**General Assembly Data Science Final Proejct**

**Alex Sherman**

**June 3, 2015**

**Problem statement and hypothesis: What factors correlate with a seller's listed price for online housing listings?**

The Washington D.C. tri-state area has a plethora of diverse housing and apartment rental options. To aid my own apartment search – which lacked sufficient criteria to compare listings - I conducted an analysis of the housing rental market for my final project in General Assembly’s Data Science course.

The focus on the analysis was housing listing data from Craigslist, used to determine the relative value of different rental criteria. Craigslist listings have text descriptions, images, preset metadata and tags, and geolocation data. This provided many avenues to learn about the housing market from the seller’s perspective.

Craigslist does not provide final sales prices; thus, Craigslist listings do not provide accurate demand-side metrics on the true value of a listing. Yet, Craigslist data does provide sufficient information on the perceived value from the seller’s standpoint. As I have never owned a property, I lack an understanding of the seller’s viewpoint. Accordingly, my project question is to determine what factor correlate with a seller’s listed price for online housing listings.

**Data collection methodology and description**

My intent for the final project was to practice the many topics covered in the General Assembly Data Science curriculum. Therefore. I choose to work with a messy user-created dataset to gain practical experience with data collection methodologies and tools. To obtain my dataset, I took a two-step process, building a web scraper to acquire online housing listings, and then collecting geographically relevant data from the Google Places API.

**Web Scraping:**

Web scraping is a systematic means to extract data from a website. Using the tag structure of HTML and CSS, a web scraper can automate the process to collect data of interest in an efficient and timely manner.

I choose to develop a data scraper for Craigslist housing listings. My initial analysis uncovered that the HTML and CSS tags on Craigslist were well formatted, lending themselves to work well with a data scraper. At the time of writing the robots.txt files at <http://www.craigslist.org/robots.txt> did not disallow the analysis of housing listing data.

Using the web scraper, I collected multiple types of provided data:

#### Housing Description:

-Seller's listed price

-Listing Title

-Listing Text

#### Housing Attributes:

-Availability: available now

-Bathroom: 1

-Bedroom: 1

-Cat: cats are OK - purrr

-Date\_available: 2015-04-15

-Dog: dogs are OK - wooof

-Housing\_type: apartment

-Laundry: w/d in unit

-Parking: attached garage

-Smoking: no smoking

-Square\_footage: 763ft2

#### Location Data

{'address': u'123 Random Street',

'city': u'Arlington',

'country': u'US',

'latitude': u'11.111111',

'location\_data\_accuracy': u'10',

'longitude': u'-22.222222',

'state': u'VA'}

#### Image data

-number of images

-size of each image (e.g. 600X450.jpg)

**Application Programming Interface**

Application Programming Interfaces **(**APIs) provide a programmatic access to stored data. In comparison, to web scraping, APIs decrease data access, but simplify the process to acquire data.

According to Google, “The Google Places API Web Service allows you to query for place information on a variety of categories, such as: establishments, prominent points of interest, geographic locations, and more. You can search for places either by proximity or a text string. A Place Search returns a list of places along with summary information about each place”

Using the latitude and longitude provided in each housing listing, I accessed the Google Places API to collect relevant information about establishments nearby each listing. My hypothesis was that, while not mentioned in most listings, nearby establishments play a role in the sellers listed price. If correct, I could use this information in my personal housing search to avoid expensive, but not personally relevant establishments; this data would also be relevant for sellers who could justify high prices by mentioning highly valued local establishments in their listings.

My analysis focused on nearby establishments within a one-mile radius, including grocery stores, gyms, movie theatres, train stations, subway stations, airports, Barnes and Nobles stores, Deloitte Consulting offices, and Starbucks coffee shops.

**Description of any pre-processing steps you took**

Online listings data require extensive data pre-processing. In my analysis, I created a separate Ipython Notebook for this endeavor.

Price

My first component of the data cleaning was to work with the listing price. Price was core to my analysis, so I removed all listings without a listed price as they were not relevant. Next, I detected and removed outliers. Most listing fell in the $2000 price range, yet, several were listed well above $10,000 with no justifiable rationale. I removed these listings as I assume these were cases of the user mistyping the price. On the contrary, many listings had a $1 price. Many of these listings had links to external sites; thus, my hypothesis, is that the $1 price was a strategy used by the seller to have their listings appear at the top when a user sorted by price. I also removed these listings to avoid skewing the data.

