2025

4Geeks Academy: data science cohort 12

DAY 17: EXPLORATORY DATA ANALYSIS

TODO

EXPLORATORY DATA ANALYSIS	EDA philosophy, workflow and techniques
LOGISTIC REGRESSION PROJECT	Finish Logistic Regression Project Tutorial (Your first ML Algorithm module), last look at model eval & next steps
EDA PROJECT	Start Data Preprocessing Project Tutorial (Exploratory data analysis project) plan to complete as much as possible by Monday - we will go over solution together.

TOPICS

01 EXPLORATORY DATA ANALYSIS

02 OBSERVING DATA

03 OBSERVING INTERACTIONS

04 CLEANING DATA

EXPLORATORY DATA ANALYSIS

WHAT

Get to know your data! Look at it from every possible angle and become an expert on your dataset.

WHY

- Identify and fix problems in the data
- Utilize the data effectively

HOW

- Look at structure of dataset
  - How much data do we have?
  - How many features are there?
  - What type of data is each feature?
- Look at composition of each feature
  - Descriptive statistics
  - Data visualizations
- Look at interactions between features
  - Statistical tests
  - Correlation coefficients
  - Data visualizations

OBSERVING DATA

STRUCTURE

```
data_df.head()
```
✓ 0.0s

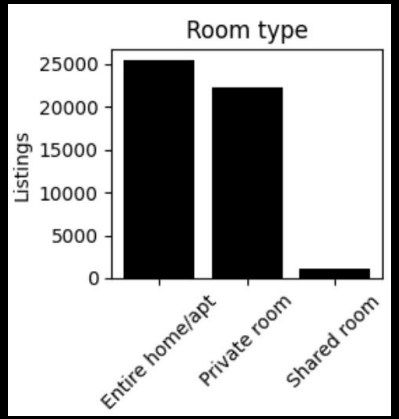
	id	price	minimum_nights	reviews_per_month	calculated_host_listings_count
0	2539	149	1	0.21	6
1	2595	225	1	0.38	2
2	3647	150	3	NaN	1
3	3831	89	1	4.64	1
4	5022	80	10	0.10	1

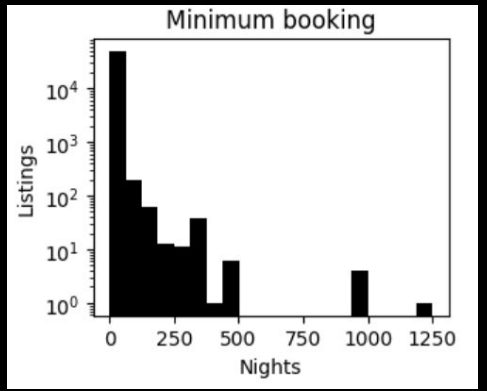
DESCRIPTIVE STATISTICS

```
data_df.describe()
```
✓ 0.0s

	id	price	minimum_nights	reviews_per_month	calculated_host_listings_count
count	4.889500e+04	48895.000000	48895.000000	38843.000000	48895.000000
mean	1.901714e+07	152.720687	7.029962	1.373221	7.143982
std	1.098311e+07	240.154170	20.510550	1.680442	32.952519
min	2.539000e+03	0.000000	1.000000	0.010000	1.000000
25%	9.471945e+06	69.000000	1.000000	0.190000	1.000000
50%	1.967728e+07	106.000000	3.000000	0.720000	1.000000
75%	2.915218e+07	175.000000	5.000000	2.020000	2.000000
max	3.648724e+07	10000.000000	1250.000000	58.500000	327.000000

DATA VISUALIZATION





OBSERVING INTERACTIONS

TESTS

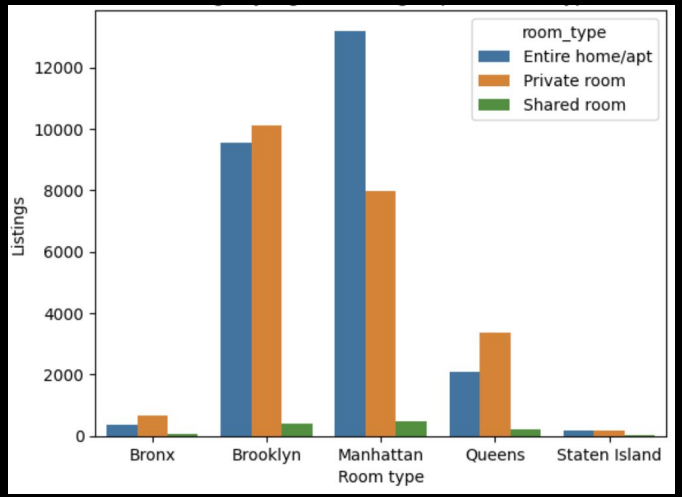
```
groups = data_df.groupby(['neighbourhood_group', 'room_type']).size()
chisquared_result = stats.chisquare(list(groups))
print(f'Chi-squared p-value = {chisquared_result.pvalue:.4f}')
```

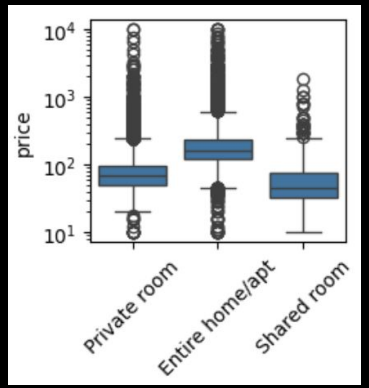
```
Chi-squared p-value = 0.0000
```

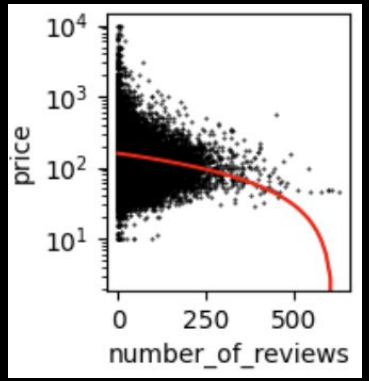
CORRELATIONS COEFS

	Feature 1	Feature 2	Spearman	Spearman p-value
0	number_of_reviews	reviews_per_month	0.706208	0.000000e+00
1	availability_365	calculated_host_listings_count	0.406606	0.000000e+00
2	availability_365	reviews_per_month	0.392126	0.000000e+00
3	availability_365	number_of_reviews	0.236664	0.000000e+00

DATA VISUALIZATION







CLEANING DATA

MISSING DATA

Missing data can hide in plain sight!

- Fill it in somehow
- Drop it

EXTREME VALUES

Extreme values (outliers) should not be arbitrarily thresholded away!

- Do nothing
- Fill it in somehow
- Drop it

ENCODING

Strings/objects must be converted into numbers

- Ordinal encoding
- One-hot encoding
- Fancy stuff: cyclical encoding with trig functions
- Something else?

WHEN ALTERING DATA THINK ABOUT THE CONTEXT & DOCUMENT EVERYTHING!