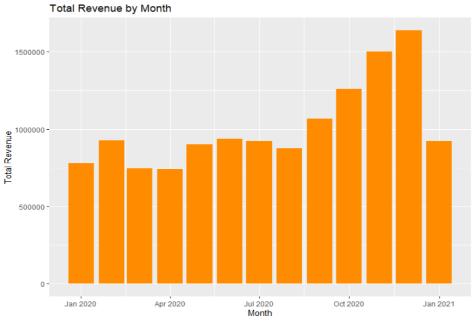
Group 8 : Assignment 3
Group Members : Chia-Ling Ni, Yi-Hsuan Tseng, Mihir Kungulwar, Tejaswini Edupuganti

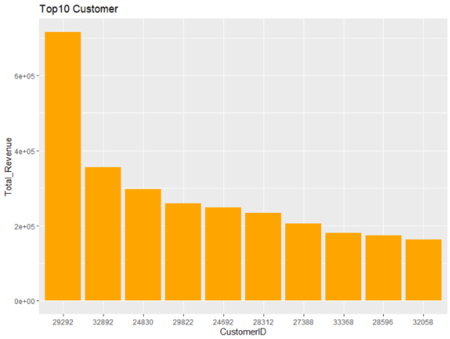
<table>
<tr><th>Questions</th><th>Chia-Ling Ni, Yi-Hsuan Tseng, Mihir Kungulwar, Tejaswini Edupuganti</th></tr>
<tr><td>What are some insights gained from exploratory data analysis? (Use visuals)</td><td>

- **Plot1: Total sales(revenue) by month**



The plot shows the total sales by month from January 2020 to January 2021, comparing the sales amount from each month, we can see that from October to December in 2020 have higher sales.
Insight from plot 1: During holiday seasons, such as Halloween in October, Thanksgiving in November and Christmas, New Year in December, the sales are relatively more than other months.

- **Plot2: Top10 customers by total sales**



CustomerID 29292 seems to be an outlier of the dataset, with a value of 715,374, much higher than other customers.
Insight from plot 2: Plot 2 shows that there is a huge sales difference from each customer, indicating some customers having higher buying capability than others, so we should segment the customers to better target each group.

</td></tr>
<tr><td>Describe steps taken to prepare the data for cluster analysis and reasoning for those steps.</td><td>

1. Remove quantities = 0 data because we want to understand the sales, and quantity 0 means 0 revenue. Create the recency, frequency, and monetary value data based on the customer behavior, and check the correlation for these three variables.

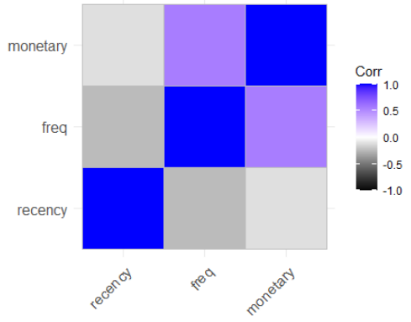
- **Recency** is calculated as the number of days since the last purchase. The smaller the value, the more recent the visit,

</td></tr>
</table>

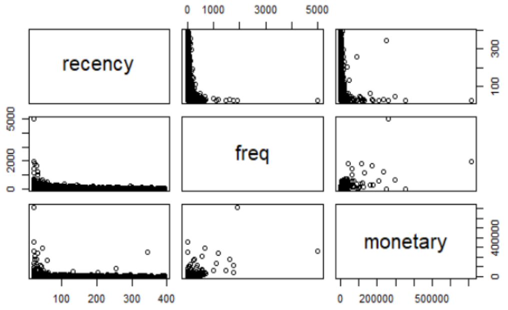
and conversely, the higher the number, the older the visit. (Set the last day as Jan. 31, 2021)

- **Frequency** is the total number of transactions per customer.
- **Monetary** includes the total revenue for each customer.
- **Correlation Plot by three variables:**

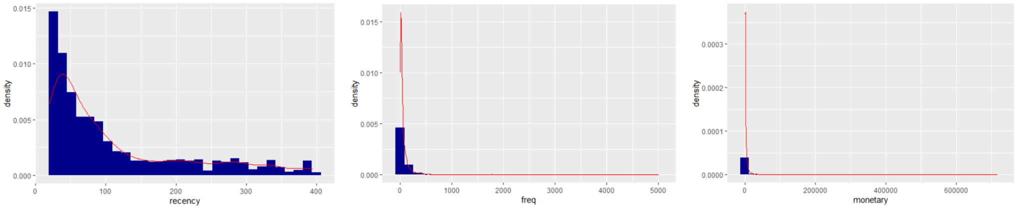
The correlation between monetary and frequency is 0.562. Recency and frequency have a correlation of -0.236. Finally, monetary and recency has a correlation of -0.107.



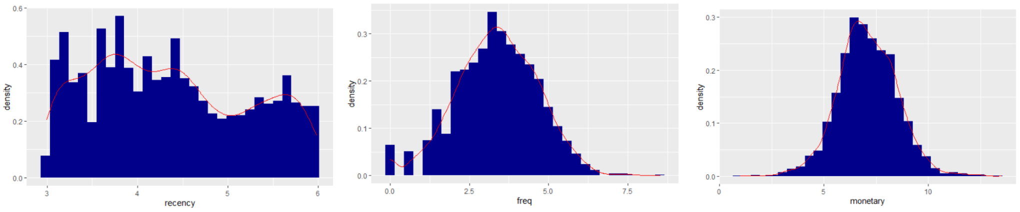
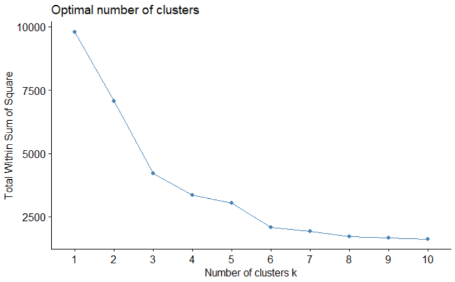
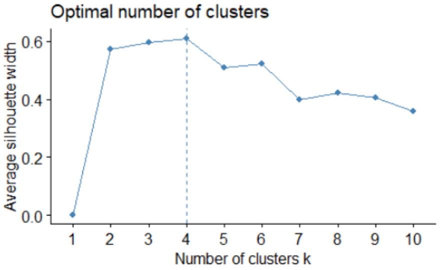
- **Pair Plot by three variables:**



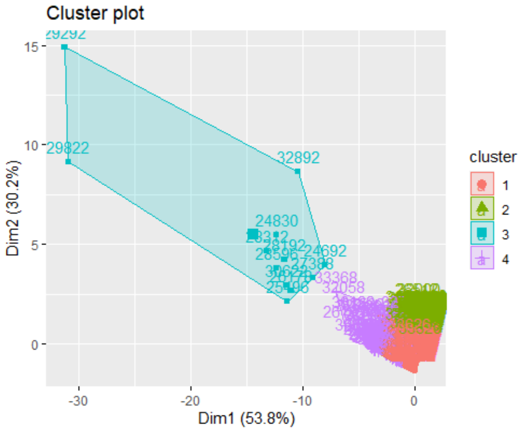
2. Since cluster analysis is sensitive to data distribution, we examined the skewness of the data and found out that the skewness is a concern, especially for Frequency and Monetary values. To correct this, we will take the log of the variables to reduce skewness in the data.

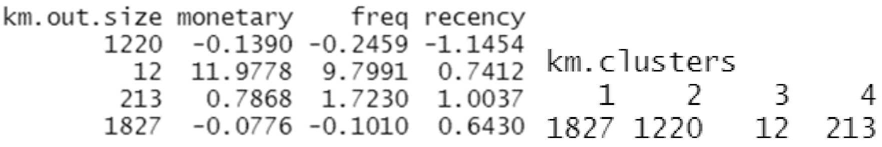


- After taking log on the data, the skewness is better now.

	 3. For cluster analysis, it's important to normalize the variables, otherwise the cluster would be off and center to the values with high values, so we scale these three numeric values. 4. After scaling the data, we also examine if there're missing values in the dataset, and there's no null data in our scaled data. But if there's null values, we have to remove them.
Describe what techniques you used and why you felt they were appropriate.	We use silhouette and elbow method to identify the optimal number of clusters for both K-means clustering and Hierarchical clustering. (example images as below)   Since the data is not explicitly labeled, we basically needed an unsupervised learning algorithm for customer segmentation based on customer visits, spending habits etc. We tried two clustering techniques. K-means clustering - We use the K-means clustering method to find common groups in the data based on certain traits with their mean values. Hierarchical clustering - We calculate the distance, then create a hierarchical cluster and plot it. Since the complete tree is too complicated, we use the optimal k=4 to separate the tree, then look at the characteristics of the 4 groups by means of different variables.
Describe insights gained from your analysis with a discussion of what the results mean	- **Cluster Model 1- Scaled data without log, using K-means=4** Based on both elbow and silhouette point, we choose K=4 for this K-means cluster model.We scaled the data first and will compare the results with the log data for clustering.

and how you assessed performance of the models.





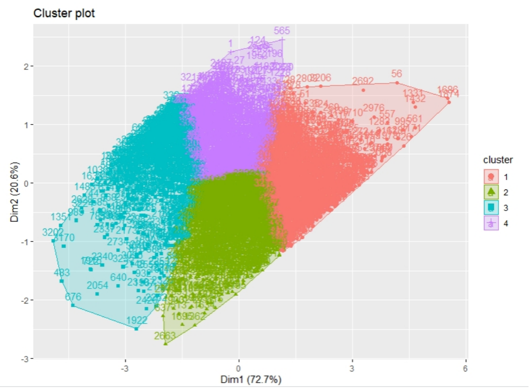
```
km.out.size monetary    freq recency
       1220  -0.1390 -0.2459 -1.1454
         12  11.9778  9.7991  0.7412
        213   0.7868  1.7230  1.0037
       1827  -0.0776 -0.1010  0.6430
```

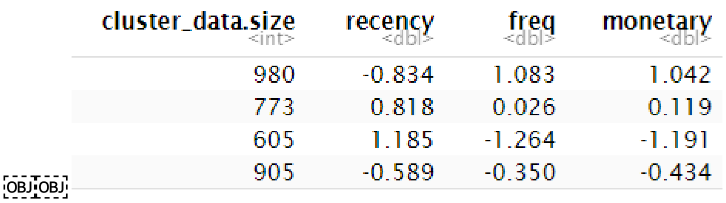
```
km.clusters
   1    2    3    4
1827 1220   12  213
```

Before logging the data, we can see that the outliers (for e.g. CustomerID 29292) have led the clustering model to an imbalance amount. For group3(blue), there are only 12 customers with the highest monetary, frequency and medium recency. Compared to plot 2(Top 10 customer sales), this graph has verified that the outliers might be an issue for clustering. For group2(green), customers have the lowest monetary, lowest frequency and the lowest recency, which means customers in this group have just visited recently but not often. For group4(purple), customers are having the medium monetary, frequency and the highest recency. The result indicates that customers have not visited the store recently.

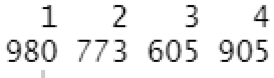
- **Cluster Model 2 - Logged data and scaled, using K-means=4**

Based on both elbow and silhouette scores, we choose K=4 for this K-means model. We still choose K=4 for this h-cluster model.





cluster_data.size <int>	recency <dbl>	freq <dbl>	monetary <dbl>
980	-0.834	1.083	1.042
773	0.818	0.026	0.119
605	1.185	-1.264	-1.191
905	-0.589	-0.350	-0.434

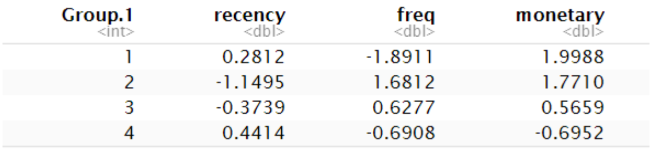


```
  1   2   3   4 
980 773 605 905 
```

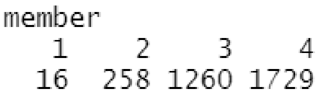
According to the 4 groups from K-means clustering result: We can identify that Group 1(red) has low recency, and higher frequency and monetary, on the contrary, Group 3(blue) has high recency, with lower frequency and monetary. Group 2(green) has median high recency and median low on both frequency and monetary, and Group 4(purple) has median low on all the recency, frequency and monetary. With these 4 groups, we're able to segment customers based on their behaviors.

- **Cluster Model 3** - **Logged and scaled data, using Hierarchy-Cluster**

Based on both elbow and silhouette point, we still choose K=4 for this hierarchy-cluster model. Since the complete tree is too complicated, we use the optimal k=4 to separate the tree, then look at the characteristics of the 4 groups by means of different variables.



Group.1 <int>	recency <dbl>	freq <dbl>	monetary <dbl>
1	0.2812	-1.8911	1.9988
2	-1.1495	1.6812	1.7710
3	-0.3739	0.6277	0.5659
4	0.4414	-0.6908	-0.6952



```
member
   1    2    3    4 
  16  258 1260 1729 
```

According to the 4 groups from Hierarchical cluster-means clustering result: We can identify that Group 2 has low recency, and higher frequency and monetary, on the contrary, Group 4 has high recency, with lower frequency and monetary. Group 2 has median high recency and median low on both frequency and monetary, and Group 4 has median low on all the recency, frequency and monetary.

Evaluate the usefulness of your analysis in terms of how helpful it is to inform marketing strategy for the company. In addition, make	Based on our plots above, because K-means cluster method is basically used for clustering the numeric types of data, and hierarchy-cluster method is primarily used for classifying the category data. Noting that our three segmentations are all in the types of numeric, we prefer to use the K-means Cluster Model 1&2 method in this project to better formulate marketing strategies relevant to each cluster.

recommendations for how the company could use these results to improve its marketing strategy.	To sum up, this cluster analysis and segmentation based on it helps the business to send targeted marketing messages, retain highly valued customers, maximize purchases by other customer bases and consequently enhance profits. This also helps in product development as we can get a better understanding of the customer purchase behavior and cater to those needs. **Recommendation for marketing strategy:** Since Higher recency values mean that the customer has not visited the store recently, and lower recent values mean that the customer has visited recently. Higher frequency means the customer generally visits the online store very often. High monetary values mean the customers spent a relatively high amount in the online store. After reviewing the four clusters we found, we could segment our customers and the value that they bring to the business into 4 groups based on these characteristics. The following are the segments of customers and the recommendations based on characteristics of each of these segments: Firstly, when we look at customers with no frequent purchases and visits to the online store and less amount in spending, a marketing strategy targeted towards developing new products may prove useful to attract them. On the other hand, the customers with medium to high recency, moderate frequency and low to moderate monetary values i.e the medium valued and valuable customers seem to visit the online store but do not end up making purchases or make relatively low purchases. The chances of a purchase by these customers can be increased by providing offers and mailing out reminders to make use of those offers and gain their attention. The fourth and most valuable customers are those with low recency, high frequency and monetary value. It is important that a business retains these customers by providing exclusive perks and member benefits.