```
#OIBSIP Data Analytics
#level1 task no.:-3- Cleaning Data
#Intern name:- Sanjana Gidwani
!pip install pandas numpy matplotlib seaborn
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (2.2.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.12/dist-packages (2.0.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.12/dist-packages (3.10.0)
Requirement already satisfied: seaborn in /usr/local/lib/python3.12/dist-packages (0.13.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (1.3.3)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (4.60.1)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (1.4.9)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (25.0)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (3.2.5)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pandas) (1.17
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
airbnb_file_path='/content/AB_NYC_2019.csv.zip'
airbnb_data=pd.read_csv(airbnb_file_path)
airbnb_data.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimu
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	

```
youtube_file_paths=[
    '/content/CAvideos.csv.zip',
    '/content/DEvideos.csv.zip',
    '/content/FRvideos.csv.zip',
    '/content/GBvideos.csv.zip',
    '/content/INvideos.csv.zip',
    '/content/JPvideos.csv.zip',
    '/content/KRvideos.csv.zip',
    '/content/MXvideos.csv.zip',
    '/content/RUvideos.csv.zip',
    '/content/USvideos.csv.zip'

]
youtube_dataframes= [pd.read_csv(file_path,encoding='ISO-8859-1')for file_path in youtube_file_paths]
youtube_data= pd.concat(youtube_dataframes, ignore_index=True)
youtube_data.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_time	t
0	n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. BeyoncÃ©	EminemVEVO	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/
1	0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	plush|"bad unboxing"|"unboxing"|"fan mail"|'
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bac
3	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you
4	2Vv-BfVoq4g	17.14.11	Ed Sheeran - Perfect (Official Music Video)	Ed Sheeran	10	2017-11-09T11:04:14.000Z	edsheeran|"ed sheeran"|"acoustic"|"live"|"co

```
print("Airbnb Dataser Information:")
print(airbnb_data.info())
print("\nYouTube Dataset Information:")
print(youtube_data.info())
```

```
Airbnb Dataser Information:
<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
None

YouTube Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 375942 entries, 0 to 375941
Data columns (total 16 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   video_id                375942 non-null  object
 1   trending_date           375942 non-null  object
 2   title                   375942 non-null  object
 3   channel_title           375942 non-null  object
 4   category_id             375942 non-null  int64
 5   publish_time            375942 non-null  object
 6   tags                    375942 non-null  object
 7   views                   375942 non-null  int64
 8   likes                   375942 non-null  int64
 9   dislikes                375942 non-null  int64
 10  comment_count           375942 non-null  int64
 11  thumbnail_link          375942 non-null  object
 12  comments_disabled       375942 non-null  bool
 13  ratings_disabled        375942 non-null  bool
 14  video_error_or_removed  375942 non-null  bool
 15  description             356464 non-null  object
dtypes: bool(3), int64(5), object(8)
memory usage: 38.4+ MB
```

```
None
```

```
#missing data
missing_airbnb=airbnb_data.isnull().sum()
print("Missing Data in Airbnb Dataset:")
print(missing_airbnb)
print("columns in airbnb dataframes",airbnb_data.columns)
airbnb_data['last_review']=airbnb_data['last_review'].fillna(airbnb_data['last_review'].mode()[0])
airbnb_data.drop(columns=['reviews_per_month'],inplace=True,errors='ignore')

missing_youtube=youtube_data.isnull().sum()
print("\nMissing Data in YouTube Dataset:")
print(missing_youtube)
print("columns in youtube dataframes",youtube_data.columns)
youtube_data['description']=youtube_data['description'].fillna('No Description')
```

```
Missing Data in Airbnb Dataset:
id                                 0
name                              16
host_id                            0
host_name                         21
neighbourhood_group                0
neighbourhood                      0
latitude                           0
longitude                          0
room_type                          0
price                              0
minimum_nights                     0
number_of_reviews                  0
last_review                        0
calculated_host_listings_count     0
availability_365                   0
dtype: int64
columns in airbnb dataframes Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'calculated_host_listings_count', 'availability_365'],
      dtype='object')

Missing Data in YouTube Dataset:
video_id                  0
trending_date             0
title                     0
channel_title             0
category_id               0
publish_time              0
tags                      0
views                     0
likes                     0
dislikes                  0
comment_count             0
thumbnail_link            0
comments_disabled         0
ratings_disabled          0
video_error_or_removed    0
description               0
dtype: int64
columns in youtube dataframes Index(['video_id', 'trending_date', 'title', 'channel_title', 'category_id',
       'publish_time', 'tags', 'views', 'likes', 'dislikes', 'comment_count',
       'thumbnail_link', 'comments_disabled', 'ratings_disabled',
       'video_error_or_removed', 'description'],
      dtype='object')
```

```
#now we remove duplicates
duplicates_airbnb=airbnb_data.duplicated().sum()
print("Duplicate Rows in Airbnb Dataset:",duplicates_airbnb)
airbnb_data.drop_duplicates(inplace=True)
duplicates_youtube=youtube_data.duplicated().sum()
print("Duplicate Rows in YouTube Dataset:",duplicates_youtube)
youtube_data.drop_duplicates(inplace=True)
```

```
Duplicate Rows in Airbnb Dataset: 0
Duplicate Rows in YouTube Dataset: 36417
```

```
#standardization

airbnb_data.colums=airbnb_data.columns.str.lower().str.replace(' ','_')
youtube_data.columns=youtube_data.columns.str.lower().str.replace(' ','_')
```

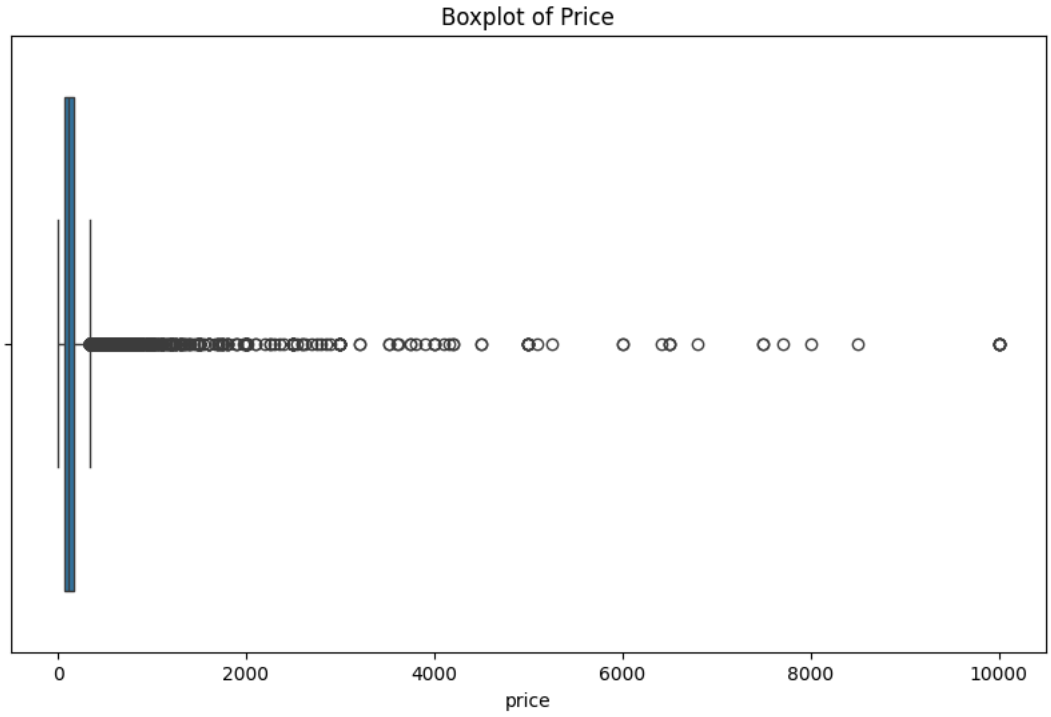
```
/tmp/ipython-input-2563928054.py:3: UserWarning: Pandas doesn't allow columns to be created via a new attribute name - see http
  airbnb_data.colums=airbnb_data.columns.str.lower().str.replace(' ','_')
```

