Data Collection and Preprocessing Phase

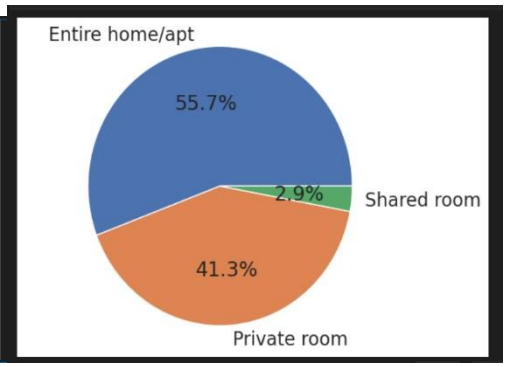
Date	18 JULY 2024
Team ID	739829
Project Title	Unveiling Airbnb Price Patterns: Machine Learning For Forecasting
Maximum Marks	6 Marks

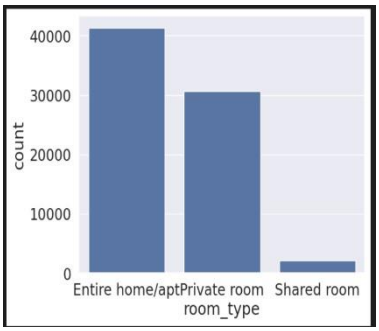
Data Exploration and Preprocessing Report

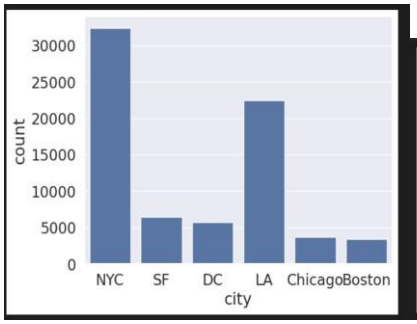
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

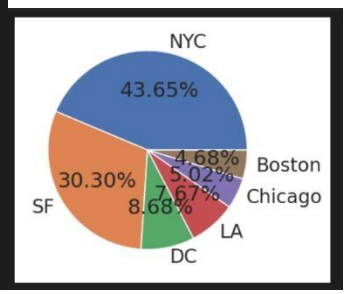
Section	Description
Data Overview	Dimension: 614 rows × 13 columns Descriptive statistics: `df.describe()`

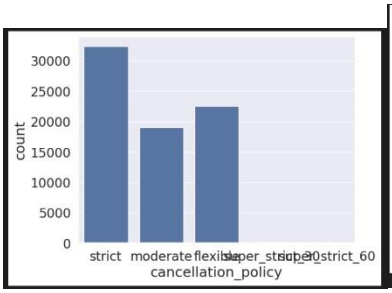
	hotel_id	log_price	accommodates	bathrooms	latitude	longitude	number_of_reviews	bedrooms	beds
count	7.411100e+04	74111.000000	74111.000000	74111.000000	74111.000000	74111.000000	74111.000000	74111.000000	74111.000000
mean	1.126662e+07	4.782069	3.155146	1.235302	38.445958	-92.397525	20.900568	1.265467	1.710677
std	6.081735e+06	0.717394	2.153589	0.582192	3.080167	21.705322	37.828641	0.851671	1.253673
min	3.440000e+02	0.000000	1.000000	0.000000	33.338905	-122.511500	0.000000	0.000000	0.000000
25%	6.261964e+06	4.317488	2.000000	1.000000	34.127908	-118.342374	1.000000	1.000000	1.000000
50%	1.225415e+07	4.709530	2.000000	1.000000	40.662138	-76.996965	6.000000	1.000000	1.000000
75%	1.640226e+07	5.220356	4.000000	1.000000	40.746096	-73.954660	23.000000	1.000000	2.000000
max	2.123090e+07	7.600402	16.000000	8.000000	42.390437	-70.985047	605.000000	10.000000	18.000000

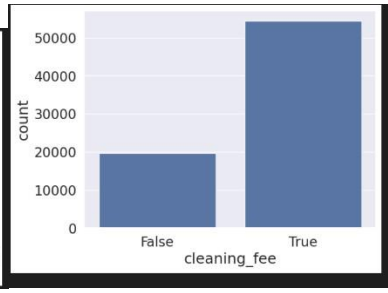


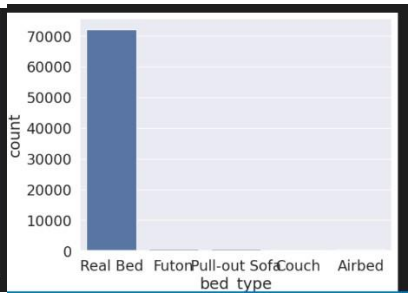


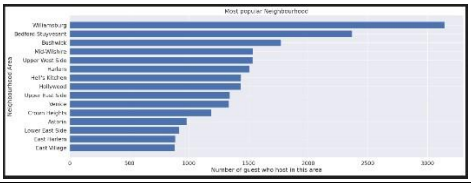




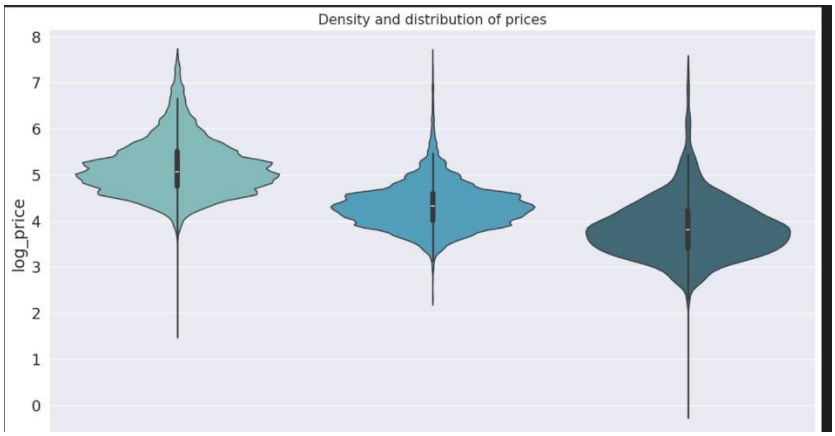


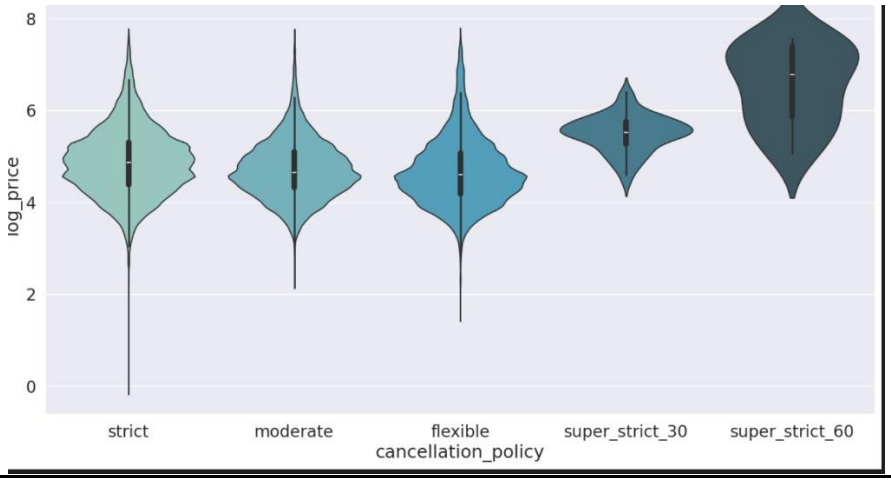




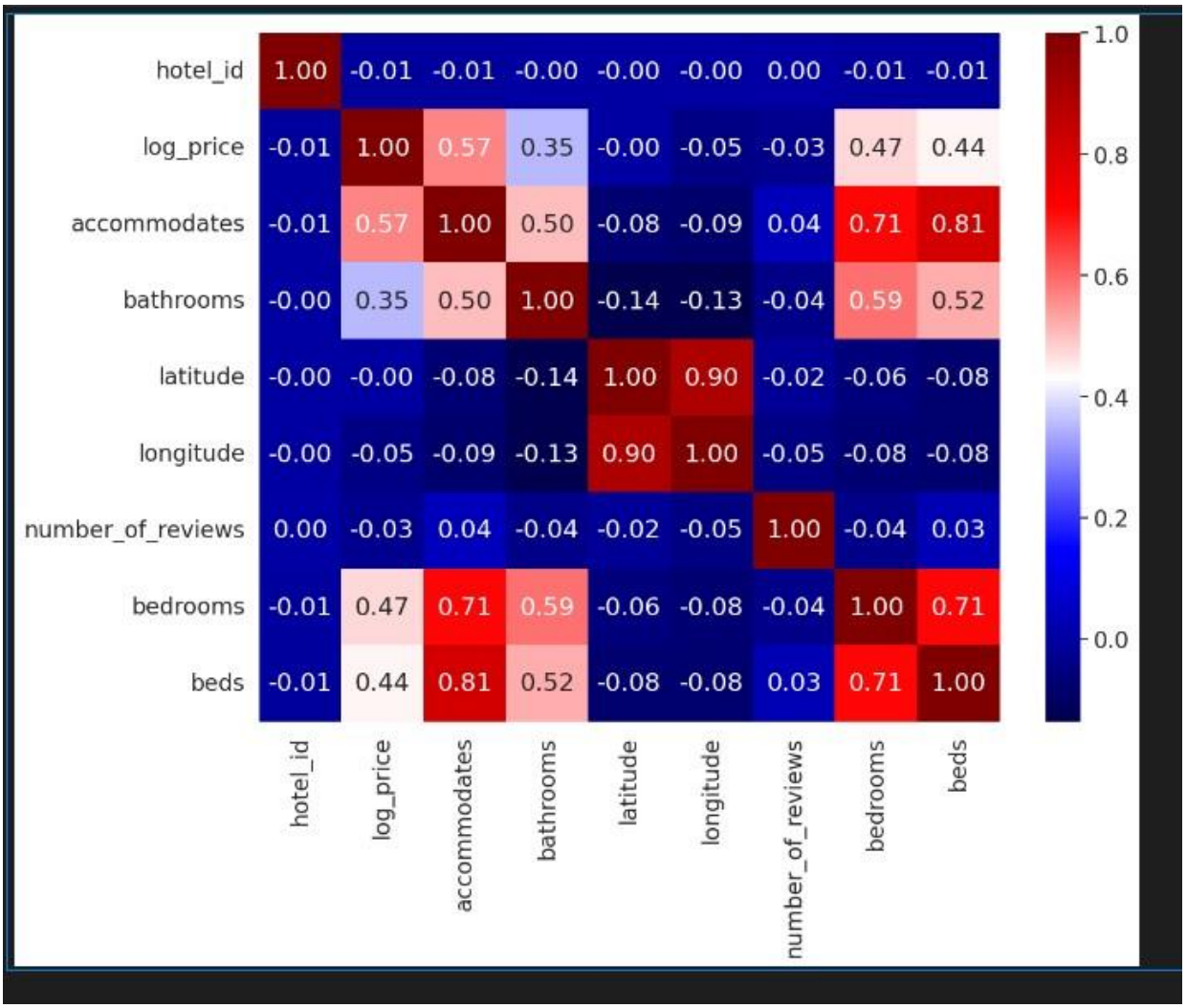


Univariate Analysis

Bivariate Analysis	 


8
6
4
2
0
log_price
strict
moderate
flexible
super_strict_30
super_strict_60
cancellation_policy

log_price
8
7
6
5
4
3
2
1
0
Real Bed
Futon
Pull-out Sofa
Couch
Airbed
bed_type

Multivariate Analysis	

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading Data	
Handling Missing Data	hotel_id 0 log_price 0 property_type 0 room_type 0 accommodates 0 bathrooms 200 bed_type 0 cancellation_policy 0 cleaning_fee 0 city 0 description 0 host_identity_verified 188 host_since 188 instant_bookable 0 latitude 0 longitude 0 name 0 neighbourhood 6872 number_of_reviews 0 bedrooms 91 beds 131 dtype: int64

Data Transformation	`from sklearn.preprocessing import StandardScaler`
	`import numpy as np`
	
	`# Generate random data with matching dimensions`
	`x_train = np.random.rand(55583, 13)`
	`y_train = np.random.rand(55583)`
	`x_test = np.random.rand(18528, 13)`
	`y_test = np.random.rand(18528)`
	
	`# Scale the data`
	`scaler = StandardScaler()`
	`x_train_scaled = scaler.fit_transform(x_train)`
	`x_test_scaled = scaler.transform(x_test)`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	-