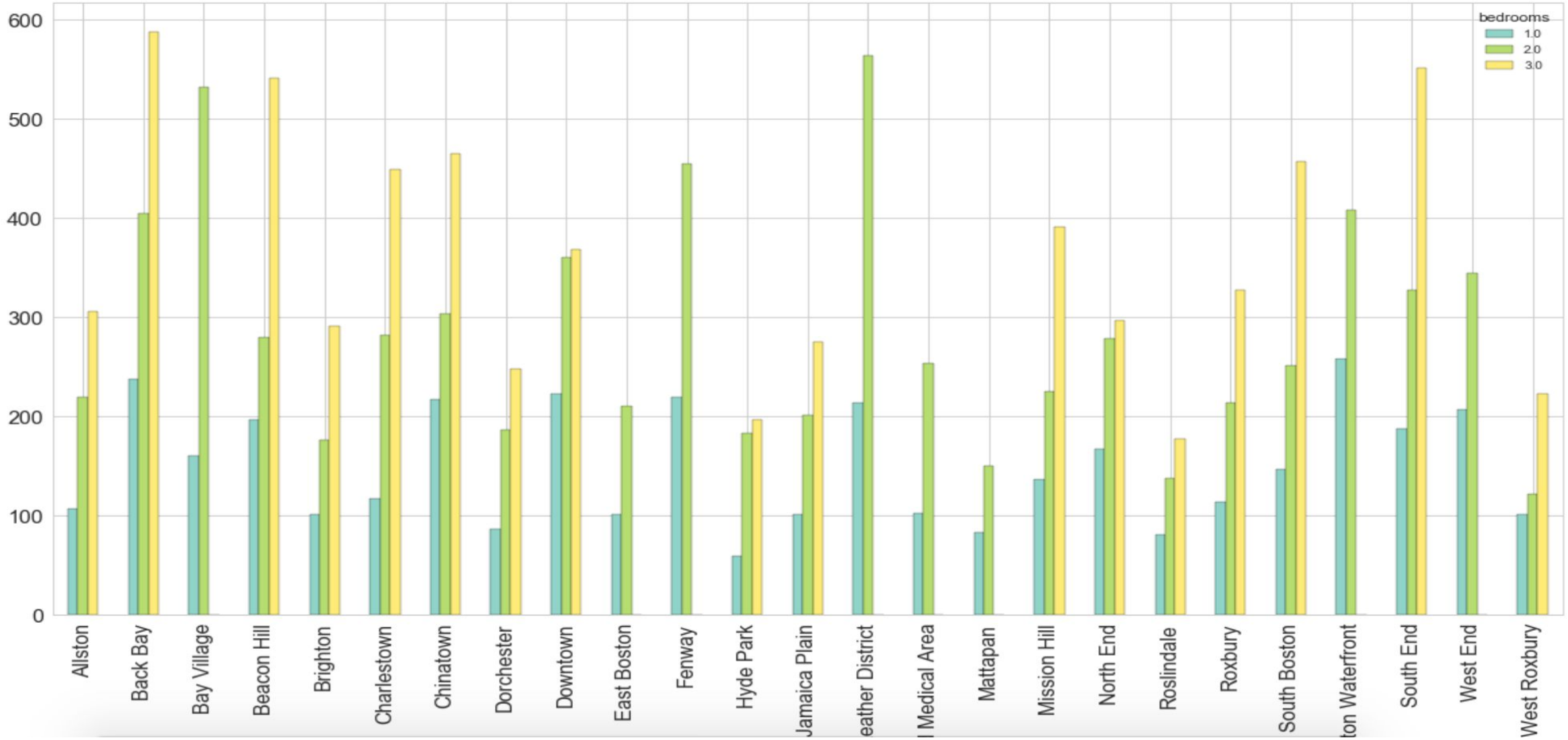




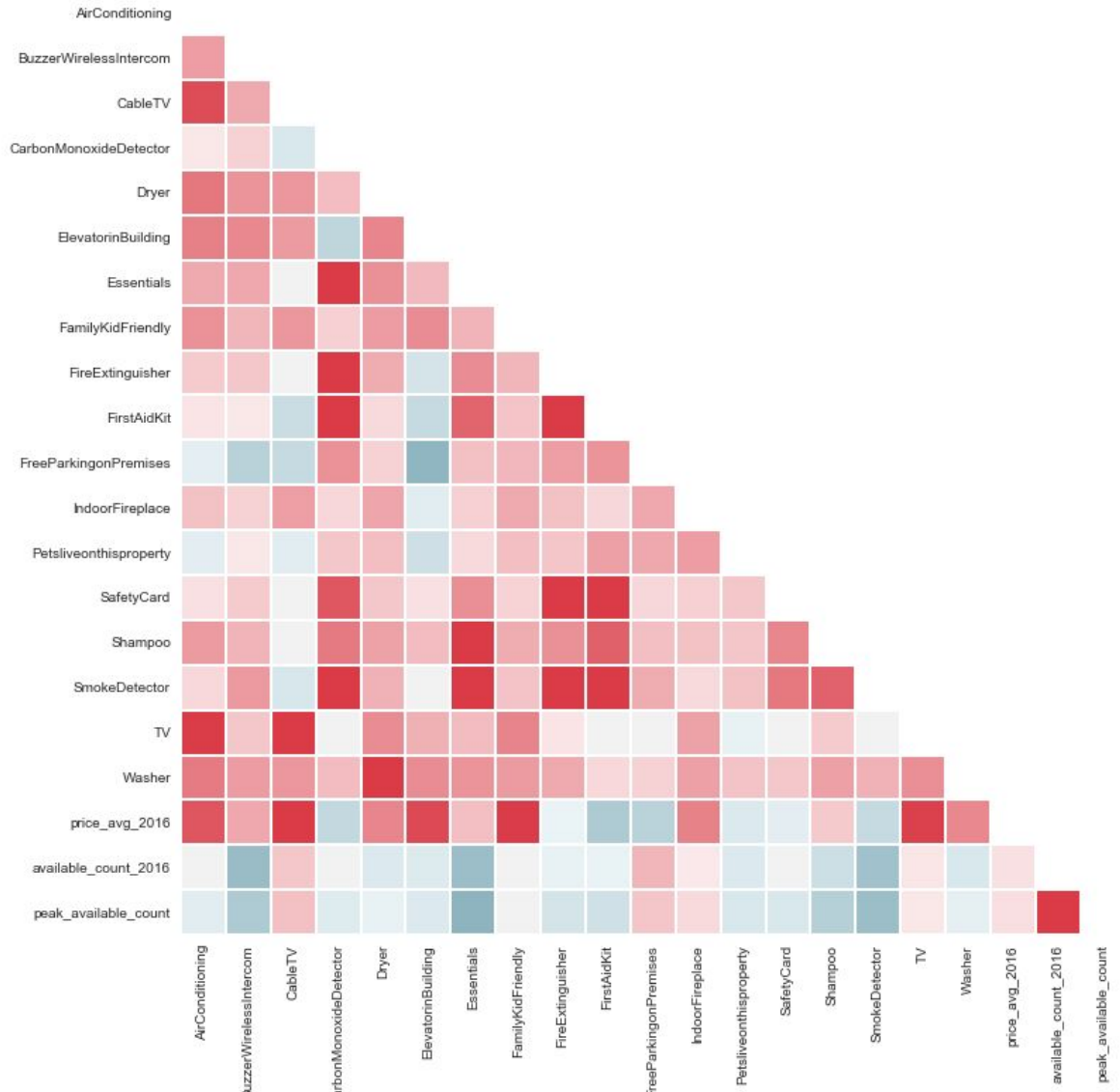


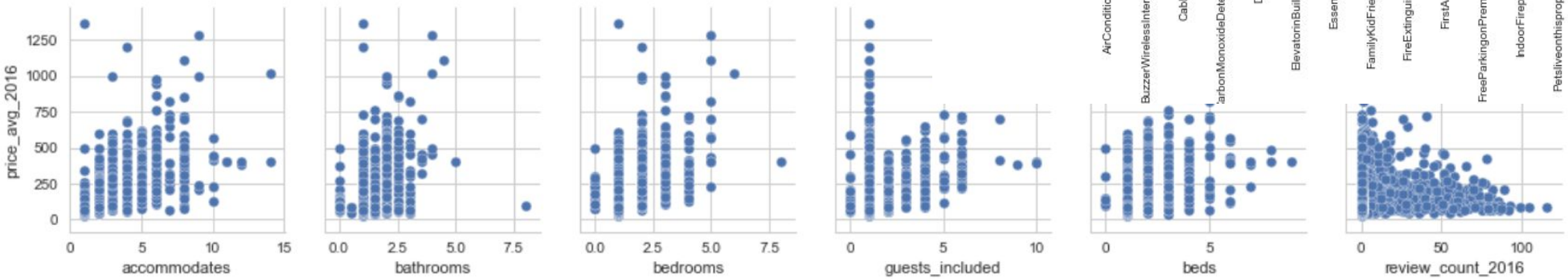


Price vs Bedrooms by Neighbourhood

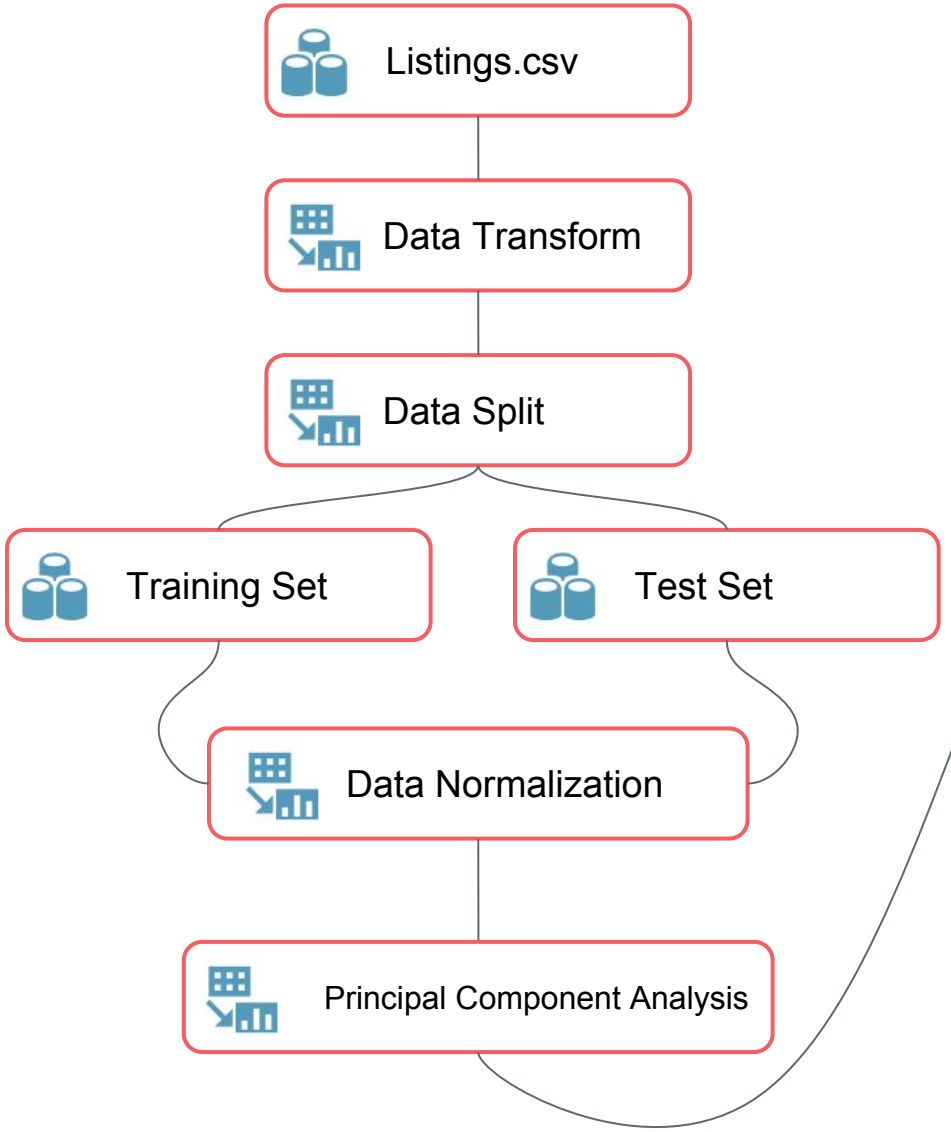


Amenities vs Price



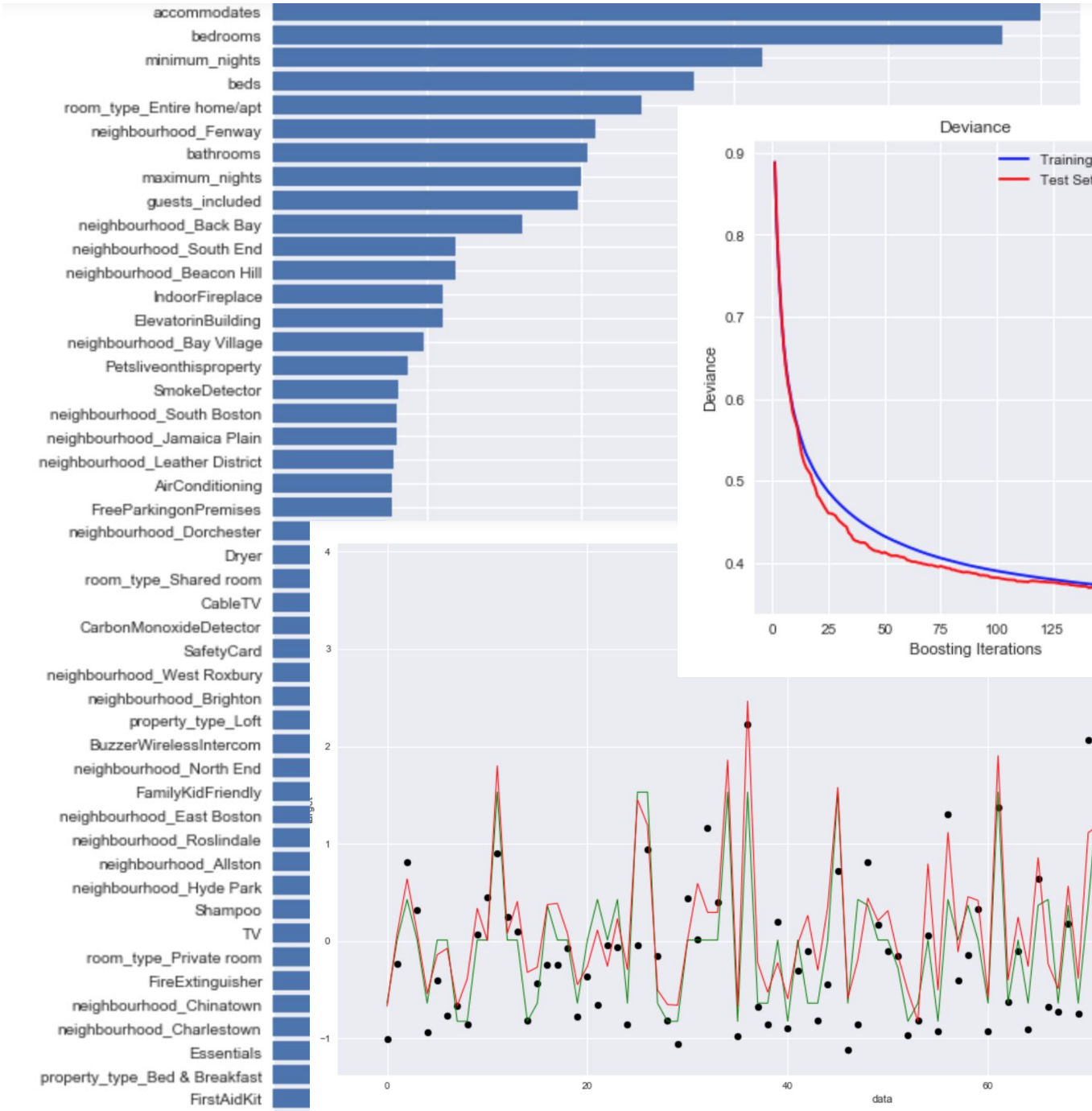


Pricing Regression Models



Stochastic Gradient Descent	MSE on training: 0.373 r^2 on training: 0.627 MSE on test: 0.347 r^2 on test: 0.620
ElasticNet	MSE on training: 0.372 r^2 on training: 0.628 MSE on test: 0.393 r^2 on test: 0.607
Decision Tree Regressor	MSE on training: 0.405 r^2 on training: 0.595 MSE on test: 0.417 r^2 on test: 0.544
Adaboost DTR	MSE on training: 0.125 r^2 on training: 0.875 MSE on test: 0.340 r^2 on test: 0.628
Neural Network	MSE on training: 0.226 r^2 on training: 0.774 MSE on test: 0.345 r^2 on test: 0.654
GradientBoostingRegressor	MSE on training: 0.364 r^2 on training: 0.636 MSE on test: 0.367 r^2 on test: 0.633

2 hidden layers, 50, 10 neurons activation rectified linear unit function stochastic gradient-based optimizer




```
Using Theano backend.

Train on 1603 samples, validate on 791 samples
