STAT 435: Midterm Report

Due on March 15, 2023

University of Waterloo
Faculty of Mathematics
Department of Statistics and Actuarial Science

Prepared by:

Gunchica Bhalla

Introduction

Watfactory is a computer-based model of an automotive camshaft manufacturing process that includes 60 inputs that can vary and 30 inputs that are fixed. The output can be measured in three locations, and the objective of the project is to minimize the variability of the final output (y300) to be within the range of -10 to 10, given a budget of $10,000. At the end of the investigation coducted for this report the remaining balance is $3,840. I used $600 to collect data for the baseline investigation, $1,125 for the measurement system analysis, $1,875 for the search of process causing variation transmission, $1,200 for search if the stream-to-stream family is home to the variation and lastly spent $1,360 to further investigate the component to find the culprit for the variation.
This report will use the findings from the initial investigation to establish the project objective and identify the primary source of variation in y300. I will also assess the measurement system's accuracy and reliability, which will serve as a foundation for future decision-making. To identify the root cause of the variation, I will utilize the elimination process outlined in Statistical Engineering (Steiner & MacKay, 2005) to trace various parts throughout the manufacturing process to find variation transmission. Additionally, I will examine the varying parts to pinpoint the specific component responsible for the variation.

I will be following the QPDAC approach for each of the 5 parts of the investigation. Each of the steps will be defined in great detail in the report.

Part 1: Baseline Investigation

Questions

In this report, several investigations will been conducted on various datasets to address different parts of the investigation. The primary objective of this report is to identify the dominant cause of variation in y_{300}. To achieve this objective, it is crucial to establish a baseline. To accomplish this, the following questions must be addressed:
Setting the baseline:

1. What is our baseline?
2. What time family does the variation belong to?
3. What is the full extent of variation (FEoV)?
4. How much should we reduce the variation to meet our goal?

Plan

Starting with a baseline investigation, I started with collecting sufficient amount of data for sampling such that the entire range of variation is captured and be able to identify the time family that the variation exists in.

To achieve this, I collected 5 parts (y_{300}) per hour for a week (5 days), capturing every shift in a day (3 shifts) (i.e., 5 parts every hour for the next 120 hours, starting at part number 1). Random and systematic sampling was used to capture 5 parts at random every hour. I chose this sampling plan to potentially identify day-to-day, shift-to-shift, hour-to-hour or part-to-part variation. In the case that we

did not see any variation in these listed time families, we would consider collecting further data to perhaps check week-to-week or collect more samples per hour, however, this was not necessary. The total cost of the plan was $600.

To identify the dominant cause, I plotted the data grouped by day, hour, shift and hour to gauge the an idea through the visualization and to confirm the results of the visualization I conducted a regression analysis and used ANOVA to confirm the significance of the time families in the output y_{300}.

Data

The data for all investigations outlined in this report was collected from the Watfactory investigation page. For the baseline establishment I collected data given the day, shift and part number for each y_{300} sample and its corresponding output value. We collected according to the above plan so we were able to investigate the potential day-to-day, shift-to-shift, hour-to-hour or part-to-part variation families. From the raw data, we then extrapolated what hour each part was from and which unique shift it came from (i.e., shift 2 on day 3). This was then used to conduct our baseline investigation.

Analysis

I chose R to conduct the investigation since it is the language we have been using in class.

To begin Figure 1 below shows the histogram of the collected data where it can be seen that the some of the observation fall outside the acceptable range [-10,10]. The histogram seems to be bimodal signifying that the data is not normally distributed.In the collected sample 4.8% of the value fall outside the required ranges, with more negative values i.e less than -10.

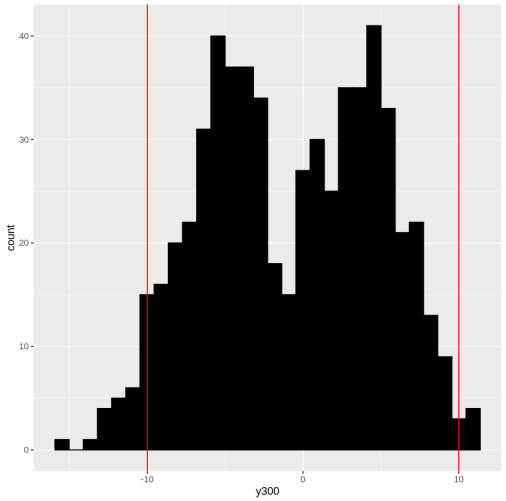


Figure 1. Histogram

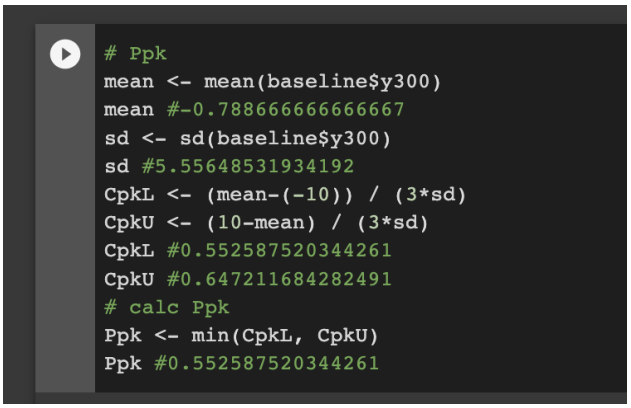
The data ranges from -15.1(LSL) to 11.3 (USL) with a mean value ($\hat{\mu}$) of -0.788 and median is -0.55. It has a $\sigma_{overall}$ of 5.556 as shown in Figure 2.

$\hat{\mu}$	$\hat{\sigma}$	Min	Q1	Median	Q3	Max	Range	FEoV

-0.788	5.556	-15.1	-5.3	-0.55	3.9	11.3	[-15.1,11.3]	26.4

Figure 2. Data Analysis

As shown in Figure 3, we calculated our P_{pk} value to be 0.55 which signifies that the process is "not capable" as the value is less that 1. One of the goals for this investigation is to improve the P_{pk} such that the value is atleast 1.



```
# Ppk
mean <- mean(baseline$y300)
mean #-0.788666666666667
sd <- sd(baseline$y300)
sd #5.55648531934192
CpkL <- (mean-(-10)) / (3*sd)
CpkU <- (10-mean) / (3*sd)
CpkL #0.552587520344261
CpkU #0.647211684282491
# calc Ppk
Ppk <- min(CpkL, CpkU)
Ppk #0.552587520344261
```

Figure 3. P_{pk} calculation

To accomplish making the P_{pk} value to be atleast 1, we can use multiple methods but I want to approach this by reducing the variance in y_{300}. In the following Figure 4, we find the maximum allowed value for the variation for the P_{pk} to be at least 1 while holding the estimated mean constant as its effect is insignificant to the estimated variation change.

$$\text{Using the } P_{pk} = min\left\{\frac{USL-\hat{\mu}}{3\hat{\sigma}_{overall}}, \frac{\hat{\mu}-LSL}{3\hat{\sigma}_{overall}}\right\}$$

$$1 \leq min\left\{\frac{USL-\hat{\mu}}{3\sigma_{overall}}, \frac{\hat{\mu}-LSL}{3\sigma_{overall}}\right\}$$

Using the values from above

$$1 \leq min\left\{\frac{10+0.788}{3\sigma_{overall}}, \frac{-0.788+10}{3\sigma_{overall}}\right\}$$

$$1 \leq \frac{-0.788+10}{3\sigma_{overall}}$$

$$1 \leq \frac{9.212}{3\sigma_{overall}}$$

$$3\sigma_{overall} \leq 9.64$$

$$\sigma_{overall} \leq 3.071$$

Figure 4: Standard Deviation Upper Limit Calculation

For the P_{pk} to be at least one, the maximum value of the variation has to be less than equal to 3.071. Our current variation is 5.556. I need to reduce our variation by 44.73% to achieve the P_{pk} value of 1.

After establishing the baseline in terms of mean, variance and P_{pk} and identifying the amount of variation reduction required to achieve the gaol of the project I continued to work on identifying the time family that the variation belongs to, starting with visualizing the data as boxplots.

As shown in Figure 5 we can see that there is not much variation within days, they have fairly comparable ranges and non fluctuating means. To verify the findings I conducted an anova analysis to find that dacount may not have a significant effect on the response variable y_{300}, as its p-value 0.0793 is greater than 0.05 as shown in Appendix B1.

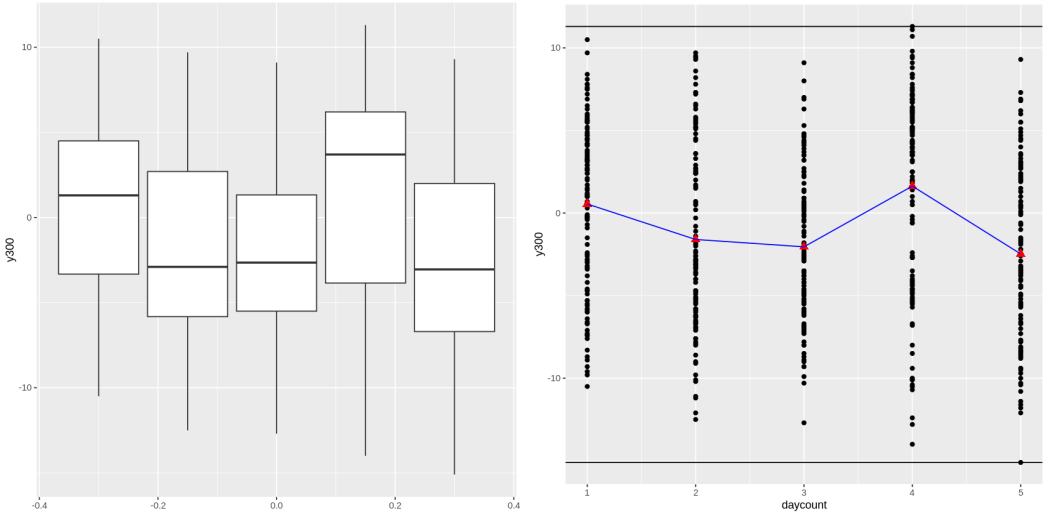


Figure 5.Plots by Day

Moving onto data grouped by shift Figure 6 shows considerable amount of variation. Shift 1 and 2 have comparable means where as shift three is very different compared to them. It is possible for the variation to be due to a shift to shift variation. To validate the observation I conducted a anova analysis to find, as shown in Appendix B2, that shift has a significant effect on the response variable y_{300}, as its p-value is much smaller than 0.05

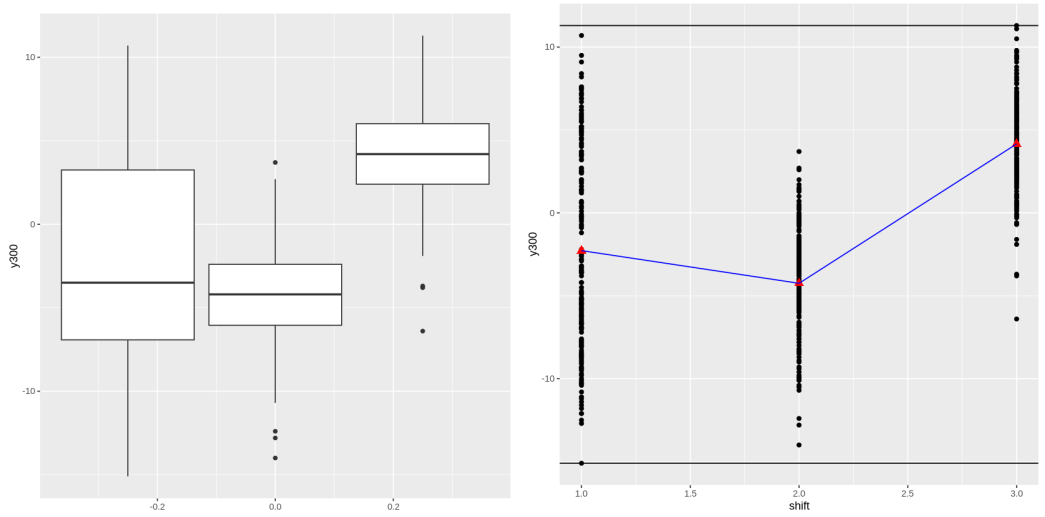


Figure 6. Plots by Shift

Figure 7 shows data visualized when grouped by the hour over the whole day. These graphs support the finding regarding the variation in shifts as variation in the below graphs can be noted between the groups of eight boxes rather than between the hours themselves. The mean is fairly similarly for each hour in the shift and variation can be seen at shift changes. To check the significance of hour on y_{300} I conducted another anova analysis as noted in Appendix B3 to find that the very small p-value signifies a high imapct on the output which differed from the visual analysis. It could be the hour to

hour variation is getting transmitted as shift to shift. To check which of the two has more impact on y_{300} I created an anova model with y_{300} ~ shift + hour as shown in Appendix B4. It can be seen that both shift and hour have p-values less that 0.5 but shift is more significant factor in the variation in y_{300} when compared to hour. Given these findings I will not be investigating the hours grouped by shift.

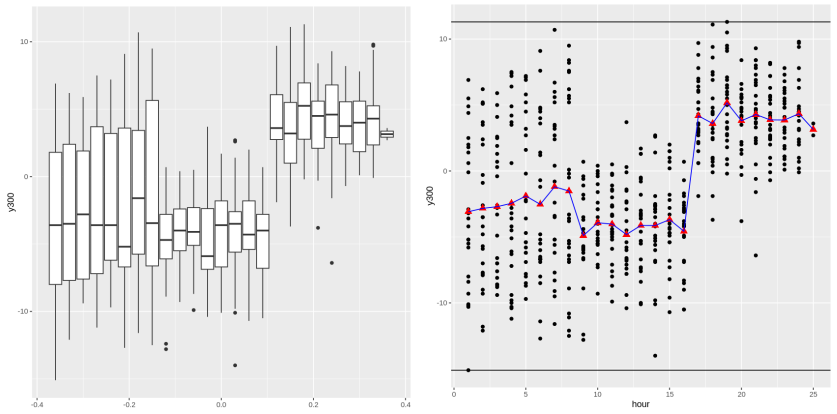


Figure 7. Plots by hour

Continuing the investigation, I plotted the data by part number as shown in Figure 8 to evaluate if the variation can be caused part-part. The graph below is a point graph rather than a box plot as the box plot did not provide additional incite. On looking closely the points seem to be 5 groups of three patterns with variation amongst the three pattern rather than between groups signifying the variation between shifts rather than days or parts.

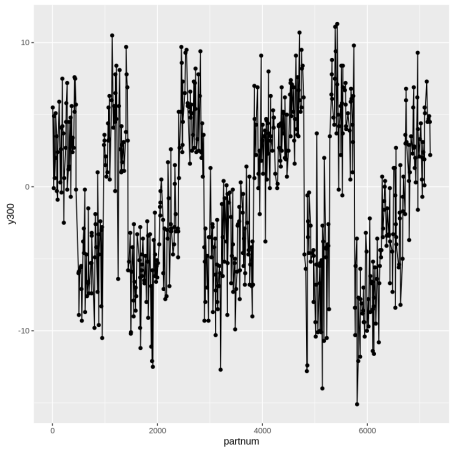


Figure 8. Plots by part number

To investigate the variation by shift further I plotted the data grouped by shift and day as show in Figure 9. The boxplots represent each shift across 5 days and it can be seen that there is large variability between the shifts every day where as the pattern is some what consistent across the days. It is will not be wrong to assume that the variation belongs to the shift to shift time family.

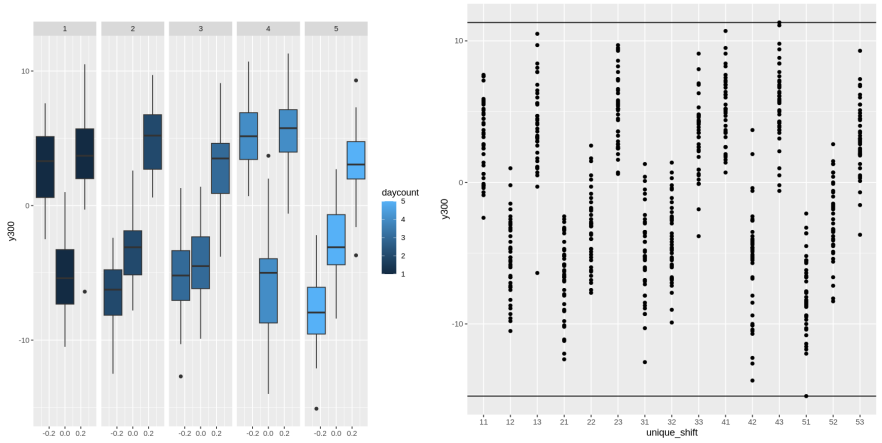


Figure 9. Plots by Day and Shift

To confirm the findings I created a anova model between the y_{300} , daycount, shift, hour and part number as shown in Appendix C. From the anova analysis it can be confirmed that the the shift-to-shift is the most significant contributor to the variation in y_{300}. According to the anova analysis hour-to-hour is not as significant as shift-to-shift hence for this project I will focus on shift-to-shift as the baseline time family.

Conclusion

In summary, the baseline investigation of the Watfactory process revealed an overall variance of 5.55, which is deemed as "not capable" with a P_{pk} value of 0.55 that will be used as a baseline for tracking progress.

It was determined that reducing the variation by 44.73% is necessary to make the process capable, and the full extent of the output variation ranges from -15.1 to 11.3.

This investigation allows for the search and testing of a solution to reduce the variation. The visual and ANOVA analysis indicate that the variation in y_{300} is due to day-to-day(8.6%) and shift-to-shift factors (43.7%). The high residual value indicates part-to-part time family also may be involved in the variation. Moving forward, I will be only considering shift-to-shift variation given the significantly higher mean square value that indicates that the shift factor has the highest impact on y_{300} as shown in Appendix C. This enables the prioritization of inputs that vary shift-to-shift or are unknown for future investigations.

Part 2: Measurement System Assesment

Question

In the previous part of this report a baseline was established and the time family of the dominant cause was determined. Moving into the next part, the measurement system will be evaluated on its efficiency and reliability.For the future investigations, it is important to utilize the measurement system effectively and ensure that it is not a source of variation. To achieve this, we will assess the discrimination ratio, which compares the standard deviation of the actual process with that of the measurement system.

$$D = \frac{SD(process)}{SD(measurement)}$$

If D is greater than 3 the measurement system gr it is not a dominant cause. It would be also beneficial to calculate the variation of the measurement system because if the measurement system contributes relatively small amount of variability to y_{300} then it is considered to be effective.

The questions that require answers in this measurement system investigation are:

1. Is the measurement system a dominant source of variation?
2. Is there a bias in the measurement system, if yes, then what?
3. What is the variation of the measurement system?

Plan

The investigation plan involves selecting one part with a true value that is near the mean, as well as two parts with high averages that are situated at the extreme values. This is done in order to assess the full range of variation and determine whether the size of the part has any effect on the accuracy of the measurement system. The areas are marked in Figure 10.

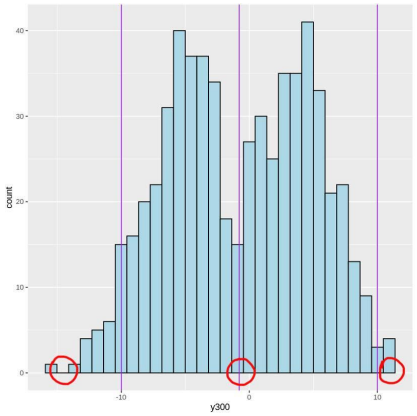


Figure 10. Histogram of previously collected data

Since the dominant time family identified in the baseline is shift-to-shift, chose to collect the measurements on that same scale. For each of the 3 parts selected, we measured each 5 times a shift for 15 shifts, resulting in a total cost of $1125 ($5 x 5 parts x 15 shifts x 5 measurements) and 225 measurements in total (75 per part). In hindsight we maybe shouldn't have chosen such a long time horizon as external/environmental impacts may have impacted the measurements but luckily in the end that was not the case.

Data

I collected data from the Watfactory measurement system page, which allowed me to measure specific components. Part 3282 was selected because its y300 value fell within the mean, whereas parts 1139 and 1914 were chosen as they represented our two extremes (even though they were not the maximum and minimum values). Figure 11 depicts the parts that were selected.

A data.frame: 3 × 7

daycount	shift	partnum	y300	hour	hr_in_shift	unique_shift
<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<chr>
1	3	1139	10.5	19	6	13
2	1	1914	-12.5	8	8	21
3	1	3282	-0.8	7	7	31

Figure 11. Selected Parts for measurement analysis

Analysis

As previously mentioned the analysis is done using R as that is the language used in class. Beginning with visual analysis I plotted the measurement of each part into a scatter plot as shown in Figure 12 below to observe that the each of the part has a similar distribution with comparable variation and a few outliers.

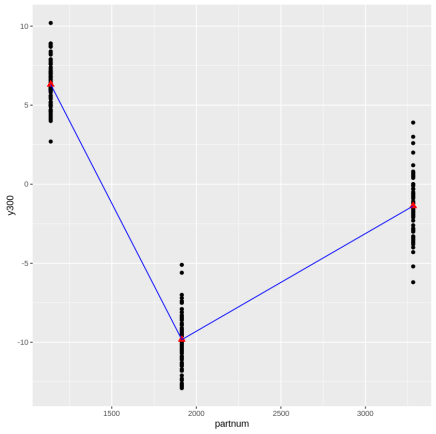


Figure 12. Scatterplot for each part

To verify the observation I conducted a statistical analysis as shown in figure 13 below. It can be seen that the standard deviation for each part is within the ball park of each other.

A tibble: 3 × 5

partnum	mean_y300	min_y300	max_y300	sd_y300
<int>	<dbl>	<dbl>	<dbl>	<dbl>
1139	6.310667	2.7	10.2	1.397640
1914	-9.830667	-12.9	-5.1	1.674690
3282	-1.377333	-6.2	3.9	1.830491

Figure 13. Statistical Summary of the selected parts.

Moving onto calculating the discriminant ratio, I starter by estimating the measurement variation using the sd_y300 values from Figure 13 to create the equation in Figure 14.

$$SD(measurement) = \frac{\sqrt{1.39^2+1.67^2+1.183}}{3} = 1.64$$

Figure14. Measurement Variation Calculation

From the Figure 14 it can be seen that the measurement standard deviation is 1.64. This information combined with the overall variation ($\sigma_{overall}$) from the baseline investigation we can calculate the SD process as shown in Figure 15. The overall variation ($\sigma_{overall}$) is the SD(total) which is equal to 5.55 as established in Figure 2.

$$SD(total) = \sqrt{SD^2(process) + SD^2(measurement)}$$
$$SD(process) = \sqrt{SD^2(total) - SD^2(measurement)} = \sqrt{5.55^2 - 1.35^2} = 5.32$$

Figure 15. SD(process) Calculation

With all the above mentioned information the discrimant ratio can be calculated as shown below in Figure 16.

$$D = \frac{SD(process)}{SD(measurement)} = \frac{5.32}{1.64} = 3.22$$

Figure 16. Discrimant ratio calculation

As the discriminant ratio is greater than 3 we can conclude that the measurement system is not a source of variation and thus is not a dominant cause of variation.

In addition, we estimated the bias of our measurement system as -0.69, as shown in Figure 16. The estimated average measurement error is -0.69, which is insignificant considering the min and max values.

```
bias = (sum(summary_stats$mean_y300 - parts$y300)) / length(parts$y300)
bias
```

```
-0.699111111111112
```

Figure 16. Estimating the bias of the measurement system

Conclusion

An average measurement error of -0.69 was estimated, which although not substantial, could be rectified in the future through recalibration. The measurement system's variance is approximated to be 1.64, whereas the process variation is 5.32. As stated earlier, the measurement system's bias is -0.69 and has a discriminant ratio of 3.22. A discrimant ration value of greater than 3 signifies that our measurement system is not the primary cause of variation, allowing us to proceed to the next phase of the StatEng algorithm.

Part 3: Search for Cause I

Question

In this part, my focus is on examining various aspects of the manufacturing process to identify the source of variation, find the process that is home to the dominant cause. The investigation process is described in more detail in the next section of the report. I will be utilizing an elimination approach to pinpoint the specific problem area. By the end of the investigation, I aim to address the following questions:

1. Between which 2 measurement locations is causing the maximum variation transmission?
2. Which phase of the entire manufacturing process is the home of the dominant cause?

Plan

The investigation plan is to obtain an adequate amount of data through sampling, covering the entire range of variation observed in the three outputs: y100, y200, and y300. This will aid in identifying the part of the manufacturing process with the highest degree of variation transmission and find the home of the dominant cause.

Based on the baseline investigation, it was discovered that most of the variation was associated with the shift-to-shift time family. Therefore, in this investigation, we concentrated on gathering data over several shifts while focusing on a lower number of parts per day. We collected data for 5 parts per shift over 15 shifts, measuring all outputs of y100, y200, and y300, to account for all variations noted during the shifts. The total cost incurred was $1875 for measuring 75 parts, each recorded thrice.

Data

Data was gathered from the WatFactory measurement system page, which measured the outputs for five parts per shift over 15 shifts. Random and systematic sampling techniques were utilized to collect the data, ensuring that parts were chosen at random throughout the shift to avoid any variation that might belong to other non-dominant time families. The collected data was analyzed to determine the mean, standard deviation, and range of each output, as shown in Figure 17. The total cost of this data collection process was $1,875.

A data.frame: 6 × 6

	daycount	shift	partnum	y100	y200	y300
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>
1	11	1	14453	3.0	5.5	-2.5
2	11	1	14475	6.3	7.1	-4.1
3	11	1	14495	4.7	5.7	-3.6
4	11	1	14557	7.7	8.2	-5.6
5	11	1	14836	5.7	4.1	-7.3
6	11	2	15007	4.0	5.2	-4.6

Figure 17. Collected raw data

Analysis

Figure 18 represents the upper limit of our FEoV; however, there is a small amount of data missing for the lower limit. This suggests that additional data may need to be collected. Nonetheless, due to the compelling evidence of variance transmission between y200 and y300, the available data is sufficient to accurately reflect the actual state of the process. In the event that the analytical analysis proves inconclusive, additional data can be collected if deemed necessary.

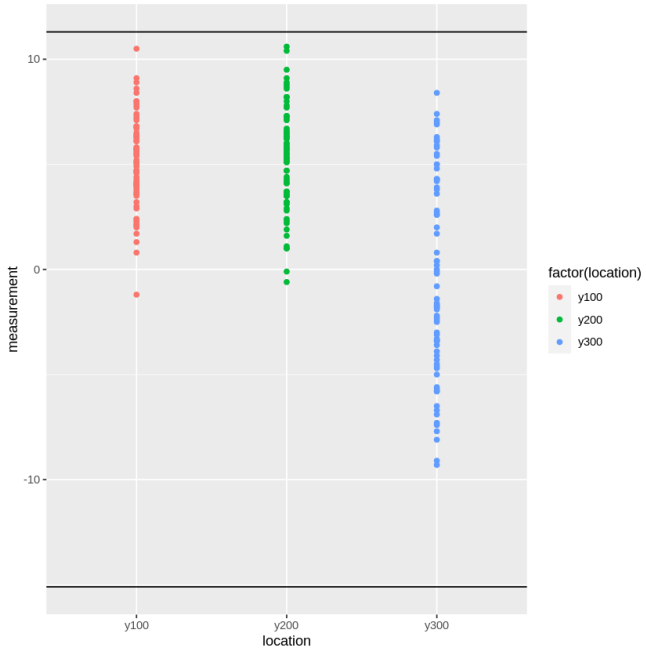


Figure 18. Scatter Plot

In Figure 19, I provided a summary of the statistical analysis conducted on the collected data. Although it did not reveal any significant new information regarding the location of where the variation increases, it did confirm that the cause of the variation increase can be attributed to the processes between y200 and y300.

	y100	y200	y300
Mean	5.1186667	5.2520000	-0.3346667
Variation	4.6685658	5.9206378	22.9344577
Range	11.7000000	11.2000000	17.7000000

Figure 19. Statistical Analysis

Therefore, we generated individual plots of the outputs for each shift to monitor the increase in variation, as shown in Appendix D. After analyzing Appendix D1, D2, and D3, we observed that y100 and y200 showed a relatively consistent level of variation across the shifts. On the other hand, there was a significant difference in the averages of y200 and y300, indicating that the process between y200 and y300 was the cause of the increased variation. This finding aligns with the results obtained from the scatter plot displayed in Figure 17.

To confirm the findings from the numerical analysis, I continued to conduct a regression analysis as shown in Appendix E where a linear regression model is created for pairs of outputs ie. $y_{300} \sim y_{100}$, $y_{300} \sim y_{200}$, and $y_{200} \sim y_{100}$. Other pairs are not considered as the scatter plot in Figure 17 implies a downstream variation transmission.

In Appendix E1, suggests there is little relationship between y_{100} and y_{300}. This means if we hold y_{100} constant, then we see most of the extent of variation. The same pattern is shown in Appendix E2. In Appendix E3,it can been that there exists a strong relationship between y_{200} and y_{100}. This means if we hold y_{100} fixed, then there is little variation in the y_{200} output, so the variation is coming from the upstream family. Hence it can be inferred that the stage between y_{200} and y_{300} - heat treatment is the home to the dominact cause.

To verify the findings from the regression analysis, I calculate the percentage of variation generated at each stage, we use the equation shown in Figure 20 below

$$\sigma_{transmitted} = \sqrt{\sigma^2_{y300} - \sigma^2_{\varepsilon}}$$

Figure 20. Variance Transmissted

As the visual and numerical analysis suggest that the process between y_{200} and y_{300} causes the variation, I used the equation in Figure 20 to calculate the variation transmiyyed from y_{200} to y_{300} as shown in Figure 21.

$$\sigma_{transmitted} = \sqrt{\sigma^2_{y300} - \sigma^2_{\varepsilon}} = 1.18 \quad \text{and} \quad \sigma_{y300} = 4.79$$

$$\Rightarrow \%\ variation\ between\ y200\ \&\ y300 = \frac{\sigma_{y300} - \sigma_{transmitted}}{\sigma_{y300}} = 75.4\%$$

Figure 21. Variation Transmission between y_{200} and y_{300}

This tells us that the variance generated between y_{200} and y_{300} is 75.4% and the remaining 24.6% comes from before y_{200}.

Next, we eliminate all x's between y_{200} and y_{300} from the suspect list and we repeat the process for between y_{100} to y_{200} as shown in Figure 22.

$$\sigma_{transmitted} = \sqrt{\sigma^2_{y200} - \sigma^2_{\varepsilon}} = 4.19 \quad \text{and} \quad \sigma_{y200} = 4.79$$

$$\Rightarrow \%\ variation\ between\ y100\ \&\ y200 = \frac{\sigma_{y200} - \sigma_{transmitted}}{\sigma_{y300}} = 12.3\%$$

Figure 22. Variation Transmission between y_{200} and y_{100}

This tells us that the variance generated between y_{100} and y_{200} is 12.3% , signifying that 12.3% of the variation comes from the start of the process before y_{100}.

It can be noted that 74.5% of the variation is generated between y_{200} and y_{300} and lives between the y_{200} to y_{300} measurement. This means that 24.5% of that variance is generated before y_{200}, out of which

12.3% comes from before y_{100} so, we may conclude that the majority of the process variation is introduced in the heat treatment process.

To further verfy the finidngs an ANOVA analysis was conducted on each linear regression model as shown in Appendix F to reveal that the highest variation was generated between y_{200} and y_{300}.

From all the above results of the visual investigation, regressional analysis, ANOVA and percent variation calculation, it can be inferred that the majority of variance generated is between y_{200} and y_{300} and can be concluded as the home to the dominant cause. We can conclude that the dominant cause lies in the heat treatment stage of the process.

Conclusion

To sum up, it has been determined that the primary factor responsible for the observed results is located in the third phase of the manufacturing process, specifically between the y_{200} and y_{300}. My findings are supported by both visual analysis of the plots and regression analysis. It was observed that 24.5% of the variation occurs prior to y_{200}, with the remaining 74.5% of the variation occurring between y_{200} and y_{300}, suggesting that the heat treatment stage is responsible for the majority of the variance.

Part 4: Search for Cause Ⅰ Ⅰ

Question

In the previous part it was determined that the dominant cause for the variation lies within phase three, between y_{200} and y_{300}, the heat treatment process, of the manufacturing. Continuing from there for this part of the report my focus is on examining various varying inputs in step three to identify the source of variation, determine the time family that is home to the dominant cause.The varying inputs in this phase are varying inputs between x_{46} and x_{60}. As the stream input is measured in the heat treatment phase I aim to answer the folowing questions:

1. Which, if any, varying inputs can be eliminated from our suspect list?
2. Is the variation caused by a part that varies stream to stream?

Plan

The plan for the investigation is to collect a sufficient amount of data for sampling such that most of the range of variation is captured as it would help identify if the dominant cause of variation is stream-to-stream in an attempt to further narrow down the varying inputs.

Since previously identified the variation is dominantly caused by the shift-to-shift time family, I collected data per shift to account for variation. I used systematic sampling to identify how the output variation varies over time and random sampling to minimise effects of confounding variables. Since it is known, the variance resides in the the heat treatment stage, I took y_{300} as the measurement output and collected x_{46} (the stream number) as a collected input to be able to conclude if measurement output depends on the stream. The collected data comprised of 10 parts every shift for 15 shifts to identify patterns, which cost me $600

Data

The investigation was conducted according to our plan (see above).I collected a total of 150 parts for this investigation. I used random and systematic sampling so I didn't capture any unexpected variation that may have been seen in consecutive sampling.

Analysis

While reviewing the Watfactory varying inputs in the heat treatment phase, between outputs y200 and y300, as it is between these outputs that the most variation is generated. This includes inputs between x46 and x60. Reviewing the associated time families (and other families) in the varying inputs and time families table (see Appendix G), where five families were identified that could be attributed to the dominant cause:

1. Part to part
2. Shift to shift
3. Stream to Stream
4. Day to Day
5. Unknown

Particular attention is paid to splitting the input into the above categories when conducting the analysis.

We will first start with a visual investigation of the data, partitioning the measurement outputs by family (ie: by day, by shift, by stream, by part and by hour) to see if there is any variation.

Since I collected the values of x_{46} for each measurement, we are able to plot each point by stream as shown in Figure 23 below.

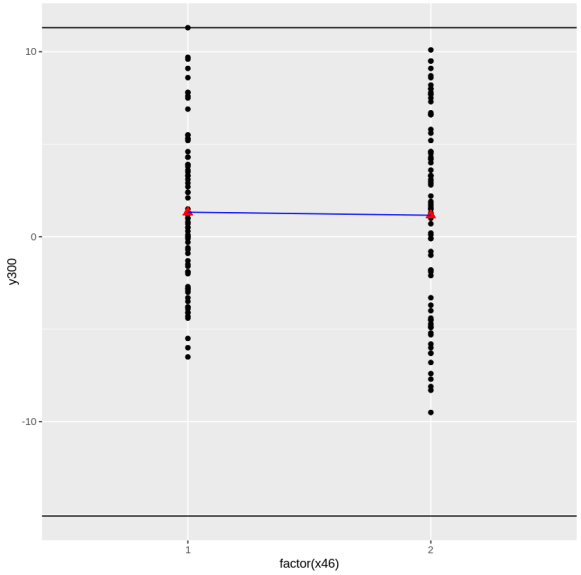


Figure 23. Boxplot of y_{300} by steam number(x_{46})

It is very evident in Figure 23 that the mean of the output y_{300} is consistent across the two streams signifying that the variation caused is not due to the steam number. Moving on I plotted the data grouped by shift as shown in Figure 24.

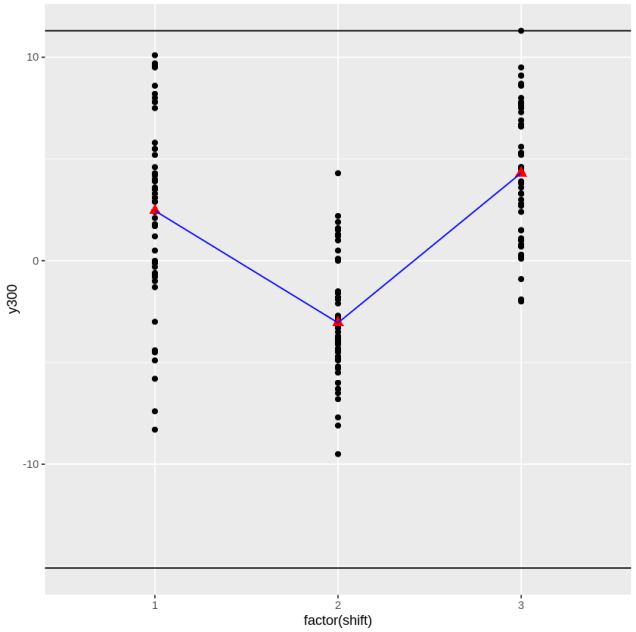


Figure 24. Boxplot for y_{300} by shift

The above graph shows significant variation in y_{300} between the shifts similar to what was seen in the baseline investigation. To determine the significance of shift on the variation in the heat treatment process I will conduct an anova analysis, but for now will move to the investigate the day to day time family as shown in Figure 25 below.

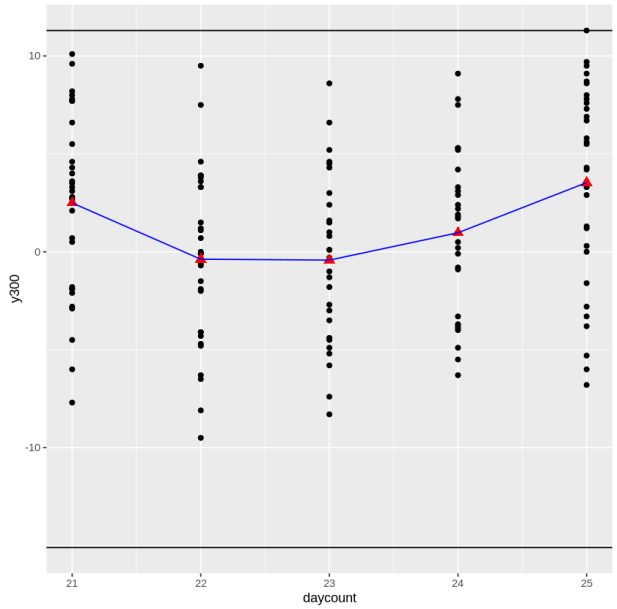


Figure 25. Boxplot of y_{300} by day

The above graph shows a slight variation in themean across all the day. There is variation in the means but compared to the variation seen in Figure 24, this variation is much lesser. Having eliminated stream to stream, I plotted the output variable by part number as shown in Figure 26 below.

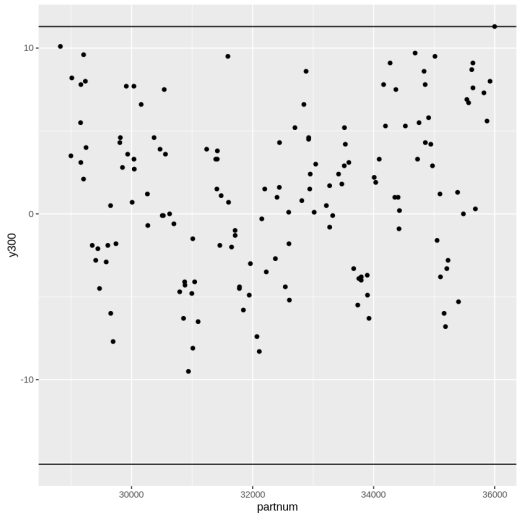


Figure 26. Boxplot of y_{300} by part

The above plot shows no specific pattern but it is evident that there is significant variation in y_{300} due to the part number. The non specific pattern in the above graph made me believe that the variation could be arising due to the hour, which could be one of the unknown families. In my pursuit to see if the variation was due to hour I plotted y_{300} grouped by hour of the shift as shown in Figure 27. There seems to be variation but due to the lack of data not much can be inferred from it.

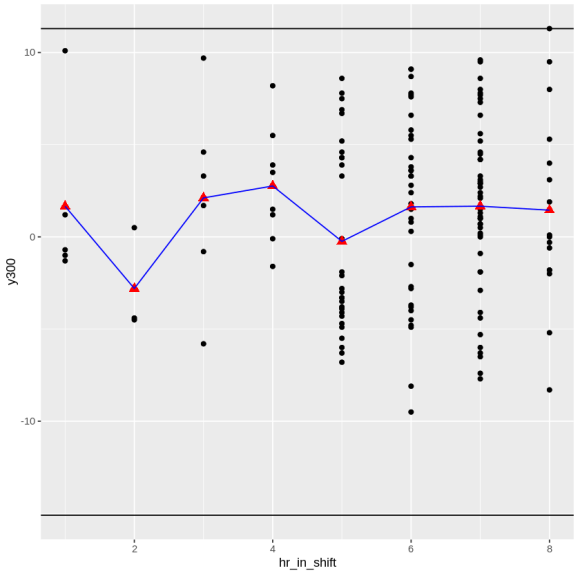


Figure 27. Boxplot of y_{300} bu hour

From the visual analysis it was found that the shift and part number have a significant impact on the observed variation. To continue this analysis, I conducted an anova analysis to get a better understanding of the impact of shift, part and day on y_{300} as shown in Appendix G.

In the anova analysis it can be seen that the both day count and shift have significant impact on the output but there is a large amount of residual which I suspect is due to part to part as seen in Figure 26. Using both the visual and anova analysis I am confident that like the baseline the variation in y_{300} is due to shift-to-shift and part-to part variation.

Conclusion

All in all, after this part of the investigation it is found that the stream number is not a contributor to the variation seen y_{300}. After investigating the different time families, I can safely eliminate three out of five suspected family, eliminating 3 components from the list of parts that could be causing the variation. Moving on for the next part I will focus on the components that belong part-part or shift-shift or unknownmtime family to find the component responsible for the the variation. There are 12 parts x_{47}, x_{48}, x_{49}, x_{50}, x_{51}, $x_{52,}x_{53}$, x_{55}, x_{56}, x_{57}, x_{58}, and x_{59} that require further investigation to find the culprit behind the variation.

Part 5: Search for Cause I I I

Question

The focus of this part is to determine the varying input that is home to the dominant cause of variation.The question that I want to answer by the end of this investigation is:

1. Which component out of x_{47}, x_{48}, x_{49}, x_{50}, x_{51}, $x_{52,}x_{53}$, x_{55}, x_{56}, x_{57}, x_{58}, and x_{59} is home to the dominant cause?

The plan for the investigation is to collect a sufficient amount of data for sampling such that the full range of variation is captured to help determine the varying input that is home to the dominant cause.

Plan

In the baseline investigation it was identified that the variation is dominantly caused by the shift-to-shift time family, so I collected data on a per shift basis. The full extent of variation in the baseline investigation was -15.1 to 11.3. I used systematic sampling to identify how the output variation varies over time and random sampling to minimise effects of confounding variables. Since we know our variance resides in the welding stage, we took y_{200} as our measurement output and collected x_{47}, x_{48}, x_{49}, x_{50}, x_{51}, $x_{52,}x_{53}$, x_{55}, x_{56}, x_{57}, x_{58}, and x_{59} to determine the varying input acting as the dominant cause of variation. We collected 8 parts per shift for each varying input for the next 5 shifts to identify patterns. 8 parts per shift were collected to ensure that there is enough data to determine groups that are significant. This data collection cost me $1,360.

Data

The investigation was conducted according to the plan. I collected a total of 40 parts for this investigation. The random and systematic sampling was used to avoid any unexpected variation that may have been seen in consecutive sampling. I re-collected data instead of using the previously collected data because the previous data does only provides enough information not to eliminate other time families, but not enough to analyze each input variable to find the main culprit for the variation.

Analysis

As established in the previous part, stream to stream is not the dominant cause and x_{47}, x_{48}, x_{49}, x_{50}, x_{51}, $x_{52,}x_{53}$, x_{55}, x_{56}, x_{57}, x_{58}, and x_{59} require further investigating. In the pursuit of finding the dominant cause I plotted each of the varying input vs the y_{300} as shown in Appendix H. As it can be seen not

much information can be gauged from visual analysis, there is no graph that shows any direct relationship between the input the and the output y_{300}.

Continuing the analysis I created a linear regression model as shown in Figure 28 below to check the significance of each of the inputs on the output y_{300}.

```
[ ] model <- lm(y300 ~ x47 + x48 + x49 + x50 + as.factor(x51) + x52 + x53 + as.factor(x55) + x56 + x57 + x58 + x59, search3)
    summary(model)

Call:
lm(formula = y300 ~ x47 + x48 + x49 + x50 + as.factor(x51) +
    x52 + x53 + as.factor(x55) + x56 + x57 + x58 + x59, data = search3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7777 -1.3160  0.3384  1.0686  3.7955

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.640366   3.039309   1.198 0.245013
x47               0.579049   0.596383   0.971 0.343178
x48               0.144929   0.117946   1.229 0.233420
x49              -0.321025   0.070829  -4.532 0.000203 ***
x50              -0.710639   0.520103  -1.366 0.186998
as.factor(x51)2 -12.352342   1.610200  -7.671 2.22e-07 ***
as.factor(x51)3  -5.772167   1.899613  -3.039 0.006486 **
as.factor(x51)4  -0.825367   1.956763  -0.422 0.677669
x52               0.043996   0.096626   0.455 0.653780
x53               0.002211   0.127844   0.017 0.986372
as.factor(x55)2   1.999998   2.085175   0.959 0.348935
as.factor(x55)3   2.549053   2.050428   1.243 0.228178
as.factor(x55)4   0.952865   1.906495   0.500 0.622671
as.factor(x55)5   1.835445   4.565358   0.402 0.691921
as.factor(x55)6   2.629024   4.203601   0.625 0.538766
as.factor(x55)7   2.526713   4.236704   0.596 0.557608
x56               0.234950   0.384680   0.611 0.548232
x57               0.022976   0.118385   0.194 0.848071
x58               0.048830   0.200647   0.243 0.810202
x59               0.078116   0.143262   0.545 0.591601
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.213 on 20 degrees of freedom
Multiple R-squared:  0.9038,    Adjusted R-squared:  0.8125
F-statistic: 9.895 on 19 and 20 DF,  p-value: 1.993e-06
```

Figure 28. Linear Regression Model

The linear regression model reveals that x_{49}, and x_{51} have significant impact on the output y_{300}. To further confirm the significance I conducted an anova analysis as shown in Figure 29 as shown below.

```
anova_model <- aov(y300 ~ x47 + x48 + x49 + x50 + as.factor(x51) + x52 + x53 + as.factor(x55) + x56 + x57 + x58 + x59, data=search3)
summary(anova_model)

                Df Sum Sq Mean Sq F value   Pr(>F)
x47              1    0.5    0.53   0.108 0.745655
x48              1  302.3  302.30  61.738 1.54e-07 ***
x49              1   93.7   93.74  19.144 0.000293 ***
x50              1   34.9   34.91   7.130 0.014704 *
as.factor(x51)   3  472.9  157.63  32.191 7.51e-08 ***
x52              1    0.5    0.47   0.096 0.760063
x53              1    0.0    0.03   0.007 0.936527
as.factor(x55)   6   11.2    1.86   0.380 0.883368
x56              1    1.9    1.93   0.394 0.537530
x57              1    0.2    0.22   0.045 0.834578
x58              1    1.0    0.96   0.195 0.663276
x59              1    1.5    1.46   0.297 0.591601
Residuals       20   97.9    4.90
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 29. ANOVA Analysis

In the above ANOVA analysis it can be seen that the along with x_{49} and x_{51}, x_{48} also has significant impact as the visual analysis does not provide any further information, it is difficult to eliminate others fron the suspect list. On close analysis of the Anova table from Figure 29, it can be noted that the F-score of x_{51} is drastically smaller in comparison to the rest signifying that x_{51} is highly significant in the variation seen in y_{300}.

Conclusion

In conclusion, the regression analysis combined with the ANOVA analysis present strong evidence that x_{51} i.e the operator, has a significant impact on the y_{300} insinuating that it is the cause of variation seen in the output. As the majority of variation is generated between y_{200} and y_{300}, it is valid to measure y_{300} and come to this conclusion.
In the baseline investigation, I discovered that the time family for variation is shift-to-shift. As the operator varies shift-to-shift, I believe that this is a reasonable finding.
I could have used retrospective analysis and saved money to as the dominant cause is x_{51} which I did not think was a likely cause.

Appendix

Appendix A: Data Analysis Code

A1. Google Collab Notebook
https://colab.research.google.com/drive/1z_XRsBWMOuLkJxjl3ielZDqAbkNJiR6r?usp=sharing

Appendix B: ANOVA Analysis for baseline investigation

```
[133] anova_model <- aov(y300 ~ as.factor(daycount), data=baseline)
      summary(anova_model)

                        Df Sum Sq Mean Sq F value   Pr(>F)
      as.factor(daycount)   4   1527   381.7   13.39 1.88e-10 ***
      Residuals           595  16967    28.5
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix B1. Anova Summary for y_{300} ~ daycount

```
[134] anova_model <- aov(y300 ~ as.factor(shift), data=baseline)
      summary(anova_model)

                      Df Sum Sq Mean Sq F value Pr(>F)
      as.factor(shift)   2   7732    3866   214.5 <2e-16 ***
      Residuals        597  10762      18
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix B2. Anova Summary for y_{300} ~ shift

```
anova_model <- aov(y300 ~ as.factor(hour), data=baseline)
summary(anova_model)

                 Df Sum Sq Mean Sq F value Pr(>F)
as.factor(hour)  24   7853   327.2   17.68 <2e-16 ***
Residuals       575  10640    18.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix B3. Anova Summary for y_{300} ~ hour

```
anova_model <- aov(y300 ~ as.factor(shift) + as.factor(hour), data=baseline)
summary(anova_model)

                 Df Sum Sq Mean Sq F value Pr(>F)
as.factor(shift)   2   7732    3866 209.306 <2e-16 ***
as.factor(hour)   23    160       7   0.377  0.997
Residuals        574  10602      18
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix B4. Anova Summary for y_{300} ~ shift + hour

Appendix C. Anova Summary for y_{300} ~ daycount + shift + hour

```
[132] anova_model <- aov(y300 ~ as.factor(daycount) + as.factor(shift) + as.factor(hour), data=baseline)
      summary(anova_model)

                         Df Sum Sq Mean Sq F value Pr(>F)
      as.factor(daycount)   4   1527     382  24.012 <2e-16 ***
      as.factor(shift)      2   7732    3866 243.190 <2e-16 ***
      as.factor(hour)      23    174       8   0.475  0.983
      Residuals           570   9061      16
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix D: Numerical Analysis

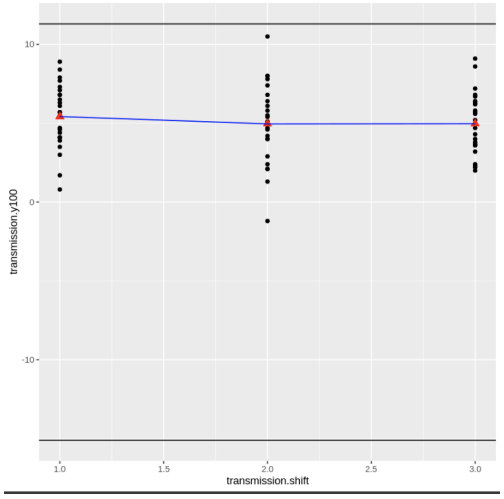


Figure D1. y_{100}

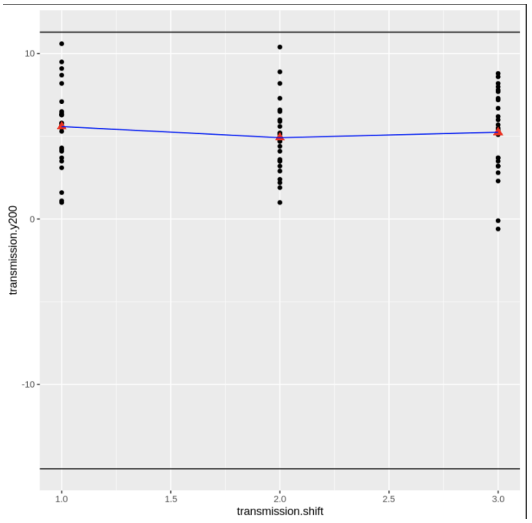


Figure D2. y_{200}

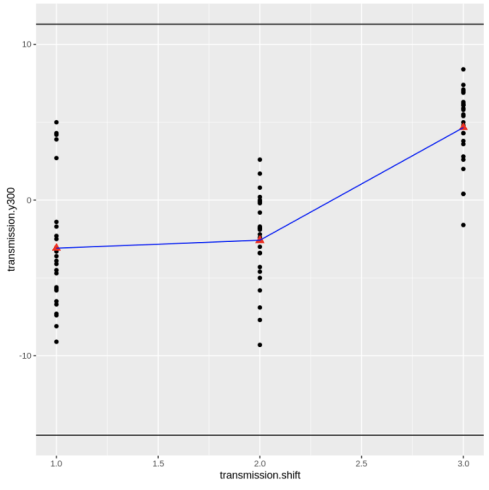


Figure D3. y_{300}

Appendix E: Regressional Analysis for process research

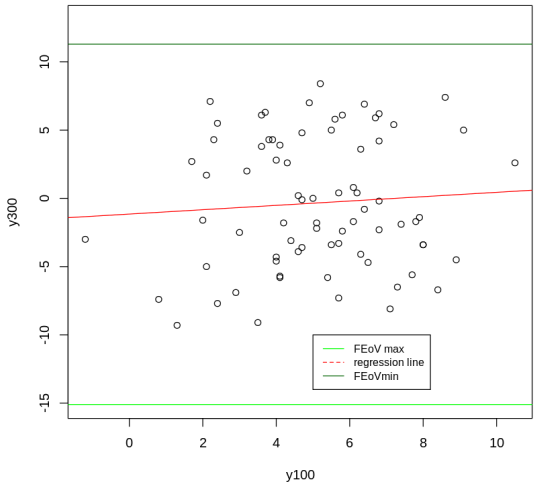


Figure E1. $y_{300} \sim y_{100}$

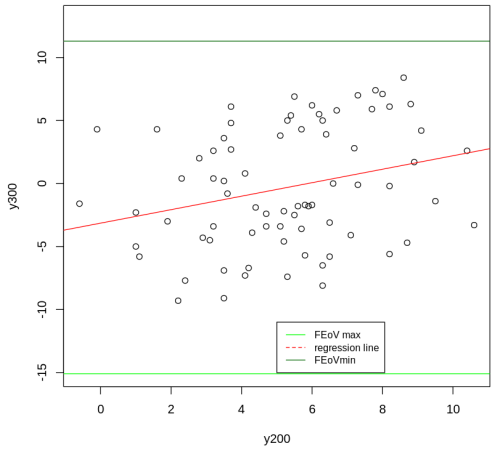


Figure E2. $y_{300} \sim y_{200}$

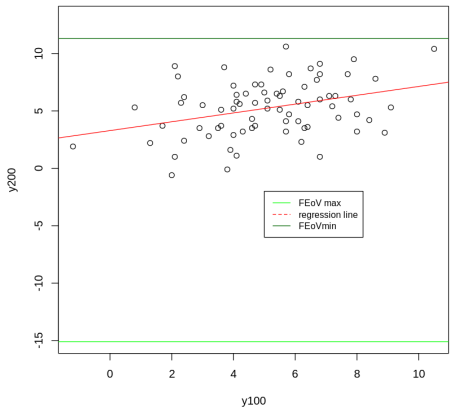


Figure E3. $y_{200} \sim y_{100}$

Appendix E: ANOVA Analysis for process search

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
y100	1	8.755622	8.755622	0.3785611	0.5402877
Residuals	73	1688.394244	23.128688	NA	NA

Figure E1. $y_{300} \sim y_{100}$

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
y200	1	125.3036	125.30363	5.819376	0.01836403
Residuals	73	1571.8462	21.53214	NA	NA

Figure E2. $y_{300} \sim y_{200}$

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
y100	1	51.04611	51.046108	9.626835	0.002726826
Residuals	73	387.08109	5.302481	NA	NA

Figure E3. $y_{200} \sim y_{100}$

Appendix F: Information on Varying Inputs in Step 3

Varying Input	Time Family	Other Families	Varying Input	Time Family	Other Families
x1	unknown	-	x31	part-to-part	-
x2	unknown	-	x32	unknown	-
x3	part-to-part	-	x33	unknown	machine-to-machine
x4	unknown	-	x34	part-to-part	-
x5	part-to-part	-	x35	unknown	-
x6	part-to-part	-	x36	unknown	machine-to-machine
x7	unknown	-	x37	unknown	-
x8	part-to-part	-	x38	part-to-part	-
x9	unknown	-	x39	part-to-part	-
x10	unknown	-	x40	unknown	machine-to-machine
x11	unknown	-	x41	part-to-part	machine-to-machine
x12	part-to-part	-	x42	unknown	-
x13	unknown	-	x43	unknown	-
x14	unknown	-	x44	shift-to-shift	-
x15	part-to-part	-	x45	unknown	-
x16	unknown	-	x46	part-to-part	-
x17	part-to-part	-	x47	unknown	-
x18	shift-to-shift	-	x48	shift-to-shift	-
x19	unknown	-	x49	part-to-part	-
x20	part-to-part	-	x50	part-to-part	stream-to-stream
x21	unknown	-	x51	shift-to-shift	-
x22	part-to-part	-	x52	unknown	-
x23	unknown	-	x53	unknown	-
x24	day-to-day	-	x54	day-to-day	stream-to-stream
x25	shift-to-shift	-	x55	part-to-part	-
x26	part-to-part	-	x56	unknown	stream-to-stream
x27	part-to-part	-	x57	unknown	-
x28	shift-to-shift	-	x58	unknown	stream-to-stream
x29	unknown	-	x59	part-to-part	-
x30	unknown	-	x60	day-to-day	-

Appendix G: ANOVA Analysis for the elimination process

```
anova_model <- aov(y300 ~ factor(daycount) + factor(shift) + factor(x46), data=stream_input)
summary(anova_model)
```

```
                  Df Sum Sq Mean Sq F value   Pr(>F)
factor(daycount)   4  368.8    92.2   8.183 5.75e-06 ***
factor(shift)      2 1466.0   733.0  65.060  < 2e-16 ***
factor(x46)        1    0.9     0.9   0.084    0.773
Residuals        142 1599.8    11.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure G1. Anova analysis $y_{300} \sim daycount + shift + stream$

Appendix H: Visual Analysis of y_{300} and input variables.