Data Cleaning
on NYC-AirBnB 2019 Project

TAKEN FROM CASE STUDY
RevoU Mini Course - Data Analytics

Dhea Amalia Lutfiani

Case Study Instructions

QUESTION

Table of interest : NYC AirBnB Dummy Data

1. Look at this data and start thinking. List down 3 trends/points that you want to show.
2. From here, try to explore the data and make changes, filter, and prepare the data that you need.
3. Create some visualizations or dashboard with the best type of chart you have learned.

   The easiest is with Google Data Studio or Google Sheets.
4. Then, make 1-2 slides from the Graphs with the insights you got to present your findings to the stakeholders (read this article from HBR)

Python & Data Cleaning

Info from the Dataset :

```
[2]  airbnb = pd.read_csv("AB_NYC_2019.csv")
     airbnb.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
```

From this information, it can be seen that:

- There are 16 columns and the amount of data varies from 38843 to 48895
  - > 38843 in the last_review and reviews_per_month columns
  - > 48874 in the hostname column
  - > 48879 in name column
- In the last_review column the data type is object, whereas in the data table it contains date

- Therefore it is necessary to do data cleaning because the data in the name and host_name columns cannot be empty, considering that there can be no AirBnB transactions if there is no name ordering.
- While in the last_review and reviews_per_month columns the occurrence of empty data is normal because of the nature of the review that is not mandatory in a transaction
- So we have to eliminate data that is empty and the number of data becomes 48874 data.

Python & Data Cleaning

Preview the Data before data cleaning:

```
[ ]  airbnb
```

	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	avai
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853	-73.94995	Private room	70	2	0	NaN	NaN	2	
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.70184	-73.93317	Private room	40	4	0	NaN	NaN	2	
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire home/apt	115	10	0	NaN	NaN	1	

Python & Data Cleaning

DATA CLEANING

Messy data is a common problem you'd likely face when you have data from sources like spreadsheets. It is important to clean your data before doing any analysis.

Things to do in data cleaning :

1. Change data type
2. Remove duplicated data
3. Remove empty data
4. Remove outliers
5. Remove unnecessary data

Cleaning Data : Python (Google Collab)

Things to do in data cleaning :

1. Menyesuaikan tipe data per-kolom
2. Menghilangkan nilai kosong
3. Menghilangkan outliers
4. Cek lagi dalam perjalanan (selama step by step) cleaning data

Python & Data Cleaning

DATA CLEANING

Messy data is a common problem you'd likely face when you have data from sources like spreadsheets. It is important to clean your data before doing any analysis.

Things to do in data cleaning :

1. Change data type : last_review column become datetime, latitude and longitude colume become objects
2. Remove duplicated data
3. Remove empty data : Delete null data so the number of data becomes 48858
4. Remove outliers : Delete data that has a value that exceeds the upper and lower limits (data outliers) in the 'price' column, so that the total data becomes 45882
5. Remove unnecessary data

Python & Data Cleaning

Preview the Data after data cleaning:

	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	avail
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3	0	NaT	NaN	1	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853	-73.94995	Private room	70	2	0	NaT	NaN	2	
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.70184	-73.93317	Private room	40	4	0	NaT	NaN	2	
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire home/apt	115	10	0	NaT	NaN	1	

Defining the Problem

1. What is AirBnB largest segment of rented properties?
2. How many properties are there in each borough?
3. How is the rental price distribution for each room type?
4. How many property does each host rents via AirBnB?
5. How is the trend of property review? Are there any properties that's left too long without any price review?

THANK YOU