General Observations on Train Set

The first objective is to find all features on which the cost depends from the train set. Then, we find which machine learning algorithms will apply depending of these features.

Let's define the features used in the training set. The description is given in the codebook.

Feature	Variable
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

Let's denote $h_\beta(x)$ our cost heuristic function of a tube assembly given by a supplier where β 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.

Per the codebook, x_6 determines on which features the cost depends. We will use 2 classes to base our cost estimation.

1. Bracket pricing ($x_6 = 1$ (Yes)) where the function $C(x)$ depends of x_7 amoung other features.
2. Non-bracket pricing ($x_6 = 0$ (No)) where the function $C(x)$ depends of x_5 and x_7.

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

- The weight of the tube
- The quantity of tubes
- The volume of the tube
- The material used to make the tube which can be referred as the density
- 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.

Other factors than the tube properties may also be helpful.

- The date when the supplier has quoted the price. The price is certainly less 20 years ago than today.
- The supplier may use different mathematical models to quote his price. We can then classify the data by supplier.
- The tube assembly ID may be used at some points in the prediction. Need to investigate why specifically these tubes are chosen in the train set.

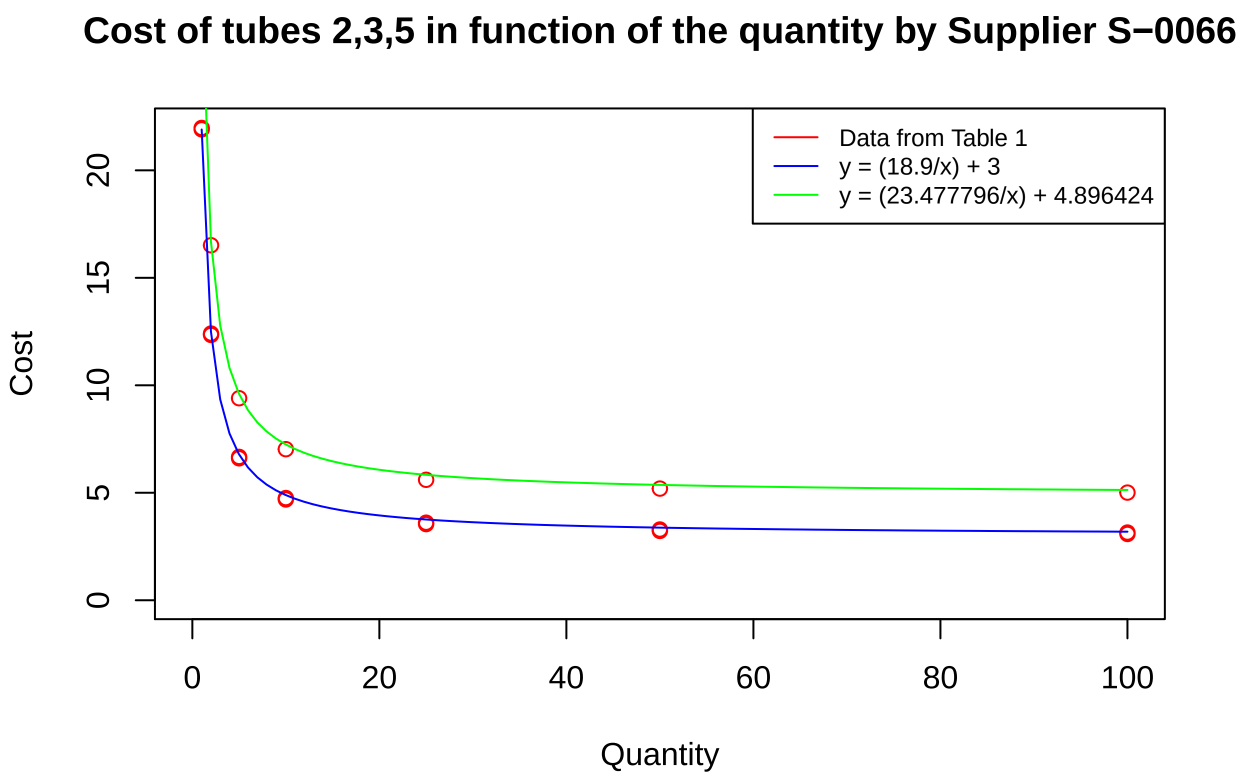
From the dataset, we note that there are a total of 2048 components. These components are spread amoung the `comp_[type].csv` files uniquely. This means that we can create a single table `Component` by merging those files together with the `merge` function. To avoid to many columns, we will remove some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

Simple Test on Comparing Quantities and Costs

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.

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

fkTubeAssembly	supplierID	quantity	cost
2	S-0066	1	21.905933
2	S-0066	2	12.341214
2	S-0066	5	6.601826
2	S-0066	10	4.687769
2	S-0066	25	3.541561
2	S-0066	50	3.224406
2	S-0066	100	3.082521
2	S-0066	250	2.999060
4	S-0066	1	21.972702
4	S-0066	2	12.407983
4	S-0066	5	6.668596
4	S-0066	10	4.754539
4	S-0066	25	3.608331
4	S-0066	50	3.291176
4	S-0066	100	3.149291
4	S-0066	250	3.065829
5	S-0066	1	28.374220
5	S-0066	2	16.514303
5	S-0066	5	9.397796
5	S-0066	10	7.027481
5	S-0066	25	5.603067
5	S-0066	50	5.194104
5	S-0066	100	5.007706
5	S-0066	250	4.896424



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 $x_7 \geq 1$, β_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 the company buy more tubes, cheaper will be the cost per tube by the supplier `S-0066`.

We need to find on which features depend β_0 and β_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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 525269 28.1     940480 50.3   743624 39.8
## Vcells 707129  5.4    1308461 10.0  1080462  8.3
```

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.