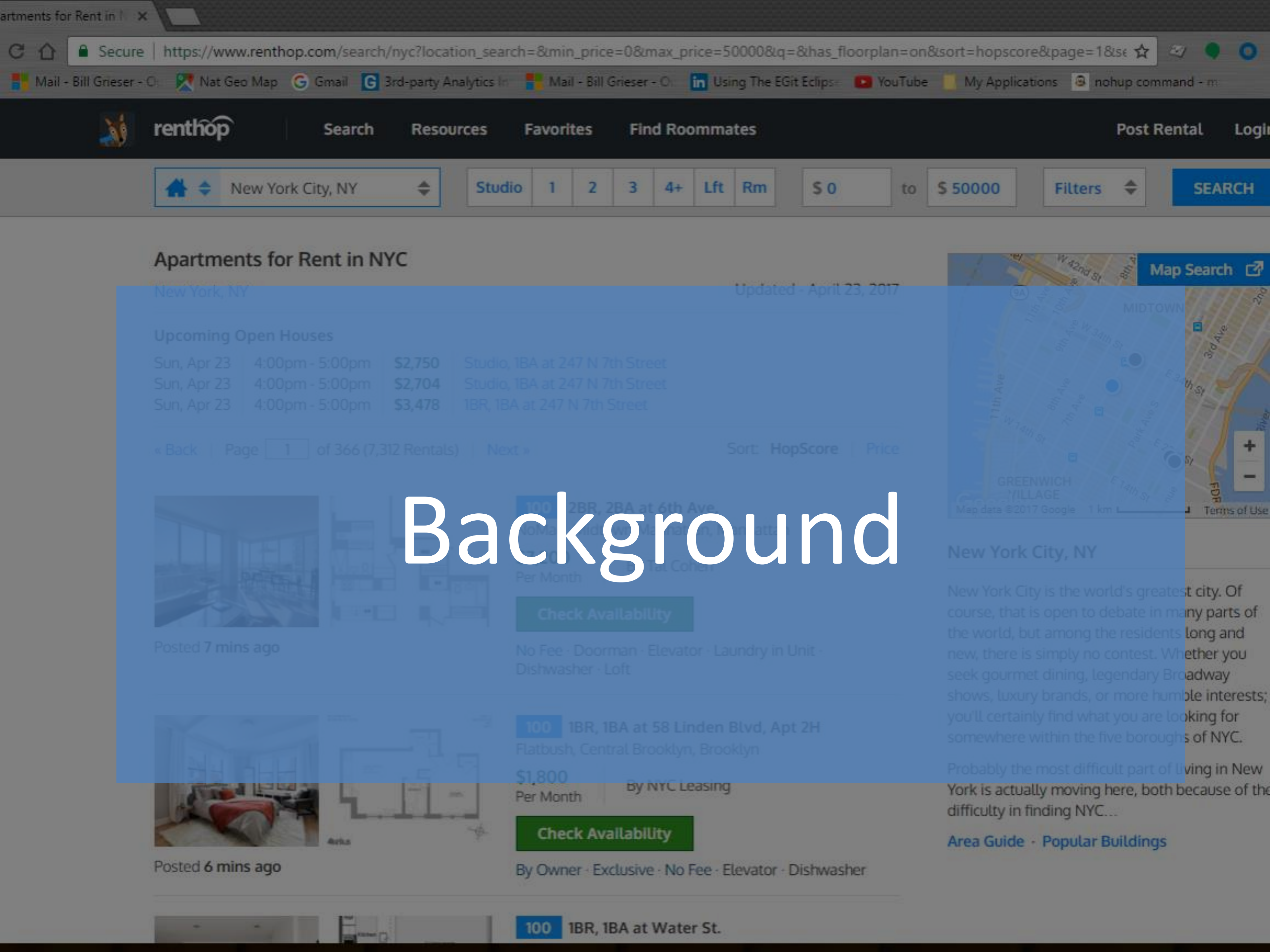
Factors Influencing Renthop.com Apartment Listing Interest

Bill Grieser
The George Washington University
Machine Learning I, DATS 6202
Final Assignment, Spring 2017

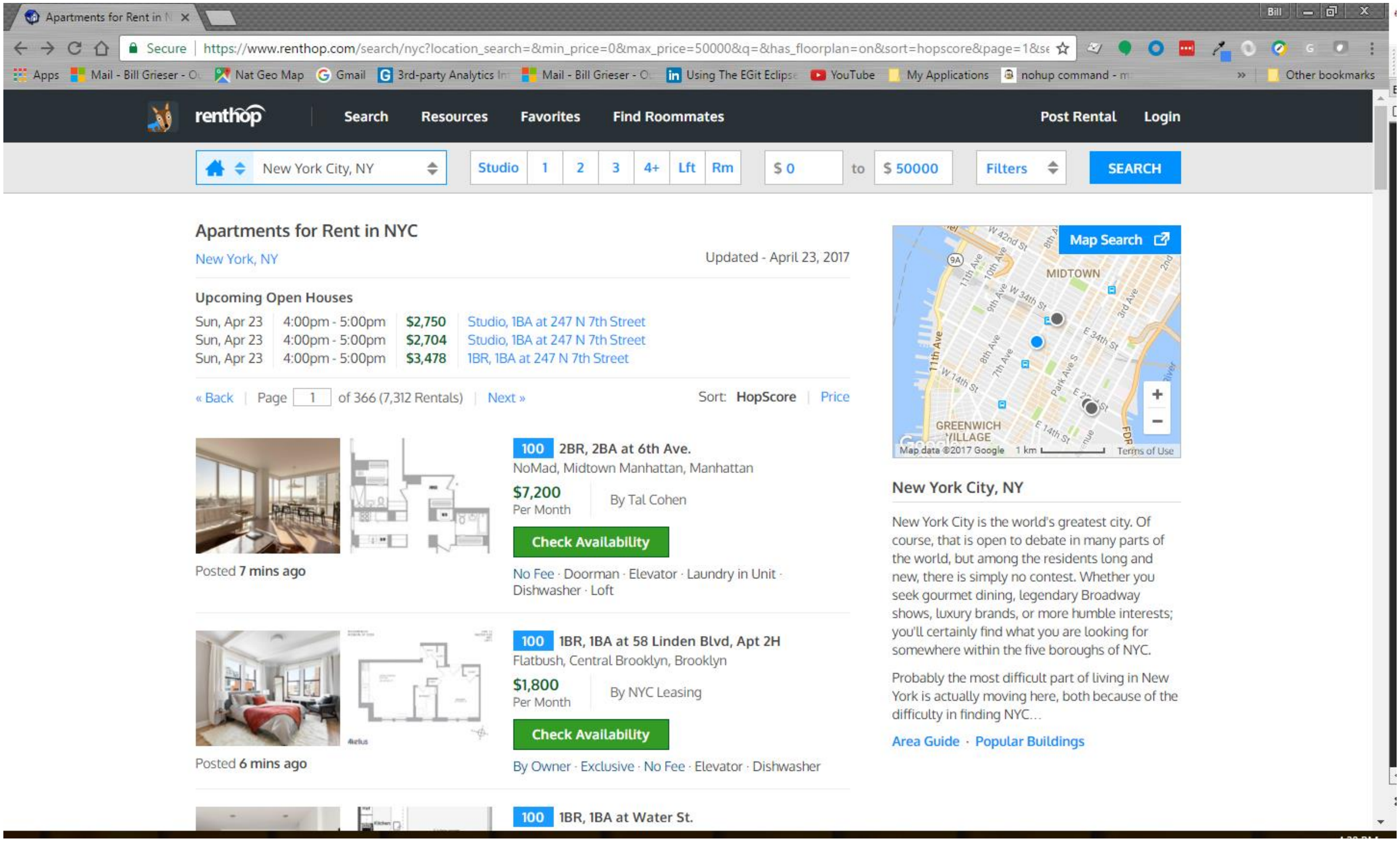
This Project

- Renthop is an apartment listing website, matching renters with available apartments and their rental agents
- They sponsored a Kaggle competition: given the information for a set of listings on the web site for NYC, **predict if there would be High, Medium, or Low interest in the apartment**
- That is what this project undertakes

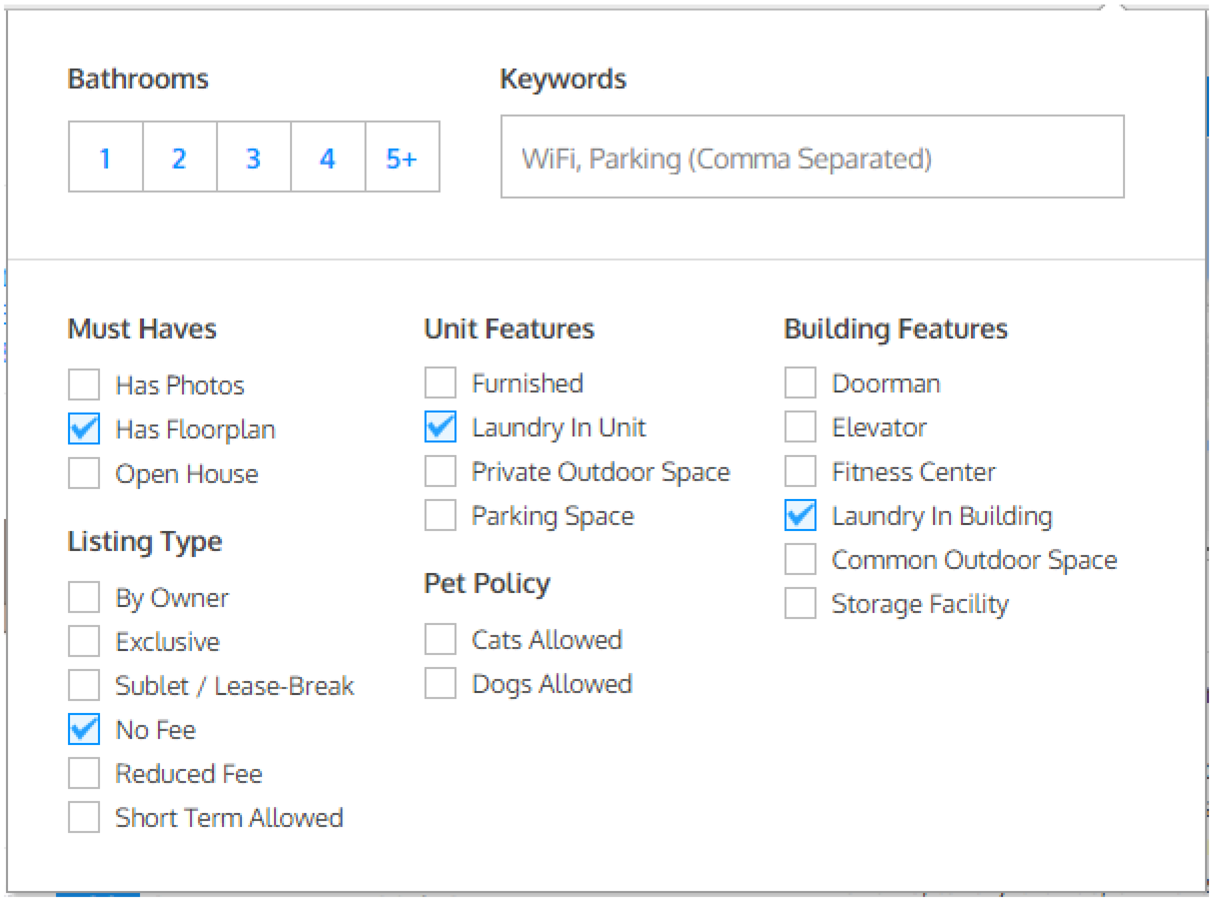
Background


https://www.renthop.com/search/nyc?location_search=&min_price=0&max_price=50000&q=&has_floorplan=on&sort=hopscore&page=1&se
renthop
Search
Resources
Favorites
Find Roommates
Post Rental
New York City, NY
Studio
1
2
3
4+
Lft
Rm
$ 0
to
$ 50000
Filters
SEARCH
Apartments for Rent in NYC
Map Search
New York City, NY
Probably the most difficult part of living in New York is actually moving here, both because of the difficulty in finding NYC...
Area Guide · Popular Buildings
$1,800
Per Month
By NYC Leasing
Check Availability
Posted 6 mins ago
By Owner · Exclusive · No Fee · Elevator · Dishwasher
100
1BR, 1BA at Water St.

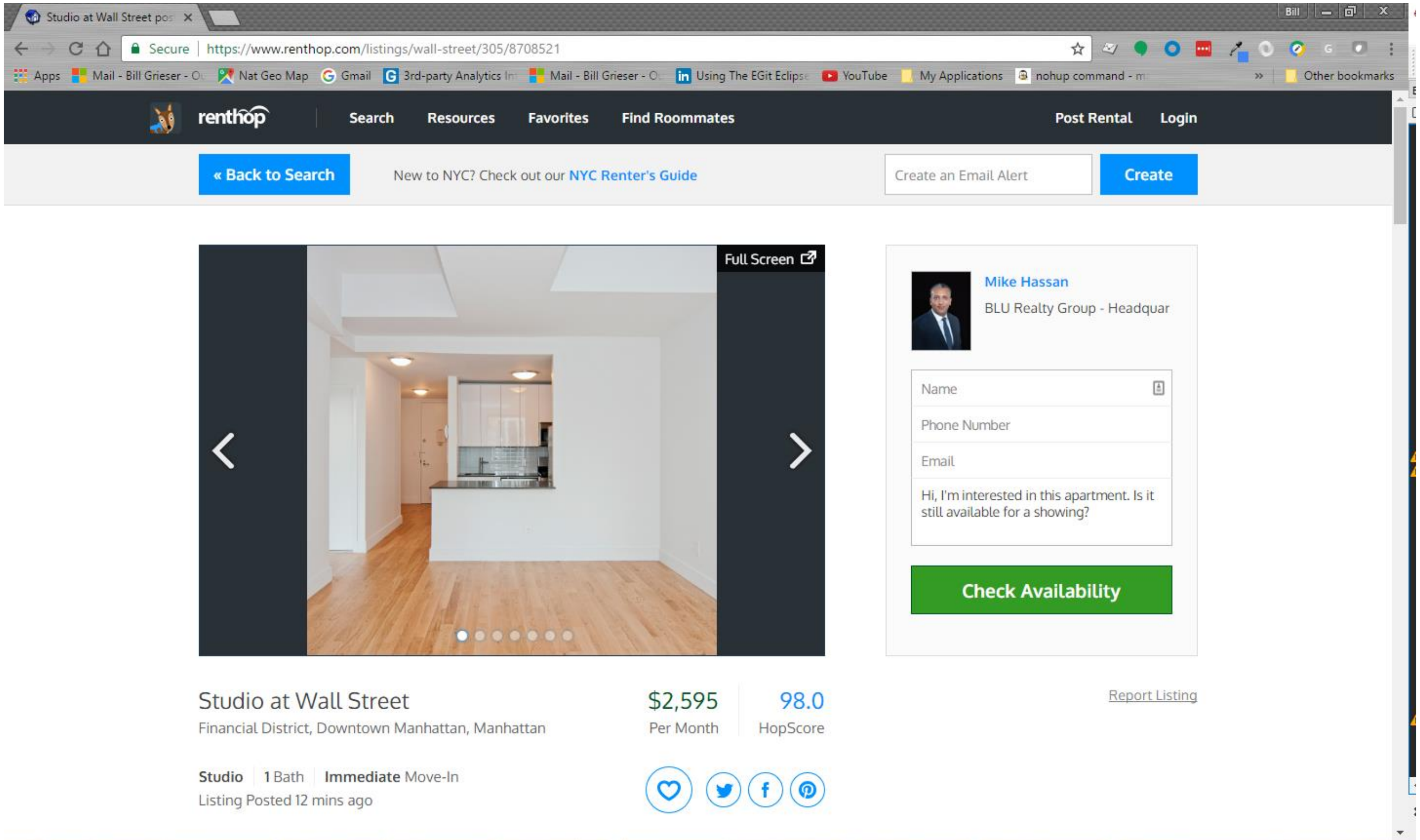
The Renthop Website



Search



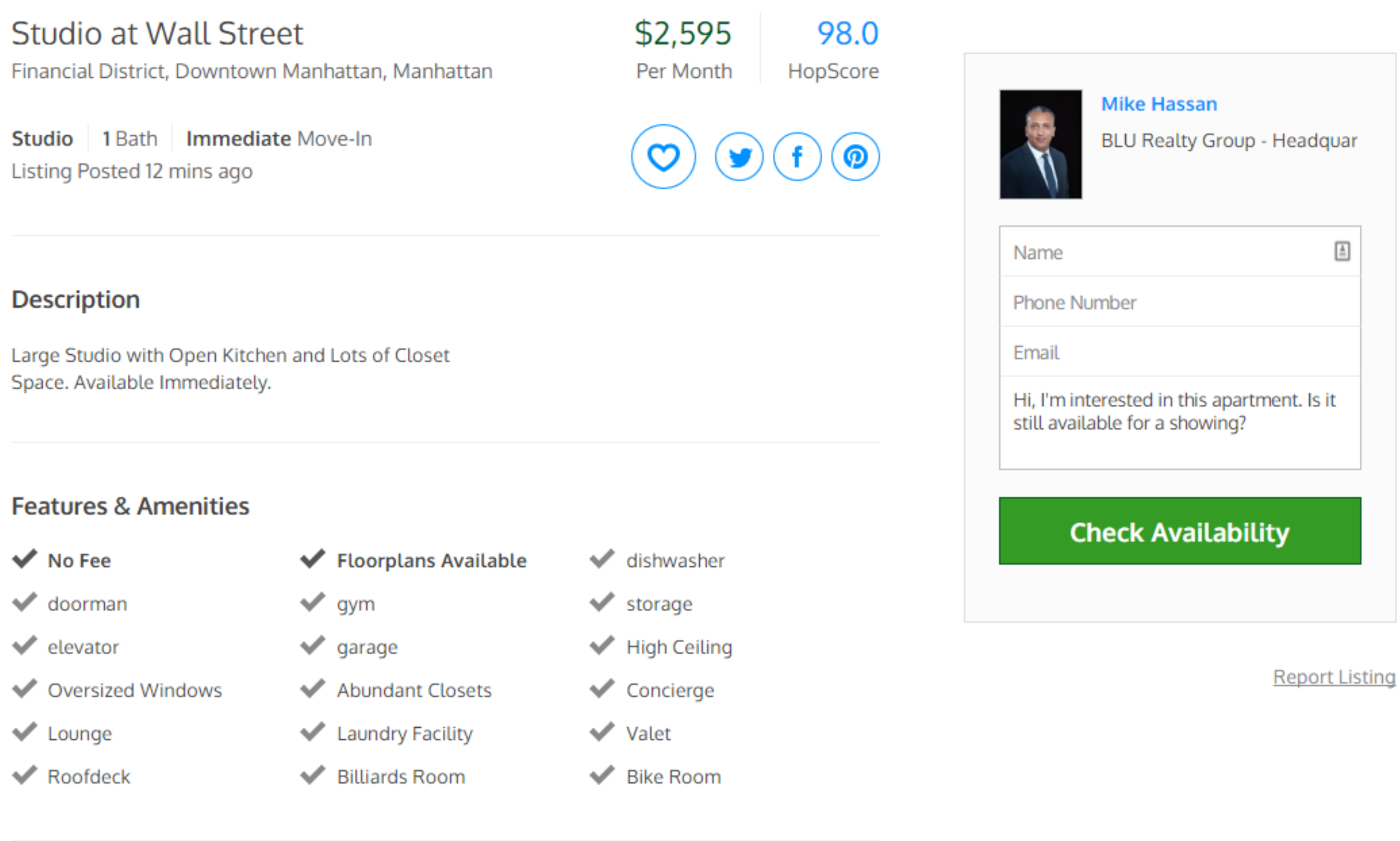
Bathrooms
1
2
3
4
5+
Keywords
WiFi, Parking (Comma Separated)
Must Haves
Has Photos
Has Floorplan
Open House
Listing Type
By Owner
Exclusive
Sublet / Lease-Break
No Fee
Reduced Fee
Short Term Allowed
Unit Features
Furnished
Laundry In Unit
Private Outdoor Space
Parking Space
Pet Policy
Cats Allowed
Dogs Allowed
Building Features
Doorman
Elevator
Fitness Center
Laundry In Building
Common Outdoor Space
Storage Facility



