Las Vegas House Price Analysis

August 7, 2019

1. Using the zillow API, provide a data set that will include sale transactions information for single family homes located in the Las Vegas MSA that achieved a sale price between $200,000 and $400,000 in the last 18 months. The information will include all the data fields of the Zillow API, including but not limited to location, age, internal and plot size, layout, amenities etc. See: https://www.zillow.com/howto/api/GetDeepSearchResults.htm We'd like also to find time on market but we're not sure if the Zillow API provides it.

Conduct data merge and clean up so that the data is made reliable for modelling purposes.

Create a data frame in Python (Pandas or equivalent). Show some initial statistics to visualize and assess the data e.g. quartiles, minimum, max and mean values for the key variables such as floor area, price and features. Snapshot data using plot.ly box plot. Show impacting features.

Interpret statistically the data, including at a minimum: Deviation from normal distribution;

Data Description

2.

Skewness and kurtosis; Seaborn (or equivalent) pairplots to visually show relationships between variables. Seaborn heat maps to plot correlation values between property variables and price.

Drop variables with negligible correlations. Zoom into variables with price effects, both log scale and non scaled.

Last sold price -- before log scale

Last sold price -- log scaled

Pairlot Chart

Seaborn Heatmap

Neiborhood Analysis:

In [24]:

Neiborhood Name	Number of Properties
Angel Park Lindell	304
Buffalo	158
Centennial Hills	2260
Charleston Heights	645
Cultural Corridor	38
Desert Shores	357
Downtown	22
Downtown East	11
East Las Vegas	150
Enterprise	4120

```
Huntridge              290
Las Vegas              790
Lone Mountain         1568
Meadows Village          1
Michael Way            719
North Cheyenne        1183
North Las Vegas          7
Paradise              2889
Pioneer Park           310
Providence             998
Queensridge             18
Rancho Charleston      359
Sheep Mountain         559
Spring Valley         3247
Summerlin North       1201
Summerlin South        503
Sun City Summerlin     544
Sunrise                102
Sunrise Manor         1701
The Lakes              722
The Strip                8
Tule Springs           665
Twin Lakes             195
UMC                     40
West Las Vegas         127
Whitney                638
Winchester             217
Name: addr_latitude, dtype: int64


  File "<ipython-input-24-ca2f20cbd256>", line 1
Neiborhood Name      Number of Properties
               ^
SyntaxError: invalid syntax
```

3.

Analysis:

Find to what extent (in %) each data field contributes to the formation of the sale price of units grouped by zip codes and number of bedrooms.

Separate data into training set and test set. Use R2 score (or equivalent) to check prediction accuracy.

Try a minimum of three different machine learning algorithms/systems, such as regression trees (random and multiple regression), neural networks, support vector machines and select the best performing one based on lowest mean absolute error or other relevant measure.

3.1 XGBoost Result with Zestimation in feature

Finale Validation MAE for multi-pass XGBoost Model : 13209.818566128175
3.2 XGBoost Result w/o Zestimation in feature
Finale Validation MAE for multi-pass XGBoost Model w/o zest : 17017.505402260635
3.3 Zest MAE as comparision : 19691.43420015762
3.4 Random Forest Model Result (w/o zest)
Finale Validation MAE for Random Forest Model w/o zest : 17266.360878336007

In [ ]:
```
Variable Importance Analysis from Random Forest.

     Variable: hf_sqft              Importance: 0.53
     Variable: addr_longitude       Importance: 0.1
     Variable: addr_latitude        Importance: 0.05
     Variable: hf_lot_size          Importance: 0.05
     Variable: hf_year_built        Importance: 0.04
     Variable: images_count         Importance: 0.04
     Variable: sch_high_Palo Verde  Importance: 0.02
     Variable: hoa_month            Importance: 0.01
     Variable: addr_zipcode         Importance: 0.01
     Variable: hf_bathrooms         Importance: 0.01
     Variable: hf_num_rooms         Importance: 0.01
     Variable: hf_bedrooms          Importance: 0.0
```

3.5 Support Vector Regression Model Result (w/o zest)
Finale Validation MAE for Support Vector Regression w/o zest : 24593.699881529938

In [ ]: