ECSA Report

Name: Daniel Kalell

Student Number: 26177099

Module: Quality Assurance 344

Index

Part 1.2:

Customer Data:

Data loading and Inspection:

Nothing out of the ordinary in first inspection, 5 x 5000 rows with a mixture of numeric and categorical data.

Summary Statistic:

```
  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
```

This summary data can be used to compare any of the data instances to the mean and categorise it within which quartile each instance lies.

```
Female   Male  Other
  2432   2350    218
> table(customerdata$City)

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

We can use this data to see that there is not one dominating gender or city, the genders and cities all have very similar amounts of data respectively. This means the data is not skewed and will not have a bias towards any specific gender or city.

Handling Missing Values:

There are no missing values in this dataset.

Data Filtering and Subsetting:

We can filter the data by age and income:

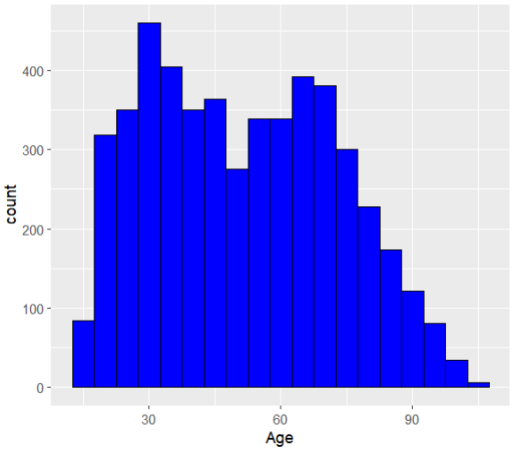
We find that there are 1850 customers over the age of 60 and 922 customers under the age of 30. This shows that our customers are more middle aged to elderly, this could urge the company to try target the youth and tap more into a younger customer base.

Chicago	Huston	Los Angeles	Miami	New York	San Fansisco	Seattle
221	209	208	203	207	228	191

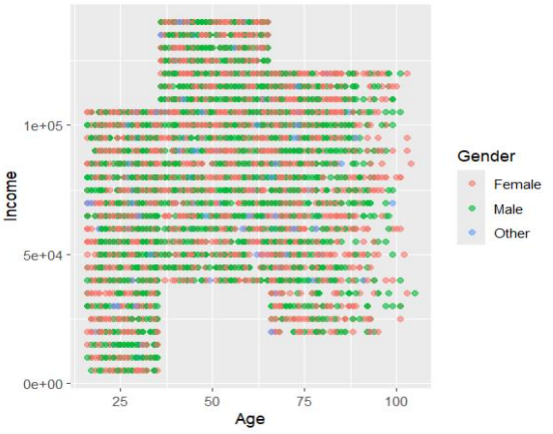
The table above shows the number of customers with an income above 100 000 per city. This shows that there is a relatively even number of high-income customers per city.

Data Visualisation:

The histogram to the right shows that our customer base is bimodal with peaks around ages 30 and 70. But as we suspected there is a big drop off below the age of 30. Unusually there is a big dip between the ages of 40 – 50, this must be investigated as there should not be such a big dip within this age bracket.



The graph to the right shows the different genders at different ages vs income. This graph shows us that gender and income have no correlation. It does show us that the highest income earners are between the ages of 40 – 60. Using the histogram above we can conclude that we are not targeting the highest income individuals, and this is somewhere the company could grow.



Exploring Relationships:

Using the data the calculated correlation between age and income is 0.15 which is low and should not be used as the main indicator for targeting customers.

Product Data / Product Head Office Data:

Data loading and Inspection:

The head office dataset has 360 rows and 5 columns wile the product data has 60 rows and 5 columns, both have categorical columns for ProductID, Category and Description and numeric for SellingPice and Markup.

Summary Statistic:

	Category	Mean_SellingPrice	SD_SellingPrice	Mean_Markup	SD_Markup
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Cloud Subscription	4387.	6366.	21.5	5.77
2	Keyboard	4380.	6469.	20.0	5.65
3	Laptop	4306.	6242.	20.5	5.29
4	Monitor	4457.	6607.	19.4	5.95
5	Mouse	4479.	6681.	20.2	5.74
6	Software	4457.	6677.	20.8	5.58

The table above shows the different mean values for all the categories within the head office data.

	Category	Mean_SellingPrice	SD_SellingPrice	Mean_Markup	SD_Markup
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Cloud Subscription	3692.	5812.	20.6	6.67
2	Keyboard	4638.	7142.	20.2	5.72
3	Laptop	5218.	7315.	20.6	6.57
4	Monitor	5014.	6984.	20.7	6.06
5	Mouse	4585.	7095.	20.7	7.08
6	Software	3814.	6144.	20.0	5.84

The table above shows the different mean values for all the categories within the products data. If we compare the two tables we can see significant differences in the selling prices, this must be investigated within the company as the head office seems to be selling items at a reduced cost as opposed to the product data.

Handling Missing Values:

There are no missing values within either dataset.

Data Filtering and Subsetting:

I chose to filter the data with only common products between the two datasets. The table below is all common products within the head office dataset.

ProductID	Category	Description	SellingPrice	Markup
<chr>	<chr>	<chr>	<dbl>	<dbl>
SOF001	Software	coral silk	522.	15.6
SOF002	Software	black silk	467.	28.4
SOF003	Software	burlywood marble	496.	20.1
SOF004	Software	black marble	389.	17.2
SOF005	Software	chartreuse sandpaper	483.	17.6
SOF006	Software	cornflowerblue marble	539.	25.6

The table below shows the common products from the products dataset.

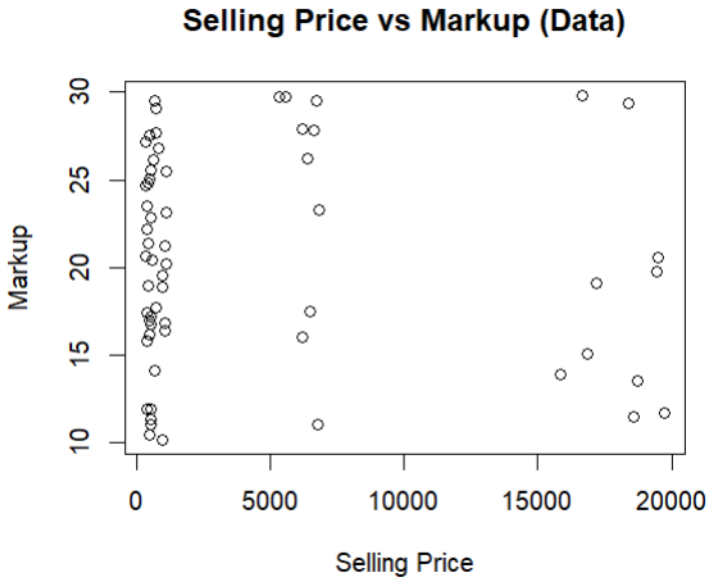
ProductID	Category	Description	SellingPrice	Markup
<chr>	<chr>	<chr>	<dbl>	<dbl>
SOF001	Software	coral matt	512.	25.0
SOF002	Cloud Subscription	cyan silk	505.	10.4
SOF003	Laptop	burlywood marble	494.	16.2
SOF004	Monitor	blue silk	543.	17.2
SOF005	Keyboard	aliceblue wood	516.	11.0
SOF006	Mouse	black silk	479.	17.0

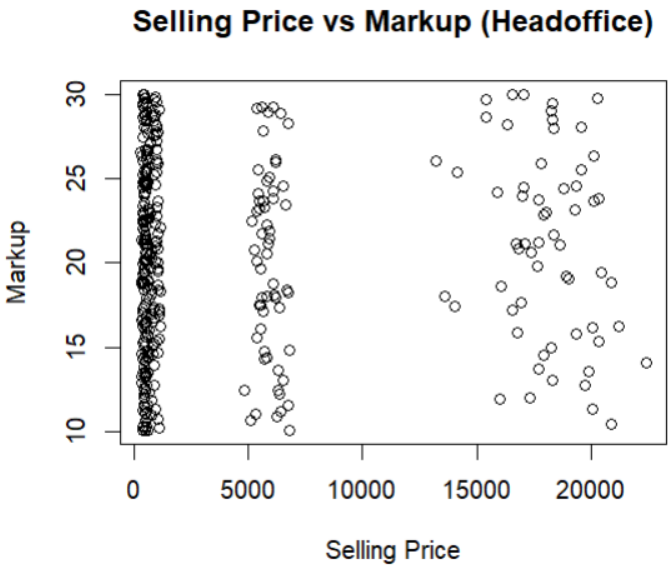
Inconsistencies within the Data:

The two datasets share ten common products, all ten show inconsistencies; differing categories (eg SOF002 is classified as Software in the head office dataset but is classified as Cloud Subscription in the products dataset), differing Descriptions (eg, ‘black silk’ as opposed to ‘cyan silk’), differing selling price and differing markups. In addition to this the head office dataset has duplicate product ID’s across categories

This problem is most likely due to data entry errors or merging issues / communication issues between the head office and the place where the other products are sold. This is a big problem that must be addressed as it will cause more errors if customers are trying to return products or if a certain product is trying to be located.

Data Visualisation:





Even though the means of selling price and markup are different for the head office and the product data once the values have been plotted, we can see they are fairly similar. The reason for the difference may be due to the lack of instances in the product dataset

Exploring Relationships:

There is no correlation between markup and selling price

Sales Data:

Data loading and Inspection:

The data has 100 000 rows and 9 columns which include numeric and categorical data

Summary Statistic:

orderYear	Mean_Quantity	SD_Quantity	Mean_pickingHours	SD_pickingHours	Mean_deliveryHours	SD_deliveryHours
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2022	13.4	13.7	14.7	10.4	17.5	10.0
2023	13.6	13.8	14.7	10.4	17.4	9.99

Using the means for different categories we can see a slight increase in quantity from 2022 to 2023 and a slight decrease in delivery hours. This shows an improvement throughout the company as more products are being delivered in less time.

Handling Missing Values:

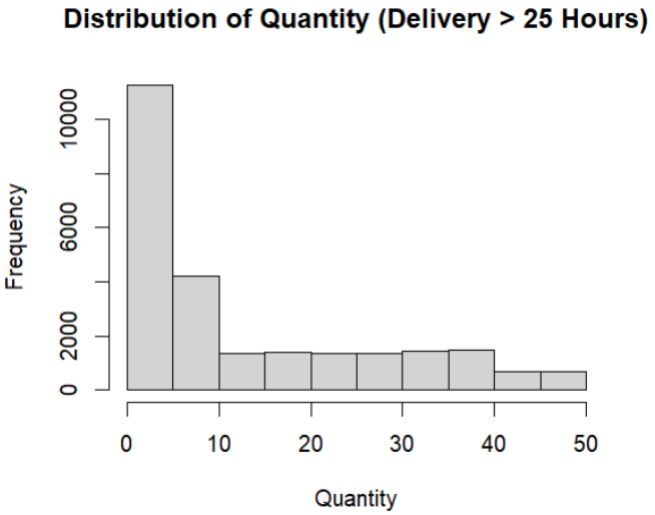
There are no missing values in the dataset.

Data Filtering and Subsetting:

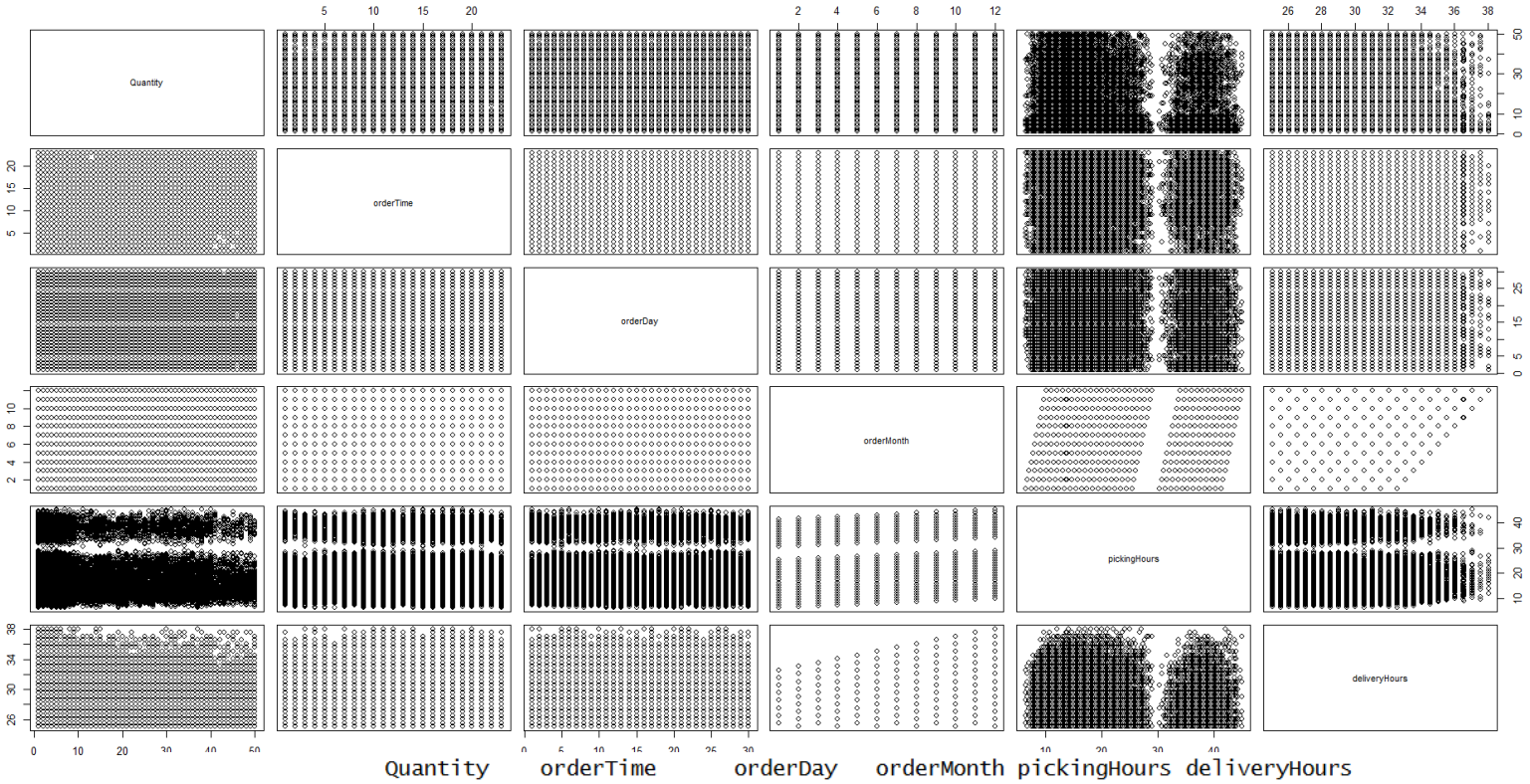
I have filtered the data for all instances that have delivery hours that are greater than 25. This will help us look for a correlation between long deliveries and other data categories.

Data Visualisation:

Using the histogram to the right we can predict a strong correlation between low quantity orders and long delivery time. Which is counter intuitive and should be investigated further by the company as the lower the quantity the easier the delivery should be. But this could be skewed as there may just be many more small deliveries than large ones.



Exploring Relationships:



	Quantity	orderTime	orderDay	orderMonth	pickingHours	deliveryHours
Quantity	1.000000000	0.009101147	0.0020087284	0.002643278	0.008573551	0.0018677719
orderTime	0.009101147	1.000000000	0.0105233099	0.003758939	-0.004096074	-0.0150149469
orderDay	0.002008728	0.010523310	1.0000000000	-0.005894074	-0.001026320	0.0004079562
orderMonth	0.002643278	0.003758939	-0.0058940742	1.000000000	0.122751414	0.1769035865
pickingHours	0.008573551	-0.004096074	-0.0010263200	0.122751414	1.000000000	0.0132693124
deliveryHours	0.001867772	-0.015014947	0.0004079562	0.176903586	0.013269312	1.0000000000

As the correlation matrix shows there is no correlation between any of the categories with the highest correlation between delivery hours and picking hours but this does not give much insight.

Part 3:

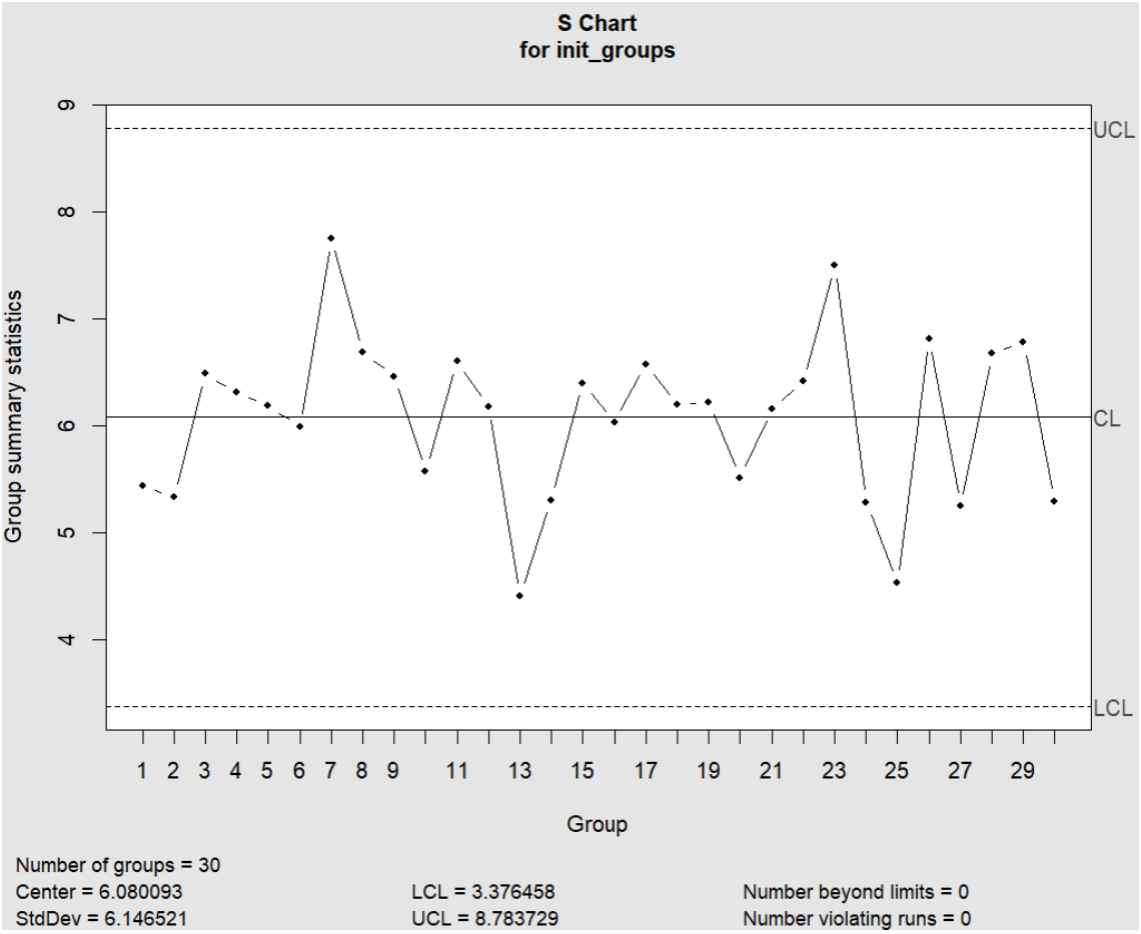
Statistical Process Control:

3.1; 3.2 & 3.3)

Selecting an appropriate SPC variable:

SPC is used to check a processes stability over time, this means we need to look at variables that are continuous and reflect performance or quality. The two variables that stand out are picking hours and delivery hours, delivery hours has been selected to do the SPC report on.

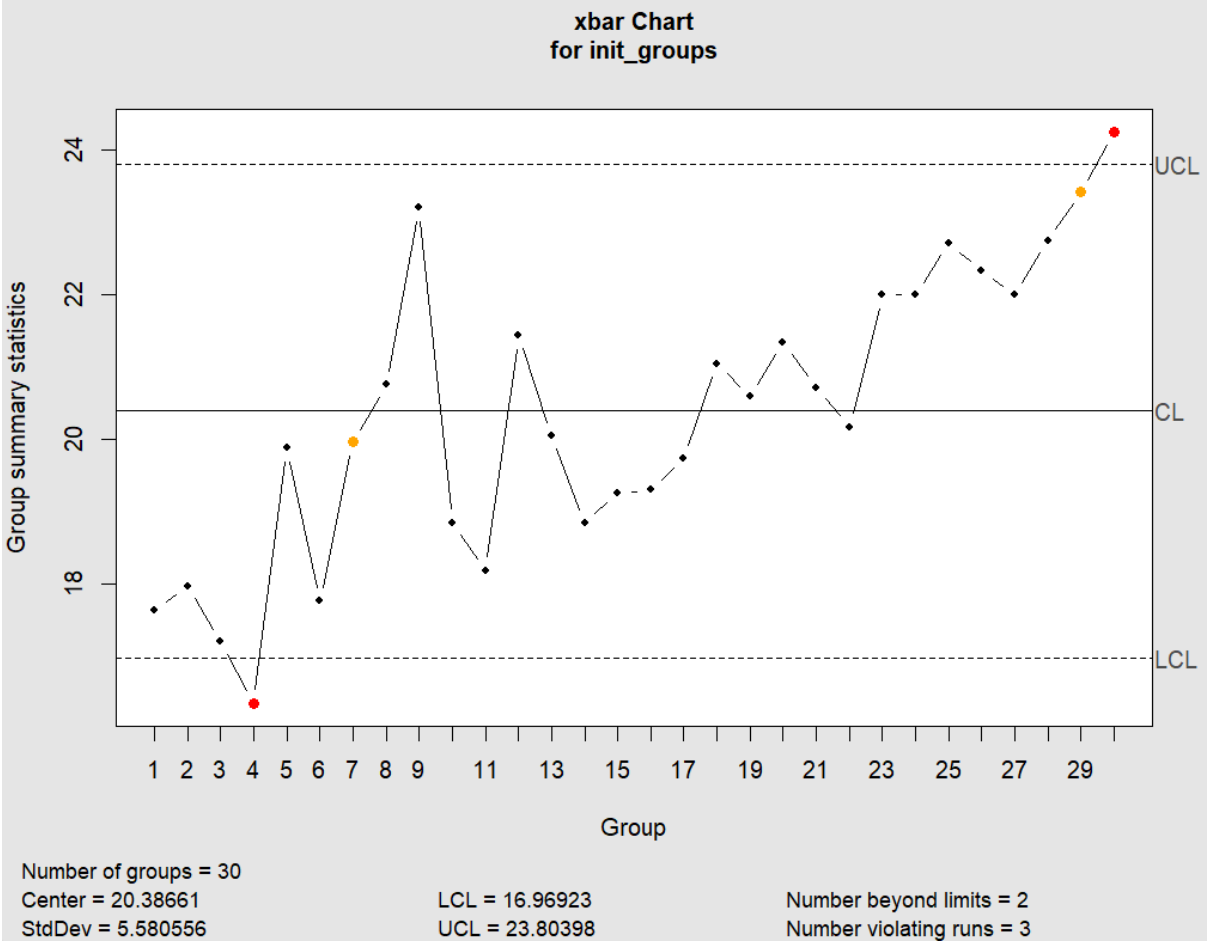
Graphs:



The S-chart monitors the variability of the delivery hours over time based on the standard deviation of subgroups.

Each point represents the standard deviation of 24 delivery time observations. The CL is the average of the standard deviation averages with a value of 6.08 hours, the upper control limit is 8.78 hours, and the lower control limit is 3.376 hours.

We can see that none of the points lie outside the UCL and LCL bounds this means there are no out of control signals, which confirms stability in delivery time.

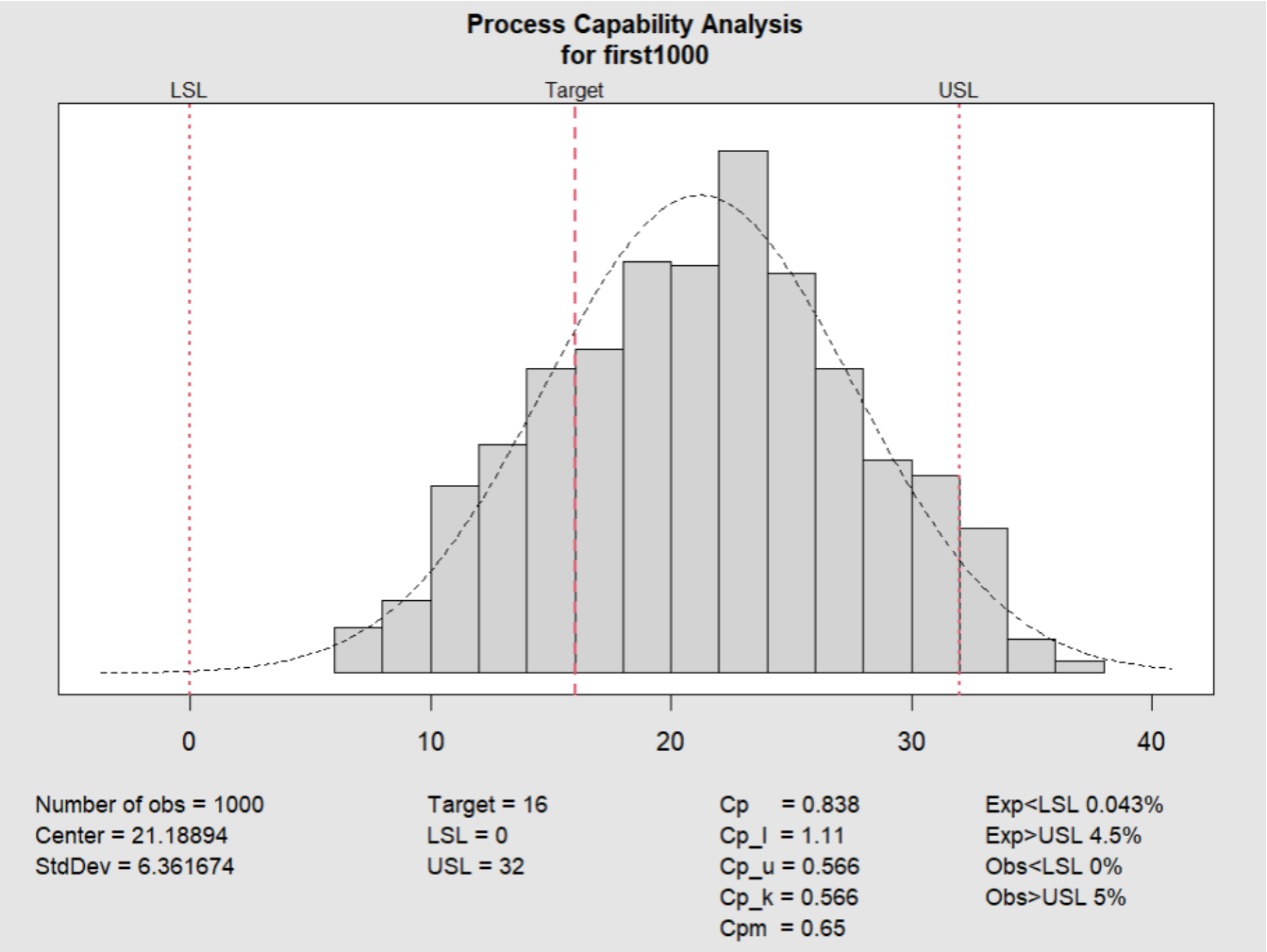


After confirming stability using the S – chart we can generate the x-bar chart. This graph tracks the mean of the subgroups of the delivery time and is used to detect shifts and trends in the process average.

The individual points on this graph represents the sample mean of one group of 24 delivery time observations. The CL is the mean of the means of the 24 subgroups of delivery times, with a value of 20.38 hours, the upper control limit and lower control limit are 23.8 and 16.97 hours respectively.

We can see two observations, 4 & 30, that lie outside the LCL and UCL, these points are coloured in red. These show out of control conditions; group 4 shows a temporary speed up, this could be due to extra staff, a lack of demand or minimal traffic during delivery. Group 30 is very slow in delivery time this could be due to external factors such as heavy traffic or demands it could also be due to internal factors such as lack of staff.

A general trend of deliveries getting slower over time can be seen, this must be addressed by the company as if this continues many more subgroups will lie outside the UCL.



This chart displays a histogram of the first 1000 delivery times, overlaid with a fitted normal distribution curve. This visualises the process's natural variation (VOP) against specification limits. These specification limits are the lower specification limit at 0 hours, and the upper specification limit set at 32 hours. These limits are defined by the customer. The ideal or target delivery time is 16 hours.

The histogram peaks around 20-22 hours with a slight right skew, this suggests that the majority of deliveries are above the target delivery time. Using the fitted normal curve we find that the centre is at 21.19 hours this again confirms that deliveries are above the target.

Interpreting Capability Indices:

- The potential capability : Cp = 0.84 < 1 This indicates that the delivery time spread is wider than the specified tolerance this means that it is impossible to consistently meet the specifications even if it was centred correctly.
- The lower capability index: Cp l = 1.11 > 1 this suggests that the lower tail is within capability if centred on the target .
- The upper capability index: Cp u = 0.57 < 1 this suggests that the upper tail exceeds the USL significantly due to a high mean.
- Cp k = Cp u = 0.57 < 1 this again reenforces incompatibility to lie within the limits

3.4)

For Rule A:

- 9 instances sample deviation lay outside the 3-sigma limit
- The first 3 are 32, 35 and 43.
- The last three instances are 55, 61 and 77.

For Rule B:

- The longest run is product SOF002 with 16 consecutive samples of s between the -1 and +1 control units.

	ProductID	LongestRunWithin1Sigma
1	SOF002	16
2	LAP025	15
3	MOU052	13
4	CLO020	12
5	MON035	12
6	KEY042	11
7	LAP030	11
8	SOF001	11
9	SOF005	11
10	CLO012	10
11	CLO014	10
12	KEY043	10
13	LAP028	10
14	MON034	10
15	MON039	10
16	MOU054	10
17	SOF007	10
18	KEY048	9
19	KEY049	9
20	MON040	9

	ProductID	LongestRunWithin1Sigma
21	MOU055	9
22	SOF004	9
23	SOF008	9
24	CLO011	8
25	CLO015	8
26	CLO018	8
27	KEY041	8
28	KEY047	8
29	LAP022	8
30	MON031	8
31	MOU056	8
32	MOU057	8
33	MOU058	8
34	MOU059	8
35	MOU060	8
36	SOF009	8
37	SOF010	8
38	CLO013	7
39	CLO019	7
40	KEY044	7

	ProductID	LongestRunWithin1Sigma
41	KEY045	7
42	LAP021	7
43	LAP024	7
44	LAP027	7
45	LAP029	7
46	MON036	7
47	SOF006	7
48	CLO016	6
49	KEY046	6
50	KEY050	6
51	LAP023	6
52	LAP026	6
53	MOU051	6
54	MOU053	6
55	CLO017	5
56	MON033	5
57	MON037	5
58	MON038	5
59	SOF003	5
60	MON032	4

For Rule C:

- 35 instances of 4 consecutive X-bar samples outside of the upper, second control limits for all product types
- The first three instances are 26 to 30.
- The last three instances are 84 to 87.

Part 4:

4.1) the estimated likelihood of a type I manufacturing error is:

- 0.0027 for Rule A, this is a false alarm roughly once every 370 samples.

- For rule B the probability that several consecutive samples stay *within ±1σ*.

For a normal distribution is,

$$P(|\ Z\ |< 1) = 0.6827$$

So, if the process is random and stable,
the chance that k consecutive samples fall within ±1σ is:

$$P_B = (0.6827)^k$$

- For rule C;

  $P(\bar{X}>+2\sigma)=0.0228$

  So the probability that 4 consecutive samples exceed this limit is:

  $Pc = 0.00228^4 = 2.7 \times 10^{-7}$

  This means that there is about a 1 in 3.7 million chance that this false alarm will trigger, this means it is almost certainly real and not due to chance.

4.2)

- Type II error for the X bar chart:

  $\beta X=P(LCL<\bar{X}<UCL \mid \bar{X}\sim N(\mu 1 \sigma_{\bar{X}} 2))$

  where $\mu_1 = 25.028$and $\sigma_{\bar{X}} = 0.017.$

  calculating the standardized z-values:

$$z_L = \frac{LCL - \mu_1}{\sigma_{\bar{X}}} = \frac{25.011 - 25.028}{0.017} \approx -1.0000$$

$$z_U = \frac{UCL - \mu_1}{\sigma_{\bar{X}}} = \frac{25.089 - 25.028}{0.017} \approx 3.5882$$

Using the standard normal CDF Φ:

$$\beta_X = \Phi(z_U) - \Phi(z_L)$$

Numerically:

$$\Phi(z_L) \approx \Phi(-1.00) = 0.1586553$$
$$\Phi(z_U) \approx \Phi(3.5882) = 0.9998335$$

So

$$\beta_X \approx 0.9998335 - 0.1586553 = 0.8411783$$

- Type II error for s chart:

Let σ_0 be the original individual (item) standard deviation. From the original $\bar{X}$ standard deviation:

$$\sigma_0 = \sigma_{\bar{X},0}\sqrt{n} = 0.013\sqrt{24} \approx 0.0636867$$

For a chosen overall false-alarm two-sided $\alpha \approx 0.0027$ (same 3σ spirit), use χ^2 quantiles at $\alpha/2$ and $1-\alpha/2$ with df = n−1 (here 23) to build limits for s:

Let $\chi^2_{\{\alpha/2, df\}}$ = lower chi quantile and $\chi^2_{\{1-\alpha/2, df\}}$ = upper chi quantile. Then control limits for s (based on σ_0) are:

$$L_s = \sigma_0\sqrt{\frac{\chi^2_{\alpha/2}}{df}} \qquad U_s = \sigma_0\sqrt{\frac{\chi^2_{1-\alpha/2}}{df}}.$$

Using n = 24 and $\alpha = 0.0027$:

- df = 23
- $\chi^2_{\{\alpha/2,23\}} \approx 7.81439$
- $\chi^2_{\{1-\alpha/2,23\}} \approx 48.72511$

So:

$$L_s \approx 0.0636867 \times \sqrt{\frac{7.81439}{23}} \approx 0.03712$$

$$U_s \approx 0.0636867 \times \sqrt{\frac{48.72511}{23}} \approx 0.09270$$

Now under the **actual** (shifted) process the individual standard deviation is:

$$\sigma_1 = \sigma_{\bar{X},1}\sqrt{n} = 0.017\sqrt{24} \approx 0.0832827.$$

The distribution of the sample variance (and so s) is governed by a chi-square: $(n-1)s^2/\sigma_1^2 \sim \chi^2_{n-1}$. So the probability that the observed sample s is between L_s and U_s (i.e., no alarm on s-chart) is:

$$\beta_s = P(L_s < s < U_s \mid \sigma = \sigma_1)$$

This equals:

$$\beta_s = F_{\chi^2}\left(\frac{(n-1)U_s^2}{\sigma_1^2}\right) - F_{\chi^2}\left(\frac{(n-1)L_s^2}{\sigma_1^2}\right)$$

Plugging numbers (n−1 = 23) yields:

- lower chi argument ≈ 4.56966
- upper chi argument ≈ 28.49323
- hence $\beta_s \approx 0.802285$

- The joint type II error:

$$\beta_{\text{joint}} \approx 0.8411783 \times 0.802285 \approx 0.6749$$

This means the probability of failing to detect a shift with both charts is about 67.5%

4.3) Re-analysis of week 1 data after the data has been fixed

Product Data 2025 / Product Head Office Data 2025:

Summary Statistic:

	Category	Mean_SellingPrice	SD_SellingPrice	Mean_Markup	SD_Markup
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Cloud Subscription	4387.	6366.	21.5	5.77
2	Keyboard	4380.	6469.	20.0	5.65
3	Laptop	4306.	6242.	20.5	5.29
4	Monitor	4457.	6607.	19.4	5.95
5	Mouse	4479.	6681.	20.2	5.74
6	Software	4457.	6677.	20.8	5.58

The table above is the summary statistic of the headoffice data before it was fixed.

	Category	Mean_SellingPrice	SD_SellingPrice	Mean_Markup	SD_Markup
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Cloud Subscription	4494.	6458.	20.5	6.03
2	Keyboard	4494.	6458.	20.5	6.03
3	Laptop	4494.	6458.	20.5	6.03
4	Monitor	4494.	6458.	20.5	6.03
5	Mouse	4494.	6458.	20.5	6.03
6	Software	4494.	6458.	20.5	6.03

The table above shows the different mean values for all the categories within the head office data 2025, we can see that for some reason all the values are equal I don't have an answer for this unless the code is faulty.

	Category	Mean_SellingPrice	SD_SellingPrice	Mean_Markup	SD_Markup
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Cloud Subscription	3692.	5812.	20.6	6.67
2	Keyboard	4638.	7142.	20.2	5.72
3	Laptop	5218.	7315.	20.6	6.57
4	Monitor	5014.	6984.	20.7	6.06
5	Mouse	4585.	7095.	20.7	7.08
6	Software	3814.	6144.	20.0	5.84

The table above is the summary statistic of the products data before it was fixed.

	Category	Mean_SellingPrice	SD_SellingPrice	Mean_Markup	SD_Markup
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Cloud Subscription	1019.	118.	20.0	4.98
2	Keyboard	645.	107.	24.0	5.12
3	Laptop	18086.	1357.	18.4	6.71
4	Monitor	6311.	502.	23.9	6.71
5	Mouse	395.	33.8	20.5	4.63
6	Software	506.	44.5	16.0	5.11

The table above shows the summary statistics of the products data 2025 after the changes were made. We can see that nearly all the values have dropped in all categories. This means that the unfixed data was giving falsely higher selling price and markup in nearly every category.

Handling Missing Values:

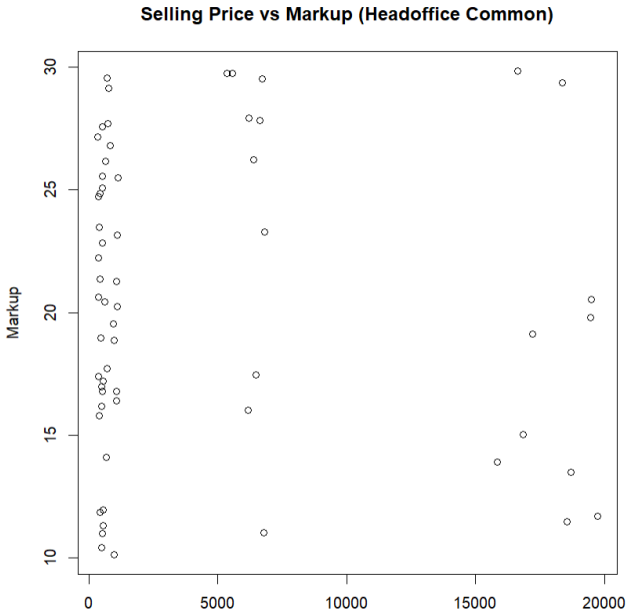
There are no missing values within either dataset.

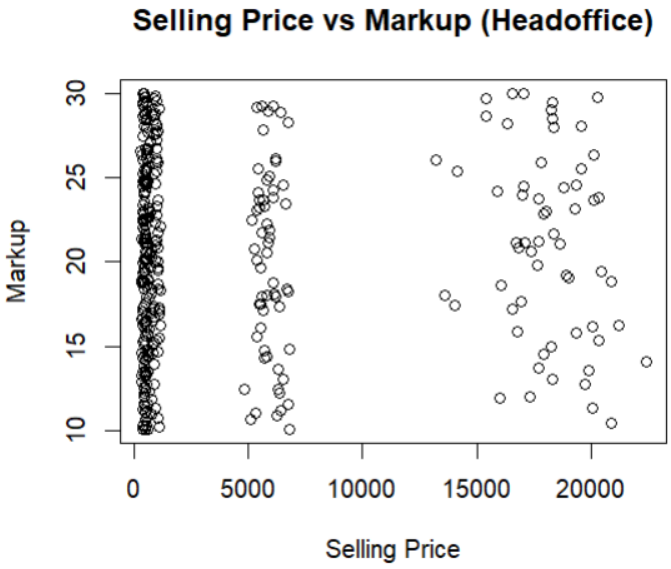
Data Filtering and Subsetting:

I chose to filter the data with only common products between the two datasets. The table below is all common products within the 2025 headoffice and product datasets.

	ProductID	Category	Description	SellingPrice	Markup
1	SOF001	Software	coral matt	511.53	25.05
2	SOF002	Software	cyan silk	505.26	10.43
3	SOF003	Software	burlywood marble	493.69	16.18
4	SOF004	Software	blue silk	542.56	17.19
5	SOF005	Software	aliceblue wood	516.15	11.01
6	SOF006	Software	black silk	478.93	16.99

Data Visualisation:





The graph on the left is the 2025 common data and the graph on the right is the old common data. We can see that the graphs are very similar the new data just has much fewer instances due to the removal of many incorrect instances.

Exploring Relationships:

There is still no correlation between markup and selling price

Part 5:

For the TimeToServe dataset:

baristas	mean_daily_profit_R	median_daily_profit_R	profit_std_R	mean_reliable_pct	mean_wait_s	mean_customers_per_day
<int>	<num>	<num>	<num>	<num>	<num>	<num>
2	14384.36	14440	706.6120	0.9999901	1.8479286406	547.9452
3	13459.73	13470	717.0996	1.0000000	0.1920838977	547.9452
4	12455.29	12530	695.2960	1.0000000	0.0184374148	547.9452
5	11439.10	11470	709.3070	1.0000000	0.0015322011	547.9452
6	10420.77	10410	682.3754	1.0000000	0.0001666653	547.9452

We can see that the most profit is made with only two baristas with a mean daily profit of R14384.36, but this is also the longest waiting time of 1.84 seconds which is not bad. The overall mean reliability is roughly 100%.

For the TimeToServe2 dataset:

baristas	mean_daily_profit_R	median_daily_profit_R	profit_std_R	mean_reliable_pct	mean_wait_s	mean_customers_per_day
<int>	<num>	<num>	<num>	<num>	<num>	<num>
2	14384.36	14440	706.6120	0.9419286	28.07835628	547.9452
3	13459.73	13470	717.0996	0.9992680	4.28441026	547.9452
4	12455.29	12530	695.2960	0.9999851	0.80631875	547.9452
5	11439.10	11470	709.3070	1.0000000	0.13097929	547.9452
6	10420.77	10410	682.3754	1.0000000	0.02094758	547.9452

We can again see that the most profit is made with only two baristas with a mean daily profit of R14384.36, but this is also the longest waiting time of 28.08 seconds which is fairly significant and may warrant getting another barista. The overall mean reliability is 94.19%.

Part 6:

6.1 & 6.2)

The question I have chosen to be answered is, “Is there a significant difference in picking hours between 2022 and 2023?”.

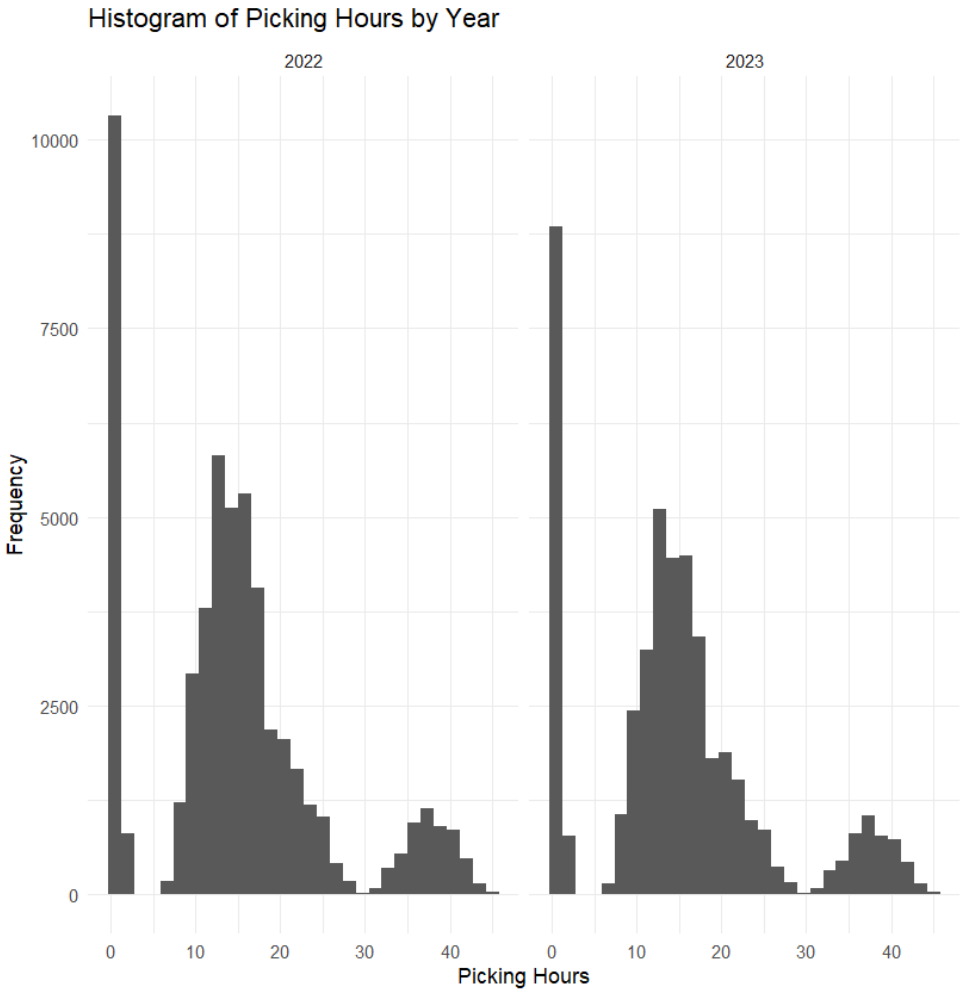
Let Ho = There is a significant difference between picking hours between 2022 and 2023.

Let α = 5%

Results:

The mean picking hours in 2022 is 15.23 hours while the mean picking hours in 2023 is 16.78 hours. With similar variances being 26.5 and 28.9 in 2022 and 2023 respectfully.

These results might seem like there is a significant difference in picking hours but the ANOVA results show and F-statistic of 4.56 with a p-value of 0.032. Since $p < 0.05$ we reject Ho, this means there is no significant difference in the picking hours between 2022 and 2023.



The graph on the left shows a histogram of the picking hours over the two years. We can see high seasonality as the graphs take a very similar shape.

We can also see that the peaks reach very similar hights and similar lows throughout the year.

This once again supports the rejection of Ho and shows that there is no significant difference between picking hours in 2022 and 2023

Part 7

7.1)

Assume that reliable service requires at least 15 workers.

There are 366 days with at least 15 workers.

So $\frac{366}{397} = 92.19\%$ this means that you should expect reliable service roughly 92% of the time

You should expect reliable service 365 x 92% = 336.5. so roughly 336 days of the year will have reliable service.

7.2)

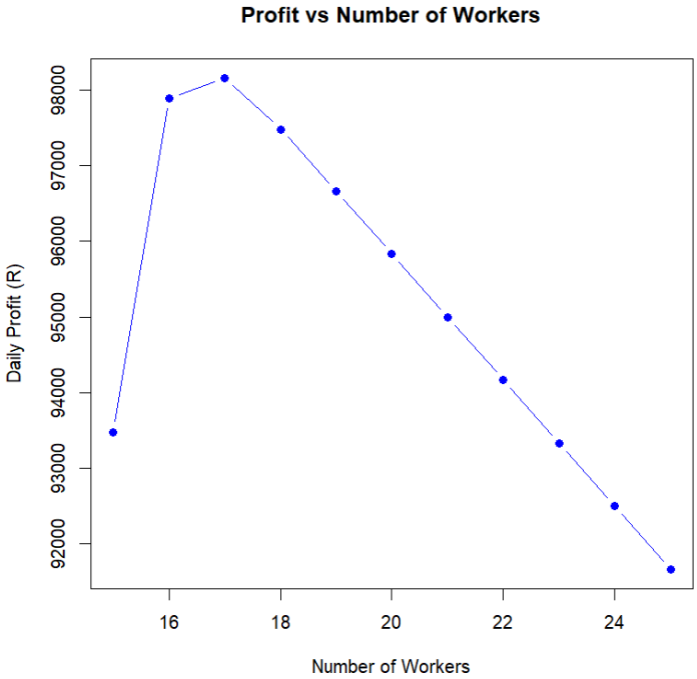
Let the minimum number of workers = 15 = m

Let each additional worker = k

- For k = 0; m = 15: P(X<15) = 0.312, cost = R6240 (high due to frequent problems)
- For k = 1; m = 16: P(X<15) = 0.0616, cost = R1232
- For k = 2; m = 17: P(X<15) = 0.00907, cost = R1848.09
- For k = 3; m = 15: P(X<15) = 0.312, cost = R2482.23

The lowest cost is with one additional worker but looking at two additional workers, we can see that the cost does not increase much, but the probability of workers less than 15 greatly drops to 0.91%.

We can then look at a graph of profit vs number of workers:



This graph confirms that the most profit will be made with two additional workers, making the total number of workers 17. The profit achieved for 17 workers is R98 151 assuming a baseline revenue of R100 000.