Tasneem, EDA Notes:

1. Load Data
    a. View First 5 Rows
2. Identify Variables
    a. Describe() Variables
    b. Identify Predictors, Target, and Variable Types
        i. List these categories
    c. Recode variables: Clean u
        i. Ensure that coding reflects ordinality as applicable
3. Univariate Analysis
    a. In this section, consider:
        i. Class Imbalances
        ii. Missing Values
        iii. Variance
        iv. Distribution
            1. Consider expected distribution vs actual distribution
        v. Outliers
            1. IQR Percentiles
            2. Do Min and Max make sense?
            3. Identify Natural Outliers
        vi. Frequency and Percentages for Categorical Variables
    b. Plot Numerical Variables and Bin as Necessary
        i. Limit Balance: Histogram
        ii. Age: Histogram
        iii. Bill Amounts (6 variables): Boxplot with bill amount variables on x axis
        iv. Payment Amounts (6 variables): Boxplot with payment amount variables on x axis
    c. Plot Categorical Variables
        i. Sex: Frequency Table, Bar Chart
        ii. Education: Frequency Table, Bar Chart
        iii. Marriage: Frequency Table, Bar Chart
        iv. Age (if binned): Frequency Table, Bar Chart
        v. PaymentHistories (6 variables): Frequency Table (pivoted?), with stacked bar chart, with payment history values inside bars
4. Multivariate Analysis
    a. Too many variables to explore all relationships. Start with Independent Variables vs Target and Dimension Reduction Techniques
    b. Plot all Independent Variables vs Target variable:
        i. Continuous vs Categorical (box plots for each level of categorical)
        ii. Categorical vs Categorical (stacked column bar chart)
    c. Dimensionality Reduction Techniques
        i. Variance: Remove independent variables with zero variance and possibly near zero variance

Tasneem, EDA Notes:

ii. Correlation: Remove independent variables that are highly correlated with each other
iii. Consider Variable Transformations
    1. Consider combining multiple variables into one new one
iv. Iterative Techniques
    1. Variable Importance Factor
        a. If needed in attempt to improve accuracy and other metrics, however:
            i. Random Forest takes forever to run. Computationally expensive.
    2. Recursive Feature Elimination
    3. Principal Component Analysis
v. Create dataset with selected features
vi. Plot selected features of datasets against target variables

d. Plot Other Relationships as necessary
    i. Continuous vs Continuous (scatterplots)
        1. Plot: Bill Amounts vs Limit Balance
        2. Plot: Paid Amounts vs Limit Balance
    ii. Categorical vs Categorical(Stacked column bar chart)
        1. Consider: Two-way tables
    iii. Continuous vs Categorical(Box plots for each level of categorical)

Other Notes:

Techniques for Addressing Class Imbalances

- Accuracy Paradox: If our model simply predicts that our customers won't default, than, it will be right most of the time since majority of our data consists of customers who did not default (22,000 compared to about 7,000). We will need to explore performance metrics in addition to accuracy:
  - confusion matrix, precision, recall, f1-score
  - kappa, roc curves
- resampling dataset
  - Sampling with Replacement: We will add copies of instances from the default class.
    - Depending on time constraints, we can try the Synthetic Minority Over-sampling Technique
  - Undersampling: Delete instances from the did not default class.
- we will favor decision tree algorithms, which do well with class imbalances

Source:
https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/