

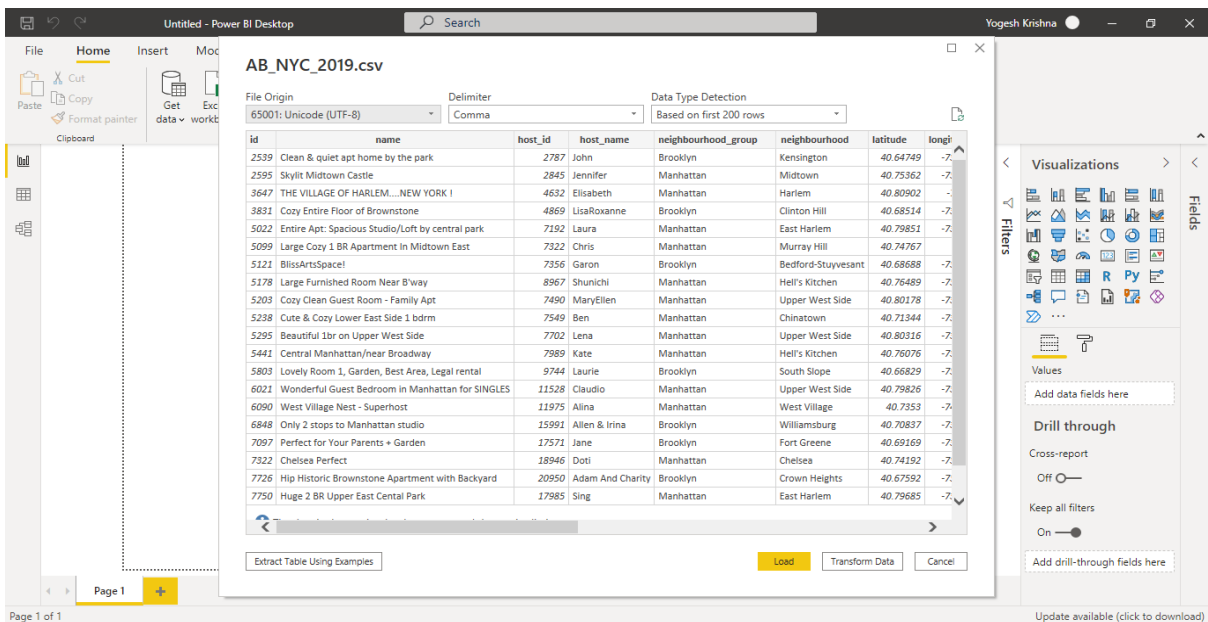
Methodology Document for Airbnb, NY Dataset

The data cleaning, transformation, and visualization were done on the **Microsoft PowerBI** tool.

All the necessary steps on the data have been performed before creating the PPT. Below are the methodology/steps followed:

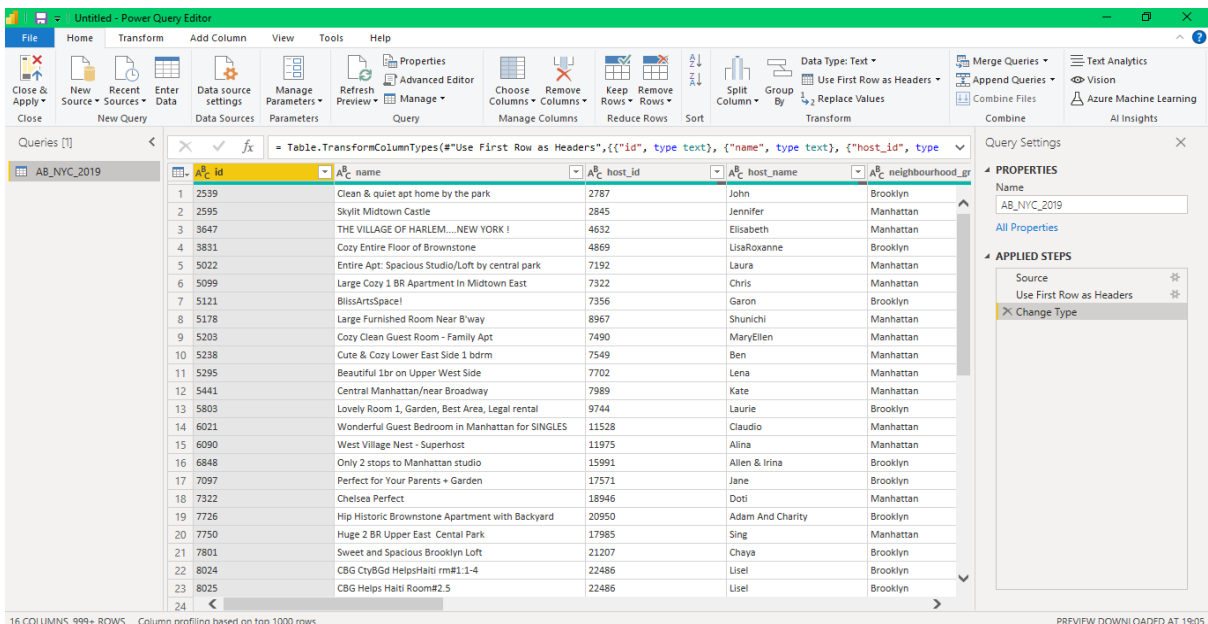
Data Cleaning:

1. Entire Airbnb data was loaded in PowerBI



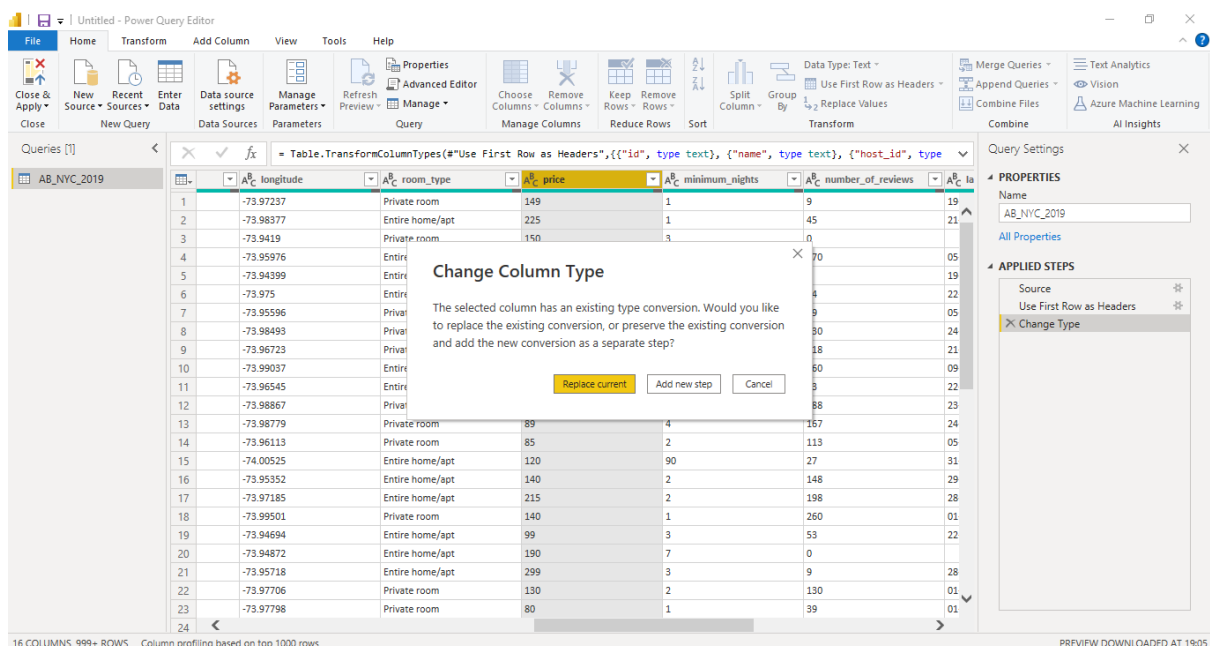
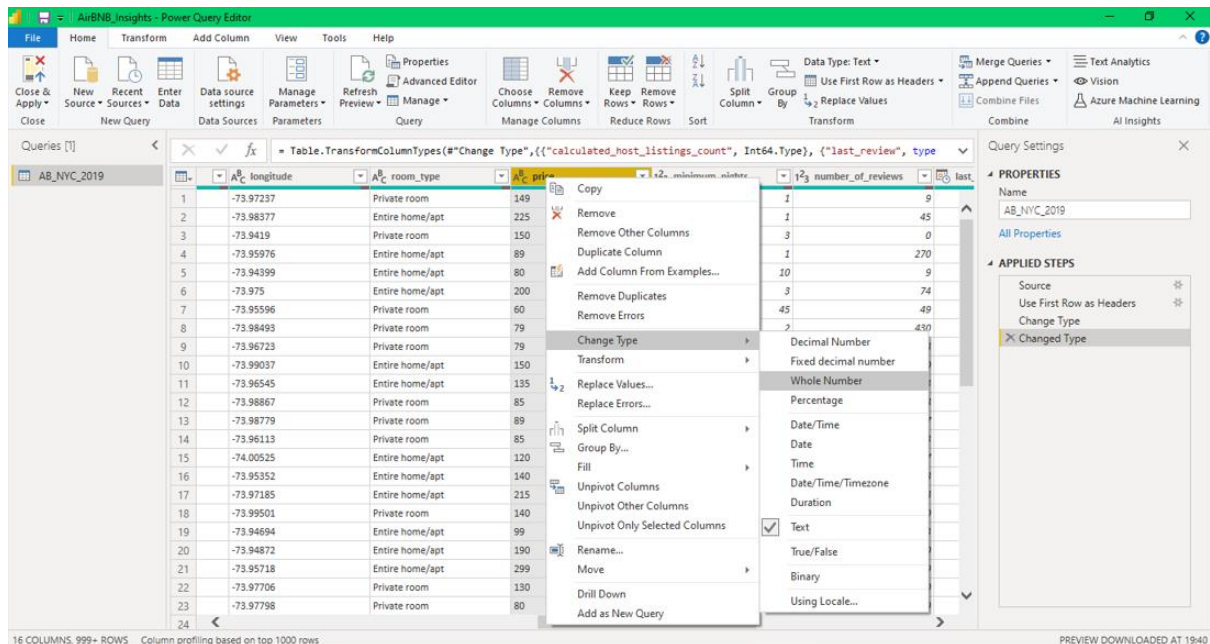
The screenshot displays the Microsoft Power BI Desktop interface. The main window shows the 'AB_NYC_2019.csv' data source loaded. The ribbon at the top includes 'File', 'Home', 'Insert', and 'Model View'. The main data view shows a table with columns: id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, and longitude. The right-hand pane shows the 'Visualizations' and 'Fields' sections. The 'Fields' section lists the columns available for use in visualizations.