A listing has:

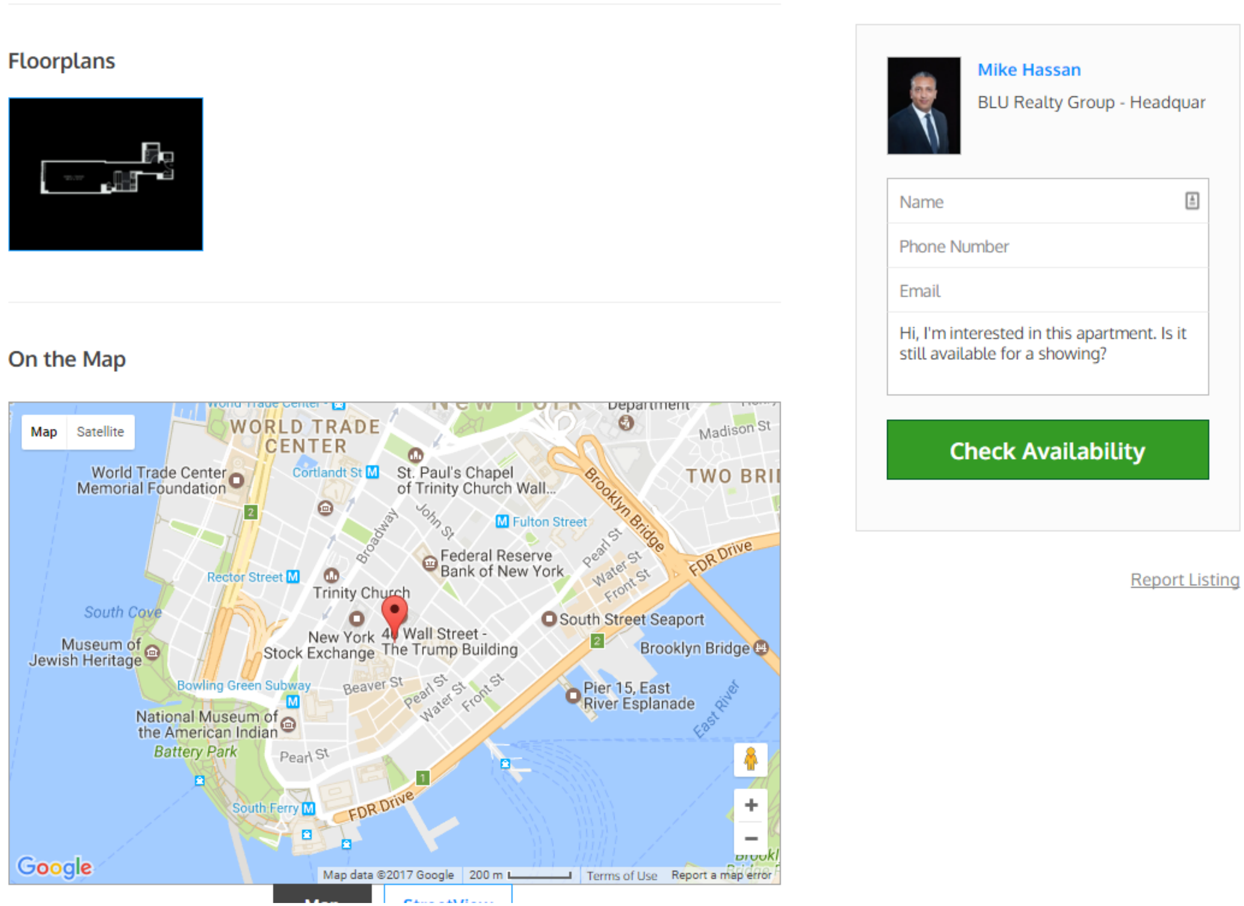
(Data available in **Bold)**

- **Photos**
- **Size**
- **Display Address**
- Listing Agent Information (**Id only**)
- **Price**
- “Hop Score”



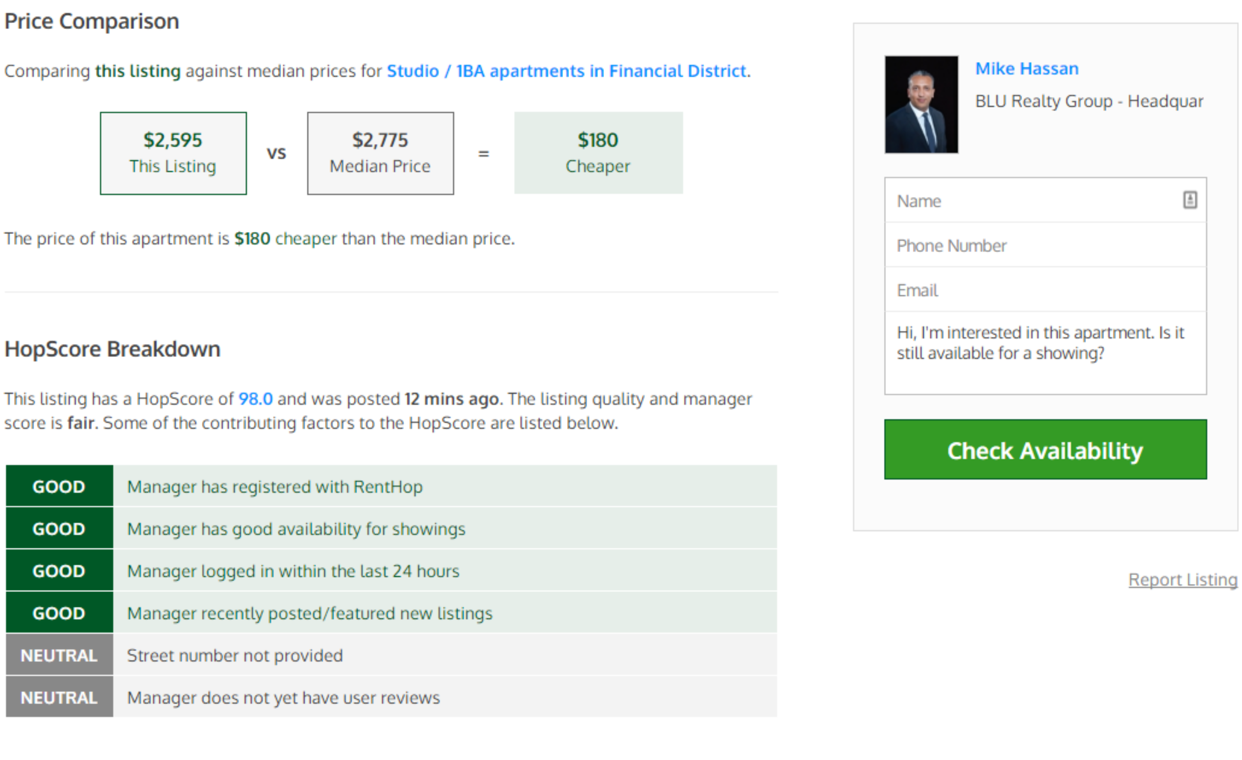
A listing has:

- How long on site **(Creation Date)**
- **Description**
- **Bulleted Feature List**
- **Check Availability** (probably our target)



A listing has:

- Optional Floorplan **(as picture)**
- Map (**Longitude / Latitude)**



Price Comparison

Comparing **this listing** against median prices for **Studio / 1BA apartments in Financial District.**

$2,595 This Listing	vs	$2,775 Median Price	=	$180 Cheaper

The price of this apartment is **$180** cheaper than the median price.

HopScore Breakdown

This listing has a HopScore of 98.0 and was posted **12 mins ago**. The listing quality and manager score is **fair**. Some of the contributing factors to the HopScore are listed below.

GOOD	Manager has registered with RentHop
GOOD	Manager has good availability for showings
GOOD	Manager logged in within the last 24 hours
GOOD	Manager recently posted/featured new listings
NEUTRAL	Street number not provided
NEUTRAL	Manager does not yet have user reviews

Mike Hassan
BLU Realty Group - Headquar

Name
Phone Number
Email
Hi, I'm interested in this apartment. Is it still available for a showing?

Check Availability

Report Listing

A listing has:

- Price Comparison
- Hop Score (really a manager score)

Similar Apartments Nearby

Studio at Financial District
Financial District, Downtown Manhat...
$2,550 Per Month | 100 HopScore
Posted 49 mins ago

Studio at Wall Street
Financial District, Downtown Manhat...
$2,745 Per Month | 100 HopScore
Posted 13 mins ago

Studio at FiDi Water
Financial District, Downtown Manhat...
$2,389 Per Month | 100 HopScore
Posted 18 mins ago

Studio at Water
Financial District, Downtown Manhat...
$2,390 Per Month | 99.9 HopScore
Posted 22 mins ago

Studio at Pine Street
Financial District, Downtown Manhat...
$2,820 Per Month | 99.0 HopScore
Posted 53 mins ago

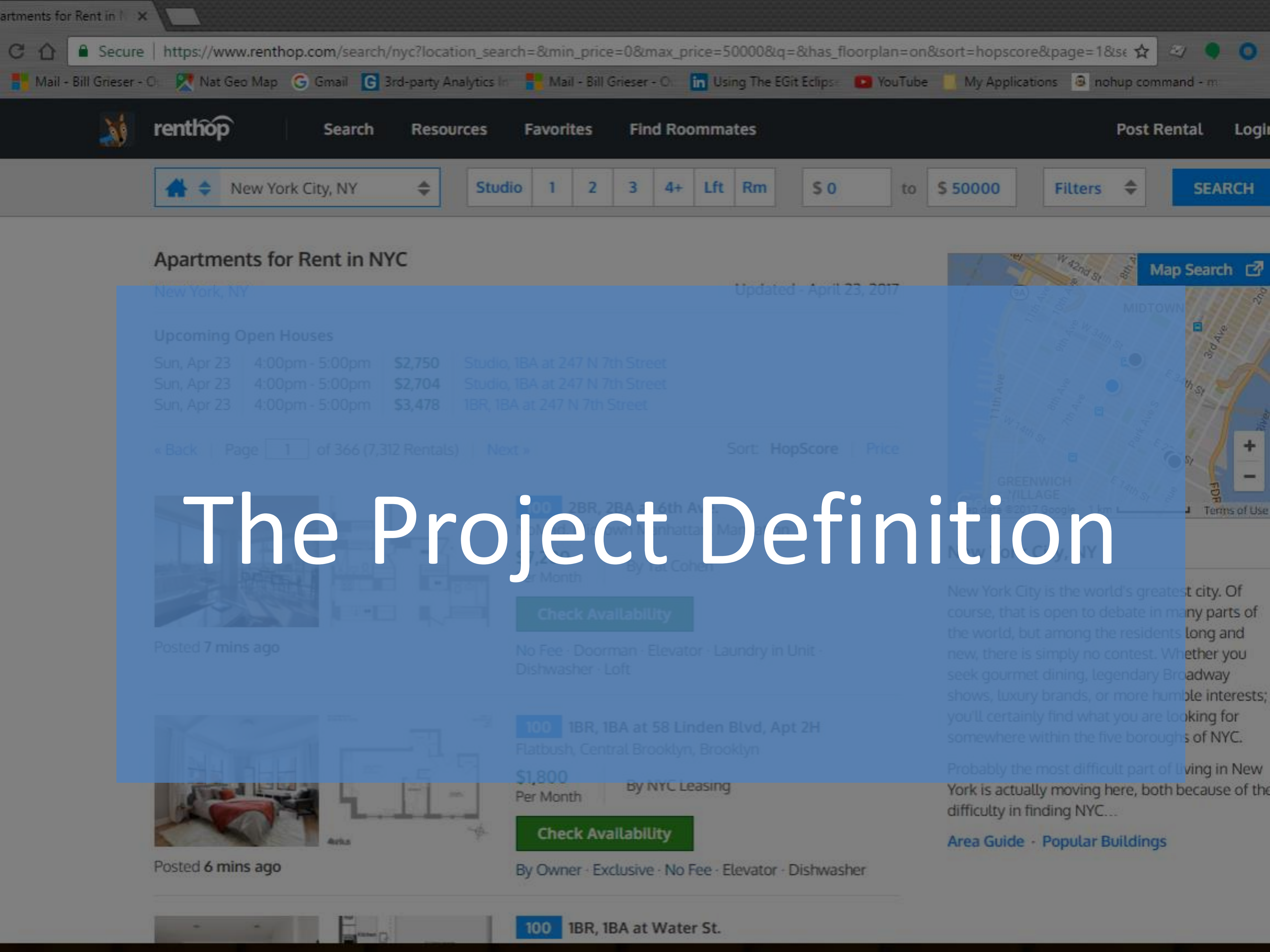
Studio at Water
Financial District, Downtown Manhat...
$2,398 Per Month | 99.1 HopScore
Posted 25 mins ago

Studio at Exchange Place
Financial District, Downtown Manhat...
$2,420 Per Month | 98.2 HopScore
Posted 33 mins ago

A listing has:

- Comparables

The Project Definition



The Question

Can we identify factors that influence the interest level specific New York City listings receive on the Renthop.com apartment listing web site . . .

. . . by fitting a Machine Learning model to predict listing interest level, using only data provided by Renthop through the Kaggle "Two Sigma Connect: Rental Listing Inquiries" competition?

kaggle

Two Sigma CONNECT renthop

S.M.A.R.T.

- Specific
  - Defined by the Kaggle Competition
- Measurable
  - Use goodness-of-fit measures of the classification model
  - Kaggle score submissions
  - Use other statistical measures (chi squared test)
- Actionable
  - Where possible, choose techniques that are interpretable, even at the expense of model accuracy
- Relevant
  - Real-life problem
- Time Bound
  - Strictly Bounded – see next chart

Bounds of the Problem

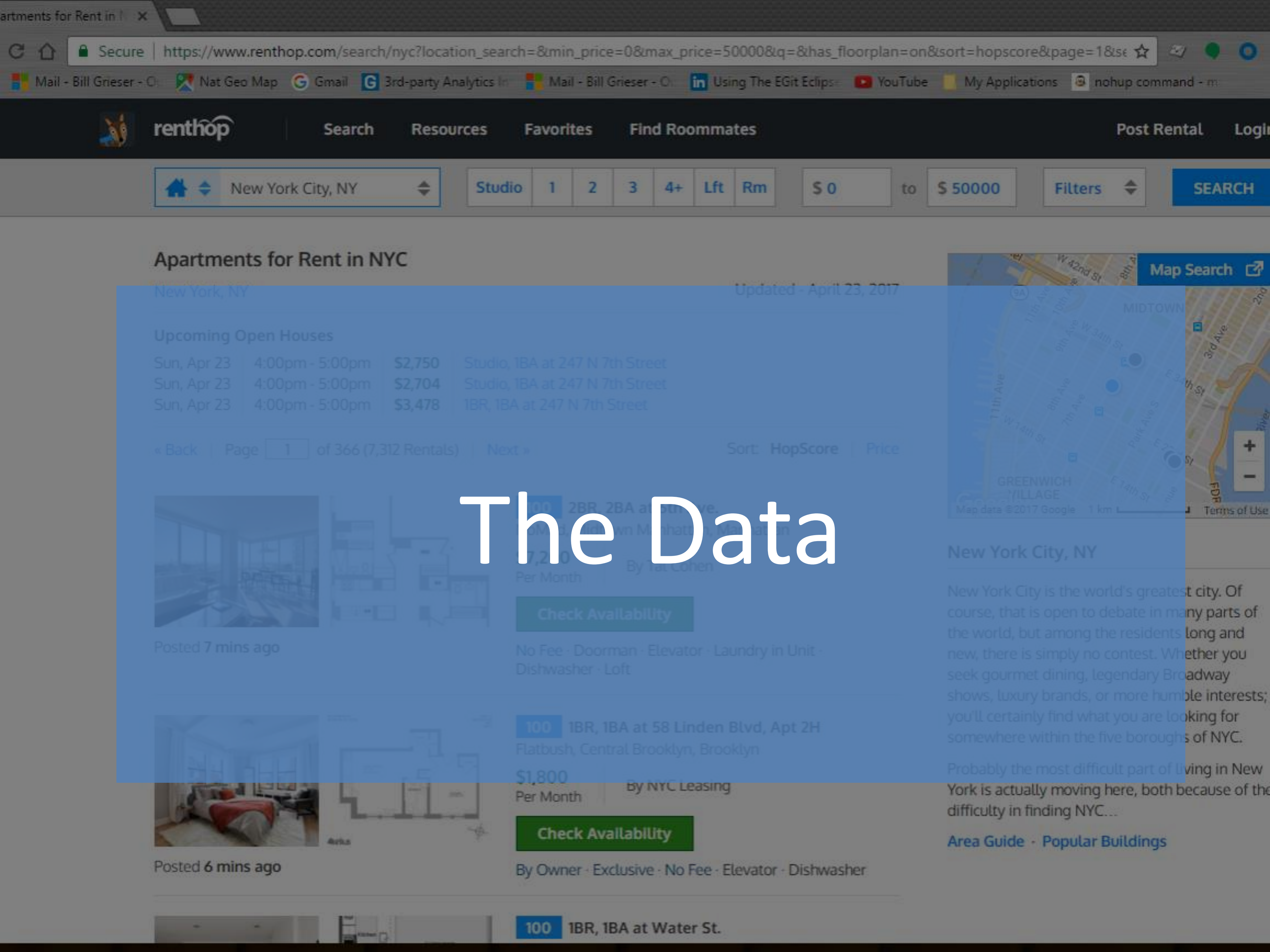
- The main bound is the due date
  - Kaggle and class coincide
- Kaggle ground rule: No outside data

It became essential to time-bound the phases in order to not get lost

What does Success look like?

- A fitted model using most of the data provided that has a reasonable goodness of fit, as measured by Log Loss and F1 measures against held out test data
- Some insights into successful listings supported by the data and model
- A Kaggle entry in the competition

The Data


Secure
https://www.renthop.com/search/nyc?location_search=&min_price=0&max_price=50000&q=&has_floorplan=on&sort=hopscore&page=1&se
Mail - Bill Grieser
Nat Geo Map
Gmail
3rd-party Analytics
Using The EGit Eclipse
YouTube
My Applications
nohup command
renthop
Search
Resources
Favorites
Find Roommates
Post Rental
New York City, NY
Studio
1
2
3
4+
Lft
Rm
$ 0
to
$ 50000
Filters
SEARCH
Apartments for Rent in NYC
Map Search
New York City, NY
100
1BR, 1BA at 58 Linden Blvd, Apt 2H
$1,800
Per Month
By NYC Leasing
Check Availability
Posted 6 mins ago
By Owner · Exclusive · No Fee · Elevator · Dishwasher
York is actually moving here, both because of the
difficulty in finding NYC...
Area Guide · Popular Buildings
100
1BR, 1BA at Water St.

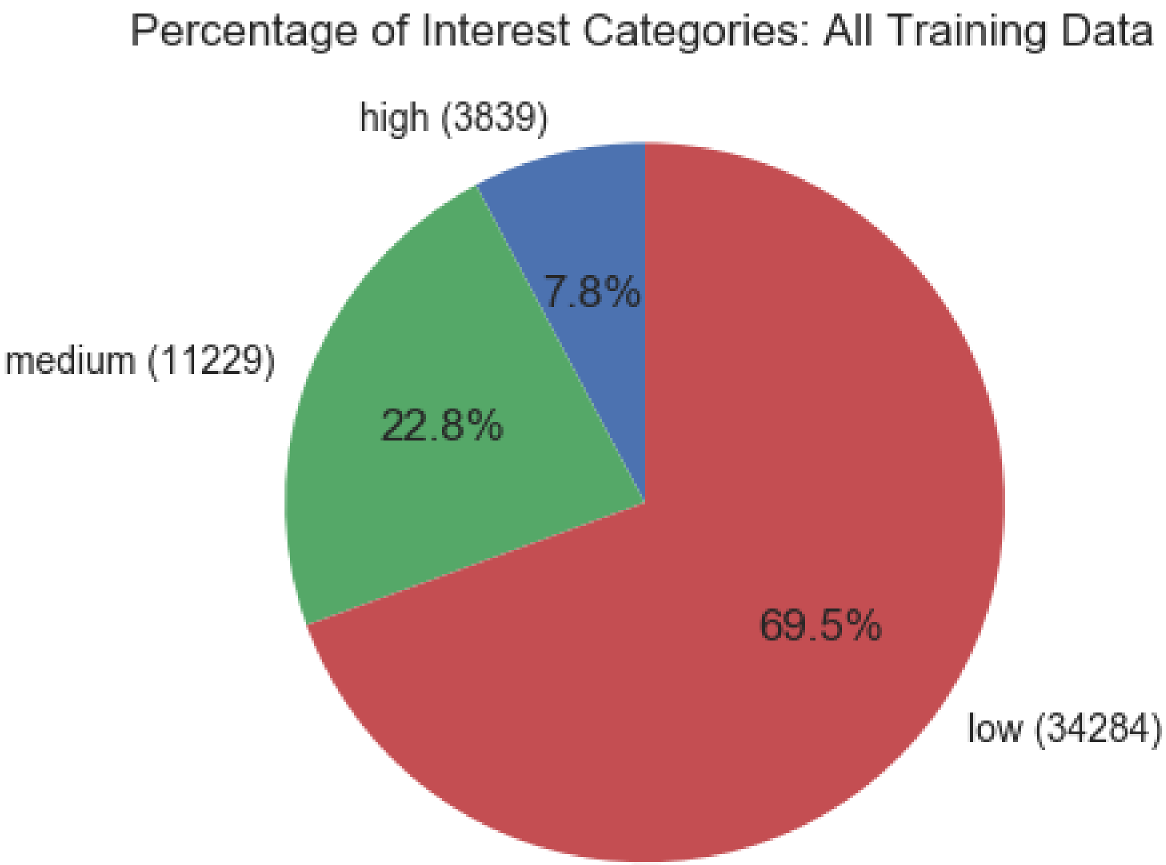
Original Data Fields

- **Listing_id** (int): Unique ID for the listing
- **Bedrooms** (int): number of bathrooms
- **Bathrooms** (decimal): number of bathrooms. Half-baths are 0.5
- **Building Id** (GUID): Unique ID of building, or 0 if no building id
- **Created** (date): Creation data of the listing
- **Description** (string): Free-form text description entered by the listing agent
- **Display Address** (string): Address of property as seen in the listing
- **Features** (list of strings): a list of features entered by the listing agent about this apartment
- **Longitude, Latitude** (decimal, degrees): Location of the apartment
- **Manager Id** (GUID): Unique ID of the listing agent
- **Photos** (list of URLS): The photos associated with the listing. Also provided in a file grouped by Listing ID.
- **Price** (int): Monthly rent in USD
- **Street Address** (string): Address of the apartment; not displayed in the listing
- **Interest Level** (string, categorical): This is the target variable. It has three categories: 'high', 'medium', 'low‘

Approach and Strategy

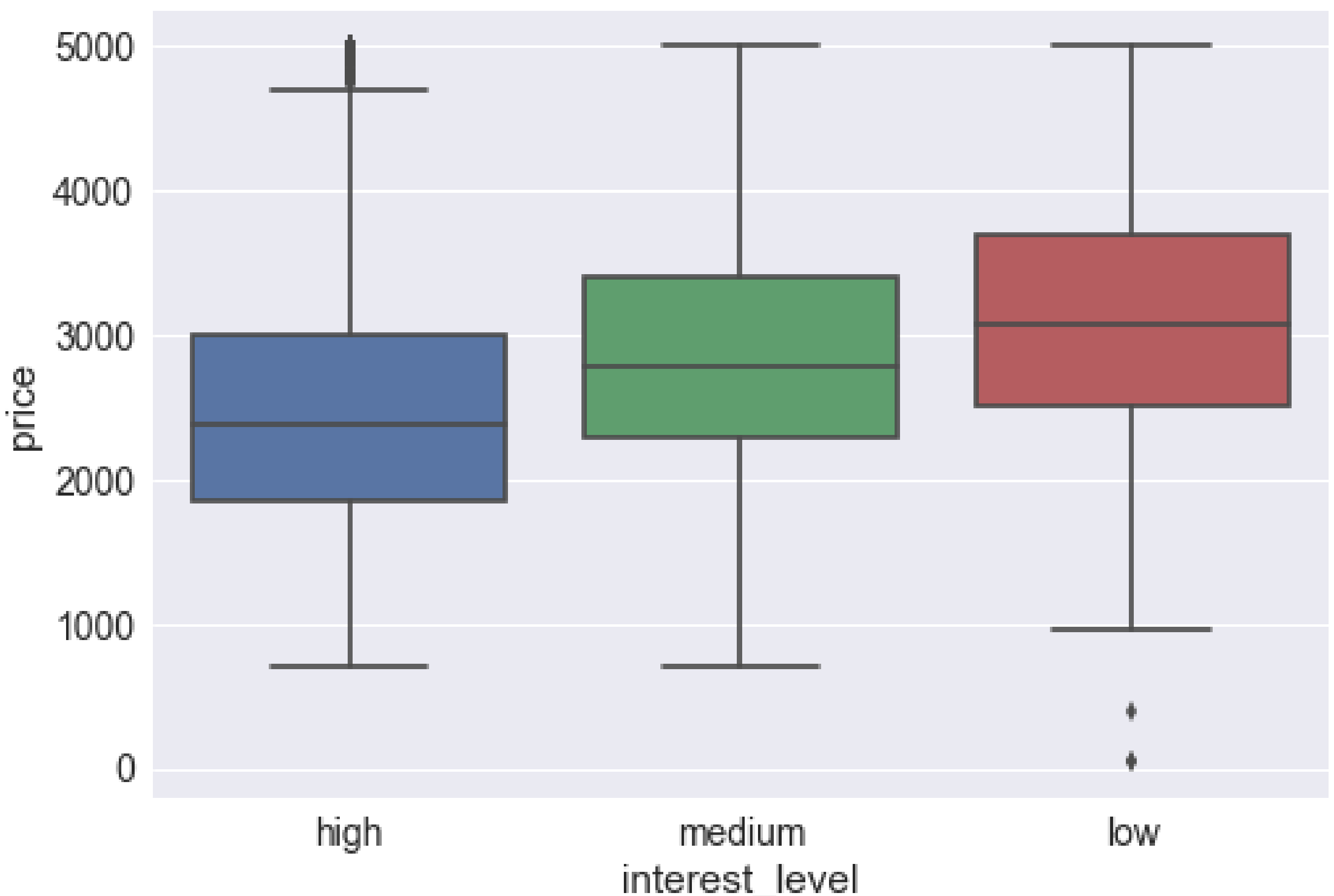
- Explore Data
- Create features usable for analysis
    1. **Variable length list of listing features**
    2. **Comparable listings**
    3. **Location**
    4. **Images**
    5. **Text description**
- Use mRMR, Random Forest, and home-grown feature selector to identify important features
- Fit competing models using different classifiers attempting to predict Interest Level
- Determine insights
- Report Results

Data Summary

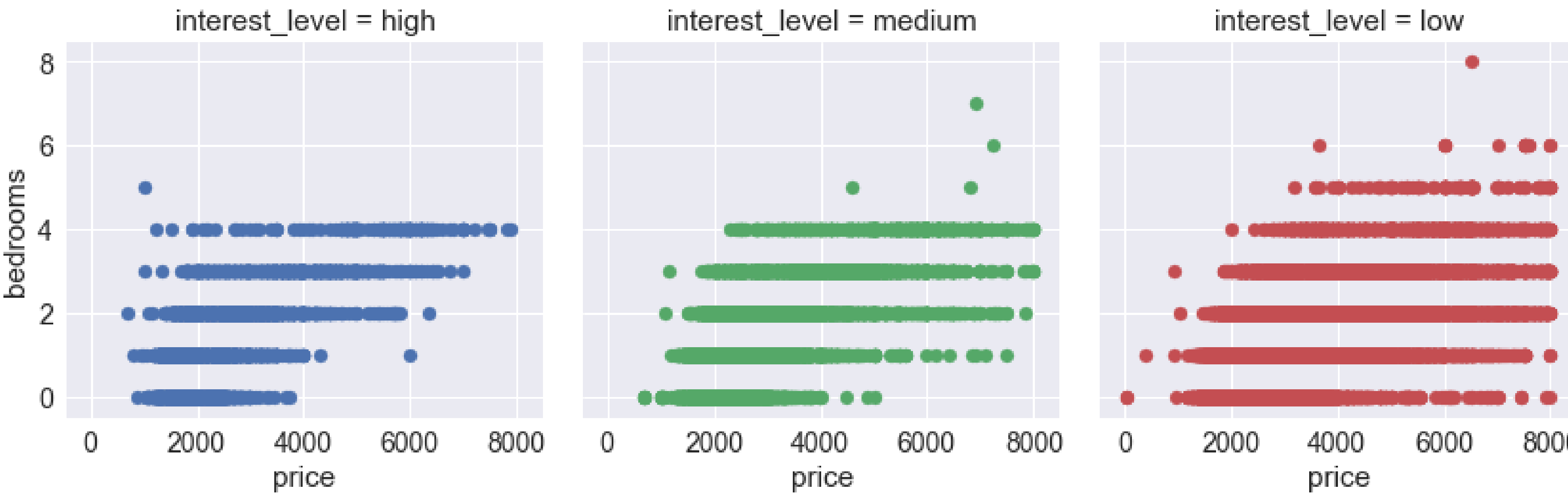


49352 Samples x 15 columns
Date range: April 1 – June 30 2016

Price Summary

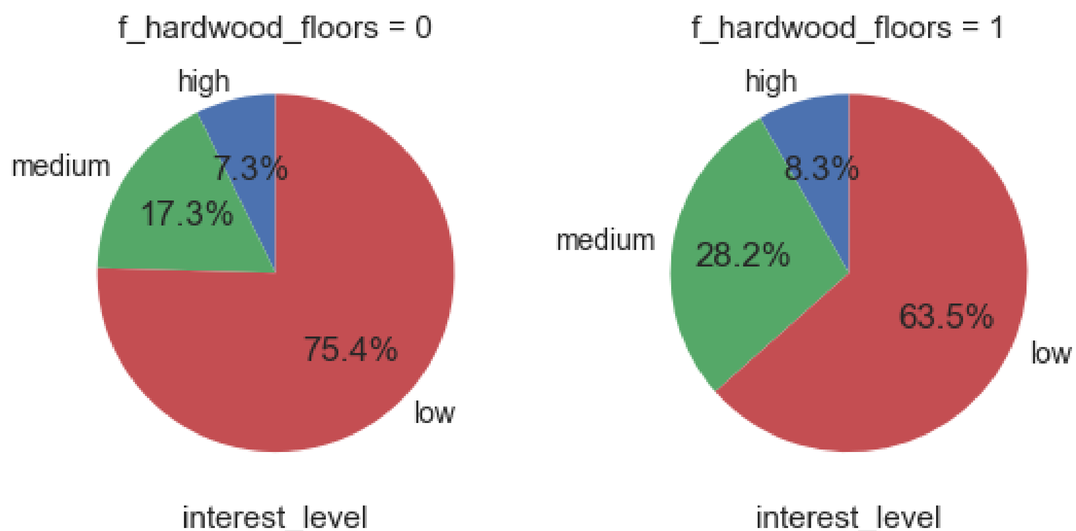


Price and Bedrooms by Interest Level



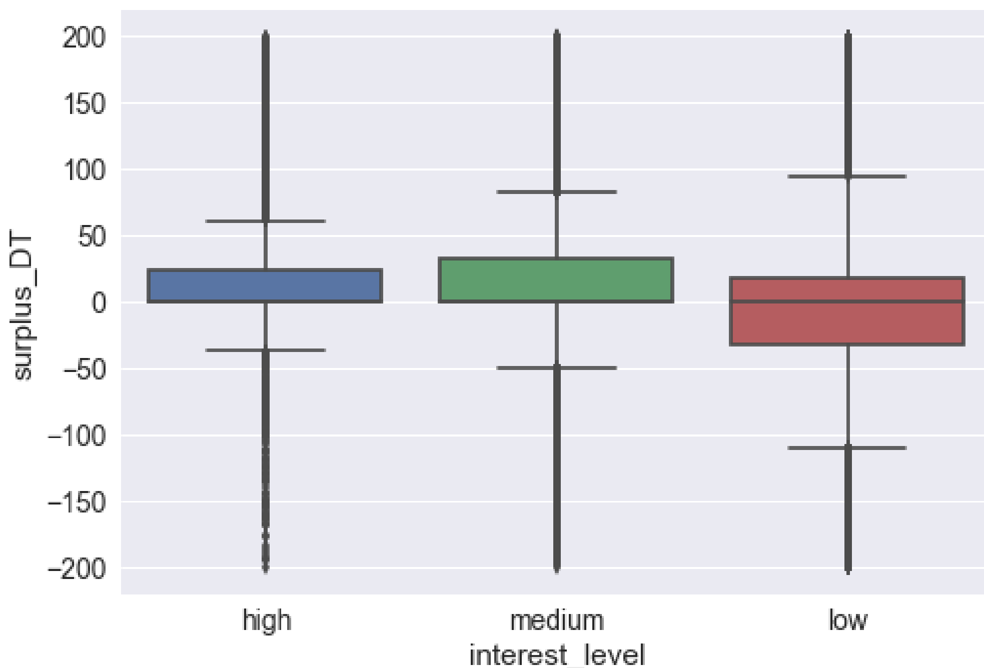
Task 1: Listing Features

- Found all features occurring > 100 times
- Used Tfidf Vectorizer to identify similar feature strings; collapsed them together
- Manually collapsed others
- Created dummy variables for remaining features (57)



