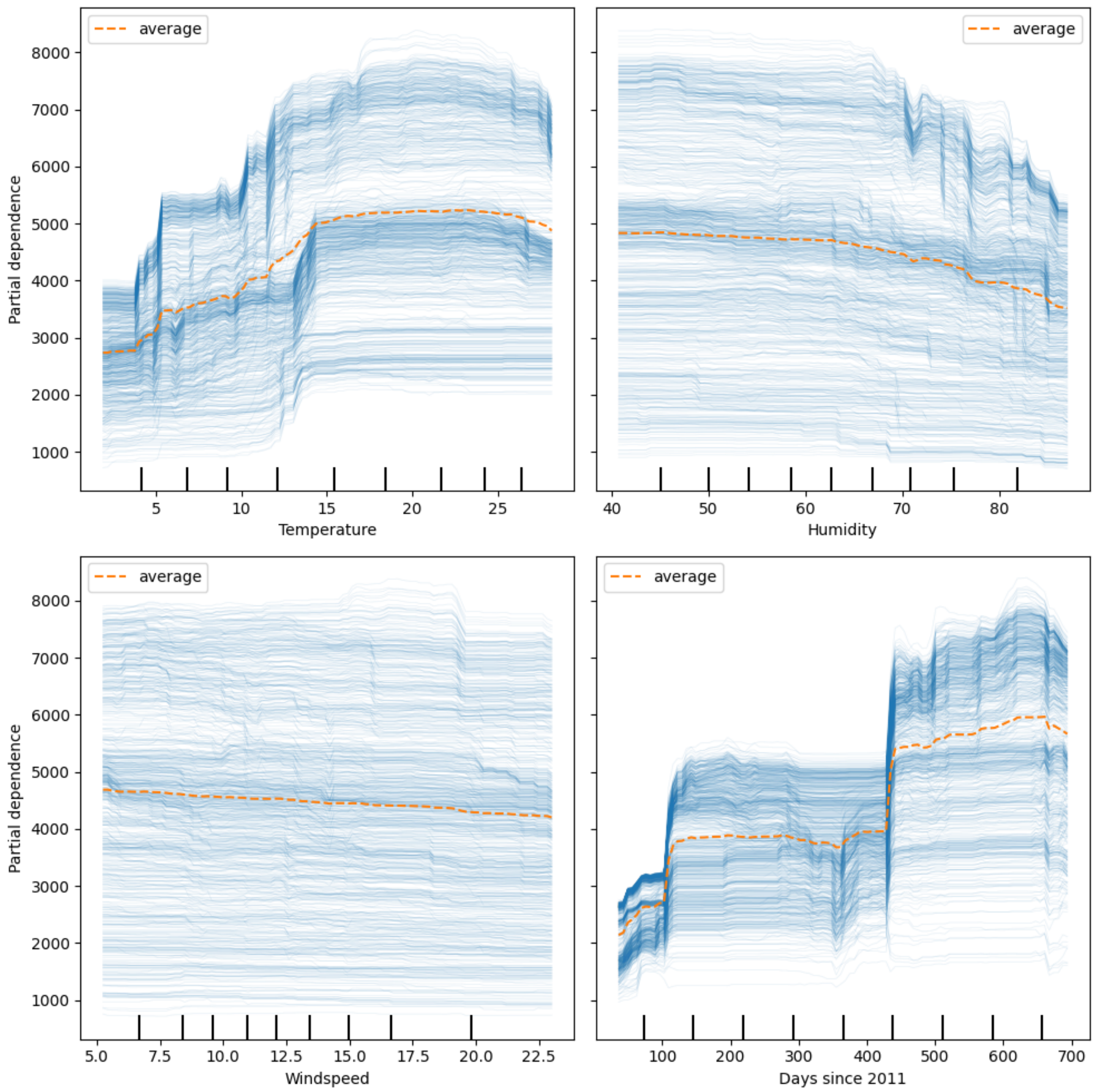
Adam Steciuk
David Borregón
Diana Haj

1.- One dimensional Partial Dependence Plot

In general, Partial Dependence Plots provide a useful tool for understanding how a model uses each feature to make predictions and can help identify important features and interactions. However, it is important to remember that PDPs only show the marginal effect of a feature on the predicted outcome, which could result in some interactions between features being missed and lead to an incomplete understanding of the predicted outcome.

Let's look at the PDP of Random Forest Regressor for features:

- temperature,
- humidity,
- windspeed,
- days since 2011.

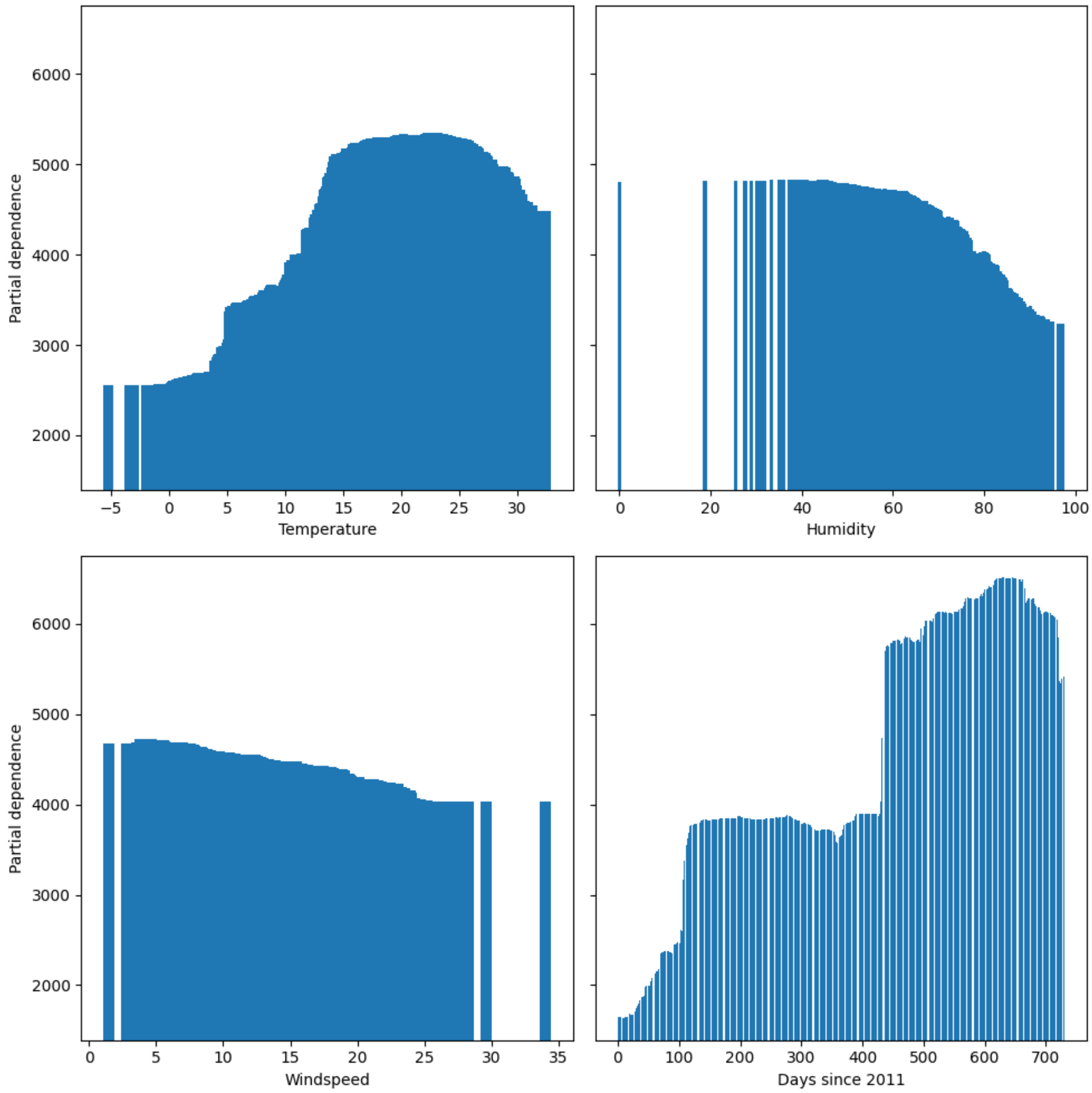


We can see that the temperature increase positively influences the average number of bikes rented until it reaches 15 degrees Celsius. After that, the number of rentals stagnates and seems not to be influenced by the temperature until it starts dropping around the value of 25 degrees Celsius. That kind of relationship seems to be logical and to correctly represent the real-world situation. 15 - 25 degrees Celsius seems to be the optimal temperature for biking and temperatures outside this range may be unpleasant.

When it comes to humidity, we can see that it negatively affects the number of bikes rented. In the beginning very light, but the higher the humidity, the faster the number of bikes rented decreases. That also somewhat makes sense as the range of 40 - 65% humidity doesn't make that much difference to the suitability of conditions for the sport, but higher values may lead to much faster tiredness. Bare in mind that this could be misleading as the influence of humidity for sports performance is also highly tied to the temperature.

Windspeed doesn't seem to influence the number of bikes rented that much. The negative influence is clearly visible, but the slope of the average is rather small.

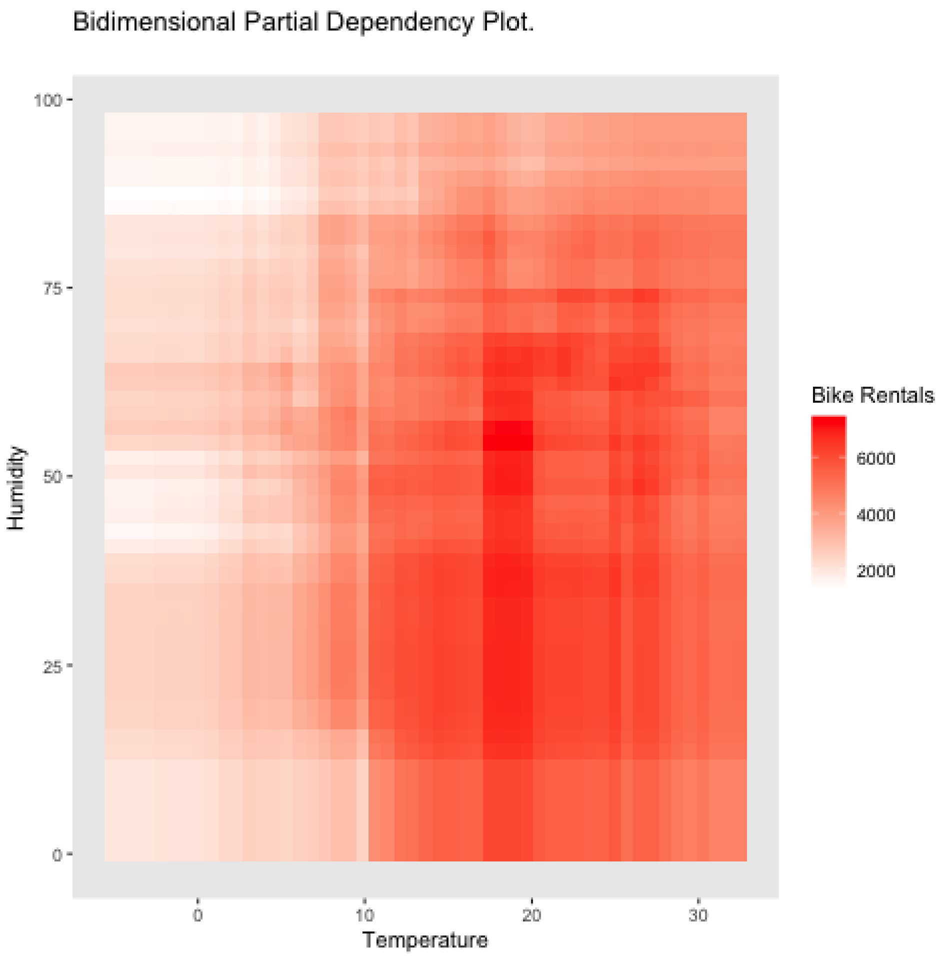
Days since 2011 plot shows the highest amplitude and lack of stability, but also the least number of singular anomalies. The highest increase can be seen around day 430. Those jumps in the number of rentals could be a result of various different factors (e.g. launch of an advertisement campaign, introducing new bike models, etc.).



If we look at the density of samples for different values of the attributes we analyze, we can see that in previous plots, the edges of the value distributions with a small number of samples were automatically cut off from the plots (pay attention to X-axis values on both plot compositions). That can make us more certain that the conclusions about the model we acquired weren't misled by the insufficient number of samples in the data set, and that we weren't inferring any information from the not-well-defined (undersampled) behaviors of the model.

2.- Bidimensional Partial Dependency Plot.

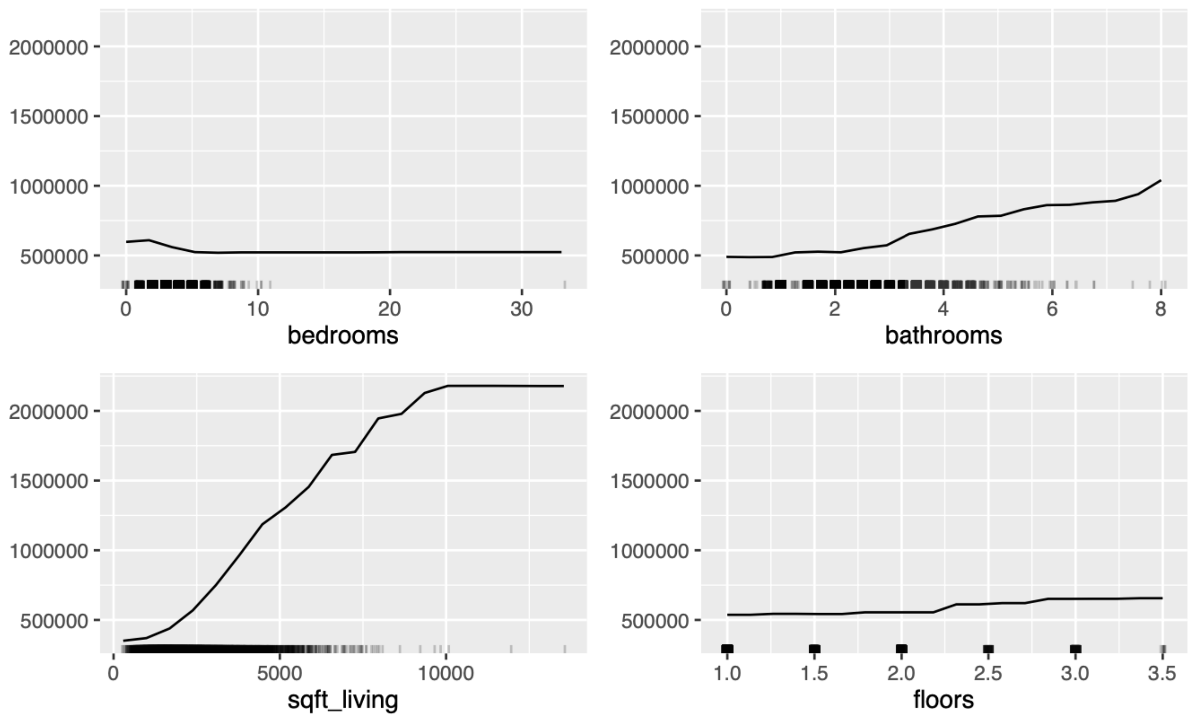
Partial dependence plots (PDPs) depict how a set of input features of interest relate to the target response while ignoring the values of all other input features. Essentially, PDPs reflect the expected target response as a function of the input features of interest. PDPs with two input features demonstrate how these features interact. For instance, the two-variable PDP illustrated in the image below, indicates the correlation between the number of bike rentals and joint values of humidity and temperature. This PDP showcases a clear interaction between the two features; when the temperature exceeds 10 degrees Celsius, humidity predominantly influences the number of bike rentals. Conversely, for lower temperatures, both temperature and humidity have an impact on the number of bike rentals. Humidity above 75% and temperatures above 25 degrees do not have a positive effect over bike rentals. We can observe that the most bike rentals occur around 20 degrees and 50% of humidity. This representation is a very concise and clear way of representing the effect of two different variables over the predicted variable, in our case temperature and humidity over bike rentals.



3.- PDP to explain the price of a house

We'll now apply the previous concepts to predict the **price** of a house from the *kc_house_data.csv* dataset.

Let's look at the PDP of Random Forest Regressor for features:
- bedrooms,
- bathrooms,
- sqft_living,
- floors,
- yr_built.



***Note**: the plot shows there are houses with up to 33 bedrooms, but we are assuming that is an outlier. The rest of the houses oscillate between 1 and 6 bedrooms.*

As it can be seen, the increase in the number of **bedrooms** influences the price of a house positively, but only until it reaches a certain number. Two, to be specific. One reason for the drop in price from 3 bedrooms onward is market demand: if the majority of buyers in a certain area are looking for smaller houses, a house with too many rooms may not be as appealing to them, which can negatively affect the price.

The number of **bathrooms** is another feature that influences the price of a house positively, but this time we don't see a drop in price for any given number of bathrooms, it just continues to grow. This can be due to two things:

- Prestige: in some cases, having more bathrooms can be a symbol of status and prestige. Larger, more luxurious homes often have more bathrooms, so a home with more bathrooms can be seen as more exclusive and desirable.

- More living space: often a home with more bathrooms also has more overall living space, which can increase its value on the real estate market.

The explanation for why the price of a house increases with the number of **square feet** is easy: a larger space translates to a higher price. A few reasons for this:

- Larger built area: the larger the space, the more construction materials are needed for its construction, which implies higher material and labor costs. Therefore, the cost of building a larger space is generally higher than the cost of building a smaller one.

- More comfort: larger spaces often offer more comfort and flexibility in terms of use and design. Plus, they often provide more storage space and activity areas, which can be attractive to many buyers with high purchasing power.

- Prestige: in some cases, having a larger space can be considered a sign of prestige or social status. Therefore, some people may be willing to pay more for a larger space simply because it is seen as a symbol of success.

- Shortage: in some areas, the supply of large spaces may be limited due to the scarcity of land available for construction. In these cases, the limited supply can drive up the price of larger spaces due to high demand.

Finally, we're looking at how the number of **floors** affects the price of a house. As it can be seen, this feature positively influences the price, although in a considerably slight way. We've come up with a couple of reasons for why the price of a multi-story house isn't significantly different from that of a house with a single floor:

- Purchasing power: in certain countries, the average purchasing power isn't high enough to be able to afford a multi-story home. Most people in Spain, for example, live in a single-floor flat. As a result, there isn't a high demand for multi-story houses. If buyers are not willing to pay more for a house with more floors, sellers will have no incentive to increase the price by a lot.

- Location: in some geographic areas, multi-story homes may be more common and therefore not perceived as a differentiating factor that justifies a higher price. In this case, the price of a house would depend more on the location and the specific characteristics of the property, rather than on the number of floors.