# Discussion of the What, Why and How:

In this problem I have tried to predict the optimal price tier for each user in the test set by using several Machine Learning Models. I have total 8000 unique individuals who has bought products at different time.

## Missing Data Handling:

1. I have started the project by understanding both the training and test data and found that the datasets have some missing values. Product, Currency, gross\_usd\_amount, and PRICE\_TIER are the variables with 68.96% of missing data.
2. Price\_tier is the dependent variable, whereas other variables are independent variables. Instead of dropping the variables with missing data, I have used a simple data engineering technique to impute the missing data.
3. I have used back fill and forward fill technique to fill up the missing rows for product and used interpolation technique along with back fill for the price tier and gross\_usd\_amont variables. I have used mode technique to fill up the currency column.
4. To fill up the missing values, I have assumed that the variables are not related and imputed the product irrespective of the price and solely based on the purchase history individuals at the individual levels. I first fill up the product missing rows.
5. For the price and revenue used the history of these two variables at the product levels. Ex. To impute the price of good X, solely used the available price records of the good X. To impute the revenue generated from good X, solely used the available revenue records from good X.

## Removal of Outliers:

I have removed the outliers from the numerical variables by trimming the extreme values (at the 3 standard deviation points from the upper end).

Exploratory Data Analysis:

1. The underlying data has both numerical and categorical variables. As a starting point I am just considering the numerical variables and first checked the correlation between the variables. I found that gross\_usd\_amount, MAX\_PRICE\_TIER, and AVG\_PRICE\_TIER variables are highly positively correlated with the target variable.
2. I then separately visualize the relationships between the target variable and the variables that are highly and positively correlated with it by drawing scatter plot.

Model Building:

1. For the model building I have used training datasets and evaluate by using the test dataset. I have used very basic models as the starting the point Ridge regression, Lasso regression, XGBoost, and random forest Model.
2. I have then compared the RMSLE and R2 values of all the models and found that the XGBoost model works the best in predicting the optimal price for each of the customers.

Few Notes:

1. I have an evening to solve this complex problem.
2. This is not the best solution I will be providing to my client and will take at least one month to finish it.
3. These results are just glimpses of my work.
4. I start my analysis by spending some time to understand the data and a significant amount of time on the EDA portion.
5. In this problem, I have used numerical variables to predict the model, however, given time I use several ways to encode the categorical variables and then use them as features as well.
6. I also often check the variance covariance matrix to see how different features are related.
7. If necessary I drop few highly correlated variables (To remove the problem of multicollinearity) and sometimes use PCA to make a new variables rather than dropping variables.
8. Technique highly depends on problem to problems and there is no one way.
9. I think some more information about the customer at the induvial levels like, gender, age will allow to make a better model. I also think that the demographic variables also play a significant role while making a model. These kinds of data I would like to ask from the clients.
10. Lastly, I were given four weeks of time I will do an extensive the data engineering and will do EDA. I will split even my training dataset as a training and validation dataset. Then, I will applying my selected model to check its performance on the test dataset.