Two Sigma Connect: Rental Listing Inquiries

Goal:

How much interest will a new rental listing on RentHop receive? you will predict how popular an apartment rental listing is based on the listing content like text description, photos, number of bedrooms, price, etc. 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.

Finding the perfect place to call your new home should be more than browsing through endless listings. RentHop makes apartment search smarter by using data to sort rental listings by quality. But while looking for the perfect apartment is difficult enough, structuring and making sense of all available real estate data programmatically is even harder. You 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.

Dataset:

Data is freely available for non commercial use at

1. Rental Listing Dataset
2. Crime Map of NYC Dataset

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. You are welcome to download the pictures yourselves from renthop's site, but they are the same as imgs.zip.
- price: in USD
- street_address

- interest_level: this is the target variable. It has 3 categories: 'high', 'medium', 'low'

Exploratory data analysis and Inferential statistics:

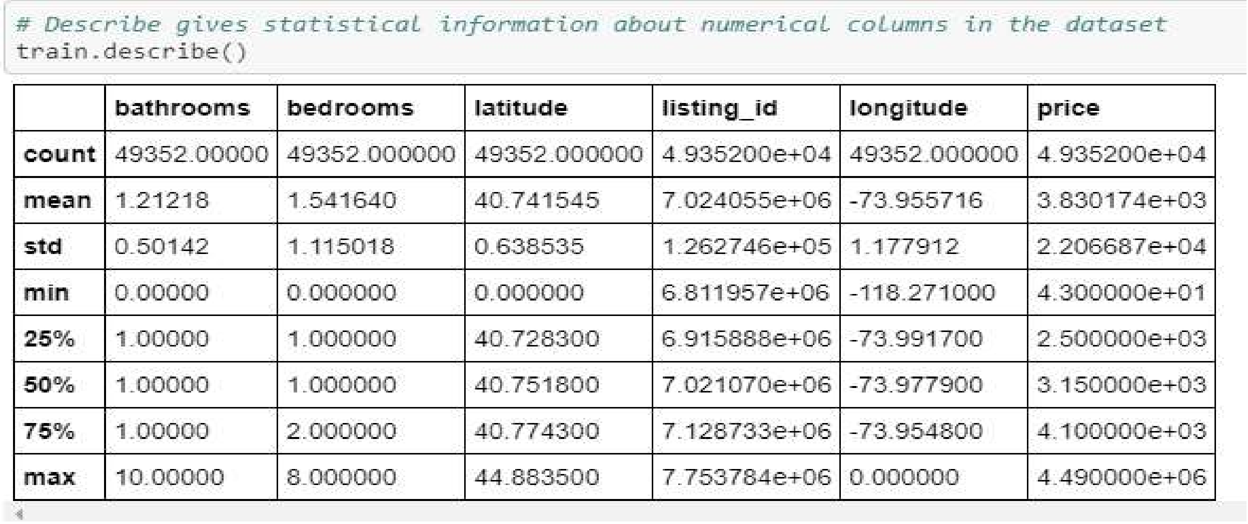
Links: Exploratory Data Analysis Jupyter Notebook

Most of the data that I received was cleaned and ready to use.

To make things easier for inferential statistics I took these steps:

1) converted interest level to a categorical variable
2) added photos_count
3) features_count variable
4) I also converted longitude, latitude to associate zip code using google's reverse geocoding API.

Testing dataset has 49352 records.



```
# Describe gives statistical information about numerical columns in the dataset
train.describe()
```

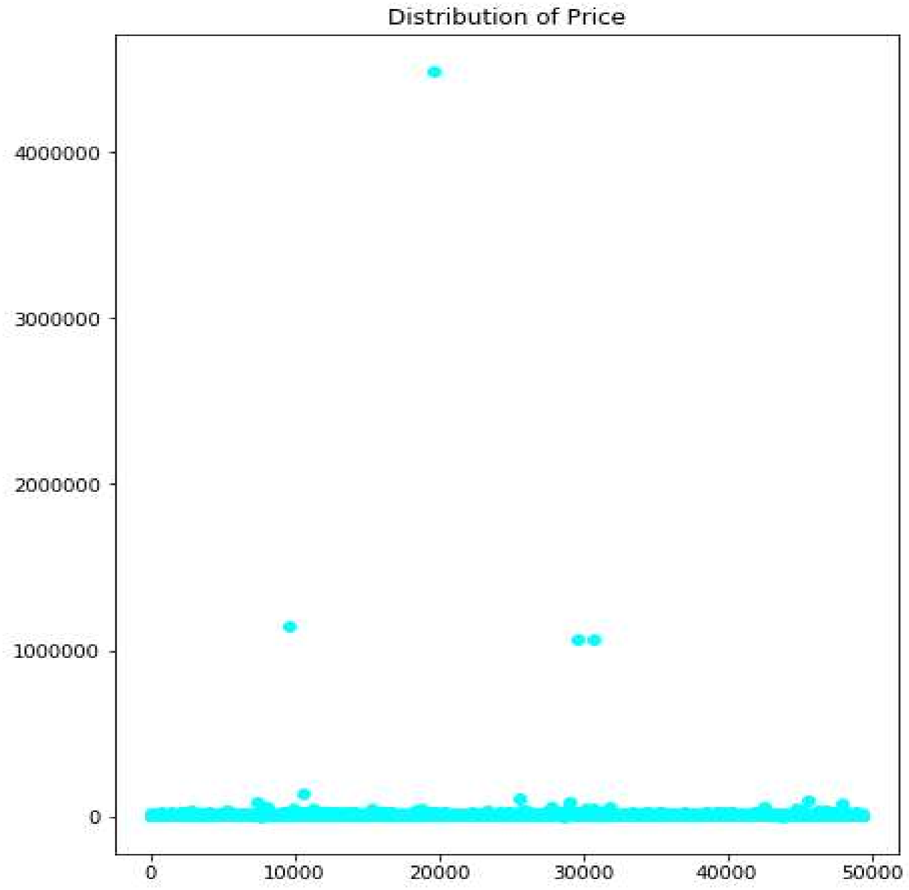
	bathrooms	bedrooms	latitude	listing_id	longitude	price
count	49352.00000	49352.000000	49352.000000	4.935200e+04	49352.000000	4.935200e+04
mean	1.21218	1.541640	40.741545	7.024055e+06	-73.955716	3.830174e+03
std	0.50142	1.115018	0.638535	1.262746e+05	1.177912	2.206687e+04
min	0.00000	0.000000	0.000000	6.811957e+06	-118.271000	4.300000e+01
25%	1.00000	1.000000	40.728300	6.915888e+06	-73.991700	2.500000e+03
50%	1.00000	1.000000	40.751800	7.021070e+06	-73.977900	3.150000e+03
75%	1.00000	2.000000	40.774300	7.128733e+06	-73.954800	4.100000e+03
max	10.00000	8.000000	44.883500	7.753784e+06	0.000000	4.490000e+06

Remove outliers:

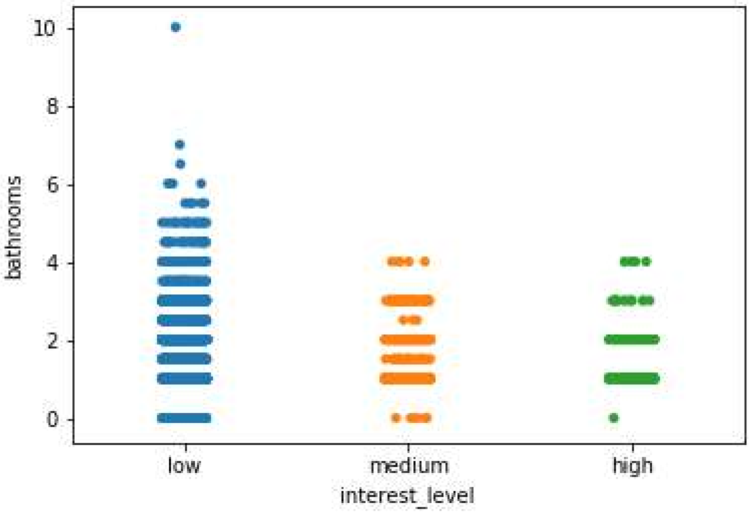
1) Max of bathrooms is 10 and the max of bedrooms is 8. The min of bathrooms and bedrooms are 0. Min of latitude is 0 and max is 44.88. Latitude is mostly around 40.7. Min of longitude is -118.27 and max is 0. Most of longitude is around -73.9. Min of prince is 43 and max is 4.49MM.

