Project Report: Cloud Cost Anomaly Detection & Forecasting Using Data Analytics

1. Background & Objective

Cloud infrastructure costs are one of the largest expenses for IT organizations. Unexpected spikes, anomalies, or inefficient usage can lead to budget overruns, SLA penalties, and reduced profitability.

This project simulates a real-world IT analytics scenario where we use data science + business intelligence to:

- Detect unusual patterns in daily cloud usage costs.
- Forecast future spend to enable proactive budget planning.
- Provide actionable insights through an interactive Power BI dashboard.

Business Objective:

- Detect anomalies in daily spend at both global and service levels.
- Forecast 30-day (operational) and 6-month (strategic) costs.
- Empower IT managers with insights to optimize cloud usage and avoid financial risk.

2. Data Description

- Synthetic dataset generated with ~1 million rows to simulate cloud usage and cost logs.

- Fields included:

    - timestamp → event time
    - service_tag → cloud service (RDS, EC2, S3, etc.)
    - team → owning team (analytics, ML, payments, etc.)
    - region → region of deployment
    - usage & usage_unit → amount of compute/storage/network used
    - cost_usd → corresponding cost in USD (with injected noise, spikes, and missing values)
    - owner, source_ip → metadata

Dataset properties:

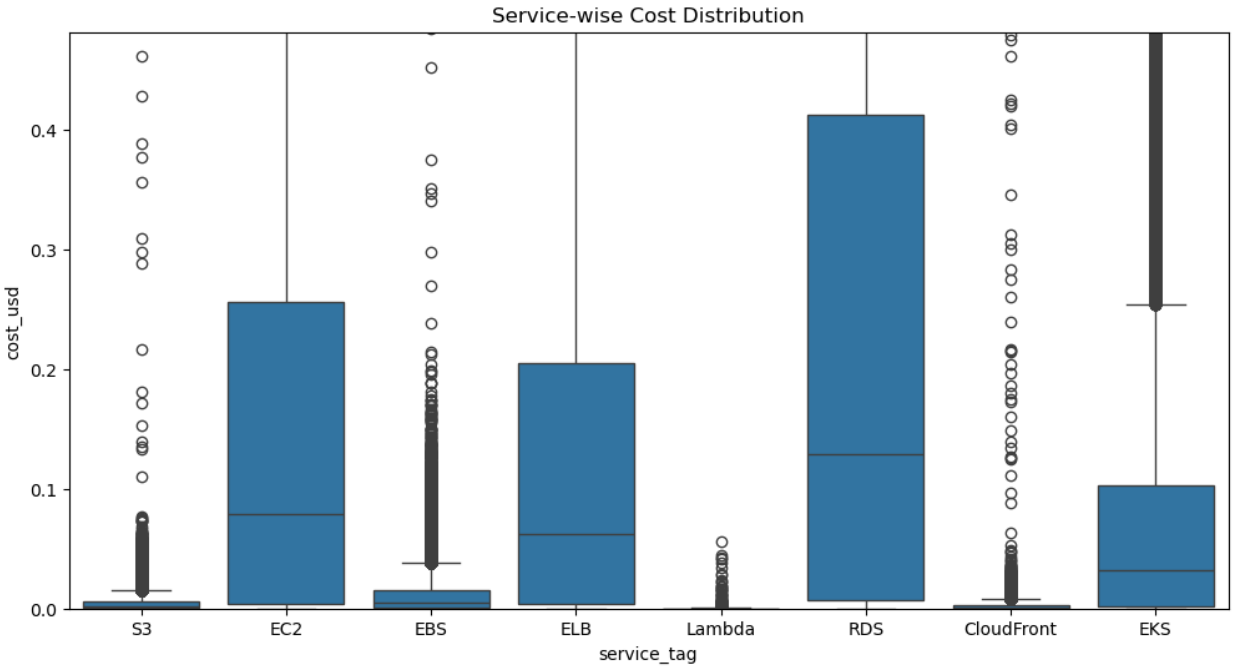
- 12 months of daily data.
- Imperfections included: missing values, duplicates, inconsistent service tags, currency conversions, outliers.
- Exported in .csv and .parquet formats.

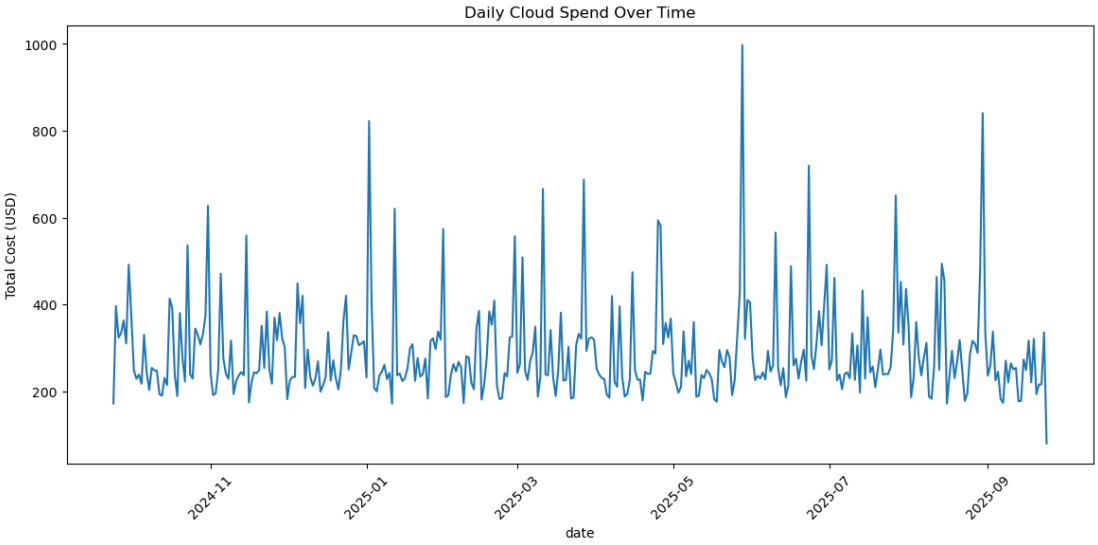
3. Methodology – PACE Framework
Phase 1: Plan

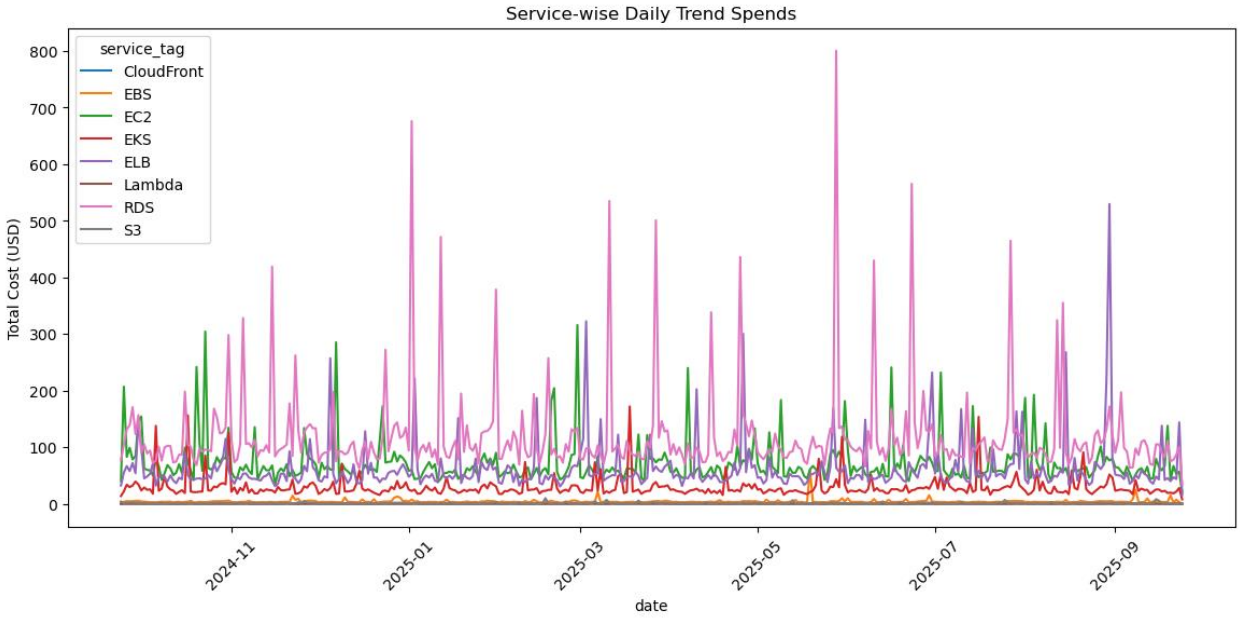
- Define the business problem: cost anomalies + forecasting.
- Identify methods: statistical + ML anomaly detection, Prophet forecasting.
- Design outputs: Jupyter notebooks + Power BI dashboard.

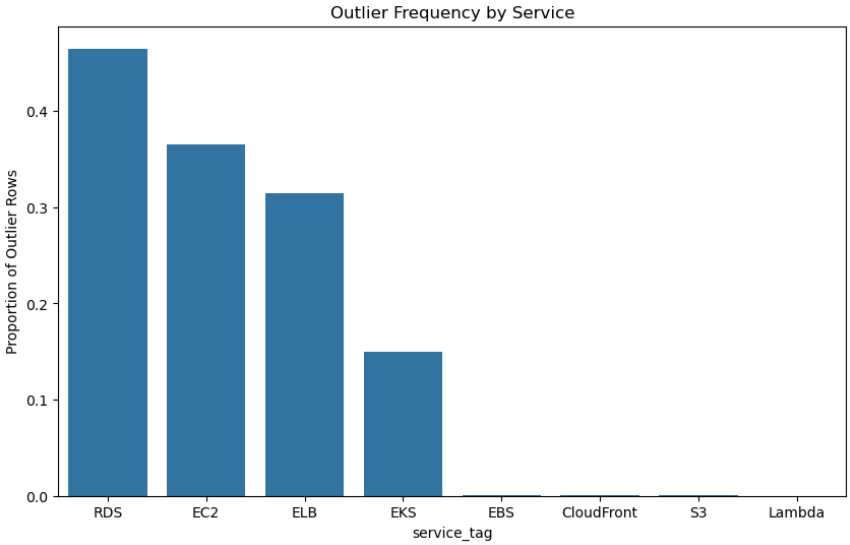
Phase 2: Analyze (EDA)

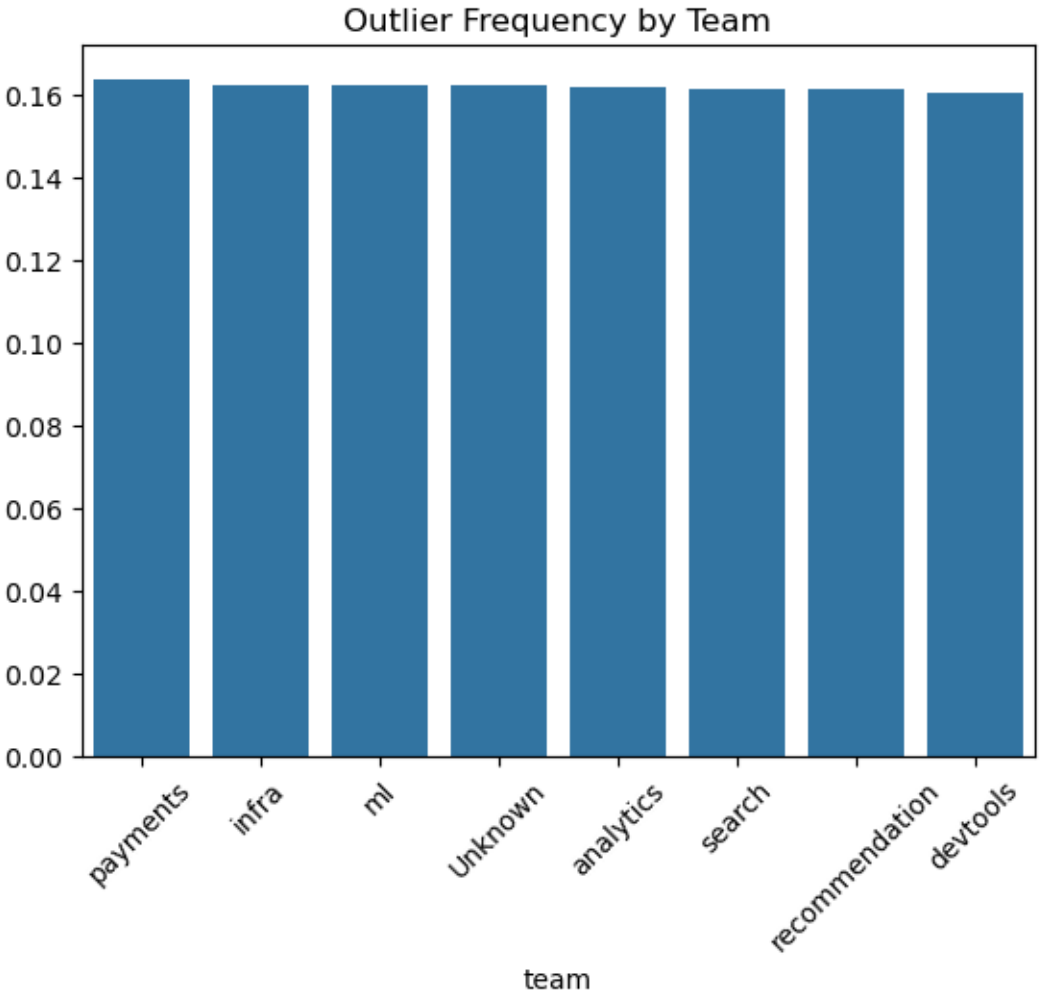
- Data cleaning: handled missing values, duplicates, invalid timestamps, and currency conversion.
- Outlier detection: identified ~16% anomalies in raw usage logs.
- Exploratory analysis:
    - RDS & ELB showed major spikes.
    - Teams contributed ~equally to anomalies, with Payments slightly higher.
    - Daily costs mostly between 200–400 USD, with extreme spikes up to 1000 USD.
- Visualized trends: daily costs, service-level patterns, anomalies by team/region.

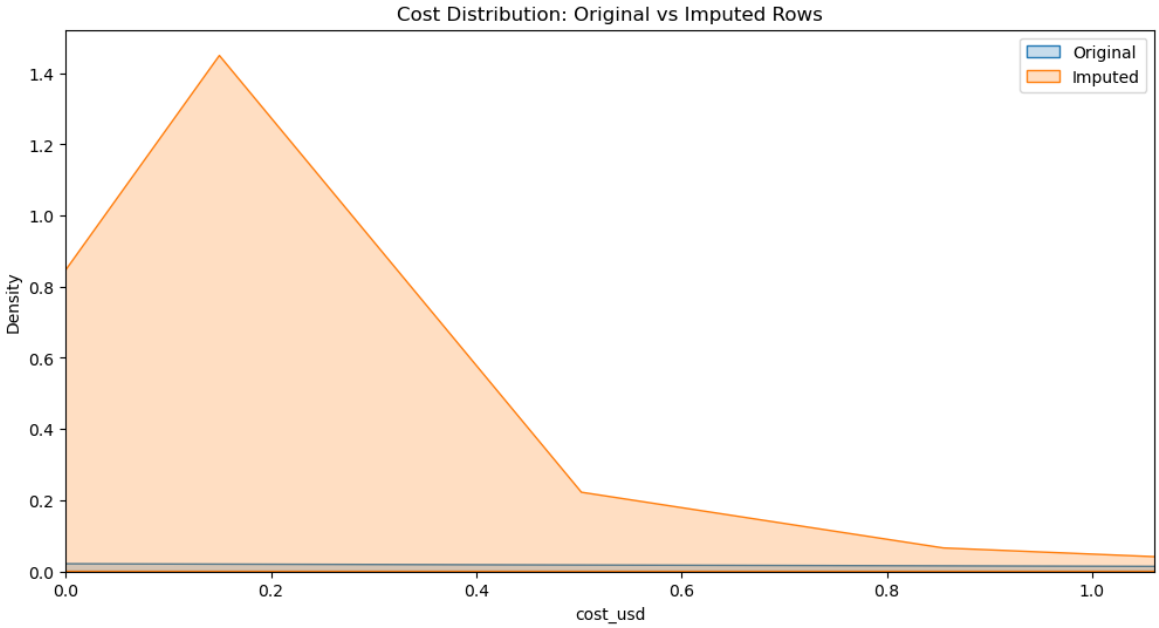


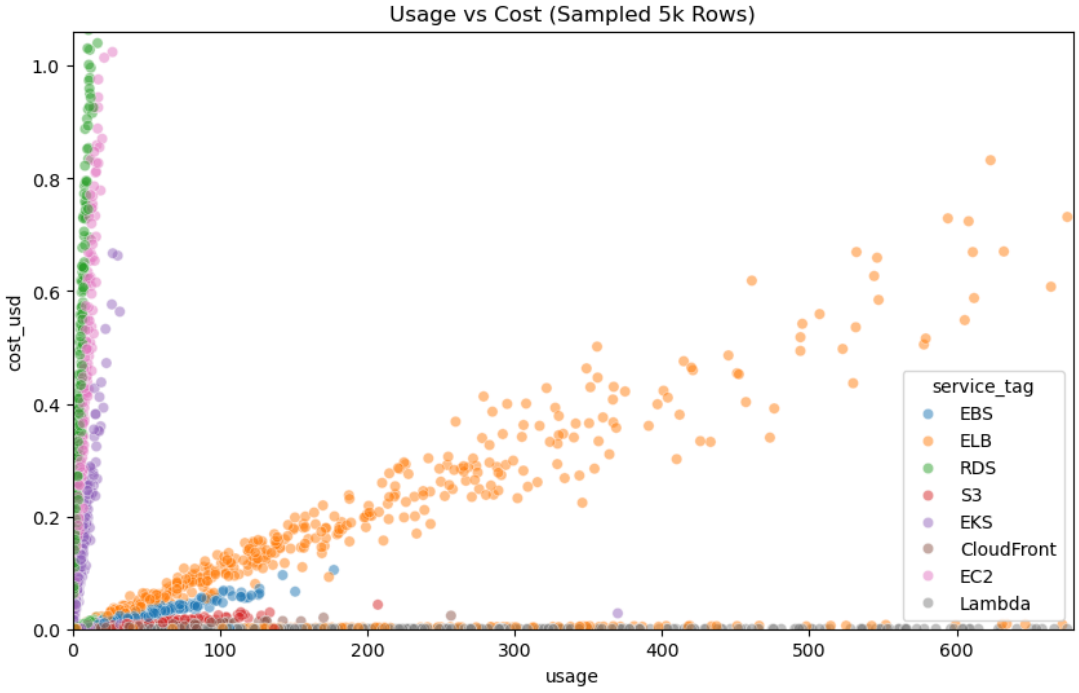

Daily Cloud Spend Over Time
Total Cost (USD)
1000
800
600
400
200
2024-11
2025-01
2025-03
2025-05
2025-07
2025-09
date

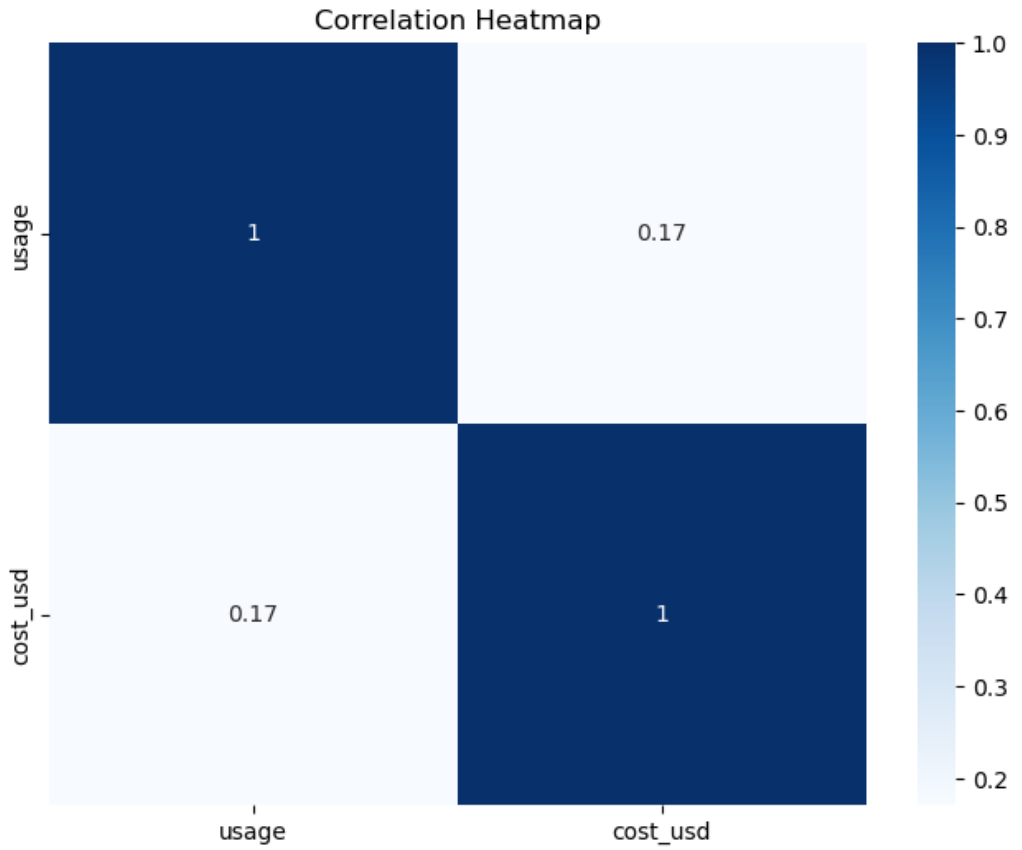

Service-wise Daily Trend Spends
service_tag
CloudFront
EBS
EC2
EKS
ELB
Lambda
RDS
S3
Total Cost (USD)
800
700
600
500
400
300
200
100
0
2024-11
2025-01
2025-03
2025-05
2025-07
2025-09
date


Outlier Frequency by Service
Proportion of Outlier Rows
0.0
0.1
0.2
0.3
0.4
RDS
EC2
ELB
EKS
EBS
CloudFront
S3
Lambda
service_tag


Outlier Frequency by Team
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
payments
infra
ml
Unknown
analytics
search
recommendation
devtools
team

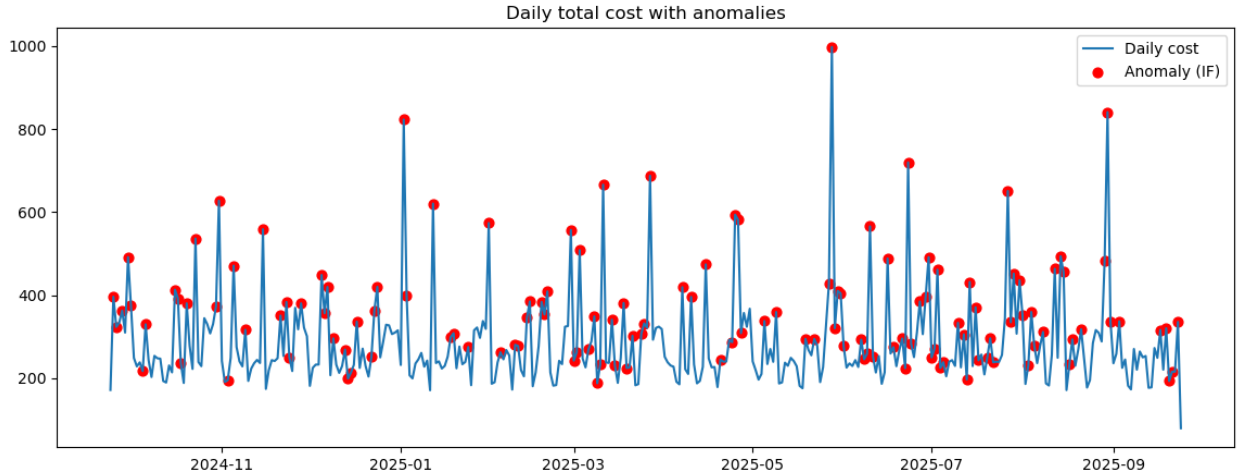

Cost Distribution: Original vs Imputed Rows
Original
Imputed
Density
cost_usd
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

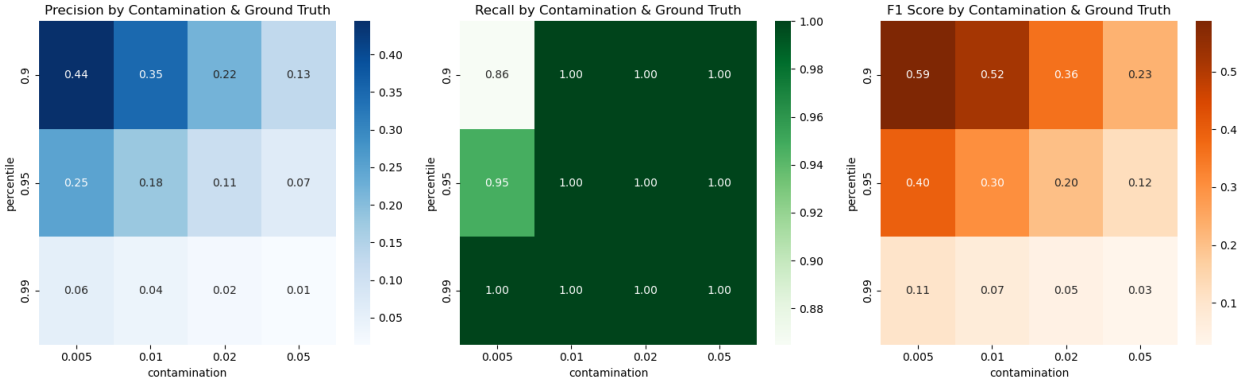

Usage vs Cost (Sampled 5k Rows)
cost_usd
usage
service_tag
EBS
ELB
RDS
S3
EKS
CloudFront
EC2
Lambda
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0

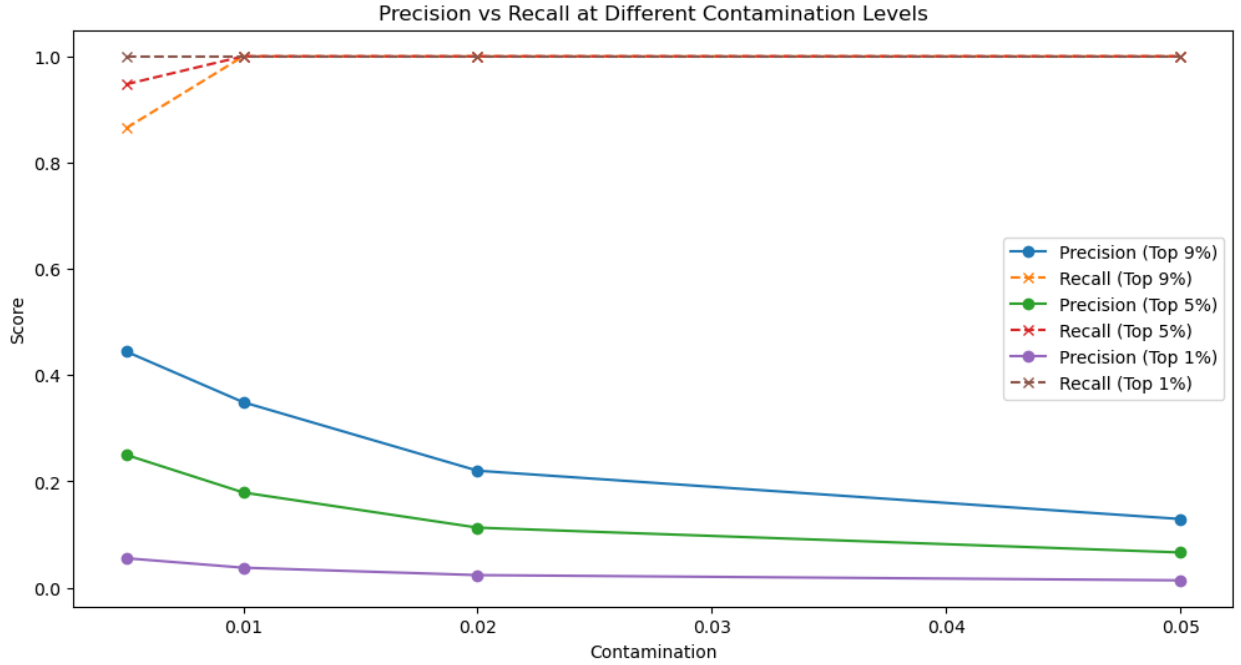


Phase 3: Construct (Modeling)
Anomaly Detection

- Statistical Baseline: Z-score using rolling mean & std.
- Machine Learning: Isolation Forest with features:
  - lag costs, rolling mean/std, delta cost, pct change, cost-per-usage, weekend flag.
- Model Tuning: Tested contamination levels (0.5–5%) vs proxy ground truths (top 1%, 5%, 10% spend days).
  - Best trade-off: 0.5% contamination, top 10% ground truth → Precision 44%, Recall 87%, F1 ≈ 0.59.
- Deliverable: service_level_anomalies.csv (service-level flagged anomalies).



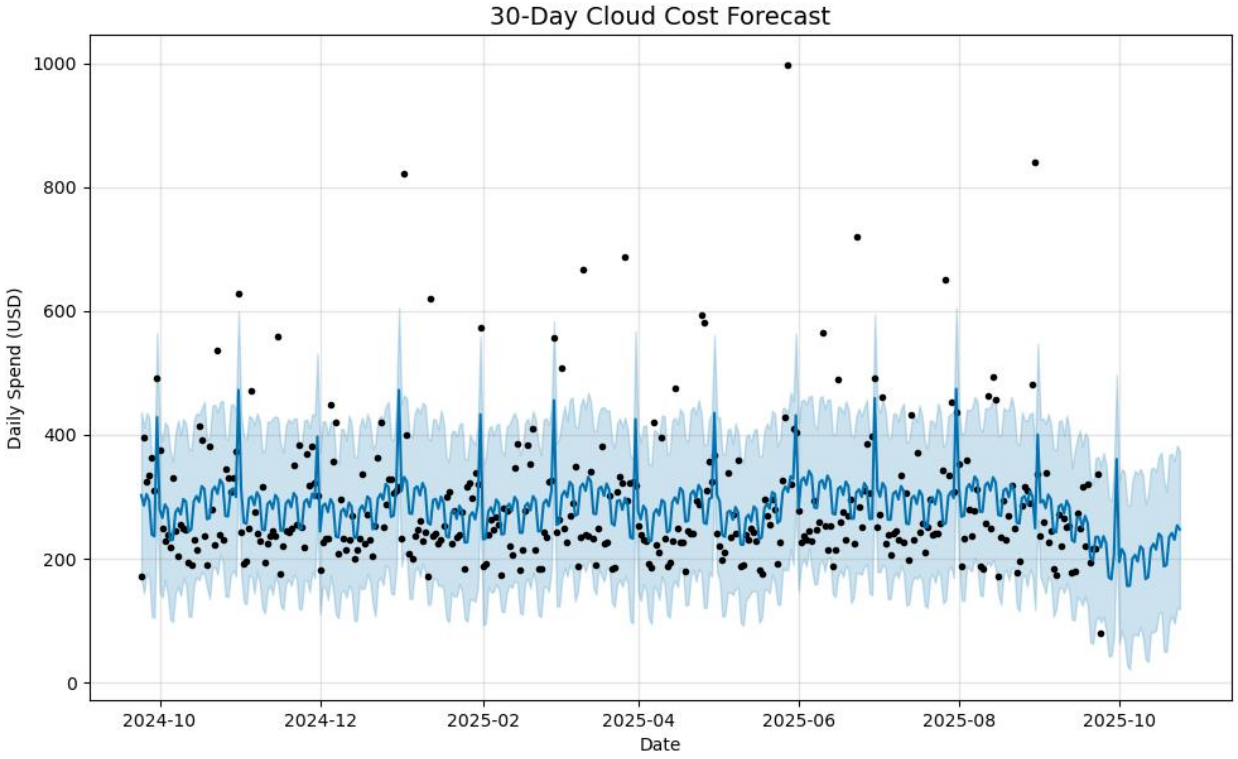


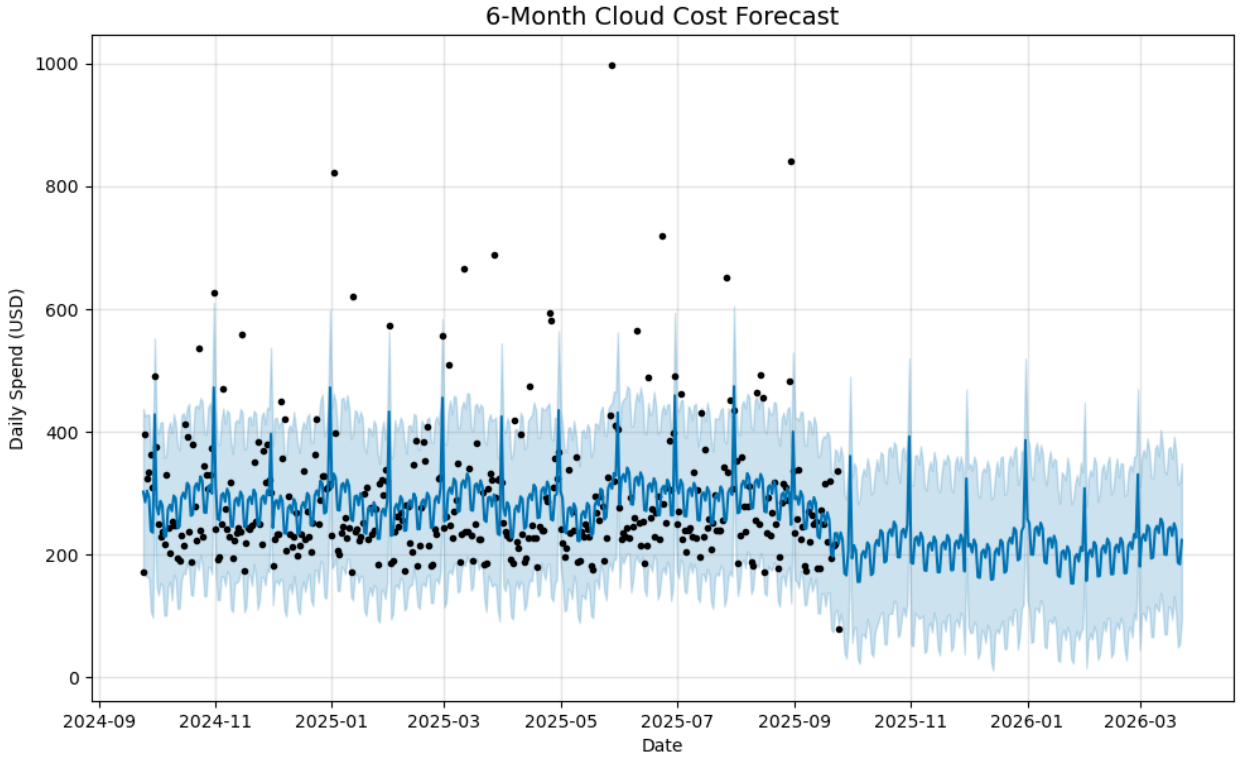


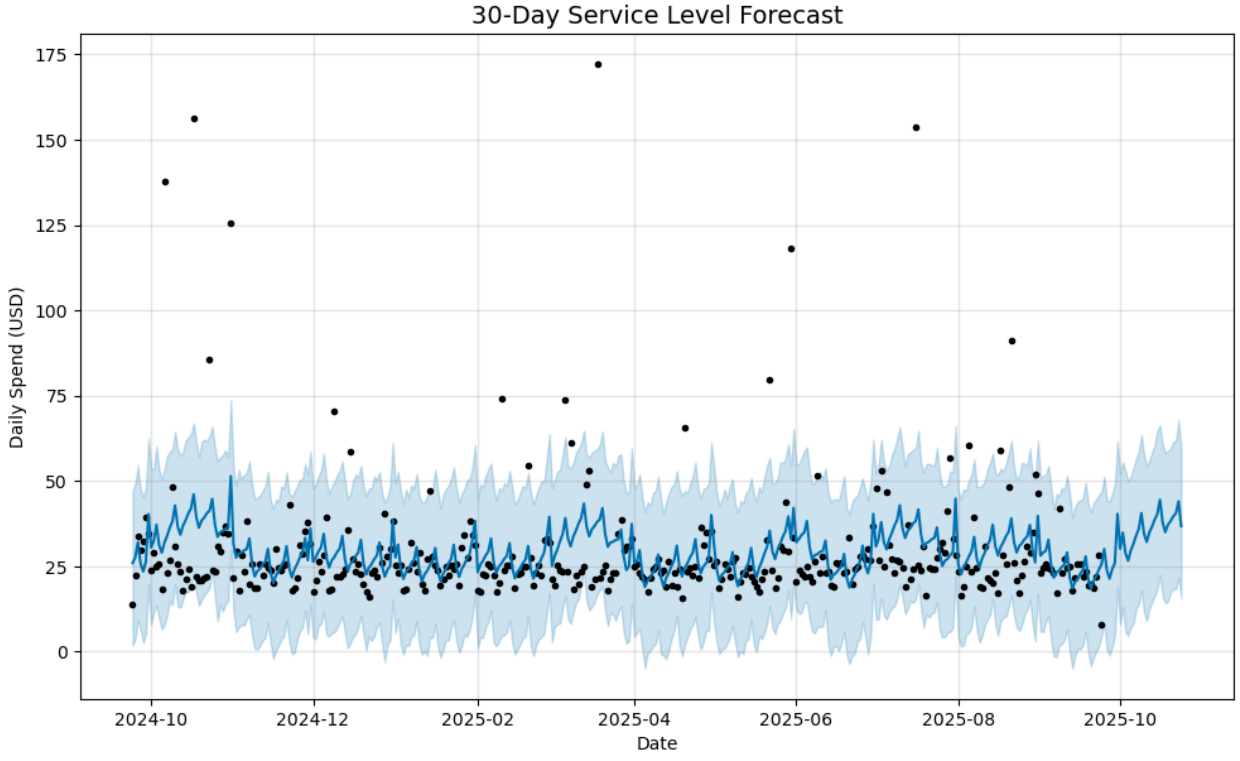
Forecasting

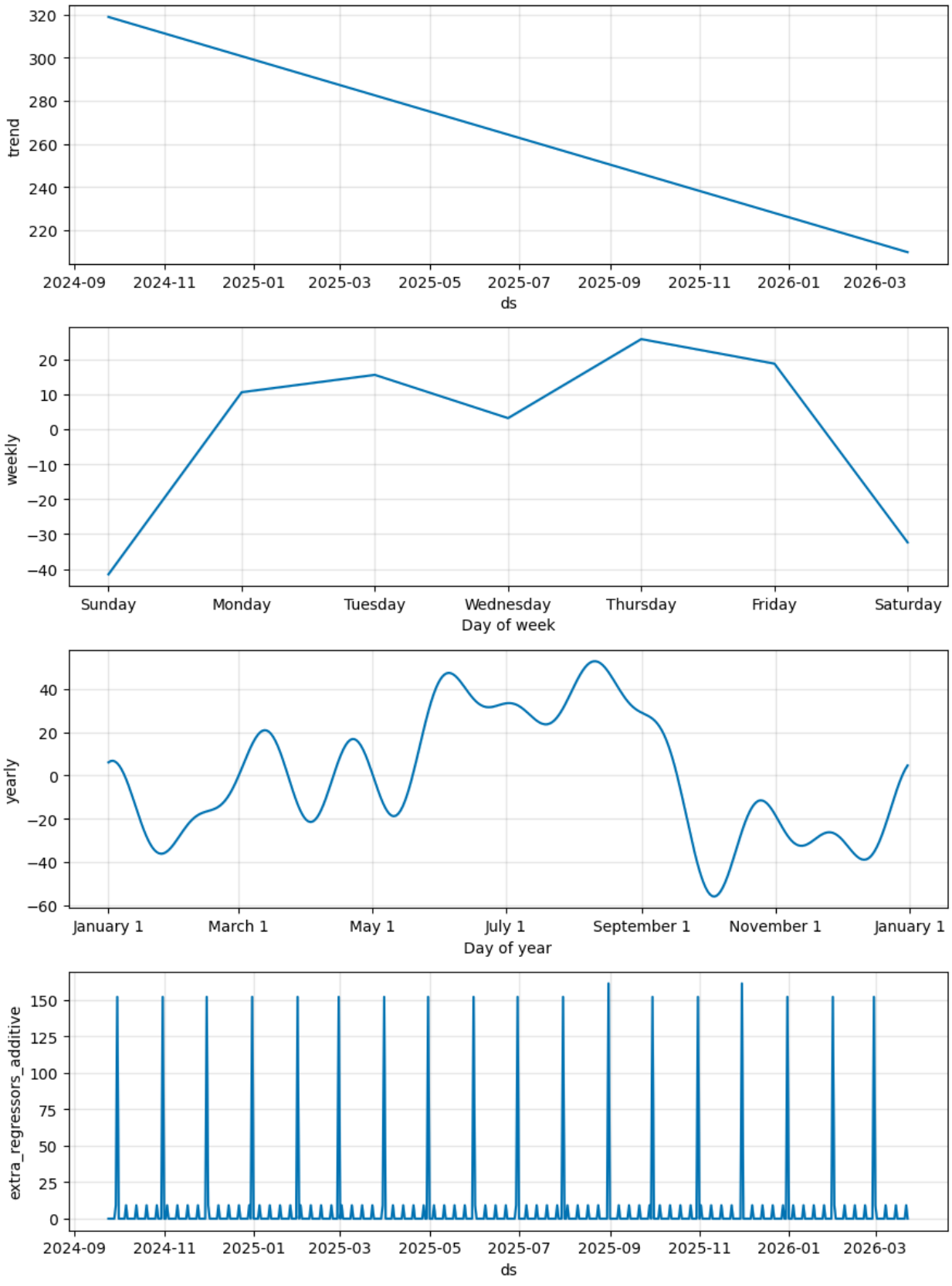
- Tool: Prophet (handles trend + seasonality + regressors).
- Global forecast:

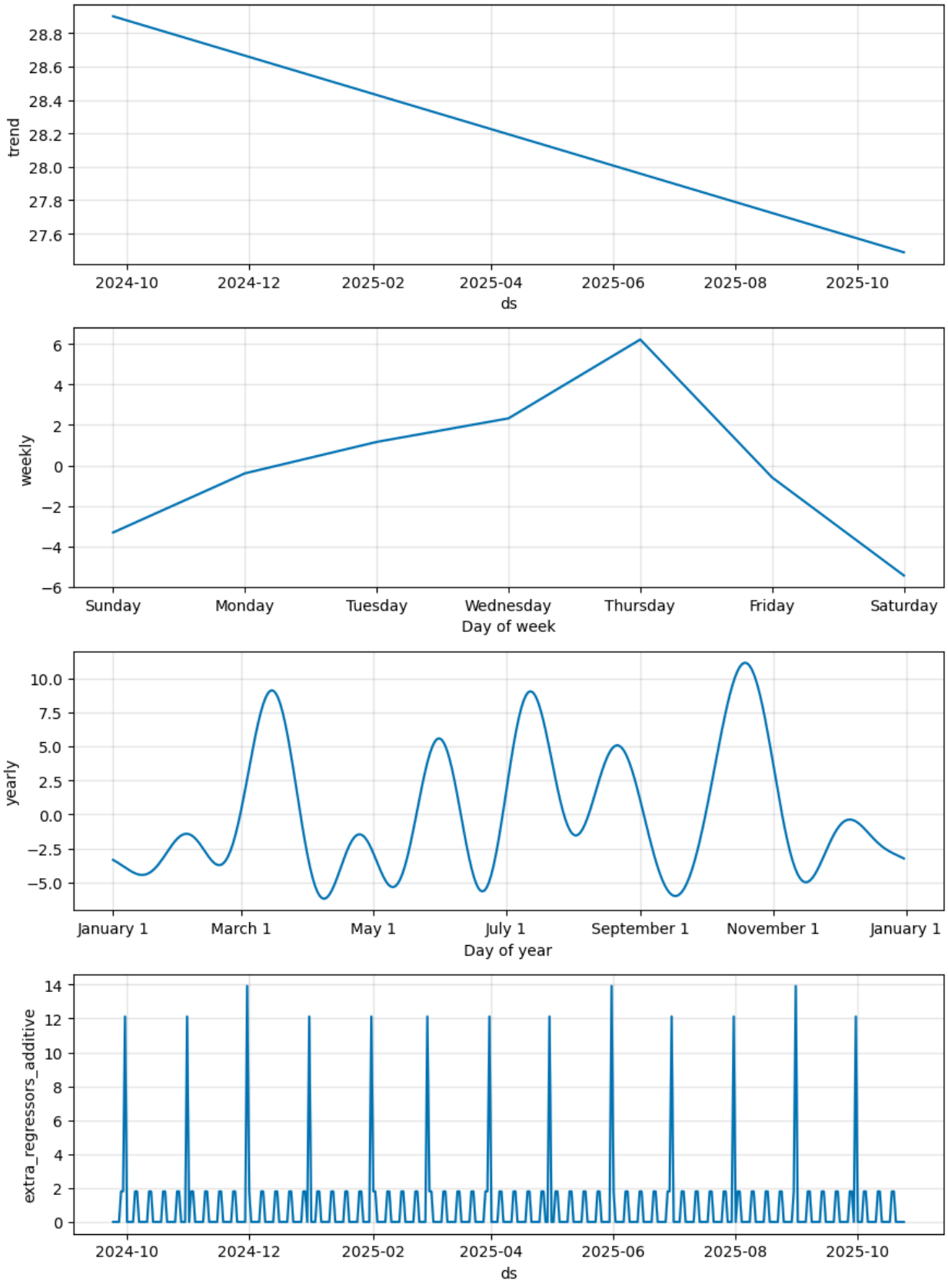
    - 30 days: stable trend, costs ~200–400 USD/day.
    - 6 months: seasonal peaks in May–July, dips in Sept–Nov.
- Regressors: weekends (lower cost), month-end (batch job spikes).
- Service-level forecasts:
    - RDS & ELB → high volatility.
    - S3 & Lambda → stable & predictable.
- Deliverables:
    - cloud_cost_forecast.csv (global).
    - service_level_forecasts.csv (per service).




6-Month Cloud Cost Forecast
Daily Spend (USD)
0
200
400
600
800
1000
2024-09
2024-11
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
2026-01
2026-03
Date


30-Day Service Level Forecast
Daily Spend (USD)
0
25
50
75
100
125
150
175
2024-10
2024-12
2025-02
2025-04
2025-06
2025-08
2025-10
Date


trend
320
300
280
260
240
220
2024-09
2024-11
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
2026-01
2026-03
ds
weekly
20
10
0
−10
−20
−30
−40
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week
yearly
40
20
0
−20
−40
−60
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year
extra_regressors_additive
150
125
100
75
50
25
0
2024-09
2024-11
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
2026-01
2026-03
ds


trend
28.8
28.6
28.4
28.2
28.0
27.8
27.6
2024-10
2024-12
2025-02
2025-04
2025-06
2025-08
2025-10
ds
weekly
6
4
2
0
−2
−4
−6
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week
yearly
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year
extra_regressors_additive
14
12
10
8
6
4
2
0
2024-10
2024-12
2025-02
2025-04
2025-06
2025-08
2025-10
ds

Phase 4: Execute (Power BI Dashboard)

- Designed interactive dashboard with 4 key pages:
    - Executive Overview: KPIs + global trends.
    - Anomaly Insights: anomalies by service, team, timeline.
    - Forecast Insights: global + per-service forecasts with seasonality patterns.
    - Team & Region Insights: spend by team, region maps, drill-through.
- Filters: service, team, region, date.
- Outputs: .pbix dashboard file + exports.

5. Key Insights

- **Anomalies**: RDS & ELB caused the largest spikes, often unrelated to usage (inefficient scaling).
- **Forecasting**: Costs are stable overall, but predictable cycles (weekends, month-end, seasonal peaks) matter for planning.
- **Business Value**:
    - Early detection of cost spikes.
    - Short-term planning for next 30 days.
    - Long-term visibility (6 months) for budgeting.
    - Accountability across services & teams.

6. Recommendations

- Set automated alerts for anomalies in RDS & ELB, with thresholds tied to historical rolling std.
- Implement cost governance policies for teams with repeated anomalies.
- Negotiate reserved capacity for predictable services (S3, Lambda) to reduce costs.
- Review weekend/month-end jobs to identify optimization opportunities.
- Use this dashboard as a cloud FinOps tool for continuous monitoring.

7. Deliverables

- Jupyter Notebooks (Data Cleaning, EDA, Anomaly Detection, Forecasting).
- Saved ML model (isolation_forest_final.joblib + scaler_final.joblib).
- Forecast & anomaly CSVs for dashboarding.
- Power BI Dashboard (Cloud_Cost_Analytics.pbix).
- GitHub repo with code, data samples, and README.