In [32]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno

#Load the CSV into a Pandas DataFrame
airbnb_dirty= pd.read_csv('/Users/herlihpj/Desktop/Data Analytics/D214 - Capstone/San Diego Airbnb.csv',
                          index_col=0, low_memory=False)
pd.set_option('display.max_columns', 100)

#Remove rows where columns are completely misaligned
print(airbnb_dirty.shape)
#(13049, 107)
# Drop rows where the value in the offset data column is not null
airbnb_dirty = airbnb_dirty[airbnb_dirty['TO DROP'].isnull()]
#Drop the misc random columns added to the end of the dataframe
airbnb_dirty = airbnb_dirty.iloc[:, :-32]

print(airbnb_dirty.shape)
#(12795, 75)
print(airbnb_dirty.head())
#Dropping columns not relavant to the analysis
airbnb_dirty = airbnb_dirty.drop(columns= ['name','listing_url','summary', 'space','neighborhood_overview',
                                    'notes', 'transit','access','interaction','house_rules','thumbnail_url',
                                   'host_id', 'host_url','host_location', 'host_about','host_neighbourhood',
                                    'host_listings_count','host_total_listings_count', 'host_has_profile_pic',
                                   'host_acceptance_rate','neighbourhood_cleansed','market','smart_location',
                                    'country_code', 'country','square_feet','host_response_time','host_response_rate',
                                    'host_identity_verified','street','description','review_scores_rating',
                                    'review_scores_accuracy','review_scores_cleanliness', 'review_scores_checkin',
                                    'review_scores_communication', 'review_scores_location','review_scores_value',
                                    'requires_license','is_business_travel_ready','zipcode',
                                    'require_guest_profile_picture','require_guest_phone_verification','price_per_stay',
                                   'bed_type','minimum_nights','maximum_nights'])
print(airbnb_dirty.shape)
#(12795, 28)
#Display all columns in the dataset
print(airbnb_dirty.columns)
print(airbnb_dirty.info())

#Remove '$' & ',' from the following list of columns and convert to float
dollar_drop=['nightly_price', 'security_deposit','cleaning_fee','extra_people']
for i in dollar_drop:
```

```
    airbnb_dirty[i]=airbnb_dirty[i].str.strip() #replace('$','')
    airbnb_dirty[i]=airbnb_dirty[i].str.replace('$','')
    airbnb_dirty[i]=airbnb_dirty[i].str.replace(',','').astype(float)
    airbnb_dirty[i].fillna(0, inplace=True)
#Check to ensure $ and , were all removed and are type float
print(airbnb_dirty['nightly_price'].head(20))

# Drop rows where there are no stays
#med_dirty = med_dirty[med_dirty['number_of_stays'].astype(int)!=0]

# Drop rows where value for neighborhood was null and rename the column title
airbnb_dirty = airbnb_dirty[airbnb_dirty['neighbourhood'].notnull()]
airbnb_dirty.rename(columns = {'neighbourhood':'neighborhood'},inplace=True)
print(airbnb_dirty.shape)
#(12364, 32)

#City column Fix la Jollas, San Diegos, Ocean Beach empty spaces and different versions
#med_dirty['city']=med_dirty['city'].astype('string')
airbnb_dirty['city']=airbnb_dirty['city'].str.strip()
keys=['La Jolla', 'Ocean Beach', 'San Diego']
for key in keys:
    airbnb_dirty.loc[airbnb_dirty['city'].str.startswith(key), 'city'] =key

#Remove outliers for all numerical columns 2.5X the Standard Deviation
#num_cols = airbnb_dirty.select_dtypes(include=[np.number]).columns
num_cols = ['nightly_price']
for col in num_cols:
    mean = airbnb_dirty[col].mean()
    std = airbnb_dirty[col].std()
    upper_bound = mean + 2 * std
    lower_bound = mean - 2 * std
    airbnb_dirty = airbnb_dirty[(airbnb_dirty[col] < upper_bound) & (med_dirty[col] > lower_bound)]
print(airbnb_dirty.shape)
(8420, 28)

airbnb_dirty.to_csv('/Users/herlihpj/Desktop/Data Analytics/D214 - Capstone/Airbnb Prepared.csv')

print('======================','Cleaned Results','======================')
airbnb_cleaned=pd.read_csv('/Users/herlihpj/Desktop/Data Analytics/D214 - Capstone/Airbnb Prepared.csv', index_col=0)
print(airbnb_cleaned.info())
print(airbnb_cleaned.columns)
print(airbnb_cleaned.shape)
#(8420, 28)

#Visualize missingness
```

```
airbnb_cleaned_sorted=airbnb_cleaned.sort_values('number_of_stays')
msno.matrix(airbnb_cleaned_sorted)
```

```
(13049, 107)
(12795, 75)
             id                              listing_url  \
index
1      17138468  https://www.airbnb.com/rooms/17138468
2      21898446  https://www.airbnb.com/rooms/21898446
3      25948680  https://www.airbnb.com/rooms/25948680
4       1756516   https://www.airbnb.com/rooms/1756516
5      33395869  https://www.airbnb.com/rooms/33395869

                                                 name  \
index
1                                        NOT AVAILABLE
2                             Townhome in Pacific Beach
3      Spacious furnished 1 BR with tons of amenaties
4                             A Spacious luxury retreat
5                            Room for Rent in Mira Mesa

                                                 summary  \
index
1      AVAILABLE ONLY IN WINTER PRIME La Jolla Villag...
2      Hi! We are basically renting this master for a...
3      All my stuff will be gone. Dishwasher, washer/...
4      NIcely furnished. Great location, 2 blocks to ...
5      The room for rent is on a monthly basis.  It i...

                                                   space  \
index
1                                                    NaN
2                                                    NaN
3      Lobby provides free coffee. Building doors loc...
4                                                    NaN
5      I am renting it out for 3 months while I am on...

                                               description  \
index
1      AVAILABLE ONLY IN WINTER PRIME La Jolla Villag...
2      Hi! We are basically renting this master for a...
3      All my stuff will be gone. Dishwasher, washer/...
4      NIcely furnished. Great location, 2 blocks to ...
5      The room for rent is on a monthly basis.  It i...

                         neighborhood_overview notes  \
index
1                                        NaN   NaN
```

```
2                                     NaN   NaN
3                             College area   NaN
4      coastal town great travel destination.   NaN
5                                     NaN   NaN

                                                transit  \
index
1                                                  NaN
2                                                  NaN
3      Across the street from the green line trolly. ...
4                                   Bus line near by.
5                                                  NaN

                                                access  \
index
1                                                  NaN
2                                                  NaN
3      Everything except second bedroom in apartment....
4                                                  NaN
5                                                  NaN

                     interaction                  house_rules thumbnail_url  \
index
1                             NaN                          NaN           NaN
2                             NaN                          NaN           NaN
3                             NaN                          NaN           NaN
4      Call/text business hours.  No pets. Clean and healthy.           NaN
5                             NaN                          NaN           NaN

         host_id                                    host_url    host_name  \
index
1       79755951   https://www.airbnb.com/users/show/79755951  Juan Carlos
2      159773487  https://www.airbnb.com/users/show/159773487      Kaitlyn
3      185758729  https://www.airbnb.com/users/show/185758729         Alex
4        3634860    https://www.airbnb.com/users/show/3634860         Anne
5       89146371   https://www.airbnb.com/users/show/89146371        Jesse

      host_since                         host_location  \
index
1        6/24/16  San Diego, California, United States
2       11/22/17                                   US
3        4/22/18                                   US
4        9/21/12  San Diego, California, United States
5        8/10/16  San Diego, California, United States
```

```
                                          host_about  host_response_time  \
index
1                              Quiet and considerate.                 NaN
2                                                 NaN                 NaN
3                                                 NaN                 NaN
4                   Life is too short, enjoy it.....\n        within a day
5     I am a Real Estate Broker in California.  \n\n...  within a few hours

      host_response_rate host_acceptance_rate host_is_superhost  \
index
1                    NaN                  NaN                 f
2                    NaN                  NaN                 f
3                    NaN                  NaN                 f
4                    67%                  NaN                 f
5                   100%                  NaN                 f

      host_neighbourhood host_listings_count host_total_listings_count  \
index
1              La Jolla                   1                         1
2          Pacific Beach                  1                         1
3           College East                  1                         1
4              La Jolla                   6                         6
5             Poblacion                   5                         5

      host_has_profile_pic host_identity_verified  \
index
1                        t                      f
2                        t                      f
3                        t                      f
4                        t                      t
5                        t                      f

                           street  neighbourhood neighbourhood_cleansed  \
index
1      San Diego, CA, United States       La Jolla              La Jolla
2      San Diego, CA, United States  Pacific Beach          Pacific Beach
3      San Diego, CA, United States   College East           College Area
4      San Diego, CA, United States       La Jolla              La Jolla
5      San Diego, CA, United States      Mira Mesa             Mira Mesa

           city state zipcode     market smart_location country_code  \
index
1      San Diego    CA   92037  San Diego  San Diego, CA           US
2      San Diego    CA   92109  San Diego  San Diego, CA           US
3      San Diego    CA   92120  San Diego  San Diego, CA           US
```

```
4      San Diego    CA   92037  San Diego  San Diego, CA           US
5      San Diego    CA   92126  San Diego  San Diego, CA           US

             country  latitude   longitude is_location_exact property_type  \
index
1      United States  32.84067  -117.27443                 t     Apartment
2      United States  32.79797   -117.2425                 t     Townhouse
3      United States  32.77545  -117.05923                 t     Apartment
4      United States  32.84619  -117.27558                 t   Condominium
5      United States  32.93033  -117.13254                 t         House

             room_type accommodates bathrooms bedrooms beds  bed_type  \
index
1      Entire home/apt            1         2        2    3  Real Bed
2         Private room            1         1        1    1  Real Bed
3      Entire home/apt            1         1        1    1  Real Bed
4         Private room            1         1        1    1  Real Bed
5         Private room            1         1        1    1  Real Bed

                                           amenities square_feet  \
index
1      {TV,Wifi,"Air conditioning",Kitchen,"Free park...         NaN
2      {TV,Wifi,Kitchen,"Free parking on premises","P...         NaN
3      {Wifi,"Air conditioning",Pool,Kitchen,"Free pa...         NaN
4      {TV,Wifi,Kitchen,"Free parking on premises",El...         NaN
5      {TV,Wifi,Pool,Kitchen,"Free parking on premise...         NaN

      nightly_price price_per_stay security_deposit cleaning_fee  \
index
1        $1,400.00      $1,400.00              NaN          NaN
2        $1,250.00      $1,250.00              NaN          NaN
3        $1,150.00      $1,150.00              NaN          NaN
4          $110.00        $110.00          $200.00          NaN
5          $645.00        $645.00              NaN          NaN

      guests_included extra_people minimum_nights maximum_nights  \
index
1                   1       $0.00             30          1125
2                   1       $0.00             30            30
3                   1       $0.00             31            40
4                   1      $50.00            180           365
5                   1       $0.00             30            90

      number_of_reviews number_of_stays first_review last_review  \
index
```

```
1                    2              4      4/22/17     8/31/17
2                    0              0          NaN         NaN
3                    0              0          NaN         NaN
4                    2              4      2/22/15     9/23/18
5                    0              0          NaN         NaN

      review_scores_rating review_scores_accuracy review_scores_cleanliness  \
index
1                     100                     10                        10
2                     NaN                    NaN                       NaN
3                     NaN                    NaN                       NaN
4                      90                      7                         7
5                     NaN                    NaN                       NaN

      review_scores_checkin review_scores_communication  \
index
1                        10                         10
2                       NaN                        NaN
3                       NaN                        NaN
4                         8                          8
5                       NaN                        NaN

      review_scores_location review_scores_value requires_license  \
index
1                         10                 10                f
2                        NaN                NaN                f
3                        NaN                NaN                f
4                          9                 10                f
5                        NaN                NaN                f

      instant_bookable is_business_travel_ready          cancellation_policy  \
index
1                    t                         f  strict_14_with_grace_period
2                    t                         f                     flexible
3                    f                         f                     flexible
4                    f                         f  strict_14_with_grace_period
5                    t                         f                     flexible

      require_guest_profile_picture require_guest_phone_verification
index
1                                 f                                f
2                                 f                                f
3                                 f                                f
4                                 f                                f
5                                 f                                f
```

```
(12795, 28)
Index(['id', 'host_name', 'host_since', 'host_is_superhost', 'neighbourhood',
       'city', 'state', 'latitude', 'longitude', 'is_location_exact',
       'property_type', 'room_type', 'accommodates', 'bathrooms', 'bedrooms',
       'beds', 'amenities', 'nightly_price', 'security_deposit',
       'cleaning_fee', 'guests_included', 'extra_people', 'number_of_reviews',
       'number_of_stays', 'first_review', 'last_review', 'instant_bookable',
       'cancellation_policy'],
      dtype='object')
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12795 entries, 1 to 13051
Data columns (total 28 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   id                   12795 non-null  int64 
 1   host_name            12791 non-null  object
 2   host_since           12791 non-null  object
 3   host_is_superhost    12791 non-null  object
 4   neighbourhood        12364 non-null  object
 5   city                 12795 non-null  object
 6   state                12795 non-null  object
 7   latitude             12795 non-null  object
 8   longitude            12795 non-null  object
 9   is_location_exact    12795 non-null  object
 10  property_type        12795 non-null  object
 11  room_type            12795 non-null  object
 12  accommodates         12795 non-null  object
 13  bathrooms            12795 non-null  object
 14  bedrooms             12795 non-null  object
 15  beds                 12795 non-null  object
 16  amenities            12795 non-null  object
 17  nightly_price        12795 non-null  object
 18  security_deposit     10153 non-null  object
 19  cleaning_fee         11309 non-null  object
 20  guests_included      12795 non-null  object
 21  extra_people         12795 non-null  object
 22  number_of_reviews    12795 non-null  object
 23  number_of_stays      12795 non-null  object
 24  first_review         10795 non-null  object
 25  last_review          10795 non-null  object
 26  instant_bookable     12795 non-null  object
 27  cancellation_policy  12795 non-null  object
dtypes: int64(1), object(27)
memory usage: 2.8+ MB
None
```

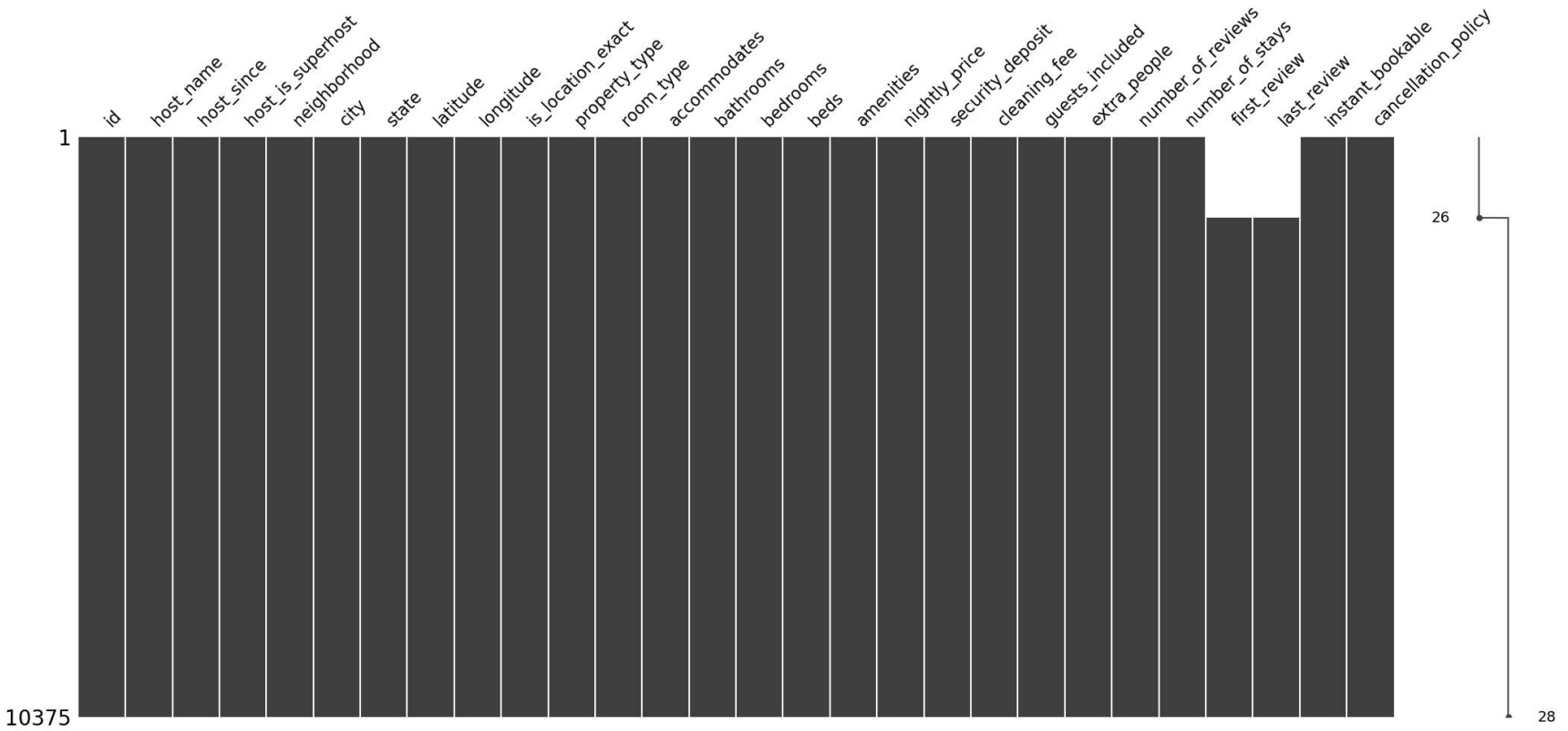
```
index
1     1400.0
2     1250.0
3     1150.0
4      110.0
5      645.0
6       38.0
7       42.0
8      250.0
9       40.0
10      75.0
11    5000.0
12      75.0
13      65.0
14     127.0
15     750.0
16      38.0
17     110.0
18     115.0
19      99.0
20      95.0
Name: nightly_price, dtype: float64
```

```
C:\Users\herlihpj\AppData\Local\Temp\ipykernel_23096\2991358282.py:46: FutureWarning: The default value of regex will ch
ange from True to False in a future version. In addition, single character regular expressions will *not* be treated as
literal strings when regex=True.
  airbnb_dirty[i]=airbnb_dirty[i].str.replace('$','')
```

```
(12364, 28)
(10375, 28)
====================== Cleaned Results ======================
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10375 entries, 4 to 13042
Data columns (total 28 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   id                   10375 non-null  int64  
 1   host_name            10375 non-null  object 
 2   host_since           10375 non-null  object 
 3   host_is_superhost    10375 non-null  object 
 4   neighborhood         10375 non-null  object 
 5   city                 10375 non-null  object 
 6   state                10375 non-null  object 
 7   latitude             10375 non-null  float64
 8   longitude            10375 non-null  float64
 9   is_location_exact    10375 non-null  object 
 10  property_type        10375 non-null  object 
 11  room_type            10375 non-null  object 
 12  accommodates         10375 non-null  int64  
 13  bathrooms            10375 non-null  float64
 14  bedrooms             10375 non-null  int64  
 15  beds                 10375 non-null  int64  
 16  amenities            10375 non-null  object 
 17  nightly_price        10375 non-null  float64
 18  security_deposit     10375 non-null  float64
 19  cleaning_fee         10375 non-null  float64
 20  guests_included      10375 non-null  int64  
 21  extra_people         10375 non-null  float64
 22  number_of_reviews    10375 non-null  int64  
 23  number_of_stays      10375 non-null  int64  
 24  first_review         8930 non-null   object 
 25  last_review          8930 non-null   object 
 26  instant_bookable     10375 non-null  object 
 27  cancellation_policy  10375 non-null  object 
dtypes: float64(7), int64(7), object(14)
memory usage: 2.3+ MB
None
Index(['id', 'host_name', 'host_since', 'host_is_superhost', 'neighborhood',
       'city', 'state', 'latitude', 'longitude', 'is_location_exact',
       'property_type', 'room_type', 'accommodates', 'bathrooms', 'bedrooms',
       'beds', 'amenities', 'nightly_price', 'security_deposit',
       'cleaning_fee', 'guests_included', 'extra_people', 'number_of_reviews',
       'number_of_stays', 'first_review', 'last_review', 'instant_bookable',
```

```
       'cancellation_policy'],
      dtype='object')
(10375, 28)
<AxesSubplot:>
```

Out[32]:



In [ ]:

In [ ]: