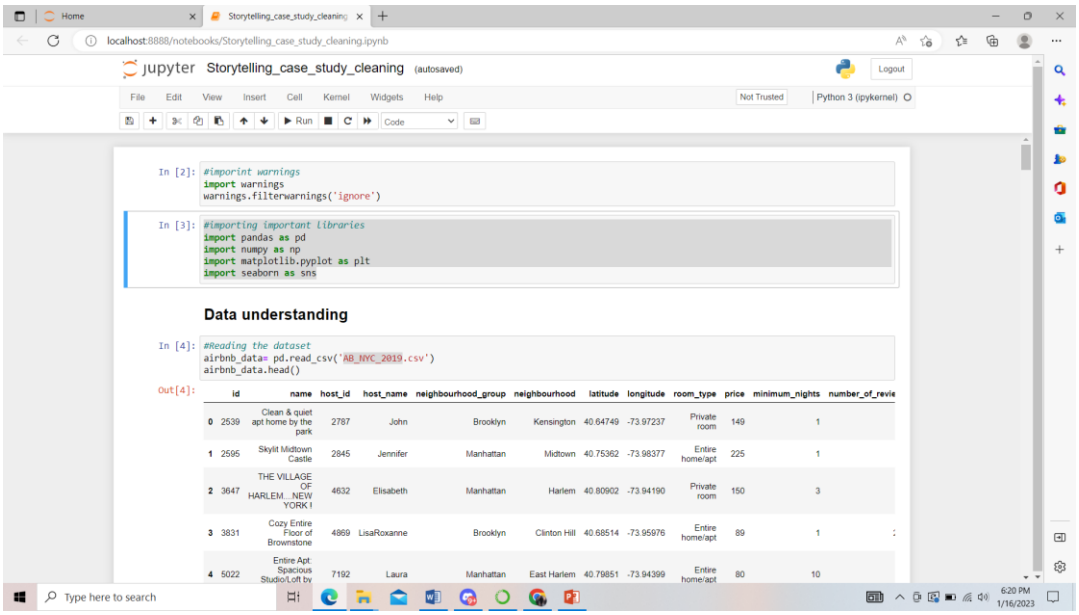
METHODOLOGY

1. We first downloaded the AB_NYC_2019 dataset and imported into a jupyter notebook using pandas for cleaning the data.

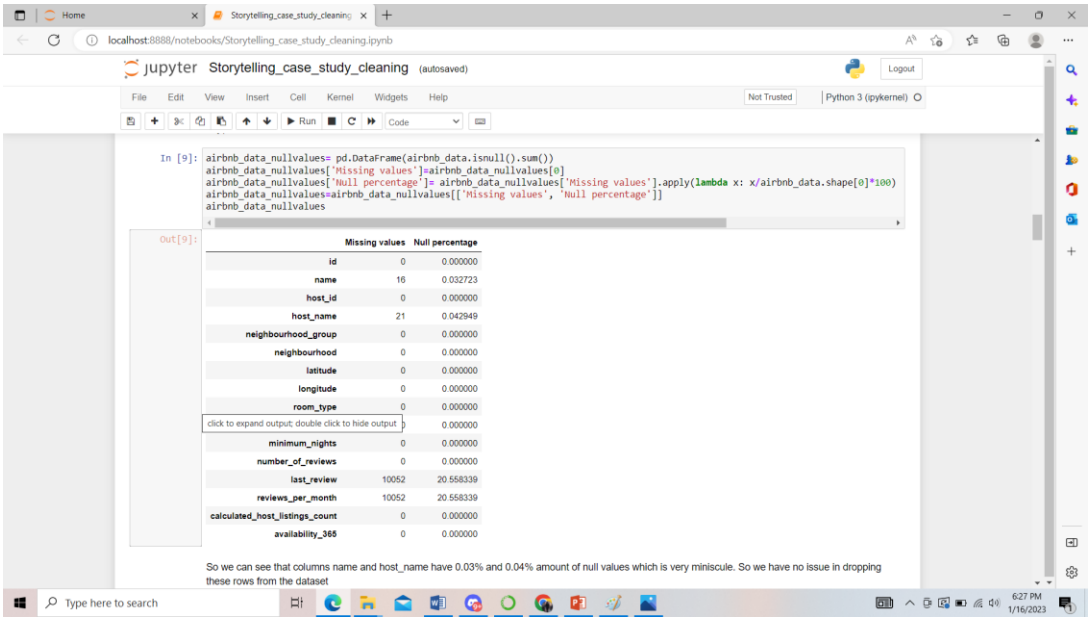
Code snippet:



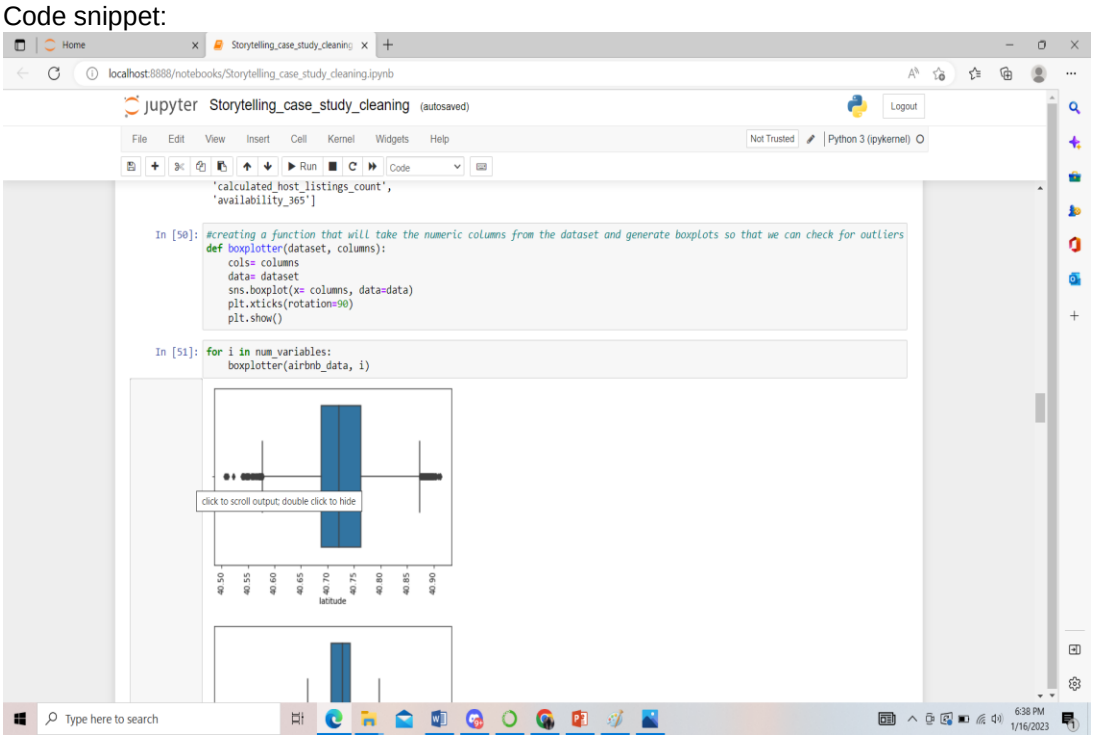
	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-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	
4	5022	Entire Apt: Spacious Studio/Loft by	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

2. We imported all the necessary libraries like numpy, pandas, matplotlib and seaborn.
3. We then quickly checked the dimensions and shape of the data for our understanding with info and shape functions and then moved on to the data cleaning step.
4. We first checked for duplicates in the data. None were found so we moved on to checking for missing values.
5. We checked for missing values in all the columns and checked percentage of the missing values for each column as well.

Code snippet:

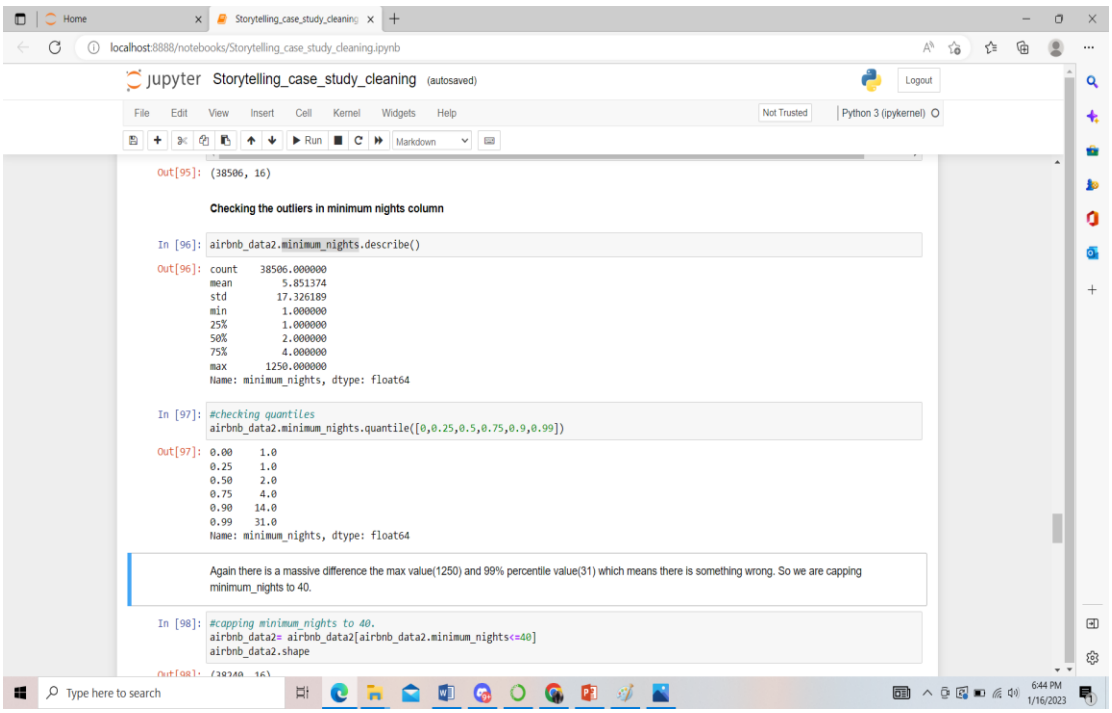


6. We saw that columns name and host_name have 0.03% and 0.04% amount of null values which is very miniscule. So we had no issue in dropping these rows from the dataset.
7. We also see that last_review and reviews_per_month has 20.54% null values. We know that last_review is supposed to be a date column and if we replace the missing values with the mode or a 'missing' category. It won't help us with the further analysis as it won't be treated as a date column. So we are dropping these rows
8. Now that the all the null values are taken care of we moved to the outlier checking process,
9. We plotted box plots for the numeric variablles to check for outliers.
   Code snippet:



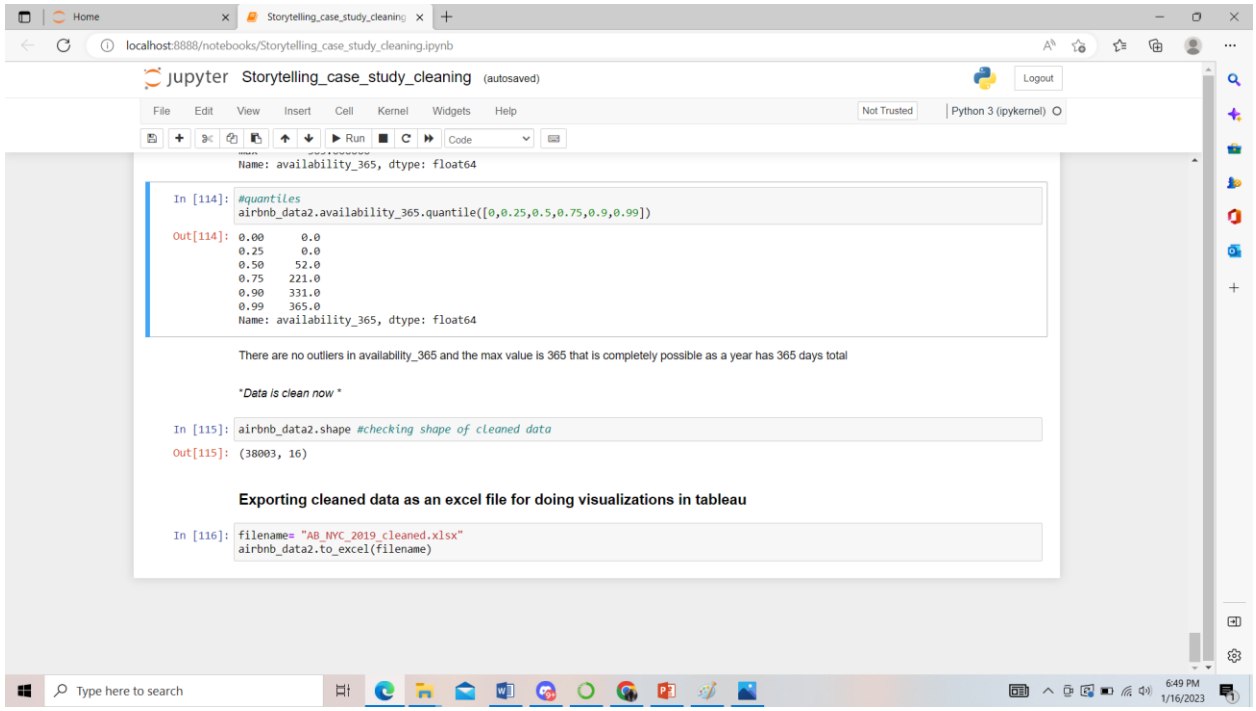
10. We identified the columns that may have outliers from the boxplots and then conducted univariate analysis on them to further investigate if there are legitimate outliers or not.
11. We checked for how big the difference in the 99th percentile value and the max value is or how big the gap between mean and median is to check for outliers. We also used business logic to determine if the outliers made sense or not.
12. For e.g. in the column minimum_nights we found that there is a massive difference between the max value(1250) and 99% percentile value(31) which means there is something wrong. So we capped minimum_nights to 40.

    Code snippet:



13. After dealing with outliers we finally arrived at the clean dataset which had 38003 rows and 16 columns.
14. We then finally exported the clean dataset as an excel workbook so that we can import the same to tableau for exploratory analysis and generating insights.

    Code snippet:



15. We then imported the clean dataset into tableau for the analysis, creating graphs and generating insights.