For this project, I wanted to do something dealing with manufacturing, since that is the field, I work in. I found my data from a Kaggle competition. Caterpillar had a competition about 5 years ago with the end goal of estimating the quoted price of tubing. Many tubes are in machines like forklifts, loaders, and bulldozers. Caterpillar provided tubing dimensions, annual usage, and past pricing for the contestants to build a predictive algorithm. My project will use this same data but instead of predicting price, I want to predict usage. I believe that there will be correlation between the many dimensions of the tubing and the annual usage.

There are a couple of files to piece together for the analysis. Below are the variables.

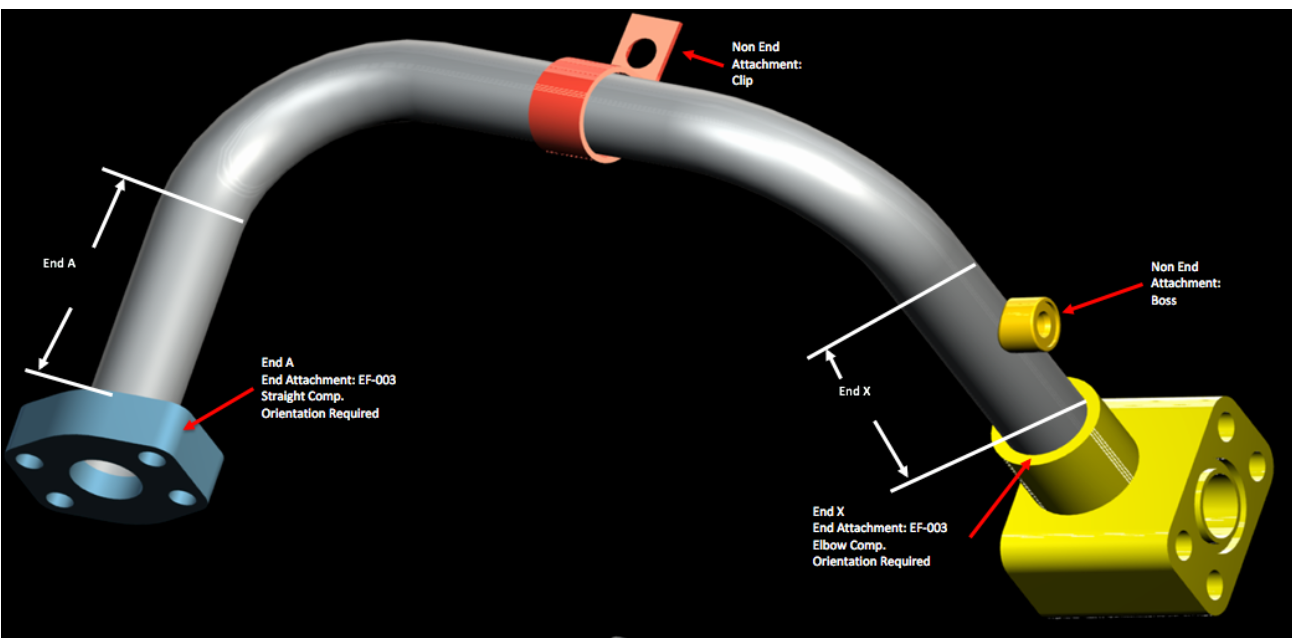
Train/Test:

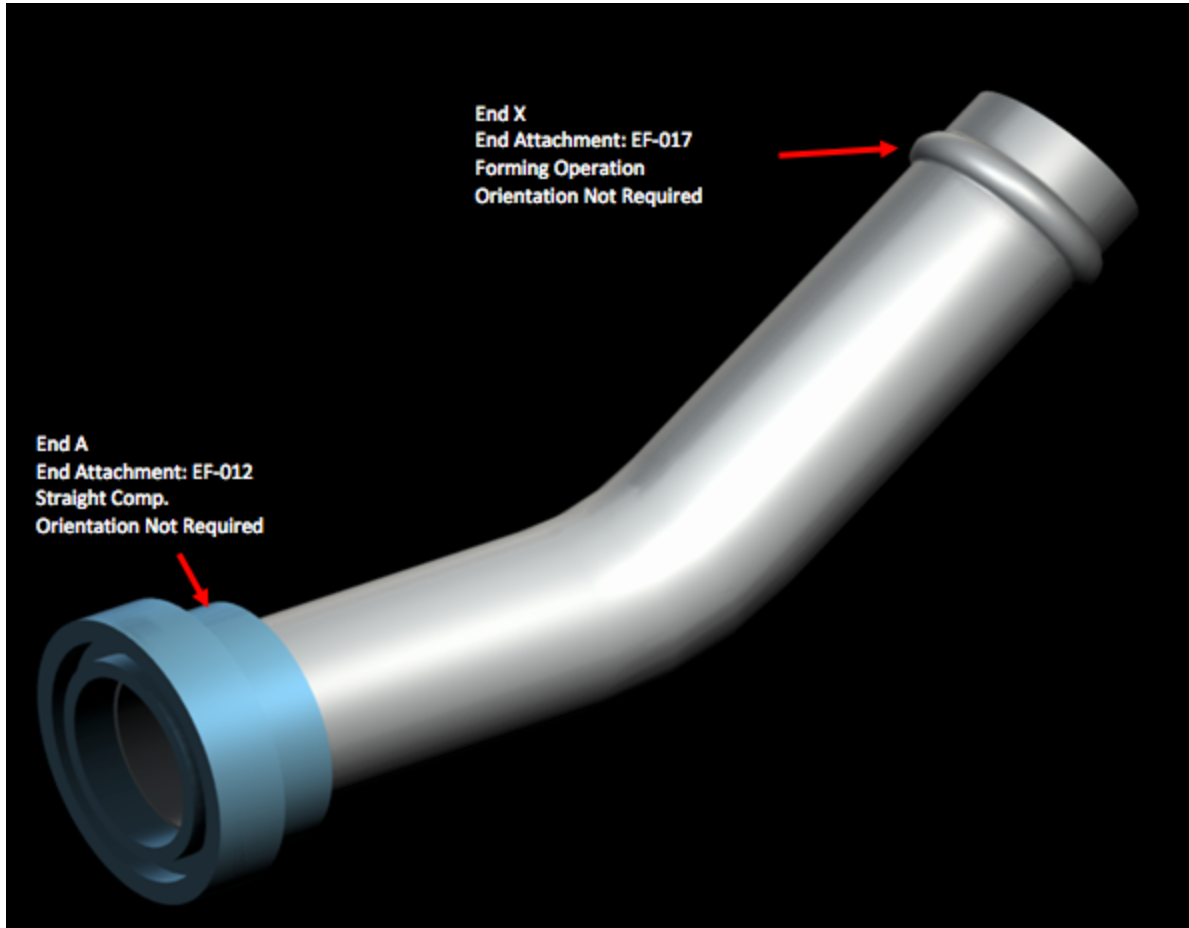
* Tube assembly id
* Supplier
* Quote date
* Annual usage
* Minimum order quantity
* Bracket pricing
* Quantity
* Cost

Tube:

* Tube Assembly ID
* Material ID
* Diameter
* Wall
* Length
* Num Bend
* Bend Radius

Below are pictures of tubing from the data. It also explains some of the aspects of the tubes.





# Part 1

1. Load Data
   1. Clean Train data so there is one line for each tube ID
   2. Left Join tube to train data.
2. Check the Dimension
3. Look at the head of the data
4. Describe the Data
5. Summarize the data
6. Create histograms for Annual Usage, quantity, cost, diameter, wall, length, num\_bends, bend radius
7. Create bar charts for end\_a\_1x, end\_a\_2x, end\_x\_1x, and end\_x\_2x
8. Pearson Ranking of numeric variables
9. Compare end\_a\_2x Yes and No against annual usage, unit cost, diameter, and wall
10. I did not create a stacked bar chart because I am focused on how a continuous variable (usage) can be predicted.

# Part 2

1. Material ID was the only column with missing data. Since the material IDs are grouped together, I performed a front fill to fill in the missing data. I removed the other and tube assembly ID columns because they were unnecessary. I deleted rows with a bend radius of 9999 because this was an extreme outlier. There were 8 rows with a bend radius of 9999. The average bend radius excluding the outliers is 45.
2. I used a log transformation on length and bend radius.
3. I converted the end\_a and end\_x variables to numeric variables.

# Part 3

1. Combine features and split up train and test data. I made usage group a categorical variable by creating group ranges.
2. Evaluate Logistic Regression model. The predictor did a much better job when the usage was between 0 and 500 units per year.