How can AirBnb improve operations?

Cursory Problem Statement:

Underlying economic, business, and socio-cultural frameworks within the lodging industry remain transient and dynamic, and Airbnb managers must remain updated towards consumer trends affecting pricing strategy. Moreover, sophisticated modeling of components driving consumer tastes allows both Airbnb corporate strategists and host managers to productively invest in productive features or divest irrelevant costs to optimize cost structures, and ultimately, increase net profits.

Thus, Airbnb has accumulated data towards helping corporate strategy fine-tune operations, and similarly, assist hosts towards maximizing lodging attractiveness while maintaining competitive pricing. Specifically, Airbnb strategists can gain clarity on strategic weaknesses within business operations, while concurrently, Airbnb hosts can discern critical room, city-specific, geographical, and socio-economic elements most valuable to consumers as reflected via pricing strategy.

More Context:

Airbnb is an online platform that allows consumers to efficiently scour, locate, and plan lodging matching their preferences during planned stay lengths. Thus, it's important for Airbnb to host lodging with characteristics that not only remain competitive with consumer trends but perhaps, pioneer new consumer paradigms.

Airbnb's powerful network effect, with a semi-monopoly within this online lodging space, commands a majority of consumer attention and provides Airbnb with tremendous competitive leverage against other consumer choices. Commensurate with pricing leverage is the online platform's ability to collect data analytics on not only consumer trends but also on spotting nascent paradigms in the constantly evolving lodging space - i.e. a small percentage of Airbnb hosts have increased investments in their own lodging to progress beyond simple rooms and into tourist attractions themselves.

Thus, analyzing the breadth of data available to Airbnb allows both Airbnb's corporate strategists and room hosts to identify current key room features in the lodging and hospitality industry, while tailoring each room to relevant consumer types, so that Airbnb can remain as one of the top online destinations and resources for temporary stays. Concurrently, this will also allow room hosts to maximize profits. Increasing these streams of business competitiveness will ultimately support Airbnb's growth in its businesses and augment the bottom line.

Data sources:
These datasets stem from the Kaggle database (https://www.kaggle.com) as well as Wikipedia, (https://en.wikipedia.org/wiki/List_of_EU_metropolitan_areas_by_GDP)

Kaggle is an online data science platform that hosts competitions and practice for data science practitioners. Although the types of data are typically limited via this source, the data itself gleaned are usually cleaned and ready for analysis.

Wikipedia is an online free encyclopedia, with readily available tables that can easily convert to CSV for the data science process.

Airbnb room features contained in the dataframe:

Price

Room Type (Private/Shared/Entire Home/Apt)

Host is Superhost?

Multiple Rooms or Not?

Biz (Business Indicator)

Guest Satisfaction

Number of Bedrooms

Distance from City Center

Person Capacity

Cleanliness Rating

Attraction Index (proximity to valuable tourist attractions)

Restaurant Index (proximity to valuable restaurants)

GDP of City

GDP per capita of City

Population of City

Shared Room?

Weekday/Weekend Rental

Data Wrangling Process:

A concise explanation of the processes necessary to manipulate the dataframes for analysis is explicated below.

Problem 1 - Aggregating all respective cities datasets into one coherent dataframe:

Sourcing Kaggle data entails extracting data for each respective city into independent city dataframes that are divided into weekday vs weekend partitions. After individual city dataframes were labeled (city, weekend/weekday), they were then concatenated coherently into one aggregate dataframe comprising all initially separated datasets.

Problem 2 - Supplementary geographic and demographic csvs sourced and checked for validity:

A new csv containing European city GDPs was pulled into pandas. Subsequently, the new dataset was checked to verify that all European Airbnb cities contained in the initial dataframe are contained with the new dataset.

Problem 3 - Merging new geographic and demographic features onto the central dataset:

An additional city GDP column was fashioned and linked to every Airbnb city using the merge function on the 'city' column. Analogously, the population and Area of each respective Airbnb city was created and merged into the central dataframe.

Problem 4 - Creation of new column - Population Density:

Creation of a new column was incorporated into main dataframe by dividing new 'population' column by new 'area' column.

5. Noteworthy: Minimal scouring for null or duplicate values, or imputation towards missing values, was necessary because Kaggle offered an ideal and clean dataset for analysis.

Exploratory Data Analysis: Concise Summary of Major Points

Overall summary: EDA actually became the most important stage for analysis due to limitations of the eventual ML model.

1. Investigation of each feature, independently:

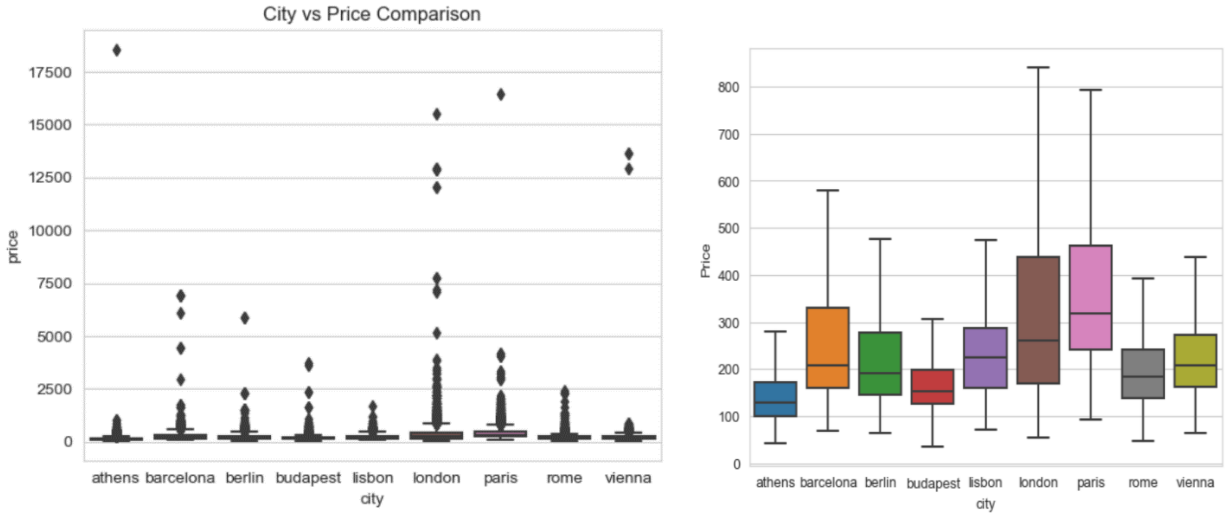
A histogram of every feature was examined to understand data distributions for each category. Unfortunately, for most categories, most data points clustered around specific segments, which is not ideal for regression analysis

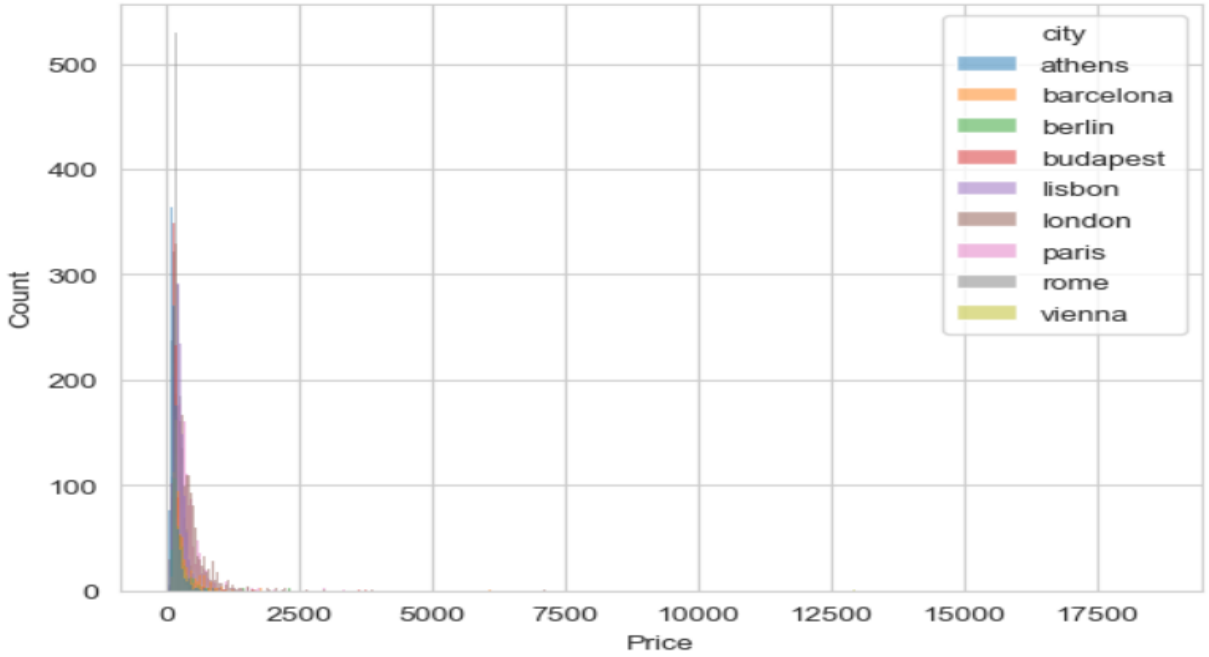
Distribution of prices – aggregating all cities together - was then plotted for visualization of patterns; it's easy to discern the mean of ~200 USD. Distributions and IQRs for each respective city were plotted; each cities' mean and dispersion varied widely signaling that each city possesses individual pricing dynamics.

2. Investigating relationships between feature/s vs price:

Every feature in this dataset was plotted against price and slotted into cities to visualize their respective relationships more conspicuously. More granular details regarding each chart is explicated below:

- Each city was plotted vs price to understand the price dynamics of each city:
- The left chart includes all data points:
- The right chart cuts out all the outliers based on IQR:
- Bottom chart shows histogram of aggregate cities' prices:

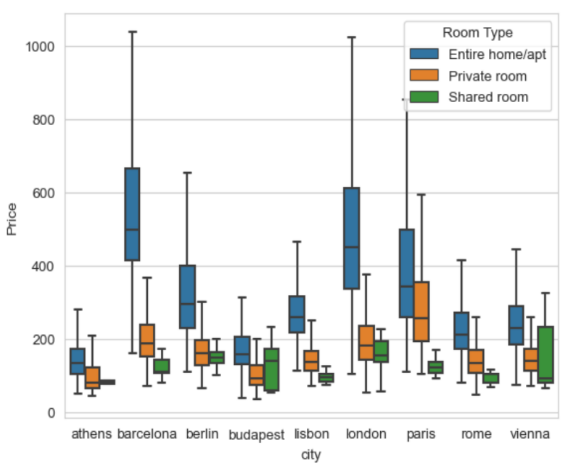


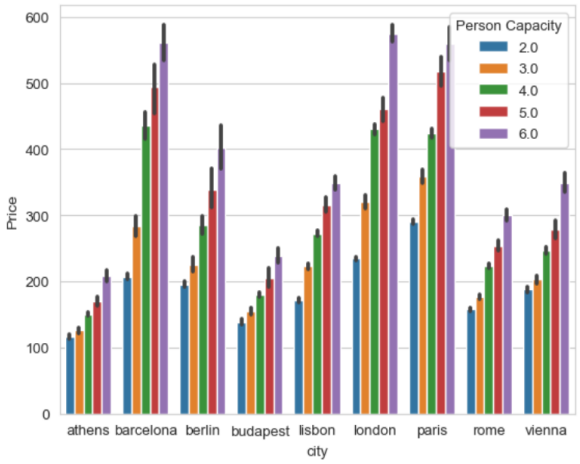


Two main patterns emerge:

1. Each city follows their own price dynamics
2. Heavy emphasis on premium segment in Airbnb price dynamics

The below chart shows the respective mean room rates by room type:

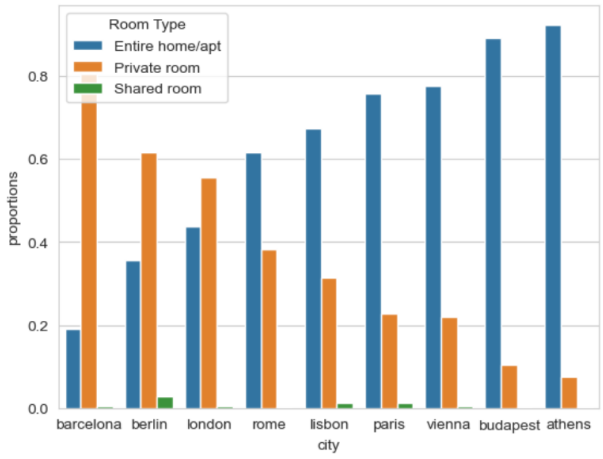


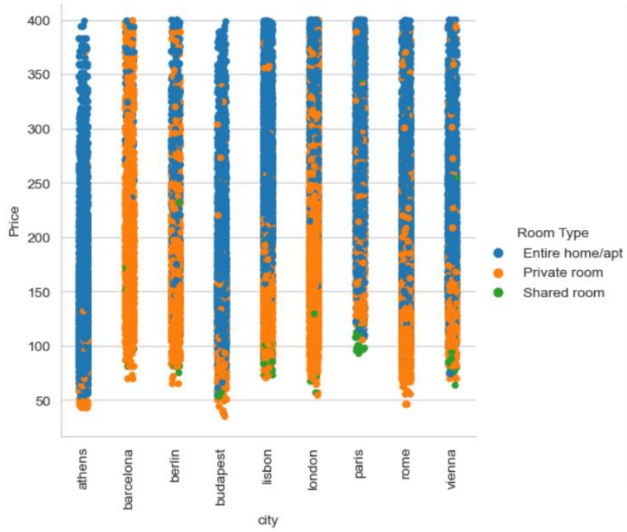


As expected, you can easily observe the strong positive correlation between larger room / person capacity vs higher price

- However, unexpectedly Barcelona possesses the largest mean home/apt price. Why? This anomaly is investigated in more depth later

Simultaneously, in the below charts, you can easily view that Barcelona – while possessing the highest mean prices for the 'Entire home/Apt' category also exhibits the lowest percentage of units within this category.

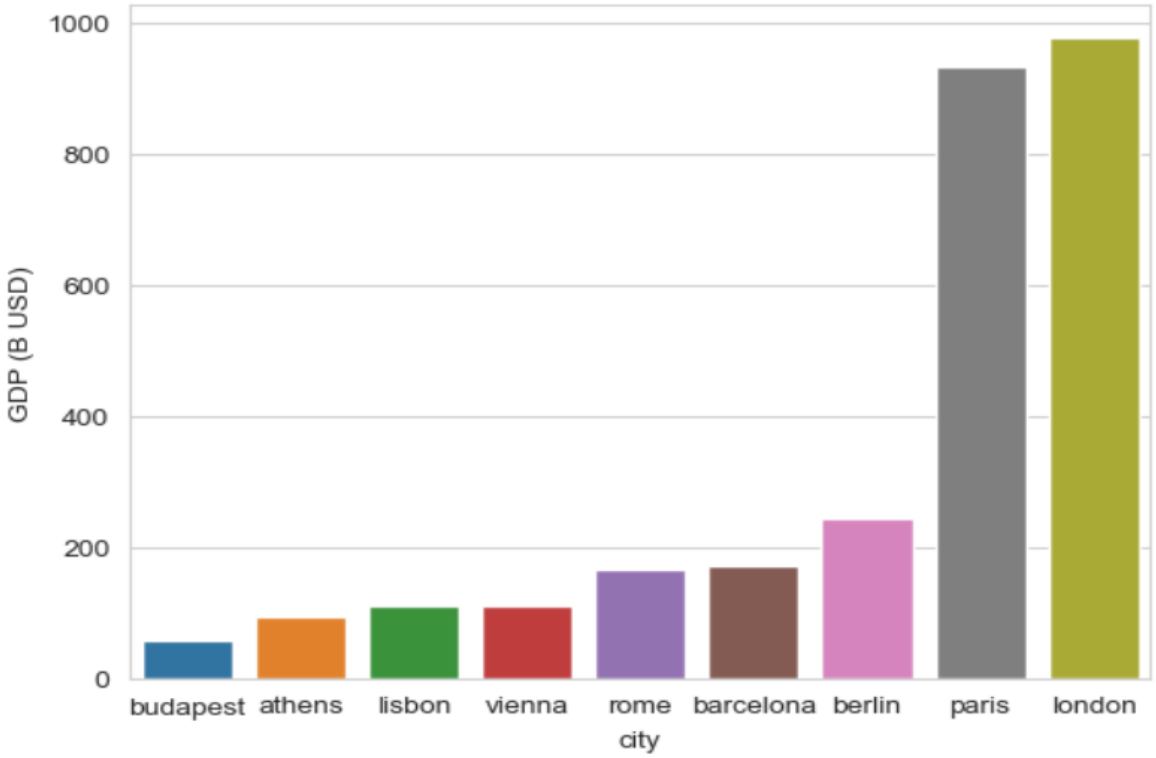




As discernable in the above graphs, this simultaneous balance of high price (high demand) vs low proportion of units in that Entire Room/Apt category (low supply) is unusual in comparison with other

cities. Furthermore, this low supply/high demand suggests a supply-demand imbalance, and is a peculiarity that Airbnb should investigate heavily into, for perhaps it suggests Airbnb can house a plethora of more 'Entire rooms/Apts' in Barcelona.

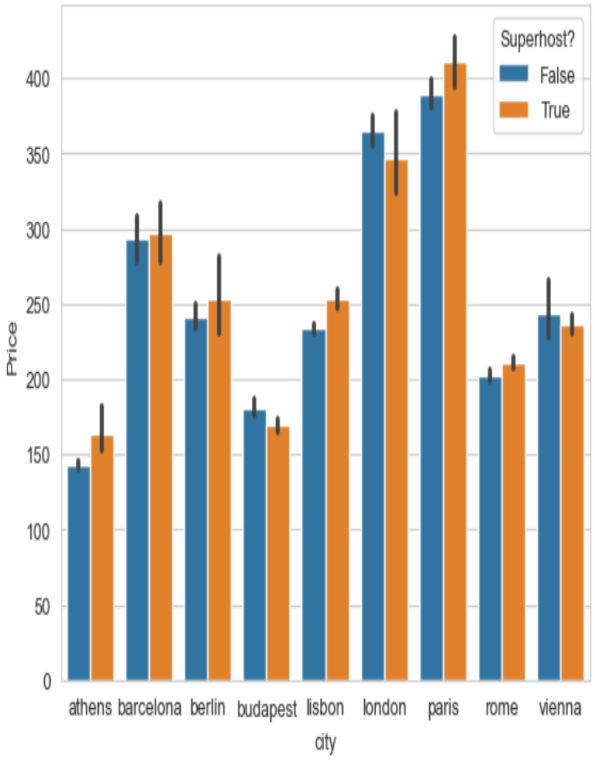
While viewing these respective cities' charts, you may have pondered if city GDP may predispose Barcelona (and other cities) towards these unique characteristics.

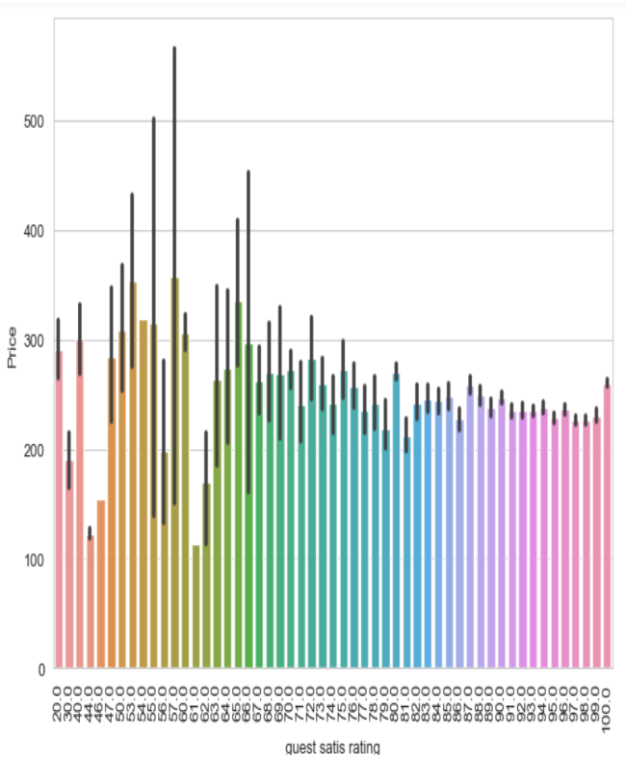


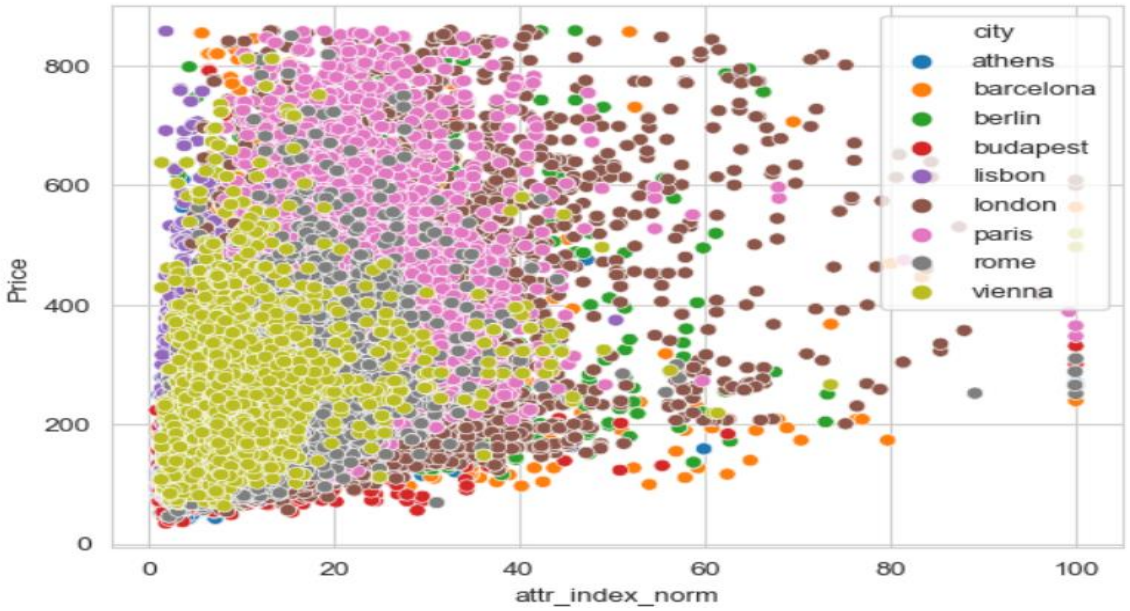
You can view through the above chart that Barcelona's GDP is in line with many of the other cities, and thus, GDP itself is not an underlying characteristic that influences Barcelona towards outlier supply-demand patterns.

These findings should prompt Airbnb to investigate further the ecosystem of variables behind Barcelona's unique supply-demand traits and unearth the bottleneck towards supply constraints for 'Entire Room/Apt' category in that city.

Furthermore, the below charts exhibit other surprising discoveries in the EDA process. More specifically, neither Superhost(?), distance towards valuable attractions, or guest satisfaction are surprisingly not valuable features towards Airbnb pricing. These insights can help Airbnb

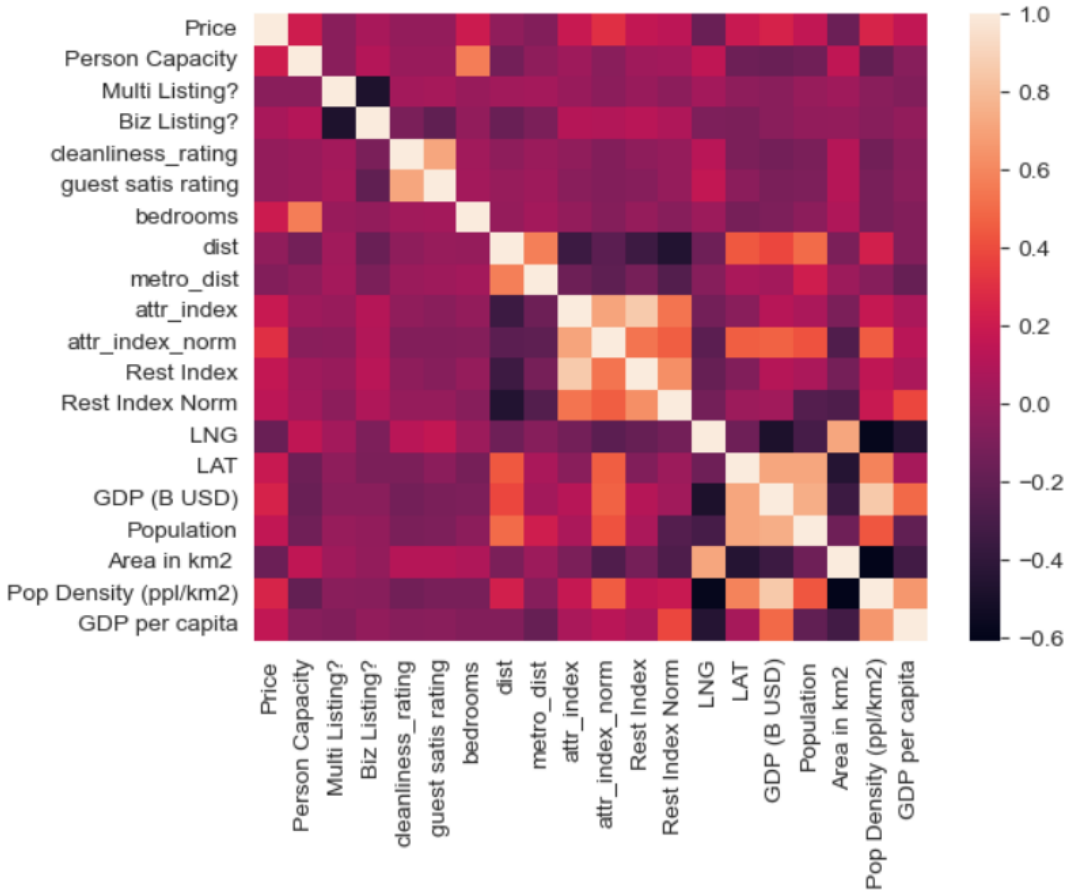






3. Heatmap/Pairplot:

In the below heatmap, you can discern that no room features correlate significantly with price. Furthermore, there is multicollinearity between many room features, which may distort the Machine Learning models in the next stage.



