# 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 |

Since we are predicting the cost value, a continuous variable, we will use a regression algorithm.

## 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.
- The number of components to assemble a 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. Is there a unique mathematical model describing the cost in function of the quantity for each supplier?
3. If a mathematical model exists, is it a linear or non-linear model?
4. Are there decisions to take? If yes, what decisions?

# 2 Preparing & Cleaning the Dataset

In this section, we will answer the 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
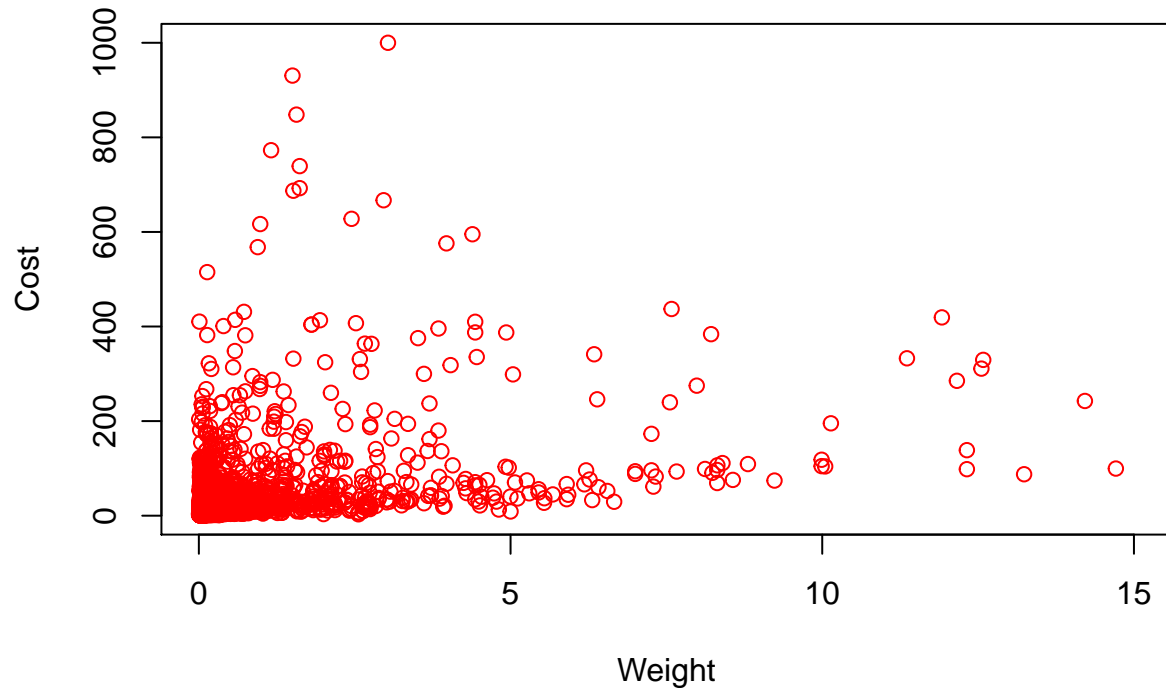


Following the Chi-Square test done and the plot, the cost is independent of the weight of a tube since the p-value is less than 0.05.

Pearson's Chi-squared test

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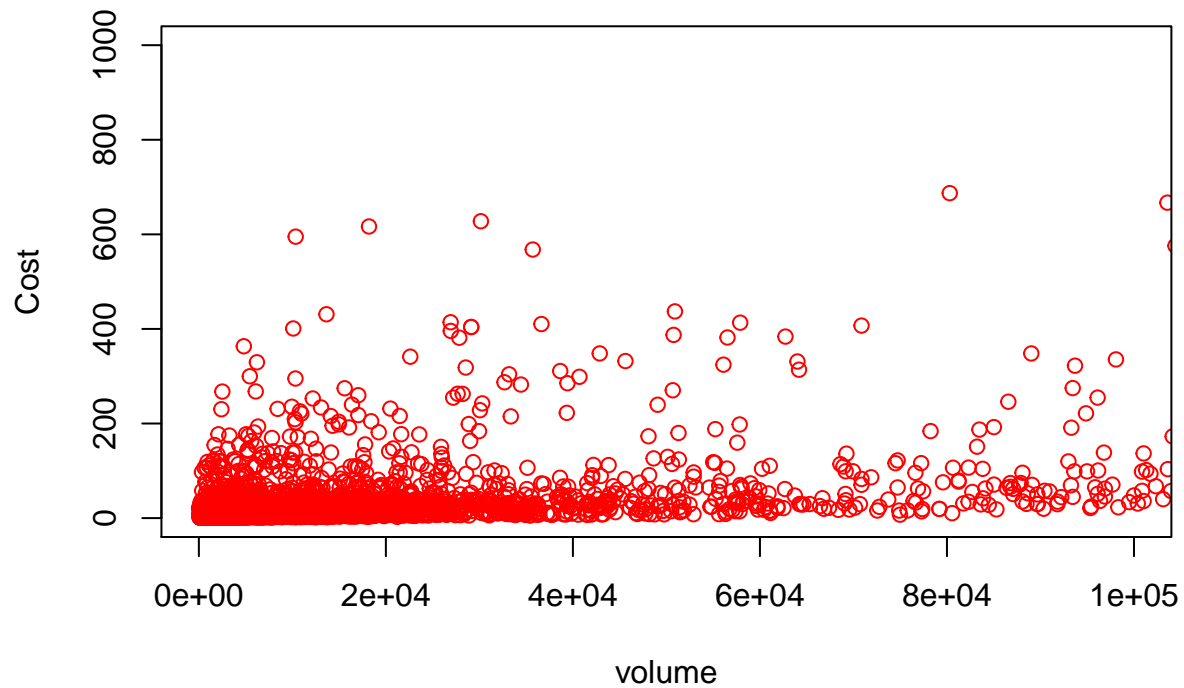


Following the Chi-Square test done and the plot, the cost is independent of the volume of a tube since the p-value is less than 0.05.

```
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 Mathematical Models

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 will then answer the question: *If a mathematical model exists, is it a linear or non-linear model?*

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

### 3.1 Existence of a Mathematical Model

We start with the few tube assemblies which are quoted by the supplier `S-0066`.



Cost of tubes in function of quantity (Supplier S−0066)
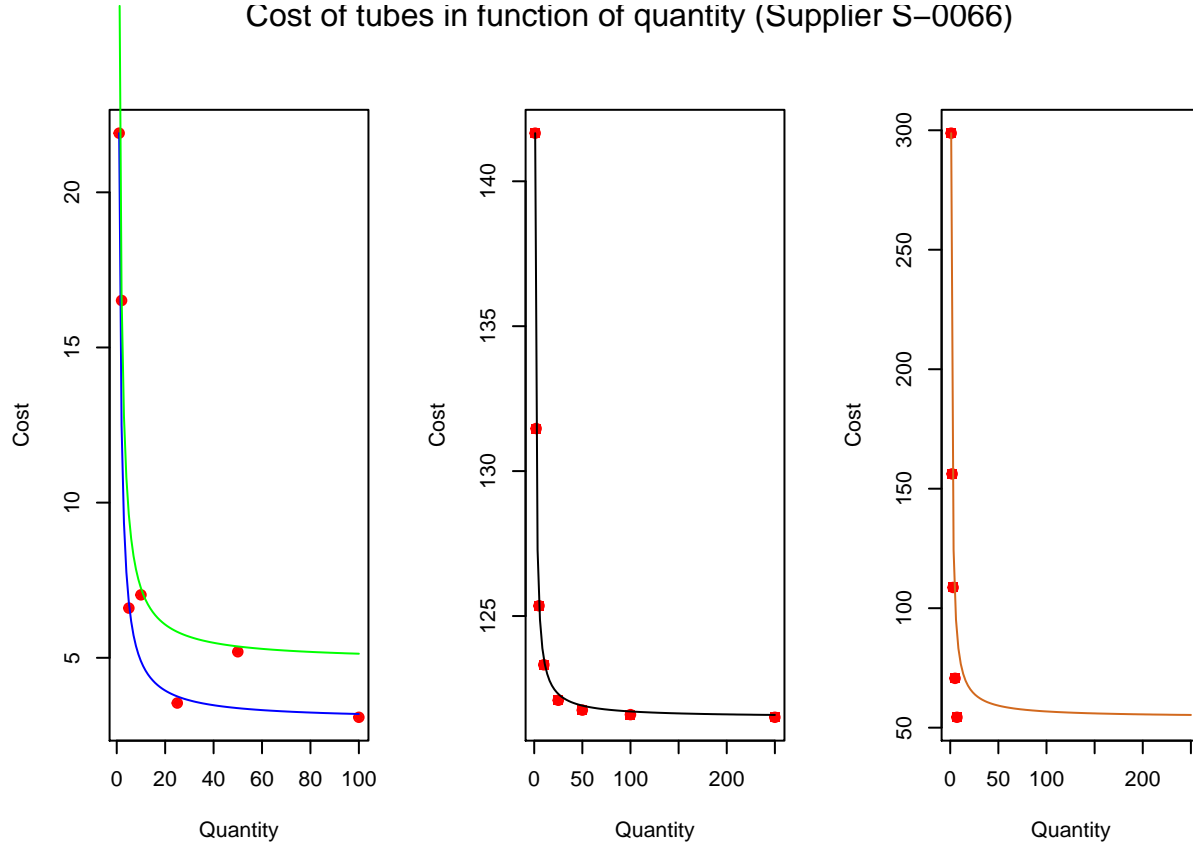
From the graphs, we see that the curves estimating the red points are clearly hyperbolas of equation

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

| | 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

where $Q \geq 1$ is the quantity for a tube assembly ID $T$, $\beta_1$ is the cost at the last level of purchase based on quantity and supplier (most of the time $Q = 250$), and $\beta_0$ 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 makes sense since

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

which means that we need to find the right $\beta_1$ to match with any quantity. We also have to find the cost of one tube which is $\beta_0$.

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, if $Q$ is small (say $Q < 25$), the model may underfit.

## 3.2 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 4: Table built from Tube 130, 280, 1892, 5013

Cost of tubes in function of quantity (Supplier S–0054)
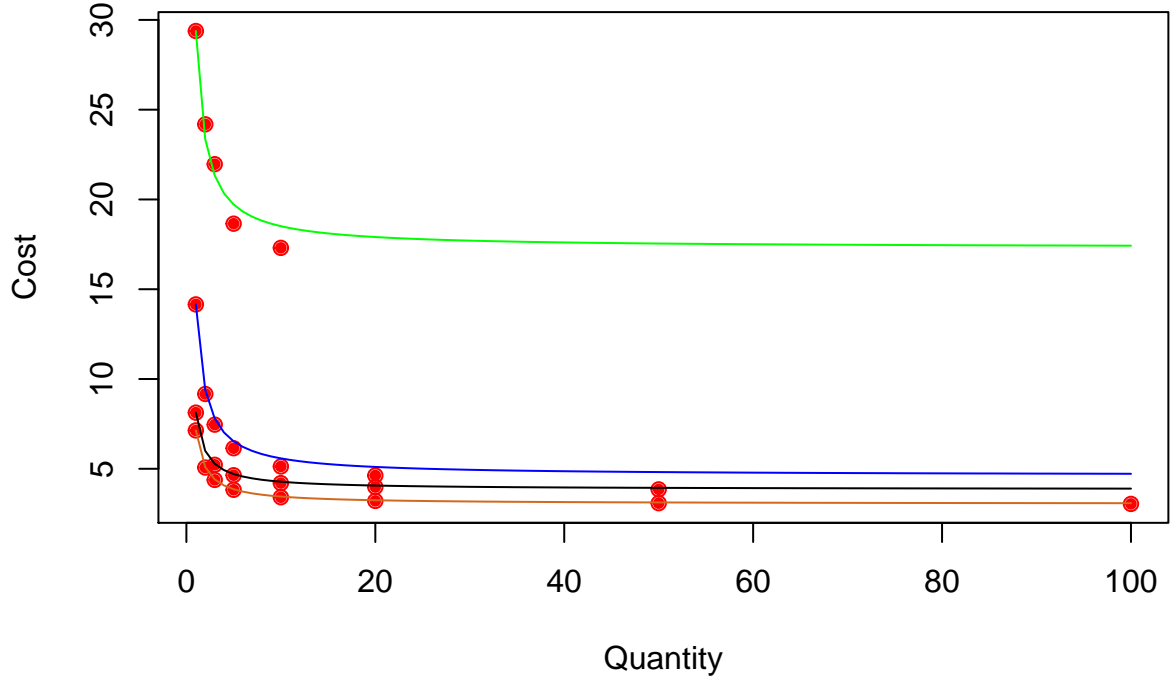
The model used by the supplier `S-0054` seems to be the same as the one used by the supplier `S-0066`, but if we look carefully the curves, we see that greater is the quantity, more accurate is the estimate. This means that the model follows the same behaviour as the model used by the supplier `S-0066`.

This doesn't seem to be the case for the tube `TA-00384` from the supplier `S-0064`. This supplier provides 6 levels of purchase where the highest quantity is $Q = 45$. We use the same model as before but this time, the model doesn't fit the points.

|   | fkTubeAssembly | supplierID | minOrderQuantity | anualUsage | quantity | cost |
|---|---|---|---|---|---|---|
| 1 | 384 | S-0064 | 0 | 0 | 20 | 5.29 |
| 2 | 384 | S-0064 | 0 | 0 | 25 | 4.97 |
| 3 | 384 | S-0064 | 0 | 0 | 30 | 4.75 |
| 4 | 384 | S-0064 | 0 | 0 | 35 | 4.60 |
| 5 | 384 | S-0064 | 0 | 0 | 40 | 4.49 |
| 6 | 384 | S-0064 | 0 | 0 | 45 | 4.40 |

Table 5: Tube 384

### Cost in function of quantity for tube 384 of supplier S−0064
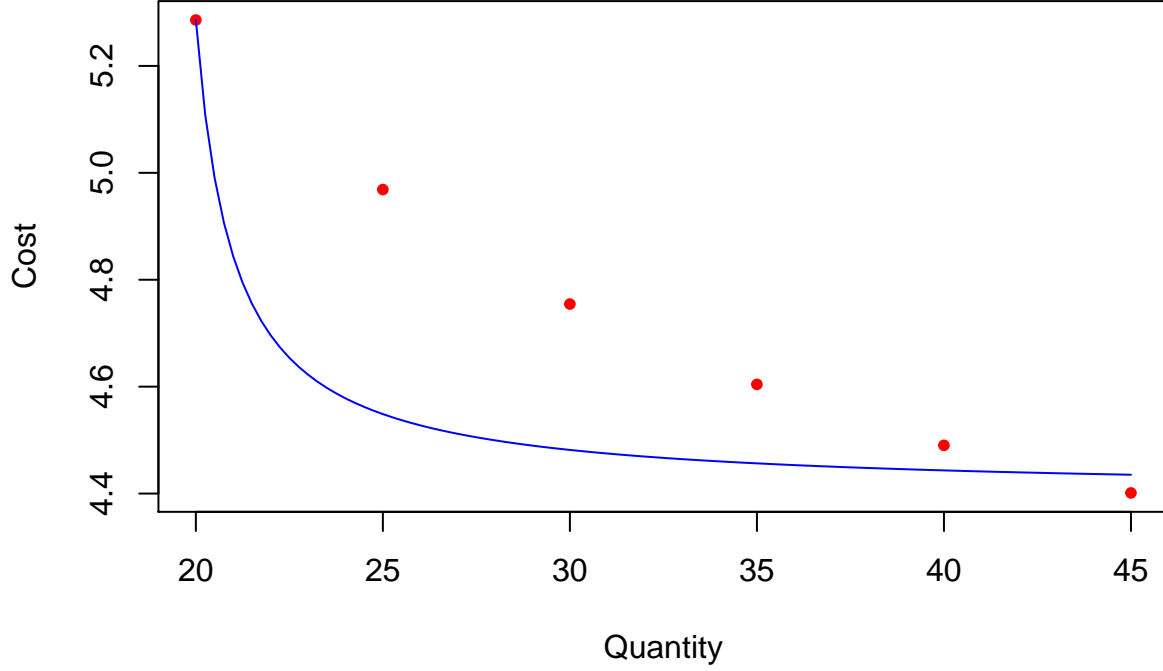
Since the first quantity level is 20, we need to translate the model by subtracting $x$ by $Q_0 - 1 = 19$. This gives the following model

$$C_T(Q) = \frac{\beta_0 - \beta_1}{Q - Q_0 - 1} + \beta_1$$

if $Q > 1$. However, the model still underfits the data because the cost decreases much slower than the model used for our previous tests. Therefore, we can assume that a model can be used to estimate the cost by one or many suppliers but not all.

## 4 Decision Tree(s)

In this section, we will identify conditional paths which will tell us if decision trees will be useful or not. We will answer the question: *Are there decisions to take? If yes, what decisions?* In the previous section, we have seen that the model can underfit if there are not enough quantity purchase levels and if the quantity is small. Otherwise, we can use the model to estimate the cost given a quantity and a tube assembly. Here are few points that identify some conditions.

- If there is only one quantity purchase level, we cannot estimate the cost. We need al least another feature on which the cost depends.
- If there are many quantity purchase levels but with $Q_0 > 1$, then we need to translate the model found at section 3.
- If the quantities are too low, then the model underfits. Thus, we need to add other cost-dependent features.
- Depending of the supplier, the model may be different. Since we have 57 suppliers in the train set, we can have at most 57 possible models.

Only with those conditions, we can build many decision trees to help us to estimate the cost.

# 5 Machine Learning Algorithms and Results

In this section, we present the machine learning algorithms we will try after the analysis given by the four previous sections. Per the section 4 (Decisions), we have seen that many decision trees can be built. This leaves us two possible algorithms: Random Forest or Boosting Trees for Regression (XGBoost).

## 5.1 Random Forest Algorithm

```
|       Out-of-bag    |
```

Tree | MSE %Var(y) | 2 | 0.6696 98.80 | 4 | 0.5368 79.21 | 6 | 0.5618 82.90 | 8 | 0.52 76.73 | 10 | 0.4755 70.17 | 12 | 0.4986 73.57 | 14 | 0.4682 69.08 | 16 | 0.4431 65.39 | 18 | 0.4283 63.20 | 20 | 0.4252 62.74 | 22 | 0.4227 62.37 | 24 | 0.4186 61.76 | 26 | 0.4022 59.34 | 28 | 0.3993 58.92 | 30 | 0.3947 58.24 | 32 | 0.3983 58.77 | 34 | 0.401 59.17 | 36 | 0.4018 59.29 | 38 | 0.397 58.58 | 40 | 0.3995 58.95 | 42 | 0.3932 58.01 | 44 | 0.4016 59.25 | 46 | 0.4094 60.41 | 48 | 0.4111 60.67 | 50 | 0.4126 60.88 | 52 | 0.4159 61.37 | 54 | 0.4143 61.13 | 56 | 0.4077 60.16 | 58 | 0.4104 60.56 | 60 | 0.4161 61.39 | 62 | 0.4181 61.69 | 64 | 0.4236 62.50 | 66 | 0.4263 62.91 | 68 | 0.4315 63.66 | 70 | 0.4258 62.83 | 72 | 0.4259 62.84 | 74 | 0.4285 63.22 | 76 | 0.4291 63.31 | 78 | 0.4318 63.71 | 80 | 0.4318 63.72 | 82 | 0.4361 64.34 | 84 | 0.4322 63.77 | 86 | 0.4277 63.11 | 88 | 0.4286 63.25 | 90 | 0.4281 63.17 | 92 | 0.4258 62.83 | 94 | 0.4239 62.55 | 96 | 0.4266 62.94 | 98 | 0.4224 62.32 | 100 | 0.4253 62.75 | 102 | 0.4249 62.70 | 104 | 0.4244 62.63 | 106 | 0.4229 62.39 | 108 | 0.422 62.26 | 110 | 0.4242 62.59 | 112 | 0.4263 62.90 | 114 | 0.4263 62.90 | 116 | 0.423 62.41 | 118 | 0.4232 62.45 | 120 | 0.4233 62.46 | 122 | 0.4219 62.25 | 124 | 0.4214 62.18 | 126 | 0.4243 62.61 | 128 | 0.4257 62.81 | 130 | 0.4259 62.84 | 132 | 0.426 62.86 | 134 | 0.4241 62.57 | 136 | 0.4232 62.44 | 138 | 0.4227 62.38 | 140 | 0.4232 62.44 | 142 | 0.4247 62.67 | 144 | 0.425 62.72 | 146 | 0.4275 63.08 | 148 | 0.4279 63.14 | 150 | 0.428 63.16 | 152 | 0.4259 62.84 | 154 | 0.4267 62.96 | 156 | 0.427 63.01 | 158 | 0.4272 63.03 | 160 | 0.4274 63.06 | 162 | 0.4273 63.05 | 164 | 0.4278 63.12 | 166 | 0.4299 63.44 | 168 | 0.4301 63.46 | 170 | 0.4291 63.31 | 172 | 0.4288 63.28 | 174 | 0.4294 63.36 | 176 | 0.4293 63.34 | 178 | 0.4304 63.50 | 180 | 0.4293 63.34 | 182 | 0.4302 63.48 | 184 | 0.4298 63.42 | 186 | 0.4278 63.13 | 188 | 0.4289 63.29 | 190 | 0.4306 63.53 | 192 | 0.4324 63.81 | 194 | 0.4326 63.84 | 196 | 0.4316 63.69 | 198 | 0.4321 63.76 | 200 | 0.4306 63.54 | 202 | 0.4291 63.32 | 204 | 0.4293 63.35 | 206 | 0.43 63.44 | 208 | 0.4296 63.39 | 210 | 0.4313 63.64 | 212 | 0.4319 63.73 | 214 | 0.4322 63.77 | 216 | 0.4322 63.77 | 218 | 0.4319 63.72 | 220 | 0.4319 63.73 | 222 | 0.4325 63.81 | 224 | 0.4316 63.68 | 226 | 0.4319 63.73 | 228 | 0.4317 63.70 | 230 | 0.431 63.59 | 232 | 0.432 63.74 | 234 | 0.4306 63.54 | 236 | 0.4303 63.49 | 238 | 0.429 63.30 | 240 | 0.4296 63.39 | 242 | 0.4301 63.46 | 244 | 0.4311 63.61 | 246 | 0.4298 63.43 | 248 | 0.4292 63.33 | 250 | 0.4294 63.35 | 252 | 0.4308 63.56 | 254 | 0.4313 63.64 | 256 | 0.4314 63.65 | 258 | 0.4312 63.63 | 260 | 0.4324 63.81 | 262 | 0.4336 63.98 | 264 | 0.4326 63.83 | 266 | 0.4329 63.87 | 268 | 0.4342 64.07 | 270 | 0.4325 63.82 | 272 | 0.4324 63.80 | 274 | 0.4336 63.99 | 276 | 0.4348 64.15 | 278 | 0.4349 64.17 | 280 | 0.4334 63.95 | 282 | 0.432 63.74 | 284 | 0.432 63.74 | 286 | 0.4312 63.62 | 288 | 0.4303 63.49 | 290 | 0.4304 63.50 | 292 | 0.4303 63.49 | 294 | 0.4288 63.28 | 296 | 0.429 63.30 | 298 | 0.4293 63.34 | 300 | 0.4295 63.37 | 302 | 0.4299 63.43 | 304 | 0.4296 63.38 | 306 | 0.4284 63.21 | 308 | 0.4276 63.10 | 310 | 0.4278 63.12 | 312 | 0.4285 63.22 | 314 | 0.4295 63.37 | 316 | 0.4282 63.19 | 318 | 0.4277 63.11 | 320 | 0.4275 63.08 | 322 | 0.428 63.15 | 324 | 0.4287 63.26 | 326 | 0.4285 63.23 | 328 | 0.4296 63.39 | 330 | 0.4304 63.51 | 332 | 0.4313 63.64 | 334 | 0.4312 63.62 | 336 | 0.4322 63.78 | 338 | 0.432 63.74 | 340 | 0.4323 63.79 | 342 | 0.4334 63.95 | 344 | 0.4327 63.85 | 346 | 0.4331 63.90 | 348 | 0.4332 63.92 | 350 | 0.4319 63.72 | 352 | 0.4329 63.87 | 354 | 0.4338 64.01 | 356 | 0.4334 63.94 | 358 | 0.434 64.03 | 360 | 0.434 64.04 | 362 | 0.4343 64.08 | 364 | 0.4351 64.20 | 366 | 0.4361 64.35 | 368 | 0.4358 64.30 | 370 | 0.4358 64.30 | 372 | 0.4367 64.44 | 374 | 0.4376 64.57 | 376 | 0.4379 64.62 | 378 | 0.4386 64.72 | 380 | 0.4381 64.64 | 382 | 0.4384 64.69 | 384 | 0.4386 64.71 | 386 | 0.4386 64.71 | 388 | 0.4389 64.76 | 390 | 0.4388 64.75 | 392 | 0.4395 64.86 | 394 | 0.4385 64.71 | 396 | 0.4383 64.67 | 398 | 0.4383 64.67 | 400 | 0.4384 64.68 | 402 | 0.4384 64.69 | 404 | 0.4383 64.67 | 406 | 0.4384 64.68 | 408 | 0.4377 64.59 | 410 | 0.4386 64.72 | 412 | 0.4383 64.68 | 414 | 0.4385 64.70 | 416 | 0.4375 64.56 | 418 | 0.4376 64.57 | 420 | 0.4384 64.68 | 422 | 0.4379 64.62 | 424 | 0.4381 64.64 | 426 | 0.4372 64.51 | 428 | 0.4361 64.35 | 430 | 0.4352 64.21 | 432 | 0.4353 64.23 | 434 | 0.4348 64.16 | 436 | 0.4355 64.26 | 438 | 0.4359 64.32 | 440 | 0.4357 64.29 | 442 | 0.4352 64.22 | 444 | 0.4354 64.24 | 446 | 0.4355 64.26 | 448 | 0.4356 64.28 | 450 | 0.4353 64.23 | 452 | 0.4344 64.10 | 454 | 0.4343 64.08 | 456 | 0.4344 64.10 | 458 | 0.4346 64.12 | 460 | 0.4345 64.11 | 462 | 0.4349 64.17 | 464 | 0.4352 64.21 | 466 | 0.4353 64.23 | 468 | 0.436 64.33 | 470 | 0.4367

64.44 | 472 | 0.4369 64.46 | 474 | 0.4373 64.52 | 476 | 0.4374 64.54 | 478 | 0.437 64.49 | 480 | 0.4371 64.50 | 482 | 0.4379 64.61 | 484 | 0.4377 64.59 | 486 | 0.4384 64.69 | 488 | 0.4385 64.71 | 490 | 0.4392 64.80 | 492 | 0.4388 64.74 | 494 | 0.4393 64.82 | 496 | 0.4395 64.85 | 498 | 0.4393 64.83 | 500 | 0.4386 64.72 |