# 1 Pre-analysis & Questions

The first objective is to observe what features may influence the cost of tubes. We start our observation with the features given in the training set where we define a variable for each feature.

| Variable | Feature |
|---|---|
| $x_1$ | tube_assembly_id |
| $x_2$ | supplier |
| $x_3$ | quote_date |
| $x_4$ | annual_usage |
| $x_5$ | min_order_quantity |
| $x_6$ | bracket_pricing |
| $x_7$ | quantity |
| $x_8$ | cost |

### 1.1 Tube Physical Properties

As a supplier, we have to think on which tube features the cost will be based. We know that a tube assembly is made with one or more components. Some numerical tube properties may be helpful to check.

- The weight
- The quantity
- The volume
- The number of bends used. Logically, it is more difficult to bend a tube than to keep it linear, so it should be more expansive.
- The component types used to assemble the tube.

### 1.2 Supplier Features

- The date when the supplier has quoted the price which is certainly less 20 years ago than today when not adjusted.
- The suppliers may use different mathematical models to quote their price.
- The supplier uses or not a bracket pricing which can influence what features to use in both cases.

### 1.3 Other Observations

- The tube assembly ID may be used at some points in the prediction. We need to investigate why and how these tubes are chosen in the train set.
- The costs have too many decimals to be a real cost. Maybe a conversion is needed in some way to get the real cost.
- The supplier may have decided to quote the tubes with very expensive or cheap price. These prices can be considered as anomalies.

### 1.4 Questions to Answer

To archieve our goal of predicting the cost with a good accuracy, we need to answer the following questions.

1. What features are used to determine the cost and what features to exclude from the analysis?

2. Do the costs presented in the training set are the real costs or do they need to be adjusted?
3. Is there a unique mathematical model describing the cost in function of the quantity for each supplier?
4. What particular features need to be classified? Why? How?
5. Are there anomalies to consider in the dataset? Why?

# 2 Preparing & Cleaning the Dataset

In this section, we will answer the first question: *What features are used to determine the cost and what features to exclude from the analysis?* We will also explain why we chose to keep and exclude features and how we will clean the dataset.

From the dataset, we note that there are a total of 2048 components. These components are spread amoung the `comp_[type].csv` files uniquely. This means that we can create a single table `Component` by merging those files together. To avoid to many columns, we will remove some features that we do not want in our analysis.

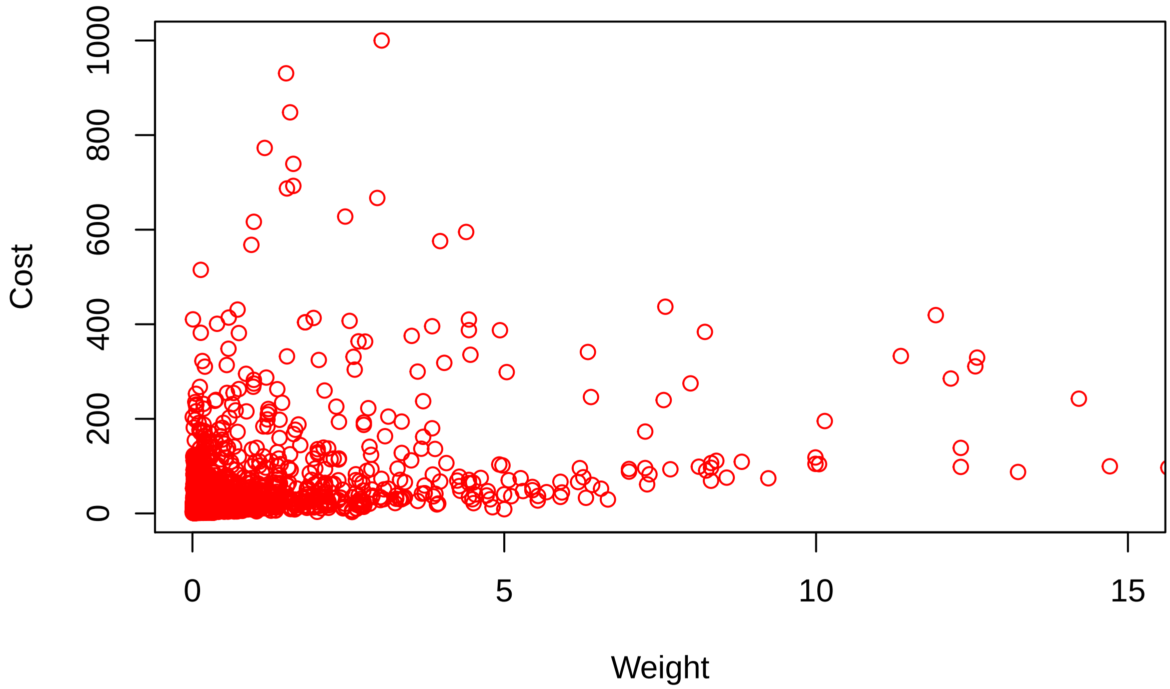
We use the Chi-Squared test to check if a feature is independent or dependent of the cost. We reject the null hypothesis of independence if the p-value is greater than 0.05.