2. Started Power Query editor for data cleaning and transformation.



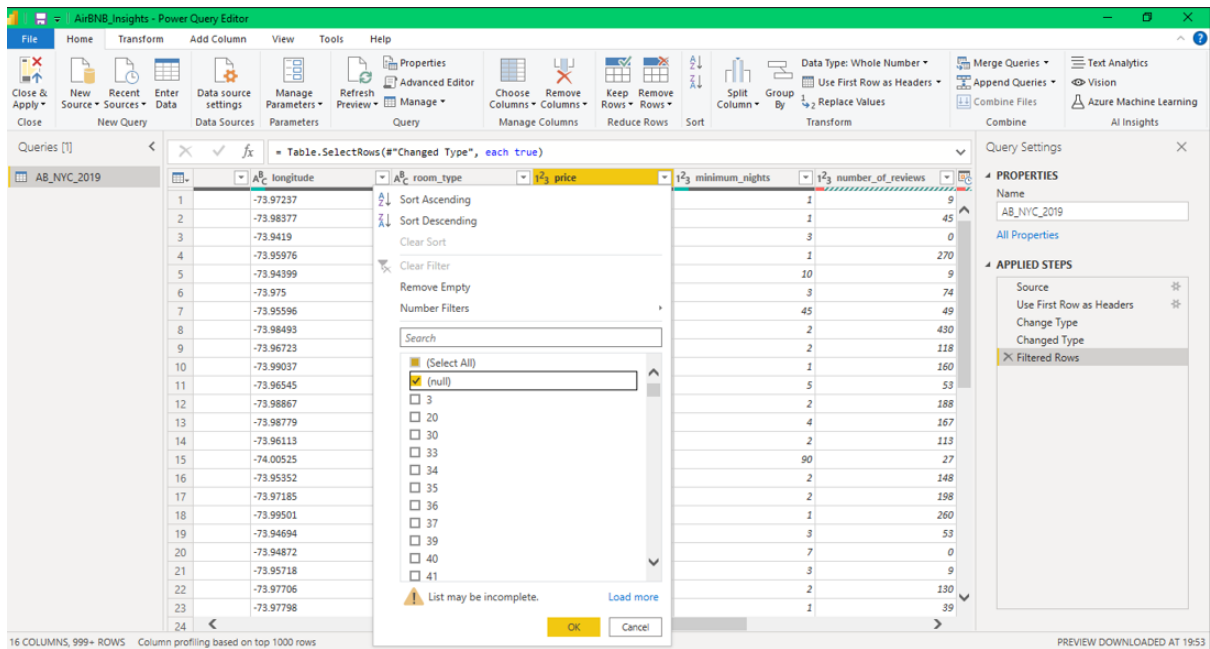
The screenshot displays the Microsoft Power Query Editor interface. The main window shows the 'AB_NYC_2019' data source loaded. The ribbon at the top includes 'File', 'Home', 'Transform', 'Add Column', 'View', 'Tools', and 'Help'. The main data view shows a table with columns: id, name, host_id, host_name, and neighbourhood_group. The right-hand pane shows the 'Properties' and 'Applied Steps' sections. The 'Applied Steps' section lists the steps applied to the data, including 'Source', 'Use First Row as Headers', and 'Change Type'.

- Changed the column type for columns **price**, **minimum_nights**, **number_of_reviews**, **calculated_host_listing_count**, **availability_365** to the whole number.



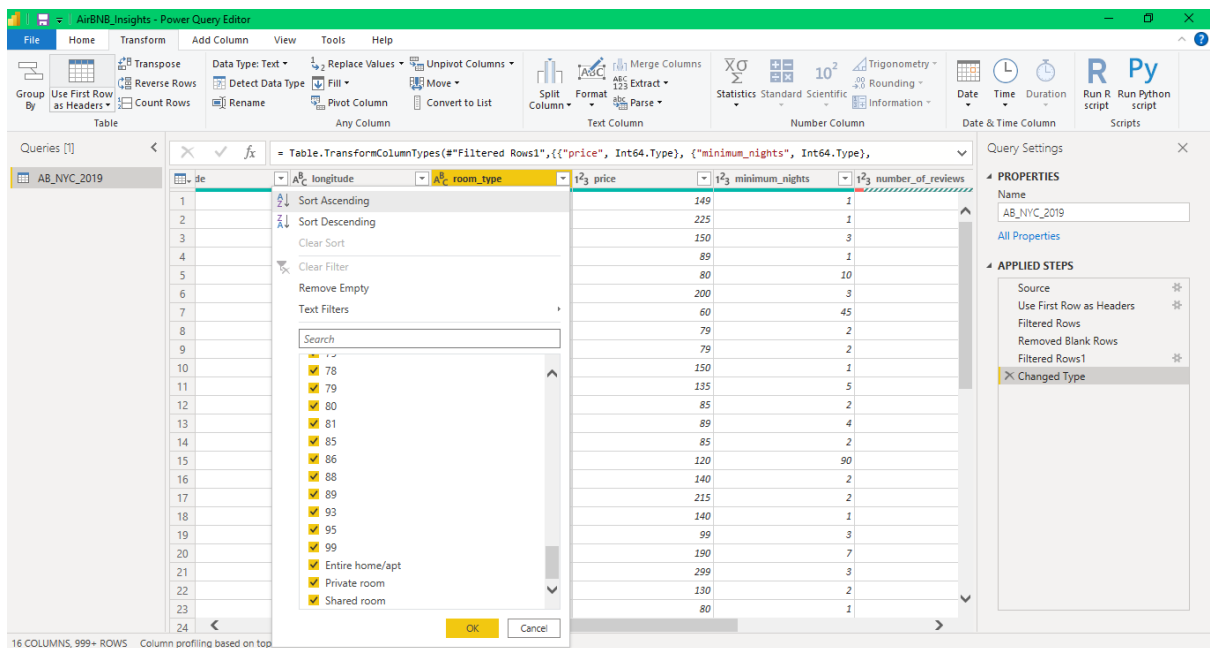
- Checked all the columns one by one for NULL values:

Found that many columns had NULL values such as **price**, **availability_365**, **number_of_reviews**, etc. Since these columns are NULL it may imply that either the property was never booked or isn't available anymore for booking. Therefore, dropping all the NULL.



5. Removed errors from columns **room_type** column:

While checking the room_type, many of the values were having numbers. Therefore filtered those and kept only Entire home/apts, Private room and shared rooms.



6. After filtering found that the data is messed up i.e., columns were having wrong values. Such as longitude column has values of room_type (Private room, shared room etc.)

Power Query Editor - AirBNB_Insights

Formulas: Table.SelectRows(#"Changed Type", each ([room_type] <> "Entire home/apt" and [room_type] <> "Private room" and

	longitude	room_type	price	minimum_nights	number_of_reviews	last_review
1	Private room	89		30	Error	Error
2	Private room	99		3	25	Error
3	Entire home/apt	225		3	134	Error
4	Private room	56		1	145	Error
5	Private room	69		2	177	Error
6	Entire home/apt	170		3	116	Error
7	Entire home/apt	298		3	16	Error
8	Private room	59		2	151	Error
9	Private room	50		2	21	Error
10	Private room	80		5	0	null
11	Private room	65		5	15	Error
12	Entire home/apt	160		1	70	Error
13	Entire home/apt	89		2	58	Error
14	40.59195	-73.94639	Error		500	30
15	Entire home/apt	140		7	65	Error
16	Entire home/apt	200		1	3	Error
17	Private room	45		5	2	Error
18	Entire home/apt	89		1	192	Error
19	Private room	67		6	22	Error
20	Private room	86		6	34	Error
21	Private room	72		1	96	Error
22	Entire home/apt	95		14	5	Error
23	Entire home/apt	230		6	9	Error
24						

16 COLUMNS, 159 ROWS Column profiling based on top 1000 rows

Query Settings: Name: AB_NYC_2019

APPLIED STEPS: Source, Use First Row as Headers, Filtered Rows, Removed Blank Rows, Filtered Rows1, Changed Type, Filtered Rows2

7. Removed errors from last_review column

Power Query Editor - AirBNB_Insights

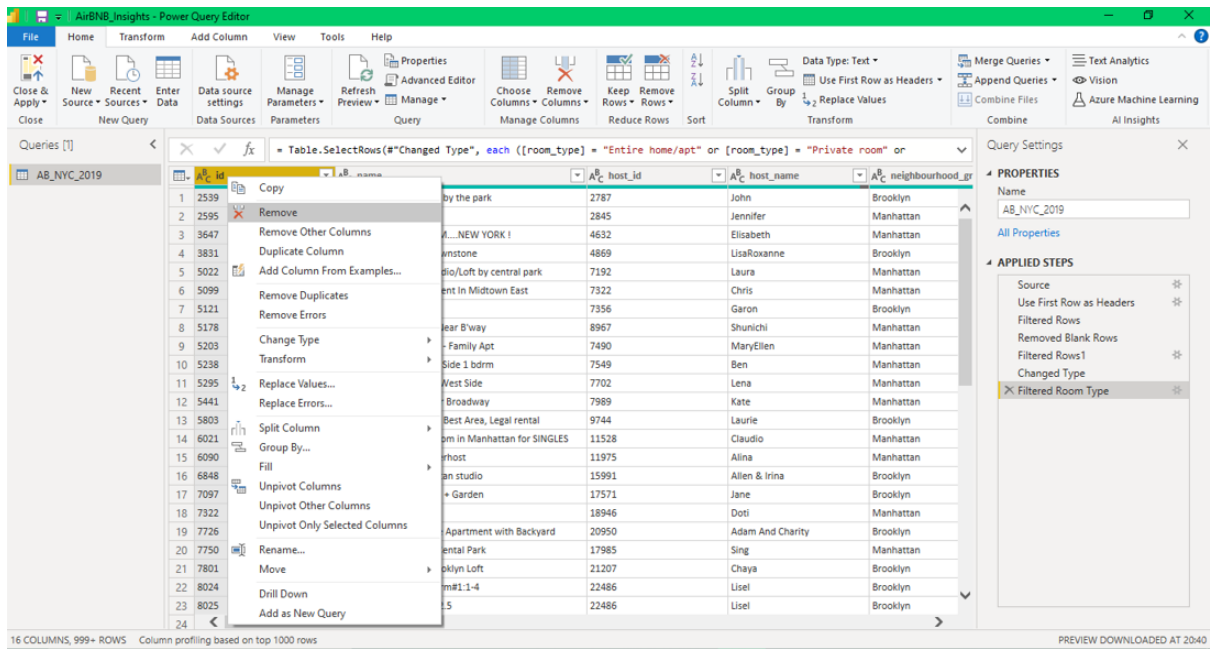
Formulas: Table.RemoveRowsWithErrors(#"Removed Errors", {"number_of_reviews"})

	room_type	price	minimum_nights	number_of_reviews	last_review
1	Private room	149	1	9	
2	Entire home/apt	225	1	45	
3	Private room	150	3	0	
4	Entire home/apt	89	1	270	
5	Entire home/apt	80	10	9	
6	Entire home/apt	200	3	74	
7	Private room	60	45	49	
8	Private room	79	2	430	
9	Private room	79	2	118	
10	Entire home/apt	150	1	160	
11	Entire home/apt	135	5	53	
12	Private room	85	2	188	
13	Private room	89	4	167	
14	Private room	85	2	113	
15	Entire home/apt	120	90	27	
16	Entire home/apt	140	2	148	
17	Entire home/apt	215	2	198	
18	Private room	140	1	260	
19	Entire home/apt	99	3	53	
20	Entire home/apt	190	7	0	
21	Entire home/apt	299	3	9	
22	Private room	130	2	130	
23	Private room	80	1	39	
24					

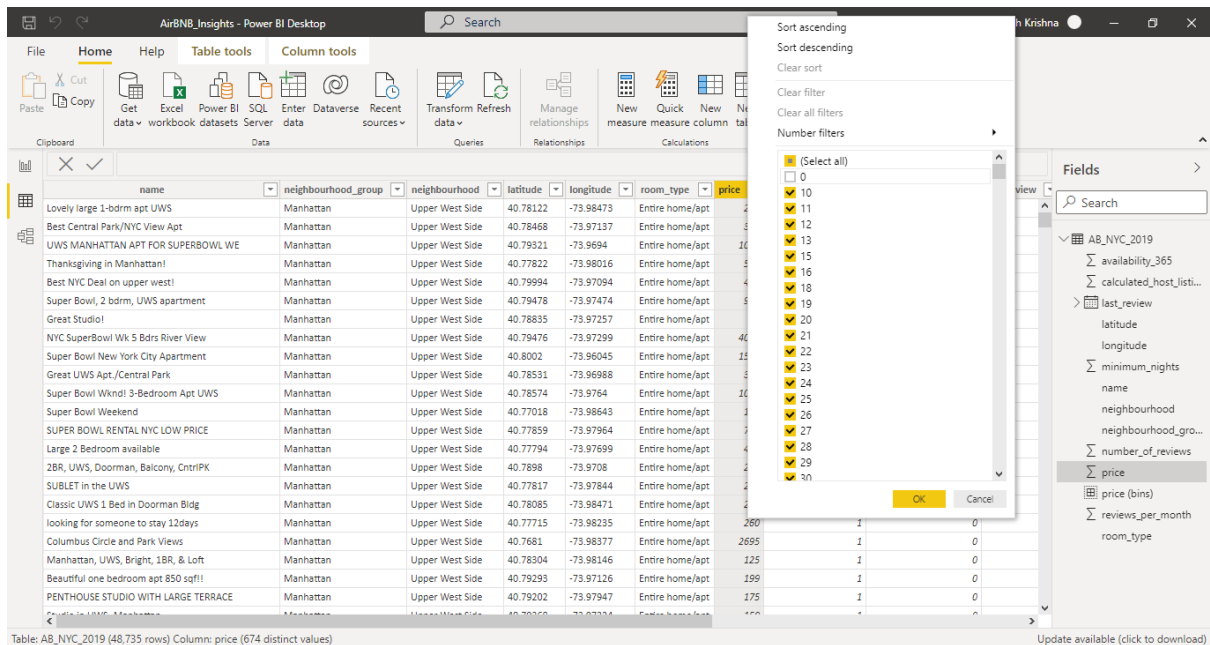
16 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

Context Menu: Copy, Remove, Remove Other Columns, Duplicate Column, Add Column From Examples..., Remove Duplicates, Remove Errors, Change Type, Transform, Replace Values..., Replace Errors..., Group By..., Fill, Unpivot Columns, Unpivot Other Columns, Unpivot Only Selected Columns, Rename..., Move, Drill Down, Add as New Query

8. Removing unwanted columns such as ID, host_id, host_name, etc which didn't give out important insights.



9. Removed the rows where the **price** column was 0.



10. Removed the **price** column outliers such as 10000, 9999, etc.

11. Removed the **minimum_nights** columns having **outliers** greater than 180 days. As nobody will book a host for more than 6 months.

12. Removed the **number_of_reviews** column having 0s. That means nobody ever stayed in that property and will not give many insights about customer preference.

Filter Rows

Apply one or more filter conditions to the rows in this table.

Basic Advanced

Keep rows where 'number_of_reviews'

is greater than 0

And Or

OK Cancel

14 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 09:55

13. Removed rows where **availability_365** column is 0

Filter Rows

Apply one or more filter conditions to the rows in this table.

Basic Advanced

Keep rows where 'availability_365'

is greater than 0

And Or

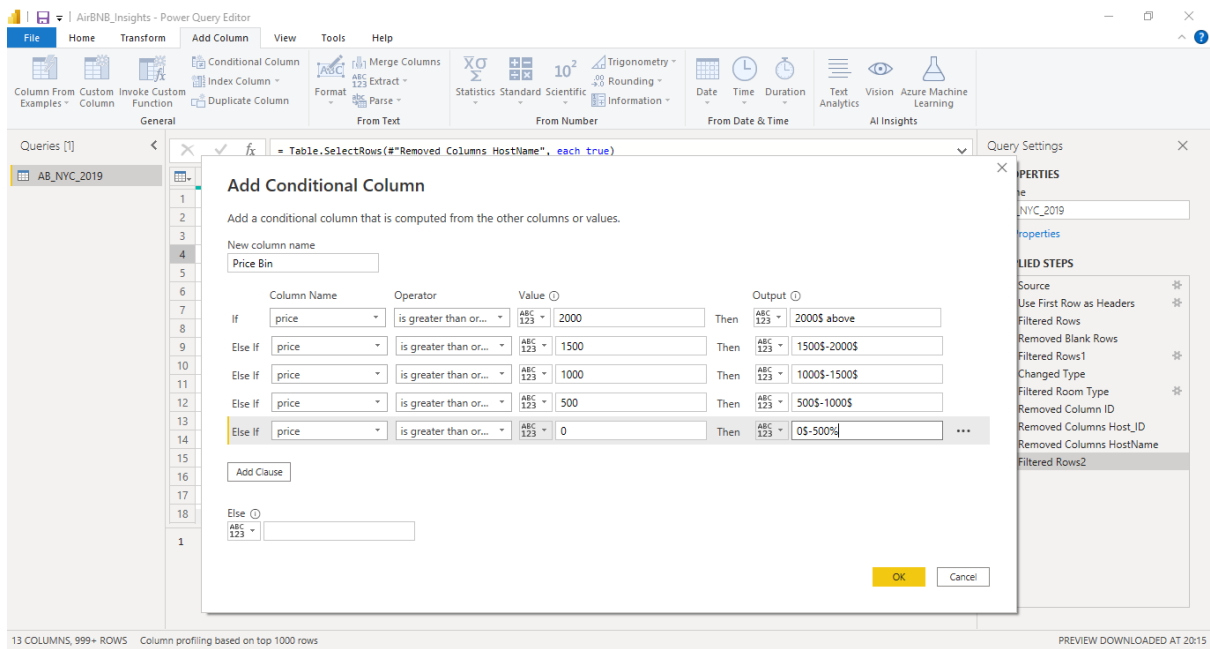
OK Cancel

15 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

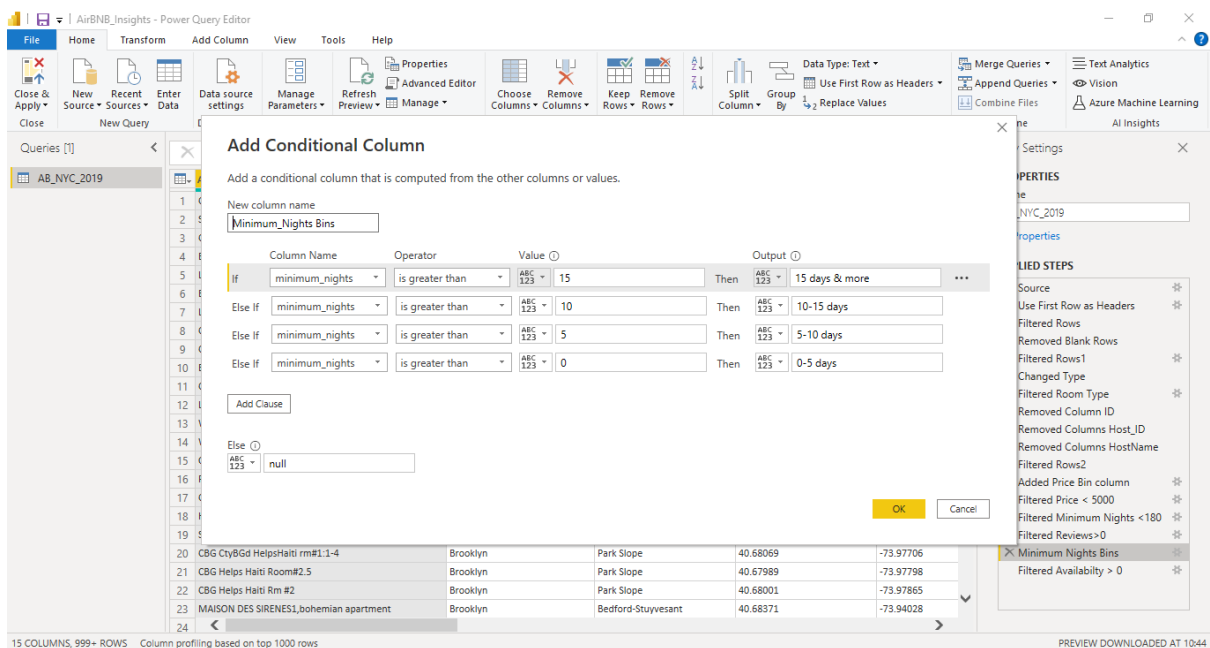
PREVIEW DOWNLOADED AT 10:20

Data Preparation:

1. Created a conditional column (Price Bin) to create price bins for generating insights.



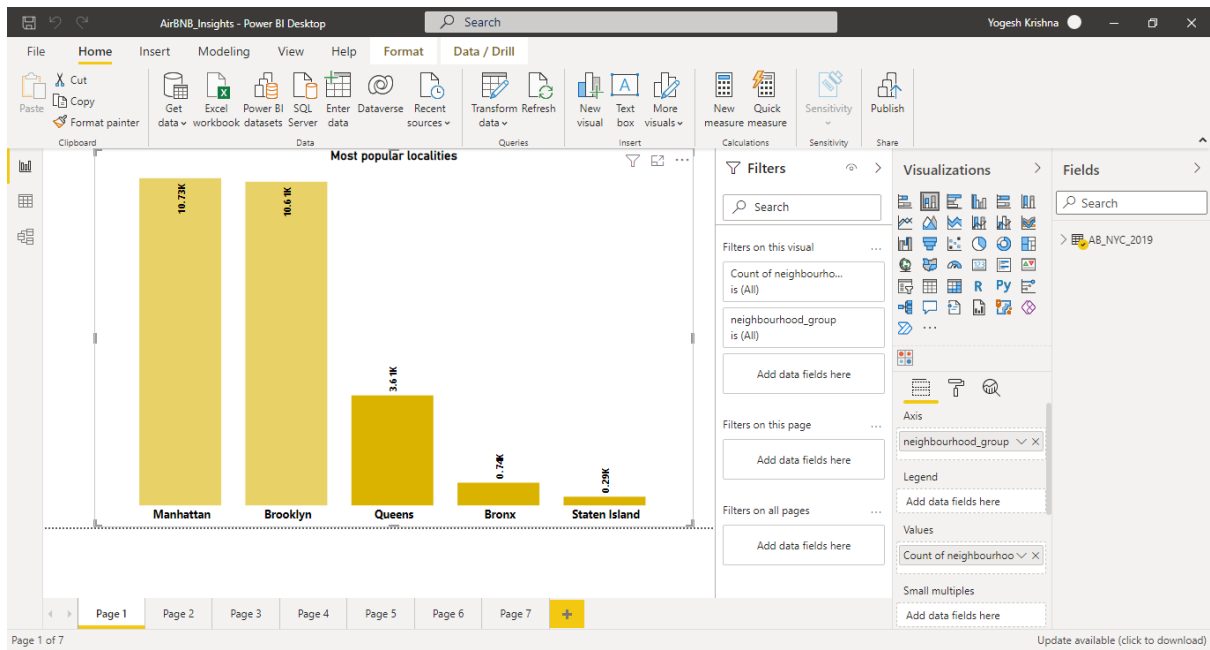
2. Created a conditional column (Minimum_Nights Bins) to create minimum_night bins for generating insights.



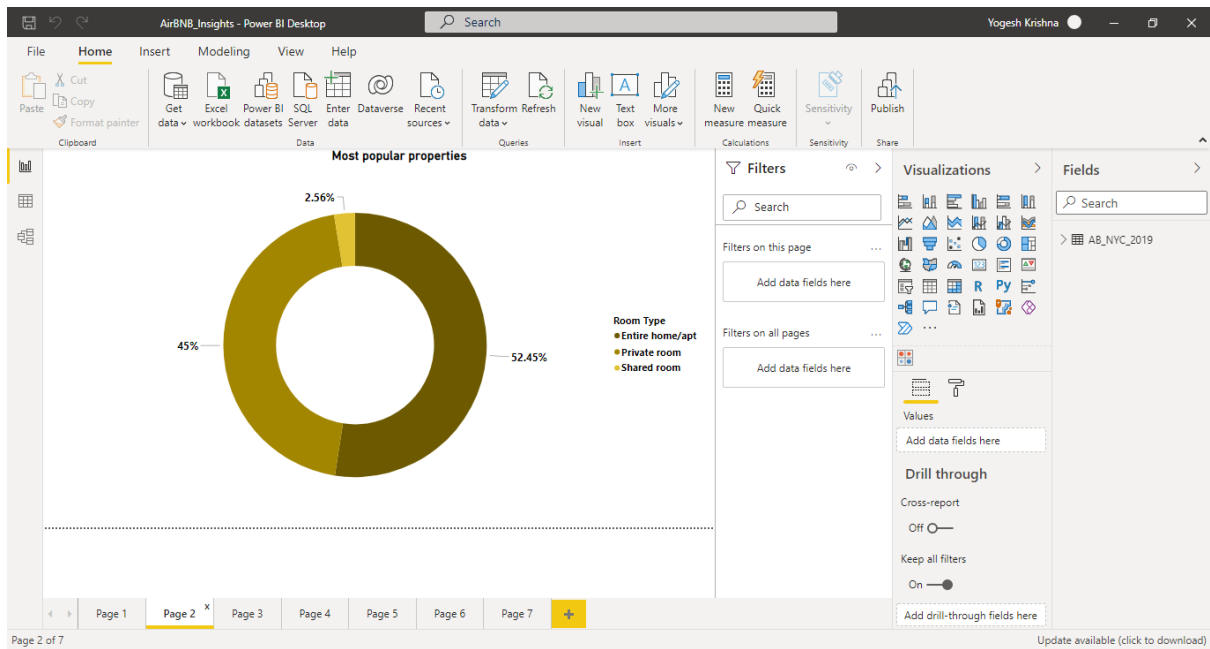
Data Visualization:

All the visualizations were created in the PowerBI report pane. Below are a few of the screenshots for reference:

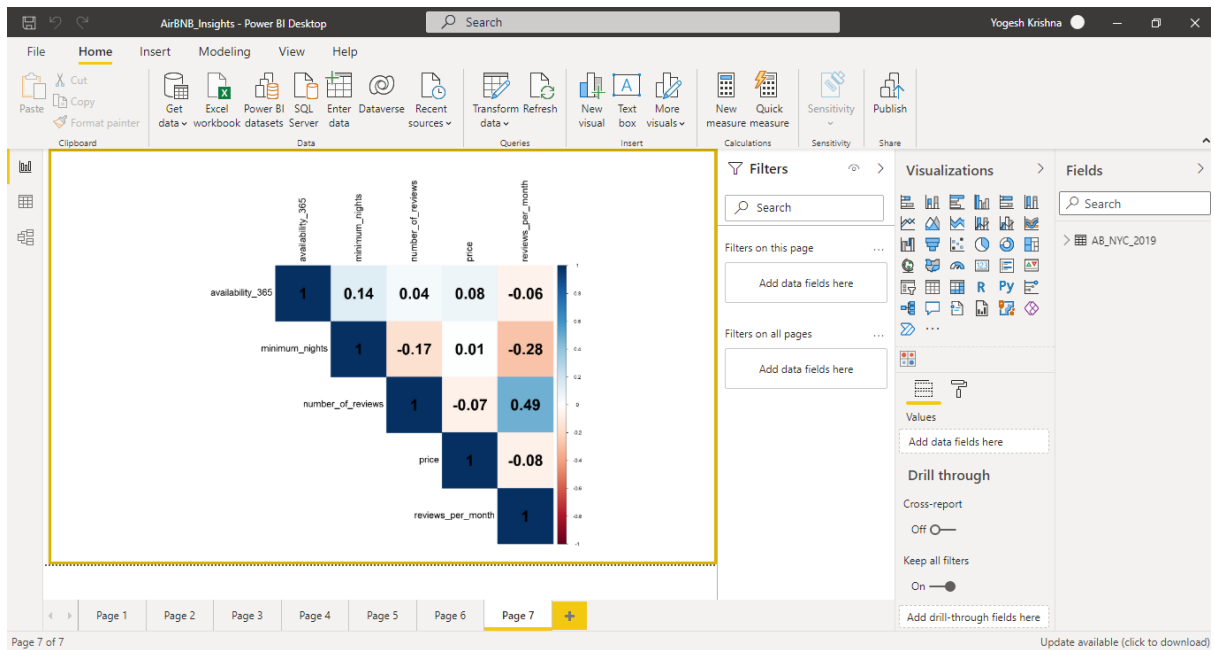
- Barchart for Most popular localities:



- Donut Chart for popular properties:



- Correlation Matrix:



- Pie Chart for price range:

