**1. What kind of information is available in the dataset?**

The dataset includes user details like age, total orders, average spend per order, user rating, and time since last order. For example, the "Spend per Order" column tells us how much a user usually spends, while "Total Orders" shows how active they are on the platform.

**2. How many rows and columns does the dataset have?**

There are 1000 rows and 7 columns. This means we have data for 1000 users and 7 different features about each one. It is a decent-sized dataset to find patterns and insights.

**3. Can you spot any unusual values or missing information?**

No there is no missing information.

**4. Choose any one user and describe their behavior.**

One user, aged 29, has placed 18 orders, spends 260 per order, and rated 4.5 stars. Their last order was just 3 days ago. This shows they are active, spend a fair amount, and are quite satisfied.

**5. What patterns or questions come to your mind after seeing the first few rows?**

Some users spend a lot, others very little. Some order often, while others rarely do. It makes me wonder do frequent users also spend more? Do happy users order more often?

**1. What is the average (mean) value of two or more key columns?**

The average age of users is around 33 years, and the average spend per order is roughly ₹283. This shows that most users are adults who prefer mid-range spending on food orders.

**2. Which column has the highest variation (difference between users)?**

The Total Orders column has the highest standard deviation. This means user activity differs a lot some order very frequently, while others hardly do.

**3. What are the minimum and maximum values in the dataset, and what do they tell you?**

* The minimum spend is 50, and the maximum is 980. So, some users order very cheap food, while others spend big.
* Ratings also go from 1 to 5, meaning we have both unhappy and very happy users.

**4. Pick any one column and describe how user behavior is spread out.**

Let us take Spend per Order:

* 25% of users spend less than 180.
* 50% (median) spend around 250.
* 75% spend up to 350.

This shows most users spend between 180 to 350 per order, with a few spending much more.

**5. Based on the summary statistics, what are 2-3 possible customer groups or behaviors you think might exist?**

Yes, we can guess a few groups:

* **Frequent users**: People with many total orders.
* **Heavy spenders**: Users who spend 500+ per order.
* **Casual users**: Those who order rarely and spend less.

These groups could be useful for targeted offers or marketing.

**1. Were there any missing values in your dataset?**

No, there is no missing value in the dataset.

**2. Why is it important to check for missing values before analyzing data?**

If we do not check for missing values, our analysis can become inaccurate or misleading. Charts may show wrong patterns, and models might give poor results. It is always better to clean the data first.

**3. If you found missing values, what would you do about them?**

Since only a few ratings are missing, I would fill them with the average rating. That way, we do not lose data, and the overall impact stays small. Removing rows is only better when a lot of data is missing.

**4. What could be some real-life reasons why data might be missing in a food delivery app?**

A user might skip giving a rating, or they just tried the app once and left. Some users may not complete their profile or may not use all features—like choosing favorite cuisine.

**5. How does knowing your data has no missing values help your project?**

If data is complete, analysis becomes easier and more reliable. We can build models, create charts, or run clustering without worrying about errors or gaps in information.

**1. Which columns did you remove from the dataset, and why?**

I removed columns like UserID**.** This do not help in clustering because UserID is just a label.

**2. Which columns did you keep in your cleaned dataset?**

I kept these columns:

* **Age** – helps us know if younger or older users behave differently.
* **Total Orders** – shows how active the user is.
* **Spend per Order** – gives an idea of how much the user usually spends.
* **Rating by User** – tells us how satisfied they are with the service.
* **FavoriteCuisine** – shows favorite cuisine of user.

These features are useful to understand and group users better.

**3. Why should unique IDs (like UserID) not be used for clustering?**

UserID is just a random label. If we use it for clustering, the model might wrongly group users just because of the ID number, even though it has no meaning. It confuses the algorithm.

**4. Why is it helpful to remove text columns (like FavoriteCuisine) in early clustering steps?**

Text columns cannot be directly used in clustering, which needs numbers. Text needs extra steps like encoding. If we use it too early, it may create errors or strange groupings.

**5. How does this data-cleaning step help you get ready for clustering users?**

Cleaning the data and keeping only **relevant numeric columns** makes the clustering process smoother and more accurate. It ensures the algorithm works on meaningful patterns, not on random or confusing data.

**1. Why do we need to standardize (scale) the data before applying clustering?**

We need to scale the data so that all features are treated equally. If one column has big numbers and another has small ones, the model might think the bigger one is more important—even if it is not.

**2. What could happen if we skip the scaling step and apply clustering directly?**

Let us say “App Usage Time” is in hours (like 50), and “Delivery Rating” is out of 5. Without scaling, the model will focus more on usage time and ignore ratings, even though both matter. This leads to unfair or incorrect clusters.

**3. Which columns in your dataset do you think had the largest numbers? And which ones had smaller numbers?**

* **Largest:** Total Orders and Spend per Order (some users order 40+ times or spend 900+).
* **Smaller:** Rating by User (only 1 to 5).

If we do not scale, the large-number columns will dominate clustering, even if small-number features like ratings are important too.

**4. After scaling, what changes in your data? What stays the same?**

The actual values change (they are now on the same scale), but the patterns stay the same. Relationships between users are not lost just made fair for the model to understand.

**5. How does standardization help you get more accurate and fair clusters?**

Scaling helps the algorithm see all features equally. It ensures that users are grouped based on real behavior patterns—not just based on large numbers. This makes the clusters more fair, balanced, and meaningful.

**1. What is the main purpose of using PCA in this project?**

The main goal of PCA here is to reduce the number of features while keeping the important patterns. By converting the data into just 2 features, we can visualize and cluster users more easily.

**2. How many features (columns) did your data have before PCA? And how many after?**

Before PCA, we had around 4–5 numeric features. After PCA, we reduced it to 2 main components. This helps make the clustering process faster and easier to visualize, without much loss of information.

**3. Do you think any important information might be lost during PCA? Why or why not?**

Yes, a small amount of information can be lost because we are simplifying. But PCA keeps the most important patterns, so it is a good trade-off when we want to make the data easier to understand and work with.

**4. What benefits does PCA provide when you want to visualize your user data?**

By turning the data into just 2 dimensions, PCA helps us see clusters or patterns more clearly in a plot. Without it, we cannot easily "see" what is going on if we have too many features.

**5. Imagine you had to explain PCA to a friend. How would you describe what it does in one or two sentences?**

PCA is like summarizing your school report card into just two scores: one for academic strength and one for hobbies. It takes many numbers and turns them into fewer, smarter features so we can spot patterns easily.

**1. What is the main purpose of using KMeans clustering in this project?**

We used KMeans to group similar users together based on their behavior. It helps us discover patterns, like who orders often, who spends more, or who is less active.

**2. How many clusters did you create, and why?**

We created 3 clusters. This number was chosen because it showed clear separation between different user types. It seemed like a good balance not too few, not too many.

**3. What does each cluster label represent?**

Each cluster represents a group of users with similar habits. For example:

* Cluster 0 might be low spenders who order rarely.
* Cluster 1 could be active users with high satisfaction.
* Cluster 2 might be new users or those who have not ordered in a while.

**4. How do you think this grouping could help a business?**

The business can use clusters to:

* Target ads to active or high-spending users.
* Send reminders or offers to less active users.  
  This helps improve marketing and customer retention.

**5. Were you surprised by anything after applying clustering? Why or why not?**

The groups mostly made sense, but it was interesting to see how clearly the users separated just based on their data. It made me curious to explore what exactly made each group different.

**1. What does each dot in your scatter plot represent?**

Each dot represents one user. Its position on the plot is based on two main features (after PCA), showing how similar or different that user is from others.

**2. How many clusters do you see in the chart, and how are they separated?**

There are 3 clusters in the chart. They are mostly well separated, but a few dots near the edges are a bit closer. This means most users fall into clear groups, but some users share traits from more than one group.

**3. Choose one cluster and describe what kind of users might belong to it.**

Let us take Cluster 1 — it’s placed away from others and has tight points. These users are likely highly active, spend more, and give good ratings. They might be regular and loyal customers.

**4. Why is it helpful to visualize clusters like this instead of just looking at raw numbers?**

The plot helps you see the shape and spread of the user groups. It is much easier to spot patterns visually than by reading through rows of numbers or stats.

**5. If you were running a food delivery app, how could this chart help you improve your service?**

If I saw a cluster of low-activity or low-rating users, I could focus on improving their experience maybe with better offers, support, or faster delivery to boost engagement.

**1. What is Agglomerative Clustering and how does it work?**

Agglomerative Clustering is a bottom-up method. It starts by treating each user as its own cluster, then slowly merges the closest one’s step by step until only a few groups remain.

**2. How is Agglomerative Clustering different from KMeans Clustering?**

* KMeans starts with several clusters and adjusts them by minimizing distances.
* Agglomerative builds clusters by merging from the bottom up.  
  When I switched to Agglomerative, the clusters looked more natural, especially for uneven user groups.

**3. How many clusters did you create, and how did the user distribution look?**

I created 3 clusters, just like KMeans. The groups were not perfectly balanced—one cluster had more users. This could mean most users have similar habits, while smaller clusters represent special behavior types.

**4. Choose one cluster and describe the possible behavior of users in that group.**

One small cluster grouped users who spent a lot per order and gave high ratings. These are likely loyal and premium users, who might order for families or prefer better-quality food.

**5. Which clustering method (KMeans or Agglomerative) do you feel gave better groupings — and why?**

I think Agglomerative Clustering gave better groupings. It handled uneven user types more smoothly and didn’t force users into strict equal-sized clusters like KMeans sometimes does.

**1. What does each dot in the scatter plot represent?**

Each dot represents one user. It shows their behavior pattern after reducing the features. It is important because it helps us visualize how similar or different users are from each other.

**2. How many clusters are shown in the plot? Do they appear clearly separated or overlapping?**

There are 3 clusters shown. They are well-separated, but not perfect some points are close to other groups. Overall, the shapes are visible, and we can tell the groups apart.

**3. Pick one cluster and describe what kind of users it might represent.**

One cluster is tightly grouped and far from others. This could represent active users who order often, spend more, and give high ratings. Their behavior is more consistent than others.

**4. How does this clustering result compare to what you saw with KMeans?**

The clusters here look more natural and flexible than in KMeans. KMeans gave clearer borders, but Agglomerative seems to respect the real differences in user behavior better. I found this method more meaningful overall.

**5. Why is visualizing user clusters important for a company like a food delivery app?**

It helps the company see who their users really are like loyal users, casual ones, or low-rating customers. This can guide personalized offers, app updates, and marketing strategies more effectively.

**1. What is a dendrogram and what does it show?**

A dendrogram is a tree-like diagram used in hierarchical clustering. It shows how users are grouped step by step from individual users to larger clusters—based on how similar they are.

**2. How many user groups (clusters) do you think are visible in the dendrogram?**

Looking at the dendrogram, I would draw a horizontal line that cuts the tree into about 3 clear clusters. That is where the branches are clearly split apart and not too close together.

**3. What does it mean when two users or groups are merged at a very low height?**

If two points merge at a low height, it means they are very similar. For example, if two users merge early, they probably spend similar amounts and have similar order habits.

**4. Was the dendrogram result similar or different compared to KMeans and Agglomerative clustering visual plots?**

It was similar to Agglomerative, since both are hierarchical methods. But the dendrogram gives a clearer picture of how close or far users are at each merge. KMeans gives fixed groups, while this shows how clusters are formed gradually.

**5. How could a business (like a food delivery company) use these cluster insights in real life?**

A food delivery app can use this to identify loyal users vs. one-time users, and send personalized offers or improve service for low-rating groups. It helps in targeted marketing and user retention.

**1. What is DBSCAN, and how is it different from KMeans or Agglomerative Clustering?**

DBSCAN groups users based on how close they are to each other (density). Unlike KMeans or Agglomerative, it doesn’t need you to choose the number of clusters and it can spot outliers that don’t belong to any group. This makes it great for finding unusual users.

**2. How many user groups (clusters) were formed by DBSCAN?**

DBSCAN formed 2 main clusters, and found some outliers. One cluster had most users, while the second group was smaller and more distinct.

**3. Did DBSCAN identify any outliers? If yes, how many and what might these users represent?**

Yes, DBSCAN marked some points as outliers (often labeled as -1). These could be users who rarely order, spend unusually high or low, or behave very differently from the majority.

**4. Compare the DBSCAN results with your earlier clustering (KMeans or Hierarchical). Were the groupings similar or different?**

The groupings were somewhat different. DBSCAN gave more flexible clusters and spotted outliers, while KMeans and Agglomerative grouped everyone, even the unusual ones. I found DBSCAN more useful for finding edge cases or rare behaviors.

**5. How could a company use the DBSCAN results to improve its food delivery service or marketing?**

The company could study the outliers to understand why those users are different—maybe they need special offers or support. The clusters can also help with targeted campaigns, like sending deals to frequent spenders or re-engaging inactive users.

**1. How many distinct groups (clusters) can you see in the scatter plot?**

Looking at the plot colors, we can see 2 main clusters. DBSCAN has grouped users based on how close they are in behavior.

**2. Are there any outliers (users marked as -1)? Where do they appear in the plot?**

Yes, there are a few outliers marked as -1. They appear far away from the main clusters, usually on the edges of the plot. These users behave very differently from most others.

**3. Does the shape or size of the cluster suggest anything about user behavior?**

Yes. One cluster is tightly packed, which means those users have similar habits. Another cluster is more spread out, suggesting more variety in that group’s behavior.

**4. How does this DBSCAN visualization compare to the KMeans or Agglomerative Clustering plot?**

The clusters look a bit different. DBSCAN is better at finding uneven shapes and detecting outliers, while KMeans and Agglomerative forced everyone into groups. DBSCAN gives a more realistic view of unusual users.

**5. If you were working in the marketing team of a food delivery app, how would you use this visual insight to improve service or engagement?**

I would look at the outliers to see why they are different maybe they had bad experiences or unusual spending. The main clusters can help us target similar users with personalized offers or app features based on their behavior.