Machine Learning Interview Questions Part 3

1. How do we ensure you're not overfitting with a model?

This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.
There are three main methods to avoid overfitting:
1- Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data.
2- Use cross-validation techniques such as k-folds cross-validation.
3- Use regularization techniques such as LASSO that penalize certain model parameters if they're likely to cause overfitting.

2. Name a few libraries in Python used for Data Analysis and Scientific Computations.

Here is a list of Python libraries mainly used for Data Analysis:

- NumPy
- SciPy
- Pandas
- SciKit
- Matplotlib

3. Which library would we prefer for plotting in Python language?

It depends on the visualization you're trying to achieve. Each of these libraries is used for a specific purpose:

- Matplotlib: Used for basic plotting like bars, pies, lines, scatter plots, etc
- Seaborn: Is built on top of Matplotlib and Pandas to ease data plotting. It is used for statistical visualizations like creating heatmaps or showing the distribution of your data.

4. How are NumPy and SciPy related?

- NumPy is part of SciPy.
- NumPy defines arrays along with some basic numerical functions like indexing, sorting, reshaping, etc.
- SciPy implements computations such as numerical integration, optimization and machine learning using NumPy's functionality.

5. How can we handle duplicate values in a dataset for a variable in Python?

Consider the following Python code:

```
bill_data=pd.read_csv("Marvellous.csv")
bill_data.shape
#Identify duplicates records in the data
Dupes = bill_data.duplicated()
sum(dupes)
#Removing Duplicates
bill_data_uniq = bill_data.drop_duplicates()
```