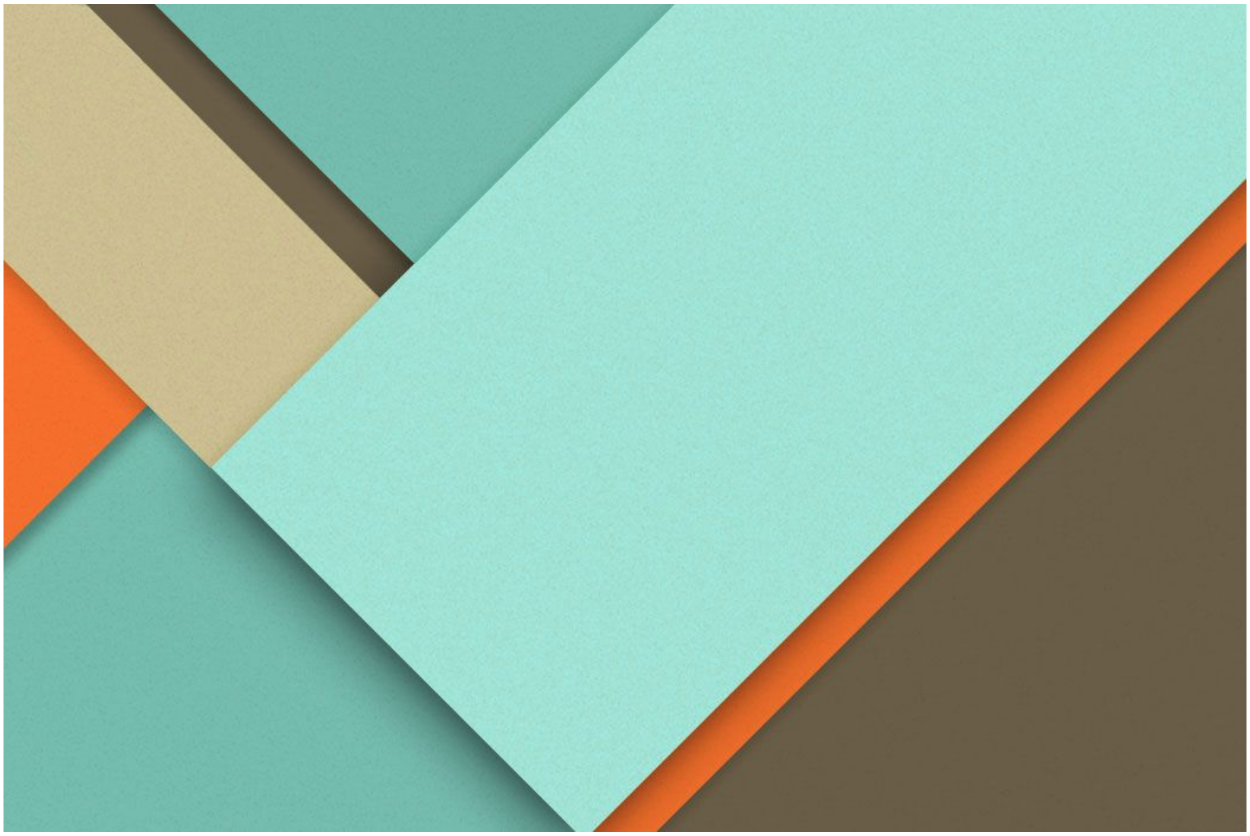


Final Project:

Two Sigma Connect

04.03.2017

Himanshu Makhija (A09845605) -- Math/Computer Science, B.S.
Hanna Goldman (A11436767) -- Math/Computer Science, B.S.
Shagun Gupta (A91068956) -- Bioinformatics, B.S.
Amey Paranjape (A53218045) -- ECE, M.S.

Table of Contents

Introduction

1.1 Background

From the kaggle competition website:

> "Two Sigma invites you to apply your talents in this recruiting competition featuring rental listing data from RentHop. Kagglers will predict the number of inquiries a new listing receives based on the listing's creation date and other features. Doing so will help RentHop better handle fraud control, identify potential listing quality issues, and allow owners and agents to better understand renters' needs and preferences."

