# XAI 3: Model-Agnostic methods

## EXERCISE:

Apply PDP to the regression example of predicting bike rentals. Fit a random forest approximation for the prediction of bike rentals (**cnt**). Use the partial dependence plot to visualize the relationships the model learned. Use the slides shown in class as model.

## QUESTION:

Analyse the influence of **days since 2011, temperature, humidity** and **wind speed** on the predicted bike counts.

· Number of bikes renteg grows as temperatures grow until 20 degrees, then it starts declining.

· As we get more far away in time from 2011, we rent more bikes.

· Finally, the number of bikes rented decreases as humodoty and windspeed increase.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plotly)
```

```
## Loading required package: ggplot2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(reshape2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

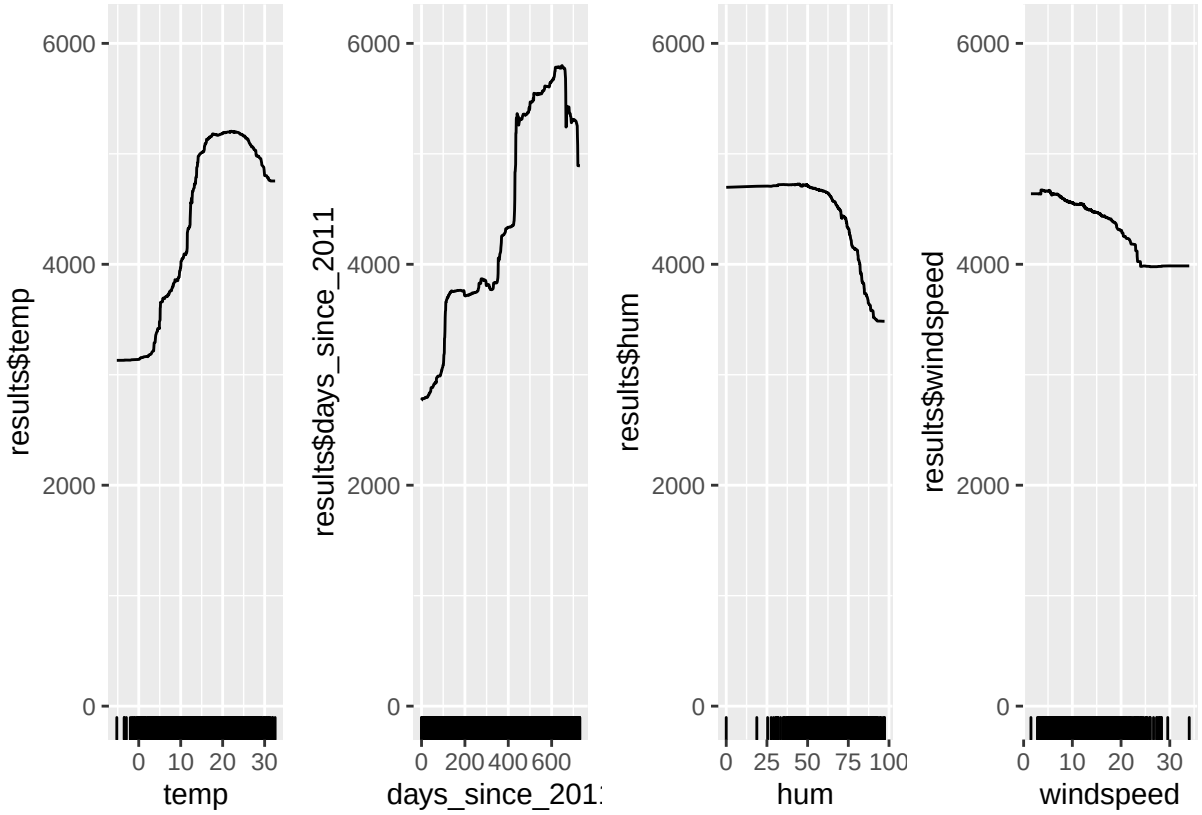
```
library(randomForestSRC)
```

```
##
##  randomForestSRC 3.1.0
##
##  Type rfsrc.news() to see new features, changes, and bug fixes.
##
```

```
#setwd("/Users/cmonserr/OneDrive - UPV/Trabajo_2/Asignaturas/Evaluacion de modelos/Practicas/Practica 3/
days <- read.csv("day.csv")
hour <- read.csv("hour.csv")

days$dteday <- as_date(days$dteday)
days_since <- select(days, workingday, holiday, temp, hum, windspeed, cnt)
days_since$days_since_2011 <- int_length(interval(ymd("2011-01-01"), days$dteday)) / (3600*24)
days_since$SUMMER <- ifelse(days$season == 3, 1, 0)
days_since$FALL <- ifelse(days$season == 4, 1, 0)
days_since$WINTER <- ifelse(days$season == 1, 1, 0)
days_since$MISTY <- ifelse(days$weathersit == 2, 1, 0)
days_since$RAIN <- ifelse(days$weathersit == 3 | days$weathersit == 4, 1, 0)
days_since$temp <- days_since$temp * 47 - 8
days_since$hum <- days_since$hum * 100
days_since$windspeed <- days_since$windspeed * 67

rf <- rfsrc(cnt~., data=days_since)

results <- select(days_since, days_since_2011, temp, hum, windspeed, cnt)
nr <- nrow(days_since)
for(c in names(results)[1:4])
{
  for(i in 1:nr){
    r <- days_since
    r[[c]] <- days_since[[c]][i]
    sal <- predict(rf, r)$predicted
    results[[c]][i] <- sum(sal) / nr
  }
}
```

```
p1 = ggplot(days_since, aes(x=temp,y=results$temp)) + geom_line() + ylim(0, 6050) + geom_rug(mapping = a
p2 = ggplot(days_since,aes(x=days_since_2011, y=results$days_since_2011)) + geom_line() + ylim(0, 6050)
p3 = ggplot(days_since,aes(x=hum, y=results$hum)) + geom_line() + ylim(0, 6050) + geom_rug(mapping = aes
p4 = ggplot(days_since,aes(x=windspeed, y=results$windspeed)) + geom_line() + ylim(0, 6050) + geom_rug(m

library(patchwork)
p1 + p2+ p3 + p4 + plot_layout(ncol=4)
```

```
## Warning: Use of ‘days_since$temp‘ is discouraged. Use ‘temp‘ instead.

## Warning: Use of ‘days_since$days_since_2011‘ is discouraged. Use
## ‘days_since_2011‘ instead.

## Warning: Use of ‘days_since$hum‘ is discouraged. Use ‘hum‘ instead.

## Warning: Use of ‘days_since$windspeed‘ is discouraged. Use ‘windspeed‘ instead.
```




## EXERCISE:

Generate a 2D Partial Dependency Plot with humidity and temperature to predict the number of bikes rented depending of those parameters.

BE CAREFUL: due to the size, extract a set of random samples from the BBDD before generating the the data for the Partial Dependency Plot.

Show the density distribution of both input features with the 2D plot as shown in the class slides.

TIP: Use geom_tile() to generate the 2D plot. Set width and height to avoid holes.

## QUESTION:

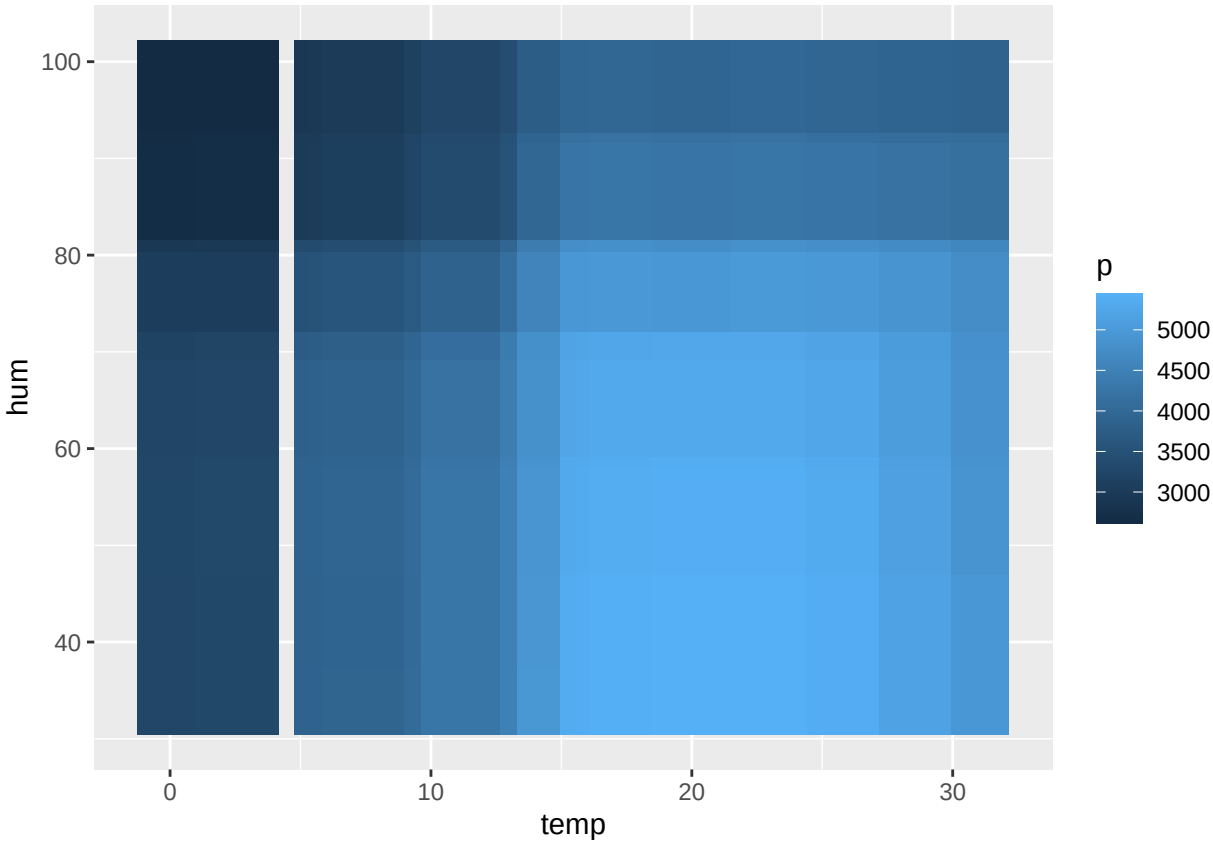
Interpret the results.

Number of bikes rented grows as temperatures grow until 20 degrees, then it starts declining.

As we get more far away in time from 2011, we rent more bikes.

Finally, the number of bikes rented decreases as humidoty and windspeed increase.

```
sampled <- sample_n(days_since, 40)
temp <- sampled$temp
hum <- sampled$hum
th <- inner_join(data.frame(temp),data.frame(hum), by=character())
th$p <- 0

for(i in 1:nrow(th)){
  r <- days_since
  r[["temp"]] <- th[["temp"]][i]
  r[["hum"]] <- th[["hum"]][i]

  sal <- predict(rf, r)$predicted
  th[["p"]][i] <- sum(sal) / nr
}

ggplot(th, aes(x=temp, y=hum, fill=p)) + geom_tile(width=3, height=10)
```




The higher the temperature and humidity, the lower the number of rentals. This is increased if the two variables increase at the same time.

## EXERCISE:

Apply the previous concepts to predict the **price** of a house from the database **kc_house_data.csv**. In this case, use again a random forest approximation for the prediction based on the features **bedrooms**, **bathrooms**, **sqft_living**, **sqft_lot**, **floors** and **yr_built**. Use the partial dependence plot to visualize the relationships the model learned.

BE CAREFUL: due to the size, extract a set of random samples from the BBDD before generating the data for the Partial Dependency Plot.

## QUESTION:

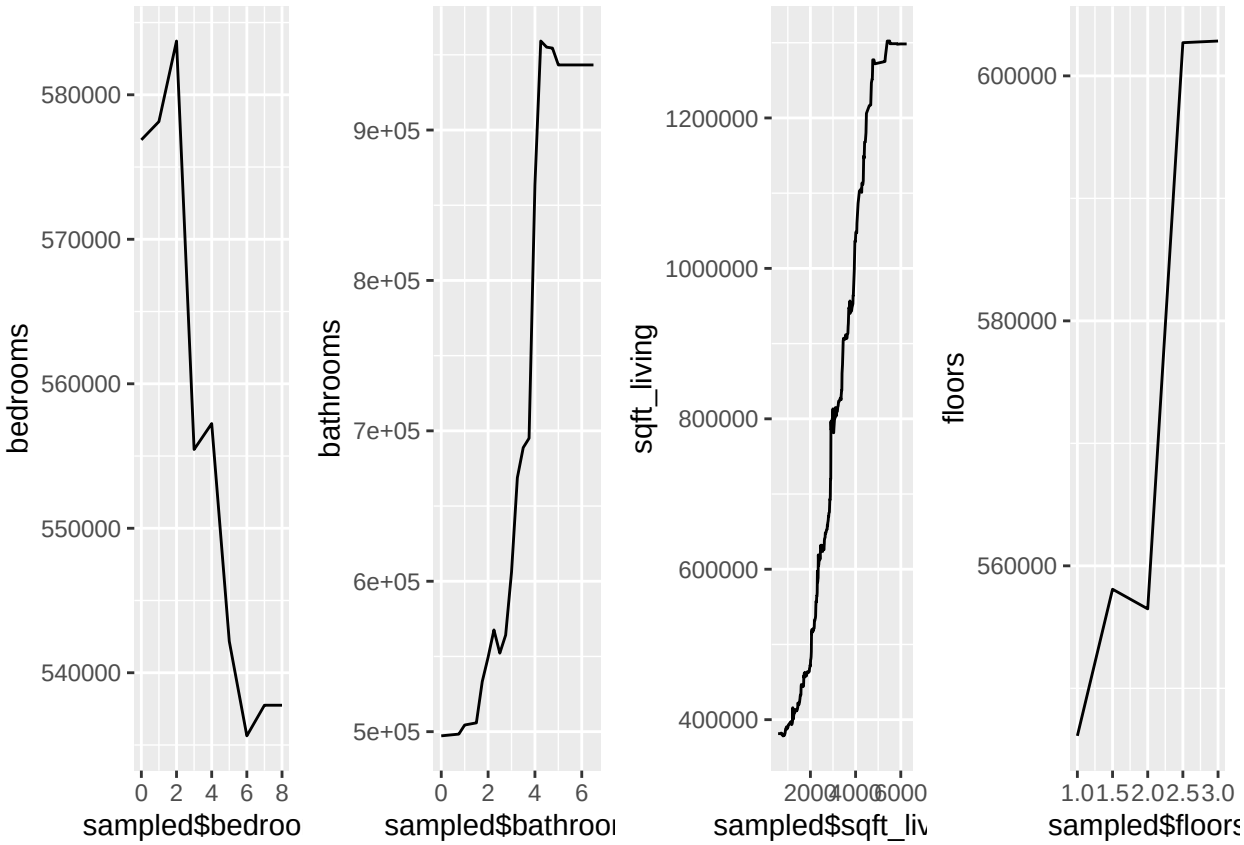
Analyse the influence of **bedrooms, bathrooms, sqft_living** and **floors** on the predicted price.

```
d <- read.csv("kc_house_data.csv")

sampled <- sample_n(d, 1000)

sampled <- select(sampled, bedrooms, bathrooms, sqft_living, sqft_lot, floors, yr_built, price)

rf <- rfsrc(price~., data=sampled)

results <- select(sampled, bedrooms, bathrooms, sqft_living, floors, price)
nr <- nrow(sampled)
for(c in names(results)[1:4])
{
  for(i in 1:nr){
    r <- sampled
    r[[c]] <- sampled[[c]][i]
    sal <- predict(rf, r)$predicted
    results[[c]][i] <- sum(sal) / nr
  }
}

library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
g1 = ggplot(results, aes(x=sampled$bedrooms,y=bedrooms)) + geom_line()
g2 = ggplot(results,aes(x=sampled$bathrooms, y=bathrooms)) + geom_line()
g3 = ggplot(results,aes(x=sampled$sqft_living, y=sqft_living)) + geom_line()
g4 = ggplot(results,aes(x=sampled$floors, y=floors)) + geom_line()

#g1 + g2+ g3 + g4 + plot_layout(ncol=4)
grid.arrange(g1, g2, g3, g4,nrow=1)
```




The number of bathrooms, the square footage and the number of floors all drive up the price of apartments.

The number of beds affects the price differently, as it increases the price if there are too many or too few beds.