Advanced Pandas Cheat Sheet – Week 9

Note: please study all the concepts before solving the questions

1. Data Exploration

```
df.head()          # First 5 rows

df.tail()          # Last 5 rows

df.info()          # Summary of data types and nulls

df.describe()      # Summary statistics for numerical columns

df.shape           # Rows and columns count

df.columns         # Column names

df.dtypes          # Data types of each column
```

2. Grouping and Aggregation

```
df.groupby('Department')['Salary'].mean()

df.groupby(['Department', 'Gender'])['Salary'].mean()

df.groupby('JoiningYear').size()
```

3. Pivot Tables

```
pd.pivot_table(df, values='PerformanceScore', index='Gender', columns='Department', aggfunc='mean')

pd.pivot_table(df, values='RemoteWork', index='Department', columns='Gender', aggfunc='mean')
```

4. Crosstab

```
pd.crosstab(df['RemoteWork'], df['LeftCompany'])
```

5. Ranking, Sorting, and Filtering

```
df.sort_values(by='Salary', ascending=False)

df['SalaryRank'] = df.groupby('Department')['Salary'].rank(ascending=False)

df[df['Age'] > 50]

df.query("Department == 'Marketing' and Salary > 100000 and PerformanceScore > 4")
```

6. MultiIndex GroupBy

```
df.groupby(['Department', 'Gender'])['WorkHoursPerWeek'].mean()
```

7. Correlation & Statistics

```
df[['Age', 'Salary', 'WorkHoursPerWeek']].corr()

df.groupby('Department')['Salary'].std()
```

8. New Columns & Calculations

```
df['Tenure'] = 2025 - df['JoiningYear']

df['ExpectedSalary'] = df['Salary'] * (1.05 ** df['Tenure'])
```

9. Binning and Categorizing

```
bins = [20, 30, 40, 50, 60]

labels = ['20s', '30s', '40s', '50s']

df['AgeGroup'] = pd.cut(df['Age'], bins=bins, labels=labels)

df.groupby('AgeGroup')['Salary'].mean()
```

10. Boolean Masks and Queries

```
mask = df['PerformanceScore'] <= 2

df[mask]
```

11. Cumulative & Aggregated Operations

```
df.groupby('Department')['Salary'].cumsum()
```

12. Outlier Detection (IQR Method)

```
Q1 = df['Salary'].quantile(0.25)

Q3 = df['Salary'].quantile(0.75)

IQR = Q3 - Q1

outliers = df[(df['Salary'] < Q1 - 1.5 * IQR) | (df['Salary'] > Q3 + 1.5 * IQR)]
```

13. Finding Max in Group

```
df.loc[df.groupby('Department')['PerformanceScore'].idxmax()]
```

14. Visualizations (Optional)

```
import matplotlib.pyplot as plt

df[df['Gender'] == 'Male']['PerformanceScore'].hist()

plt.title("Performance Score for Male Employees")

plt.show()
```

15. Exporting and Importing

```
df.to_csv('filename.csv', index=False)

pd.read_csv('filename.csv')
```