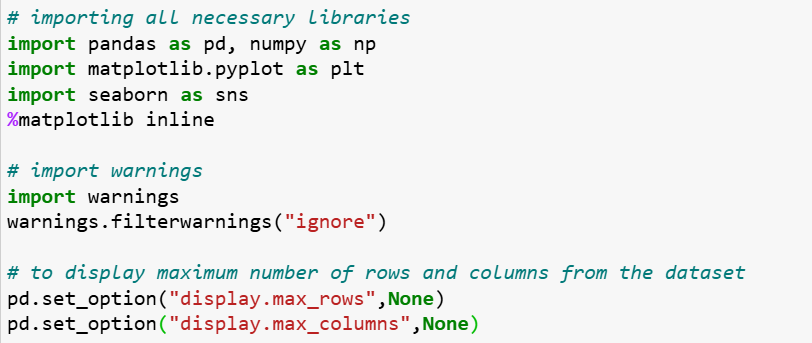
**Methodology Document**

1. **Introduction:**

* In this project, we analyzed a comprehensive Airbnb dataset to uncover patterns and trends related to various factors such as neighbourhood groups, room types, number of reviews, and pricing of listings.
* Our goal was to derive insights that could support key stakeholders in making data-driven decisions.
* The following methodology details the steps taken for data cleaning, exploratory data analysis (EDA), and key visualizations, utilizing Python, Excel, and Tableau.

1. **Data overview:**

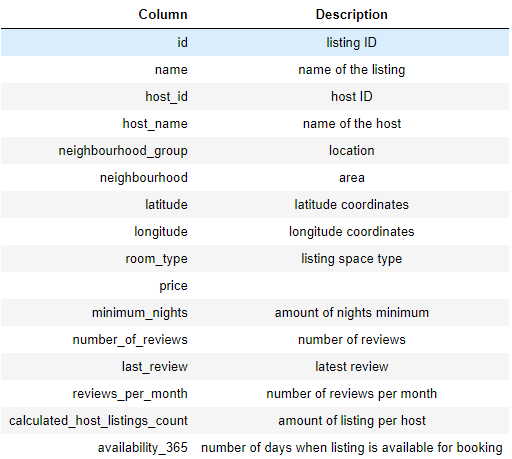
* Initial Dataset Understanding
  + - Importing libraries and warnings and setting the display to maximum



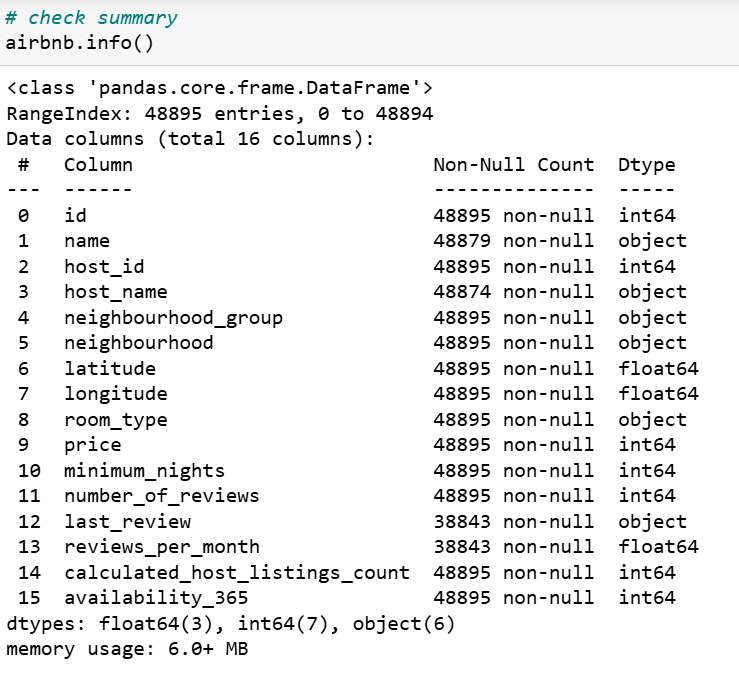
* + - Loading the data



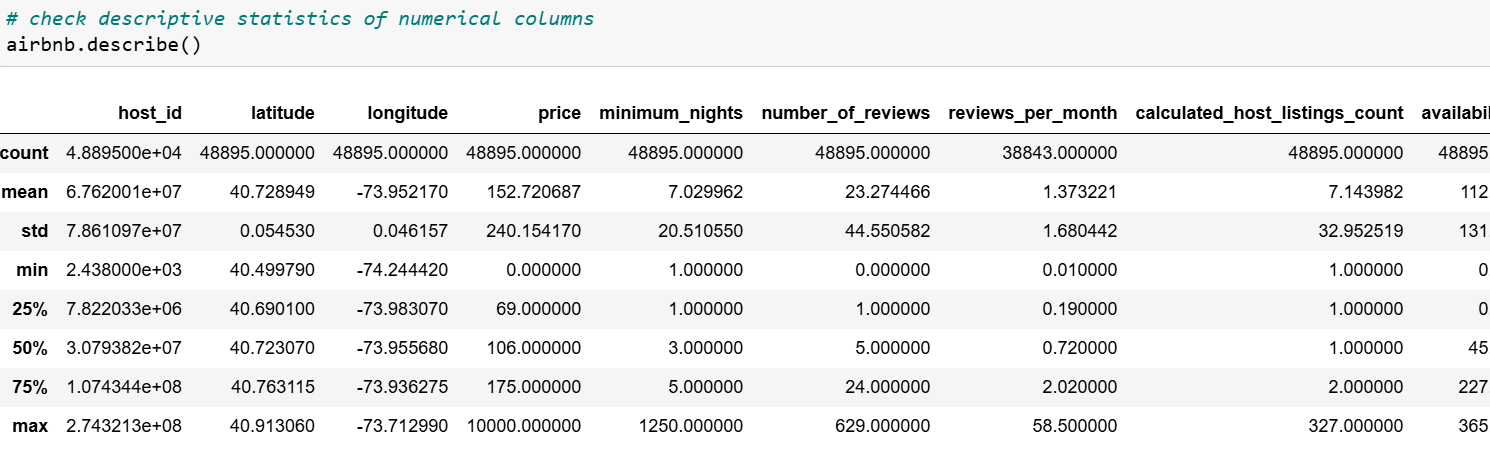
* + - Columns description



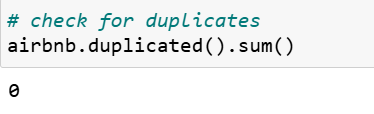
* + - Summary of dataset



* + - * Total entries are 48895 and there are 16 columns.
      * Looks like there are few missing values.
      * last\_review should be of date type.
    - Descriptive summary

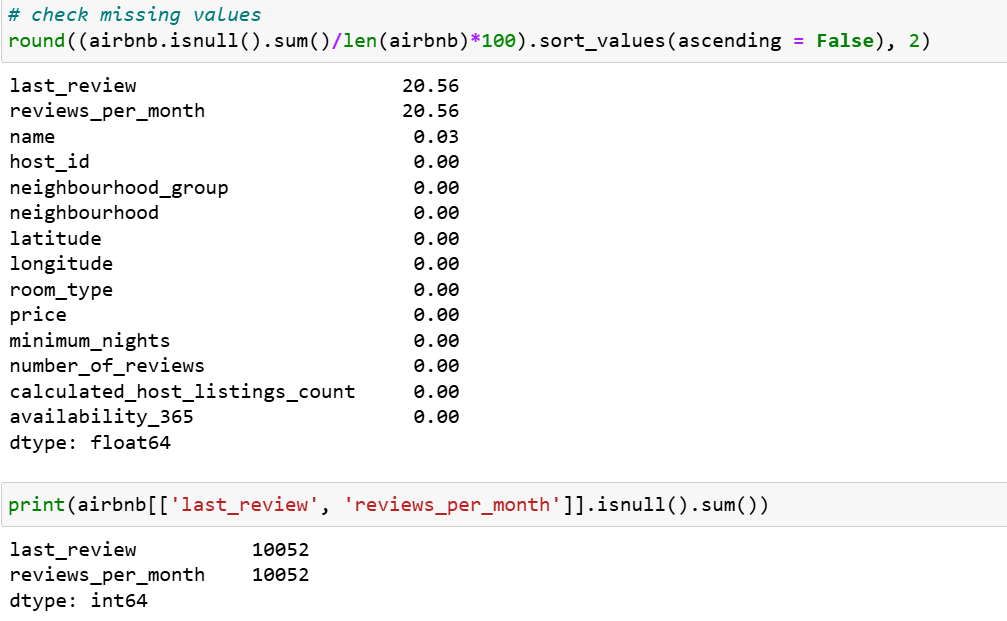


1. **Data wrangling**
2. Handling duplicate entries:

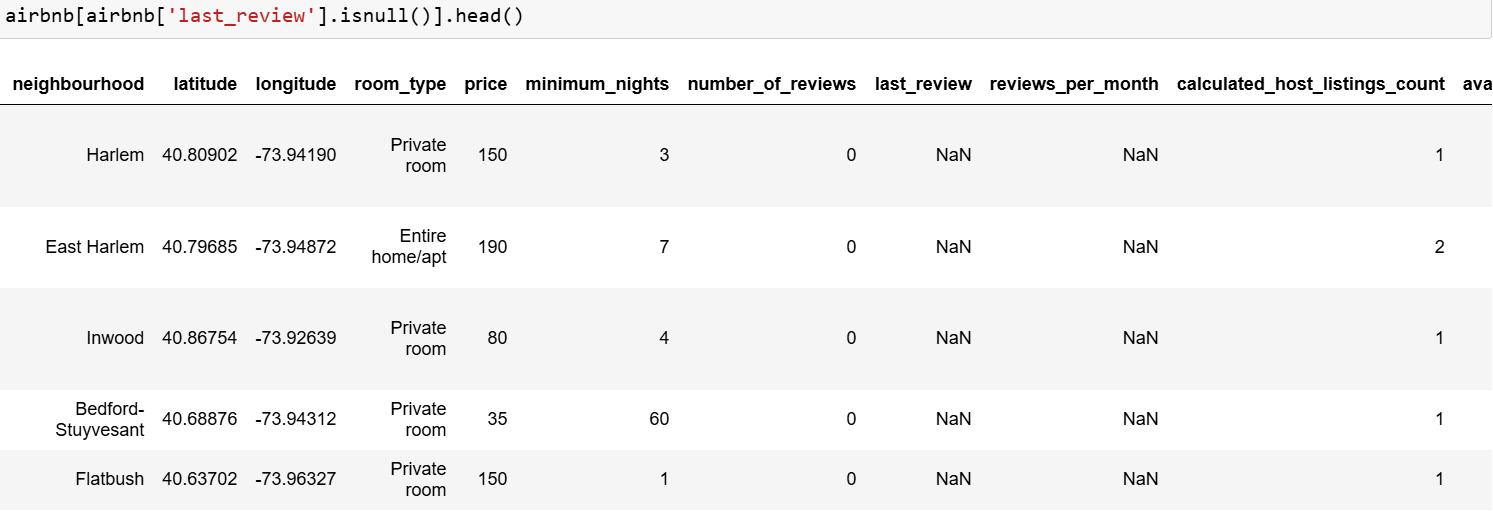


1. DATA CONSTRAINTS:

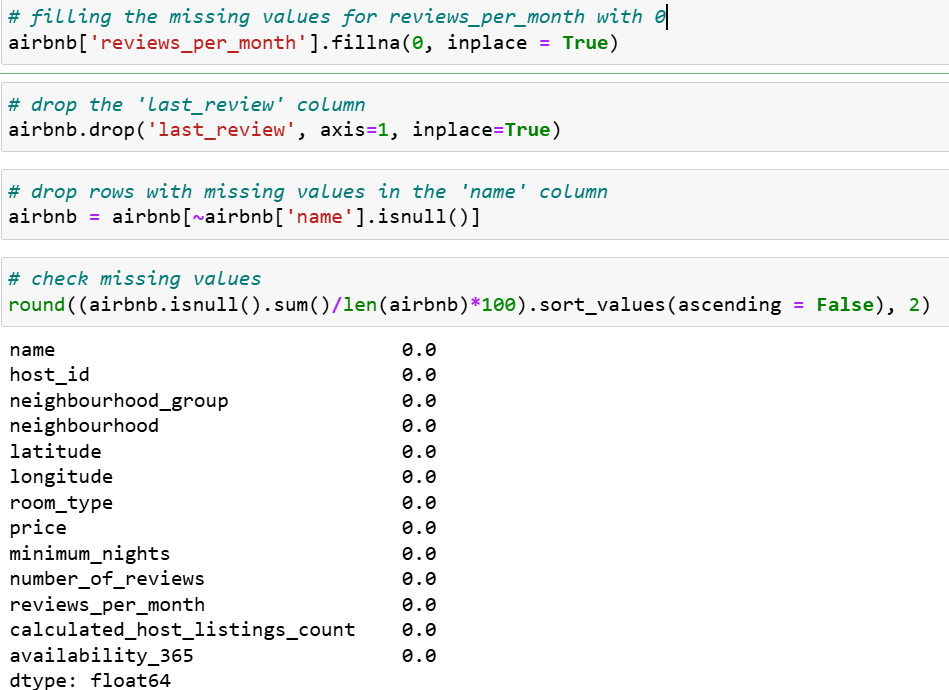
* Removing or imputing missing values:



* ASSUMPTION:
* Count of missing values for last\_review and reviews\_per\_month are the same, it likely indicates that these listings have no recorded activity.

