After the test data is generated using the faker library, the next logical step is to anonymize it. Before anonymizing the data, we need to generate the hierarchies for each of the columns in the data. The given data contains:

| Name | (String) Name of the consumer |
| --- | --- |
| Email | (string) Their Email address |
| Gender | (String) Gender of the customer making the purchase |
| Amount Spent | (Int) The amount being spent by the customer |
| Zip Code | (Int) The zipcode of the consumer |
| Age | (Int) The age of the consumer |

* Data Pre Processing:

If the data contains any missing values or floating point numbers, this would be a good stage to remove them and also convert them to int as pyarxaas does not support missing values and floating point numbers. Also storing them as int instead of floating point numbers makes a lot more sense as snowflake works on a per second billing model and converting the column from float to int reduces the data storage size in half thereby saving you the unnecessary costs

After the preprocessing is done, now is the time to create hierarchies for the data. While making hierarchies for the data, it is crucial to make sure that every possible data point is covered or else the dataset will not be anonymized.

### Hierarchy Generation for the dataset:

For every column of the data, hierarchy generation would look like a different problem as not all columns of the data are created equally and not all of them store data equally

The name column is highly sensitive and hence should not be released at all as the identity of the users is correlated with the name and all of the sensitive information will be released by their name. Hence name is a sensitive attribute

The email column is also very sensitive but not as sensitive as name since people tend to use pseudonyms for their emails. Hence email is a quasi identifying column and can build the hierarchy for the same using redaction based hierarchy

The age column is a quasidenifying column as it can be combined with gender, zip code or any other column of the same or different dataset and get information out of the same

The gender column does not reveal a lot of information and needs to be present for the analysis and hence the gender column is an insensitive column

The amount spent is not a sensitive column as it is very tough to look at the prices spent by a user and narrow it down to a specific user. It can still be done but this is a crucial column with the analysis hence instead of specifically mentioning the amount spent, we can specify the range so that the data is still sound for analysis. We can use a interval based hierarchy for the same

The hierarchy generation for the columns is as follows: