Kaggle - Two Sigma Connect: Rental Listing Inquiries

-Tarantino

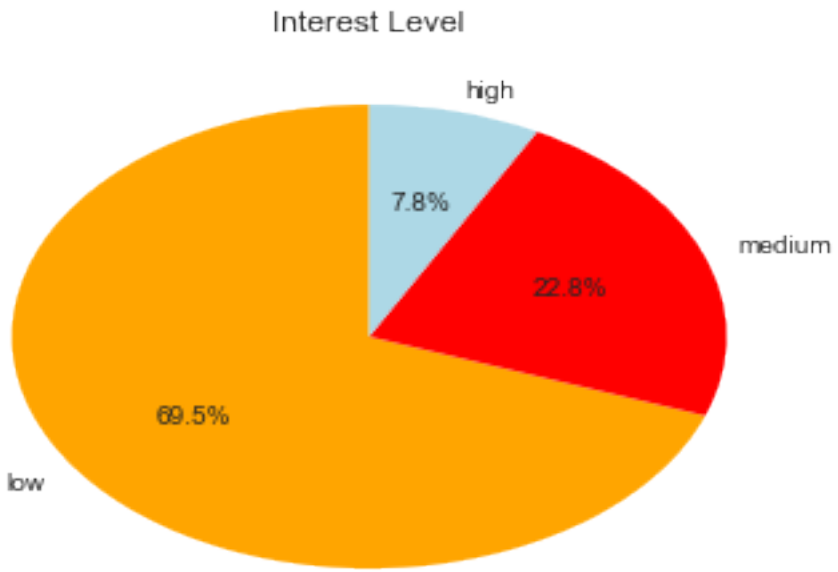
Overview:

- To determine on what criteria the interest level showed by the tenants depends
- The outcome is what would be a “good listing” that would get the maximum interest
- Performed exploratory data analysis
- Performed data cleaning, preprocessing and generated new features
- Use the different ML algorithms to generate accuracy and log loss
- Compared the log loss score and selected the best approach
- Tools and Libraries used: Python, Scikit Learn, Numpy, Pandas, NLTK, Matplotlib, Plotly, Gpxpy

Data Cleaning and Preprocessing:

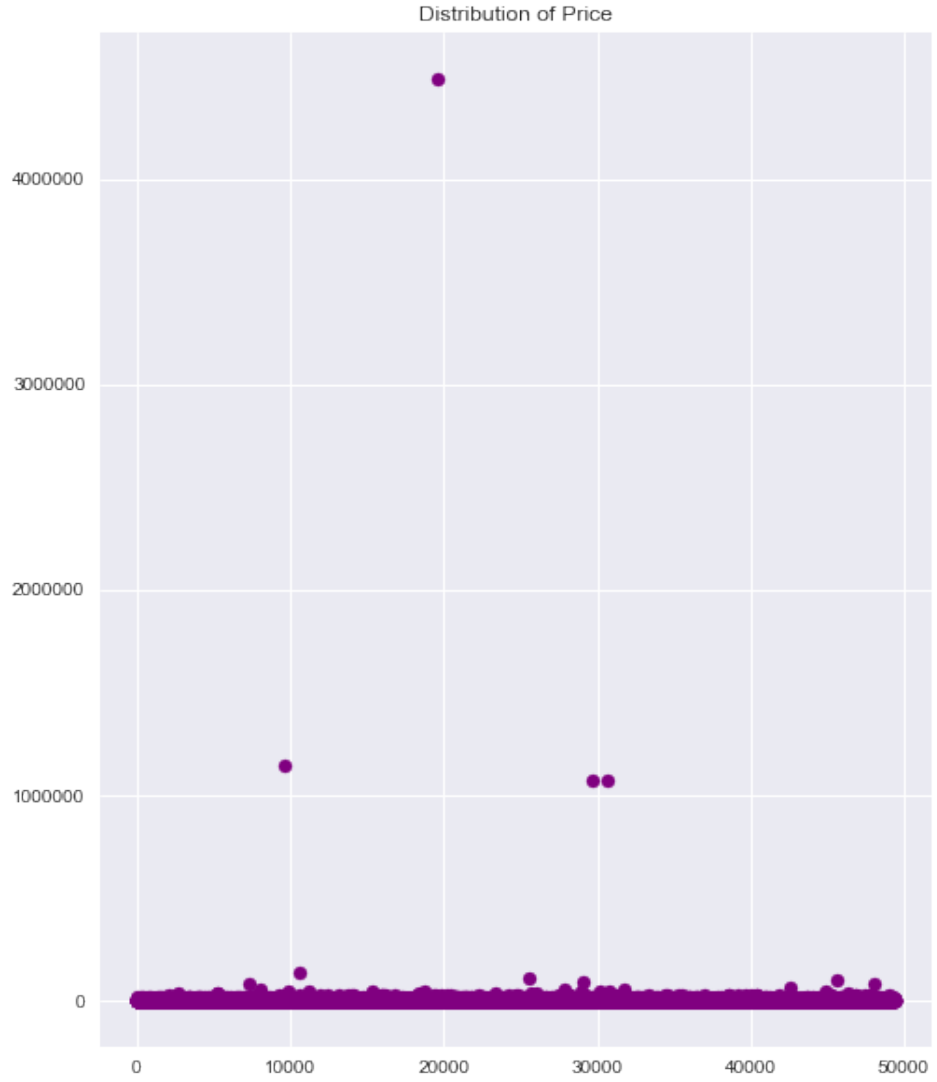
- Data cleaning
  - Removed usernames, URL, HTML tags, hashtags, punctuation, stop words and special characters
  - Stemming
  - Slang Conversion
  - Tokenization
  - Tf-Idf Vectorization
  - Label Encoding using Preprocessing
- Training and Testing
  - K – Fold Cross Validation (k = 10/100)
  - Stratified K – Fold Cross Validation (k = 10/100)

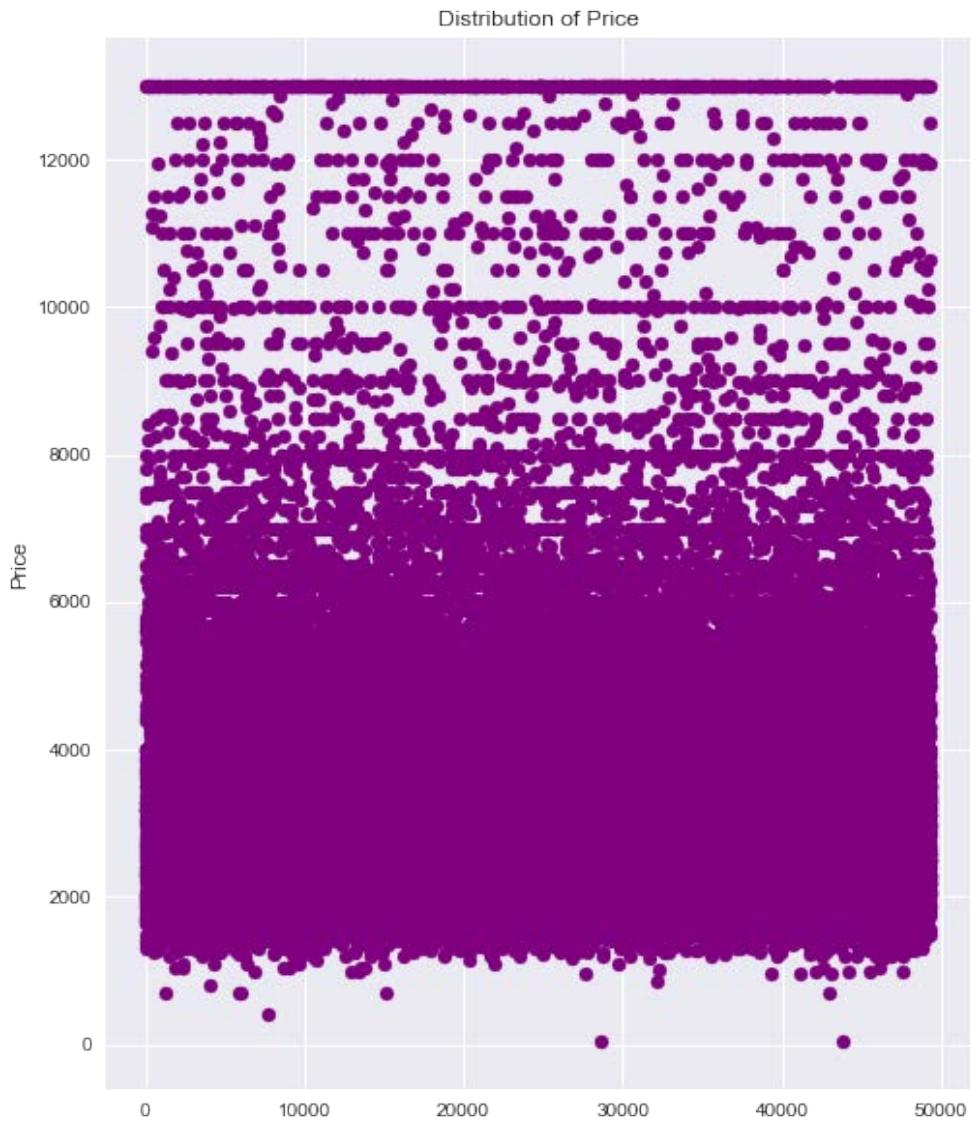
Disparity in the data:



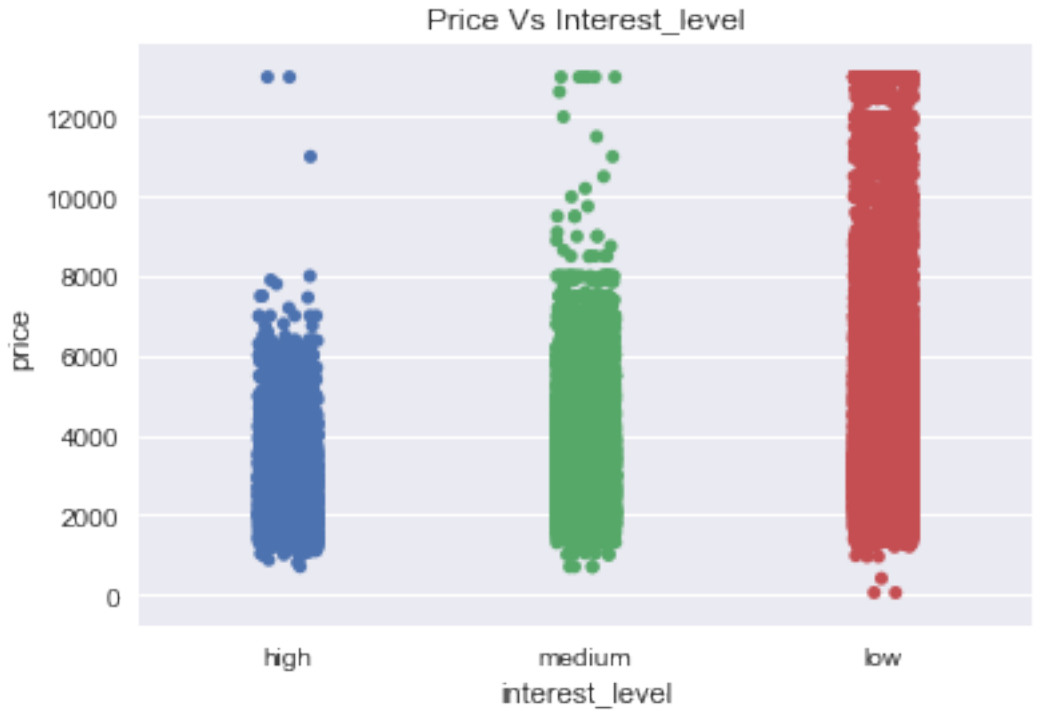
- The data given is highly skewed
- Use stratified k-fold approach for taking equal portions of data of each class
- Make the class names into labels for easier processing –
  1 - high , 2 - medium, 3 - low
- Having unbalanced data might cause performance variation and results tend to go towards the label with more training examples

Removing outliers based on price:

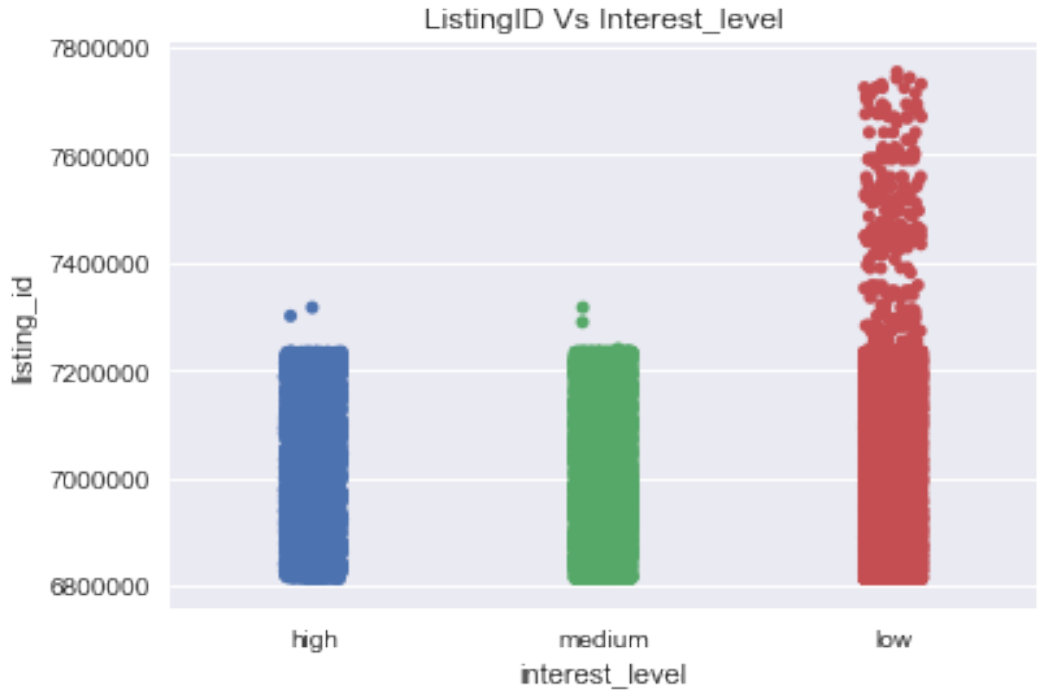




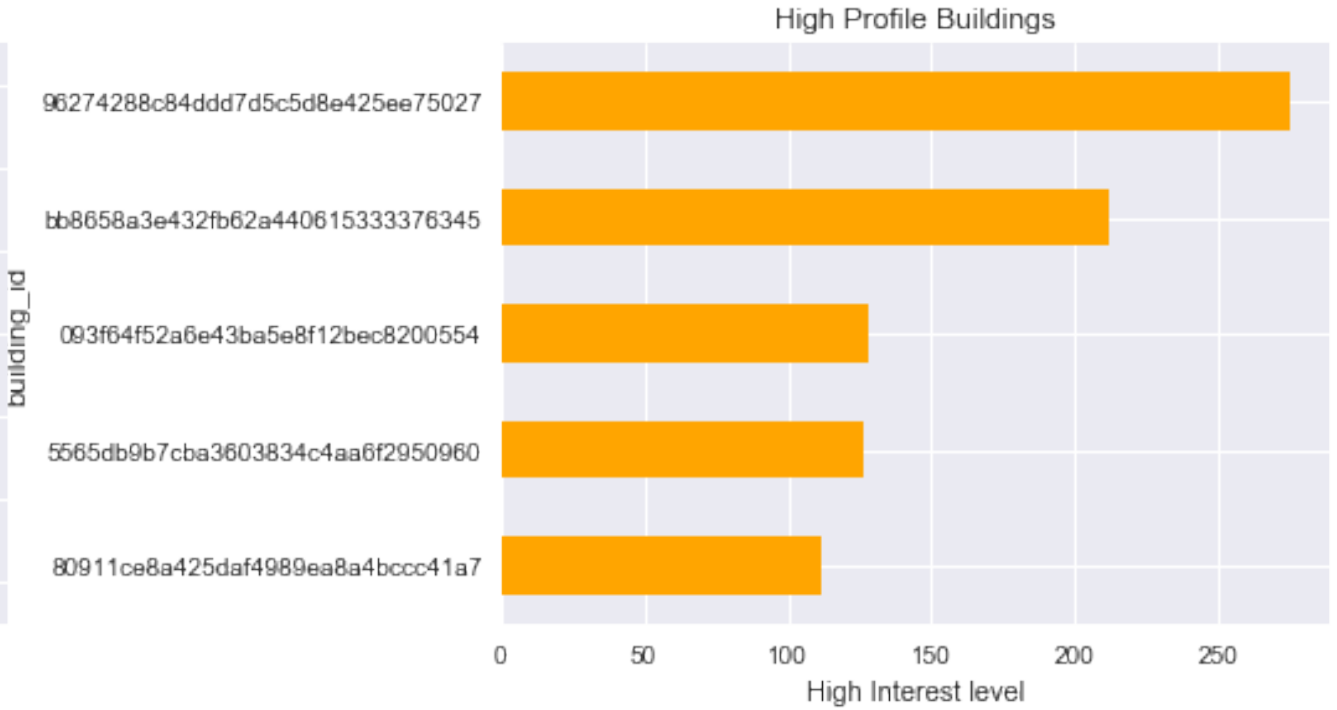
Price vs Interest:



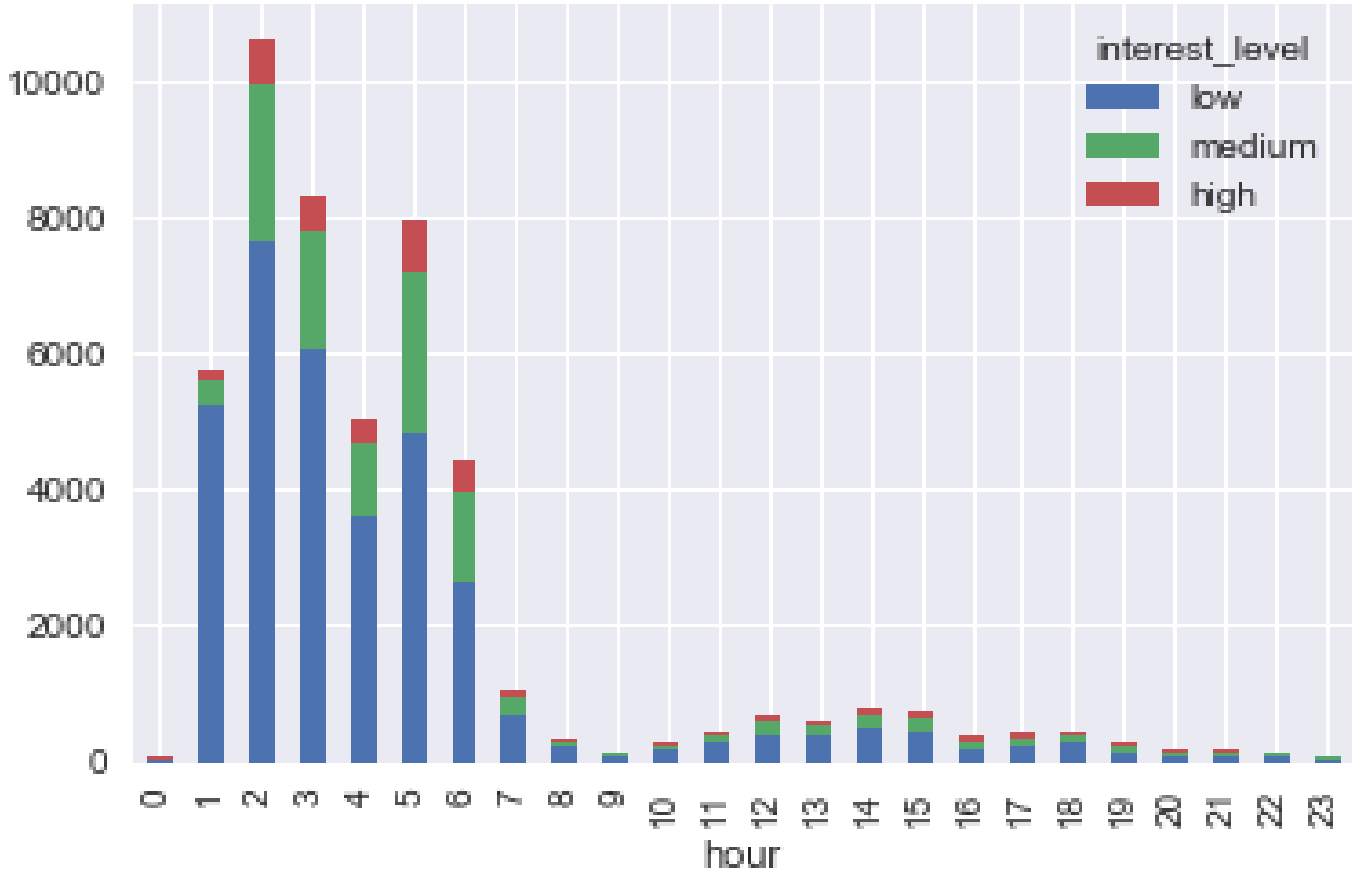
Listing ID vs Interest:



Repeated Managers and Buildings in the data:

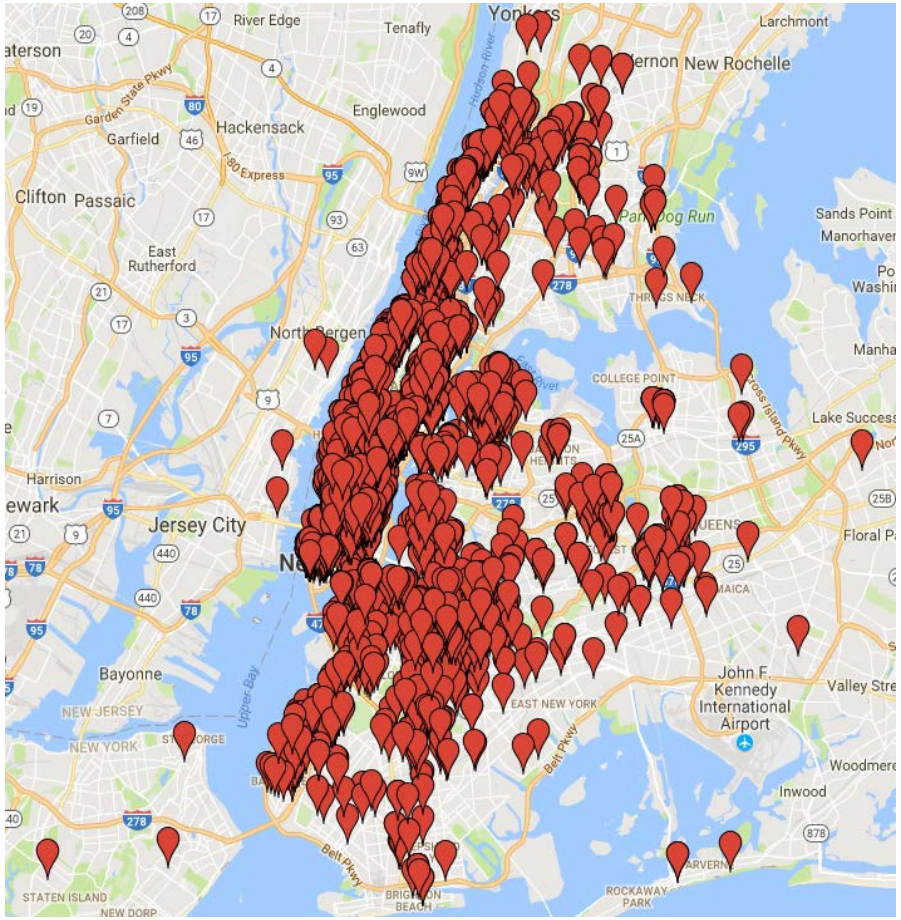


Time of the Day vs Interest Level:

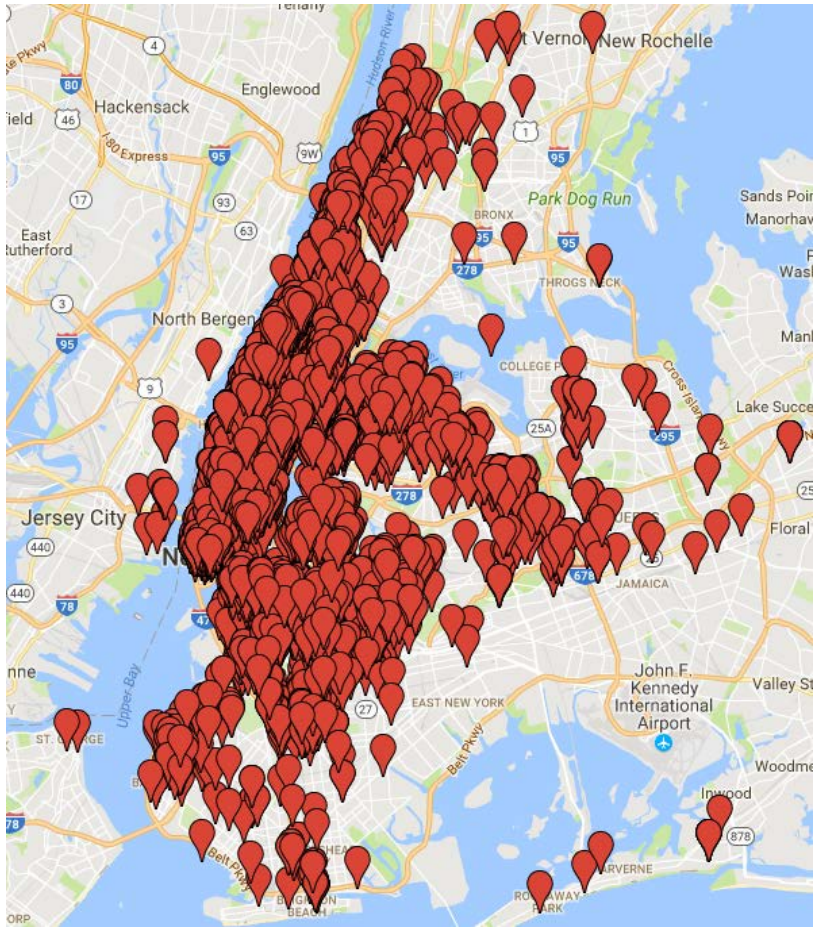


Interest Level based on Latitude and Longitude:

High:

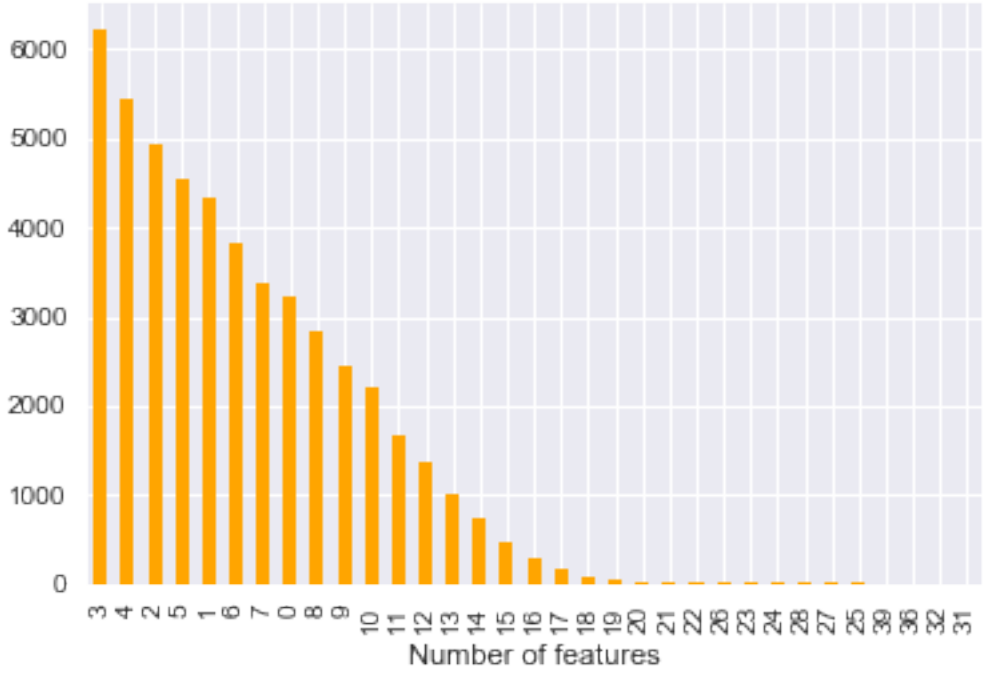


Medium:

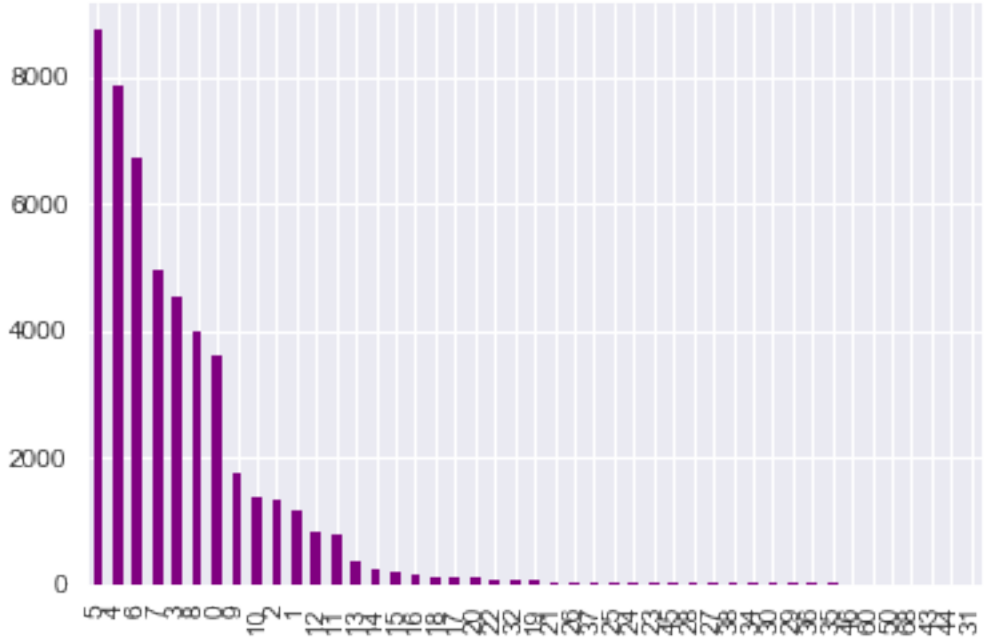


Interest based on features and photos:

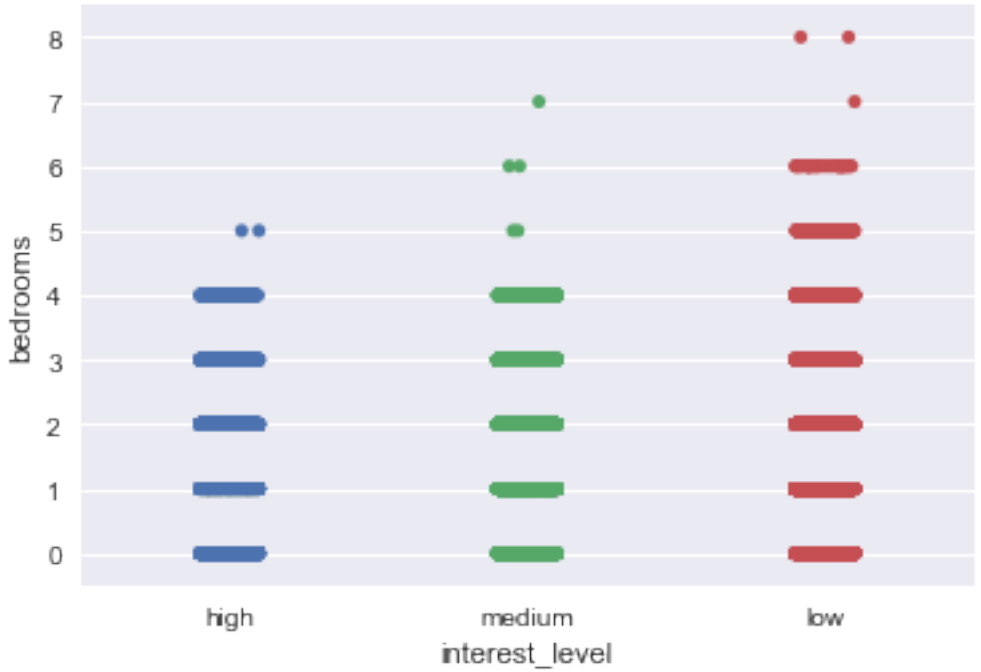
No. of Features:

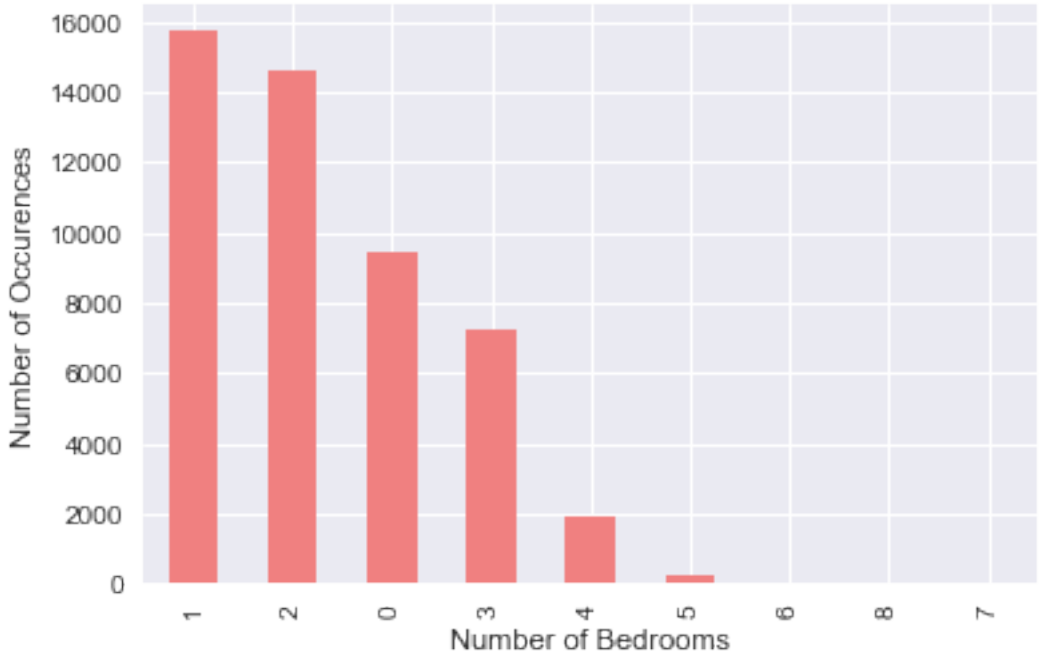


No. of Photos:



Analysis on the number of bedrooms:





Classifiers used and Results:

Classifier	Accuracy	Log Loss
Random Forests Classifier (n = 100)	78.90%	0.61
Extra Trees Classifier (n = 100)	75.96%	1.1
AdaBoost Classifier	76.59%	0.92
Logistic Regression	72.32%	0.69
Soft Voting Classifier (RF, ET, AdaB & LR)	79.02%	0.62
XGBoost Classifier	81.2%	0.57