Lab 8 Assignment

July 30, 2025

1 Lab 8: Define and Solve an ML Problem of Your Choosing

[1]:

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

In this lab assignment, you will follow the machine learning life cycle and implement a model to solve a machine learning problem of your choosing. You will select a data set and choose a predictive problem that the data set supports. You will then inspect the data with your problem in mind and begin to formulate a project plan. You will then implement the machine learning project plan.

You will complete the following tasks:

1. Build Your DataFrame
2. Define Your ML Problem
3. Perform exploratory data analysis to understand your data.
4. Define Your Project Plan
5. Implement Your Project Plan:
    - Prepare your data for your model.
    - Fit your model to the training data and evaluate your model.
    - Improve your model's performance.

1.1 Part 1: Build Your DataFrame

You will have the option to choose one of four data sets that you have worked with in this program:

- The ”census” data set that contains Census information from 1994: `censusData.csv`
- Airbnb NYC ”listings” data set: `airbnbListingsData.csv`
- World Happiness Report (WHR) data set: `WHR2018Chapter2OnlineData.csv`
- Book Review data set: `bookReviewsData.csv`

Note that these are variations of the data sets that you have worked with in this program. For example, some do not include some of the preprocessing necessary for specific models.

Load a Data Set and Save it as a Pandas DataFrame The code cell below contains filenames (path + filename) for each of the four data sets available to you.

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

You can load each file as a new DataFrame to inspect the data before choosing your data set.

[2]:
```python
import os
import pandas as pd
# File names of the four data sets
adultDataSet_filename = os.path.join(os.getcwd(), "data", "censusData.csv")
airbnbDataSet_filename = os.path.join(os.getcwd(), "data", "airbnbListingsData.
 ↪csv")
WHRDataSet_filename = os.path.join(os.getcwd(), "data",␣
 ↪"WHR2018Chapter2OnlineData.csv")
bookReviewDataSet_filename = os.path.join(os.getcwd(), "data", "bookReviewsData.
 ↪csv")


df = pd.read_csv(airbnbDataSet_filename)

df.head()
```

[2]:
```
                                                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

                                         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 don’t expect the luxury here just a bas…
4  Our best guests are seeking a safe, clean, spa…

                               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

                       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
```

```
                                          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   

   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  …   

   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   

  instant_bookable calculated_host_listings_count  \
0            False                              3   
1            False                              1   
2            False                              1   
3            False                              1   
4            False                              1   

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

   calculated_host_listings_count_private_rooms  \
0                                             0   
1                                             0   
2                                             0   
3                                             1   
4                                             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

  n_host_verifications
0                    9
1                    6
2                    3
3                    4
4                    7

[5 rows x 50 columns]
```

1.2 Part 2: Define Your ML Problem

Next you will formulate your ML Problem. In the markdown cell below, answer the following questions:

1. List the data set you have chosen.
2. What will you be predicting? What is the label?
3. Is this a supervised or unsupervised learning problem? Is this a clustering, classification or regression problem? Is it a binary classificaiton or multi-class classifiction problem?
4. What are your features? (note: this list may change after your explore your data)
5. Explain why this is an important problem. In other words, how would a company create value with a model that predicts this label?

Dataset Chosen: Airbnb NYC Listings (airbnbListingsData.csv)

Prediction Goal / Label: I will be predicting the price of a listing per night.

Type of ML Problem:

This is a supervised learning problem.

It is a regression problem because the label (price) is a continuous numeric value.

Initial Feature Set (subject to change after EDA):

neighbourhood_group (e.g., Manhattan, Brooklyn)

neighbourhood

room_type

minimum_nights

number_of_reviews

reviews_per_month

availability_365

latitude, longitude

1.3 Part 3: Understand Your Data

The next step is to perform exploratory data analysis. Inspect and analyze your data set with your machine learning problem in mind. Consider the following as you inspect your data:

1. What data preparation techniques would you like to use? These data preparation techniques may include:

   - addressing missingness, such as replacing missing values with means
   - finding and replacing outliers
   - renaming features and labels
   - finding and replacing outliers
   - performing feature engineering techniques such as one-hot encoding on categorical features
   - selecting appropriate features and removing irrelevant features
   - performing specific data cleaning and preprocessing techniques for an NLP problem
   - addressing class imbalance in your data sample to promote fair AI

2. What machine learning model (or models) you would like to use that is suitable for your predictive problem and data?

   - Are there other data preparation techniques that you will need to apply to build a balanced modeling data set for your problem and model? For example, will you need to scale your data?

3. How will you evaluate and improve the model's performance?

   - Are there specific evaluation metrics and methods that are appropriate for your model?

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. You can import additional packages that you have used in this course that you will need to perform this task.

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.

[3]:

```
# Import required libraries
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the Airbnb dataset
```

```
airbnbDataSet_filename = os.path.join(os.getcwd(), "data", "airbnbListingsData.
 ↪csv")
df = pd.read_csv(airbnbDataSet_filename)

# Display basic info
print("Shape of dataset:", df.shape)
display(df.head())

# Check for missing values
print("\nMissing Values:\n", df.isnull().sum())

# Check data types
print("\nData Types:\n", df.dtypes)

# Summary statistics
print("\nDescriptive Stats:\n")
display(df.describe())

# Check number of unique values in each column
print("\nUnique Values:\n", df.nunique())

# Check price distribution
plt.figure(figsize=(8, 4))
sns.histplot(df['price'], bins=100, kde=True)
plt.xlim(0, 500)  # Clip outliers for better visualization
plt.title("Price Distribution")
plt.show()

# Check for outliers in minimum_nights
plt.figure(figsize=(8, 2))
sns.boxplot(x=df['minimum_nights'])
plt.xlim(0, 100)  # Clip for visualization
plt.title("Minimum Nights")
plt.show()

# Boxplot: Price by room_type
plt.figure(figsize=(6, 4))
sns.boxplot(x='room_type', y='price', data=df)
plt.ylim(0, 500)
plt.title("Price by Room Type")
plt.show()

# Correlation heatmap
plt.figure(figsize=(10, 6))
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()

# Basic cleaning steps:

# Fill missing reviews_per_month with 0
df['reviews_per_month'] = df['reviews_per_month'].fillna(0)

# Drop unnecessary columns
df.drop(columns=['name', 'host_name', 'last_review', 'id'], inplace=True)

# Remove listings with extreme prices and minimum_nights
df = df[(df['price'] <= 500) & (df['minimum_nights'] <= 365)]

# One-hot encode categorical features
df = pd.get_dummies(df, columns=['neighbourhood_group', 'room_type'],␣
 ↪drop_first=True)

# Confirm final structure
print("\nCleaned DataFrame Shape:", df.shape)
display(df.head())
```

```
Shape of dataset: (28022, 50)
                                                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

                                         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 don’t expect the luxury here just a bas...
4  Our best guests are seeking a safe, clean, spa...

                               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

                       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

                                          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

   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  ...

   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

  instant_bookable calculated_host_listings_count  \
0            False                              3
1            False                              1
2            False                              1
3            False                              1
4            False                              1

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

   calculated_host_listings_count_private_rooms  \
0                                             0
1                                             0
2                                             0
3                                             1
4                                             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

  n_host_verifications
0                    9
1                    6
2                    3
3                    4
4                    7

[5 rows x 50 columns]


Missing Values:
 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
```

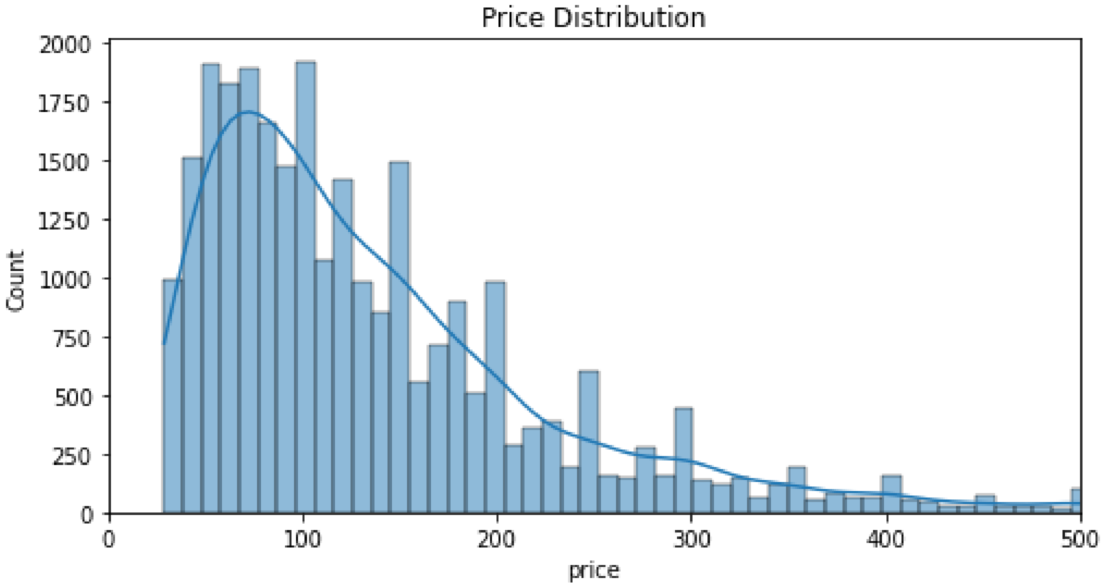
```
availability_90                                     0
availability_365                                    0
number_of_reviews                                   0
number_of_reviews_ltm                               0
number_of_reviews_l30d                              0
review_scores_rating                                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

Data Types:
 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
```

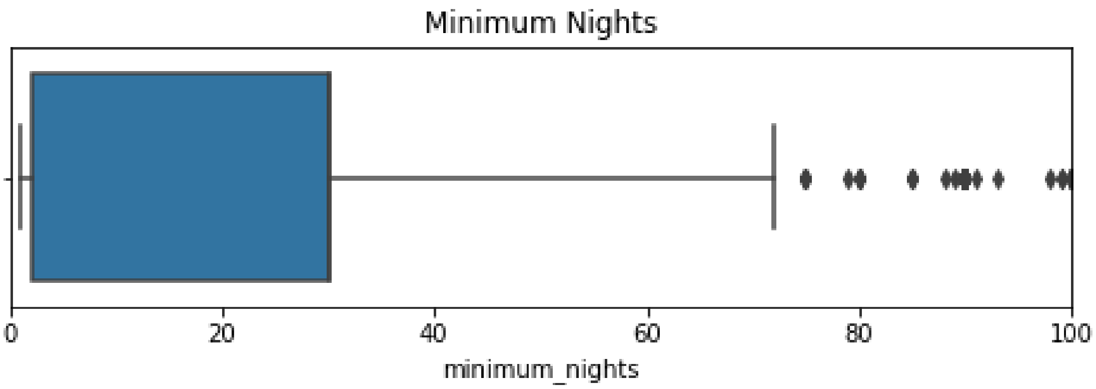
```
minimum_nights_avg_ntm                          float64
maximum_nights_avg_ntm                          float64
has_availability                                   bool
availability_30                                   int64
availability_60                                   int64
availability_90                                   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

Descriptive Stats:

       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
```

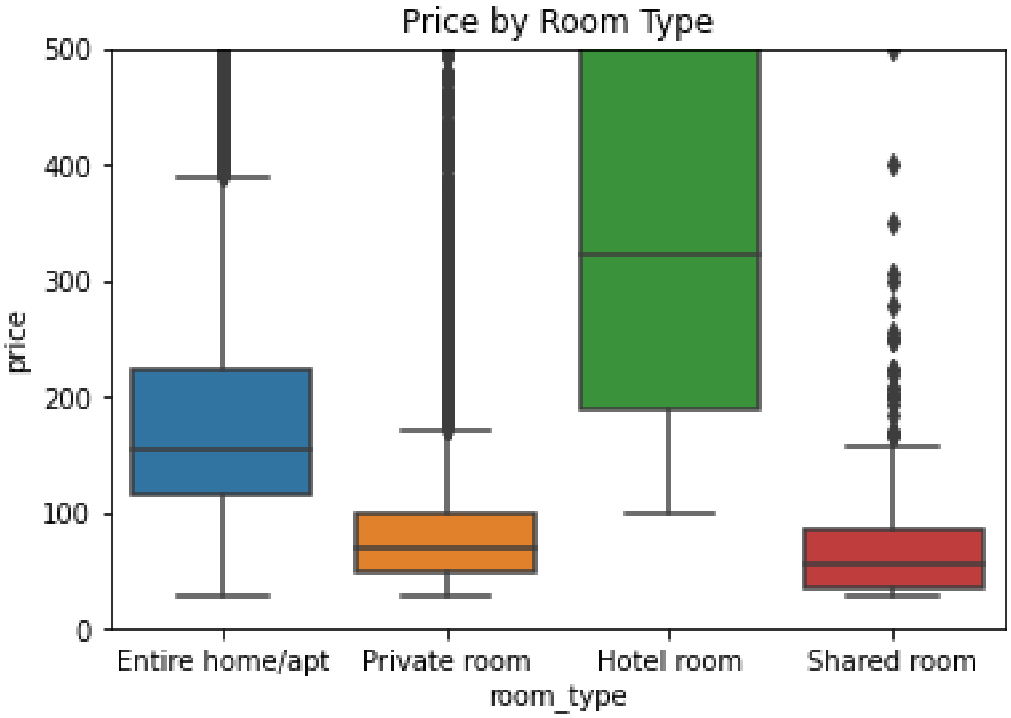
```
               beds         price  minimum_nights  ...  review_scores_checkin  \
count  26668.000000  28022.000000    28022.000000  ...           28022.000000   
mean       1.629556    154.228749       18.689387  ...               4.814300   
std        1.097104    140.816605       25.569151  ...               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]


Unique Values:
 name                                            27386
description                                     25952
neighborhood_overview                           15800
host_name                                        7566
host_location                                    1364
host_about                                      11962
host_response_rate                                 85
host_acceptance_rate                              101
host_is_superhost                                   1
host_listings_count                                73
host_total_listings_count                          73
host_has_profile_pic                                1
host_identity_verified                              1
neighbourhood_group_cleansed                        5
room_type                                           4
accommodates                                       16
bathrooms                                          16
bedrooms                                           11
beds                                               16
amenities                                       25020
price                                             684
```

```
minimum_nights                                     95
maximum_nights                                    229
minimum_minimum_nights                             98
maximum_minimum_nights                            102
minimum_maximum_nights                            206
maximum_maximum_nights                            206
minimum_nights_avg_ntm                            329
maximum_nights_avg_ntm                            452
has_availability                                    2
availability_30                                    31
availability_60                                    61
availability_90                                    91
availability_365                                  366
number_of_reviews                                 418
number_of_reviews_ltm                             140
number_of_reviews_l30d                             29
review_scores_rating                              154
review_scores_cleanliness                         196
review_scores_checkin                             135
review_scores_communication                       141
review_scores_location                            153
review_scores_value                               164
instant_bookable                                    2
calculated_host_listings_count                     59
calculated_host_listings_count_entire_homes        44
calculated_host_listings_count_private_rooms       45
calculated_host_listings_count_shared_rooms         9
reviews_per_month                                1357
n_host_verifications                               13
dtype: int64
```







```
        ␣
↪---------------------------------------------------------------------------

       TypeError                                 Traceback (most recent call␣
↪last)
```

```
    <ipython-input-3-0f846d227e34> in <module>()
     50 # Correlation heatmap
     51 plt.figure(figsize=(10, 6))
---> 52 corr = df.corr(numeric_only=True)
     53 sns.heatmap(corr, annot=True, cmap='coolwarm')
     54 plt.title("Correlation Heatmap")


    TypeError: corr() got an unexpected keyword argument 'numeric_only'
```

```
<Figure size 720x432 with 0 Axes>
```

