Data Scientist certification

Author: Dmitry Moskalev

Table of contents

Input data

Column name	Details
id	Numeric, the unique identification number of the property
latitude	Numeric, the latitude of the property
longitude	Numeric, the longitude of the property
property_type	Character, the type of property (e.g., apartment, house, etc)
room_type	Character, the type of room (e.g., private room, entire home, etc)
bathrooms	Numeric, the number of bathrooms
bedrooms	Numeric, the number of bedrooms
minimum_nights	Numeric, the minimum number of nights someone can book
price	Character, the dollars per night charged

Inn the Neighborhood company wants to avoid estimating prices that are more than 25 dollars off of the actual price, as this may discourage people.

Input data contains in “*rentals.csv*” file with **9 columns (features)** and **8111 rows.**

Items colors:

- The “*green*” items do not require changes
- The “*yellow*” items require calculations
- The “*orange*” items require changes due to type incompatibility with “Numeric” one

Example of raw (input) data

id	latitude	longitude	property_type	room_type	bathrooms	bedrooms	minimum_nights	price
958	37.76931	-122.43386	Apartment	Entire home/apt	1	1	1	$170.00

Example of data for processing

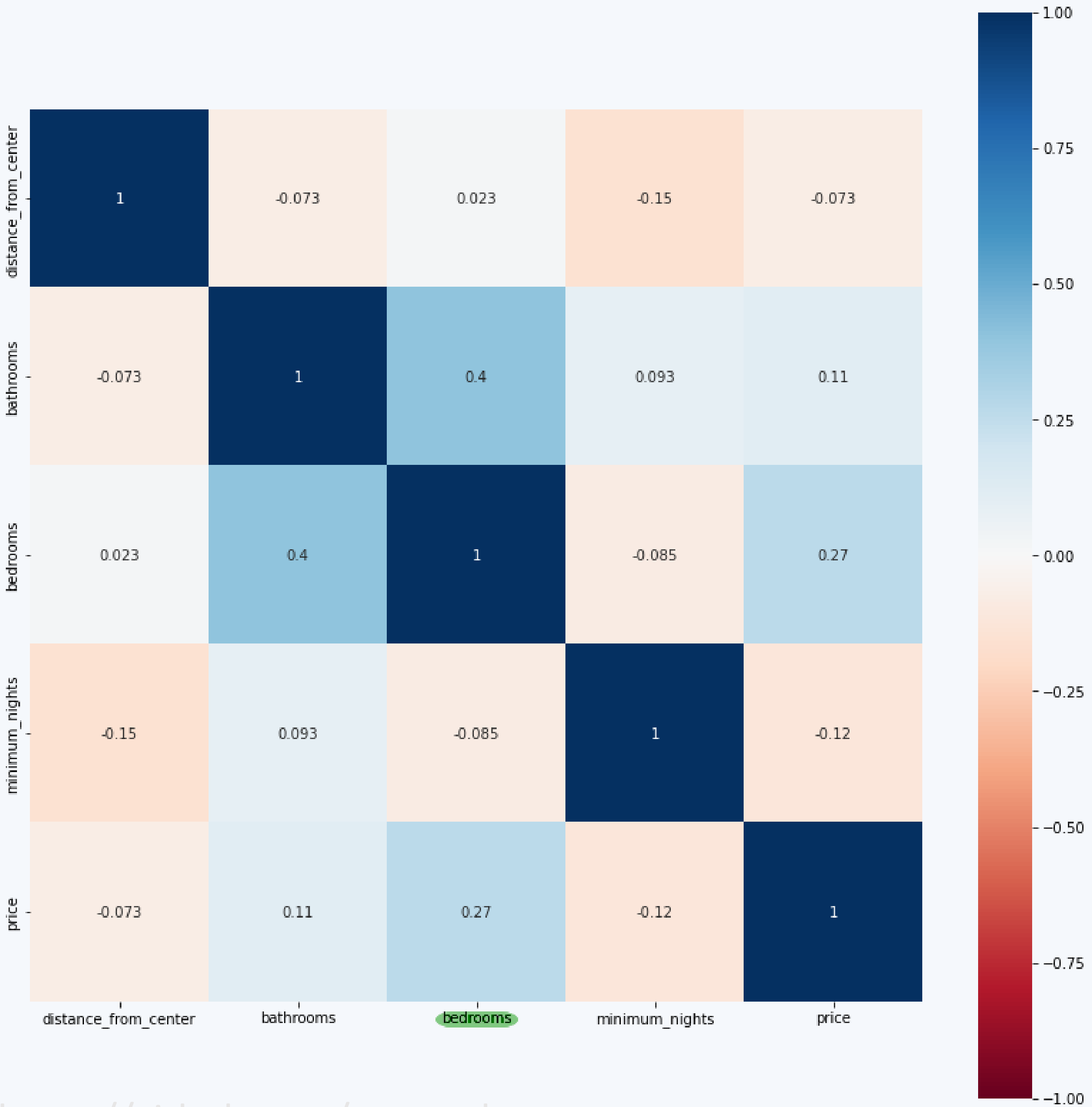
distance_from_center	property_type	room_type	bathrooms	bedrooms	minimum_nights	price
2.45	1	1	1	1	1	170

Data analysis

Let’s look to the target feature “***Price***” in current report using the Pearson’s correlation matrix.

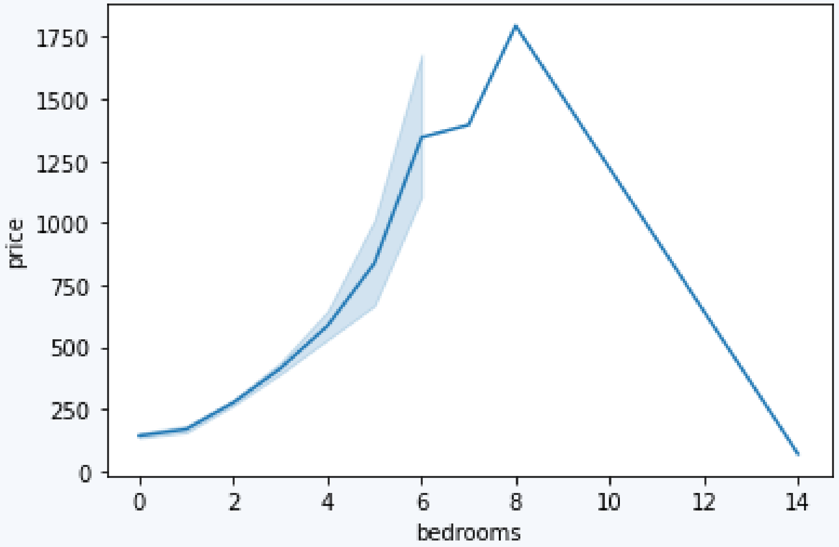
Coefficients with value greater that zero indicates direct dependence and values less that zero indicates reverse dependence, zero value means missing of any dependence between features. When dependence is direct the greater X values is, the greater Y values will be. Reverse dependence means for a larger X values match a smaller Y values.

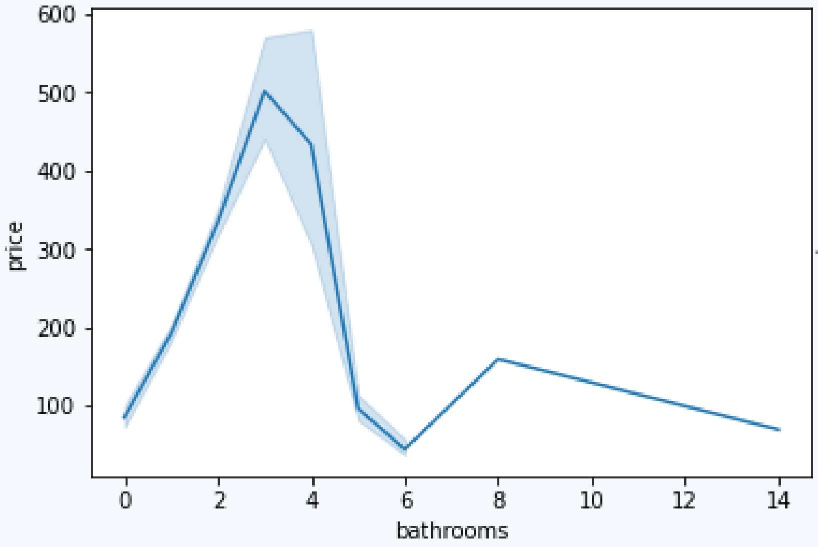
The greatest absolute value for “***price***” correlated with the ***“bedrooms”*** feature. It means, that this column is also target feature, because of making the most feature importance for predicting accommodation price per night.

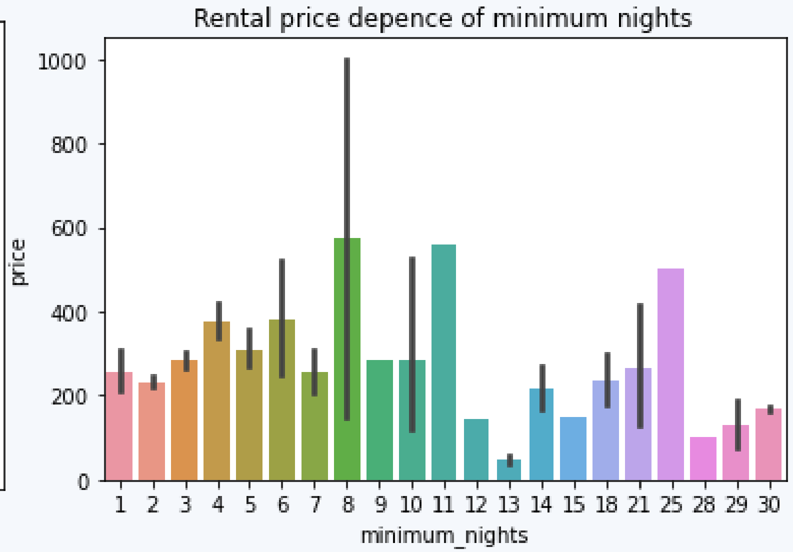


	distance_from_center	bathrooms	bedrooms	minimum_nights	price
distance_from_center	1	-0.073	0.023	-0.15	-0.073
bathrooms	-0.073	1	0.4	0.093	0.11
bedrooms	0.023	0.4	1	-0.085	0.27
minimum_nights	-0.15	0.093	-0.085	1	-0.12
price	-0.073	0.11	0.27	-0.12	1

Data analysis



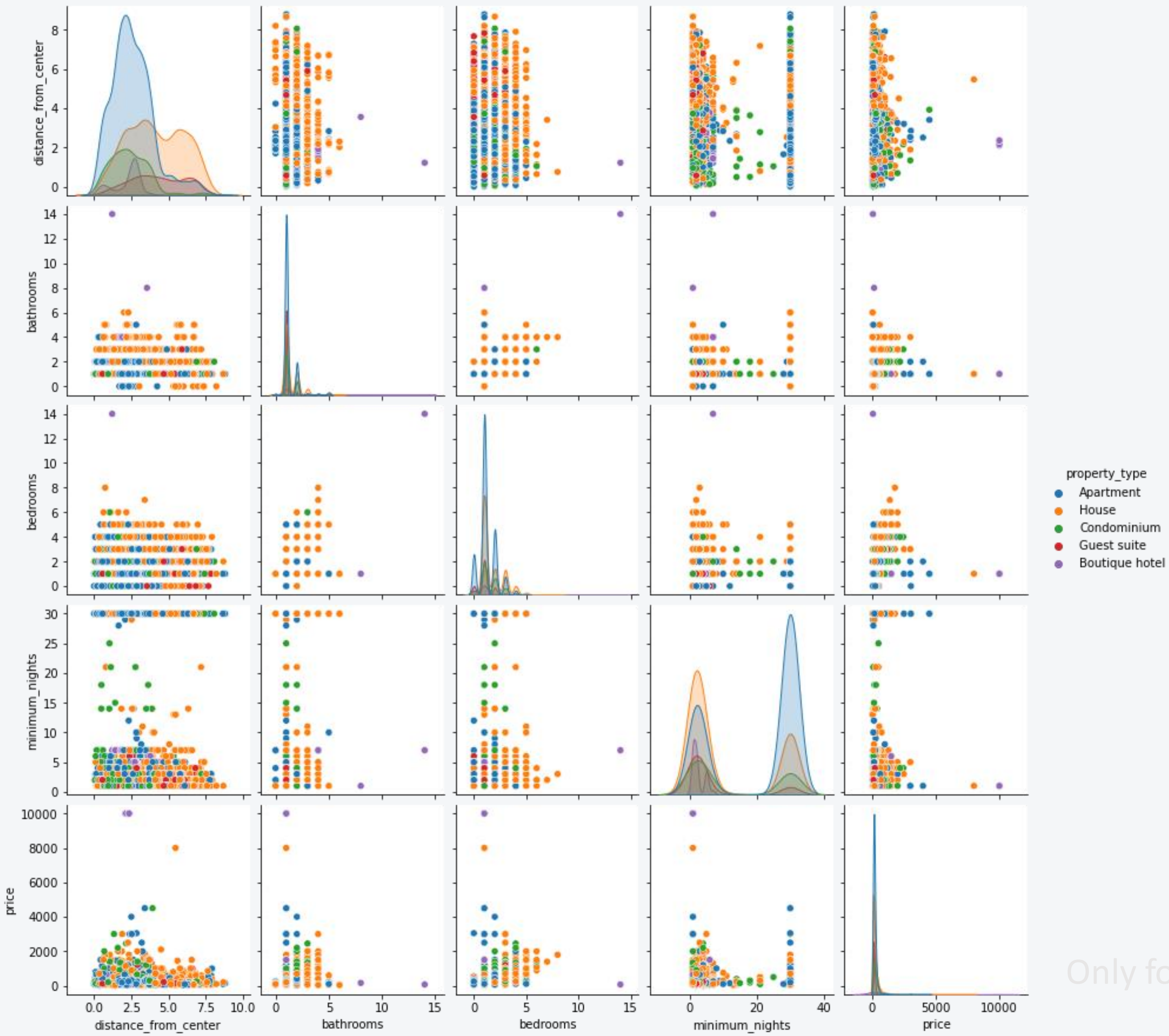


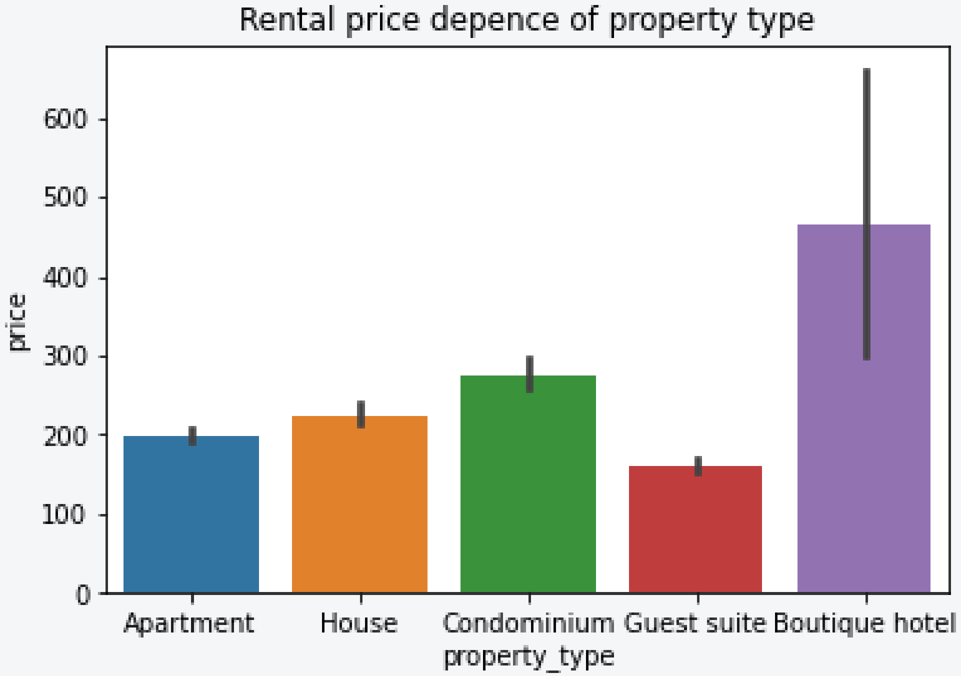

Rental price depence of minimum nights

Some insights of input data:

- Count of bedrooms until 8 raises a price per night and after that there is some accommodation with more cheaper price
- Count of bathrooms until 3 raises a price per night and after that there is some accommodation with more cheaper price
- The price for 8, 11 and 25 nights of rental are most expensive

Data analysis

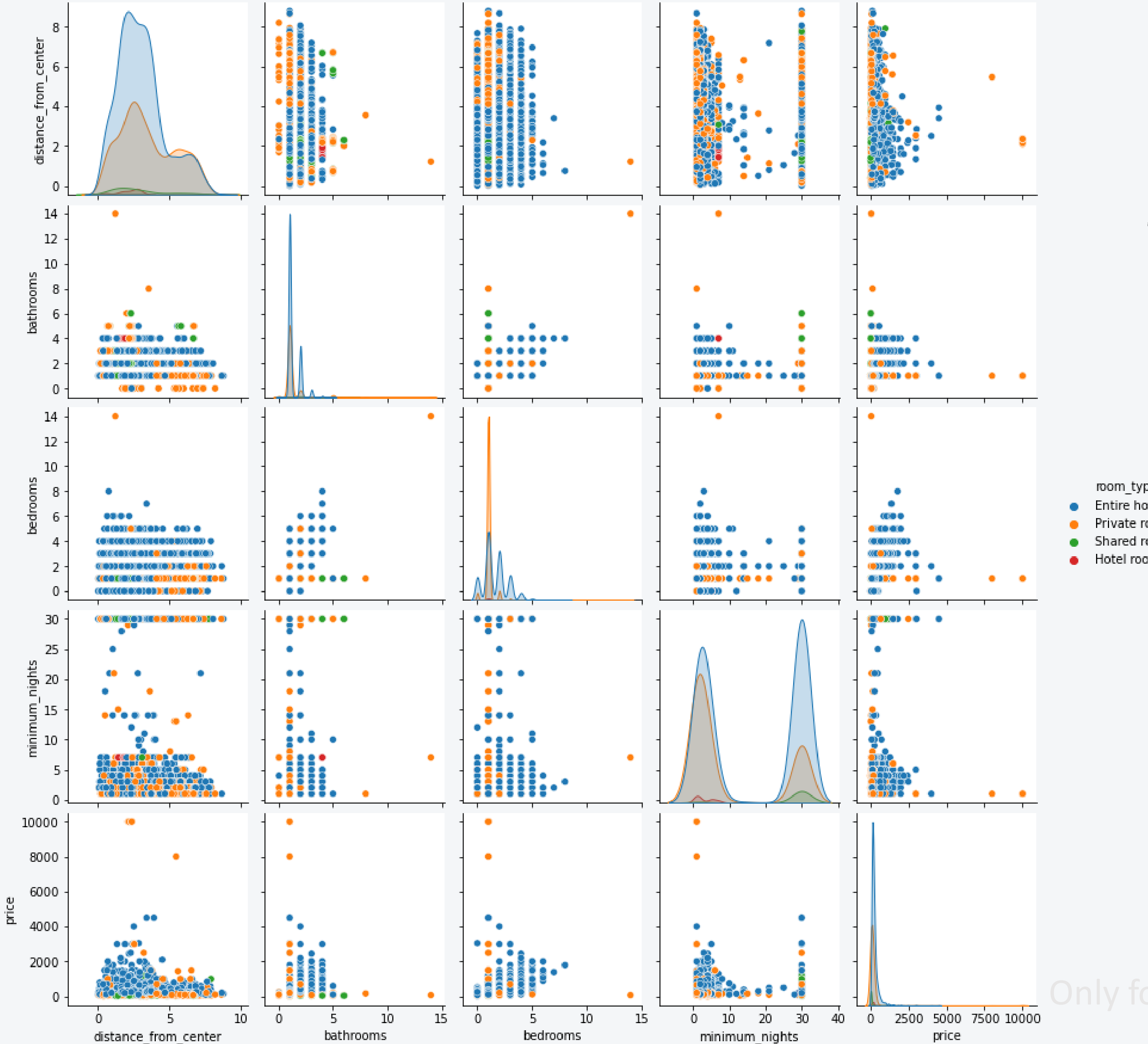


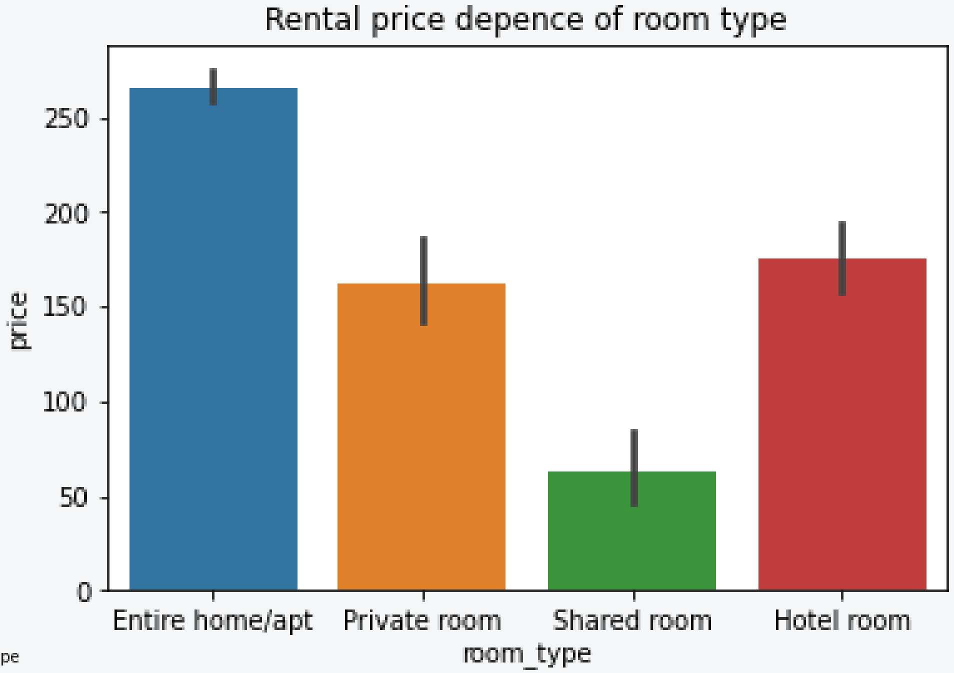


Some insights about property types:

- "Guest suite" is the cheapest in price per night property type and "boutique hotel" is the most expensive
- "Apartment" type has the greatest distance from the center of San Francisco and offers more nights for guests with the highest average price

Data analysis

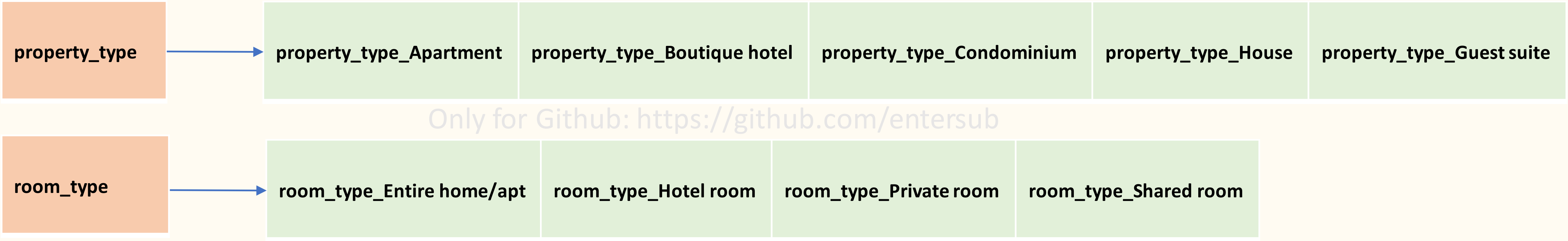


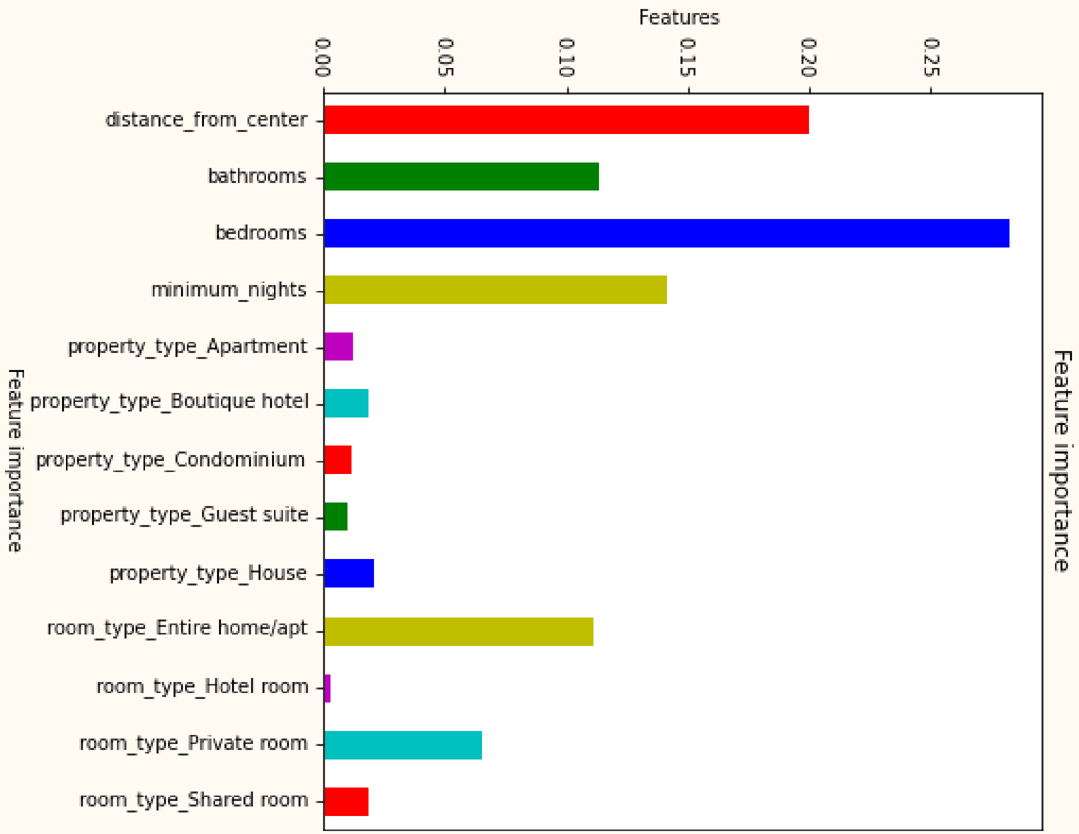

Rental price depence of room type

Some insights about room types:

- “Shared room” is the cheapest in price per night room type and “Entire home/apt” is the most expensive
- “Entire home/apt” type has the greatest distance from the center of San Francisco and offers more nights for guests with the highest average price

Model and application





Evaluate price using Random Forest Regressor as the most suitable method.

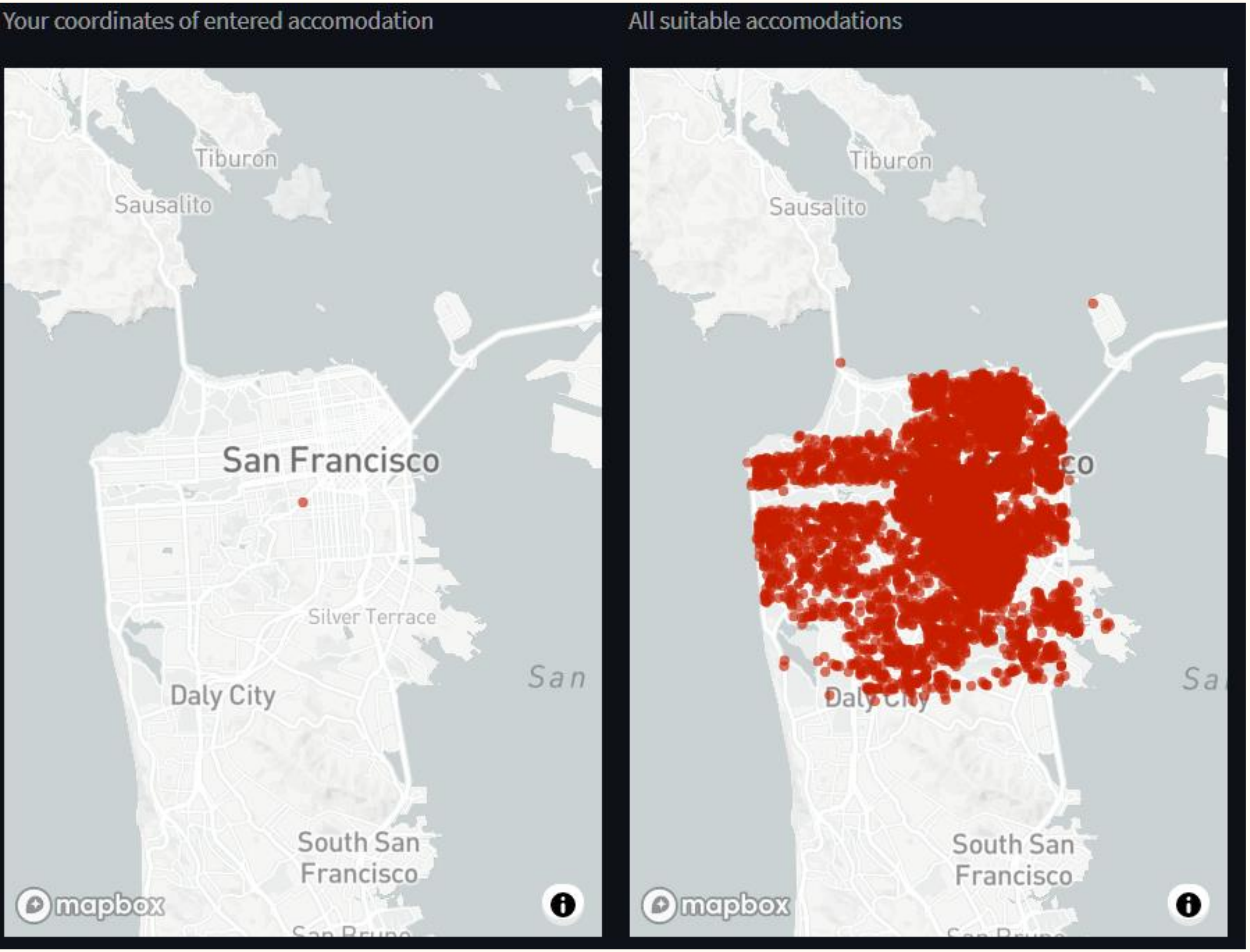
Input data contains in X_train_pr variable with **13 columns (features)** and **5613 rows.**

Items colors:

- The "*green*" items do not require changes
- The "*orange*" items was normalized due to better compability

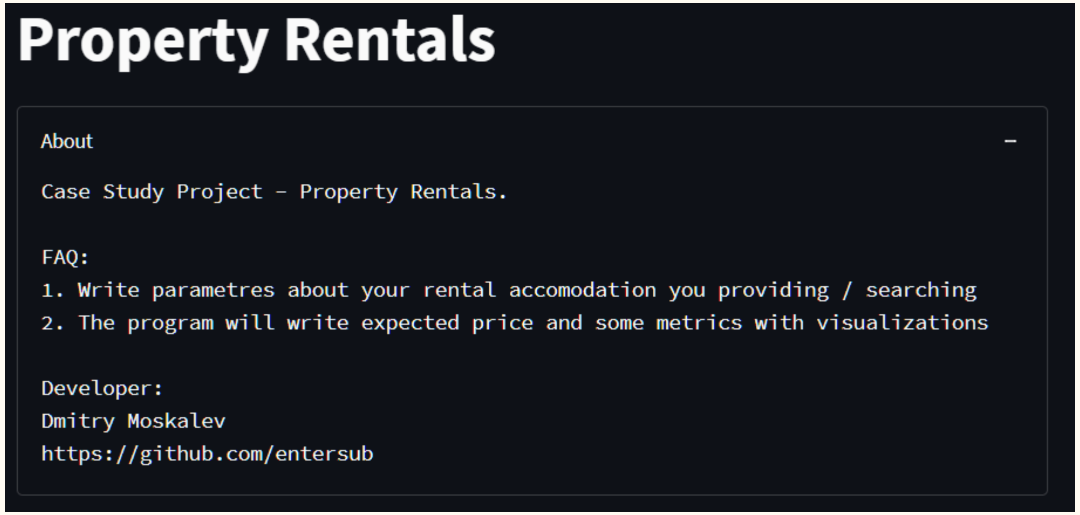
Top features by importance after fitting model and tuning hyperparametres using *RandomizedSearchCV* function: **bedrooms**, **distance_from_center**

Model and application



Application: Property Rentals

Input data for rental accommodation: latitude, longitude, property_type, room_type, bathrooms, bedrooms, minimum_nights, price
Ouput data: Calculated price with visualization and metrics. Pickle model can be also saved and then implemented for better accuracy.
Application has a tutorial and a user-friendly filling form with auto calculating.



Property Rentals

About –

```
Case Study Project - Property Rentals.

FAQ:
1. Write parametres about your rental accomodation you providing / searching
2. The program will write expected price and some metrics with visualizations

Developer:
Dmitry Moskalev
https://github.com/entersub
```

Model and application

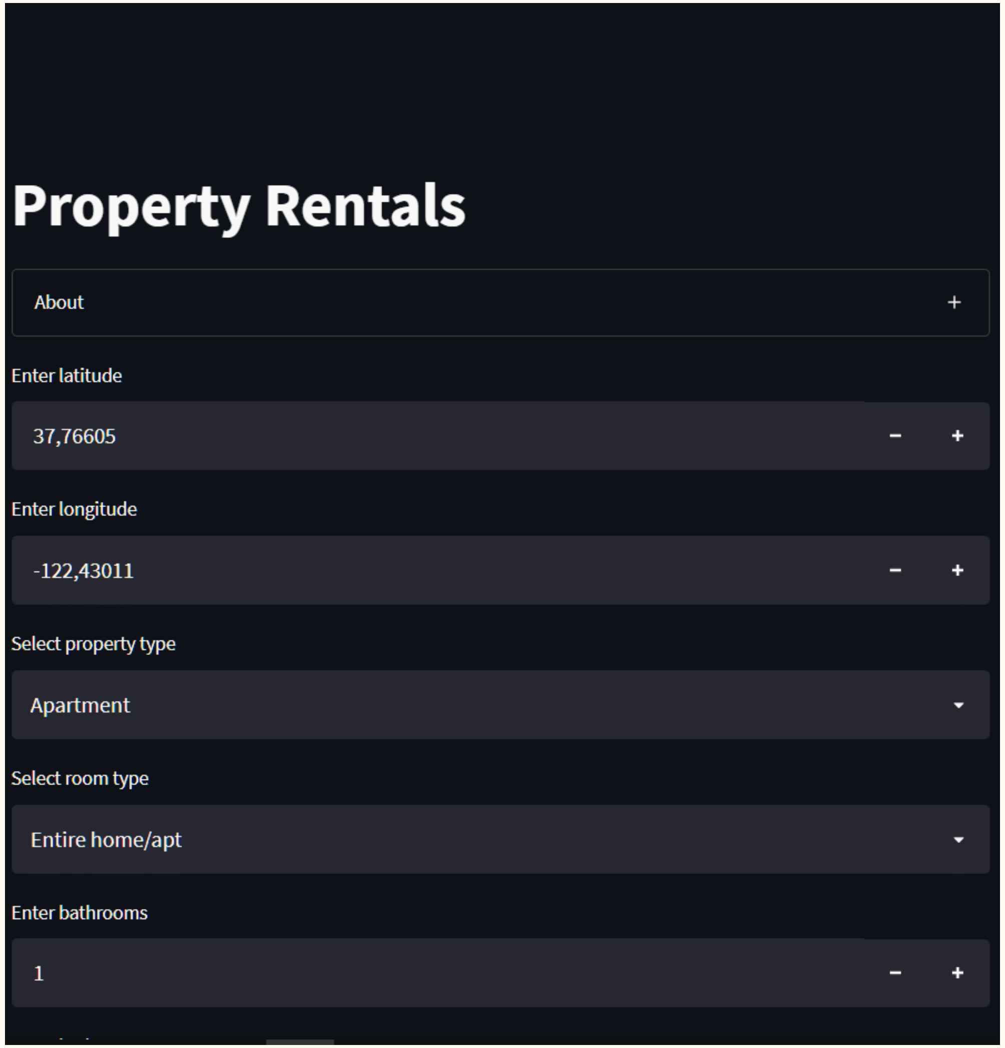
You can watch “Application.gif” demonstration:
https://github.com/EnterSub/Other_Projects/blob/main/certifications/Application.gif

Input price: 270
Calculated price: 274

The user or accommodation owner selected right price that is in a range of 25$.

distance_for_center	bedrooms	Calculated price
0.89	1	274
0.89	2	295
0.89	5	318

Choosing important features: bedrooms, distance_for_center (latitude and longitude)



Property Rentals

Enter latitude
37,76605

Enter longitude
-122,43011

Select property type
Apartment

Select room type
Entire home/apt

Enter bathrooms
1

Results

The main **insights** of calculating prices are:

- The greatest absolute value for "***price***" correlated with the ***"bedrooms"*** feature
- Rental prices for less that a week more often have more expensive price
- Count of bedrooms and bathrooms increase the price for some amount and then during the changing type of room or property theprice can be lower

If you are looking for cheap accommodation price per night:

- Take more than a week rental
- Prefer accommodation not in city center
- Live in shared room in a guest suite

If you can pay expensive price per night for accommodation:

- Live in entire home/apt in a guest suite in boutique hotel
- Choose accommodation in a city center

Outcome:

The current model and application provide calculated price that is almost similar to actual and inform if user typed not similar target value, so it can be a solution to avoid estimating prices that are more than 25 dollars off of the actual price not to discourage people.

Future work:

- Deploy the application on VPS or static site
- Upgrade metrics

Thanks for your attention!