DATA PREPARATION & EXPLORATION STEPS	CHECKLIST
DEFINING THE PROBLEM	
Is there a working hypothesis or a well-defined problem?	□
What type of analysis is needed?	□
Explorative?	□
Causative?	□
Predictive analytics?	□
Prescriptive analytics?	□
Is there staff and C-suite buy-in?	□
Is there a business case for finding a solution to the problem?	□
Is there a practical and affordable approach?	□
LOCATING AND RETRIEVING DATA	
Are there data privacy concerns?	□
What is the data format?	□
Structured data?	□
Unstructured data?	□
Who owns the data?	□
Did you consider confounding variables?	□
Does the data require extract, transfer, and load (ETL)?	□
Does the data require SQL queries?	□
DATA PREPARATION	
What is the file type?	□
Will the file type need conversion?	□
Are the data in a flat file or database?	□
How many rows and columns are there in this dataset?	□
What are the data types (numerical, categorical, etc.) in the dataset?	□
Is there a data dictionary?	□
Does each column have a header that makes sense with no spaces?	□
Did you visualize the data with univariate and bivariate plots?	□
Does the dataset include the units/values of the data?	□
Is there a target or outcome column?	□
Descriptive statistics?	□
Do the minimum and maximum values make sense?	□
Are the mean and median similar?	□
How large is each variable's standard deviation?	□
What is the distribution of the data?	□
Normally distributed?	□
Skewed to the right or left?	□
If skewed, do you need to transform the data?	□
Are there outliers?	□
Are there duplicates?	□
Are the data on different scales?	□
Do the data require standardization?	□

Do the data need to be combined by stacking or joining?	□
Do the data require a pivot table?	□
Did you cleanup incorrect cell data?	□
Did you prevent data leakage?	□
How did you handle missing data?	□
DATA EXPLORATION	
Do you have high correlations between predictors?	□
Do you have high correlations between the predictors and the outcome?	□
Do data need to be encoded?	□
Dummy encoding?	□
One hot encoding?	□
Do data need to be placed in bins?	□
Do you have too many predictors and need dimension reduction?	□
Do you have a class imbalance problem?	□