**Step 1: Criteria for Selecting Cities for Analysis**

1. **Geographical Distribution**: The dataset from Inside Airbnb includes cities across North America, Europe, Asia, and Australia, though data for other continents required a fee. Europe had the largest representation, so we selected three European cities. Additionally, we chose two cities from the U.S. and Canada, the second-largest region in the dataset. Cities from Asia and Australia were excluded due to limited data.
2. **Airbnb’s Travel Trends**: We consulted a detailed article on Airbnb’s impact on sustainable tourism and travel trends. Key insights influencing our selection included:
   * **Redistributing Tourism: Airbnb promotes less-visited neighborhoods in major European cities, reducing congestion in traditional hotspots.**
   * **Market Share: Despite Airbnb’s growing influence, hotels still dominate the accommodation market in European cities.**
   * **Pandemic-Induced Shifts: Post-COVID, travelers are opting for domestic and rural destinations, with increasing interest in less-crowded areas as international travel rebounds.**
   * **Economic Benefits: In cities like Amsterdam and London, Airbnb is expanding affordable lodging options in underserved neighborhoods, contributing to local economies.**
   * **Tech-Driven Discovery: Features like "Airbnb Categories" and "I’m Flexible" are helping travelers discover less-known areas, supporting more sustainable tourism.**
   * **Emerging Neighborhoods: Areas like Volewijck in Amsterdam, Campolide in Lisbon, and Nou Barris in Barcelona are gaining popularity as Airbnb guests seek more local, authentic experiences.  
     Based on these insights, we selected Amsterdam and London from Europe for further analysis.  
     Source:** [**How Airbnb Disperses Travel in Europe**](https://news.airbnb.com/how-airbnb-disperses-travel-in-europes-major-cities/)**.**
3. **U.S. City Selection**: For the U.S. and Canada, we focused on cities with high Airbnb popularity and revenue generation. **Based on these insights, we selected San Francisco, New York and Canada from US and Canada respectively**  
   **Source**: [**Best Airbnb Cities in the USA**](https://masterhost.ca/best-airbnb-cities-usa/).

**Step 2: Data Pre-Processing**

1. **Data Preparation**: We performed the following steps to preprocess both the listing and review datasets:
   * **Created a table associating city IDs with five selected cities for easier categorization and future merging of the datasets.**
   * **Loaded the .csv files into data frames to facilitate analysis.**
   * **Removed irrelevant rows from both datasets.**
   * **Identified and eliminated rows with null values in key features (e.g., price and review score rating for Amsterdam, New York, and London).**
   * **Cleaned the price feature by removing the "$" symbol and commas, converting the data to floats.**
   * **For the review dataset, visualized the "comments" feature using a word cloud to identify frequent terms.**
   * **Preprocessed text data by converting to lowercase, removing special characters, HTML tags, and numerical values for a cleaner analysis.**
   * **Saved the pre-processed data into a .csv file for future use.**

**Step 3: Task 1 and Task 2 and Task 6**

1. **Data Preprocessing for Task 1**:  
   As a prerequisite, we addressed discrepancies in the minimum\_nights and maximum\_nights features. In some cases, the values exceeded one billion, indicating data corruption. These erroneous values were removed to ensure accurate numerical analysis. Additionally, the data types of these features were corrected to align with the requirements of Task 1.
2. **Numerical Analysis for Task 1**:  
   We conducted a numerical analysis on key features, including price, minimum\_nights, maximum\_nights, number\_of\_reviews, and review\_score\_rating. The analysis generated the following statistical metrics:
   * **Count**
   * **Mean**
   * **Standard Deviation**
   * **Minimum Value**
   * **25th Percentile**
   * **50th Percentile (Median)**
   * **75th Percentile**
   * **Maximum Value**
3. **Insights from the Analysis**:  
   These statistics provided valuable insights into the data’s distribution, variability, and central tendency. However, the results were significantly affected by the presence of outliers in the dataset, which skewed the overall analysis.
4. **Numerical statistics before removing outliers:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Count** | **Mean Price** | **Std Dev Price** | **Min Price** | **25th Percentile Price** | **Median Price** | **75th Percentile Price** | **Max Price** |
| Amsterdam | 5,689 | € 289.64 | 1216.78 | € 11 | € 164 | € 225 | € 321 | € 91,466 |
| New York | 22,789 | $232.89 | 1019.76 | $8 | $85 | $147 | $250 | $100,000 |
| London | 63,203 | £197.15 | 839.64 | £1 | £76 | £130 | £211 | £80,000 |
| San Francisco | 6,155 | $239.71 | 1000.32 | $25 | $100 | $150 | $243 | $50,000 |
| Toronto | 16,536 | $196.53 | 308.19 | $3 | $80 | $139 | $235 | $12,400 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Mean Nights (Min)** | **Std Dev (Min)** | **Min Nights** | **25th Percentile (Min)** | **Median (Min)** | **75th Percentile (Min)** | **Max Nights (Min)** |
| Amsterdam | 4.22 | 19.85 | 1 | 2 | 3 | 4 | 1,001 |
| New York | 28.97 | 30.62 | 1 | 30 | 30 | 32 | 1,250 |
| London | 3.54 | 23.06 | 1 | 2 | 3 | 4 | 1,125 |
| San Francisco | 20.69 | 48.71 | 1 | 1 | 3 | 30 | 1,125 |
| Toronto | 24.37 | 37.98 | 1 | 28 | 28 | 30 | 1,239 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Mean Nights (Max)** | **Std Dev (Max)** | **Min Nights (Max)** | **25th Percentile (Max)** | **Median (Max)** | **75th Percentile (Max)** | **Max Nights (Max)** |
| Amsterdam | 320.72 | 417.47 | 1 | 20 | 31 | 365 | 1,125 |
| New York | 578.89 | 453.52 | 1 | 150 | 365 | 1,125 | 3,000 |
| London | 476.47 | 433.66 | 1 | 60 | 365 | 1,125 | 1,682 |
| San Francisco | 390.24 | 413.11 | 1 | 30 | 360 | 365 | 3,209 |
| Toronto | 466.74 | 413.66 | 1 | 90 | 365 | 1,124 | 3,650 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Mean Rating** | **Std Dev Rating** | **Min Rating** | **25th Percentile (Rating)** | **Median Rating** | **75th Percentile (Rating)** | **Max Rating** |
| Amsterdam | 4.84 | 0.26 | 1 | 4.78 | 4.92 | 5 | 5 |
| New York | 4.73 | 0.44 | 1 | 4.64 | 4.85 | 5 | 5 |
| London | 4.69 | 0.4 | 1 | 4.64 | 4.89 | 5 | 5 |
| San Francisco | 4.79 | 0.29 | 1 | 4.75 | 4.91 | 5 | 5 |
| Toronto | 4.78 | 0.38 | 1 | 4.73 | 4.89 | 5 | 5 |