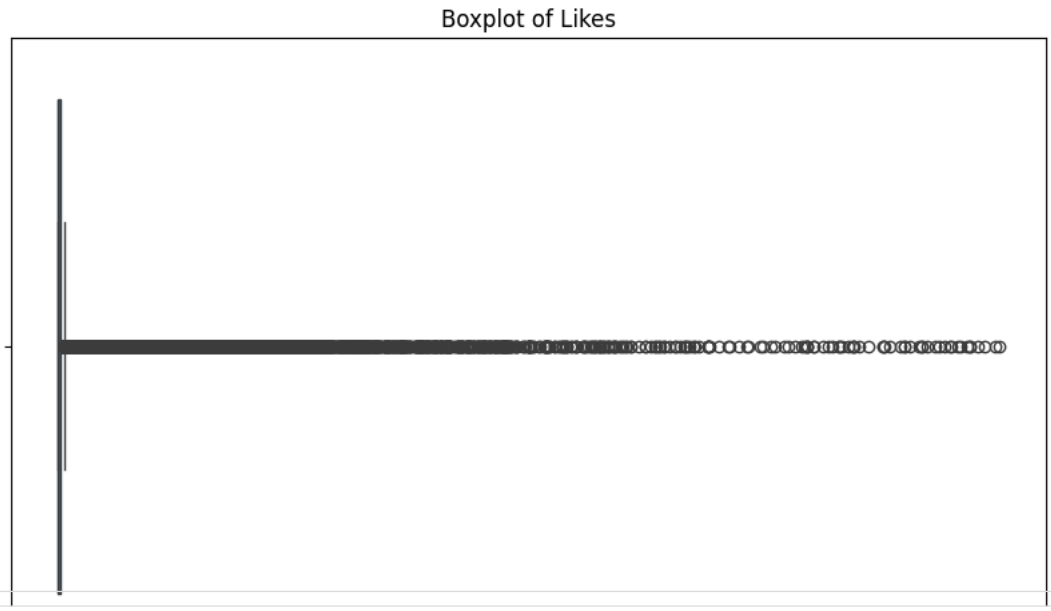
```
#outlier
plt.figure(figsize=(10,6))
sns.boxplot(x=airbnb_data['price'])
plt.title('Boxplot of Price')
plt.show()

q1=airbnb_data['price'].quantile(0.25)
q3=airbnb_data['price'].quantile(0.75)
iqr=q3-q1
airbnb_data = airbnb_data[(airbnb_data['price'] >= q1 - 1.5 * iqr) & (airbnb_data['price'] <= q3 + 1.5 * iqr)]

#outlier for likes in youtube dataset
plt.figure(figsize=(10,6))
sns.boxplot(x=youtube_data['likes'])
plt.title('Boxplot of Likes')
plt.show()

q1=youtube_data['likes'].quantile(0.25)
q3=youtube_data['likes'].quantile(0.75)
iqr=q3-q1
youtube_data = youtube_data[(youtube_data['likes'] >= q1 - 1.5 * iqr) & (youtube_data['likes'] <= q3 + 1.5 * iqr)]
```






```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
```

```
from sklearn.preprocessing import StandardScaler

airbnb_data=pd.DataFrame({
    'price':[100,150,200,300,400,500,600,1200,2000,3000],
    'other_feature':[1,2,1.5,2.5,3,3.5,4,2,3,4]

})
youtube_data=pd.DataFrame({
    'likes':[100,150,200,300,400,500,600,1200,2000,3000],
    'other_feature':[1,2,1.5,2.5,3,3.5,4,2,3,4]

})

def visualize_and_remove_outliers(df, feature, dataset_name):
  plt.figure(figsize=(10, 6))
  sns.boxplot(x=df[feature])
  plt.title(f'Boxplot of {feature.capitalize()} in {dataset_name} Dataset')
  plt.show()

  q1=df[feature].quantile(0.25)
  q3=df[feature].quantile(0.75)
  iqr=q3-q1
  filtered_df=df[(df[feature] >= q1 - 1.5 * iqr) & (df[feature] <= q3 + 1.5 * iqr)]
  return filtered_df
  #K-Means clustering
  scaler=StandardScaler()
  scaled_data=scaler.fit_transform(filtered_df[[feature]])
  KMeans=KMeans(n_clusters=3,random_state=42)
  filtered_df['cluster']=KMeans.fit_predict(scaled_data)

  #clustered data
  plt.figure(figsize=(10,6))
  sns.scatterplot(data=filtered_df,x=feature,y='other_feature',hue='cluster',palette='Set2', s=100)
  plt.title(f'Clustered Data with {feature.capitalize()} in {dataset_name} Dataset')
  plt.xlabel(feature.capitalize())
  plt.ylabel('Other Feature')
  plt.legend()
  plt.show()
  return filtered_df

  airbnb_filtered=visualize_and_remove_outliers(airbnb_data,'price','Airbnb')
  youtube_filtered=visualize_and_remove_outliers(youtube_data,'likes','YouTube')
```

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