ECSA Graduate Attribute Project

E Commaille

Student number: 27403785

Beng Industrial Y3

Quality Assurance 344

Stellenbosch University

24 October 2025

Contents

List of Figures and Tables

Introduction: QA ECSA project

This project is undertaken in fulfilment of the module of Quality Assurance 344, with the goal of recognising the ECSA Graduate Attribute 4. This report is accompanied by an R Markdown file named ECSA_27403785.Rmd, in which the full set of plots and supporting code can be found. This report is compiled to provide a meaningful overview of the insights to the various provided datasets and problems.

Insights are created through statistical modelling, and mathematical manipulation.

Part 1

Data is provided detailing the sales and customer base of a technology retailer, from 2022 to 2023. An initial inspection of data is done to investigate correlation between sales and other features and determine what valuable market strategies could be considered or investigated. Relationships were tested between product types, city of sales, and the customer market.

1.1 Exploration of product data

From a cursory analysis of product sales distribution, it can be seen that there is a rather uniform distribution of product sales by type and by city (Figure 1). This indicates type and city to be inconclusive in determining strategy and sales values. Likewise, sales show small fluctuation between the years 2022 and 2023 (Figure 2). Even the quantity of products sold is quite similar (Figure 3).

An initial conclusion that can be drawn here is to cast suspicion on the legitimacy of the analytic data provided by the company, as distribution of number of sales is basically uniform, between products, cities, and years.

Figure 1 Sales volume and number by Product Category

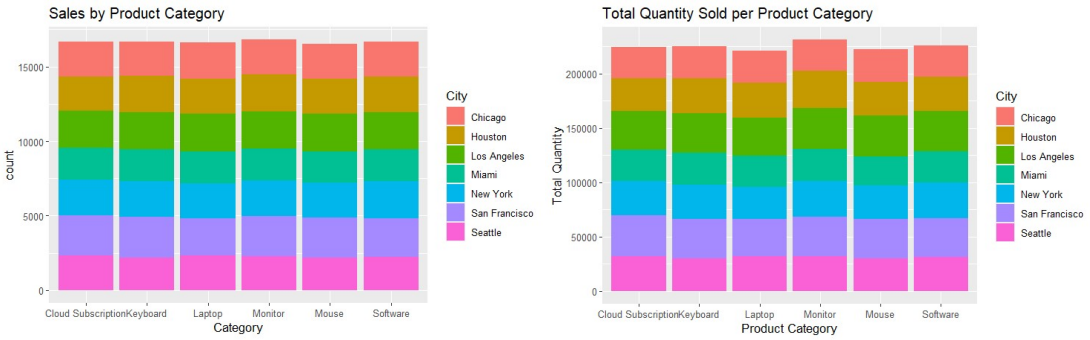


Figure 2 Year on comparison of number of Products sold

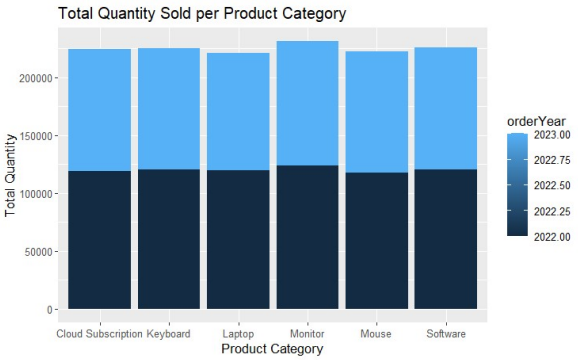
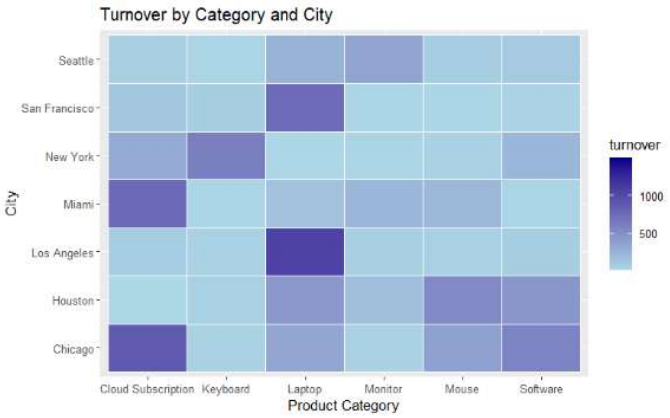


Figure 3 Heat map of Turnover by City and Category

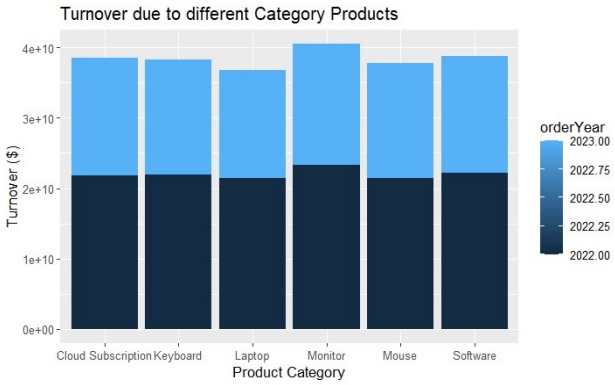


1.2 Turnover and Sales

Of great interest is which products produce the most value in sales for the company.

There is a substantial decrease in sales from 2022 to 2023 (Figure 4). However, products can be seen to occupy an almost equal share of the turnover of the company as reckoned over the past 2 years (Figure 5). Interesting insight is provided by Figure 4 to show where certain products were responsible for a large portion of the regional turnover. For example, the Laptops in LA, or Cloud services in Chicago.

Figure 4 Yearly Turnover by Category



Individual products have more fluctuating data as turnover is calculated directly in combination of quantity sold and markup. Low markup, High volume goods can bring more market value, and high value items may be slow moving and not contributing to daily profits.

The highest turnover individual products are software, a mouse, and monitors (Figure 6).

Figure 5 Turnover percentage by Category

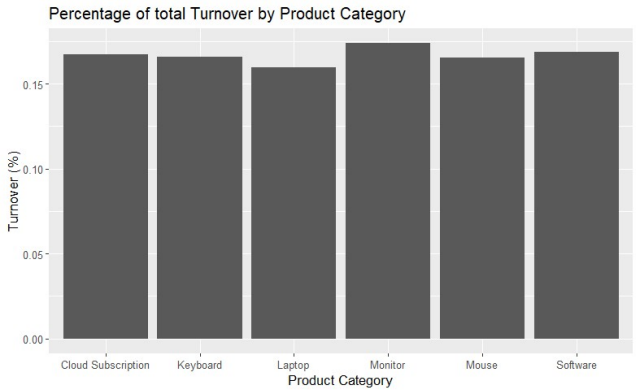
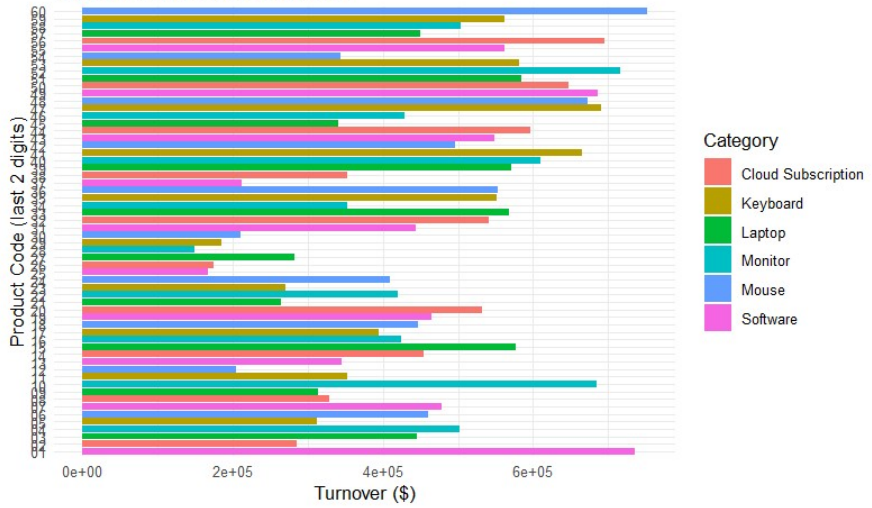


Figure 6 Individual products contribution to Turnover



1.2 Customer Profiles

Understanding the distribution and economic status of customers is valuable for retaining customers, planning expansion, and recognising gaps in market strategy.

The income distribution of customers is the same across the top turnover cities (Figure 7). The top 7 turnover cities are Chicago, Houston, Los Angeles, Miami, New York, San Francisco and Seattle.

Customer age is slightly binomially distributed, with majority of customers between the ages of 25 and 30, and 65 to 75 (Figure 8). This corresponds to the age of the most valuable customers (Figure 9) – 20-40 and 50-70. Noteworthy is the competitive turnover as a result of the 70-80 age group. These insights aid the company in considering their marketing techniques and support services, based on who they are targeting and trying to retain.

Figure 10 gives us the ability to further build customer social economic status. The highest income and average earning customers are between 36 and 65. This does not coincide with the highest value customers.

San Francisco and Los Angeles bring in the most value by a small margin (Figure 11).

Figure 7 Box plot of Income distribution by City

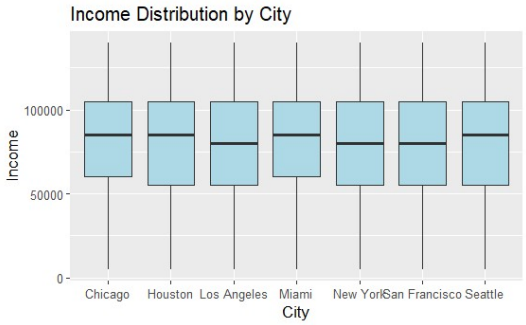


Figure 8 Distribution of Customer ages

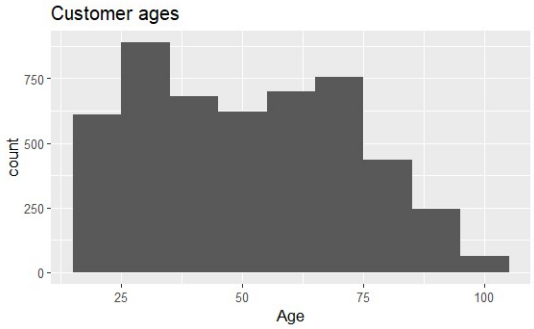


Figure 9 Customer value by age

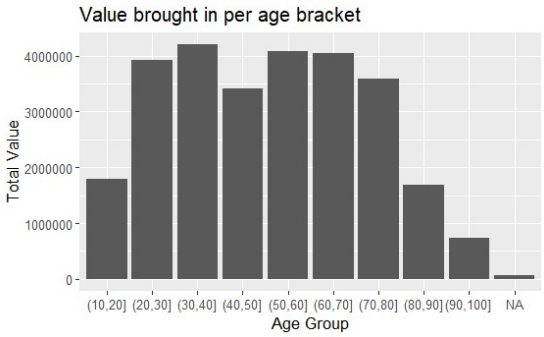
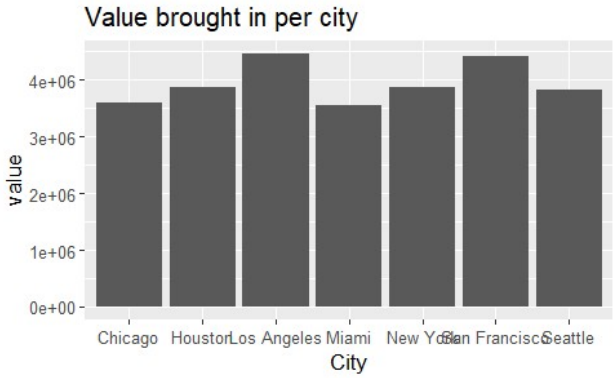


Figure 10 Customer social profile

Figure 11 City Value



Part 2 & 3: SPC

Statistical Process control is used to identify the quality and variation of a process – to detect where deviation falls within the expected boundaries or where it indicates a weakness in a process that will yield improvements if investigated. The statistical process control charts were constructed on the *sales2026and2027* data, where the first 30 samples (of 24 instances) established the expected mean and standard deviation of the delivery hours across the product types.

Process performance is seen as acceptable when the Voice of the Customer (VOC) specifications are met (these establish the accepted level of deviation on an individual instance as agreed with the customer).

Process Capability Indices indicate ability to meet the customers needs. The Cp ratio gives the spread of the process.

ChatGPT (GPT-5) was used to troubleshoot the SPC code and identify syntax errors. The implementation was verified and tested independently by the author.

In general, the samples did not fit the historical data and its bounds well. SPC charts are sampled and here discussed, you may however want to refer to the RMD project for full range of charts and graphs.

It is relevant to note that products of type “SOF” (*Software)* have an incredibly high CP and low standard deviation. This is due to the online ordering and distribution of software occurring automatically and with less manual processing. Software sales typically all take a similar amount of time, are thus densely grouped and faster to process than physical goods. Software has very positive distribution behaviour. In fact, it is outperforming so much that the company could consider reducing costs while still producing the product well-within the time.

System issues are indicated by the very Cpk values exhibited by all but the Software. Low Cpk values indicate that products in the “KEY”, “CLO”, “MOU”, “MON”, and “LAP” types are largely falling outside of the Upper Service Level. This, and the fact that the mean approaches the USL, indicates high error rate in these products. These products need to be investigated and acted on. This indicates a more fundamental issue in the processes, and can lead to suggestions like improved technology or alternate machine use.

From the Capability Summary Table (*TABLE 1),* only software is meeting the voice of the customer requirements (due to a larger than 1 Cpk). Software is currently a “better” product, while the others are bad because they are unreliable and not within Customer agreements. This is a cause for concern.

According to the Taguchi Loss, even some samples withing the agreed realm lose value to customers as they have failed to deliver on target.

Table 1 Summary of Capability Indices

ProductType <chr>	Cp <dbl>	Cpl <dbl>	Cpu <dbl>	Cpk <dbl>	Mean <dbl>	StdDev <dbl>
SOF	18.1546726	1.086642	35.2227029	1.0866423	0.957675	0.2937719
KEY	0.9169206	1.104030	0.7298115	0.7298115	19.265000	5.8165704
CLO	0.8971579	1.077375	0.7169413	0.7169413	19.214000	5.9446984
MOU	0.9151921	1.104951	0.7254328	0.7254328	19.317500	5.8275559
MON	0.8897044	1.079545	0.6998637	0.6998637	19.414000	5.9945003
LAP	0.8987584	1.100923	0.6965939	0.6965939	19.599000	5.9341123

The outcomes of finding the statistical boundaries (plus 2 sigma, etc) are beneficial because they allow us to predict a confidence interval. These values can be found I Table 2. For cases where the data can be approximated as following a normal distribution, 99,7% of all data points is contained within the Upper (+3σ) and lower level (-3σ). It is desirable to identify distributions which can be modelled as normal.

Fortunately, by the bootstrap method it can be found that the product sales can all be modelled normally, however they are statistically different from the first 30 samples which established the initial parameters. This means that future diagnosis will be inaccurate if not changed to reflect the later data. This is visible in the xbar SOF normal histogram (Figure 12). The data sits normally distributed, but not with the specs of the previous data, denoted by a red line. This shows great change in the data and requires changing parameters to continue meaningful monitoring of the process.

More than 200 samples in both CLOUD and MONITOR products xbar data shown here are outside of the limits. As can be seen in Figure 13 and Figure 14, the data changed notably. This indicates real change in behaviour in the system, and most of the xbar SPC charts behaved this way.

It is the trend across all the categories that the standard deviation became smaller with the introduction of new data, hence the more closely packed structure (Figure 15). Data that is out of bounds can be summarised and dealt with in smaller batches, as contained in Table 3.

Table 2 SPC sigma values and bounds

ProductTy... <chr>	Center <dbl>	xbar_LCL <dbl>	xbar_UCL <dbl>	Minus2Sig... <dbl>	Minus1Sig... <dbl>	Plus1Sigma <dbl>	Plus2Sigma <dbl>
SOF	0.95	0.79	1.12	0.84	0.90	1.01	1.06
KEY	19.19	15.80	22.59	16.93	18.06	20.32	21.46
CLO	19.15	15.80	22.49	16.91	18.03	20.26	21.38
MOU	19.23	15.96	22.51	17.05	18.14	20.33	21.42
MON	19.46	16.14	22.78	17.24	18.35	20.56	21.67
LAP	19.54	16.13	22.95	17.27	18.40	20.68	21.81

Figure 13 Statistical process control chart showing the distribution of CLOUD product delivery times

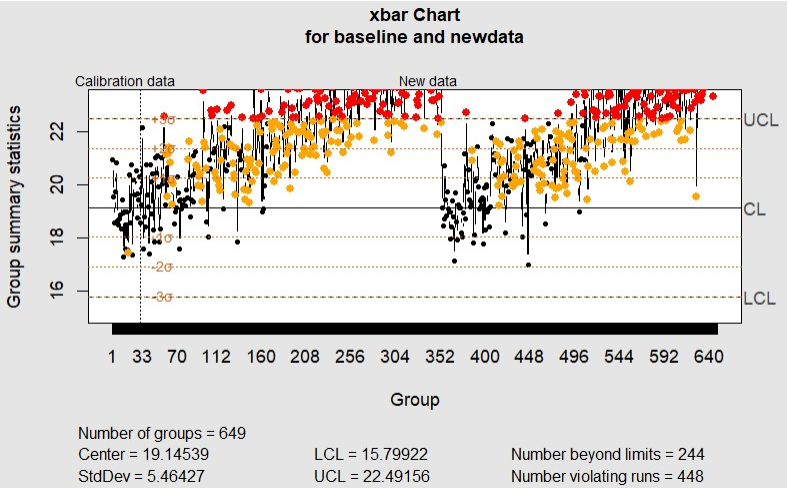


Figure 12 SOF xbar delivery hours distribution against initial normal distribution

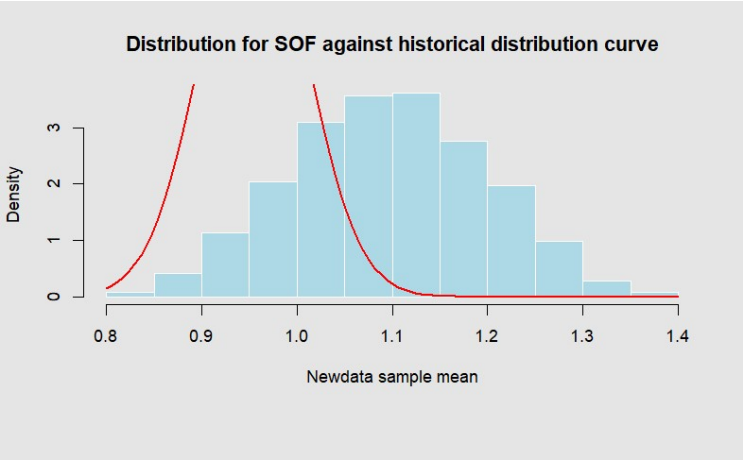


Figure 14 Statistical Process Control chart 2, showing Delivery Hour xbar across multiple samples for MONITOR products

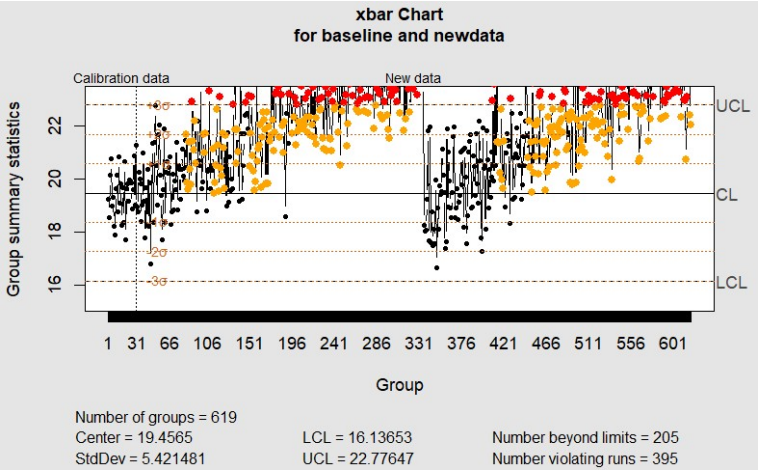


Figure 15 s distribution for a Keyboard

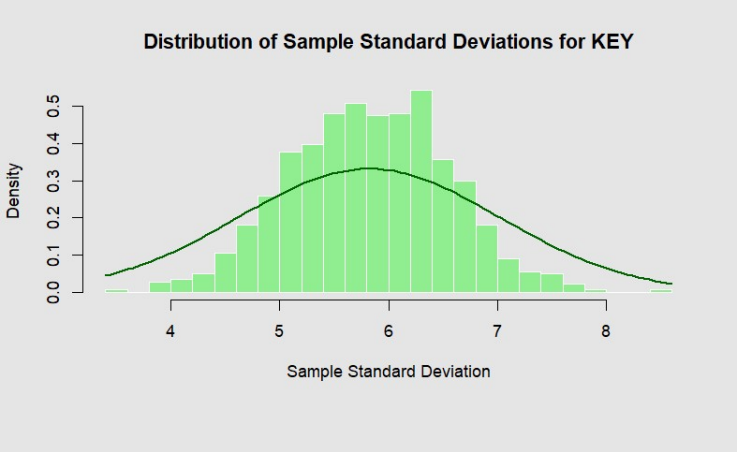


Table 3 Summary of some out of bounds samples in SPC

ProductType <chr>	SampleIndex <int>	Xbar_Value <dbl>	UCL <dbl>	LCL <dbl>	Status <chr>
SOF	306	NA	1.120	0.788	Out of Control
SOF	307	NA	1.120	0.788	Out of Control
SOF	308	NA	1.120	0.788	Out of Control
SOF	310	NA	1.120	0.788	Out of Control
SOF	311	NA	1.120	0.788	Out of Control
SOF	312	NA	1.120	0.788	Out of Control
SOF	313	NA	1.120	0.788	Out of Control
SOF	314	NA	1.120	0.788	Out of Control
SOF	316	NA	1.120	0.788	Out of Control
SOF	320	NA	1.120	0.788	Out of Control

Part 4: Risk & Data Correction

4.1 Estimating the likelihood of Type 1 errors

In the previous section, it is shown that average delivery hours across products follow a normal distribution. This means we expect a central tendency for data to be clustered around the mean, with a known amount of data (99.7 %) falling within 3 standard deviations of the mean.

A normal distribution is symmetrical, hence the expectation that half of the sample means should be above the mean, and half below.

A type 1 error is the case of a false negative, where H0 is rejected despite being true. The probability of a type 1 error is α, where (1- α) represents the confidence of our expectation that a type 1 error will not occur.

For case A:

For 6 product types, each following a normal distribution, 1 sample in each has a standard deviation outside of the upper +3 sigma-control limits. 99.7 % of data falls within 3σ of the mean. 0.15 % is expected above and below the 3σ bounds, each.

$P(\text{instance} > \text{xbar} + 3\sigma) = 0.015$ $P(\text{6 products with instance} > \text{xbar} + 3\sigma) = (0.015)^6 = 1.139062e\text{-}11$

Case B:

In a normal distribution, 68.2 % (van den Berg, 2025) of data falls within 1σ of the mean. For each product, you can calculate the probability of m standard deviations falling within 1σ consecutively. The probability of each product achieving those streaks collectively, would be the product of the individual probabilities.

Individual probability of product 1 achieving m suitable points in a row:

$P(\text{ m consecutive pts within } 1\sigma) = (0.68)^m$

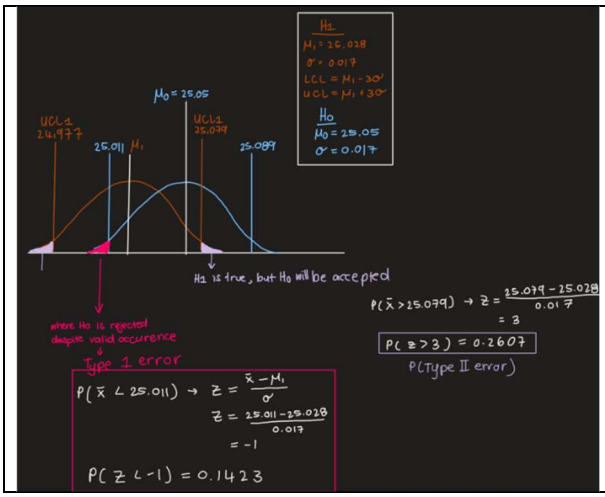
Case C:

In a normal distribution, 95.4% of data falls within 2σ of the mean. Due to the symmetricity of a normal distribution, 2.3% of data falls above (xbar + 2σ).

Let event A be the occurrence of 4 consecutive X-bar samples outside of the upper, second control limits for 1 product. $P(A) = 0.023$ $P(\text{A for 6 products}) = (0.023)^6 = 1.480359e\text{-}10$

4.2 Estimating the likelihood of Type 2 errors

In a type 2 error, a false case is accepted. The probability of not rejecting an incorrect H0 is called β.



The probability of a type 2 error – one where H0 is accepted though it should be rejected – is 26.07%.

The probability of a suitable value being misclassified and rejected due to the changed normal distribution is 14.23 % (Type 1 error).

4.3 Data Re-analysis

The corrected data yielded more clear results than the very uniform erroneous data.

Customer profile stayed the same as in the first analysis.

There is a clear higher volume of sales of Mouse and Software products (Figure 16) and a decline in sales from 2022 (Figure 17)

Figure 16 Corrected quantity sales

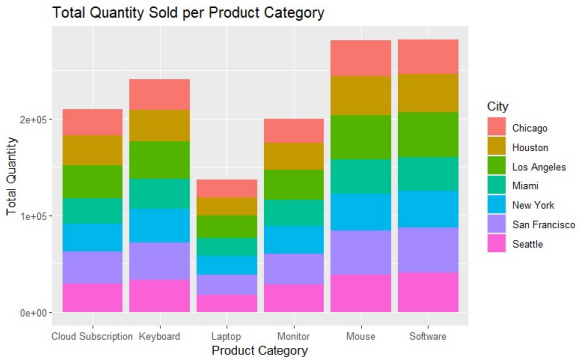
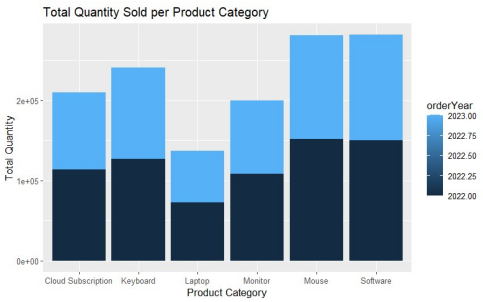


Figure 17 Corrected quantity sales by year



Turnover from specific products now valuably shows which products dominate the turnover. Figure 18 clearly shows Laptops to be the lowest turnover achieving and Keyboards and Mouse to be the highest – this is summarised in Figure 20.

Figure 18 Corrected Turnover by Product

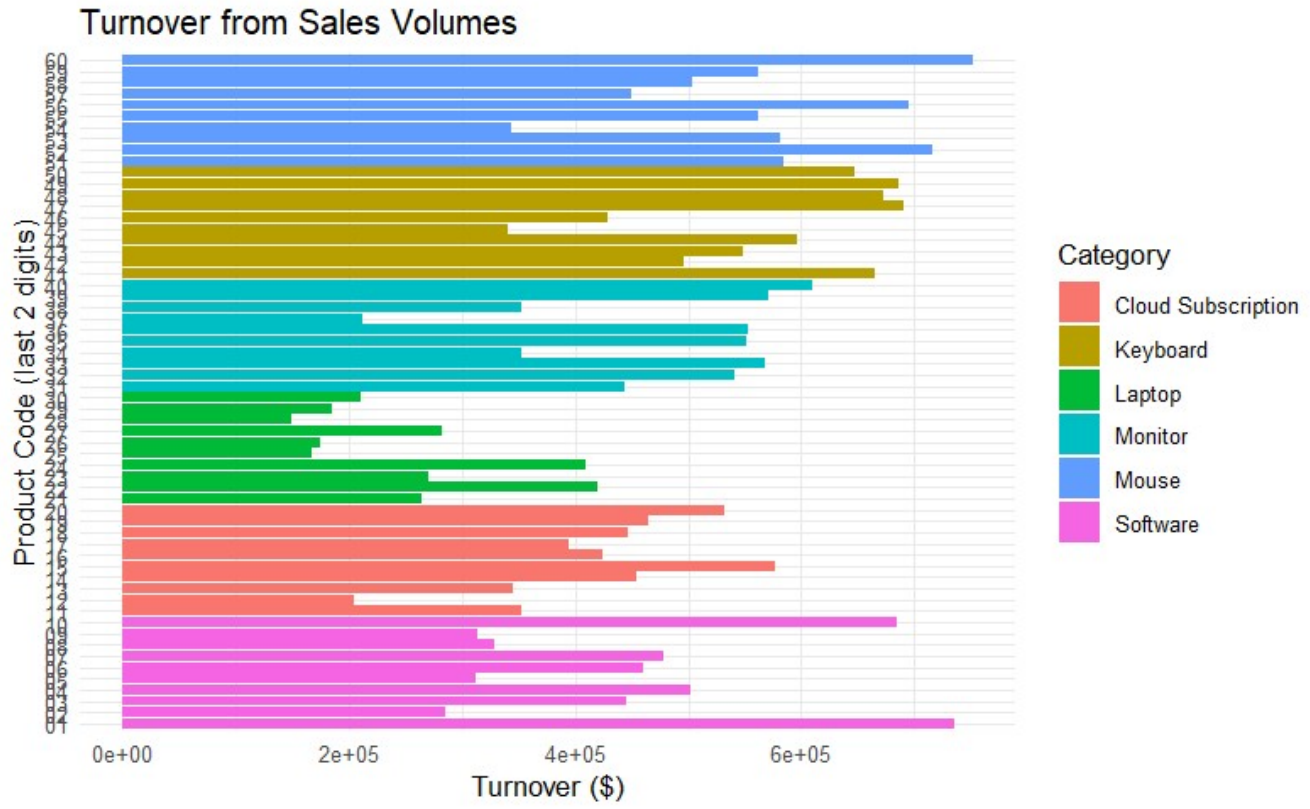
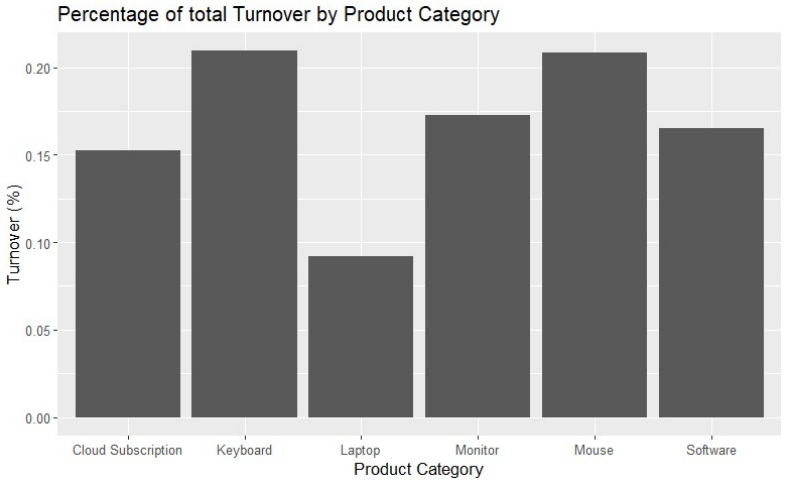


Figure 19 Corrected turnover by Category



Part 5: Optimise Profit

A model is built to show the optimal number of baristas to employ in shops 1 and 2 for optimal profit, *Figure 20*. This expects a return of R30 per customer and a cost of R1000 per barista. The assumption is made of an 8 hour work day. Additionally, reliable service is determined at service under 2 minutes.

For shop 1, a maximum profit is achieved at 6 workers, at a profit of $19 903 and a service time of 33.35 sec, *Figure 21*. Additionally, the percentage of customers who experience reliable service (under 2 minutes) for 6 workers is 100%.

For shop 2, a maximum profit is achieved at 5 workers, at a profit of $ 4661 and a service time of 89.44 sec, *Figure 23*. Additionally, the percentage of customers who experience reliable service (under 2 minutes) for 6 workers is 100%.

As expected, an increased number of baristas increases the speed of service. This improves the associated reliable service, which is service that runs for under 2 minutes. The relationship of baristas to profit is linear in shop 1 but parabolic in shop 2. Reliability increases with baristas, but this does not guarantee in a profit increase.

Figure 20 Time to Serve Shop 1

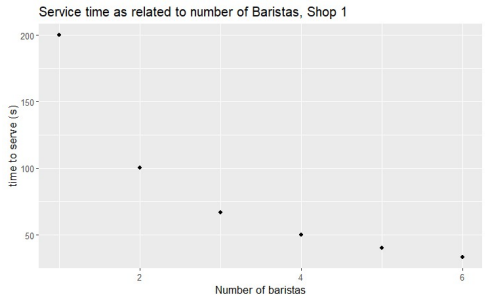


Figure 21 Profit Shop 1

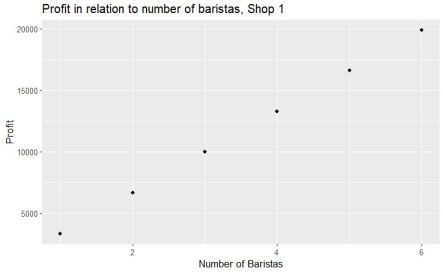


Figure 22 Time to Serve Shop 2

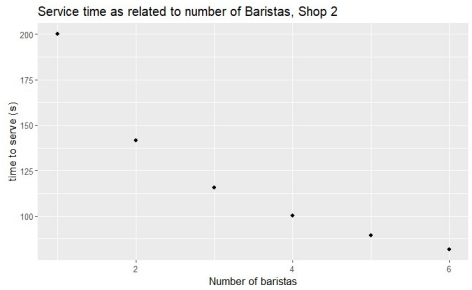
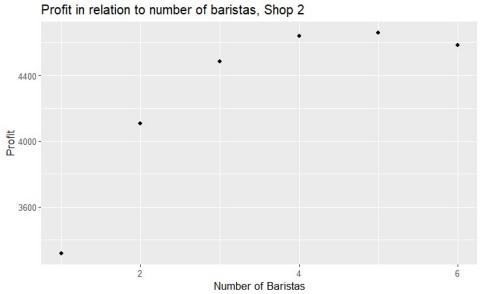


Figure 23 Profit Shop 2



Part 6: ANOVA

An ANOVA test is conducted to investigate the deliveryHours of products of the keyboard category type.

This was formulated with the principles of hypothesis testing and the **aov function** in R.

A 2 way hypothesis test is performed, to consider products of the years 2022 and 2023, respectively.

---

1. **Hypothesis:**

H0 : there is no interaction

The average means are equal from between 2022 and 2023.

H1: there is interaction

At least one mean differs.

2. **Select alpha**
   Alpha = 10%

3. **Calculate $F_{critical}$, and compute F_{test}**

   $F_{critical}$ = 1.008138

   F_{test} = 1.37644

   And p = 0.241 (if this were less than alpha, H0 would be rejected)

4. **Compare test and critical value**
   Reject H0 if **$F_{test} > F_{critical}$**

   Therefore, H0 is rejected

5. **Interpretation**
   Mean deliveryHours differ between 2022 and 2023 for Keyboards, with alpha of 10%.

Part 7: Reliability of Service

7.1 Estimation of reliable service

It is given that there are problems when there are fewer than 15 employees working on a day.

“Reliable service” is a function of the number of employees, and is thus only provided when either 15, or the maximum 16 employees are present.

Data is given over 397 days. 15 and 16 workers are present on respectively 96 and 270 days.

This means 366 out of 397 days have reliable service.

P(reliable service) = 366/397 = 0.9219

7.2 Profit optimisation

To propose an optimal solution, an activity was done to minimise cost to business.

Potential costs are employees (R25 000 per month) and R20 000 daily lost sales when staff is insufficient. Solving to minimise cost is not dependent on the business’s revenue, but purely analyses expense to company.

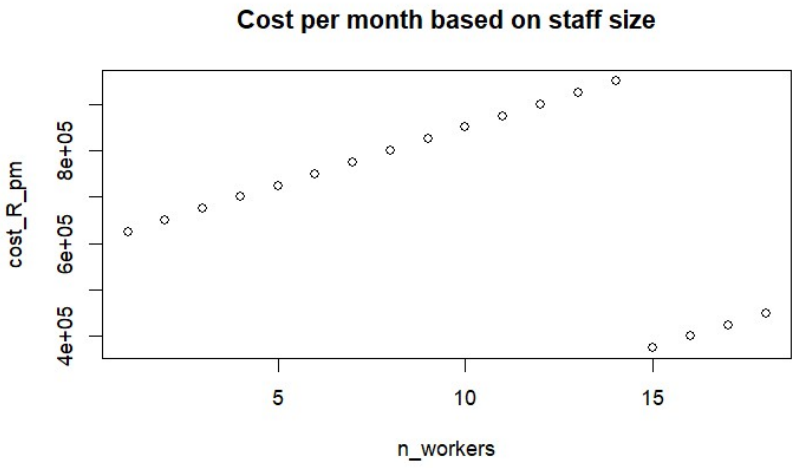
As it is a rental car agency, the assumption is made that there are 30 working days in a month, and hiring an employee adds a staff number for the whole month.

Practically, you would not hire someone to work 30 days consecutively, and you would arrange shifts. However, this solution yields the same optimal suggestion for even fewer days so it does not lose validity.

The minimal cost occurs at a total of 15 employees, and it can be seen to be about R200 000 less than the next lowest monthly cost. This large bias towards 15 as the ideal number is largely due to the large “penalties” (loss of sales) that are being accrued daily.

The cost progression can be seen in Figure 24.

Figure 24 Cost as related to number of workers



Conclusion

In depth analysis has been conducted on the product sales data of a company. This includes identifying high performing products, customer market, and correcting weaknesses in the data.

A statistical process control review was done based on the delivery hours taken for each product sale. This delivery time was further explored in an ANOVA analysis.

Optimal solutions to optimisation problems were proposed to offer practical solutions on employee number and service level.

Some analysis was conducted with the support of code taught in Quality Assurance class.

References

van den Berg, R.G. (2025) *Normal distribution – quick introduction*. SPSS Tutorials. Available at: https://www.spss-tutorials.com/normal-distribution/ (Accessed: 22 October 2025).