ImplementMLProjectPlan

August 11, 2023

1 Lab 8: Implement Your Machine Learning Project Plan

In this lab assignment, you will implement the machine learning project plan you created in the written assignment. You will:

1. Load your data set and save it to a Pandas DataFrame.
2. Perform exploratory data analysis on your data to determine which feature engineering and data preparation techniques you will use.
3. Prepare your data for your model and create features and a label.
4. Fit your model to the training data and evaluate your model.
5. Improve your model by performing model selection and/or feature selection techniques to find best model for your problem.

1.0.1 Import Packages

Before you get started, import a few packages.

[131]:
```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import scipy.stats as stats

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Task: In the code cell below, import additional packages that you have used in this course that you will need for this task.

[ ]:
```
```

1.1 Part 1: Load the Data Set

You have chosen to work with one of four data sets. The data sets are located in a folder named "data." The file names of the three data sets are as follows:

- The "adult" data set that contains Census information from 1994 is located in file `adultData.csv`
- The airbnb NYC "listings" data set is located in file `airbnbListingsData.csv`
- The World Happiness Report (WHR) data set is located in file `WHR2018Chapter2OnlineData.csv`
- The book review data set is located in file `bookReviewsData.csv`

Task: In the code cell below, use the same method you have been using to load your data using `pd.read_csv()` and save it to DataFrame `df`.

[132]:
```
#Loading the dataset into the notebook
filename = os.path.join(os.getcwd(), "data", "airbnbListingsData.csv")
df = pd.read_csv(filename, header = 0)
```

1.2 Part 2: Exploratory Data Analysis

The next step is to inspect and analyze your data set with your machine learning problem and project plan in mind.

This step will help you determine data preparation and feature engineering techniques you will need to apply to your data to build a balanced modeling data set for your problem and model. These data preparation techniques may include: * addressing missingness, such as replacing missing values with means * renaming features and labels * finding and replacing outliers * performing winsorization if needed * performing one-hot encoding on categorical features * performing vectorization for an NLP problem * addressing class imbalance in your data sample to promote fair AI

Think of the different techniques you have used to inspect and analyze your data in this course. These include using Pandas to apply data filters, using the Pandas `describe()` method to get insight into key statistics for each column, using the Pandas `dtypes` property to inspect the data type of each column, and using Matplotlib and Seaborn to detect outliers and visualize relationships between features and labels. If you are working on a classification problem, use techniques you have learned to determine if there is class imbalance.

Task: Use the techniques you have learned in this course to inspect and analyze your data.

Note: You can add code cells if needed by going to the Insert menu and clicking on Insert Cell Below in the drop-drown menu.

[133]:
```
df.head(20)
```

[133]:
```
                                                 name  \
0                             Skylit Midtown Castle
1   Whole flr w/private bdrm, bath & kitchen(pls r...
2           Spacious Brooklyn Duplex, Patio + Garden
3                   Large Furnished Room Near B'way
4                 Cozy Clean Guest Room - Family Apt
5     Lovely Room 1, Garden, Best Area, Legal rental
6                   Only 2 stops to Manhattan studio
7                            UES Beautiful Blue Room
8   Amazing location! Wburg. Large, bright & tranquil
9         Perfect for Your Parents:  Privacy + Garden
10                  Sweet and Spacious Brooklyn Loft
```

```
11   Maison des Sirenes1,bohemian, luminous apartment
12                               Midtown Pied-a-terre
13                   Modern 1 BR / NYC / East Village
14                Spacious 1 bedroom in luxe building
15                              Large B&B Style rooms
16           Lovely Room 2; Garden; Best area, Legal
17                               ENJOY Downtown NYC!
18                                BEST BET IN HARLEM
19  1 Stop fr. Manhattan! Private Suite,Landmark B...

                                           description  \
0   Beautiful, spacious skylit studio in the heart...
1   Enjoy 500 s.f. top floor in 1899 brownstone, w...
2   We welcome you to stay in our lovely 2 br dupl...
3   Please dont expect the luxury here just a bas...
4   Our best guests are seeking a safe, clean, spa...
5   Beautiful house, gorgeous garden, patio, cozy ...
6   Comfortable studio apartment with super comfor...
7   Beautiful peaceful healthy home<br /><br /><b>...
8   Large, private loft-like room in a spacious 2-...
9   Parents/grandparents coming to town, or are yo...
10  A true open-plan loft in a repurposed factory ...
11  <b>The space</b><br />I am the lucky owner of ...
12  HELLO.  PLEASE DO NOT HIT "REQUEST TO BOOK". H...
13  Awesome, spacious & clean 1 bedroom with a coz...
14  The room is spacious, the neighborhood is safe...
15  Great location.<br /><br /><b>The space</b><br...
16  Lovely room,  gorgeous garden, helpful host in...
17  Please be vaccinated and responsible if you ar...
18  <b>The space</b><br />Centrally located in the...
19  Private room, dedicated bath and a separate en...

