Data Wrangling II - Full Cheatsheet

Full Code with Line-by-Line Comments

```
# Step 1: Import Required Libraries
import pandas as pd      # pandas for data manipulation
import numpy as np       # numpy for numerical operations
import matplotlib.pyplot as plt  # plotting
import seaborn as sns    # statistical plots

# Step 2: Load the Dataset
df = pd.read_csv('StudentsPerformance.csv')  # load CSV
df  # display data

# Step 3: Check for Missing Values
df.isnull().sum()  # count missing per column

# Step 4: Summary Statistics & Info
df.describe()  # stats for numeric columns
df.info()      # types and non-null counts

# Step 5: Fill Missing with Median
median = df['math score'].median()  # compute median
df['math score'] = df['math score'].fillna(median)  # fill NaNs

# Step 6: Detect Outliers (Boxplot)
df.boxplot()  # visualize outliers

# Step 7: Remove Outliers via IQR
Q1 = df['math score'].quantile(0.25)
Q3 = df['math score'].quantile(0.75)
IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR
newdf = df[(df['math score'] >= lower) & (df['math score'] <= upper)]
newdf.boxplot()

# Step 8: Skewness Before Transformation
sns.histplot(df['math score'], kde=True)
plt.title('Before Transformation')
plt.show()
print("Skewness:", df['math score'].skew())

# Step 9: Log Transform
df['math score log'] = np.log1p(df['math score'])

# Step 10: Skewness After
sns.histplot(df['math score log'], kde=True)
plt.title('After Log Transformation')
plt.show()
print("Skewness:", df['math score log'].skew())
```

```
# Step 11: Compare
print(df[['math score', 'math score log']])
```

Theory & Key Concepts

Missing Values:

Use dropna() or fillna() (mean/median/mode) to handle missing data.

IQR Method:

IQR = Q3 - Q1; outliers outside [Q1-1.5*IQR, Q3+1.5*IQR].

Skewness:

Measure of asymmetry; positive (>0), negative (<0), zero symmetric.

Log Transformation:

np.log1p(x) = log(1+x); reduces skew, handles zeros.