The Amazing Bank (AB) is one of the leading financial institutions in the world. You have been recruited as a freelance data science consultant by AB to help the bank design their credit risk strategy, enabling data driven decisions. The CEO of the bank writes the following email to you:

"Welcome aboard and we are very proud to have you. Our existing credit strategy need some serious fine tuning as it has completely failed to identify potential default behaviors in the post covid world. We need your expertise and help to support us with redesigning our credit risk strategy in this new covid world. We have shared a sample data for you to get started which has data from Mar 2020 till current date; please let us know if you might need more data or any other requirements in specific for you to get started. ***More than identifying defaulters, we also want to understand why they would default; that's the key****. Very excited to look forward to what you can bring to the table!"*

- Bravo, CEO, Amazing Bank

You skimmed through the data and learnt that there are 1 million customers, 1000 features, with 700 numerical and 300 categorical and 5% defaulters, and there are quite a few missing values as well in different levels. *Think aloud and help us understand your approach towards solving this problem!*

a) What would be your first step? List different EDA you would like to do with the data before you get started.
b) How are you going to handle missing values? Ideate and list them.
c) Before getting into modeling, apart from points a. and b., do you want to do anything else with the data to understand default behavior?
d) The default labeling is based on customers who did not pay 3 installments continuously. Do you want to rethink about this labelling strategy for the target? How will you validate the labelling strategy is correct?
e) What will be your X and Y?
f) How are you going to handle outliers, numerical columns and categorical columns?
g) Do you want to include the entire 1000 features?
h) What model do you want to choose and why?
i) What is your validation strategy?
j) How are you going to handle class imbalance?
k) What will be your experiments and how are you going to choose the best model?
l) What metrics are important for you in evaluating your best model?
m) "***More than identifying defaulters, we also want to understand why they would default; that's the key"*** – The CEO specifically mentions this in his email. What is your strategy to address this concern?