Valentina Rizzati

Axelrod launches in NYC

May 11, 2021

Introduction

In 2022, Axelrod Development Group (Axelrod) is planning to allocate a significant portion of its new investment fund to launching the NYC market. For this purpose, the Chief Investment Officer asked me to analyze the NY Real Estate market on the sellers side and:

1. **Interpretative goal:** understand what are the factors driving the largest impact on property prices so that the Investment team will know what features (e.g. zip code, number of rooms) to prioritize when developing new projects
2. **Predictive goal:** build a model that will allow the Investment Team to get an estimate of property price by simply inputting the features (e.g. 11249 for zipcode, and 5 for number of rooms)

Process

To achieve the goals, I adopted the following strategy:

1. **Scrape Zillow** with BeautifulSoup and Selenium in order to gather data on a random sample of at least 1000 listings in the five boroughs of NYC across all prices.

2. **Map zip code to areas** with a triangulation of NYC Gov, ZipDataMaps, and Google Maps data sources. Because of the relatively contained sample of observations I'll be working with, areas will be used instead of zip codes in the linear regression modeling.

3. **Clean the data** by removing duplicates, dealing with missing values and zeros

4. **Conduct exploratory data analysis** by visualizing data through pairplot, correlation matrix and other relevant visualizations. See some examples below.

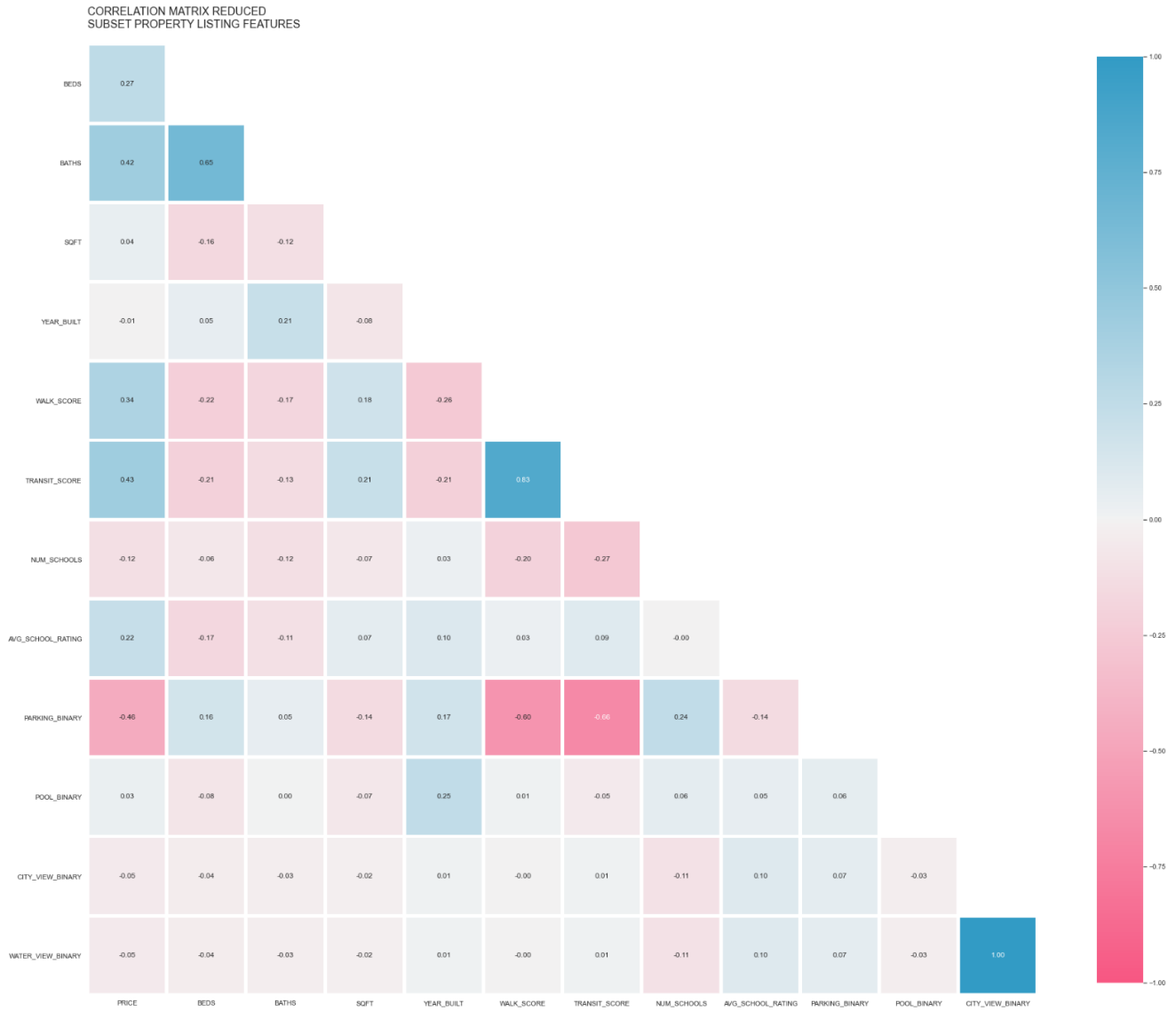
5. **Baseline regression model** by running the linear regression modeling on a smaller dataframe not including all the columns related to areas and property type.

