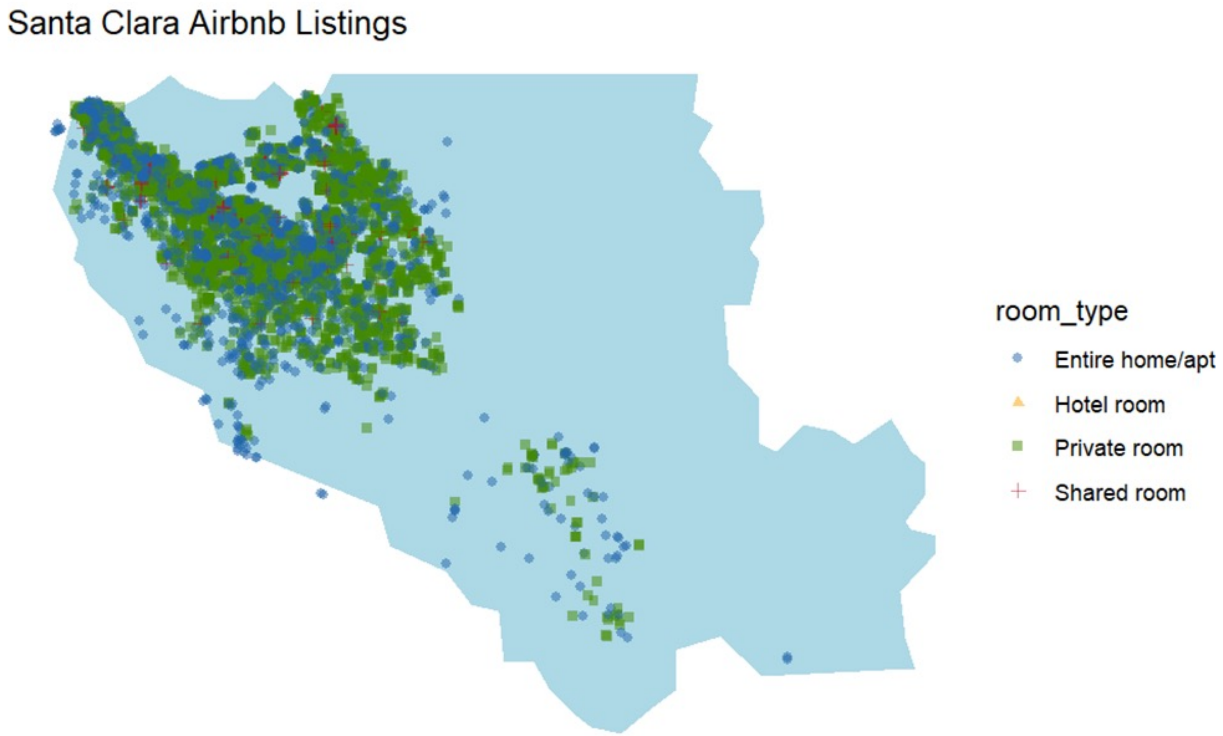
Bay Area Airbnb Analysis/Modeling

Yan He

Introduction

- Santa Clara County is one of the major counties in the Bay Area and home to many prominent tech companies and startups in the Silicon Valley. As such, home prices and cost of living are exceptionally high in this area
- In this project, I am going to use the Airbnb publicly available data to answer a couple business questions with some statistical analyses.
- The map on the left shows that most Airbnb listings locate in the Northwest region of Santa Clara



Outline

- Data Overview & Business Problems
- Exploratory Data Analysis (EDA)
- Regression with cross-validation (Linear, Ridge, Lasso, Random Forest, Boosting)
- Clustering (k-Means)
- Summary & Conclusion
- Further Discussion

Data Overview & Business Problems

Data Overview

- 3 datasets were given for this project, Airbnb Listings, Reviews and Neighborhoods. The Airbnb Listings data is the main data used for the analysis.

- There are around ~7000 listings listed by ~3500 hosts.

- There are ~100 variables in the given data, including various features of the listings such as price, room type, neighborhood, etc.

- To reduce the price skewness, when listing price is used as dependent variable in regression models, it is log transformed.

Data Overview

Initial Data cleansing/processing

1. Drop variables that are for sure not useful

2. Clean variable values (remove $ signs from currency variables; convert data type of date variables; clean the zip codes to extract valid 5-digit zip codes; remove unnecessary characters from the long string variables; change Boolean variable values from t/f to Y/N)

3. Check missing variables & impute some missing data (most variables of interest have valid data, missing reviews per month, bedrooms and bathrooms were filled with 0s)

4. Create new variables that might be useful for analysis (e.g. since it is unknown that the cleaning fee and security fee is per night or not, indicators of whether cleaning fee and security fee is required were created. Other new variables include bath_per_cap (bathrooms/accommodates), price_per_guest (price/guests_included).

Business Problems

A. Create a price-suggestion model for new Airbnb hosts & identify the important features that adds to the value (in terms of price).

B. Provide a listing-area suggestion algorithms for visitors based on their preferences.

Methods:

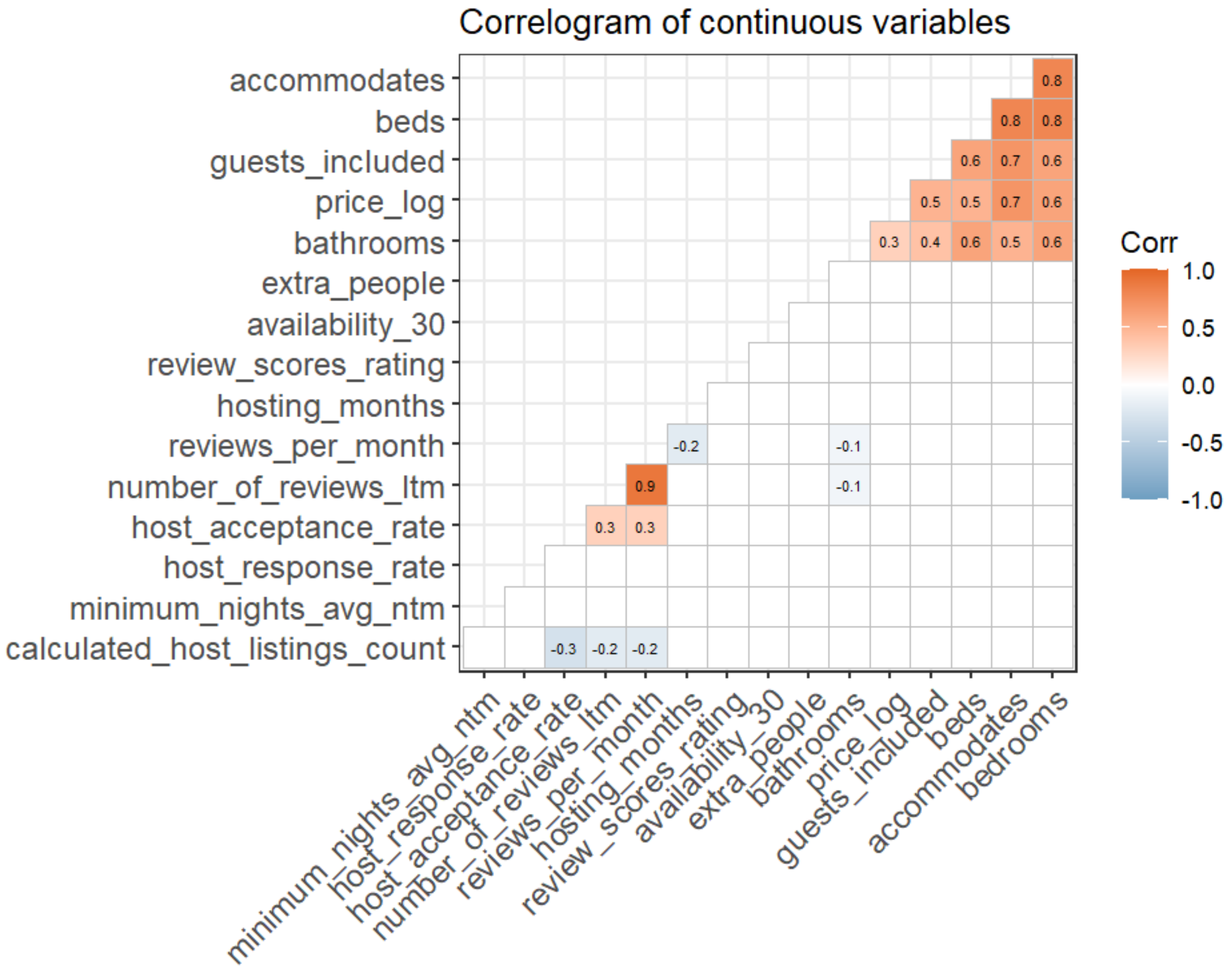
- For problem A, various regression models will be fitted with price (log-transformed) as dependent variables.

- For problem B, K-means will be employed to cluster zip-code areas based on various features.

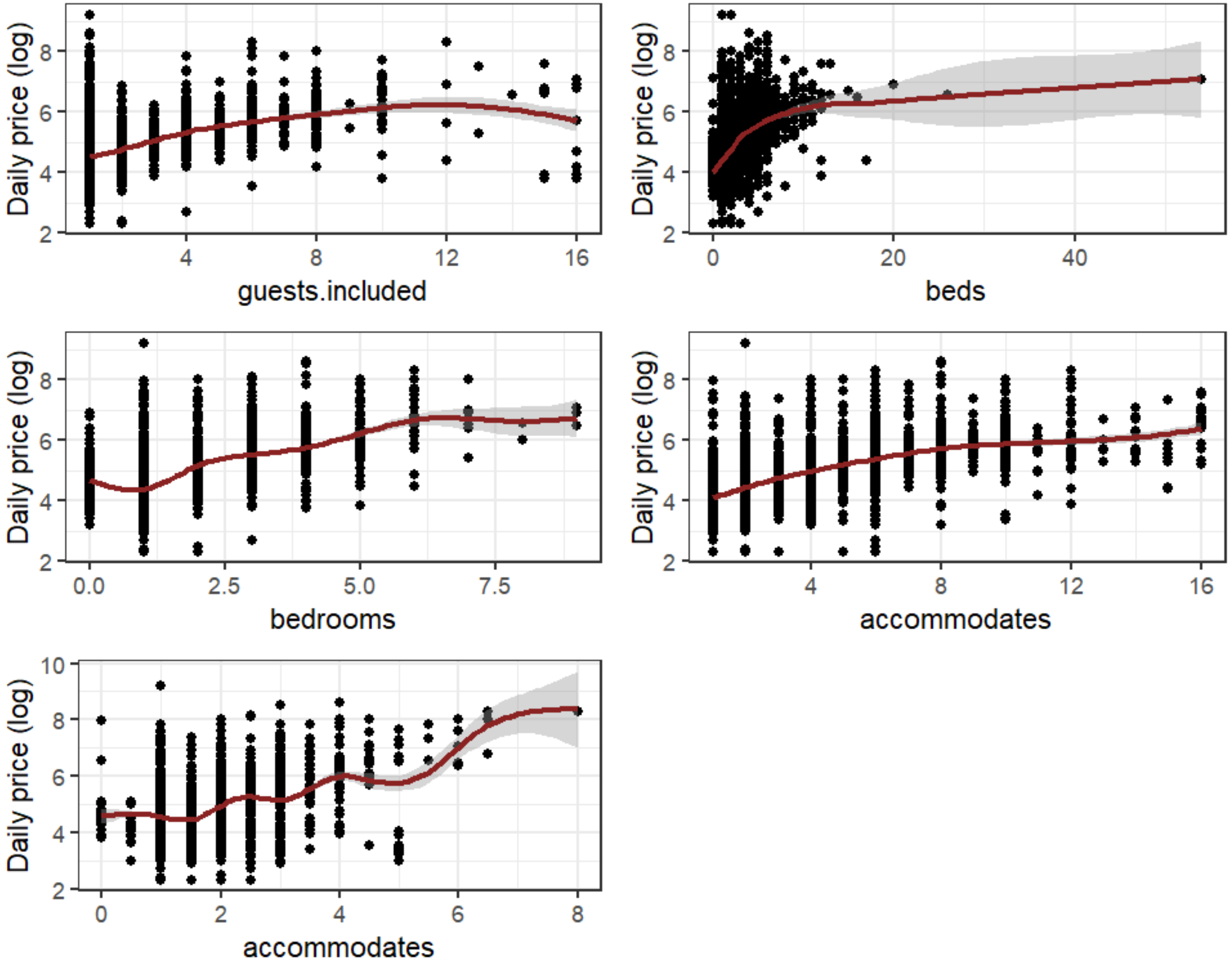
Exploratory Data Analysis (EDA)

Correlogram
(Correlation of numeric variables)

- This plot presents the correlation of all the numeric variables of interest (should focus on the bottom-right grids below the diagonal)
- Labels on each grid are the correlation numbers, blank grids (among the ones below the diagonal) indicate non-significant correlation



Scatter plots

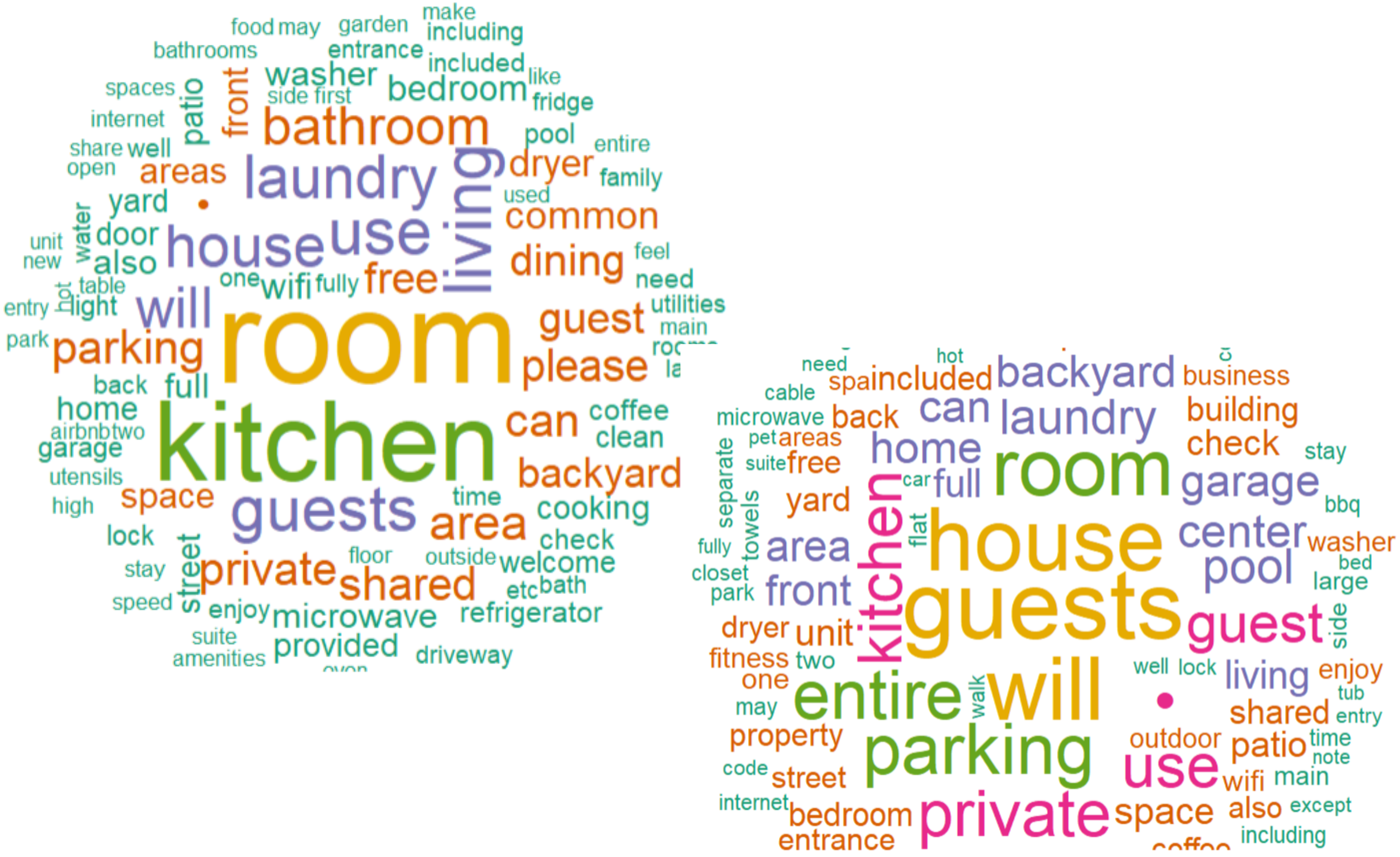


- The scatter plots of the log-transformed daily price vs other numeric variables that were indicated correlated with the dependent variable (log price).

- It seems all these 5 variables have some positive relationship with (log) price.

Word cloud -- “access” of the listings

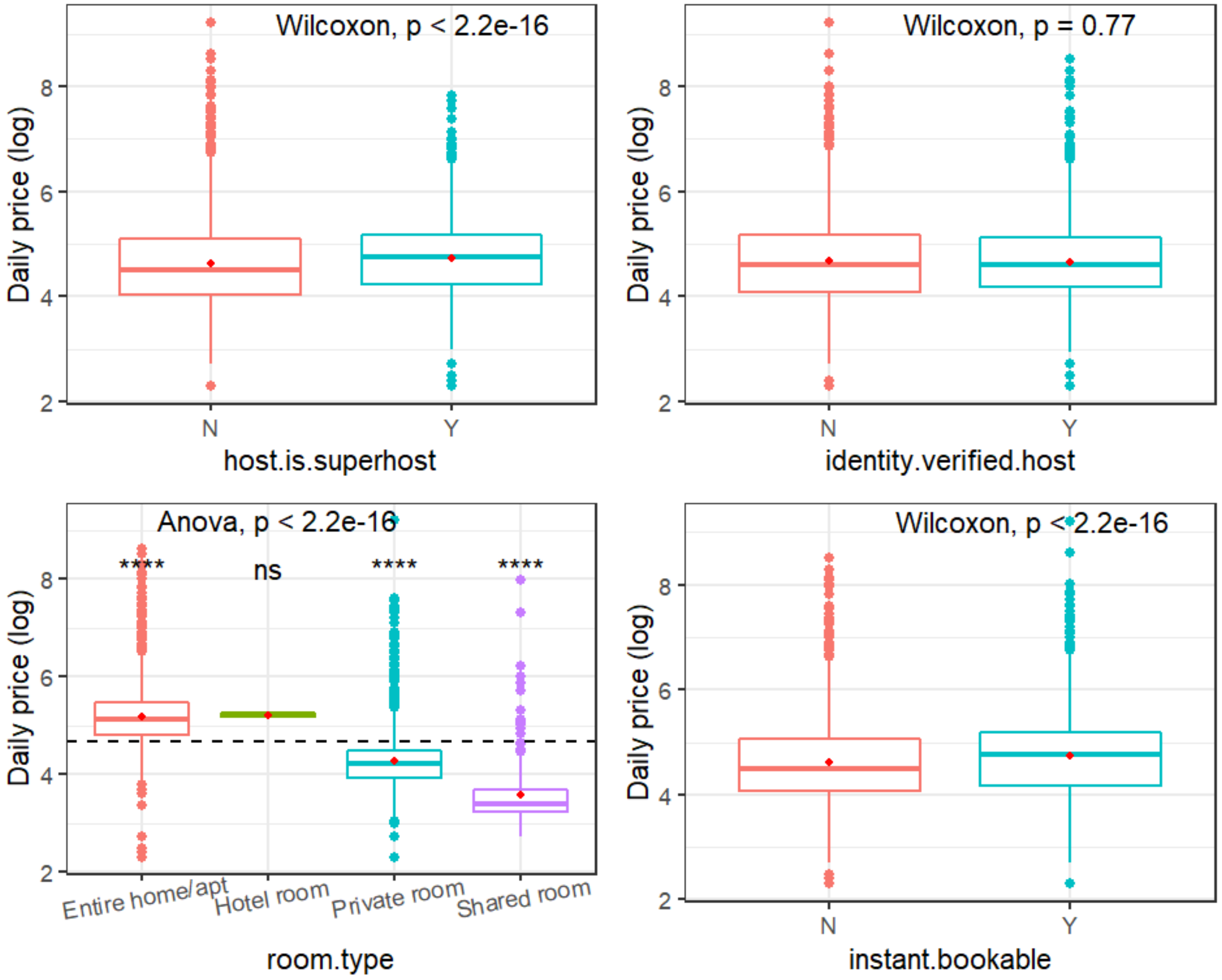
- World cloud of the “access notes” for listings whose (log) price is above median and below median.
- For the listings that have higher price, the visitors can have access to pool and garage.
- Create another new variable indicating whether the listing is pool/garage accessible which might be useful in the price suggestion model



Boxplot
(price difference among categorical variables)

- Boxplots to compare the mean (log) price among different groups.

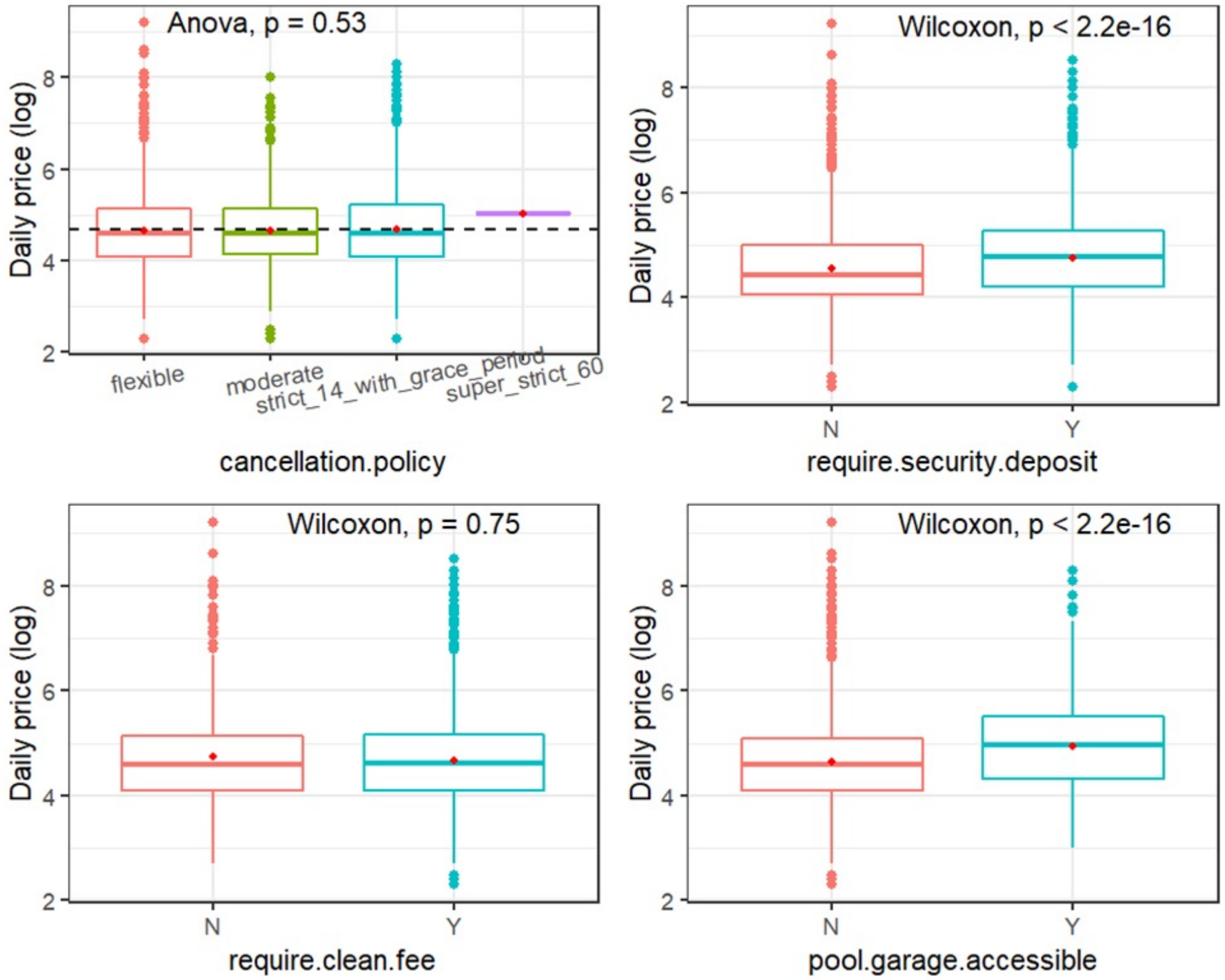
- On top of each plot, it presents the test result of the difference--for 2-level categorical variables, Wilcoxon tests were performed, for groups with more than 2 levels, ANOVA tests were performed (t-test results of the comparison between group mean and overall mean were show on top of the boxplot as well for groups with more than 2 levels)



Boxplot
(price difference among categorical variables)

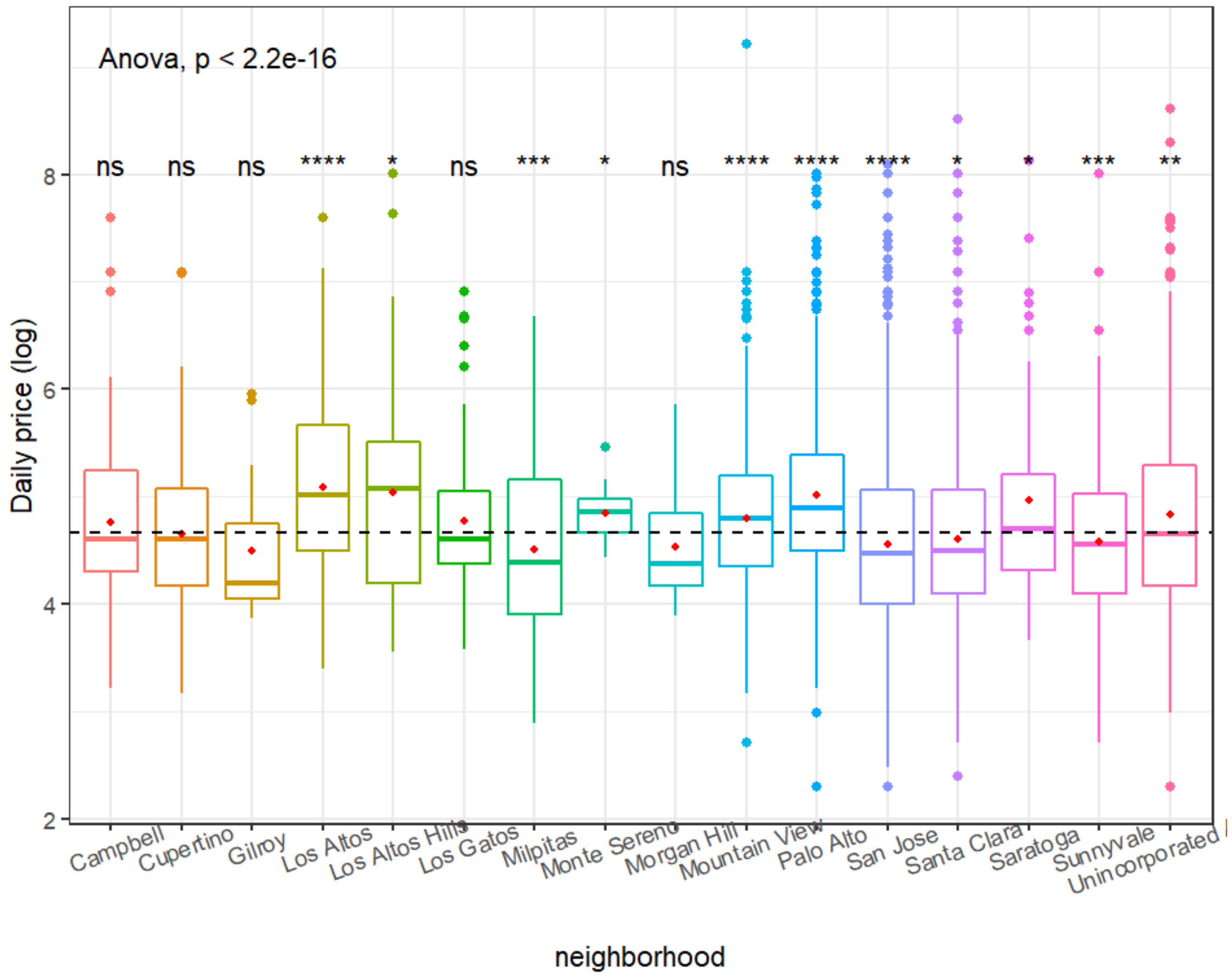
- Whether the host identity is verified or not does not affect the (log) price (p=0.77)

- Cancellation policy and whether require cleaning fee do not impact the (log) price, either (p=0.53, 0.75, respectively)

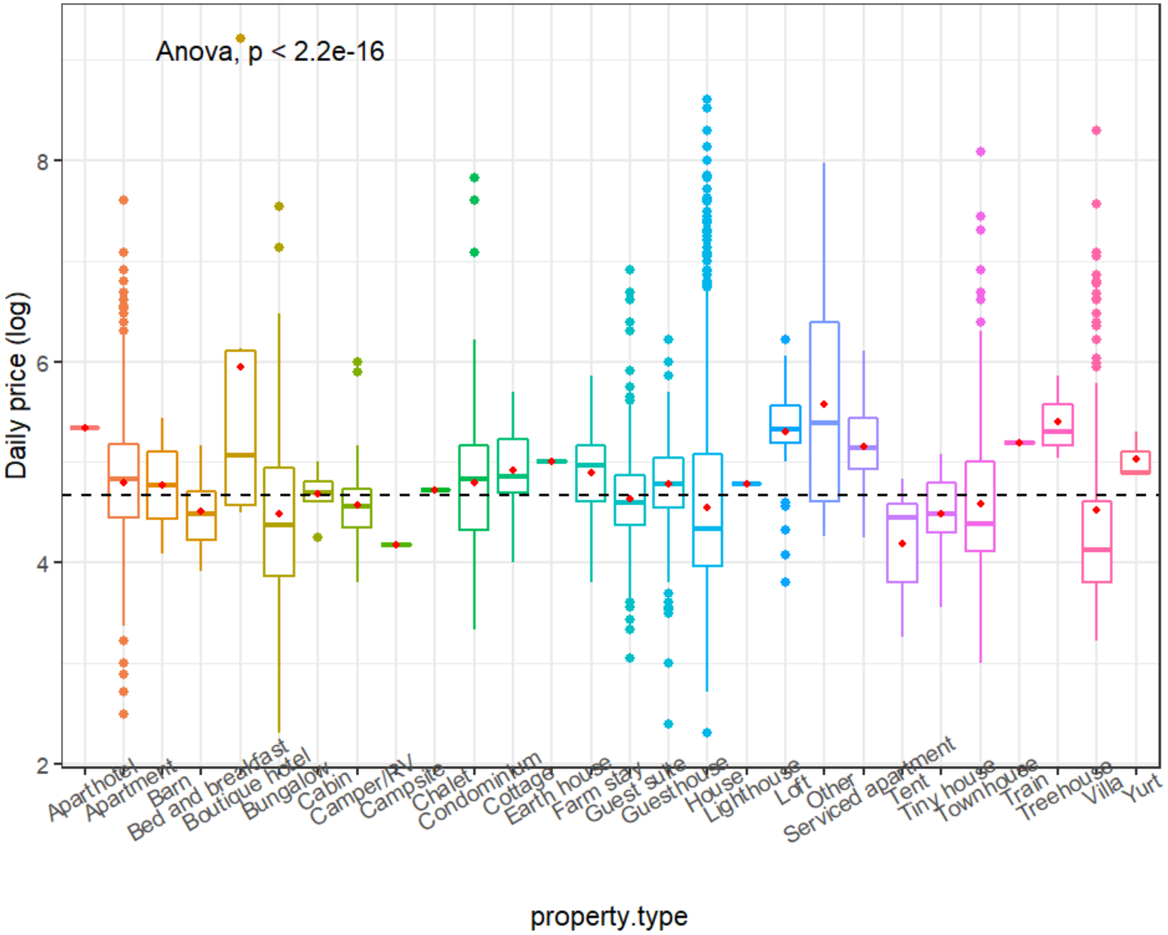


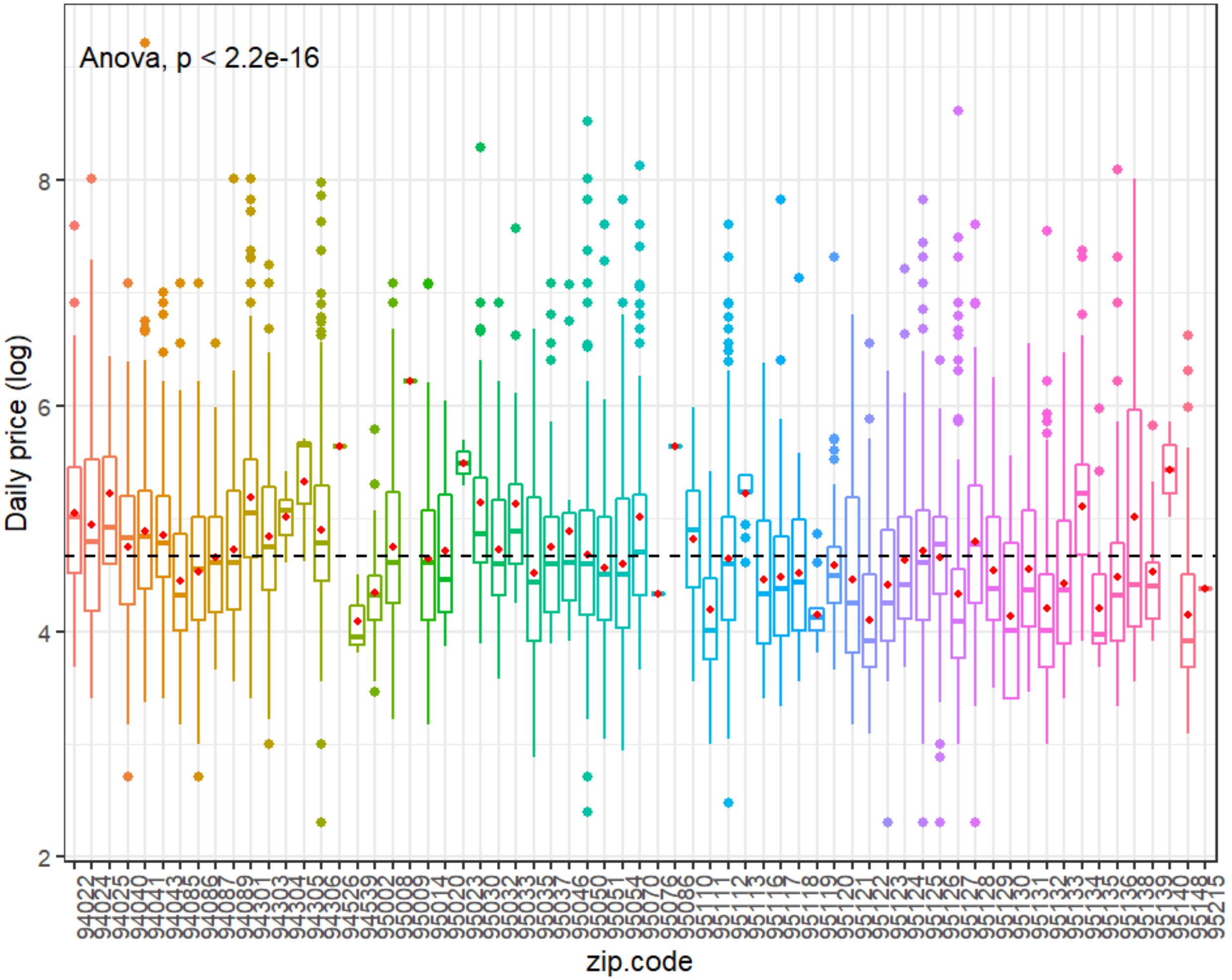
Boxplot
(price difference among categorical variables)

