ECSA ANALYSIS

Pieter Fourie - 27127966

QUALITY ASSURANCE 344

Table of Contents

Introduction

This report presents a comprehensive statistical analysis of business processes to enhance operational efficiency, quality control, and profitability. Drawing on multiple datasets, including sales records, product information, and service times, the project employs R programming for rigorous data wrangling, descriptive statistics, and advanced analytical techniques. The body of the report is structured to guide the analysis. Beginning with an initial data exploration, it progresses to establish Statistical Process Control (SPC) charts and Process Capability indices for delivery times, evaluates Type I and Type II error risks, and corrects data integrity issues. Subsequently, it optimizes staffing models for service operations to maximize profit and applies Multivariate Analysis of Variance (MANOVA) to investigate significant differences within the data. Finally, the reliability of service is modelled and optimized. The end goal is to transform raw data into usable information that shows insights into the data.

1.2: Descriptive statistics

Customer Data

1. **Data Loading and Inspection**

   The Head function was used to allow us to see the structure, dimensions, and variable types. The customer data set includes a CustomerID, Gender, Age, Income and City feature for each of its 5000 rows of data. The customer ID is a mix of characters (CUST), indicating that the record is of a customer and a unique number, making cross-referencing easy.

	CustomerID <chr>	Gender <chr>	Age <int>	Income <dbl>	City <chr>
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

2. **Summary Statistics**

   A summary of each feature allows us to see the range and later use that information to create graphs. We can see more in-depth data about each feature. We see that customerID, Gender and City are all of type character and Age and Income are integer values. Here we once again see that there are 5000 records.

```
  CustomerID          Gender              Age             Income          City
 Length:5000        Length:5000        Min.   : 16.00   Min.   :  5000   Length:5000
 Class :character   Class :character   1st Qu.: 33.00   1st Qu.: 55000   Class :character
 Mode  :character   Mode  :character   Median : 51.00   Median : 85000   Mode  :character
                                       Mean   : 51.55   Mean   : 80797
                                       3rd Qu.: 68.00   3rd Qu.:105000
                                       Max.   :105.00   Max.   :140000

Female   Male  Other
  2432   2350    218

     Chicago      Houston  Los Angeles        Miami     New York San Francisco      Seattle
         724          724          726          647          726          780          673
```

3. **Handling Missing Values**

   This function allows us to make changes to our calculations if missing values are found. Here we see that the dataset contains no missing values, and this allows us to do calculations on the data without having to exclude any records and possibly skewing the outcomes.

```
CustomerID     Gender        Age     Income       City
         0          0          0          0          0
[1] 0
```

4. **Data Filtering and Subsetting**

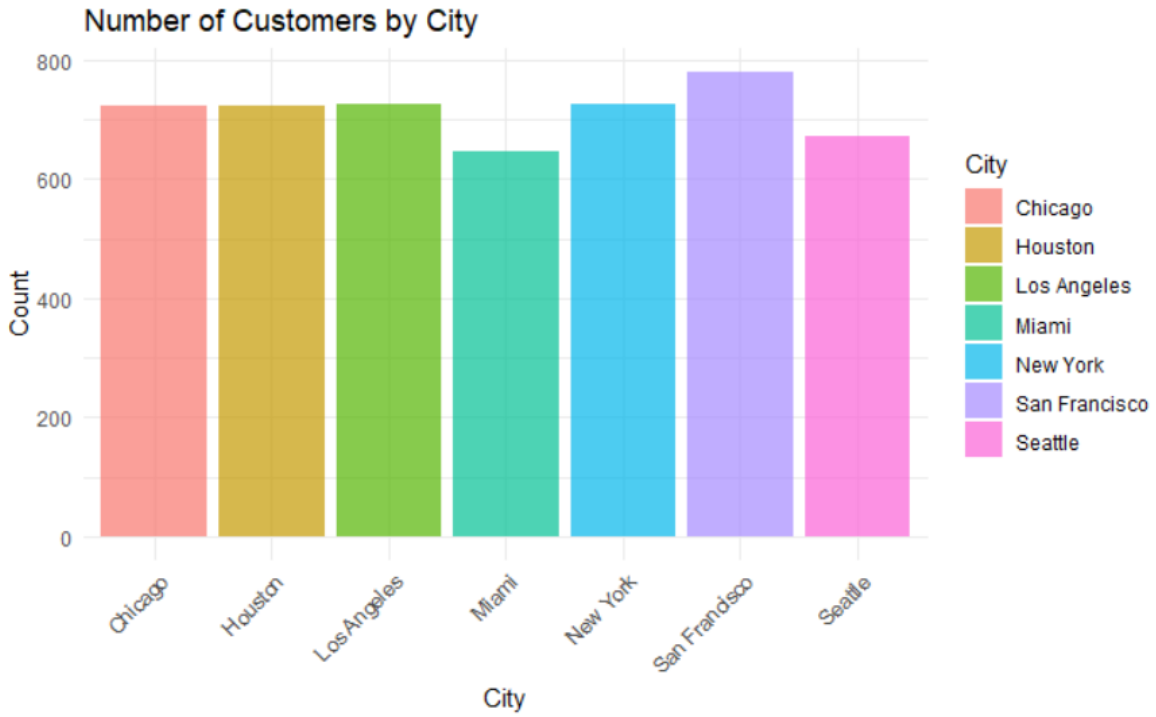
   From using filtering, we can see that roughly 20% of our customers are high-income individuals, having an income of more than 110 000 per year. Other interesting observations that can be made are that most of our top-earning customers seem to be men above the age of 40. We also see that there are a variety of cities in the top 10

with Houston and Los Angeles making up the majority.

CustomerID <chr>	Gender <chr>	Age <int>	Income <dbl>	City <chr>
CUST1502	Male	50	120000	Houston
CUST1503	Male	62	120000	San Francisco
CUST1511	Male	41	140000	Los Angeles
CUST1514	Female	48	140000	San Francisco
CUST1515	Male	49	140000	Houston
CUST1517	Male	62	140000	Chicago
CUST1518	Female	58	130000	Los Angeles
CUST1519	Male	47	115000	Los Angeles
CUST1520	Female	63	120000	Seattle
CUST1527	Male	43	120000	Houston

1-10 of 956 rows Previous 1 2

5. **Data Visualization**

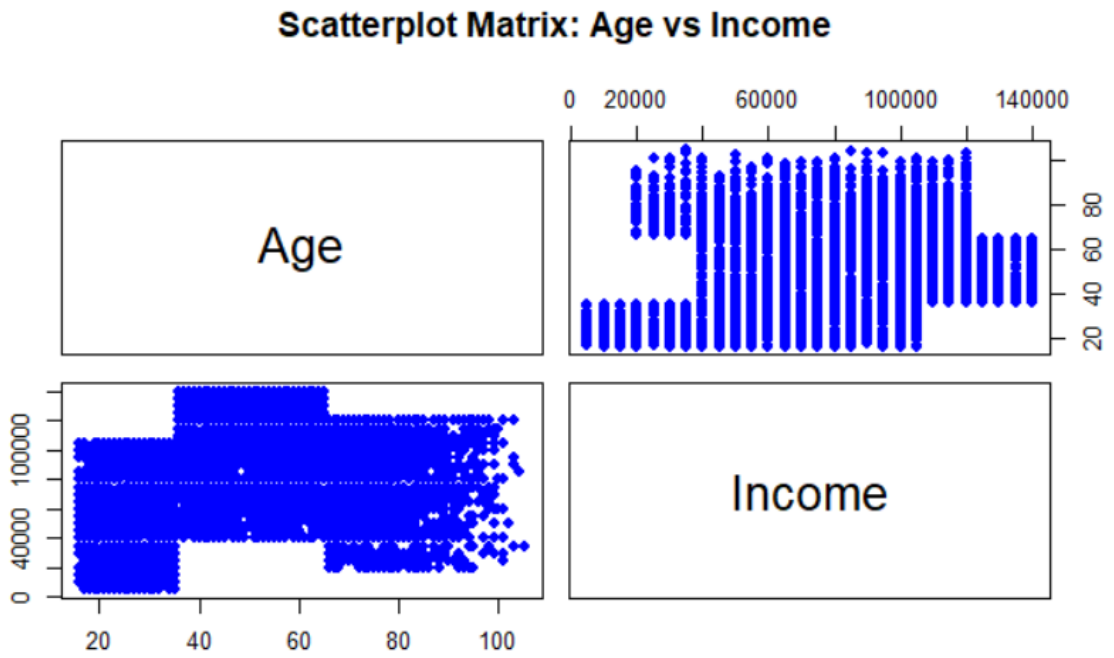


We have the most customers in San Francisco, thus, increasing our presence in San Francisco could be beneficial. We also see that we have the least customers in Miami, thus increasing marketing in Miami might lead to more customers in the future.

6. **Exploring Relationships**

Here we see the relationship between the age and income of our customers, and from this we can see that our market should be the ages of 40 to 60, as these are our

wealthiest customers who are more likely to buy more products.



Products Data

1. **Data Loading and Inspection**

   The Head function was used to allow us to see the structure, dimensions, and variable types. The product's data set includes a ProductID, Category, Description, SellingPrice, and Markup feature for each of its 60 rows of data. The product ID is a mix of characters (3 characters indicating the type of product) and a unique number, making cross-referencing easy.

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
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

2. **Summary Statistics**

   A summary of each feature allows us to see the range and later use that information to create graphs. Here we can see more in-depth data about each feature. We see that ProductID, Category and Description are all of type character and SellingPrice and Markup are decimal values. Here we once again see that there are 60 records.

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

3. **Handling Missing Values**

   This function allows us to make changes to our calculations if missing values are found. Here we see that the dataset contains no missing values, and this allows us to do calculations on the data without having to exclude any records and possibly skewing the outcomes.

```
   ProductID     Category  Description SellingPrice       Markup
           0            0            0            0            0
[1] 0
```

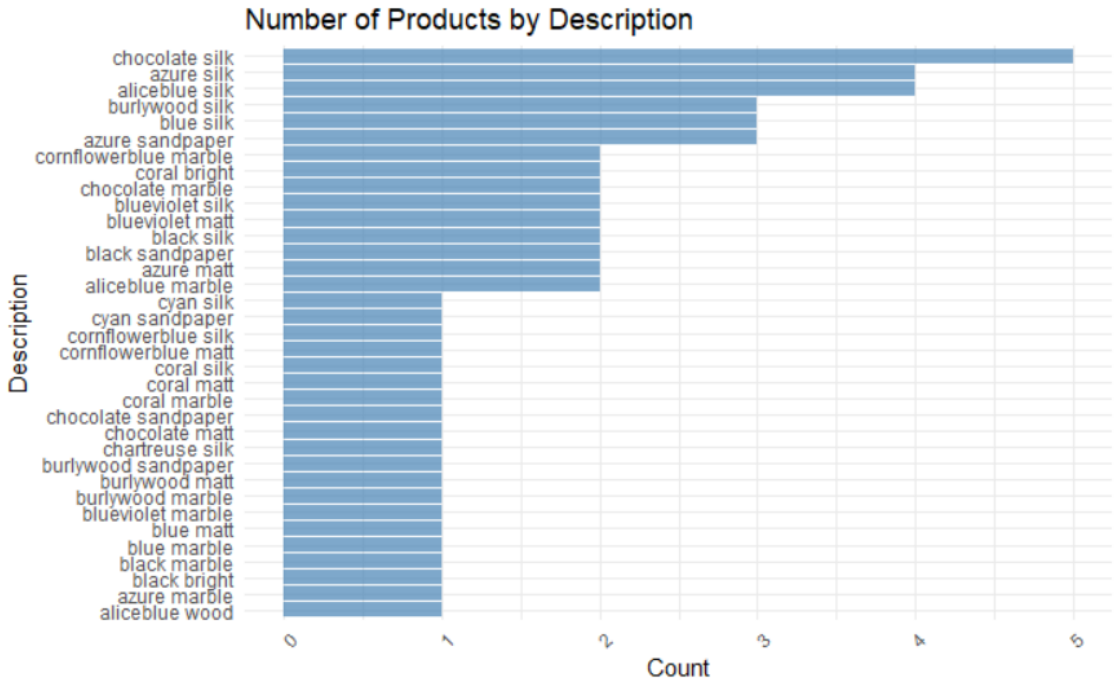
4. **Data Filtering and Subsetting**

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
SOF001	Software	coral matt	511.53	25.05
SOF010	Monitor	chocolate sandpaper	396.72	23.47
CLO014	Cloud Subscription	burlywood silk	1083.11	21.25
CLO015	Laptop	azure silk	728.26	27.70
CLO018	Mouse	chocolate matt	1105.66	20.23
CLO019	Software	aliceblue silk	1092.07	23.14
CLO020	Cloud Subscription	chocolate silk	1128.98	25.48
LAP021	Laptop	black marble	19494.91	20.54
LAP022	Monitor	chocolate marble	16644.21	29.84
LAP024	Mouse	blueviolet marble	18366.92	29.35

1-10 of 31 rows Previous 1 2 3 4 Next

   By applying data filtering, we can see that 50% of our products have a markup of 20% or more. We can also see that in this top 10, we have products of different kinds and selling prices, letting us know there is no overcharging on a specific type of product.

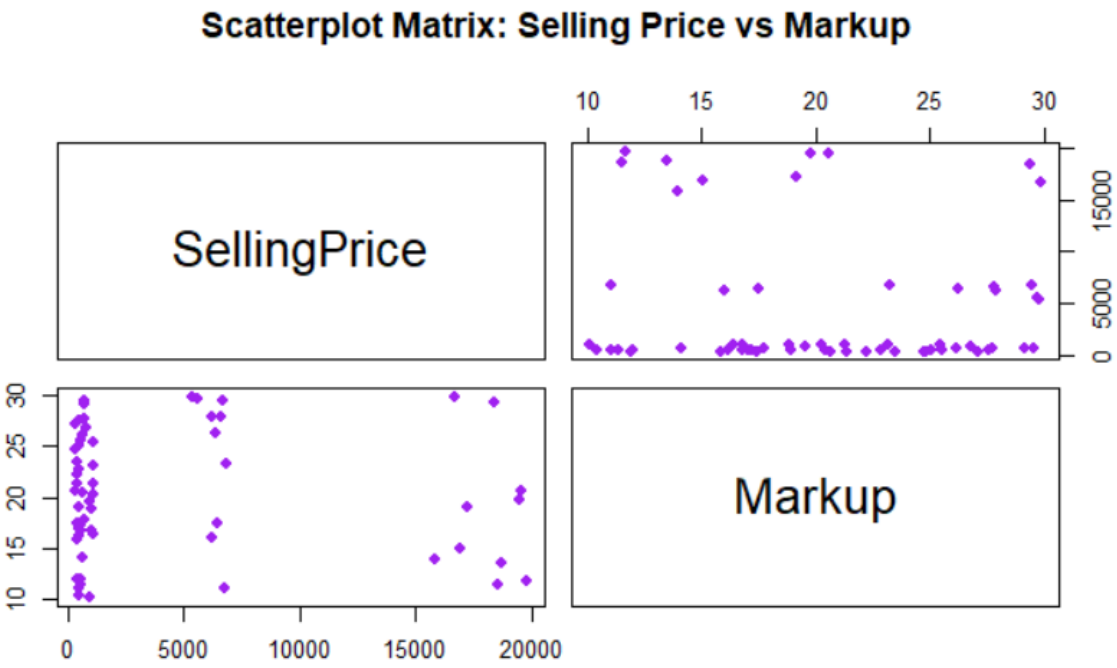
5. **Data Visualization**



   From this graph, we can clearly see that Chocolate silk is our most popular product, with silk products being the most popular category. After silk, most other categories of products only contain one type.

6. **Exploring Relationships**

From this scatterplot, we can see that there seems to be no connection between Selling Price and Markup. This backs up the previous claim of no overpricing of a single category or product.



Products Head Office

1. **Data Loading and Inspection**

   The Head function was used to allow us to see the structure, dimensions, and variable types. The product's head office data set includes a ProductID, Category, Description, SellingPrice, and Markup feature for each of its 360 rows of data. The product ID is a mix of characters (3 characters indicating the type of product) and a unique number, making cross-referencing easy.

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
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

2. **Summary Statistics**

   A summary of each feature allows us to see the range and later use that information to create graphs. Here we can see more in-depth data about each feature. We see that ProductID, Category and Description are all of type character and SellingPrice and Markup are decimal values. Here we once again see that there are 360 records.

```
  ProductID          Category         Description        SellingPrice        Markup      
 Length:360         Length:360         Length:360         Min.   :  290.5   Min.   :10.06  
 Class :character   Class :character   Class :character   1st Qu.:  495.9   1st Qu.:15.84  
 Mode  :character   Mode  :character   Mode  :character   Median :  797.2   Median :20.58  
                                                          Mean   : 4411.0   Mean   :20.39  
                                                          3rd Qu.: 5843.3   3rd Qu.:24.84  
                                                          Max.   :22420.1   Max.   :30.00  
```

3. **Handling Missing Values**

   This function allows us to make changes to our calculations if missing values are found. Here we see that the dataset contains no missing values, and this allows us to do calculations on the data without having to exclude any records and possibly skewing the outcomes.

```
   ProductID     Category  Description SellingPrice       Markup 
           0            0            0            0            0 
[1] 0
```

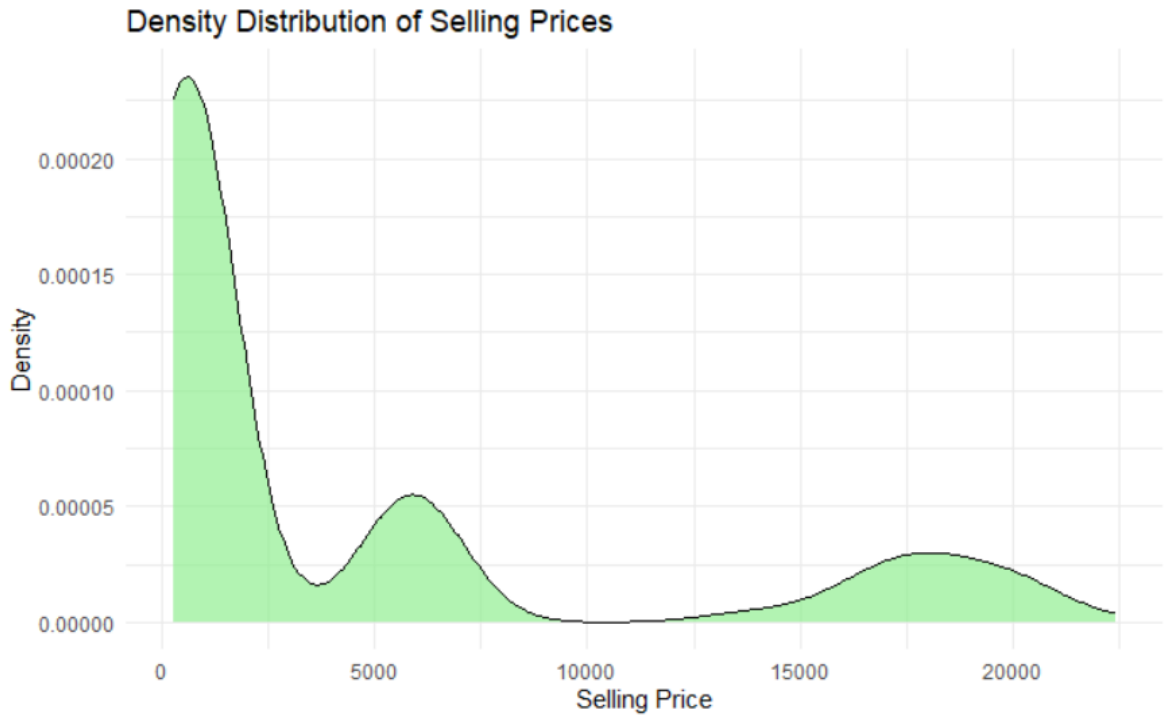
4. **Data Filtering and Subsetting**

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
NA026	Keyboard	blueviolet silk	16569.13	30.00
NA059	Cloud Subscription	cornflowerblue sandpaper	400.32	29.96
NA028	Software	coral silk	17041.57	29.94
LAP008	Laptop	blue silk	472.86	29.94
NA015	Keyboard	black bright	949.88	29.79
NA055	Software	coral sandpaper	432.53	29.78
NA022	Software	cornflowerblue silk	20261.34	29.77
NA056	Mouse	blueviolet sandpaper	425.55	29.73
KEY008	Keyboard	cornflowerblue sandpaper	496.14	29.72
MOU008	Mouse	cornflowerblue sandpaper	435.50	29.72

   By applying data filtering, we can see that some of our highest markups are on silk and sandpaper products. Interestingly enough, most of our products with the highest

markups are not on our most expensive products but rather on our less expensive products.

5. **Data Visualization**

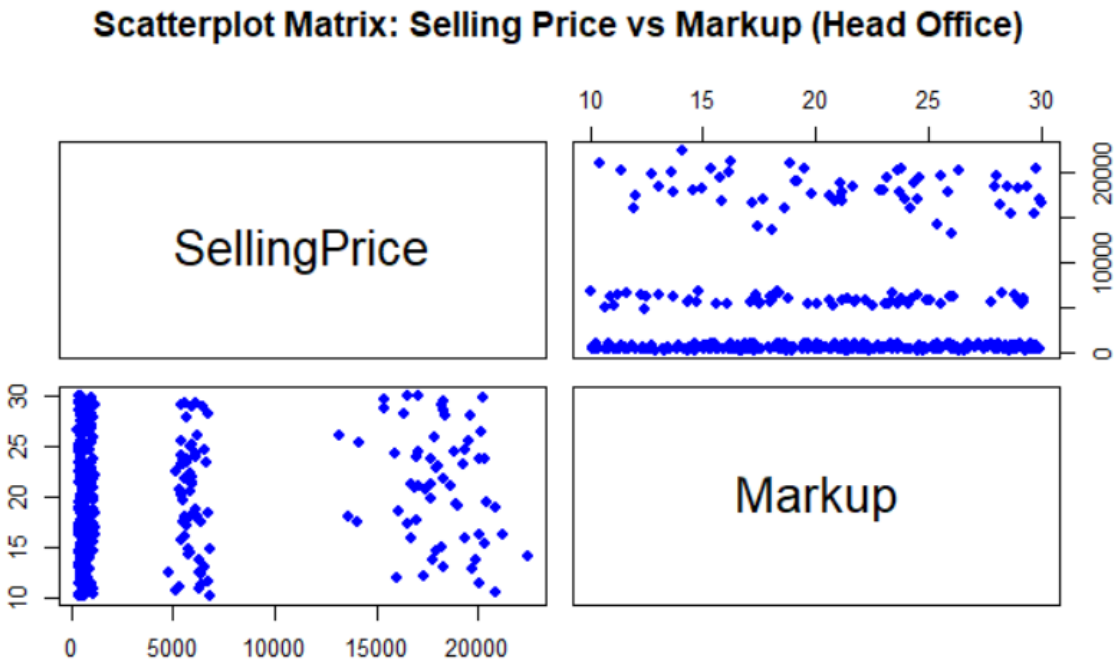


From this graph, we see that the majority of products sold were less than R 3000, meaning almost all of our bulk orders were for products that cost less than R 3000. We can also see that there are 2 other distinct spikes in density at roughly R 6000 and R 17000.

6. **Exploring Relationships**

From this scatterplot, we can see that there seems to be no connection between Selling Price and Markup. This backs up the previous claim of no overpricing of a

single category or product.



Sales 2022 and 2023

1. **Data Loading and Inspection**

   The Head function was used to allow us to see the structure, dimensions, and variable types. The sales 2022 and 2023 data set includes a CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours, and deliveryHours feature for each of its 100 000 rows of data. The product and customer ID is a mix of characters (3 characters indicating the type of product/ if it's a customer) and a unique number, making cross-referencing easy.

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

2. **Summary Statistics**

   A summary of each feature allows us to see the range and later use that information to create graphs. Here we can see more in-depth data about each feature. We see that CustomerID and ProductID are of type character, while Quantity, orderTime, orderDay, orderMonth, and orderYear are of type integer, and pickingHours and deliveryHours are decimal values. Here we once again see that there are 100 000 records.

```
 CustomerID         ProductID          Quantity      orderTime       orderDay      orderMonth      orderYear
Length:100000      Length:100000      Min.   : 1.0   Min.   : 1.00   Min.   : 1.0   Min.   : 1.000   Min.   :2022
Class :character   Class :character   1st Qu.: 3.0   1st Qu.: 9.00   1st Qu.: 8.0   1st Qu.: 4.000   1st Qu.:2022
Mode  :character   Mode  :character   Median : 6.0   Median :13.00   Median :15.0   Median : 6.000   Median :2022
                                      Mean   :13.5   Mean   :12.93   Mean   :15.5   Mean   : 6.448   Mean   :2022
                                      3rd Qu.:23.0   3rd Qu.:17.00   3rd Qu.:23.0   3rd Qu.: 9.000   3rd Qu.:2023
                                      Max.   :50.0   Max.   :23.00   Max.   :30.0   Max.   :12.000   Max.   :2023

 pickingHours     deliveryHours      orderDate
Min.   : 0.4259   Min.   : 0.2772   Min.   :2022-01-01
1st Qu.: 9.3908   1st Qu.:11.5460   1st Qu.:2022-06-17
Median :14.0550   Median :19.5460   Median :2022-11-25
Mean   :14.6955   Mean   :17.4765   Mean   :2022-12-15
3rd Qu.:18.7217   3rd Qu.:25.0440   3rd Qu.:2023-06-16
Max.   :45.0575   Max.   :38.0460   Max.   :2023-12-30
                                    NA's   :560
```

3. **Handling Missing Values**

   This function allows us to make changes to our calculations if missing values are found. Here we see that the dataset contains 560 missing values in the orderDate field. Now that we know this, it allows us to take the necessary steps in removing these records from our calculations, ensuring our results aren't skewed by the missing values.

```
  CustomerID    ProductID     Quantity    orderTime     orderDay   orderMonth    orderYear  pickingHours
           0            0            0            0            0            0            0            0
deliveryHours    orderDate
           0          560
[1] 560
```

4. **Data Filtering and Subsetting**

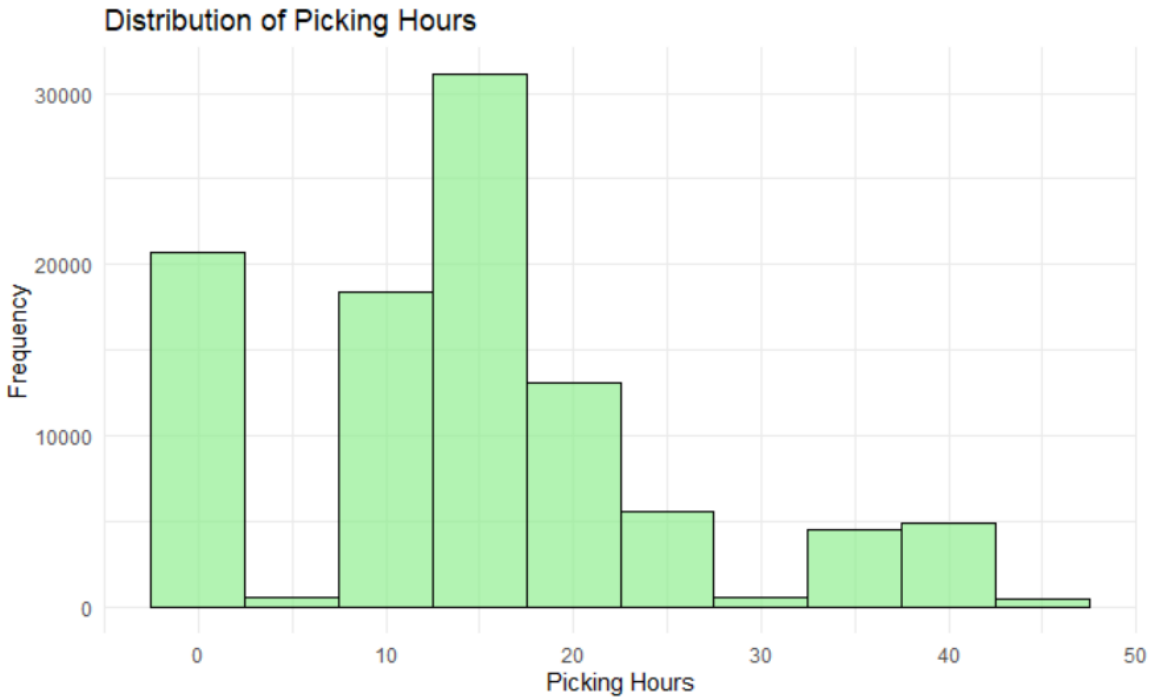
   By applying data filtering, we see that 70% of our orders have fast delivery times, meaning customers get their products within 24 hours.

CustomerID <chr>	ProductID <chr>	Quantity <int>	orderTime <int>	orderDay <int>	orderMonth <int>	orderYear <int>	pickingHours <dbl>	deliveryHours <dbl>	orderDate <date>
CUST1022	KEY046	11	16	23	5	2022	14.7216667	21.5440	2022-05-23
CUST582	MON032	1	19	9	6	2023	17.0575000	22.0460	2023-06-09
CUST4331	KEY049	1	18	30	4	2022	15.3883333	20.0440	2022-04-30
CUST1628	CLO015	5	10	9	8	2023	13.7241667	14.0460	2023-08-09
CUST1501	CLO019	6	9	23	10	2022	20.3883333	13.0440	2022-10-23
CUST3625	MON033	1	9	24	3	2022	23.0550000	17.5440	2022-03-24
CUST574	MOU051	3	3	26	6	2022	9.0550000	22.0440	2022-06-26
CUST4488	CLO019	19	10	8	2	2022	15.7216667	21.0440	2022-02-08
CUST4073	SOF002	1	7	29	2	2022	0.9814444	1.3522	<NA>
CUST2948	MOU053	8	11	16	1	2022	14.3883333	16.5440	2022-01-16

1-10 of 70,253 rows Previous 1 2 3 4 5 6 ... 100 Next

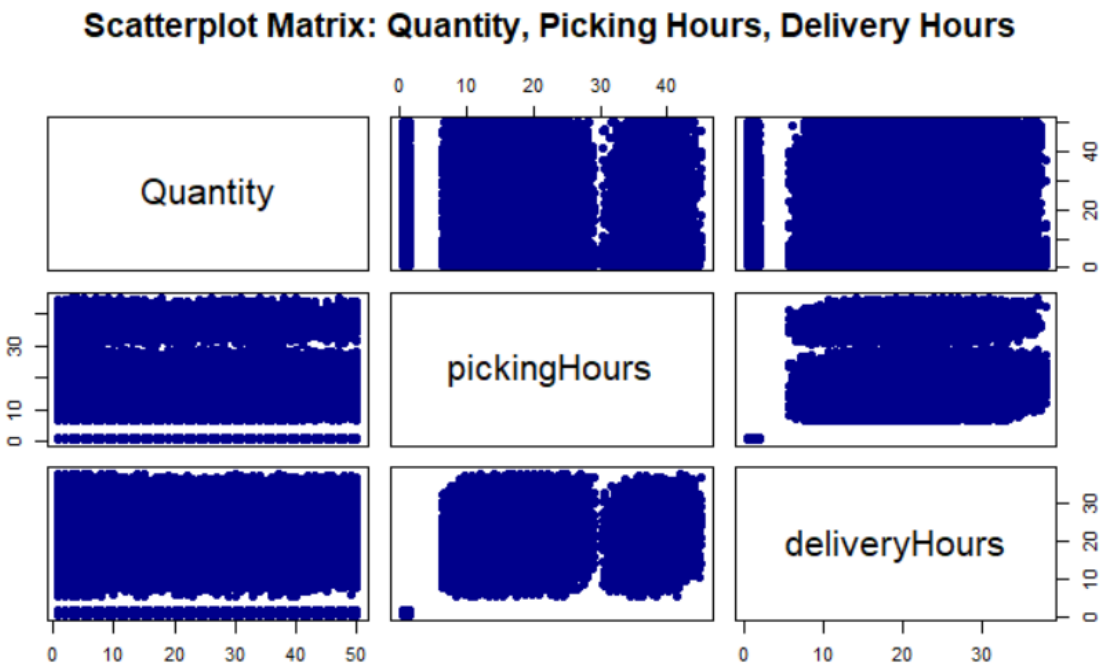
Data Visualization

From this graph, we can see that the favourite picking hours were between 13 and 18.



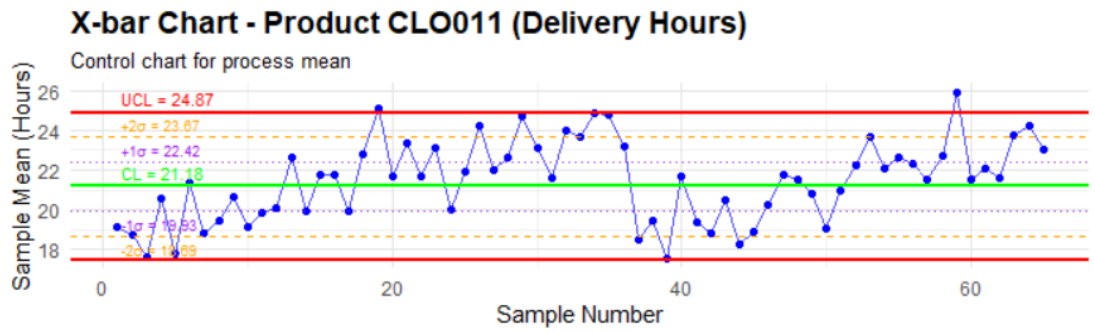
5. **Exploring Relationships**

   From these scatterplots, we can see that there is no clear relationship between Quantity, Picking hours, or Delivery hours.


Scatterplot Matrix: Quantity, Picking Hours, Delivery Hours
Quantity
pickingHours
deliveryHours
0
10
20
30
40
50

Part 3.1: Control charts

In part 3.1, Statistical Process Control (SPC) is applied to each product. Each product was sorted from oldest to newest, where SPC was then applied to the first 720 records. The records were divided into 30 samples of 24 records each, X-bar and S control charts could then be constructed with these samples. With these samples, we were then able to determine centre lines, outer control limits, the 2-sigma control limits and the 1-sigma-control limits for the charts. From the graphs below, we can see that the s-Chart shows that we achieved a low standard deviation in delivery hours, as most points on the graph falls between two standard deviations, showing that these times are very predictable. From the X-bar chart, we see that the graph did not stay as stable, meaning that the average performance is still lacking.

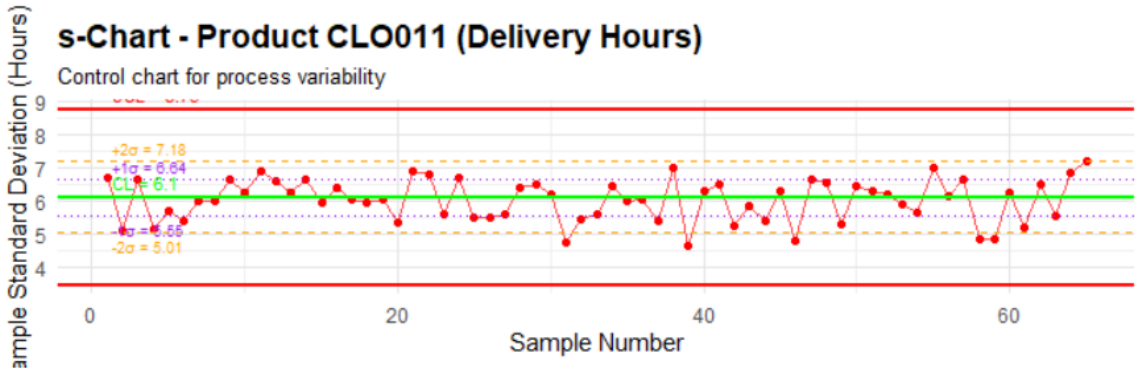


This plot shows the average delivery time for each sample, with the CL being the average for all samples and UCL and LCL being 3 sigma above and below the CL. For CLO011, we see that there are only 2 points outside the control limits.

Centre Line = 21.18

Upper Control Limit = 24.87

Lower Control Limit = 17.48



This plot shows the variance between each sample delivery time, with CL being the average for all samples and UCL and LCL being 3 sigma above and below the CL. For the s-chart of CLO011, we see that there are no points outside the control limits.

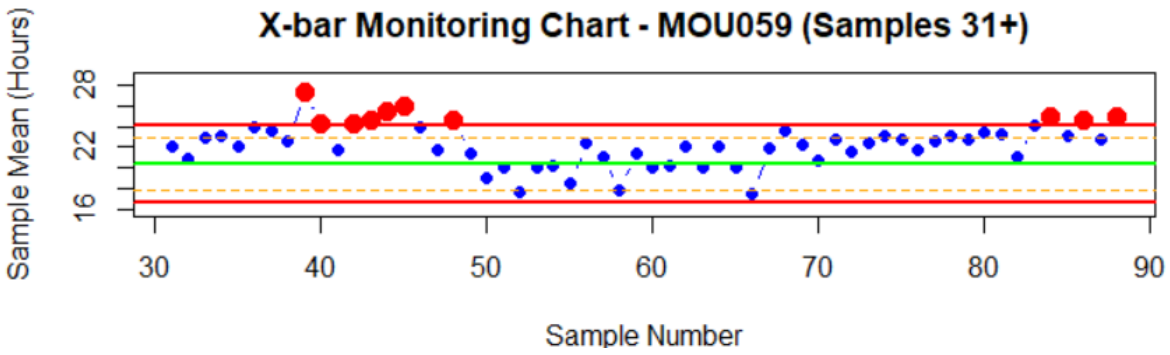
Centre Line = 6.10

Upper Control Limit = 8.75

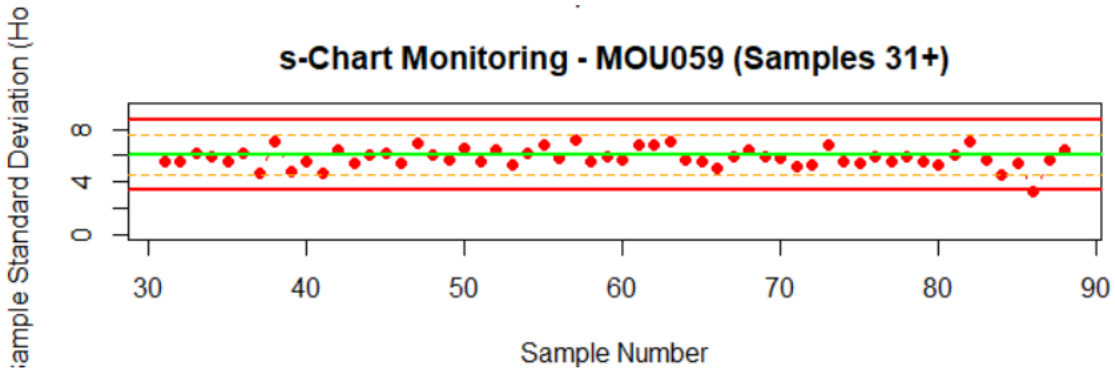
Lower Control Limit = 3.44

3.2 Drawing more samples

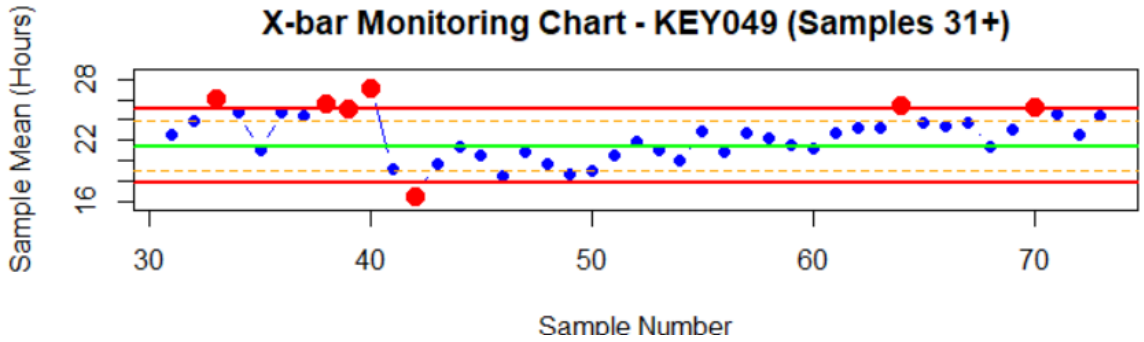
Continuing with the other products and now taking up to 90 samples of each product, we see that some trends emerge. We see from the X-bar charts that there continue to be points outside of the control limits, showing signs that there might be a problem with the delivery methods, since there are problems getting every product out on time. We see the opposite happening on the s-chart, the chart seems to stay very stable for each product, showing that the standard deviation between deliveries is not the problem.



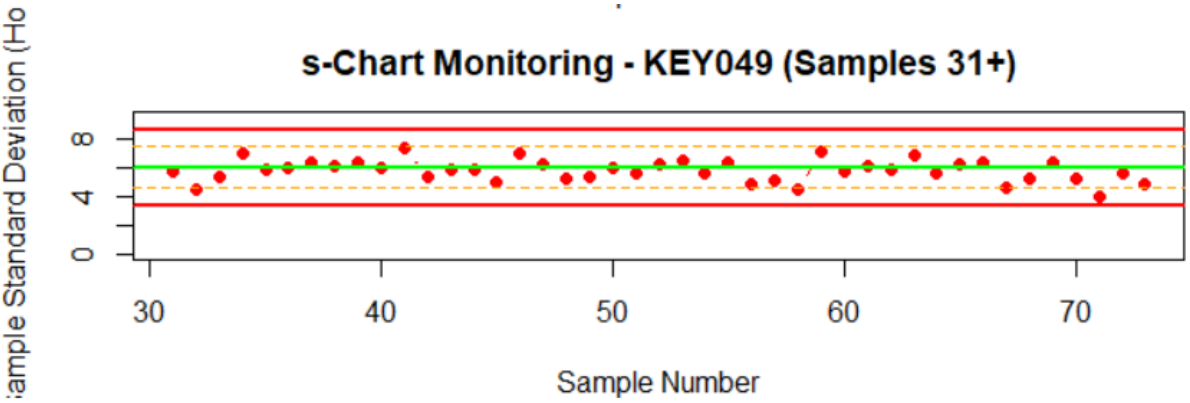
For the x-bar chart of MOU059, we see that there are 10 points outside the control limits, indicating a very unstable average delivery time.



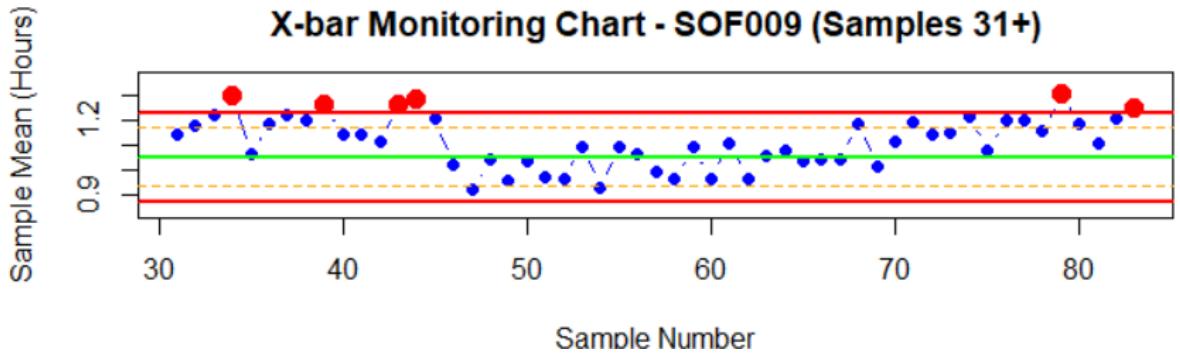
For the s-chart of MOU059 we see that there are no points outside the control limits, indicating a very stable standard deviation in the delivery times.



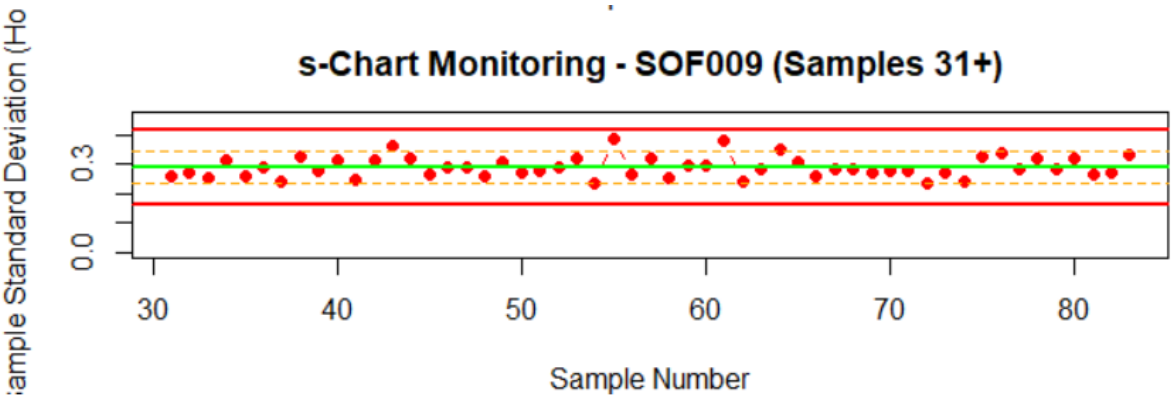
For the x-bar chart of KEY049, we see that there are 7 points outside the control limits, indicating a very unstable average delivery time. Interestingly, for this product, we see one point on the graph below the LCL, which is technically a good thing because that customer got a fast response but for making future predictions we would still like there not to be any points outside of the control limits as it makes the graph more stable and we can then make more accurate predictions from a stable graph.



For the s-chart of KEY049, we see that there are no points outside the control limits, indicating a very stable standard deviation in the delivery times.



For the x-bar chart of SOF009, we see that there are 6 points outside the control limits, indicating a very unstable average delivery time.



For the s-chart of SOF009, we see that there are no points outside the control limits, indicating a very stable standard deviation in the delivery times.

3.3 Process Capabilities

By using the given data, LSL = 0, UCL = 32, and looking at the first 1000 deliveries for each product, we found that none of the products were capable of meeting the Voice of the Customer. Most products didn't even come close to the required Cpk of 1.33 to be considered, furthermore, only 10 products had a Cpk of over 1, were the other 50 were all under 0.7. These shocking numbers tell us that management needs to make immediate changes to the current processes. Investigations need to be put into the causes so that they can be monitored, and solutions can be found.

3.4 Process Control Issues

After going through all the data, we found none of the product data breaking rule A (1 s sample outside of the upper +3 sigma-control limits for all product types).

For rule B, we had to find the most consecutive samples of s between the -1 and +1 sigma-control limits for each product type and found: SOF002 – 16, LAP025 – 15, CLO020 – 12, MON035 – 12, MOU052 – 12 consecutive instances between -1 and +1 sigma. This high number of consecutive instances indicates a great deal of control.

Rule C, which tested if any products had any 4 consecutive X-bar samples outside of the upper, second control limits. After running some tests, we found that 35 products did not adhere to this rule. These products with the amount of rule breakings were: MOU053 – 20, MOU055 – 19, SOF022 – 18 and the last 3 on the list were MOU056 – 4, MON039 – 4, LAP027 – 4

4.1 Estimate the likelihood of making a Type I Error

The probability (1 sample > centreline) = 0.5, this is because we are working with a normal distribution. In a normal distribution, by definition, the mean must be 0 and the standard deviation 1. If we then look at the Z score of 0, we see that the probability is 0.5.

A type 1 error is defined as the incorrect rejection of a true null hypothesis, also known as a false positive.

Rule A: 1 s sample outside of the upper +3 sigma-control limits for all product types.

UCL = $\mu+3\sigma = 0 + 3(1) = 3$

Thus $P(Z>3) = 1-0.99865 = 0.00135$

Rule B: Find the most consecutive samples of s between the -1 and +1 sigma-control limits for all product types.

The probability of one point being between -1 and +1 sigma is, $P(-1<Z<1) = 0.841345 – 0.158655 =0.68269$

Thus, the probability of 16, the highest number of consecutive samples found in product SOF002, is 0.68269^16 = 0.002226

Rule C: 4 consecutive X-bar samples outside of the upper, second control limits for all product types.

The probability of one point being between -2 and +2 sigma is, $P(-2<Z<2) = 0.97725 – 0.02275 =0.9545$

Thus, the probability of a point falling outside of the upper second control limit is = (1 – 0.9545)/2 = 0.02275, and the probability of 4 consecutive samples outside the upper second control limit is = 0.02275^4 = 0.26787 x 10^ (-6). From these values, we can conclude that there is a very low chance of a type I error occurring, and that they give a high confidence that the current rules are effective at spotting process shifts without the need for unnecessary interventions.

4.2 Estimate the likelihood of making Type II Errors

H_0:

$\mu_0 = 25.05$

$\sigma_0 = 0.013$

H_a:

$\mu_1 = 25.028$

$\sigma_1 = 0.017$

$\beta = P(LCL < \bar{X} < UCL \mid \text{True Mean} = \mu_1 \text{ and True Std Dev} = \sigma_1)$

$P(25.011 < X < 25.089) = P((25.011-25.028)/0.017 < Z < (25.089-25.028)/0.017) = P(-1<Z<3.59) = 0.999835-0.158655$

$=0.84118$

Thus, the estimated likelihood of a type 2 error is 84.12%. This means that the system only has a 15.88% chance of detecting process degradation and is a clear indicator that the current monitoring system has to be made more sensitive, as the current one can easily allow quality issues to go unnoticed. We would strongly recommend adding more rules or using tighter control limits.

4.3 Correcting Errors and Data Analysis

Products Head Office

1. Data Loading and Inspection

Old

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
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

New

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
1	SOF001	Software	coral silk	511.53	25.05
2	SOF002	Software	black silk	505.26	10.43
3	SOF003	Software	burlywood marble	493.69	16.18
4	SOF004	Software	black marble	542.56	17.19
5	SOF005	Software	chartreuse sandpaper	516.15	11.01
6	SOF006	Software	cornflowerblue marble	478.93	16.99

Here we can see that the updated prices and markups have been implemented. We can see that all the values in the table have vastly different selling prices and markups after the data has been updated.

2. Summary Statistics

Old

```
  ProductID            Category         Description         SellingPrice        Markup     
 Length:360         Length:360         Length:360         Min.   :  290.5   Min.   :10.06  
 Class :character   Class :character   Class :character   1st Qu.:  495.9   1st Qu.:15.84  
 Mode  :character   Mode  :character   Mode  :character   Median :  797.2   Median :20.58  
                                                          Mean   : 4411.0   Mean   :20.39  
                                                          3rd Qu.: 5843.3   3rd Qu.:24.84  
                                                          Max.   :22420.1   Max.   :30.00  
```

New

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

The big difference we see here is that the size of the file is now only 60 records instead of the previous 360. This has also affected other parts of the file, like min, max, median, and mean. Yet the structure of the file has stayed the same, as there were no features added or removed from the data set. If we look at the changes for selling price and markup, we see that markup had very small changes. If we look at the selling price, we can see some big changes that happened, like min going from 290.5 to 350.4 and max going from 22 420.1 to 19 725.2. These changes are most likely because of the loss of data after the dataset was updated.

3. Handling Missing Values

Old

```
   ProductID     Category  Description SellingPrice       Markup
           0            0            0            0            0
[1] 0
```

New

```
   ProductID     Category  Description SellingPrice       Markup
           0            0            0            0            0
```

We can see that the missing values show no difference, the file still has no missing values.

4. Data Filtering and Subsetting

Old

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
NA026	Keyboard	blueviolet silk	16569.13	30.00
NA059	Cloud Subscription	cornflowerblue sandpaper	400.32	29.96
NA028	Software	coral silk	17041.57	29.94
LAP008	Laptop	blue silk	472.86	29.94
NA015	Keyboard	black bright	949.88	29.79
NA055	Software	coral sandpaper	432.53	29.78
NA022	Software	cornflowerblue silk	20261.34	29.77
NA056	Mouse	blueviolet sandpaper	425.55	29.73
KEY008	Keyboard	cornflowerblue sandpaper	496.14	29.72
MOU008	Mouse	cornflowerblue sandpaper	435.50	29.72

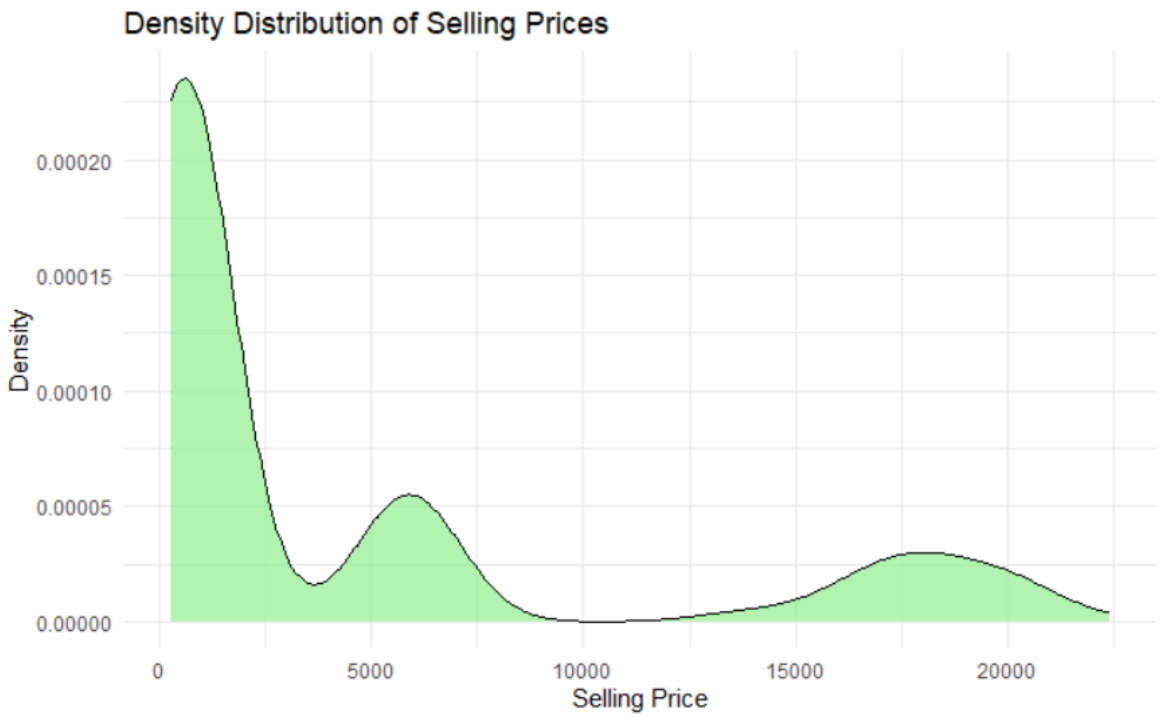
New

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
LAP002	Laptop	black bright	16644.21	29.84
MON010	Monitor	azure silk	5346.14	29.74
MON003	Monitor	black sandpaper	5572.82	29.72
KEY007	Keyboard	aliceblue matt	693.24	29.53
MON009	Monitor	chocolate matt	6711.03	29.50
LAP004	Laptop	azure matt	18366.92	29.35
KEY009	Keyboard	cornflowerblue bright	752.75	29.11
MON006	Monitor	blueviolet marble	6192.01	27.92
MON002	Monitor	blue marble	6634.13	27.80
CLO005	Cloud Subscription	chocolate marble	728.26	27.70

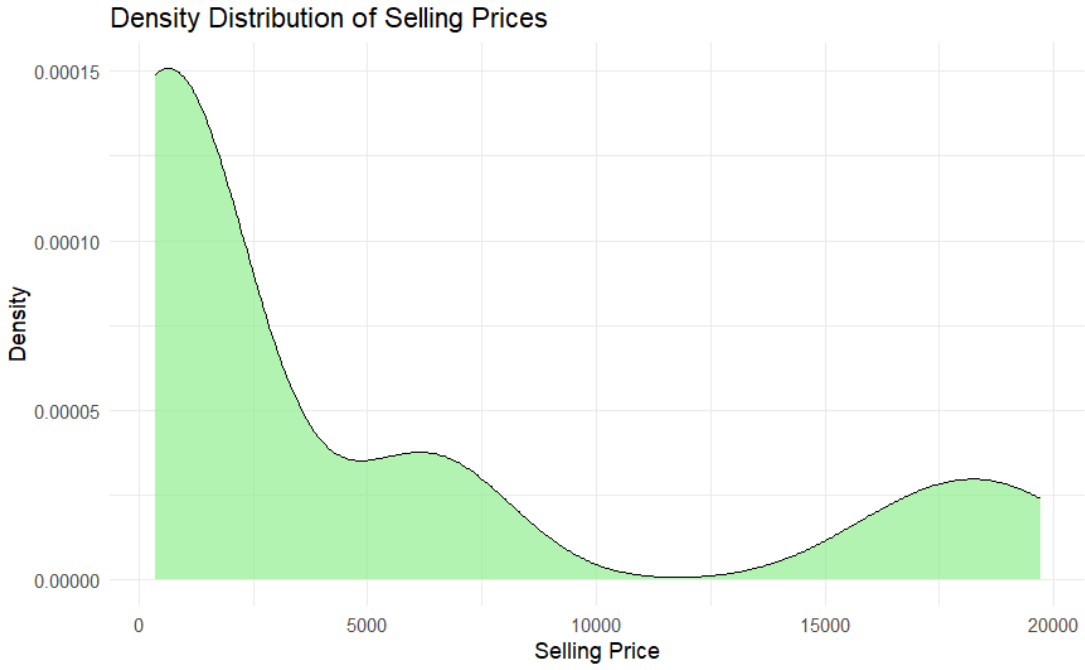
Here we see more evidence of the updated values as NA026 is no longer the product with the highest markup and has been replaced by LAP002, which has a markup of only 29.84. Similar differences can be seen throughout the top 10 highest markups.

5. Data Visualization

Old



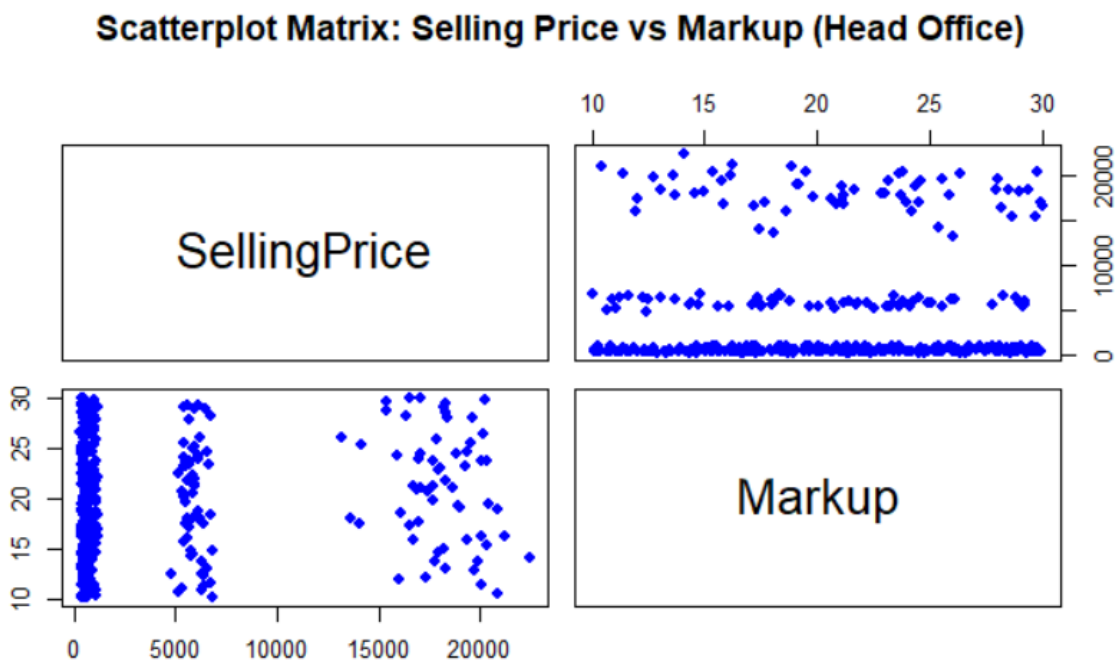
New



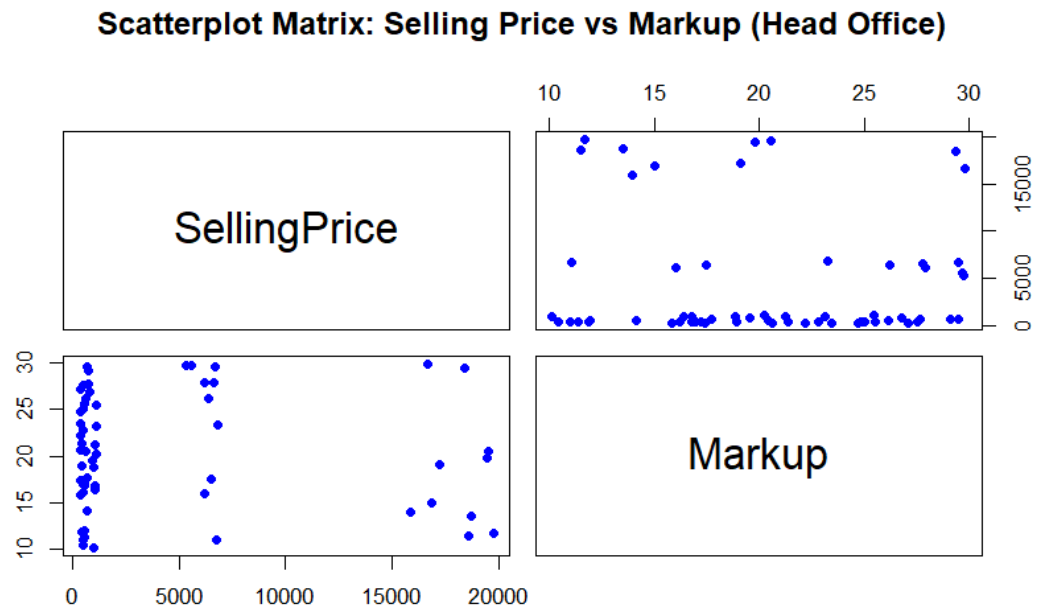
Due to the lower number of records, we see that the density distribution has also changed. The new highest density is roughly 0.00015 compared to the previous 0.00023. The new graph also has a lower highest selling price at 20000. We see that the new graph is less volatile, most likely due to the loss of data from the dataset updating.

6. Exploring Relationships

Old



New



Here we once again see the effects of the lower number of records, but the correlations seem to have stayed the same.

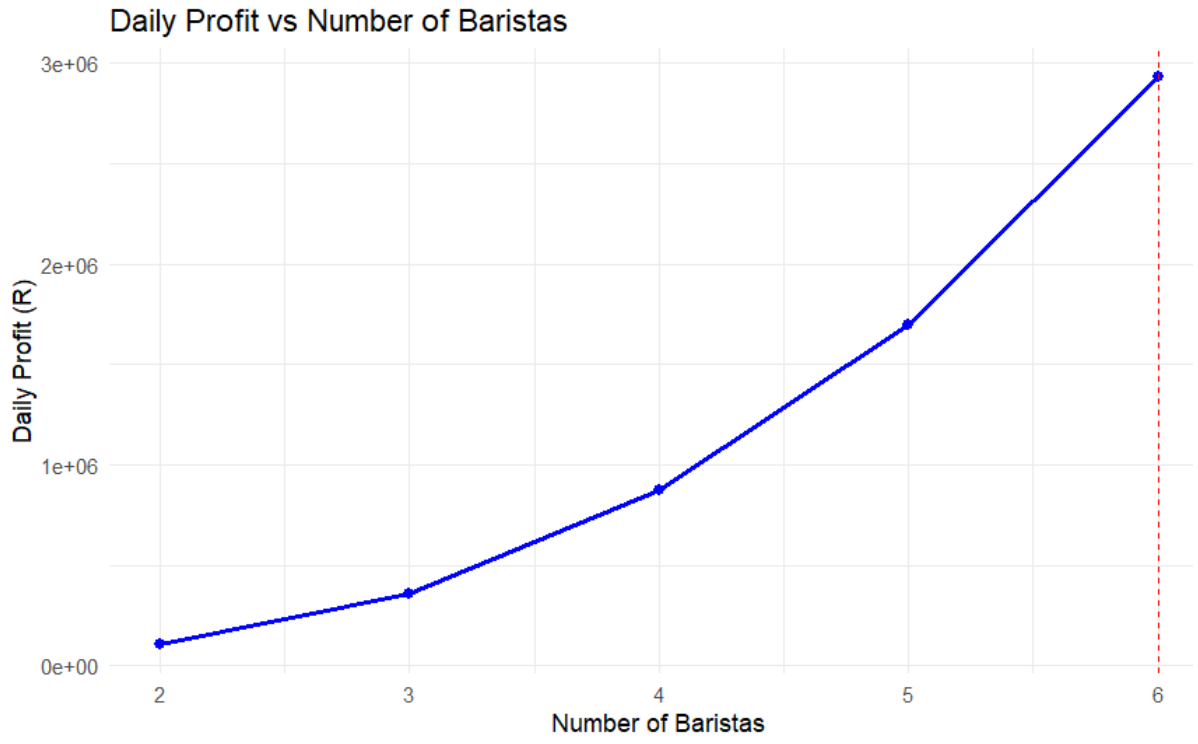
Category <chr>	TotalSalesValue <dbl>
Laptop	1163889479
Monitor	578385570
Cloud Subscription	98715482
Keyboard	73499067
Software	66468485
Mouse	51219577

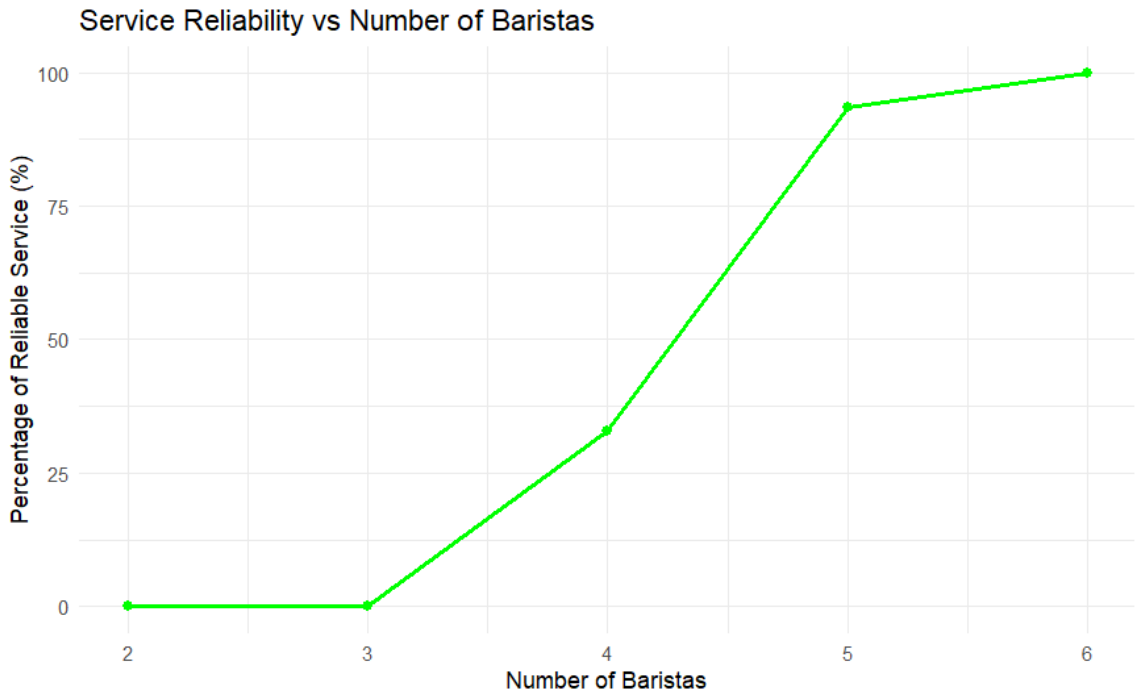
Here we can see the total sales value made in 2023 from each category, with laptops and monitors being our biggest sellers. These two categories make up 85% of our total sales value, with laptops being 57% of our total sales value. From this, we can conclude that most of our marketing budget should be spent on these two products.

5. Optimise profit

5.1 Time To Serve

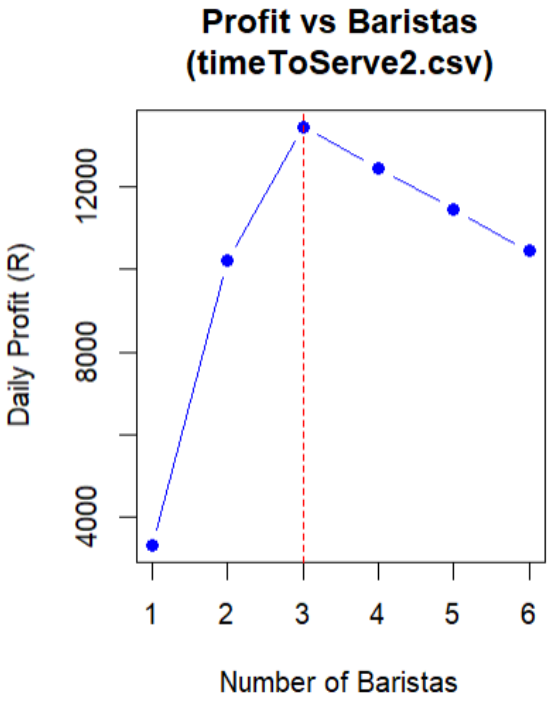
From the previous data analyst's note, we know there are problems if there are less than 2 baristas, so we can automatically eliminate 1 barista being an optimal answer. From the data, we were able to make 2 graphs, each seeming to lead to the same conclusion, the more baristas we add, the better the business will do. By adding baristas, we can clearly see that it increases the daily profit, increases the service reliability, and decreases the average service time. From the model suggested by the previous data analyst, we found that the optimal number of baristas to have is 6. With 6 baristas, the expected daily revenue will be R 2 936 850 and a daily profit of R 2 930 850. The expected service reliability will be 99.9% with an average service time of 33.4 seconds.

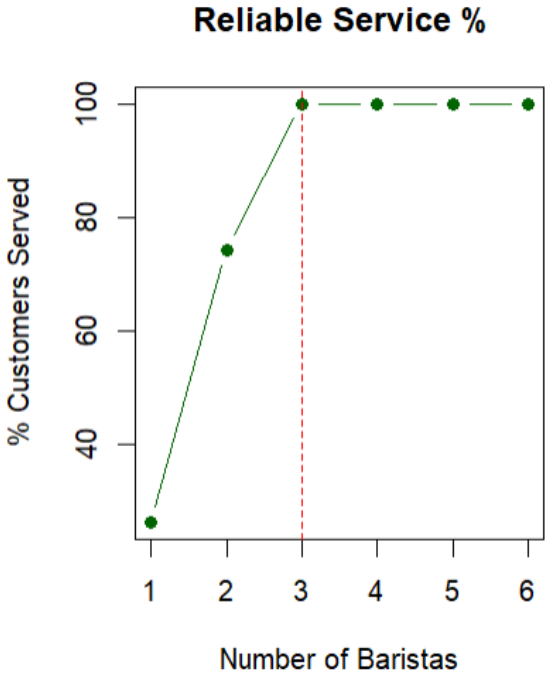




5.2 Time To Serve 2

Here we have the outcomes for the corrected dataset, timeToServe2. From the previous data analyst's note, we know there are problems if there are less than 2 baristas, so we can automatically eliminate 1 barista being an optimal answer. From the data, we were able to make 2 graphs, both leading to the same conclusion, three baristas will give the optimal profit. We see that by adding baristas, there is a sharp increase in profit until 3 baristas, where the profit peaks, and after that, there is a slow decrease in profit. The same can be seen at reliable service, service increases to 100% at three baristas and then stays there, the more you add, indicating no real advantages for adding additional baristas. With 3 baristas, the expected daily profit will be R 13 438.36, the expected service reliability will be 100%.


Profit vs Baristas
(timeToServe2.csv)
Daily Profit (R)
4000
8000
12000
1
2
3
4
5
6
Number of Baristas


Reliable Service %
% Customers Served
40
60
80
100
1
2
3
4
5
6
Number of Baristas

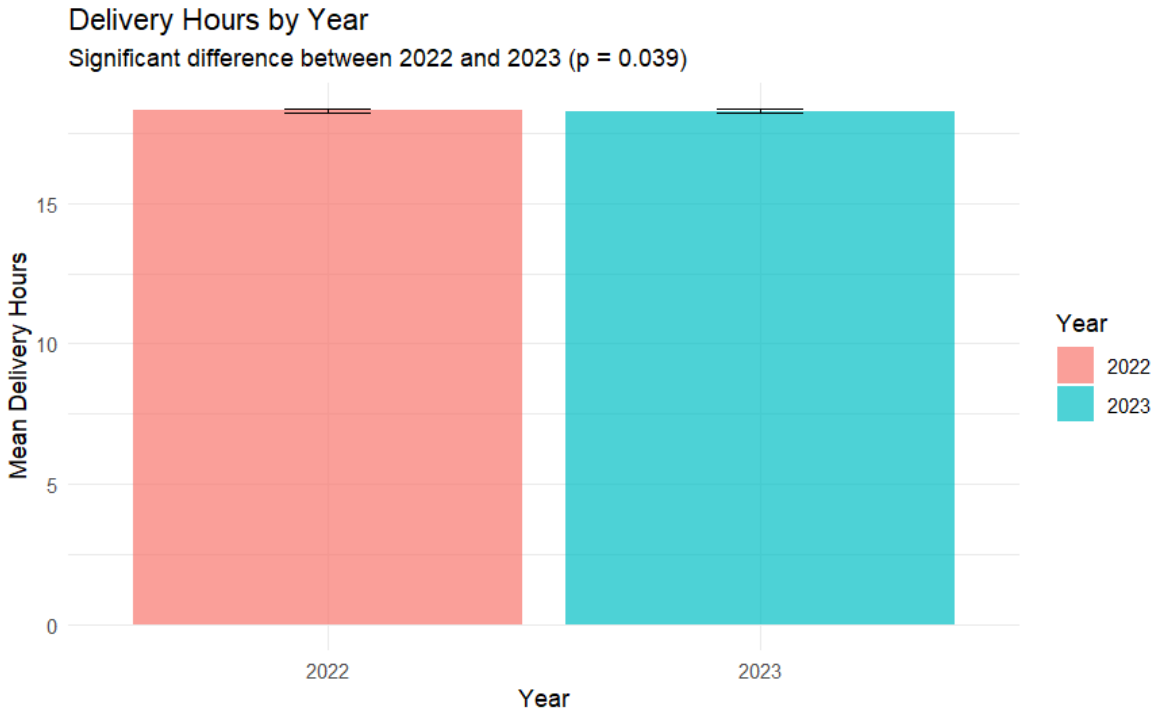
6. DOE and MANOVA or ANOVA

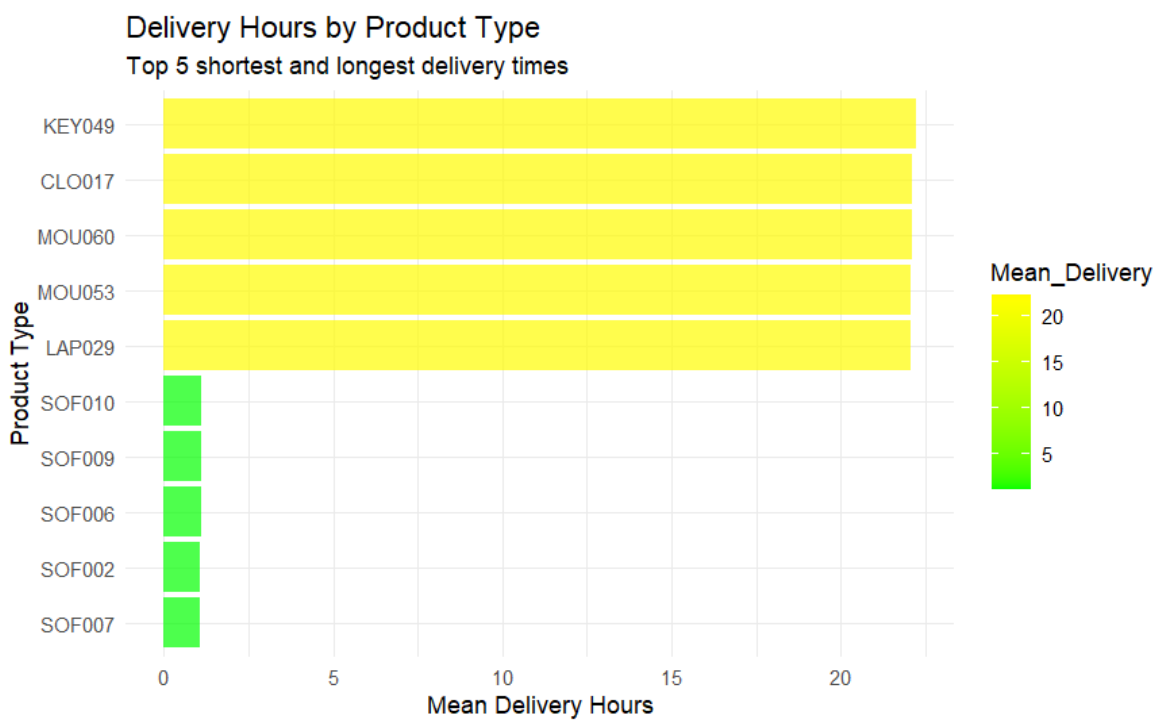
6.1 MANOVA Analysis

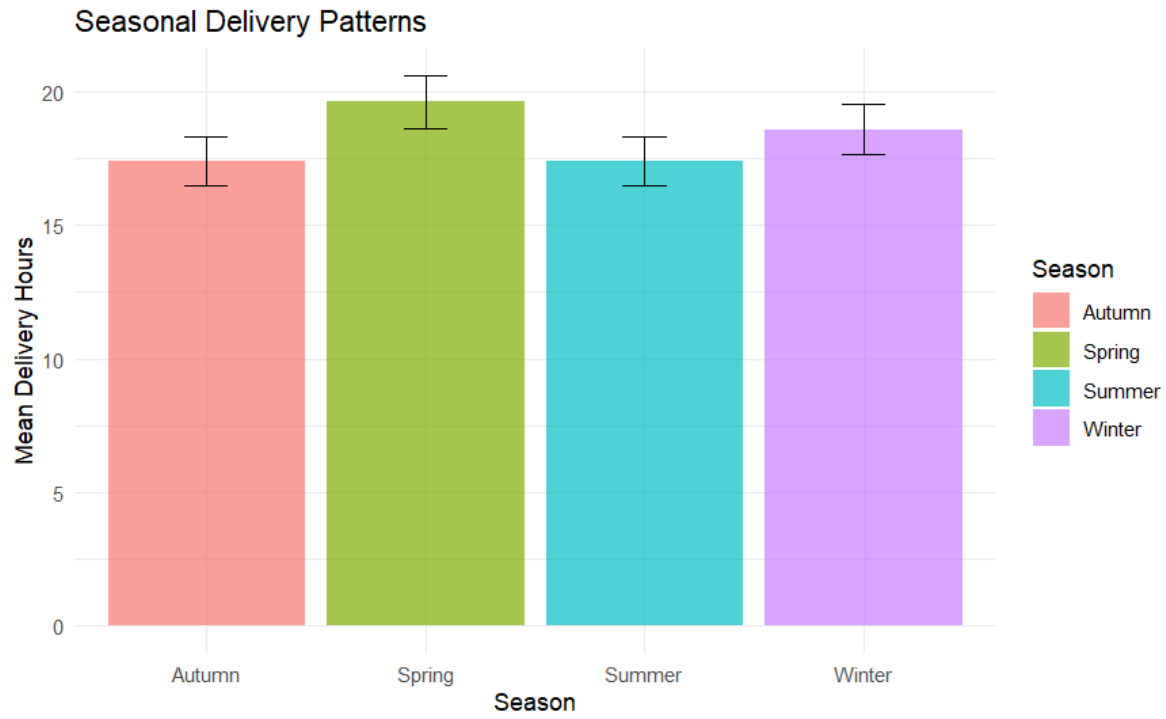
MANOVA has been chosen over ANOVA as we are working with multiple interrelated dependent variables (delivery hours, picking hours, and quantity). With this method, we can better understand their collective relationships with product types. Whereas ANOVA can only examine each response separately.

6.2 Time-Based ANOVA Analyses

After analysing delivery performance data from 2022 to 2023, key patterns emerged that can be used for understanding and improving future data. The MANOVA and ANOVA analyses show that the biggest factor in delivery times is product type (shown in Delivery hours by product type graph), with highly significant differences across all models ($p < 0.001$). While looking at the year-to-year changes, no significant differences ($p = 0.724$) appeared (shown in delivery hours by year graph), indicating stability and strong seasonal patterns (shown in seasonal delivery patterns graph). Spring demonstrates the longest mean delivery times at 19.62 hours, significantly higher than other seasons, particularly Summer and Autumn, which show the shortest delivery times at approximately 17.39-17.40 hours. This seasonal variation is further reinforced by significant monthly patterns ($p < 0.001$) and a notable time period effect ($p = 0.004$), where the second half of the dataset shows different performance characteristics than the first half.




Delivery Hours by Product Type
Top 5 shortest and longest delivery times
KEY049
CLO017
MOU060
MOU053
LAP029
SOF010
SOF009
SOF006
SOF002
SOF007
Product Type
0
5
10
15
20
Mean Delivery Hours
Mean_Delivery
20
15
10
5


Seasonal Delivery Patterns
20
15
10
5
0
Mean Delivery Hours
Autumn
Spring
Summer
Winter
Season
Season
Autumn
Spring
Summer
Winter

7. Reliability of service

7.1 Reliable days of service per year

Reliable days per year = 365 * total reliable days / total days

= 365 * (96+270)/(397)

= 336.5 ~ 337 ~ 92.2%

From the calculations above, we see that with the company's current layout of 16 workers, it is already very reliable. In the next part, we will do more calculations to find out if there is a way to further improve this result.

7.2 Optimise the profit for the company

From the information provided, we can already eliminate any values below 15 for workers present, as this results in a R20 000 loss in sales for that day. Using the value calculated in 7.1, we obtain a p-value of 0.97375. We then employ the binomial distribution to develop a model for worker attendance. Through this model, we can optimise the results by testing various staffing levels and comparing the reduction in lost sales against the additional personnel costs. This allows us to determine the N (number of personnel) that maximises the annual net savings. In this case, the optimal number of workers is 17, which yields an annual savings of R180 513 compared to the current 16 workers.

Conclusion

From the statistical analysis done in this report, raw data was converted into actionable information that can improve business performance. Some parts of the report found that urgent intervention from management will be needed, while other parts showed strengths and opportunities for growth.

The most pressing issue identified by the control charts demonstrated an unstable average delivery time across numerous products, and process capability analysis confirmed that none of the products meet customer specifications, with Cpk values significantly below the acceptable threshold. This failure presents a substantial risk to customer satisfaction and requires urgent process redesign.

On the other hand, the analysis of service operations yielded positive and actionable insights. The staffing model determined that three baristas would maximise daily profit while ensuring a 100% service reliability. Furthermore, the model for worker attendance indicated that increasing personnel from 16 to 17 would result in a significant increase in annual savings.

Data integrity was another key focus. The correction of pricing and markup data in the head office products file has ensured the accuracy of financial reporting and sales analysis, which correctly identifies laptops and monitors as the primary drivers of revenue.

The MANOVA and ANOVA tests pinpointed product type and seasonal trends, particularly slower spring deliveries, thus this gives us a clear objective of improving productivity year-round.

In short, if we want to increase efficiency and profitability, management should prioritise a redesign of the delivery logistics, implement the optimised staffing models for service operations, and make use of the insights gathered on seasonal and product-based variations for strategic planning. By acting on these data-driven recommendations, the company can significantly improve its operational reliability and gain a competitive edge.

References:

- Datasets provided by course (Quality assurance 344)