                                  neighborhood_overview      host_name  \
0   Centrally located in the heart of Manhattan ju...       Jennifer
1   Just the right mix of urban center and local n...    LisaRoxanne
2                                                 NaN        Rebecca
3     Theater district, many restaurants around here.       Shunichi
4   Our neighborhood is full of restaurants and ca...      MaryEllen
5   Neighborhood is amazing!<br />Best subways to ...         Laurie
6                                                 NaN  Allen & Irina
7   Location: Five minutes to Central Park, Museum...            Cyn
8   - One stop from the East Village, Lower East S...         Joelle
9   Residential, village-like atmosphere. Lots of ...           Jane
10  We've lived here for over 10 years and watched...          Chaya
11                                                NaN       Nathalie
12  Quiet residential block near many restaurants ...          Tommi
13  The east village offers a mixture of old schoo...           Dana
```

```
14                                                  NaN           Teri
15                                                  NaN         Angela
16  Neighborhood is wonderful, a great walking nei...          Laurie
17  Enjoy great food, music, unique shops, night-l...          Edward
18                                                  NaN           Earl
19  Long Island City is the hottest neighborhood i...         Orestes

                       host_location  \
0   New York, New York, United States
1   New York, New York, United States
2   Brooklyn, New York, United States
3   New York, New York, United States
4   New York, New York, United States
5   New York, New York, United States
6   New York, New York, United States
7   New York, New York, United States
8   New York, New York, United States
9   New York, New York, United States
10  New York, New York, United States
11  New York, New York, United States
12  New York, New York, United States
13  New York, New York, United States
14  New York, New York, United States
15  New York, New York, United States
16  New York, New York, United States
17  New York, New York, United States
18  New York, New York, United States
19  New York, New York, United States

                                           host_about  host_response_rate  \
0   A New Yorker since 2000! My passion is creatin...                0.80
1   Laid-back Native New Yorker (formerly bi-coast...                0.09
2   Rebecca is an artist/designer, and Henoch is i...                1.00
3   I used to work for a financial industry but no...                1.00
4   Welcome to family life with my oldest two away...                 NaN
5   Hello, \r\nI will be welcoming and helpful,  w...                1.00
6   We love to travel. When we travel we like to s...                1.00
7   Capturing the Steinbeck side of life in its Fi...                1.00
8   I have lived in the same apartment in Brooklyn...                1.00
9   I have been an Airbnb host since 2009 -- just ...                1.00
10  We're a couple in our thirties who love to tra...                1.00
11   I am French and have been living in Ny for 10...                1.00
12  I am a spirit-minded shoe model and alternativ...                0.00
13  I am an industrial designer and native new yor...                0.80
14  I'm a citizen of the world.  I love to travel ...                0.00
15                          Loves to travel and host.                1.00
16  Hello, \r\nI will be welcoming and helpful,  w...                1.00
```

```
17  I am an actor & adventurer originally from Tor...                1.00
18  My reviews are a more accurate description of ...                 NaN
19  Photographer/Real Estate Developer. \r\nI enjo...                1.00

    host_acceptance_rate  host_is_superhost  host_listings_count  ...  \
0                   0.17              True                  8.0  ...
1                   0.69              True                  1.0  ...
2                   0.25              True                  1.0  ...
3                   1.00              True                  1.0  ...
4                    NaN              True                  1.0  ...
5                   1.00              True                  3.0  ...
6                   1.00              True                  1.0  ...
7                   1.00              True                  3.0  ...
8                   0.00              True                  2.0  ...
9                   0.99              True                  1.0  ...
10                  0.61              True                  4.0  ...
11                  0.98              True                  2.0  ...
12                   NaN              True                  1.0  ...
13                  0.54              True                  1.0  ...
14                  0.00              True                  1.0  ...
15                  0.84              True                  0.0  ...
16                  1.00              True                  3.0  ...
17                  0.75              True                  2.0  ...
18                   NaN              True                  1.0  ...
19                  0.94              True                  1.0  ...

    review_scores_communication  review_scores_location  review_scores_value  \
0                          4.79                    4.86                 4.41
1                          4.80                    4.71                 4.64
2                          5.00                    4.50                 5.00
3                          4.42                    4.87                 4.36
4                          4.95                    4.94                 4.92
5                          4.82                    4.87                 4.73
6                          4.80                    4.67                 4.57
7                          4.95                    4.84                 4.84
8                          5.00                    5.00                 5.00
9                          4.91                    4.93                 4.78
10                         4.60                    5.00                 4.80
11                         4.87                    4.61                 4.75
12                         5.00                    4.95                 4.58
13                         4.84                    4.87                 4.34
14                         4.84                    4.84                 4.90
15                         4.85                    4.34                 4.63
16                         4.77                    4.88                 4.75
17                         4.85                    4.70                 4.55
18                         4.90                    4.51                 4.70
19                         4.90                    4.88                 4.84
```

```
   instant_bookable calculated_host_listings_count  \
0             False                               3
1             False                               1
2             False                               1
3             False                               1
4             False                               1
5             False                               3
6              True                               1
7              True                               1
8             False                               2
9              True                               2
10            False                               1
11            False                               2
12            False                               1
13            False                               1
14            False                               1
15            False                               4
16            False                               3
17            False                               2
18             True                               1
19            False                               1

    calculated_host_listings_count_entire_homes  \
0                                             3
1                                             1
2                                             1
3                                             0
4                                             0
5                                             1
6                                             1
7                                             0
8                                             0
9                                             1
10                                            1
11                                            2
12                                            1
13                                            1
14                                            0
15                                            0
16                                            1
17                                            0
18                                            1
19                                            0

    calculated_host_listings_count_private_rooms  \
0                                               0
```

1	0
2	0
3	1
4	1
5	2
6	0
7	1
8	2
9	1
10	0
11	0
12	0
13	0
14	1
15	4
16	2
17	2
18	0
19	1

	calculated_host_listings_count_shared_rooms	reviews_per_month	\
0	0	0.33	
1	0	4.86	
2	0	0.02	
3	0	3.68	
4	0	0.87	
5	0	1.48	
6	0	1.24	
7	0	1.82	
8	0	0.07	
9	0	3.05	
10	0	0.06	
11	0	1.17	
12	0	0.55	
13	0	0.75	
14	0	1.36	
15	0	0.66	
16	0	2.12	
17	0	1.99	
18	0	0.79	
19	0	3.15	

	n_host_verifications
0	9
1	6
2	3
3	4

