Quality Assurance 345 – ECSA GA4 Project Report

Author: Rebecca Cook

To: The Department of Industrial Engineering

Stellenbosch University

Date: October 2025

Abstract

This report presents a comprehensive quality assurance and process optimisation analysis, using a data-driven approach. The study follows integrated statistical and analytical methods to assess data integrity, process capability, and operational performance. Statistical Process Control and risk analysis on process variation is performed. Design of Experiments and ANOVA are used to identify significant differences across product types and further identify process control issues. Optimisation models are also developed to maximise profitability, while maintaining reliability, to ensure optimal staffing level decisions are made. The results of the studies highlight how statistical thinking and engineering methods can enhance process stability, operational performance and decision-making.

Table of Contents

List of Figures

Introduction

The ECSA GA4 Project, completed as part of the Quality Assurance 344 module, applies statistical and analytical engineering tools to real-world data. The focus is on improving operational performance by extracting useful insights, assessing process stability, and supporting data-driven decisions that enhance quality and profitability. The five main objectives of the project are: (1) to perform a data integrity analysis and descriptive analysis, (2) to perform a Statistical Process Control analysis, (3) to identify the risks of false or missed alarms associated with making a type 1 and type 2 error, (4) to design and evaluate experiments that signify process differences across products and time periods, and (5) to build profit-optimisation models that enhance decision-making. Overall, the project showcases a range of analytical skills and process improvement tools used in quality assurance; highlighting how statistical reasoning and data-driven approaches can help managerial decision-making in the context of Industrial Engineering and continuous improvement.

Basic Data Analysis (Part 1.2)

Objective

This section of the report presents a basic data analysis on the given datasets: products, products_headoffice, customers and sales (2022 and 2023). The goal is to understand the structure, quality and content of the data and explore relationships and trends within the data which can provide insights to aid management in making data-driven decisions.

Data Overview

Dataset	Rows	Columns	Description (column names)
products	60	5	ProductID Category Description SellingPrice Markup
products_headoffice	360	5	ProductID Category Description SellingPrice Markup
customers	5 000	5	CustomerID Gender Age Income City
sales	100 000	9	CustomerID ProductID Quantity orderTime orderDay orderMonth orderYear pickingHours deliveryHours

Table 1: Data Overview

Data Quality Issues

There are no missing values in any of the datasets. There are also no duplicated instances. There are no rows with negative delivery or picking times. Data quality issues (of which there are many) relating to the products_headoffice dataset, and products dataset are explored later.

Descriptive/Summary Statistics

```
── Data Summary ────────────────────────
                           Values
Name                       products
Number of rows             60
Number of columns          5
_________________________
Column type frequency:
  character                3
  numeric                  2
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 ProductID             0             1   6   6     0       60          0
2 Category              0             1   5  18     0        6          0
3 Description           0             1   9  21     0       35          0

── Variable type: numeric ────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd    p0   p25   p50    p75    p100 hist
1 SellingPrice          0             1 4494.  6504.   350.  512.  794.  6417.  19725. ▇▁▁▁▁
2 Markup                0             1   20.5    6.07  10.1  16.1  20.3   25.7    29.8 ▆▆▇▆▇
```

Figure 1: Summary statistics of the products dataset

Comments:

- The SellingPrice has a mean of 4494 which is much greater than the median of 794. This indicates a right-skewed distribution, with potential outliers bringing up the mean, but majority of the data instances having lower values. This indicates that the majority of the products are in the low- to mid-price range, with a few very high-priced products skewing the average upward. These premium products should be handled separately to understand their revenue contribution and to tailor marketing and inventory strategies accordingly.

```
── Data Summary ────────────────────────
                           Values
Name                       products_ho_raw
Number of rows             360
Number of columns          5
_______________________
Column type frequency:
  character                3
  numeric                  2
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 ProductID             0             1   5   6     0      110          0
2 Category              0             1   5  18     0        6          0
3 Description           0             1   9  24     0       60          0

── Variable type: numeric ────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd    p0   p25   p50    p75   p100 hist
1 SellingPrice          0             1 4411.  6464.   291.  496.  797.  5843.  22420. ▇▁▁▁▁
2 Markup                0             1   20.4    5.67  10.1  15.8  20.6   24.8    30  ▆▆▇▇▇
```

Figure 2: Summary statistics of the products_headoffice dataset

The products_headoffice dataset reveals very similar summary statistics as the products dataset, however there are 360 rows in the products_headoffice dataset, compared to 60 in the products_dataset.

The number of unique IDs being 110 in the productID column is an indication of errors in the dataset, as 360 unique IDs are expected.

```
> head(products_raw)
# A tibble: 6 × 5
  ProductID Category           Description      SellingPrice Markup
  <chr>     <chr>              <chr>                   <dbl>  <dbl>
1 SOF001    Software           coral matt               512.   25.0
2 SOF002    Cloud Subscription cyan silk                505.   10.4
3 SOF003    Laptop             burlywood marble         494.   16.2
4 SOF004    Monitor            blue silk                543.   17.2
5 SOF005    Keyboard           aliceblue wood           516.   11.0
6 SOF006    Mouse              black silk               479.   17.0
> head(products_ho_raw)
# A tibble: 6 × 5
  ProductID Category Description           SellingPrice Markup
  <chr>     <chr>    <chr>                        <dbl>  <dbl>
1 SOF001    Software coral silk                    522.   15.6
2 SOF002    Software black silk                    467.   28.4
3 SOF003    Software burlywood marble              496.   20.1
4 SOF004    Software black marble                  389.   17.2
5 SOF005    Software chartreuse sandpaper          483.   17.6
6 SOF006    Software cornflowerblue marble         539.   25.6
```

Figure 3: Glimpse of the products and products_headoffice datasets, showing the discrepencies in the datasets

Products_headoffice has ten times the amount of instances as products – likely indicating some discrepancies or errors in the data. Viewing the first few instances of both datasets, as in Figure 3, shows an issue in the datasets that the common ProductID in both datasets has different descriptions, selling price and markups.

The products dataset also has errors, where the key on the product ID indicates SOF (i.e. software), but the category is not always Software when it should be, and similarly for the other product types.

	ProductID	Category	Description	SellingPrice	Markup	type
1	SOF001	Software	coral silk	521.72	15.65	SOF
2	SOF002	Software	black silk	466.95	28.42	SOF
3	SOF003	Software	burlywood marble	496.43	20.07	SOF
4	SOF004	Software	black marble	389.33	17.25	SOF
5	SOF005	Software	chartreuse sandpaper	482.64	17.60	SOF
6	SOF006	Software	cornflowerblue marble	539.33	25.57	SOF
7	SOF007	Software	blue marble	495.13	10.23	SOF
8	SOF008	Software	cornflowerblue marble	465.73	21.89	SOF
9	SOF009	Software	black bright	452.40	19.64	SOF
10	SOF010	Software	cornflowerblue matt	399.43	17.08	SOF
11	NA011	Software	aliceblue silk	823.51	14.59	NA0
12	NA012	Software	coral marble	987.13	27.59	NA0
13	NA013	Software	cornflowerblue sandpaper	1176.31	18.30	NA0

Figure 4: Error in products_headoffice productID column

ProductID begins with "NA", when it should begin with "SOF" as it is in the Software category. This error is repeated for the 11-60th instances of each product type.

Further investigation reveals that the number of rows where ProductID in the sales dataset is not in the products data is 0, but it is 79251 for the products_headoffice dataset. This shows that the values in the products_headoffice data are likely not correct, and should be replaced with the correct values in the products dataset.

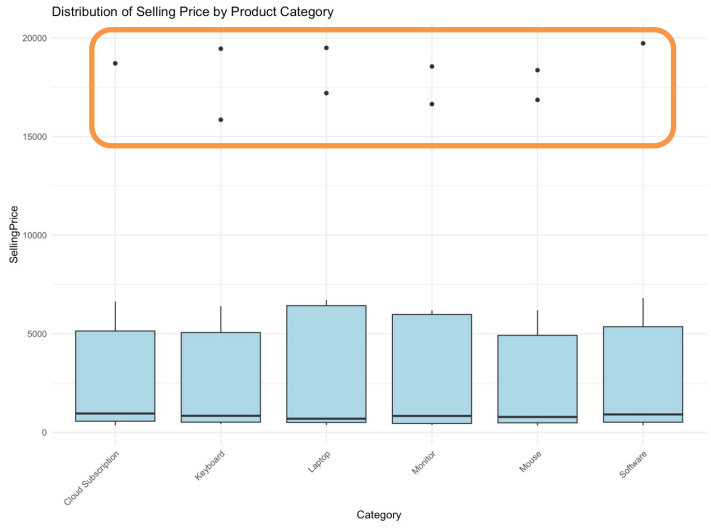

Figure 5: Distributions of Selling Price per product category

Comments:

- High outliers can be seen in each category, indicating the premium products that are much more expensive than most of the products within each category. Figure 5 proves the interpretation above of a few high-priced items that are pulling the mean selling price up. These products should be treated separately or should be aimed towards a specific customer base with higher income.

```
── Data Summary ────────────────────────
                           Values
Name                       customers
Number of rows             5000
Number of columns          5
_________________________
Column type frequency:
  character                3
  numeric                  2
_________________________
Group variables            None

── Variable type: character ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 CustomerID            0             1   7   8     0     5000          0
2 Gender                0             1   4   6     0        3          0
3 City                  0             1   5  13     0        7          0

── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean      sd   p0   p25   p50    p75   p100 hist
1 Age                   0             1    51.6    21.2   16    33    51     68    105 ▇▇▆▃▂
2 Income                0             1 80797   33150.  5000 55000 85000 105000 140000 ▂▅▇▃▂
```

Figure 6: Summary statistics of the customers dataset

Comments:

- The Income for the customers has a left-skewed distribution, with majority of customers around the median value of R85 000, but a few customers with a much lower income, causing the mean value to be R80 797.

```
── Data Summary ────────────────────────
                           Values
Name                       sales
Number of rows             100000
Number of columns          9
_________________________
Column type frequency:
  character                2
  numeric                  7
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 CustomerID            0             1   7   8     0     5000          0
2 ProductID             0             1   6   6     0       60          0

── Variable type: numeric ────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean     sd       p0     p25     p50    p75   p100 hist
1 Quantity              0             1   13.5   13.8        1       3       6     23     50
2 orderTime             0             1   12.9    5.50       1       9      13     17     23
3 orderDay              0             1   15.5    8.65       1       8      15     23     30
4 orderMonth            0             1    6.45   3.28       1       4       6      9     12
5 orderYear             0             1 2022.     0.499 2022      2022    2022   2023   2023
6 pickingHours          0             1   14.7   10.4        0.426   9.39   14.1   18.7   45.1
7 deliveryHours         0             1   17.5   10.00       0.277  11.5    19.5   25.0   38.0
> |
```

Figure 7: Summary statistics of the sales dataset

Comments:

- Quantity has a right-skewed distribution with the median quantity per order of 6, compared the mean of 13.5. Although the average order size is 13.5 units, half of all orders are 6 units or less, indicating that the majority of the business consists of small orders, with a minority of very large orders skewing the average upward. This suggests different customer segments or ordering patterns which may require further investigation and tailored handling. This distribution is shown below in Figure 8.

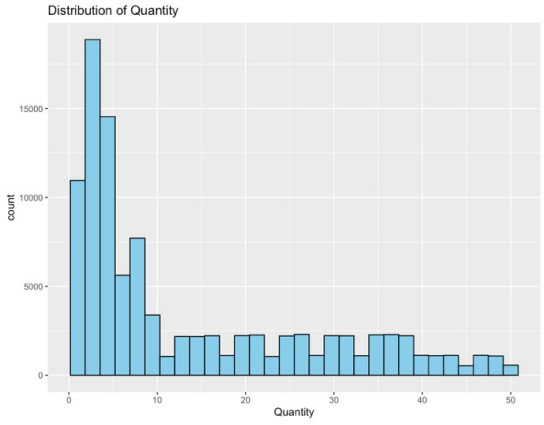

Figure 8:Distribution of order quantity

- DeliveryHours has a multimodal distribution for both 2022 and 2023. This is because the Software product type has a much quicker delivery time than the other physical products (it can be downloaded remotely in minutes). There is no clear difference in delivery times across the two years.

Relationships/Trends

Figure 9: Monthly sales trend

Figure 9 shows clear **seasonal dips** in monthly revenue during **December and January** in both 2022/2023 and December 2023. This likely reflects a combination of holiday closures in December and early January, and internal capacity reductions. Management should plan for these slowdowns — for example by adjusting inventory, staffing, or cash flow expectations — and consider pre-emptive promotions to smooth demand around the holiday period.

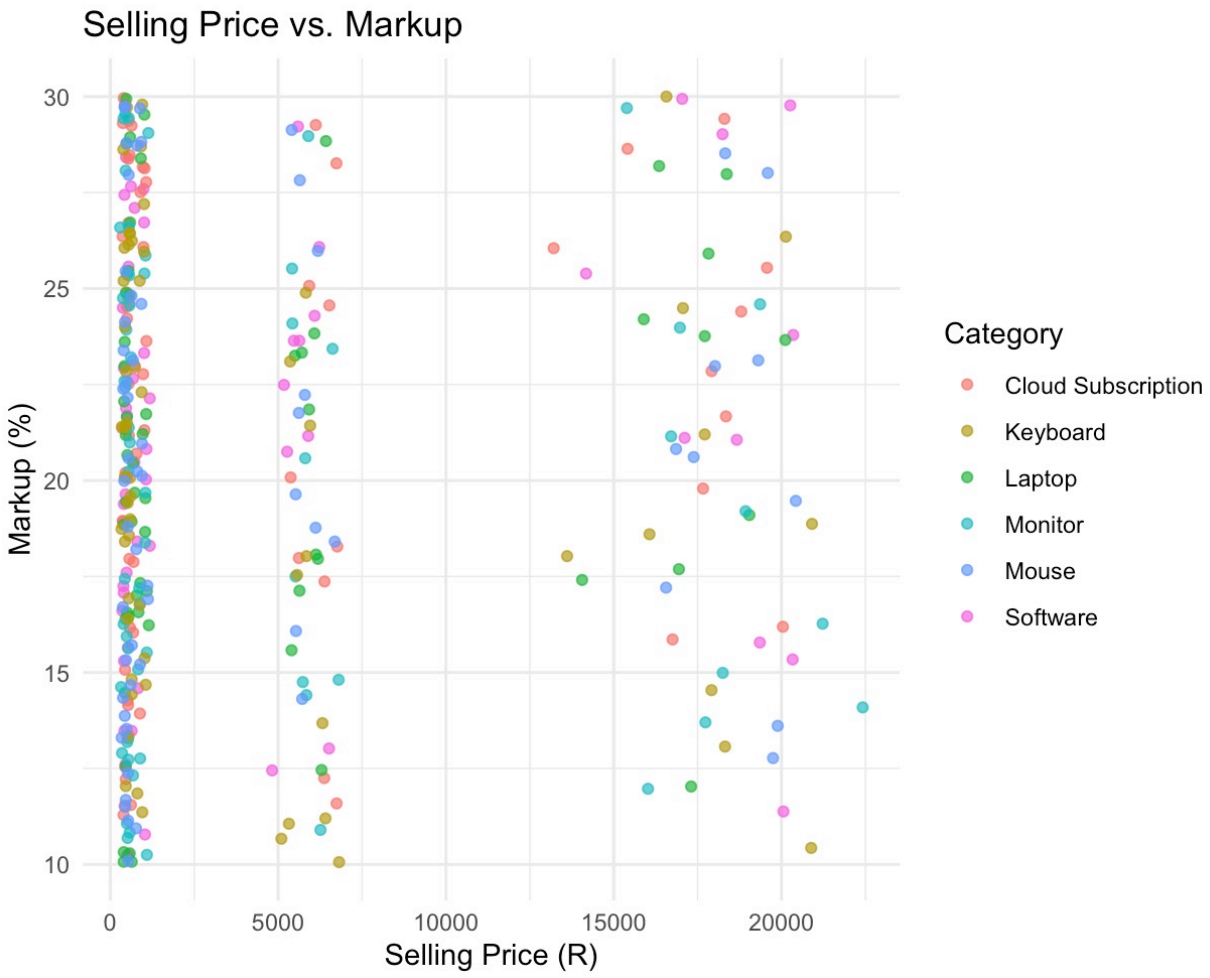

Figure 10: Selling Price vs Markup

Figure 10 shows no clear relationship between selling price and mark-up but rather shows clusters of different products across all the categories, that have low, high and extremely high selling prices, with markup varying largely for each SKU. It seems that the company does not have a strategy of higher mark-ups for certain categories, or for higher selling prices.

	ProductID	total_revenue	pct_of_total
1	LAP025	281754471	0.11403304
2	LAP023	265237837	0.10734835
3	LAP024	256255268	0.10371288
4	LAP027	254026069	0.10281066
5	LAP021	250568078	0.10141113
6	LAP026	241231494	0.09763238
7	LAP028	241001543	0.09753932
8	LAP030	236466128	0.09570372
9	LAP022	233984304	0.09469926
10	LAP029	210289183	0.08510926

Table 2: Highest revenue contributing SKUs

Revenue is calculated by multiplying order quantity for each order by the selling price of the product ordered. The proportion that each product contributes to the total revenue is then also calculated.

Table 2 shows that the top 10 products (in the products dataset), which are all variants of laptops, contribute to about 57% of the total revenue.

This high concentration represents both an opportunity (focus on top sellers) and a risk (excess dependence on few SKUs). Prioritise stock and supplier relationships for these items.

Key findings and recommendations

1. **Pricing structure:**

   Most products are in the low-mid selling price range, but there are a few products with a much higher selling price.

   These premium products across all categories should be marked separately to high-income customers.

2. **Customer characteristics:**

   Customer income is left-skewed with majority of customers earning around R85000, and a few low-earning customers.

   Segmentation by income could support target marketing, pricing and credit terms.

3. **Sales characteristics:**

   Order quantities are mostly around 6 units, but larger orders bring up the mean order size.

   This indicates many small-order buyers, and a few large-order buyers.

   Different order fulfilment policies may be needed for the different groups.

4. **Pricing strategy and mark-up:**

   There is no consistent pattern of higher-markups on higher-priced products.

   This could indicate deliberate flexibility or a suboptimal mark-up policy which may require reviewing.

5. **Revenue concentration:**

   The top 10 revenue contributors contribute to nearly 60% of total revenue.

   These SKUs are all laptops. There is an opportunity to focus on these SKUs for marketing opportunities, yet also an increased risk if the supply or demand of one of these products decreases).

6. **Inconsistent datasets:**

   Data issues need to be investigated to determine which datasets are correct and where the issues lie.

Data analysis on fixed datasets (Q4.3)

```
── Data Summary ────────────────────────
                           Values
Name                       products_correct
Number of rows             60
Number of columns          5
_______________________
Column type frequency:
  character                3
  numeric                  2
_______________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 ProductID             0             1   6   6     0       60          0
2 Category              0             1   5  18     0        6          0
3 Description           0             1   9  21     0       35          0

── Variable type: numeric ────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd    p0   p25   p50    p75   p100 hist
1 SellingPrice          0             1 4494.  6504.   350.  512.  794.  6417.  19725. ▇▁▁▁▁
2 Markup                0             1   20.5    6.07  10.1  16.1  20.3   25.7   29.8 ▇▇▇▇▇
```

Figure 11: Corrected summary statistics of the products dataset

```
── Data Summary ────────────────────────
                           Values
Name                       products_ho_correct
Number of rows             360
Number of columns          5
_______________________
Column type frequency:
  character                3
  numeric                  2
_______________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 ProductID             0             1   6   6     0      360          0
2 Category              0             1   5  18     0        6          0
3 Description           0             1   9  24     0       60          0

── Variable type: numeric ────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd    p0   p25   p50    p75   p100 hist
1 SellingPrice          0             1 4494.  6458.   350.  512.  794.  6417.  19725. ▇▁▁▁▁
2 Markup                0             1   20.5    6.03  10.1  16.1  20.3   25.7   29.8 ▇▇▇▇▇
```

Figure 12: Corrected summary statistics of the products_headoffice dataset

Comparison to before data cleaning and fixing errors:

- The number of unique productIDs in the products head_office dataset is now 360, as it should be.
- The distributions of SellingPrice and Markup remained unchanged in the products dataset, however the distribution in the headoffice dataset was corrected to reflect the same values as the original products dataset.
- Comments made on the pricing structure and customer characteristics were correct
- Previous revenue analysis on the products dataset was not correct:

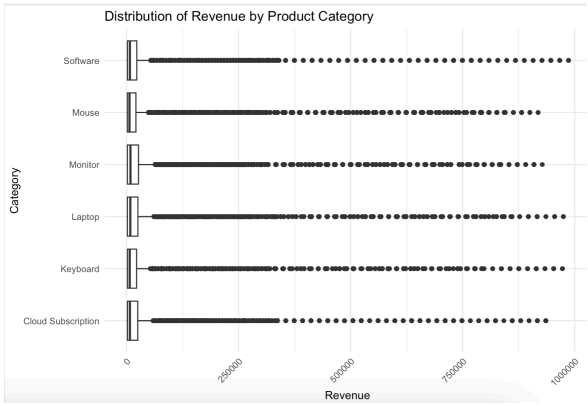

Figure 13: Revenue per product type (old dataset)

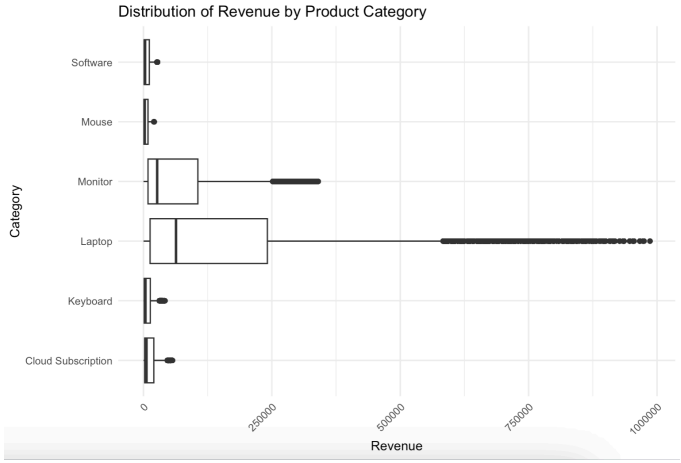

Figure 14: Revenue per product type (updated dataset)

- Figure 14 shows the correct revenue distribution, and highlights the fact that laptops contribute to a significantly large portion of the company's revenue, with monitors yielding the second most revenue.
- Revenue seasonality remains unchanged, with seasonal variation

- Distribution of selling price per category was incorrectly interpreted with the erroneous dataset. The distribution is not as uniform as Figure 5 suggested, however it shows that the selling price is very distinct per product category (laptop has the highest, and the monitor, and then cloud subscription, with the other distributions being relatively similar and lower).

Statistical Process Control (Part 3)

Objective

This section of the report analyses the delivery time process for each of the product types, using Statistical Process Control (SPC) methods. Construct X-bar and S charts for each product type. Identify control issues according to standard rules. Calculate process capability indices (Cp, Cpk, Cpu, Cpl) for the first 1000 deliveries per product type. Assess which products meet the Voice of the Customer (VOC) requirements.

Data Preparation

Load the *sales2026and2027* dataset, and arrange chronologically by year, month, day and time. Plotting the distribution of the *deliveryHours* reveals a multimodal distribution, but this is due to the Software product types that are delivered very quickly (due to the nature of the products). Data transformations are not required, as distributions do approximate a normal distribution.

Figure 15: Distribution of delivery times

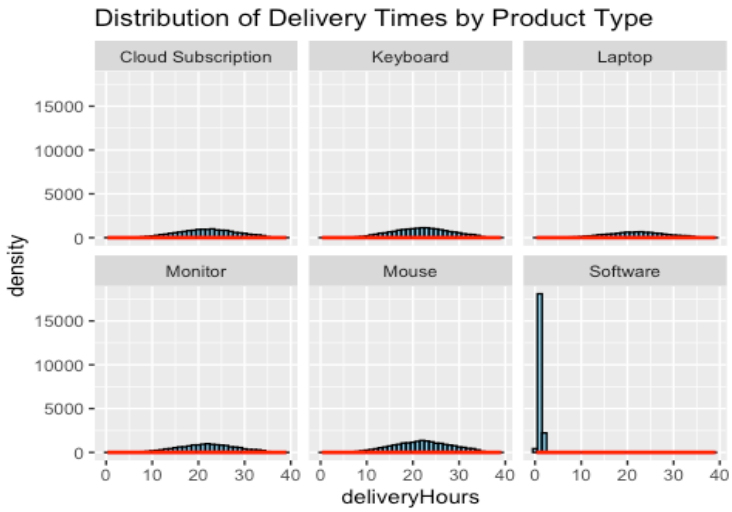

Figure 16: Distribution of delivery times per product category

SPC charts (x-bar and s)

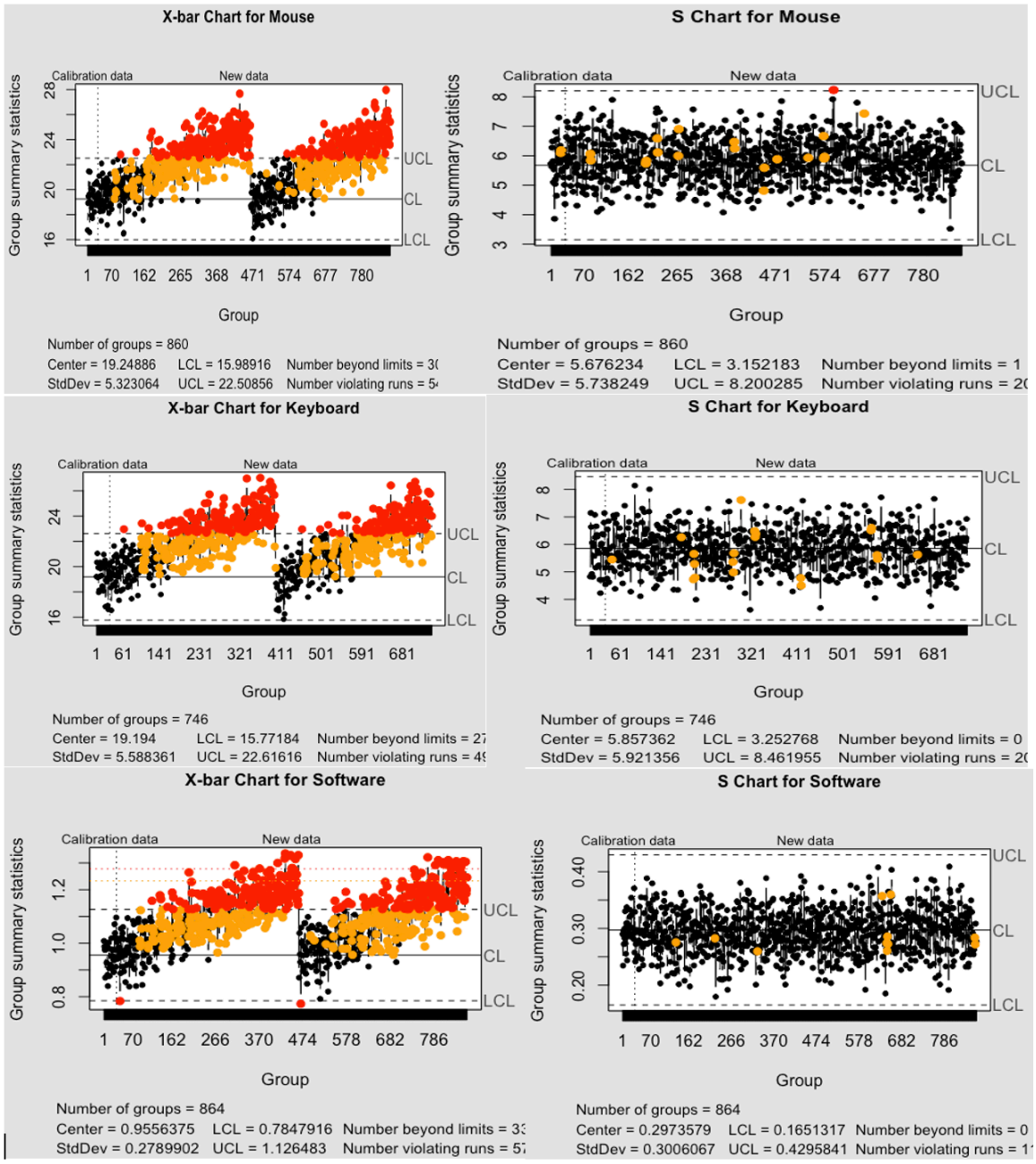

Figure 17: s and x-bar charts for Mouse, Keyboard and software

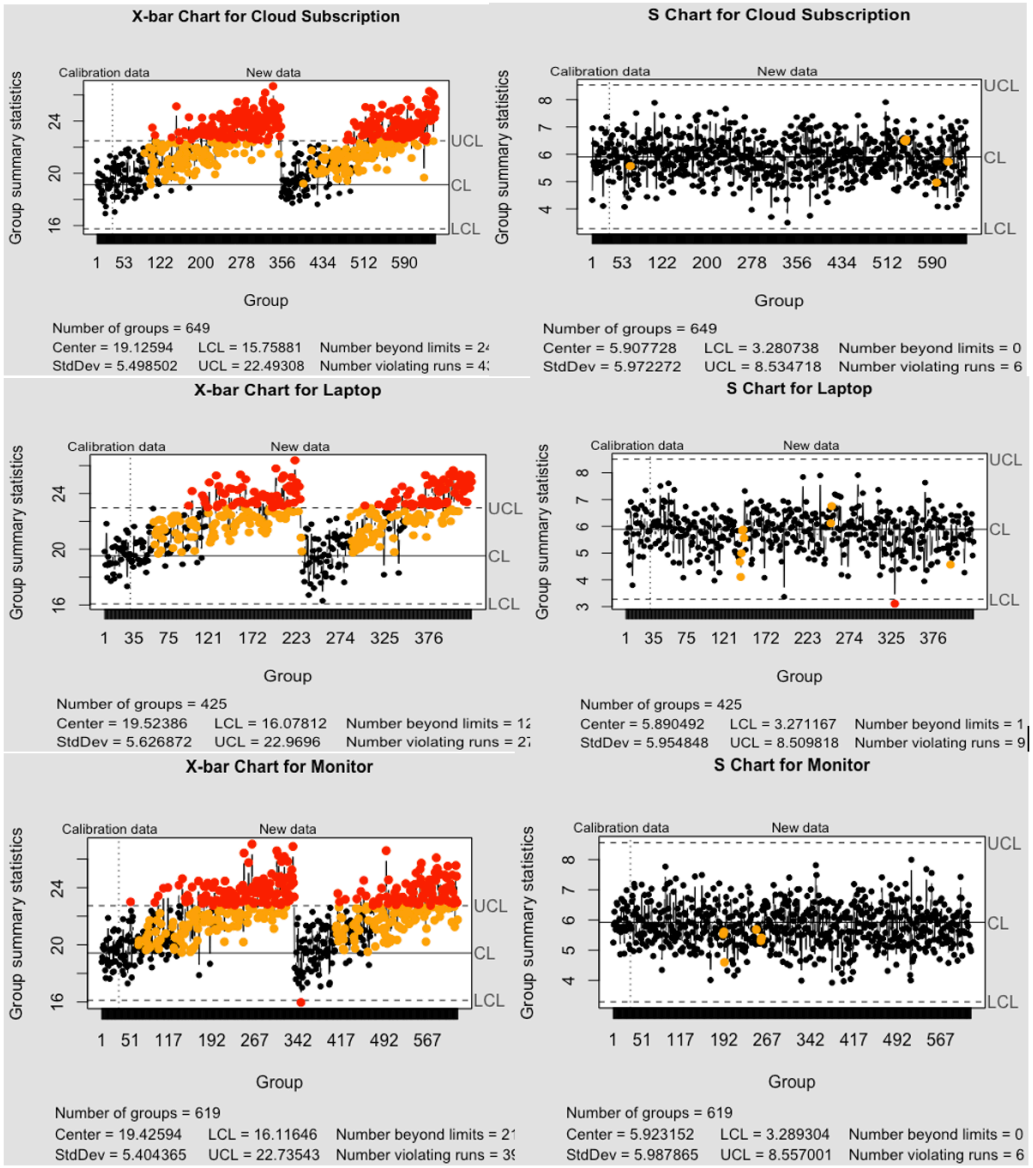

Figure 18: x-bar and s charts for Cloud Subscription, Laptop and Monitor

Notes on the SPC charts (answers to question 3.3):

- SPC charts show that all processes are in-control for the first 30 samples, but then run out of control, before being reset (at the beginning of a new year, potentially due to recalibration and training at the start of 2023).
- Shows poor management in not monitoring the process real-time
- Management should be on the floor assessing the reasons for any out-of-control signals as they occur, so that changes can be made before the process continually grows out of control as workers slack off, or machines need re-calibration.

Process Capability Analysis (Part 3.3)

ProductType	Cp	Cpl	Cpu	Cpk	Capability
Mouse	0.9151848	1.103799	0.7265710	0.7265710	Not Capable
Keyboard	0.9171375	1.104921	0.7293536	0.7293536	Not Capable
Software	18.1352369	1.082872	35.1876018	1.0828720	Not Capable
Cloud Subscription	0.8977458	1.078754	0.7167378	0.7167378	Not Capable
Laptop	0.8987816	1.101345	0.6962187	0.6962187	Not Capable
Monitor	0.8890490	1.078528	0.6995705	0.6995705	Not Capable

Figure 19: Process capability indices

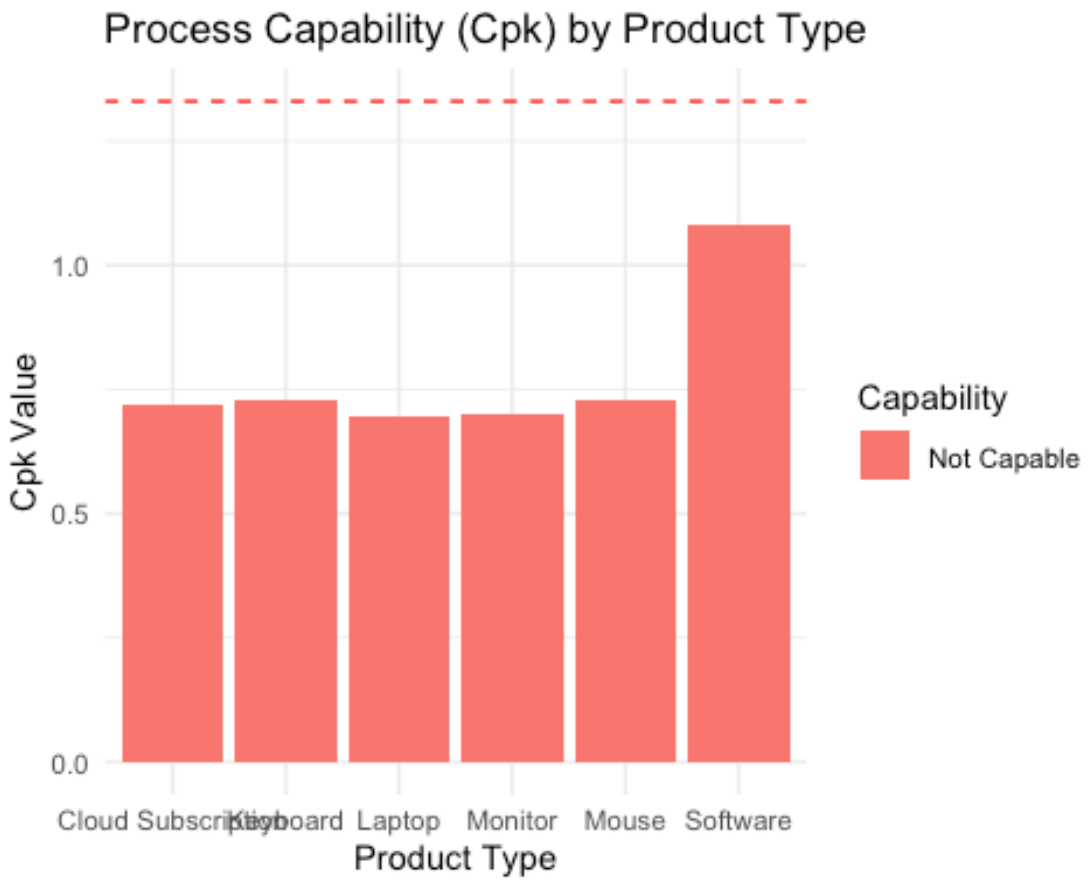

Figure 20: Visualisation of Cpk per product type

Using the first 1000 deliveries per product type, and calculating capacity indices, it can be seen that the process delivery times of all product types are below an acceptable industry standard Cpk of 1.33. Software has a CPK >1, indicating a marginally capable process. None of the product types are therefore capable of meeting the VOC requirements. However, the CPK looks at both the upper and lower limits, whereas it is only practical to look at the upper limits of a delivery time process. Software is capable on the upper side, CPU=35.188, indicating very capable of

meeting delivery requirements quick enough. On the lower side, all products are just capable (but less than 1.33), this is not an important factor to consider as there should not be a lower limit on how fast the process delivers the products. SOFTWARE limits do not make sense – the upper limit should NOT be 32 hours for delivery of software, the delivery process should take a couple of minutes only.

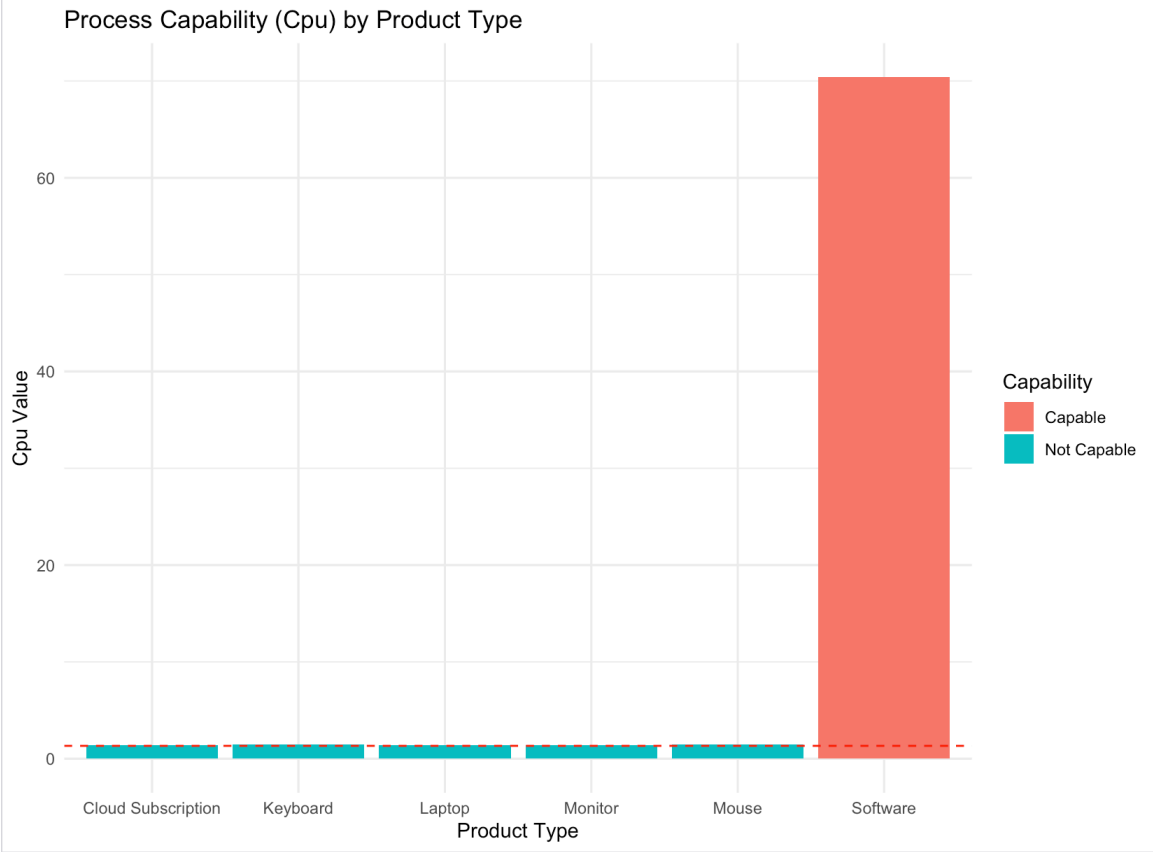

Figure 21: CPU indices per product type

Figure 21 shows Process Capabilities based only on the upper limit of 32hours. This figure shows that Software is being delivered well under 32hours consistently (the only capable delivery time process).

Control Issue Analysis (Part 3.4)

A) S samples above +3σ

```
Product: Mouse
First 3 samples above +3σ: 77 78 79
Last 3 samples above +3σ: 858 859 860
Total samples above +3σ: 540

Product: Keyboard
First 3 samples above +3σ: 100 101 102
Last 3 samples above +3σ: 744 745 746
Total samples above +3σ: 492

Product: Software
First 3 samples above +3σ: 86 87 88
Last 3 samples above +3σ: 862 863 864
Total samples above +3σ: 578

Product: Cloud Subscription
First 3 samples above +3σ: 92 101 102
Last 3 samples above +3σ: 647 648 649
Total samples above +3σ: 434

Product: Laptop
First 3 samples above +3σ: 58 59 60
Last 3 samples above +3σ: 423 424 425
Total samples above +3σ: 272

Product: Monitor
First 3 samples above +3σ: 65 66 77
Last 3 samples above +3σ: 617 618 619
Total samples above +3σ: 395
```

Figure 22

Notes:

- Each sample above 3σ, is an indicator of unusually high variability, and an instable process.
- As soon as the first of these “out-of-control” signals occurred, management should investigate the cause and fix it.
- The data shows that this has not been the case, with a very large number of samples above 3σ, showing that the delivery process is NOT under statistical control for any of the product categories. This is likely due to external causes like machine wear, or inconsistent materials and labour methods.
- Some products exhibit more unstable variation and behaviour than others, with the Mouse being the most unstable (540 samples above 3σ), while the Laptop has 272 samples above 3σ, still not a good indicator at all, but shows that the process has slightly less variation and is more stable than the delivery of the mouse and other products.
- The process capability estimates from these s-charts are unreliable, until stability is restored
- Consider recalibrating equipment, implementing consistent and preventative maintenance, as well as process standardisation techniques.
- The production environment or data collecting system may be inconsistent across time or locations

B) Longest consecutive S samples within ±1σ

```
Product: Mouse - Max consecutive samples within ±1σ: 4
Product: Keyboard - Max consecutive samples within ±1σ: 7
Product: Software - Max consecutive samples within ±1σ: 4
Product: Cloud Subscription - Max consecutive samples within ±1σ: 6
Product: Laptop - Max consecutive samples within ±1σ: 7
Product: Monitor - Max consecutive samples within ±1σ: 6
```

Figure 23

Notes:

- A maximum of 7 consecutive samples within 1 standard deviation of the mean for the keyboards, and even lower maximums for the other products, clearly highlights process control issues.

C) Four consecutive X-bar samples above +2σ

```
Product: Mouse
  Product RunNumber StartIndex EndIndex RunLength
1   Mouse         1          1       30        30
Total runs of 4+ consecutive samples above +2σ: 1

Product: Keyboard
   Product RunNumber StartIndex EndIndex RunLength
1 Keyboard         1          1       30        30
Total runs of 4+ consecutive samples above +2σ: 1

Product: Software
    Product RunNumber StartIndex EndIndex RunLength
30 Software         1          1       29        29
Total runs of 4+ consecutive samples above +2σ: 1

Product: Cloud Subscription
             Product RunNumber StartIndex EndIndex RunLength
1 Cloud Subscription         1          1       30        30
Total runs of 4+ consecutive samples above +2σ: 1

Product: Laptop
  Product RunNumber StartIndex EndIndex RunLength
1  Laptop         1          1       30        30
Total runs of 4+ consecutive samples above +2σ: 1

Product: Monitor
   Product RunNumber StartIndex EndIndex RunLength
12 Monitor         1          1       11        11
   Monitor         2         13       30        18
Total runs of 4+ consecutive samples above +2σ: 2
```

Figure 24

Notes:

- It is very concerning to see at least one run of 4+ consecutive samples for each product
- Monitor has two runs of consecutively being above the 2sigma control limit, which indicates extremely poor process control in delivery times, however the run length is lower than that of the other products, which are all about 30 consecutive runs above the 2sigma control limit, further indicating poor process control.

Risk

Probability of making a type 1 error

- H_0 : The process is in control, centred, and stable.
- H_1 : The process is out of control (mean shift or increase in variation)

The Type I error is the probability of rejecting H_0 when it is true – in other words, signalling that the process is "out of control" when it is not.

Rule A: One sample above +3σ

Rule B: k consecutive samples between ±1σ (k = most consecutive samples of s between the -1 and +1 sigma-control limits for each product type)

Rule C: Four consecutive samples above +2σ

Table (Type I probabilities)

Product	Rule A: P(one > +3σ)	Rule B: P(k consecutive inside ±1σ)	Rule B (%)	Rule C: P(4 consecutive > +2σ)
Mouse	0.0013498980	0.2172165308	21.721653%	0.00000026787715598040045
Keyboard	0.0013498980	0.0691134429	6.911344%	0.00000026787715598040045
Software	0.0013498980	0.2172165308	21.721653%	0.00000026787715598040045
Cloud Subscription	0.0013498980	0.1012370100	10.123701%	0.00000026787715598040045
Laptop	0.0013498980	0.0691134429	6.911344%	0.00000026787715598040045
Monitor	0.0013498980	0.1012370100	10.123701%	0.00000026787715598040045

Figure 25: Type 1 error probabilities

Figure 25, generated by OpenAI, summarises the theoretical likelihood of making a type 1 error (manufacturer's error) for each SPC rule. These percentages represent the likelihood of false alarms for each process. If an alarm goes off for Rule C, it is extremely unlikely that it is a false

alarm. The probability of 4 consecutive samples *above* +2σ is extremely tiny under H_0 indicating that if such a run occurs it is a very strong signal. For Rule B, there is quite a high chance that the out-of-control alarm is not correct. Rule B values vary by product because it uses the observed *maximum run length k* within ±1σ in p^k, where p=P(-1<Z<1) = 0.6826894921. Rule A is the same for every product because it's the theoretical chance a sample mean lies above +3σ, which is ~0.00135 (≈0.135%). This indicates a low likelihood of a false alarm if such a run occurs in a centred and stable process.

Probability of making a type 2 error

A Type II error (β) is the probability the chart fails to signal (i.e., sample statistic falls inside the control limits) when the process has shifted to a new mean (or new σ).

	mu1	sigma_xbar	beta	power	beta_pct	power_pct
1	25.028	0.017	0.8411783	0.1588217	84.1178	15.8822
2	25.028	0.013	0.9045098	0.0954902	90.4510	9.5490

Figure 26: Type 2 error probabilities

Figure 26 shows the type 2 errors for the increased variation (row 1 in the table) and the original scenario (row 2). Beta is the probability of making a type 2 error (Consumer's error), and the power is (1- β), which is the probability of a correct decision – in other words, the probability the shift is detected correctly.

Both scenarios have a high beta value, meaning the chart is highly unlikely to detect a shift in the mean or variation (low sensitivity to each shift). If detecting a shift is crucial: sample size can be increased, tighter control rules can be used, or process variability can be reduced to improve sensitivity.

Optimising profit (Part 5)

Objective

This section develops a simple analytical model to determine the optimal number of baristas at two coffee shops in order to maximise profits, while maintaining efficient and reliable service. The model uses historical data of the number of baristas and the time taken for a customer to be served (found in TimeToServe.csv and TimeToServe2.csv), calculates the mean service time per barista, and converts this to an average number of customers served per day (throughput). The daily profits are then calculated to evaluate the trade-off between operational costs (R1000 per barista per day) and customer revenue (R30 profit per customer served). The optimal number of baristas is then identified for each shop. The differences between the two shops show that the optimal number of baristas depends on both process speed and labour cost structure.

Results

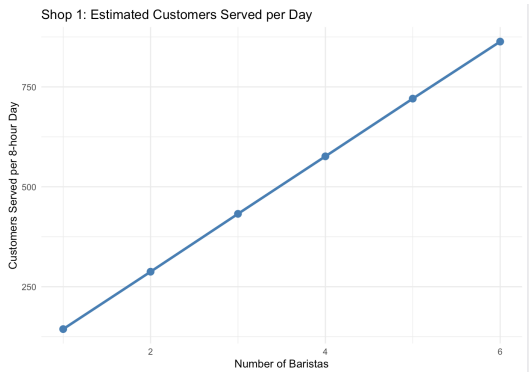

Figure 27: Customers served per number of baristas at shop 1

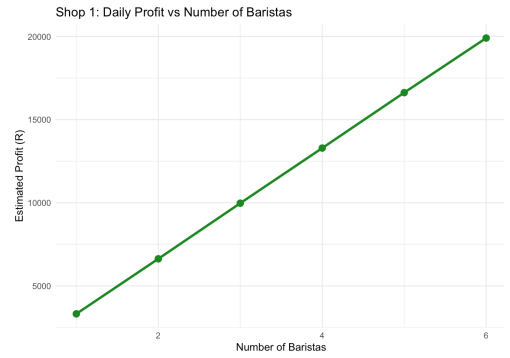

Figure 28: Daily profit per number of baristas at shop 1

Shop 1 optimal number of baristas = 6

Shop 1 reveals a linear relationship between profits and number of baristas, with the maximum permitted number of baristas yielding the highest profit. Each barista added is still adding more value than their cost. It is likely that if an additional barista was added, profits would continue to increase. Shop 1 is likely capacity-constrained, where customers are waiting long or being lost due to slow service.

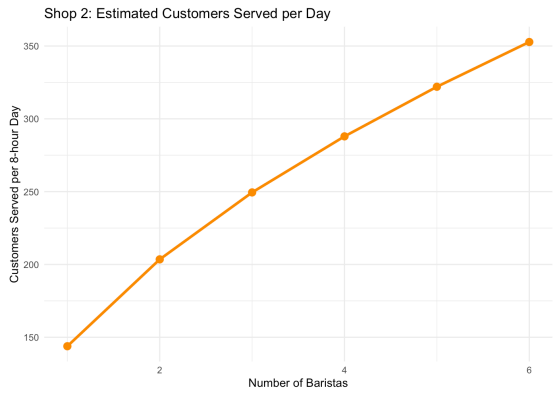

Figure 29: Customers served per number of baristas at shop 2

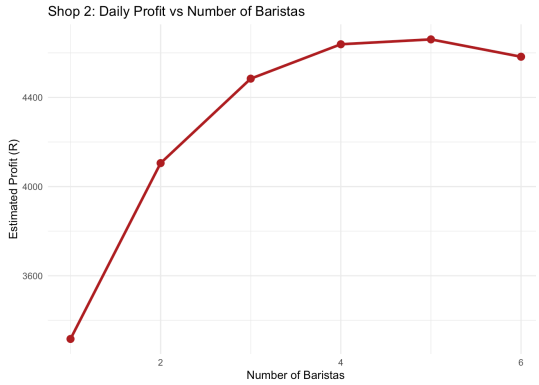

Figure 30: Daily profit per number of baristas at shop 2

Shop 2 optimal number of baristas = 5

Shop 2 shows a non-linear relationship between profit and number of baristas, where the curve flattens and then decreases after 5 baristas. This shows the diminishing returns, where beyond 5 baristas the extra labour cost outweighs the marginal improvement in service level, causing profits to begin decreasing. Shop 2 may be more efficient, or have lower customer arrival rates, so additional staff add less value.

Design of Experiments (Part 6)

Objective

The purpose of this analysis was to determine whether significant differences exist in delivery times across time periods and product types. Both ANOVA and MANOVA were conducted (using R packages), to determine the results.

Methodology

1. The DOE treats the following as factors:

   - orderYear (2022 and 2023)

   - orderMonth

   - ProductType

   And the following are treated as the responses:

   -deliveryHours

   - pickingHours

2. The hypotheses tested are:

   H_0: There is no significant difference in mean response(s) between groups.

   H_1: At least one group differs significantly

3. Analyses performed:

- One way ANOVA to test differences in deliveryHours only (1 – mean delivery times between 2022 and 2023; 2 – mean delivery times between months; 3 – mean delivery times between product types)
- MANOVA on both deliveryHours and pickingHours to test multi-response differences (between the different product types)

4. Significance level:

   $\alpha = 0.05$

Results and Interpretations

1. **Compare mean delivery times between 2022 and 2023:**

```
                Df  Sum Sq Mean Sq F value Pr(>F)
orderYear        1     138   137.6   1.376  0.241
Residuals    99998 9999650   100.0
```

Figure 31: ANOVA results experiment 1

The one-way ANOVA returned a p=0.241 < 0.05. Therefore, the null hypothesis cannot be rejected, indicating there is no significant differences between the mean delivery times in 2022 and 2023.

This suggests operational efficiency has not improved across the years – consider process improvements to decrease delivery times and improve customer satisfaction.

2. **Compare mean delivery times between months:**

```
                      Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(orderMonth)    11  172589   15690   159.6 <2e-16 ***
Residuals             99988 9827199      98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 32: ANOVA results experiment 2

The one-way ANOVA confirms a significant difference in monthly delivery times between months, confirming a seasonal variation in delivery performance. This trend can be visualized in Figure 29 below. The delivery times get larger at later months, likely indicating a performance drop throughout the year. Management needs to investigate why this is happening and implement processes and procedures to ensure delivery times remain keep improving each month, rather than getting longer.

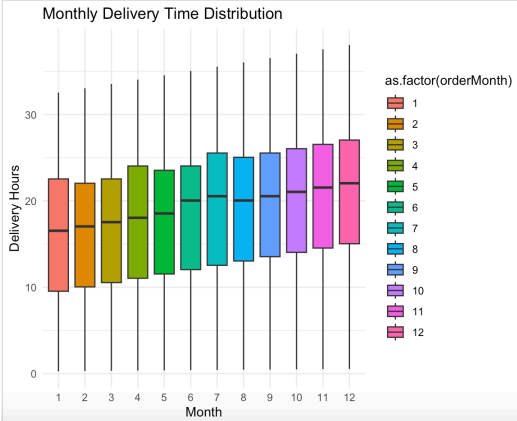

Figure 33: Mean delivery times per month

3. **Compare mean delivery times across product types:**

```
                Df  Sum Sq Mean Sq F value Pr(>F)
ProductType      5 7031007 1406201   47363 <2e-16 ***
Residuals    99994 2968781      30
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 34: ANOVA results experiment 3

The one-way ANOVA confirmed a significant difference between at least one product's mean delivery times. This is expected, as the distributions in Part 3 showed that software had a much shorter delivery time than the other products.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = deliveryHours ~ ProductType, data = sales_full)

$ProductType
                  diff          lwr         upr     p adj
KEY-CLO   0.023863154  -0.14617158   0.1938979 0.9986885
LAP-CLO   0.062521349  -0.13516218   0.2602049 0.9462939
MON-CLO   0.017504219  -0.16047896   0.1954874 0.9997682
MOU-CLO   0.070894633  -0.09380602   0.2355953 0.8239382
SOF-CLO -20.642662736 -20.80721479 -20.4781107 0.0000000
LAP-KEY   0.038658195  -0.15389277   0.2312092 0.9928287
MON-KEY  -0.006358934  -0.17862359   0.1659057 0.9999982
MOU-KEY   0.047031479  -0.11147216   0.2055351 0.9589118
SOF-KEY -20.666525889 -20.82487512 -20.5081767 0.0000000
MON-LAP  -0.045017129  -0.24462193   0.1545877 0.9877538
MOU-LAP   0.008373284  -0.17948403   0.1962306 0.9999954
SOF-LAP -20.705184084 -20.89291113 -20.5174570 0.0000000
MOU-MON   0.053390414  -0.11361140   0.2203922 0.9438416
SOF-MON -20.660166955 -20.82702222 -20.4933117 0.0000000
SOF-MOU -20.713557369 -20.86616462 -20.5609501 0.0000000
```

Figure 35: Post-hoc test for experiment 3

Results from Tukey's HSD post-hoc test revealed that the only significant different product's delivery time is indeed the Software, whereas there is not a significant difference between the other product types.

4. **Compare mean delivery times AND picking hours across product types:**

```
              Df     Pillai approx F num Df den Df Pr(>F)
orderYear      1 3.3313e-05   1.6657      2  99997 0.1891
Residuals  99998
```

Figure 36: MANOVA results experiment 4

The p-value of 0.1891 > 0.05 indicates that neither of the dependent variables (delivery hours or picking hours) differs significantly between 2022 and 2023.

Reliability of Service (Part 7)

Objective

The objective of this analysis is to evaluate reliability as a function of personnel assigned and to identify the optimal staffing level that maximises profits for a car company.

Expected reliable service (7.1)

The probability distribution (based on 397 days of observed data) is modelled as a binomial problem, where success is defined as having 15 workers or more available on a given day, and failure is defined as having less than 15 workers on duty.

Probability of a "problem day": P(#workers <15) = 0.0781

Probability of a "reliable day": P(#workers ≥ 15) = 0.9219

Therefore, the expected number of days per year where reliable service can be expected is calculated as (0.9219) *(365) = 336.4987 ≈ **336 days of reliable service per year.**

Optimising the profit (7.2)

Daily staffing presence is modelled as a binomial distribution, X ~ Binomial(R, p), where R is the number of staff appointed per day (roster size) and p is the attendance probability per employee, estimated from historical counts (12–16 present over 397 days).

For each candidate roster size, the probability of a problem day is calculated as P(X <15).

Expected monthly lost sales are calculated as R20 000 x 30 x P(X <15).

Payroll costs are calculated as R x R25 000.

Total costs are a summation of expected lost sales and payroll.

A cost curve is then plotted to visualize the results.

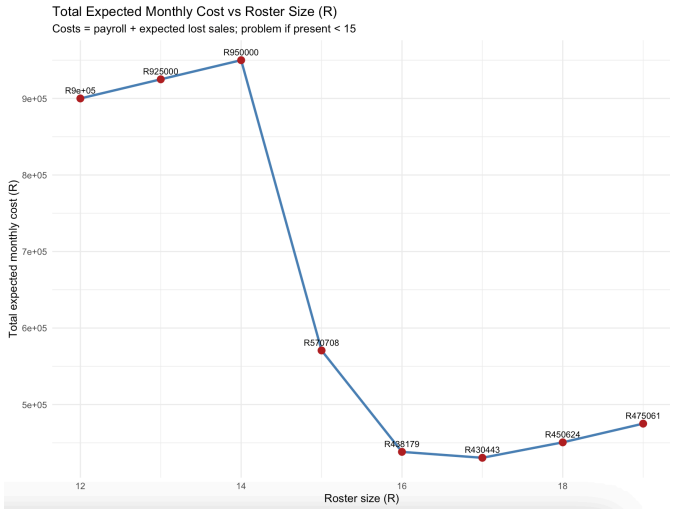

Figure 37: Total expected costs at different roster sizes

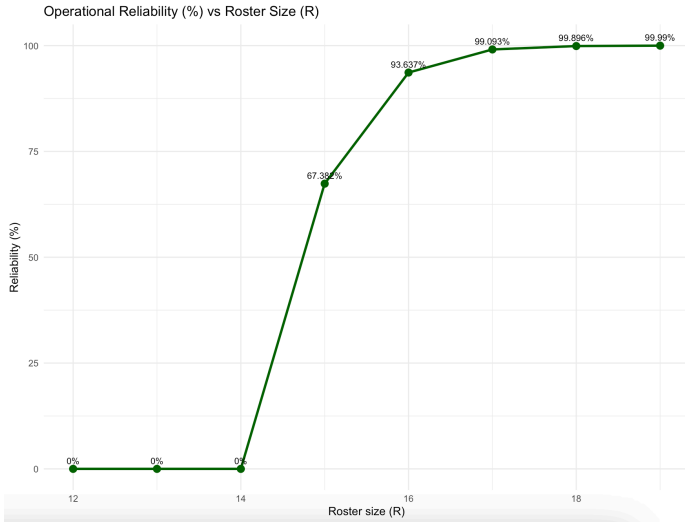

Figure 38: Reliability levels at different roster sizes

Figure 35 shows the U-shaped cost curve, where a lower R causes high lost sales costs (due to low reliability) and a higher R causes a high payroll that increases costs. The minimum total expected monthly cost is found to be at R=17. Therefore **17 workers should be assigned at any given time.**

Figure 36 shows how reliability continues to increase as more personnel are assigned. AT R=17, the **expected reliability is about 99%,** which makes it an excellent choice of the optimal number of personnel to hire.

Conclusion

The ECSA GA4 project successfully applied a range of statistical and analytical techniques to evaluate and improve processes and operational performance in a practical manner. Data validation, cleaning and analyses were performed to ensure reliable results. Statistical Process Control revealed notable process variation, and a risk analysis quantified the risks associated with making type 1 and type 2 errors. Experimental design methods were used to identify significant differences between product types to pinpoint problems more specifically and make decisions that will have a greater impact on operational performance. Profit and reliability models were built to optimise profits and enhance decision-making. The project reveals the importance of data integrity, process control and optimisation within the areas of quality management and continuous improvement.

References

Bhandari, P. (2023). *Type I & Type II Errors | Differences, Examples, Visualizations*. [online] Scribbr. Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/ [Accessed 18 Oct. 2025].

OpenAI, 2025. ChatGPT [GPT-4] \[large language model]. Available at: https://openai.com/.