Project 1: Write a data science blog post

Introduction

Airbnb is a popular online marketplace that facilitates short-term home rentals around the world. In this project, we will analyze the Airbnb data for the Boston city to understand the key factors that influence the prices of Airbnb listings. *After this, we will present this blog follow the CRISP-DM process.*

Business Understanding

Brief description

The Airbnb dataset for Boston contain information on thousands of listings, including the location, property type, number of bedrooms and bathrooms, and amenities. The data also includes the prices of the listings, as well as reviews from previous guests. By analyzing this data, we hope to gain insights into the characteristics of successful Airbnb listings and identify strategies for hosts to optimize their listings and increase their revenue.

Business Question

1. Is there a trend in house prices?
2. which neighborhood has highest price?

3. What are factors that have most impact on a house’s price?

Data Understanding

Raw Airbnb Boston data contains three files:

- calendar.csv, which has availability information of listing by date from Sep 2016 to Sep 2017
- listings.csv, which includes detail information of each listing, such as the location, property type, amenities… and average review score. There are a total of 3,585 listings. Each listing has information about:
    - location(neighborhood, city, address);
    - amenities(TV, Wifi, number of bedrooms …);
    - responsiveness (host_acceptance_rate, host_response_rate, host_response_time, cancellation_policy …);
    - some text features like: name, summary, description, …
- reviews.csv, contains information about comments from previous customers in nature language form.

Prepare Data

Before we can analyze the Airbnb datasets, we need to clean the data to ensure that it is accurate and consistent. This will involve removing incomplete or irrelevant data, correcting errors, and standardizing the format of the data.

We drop some less information features.

['scrape_id','last_scraped','country', 'country_code', 'smart_location', 'street','thumbnail_url', 'medium_url', 'picture_url', 'xl_picture_url', 'listing_url', 'host_url', 'host_thumbnail_url', 'host_picture_url', 'country', 'country_code', 'smart_location', 'street', 'market', 'first_review', 'last_review', 'state', 'calendar_last_scraped', 'city', 'scrape_id', 'last_scraped', 'space', 'host_listings_count', 'zipcode', 'is_location_exact', 'host_location', 'host_total_listings_count']

We also drop columns with missing rate more than 50%. Then, fill other missing column with appropriate function.

The remain category features are transformed to one hot category by pandas.get_dummies function.

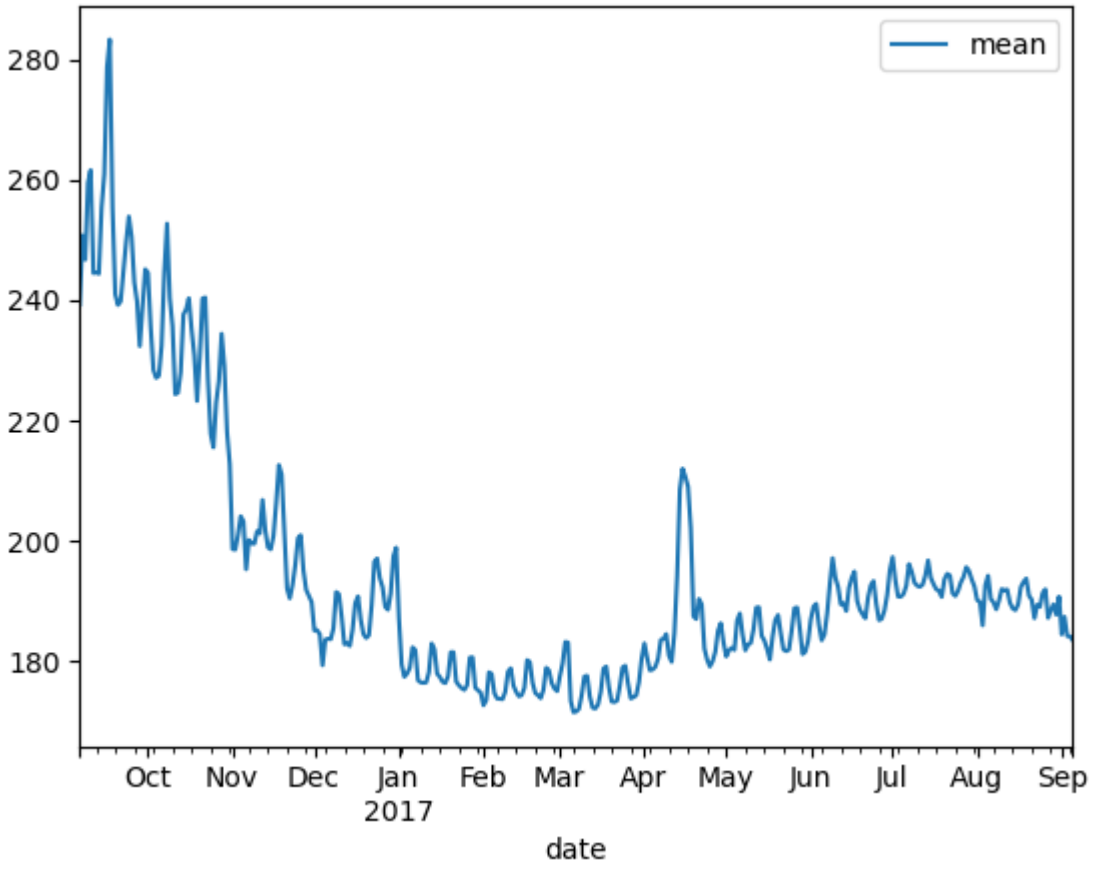
Exploratory Data Analysis

Once we have cleaned the data, we will perform exploratory data analysis to gain insights into the characteristics of the Airbnb listings in Boston. This will involve creating visualizations and statistical summaries of the data to identify patterns and trends.

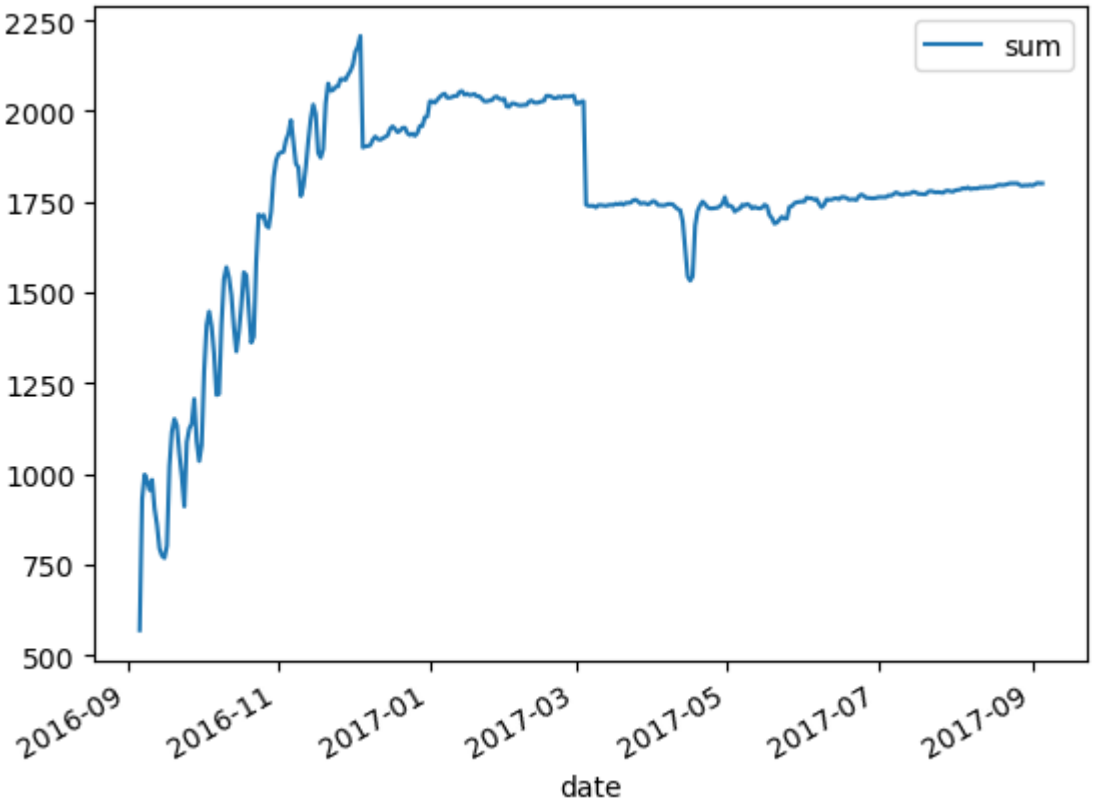
To answer the first two questions, we use EDA and visualize corresponding pattern in data to obtain some observation and conclusion.

Analyse temporal pattern to detect trend in listing price:

1. First, we plot the sum of price by date:

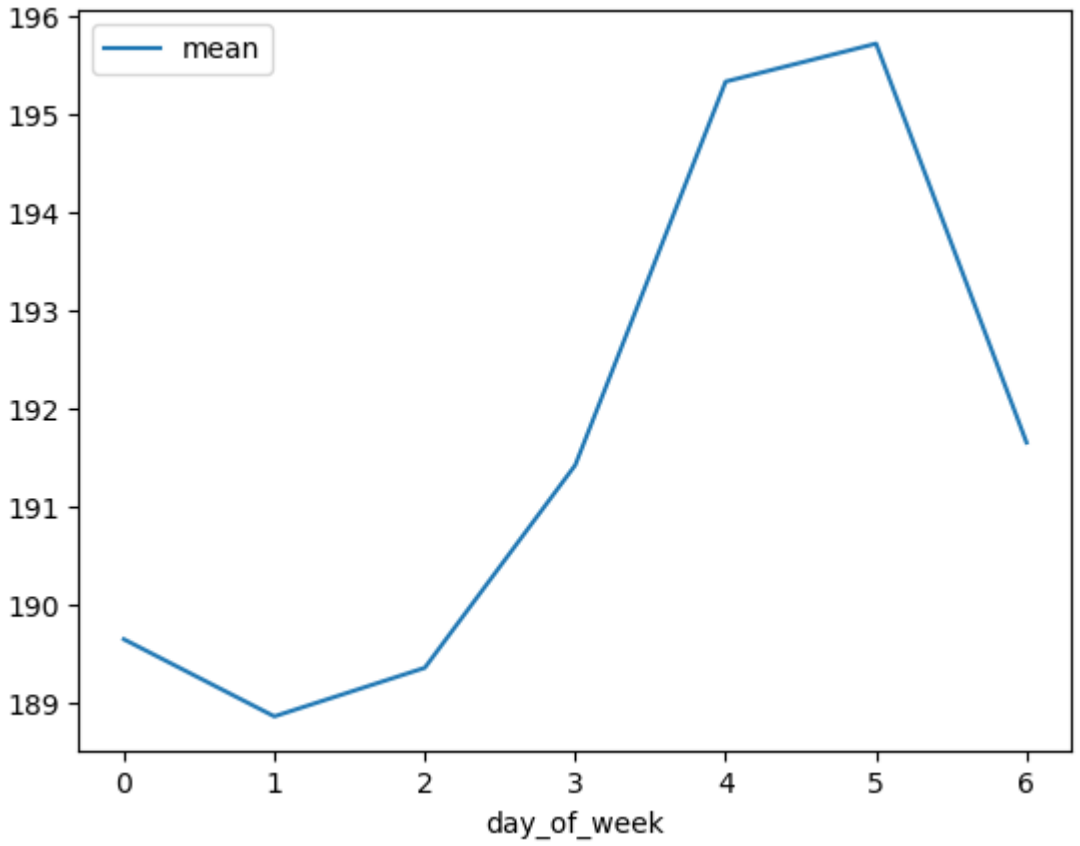


2. Second, we plot the available rate of total listings



- We can see a downtrend about listing price in the late of 2016, but the available rate is increase at the same time. From the average price graph and the above available rate graph, we have an assumption that there was an abnormal event occur in the last months of 2016 at Boston, then the accommodation supply is not enough and the listed price was pushed up.
- The average price in summer months(June, July, August of 2017) is higher than spring months (January, February, March of 2017). Available rate in summer is also lower than spring.
- There was a peak of price in mid of April 2017, this can help for future analyze.

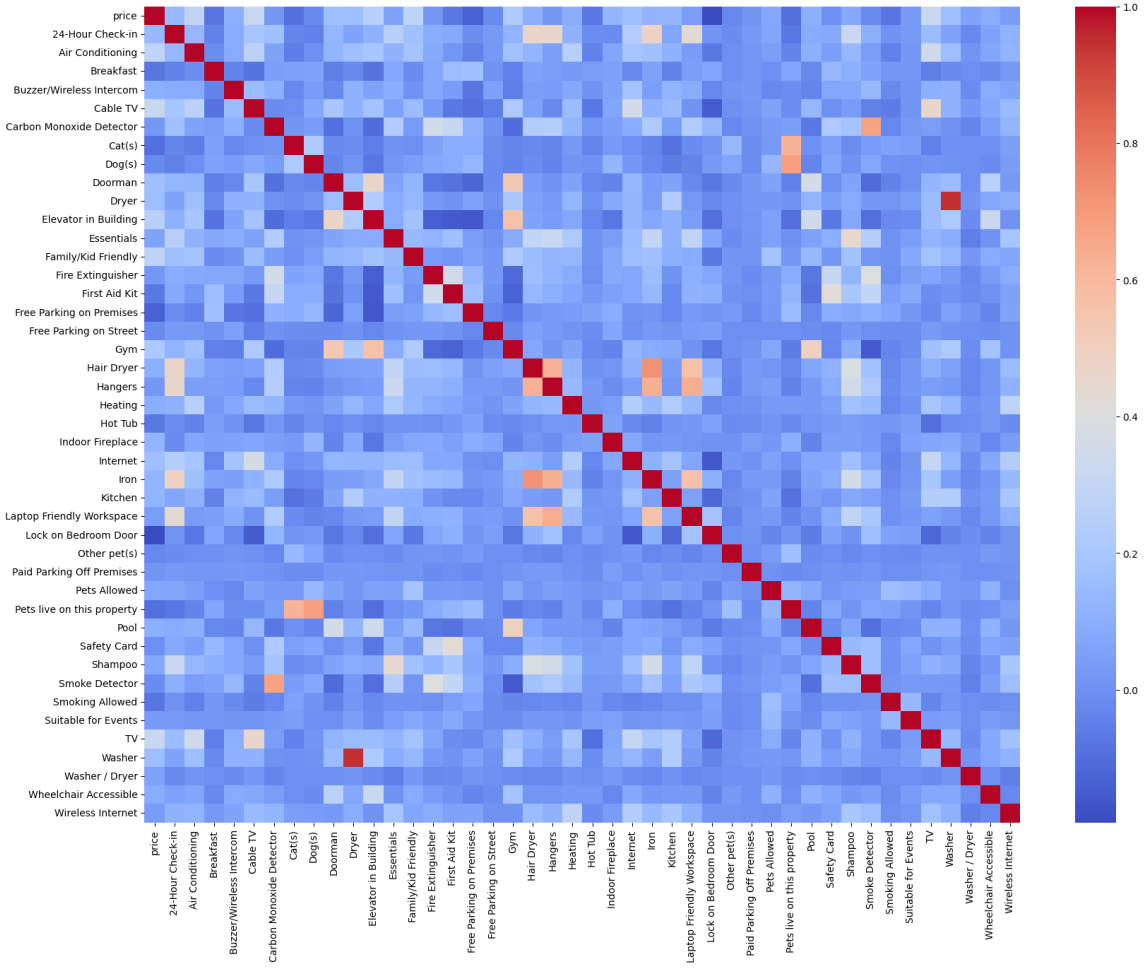
3. Next, we create plot of pricing by day_of_week



- The average price by day_of_week graph show that the listed price in highest at Friday and Saturday. The general trend is price increase from Tuesday and get highest at Saturday, then it decrease from Sunday to the next Tuesday.

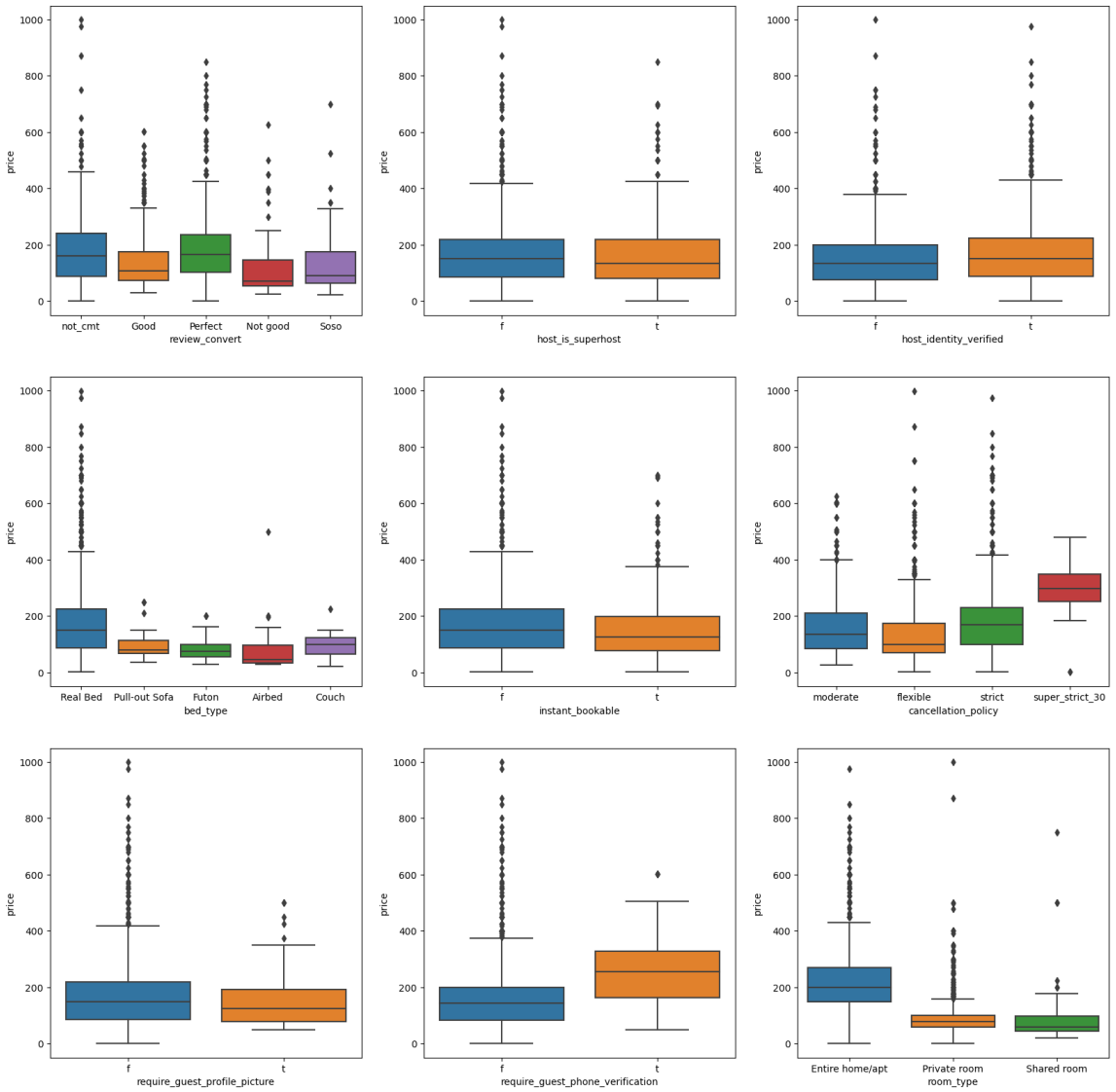
Analyse correlation between features and price to determine factors which most effect to price

1. Next, I will analyze amenities of all listings to find which is factor most influence to price.



- The amenities that appear the most are: Wifi (>95%), Heating (>94%), Kitchen (>91%).
- The amenities that have the highest correlation with price are TV, Air Conditioning, Elevator in Building, and Gym. Conversely, 'Lock on Bedroom Door' and 'Free Parking on Premises' correlate most negatively with the price, which mean that the room seem to be cheaper if it doesn't have a lock for bedroom and more expensive if it has free parking on premises.

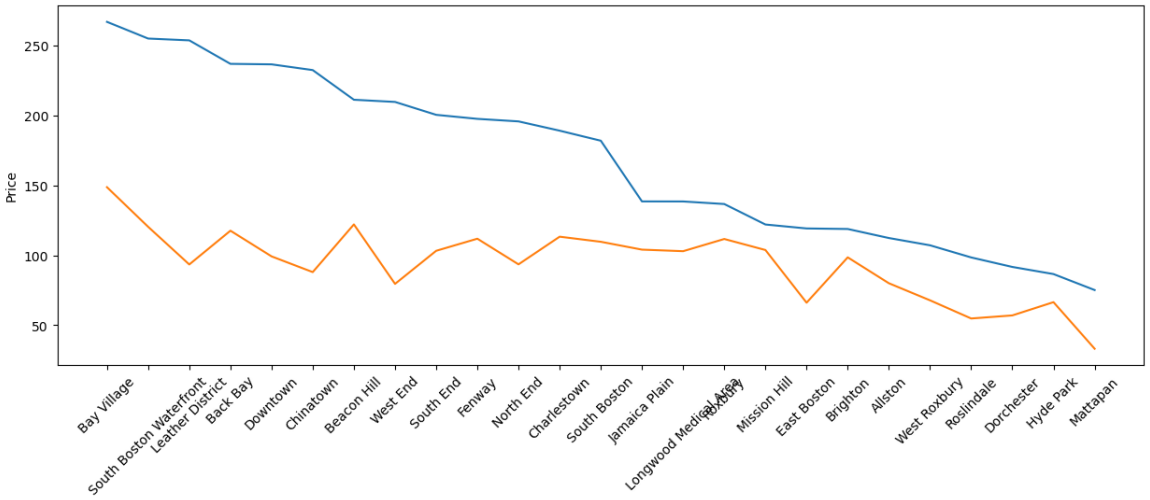
6. We analyze some category features by using box plot



- From above graph, we have some observes about categorical variables. According to room type, price of 'Entire home/apt' is much expensive than price of 'Private room' and 'Shared room'.
- cancellation_policy is seem to be more strict if the price is higher.
- The more expensive room have a slightly higher in host identity verification ratio and often require guest phone verification.
- Finally, the 'bed_type' may effect a lot to the price, room with a 'real bed' is often expensive than the without one.

Analyse Listing Price by Address

1. Finally, we plot the average Listing Price by Neighborhood



- There is a noticeable variation in room rates between neighborhood. The neighborhood has highest average price is Bay Village and the lowest is Mattapan.
  Higher-priced areas often have higher price volatility.

Modeling

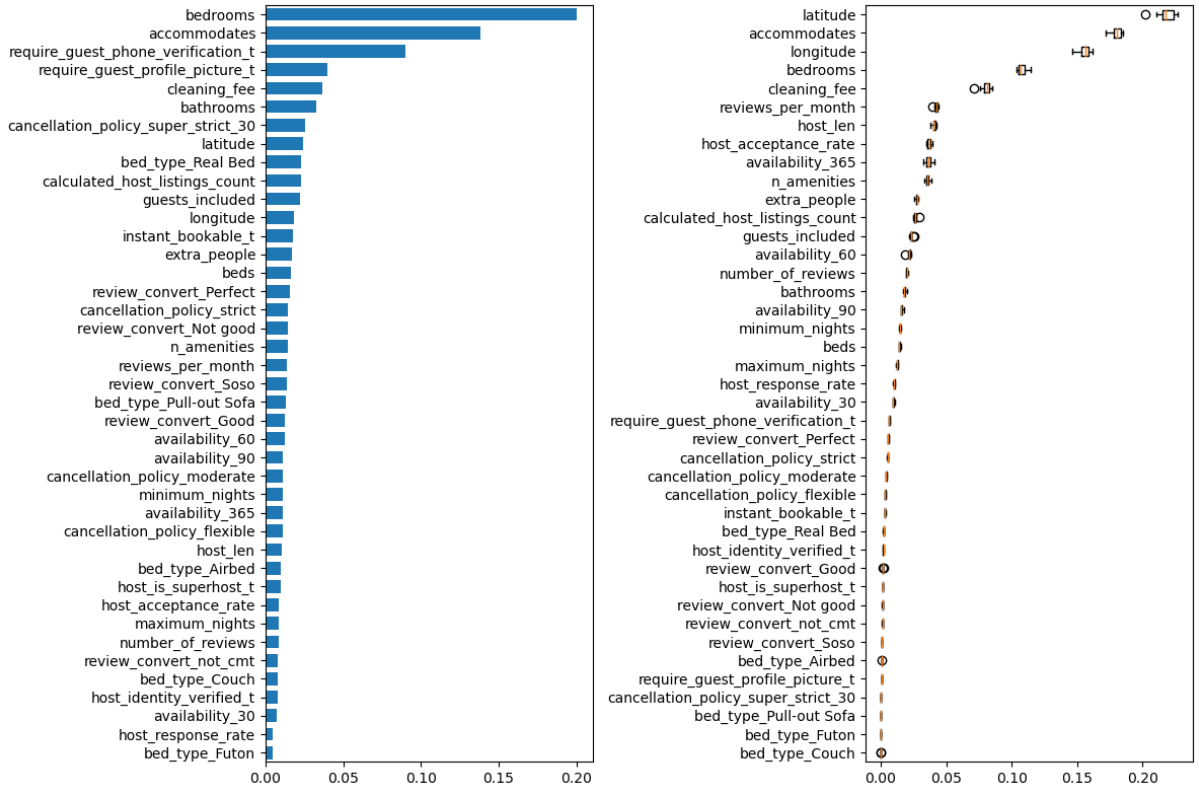
Build model

To identify the key factors that influence the prices of Airbnb listings, we also build predict models on data such as Linear Regression and XGBoost. The Average R2 score of these model through *cross validation* are shown below:

model xgb:
Mean score: **0.707** (0.025)
model Lasso:
Mean score: 949.654 (2970.267)
model LR:
Mean score: 1741.846 (4683.861)

Validate model, importance_features

Next, I use permutation_importance to show importance score of each feature in average!



According to above figure, we have some comments:

1. Longitude and Latitude are related to the geographic location of the Airbnb listings at Boston, so it is clear that this is most factor that impact to the price. Location is a crucial factor for travelers when choosing accommodations. The proximity of a listing to popular attractions, transportation hubs, or specific neighborhoods can greatly impact its desirability and, consequently, its price. For example, listings at Bay Vilagge may command a higher price. In addition, listings that close to public transportation, restaurants, or shopping areas may be more expensive.
2. the Accommodate has a direct correlation to listing price. This is easy to explain because apartments with more capacity will cost more. The number of bedrooms and bedroom types are also similar.
3. Some features related to the listing's feedback and review score, such as reviews_per_month, host_acceptance_rate, are also worth noting. Positive feedback and high approval rate will increase the reputation of the listing and directly affect the price.
4. Feature importance analysis are quite consist with correlation analysis

Answer the Business Questions

1. Is there a trend in house prices?
    - Yes, Listing prices depend on time factors and the law of supply and demand. It changes seasonally or is affected by non-annual events.
2. which neighborhood has highest price?
    - Three highest are: Bay Village, South Boston Waterfront and Leather District
    - If you want to rent cheaper home in Boston, you may looking to Mattapan, Hyde Park and Dorchester.
3. What are factors that have most impact on a house's price?
    - The most affect factor of a listing is its address, then the accommodation and followed by several factor: cleaning_fee, reviews_per_month, host_len, host_acceptance_rate.
    - The importance conclusion is online reviews will affect the value of the listing and in turn will affect the price.

Conclusion

Based on our analysis of the Airbnb data for Boston, we can draw several conclusions about the key factors that influence the prices of Airbnb listings. We find that location and reviews type are the most important factors, as well as the accommodation, number_of_bedroom and amenities. Our findings can help hosts optimize their listings and pricing strategies to increase bookings and revenue.

Future plan

1. Answer the question which factor is importance for home rental investor in Boston?
2. Using NLP technique to analyse reviews data and text listings data