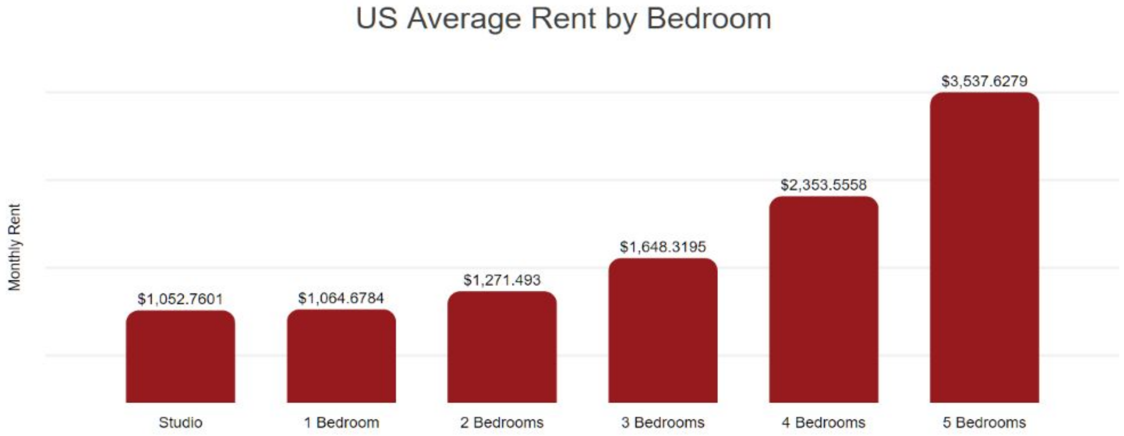
Buying or renting real estate properties has been a one of the toughest as well as critical decisions over the years, as it affects almost every aspect of human life. Usually, features that one sees in the house differ person to person. But there are a few common features that almost everyone looks for. There can be a few minor variations according to personal needs. It is a well-known fact that apartment hunting is very difficult and time consuming process. So often people tend to use services which help them to find an apartment that is ideal according to their needs. Also often people tend to have a look at the statistics available to get an idea of the current market trends.

Usually following are the parameters which people set as the look to rent an apartment:

- Location of property (Building, Street, Neighborhood, Area etc.)
- Monthly rent
- Number of rooms (bedrooms, bathrooms etc.)
- Neighboring communities
- Nearby supermarkets, eateries, medical services, schools, public transit centers, corporate offices etc.
- Garage parking or storage
- Other amenities like swimming pool, gym, Wi-Fi access etc.
- Access to freeways, state highways etc.
- Allowance for pets, rent for pets
- Furnished or unfurnished property

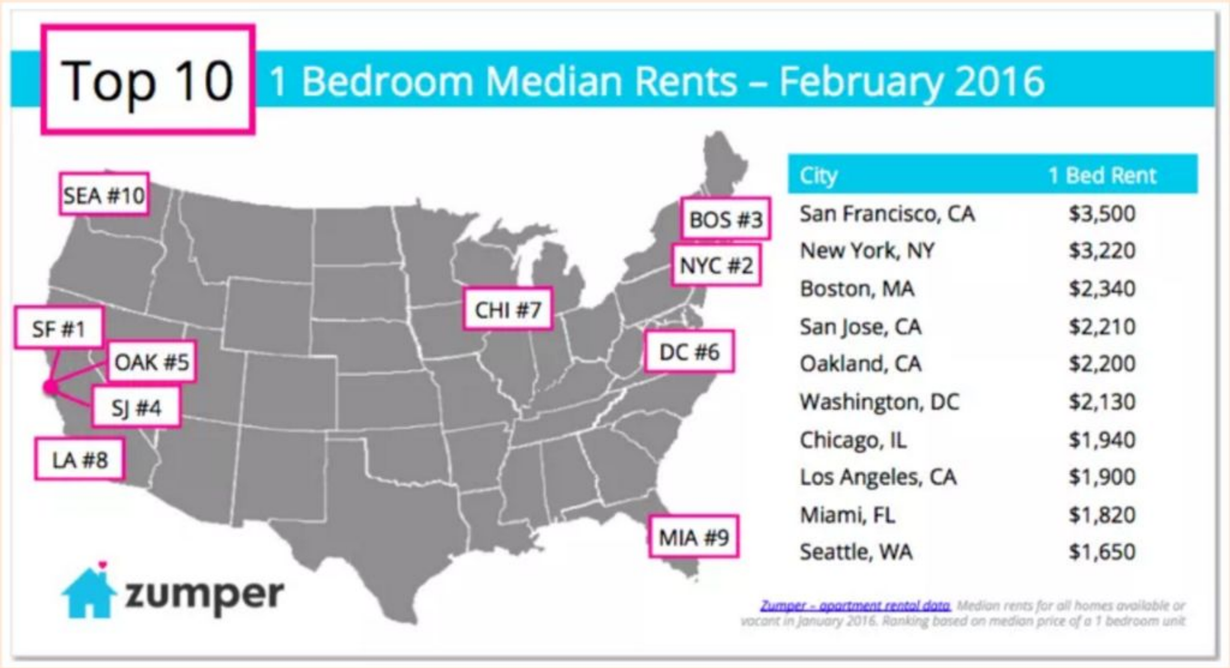
There are a lot of agencies which publish data/statistics related to apartment rental in various cities of United States of America. Some of them represent the real trends but also some of them are made up numbers for some organizational benefits. Both journalists and consumers should be careful while considering or using any of such available data. One should look at the source of such data and then only trust it.

