Airbnb Price Prediction

Motivation: I am a huge fan of travelling across the globe especially the places that I like to visit are the ones that hold historical significance. Also, I believe that on a personal level, traveling helps us grow personally because it exposes us to diverse cultures, and diverse mindsets. But before traveling to any place, I love to explore more in-depth about that place and create an itinerary that helps me plan and budget my stays at Airbnb.

What is Airbnb:

Before diving deep into the project, it is important to know what Airbnb is.

Airbnb is an online marketplace to connect people looking for a place to stay with those willing to rent out their apartments. The listings on Airbnb provide a wide variety of choices, from shared beds to luxurious residences, all on one platform. After making reservations, customers may submit recommendations through a peer-review system. The organization gathers a ton of data, from reviews by users to listing facilities. We are evaluating these listings using Spark on readily available Airbnb listing data on the internet to forecast prices for a listing.

Data Source: This dataset was taken from http://insideairbnb.com/get-the-data/ and consisted of data for roughly 80,000 Airbnb listings across the United States.

Tools: Python programming. I used the PySpark library to build ETL pipelines that can handle the data extraction part. I used pandas, and matplotlib for exploratory analysis and data visualization.

Project:

This section is divided into steps.

1. **Data Extraction and Loading:**

   The SparkSession and SparkContext objects are used to initiate the Spark session for ETL

operations and Spark. Spark context serves as the entry point to connect with Spark environment. SparkConf is used to connect with available spark clusters.

The famous line **SparkSession.builder.appName("Airbnb-Prediction").getOrCreate()**

- **Builder creates a new Spark session the main interface for working with Spark.**
- **appName() sets the name of the Spark application**
- **getOrCreate() creates a Spark Session if does not exist, else retrieves the new object.**
- **Dataset:**

```
In [ ]: spark=SparkSession.builder.appName("Airbnb-Prediction").getOrCreate()

In [ ]: df=pd.read_csv("/content/sample_data/train.csv")

In [ ]: pd.set_option('display.max_columns',100)
        df.head(10)
```

Out[ ]:

	id	log_price	property_type	room_type	amenities	accommodates	bathrooms	bed_type	ca
0	6901257	5.010635	Apartment	Entire home/apt	{"Wireless Internet","Air conditioning",Kitche...	3	1.0	Real Bed	
1	6304928	5.129899	Apartment	Entire home/apt	{"Wireless Internet","Air conditioning",Kitche...	7	1.0	Real Bed	
2	7919400	4.976734	Apartment	Entire home/apt	{TV,"Cable TV","Wireless Internet","Air condit...	5	1.0	Real Bed	
3	13418779	6.620073	House	Entire home/apt	{TV,"Cable TV",Internet,"Wireless Internet",Ki...	4	1.0	Real Bed	

- The dataset consists details about:
  - Price
  - Amenities
  - Accommodation Capacity
  - Bathrooms/Bedrooms
  - Property Type
  - City, State, Address
  - Cancellation/Cleaning Fees
  - Reviews, Images

2. Data Cleaning/Transformation:

Dealing with Null Values:

- Given the diverse datatype of these columns, I dealt with replacing the null values according to the values

Columns	**Datatype**	**Type of Values replacing nulls**
host_has_profile_pic,host_host_identity_verified,instant_bookable	Object	True, False
bedrooms,beds,review_scores_rating,host_response_rate	int64,float64	Mean, Median Values
first_review,last_review	datetime	Difference of mean() and least value

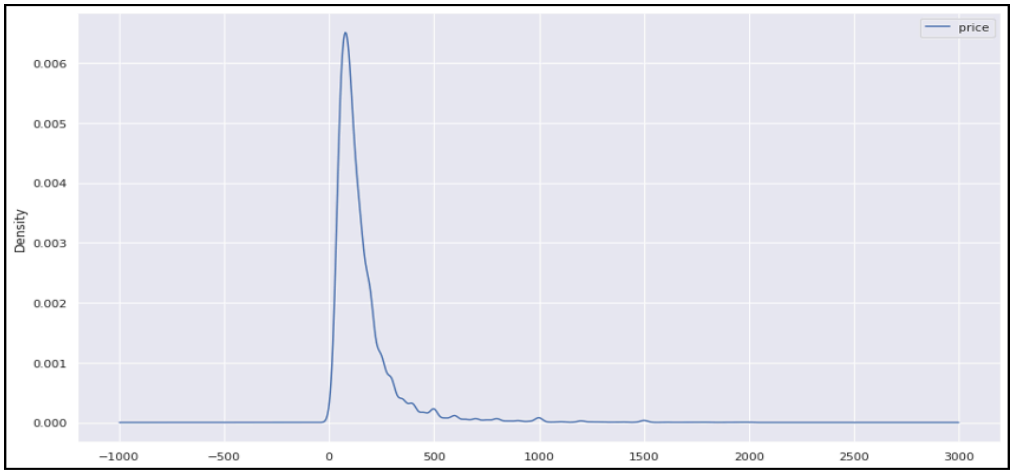
Transformation Values: Columns like Amenities consisted of values that have values not in the intended format for analysis.
To achieve these transformations and data cleaning,

3. **Data Analysis:**
   In this step, we deeply explore the dataset to analyze the different Airbnb rentals and examine the relation between the prices of Airbnb rentals and the different features of Airbnb rentals.
   These factors include:
   Location, property types, prices, temporary fees, review scores, amenities, number of Rooms, first and last review dates.

Correlation Matrix:

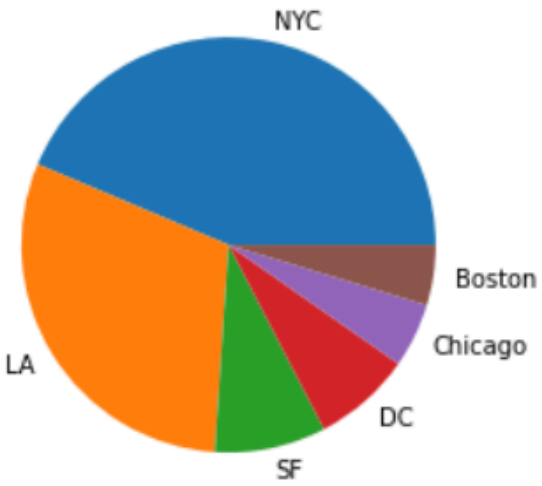
	log_price	accommodates	bathrooms	cleaning_fee	host_has_profile_pic	host_identity_verified	latitude	longitude	number_of_reviews	review_scores_rating	zipcode	bedrooms	beds	first_review_days
log_price	1.00	0.57	0.35	0.11	-0.01	0.02	-0.00	-0.05	-0.03	0.07	0.03	0.47	0.44	-0.05
accommodates	0.57	1.00	0.49	0.18	0.01	0.06	-0.08	-0.09	0.04	-0.01	0.10	0.71	0.81	-0.01
bathrooms	0.35	0.49	1.00	0.06	0.01	0.01	-0.13	-0.13	-0.04	0.01	0.13	0.57	0.52	0.02
cleaning_fee	0.11	0.18	0.06	1.00	0.02	0.16	-0.06	-0.07	0.11	0.03	0.07	0.11	0.13	-0.07
host_has_profile_pic	-0.01	0.01	0.01	0.02	1.00	0.10	-0.02	-0.02	0.02	0.00	0.03	0.01	0.01	-0.01
host_identity_verified	0.02	0.06	0.01	0.16	0.10	1.00	-0.05	-0.06	0.16	0.05	0.07	0.03	0.04	-0.19
latitude	-0.00	-0.08	-0.13	-0.06	-0.02	-0.05	1.00	0.90	-0.02	-0.03	-0.87	-0.06	-0.08	0.05
longitude	-0.05	0.09	-0.13	0.07	0.02	0.06	0.90	1.00	-0.05	0.04	-0.99	-0.08	0.08	0.06
number_of_reviews	-0.03	0.04	-0.04	0.11	0.02	0.16	-0.02	-0.05	1.00	0.01	0.05	-0.04	0.03	-0.49
review_scores_rating	0.07	-0.01	0.01	0.03	0.00	0.05	-0.03	-0.04	0.01	1.00	0.05	0.01	-0.02	0.03
zipcode	0.03	0.10	0.13	0.07	0.03	0.07	-0.87	-0.99	0.05	0.05	1.00	0.08	0.09	-0.06
bedrooms	0.47	0.71	0.57	0.11	0.01	0.03	-0.06	-0.08	-0.04	0.01	0.08	1.00	0.71	-0.01
beds	0.44	0.81	0.52	0.13	0.01	0.04	-0.08	-0.08	0.03	-0.02	0.09	0.71	1.00	-0.00
first_review_days	-0.05	-0.01	0.02	-0.07	-0.01	-0.19	0.05	0.06	-0.49	0.03	-0.06	-0.01	-0.00	1.00
last_review_days	-0.01	0.04	-0.02	0.07	0.00	-0.05	0.16	0.13	0.21	0.06	-0.14	-0.01	0.04	0.27
host_since_days	-0.08	0.01	0.02	-0.09	-0.03	-0.33	-0.01	0.01	-0.20	-0.04	-0.01	-0.01	0.02	0.49
Price	0.84	0.52	0.44	0.03	-0.01	-0.01	-0.03	-0.06	-0.07	0.05	0.05	0.49	0.43	-0.00

Price Distribution:
More than 90% of the listing prices for a day are between $0 and $500, according to the density distribution. The highest peaks were found to be around $100. estimating the cost of reserving a property for one day in e in each of the five cities under consideration.

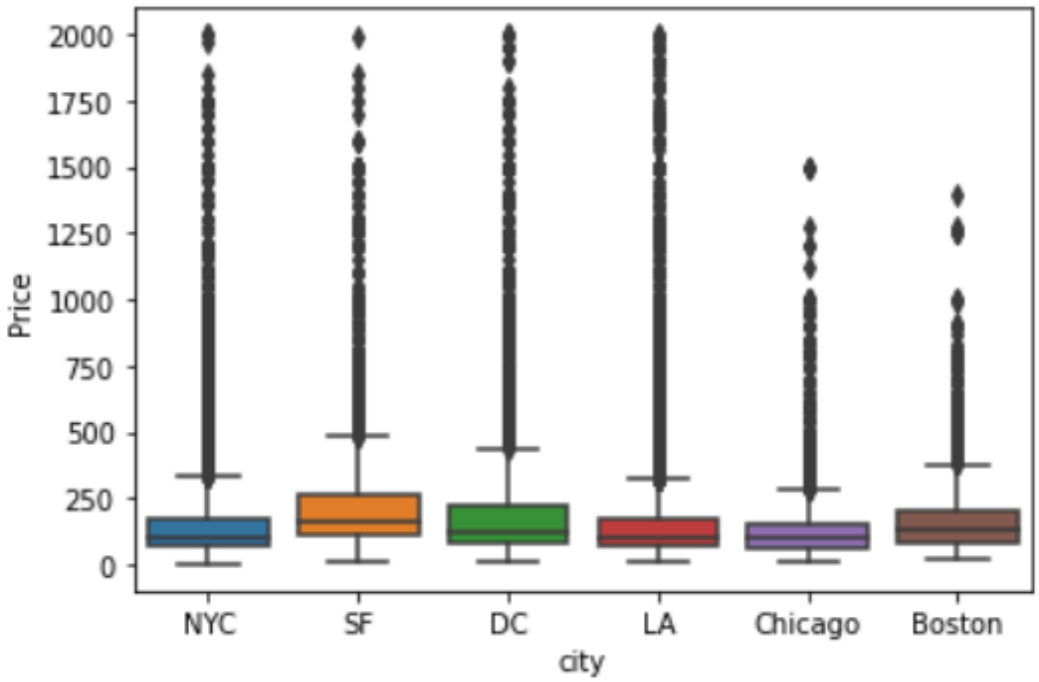


Top 10 amenities:

Top 6 cities that brought the most revenue:
Cities like NYC, LA, DC, and SF brought the majority of profits for Airbnb rentals followed by Chicago and Boston.

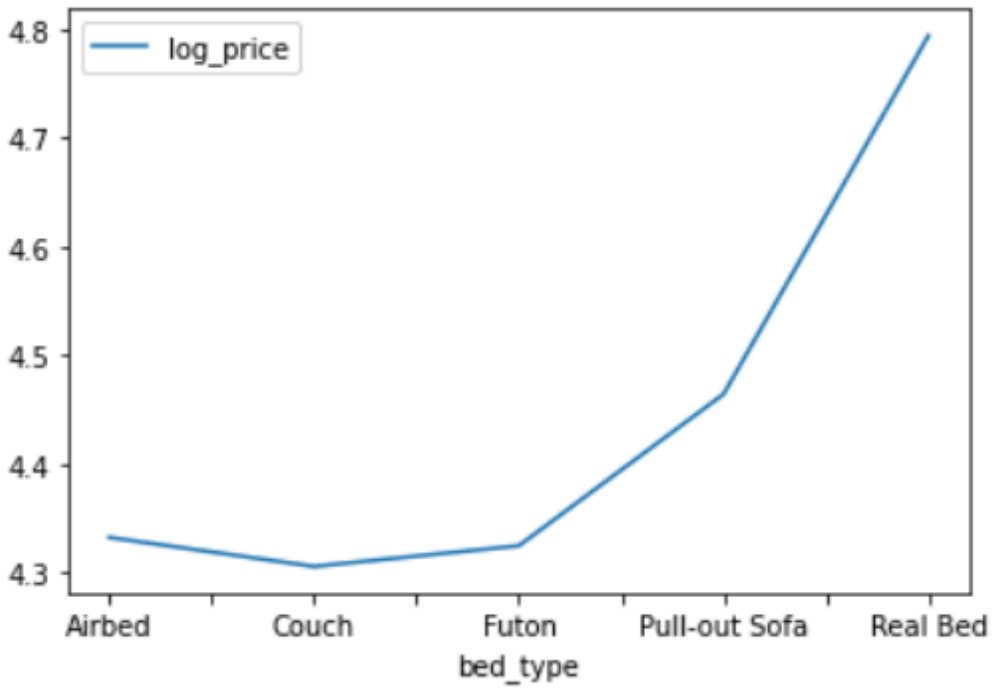


This graph gives an idea about the housing market in 6 cities across the USA and also tells us about the affordability of having Airbnb rentals in these cities.



For the prices of Airbnb listings across the 6 cities, NYC has the most price variation whereas San Francisco has the highest median listing price followed by DC, LA, and NYC. Chicago had the lowest median price. This suggests that SF has the most expensive and also the most profitable market for Airbnb and that Airbnb should also focus on improving the markets in other cities.

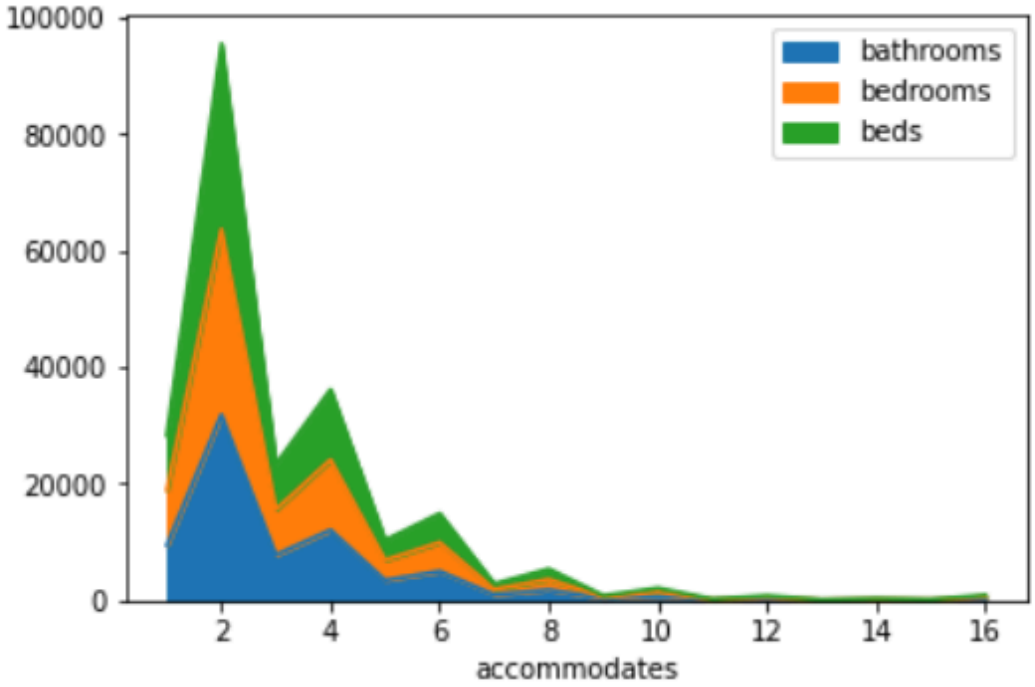
Maybe this can be achieved by providing more cheap and affordable Airbnb in Chicago, and Boston areas through campaigns that motivate more people to visit these cities and make more use of Airbnb listings. They can partner with the local government and tourism boards of Chicago, and Boston or event planners that can promote event related purchase offers with travelers who opt-in to stay at Airbnb listings. Also, they can partner up with more listing owners and extend their support to have more affordable Airbnb accommodations for travelers.



Bedroom Types and Price Type and Accomodations

We represent price in the form of log values while comparing prices gainst both the accommodations and bed types.

We can see that rentals with Airbeds have the lowest price whereas the Real Beds and Pull-out sofas have the highest price. This indicates that as the bed types with much more handy furniture or the ones not fully-furnished are the cheapest whereas the ones wth mattresses have the highest price.



From this graph, the accommodations that accommodate 2 people with all the amenities like bathrooms, bedrooms, and beds had the highest price, and the ones with accommodations up to 16 people had the lowest price but also provided fewer bedrooms and bathrooms.
Recommendations: Airbnb should focus on adding more alternate Airbnb rentals to attract more travelers and also adding more basic amenities like bedrooms, and bathrooms to improve customer experience.

4. **Model Building and Feature Engineering:**

From exploratory analysis, we discovered that there are certain variables that affect the price of an AirBnb property listing. To exactly know which variables had a significant effect, we plotted the correlation matrix. From the results of the correlation of amenities with respect to each other and **log_price**, we selected amenities that had a significant effect on **log_price.** Therefore, we manually selected amenities that had a correlation value greater than 0.1 with respect to **log_price**.

Selected Amenities		
Kitchen	**Iron**	**Indoor_fireplace**
Heating	**Familykid_friendly**	**Gym**
TV	**Cable_TV**	**Private_entrance**
Hair_dryer	**24-hour_check-in**	**Doorman**
Washer	**Elevator**	**Suitable_for_events**
Dryer	**Pets_live_on_this_property**	

Model Selection:
The following are the regression models that we used for our analysis: Linear Regression, Decision Tree, Random Forest, Polynomial Regression with different features selected. We also used cross-validation on Random Forest model.
The best model gave a score of 0.60 for decision trees.

```
In [ ]: decision_model.score(X_test,Y_test)

Out[ ]: 0.600243050106555
```