Q1): What preprocessing steps would you perform on the train dataset before using it for modelling.

* I handle missing values perform Text cleaning (Preprocessing) [Punctuation, special characters, and numbers were removed from the text. Text was converted to lowercase. Tokenization and stemming were performed on the text data.
* I encoded categorical column[“drugName] to numerical column using LabelEncoder
* I perform Feature Engineering, Scaling Numerical Features & Removing Unnecessary Column’s.

Q2) Are there any missing values in the dataset? If yes, what is the nature of missing values and how might this affect model performance. What are some techniques to mitigate this issue.

* Yes, there are missing values, I go with dropping the null values, because more than 90% data are missing. We can’t fill the missing based on 10% data.
* If there will be an option for filling the missing values which are not “particularly in the data”. I will go with filling with mode because it’s categorical data

Q4) Do you see any potential features in the review column that could be useful for predicting the drug name and condition.

* Yes the Potential features derived from the 'review' column include Word Frequencies’, Stem Word, Contextual Information, Length of Review,TDF-IDF Vectorixation, NER, Topic Modelling etc.

Q5) Can you provide some statistics and visualizations to summarize the distribution of categories (drug name)

* Yes, my provided Jupyter notebook (code) includes some exploratory data analysis and visualizations related to the distribution of drug categories ('drugName'). Here are the key statistics and visualizations:
* Top 20 Drug’s,Distribution of Drugs By Categories, No of Drug per Drug Category, Bar Plot for Drug Categories and so on.. have a look to notebook which I attached In this file.

Q6) Are there classes which are underrepresented or overrepresented? How might this skew the model’s prediction.

* Yes, the given data is very big so the prediction accuracy will not be perfect(best) I give problems like underfitting, overfitting, But the problem is not related skewness it is good not a problem.
* that's why we are going with sample (-TOP 4 DRUG)

Q7) What are some techniques that can be used to address this? Does the model performance improve after incorporating such strategies.

* Resampling, weight loss,Data Agumentation,Ensemble Methods,Transfer Learning, Anomaly Detection, Cost Loss Function etc.

Q8) What Modelling strategy would you consider for predicting the drug name from the dataset and why.

* Muticlassification, Ensenmble Methods,NLP,Feature Engineering and Text Preprocessing,Hyperparameter Tuning.

Q9) How can we assess model robustness (Cross Validation Strategies).

* By consistent performance,variability,overfittig detection

Q10) What evaluation metrics would you use to assess the performance of your model.

* Accuracy ,Precession And Recall, F1 Score, Confusion Matrix,
* Selection Criteria:= Imbalance , Threshold Sensitivity ,overall performance.

Q11) What considerations would you have for deploying this model in a production environment.

* Scalilbility, Data Input & Integration, Model Monitoring and maintainence, Cost Considerations,Error handling and Loggging, Model Versioning, Deployement Rollout Strategy and so on..