The website www.abodo.com, has published statistics for average monthly rental in the United States by bedroom.



These numbers are very generic and they change according to the state, city, neighborhood and street. Usually apartments on streets or neighborhoods which are closer to corporate offices are more pricey. Apartments near famous tourist places, landmarks etc. also have high rent. Despite having high property rents, people are interested in renting such places due to obvious reasons.

To get an idea about how the rents differ from city to city many agencies publish their data according to cities. States with more number of corporate offices and industries such as California, New York, Massachusetts etc. Are costly in terms of living expenses. The agency 'Zumper' has published some interesting statistics regarding the same.

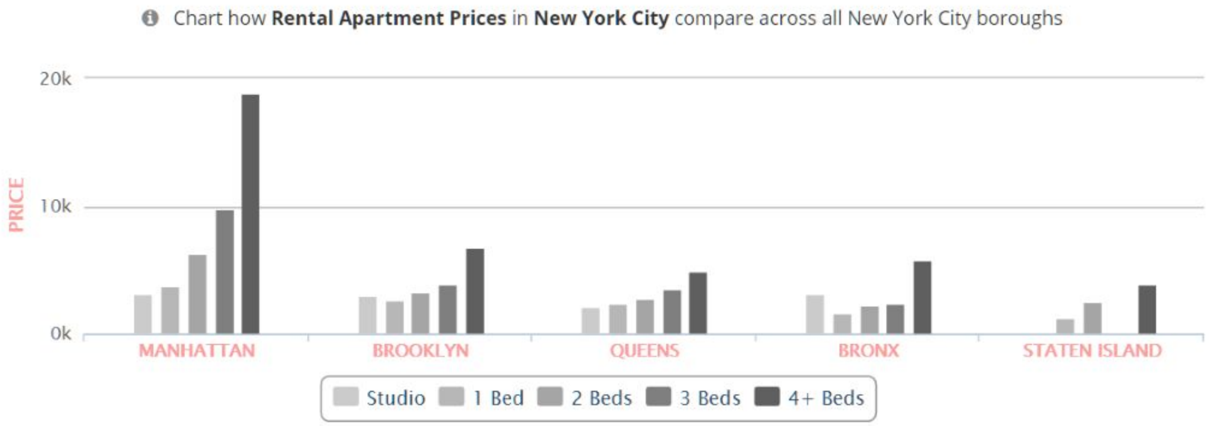

Top 10 1 Bedroom Median Rents – February 2016

City	1 Bed Rent
San Francisco, CA	$3,500
New York, NY	$3,220
Boston, MA	$2,340
San Jose, CA	$2,210
Oakland, CA	$2,200
Washington, DC	$2,130
Chicago, IL	$1,940
Los Angeles, CA	$1,900
Miami, FL	$1,820
Seattle, WA	$1,650

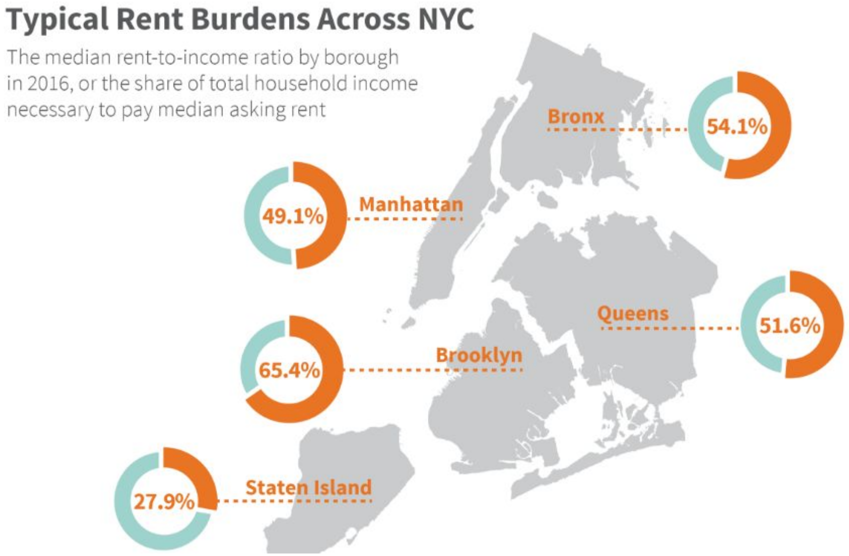
Zumper - apartment rental data. Median rents for all homes available or vacant in January 2016. Ranking based on median price of a 1 bedroom unit

An interesting thing to note in this data is that it is a good practice to consider median instead of mean, while observing such data, because outliers have more direct effect on mean than on median. So, median gives us more precise idea than mean, in these cases.

