5 Must-Know Resources & Concepts for Data Scientists

1. Pandas Tips for Efficient Data Analysis

- Use `.loc[]` and `.iloc[]` for efficient row/column selection.
- Avoid loops; use vectorized operations like `df['col'].apply()`.
- Use `groupby()` and `pivot_table()` for summaries.
- Use `df.info()` and `df.describe()` for data inspection.
- Handle missing data with `fillna()`, `dropna()` wisely.

2. Matplotlib vs Seaborn: Visualization Cheat Sheet

- Matplotlib: More control, base plotting.
- Seaborn: Simplifies plotting with themes and grouped data.

Examples:

- `plt.plot()` for line charts (Matplotlib).
- `sns.barplot()` or `sns.heatmap()` for grouped/heatmap visuals.
- Use `%matplotlib inline` in notebooks.

3. Most Used Machine Learning Algorithms

- Linear Regression - Used in predicting continuous values.
- Logistic Regression - For binary classification problems.
- Decision Trees/Random Forest - Classification and regression.
- K-Means Clustering - Unsupervised clustering.
- XGBoost - Efficient boosting algorithm for competitions.

4. Data Preprocessing Techniques

- Scaling: Use `StandardScaler` or `MinMaxScaler` from sklearn.
- Encoding: Convert categories using One-Hot or Label Encoding.
- Feature Engineering: Combine/create new useful features.
- Imputation: Fill missing values using mean/median/model-based methods.

5 Must-Know Resources & Concepts for Data Scientists

5. Best GitHub Repos for Data Scientists

- `awesome-datascience`: Curated list of data science tools.
- `fastai/fastbook`: Deep learning lessons from the FastAI course.
- `jakevdp/PythonDataScienceHandbook`: Python-based data science reference.
- `ageron/handson-ml`: Practical ML with Scikit-Learn, Keras, and TensorFlow.
- `explosion/spacy`: NLP library used widely in the industry.