Homework 4: Data Cleaning and Pre-Processing

Complete the following exercises. Remember to explain your answers.

You are working with a dataset of Airbnb rental listings from a major city. The dataset contains information about various properties, including location, price, amenities, host information, and guest reviews. Your task is to analyze this dataset and plan the appropriate data cleaning steps before using it for machine learning.

1. **Importance of Data Cleaning and Preprocessing**: During class we discussed that while data cleaning takes a lot of time and effort, it is the most impactful thing you can do to improve performance of your models.

   (a) Explain two specific techniques for data cleaning, and how they can improve machine learning model performance. Keep your explaination to 1-2 sentences.

   -

   -

   (b) Describe a scenario with Airbnb rental data where poor data cleaning might lead to biased or inaccurate predictions. Explain how this could impact real-world decision-making for either hosts or guests.

2. **Exploratory Data Analysis**: You've been provided with the following summary information about the Airbnb dataset:

   - The dataset has 10,000 listings
   - 15% of listings are missing review scores
   - 5% of listings have prices listed as "$0" or "Contact host"
   - There are 20 duplicate listings (same property listed multiple times)
   - The "neighborhood" column has 45 unique values, but 3 values appear as both "Downtown" and "City Center"
   - Host response times are categorized as "within an hour", "within a day", "within a few days", and "N/A"

   For each of the following issues, explain how you would handle it and justify your approach based on the data cleaning principles discussed in class:

   (a) Missing review scores

   (b) Problematic price values

   (c) Duplicate listings

3. **Feature Selection and Transformation**: The following columns are just some of the columns in the Airbnb dataset:

   - listing_id (unique identifier)
   - host_id (unique identifier for each host)
   - property_type (apartment, house, etc.)
   - room_type (entire home, private room, shared room)
   - accommodates (number of people)
   - bathrooms (number of bathrooms)
   - bedrooms (number of bedrooms)
   - price (per night in USD)
   - minimum_nights (minimum stay required)
   - maximum_nights (maximum stay allowed)
   - availability_365 (number of days available in the year)
   - number_of_reviews (total reviews received)
   - reviews_per_month (average reviews per month)
   - review_scores_rating (overall rating 0-100)
   - latitude and longitude (property coordinates)
   - calculated_host_listings_count (number of properties the host has)
   - host_response_rate (percentage)
   - host_is_superhost (yes/no)

   (a) Identify 3 columns you would consider dropping and explain your reasoning.

   (b) For the "price" column which shows a heavily right-skewed distribution (most properties are affordable but some are extremely expensive), suggest an appropriate transformation and explain why it would be beneficial.

   (c) Suggest 2 new features you could engineer from the existing columns that might be helpful for predicting rental prices or popularity. Explain how you would create them and why they would be valuable.

   -

   -

4. **Handling Distributions and Outliers**: You notice the following patterns in the Airbnb dataset:

   - 80% of listings are for "entire home/apt", 18% for "private room", and only 2% for "shared room"
   - There are 10 properties with prices over $1000/night, compared to a median price of $120/night
   - Most properties have between 0-100 reviews, but a few popular ones have over 500 reviews

   (a) How would you address the imbalance in the room_type distribution if you wanted to build a model that performs well across all room types? Discuss one approach and its potential benefits and drawbacks.

   (b) Describe your approach to handling the price outliers. Would you remove them, transform them, or keep them as is? Justify your answer considering both statistical modeling needs and business requirements.