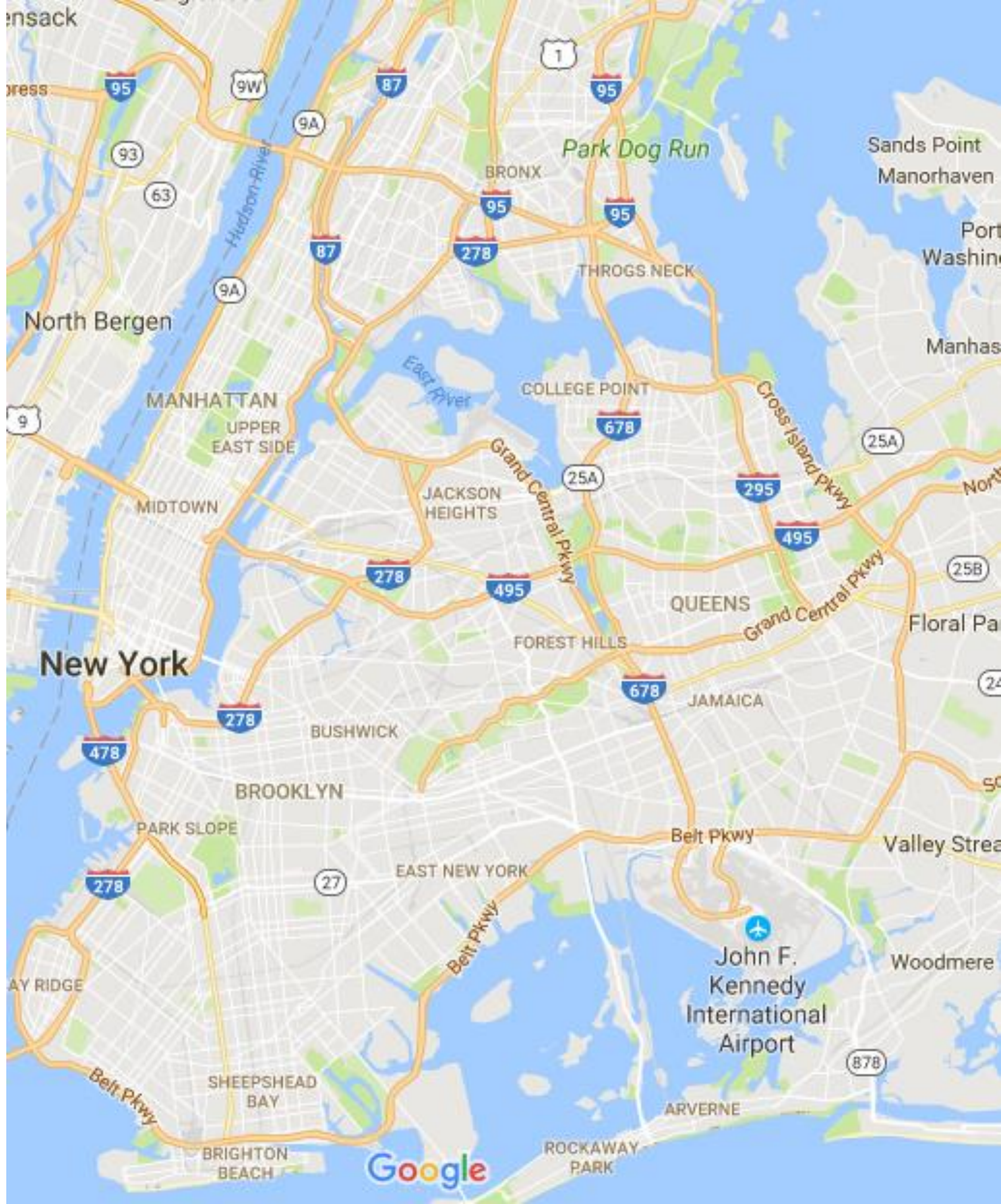
Task 2: Comparables

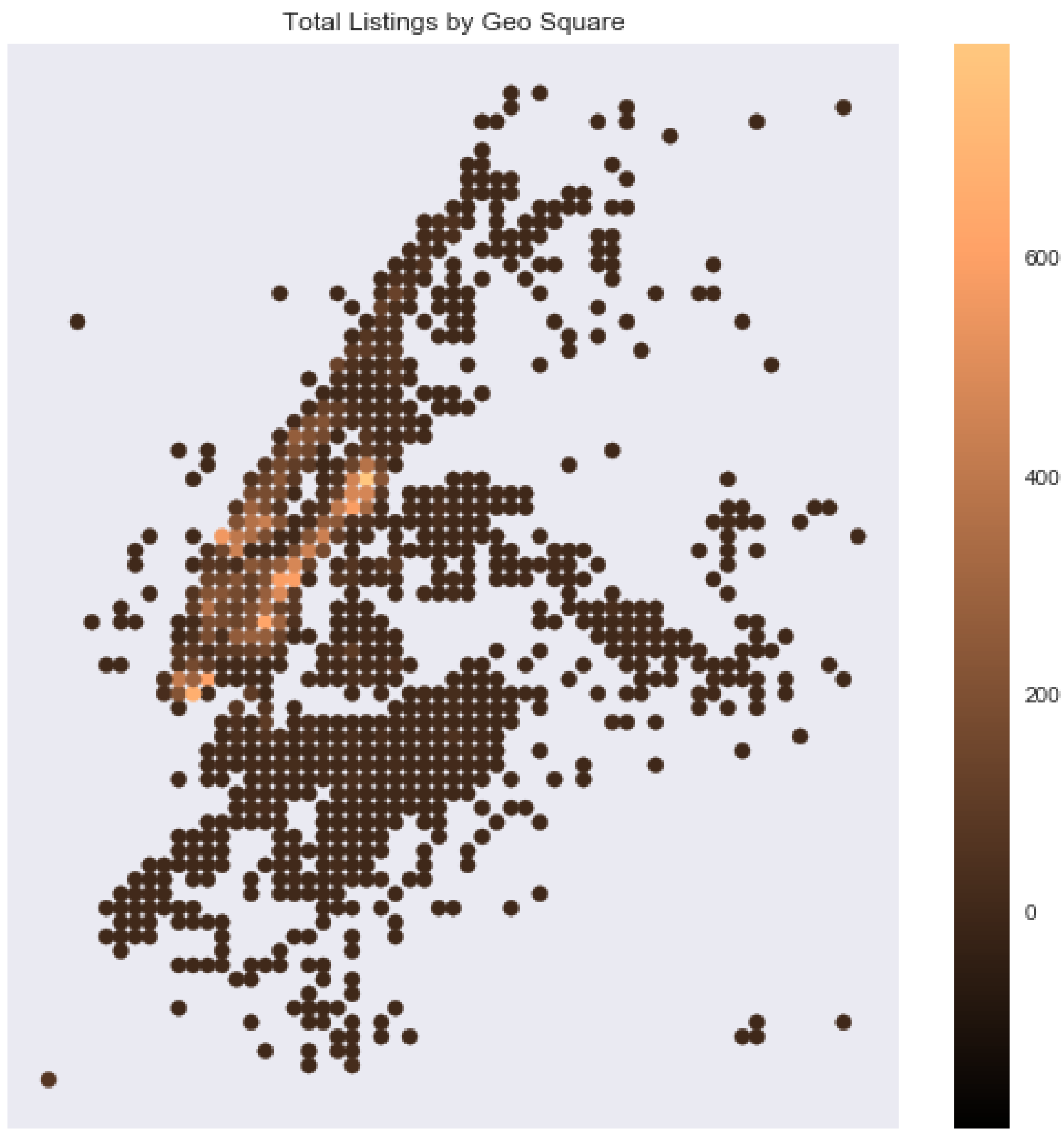
- In the training data, ran a regression to predict rent for an apartment that considered only:
  - Bedrooms
  - Bathroom
  - Lon/Lat
- Scored each listing's rent against their predicted rent and calculated a renter's surplus (+ is bargain, - is pricey)

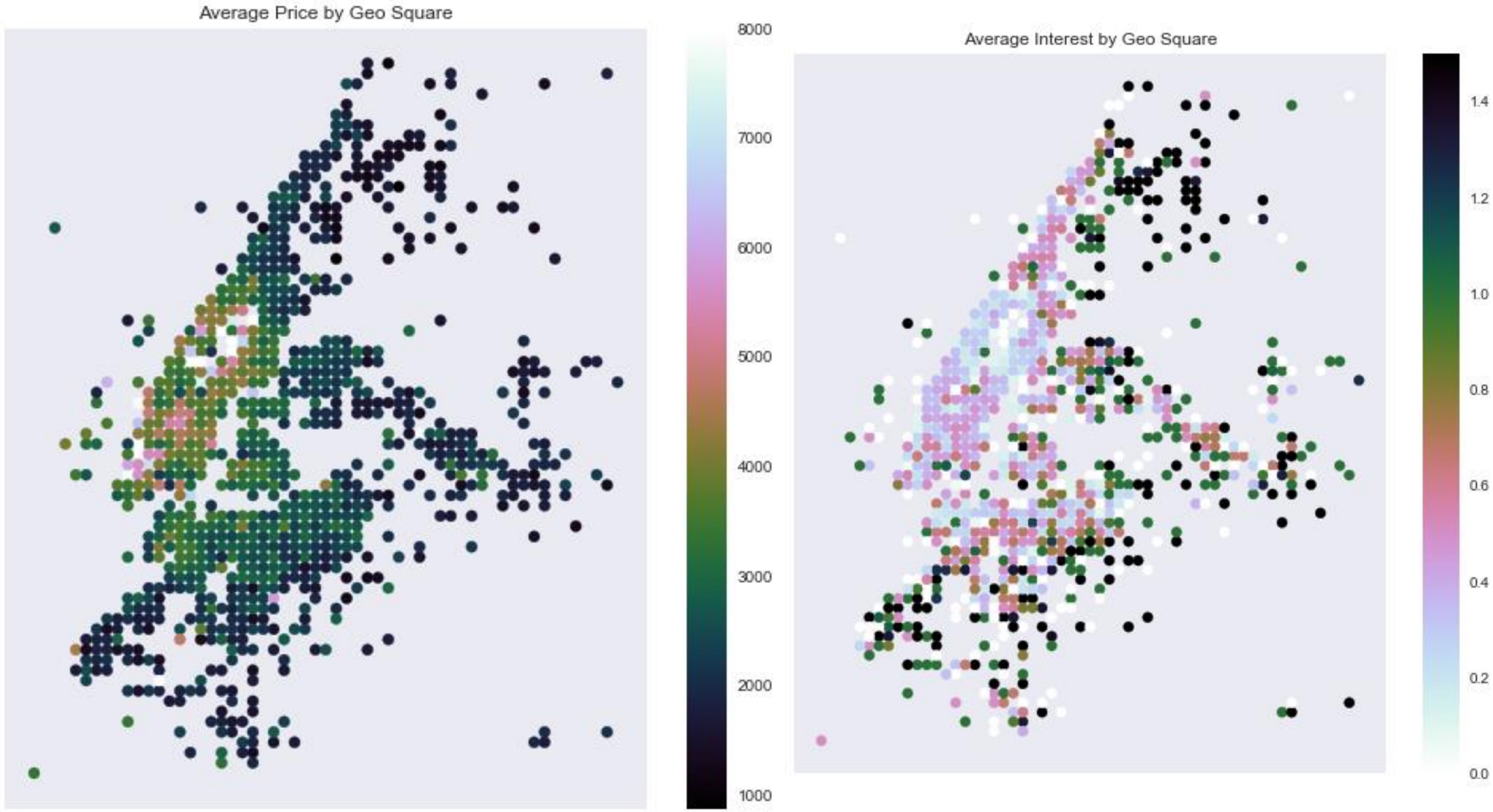


Task 3: Location

- Overlayed a lon/lat grid on NYC and determined a grid cell for each listing
- Calculated an average neighborhood desirability and used it to boost/penalize the listings in the cell




Total Listings by Geo Square
600
400
200
0


Average Price by Geo Square
8000
7000
6000
5000
4000
3000
2000
1000
Average Interest by Geo Square
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

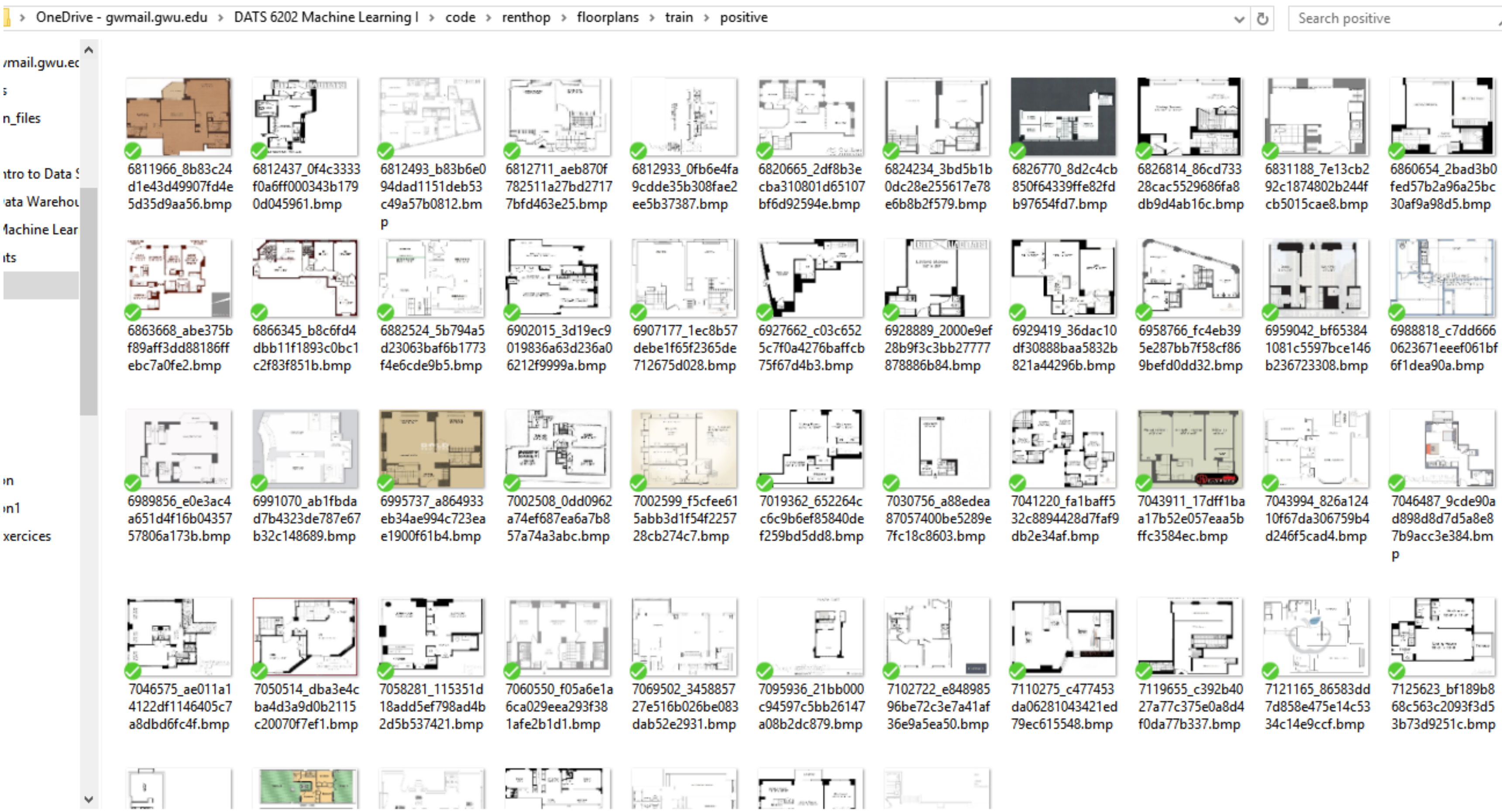
Task 4: Images

- Goal: Determine if a listing has a floorplan
- Converted every image (millions) to same size (160 x 140) and 256 colors
- Tried training images, running them trough PCA and then training on floorplans: No success

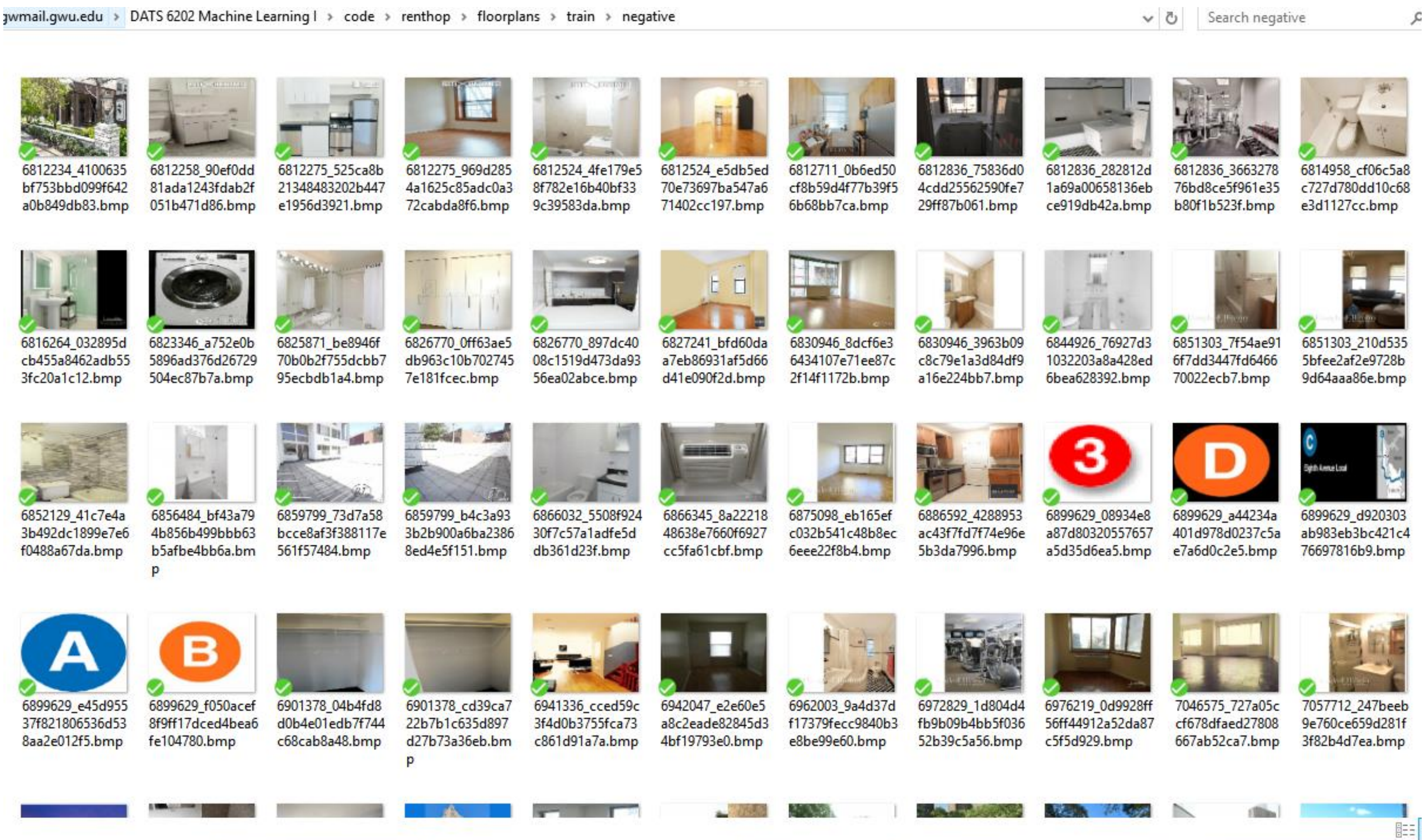
- **Found a method that worked**: Train a neural network on the image color **histograms**
- Manually bootstrapped the training data 50 listings at a time until model could process 150 listings with no errors
- Used an MLP Classifier (1 hidden layer with 50 nodes)

- Created a dummy variable for each listing: "Has Floorplan"