In our case, our data is limited only to New York City apartment rentals. According to the data available on the website of www.blocksy.com, "the average New York apartment rent is $6,473 per month. The average NY apartment is 1,182 square feet and has 1 bedroom and 1 bathroom. The majority of rentals in NYC are in post war buildings. New York City currently has 15033 apartment rentals available (2130 are furnished apartments and the rest are unfurnished apartments). Of these, 2722 apartments are single room studios, 5924 are one bedroom apartments, 4214 are two bedroom apartments, 1643 are three bedroom apartments and the rest are larger. While data is not available for all New York City apartment buildings, 1594 rentals are known to have been constructed or renovated recently and are therefore likely to be relatively newer in appearance and amenities. A number of rental properties in New York City are luxury apartments or contain premium amenities. Among them, 3955 have air conditioning, 2550 offer city views, 4810 have concierge services inside the building, 7103 have a doorman, 6335 have elevator service, 236 have a fireplace, 2130 of the listings are available furnished, 4031 have a gym or fitness center, 4411 have hardwood floors, 2404 boast high ceilings, 308 have a jacuzzi, 6484 offer laundry services inside the building, 5484 have public outdoor space, 391 have park views, 4693 offer parking, 367 are penthouse apartments, 4468 allow pets, 1736 have private outdoor space, 4158 offer storage, 1626 have a swimming pool, 2386 have a washer / dryer inside the home, 1028 have water views." So here we can observe that while searching for an apartment there can be multiple factors which can direct consumers' final decision. Following bar graph shows the variation of rental prices according to the different areas of NY city.



Renting an apartment is also one of the critical decisions because it directly affects household economy. Following is one more interesting statistics that shows the percentage of total income dedicated towards payments of rent.


Typical Rent Burdens Across NYC
The median rent-to-income ratio by borough in 2016, or the share of total household income necessary to pay median asking rent
Bronx
54.1%
Manhattan
49.1%
Queens
51.6%
Brooklyn
65.4%
Staten Island
27.9%

So, we can conclude that, renting an apartment is a very complex problem, depending upon a number of factors. So, people always tend to seek help of different sources such as real estate agents, real estate websites. These sources use previous data to analyze market trends and suggest suitable properties depending on the needs of customers. Technology has always tried to replace these sources with computer software and tools, but no software has yet successfully replaced these real estate industry.

1.2 Goal

Our goal is to find the features of an apartment which play the biggest roles in determining its interest level. Then we wish to create a model which given apartments and their features, predicts an apartment's interest level.

1.3 Data Provided

The data comes from renthop.com, an apartment listing website. These apartments are located in New York City. The target variable, interest_level, is defined by the number of inquiries a listing has in the duration that the listing was live on the site.

File Descriptions:

- train.json - the training set
- test.json - the test set
- sample_submission.csv - a sample submission file in the correct format
- images_sample.zip - listing images organized by listing_id (a sample of 100 listings)
- Kaggle-renthop.7z - (optional) listing images organized by listing_id. Total size: 78.5GB compressed. Distributed by BitTorrent (Kaggle-renthop.torrent).

Data Fields:

- bathrooms: number of bathrooms
- bedrooms: number of bathrooms
- building_id
- created
- description
- display_address
- features: a list of features about this apartment
- latitude
- listing_id
- longitude
- manager_id
- photos: a list of photo links.
- price: in USD
- street_address
- interest_level: this is the target variable. It has 3 categories: ’high’, ’medium’, ’low’

1.4 Hypothesis

Our hypothesis is that the biggest contributions to an interest level for an apartment will be the price and the ratio of bedrooms to bathrooms. There will be a negative correlation between price and interest level, the lower the price of an apartment the higher the interest level it will be.

