

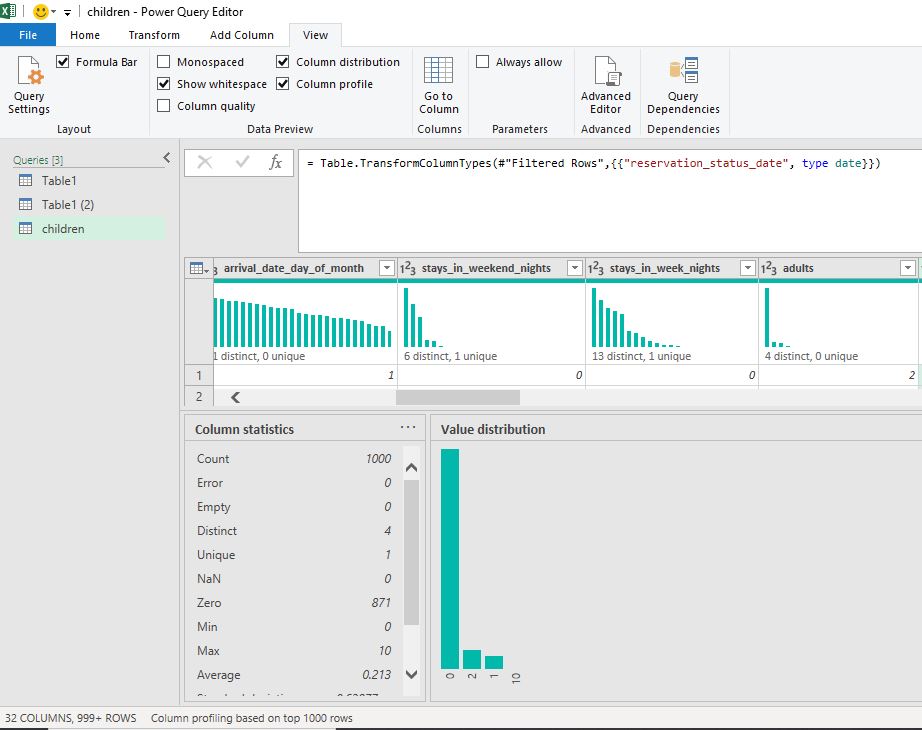
Hotel Booking Dataset Cleaning Using Power Query

**Introduction**

Booking a staff in a hotel allows a guest to have a reservation and accommodation which it has to be booked either through phone call or through an online system. This dataset could have some irregularities and unorganized structure that won’t be fit for the analysis. I am applying my data analysis skill in the cleaning of hotel dataset using power query in excel. During my data quality assessment, I have identified columns with missing values, fixing errors and data that needs to be formatted to datatypes with standardization.

**Data Assessment:**

Data assessment was carried out by checking through the columns one after the other, to do this the view tab at the power query editor allows you to check the box beside the column distribution, column profile and show with reference. This help to view the statistics of the column and also know the empty and null cells available in the columns or rows.

