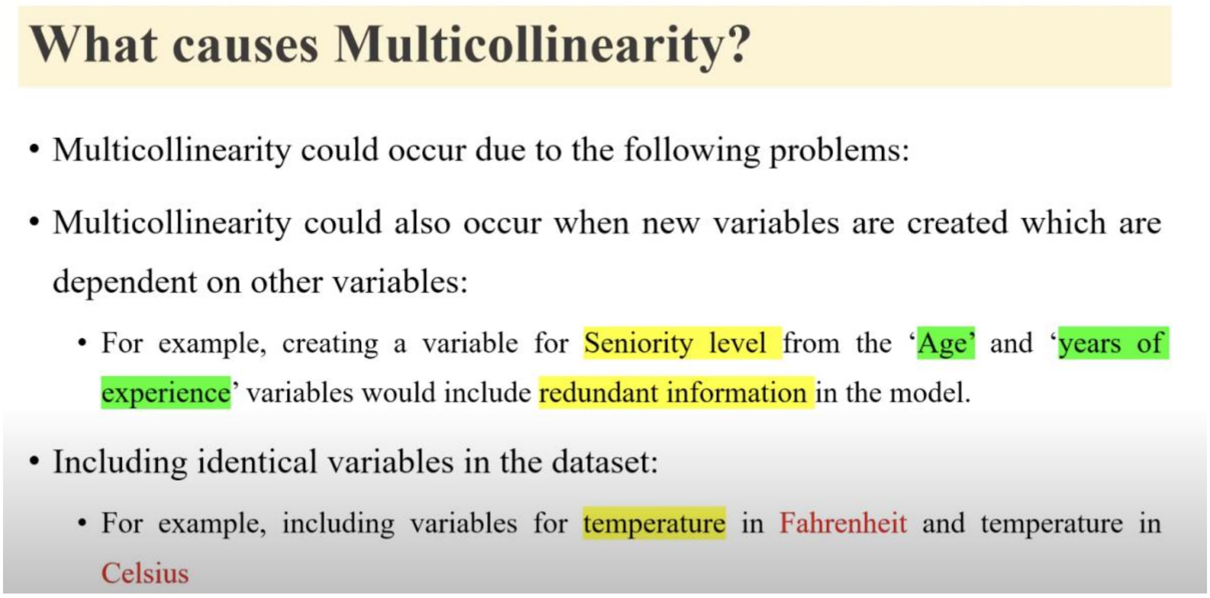
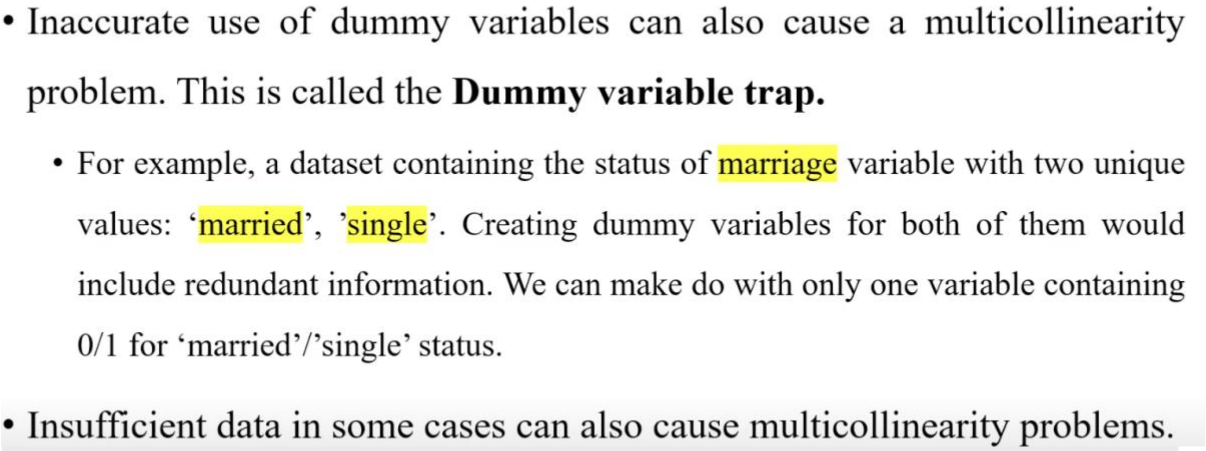
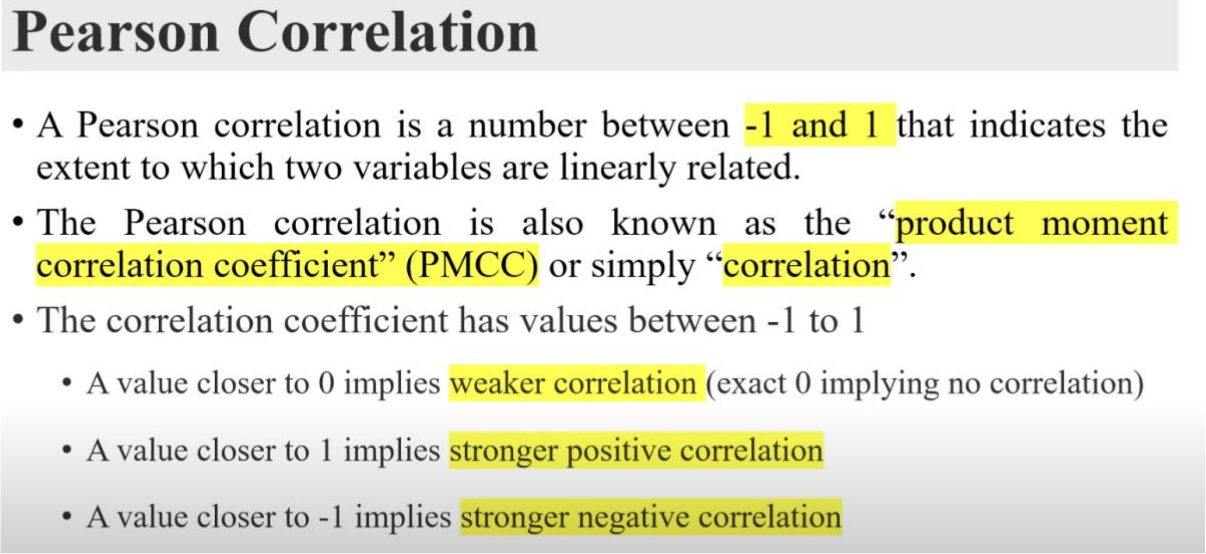


What causes Multicollinearity?

- Multicollinearity could occur due to the following problems:
- Multicollinearity could also occur when new variables are created which are dependent on other variables:
  - For example, creating a variable for Seniority level from the 'Age' and 'years of experience' variables would include redundant information in the model.
- Including identical variables in the dataset:
  - For example, including variables for temperature in Fahrenheit and temperature in Celsius

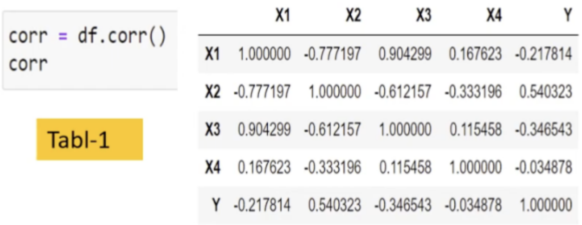


- Inaccurate use of dummy variables can also cause a multicollinearity problem. This is called the **Dummy variable trap.**
  - For example, a dataset containing the status of marriage variable with two unique values: 'married', 'single'. Creating dummy variables for both of them would include redundant information. We can make do with only one variable containing 0/1 for 'married'/'single' status.
- Insufficient data in some cases can also cause multicollinearity problems.



Pearson Correlation

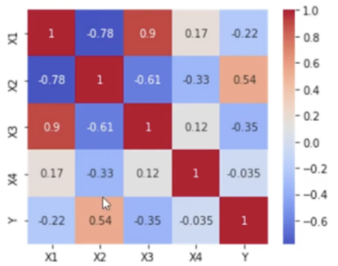
- A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related.
- The Pearson correlation is also known as the "product moment correlation coefficient" (PMCC) or simply "correlation".
- The correlation coefficient has values between -1 to 1
  - A value closer to 0 implies weaker correlation (exact 0 implying no correlation)
  - A value closer to 1 implies stronger positive correlation
  - A value closer to -1 implies stronger negative correlation



```
corr = df.corr()
corr
```

	X1	X2	X3	X4	Y
X1	1.000000	-0.777197	0.904299	0.167623	-0.217814
X2	-0.777197	1.000000	-0.612157	-0.333196	0.540323
X3	0.904299	-0.612157	1.000000	0.115458	-0.346543
X4	0.167623	-0.333196	0.115458	1.000000	-0.034878
Y	-0.217814	0.540323	-0.346543	-0.034878	1.000000

Tabl-1



Heatmap

Here our target variable is "Y" and from above figure we find out strong and weak correlation with independent variable and

How do we detect and remove multicollinearity?

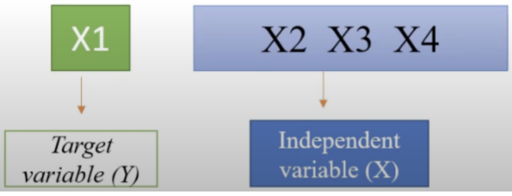
- The best way to identify the multicollinearity is to calculate the Variance Inflation Factor (VIF) corresponding to every independent Variable in the Dataset.
- VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable
- VIF score of an independent variable represents how well the variable is explained by other independent variables.
- R^2 value is determined to find out how well an independent variable is described by the other independent variables.
- A high value of R^2 means that the variable is highly correlated with the other variables. This is captured by the VIF which is denoted below:

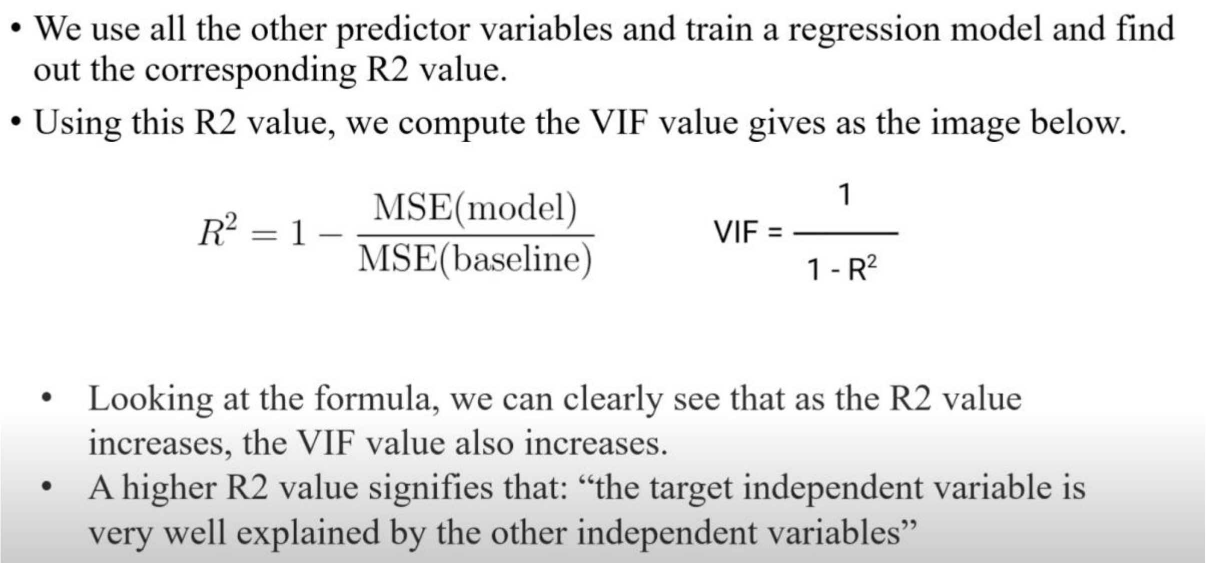
$$\text{VIF} = \frac{1}{1 - R^2}$$

- VIF tells us about how well an independent variable is predictable using the other independent variables.
- Let's understand this with the help of an example.

X1 X2 X3 X4

- Consider that we have 4 independent variables as shown.
- To calculate the VIF of variable X1, we isolate the variable X1 and consider