Essential Data Science Practices with Python

1. Data Collection / Loading

Practice: Load data from CSV, Excel, API, or SQL.

```python
import pandas as pd

# Load CSV
df = pd.read_csv("data.csv")

# Load Excel
# df = pd.read_excel("data.xlsx")

# Display first 5 rows
print(df.head())
```

2. Data Cleaning & Preprocessing

Practice: Handle missing values, remove duplicates, encode categories.

```python
# Check for missing values
print(df.isnull().sum())

# Fill missing values
df['column_name'].fillna(df['column_name'].mean(), inplace=True)

# Drop duplicates
df.drop_duplicates(inplace=True)

# Label encoding
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['category'] = le.fit_transform(df['category'])
```

3. Exploratory Data Analysis (EDA)

Practice: Understand distribution, correlation, outliers.

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Histogram
sns.histplot(df['column_name'])
```

```
# Correlation heatmap
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

# Boxplot for outlier detection
sns.boxplot(x=df['column_name'])
plt.show()
```

4. Feature Engineering

Practice: Create new features, scaling, polynomial features.

```python
# Create a new feature
df['total_amount'] = df['quantity'] * df['price']

# Scaling
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[['col1', 'col2']] = scaler.fit_transform(df[['col1', 'col2']])
```

5. Model Building (ML)

Practice: Train/test split, fit models, evaluate.

```python
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

# Split
X = df.drop('target', axis=1)
y = df['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Model
model = LogisticRegression()
model.fit(X_train, y_train)

# Prediction
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
```

6. Model Evaluation & Tuning

Practice: Use cross-validation, hyperparameter tuning.

```python
from sklearn.model_selection import cross_val_score, GridSearchCV

# Cross-validation
scores = cross_val_score(model, X, y, cv=5)
print("Cross-validation accuracy:", scores.mean())

# Grid Search
param_grid = {'C': [0.1, 1, 10]}
grid = GridSearchCV(LogisticRegression(), param_grid, cv=3)
grid.fit(X_train, y_train)
print("Best parameters:", grid.best_params_)
```

7. Model Deployment (Basic)

Practice: Save and load models.

```python
import joblib

# Save model
joblib.dump(model, 'logistic_model.pkl')

# Load model
loaded_model = joblib.load('logistic_model.pkl')
print(loaded_model.predict(X_test[:5]))
```

8. Practice Projects (Suggestions)

Try these to apply what you've learned:

- Titanic Survival Prediction (Kaggle)
- House Price Prediction
- Customer Churn Analysis
- Sentiment Analysis on Tweets
- Sales Forecasting using Time Series