Epoch 1/35
0s - loss: 0.6431 - val_loss: 0.3589
Epoch 2/35
0s - loss: 0.3820 - val_loss: 0.3320
Epoch 3/35
0s - loss: 0.3496 - val_loss: 0.3289
Epoch 4/35
0s - loss: 0.3422 - val_loss: 0.3456
Epoch 5/35
0s - loss: 0.3217 - val_loss: 0.3223
Epoch 6/35
0s - loss: 0.3014 - val_loss: 0.3383
Epoch 7/35
0s - loss: 0.3010 - val_loss: 0.3338
Epoch 8/35
0s - loss: 0.3097 - val_loss: 0.3384
Epoch 9/35
0s - loss: 0.2905 - val_loss: 0.3640
Epoch 10/35
0s - loss: 0.2689 - val_loss: 0.4152
Epoch 11/35
0s - loss: 0.2680 - val_loss: 0.4073
Epoch 12/35
0s - loss: 0.2662 - val_loss: 0.3563
Epoch 13/35
0s - loss: 0.2622 - val_loss: 0.3506
Epoch 14/35
0s - loss: 0.2533 - val_loss: 0.3532
Epoch 15/35
0s - loss: 0.2487 - val_loss: 0.3908
Epoch 16/35
0s - loss: 0.2343 - val_loss: 0.3361
Epoch 17/35
0s - loss: 0.2358 - val_loss: 0.3783
Epoch 18/35
0s - loss: 0.2367 - val_loss: 0.3767
Epoch 19/35
0s - loss: 0.2224 - val_loss: 0.3443
Epoch 20/35
0s - loss: 0.2240 - val_loss: 0.3603
```

Is your listing popular?

Number of Reviews=
More people have stayed=
Popularity !

"80% of hosts leave a review for their guests.
72% of guests leave a review for hosts."

- Count number of reviews by listing_id
- Assign 1 to listings which number of reviews are at 90 percentile and above in 2016. (Popular)
- Assign 0 to the listings which number of reviews are below 90 percentile in 2016. (Not Popular)

Is your listing popular? Model Selection

70% Training
30% Testing

```
Algo: Perceptron              and Score: 0.896
Algo: LogisticRegression      and Score: 0.898
Algo: Decision Tree           and Score: 0.906
Algo: Random Forest           and Score: 0.898
Algo: AdaBoost                and Score: 0.939
Algo: Neural Net              and Score: 0.877
Algo: Naive Bayes             and Score: 0.896
Algo: Nearest Neighbors       and Score: 0.887
```

```
Algo: Perceptron              and Score: 0.785
Algo: LogisticRegression      and Score: 0.900
Algo: Decision Tree           and Score: 0.790
Algo: Random Forest           and Score: 0.900
Algo: AdaBoost                and Score: 0.835
Algo: Neural Net              and Score: 0.788
Algo: Naive Bayes             and Score: 0.900
Algo: Nearest Neighbors       and Score: 0.658
```

