# **40 Python Data Science & ML Practice Questions**

**Dataset:** [Human Resources Dataset (Kaggle)](https://www.kaggle.com/datasets/rhuebner/human-resources-data-set)

## **Part A: Basic EDA (Exploratory Data Analysis)**

1. Load the dataset into a Pandas DataFrame and display the first 5 rows.  
    [W3Schools – Pandas Intro](https://www.w3schools.com/python/pandas/pandas_intro.asp)
2. Check the shape of the dataset (rows, columns).  
    [W3Schools – Pandas DataFrame](https://www.w3schools.com/python/pandas/pandas_dataframe.asp)
3. Display the column names and their data types.  
    [Pandas dtypes Documentation](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dtypes.html)
4. Find the number of unique values in each column.  
    [Pandas nunique()](https://www.geeksforgeeks.org/python-pandas-dataframe-nunique/)
5. Identify missing values in the dataset and list columns with nulls.  
    [Pandas isnull()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.isnull.html)
6. Describe the numerical columns (mean, median, std).  
    [Pandas describe()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html)
7. Find the distribution of employee salaries.  
    [Matplotlib Histograms](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html)
8. What is the average age of employees at the company? (Use DOB column).  
    [Convert Dates in Pandas](https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html)
9. How many employees are still employed vs terminated?  
    [Pandas value\_counts()](https://pandas.pydata.org/docs/reference/api/pandas.Series.value_counts.html)
10. Which departments have the most employees?  
     [Seaborn Countplot](https://seaborn.pydata.org/generated/seaborn.countplot.html)

## **Part B: Business Analysis**

1. What is the average salary per department?  
    [Pandas groupby()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html)
2. Find the distribution of employment status (Active, Voluntarily Terminated, etc.).  
    [Seaborn Pie Charts with Matplotlib](https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html)
3. Compare salary levels between Male and Female employees.  
    [Seaborn Boxplot](https://seaborn.pydata.org/generated/seaborn.boxplot.html)
4. Which recruitment source brings in the most employees?  
    [Pandas value\_counts()](https://pandas.pydata.org/docs/reference/api/pandas.Series.value_counts.html)
5. What percentage of employees attended a Diversity Job Fair?  
    [Pandas mean() on Boolean](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.mean.html)
6. Compare engagement survey scores across different departments.  
    [Seaborn Barplot](https://seaborn.pydata.org/generated/seaborn.barplot.html)
7. Which race has the highest average salary?  
    [Pandas groupby() with mean](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html)
8. What is the relationship between number of projects (SpecialProjectsCount) and salary?  
    [Seaborn Scatterplot](https://seaborn.pydata.org/generated/seaborn.scatterplot.html)
9. Do married employees earn more on average than single employees?  
    [Seaborn Barplot](https://seaborn.pydata.org/generated/seaborn.barplot.html)
10. Which managers have the largest teams?  
     [Pandas groupby size](https://pandas.pydata.org/docs/reference/api/pandas.core.groupby.DataFrameGroupBy.size.html)

## **Part C: Data Visualization**

1. Plot the salary distribution using histograms.  
    [Matplotlib Hist](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html)
2. Show the count of employees by department.  
    [Seaborn Countplot](https://seaborn.pydata.org/generated/seaborn.countplot.html)
3. Compare average satisfaction score by department.  
    [Seaborn Barplot](https://seaborn.pydata.org/generated/seaborn.barplot.html)
4. Visualize employee terminations over time.  
    [Pandas Datetime Plot](https://pandas.pydata.org/docs/reference/api/pandas.Series.dt.html)
5. Plot average salary by gender using a boxplot.  
    [Seaborn Boxplot](https://seaborn.pydata.org/generated/seaborn.boxplot.html)
6. Visualize performance scores vs salary.  
    [Seaborn Stripplot](https://seaborn.pydata.org/generated/seaborn.stripplot.html)
7. Create a heatmap of correlations between numeric variables.  
    [Seaborn Heatmap](https://seaborn.pydata.org/generated/seaborn.heatmap.html)
8. Plot engagement survey score vs satisfaction score.  
    [Seaborn Scatterplot](https://seaborn.pydata.org/generated/seaborn.scatterplot.html)
9. Show a stacked bar chart of employee status across departments.  
    [Matplotlib Stacked Bar](https://matplotlib.org/stable/gallery/lines_bars_and_markers/bar_stacked.html)
10. Plot absenteeism (Absences) distribution among employees.  
     [Seaborn Histplot](https://seaborn.pydata.org/generated/seaborn.histplot.html)

## **Part D: PCA (Dimensionality Reduction)**

1. Standardize numerical features before applying PCA.  
    [Scikit-learn StandardScaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html)
2. Perform PCA on the dataset and explain the first 2 components.  
    [Scikit-learn PCA](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html)
3. Plot the explained variance ratio for the PCA components.  
    [Matplotlib Scree Plot](https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html)
4. Reduce the dataset to 2 dimensions with PCA and plot employees colored by department.  
    [Seaborn Scatterplot](https://seaborn.pydata.org/generated/seaborn.scatterplot.html)
5. Which variables contribute most to the first principal component?  
    [PCA Components](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html)
6. Apply PCA to engagement + satisfaction + absences. Do these condense into 1 dimension?  
    [PCA in Practice – TowardsDataScience](https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad)
7. Visualize employees in PCA-reduced 2D space grouped by PerformanceScore.  
    [Seaborn Hue in Scatterplot](https://seaborn.pydata.org/generated/seaborn.scatterplot.html)
8. Compare clustering before and after PCA using k-means on numeric data.  
    [Scikit-learn KMeans](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)
9. Plot loadings of salary, absences, and engagement on the first 2 PCs.  
    [PCA Biplot Guide](https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d)
10. Discuss when PCA is appropriate in HR analytics (e.g., reducing survey dimensions).  
     [Analytics Vidhya – PCA Use Cases](https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/)