

Spatial Data Science Applied: **ArcPy & Scikit learn for predicting Hotel Room prices**

Nacho Moreno

Index

1 – Workflow

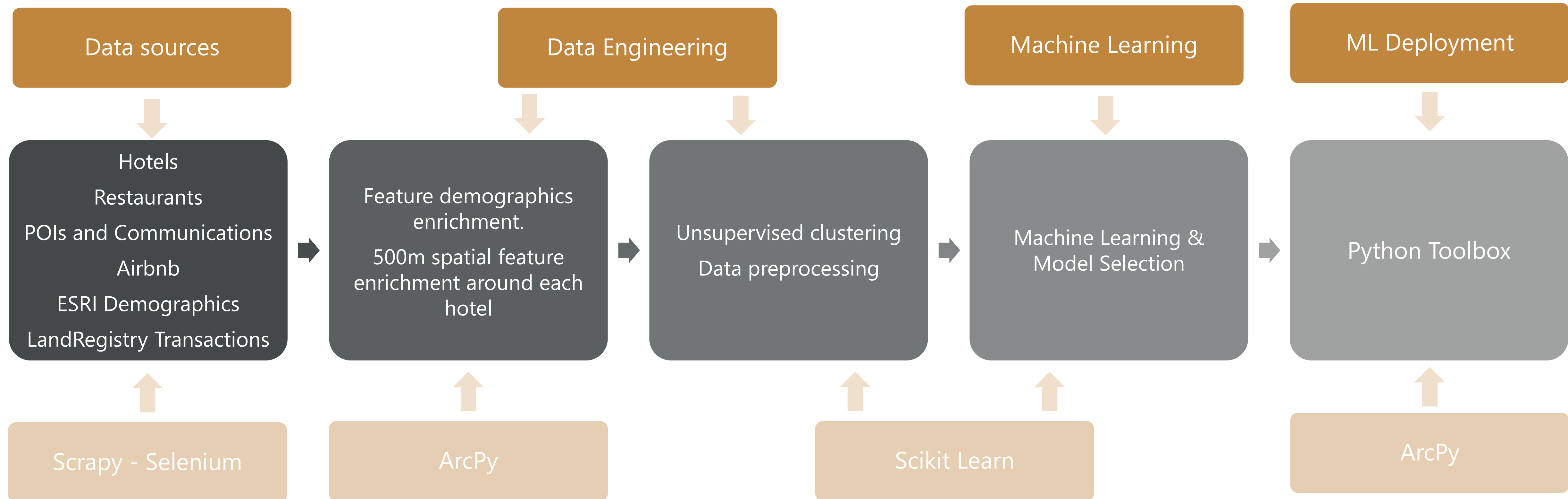
2 – Data Engineering

3 – Clustering & EDA

4 – Machine Learning modelling

5 – Tool deployment (Python Toolbox)

1- Project workflow and techonologies



2- Data Engineering

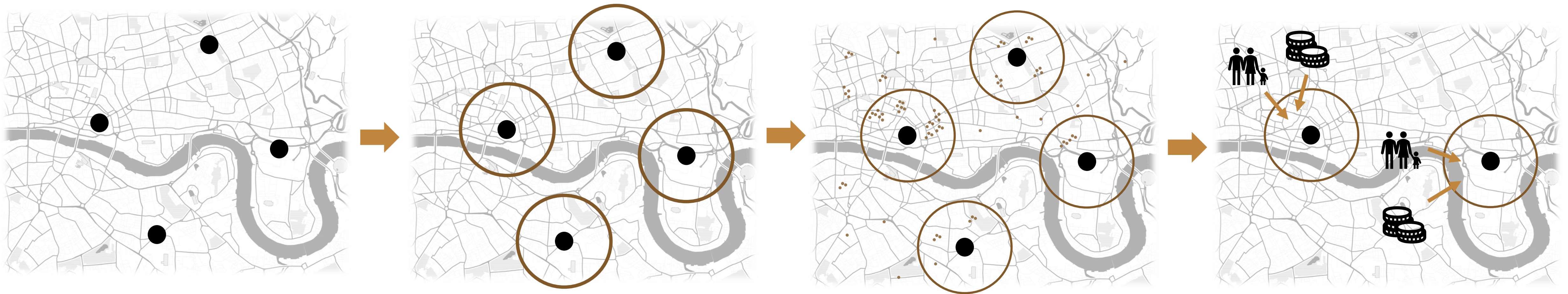
DATA SCRAPER COLLECTION

Hotel room price data information:

- Price per room per night (22nd of November 2018)
- Hotels classification (1 to 5 stars)
- Number of reviews and ranking
- Main hotel amenities: Internet connection, Room Service, Gym facilities, Swimming Pool, Parking and Air Conditioning.

TOTAL of 9 MAIN FEATURES

DATA FEATURE ENRICHMENT



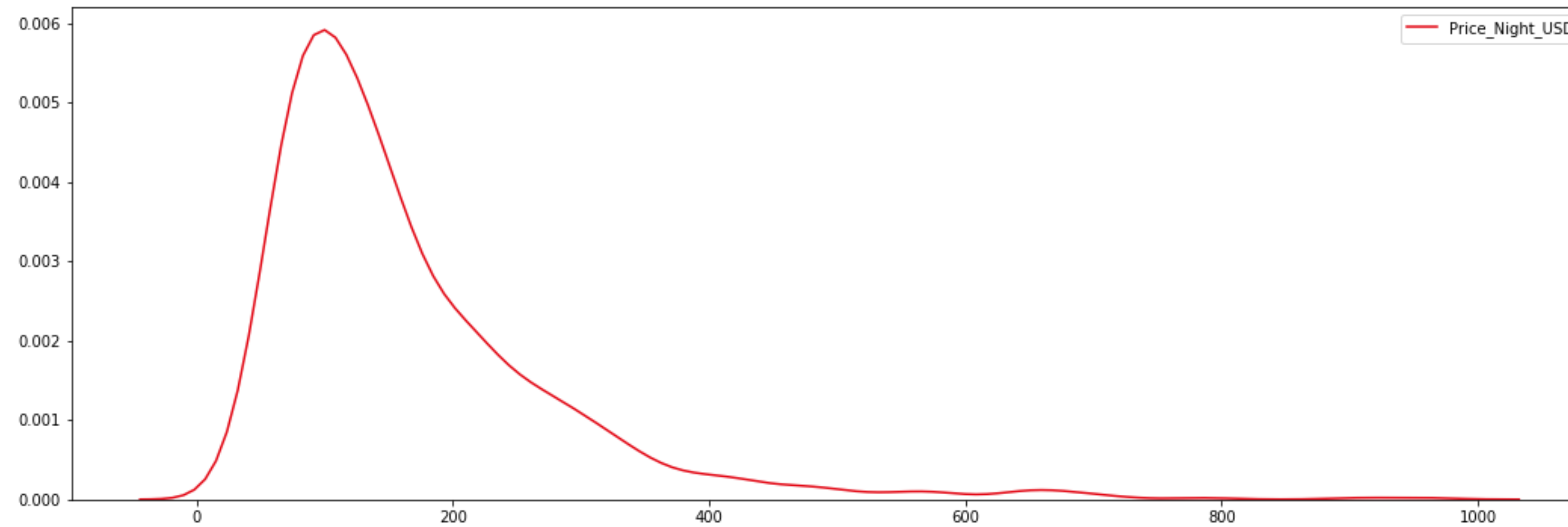
Each hotel is enriched with its vicinity (500m) data:

- Restaurants location, quality and reviews
- Airbnb number of beds available and median price
- House sale transactions number (last 12 months) and median price.
- Demographics: Purchasing Power, Household Composition and Total Population
- Average travel time to tourist POIs, Business centres and Airport accessibility.

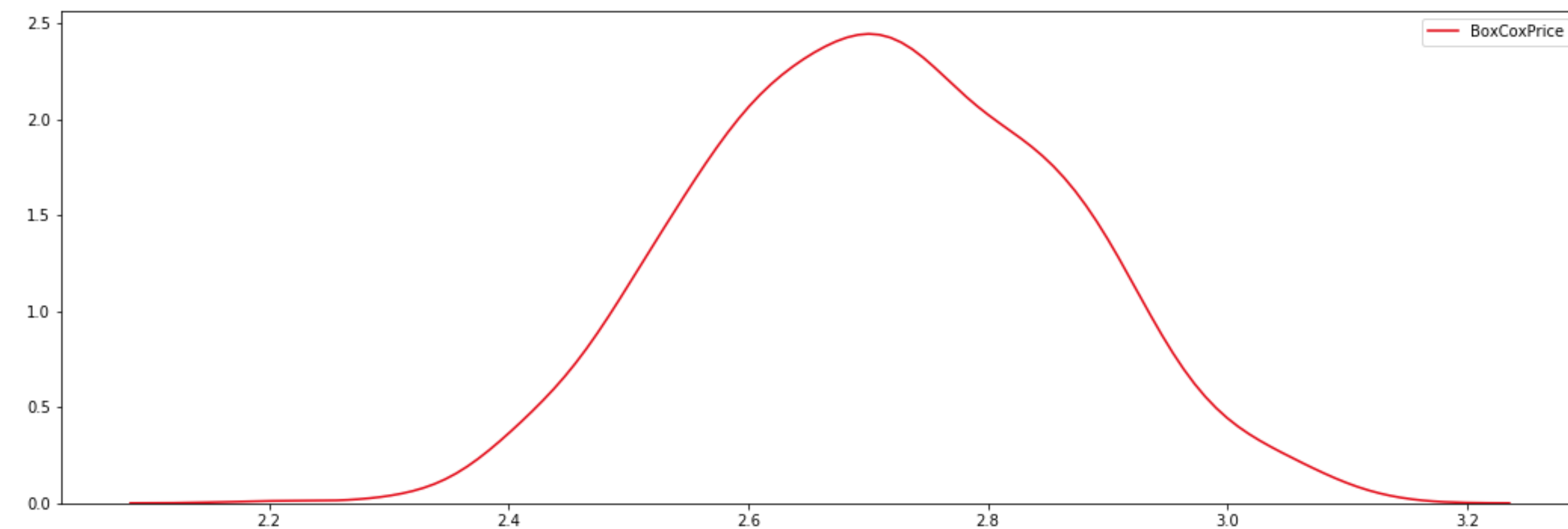
ADDITIONAL 15 FEATURES PER HOTEL

3- Clustering & EDA

ORIGINAL DATA DISTRIBUTION



BOXCOX TRANSFORMED DATA DISTRIBUTION



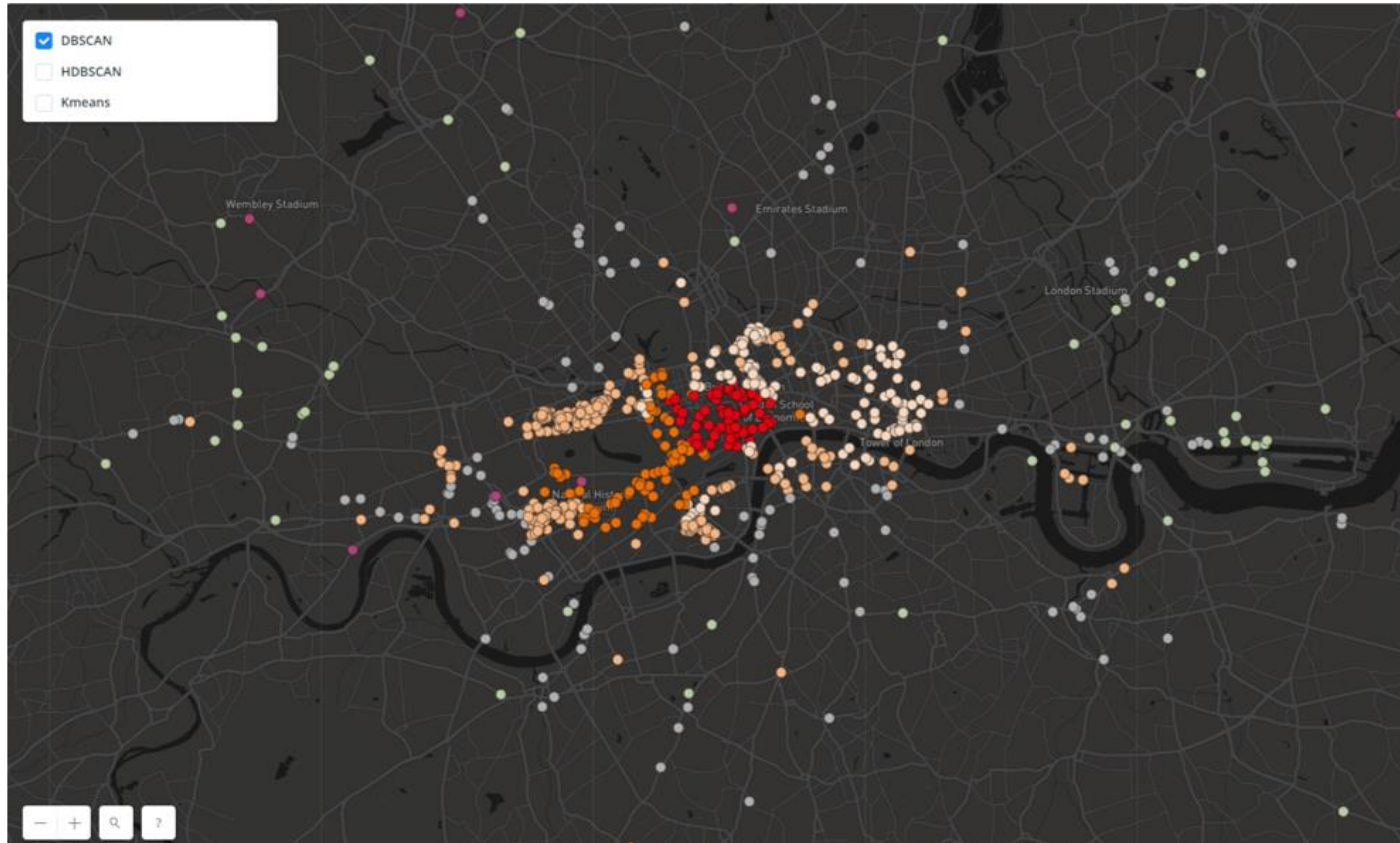
Skewness

DATA TRANSFORMATION

The hotel room price shows a non normal distribution reason why we apply a BoxCox transformation.

MAIN CORRELATION BETWEEN VARIABLES

Average_Household_Size																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
------------------------	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--



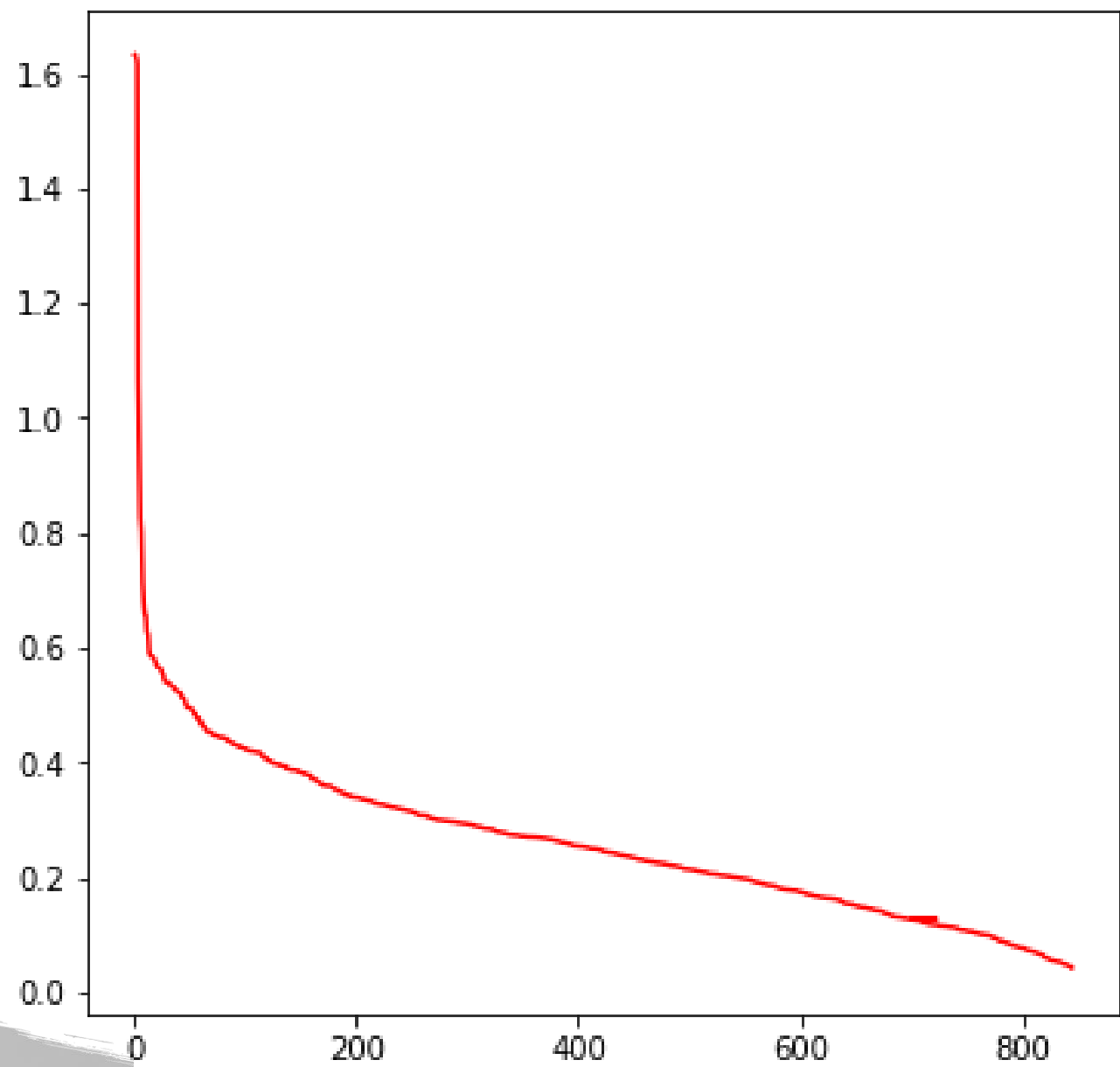
Clustering

We explore the data with unsupervised techniques in order to get an idea of the distribution of the different variables we will be studying: We will be focused in the Hotels dataset and how it is distributed in terms of different variables.

Using the following methods: K-means clustering, DBSCAN, HDBSCAN

CLUSTERING MAIN FINDINGS

DBSCAN Clusters	Price per room per night (\$)	Median House Price (£)	Average travel time to main Tourist POIs	Average Purchasing Power per Capita	Average number of sorrounding restaurants	Median Airbnb Price (\$)	Average Household size
-1	\$86	£731,700	50.9	£26,746	9	\$57	2.5
5	\$83	£461,704	41.1	£23,581	9	\$54	2.4
3	\$292	£1,732,278	21.8	£43,947	562	\$153	1.8
1	\$286	£2,017,518	26.3	£50,580	155	\$163	1.9
0	\$200	£722,277	23.2	£35,104	213	\$113	2.1
2	\$131	£866,737	29.5	£43,855	110	\$104	1.9
4	\$105	£628,660	36.2	£29,842	40	\$71	2.2



- Initially we create an elbow graph using the nearest neighbours that will allow us to determine the best epsilon for our DBSCAN model. The value of epsilon will ultimately determine the number of clusters we will have.
- We finally obtain a total of 6 clusters with some items misclassified (cluster -1). We see that the clusters 1 and 3 with the highest price per room are located in areas with a great number of restaurants and in expensive residential areas.
- We also observe lower travel times to Tourist Points of Interest as well as a higher cost of Airbnb prices.
- The cluster englobing the non classified hotels are located quite far from tourist locations and in what seems to be more family residential areasaccoridng to the average household size.

4 – Machine Learning modelling

We will implement the following ML models for predicting the hotel room price:

- Linear Model: Elastic Net
- Tree modelling: XGBoost

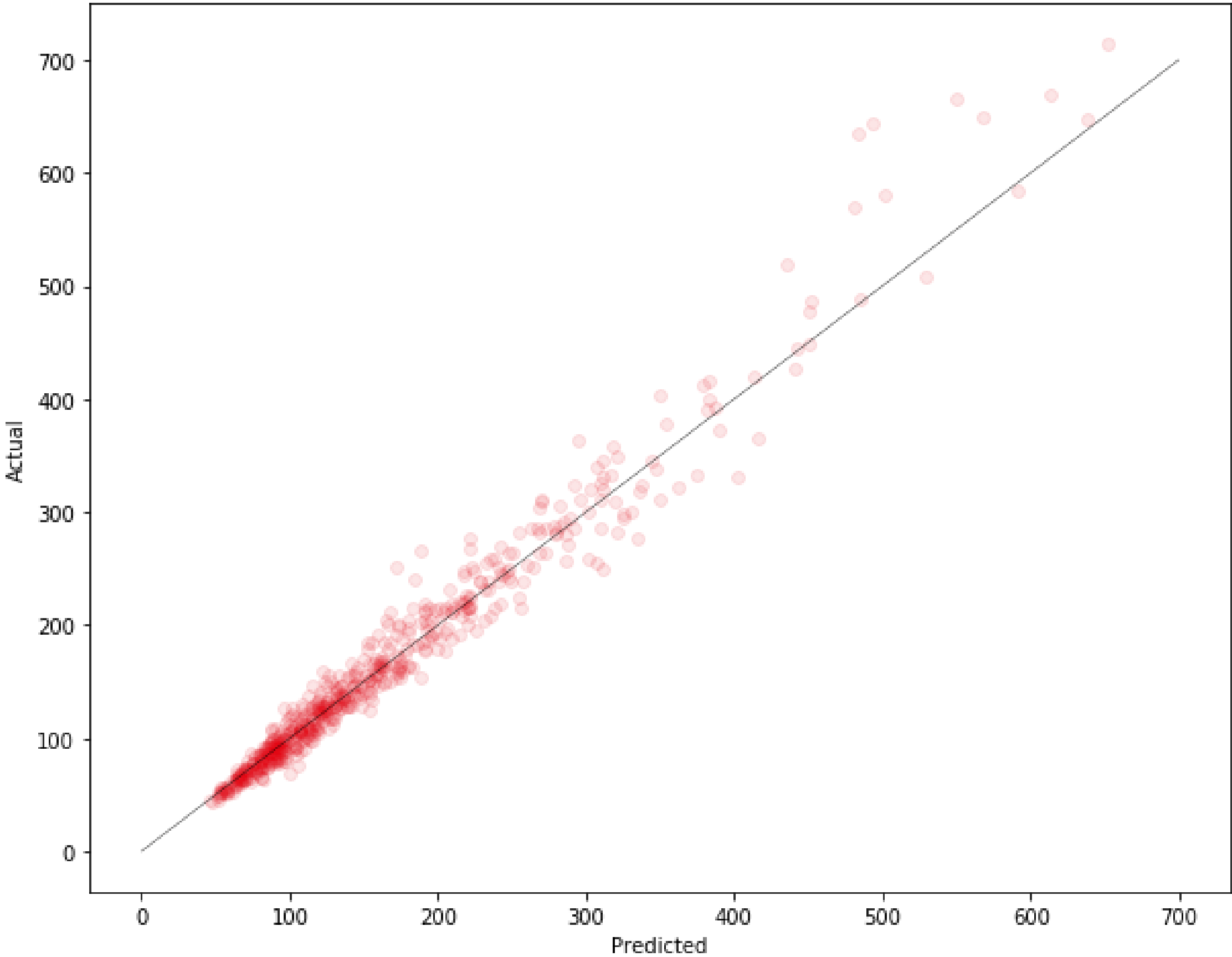
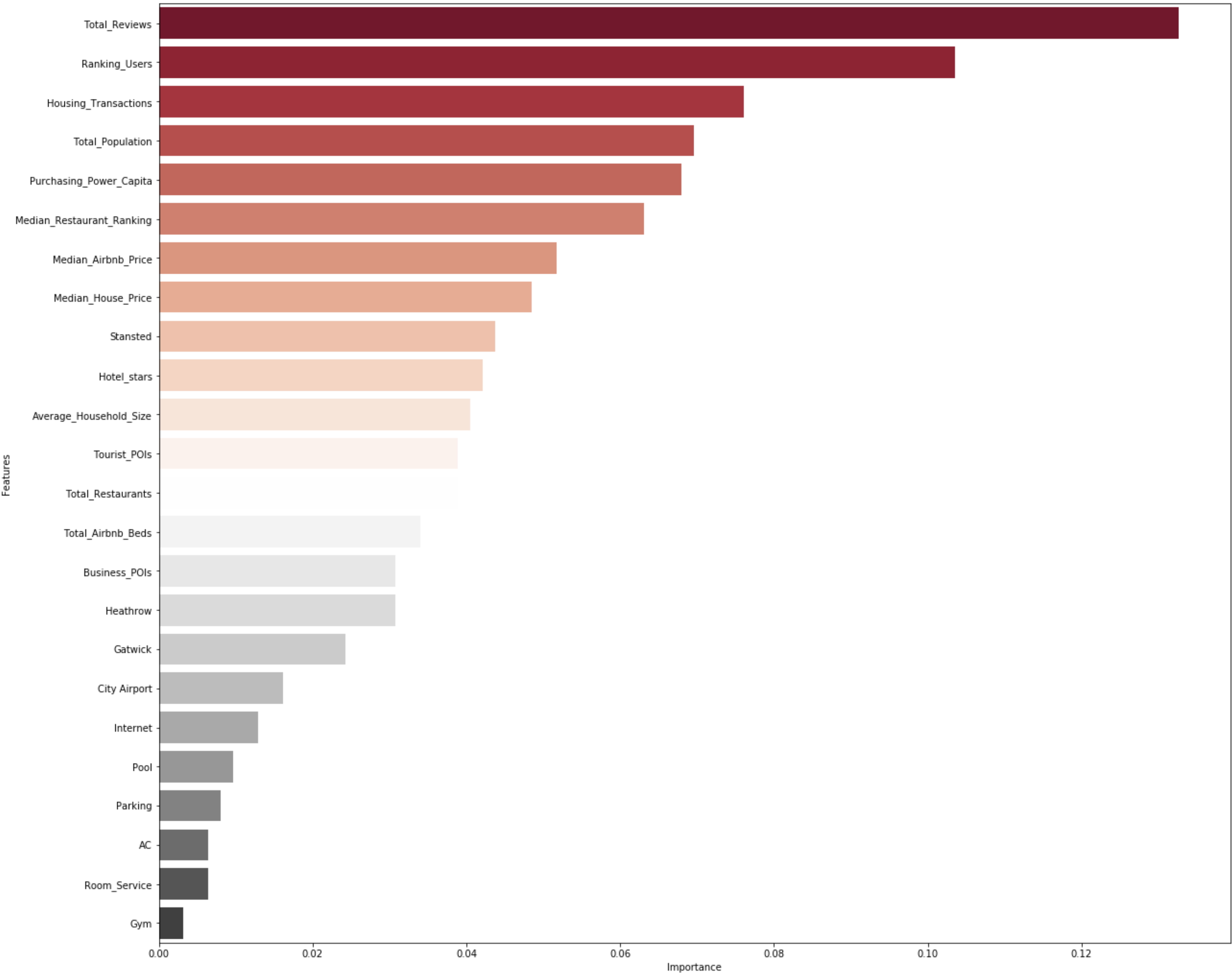
For that we will use the feature engineered dataset, which has been treated to avoid multicollinearity as well as skewness. We split the dataset in a 70-30 for cross validation.

Our main goal is not to get the greatest accuracy but to be able to create a model reproducible via an ArcGIS Python Toolbox.

After running several models we conclude that XGBoost is the one that better predicts the hotel room prices so this is the one discussed here.

XGBoost

ASSESSING FEATURE IMPORTANCES AND MODEL FIT



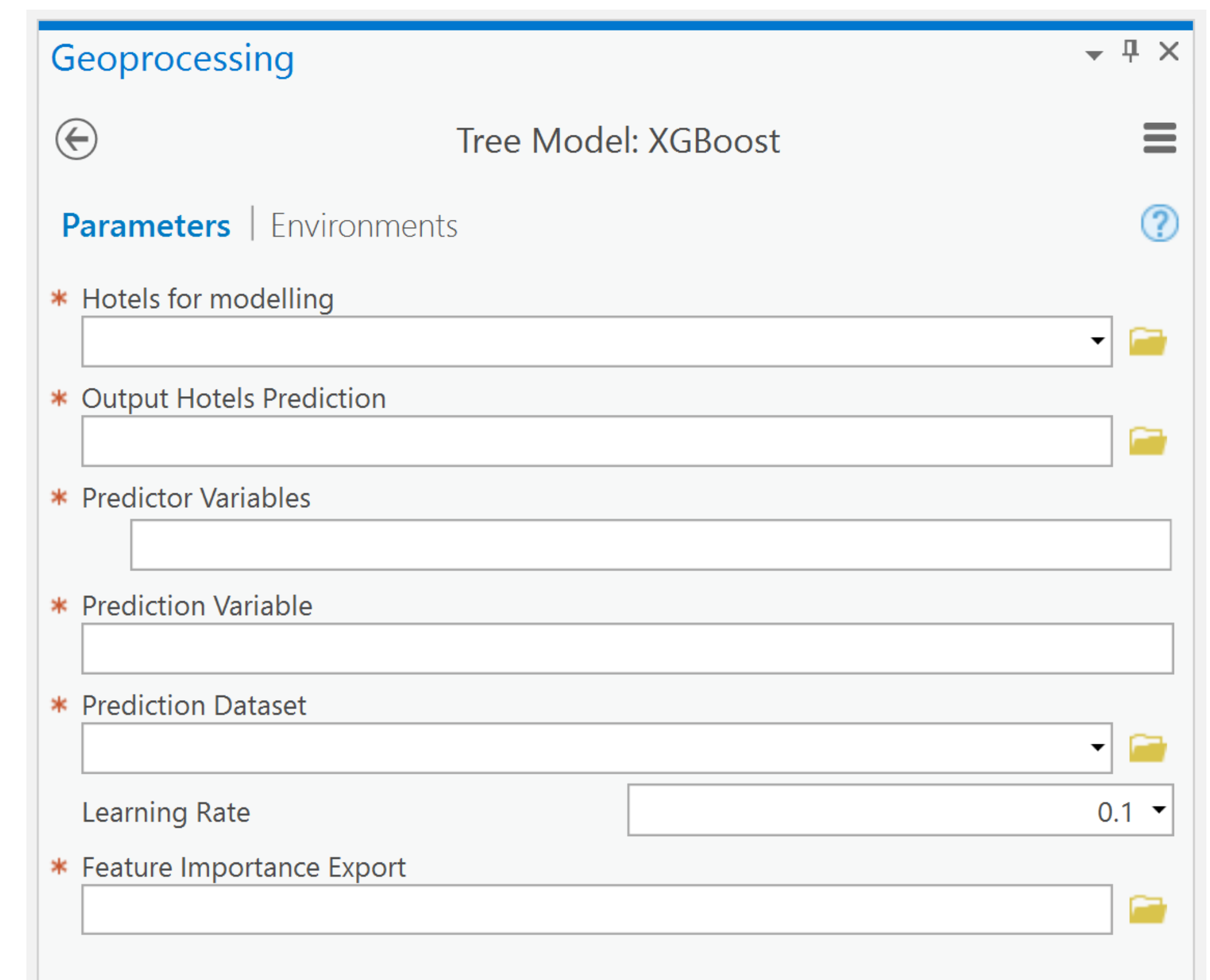
4 – Tool deployment (Python Toolbox)

Once we have validated and tested different ML models we will implement it into an easy to use GUI for non expert users to use. For this purpose we will use ArcGIS Pro and ArcPy to create a Python Toolbox

Python toolbox model parameters

These toolboxes can be distributed, user only need to indicate the input, output and other parameters of the model and the ML will be implemented on the desired data.

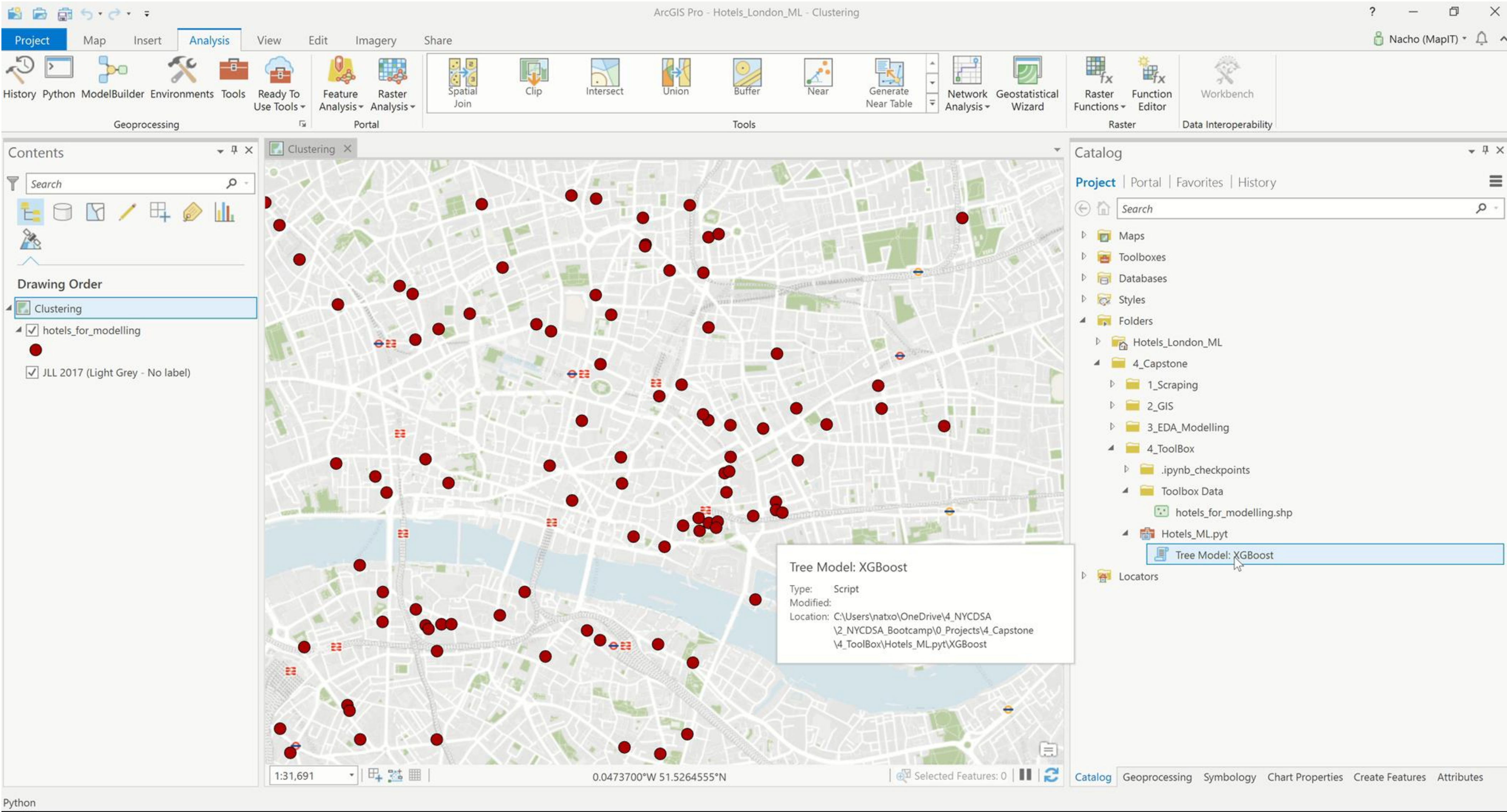
1. Dataset we will use to fit the model
2. Output location
3. Features we will be using for prediction (x)
4. Features to be predicted (y)
5. Data on which we want to use our fitted model
6. ML parameters, in this case the Tree learning rate
7. Text document with the feature importance of the model



The screenshot shows the 'Geoprocessing' window in ArcGIS Pro. The title bar says 'Geoprocessing'. Below the title bar, there is a back arrow icon and the text 'Tree Model: XGBoost'. To the right of this text is a hamburger menu icon. Below this, there are two tabs: 'Parameters' (selected) and 'Environments'. To the right of the tabs is a help icon (a question mark in a circle). The 'Parameters' tab contains several parameter fields, each preceded by a red asterisk (*):

- 'Hotels for modelling': A text box with a dropdown arrow and a folder icon to its right.
- 'Output Hotels Prediction': A text box with a folder icon to its right.
- 'Predictor Variables': A text box.
- 'Prediction Variable': A text box.
- 'Prediction Dataset': A text box with a dropdown arrow and a folder icon to its right.
- 'Learning Rate': A text box with the value '0.1' and a dropdown arrow to its right.
- 'Feature Importance Export': A text box with a folder icon to its right.

Python tool demo



SUMMARY

- Main objective of this capstone project was to develop a ML algorithm that could be easily reproducible by non technical colleagues.
- Gathering and cleaning the data was (as expected) one of the most time consuming task. In this particular project the diversity of datasets and sources used added extra complexity.
- The use of different technologies and software (proprietary and open source) it's usually the best approach since you can leverage the flexibility of open source software with the data and workflows of proprietary software.
- In the future we would like to develop a set of tools that will allow my colleagues and **clients** to put in practice ML algorithms