EDA CHEETSHEET

**Import Libraries**

* import numpy as np
* import pandas as pd
* import seaborn as sns
* import matplotlib.pyplot as plt
* import scipy.stats as stats
* from sklearn.preprocessing import StandardScaler, MinMaxScaler, LabelEncoder

**Load Data**

* df = pd.read\_csv('your\_data.csv')
* df.head()
* df.tail()
* df.sample(5)

**Basic Info**

* df.shape
* df.columns
* df.dtypes
* df.info()
* df.describe()
* df.isnull().sum()
* df.duplicated().sum()

**Missing Values**

* df.isnull().sum()/len(df)\*100
* df.fillna(value)
* df.dropna()

**Data Cleaning**

* df['column'].str.strip()
* df['column'] = df['column'].replace('-', np.nan)
* df.drop\_duplicates(inplace=True)
* df.rename(columns={'old': 'new'}, inplace=True)

**Univariate Analysis**

* df['column'].value\_counts()
* df['column'].value\_counts(normalize=True)
* df['column'].unique()
* df['column'].nunique()

**Bivariate / Multivariate**

* sns.scatterplot(x='feature1', y='feature2', data=df)
* sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
* sns.pairplot(df, diag\_kind='kde')
* sns.barplot(x='category', y='numeric', data=df)

**Outliers**

* sns.boxplot(x=df['column'])
* Q1 = df['column'].quantile(0.25)
* Q3 = df['column'].quantile(0.75)
* IQR = Q3 - Q1
* lower = Q1 - 1.5 \* IQR
* upper = Q3 + 1.5 \* IQR
* df\_outliers = df[(df['column'] < lower) | (df['column'] > upper)]

**Skewness & Transformation**

* df['column'].skew()
* df['column'].kurt()
* df['new\_column'] = np.log1p(df['column'])
* df['new\_column'] = np.sqrt(df['column'])

**Categorical Encoding**

* pd.get\_dummies(df['category\_column'], drop\_first=True)
* le = LabelEncoder()
* df['encoded'] = le.fit\_transform(df['category\_column'])

**Feature Scaling**

* scaler = StandardScaler()
* df\_scaled = scaler.fit\_transform(df[['col1', 'col2']])
* scaler = MinMaxScaler()
* df\_scaled = scaler.fit\_transform(df[['col1', 'col2']])

**Time Series EDA**

* df['date'] = pd.to\_datetime(df['date'])
* df.set\_index('date', inplace=True)
* df.resample('M').mean().plot()

**Correlation**

* df.corr()
* sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

**Groupby & Aggregation**

* df.groupby('category\_column')['numeric\_column'].mean()
* df.pivot\_table(values='numeric', index='category1', columns='category2', aggfunc='mean')
* Save / Export
* df.to\_csv('cleaned\_data.csv', index=False)

**Other Useful Pandas Commands**

* df.sort\_values(by='column', ascending=False)
* df['new\_col'] = df['col1'] / df['col2']
* df['year'] = df['date'].dt.year
* df['month'] = df['date'].dt.month