**AIR BNB CASE STUDY**

**Understanding the data**:

* Started with understanding the problem statement.
* Went through the data set quickly to understand the available fields.
* Noted down some fields that had to be binned further

**Data Wrangling:**

1. Loade the data set into python under the name abnb
2. Started off with looking at the details of the available attributes.  
   abnb.info()
3. Converted the Last\_review data type to appropriate format i.e date time format so that we would be able to extract the year and month in later stage for our analysis.
4. Checked the summary stats for both continuous and categorical columns.  
   abnb.describe()  
   abnb.describe(include=’object’)
5. Checked for missing values in the data.  
   print(abnb.isnull().mean().sort\_values(ascending = False)\*100)

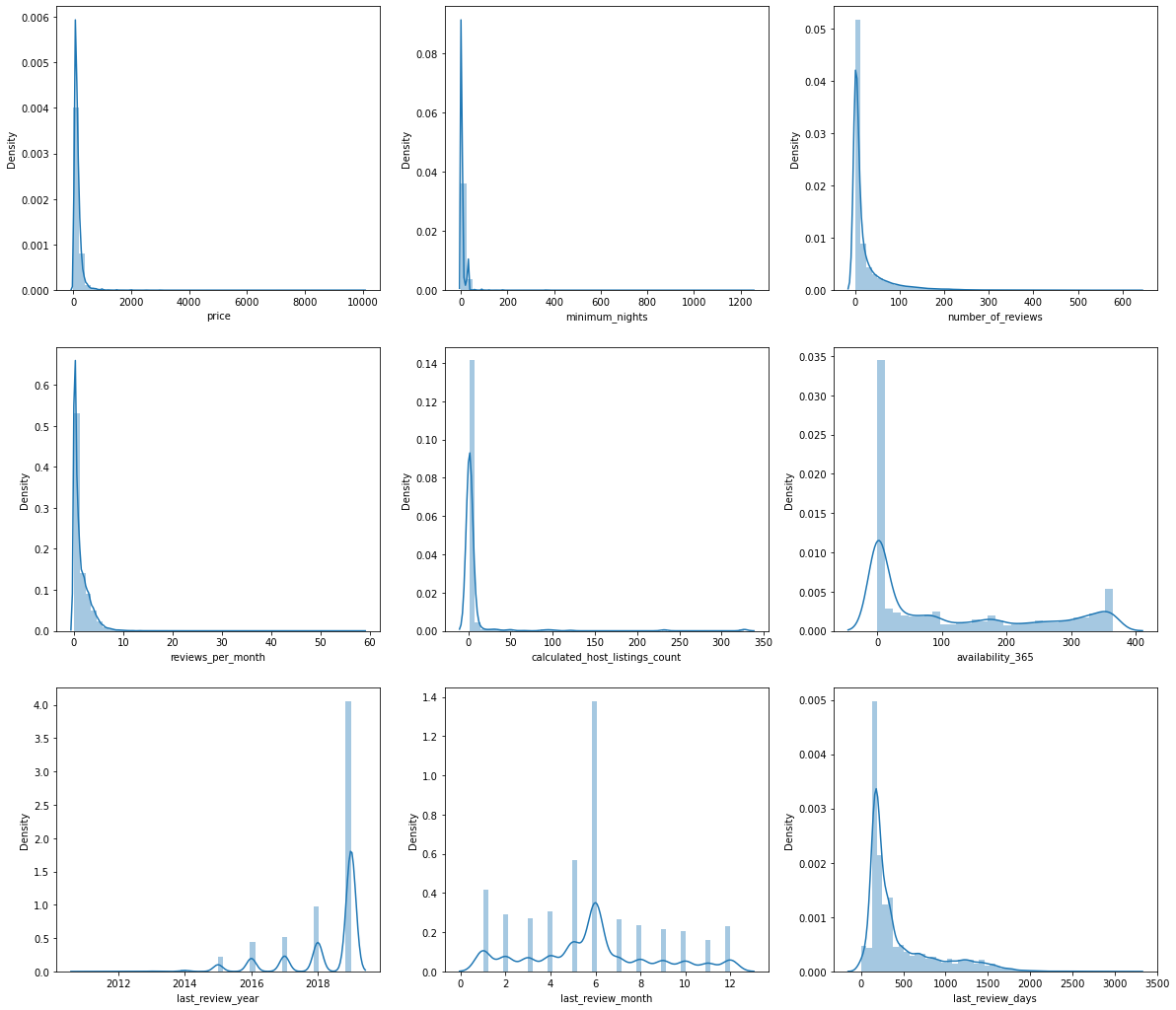
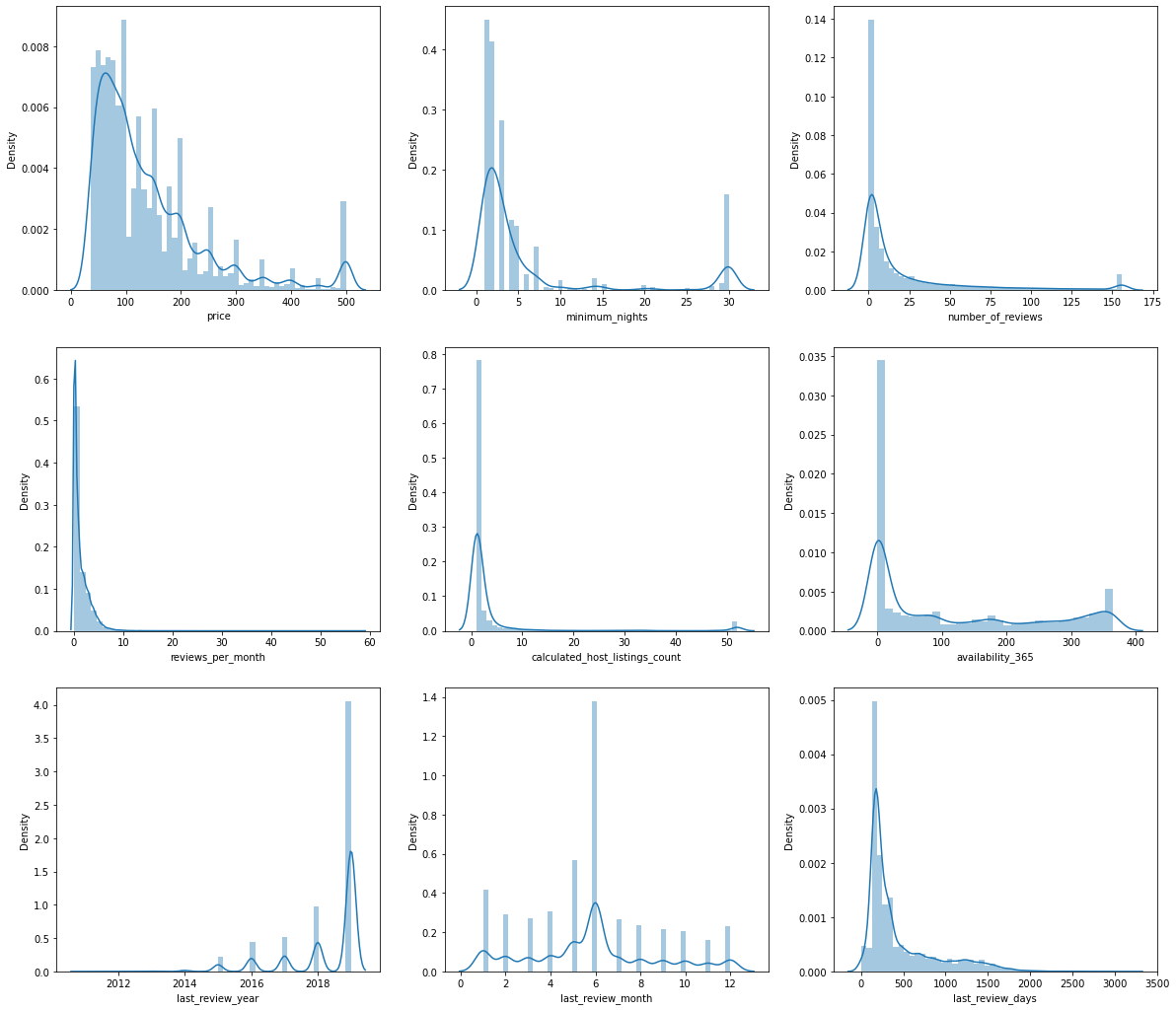
We saw that last\_review, reviews\_per\_month have high and exactly same count of missing values (20.56%). Since these columns are related to reviews so as per my understanding the reason can be that the place never recieved any review/ none has ever visited the place. We will not remove the rows and proceed as it is. The same is the case with host\_name and name as they have extremely low missing values(0.04% & 0.03%) and will not cause any issues in visualisation.