   It is very likely latitude, longitude and price has outliers.

   a) The 1st and 99th percentile of latitude is 40.6404 and 40.862047. So it is reasonable of thinking 0 and 44.8835 as outliers. We will floor and cap latitude by its 1st and 99th percentile.
   b) The 1st and 99th percentile of longitude is -74.0162 and -73.852651. So we will also floor and cap at these two numbers.



2) The 1st and 99th percentile of price is 1475 and 13000. To be reasonable, we can floor and cap at 500 to 20000 for price.

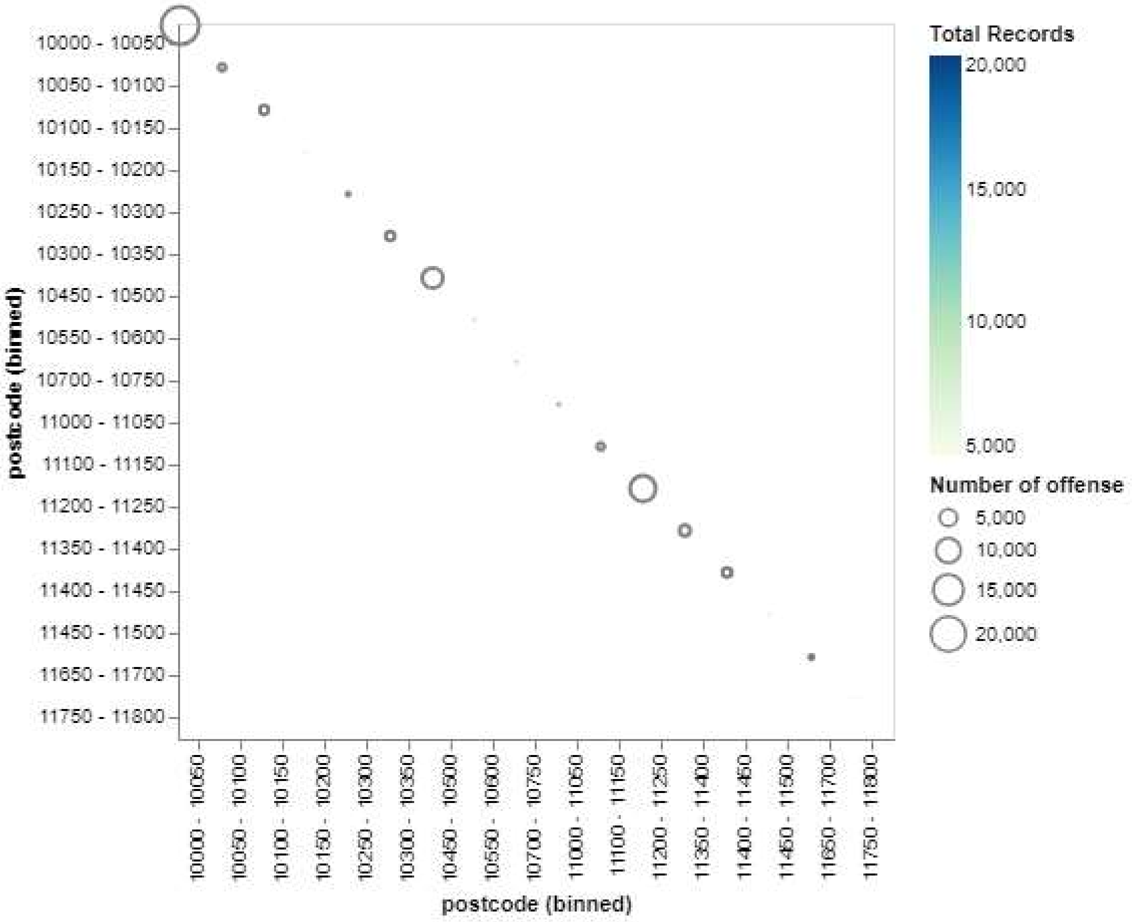


3) There is 1 house listing with 10 bathrooms. I think we can treat that as outlier.