K-Fold Validation, cv=10

Is your listing popular? Model Selection

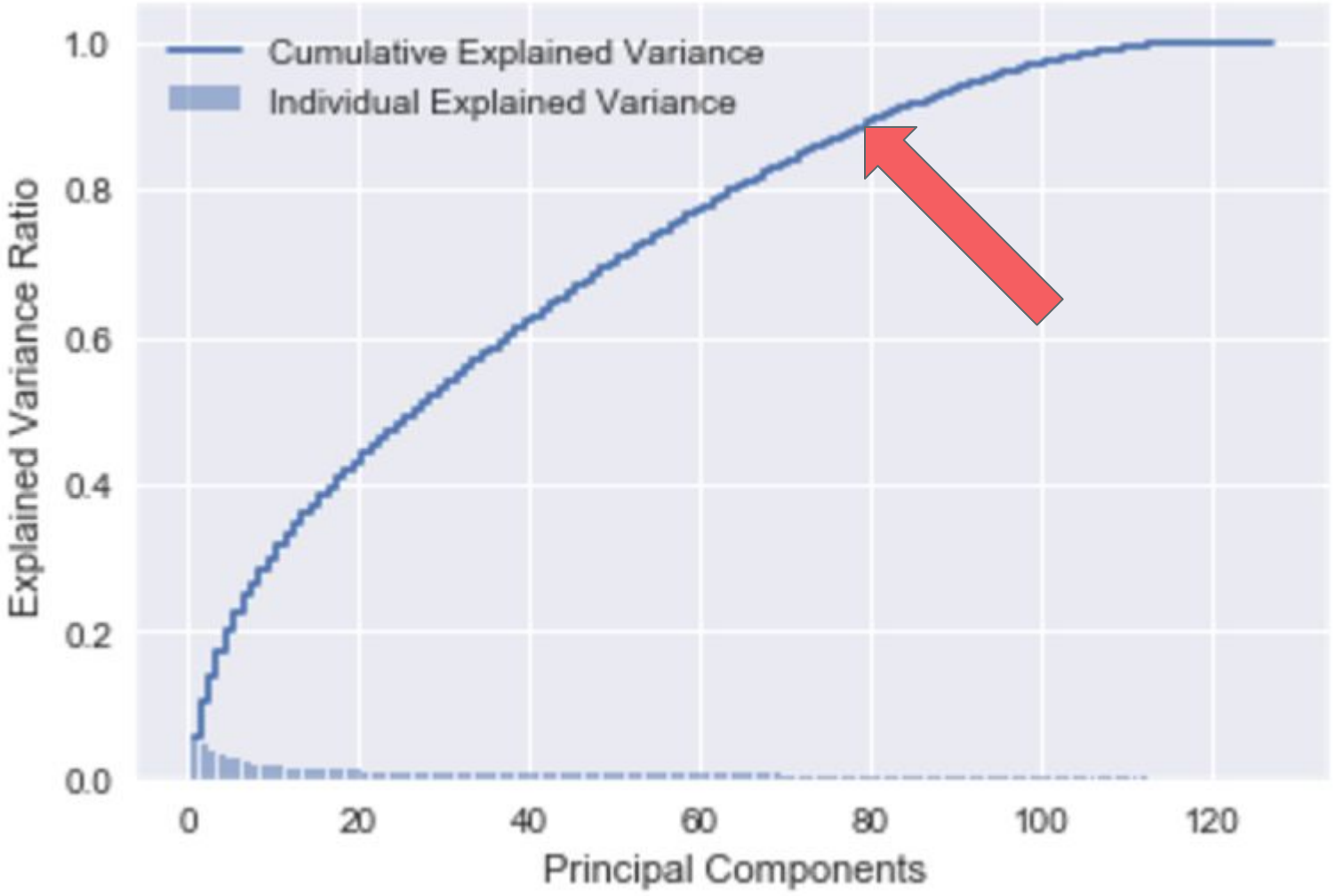
Standardization

```
Algo: Perceptron              and Score: 0.893
Algo: LogisticRegression      and Score: 0.925
Algo: Decision Tree           and Score: 0.910
Algo: Random Forest           and Score: 0.898
Algo: AdaBoost                and Score: 0.939
Algo: Neural Net              and Score: 0.924
Algo: Naive Bayes             and Score: 0.242
Algo: Nearest Neighbors       and Score: 0.898
```

PCA and Standardization

```
Algo: Perceptron              and Score: 0.893
Algo: LogisticRegression      and Score: 0.925
Algo: Decision Tree           and Score: 0.882
Algo: Random Forest           and Score: 0.898
Algo: AdaBoost                and Score: 0.909
Algo: Neural Net              and Score: 0.928
Algo: Naive Bayes             and Score: 0.814
Algo: Nearest Neighbors       and Score: 0.898
```

PCA



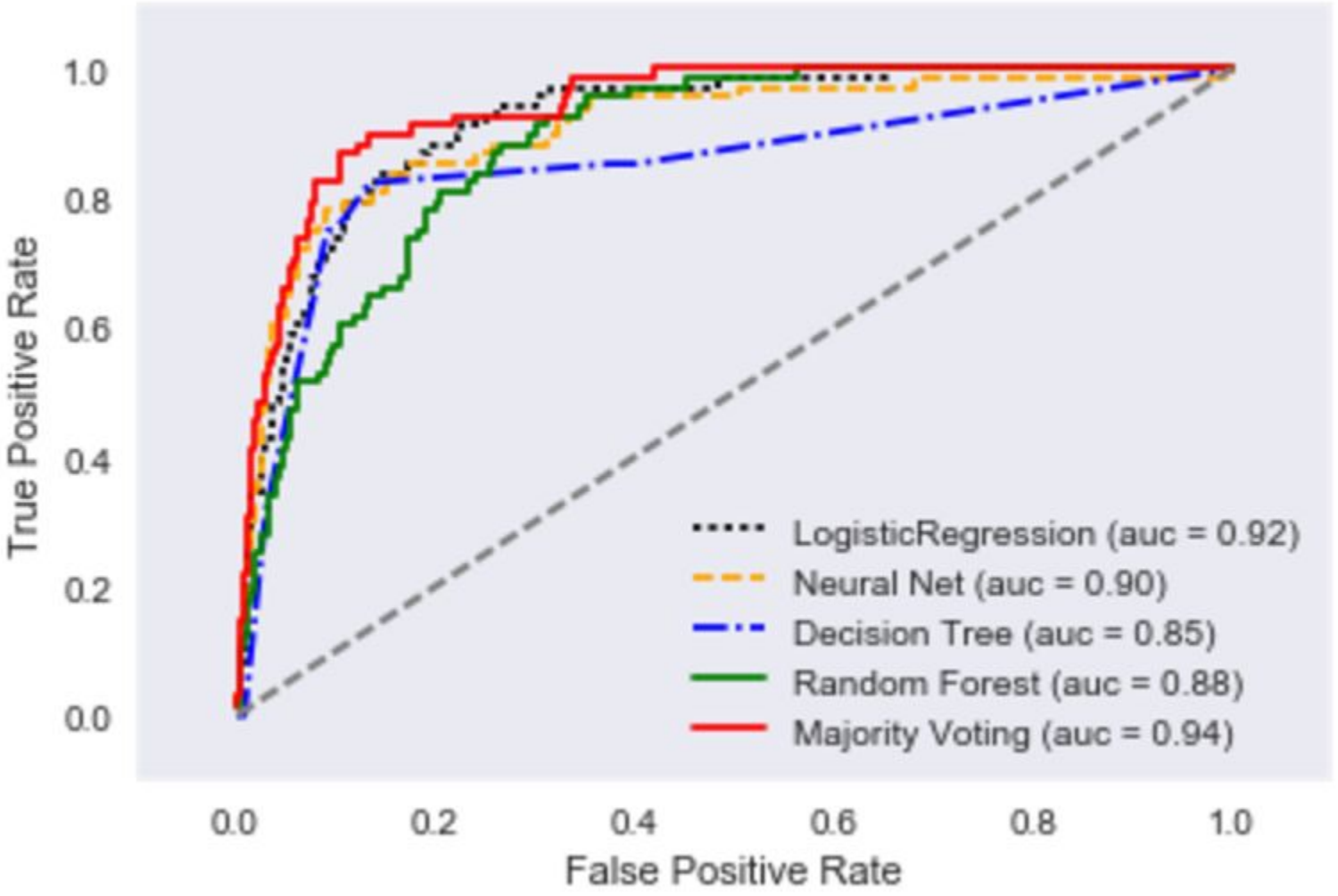
- It is a complex model: no single feature could explain the majority of variance in this model.
- We chose 80 features to be included in our model, which explain about 90% variance in the dataset

