initial_eda

September 22, 2023

[ ]:
```
import os
import glob
import plotly.express as px
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[ ]:
```
# Get CSV files list from a folder
path = r'C:/Users/151ha/Downloads/DATA606/Data'
csv_files = glob.glob(os.path.join(path, "*.csv"))
```

[ ]:
```
df = pd.concat((pd.read_csv(f) for f in csv_files), ignore_index=True)
df.head()
```

[ ]:
```
        id                            listing_url       scrape_id last_scraped  \
0   108061   https://www.airbnb.com/rooms/108061  20230618033311   2023-06-18
1  4394761  https://www.airbnb.com/rooms/4394761  20230618033311   2023-06-18
2  4448604  https://www.airbnb.com/rooms/4448604  20230618033311   2023-06-18
3  4515772  https://www.airbnb.com/rooms/4515772  20230618033311   2023-06-18
4  4587394  https://www.airbnb.com/rooms/4587394  20230618033311   2023-06-18

        source                                                name  \
0  city scrape  Rental unit in Asheville · ★4.51 · 1 bedroom ·…
1  city scrape  Rental unit in Asheville · ★4.87 · 2 bedrooms …
2  city scrape  Guest suite in Asheville · ★4.77 · 1 bedroom ·…
3  city scrape  Home in Asheville · ★4.79 · 2 bedrooms · 2 bed…
4  city scrape  Cabin in Asheville · ★4.94 · 1 bedroom · 1 bed…

                                         description  \
0  Walk to town in ten minutes! Monthly rental in…
1  Clean, adorable space w/ small deck overlookin…
2  Our beautiful home is on a quiet hillside, 2 m…
3  2br,1b, bugalow in Downtown Asheville. Enjoy t…
4  CHANGE YOUR VIEW FOR WORK or RETIREMENT!<br />…

                              neighborhood_overview  \
0  I love my neighborhood! Its friendly, easy-goi…
1  The neighborhood is a friendly, mellow, safe, …
```

```
2  This home is perched on a hill with some views…
3  Close to resturants, grocery stores, mall, hos…
4  Homeland Park is a neighborhood built as a res…

                                        picture_url   host_id  …  \
0  https://a0.muscache.com/pictures/miso/Hosting-…    320564  …
1  https://a0.muscache.com/pictures/76195c9a-68ee…  20447538  …
2  https://a0.muscache.com/pictures/2da78921-28aa…   5656896  …
3  https://a0.muscache.com/pictures/56710558/bab1…  17363326  …
4  https://a0.muscache.com/pictures/57609340/7c75…  16339042  …

  review_scores_communication review_scores_location review_scores_value  \
0                        4.80                    4.84                4.49
1                        4.97                    4.88                4.85
2                        4.94                    4.89                4.89
3                        4.99                    4.62                4.69
4                        5.00                    4.83                4.92

  license instant_bookable calculated_host_listings_count  \
0     NaN                f                               2
1     NaN                t                               2
2     NaN                f                               2
3     NaN                f                               2
4     NaN                f                               1

  calculated_host_listings_count_entire_homes  \
0                                           2
1                                           2
2                                           2
3                                           2
4                                           1

  calculated_host_listings_count_private_rooms  \
0                                            0
1                                            0
2                                            0
3                                            0
4                                            0

  calculated_host_listings_count_shared_rooms reviews_per_month
0                                           0              0.64
1                                           0              3.12
2                                           0              5.54
3                                           0              0.73
4                                           0              1.55