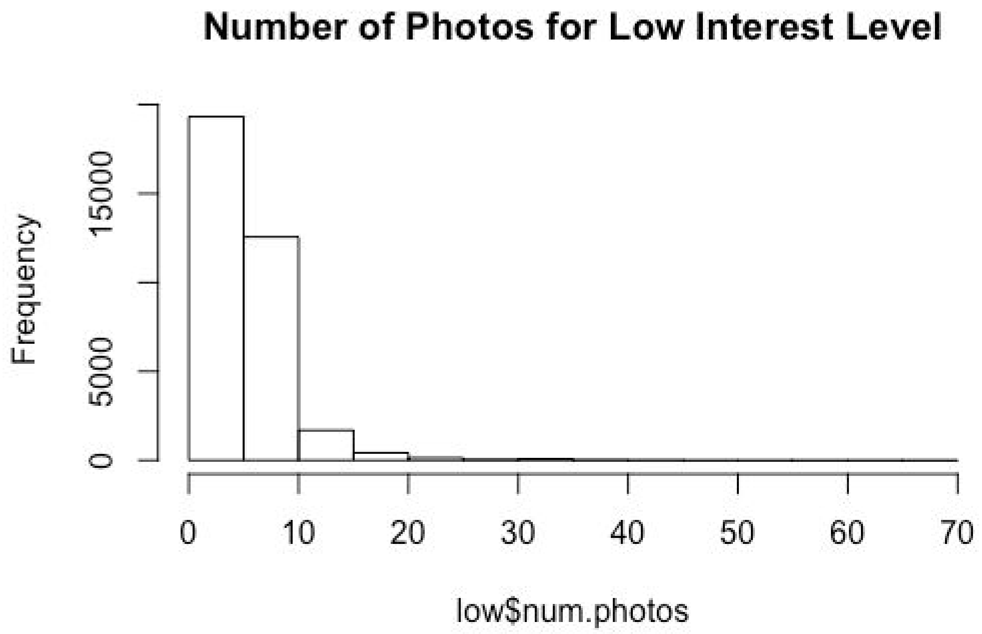
Analysis

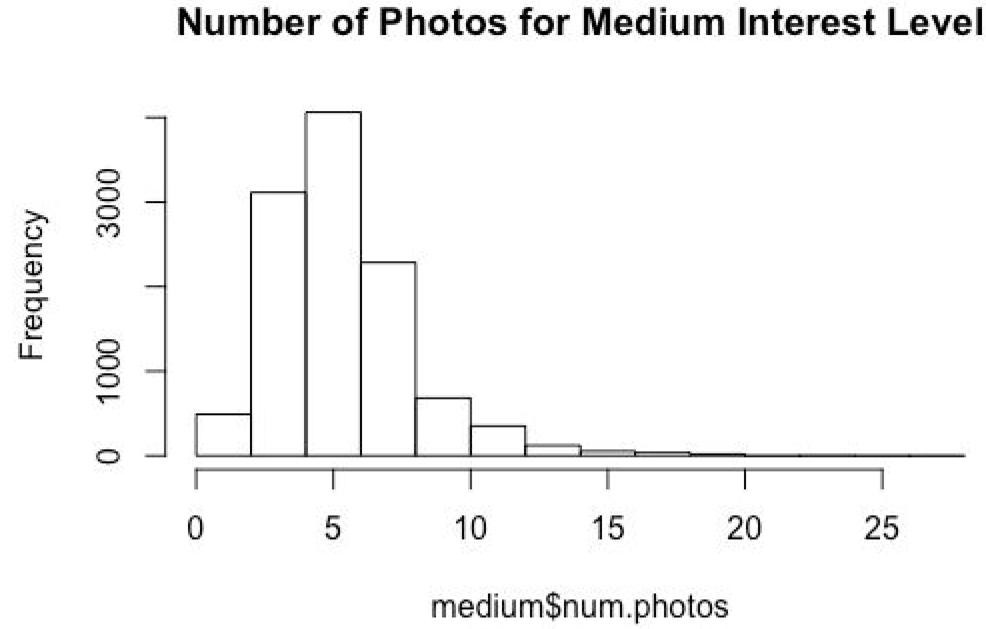
2.1 Description of Data

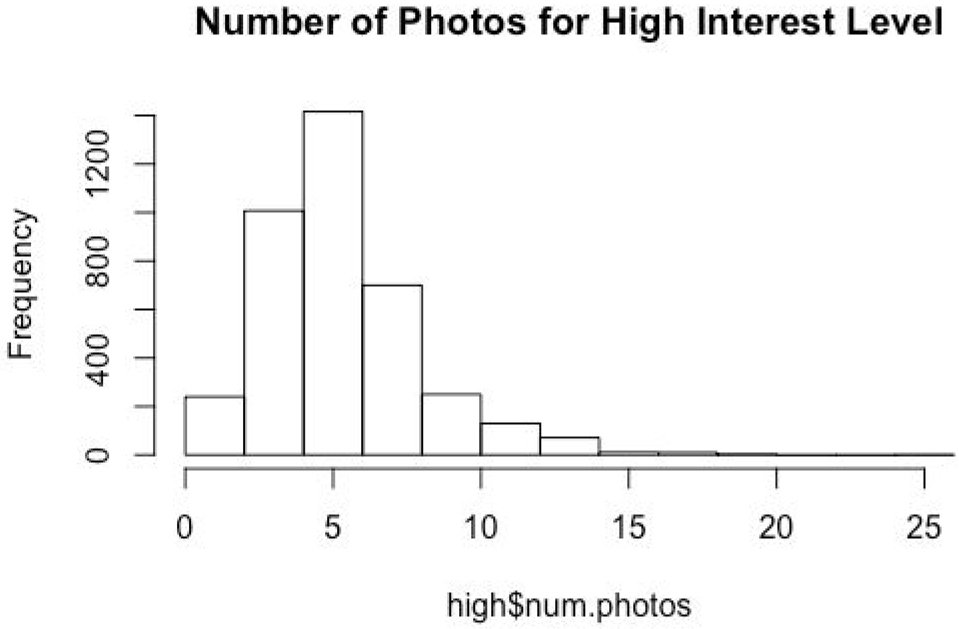
The train data consists of 49352 observations and 15 variables. The test set consists of 74659 observations and 14 variables. The final variable of the train data is interest_level which is not present in the test data set because it needs to be predicted.

The breakdown of the 49352 training data observations is: 34284 low interest level, 11229 medium interest level and 3839 high interest level. Thus if I was given no further information, I would say the probability of data to be low interest level is 0.6943831, the probability for medium interest level is 0.2275288, the probability for high interest level is 0.07778813. We will set this as our benchmark.

2.2 Number of Photos





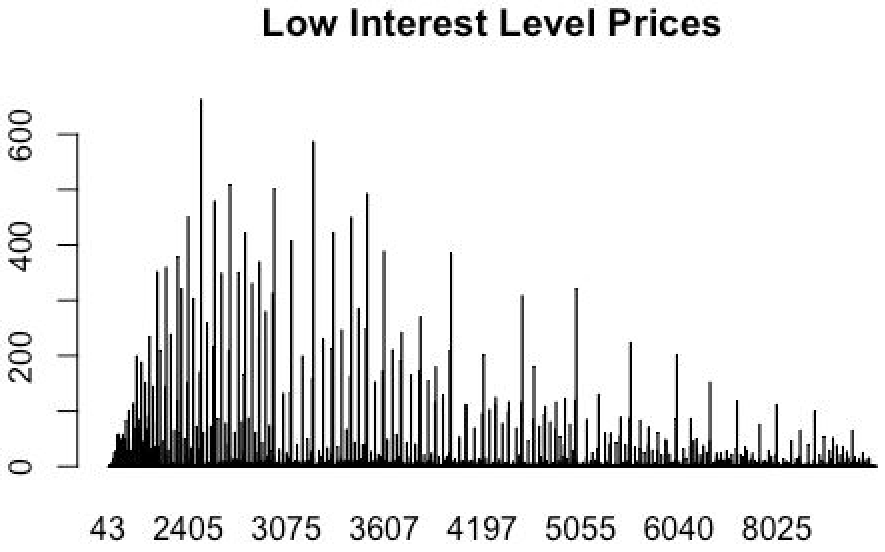


The distribution for the number of photos is similar for all three interest levels. One result we can see though is that the majority of apartments with zero photos tend to be of low interest level. This could be relevant in developing our model.

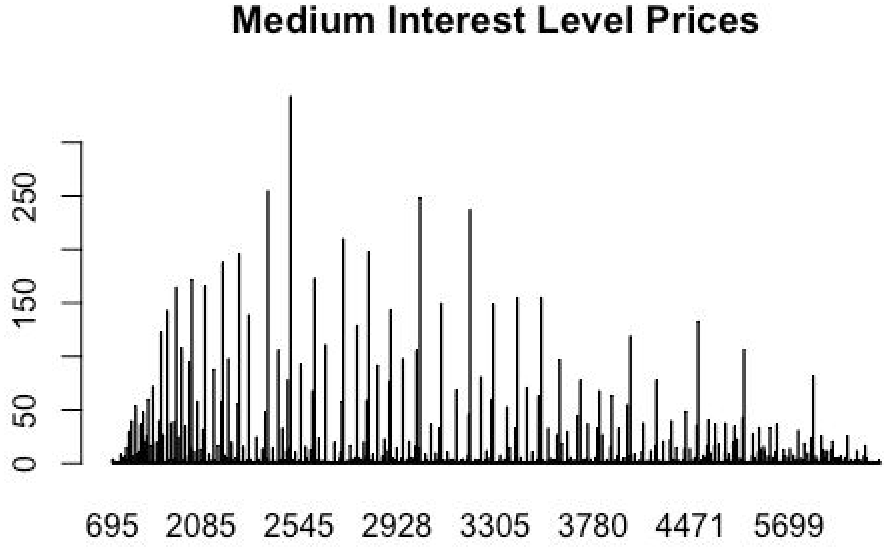
The percentage of apartments with zero photos that are of low interest level are 3442/3615 = .95214385. The percentage that are medium interest level are 123/3615 = 0.0340249. The percentage that are high interest level are 50/3615 = 0.01383126.