| | tubeAssemblyID | totalWeight | volume | numberOfBends | cntQty | cost |
|---|---|---|---|---|---|---|
| 1 | 2 | 0.02 | 1723.49 | 8 | 8.00 | 21.91 |
| 2 | 4 | 0.02 | 1723.49 | 9 | 8.00 | 21.97 |
| 3 | 5 | 0.21 | 7562.44 | 4 | 8.00 | 28.37 |
| 4 | 12 | 0.02 | 2604.10 | 7 | 8.00 | 22.42 |
| 5 | 13 | 0.21 | 20027.98 | 3 | 1.00 | 10.00 |
| 6 | 14 | 0.10 | 3170.01 | 4 | 8.00 | 21.99 |
| 7 | 21 | 0.05 | 2404.38 | 6 | 1.00 | 3.43 |
| 8 | 22 | 0.02 | 3335.12 | 6 | 1.00 | 8.56 |
| 9 | 24 | 0.03 | 1945.45 | 2 | 8.00 | 20.93 |
| 10 | 25 | 0.03 | 1277.31 | 3 | 8.00 | 20.78 |

Table 2:

Pearson's Chi-squared test

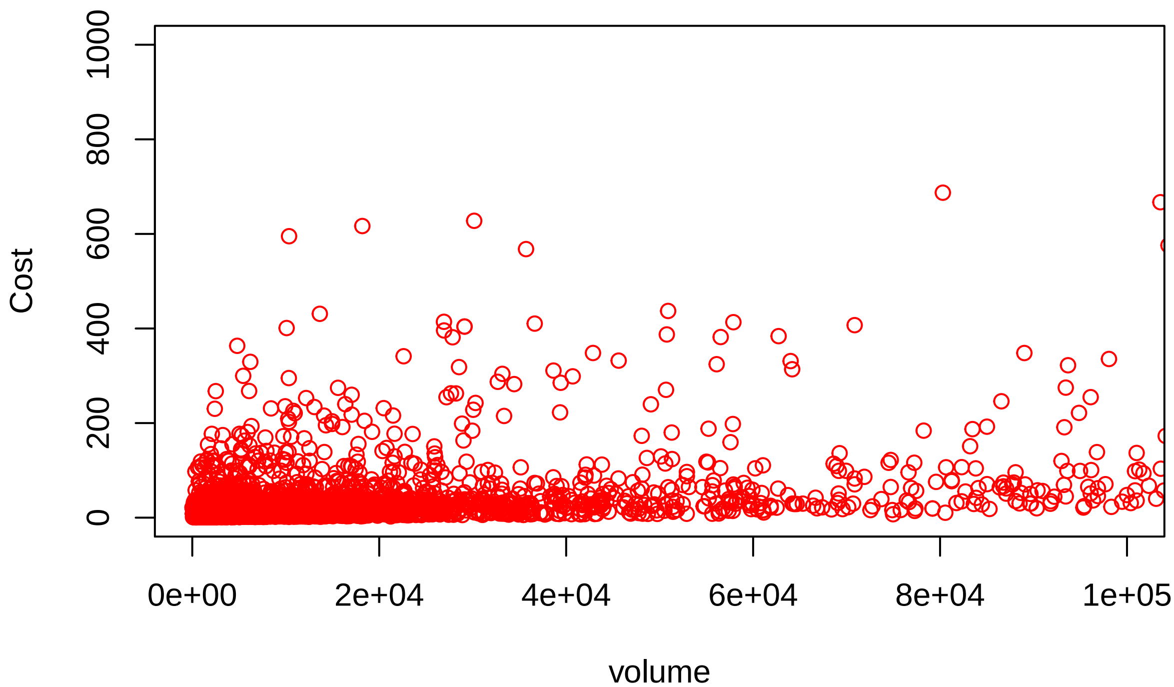
data: tbl X-squared = 5244100, df = 4262600, p-value $< 2.2\text{e-}16$




Following the Chi-Square test done and the plot, the cost is independent of the weight of a tube since the p-value is less than 0.05.

```
Pearson's Chi-squared test
```

data: tbl X-squared = 22286000, df = 20737000, p-value < 2.2e-16




Following the Chi-Square test done and the plot, the cost is independent of the volume of a tube since the p-value is less than 0.05.

```
Pearson's Chi-squared test
```

data: tbl X-squared = 87527, df = 93262, p-value = 1

Following the Chi-Square tests done, the cost is dependent of the number of bends since the p-value is greater than 0.05.

# 3 Real Cost vs Adjusted Cost

# 4 Supplier's Model

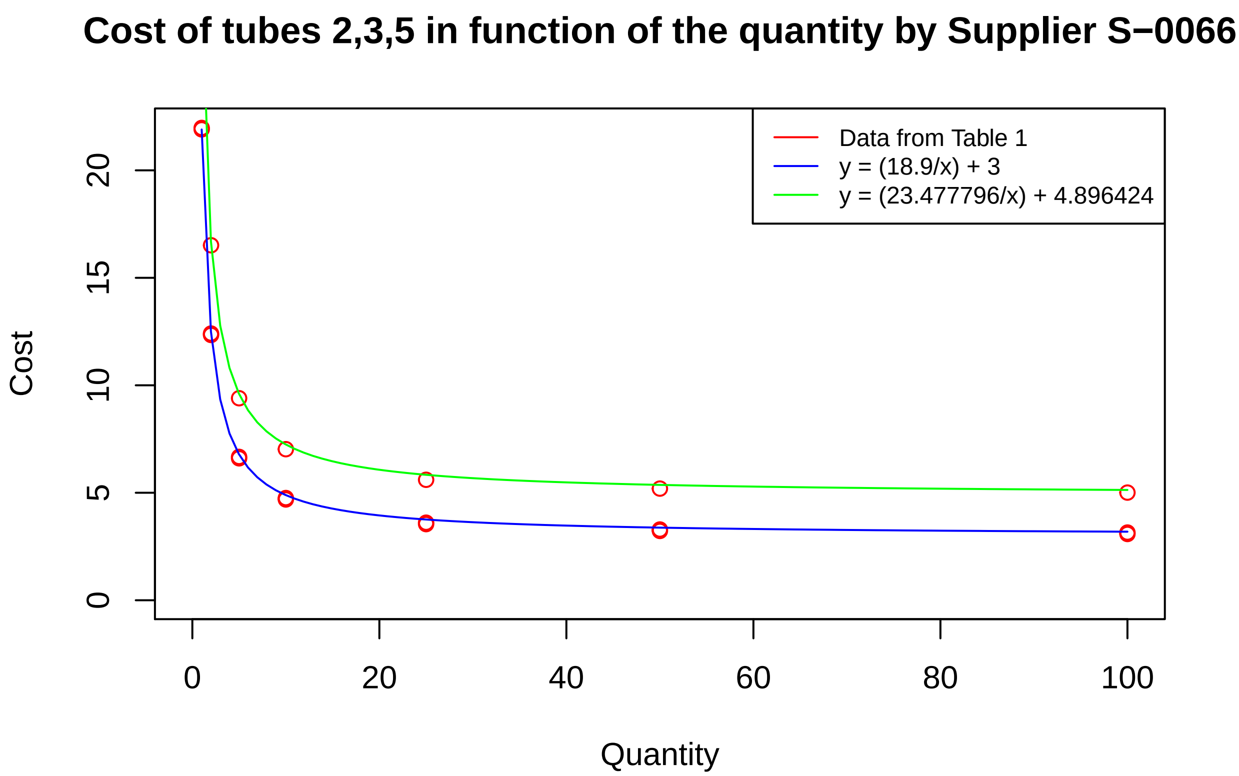
In this section, we simplify the dataset where we use the bracket pricing with the supplier S-0066. The objective is to find a mathematical model representing the cost in function of the quantity.

Let's denote $h_\beta(x)$ our cost heuristic function of a tube assembly given by a supplier where $\beta$ is our learning parameters. Since the output of this analysis is known (the cost a supplier will quote for a given tube assembly), then we will use a supervised algorithm.

We start with the 3 first tube assemblies (TA-00002, TA-00004 and TA-00005) which have the bracket pricing and supplier S-0066.

| | fkTubeAssembly | supplierID | quantity | cost |
|---|---|---|---|---|
| 1 | 2 | S-0066 | 1 | 21.91 |
| 2 | 2 | S-0066 | 2 | 12.34 |
| 3 | 2 | S-0066 | 5 | 6.60 |
| 4 | 2 | S-0066 | 10 | 4.69 |
| 5 | 2 | S-0066 | 25 | 3.54 |
| 6 | 2 | S-0066 | 50 | 3.22 |
| 7 | 2 | S-0066 | 100 | 3.08 |
| 8 | 2 | S-0066 | 250 | 3.00 |
| 9 | 4 | S-0066 | 1 | 21.97 |
| 10 | 4 | S-0066 | 2 | 12.41 |
| 11 | 4 | S-0066 | 5 | 6.67 |
| 12 | 4 | S-0066 | 10 | 4.75 |
| 13 | 4 | S-0066 | 25 | 3.61 |
| 14 | 4 | S-0066 | 50 | 3.29 |
| 15 | 4 | S-0066 | 100 | 3.15 |
| 16 | 4 | S-0066 | 250 | 3.07 |
| 17 | 5 | S-0066 | 1 | 28.37 |
| 18 | 5 | S-0066 | 2 | 16.51 |
| 19 | 5 | S-0066 | 5 | 9.40 |
| 20 | 5 | S-0066 | 10 | 7.03 |
| 21 | 5 | S-0066 | 25 | 5.60 |
| 22 | 5 | S-0066 | 50 | 5.19 |
| 23 | 5 | S-0066 | 100 | 5.01 |
| 24 | 5 | S-0066 | 250 | 4.90 |

Table 3: Table built from Tube 2, 4 and 5




From the plot, we see that the curve representing the points is clearly an hyperbola of equation

$$h_\beta(x_7) = \frac{\beta_0}{x_7} + \beta_1$$

where $x_7 \geq 1$, $\beta_1$ is the cost at the last level of purchase based on quantity and $\beta_0 = h_\beta(1) - \beta_1$. This equation indicates that if Caterpillar buy more tubes, cheaper will be the cost per tube by the supplier `S-0066`.

We need to find on which features depend $\beta_0$ and $\beta_1$.

### Physical Properties of Tubes

The goal of this section is to get the total weight for each tube assembly and check if the cost depends of the weight. If yes, then we have to modelize this dependency to help our estimation of the cost.

The file `bill_of_materials.csv` gives us the list of components with their respective quantity used to assemble a tube. Thus, to calculate the total weight for each tube, we use the formula

$$W_T = \sum_{i=0}^{n} W_i * Q_i$$

where $W_T$ is the total weight of the tube $T$, $W = (W_1, \ldots, W_n)$ is the vector of component weights, $Q = (Q_1, \ldots, Q_n)$ the vector of component quantities and $n \leq 8$ the number of possible components used to assemble a tube $T$.

Let the total volume estimation of a tube assembly be denoted by $V_T$. The volume is function of the length, the wall thickness and the diameter of the tube and its formula is

$$V_T = \pi L t(d - t)$$

, where $t$ is the wall thickness, $d$ the outside diameter and $L$ the developed length of the tube.

| TubeAssemblyID | supplierID | totalWeight | volume | numberOfBends | quantity | cost |
|---|---|---|---|---|---|---|
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 1 | 21.905933 |
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 2 | 12.341214 |
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 5 | 6.601826 |
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 10 | 4.687769 |
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 25 | 3.541561 |
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 50 | 3.224406 |
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 100 | 3.082521 |
| 2 | S-0066 | 0.018 | 1723.487 | 8 | 250 | 2.999060 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 1 | 21.972702 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 2 | 12.407983 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 5 | 6.668596 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 10 | 4.754539 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 25 | 3.608331 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 50 | 3.291176 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 100 | 3.149291 |
| 4 | S-0066 | 0.018 | 1723.487 | 9 | 250 | 3.065829 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 1 | 28.374220 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 2 | 16.514303 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 5 | 9.397796 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 10 | 7.027481 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 25 | 5.603067 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 50 | 5.194104 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 100 | 5.007706 |
| 5 | S-0066 | 0.210 | 7562.441 | 4 | 250 | 4.896424 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 1 | 22.415050 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 2 | 12.850331 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 5 | 7.113725 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 10 | 5.199668 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 25 | 4.050678 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 50 | 3.736305 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 100 | 3.594420 |
| 12 | S-0066 | 0.018 | 2604.100 | 7 | 250 | 3.510959 |
| 13 | S-0026 | 0.215 | 20027.983 | 3 | 1 | 10.004284 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 1 | 21.994959 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 2 | 12.430240 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 5 | 6.690852 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 10 | 4.779578 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 25 | 3.630587 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 50 | 3.316214 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 100 | 3.174329 |
| 14 | S-0066 | 0.096 | 3170.013 | 4 | 250 | 3.088086 |

**Number of Bends Dependency**

From the table, we can see that for a fixed quantity, the cost varies. This implies that the cost depends also on tube assembly features. By comparing the tube TA-00002 and TA-00004 from the table `TubeAssembly`, we see that the only feature that varies is the number of bends. The tube TA-00002 is made with 8 bends and the tube TA-00004 is made with 9 bends. Thus, we can find the variation of the cost per bend for the supplier S-0066. From the table, we have $21.9727024365 - 21.9059330191 = 0.066769417$ for the minimal quantity (which is 1). Therefore, the equation for the cost of bends is given by $C(B) = 0.066769417B$ where $B$ is the number of bends in a tube.

# 5 Classified Features

# 6 Anomalies Detection

| fkTubeAssembly | supplierID | minQty | cntQty | cost |
|---:|---|---:|---:|---:|
| 5013 | S-0054 | 1 | 8 | 7.135981 |
| 1243 | S-0066 | 15 | 8 | 14.942439 |
| 18244 | S-0054 | 1 | 8 | 16.781380 |
| 20621 | S-0066 | 1 | 8 | 16.920483 |
| 20557 | S-0066 | 1 | 8 | 16.953868 |
| 20558 | S-0066 | 1 | 8 | 17.051240 |
| 8661 | S-0066 | 1 | 8 | 17.056804 |
| 19148 | S-0066 | 1 | 8 | 17.129138 |

| | fkTubeAssembly | supplierID | minQty | cntQty | cost |
|---|---:|---|---:|---:|---:|
| 2194 | 19143 | S-0066 | 1 | 8 | 55.65231 |
| 2195 | 20477 | S-0066 | 1 | 8 | 59.63344 |
| 2196 | 20619 | S-0066 | 1 | 8 | 60.37903 |
| 2197 | 20639 | S-0066 | 1 | 8 | 63.07206 |
| 2198 | 18838 | S-0066 | 1 | 8 | 66.51625 |
| 2199 | 20272 | S-0066 | 1 | 8 | 68.06029 |
| 2200 | 20273 | S-0066 | 1 | 8 | 68.11037 |
| 2201 | 5000 | S-0066 | 1 | 8 | 141.65967 |

```
used (Mb) gc trigger   (Mb)  max used   (Mb)
```

Ncells 543894 29.1 940480 50.3 940480 50.3 Vcells 798004 6.1 150343938 1147.1 187920911 1433.8