1. Created bins for 3 columns namely
   1. Minimum\_nights
   2. number\_of\_reviews
   3. availability\_356
2. Extracted month and year from the column last\_review. Extracting year and month from the last\_review column will help in analyzing the trend in Airbnb listings over time. It will help us identify if there are any monthly/seasonal/Yearly popularity trends.These information will be useful in making business decisions such as pricing, marketing, and promotions.  
   abnb['last\_review\_year'] = abnb['last\_review'].dt.year # extract year and create new column

abnb['last\_review\_month'] = abnb['last\_review'].dt.month # extract year and create new column

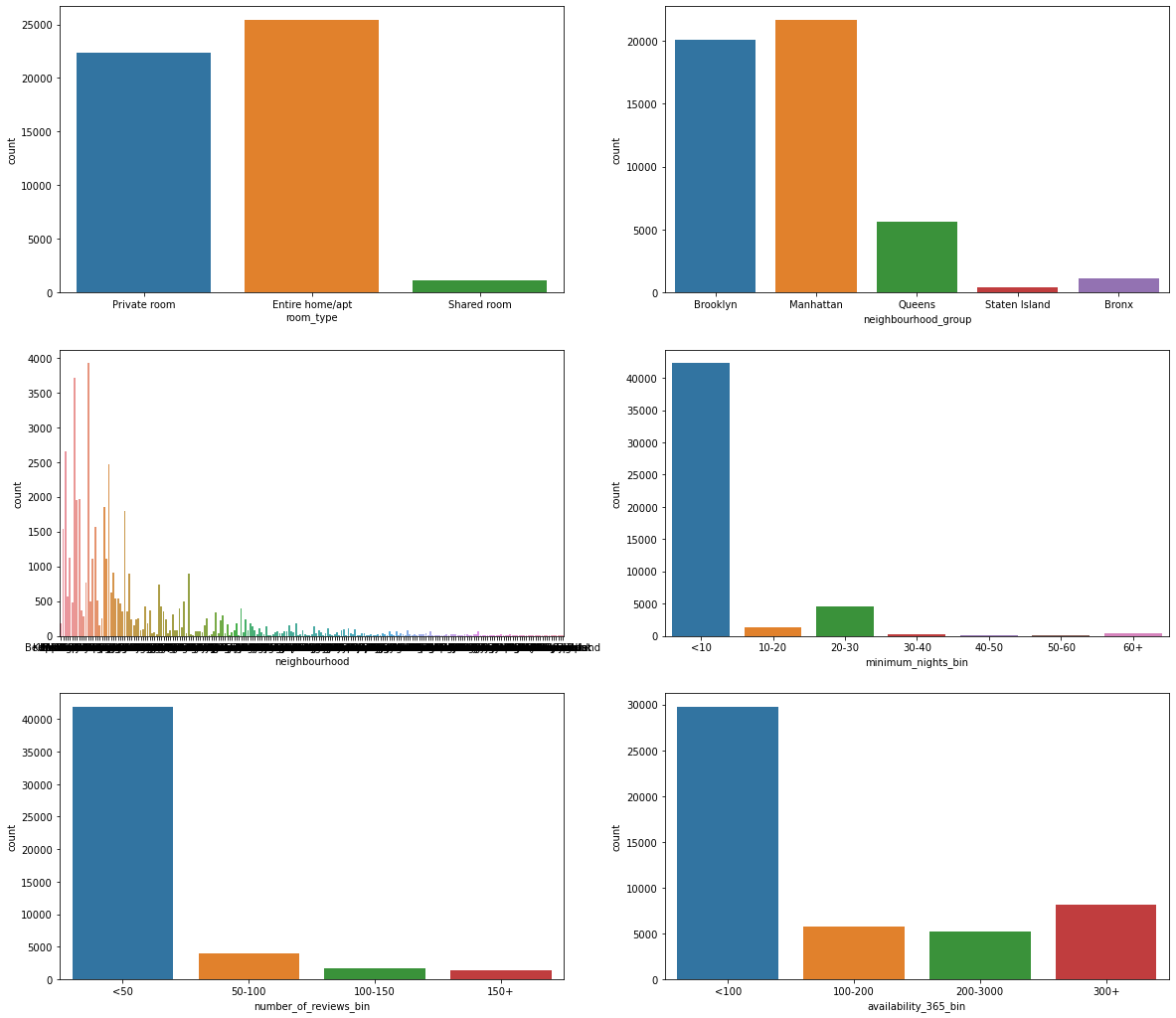
Now our data set was ready for visualisations.