* **Price Insights:**
  + Amsterdam has the highest average price (€289.64), with a large standard deviation indicating a wide range of prices, from as low as €11 to as high as €91,466.
  + New York follows with an average price of $232.89, but with extreme variability, as the maximum price reaches $100,000.
  + London has a lower average price (£197.15) but also shows a significant spread in prices, with a maximum of £80,000.
  + San Francisco has an average price of $239.71, while the prices range from $25 to $50,000.
  + Toronto has the lowest maximum price ($12,400) with a moderate average price of $196.53.
  + **Key Takeaway:** Prices vary widely, with Amsterdam and New York exhibiting the highest levels of price variability, potentially driven by a few extremely high-priced listings.
* **Minimum and Maximum Nights Insights:**
  + New York has the highest average minimum nights (29 days), with 75% of listings requiring at least 30 nights, likely reflecting more long-term rental tendencies.
  + San Francisco also has a high minimum nights average (21 days), with a significant portion of listings requiring a minimum of 30 nights.
  + In contrast, Amsterdam, London, and Toronto have lower average minimum nights (4 to 5 days), suggesting more flexibility for short-term stays.
  + Maximum nights statistics show that New York listings have an average maximum nights stay of 579, with a maximum of 3,000 days, while Toronto has the highest maximum of 3,650 days (around 10 ye**ars).**
  + **Key Takeaway:** New York and San Francisco are more geared towards longer-term rentals, while Amsterdam, London, and Toronto offer more short-term rental flexibility.
* **Number of Reviews Insights:**
  + Amsterdam has the highest average number of reviews (45.64), indicating that listings in Amsterdam may have a more active user base or higher demand for reviews.
  + San Francisco also shows a high number of reviews on average (50.35), with a maximum of 1,239 reviews.
  + London, New York, and Toronto have fewer reviews on average, with New York’s median number of reviews being just 3, suggesting a higher proportion of new or less frequently reviewed listings.
  + **Key Takeaway:** Amsterdam and San Francisco listings are more frequently reviewed, which may imply higher occupancy rates or a more engaged user base.
* **Review Ratings Insights:**
  + Amsterdam has the highest average review score (4.84 out of 5), indicating high satisfaction among guests.
  + San Francisco and Toronto also show high satisfaction with average ratings around 4.78-4.79.
  + New York and London have slightly lower average ratings (4.73 and 4.69, respectively), but still exhibit strong satisfaction levels.
  + Across all cities, the 75th percentile review score is 5.00, indicating that many listings achieve perfect scores.
  + **Key Takeaway:** Guest satisfaction is high across all cities, with Amsterdam and San Francisco standing out slightly for higher average ratings.
* **Overall Summary:**
  + Price Variability: Amsterdam and New York have the most extreme price ranges, with a few very high-priced listings.
  + Stay Duration: New York and San Francisco cater more to long-term rentals, while the other cities are more flexible with short-term stays.
  + Review Activity: Amsterdam and San Francisco listings tend to receive more reviews on average.
  + Guest Satisfaction: Review scores are high across the board, with Amsterdam leading in average guest satisfaction.
  + These insights can help in understanding the market dynamics of short-term rentals in each city, particularly regarding pricing, duration of stay requirements, and customer satisfaction levels.