6. **Establish a validation and testing scheme** by setting up a data splitting structure for validating and testing the model with cross-validation.

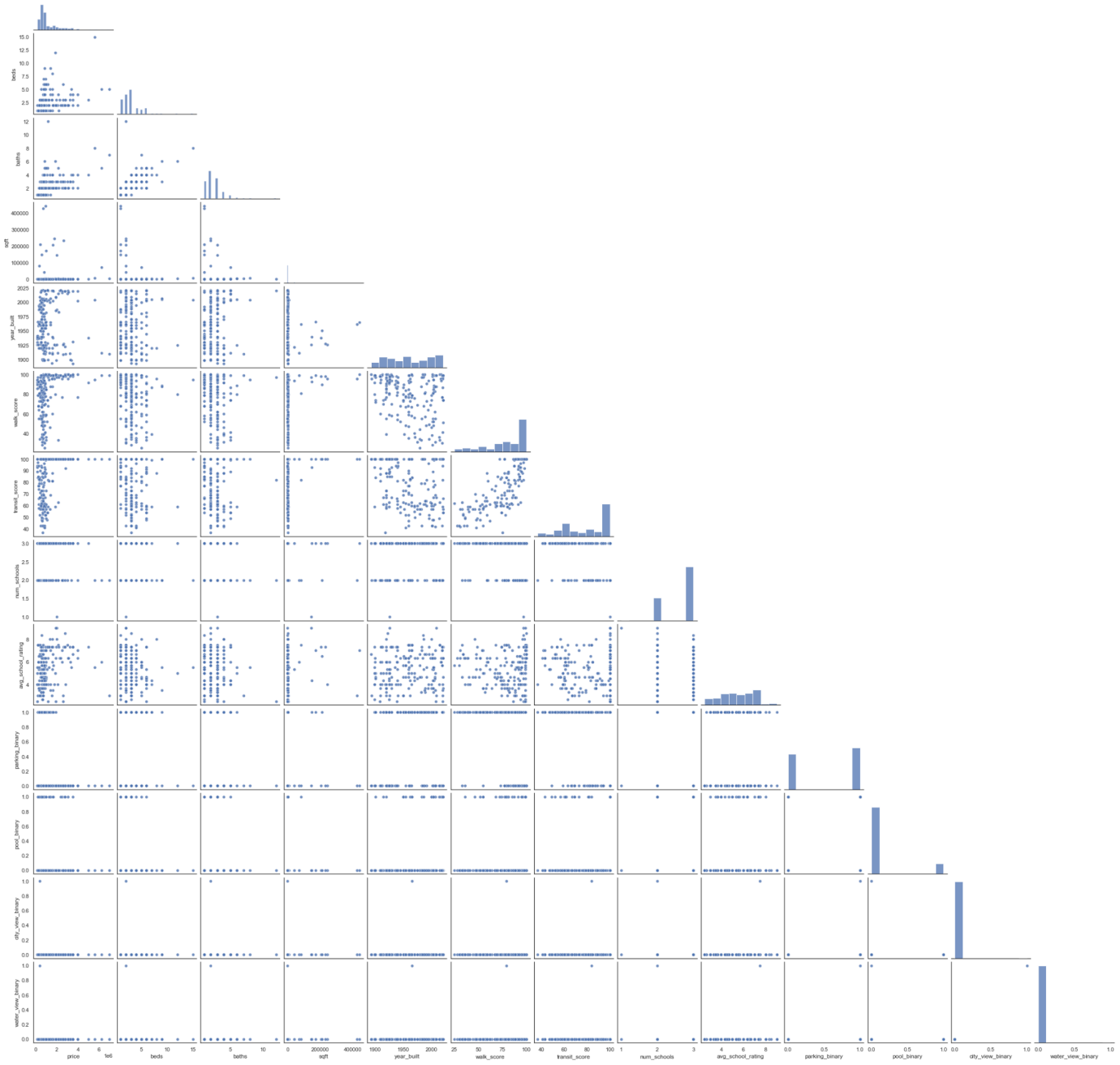
7. **Expand and refine the model** by first evaluating if the complexity of the model needs adjustment. Also, additional feature engineering might be needed. In fact, my hypothesis is that location is a necessary predictor of price (target), especially for a city like New York. If the model doesn't yield satisfactory results when all areas are included I might represent location by a higher level of aggregation (i.e. borough). However, if I was to implement this last solution, I would probably improve the model's predictability at the expense of its interpretability (i.e. NYC boroughs are so different that any recommendation at a level of borough would be significantly less actionable than a neighbourhood-based recommendation). Finally, I might consider regularization if needed.

8. **Finalize, test and interpret the model** by establishing the model choices and interpreting the results based on the context of Residential Real Estate in NYC and the model's use case for Axelrod.

The following visualizations and baseline results are based on a sample of **192 observations**. The remaining observations will be integrated once the scraping is completed.



4


price
beds
baths
sqft
year_built
walk_score
transit_score
num_schools
avg_school_rating
parking_binary
pool_binary
city_view_binary
water_view_binary
1e6

Baseline the model

[126]:
```
# Fit a simple model first and do not include property type and area dummies
features_simple = house_lean_df[['beds', 'baths', 'sqft', 'year_built',
                                 'walk_score', 'transit_score', 'num_schools', 'avg_school_rating',
                                 'parking_binary', 'pool_binary', 'city_view_binary',
                                 'water_view_binary']]
target = house_lean_df['price']
```

[127]:
```
# Fit a linear regression model on this preliminary data set
house_lr = LinearRegression()
house_lr.fit(features_simple, target)
```

[127]:
```
LinearRegression()
```

[128]:
```
# Check the R-squared value of the model on this preliminary data set
house_lr.score(features_simple, target)
```

[128]:
```
0.5315856144731066
```

[129]:
```
house_lr.coef_
```

[129]:
```
array([ 8.41924350e+04,  3.34233976e+05, -1.88852738e-02, -7.12413410e+02,
        2.73604440e+03,  1.71625975e+04,  2.07489510e+05,  1.64252974e+05,
       -5.64128726e+05,  2.06662511e+05, -2.48030370e+05, -2.48030370e+05])
```

Next Steps

1. Expand sample size

   Integrate additional ~1000 observations from additional scraping.

2. Baseline model with larger sample size

   This should boost the model's predicting and interpretative power.

3. Establish cross-validation strategy, refine the model and finalize recommendations

   Ensure that recommendations are actionable from a business perspective.