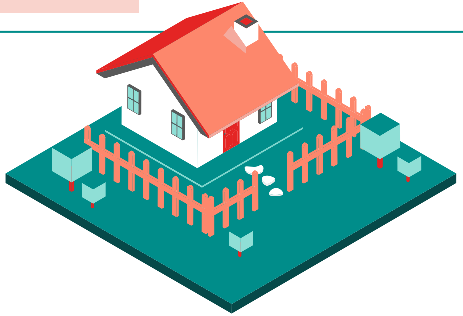
Airbnb: Price Prediction

Vivian Xia

TABLE OF CONTENTS

Business Objective

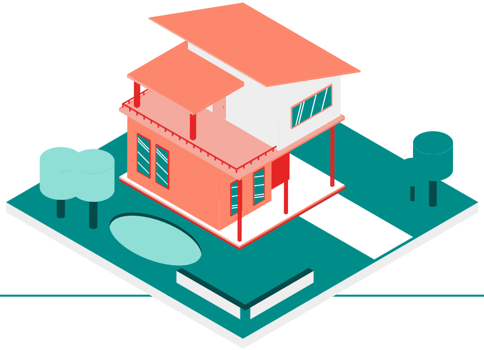


Problem

AirBnb wants to provide Seattle hosts with a pricing strategy to set competitive rates and maximize profits.

Solution

A model to predict Seattle Airbnb prices with transparent features.



Data

- Airbnb: Seattle listings
- ZipAtlas: Median household income by Seattle zip codes

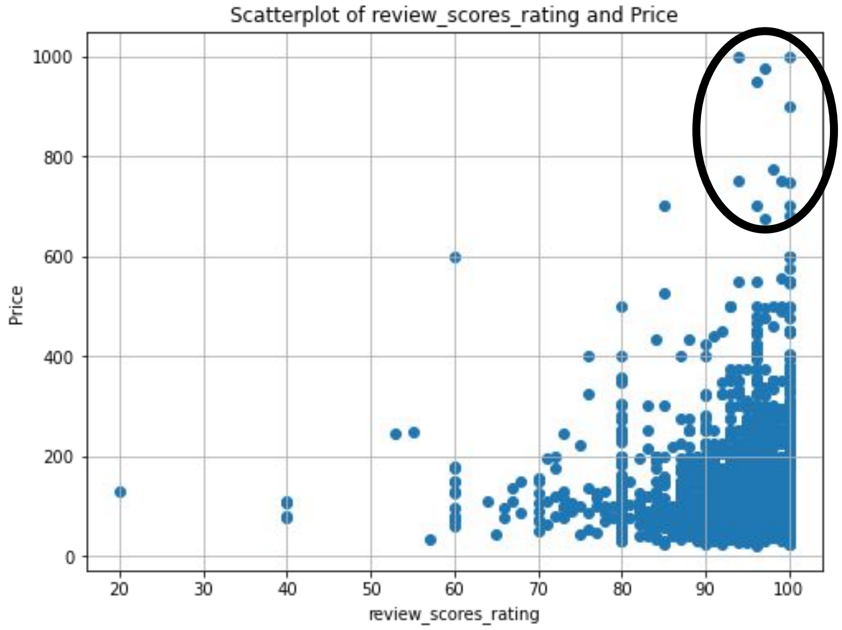
Data Cleaning

Non-unique variables (city, state) &
only unique variables (id, lisiting_url)

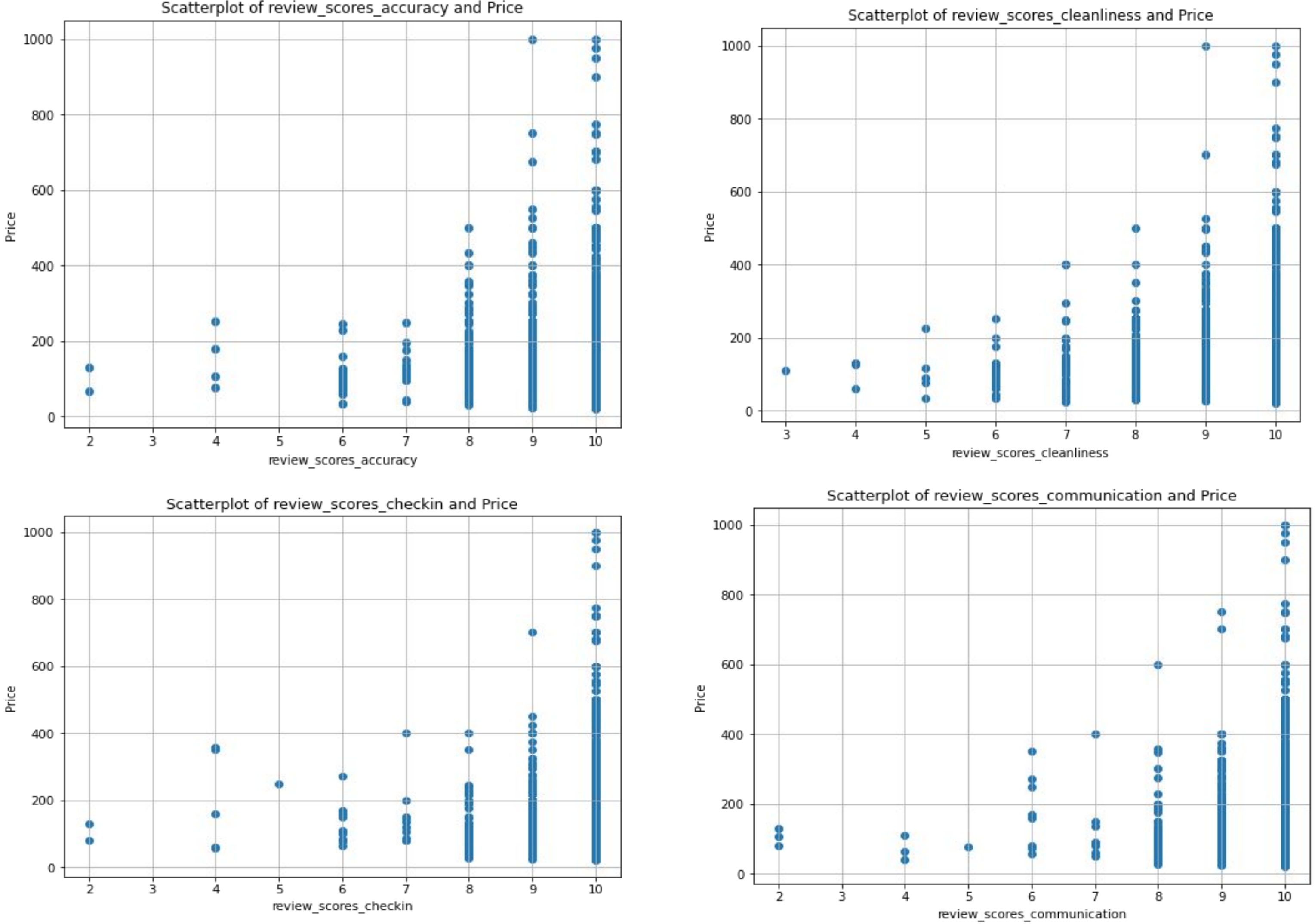
Missing values

Outliers

Review Scores vs. Price

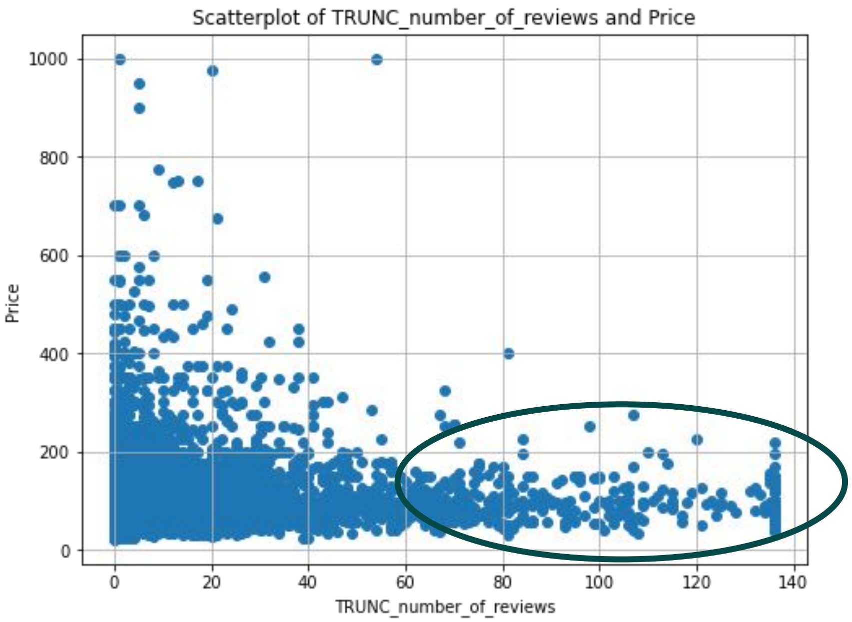


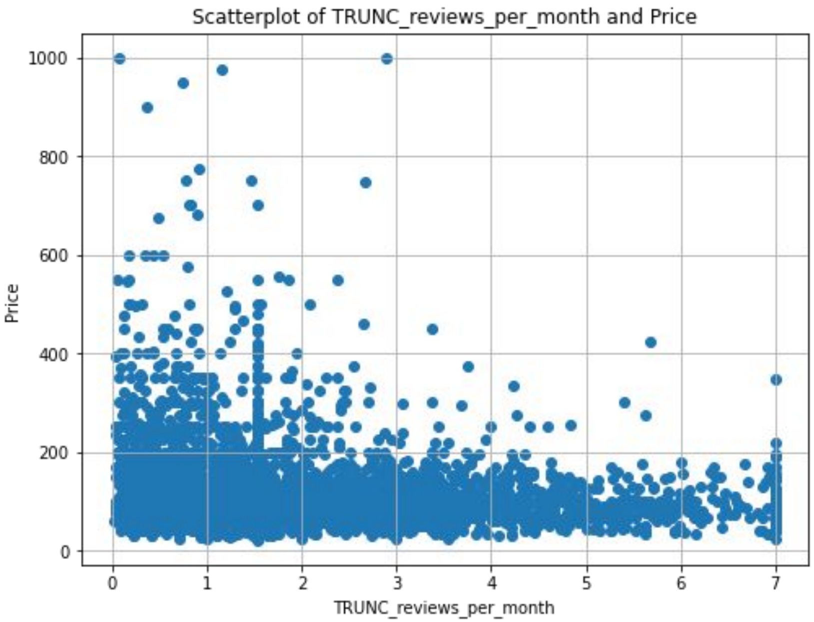
- Listings with high review scores have higher prices
- Lower rated listings have lower range of prices



Scatterplot of review_scores_accuracy and Price
Price
review_scores_accuracy
Scatterplot of review_scores_cleanliness and Price
Price
review_scores_cleanliness
Scatterplot of review_scores_checkin and Price
Price
review_scores_checkin
Scatterplot of review_scores_communication and Price
Price
review_scores_communication

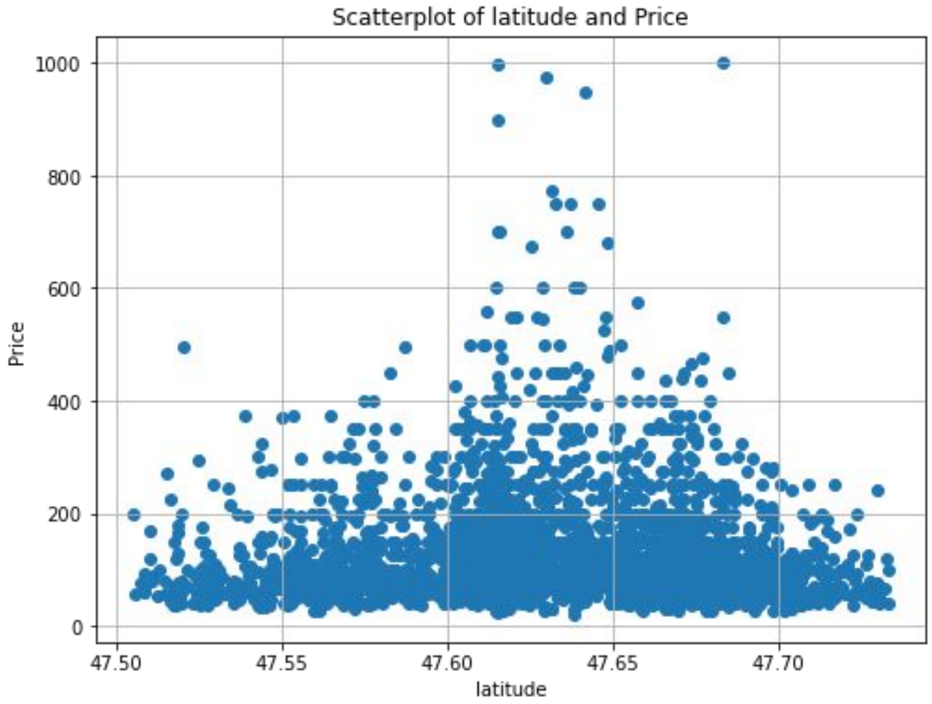
Number of Reviews vs. Price





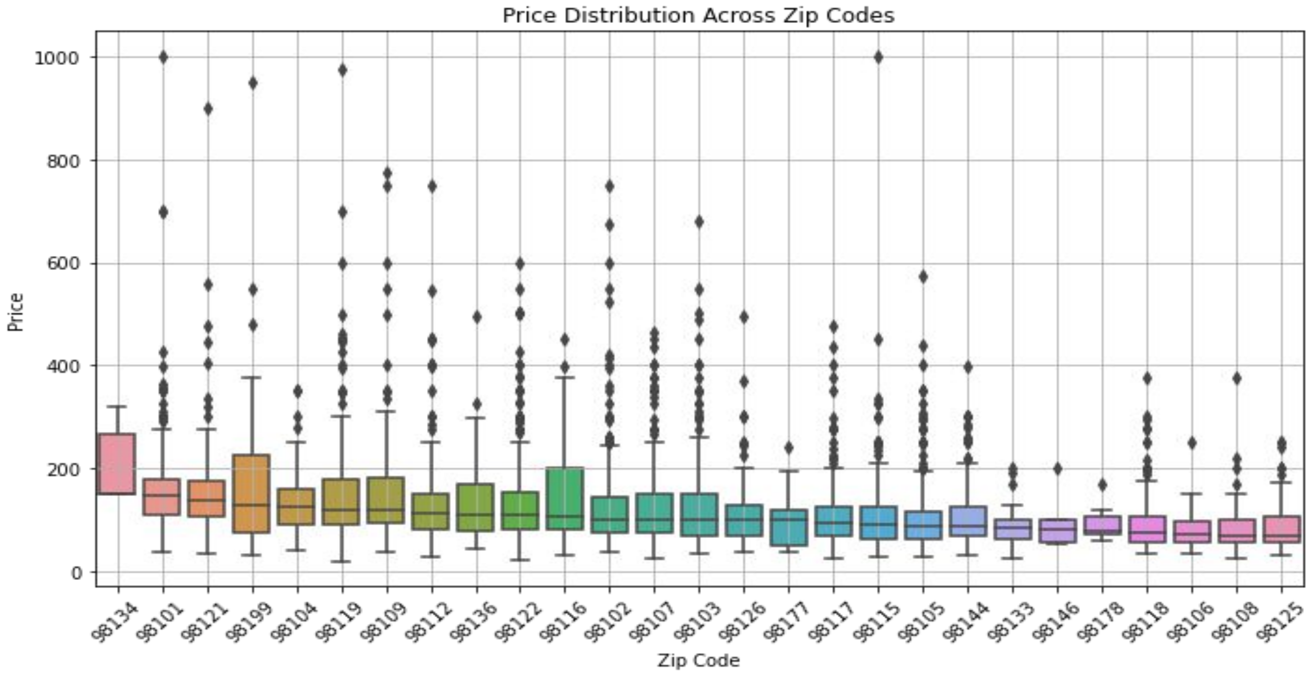
- Listings with higher number of reviews tend to have lower prices than the listings with lower number of reviews

Latitude vs. Price



- Higher priced listings are within the latitude of 47.60-47.65

Price Distribution Across Zip Codes



- 98134 and 98101 zip codes have the highest median price at $150 and $149, respectively

Features: Calendar Updated

```
listings_df['calendar_updated'].head(5)
```

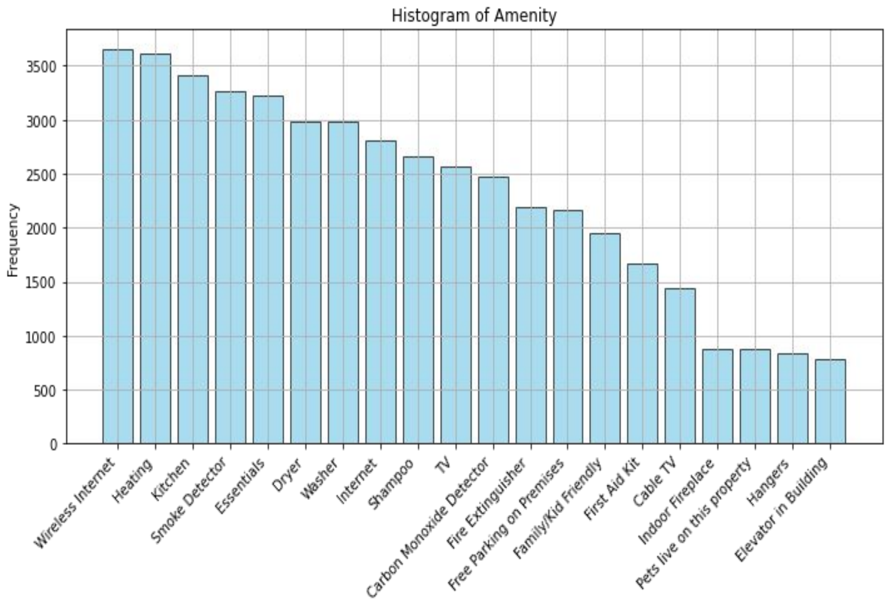
```
0      4 weeks ago
1            today
2      5 weeks ago
3     6 months ago
4      7 weeks ago
Name: calendar_updated, dtype: object
```

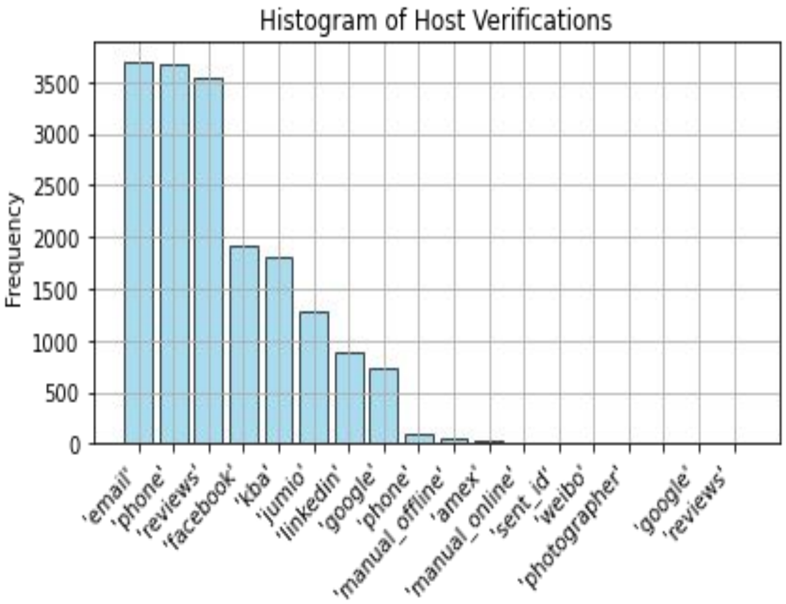
```
listings_df['days_ago'].head(5)
```

```
0      28.0
1       0.0
2      35.0
3     180.0
4      49.0
Name: days_ago, dtype: float64
```

- calendar_updated converted to days_ago

Features: Amenities and Host Verifications





- Took the top 15 amenities and top 9 host verifications and converted them into features

Further filtering of variables

Multicollinearity: Two or more features are highly related to each other, providing redundant information.

Statistical Significance: The variable does have an impact on the price.

Types of Models

Tree-based Models

Simple Decision Tree	Random Forest	Gradient Boosting

Regression Model

Linear Regression

Results

Model	RMSE	Number of Variables
Gradient Boosting with Income	50.47	24
Gradient Boosting	52.33	25
Linear Regression with Random Forest Variables and Income	54.31	34
Random Forest	54.46	33
Linear Regression with Random Forest Variables	56.09	33
Linear Regression with Gradient Boosting Variables	56.21	25

Recommendation:

Linear Regression
with
Gradient Boosting
Variables

```
LOSS AMOUNT
----------
Total Variables:  26
INTERCEPT  =  -18377.15587176956
TRUNC_bedrooms  =  22.186589518871884
TRUNC_accommodates  =  7.66286797694578
TRUNC_bathrooms  =  26.100763649437564
TRUNC_cleaning_fee  =  0.3241751673044986
room_type_Entire home/apt  =  35.17302348495243
latitude  =  -29.4042257179941
TRUNC_security_deposit  =  0.050007768495403226
TRUNC_reviews_per_month  =  -0.32734013531011075
neighbourhood_cleansed_Roosevelt  =  27.85373644516014
availability_365  =  0.029985155743971204
TRUNC_days_ago  =  0.1454742499394625
neighbourhood_cleansed_Southeast Magnolia  =  85.44292363476133
0_bedrooms  =  94.75780686137982
neighbourhood_cleansed_Belltown  =  8.665022263554167
missing_host_acceptance_rate  =  17.582700956157264
neighbourhood_group_cleansed_Downtown  =  26.2512169640533
0_security_deposit  =  28.09248521502876
property_type_Boat  =  136.55098947270224
longitude  =  -161.29162593652777
TRUNC_extra_people  =  0.07987849915654932
zipcode_98101  =  4.292004253479025
neighbourhood_group_cleansed_Capitol Hill  =  23.06621507971249
TRUNC_guests_included  =  3.6825542581678854
neighbourhood_group_cleansed_Cascade  =  27.85856132647873
0_bathrooms  =  3.796372591013543
```

Future Work

- Explore deeper into the calendar dataset
- Explore the relationship between neighbourhoods and price as well as other relationships to confirm intuitiveness of model
- Build a linear regression model with gradient boosting variables and income
- Add more features

Thank You!

Q&A