```
4                    7
5                    7
6                    7
7                    5
8                    5
9                    8
10                   4
11                   5
12                   4
13                   4
14                   8
15                   6
16                   7
17                   4
18                   5
19                   4

[20 rows x 50 columns]
```

[134]:
```
#Examining the features present in the dataset and the shape to determine the␣
 ↪size
features = list(df.columns)
print(features)
print("This is the shape of my data: ", df.shape)
```

```
['name', 'description', 'neighborhood_overview', 'host_name', 'host_location',
'host_about', 'host_response_rate', 'host_acceptance_rate', 'host_is_superhost',
'host_listings_count', 'host_total_listings_count', 'host_has_profile_pic',
'host_identity_verified', 'neighbourhood_group_cleansed', 'room_type',
'accommodates', 'bathrooms', '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', 'has_availability',
'availability_30', 'availability_60', 'availability_90', 'availability_365',
'number_of_reviews', 'number_of_reviews_ltm', 'number_of_reviews_l30d',
'review_scores_rating', 'review_scores_cleanliness', 'review_scores_checkin',
'review_scores_communication', 'review_scores_location', 'review_scores_value',
'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',
'n_host_verifications']
This is the shape of my data:  (28022, 50)
```

[135]:
```
df.describe()
```

[135]:
```
       host_response_rate  host_acceptance_rate  host_listings_count  \
count        16179.000000          16909.000000         28022.000000
```

mean	0.906901	0.791953	14.554778
std	0.227282	0.276732	120.721287
min	0.000000	0.000000	0.000000
25%	0.940000	0.680000	1.000000
50%	1.000000	0.910000	1.000000
75%	1.000000	1.000000	3.000000
max	1.000000	1.000000	3387.000000

	host_total_listings_count	accommodates	bathrooms	bedrooms	\
count	28022.000000	28022.000000	28022.000000	25104.000000	
mean	14.554778	2.874491	1.142174	1.329708	
std	120.721287	1.860251	0.421132	0.700726	
min	0.000000	1.000000	0.000000	1.000000	
25%	1.000000	2.000000	1.000000	1.000000	
50%	1.000000	2.000000	1.000000	1.000000	
75%	3.000000	4.000000	1.000000	1.000000	
max	3387.000000	16.000000	8.000000	12.000000	

	beds	price	minimum_nights	...	review_scores_checkin	\
count	26668.000000	28022.000000	28022.000000	...	28022.000000	
mean	1.629556	154.228749	18.689387	...	4.814300	
std	1.097104	140.816605	25.569151	...	0.438603	
min	1.000000	29.000000	1.000000	...	0.000000	
25%	1.000000	70.000000	2.000000	...	4.810000	
50%	1.000000	115.000000	30.000000	...	4.960000	
75%	2.000000	180.000000	30.000000	...	5.000000	
max	21.000000	1000.000000	1250.000000	...	5.000000	

	review_scores_communication	review_scores_location	\
count	28022.000000	28022.000000	
mean	4.808041	4.750393	
std	0.464585	0.415717	
min	0.000000	0.000000	
25%	4.810000	4.670000	
50%	4.970000	4.880000	
75%	5.000000	5.000000	
max	5.000000	5.000000	

	review_scores_value	calculated_host_listings_count	\
count	28022.000000	28022.000000	
mean	4.647670	9.581900	
std	0.518023	32.227523	
min	0.000000	1.000000	
25%	4.550000	1.000000	
50%	4.780000	1.000000	
75%	5.000000	3.000000	
max	5.000000	421.000000	

```
       calculated_host_listings_count_entire_homes  \
count                                 28022.000000
mean                                      5.562986
std                                      26.121426
min                                       0.000000
25%                                       0.000000
50%                                       1.000000
75%                                       1.000000
max                                     308.000000

       calculated_host_listings_count_private_rooms  \
count                                  28022.000000
mean                                       3.902077
std                                       17.972386
min                                        0.000000
25%                                        0.000000
50%                                        0.000000
75%                                        1.000000
max                                      359.000000

       calculated_host_listings_count_shared_rooms  reviews_per_month  \
count                                 28022.000000       28022.000000
mean                                      0.048283           1.758325
std                                       0.442459           4.446143
min                                       0.000000           0.010000
25%                                       0.000000           0.130000
50%                                       0.000000           0.510000
75%                                       0.000000           1.830000
max                                       8.000000         141.000000

       n_host_verifications
count          28022.000000
mean               5.169510
std                2.028497
min                1.000000
25%                4.000000
50%                5.000000
75%                7.000000
max               13.000000

[8 rows x 36 columns]
```

[136]:
```
#Data types of columns
colTypes = df.dtypes
print(colTypes)
```

