**Multicollinearity in Data**

The variable should have a robust relationship with independent variables. However, any unbiased variables shouldn’t have robust correlations among other independent variables. Collinearity can be a linear affiliation among explanatory variables. Two variables are perfectly collinear if there’s a particular linear relationship between them.   
Multicollinearity refers to a situation at some stage in which two or greater explanatory variables in the course of a multiple correlation model are pretty linearly related. We’ve perfect multicollinearity if the correlation between impartial variables is good to 1 or -1. In practice, we do not often face ideal multicollinearity for the duration of an information set. More commonly, the difficulty of multicollinearity arises when there’s an approximately linear courting between two or more unbiased variables.   
In easy words, Multicollinearity can be defined as it’s far an event wherein one or greater of the unbiased variables are strongly correlated with one another. In such incidents, we ought to usually use just one in every correlated impartial variable.  
VIF(Variance Inflation Factor) is a hallmark of the life of multicollinearity, and *statsmodel* presents a characteristic to calculate the VIF for each experimental variable and worth of greater than 10 is that the rule of thumb for the possible lifestyles of high multicollinearity. The excellent guiding principle for VIF price is as follows, VIF = 1 manner no correlation exists, VIF > 1, but < 5 then correlation exists.    
**What Causes Multicollinearity?**  
**The principal types are:**

* Data-based multicollinearity: as a result of poorly designed experiments, statistics that’s 100% observational, or data collection methods that can’t be manipulated. In some cases, variables could also be particularly correlated (usually way to collecting facts from purely observational studies) and there’s no error on the researcher’s part. For this reason, you ought to behaviour experiments every time possible, placing the extent of the predictor variables beforehand.
* Structural multicollinearity: because of you, the researcher, when they are attempting to create new predictor variables.

**Causes for multicollinearity also can consist of:**

* Insufficient facts. In some cases, collecting extra statistics can resolve the issue.
* Dummy variables could also be incorrectly used. For instance, the researcher may fail to exclude one category, or add a dummy variable for every category (e.g. Spring, summer, autumn, winter.
* Including a variable within the regression that’s a mixture of other variables. For instance, consisting of “general investment profits” while total investment income = earnings from stocks and bonds + profits from savings interest.
* Including identical (or almost identical) variables. For instance, weight in pounds and weight in kilos, or investment earnings and savings/bond earnings.

**Example: You can also locate that multicollinearity may be a characteristic of the making plans of the test.**   
In the material manufacturer case, we can without problems see that advertising and marketing and volume are correlated predictor variables, main to foremost swings inside the impact of marketing while the quantity is and aren’t included within the version. In a similar test, you’ll find out that the product producer may additionally introduce multicollinearity between volume and advertising as it’s far part of the experimental design using assigning an excessive ad price range to cities with smaller stores and an espresso ad finances to cities with larger shops.   
If you are geared up to re-do the market test, you’ll address this difficulty via restructuring the experiment to make sure an honest aggregate of excessive ad/low volume, excessive ad/high quantity, low ad/high quantity, and low ad/low quantity stores. this may let you remove the multicollinearity in the facts set. It is regularly not possible though, to re-do an experiment. that is regularly why it’s crucial to very cautiously analyze the planning of a controlled experiment before starting so that you’ll avoid by chance causing such problems. If you’ve got located multicollinearity because of the experimental design and also you can’t re-do the experiment, you’ll deal with the multicollinearity by which include controls. inside the case of the cloth producer, it’ll be vital to incorporate extent inside the version as an effect to urge a much higher proper estimate for the impact of advertising. Other answers to addressing multicollinearity in instances like this consist of shrinkage estimations like principal additives regression or partial least-squares analysis

**Effects of Multi-Collinearity**

Due to the presence of collinearity/multi-collinearity, it becomes difficult to isolate the individual effects of explanatory variables on the response variable.

Multi-collinearity results in the following:

1. **Uncertainty in coefficient estimates or unstable variance:** Small changes (adding/removing rows/columns) in the data results in change of coefficients.
2. **Increased standard error:** Reduces the accuracy of the estimates and increases the chances of detection.
3. **Decreased statistical significance:** Due to increased standard error, t-statistic declines which negatively impacts the capability of detecting statistical significance in coefficient leading to type-II error.
4. **Reducing coefficient & p-value:**The importance of the correlated explanatory variable is masked due to collinearity.
5. **Overfitting:**Leads to overfitting as is indicated by the high variance problem.

**How to detect Multi-Collinearity?**

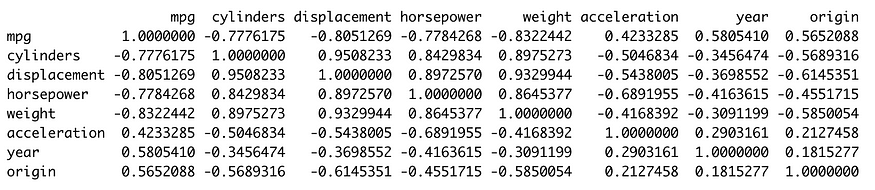
In addition to observing the model’s behavior for the above stated effects, mutli-collinearity is quantitatively captured in correlation values too. Thus, the following can be used:

**Correlation matrix:**

Pearson’s correlation between two variables in the data varies between -1 to 1. The two variables data type should be numeric for calculation of pearson’s correlation value.

Here, the correlation matrix for the[Auto MPG](http://archive.ics.uci.edu/ml/datasets/Auto+MPG) dataset using R. The column name contains string and hence is eliminated. The response variable is mpg which represents fuel consumption efficiency.

#data = Auto  
#generate correlation matrix in Rcorrelation\_matrix <- cor(Auto[, -which(names(Auto) == "name")])



The high correlation between cylinder and displacement, horsepower and weight can be observed. Additionally, there are several pairs of explanatory variables with high positive/negative correlation. Thus, there is multi-collinearity.

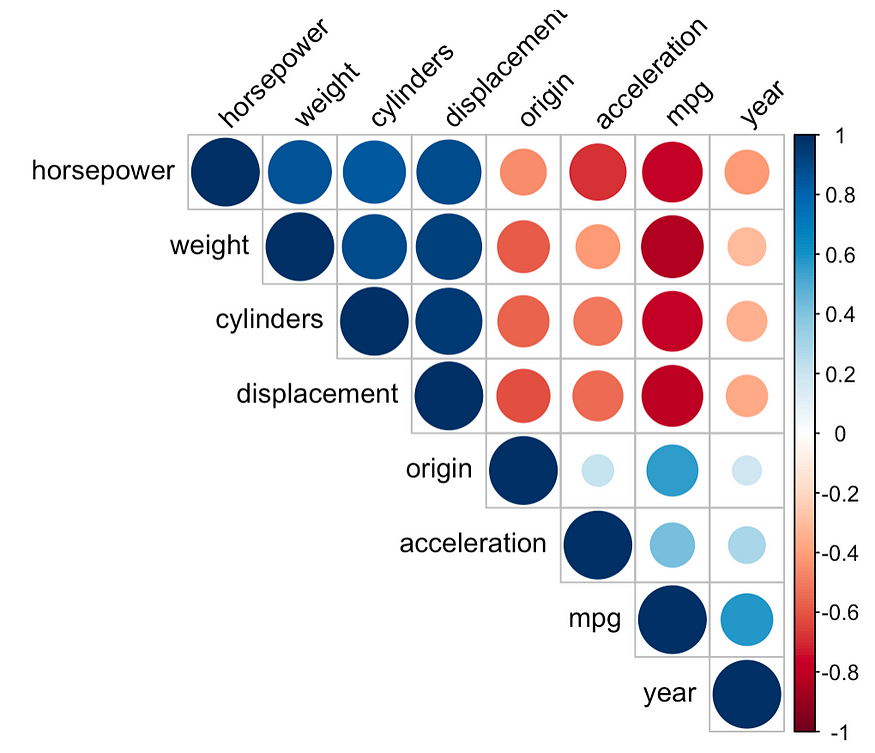
However, as one would notice, going through the table and identifying these variables is tiresome for 8 variables and it will only get worse as the number of variables increase. Thus, a heatmap of correlation is a more intuitive representation of correlation.

**Heatmap of correlations:**

Heatmap of correlations helps visualize the data better by adjusting the color for positive and negative correlation and size for magnitude.

In R, corrplot package can be used to create a heatmap of correlations.

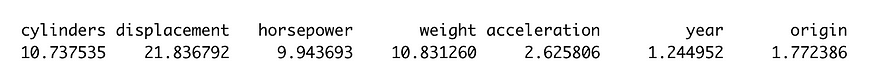
library(corrplot)  
corrplot(correlation\_matrix, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)



The heatmaps are definitely more intuitive & visual. However, it helps identify correlation between 2 variables strictly and **fails to identify collinearity which exists between 3 or more variables, for which Variance Inflation Factor can be used.**

**Variance Inflation Factor (VIF):**VIF is the ratio of variance of coefficient estimate when fitting the full model divided by the variance of coefficient estimate if fit on its own. The minimum possible value is 1 which indicates no collinearity. If value exceeds 5, then collinearity should be addressed.

library(rms)  
multiple.lm <- lm(mpg ~ . -name, data = Auto)  
summary(multiple.lm)  
vif(multiple.lm)

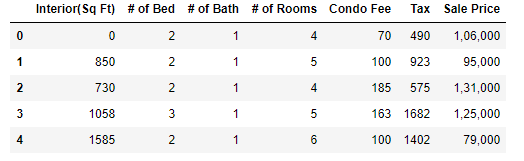


The VIF value for cylinders, displacement, horsepower, and weight are way higher than 5 and hence should be handled as the collinearity is high in the data.

**Dealing with Multi-Collinearity**

Multi-collinearity can be handled with the following two methods. Note that this correlation between independent variable leads to data redundancy, eliminating which can help get rid of multi-collinearity.

1. **Introduce penalization or remove highly correlated variables:**Use lasso and ridge regression to eliminate variables which provide information which is redundant. This can also be achieved by observing the VIF.
2. **Combine highly correlated variables:**Since the collinear variables contain redundant information, combining them into a single variable using methods such as PCA to generate independent variables.
3. The features in the dataset are as below. Sale Price is the target variable and the remaining are independent features.

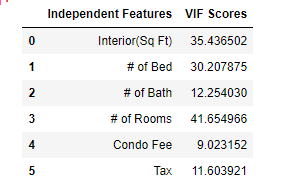


First few rows of the dataset

**Identifying Multicollinearity with numerical variables**

Now we will calculate the VIF scores for each independent variable.

def vif\_scores(df):  
 VIF\_Scores = pd.DataFrame()  
 VIF\_Scores["Independent Features"] = df.columns  
 VIF\_Scores["VIF Scores"] = [variance\_inflation\_factor(df.values,i) for i in range(df.shape[1])]  
 return VIF\_Scoresdf1 = df.iloc[:,:-1]  
vif\_scores(df1)



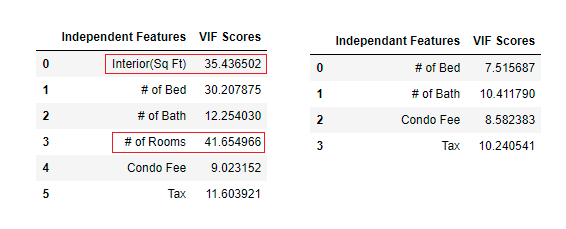
VIF for each independent feature

The VIF scores are higher than 10 for most of the variables. The individual coefficients and the p-values will be greatly impacted if we build a regression model with this dataset. We will proceed on how to fix this issue.

**Fixing Multicollinearity — Dropping variables**

We will consider dropping the features *Interior(Sq Ft)*and *# of Rooms* which are having high VIF values because the same information is being captured by other variables. Also, it helps to reduce the redundancy in the dataset.

Let us compare the VIF values before and after dropping the VIF values.



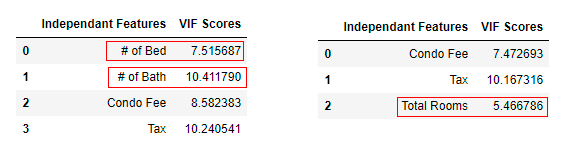
The highlighted features are removed and calculated the VIF scores

From the above, we can notice that the VIF scores have reduced for other variables also after dropping the high-value (*Interior(Sq Ft)*and *# of Rooms)*VIF features.

**Fixing Multicollinearity — Combining variables**

Next, we can observe that #of Bed and #of Bath can be combined to a single variable and it helps us to capture more information from both variables.

df5 = df4.copy()  
df5['Total Rooms'] = df4.apply(lambda x: x['# of Bed'] + x['# of Bath'],axis=1)  
X = df5.drop(['# of Bed','# of Bath'],axis=1)  
vif\_scores(X)



#of Bed and #of Bath features combined to a single feature Total Rooms

From the above, we can notice that all three variables (Condo Fee, Tax, and Total Rooms) came up with satisfying VIF values and we can proceed further to build a regression model.

**Summary**: we learned how to identify and the ways to fix the multicollinearity issue with numerical values in regression analysis. This is an iterative process and we also need domain knowledge to decide on which variables to take an appropriate action. There are other techniques such as PCA and Regularization methods also to address this issue.

[What is multicollinearity and how to remove it? | by Sharoon Saxena | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/what-is-multicollinearity-and-how-to-remove-it-413c419de2f)

[Multicollinearity in Data - GeeksforGeeks](https://www.geeksforgeeks.org/multicollinearity-in-data/)

[Multicollinearity: Problem, Detection and Solution | Analytics Vidhya](https://www.analyticsvidhya.com/blog/2021/02/multicollinearity-problem-detection-and-solution/)

[What is Multicollinearity? | Causes, Effects and Detection Using VIF (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/)