Final Project

Pre-requisite

- Understanding of Python
- Understanding of Data Cleaning

Level of Exercise: Beginner

Duration: 4 hours

Project Details:

Objective:

In this exericise, you will be working an Open Dataset dataset coming from Airbnb. Some of the tasks include

- Data Cleaning.
- Data Transformation

Overview of Airbnb Data:

People's main criteria when visiting new places are reasonable accommodation and food. Airbnb (Air-Bed-Breakfast) is an online marketplace created to meet this need of people by renting out their homes for a short term. They offer this facility at a relatively lower price than hotels. Further people worldwide prefer the homely and economical service offered by them. They offer services across various geographical locations

Dataset Source

You can get the dataset for this assessment using the following link:
https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata

This dataset contains information such as the neighborhood offering these services, room type, price,avaliabilty, reviews, service fee, cancellation policy and rules to use the house. This analysis will help airbnb in improving its services.

So all the best for your Data Journey on Airbnb data!!!

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```

Task 1: Data Loading (Python)

1. Read the csv file and load it into a pandas dataframe.
2. Display the first five rows of your dataframe.
3. Display the data types of the columns.

```
## Import Libraries
import pandas as pd
import numpy as np


## Read the csv file
readCSV = pd.read_csv('/Users/sandilyamekala/Downloads/Airbnb_Open_Data.csv', low
```

```
df = pd.DataFrame(data=readCSV)
print(df)
```

```
...                    ...       ...       ...            ...  ...
102594         Williamsburg  40.70862 -73.94651  United States  ...
102595  Morningside Heights  40.80460 -73.96545  United States  ...
102596           Park Slope  40.67505 -73.98045  United States  ...
102597     Long Island City  40.74989 -73.93777  United States  ...
102598      Upper West Side  40.76807 -73.98342  United States  ...

       service fee minimum nights number of reviews last review  \
0             $193           10.0               9.0  10/19/2021
1              $28           30.0              45.0   5/21/2022
2             $124            3.0               0.0         NaN
3              $74           30.0             270.0    7/5/2019
4              $41           10.0               9.0  11/19/2018
...             ...            ...               ...         ...
102594        $169            1.0               0.0         NaN
102595        $167            1.0               1.0    7/6/2015
102596        $198            3.0               0.0         NaN
102597        $109            2.0               5.0  10/11/2015
102598        $206            1.0               0.0         NaN

        reviews per month review rate number calculated host listings count  \
0                    0.21                4.0                             6.0
1                    0.38                4.0                             2.0
2                     NaN                5.0                             1.0
3                    4.64                4.0                             1.0
4                    0.10                3.0                             1.0
...                   ...                ...                             ...
```

```
102594                NaN                3.0                            1.0
102595               0.02                2.0                            2.0
102596                NaN                5.0                            1.0
102597               0.10                3.0                            1.0
102598                NaN                3.0                            1.0

        availability 365                                        house_rules  \
0                 286.0  Clean up and treat the home the way you'd like...
1                 228.0  Pet friendly but please confirm with me if the...
2                 352.0  I encourage you to use my kitchen, cooking and...
3                 322.0                                                NaN
4                 289.0  Please no smoking in the house, porch or on th...
...                 ...                                                ...
102594            227.0  No Smoking No Parties or Events of any kind Pl...
102595            395.0  House rules: Guests agree to the following ter...
102596            342.0                                                NaN
102597            386.0                                                NaN
102598             69.0                                                NaN

       license
0          NaN
1          NaN
2          NaN
3          NaN
4          NaN
...        ...
102594     NaN
102595     NaN
102596     NaN
102597     NaN
102598     NaN

[102599 rows x 26 columns]
```

```
## Display the first 5 rows
print(df.head())
```

```
         id                                                 NAME      host id  \
0  1001254                Clean & quiet apt home by the park  80014485718
1  1002102                           Skylit Midtown Castle  52335172823
2  1002403              THE VILLAGE OF HARLEM....NEW YORK !  78829239556
3  1002755                                              NaN  85098326012
4  1003689  Entire Apt: Spacious Studio/Loft by central park  92037596077

  host_identity_verified host name neighbourhood group neighbourhood  \
0            unconfirmed  Madaline            Brooklyn    Kensington
1               verified     Jenna           Manhattan       Midtown
2                    NaN     Elise           Manhattan        Harlem
3            unconfirmed     Garry            Brooklyn  Clinton Hill
4               verified    Lyndon           Manhattan   East Harlem

        lat      long        country  ... service fee minimum nights  \
0  40.64749 -73.97237  United States  ...       $193            10.0
1  40.75362 -73.98377  United States  ...        $28            30.0
2  40.80902 -73.94190  United States  ...       $124             3.0
3  40.68514 -73.95976  United States  ...        $74            30.0
4  40.79851 -73.94399  United States  ...        $41            10.0

  number of reviews last review  reviews per month review rate number  \
0               9.0  10/19/2021                0.21                4.0
1              45.0   5/21/2022                0.38                4.0
2               0.0         NaN                 NaN                5.0
3             270.0    7/5/2019                4.64                4.0
4               9.0  11/19/2018                0.10                3.0

  calculated host listings count  availability 365  \
0                            6.0             286.0
1                            2.0             228.0
2                            1.0             352.0
3                            1.0             322.0
4                            1.0             289.0

                                        house_rules license
0  Clean up and treat the home the way you'd like...     NaN
1  Pet friendly but please confirm with me if the...     NaN
2  I encourage you to use my kitchen, cooking and...     NaN
3                                                NaN     NaN
4  Please no smoking in the house, porch or on th...     NaN

[5 rows x 26 columns]
```

```
## Display the data types
print(df.dtypes)
```

```
id                                int64
NAME                             object
host id                           int64
host_identity_verified           object
host name                        object
neighbourhood group              object
neighbourhood                    object
lat                             float64
long                            float64
country                          object
country code                     object
instant_bookable                 object
cancellation_policy              object
room type                        object
Construction year               float64
price                            object
service fee                      object
minimum nights                  float64
number of reviews               float64
last review                      object
reviews per month               float64
review rate number              float64
calculated host listings count  float64
availability 365                float64
house_rules                      object
license                          object
dtype: object
```

Task 2a: Data Cleaning

1. Drop some of the unwanted columns. These include `host id`, `id`, `country` and `country code` from the dataset.

Please include the code in the cells below.

```
df = df.drop(columns=['host id', 'id', 'country', 'country code'])
```

```
print(df)
```

```
                                                NAME  \
```

```
0                    Clean & quiet apt home by the park
1                                 Skylit Midtown Castle
2                  THE VILLAGE OF HARLEM....NEW YORK !
3                                                  NaN
4       Entire Apt: Spacious Studio/Loft by central park
...                                                 ...
102594                        Spare room in Williamsburg
102595                   Best Location near Columbia U
102596                  Comfy, bright room in Brooklyn
102597                Big Studio-One Stop from Midtown
102598                             585 sf Luxury Studio

       host_identity_verified    host name neighbourhood group  \
0                  unconfirmed     Madaline            Brooklyn
1                     verified        Jenna           Manhattan
2                          NaN        Elise           Manhattan
3                  unconfirmed        Garry            Brooklyn
4                     verified       Lyndon           Manhattan
...                        ...          ...                 ...
102594                verified         Krik            Brooklyn
102595             unconfirmed        Mifan           Manhattan
102596             unconfirmed        Megan            Brooklyn
102597             unconfirmed  Christopher              Queens
102598             unconfirmed      Rebecca           Manhattan

              neighbourhood       lat      long instant_bookable  \
0                 Kensington  40.64749 -73.97237            False
1                    Midtown  40.75362 -73.98377            False
2                     Harlem  40.80902 -73.94190             True
3               Clinton Hill  40.68514 -73.95976             True
4               East Harlem  40.79851 -73.94399            False
...                      ...       ...       ...              ...
102594          Williamsburg  40.70862 -73.94651            False
102595  Morningside Heights  40.80460 -73.96545             True
102596            Park Slope  40.67505 -73.98045             True
102597     Long Island City  40.74989 -73.93777             True
102598      Upper West Side  40.76807 -73.98342            False

       cancellation_policy        room type  ...  service fee minimum nights
0                   strict     Private room  ...        $193            10.0
1                 moderate  Entire home/apt  ...         $28            30.0
2                 flexible     Private room  ...        $124             3.0
3                 moderate  Entire home/apt  ...         $74            30.0
4                 moderate  Entire home/apt  ...         $41            10.0
...                    ...              ...  ...         ...             ...
102594            flexible     Private room  ...        $169             1.0
102595            moderate     Private room  ...        $167             1.0
102596            moderate     Private room  ...        $198             3.0
102597              strict  Entire home/apt  ...        $109             2.0
102598            flexible  Entire home/apt  ...        $206             1.0
```

```
        number of reviews  last review  reviews per month review rate number  \
0                     9.0   10/19/2021               0.21                 4.0
1                    45.0    5/21/2022               0.38                 4.0
2                     0.0          NaN                NaN                 5.0
3                   270.0     7/5/2019               4.64                 4.0
4                     9.0   11/19/2018               0.10                 3.0
```

```
print(df.dtypes)
```

```
NAME                             object
host_identity_verified           object
host name                        object
neighbourhood group              object
neighbourhood                    object
lat                             float64
long                            float64
instant_bookable                 object
cancellation_policy              object
room type                        object
Construction year               float64
price                            object
service fee                      object
minimum nights                  float64
number of reviews               float64
last review                      object
reviews per month               float64
review rate number              float64
calculated host listings count  float64
availability 365                float64
house_rules                      object
license                          object
dtype: object
```

Task 2b: Data Cleaning

- Check for missing values in the dataframe and display the count in ascending order. **If the values are missing, impute the values as per the datatype of the columns.**
- Check whether there are any duplicate values in the dataframe and, if present, remove them.
- Display the total number of records in the dataframe before and after removing the duplicates.

```
## Check for missing values in the dataframe and display the count in ascending o
print(df.isnull().sum().sort_values(ascending=True))
```

```
room type                           0
lat                                 8
long                                8
neighbourhood                      16
neighbourhood group                29
cancellation_policy                76
instant_bookable                  105
number of reviews                 180
Construction year                 214
price                             247
NAME                              250
service fee                       273
host_identity_verified            289
calculated host listings count    308
review rate number                315
host name                         392
minimum nights                    398
availability 365                  448
reviews per month                15371
last review                      15385
house_rules                      50794
license                          99161
dtype: int64
```

```
## Check whether there are any duplicate values in the dataframe and if present r
df.drop_duplicates(inplace=True)
print(df)
```

```
        ...                 ...       ...       ...              ...
102053            Flatbush  40.64945 -73.96108             True
102054            Bushwick  40.69872 -73.92718            False
102055  Bedford-Stuyvesant  40.67810 -73.90822             True
102056              Harlem  40.81248 -73.94317             True
102057              Harlem  40.81315 -73.94747            False

       cancellation_policy        room type  ...  service fee minimum nights
0                   strict     Private room  ...         $193           10.0
1                 moderate  Entire home/apt  ...          $28           30.0
2                 flexible     Private room  ...         $124            3.0
3                 moderate  Entire home/apt  ...          $74           30.0
4                 moderate  Entire home/apt  ...          $41           10.0
...                    ...              ...  ...          ...            ...
102053            moderate     Private room  ...          NaN            7.0
102054            flexible     Private room  ...          NaN            1.0
102055            moderate  Entire home/apt  ...          NaN            2.0
102056              strict     Private room  ...          NaN            2.0
```

```
102057            flexible  Entire home/apt  ...          NaN            4.0

       number of reviews  last review  reviews per month review rate number  \
0                    9.0   10/19/2021               0.21                4.0
1                   45.0    5/21/2022               0.38                4.0
2                    0.0          NaN                NaN                5.0
3                  270.0    7/5/2019               4.64                4.0
4                    9.0   11/19/2018               0.10                3.0
...                  ...          ...                ...                ...
102053              12.0    3/27/2019               0.44                5.0
102054              19.0    8/31/2017               0.72                3.0
102055              50.0    6/26/2019               3.12                4.0
102056               0.0          NaN                NaN                1.0
102057              22.0    6/15/2019               0.85                4.0

        calculated host listings count  availability 365  \
0                                  6.0             286.0
1                                  2.0             228.0
2                                  1.0             352.0
3                                  1.0             322.0
4                                  1.0             289.0
...                                ...               ...
102053                             1.0               0.0
102054                             2.0               0.0
102055                             2.0             235.0
102056                             1.0               0.0
102057                             1.0             238.0

                                           house_rules  license
0       Clean up and treat the home the way you'd like...      NaN
1       Pet friendly but please confirm with me if the...      NaN
2       I encourage you to use my kitchen, cooking and...      NaN
3                                                     NaN      NaN
4       Please no smoking in the house, porch or on th...      NaN
...                                                   ...      ...
102053  Shoes off Clean After yourself Turn Lights and...      NaN
102054                                          #NAME?     NaN
102055  * Check out: 10am * We made an effort to keep ...      NaN
102056  Each of us is working and/or going to school a...      NaN
102057  Please remember that this is a residential bui...      NaN

[99163 rows x 22 columns]
```

```
## Display the total number of records in the dataframe after removing the duplic
print(len(df))
```

```
99163
```