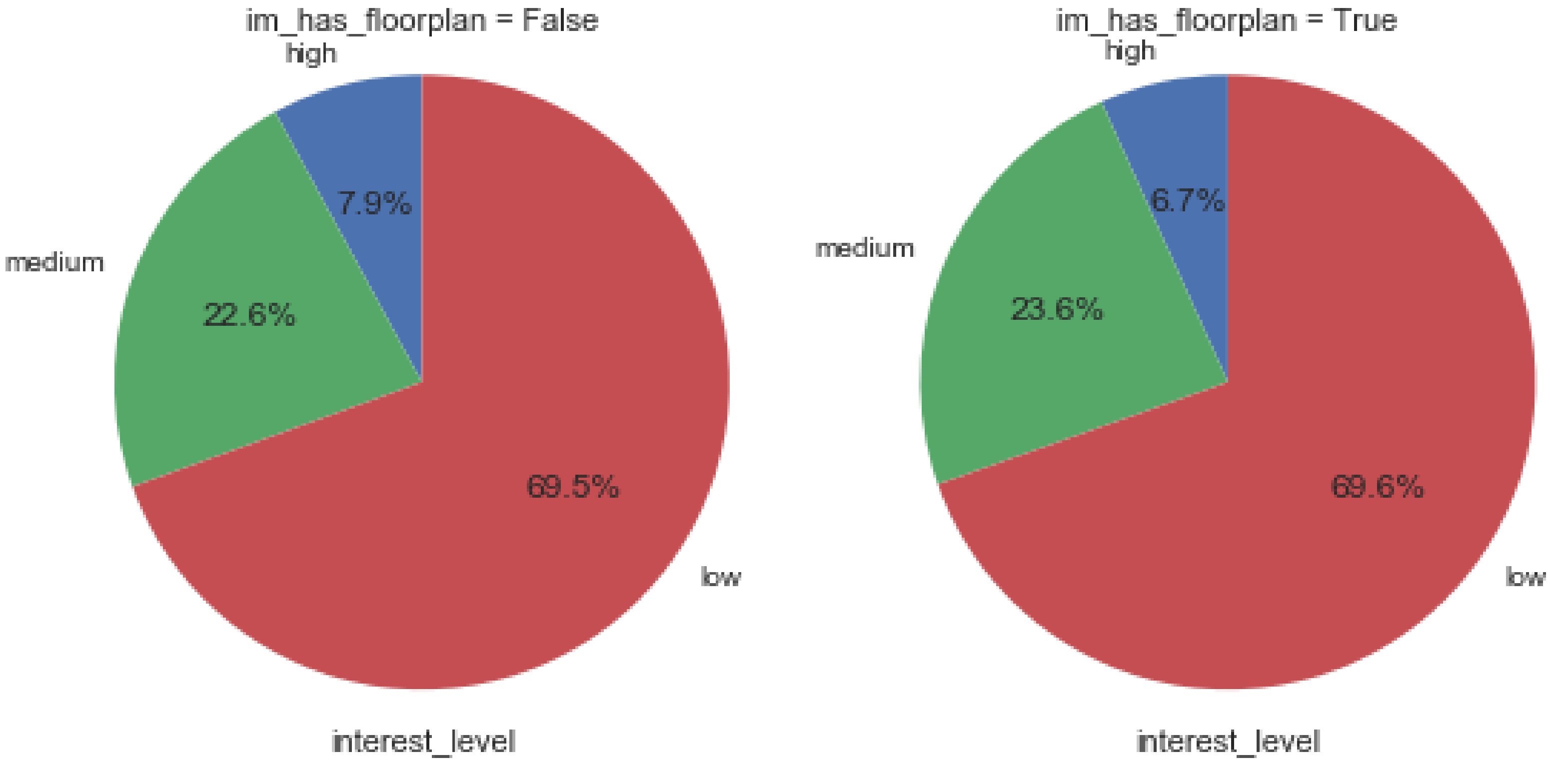
Positive Training Data



Negative Training Data



Floor Plan Results

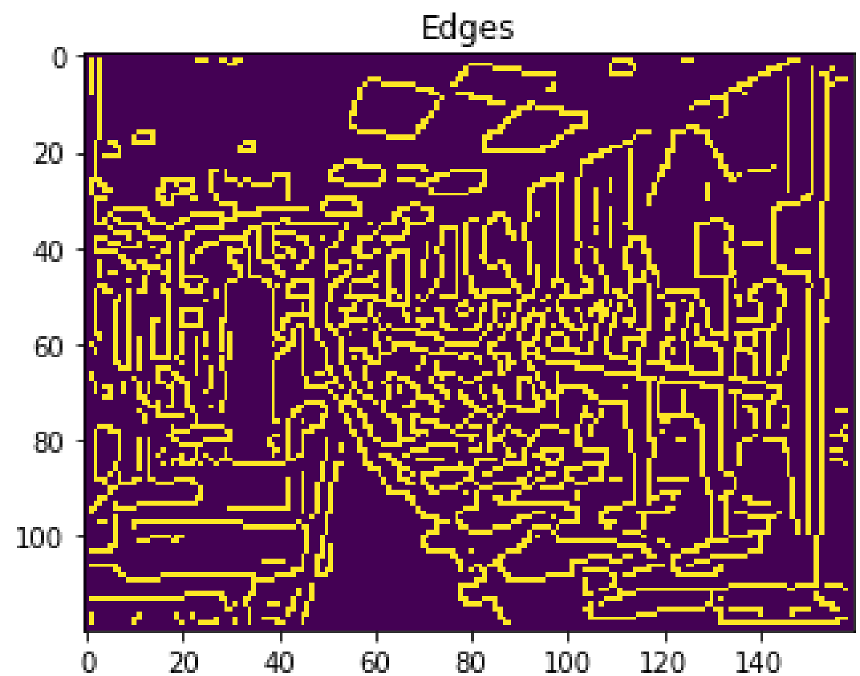


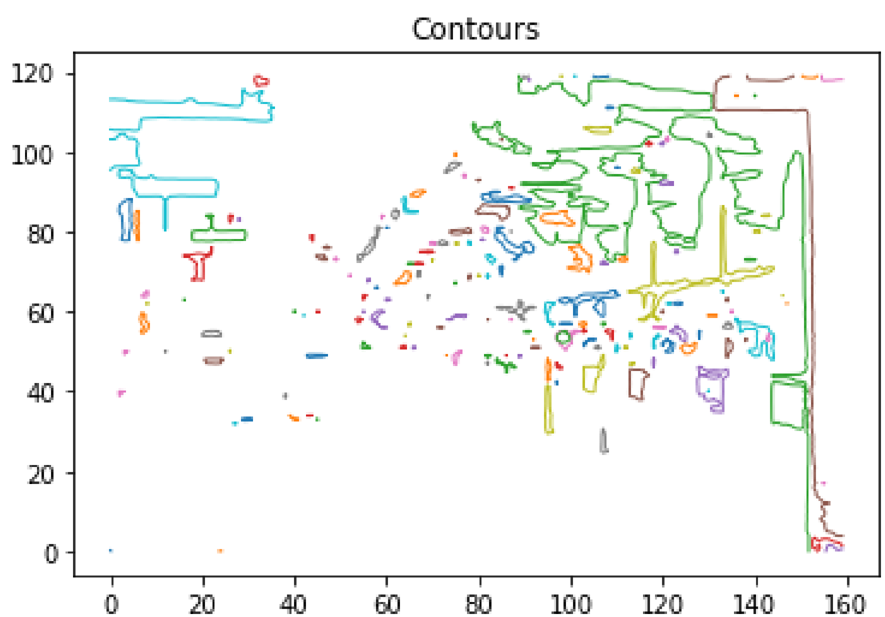
Busy vs. Plain Pictures

- What works better? Dull pictures of the apartment rooms or vibrant pictures of the view, the gym, the building itself, the neighborhood, etc?
- For each image, used sci-kit image to determine:
  - Number of contours
  - Number of points in edges
- Fed that to a KMeans unsupervised learner
  - Fit it with a subset of the training data
  - Selecting for 2 clusters gave sets of images with a perceived difference when run against larger sets of images
  - Better results not running through PCA first
  - Clustering distinguished “Plain” vs. “Busy” images
- Totals in training data after clustering:

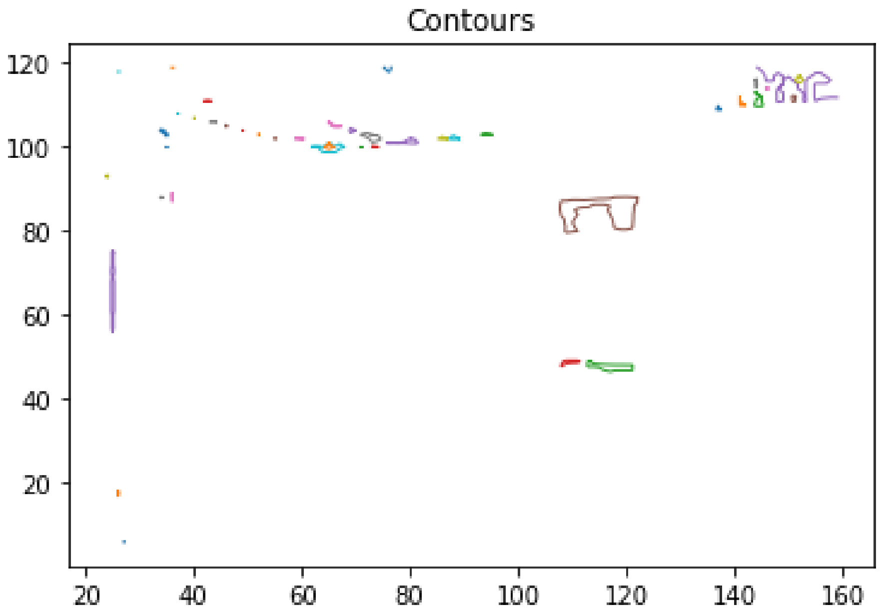
```
   30,032 Floorplans (excluded from the clusters)
1,252,704 “Plain” images
  930,136 “Busy” images
```

