ENGINEERING COUNSEL OF SOUTH AFRICA (ECSA) REPORT

Loubser, PE, Mr [24146781@sun.ac.za]

Contents

Introduction

Data has always been a valuable resource for a company, especially if utilised correctly. In today's modern age of information and technology, that has become even more true.

This report covers an in-depth data analysis of a company, with a brief overview of general information, followed by an extensive report of the delivery times. The general overview consists of descriptive statistics, such as trends and patterns. Statistical Process Control is performed on the Delivery Times, and this is done using different control limits and time frames. The data is then studied, and recommendations on optimisations is made.

Part 1: Data Wrangling

In the dataset, there were values that were not possible, such as NA, and negative values in the Price category. These values were removed (cleaned), and the data was ordered according to the date, from oldest to newest data.

Part 2. Descriptive statistics.

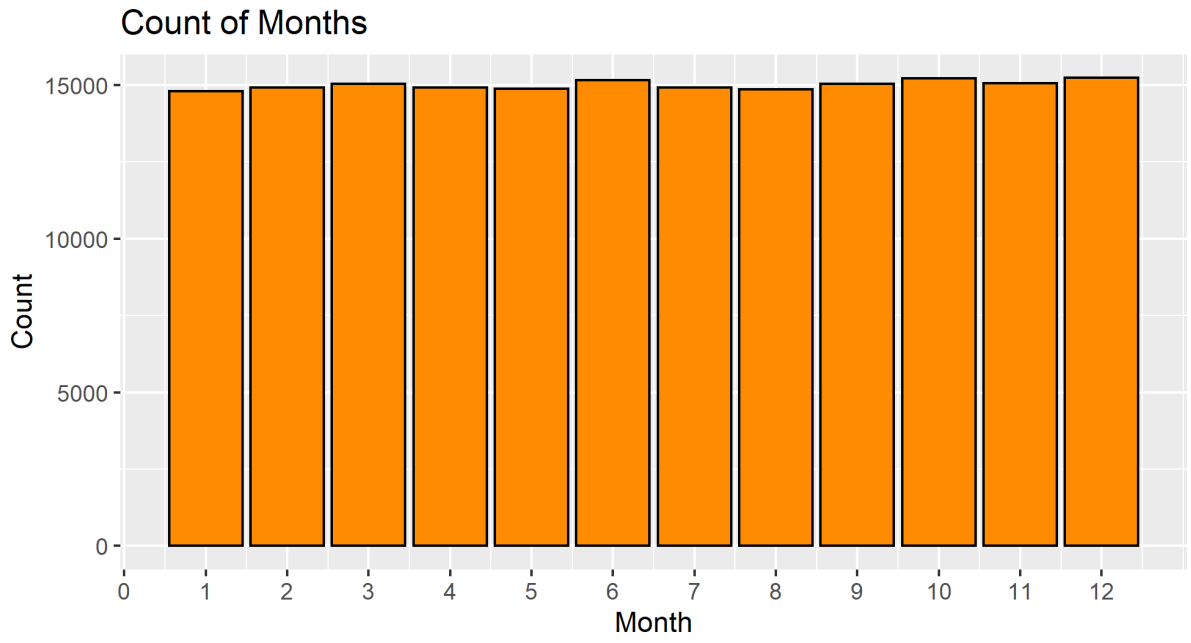


Figure 1

Figure 1 simply counts how many products were bought in each month. There are slight peaks in June, October and December. The most likely reason for June and December is that it is school holidays, however, this does not explain October. The next plot might give us more detail.

Figure 2

Comparing the previous plot in more detail, namely viewing the sales per month by class might allow us to see what drives the peak months as discussed in figure 1. It does not provide more clarity on why there is a peak in October, but it does still give general insight into the shopping habits of customers. This figure confirms by splitting figure 1 into classes that there is very little variation across the classes from month to month, which allows us to assume that from year to year it will follow the same trend across different classes.

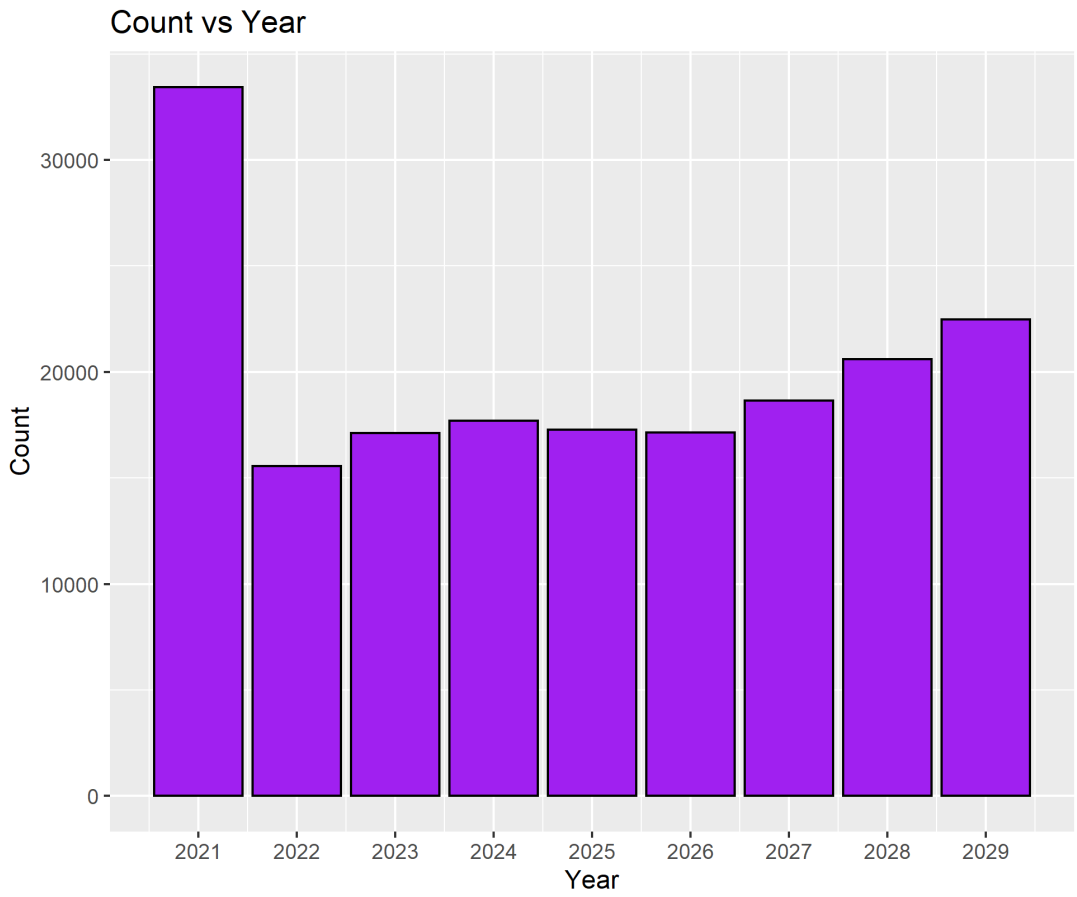


Figure 3

The figure above shows a dramatic drop after 2021. It is difficult to know whether this decline could have been expected, but it is unlikely. It is more likely that this was because of unforeseen circumstances (like a possible pandemic), and the gradual incline after this is possible because of economic growth. Another possibility is that the data was incorrect in 2021, or incomplete after 2021, although this is unlikely.

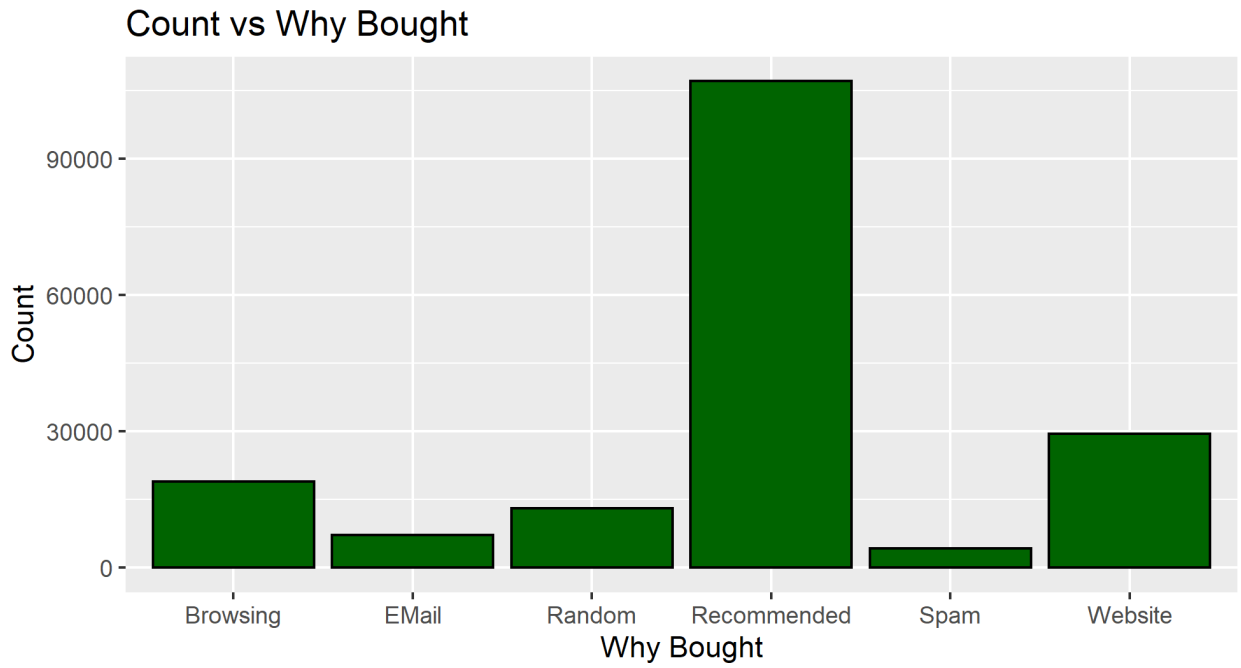


Figure 4

This graph gives an overview of the reasons why customers bought what they did, and it clearly shows that recommended sales are by far the most recorded. This can be exploited in advertising, knowing what specific prompts customers respond to well. The next figure will give more detail of figure 4, and will allow us to have a better understanding of the reason for buying and relating this to yearly trends.

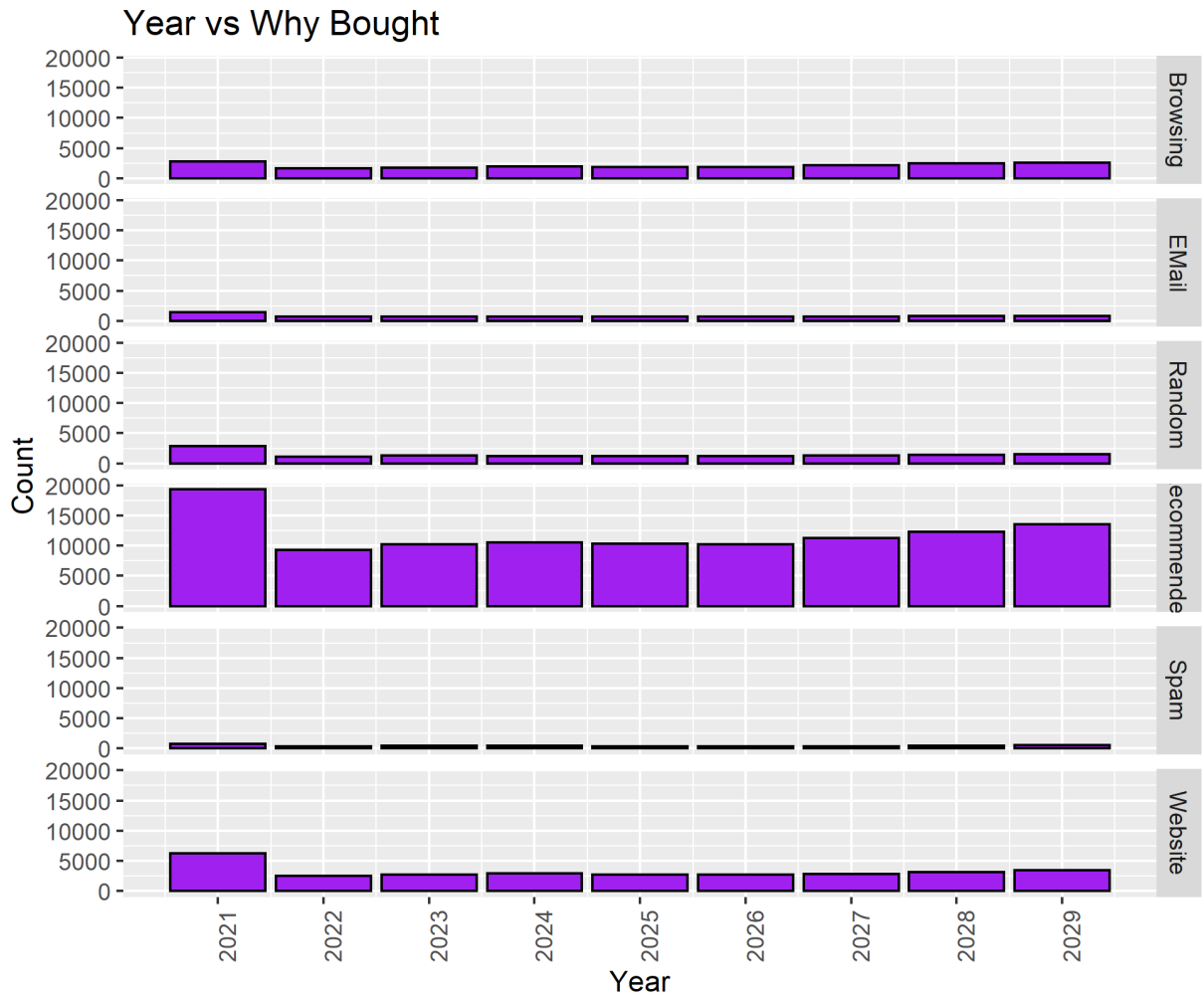


Figure 5

The figure clearly supports the figure 4 in terms of trends, with some extra information, such as an outlier in 2021, then a dramatic drop, and then a gradual increase. What this figure informs us of is that this gradual increase is largely driven by recommended sales. A possible reason for this is the growing influence of social media, people interact more and more. The age of information can be a very powerful tool if used correctly. The age demographic will later be discussed, and hopefully that data will support the previous statements.

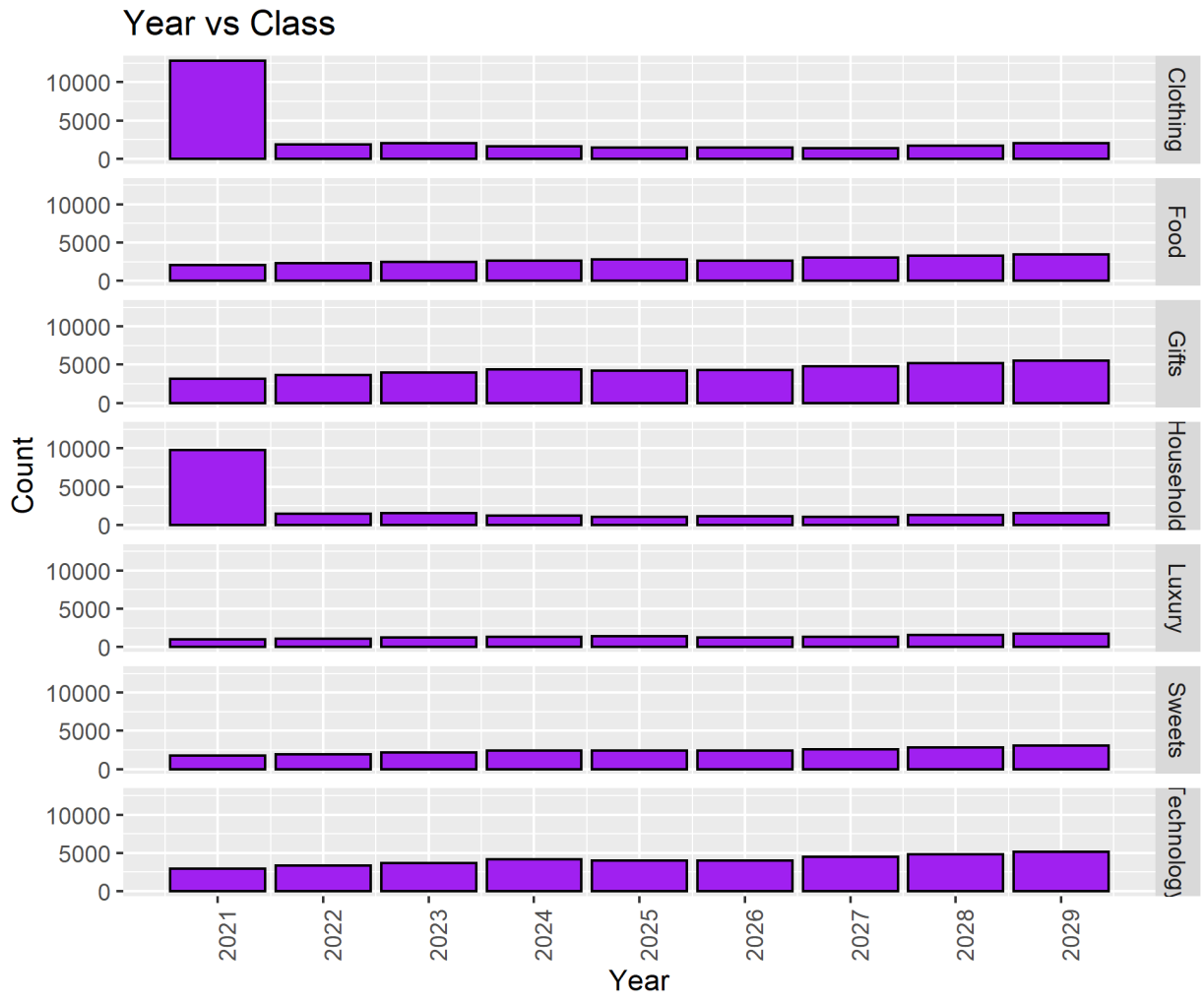


Figure 6

Figure 6 shows that sales in 2021 was driven to a large extent by household and clothing items, which plummeted after 2021 and stayed flat, not following the increase the rest of the classes followed.

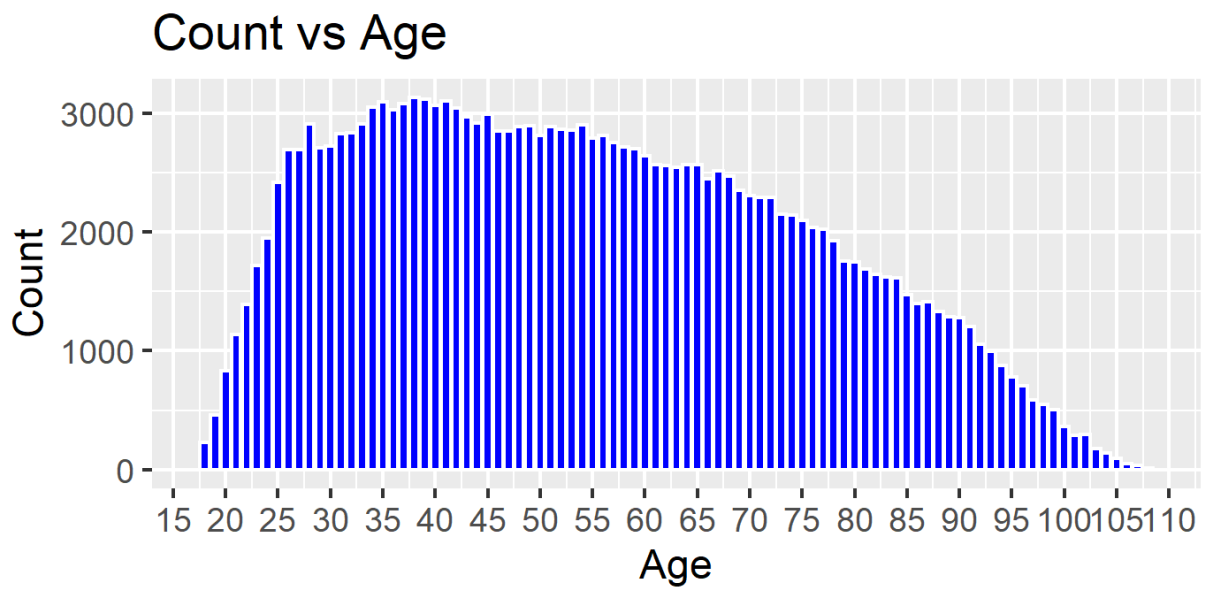


Figure 7

Age is one of the 1st plots with a clear, predictable distribution. The number of sales per age is skewed to the right, which means the majority of the value is cluttered to the left end of the distribution. The peak is in the age group 35-45, which supports the logical assumption of middle-age spending, since the middle age person typically buys for two generations, themselves and their children.

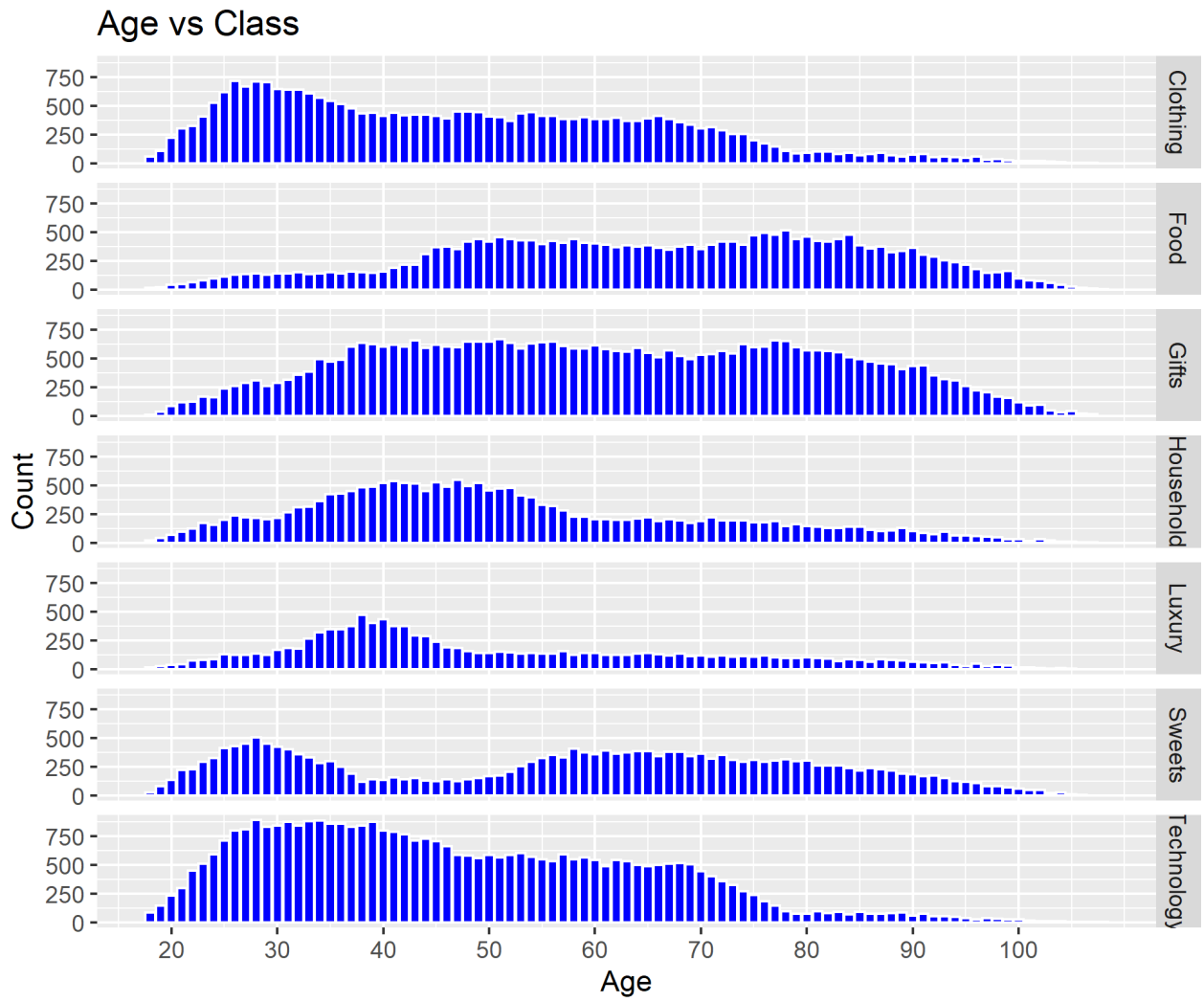


Figure 8

This figure highlights some interesting spending habits. Most of the data is skewed left, such as clothing and technology, but then there are also different distributions. Sweets has a bimodal distribution, peaking around age 28, and then again from 55-65. This indicates that the younger population, up to around 30, have a sweet tooth. This largely disappears for a few years, and then reappears at around ages 55 to 75. Food and gifts are more widely distributed, but both are bimodally distributed, with distinct peaks that follow a loose normal distribution around the peaks.

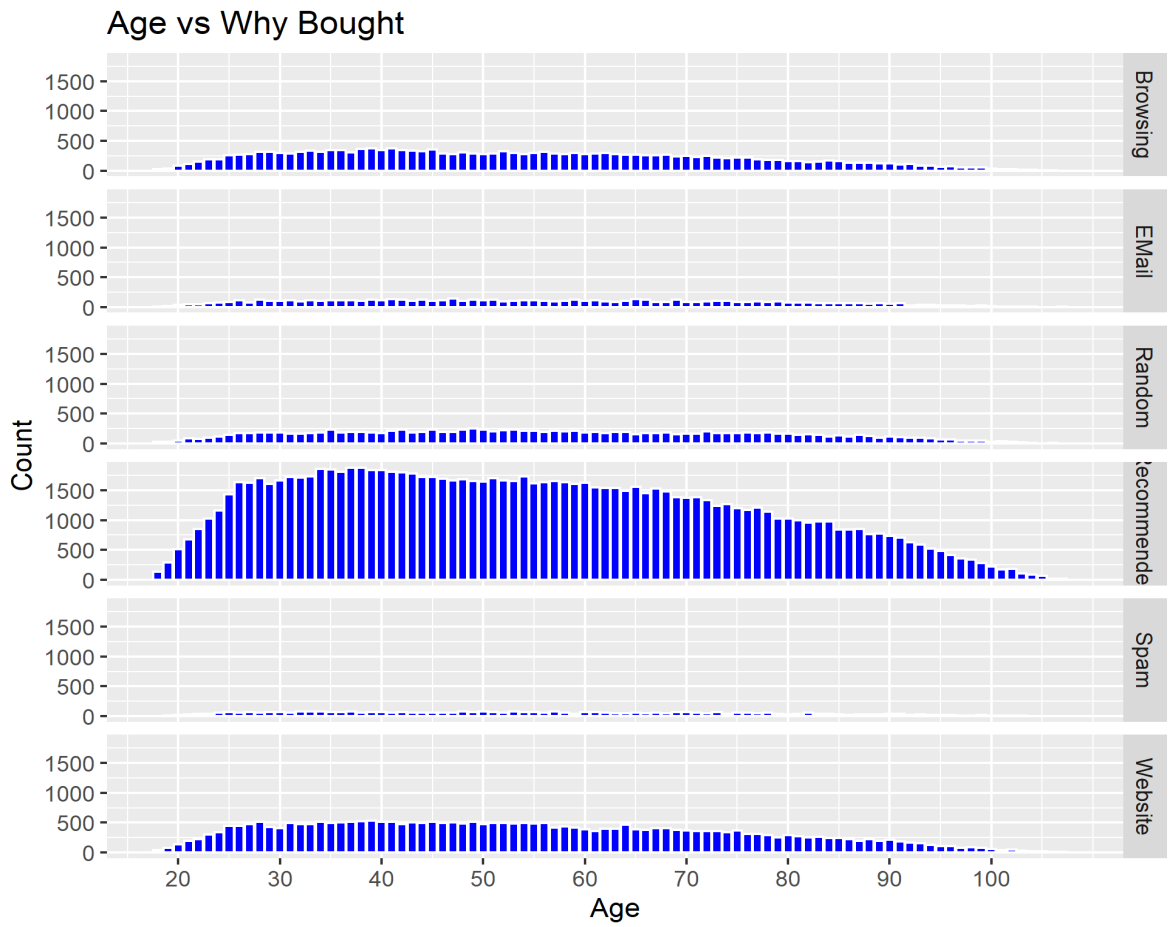


Figure 9

Age plotted against the reason for buying is very predictable, following the same distribution as the count of ages with no exceptions, as well as following the trend of recommended being by far the biggest contributor.

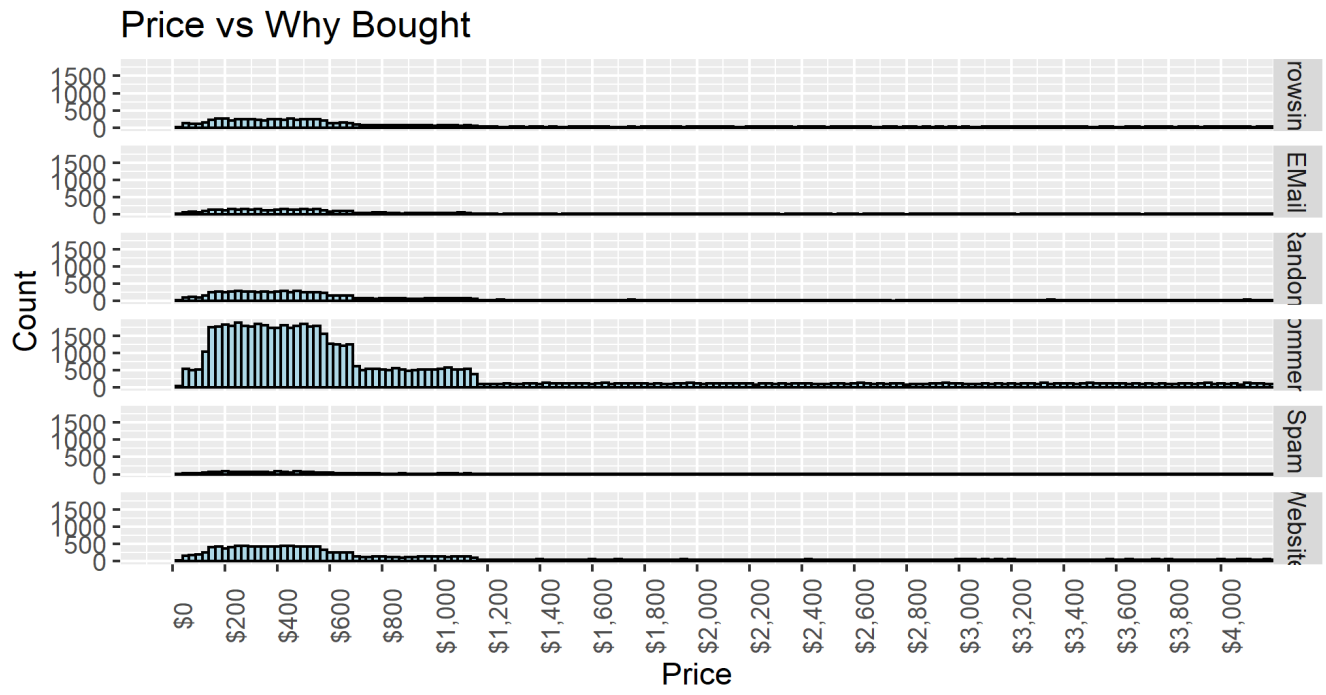
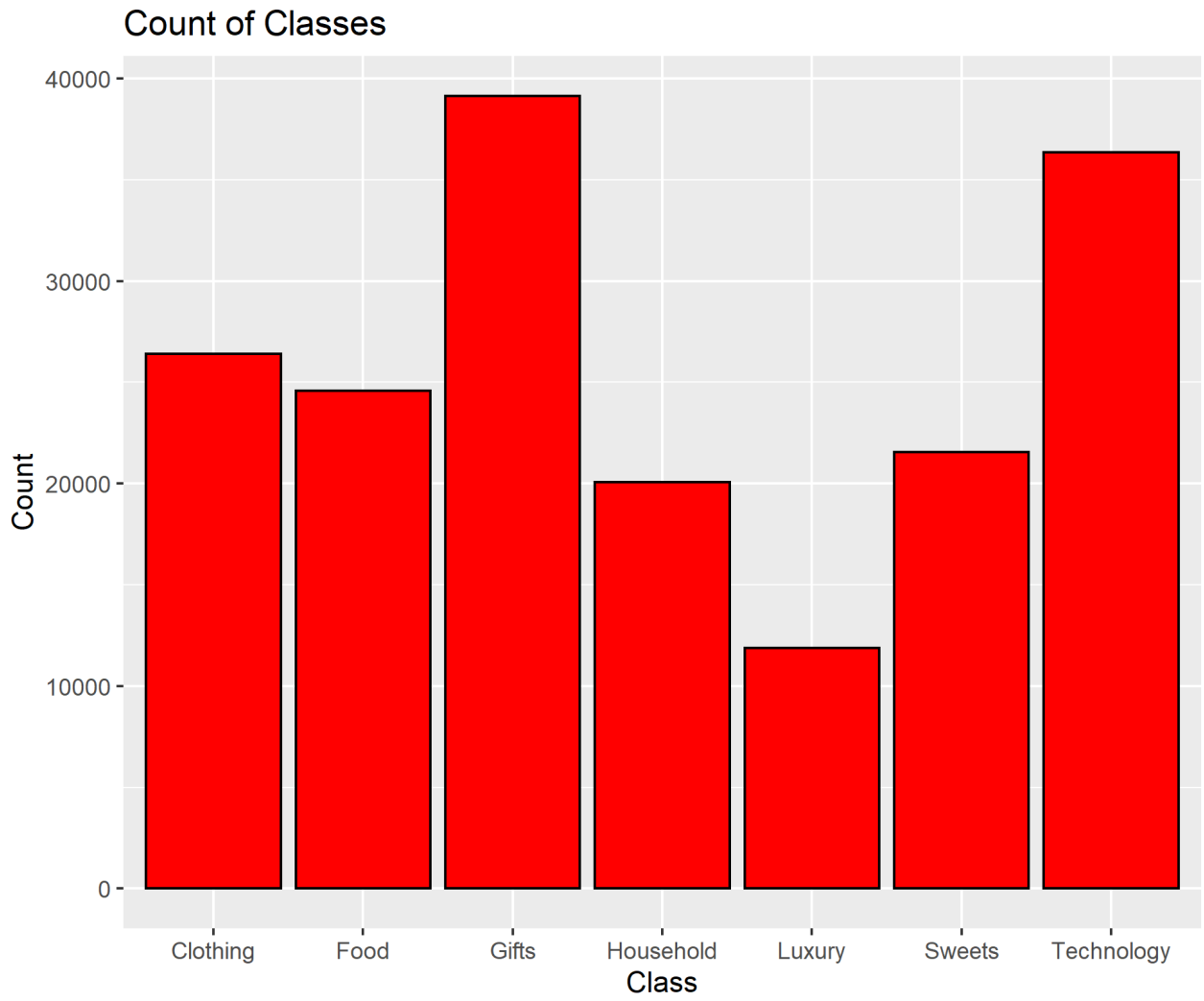
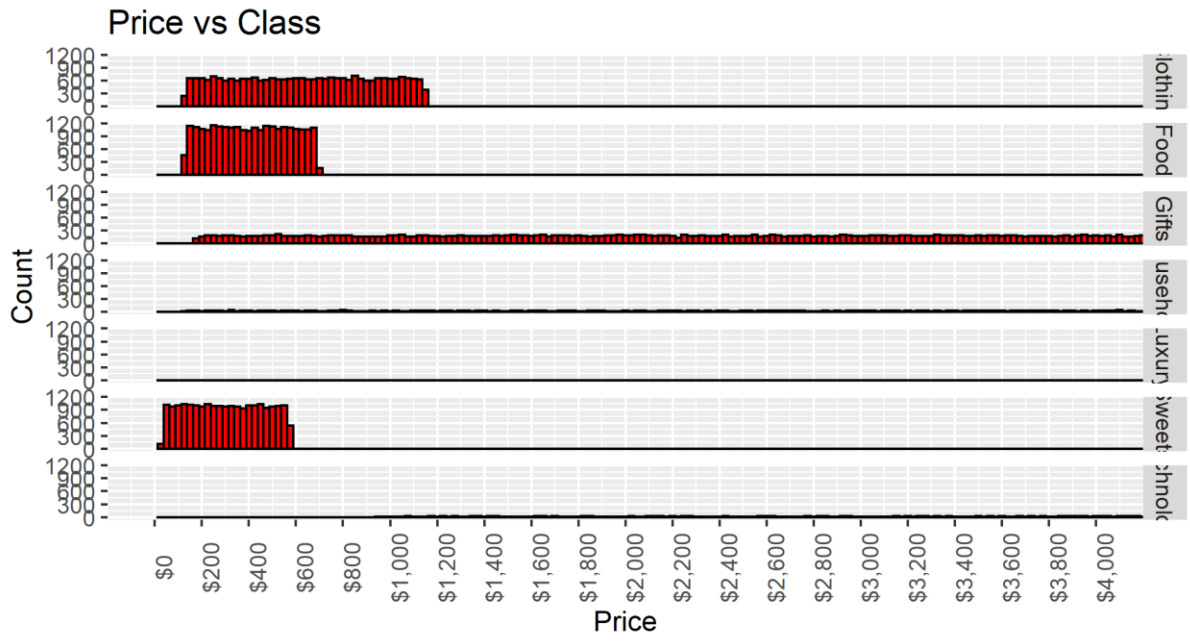


Figure 10

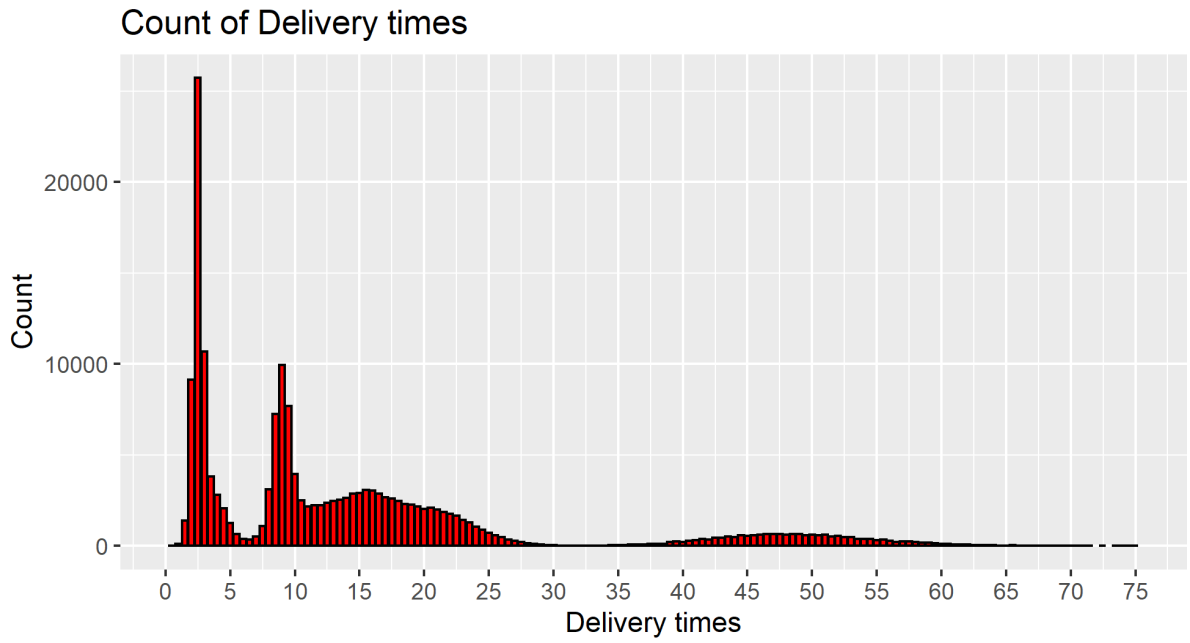
Price compared to the reason for buying is skewed to the right, in proportion to the reason for buying as previously discussed. The price ranges will later be discussed in more detail.



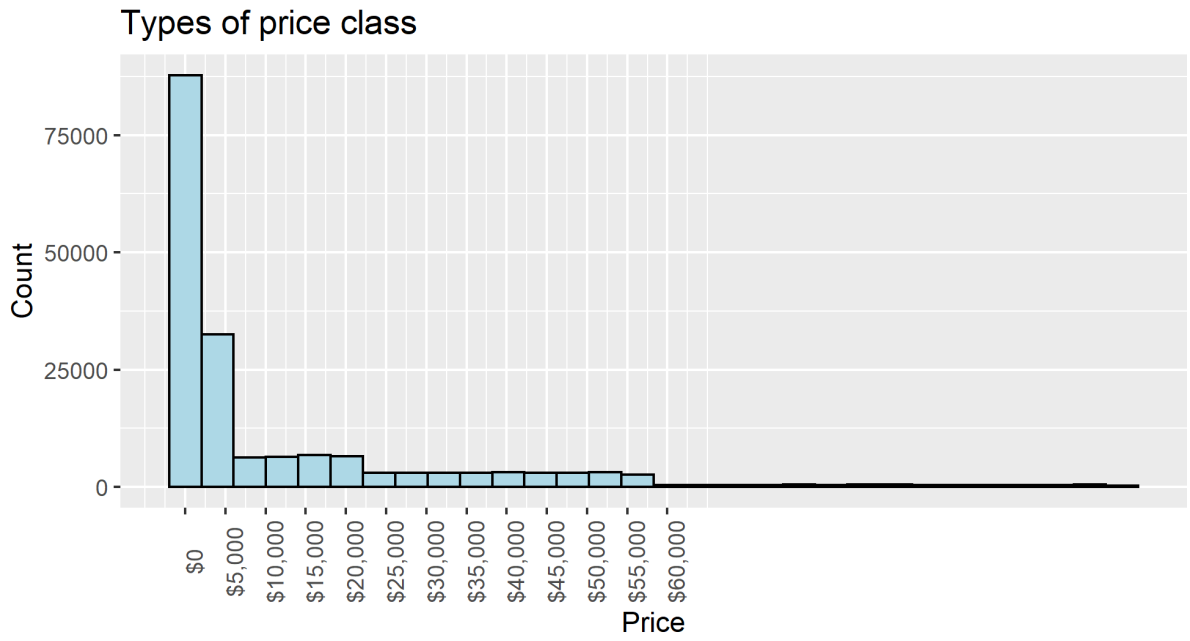
The above figure counts the number of sales that were recorded in each class. The most sales were recorded in gifts, followed by technology, followed by clothing, then food, sweets, household items and luxury goods. This distribution might be explained by a graph detailing the prices of each class, as it is possible that expensive items have lower volumes (such as luxury).



It is interesting to note that the count of different price classes per class is evenly (uniformly) distributed over different ranges, i.e., gifts is distributed from around $200 all the way to $4000, whereas sweets is evenly distributed from very cheap up to around $6000.



The count of delivery times is an interesting distribution. It is skew to the right, with 4 distinct, decreasing peaks from left to right. This is likely caused by the different types/distances of delivery. The majority of deliveries are short, quick deliveries, then as ek gets longer the mode of transportation might change, and it might be necessary to wait for a shipment, etc.



Comparing the volumes of different price classes shows us that the data is strongly skew to the right, showing that a strong majority of sales lie below $5000, which supports logic. This information is important for companies to utilise, as it clearly shows in which price range lies the most sales, but also the most opportunity for growth.

Cp Values:

Cp = 1.142

Cpu = 0.404

Cpl = 1.881

Cpk = 0.404

The Cp (process capability ratio) indicates how the distribution compares to the width of the specifications. Cpk (process capability index) indicates the conformance to specifications.

A low Cpk value usually mean that improvements are necessary. A industry standard is 1.33, and considering this company has a Cpk value of 0.404, it indicates that the process should be improved.

Part 3: Statistical Process Control

Control charts have a specific goal, which=h is splitting variation into common causes and special causes. We used the first 30 samples to set the control limits.

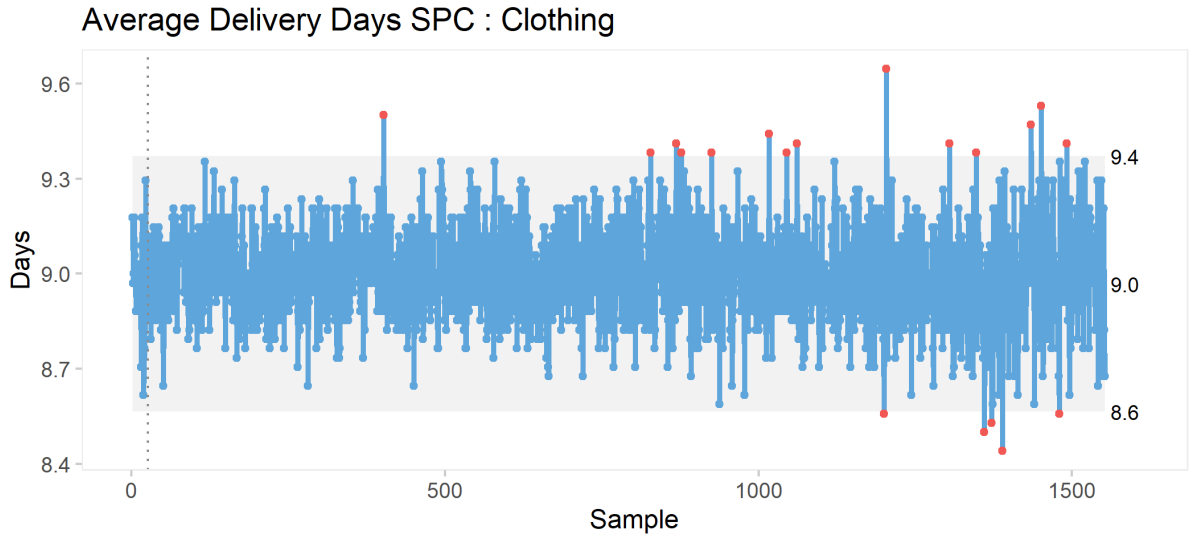
Chart tables

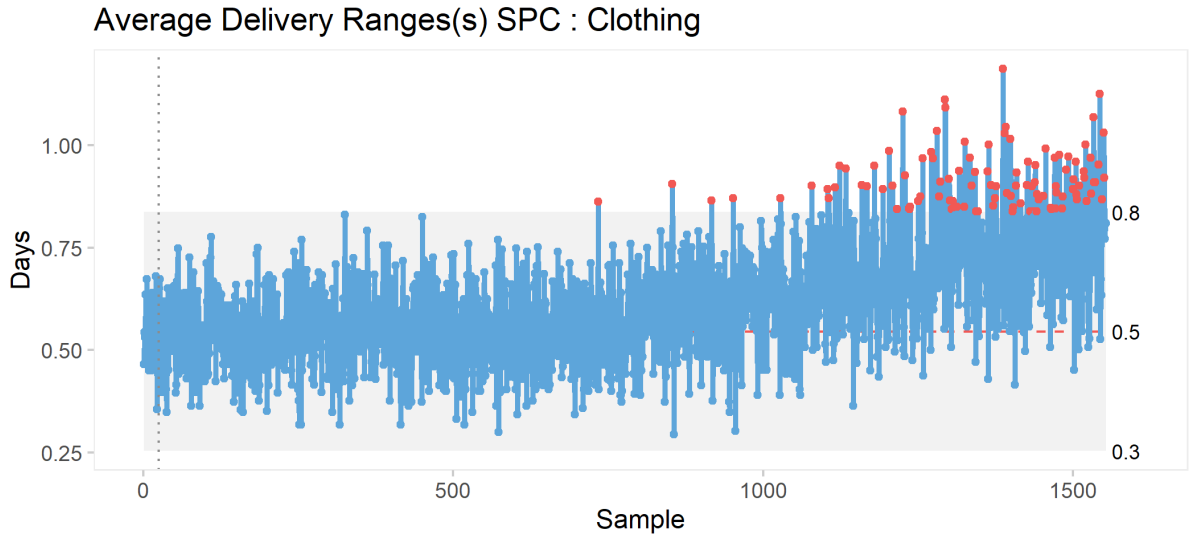
Table 1

Class	UCL	UCL2	UCL1	CL	LCL1	LCL2	LCL
Clothing	0.8363	0.7392	0.6421	0.5451	0.448	0.3509	0.2538
Household	7.1834	6.3495	5.5156	4.6818	3.8479	3.0141	2.1802
Food	0.4282	0.3785	0.3288	0.2791	0.2294	0.1797	0.13
Technology	5.0931	4.5019	3.9107	3.3194	2.7282	2.137	1.5458
Sweets	0.8149	0.7203	0.6257	0.5311	0.4365	0.3419	0.2473
Gifts	2.2054	1.9494	1.6934	1.4374	1.1814	0.9253	0.6693
Luxury	1.4722	1.3013	1.1304	0.9595	0.7886	0.6177	0.4468

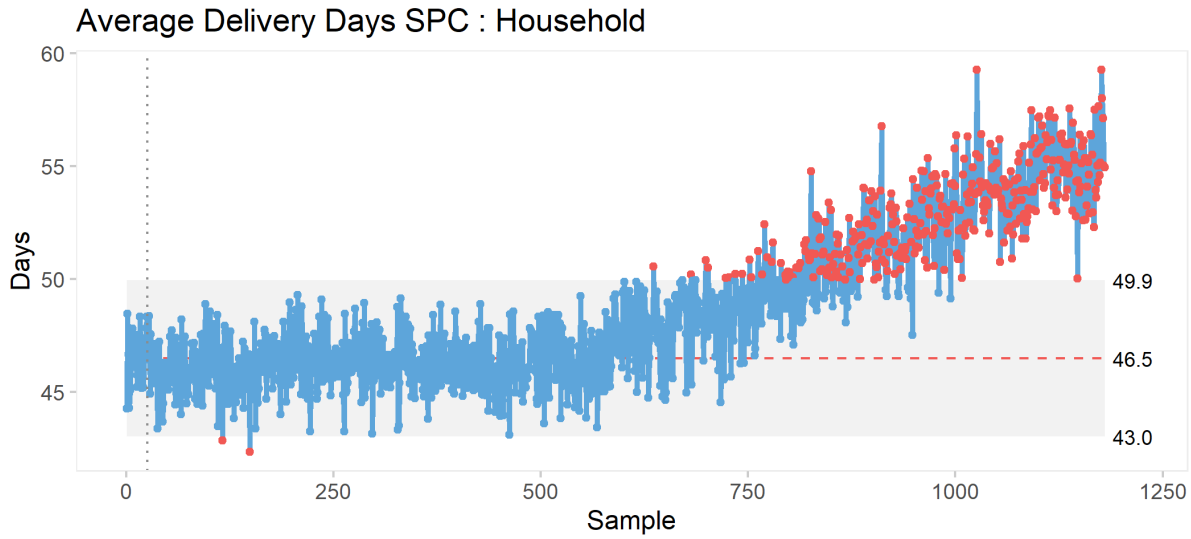
SPC Charts

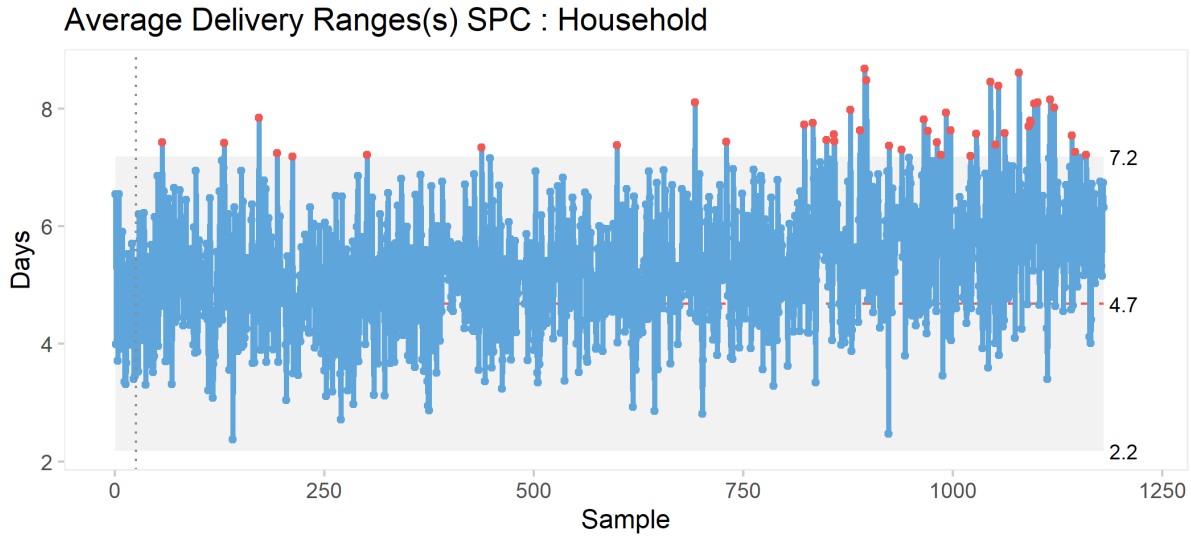
Clothing:



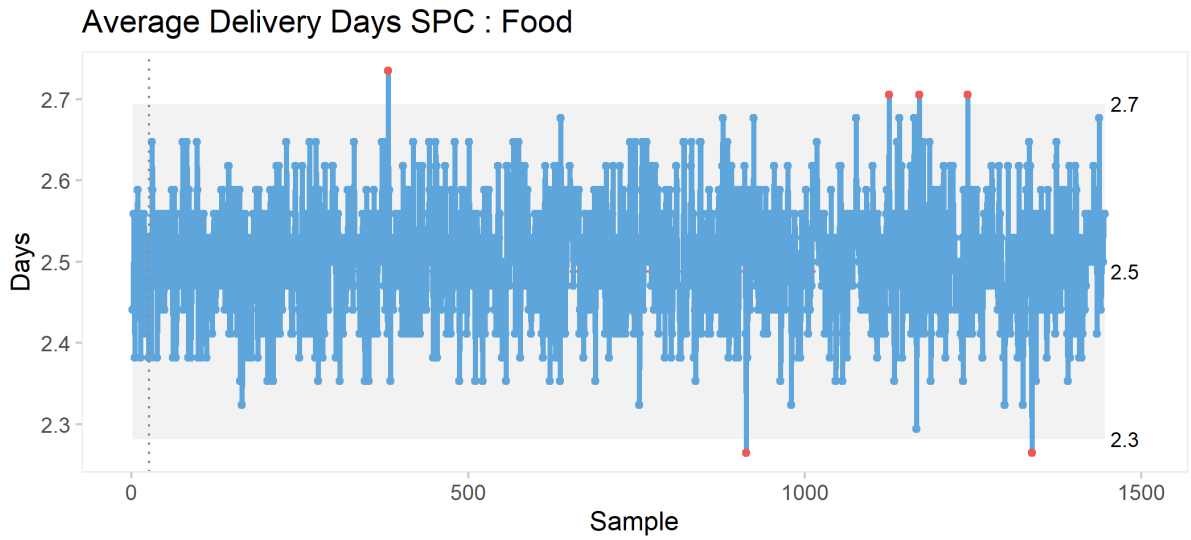


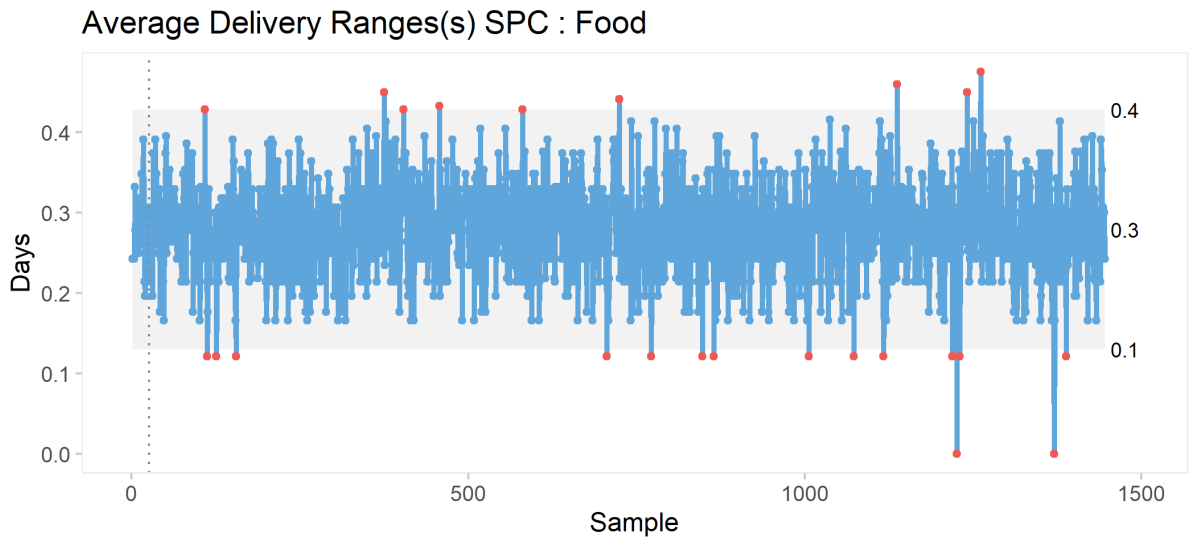
Household:


Average Delivery Days SPC : Household
Days
60
55
50
45
49.9
46.5
43.0
0
250
500
750
1000
1250
Sample

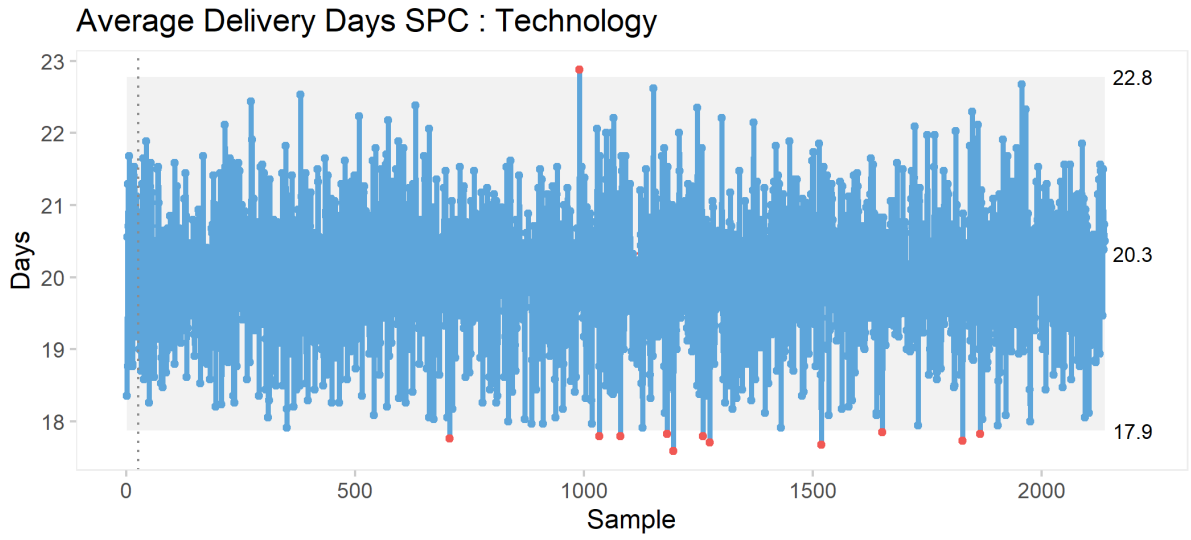

Average Delivery Ranges(s) SPC : Household
Days
8
6
4
2
7.2
4.7
2.2
0
250
500
750
1000
1250
Sample

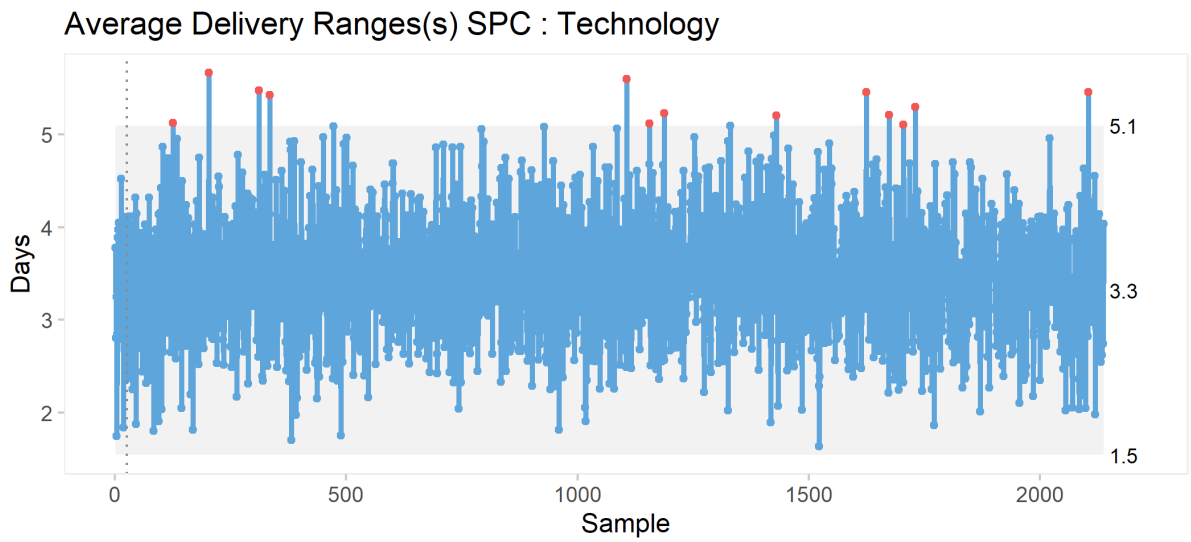
Food:



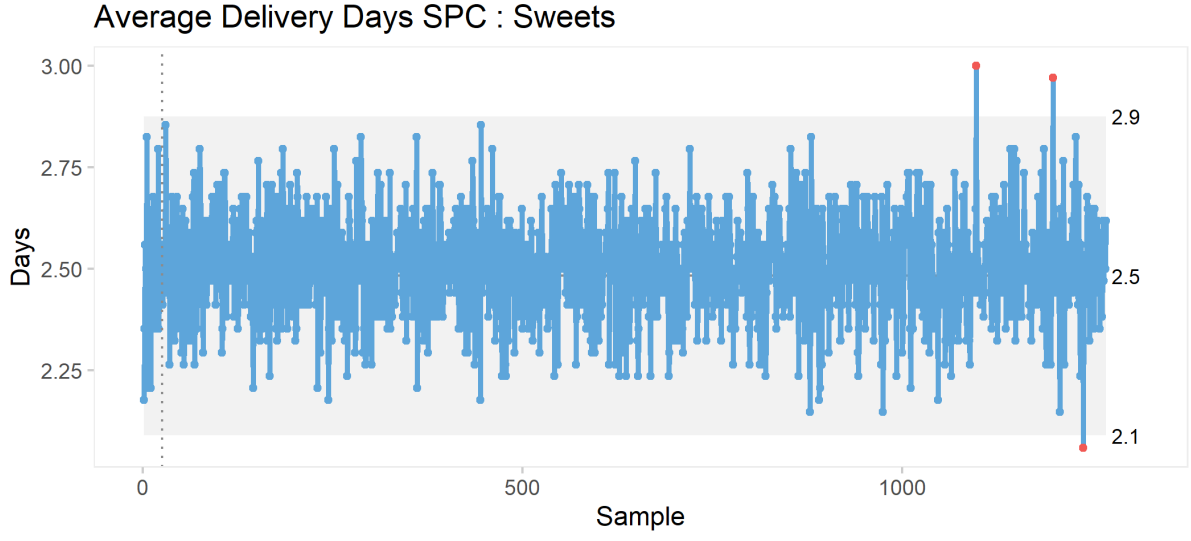


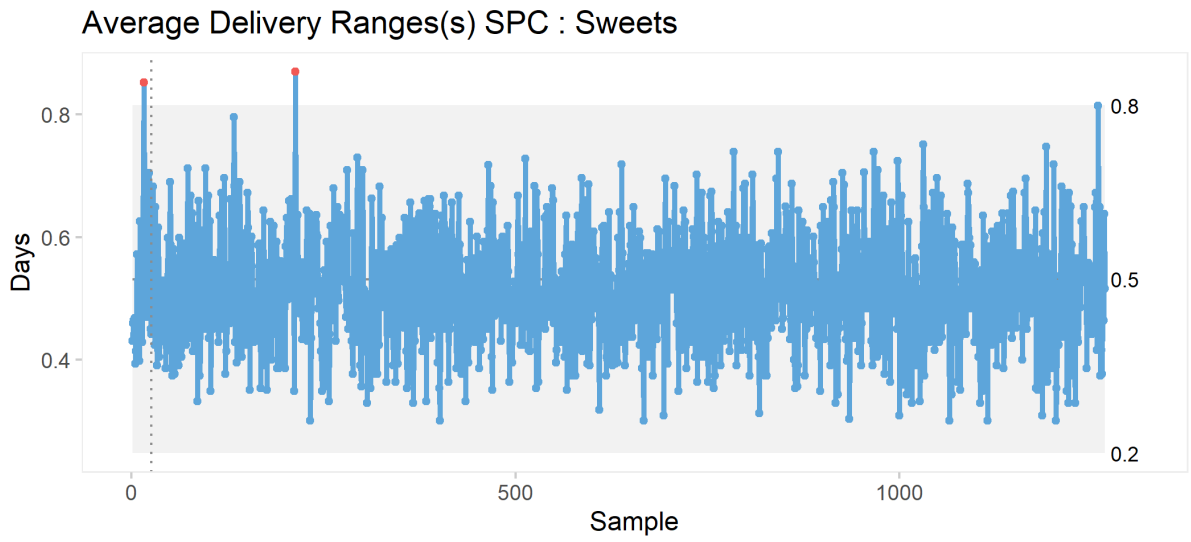
Technology:


Average Delivery Days SPC : Technology
Days
Sample
22.8
20.3
17.9

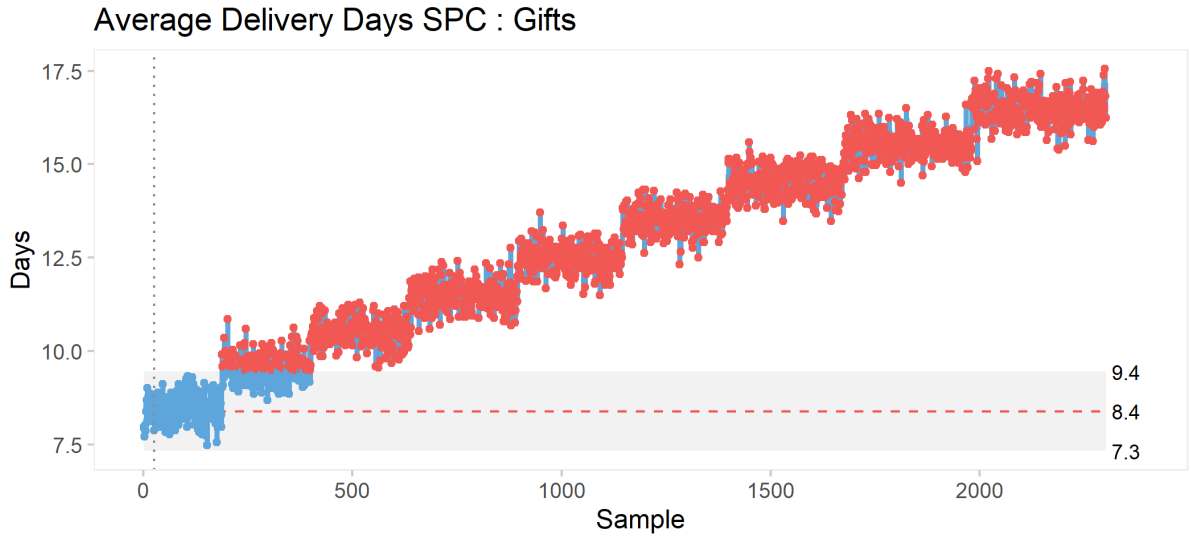

Average Delivery Ranges(s) SPC : Technology
Days
Sample
5.1
3.3
1.5

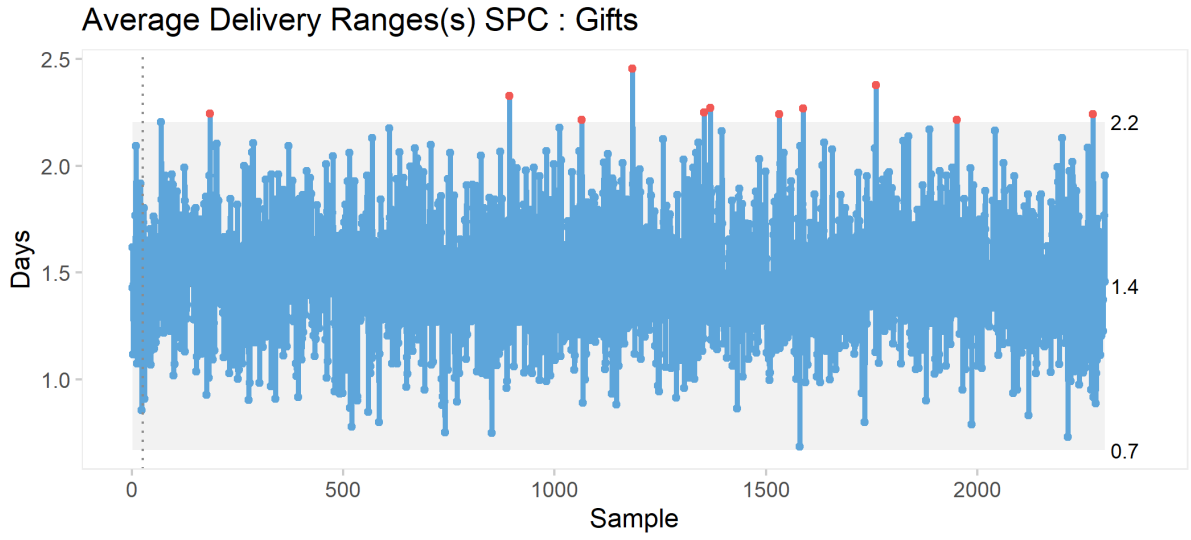
Sweets:


Average Delivery Days SPC : Sweets
Days
Sample
2.9
2.5
2.1

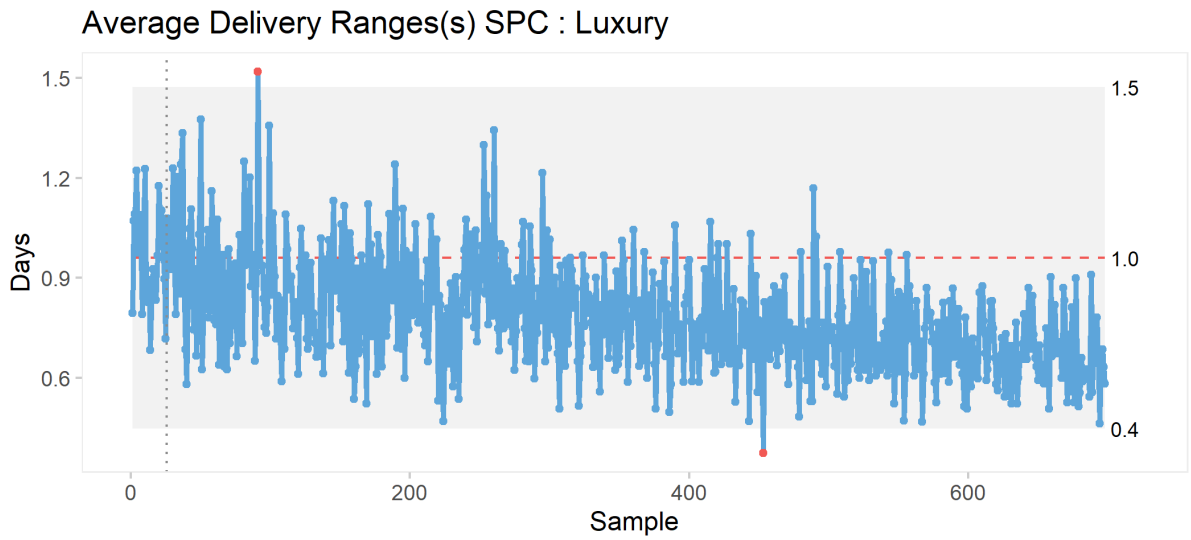


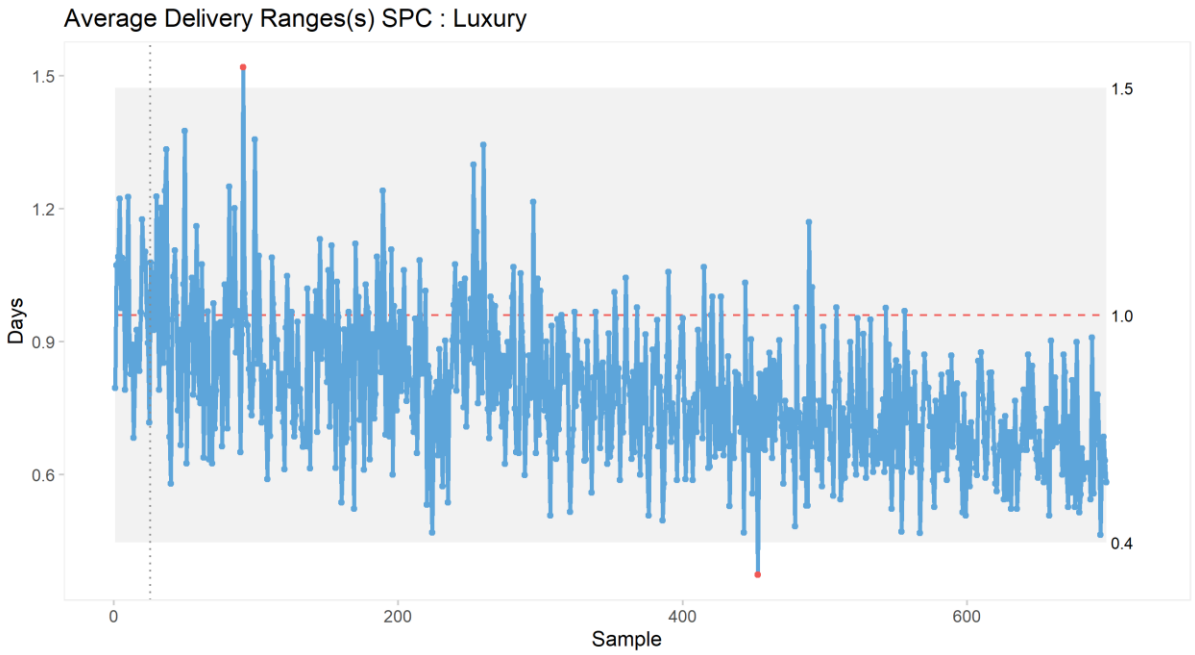
Gifts:





Luxury:


Average Delivery Ranges(s) SPC : Luxury
Days
1.5
1.2
0.9
0.6
1.5
1.0
0.4
0
200
400
600
Sample


Average Delivery Ranges(s) SPC : Luxury
Days
1.5
1.2
0.9
0.6
1.5
1.0
0.4
0
200
400
600
Sample

Part 4: Optimising the delivery process:

Table 2

Class	Total found	1st	2nd	3rd	3rd Last	2nd Last	Last
Clothing	19	402	828	869	1451	1480	1492
Household	360	116	149	636	1178	1179	1180
Food	6	381	913	1125	1170	1242	1337
Technology	12	707	990	1033	1652	1827	1866
Sweets	3	1097	1198	1238	NA	NA	NA
Gifts	2028	188	190	192	2300	2301	2302
Luxury	415	126	151	162	696	697	698

Gifts have the most values outside the control chart, with 2028, which means it definitely has to be inspected. This is followed by Luxury and household items, with 415 and 360 items respectively, which should also warrant an inspection. The rest of the items are of an acceptable level.

Part 5: DOE and Manova

Hypothesis

H0: The class of a product sold does not have an effect on the price of the product and the delivery time of the product.

H1: The class of a product has an effect on at least the price of the product or the delivery time of the product .

Independent variable: Class

Dependant variables: Price & Delivery time

P- value: 0.05

Table 3

```
 Response Price :
               Df     Sum Sq    Mean Sq F value    Pr(>F)
Class           6 5.7168e+13 9.5281e+12   80258 < 2.2e-16 ***
Residuals  179971 2.1366e+13 1.1872e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Delivery.time :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Class           6 33458565 5576427  629429 < 2.2e-16 ***
Residuals  179971  1594452       9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A Manova table is used to generate a p-value for the dependant variables that can indicate if there is any correlation between the dependant variables and the independent variables. In this case the dependant variable is the class of the product and the two independent variables are the price and the delivery time of the product. The p values for price and delivery time is 2.2e-16 and is smaller than 0.05, indicating that the price and delivery time differs, depending on the class.

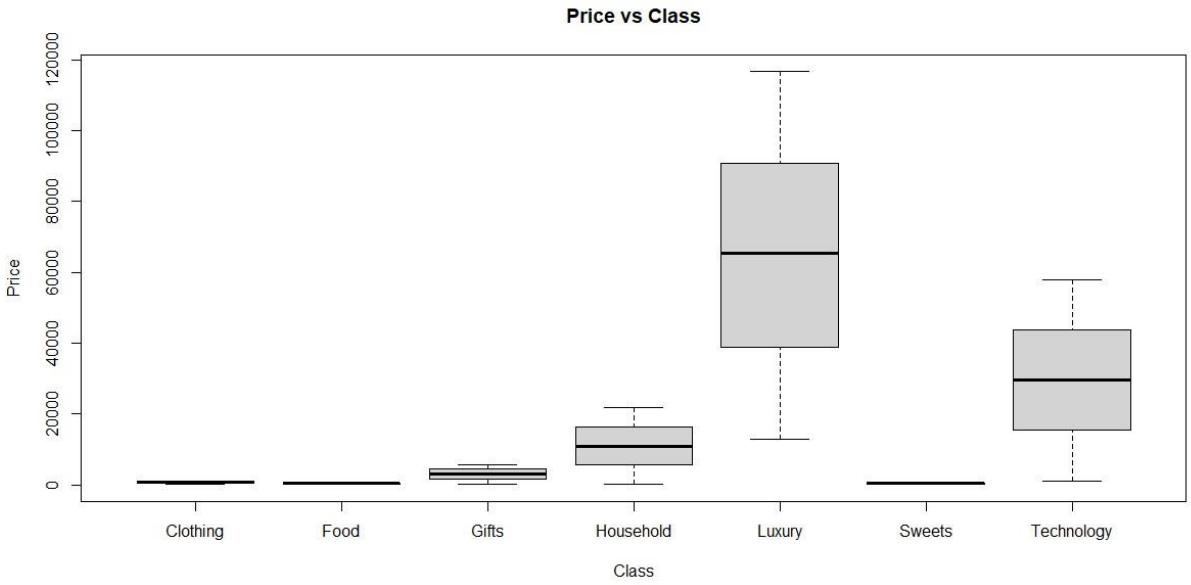


Figure 11

Figure 12

The box plots for the price of the products and the different classes and the delivery time of the products and the different classes also show that there are clear differences in price and delivery times for the different classes. Thus, the null hypothesis is rejected. The class of a product has an effect on the price and the delivery time.

Part 6: Reliability of services and products:

The Taguchi loss function states that variation away from the target feature will cause dissatisfaction, and attempts to quantify this dissatisfaction. With and increase in variation, there is an increase in dissatisfaction.

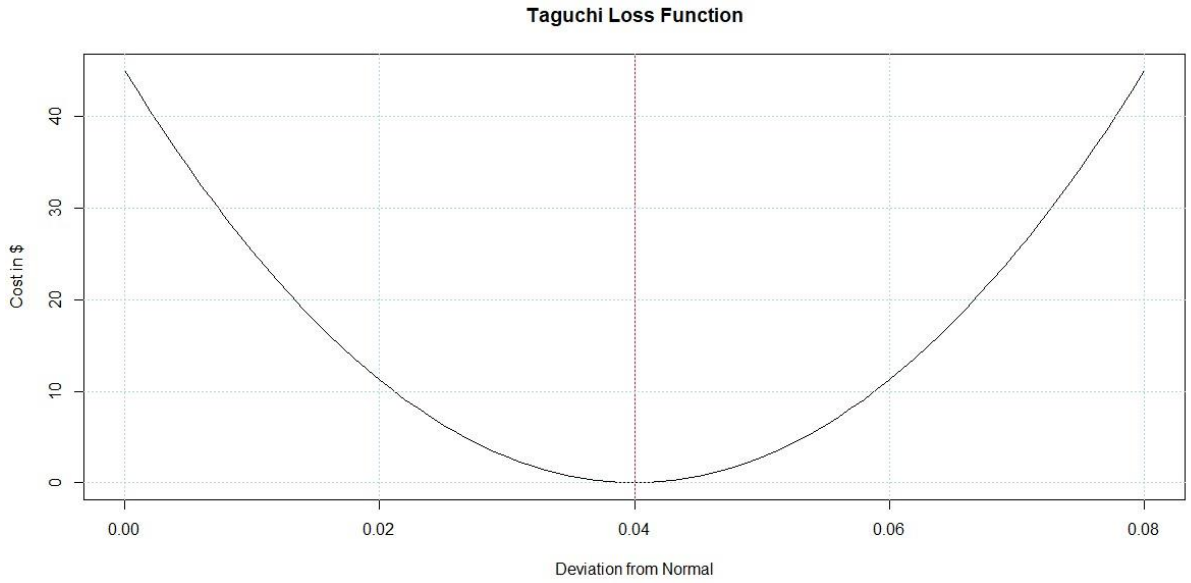
Problem 6:

t <- 0.06 #target

D <- 0.04 #deviation/tolerance

L <- 45 #loss k <- L/(D^2) #constant

Loss function(x) = $k(x - T)^2$



If the refrigerator part is within the limits of 0 to 0.06, the customer will be satisfied.

Problem 7:

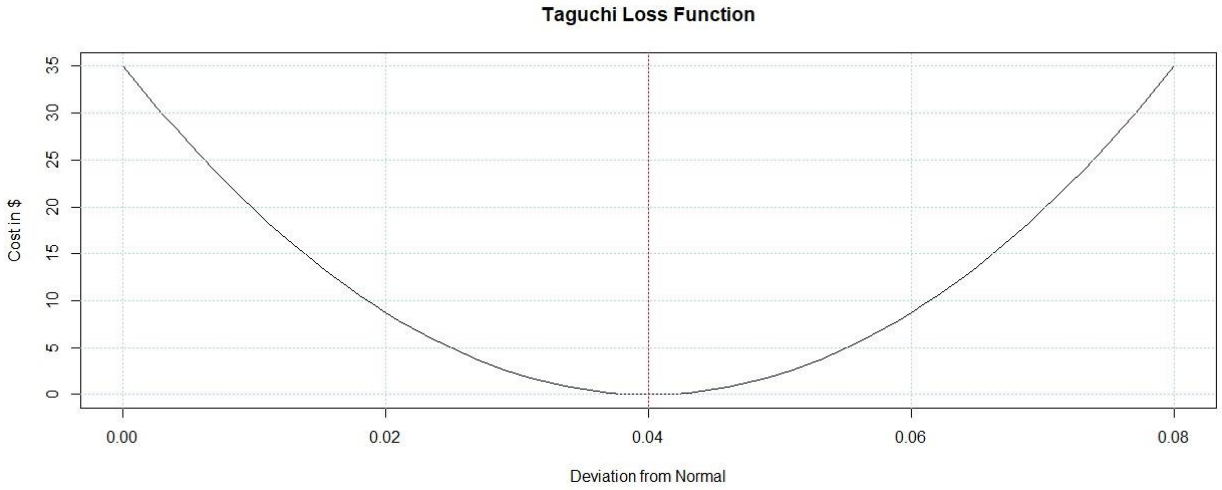
t <- 0.06 #target

D <- 0.04 #deviation/tolerance

L <- 35 #loss

k <- L/(D^2) #constant

Loss function(x) = $k(x - T)^2$



If the part falls within the specs of 0 to 0.08, the customer will be satisfied.

Problem 27:

The probability for success of the system with 1 machine at a station is 0.7038 (70.38%), and the reliability of the system with 2 machines at a station is 0.9615 (96.15%).

Problem 6.3:

Probability for reliable number of vehicles = (1556/1560) = 0.9984359

Probability for reliable number of drivers = (1559/1560) = 0.99935897

Calculations:

```
  pv<-(1556/1560)
n_v<-20
Car.20<-dbinom(20, size = n_v, prob= pv)
Car.19<-dbinom(19, size = n_v, prob= pv)
Car.18<-dbinom(18, size = n_v, prob= pv)
Car.17<-dbinom(17, size = n_v, prob= pv)
p_enough_vehicle<-Car.20+Car.19
#drivers
pd<-(1559/1560)
n_d<-21
d.21<-dbinom(21, size = n_d, prob= pd)
d.20<-dbinom(20, size = n_d, prob= pd)
d.19<-dbinom(19, size = n_d, prob= pd)
```

```
d.18<-dbinom(18, size = n_d, prob= pd)
d.17<-dbinom(17, size = n_d, prob= pd)
p_enough_drivers<-d.21+d.20+d.19
# delivery days per year
p_enough_vehicle*p_enough_drivers*365
#364.5577
##############################
#more_vehicles
pv<-(1556/1560)
n_v<-21
Car.21<-dbinom(21, size = n_v, prob= pv)
Car.20<-dbinom(20, size = n_v, prob= pv)
Car.19<-dbinom(19, size = n_v, prob= pv)
Car.18<-dbinom(18, size = n_v, prob= pv)
Car.17<-dbinom(17, size = n_v, prob= pv)
p_enough_vehicle<-Car.20+Car.19
p_enough_vehicle_more<-Car.21+Car.20+Car.19
# delivery days per year [more vehicles]
p_enough_vehicle_more*p_enough_drivers*365
#364.992
```

Reliable delivery days per year with 20 vehicles and 21 drivers = 364.557 = 364 days

Reliable delivery days per year with 21 vehicles and 21 drivers = 364.992 = 364 days

Conclusion:

After the provided data about the online business is cleaned, visual representations were generated to better understand and interpret the data. The process capability indices were also calculated and used as an indication of how the distribution compares to the width of the specifications.

The statistical process control calculations were performed, where the control limits were set by using the first 30 samples. Xbar charts and S charts were generated for each of the different classes, the different classes that are out of control are identified and the delivery times for technology is optimized.

A Manova table is generated to see if there is any correlation between the chosen dependant variables and the chosen independent variables, and a visual representation of the two independent variables are also showcased. This data analysis will provide the online company with valuable information that will help them improve their business and maximize their profit.

References:

- ESCA, n.d. Additional notes for ECSA Graduate Attributes Project. [Online] Available at: https://learn.sun.ac.za/pluginfile.php/2628705/mod_resource/content/3/SPC2020.pdf [Accessed 02/10/2022].
- Tague, N. R., 2021. CONTROL CHART. [Online] Available at: https://asq.org/quality-resources/control-chart [Accessed 05/10/2022]
- Anon, (n.d.). *Taguchi Loss Function – Lean Manufacturing and Six Sigma Definitions*. [online] Available at: https://www.leansixsigmadefinition.com/glossary/taguchi-loss-function/. [Accessed 14/10/2022]
- www.rdocumentation.org. (n.d.). *QIC function - RDocumentation*. [online] Available at: https://www.rdocumentation.org/packages/MuMIn/versions/1.43.17/topics/QIC [Accessed 18/10/2022].