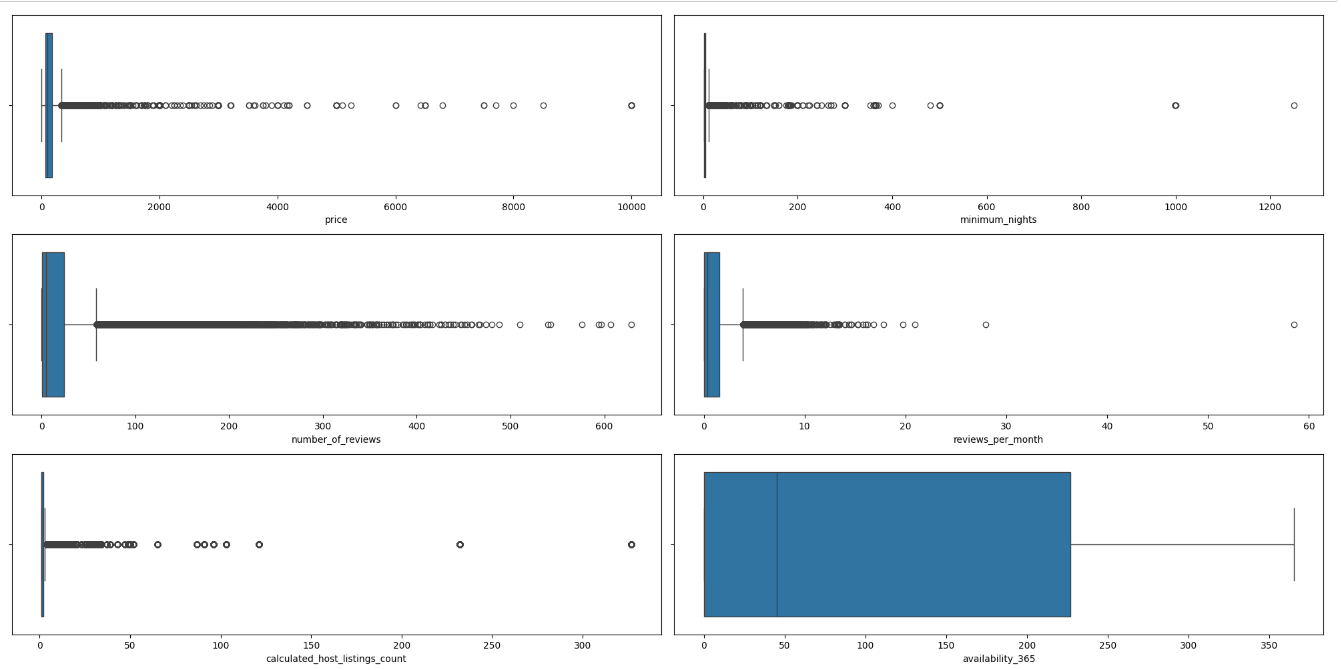


* Some listings have NaN values in the last\_review and reviews\_per\_month columns despite having activity recorded in the calculated\_host\_listings\_count column.
* This suggests that while these listings are active and available for booking, they have not yet received any reviews.
* These listings might be relatively new or less popular, which is why they haven't received any reviews yet.
* Since we already have other metrics like number\_of\_reviews and reviews\_per\_month, we can drop the last\_review column.
* Handling the missing values and rechecking:

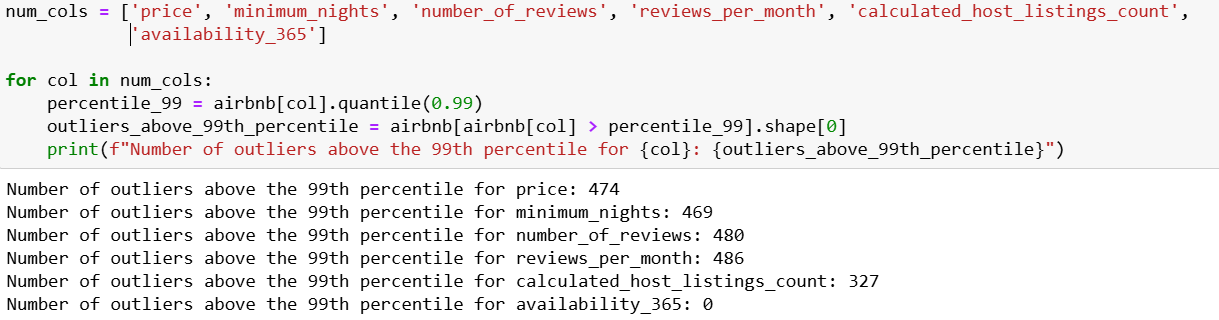


1. Filtering out irrelevant or outlier data:



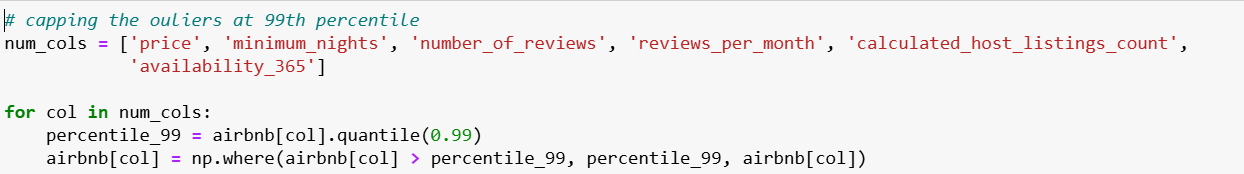


* The datapoints above the 99th percentile:

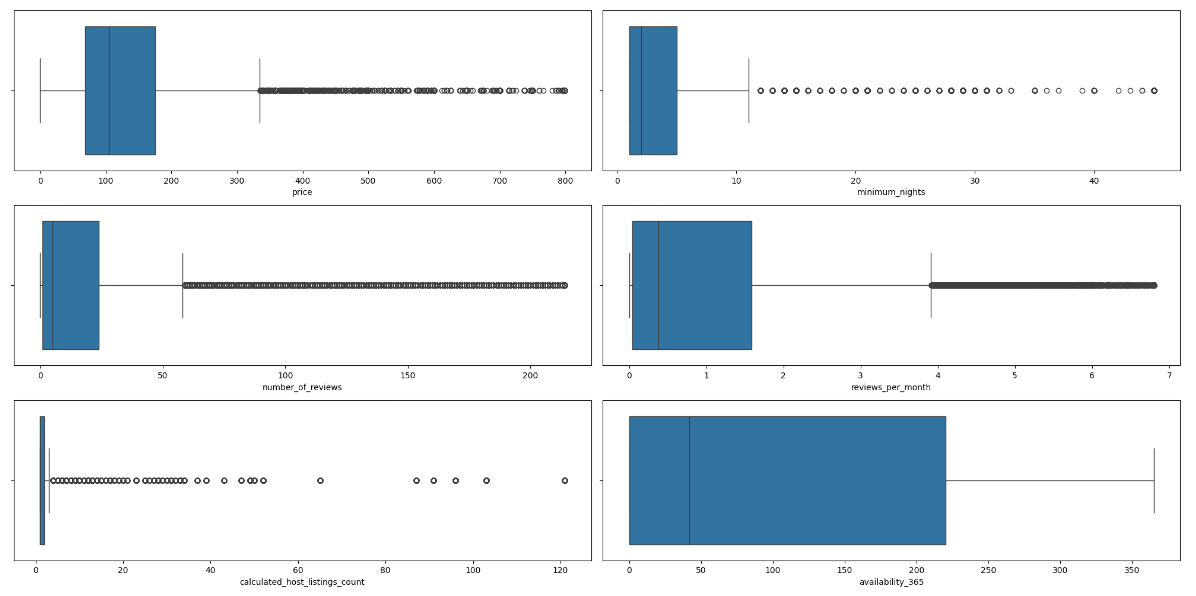


* Assumption:

We assumed that values above the 99th percentile represent extreme outliers, which could skew the analysis. To address this, we decided to cap values above the 99th percentile. This approach helps in normalizing the data without completely removing valuable information.

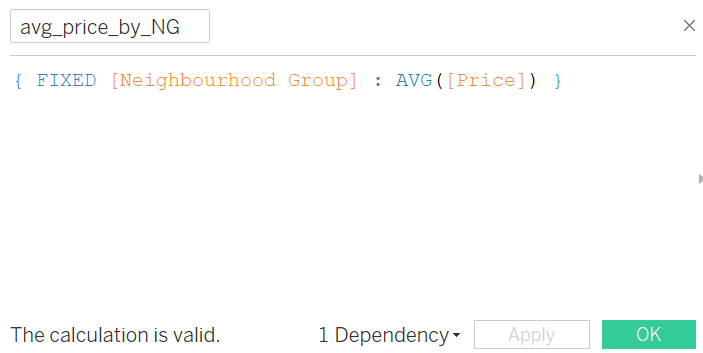


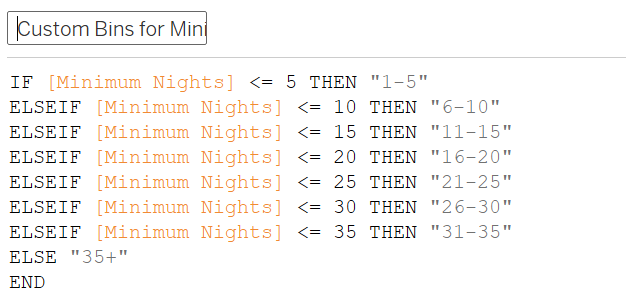
* After capping values:



1. Calculated Fields:

* To calculate average price across the neighbourhood groups to see aggregate



* Binning minimum nights
* 

1. **DATA ASSUMPTIONS:**

* We assumed that a higher number of reviews reflects the popularity and demand for a listing, suggesting higher occupancy rates.
* We assumed that hosts' pricing and availability strategies are consistent over time.
* We assumed that higher pricing generally correlates with higher quality.
* We assumed that historical data on reviews and occupancy would be indicative of future trends.

1. **Supporting Tools:**

* Python
* Excel
* Tableau