Task 3: Data Transformation

- Rename the column `availability 365` to `days_booked`
- Convert all column names to lowercase and replace the spaces in the column names with an underscore "_".

Please include the code in the cells below.

```
## Rename the column.
df = df.rename(columns={'availability 365': 'days_booked'})
print(df.keys())
```

```
Index(['name', 'host_identity_verified', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'lat', 'long', 'instant_bookable',
       'cancellation_policy', 'room_type', 'construction_year', 'price',
       'service_fee', 'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'review_rate_number',
       'calculated_host_listings_count', 'days_booked', 'house_rules',
       'license'],
      dtype='object')
```

```
## Convert all column names to lowercase and replace the spaces with an underscor
df.columns = df.columns.str.lower().str.replace(' ', '_')
print(df.columns)
```

```
Index(['name', 'host_identity_verified', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'lat', 'long', 'instant_bookable',
       'cancellation_policy', 'room_type', 'construction_year', 'price',
       'service_fee', 'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'review_rate_number',
       'calculated_host_listings_count', 'days_booked', 'house_rules',
       'license'],
      dtype='object')
```

Task 4: Exploratory Data Analysis

- List the count of various room types avaliable in the dataset.
- Which room type has the most strict cancellation policy?

Please include the code in the cells below.

```
## List the count of various room types avaliable with Airbnb
print(df['room_type'].value_counts())
```

```
room_type
Entire home/apt    52003
Private room       44895
Shared room         2150
Hotel room           115
Name: count, dtype: int64
```

```
## Which room type adheres to more strict cancellation policy
cancellation = df[df['cancellation_policy'] == 'strict']
print(cancellation[['room_type', 'cancellation_policy']])
```

```
              room_type cancellation_policy
0          Private room              strict
8          Private room              strict
9          Private room              strict
12         Private room              strict
24         Private room              strict
...                 ...                 ...
102037     Private room              strict
102040     Private room              strict
102042     Private room              strict
102049  Entire home/apt              strict
102056     Private room              strict

[32930 rows x 2 columns]
```