Shipping Data Modelling Project

Amy Lai

Spring 2021

1 Problem

(1)

The target population or process for this study is all purchases made by customers at GoodBuy in 2020. The units are each individual shipping ordered by customers at GoodBuy in 2020.

(2)

In this study, we will be analyzing the Customer rating, Mode of shipment, Discount offered, Customer care calls, Weight in grams, and Reached on time variates. When considering the variate Customer rating, we note that it is recorded to take on integer values from 1 to 5 and as such this makes it a discrete variate. An attribute of interest for this variate is the mean of the target population's customer rating. When considering the variate mode of shipment, we note that it is recorded to take on values from one of the three types: ship, flight and road. As such, this makes it a categorical variate. An attribute of interest for this variate is the proportion of shipments through flight in the target population. When considering the variate Discount offered, we note that it is recorded to take on integer values, hence it is a discrete variate. An attribute of interest for this variate is the mean of the target population's discount offered. When considering the variate Customer care calls, we note that it is recorded to take on integer values, hence it is a discrete variate. An attribute of interest for this variate is the average of the target population's customer rating. When considering the variate Weight in grams, we note that it is recorded to take on integer values, hence it is a discrete variate. An attribute of interest for this variate is the mean of the target population's shipments' weight. When considering the variate Reached on time, we note that it is recorded to take on values of 1 or 0 and as such this makes it a binary variate. An attribute of interest for this variate is the proportion product delivered by the expected delivery date in the target population (shown as 1 in the dataset).

(3)

We will be looking at several motivating questions for the variates of interest, including:

1. "Is the distribution of customer rating similar for both genders in the target population?" This is a descriptive problem as it aims to compare and describe the distribution of the target population's customer rating variate between male and female.

2. "Is the exponential model a reasonable model for modeling the distribution of discount offered?" This is a descriptive problem as it aims to describe the distribution of Discount offered variate using a statistical model.

3. "Is the Poisson model a reasonable model for modelling th edistribution of the number of customer care calls int eh target population?" This is a descriptive problem as it aims to describe the distribution of Customer care calls variate using a statistical model.

4. "Is the Gaussian model reasonable for modelling the weight of the products in the target population?" This is a descriptive problem as it aims to describe the distribution of Weight in grams variate using a statistical model.

5. "What proportion of high importance products were delivered by the expected delivery time?" This is a descriptive problem because it aims to discribe the Reached on time variate by examining high importance products.

2 Plan

(1) If the target population is all purchases made by customers at GoodBuy in 2020 and the study population is all purchases recorded in January to February of 2020, then a possible source of study error is the timing. January to February is a short period of time with special weather conditions like snow. Snow can cause more troubles for shipping during January to Feburary than other months such as June and July. It can also affect the mood of the customer. Thus, collecting datas from such a short period of time can cause errors and biases in the data collected (ex. reached on time, customer rating).

(2) Notice that in the sample of 500 shipments, the number of Ship shipments selected exceeds the other two types of shipments. This can affect the analysis of the variates of interest. For example, Ship shipments tend to have higher weight than Flight shipments. Hence, the average weight among all shipments sampled would be relatively high, and and the result of weight related analysis may not reflect the actual behaviour in the target population.

(3) When measuring the weight of the product, different scales are used to weigh different products in order to increase efficiency. This would cause a random measurement error in the weight variate if the scales aren't all accurate. When measuring the Reached on Time variate, customers will sign the date that they received the package and datas are collected by comparing this time with the expected delivery time. Note that there is a difference between delivery time and the time that customers actually picks up the shipment. This would cause a measurement error in the reached on time variate since some customers may pick up shipments after the expected delivery date and assumed that it was late on delivery. Additionally, the customers will give an online rating for the purchase after they received the shipment. We will be using these ratings for the Customer rating variate. Notice that customers' emotions coming from other incidents in life can create a measurement error in the customer rating variate. Some customers may be angry at receiving low test grades at the time of rating. They may give 1 star to express anger. Meanwhile, others may be happy due to a friends' visit, and they decide to give out 5 stars to express joy. These emotions can greatly impact the accuracy of the shipment rating, thus not reflecting their true satisfaction with the delivery.

3 Data

It should be taken in to discretion that there are missing data for customer rating and gender. When analyzing these variates, we should eliminate the units with missing data. Another thing to note is that there are some large outliers in the discount sample data. Most purchases have a discount around and below 10%. However, there are a few purchases where the discount reached above 60%. It is important to consider these missing data and outliers when analyzing the associated variates of interest.

4 Analysis

Part 1

a)

The bar graph below reports the modal customer rating.

Figure 1: Bar Chart for the Customer Rating of Shipments

Notice that the rating 1 is the mode of the distribution, meaning that most customers in the sampled data gave a rating of 1.

b)

The table below summarizes the frequency of each customer's rating level by gender.

		Customer Rating				
		1	2	3	4	5
Gender	M	60	55	42	48	48
	F	55	47	43	59	38

Table 1. Frequency for Each Customer's Rating Level by Gender

c)

The figure below is a side-by-side bar graph for Customer rating by gender

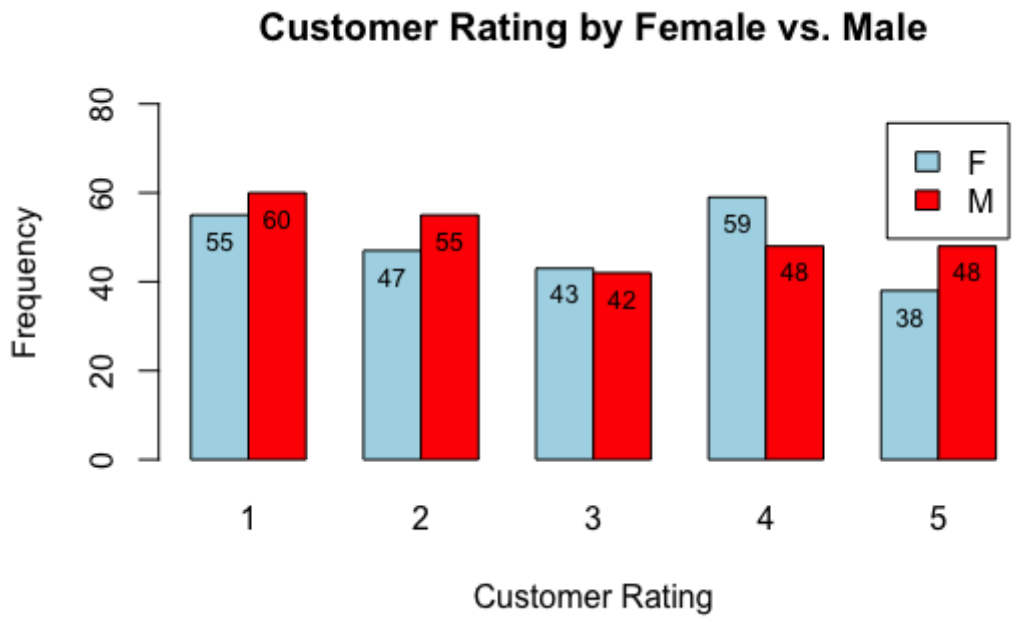

Figure 2. Customer Rating by Female vs Male

Observe that the mode of male customer rating is 1 and male rating are more concentrated on the two ends (1 and 5). On the other hand, the mode of female customer's ratings is 4 and the female ratings are more randomnized and display a less obvious trend than male ratings.

d)

Notice that the Rating variate classifies each event into success and failures indicated by 1 and 0. It is also repeated independently as each cutomer gives out his/her rating independently without the influence of one another. Thus, we conclude that the binomial model is suitable for the Rating variate.

e)

The table below summarizes the relative frequency of each Rating level by whether the item was delivered on time or not. A Rating of 1 means that the shipment received a rating above or equal to 4, otherwise it receives a Rating of 0.

		Rating	
		0	1
Reached on Time	Y=1	0.36	0.23
	N=0	0.25	0.16

Table 2. Relative Frequencies of Rating Level and Reached on Time Variates

f)

Let θ be the probability of a shipment receiving an excellent rating. Since we assume a Binomial(n, θ) model where the total number of ratinngs is n=60+55+42+48+48+55+47+43+59+38=495 and the total number of excellent rating is 48+48+50+38=184, then the maximum likelihood estimate is given by $\hat{\theta} = 184/495 = 0.372$.

g)

The figure below graphs the relative likelihood function of θ with the 15% likelihood line.

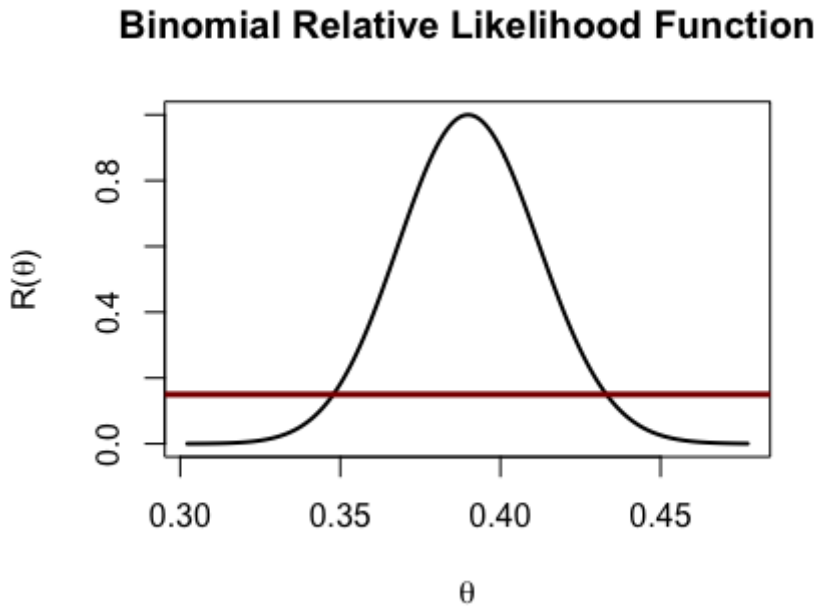

Figure 3. Binomial Relative Likelihood Function of Excellent Rating

Based on the graph and the use of R, we conclude that the 15% likelihood interval for θ is [0.348, 0.433].

Part 2

a)

For this section, we assume Exponential(θ) for the distribution of Discount offered variate, where θ represents the mean discount offered on one purchase recorded in January to February of 2020.

b)

The table below provides the numerical summaries for discount offered.

Minimum	1.000	IQR	6.000
1st Quartile	4.000	Range	63.000
Sample Median	7.000	Sample Mean = $\overline{y}$	11.774
3rd Quartile	10.000	Sample Standard Deviation = s	14.281
Maximum	64.000	Sample Skewness	2.113

Table 3. Numerical Summaries of Discount Offered

Notice that the mean is greater than the median. and the sample skewness is positive. Hence, we conclude that the Exponential model would be suitable for the distribution of discount offered.

c)

The figure below graphs the relative likelihood function for θ.

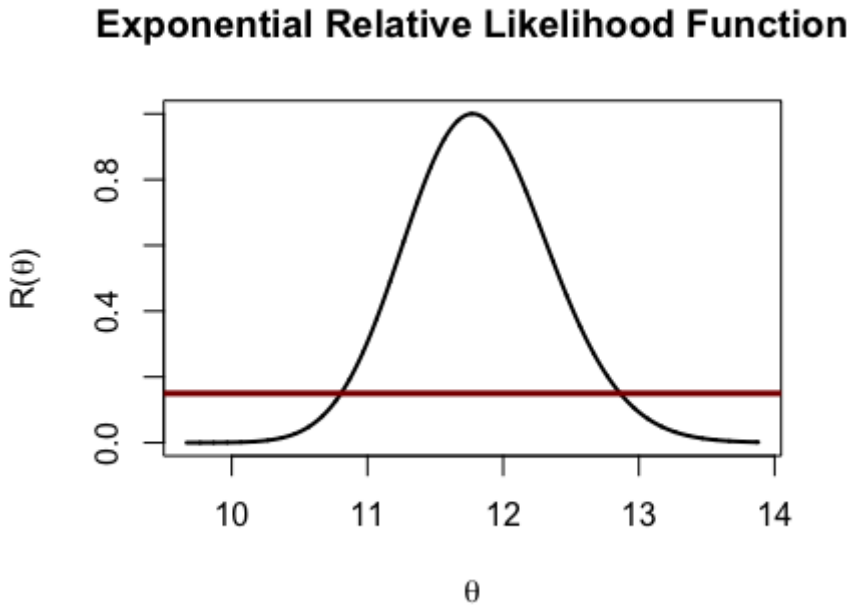

Figure 4. Exponential Relative Likelihood Function for Discount Offered

Based on the graph and the use of R, we conclude that the 15% likelihood interval for θ is [10.805, 12.862].

d)

To find the confidence interval, we make use of the asymptotic Gaussian pivotal quantity $\frac{\bar{Y}-\theta}{\frac{\bar{Y}}{\sqrt{n}}}$, where $\frac{\bar{Y}-\theta}{\frac{\bar{Y}}{\sqrt{n}}}$ has a distribution of G(0,1).

$$\begin{aligned}
0.90 &= P(-a \leq Z \leq a) \\
&= P(-a \leq \frac{\bar{Y}-\theta}{\frac{\bar{Y}}{\sqrt{n}}} \leq a) \\
&= P(-a\frac{\bar{Y}}{\sqrt{n}} \leq \bar{Y}-\theta \leq a\frac{\bar{Y}}{\sqrt{n}}) \\
&= P(\bar{Y}-a\frac{\bar{Y}}{\sqrt{n}} \leq \theta \leq \bar{Y}+a\frac{\bar{Y}}{\sqrt{n}})
\end{aligned}$$

Hence, the 100p% confidence interval for θ is $[\bar{Y}-a\frac{\bar{Y}}{\sqrt{n}}, \bar{Y}+a\frac{\bar{Y}}{\sqrt{n}}]$= [10.908, 12.640].

Part 3.

a)

For this section, we assume a Poisson(θ) model for the Customer care calls variate, where θ represents the average customer care calls made on one purchase recorded in January to February of 2020.

b)

The table below provides the numerical summaries for Customer care calls.

Minimum	0	IQR	2.000
1st Quartile	3.000	Range	7.000
Sample Median	4.000	Sample Mean = $\overline{y}$	3.908
3rd Quartile	5.000	Sample Standard Deviation = s	1.261
Maximum	7.000	Sample Skewness	-0.199

Table 4. Numerical Summaries of Customer Care Calls

c)

The table below compares the observed frequency and expected frequency using a Poisson(θ) model.

Number of Customer Care Calls	Observed Frequency	Expected Frequency
0	7.000	10.040
1	10.000	39.238
2	28.000	76.670
3	141.000	99.876
4	163.000	97.579
≥ 5	151.000	176.598
Total	500.000	500.000

Table 5. Observed Frequencies vs Expected Frequencies for Customer Care Calls

d)

From the two tables above, we determine that Poisson distribution is not a suitable distribution for the data. First of all, theoretically, the mean and variance should be identical. However, the observed data demonstrates that sample mean is 3.908, whereas sample variance is $1.261236^2 = 1.591$. The difference between the two values is quite large. Moreover, by

observing the second table, we notice that the expected frequencies are quite different from the observed frequencies. For example, we would expect about 98 of the 500 shipments to receive 4 calls from customers, but observation suggests that 163 of the 500 shipments received 4 calls from customers. The two differed by around 65 shipments, which is quite large. Altogether, we conclude that the Poisson distribution is not a suitable distribution for the data.

Part 4

For each shipping method, we assume that the weight of products can be modeled as a $G(\mu_i, \sigma_i)$ for i=1,2,3 where 1=Ship, 2=Flight, 3=Road.

a)

μ_1 represents the mean weight in grams of Ship shipments recorded in January to February of 2020 and σ_1 represents the standard deviation of weight in grams of Ship shipments recorded in January to February of 2020.

b)

The table below provides the numerical summaries for the weight of products shipped by Ship only.

Minimum	1001.000	IQR	3243.000
1st Quartile	1858.500	Range	4991.000
Sample Median	4258.000	Sample Mean = $\overline{y}$	3718.553
3rd Quartile	5101.500	Sample Standard Deviation = s	1630.805
Maximum	5992.000	Sample Skewness	-0.388
		Sample Kurtosis	1.633

Table 6. Numerical Summaries for the Weight of Shipped Only Shipments

c)

The following figure is a qqplot fo the weight of products shipped by Ship.

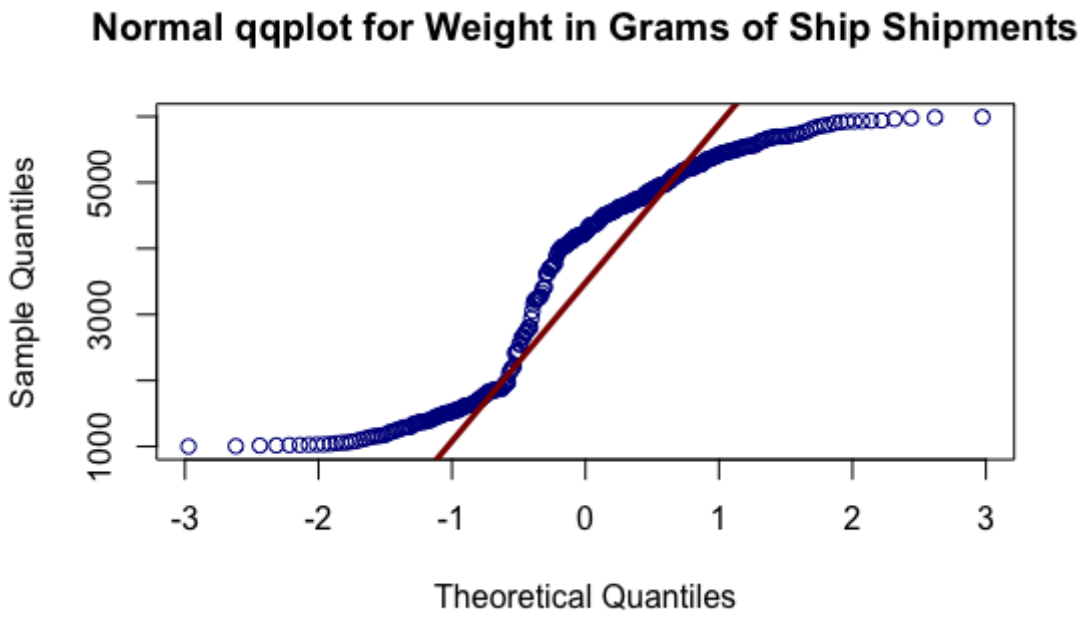

Figure 5. Qqplot of the Weight of Shipped Only Shipments

d)

The Gaussian model would not be a suitable model for this data. Firstly, the sample mean and median of a gaussian model should be identical. However, the observed sample mean and sample median differed by 4258-3718.553 = 539.447, which is quite large. Secondly, the sample skewness of a Gaussian model should be 0 and the model should be symmetrical, though the sample skewness is -0.388, indicating negative skewness. Hence, we notice that there exists a small difference in skewness between the observed and expected skewness even. Thirdly, the sample kurtosis should be 3. For the observed data, the observed kurtosis is 1.633, which is quite different from 3. Fourthly, the qqplot of a Gaussian distribution should be a straight line. The qqplot of the observed data exhibits a s-shape. Based on these characteristics, we conclude that the Gaussian model is not a good fit for this data.

e)

The 95% confidence interval for μ_1 is $\bar{y} \pm bs/\sqrt{n}$ = [3544.587, 3892.519]. The 90% confidence interval for σ_1 is $[\sqrt{\frac{(n-1)s^2}{d}}, \sqrt{\frac{(n-1)s^2}{c}}]$ = [1534.401,1741.315].

f)

The table below provides the numerical summaries for the weight of products shipped by Road only.

Minimum	1017.000	IQR	3206.000
1st Quartile	1806.000	Range	4983.000
Sample Median	4131.000	Sample Mean = $\overline{y}$	3581.447
3rd Quartile	5012.000	Sample Standard Deviation = s	1694.396
Maximum	6000.000	Sample Skewness	-0.171
		Sample Kurtosis	1.469

Table 7. Numerical Summaries for the Weight of Road Only Shipments

g)

The following figure is a qqplot fo the weight of products shipped by Ship

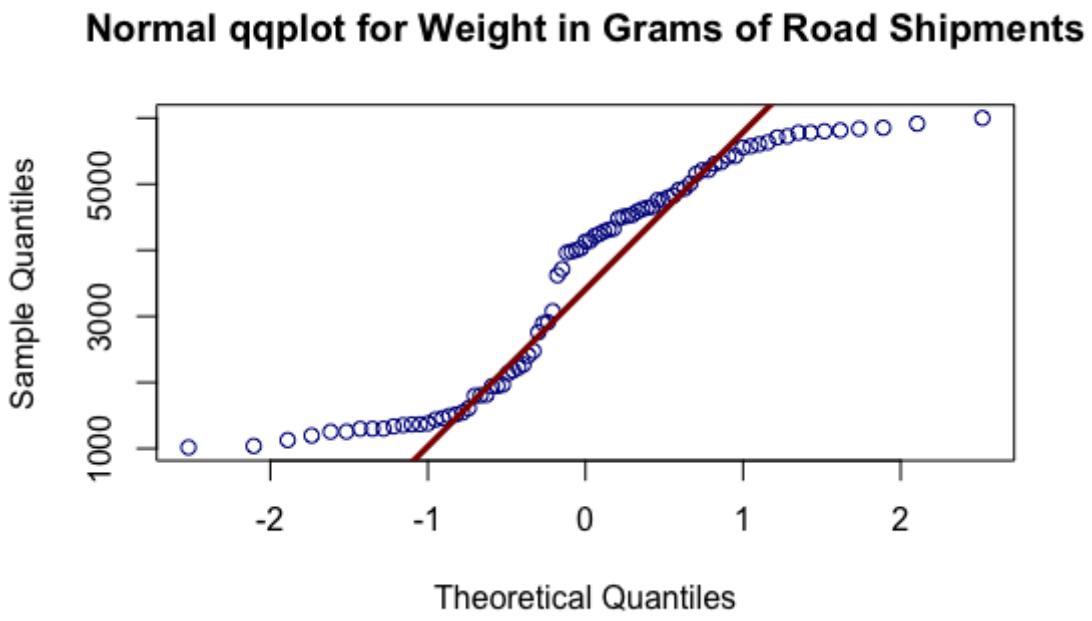

Figure 6. Qqplot of the Weight of Road Only Shipments

h)

The Gaussian model would not be a suitable model for this data. Firstly, the sample mean and median of a gaussian model should be identical. However, the observed sample mean and sample median differed by 4131-3581.447 = 549.553, which is quite large. Secondly, the sample skewness of a Gaussian model should be 0. For my data, the observed skewness is -0.171, which is close to 0 but a difference exists. Thirdly, the sample kurtosis for the Gaussian model should be 3, but the observed sample kurtosis is 1.469, which is quite different from the expecation with a difference of 1.531. Lastly, the qqplot of a Gaussian distribution should be a straight

line. The qqplot of the observed data exhibits a s-shape. Based on these characteristics, we conclude that the Gaussian model is not a good fit for this data.

i)

The 95% confidence interval for μ_3 is $\bar{y} \pm bs/\sqrt{n}$ = [3215.974, 3946.920]. The 90% confidence interval for σ_3 is $[\sqrt{\frac{(n-1)s^2}{d}}, \sqrt{\frac{(n-1)s^2}{c}}]$ = [1505.547,1943.054].

j)

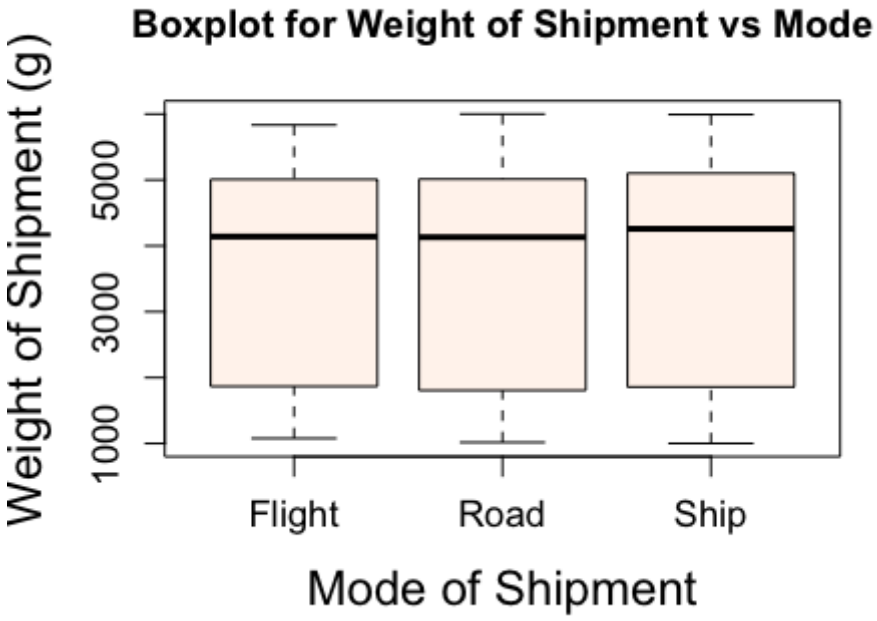

Figure 7. Side-by-Side Boxplot of Weight in Grams vs Mode of Shipments

The above figure displays the side-by-side boxplots of shipment's weight categrozied by the mode of shipment. The shape of the three boxplots are similar. All three boxplots has its sample median in the upper region of the box, indicating that the distributions of all three data sets are negatively skewed. Moreover, the three boxplot all have its upper whisker and lower whisker close in length. The difference is that road shipments has a smaller range than the other two shipment modes, indicating that flight shipments' weight varies less than shipments by other modes. Additionally, ship shipments has a median higher than flight and road shipments. This means that ship shipments are more negatively skewed and ship shipments' tend to have higher weight. Finally, there are no outliers for all three data sets.

Part 5

a)

For this section, we assume a Binomial(n, θ) model for the distribution of Reached on Time variate where θ represents the probability of a GoodBuy shipment recorded in January to February of 2020 arriving on time.

b)

Since this is a Binomial distribution, then the maximum likelihood estimate of $\hat{\theta}$ is given by $y/n = 0.594$.

c)

The approximate 95% confidence interval for the proportion of the population that had deliveries on time is given by $\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$, which is [0.551, 0.637].

d)

Assume that fifty products are to be shipped on a certain day. The maximum likelihood estimate of the proabbility htat at least 23 are delivered on time. In other words, this probability is defined as $1 - P(X \leq 22) = 0.980$

e)

The proportion of high importance products that arrived by the expected delivery date is 0.667.

f)

The following table provides the relative frequency of each Product importance level by whether the item was delivered on time or not.

		Product Importance		
		Low	Medium	High
Reached on Time	Y=1	0.20	0.18	0.03
	N=0	0.28	0.25	0.06

Table 8. Relative Frequencies of Product Importance by Reached on Time

Notice that the total proportion of products reached on time is 0.20+0.18+0.03=0.41. The total proportion of products that did not reach on time is 0.28+0.25+0.06 = 0.59.

g)

The following figure is a side by side bar graph for the proportion of products delivered on time by product importance.

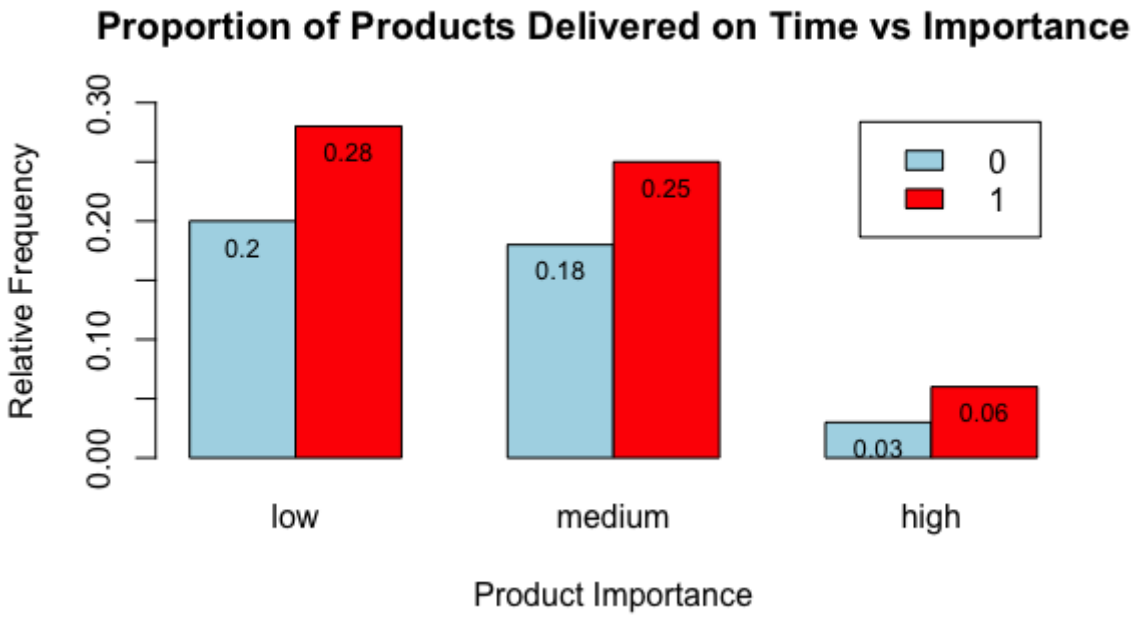

Figure 8. Side-by-Side Bar Graph of Proportion of Products Delievred on Time vs Importance

From the figure above, it can be observed that for all categories of importance, the proportion of products that arrived on time is slightly larger than the proportion of products that did not arrive on time. In particular, there is a higher proportion of high importance products arriving on time (2 on time : 1 not on time, in other words 2/3 arrived on time) than medium and low importance products.

5 Conclusion

Based on the analysis, several characteristics of the variates of interest can be detected. Through the "Bar Char for the Customer Rating of Shipment", we notice that the modal customer rating is 1. A closer examination of the "Customer Rating by Female vs Male" graph reveals that the modal rating of male customers is 1 and the modal rating of female customers is 4. From this figure, it can be observed that the number of males and females in the target population giving out a particular rating are close in value. Although, male ratings are more concentrated on the two ends 1 and 5, whereas female ratings are more randomized. Moreover, it has also been discovered that among the 500 samples, 23% of the purchases arrived on time and received a rating of at least 4 25% of the purchases arrived after the expected delivery date and have a customer rating of no more than 3. We notice that the proportion of purchases that arrived on time and received a high rating is not high, thus we conclude that the arrival time does not greatly impact a customer's rating. One thing to note in the analysis of customer rating is the response bias and missing data when customers rate their purchase. It is possible that the ratings collected do not reflect the true satisfaction of the customers.

Through analysis, it has also been obtained that the mean of the Discount offered variate is 11.774 and the median is 7. This reflects that the mean is heavily affected by the large outliers. Although the average discount is 11.774, most customers are receiving much lower discounts. Additionally, since the sample mean is greater than the sample median and the sample skewness is positive, we conclude that the Discount offered variate is best represented by an exponential model.

For the Customer care calls variate, we found that the mean number of customer care calls to the target population is 3.908. Through further numerical analysis, it can be observed that the Poisson model is not a reasonable model for modeling the distribution of the number of customer care calls in the target population. Not only does the sample mean and sample variance differ in value, the expected frequency and observed frequency of customer calls also vary. Therefore, we would not use a Poisson model to represent the distribution of the customer care calls variate.

Similarly, the Gaussian model is not reasonable for modeling the weight of the products in the target population. This has been confirmed by comparing the observed sample skewness

and kurtosis to the expectance as well as comparing the similarity between a sample mean and sample median. The qqplot, once again, confirms that the Gaussian model is not suitable for the weight in grams variate. In addition, we notice that the mean and standard deviation for weight among items shipped by road in the target population are 3581.447 and 1694.396 respectively. Likewise, the mean and standard deviation for weight among items shipped by ship in the target population are 3718.553 and 1630.805 respectively. In fact, through the "Boxplot for Weight of Shipment vs Mode" graph, we notice that all three modes of shipments have a similar distribution of product weight. All three modes have a similar sample median and are all negatively skewed in weight distribution, though the range of flight shipments' weight is slightly smaller than the other two modes.

Lastly, we will examine the Reached on Time variate. We recognize that approximately 0.41 of the target population was delivered by the expected delivery time. Moreover, if we choose 50 random purchases from the target population, the probability that at least 23 are delivered by the expected delivery time is 0.980. It has also been calculated that 0.667 of the high-importance products were delivered by the expected delivery time. Finally, through graphical analysis, we can observe that the proportion of products arriving on time is similar across all product importance in which more products reached on time. Though, we have also noticed that a higher proportion of high-quality products reached on time in comparison to low and medium quality products. It should be noted that the arrival time may not be the time that the customer receives the purchase. Thus, the Reached on Time variate may not be accurate.

There are several limitations to these conclusions other than the measurement errors discussed above. First of all, the study population includes all purchases recorded from January to February of 2020, which is a very short period around Winter. Since the target population is all purchases made by customers at GoodBuy in 2020, the data over two months is insufficient to analyze purchases throughout the year. Moreover, January and February are in Winter, thus the type of items customers purchase are different from the items purchased in Summer. Additionally, the weather conditions can also affect shipping and even mood when giving out customer ratings. Hence, more data are required to better analyze the variates of interest throughout the year.