XAI: Model-agnostic methods. Partial Dependency Plot (PDP).

Yassmina Jebbour Maamri
Qilu Diana Wu
(Group 21)
3º GCD 2

CONTENT INDEX.

FIGURES INDEX.

Introduction.

The main purpose of this report is to analyze two model-agnostic methods done with R studio, one for the bike rentals dataset and the other one for the housing dataset. On the one hand, we want to study the bike rentals according to certain variables such as temperature and humidity. And, on the other hand, we want to study the housing price according to some variables such as the number of bedrooms and the number of bathrooms.

Access to the GitHub repository that stores all updates can be done through this link: ddidiadiandiana/XAI3 (github.com)

One dimensional Partial Dependence Plot.

Firstly, we will analyze the relationship between some specific variables and the predicted bike rentals. This will be done in order to see how the change in the value of this features could affect the bike rentals in a day. Those variables are: “days_since_2011”, “temp”, “hum” and “windspeed”. In *Figure 1*, we begin with the first variable, the days since 2011, and we can see that it has several trends depending on the section:

- Start (0 to 100 days): At the beginning of the graph (days 0 to 100), the number of rentals predicted is relatively low, around 3000. This could correspond to the winter months, when the demand for bicycles is lower due to weather conditions.
- Increase (100 to 200 days): There is a notable increase in the rentals predicted around day 100, reaching a value close to 4000 rentals. This could coincide with the arrival of spring, when the weather improves, and more people start using bicycles.
- Stabilization (200 to 400 days): After the initial increase, the number of rentals remains relatively stable around 4000 rentals during this period. This suggests a constant demand for bicycles during this time, probably due to favorable weather conditions and stable usage patterns.
- Second increase (400 to 600 days): Another significant increase is observed in the predicted rents, which reaches a maximum of more than 5000 rents around the day 600. This period could represent the summer of the second year, when the demand for bicycles is usually higher.

Decrease (after 600 days): At the end of the chart, there is a decrease in the number of predicted rentals, falling back below 5000 rentals. This could be related to the arrival of autumn or winter, when demand tends to decline.

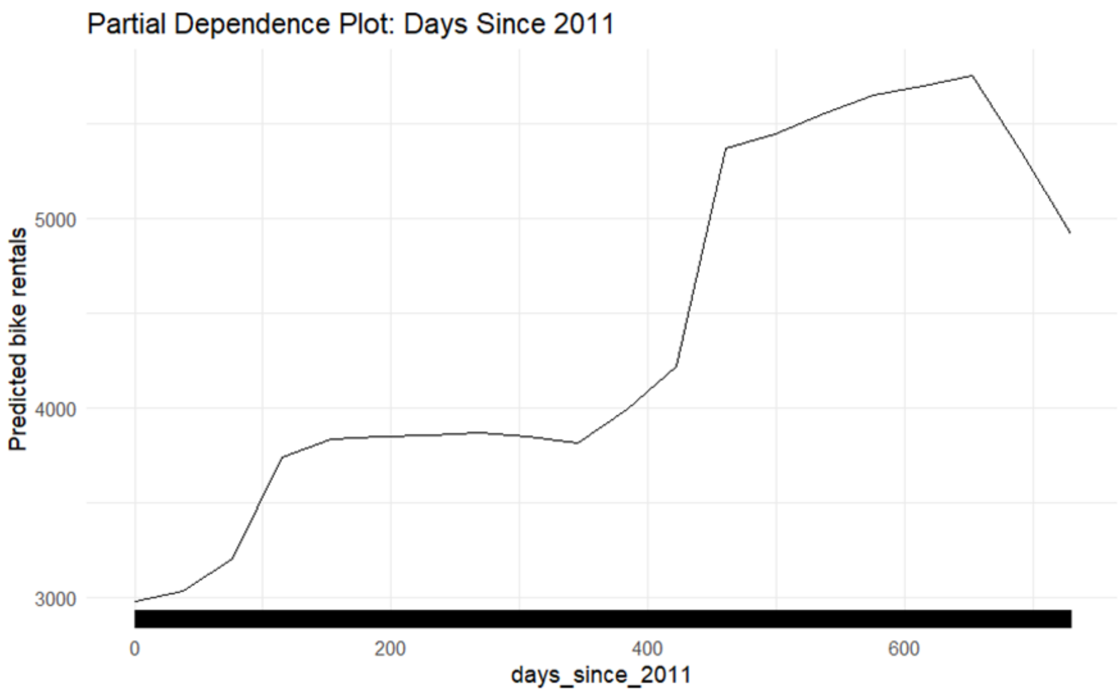

Figure 1. Partial Dependence Plot for the variable "days_since_2011" versus the predicted bike rentals.

Next, in *Figure 2*, for the variable temperature, we can see the next tendencies:

- Low temperatures (up to 10°C): At low temperatures (less than 10°C), the number of rentals predicted is relatively low, ranging from 3000 to 3500 rentals. This is understandable as cold conditions are not ideal for cycling.
- Increase (10°C to 20°C): There is a notable increase in predicted rents as the temperature increases from 10°C to 20°C. This interval shows a steep slope in the curve, reaching about 5000 rents. This temperature range is probably ideal for cycling, comfortable and without excessive heat.
- Stabilization (20°C to 25°C): Between 20°C and 25°C, the number of predicted rentals stabilizes at their peak, around 5000 to 5500 rentals. This indicates that these temperatures are optimal for bicycles.
- Decrease (more than 25°C): At temperatures above 25°C, the number of rentals predicted begins to decrease, falling below 5000 rentals when temperatures approach 30°C and beyond. This suggests that very high temperatures can deter people from renting bikes due to excessive heat.

The predictions for the temperature between 0 and 30ºC can be trusted, but the predicted values for temperatures below 0ºC or above 30ºC should not be credible, due to the low number of instances that were used.

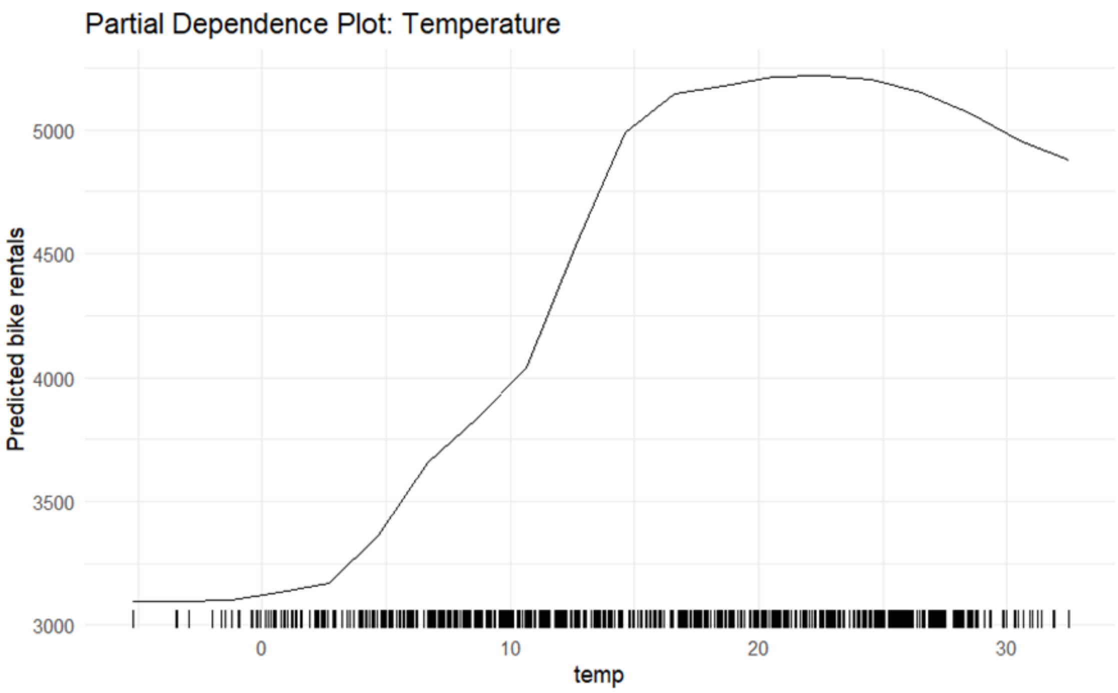

Figure 2. Partial Dependence Plot for the variable "temp" versus the predicted bike rentals.

To continue, in *Figure 3*, with the variable humidity, we can identify the sections explained below:

- Low humidity (0% to 40%): At low humidity levels (0% to 50%), the number of rentals predicted remains fairly stable, around 4750 to 5000 rentals. This suggests that low humidity does not have a significant negative effect on the use of bicycles.
- Moderate humidity (40% to 75%): When reaching moderate humidity levels (50% to 75%), the number of rents begins to decrease slowly. Although the decline is not very steep, it is notable that the amount of rents falls slightly.
- High humidity (more than 75%): From a humidity of 75%, a more pronounced decrease in predicted rents is observed, falling from 4750 to less than 4000 rentals as humidity approaches 100%. This indicates that high humidity levels significantly discourage the use of bicycles.

We should not really rely on the low humidity values, as there are low number of instances that were used to predict them; same as the very high humidity percentages (90%-100%). However, the predicted values for the middle part of the graphic (40%-90%) can be trustful, given the high number of instances used.

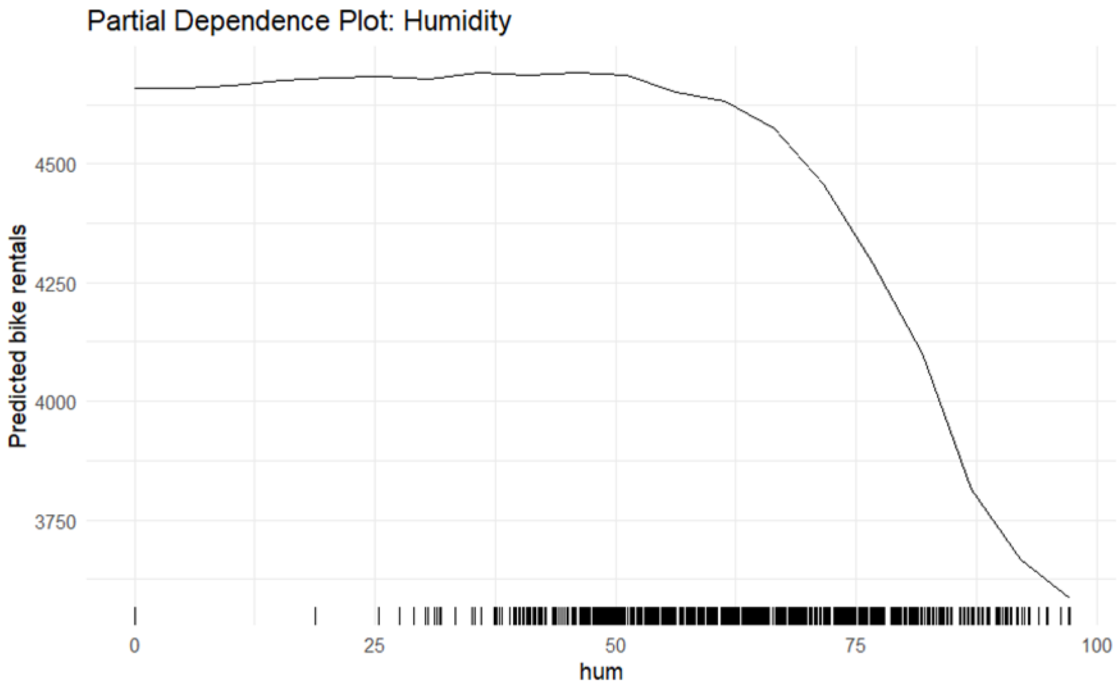

Figure 3. Partial Dependence Plot for the variable "hum" versus the predicted bike rentals.

And, finally, in *Figure 4* with the variable wind speed, we can see the next sections:

- Low wind speed (0 to 10 km/h): At low wind speed levels (0 to 10 km/h), the number of bike rentals predicted remain around 4500 to 4700 rentals. This suggests that low wind speed does not have a significant effect on the use of bicycles and it is even a positive thing, since the values are the highest ones.
- Moderate wind speed (10 to 20 km/h): When reaching moderate wind speed levels (10 to 20 km/h), the number of rents begins to decrease. Although the decline is not very steep, it is notable that the number of rents falls slightly.
- High wind speed (more than 20 km/h): From a wind speed of 20 km, a very pronounced decrease in predicted rents is observed, falling from 4300 to 4000 all at once until 30 km/h and, after that, it remains stable (around 4000 rentals). This indicates that high wind speed levels significantly reduce the use of bicycles.

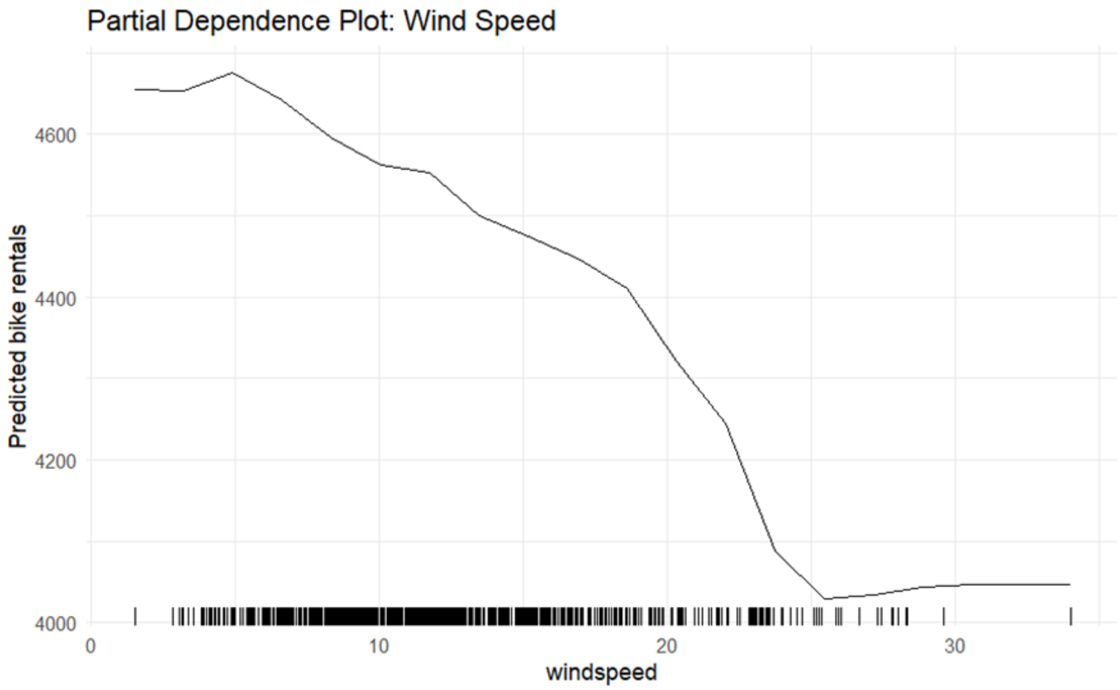

Figure 4. Partial Dependence Plot for the variable "windspeed" versus the predicted bike rentals.

Bidimensional Partial Dependency Plot.

Secondly, with *Figure 5* we want to analyze how the humidity and temperature are related and determine the number of bike rentals. As we can see, when temperature is low, around 0 ºC, and humidity is high, around 75 %, the number of predicted bikes is also too low, around 2000 units (although the density shows that is not a common situation). However, this number begins to increase as the temperature increases, until it reaches a value around 3000 bikes. On the other hand, the values that obtain the highest number of bike rentals are 50 % for humidity and 20 ºC for temperature, there we reach 7000 units. With this information it could be concluded that the ideal weather is to have intermediate conditions, not too high and not too low for both, temperature and humidity.

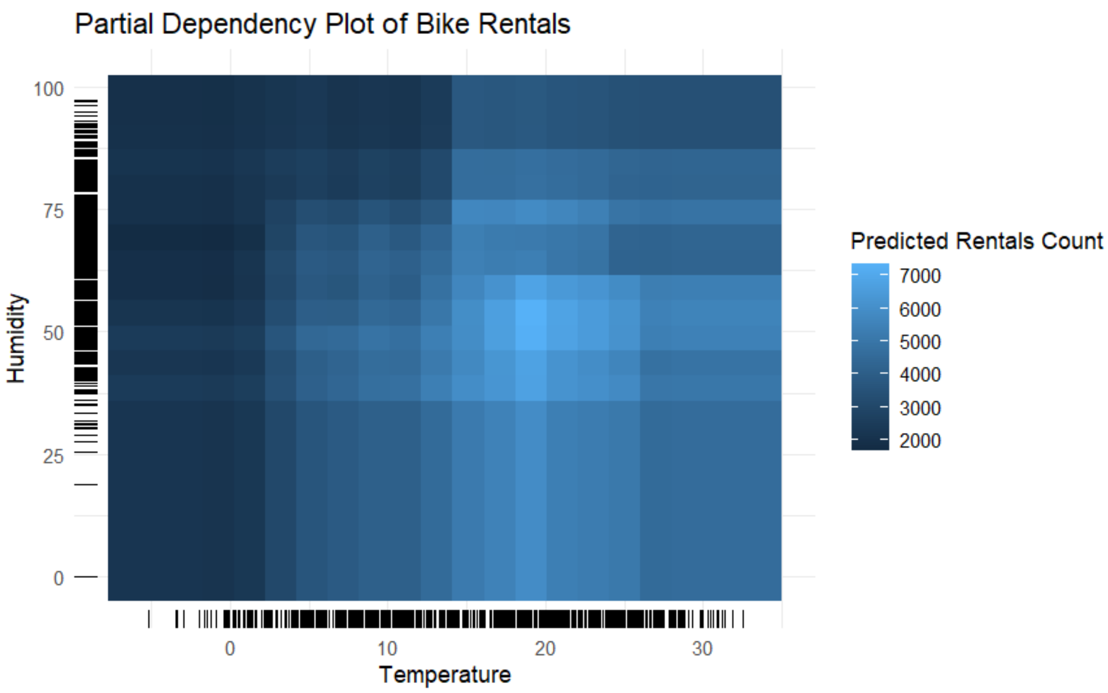

Figure 5. 2D Partial Dependency Plot with humidity and temperature to predict bike rentals.

PDP to explain the price of a house.

Lastly, we will explain the relationships between some specific features and the predicted house price, by approximating with random forest and plotting partial dependence plots. Due to the large size of the dataset, we extracted the 20% of the data randomly and did the plot. We put seeds for reproducibility.

In *Figure 6*, we observe:

- Peak (0 bedrooms): The predicted price of the house there are no bedrooms is higher than 540000; it is also the peak price in our sampled data.
- Decrease (1-3 bedrooms): The predicted price starts to decrease notably, falling to below 536000.
- Increase (3-7 bedrooms): The price starts to rise again, until around 540000.
- Stabilization (more than 8 bedrooms): The price is more stable, around 540000.

The data density is concentrated between 0 and 11 bedrooms, which suggests that the explanations we gave to between these number of bedrooms are reliable. It seems that there are also quite a few instances for houses with around 33 bedrooms, wo this may be reliable as well.

Conversely, no instances can be seen from the plot between 11 and 33 bedrooms, so we cannot confirm anything from there.

The fall of the price can be related to the fact that houses with between 1 and 7 bedrooms are the most demanded. This can be also seen in the density of the data (even though we sampled randomly).

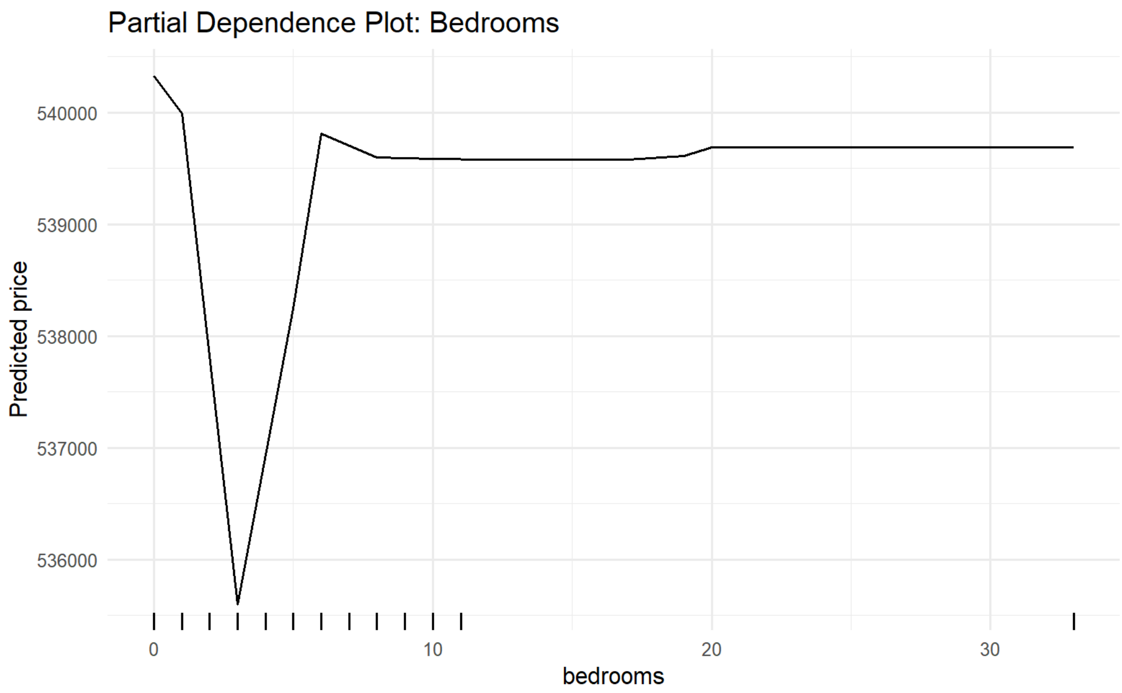

Figure 6. Partial Dependence Plot for the variable "bedrooms" versus the predicted house price.

As for *Figure 7*, we visualize these patterns:

- Stabilization (0-2 bathrooms): The predicted price between 0 and 2 bathrooms is quite stable, around 530000.
- Increase (2-7 batrooms): The price starts to increase and relies between 530000 and 580000.
- Even higher increase (more than 7 bathrooms): there is a drastic increase from 7 bathrooms, where the price increases until around 660000.

This plot indicates that the price increases with the additional number of bathrooms, although we cannot confirm that the approximate price for the 7 or more bathrooms cannot be reliable enough, as they are represented with few instances.

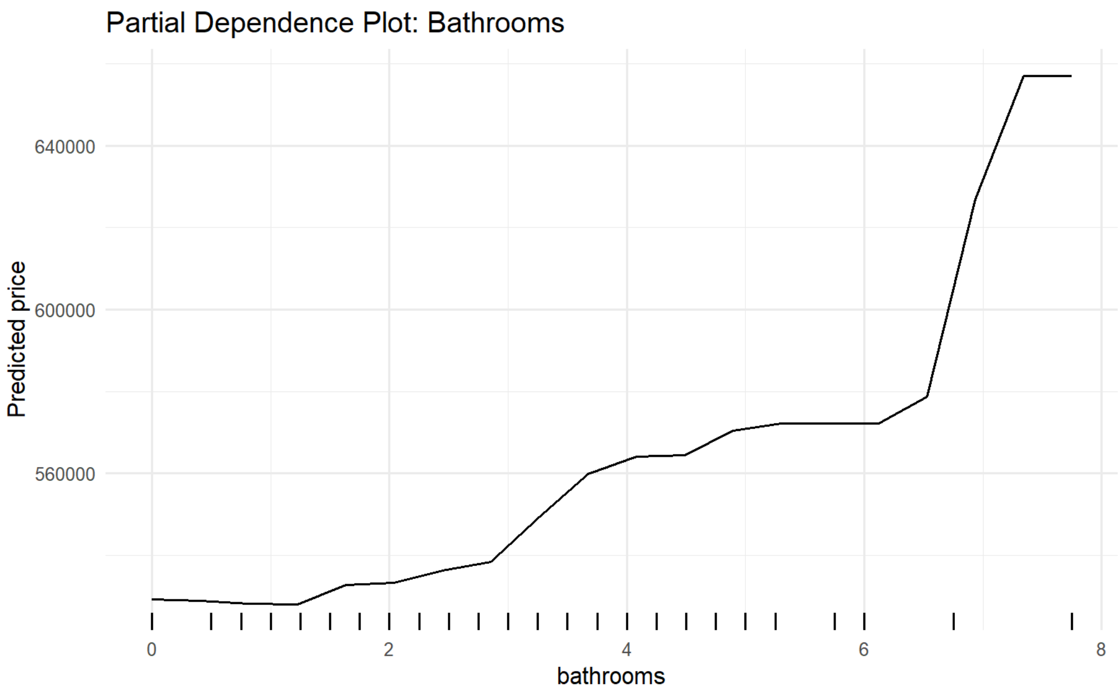

Figure 7. Partial Dependence Plot for the variable "bathrooms" versus the predicted house price.

In *Figure 8*, the following can be observed:

- Constant increase: From around 0 to 10000 square feet houses, the house price keeps increasing, suggesting a positive direct relationship between the living area size and its price (just as the number of bathrooms).

The high concentration of instances between values 0 and 5000 indicates that the increase we have explained is reliable enough. However, the living area size that goes 5000 onwards, especially between 7500 and 10000 square feet, has very few instances for us to trust the predicted price.

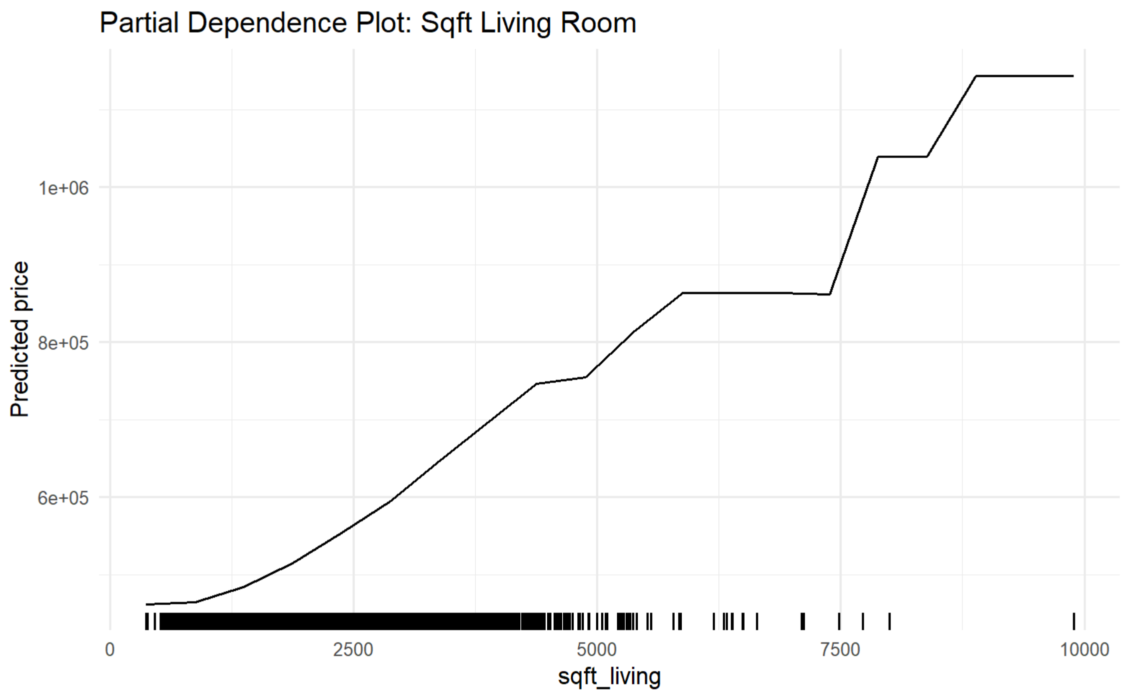

Figure 8. Partial Dependence Plot for the variable "sqft_living" versus the predicted house price.

In *Figure 9*, we can see the next situations:

- Peak (1 floor): The predicted price of the house when there is only 1 floor is higher than 537000, which seems a pretty intermediate price in this context.
- Decrease (1.5 floors): The predicted price starts to decrease notably after 1 floor, falling to below 536000.
- Increase (2 floors or more): The price starts to rise again when there are 2 floors or more, until above 540000. This suggests that having a house of 2 floors or more increases the price notably.

Regarding the density, it seems to be very well distributed as it is not concentrated in a specific side of the graph, so it could be said that the information given is reliable.

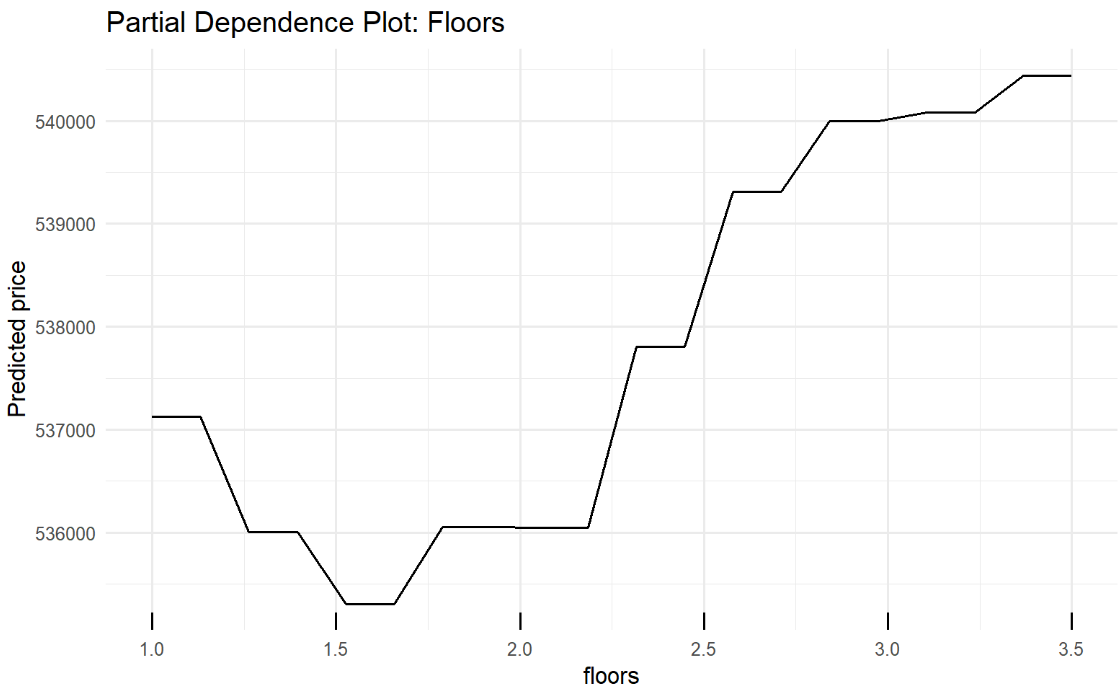

Figure 9. Partial Dependence Plot for the variable "floors" versus the predicted house price.

Conclusion.

In conclusion, bike rental analysis reveals that demand is seasonal and is significantly influenced by temperature and humidity. Moderate temperatures (15°C to 25°C) and low to moderate humidity levels are optimal for bicycle use. In contrast, extreme conditions reduce demand.

Regarding the price of housing, we note that the number of bedrooms has a notable impact. Houses without bedrooms have the highest prices, while the price decreases with 1 to 3 bedrooms and then increases again until it stabilizes with more than 8 bedrooms. These variations reflect market preferences and demand. Something similar happens with the number of floors. In addition, we observed the direct positive relation between the house price and the number bathrooms or living area size.