model2

November 27, 2024

1 Model 2

1.1 KNN -> K-means -> XGBoost on entire dataset with one-hot cluster encodings

[ ]:
```python
import pandas as pd
import numpy as np
import sklearn as sk
import random
from sklearn.neighbors import NearestNeighbors
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics.pairwise import euclidean_distances
from sklearn.preprocessing import MinMaxScaler
import xgboost as xgb
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import KFold
```

1.2 Merge the train and test datsets

[ ]:
```python
pd.set_option('display.max_columns', None)

# Load the dfs
train_df = pd.read_csv('train_processed.csv')
test_df = pd.read_csv('test_processed.csv')

# Find the lengths of the train and test dataframes
train_size = len(train_df)
test_size = len(test_df)

# combine them for unsupervised learning
combined_df = pd.concat([train_df, test_df], axis=0)
```

```python
# save the price and id columns for later
price_column = combined_df['price']
id_column = combined_df['id']

combined_df = combined_df.drop(columns=['price','id'])

combined_df = combined_df.reset_index(drop=True)
price_column = price_column.reset_index(drop=True)
id_column = id_column.reset_index(drop=True)

combined_df
```

[ ]:

```
        latitude  longitude  host_since  host_response_time  \
0      40.684560 -73.939870       16578                 2.0
1      40.638991 -73.965739       19614                 1.0
2      40.618810 -74.032380       19204                 1.0
3      40.673970 -73.953990       15563                 1.0
4      40.747180 -73.985390       16427                 1.0
…            …          …         …                   …
22418  40.637960 -73.951360       16690                 1.0
22419  40.823720 -73.945460       16877                 4.0
22420  40.755094 -73.937260       15563                 1.0
22421  40.781580 -73.984780       15351                 1.0
22422  40.695890 -73.760830       17286                 1.0

       host_response_rate  host_acceptance_rate  host_is_superhost  \
0                   100.0                 100.0                  1
1                   100.0                  98.0                  1
2                   100.0                 100.0                  0
3                    99.0                  23.0                  0
4                    93.0                  95.0                  0
…                     …                     …                 …
22418               100.0                 100.0                  0
22419                 0.0                  33.0                  0
22420                99.0                  23.0                  0
22421               100.0                 100.0                  1
22422               100.0                 100.0                  0

       host_listings_count  host_total_listings_count  host_has_profile_pic  \
0                      2.0                        2.0                     1
1                      1.0                        1.0                     1
2                     52.0                       55.0                     1
3                    727.0                     1336.0                     1
4                    707.0                     2453.0                     1
…                      …                        …                    …
22418                  2.0                        2.0                     1
22419                  7.0                        8.0                     1
```

```
22420             727.0                        1336.0                    1
22421               1.0                           3.0                    1
22422               3.0                           3.0                    1

       host_identity_verified  calculated_host_listings_count  accommodates  \
0                           1                               1             4
1                           1                               1             2
2                           1                              52             2
3                           1                             719             1
4                           1                              73             2
…                         …                             …             …
22418                       1                               2             2
22419                       0                               7             1
22420                       1                             719             1
22421                       1                               1             2
22422                       1                               3             2

       bathrooms  bedrooms  beds  has_availability  availability_30  \
0            2.0       2.0   2.0                 1               12
1            1.0       1.0   2.0                 1               10
2            1.0       0.0   1.0                 1               17
3            1.5       4.0   1.0                 1                0
4            1.0       1.0   1.0                 1                4
…            …         …     …               …                …
22418        1.0       2.0   6.0                 1               30
22419        1.0       1.0   1.0                 1                0
22420        3.0       5.0   1.0                 1                3
22421        1.0       1.0   1.0                 1               18
22422        1.0       1.0   1.0                 1               30

       availability_60  availability_90  availability_365  instant_bookable  \
0                   42               70                70             False
1                   20               49               324             False
2                   44               70               146              True
3                    0                0               111             False
4                   13               22               241              True
…                …                …                 …               …
22418               60               90               365             False
22419                1                1                97             False
22420               33               63               159             False
22421               35               35                35             False
22422               60               90               180             False

       minimum_nights  maximum_nights  number_of_reviews  \
0                  30            1125                 34
1                   1              29                 30
2                   1              29                  5
```

3	30	365	0
4	1	1125	0
…	…	…	…
22418	30	1125	325
22419	30	365	11
22420	30	365	0
22421	30	30	33
22422	30	90	19

	number_of_reviews_ltm	number_of_reviews_l30d	first_review \
0	5	1	18014
1	30	6	19735
2	5	2	19901
3	0	0	20049
4	0	0	20049
…	…	…	…
22418	9	0	16729
22419	2	0	17774
22420	0	0	20049
22421	6	0	18962
22422	2	0	19121

	last_review	review_scores_rating	review_scores_accuracy \
0	19945	5.000000	5.000000
1	19968	4.830000	4.870000
2	19952	4.600000	4.800000
3	20049	4.719393	4.742812
4	20049	4.719393	4.742812
…	…	…	…
22418	19679	4.710000	4.750000
22419	19875	4.360000	4.550000
22420	20049	4.711449	4.748402
22421	19906	4.880000	4.940000
22422	19813	4.740000	4.680000

	review_scores_cleanliness	review_scores_checkin \
0	4.970000	5.000000
1	4.930000	4.800000
2	4.200000	4.800000
3	4.679642	4.826310
4	4.679642	4.826310
…	…	…
22418	4.640000	4.880000
22419	4.550000	4.910000
22420	4.673496	4.813213
22421	4.730000	4.940000
22422	4.790000	4.630000

```
       review_scores_communication  review_scores_location  \
0                         5.000000                4.710000
1                         4.900000                4.900000
2                         4.800000                4.800000
3                         4.808233                4.721844
4                         4.808233                4.721844
…                              …                     …
22418                     4.900000                4.650000
22419                     4.550000                4.910000
22420                     4.800491                4.710323
22421                     5.000000                4.850000
22422                     4.680000                4.740000

       review_scores_value  reviews_per_month  room_Entire home/apt  \
0                 4.940000           0.520000                     1
1                 4.630000           3.810000                     0
2                 4.200000           2.140000                     1
3                 4.609505           1.245801                     0
4                 4.609505           1.245801                     0
…                      …                  …                  …
22418             4.710000           3.010000                     0
22419             4.450000           0.150000                     0
22420             4.604796           1.234094                     0
22421             4.760000           0.980000                     1
22422             4.680000           0.670000                     0

       room_Hotel room  room_Private room  room_Shared room  Air conditioning  \
0                    0                  0                 0                 0
1                    0                  1                 0                 1
2                    0                  0                 0                 1
3                    0                  1                 0                 0
4                    1                  0                 0                 1
…                 …                  …                 …                 …
22418                0                  1                 0                 0
22419                0                  1                 0                 1
22420                0                  1                 0                 1
22421                0                  0                 0                 1
22422                0                  1                 0                 0

       Kitchen  Dedicated workspace  Heating  Hot water  Refrigerator  \
0            1                    0        1          1             1
1            1                    0        0          1             0
2            0                    1        1          1             0
3            1                    1        1          1             1
4            0                    1        1          1             0
…         …                  …       …         …            …
```

```
22418        1                    1        1          1             1
22419        0                    1        1          0             0
22420        1                    1        1          1             1
22421        1                    1        0          1             1
22422        0                    1        0          1             1

       Free street parking  Self check-in  Shampoo  Washer
0                        1              1        1       0
1                        1              1        1       0
2                        0              1        1       0
3                        0              0        0       1
4                        0              1        1       0
…                   …            …       …     …
22418                    1              1        1       0
22419                    0              1        0       0
22420                    0              0        0       0
22421                    0              0        1       0
22422                    1              1        1       0

[22423 rows x 51 columns]
```

1.3 Normalize the dataset

[ ]:
```python
# Initialize MinMaxScaler
scaler = MinMaxScaler()

# Normalize all columns
combined_df[:] = scaler.fit_transform(combined_df)

combined_df
```

```
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.42223738 0.93997271 0.87005457 …
0.24914734 0.21299454 0.54297408]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.         0.         0.05828571 …
0.82057143 0.         0.00228571]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.2        0.06666667 0.06666667 …
0.         0.06666667 0.06666667]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
```

```
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.4        0.33333333 0.56666667 …
0.1        0.6        1.        ]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.7        0.33333333 0.73333333 …
0.55       0.58333333 1.        ]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.77777778 0.54444444 0.77777778 …
0.7        0.38888889 1.        ]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.19178082 0.88767123 0.4        …
0.43561644 0.09589041 0.49315068]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0. 0. 1. … 0. 0. 0.]' has dtype
incompatible with bool, please explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.05811623 0.         0.         …
0.05811623 0.05811623 0.05811623]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.11241124 0.00280028 0.00280028 …
0.03640364 0.00290029 0.00890089]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.01751674 0.01545595 0.00257599 …
0.         0.01700155 0.00978877]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
```

```
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.00282167 0.01693002 0.00282167 …
0.         0.003386   0.00112867]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.00680272 0.04081633 0.01360544 …
0.         0.         0.        ]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.64045936 0.94452297 0.97385159 …
1.         0.80795053 0.8360424 ]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/2371048953.py:5
: FutureWarning: Setting an item of incompatible dtype is deprecated and will
raise in a future error of pandas. Value '[0.97801733 0.98287888 0.97949694 …
1.         0.96977383 0.95011625]' has dtype incompatible with int64, please
explicitly cast to a compatible dtype first.
  combined_df[:] = scaler.fit_transform(combined_df)
```

[ ]:

	latitude	longitude	host_since	host_response_time	\
0	0.448134	0.579717	0.422237	0.333333	
1	0.337268	0.531656	0.939973	0.000000	
2	0.288168	0.407848	0.870055	0.000000	
3	0.422369	0.553485	0.249147	0.000000	
4	0.600485	0.495148	0.396487	0.000000	
…	…	…	…	…	
22418	0.334759	0.558371	0.441337	0.000000	
22419	0.786703	0.569332	0.473226	1.000000	
22420	0.619740	0.584566	0.249147	0.000000	
22421	0.684179	0.496282	0.212995	0.000000	
22422	0.475699	0.912347	0.542974	0.000000	

	host_response_rate	host_acceptance_rate	host_is_superhost	\
0	1.00	1.00	1	
1	1.00	0.98	1	
2	1.00	1.00	0	
3	0.99	0.23	0	
4	0.93	0.95	0	
…	…	…	…	
22418	1.00	1.00	0	
22419	0.00	0.33	0	

```
22420             0.99               0.23               0
22421             1.00               1.00               1
22422             1.00               1.00               0

       host_listings_count  host_total_listings_count  host_has_profile_pic  \
0                 0.000223                   0.000111                     1
1                 0.000000                   0.000000                     1
2                 0.011351                   0.005988                     1
3                 0.161585                   0.148037                     1
4                 0.157133                   0.271901                     1
…                    …                          …                      …
22418             0.000223                   0.000111                     1
22419             0.001335                   0.000776                     1
22420             0.161585                   0.148037                     1
22421             0.000000                   0.000222                     1
22422             0.000445                   0.000222                     1

       host_identity_verified  calculated_host_listings_count  accommodates  \
0                           1                        0.000000      0.200000
1                           1                        0.000000      0.066667
2                           1                        0.058286      0.066667
3                           1                        0.820571      0.000000
4                           1                        0.082286      0.066667
…                        …                              …             …
22418                       1                        0.001143      0.066667
22419                       0                        0.006857      0.000000
22420                       1                        0.820571      0.000000
22421                       1                        0.000000      0.066667
22422                       1                        0.002286      0.066667

       bathrooms  bedrooms    beds  has_availability  availability_30  \
0       0.173913  0.222222  0.1250                 1         0.400000
1       0.086957  0.111111  0.1250                 1         0.333333
2       0.086957  0.000000  0.0625                 1         0.566667
3       0.130435  0.444444  0.0625                 1         0.000000
4       0.086957  0.111111  0.0625                 1         0.133333
…          …         …     …                 …              …
22418   0.086957  0.222222  0.3750                 1         1.000000
22419   0.086957  0.111111  0.0625                 1         0.000000
22420   0.260870  0.555556  0.0625                 1         0.100000
22421   0.086957  0.111111  0.0625                 1         0.600000
22422   0.086957  0.111111  0.0625                 1         1.000000

       availability_60  availability_90  availability_365  instant_bookable  \
0             0.700000         0.777778          0.191781               0.0
1             0.333333         0.544444          0.887671               0.0
2             0.733333         0.777778          0.400000               1.0
```

3	0.000000	0.000000	0.304110	0.0
4	0.216667	0.244444	0.660274	1.0
…	…	…	…	…
22418	1.000000	1.000000	1.000000	0.0
22419	0.016667	0.011111	0.265753	0.0
22420	0.550000	0.700000	0.435616	0.0
22421	0.583333	0.388889	0.095890	0.0
22422	1.000000	1.000000	0.493151	0.0

	minimum_nights	maximum_nights	number_of_reviews \
0	0.058116	0.112411	0.017517
1	0.000000	0.002800	0.015456
2	0.000000	0.002800	0.002576
3	0.058116	0.036404	0.000000
4	0.000000	0.112411	0.000000
…	…	…	…
22418	0.058116	0.112411	0.167439
22419	0.058116	0.036404	0.005667
22420	0.058116	0.036404	0.000000
22421	0.058116	0.002900	0.017002
22422	0.058116	0.008901	0.009789

	number_of_reviews_ltm	number_of_reviews_l30d	first_review \
0	0.002822	0.006803	0.640459
1	0.016930	0.040816	0.944523
2	0.002822	0.013605	0.973852
3	0.000000	0.000000	1.000000
4	0.000000	0.000000	1.000000
…	…	…	…
22418	0.005079	0.000000	0.413428
22419	0.001129	0.000000	0.598057
22420	0.000000	0.000000	1.000000
22421	0.003386	0.000000	0.807951
22422	0.001129	0.000000	0.836042

	last_review	review_scores_rating	review_scores_accuracy \
0	0.978017	1.000000	1.000000
1	0.982879	0.957500	0.967500
2	0.979497	0.900000	0.950000
3	1.000000	0.929848	0.935703
4	1.000000	0.929848	0.935703
…	…	…	…
22418	0.921792	0.927500	0.937500
22419	0.963221	0.840000	0.887500
22420	1.000000	0.927862	0.937101
22421	0.969774	0.970000	0.985000
22422	0.950116	0.935000	0.920000

```
       review_scores_cleanliness  review_scores_checkin  \
0                       0.992500               1.000000   
1                       0.982500               0.950000   
2                       0.800000               0.950000   
3                       0.919910               0.956577   
4                       0.919910               0.956577   
…                            …                    …   
22418                   0.910000               0.970000   
22419                   0.887500               0.977500   
22420                   0.918374               0.953303   
22421                   0.932500               0.985000   
22422                   0.947500               0.907500   

       review_scores_communication  review_scores_location  \
0                         1.000000                0.927500   
1                         0.975000                0.975000   
2                         0.950000                0.950000   
3                         0.952058                0.930461   
4                         0.952058                0.930461   
…                              …                     …   
22418                     0.975000                0.912500   
22419                     0.887500                0.977500   
22420                     0.950123                0.927581   
22421                     1.000000                0.962500   
22422                     0.920000                0.935000   

       review_scores_value  reviews_per_month  room_Entire home/apt  \
0                 0.985000           0.004633                     1   
1                 0.907500           0.034517                     0   
2                 0.800000           0.019348                     1   
3                 0.902376           0.011225                     0   
4                 0.902376           0.011225                     0   
…                      …                  …                   …   
22418             0.927500           0.027250                     0   
22419             0.862500           0.001272                     0   
22420             0.901199           0.011119                     0   
22421             0.940000           0.008811                     1   
22422             0.920000           0.005995                     0   

       room_Hotel room  room_Private room  room_Shared room  Air conditioning  \
0                    0                  0                 0                 0   
1                    0                  1                 0                 1   
2                    0                  0                 0                 1   
3                    0                  1                 0                 0   
4                    1                  0                 0                 1   
…                  …                  …                …                …   
```

```
22418                    0                    1                  0                  0
22419                    0                    1                  0                  1
22420                    0                    1                  0                  1
22421                    0                    0                  0                  1
22422                    0                    1                  0                  0

       Kitchen  Dedicated workspace  Heating  Hot water  Refrigerator  \
0            1                    0        1          1             1
1            1                    0        0          1             0
2            0                    1        1          1             0
3            1                    1        1          1             1
4            0                    1        1          1             0
…          …                  …      …        …          …
22418        1                    1        1          1             1
22419        0                    1        1          0             0
22420        1                    1        1          1             1
22421        1                    1        0          1             1
22422        0                    1        0          1             1

       Free street parking  Self check-in  Shampoo  Washer
0                        1              1        1       0
1                        1              1        1       0
2                        0              1        1       0
3                        0              0        0       1
4                        0              1        1       0
…                      …            …      …     …
22418                    1              1        1       0
22419                    0              1        0       0
22420                    0              0        0       0
22421                    0              0        1       0
22422                    1              1        1       0

[22423 rows x 51 columns]
```

2 Unsupervised Learning: Neighborhood Analysis

2.1 Use KNN to find averaged neighborhood representations

[ ]:
```python
# Find k1 nearest neighbors
# Compute average representations of each property according to its k1 nearest
 ↪neighbors
def knn(combined_df, k1):
    # Extract longitude and latitude for KNN (ignore them when averaging
 ↪features)
    coordinates = combined_df[['longitude', 'latitude']].values

    # Initialize the KNN model (for 500 neighbors)
```

```
    knn = NearestNeighbors(n_neighbors=k1)

    # Fit the KNN model
    knn.fit(coordinates)

    # Step 5: Find the 20 nearest neighbors for each property (including itself)
    distances, indices = knn.kneighbors(coordinates)

    # Identify the feature columns
    feature_columns = combined_df.columns[1:]  # All columns starting from␣
 ↪index 2 onward

    # Create an empty array to store the averaged features
    averaged_features = np.zeros((combined_df.shape[0], len(feature_columns)))

    # For each property, average its own features and those of the 200 nearest␣
 ↪neighbors
    for i in range(combined_df.shape[0]):
        # Get the indices of the neighbors (including the property itself)
        neighbor_indices = indices[i]

        # Extract the features for the neighbors
        neighbor_features = combined_df.iloc[neighbor_indices][feature_columns]

        # Ensure the neighbor features are in the same order as feature_columns
        neighbor_features = neighbor_features[feature_columns]

        # Average the features of the property and its neighbors, handling␣
 ↪missing values
        averaged_features[i] = neighbor_features.mean(axis=0, skipna=True)

    # Create a DataFrame for the averaged features
    averaged_df = pd.DataFrame(averaged_features, columns=feature_columns)

    return averaged_df

# Run an example with k1 = 500
averaged_df = knn(combined_df, 500)
```

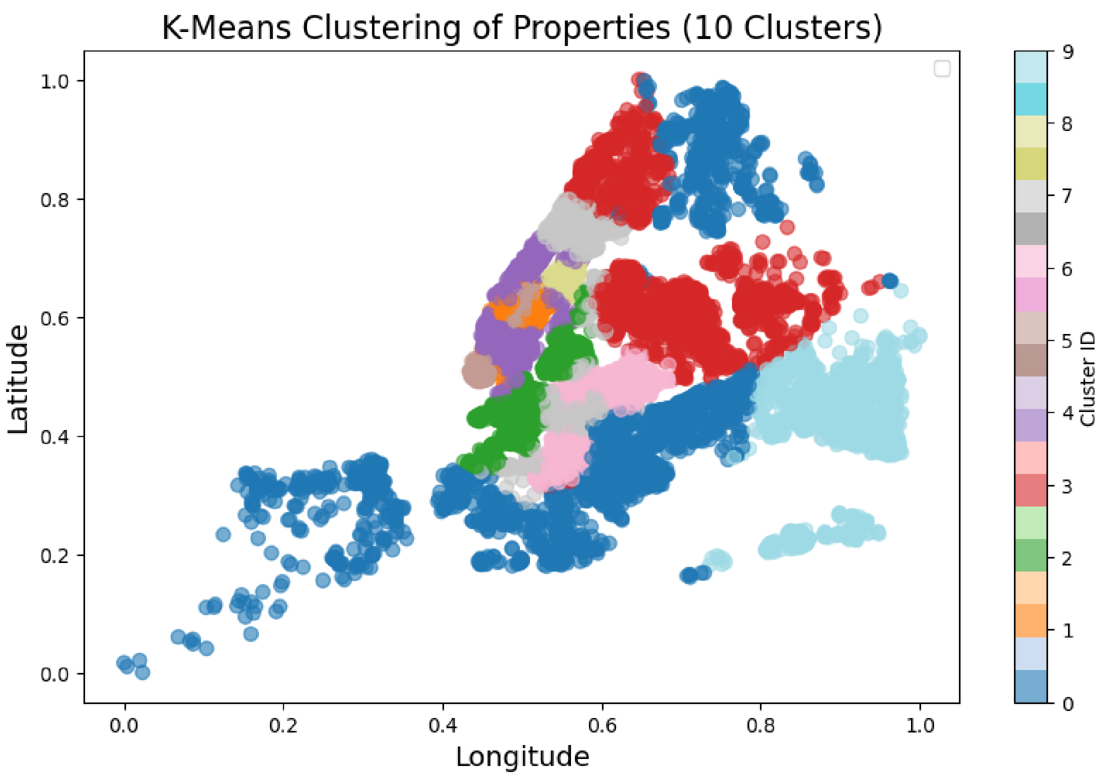
2.2 K-means clustering on averaged neighbor representations

[ ]:
```
# K-means on averaged neighborhood representations, computes k2 clusters and␣
 ↪returns the new combined dataframe with

def k_means(k2, averaged_df):

    # Step 1: Apply KMeans clustering on the averaged features
```

```
    kmeans = KMeans(n_clusters=k2, random_state=42)  # 10 clusters
    averaged_df['cluster'] = kmeans.fit_predict(averaged_df)  # Assign clusters␣
↪to averaged_df

    # Map the clusters back to the original dataframe (df)
    df = combined_df.merge(averaged_df[['cluster']], left_index=True,␣
↪right_index=True)

    # Step 3: Retrieve cluster centroids
    centroids = kmeans.cluster_centers_

    return df, centroids

def plot_cluster(df,k2):
    # Plot the clusters and their centroids
    plt.figure(figsize=(10, 6))

    # Define a colormap with more distinct colors
    cmap = plt.cm.get_cmap('tab20', 20)  # Use the 'tab20' colormap with 20␣
↪distinct colors

    # Plot each property with its assigned cluster (color-coded)
    scatter = plt.scatter(df['longitude'], df['latitude'], c=df['cluster'],␣
↪cmap=cmap, alpha=0.6, s=50)

    # Add a colorbar for reference
    plt.colorbar(scatter, label='Cluster ID')

    # Add labels and title
    plt.xlabel('Longitude', fontsize=14)
    plt.ylabel('Latitude', fontsize=14)
    plt.title(f'K-Means Clustering of Properties ({k2} Clusters)', fontsize=16)
    plt.legend()

    # Display the plot
    plt.show()

# Run an example with k2 = 10
df, centroids = k_means(10, averaged_df)
plot_cluster(df,10)
```

```
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/85324026.py:22:
MatplotlibDeprecationWarning: The get_cmap function was deprecated in Matplotlib
3.7 and will be removed in 3.11. Use ``matplotlib.colormaps[name]`` or
``matplotlib.colormaps.get_cmap()`` or ``pyplot.get_cmap()`` instead.
  cmap = plt.cm.get_cmap('tab20', 20)  # Use the 'tab20' colormap with 20
distinct colors
```

```
/var/folders/2f/7sxq51rj0xd8mgyvzzj2tgy80000gn/T/ipykernel_82843/85324026.py:34:
UserWarning: No artists with labels found to put in legend.  Note that artists
whose label start with an underscore are ignored when legend() is called with no
argument.
  plt.legend()
```



2.3 One-hot Encoding of Cluster Columnn and Dataframe Split

[ ]:
```python
def one_hot_cluster(df):
    # One-hot Encode Cluster Column
    df = pd.get_dummies(df, columns=['cluster'], prefix='cluster',␣
 ↪dummy_na=False)
    return df

# Example
df = one_hot_cluster(df)
df
```

[ ]:
```
        latitude  longitude  host_since  host_response_time  \
0       0.448134   0.579717    0.422237            0.333333
1       0.337268   0.531656    0.939973            0.000000
2       0.288168   0.407848    0.870055            0.000000
```

```
3      0.422369  0.553485   0.249147          0.000000
4      0.600485  0.495148   0.396487          0.000000
…         …        …        …                  …
22418  0.334759  0.558371   0.441337          0.000000
22419  0.786703  0.569332   0.473226          1.000000
22420  0.619740  0.584566   0.249147          0.000000
22421  0.684179  0.496282   0.212995          0.000000
22422  0.475699  0.912347   0.542974          0.000000

       host_response_rate  host_acceptance_rate  host_is_superhost  \
0                    1.00                  1.00                  1
1                    1.00                  0.98                  1
2                    1.00                  1.00                  0
3                    0.99                  0.23                  0
4                    0.93                  0.95                  0
…                     …                     …                 …
22418                1.00                  1.00                  0
22419                0.00                  0.33                  0
22420                0.99                  0.23                  0
22421                1.00                  1.00                  1
22422                1.00                  1.00                  0

       host_listings_count  host_total_listings_count  host_has_profile_pic  \
0                 0.000223                   0.000111                     1
1                 0.000000                   0.000000                     1
2                 0.011351                   0.005988                     1
3                 0.161585                   0.148037                     1
4                 0.157133                   0.271901                     1
…                    …                         …                       …
22418             0.000223                   0.000111                     1
22419             0.001335                   0.000776                     1
22420             0.161585                   0.148037                     1
22421             0.000000                   0.000222                     1
22422             0.000445                   0.000222                     1

       host_identity_verified  calculated_host_listings_count  accommodates  \
0                           1                        0.000000      0.200000
1                           1                        0.000000      0.066667
2                           1                        0.058286      0.066667
3                           1                        0.820571      0.000000
4                           1                        0.082286      0.066667
…                       …                               …             …
22418                       1                        0.001143      0.066667
22419                       0                        0.006857      0.000000
22420                       1                        0.820571      0.000000
22421                       1                        0.000000      0.066667
22422                       1                        0.002286      0.066667
```

```
       bathrooms  bedrooms    beds  has_availability  availability_30  \
0       0.173913  0.222222  0.1250                 1         0.400000
1       0.086957  0.111111  0.1250                 1         0.333333
2       0.086957  0.000000  0.0625                 1         0.566667
3       0.130435  0.444444  0.0625                 1         0.000000
4       0.086957  0.111111  0.0625                 1         0.133333
…            …         …       …                 …                …
22418   0.086957  0.222222  0.3750                 1         1.000000
22419   0.086957  0.111111  0.0625                 1         0.000000
22420   0.260870  0.555556  0.0625                 1         0.100000
22421   0.086957  0.111111  0.0625                 1         0.600000
22422   0.086957  0.111111  0.0625                 1         1.000000

       availability_60  availability_90  availability_365  instant_bookable  \
0             0.700000         0.777778          0.191781               0.0
1             0.333333         0.544444          0.887671               0.0
2             0.733333         0.777778          0.400000               1.0
3             0.000000         0.000000          0.304110               0.0
4             0.216667         0.244444          0.660274               1.0
…                    …                …                 …                 …
22418         1.000000         1.000000          1.000000               0.0
22419         0.016667         0.011111          0.265753               0.0
22420         0.550000         0.700000          0.435616               0.0
22421         0.583333         0.388889          0.095890               0.0
22422         1.000000         1.000000          0.493151               0.0

       minimum_nights  maximum_nights  number_of_reviews  \
0            0.058116        0.112411           0.017517
1            0.000000        0.002800           0.015456
2            0.000000        0.002800           0.002576
3            0.058116        0.036404           0.000000
4            0.000000        0.112411           0.000000
…                   …               …                  …
22418        0.058116        0.112411           0.167439
22419        0.058116        0.036404           0.005667
22420        0.058116        0.036404           0.000000
22421        0.058116        0.002900           0.017002
22422        0.058116        0.008901           0.009789

       number_of_reviews_ltm  number_of_reviews_l30d  first_review  \
0                   0.002822                0.006803      0.640459
1                   0.016930                0.040816      0.944523
2                   0.002822                0.013605      0.973852
3                   0.000000                0.000000      1.000000
4                   0.000000                0.000000      1.000000
…                          …                       …             …
```

```
22418              0.005079               0.000000     0.413428
22419              0.001129               0.000000     0.598057
22420              0.000000               0.000000     1.000000
22421              0.003386               0.000000     0.807951
22422              0.001129               0.000000     0.836042

       last_review  review_scores_rating  review_scores_accuracy  \
0         0.978017              1.000000                1.000000
1         0.982879              0.957500                0.967500
2         0.979497              0.900000                0.950000
3         1.000000              0.929848                0.935703
4         1.000000              0.929848                0.935703
…            …                    …                       …
22418     0.921792              0.927500                0.937500
22419     0.963221              0.840000                0.887500
22420     1.000000              0.927862                0.937101
22421     0.969774              0.970000                0.985000
22422     0.950116              0.935000                0.920000

       review_scores_cleanliness  review_scores_checkin  \
0                       0.992500               1.000000
1                       0.982500               0.950000
2                       0.800000               0.950000
3                       0.919910               0.956577
4                       0.919910               0.956577
…                          …                      …
22418                   0.910000               0.970000
22419                   0.887500               0.977500
22420                   0.918374               0.953303
22421                   0.932500               0.985000
22422                   0.947500               0.907500

       review_scores_communication  review_scores_location  \
0                         1.000000                0.927500
1                         0.975000                0.975000
2                         0.950000                0.950000
3                         0.952058                0.930461
4                         0.952058                0.930461
…                            …                      …
22418                     0.975000                0.912500
22419                     0.887500                0.977500
22420                     0.950123                0.927581
22421                     1.000000                0.962500
22422                     0.920000                0.935000

       review_scores_value  reviews_per_month  room_Entire home/apt  \
0                 0.985000           0.004633                     1
```

```
1               0.907500        0.034517                  0
2               0.800000        0.019348                  1
3               0.902376        0.011225                  0
4               0.902376        0.011225                  0
…                  …              …                 …
22418           0.927500        0.027250                  0
22419           0.862500        0.001272                  0
22420           0.901199        0.011119                  0
22421           0.940000        0.008811                  1
22422           0.920000        0.005995                  0

       room_Hotel room  room_Private room  room_Shared room  Air conditioning  \
0                    0                  0                 0                 0
1                    0                  1                 0                 1
2                    0                  0                 0                 1
3                    0                  1                 0                 0
4                    1                  0                 0                 1
…                …                …                …                …
22418                0                  1                 0                 0
22419                0                  1                 0                 1
22420                0                  1                 0                 1
22421                0                  0                 0                 1
22422                0                  1                 0                 0

       Kitchen  Dedicated workspace  Heating  Hot water  Refrigerator  \
0            1                    0        1          1             1
1            1                    0        0          1             0
2            0                    1        1          1             0
3            1                    1        1          1             1
4            0                    1        1          1             0
…       …                  …      …        …           …
22418        1                    1        1          1             1
22419        0                    1        1          0             0
22420        1                    1        1          1             1
22421        1                    1        0          1             1
22422        0                    1        0          1             1

       Free street parking  Self check-in  Shampoo  Washer  cluster_0  \
0                        1              1        1       0      False
1                        1              1        1       0      False
2                        0              1        1       0       True
3                        0              0        0       1      False
4                        0              1        1       0      False
…                   …              …      …      …        …
22418                    1              1        1       0      False
22419                    0              1        0       0      False
22420                    0              0        0       0      False
```

```
22421                    0              0         1       0      False
22422                    1              1         1       0      False

       cluster_1  cluster_2  cluster_3  cluster_4  cluster_5  cluster_6  \
0          False      False      False      False      False      False
1          False      False      False      False      False       True
2          False      False      False      False      False      False
3          False      False      False      False      False      False
4          False      False      False      False       True      False
…          …         …         …         …         …         …
22418      False      False      False      False      False       True
22419      False      False      False      False      False      False
22420      False      False      False      False      False      False
22421      False      False      False       True      False      False
22422      False      False      False      False      False      False

       cluster_7  cluster_8  cluster_9
0           True      False      False
1          False      False      False
2          False      False      False
3           True      False      False
4          False      False      False
…          …         …         …
22418      False      False      False
22419       True      False      False
22420       True      False      False
22421      False      False      False
22422      False      False       True

[22423 rows x 61 columns]
```

[ ]:
```python
def train_test_split(df, train_size):

    # Add 'price' and 'id' columns back after manipulation
    df['price'] = price_column
    df['id'] = id_column

    # Split the combined_df back into train_df and test_df
    train_df = df.iloc[:train_size].reset_index(drop=True)
    test_df = df.iloc[train_size:].reset_index(drop=True)

    train_df = train_df.drop(columns='id').reset_index(drop=True)
    test_df = test_df.drop(columns='price').reset_index(drop=True)

    return train_df, test_df

# Example
```

```
train_df, test_df = train_test_split(df, train_size)

train_df
```