Note: We actually find a similar idea with the number of bathrooms. If an apartment has zero bathrooms it had a 0.97763578 chance of being low interest level, a 0.01916933 chance of being medium interest level and a 0.00319489 chance of being high interest level.

2.3 Prices by Interest Level

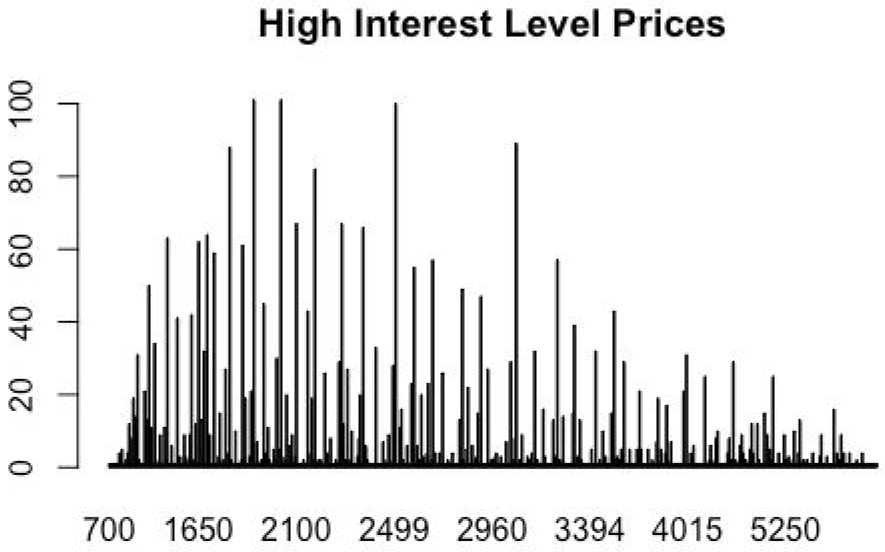


mean = 4176.599
median = 3300



mean = 3158.767

median = 2895



mean = 2700.293
median = 2400

We can see a negative trend that the lower the prices, the higher the interest level they tend to be.

2.4 Managers

If we observe the managers of the apartments in each interest level, we can hope to see whether a certain manager has an effect on the interest level of an apartment.

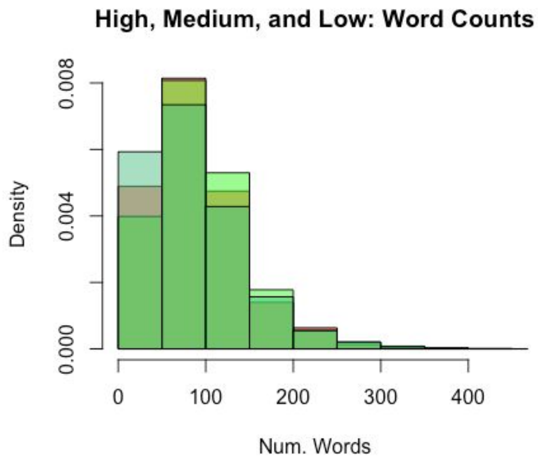
Table of Managers for Low Interest Level
3136 1500 1955 1339 2534 2374 2765 2146 2339 1338
1739 644 405 402 316 280 274 273 232 226

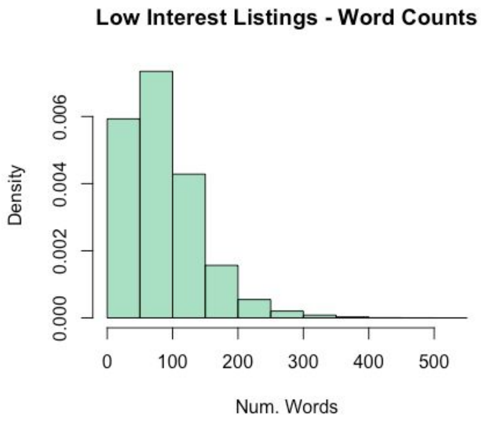
Table of Managers for Medium Interest Level
3136 2785 458 608 2899 3004 3205 1317687 602
622 183 120 115 102 92 85 78 77 71

Table of Managers for High Interest Level
3136 458 2785 262 2605 608 687 3205 602
1804 172 68 55 51 43 38 36 34 32 32

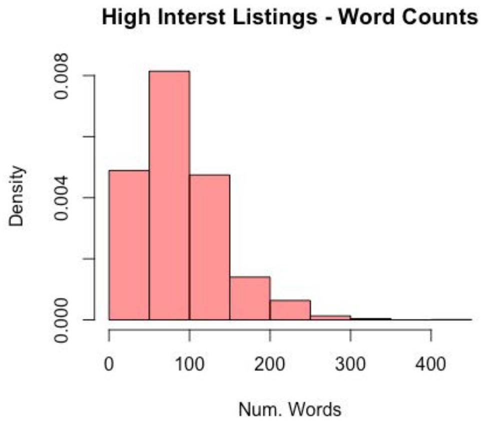
We can see that the same manager, id number 3136, has highest number of apartments for all three interest levels.