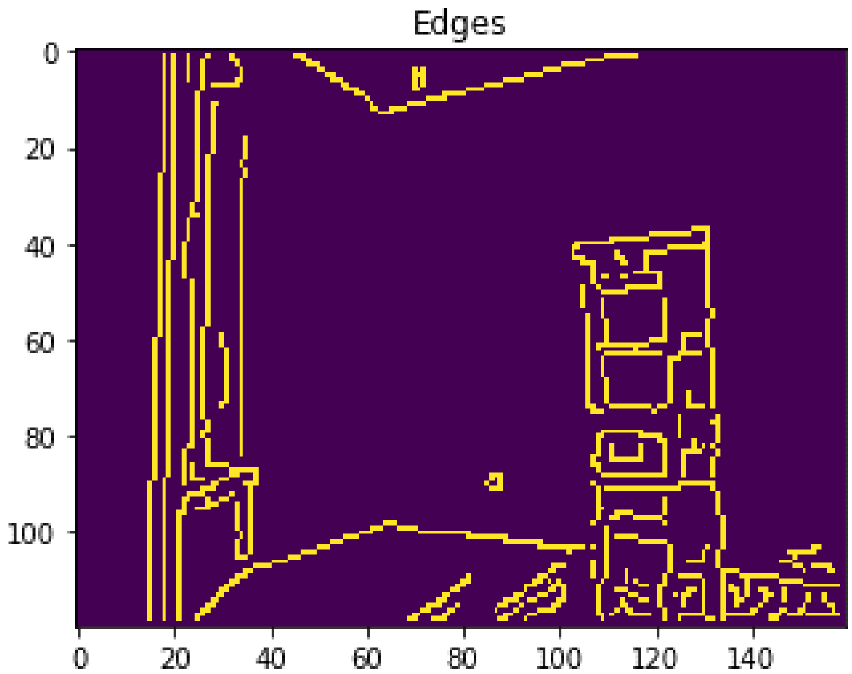
Busy Example



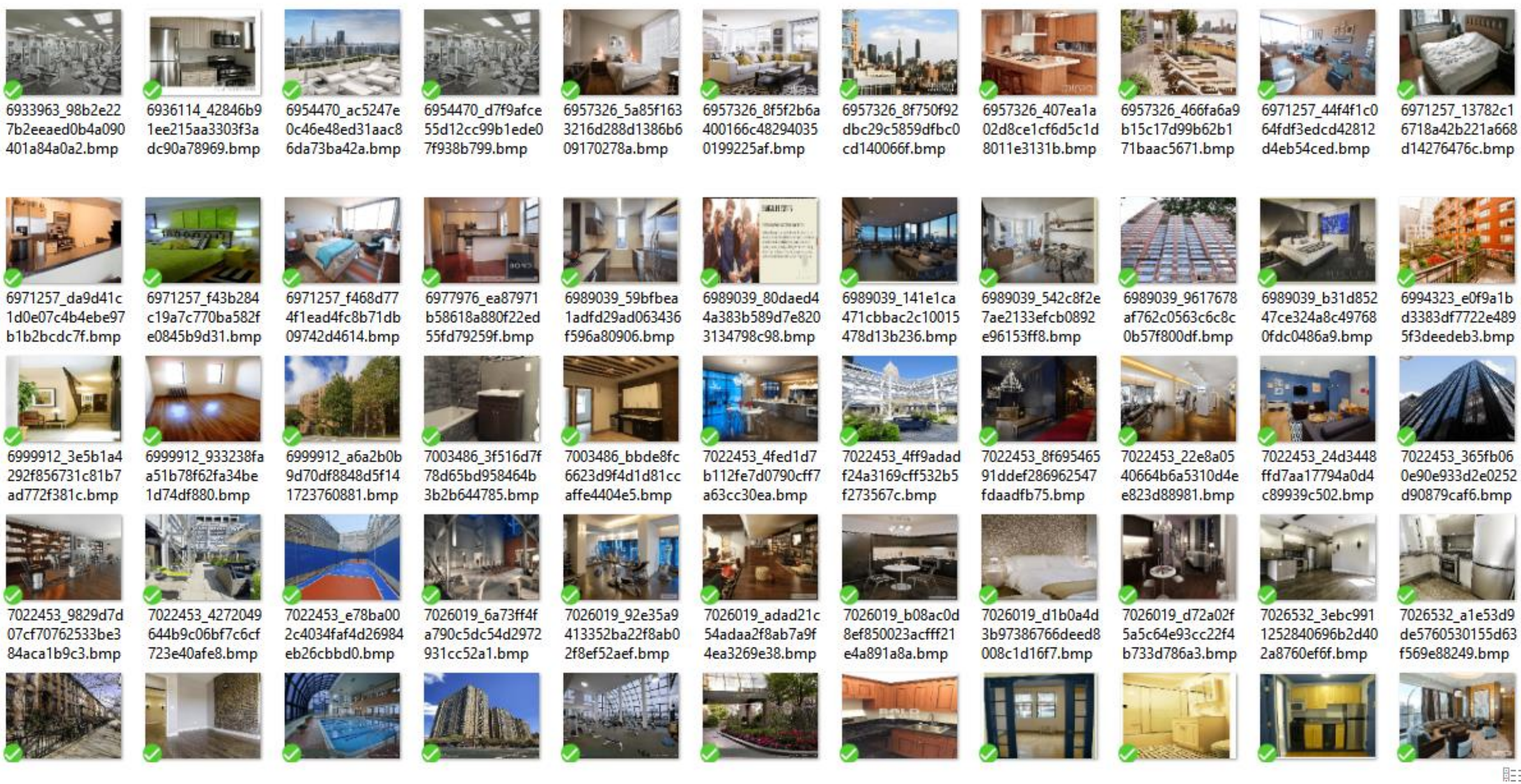


Plain Example

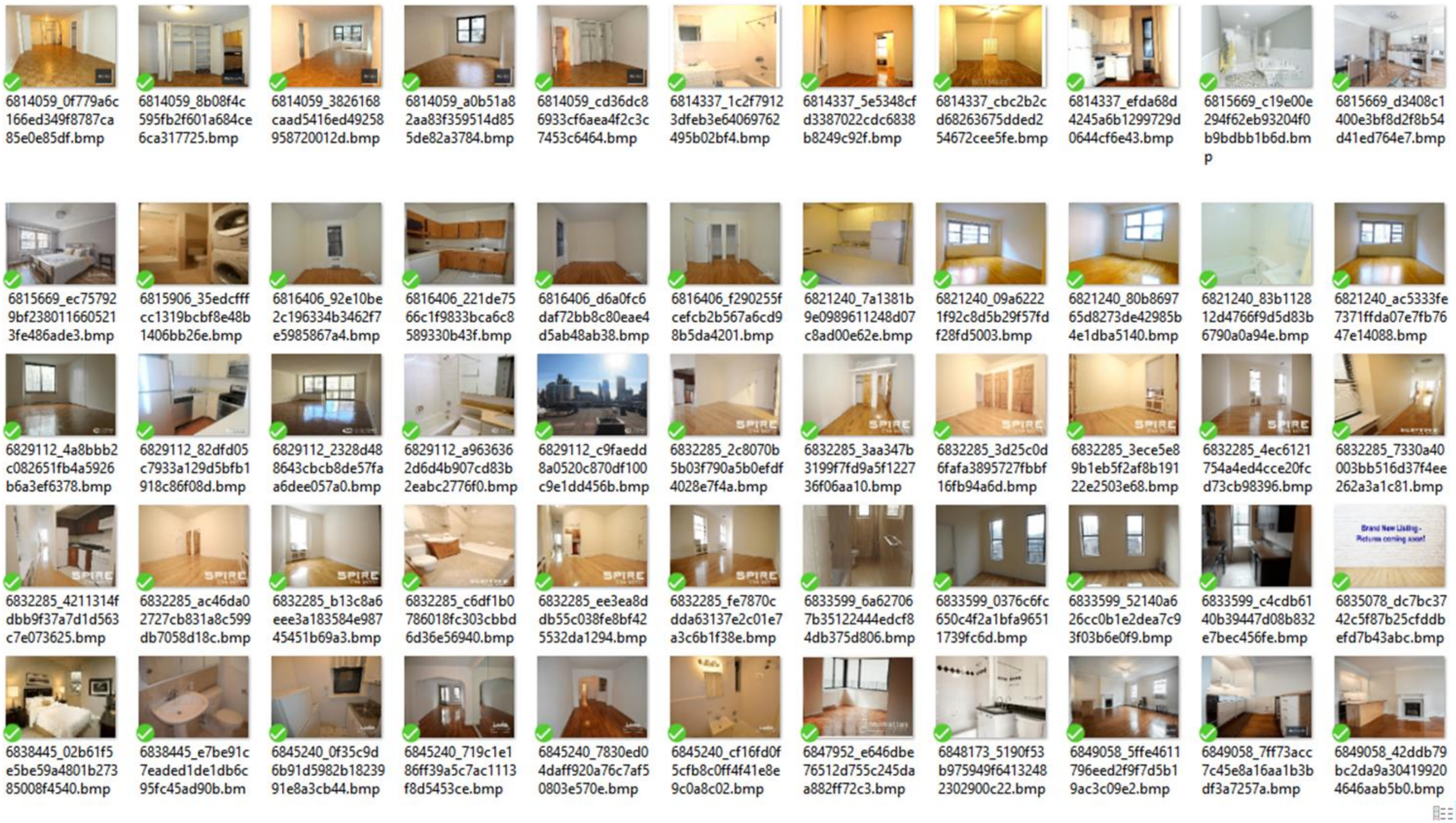




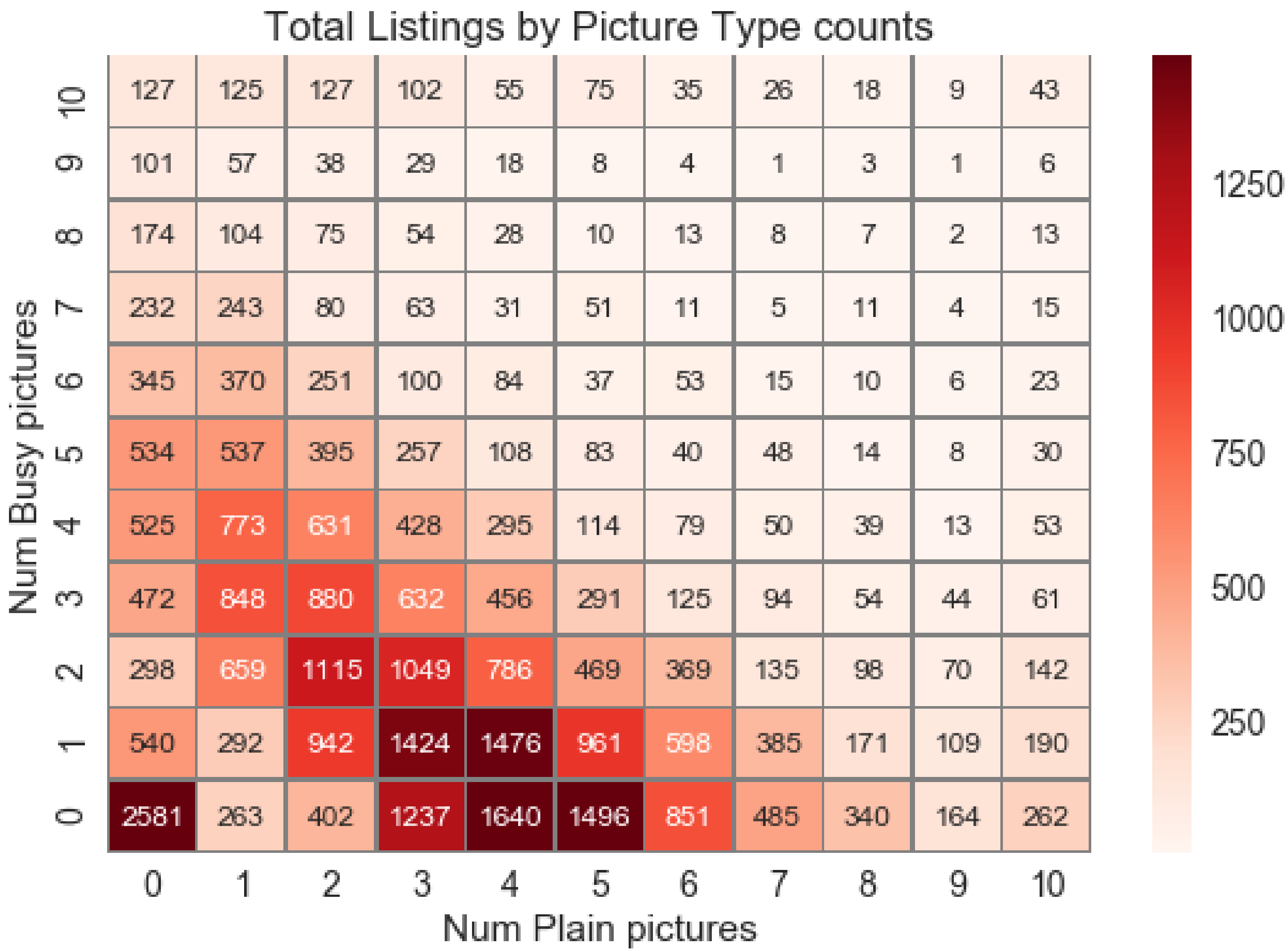
Some Busy Pictures



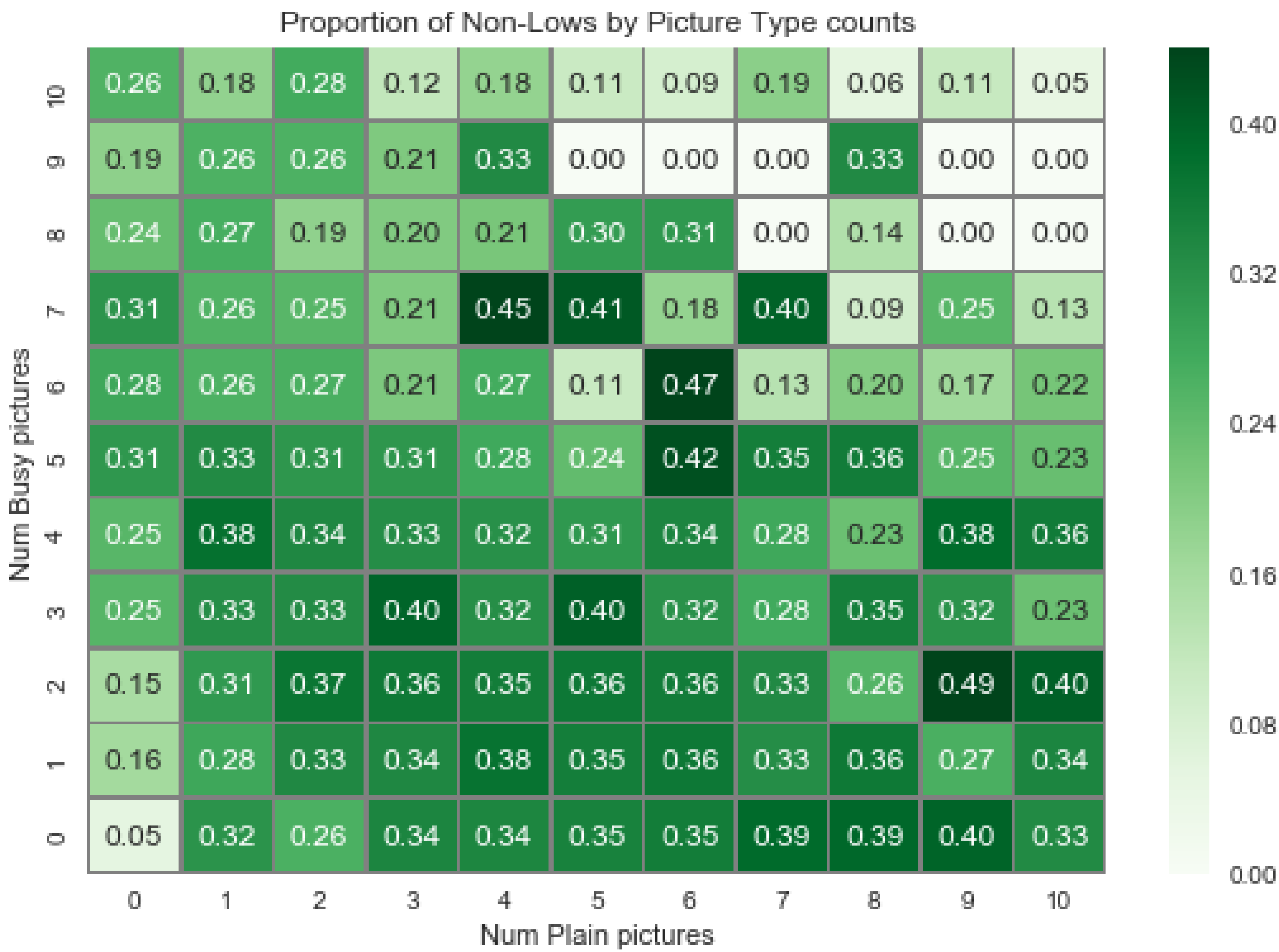
Some Plain Pictures



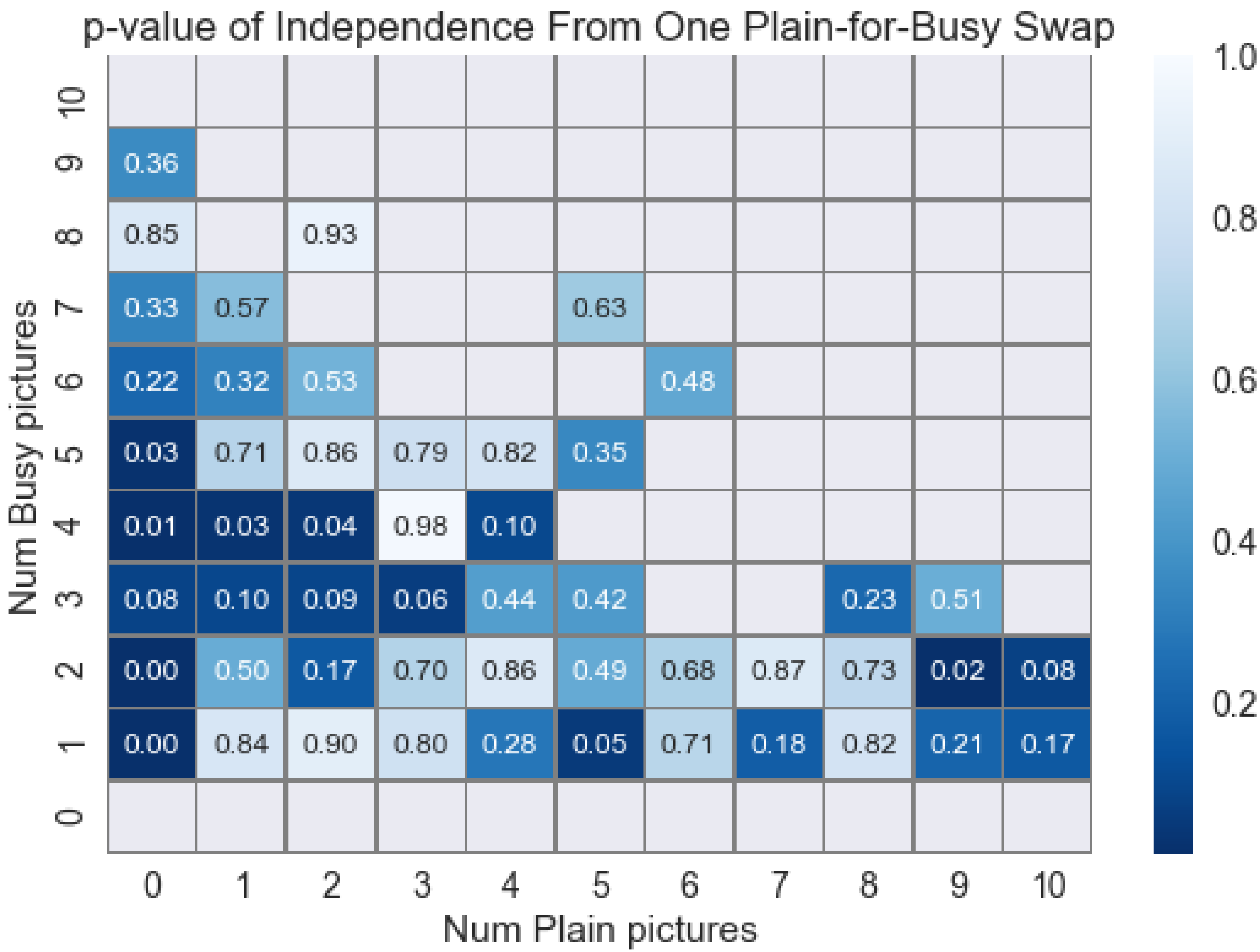
How Pictures Were Used



Success Metric for Picture Combos



Plain-for-Busy Chi Squared Test



Task 5: Text Description

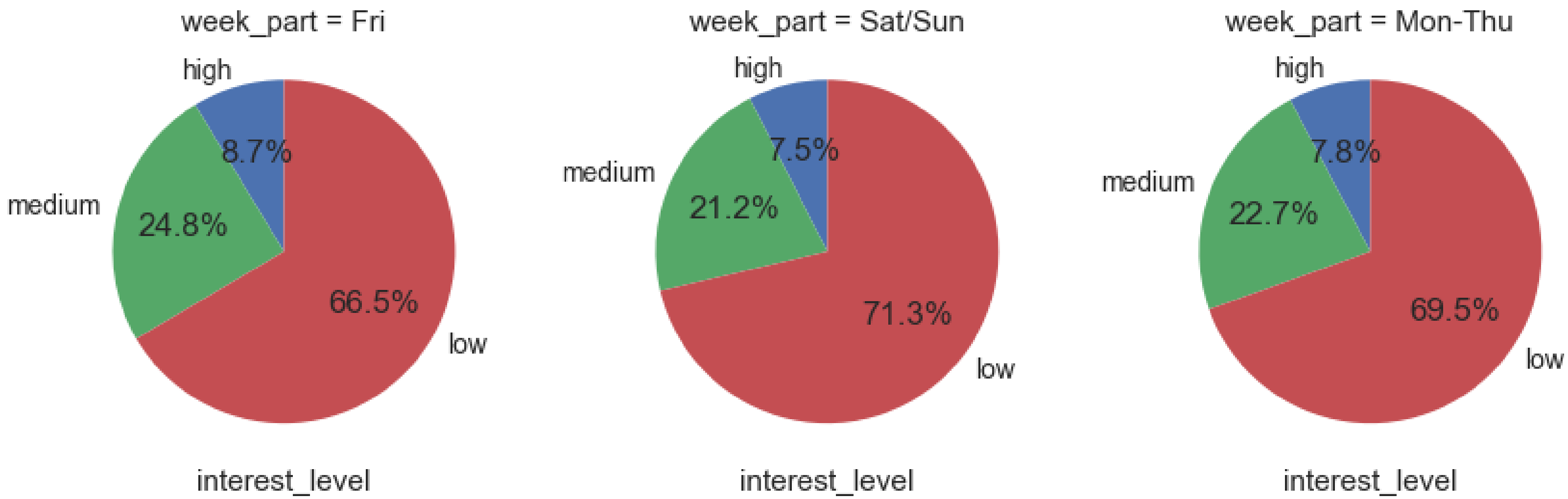
- Used a tf-idf Vectorizer to identify 1-, 2-, 3-grams in the text description
- Min frequency: 1%, max Frequency 98%
- Approximately 800 words in the vocabulary

- Using only the description vector to predict the label – log loss score of 0.74
  - Null Log loss model: scored 0.787

- Created a feature for predicted non-low based on text alone and included it in the main model

Miscellaneous

- Created dummies for day-of-week the listing was created

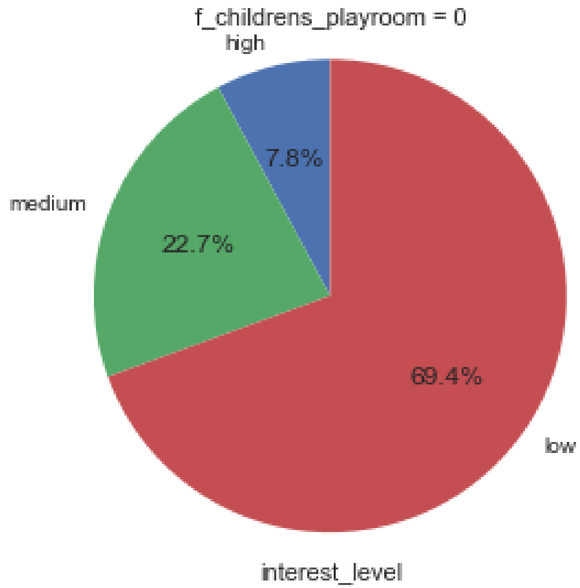


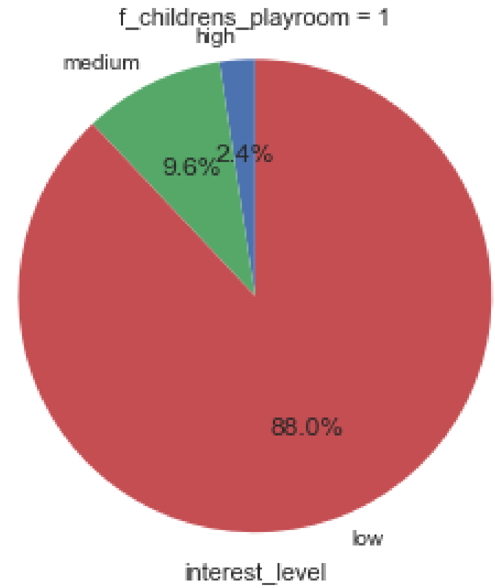
- Created a boost for successful property managers – related to the HOP score and user ratings of the agents
- Created metadata about the listing such as:
  - Percentage Upper Case in the description
  - Log of description length
  - HTML formatting in description; website URL in description
- After feature engineering, new training data shape: 49352 rows x 123 columns + 2 label columns (original categorial converted to integer)

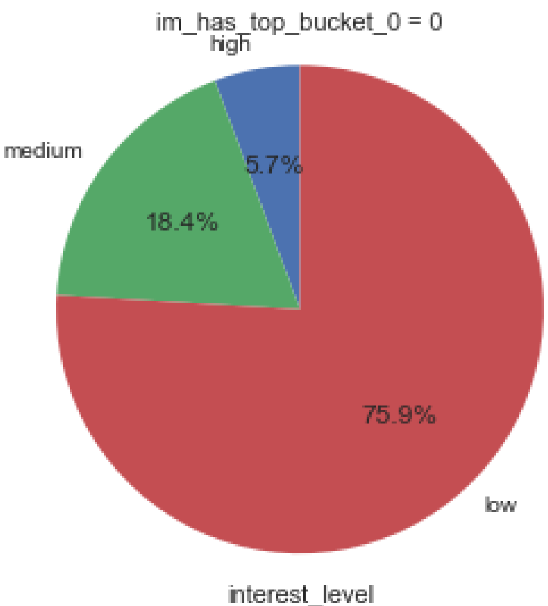
Top Binary Features Selected by MRMR

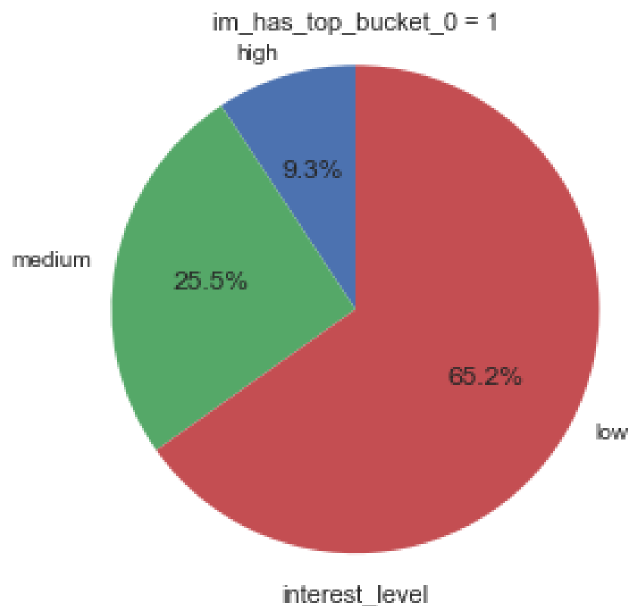
Feature	Feature
Has Building ID	Has More Plain Pictures
No Rental Fee	Display Address Starts with a Digit
Reduced Rental Fee	Has 5 bedrooms
Furnished	Has 1 bedroom
Dogs Allowed	Laundry Room
Renovated	Public Outdoor Space
Hardwood Floors	Created on Friday
Washer in Unit	Fireplace
Has 6 Bedrooms	Newly Renovated
Has 8 Bedrooms	Children’s Playroom