- There seems to be significant difference of the mean (log) price among different neighborhoods, among different property types, and among different zip codes.



Boxplot
(price difference among categorical variables)



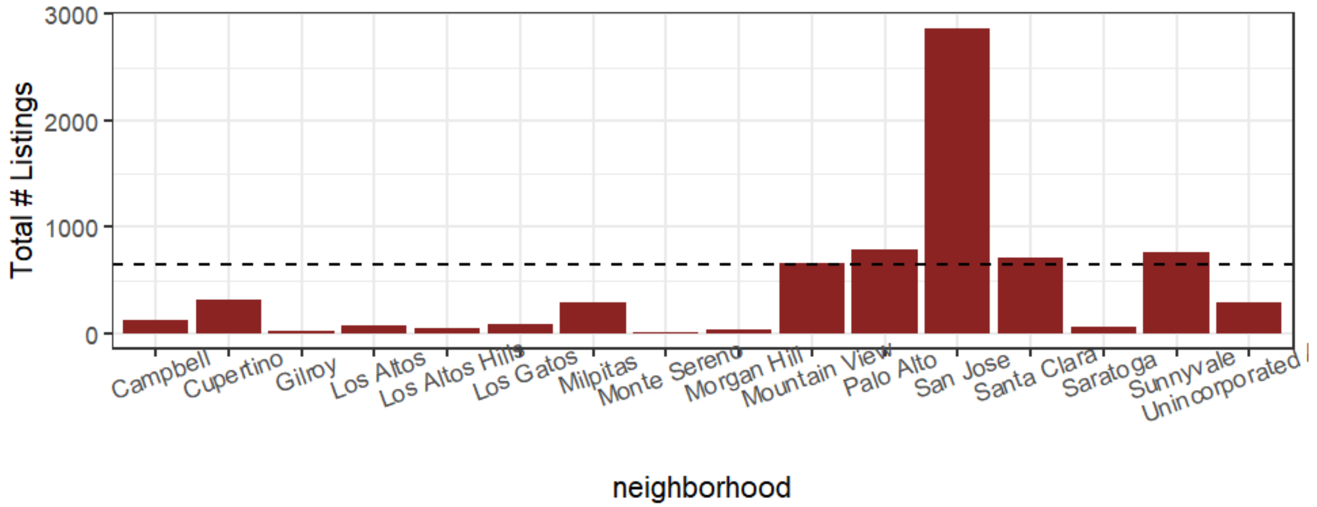


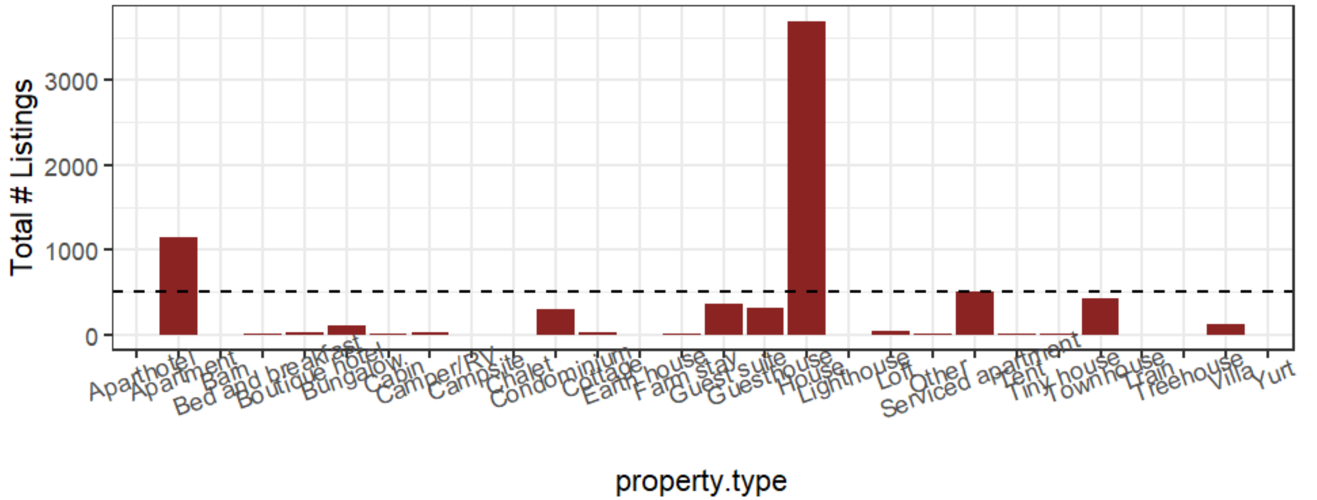
Bar plot
## (# listings by neighborhood & property type)

- Check the frequency (# listings) of each categories for neighborhood and property type

- Combine some categories with very small number of listings (this would help with the cross-validation in the modeling)

- Did not keep zip codes in the modeling process, considering it has too many levels, and it may provide overlapping information with neighborhood

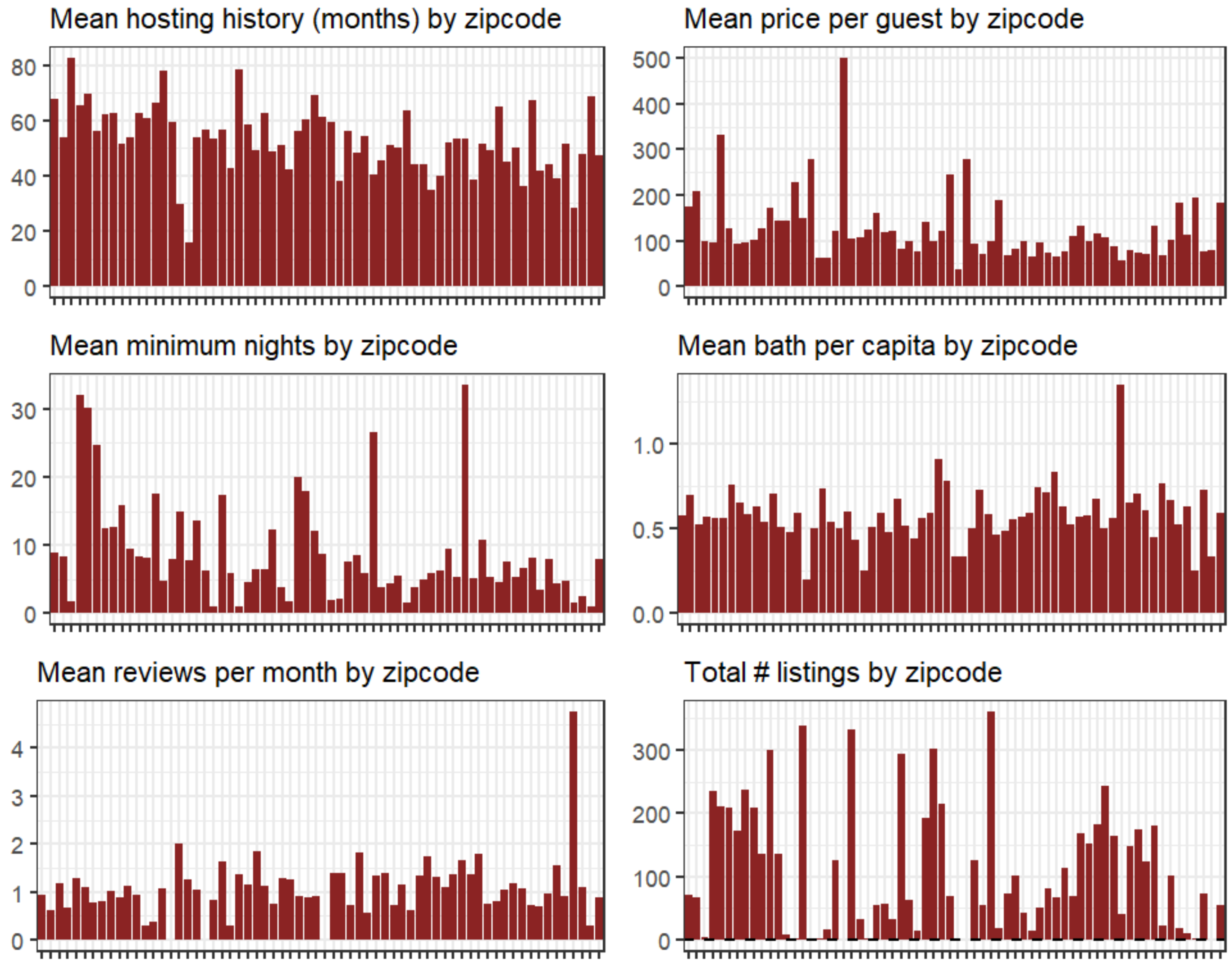




Average/overall features within zip codes

- Bar plots present the average hosting age of listings (months), price per guest, minimum nights, bath per accommodate/capita, reviews per month by zip code.

- Also, total number of listings within each zip code



Mean hosting history (months) by zipcode

Mean price per guest by zipcode

Mean minimum nights by zipcode

Mean bath per capita by zipcode

Mean reviews per month by zipcode

Total # listings by zipcode

Regression Modeling with Cross-validation

(Linear regression, Ridge, Lasso,
Random Forest, Boosting)

Cross-validation (CV)

- For all the regression models, 10-fold cross-validation (CV) was applied to estimate the test error and compare the performance across all the models.

- The data was split into 10 folds, the models were fitted 10 times, each time 9/10 folders were taken as training data and the rest 1 folder of data as testing data.

- For Ridge & Lasso, every time during the 10 times the models were fitted, there would be a best lambda identified. The average of the 10 (best) lambdas would be considered the final best lambda for the corresponding model.

- For Random Forest and Boosting, different sets of tuning parameters were tried along with cross validation to find the best parameters for the model.

Model	Test Error (from CV)
Linear Regression	0.251172
Ridge	0.251999
Lasso	0.251153
Random Forest	**0.216017**
Boosting	0.224952

Random Forest performs best since it has the smallest mean test error.

Linear Regression, Ridge & Lasso

Ridge - CV

k	bestlam	error
1	0.047526	0.236955
2	0.048176	0.20259
3	0.048362	0.288564
4	0.048256	0.29287
5	0.049081	0.236959
6	0.048123	0.282706
7	0.048098	0.233371
8	0.04799	0.246793
9	0.047991	0.259913
10	0.047548	0.23927

Lasso - CV

k	bestlam	error
1	0.000586	0.234485
2	0.000594	0.20123
3	0.000654	0.286889
4	0.000595	0.292429
5	0.000605	0.236945
6	0.000715	0.285281
7	0.000651	0.233959
8	0.000592	0.245186
9	0.000592	0.257165
10	0.000586	0.237961

The lambda is pretty close to 0, so the Lasso and Ridge models should be very close to (OLS) Linear Regression

Best lambda for Ridge regression remains consistent during the process of CV, ~0.048

Best lambda for Lasso regression remains consistent as well during the process of CV, ~0.0006

Linear Regression, Ridge & Lasso

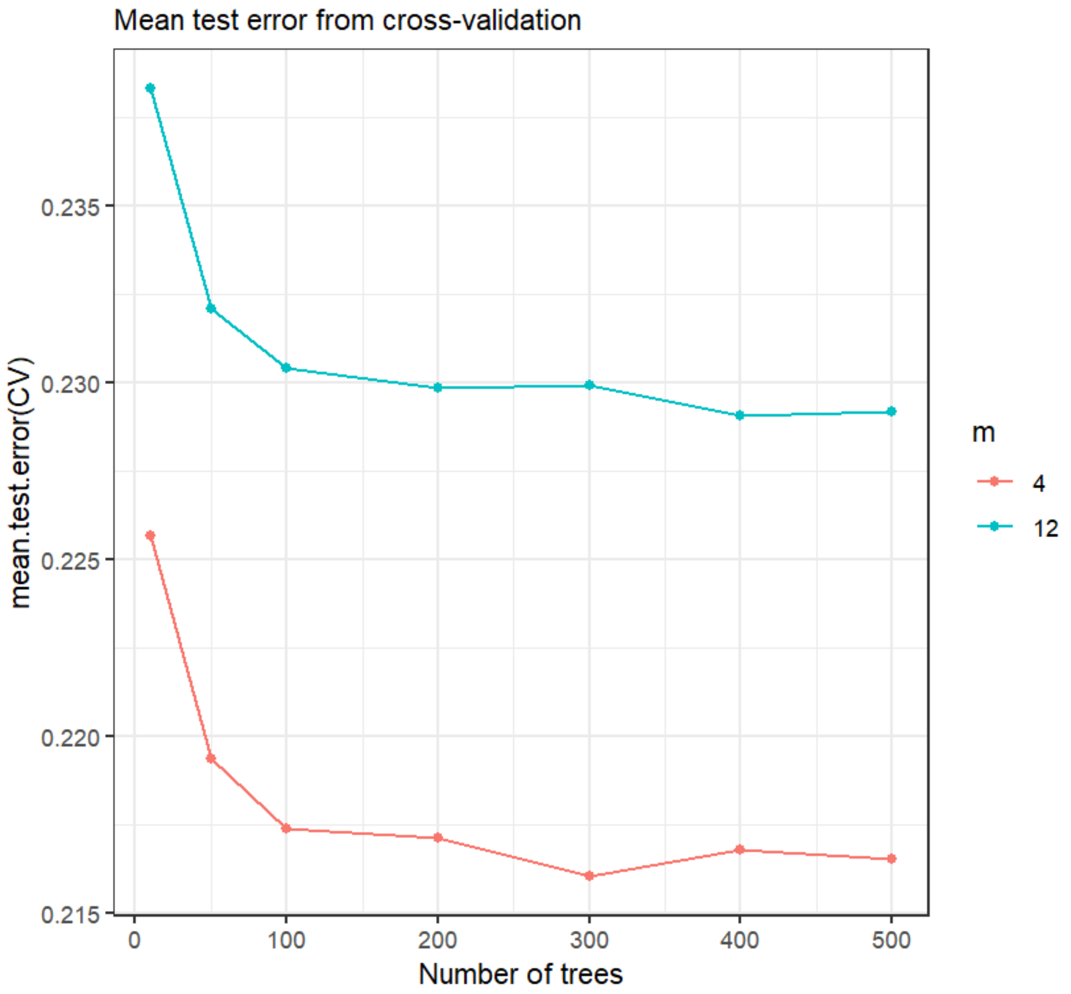
	Lasso	Linear Regression	p-val (linear)
(Intercept)	4.63711967	4.6456***	< 2e-16
guests_included	0.01111772	0.0114*	0.0173
beds	-0.0124261	-0.0147*	0.0194
accommodates	0.07641555	0.0771***	< 2e-16
bedrooms	0.17099742	0.1722***	< 2e-16
bathrooms	0.02543213	0.0266*	0.0260
host_is_superhost: Y	-0.04634464	-0.0481***	0.0001
property.type: House	-0.06733584	-0.0656***	0.0007
property.type: Other	0.0227837	0.0258	0.1759
room_type: Other	-1.23925429	-1.2418***	< 2e-16
room_type: Private room	-0.52451087	-0.5256***	< 2e-16
instant_bookable: Y	**0**	-0.0005	0.9705
require_security_deposit: Y	-0.07097435	-0.0731***	0.0000
pool_garage_access: Y	0.07088259	0.0725***	0.0001
neighbourhood: Other	-0.07445506	-0.0858***	0.0004
neighbourhood: Palo Alto	0.14883473	0.1407***	0.0000
neighbourhood: San Jose	-0.18731243	-0.1979***	< 2e-16
neighbourhood: Santa Clara	-0.06818132	-0.0794**	0.0039
neighbourhood: Sunnyvale	-0.09596037	-0.1071***	0.0001

The table presents the coefficients from Linear regression and Lasso. The regression results from both models are pretty similar.

The coefficient for most of the variables are significant. E.g. **The listings requiring security deposits are normally cheaper**. Compared to Mountain View, only listings in Palo Alto are more expensive.

Reference group for property type is Apartment, for room type is Entire home/apt, for neighborhood is Mountain View

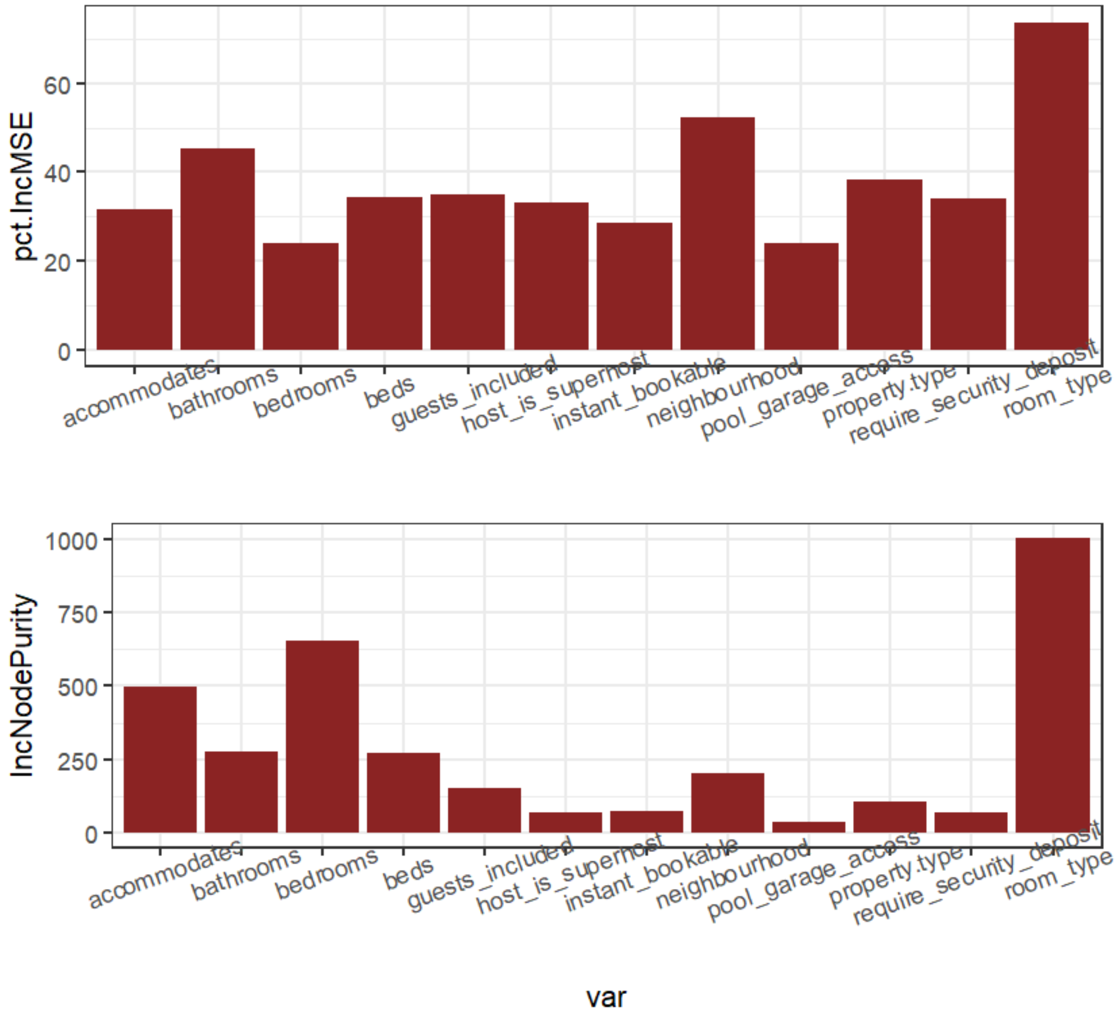
Random Forest



In Random Forest model, the tuning parameters include m (# predictors as split candidates during the tree splitting), and number of trees.

When m is 4 and number of trees is 300, the resulting cross-validation (mean) test error is the smallest, which indicates the model performs best with these parameter selection

Random Forest

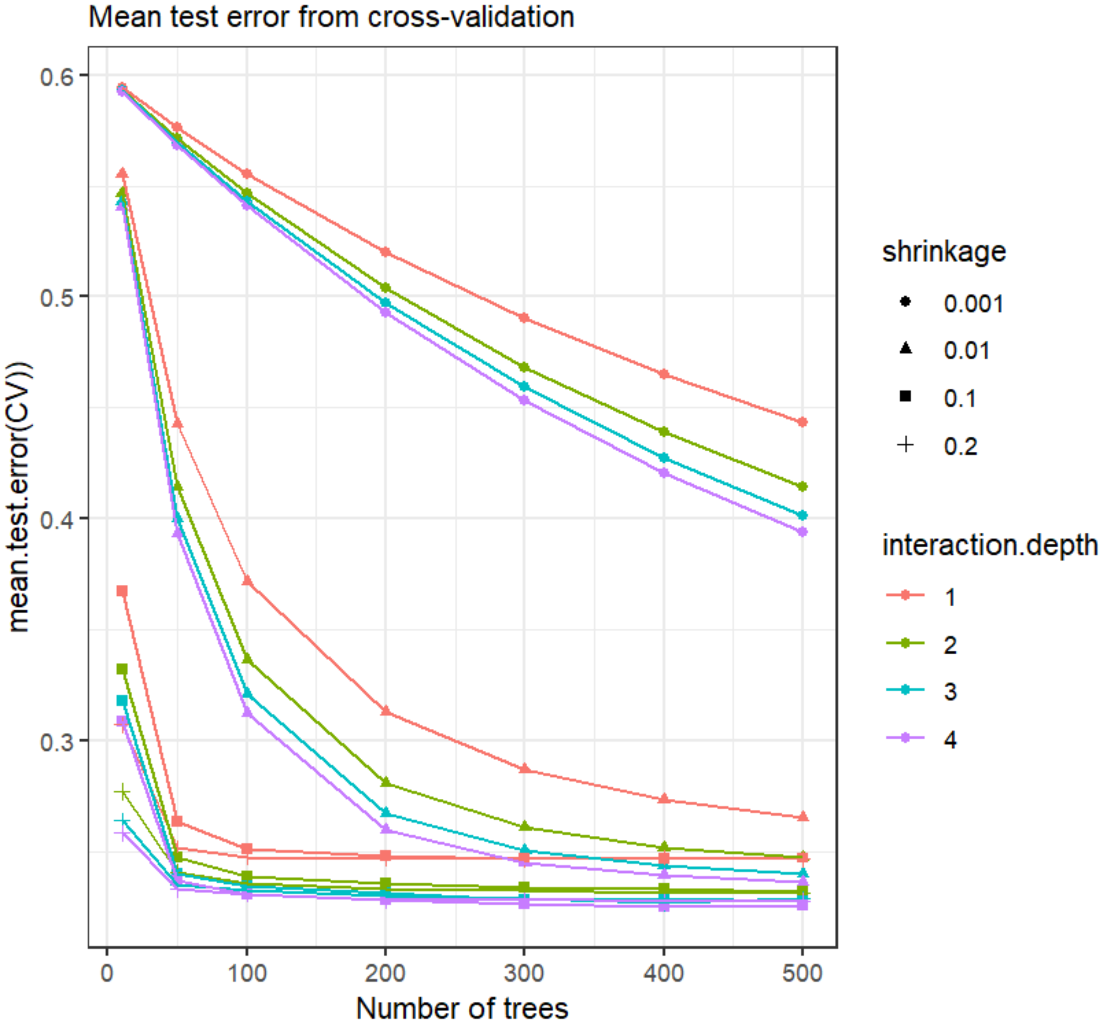


The plots on the left plot the variable importance in terms of %IncMSE (mean decrease of accuarcy in predictions on OOB samples without given variable), and IncNodePurity (training RSS).

Both measures are the bigger the better.

Room type seems to be the most important variable, accommodates and bedrooms are also relatively important.

Boosting



In Boosting model, the tuning parameters include shrinkage, interaction depth, and number of trees.

