COMP 4211

Start

AIRBNB IMPROVEMENT RECOMMENDATION SYSTEM

Group Project COMP 4211: Machine Learning

YouTube Link: https://www.youtube.com/watch?v=nMKmVBmuUDg

Submitted by: CHAN YEE KI (20860858) | LOKESWARA LOVERA (20798275)

PROJECT BACKGROUND

Because of the competitive market in the Airbnb industry, we are making a model to predict the listing's score based on its features.
Ultimately, we are giving recommendations for Airbnb owners on how to increase their properties review scores.

SIZE OF DATASET 1

- 28 columns and 6173 entries
- 26 features and 2 targets
- 16 numerical features and 10 categorical features

Features	'Account_life', 'host_response_time', 'host_is_superhost', 'host_listings_count', 'host_identity_verified', 'neighbourhood_cleansed', 'property_type', 'room_type', 'accommodates', 'bathrooms_total', 'Bathroom_type', 'bedrooms', 'beds', 'amenities', 'price', 'minimum_nights', 'maximum_nights', 'has_availability', 'number_of_reviews', 'Latest_review', 'instant_bookable', 'calculated_host_listings_count', 'calculated_host_listings_count_entire_homes', 'calculated_host_listings_count_private_rooms', 'calculated_host_listings_count_shared_rooms', 'reviews_per_month'
Targets	‘review_scores_rating’, ‘classification_target’

MISSING VALUES IN DATASET 1

Column Name	Number of Null	Portion of Null
host_response_time	1481	0.23991576219018307
bathrooms_total	15	0.0024299368216426373
Bathroom_type	4748	0.7691560019439495
bedrooms	314	0.05086667746638587
beds	91	0.014741616717965332
Latest_review	590	0.0955775149846104
reviews_per_month	590	0.0955775149846104
review_scores_rating	590	0.0955775149846104
classification_target	590	0.0955775149846104

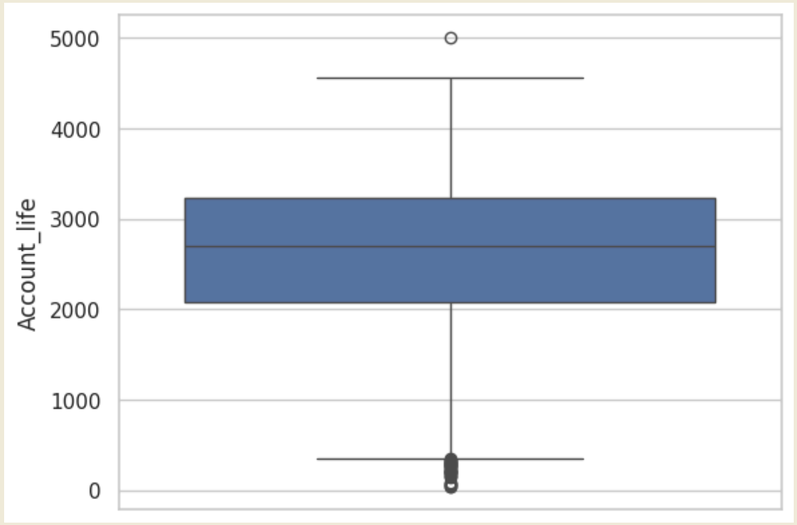
NUMERICAL FEATURE DISTRIBUTION

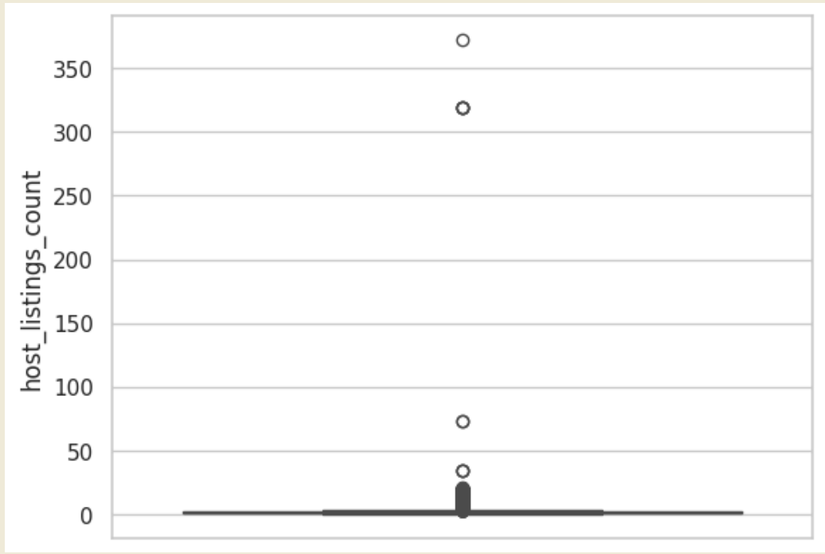
Continuous	'bathrooms_total', 'bedrooms', 'beds', 'price','latest_review ', 'reviews_per_month '
Discrete	' Account_life', ' host_listings_count', 'accommodates', 'minimum_nights', 'maximum_nights', 'number_of_reviews', 'calculated_host_listings_count', 'calculated_host_listings_count_entire_homes', 'calculated_host_listings_count_private_rooms', 'calculated_host_listings_count_shared_rooms'
Total	16 numerical features

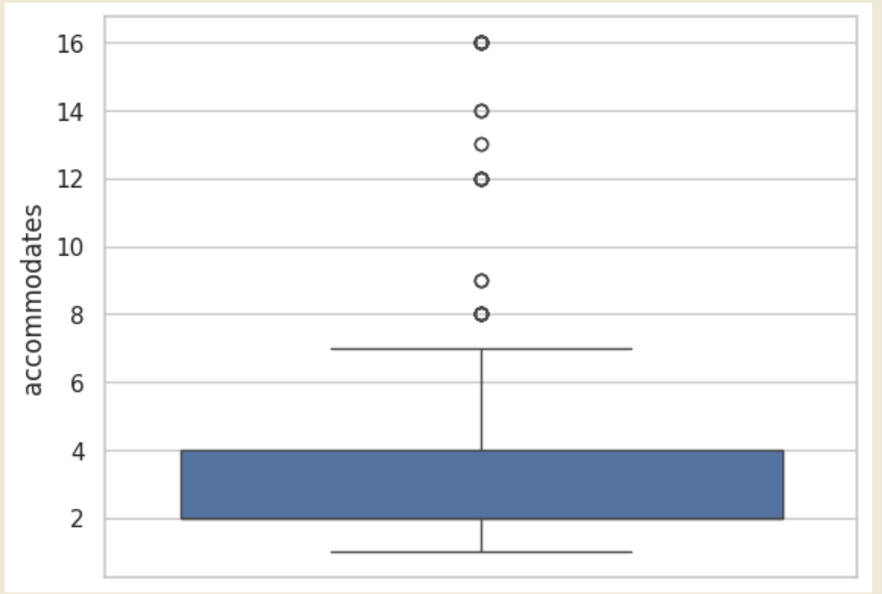
NUMERICAL FEATURE DISTRIBUTION

The distribution of the first 3 numerical features:

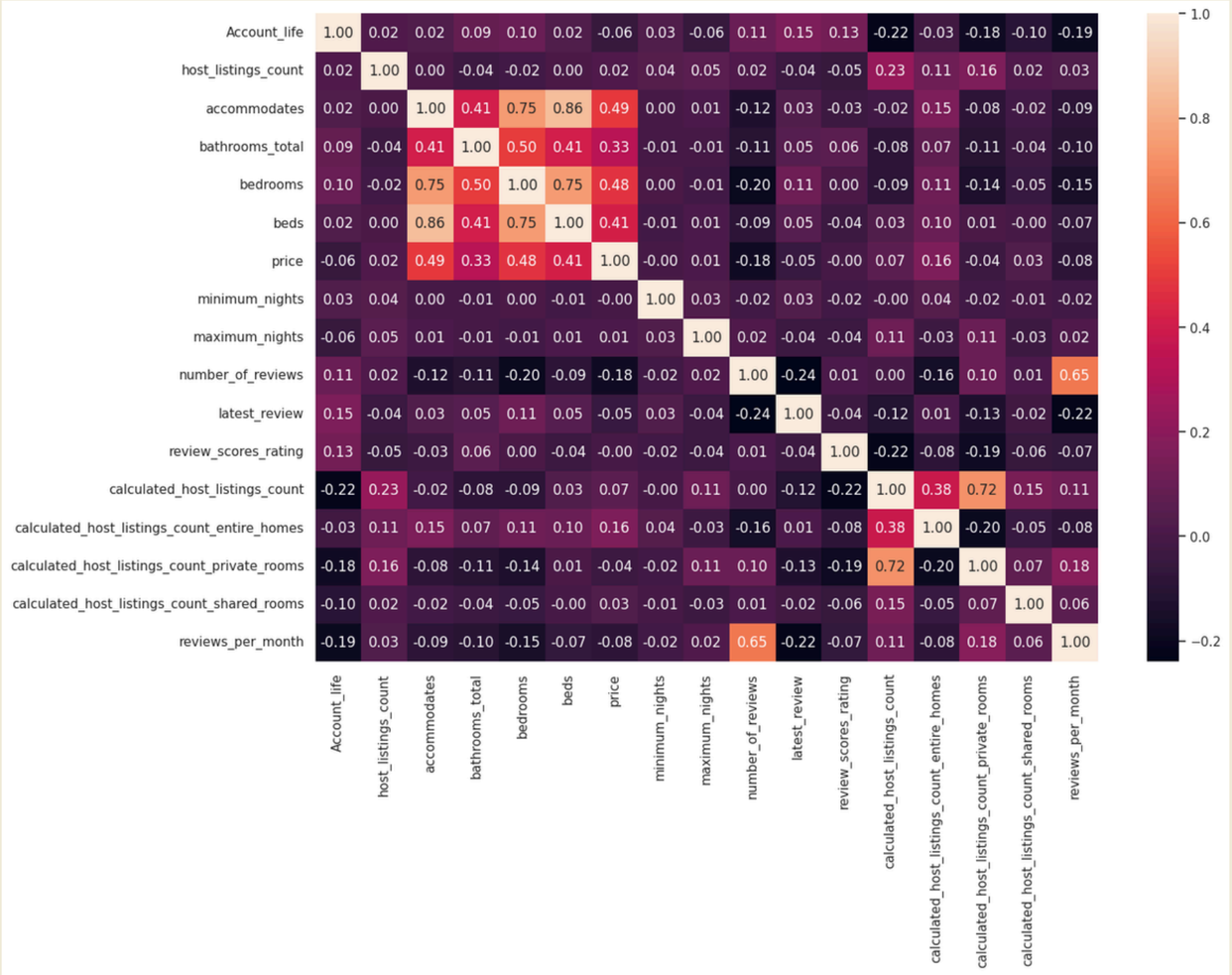
	Account_life	host_listings_count	accommodates
count	3657	3657	3657
mean	2559.948045	2.441072	2.980859
std	884.920354	12.632899	1.547096
min	35	0	1
25%	2078	1	2
50%	2695	1	2
75%	3228	2	4
max	5002	372	16


Account_life
5000
4000
3000
2000
1000
0


host_listings_count
350
300
250
200
150
100
50
0


accommodates
16
14
12
10
8
6
4
2

CORRELATION



CATEGORICAL FEATURE DISTRIBUTION

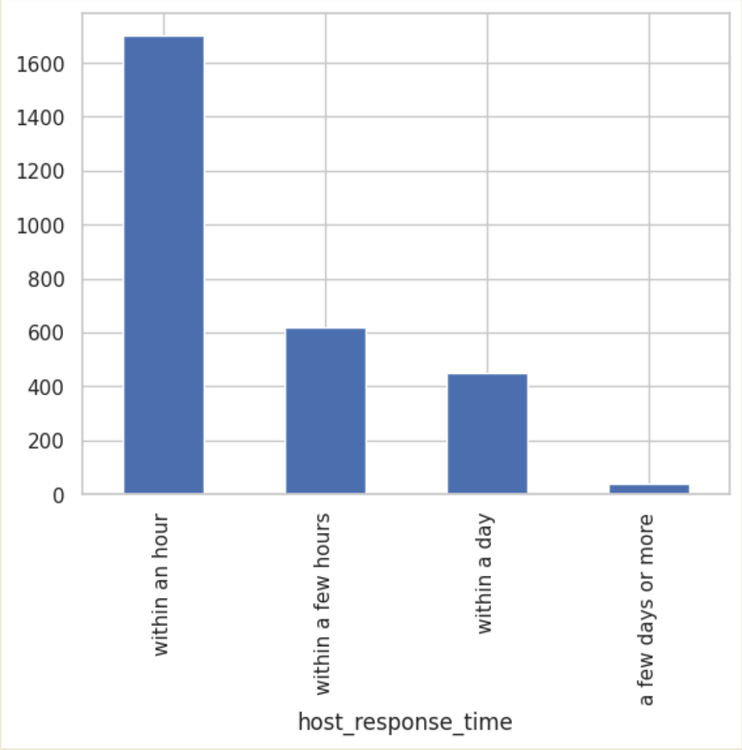
Binary	'host_is_superhost', 'host_identity_verified', 'has_availability', 'instant_bookable'
Nominal	'neighbourhood_cleansed', 'property_type', 'room_type', 'bathroom_type', 'amenities'
Ordinal	'host_response_time'
Total	10 categorical features

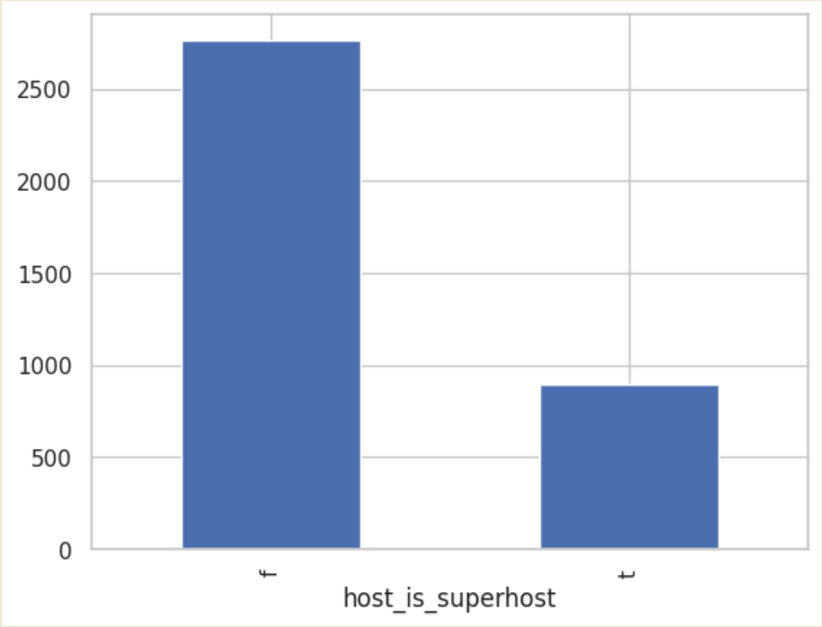
CATEGORICAL FEATURE DISTRIBUTION

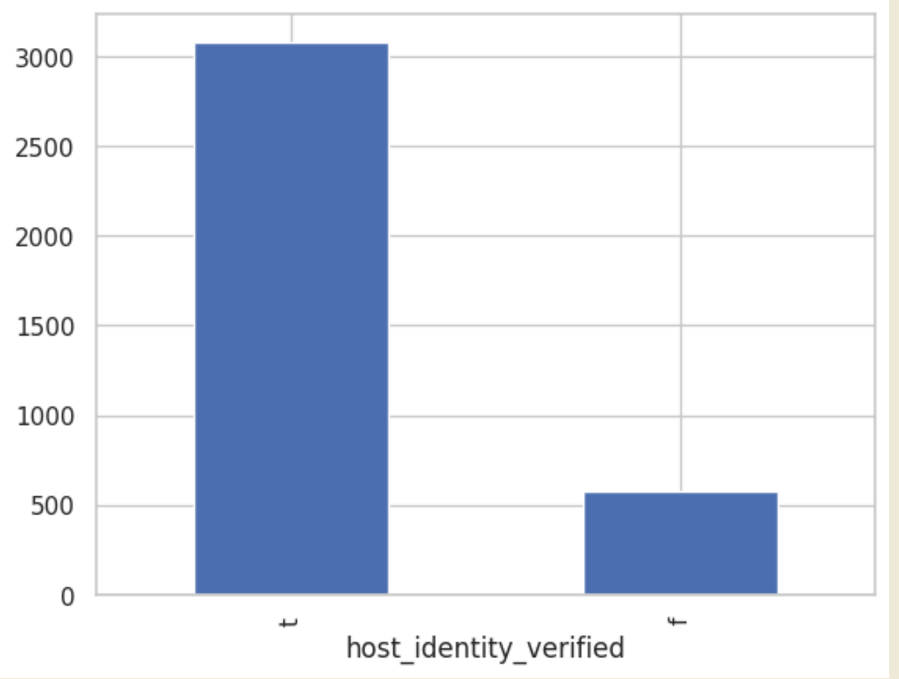
The distribution of the first 3 categorical features:

	host_response_time	host_is_superhost	host_identity_verified
count	2341	3132	3132
unique	4	2	2
top	within an hour	f	t
freq	1337	2415	2585

CATEGORICAL FEATURES







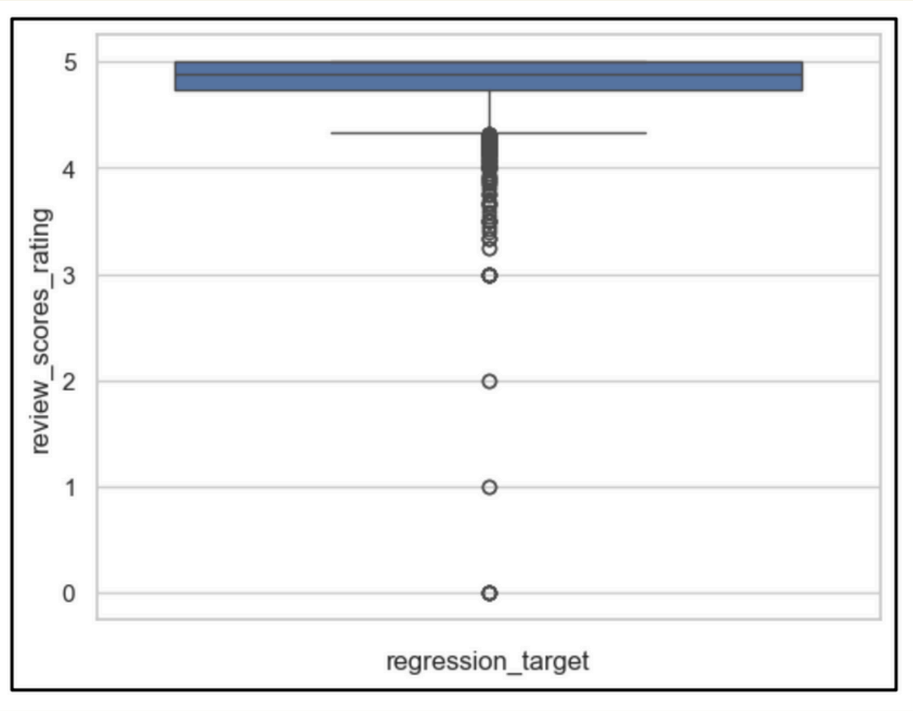
OUTLIERS

Column Name	Number of Outliers
host_listings_count	6
accommodates	59
price	105
minimum_nights	9
number_of_reviews	130
latest_review	86
calculated_host_listings_count	141
calculated_host_listings_count_entire_homes	102
calculated_host_listings_count_private_rooms	136
calculated_host_listings_count_shared_rooms	39
reviews_per_month	41

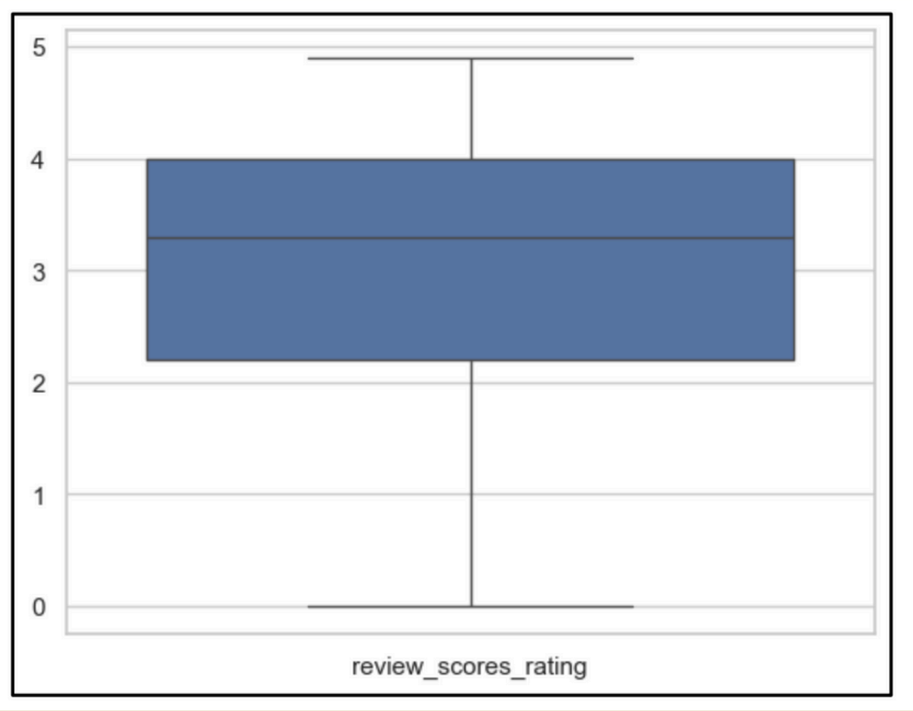
ADJUSTMENT FOR TARGET

- Removed some data with a rating of less than 4.5
- Scale the numerical data to the range (0,5) by MinMaxScaler()

before



after



SIZE OF DATASET 2

- 20 columns and 40078 entries
- 18 features and 2 targets
- 12 numerical features and 6 categorical features
- No Missing Values in the this dataset

Features	'City', 'Price', 'Day', 'Room Type', 'Shared Room', 'Private Room', 'Person Capacity', 'Superhost', 'Multiple Rooms', 'Business', 'Cleanliness Rating', 'Bedrooms', 'City Center (km)', 'Metro Distance (km)', 'Attraction Index', 'Normalised Attraction Index', 'Restraunt Index', 'Normalised Restraunt Index'
Targets	‘regression_target’ and ‘classification_target__2’

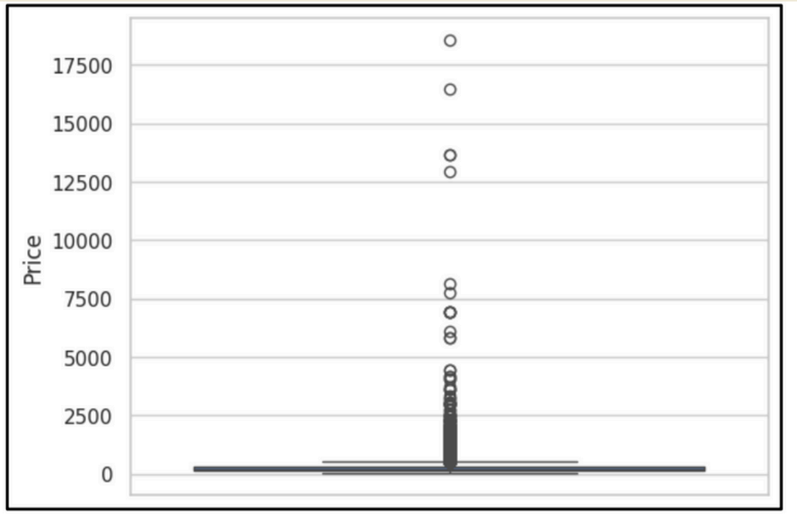
NUMERICAL FEATURE DISTRIBUTION

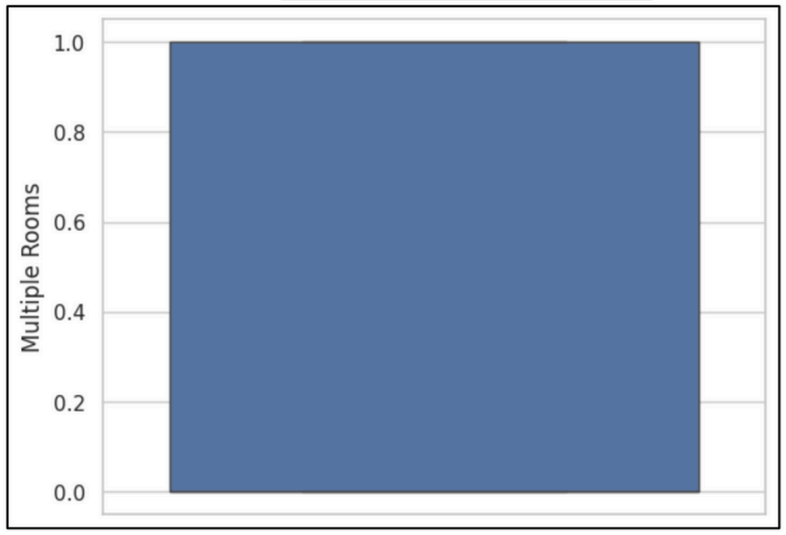
Continuous	'Price', 'Person Capacity', 'Cleanliness Rating', 'City Center (km)', 'Metro Distance (km)', 'Attraction Index', 'Normalised Attraction Index', 'Restraunt Index', 'Normalised Restraunt Index'
Discrete	'Multiple Rooms', 'Business', 'Bedrooms'
Total	12 numerical features

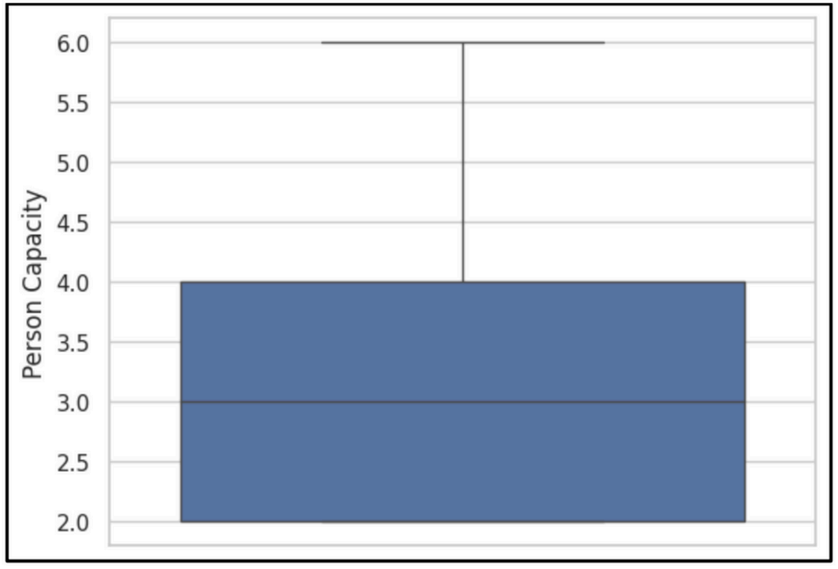
NUMERICAL FEATURE DISTRIBUTION

The distribution of the first 3 numerical features:

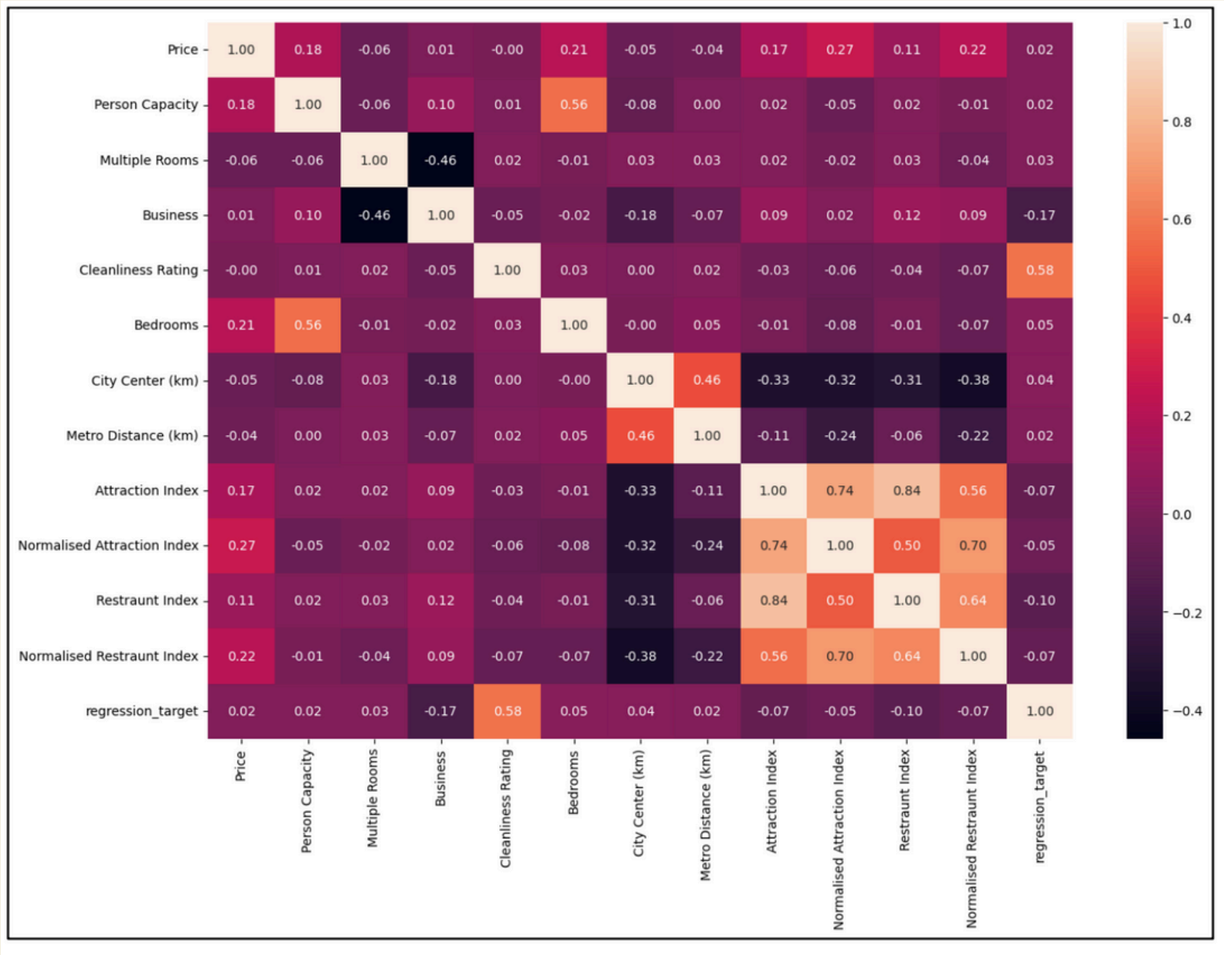
	Price	Person Capacity	Multiple Rooms
count	40078	40078	40078
mean	260.385731	3.241654	0.298318
std	282.724361	1.300120	0.457526
min	34.779339	2	0
25%	144.016085	2	0
50%	203.881325	3	0
75%	296.389566	4	1
max	18545.450285	6	1


Price
17500
15000
12500
10000
7500
5000
2500
0


Multiple Rooms
1.0
0.8
0.6
0.4
0.2
0.0


Person Capacity
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0

CORRELATION



CATEGORICAL FEATURE DISTRIBUTION

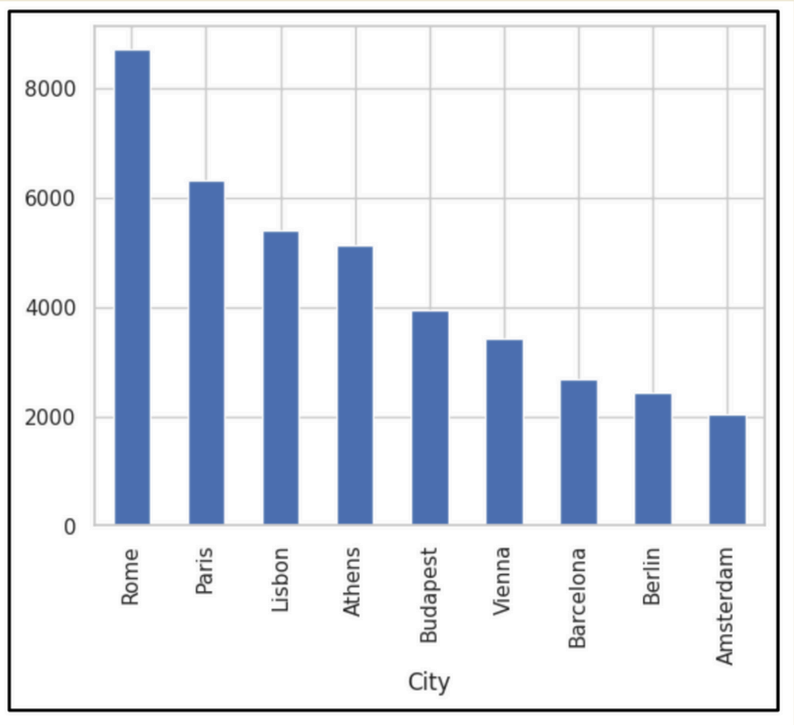
Binary	'City', 'Day', 'Room Type'
Nominal	'Shared Room', 'Private Room', 'Superhost'
Total	6 categorical features

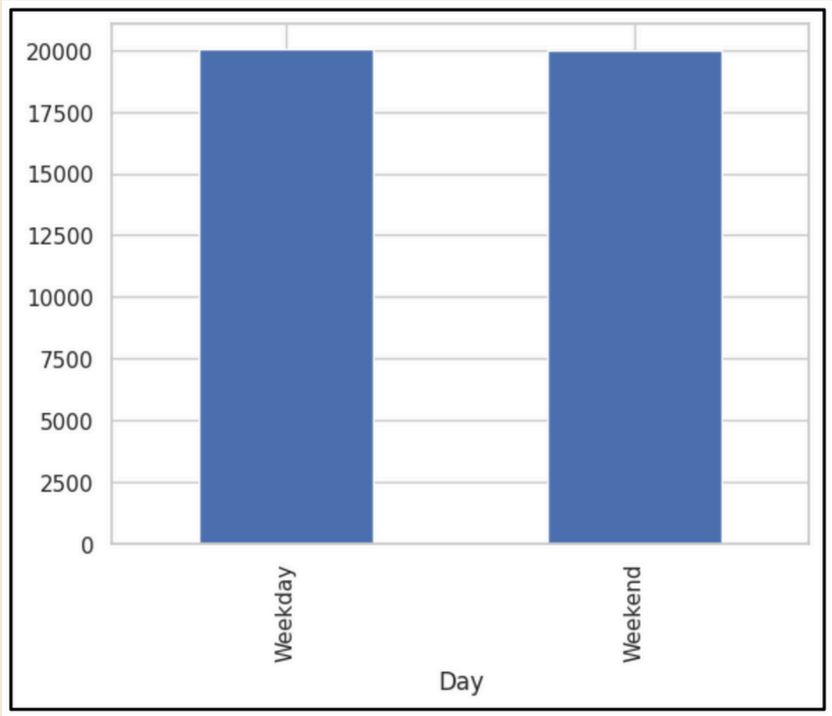
CATEGORICAL FEATURE DISTRIBUTION

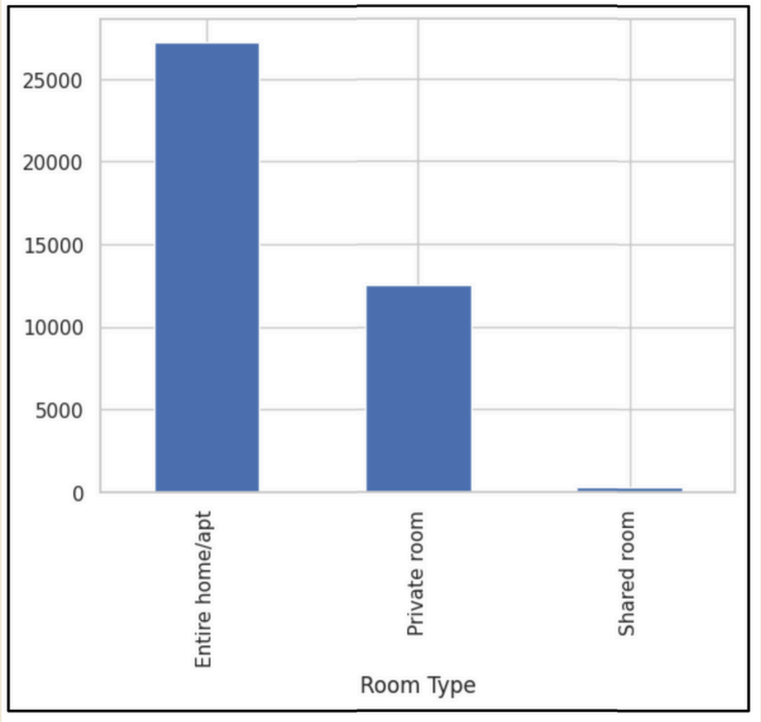
The distribution of the first 3 categorical features:

	City	Day	Room Type
count	40078	40078	40078
unique	9	2	3
top	Rome	Weekday	Entire home/apt
freq	8713	20063	27237

CATEGORICAL FEATURES







OUTLIERS

Column Name	Number of Outliers
Price	356
Cleanliness Rating	520
Bedrooms	86
City Center (km)	598
Metro Distance (km)	833
Attraction Index	540
Normalised Attraction Index	502
Restraunt Index	610
Normalised Restraunt Index	236

ADJUSTMENT FOR TARGET

- Removed some data with a rating of less than 4
- Scale the numerical data to the range (0,5) by MinMaxScaler()

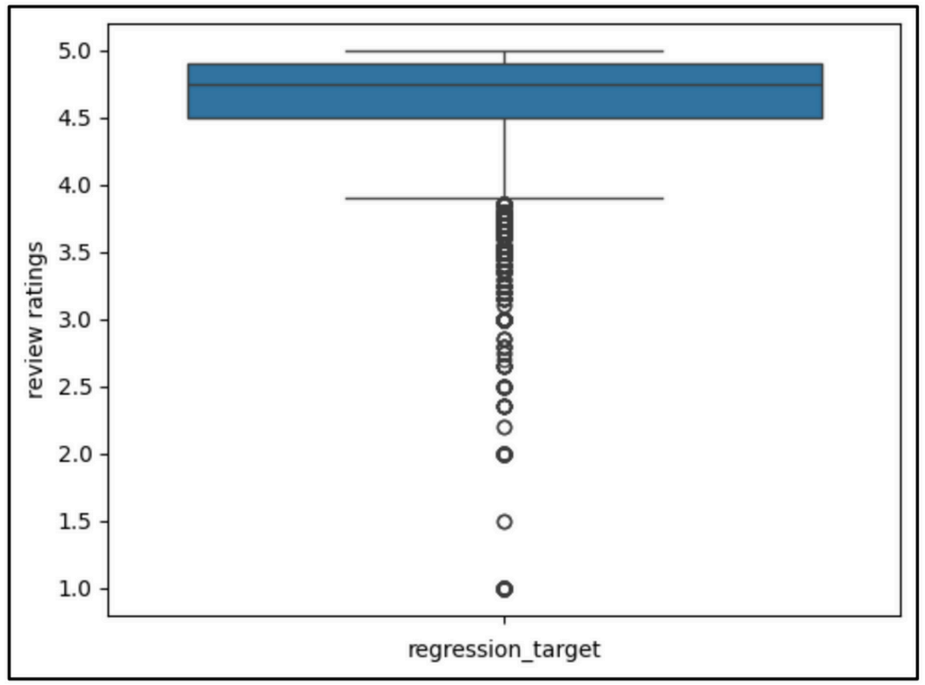


Figure 1. Before Adjustments

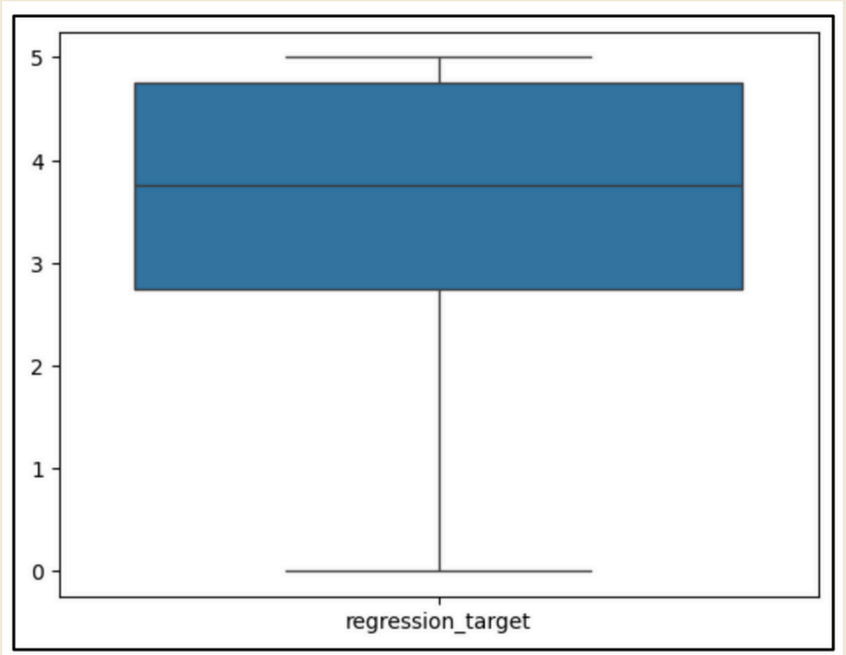


Figure 2. After Adjustments

PREPROCESSING DATASET 1

MISSING VALUES

Column Name	Strategy Used
host_response_time	constant
bathrooms_total	mean
bathroom_type	most_frequent
bedrooms	mean
beds	mean

NORMALIZATION

RobustScaler()

ENCODING CATEGORICAL FEATURES

One-Hot Encoder	'host_is_superhost', 'host_identity_verified', 'has_availability', 'instant_bookable', 'neighbourhood_cleansed', 'property_type', 'room_type', 'bathroom_type', 'amenities'
Ordinal Encoder	'host_response_time'

FEATURE SELECTION

AMENITIES

Select the best 50 amenities using SelectKBest, where K=50

FEATURES

Select the best features using SelectKBest, where
K = 50 for model 1 and
K = 40 for model 5

PREPROCESSING DATASET 2

MISSING VALUES

No Missing Values

NORMALIZATION

RobustScaler()

ENCODING CATEGORICAL FEATURES

One-Hot Encoder	'City', 'Day', 'Room Type', 'Shared Room', 'Private Room', 'Superhost'

FEATURE SELECTION

FEATURES

Select the best features using SelectKBest, where
K = 20 for model 1 and model 5

MODEL

MODEL 1

	Input Feature Description	Total Input
Dataset 1	Best 50 features among the best 50 amenities and all other features	50 features
Dataset 2	Best 20 features among all the features	20 features

MODEL 2

	Input Feature Description	Total Input
Dataset 1	Neighbourhood features	22 features
Dataset 2	City features	9 features

MODEL 3 (1)

	Input Feature Description	Total Input
Dataset 1	the whole amenities features	884 features

MODEL 3 (2)

	Input Feature Description	Total Input
Dataset 1	Best 50 Amenities	50 features

MODEL 4

	Input Feature Description	Total Input
Dataset 1	All features, excluding the neighborhood and amenities	78 features
Dataset 2	all the features, excluding the city	22 features

MODEL 5

	Input Feature Description	Total Input
Dataset 1	best 40 features out of the all features, excluding the neighborhood and amenities	40 features
Dataset 2	best 20 features out of all the features, excluding the city	20 features

MODEL 6

	Input Feature Description	Total Input
Dataset 1	[calculated_host_listings_count, calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms, price, bathrooms_total, bedrooms, latest_review, and the categorical features (excluding the neighborhood and amenities)]	10 features
Dataset 2	['Price', 'Person Capacity', 'Superhost', 'Multiple Rooms', 'Business', 'Cleanliness Rating', 'Bedrooms', 'City Center (km)', 'Metro Distance (km)', 'Normalised Attraction Index', 'Normalised Restraunt Index', 'Day_Weekday', 'Day_Weekend', 'Room Type_Entire home/apt', 'Room Type_Private room', 'Room Type_Shared room', 'Shared Room_True', 'Private Room_True']	18 features

MODEL 7

	Input Feature Description	Total Input
Dataset 1	SKIP	NONE
Dataset 2	All features	31 features

LINEAR REGRESSION

Here is the R2 Score and the MSE of each model:

	Dataset 1	Dataset 2
Model 1	**0.2152; 1.1894**	**0.3663; 1.0581**
Model 2	-0.0204; 1.5464	0.0440; 1.5964
Model 3 (1)	-3.8784; 5.8779	-
Model 3 (2)	0.0743; 1.4029	-
Model 4	0.1399; 1.3036	0.3619; 1.0655
Model 5	0.1606; 1.2722	**0.3663; 1.0581**
Model 6	0.1574; 1.2771	0.3574; 1.0730
Model 7	-	0.3747; 1.0441

POLYNOMIAL (LOGARITHM) REGRESSION

Here is the R2 Score and the MSE of each model:

	Dataset 1	Dataset 2
Model 1	-0.0054; 1.5238	**0.3584; 1.0713**
Model 2	-0.0072; 1.5265	0.0435; 1.5972
Model 3 (1)	0.07437; 1.4029	-
Model 3 (2)	0.0653; 1.4167	-
Model 4	-0.0054; 1.5238	0.3545; 1.0778
Model 5	-0.0054; 1.5238	**0.3584; 1.0713**
Model 6	**0.1120; 1.3458**	0.3506; 1.0843
Model 7	-	0.3473; 1.0898

STOCHASTIC GRADIENT DESCENT REGRESSION

Here is the R2 Score and the MSE of each model:

	Dataset 1	Dataset 2
Model 1	-1.2950; 1.9567	0.3654; 1.0596
Model 2	0.0046; 1.5351	0.0413; 1.6009
Model 3 (1)	**0.1298**; 1.4140	-
Model 3 (2)	0.0659; 1.4053	-
Model 4	-4.3218e+30; 6.5300e+30	0.3574; 1.0730
Model 5	-5.9373e+30; 8.9710e+30	0.3634; 1.0630
Model 6	0.1198; **1.2805**	0.3557; 1.0758
Model 7	-	**0.3747; 1.0441**

FEEDFORWARD NEURAL NETWORKS

Here is the R2 Score and the MSE of each model:

	Dataset 1	Dataset 2
Model 1	-0.0661; 1.4660	0.3726; 1.0477
Model 2	-0.0243; 1.4086	0.0437; 1.5969
Model 3 (1)	0.0497; 1.3067	-
Model 3 (2)	0.0540; 1.3008	-
Model 4	-0.0345; 1.4226	0.3709; 1.0504
Model 5	-0.0971; 1.5087	0.3770; 1.0404
Model 6	**0.1097; 1.2242**	0.3618; 1.0657
Model 7	-	**0.3864; 1.0246**

DEEP NEURAL NETWORKS WITH DROPOUT

Here is the R2 Score and the MSE of each model:

	Dataset 1	Dataset 2
Model 1	0.0470; 1.4533	-0.0025; 1.8066
Model 2	-0.0395; 1.5717	-0.0212; 1.7921
Model 3 (1)	0.0106; 1.4606	-
Model 3 (2)	0.0353; 1.5155	-
Model 4	0.0367; 1.5243	-0.2355; 2.2054
Model 5	0.0265; 1.5224	-0.0289; 1.8348
Model 6	**0.1252; 1.3016**	**0.3108; 1.2314**
Model 7	-	-0.04702; 1.7459

CONCLUSION REGRESSION

DATASET 1

The best model is **Model 1** using the Linear Regression:

R2 Score: **0.2152**
MSE: **1.1894**

DATASET 2

The best model is **Model 7** using the Feedforward Neural Networks:

R2 Score: **0.3864**
MSE: **1.0246**

LOGISTIC REGRESSION

Here is the Accuracy and F1 Score of each model:

	Dataset 1	Dataset 2
Model 1	0.4705; 0.6399	0.7888; 0.8363
Model 2	0.4896; 0.5137	0.6290; 0.7380
Model 3 (1)	0.5726; 0.3431	-
Model 3 (2)	0.5742; **0.6616**	-
Model 4	0.5295; 0.0	0.7767; 0.8355
Model 5	0.4705; 0.6399	0.7887; 0.8332
Model 6	**0.6061;** 0.6308	0.7919; **0.8387**
Model 7	-; -	**0.8312;** 0.2977

FEEDFORWARD NEURAL NETWORKS

Here is the Accuracy and F1 Score of each model:

Model	Dataset 1	Dataset 2
1	0.5396; 0.5417	0.7921; 0.8378
2	0.4805; 0.4577	0.6287; 0.7457
3 (1)	0.5773; 0.6220	-; -
3 (2)	0.5810; 0.5496	-; -
4	0.5305; 0.5886	0.7916; **0.8401**
5	0.4795; **0.6418**	0.7902; 0.8360
6	**0.6220**; 0.5861	0.7919; 0.8387
7	-; -	**0.7938;** 0.8396

DEEP NEURAL NETWORKS WITH DROPOUT

Here is the Accuracy and F1 Score of each model:

	Dataset 1	Dataset 2
Model 1	0.4508; 0.6215	0.6163; 0.7626
Model 2	0.5232; 0.0	0.6125; 0.7597
Model 3 (1)	0.5341; 0.0	-; -
Model 3 (2)	**0.5396;** 0.0	-; -
Model 4	0.4809, **0.6494**	0.6286; **0.7719**
Model 5	0.5382, 0.0	0.6202; 0.7656
Model 6	0.5368; 0.0.	0.6181; 0.7640
Model 7	-; -	**0.7742;** 0.6316

CONCLUSION CLASSIFICATION

DATASET 1

The best model is **Model 6** using the Feedforward Neural Networks:

Accuracy: **0.6220**

DATASET 2

The best model is **Model 7** using the Logistic Regression:

Accuracy: **0.8312**

PERFORMANCE ENHANCEMENT

REGRESSION (MODEL 1)

- Linear Regression from scikit-learn
- Best Parameter: {'copy_X': True, 'fit_intercept': False, 'n_jobs': None}
- Resulting R2 score = 0.2151
- Resulting MSE score = 1.1894

Tuned Parameters	Value
'fit_intercept’	True, False
'copy_X’	True, False
'n_jobs’	None, 3, 10

CLASSIFICATION (MODEL 6)

- MLPClassifier from scikit-learn
- Best Parameter: {'activation': 'relu', 'learning_rate': 'constant', 'learning_rate_init': 0.1, 'max_iter': 100}
- Resulting Accuracy = 0.5946
- Resulting F1 score = 0.6184

Trained Parameters	Value
‘activation’	'logistic', 'relu'
‘learning_rate’	‘constant’, ‘invscaling’
‘learning_rate_init’	0.001, 0.01, 0.1
‘max_iter’	100, 300

REGRESSION (MODEL 7)

- MLPRegressor from scikit-learn
- Best Parameter: {'activation': 'relu', 'learning_rate': 'constant', 'learning_rate_init': 0.01, 'max_iter': 100}
- Resulting R2 score = 0.3892
- Resulting MSE score = 1.0199

Tuned Parameters	Value
'activation’	'relu','tanh'
' learning_rate’	'constant’
' learning_rate_init'	0.001, 0.01
‘max_iter’	100, 200, 300

CLASSIFICATION (MODEL 7)

- Logistic Regression from scikit-learn
- Best Parameter: {'C': 0.01, 'penalty': 'l2', 'solver': 'newton-cg'}
- Resulting Accuracy = 0.80
- Resulting F1 score = 0.84

Tuned Parameters	Value
‘C’	0.01, 0.1, 1, 10, 100
'solver'	'newton-cg', 'lbfgs', 'liblinear'
'penalty'	‘l2’

CONCLUSION

Overall best model:

- Dataset 1: Model 1
- Dataset 2: Model 7

How to implement the models:

1. Use classification model to get the classification: “Good” or “Bad”.
2. Use the coefficients in Regression model to give suggestions, accordingly.
3. Use regression model to get the predicted rating score.

```
Feature: Account_life, Coefficient: 1.3325456798804486e-05
Feature: host_listings_count, Coefficient: -0.02998798485239862
Feature: accommodates, Coefficient: -0.230100262965358 82
Feature: bathrooms_total, Coefficient: 0.044980577486611655
Feature: beds, Coefficient: -0.016591327996837112
Feature: price, Coefficient: 0.083015127189776 95
Feature: maximum_nights, Coefficient: -0.008872502411750595
Feature: latest_review, Coefficient: 0.016269996716158235
Feature: calculated_host_listings_count, Coefficient: -0.18787118040177792
Feature: calculated_host_listings_count_private_rooms, Coefficient: -0.10644043912490812
```

Negative coefficients mean that it affect the result negatively.
Positive coefficients mean that it affect the result positively.

THANK YOU

Submitted by: CHAN YEE KI (20860858) | LOKESWARA LOVERA (20798275)