[5 rows x 75 columns]
```

[ ]:
```
df.shape
```

[ ]:
```
(62330, 75)
```

[ ]:
```
df.columns
```

[ ]:
```
Index(['id', 'listing_url', 'scrape_id', 'last_scraped', 'source', 'name',
       'description', 'neighborhood_overview', 'picture_url', 'host_id',
       'host_url', 'host_name', 'host_since', 'host_location', 'host_about',
       'host_response_time', 'host_response_rate', 'host_acceptance_rate',
       'host_is_superhost', 'host_thumbnail_url', 'host_picture_url',
       'host_neighbourhood', 'host_listings_count',
       'host_total_listings_count', 'host_verifications',
       'host_has_profile_pic', 'host_identity_verified', 'neighbourhood',
       'neighbourhood_cleansed', 'neighbourhood_group_cleansed', 'latitude',
       'longitude', 'property_type', 'room_type', 'accommodates', 'bathrooms',
       'bathrooms_text', 'bedrooms', 'beds', 'amenities', 'price',
       'minimum_nights', 'maximum_nights', 'minimum_minimum_nights',
       'maximum_minimum_nights', 'minimum_maximum_nights',
       'maximum_maximum_nights', 'minimum_nights_avg_ntm',
       'maximum_nights_avg_ntm', 'calendar_updated', 'has_availability',
       'availability_30', 'availability_60', 'availability_90',
       'availability_365', 'calendar_last_scraped', 'number_of_reviews',
       'number_of_reviews_ltm', 'number_of_reviews_l30d', 'first_review',
       'last_review', 'review_scores_rating', 'review_scores_accuracy',
       'review_scores_cleanliness', 'review_scores_checkin',
       'review_scores_communication', 'review_scores_location',
       'review_scores_value', 'license', 'instant_bookable',
       'calculated_host_listings_count',
       'calculated_host_listings_count_entire_homes',
       'calculated_host_listings_count_private_rooms',
       'calculated_host_listings_count_shared_rooms', 'reviews_per_month'],
      dtype='object')
```

[ ]:
```
df.describe()
```

[ ]:
```
                 id     scrape_id       host_id  host_listings_count  \
count  6.233000e+04  6.233000e+04  6.233000e+04         62325.000000
mean   2.776374e+17  2.023080e+13  1.561000e+08           163.677465
std    3.761961e+17  1.360416e+08  1.625461e+08           712.815925
min    2.595000e+03  2.023062e+13  2.234000e+03             0.000000
25%    2.206646e+07  2.023063e+13  1.937045e+07             1.000000
50%    4.667852e+07  2.023091e+13  8.986868e+07             2.000000
75%    7.052017e+17  2.023091e+13  2.712750e+08            10.000000
max    9.733282e+17  2.023091e+13  5.355317e+08          4751.000000

       host_total_listings_count      latitude     longitude  accommodates  \
```

```
count               62325.000000  62330.000000  62330.000000  62330.000000
mean                  248.807541     40.469618    -74.283801      3.340671
std                  1016.574648      1.390300      2.433663      2.412306
min                     1.000000     35.422810    -82.691050      0.000000
25%                     1.000000     40.674400    -73.999957      2.000000
50%                     3.000000     40.726855    -73.956195      2.000000
75%                    14.000000     40.793048    -73.897806      4.000000
max                  9176.000000     42.400180    -70.996000     16.000000

       bathrooms      bedrooms  …  review_scores_cleanliness  \
count        0.0  39169.000000  …                47232.000000
mean         NaN      1.788072  …                    4.681624
std          NaN      1.116455  …                    0.501259
min          NaN      1.000000  …                    0.000000
25%          NaN      1.000000  …                    4.600000
50%          NaN      1.000000  …                    4.840000
75%          NaN      2.000000  …                    5.000000
max          NaN     50.000000  …                    5.000000

       review_scores_checkin  review_scores_communication  \
count           47216.000000                 47226.000000
mean                4.830801                    4.823333
std                 0.388710                    0.413004
min                 0.000000                    0.000000
25%                 4.830000                    4.820000
50%                 4.950000                    4.960000
75%                 5.000000                    5.000000
max                 5.000000                    5.000000

       review_scores_location  review_scores_value  \
count            47212.000000         47214.000000
mean                 4.741157             4.648531
std                  0.398710             0.479050
min                  0.000000             0.000000
25%                  4.660000             4.560000
50%                  4.860000             4.770000
75%                  5.000000             4.920000
max                  5.000000             5.000000

       calculated_host_listings_count  \
count                    62330.000000
mean                        32.526328
std                         95.972190
min                          1.000000
25%                          1.000000
50%                          2.000000
75%                          8.000000
```

```
max                          597.000000

       calculated_host_listings_count_entire_homes  \
count                                 62330.000000
mean                                     17.565875
std                                      68.967118
min                                       0.000000
25%                                       0.000000
50%                                       1.000000
75%                                       3.000000
max                                     597.000000

       calculated_host_listings_count_private_rooms  \
count                                  62330.000000
mean                                      14.846783
std                                       65.067625
min                                        0.000000
25%                                        0.000000
50%                                        0.000000
75%                                        2.000000
max                                      519.000000

       calculated_host_listings_count_shared_rooms  reviews_per_month
count                                  62330.000000       47685.000000
mean                                       0.062362           1.406294
std                                        0.602723           1.817841
min                                        0.000000           0.010000
25%                                        0.000000           0.180000
50%                                        0.000000           0.730000
75%                                        0.000000           2.060000
max                                       15.000000          79.820000

[8 rows x 39 columns]
```

[ ]:
```
df.isnull().sum()
```

[ ]:
```
id                                                0
listing_url                                       0
scrape_id                                         0
last_scraped                                      0
source                                            0
                                                 ..
calculated_host_listings_count                    0
calculated_host_listings_count_entire_homes       0
calculated_host_listings_count_private_rooms      0
calculated_host_listings_count_shared_rooms       0
reviews_per_month                             14645
```

```
Length: 75, dtype: int64
```

[ ]:
```
fig = px.scatter_mapbox(df, lat="latitude", lon="longitude",
                        color_continuous_scale=px.colors.cyclical.IceFire,␣
 ↪size_max=15, zoom=10, height=800)

fig.update_layout(mapbox_style="carto-positron")
fig.show()
```

[ ]:
```
numerical_features = ['accommodates', 'bedrooms', 'beds', 'price',␣
 ↪'number_of_reviews']
for feature in numerical_features:
    fig = px.histogram(df, x=feature, title=f'Distribution of {feature}')
    fig.show()
```

[ ]:
```
categorical_features = ['room_type', 'property_type', 'host_response_time',␣
 ↪'neighbourhood_group_cleansed']
for feature in categorical_features:
    fig = px.bar(df[feature].value_counts(), title=f'Distribution of {feature}')
    fig.show()
```

[ ]:
```
fig = px.box(df, x='room_type', y='price', title='Price Distribution by Room␣
 ↪Type')
fig.show()
```

[ ]:
```
df['host_since'] = pd.to_datetime(df['host_since'])
df['first_review'] = pd.to_datetime(df['first_review'])
df['last_review'] = pd.to_datetime(df['last_review'])
```

[ ]:
```
review_scores = ['review_scores_rating', 'review_scores_accuracy',␣
 ↪'review_scores_cleanliness',
                 'review_scores_checkin', 'review_scores_communication',␣
 ↪'review_scores_location', 'review_scores_value']
for score in review_scores:
    fig = px.histogram(df, x=score, title=f'Distribution of {score}')
    fig.show()
```

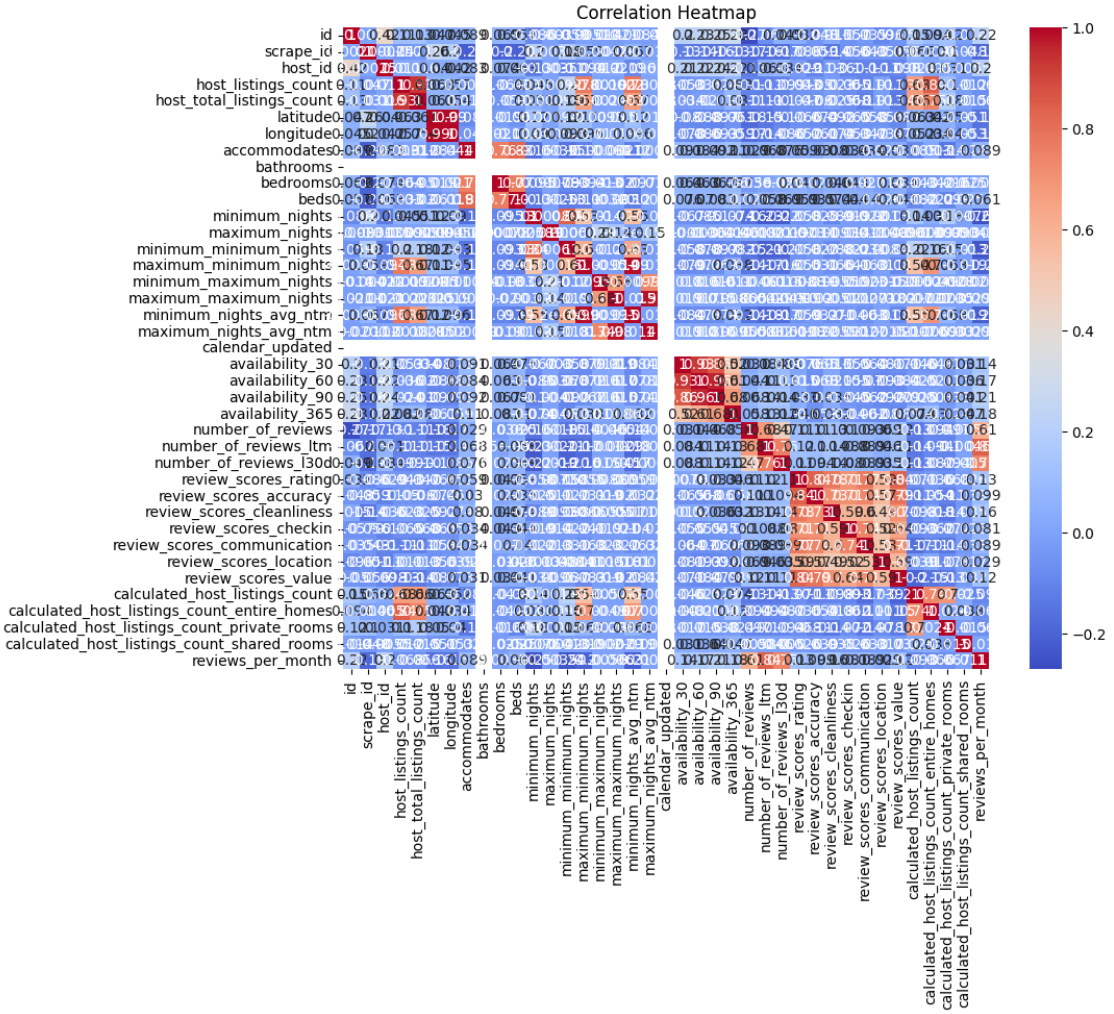
[ ]:
```
fig = px.scatter(df, x='host_listings_count',␣
 ↪y='calculated_host_listings_count', title='Host Listings vs. Calculated Host␣
 ↪Listings')
fig.show()
```

[ ]:
```
fig = px.scatter_geo(df, lat='latitude', lon='longitude', title='Geospatial␣
 ↪Distribution of Listings')
fig.show()
```

[ ]:
```
correlation_matrix = df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

```
C:\Users\151ha\AppData\Local\Temp\ipykernel_14680\3106231958.py:1:
FutureWarning:

The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
```



[ ]:
```
host_verifications = df['host_verifications'].str.replace('[\[\]\'\"]', '',␣
↪regex=True).str.split(', ')
```

```
host_verification_counts = host_verifications.explode().value_counts()
fig = px.bar(host_verification_counts, title='Host Verification Methods')
fig.show()
```

[ ]:
```
fig = px.histogram(df, x='minimum_nights', title='Distribution of Minimum␣
↪Nights')
fig.show()
```

[ ]:
```

```