Partial Dependence Plots

It is not necessary for a relationship to exist between a dependent and an independent variable and even if it does, it may be INFLUENCED BY OTHER FEATURES like two variables which are working together.
It should be noted that a univariate chart like below-

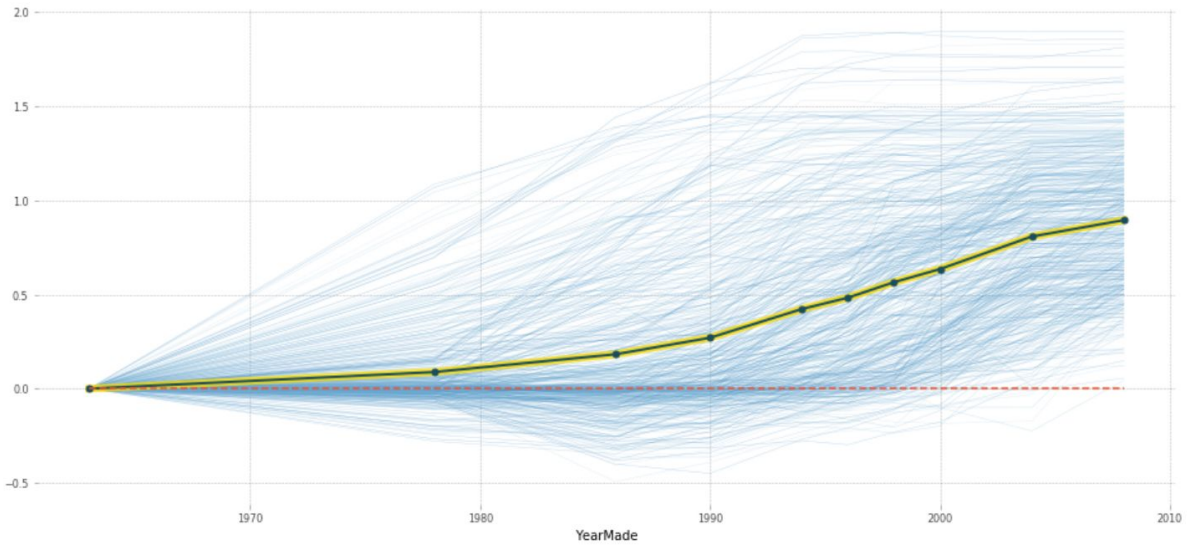
The above plot doesn't really capture the relationship between YearMade and SalePrice as there may be some other features and interactions that are driving the prices between the period 1990 to 1995 and so this is where the partial dependence plot comes in.

What are PDPs?

All other things being constant, how does my Dependent Variable change with the feature I am plotting Partial Dependence for.

How is the plot calculated?

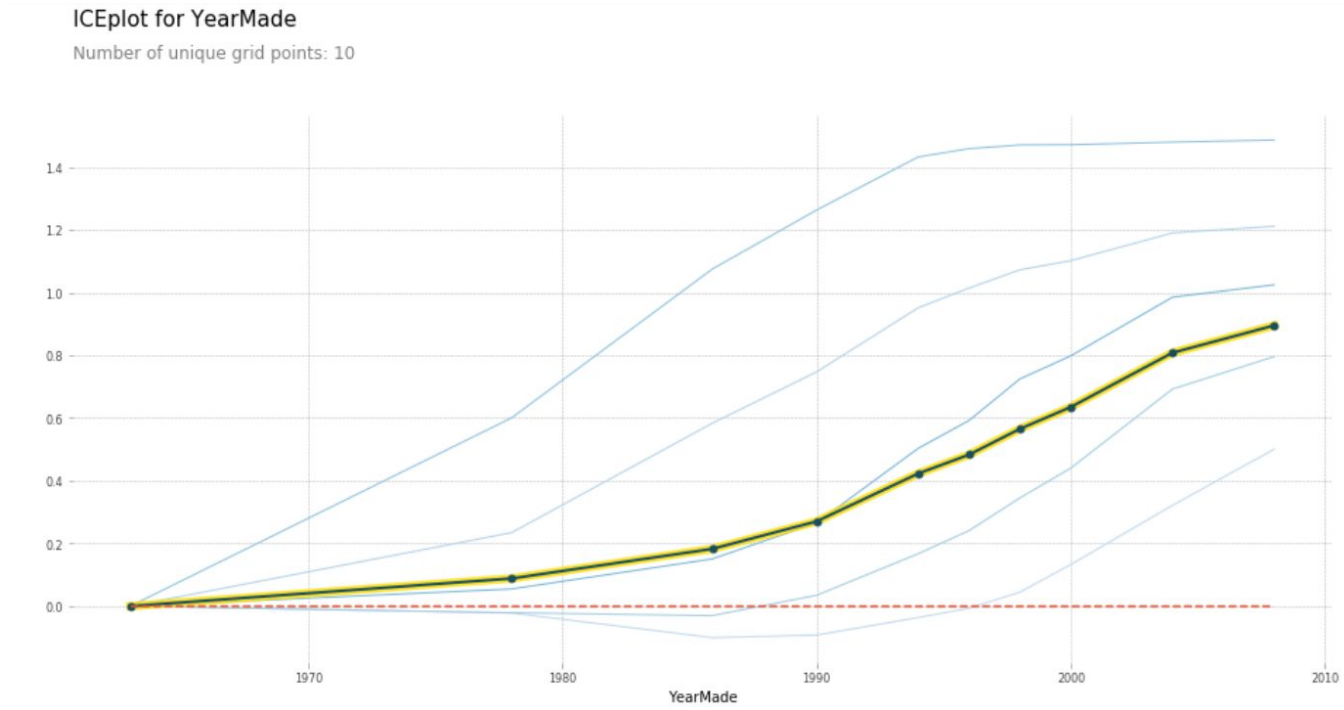
- Suppose we want the PDP for YEAR MADE.
- Leave the rest of the data as it is.
- Now replace every single value of year made column with exactly the same thing as 1960.
- Now we calculate the PREDICTIONS for the rows and then plot the average with 1960.
- Then we do the same thing for 1961 and 1962 and so on until the last value.
- The **PLOT FOR THESE POINTS GIVE US THE RELATIONSHIP BETWEEN THE FEATURE AND THE DEPENDENT VARIABLE KEEPING ALL OTHER THINGS CONSTANT.**
- **This could have also been done for EACH of the rows. You know, plotting each of their Sale Price values corresponding to 1960, 1961, 1962… and then plotting them on the graph. See image->**



- **Each of the BLUE LINE IS THE PREDICTION PLOT FOR A PARTICULAR ROW AND THE YELLOW LINE IS THE MEDIAN OF ALL THE PREDICTION PLOTS.**

WHY DOES THIS GIVE YOU THE RELATIONSHIP?

- A simple way to do PDPs would be get the largest frequency of each variable and locate a row with all of those MOST FREQUENT FEATURES and then change its year made to 1960,1961,1962… and then see its sale price plot.
- That would be one way to do it and it would give a version of the relationship between year made and sale price with all other things being equal.
- The problem with the above approach is that we won't be able to see OTHER TYPES OF RELATIONSHIPS FOR DIFFERENT TYPES OF PRODUCTS OR DIFFERENT CATEGORIES OF PRODUCTS.
- So instead of JUST PLOTTING THE AVERAGE, we plot all of the data and then take the median.
- If we really wanna see the different types of relationships, we can also plot the graph showing clusters like below->



- The above plot tells us the different kinds of clusters that our relationships have and on close looking we see that they are basically the same graph with different slopes so on average the **relationship is correct.**
- **Note in this case the graph was AGAINST YEARMADE AND LOG OF SALE PRICE.** This linear relationship means that the actual relationship is exponential.