**SOCIAL NETWORK ADS**

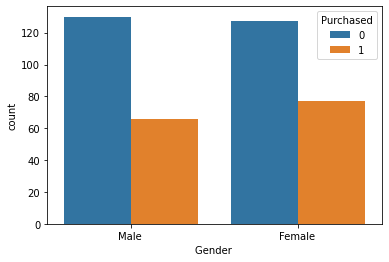
**Problem Statement:**

We have a real-life data of whether a product is sold when an ad is shown to people. The dataset contains 401 rows. The independent variables are user\_id, gender, age and estimated salary. The dependent variable purchased contains 0 – for not purchased and 1- for purchased.

**Data Pre-processing:**

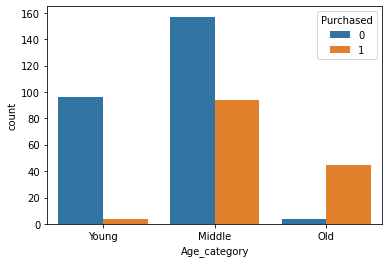
By observing the data, there is no missing values or 0 values. Then we plot between different variables to get additional information from the data.

Male vs Female:



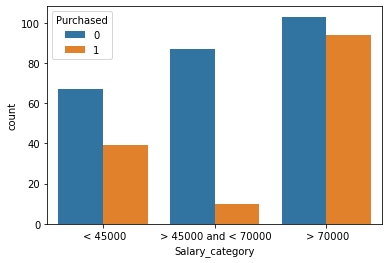
From the plot we can note that females are more influenced by these ads and tend to purchase more than men.

Age vs purchased:



From this plot we can observe that people between the age 30 and 50 tend to purchase the products. Also, old people whose age is greater than 50 tend to buy less. The young people who are below 30 stands between the old and the middle aged.

Salary vs Purchased:



Here we can strangely observe that people with salary less than 45000 tend to but more than people with salary between 45000 and 70000. And obviously people with higher salary than 75000 purchased the most.

Encoding:

The gender column contains string. So, we have to encode it to numeric. So, we assigned the value 0 for female and 1 for male.

Remove unwanted column:

In our dataset we have a column with user id. It is needed for dbms only. For our analysis we don’t need that, so we removed that column.

Scaling:

The StandardScaler assumes your data is normally distributed within each feature and will scale them such that the distribution is now centered around 0, with a standard deviation of 1. This process increases the accuracy of our model.

**KNN Algorithm:**

* K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

**ACCURACY:**

The accuracy for this model is 92.5 %. This is a high accuracy, which means we can use this model for predictions which would give results which have the probability of being 92.5 % correct.