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.

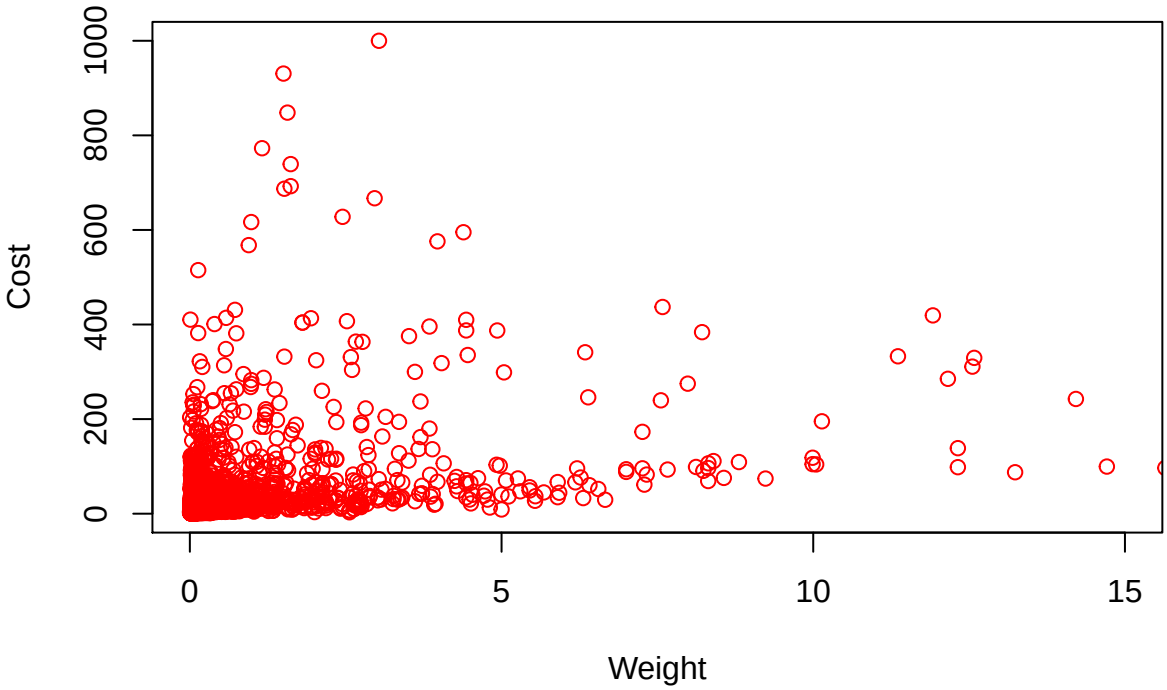
We use the Chi-Squared test to check if a feature is independent or dependent of the cost. We reject the hypothesis of independence if the p-value is greater than 0.05.

	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

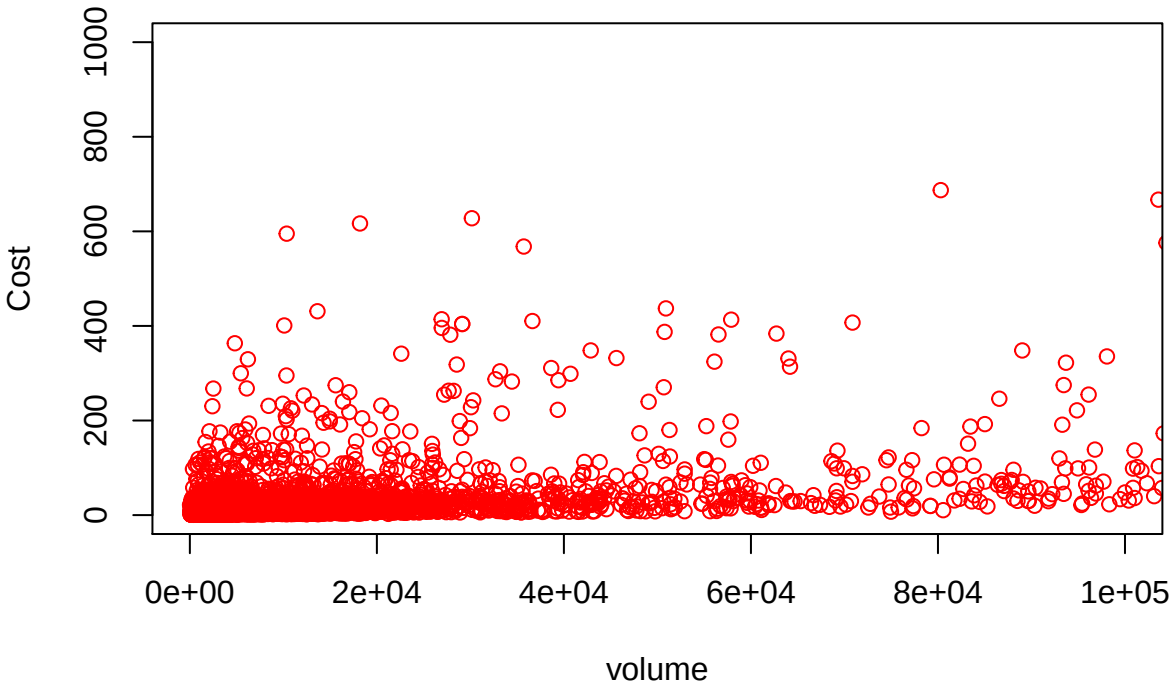
data: tbl X-squared = 5244100, df = 4262600, p-value < 2.2e-16



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.

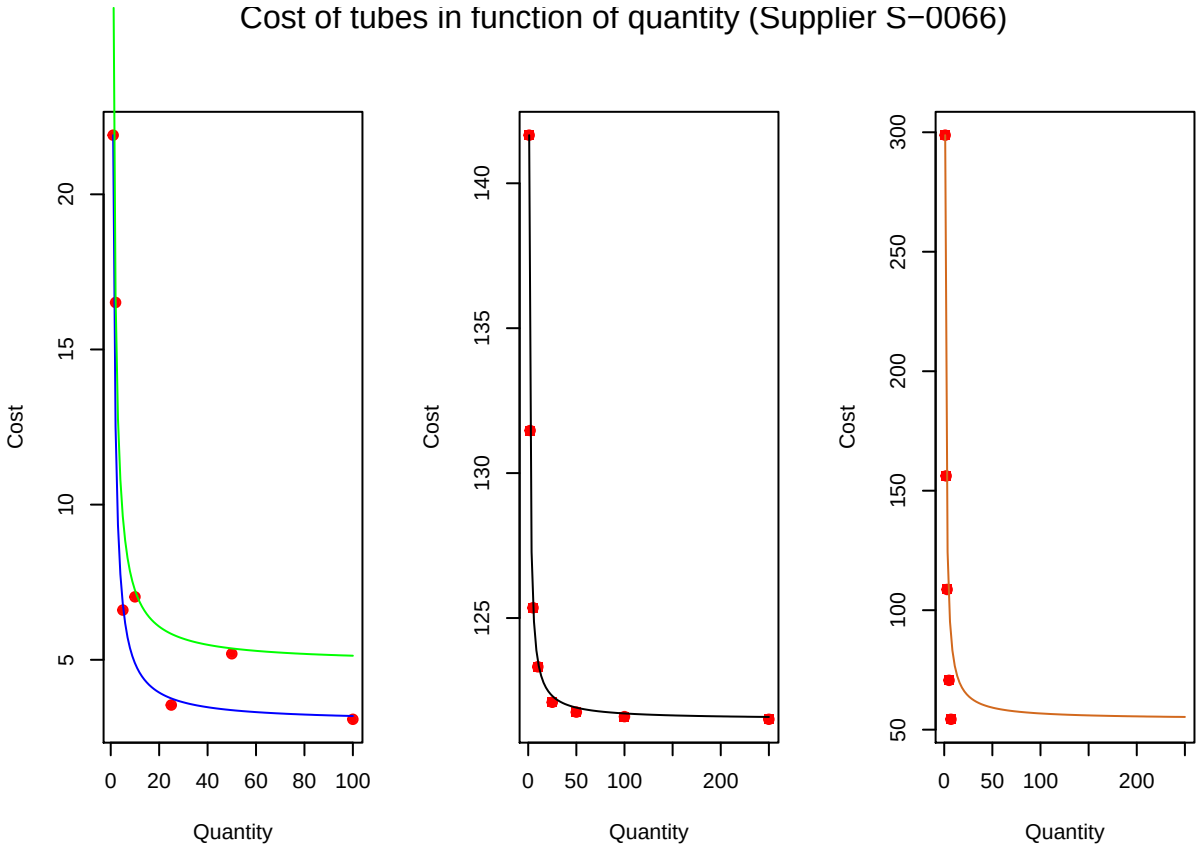
3 Mathematical Models

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 β 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`.



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, 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 makes 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.

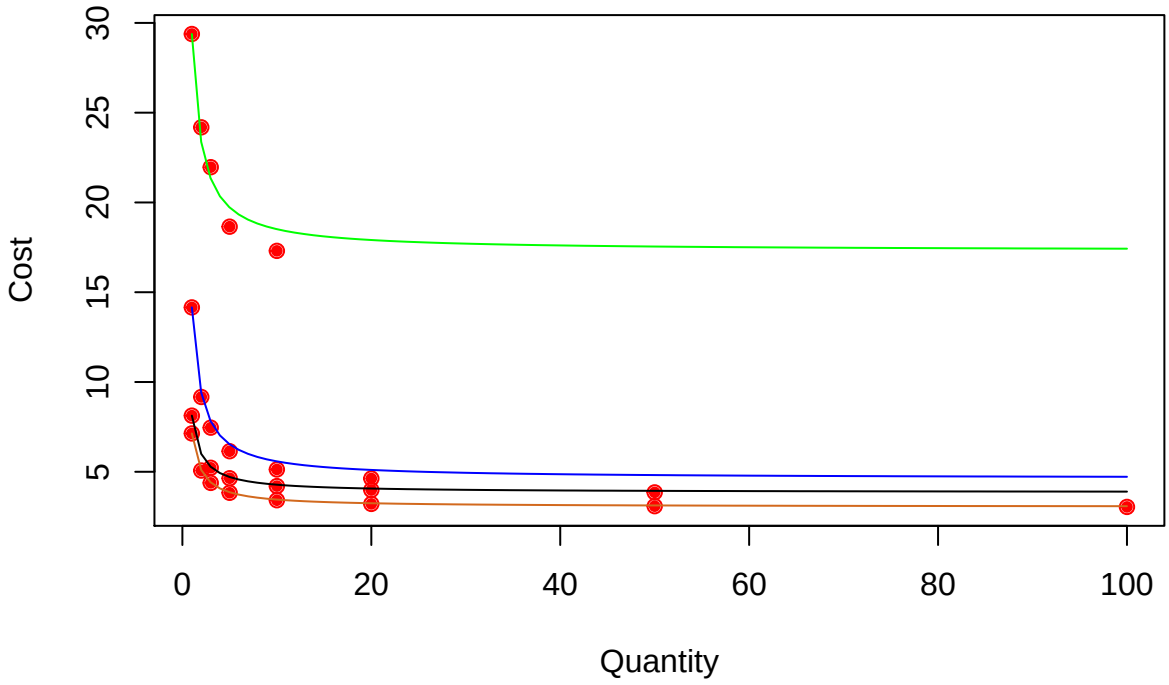
For example, if we take the tube `TA-19365`, we have $C_T(1) = \beta_0 = 298.7820145446$. We know that $C_T(2) = \frac{\beta_0+\beta_1}{2} = 156.1959237271 \Leftrightarrow \beta_1 = 13.60983291$. Therefore, the model for the tube `TA-19365` is $C_T(Q) = \frac{285.172181635}{Q} + 13.60983291$. With $Q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

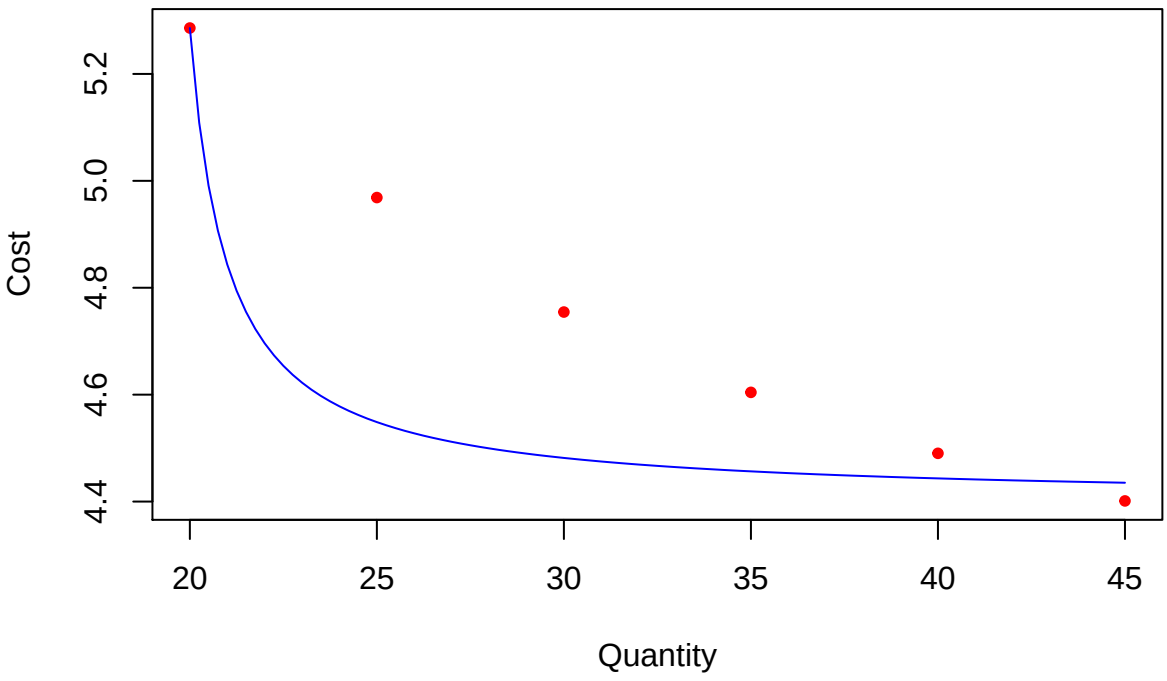


The model used by the supplier `S-0054` seems to be the same as the one used by the supplier `S-0066`, but if we look carefully the curves, we see that greater is the quantity, more accurate is the estimate. This means that the model follows the same behaviour as the model used by the supplier `S-0066`.

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



used (Mb) gc trigger (Mb) max used (Mb) Ncells 540547 28.9 940480 50.3 940480 50.3 Vcells 804461 6.2 150348406 1147.1 187925565 1433.8

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 |   0.2443    36.05 |
##    4 |   0.2327    34.34 |
##    6 |   0.2133    31.47 |
##    8 |   0.1917    28.29 |
##   10 |   0.1785    26.34 |
##   12 |   0.1696    25.02 |
##   14 |   0.1623    23.94 |
##   16 |   0.1589    23.44 |
##   18 |    0.156    23.02 |
##   20 |   0.1539    22.71 |
##   22 |   0.1518    22.40 |
##   24 |   0.1504    22.19 |
##   26 |   0.1487    21.93 |
##   28 |   0.1471    21.71 |
##   30 |   0.1452    21.42 |
##   32 |   0.1441    21.27 |
##   34 |   0.1435    21.17 |
##   36 |    0.143    21.11 |
##   38 |   0.1424    21.01 |
##   40 |   0.1418    20.92 |
##   42 |   0.1408    20.78 |
##   44 |   0.1403    20.70 |
##   46 |   0.1402    20.69 |
##   48 |   0.1404    20.71 |
##   50 |   0.1403    20.70 |
##   52 |     0.14    20.65 |
##   54 |   0.1397    20.62 |
##   56 |   0.1395    20.58 |
##   58 |   0.1398    20.63 |
##   60 |   0.1391    20.52 |
##   62 |    0.139    20.50 |
##   64 |   0.1389    20.50 |
##   66 |   0.1388    20.49 |
##   68 |   0.1386    20.45 |
##   70 |   0.1386    20.46 |
##   72 |   0.1381    20.38 |
##   74 |   0.1383    20.41 |
##   76 |   0.1382    20.40 |
##   78 |   0.1382    20.39 |
##   80 |   0.1382    20.40 |
```

```
##   82 |   0.1382    20.39 |
##   84 |   0.1386    20.45 |
##   86 |   0.1384    20.42 |
##   88 |   0.1383    20.40 |
##   90 |   0.1384    20.42 |
##   92 |   0.1383    20.40 |
##   94 |   0.1381    20.38 |
##   96 |   0.1379    20.35 |
##   98 |   0.1377    20.32 |
##  100 |   0.1377    20.32 |
##  102 |   0.1376    20.31 |
##  104 |   0.1378    20.33 |
##  106 |   0.1375    20.30 |
##  108 |   0.1374    20.28 |
##  110 |   0.1374    20.28 |
##  112 |   0.1372    20.24 |
##  114 |   0.1373    20.26 |
##  116 |   0.1374    20.28 |
##  118 |   0.1375    20.28 |
##  120 |   0.1375    20.28 |
##  122 |   0.1374    20.28 |
##  124 |   0.1374    20.27 |
##  126 |   0.1376    20.30 |
##  128 |   0.1377    20.32 |
##  130 |   0.1377    20.32 |
##  132 |   0.1376    20.30 |
##  134 |   0.1376    20.31 |
##  136 |   0.1377    20.31 |
##  138 |   0.1375    20.28 |
##  140 |   0.1375    20.28 |
##  142 |   0.1374    20.28 |
##  144 |   0.1375    20.29 |
##  146 |   0.1374    20.28 |
##  148 |   0.1376    20.30 |
##  150 |   0.1375    20.29 |
##  152 |   0.1375    20.28 |
##  154 |   0.1375    20.29 |
##  156 |   0.1374    20.28 |
##  158 |   0.1375    20.28 |
##  160 |   0.1373    20.26 |
##  162 |   0.1372    20.25 |
##  164 |   0.1372    20.25 |
##  166 |   0.1372    20.25 |
##  168 |   0.1372    20.24 |
##  170 |   0.1374    20.27 |
##  172 |   0.1374    20.27 |
##  174 |   0.1372    20.24 |
##  176 |   0.1371    20.23 |
##  178 |    0.137    20.22 |
##  180 |   0.1369    20.20 |
##  182 |    0.137    20.22 |
##  184 |   0.1372    20.24 |
##  186 |   0.1372    20.24 |
##  188 |   0.1371    20.23 |
```

```
##  190 |   0.1371    20.23 |
##  192 |   0.1371    20.23 |
##  194 |    0.137    20.22 |
##  196 |    0.137    20.21 |
##  198 |   0.1369    20.20 |
##  200 |   0.1368    20.19 |
##  202 |   0.1367    20.18 |
##  204 |   0.1368    20.18 |
##  206 |   0.1367    20.17 |
##  208 |   0.1366    20.16 |
##  210 |   0.1365    20.14 |
##  212 |   0.1365    20.14 |
##  214 |   0.1365    20.14 |
##  216 |   0.1364    20.12 |
##  218 |   0.1363    20.12 |
##  220 |   0.1364    20.13 |
##  222 |   0.1364    20.13 |
##  224 |   0.1363    20.12 |
##  226 |   0.1363    20.11 |
##  228 |   0.1363    20.10 |
##  230 |   0.1364    20.12 |
##  232 |   0.1364    20.12 |
##  234 |   0.1363    20.11 |
##  236 |   0.1362    20.09 |
##  238 |   0.1362    20.10 |
##  240 |   0.1362    20.10 |
##  242 |   0.1362    20.10 |
##  244 |   0.1362    20.10 |
##  246 |   0.1362    20.10 |
##  248 |   0.1361    20.09 |
##  250 |    0.136    20.07 |
##  252 |    0.136    20.07 |
##  254 |   0.1359    20.05 |
##  256 |   0.1359    20.05 |
##  258 |   0.1359    20.06 |
##  260 |   0.1359    20.05 |
##  262 |   0.1359    20.05 |
##  264 |   0.1358    20.04 |
##  266 |   0.1358    20.04 |
##  268 |   0.1358    20.04 |
##  270 |   0.1358    20.03 |
##  272 |   0.1358    20.04 |
##  274 |   0.1358    20.04 |
##  276 |   0.1357    20.03 |
##  278 |   0.1356    20.01 |
##  280 |   0.1356    20.01 |
##  282 |   0.1357    20.02 |
##  284 |   0.1357    20.02 |
##  286 |   0.1356    20.01 |
##  288 |   0.1355    19.99 |
##  290 |   0.1355    19.99 |
##  292 |   0.1354    19.98 |
##  294 |   0.1356    20.01 |
##  296 |   0.1355    20.00 |
```

```
##  298 |   0.1356    20.01 |
##  300 |   0.1356    20.00 |
##  302 |   0.1355    20.00 |
##  304 |   0.1356    20.01 |
##  306 |   0.1357    20.02 |
##  308 |   0.1357    20.02 |
##  310 |   0.1356    20.01 |
##  312 |   0.1357    20.02 |
##  314 |   0.1357    20.02 |
##  316 |   0.1356    20.01 |
##  318 |   0.1356    20.01 |
##  320 |   0.1356    20.00 |
##  322 |   0.1356    20.00 |
##  324 |   0.1356    20.00 |
##  326 |   0.1356    20.00 |
##  328 |   0.1355    19.99 |
##  330 |   0.1355    20.00 |
##  332 |   0.1355    20.00 |
##  334 |   0.1355    19.99 |
##  336 |   0.1355    20.00 |
##  338 |   0.1355    20.00 |
##  340 |   0.1354    19.98 |
##  342 |   0.1354    19.98 |
##  344 |   0.1353    19.97 |
##  346 |   0.1354    19.97 |
##  348 |   0.1354    19.98 |
##  350 |   0.1354    19.98 |
##  352 |   0.1354    19.98 |
##  354 |   0.1354    19.97 |
##  356 |   0.1353    19.96 |
##  358 |   0.1353    19.96 |
##  360 |   0.1353    19.96 |
##  362 |   0.1353    19.96 |
##  364 |   0.1353    19.96 |
##  366 |   0.1352    19.95 |
##  368 |   0.1352    19.95 |
##  370 |   0.1352    19.95 |
##  372 |   0.1351    19.94 |
##  374 |   0.1351    19.94 |
##  376 |   0.1351    19.93 |
##  378 |    0.135    19.93 |
##  380 |   0.1351    19.93 |
##  382 |    0.135    19.92 |
##  384 |    0.135    19.93 |
##  386 |   0.1351    19.93 |
##  388 |   0.1351    19.94 |
##  390 |   0.1351    19.94 |
##  392 |   0.1352    19.94 |
##  394 |   0.1352    19.95 |
##  396 |   0.1352    19.94 |
##  398 |   0.1351    19.94 |
##  400 |   0.1351    19.93 |
##  402 |   0.1351    19.93 |
##  404 |   0.1351    19.93 |
```

```
##  406 |   0.1351    19.93 |
##  408 |    0.135    19.92 |
##  410 |    0.135    19.92 |
##  412 |    0.135    19.92 |
##  414 |    0.135    19.91 |
##  416 |    0.135    19.92 |
##  418 |    0.135    19.92 |
##  420 |    0.135    19.92 |
##  422 |    0.135    19.92 |
##  424 |   0.1349    19.91 |
##  426 |   0.1349    19.90 |
##  428 |   0.1349    19.90 |
##  430 |   0.1349    19.91 |
##  432 |    0.135    19.92 |
##  434 |    0.135    19.92 |
##  436 |    0.135    19.91 |
##  438 |    0.135    19.91 |
##  440 |    0.135    19.92 |
##  442 |    0.135    19.92 |
##  444 |   0.1349    19.91 |
##  446 |   0.1349    19.91 |
##  448 |   0.1349    19.91 |
##  450 |   0.1349    19.90 |
##  452 |    0.135    19.92 |
##  454 |    0.135    19.92 |
##  456 |   0.1349    19.91 |
##  458 |   0.1349    19.91 |
##  460 |   0.1349    19.91 |
##  462 |   0.1349    19.90 |
##  464 |   0.1348    19.89 |
##  466 |   0.1348    19.89 |
##  468 |   0.1348    19.89 |
##  470 |   0.1348    19.89 |
##  472 |   0.1348    19.89 |
##  474 |   0.1347    19.88 |
##  476 |   0.1347    19.88 |
##  478 |   0.1347    19.88 |
##  480 |   0.1347    19.88 |
##  482 |   0.1347    19.88 |
##  484 |   0.1347    19.88 |
##  486 |   0.1347    19.88 |
##  488 |   0.1347    19.87 |
##  490 |   0.1347    19.87 |
##  492 |   0.1346    19.86 |
##  494 |   0.1346    19.86 |
##  496 |   0.1346    19.86 |
##  498 |   0.1346    19.87 |
##  500 |   0.1346    19.87 |
```