Cheat Sheet for comprehensive Open Data Science Certification (ODSC)

Python Basics

Data Types

- **Numeric**: `int`, `float`, `complex`
- **Sequence**: `list`, `tuple`, `range`
- **Text**: `str`
- **Mapping**: `dict`
- **Set**: `set`, `frozenset`
- **Boolean**: `bool`

Variables and Assignment

- **Assignment**: `x = 10`
- **Multiple Assignment**: `x, y, z = 1, 2, 3`
- **Swapping**: `x, y = y, x`

Control Structures

- **If-Else**:

```
if condition:
    # code
elif another_condition:
    # code
else:
    # code
```

- **For Loop**:

```
for i in range(5):
    # code
```

- **While Loop**:

```
while condition:
    # code
```

Data Manipulation with Pandas

Creating DataFrames

- **From Dictionary**:

```
import pandas as pd
data = {'Name': ['Alice', 'Bob'], 'Age': [25, 30]}
df = pd.DataFrame(data)
```

- **From CSV**:

```
df = pd.read_csv('file.csv')
```

Basic Operations

- **Selecting Columns**: `df['Name']`
- **Filtering Rows**: `df[df['Age'] > 25]`
- **Adding Columns**: `df['NewCol'] = df['Age'] + 5`
- **Grouping**: `df.groupby('Name').sum()`

Data Visualization with Matplotlib and Seaborn

Matplotlib Basics

- **Line Plot**:

```
import matplotlib.pyplot as plt
plt.plot([1, 2, 3], [4, 5, 6])
plt.show()
```

- **Scatter Plot**:

```
plt.scatter([1, 2, 3], [4, 5, 6])
plt.show()
```

Seaborn Basics

- **Histogram**:

```
import seaborn as sns
sns.histplot(df['Age'])
plt.show()
```

- **Pair Plot**:

```
sns.pairplot(df)
plt.show()
```

Machine Learning with Scikit-Learn

Model Training

- **Importing**:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

- **Splitting Data**:

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2)
```

- **Training Model**:

```
model = LinearRegression()
model.fit(X_train, y_train)
```

Model Evaluation

- **Predicting**: `y_pred = model.predict(X_test)`

- **Metrics**:

```
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, y_pred)
```

Advanced Topics

Feature Engineering

- **One-Hot Encoding**:

```
from sklearn.preprocessing import OneHotEncoder
encoder = OneHotEncoder()
encoded = encoder.fit_transform(df[['Category']])
```

- **Standard Scaling**:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled = scaler.fit_transform(df[['NumericCol']])
```

Cross-Validation

- **K-Fold**:

```
from sklearn.model_selection import KFold
kf = KFold(n_splits=5)
for train_index, test_index in kf.split(X):
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]
```

Tips and Tricks

Pythonic Code

- **List Comprehension**: `[x**2 for x in range(10)]`
- **Lambda Functions**: `square = lambda x: x**2`

Debugging

- **Print Statements**: `print(variable)`
- **Logging**:

```
import logging
logging.basicConfig(level=logging.DEBUG)
logging.debug('Debug message')
```

Performance Optimization

- **Profiling**:

```
import cProfile
cProfile.run('function_to_profile()')
```

- **Vectorization**: Use NumPy for operations on arrays

Resources

Documentation

- **Pandas**: pandas.pydata.org
- **Scikit-Learn**: scikit-learn.org
- **Matplotlib**: matplotlib.org
- **Seaborn**: seaborn.pydata.org

Communities

- **Stack Overflow**: stackoverflow.com
- **Kaggle**: [kaggle.com](https://www.kaggle.com)

Conclusion

This cheat sheet provides a comprehensive overview of essential concepts and tools for the Open Data Science Certification. Use these tips and tricks to enhance your data science skills and efficiently solve real-world problems.

By Ahmed Baheeg Khorshid

ver 1.0