Other datasets:
To get more insights along with rental dataset, we can use crime dataset, tourist attractions, and school rating datasets can be used.

To try this out I used NYC crime data along with rental data. With the use of reverse geocoding I identified zip code from longitude and latitude. I created a heat map of crime data and histogram of interest level.

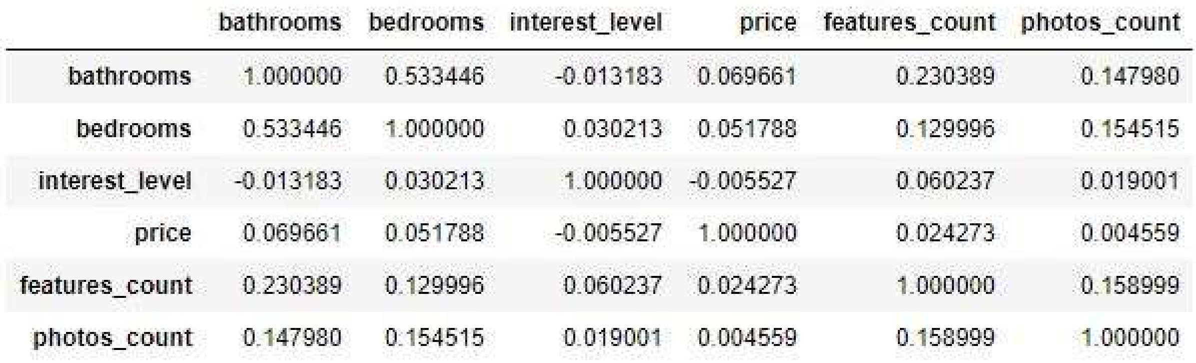
Links:
Reverse Geocoding using Google API jupyter notebook



Zip codes 10000 to 10050 are from Manhattan. Zip codes 11000 to 11200 are from Brooklyn.

Observation: Even though Manhattan has high crime rate, it's most popular location. Brooklyn has relatively low crime rate. It's 2nd best location for rental.

Inferential Statistics



	bathrooms	bedrooms	interest_level	price	features_count	photos_count
bathrooms	1.000000	0.533446	-0.013183	0.069661	0.230389	0.147980
bedrooms	0.533446	1.000000	0.030213	0.051788	0.129996	0.154515
interest_level	-0.013183	0.030213	1.000000	-0.005527	0.060237	0.019001
price	0.069661	0.051788	-0.005527	1.000000	0.024273	0.004559
features_count	0.230389	0.129996	0.060237	0.024273	1.000000	0.158999
photos_count	0.147980	0.154515	0.019001	0.004559	0.158999	1.000000

Here are my observations:

1. Number of bedrooms has positive correlation with interest level, but it’s hard to conclude anything based on the value number of features and number of photos has weak positive correlation with interest level
2. Interestingly, number of bathrooms and price has weak negative correlation with interest level number of bedrooms and number of bathrooms are directly proportional
3. Number of features, number of photos, number of bedrooms and number of bathrooms has positive correlation

Hypothesis testing:

- Null Hypothesis: Price has no effect on interest level. Mean rent is 3830
- Alternative Hypothesis: Price has an effect on interest level. In other words mean rent != 3830

I divided dataset into two sets:

- Set1 - Listing below mean rent 3830
- Set2 - Listing above or equal to 3830

After applying T-test using stats.ttest_ind, I got p-value as 1.1716559094536672e-07.

With such a low p-value we can safely deny null hypothesis. In other words, listing price has an effect on interest level.

Links:
inferential_statistics.ipynb