1. **Histograms and Density Plots:** We used histograms and density plots to visualize the results obtained during after performing numerical statistics before outlier detection
2. **Outlier Detection Methods:** 
   1. **Interquartile Range (IQR)**: We calculated the IQR to capture the middle 50% of the data (25th to 75th percentile). This allowed us to identify and visualize potential outliers falling outside this range. This technique however eliminates most of the records and will not help us produce accurate results
   2. **Z-Score Technique**: The Z-score method examines each data point to determine how many standard deviations it deviates from the mean. Data points with Z-scores between -3 and 3 are considered normal, while those outside this range are flagged as outliers. This technique identifies only few records as outliers leaving behind a huge bunch of them
   3. **Binning Technique**: The data was divided into bins based on key percentiles (0th, 25th, 50th, 75th, and max values), derived from the numerical statistics. Based on the counts in each bin, we further refined the binning ranges, allowing for more precise outlier detection and elimination. This required manual intervention and hence could not derive a generalized formula
   4. Data between 5th to 95th percentile: We calculated the lower and upper bound using the 5th and 95th percentile and eliminated records which were below and above that range
   5. **Iterative Refinement**: Based on the counts in each bin, we further refined the binning ranges, allowing for more precise outlier detection and elimination.
      * Price Insights:
        + Amsterdam: The price range is between 92.4 and 585.6. There are 5,281 data points within this range, with prices such as 269.0, 254.0, 203.0, and 375.0 being some examples.
        + New York: The price range is broader, from 44 to 614. A larger number of data points (20,538) fit within this range, with prices like 89.0, 45.0, 107.0, 340.0, and more.
        + London: The price range is 40 to 481, with 57,482 data points fitting within this. Example prices include 59.0, 120.0, 190.0, and 140.0.
        + San Francisco: The price range here is from 55 to 576.2, with 5,557 data points. Prices such as 132.0, 78.0, and 85.0 are within the range.
        + Toronto: The price range is between 45 and 499, with 14,942 data points. Example prices include 172.0, 75.0, 79.0, 126.0, and more.
      * Minimum and Maximum Nights Insights:
        + Amsterdam: For minimum nights, the range is 1 to 7, with 9,279 data points (examples: 2, 3, 4). For maximum nights, the range is 5 to 1,125, with 8,956 data points (examples: 1125, 365).
        + New York: The minimum night range is 1 to 31, with 36,512 data points (examples: 31, 30, 2). The maximum nights range is 30 to 1,125, with 35,114 data points (examples: 33, 365, 1125).
        + London: Minimum nights range is 1 to 14, with 91,578 data points (examples: 1, 2, 3). The maximum nights range is 7 to 1,125, with 89,055 data points (examples: 365, 1125).
        + San Francisco: The minimum night range is 1 to 31, with 7,525 data points (examples: 30, 3, 1). The maximum night range is 10 to 1,125, with 7,181 data points (examples: 365, 1125).
        + Toronto: Minimum nights range is 1 to 31, with 20,908 data points (examples: 28, 30, 1). Maximum nights range is 27 to 1,125, with 20,033 data points (examples: 730, 365, 1100).
      * Review Scores Insights:

* Amsterdam: The review score range is 4.42 to 5.0, with 8,206 data points (examples: 4.84, 4.71, 5.00).
* New York: The review score range is 4.0 to 5.0, with 25,275 data points (examples: 4.83, 4.00, 5.00).
* London: The review score range is 4.0 to 5.0, with 68,801 data points (examples: 4.82, 4.77, 5.00).
* San Francisco: The review score range is 4.106 to 5.0, with 5,731 data points (examples: 5.00, 4.62, 4.87).
* Toronto: The review score range is 4.17 to 5.0, with 15,792 data points (examples: 5.00, 4.84, 4.79).
  + - Key Takeaways:
* New York has the widest range for prices, indicating a diverse set of accommodations.
* London and Toronto have a large number of data points for both prices and nights, suggesting a significant Airbnb presence in these cities.
* Review scores are generally high across all cities, with most cities having upper bounds of 5.0, indicating overall customer satisfaction.
* San Francisco has tighter bounds for review scores compared to other cities, with a minimum bound starting slightly higher at 4.106.
* This provides a high-level view of Airbnb listings' price, stay durations, and review scores across major cities.