Playroom is a penalty

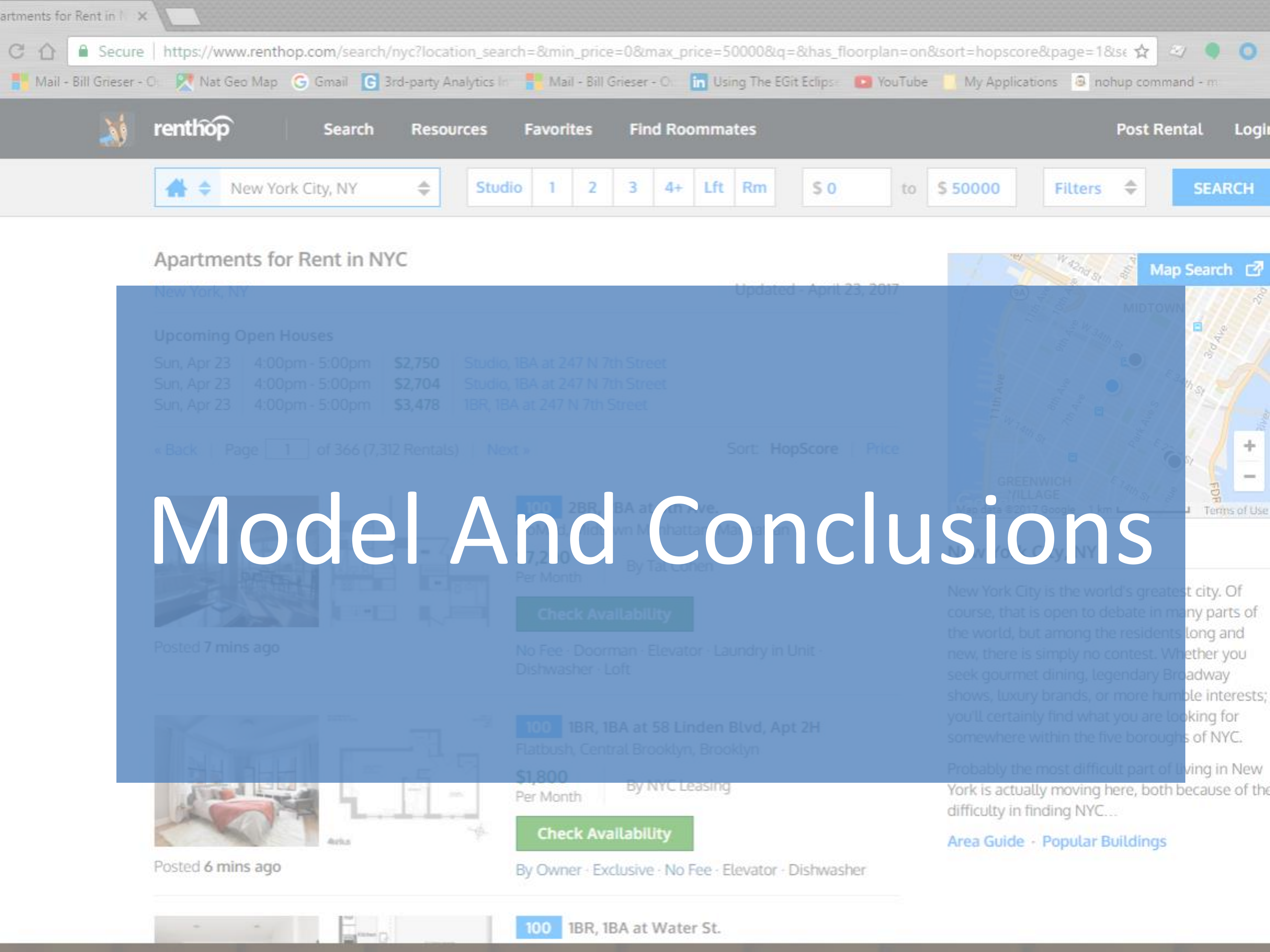








Plain pictures are a boost

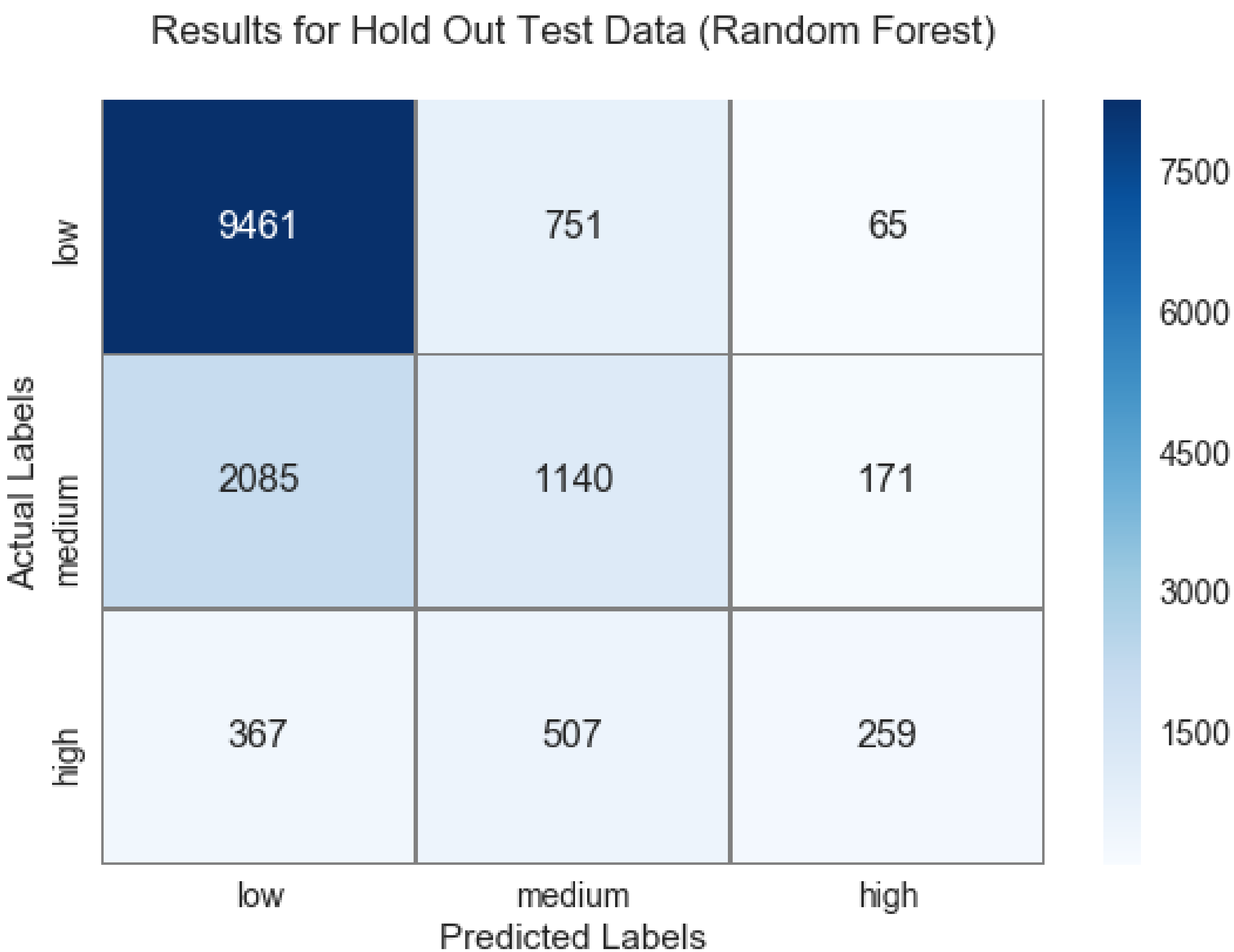


Model And Conclusions

Results of Modeling

- Three best performing models:
  - Random Forest (1000 classifiers, auto n_features)
  - MLP Neural Network (1 hidden layer 10 nodes)
  - Logistic Regression (C=1, l2 normalization, multinomial)

Random Forest

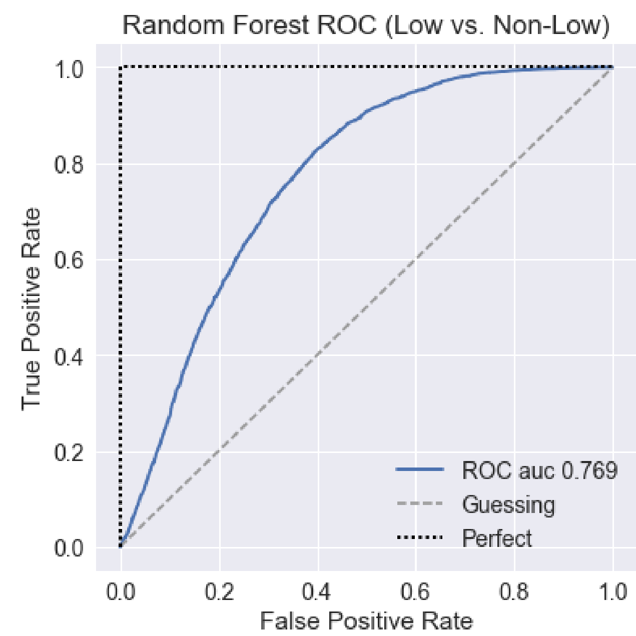


Random Forest

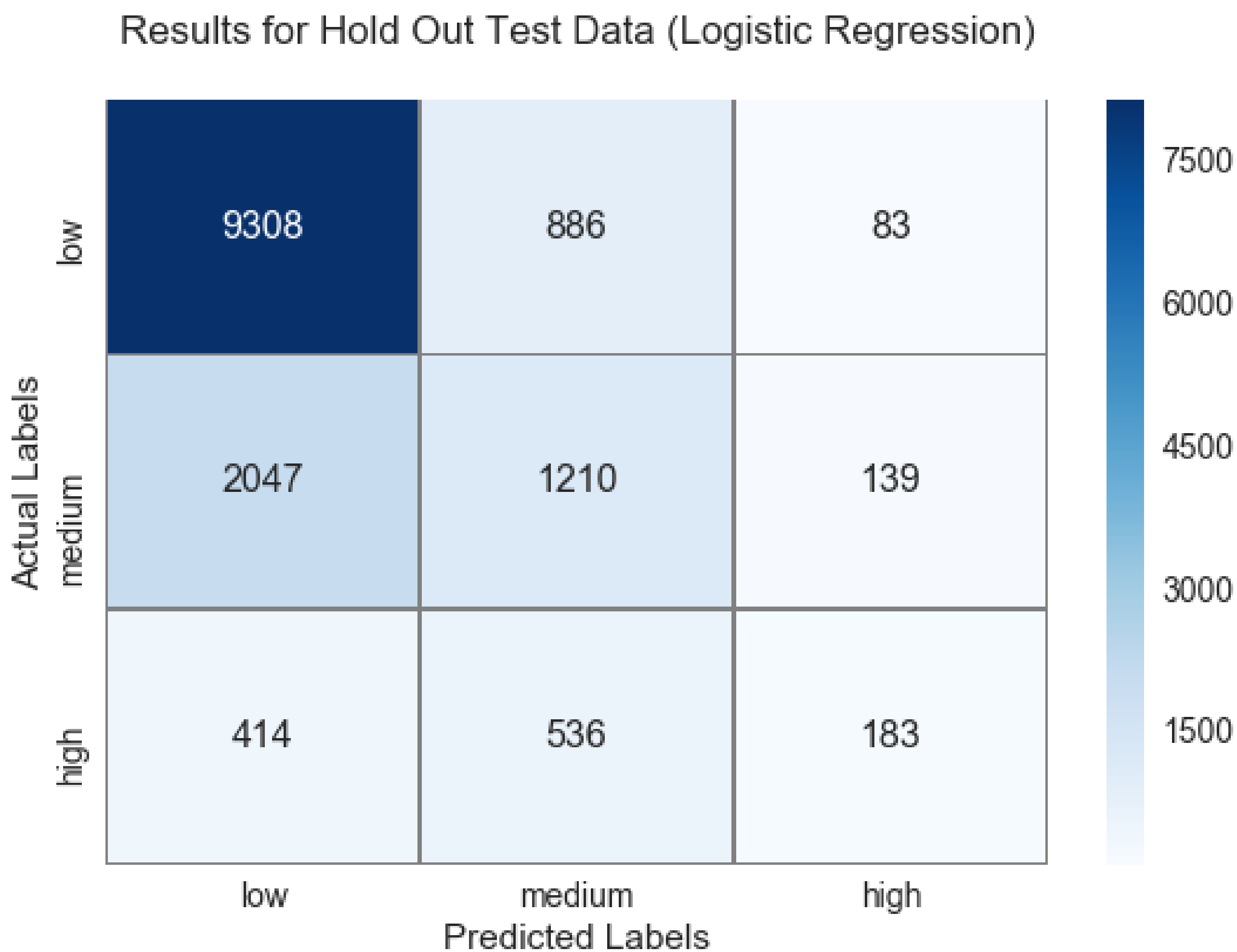
```
Feature importance:
manager_boost 0.207346114229
price 0.094255241664
desc_prob_0 0.0934663652871
neighborhood_boost 0.0674976774421
surplus_DT 0.0663217209287
m_description_length_modlog 0.0658757370221
m_description_UC_pct 0.0649307769437
grid_y 0.0468302964799
listing_feature_count_modlog 0.0453032392492
m_photos_count_modlog 0.0440506134713
grid_x 0.0434702193913
m_has_building_id 0.0208387698602
created_on_wed 0.0110780265638
bedroom_2 0.0102545084573
created_on_tue 0.0102157632852
created_on_thu 0.00992834795607
created_on_fri 0.00974111421575
bathrooms 0.0094823578633
im_has_bucket_1 0.00941038349117
bedroom_1 0.00940995126481
created_on_sat 0.00894835240923
im_has_top_bucket_0 0.00780621686129
m_display_addr_digit_start 0.00768078551956
im_has_top_bucket_1 0.00728999076254
bedroom_3 0.00717192996728
created_on_sun 0.00652371449416
im_has_bucket_0 0.00636962516371
im_has_floorplan 0.00483609621372
bedroom_4 0.00314354338473
bedroom_5 0.000476242460198
bedroom_6 4.61494481599e-05
bedroom_8 1.28249479584e-07
bedroom_7 0.0
```

- Cross-validating Grid Search selected the non-zero probability from the description over the Listing Features list – but it was very close
- Best run against hold-out test data was with the list of features

```
Random Forest

Log loss: 0.593958277915
F1 score: 0.706492740422

[[9461  751   65]
 [2085 1140  171]
 [ 367  507  259]]
Random Forest Adjusted)
```



