**Airbnb Project Outputs:**

**Problem: AirBnB**  
 Use the dataset provided (CSV file) to analyze data and answer the following questions:  
After Basic cleaning of the data set, started doing calculations and predictions for the following questions and visualization.

1. What is the average price of a listing in each borough? Is the difference statistically  
significant?

Ans:

A screen shot of a computer

AI-generated content may be incorrect.

Null Hypothesis: There is no statistically significant difference between the average Airbnb listing prices across all boroughs.

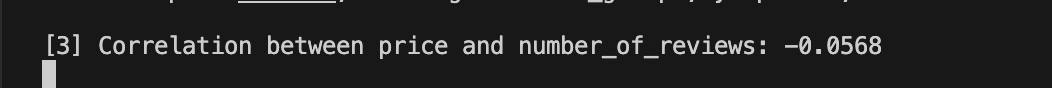
The extremely low p-value (0.000000) provides strong evidence against the null hypothesis. This means that the observed differences in the average Airbnb listing prices between the boroughs are **not due to random chance**. There is a statistically significant difference in the average Airbnb listing prices across the boroughs.

In practical terms, this confirms what the bar chart (In Data visualization sections - Visualization 1) visually suggests: Manhattan has a significantly higher average listing price, and the Bronx has a significantly lower one, with the other boroughs falling in between. The ANOVA test statistically validates that **these differences are real and not just a result of random variation in the data**.

Note: The visualization for each question is added to the Data visualization section separately.

2. Are listings with more reviews cheaper or more expensive on average?

Ans:



Based on the correlation coefficient of -0.0568:

1. **Direction:** There is a **very slight negative linear relationship** between the price of an Airbnb listing and the number of reviews it has. This means that, on average, listings with more reviews tend to have slightly lower prices.
2. **Strength:** The relationship is **extremely weak**. A correlation coefficient of -0.0568 is very close to zero, indicating that number\_of\_reviews explains very little of the variation in price. In practical terms, this correlation is so weak that it suggests that the number of reviews has almost no linear impact on the pricing of an Airbnb listing.

**Does this mean "Listings with more reviews cost less?"**

Technically Yes, the negative sign supports that direction. However, the effect is **negligible** due to the extremely weak strength of the correlation. It's highly unlikely that the number of reviews is a primary driver for a listing's price. Other factors (like location, size, amenities, host type, etc.) would almost certainly have a much stronger influence on price.

Even if the correlation were stronger, correlation does not imply causation. **A negative correlation here would not necessarily mean that having more reviews *causes* a listing to be cheaper.** There could be other underlying reasons (e.g., cheaper listings might naturally attract more bookings and thus more reviews over time, or hosts of cheaper listings might prioritize getting more bookings over maximizing price).

3. Does the average availability vary significantly between room types?

Ans:

A computer screen with numbers and letters

AI-generated content may be incorrect.

Yes, the average availability varies **statistically significantly** between room types. an **ANOVA p-value: 0.00000** for "Average Availability by Room Type" represents that there is **overwhelming statistical evidence** to conclude that the average availability **does vary significantly** between the different room types. The observed differences (e.g., Shared rooms having a notably higher average availability than Entire homes/apts or Private rooms) are not likely due to random chance.

From the averages, it appears that Shared rooms tend to be available for a significantly greater number of days in the year compared to Entire home/apt and Private room listings, which have very similar average availability. The ANOVA test confirms that this difference is statistically meaningful.

4. Which borough has the most variability in price?

Ans:

A black screen with white text

AI-generated content may be incorrect.

Based on the calculated variability measures, **Manhattan has the most variability in price**, with a value of 129.112217.

This means that the prices of Airbnb listings in Manhattan are, on average, spread out more widely from their mean compared to listings in other boroughs. This makes sense intuitively, as Manhattan hosts a wide range of properties, from small studios to luxurious penthouses, leading to a broader distribution of prices.

5. Can we statistically model the price based on a few key predictors?

Ans:

A screen shot of a computer

AI-generated content may be incorrect.

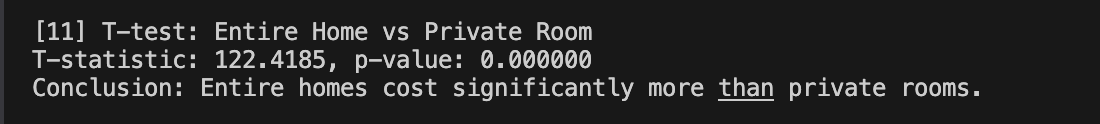
**Yes, we can statistically model the price based on a few key predictors, particularly Room Type and, to a lesser extent, availability\_365.**

While a model including number\_of\_reviews alone is ineffective, a multiple regression model incorporating Room Type and availability\_365 demonstrates a moderate ability to explain and predict price variations (R2 of nearly 25%). This R2 suggests that while these predictors are good, there are still other unexamined factors (like borough, amenities, size, specific neighborhood, host reputation, etc.) that contribute to the remaining 75% of price variability.

Therefore, while not a perfect prediction, this model provides valuable insights into what drives Airbnb prices and indicates that **Room Type is a particularly powerful predictor**.

6. Complete Hypothesis Testing  
Do homes cost significantly more than private rooms? Use t- test. Explain your findings in  
detail.

Ans:



Null Hypothesis: There is no statistically significant difference in the average price between Entire home/apt listings and Private room listings. Or, specifically, the average price of Entire home/apt is not greater than or equal to the average price of Private room.

p-value is **0.000000**, which is much, much smaller than 0.05: 0.000000≤0.05

This means we can **reject the null hypothesis (H0​)**.

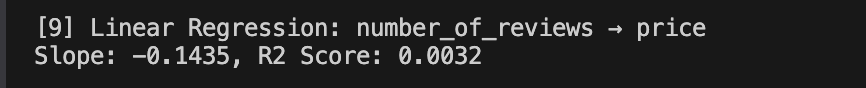
This means there is **overwhelming statistical evidence** to conclude that **"Entire home/apt" listings cost significantly more than "Private room" listings.**

**In summary:**

The t-test result formally confirms what is intuitively expected and visually suggested by the box plot of Room Type vs Price(Visualization 2). "Entire home/apt" listings are indeed priced at a substantially higher average than "Private room" listings in this Airbnb dataset, and this difference is statistically robust. This makes Room Type a very strong predictor of price, as also indicated by its coefficients(R2) in the multiple regression model.

7. Complete Regression Analysis  
a. Predict price from number\_of\_reviews

Ans:



The results of this regression are:

* **Slope**: -0.1435
* **R-squared (R2)**: 0.0032

**Observation and Interpretation**:

This analysis indicates that Price **cannot be accurately predicted** based solely on Number of Reviews.

* The **R-squared value of 0.0032** is extremely low. This means that only 0.32% of the variability in Airbnb listing prices can be explained by the number of reviews a listing has. The vast majority (over 99.6%) of the price variation remains unexplained by this single predictor.
* The **slope of -0.1435** indicates a very slight negative relationship. For every additional review a listing receives, its predicted average price decreases by approximately $0.14 (14 cents). This is a negligible change in the context of typical Airbnb prices, further highlighting the lack of practical significance in this relationship.

In summary, while there is a very weak negative linear trend, **Number of Reviews is an exceptionally poor predictor of Price in this dataset**, as it explains almost no variance in pricing.

b. Predict price from room\_type, availability\_365

Ans:

A black screen with white text

AI-generated content may be incorrect.

The regression results are as follows:

* **R2 (R-squared)**: 0.2498
* **availability\_365 coefficient**: 0.1064
* **room\_type\_Private room coefficient**: -108.3482
* **room\_type\_Shared room coefficient**: -132.2565

**Interpretation for Predicting Price**:

This regression model indicates that room\_type and availability\_365 together can explain approximately **24.98%** of the variability in Airbnb listing prices. This is a moderate level of explanatory power, suggesting that these are significant factors in determining price, though other variables not included in this specific model also influence price.

Interpretation of how each variable contributes to the predicted price:

* **availability\_365**: For every additional day a listing is available throughout the year, the predicted price is expected to increase by approximately $0.11, assuming the room\_type remains constant. This indicates **a positive but relatively small impact on price**.
* **room\_type**: This categorical variable has a substantial impact on price. Assuming "Entire home/apt" is the baseline (which is common for categorical variables in regression when other categories are listed with negative coefficients), the coefficients for the other room types mean:
  + A Private room is predicted to cost approximately **$108.35 less** than an "Entire home/apt", holding availability\_365 constant.
  + A Shared room is predicted to cost approximately **$132.26 less** than an "Entire home/apt", holding availability\_365 constant.

These large negative coefficients for Private room and Shared room demonstrate that **room\_type is a very strong predictor of price**, with "Entire home/apt" listings generally being the most expensive, followed by "Private room" and then "Shared room" listings.

8. Look for outliers:  
There are some very low, and very high prices  
Also, some interesting constraints (ie minimum nights close to 300)  
What do we do with these outliers?

Ans:

Outlier detection and it is handled by removing the mentioned outliers

A black background with white text

AI-generated content may be incorrect.

Outliers are present in the dataset, specifically concerning price and minimum\_nights.

**Observations regarding outliers:**

* There are listings with very low and very high prices.
* Some listings have interesting constraints, such as a minimum nights requirement close to 300.
* Quantitatively, there are 122 listings with a price greater than $800.
* Additionally, 53 listings have a minimum nights requirement of 300 or more.

**What to do with these outliers?** The recommendation is to **filter or transform these outliers before modeling**.

Filtering would involve removing these extreme data points from the dataset. Transformation could involve applying mathematical functions (e.g., logarithmic transformation) to reduce the impact of extreme values. The choice depends on the specific analysis goals and the nature of the outliers, but addressing them is crucial for building more robust and accurate statistical models.

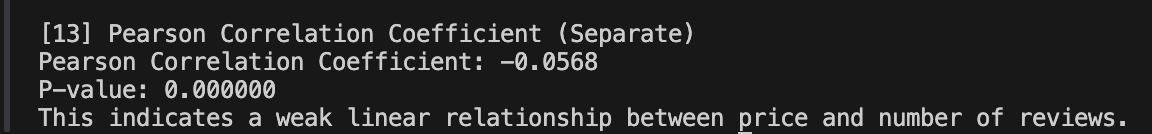
Here we have created a separate cleaned data set after excluding the outliers.

# Created a new cleaned dataset with outliers removed for future analysis

df\_cleaned = df[(df["price"] <= 800) & (df["minimum\_nights"] < 300)]

print(f"Cleaned dataset size: {df\_cleaned.shape[0]} listings (after removing outliers)")

9. Measure linear relationship:  
Use Pearson Correlation Coefficient



**Pearson Correlation Coefficient**:

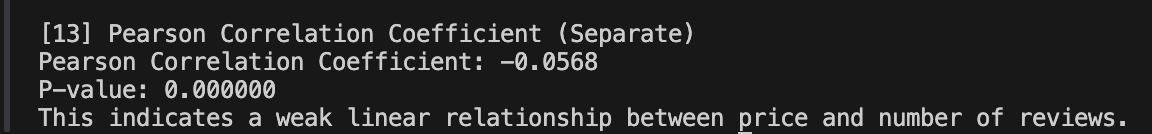
* **Pearson R**: -0.0568
* **p-value**: 0.000000

**Interpretation of the Linear Relationship**: The Pearson correlation coefficient (R) of **-0.0568** indicates an **extremely weak negative linear relationship** between Price and Number of Reviews.

* The negative sign suggests that as the Number of Reviews increases, the Price tends to slightly decrease.
* However, the value itself, being very close to zero (magnitudes below 0.1), signifies that there is **virtually no linear association** between these two variables. This means that Number of Reviews is not a meaningful predictor of Price in a linear fashion.
* The p-value of 0.000000, while statistically significant (less than 0.05), merely indicates that this extremely weak correlation is unlikely to be due to random chance. It does **not** imply a strong or practically important relationship.

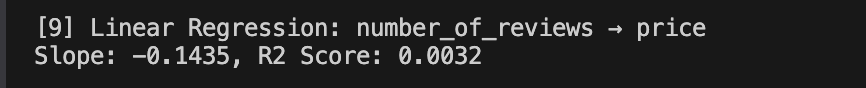
10. Find Relationship between price and number\_of\_reviews.

1. **Correlation Coefficient (Linear Relationship):**



* + The Pearson Correlation for Price vs Number of Reviews is **-0.0568**.
  + This indicates an **extremely weak negative linear relationship** between the two variables. A value close to 0 suggests almost no linear correlation. The negative sign implies that as the number of reviews increases, the price tends to slightly decrease, but this effect is very minor.
  + The p-value associated with this correlation is 0.000000, which means this extremely weak correlation is statistically significant (unlikely to be due to random chance). However, statistical significance does not imply practical significance or a strong relationship.

1. **Regression Analysis (Predictive Power):**



* + A regression model to predict Price from Number of Reviews yields a Slope of **-0.1435** and an R2 (R-squared) of **0.0032**.
  + The Slope of -0.1435 means that for every additional review a listing receives, its predicted price decreases by approximately $0.14. This is a very small change, confirming the weak practical impact.
  + The R2 value of 0.0032 is exceptionally low. This means that Number of Reviews explains only **0.32%** of the variability in Price. The vast majority (over 99%) of the variation in price is not explained by the number of reviews.

1. **Visual Observation (Scatter Plot – Visualization task 3):**
   * The scatter plot of Price vs Number of Reviews visually confirms this weak relationship. The data points are widely scattered, and while a faint downward trend line is present, it's clear that Number of Reviews does not strongly predict Price. There's a wide range of prices for any given number of reviews.

**Conclusion:** There is an **extremely weak negative linear relationship** between the Price of an Airbnb listing and its Number of Reviews. While a statistical correlation and regression line can be calculated, Number of Reviews has negligible explanatory and predictive power for Price. Therefore, Price **cannot be accurately predicted** based solely on Number of Reviews.

For all questions, please explain your findings in detail.

