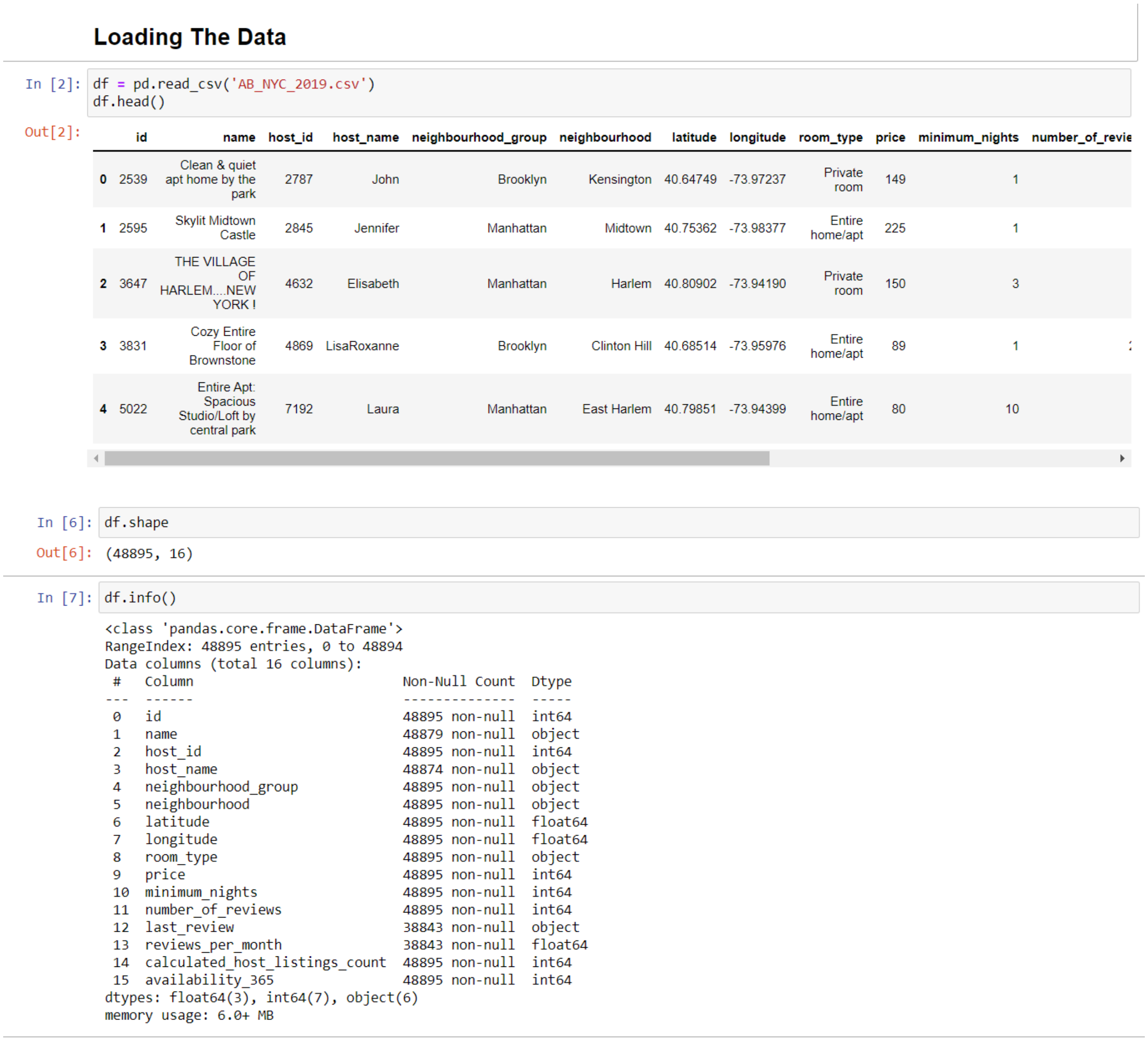
Methodology

Data Processing:

- We went through the data and did the data study thoroughly. We made a rough document that included various graphs.
- We have used python for finding out the null values and the shape of the data.



Loading The Data

```
In [2]: df = pd.read_csv('AB_NYC_2019.csv')
        df.head()
```

Out[2]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	2
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

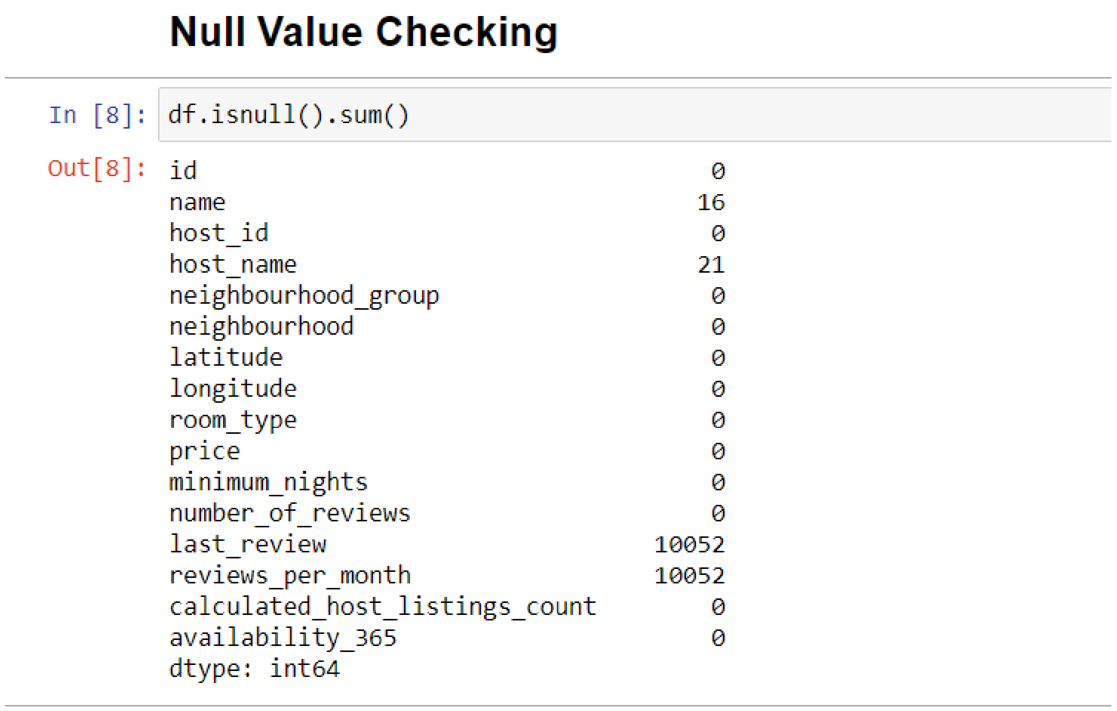
```
In [6]: df.shape

Out[6]: (48895, 16)
```

```
In [7]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

- There are total 48895 rows and 16 columns.
- The data types of each column are correctly given so we don't need to change the data type of any column



Null Value Checking

```
In [8]: df.isnull().sum()
```

```
Out[8]: id                                    0
        name                                 16
        host_id                               0
        host_name                            21
        neighbourhood_group                   0
        neighbourhood                         0
        latitude                              0
        longitude                             0
        room_type                             0
        price                                 0
        minimum_nights                        0
        number_of_reviews                     0
        last_review                       10052
        reviews_per_month                 10052
        calculated_host_listings_count        0
        availability_365                      0
        dtype: int64
```

- In the data set, two columns have more than 10k nulls values, ‘reviews per month’ and ‘last review’.
- For analysis purpose, column ‘reviews per month’ is fairly an important KPI to deduce insights. Hence, the null values were replaced with O so as to keep all listings in consideration
- column last review has more than 10,000 data points with null values and is not taken into consideration to generate insights as it could cause anomalies
- Name and host_name column have few null values that will not impact our analysys so we are not imputing them.
- No irrelevant columns were removed from the dataset.

- For Visualisation purposes, we have used Tableau and performed the EDA, and extract the insights.
- Used variety of charts to determine user preferences based on multiple KPIs or parameters

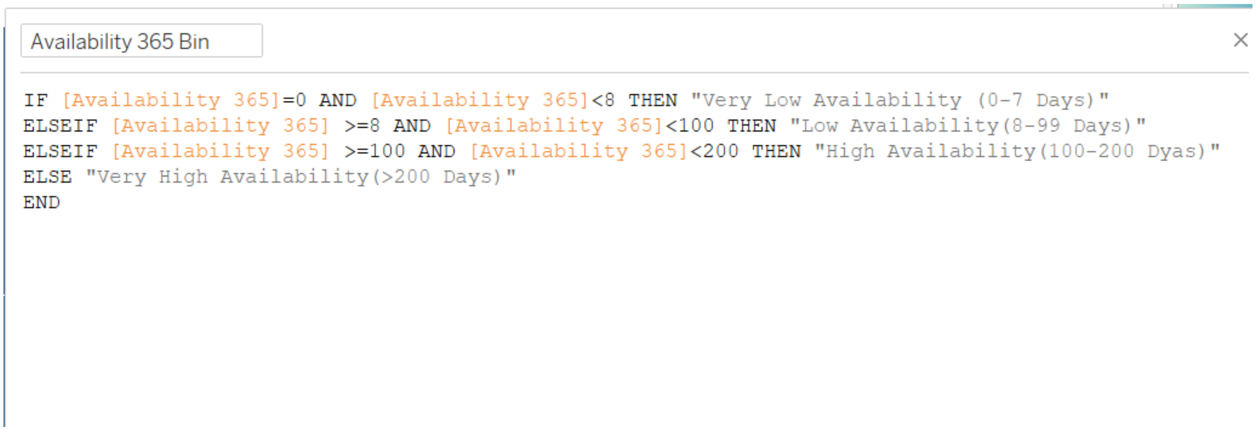
Assumptions:

- The Price column indicates the price per night for listings and we assume currency is in USD.
- We assumed that the higher number of reviews means the most preferred property.

- The low availabilty_365 means the property is booked for most of the days, which makes it popular among users.
- We assumed that airbnb is focused primarily at the mentioned neighbourhood groups only and not any other.
- We assumed that post covid restrictions, travel industry will boost.

Data Visualisation Done With Tableau Tool

- We have performed the Exploratory Data Analysis, created various graphs and extracted the insights.
- We have created bins for Availability_365 column through calculations.



Availability 365 Bin

```
IF [Availability 365]=0 AND [Availability 365]<8 THEN "Very Low Availability (0-7 Days)"
ELSEIF [Availability 365] >=8 AND [Availability 365]<100 THEN "Low Availability(8-99 Days)"
ELSEIF [Availability 365] >=100 AND [Availability 365]<200 THEN "High Availability(100-200 Dyas)"
ELSE "Very High Availability(>200 Days)"
END
```

Key findings:

1. User preferred the Entire home/apt (51.97%) & Private room(45.66%).
2. The avg price of listed properties is higher (196.9) for the Manhattan area, which is the highest among others. Brooklyn comes second in that list (124.4)
3. The customer opted to stay with the property that provides a minimum night stay of 1-7.
4. The Entire home/apt received 51.0% reviews, which makes it the most preferable room type.
5. Host ‘Michael’ is the reviewed highest when it comes to Entire Home/Apartment bookings in Manhattan
6. Higher Number of listings are available in Neighbour like Manhattan and Brooklyn compared to Bronx and Staten Island.