

Prescient Price:

Using Machine Learning to Predict the Price
of Airbnb Listings



Claire Miles

Introduction

Airbnb

- Service that allows people to rent their homes or rooms to tourists
- Offerings and experiences in listings vary widely
- What makes a successful listing?
- Successful = you can charge more = more **\$\$\$\$**

How can we predict
an Airbnb listing
price based on other
information about
the listing?

This Project

- Create a predictive model using machine learning
- Tools: Python - pandas, numpy, matplotlib, scikit-learn, BeautifulSoup, vaderSentiment
- Steps:
 - a. Data collection
 - b. Data Cleaning
 - c. Exploratory Analysis
 - d. Feature Engineering and ML Modeling

Data Collection

[show](#) archived data

Los Angeles, California, United States

See [Los Angeles data visually here](#).

Date Compiled	City	File Name	Description
06 March, 2019	Los Angeles	listings.csv.gz	Detailed Listings data for Los Angeles
06 March, 2019	Los Angeles	calendar.csv.gz	Detailed Calendar Data for listings in Los Angeles
06 March, 2019	Los Angeles	reviews.csv.gz	Detailed Review Data for listings in Los Angeles
06 March, 2019	Los Angeles	listings.csv	Summary information and metrics for listings in Los Angeles (good for visualisations).
06 March, 2019	Los Angeles	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Los Angeles	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from <small>city or open source GIS</small>

Methods

- Web Scraping ([insideairbnb.com](https://www.insideairbnb.com))
- Consolidate monthly data into single file, delete duplicate rows
- Save file

Data Cleaning

Methods

- Dropping unnecessary columns
 - Non-categorical text data (except for review text)
 - Columns with redundant information (e.g. city, region, state)
- Fixing spelling and formatting inconsistencies for categorical data
- Amenities feature: list -> multiple features
- Changing columns to appropriate data types
- One hot encoding
- Add review text -> single dataframe & file for all data

Null/Missing Values

- Target variable: Price
 - Delete all rows with no price
- Other columns
 - Zip code: fill with mode of column
 - Non-boolean: fill with median
 - Boolean: fill with "False"



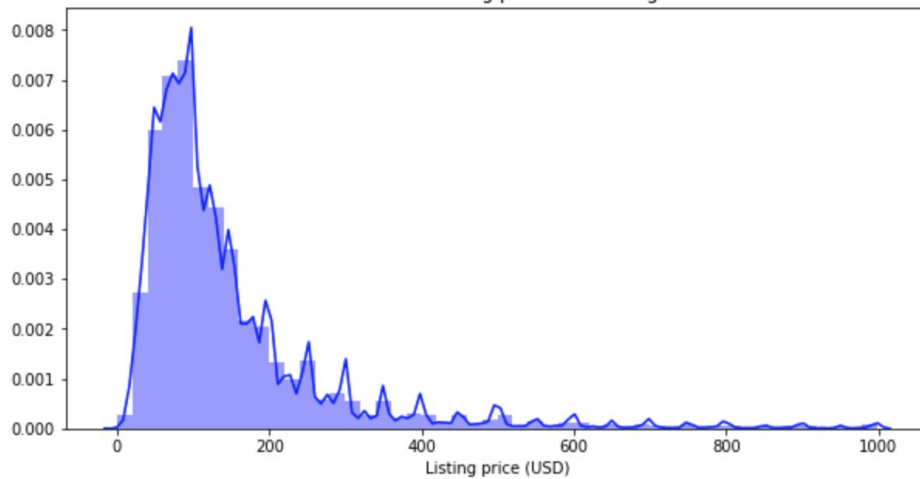
Exploratory Data Analysis



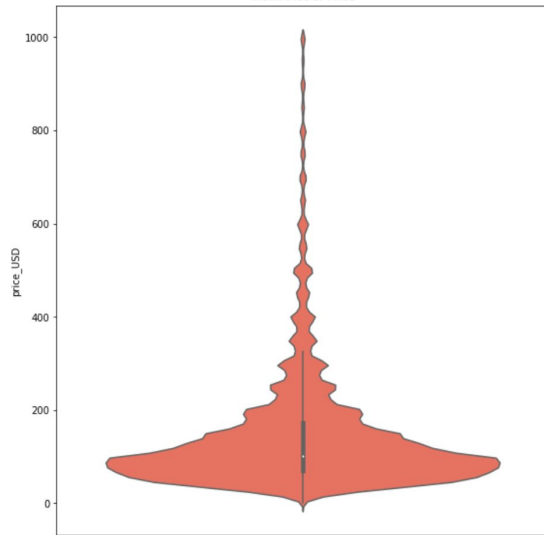
Methods

- Use histograms, box plots, violin plots, and maps to visualize distributions of variables:
 - Price
 - Accommodates
 - Property Type
 - Square Feet
- Preliminary Feature Importances