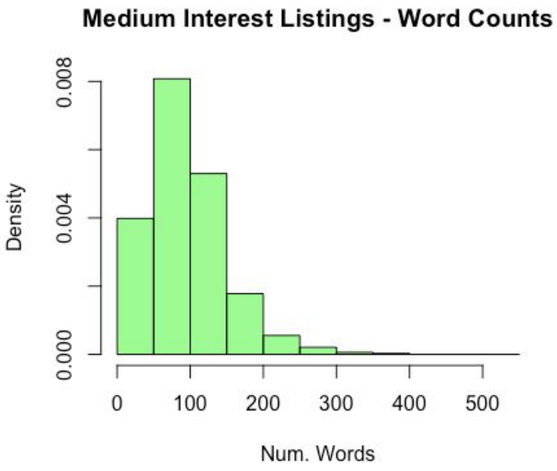
2.5 Listing Description





→ Mean: 84.66 Words/Description
→ Median: 78.00





→ Mean: 88.6 Words/Description
→ Median: 81.00

→ Mean: 94.91 Words/Description
→ Median: 88.00

The main idea here is to investigate if the number of words in a description can influence the ranking given for the listing. The 49352 entries were split into their respective trisection (low, medium, and high interest) and plotted against density.

The first graph displays similarity across the three distributions. However, we can see that the low interest histogram has a higher density of descriptions of lengths less than 50. Therefore, it is more likely that if a listing has a short description the interest will also be low.

Conclusion

In conclusion we predict that the greater the number of photos associated with the listing, the more likely it is to generate interest. While we didn’t expect the number of photos to have such an effect on the interest level, it does follow because from our readings into the difficulties and complexities of renting in New York. As expected, prices were found to have a negative correlation with interest. Managers with a history of closing deals or being involved in large number of these listings were found to be more likely to be trusted by the public for handling of their listing requests. The greater in length and more in depth is the associated description for each listing, the more interest it is likely to garner from the public. We decided that the biggest factors were bathrooms, bedrooms, price and number of photos with each listing. These features were then subsequently used in multinomial regression to generate our submission file. Results are attached in the csv file attached with the report.

Theory

3.1 Multinomial Linear Regression

Multinomial Logistic Regression is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Thus it is an extension of logistic regression, which analyzes dichotomous (binary) dependents. Since the SPSS output of the analysis is somewhat different to the logistic regression output, multinomial regression is sometimes used instead.

Like all linear regressions, the multinomial regression is a predictive analysis. Multinomial regression is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous-level(interval or ratio scale) independent variables.

Standard linear regression requires the dependent variable to be of continuous-level(interval or ratio) scale. Logistic regression jumps the gap by assuming that the dependent variable is a stochastic event. And the dependent variable describes the outcome of this stochastic event with a density function (a function of cumulated probabilities ranging from 0 to 1). Statisticians then argue one event happens if the probability is less than 0.5 and the opposite event happens when probability is greater than 0.5.

Our interest level is a nominal dependent variable.

How do we get from logistic regression to multinomial regression? Multinomial regression is a multi-equation model, similar to multiple linear regression. For a nominal dependent variable with k categories the multinomial regression model estimates k-1 logit equations. Although SPSS does compare all combinations of k groups it only displays one of the comparisons. This is typically either the first or the last category. The multinomial regression procedure in SPSS allows selecting freely one group to compare the others with.

What are logits? The basic idea behind logits is to use a logarithmic function to restrict the probability values to (0,1). Technically this is the log odds (the logarithmic of the odds of y = 1). Sometimes a probit model is used instead of a logit model for multinomial regression. The following graph shows the difference for a logit and a probit model for different values (-4,4). Both models are commonly used as the link function in ordinal regression. However, most multinomial regression models are based on the logit function. The difference between both functions is typically only seen in small samples because probit assumes normal distribution of the probability of the event, when logit assumes the log distribution.

Works Cited

- http://stackoverflow.com/questions/8920145/count-the-number-of-words-in-a-string-in-r
    - Helped to count number of words in a string
- http://www.statisticssolutions.com/mlr/
- https://www.r-bloggers.com/how-to-multinomial-regression-models-in-r/
- www.abodo.com
- www.zumper.com
- www.blocksy.com