Capstone 3 - Final Report - Michael Ruston

Data: Real Estate Data was taken from Kaggle.com. The exact data source can be found here: https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

The data was scraped from realtor.com using python libraries.
The dataset comes as a csv with the following columns:
-status
-price
-bed
-bath
-acre_lot
-full_address
-street
-city
-state
-zip_code
-house_size
-sold_date

**Exploratory Data Analysis**

The 'status' column had two values: 'for_sale' and 'ready_to_build'. Over 99% of rows were 'for_sale' status, so the 'ready_to_build' rows were dropped.

An initial look at the 'state' column revealed the following counts per state:

| | |
|---|---|
| Massachusetts | 175,248 |
| Connecticut | 89,776 |
| New Hampshire | 51,394 |
| Vermont | 46,460 |
| Maine | 36,650 |
| New York | 29,990 |
| Rhode Island | 29,596 |
| New Jersey | 25,662 |
| Puerto Rico | 24,679 |
| Virgin Islands | 2,573 |
| Georgia | 48 |
| South Carolina | 25 |
| Virginia | 20 |
| Tennessee | 20 |
| Pennsylvania | 14 |
| Wyoming | 3 |
| West Virginia | 1 |

We note that this data set does not cover all 50 states, but rather a handful of states in the New England region, Puerto Rico and the Virgin Islands, and a handful of other states.

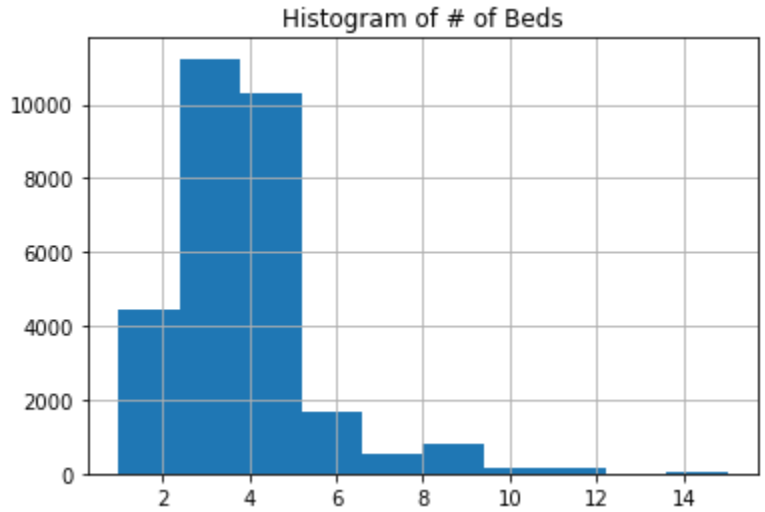
Because Puerto Rico and the Virgin Islands seem to be in a different region, and some states have only a few properties in the data set, the data was filtered here to include just the few top states:
**Massachusetts, Connecticut, New Hampshire, Vermont, Maine, New York, Rhode Island, and New Jersey**

At this stage, there were 2460 unique zip codes in the data.

After looking at the 'city' column, we see that 'New York' and 'New York City' both appear in the data with high frequency. 'New York City' is used instead of 'New York' in a minor data cleaning effort.

A look at the 'bed' column showed that some property listings have well in excess of 10 bedrooms. Some of these appeared to be whole apartment buildings for sale, which is a different type of property. To remove outliers, only rows with number of beds <= 15 were kept. Rows with 16 or more beds were removed at this stage.

A histogram of the 'beds' column shows that most listings have 6 or fewer beds.




We looked deeper at the 'house_size' variable and a few rows with erroneous data were removed from the data set.

Next, all duplicate rows were removed. A high percentage of the data was duplicate rows, which indicates the importance of this step.

