**Part 1: Data Cleaning (7.5 Marks)**

1. **Share your findings for data cleaning.**

Here are some findings for data cleaning of Airbnb dataset:

* + - * 1. **Missing Values:** Certain columns have missing values.
      * 'Review Scores Rating (bin)' and 'Review Scores Rating' both have 8,309 missing values.
      * 'Beds' have 85 missing entries.
      * 'Zipcode' has134 missing entries.
      * 'Host Since' and 'Property Type' are almost complete, with only 3 missing entries each.

1. **Unsuitable Data Types:** There are some columns we need to change its data type, as below:
   * + - ‘Host Since’ should be in **datetime** format.
       - ‘Price’, ‘Beds’, ‘Number of Reviews’, ‘Review Scores Rating’, and ‘Price’ should be in the **integer** format.
       - ‘Zipcode’ should be in the **string** format.

* + - * 1. **Missing Values**: Certain columns have missing entries:

|  |  |  |
| --- | --- | --- |
| **Column** | **Row Counts of Missing Entries** | **How to solve?** |
| Review Scores Rating (bin) | 8,309 | Create separate columns to categorize the rows where their values are missing. Then, replace those missing entries as zero. |
| Review Scores Rating | 8,309 |
| Beds | 85 |
| Host Since | 3 | Remove those 3 rows |
| Property | 3 | Replace it with the value “Other” |

* + - * 1. **Unsuitable Data Types:** There are some columns we need to change its data type, as below:

|  |  |
| --- | --- |
| **Column** | **Correct Data Type** |
| Review Scores Rating (bin) | Integer |
| Review Scores Rating | Integer |
| Beds | Integer |
| Host Since | datetime |
| Property | string - categorical |
| Price | integer |

**Redundant or Irrelevant Column:** The 'Number of Records' column contains only the value 1 for all entries, making it redundant for any analysis. And that’s why we will remove this column.

1. **How are you dealing with null values?**

For 'Review Scores Rating', 'Review Scores Rating (bin)' and ‘Beds’ columns, I would create separate columns to identify the rows where their values are missing. After this, I will replace the missing values in these 3 columns with 0.

For ‘Host Since’ column, I simply remove the 3 missing rows, given its insignificance in the dataset's size.

For ‘Property Type' and “Zipcode” column, I replace missing values with the “Other” value.

1. **Do you think some item data types need to be changed? Why do you think so.**
   * 'Host Since' should be changed to a datetime type to analyze trends over time properly.
   * 'Price' must be converted from object type to float after removing any non-numeric characters like currency symbols.
   * 'Zipcode' could be treated as a string since it is a categorical variable rather than numeric.

**Part 2: Understanding your dataset (15 Marks)**

1. **Why have you chosen this dataset?**
2. **Explain your dataset in detail.**
   * The dataset appears to contain information related to Airbnb listings, including:
     + Host-related information (ID and registration date on Airbnb).
     + Listing-related information (name, neighborhood, property type, room type, number of beds, and the number of reviews).
     + Pricing information for the listings.
     + Ratings information from customers (review scores overall).
3. **What key performance indicators you can identify of guests?**
   * Key performance indicators regarding guests could include:
     + Average price per room type and neighborhood.
     + Average number of reviews per listing as a proxy for occupancy rate.
     + Review Scores Rating as an indicator of guest satisfaction.
4. **What new information, indicators can be drawn through this dataset?**
   * Potential occupancy rates based on the number of reviews.
   * Seasonality in booking rates based on 'Host Since' dates and possibly linking to pricing trends.
   * Influence of property type, room type, and location on customer satisfaction and price.
5. **What analysis can make this a purposeful report?**
   * Correlation analysis between price, number of reviews, and review scores.
   * Comparative analysis of different neighborhoods and room types regarding price and reviews.
   * Time-series analysis to understand trends and seasonality effects.
6. **What kind of modeling are you planning to use? Please elaborate reasons to choose the model.**
   * Regression models could be useful to predict factors influencing price or satisfaction scores.
   * Time-series analysis for trend identification of prices and reviews over time.
   * Cluster analysis to group similar listings and identify unique characteristics.

**Part 3: Identify focused Area of Analysis (15 Marks)**

1. **Identify and select which fields/columns your will focus your analysis?**
   * Price: It is essential to understand the economic aspect of Airbnb listings.
   * Neighbourhood: This impacts the desirability and price of rentals.
   * Property Type and Room Type: They are crucial in categorizing and comparing different kinds of listings.
   * Number Of Reviews: This can be a proxy for the popularity or occupancy rate of a listing.
   * Review Scores Rating: Provides insight into guest satisfaction and potentially the quality of a listing.
2. **Please provide reason for choosing these fields/columns? Why do you think they should be more focused?**
   * These fields are directly related to guest preferences, market pricing strategies, and quality indicators for the listings. They provide a detailed picture of the competitive landscape of Airbnb rentals and user satisfaction, which can drive business and operational strategy.
3. **What is your current understanding of these selected fields/column?**
   * *Price*: Reflects how much guests are expected to pay, which could be influenced by several factors such as location, room type, and quality of the listing.
   * *Neighbourhood*: Represents the area where the property is located. Different neighborhoods may have different attributes that appeal to guests and can also influence price and number of bookings.
   * *Property Type and Room Type*: Indicate the kind of accommodation being offered, which could be a deciding factor for guests and affect pricing.
   * *Number Of Reviews*: Serves as a proxy for how often a listing is booked; frequent bookings often result in more reviews.
   * *Review Scores Rating*: Illustrates guest satisfaction and can be a predictor of repeat bookings and recommendations.
4. **What conclusion you can draw after drilling down these columns/fields?**
   * I could draw conclusions about the factors that most strongly correlate with high prices and guest satisfaction. For example, certain neighborhoods or property types may consistently rate higher or command higher prices.
5. **Which other fields can be directly or indirectly impacted with your conclusion?**
   * 'Host Since' may indirectly show the experience level of the host, which could impact guest satisfaction.
   * 'Beds' and 'Zipcode' could further refine price predictions, as more beds typically mean a higher price, and zip codes can denote specific parts of neighborhoods with varying characteristics.
   * Indirect relationships could also exist between these focused fields and other unexamined fields, such as specifics of the listing descriptions or amenities offered, which might impact guest satisfaction and reviews.

Next steps would typically involve a deeper statistical analysis of these fields to establish correlations and causations, and possibly predictive modeling based on the observed patterns. The resultant insights would then be used to build a comprehensive and purposeful report. Shall we go ahead and carry out this analysis, or is there anything specific you would like to address?

**Part 4: Preparing for Dashboard Report (20 marks)**

1. **What platform you are using for dashboard report? Explain in 5 points that why you have given preference to your chosen platform over others?**
   * The platform chosen for the dashboard report is Power BI, and here are five advantages it offers:
     + **Cost-effective**: Power BI offers a robust free version and the Pro version is relatively inexpensive compared to some alternatives, which can be a decisive factor for many organizations.
     + **DAX and Advanced Analytics**: Power BI's Data Analysis Expressions (DAX) language is powerful for creating custom calculations.
     + **Custom Visuals**: The platform supports a wide range of custom visuals which can be imported or created, allowing for highly personalized reports.
     + **Regular Updates and Strong Community**: Microsoft frequently updates Power BI with new features and improvements, and there is a strong community of users that provide support and share knowledge.
2. **Can you draft a sketch of how do you want your dashboard to look like?**
   * + **Top Section**: A title banner with the report name and any overarching filters that affect the entire dashboard (e.g., date range, specific areas).
     + **Left Section**: Slicers for interacting with the dashboard in the form of dropdown filters for neighbourhood, property type, room type, etc.
     + **Central Section**: Consisting of key visual highlights such as KPIs for average price and satisfaction score, a map visualization for geographical data, and a bar chart for average prices by neighbourhood.
     + **Right Section**: A line chart for trends over time (price and reviews) and a scatter plot for price vs. satisfaction rating.
     + **Bottom Section**: Detailed data grid with individual listings and the ability to drill down for more information.
3. **Where do you want your audience to put attention to? How are you planning to do that?**
   * The attention should be drawn to key insights like overall performance, customer satisfaction, and price trends. Using bold headers, strategic placement (top and center), and contrast colors will help direct attention.
4. **How are you planning to create data story through your visuals?**
   * The dashboard will start with high-level insights and progressively allow users to drill down into specifics. The story starts with overall market health (KPIs), moves to spatial distribution (map), analyzes trends (line chart), and ends with detailed transactional data (data grid).
5. **Can you elaborate the narrative you are using for data story?**
   * The narrative will explore how various factors such as location, property type, and satisfaction influence rental prices and popularity on Airbnb. The goal is to provide actionable insights into market dynamics and identify opportunities or areas for improvement.
6. **How are you ensuring that your visuals are interactive?**
   * By using slicers, drill-throughs, and other interactive elements in Power BI, the visuals will allow users to filter and explore data dynamically. Tooltips will provide additional context on hover.
7. **What formulas/DAX/Calculated fields you have used? What visuals you have used for them?**
   * One might use DAX to calculate average price per neighbourhood, occupancy rate (via Number of Reviews), and guest satisfaction index (using Review Scores Rating). These calculations could be represented as cards (for KPIs), bar charts (for average price by neighbourhood), and a matrix (for occupancy and satisfaction).
8. **How did you verify your formulas if they are returning correct values?**
   * Formulas can be verified by checking against known values, creating test cases, or comparing results from Power BI with those calculated externally (e.g., in Excel). 'DAX Studio' or querying in Power BI's data view can also help in validating calculations.

**Part 6: Final Report (12.5 Marks)**

1. **Briefly explain your dashboard workbook layout and legends.**
   * The dashboard layout organizes the information logically from general to specific, allowing users to navigate easily. The top section provides an overall summary, the central section visualizes spatial and trend data, and the bottom part offers detailed listing information. Legends are included alongside each visual to allow for quick interpretation of colors, symbols, and chart elements.
2. **Please explain what information your visuals are sharing?**
   * *Map Visualization*: Showcases the spatial distribution of Airbnb listings and how factors like neighbourhood impact price and popularity.
   * *Key Metrics*: Highlight important data points such as the average price, average review score, and total number of reviews, providing a snapshot of the market.
   * *Trend Analysis*: Illustrates how prices and guest satisfaction change over time, which could indicate market trends or seasonal effects.
   * *Bar Chart*: Comparing average prices across neighbourhoods, highlighting the areas that are premium or more affordable.
   * *Line Chart*: Occupancy trend indicated by the number of reviews, which can signify high-demand periods.
   * *Scatter Plot*: The relationship between price and guest satisfaction, exploring whether higher satisfaction is associated with higher prices.
   * *Data Grid*: Detailed listing data with the ability to sort and filter, allowing users to analyze individual listings.
3. **Please explain what problems and challenges you have faced during this project.**
   * Challenges can include cleaning and preparing the data, ensuring accuracy in the DAX formulas, and designing a dashboard that communicates insights effectively. Data may have issues such as missing values, duplicates, or inconsistencies that need addressing. Creating compelling visuals that accurately represent the data and are also user-friendly can be another hurdle.
4. **What are your key learning points during this assignment? Please mention any three.**
   * **Data Preparation**: The importance of thoroughly cleaning and structuring data before analysis cannot be overstated.
   * **Storytelling with Data**: Learning how to best craft a narrative with visualizations to guide the audience through the analysis.
   * **Understanding the Audience**: Designing the report with the end-user in mind, creating an easy-to-navigate layout, and providing clear and actionable insights.
5. **How are you going to apply these skills in your daily routine tasks? Briefly mention how this project will help you?**
   * The skills acquired can be utilized across various tasks that involve data analysis, such as monitoring KPIs, report generation, and decision-making processes. The ability to cleanse data, synthesize information into insights, and communicate findings clearly will be invaluable in supporting data-driven business strategies.

**Part 7: Create Problem Statement (10 Marks)**

1. **Please clearly state and explain what problem or problems you have drawn from your analysis?**
   * One possible problem is a discrepancy in guest satisfaction across different neighbourhoods or property types, which might suggest a need for improved standards or targeted investments.
2. **What are the reasons for these problems? How did you identify the problem root cause?**
   * The root causes could include factors such as inadequate amenities, mispricing, or location drawbacks. These were identified through the correlation of low review scores with particular attributes in the visualizations.
3. **What actions can be recommended through your analysis as a remedy for problems? Please mention a remedy against each problem.**
   * To address gaps in guest satisfaction, hosts could be provided with training or resources to enhance the guest experience. For mispricing, dynamic pricing tools could be recommended to better match prices with market expectations.
4. **What factors are putting negative and positive impact on key performance indicators? Please list down.**
   * **Negative Impacts**: High prices not justified by value, poor location accessibility, and lack of essential amenities.
   * **Positive Impacts**: Competitive pricing, excellent host communication and service, prime location, and well-furnished and maintained properties.

Through analyzing data and creating a dashboard, you would have been able to highlight these insights and propose data-driven solutions to optimize performance on the Airbnb platform. This exercise offers a guideline on how the findings could be presented in the report and actionable outcomes derived from the analysis.