ECSA GA4 Report

Prepared for: Quality Assurance 344 and Engineering Counsel of South Africa (ECSA)

Prepared by: Jodi De Lange

Student number:27043061

Date: 24 October 2025

Table of Contents

1. Introduction

Conducting investigations, analysing data and designing experiments or models is vital to the field of industrial engineering. R programs are written to handle and store datasets. This report includes an attached Rmd file containing the written user interface that allows a user to load process data and make control decisions. Raw data undergoes a complete descriptive and statistical analysis. This meets the criteria of Graduate Attribute 4 being assessed in Quality Assurance 344.

This report begins with a basic data analysis of 4 provided csv files. These files include customer, product, sales and head office data. The steps in the data analysis process are followed. Control Charts are then created to analyse future sales and identify any control issues. Risk and data correction are then attended to.

The report then discusses a model designed by the student which optimises profit for two coffeeshops. An ANOVA is created and interpreted in R using hypotheses generated by the student. Lastly the reliability of a service is considered followed by a conclusion.

2. Data analysis Process

2.1. Data loading and Inspection

The provided data sets are loaded into R. Four csv files are provided for the data analysis. The customer_data.csv describes each customer by a unique customer ID, gender, age, income and city of residence. The data set contains 5000 customer records. The sales2022and2023.csv describe the sales which occurred in 2022 and 2023. It includes a customer ID, purchase ID, quantity ordered, time of the order, day of the order, month of the order as well as hours spent picking and for delivery for every sale made. The data set contains 100 000 sale records. The products_Headoffice.csv contains the product ID, category, description, selling price and markup for 360 different products. The products_data.csv lists the products which the company sells. It contains a product ID, category, description, selling price and markup for 60 different products.

The data is then inspected. It is found that the company offers 60 different products[1] namely 10 software products, 10 cloud subscription products, 10 laptop products,10 monitor products, 10 keyboard products and 10 mouse products. It is therefore determined that the company specializes in technology.

After initial inspection, clear anomalies are visible such as differences in the csv files data. This will be addressed in Section 4.3. More valuable insight will be gained from summary statistics.

[1] Note: throughout the report the following prefixes refer to the following product categories:
SOF is software, CLO is cloud subscription, LAP is laptop, MON is monitor, KEY is keyboard and MOU is mouse.

Below is an extract[2] from the products dataset (left) and the head office products dataset (right).

	product_id	category	description	selling_price	markup
1	SOF001	Software	coral matt	511.53	25.05
2	SOF002	Cloud Subscription	cyan silk	505.26	10.43
3	SOF003	Laptop	burlywood marble	493.69	16.18
4	SOF004	Monitor	blue silk	542.56	17.19
5	SOF005	Keyboard	aliceblue wood	516.15	11.01
6	SOF006	Mouse	black silk	478.93	16.99
7	SOF007	Software	black bright	527.56	16.79
8	SOF008	Cloud Subscription	burlywood silk	549.02	11.95
9	SOF009	Laptop	azure sandpaper	540.41	11.34
10	SOF010	Monitor	chocolate sandpaper	396.72	23.47
11	CLO011	Keyboard	burlywood silk	1070.54	16.41
12	CLO012	Mouse	azure silk	963.14	10.13
13	CLO013	Software	chartreuse silk	1067.54	16.80
14	CLO014	Cloud Subscription	burlywood silk	1083.11	21.25

	product_id	category	description	selling_price	markup
1	SOF001	Software	coral silk	521.72	15.65
2	SOF002	Software	black silk	466.95	28.42
3	SOF003	Software	burlywood marble	496.43	20.07
4	SOF004	Software	black marble	389.33	17.25
5	SOF005	Software	chartreuse sandpaper	482.64	17.60
6	SOF006	Software	cornflowerblue marble	539.33	25.57
7	SOF007	Software	blue marble	495.13	10.23
8	SOF008	Software	cornflowerblue marble	465.73	21.89
9	SOF009	Software	black bright	452.40	19.64
10	SOF010	Software	cornflowerblue matt	399.43	17.08
11	NA011	Software	aliceblue silk	823.51	14.59
12	NA012	Software	coral marble	987.13	27.59
13	NA013	Software	cornflowerblue sandpaper	1176.31	18.30
14	NA014	Software	azure silk	1061.21	20.03

Below is an extract from the sales dataset and customer data below that.

	customer_id	product_id	quantity	order_time	order_day	order_month	order_year	picking_hours	delivery_hours
1	CUST1791	CLO011	16	13	11	11	2022	17.7216667	24.5440
2	CUST3172	LAP026	17	17	14	7	2023	38.3908333	31.5460
3	CUST1022	KEY046	11	16	23	5	2022	14.7216667	21.5440
4	CUST3721	LAP024	31	12	18	7	2023	41.3908333	24.5460
5	CUST4605	CLO012	20	14	7	2	2022	15.7216667	24.0440
6	CUST2766	MON035	32	21	24	12	2022	21.0550000	24.0440
7	CUST4454	MOU052	29	5	23	1	2022	12.3883333	25.5440
8	CUST582	MON032	1	19	9	6	2023	17.0575000	22.0460
9	CUST3343	MON040	10	19	13	12	2023	24.0575000	24.0460
10	CUST4331	KEY049	1	18	30	4	2022	15.3883333	20.0440

	customer_id	gender	age	income	city
1	CUST001	Male	16	65000	New York
2	CUST002	Female	31	20000	Houston
3	CUST003	Male	29	10000	Chicago
4	CUST004	Male	33	30000	San Francisco
5	CUST005	Female	21	50000	San Francisco
6	CUST006	Male	32	80000	Miami
7	CUST007	Female	31	100000	Los Angeles
8	CUST008	Male	27	90000	Los Angeles
9	CUST009	Female	26	35000	Chicago
10	CUST010	Male	28	105000	San Francisco

2.2. Summary statistics

The summary statistics are then generated for each dataset provided. This will provide a better understanding of the data and the relationships between features.

[2] Tables in this format are directly inserted from the R code output.

The summary statistics for the sales data is seen in the figure below. It should be noted that the product and customer ID for all the datasets are unique and therefore their summary statistics are not relevant. It is seen that the quantity per order varies from 1 to a maximum of 50 with a mean of 13.5 and standard deviation of 13.8 which indicates majority of customers order multiple (more than 13) products in one sale. The orders are placed between 1:00 and 23:00 while most occur between 12:00 and 13:00. The order day reveals that most in the middle of the month and order year reveals that most occur between June and July (middle of year). The minimum and maximum of these features corresponds to the number of days in a month and months in a year respectively. The average time spent picking an order is 14.7 hours and time till delivery complete is 17.5 hours.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
customer_id*	1	1e+05	2492.33848	1444.5778106	2503.000	2491.191988	1862.1456	1.0000000	5000.0000	4999.00000	0.002096356	-1.2107656	4.568156139
product_id*	2	1e+05	32.43610	18.0302099	35.000	32.819100	23.7216	1.0000000	60.0000	59.00000	-0.160340655	-1.3178154	0.057016530
quantity	3	1e+05	13.50347	13.7601316	6.000	11.458100	5.9304	1.0000000	50.0000	49.00000	1.044341146	-0.2185180	0.043513357
order_time	4	1e+05	12.93230	5.4951268	13.000	13.117888	5.9304	1.0000000	23.0000	22.00000	-0.227168462	-0.7101693	0.017377117
order_day	5	1e+05	15.49683	8.6465055	15.000	15.495088	10.3782	1.0000000	30.0000	29.00000	0.002772591	-1.2007412	0.027342651
order_month	6	1e+05	6.44813	3.2834460	6.000	6.445537	4.4478	1.0000000	12.0000	11.00000	0.006928166	-1.1764404	0.010383168
order_year	7	1e+05	2022.46273	0.4986115	2022.000	2022.453413	0.0000	2022.0000000	2023.0000	1.00000	0.149493651	-1.9776714	0.001576748
picking_hours	8	1e+05	14.69547	10.3873345	14.055	13.543098	6.9188	0.4258889	45.0575	44.63161	0.735709308	0.4143469	0.032847636
delivery_hours	9	1e+05	17.47646	9.9999440	19.546	17.775077	8.8956	0.2772000	38.0460	37.76880	-0.470487992	-0.8716457	0.031622600

The summary statistics for head office product data is seen in the figure below. The category and description statistics are irrelevant as these are not numerical features but rather categorical features. Selling price of the products vary from R290.52 to R22 420.14. The average selling price is R4 410.96. These costs make logical sense as it is known that technology tends to have high costs. The percentage markup ranges from 10 to 30 %. This impacts how much money the company will make. It is crucial for later profit calculations.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
product_id*	1	360	69.38889	23.217847	72.000	71.88542	22.239000	1.00	110.00	109.00	-0.86651723	0.4912730	1.22368796
category*	2	360	3.50000	1.710202	3.500	3.50000	2.223900	1.00	6.00	5.00	0.00000000	-1.2781771	0.09013556
description*	3	360	30.68611	17.319505	29.500	30.76736	22.980300	1.00	60.00	59.00	-0.02778185	-1.3900365	0.91281808
selling_price	4	360	4410.96186	6463.822788	797.215	3054.22903	515.752062	290.52	22420.14	22129.62	1.52790958	0.7789339	340.67337335
markup	5	360	20.38550	5.665949	20.580	20.42868	6.664287	10.06	30.00	19.94	-0.04776921	-1.0739041	0.29862173

The summary statistics for the product data is seen in the figure below. The selling price and markup is similar to that in the head office product data which makes logical sense as the data is for the same company.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
product_id*	1	60	30.50000	17.464249	30.500	30.50000	22.239000	1.00	60.00	59.00	0.00000000	-1.2601448	2.2546249
category*	2	60	3.50000	1.722237	3.500	3.50000	2.223900	1.00	6.00	5.00	0.00000000	-1.3258048	0.2223399
description*	3	60	16.40000	10.078001	16.000	16.20833	13.343400	1.00	35.00	34.00	0.10295987	-1.2935763	1.3010643
selling_price	4	60	4493.59283	6503.770150	794.185	3189.25479	525.722547	350.45	19725.18	19374.73	1.42617520	0.4338057	839.6331159
markup	5	60	20.46167	6.072598	20.335	20.51187	7.309218	10.13	29.84	19.71	-0.03670775	-1.2380989	0.7839690

The summary statistics for the customer data is seen in the figure below. Once again categorical features statistics are ignored. The gender of a customer is classified as male, female or other. The age of customers ranges from 16 to 105 years with most being 51.6 years old. The income has a very large standard deviation which is expected as there is a large variation of customer incomes. The average income however is R80 797.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
customer_id*	1	5000	2500.5000	1443.520003	2500.5	2500.50000	1853.2500	1	5000	4999	0.0000000	-1.2007200	2.041446e+01
gender*	2	5000	1.5572	0.577923	2.0	1.51700	1.4826	1	3	2	0.4538869	-0.7240466	8.173065e-03
age	3	5000	51.5538	21.216096	51.0	50.88275	26.6868	16	105	89	0.2041739	-0.9874439	3.000409e-01
income	4	5000	80797.0000	33150.106741	85000.0	81665.00000	37065.0000	5000	140000	135000	-0.2135307	-0.7456542	4.688133e+02
city*	5	5000	3.9918	2.002232	4.0	3.98975	2.9652	1	7	6	-0.0108635	-1.2745838	2.831584e-02

2.3. Handling Missing Values

Missing values may negatively impact on the results of the data analysis unless dealt with appropriately. From the previous section of this report, it is discovered that order date in the sales data contains missing values. There are very few, only 560, relative to the size of the dataset,10 000, therefore it can still be used as a descriptive feature in this analysis. The order data is also not critical to the data analysis process.

2.4. Data filtering and Sub setting

The data can be divided into categories of products. The data contains 6 categories. Valuable information such as most revenue generating category or category with the least sales may be useful for the company when planning marketing schemes or production plans. The revenue per category can be seen below. Clearly the laptops generate the most revenue for the company.

	category	total_rev
1	Laptop	821533851
2	Monitor	809104952
3	Keyboard	723693159
4	Mouse	721090260
5	Software	655365933
6	Cloud Subscription	621799523

Another useful division is by city in which purchase occurred. Here can be seen that San Francisco places the most orders and Miami the least.

	city	avg_deliverytime	orders
1	Houston	17.60277	14325
2	Los Angeles	17.58345	14978
3	New York	17.55999	14423
4	Chicago	17.51673	14062
5	Miami	17.42527	12816
6	San Francisco	17.38119	15803
7	Seattle	17.25421	13593

The delivery time and number of orders can also be calculated per city. This indicates that the most orders do not correlate to the longest lead time as one might expect from a busy store.

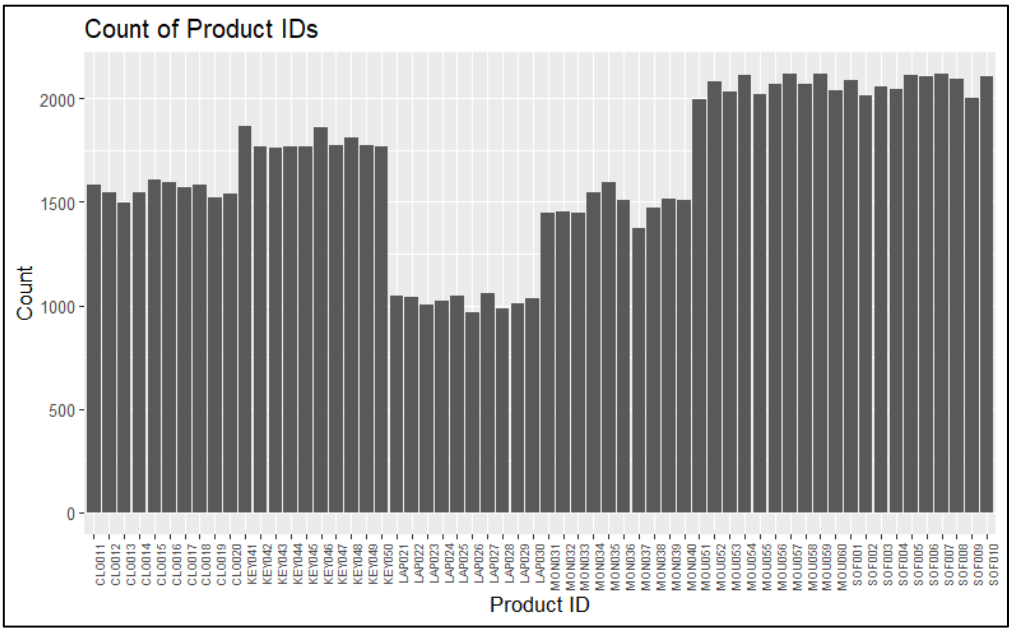
	city	avg_deliverytime	orders
1	Houston	17.60277	14325
2	Los Angeles	17.58345	14978
3	New York	17.55999	14423
4	Chicago	17.51673	14062
5	Miami	17.42527	12816
6	San Francisco	17.38119	15803
7	Seattle	17.25421	13593

Subsetting the data may be useful for simple analysis such as the two instances above. However much more valuable information will be gained by analysing visual representations of the data available as done in the next section of this report.

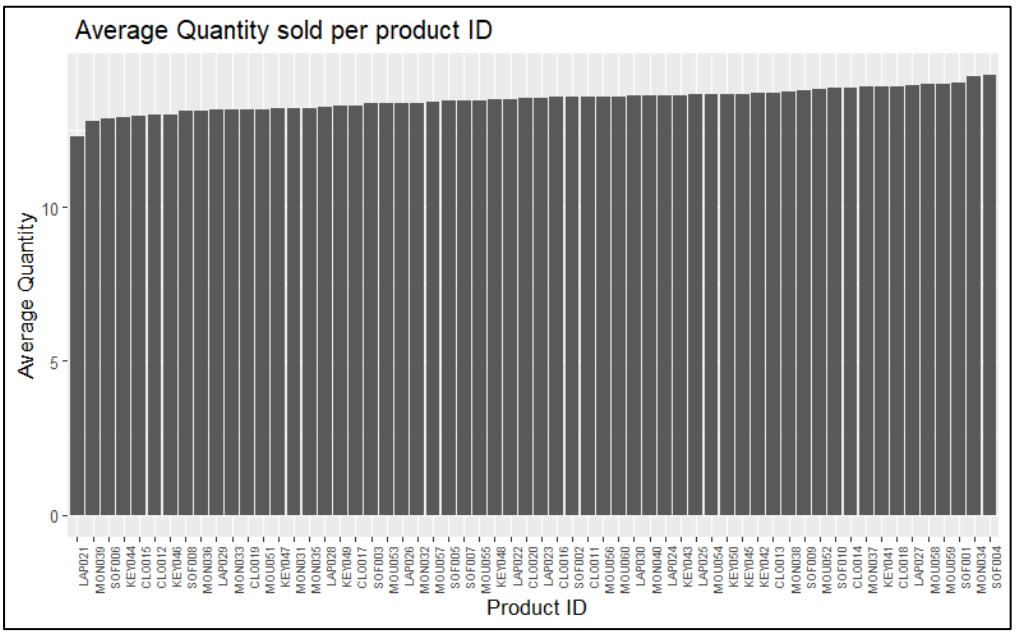
2.5. Data Visualization

The next step in the data analysis process is data visualization. This is critical for understanding how feature impact on each other and to gain a good understanding of the data provided and identify any underlying patterns which can be utilized.

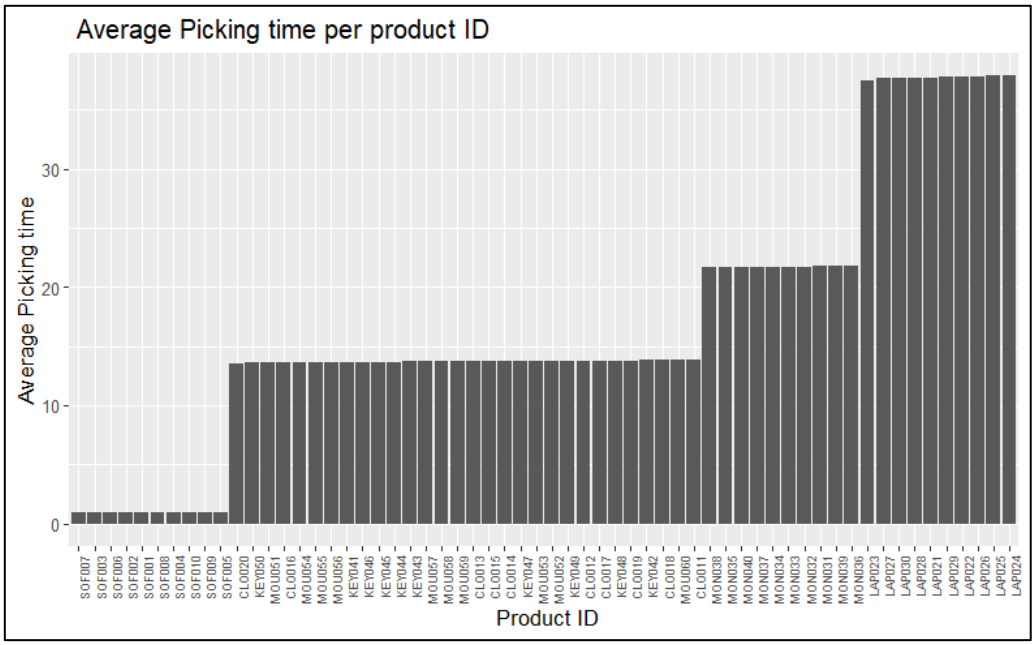
The **sales data** will be visualized first. Firstly, the count for each product ID must be plotted. This indicates which products are more popular in other words which product ID occurs in the most sales. Here is it clearly seen that mouse and software are most purchased while the least being laptops.



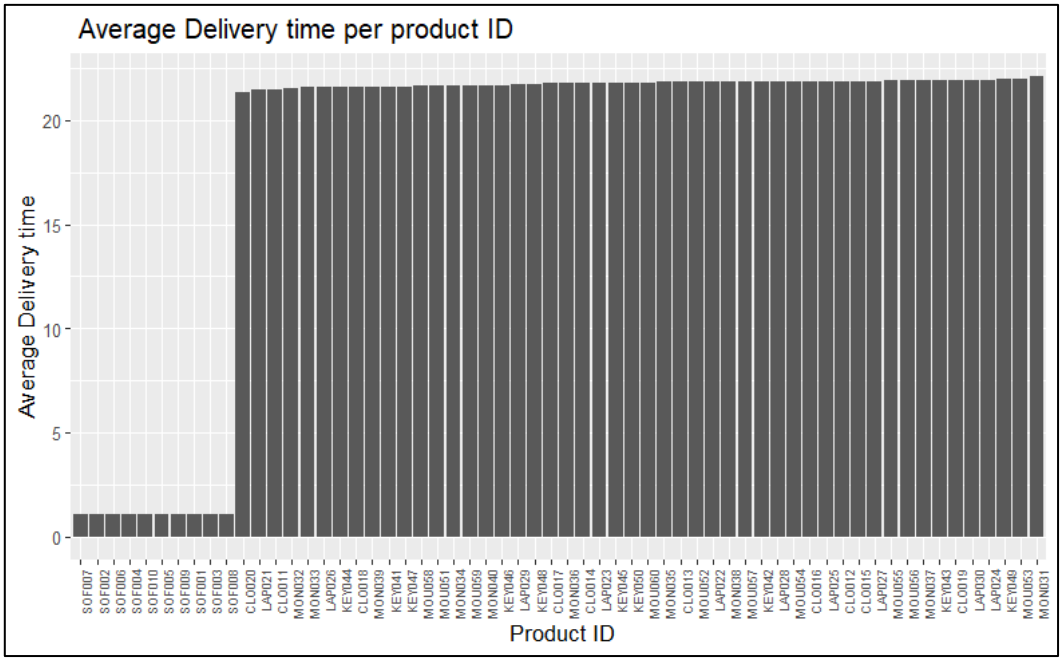
Below is a bar graph of the average quantity purchased per order for each product ID. It is seen that the quantity is quite similar for all product IDS, about 13. This agrees with the mean of 13.5 from Section 2.2 of this report.



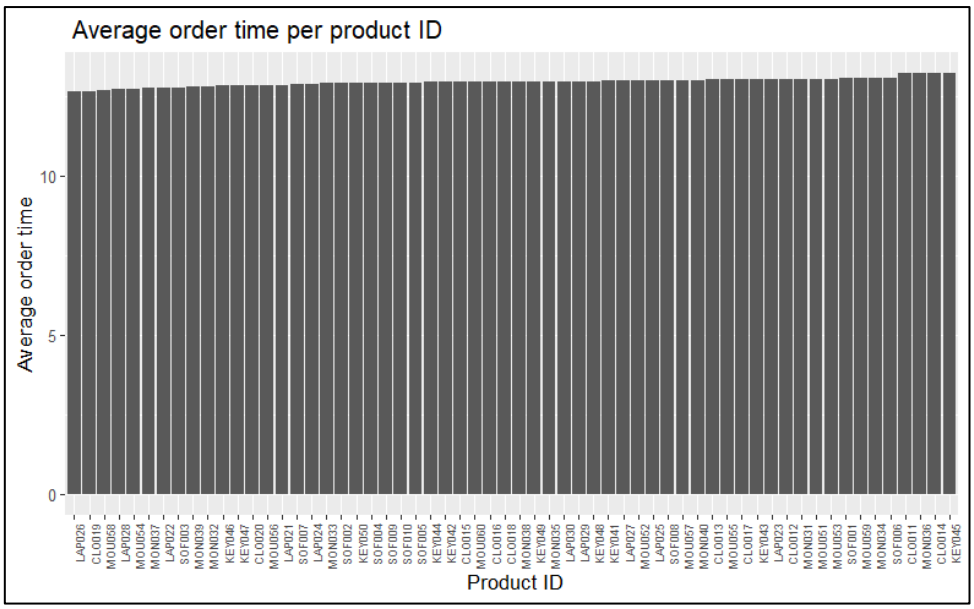
The average picking time per product ID shows a clean picking time for each category of product. Most laptops take approximately 38 hours while software takes 2 hours. This makes logical sense as software is often stored electronically and therefore can be accessed quickly whereas a laptop must be picked from a warehouse or storage facility.

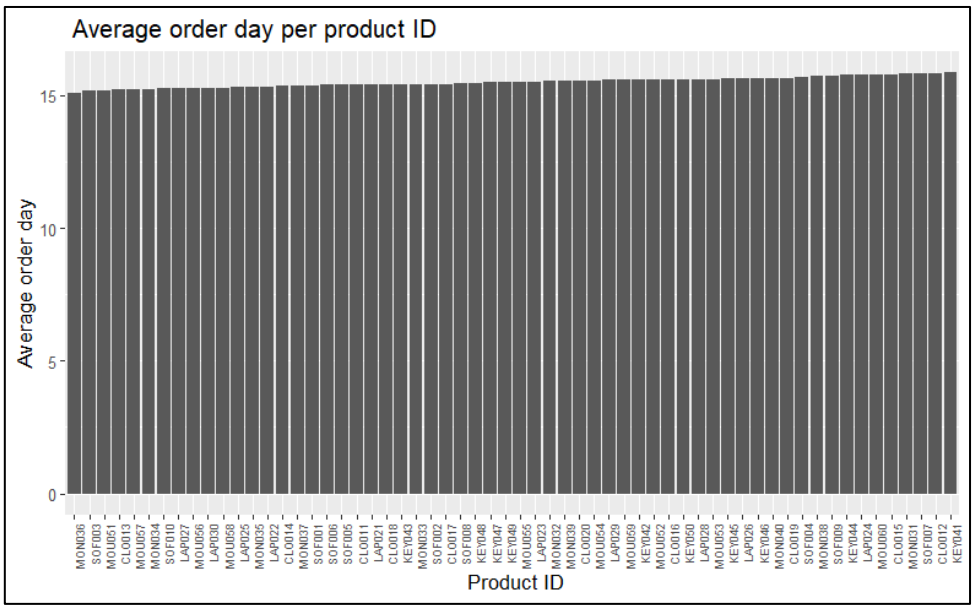


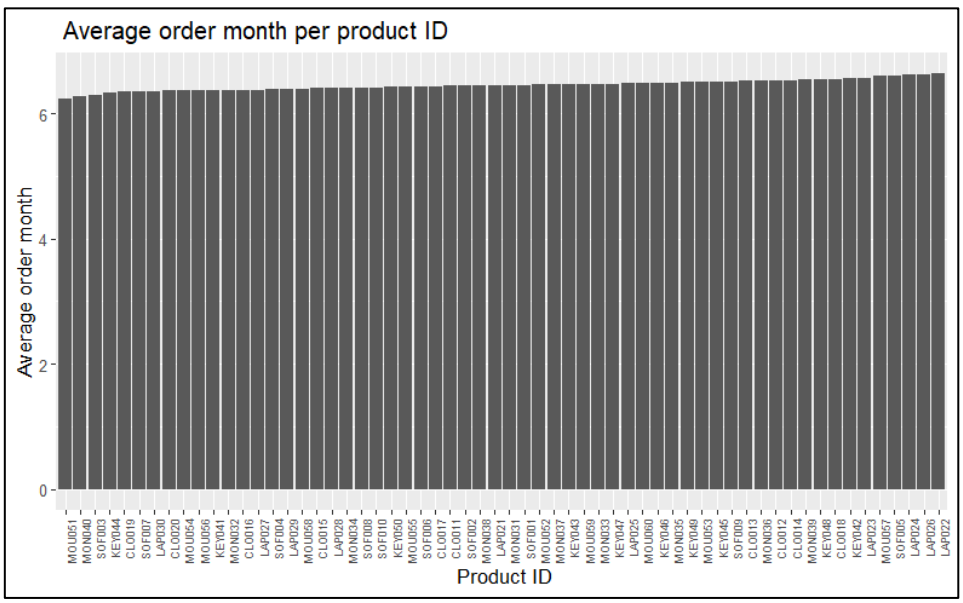
The same can be said for the figure below. The software is delivered almost instantaneously as it is simply uploaded whereas the physical items such as laptops have longer delivery times as they are transported from storage to customer by road.



The next three bar graphs depict order time, day and month. All the values are roughly the same which agrees with the summary statistics for the sales data in Section 2.2.; this data provides no valuable insight and can be classified as redundant.

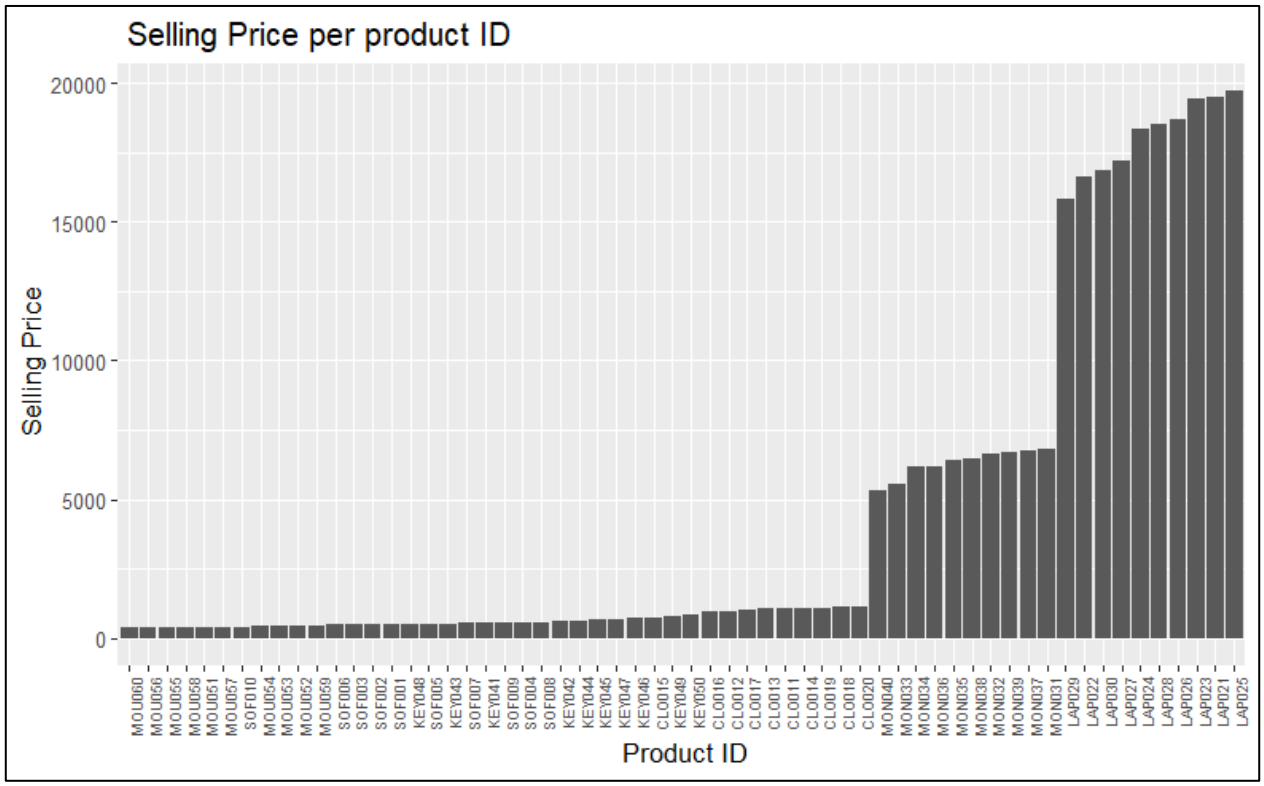

Average order time per product ID
Average order time
0
5
10
LAP026
CLO019
MOU058
LAP028
MOU054
MOND37
LAP022
SOF003
MOND39
MOND32
KEY046
KEY047
CLO020
MOU056
LAP021
SOF007
LAP024
MOND33
SOF002
KEY050
SOF004
SOF009
SOF010
SOF005
KEY044
KEY042
CLO015
MOU060
CLO016
CLO018
MOND38
KEY040
MOND35
LAP030
LAP029
KEY048
KEY041
LAP027
MOU052
LAP025
SOF008
MOU057
MOND40
CLO013
MOU055
CLO017
KEY043
LAP023
CLO012
MOND31
MOU051
MOU053
SOF001
MOU059
MOND34
SOF006
CLO011
MOND36
CLO014
KEY045
Product ID


Average order day per product ID
Average order day
0
5
10
15
MOND36
SOF003
MOU051
CLO013
MOU057
MOND34
SOF010
LAP027
MOU056
LAP030
MOU058
LAP025
MOND35
LAP022
CLO014
MOND37
SOF001
SOF006
SOF005
CLO011
LAP021
CLO018
KEY043
MOND33
SOF002
CLO017
SOF008
KEY048
KEY047
KEY049
MOU055
LAP023
MOND32
MOND39
CLO020
MOU054
LAP029
MOU059
KEY042
MOU052
CLO016
KEY050
LAP028
MOU053
KEY045
LAP026
KEY046
MOND40
CLO019
SOF004
MOND38
SOF009
KEY044
LAP024
MOU060
CLO015
MOND31
SOF007
CLO012
KEY041
Product ID

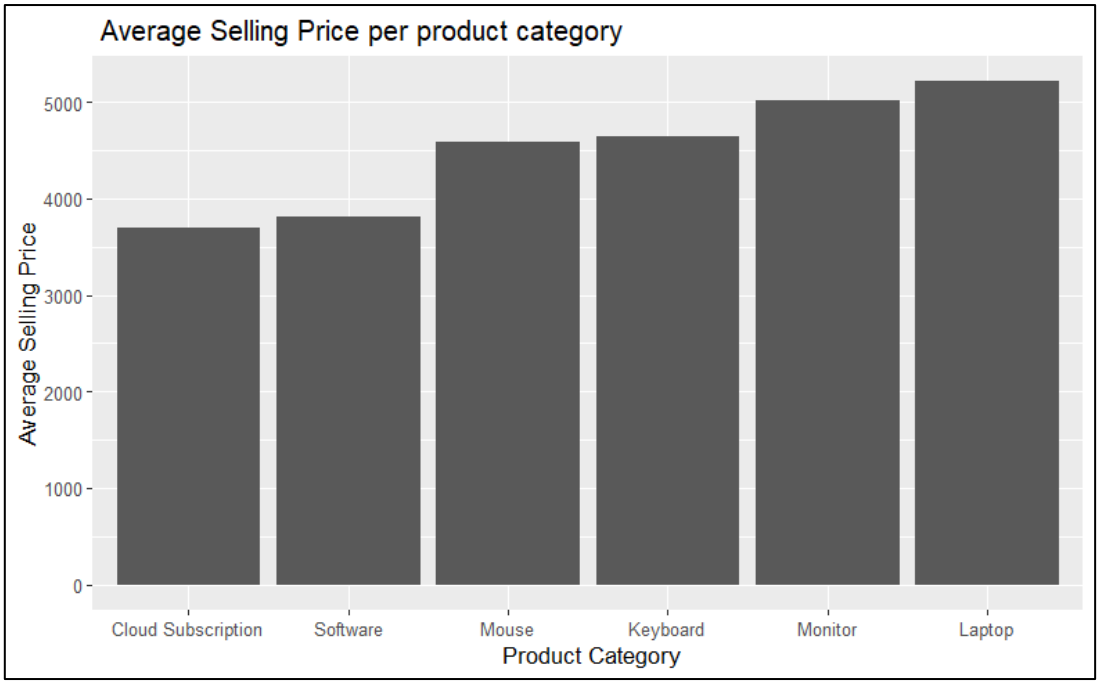

Average order month per product ID
Average order month
0
2
4
6
MOU051
MOND40
SOF003
KEY044
CLO019
SOF007
LAP030
CLO020
MOU054
MOU056
KEY041
MOND32
CLO016
LAP027
SOF004
LAP029
MOU058
CLO015
LAP028
MOND34
SOF008
SOF010
KEY050
MOU055
SOF006
CLO017
CLO011
SOF002
MOND38
LAP021
MOND31
SOF001
MOU052
MOND37
KEY043
MOU059
MOND33
KEY047
LAP025
MOU060
KEY046
MOND35
KEY049
MOU053
KEY045
SOF009
CLO013
MOND36
CLO012
CLO014
MOND39
KEY048
CLO018
KEY042
LAP023
MOU057
SOF005
LAP024
LAP026
LAP022
Product ID

.

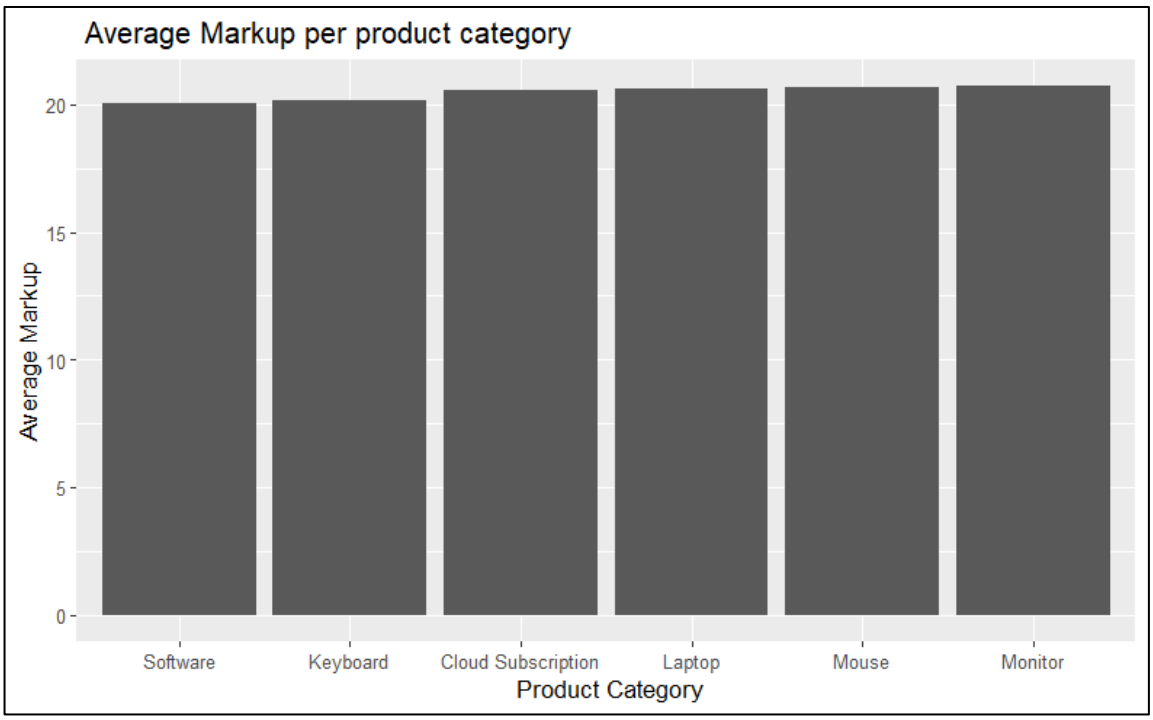
The next data set that will be visualized is **product data**. In this dataset 60 products are listed which each have a unique ID. In the first plot, a bar graph of selling price per product can be seen. Laptops are clearly the category with the highest selling price and mouse the lowest selling price.



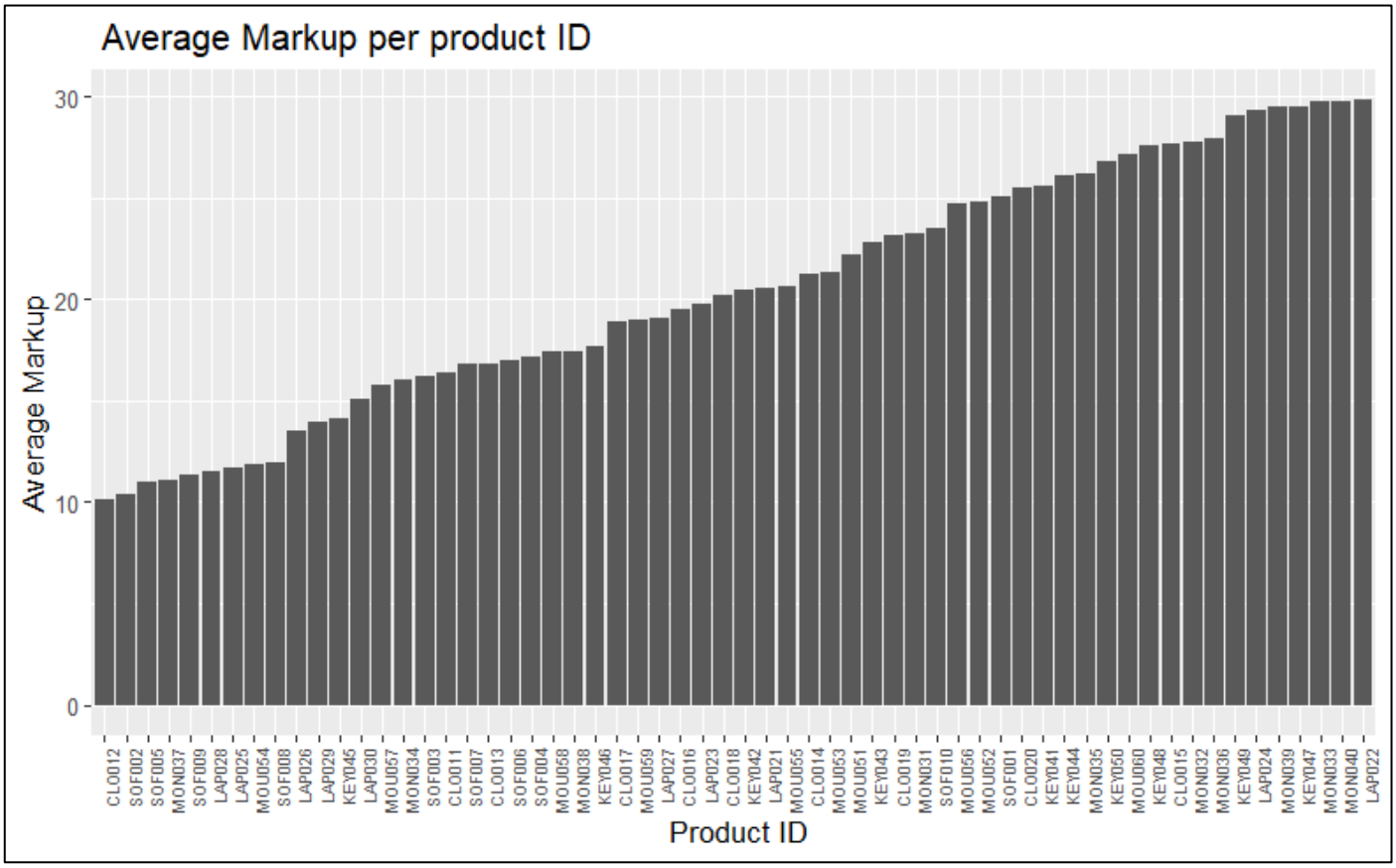
The next figure indicates which categories have the highest average selling price. It can clearly be seen once again that laptops have the highest selling price and cloud subscriptions now have the lowest.



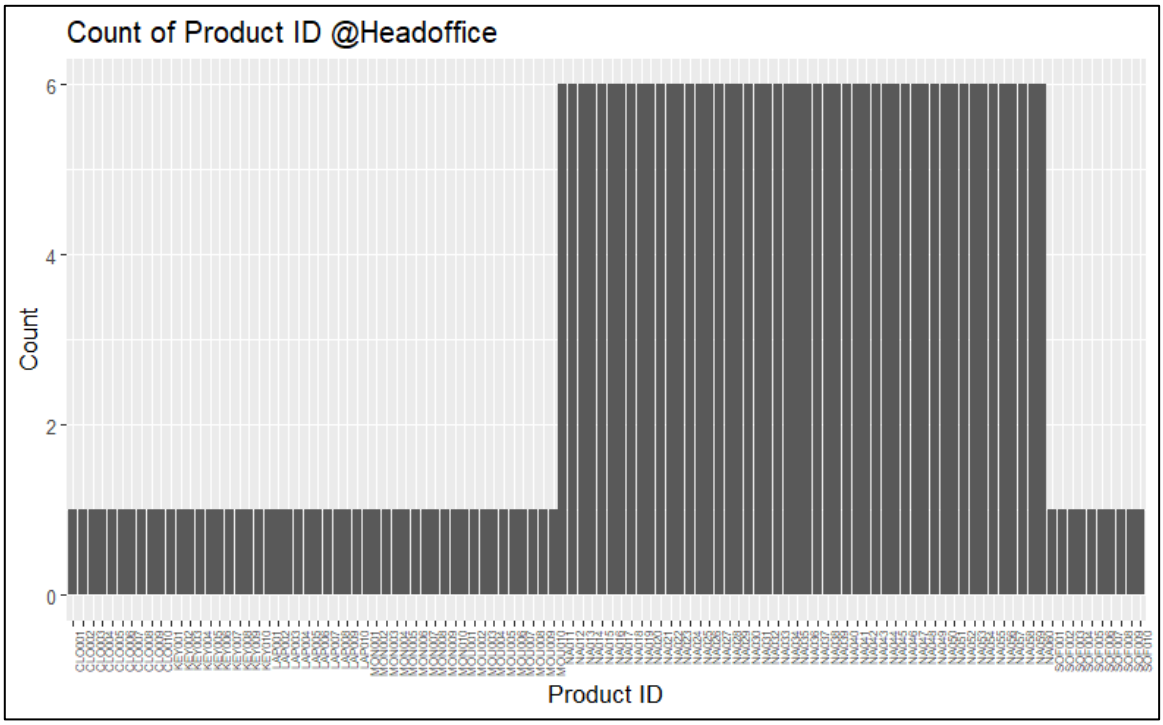
The average markup per product category is shown below. It is seen that the markups do not vary greatly between categories. The average being roughly 20 % markup.



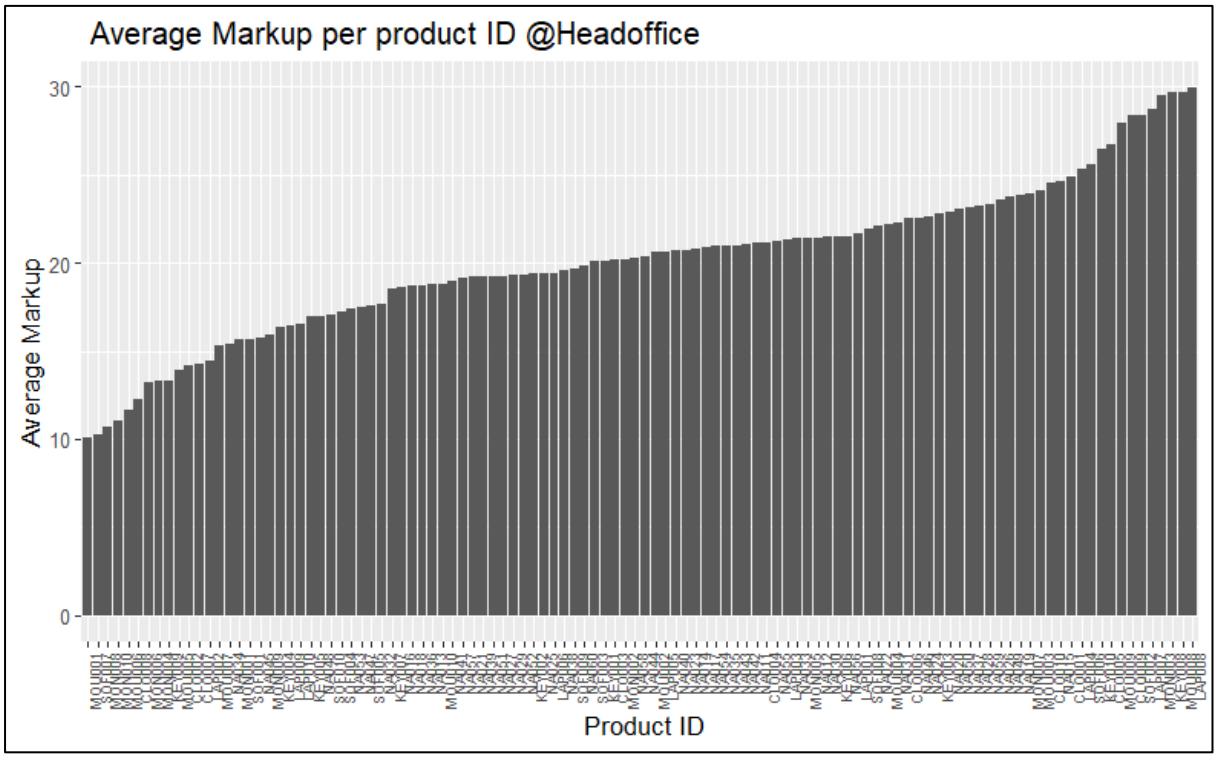
The markup for each individual product is shown below. The LAP002 has the highest and CLO012 has the lowest markup.



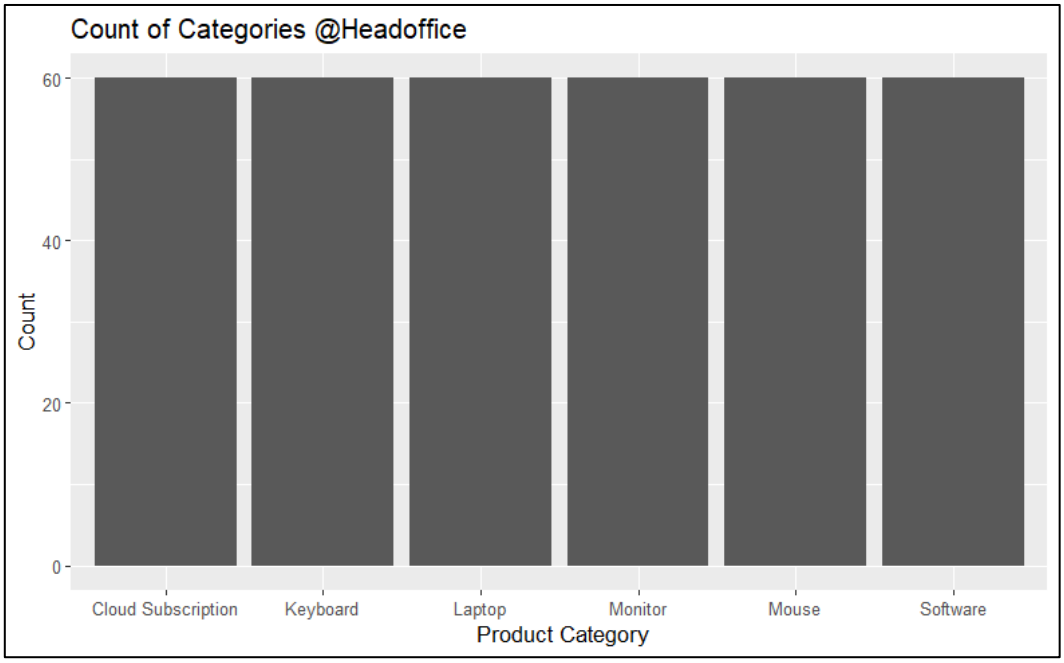
The next data set that will be visualized is **head office product data**. The count of product ID in the head office data shows that many NA product IDs exist, specifically 6 of each numerical value from 11 to 60. This indicates that these were mislabelled and should have a prefix of the appropriate category. This is addressed in Section 4 of this report.



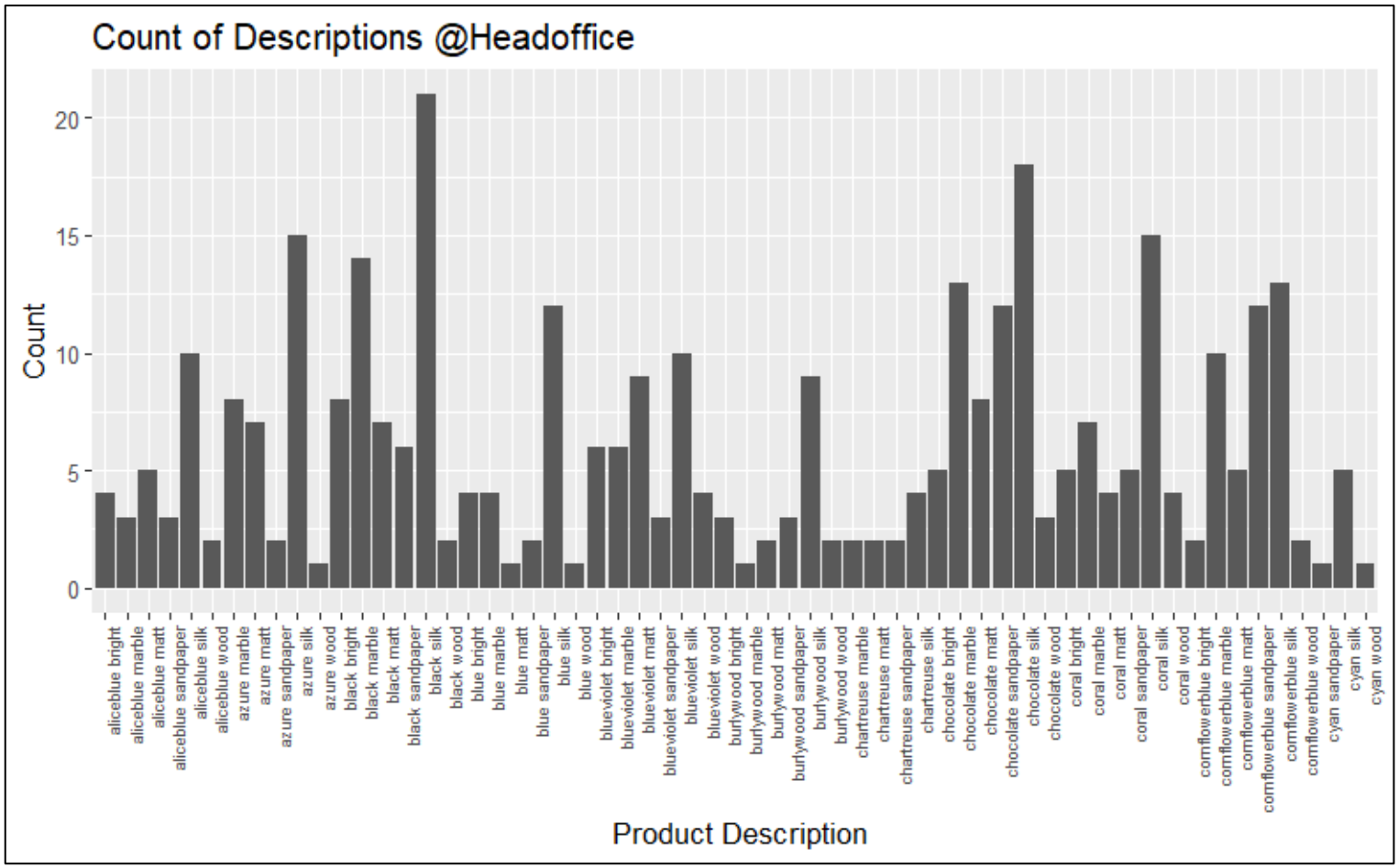
The average markup per product ID is also plotted. The products are plotted in order of increasing average markup. We can see that LAP008 has the largest markup and MOU001 the smallest mark-up. There is no visible relationship between category type and markup.



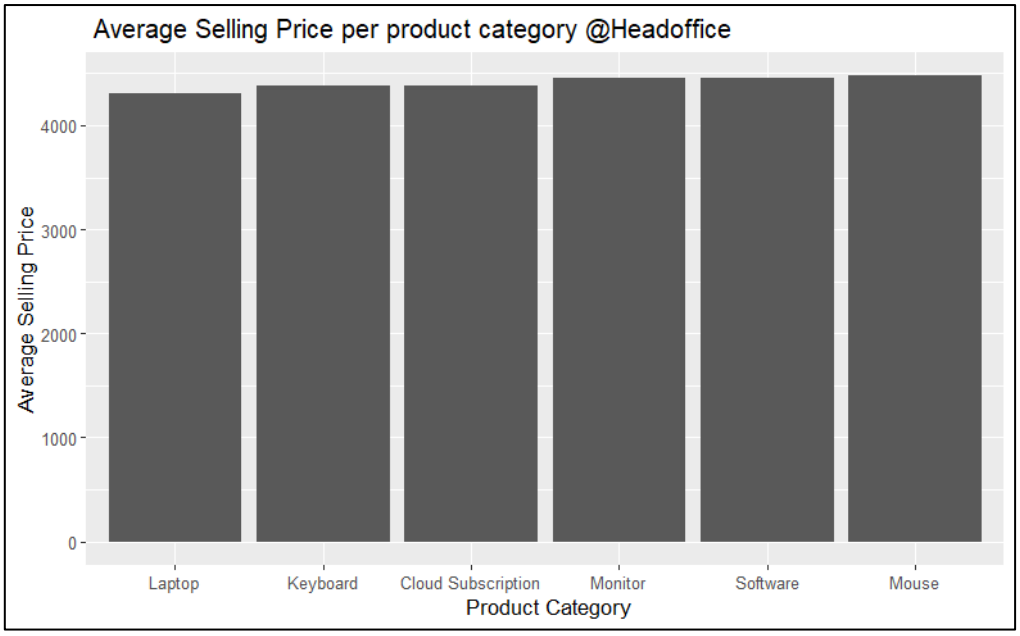
The below plot shows that there are 60 product IDs per category. This means that there are 60 kinds of each category sold by the company.



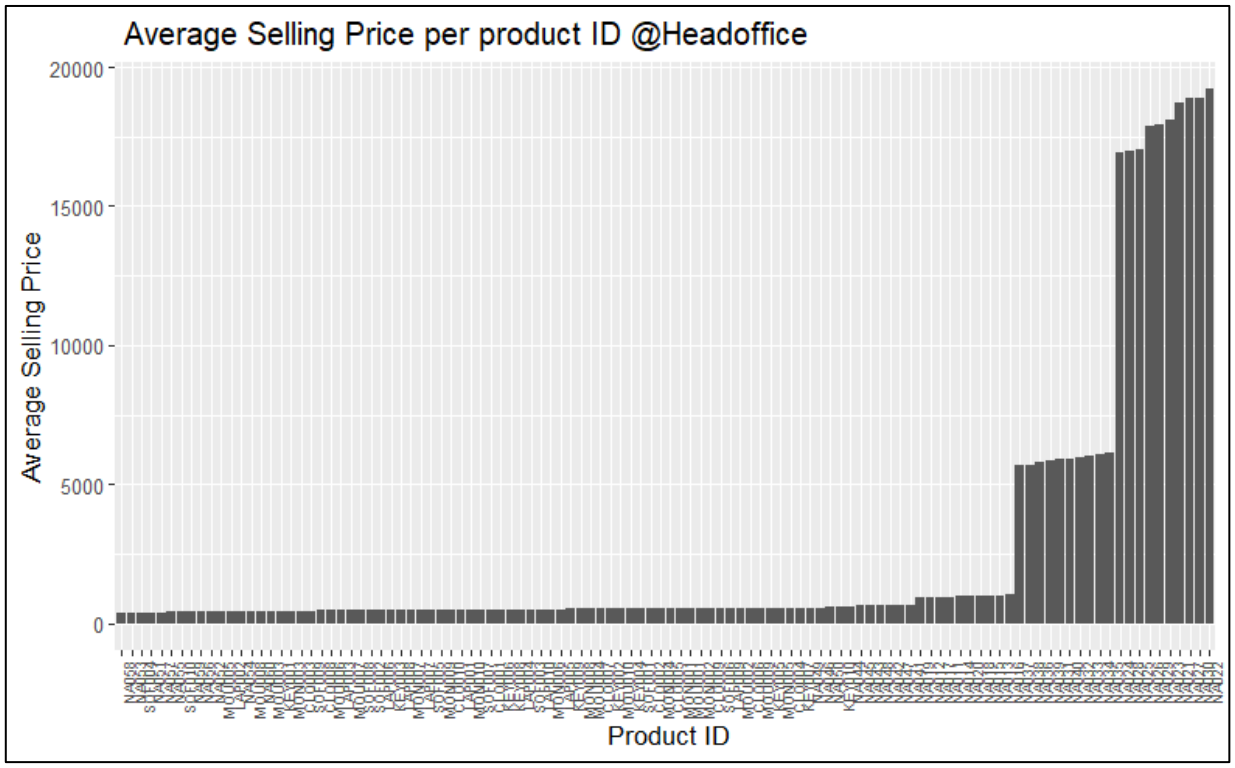
The next plot shows the count of each product description. Black silk is the most common description for a product sold by the company.



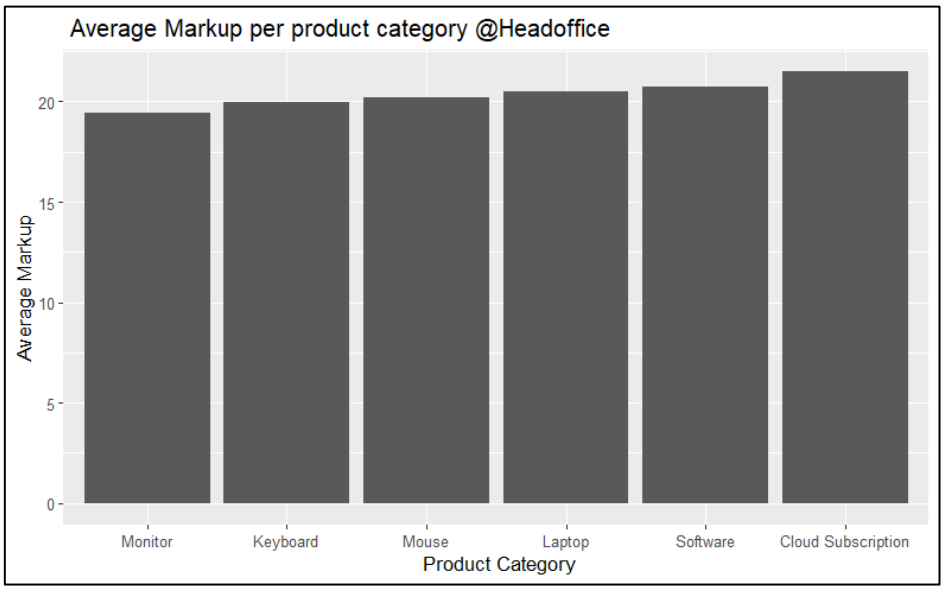
The bar graph below indicates the average selling price per category. All the prices are similar however laptops cost the least and mouse the most. This contradicts the natural assumption that a mouse would be cheaper than a laptop therefore this should be investigated. This plot contradicts the conclusions drawn from the products dataset.



The next plot shows that the NA product IDs have the highest selling price. This part of the data analysis should be repeated once the head office product data is corrected. This will be covered in Section 4. From the above bar graph, we suspect the NAs with the highest selling price will be Mouse products.

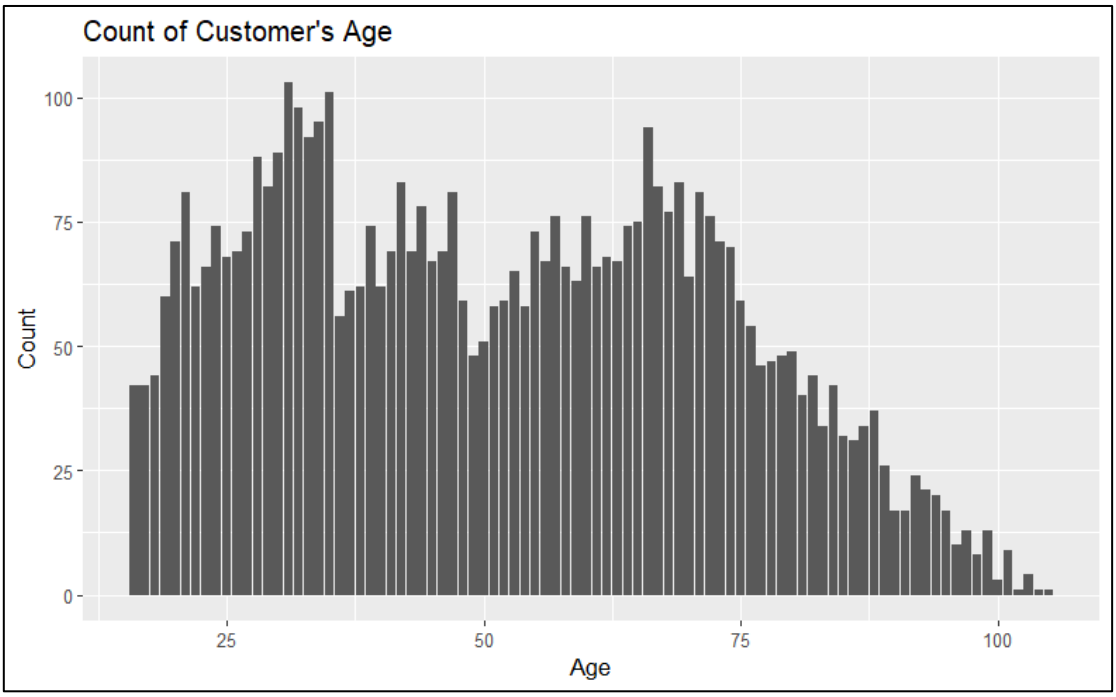


The bar graph below shows the average markup per product category. The cloud subscription has the biggest markup.

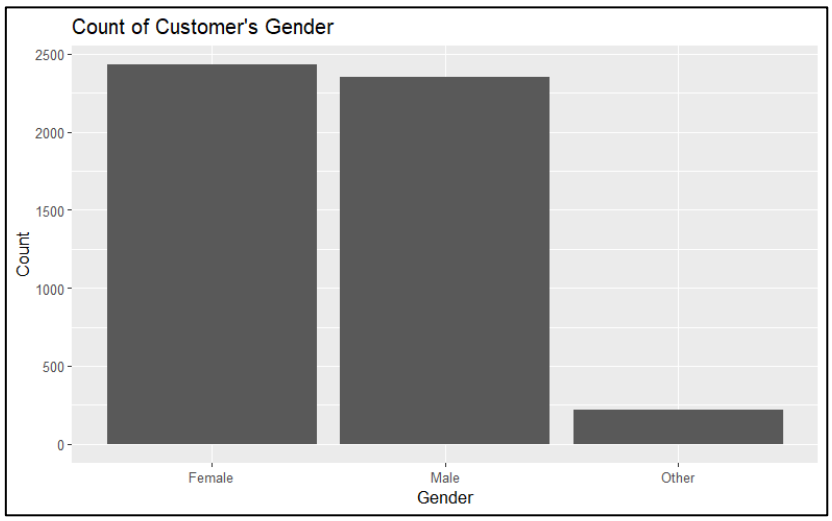


The next dataset that will be visualized is the **customer data**. This is arguably the most important data as understanding the target market of the company can make the biggest difference in sales.

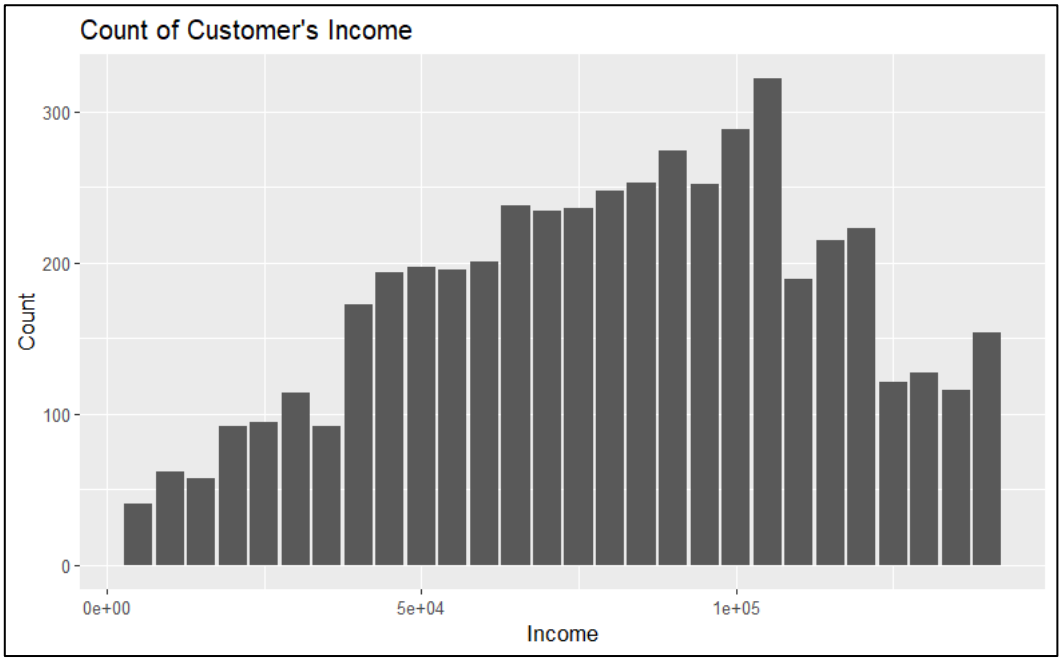
The plot below shows the age distribution of customers. This is useful for the marketing department to determine their target market. Advertising will increase sales if the advertisement is geared towards the correct age groups. It can clearly be seen that people of age 30 are very likely to buy from this company compared to 90-year-olds. This is expected as the company sells technology, often used by the young adults to middle aged groups. It is seen however that every age buys technology which correlates to the technology driven world of today. Below the multi-modal distribution is seen.



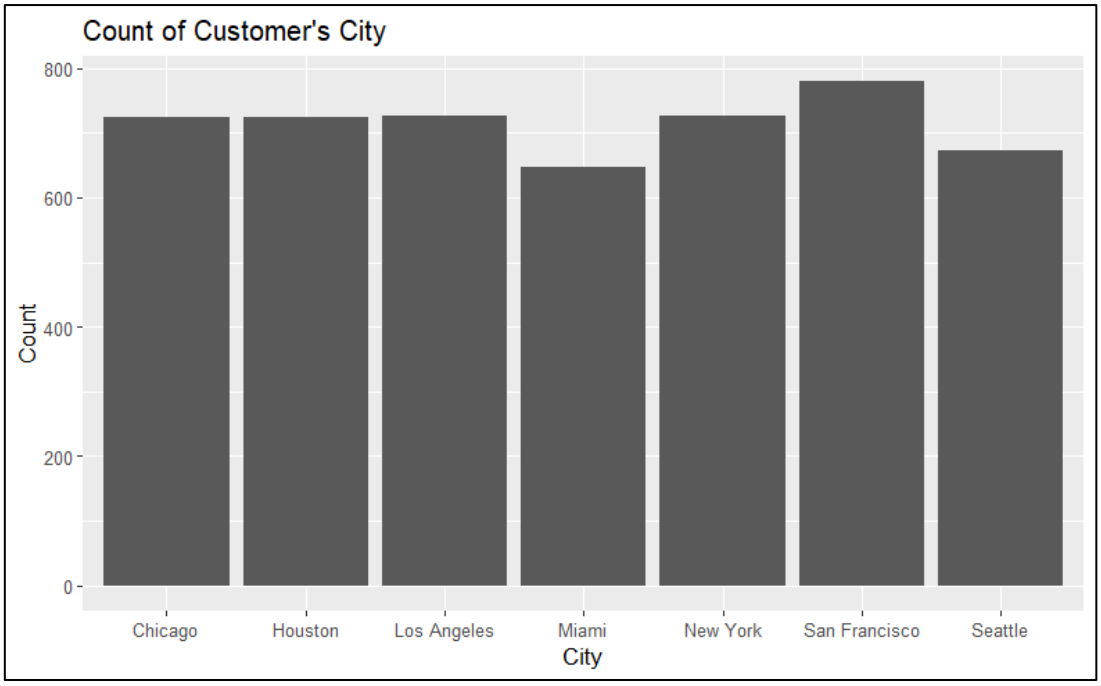
The gender of the customer is relatively evenly distributed between male and female.



The next plot shows that as a customer's income increases, the more likely they are to buy from the technology company. This makes logical sense as income often dictates the ability to buy luxury items like laptops. A rough uniform distribution is seen.

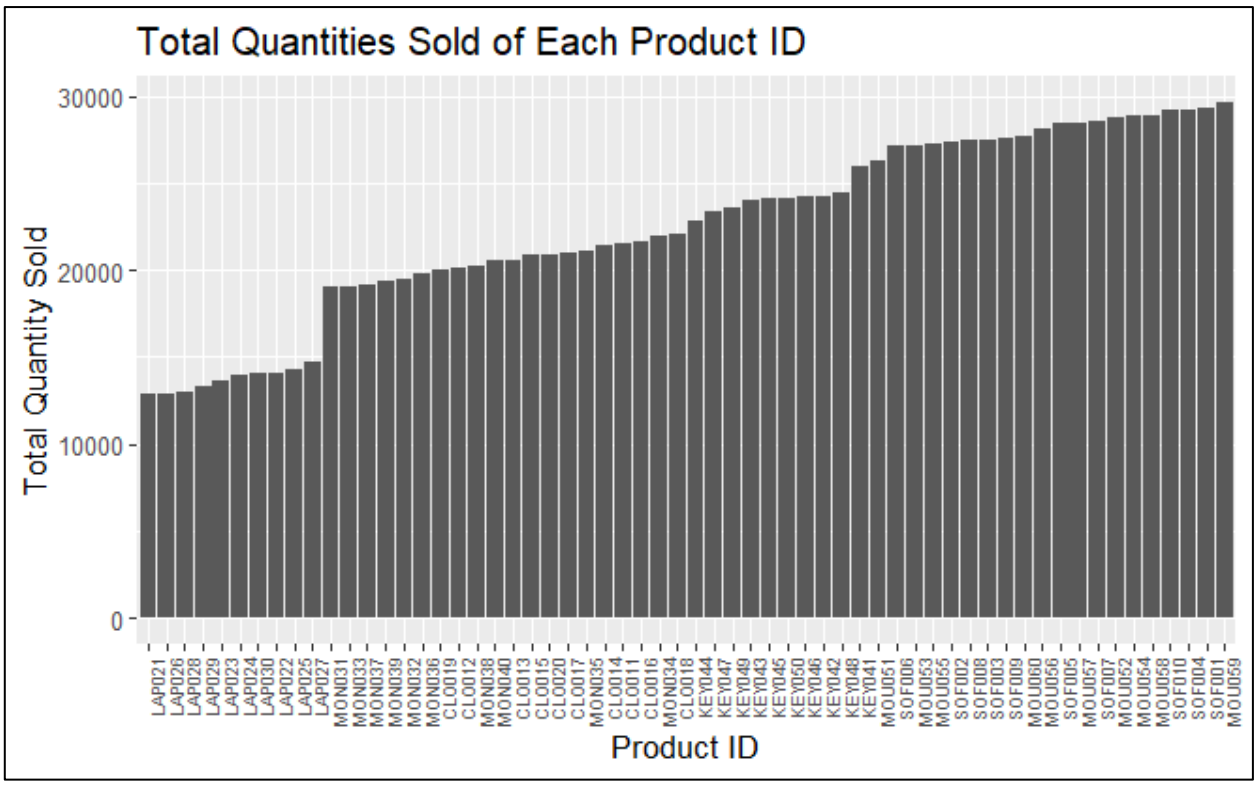


The bar graph below agrees with the previous conclusion, in Section 2.4., which revealed that Miami has the least sales and San Francisco the most.



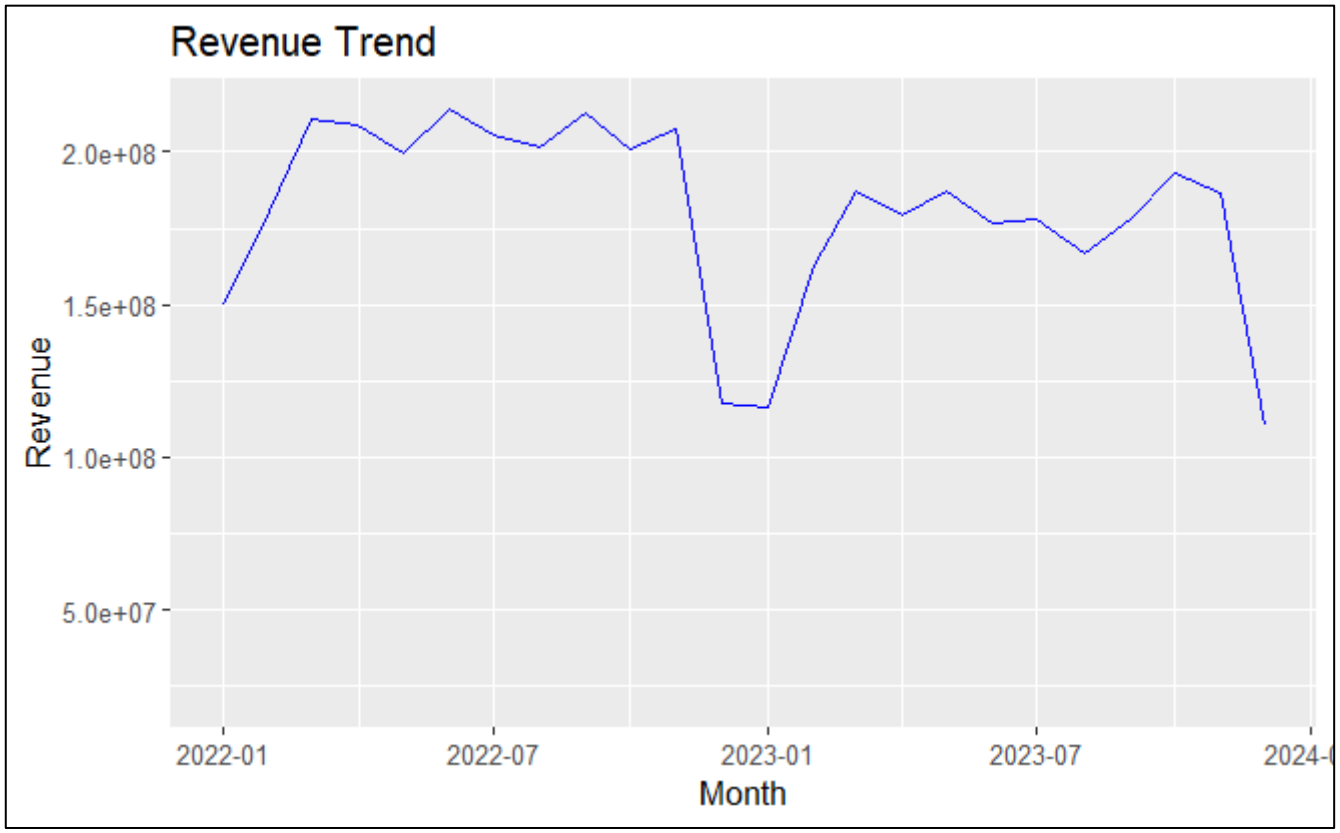
Now that the individual datasets have been visualized, some plots will be used to show derived features. The first derived feature sums the quantity sold from the sales data per product ID. The product IDs are first ranked from most sold to least number sold. The MOU059 it the product for which the most amount has been ordered and sold. Below is a table indicating the top ten products according to total quantity sold. Below that the new dataset is plotted.

	product_id	total_sold
1	MOU059	29675
2	SOF001	29336
3	SOF004	29219
4	SOF010	29168
5	MOU058	28924
6	MOU054	28875
7	MOU052	28804
8	SOF007	28517
9	MOU057	28423
10	SOF005	28412



The datasets are then combined using R functions and packages. The sales, customer and product data are joined. In the figure below the price per order for every customer order is placed. The x-axis represents each of the 100 000 orders placed at the company for 2022 and 2023. There is a random distribution however we do see a dense area at the lower order cost. This indicates that there are cheaper than expensive orders placed.

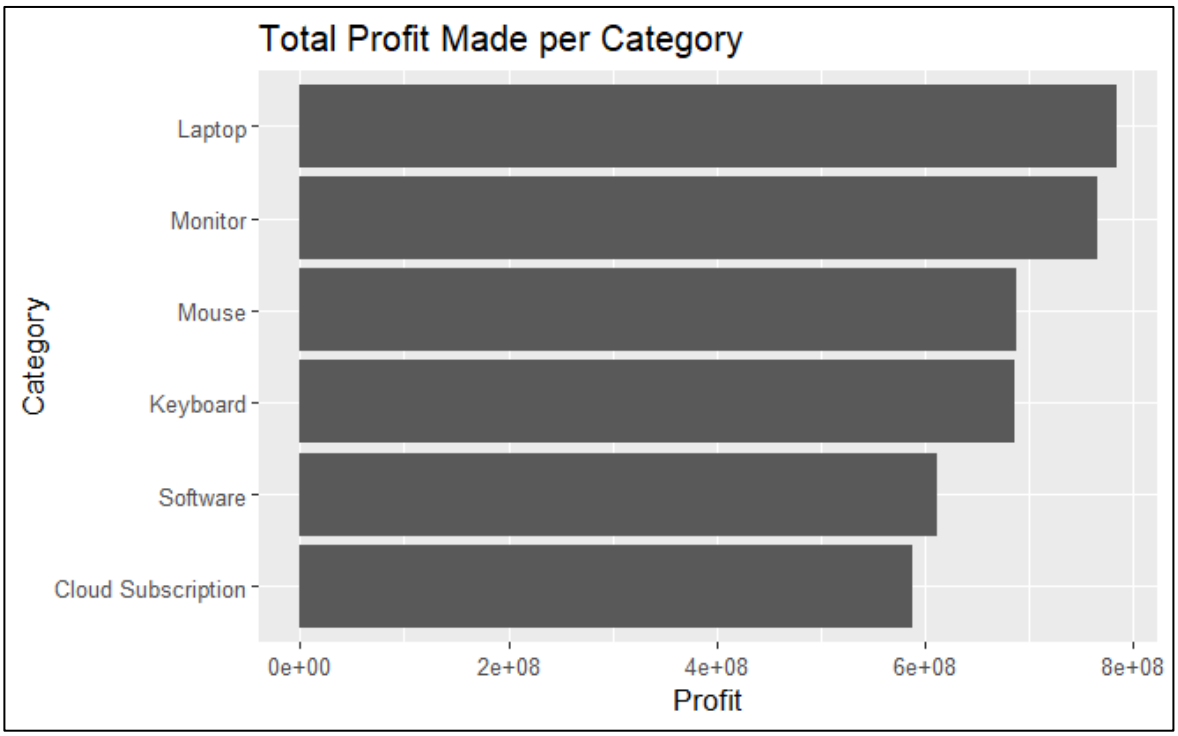
The next plots is composed of the revenue generated for the company across time using order dates and summed sales data. It plots time versus revenue in order to show the trend of monthly sales. It can be seen that there is an extreme dip in the period of December 2022 and January 2023. This should be investigated to prevent future declines. It is however seen that every year around this time the sales drastically decrease indicating a seasonal trend. This is important to allow production to compensate for this sudden decrease in demand and revenue.



As discussed in section 2.2. of this report, laptops generate the most revenue for the company. The company should therefore take the most care with this products' forecasting and inventory management to ensure its most profitable product is always available.

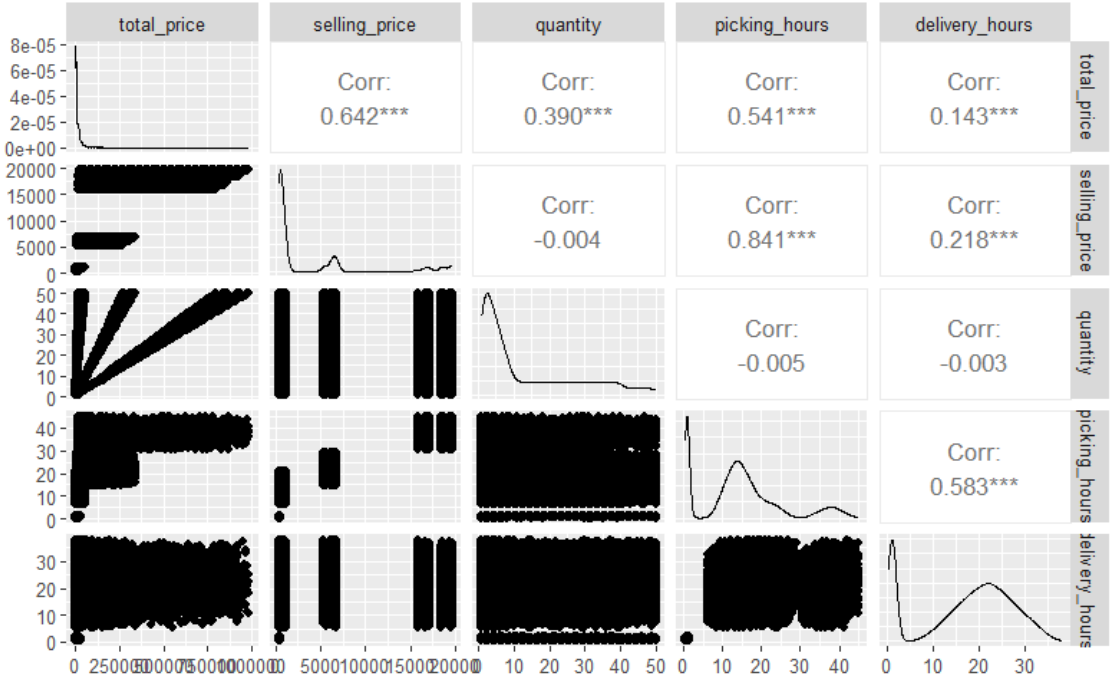
	category	total_rev
1	Laptop	821533851
2	Monitor	809104952
3	Keyboard	723693159
4	Mouse	721090260
5	Software	655365933
6	Cloud Subscription	621799523

After the total revenue per category is determined, markup is used to calculate how much profit the company actually made per product category. This can be seen below. The laptop and then monitor generate the most profit for the company and cloud subscriptions the least.



2.6. Exploring Relationships

The above SPLOM (scatterplot matrix) created in R summarizes the relationship between all the features in the data set. It can be seen that features like selling price and total price are strongly correlated while quantity and picking hours have no correlation. This information is useful for the analysts in the company to suggest what decisions impact which features.

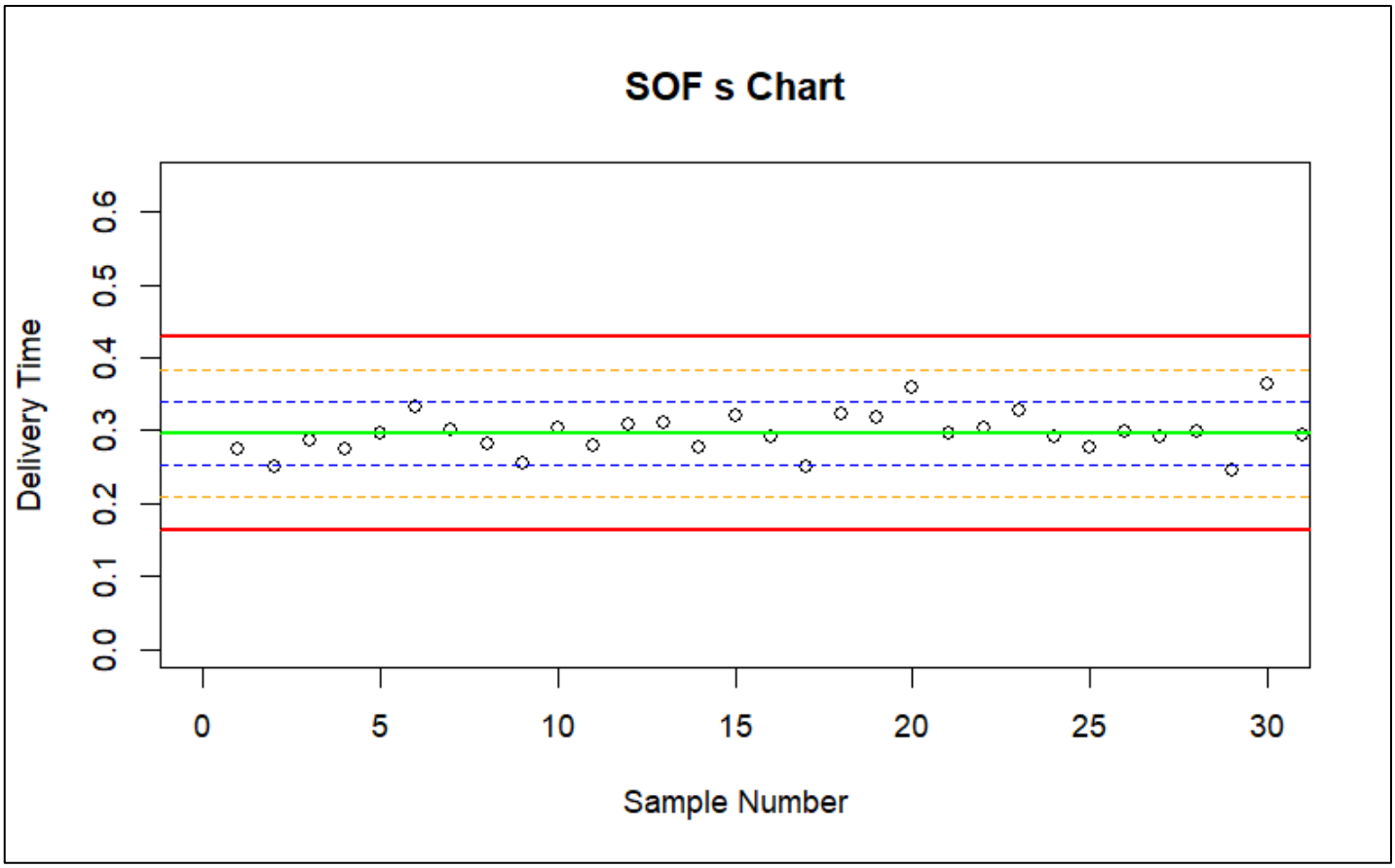


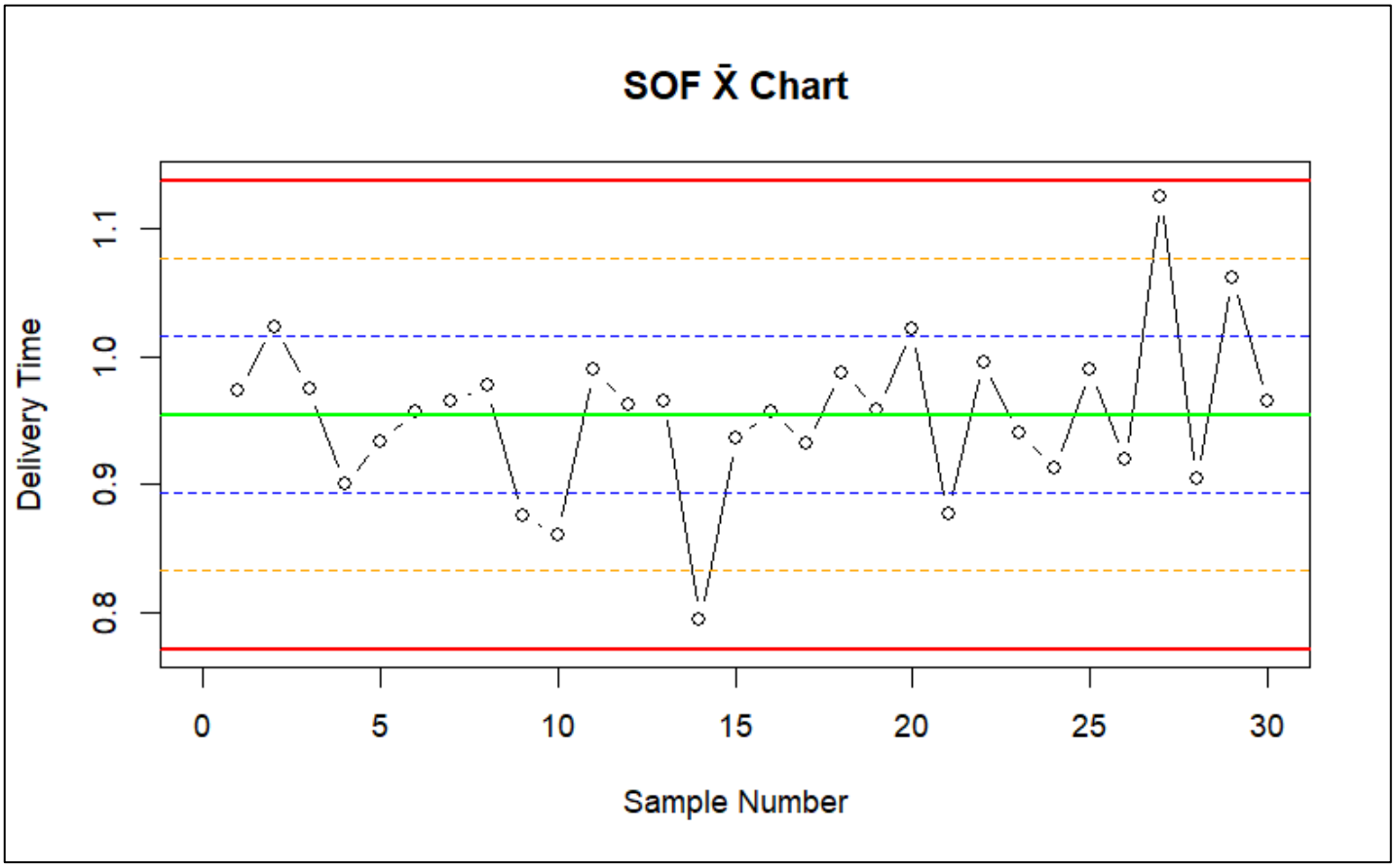
3. Statistical Process Control

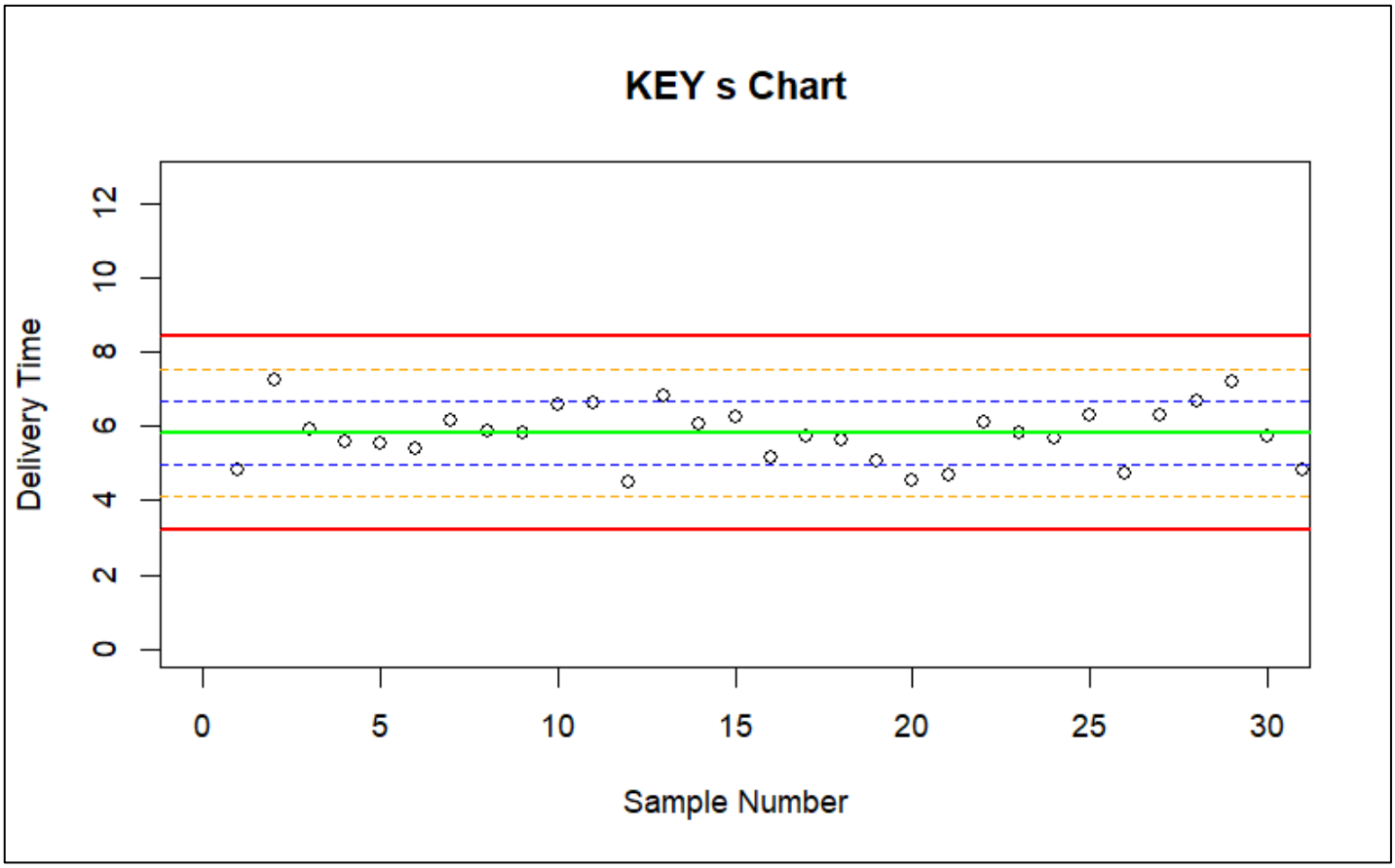
R is used to do statistical process control. X-s charts are generated using the "sales2026and2027Future.csv" file. The data is ordered and divided into subsets. In this report samples of 24 are used. The first 30 samples per process is used and then the rest of the samples are run in an accelerated simulation in Section 3.2 and 4.

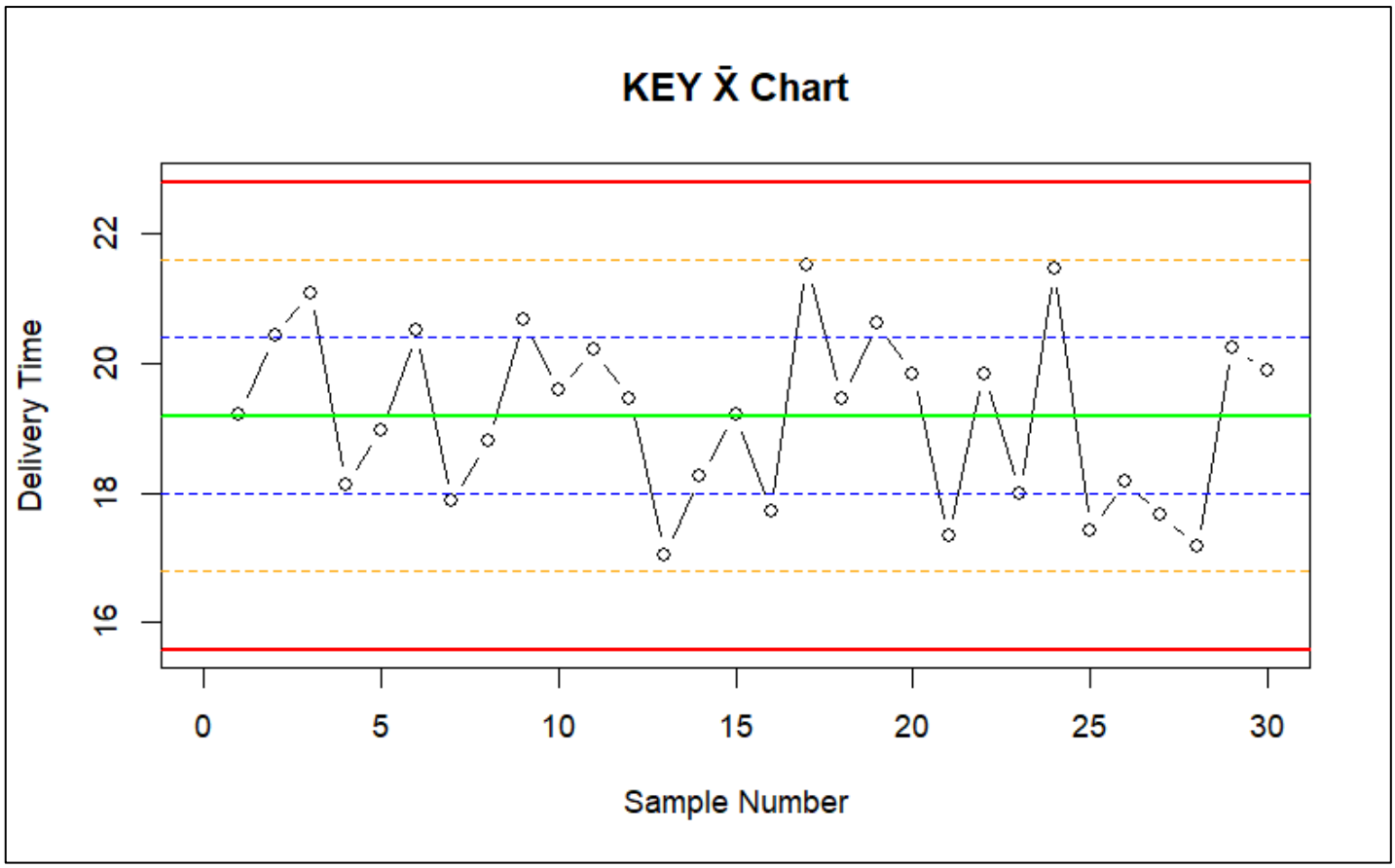
3.1. Control Charts Per Product Type for Delivery Times

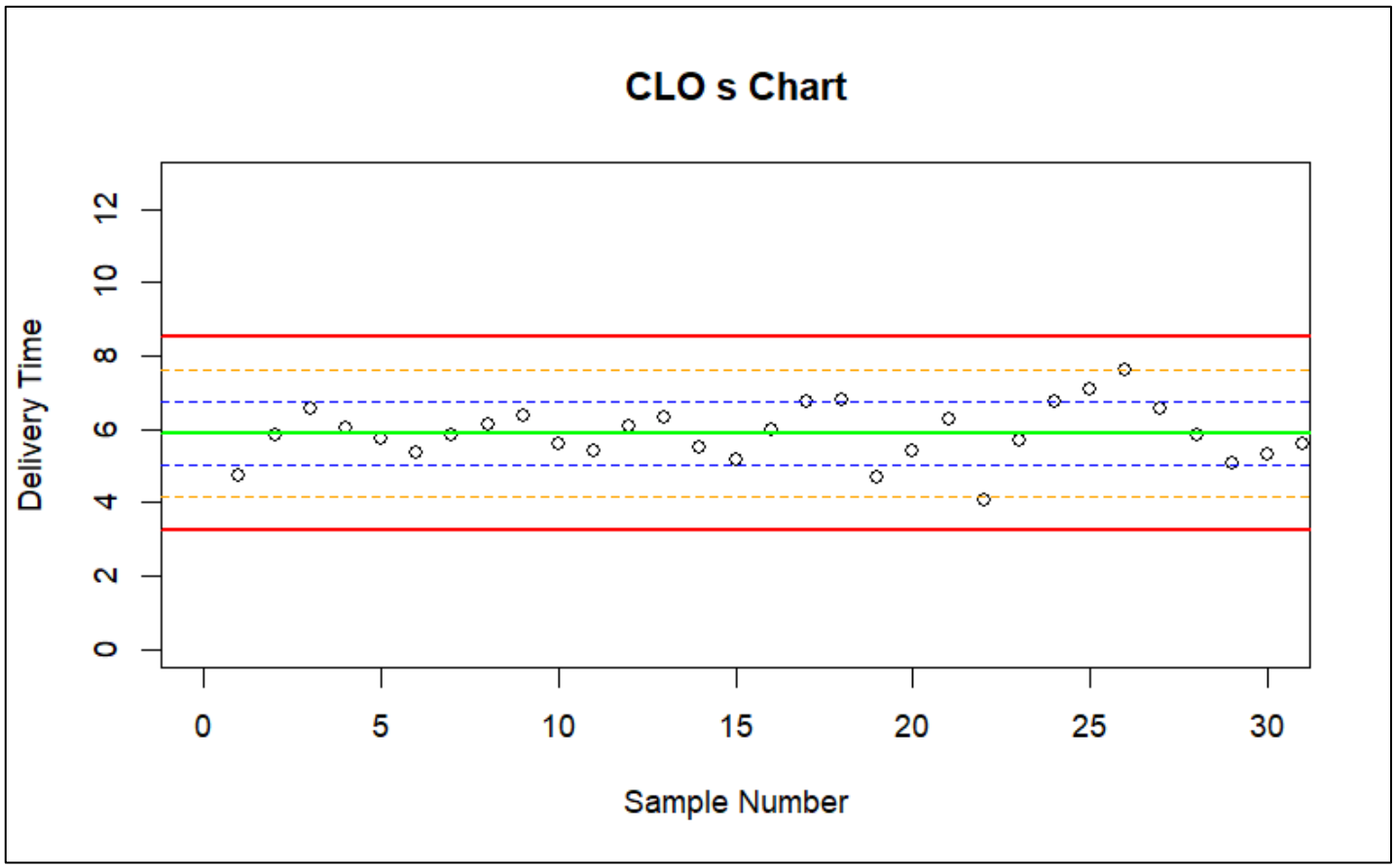
Control charts are generated in R for delivery times for every product type. 30 samples of size 24 (oldest data used first) are used to determine the centre lines (green line), outer control limits (red lines), 2-sigma control limits (yellow lines) and the 1-sigma control limits (blue lines) for each chart. Due to this, the points plotted fall within the limits indicating a good and in-control process. The s charts show the standard deviation while the X bar charts show the mean of each sample. This is done to initialize the statistical process control.

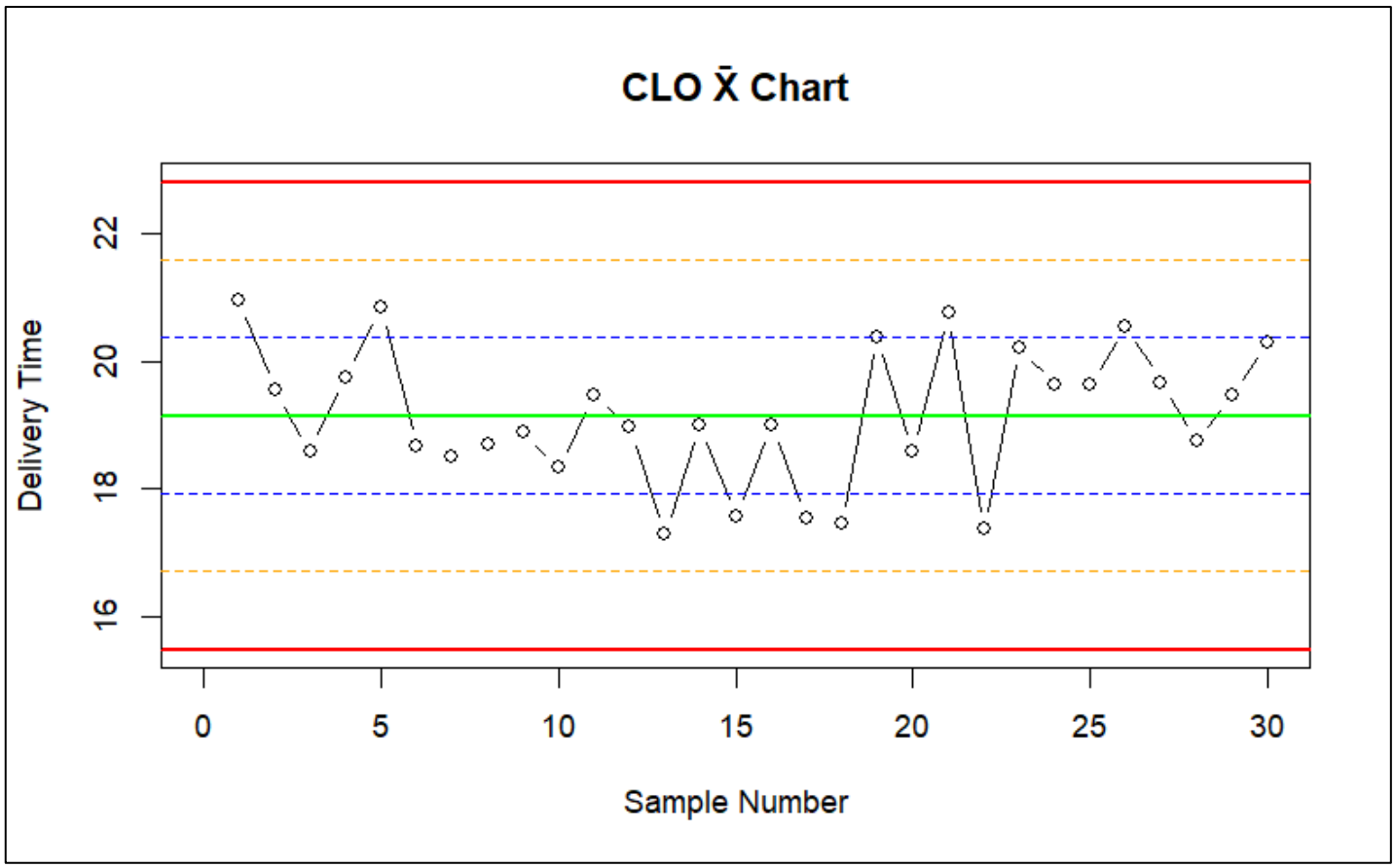


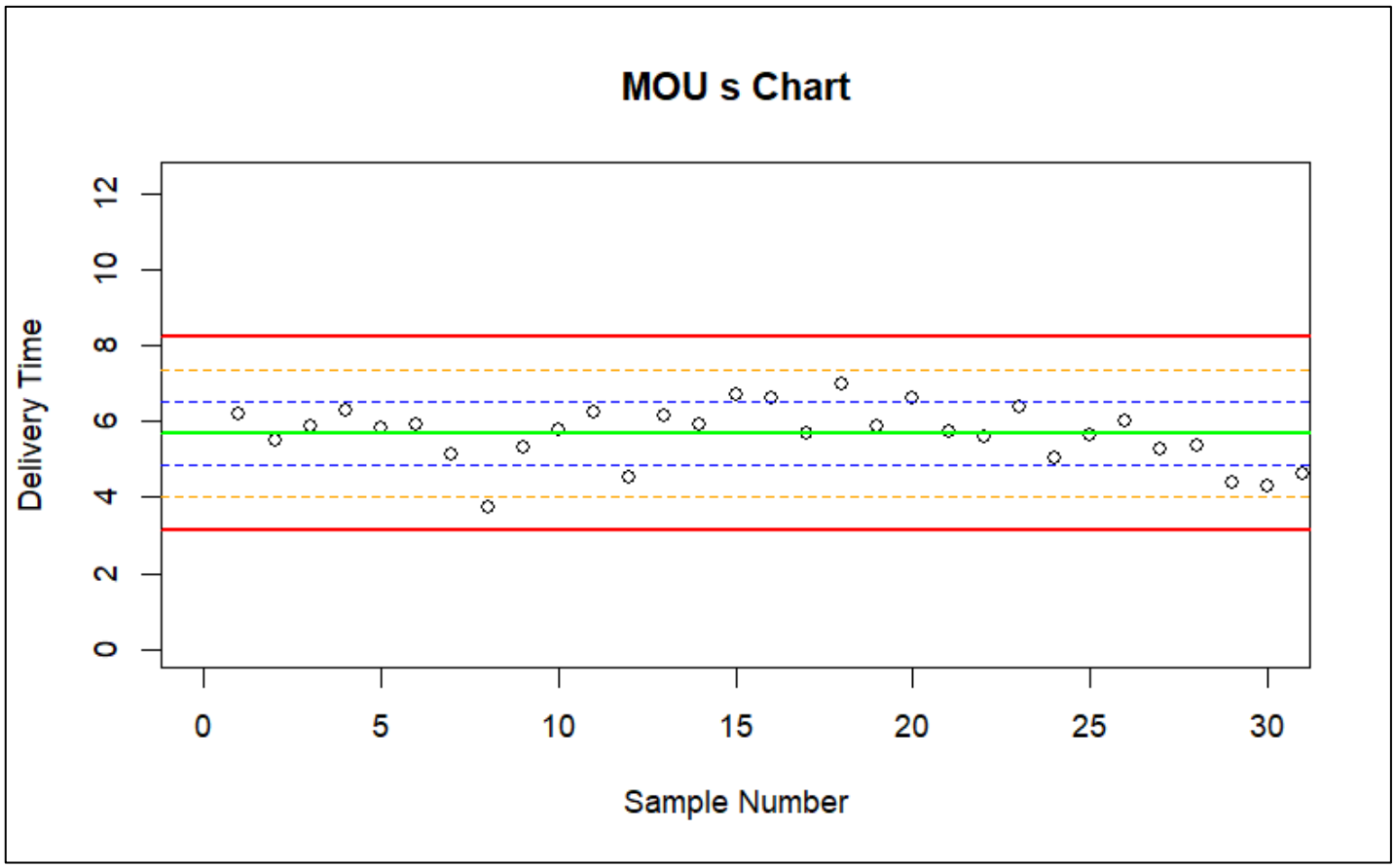

SOF X̄ Chart
Delivery Time
0.8
0.9
1.0
1.1
0
5
10
15
20
25
30
Sample Number

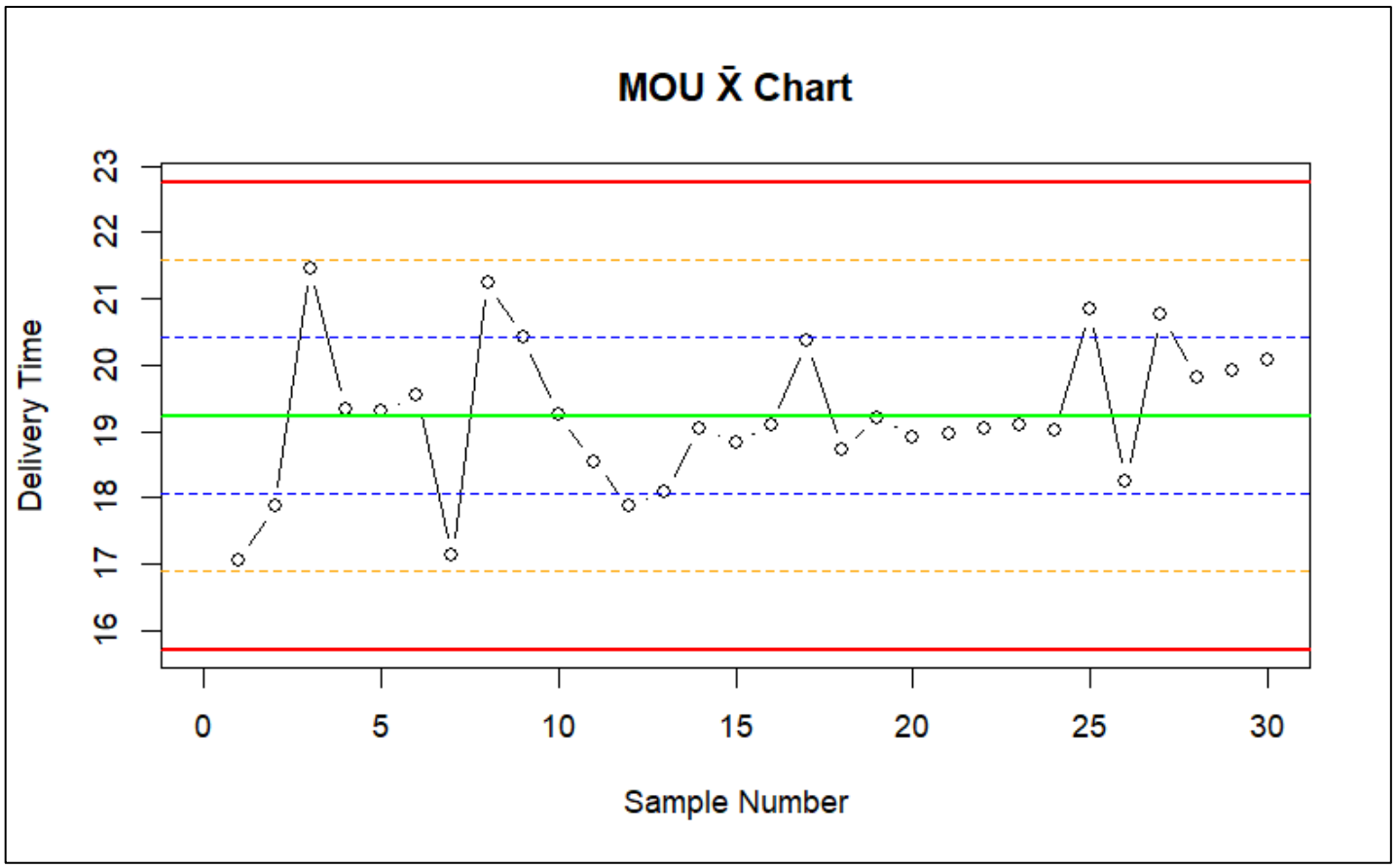

KEY s Chart
Delivery Time
0
2
4
6
8
10
12
0
5
10
15
20
25
30
Sample Number

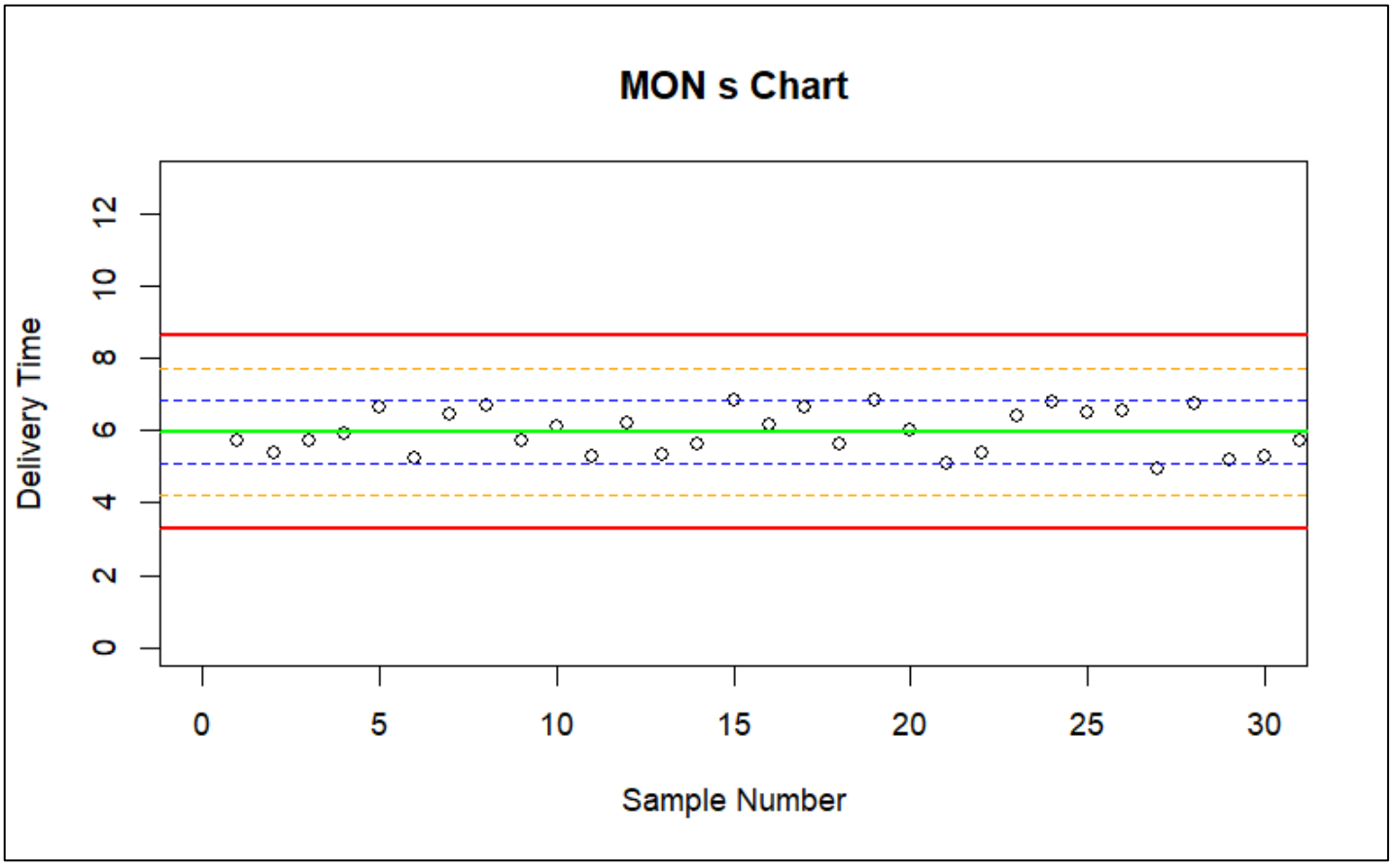

KEY X̄ Chart
Delivery Time
16
18
20
22
0
5
10
15
20
25
30
Sample Number

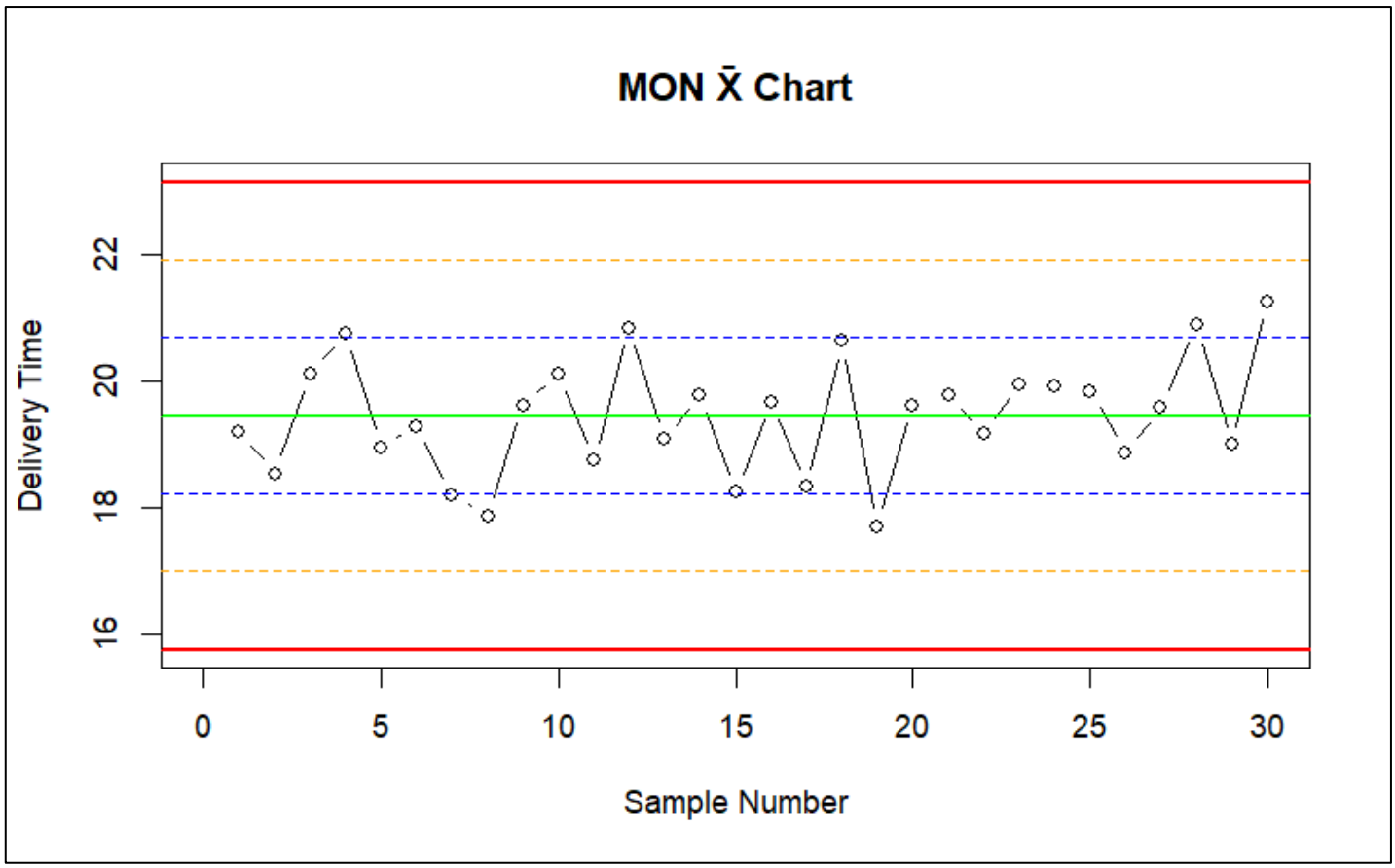

CLO s Chart
Delivery Time
0
2
4
6
8
10
12
0
5
10
15
20
25
30
Sample Number

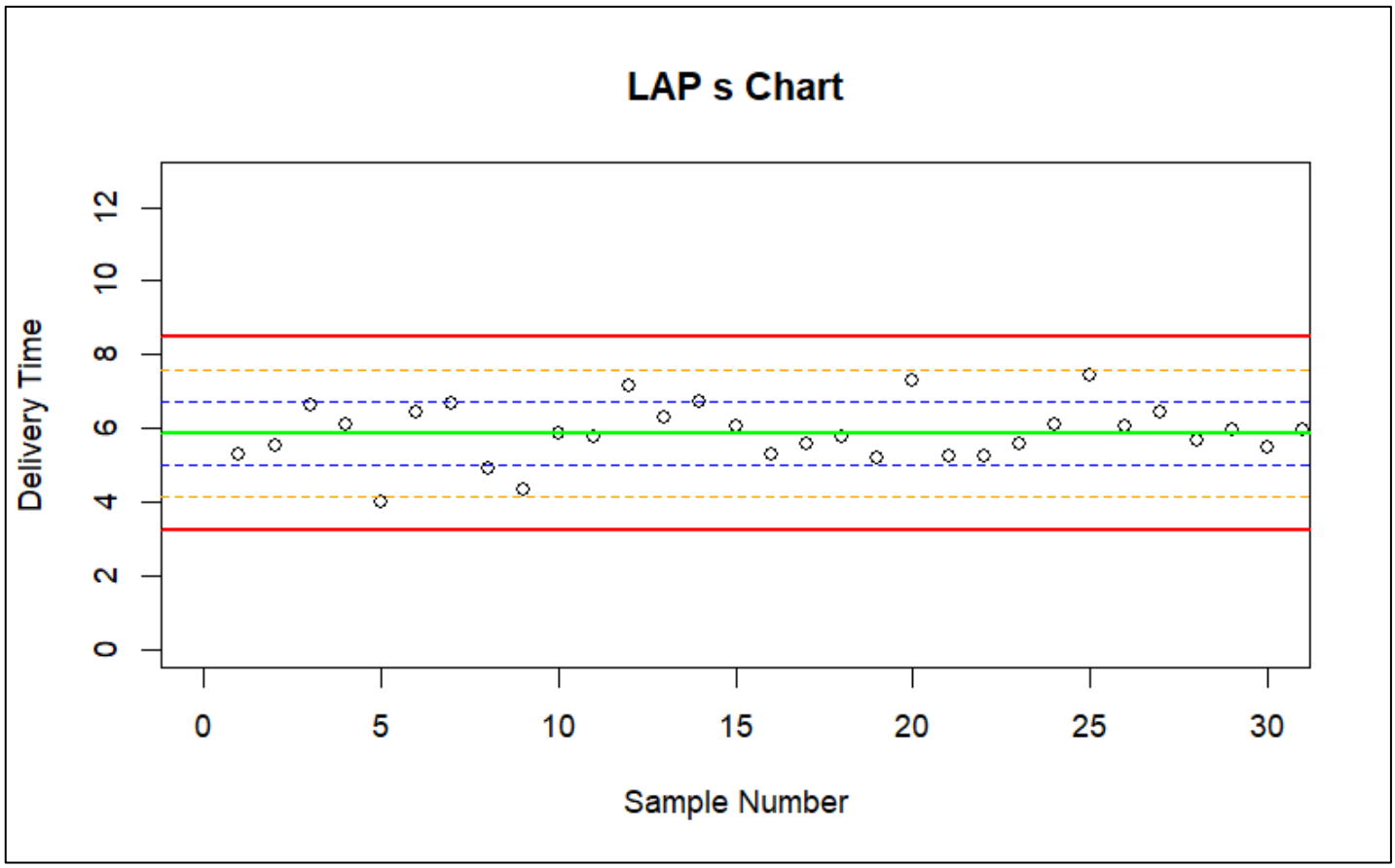

CLO X̄ Chart
Delivery Time
16
18
20
22
0
5
10
15
20
25
30
Sample Number

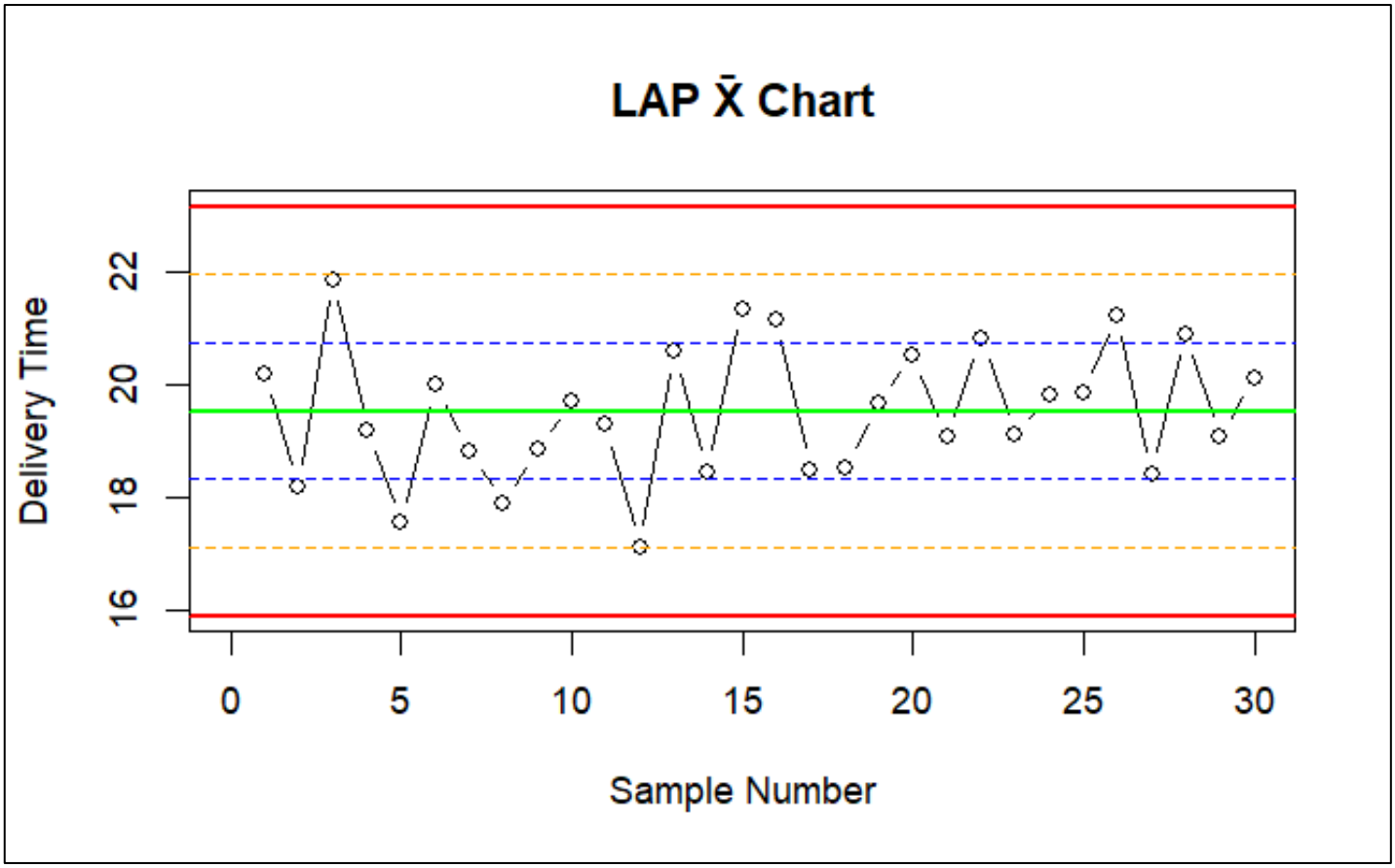

MOU s Chart
Delivery Time
0
2
4
6
8
10
12
0
5
10
15
20
25
30
Sample Number


MOU X̄ Chart
Delivery Time
16
17
18
19
20
21
22
23
0
5
10
15
20
25
30
Sample Number


MON s Chart
Delivery Time
0
2
4
6
8
10
12
0
5
10
15
20
25
30
Sample Number


MON X̄ Chart
Delivery Time
16
18
20
22
0
5
10
15
20
25
30
Sample Number

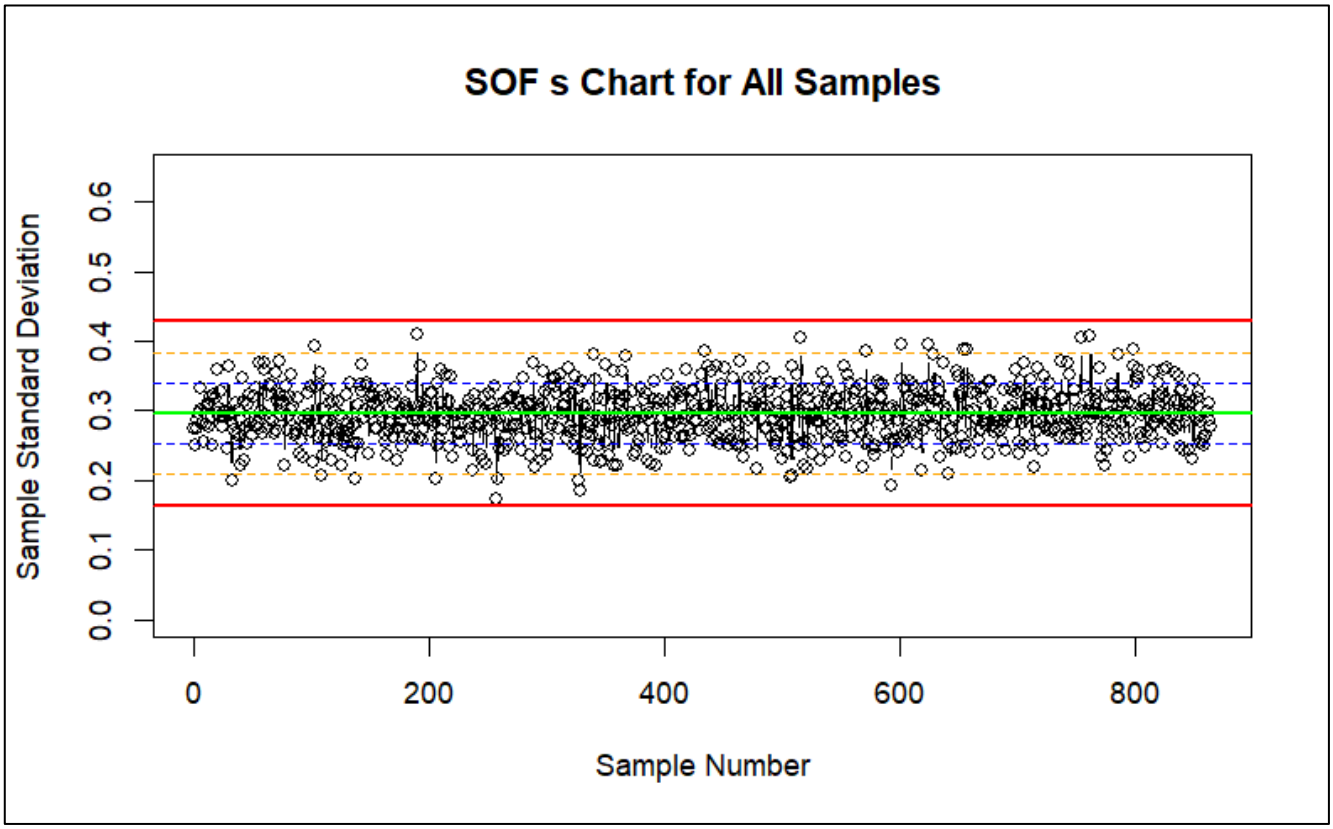

LAP s Chart
Delivery Time
0
2
4
6
8
10
12
0
5
10
15
20
25
30
Sample Number

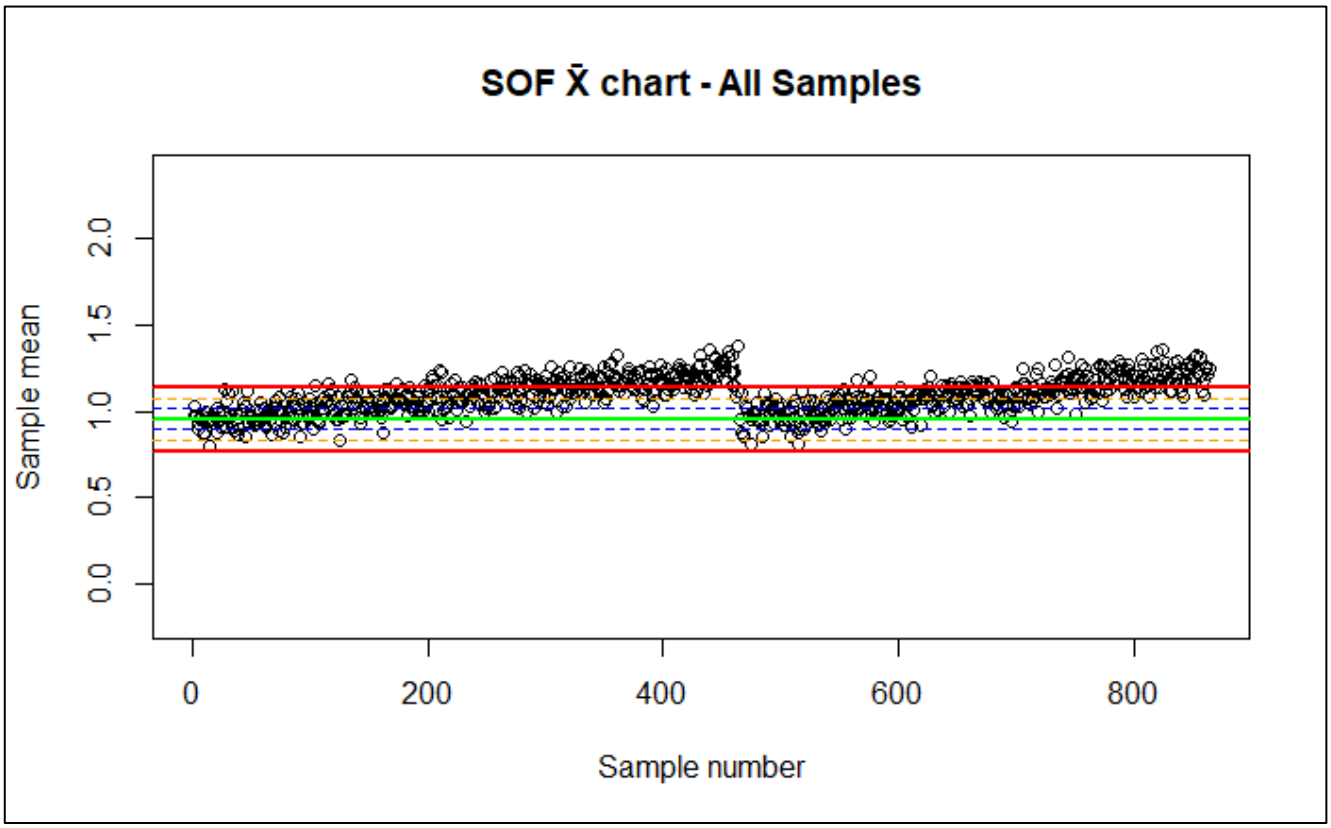


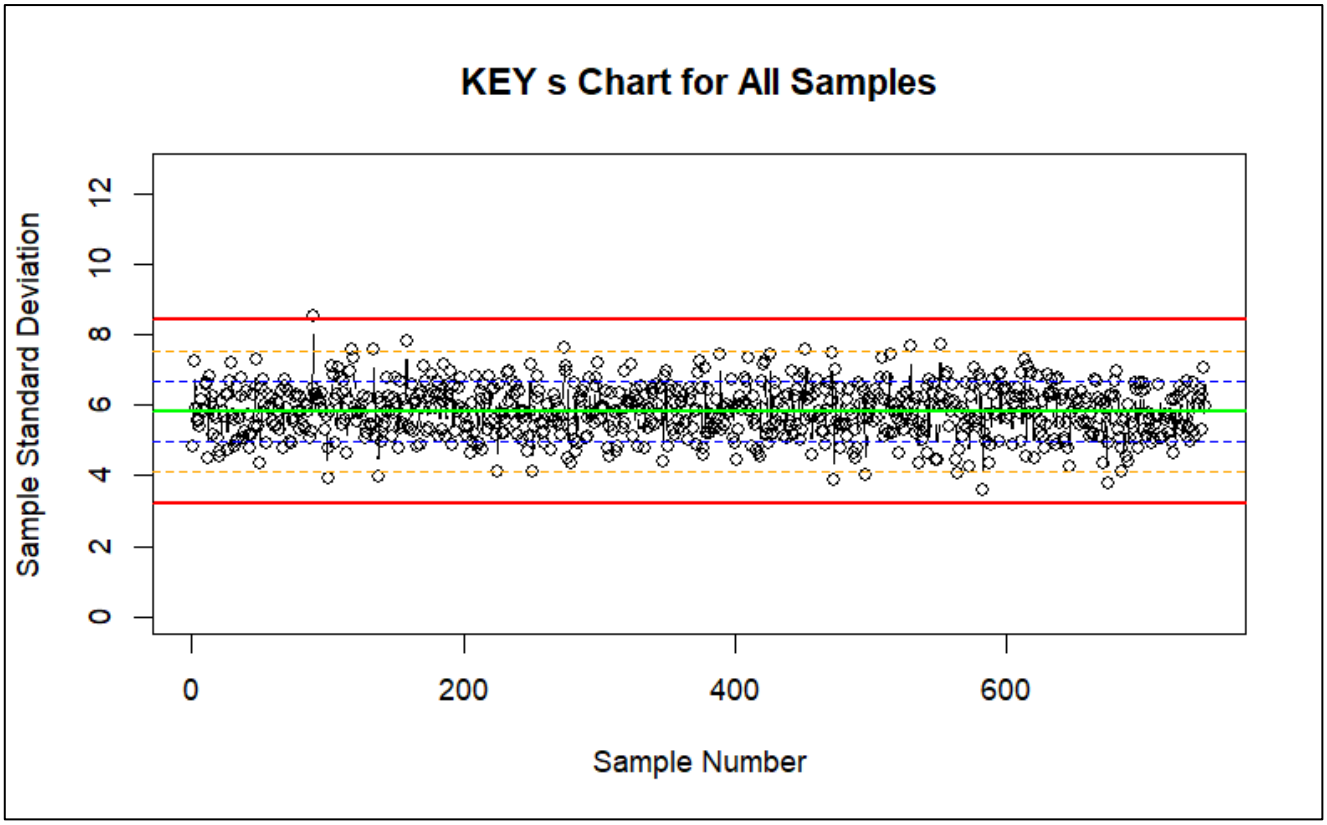
The s and x bar charts indicate in control processes as no points or samples fall outside the upper or lower control limits. There is random variation in the samples but yet the process remains stable. Now in order to discover trends and valuable insight, the samples will continue to be generated and monitored for all the data.

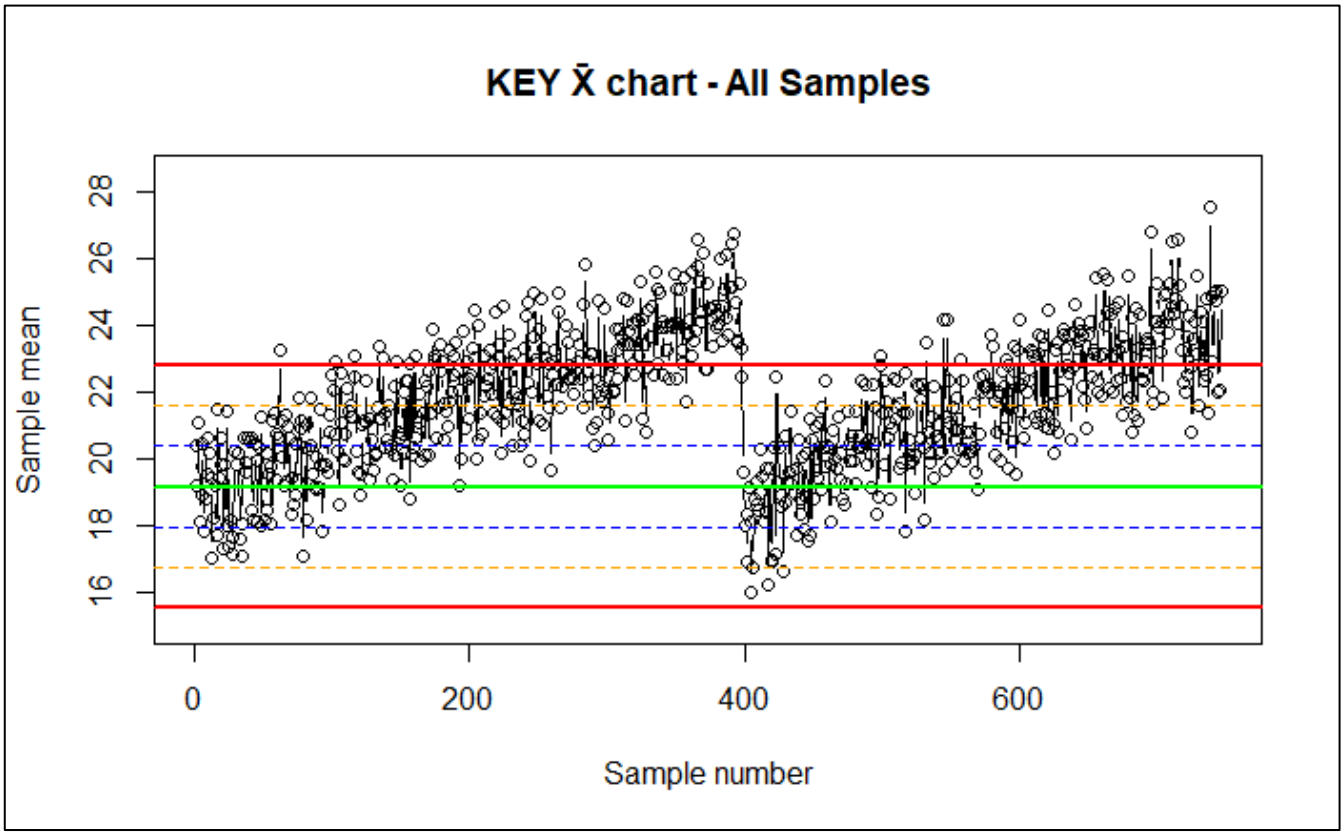
3.2. Using More Samples to Control

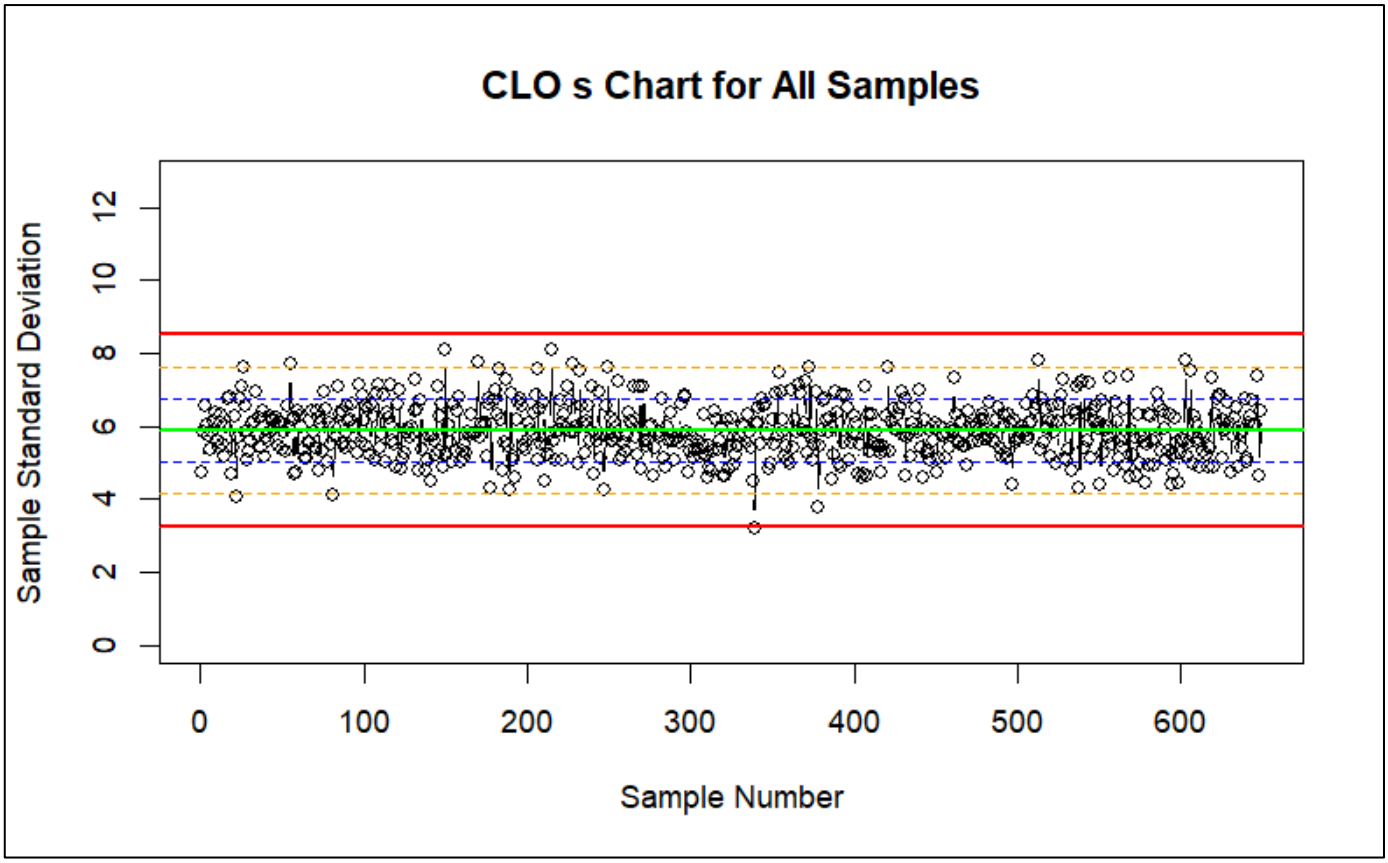
Samples of 24 are then continued to be drawn. This allows us to better understand the whole dataset with respect to delivery time for each product type. Below it can be seen that when the points plotted fall out of the control limits, it indicates the process in out-of-control.

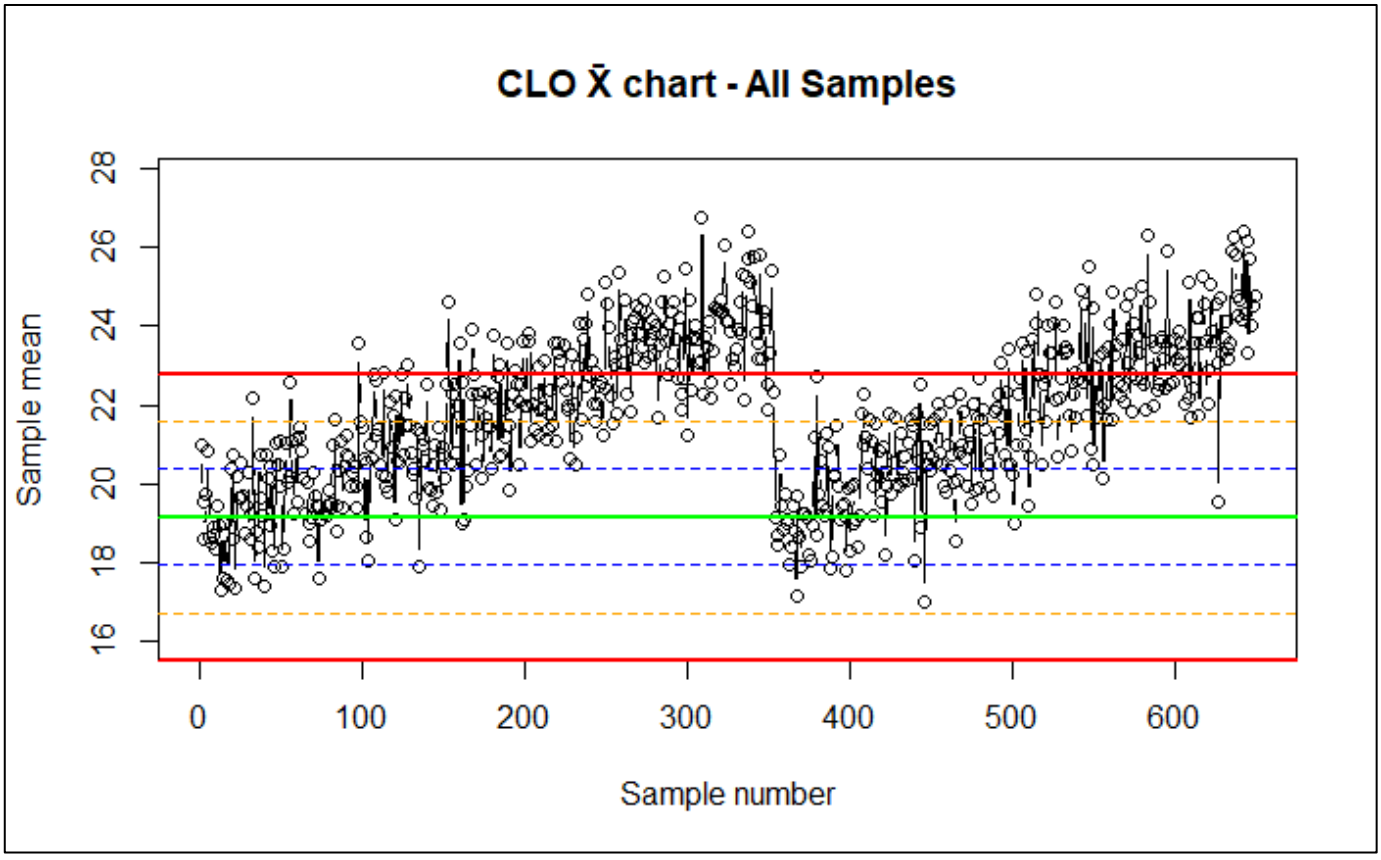


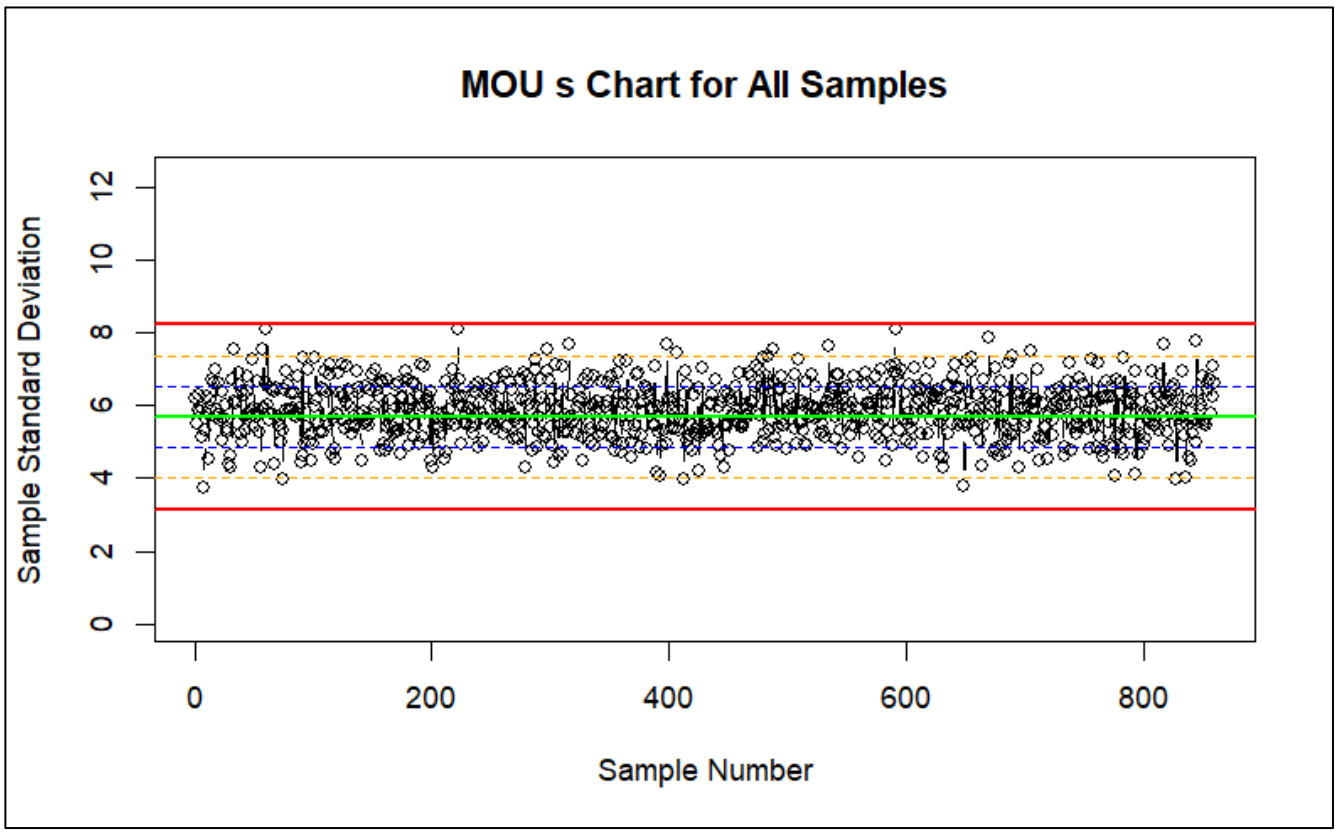


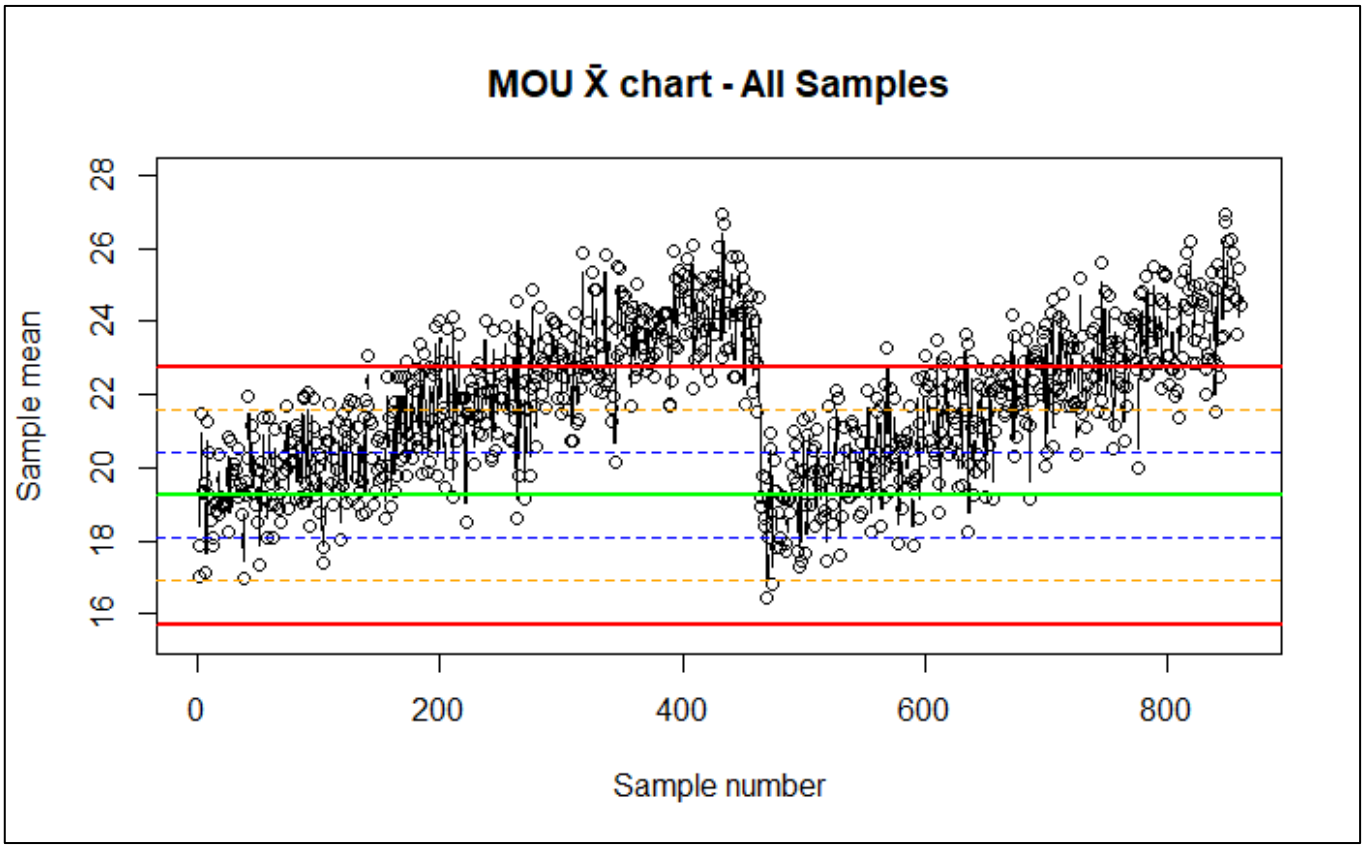

KEY s Chart for All Samples
Sample Standard Deviation
0
2
4
6
8
10
12
0
200
400
600
Sample Number

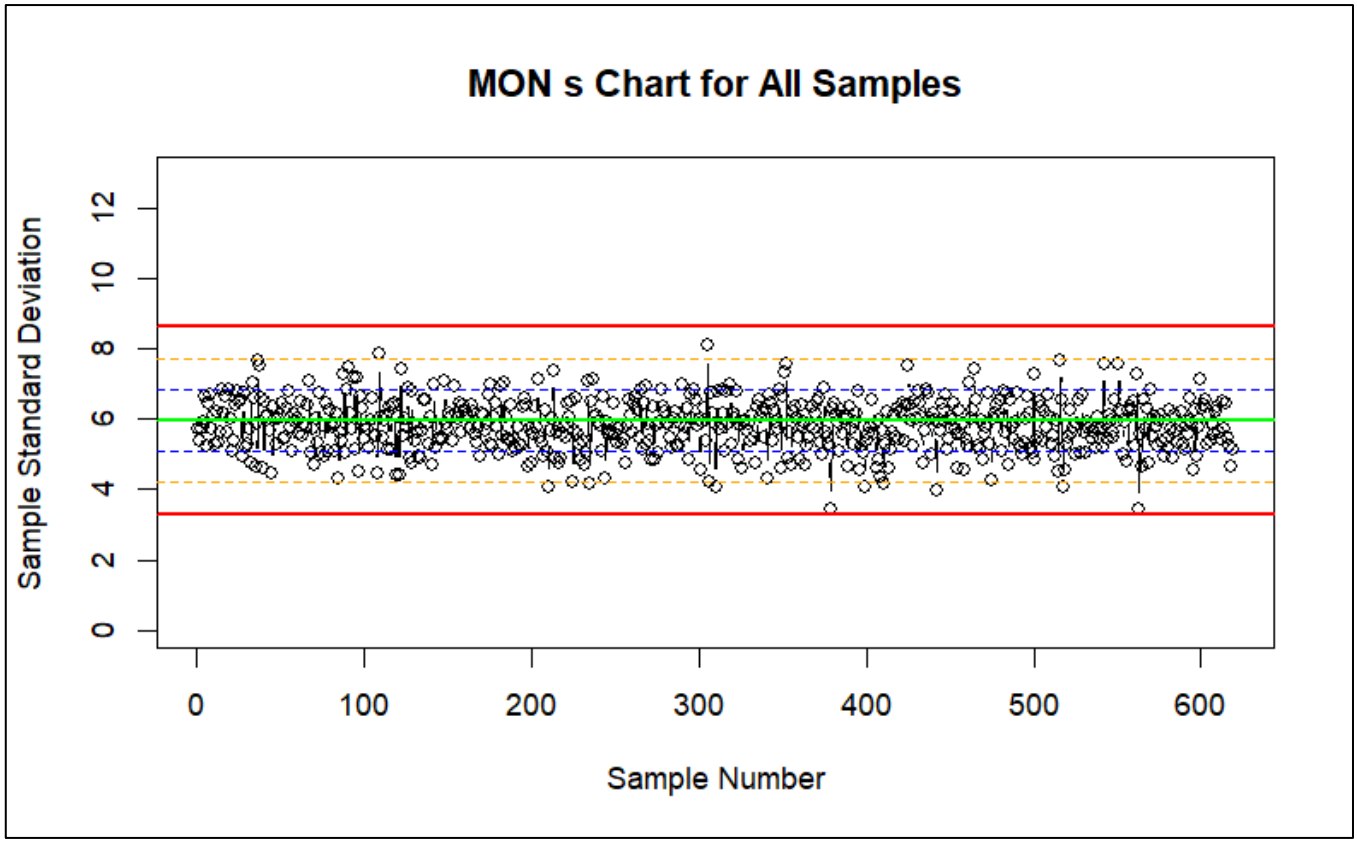

KEY X̄ chart - All Samples
Sample mean
16
18
20
22
24
26
28
0
200
400
600
Sample number

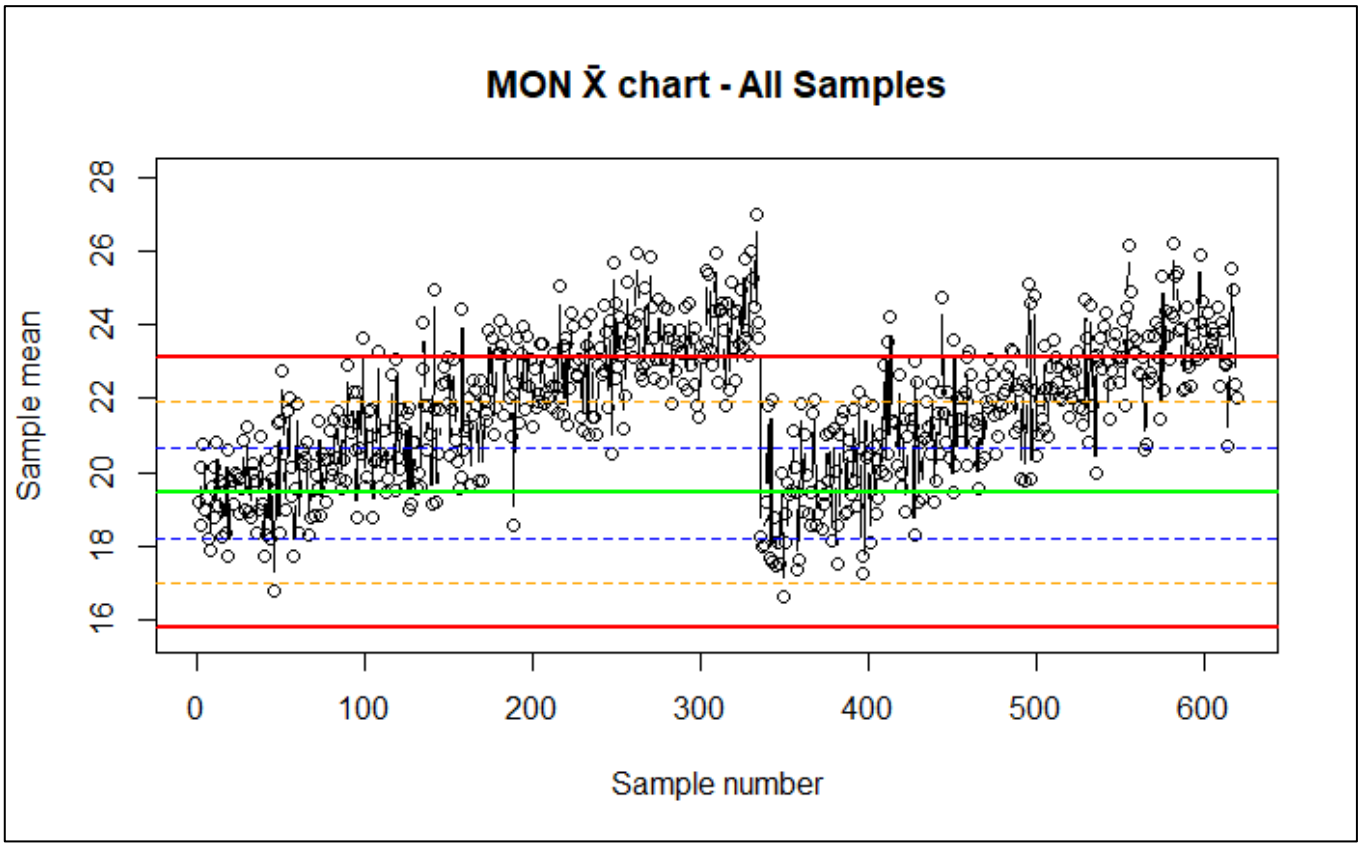

CLO s Chart for All Samples
Sample Standard Deviation
0
2
4
6
8
10
12
0
100
200
300
400
500
600
Sample Number

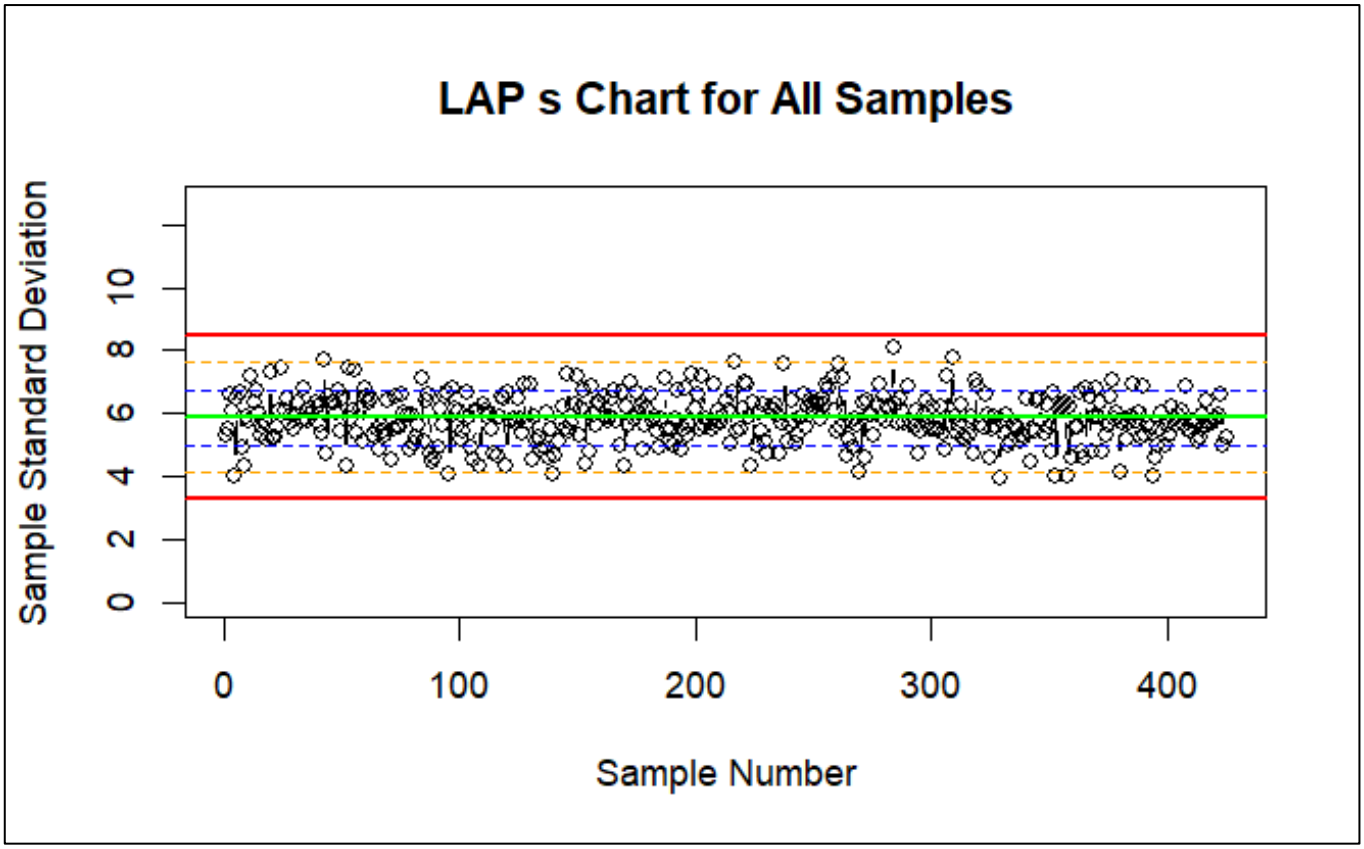

CLO X̄ chart - All Samples
Sample mean
16
18
20
22
24
26
28
0
100
200
300
400
500
600
Sample number

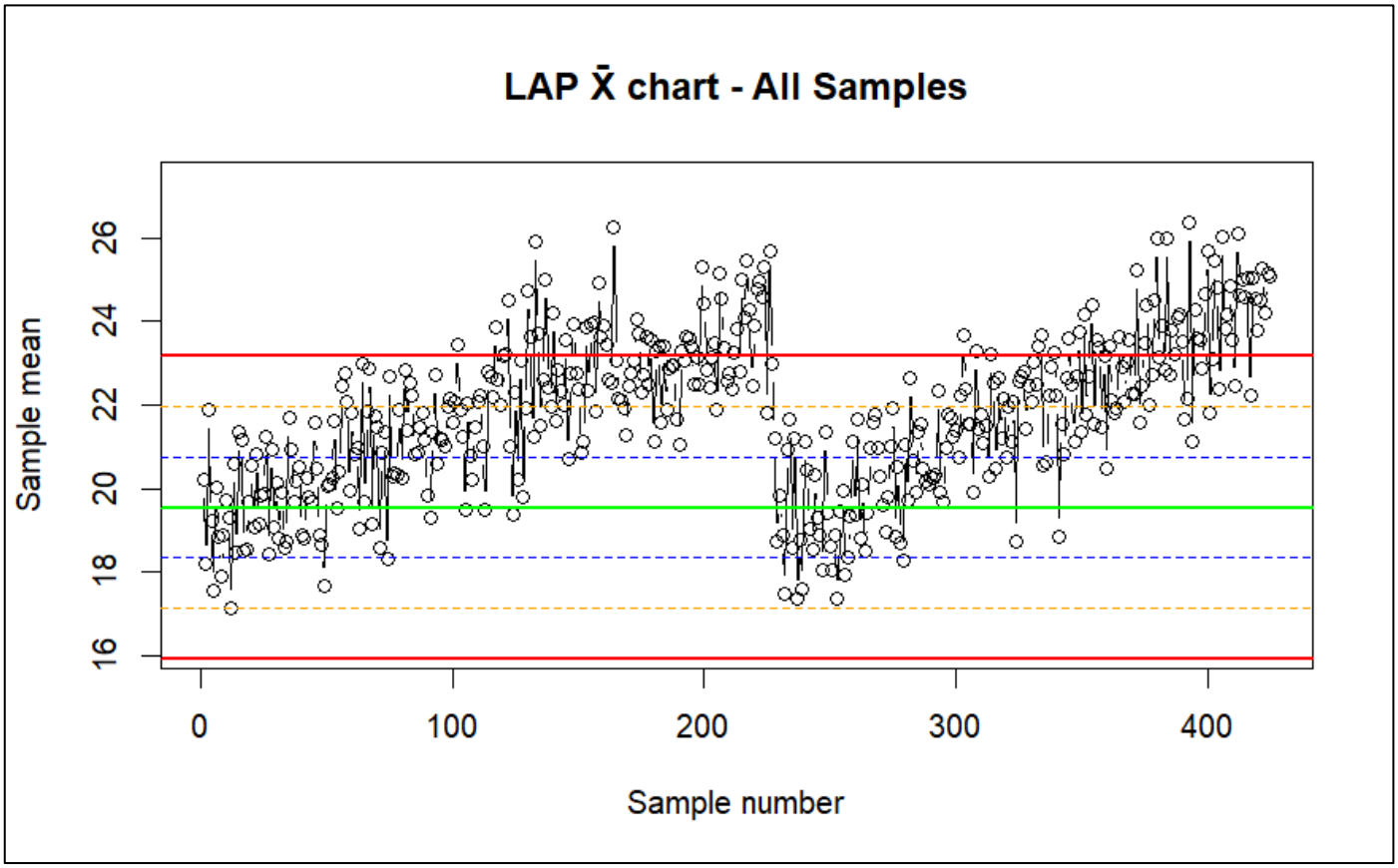

MOU s Chart for All Samples
Sample Standard Deviation
0
2
4
6
8
10
12
0
200
400
600
800
Sample Number


MOU X̄ chart - All Samples
Sample mean
16
18
20
22
24
26
28
0
200
400
600
800
Sample number


MON s Chart for All Samples
Sample Standard Deviation
0
2
4
6
8
10
12
0
100
200
300
400
500
600
Sample Number


MON X̄ chart - All Samples
Sample mean
16
18
20
22
24
26
28
0
100
200
300
400
500
600
Sample number


LAP s Chart for All Samples
Sample Standard Deviation
0
2
4
6
8
10
0
100
200
300
400
Sample Number


LAP X̄ chart - All Samples
Sample mean
16
18
20
22
24
26
0
100
200
300
400
Sample number

Conclusions drawn from s charts: Most of the points/ sample standard deviations fall between the red control limits indicating the process variability is in control/stable. A few points approach the upper limits which represents sudden increases to a higher variation. The variability does not show any specific trends indicating that the delivery times are predictable and consistent. Overall, the standard deviations have a random variation and remain steady. The process is in control.

Conclusions drawn from x bar charts: Most points fall within the control limits. This indicates that delivery times are mostly consistent (predictable), and the process is stable. There are small fluctuations around the centreline indicating a normal low variation in delivery times due to normal variations in the process. It should be noted that around sample 250 to 450 each of the x bar charts experience an upward drift of samples (slower deliveries) and then a spike where the mean temporarily rises/shifts which may indicate a backlog, but the sample recovers back to the average mean after 500th sample indicating action was taken to correct the control. Points near or beyond the control limits indicate delays. The spikes in the mean will be investigated in the following sections.

3.3. Process Capability

In reality the sales would occur and then the data from the process will be obtained. In this report the assumption is made that the samples of 24 are obtained as they are sampled in the R code. If the distribution is too large an X bar chart will not be accurate. The process control needs to be checked by the product manager when the process control rules discussed in the next section are not met. More specifically, all points in the control charts above which are outside the red lines (outer control limits) are problematic and should be addressed as they deviate greatly from the centre line.

Below the capability indices are stated for each product type. These were calculated using the first 1000 deliveries of each product type. Assumptions made include LSL of 0 and USL of 32 hours. The last column of the table states whether the product is capable of meeting the Voice of the Customer (VOC).

	ProdType	Cp	Cpu	Cpl	Cpk	capable
1	MOU	0.9151848	0.7265710	1.103799	0.7265710	FALSE
2	KEY	0.9171375	0.7293536	1.104921	0.7293536	FALSE
3	SOF	18.1352369	35.1876018	1.082872	1.0828720	TRUE
4	CLO	0.8977458	0.7167378	1.078754	0.7167378	FALSE
5	LAP	0.8987816	0.6962187	1.101345	0.6962187	FALSE
6	MON	0.8890490	0.6995705	1.078528	0.6995705	FALSE

In the figure above, the Cp value of SOF is greater than one indicating a capable process. The other products have a value lower than 1 indicating it is not capable. The high Cpu of SOF indicates the process is leaning far away from the upper specification limit unlike for the other products (Cp of less than 1 which is not acceptable due to unmet specifications). All the products have low Cpl values indicating all are far enough away from the lower specification limit. The Cpk of SOF indicates that it meets the specifications.

3.4. Process Control Issues

Control issues are identified based on 3 rules. The first rule (A) is 1 s sample outside the upper +3 sigma-control limits. The second rule (B) is the most consecutive samples of s between the -1 and +1 sigma-control limits. The third rule (C) is 4 consecutive x-bar samples outside the upper second control limits.

Product name	Rule 1	Rule 2	Rule 3
MOU	1 sample namely 592	16 consecutive samples namely 672-687	369 total namely 194,195,196 (first 3) and 858, 859, 860 (last 3)
KEY	0	10 consecutive samples namely 224-233	326 total namely 99, 100, 101 (first3) and 744, 745, 746 (last 3)
SOF	0	13 consecutive samples namely 466-478	384 total namely 133, 134, 135 (first3) and 862, 863, 864 (last 3)
CLO	0	29 consecutive samples namely 476-504	283 total namely 122, 123, 124 (first3) and 647, 648, 649 (last 3)
LAP	0	18 consecutive samples namely 116-133	172 total namely 119, 120, 1215 (first3) and 423, 424, 425 (last 3)
MON	0	34 consecutive samples namely 238-271	263 total namely 134, 135, 136(first3) and 617, 618, 619 (last 3)

4. Risk and Data Correction

4.1. Manufacturer's Error

Type I error also known as manufacturer's error is when the null hypothesis is rejected but the null hypothesis is correct. In context of the previous section, a type I error would be deducing that the process is out of control, but it is in control. It should be noted when calculating the type I error for the three rules discussed in the previous section, R code was used to find the final values. Equations are only included as part of the explanation and logic used when writing the code. It is however possible to convert these equations to the Z domain and used Z tables to calculate it by hand.

The first rule (A) is 1 s sample outside the upper +3 sigma-control limits. In a normal distribution the probability of a Type I error for A is:

$$P_A = 2 \times [1 - P(X < 3)] = P(X > 3) + P(X < -3)$$

The probability of a type I error for B is 0.002699796 or 0.27 %. This indicates that the likeliness of making such an error with regard to A is very low.

The second rule(B) is to identify the most consecutive samples of s between the -1 and +1 sigma-control limits. Type I error for B.in a normal distribution is:

$$P_B = P(-1 < X < 1) = P(X < 1) - P(X < -1)$$

The probability of a type I error for B is 0.6826895 or 68 %. This indicates there is a good chance of making such a type I error. When identifying the probability of consecutive samples (for n consecutive samples) the probability of one sample is raised to the exponent of n. For example, if n = 13, P = 0.006996838 as shown in the code. As n increases the probability decreases.

The third rule is 4 consecutive x-bar samples outside the upper second control limits. In a normal distribution the probability of a type I error for C is:

$$P_c = [P(x > 2) + P(x < -2)]^4 = [2 \times \big(1 - P(X < 2)\big)]^4$$

The probability of a type I error for C is 4.286034e-06. This is extremely unlikely to make this kind of error with regards to rule C.

Another important aspect to note is that on the above control charts the centreline is the process mean therefore the probability that a sample is above the centreline is 0.5 and 0.5 for below the centreline. This is true for a normal distribution due to the symmetry.

4.2. Consumer's Error

Type II error is when the process is out of control but from the control chart rules, it is predicted to be in control. This means the null hypothesis is not rejected, but it should have been.

The null hypothesis is that the process is centred at 25.05 L (still, same as previously). The alternative hypothesis is that the process mean has shifted to 25.028 L.

$$P(type\ II\ error) = P(LCL \leq \bar{X} \leq UCL | \mu = \mu_{new})$$

The probability of a type II error is 0.8411783 or 84 %. This indicates that the chance the shift is not detected from the charts is highly likely.

4.3. Correction of Product Data

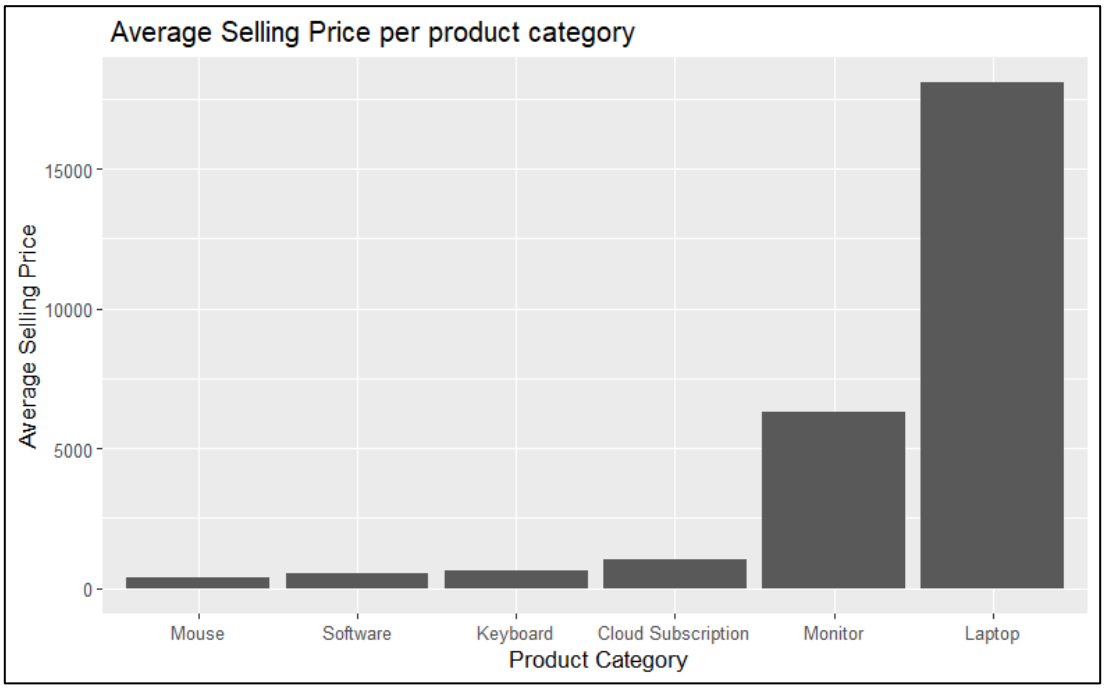
Errors were identified in both the product and head office products datasets. This explains some of the inconsistencies between these two datasets during the initial data analysis. In this section products_data.csv is corrected by matching the right category with the listed product ID. The corrected dataset is saved to products_data2025.csv. The head office data in products_Headoffice.csv was also corrected by fixing NA product IDs and replacing them with the given category's respective product ID prefix. Then the selling price and markup from the products data was used to replace those in the head office data. The corrected data is saved to products_Headoffice2025.csv.

The data analysis conducted earlier on these two datasets are now repeated using the corrected data. The **corrected product data** will be analysed first. Below are the summary statistics.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
product_id*	1	60	30.50000	17.464249	30.500	30.50000	22.239000	1.00	60.00	59.00	0.00000000	-1.2601448	2.2546249
category*	2	60	3.50000	1.722237	3.500	3.50000	2.223900	1.00	6.00	5.00	0.00000000	-1.3258048	0.2223399
description*	3	60	16.40000	10.078001	16.000	16.20833	13.343400	1.00	35.00	34.00	0.10295987	-1.2935763	1.3010643
selling_price	4	60	4493.59283	6503.770150	794.185	3189.25479	525.722547	350.45	19725.18	19374.73	1.42617520	0.4338057	839.6331159
markup	5	60	20.46167	6.072598	20.335	20.51187	7.309218	10.13	29.84	19.71	-0.03670775	-1.2380989	0.7839690

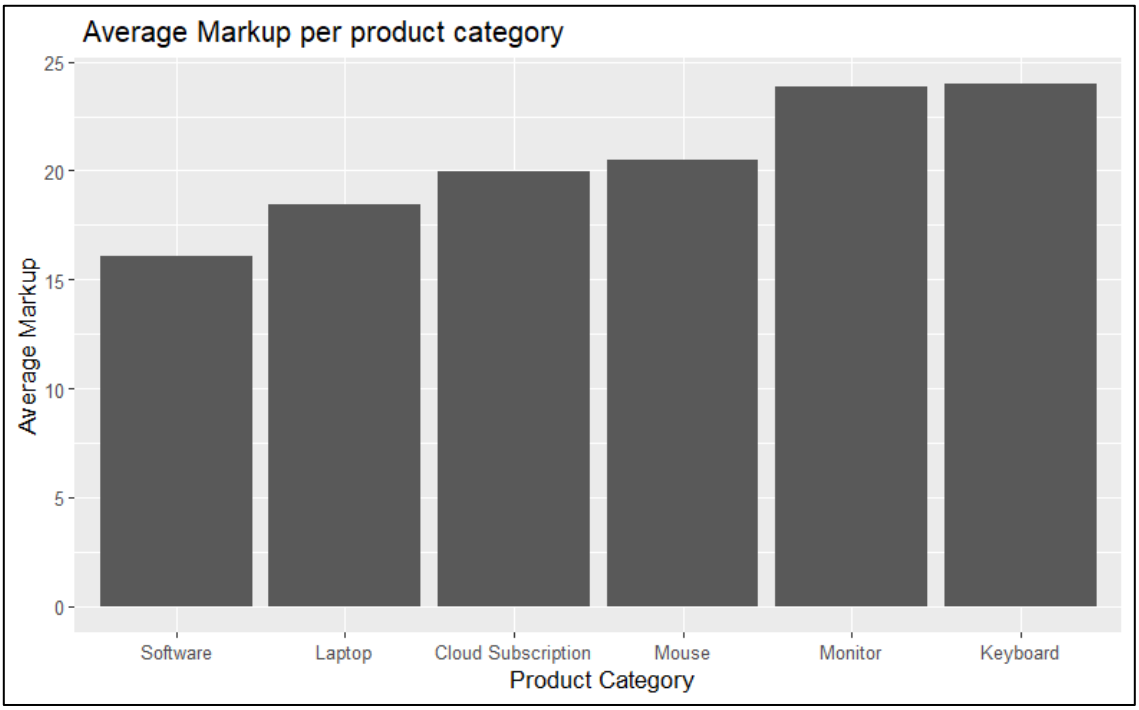
The selling price did not change as well as the product ID therefore the only data analysis that will change is that regarding product category.

It can now be seen that the previous data analysis was very different. It is clear that once average selling price per category is plotted (now that the categories of the product data has been corrected) that laptops have a much higher selling price and mouse has a much lower selling price.



The markup per category also changed but not as drastically. It can now be seen that the category markups vary more than originally identified in the first data analysis.

In the previous anlysis, almost all product categories has an avearge markup up 20%. Now each category varies with keyboard having the highest markup and software the lowest.



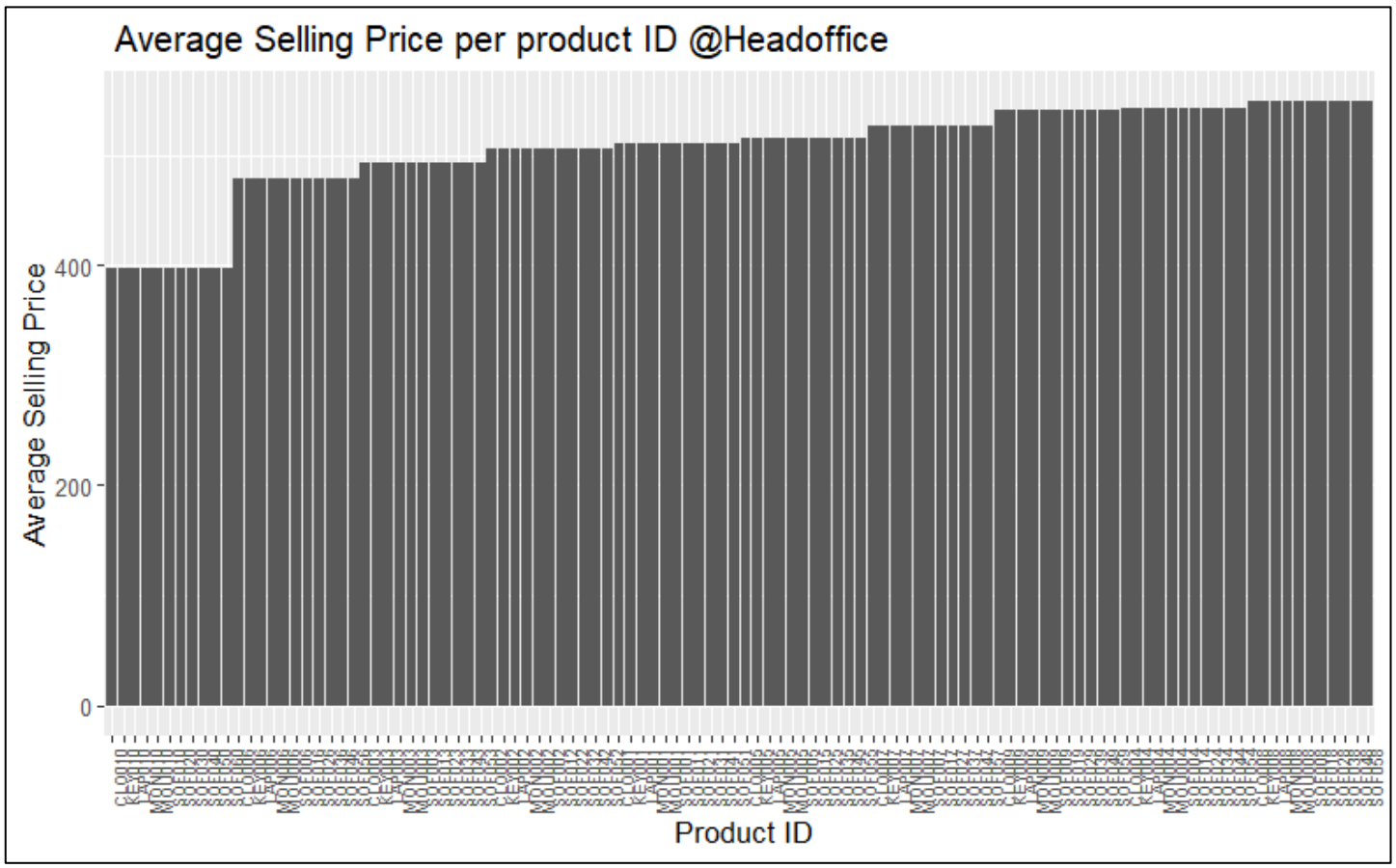
Using the corrected product data, the sales per category for 2023 is calculated by multiplying the quantity and selling price over all products. In 2023 the laptops made the most money according to the sales data (R 1 163 889 479).

	category	SALE2023
1	Cloud Subscription	98715482
2	Keyboard	73499067
3	Laptop	1163889479
4	Monitor	578385570
5	Mouse	51219577
6	Software	66468485

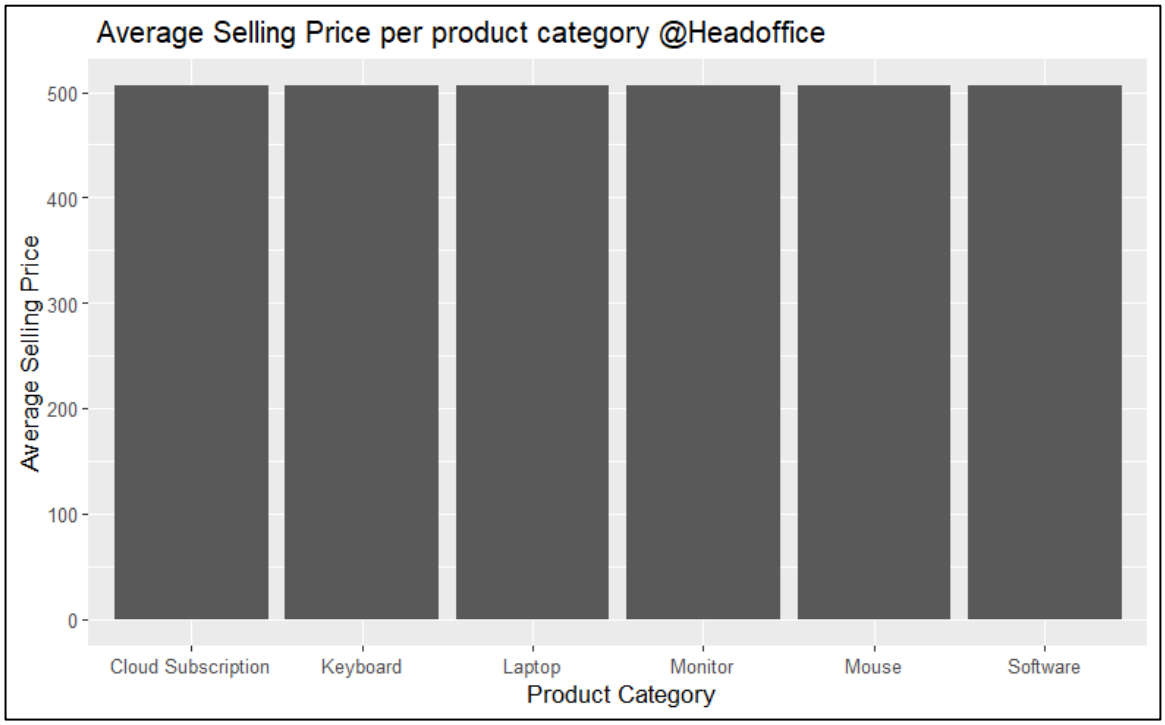
The **corrected head office product data** will now be looked at in depth. It is noted that the product category and description did not change. The summary statistics are seen below. There are now no missing values in the dataset.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
product_id*	1	360	76.33333	25.405897	80.500	79.66667	22.239000	1.00	110.00	109.00	-1.06752495	0.7225690	1.33900835
category*	2	360	3.50000	1.710202	3.500	3.50000	2.223900	1.00	6.00	5.00	0.00000000	-1.2781771	0.09013556
description*	3	360	30.68611	17.319505	29.500	30.76736	22.980300	1.00	60.00	59.00	-0.02778185	-1.3900365	0.91281808
selling_price	4	360	506.18300	42.244602	513.840	514.51125	34.633536	396.72	549.02	152.30	-1.53341631	1.7367866	2.22648600
markup	5	360	16.04000	4.852187	16.485	15.61500	7.175784	10.43	25.05	14.62	0.58504345	-0.8280804	0.25573271

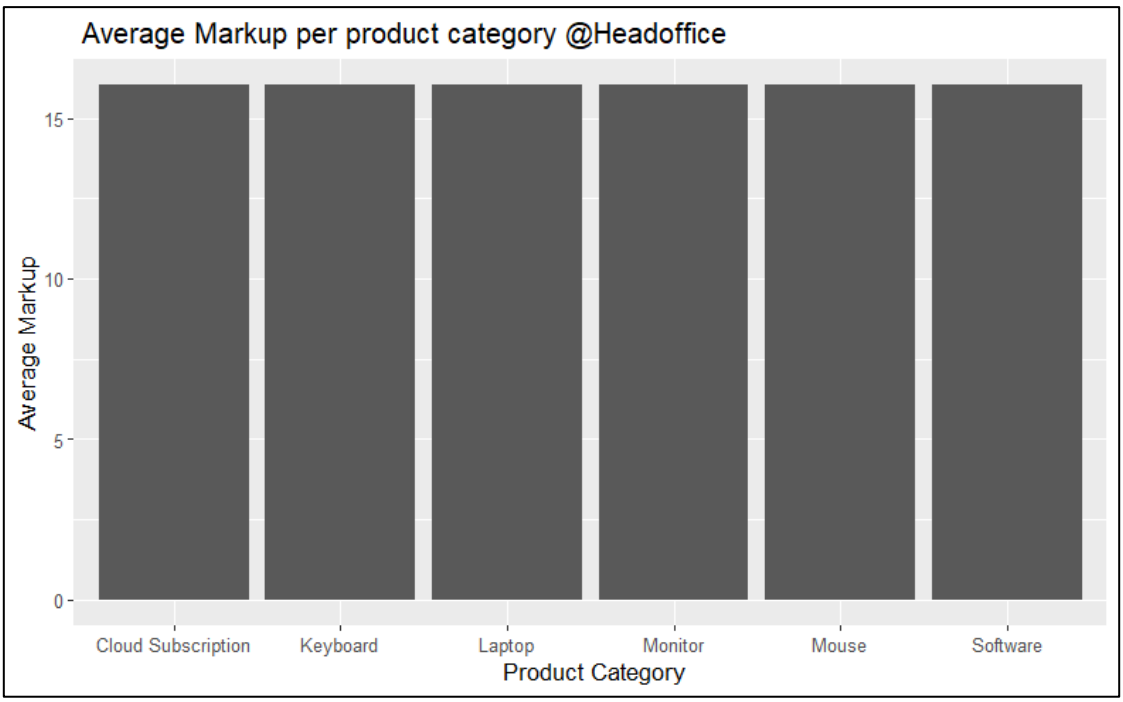
The average selling price per product ID is plotted below. The average selling price is now almost uniformly distributed whereas previous there was an exponential distribution.

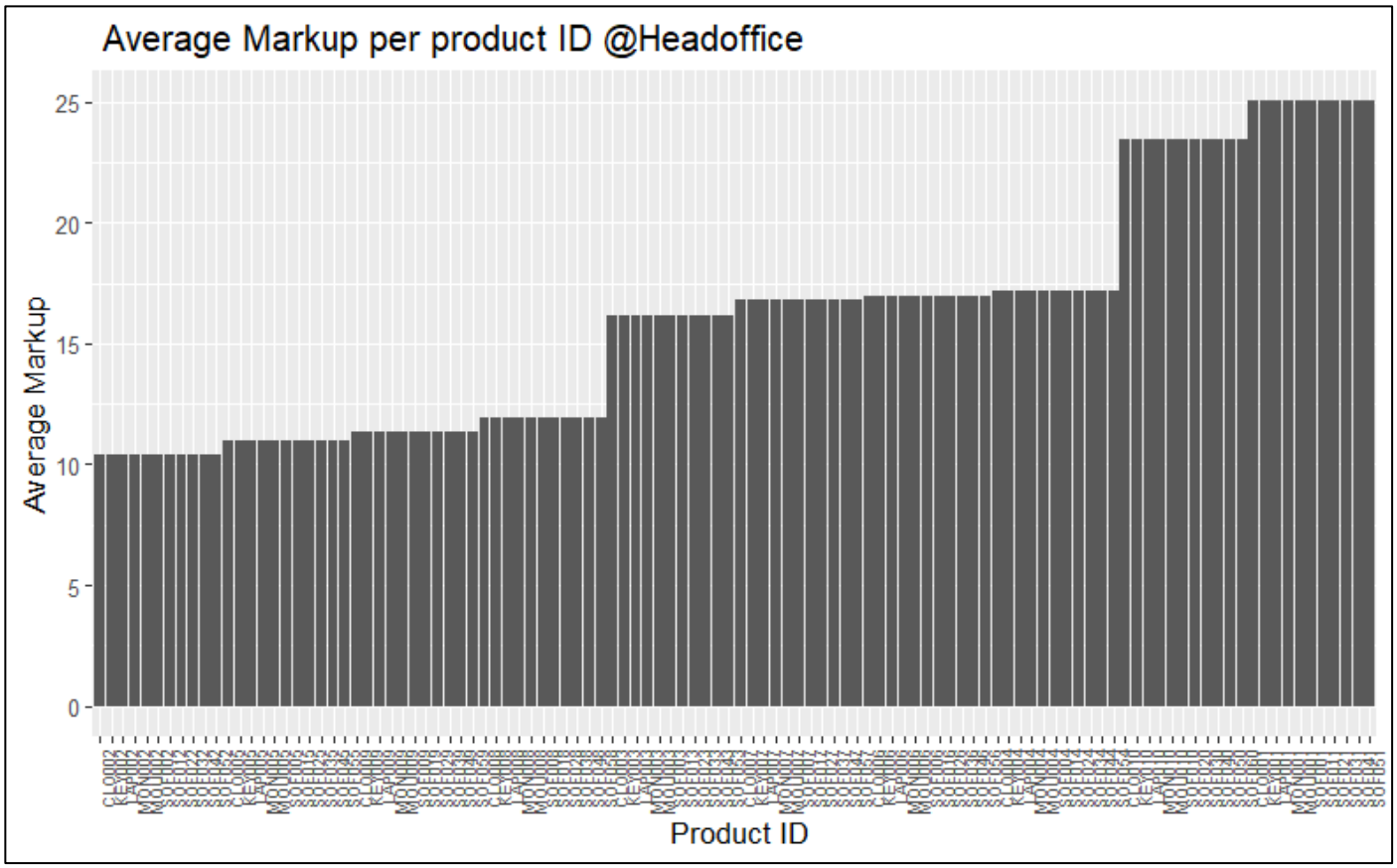


The average selling price per product category is plotted below. The average selling price is now 500 for every product category.



The average markup was corrected so both the plots below have changed. The average markup is now the same (16%) for all the product categories. The products with the higher markups appear to be in the Software category.





5. Optimising for Maximum Profit

In this section of the report, data is provided for two different coffeeshops. The goal of this section is to build a model suggested by a previous analyst which will optimise the profit by choosing the optimal number of baristas at each shop. The assumption is made that a maximum of six baristas may be employed. The data spans one year.

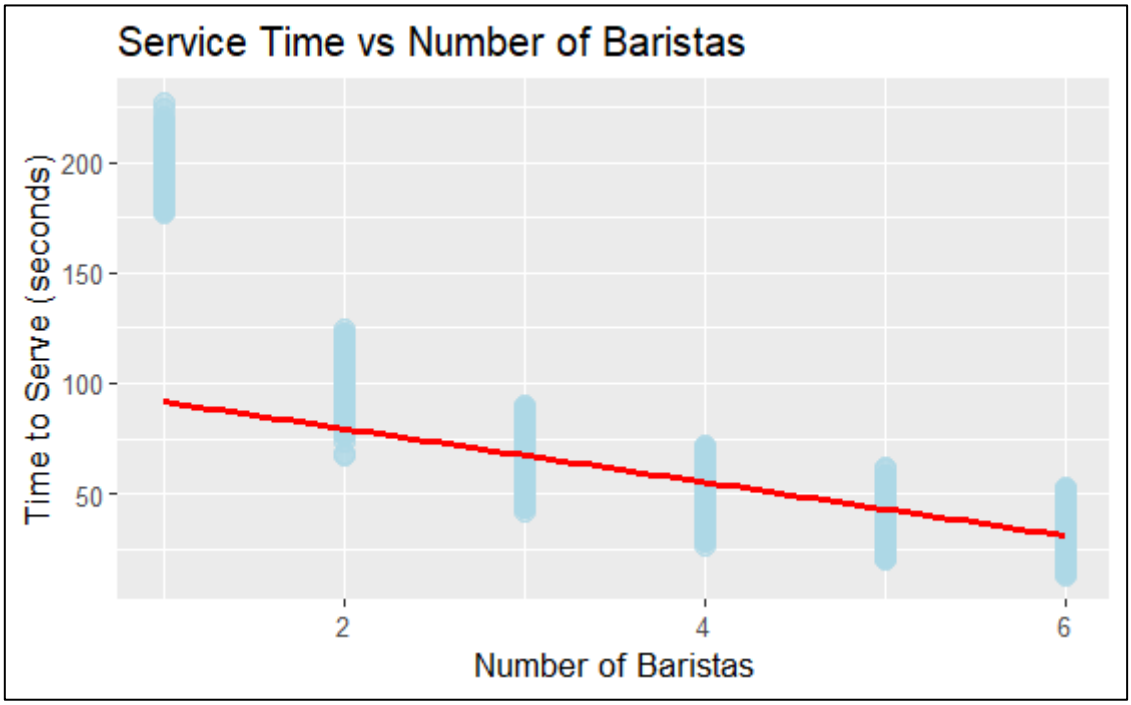
5.1. The First Coffeeshop Dataset

The profit for the given dataset namely timeToServe.csv is calculated using individual service times. The number of baristas is then analysed to determine the optimal or maximum profit that can be made at the coffee shop. The main factors to consider is speed, service quality and operational costs when making decisions regarding the model.

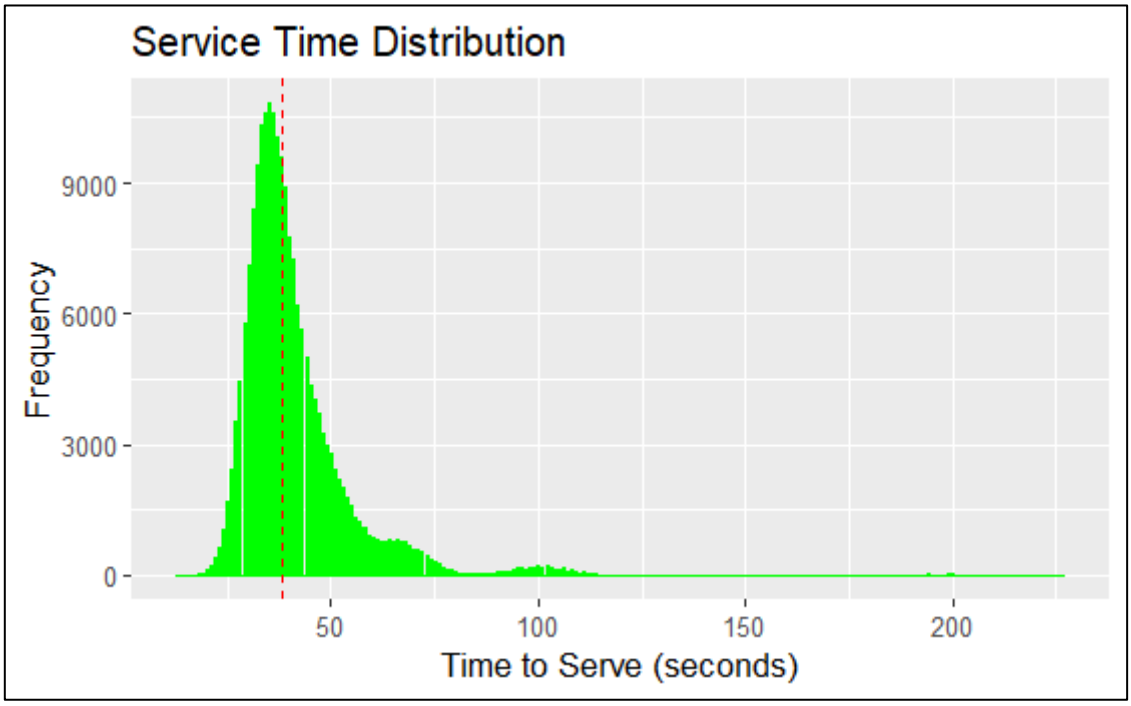
First to better understand the data provided, a simple data analysis is done to gain summary statistics and figures to identify the distribution. The data is cleaned up in R and is given appropriate column names, namely baristas and timetoserve.

baristas	timetoserve
Min. :1.00	Min. : 13.00
1st Qu.:5.00	1st Qu.: 33.00
Median :5.00	Median : 38.00
Mean :5.16	Mean : 41.22
3rd Qu.:6.00	3rd Qu.: 45.00
Max. :6.00	Max. :227.00

The time taken to serve a customer is then plotted against the number of baristas working that day. Here a clear decreasing exponential trend can be seen. The time taken decreases exponentially as the number of baristas increases. This is a logical conclusion as more baristas can work faster and more efficiently allowing for time saving and cost saving strategies to be implemented.



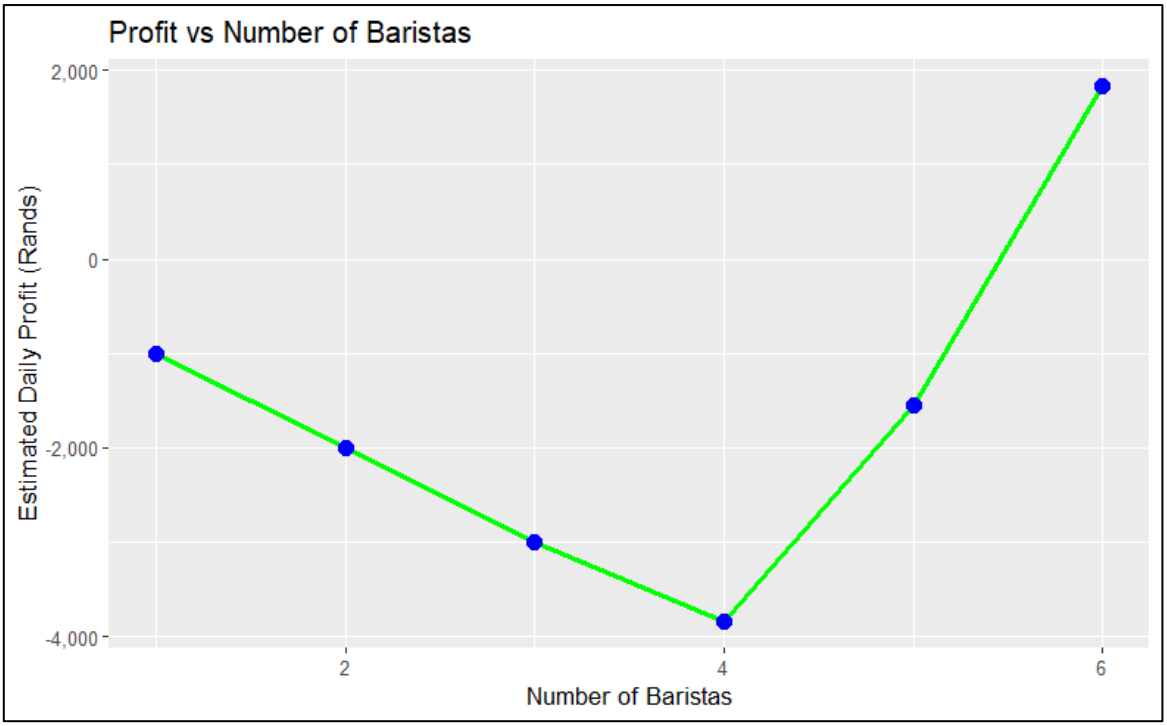
The distribution of the service time is then further investigated. In the figure below a normal distribution can be seen which is skewed to the right. There are also some visible outliers which accounts for baristas taking longer to serve for reasons like getting an order wrong or taking a break to have a social interaction. This is seen by the long tail on the right of the histogram below. These outliers should be investigated as the times that are far from the mean indicate issues that should be resolved. The assumption made in the figure below (as seen by the dashed red line) is that to be considered reliable, the time to serve must be 38 seconds or less.



The data is modelled in R in order to find this optimal point that balances performance and efficiency allowing for maximum profit. Assumptions made in the model include a profit of R30 per customer served and a cost per employee of R1000 per day. The personnel costs, expected revenue and profit are then calculated. The profit is also plotted below.

	baristas	meanTime	Reliability	count	ExpectedRevenue	PersonnelCost	Profit
1	1	200.15588	0.00000000	417	0.0000	1000	-1000.000
2	2	100.17098	0.00000000	3556	0.0000	2000	-2000.000
3	3	66.61174	0.00000000	12126	0.0000	3000	-3000.000
4	4	49.98038	0.01859751	29305	167.3776	4000	-3832.622
5	5	39.96183	0.38415548	56701	3457.3993	5000	-1542.601
6	6	33.35565	0.87017723	97895	7831.5951	6000	1831.595

It can then be seen by the results the model has identified (as seen below), the optimal number of baristas as six as the reliability for this number of baristas is the highest at 87% and the largest profit of R1 831.60.



The output produced by the R model is seen below. The average time to serve is 33 seconds and reliability 0.87. The personnel cost is R6000 and expected revenue is R7831.60 and therefore R1831.60.

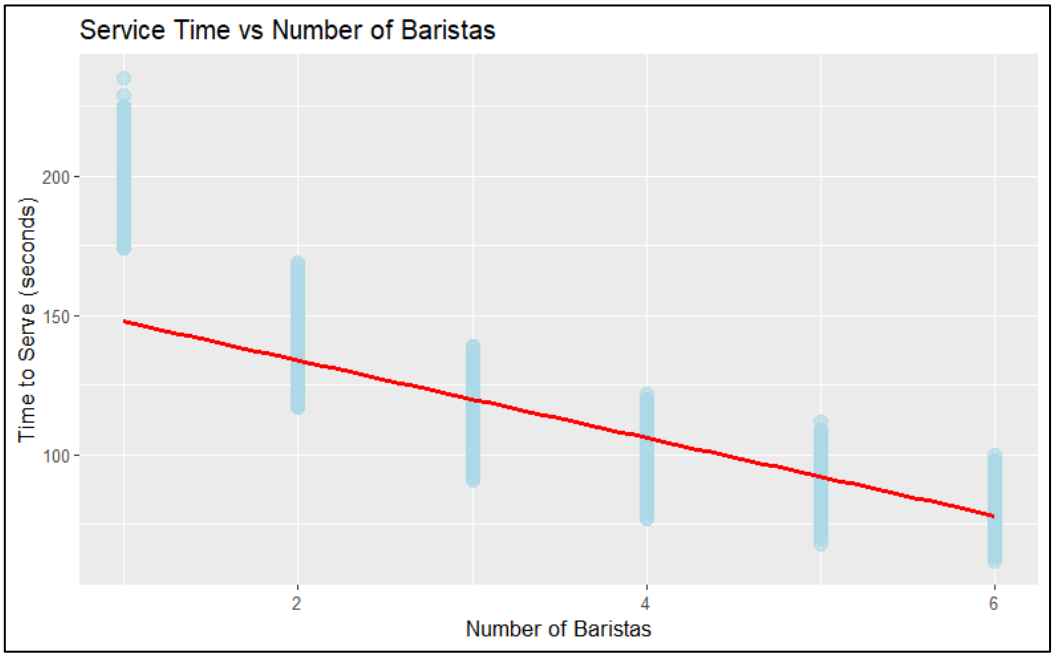
	baristas	meanTime	Reliability	count	ExpectedRevenue	PersonnelCost	Profit
1	6	33.35565	0.8701772	97895	7831.595	6000	1831.595

5.2. The Second Coffeeshop Dataset

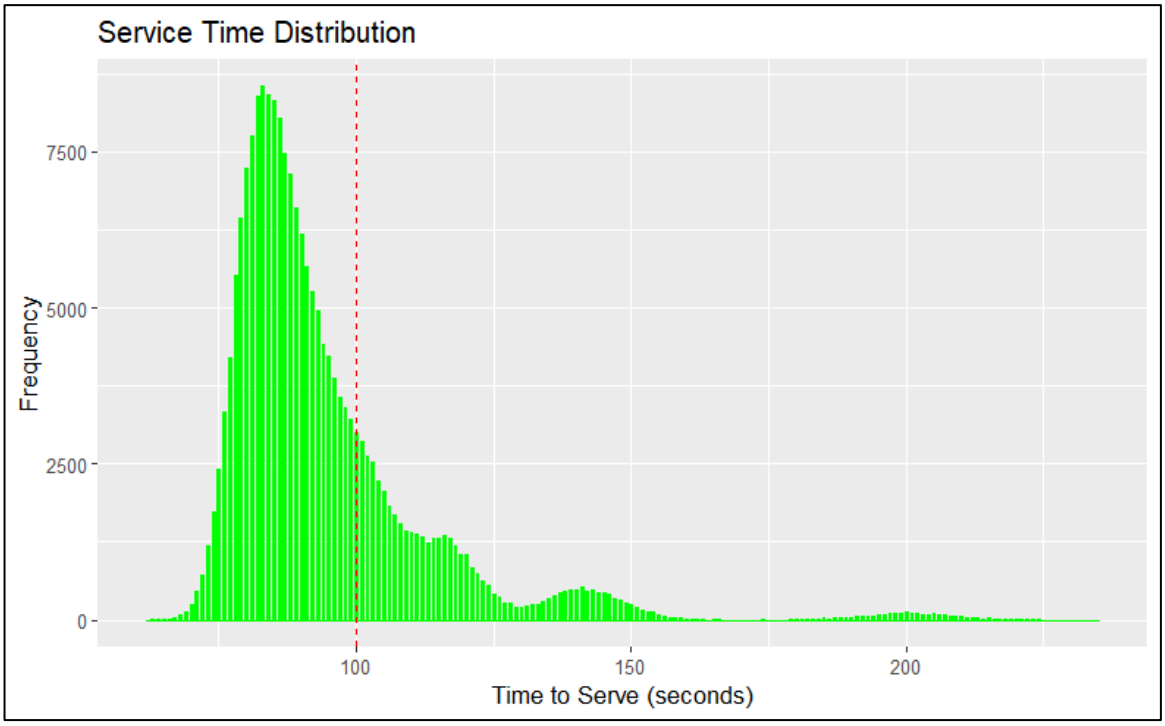
First to better understand the data provided, a simple data analysis is done to gain summary statistics and figures to identify the distribution. The data is cleaned up in R and is given appropriate column names, namely Baristas2 and TimeToServe2.

```
  baristas2      timetoserve2
Min.   :1.000   Min.   : 62.00
1st Qu.:4.000   1st Qu.: 83.00
Median :5.000   Median : 89.00
Mean   :4.844   Mean   : 94.32
3rd Qu.:6.000   3rd Qu.:100.00
Max.   :6.000   Max.   :235.00
```

The time taken to serve in then plotted for each number odf baristas. There is a linear relationship between time to serve and number of baristas, indicating that the more baristas employed, the less the time taken to serve customers, as expected.



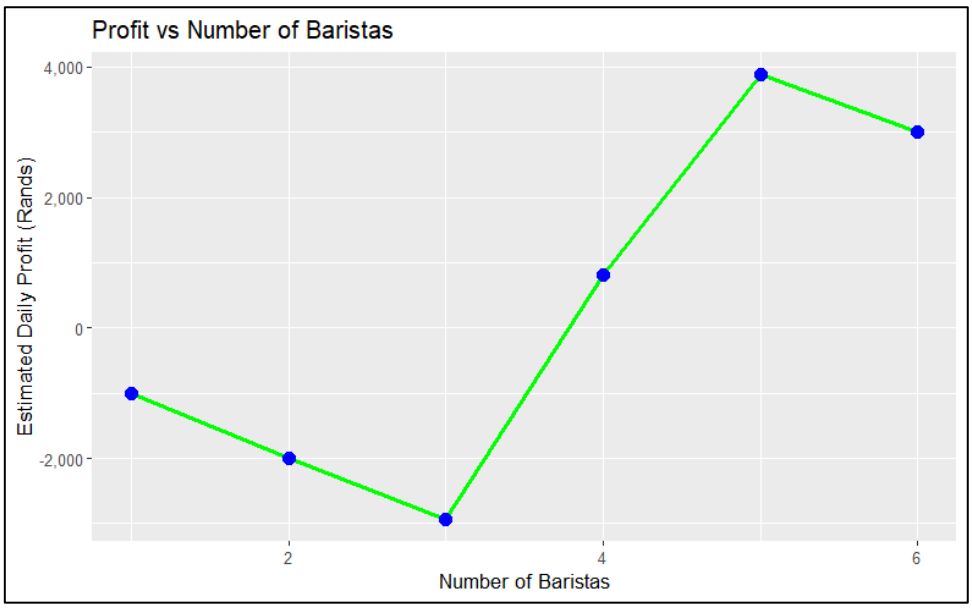
The distribution of the service time is then again further investigated. In the figure below a normal distribution can be seen which is skewed to the right. There are even more visible outliers in the second dataset which accounts for baristas taking longer to serve for reasons like getting an order wrong or taking a break to have a social interaction. This is seen by the long tail on the right of the histogram below. These outliers should be investigated as the times that are far from the mean indicate issues that should be resolved. The assumption made in the figure below (as seen by the dashed red line) is that to be considered reliable, the time to serve must be 100 seconds or less.



The data is modelled in R in order to find this optimal point that balances performance and efficiency allowing for maximum profit. Assumptions made in the model include a profit of R30 per customer served and a cost per employee of R1000 per day. The personnel costs, expected revenue and profit are then calculated. The profit is also plotted below.

	baristas2	meanTime2	Reliability2	count	ExpectedRevenue2	PersonnelCost2	Profit2
1	1	200.16894	0.000000000	2196	0.00000	1000	-1000.0000
2	2	141.51462	0.000000000	8859	0.00000	2000	-2000.0000
3	3	115.44091	0.007891542	19768	71.02388	3000	-2928.9761
4	4	100.01527	0.534472499	35289	4810.25249	4000	810.2525
5	5	89.43597	0.986753521	54958	8880.78169	5000	3880.7817
6	6	81.64272	1.000000000	78930	9000.00000	6000	3000.0000

It can then be seen by the results the model has identified (as seen below), **the optimal number of baristas as five** as the reliability for this number of baristas is the highest at 98.68% and the largest profit of R13880.78 per day.



The output produced by the R model is seen below. The average time to serve is 89 seconds and reliability 0.9868. The personnel cost is R5000 and expected revenue is R8880.78 and therefore R3880.78.

	baristas2	meanTime2	Reliability2	count	ExpectedRevenue2	PersonnelCost2	Profit2
1	5	89.43597	0.9867535	54958	8880.782	5000	3880.782

6. ANOVA to Test Two Hypotheses

An ANOVA or Analysis of Variance is a useful statistical tool which can be used to compare the means of different groups of data. The ANOVA format shows the main effects two features have by themselves and then the effect of the interaction between the two features.

Two two-way ANOVAs are created to test various hypotheses. For the first ANOVA the hypotheses are as follows:

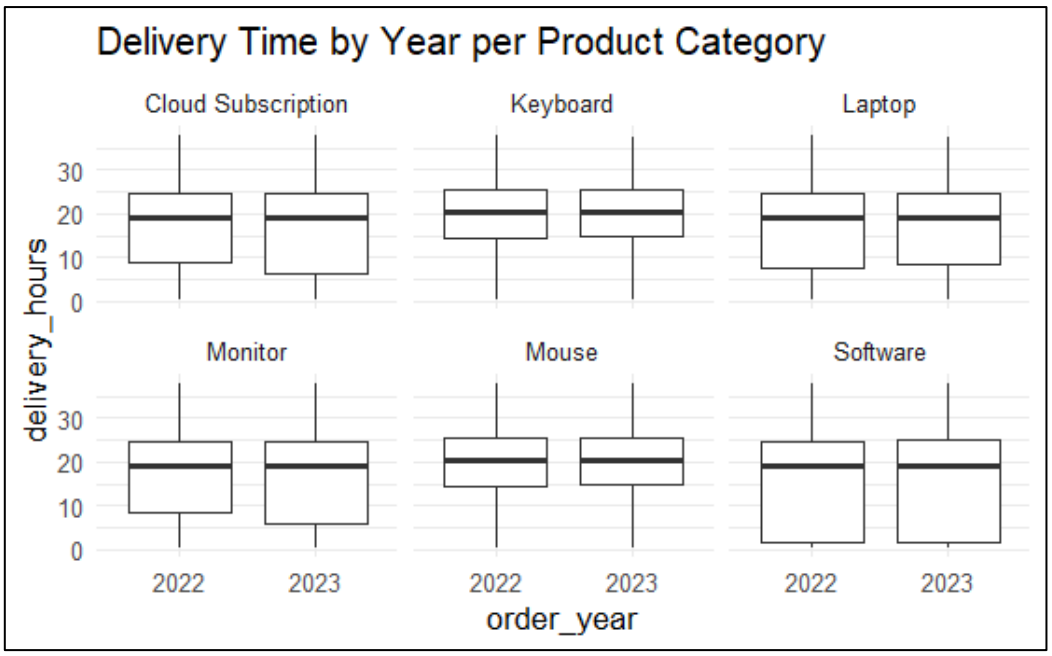
- First null hypothesis: The mean of the dependent variable is the same across all years
- First alternative hypothesis: At least one year has a different mean.
- Second null hypothesis: The mean of the dependent variable is the same across all product categories
- Second alternative hypothesis: At least one product category has a different mean.
- Third null hypothesis: The effect of category does not depend on year.
- Third alternative hypothesis: The effect of category does depend on year

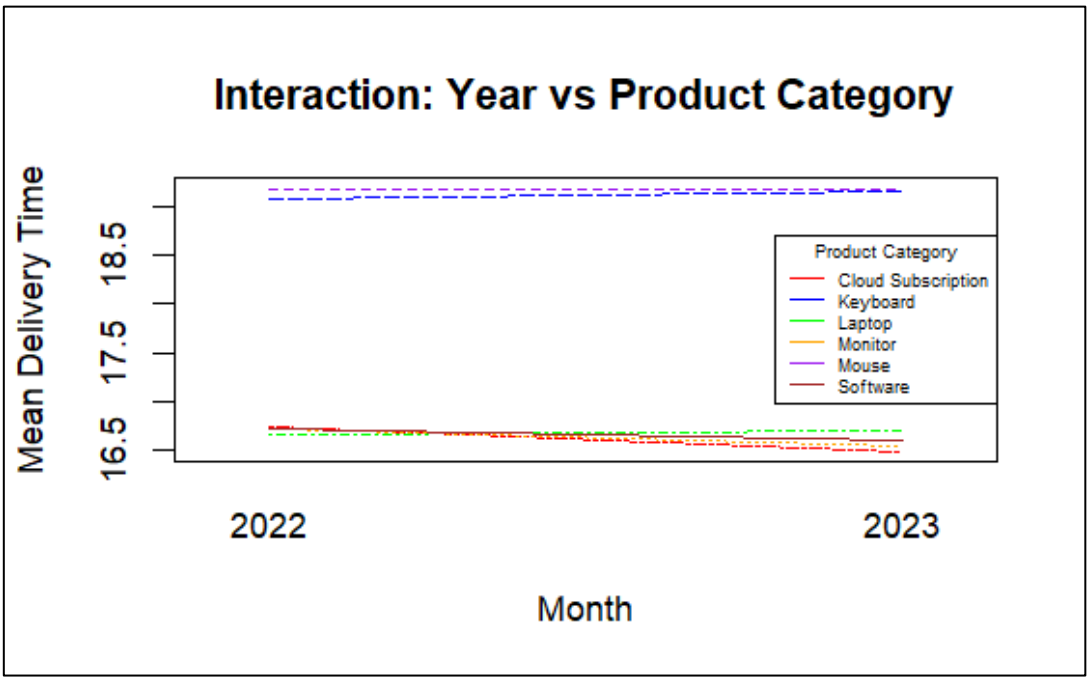
Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Order year	1	138	138	1.395	0.237
Category	5	136825	27365	277.433	<2e-16
Order year: Category (Interaction)	5	349	70	0.707	0.618
Residuals	99988	9862476	99		

The following conclusions can be drawn from the ANOVA above:

- Order year: $p = 0.237$ indicates that there is no significant difference between years. Delivery time stayed roughly the same from one year to the next. Do not reject the first null hypothesis.
- Category: $p<0.001$ indicates that there are big differences in delivery time between categories. Some categories have longer delivery times than others. Reject the second null hypothesis.
- Interaction: $p = 0.618$ indicates that there is no interaction. The pattern of categories' delivery times stayed the same over the years considered (as seen in the interaction plot below). Do not reject the third null hypothesis.

The box plot diagram below supports the findings of the ANOVA above. The delivery times change from product to product however remain the same from year to year.





For the second ANOVA the hypotheses are as follows:

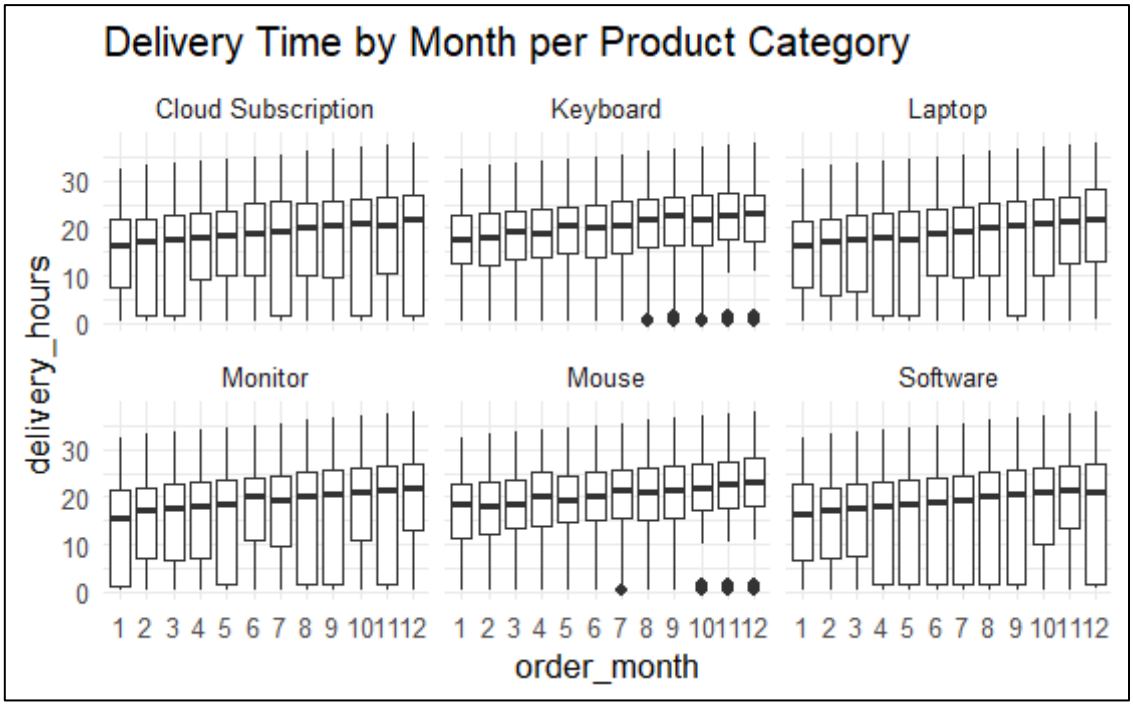
- First null hypothesis: The mean of the dependent variable is the same across all months
- First alternative hypothesis: At least one month has a different mean.
- Second null hypothesis: The mean of the dependent variable is the same across all product categories
- Second alternative hypothesis: At least one product category has a different mean.
- Third null hypothesis: The effect of category does not depend on month.
- Third alternative hypothesis: The effect of category does depend on month

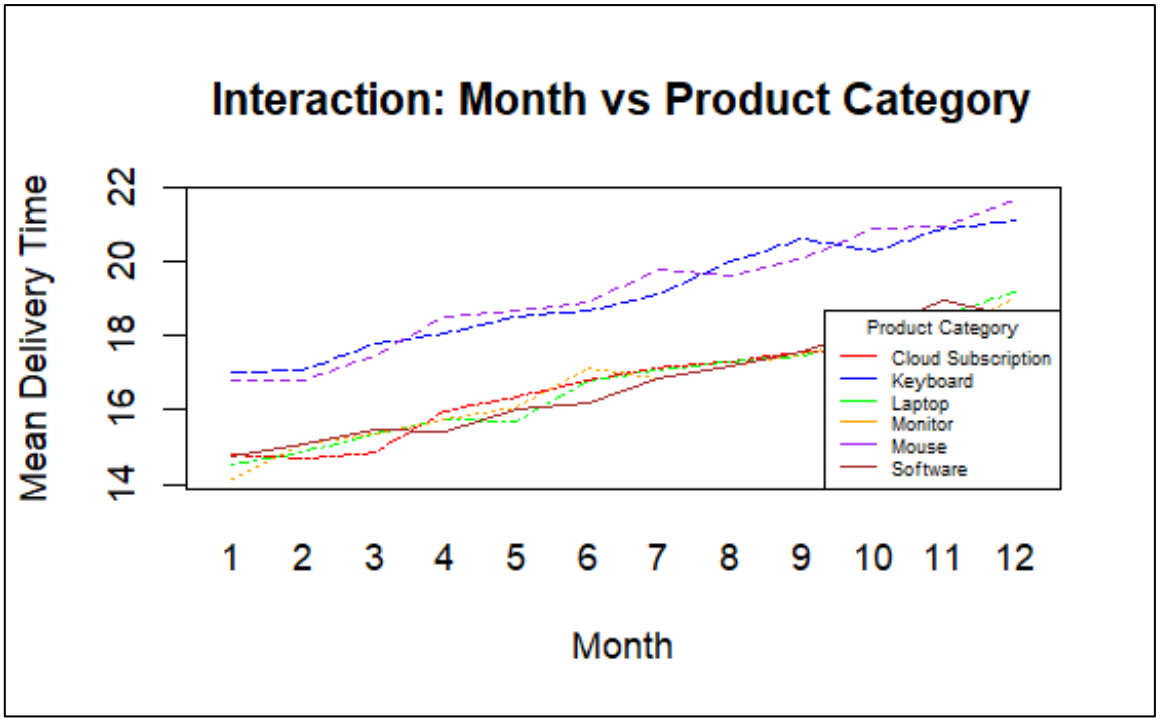
Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Order month	11	172589	15690	161.858	<2e-16
Category	5	134534	26907	277.572	<2e-16
Order month: Category (Interaction)	55	6006	109	1.126	0.242
Residuals	99928	9686659	97		

The following conclusions can be drawn from the ANOVA above:

- Order month: p <0.001 indicates that there is a significant and strong monthly trend in delivery times. Some months have higher or lower delivery times due to seasonality in the data. Reject the first null hypothesis.
- Category: p<0.001 indicates that there are big differences between categories. Some categories have longer delivery times than others. Reject the second null hypothesis.
- Interaction: p = 0.242 indicates that there is no interaction. The pattern of categories' delivery times stayed the same over the months considered (as seen in the interaction plot below). Do not reject the third null hypothesis.

The box plot diagram below supports the findings of the ANOVA above. The delivery times change from product to product and increase as the months of the year increase.



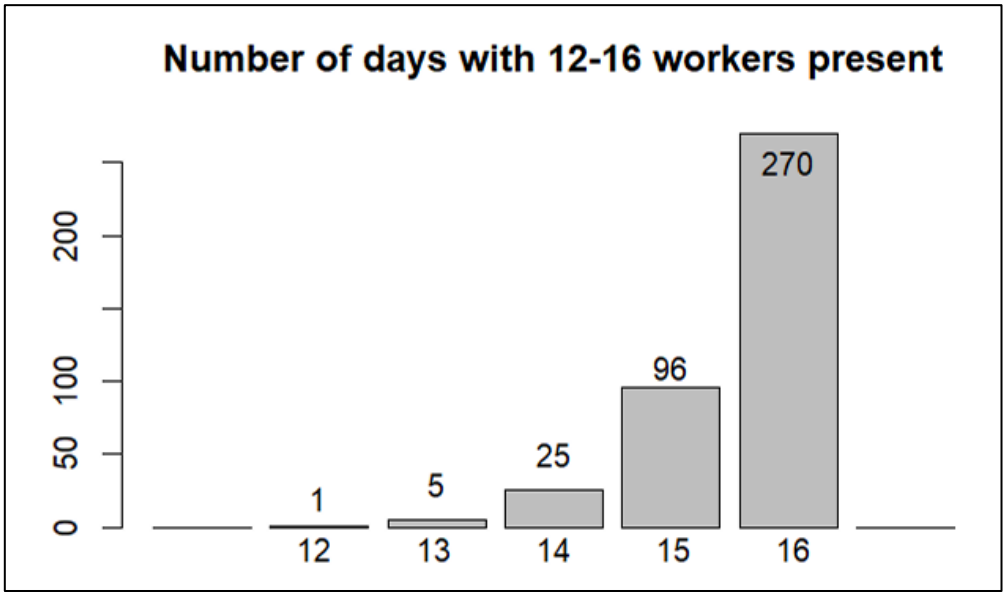


In conclusion the product category and month both greatly impact delivery times. This may be indicative of seasonal trends in data. The way categories change over the years or months is insignificant over time. The delivery times also remain the same over years but change for the months in those years. Understanding the results of the ANOVA will assist the technology company by allowing for a better understanding of the impact of features in the data.

Knowing that the category and month impacts delivery times, planning for production orders and delivery becomes simplified. Having this information will allow for the company to reduce delivery times or possibly warn customers when there are longer delivery times or even promote shorter delivery times. The company can also identify problem areas such as categories that require adjustments in supply chain to lower delivery times.

7. Reliability of a Service

In this section of the report, information regarding a car rental agency is provided. The data given the number of people on duty over 397 days. The data is given in the form of a bar graph indicating the number of days for which a specific number of workers were present as seen in the figure below.



7.1. Reliable Service Estimation

In order to estimate how many days per year it is expected to have reliable service at the car rental agency, the assumption that less than 15 workers will result in unreliable service.

We can therefore calculate the expected number of reliable days

$$R_{days\ in\ sample(workers=15,16)} = 96 + 270 = 366\ days$$

$$Relaible_{days\ in\ year} = \frac{366}{397} \times 365 = 336\ days$$

7.2. Profit optimization

The number of workers that allows for the best reliability and maximum profit needs to be determined. Every day that problems occur at the agency results in a loss of R20 000 sales per day. Appointing additional workers cost R25 000 per person per month or R300 000 per year. The data is modelled as a binomial problem. The formula for a binomial distribution can be seen below.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The mean needs to be calculated.

$$\mu = \frac{1(12) + 5(13) + 25(14) + 96(19) + 270(16)}{397} = \frac{6187}{397} = 15.59$$

However: $p = \frac{\mu}{n}$ therefore $p = \frac{15.59}{16} = 0.974$. This is a good estimation for p in the model.

The p value calculated is the probability that each worker has of being present on that day. Therefore, the number of workers on duty per day ~Binomial (n, p).

This binomial problem is modelled in R. The following is generated as a result of running the model on the given data. Where n is the number of workers on a given day.

	n	prob_problem_day	expected_annual_loss	extra_staff_cost	total_cost
3	17	0.0091	66386.29	300000	366386.3
2	16	0.0637	465259.76	0	465259.8
4	18	0.0010	7618.93	600000	607618.9
5	19	0.0001	743.15	900000	900743.1
6	20	0.0000	63.82	1200000	1200063.8
1	15	0.3264	2382920.69	0	2382920.7

We can therefore conclude that the optimal number of workers is 17 as the total cost is the least therefore profit will be the greatest. The probability that there is a problem on any given day is also low, 0.9 %.

Understanding the reliability of a service is vital. Using this model the car rental agency can now make more informed decisions regarding reliability. Delivering a reliable service will result in the company making more profit, not just short term but long term too. If customers are satisfied, they will return, and the agency will benefit. Positive word of mouth is another result of providing a reliable service.

8. Conclusion

This report meets the criteria of Graduate Attribute 4 being assessed in Quality Assurance 344 because R was used to write programs that contains understandable and reusable code and performs the handling of multiple raw datasets.

Data analysis was carried out on multiple datasets and the relationships between features were identified as well as incorrect data corrected. The report gave great insight in optimising profits and reliability of companies.

In conclusion the basic data analysis of the technology company has provided insights into the sales and customers. Various statistical techniques were applied throughout the report like SPC, ANOVA and more.

9. Bibliography

Dirkse van Schalwyk, T., n.d. *QA344 Statistics*. [Online]
Available at: https://stemlearn.sun.ac.za
[Accessed October 2025].

10. AI Declaration

AI system used	What was it used for?	Where in the work was it used?
ChatGPT	Code assistance	The Rmd file
To what extent did you use AI and why do you consider the work as your own?		
AI was used to help with coding requiring above the expected level of a third-year industrial student. I consider the report and idea generation to be completely my own work. AI was only used when the coding required was above my skill level.		