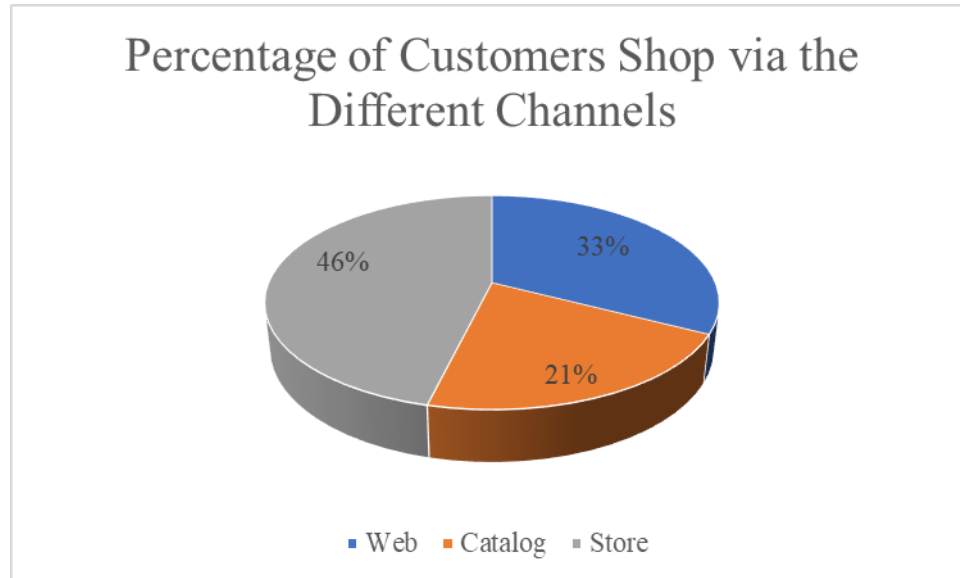


Name: Xuhui Ying WFU ID: 06648543

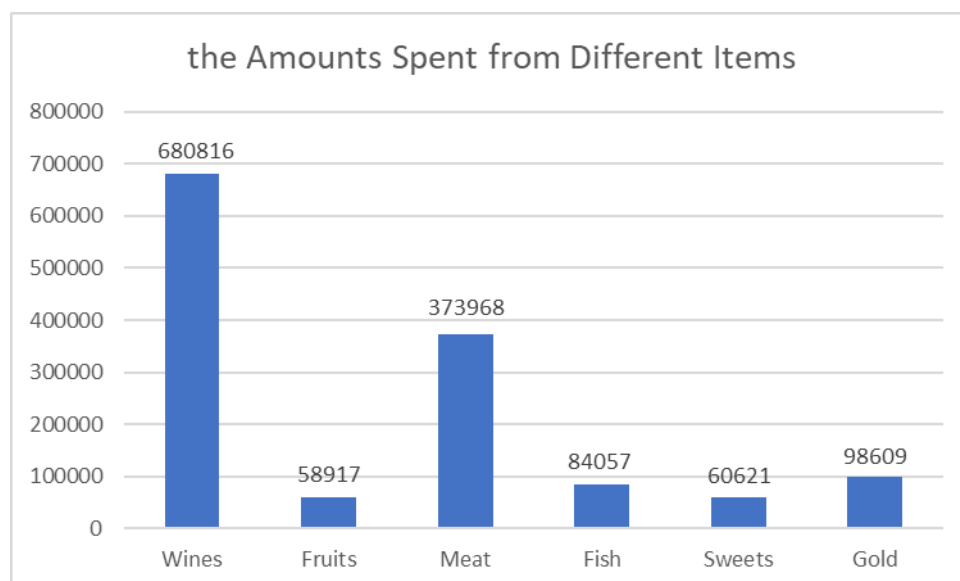
1. An overview of the customers that should address issues such as (but not limited to) the following:

(1) What percentage of customers shop via the different channels?



As is seen in the dataset, there are 2240 pieces of customer data in total; each customer uses one or more channels to shop. There are three different channels for customers to shop - the company's website (Web), catalog (Catalog), and physical stores (Store). Among all these three channels, the company's website is the most popular source of customers' shops, occupying 46% of total customers' shops, followed by catalogs, which occupy 33%. By contrast, only 21% of customers shop from the catalog, which has the most negligible proportion among all these channels.

(2) What are the most popular items (in terms of amount spent) that are purchased from the retailer?

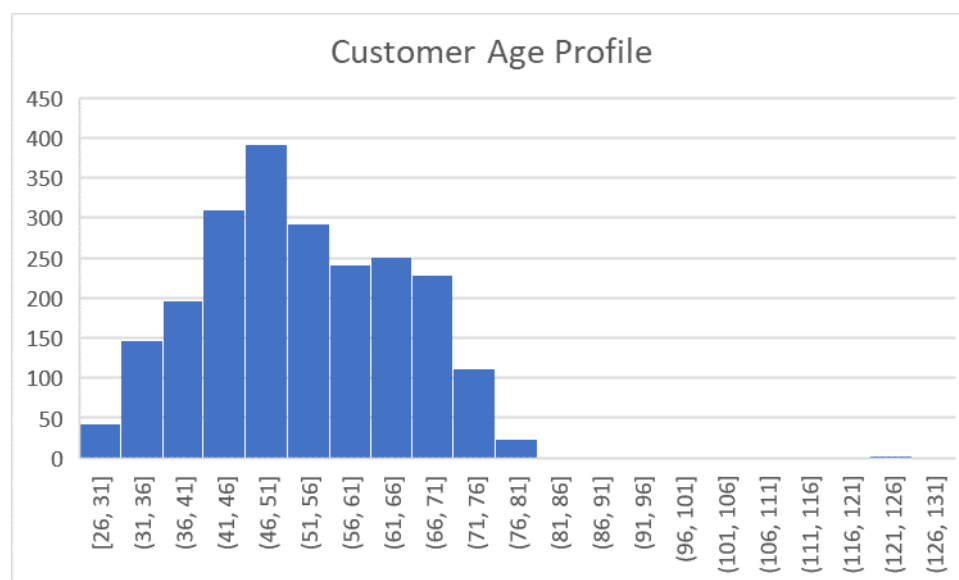
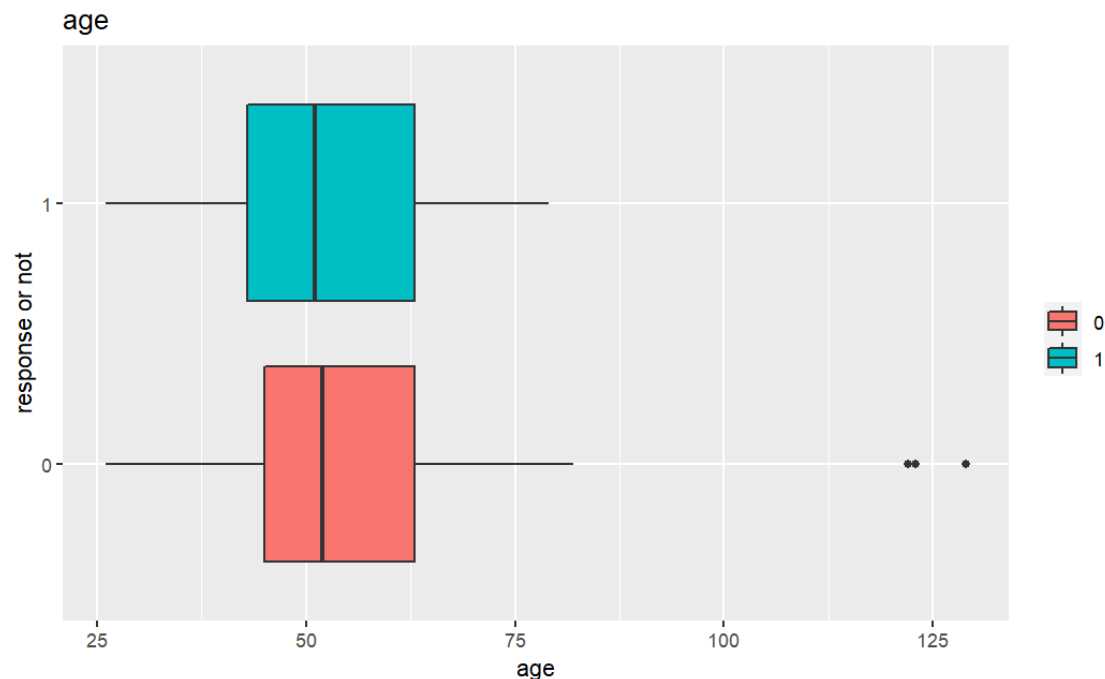


In the given dataset, six columns represent the amount spent on a particular kind of product in

the last two years, namely, 'wines', 'fruits', 'meat', 'fish', 'sweets', and 'gold'. Among them, the most popular item (in terms of amount spent) purchased from the retailer is wine. As recorded by the dataset, these 2240 customers spend \$680,816 on wines in total, which is significantly higher than all other five items in amounts of dollars.

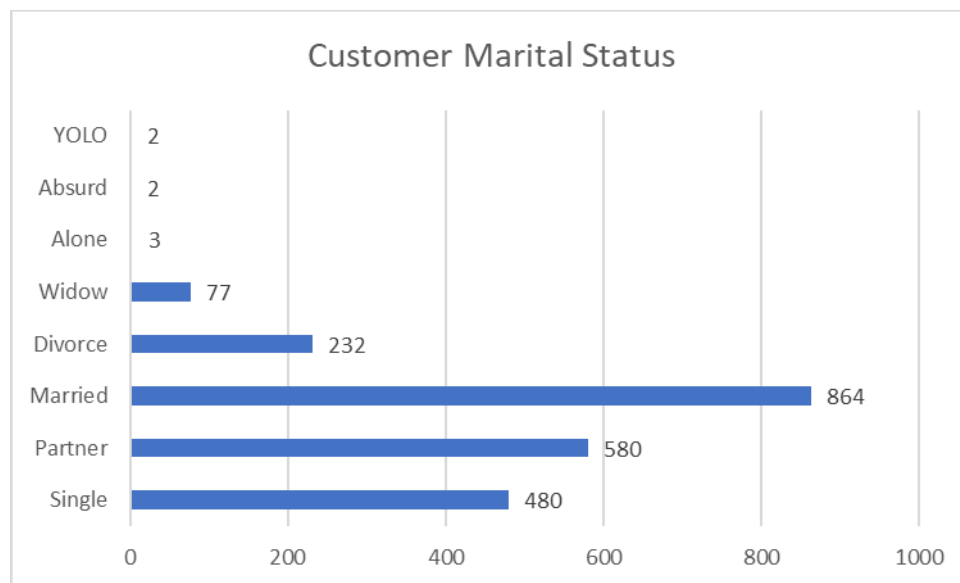
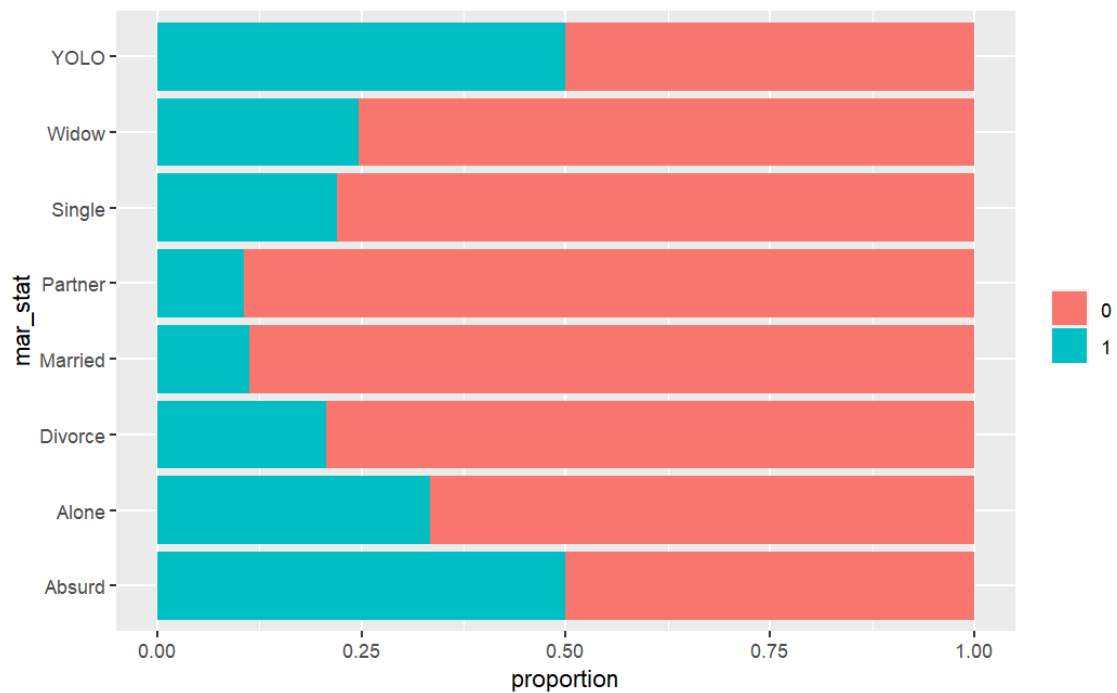
(3) What is the overall demographic makeup of our customer base (age? married? Single? Kids at home? Educated?)?

According to bar charts and boxplots drawn in R Notebook and bar/column charts drawn in Excel, our customer base shows diverse demographic makeup from perspectives of age, marital status, whether kids are at home, and education level.

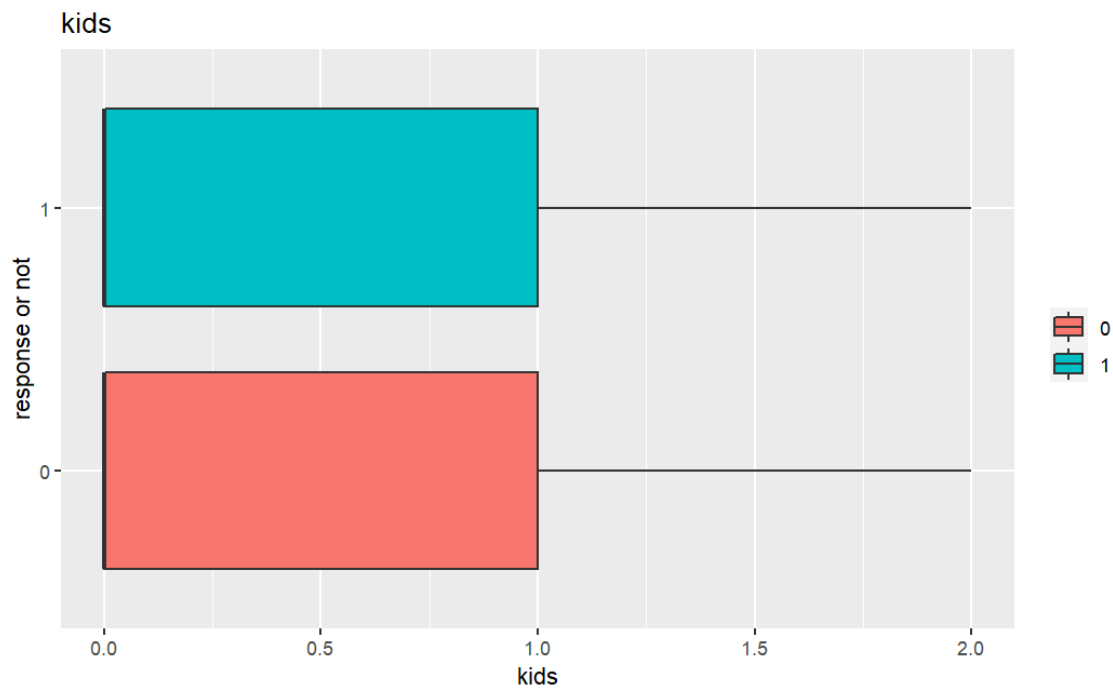


As for the customer age profile, we use a histogram (bin = 5) to demonstrate the distribution of the age of customers (2022 - birthyear). The distribution curve is a little right-skewed, with an outlier. Surprisingly, we can find that there is a customer who is more than 120 years old. Customers who accepted the offer in the last campaign are slightly younger than those who

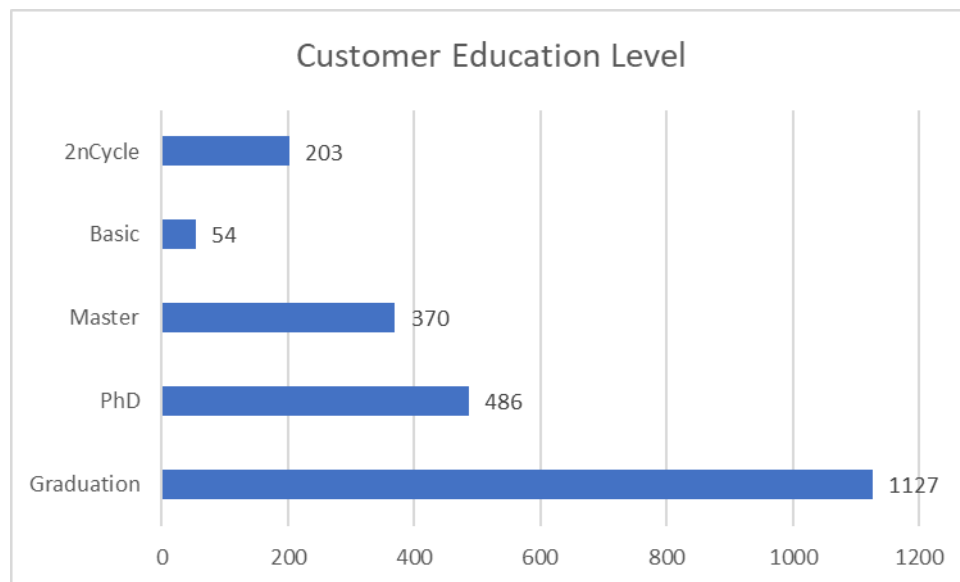
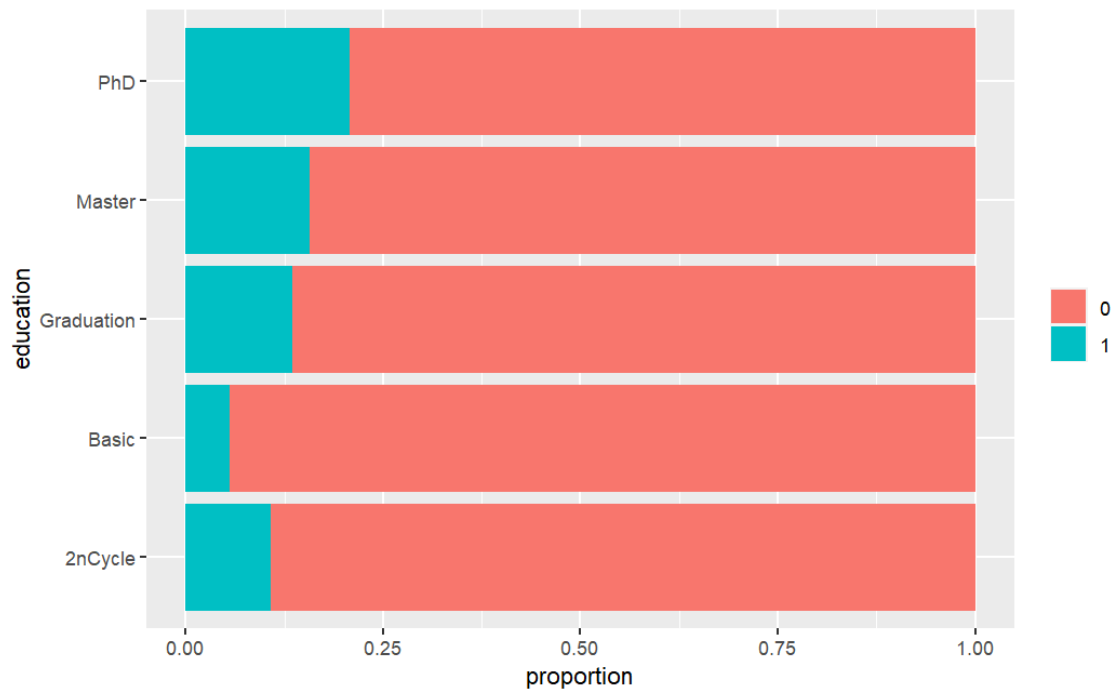
did not accept the offer when comparing medians.



As is shown by the above bar charts, marital customer status is divided into eight categories - YOLO (You only live once), absurd, alone, widow, divorced, married, partner, and single. Among all these 2240 customers, married customers have the highest number (864), while customers in YOLO, absurd, alone categories, only have negligible numbers. As to whether customers accepted the offer in the last campaign, partners and married customers were least likely to accept the offer, which only has around 10% for each category. By comparison, YOLO and absurd customers were most likely to accept offers in the last campaign, taking up about half of the numbers in total among these two groups of customers.



The above boxplot and bar chart compare the numbers of customers with/without kids and the likelihood they accepted the offer in the last campaign. Among these 2240 customers, 899 have kids, and 1293 have no kids. However, when we look at the boxplot, we can see no difference in the probability of customer acceptance of offers. Therefore, variable kids should not be the predictor of my classification model.



As is shown by the above two bar charts drawn by R Notebook and Excel separately, customers' education level can be divided into five different categories - second cycle (graduate level or master's level studies), basic (primary education and lower secondary education), master, Ph.D., graduation. Nearly half of the 2240 customers have a doctoral degree (graduation level). As to whether customers accepted the offer in the last campaign, the higher education a customer received, the more likely they would accept the offer. Around 20% of Customers with a Ph.D. degree accepted the offer in the last campaign, which ranks highest among these five categories, followed by customers with a master's degree (15% or so).

2. An overview of the customer segments created via k-means clustering which should include:

(1) A discussion of the unique characteristics of customers in that segment

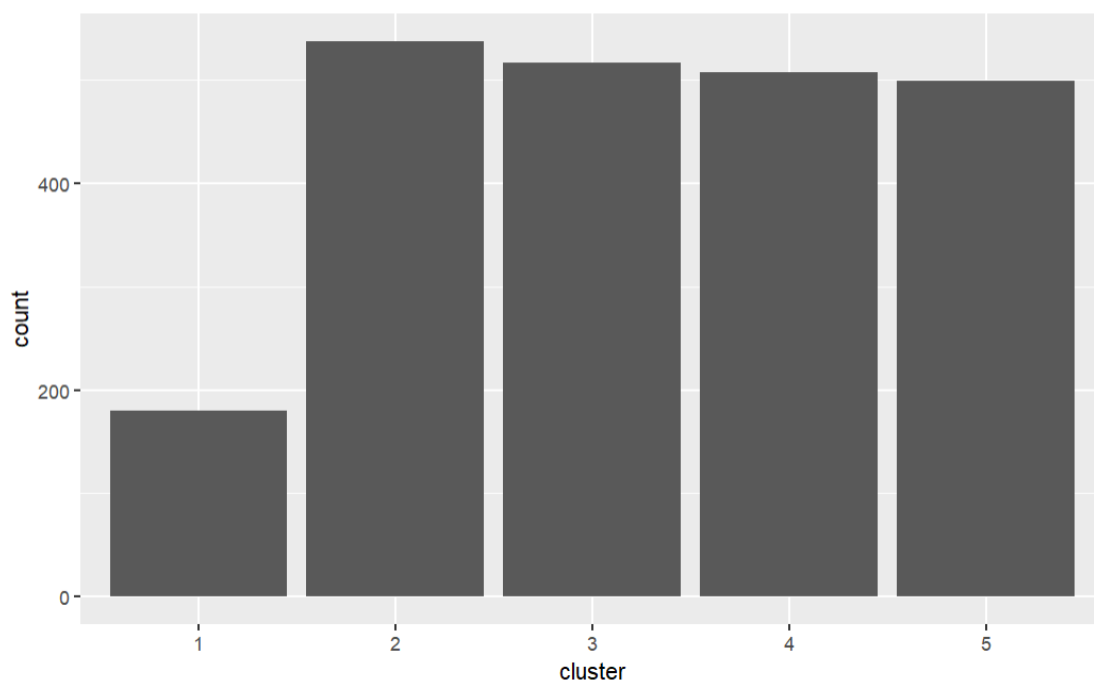
Group 1: This group has the lowest number of customers (less than 200 people) among the five clusters, containing the lowest number of customers without kids.

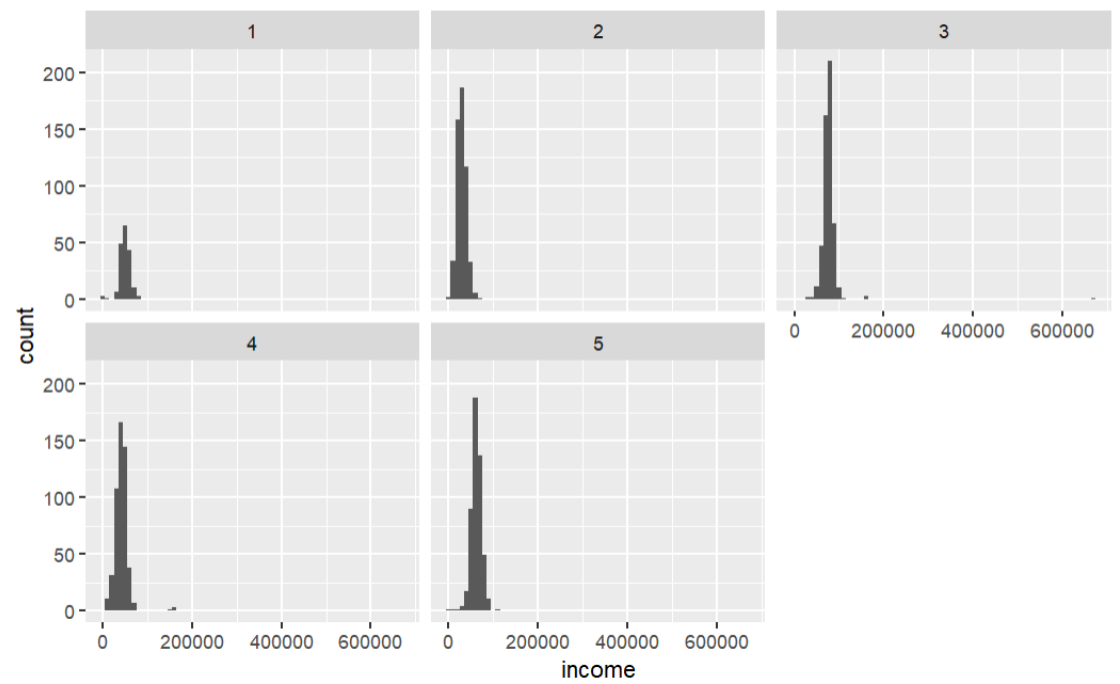
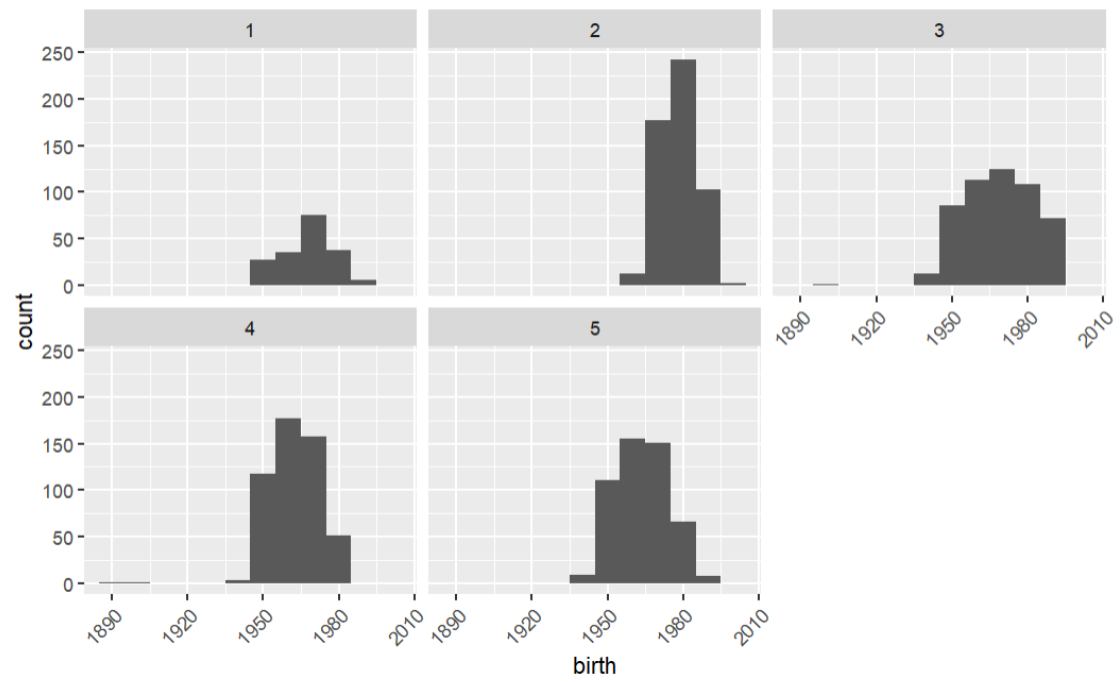
Group 2: The average age of customers in this customer group is the youngest among all five clusters. This cluster contains the most significant number of married customers and the greatest number of teenagers in the customer's household. In addition, this group has the greatest number of customers who only receive basic education, and their incomes are the lowest among all five groups of customers. Besides, this group of customers has spent the least money on wines and meat in the last two years, and they are least likely to buy goods via company websites and catalogs.

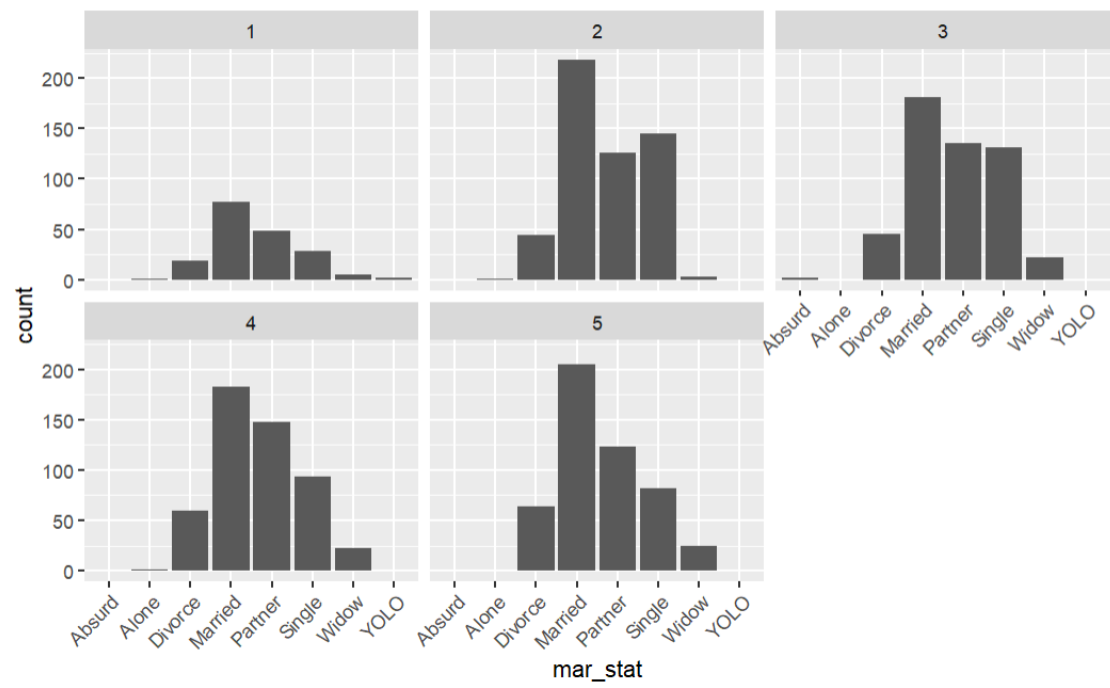
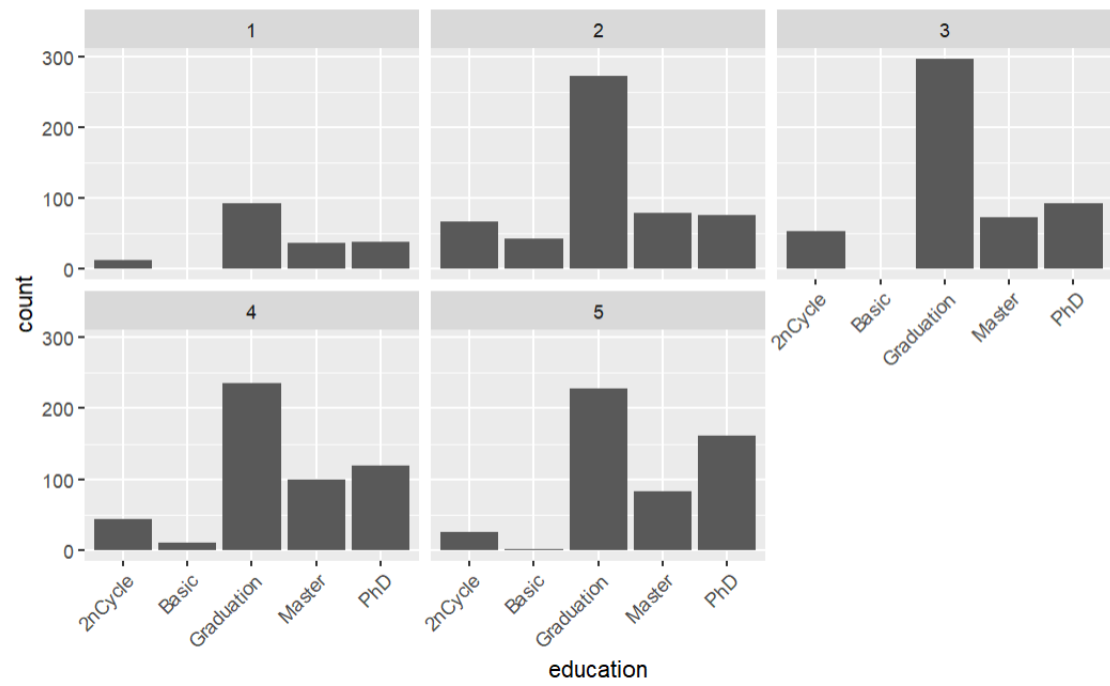
Group 3: The group contains the most significant number of customers without kids and the least number of purchases made with a discount. Besides, they have had minor visits to the company's website in the last month.

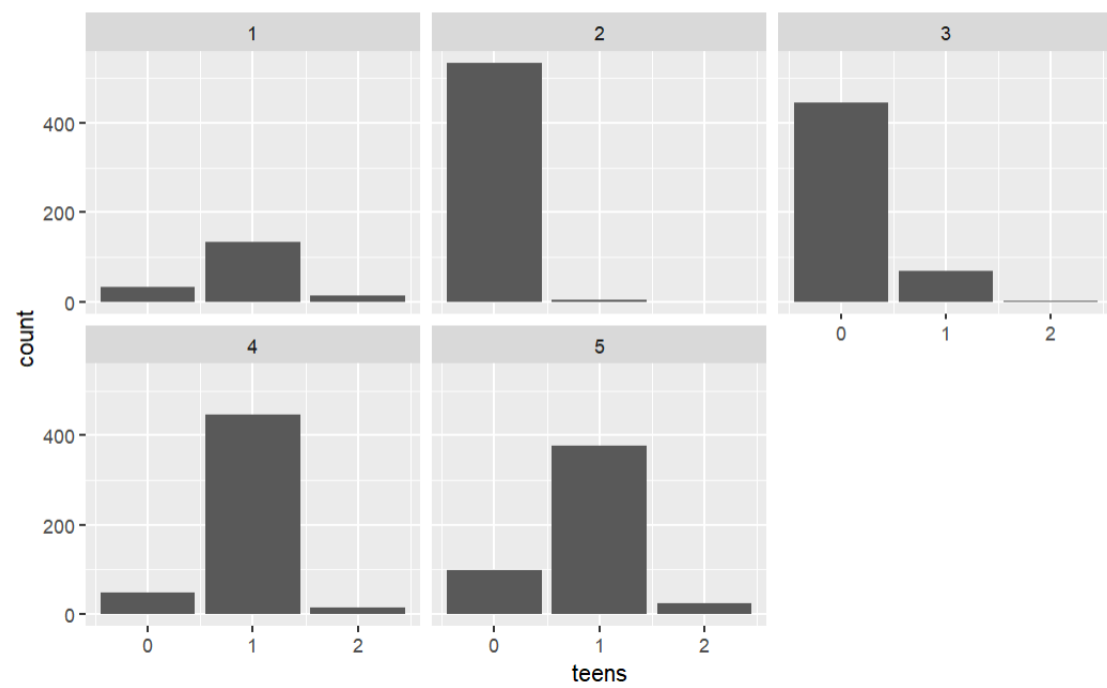
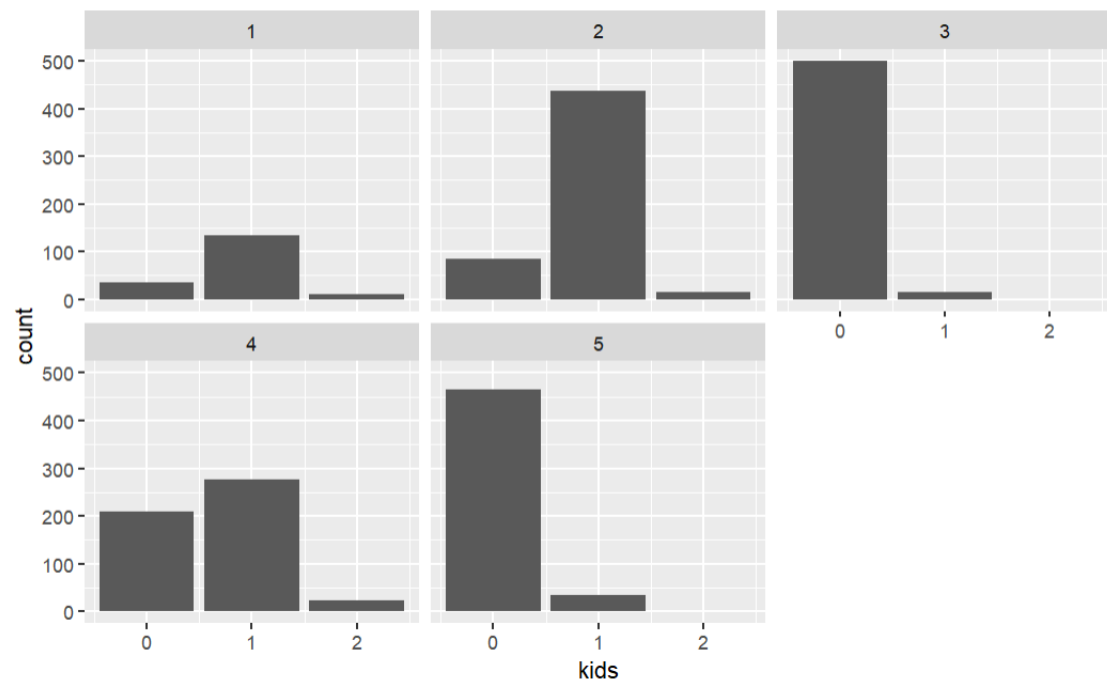
Group 4: This group of customers has spent the least money purchasing fruits, fish, sweets, and gold in the last two years.

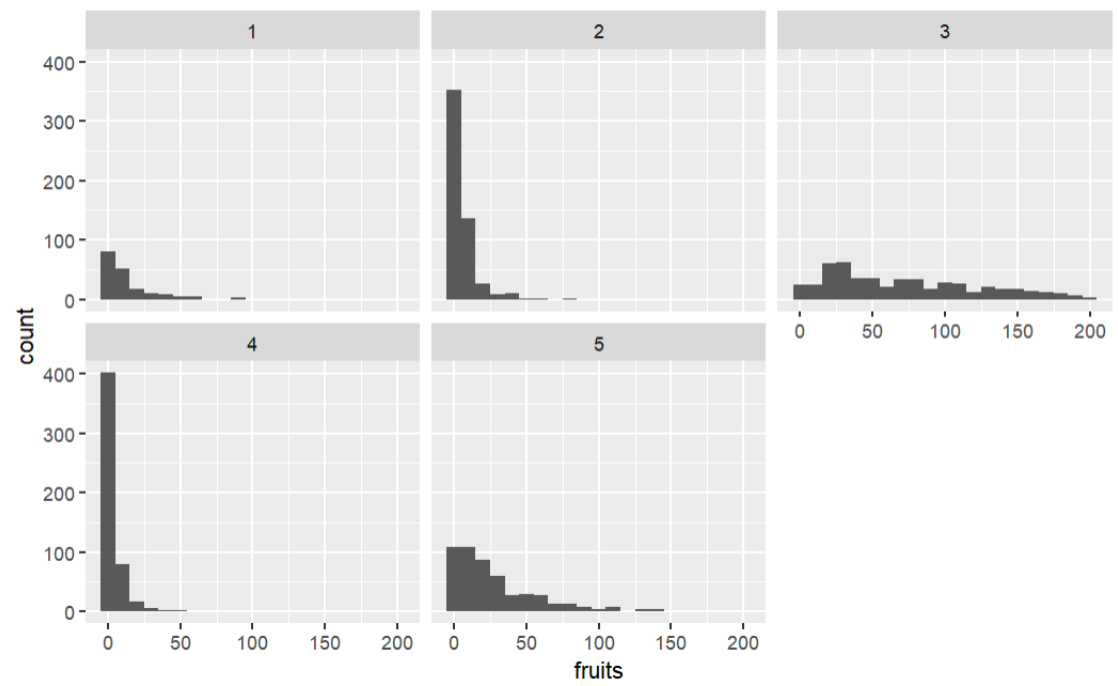
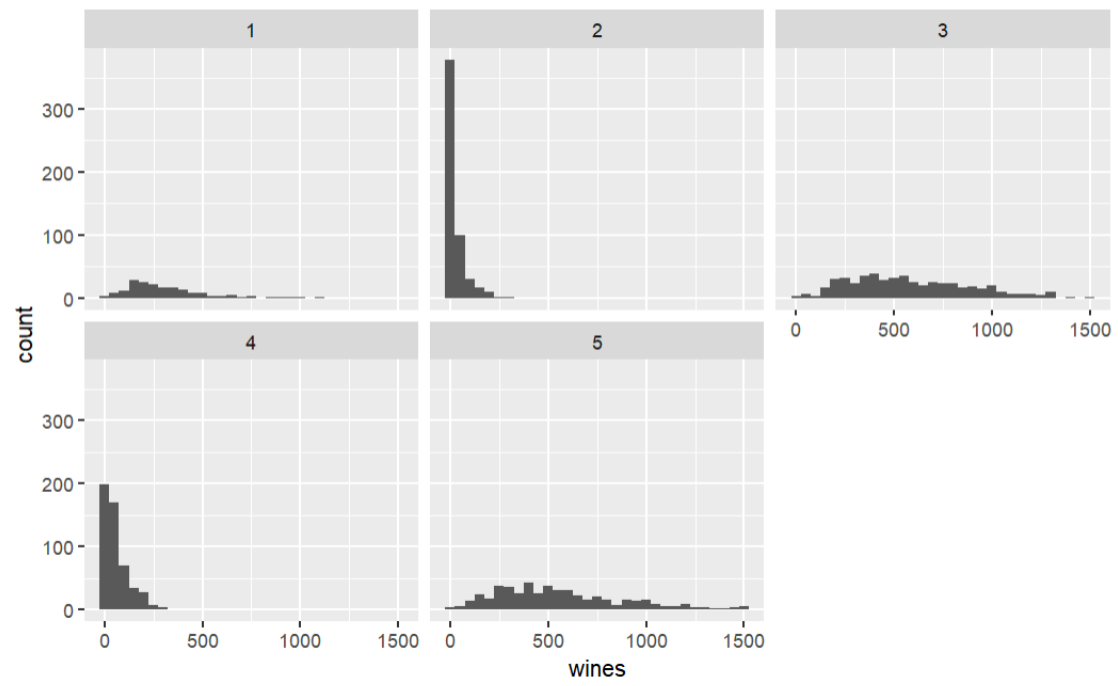
Group 5: This cluster of customers has the highest number of PhDs.

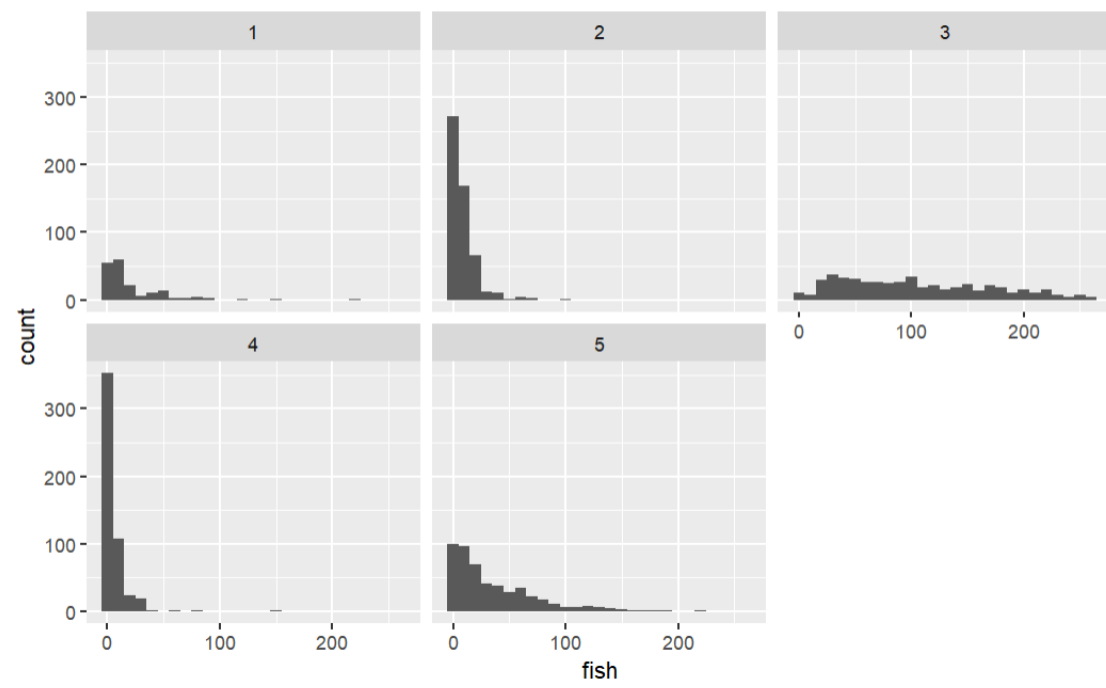
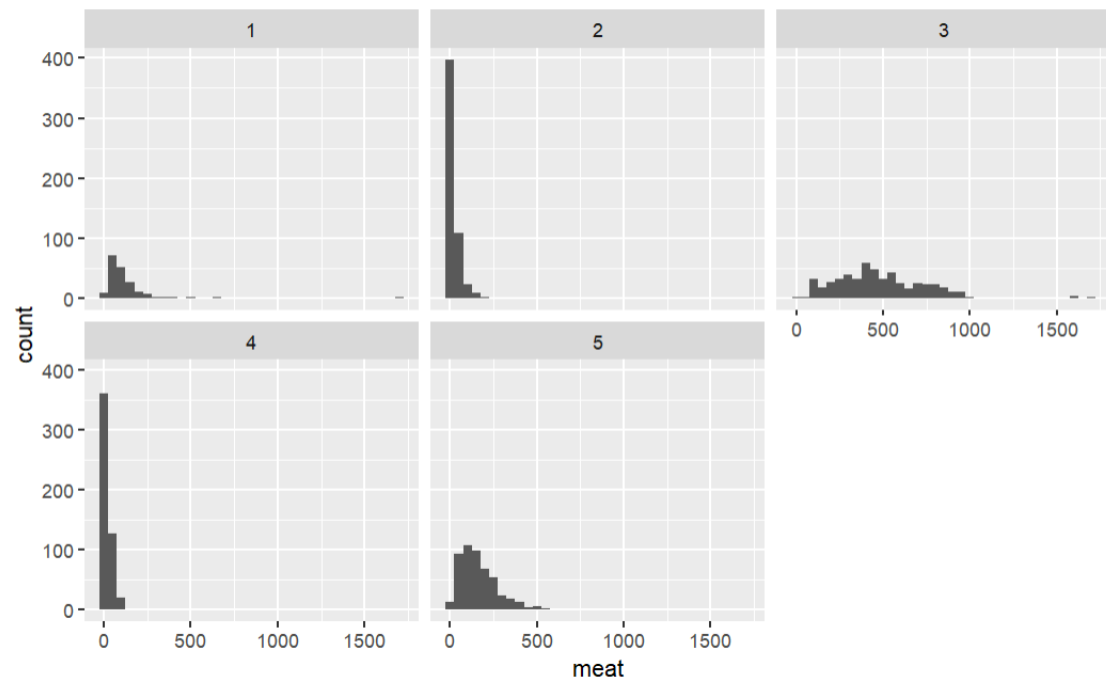


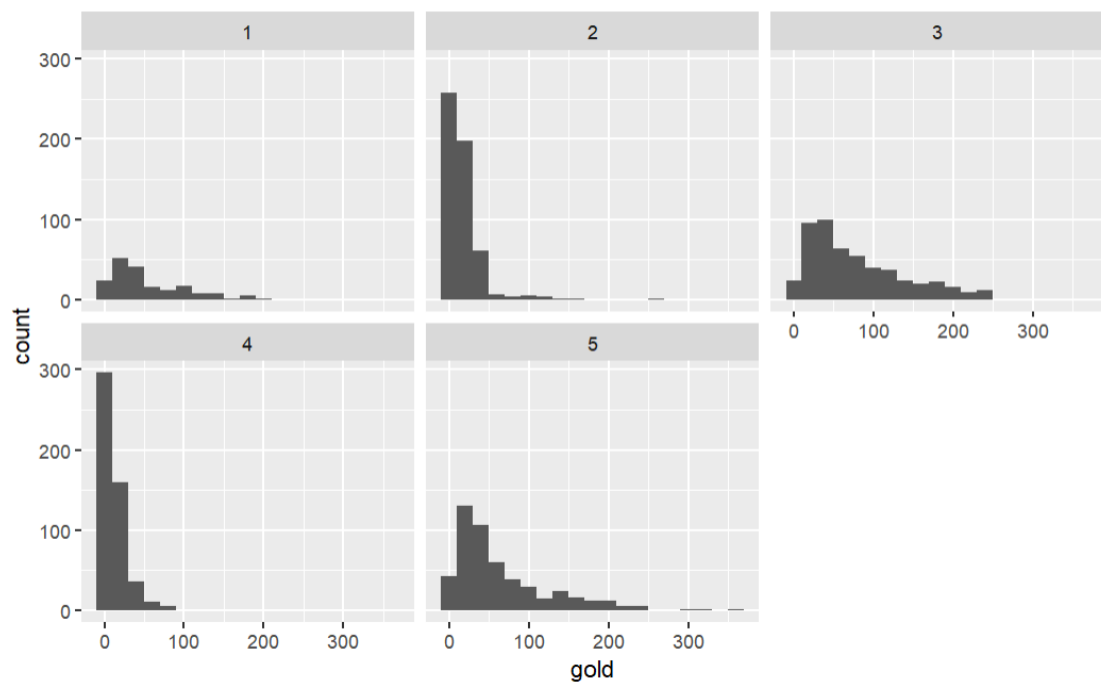
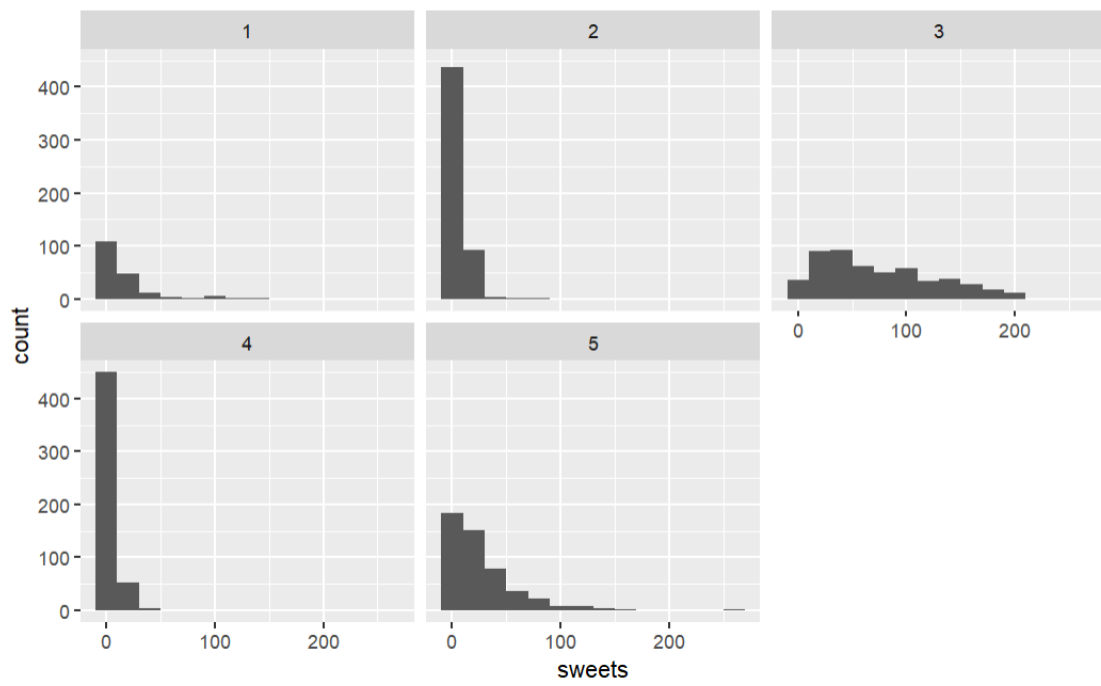


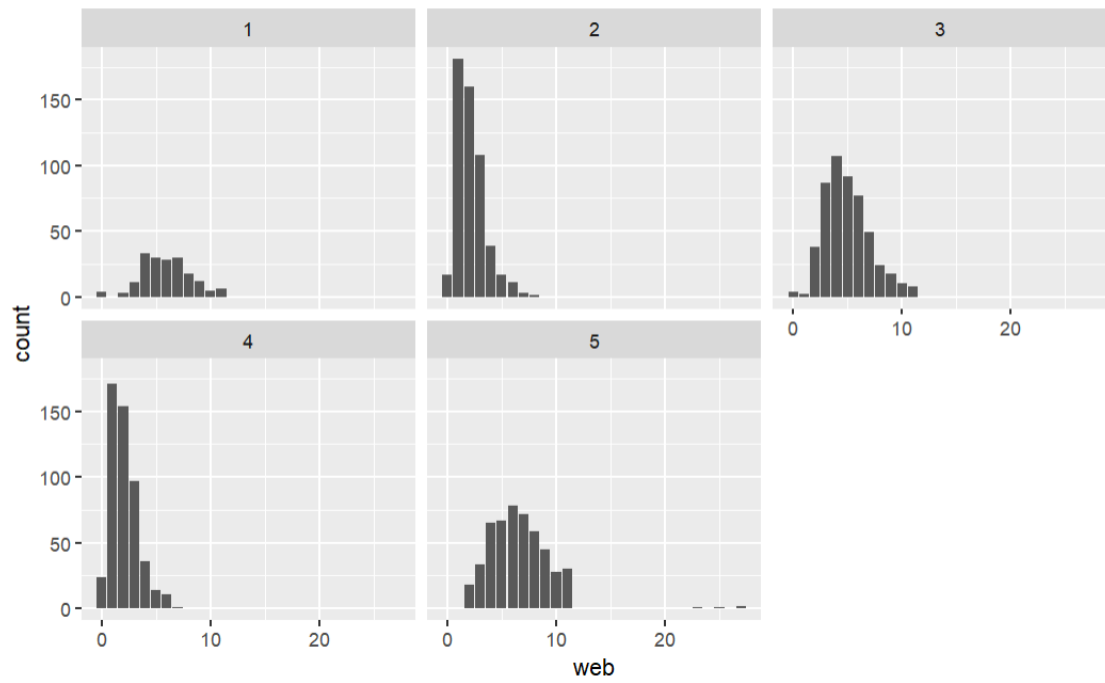
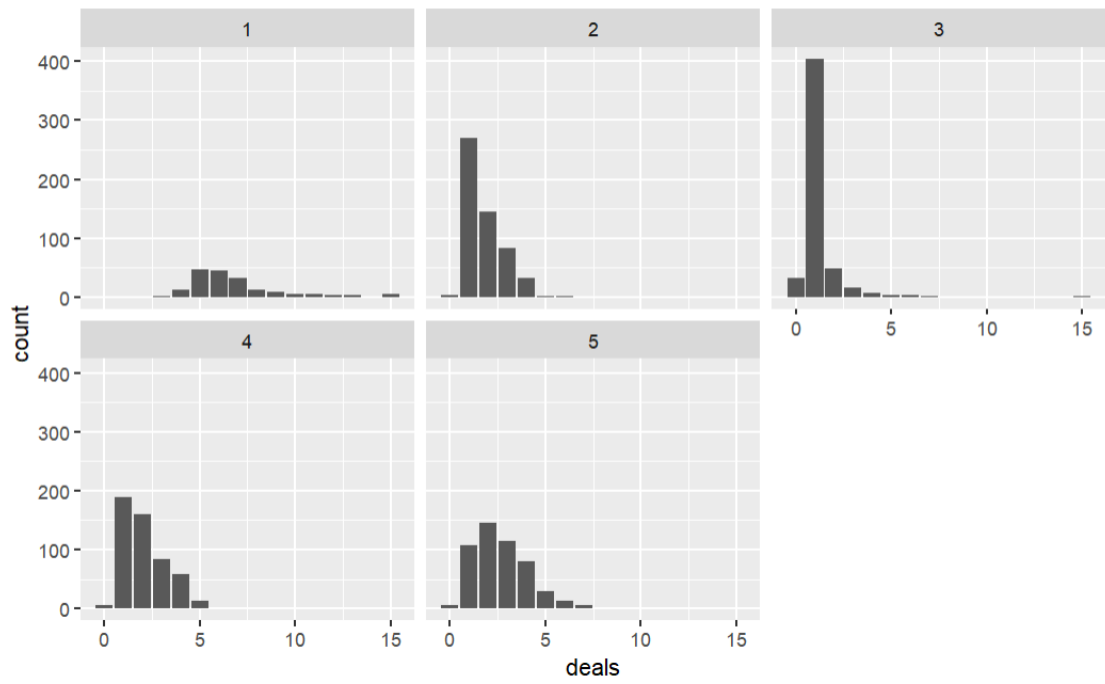


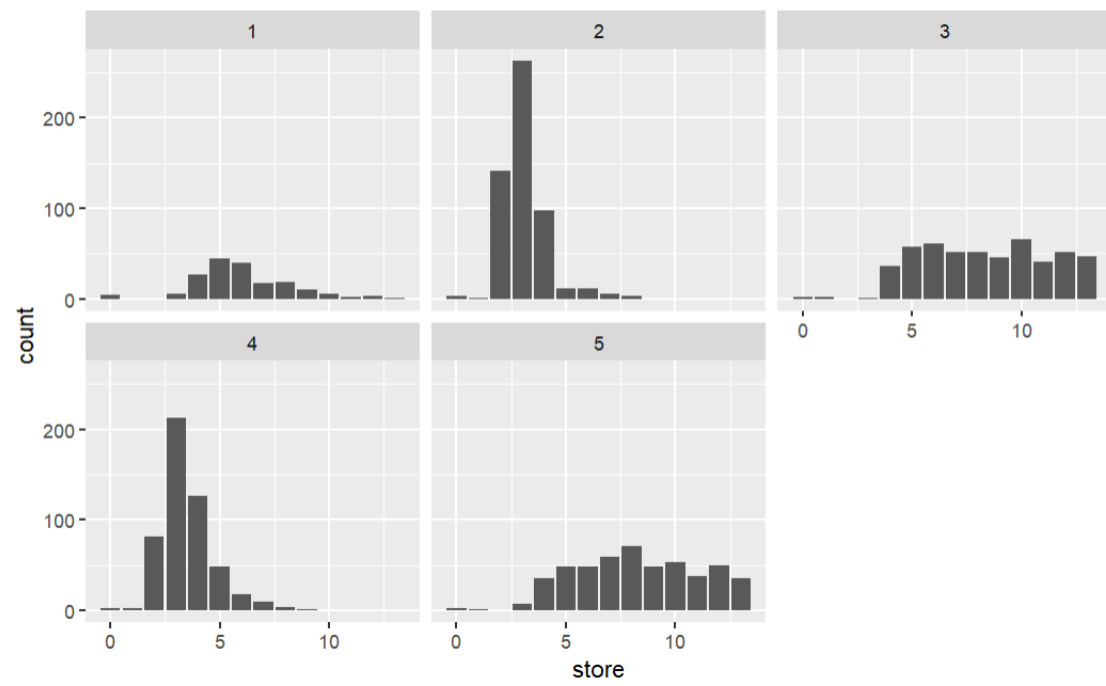
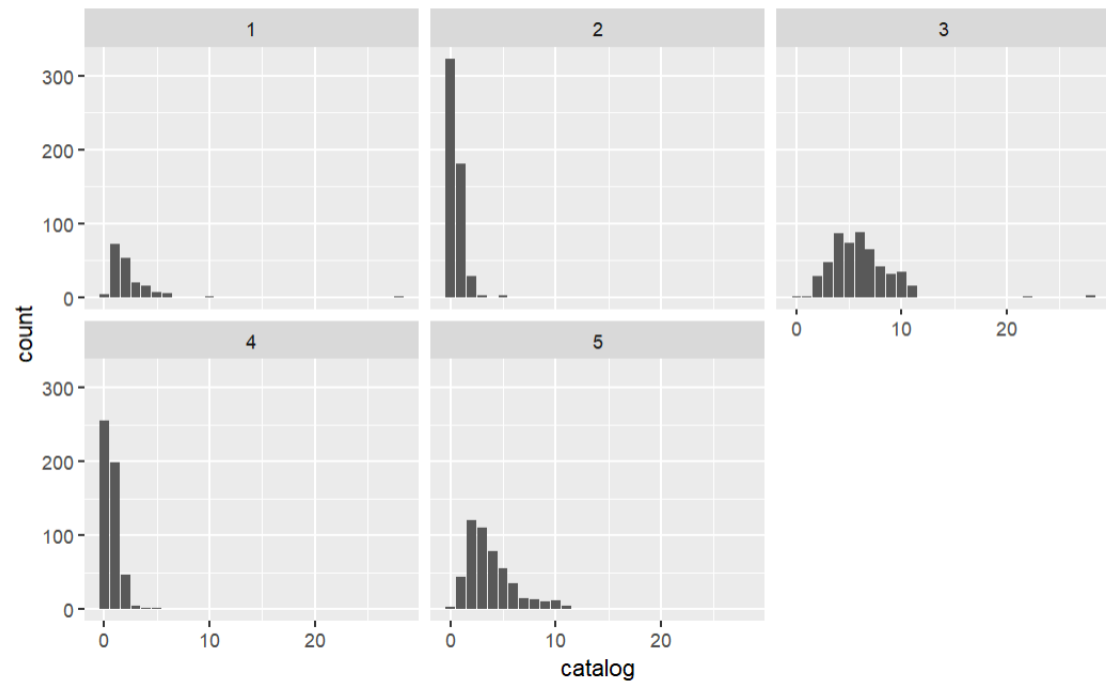


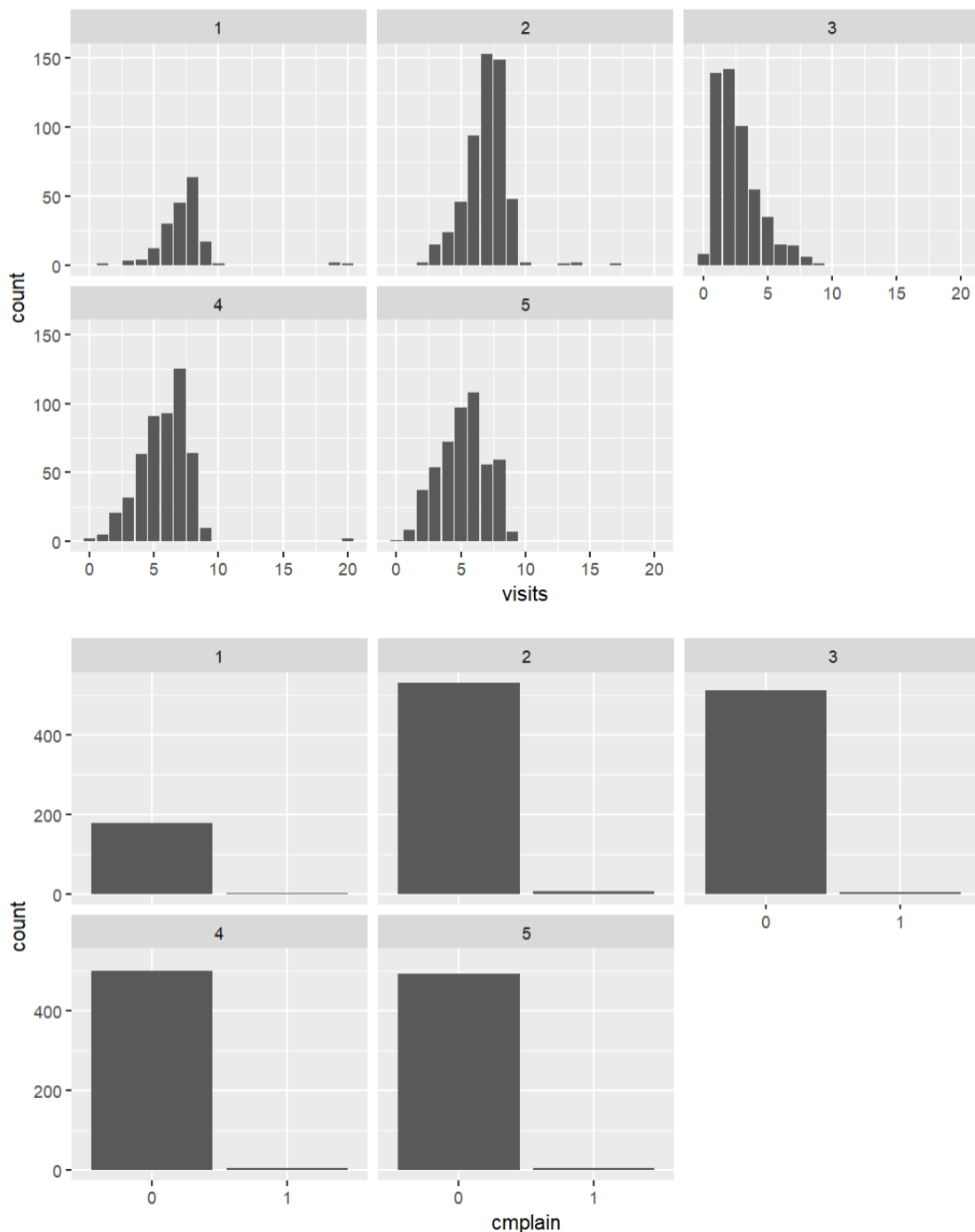












(2) Suggestions of how to develop a marketing message targeted at each group

Group 1: Since this cluster contains the highest number of customers, the store can sell more children's products to these customers, like milk powder, diapers, and all kinds of kid clothes. Besides, they can sell more playful toys and beautiful products that appeal to kids, which can significantly increase sales to this group of customers.

Group 2: The store can use discounts and promotions to sell inexpensive goods to simulate impulse buying because the customers are younger, have a lower education level, and have lower incomes. Also, the store can sell more products except for wines and meat, and they would better sell products in brick-and-mortar stores.

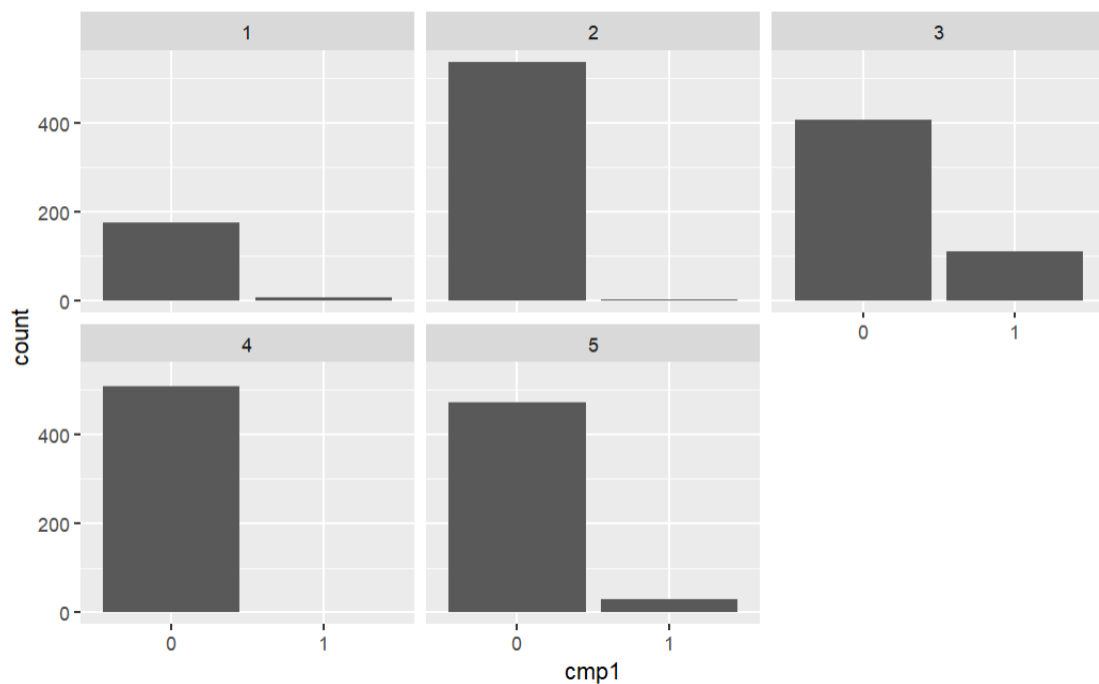
Group 3: As this cluster contains the highest number of customers without kids, the store does

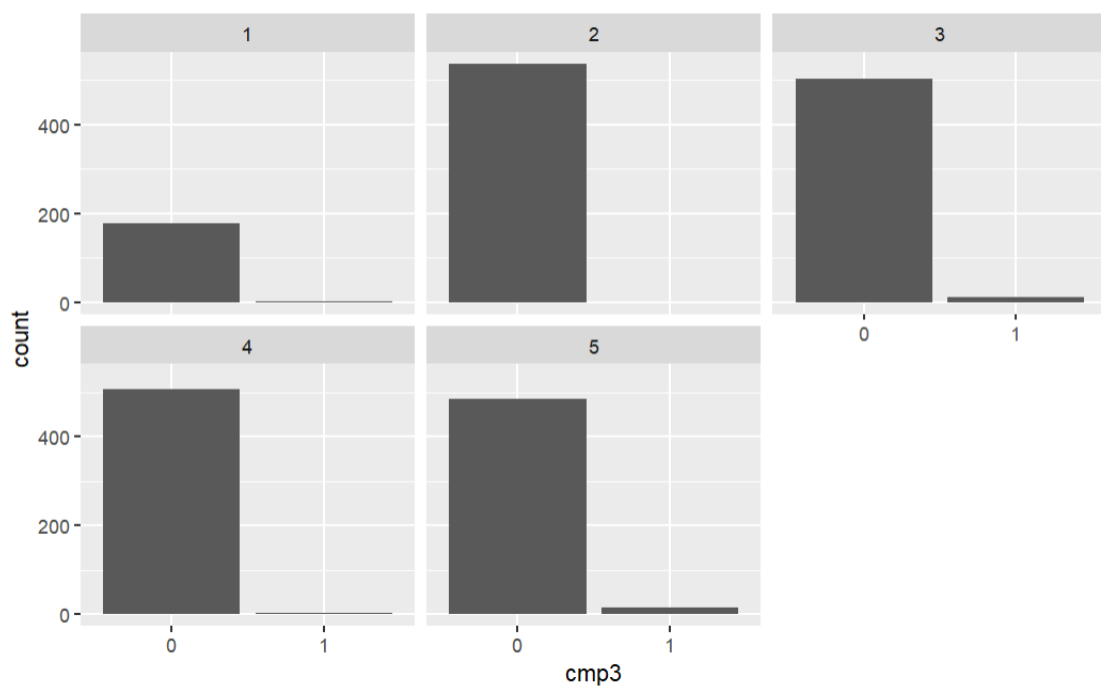
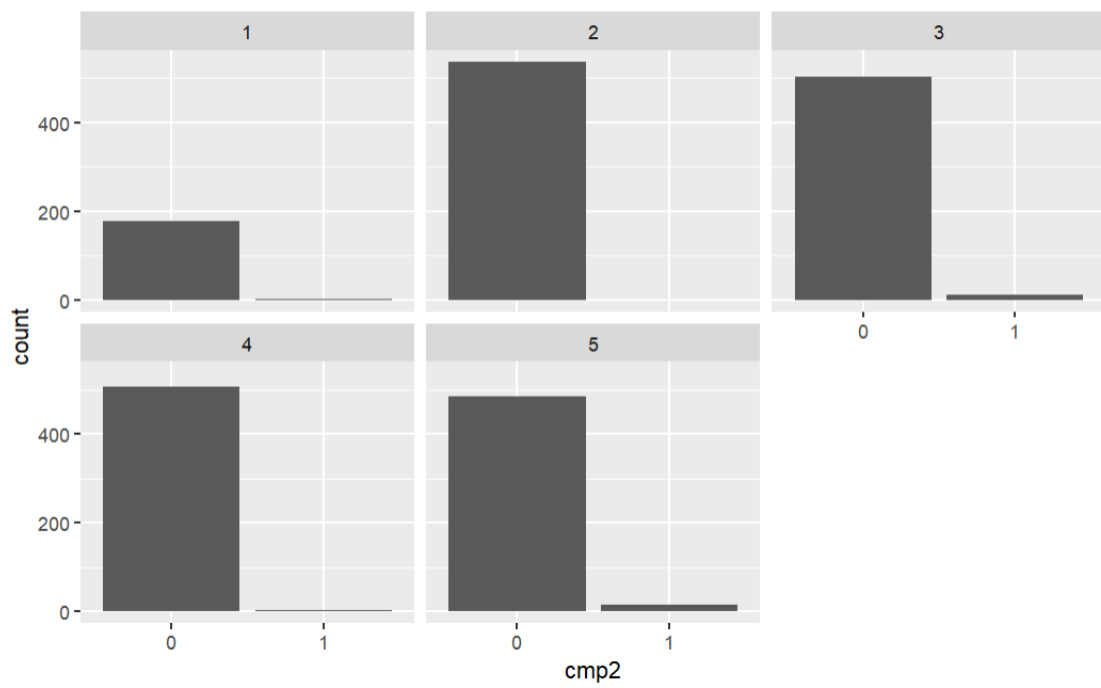
not need to sell children's products to this group of customers. For example, they don't need to spend much time and effort selling children's toys, milk powder, diapers, and so forth. Additionally, the store does not need to promote its products on its website for these customers because of the low website traffic created by this cluster.

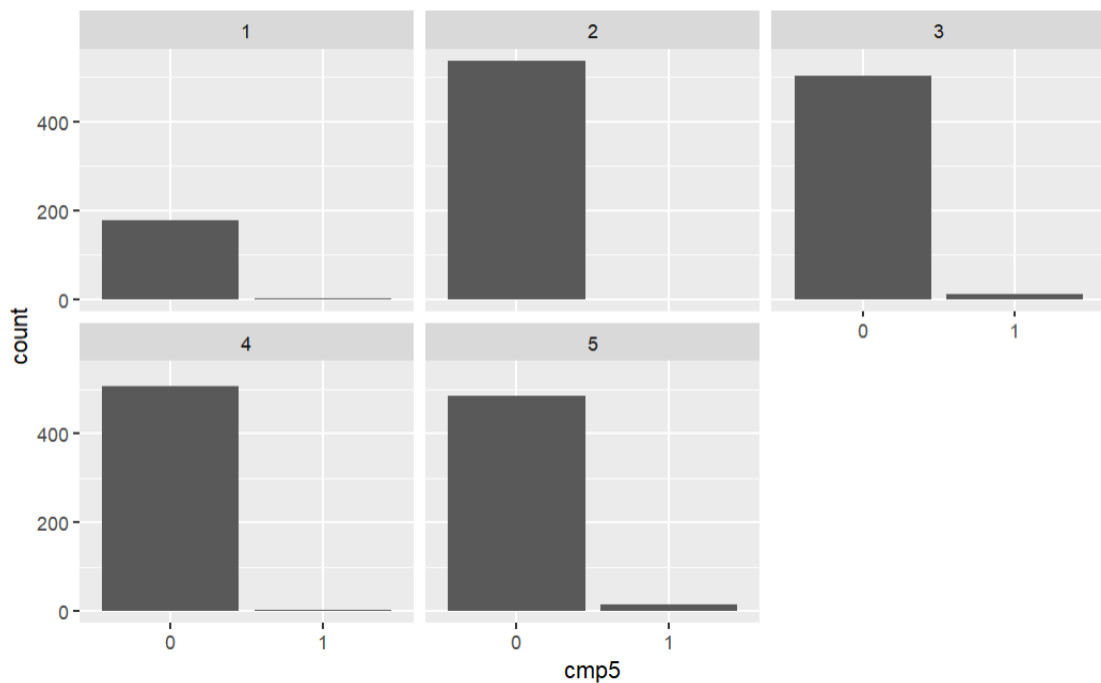
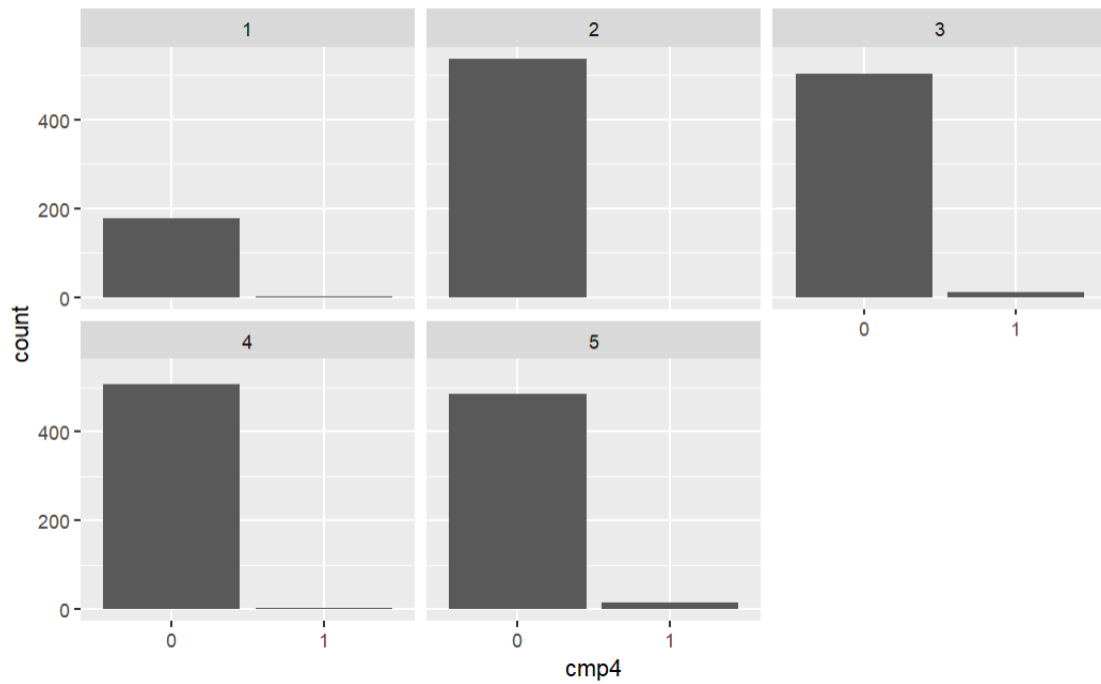
Group 4: Since this group of customers has spent the least money purchasing fruits, fish, sweets, and gold in the last two years, the store can sell them other kinds of products like wines and meats, and their complementary products, which can increase their willingness to come to this retail chain store and keep the customers.

Group 5: Since this cluster of customers is highly educated, with the highest number of Ph.D. education levels. The retail chain store should use data and scientific research results to convince customers rather than alluring them to impulse shopping. Also, they can sell some knowledge products to these highly-educated people, like books, journals, periodicals, software, etc.

(3) An indication of whether the segment includes a high percentage of customers who responded positively to the previous campaign (based on the variable response).







According to the five above column charts, No customers in cluster 4 accepted the offer in the first campaign, and no customers in cluster 2 accepted the offer in the second, third, fourth, and fifth campaigns. Cluster 3 has the most significant number of customers who accepted the offer in the first campaign, which is significantly higher than any other group of customers.

3. A recommendation for each of the customers included in the attached “new customers” csv file of whether the retailer should solicit them in the next marketing campaign. This should be presented as a table that includes the customer’s id number, their probability of responding positively, and their recommended classification (e.g., if predicted response = 1, then contact / predicted response = 0, then do not contact)

My recommendation:

(1) According to the overall demographic makeup of our customer base, the retailer should solicit younger customers with higher education levels and higher income to the next marketing campaign since they are more likely to accept the offer in the incoming campaign. Therefore, the retailer needs to follow the trend, sell high-quality products, and use data analysis and scientific research to convince highly-educated customers to buy their products.

(2) Since the web and store are both predictors of my model, the retailer needs to spend time and effort on the company website and physical store and conduct both online and offline marketing to attract as many customers as possible. For example, they can design their website to increase customers' click-through and conversion rates, use posters and banners, and recruit more experienced salespeople to improve customers' shopping experience in brick-and-mortar stores.

(3) Since recency is a significant predictor of the final model, the retailer should spend more effort on retaining customers who recently bought products from this retail store to reduce customer churn. For example, they can use phone calls, emails, and leaflets to remind customers to repurchase and notify them about the incoming marketing campaign.

(4) The retail should focus on customers who previously accepted the offer in the previous campaign, especially in the first and the third campaign. The retailer can design questionnaires to better understand customers' feedback during the campaign and well-design the details and campaign. In this way, the retailer can make the best use of the incoming campaign to attract more new customers, retain more old customers, and increase sales.

My prediction:

id	.pred_1	response
111	0.000454	0
112	0.023911	0
113	0.002876	0
114	0.005759	0
115	0.005219	0
116	0.002920	0
117	0.007623	0
118	0.000992	0
119	0.001582	0
120	0.000720	0
121	0.003855	0
122	0.000048	0
123	0.003930	0
124	0.000239	0
125	0.479381	0

126	0.000014	0
127	0.980860	1
128	0.007176	0
129	0.000130	0
130	0.002065	0
131	0.959956	1
132	0.002845	0
133	0.968114	1
134	0.013353	0
135	0.004531	0
136	0.003586	0
136	0.002838	0
138	0.000040	0
139	0.899728	1
140	0.019616	0
141	0.000038	0
142	0.987867	1
143	0.000096	0
144	0.000975	0
145	0.000054	0
146	0.000041	0
147	0.017063	0
148	0.000569	0
149	0.035508	0
150	0.000556	0