1.4 Part 4: Define Your Project Plan

Now that you understand your data, in the markdown cell below, define your plan to implement the remaining phases of the machine learning life cycle (data preparation, modeling, evaluation) to solve your ML problem. Answer the following questions:

- Do you have a new feature list? If so, what are the features that you chose to keep and remove after inspecting the data?
- Explain different data preparation techniques that you will use to prepare your data for modeling.
- What is your model (or models)?
- Describe your plan to train your model, analyze its performance and then improve the model. That is, describe your model building, validation and selection plan to produce a model that generalizes well to new data.

For this project, I chose to retain features that are most relevant to predicting Airbnb listing prices, including latitude, longitude, minimum_nights, number_of_reviews, reviews_per_month, and availability_365, along with one-hot encoded versions of neighbourhood_group and room_type. I removed non-predictive or high-cardinality columns such as id, name, host_name, last_review, and neighbourhood. To prepare the data for modeling, I handled missing values by filling reviews_per_month with 0, removed extreme outliers from price and minimum_nights, and applied one-hot encoding to categorical variables. My primary model will be a Random Forest Regressor due to its robustness with non-linear data and mixed feature types. I will also use a Linear Regression model as a baseline and may explore Gradient Boosting for further improvement. I plan to split the data into training and test sets, train and evaluate the models using RMSE, MAE, and R^2 metrics, and perform hyperparameter tuning through cross-validation to optimize model performance. My goal is to select a model that generalizes well to new data and provides actionable insights into what factors most impact Airbnb pricing.

1.5 Part 5: Implement Your Project Plan

Task: In the code cell below, import additional packages that you have used in this course that you will need to implement your project plan.

[4]:

```
# Data manipulation and analysis
import pandas as pd
import numpy as np

# Data visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Data preprocessing
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler  # Optional for linear models
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Machine learning models
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor

# Model tuning (optional)
from sklearn.model_selection import GridSearchCV, cross_val_score

# Set plot style
sns.set(style='whitegrid')
```

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

You will:

1. Prepare your data for your model.
2. Fit your model to the training data and evaluate your model.
3. Improve your model's performance 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.

[7]:

```
# 1. Import all required packages
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Set plotting style
```

```
sns.set(style='whitegrid')

# 2. Load the Airbnb dataset
airbnbDataSet_filename = os.path.join(os.getcwd(), "data", "airbnbListingsData.
 ↪csv")
df = pd.read_csv(airbnbDataSet_filename)
print("Original shape:", df.shape)
display(df.head())

# 3. Basic cleaning

# Fill missing reviews_per_month with 0
if 'reviews_per_month' in df.columns:
    df['reviews_per_month'] = df['reviews_per_month'].fillna(0)

# Drop irrelevant columns if they exist
cols_to_drop = ['id', 'name', 'host_name', 'last_review']
existing_cols_to_drop = [col for col in cols_to_drop if col in df.columns]
df.drop(columns=existing_cols_to_drop, inplace=True)

# Remove outliers
df = df[(df['price'] <= 500) & (df['minimum_nights'] <= 365)]

# One-hot encode only the columns that exist
cols_to_encode = ['neighbourhood_group', 'room_type']
existing_cols = [col for col in cols_to_encode if col in df.columns]
df = pd.get_dummies(df, columns=existing_cols, drop_first=True)

print("Cleaned shape:", df.shape)
display(df.head())

# 4. Split data into features and label
X = df.drop('price', axis=1)
y = df['price']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)

# Scale data (only for Linear Regression)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# 5. Train and evaluate Linear Regression
lr = LinearRegression()
lr.fit(X_train_scaled, y_train)
```

```
y_pred_lr = lr.predict(X_test_scaled)

print("Linear Regression Performance:")
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred_lr)))
print("MAE:", mean_absolute_error(y_test, y_pred_lr))
print("R^2:", r2_score(y_test, y_pred_lr))

# 6. Train and evaluate Random Forest Regressor
rf = RandomForestRegressor(random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

print("\nRandom Forest Performance:")
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred_rf)))
print("MAE:", mean_absolute_error(y_test, y_pred_rf))
print("R^2:", r2_score(y_test, y_pred_rf))

# 7. Visualize feature importances
importances = rf.feature_importances_
features = X.columns
importance_df = pd.DataFrame({'Feature': features, 'Importance': importances})
importance_df.sort_values(by='Importance', ascending=False, inplace=True)

plt.figure(figsize=(10, 6))
sns.barplot(data=importance_df.head(10), x='Importance', y='Feature')
plt.title('Top 10 Feature Importances (Random Forest)')
plt.show()

# 8. Hyperparameter tuning for Random Forest
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20, None],
    'min_samples_split': [2, 5],
}

grid_search = GridSearchCV(RandomForestRegressor(random_state=42), param_grid,
                           cv=3, scoring='neg_root_mean_squared_error',␣
 ↪n_jobs=-1)
grid_search.fit(X_train, y_train)
best_rf = grid_search.best_estimator_
y_pred_best = best_rf.predict(X_test)

print("\nTuned Random Forest Performance:")
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred_best)))
print("MAE:", mean_absolute_error(y_test, y_pred_best))
print("R^2:", r2_score(y_test, y_pred_best))
```

```python
# 9. Final Summary
print("\nModel Comparison Summary")
print("--------------------------------------------------")
print(f"Linear Regression R^2: {r2_score(y_test, y_pred_lr):.3f}")
print(f"Random Forest R^2: {r2_score(y_test, y_pred_rf):.3f}")
print(f"Tuned Random Forest R^2: {r2_score(y_test, y_pred_best):.3f}")
```

```
Original shape: (28022, 50)
                                                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

                                         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 don’t expect the luxury here just a bas...
4  Our best guests are seeking a safe, clean, spa...

                               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

                       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

                                          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

   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  ...

   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

  instant_bookable calculated_host_listings_count  \
0            False                              3
1            False                              1
2            False                              1
3            False                              1
4            False                              1

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

   calculated_host_listings_count_private_rooms  \
0                                             0
1                                             0
2                                             0
3                                             1
4                                             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

  n_host_verifications
0                    9
1                    6
2                    3
3                    4
4                    7

[5 rows x 50 columns]
```

```
Cleaned shape: (27184, 50)

                                         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 don’t expect the luxury here just a bas...
4  Our best guests are seeking a safe, clean, spa...

                               neighborhood_overview  \
0  Centrally located in the heart of Manhattan ju...
1  Just the right mix of urban center and local n...
2                                                NaN
3    Theater district, many restaurants around here.