1. **Numerical Statistic After Outlier Detection:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **Median** | **75%** | **Max** |
| Amsterdam | 5281 | 249.53 | 104.87 | 93 | 171 | 225 | 300 | 585 |
| New York | 20538 | 181.31 | 123.5 | 44 | 90 | 147 | 238 | 614 |
| London | 57482 | 152.95 | 93.44 | 40 | 80 | 130 | 200 | 481 |
| San Francisco | 5557 | 180.26 | 104.93 | 55 | 105 | 150 | 229 | 575 |
| Toronto | 14942 | 165.89 | 100.87 | 45 | 87 | 139 | 221 | 499 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Metric** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **Median** | **75%** | **Max** |
| Amsterdam | Minimum Nights | 18235 | 2.78 | 1.44 | 1 | 2 | 3 | 3 | 7 |
|  | Maximum Nights | 18235 | 328.57 | 421.66 | 1 | 20 | 31 | 365 | 1125 |
|  | Number of Reviews | 18235 | 47.17 | 121.29 | 0 | 3 | 10 | 33 | 4130 |
| New York | Minimum Nights | 71626 | 26.51 | 9.32 | 1 | 30 | 30 | 30 | 31 |
|  | Maximum Nights | 71626 | 591.44 | 452.57 | 1 | 180 | 365 | 1125 | 3000 |
|  | Number of Reviews | 71626 | 24.75 | 59.36 | 0 | 0 | 3 | 21 | 1941 |
| London | Minimum Nights | 180633 | 2.79 | 2.35 | 1 | 1 | 2 | 3 | 14 |
|  | Maximum Nights | 180633 | 479.34 | 432.9 | 1 | 60 | 365 | 1125 | 2000 |
|  | Number of Reviews | 180633 | 20.15 | 47.43 | 0 | 1 | 5 | 19 | 1682 |
| San Francisco | Minimum Nights | 14706 | 14.18 | 13.91 | 1 | 2 | 3 | 30 | 31 |
|  | Maximum Nights | 14706 | 388.75 | 409.14 | 1 | 30 | 360 | 365 | 3000 |
|  | Number of Reviews | 14706 | 51.19 | 106.55 | 0 | 1 | 8 | 48 | 1239 |
| Toronto | Minimum Nights | 40941 | 19.42 | 12.47 | 1 | 2 | 28 | 28 | 31 |
|  | Maximum Nights | 40941 | 472.66 | 413.09 | 1 | 100 | 365 | 1125 | 3000 |
|  | Number of Reviews | 40941 | 26.09 | 54.35 | 0 | 1 | 6 | 27 | 1116 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **Median** | **75%** | **Max** |
| Amsterdam | 8206 | 4.88 | 0.14 | 4.42 | 4.8 | 4.93 | 5 | 5 |
| New York | 25275 | 4.78 | 0.26 | 4 | 4.67 | 4.86 | 5 | 5 |
| London | 68801 | 4.76 | 0.28 | 4 | 4.63 | 4.85 | 5 | 5 |
| San Francisco | 5731 | 4.86 | 0.18 | 4.11 | 4.79 | 4.92 | 5 | 5 |
| Toronto | 15792 | 4.85 | 0.18 | 4.17 | 4.75 | 4.91 | 5 | 5 |

* Pricing Insights
* Average Prices:
  1. Amsterdam has the highest average price per stay ($249.53), followed by New York ($181.31) and San Francisco ($180.26).
  2. London has the lowest average price ($152.95), while Toronto is slightly higher at $165.89.
* Price Range:
  1. The maximum price in Amsterdam ($585) is significantly higher than in other cities, particularly compared to London ($481) and New York ($614).
  2. Toronto has a maximum price of $499, indicating a competitive price range among major cities.
* Minimum and Maximum Nights Insights
* Minimum Nights:
  + 1. New York has the highest average minimum nights required (26.51), which may indicate stricter rental policies or demand for longer stays.
    2. Amsterdam, London, and San Francisco have relatively low minimum nights (around 2-3), making them more accessible for shorter stays.
  + Maximum Nights:
    1. New York also has the highest average maximum nights (591.44), suggesting a higher capacity for long-term stays compared to others.
    2. Amsterdam has a much lower average maximum nights (328.57), indicating that properties might be less geared toward long-term stays.
* Review Metrics Insights
  + Number of Reviews:
    1. London has the highest count of reviews (180,633), which may suggest a higher level of engagement or activity in the rental market, potentially indicating a greater number of listings or tourist activity.
    2. Amsterdam (18,235 reviews) and Toronto (40,941 reviews) also show significant engagement but less than London.
* Review Scores:
  + Amsterdam boasts the highest average review score (4.88), suggesting a high level of satisfaction among guests.
  + New York has the lowest average rating (4.78), while London follows closely with 4.76. These scores indicate that while all cities maintain high ratings, Amsterdam stands out for customer satisfaction.
* Summary of Insights
  + Amsterdam is the most expensive but also has high customer satisfaction. Its policies seem favorable for shorter stays.
  + New York demands longer stays and has a high maximum capacity but slightly lower customer satisfaction.
  + London shows robust activity in terms of reviews but ranks lower in customer satisfaction than Amsterdam.
  + San Francisco has similar pricing to New York, with a decent number of reviews but lower average review scores than Amsterdam.
  + Toronto shows competitive pricing with a moderate level of customer satisfaction and engagement.
* Recommendations
  + Property Managers: Consider strategies to enhance customer satisfaction, especially in cities like New York and London, where reviews are lower compared to Amsterdam.
  + Market Analysts: Focus on the unique market dynamics in each city, as the pricing and review patterns differ significantly, which could impact investment decisions or marketing strategies.

1. **Histograms and Density Plots:** We used histograms and density plots to visualize the results obtained during after performing numerical statistics after outlier detection

**Step 4: Task 3**

1. Correlation Analysis:

* **Amsterdam:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Amsterdam** | **Price** | **Accommodates** | **Bedrooms** | **Number of Reviews** | **Availability 365** | **Review Scores Rating** |
| **Price** | 1 | 0.450909 | 0.451183 | -0.242996 | 0.034868 | 0.135983 |
| **Accommodates** | 0.450909 | 1 | 0.682794 | -0.059211 | -0.025003 | -0.073001 |
| **Bedrooms** | 0.451183 | 0.682794 | 1 | -0.157943 | -0.106543 | 0.040667 |
| **Number of Reviews** | -0.242996 | -0.059211 | -0.157943 | 1 | 0.115539 | -0.044641 |
| **Availability 365** | 0.034868 | -0.025003 | -0.106543 | 0.115539 | 1 | -0.105473 |
| **Review Scores Rating** | 0.135983 | -0.073001 | 0.040667 | -0.044641 | -0.105473 | 1 |