```
name                                            object
```

```
description                                    object
neighborhood_overview                          object
host_name                                      object
host_location                                  object
host_about                                     object
host_response_rate                            float64
host_acceptance_rate                          float64
host_is_superhost                                bool
host_listings_count                           float64
host_total_listings_count                     float64
host_has_profile_pic                             bool
host_identity_verified                           bool
neighbourhood_group_cleansed                   object
room_type                                      object
accommodates                                    int64
bathrooms                                     float64
bedrooms                                      float64
beds                                          float64
amenities                                      object
price                                         float64
minimum_nights                                  int64
maximum_nights                                  int64
minimum_minimum_nights                        float64
maximum_minimum_nights                        float64
minimum_maximum_nights                        float64
maximum_maximum_nights                        float64
minimum_nights_avg_ntm                        float64
maximum_nights_avg_ntm                        float64
has_availability                                 bool
availability_30                                 int64
availability_60                                 int64
availability_90                                 int64
availability_365                                int64
number_of_reviews                               int64
number_of_reviews_ltm                           int64
number_of_reviews_l30d                          int64
review_scores_rating                          float64
review_scores_cleanliness                     float64
review_scores_checkin                         float64
review_scores_communication                   float64
review_scores_location                        float64
review_scores_value                           float64
instant_bookable                                 bool
calculated_host_listings_count                  int64
calculated_host_listings_count_entire_homes     int64
calculated_host_listings_count_private_rooms    int64
calculated_host_listings_count_shared_rooms     int64
reviews_per_month                             float64
```

```
n_host_verifications                            int64
dtype: object
```

Note: There are 4 types of data that I am observing: object, float64, int64, and bool. In order to get accurate data and features, I would have to change the bool to 0 or 1 (False or True).

[137]:
```
anyNull = df.isnull().values.any()
print(anyNull)
```

```
True
```

[138]:
```
nullCount = np.sum(df.isnull(), axis = 0)
print("This is the number of missing data by column:\n", nullCount)

#Aggregating only columns with missing values
#All columns with missing values
nullCol = nullCount != 0
colNames = nullCount[nullCol].index
nan_cols = list(colNames)
print("\nAll features with missing values: \n", nan_cols)
```

```
This is the number of missing data by column:
 name                                           5
description                                  570
neighborhood_overview                       9816
host_name                                      0
host_location                                 60
host_about                                 10945
host_response_rate                         11843
host_acceptance_rate                       11113
host_is_superhost                              0
host_listings_count                            0
host_total_listings_count                      0
host_has_profile_pic                           0
host_identity_verified                         0
neighbourhood_group_cleansed                   0
room_type                                      0
accommodates                                   0
bathrooms                                      0
bedrooms                                    2918
beds                                        1354
amenities                                      0
price                                          0
minimum_nights                                 0
maximum_nights                                 0
minimum_minimum_nights                         0
maximum_minimum_nights                         0
minimum_maximum_nights                         0
```

```
maximum_maximum_nights                          0
minimum_nights_avg_ntm                          0
maximum_nights_avg_ntm                          0
has_availability                                0
availability_30                                 0
availability_60                                 0
availability_90                                 0
availability_365                                0
number_of_reviews                               0
number_of_reviews_ltm                           0
number_of_reviews_l30d                          0
review_scores_rating                            0
review_scores_cleanliness                       0
review_scores_checkin                           0
review_scores_communication                     0
review_scores_location                          0
review_scores_value                             0
instant_bookable                                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                               0
n_host_verifications                            0
dtype: int64

All features with missing values:
 ['name', 'description', 'neighborhood_overview', 'host_location', 'host_about',
'host_response_rate', 'host_acceptance_rate', 'bedrooms', 'beds']
```

[139]:
```
#Data types of columns with missing data
nanColTypes = df[nan_cols].dtypes
nanColTypes
```

[139]:
```
name                    object
description             object
neighborhood_overview   object
host_location           object
host_about              object
host_response_rate      float64
host_acceptance_rate    float64
bedrooms                float64
beds                    float64
dtype: object
```

[140]:
```
df.loc[df['bedrooms'].isnull()]
```

[140]:
```
                                                 name  \
0                              Skylit Midtown Castle
```

```
6                       Only 2 stops to Manhattan studio
10                      Sweet and Spacious Brooklyn Loft
12                                Midtown Pied-a-terre
59                               East Village Sanctuary
...                                                 ...
27776                                            121816
27787                                  KO LOFT & LOUNGE
27923  Spacious Brooklyn Loft w/ Private Rooftop City...
27958  Cozy Alcove Studio at the Heart of Astoria, Qu...
27994                Lovely studio apartment in New York

                                             description  \
0      Beautiful, spacious skylit studio in the heart...
6      Comfortable studio apartment with super comfor...
10     A true open-plan loft in a repurposed factory ...
12     HELLO.  PLEASE DO NOT HIT "REQUEST TO BOOK". H...
59           Sorry, this listing is no longer available.
...                                                  ...
27776                  12
27787  Keep it simple at this peaceful and centrally-...
27923  This large stylish Brooklyn loft is perfect fo...
27958  Cozy Spacious Alcove Studio at the Heart of As...
27994  Cozy and sunny studio apartment in a walk up b...