When shrinkage is 0.1, interaction depth is 4, and number of trees is 400, the resulting cross-validation (mean) test error is the smallest, which indicates the model performs best with these parameter selection

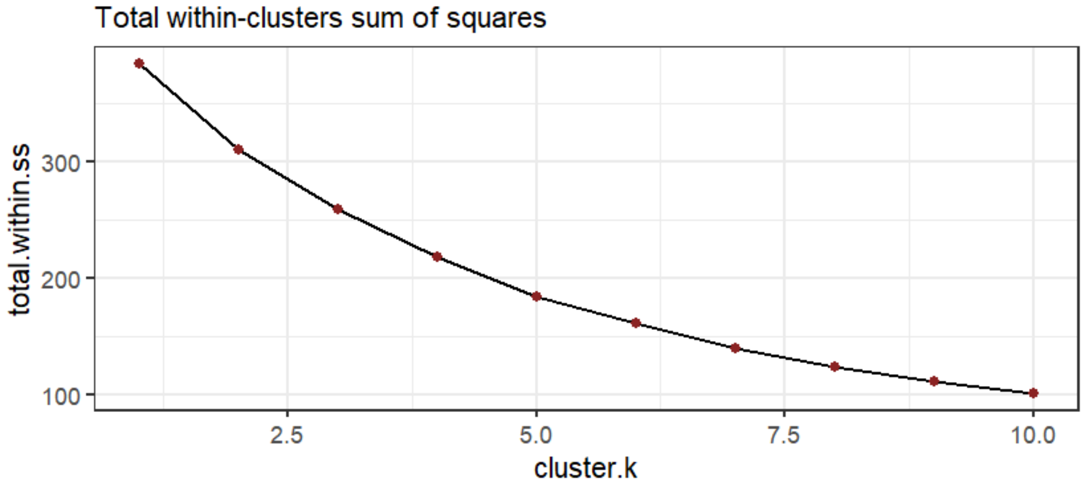
Clustering (K-Means)

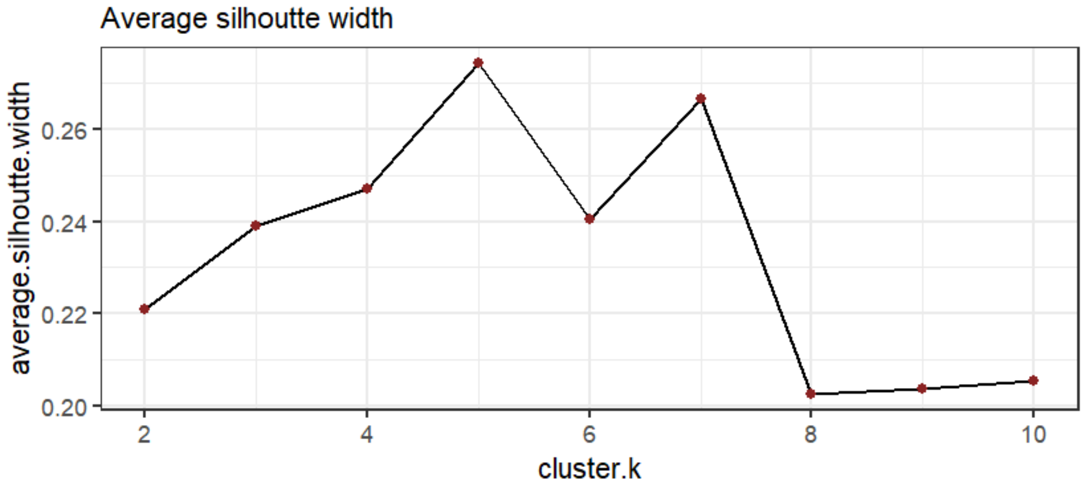
Clustering zip code areas

- First derive some desired zip-code level features based on some aggregation by zip codes (most features are averages, except for total number of listings in a zip code). All features are scaled/standardized before clustering.

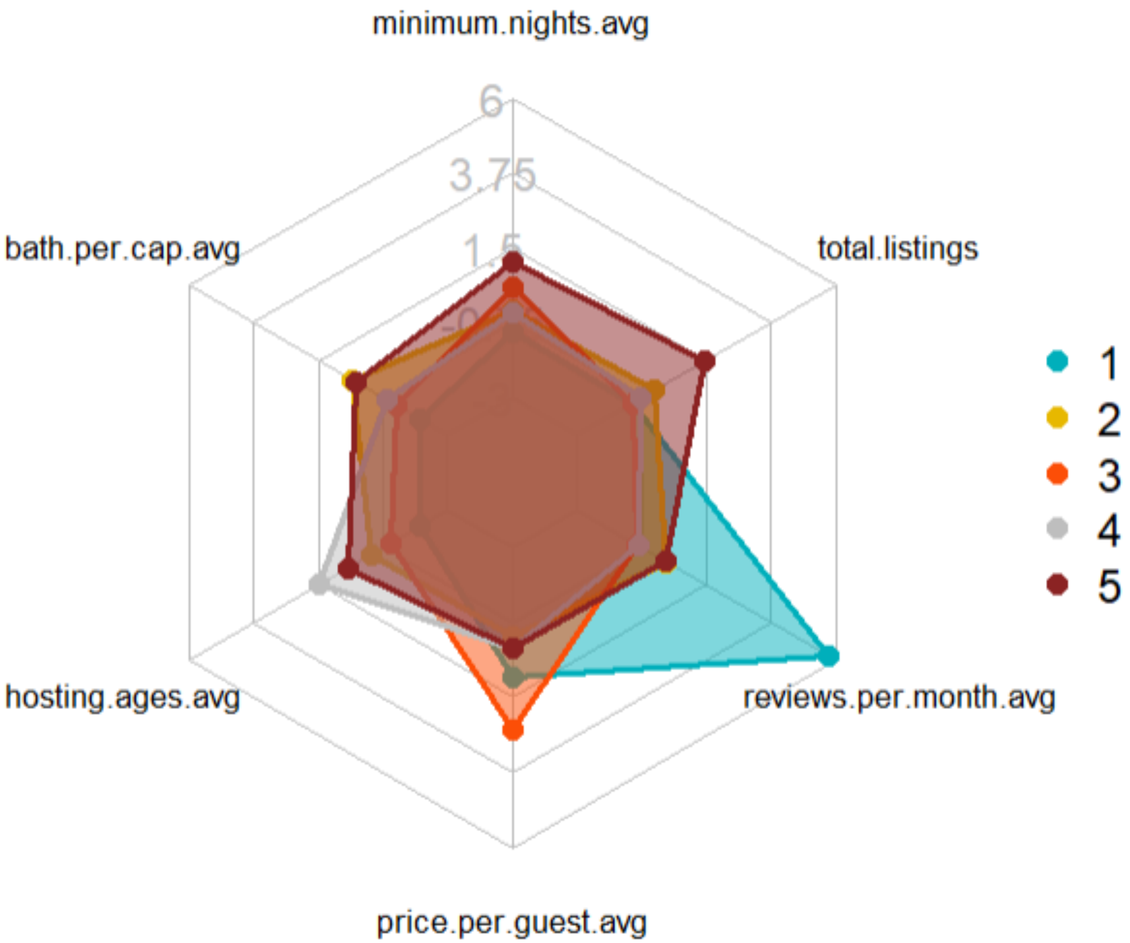
- Use K-means clustering to cluster the zip codes. Use both Elbow Method (minimize the total within-cluster variation/ss) and Silhouette method (cluster-quality measurement. A high average silhouette width indicates a good clustering) to find the optimal number of clusters.

- Seems 5 is the optimal number of clusters





Clustering zip code areas



- The Radar chart plots the mean of the features in each cluster. Based on the plot, cluster 5 has the most listings per zip code.
- If the visitors want to live in younger listings, they should check zip code areas in cluster 1,3 or 2.
- If they care about price per guest, areas in cluster 4 or 5 have better deals.
- If they prefer more bathroom space per person, listings in 5 or 2 are more likely to fit the visitors.

Clusters

- **Cluster 1**: "95140"
- **Cluster 2**: "94022" "94025" "94304" "94305" "95023" "95076" "95134" "95215"
- **Cluster 3**: "94526" "95009" "95086" "95113“
- **Cluster 4**: "94040" "94041" "94043" "94085" "94086" "94087" "94301" "94306" "95014"
- "95035" "95050" "95051" "95054" "95112" "95126" "95128“
- **Cluster 5**: all other 36 zip codes.

Summary & Conclusions

- Room type, accommodates, bedrooms are important factors that would affect the listing price.
- The performance of all the regression models are similar. Random Forest outperform over other models slightly based on the mean test error from cross-validation.
- There are more than 60 zip code areas in Santa Clara, but they can be clustered based on similar features. Visitors can choose or the platform can suggest listings in certain clusters based on visitors' preference.

Further Discussion

Clustering?

- Look into more details about each cluster of the zip codes.
- Maybe can cluster all the listings instead of cluster zip codes, so the suggestion system can be more specific (to specific listings). Or identify the listings with relatively lower review score.

Text mining of review comments

- Analyze the comments can help understand what impact visitors' satisfaction or help predict the review score.

Model improvement

- More testing with the models using public data updated in later months
- Directly predict the non-log-transformed price and try other advanced ML models such as XGBoost.

Thank you!

Yan He

****@******.edu

https://yanhe.me/