Majority Voting

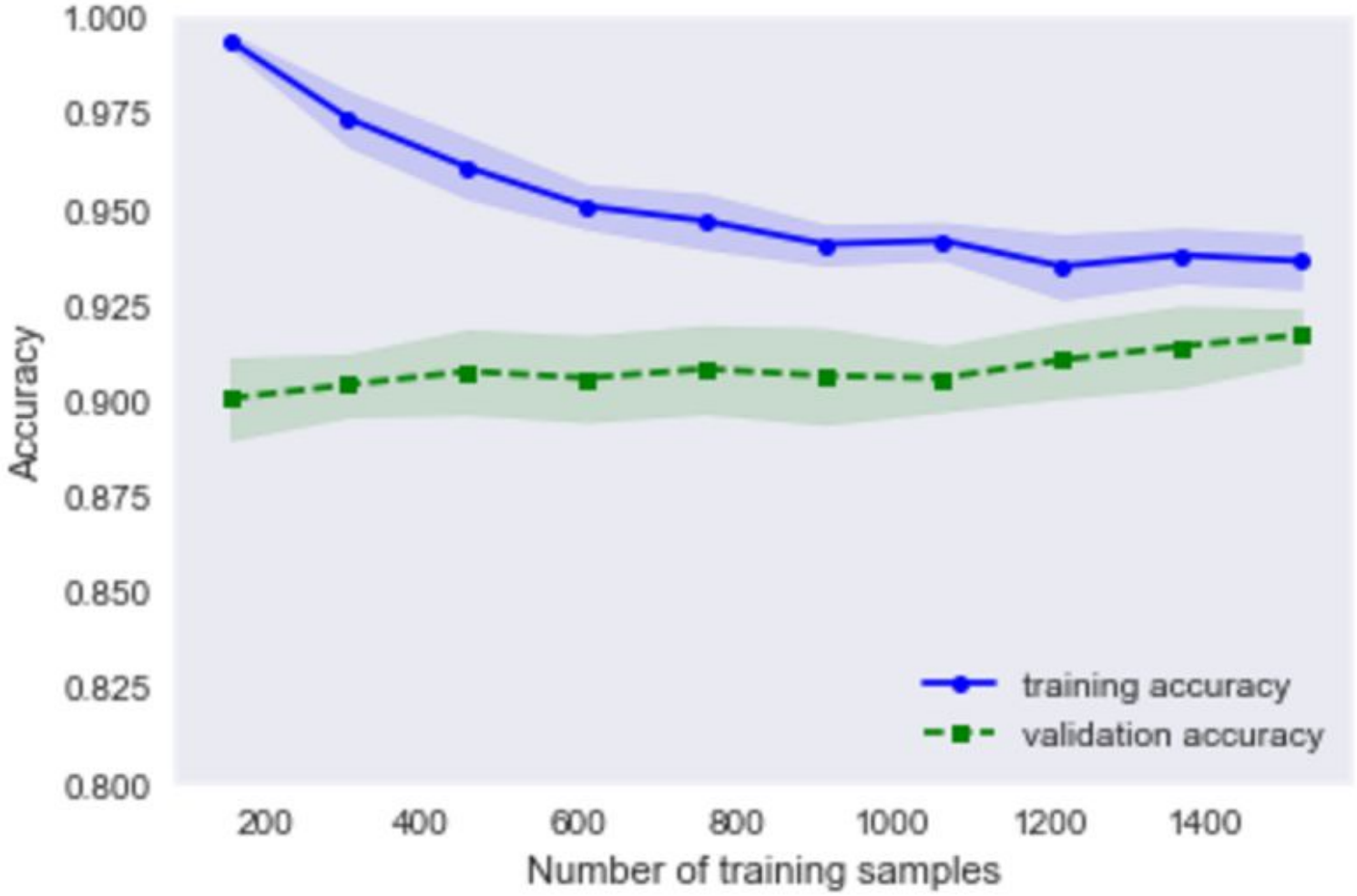
```
LogisticRegression:
Accuracy: 0.917
Precision: 0.571
Recall: 0.471
F1 Score: 0.516
ROC_AUC score: 0.717
================================================
Neural Net:
Accuracy: 0.916
Precision: 0.556
Recall: 0.515
F1 Score: 0.534
ROC_AUC score: 0.736
================================================
```

```
Decision Tree:
Accuracy: 0.912
Precision: 0.553
Recall: 0.309
F1 Score: 0.396
ROC_AUC score: 0.641
================================================
Random Forest:
Accuracy: 0.906
Precision: 0.000
Recall: 0.000
F1 Score: 0.000
ROC_AUC score: 0.500
================================================
Majority Voting:
Accuracy: 0.919
Precision: 0.714
Recall: 0.221
F1 Score: 0.337
ROC_AUC score: 0.606
================================================
```

ROC_AUC Curve



Learning Rate Curve-Majority Voting



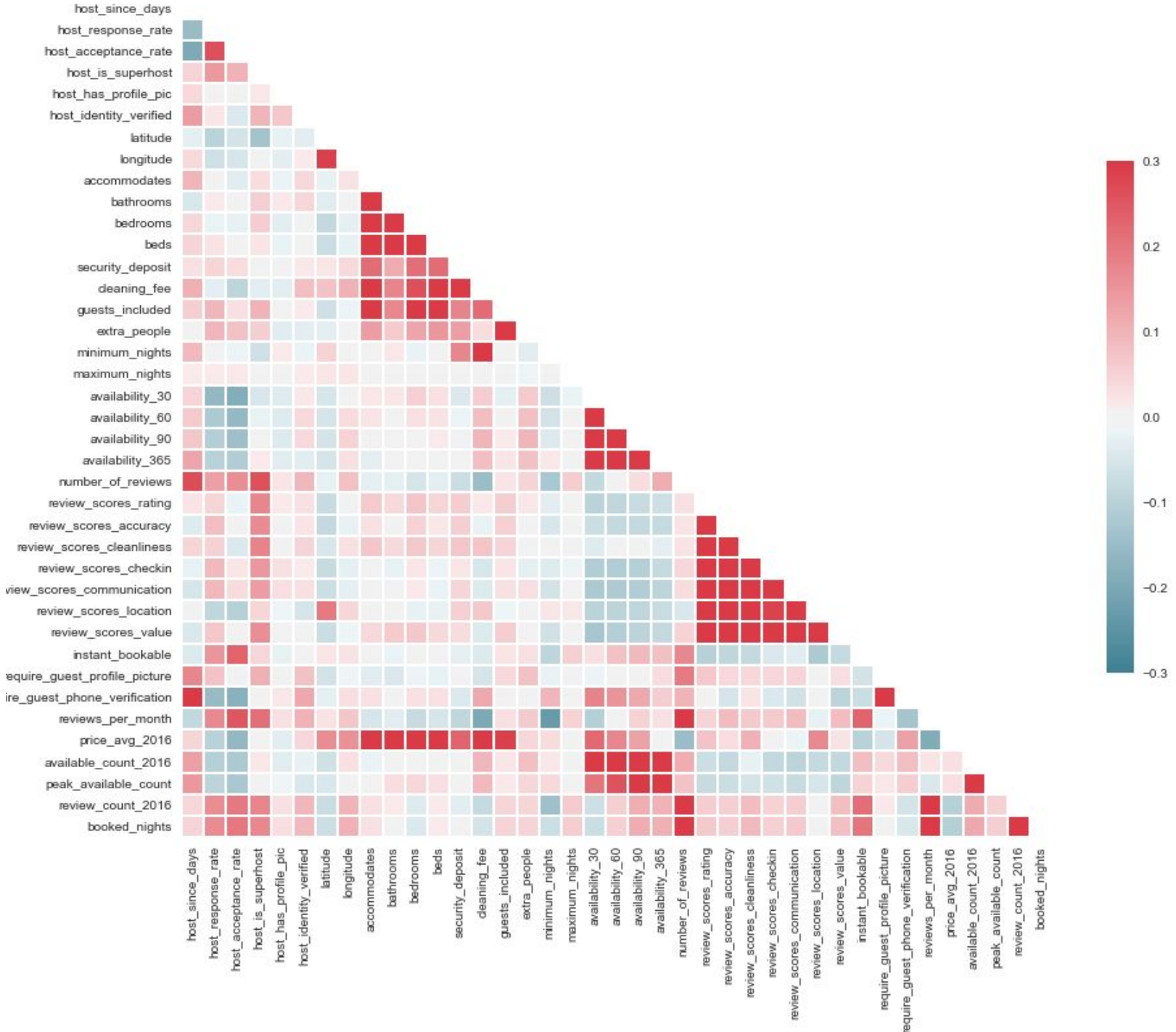
Hyper-Parameter Tuning

```
print('Best parameters: %s' % gs.best_params_)
```

```
Best parameters: {'pipeline-2_clf__C': 1.0, 'pipeline-2_clf__penalty': 'l2', 'pipeline-3_clf__alpha': 0.1, 'pipeline-3_clf__hidden_layer_sizes': (100, 100)}
```

```
print('Accuracy: %.2f' % gs.best_score_)
```

```
Accuracy: 0.94
```



Covariance Matrix: Which Features are important?

How to determine booked nights

- A **review rate** of **70%** for the number of guests making a booking who leave a review
- An average booking of **3 nights** or a higher minimum nights configured for a listing.

Rental Multi-class Classification

Estimated Rental = MIN (Average Stay, Minimum Nights Required) * Number of Reviews/ 0.7

"80% of hosts leave a review for their guests.
72% of guests leave a review for hosts."

For popular 1b listings whose review numbers are above 90 percentile, the estimated earning will be

Estimated Rental (147) * Avg Price (120) /12 = 1470 per month.

For popular 2b listings whose review numbers are above 90 percentile, the estimated earning will be

Estimated Rental (147) * Avg Price (263) /12 = 3200 per month.

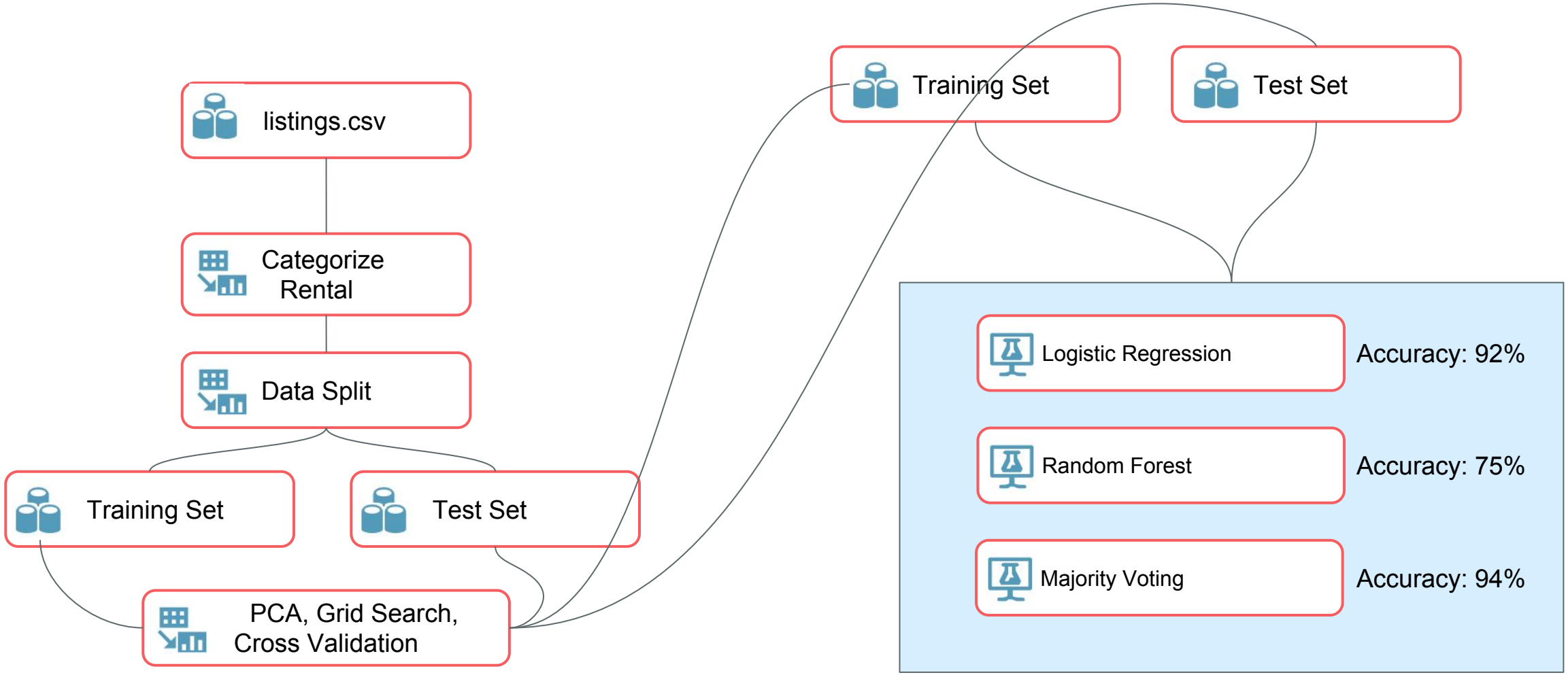
Next: Use Multi-class Classification to get more accurate estimation.

Estimated Rental =< 75 Percentile

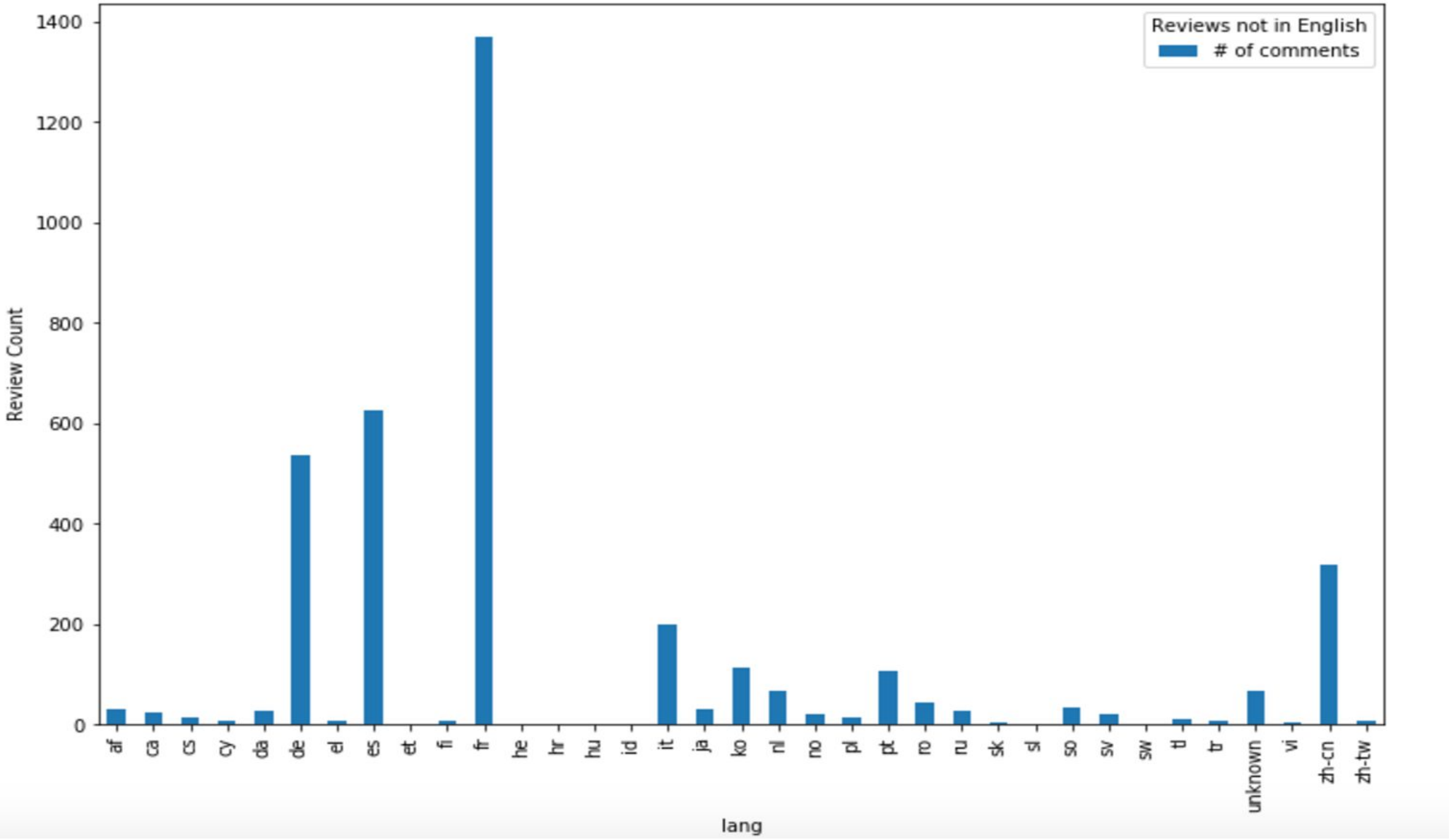
75 Percentile < Estimated Rental < 95 Percentile

Estimated Rental >= 95 Percentile

Rental Multi-class Classification



Sentiment Analysis-Language in Reviews

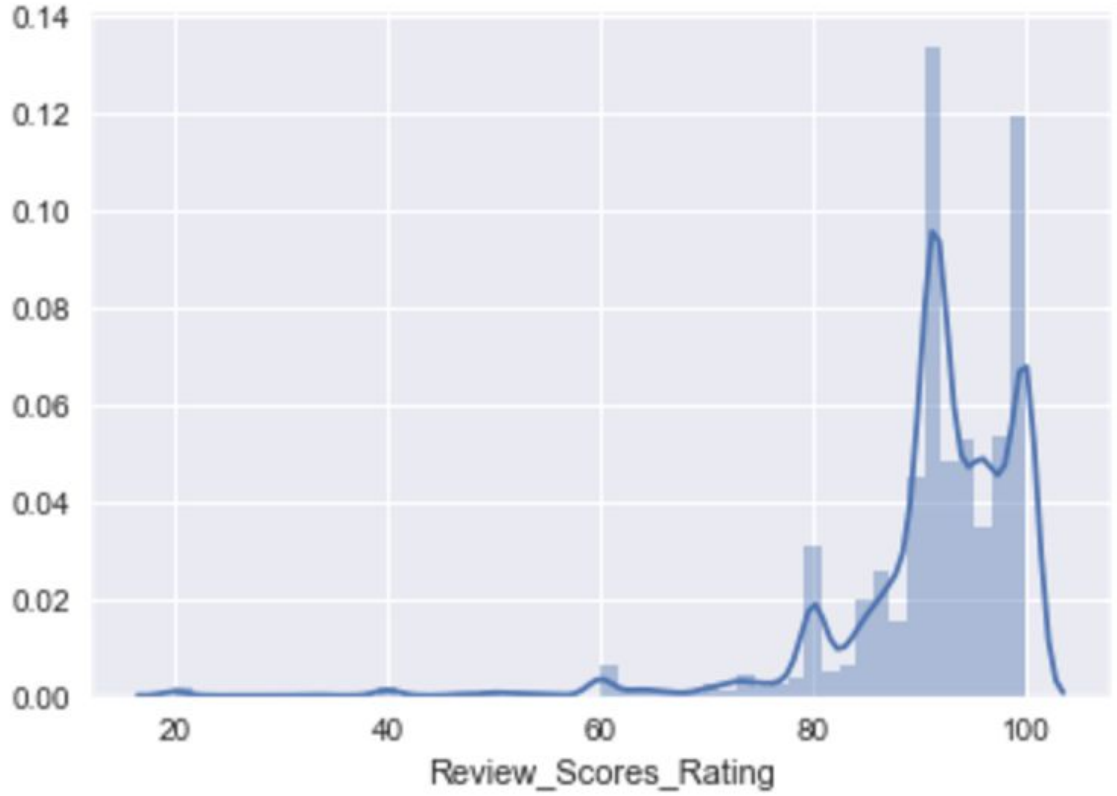


Sensitive Analysis-Host Description

Sentiment Analysis-Reviews and Review Scores

- Study the relationship between reviews that guests left and review scores to see if review scores could really represent/reflect the sentiments of guest reviews

	review_scores_rating
count	49287.000000
mean	92.046627
std	6.302865
min	20.000000
25%	88.000000
50%	93.000000
75%	96.000000
max	100.000000



Can Reviews really reflect how people would rate the listings/hosts?

Review 1

It was a nice experience, the location is convenient as it was close to the subway station. My mom like the kitchen and the room is pretty comfortable

Review 3

The room is very nice, just as described, clean and beautifully decorated. Kaitlyn is very supportive and punctual in replying messages. When I visited Boston, she wasn't here but she always texted me to make sure everything is fine and good. Overall speaking, I highly recommend her place due to the hospitality of Kaitlyn, the convenience of the place to the subway station (5 mins walk~), and the homely feeling this room has.

Review 2

Thanks Kaitlin for being such a great host! Makes my first Airbnb experience a positive one! The place is near to the train station, easy to get around, and for any animal-lovers out there, there's a zoo just a few miles away - not that you can see any wild zebras in the 'hood though! I did not visit the zoo; just saw it on the map and signs at the train station! =)

No, reviews cannot reflect review scores!

Sentiment Analysis-Reviews and Review Scores

- In Review Scores Rating column, 5% Percentile is 78; 95% Percentile is 100
- Then map ‘Review_Scores_Rating’ column into three ranges:
  - If ‘Review_Scores_Rating’ <=78, assign -1 (Bad)
  - If ‘Review_Scores_Rating’ >78 & <100, assign 0 (Good)
  - Else, Assign 1 (Excellent)