* **Price vs. Accommodates & Bedrooms**: Moderate positive correlations (0.450 and 0.451) suggest that higher-priced listings tend to have more capacity for guests and more bedrooms.
* **Price vs. Number of Reviews**: A negative correlation (-0.243) indicates that higher-priced listings may receive fewer reviews, potentially reflecting a premium offering or niche market.
* **Accommodates vs. Bedrooms**: Strong correlation (0.683) indicates that listings that accommodate more guests generally have more bedrooms.
* **Number of Reviews**: Weak positive correlations with availability and review scores show limited influence on these metrics.

**New York**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **New York** | **Price** | **Accommodates** | **Bedrooms** | **Number of Reviews** | **Availability 365** | **Review Scores Rating** |
| **Price** | 1 | 0.401678 | 0.185254 | -0.033578 | 0.011818 | 0.022942 |
| **Accommodates** | 0.401678 | 1 | 0.488343 | 0.016404 | 0.0473 | -0.014503 |
| **Bedrooms** | 0.185254 | 0.488343 | 1 | -0.044114 | 0.052876 | 0.021741 |
| **Number of Reviews** | -0.033578 | 0.016404 | -0.044114 | 1 | -0.091458 | 0.072918 |
| **Availability 365** | 0.011818 | 0.0473 | 0.052876 | -0.091458 | 1 | -0.082676 |
| **Review Scores Rating** | 0.022942 | -0.014503 | 0.021741 | 0.072918 | -0.082676 | 1 |

* **Price vs. Accommodates**: Moderate positive correlation (0.402) suggests that as the price increases, the accommodation capacity slightly increases.
* **Bedrooms**: Weaker positive correlation with price (0.185) and accommodates (0.488) suggests that the number of bedrooms has some relationship with price but is less significant.
* **Number of Reviews**: Weak negative correlations with price (-0.034) and accommodates (0.016) suggest minimal impact on pricing.
* **Overall**: Limited correlations with availability and review scores indicate they do not significantly impact pricing.

**London**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **London** | **Price** | **Accommodates** | **Bedrooms** | **Number of Reviews** | **Availability 365** | **Review Scores Rating** |
| **Price** | 1 | 0.580305 | 0.495223 | -0.108523 | 0.050362 | 0.019374 |
| **Accommodates** | 0.580305 | 1 | 0.7344 | -0.066117 | 0.021613 | -0.061507 |
| **Bedrooms** | 0.495223 | 0.7344 | 1 | -0.087928 | 0.00885 | 0.003285 |
| **Number of Reviews** | -0.108523 | -0.066117 | -0.087928 | 1 | -0.058458 | 0.074378 |
| **Availability 365** | 0.050362 | 0.021613 | 0.00885 | -0.058458 | 1 | -0.082454 |
| **Review Scores Rating** | 0.019374 | -0.061507 | 0.003285 | 0.074378 | -0.082454 | 1 |

* **Price vs. Accommodates**: Strong correlation (0.580) suggests that higher-priced listings generally accommodate more guests.
* **Bedrooms**: Moderate correlation with price (0.495) indicates a relationship between the number of bedrooms and price.
* **Negative Correlation with Reviews**: Number of reviews shows a weak negative correlation (-0.109) with price, indicating a potential trend where higher-priced listings receive fewer reviews.
* **Availability**: Weak correlations with other metrics show little impact on pricing decisions.

**San Francisco**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **San Francisco** | **Price** | **Accommodates** | **Bedrooms** | **Number of Reviews** | **Availability 365** | **Review Scores Rating** |
| **Price** | 1 | 0.572651 | 0.467289 | -0.086589 | -0.060538 | 0.171285 |
| **Accommodates** | 0.572651 | 1 | 0.701108 | -0.019202 | -0.048932 | 0.084917 |
| **Bedrooms** | 0.467289 | 0.701108 | 1 | -0.05426 | -0.017921 | 0.084476 |
| **Number of Reviews** | -0.086589 | -0.019202 | -0.05426 | 1 | -0.124058 | 0.079608 |
| **Availability 365** | -0.060538 | -0.048932 | -0.017921 | -0.124058 | 1 | -0.166664 |
| **Review Scores Rating** | 0.171285 | 0.084917 | 0.084476 | 0.079608 | -0.166664 | 1 |

* **Price vs. Accommodates**: Strong correlation (0.573) indicates that higher-priced listings often accommodate more guests.
* **Bedrooms**: Moderate correlation (0.467) suggests more bedrooms correlate with higher prices.
* **Number of Reviews**: Weak negative correlation (-0.087) shows limited relationship with price, with stronger negative correlation with availability (-0.124).
* **Review Scores**: Shows a stronger positive correlation with price (0.171), suggesting that better-rated properties are generally priced higher.

