

DATA ANALYSIS FOR COMPANY

Quality Assurance Project 2025

Abstract

This project used R to analyze operational performance across customer, product and sales data to increase operational efficiency and ultimately increase profit across all business sectors

Kira van Peer (26870991)
26870991@sun.ac.za

Table of contents

Table of contents	1
Table of figures.....	1
Introduction:.....	2
Part 1: Descriptive Statistics:	2
1.1 Data loading and inspection	3
1.2 Summary statistics	4
1.3 Missing values	7
1.4 Data filtering /Subsetting.....	7
1.5 Data visualization	9
1.6 Exploring Relationships.....	11
Part 3: Statistical Process Control (SPC)	13
Part 4: Risk and Data Correction.....	22
4.1 Type 1 Error	22
4.2 Type II Error (β).....	23
4.3 Correcting Head-office File	24
Part 5: Performance analysis – Barista service times.....	25
Part 6 : Comparative Manova of delivery and performance analysis	29
Part 7: Reliability of service	30
7.1 Reliability Estimation	30
7.2 Profit optimization	30
Conclusion.....	32
References	32

Table of figures

Figure 1: Histogram of Age	9
Figure 2: Histogram of quantity.....	10
Figure 3: Histogram of delivery hours	10
Figure 4: Age vs income	11
Figure 5: Marking vs Selling Price	12
Figure 6:Picking vs delivery hours	12

Figure 7: Profit and reliability vs Baristas	27
Figure 8: Profit and reliability vs Baristas	28
Figure 9: Product category vs delivery hours.....	29
Figure 10: Average Profit vs Staff on Duty.....	31

Introduction:

In industrial environments data plays a critical role in achieving operational excellence, quality assurance and maximizing profit. Therefore, it is of utmost importance to analyze data in a way that is helpful to a business to achieve their operational goals. In this report it is highlighted that analyzing data is important in all industries such as manufacturing, service and retail. By collecting, cleaning and interpreting data, organizations can identify inefficiencies, monitor the performance of processes and therefore make evidence-based improvements.

This report applies statistical and optimization techniques to multiple company scenarios, each representing a different sector, previously mentioned. Using R, various analyses are performed such as descriptive statistics, process capability assessment, statistical process control, hypothesis testing using ANOVA and MANOVA and profit optimization modelling

The report focuses on three distinct companies:

- 1) A manufacturing company, where production and sales data are analyzed to understand process behavior and control.
- 2) A coffee shop where staffing levels and service times are modelled to determine optimal configurations to obtain maximum profit
- 3) A car rental agency, where workforce reliability and profit optimization are examined using binomial modelling.

Part 1: Descriptive Statistics:

Descriptive Statistics provide a foundation for understanding the characteristics, features and behavior of a dataset before conducting a more advanced analysis.

In manufacturing descriptive statistics provide insight into customer demand, product performance and delivery efficiency

1.1 Data loading and inspection

Three crucial datasets are analyzed

File: Customers_data

- 1) Dimensions: 5000x5
- 2) Structure

Table 1: sample of customer data

CustomerID <chr>	Gender <chr>	Age <dbl>	Income <dbl>	City <chr>
CUST001	Male	16	65000	New York
CUST002	Female	31	20000	Houston
CUST003	Male	29	10000	Chicago
CUST004	Male	33	30000	San Francisco
CUST005	Female	21	50000	San Francisco
CUST006	Male	32	80000	Miami

- 3) Column names :
“CustomerID”; “Gender”; “Age”; “Income”; “City”

File: Products_data

- 1) Dimensions: 60x5
- 2) Structure

Table 2: Sample of products data

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
SOF001	Software	coral matt	511.53	25.05
SOF002	Cloud Subscription	cyan silk	505.26	10.43
SOF003	Laptop	burlywood marble	493.69	16.18
SOF004	Monitor	blue silk	542.56	17.19
SOF005	Keyboard	aliceblue wood	516.15	11.01
SOF006	Mouse	black silk	478.93	16.99

- 3) Column names
“ProductID”; “Category”; “Description”; “SellingPrice”; “Markup”

File : Sales2022and2023

- 1) 100000x9

Table 3: Sample of sales data

2) Structure

CustomerID <chr>	ProductID <chr>	Quantity <dbl>	orderTime <dbl>	orderDay <dbl>	orderMonth <dbl>	orderYear <dbl>	pickingHours <dbl>	deliveryHours <dbl>
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544
CUST3172	LAP026	17	17	14	7	2023	38.39083	31.546
CUST1022	KEY046	11	16	23	5	2022	14.72167	21.544
CUST3721	LAP024	31	12	18	7	2023	41.39083	24.546
CUST4605	CLO012	20	14	7	2	2022	15.72167	24.044
CUST2766	MON035	32	21	24	12	2022	21.05500	24.044

3) Column names

“Customer ID”; “ProductID”; “Quantity”; “orderTime”; “orderDay”; “orderMonth”; “orderYear”; “pickingHours”; “deliveryHours”

By inspecting data files like this, obvious mistakes can be spotted and ideas can be formed of how the data should be analyzed and read. Here it can be seen that both the customers data and products data can be matched to the sales data file through customerID and productID respectively. This can contribute to the comparison of the two data sets. It is also helpful to note that there are differences in the dimensions of the datasets which could be problematic when matching them.

1.2 Summary statistics

Customers

Table 4: Customer summary statistics

CustomerID	Gender	Age	Income	City
Length: 5000	Length:5000	Min: 16.00	Min: 5000	Length: 5000
Mode:Character	Mode:Character	1 st Q : 33.00	1 st Q :55000	Class: character
		Median: 51.00	Median: 85000	Mode:character
		Mean: 51.00	Mean: 80797	
		3 rd Q: 68.00	3 rd Q: 105000	
		Max: 105.00	Max: 140000	

This shows the demographic information of the customer base: The customer dataset contains 5000 entries each representing an individual buyer. The ages range from 16 to 105 years with a mean and median of 51 years, indicating that the customer base is generally middle aged with a roughly symmetrical age distribution when looking at the

1st and 3rd quartile. This suggests that both younger and older adults are represented but the majority of the customers are likely between the ages of 30 and late 60s.

Customer income values range from R 5000 to R140000, with a mean income of around R80000 and a median of R85000. The slightly higher median relative to the mean suggests a mild right skew, meaning lower income customers bring the average down. Overall, the income distribution indicates mostly middle-to upper class as the 1st and 3rd quartiles suggest.

The Gender and city columns are categorical which shows that the data was collected from multiple regions and across male and female which makes the data unbiased. These demographic variables can later be used to identify whether location or gender influences purchasing behavior or delivery performance.

Products

Table 5: Products summary statistics

ProductID	Category	Description	SellingPrice	Markup
Length:60	Length:60	Length:60	Min: 350.4	Min: 10.13
Mode:character	Mode:character	Mode:character	1 st Q :512.2	1 st Q :16.14
			Median: 794.2	Median: 20.34
			Mean: 4493.6	Mean: 20.46
			3 rd Q: 6416.7	3 rd Q: 25.71
			Max: 19725.2	Max: 29.84

The product dataset consists of 60 unique items across various categories such as software, laptops and peripherals. The selling price ranges from R350.40 to R19725.20 with a mean price of R4493.30. The large gap between minimum and maximum prices, along with the high mean value relative to the median indicates a right – skewed distribution. This means that there are a few products that are high-end and significantly increase the average.

The markup percentage varies from 10.13% to 29.84%, with an average of about 20.46%. The quartile range (16.14% to 25.71%) shows that most products are priced within a consistent profit margin, suggesting a controlled pricing policy across categories.

Together, the selling price and markup data suggests that the company has a diverse mix of products, meaning more affordable high volume products to higher end products with higher margins. This means the company applies to all markets

Sales2022and2023

Table 6: Sales statistics

CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours
Length:10000	Length: 10000	Min: 1.0	Min: 1.0	Min: 1.0	Min: 1.0	Min: 2022	Min: 0.4259	Min: 0.2772
Mode:character	Class: character	1 st Q :3.0	1 st Q :9.0	1 st Q :8.0	1 st Q :4.0	1 st Q :2022	1 st Q :9.3908	1 st Q :11.5460
		Median: 6.0	Median: 13.0	Median: 15.0	Median: 6.0	Median: 2022	Median: 14.0550	Median: 19.5460
		Mean: 13.5	Mean: 12.93	Mean: 15.5	Mean: 6.448	Mean: 2022	Mean: 14.6955	Mean: 17.4765
		3 rd Q: 23.0	3 rd Q: 17.00	3 rd Q: 23.00	3 rd Q: 9.00	3 rd Q: 2022	3 rd Q: 18.7217	3 rd Q: 25.0440
		Max: 50.0	Max: 23.00	Max: 30.00	Max: 12.00	Max: 2022	Max: 45.0575	Max: 38.0460

The sales dataset contains 10000 transaction records in the years 2022 and 2023, with information on customer purchases, product ID's, quantities, order timing, and processing durations.

The quantity sold per transaction ranges from 1 to 23 units, with a median of 6 units and a mean of 13.5 units. The mean is higher than the median so there are a few bulk or “larger” orders but most of the orders are small.

The order time and order day values indicate that the purchases are distributed evenly throughout each month, while the orderMonth data (1-12) confirms that sales are active throughout the year. The order year is always 2022 to 2023.

For the process variables, picking hours range from 0.43 to 18.72 hours, and delivery hours range from 0.28 to 25.04 hours, with means of approximately 14.70 hours and 17.48 hours respectively. These measures suggest that, on typical orders picking times and delivery times are within one day of ordering but can be longer for more complex or delayed deliveries.

Summary statistics provide a concise overview of the main characteristics of the dataset and help the company understand and interpret data before analyzing the data in depth. Issues with the quality of the data can be identified, like missing data, outliers or unusual data entries. The central tendency can also be understood because values

like the minimum, maximum, mean and standard can help identify where most of the values lie and how much variation there is in the data. It can also be easier to identify and compare patterns between variables as well as datasets.

1.3 Missing values

There are no missing values in the dataset which means the data is accurate and worth analyzing. The results will be accurate and can help the business make conclusions that are correct.

1.4 Data filtering /Subsetting

By subsetting and filtering data you can select specific rows and columns of the dataset that meet certain constraints which are relevant to your analysis

Customers (Gender vs Income)

Table 7: Customers (Gender vs Income)

Gender <chr>	Income <dbl>
Female	80816.20
Male	80770.21
Other	80871.56

The average income across gender categories is very similar, with all values around R80000. This means balanced income distribution among all genders in the customer base, and this means gender does not influence purchasing power. It can also be seen that the income is relatively high as previously confirmed.

Products (Filtered Counts)

The number of instances that meet the following conditions

- 1) high income customers: customers with an income higher than 60 000:
3490
 - 2) high markup products: products with a markup of above 25%
18 products
 - 3) sales 2023: Sales transactions from the year 2023
46273
- Of the 5000 customers, approximately 70% earn above R60000, which once again highlights that the company's high income customer base. Only 18 of the

60 products have a markup of higher than 25% which means that higher end products are few and far between in the product list.

The majority of the sale in sales2023and2023 occurred in 2023 which indicates that the company grew in 2023.

Categories of products and selling prices

Table 8: Categories of products and selling prices

Category <chr>	SellingPrice <dbl>
Cloud Subscription	3691.861
Keyboard	4638.172
Laptop	5217.545
Monitor	5014.170
Mouse	4585.465
Software	3814.344

Among the product categories, laptops and monitors have the highest average selling prices while software and cloud subscriptions are the least expensive. Once again, the mix of selling prices ensures that the company is part of a wide range of markets and appeal to a wide variety of customers with different customer incomes.

Top 5 customers by Quantity purchased

Table 9: Top five customers by quantity purchased

CustomerID <chr>	TotalQuantity <dbl>
CUST596	7214
CUST1791	7165
CUST1193	6917
CUST3721	6394
CUST2277	6381

Here we can see that there are 5 customers that account for a large portion of the total sales volumes. These are most likely bulk buyers or long term clients. This indicates the power of strong customer relationships.

Average of delivery and picking

Table 10: Average delivery and picking

AvgPicking	AvgDelivery

<dbl>	<dbl>
14.69547	17.47646

On average products take about 14,7 hours to be picked and 17, 5 hours to be delivered. This suggests that the delivery process adds roughly 3 hours beyond picking likely due to logistics or transportation. This process is generally efficient but may still benefit from optimizing the delivery stage to reduce the total turnaround time

1.5 Data visualization

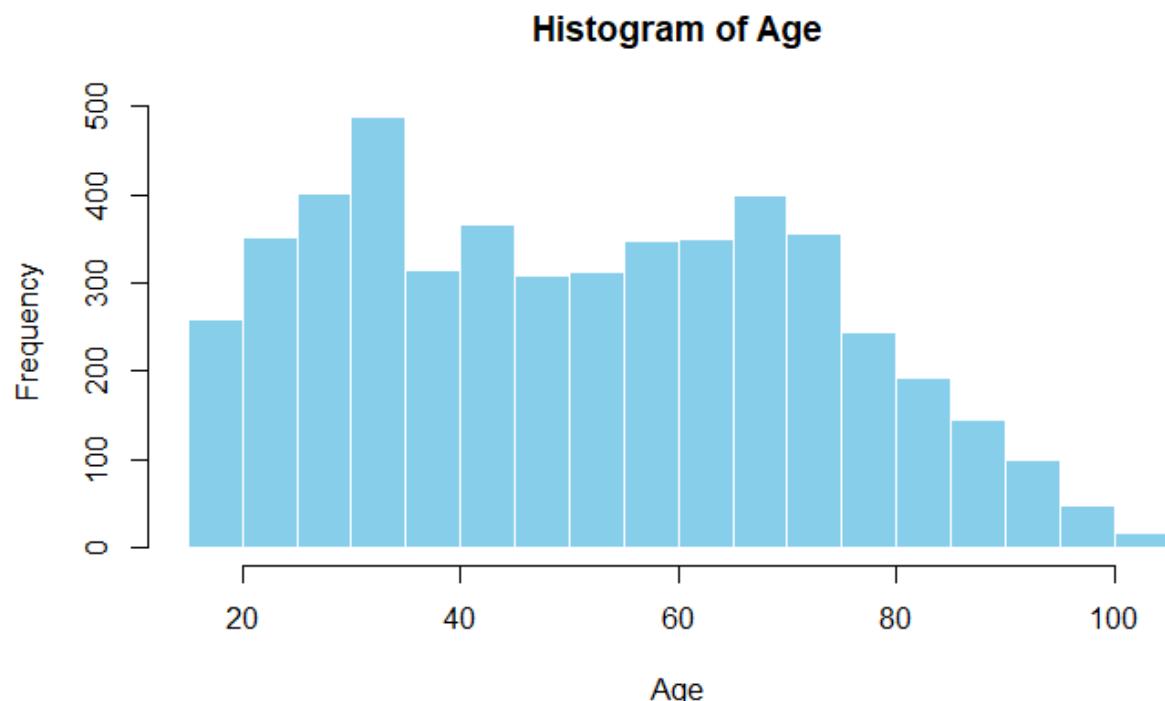


Figure 1: Histogram of Age

The histogram of customer ages shows a broad distribution from 16 to 100 years with most customers between 30 and 70. A noticeable peak occurs around age 30, which could be due to young professionals investing in equipment. This shows a strong appeal for that group but still holds appeal across older age groups

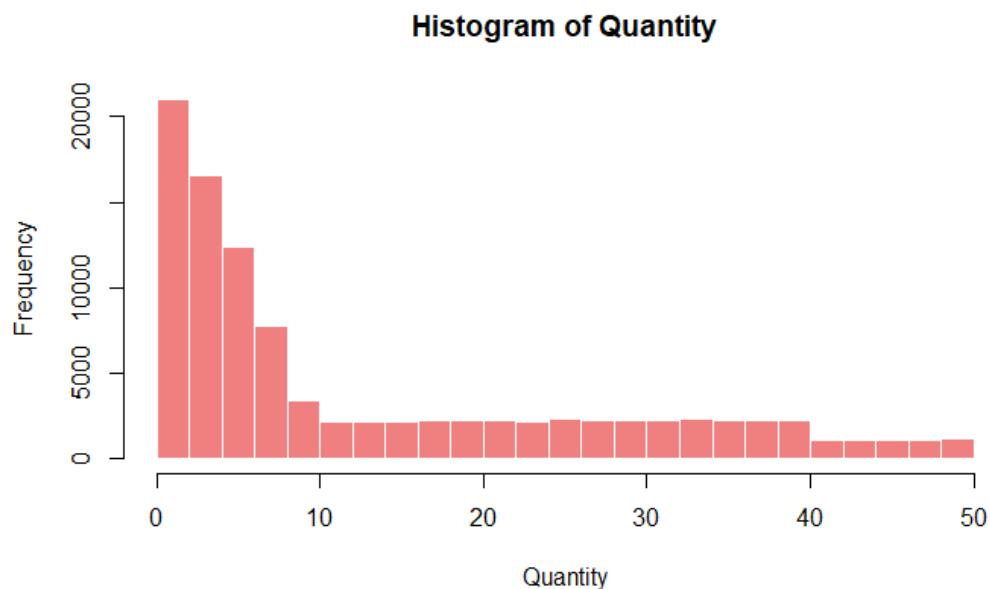


Figure 2: Histogram of quantity

This histogram is heavily right-skewed which further indicates that most orders are small orders . Only a few sales have large quantities.

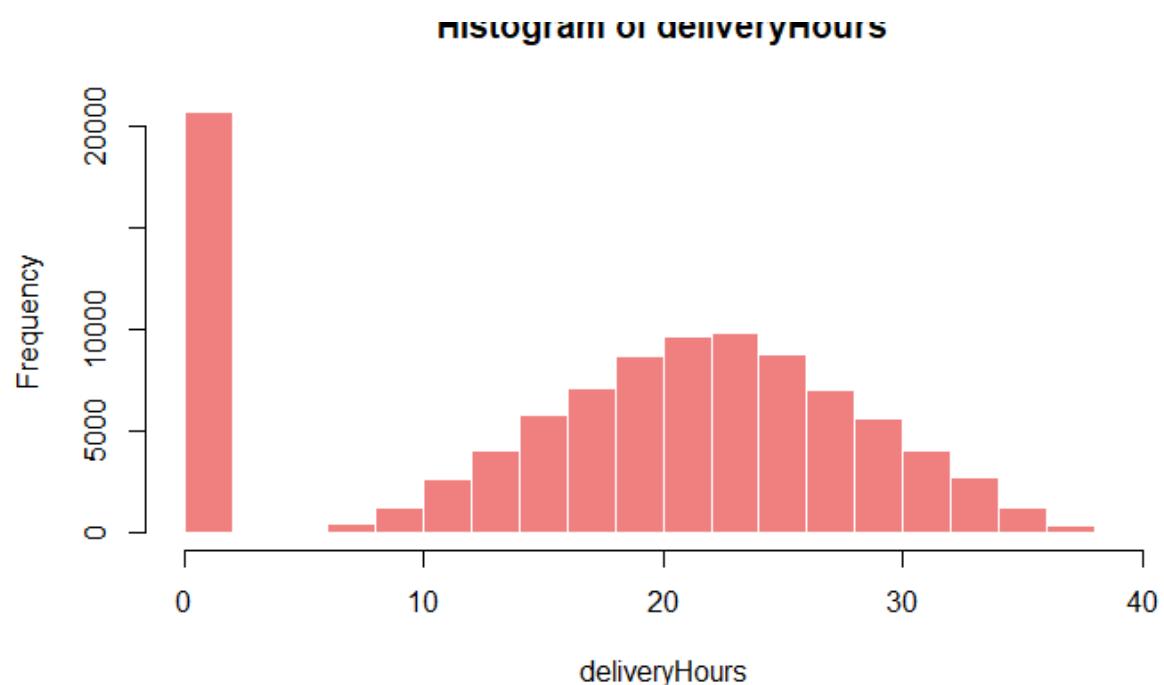


Figure 3: Histogram of delivery hours

The histogram shows a peak near zero and this means that most orders were delivered immediately (bought in store) . Beyond that , the rest of the data forms a roughly bell – shaped curve distribution with a center at around twenty hours. This means most deliveries take a day to complete

A few similar data visualizations were done but only the above mentioned were worth mentioning and the other visualizations were trivial since the data was discussed in detail previously.

1.6 Exploring Relationships

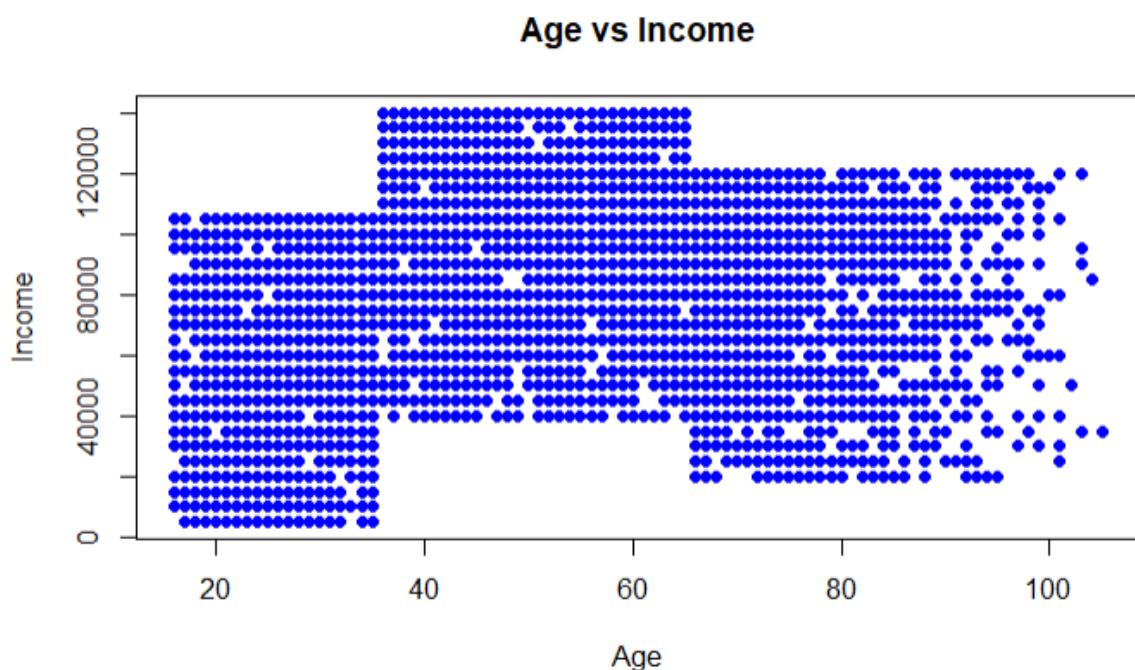


Figure 4: Age vs income

This relationship is exactly to be expected. Because income typically increases when a person is earning money. A person works from about 30 to 60 which is where the data points are at the highest income, lower ages earn less money since children cannot earn money and then the income values drops as soon as people reach retirement.

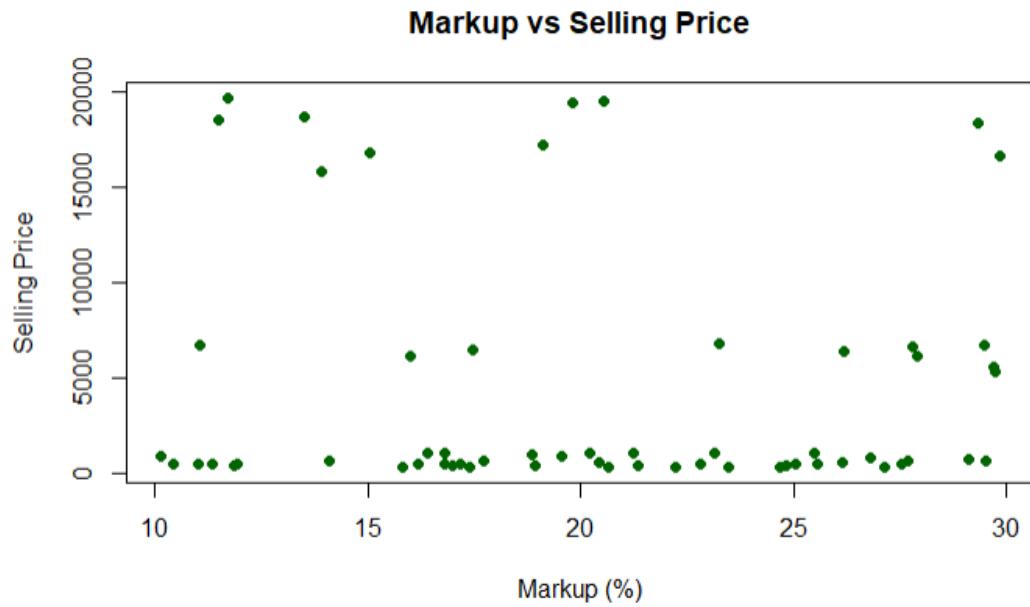


Figure 5: Markup vs Selling Price

There is no clear relationship here. The selling prices are widely spread across all markup values, suggesting that higher markups do not necessarily correspond to higher selling prices. This shows that product pricing may depend more on product type or category rather than just markup percentage.



Figure 6:Picking vs delivery hours

So far using the descriptive statistics the following recommendations can be made

- 1) Targeted Marketing: Customer income and age distribution clearly shows the company's main customer base. The marketing strategies can be targeted towards high-income and mid-age groups. This will increase engagement and ultimately increase sales. Generational marketing includes dividing audiences up according to their age or generation like Baby Boomers, Generation X, Millennials and Generation Z and it is helpful to create strategies for marketing that align with the characteristics of each group. (Porch Group Media)
- 2) Operational Efficiency: The picking vs delivery Hours plot suggests that there are some inefficiencies or clustering of operations at certain times. The company should analyze workflow timing to balance workloads and reduce delays in delivery.
- 3) Customer Retention: With clearer patterns of who the main buyers are, loyalty programs or personalized offers can be developed to maintain repeat purchases among the most profitable customer segments.

Part 3: Statistical Process Control (SPC)

Product: SOF

Control Limits for Product SOF

X-bar Chart: UCL = 1.14 CL = 0.96 LCL = 0.77

S Chart: UCL = 0.43 CL = 0.3 LCL = 0.17

Out-of-control X-bar samples: 115 134 136 163 166 181 193 203 209 210 223 230 237
239 244 245 246 254 256 261 262 263 266 267 268 269 270 272 274 276 278 281 283
284 285 286 287 288 292 296 298 300 302 304 305 306 307 310 312 315 317 320 321
323 325 326 327 328 329 331 334 335 336 337 341 343 345 348 350 351 353 354 355
356 357 358 359 360 361 362 363 366 367 369 371 372 373 375 376 377 378 379 380
382 383 386 387 389 390 394 395 396 397 398 399 402 403 404 406 407 408 410 411
412 413 414 415 416 417 418 419 420 421 422 423 425 428 429 430 431 432 433 435

436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455
456 457 458 459 460 461 462 463 464 469 550 575 576 582 589 618 622 626 627 637
641 642 644 651 652 653 656 664 669 670 674 677 680 700 705 706 707 710 717 719
722 725 729 730 732 733 739 740 741 743 744 745 746 747 750 753 755 756 757 758
760 761 762 765 766 767 768 771 772 774 775 776 777 778 780 781 782 783 784 785
786 787 788 789 791 792 793 794 796 797 799 800 801 803 804 806 807 808 809 810
811 812 814 815 816 817 819 820 821 822 823 824 825 826 827 829 830 833 834 836
837 838 839 840 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 860
861 862 863 864 865

Out-of-control S samples:

Process Capability Indices for SOF

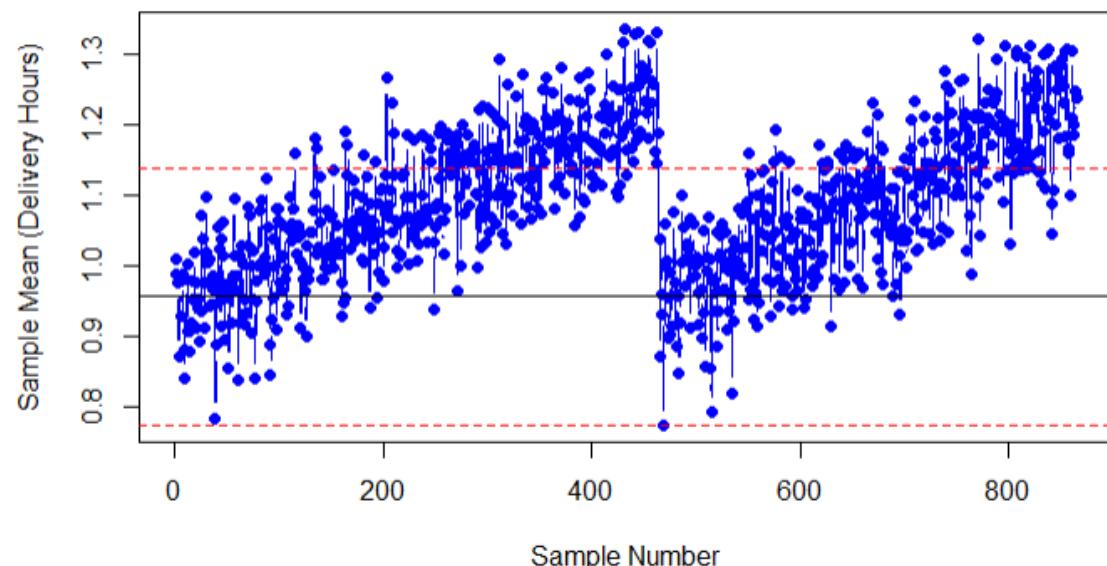
$C_p = 18.135$

$C_{pk} = 1.083$

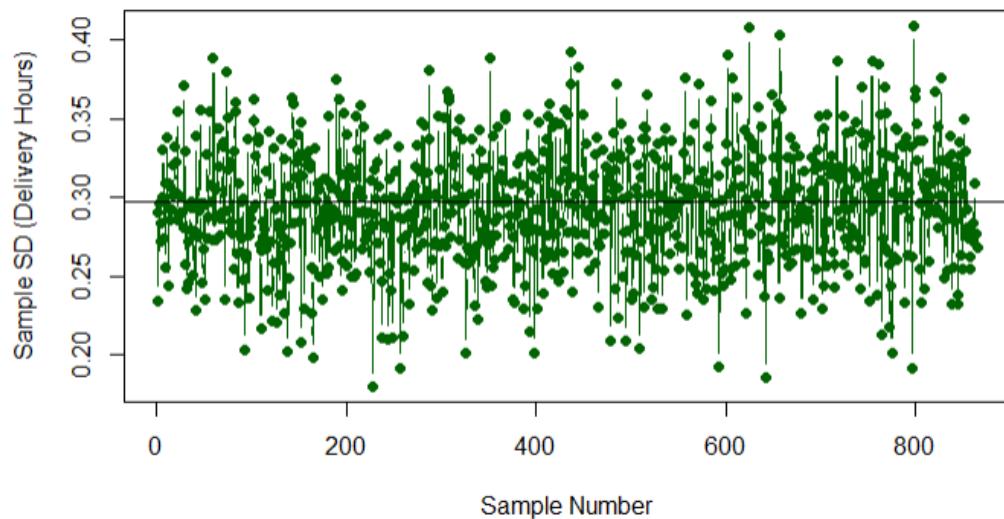
$C_{pu} = 35.188$

$C_{pl} = 1.083$

X-bar Chart for Product SOF



S Chart for Product SOF



Product LAP

Control Limits for Product LAP

X-bar Chart: UCL = 23.13 CL = 19.52 LCL = 15.92

S Chart: UCL = 8.5 CL = 5.89 LCL = 3.28

Out-of-control X-bar samples: 102 114 116 117 122 130 132 133 136 137 140 145 148
154 155 156 158 159 160 162 164 165 171 173 174 175 177 179 181 184 186 188 193
194 197 199 200 201 203 206 208 212 214 215 216 217 218 220 221 222 223 224 226
227 318 331 337 339 345 349 351 354 356 359 367 368 369 372 374 375 376 378 379
380 381 382 384 385 386 387 388 389 390 393 395 396 397 399 400 403 404 406 407
408 409 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426

Out-of-control S samples: 329

Process Capability Indices for LAP

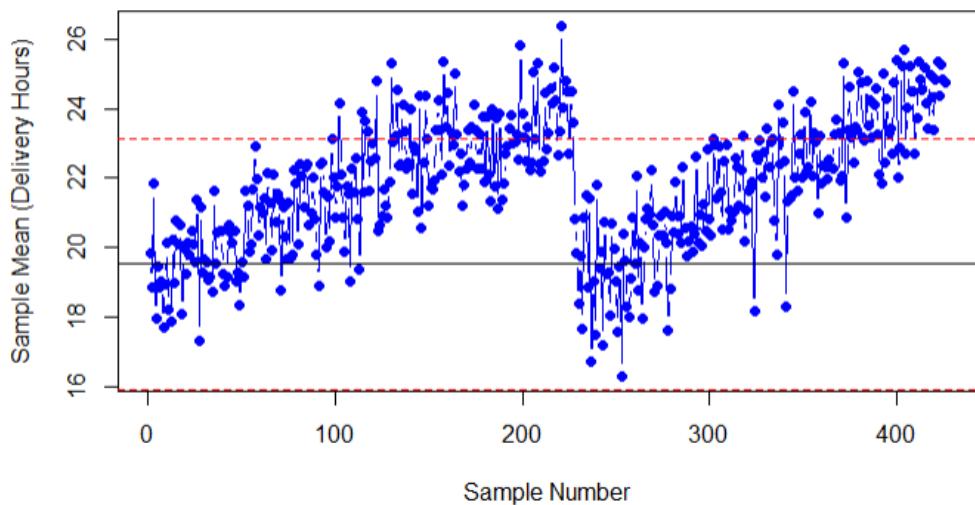
C_p = 0.899

C_{pk} = 0.696

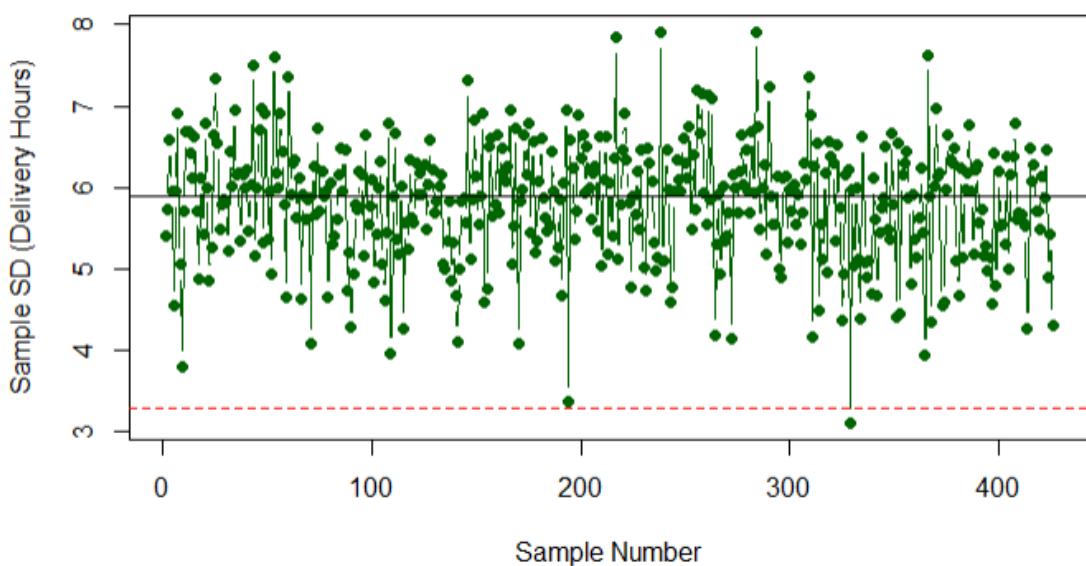
C_{pu} = 0.696

Cpl = 1.101

X-bar Chart for Product LAP



S Chart for Product LAP



Product code: KEY

Control Limits for Product KEY

X-bar Chart: UCL = 22.78 CL = 19.19 LCL = 15.61

S Chart: UCL = 8.45 CL = 5.86 LCL = 3.27

Out-of-control X-bar samples: 62 102 117 145 160 163 172 173 174 178 180 188 189
190 194 196 197 200 202 203 206 210 212 213 216 218 219 220 222 224 227 229 232
233 240 241 242 248 250 252 254 256 262 263 264 265 268 269 271 272 275 277 278
279 281 283 284 287 288 293 294 295 297 298 303 306 308 309 310 311 312 314 315
316 317 319 320 321 322 323 324 325 326 330 331 333 335 337 338 339 340 341 342
343 344 345 346 347 348 349 351 352 353 354 355 356 357 358 360 361 362 363 364
365 366 367 368 369 370 372 373 374 375 376 377 378 379 380 381 382 383 384 385
386 387 388 389 390 391 392 393 394 395 396 398 458 498 516 547 558 578 579 581
596 598 601 606 608 609 610 614 616 617 620 626 628 629 631 632 633 635 638 639
640 641 642 643 644 646 649 650 651 653 654 655 657 659 660 661 662 663 664 666
668 669 670 671 672 673 676 677 680 681 682 684 685 687 689 691 692 694 695 696
698 699 700 701 702 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718
721 722 723 724 726 727 728 729 730 731 732 734 735 736 738 739 740 741 742 743
746 747

Out-of-control S samples:

Process Capability Indices for KEY

Cp = 0.917

Cpk = 0.729

Cpu = 0.729

Cpl = 1.105

Product code: MON

Control Limits for Product MON

X-bar Chart: UCL = 23.05 CL = 19.43 LCL = 15.8

S Chart: UCL = 8.54 CL = 5.92 LCL = 3.3

Out-of-control X-bar samples: 99 108 116 135 141 146 150 153 158 169 174 179 181
182 185 191 193 197 200 205 210 211 213 215 218 223 224 225 226 228 234 235 241
242 243 246 248 249 250 251 252 253 256 258 259 260 262 263 267 268 269 270 271

273 274 275 277 278 280 281 282 284 286 287 288 291 292 293 296 299 301 304 305
306 308 309 311 313 315 318 319 320 322 324 325 326 327 328 329 330 331 332 333
334 335 411 424 444 451 456 460 484 487 495 496 499 505 510 511 513 518 523 524
529 537 538 540 541 542 545 547 554 555 556 557 558 559 560 562 568 570 573 574
575 577 578 580 581 582 583 584 585 586 589 590 593 594 595 596 597 598 599 600
601 602 604 605 607 608 610 611 615 616 617 620

Out-of-control S samples:

Process Capability Indices for MON:

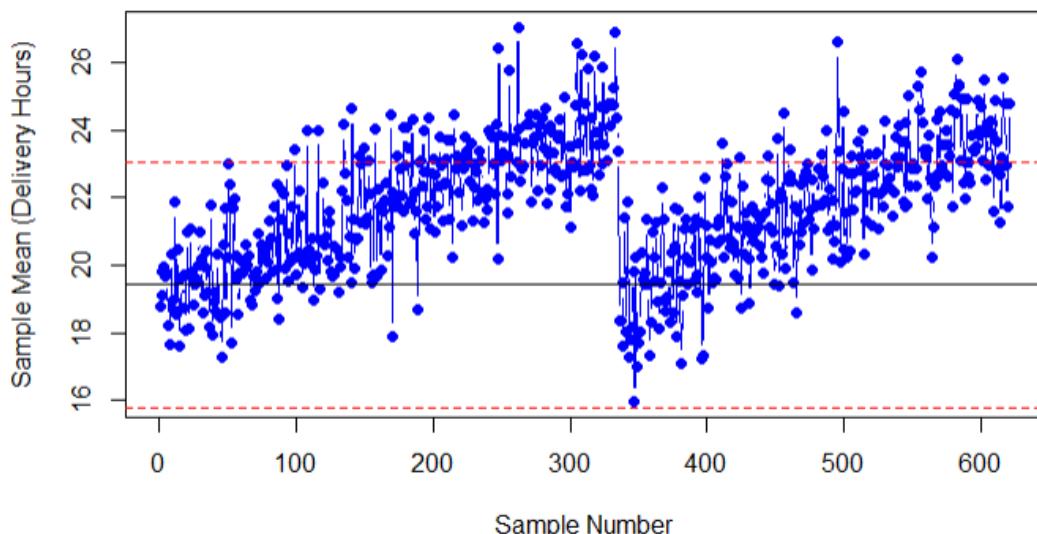
$C_p = 0.889$

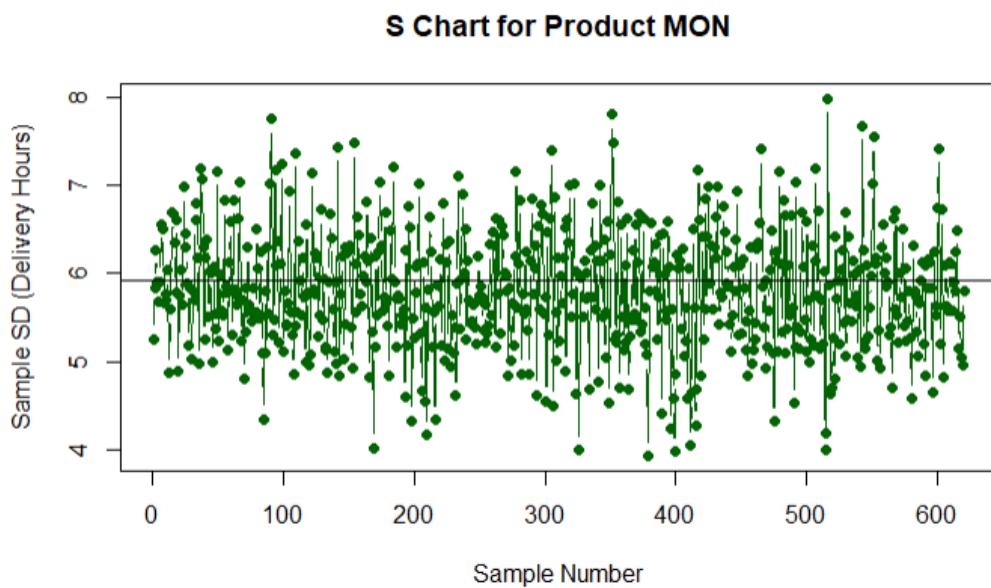
$C_{pk} = 0.7$

$C_{pu} = 0.7$

$C_{pl} = 1.079$

X-bar Chart for Product MON





Product Code : CLO

Control Limits for Product CLO

X-bar Chart: UCL = 22.74 CL = 19.13 LCL = 15.51

S Chart: UCL = 8.52 CL = 5.91 LCL = 3.29

Out-of-control X-bar samples: 107 112 144 153 157 168 171 177 179 181 183 185 189
 194 198 200 202 203 204 209 211 212 216 217 219 221 223 224 225 229 232 236 237
 239 240 243 244 250 251 252 256 257 258 259 260 261 262 264 266 267 268 270 271
 272 273 274 275 276 277 278 279 280 281 282 284 285 286 288 289 290 291 292 293
 295 296 298 299 301 302 303 305 306 307 308 309 310 311 312 314 316 317 318 319
 320 321 322 323 324 325 326 329 330 331 332 333 334 335 336 337 338 339 340 341
 342 343 344 345 346 347 348 350 351 352 353 482 493 498 506 508 512 513 515 516
 519 522 523 524 525 526 530 531 533 535 538 540 542 543 545 547 554 555 559 561
 562 563 565 567 569 570 571 572 575 577 578 579 581 583 584 585 586 588 590 591
 592 593 594 595 596 597 599 600 601 602 604 605 606 607 608 610 611 612 613 614
 616 617 618 619 620 621 622 623 624 625 628 629 631 632 634 635 636 637 638 639
 640 641 642 643 644 646 647 648 649 650

Out-of-control S samples:

Process Capability Indices for CLO

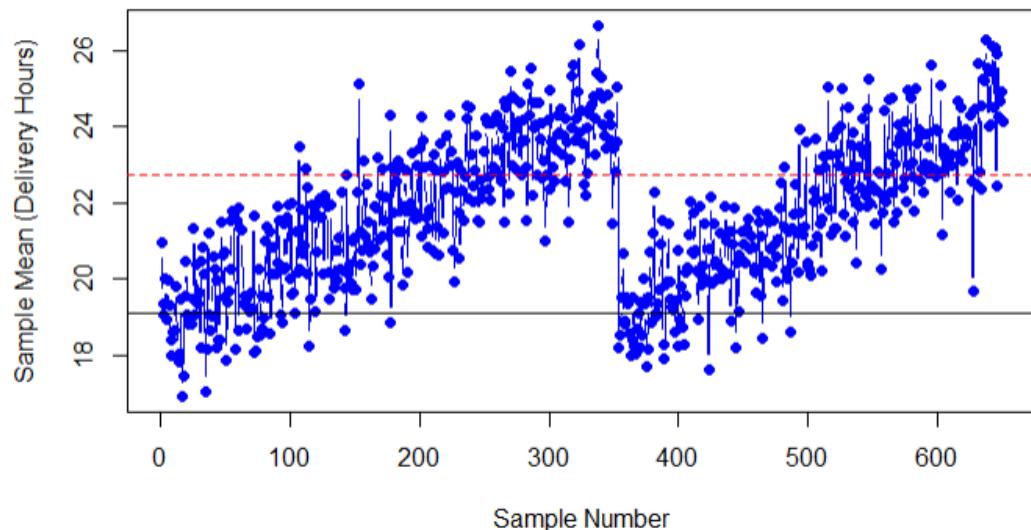
$C_p = 0.898$

$C_{pk} = 0.717$

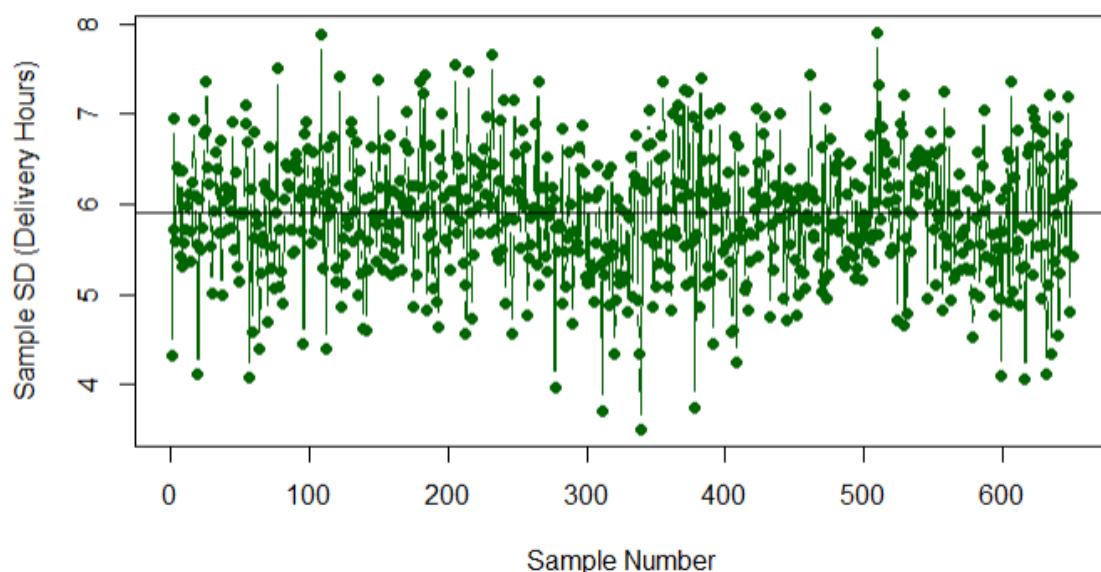
$C_{pu} = 0.717$

$C_{pl} = 1.079$

X-bar Chart for Product CLO



S Chart for Product CLO



An easier way to visualize the data:

Table 3.1: Control limits per product type

Product	X-bar UCL	X-bar CL	X-bar LCL	S UCL	S CL	S LCL
SOF	1.14	0.96	0.77	0.43	0.30	0.17
LAP	23.13	19.52	15.92	8.50	5.89	3.28
KEY	22.78	19.19	15.61	8.45	5.86	3.27
MON	23.05	19.43	15.80	8.54	5.92	3.30
CLO	22.74	19.13	15.51	8.52	5.91	3.29

Table 3.2 Process Capability indices

Product	CP	Cpk	Cpu	Cpl
SOF	18.135	1.083	35.188	1.083
LAP	0.899	0.696	0.696	1.101
KEY	0.917	0.729	0.729	1.105
MON	0.889	0.700	0.700	1.079
CLO	0.898	0.717	0.717	1.079

Table 3.3: Out of control Samples

Product	Number of out of control x-bar samples	Out of control S samples
SOF	Very high ,multiple 100+ points	None
LAP	70+	1 (sample 329)
KEY	150 +	None
MON	120 +	None
CLO	130 +	None

These results can be interpreted in the following way:

SOF(software) has extremely high CP(18.1) which indicates very little variation relative to specific limits. However, the large number of out-of-control points on the X-bar chart suggests instability due to possible process shifts or measurement inconsistencies

LAP, KEY, MON and CLO all have Cp Values below 1 which means their process spread exceeds specification limits (These are not capable of consistently meeting target performance)

For all products the Cpk is less than 1 which confirms that centering or variability issues exist, reducing process capability

S-Charts show generally stable spread (no out of control points), meaning variability within sample is consistent but x-bar charts show mean shifts over time

Overall, the manufacturing process is somewhat stable in variation but not fully capable with multiple mean shifts requiring process investigation or adjustment.

Using the Statistical Process Control analysis a few recommendations can be made:

While the variability within samples (as shown by the S charts) is stable across most product types, the x-bar charts highlight frequent mean shifts. This indicates that even though the process operates consistently in terms of variation, the average delivery times fluctuate over time, which suggests that there are periodic inefficiencies, workload surges or scheduling imbalances.

Instability reduces reliability of a business as delivery targets are not met and leads to inconsistent customer service. Products such as SOF show a particularly high process capability ($C_p > 1$) meaning they have the potential to perform well if properly centered, but the large number of out of control points indicates poor day to day control. In contrast the other product lines (LAP, KEY, MON, CLO) show C_p and C_{pk} values below 1, implying that the current process spread and centering are not sufficient to consistently meet the 0-32 hour delivery.

Part 4: Risk and Data Correction

4.1 Type 1 Error

A type 1 error occurs when a stable process is incorrectly signalled as “out of control”. This is also seen as a false alarm while the process is still operating as expected. Using a standard model with independent samples, false alarm probabilities for each rule are as follows:

A) 1 s-sample outside the upper $+3\sigma$ limit

$$\alpha_A = P(Z > 3) = 0.00135 (0.135\%)$$

This means that approximately 1 in every 740 samples will exceed the $+3\sigma$ limit purely by chance

B) For a run of seven consecutive samples above the centre line

$$\alpha B = 0.5^7 = 0.0078125 (0.781\%)$$

This means that approximately 1 in every 128 sequences of seven points will occur randomly under stable conditions.

C) 4 consecutive X-bar samples outside the upper $+2\sigma$ limits

$$\alpha C = 0.0228^4 = 2.68 \times 10^{-7} (= 0.0000268\%)$$

This is an extremely low probability which is about 1 in 3.7 million and it can be concluded that when this is the outcome the process is almost certainly out of control

These probabilities are calculated under the assumption that the process is stable (H_0 is true). If both sides of the chart are monitored (upper and lower limits), the total alpha should be doubled. In real applications, the expected number of false alarms is approximately equal to the number of opportunities multiplied by alpha.

4.2 Type II Error (β)

A type II error occurs when the process has actually shifted (out of control), but the sample mean (\bar{x} -bar) still falls inside the control limits, causing the shift to go undetected. The probability detecting the shift is called the power, given by $1 - \beta$

It is given that:

- X-bar chart limits: UCL = 25.089 L, LCL = 25.011
- In control centre line: CL = 25.050
- Shifted mean: $\mu_1 = 25.028$ L
- Standard deviation of X-bar under shift: $\sigma_x = 0.017$ L

Goal: $\beta = P(LCL \leq \bar{x} \leq UCL | \mu = \mu_1, \sigma_x = 0.017)$

Calculation: $\beta = \Phi\left(\frac{UCL - \mu_1}{\sigma_x}\right) - \Phi\left(\frac{LCL - \mu_1}{\sigma_x}\right)$

$$ZU = \frac{25.089 - 25.028}{0.017} = 3.588 \quad ZL = \frac{25.011 - 25.028}{0.017} = -1.000$$

$$B = \Phi(3.588) - \Phi(-1.000) = 0.9998 - 0.1587 = 0.841$$

$$\text{Power} = 1 - \beta = 0.159$$

Interpretation: with the process mean shifted to 25.028 L and the $\sigma_x = 0.017$ L, there is about an 84% chance that a sample mean will still fall inside the control limits. This means the chart fails to detect the shift most of the time. ($\beta = 0.841$). The corresponding power of 0.159 (15.9%) indicates that the current xbar chart is not very sensitive to this small process shift

In order to improve detection, β needs to be reduced and sensitivity needs to be increased. The sample size can be increased (reducing σ_x), or additional detection rules such as 2σ or 1σ rules, EWMA or CUSUM (de Vargas, Dias Lopes and Mendonça Souza, 2004) charts can be applied.

4.3 Correcting Head-office File

To correct the head-office file, the data was first examined and found to contain repeated pricing errors and incorrect product identifiers (it contained prefixes NA instead of SOF, LAP, MON, KEY, CLO). The markup and selling-price values after the first ten rows of each product type was also misaligned.

Before being fixed (products vs products2025) and (head-office vs headoffice2025)

Category <chr>	Local_Mean_Price <dbl>	Local_Mean_Markup <dbl>	HO_Mean_Price <dbl>	HO_Mean_Markup <dbl>
Cloud Subscription	3691.861	20.553	4386.710	21.50000
Keyboard	4638.172	20.161	4380.485	19.95417
Laptop	5217.545	20.623	4305.739	20.47517
Monitor	5014.170	20.727	4456.745	19.44250
Mouse	4585.465	20.668	4478.900	20.17967
Software	3814.344	20.038	4457.193	20.76150

After being fixed (products vs products2025) and head-office vs headoffice2025

Category <chr>	Local_Mean_Price <dbl>	Local_Mean_Markup <dbl>	HO_Mean_Price <dbl>	HO_Mean_Markup <dbl>
Cloud Subscription	1019.062	19.956	1019.062	19.956
Keyboard	644.660	23.981	644.660	23.981
Laptop	18086.429	18.430	18086.429	18.430
Monitor	6310.525	23.868	6310.525	23.868
Other	394.698	20.495	394.698	20.495
Software	506.183	16.040	506.183	16.040

When comparing products_2025 with products_Headoffice2025 by summary statistics:

After applying the correction, the mean selling prices and markups between products_data2025 and products_headoffice2025 were identical, confirming that the data is now synchronised. This ensures consistency cross local and head-office systems.

Old vs new head-office dataset

Category	Old_Mean_Price	Old_Mean_Markup	New_Mean_Price	New_Mean_Markup
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Cloud Subscription	4386.710	21.50000	1019.062	19.956
Keyboard	4380.485	19.95417	644.660	23.981
Laptop	4305.739	20.47517	18086.429	18.430
Monitor	4456.745	19.44250	6310.525	23.868
Mouse	4478.900	20.17967	NA	NA
Software	4457.193	20.76150	506.183	16.040

Other summary statistics

When it comes to other parts of analysis of the documents such as handling missing values, data filtering and sub setting, data visualization and exploring relationships, it was not done since it will not yield any crucial information. For example, once a dataset is cleaned it does not need to be cleaned again therefore missing values were previously handled. In the case of Exploring relationships, there is no need to explore the relationship between products_2025 with products_Headoffice2025, since their values are identical and the relationships are identical.

Part 5: Performance analysis – Barista service times

Shop 1

For shop 1 the average service time decreased rapidly as the number of baristas increased, indicating improved efficiency with added staff. Reliability rose sharply, reaching 100% from three baristas onwards, showing that service times consistently met the target threshold. Profit also increased steadily since more baristas enabled a higher number of customers to be served without compromising service speed. The most profitable and fully reliable setup was achieved with six baristas, generating approximately R2.93 million in profit

Baristas	Avg_Service	Total_Customers	Reliable_Pct	Profit
<int>	<dbl>	<int>	<dbl>	<dbl>
1	200.17	2196	0.00	64880
2	141.51	8859	0.12	263770
3	115.44	19768	79.35	590040
4	100.02	35289	100.00	1054670
5	89.44	54958	100.00	1643740
6	81.64	78930	100.00	2361900

Baristas	Avg_Service	Total_Customers	Reliable_Pct	Profit
<int>	<dbl>	<int>	<dbl>	<dbl>
1	200.16	417	0.00	11510
2	100.17	3556	99.72	104680
3	66.61	12126	100.00	360780
4	49.98	29305	100.00	875150
5	39.96	56701	100.00	1696030
6	33.36	97895	100.00	2930850

Shop 2

For shop 2, the average service time steadily decreased as the number of baristas increased, leading to faster and more consistent service. Reliability improved gradually, reaching 100% at four baristas and remaining stable thereafter, indicating that the shop achieved full service dependability beyond this point. Profit continued to rise with each additional barista added, as higher staffing allowed more customers to be served efficiently. The most profitable configuration was achieved with 6 baristas which generated about R 2.36 million in profit.

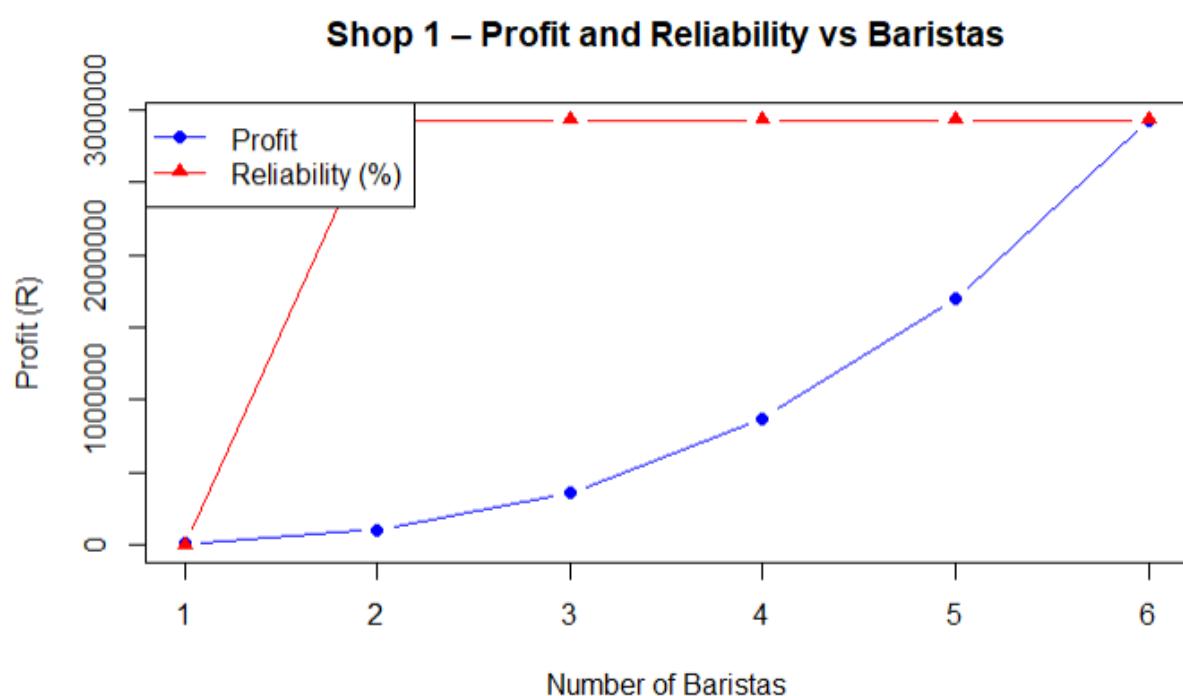


Figure 7: Profit and reliability vs Baristas

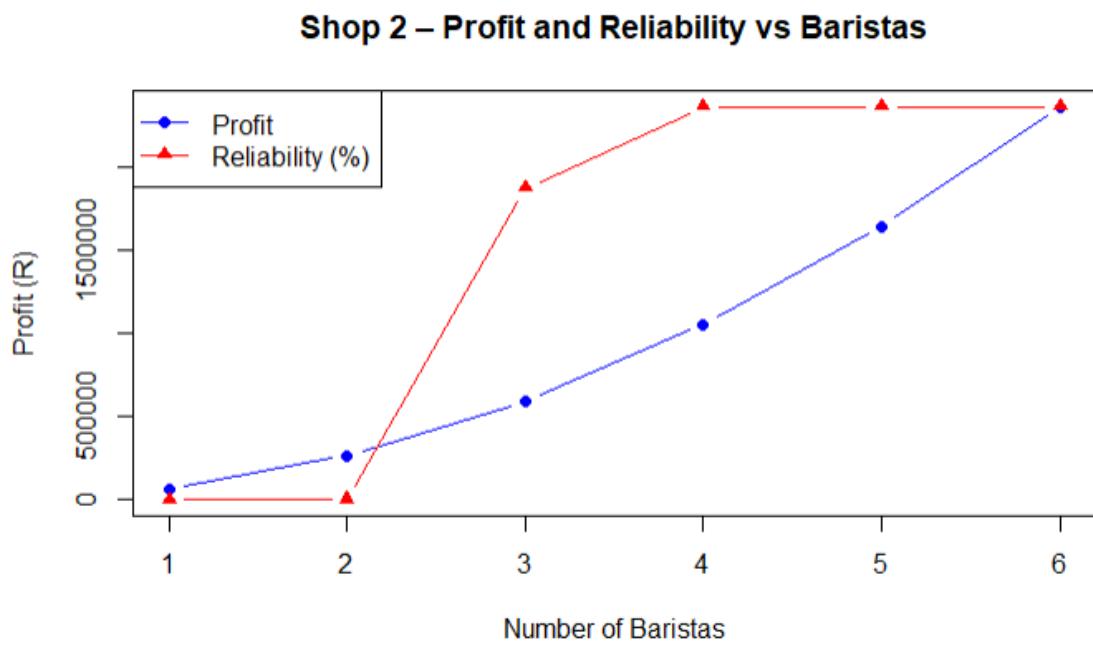


Figure 8: Profit and reliability vs Baristas

Linear regression confirmed a strong positive relationship between the number of baristas and total profit. Operationally this means both shops gain efficiency and customer satisfaction as staffing rises to the point of full reliability. After that, extra staff mainly boost capacity and revenue rather than service quality

It is therefore recommended that each shop maintain six baristas per shift in both shops to maximize profit while maintaining full reliability.

Part 6 : Comparative Manova of delivery and performance analysis

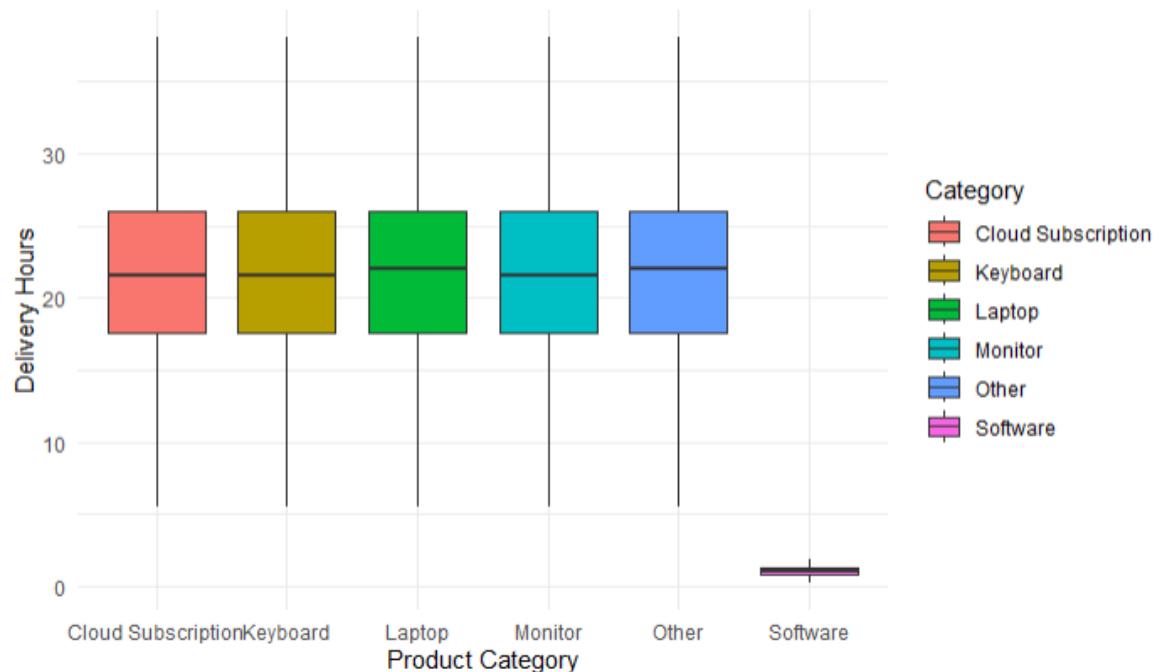


Figure 9: Product category vs delivery hours

A one-way ANOVA was conducted to determine whether the product category has a significant effect on delivery hours and selling price. The results show that product category strongly influences both delivery performance and pricing ($p<0.001$ for both tests). The F – value for delivery hours ($F = 47\ 363$) and for selling price ($F = 2\ 796\ 296$) are both extremely large, confirming that differences among product types are significant and worth noting.

The boxplot of delivery hours supports these results as it shows that most product categories (LAP, SOF, MON, KEY, CLO) have similar median delivery times (about 20 hours) which indicates standard logistics across all products. In contrast Software category displays near zero delivery hours which reflects that software is installed immediately as it is a digital product.

Therefore, it is clear that physical products require more handling and transport whereas digital products are delivered to customers instantly. Similarly, the selling price differences reflect how much value a product has, where higher priced items require more logistical effort and longer delivery cycles as it needs to be handled with caution as delivery mistakes could lead to big losses

Overall, the ANOVA results validate that the product category is a key driver of both pricing and operational efficiency. This emphasizes the need for tailored delivery for each product type and that products should not be handled uniformly.

Part 7: Reliability of service

7.1 Reliability Estimation

Based on the data provided for the number of staff on duty over 397 days, the average staffing level was 15.58 employees, with a standard deviation of 0.69. Using the $\pm 3\sigma$ control limit method, the upper control limit (UCL) was calculated as 17.65 and the lower control limit (LCL) as 13.52. 391 of the 397 days (94.49%) fell within these limits.

It is therefore clear that the company operates under reliable service conditions around 95% of the time as the daily staffing remains stable and almost never falls out of the expected range. The values correspond with the histogram in the brief, which also shows that most days have between 14 and 16 staff members on duty.

It can be concluded that the company keeps a consistent staffing pattern which ensures stable service quality throughout the year.

7.2 Profit optimization

A profit model was developed to determine the optimal staffing level that maximizes the profit, daily. Assuming each staff member serves 25 rental cars per day and generates R1200 in revenue per rental and costs R1000 per day

Profit of each staffing level:

$$\text{Profit} = (\text{staff} \times 25 \times 1200) - (\text{staff} \times 1000)$$

Staff	Profit
<dbl>	<dbl>
12	348000
13	377000
14	406000

15	435000
16	464000

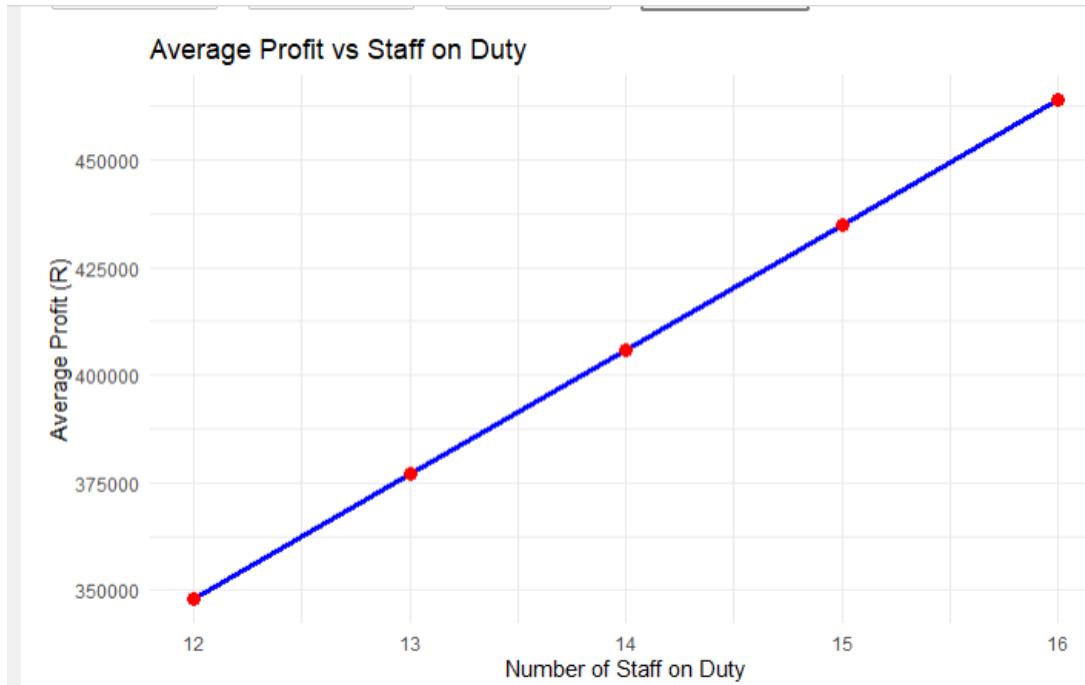


Figure 10: Average Profit vs Staff on Duty

The “Average profit vs staff on duty” graph indicates that profit increases consistently as staff increases. It rises from R34800 at 12 staff members to R464000 at 16 staff members. This indicated that higher staffing improves service levels and revenue without reaching overstaffing.

The results clearly show that profit increases consistently with the number of staff on duty which ultimately reaches a maximum at 16 employees. This has a clear correlation with the histogram given in the assignment brief.

Therefore, it is clear that the optimal staffing level for maximising profitability while maintaining service reliability is 16 staff members per day. Beyond this point it is likely that additional staff members would cost more money as the demand needs to align with the amount of staff members. Also, the revenue gained per rental would not increase proportionally when more staff members are employed beyond the demand for rental cars. Maintaining 16 employees per day achieves the best balance between profit and service reliability

Conclusion

This project provided an in-depth analysis of multiple aspects of businesses in different industries. It included the operational and performance aspects of the retail, manufacturing and service sectors across different datasets using R. During data preparation and descriptive analysis, customer, product and sales data were cleaned, summarised and visualized to understand the obvious trends and patterns. Statistical process control (SPC) using x-bar and s charts showed that while most processes were stable, several product types experienced mean shifts which indicates opportunity for improvement. ANOVA results confirmed that product category significantly affects both delivery time and selling price. Digital products had instant delivery where physical products had longer delivery times, which is expected. The car rental company showed that keeping 16 staff members achieved maximum profit and good service reliability (95%) while not overstaffing. Overall, the project combined data preprocessing, quality control, and optimisation models to deliver results that can help a business take action to improve efficiency and profitability.

References

Porch Group Media. “Generational Marketing: How to Target Different Age Groups | Porch Group Media.” *Porch Group Media*, 25 Aug. 2023, porchgroupmedia.com/blog/generational-marketing/.

de Vargas, V. do C.C., Dias Lopes, L.F. and Mendonça Souza, A. (2004). Comparative study of the performance of the CuSum and EWMA control charts. *Computers & Industrial Engineering*, 46(4), pp.707–724. Doi: <https://doi.org/10.1016/j.cie.2004.05.025>.

The R scripts were written with guidance from ChatGPT (OpenAI, 2025).

OpenAI. 2025. *ChatGPT*. [online] Available at: <https://chat.openai.com> [Accessed 24 Oct. 2025].