Answered every question and visualization tasks in detail.

**DATA VISUALIZATION TASKS**  
1. **Borough vs. Average Price**Create a bar chart showing the average listing price per borough. Based on the chart, which  
boroughs have the highest and lowest prices? How might this influence a customer’s decision?

Ans:

A graph of a number of blue bars

AI-generated content may be incorrect.

**Bar chart analysis:**

* **Highest Price:** Based on the bar chart, **Manhattan** has the highest average daily Airbnb listing price, appearing to be around **$180**.
* **Lowest Price:** The **Bronx** has the lowest average daily Airbnb listing price, appearing to be around **$85**.
* Brooklyn appears to be around $120, Queens around $88, and Staten Island around $90.

**Influence on a customer's decision:** The significant difference in average daily listing prices across boroughs would heavily influence an Airbnb customer's decision, primarily based on their budget, desired experience, and travel needs:

* **Budget:** Customers on a tighter budget would gravitate towards the Bronx, Staten Island, or Queens, while those with more flexibility would consider Brooklyn or Manhattan.
* **Location and Convenience:** Manhattan's higher price reflects its central location, proximity to major attractions, business districts, and extensive public transport. Customers might pay more for the convenience of staying in Manhattan.
* **Type of Trip:** Tourists focused on museums, Broadway, and iconic landmarks might prioritize Manhattan despite the cost. Visitors looking for a more local, neighborhood experience, or those with specific destinations in other boroughs, might find better value and suitability elsewhere.
* **Value for Money:** A customer might weigh whether paying almost double in Manhattan (compared to the Bronx) offers a proportionate increase in amenities, space, or convenience for their specific travel goals.
* **Availability:** In high-demand periods, customers might be forced to consider less expensive boroughs if Manhattan options are limited or exorbitantly priced.

2. **Room Type vs. Price**Create a box plot comparing the distribution of prices across different room types. What does  
this tell you about the pricing spread and outliers in each room category?

Ans:

A graph of different types of rooms

AI-generated content may be incorrect.

**Box Plot Analysis: Price Distribution by Room Type**

A box plot effectively summarizes the distribution of a dataset using five key numbers: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum (within 1.5 IQR of the quartiles), along with displaying outliers.

**1. Private room:**

* **Median (thick line inside the box):** Around $70-$80. This indicates that half of all private room listings are priced below this amount.
* **Pricing Spread (Box Height / IQR):** The box is relatively narrow, indicating that the middle 50% of private room prices are fairly concentrated, likely within a range of about $50 to $100.
* **Whiskers:** Extend upwards, showing a range of prices beyond the interquartile range.
* **Outliers (Individual points):** There is a **very large number of outliers** extending significantly upwards, reaching prices of $1000 or more. This tells us that while most private rooms are affordably priced, there's a substantial segment of highly-priced private rooms that are considered outliers.

**2. Entire home/apt:**

* **Median (thick line inside the box):** Significantly higher than private or shared rooms, around $150.
* **Pricing Spread (Box Height / IQR):** The **box for "Entire home/apt" is considerably taller/wider** than the other two room types. This indicates a much larger spread in the middle 50% of prices for entire homes/apartments (likely ranging from around $120 to over $200). This category has the broadest typical price range.
* **Whiskers:** Extend upwards, covering a wider range than private rooms.
* **Outliers (Individual points):** Similar to private rooms, there is a **very large number of outliers** extending upwards to $1000 and beyond. This suggests that while many entire homes/apts are in the mid-range, there are numerous luxury or premium listings that significantly increase the upper end of the price spectrum. This category clearly has the largest spread of prices overall.

**3. Shared room:**

* **Median (thick line inside the box):** The lowest median among all room types, around $40-$50.
* **Pricing Spread (Box Height / IQR):** The box is **the narrowest**, indicating the tightest concentration of prices for the middle 50% of shared rooms (likely between $30 and $60).
* **Whiskers:** Relatively short, showing a smaller typical range.
* **Outliers (Individual points):** There are outliers, but **fewer in number and generally not extending as high** as those seen in "Private room" or "Entire home/apt" categories (most are below $500, with a few reaching up to $800). This suggests that while there are some more expensive shared room options, the extreme high-end outliers are less frequent and less extreme compared to the other categories.

**Interpretation of Data about the pricing spread and outliers in each room category**

1. **Pricing Spread:**
   * **Entire home/apt** has the **largest pricing spread** for its core prices (middle 50%), reflecting a wide variety of properties and amenities available in this category.
   * **Private room** has a moderate core pricing spread.
   * **Shared room** has the **tightest pricing spread** for its core, indicating a more consistent and generally lower price point.
2. **Outliers:**
   * **Entire home/apt** and **Private room** categories both exhibit a **significant presence of high-priced outliers**, extending to $1000 and beyond. This indicates that while the majority of listings in these categories fall within a certain price range, there's a substantial tail of very expensive, premium, or luxury listings.
   * **Shared room** also has outliers, but they are **less numerous and generally do not reach the extremely high price points** seen in the other two categories. This suggests a more uniform upper limit to what people are willing to pay for a shared space.

