🧠 Machine Learning Pipeline Cheat Sheet

1. Data Collection

- ✅ What: Gather raw data from files, APIs, sensors, etc.
- 🛠️ Tools: `pandas`, `requests`, `SQL`, `scrapy`, `BeautifulSoup`

```
import pandas as pd
df = pd.read_csv('data.csv')
```

---

2. Data Preprocessing

a. Cleaning

- Handle missing values (`NaN`)
- Fix data types
- Remove duplicates
- Normalize formats

```
df.dropna()
df['price'] = df['price'].astype(float)
```

b. Text Cleanup (NLP-specific)

- Remove punctuation, stopwords, URLs
- Lowercase conversion
- Tokenization

```
from sklearn.feature_extraction.text import ENGLISH_STOP_WORDS
from nltk.tokenize import word_tokenize
```

---

3. Exploratory Data Analysis (EDA)