as the target variable and all the other variables will be treated as the predictor (Independent) variables.

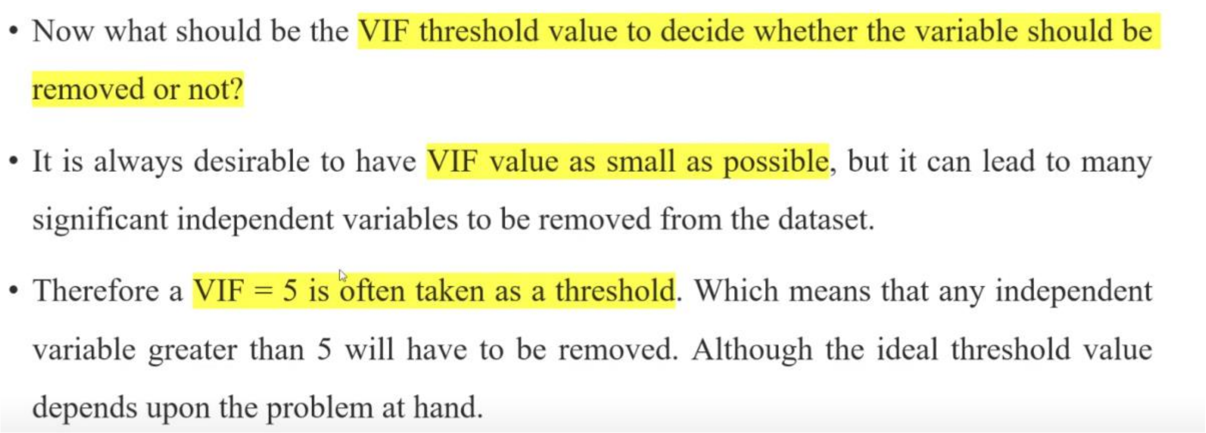




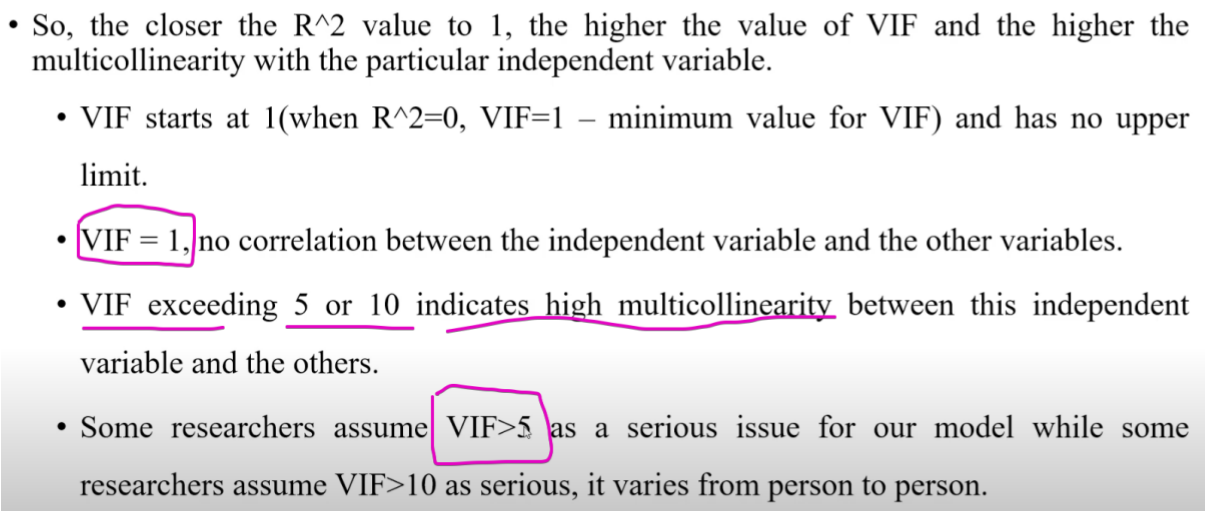
- We use all the other predictor variables and train a regression model and find out the corresponding R2 value.
- Using this R2 value, we compute the VIF value gives as the image below.

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}} \qquad \text{VIF} = \frac{1}{1 - R^2}$$

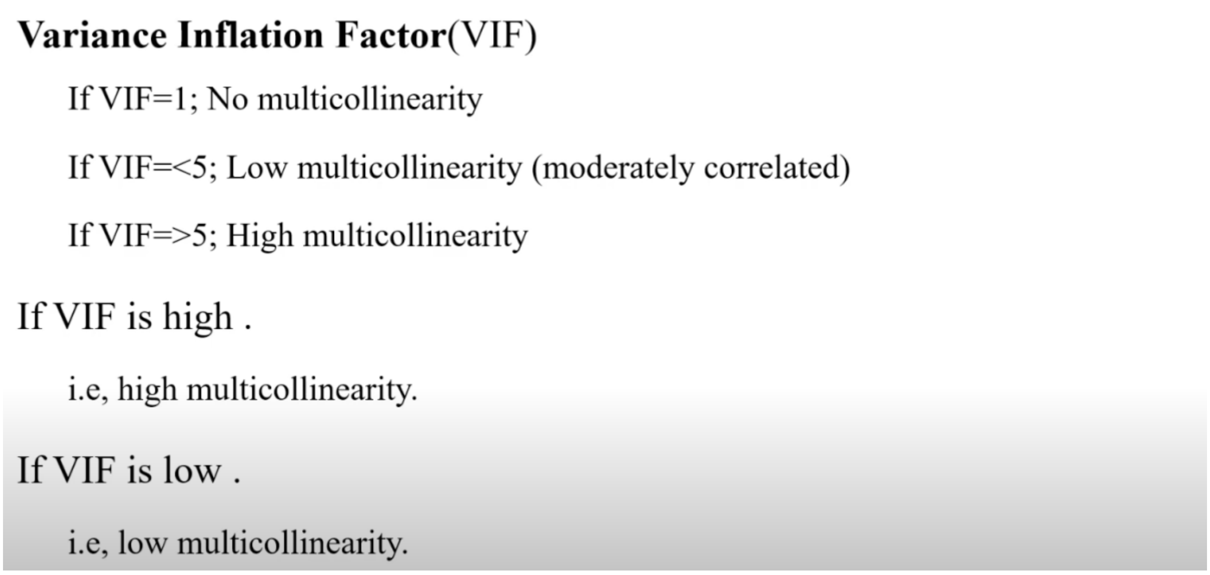
- Looking at the formula, we can clearly see that as the R2 value increases, the VIF value also increases.
- A higher R2 value signifies that: "the target independent variable is very well explained by the other independent variables"



- Now what should be the VIF threshold value to decide whether the variable should be removed or not?
- It is always desirable to have VIF value as small as possible, but it can lead to many significant independent variables to be removed from the dataset.
- Therefore a VIF = 5 is often taken as a threshold. Which means that any independent variable greater than 5 will have to be removed. Although the ideal threshold value depends upon the problem at hand.



- So, the closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.
  - VIF starts at 1(when R^2=0, VIF=1 – minimum value for VIF) and has no upper limit.
  - VIF = 1, no correlation between the independent variable and the other variables.
  - VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others.
  - Some researchers assume VIF>5 as a serious issue for our model while some researchers assume VIF>10 as serious, it varies from person to person.



Variance Inflation Factor(VIF)

If VIF=1; No multicollinearity

If VIF=<5; Low multicollinearity (moderately correlated)

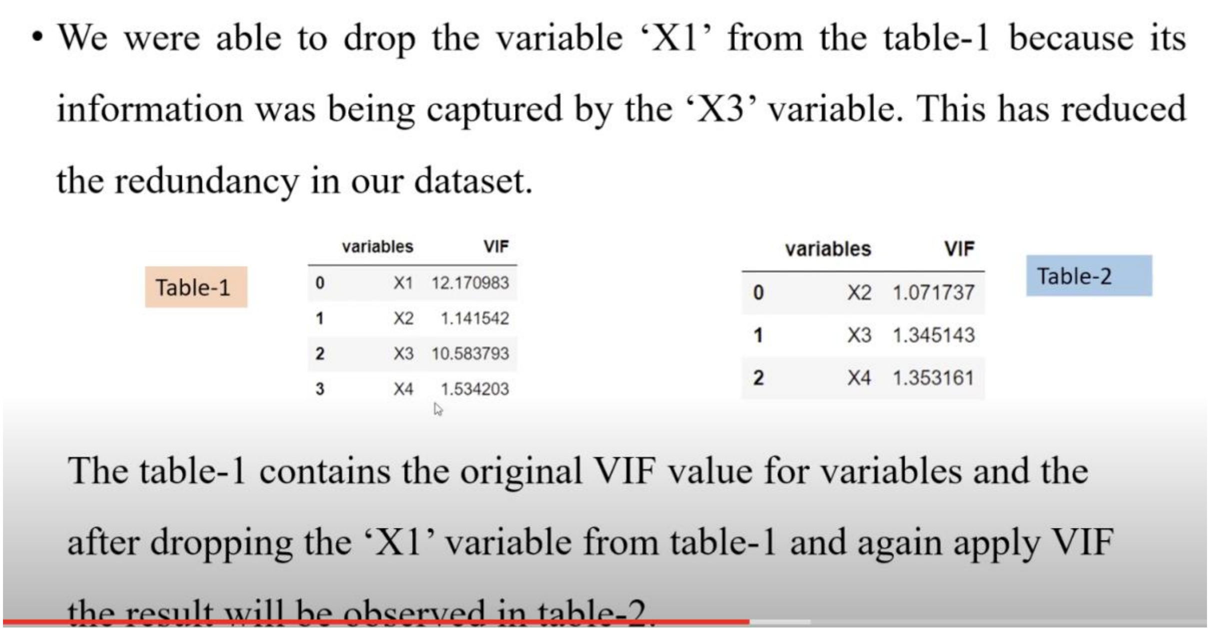
If VIF=>5; High multicollinearity

If VIF is high .

i.e, high multicollinearity.

If VIF is low .

i.e, low multicollinearity.



- We were able to drop the variable ‘X1’ from the table-1 because its information was being captured by the ‘X3’ variable. This has reduced the redundancy in our dataset.

Table-1

	variables	VIF
0	X1	12.170983
1	X2	1.141542
2	X3	10.583793
3	X4	1.534203

Table-2

	variables	VIF
0	X2	1.071737
1	X3	1.345143
2	X4	1.353161

The table-1 contains the original VIF value for variables and the after dropping the ‘X1’ variable from table-1 and again apply VIF the result will be observed in table-2.

References:
https://www.youtube.com/watch?v=qk7M749HKBs