Capstone Project

-Project approaches for Data Cleansing-

Mentor:
Maryam Najimigoshtasb

Member:
Pablo Martin Ruiz Diaz

Content

Data Cleansing..... 3

Listings_datailed Table:..... 3

Reviews_detailed Table:..... 4

Calendar Table:.....4

Data Transformation..... 5

Listings_datailed Table:..... 5

Data Cleansing

Listings_datailed Table:

1. Price column doesn't show outliers in situ. They are affected by location, amenities, quality and so on. Nevertheless, there are values less than usd20, thus we can drop them because they represent only 3% of all.
2. We consider listings as incorrect when minimum nights are greater than maximum nights.
3. AirBnB's main market is for travelers, short-term students and business trips. For this we consider a rent when the length of stay is equal or more than a year. For this reason:
    3.1. Analysing minimum nights, a 75% up of upper whisker is 180. It represents a semester and is possible for its main market.
    3.2. For maximum nights, we talk about years in some cases. From my point of view, hosts just write a random number, then they decide with the user about the total nights. Therefore, the maximum night column does not represent an important column to take into consideration in our analysis. Note: an outlier is when a maximum night is 2700 nights (around 7 years).
4. Because in our data we count hotels, hostels and also castles, a maximum of 16 accommodates is reasonable. Nevertheless, if this is the max number it's impossible to have almost 35 bathrooms, more than 100 beds or 15 bedrooms. For this reason, we consider outliers when the numbers go out of maximum accommodations.
5. Listings where they don't have any first_review value, mean they're new. As a consequence, there aren't any review scores and missing values on them.
6. The following variables represent almost 5% of all data. Besides, there are categorical and numeric columns. Therefore, I decided to drop them.
    - Name, beds, description, review_scores_value, review_scores_location, review_scores_checkin, review_scores_accuracy, review_scores_communication, review_scores_cleanliness, bathrooms and bathrooms_type.
7. Null values in categorical columns may be treated by the world cloud. Nevertheless, we should remember this technique is to recommend to a user possible words but not a prediction to get a value, such as a name.
8. Missing values in the bedroom variable may be filled by some machine learning technique. We can try a variety of them to check which one has better results.
    8.1. Approximation of correlation as first filter on dependency/independency variables: *bedrooms ~ price + property_type + room_type + accommodates + bathrooms + bathrooms_type + bedrooms + beds + city + state*
    8.2. Shapiro test, all the p-values are less than 5%. Thus, it can be assumed that the data are not normally distributed. Besides, we compared the Anderson-Darling test too, because the Shapiro test has a maximum of 5000 values to evaluate.

8.3. Because the correlation graph, we sum up the following assumptions:
  8.3.1. We go through variables with high correlation, this means results higher than 0.5.
  8.3.2. Our bedrooms variable is dependent on price, accommodates, bathrooms and beds. Nevertheless, we have to select one-to-one variable relation because of the VIF (variance inflation factor). So we can just select the highest correlated variable.
  8.3.3. We analyze VIF because of multicollinearity. However, it does not affect the regression model because we select just one variable as independent.

8.4. Finally, after all we analyze we may conclude on applying two different models and choose which one is better.
  8.4.1. *Bedrooms ~ Accommodates*
  8.4.2. Bedrooms ~ Beds

8.5. Feature selection is the process of reducing the number of input variables when developing a predictive model. However, we just use one variable that is not necessary to apply this technique.

8.6. Last but not least, it is not necessary to apply PCA, because we just have one variable as independent to apply for our model and we do not need to apply any variable-reduction technique.

8.7. After checking linear and polynomial regressions, and of course cross validation, we assume Bedrooms ~ Accommodates is the best with the following model features:
  8.7.1. Polynomial, degree = 10;
  8.7.2. Mean Square Error = 0.3090
  8.7.3. Coefficient of Determination = 0.7300

Reviews_detailed Table:

1. Reviews represent comments from diverse users who booked listings previously. Thus, there are not missing values or outliers.

Calendar Table:

1. Calendar represents from March 2023 to May 2024 all the listings not available booked for users. Thus, there are not missing values or outliers.

Data Transformation

Listings_datailed Table:

1. Because I found some mistakes in the room_type column (property_type=='Room in hotel'), it is necessary to create a new one. We can differentiate or categorize property_type variables and name it as “room_category”.
   This new column has three variables: House, Room and Others, and each one is made from specific words in the property_type variable.
2. There is not any insight from price variable vs others to apply some clustering technique or something else.
3. From minimum_nights variable we create a new column names minimum_duration with the following rules:
   3.1. minimum_nights < 7 is 'short_term';
   3.2. minimum_nights < 30 is 'mid_term';
   3.3. else 'long_term'.
4. Analyzing correlation with heatmap over all variables, we cannot find any interesting insight apart from all of them we did.