4  Our neighborhood is full of restaurants and ca...

                       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

                                          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

   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

   host_total_listings_count  host_has_profile_pic  ...  instant_bookable  \
0                        8.0                  True  ...             False
1                        1.0                  True  ...             False
2                        1.0                  True  ...             False
3                        1.0                  True  ...             False
4                        1.0                  True  ...             False

  calculated_host_listings_count  calculated_host_listings_count_entire_homes  \
0                              3                                            3
1                              1                                            1
2                              1                                            1
```

```
3                              1                                             0
4                              1                                             0

   calculated_host_listings_count_private_rooms  \
0                                             0
1                                             0
2                                             0
3                                             1
4                                             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

  n_host_verifications  room_type_Hotel room  room_type_Private room  \
0                    9                     0                       0
1                    6                     0                       0
2                    3                     0                       0
3                    4                     0                       1
4                    7                     0                       1

   room_type_Shared room
0                      0
1                      0
2                      0
3                      0
4                      0

[5 rows x 50 columns]
```

```
       ␣
 ↪---------------------------------------------------------------------------

        ValueError                                Traceback (most recent call␣
 ↪last)

        <ipython-input-7-167920ad036d> in <module>()
         52 # Scale data (only for Linear Regression)
         53 scaler = StandardScaler()
    ---> 54 X_train_scaled = scaler.fit_transform(X_train)
         55 X_test_scaled = scaler.transform(X_test)
         56
    /usr/local/lib/python3.6/dist-packages/sklearn/base.py in␣
↪fit_transform(self, X, y, **fit_params)
    569         if y is None:
    570             # fit method of arity 1 (unsupervised transformation)
--> 571             return self.fit(X, **fit_params).transform(X)
    572         else:
    573             # fit method of arity 2 (supervised transformation)


    /usr/local/lib/python3.6/dist-packages/sklearn/preprocessing/_data.py in␣
↪fit(self, X, y)
    667         # Reset internal state before fitting
    668         self._reset()
--> 669         return self.partial_fit(X, y)
    670
    671     def partial_fit(self, X, y=None):


    /usr/local/lib/python3.6/dist-packages/sklearn/preprocessing/_data.py in␣
↪partial_fit(self, X, y)
    698         X = check_array(X, accept_sparse=('csr', 'csc'),
    699                         estimator=self, dtype=FLOAT_DTYPES,
--> 700                         force_all_finite='allow-nan')
    701
    702         # Even in the case of `with_mean=False`, we update the mean␣
↪anyway


    /usr/local/lib/python3.6/dist-packages/sklearn/utils/validation.py in␣
↪check_array(array, accept_sparse, accept_large_sparse, dtype, order, copy,␣
↪force_all_finite, ensure_2d, allow_nd, ensure_min_samples,␣
↪ensure_min_features, warn_on_dtype, estimator)
    529                     array = array.astype(dtype, casting="unsafe",␣
↪copy=False)
    530                 else:
--> 531                     array = np.asarray(array, order=order,␣
↪dtype=dtype)
    532             except ComplexWarning:
    533                 raise ValueError("Complex data not supported\n"


    ~/.local/lib/python3.6/site-packages/numpy/core/_asarray.py in␣
↪asarray(a, dtype, order)
     81
     82     """
---> 83     return array(a, dtype, copy=False, order=order)
```

```
        84
        85


      ~/.local/lib/python3.6/site-packages/pandas/core/generic.py in␣
↪__array__(self, dtype)
      1779
      1780     def __array__(self, dtype=None) -> np.ndarray:
   -> 1781         return np.asarray(self._values, dtype=dtype)
      1782
      1783     def __array_wrap__(self, result, context=None):


      ~/.local/lib/python3.6/site-packages/numpy/core/_asarray.py in␣
↪asarray(a, dtype, order)
        81
        82     """
   ---> 83     return array(a, dtype, copy=False, order=order)
        84
        85


      ValueError: could not convert string to float: "Fantastic Hudson Yards 3␣
↪Bed 1 Bath located just a short walk to Times Square 42nd Street. Amazing␣
↪access to all that makes NYC great. Just two flights in a traditional NYC walk␣
↪up building. Beautiful character throughout with original brick wall and␣
↪hardwood accents. Two queen beds and one full each with space to store␣
↪belongings. High Speed WiFi and Flat Screen TV with Netflix provided. Dining␣
↪table seats 6 and kitchen has all you'll need to prepare meal at home."
```