Machine Learning Modeling:

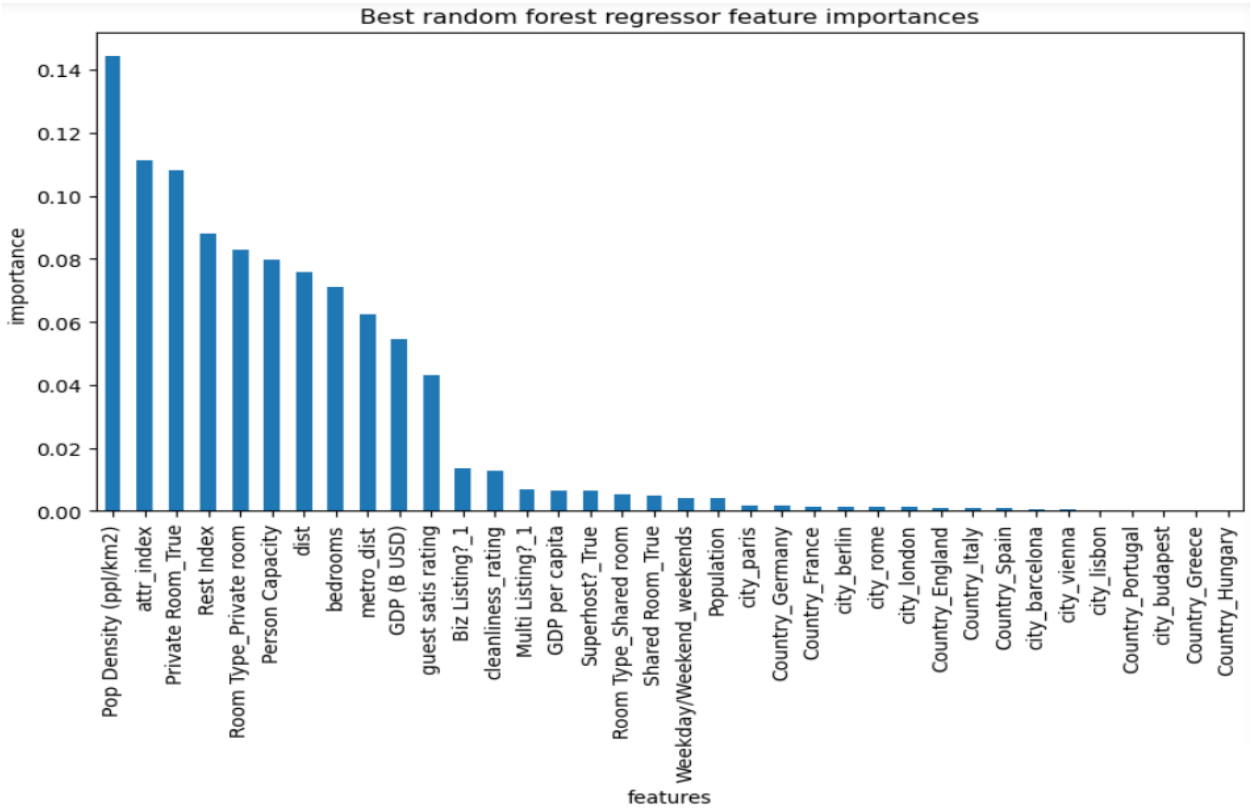
Five different Machine Learning Regressors – Linear Regression, Lasso, Random Forest, Gradient Boost, KNN Regressor - were fitted onto the dataset to formulate a predictive model and determine feature importance. Below are the performance metrics for each regressor:

Performance Metrics:

	Train			Test		
	R2 score	MAE	RMSE	R2 score	MAE	RMSE
	Train					
Linear Regression	0.5576	69.0400	95.7700	0.5528	69.0402	95.7781
Lasso Regression	0.5564	67.3681	95.8944	0.5564	68.8308	95.8944
Random Forest Regressor	0.9750	13.9100	22.2674	0.8312	37.4394	58.8542
GradientBoost	0.6699	55.6866	81.0265	0.6590	57.0701	83.6374
KNN Regressor	0.6440	52.6397	78.0558	0.6506	57.0430	78.0558

Random Forest Best Estimator:

```
{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 150}
RandomForestRegressor(n_estimators=150, random_state=44)
```



Analysis:

Random Forest regression is the unequivocal performance front-runner across all regressors. It's fascinating that this algorithm can ascertain city significance on Airbnb room prices - albeit indirectly via the feature-engineered 'Population' and GDP variables - and with remarkable accuracy (R-squared ~.84 on the test case) and low MSE (39).

If compelled to use this regression model without levity to re-engineer features, including performing separate multiple regressions for respective cities (this topic detailed more below), Random Forest model is the clear choice. Airbnb hosts can benefit from this analysis by grasping the importance of these features, in order of significance:

1. Attr - Index (room proximity to attractions)

2. Private rooms are significantly valued

3. Rest-Index (proximity to valuable restaurants)

4. Person Capacity - # of people per unit, which correlates with Room Type (entire room/apt vs private room)

5. Dist (distance from city center), which correlates with distance from attractions

6. Bedrooms

7. Distance from Metro

8. Guest - Satisfaction

Cautionary Context: The limits to this machine Learning model and dataset

Limited macroeconomic variables for robust analysis of overall supply-demand characteristics:

Overall, the datasets contain limited data points regarding the macroeconomic dynamics of the lodging industry to assess overall supply and demand characteristics. These include the price elasticity of different competitors to Airbnb, the consumer attractiveness of the tourist attractions in each respective city, the cost of transportation and logistics to and within the cities, the average income of travelers, cities' Gini-coefficient, and much more. Far more dimensions are needed to conduct a thorough multi-variate, econometric analysis of this industry.

Limited microeconomic / room features for robust analysis of Airbnb rooms competitive advantage

The current datasets attached contain mostly microeconomic parameters, mostly pertaining to the room features themselves. However, even within these parameters, a lack of access to data on many important dimensions remains. Of most noticeable is tourism-related information, especially room proximity to tourist attractions, since tourism proximity plays an important choice for most Airbnb users. Other important demand-side attributes include tourist attraction

attractiveness within the city, including the overall price elasticity of these attractions. Supply-side factors overlapping with these sentiments include the overall supply, of both Airbnb and other consumer lodging choices, especially in proximity to tourist attractions.

Limited business dimensions to assess business operations:
On the business side: Host reputation, hospitality, and service are other critical attributes missing. The dataset does not contain information pertaining to valuable services or host personality characteristics provided by hosts/super hosts that may allow for increased pricing or even the sustainable competitiveness of the room itself.

Lack of clarity of socio-cultural dimensions within consumer psychology for each respective city:
On Airbnb's online platform: There may exist latent inefficiencies in how Airbnb presents and markets its room information across respective cities' socio-culture milieu. For instance, there exist variances of social psychologies across Europe's plurality of cultural experiences. For example, across regions, perhaps there exists diminished trust towards online platforms, or differences in psychologies underlying how rooms are photographed for appeal, and much more. It's important to become sensitive to cultural perceptual differences and become cognizant of hidden cultural assumptions, including the potential projection of Airbnb's headquartered (American) culture as an objective representation of reality within this online platform. Likewise, varying cultures may project their values, beliefs, and social psychologies onto this platform, and it's important to remain aware of unintended latent inefficiencies from these more inconspicuous factors that may affect how Airbnb should design its online platform for consumption for each respective culture.

Miscellaneous factors:
Other factors include limited city, country, or geographical features pertaining to each respective room in these datasets. These include housing or zoning policies that may limit Airbnb, or overall lodging supply within the respective cities, or capped investment or capacity of each respective room.

It's important to remember that this dataset contains only a subset of important attributes within the entire ecosystem of variables that determine price.

Overall Analysis of the Machine Learning Model:

The dataset features and consequent machine learning model is reductionist in its interpretation of the overall socio-economic and cultural ecosystem that influence Airbnb pricing, and thus, may become misleading in its predictions. Although this study produced some valuable insights, more rigorous studies are recommended.

Future Recommendations:

Future Studies:

Separate linear regressions should be conducted for each individual city: As easily discerned via the EDA stage, each city follows a distinct price dynamic. Additionally, the geographic diaspora of Europe can lead to distinct socio-cultural and economic elements that affect pricing for each respective city.

Furthermore, for each additional city, separate linear regressions should be computed for each price segment - i.e. bargain, economic, luxury, ultra-luxury etc. During the EDA stage, it can be discerned that the price gradient vs different room types and person capacities differed for each segment in the typical consumer expense categories. Separate regressions should be performed for each consumer category to achieve a more accurate analysis of consumer psychology within each category.

Implement more methods to collect and increase variety of features in dataset: As previously discussed, the current model is reductionist in its approach to pattern detection for it does not capture the entire ecosystem of variables that determine pricing.

Airbnb operations-specific recommendations:

Insert Random Forest's feature importance list as a section within Airbnb host's platform: Although the ML models are reductionist, they still do possess interesting discoveries. For instance, Airbnb can benefit from knowing what features are more important than others.

Deploy more resources to investigate the supply-demand imbalance in Barcelona: As discussed in the EDA stage

Similarly, investigate the premium category for all cities with low proportion of 'Entire Room/Apt': The 'Entire Room/Apt' category in is the main cash cow for Airbnb. Thus, Airbnb should deploy more resources into investigating why the cities with such low proportion of this category exhibit this characteristic.