Review 1

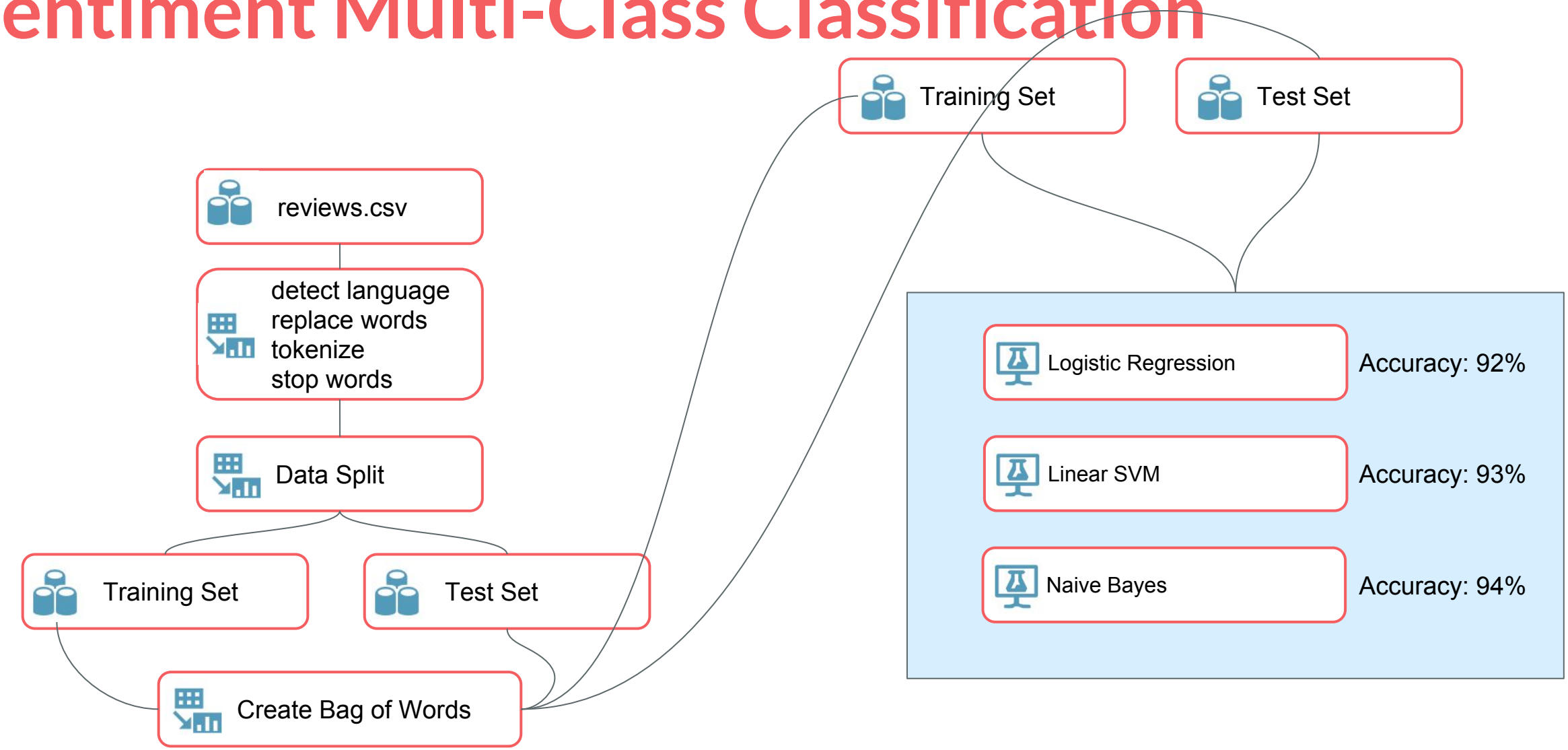
Kaitlin welcomed me in her home and made sure I have what I needed. She even borrowed me her hair dryer. The room was fine and I really liked the kitchen as well. The only thing is that there is no heating in the room but Kaitlin helped with that as well. Also, the house is in a convenient and quiet neighbourhood, very close to an orange line station.One of the things I didn't like the most is the fact that the room is on the third floor while the bathroom is on the first floor. The stairs you have to climb are quite old, slippery and very noisy. The other thing is that the bathroom wasn't clean all the time, only on the day the cleaner came by.

Review 2

I stay two nights there, the severals hosts are welcoming. But the dirt of the place make me very uncomfortable.'

Sentiment Analysis-Customer Reviews

Sentiment Multi-Class Classification



Summary

Find Candidates

Predict Review Score

Predict Price

Predict Rental Class

Calculate ROI

Don’t forget to regularize candidate data before prediction!


Buy or Rent?
North End
2 Bedrooms + 2 Bathrooms
Harvard University
Gyu-Kaku Japanese BBQ
Cambridge St
EAST CAMBRIDGE
H Mart
River St
Main St
Massachusetts General Hospital
TD Garden
Pier 6
EAST BOSTON
JEFFRIES POINT
NORTH END
Boston Logan International Airport
Boston
Massachusetts Institute of Technology
Brookline St
DCR's Hatch Memorial Shell
Boston Common
Zipcar Boston
The Institute Of Contemporary Art
Storrow Dr
BACK BAY
Dexter Park
Urban Outfitters
Blue Hills Bank Pavilion
Boston Convention and Exhibition Center
Summer St
Museum of Fine Arts, Boston
Tremont St
Northeastern University
SOUTH BOSTON
Brookline
E 8th St
William J Day Blvd
South Bay Center
Centre St
Spectacle Island

Renting?

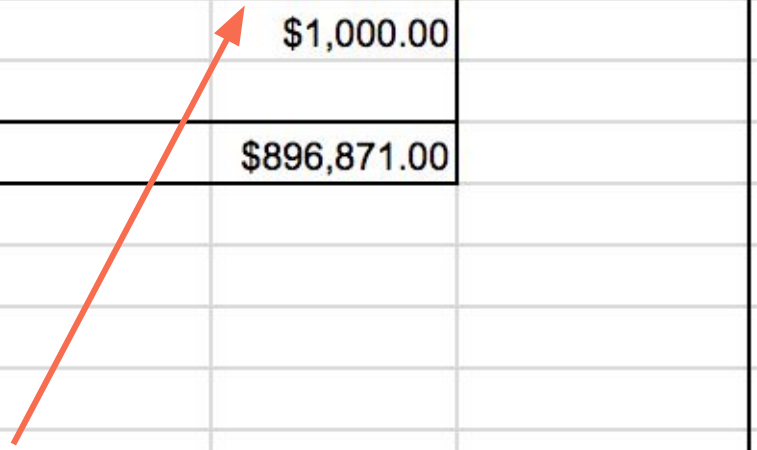
	2 Bedroom + 2 Bathroom	
RENTING	**North End: Christopher Columbus Plaza**	
First Month		
Revenue		
Expenses		
	Upfront Expenses	
	Security Deposit	$2,750.00
	Application Fee	$300.00
	Furniture	$1,000.00
Sum		$4,050.00

Following Each Month		
Revenue		$3,200.00
Expenses		
	Variable and Fixed Costs	
	Rent	$2,750.00
	Cable/ Internet	$70.00
	Renter's Insurance	$11.67
	Parking	$325.00
	Administration Fee	$29.00
	Sewer	$8.00
	Trash	$0.00
	Gas	$0.00
	Electricity	$0.00
Sum		$6.33

Buying?

BUYING	North End: Christopher Columbus Plaza	
First Month		
Revenue		
Expenses		
	Upfront Expenses	
	Buying	$895,871.00
	Furniture	$1,000.00
Sum		$896,871.00

Following Each Month		
Revenue		$3,200.00
Expenses		
	Variable and Fixed Costs	
	Cable/ Internet	$70.00
	Renter's Insurance	$11.67
	Parking	$325.00
	Administration Fee	$29.00
	Sewer	$8.00
	Trash	$14.00
	Gas	$60.00
	Electricity	$150.00
Sum		$2,532.33
		29.51



Property increasing 4% next year

Could you guide me to find properties?

How about San Francisco?

How about Seattle?

Could you make me a website?

Take a look at San Francisco

THANK YOU