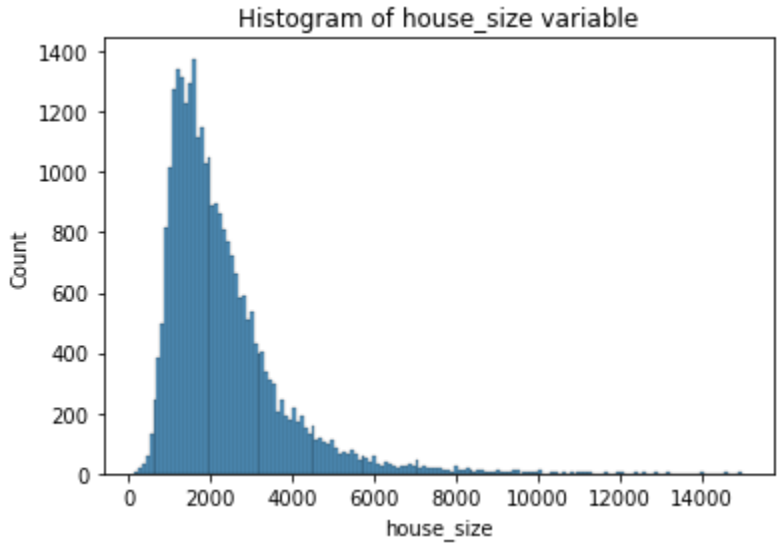
One listing had a missing value for zip code and was dropped. Next, zip code was formatted correctly, as a 5-digit string, not an integer.

The ‘acre_lot’ column was examined and some erroneous values such as 99999 were removed from the data.

A new column was created: ‘prop_type’ to indicate if the property is a house or an apartment or condo. This was set to ‘house’ by default and changed to ‘apt’ if the full_address column contained ‘Apt’ or ‘Unit’, while ignoring case.

In another attempt at removing outliers from the data, rows with house_size of 15,000 or more were removed from the data.

The majority of the data had a house_size well under 10,000. Just a small percentage of extreme values were removed at this stage by filtering on house_size.




A new column was created: ‘log_price’ that’s simply the natural log of the price column. All prices were positive, so this didn’t create any null values.

38 properties (rows) had an acre_lot value of exactly 0. This seemed strange, and a judgement call was made to replace these zeros with 0.01, the lowest value for acre_lot besides 0.

After these steps, EDA (Exploratory Data Analysis) was considered finished and I moved on to data pre-processing.

**Data Pre-Processing**

The status column contains just ‘for_sale’ and so was dropped as it contained no variation.

123 rows had NaN values for ‘bath’. These rows were dropped.

A new column: ‘sold_flag’ was created. It’s 1 if the sold_date is not missing and 0 if the sold_date value is missing.

A ‘log_acre’ value was created as the natural log of the ‘acre_lot’ column. All acre_lot values were positive, so no errors were created at this step.

I next looked at duplicate values in the ‘full_address’ column. Some rows had the same address in the full address column, but differing values for bed and bath, for instance. This looked to be erroneous and so these duplicates were dropped from the data set.

Dummies were created for ‘prop_type’ (house or apt) and for each of the 8 states in the data set.

The data at this stage had 29,075 rows and 26 columns.
The data was again saved to a .csv to be used in the next step.

**Problem/Goal:** This project has two main goals: one is simply to produce a useful dashboard using Tableau Public. The other is to predict the price column based on the other features of the data set. I will try two approaches: linear regression and random forests to predict price.

## Modeling

The data was split into 80% training / 20% testing chunks.

A linear regression was fit on the training data. This model had an R-squared around 0.31 and an in-sample RMSE of around 1,363,000. The RMSE on the test data was just over 1,030,000.

A random forest regressor model was also fit on the training data. This model had an in-sample RMSE of just about 405,500, which appears to be much better than the linear regression model. The RMSE of the random forest model on the test data was just under 956,000.

This indicates that the random forest model is overfitting the in-sample data. The linear regression model actually gives a lower RMSE on the test data, indicating no problem with overfitting.

**Dashboard**

A dashboard was created using Tableau Public.

It can be viewed at the following location:

https://public.tableau.com/views/MichaelRustonRealEstateDashboard/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link