[ ]:

```
       latitude  longitude  host_since  host_response_time  \
0      0.448134   0.579717    0.422237            0.333333
1      0.337268   0.531656    0.939973            0.000000
2      0.288168   0.407848    0.870055            0.000000
3      0.422369   0.553485    0.249147            0.000000
4      0.600485   0.495148    0.396487            0.000000
…           …          …          …                   …
15691  0.497321   0.456068    0.812415            0.000000
15692  0.927255   0.634970    0.619372            0.000000
15693  0.644912   0.511651    0.595839            0.000000
15694  0.572396   0.457341    0.519952            0.000000
15695  0.310940   0.320027    0.742838            0.000000

       host_response_rate  host_acceptance_rate  host_is_superhost  \
0                    1.00                  1.00                  1
1                    1.00                  0.98                  1
2                    1.00                  1.00                  0
3                    0.99                  0.23                  0
4                    0.93                  0.95                  0
…                     …                     …                   …
15691                0.99                  0.99                  1
15692                1.00                  0.67                  0
15693                1.00                  0.98                  1
15694                1.00                  0.96                  0
15695                1.00                  0.83                  1

       host_listings_count  host_total_listings_count  host_has_profile_pic  \
0                 0.000223                   0.000111                     1
1                 0.000000                   0.000000                     1
2                 0.011351                   0.005988                     1
3                 0.161585                   0.148037                     1
4                 0.157133                   0.271901                     1
…                      …                          …                       …
15691             0.003116                   0.001552                     1
15692             0.000223                   0.000111                     1
15693             0.006009                   0.003770                     1
15694             1.000000                   0.530384                     1
15695             0.001113                   0.000887                     1

       host_identity_verified  calculated_host_listings_count  accommodates  \
0                           1                        0.000000      0.200000
1                           1                        0.000000      0.066667
```

```
2                            1                         0.058286      0.066667
3                            1                         0.820571      0.000000
4                            1                         0.082286      0.066667
…                        …                               …             …
15691                        1                         0.016000      0.200000
15692                        1                         0.001143      0.066667
15693                        1                         0.017143      0.133333
15694                        1                         1.000000      0.066667
15695                        1                         0.005714      0.000000

       bathrooms  bedrooms    beds  has_availability  availability_30  \
0       0.173913  0.222222  0.1250                 1         0.400000
1       0.086957  0.111111  0.1250                 1         0.333333
2       0.086957  0.000000  0.0625                 1         0.566667
3       0.130435  0.444444  0.0625                 1         0.000000
4       0.086957  0.111111  0.0625                 1         0.133333
…            …         …       …                …                …
15691   0.086957  0.111111  0.2500                 1         0.200000
15692   0.086957  0.111111  0.0625                 1         0.000000
15693   0.086957  0.111111  0.0625                 1         0.800000
15694   0.086957  0.111111  0.0625                 1         0.000000
15695   0.086957  0.111111  0.0625                 1         0.966667

       availability_60  availability_90  availability_365  instant_bookable  \
0             0.700000         0.777778          0.191781               0.0
1             0.333333         0.544444          0.887671               0.0
2             0.733333         0.777778          0.400000               1.0
3             0.000000         0.000000          0.304110               0.0
4             0.216667         0.244444          0.660274               1.0
…                  …                …                …                …
15691         0.200000         0.244444          0.380822               1.0
15692         0.000000         0.000000          0.591781               0.0
15693         0.450000         0.633333          0.906849               1.0
15694         0.000000         0.000000          0.676712               0.0
15695         0.983333         0.988889          0.490411               0.0

       minimum_nights  maximum_nights  number_of_reviews  \
0            0.058116        0.112411           0.017517
1            0.000000        0.002800           0.015456
2            0.000000        0.002800           0.002576
3            0.058116        0.036404           0.000000
4            0.000000        0.112411           0.000000
…                 …               …                 …
15691        0.000000        0.036404           0.016486
15692        0.058116        0.112411           0.003091
15693        0.000000        0.036404           0.000515
15694        0.060120        0.112411           0.000000
```

```
15695       0.058116       0.015902        0.010819

       number_of_reviews_ltm  number_of_reviews_l30d  first_review  \
0                   0.002822                0.006803      0.640459
1                   0.016930                0.040816      0.944523
2                   0.002822                0.013605      0.973852
3                   0.000000                0.000000      1.000000
4                   0.000000                0.000000      1.000000
…                     …                      …             …
15691               0.007336                0.006803      0.880389
15692               0.001129                0.000000      0.896466
15693               0.000000                0.000000      0.918905
15694               0.000000                0.000000      1.000000
15695               0.000000                0.000000      0.737809

       last_review  review_scores_rating  review_scores_accuracy  \
0         0.978017              1.000000                1.000000
1         0.982879              0.957500                0.967500
2         0.979497              0.900000                0.950000
3         1.000000              0.929848                0.935703
4         1.000000              0.929848                0.935703
…            …                    …                      …
15691     0.978863              0.985000                0.985000
15692     0.921792              0.832500                0.832500
15693     0.902980              1.000000                1.000000
15694     1.000000              0.929848                0.935703
15695     0.896428              0.952500                0.975000

       review_scores_cleanliness  review_scores_checkin  \
0                        0.99250               1.000000
1                        0.98250               0.950000
2                        0.80000               0.950000
3                        0.91991               0.956577
4                        0.91991               0.956577
…                          …                     …
15691                    0.97750               0.992500
15692                    0.79250               0.792500
15693                    1.00000               1.000000
15694                    0.91991               0.956577
15695                    0.98750               0.965000

       review_scores_communication  review_scores_location  \
0                         1.000000                0.927500
1                         0.975000                0.975000
2                         0.950000                0.950000
3                         0.952058                0.930461
4                         0.952058                0.930461
```

```
…                           …                    …
15691                0.952500             1.000000
15692                0.832500             0.750000
15693                1.000000             1.000000
15694                0.952058             0.930461
15695                0.965000             0.905000

       review_scores_value  reviews_per_month  room_Entire home/apt  \
0                 0.985000           0.004633                     1
1                 0.907500           0.034517                     0
2                 0.800000           0.019348                     1
3                 0.902376           0.011225                     0
4                 0.902376           0.011225                     0
…                      …                  …                   …
15691             0.930000           0.014443                     1
15692             0.832500           0.003088                     0
15693             1.000000           0.000636                     0
15694             0.902376           0.011225                     1
15695             0.952500           0.003997                     0

       room_Hotel room  room_Private room  room_Shared room  Air conditioning  \
0                    0                  0                 0                 0
1                    0                  1                 0                 1
2                    0                  0                 0                 1
3                    0                  1                 0                 0
4                    1                  0                 0                 1
…                …                …                …                …
15691                0                  0                 0                 1
15692                0                  1                 0                 0
15693                0                  1                 0                 1
15694                0                  0                 0                 1
15695                0                  1                 0                 1

       Kitchen  Dedicated workspace  Heating  Hot water  Refrigerator  \
0            1                    0        1          1             1
1            1                    0        0          1             0
2            0                    1        1          1             0
3            1                    1        1          1             1
4            0                    1        1          1             0
…         …                  …      …        …           …
15691        0                    1        1          1             1
15692        1                    1        0          1             1
15693        1                    0        1          0             0
15694        1                    0        1          1             1
15695        1                    1        1          1             0

       Free street parking  Self check-in  Shampoo  Washer  cluster_0  \
```

```
0                        1              1        1       0      False
1                        1              1        1       0      False
2                        0              1        1       0       True
3                        0              0        0       1      False
4                        0              1        1       0      False
…                    …            …      …     …        …
15691                    0              1        1       0      False
15692                    1              1        1       0      False
15693                    0              1        1       0      False
15694                    0              1        1       0      False
15695                    0              1        0       0       True

       cluster_1  cluster_2  cluster_3  cluster_4  cluster_5  cluster_6  \
0          False      False      False      False      False      False
1          False      False      False      False      False       True
2          False      False      False      False      False      False
3          False      False      False      False      False      False
4          False      False      False      False       True      False
…          …        …        …        …        …        …
15691      False      False      False      False       True      False
15692      False      False       True      False      False      False
15693      False      False      False      False       True      False
15694      False      False      False       True      False      False
15695      False      False      False      False      False      False

       cluster_7  cluster_8  cluster_9  price
0           True      False      False    4.0
1          False      False      False    3.0
2          False      False      False    3.0
3           True      False      False    0.0
4          False      False      False    2.0
…          …        …        …     …
15691      False      False      False    5.0
15692      False      False      False    0.0
15693      False      False      False    5.0
15694      False      False      False    5.0
15695      False      False      False    1.0

[15696 rows x 62 columns]
```

[ ]:
```
test_df
```

[ ]:
```
      latitude  longitude  host_since  host_response_time  host_response_rate  \
0     0.594257   0.645392    0.295020            0.233668             0.91509
1     0.615634   0.588774    0.249147            0.000000             0.99000
2     0.429960   0.572398    0.233970            0.233668             0.91509
3     0.718678   0.520824    0.787858            0.333333             0.70000
```

4	0.518763	0.550865	0.383356	0.666667	1.00000
…	…	…	…	…	…
6722	0.334759	0.558371	0.441337	0.000000	1.00000
6723	0.786703	0.569332	0.473226	1.000000	0.00000
6724	0.619740	0.584566	0.249147	0.000000	0.99000
6725	0.684179	0.496282	0.212995	0.000000	1.00000
6726	0.475699	0.912347	0.542974	0.000000	1.00000

	host_acceptance_rate	host_is_superhost	host_listings_count	\
0	0.785925	0	0.000000	
1	0.230000	0	0.161585	
2	0.785925	0	0.000000	
3	0.370000	0	0.007790	
4	0.750000	0	0.000000	
…	…	…	…	
6722	1.000000	0	0.000223	
6723	0.330000	0	0.001335	
6724	0.230000	0	0.161585	
6725	1.000000	1	0.000000	
6726	1.000000	0	0.000445	

	host_total_listings_count	host_has_profile_pic	host_identity_verified	\
0	0.001220	1	1	
1	0.148037	1	1	
2	0.000000	1	1	
3	0.008649	1	1	
4	0.000000	1	1	
…	…	…	…	
6722	0.000111	1	1	
6723	0.000776	1	0	
6724	0.148037	1	1	
6725	0.000222	1	1	
6726	0.000222	1	1	

	calculated_host_listings_count	accommodates	bathrooms	bedrooms	\
0	0.000000	0.333333	0.130435	0.333333	
1	0.820571	0.000000	0.260870	0.444444	
2	0.000000	0.066667	0.086957	0.111111	
3	0.040000	0.000000	0.260870	0.000000	
4	0.000000	0.066667	0.086957	0.000000	
…	…	…	…	…	
6722	0.001143	0.066667	0.086957	0.222222	
6723	0.006857	0.000000	0.086957	0.111111	
6724	0.820571	0.000000	0.260870	0.555556	
6725	0.000000	0.066667	0.086957	0.111111	
6726	0.002286	0.066667	0.086957	0.111111	

	beds	has_availability	availability_30	availability_60	\
0	0.2500	1	0.966667	0.983333	
1	0.0625	1	0.966667	0.983333	
2	0.0625	1	0.966667	0.983333	
3	0.0625	1	0.000000	0.000000	
4	0.0625	1	0.100000	0.200000	
…	…	…	…	…	
6722	0.3750	1	1.000000	1.000000	
6723	0.0625	1	0.000000	0.016667	
6724	0.0625	1	0.100000	0.550000	
6725	0.0625	1	0.600000	0.583333	
6726	0.0625	1	1.000000	1.000000	

	availability_90	availability_365	instant_bookable	minimum_nights	\
0	0.988889	0.243836	0.0	0.058116	
1	0.988889	0.997260	0.0	0.058116	
2	0.988889	0.243836	0.0	0.058116	
3	0.000000	0.430137	0.0	0.058116	
4	0.266667	0.309589	0.0	0.058116	
…	…	…	…	…	
6722	1.000000	1.000000	0.0	0.058116	
6723	0.011111	0.265753	0.0	0.058116	
6724	0.700000	0.435616	0.0	0.058116	
6725	0.388889	0.095890	0.0	0.058116	
6726	1.000000	0.493151	0.0	0.058116	

	maximum_nights	number_of_reviews	number_of_reviews_ltm	\
0	0.008901	0.014426	0.001129	
1	0.036404	0.000000	0.000000	
2	0.036404	0.015971	0.000000	
3	0.049905	0.002061	0.001693	
4	0.005901	0.080886	0.001129	
…	…	…	…	
6722	0.112411	0.167439	0.005079	
6723	0.036404	0.005667	0.001129	
6724	0.036404	0.000000	0.000000	
6725	0.002900	0.017002	0.003386	
6726	0.008901	0.009789	0.001129	

	number_of_reviews_l30d	first_review	last_review	review_scores_rating	\
0	0.000000	0.864488	0.908687	1.000000	
1	0.000000	1.000000	1.000000	0.927862	
2	0.000000	0.800000	0.840837	0.985000	
3	0.006803	0.904064	0.980342	0.937500	
4	0.006803	0.370848	0.982456	0.955000	
…	…	…	…	…	
6722	0.000000	0.413428	0.921792	0.927500	

```
6723              0.000000      0.598057     0.963221               0.840000
6724              0.000000      1.000000     1.000000               0.927862
6725              0.000000      0.807951     0.969774               0.970000
6726              0.000000      0.836042     0.950116               0.935000

      review_scores_accuracy  review_scores_cleanliness  \
0                   1.000000                   0.990000
1                   0.937101                   0.918374
2                   1.000000                   0.975000
3                   0.875000                   0.937500
4                   0.972500                   0.922500
…                        …                        …
6722                0.937500                   0.910000
6723                0.887500                   0.887500
6724                0.937101                   0.918374
6725                0.985000                   0.932500
6726                0.920000                   0.947500

      review_scores_checkin  review_scores_communication  \
0                  1.000000                     1.000000
1                  0.953303                     0.950123
2                  0.975000                     0.992500
3                  1.000000                     0.875000
4                  0.992500                     0.990000
…                       …                          …
6722               0.970000                     0.975000
6723               0.977500                     0.887500
6724               0.953303                     0.950123
6725               0.985000                     1.000000
6726               0.907500                     0.920000

      review_scores_location  review_scores_value  reviews_per_month  \
0                   0.990000             0.990000           0.010991
1                   0.927581             0.901199           0.011119
2                   0.920000             0.975000           0.007903
3                   1.000000             0.875000           0.002271
4                   0.985000             0.932500           0.012172
…                        …                    …                  …
6722                0.912500             0.927500           0.027250
6723                0.977500             0.862500           0.001272
6724                0.927581             0.901199           0.011119
6725                0.962500             0.940000           0.008811
6726                0.935000             0.920000           0.005995

      room_Entire home/apt  room_Hotel room  room_Private room  \
0                        1                0                  0
1                        0                0                  1
```

```
2                       1                0                  0
3                       1                0                  0
4                       1                0                  0
…                     …              …                …
6722                    0                0                  1
6723                    0                0                  1
6724                    0                0                  1
6725                    1                0                  0
6726                    0                0                  1

      room_Shared room  Air conditioning  Kitchen  Dedicated workspace  \
0                    0                 0        1                    1
1                    0                 1        1                    1
2                    0                 1        1                    0
3                    0                 1        1                    1
4                    0                 0        1                    0
…                  …               …      …                    …
6722                 0                 0        1                    1
6723                 0                 1        0                    1
6724                 0                 1        1                    1
6725                 0                 1        1                    1
6726                 0                 0        0                    1

      Heating  Hot water  Refrigerator  Free street parking  Self check-in  \
0           0          1             1                    1              1
1           1          1             1                    0              0
2           1          0             0                    0              0
3           1          1             0                    0              0
4           1          1             1                    1              0
…         …        …           …                  …            …
6722        1          1             1                    1              1
6723        1          0             0                    0              1
6724        1          1             1                    0              0
6725        0          1             1                    0              0
6726        0          1             1                    1              1

      Shampoo  Washer  cluster_0  cluster_1  cluster_2  cluster_3  cluster_4  \
0           1       0      False      False      False       True      False
1           0       0      False      False      False      False      False
2           0       0      False      False      False      False      False
3           0       1      False      False      False      False       True
4           1       0      False      False       True      False      False
…         …     …        …        …        …        …        …
6722        1       0      False      False      False      False      False
6723        0       0      False      False      False      False      False
6724        0       0      False      False      False      False      False
6725        1       0      False      False      False      False       True
```

```
6726         1        0      False      False      False      False      False

      cluster_5  cluster_6  cluster_7  cluster_8  cluster_9       id
0         False      False      False      False      False   3917.0
1         False      False       True      False      False   1885.0
2         False      False       True      False      False   1305.0
3         False      False      False      False      False  19328.0
4         False      False      False      False      False  16511.0
…           …         …         …         …          …        …
6722      False       True      False      False      False   7205.0
6723      False      False       True      False      False   3954.0
6724      False      False       True      False      False   1358.0
6725      False      False      False      False      False   2793.0
6726      False      False      False      False       True    865.0

[6727 rows x 62 columns]
```

3 Supervized Learning

Here, I implement homeade grid search cross validation to tune the hyperparameters of the unsupervised neighborhood analysis and the supervised model simultaneously

3.1 XG Boost

[ ]:
```
knn_neighbors = [100, 400, 1000]
kmeans_clusters = [1, 5, 10, 20]

# XGBoost hyperparameters to tune
learning_rates = [0.01, 0.1, 0.3]
max_depths = [8, 10, 15]
n_estimators = [100, 500, 1000] # Keeping it at 100 for now actually much␣
↪higher is better

n_folds = 5
# Generate a new random state dynamically
kf = KFold(n_splits=n_folds, shuffle=True, random_state=42)


# Placeholder for storing best results
best_params = None
best_score = float('inf')  # MSE should be minimized
```

[ ]:
```
def grid_search_cv(combined_df):
    global best_params, best_score
    k = 0
    # Iterate over all combinations of XGBoost hyperparameters
    for k1 in knn_neighbors:
```

```
        # Run KNN
        averaged_df = knn(combined_df, k1)

        for k2 in kmeans_clusters:

            # Run K-means on averaged neighborhood data
            df, centroids = k_means(k2, averaged_df)

            # One-hot on cluster
            df = one_hot_cluster(df)

            # Train-test split
            train_df, test_df = train_test_split(df, train_size)

            y = train_df['price']
            X = train_df.drop(columns='price')

            for lr in learning_rates:
                for depth in max_depths:
                    for ne in n_estimators:

                        # Set XGBoost regression hyperparameters
                        params = {
                            'objective': 'reg:squarederror',  # Regression␣
↪objective
                            'eta': lr,
                            'max_depth': depth,
                            'n_estimators': ne,
                            'verbosity': 0
                        }

                        # Store RMSE for each fold
                        fold_rmse = []

                        # 5-fold Cross Validation
                        for train_index, val_index in kf.split(X):
                            X_train, X_val = X.iloc[train_index], X.
↪iloc[val_index]
                            y_train, y_val = y.iloc[train_index], y.
↪iloc[val_index]

                            # Train XGBoost regression model
                            model = xgb.XGBRegressor(**params)
                            model.fit(X_train, y_train)

                            # Predict continuous values on validation set
```

```
                        y_pred_continuous = model.predict(X_val)

                        # Round predictions and clip to valid class range␣
↪(e.g., 0-5)
                        y_pred = np.round(y_pred_continuous).clip(0, 5)

                        # Calculate Root Mean Squared Error (RMSE) for this␣
↪fold
                        rmse = np.sqrt(mean_squared_error(y_val, y_pred))
                        fold_rmse.append(rmse)

                    # Calculate average RMSE for this set of hyperparameters
                    avg_rmse = np.mean(fold_rmse)
                    print(f"Iteration {k}: Avg RMSE = {avg_rmse}")
                    k += 1

                    # Update best parameters based on the lowest average␣
↪RMSE
                    if avg_rmse < best_score:
                        params['k1'] = k1
                        params['k2'] = k2
                        best_score = avg_rmse
                        best_params = params

                        # Print the current best hyperparameters and score
                        print(f"Current Best Parameters: {best_params}")
                        print(f"Current Best RMSE: {best_score}")

    # Print the best hyperparameters and score
    print(f"FINAL Best Parameters: {best_params}")
    print(f"FINAL Best RMSE: {best_score}")

# Run the grid search cross-validation
grid_search_cv(combined_df)
```

```
Iteration 0: Avg RMSE = 1.0379314813461833
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.01,
'max_depth': 8, 'n_estimators': 100, 'verbosity': 0, 'k1': 100, 'k2': 1}
Current Best RMSE: 1.0379314813461833
Iteration 1: Avg RMSE = 0.8070776227147178
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.01,
'max_depth': 8, 'n_estimators': 500, 'verbosity': 0, 'k1': 100, 'k2': 1}
Current Best RMSE: 0.8070776227147178
Iteration 2: Avg RMSE = 0.7936924033782632
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.01,
'max_depth': 8, 'n_estimators': 1000, 'verbosity': 0, 'k1': 100, 'k2': 1}
Current Best RMSE: 0.7936924033782632
```

```
Iteration 3: Avg RMSE = 1.0237346244441023
Iteration 4: Avg RMSE = 0.8082156132321602
Iteration 5: Avg RMSE = 0.7963513903423973
Iteration 6: Avg RMSE = 1.0370983663449125
Iteration 7: Avg RMSE = 0.85000871303642
Iteration 8: Avg RMSE = 0.8490787507929085
Iteration 9: Avg RMSE = 0.7964216327400575
Iteration 10: Avg RMSE = 0.7900231392740771
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.1,
'max_depth': 8, 'n_estimators': 500, 'verbosity': 0, 'k1': 100, 'k2': 1}
Current Best RMSE: 0.7900231392740771
Iteration 11: Avg RMSE = 0.7887229348527974
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.1,
'max_depth': 8, 'n_estimators': 1000, 'verbosity': 0, 'k1': 100, 'k2': 1}
Current Best RMSE: 0.7887229348527974
Iteration 12: Avg RMSE = 0.8017318441715311
Iteration 13: Avg RMSE = 0.798047673558579
Iteration 14: Avg RMSE = 0.7977758157324443
Iteration 15: Avg RMSE = 0.8464219806856452
Iteration 16: Avg RMSE = 0.8456618530327515
Iteration 17: Avg RMSE = 0.8456618530327515
Iteration 18: Avg RMSE = 0.8233618762701029
Iteration 19: Avg RMSE = 0.8256773145911982
Iteration 20: Avg RMSE = 0.8253669082846937
Iteration 21: Avg RMSE = 0.8319279782356895
Iteration 22: Avg RMSE = 0.8319658863695235
Iteration 23: Avg RMSE = 0.8319658863695235
Iteration 24: Avg RMSE = 0.8635473636428254
Iteration 25: Avg RMSE = 0.8635473636428254
Iteration 26: Avg RMSE = 0.8635473636428254
Iteration 27: Avg RMSE = 1.0355931276496342
Iteration 28: Avg RMSE = 0.804968304009777
Iteration 29: Avg RMSE = 0.7896527867668771
Iteration 30: Avg RMSE = 1.0216922891345324
Iteration 31: Avg RMSE = 0.8038504974288593
Iteration 32: Avg RMSE = 0.7934083072257474
Iteration 33: Avg RMSE = 1.033291960606055
Iteration 34: Avg RMSE = 0.8461894106387955
Iteration 35: Avg RMSE = 0.8407332725099138
Iteration 36: Avg RMSE = 0.8004523663968618
Iteration 37: Avg RMSE = 0.7926208694550663
Iteration 38: Avg RMSE = 0.791229569846747
Iteration 39: Avg RMSE = 0.7982000415022544
Iteration 40: Avg RMSE = 0.794206958270487
Iteration 41: Avg RMSE = 0.7935263513313136
Iteration 42: Avg RMSE = 0.8469455494933362
Iteration 43: Avg RMSE = 0.8465304938054443
Iteration 44: Avg RMSE = 0.8465304938054443
```

```
Iteration 45: Avg RMSE = 0.8278141101588867
Iteration 46: Avg RMSE = 0.826407740426738
Iteration 47: Avg RMSE = 0.8264494084912783
Iteration 48: Avg RMSE = 0.8301787293892365
Iteration 49: Avg RMSE = 0.8304134032941277
Iteration 50: Avg RMSE = 0.8304134032941277
Iteration 51: Avg RMSE = 0.8682736324518947
Iteration 52: Avg RMSE = 0.8682736324518947
Iteration 53: Avg RMSE = 0.8682736324518947
Iteration 54: Avg RMSE = 1.036900817802065
Iteration 55: Avg RMSE = 0.8051670425709704
Iteration 56: Avg RMSE = 0.791623741487332
Iteration 57: Avg RMSE = 1.0241698794931342
Iteration 58: Avg RMSE = 0.807326285323866
Iteration 59: Avg RMSE = 0.7965655745207951
Iteration 60: Avg RMSE = 1.0335890629582674
Iteration 61: Avg RMSE = 0.8503697849917831
Iteration 62: Avg RMSE = 0.8453408656245053
Iteration 63: Avg RMSE = 0.7962368760191794
Iteration 64: Avg RMSE = 0.7898635431883503
Iteration 65: Avg RMSE = 0.7896705840719604
Iteration 66: Avg RMSE = 0.8013506366618215
Iteration 67: Avg RMSE = 0.8006444667731666
Iteration 68: Avg RMSE = 0.8018364666337495
Iteration 69: Avg RMSE = 0.84131646678618
Iteration 70: Avg RMSE = 0.8406920514876856
Iteration 71: Avg RMSE = 0.8406920514876856
Iteration 72: Avg RMSE = 0.8261011208056708
Iteration 73: Avg RMSE = 0.8290461828885147
Iteration 74: Avg RMSE = 0.8288937926034571
Iteration 75: Avg RMSE = 0.8278671830913134
Iteration 76: Avg RMSE = 0.8273240369148065
Iteration 77: Avg RMSE = 0.8273240369148065
Iteration 78: Avg RMSE = 0.863855954091463
Iteration 79: Avg RMSE = 0.863855954091463
Iteration 80: Avg RMSE = 0.863855954091463
Iteration 81: Avg RMSE = 1.0360697980207194
Iteration 82: Avg RMSE = 0.8054652304180479
Iteration 83: Avg RMSE = 0.7882670098904634
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.01,
'max_depth': 8, 'n_estimators': 1000, 'verbosity': 0, 'k1': 100, 'k2': 20}
Current Best RMSE: 0.7882670098904634
Iteration 84: Avg RMSE = 1.0233074216726514
Iteration 85: Avg RMSE = 0.8051369151170821
Iteration 86: Avg RMSE = 0.795634365768535
Iteration 87: Avg RMSE = 1.040932418297579
Iteration 88: Avg RMSE = 0.8461380676169737
Iteration 89: Avg RMSE = 0.8421531593409146
```

```
Iteration 90: Avg RMSE = 0.7965441776664596
Iteration 91: Avg RMSE = 0.790217524669009
Iteration 92: Avg RMSE = 0.7891397474386073
Iteration 93: Avg RMSE = 0.7984395079025206
Iteration 94: Avg RMSE = 0.7967190683124244
Iteration 95: Avg RMSE = 0.796445340879253
Iteration 96: Avg RMSE = 0.8424365046805417
Iteration 97: Avg RMSE = 0.8422479217272132
Iteration 98: Avg RMSE = 0.8422479217272132
Iteration 99: Avg RMSE = 0.8261971394788711
Iteration 100: Avg RMSE = 0.8294753398193171
Iteration 101: Avg RMSE = 0.8298577931825205
Iteration 102: Avg RMSE = 0.8308327194559693
Iteration 103: Avg RMSE = 0.8317565805399696
Iteration 104: Avg RMSE = 0.8317565805399696
Iteration 105: Avg RMSE = 0.8694966143753644
Iteration 106: Avg RMSE = 0.8694966143753644
Iteration 107: Avg RMSE = 0.8694966143753644
Iteration 108: Avg RMSE = 1.0379314813461833
Iteration 109: Avg RMSE = 0.8070776227147178
Iteration 110: Avg RMSE = 0.7936924033782632
Iteration 111: Avg RMSE = 1.0237346244441023
Iteration 112: Avg RMSE = 0.8082156132321602
Iteration 113: Avg RMSE = 0.7963513903423973
Iteration 114: Avg RMSE = 1.0370983663449125
Iteration 115: Avg RMSE = 0.85000871303642
Iteration 116: Avg RMSE = 0.8490787507929085
Iteration 117: Avg RMSE = 0.7964216327400575
Iteration 118: Avg RMSE = 0.7900231392740771
Iteration 119: Avg RMSE = 0.7887229348527974
Iteration 120: Avg RMSE = 0.8017318441715311
Iteration 121: Avg RMSE = 0.798047673558579
Iteration 122: Avg RMSE = 0.7977758157324443
Iteration 123: Avg RMSE = 0.8464219806856452
Iteration 124: Avg RMSE = 0.8456618530327515
Iteration 125: Avg RMSE = 0.8456618530327515
Iteration 126: Avg RMSE = 0.8233618762701029
Iteration 127: Avg RMSE = 0.8256773145911982
Iteration 128: Avg RMSE = 0.8253669082846937
Iteration 129: Avg RMSE = 0.8319279782356895
Iteration 130: Avg RMSE = 0.8319658863695235
Iteration 131: Avg RMSE = 0.8319658863695235
Iteration 132: Avg RMSE = 0.8635473636428254
Iteration 133: Avg RMSE = 0.8635473636428254
Iteration 134: Avg RMSE = 0.8635473636428254
Iteration 135: Avg RMSE = 1.0343948781856789
Iteration 136: Avg RMSE = 0.8056996741608711
Iteration 137: Avg RMSE = 0.7928814271404455
```

```
Iteration 138: Avg RMSE = 1.021070650178281
Iteration 139: Avg RMSE = 0.8008760740007522
Iteration 140: Avg RMSE = 0.7942087299133169
Iteration 141: Avg RMSE = 1.0360239745004012
Iteration 142: Avg RMSE = 0.8507932444591997
Iteration 143: Avg RMSE = 0.846969067145406
Iteration 144: Avg RMSE = 0.7928465681170322
Iteration 145: Avg RMSE = 0.7874339995188515
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.1,
'max_depth': 8, 'n_estimators': 500, 'verbosity': 0, 'k1': 400, 'k2': 5}
Current Best RMSE: 0.7874339995188515
Iteration 146: Avg RMSE = 0.7873586937715451
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.1,
'max_depth': 8, 'n_estimators': 1000, 'verbosity': 0, 'k1': 400, 'k2': 5}
Current Best RMSE: 0.7873586937715451
Iteration 147: Avg RMSE = 0.7999685983862259
Iteration 148: Avg RMSE = 0.7985325477604464
Iteration 149: Avg RMSE = 0.7985335834254023
Iteration 150: Avg RMSE = 0.8513599424158503
Iteration 151: Avg RMSE = 0.8513941053632823
Iteration 152: Avg RMSE = 0.8513941053632823
Iteration 153: Avg RMSE = 0.821084600960123
Iteration 154: Avg RMSE = 0.8244795250419112
Iteration 155: Avg RMSE = 0.8247845571120005
Iteration 156: Avg RMSE = 0.8336216570483307
Iteration 157: Avg RMSE = 0.8346546713216618
Iteration 158: Avg RMSE = 0.8346546713216618
Iteration 159: Avg RMSE = 0.8699543388069492
Iteration 160: Avg RMSE = 0.8699543388069492
Iteration 161: Avg RMSE = 0.8699543388069492
Iteration 162: Avg RMSE = 1.0341181632528753
Iteration 163: Avg RMSE = 0.8075102207035488
Iteration 164: Avg RMSE = 0.7913613194131884
Iteration 165: Avg RMSE = 1.0207347995416238
Iteration 166: Avg RMSE = 0.8037686093508469
Iteration 167: Avg RMSE = 0.7955276564229278
Iteration 168: Avg RMSE = 1.0362132324090996
Iteration 169: Avg RMSE = 0.8462926065489234
Iteration 170: Avg RMSE = 0.8434324451566824
Iteration 171: Avg RMSE = 0.7997025047122065
Iteration 172: Avg RMSE = 0.789129757911412
Iteration 173: Avg RMSE = 0.7916040617157946
Iteration 174: Avg RMSE = 0.7990848406651763
Iteration 175: Avg RMSE = 0.7970511507745777
Iteration 176: Avg RMSE = 0.7966526982959354
Iteration 177: Avg RMSE = 0.8482882833129075
Iteration 178: Avg RMSE = 0.848738844986036
Iteration 179: Avg RMSE = 0.848738844986036
```

```
Iteration 180: Avg RMSE = 0.8247269945544808
Iteration 181: Avg RMSE = 0.824748988411596
Iteration 182: Avg RMSE = 0.82455369210004
Iteration 183: Avg RMSE = 0.8355069153493215
Iteration 184: Avg RMSE = 0.8346913365983214
Iteration 185: Avg RMSE = 0.8346913365983214
Iteration 186: Avg RMSE = 0.8659863543336013
Iteration 187: Avg RMSE = 0.8659863543336013
Iteration 188: Avg RMSE = 0.8659863543336013
Iteration 189: Avg RMSE = 1.0358719588357208
Iteration 190: Avg RMSE = 0.8075225921513083
Iteration 191: Avg RMSE = 0.792061228599591
Iteration 192: Avg RMSE = 1.021626976276774
Iteration 193: Avg RMSE = 0.8064731991317245
Iteration 194: Avg RMSE = 0.7963799327873764
Iteration 195: Avg RMSE = 1.035911669366333
Iteration 196: Avg RMSE = 0.8462446408677359
Iteration 197: Avg RMSE = 0.8443236155557097
Iteration 198: Avg RMSE = 0.7994965674304495
Iteration 199: Avg RMSE = 0.7881101315299694
Iteration 200: Avg RMSE = 0.7872721819931525
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.1,
'max_depth': 8, 'n_estimators': 1000, 'verbosity': 0, 'k1': 400, 'k2': 20}
Current Best RMSE: 0.7872721819931525
Iteration 201: Avg RMSE = 0.804998186662166
Iteration 202: Avg RMSE = 0.8021423875312397
Iteration 203: Avg RMSE = 0.8024180435979135
Iteration 204: Avg RMSE = 0.8468494193736206
Iteration 205: Avg RMSE = 0.8470357258893225
Iteration 206: Avg RMSE = 0.8470357258893225
Iteration 207: Avg RMSE = 0.8253775965002479
Iteration 208: Avg RMSE = 0.8260589922811696
Iteration 209: Avg RMSE = 0.8265208704154222
Iteration 210: Avg RMSE = 0.8352390100284983
Iteration 211: Avg RMSE = 0.835242558562651
Iteration 212: Avg RMSE = 0.835242558562651
Iteration 213: Avg RMSE = 0.8625168294604719
Iteration 214: Avg RMSE = 0.8625168294604719
Iteration 215: Avg RMSE = 0.8625168294604719
Iteration 216: Avg RMSE = 1.0379314813461833
Iteration 217: Avg RMSE = 0.8070776227147178
Iteration 218: Avg RMSE = 0.7936924033782632
Iteration 219: Avg RMSE = 1.0237346244441023
Iteration 220: Avg RMSE = 0.8082156132321602
Iteration 221: Avg RMSE = 0.7963513903423973
Iteration 222: Avg RMSE = 1.0370983663449125
Iteration 223: Avg RMSE = 0.85000871303642
Iteration 224: Avg RMSE = 0.8490787507929085
```

```
Iteration 225: Avg RMSE = 0.7964216327400575
Iteration 226: Avg RMSE = 0.7900231392740771
Iteration 227: Avg RMSE = 0.7887229348527974
Iteration 228: Avg RMSE = 0.8017318441715311
Iteration 229: Avg RMSE = 0.798047673558579
Iteration 230: Avg RMSE = 0.7977758157324443
Iteration 231: Avg RMSE = 0.8464219806856452
Iteration 232: Avg RMSE = 0.8456618530327515
Iteration 233: Avg RMSE = 0.8456618530327515
Iteration 234: Avg RMSE = 0.8233618762701029
Iteration 235: Avg RMSE = 0.8256773145911982
Iteration 236: Avg RMSE = 0.8253669082846937
Iteration 237: Avg RMSE = 0.8319279782356895
Iteration 238: Avg RMSE = 0.8319658863695235
Iteration 239: Avg RMSE = 0.8319658863695235
Iteration 240: Avg RMSE = 0.8635473636428254
Iteration 241: Avg RMSE = 0.8635473636428254
Iteration 242: Avg RMSE = 0.8635473636428254
Iteration 243: Avg RMSE = 1.0339077613579921
Iteration 244: Avg RMSE = 0.8073893674347733
Iteration 245: Avg RMSE = 0.7908571583533771
Iteration 246: Avg RMSE = 1.022553965620227
Iteration 247: Avg RMSE = 0.8046901036159518
Iteration 248: Avg RMSE = 0.7936862848776916
Iteration 249: Avg RMSE = 1.0357245566153614
Iteration 250: Avg RMSE = 0.841542289323338
Iteration 251: Avg RMSE = 0.8408992183732862
Iteration 252: Avg RMSE = 0.7949666653219241
Iteration 253: Avg RMSE = 0.7882445314622666
Iteration 254: Avg RMSE = 0.7896914746734454
Iteration 255: Avg RMSE = 0.8020381885842408
Iteration 256: Avg RMSE = 0.7966232082641084
Iteration 257: Avg RMSE = 0.7975421568778838
Iteration 258: Avg RMSE = 0.8453205009272265
Iteration 259: Avg RMSE = 0.8447120968206834
Iteration 260: Avg RMSE = 0.8447120968206834
Iteration 261: Avg RMSE = 0.8268016383861164
Iteration 262: Avg RMSE = 0.8234580959053703
Iteration 263: Avg RMSE = 0.8233426528373876
Iteration 264: Avg RMSE = 0.8320557231323906
Iteration 265: Avg RMSE = 0.8322953536680515
Iteration 266: Avg RMSE = 0.8322953536680515
Iteration 267: Avg RMSE = 0.8642287312533604
Iteration 268: Avg RMSE = 0.8642287312533604
Iteration 269: Avg RMSE = 0.8642287312533604
Iteration 270: Avg RMSE = 1.0340814295677656
Iteration 271: Avg RMSE = 0.8053312823662694
Iteration 272: Avg RMSE = 0.7925241255947568
```

```
Iteration 273: Avg RMSE = 1.0221116143241487
Iteration 274: Avg RMSE = 0.8068917336349664
Iteration 275: Avg RMSE = 0.7973847482363821
Iteration 276: Avg RMSE = 1.034604991141387
Iteration 277: Avg RMSE = 0.8438274820983164
Iteration 278: Avg RMSE = 0.842435746939902
Iteration 279: Avg RMSE = 0.7968581321221392
Iteration 280: Avg RMSE = 0.7845074004346797
Current Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.1,
'max_depth': 8, 'n_estimators': 500, 'verbosity': 0, 'k1': 1000, 'k2': 10}
Current Best RMSE: 0.7845074004346797
Iteration 281: Avg RMSE = 0.7870195445763264
Iteration 282: Avg RMSE = 0.8003740295342162
Iteration 283: Avg RMSE = 0.7992525022977841
Iteration 284: Avg RMSE = 0.79913552658912
Iteration 285: Avg RMSE = 0.8456196735823397
Iteration 286: Avg RMSE = 0.8453886996330343
Iteration 287: Avg RMSE = 0.8453886996330343
Iteration 288: Avg RMSE = 0.8232352840362337
Iteration 289: Avg RMSE = 0.8211410612262803
Iteration 290: Avg RMSE = 0.8213003601487019
Iteration 291: Avg RMSE = 0.8344512037487984
Iteration 292: Avg RMSE = 0.8346830079346436
Iteration 293: Avg RMSE = 0.8346830079346436
Iteration 294: Avg RMSE = 0.8606388401899986
Iteration 295: Avg RMSE = 0.8606388401899986
Iteration 296: Avg RMSE = 0.8606388401899986
Iteration 297: Avg RMSE = 1.0359917805903158
Iteration 298: Avg RMSE = 0.8070424627190127
Iteration 299: Avg RMSE = 0.7905437916324617
Iteration 300: Avg RMSE = 1.021923109945403
Iteration 301: Avg RMSE = 0.8028998845670131
Iteration 302: Avg RMSE = 0.7928648657139703
Iteration 303: Avg RMSE = 1.03574458406011
Iteration 304: Avg RMSE = 0.8420687718560103
Iteration 305: Avg RMSE = 0.8380909513024528
Iteration 306: Avg RMSE = 0.7942546858789983
Iteration 307: Avg RMSE = 0.7868859513774003
Iteration 308: Avg RMSE = 0.785867891674233
Iteration 309: Avg RMSE = 0.7985359653623743
Iteration 310: Avg RMSE = 0.7972403548440071
Iteration 311: Avg RMSE = 0.7977646413197067
Iteration 312: Avg RMSE = 0.8441356119127603
Iteration 313: Avg RMSE = 0.8433059443037602
Iteration 314: Avg RMSE = 0.8433059443037602
Iteration 315: Avg RMSE = 0.8185464626815252
Iteration 316: Avg RMSE = 0.818074289539273
Iteration 317: Avg RMSE = 0.8185041326052058
```

```
Iteration 318: Avg RMSE = 0.8306806611040548
Iteration 319: Avg RMSE = 0.8300218268819182
Iteration 320: Avg RMSE = 0.8300218268819182
Iteration 321: Avg RMSE = 0.8594529180281933
Iteration 322: Avg RMSE = 0.8594529180281933
Iteration 323: Avg RMSE = 0.8594529180281933
FINAL Best Parameters: {'objective': 'reg:squarederror', 'eta': 0.1,
'max_depth': 8, 'n_estimators': 500, 'verbosity': 0, 'k1': 1000, 'k2': 10}
FINAL Best RMSE: 0.7845074004346797
```

3.2 Train on the Entire Dataset with the Best Hyperparameters

[ ]:
```
# Run KNN
averaged_df = knn(combined_df, best_params['k1'])

# Run K-means on averaged neighborhood data
df, centroids = k_means(best_params['k2'], averaged_df)

# One-hot on cluster
df = one_hot_cluster(df)

# Train-test split
train_df, test_df = train_test_split(df, train_size)

y = train_df['price']
X = train_df.drop(columns='price')

params = {
        'objective': 'reg:squarederror',  # regression
        'eta': best_params['eta'],
        'max_depth': best_params['max_depth'],
        'n_estimators': best_params['n_estimators'],
        'verbosity': 0
    }

# Train XGBoost model
model = xgb.XGBRegressor(**params)
model.fit(X, y)
```

[ ]:
```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, device=None, early_stopping_rounds=None,
             enable_categorical=False, eta=0.1, eval_metric=None,
             feature_types=None, gamma=None, grow_policy=None,
             importance_type=None, interaction_constraints=None,
             learning_rate=None, max_bin=None, max_cat_threshold=None,
             max_cat_to_onehot=None, max_delta_step=None, max_depth=8,
```

```
              max_leaves=None, min_child_weight=None, missing=nan,
              monotone_constraints=None, multi_strategy=None, n_estimators=500,
              n_jobs=None, num_parallel_tree=None, …)
```

[ ]:
```
# Get feature importance scores
feature_importances = model.get_booster().get_score(importance_type='gain')  #␣
↪'weight', 'gain', 'cover', etc.

# Convert to a sorted list of tuples for better readability
sorted_importances = sorted(feature_importances.items(), key=lambda x: x[1],␣
↪reverse=True)

print("Feature Importances (sorted by gain):")
for feature, importance in sorted_importances:
    print(f"{feature}: {importance}")
```

```
Feature Importances (sorted by gain):
room_Private room: 325.2705993652344
room_Entire home/apt: 64.36674499511719
minimum_nights: 47.259891510009766
calculated_host_listings_count: 9.63469409942627
accommodates: 7.497044086456299
room_Shared room: 6.880578517913818
bathrooms: 6.248887062072754
host_total_listings_count: 6.155507564544678
bedrooms: 6.110628604888916
beds: 4.61558198928833
cluster_4: 3.787069320678711
host_listings_count: 3.0366976261138916
availability_90: 2.767470598220825
longitude: 2.659891128540039
availability_60: 2.3329100608825684
number_of_reviews_l30d: 2.3114709854125977
cluster_7: 2.1329452991485596
availability_30: 1.7847083806991577
Washer: 1.5363706350326538
Shampoo: 1.492868185043335
Kitchen: 1.470493197441101
host_response_rate: 1.4136143922805786
cluster_5: 1.29920494556427
host_has_profile_pic: 1.2659605741500854
host_response_time: 1.2141882181167603
Self check-in: 1.1893988847732544
review_scores_rating: 1.1297128200531006
cluster_6: 1.127785563468933
host_is_superhost: 1.0882351398468018
host_acceptance_rate: 1.081351399421692
number_of_reviews_ltm: 1.0798776149749756
```

```
latitude: 1.071574330329895
cluster_0: 1.0615819692611694
review_scores_cleanliness: 1.0444649457931519
Heating: 0.9837900996208191
review_scores_location: 0.9168003797531128
Dedicated workspace: 0.8862706422805786
last_review: 0.8702533841133118
instant_bookable: 0.8698781728744507
first_review: 0.8621521592140198
room_Hotel room: 0.8530785441398621
has_availability: 0.8246044516563416
host_since: 0.8052897453308105
cluster_2: 0.7942496538162231
cluster_9: 0.7869352698326111
reviews_per_month: 0.768190324306488
host_identity_verified: 0.7632337808609009
number_of_reviews: 0.7528233528137207
cluster_1: 0.7403284311294556
review_scores_value: 0.6961389183998108
Free street parking: 0.6745784282684326
maximum_nights: 0.6696850657463074
review_scores_accuracy: 0.6617966294288635
review_scores_communication: 0.6433504819869995
cluster_8: 0.6397855281829834
Refrigerator: 0.6191384196281433
availability_365: 0.6159202456474304
Hot water: 0.6122521758079529
cluster_3: 0.6109626293182373
Air conditioning: 0.5454719662666321
review_scores_checkin: 0.5442160964012146
```

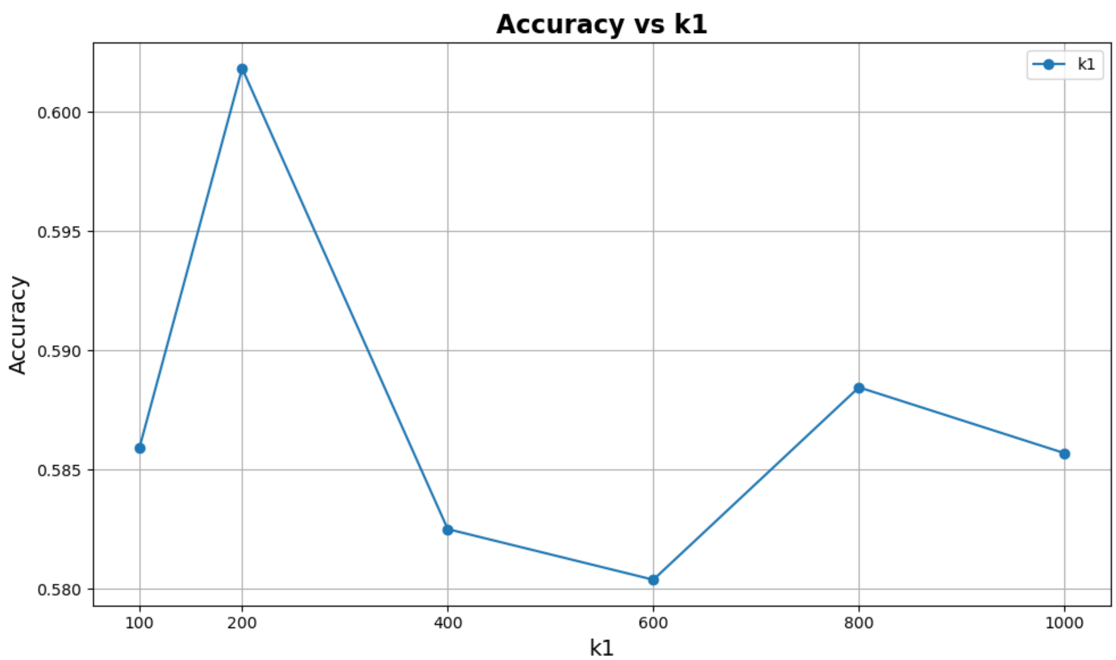
3.3 Hyperparameter Plots

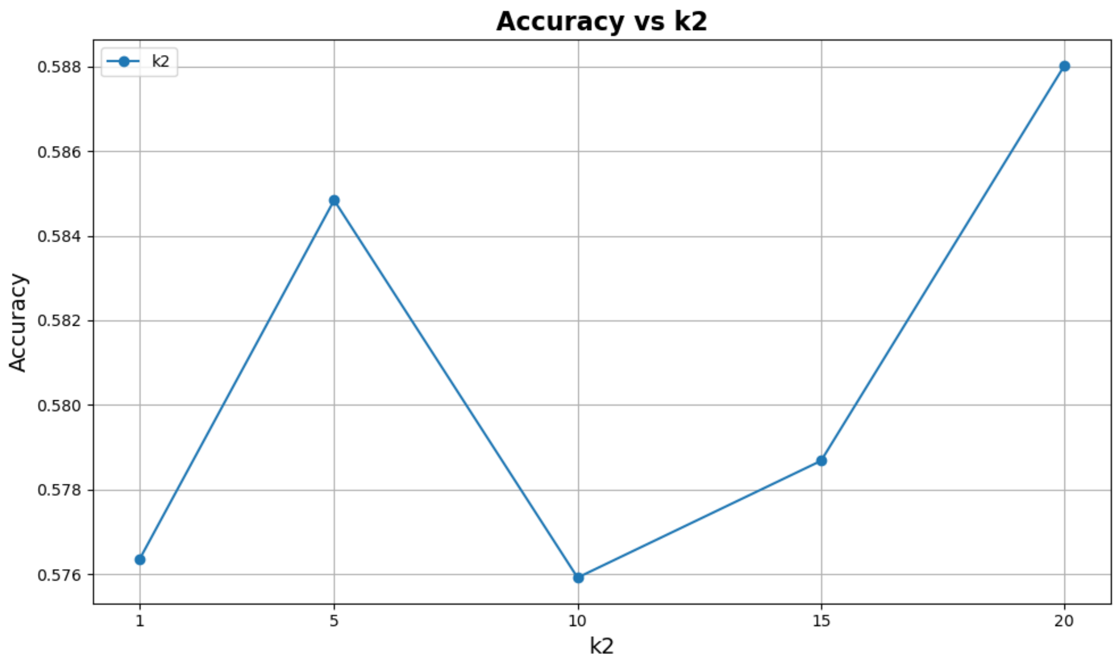
Vary one hyperparamter at a time keeping the others at the best value, split the training dataframe to compute predictve accuracy on a labled test set

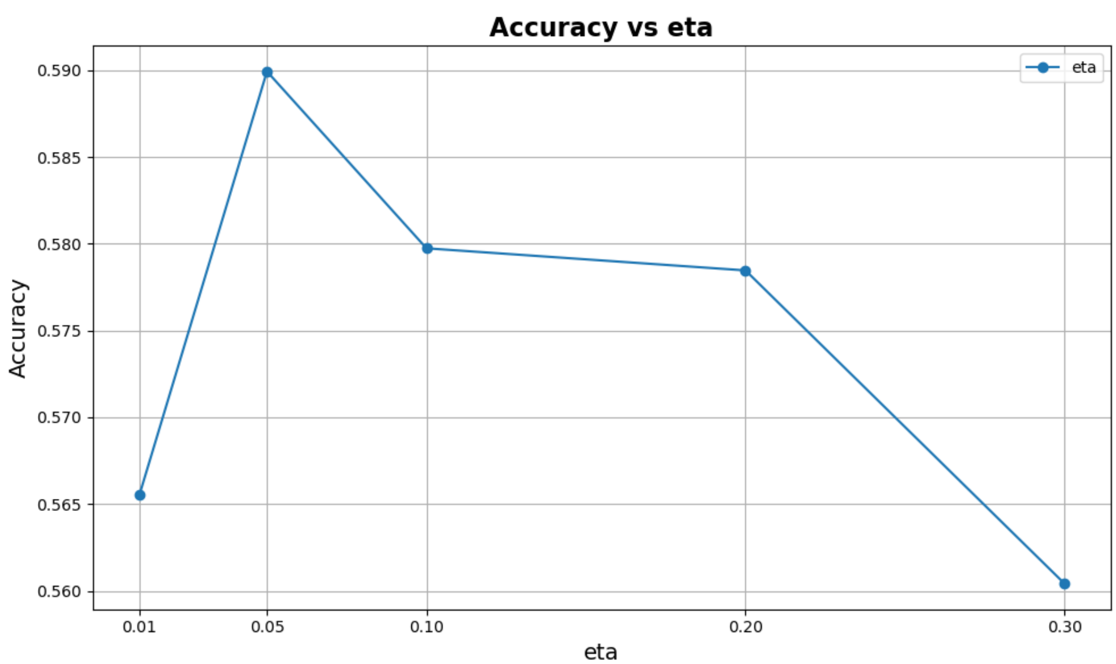
[ ]:
```
import sklearn.model_selection as skm
from sklearn.metrics import accuracy_score

hyperparameter_grids = {
    'k1': [100, 200, 400, 600, 800, 1000],
    'k2': [1, 5, 10, 15, 20],
    'eta': [0.01, 0.05, 0.1, 0.2, 0.3],
    'max_depth': [6, 8, 10, 12, 15],
    'n_estimators': [100, 200, 500, 800, 1000],
}

# Base hyperparameters
```

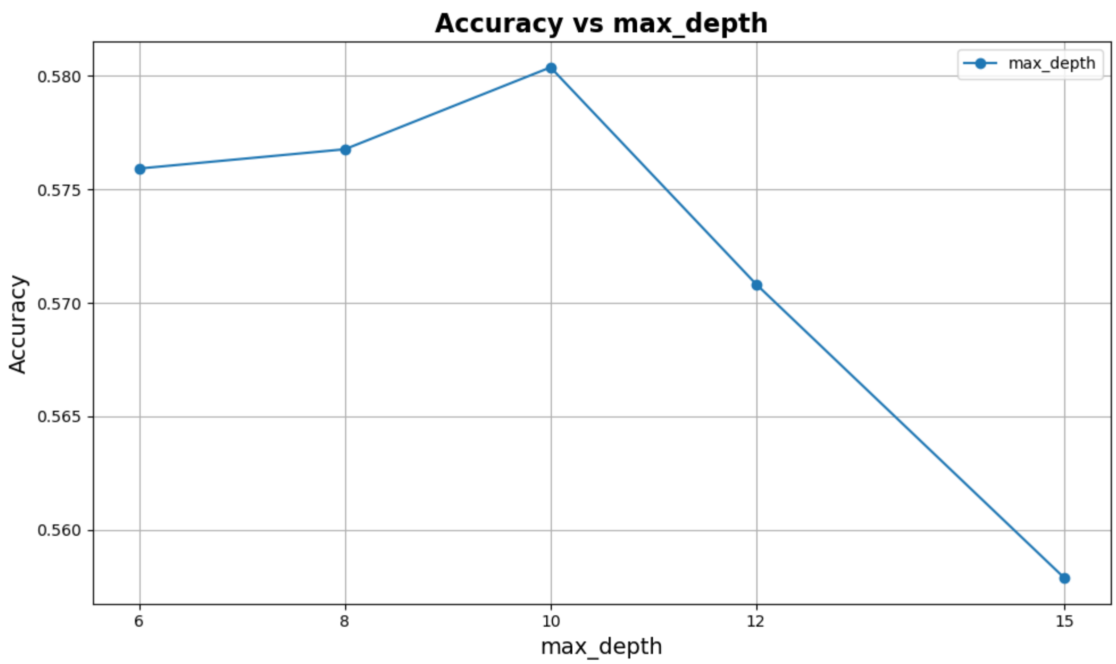
```python
best_params = {
    'objective': 'reg:squarederror',
    'eta': 0.1,
    'max_depth': 8,
    'n_estimators': 500,
    'verbosity': 0,
    'k1': 1000,
    'k2': 10,
}

def evaluate_model(params, train_df, test_df):
    # Split train_df into features and target
    y_train = train_df['price']
    X_train = train_df.drop(columns='price')

    # Train the model
    model = xgb.XGBRegressor(**params)
    model.fit(X_train, y_train)

    # Prepare test data
    y_test = test_df['price']
    X_test = test_df.drop(columns='price')

    # Make predictions on the test dataset
    y_pred = model.predict(X_test)

    # Round predictions to the nearest integer in the range 0-5
    y_pred_rounded = np.clip(np.round(y_pred), 0, 5).astype(int)

    # Calculate accuracy (percentage of correct predictions)
    accuracy = accuracy_score(y_test, y_pred_rounded)
    return accuracy

# Store results
results = {}

for param_name, values in hyperparameter_grids.items():
    accuracies = []

    for value in values:
        temp_params = best_params.copy()
        temp_params[param_name] = value

        # Run KNN
        averaged_df = knn(combined_df, temp_params['k1'])

        # Run K-means
```

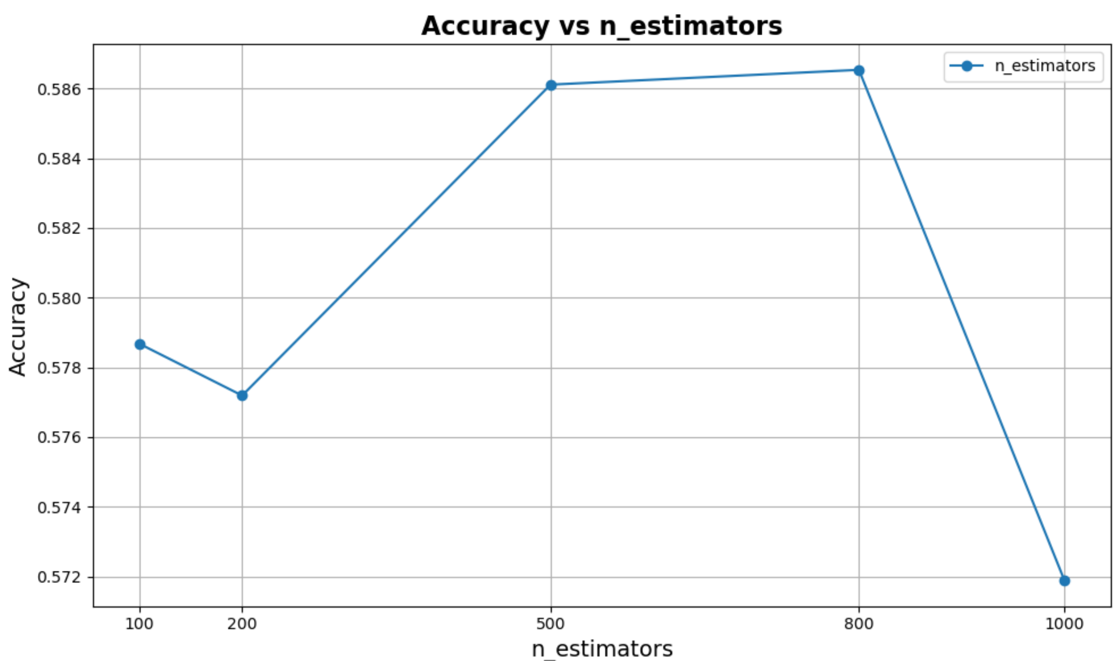
```python
        df, centroids = k_means(temp_params['k2'], averaged_df)
        df = one_hot_cluster(df)

        # Train-test split after K-means (using your custom function)
        train_df, test_df = train_test_split(df, train_size)

        # Split train_df again into final training and validation sets (
        final_train_df, final_test_df = skm.train_test_split(train_df,␣
↪train_size=0.7)  # 70% train, 30% validation

        # Evaluate XGBoost model on final_train_df
        accuracy = evaluate_model(temp_params, final_train_df, final_test_df)
        accuracies.append(accuracy)

    results[param_name] = {'values': values, 'accuracies': accuracies}

    # Plot results
    plt.figure(figsize=(10, 6))
    plt.plot(values, accuracies, marker='o', label=f'{param_name}')
    plt.title(f'Accuracy vs {param_name}', fontsize=16, fontweight='bold')
    plt.xlabel(param_name, fontsize=14)
    plt.ylabel('Accuracy', fontsize=14)
    plt.grid(True)
    plt.xticks(values)
    plt.legend()
    plt.tight_layout()
    plt.show()
```




Accuracy vs k2
k2
0.588
0.586
0.584
0.582
0.580
0.578
0.576
Accuracy
1
5
10
15
20
k2

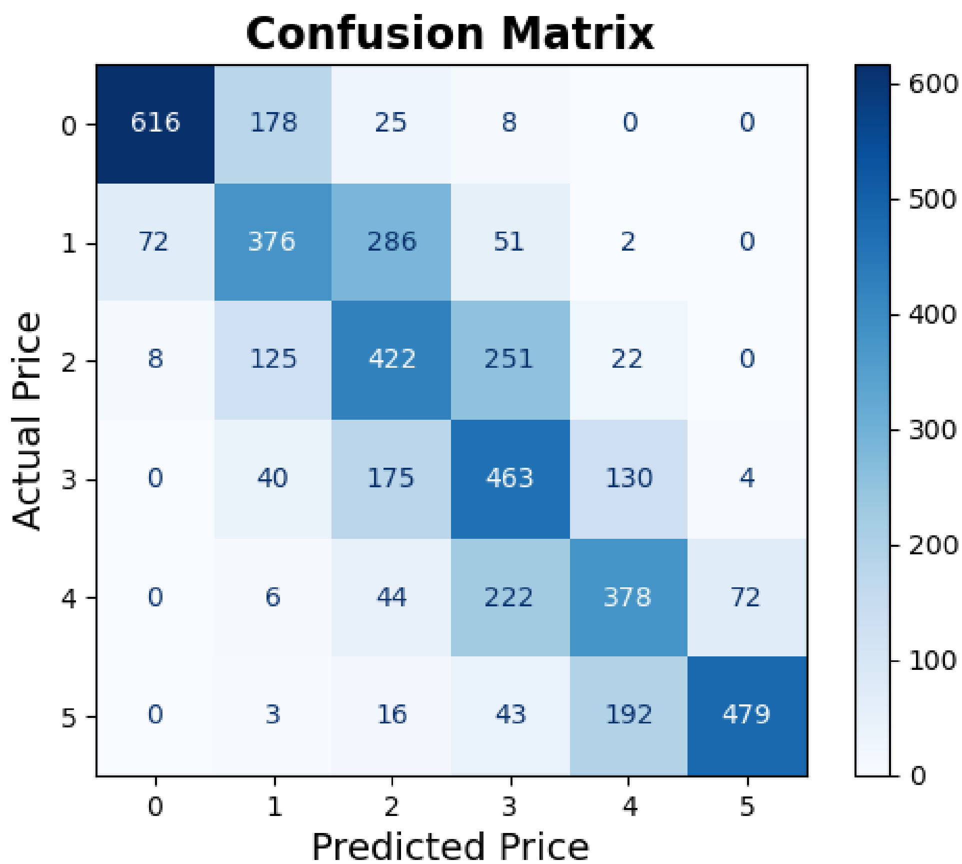

Accuracy vs eta
eta
0.590
0.585
0.580
0.575
0.570
0.565
0.560
Accuracy
0.01
0.05
0.10
0.20
0.30
eta


Accuracy vs max_depth
max_depth
Accuracy
0.580
0.575
0.570
0.565
0.560
6
8
10
12
15
max_depth


Accuracy vs n_estimators
n_estimators
Accuracy
0.586
0.584
0.582
0.580
0.578
0.576
0.574
0.572
100
200
500
800
1000
n_estimators

3.4 Crate Confusion Matrix

[ ]:
```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay,␣
 ↪accuracy_score

# kNN
averaged_df = knn(combined_df, best_params['k1'])

# K-means
df, centroids = k_means(best_params['k2'], averaged_df)
df = one_hot_cluster(df)

# Train-test split after K-means (using your custom function)
train_df, test_df = train_test_split(df, train_size)

# Split train_df again into final training and validation sets (
final_train_df, final_test_df = skm.train_test_split(train_df, train_size=0.7) ␣
 ↪# 70% train, 30% validation

# Extract Features and Targets
y_train = final_train_df['price']
X_train = final_train_df.drop(columns='price')

y_test = final_test_df['price']
X_test = final_test_df.drop(columns='price')

# Train XGBoost Model
model = xgb.XGBRegressor(**best_params)
model.fit(X_train, y_train)

# Make Predictions
y_pred = model.predict(X_test)

# Round predictions to integers in the range [0, 5]
y_pred_rounded = np.clip(np.round(y_pred), 0, 5).astype(int)

# Create Confusion Matrix
cm = confusion_matrix(y_test, y_pred_rounded, labels=[0, 1, 2, 3, 4, 5])

# Display the Confusion Matrix
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=[0, 1, 2, 3,␣
 ↪4, 5])
disp.plot(cmap='Blues', values_format='d')
plt.xlabel("Predicted Price", fontsize=14)
plt.ylabel("Actual Price", fontsize=14)
plt.title("Confusion Matrix", fontsize=16, fontweight='bold')
plt.show()
```



3.5 Predict on the Test Dataset

[ ]:
```
id_col = test_df['id']
predict_df = test_df.drop(columns='id')

pred_cont = model.predict(predict_df)
y_pred = np.round(pred_cont).clip(0, 5)

predicted_prices = y_pred.astype(int)
ids = id_col.astype(int)

# Create a DataFrame with 'id' and 'price' columns
result_df = pd.DataFrame({
    'id': ids,
    'price': predicted_prices
})

# Save the DataFrame to a CSV file
result_df.to_csv('predictions2.csv', index=False)
```