**Toronto**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Toronto** | **Price** | **Accommodates** | **Bedrooms** | **Number of Reviews** | **Availability 365** | **Review Scores Rating** |
| **Price** | 1 | 0.571784 | 0.460141 | 0.011596 | -0.030378 | 0.088351 |
| **Accommodates** | 0.571784 | 1 | 0.746255 | -0.026431 | 0.00664 | 0.01597 |
| **Bedrooms** | 0.460141 | 0.746255 | 1 | -0.079763 | 0.036983 | 0.013706 |
| **Number of Reviews** | 0.011596 | -0.026431 | -0.079763 | 1 | -0.117975 | 0.059333 |
| **Availability 365** | -0.030378 | 0.00664 | 0.036983 | -0.117975 | 1 | -0.055746 |
| **Review Scores Rating** | 0.088351 | 0.01597 | 0.013706 | 0.059333 | -0.055746 | 1 |

* **Price vs. Accommodates & Bedrooms**: Moderate positive correlations (0.572 and 0.460) suggest that higher-priced listings tend to accommodate more guests and have more bedrooms.
* **Number of Reviews**: Shows a weak positive correlation (0.012) with price, indicating minimal influence.
* **Availability**: Negative correlation with number of reviews (-0.118) suggests that more available listings may receive fewer reviews, reflecting less market demand.
* **Overall Insights**
  + **Strong Relationships**: Across all cities, price consistently correlates positively with the number of accommodations and bedrooms, indicating larger, more spacious listings tend to command higher prices.
  + **Weak Negative Correlations with Reviews**: Many cities exhibit weak negative correlations between price and number of reviews, suggesting that premium listings might not engage with reviewers as frequently, possibly due to their higher pricing or niche appeal.
  + **Influence of Availability**: The availability of listings often has weak to moderate correlations with other metrics, indicating that while availability may affect the market dynamics, its direct impact on pricing is limited.
* These insights can guide stakeholders in understanding the rental landscape in each city and may inform pricing strategies, marketing approaches, and inventory management decisions.

**Step 4: Task 4**

1. Similar Trends across Cities:

· Neighborhood Dynamics: Rental prices vary significantly by neighborhood, influenced by proximity to amenities, transportation, and desirability. Established, wealthier areas tend to have higher rates than less developed ones.

· Stay Duration Flexibility: Entire homes and shared rooms are better suited for long stays, while hotels and private rooms cater to short visits. Each accommodation type serves different traveler needs, such as leisure or work.

· Traveler Preferences: Longer stays attract travelers to homes or shared spaces for privacy and comfort, while shorter trips favor hotels or private rooms for convenience and location.

· Market Trends: Short-term stays dominate most markets, though long-term stays hold importance in areas with transient populations or remote workers.

2. Contrasting Trends across Cities:

· Price Variation: Price differences across neighborhoods vary widely between cities, depending on factors like living costs and tourism demand.

· Stay Type Dominance: Cities differ in the prevalence of long- or short-term stays, with places like New York seeing more long-term visitors, while others like Amsterdam or London cater to short-term tourism.

· Regulatory Impact: Local regulations affect the rental market by setting stay limits and turnover rates, influencing hosts' behavior and traveler options.

3. Implications for Stakeholders:

· Travelers: Awareness of price variations and stay options helps travelers choose suitable accommodations based on budget and length of stay.

· Hosts: Hosts can optimize bookings by adjusting pricing and stay limits to match neighborhood trends and local demand, staying mindful of regulations.

· Market Analysts: Analysts can assess rental trends, regulations, and traveler preferences to predict market changes, uncover risks, and identify growth opportunities.

**Step 5: Task 5**

Amsterdam

* Key Insights:

1. Top-Rated Areas: Neighborhoods like CIT 6 DOWNT, Midtown, and New Amsterda have consistently high review scores with little variability, indicating reliable, high-quality experiences.
2. Low-Rated Areas: SOC SFERM and Mission Area South have the lowest ratings with high variability, showing inconsistent services and user dissatisfaction.
3. High Variability: Neighborhoods like Sunset Park and Bay Ridge show a broad range of review scores, reflecting inconsistent experiences.
4. Outliers: Some areas, such as Upper West in Lenox Hill, have outliers, indicating some extremely different experiences compared to the norm.

* In summary, high-rated areas offer consistent quality, while low-rated ones show more variability and need improvement.
* The ANOVA test result shows that the F-statistic is 1.397 with a p-value of 0.021. Since the p-value is less than 0.05, the differences in ratings across neighborhoods are statistically significant. This suggests that not all neighborhoods have similar ratings, and there are meaningful variations in guest satisfaction depending on the neighborhood.

New York

* **Key Insights:**

1. **Top-Rated Neighborhoods:** Neighborhoods like **Zonne / Bosrijk** and **CIT 6 DOWNT** are at the top with the highest review scores, indicating high satisfaction levels.
2. **Lowest-Rated Neighborhoods:** Areas like **SOC SFERM** and **Mission Area South** are ranked lowest with scores trending toward red, reflecting dissatisfaction or poor experiences.
3. **Middle-Rated Neighborhoods:** A large number of neighborhoods fall in the middle range (gray/neutral shades), suggesting average experiences, neither extremely positive nor negative.

* In summary, top neighborhoods maintain high quality and consistent reviews, while the lowest-rated ones need improvement. The majority of areas sit in the average review range.
* The ANOVA test results indicate a statistically significant difference in ratings across neighborhoods, with an F-statistic of 1.962 and a p-value of 0.000 (p < 0.05).

London

* **Key Insights:**

1. **High Ratings**: The top 5 neighborhoods, all rated 5.0, show excellent guest experiences, indicating strong property management or appealing surroundings.
2. **Low Ratings Still Good**: The bottom 5 neighborhoods, with 4.0 ratings, still offer favorable experiences but could improve on factors like amenities or cleanliness.
3. **Narrow Rating Range**: A small gap between the highest (5.0) and lowest (4.0) ratings suggests that overall, London neighborhoods offer a consistently high standard of accommodation.
4. **Improvement Potential**: Lower-rated areas can boost ratings by learning from the top-rated neighborhoods’ success.

* The ANOVA test results show a statistically significant difference in ratings across neighborhoods, with an F-statistic of 4.382 and a p-value of 0.000 (p < 0.05).

San Francisco

* **Key Insights:**
  + **High Satisfaction**:
    1. The top five neighborhoods, with perfect ratings of 5.0, indicate exceptional guest experiences, likely due to strong amenities and services.
  + **Areas for Improvement**:
    1. The bottom five neighborhoods, with ratings as low as 4.50, suggest potential dissatisfaction. Factors like service quality or location may contribute to these lower ratings.
  + **Strategic Recommendations**:
    1. **Leverage High Ratings**: Market high-performing neighborhoods to attract more guests.
    2. **Target Improvements**: Invest in enhancements for lower-rated neighborhoods based on guest feedback.
    3. **Cross-Promotion**: Encourage exploration of lower-rated areas by promoting events in top-rated neighborhoods.
* By capitalizing on strengths and addressing weaknesses, stakeholders can enhance guest experiences and improve overall neighborhood ratings.
* The ANOVA test results indicate a statistically significant difference in ratings across neighborhoods, with an F-statistic of 2.705 and a p-value of 0.000 (p < 0.05).

Toronto

* Key Insights

1. **High Satisfaction**: Top neighborhoods like Puntarenas residence and Venice, with perfect ratings of 5.0, show exceptional guest experiences, likely due to superior amenities and attractions.
2. **Improvement Needed**: Lower-rated neighborhoods, like Sunnybrook (4.20), may need enhancements in services or facilities to boost guest satisfaction.
3. **Strategic Opportunities**: High-performing areas can be leveraged in marketing, while investment in underperforming neighborhoods can drive improvements.

* As Abraham Lincoln said, “The best way to predict the future is to create it.” By addressing weaknesses and enhancing strengths, stakeholders can create better guest experiences and improve ratings across Toronto.
* The ANOVA test results show a statistically significant difference in ratings across neighborhoods, with an F-statistic of 1.849 and a p-value of 0.000 (p < 0.05).

**Step 7: Task 7**

**Create a new feature that measures the length of each review (number of words or characters). Determine if the length of a review correlates with its sentiment or the review scores.**

**Amsterdam-**

1: Sentiment Analysis

* Attributes: comment\_length, sentiment\_polarity, sentiment\_subjectivity
* Key Correlations:
  + Sentiment Polarity & Subjectivity: Strong positive correlation (0.79), indicating that as the sentiment polarity increases (more positive), the subjectivity also tends to increase.
  + Comment Length & Sentiment Polarity: Weak negative correlation (-0.17), suggesting a slight inverse relationship.
  + Comment Length & Sentiment Subjectivity: Very weak negative correlation (-0.0015), indicating almost no relationship.

2: Comment Length and Compound sentiment score

* Attributes: comment\_length, compound
* Key Correlations:
  + Comment Length & Compound: Weak positive correlation (0.062), suggesting a minimal relationship between comment length and the compound sentiment score.

3: Comment Length and Review Scores

* Attributes: comment\_length, review\_scores\_rating
* Key Correlations:
  + Comment Length & Review Scores Rating: Weak positive correlation (0.1), indicating a slight tendency for longer comments to have higher review scores.

**New York-**

1: Sentiment Analysis

* Attributes: comment\_length, sentiment\_polarity, sentiment\_subjectivity
* Key Correlations:
  + Sentiment Polarity & Subjectivity: Strong positive correlation (0.72), indicating that as sentiment polarity increases, subjectivity also tends to increase.
  + Comment Length & Sentiment Polarity: Moderate negative correlation (-0.28), suggesting that longer comments might be less positive in sentiment.
  + Comment Length & Sentiment Subjectivity: Weak negative correlation (-0.099), indicating a slight inverse relationship.

2: Comment Length and Compound Score

* Attributes: comment\_length, compound
* Key Correlations:
  + Comment Length & Compound: Very weak positive correlation (0.014), suggesting almost no relationship between comment length and the compound sentiment score.

3: Comment Length and Review Scores

* Attributes: comment\_length, review\_scores\_rating
* Key Correlations:
  + Comment Length & Review Scores Rating: Very weak positive correlation (0.038), indicating minimal relationship between comment length and review scores.

**London-**

1: Sentiment Analysis

* **Attributes**: comment\_length, sentiment\_polarity, sentiment\_subjectivity
* **Key Correlations**:
  + **Sentiment Polarity & Subjectivity**: Strong positive correlation (0.76), indicating that as sentiment polarity increases, subjectivity also tends to increase.
  + **Comment Length & Sentiment Polarity**: Moderate negative correlation (-0.25), suggesting that longer comments might have slightly less positive sentiment.
  + **Comment Length & Sentiment Subjectivity**: Weak negative correlation (-0.068), indicating a slight inverse relationship.

2: Comment Length and Compound Score

* **Attributes**: comment\_length, compound
* **Key Correlations**:
  + **Comment Length & Compound**: Very weak positive correlation (0.0099), suggesting almost no relationship between comment length and the compound sentiment score.

3: Comment Length and Review Scores

* **Attributes**: comment\_length, review\_scores\_rating
* **Key Correlations**:
  + **Comment Length & Review Scores Rating**: Very weak positive correlation (0.023), indicating minimal relationship between comment length and review scores.

**San Francisco-**

1: Sentiment Analysis

* Attributes: comment\_length, sentiment\_polarity, sentiment\_subjectivity
* Key Correlations:
  + Sentiment Polarity & Subjectivity: Strong positive correlation (0.7), indicating that as sentiment polarity increases, subjectivity also tends to increase.
  + Comment Length & Sentiment Polarity: Moderate negative correlation (-0.28), suggesting that longer comments might have slightly less positive sentiment.
  + Comment Length & Sentiment Subjectivity: Weak negative correlation (-0.12), indicating a slight inverse relationship.

2: Comment Length and Compound Score

* Attributes: comment\_length, compound
* Key Correlations:
  + Comment Length & Compound: Very weak positive correlation (0.015), suggesting almost no relationship between comment length and the compound sentiment score.

3: Comment Length and Review Scores

* Attributes: comment\_length, review\_scores\_rating
* Key Correlations:
  + Comment Length & Review Scores Rating: Very weak positive correlation (0.048), indicating minimal relationship between comment length and review scores.

**Toronto-**

1: Sentiment Analysis

* Attributes: comment\_length, sentiment\_polarity, sentiment\_subjectivity
* Key Correlations:
  + Sentiment Polarity & Subjectivity: Strong positive correlation (0.73), indicating that as sentiment polarity increases, subjectivity also tends to increase.
  + Comment Length & Sentiment Polarity: Moderate negative correlation (-0.27), suggesting that longer comments might have slightly less positive sentiment.
  + Comment Length & Sentiment Subjectivity: Weak negative correlation (-0.089), indicating a slight inverse relationship.

2: Comment Length and Compound Score

* Attributes: comment\_length, compound
* Key Correlations:
  + Comment Length & Compound: Very weak positive correlation (0.0086), suggesting almost no relationship between comment length and the compound sentiment score.

3: Comment Length and Review Scores

* Attributes: comment\_length, review\_scores\_rating
* Key Correlations:
  + Comment Length & Review Scores Rating: Very weak positive correlation (0.013), indicating minimal relationship between comment length and review scores.

Summary

* Sentiment Polarity & Subjectivity: Amsterdam shows the strongest correlation (0.79), indicating a strong link between these two attributes across all cities.
* Comment Length & Sentiment Polarity: The correlation is moderately negative across all cities, with New York and San Francisco showing the strongest negative correlation (-0.28).
* Comment Length & Sentiment Subjectivity: Very weak negative correlations are observed across all cities, indicating minimal relationships.
* Comment Length & Compound Score: All cities show very weak positive correlations, suggesting almost no relationship.
* Comment Length & Review Scores Rating: Amsterdam shows a slightly stronger positive correlation (0.1) compared to other cities, which have very weak correlations.

**Step 8: Task 8**

**Identify outliers in the dataset, particularly in price, minimum\_nights, and review\_scores\_rating.**

Keywords and Sentiment Impact

* **Clean**
  + Absent Score: 0.60
  + Present Score: 0.75
  + Impact: +0.15
* **Comfortable**
  + Absent Score: 0.55
  + Present Score: 0.80
  + Impact: +0.25
* **Noisy**
  + Absent Score: 0.45
  + Present Score: 0.60
  + Impact: +0.15
* **Spacious**
  + Absent Score: 0.60
  + Present Score: 0.70
  + Impact: +0.10
* **Great**
  + Absent Score: 0.65
  + Present Score: 0.70
  + Impact: +0.05
* **Highly**
  + Absent Score: 0.55
  + Present Score: 0.72
  + Impact: +0.17
* **Recommend**
  + Absent Score: 0.60
  + Present Score: 0.69
  + Impact: +0.09
* **Friendly**
  + Absent Score: 0.50
  + Present Score: 0.65
  + Impact: +0.15
* **Problem**
  + Absent Score: 0.40
  + Present Score: 0.70
  + Impact: +0.30
* **Issue**
  + Absent Score: 0.50
  + Present Score: 0.68
  + Impact: +0.18

Key Observations

* The presence of positive words like "comfortable," "clean," and "friendly" significantly increases the sentiment score.
* Negative words like "problem" also show a notable increase, possibly indicating that context or resolution of problems might affect sentiment positively.
* The keyword "comfortable" has the highest positive impact on sentiment when present.

This analysis demonstrates that specific keywords can significantly influence sentiment scores, with positive terms generally enhancing perceived positivity more than negative terms reduce it, possibly due to context or resolution narratives in reviews or comments.