**Data Visualisations:**

1. Created a distplot for all the present integer columns, by which we found that the data was right skewed.  
     
   Also cross verified by creating box and whisker chart for these.
2. Used Winsoriation method for Outlier treatment and plotted distplot again:  
   

\*\*Observation\*\*

* + The price is heavily skewed, with most listings priced under $200 per night.
  + The minimum\_nights is also heavily skewed, with most listings having a minimum stay of less than 7 nights.
  + number\_of\_reviews is heavily skewed to the right, with most listings having less than 50 reviews.
  + distribution of reviews\_per\_month is also heavily skewed, with almost all listings having less than 10 reviews per month.
  + The majority of the listings are managed by hosts with a relatively small number of listings.
  + Distribution of availability\_365 is bimodal, i.e., a large number of listings have either high or low availability throughout the year.
  + Most of the reviews in the dataset were left in 2019.But the number of reviews is significantly low for years before 2019, we can also observe a dropping trend in the previous years.
  + June is the month when majority of the reviews are posted.
  + Most of the reviews in the dataset are fresh, with a significant number of reviews in less than 300 days. Also, there are reviews that are more than a year old, which means that property is not much popular among the guests.

1. Checked the availability of the properties throughout the year  
   abnb.availability\_365.value\_counts(normalize=True)  
   🔷 Almost 36% of the properties listed are not available whole year
2. Plotted count plot to get initial idea on the categorical columns:  
   

**1**. The most common room\_type listed on Airbnb is Entire home/apt, followed by Private room, and Shared room.

**2.** The neighbourhood group with the highest number of listings is Manhattan, followed by Brooklyn, Queens, Bronx, and Staten Island.

**3**. The majority of the listings in the dataset have a minimum stay requirement of less than a 10 days.

**4.** The majority of listings in the dataset have a relatively low number of reviews.

**5.** The availability of a listing is the number of days it is available for booking in a year. The majority of listings in the dataset are available for less than 100 days

Similarly did initial visualization to get an idea on what we will have to prepare and then finally used Tableau with combination of Power BI to create proper visualisation that has been presented in the presentation file.

**PRESENTATION:**

* Made the presentation using the visualisations created from tableau and power BI.
* Added recommendation and per slide insights.