Cities and States

Craigslist listings often include detailed geolocation data. Most of the listings included both map from Google and Yahoo, which included latitude, longitude, country, state, and city data. The country and city data required the most cleaning. For both, I followed a four step data cleaning process:

* Turn all instances to lower case to reduce differences in capitalization.
  + e.g. Maryland and MARYLAND both became maryland
* Replace all encodings with blank string
  + e.g. washington%2edc became washingtondc
* Correct misspellings
  + e.g. mayland and marylnd became maryland
* Group logical instances
  + e.g. dc and washington became washingtondc

There variables were not used in the final model.

Square Footage

Most listings provided a square footage metric indicating the area of the housing. These were all provided as strings (e.g. 450ft2), so I replaced the ‘ft2’ for each to make them numeric. I removed several listings which did not follow the craigslist template and included square footage values such as loft. To fill in missing values I calculated and imputed the mean value for each grouping of housing type and number of bathrooms.

**Feature Engineering**

Feature engineering is the process of using existing fields in a dataset to create new variables. Feature engineering adds value to a data science analysis by extracting information from a data set and making it machine accessible.

Each listing included a date and time of posting. I collected this data with the hypothesis that more reputable rental agencies would post listings during work hours and private owners at off-peak hours. The data and time metrics were provided together. As a simple feature engineering task, I split these variables to evaluate date and time separately in my model. Additionally, I determined which weekday each listing was posted.

**Impute Missing Values**

Imputation is the process of filling in missing values with predicted values.

Due to the large amount of variables collected, most listings were missing a portion of data. Modeling libraries in Python, such as scikit-learn, do not accept blank values. Thus, imputation was required for most variables.

I took a simple approach to impute missing values.

1. For categorical values, I created a field specific ‘No\_Field\_Data’ variable (e.g. No\_City\_Data)
2. For numeric values, I filled all missing values as zero except for those that I could estimate such as square footage

**Data Transformation**

Python modeling libraries often require numeric data. Variables without a natural ordering must be converted to dummy variables with True (1) or False (0) values. For instance, craigslist provided data whether cats were allowed as text. I converted these values to (1) = yes, cat are allowed or (0) no, cats are not allowed. I conducted this process for a variables with binary answers and used pandas.get\_dummies() function to convert all other categorical variables with more than two options, such as city.

**Outlier Detection**

Outliers are values that values that significantly differ from other values in a group.

The data included many outliers. Some listings had zero bedrooms, bathrooms, or pictures. Other listing were for less than 100 square feet. I choose to remove unrealistic listings similar to these.

**What you learned from exploring the data, including visualizations**

In practice, my data cleaning process was conducted in parallel with the project data visualization. Many of the data cleaning choices were initiated from data visualization discoveries. Nonetheless, I choose to split these to topics into separate Ipython Notebooks for ease of data reproducibility.

To learn about the fields in my data set, I created histograms for the distributions of each variable. One prevalent trend was that the majority of the data from the Google Places API was skewed to the right. Due to these distributions, I became less certain that this data would provide value to the analysis.

**How you chose which features to use in your analysis**

My criteria to determine if a variable was useful for the analysis was if could provide significant differences in price. If the inclusion or exclusion of the variable provided no change in predicted price, then it would not help the analysis. To visually examine this criteria, I created a function to group each variable by price and then plot a line chart with price on the y-axis and the variable of interest on the x-axis.

My final model included the following features:

attached garage,

bathroom,

bedroom,

grocery\_list,

gym\_list,

image\_number,

latitude,

longitude,

off-street parking,

Posting\_Day,

Posting\_Time,

Posting\_Time\_AM\_or\_PM,

square\_footage,

square\_footage\_cleaned,

street parking,

weekday\_of\_posting

**Details of your modeling process, including how you selected your models and validated them**

My most successful model was a random forest using n\_estimators=200 and max\_features=5. In practice, most regression models, multivariate, ridge, and lasso were ineffective at predicting price given the available and engineered features. I expect that the prediction will improve given an expanded data set.

**Conclusions and key learnings**

My most significant finding was the distinction between provided data and data that influences the price. My analysis concluded that there are significant location specific data that influence the listing price, but are rarely mentioned in listings. In particular, information about nearby gyms and grocery stores both were in my top eight features throughout my analysis.