50 EDA interview questions

Dataframe Basics

How do you create a DataFrame from a dictionary?

import pandas as pd data = {'name':['A' , 'B'], 'age':[2, 3]} df = pd.DataFrame(data)

How to check the shape, size, and data types of a DataFrame?

df.shape, df.size, df.dtypes

How do you get the first and last 5 rows?

df.head(), df.tail()

How to rename columns in a DataFrame?

df.rename(columns={'old_name': 'new_name'}, inplace=True)

How to reset and set the index of a DataFrame?

df.reset_index(drop=True, inplace=True) df.set_index('column_name' , inplace=True)

Missing and Duplicate data

How to detect and count missing values?

df.isnull().sum()

How to fill missing values with mean/median/mode?

df['col'].fillna(df['col'].mean(), inplace=True)

How to drop rows or columns with missing values?

df.dropna(axis=0), df.dropna(axis=1)

How to detect and remove duplicates?

df[df.duplicated()] df.drop_duplicates(inplace=True)

How to replace values in a DataFrame?

df.replace({'old': 'new'}, inplace=True)

Filtering and conditions

How to filter rows based on a condition?

df[df['age'] > 30]

How to filter rows using multiple conditions?

df[(df['age'] > 30) & (df['gender'] == 'Male')]

How to query rows using query()?

df.query("age > 30 and gender == 'Male'")

How to use isin() to filter values?

df[df['country'].isin(['India' , 'USA'])]

How to apply a custom function row-wise?

df.apply(lambda row: row['a'] + row['b'], axis=1)

Aggregations and grouping

How to detect and count missing values?

df.isnull().sum()

How to perform multiple aggregations?

df.groupby('region').agg({'sales': ['sum' , 'mean']})

How to get group size and count?

df.groupby('category').size() df.groupby('category')['item'].count()

How to apply transformations to groups?

df.groupby('region') ['sales'].transform('mean')

How to rank values within groups?

df['rank'] = df.groupby('region') ['sales'].rank(ascending=False)

Merging and reshaping

How to merge two DataFrames (like SQL JOIN)?

pd.merge(df1, df2, on='id' , how='left')

How to concatenate DataFrames?

pd.concat([df1, df2], axis=0) # vertical pd.concat([df1, df2], axis=1) # horizontal

How to pivot data?

df.pivot_table(values='sales' , index='region' , columns='month' , aggfunc='sum')

How to unpivot (melt) data?

pd.melt(df, id_vars=['id'], value_vars= ['score1' , 'score2'])

How to join based on index?

df1.join(df2, how='inner')

Datetime operations

How to convert a column to datetime?

df['date'] = pd.to_datetime(df['date'])

How to extract year, month, day?

df['year'] = df['date'].dt.year

How to filter rows based on date range?

df[(df['date'] >= '2023-01-01') & (df['date'] <= '2023-12-31')]

How to create a new column for day of week?

df['day_of_week'] = df['date'].dt.day_name()

How to set datetime column as index?

df.set_index('date' , inplace=True)

Advanced Column Operations

How to create new columns based on other columns?

df['total'] = df['price'] * df['quantity']

How to use np.where() for conditional columns?

import numpy as np df['grade'] = np.where(df['score'] > 90, 'A' , 'B')

How to use map() or replace() for value mapping?

df['gender'] = df['gender'].map({'M': 'Male' , 'F': 'Female'})

How to apply string methods to a column?

df['name'] = df['name'].str.lower()

How to split a column into multiple columns?

df[['first' , 'last']] = df['full_name'].str.split(' ' , expand=True)

Statistical & window Functions

How to calculate correlation between features?

df.corr()

How to calculate cumulative sum and product?

df['cumsum'] = df['sales'].cumsum() df['cumprod'] = df['returns'].cumprod()

How to calculate rolling mean?

df['rolling_avg'] = df['sales'].rolling(window=7).mean()

How to use diff() and pct_change()?

df['diff'] = df['sales'].diff() df['pct_change'] = df['sales'].pct_change()

How to detect outliers using IQR?

Q1 = df['value'].quantile(0.25) Q3 = df['value'].quantile(0.75) IQR = Q3 - Q1 outliers = df[(df['value'] < Q1 - 1.5*IQR) | (df['value']

Exploratory Data Analysis

How to get summary statistics for numeric columns?

df.describe()

How to get value counts for categorical column?

df['category'].value_counts()

How to find unique values and their count?

df['column'].unique(), df['column'].nunique()

How to identify skewness and kurtosis?

df['column'].skew(), df['column'].kurt()

How to use .info() and .memory_usage()?

df.info() df.memory_usage(deep=True)

Visualization

How to plot histogram and boxplot?

df['sales'].hist() df.boxplot(column='sales')

How to create a bar plot?

df['category'].value_counts().plot(kind=' bar')

How to plot a time series?

df.set_index('date')['sales'].plot()

How to use seaborn for correlation heatmap?

import seaborn as sns sns.heatmap(df.corr(), annot=True)

How to use matplotlib for multiple plots?

import matplotlib.pyplot as plt plt.figure(figsize=(10,5)) plt.plot(df['date'], df['sales']) plt.show()