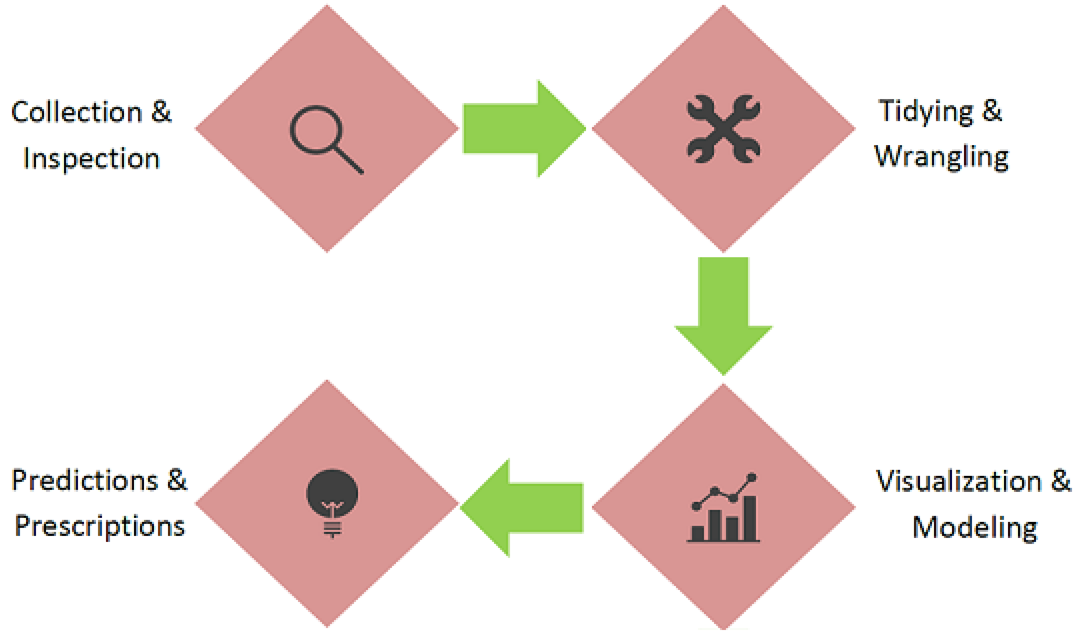
Sishen DMS Blockages - Exploratory Analytics

This document is intended to aid the exploration of the blockage data for the Sishen mine. The analysis is to provide insight that will guide the more efficient root cause analysis and identification of problem hotspots.

Introduction

Exploratory Data Analytics (EDA) is a critical component of the descriptive analytics task that engineers have to solve when dealing with data. The data can come in any format; the engineers normally have to wrangle the data, and then perform various transformations and functions to the data to draw insights. The cycle of analytics can be summarized in the following diagram:



Data Analytics diagram

Scope

An opportunity is available to have a look at the blockage and spillage data from July 2018 to June 2019. The anticipated outputs of the analysis are:

1. Understand the main causes of blockages and spillages objectively.
2. Visualize data to gain a better insight into the problem.
3. Quantify the losses by means of a pivot table; and
4. Build a regression model to predict the trajectory of stoppages. This will help in understanding the effort needed to address major causes for unnatural variation.

A Look at Blockage Data

The dataset has 16 variables and 1830 observations. The variables are of different types including character (words), numeric and factors. Factors can be either numeric or character based, and represent categories.

The structure of the data is shown below:

```
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':    1830 obs. of  16
variables:
 $ Parent_Area     : Factor w/ 5 levels "A1","A2","B",..: 1 1 1 1 1 1 1 1 1 1
...
 $ process_affected: chr  " Conveyor 702-2500 " " Conveyor 702-2500 " "
Conveyor 702-2500 " " Conveyor 702-2500 " ...
 $ process_cause   : chr  " Conveyor 702-2500 " " Conveyor 702-2500 " "
Conveyor 702-2500 " " Conveyor 702-2500 " ...
 $ TimeCategory    : Factor w/ 3 levels "D100","D300",..: 2 2 2 2 2 2 2 2 2 2
...
 $ Duration_Hours  : num  0.25 0.25 0.695 1.167 2.424 ...
 $ Delay_Start     : POSIXct, format: "2018-12-26 19:30:00" "2018-12-27
00:30:00" "2019-06-25 07:30:00" "2018-12-19 01:00:00" ...
 $ Delay_End       : POSIXct, format: "2018-12-26 19:45:00" "2018-12-27
00:45:00" "2019-06-25 08:12:00" "2018-12-19 02:10:00" ...
 $ Class           : Factor w/ 2 levels "External","Internal": 2 2 2 2 2 2 2
2 2 2 ...
 $ Scheduled       : Factor w/ 1 level "Unscheduled": 1 1 1 1 1 1 1 1 1 1 ...
 $ Responsibility  : Factor w/ 3 levels "Electrical","Mechanical",..: 3 3 3 3
3 3 3 3 3 3 ...
 $ Remark          : chr  "702-2500/2600 CLEAN APEX 19:30 TILL 19:45 ." ...
 $ Major           : Factor w/ 6 levels "Chute","Conveyor",..: 1 1 1 1 1 1 1
1 1 1 ...
 $ Minor           : Factor w/ 10 levels "Belt Skew","Belt Skew (SEP)",..: 5
5 5 5 5 5 5 5 5 5 ...
 $ Detail          : Factor w/ 10 levels "-","Blockage / Buildup",..: 1 1 1 1
1 1 1 1 1 1 ...
 $ Stop_Percent    : num  100 100 100 100 100 100 100 100 100 100 ...
 $ X_Effect        : chr  "100" "100" "100" "100" ...
```

From the structure, it can be seen that most of the data is in factor format, with one numerical variable (duration). Another important observation is that the “Scheduled” parameter has only one level, i.e. it does not change. This makes sense as blockages and spillages are not planned events. Therefore, from a data perspective, we can exclude the constant variable from our analysis.

A summary of the data is given in the following table:

```
Explore
Data        : DMS_Blockages_Spillages
Grouped by  : Parent_Area
```

```
Functions   : n_obs, mean, median, min, max, sd
Top         : Function

 Parent_Area       variable n_obs mean median  min    max    sd
          A1 Duration_Hours   832 0.73   0.43 0.00  39.16  1.96
          A2 Duration_Hours   195 1.70   0.67 0.00 119.42  8.64
           B Duration_Hours   551 0.90   0.53 0.00   6.83  1.16
           C Duration_Hours   149 4.33   1.64 0.01 166.60 14.64
           D Duration_Hours   103 4.06   3.25 0.01  22.50  4.06
```

This distibution is based on the delay duration as the target variable. On this exploration, the following observations are made:

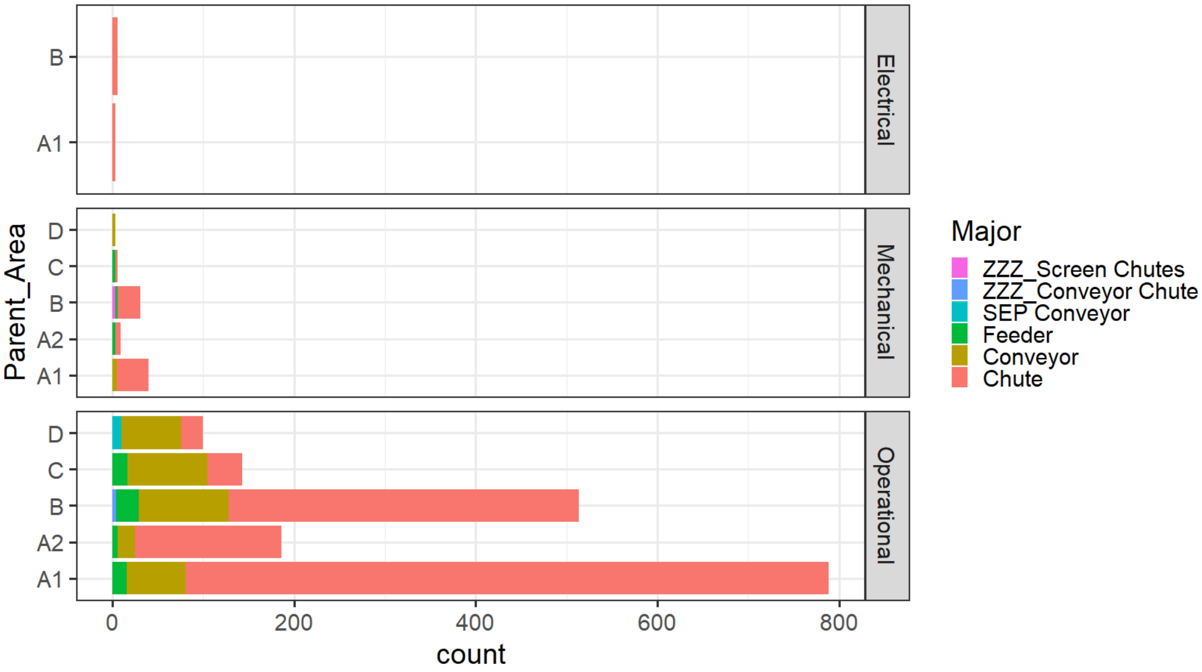
1. The data is skewed to the left (median < mean) for all parent areas. The skewness indicates a special cause in the downtime data. One possible suggestion is that the reaction time to a blockage is improving. While this is a good thing, it would be ideal if it shifted the mean rather than skewing the distribution. Hence there is an opportunity to improve the duration of a stoppage by introducing more inherent means of either detecting or responding to blockages and spillages.
2. Most of the incidents happen at parent area A1, followed by B. Plant area D has the lowest incident rate.
3. The duration at plant area C has the highest variance (standard deviation). This might suggest that the types of stoppages that happen at plant area C are relatively very different. Conversely, the duration at plant area B has the lowest variance.

Problem Description

We now look at some visualizations to get a better sense of the behaviour of the data.

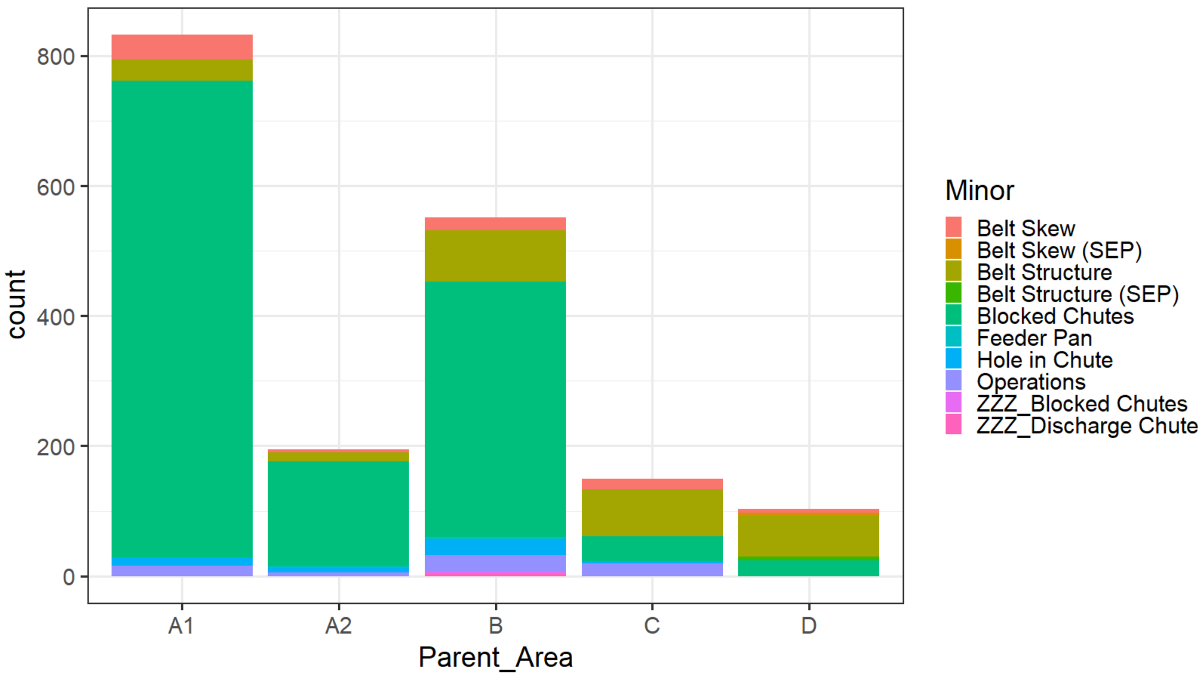
What is the Biggest Cause of Stoppages?

To answer this question, we create a visualization of the distribution of the delays across the different plant areas. We then fill the histograms with the relative proportions of the major causes. Lastly, we create facet rows to separate the data by the different responsibilities.



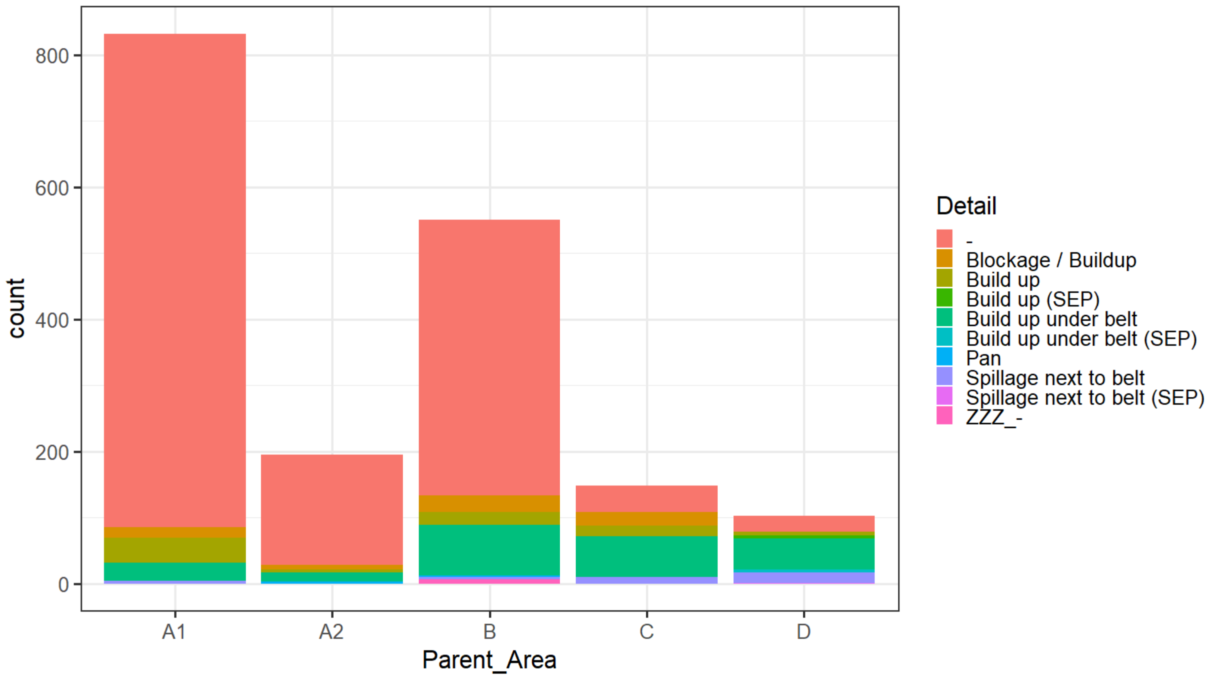
It can be immediately seen that the bulk of the information is in the Operational responsibility, which is much more than in the Mechanical and even more than the Electrical facets. Therefore most of the issues are operational. This indicates that the conditions of the machinery are not the main cause of problems, and that more attention could be paid to processes and personnel. Another thing to note is that the most incidents took place in plant area A1, followed by plant area B and then A2.

Another immediate observation is that the major problem description is the chute. This indicates that the chute problem is responsible for the most number of incidents and consequently the most amount of delays. This is especially true for plant areas A1 and B. Plant areas C and D are mostly struggling with conveyor related issues instead of chute. Although plant area B has a similar size of conveyor related issues to plant areas C and D, this is shadowed by the high number of chute incidents.



Based on the previous establishments, we take a closer look at the reasons for stoppages. We zoom inside the data by looking at the minor reason, which should be a more detailed explanation of the reasons given in the previous figure. The figure above shows that the biggest cause of delays is blocked chutes. Following this is the belt structure problem dominant in plant areas C and D. This is consistent with our earlier finding of the conveyor problem. Again, it is equally present in plant area B but it is shadowed by the blocked chutes problem.

Our next step is to look for even finer detail from the data in order to simplify the process of intervention. We look at the following distribution:

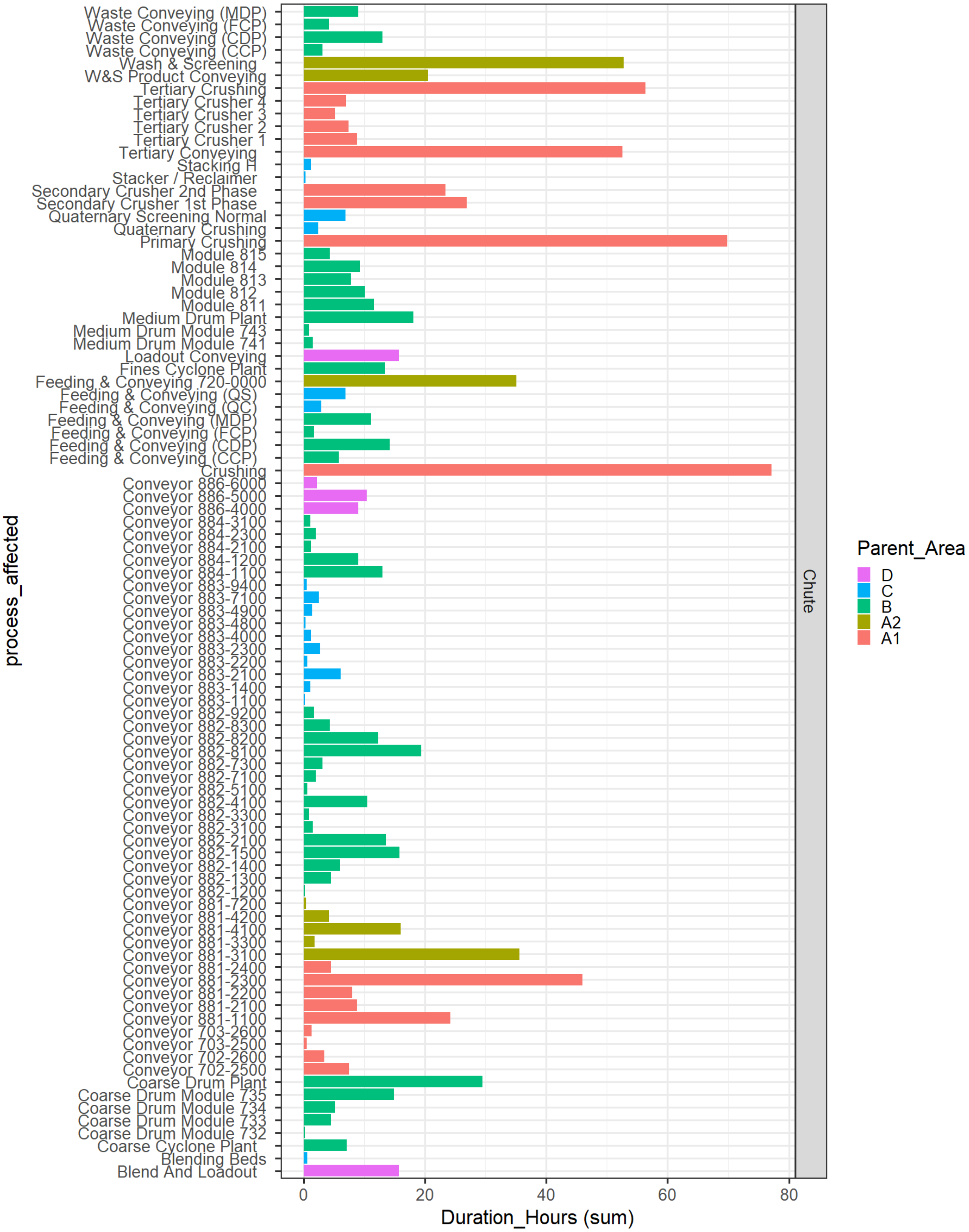


The distribution indicates that the blocked chutes problem is the dominant problem, although the detail does not specify it. We can confidently assume that the lowest level of detail available for chutes is blocked chutes. The second most dominant problem is the build up under the conveyor belt, indicated by the turquoise colour in the histogram. More domain knowledge should be able to define the build up problem.

Where Are The Problems Concentrated?

Chute Related Problems

As blocked chutes consitute the highest delays during stoppages, we look at the affected processes in the plant areas due to blocked chutes.

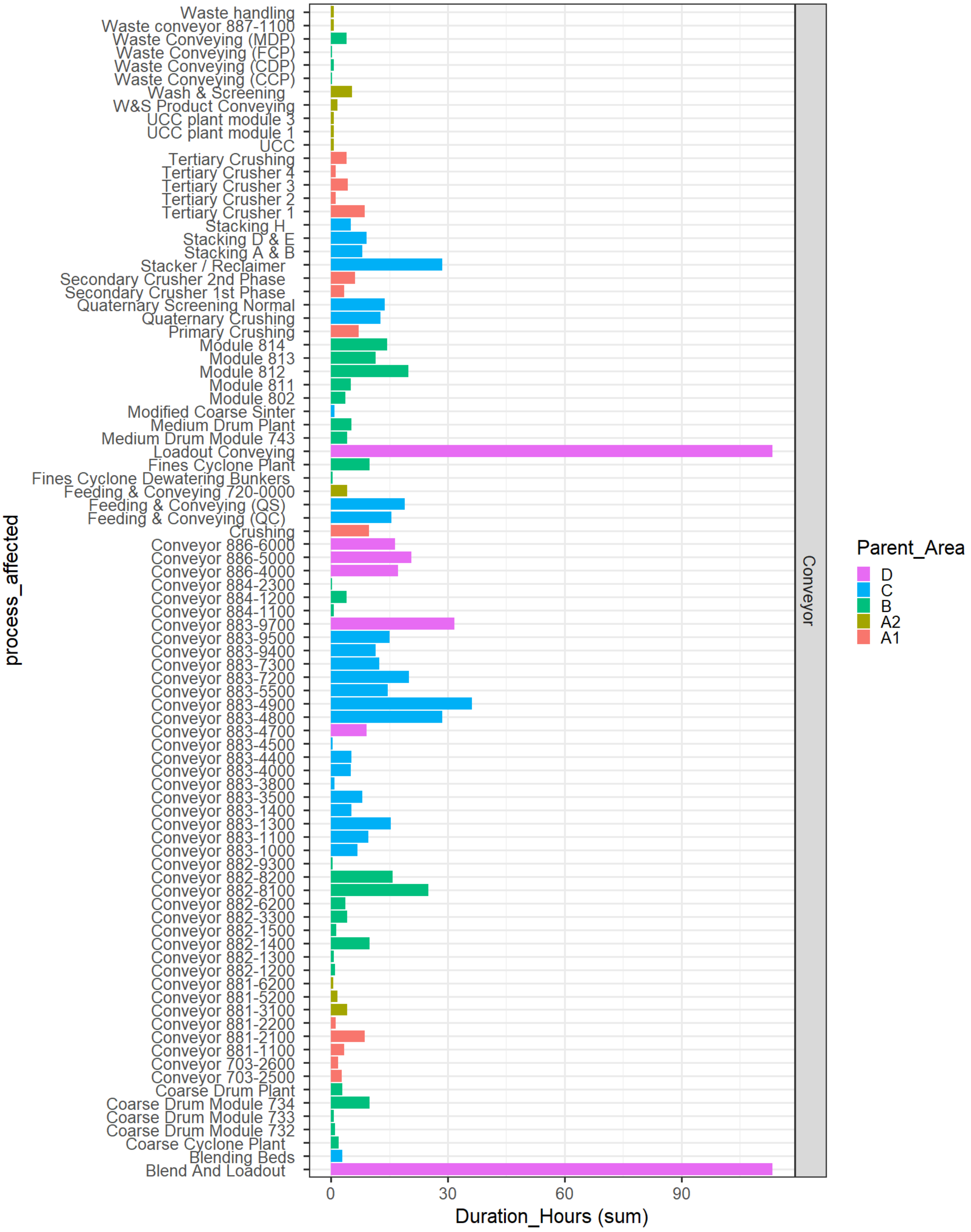


The most affected processes are crushing (Crushing, Primary Crushing, Tertiary Crushing) processes in parent area A1, as is shown in the bar plot. The next most affected process is Wash and Screening in parent area A2.

The third most affected process is Conveyor 881, which is apread between parent areas A1 and A2.

Conveyor Related Problems

The affected processes based on conveyor related problems are shown in the following bar plot.

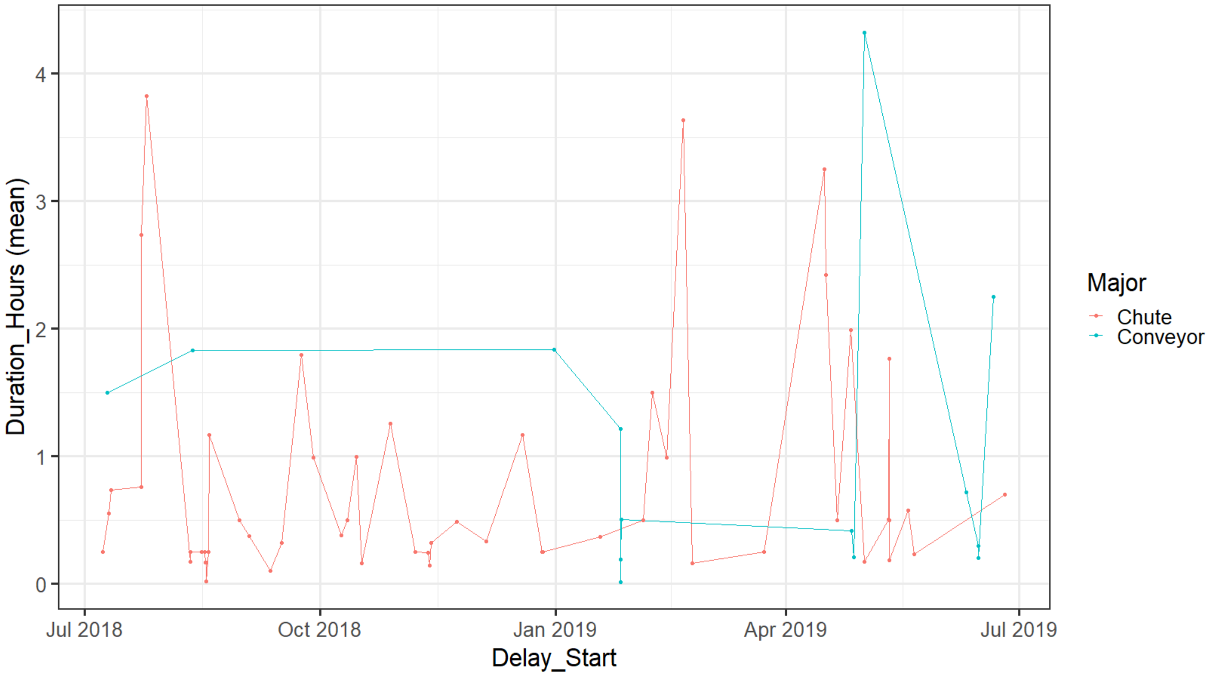


The most affected process in this case is the Bend And Loadout and Loadout Conveying in parent area D. They are followed by Conveyor 883-4800 to Conveyor 883-9700. The third most affected process is stacking (Stacker /

Reclaimer, Stacking D&E and Stacking A&B). These processes are where interventions would yield the most impact.

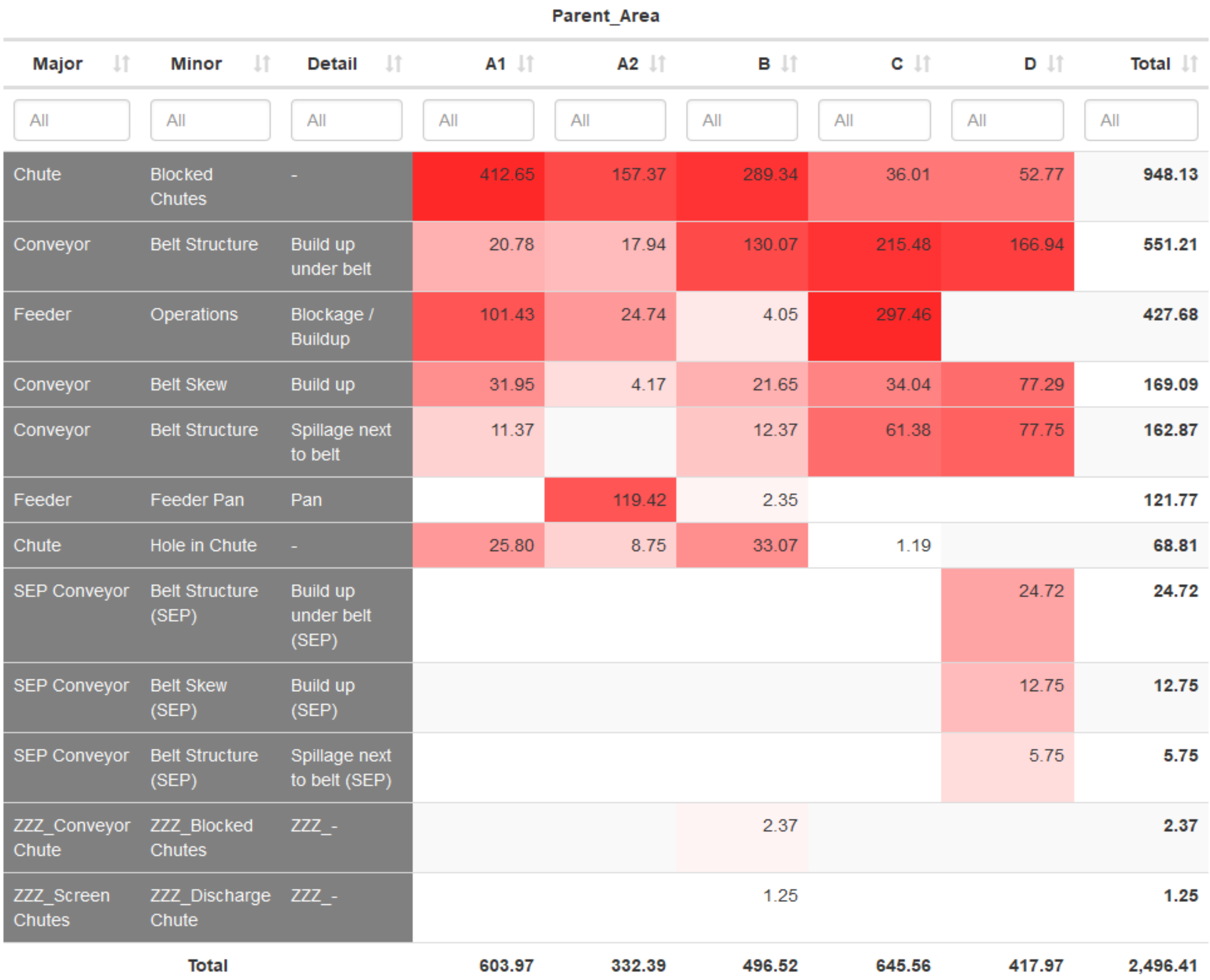
Delay Trends

Another important factor to consider is the pattern of the delays for the major problems (chutes and conveyor belts). To aid us in observing this we plot a time series trend of the durations:



As shown in the figure, the stoppage duration for conveyor related causes is much smoother than that for chute related causes. This can be based on several reasons, including that there are different types of detection mechanisms for the two. It might mean that the under build up at the conveyor belts is more detectable than the blocked chutes. It could also mean that the nature of the blockages is such that the time needed to unblock is not as deterministic.

To further aid our understanding of the delay behaviour of the data, we pivot the data:



			Parent_Area					
Major	Minor	Detail	A1	A2	B	C	D	Total
All	All	All	All	All	All	All	All	All
Chute	Blocked Chutes	-	412.65	157.37	289.34	36.01	52.77	948.13
Conveyor	Belt Structure	Build up under belt	20.78	17.94	130.07	215.48	166.94	551.21
Feeder	Operations	Blockage / Buildup	101.43	24.74	4.05	297.46		427.68
Conveyor	Belt Skew	Build up	31.95	4.17	21.65	34.04	77.29	169.09
Conveyor	Belt Structure	Spillage next to belt	11.37		12.37	61.38	77.75	162.87
Feeder	Feeder Pan	Pan		119.42	2.35			121.77
Chute	Hole in Chute	-	25.80	8.75	33.07	1.19		68.81
SEP Conveyor	Belt Structure (SEP)	Build up under belt (SEP)					24.72	24.72
SEP Conveyor	Belt Skew (SEP)	Build up (SEP)					12.75	12.75
SEP Conveyor	Belt Structure (SEP)	Spillage next to belt (SEP)					5.75	5.75
ZZZ_Conveyor Chute	ZZZ_Blocked Chutes	ZZZ_-			2.37			2.37
ZZZ_Screen Chutes	ZZZ_Discharge Chute	ZZZ_-			1.25			1.25
	Total		603.97	332.39	496.52	645.56	417.97	2,496.41

The pivot table allows us to see the totals of the stoppage durations per parent area and per problem. Here we used two tiers of the problem descriptions (major and minor) as it was previously noted that they describe the data adequately. We can see the total amount of time across the rows (problem descriptions) and columns (parent areas). The grand total amount of stoppage time is 2, 496.41 hours. This time, based on the production cost rates of the different areas of the plant, can be quantified as waste and lost opportunity in monetary terms. This would then justify the need for capital expenditure in order to employ permanent solutions to the major causes for stoppages and spillages.

Data Modeling

Data modeling is a pivotal part of analytics because it allows us to describe the data mathematically. This means that an accurate model of the observed behaviour can be produced, and this model will allow us to accurately predict outputs based on given inputs. We can then assess the error margins (residuals) in order to judge the accuracy of the model.

Linear Regression

The most intuitive model to use is the linear regression model. It is also a good starting point as it is relatively easy to implement, and many datasets can be accurately represented by linear models. In addition, the behaviour of the residuals can tell when the linear model is not a good fit.

The linear regression model is fitted on the data and the results are shown in the following table:

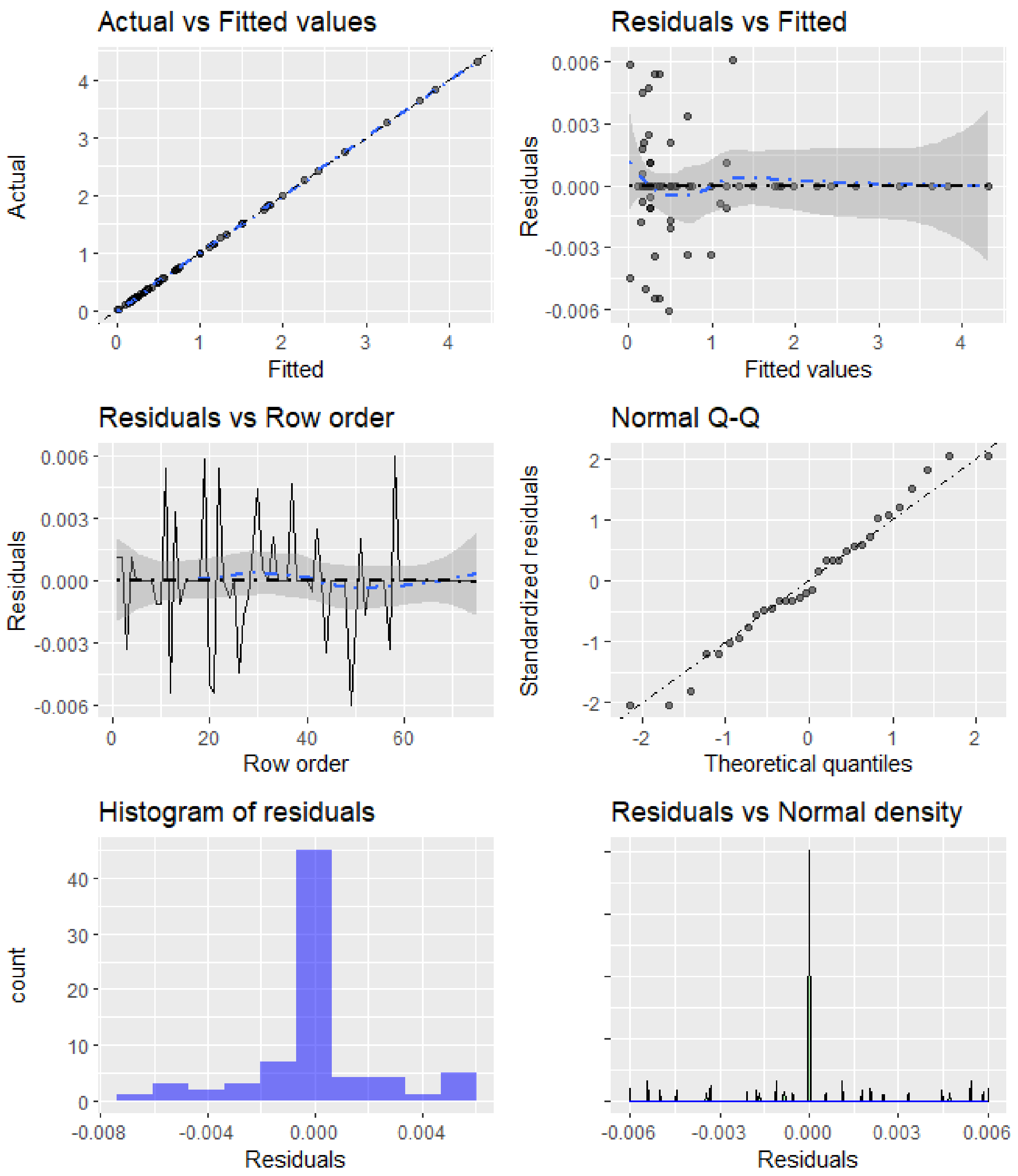
```
Linear regression (OLS)
Data     : DMS_Blockages_Spillages
Response variable    : Duration_Hours
Explanatory variables: process_affected, process_cause, TimeCategory,
Delay_Start, Delay_End, Responsibility, Major, Minor, Detail
Null hyp.: the effect of x on Duration_Hours is zero
Alt. hyp.: the effect of x on Duration_Hours is not zero

                                      coefficient std.error  t.value p.value
 (Intercept)                               -0.094     0.148   -0.639   0.525
 process_affected| Conveyor 702-2600        0.002     0.004    0.588   0.559
 process_affected| Conveyor 703-2500       -0.002     0.005   -0.456   0.650
 process_affected| Conveyor 703-2600        0.005     0.005    1.067   0.290
 process_affected| Conveyor 881-1100       -0.000     0.003   -0.130   0.897
 process_affected| Conveyor 881-2100        0.001     0.004    0.356   0.723
 TimeCategory|D300                          0.005     0.004    1.233   0.222
 Delay_Start                               -0.000     0.000 -914.886  < .001
***
 Delay_End                                  0.000     0.000  914.859  < .001
***
 Major|Conveyor                             0.012     0.008    1.410   0.164
 Minor|Belt Structure                      -0.014     0.009   -1.670   0.100
.
 Minor|Blocked Chutes                       0.008     0.008    1.032   0.306
 Detail|Build up under belt                 0.007     0.008    0.849   0.399

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 1,  Adjusted R-squared: 1
F-statistic: 120251.9 df(12,62), p.value < .001
Nr obs: 75

The set of explanatory variables exhibit perfect multicollinearity.
One or more variables were dropped from the estimation.
Model 1: Duration_Hours ~ process_affected + process_cause + TimeCategory +
    Delay_Start + Delay_End + Responsibility
Model 2: Duration_Hours ~ process_affected + process_cause + TimeCategory +
    Delay_Start + Delay_End + Responsibility + Major + Minor +
    Detail
```

```
R-squared, Model 1 vs 2: 1 1
F-statistic: 1.288 df(4,62), p.value 0.284
```

The linear regression model uses the null hypothesis to determine if each variable has an influence on the target parameter, which is the stoppage duration. The null hypothesis would therefore state that the variable has no effect on the stoppage duration, and an alternative hypothesis would be that the variable has an effect on the stoppage duration. Based on the model results, the most important contributors to the duration of the stoppage include the belt structure, blocked chutes, build up under belt and conveyor. These are consistent with the initial observations about the problem hotspots.

Regression-Based Predictions

In order to assess our model, we perform a fit and analyze the residuals. A fit is the model result when a predictor (input) is applied to the model. Residuals are the differences between observed and fitted values, or the errors. These can be positive if the fit is less than the observation, or negative if vice versa. The residuals are shown below:



The results indicate that the model is an accurate representation of the data and can be used to make predictions. The following observations are made:

1. There is a linear, 1 - 1 relationship between the actual and fitted values.
2. The residuals are randomly spread about the horizontal axis, which implies that the linear model is the right choice.

3. The distributions of the fitted values and observed values are linearly related (Q-Q plot).
4. The residuals are centered about zero, meaning that the model does not introduce any bias.
5. The relative proportions of the residuals are small, so the predictions will have small errors.

Recommendations

Based on the analysis of the stoppages dataset, the following recommendations are made:

1. A financial losses analysis that is based on the stoppage hours. This would justify any subsequent efforts for intervention and problem-solving.
2. The biggest stoppages are caused by blocked chutes followed by build up under conveyor belt.
3. Improved detection of problems, e.g. a blocked chute sensor. These would reduce the time-to-reaction, provide a basis for solution automation, and aid in automated reporting of chute related stoppages.
4. Improved documentation of problems. The wrangling stage revealed that there are characters like “ZZZ” that are used to prepend and append parameters. It was also noted that the detail for blocked chutes is “-”. This is not good practice for documentation.
5. An improved categorization of data, e.g. Remark, would improve the statistical insights that could be gained from analysis efforts.
6. Most stoppages occurred in parent area A1.
7. Most stoppages occurred in the Operational responsibility. This could mean operators need training, or that the mill conditions need better care and more automation.
8. A predictive model can be built to aid in better preparations for stoppages, and for experimenting with possible solutions.