Problem statements and solutions Q&A:

Programmer: Souparna Bose

1. Do a descriptive analysis of all the variables

Using the dataset ‘**trainingData’**, I have performed a descriptive analysis for individual features as well as their co-relation and variation with other features as well.
Using **pandas** I have loaded the dataset into a dataframe and described important details regarding the same. Then using **seaborn** and **matplotlib**, I have shown visualization of various features and their relations.
Please find the notebook **Data-Analysis-and-Visualization.ipynb** or **Data-Analysis-and-Visualization.pdf** for further details.

2. There is a new customer who needs a loan. Which models will be best suited to predict the loan_amount that can be granted to the customer?

Since the **loan_amount** is a continuous value and not a category, **Linear Regression** model and Deep Learning using **Artificial Neural Network** (perceptron) model will be best suited to predict the outcome.

3. Build a model to predict the maximum loan_amount that can be granted to the customer. Which all variables are good predictors?

Please find the notebook **Data-Analysis-and-Visualization.ipynb** or **Data-Analysis-and-Visualization.pdf** & **Model-creation-training-prediction-evaluation.ipynb** or **Model-creation-training-prediction-evaluation.pdf** for further details.
The variables **annual_income, monthly_expenses, home_ownership, house_area** and **loan_amount** are the best predictor variables.

4. Is loan_purpose a significant predictor? The business has insisted on using loan_purpose as a predictor. If it is not already a significant contributor, can we still modify the model to include it?

loan_purpose might be an important thing with respect to understanding the customer’s reason for seeking a loan. But for model prediction and evaluation, it does not have a significant importance, compared to other features.

We can still however modify the model to include it. But for the model to understand and interpret it properly, since it is a categorical data, we need to convert it to **dummy variables** for model training.

5. How will you measure the fitness of the model? Which metrics (accuracy, recall, etc.) are most relevant?

We can measure the fitness of the model by using scikit-learn metrics for evaluation and error calculation.
The important metrics include **mean_absolute_error(MAE), mean_squared_error(MSE), root_mean_squared_error(RMSE)** and **r2_score(R2 score)**.
The fitness and metrics for the model created has been documented, please refer to **Model-creation-training-prediction-evaluation.ipynb** or **Model-creation-training-prediction-evaluation.pdf** for more details.