In summary, the box plot clearly illustrates that **"Entire home/apt" listings generally command the highest prices and exhibit the greatest price variability and the largest number of high-end outliers**. "Private room" listings are a step down in median price but still have many high-priced outliers. "Shared room" listings are the most affordable and have the least price variability, with fewer extreme outliers. This reflects the different value propositions and target markets for each room type on Airbnb.

3. **Reviews vs. Price**  
Create a scatter plot of number\_of\_reviews vs price. Add a regression trend line. Do you  
observe any trend between price and number of reviews?

Ans:

A graph with a red line

AI-generated content may be incorrect.

**1. Visualizing the Trend Line:** The red line on the scatter plot is the regression trend line. This line visually represents the general direction of the relationship between Price and Number of Reviews.

* **Observation:** The trend line slopes downwards from left to right. The downward slope of the regression line indicates that as the Number of Reviews increases, the Price tends to slightly decrease.

**2. Interpreting the Data Distribution (Scatter of Points):** The individual points on the scatter plot represent each Airbnb listing.

* Most of the data points are clustered towards the lower end of the "Number of Reviews" axis (0 to about 100-150 reviews) and the lower end of the "Price ($)" axis (0 to about $200-$300). This indicates that many listings have relatively few reviews and lower prices.
* As the Number of Reviews increases, the density of points generally decreases, especially for higher prices.
* There are some outliers, particularly listings with very high prices (e.g., $800-$1000 and above) that have very few reviews.

**Do you observe any trend between price and number of reviews?**

Based on the scatter plot and the regression trend line:

Yes, a **weak negative linear trend** is observed between the Price of an Airbnb listing and the Number of Reviews it has received.

**Explanation of the Trend:**

* **Direction (Negative):** The downward slope of the regression line indicates that as the Number of Reviews increases, the Price tends to slightly decrease. In simpler terms, listings with a higher number of reviews generally have a slightly lower average price.
* **Strength (Weak):** Despite the clear downward slope of the line, the individual data points (the scatter) are very widely dispersed around this trend line. They do not tightly cluster around the line. This wide spread signifies a **very weak** relationship. Many listings deviate significantly from the trend. For example, many listings with few reviews that have both very low and very high prices, and listings with many reviews that still command moderate prices.

This visual observation aligns perfectly with the correlation coefficient of -0.0568 that was previously discussed. A correlation near zero, even if negative, points to an extremely weak linear relationship, meaning **number\_of\_reviews is not a strong predictor of price.**

4. **Price Variability by Borough**  
Create a box plot for each borough showing the price distribution. Which borough shows the  
most variability (largest spread or presence of outliers)?

Ans:

A screenshot of a graph

AI-generated content may be incorrect.

**Box Plot Analysis: Price Distribution by Borough**

A box plot visually represents the distribution of data through quartiles and highlights outliers.

* **Box:** Represents the interquartile range (IQR), from the 25th percentile (Q1) to the 75th percentile (Q3). The line inside the box is the median (50th percentile). A wider box indicates a larger spread for the middle 50% of the data.
* **Whiskers:** Extend from the box to the minimum and maximum values within 1.5 times the IQR. The longer the whiskers, the greater the spread of data beyond the central 50%.
* **Outliers (Individual points):** Data points that fall outside the whiskers are considered outliers. A greater number of outliers, especially those far from the main distribution, indicate high variability or extreme values.

1. **Brooklyn:**
   * The box is relatively compact.
   * The whiskers extend a fair bit.
   * There's a significant number of outliers, extending up to $1000.
2. **Manhattan:**
   * The **box itself is notably taller and wider** than all other boroughs, indicating a much larger IQR (spread of the middle 50% of prices).
   * The **whiskers are also longer** compared to most other boroughs.
   * There is an **extremely dense and extensive collection of outliers** extending all the way up to $1000 and beyond. This indicates a very wide range of prices and a large number of listings at the higher end of the spectrum that are considered extreme relative to the bulk of Manhattan prices.
3. **Queens:**
   * The box is relatively narrow and at a lower price range.
   * The whiskers are shorter.
   * There are outliers, but fewer and generally not extending as high as Manhattan or Brooklyn.
4. **Staten Island:**
   * The box is narrow and at a lower price range.
   * The whiskers are shorter.
   * There are some outliers, but relatively few and not extending as high as Manhattan or Brooklyn.
5. **Bronx:**
   * The box is the narrowest and at the lowest price range.
   * The whiskers are short.
   * There are some outliers, but similar to Staten Island, they are fewer and don't extend as high as the most expensive boroughs.

**Conclusion:**

Based on the box plot, **Manhattan** clearly shows the most variability in price.

**Reasons:**

* **Largest Interquartile Range (IQR):** The box for Manhattan is the tallest, indicating the largest spread for the middle 50% of its listing prices.
* **Extensive Range of Values (Whiskers and Outliers):** Manhattan has the longest "whisker" extending upwards (even if truncated by the plot's upper limit) and by far the highest density and furthest-reaching individual outlier points. This means it has a much wider overall range of prices, from very low to extremely high.
* **Presence of Outliers:** While all boroughs have outliers, Manhattan has an exceptionally high number of them, concentrated at very high price points, which contributes significantly to its overall variability.

This visual confirmation from the box plot aligns perfectly with the quantitative measure of variability (standard deviation) discussed earlier in Que 4, where Manhattan also had the highest standard deviation (129.11), indicating the greatest spread of prices.

5. **Room Type vs. Availability**  
Create a violin or box plot showing the distribution of availability (availability\_365) by room  
type. Are certain room types available more consistently throughout the year?

A screenshot of a computer

AI-generated content may be incorrect.

**Violin Plot Analysis: Availability by Room Type**

A violin plot is a hybrid of a box plot and a kernel density plot. It shows:

* **Box plot elements (inside the "violin"):** The thick bar in the middle is the median, the box represents the interquartile range (IQR, 25th to 75th percentile), and the thin line extends to the "whiskers" (usually 1.5 times the IQR from the box edges).
* **Kernel Density Estimation (the "violin" shape):** The width of the shaded area at any given point indicates the density of data points at that availability (days/year) value. A wider section means more listings have that availability.

**1. Private room:**

* **Median (thick bar):** Appears to be very low, close to 0 days.
* **Density (violin shape):** The widest part of the violin is at 0 days, indicating a very high concentration of private rooms are available for 0 days (meaning they are always booked or not available for booking at that time). There's a thinner but significant spread of data points across higher availability values, even reaching up to 400 days, but the density is much lower at these higher values.
* **Consistency:** Primarily inconsistent, with a strong tendency to be not available or available for very few days.

**2. Entire home/apt:**

* **Median (thick bar):** Also appears to be very low, close to 0 days.
* **Density (violin shape):** Similar to private rooms, the widest part of the violin is at 0 days, indicating a very high concentration of entire homes/apartments are available for 0 days. There's a thinner spread of data points across higher availability values, up to 400 days, but again, the density is much lower at these higher values.
* **Consistency:** Primarily inconsistent, with a strong tendency to be not available or available for very few days.

**3. Shared room:**

* **Median (thick bar):** Appears to be significantly higher than private rooms and entire homes/apts, roughly around 100 days or more.
* **Density (violin shape):** While there's still a peak at 0 days (meaning some shared rooms are also not available), the violin shape is much wider and extends to higher availability values with more substantial density. There's a notable concentration of shared rooms available for a moderate number of days (e.g., 50-150 days), and also a significant density at higher availability values, including many listings available for almost the entire year (near 365 days).
* **Consistency:** Shared rooms show a greater tendency to be available more consistently, with a substantial portion of listings having moderate to high availability throughout the year, as well as a noticeable peak around the 365-day mark.

**Are certain room types available more consistently throughout the year?**

Yes, based on the violin plot:

**Shared rooms appear to be available more consistently throughout the year compared to Private rooms and Entire homes/apts.**

**Reasons for this observation:**

* The median availability for Shared rooms is considerably higher.
* The "violin" shape for Shared rooms is much fuller across the higher availability range (e.g., 50-400 days), indicating a greater proportion of these listings are available for longer periods.
* The clear peak in density around 365 days for Shared rooms suggests that many shared room listings are available almost every day of the year, which is less pronounced for the other two room types.
* For "Private room" and "Entire home/apt," the extremely wide base of the violin at 0 days indicates that a very large number of these listings have little to no availability, suggesting they are either short-term rentals that get booked quickly, or they are not available for continuous booking.

This visual analysis aligns with the previous(Que 3) numerical average of 162.21 days for Shared rooms, which was significantly higher than the ~111 days for Private rooms and Entire homes/apts, and the ANOVA result confirmed that these differences are statistically significant.

6. **Identify Outliers**  
Create a histogram or box plot of the price variable. Are there visible outliers? How many  
listings are priced far above or below the median?

Ans:

A graph of a price distribution

AI-generated content may be incorrect.

**1. Is there visible outlier?**

* **Yes, there are clearly visible outliers.**
* The histogram shows a very strong right skew. The vast majority of listings are concentrated at lower prices (below $200-$300).
* However, the distribution extends significantly to the right, with a long "tail" of bars, reaching up to at least $1000. These bars, representing listings at higher prices, are much less frequent but still present.
* The outlier detection analysis also explicitly states that there are 122 listings with a price greater than $800, which are considered outliers.

**2. How many listings are priced above or below the median?**

By definition, the median is the value that divides a dataset into two equal halves. Therefore, approximately **50% of the listings are priced at or below the median**, and approximately **50% of the listings are priced at or above the median**. The histogram visually indicates the median price with a red dashed line, which appears to be around $100-$150, given the high concentration of listings in the lower price bins.

A black background with white text

AI-generated content may be incorrect.  
7. **Bonus: Correlation Heatmap**  
Generate a heatmap showing correlation values between numeric variables (price,  
number\_of\_reviews, minimum\_nights, availability\_365). Which variables are strongly or weakly  
correlated?

Ans:

A screenshot of a computer screen

AI-generated content may be incorrect.

The heatmap displays the Pearson correlation coefficients between the following numeric variables: price, number\_of\_reviews, minimum\_nights, and availability\_365. The color intensity and the correlation value within each cell indicate the strength and direction of the linear relationship.

**1. Price vs. Other Variables:**

* **price vs. number\_of\_reviews:** **-0.06**
  + **Interpretation:** This is an **extremely weak negative correlation**. As the number of reviews increases, the price has a very slight tendency to decrease. This confirms the visual observation from the scatter plot and the previously given correlation value. In practical terms, number\_of\_reviews has almost no linear relationship with price.
* **price vs. minimum\_nights:** **0.02**
  + **Interpretation:** This is an **extremely weak positive correlation**, very close to zero. It suggests virtually no linear relationship between the price of a listing and its minimum required nights.
* **price vs. availability\_365:** **0.12**
  + **Interpretation:** This is a **weak positive correlation**. As the availability (days/year) increases, the price tends to slightly increase. While weak, it's the strongest positive correlation with price among these variables.

**2. Number of Reviews vs. Other Variables:**

* **number\_of\_reviews vs. minimum\_nights:** **-0.08**
  + **Interpretation:** This is an **extremely weak negative correlation**. Listings with more reviews tend to have a very slightly lower minimum nights requirement.
* **number\_of\_reviews vs. availability\_365:** **0.17**
  + **Interpretation:** This is a **weak positive correlation**. Listings with more reviews tend to have slightly higher availability. **This is the strongest correlation observed in the entire heatmap (excluding self-correlation).**

**3. Minimum Nights vs. Other Variables:**

* **minimum\_nights vs. availability\_365:** **0.14**
  + **Interpretation:** This is a **weak positive correlation**. Listings with a higher minimum nights requirement tend to have slightly higher availability.

### Which variables are strongly or weakly correlated?

**Strongly Correlated Variables:** Based on this heatmap, there are **no strongly correlated variables** among price, number\_of\_reviews, minimum\_nights, and availability\_365. A strong correlation would typically have an absolute value of 0.7 or higher. All values are well below this threshold.

**Weakly Correlated Variables (and their interpretations):**

* **number\_of\_reviews and availability\_365 (0.17):** This is the strongest of the weak correlations. It suggests that listings that are available more often tend to accumulate more reviews. This makes intuitive sense: more availability leads to more bookings, which leads to more opportunities for reviews.
* **minimum\_nights and availability\_365 (0.14):** Weak positive. Listings with longer minimum stays tend to be available for more days per year. Perhaps hosts with higher minimums are less reliant on frequent turnover and thus keep their calendars more open.
* **price and availability\_365 (0.12):** Weak positive. More expensive listings have a slight tendency to be available more often. This could be because they are harder to book, or hosts of premium listings might offer longer availability.
* **price and number\_of\_reviews (-0.06):** Extremely weak negative. As previously discussed, this relationship is almost non-existent.
* **price and minimum\_nights (0.02):** Extremely weak positive. This relationship is essentially non-existent.
* **number\_of\_reviews and minimum\_nights (-0.08):** Extremely weak negative.

**Overall Summary:**

The heatmap reveals that while there are some discernible weak relationships, **none of the selected numeric features are strongly linearly correlated with each other in this Airbnb dataset.** **The strongest linear relationship observed is between number\_of\_reviews and availability\_365, but even this is relatively weak**. This suggests that these variables largely act independently in terms of linear relationships, and other factors not included in this heatmap might be stronger drivers of price or other characteristics.

--------------------------------🡪>>>> End of Assignment <<<<<<<🡨----------------------------------

**Additional Business Insight section:**

**Business insights based on Revenue and other variables**

Insight 1: Which borough generates the highest total revenue?

A screenshot of a graph

AI-generated content may be incorrect.

Looking at the bar chart, the bar for **Manhattan** is significantly taller than the others, reaching approximately $4.2 million.

The next highest is Brooklyn, followed by Queens, Bronx, and Staten Island.

While the Manhattan bar towers at ~$4.2M, the lower rental rates and still-substantial revenue in Brooklyn and Queens make them appealing for **high-ROI, broader-market business opportunities**.

Insight 2: What room types deliver the best revenue per availability day?

A graph of a bar chart

AI-generated content may be incorrect.

* **X-axis:** room\_type (categorized as "Entire home/apt", "Private room", "Shared room").
* **Y-axis:** Is labeled "Mean of Price if Available".

The chart clearly shows that:

* **"Entire home/apt"** has the highest mean price (revenue per availability day), with its bar reaching approximately $210-$220.
* **"Private room"** is next, with a mean price around $90-$100.
* **"Shared room"** has the lowest mean price, around $70-$80.

Therefore, this plot definitively indicates that **"Entire home/apt" delivers the best revenue per availability day.**

Insight 3: How does Airbnb supply concentration compare across boroughs?

A screenshot of a graph with Crust in the background

AI-generated content may be incorrect.

The chart shows the following supply concentration (number of listings) by borough:

* **Manhattan** has the highest concentration of Airbnb supply, with over 20,000 listings.
* **Brooklyn** is the second-highest, with a significant number of listings, approaching 20,000 but still less than Manhattan.
* **Queens** has a substantially lower supply concentration compared to Manhattan and Brooklyn, with around 5,000 listings.
* **The Bronx** has a very low supply, with less than 1,000 listings.
* **Staten Island** has the lowest supply concentration, with just a few hundred listings.

In summary, Airbnb supply is heavily concentrated in **Manhattan and Brooklyn**, with significantly lower concentrations in Queens, and very minimal supply in the Bronx and Staten Island.

### Business Ideas for Airbnb Hosts based on the insights from revenue vs other factors:

1. **Elevate "Entire Home/Apt" Experience for Premium Guests (Manhattan & Brooklyn)**
   * **Insight:** "Entire home/apt" commands the highest average daily prices, and Manhattan/Brooklyn are the highest revenue/supply boroughs.
   * **Idea for Hosts:** If you own or manage an "Entire home/apt" in Manhattan or Brooklyn, double down on premium features and services. This includes high-end amenities (luxury linens, smart home tech), professional interior design, offering personalized welcome baskets, and potentially partnerships for unique local experiences (e.g., private chef, guided tours). This strategy aims to capture the top tier of the market and maximize your already high daily rate potential.
2. **Strategic Niche-Focused "Private Room" Hosting (All Boroughs)**
   * **Insight:** "Private room" still generates good average daily revenue. In Queens, the Bronx, and Staten Island, there's less supply and competition.
   * **Idea for Hosts:** Differentiate your "private room" offering by catering to a specific niche. Examples:
     + **Digital Nomad Haven:** Offer dedicated workspace, high-speed internet, comfortable desk chair, quiet hours.
     + **Artist/Creative Retreat:** Unique decor, good lighting, maybe even access to a shared studio space.
     + **Family-Friendly Stay:** Provide a crib, high chair, child-safe amenities, and recommendations for family activities.
     + **Long-Term Stay Discounts:** Offer attractive weekly/monthly rates to secure longer, more stable bookings, appealing to those relocating or on extended visits.
   * **Value Proposition:** Attract a dedicated segment of guests willing to pay for specific features, even if the base daily rate is lower than an entire apartment.
3. **Optimize Pricing and Availability with Data Tools**
   * **Insight:** Revenue and supply vary significantly across boroughs and room types. Maximizing earnings requires smart pricing.
   * **Idea for Hosts:** Leverage dynamic pricing tools and platforms to automatically adjust your listing's price based on demand, seasonality, local events, and competitor pricing. Also, actively manage your minimum\_nights and availability\_365 to fill gaps and capitalize on peak periods.
   * **Value Proposition:** Ensure your listing is always priced competitively to maximize occupancy and "revenue per availability day" without constant manual adjustments.
4. **Explore "Shared Room" Viability with Unique Value Adds (Budget-Conscious Travelers)**
   * **Insight:** "Shared room" has the lowest average daily price, but there's always a budget travel market.
   * **Idea for Hosts:** If you have a "shared room," focus on providing exceptional value beyond just a bed. This could include:
     + Superb cleanliness and security.
     + Communal spaces that encourage interaction.
     + Basic amenities like free breakfast or laundry access.
     + Being located very close to public transport hubs or major attractions.
   * **Value Proposition:** Attract budget travelers and backpackers by offering a highly efficient, clean, and well-located option, focusing on maximizing occupancy through volume.