EDA project

You are required to perform an EDA on Airbnb dataset, This dataset (raw.csv) contains 30k+ records on hotels in the top-10 tourist destinations and major US metropolitan areas sraped from Airbnb.com.

Each data record has 40 attributes including the number of bedrooms, price, location, etc.
The attribute "pop2016" means population of the zipcode location (area) in year 2016.
Demographic and economic attributes were scraped from city-data.com.

Attributes description:

- House specific features, collected from Airbnb.com:

1. **Bathrooms:** The number of bathrooms in the listing
2. **Bedrooms:** The number of bedrooms
3. **Beds:** The number of bed(s)
4. **LocationName:** Location of the house
5. **NumGuests:** Maximum number of guests can hold
6. **NumReviews:** number of reviews received
7. **Price:** daily price in local currency
8. **Rating:** Y/N - whether the rating of each house is 5 or not
9. **latitude:** location information latitude
10. **longitude:** location information longitude
11. **zipcode:** zipcode of the house

- demographic and economic attributes based on zipcode, collected from city-data.com (means the same zipcode should share the same value for each of the following attributes)

1. **pop2016:** popularity of the area reported in 2016
2. **pop2010:** popularity of the area reported in 2010
3. **pop2000:** popularity of the area reported in 2000
4. **cost_living_index:** a U.S standarded index for cost living measurement
5. **land_area:** space of land
6. **water_area:** space of water area
7. **pop_density:** density of population
8. **number of males:** within the area population
9. **number of females:** within the area population
10. **prop taxes paid 2016:** Median real estate property taxes paid for housing units in 2016
11. **median taxes:** median of taxes paid by house owners in the area
12. **median house value:** median of house value in the area
13. **median household income:** median of income of house owners in the area
14. **median monthly onwer costs (with mortgage):** median monthly cost of house owner including mortgage

15. **median monthly onwer costs (no mortgage):** median monthly cost of house owner without considering mortgage
16. **median gross rent:** the monthly rent agreed or contracted for plus the estimated monthly cost of utilities and fuels.
17. **median asking price for vacant for-sale home/condo:** median asking price for for-sale home in the area
18. **unemployment:** umemployment ratio of the area

Aggregated features for Abnb by zipcode

1. **Number of Homes Count of Abnb:** number of Abnb houses in this area
2. **Density of Abnb (%):** ratio of Abnb houses in this area
3. **Average Abnb Price (by zipcode):** aggregated by zipcode
4. **Average NumReviews (by zipcode):** aggregated by zipcode
5. **Average Rating (by zipcode):** aggregated by zipcode
6. **Average Number of Bathrooms (by zipcode):** aggregated by zipcode
7. **Average Number of Bedrooms (by zipcode):** aggregated by zipcode
8. **Average Number of Beds (by zipcode):** aggregated by zipcode
9. **Average Number of Guests (by zipcode):** aggregated by zipcode

The prediction label is **Rating of house**.

Given the dataset, write code and answer the following three questions.

(1) (5 marks) Propose 4 questions (non-predictive and non-trivial) that you believe are interesting to explore and can be answered using the provided dataset (at least 3 question should be answered using hypothesis test). Briefly describe why you think those questions are interesting to whom. You can answer this question in a markdown cell of your ipynb file.

(2) (15 marks) Analyze the quality of data (all columns) and report statistics of missing data and their missing mechanism and why did you choose this mechanism, and how will handle these nulls (there's no one true solution but you must explain your reasons in a separate markdown and you must missingno package to support your claims), the report must answer the following questions

2.1 does missing value exit in the table?
2.2 Where are the missing data?
2.3 How much data is missing?
2.4 Are there any variables often missing together?

- Explore outliers and duplicated, and how would you handle them (there's no one true solution but you must explain your reasons in a separate markdown), Briefly describe your step and findings.

(3) (15 marks) For the 4 questions you proposed in the first subquestion, what are the null hypothesis and alternative hypothesis? Perform statistical test to answer your question and report your findings.

(3) (10 marks) Make some visualizations about your columns and explain what did you got from each visualization.

(5) (5 marks) Briefly describe your feature engineering plan and code it (at least 3 columns should be involved).