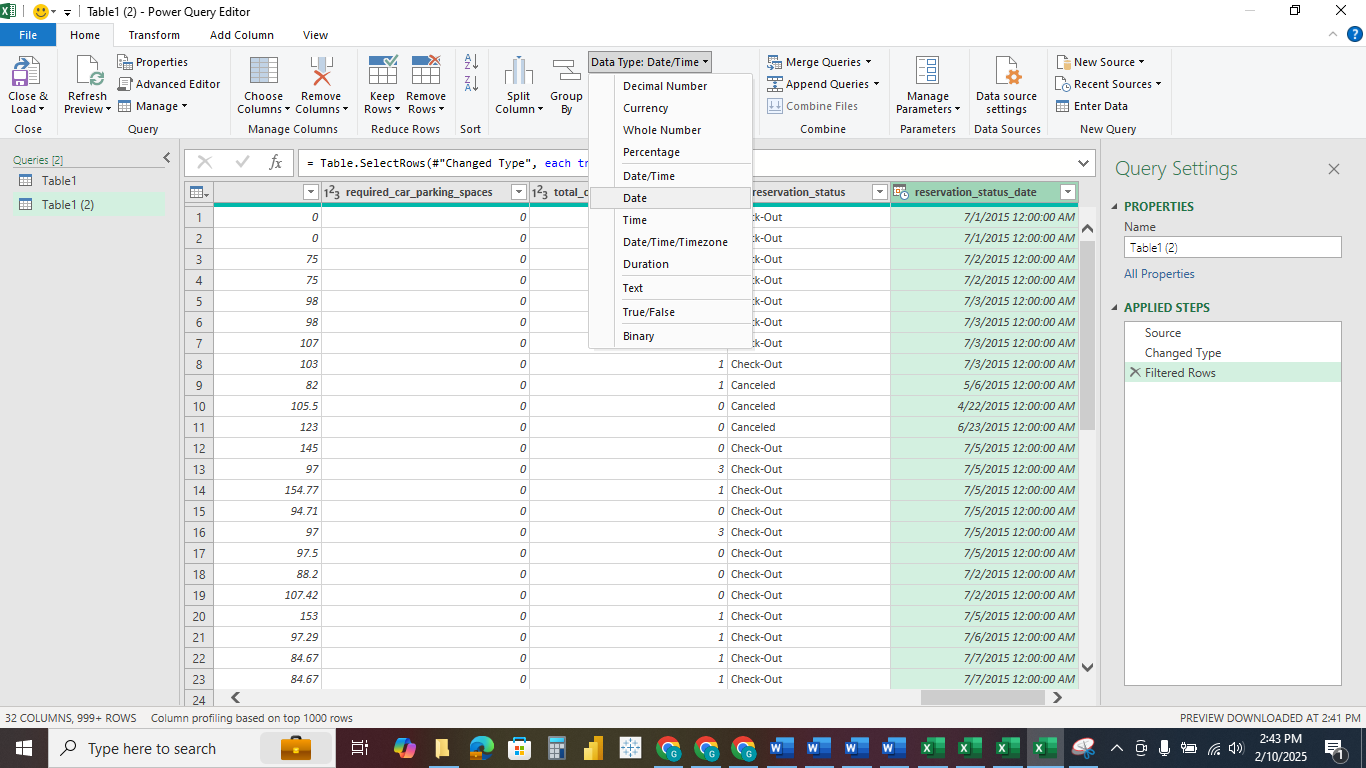
****

**Handling Missing Values:**

To ensure more clarity and accurate analysis, rows containing null values in the dataset were removed. The Rows with Null Values are In Children, Country, and Agent Columns.

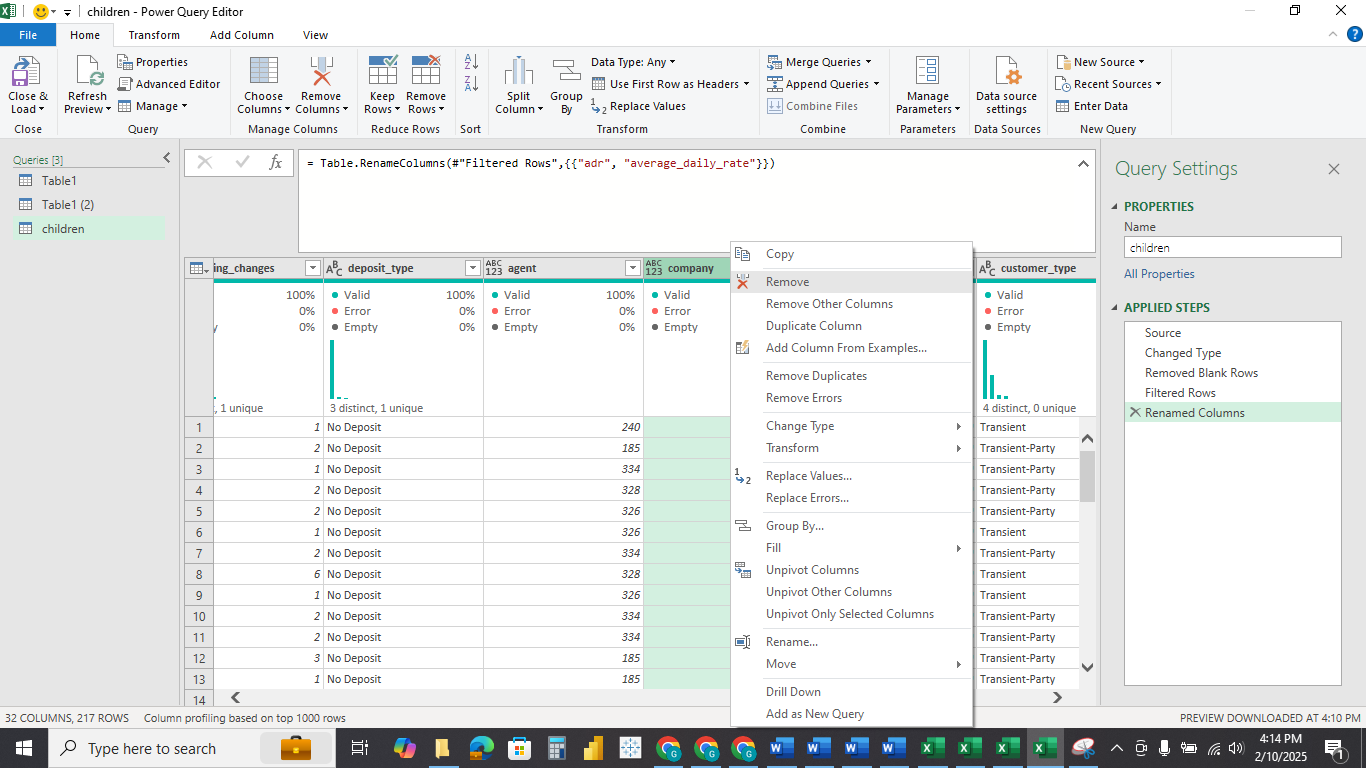
**Column Formatting:**

During the checks conducted on the datasets I discovered the column reservation\_status\_date was not in date format datatype which its needs to be change into correct data format.



**Drop Columns:**

Identification of columns that needs to be dropped from the data; The application of a technical strategy was adopted. The company column as its null values which takes up to 70% of the rows was first dropped for the data since it is not relevant and could affect the quality of the data analysis that will be carried out.

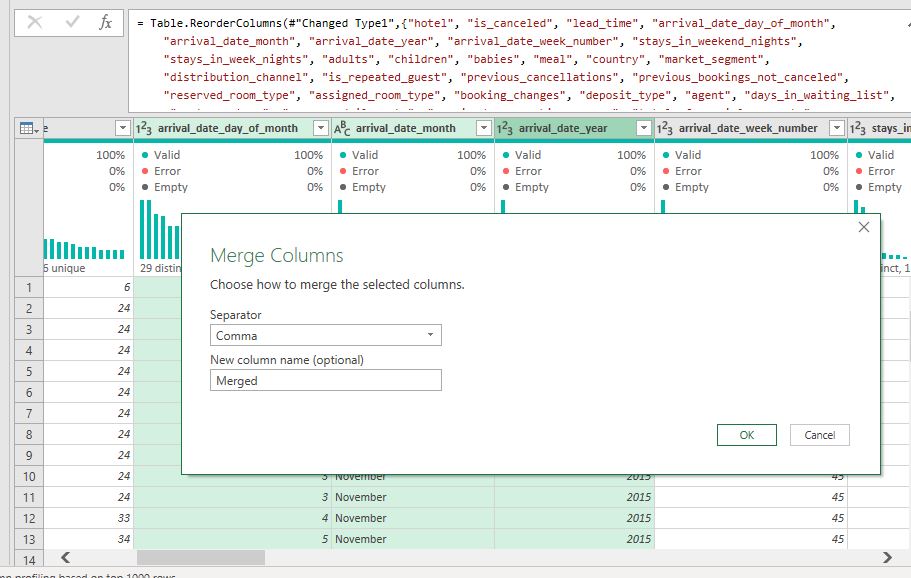


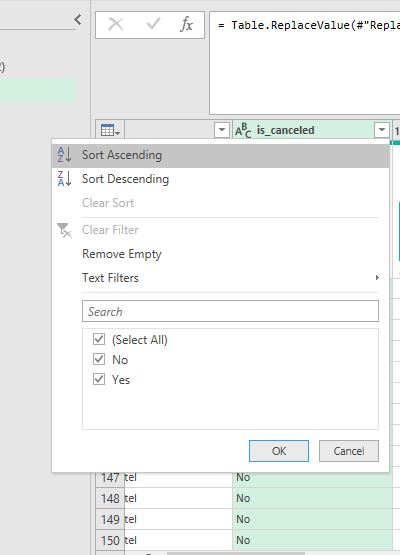
The columns arrival\_date\_year, arrival\_date\_month, arrival\_date\_day\_of\_month was also dropped after it was merged into one column. As I was able to merged and covert it into datatype, having these columns may still make the datasets to be messy.

**Find and Replace:**

To ensure that data is being interpreted well, some columns with numbers needs to be replaced with either true or false or yes or no but to be more explicit and for the stakeholders to have a better clarity when interpreting the insights. I converted the column is\_canceled with numbers to be replaced to ‘Yes' or 'No’, Change values in is\_repeated\_guest to 'Yes' or 'No' and Change values in required\_car\_parking\_spaces to 'Yes' or 'No'.

The first step taken is to convert the format of the column to text datatype before it could accept the replacement to Yes or No. after this was done, I then replaced the number from 0 and 1 to Yes or No as expected.



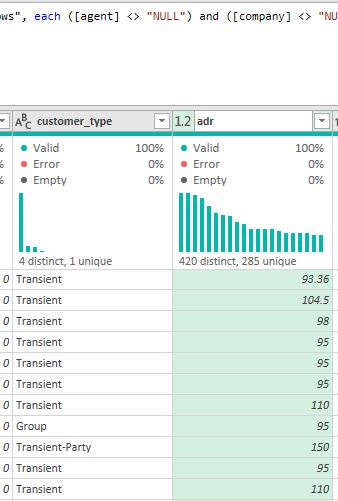


**Column Merge:**

As I was able to merge the columns (arrival\_date\_year, arrival\_date\_month, arrival\_date\_day\_of\_month) together into one arrival data column this will help when carrying out analysis on the days, and months having the highest bookings and presenting more clearer view of the importance of the date type in the datasets because for every customer that need an accommodation their date booking is very crucial ibn determining the availability of reservation in a hotel.

**Rename Column and reposition:**

The description of some of the data columns was strictly checked to know the right standard they need to be and I discovered that the column for average daily rate was abbreviated to be adr as it wasn’t in a professional format with the other columns. It was renamed while arrival\_date was also repositioned to the right side of lead\_time. This was done in order to get the data ready for proper analysis and arrangement.



**Conclusion:**

To conclude on the process, Power Query have been an excellent tool in Microsoft excel for cleaning datasets as it is equipped with various dimensions of functionality. The hotel booking dataset was carefully cleaned with clear thoughtful decision making ensuring that it was prepared for further analysis. My experience in data handling with missing values was skilled up by ensuring risk of using unstructured datasets for analysis is prevented.