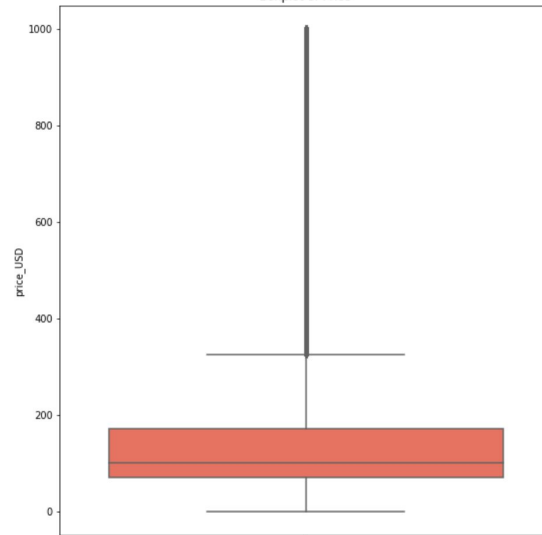
Distribution of listing prices in Los Angeles



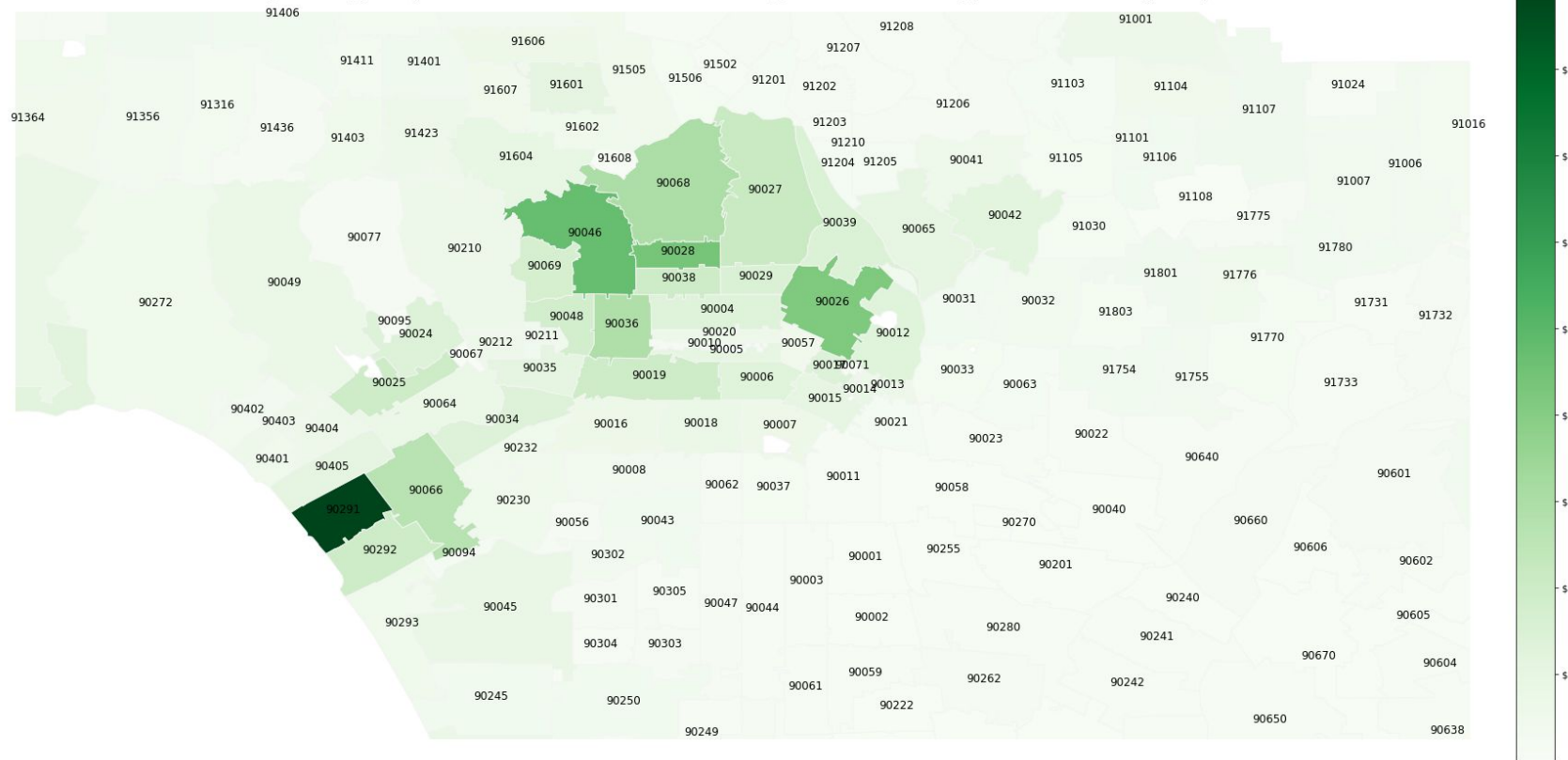
Violin Plot of Price



