</talentlabs>

Assignment 3 - Data Wrangling with Google Sheets

Instructions

1. Download the Airbnb NYC.csv data file and import it into Google Sheets.
2. Complete the tasks and answer the question
3. Please submit the assignment through TalentLabs Learning System. You will need to submit a zip file which contains this word document (with answers) and the data wrangled submission.csv file.

Question 1:
Organise the dataset and paste a screenshot of the last 3 columns and the first 3 rows below:

The first 3 rows:

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2539	Clean & quiet ap	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
2595	Skylit Midtown C	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
3647	THE VILLAGE C	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3	0			1	365

The last 3 rows:

36485431	Sunny Studio at	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire home/apt	115	10	0			1	27
36485609	43rd St. Time Sc	30985759	Taz	Manhattan	Hell's Kitchen	40.75751	-73.99112	Shared room	55	1	0			6	2
36487245	Trendy duplex in	68119814	Christophe	Manhattan	Hell's Kitchen	40.76404	-73.98933	Private room	90	7	0			1	23

Question 2:
Inspect the dataset. What data cleaning operations need to be carried out from the ones below? Select all that apply.

☒ A – Remove duplicate rows
☒ B – Handle missing values
☐ C – Correct data formats
☒ D – Drop irrelevant columns
☒ E – Fix inconsistent data entry
☒ F – Trim whitespaces
☐ G – Correct spelling errors
☐ H – Correct numerical errors

Question 3:
Columns M and N (last review and reviews per month) have missing values. Why do you think those values are missing? How would you suggest handling those missing values?

Both the 'last review' and 'reviews per month' columns display missing values because they're the posts that have not received comments since their publication or listing date.

For "last review" column, I will delete this column as it was irrelevant. It doesn't have any relationship with the subject matter of this project analysis: house rental price.

Meanwhile for "reviews per month" column, I will replace the missing values with 0, meaning it doesn't have any comments every month.

Question 4:
Clean the dataset and handle the missing values. Next, complete the following data enrichment operations:

1. Make all the ID column values 10 numbers in length (custom formatting – pad with 0s).
2. Make a new column. Name this column Clean. Search the Name column for the text "Clean". If the value is present, amend your new Clean column with the value 1. Else, 0.
3. Change the room type column, to only include one word - Private or Entire (instead of Private room and Entire home).
4. Filter the dataset for Brooklyn neighbourhood group only.

Once finished, export the dataset as a CSV file. Name the file submission.csv. Submit both this word document and the CSV file together in a zipped folder. Name the zip folder submission.zip.