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 with the bend radius. 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.

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.

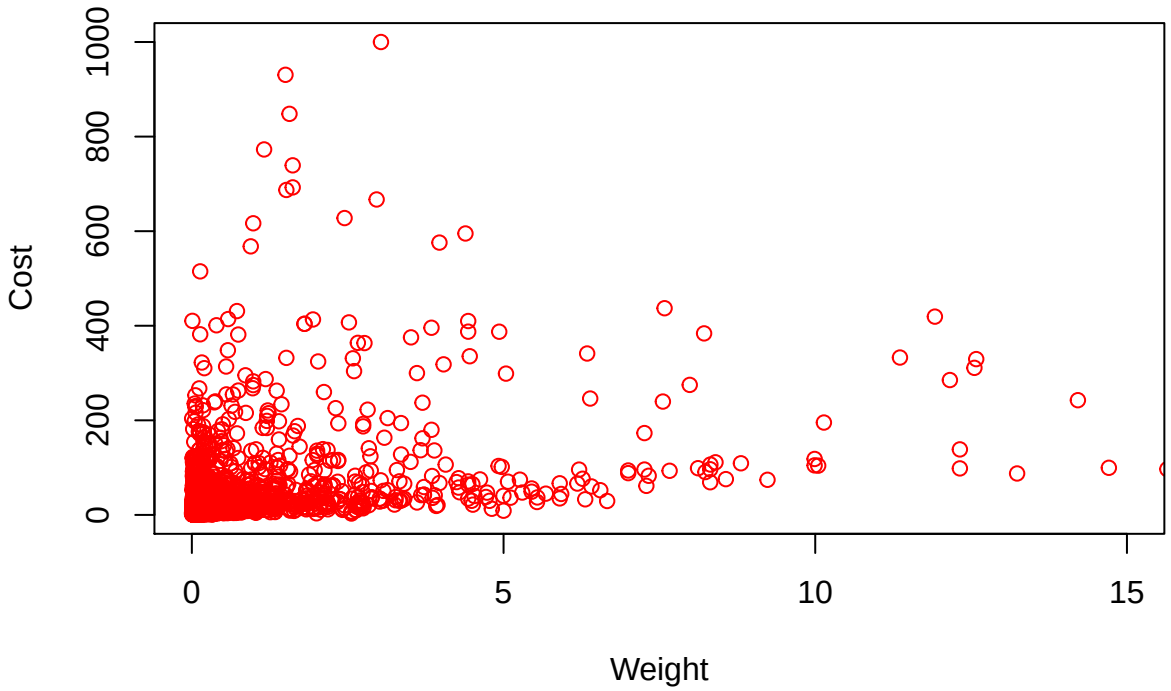
We use the Chi-Squared test to check if a feature is independent or dependent of the cost. We reject the hypothesis of independence if the p-value is greater than 0.05.

	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

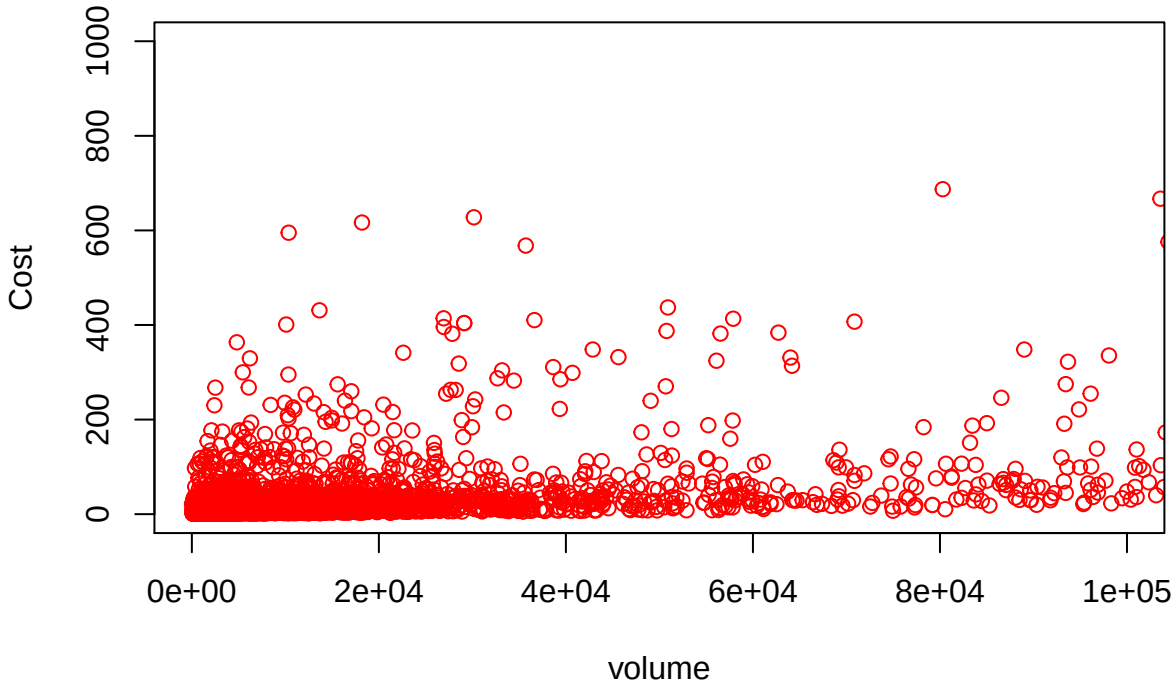
data: tbl X-squared = 5244100, df = 4262600, p-value < 2.2e-16



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

In this section, we will answer the question: *Do the costs presented in the training set are the real costs or do they need to be adjusted?*

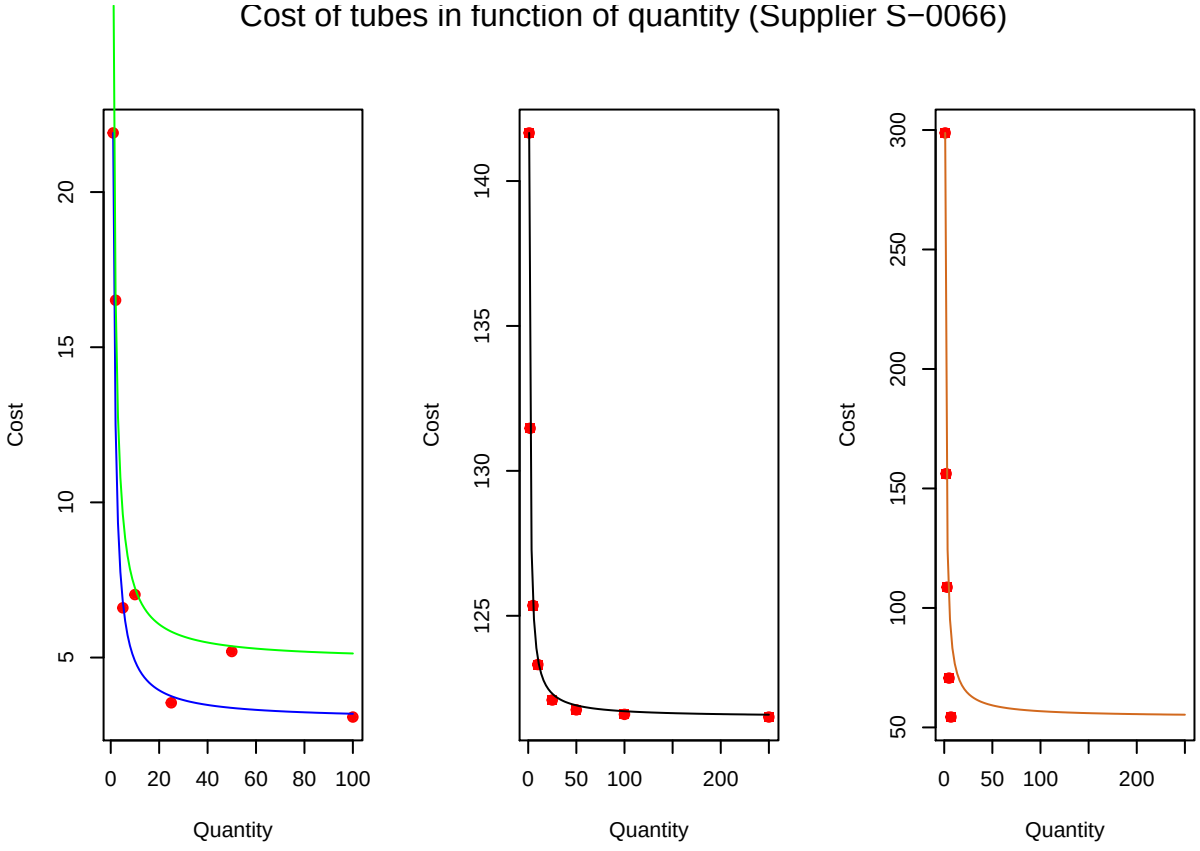
4 Supplier's Model

In this section, we will answer the question: *Is there a unique mathematical model describing the cost in function of the quantity for each supplier?* The first objective is to check the existence of a mathematical model representing the cost in function of the quantity. The second objective is to show if the model is applied by a unique supplier. The last objective is to show if each supplier has its own model. If the unicity does not hold, then we have to check if a model is applied by more than one supplier or if a supplier can apply more than one model depending of other features.

We denote $C_\beta(Q)$ our cost heuristic function of a tube assembly given by a supplier where β is our learning parameters and Q the vector of quantities.

4.1 Existence of a Mathematical Model

We start with the few tube assemblies which are quoted by the 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	5	S-0066	1	28.37
10	5	S-0066	2	16.51
11	5	S-0066	5	9.40
12	5	S-0066	10	7.03
13	5	S-0066	25	5.60
14	5	S-0066	50	5.19
15	5	S-0066	100	5.01
16	5	S-0066	250	4.90
17	5000	S-0066	1	141.66
18	5000	S-0066	2	131.47
19	5000	S-0066	5	125.35
20	5000	S-0066	10	123.32
21	5000	S-0066	25	122.09
22	5000	S-0066	50	121.75
23	5000	S-0066	100	121.60
24	5000	S-0066	250	121.51
25	19365	S-0066	1	298.78
26	19365	S-0066	2	156.20
27	19365	S-0066	3	108.67
28	19365	S-0066	5	70.64
29	19365	S-0066	7	54.35

Table 3: Table built from tubes 2, 5, 5000, 19365

From the graphs, we see that the curves estimating the red points are clearly hyperbolas of equation

$$C_T(Q) = \frac{\beta_0 - \beta_1}{Q} + \beta_1$$

where $Q \geq 1$ is the quantity for a tube assembly ID T, β_1 is the cost at the last level of purchase based on quantity and supplier (most of the time $Q = 250$), and β_0 is the cost at the first level of purchase based on quantity and supplier (most of the time $Q = 1$). This equation indicates that if Caterpillar buy more tubes, cheaper will be the cost per tube. This proves the existence of a mathematical model representing the cost in function of the quantity.

If we take a look at the right most graph, we see that our curve doesn't seem to fit the points. However, the maximum quantity is 7 (not 250) for this tube which make the model less accurate assuming the same model is used. This assumption make sense since

$$\lim_{Q \to \infty} C_T(Q) = \beta_1$$

which means that we need to find the right β_1 to match with any quantity. We also have to find the cost of one tube which is β_0.

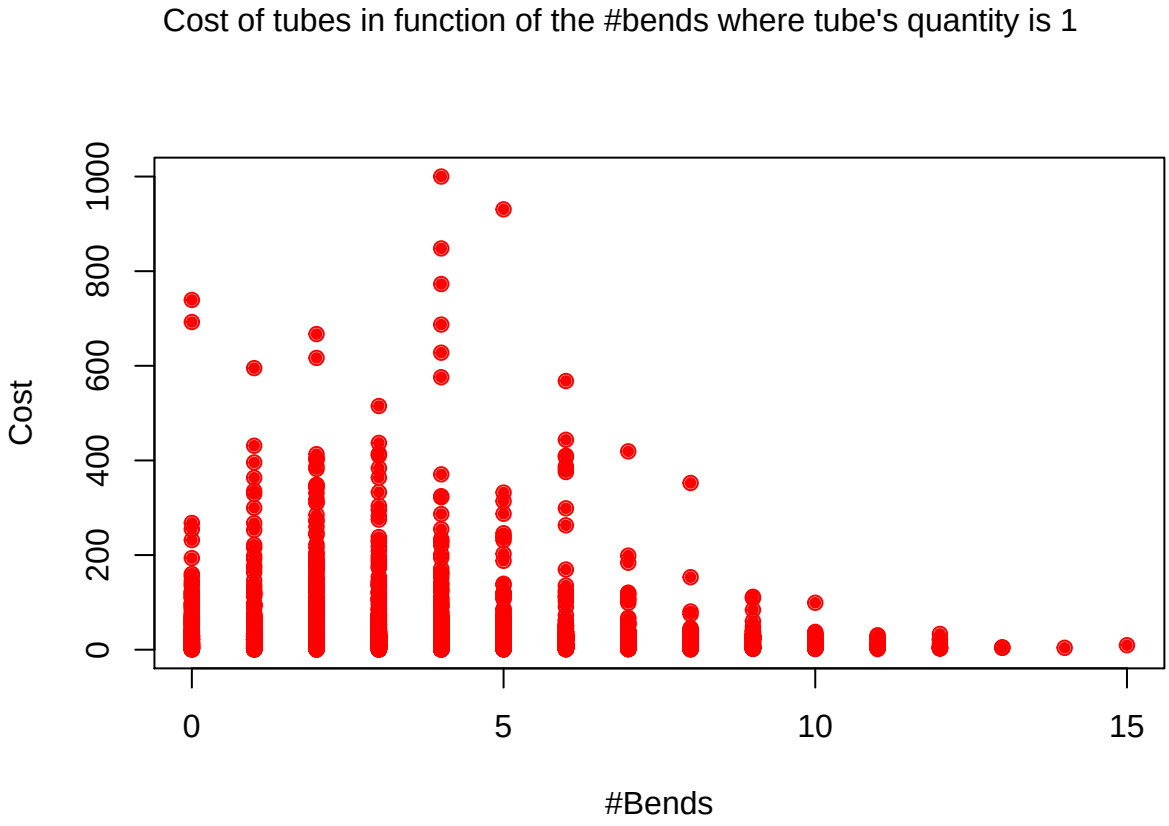
For example, if we take the tube `TA-19365`, we have $C_T(1) = \beta_0 = 298.7820145446$. We know that $C_T(2) = \frac{\beta_0+\beta_1}{2} = 156.1959237271 \Leftrightarrow \beta_1 = 13.60983291$. Therefore, the model for the tube `TA-19365` is $C_T(Q) = \frac{285.172181635}{Q} + 13.60983291$. With $Q = 7$, we obtain $C_T(7) = 54.348716001$ which has a square error

of 0.000001422 from the original cost. With our estimated, i.e. $C_T(Q) = (244.434490855/Q) + 54.3475236892$, we have $C_T(7) = 89.266736668$ which has a square error of 1219.351435073. Thus, we will gain accuracy by searching for the real value of β_1.

We start by analyzing on what features β_0 depends.

	fkTubeAssembly	supplierID	numberOfBends	cost
1	2	S-0066	8	21.91
2	4	S-0066	9	21.97
3	5	S-0066	4	28.37
4	12	S-0066	7	22.42
5	13	S-0026	3	10.00
6	14	S-0066	4	21.99
7	21	S-0030	6	3.43
8	22	S-0013	6	8.56

Table 4: First tubes where quantity is 1

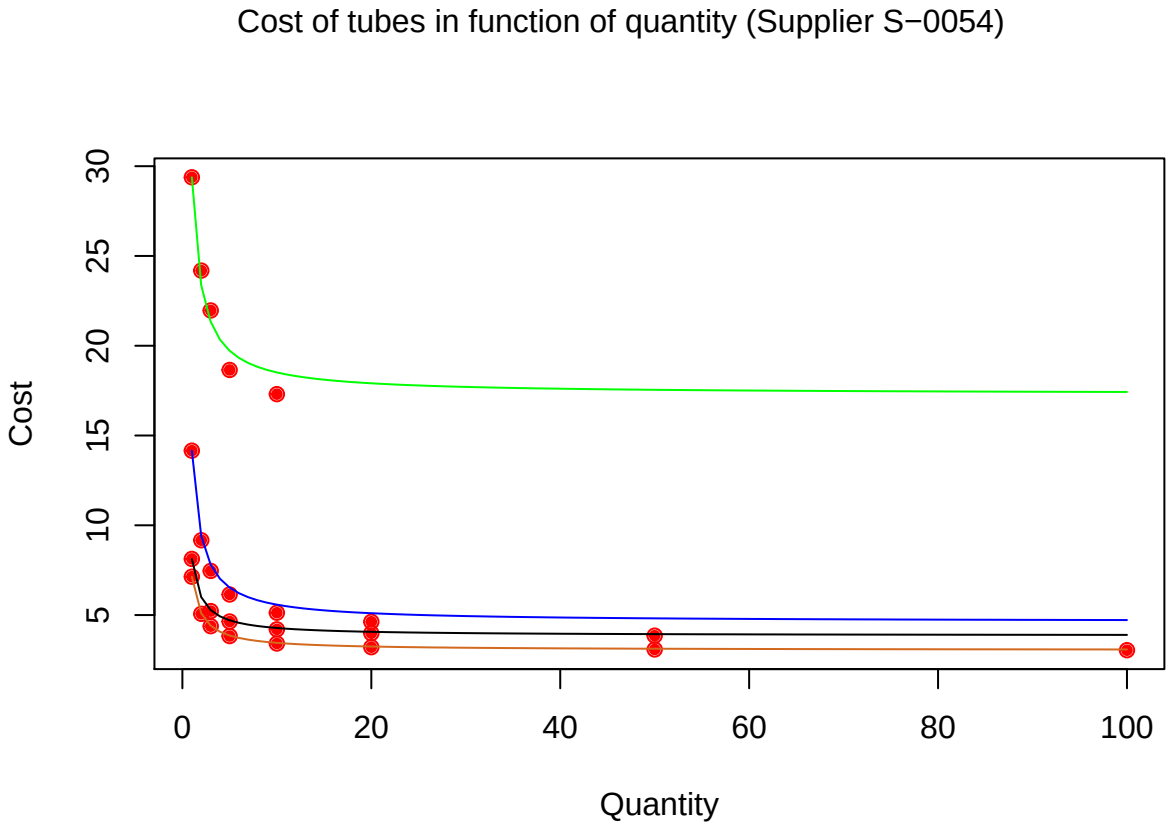


4.2 Unicity of the Model per Supplier

We verify with few tube assemblies which are quoted by the supplier `S-0054` if the same model used for the supplier `S-0066` applies.

	fkTubeAssembly	supplierID	quantity	cost
1	130	S-0054	1	14.16
2	130	S-0054	2	9.17
3	130	S-0054	3	7.46
4	130	S-0054	5	6.15
5	130	S-0054	10	5.13
6	130	S-0054	20	4.63
7	280	S-0054	1	29.38
8	280	S-0054	2	24.18
9	280	S-0054	3	21.96
10	280	S-0054	5	18.65
11	280	S-0054	10	17.30
12	1892	S-0054	1	8.13
13	1892	S-0054	3	5.22
14	1892	S-0054	5	4.64
15	1892	S-0054	10	4.20
16	1892	S-0054	20	3.99
17	1892	S-0054	50	3.85
18	5013	S-0054	1	7.14
19	5013	S-0054	2	5.07
20	5013	S-0054	3	4.38
21	5013	S-0054	5	3.83
22	5013	S-0054	10	3.42
23	5013	S-0054	20	3.21
24	5013	S-0054	50	3.09
25	5013	S-0054	100	3.04

Table 5: Table built from Tube 130, 280, 1892, 5013



The model used by the supplier `S-0054` seems to be the same as the one used by the supplier `S-0066`, but if we look carefully the curves, we see that greater is the quantity, more accurate is the estimate. This means that the model follows the same behaviour as both as the model used by the supplier `S-0066`. This shows that a model can be used by more than one supplier.

5 Classified Features

6 Anomalies Detection