                                    neighborhood_overview      host_name  \
0      Centrally located in the heart of Manhattan ju...       Jennifer
6                                                    NaN  Allen & Irina
10     We've lived here for over 10 years and watched...          Chaya
12     Quiet residential block near many restaurants ...          Tommi
59                                                   NaN            Jen
...                                                  ...            ...
27776                  123         Qiulan
27787                                                NaN        Gabriel
27923  One block away from Broadway which features lo...           Zach
27958                                                NaN          Jason
27994                                                NaN           Olga

                          host_location  \
0      New York, New York, United States
6      New York, New York, United States
10     New York, New York, United States
12     New York, New York, United States
59        Portland, Maine, United States
...                                  ...
27776                                 US
27787  Brooklyn, New York, United States
27923  Brooklyn, New York, United States
```

```
27958  New York, New York, United States
27994  New York, New York, United States

                                            host_about  host_response_rate  \
0      A New Yorker since 2000! My passion is creatin...                0.80
6      We love to travel. When we travel we like to s...                1.00
10     We're a couple in our thirties who love to tra...                1.00
12     I am a spirit-minded shoe model and alternativ...                0.00
59     Creative guru of media and fine art and good c...                1.00
...                                                  ...                 ...
27776                                  Qiu lan Lin                0.94
27787                                          NaN                 NaN
27923                                          NaN                0.98
27958  New Yorker who travels extensively. Spiritual ...                1.00
27994                                          NaN                1.00

       host_acceptance_rate  host_is_superhost  host_listings_count  ...  \
0                      0.17               True                  8.0  ...
6                      1.00               True                  1.0  ...
10                     0.61               True                  4.0  ...
12                      NaN               True                  1.0  ...
59                      NaN               True                  1.0  ...
...                     ...                ...                  ...  ...
27776                  0.78               True                 15.0  ...
27787                   NaN               True                  3.0  ...
27923                  0.90               True                  0.0  ...
27958                  1.00               True                  0.0  ...
27994                  1.00               True                  0.0  ...

       review_scores_communication  review_scores_location  \
0                             4.79                    4.86
6                             4.80                    4.67
10                            4.60                    5.00
12                            5.00                    4.95
59                            4.89                    4.79
...                            ...                     ...
27776                         2.00                    2.00
27787                         4.75                    5.00
27923                         4.00                    4.67
27958                         5.00                    5.00
27994                         5.00                    4.00

       review_scores_value instant_bookable calculated_host_listings_count  \
0                     4.41            False                              3
6                     4.57             True                              1
10                    4.80            False                              1
12                    4.58            False                              1
```

```
59                    4.74            False                                1
...                    ...              ...                              ...
27776                 1.00            False                                7
27787                 5.00             True                                3
27923                 4.67             True                                1
27958                 5.00             True                                1
27994                 4.00             True                                1

       calculated_host_listings_count_entire_homes  \
0                                                3
6                                                1
10                                               1
12                                               1
59                                               1
...                                            ...
27776                                            0
27787                                            3
27923                                            1
27958                                            1
27994                                            1

       calculated_host_listings_count_private_rooms  \
0                                                 0
6                                                 0
10                                                0
12                                                0
59                                                0
...                                             ...
27776                                             7
27787                                             0
27923                                             0
27958                                             0
27994                                             0

       calculated_host_listings_count_shared_rooms  reviews_per_month  \
0                                                0               0.33
6                                                0               1.24
10                                               0               0.06
12                                               0               0.55
59                                               0               0.20
...                                            ...                ...
27776                                            0               1.00
27787                                            0               4.00
27923                                            0               3.00
27958                                            0               1.00
27994                                            0               1.00
```

```
      n_host_verifications
0                        9
6                        7
10                       4
12                       4
59                       6
...                    ...
27776                    5
27787                    2
27923                    4
27958                    5
27994                    5

[2918 rows x 50 columns]
```

[141]:
```
df.loc[df['beds'].isnull()]
```

[141]:
```
                                                  name  \
5        Lovely Room 1, Garden, Best Area, Legal rental
16              Lovely Room 2; Garden; Best area, Legal
23                          * ORIGINAL BROOKLYN LOFT *
46                 Sunny room+Pvte office in huge loft
48                   Light-filled classic Central Park
...                                                ...
27991    Lovely Studio-apartment unit in UES, New York.
27994               Lovely studio apartment in New York
27999           Charming bedroom with private bathroom.
28000  A peaceful Brooklyn gem with a private bathroom.
28011     Private Entire Suite- 5 mins from LGA Airport

                                              description  \
5      Beautiful house, gorgeous garden, patio, cozy ...
16     Lovely room,  gorgeous garden, helpful host in...
23     Original factory building loft,  lots of natur...
46     FOR RENT is 400sqf brand new renovated room co...
48     An adorable, classic, clean, light-filled one-...
...                                                  ...
27991  Your family will be close to everything when y...
27994  Cozy and sunny studio apartment in a walk up b...
27999  Come spread love the Brooklyn way in this peac...
28000  A sun filled room with space for yoga and work...
28011  This is a cozy renovated  semi-basement Suite ...

                                  neighborhood_overview host_name  \
5      Neighborhood is amazing!<br />Best subways to ...    Laurie
16     Neighborhood is wonderful, a great walking nei...    Laurie
23     Bushwick is a constantly changing area,  new o...     James
46                                                 NaN  Augustin
```

```
48     Diverse. Great coffee shops and restaurants, n...      Dana
...                                                  ...       ...
27991                                                NaN   Sabrina
27994                                                NaN      Olga
27999                                                NaN     Amani
28000                                                NaN     Amani
28011  It is located in a very family friendly & Welc...  Mohammad

                         host_location  \
5      New York, New York, United States
16     New York, New York, United States
23     New York, New York, United States
46     Malibu, California, United States
48     New York, New York, United States
...                                  ...
27991  New York, New York, United States
27994  New York, New York, United States
27999  New York, New York, United States
28000  New York, New York, United States
28011  New York, New York, United States

                                              host_about  host_response_rate  \
5      Hello, \r\nI will be welcoming and helpful,  w...                1.00
16     Hello, \r\nI will be welcoming and helpful,  w...                1.00
23     \r\nPhotographer and Designer\r\n\r\n I've ren...                0.77
46            French filmmaker based in NY since 6 years                0.00
48     I'm an arts consultant, personal trainer and a...                0.38
...                                                  ...                 ...
27991  I am a Latina young woman. Love to travel and ...                0.97
27994                                                NaN                1.00
27999                                                NaN                1.00
28000                                                NaN                1.00
28011  My name is Mohammad and my wifes name is Minu...                NaN

       host_acceptance_rate  host_is_superhost  host_listings_count  ...  \
5                      1.00               True                  3.0  ...
16                     1.00               True                  3.0  ...
23                     0.80               True                  3.0  ...
46                     0.14               True                  1.0  ...
48                     0.50               True                  1.0  ...
...                     ...                ...                  ...  ...
27991                  1.00               True                  0.0  ...
27994                  1.00               True                  0.0  ...
27999                  0.82               True                  0.0  ...
28000                  0.82               True                  0.0  ...
28011                  1.00               True                  0.0  ...
```

```
       review_scores_communication  review_scores_location  \
5                             4.82                    4.87
16                            4.77                    4.88
23                            4.61                    4.77
46                            4.24                    4.52
48                            4.91                    4.76
...                            ...                     ...
27991                         5.00                    5.00
27994                         5.00                    4.00
27999                         5.00                    2.00
28000                         5.00                    5.00
28011                         5.00                    5.00

       review_scores_value instant_bookable calculated_host_listings_count  \
5                     4.73           False                               3
16                    4.75           False                               3
23                    4.75           False                               1
46                    4.52           False                               1
48                    4.76           False                               1
...                    ...             ...                             ...
27991                 4.67            True                               1
27994                 4.00            True                               1
27999                 5.00            True                               2
28000                 4.00            True                               2
28011                 5.00           False                               1

       calculated_host_listings_count_entire_homes  \
5                                                1
16                                               1
23                                               1
46                                               0
48                                               0
...                                            ...
27991                                            1
27994                                            1
27999                                            0
28000                                            0
28011                                            1

       calculated_host_listings_count_private_rooms  \
5                                                 2
16                                                2
23                                                0
46                                                1
48                                                1
...                                             ...
27991                                             0
```

```
27994                                                 0
27999                                                 2
28000                                                 2
28011                                                 0

       calculated_host_listings_count_shared_rooms  reviews_per_month  \
5                                                  0               1.48
16                                                 0               2.12
23                                                 0               2.16
46                                                 0               0.61
48                                                 0               1.30
...                                              ...                ...
27991                                              0               3.00
27994                                              0               1.00
27999                                              0               1.00
28000                                              0               1.00
28011                                              0               1.00

      n_host_verifications
5                        7
16                       7
23                       8
46                       6
48                       4
...                    ...
27991                    1
27994                    5
27999                    2
28000                    2
28011                    2

[1354 rows x 50 columns]
```

[142]:
```
# compute mean for all non null values
mean_resRate = df['host_response_rate'].mean()
print("mean value for all age columns: " + str(mean_resRate))

mean_accRate = df['host_acceptance_rate'].mean()
print("Mean acceptance rate: " + str(mean_accRate))

mean_bedR = df['bedrooms'].mean()
print("Mean bedrooms: " + str(mean_bedR))

mean_beds = df['beds'].mean()
print("Mean beds: " + str(mean_beds))

# fill all missing values with the mean
```

```
df['host_response_rate'].fillna(value = mean_resRate, inplace = True)
df['host_acceptance_rate'].fillna(value = mean_accRate, inplace = True)
df['bedrooms'].fillna(value = mean_bedR, inplace = True)
df['beds'].fillna(value = mean_beds, inplace = True)

#Check if means replaced the missing values
print("Row 4 - Host Response:  " + str(df['host_response_rate'][4]))
print("Row 4 - Host Acceptance:  " + str(df['host_acceptance_rate'][4]))
print("Row 6 - Bedrooms:  " + str(df['bedrooms'][6]))
print("Row 5 - Beds:  " + str(df['beds'][5]))
```

```
mean value for all age columns: 0.9069009209469064
Mean acceptance rate: 0.7919528061978829
Mean bedrooms: 1.3297084130019121
Mean beds: 1.62955602219889
Row 4 - Host Response:  0.9069009209469064
Row 4 - Host Acceptance:  0.7919528061978829
Row 6 - Bedrooms:  1.3297084130019121
Row 5 - Beds:  1.62955602219889
```

[143]:
```
#Getting all column names with boolean values
boolVal = list(df.select_dtypes(include = ['bool']).columns)
print(boolVal)
```

```
['host_is_superhost', 'host_has_profile_pic', 'host_identity_verified',
'has_availability', 'instant_bookable']
```

[144]:
```
#Changing boolean values to 0 or 1
df['host_is_superhost'] = df['host_is_superhost'].astype(int)
df['host_has_profile_pic'] = df['host_has_profile_pic'].astype(int)
df['host_identity_verified'] = df['host_identity_verified'].astype(int)
df['has_availability'] = df['has_availability'].astype(int)
df['instant_bookable'] = df['instant_bookable'].astype(int)

#Checking to see if values changed
df[boolVal].dtypes
```

[144]:
```
host_is_superhost         int64
host_has_profile_pic      int64
host_identity_verified    int64
has_availability          int64
instant_bookable          int64
dtype: object
```

[145]:
```
'''
df['host_acceptance_rate_win'] = stats.mstats.
 ↪winsorize(df['host_acceptance_rate'], limits = [0.01, 0.01])
```

```
df['host_listings_count_win'] = stats.mstats.
 ↪winsorize(df['host_listings_count'], limits = [0.01, 0.01])
df['host_total_listings_count_win'] = stats.mstats.
 ↪winsorize(df['host_total_listings_count'], limits = [0.01, 0.01])
df['host_has_profile_pic_win'] = stats.mstats.
 ↪winsorize(df['host_has_profile_pic'], limits = [0.01, 0.01])
df['host_identity_verified_win'] = stats.mstats.
 ↪winsorize(df['host_identity_verified'], limits = [0.01, 0.01])
df['review_scores_cleanliness_win'] = stats.mstats.
 ↪winsorize(df['review_scores_cleanliness'], limits = [0.01, 0.01])
df['review_scores_checkin_win'] = stats.mstats.
 ↪winsorize(df['review_scores_checkin'], limits = [0.01, 0.01])
df['review_scores_communication_win'] = stats.mstats.
 ↪winsorize(df['review_scores_communication'], limits = [0.01, 0.01])
df['review_scores_value_win'] = stats.mstats.
 ↪winsorize(df['review_scores_value'], limits = [0.01, 0.01])
df.head(20)
'''
```

[145]:
```
"\ndf['host_acceptance_rate_win'] =
stats.mstats.winsorize(df['host_acceptance_rate'], limits = [0.01,
0.01])\ndf['host_listings_count_win'] =
stats.mstats.winsorize(df['host_listings_count'], limits = [0.01,
0.01])\ndf['host_total_listings_count_win'] =
stats.mstats.winsorize(df['host_total_listings_count'], limits = [0.01,
0.01])\ndf['host_has_profile_pic_win'] =
stats.mstats.winsorize(df['host_has_profile_pic'], limits = [0.01,
0.01])\ndf['host_identity_verified_win'] =
stats.mstats.winsorize(df['host_identity_verified'], limits = [0.01,
0.01])\ndf['review_scores_cleanliness_win'] =
stats.mstats.winsorize(df['review_scores_cleanliness'], limits = [0.01,
0.01])\ndf['review_scores_checkin_win'] =
stats.mstats.winsorize(df['review_scores_checkin'], limits = [0.01,
0.01])\ndf['review_scores_communication_win'] =
stats.mstats.winsorize(df['review_scores_communication'], limits = [0.01,
0.01])\ndf['review_scores_value_win'] =
stats.mstats.winsorize(df['review_scores_value'], limits = [0.01,
0.01])\ndf.head(20)\n"
```

1.3 Part 3: Implement Your Project Plan

Task: Use the rest of this notebook to carry out your project plan. You will:

1. Prepare your data for your model and create features and a label.
2. Fit your model to the training data and evaluate your model.
3. Improve your model by performing model selection and/or feature selection techniques to find best model for your problem.

Add code cells below and populate the notebook with commentary, code, analyses, results, and figures as you see fit.

[146]:
```
#Selected features in a list
featureList = list(["host_acceptance_rate",
                "host_listings_count", "host_total_listings_count",
                "host_has_profile_pic", "host_identity_verified",
                 "review_scores_cleanliness",
                "review_scores_checkin", "review_scores_communication",
                "review_scores_value"])
```

[147]:
```
#Setting up the X and y values
y = df['review_scores_rating']
X = df[featureList]

print("Number of examples: " + str(X.shape[0]))
print("\nNumber of Features:" + str(X.shape[1]))
```

```
Number of examples: 28022

Number of Features:9
```

[148]:
```
#Splitting the data into training data and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20,␣
 ↪random_state = 42)
```

[149]:
```
#Checking the shape of the data
print(X_train.shape)
print(X_test.shape)
```

```
(22417, 9)
(5605, 9)
```

[150]:
```
#Defining the Linear Regression Model
model = LinearRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)

def LR_model(X_train, y_train, X_test, y_test):

    mse = mean_squared_error(y_test, predictions)
    rmse = np.sqrt(mse)
    r2 = r2_score(y_test, predictions)


    return mse, rmse, r2
```

[151]:
```
#Calculate and print Mean Squared Error, Root Mean Squared Error, R2
mse_airbnb, rmse_airbnb, r2_airbnb = LR_model(X_train, y_train, X_test, y_test)
```

```
print('Mean Squared Error: ' + str(mse_airbnb))
print('Root Mean Squared Error: ' + str(rmse_airbnb))
print('R2 score: ' + str(r2_airbnb))
```

```
Mean Squared Error: 0.06791523481668077
Root Mean Squared Error: 0.2606055157065575
R2 score: 0.732631567144953
```

[152]:
```
#Get the coefficients to determine the feature importance in model
coefficients = model.coef_
feature_names = X.columns

# Print the coefficients and corresponding feature names
for feature, coef in zip(feature_names, coefficients):
    print(f"{feature}: {coef}")
```

```
host_acceptance_rate: -0.01680691434760851
host_listings_count: 639855841.7313513
host_total_listings_count: -639855841.7313002
host_has_profile_pic: -74283.55010683852
host_identity_verified: 1114.2146588998214
review_scores_cleanliness: 0.2761572250165045
review_scores_checkin: 0.13531166472967016
review_scores_communication: 0.21553491093072807
review_scores_value: 0.3935187637762283
```

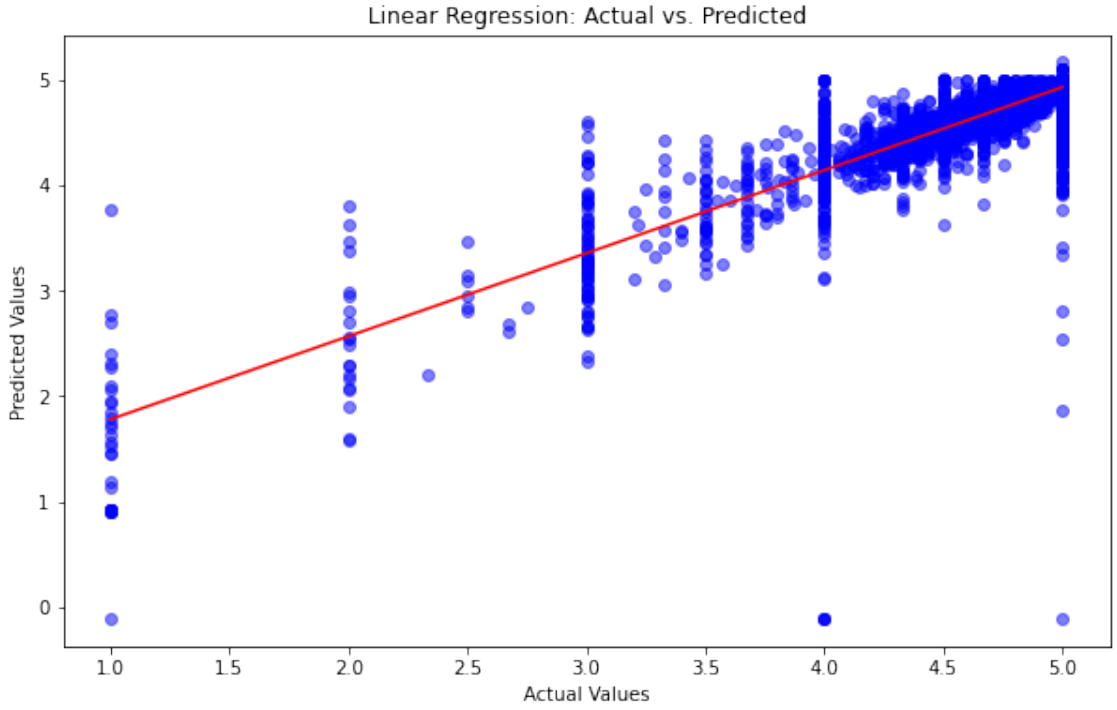
The features coefficients on the first run of the model (listed below) show that several features have 0.001 or negative coefficients, meaning that they have far less effect on the model than others like host listings count and profile pic. With this information I have removed the highlighted features to test the model performance and its improvement.

Feature Coefficients (First Run): host_is_superhost: 0.0 host_acceptance_rate: -0.017870799378971205 host_listings_count: 15909493113.91108 host_total_listings_count: -15909493113.911064 host_has_profile_pic: 43865806.82442311 host_identity_verified: 538216.3793221746 review_scores_cleanliness: 0.2733488389741146 review_scores_checkin: 0.12533477924255887 review_scores_communication: 0.21741720520857047 review_scores_value: 0.39600043759280285 calculated_host_listings_count: -0.001724280749587267 calculated_host_listings_count_entire_homes: 0.002142537521598753 calculated_host_listings_count_private_rooms: 0.001779032040259702 calculated_host_listings_count_shared_rooms: -0.006216212643633236 n_host_verifications: 0.002193321128857219

[153]:
```
plt.figure(figsize = (10, 6))
plt.scatter(y_test, predictions, color = 'blue', alpha = 0.5)
plt.plot(np.unique(y_test), np.poly1d(np.polyfit(y_test, predictions, 1))(np.
 ↪unique(y_test)), color='red')
plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.title('Linear Regression: Actual vs. Predicted')
```

```
plt.show()

#While the earlier values have a wide range, then the clusters get less␣
 ↪disselved
```



The model is working efficiently as the R2 value is at a 0.73 after removing the features with little importance, which means that 73% of the data is contributing to the predictions. The other data might be outliers or the model is not able to pick on the patterns within the data. To rectify this I will go back to the data prep stage and remove or winzsorize outliers and examine the resulting model performance.

After winsorizing the data values, they made no change to the performance of the model. This is because, as I now realize, that since the data values are on a 1-5 scale, there would be no outliers. I have now commented that section.

This is the best model I could produce by using Linear Regression. In the future, as a further analysis, it would be beneficial to conduct a more thorough investigation. Perhaps a different, more elegant model like gradient boosted regression or other types of numerical models.

[ ]: