# Report

## Data Preprocessing:

### Identifying Useless Features:

Imagine you have a bunch of data about customers, like their ID numbers or the date they became customers. While these details might seem important at first, they don't actually tell us anything useful about who the customers are or what they might want. So, we decide to toss out these irrelevant features to focus only on the ones that really matter in understanding our customers.

### Handling Missing Information:

Now, let's say we're missing some key info, like the income of some customers. This gap in our data could throw off our analysis. To fill in these blanks, we come up with a clever solution. We look at what income information we do have and combine it with other details, like the level of education, to estimate the missing incomes as accurately as possible.

### Dealing with Redundant Data:

Sometimes, there are details that are exactly the same for every single customer. It's like having a column in our data that just repeats the same thing over and over again. Since this redundant information doesn't give us any new insights into our customers, we decide to get rid of it to streamline our analysis.

### Eliminating Duplicates:

Imagine accidentally counting a customer twice in our data. This duplication could mess up our analysis by making it seem like we have more customers than we actually do. To avoid this confusion, we carefully check for and remove any duplicate entries to ensure our data is clean and accurate.

### Focusing on Main Customer Groups:

Now, when we're trying to understand our customers, we're not interested in outliers or extreme cases. Instead, we want to focus on the main groups of customers—the ones that represent the majority of our business. So, even if a few customers make unusually large purchases, we choose to overlook them in favor of understanding what the bulk of our customers want and need.

### Simplifying Categories:

Some details in our data might be a bit complex or hard to interpret. For example, the marital status of a customer could be expressed in various ways. To make things clearer and easier to analyze, we decide to simplify these categories. For instance, instead of having multiple options for marital status, we might just classify customers as either living alone or with a partner.

### Creating Easier Metrics:

We're always on the lookout for ways to simplify our analysis. So, if there are certain metrics that could make our job easier, we're all for it. For example, by combining the number of kids a customer has with their living situation, we can create a new metric called "family size." This makes it easier for us to understand the dynamics of each customer's household.

**Summarizing Important Information:**
Lastly, we're interested in getting a big-picture view of our customers. That's why we might sum up certain details, like how much they've spent on our products or their age based on their birthdates. These summary metrics help us quickly grasp the key characteristics of each customer without getting bogged down in too much detail.

By following these steps, we're able to clean up our data and extract meaningful insights that can drive our business forward.

**What is Customer Identity Analysis:** Customer Identity Analysis is utilized to understand the customer profile. This understanding assists in making informed decisions regarding product purchases, product balance, and other related factors. For instance, consider a scenario where you purchased 200 copies of product X, but only 100 were sold, and the remaining 100 expired and had to be discarded. In this case, you incur a loss. However, by understanding the customer identity and determining the optimal quantity to sell on average, you can adjust your purchase quantity to 110 units, thereby avoiding losses.

By understanding our customers on a granular level, we can effectively target our marketing efforts. This means delivering personalized advertisements, promotions, and recommendations that resonate with each customer's preferences, leading to higher engagement and conversion rates.

Numerous companies across various industries have successfully leveraged Customer Identity Analysis to drive significant improvements in profitability and customer satisfaction.

For instance, e-commerce giants like Amazon and Alibaba have revolutionized the retail industry by harnessing the power of data analytics to deliver personalized shopping experiences tailored to each customer's preferences. In summary, Customer Identity Analysis is not just a tool; it's a strategic imperative for any business looking to thrive in today's data-driven world. By investing in this analytical approach, we can unlock hidden opportunities, drive revenue growth, and cement our position as a market leader.

In the modern world, even when we talk about a specific product, we begin to see advertisements for that particular product and similar products. :))

**Data Review and Disclosure:**

| Cluster 2 | Cluster 1 | Cluster 0 |
| --- | --- | --- |
| Half single half together. | Half single half together. | Most of them single. |
| University graduated. PhD. Master. | Most of them university graduated 2n cycle and phd. | Most of them are not university graduated. |
| Most of them have no children. If have it is most likely to be teen. | Most of them have children. | Most of them have no children. |
| Have the highest income. | Have the 2nd the highest income. | Have the lowest income. |
| Makes highest spending. | Makes 2nd highest spending. | Makes the lowest spending. |
| Age ranges between 30 - 80 | Age ranges between 40 - 80 | Age ranges between 40 - 80 |
| Highest consume ratio: Meat | Highest consume ratio: Gold | Highest consume ratio: Gold |

Based on the data review, it's evident that Cluster 2 exhibits the highest spending behavior and boasts the largest income among the distinct main clusters. Additionally, a majority of individuals in this cluster hold advanced degrees such as PhD and master's degrees. Roughly half of them are single, while the other half are in relationships. Most of them do not have children, and if they do, their children are typically teenagers. Their ages range between 30 and 80, and they predominantly consume meat.

Cluster 1 follows closely behind Cluster 2 in terms of spending, and they possess the second-highest income. Many individuals in this cluster have completed their second cycles of education and hold PhDs. About half of them are single, and a significant portion have children. They tend to purchase gold products frequently.

On the other hand, Cluster 3 demonstrates the lowest spending behavior and has the lowest income among the main clusters. The majority of individuals in this cluster have completed high school. Most of them are single, and a large portion do not have any children. Their ages range from the late 20s to 80s, and they also tend to purchase gold products.

By the way, wine is the most purchased stuff.

**Advises:**
It's crucial to cater to the needs and preferences of these main customer clusters to maximize profits. For instance, you could tailor your product offerings based on whether customers have children. For example, if you observe that most individuals in Cluster 1 have children, you may consider adding diapers to your product inventory. Similarly, if gold products are popular among multiple clusters, you could increase your stock of gold items to meet.

Since the main customer base consists of adults, and wine is the most purchased item, it might be beneficial to experiment with purchasing different types of alcohol to gauge their popularity. Considering that Cluster 0 has the lowest overall income, it would be prudent to focus on acquiring inexpensive products to cater to this demographic. Conversely, for Cluster 2, which has the highest income, investing in expensive, luxury products may yield favorable results.

Moreover, considering that Cluster 2 has a significant number of teenagers as children, you could also explore purchasing items that appeal to this demographic to meet their needs and preferences. By adapting your product offerings based on the characteristics and preferences of each customer cluster, you can enhance customer satisfaction and potentially increase profits.