Quality Assurance Project

Corne Mouton (27516792)

Contents

Introduction

The following report covers the ECSA graduate attribute project completed for Quality Assurance 344. The report covers the questions answered from the ECSA report brief. The sections are ordered in the same manner as they were completed.

Initial Data Analysis

Three data sets were provided to perform data analysis on. These data sets are customer_data, products_data and products_Headoffice respectively. The analysis was an attempt to discover unique insights to the data and spot useful or informative trends or patterns within the data.

Summary Statistics

The following are basic statistics calculated for each continuous feature of the data sets. These statistics are referenced throughout the report.

Summary of customer_data

Feature Name	Age	Income
Min	16.00	5000
1st Quartile	33.00	55000
Median	51.00	85000
Mean	51.55	80797
3rd Quartile	68.00	105000
Max	105.00	140000

Summary of product_data

Feature Name	SellingPrice	Markup
Min	350.4	10.13
1st Quartile	512.2	16.14
Median	794.2	20.34
Mean	4493.6	20.46
3rd Quartile	6416.7	25.71
Max	19725.2	29.84

Summary of products_Headoffice

Feature Name	SellingPrice	Markup
Min	290.5	10.06
1st Quartile	495.9	15.84
Median	797.2	20.58
Mean	4411.0	20.39
3rd Quartile	5843.3	24.84
Max	22420.1	30.00

Summary of Sales

Feature Name	Quantity	pickingHours	deliveryHours
Min	1.0	0.4259	0.2772
1st Quartile	3.0	9.3908	11.5460
Median	6.0	14.0550	19.5460
Mean	13.5	14.6955	17.4765
3rd Quartile	23.0	18.7217	25.0440
Max	50.0	45.0575	38.0460

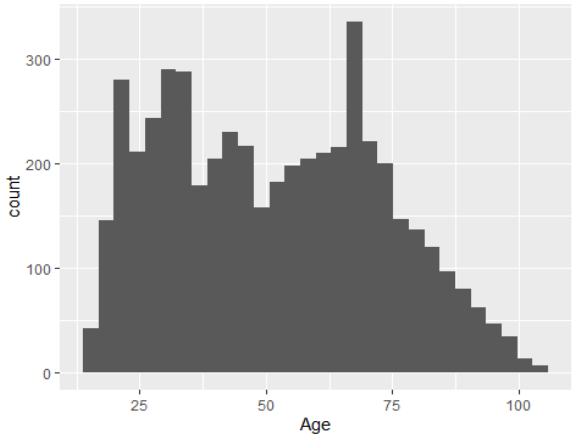
Data Plots

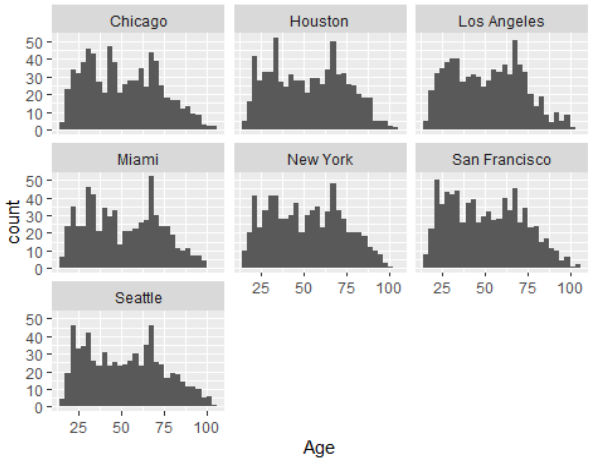
The following plots were used to visualize the data to attempt to find useful relationships within the data.

Age Demographic

Below is a plot indicating the distribution of client ages. This distribution appears to have a bimodal distribution with an expected peak between the ages of 25 and 50. However, an unexpected peak can be seen between the ages of 50 and 75. This peak is larger than any other single age interval. This information can be useful when releasing a new product or when designing new products. This age group can be a focus point, and their needs can be prioritized by the company accordingly as the contribution from this group will have a significant impact on the company's performance.

This distribution is near identical when plotted for each city. This shows that the distribution is independent of the city.

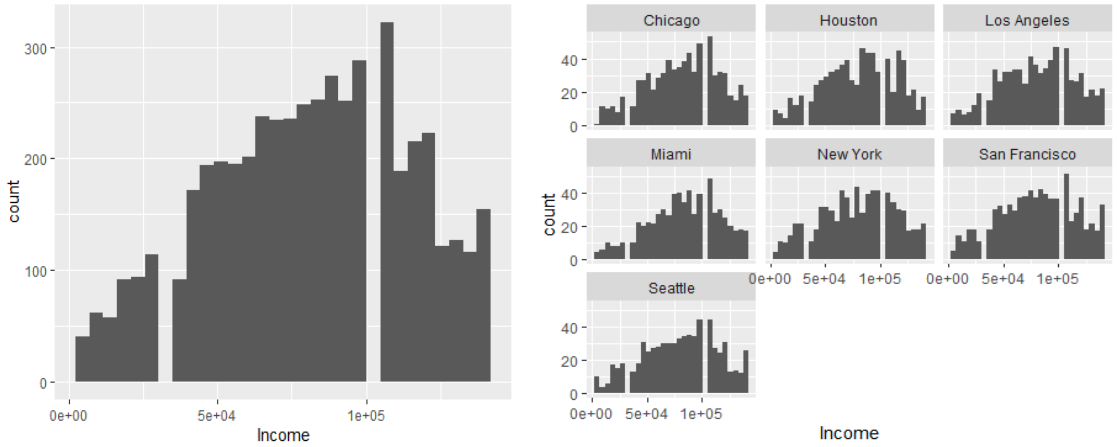




Income Demographic

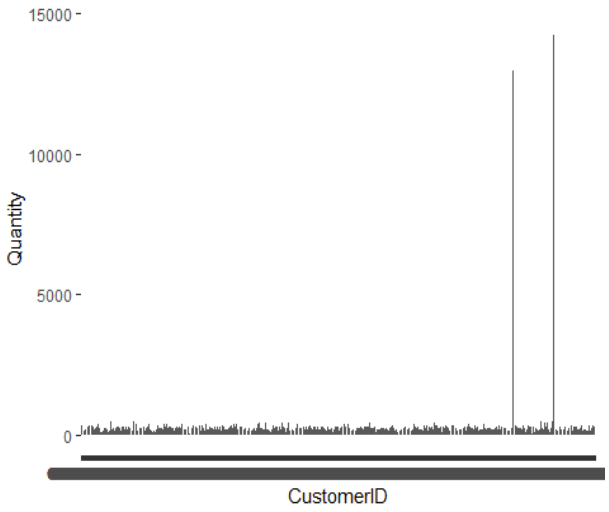
The income demographic shows a possible unimodal distribution skewed to the left. The mode or peak is positioned to the right, showing a larger number of customers coming from higher income backgrounds. This is insightful as the company focuses to a greater extent on this income demographic and its tastes, garnering a larger number of sales.

Plotting the income distribution for each city yields a near identical distribution, indicating that the distribution is independent of the city.



Customer Purchase Quantity

The plot below indicates the product quantities purchased by each customer. The plot shows a small number of customers purchasing an enormous amount more than any of the other customers. This plot shows that these customers have a sunstancial effect on the performance of the company. These customers were identified to be CUST596, CUST1193, CUST1427, CUST1501, CUST1791, CUST2277, CUST2527, CUST3721, CUST3944 and CUST4729. Together, these customers makeup **10%** of the total sales quantity. This information can be used for advertising to these customers to improve the sales to them as an improvement to their sales will have a substantial impact to the financial performance of the business due to them making up such a significant proportion of the sales volume.

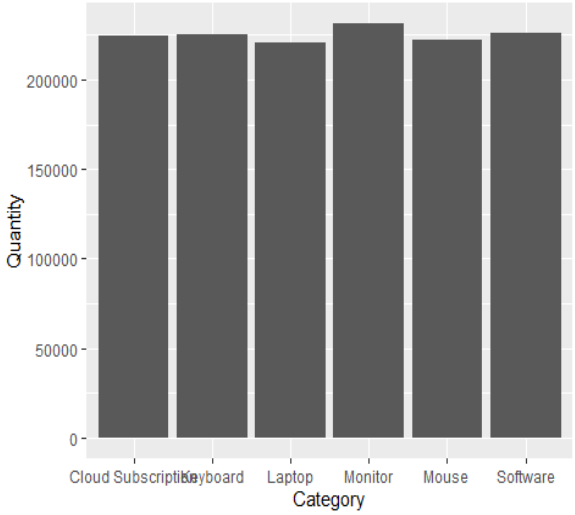


Information on each of these customers are as follows:

Customer ID	Gender	Age	Income	City
CUST596	Female	25	90000	Los Angeles
CUST1193	Female	20	65000	San Francisco
CUST1427	Female	18	55000	Seattle
CUST1501	Male	53	70000	New York
CUST1791	Male	39	100000	Los Angeles
CUST2277	Male	60	140000	Seattle
CUST2527	Female	54	50000	Los Angeles
CUST3721	Female	66	60000	Miami

CUST3944	Male	77	100000	San Francisco
CUST4729	Female	71	60000	Houston

Their income is spread over all quartiles of the income distribution with most data in the 2nd quartile. Most of these customers are Female, however the split between male and female is almost equal. Their ages are spread over all quartile with the majority falling within the 3rd quartile range. The city with the highest concentration of these customers is Los Angeles with 3 customers.

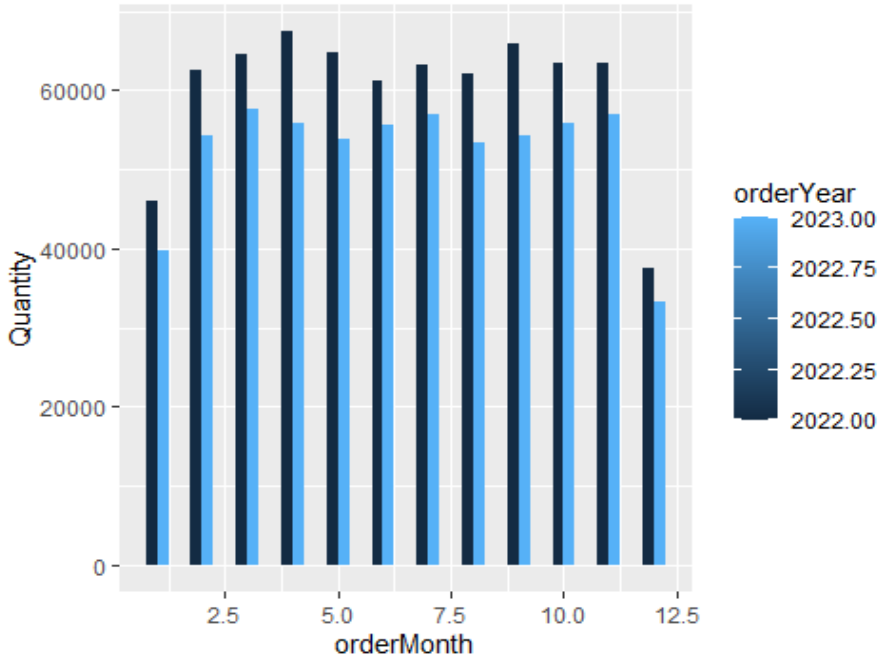


Product category distribution

The plot below shows the distribution of the quantities of each product category purchased. This plot is near uniform. Thus, the quantities purchased in each product category is unaffected by the product and all products appear to be in equal demand.

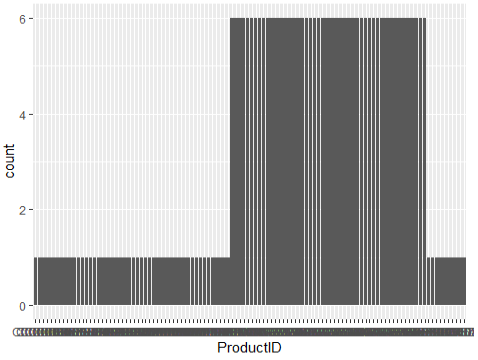
Time Plot of Sales Quantities

The plot below shows the change in the number of products purchased in each month over the two years that the data was provided for. The plot shows that there was a universal decline in the number of products purchased from 2022 to 2023. However, the purchasing pattern remained unchanged, with purchases peaking in April and September and falling to their lowest in January and December. This pattern can be useful to determine during which months extra stocks should be kept by the company or during which business will be less/slower. This can thus aid in planning.



Quality/Accuracy of data

The following plot illustrates data errors. As evident on the plot, the ProductID feature has multiple identical values which should be impossible as each ID is a unique identification. This means a sizable portion of products have been incorrectly allocated their ID numbers. This is a clear problem to be rectified in the data sets before further data analysis or use of data.



Statistical Process Control

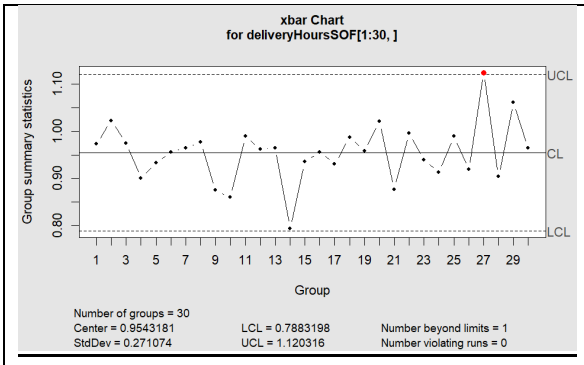
Statistical process control is a key concept to ensure quality. For a previously specified company, statistical process control charts were set up for the 6 different product classes to determine the performance of the company in terms of its delivery time reliability. These charts will aid the company by notifying them in the event of their system going unstable.

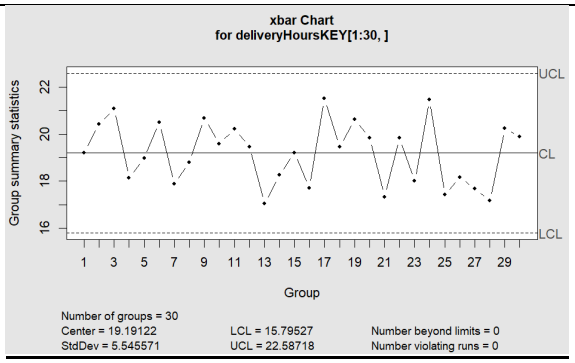
These charts are shown below for each product class. The capability of the company to supply each product class within the assigned delivery time is also evaluated. Each section concludes with a short summary of the results.

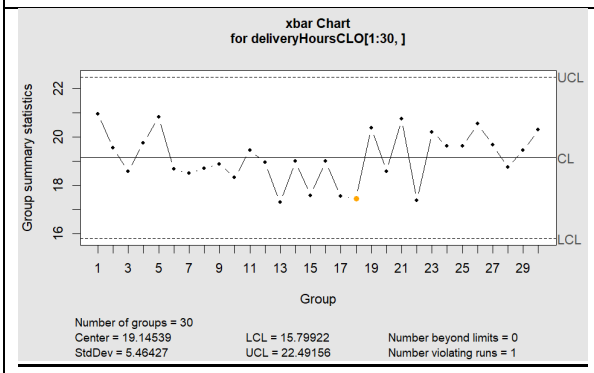
Control Charts

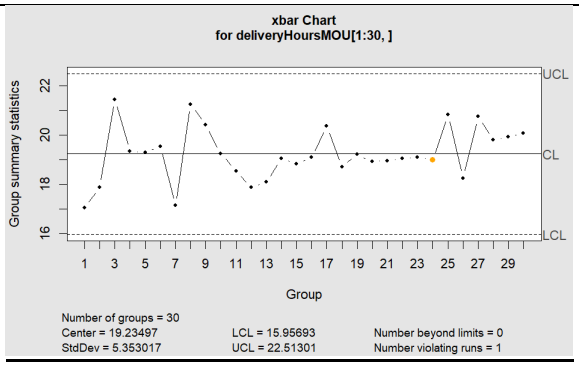
X-bar Chart Tables

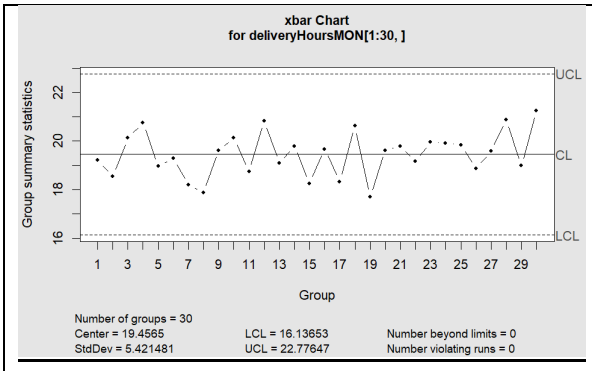
Initial Charts of 30 samples

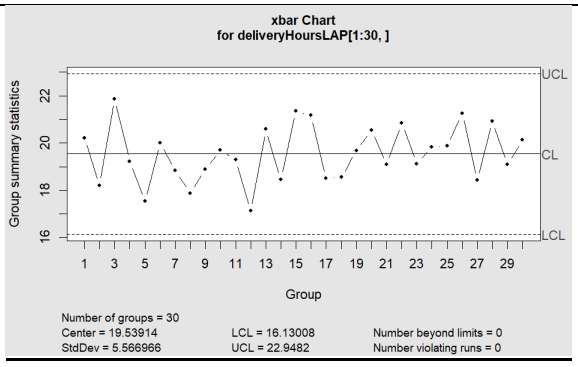






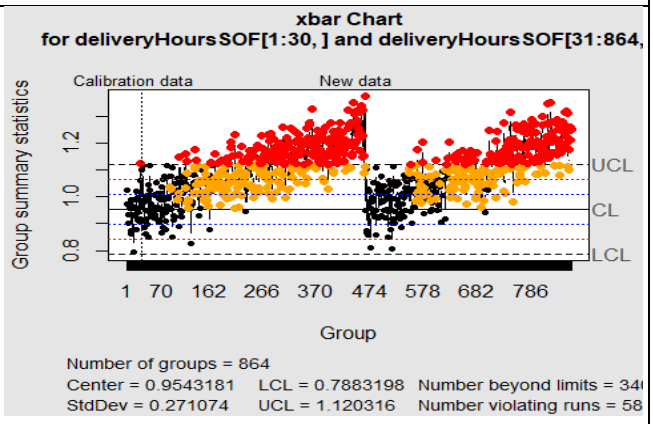
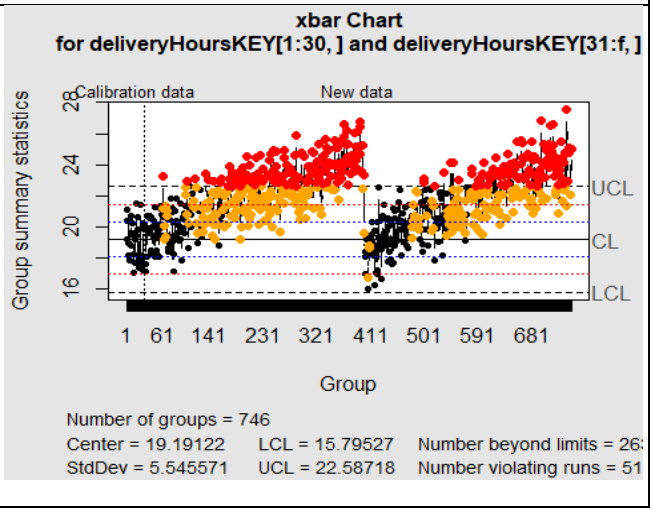


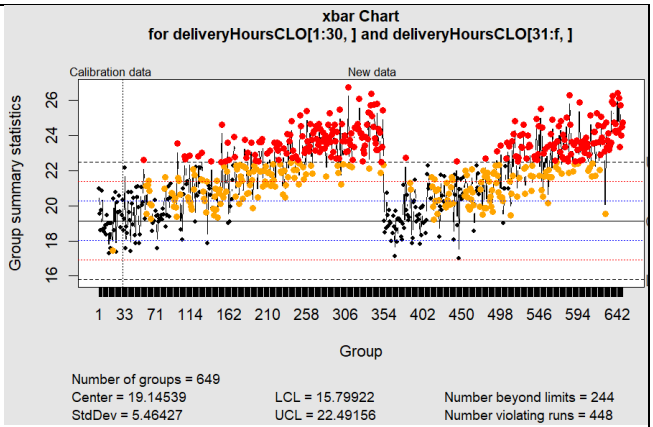
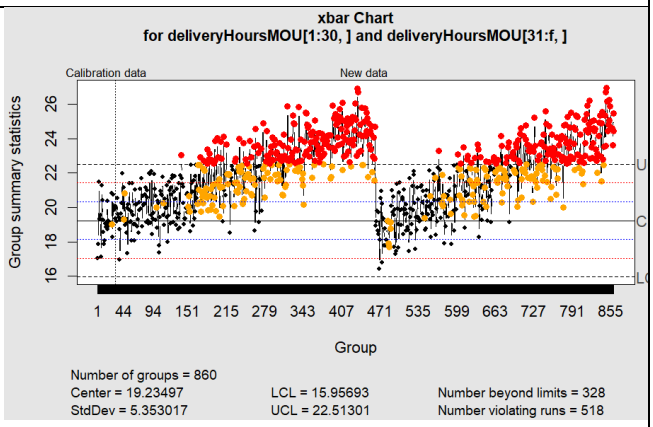
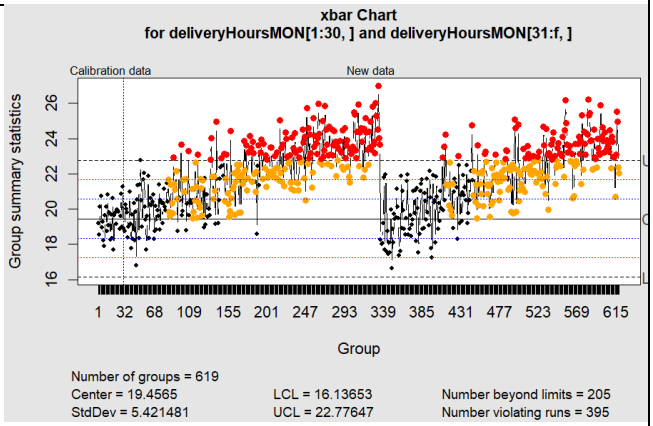
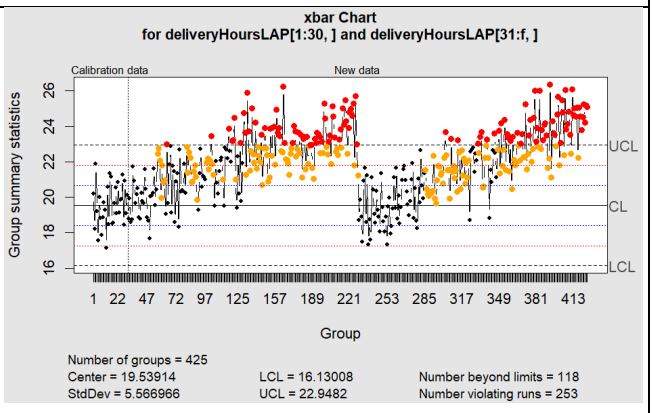




*The above charts show the control charts created from the initial 30 samples for each produ ct class. One notable feature is in the SOF product class, with a single instance of the initial 30 being outside the control bounds.

Control Charts with All future Instances

Product Class	Total out o utsid e bou nds	First 3 Sa mples out side boun ds	Last 3 Sa mples out side boun ds	Chart
SOF	340	104 115 119	862 863 864	
KEY	263	62 102 116	742 743 746	

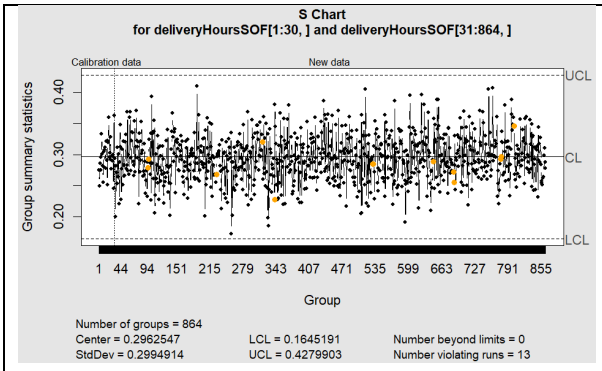
CLO	244	56 98 107	647 648 649	
MOU	328	140 172 178	858 859 860	
MON	205	98 99 108	615 616 617	
LAP	118	64 102 117	423 424 425	

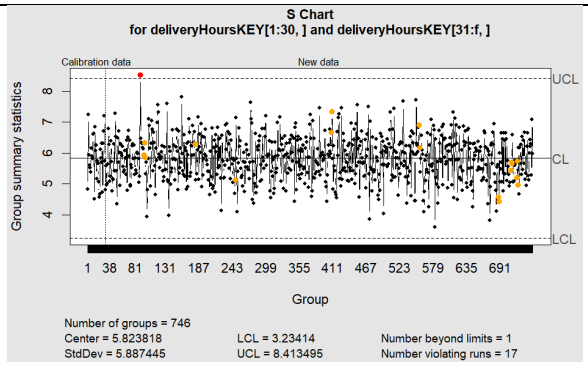
S-chart Tables

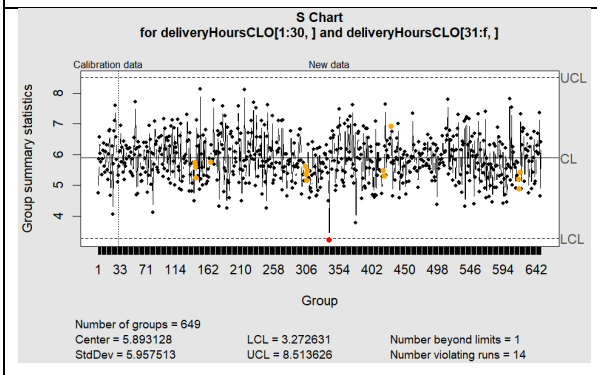
S value table summary

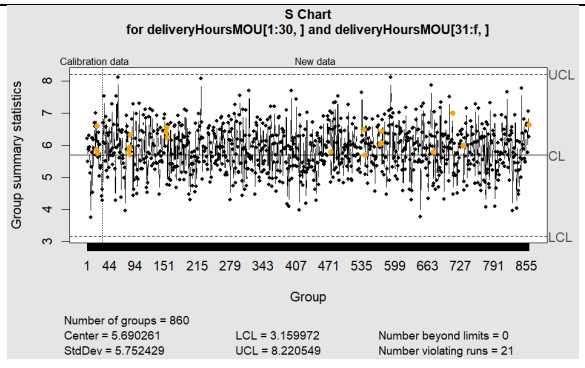
Product Class	S values outside bounds	Largest consecutive values between ±1 σ bounds
SOF	0	16
KEY	1	20
CLO	1	27
MOU	0	14
MON	0	26
LAP	0	17

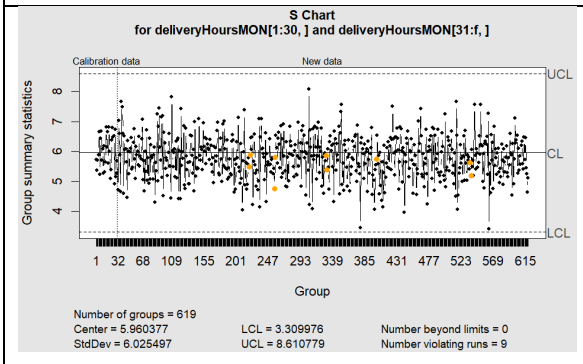
S Chart Table

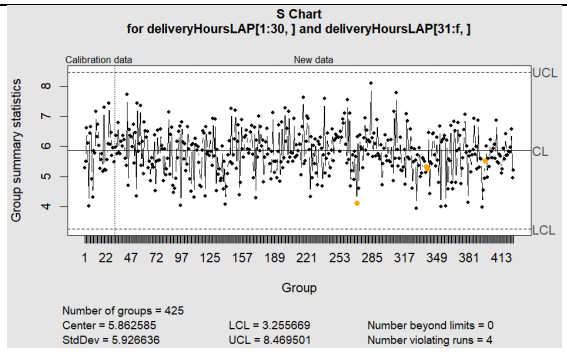




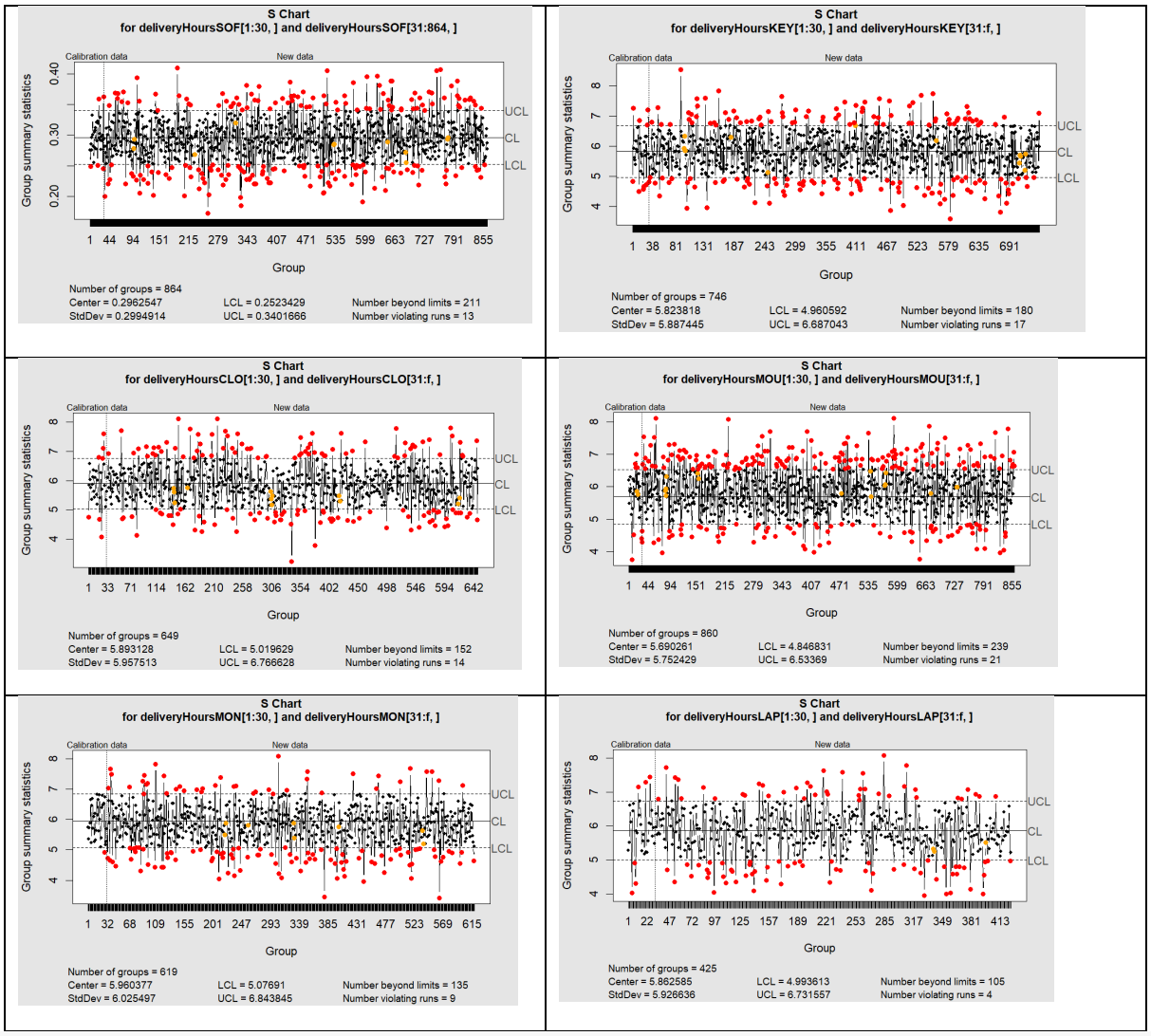








±1σ Chart Table



*(Note: All values in the above table were either directly taken from objects created by the qc c functions or collected from code written to extract the information)

Summary

From the indicated diagrams, every product class showed an upward trend in the mean of each sample delivery time. The distribution visible in the diagrams for each class is nearly identical. This indicates a companywide variable or factor affecting the delivery times of products being delivered. An interesting feature visible on the distributions is the sudden drop in the average delivery times of samples in the middle of the chart. This is assumed to correspond to the start of a new year. Thus, delivery times appear to have a steady upward trend throughout the year. This trend may be attributed to factors such as employee fatigue leading to increasingly long lead times.

The S graphs with the standard control limits and additionally the 1σ control limits indicate a very steady variation in the values of samples collected from delivery times, which is indicative of very good control in the process. The KEY and CLO product classes each have only a singular sample value outside the control bounds. These samples are assumed to be outliers given their singular occurrence. The CLO product class appears most stable as it has the highest number of consecutive values between the ±1σ sigma bounds.

The sample numbers, indicated in the tables presenting the x-bar charts for delivery times, that fell outside the control bounds, are indicators to the process managers to investigate their processes and make the necessary changes to the system. Within the context of a delivery system, this could entail a physical inspection of the system to determine whether operations are performed in the most efficient manner according to some set of operating procedures or best practices. A deviation from such activities can easily account for the increase in delivery lead times.

Process Capabilities

The capabilities of the delivery system were evaluated by calculating its Cp and Cpk values for each product class. These calculations are summarized in the following table below.

Product Class	**C_P**	**C_{PK}**
SOF	18.1546726	1.0866423
KEY	0.9169206	0.7298115
CLO	0.8971579	0.7169413
MOU	0.9151921	0.7254328
MON	0.8897044	0.6998637
LAP	0.8987584	0.6965939

Results Summary

From the C_p values we gauge what the potential of a system to meet the requirements asked for is. In these values we see the only product delivery system with the necessary capabilities to meet the requirements asked for is that of SOF-class products. All other delivery processes, from the start, do not have the capability to meet the requirements asked for. This is carried forward into the C_{pk} values.

C_{pk} values indicate how well a system can meet specifications. Here, we can predictably see that all but one of the delivery processes is completely incapable of meeting demand. The delivery process of the SOF class has marginal capability to meet requirements, however, until its C_{pk} value rises more, this process cannot be considered stable, and any planned improvements should be conducted here as well.

Error Evaluation

During statistical process control, there is always a likelihood of making an error. Two prominent errors are Type I and Type II errors. Type I errors, in this context, involve the incorrect rejection of a stable process as unstable where Type 2 errors involve the failure to reject an unstable process. An example of each was calculated to illustrate this.

Type I (α)

Type I error is the probability that an in-control process is rejected as being out of control. In the context of using control limits for delivery times in the previously mentioned company, a

process is considered out of control if it has four consecutive samples outside the specified ±3σ control limits. Thus, for a normal distribution, the probability to of making a Type I error is the probability of a sample falling above or below 3σ from the process mean and raising it to the power of 4 to find the probability for 4 consecutive samples. Taking the values from a standard normal table, the result is:

$\alpha = [P(Z \leq -3\sigma) + P(Z \geq +3\sigma)]^4 = [2*P(Z \leq -3\sigma)]^4 = [0.0027]^4 = 5.31441 \times 10^{-11}$

Type II (β)

Type II errors are the probability that we fail to reject an out-of-control process. We calculate β for a process where its assumed mean was 25.05 with a UCL and LCL of 25.089 and 25.011 respectively. We assume the same conditions as previously to reject the process as out-of-control, i.e., 4 consecutive samples outside control limits lead to rejection of the process.

The process average has shifted from 25.05 to 25.028 with a standard deviation of 0.017. The probability of making a Type II error is the probability that 1, 2 or 3 consecutive samples fall outside the control limits with the next sample falling within the control limits. This will lead to a failure to reject the process as out of control as the process will never reach the requisite 4 samples outside the control limits.

This probability is found by using the negative binomial equation. This equation gives the probability of the number of failures until the nth success using the following equation:

$$P(X - x) = \binom{n + x - 1}{n - 1} p^n (1 - p)^x.$$

*p denotes the probability of a success. From the standard normal distribution this probability is calculated as:

$p = P(25.011 \leq X \leq 25.089) = P(X \leq 25.089) - P(X \leq 25.011)$

$= P(Z \leq \frac{25.089 - 25.028}{0.017}) - P(Z \leq \frac{25.011 - 25.028}{0.017}) = P(Z \leq 3.59) - P(Z \leq -1) = 0.84117$

For the above, the equation is formulated as:

$$\beta = \sum_{x=1}^{3} \binom{1 + x - 1}{1 - 1} 0.84117^1 (1 - 0.84117)^x$$

The above then yields an answer of 0.1581936. This is the probability that the control charts fail to identify an out-of-control process.

Fixed Data Quality Issues

During earlier data analysis, some data discrepancies were identified between the products_data.csv and products_Headoffice.csv files. These included incorrect product IDs, mismatching Selling Price and Markup values between the products_data.csv and products_Headoffice.csv files and a mismatch within the products_data.csv file between product category and product ID's.

Using Microsoft Excell functions such as LEFT, these discrepancies were rectified and replaced as instructed by the project brief. As a note, the Selling Price and Markup values

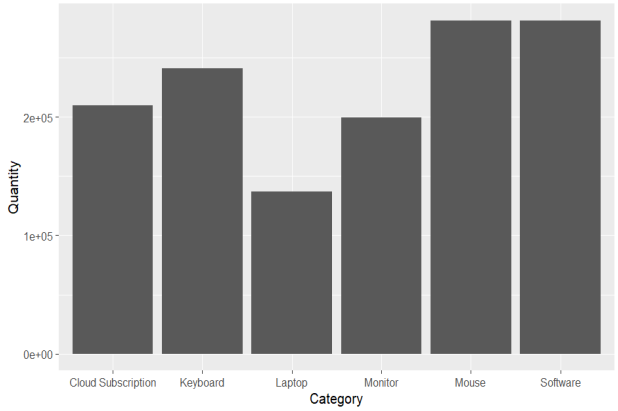
were rectified using R, with the correct values of each products category taken from the product_data.csv file and repeated within the products_Headoffice.csv file to give correct values.

Data Analysis Re-done

As the data quality issues have been rectified, the data analysis performed earlier was repeated to ascertain whether changes have occurred within the data patterns.

Product Sales Quantities Changed

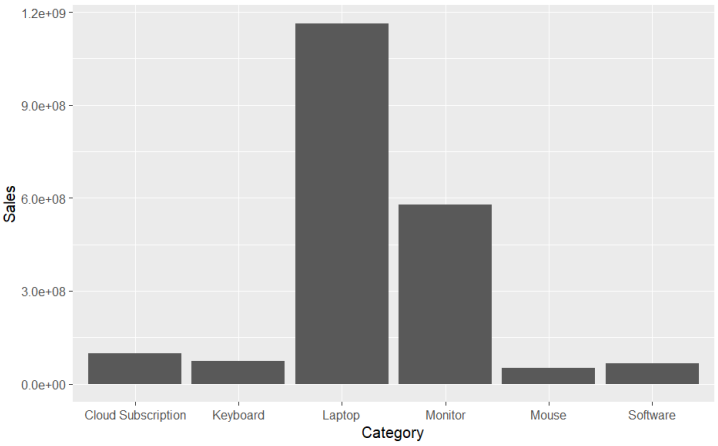
The most notable change to the data analysis results was a change in the distribution of the number of products bought for each category. As can be seen in the bar chart, the distribution that was near uniform previously, now shows some significant variation in the quantity distribution. This distribution shows Mouse and Software products which make-up the largest proportion with Laptops making up the smallest proportion of sold products.



This information is useful as it can help the company identify which product's processes to improve first if only limited improvements can be made or resources are constrained. This will ensure the company allocates its limited resources to where they are most effective. In this case, allocation would be done for either Mouse or Software products as an improvement to these two product categories will net the most substantial improvement to the company's operational performance.

Sales Aggregation by Product

As part of further data analysis, the total sales value of each product for the year 2023 was aggregated and plotted. This yielded the following plot.



This plot shows, that out of all products, laptops have the largest sales value of any product. This is noteworthy as the previous plot showed how laptop sales quantities are the lowest amongst all products. This indicates that even though laptop sales are relatively low-volume products, they are the products bringing the highest value per unit sold. This indicates that any factor affecting laptop sales will be critical for the financial success of the company.

Barista Optimization

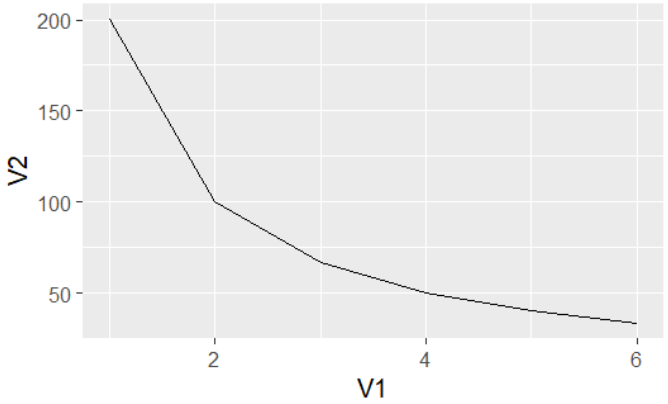
The project brief outlines an optimization problem in where data analytics is to be used to determine the number of baristas needed by two coffee shops to maximize their profit, while ensuring that their customer service is satisfactory.

Two documents outlining the time taken to serve each customer with a specified number of baristas present was given. Additionally, each barista costs R1000 per day, while each customer served nets a before-labour profit of R30 per customer. Here follows an outline of the process to find the optimal number of baristas.

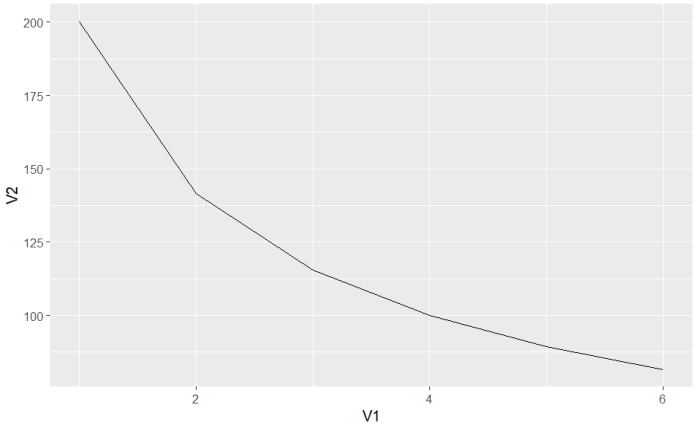
The mean serving time for each number of baristas was calculated and recorded. These values were then plotted on line-graphs (visible to the right).

*V1 is the number of Baristas

*V2 is the serving time in seconds



Shop 1 Serving Time vs Number of Baristas



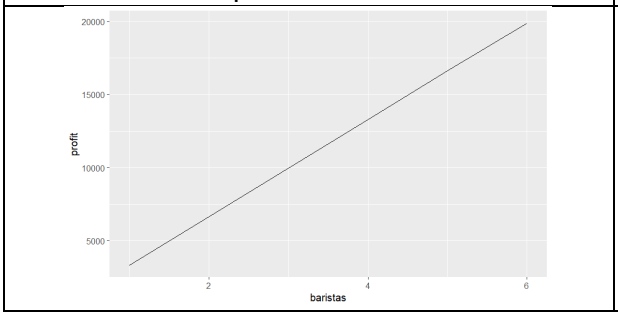
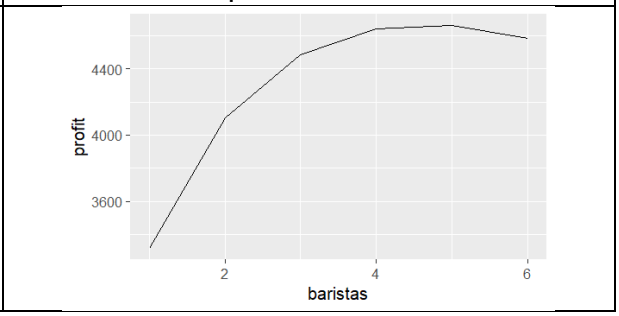
Shop 2 Serving Time vs Number of Baristas

These graphs clearly show the diminishing returns from increasing the number of baristas further. The curves appear to plateau around 6 baristas. This gives strong evidence that the optimal value for the number of baristas w.r.t cost can be found around the values 5 or 6. A further analysis of the potential profit earned per day was done to ascertain exactly how many baristas are optimal. The following formula was used to calculate the profit before labour for each number of baristas:

$$Profit\ Before\ lab. = \frac{R30\ Profit\ per\ Customer \times 3600\ seconds\ per\ hour \times 8\ hours\ per\ day}{Serving\ Time\ per\ Customer\ (seconds)}$$

The profit after labour was then calculated by subtracting the number of baristas multiplied by R1000. The following table shows the estimated profit per day for each number of baristas along with a graph showing the same.

Number of Baristas	1	2	3	4	5	6
Shop 1 Profit per day	R3316.64	R6625.25	R9970.69	R13286.78	R16620.63	R19902.66
Shop 2 Profit per day	3316.354	4105.38	4484.35	4638.68	4660.54	4582.70

Shop 1 Profit Curve	Shop 2 Profit Curve
	

The above table and graphs show that, given the current data, 6 baristas are the optimal number of baristas for shop 1 and 5 baristas for shop 2 respectively. This is known as these numbers yield the highest profit by balancing customer serving time and the cost of additional staff.

ANOVA analysis

From the previous data analysis graphs for the personal electronics company, it could be seen that the total quantities of products purchased for mouse and software products appear equal. This may not be reflective of variance in the quantities purchased for each product for each month. Consequently, the following question remains:

Are the sales patterns of mouse and software products statistically dissimilar for this company?

An Analysis of Variance (ANOVA) hypothesis test was carried out to investigate this question.

The NULL hypothesis was that there is no statistically significant dissimilarity between the sales pattern of these products and the alternative hypothesis being that the sales pattern of these products is statistically dissimilar.

The following data set was created for use in the ANOVA, with the columns being the sales for a particular month and each row representing a product category.

Product	1	2	3	4	5	6	7	8	9	10	11	12
Mouse	17676	24343	25304	26297	24402	24518	25324	23867	24586	24773	24937	15273
Software	17247	25098	26557	25345	24516	23814	25106	24257	25863	24918	24324	14658

This table was used to construct the following ANOVA table to perform the hypothesis test.

ANOVA Table

Variation Source	Sum of Squares	Degrees of Freedom	Mean Squares	F-value
Product Type	6767.0417	1	6767.0417	0.0005541397
Error	268659539	22	12211797	
Total	**268666306**	**23**		

For the above test, to find the critical F value, an alpha of 0.01 was chosen. Using the above specified degrees of freedom, the critical F value was determined as 7.95

As the calculated F =0.0005541397 is smaller than 7.95, we fail to reject the NULL hypothesis and cannot conclude that the sales pattern between the mouse and software product classes are statistically dissimilar.

Reliability of Service

At a car rental agency, an inquiry was made into how reliable their customer service is given their employee availability. A graph detailing employee availability over a 397-day period was provided. At least 15 employees are needed for smooth operations, where sub-15 employees lead to a daily loss of R20 000. Employees are paid a monthly salary of R25 000. Evaluating the given data, it can be projected that out of 365 days a year, it can be expected that reliable service will be available on 336 days.

To find the optimal employee quantity, the probability of an employee showing up at work had to be. This was done by reversing the binomial equation to find the probability of an employee showing up on a particular day. It was assumed that the total number of current employees is 16. The probability of all 16 employees showing up on a particular day is 270/397 = 0.680101. This is equivalent to the equation below:

*p is the probability of an employee coming to work on any given day

$$\binom{16}{16}(p)^{16}(1-p)^{16-16} = 0.680101$$

Solving the above for p yields p = 0.9761932457

The optimal point yielding the most profit is the point where costs of labour and potential costs of poor service are at their combined lowest. The following equation demonstrates this:

*It was assumed each month has 30 days

$$Total\ Cost = Labour\ Cost + Cost\ of\ Poor\ Service$$
$$= \left(\frac{Employee\ Cost\ per\ month}{30\ days\ pe\ month}\right) \times (Number\ of\ Employees) + 20000 \times (Probability\ of\ poor\ service)$$

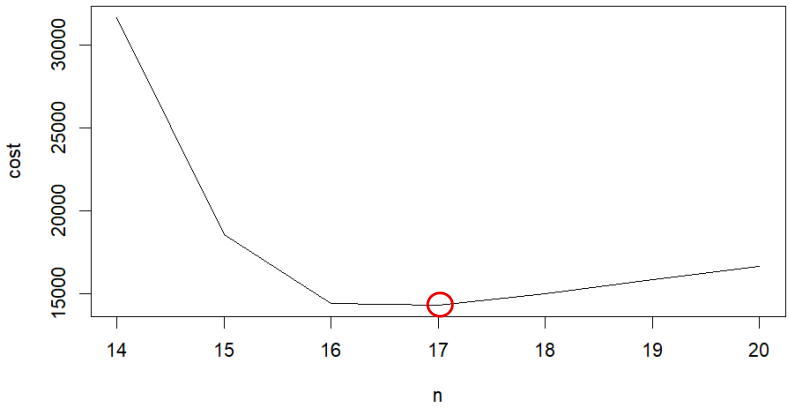
$$Probability\ of\ poor\ service = P(X \leq 14) = \sum_{i=0}^{14} \binom{n}{i}(p)^{i}(1-p)^{n-i}$$

*X is the number of employees showing up on any given day

*n is the number of employees

Using the above equations, the following cost graph of cost vs number of employees was drawn to identify the optimal point.

	n	cost
1	14	31666.67
2	15	18566.28
3	16	14423.87
4	17	14309.55
5	18	15015.04
6	19	15834.68
7	20	16666.77



From the graph and the table, it is seen the total cost is minimized when 17 employees are hired. This minimizes the risk of poor service while simultaneously minimizing labour cost.

Conclusion

This concludes the report covering the ECSA graduate attribute report completed for Quality Assurance 354. All questions and prompts were duly completed and all relevant code submitted with the report in separate files.