Week 9 Activity: Exploratory Analysis of EC2 & S3 Usage

Objective

Perform comparative exploratory data analysis (EDA) on EC2 and S3 datasets to understand:

- Cost distribution across AWS regions.
- Usage efficiency (CPU vs cost for EC2, Storage vs cost for S3).
- Optimization opportunities.

EC2 Dataset

Column	Description
InstanceId	Unique ID of EC2 instance
InstanceType	Type (e.g., t3.small, m5.large)
Region	AWS region
State	running/stopped/terminated
CPUUtilization	Avg. CPU usage (%)
MemoryUtilization	Avg. memory usage (%)
NetworkIn_Bps	Avg. incoming network (bytes/sec)
NetworkOut_Bps	Avg. outgoing network (bytes/sec)
CostPerHourUSD	On-demand cost per hour
Tags	Key=value pairs for Owner, Environment
LaunchTime	Instance launch datetime

S3 Bucket Dataset

Column	Description
BucketName	Name of S3 bucket
Region	AWS region
StorageClass	STANDARD, STANDARD_IA, GLACIER
ObjectCount	Total objects in bucket
TotalSizeGB	Total bucket size in GB

MonthlyCostUSD	Estimated monthly storage cost
VersioningEnabled	True/False
Encryption	e.g., AES256, None
CreatedDate	Bucket creation date
Tags	Owner and Purpose tags

Tasks

- Load both datasets into pandas.
- Display info, shape, and summary statistics.
- Handle missing data and detect outliers.

- Visualize:
  - EC2: Histogram of CPU utilization.
  - EC2: CPU vs Cost scatter.
  - S3: Bar chart of total storage by region.
  - S3: Cost vs Storage scatter.

- Identify:
  - Top 5 most expensive EC2 instances.
  - Top 5 largest S3 buckets.

- Compute:
  - Average EC2 cost per region.
  - Total S3 storage per region.

- Suggest **two optimization actions** for EC2 and S3 based on insights.
- Build an interactive Streamlit dashboard combining both analyses.

Demo and Submissions: Show the demo and submit your code in zip