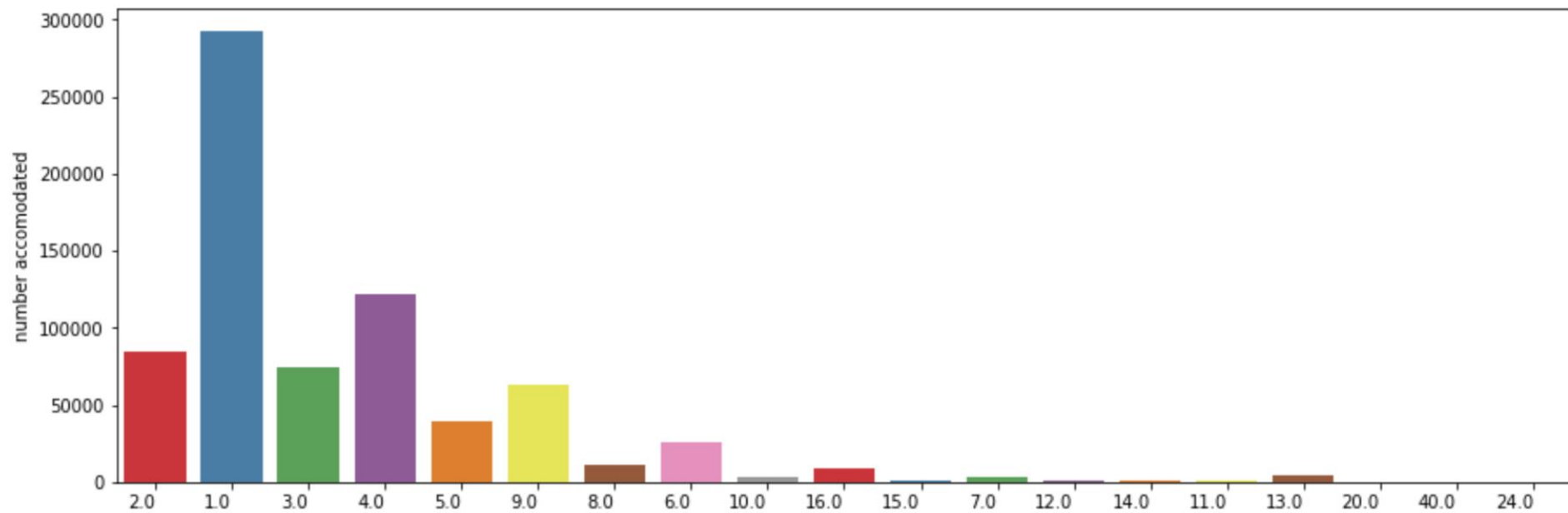
Boxplot of Price



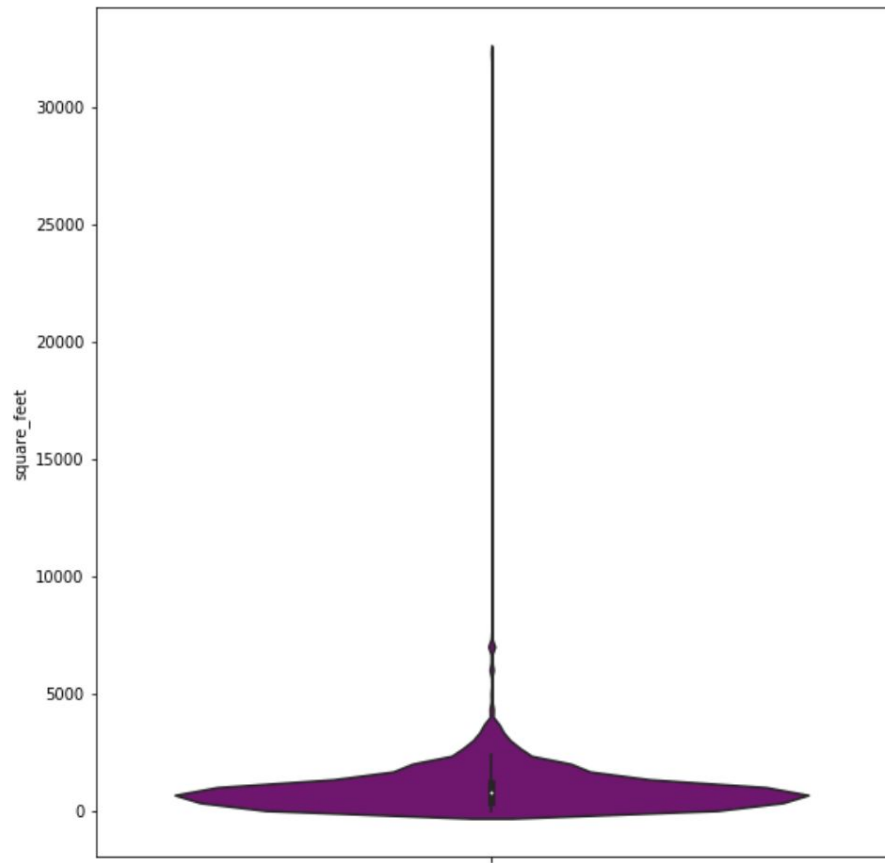
Average Nightly Rate of Airbnb Listings in Los Angeles, CA by Zip Code



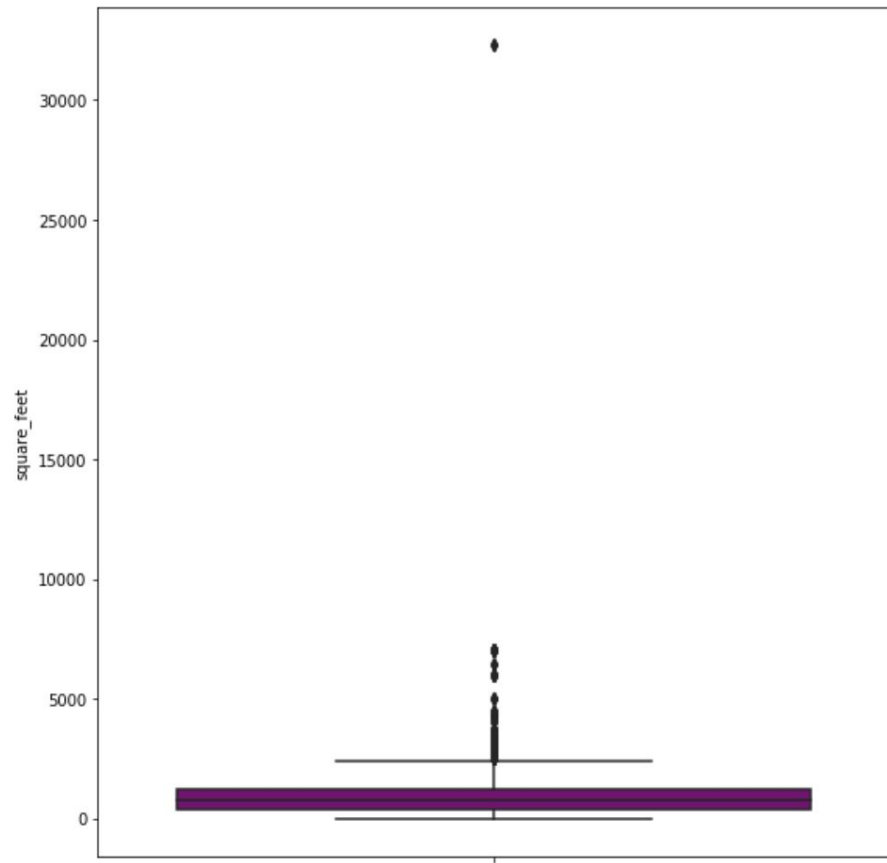
Accommodates



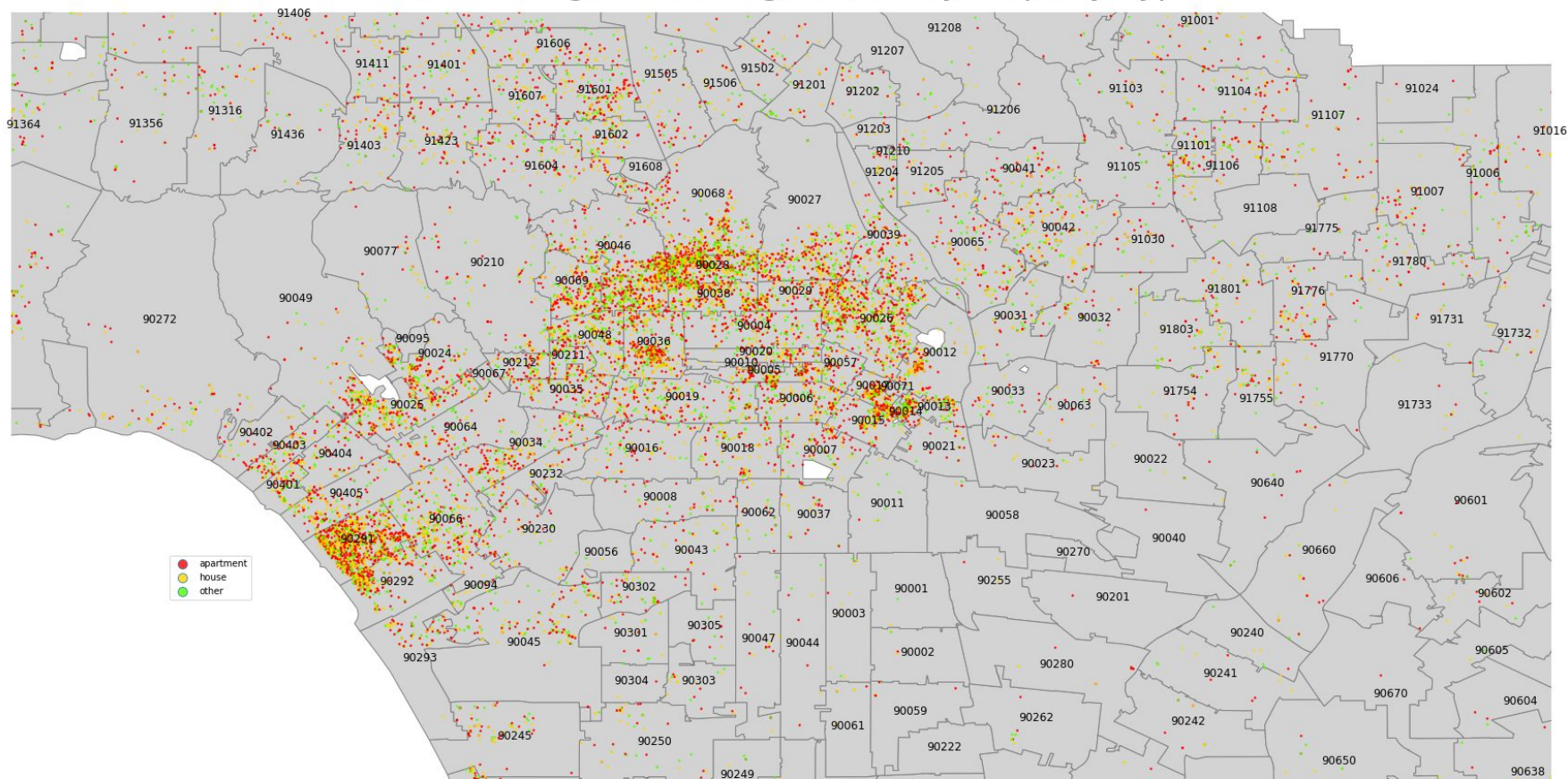
Violin Plot of Square Feet

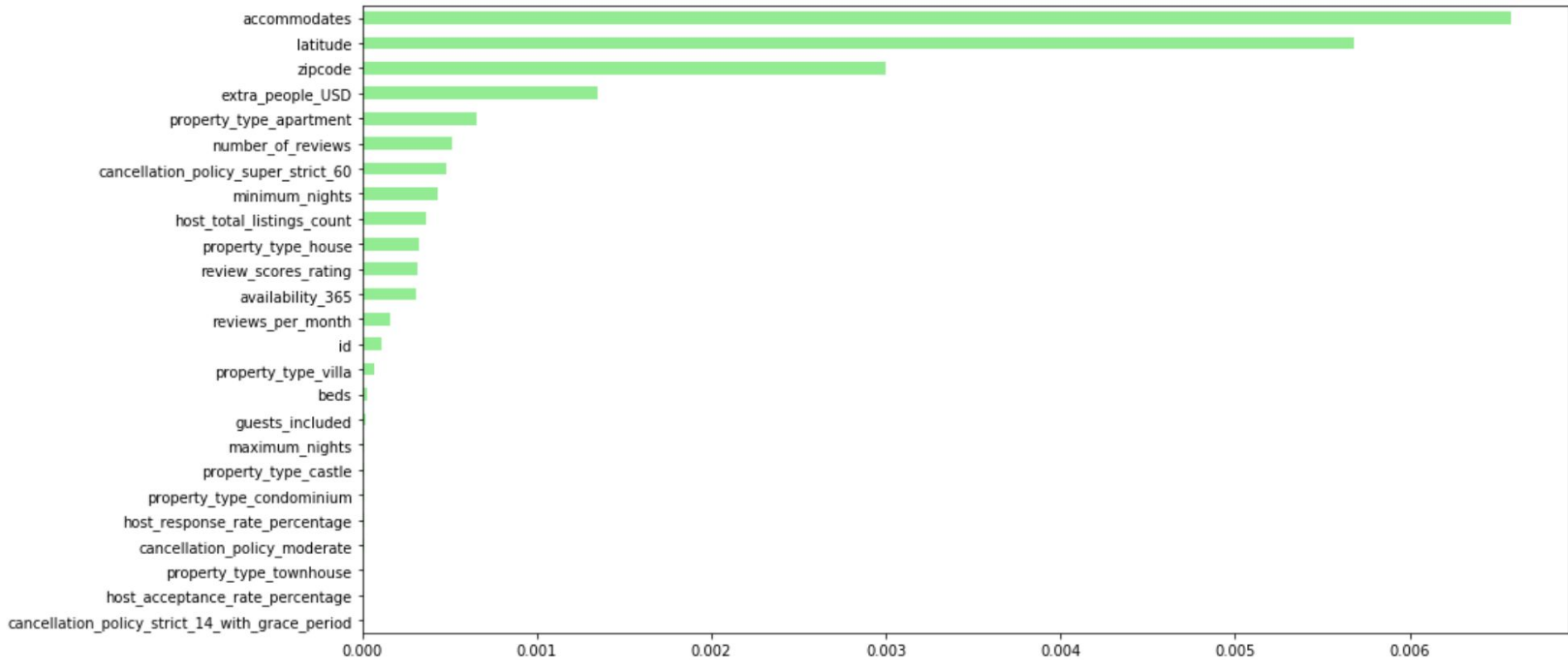


Boxplot of Square Feet



Airbnb Listings in Los Angeles, CA by Property Type





Data Story

- Most listings are very similar!
 - Apartments
 - For one person
 - Small in size
 - Downtown or by the beach



Machine Learning Modeling



Methods

- Text data -> numerical features
- Try three types of models (root mean squared error metric)
 - Random Forest: 71.9
 - Stochastic Gradient Boosting: 43.9
 - XGBoost: 44.7
- Tune SGB and XGBoost with cross validation
- Test on never-before-seen dataset
- Extract tangible insights about target variable: **price**

Most positive reviews:

Great

Perfect 🍷

Marcy was great very sweet!

Amazing spectacular experience , highly recommend

Awesome host. Great communication and accommodation skills. Made sure I was taking care of.

Least positive reviews:

普段通りの生活ができ、とてもすごしやすかった

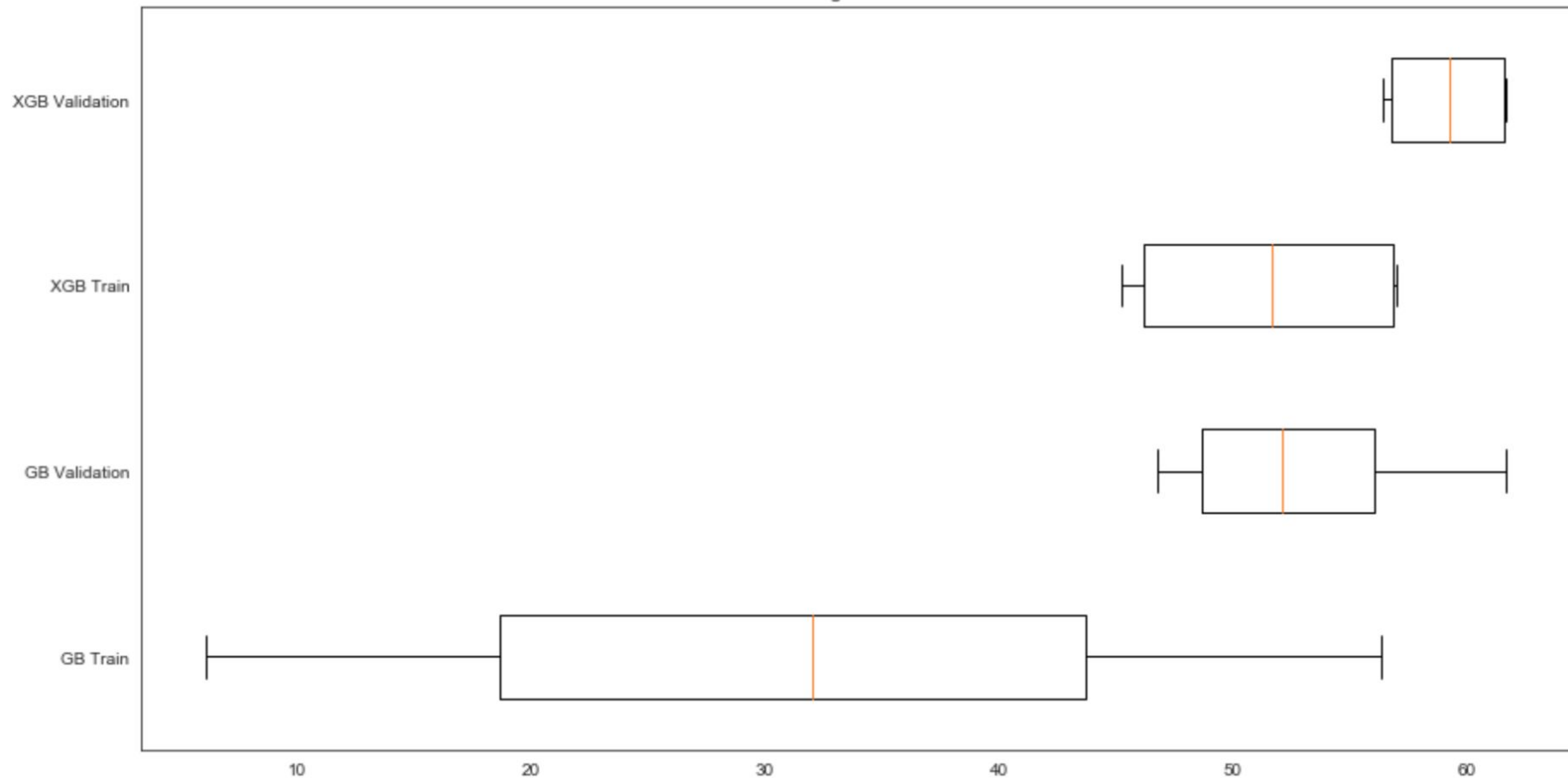
房间和照片中的一样，房东想的很周到，并且带我们去了附近的亚洲超市，非常感谢！下次来洛杉矶还会选择住在这里。

The host canceled this reservation 4 days before arrival. This is an automated posting.

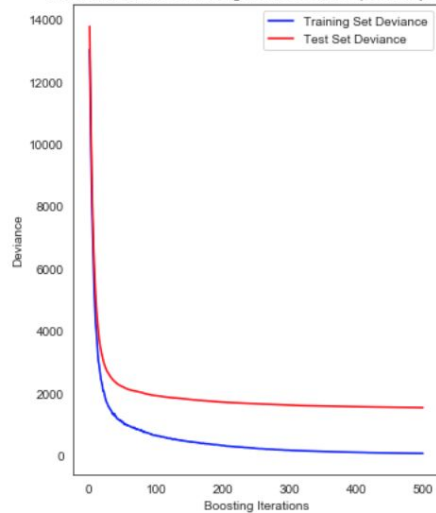
Hillary很热情，给予我们很大的帮助。地理位置非常好，靠近好莱坞影视城，房间非常干净、舒适、温馨，房间内的生活用品非常齐全。这是我们来美国住过民宿中的最好的一次。Hillary很热情，给予我们很大的帮助。地理位置非常好，靠近好莱坞影视城，房间非常干净、舒适、温馨，房间内的生活用品非常齐全。这是我们来美国住过民宿中的最好的一次。Hillary非常好，房子非常干净，周围很安全，去环球影视城很近。谢谢你

Ya es la segunda vez que elijo esta casa en West Hollywood, la elijo porque es muy linda, tiene todas las comodidades, todo funciona muy bien, tiene muchos cuartos y está en un barrio muy bonito y estratégico para moverse con comodidad en Los Angeles. Nina y Kevin son muy serviciales, están disponibles en todo momento y a cualquier hora para solucionar lo que se necesite y dar consejos para que la estadía sea perfecta. Son tan adorables y confiables que lo hacen sentir a uno como si fueran grandes amigos en esa ciudad. Gracias por todo! The host canceled this reservation the day before arrival. This is an automated posting., The host canceled this reservation 42 days before arrival. This is an automated posting.

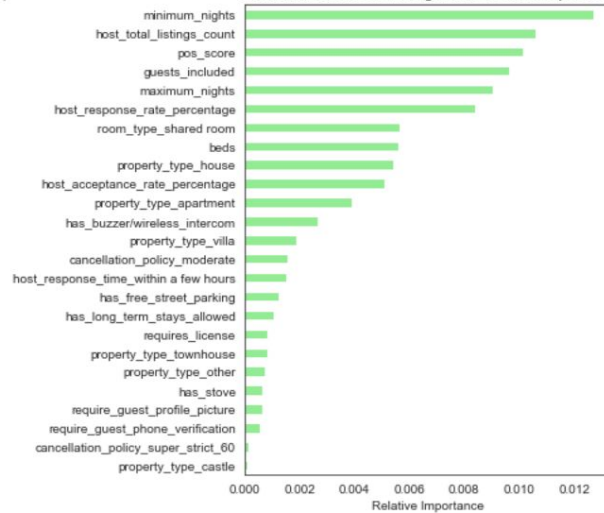
Cross Validation Root Mean Squared Error for
Gradient Boosting and XGBoost Models



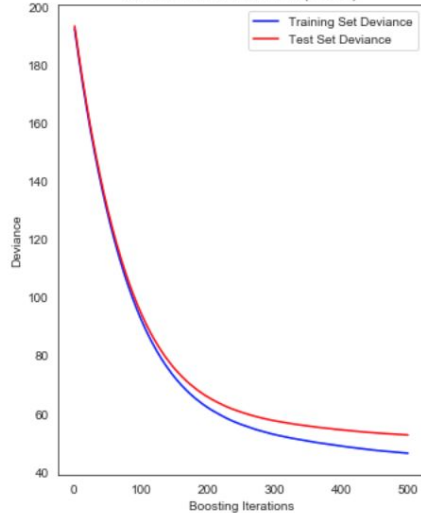
Stochastic Gradient Boosting Model Deviance (Least Squares)



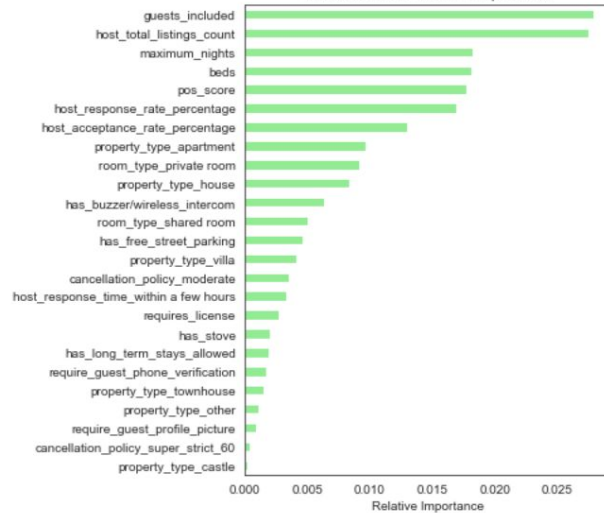
Stochastic Gradient Boosting Model Variable Importance



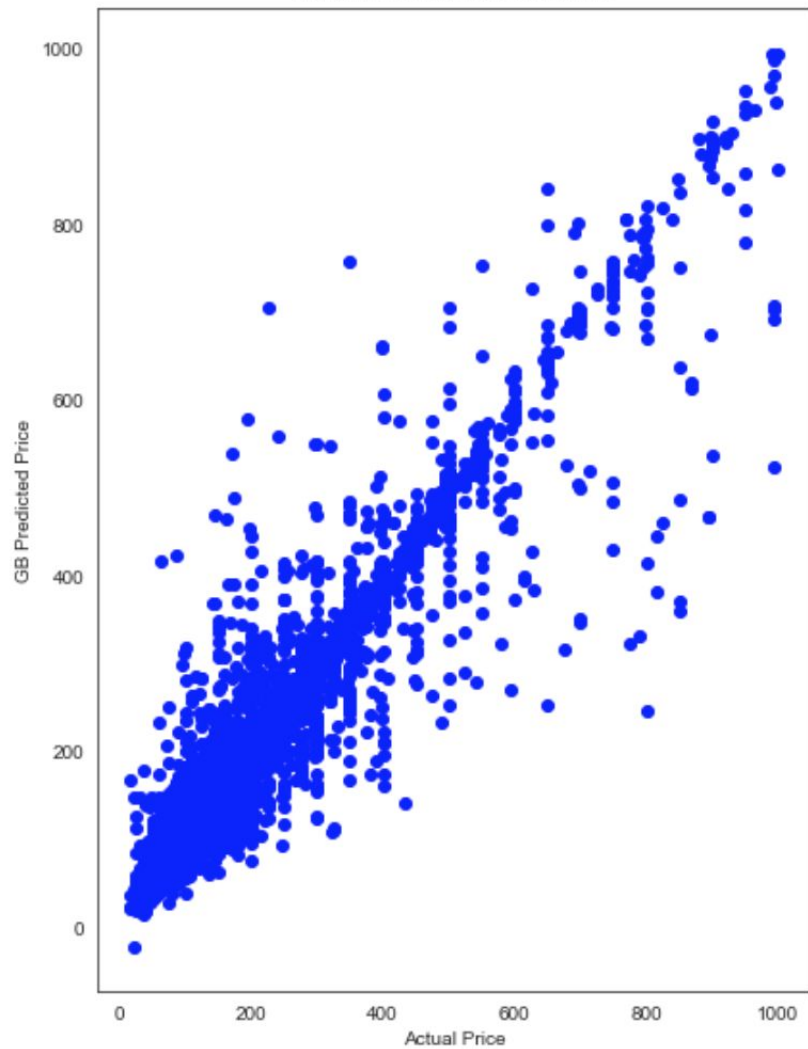
XGBoost Model Deviance (RMSE)



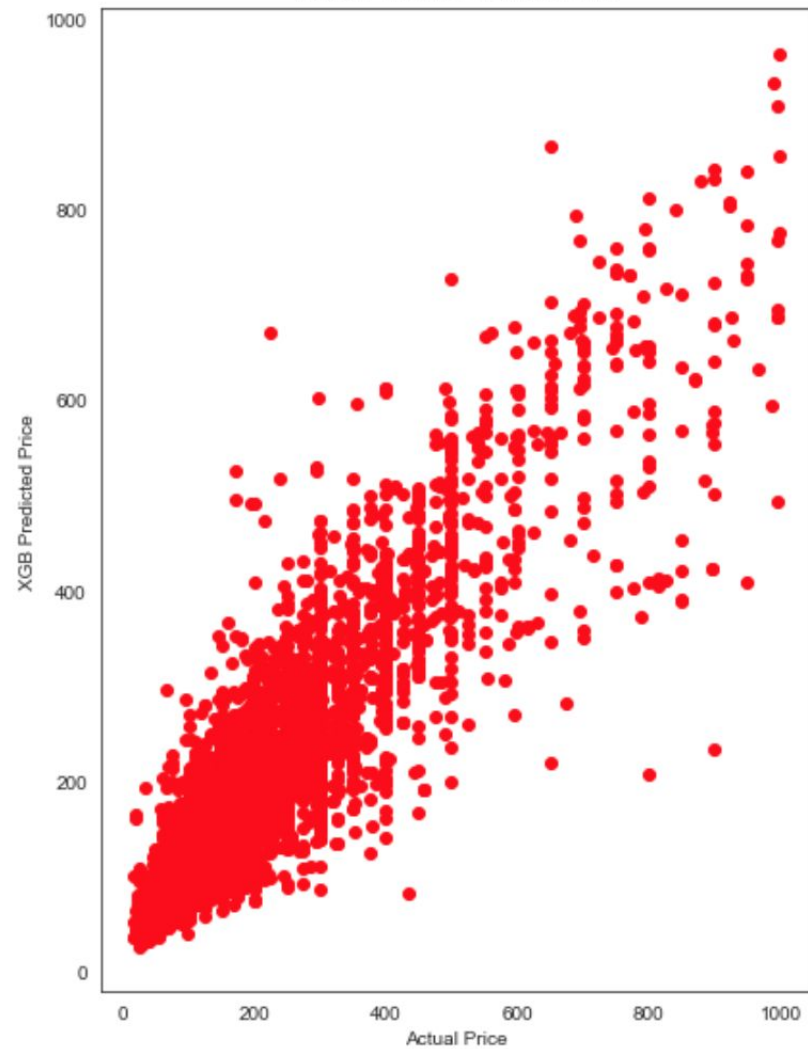
XGBoost Model Variable Importance



Actual Price vs. Predicted Price



Actual Price vs. Predicted Price

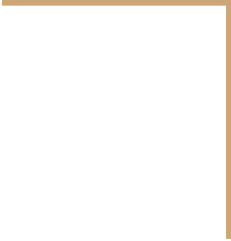


Stochastic Gradient Boosting Model:


61.4% price predictions are within 10% of actual prices.

XGBoost Model:

32.5% price predictions are within 10% of actual prices.



Conclusions & Next Steps



Conclusions

1. Provide an estimate for how much a host can charge for their listing.
2. Marketing tactics to approach potential hosts with targeted advertising
 - a. "Have an extra private bedroom in Manhattan? You can make X dollars per month by putting your room on Airbnb!"
3. Suggest improvements for hosts of existing listings to easily increase their value.
 - a. "Add a buzzer or wireless intercom to your listing and charge \$10 more per night"

Future Improvements

1. Make model production ready
 - a. Implement a pipeline in which the data could be read directly into the program, automatically cleaned and processed for feature creation, and run through the machine learning model
2. Sentiment analysis preprocessing
 - a. Remove non-English reviews from dataset
 - b. Add description of listing as another text-based sentiment feature
 - c. Use other NLP methods (word2vec or doc2vec)
3. Train models on entire LA dataset, generalize model to multiple cities
4. More extensive model tuning

Thank you!

Questions?