Logistic Regression

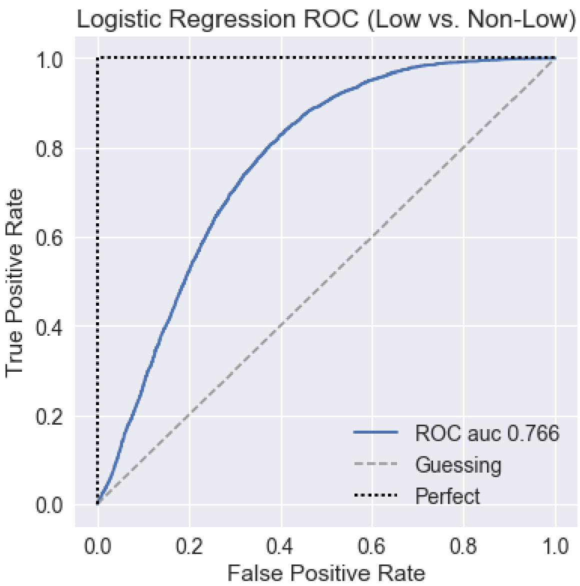


Logistic Regression

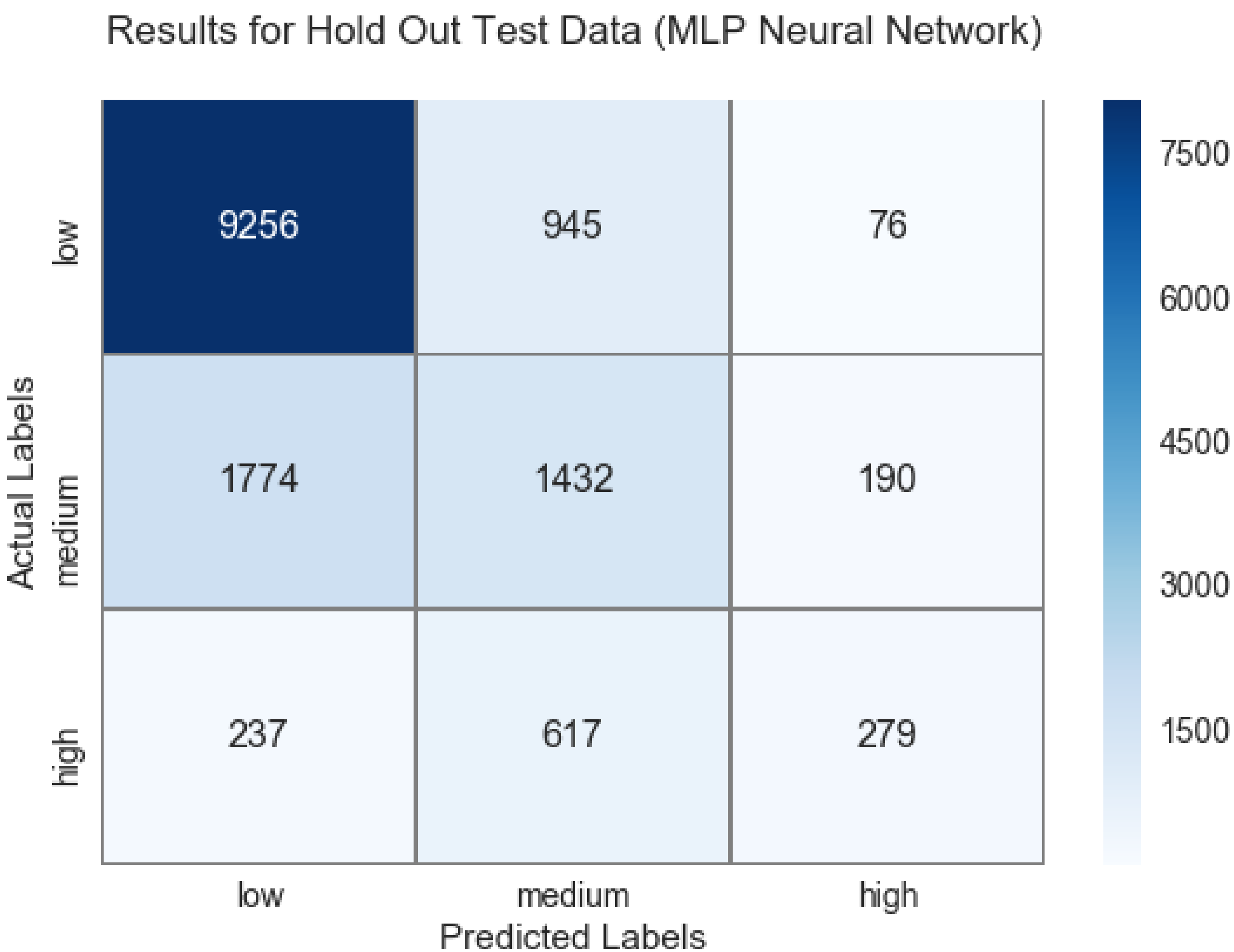
```
Coefficients:
bedroom_4          -1.38585392671
manager_boost      -1.24141179436
f_subway           1.13222479516
f_furnished        -0.978486168256
price              0.837056826108
bedroom_3          -0.783683585396
im_has_top_bucket_1   -0.769325882987
im_has_top_bucket_0   -0.722079317938
m_has_building_id   -0.704392244677
f_simplex          0.530237447817
bedroom_5          0.451675279184
f_onsite_garage    0.409496586251
f_newly_renovated   0.407701458373
f_reduced_fee      -0.383760570101
bedroom_2          -0.368601305702
f_multilevel       -0.339567376893
f_no_fee           -0.31637979972
f_common_outdoor_space   -0.305412302998
f_renovated        -0.295512834513
f_stainless_steel_appliances   -0.288331906157
f_exclusive        -0.287969207776
neighborhood_boost   -0.26352562922
f_concierge        0.240088720421
f_laundry_room     0.226048576103
f_laundry_in_building   -0.211937225452
f_private_outdoor_space   -0.208576042196
f_bike_room        0.20063921899
f_live_in_super    0.190363592592
f_washer_in_unit   0.188912537562
f_washerdryer      -0.188478297162
```

- Predicting the negative case, so negative coefficients are related improved success

```
Logistic Regression

Log loss: 0.620872752432
F1 score: 0.696409884514

[[9308  886   83]
 [2047 1210  139]
 [ 414  536  183]]
```



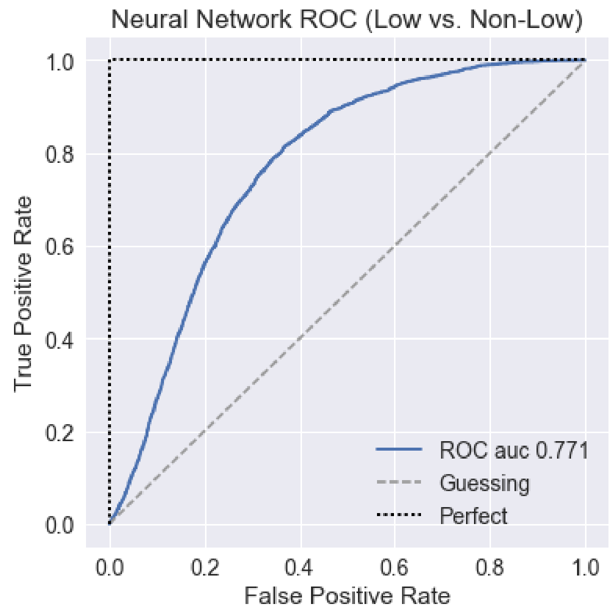
MLP Neural Network



MLP Neural Network

- Log Loss almost as good as Random Forest
- Did the best on F1 score
- Grid Search selected the simplest network (1 hidden layer, 10 modes)

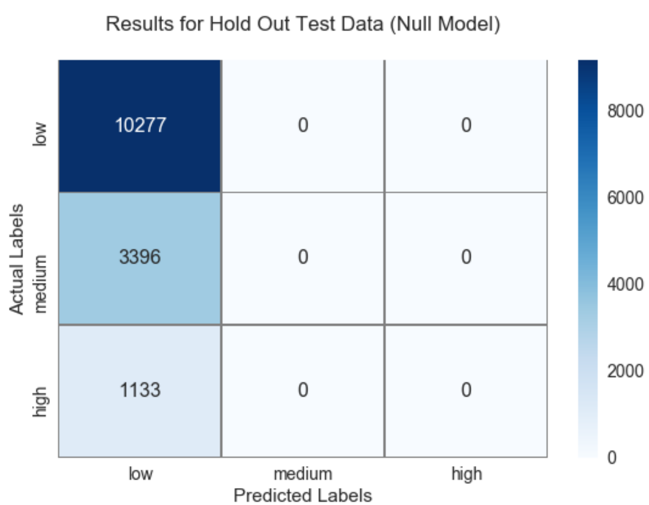
```
Neural Network

Log loss: 0.605959498991
F1 score: 0.724673686007

[[9256  945   76]
 [1774 1432  190]
 [ 237  617  279]]
```

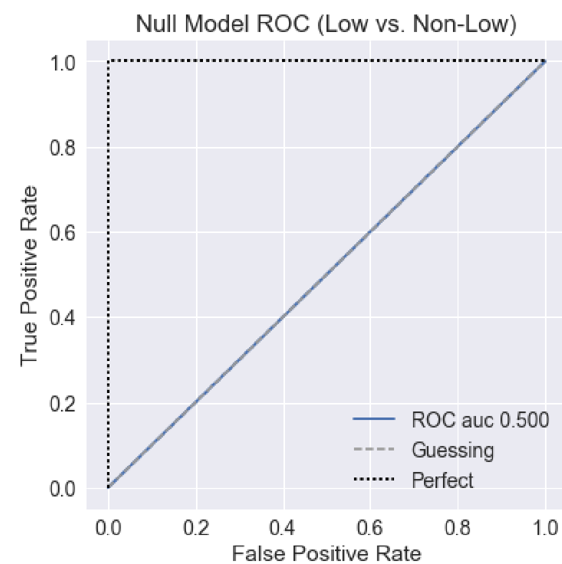


Null Model

(Predicting every listing has the same probabilities as the population average)

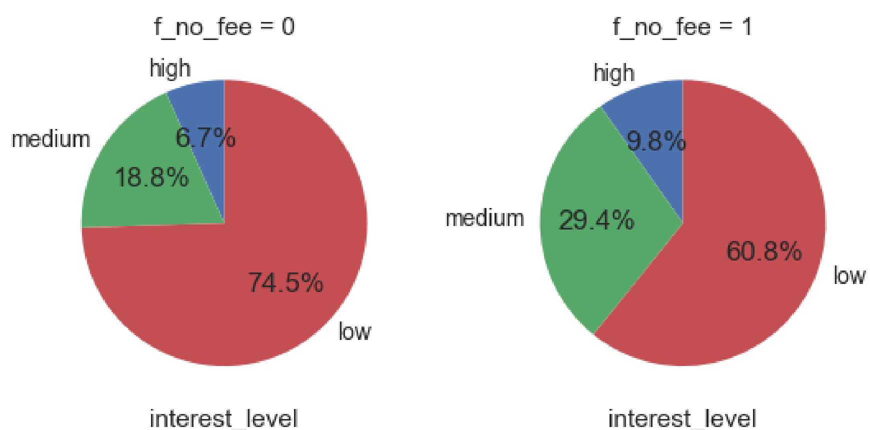
```
Null Model

Log loss: 0.787858212727
F1 score: 0.5687815305

[[10277     0     0]
 [ 3396     0     0]
 [ 1133     0     0]]
```





Insights

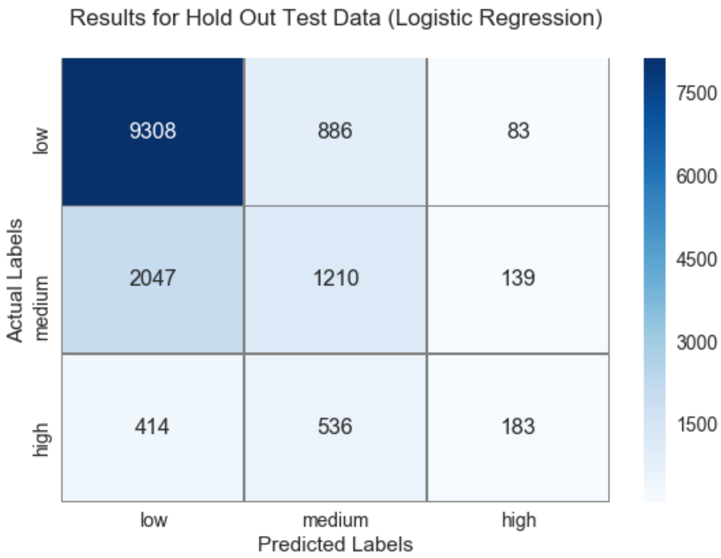
- For listings, pictures are important. At least one should be a picture of the apartment
- Try and have the listing hit on a Friday
- Networking / Social Profile is important for agents
- Keep the price down, especially the rental fee
- The description may not be that important if the pictures and feature list tell the story. There were very close results substituting description score for the list of binary features

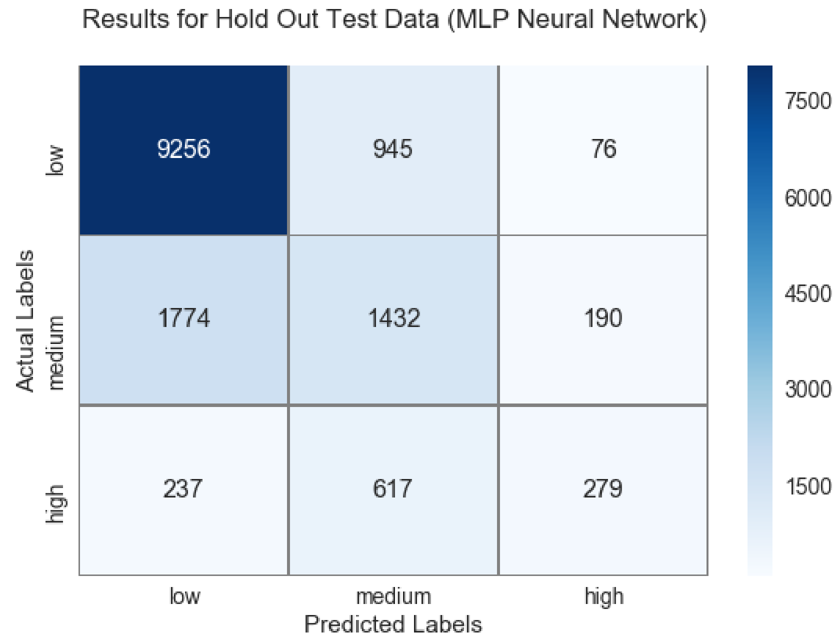


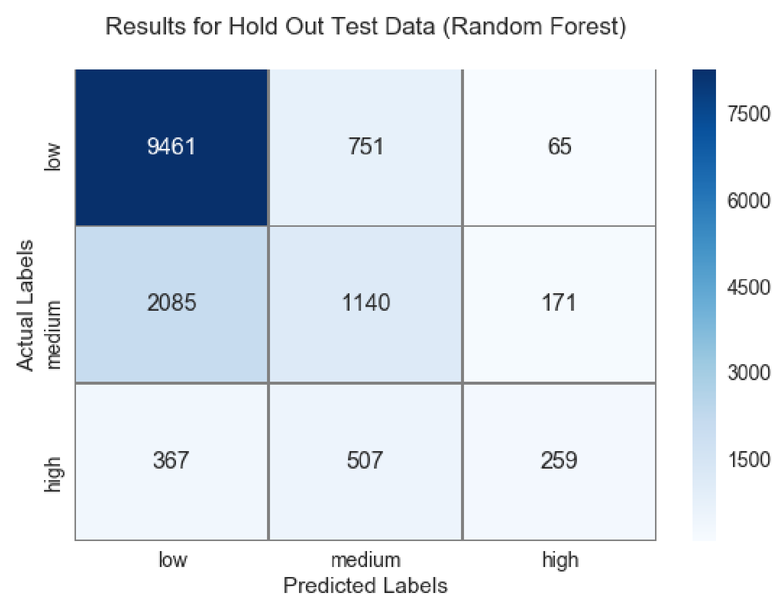
Further Study

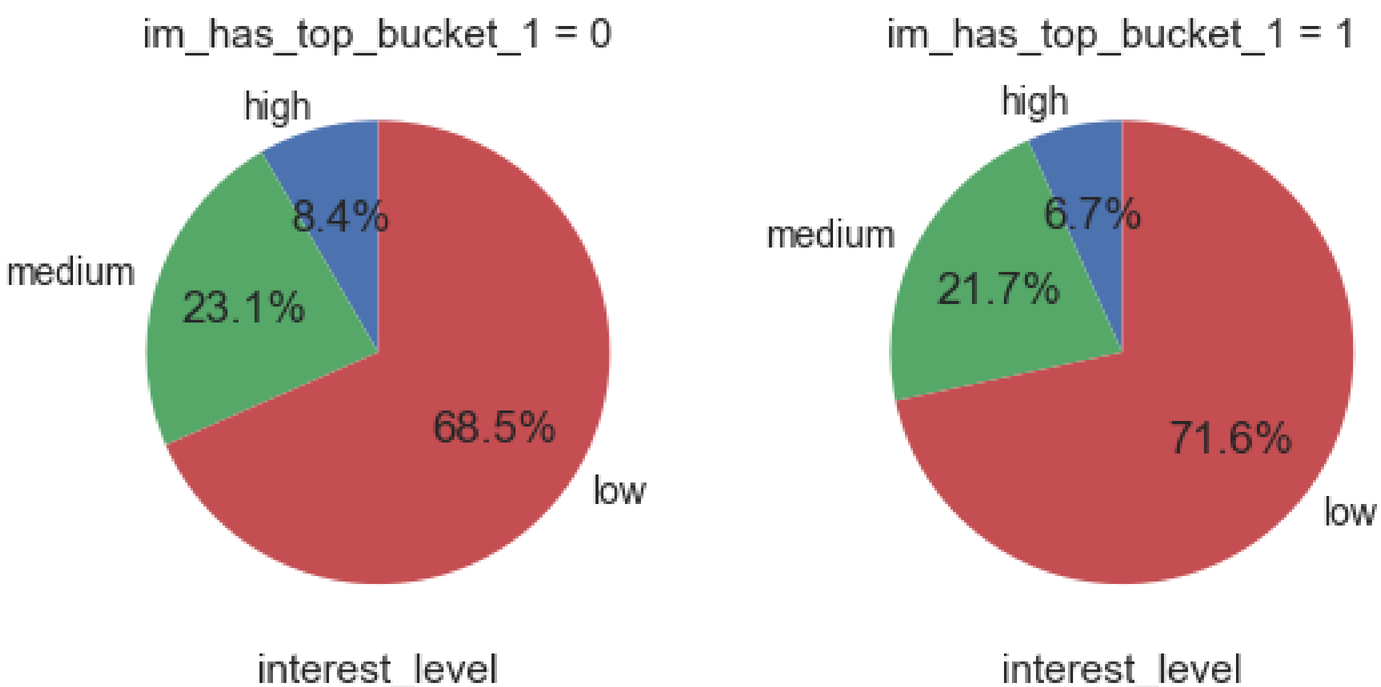
- Images
  - Identify kitchens and bathrooms
- Description
  - More sophisticated analysis needed
- More data on the relationship to the property manager would be desirable
- Model of Models

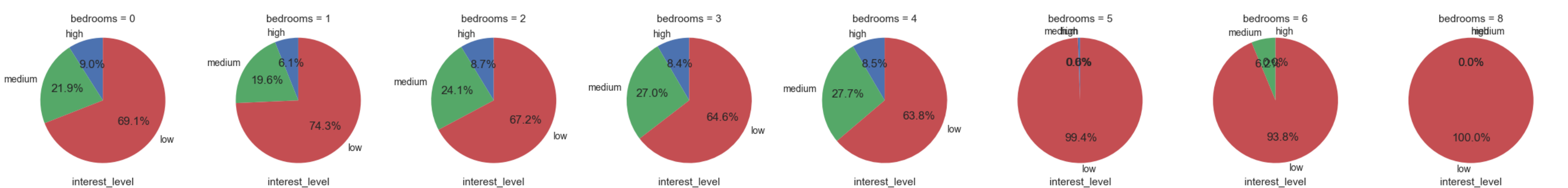
Backup Charts


Results for Hold Out Test Data (Logistic Regression)
9308
886
83
2047
1210
139
414
536
183
low
medium
high
Actual Labels
low
medium
high
Predicted Labels
7500
6000
4500
3000
1500


Results for Hold Out Test Data (MLP Neural Network)
9256
945
76
1774
1432
190
237
617
279
low
medium
high
Actual Labels
low
medium
high
Predicted Labels
7500
6000
4500
3000
1500


Results for Hold Out Test Data (Random Forest)
9461
751
65
2085
1140
171
367
507
259
low
medium
high
Actual Labels
low
medium
high
Predicted Labels
7500
6000
4500
3000
1500


im_has_top_bucket_1 = 0
high
medium
8.4%
23.1%
68.5%
low
interest_level
im_has_top_bucket_1 = 1
high
medium
6.7%
21.7%
71.6%
low
interest_level


bedrooms = 0
high
medium
9.0%
21.9%
69.1%
low
interest_level
bedrooms = 1
high
medium
6.1%
19.6%
74.3%
low
interest_level
bedrooms = 2
high
medium
8.7%
24.1%
67.2%
low
interest_level
bedrooms = 3
high
medium
8.4%
27.0%
64.6%
low
interest_level
bedrooms = 4
high
medium
8.5%
27.7%
63.8%
low
interest_level
bedrooms = 5
99.4%
low
interest_level
bedrooms = 6
medium
high
93.8%
low
interest_level
bedrooms = 8
0.0%
100.0%
low
interest_level