# Case Study on Donut Data Set using Linear Regression

The Donut data set was provided in an interview round and it was asked to solve the problem statement hands on. The data set can be accessed through the following link.

<https://s3.amazonaws.com/acadgildsite/wordpress_images/datasets/donut/train.csv>

<https://s3.amazonaws.com/acadgildsite/wordpress_images/datasets/donut/test.csv>

## Read the data and import all necessary libraries

## Check for missing values

## Handling categorical features

1. create dummy variables of train data
2. drop the first column
3. concatenate the original DataFrame and the dummy DataFrame
4. print 5 random rows

Repeat the above steps for the test data as well.

## Visualizing the data

1. Import the libraries
2. Set the figure sizes
3. Plot multiple scatter plots
4. Line plots
5. **Check for multi-collinearity**
6. Create a function named remove\_collinear\_features(x, target , threshold)

**Objective:**

Remove collinear features in a dataframe with a correlation coefficient

greater than the threshold. Removing collinear features can help a model

to generalize and improves the interpretability of the model.

Inputs:

threshold: any features with correlations greater than this value are removed

Output:

dataframe that contains only the non-highly-collinear features

1. Define y and x
2. Calculate the correlation matrix
3. Iterate through the correlation matrix and compare correlations

# Print the correlated features and the correlation value

1. Drop one of each pair of correlated columns
2. Add the score back in to the data
3. **Creating baseline with null RMSE**

Null RMSE is the RMSE that could be achieved by **always predicting the mean response value**. It is a benchmark against which you may want to measure your regression model.

1. define a function that accepts a list of features and returns testing RMSE [def get\_baseline(feature\_cols,target)]
2. create a NumPy array with the same shape as y\_test
3. fill the array with the mean value of y\_test
4. compute null RMSE

Output: The baseline guess for Donut volume is a score of 83.40

Baseline Performance on the test set for Donut volume: RMSE = 10.5856

1. **Fitting on entire features**
2. ***define a function that accepts a list of features and returns testing RMSE***
3. ***compare different sets of features***
4. ***'Donut Density', 'Donut volume'***
5. ***Donut Density***
6. ***pair the feature names with the coefficients***
7. ***Donut volume'***
8. ***pair the feature names with the coefficients***
9. **Check and treat outliers**

## Treat multi-collinearity

1. visualize correlation matrix in Seaborn using a heatmap

**What is your conclusion on this project?**