Understanding Structure and Dynamics in an Unlabelled Time-Series Dataset

Context & Problem Statement

As part of a data analysis assignment, I was provided with a large, unlabelled multivariate dataset containing **8,688 observations and 56 columns**. The only instruction given was that the data represents **time-series observations**. No timestamps, feature names, units, or business context were supplied.

The challenge was therefore not to optimize a known metric, but to understand what kind of system this data represents. This required forming assumptions carefully, validating them empirically, and extracting insight purely from observed behavior.

This mirrors real-world analytical settings, particularly in fast-growing or acquisition-driven environments, where data is often inherited without documentation and must be understood before it can be trusted or acted upon.

Overview of the Analytical Approach

The analysis followed a structured but exploratory workflow, designed to progressively reduce uncertainty and avoid premature modeling.

The work unfolded in four main phases:

1. Investigation and exploration of the raw data
2. Hypothesis formation driven by visual and statistical patterns
3. Hypothesis testing using formal statistical tools
4. Evaluation of predictive potential through a simple forecasting model

Each phase informed the next, allowing conclusions to emerge naturally from the data.

Phase 1: Investigating the Dataset

Initial Exploration and Assumptions

The analysis began with basic exploration and cleaning to understand scale, continuity, and integrity. I checked whether any columns were strictly increasing or decreasing, which could naturally represent time.

No such columns were found. Since the dataset was stated to be time-series but lacked timestamps, I introduced a sequential time index under the assumption that rows were ordered chronologically. This assumption was later validated through strong temporal dependencies observed across most features.

Searching for Categories or Labels

Next, I examined whether any columns appeared to encode categories or labels. Columns with very few unique values were flagged for inspection.

One column stood out clearly. **Column 10** contained only the values 0, 1, and 2. This suggested it might represent segmentation or grouping. Rather than assuming this, it was treated as a hypothesis and tested explicitly later.

Identifying Informative Signals Through Outliers

With no labels or context, I needed a principled way to identify features worth deeper investigation. An outlier analysis revealed a small set of columns with an unusually high number of extreme values relative to the rest of the dataset.

More importantly, when outliers were examined over time, they were **not evenly distributed**. For several features, the number of extreme observations increased sharply after a specific point in the series. This concentration of anomalies over time was the first strong signal that the system might be undergoing a **structural change**, rather than behaving stably throughout.

This observation directly motivated the later hypothesis of a regime shift.

Relationship Discovery Through Correlation

To uncover structure beyond individual features, I examined correlations across all columns. This revealed that the same outlier-heavy columns were **almost perfectly correlated** with one another, indicating redundancy and a shared underlying driver.

The correlation matrix also surfaced another pair of columns with near-perfect correlation that had not been flagged earlier. From this point onward, the analysis focused primarily on these subsets, as they appeared to capture the core dynamics of the dataset.

Phase 2: Visual Analysis & Hypothesis Formation

With a working time index in place, I examined how the selected features evolved over time.

Two distinct behavioral patterns emerged.

Columns 1, 11, 30, 39, and 48

- Move almost identically over time, forming tight bands
- Exhibit a clear and coordinated decline around the midpoint of the dataset
- Show non-stationary behavior, meaning their average level changes over time

Columns 20 and 29

- Follow a different pattern with lower overall variance
- Fluctuate around a stable mean
- Appear stationary and tightly constrained

Autocorrelation Function plots reinforced these observations. The first group showed sustained high autocorrelation across many lags, indicating strong persistence where past values strongly influence future values. The second group showed rapid decay in autocorrelation, consistent with already stabilized or transformed metrics.

Crucially, the timing of increased outlier frequency aligned closely with the observed decline in rolling averages for the first group of features. This convergence of evidence strengthened the hypothesis that the system undergoes a meaningful change in behavior partway through the dataset.

Phase 3: Hypothesis Statistical Testing

Based on the patterns observed so far, four hypotheses were formalized and tested.

Hypothesis 1: The system undergoes a structural regime change

This was tested using change-point detection techniques- where data was split at specific time intervals to see if any significant differences in mean, variance and other properties were noticed in the groups created by this splitting. The results confirmed a structural break affecting multiple correlated features at approximately the same time.

Hypothesis 2: A subset of variables share long-run equilibrium relationships

To test whether the highly correlated features were linked beyond short-term movement, I applied the **Johansen cointegration test** (a test to see if two variables have a long term equilibrium relationship). Cointegration indicates that variables may move apart sporadically in the short term but stay together in the long run.

The test confirmed the presence of multiple long-run equilibrium relationships, meaning these features are structurally linked rather than coincidentally correlated.

Hypothesis 3: The dataset exhibits genuine time-series dependence

An assessment was done to see if past values carry information about future values across multiple lags (intervals). The results were statistically significant for the vast majority of features, confirming that the dataset represents a true time-evolving system rather than independent observations.

Hypothesis 4: Column 10 represents meaningful segmentation

This hypothesis was tested using **ANOVA** (a statistical test commonly used to identify difference in behaviour of groups) to compare group means, along with **visualizations** to assess separation of the data.

Both statistical and visual evidence showed no meaningful differences between groups. Column 10 was therefore concluded to be an arbitrary partition rather than a true label or cluster.

Phase 4: Evaluating Predictive Potential

After establishing structure and stability, I explored whether the data supported short-term forecasting. The goal was not maximum accuracy, but to assess whether historical behavior contained usable predictive signal.

A **Random Forest regression model** was applied to a stable, stationary feature that exhibited meaningful temporal dependence. The model achieved strong out-of-sample performance, capturing short-term fluctuations without systematic bias.

This demonstrated that, despite regime changes and non-stationarity in parts of the system, certain metrics are predictable over short horizons and suitable for monitoring or early warning purposes.

Key Results and Takeaways

Several important conclusions emerged from this analysis:

- A small number of latent processes drive a large portion of the dataset
- Extreme events and anomalies are concentrated in specific time periods rather than randomly distributed
- The system undergoes a coordinated structural shift that affects multiple related variables simultaneously
- Variables fall into distinct behavioral categories that should play different analytical roles
- Temporal dependence is strong but uneven, making short-horizon analysis more reliable than long-term forecasting
- Structural understanding provided more value than complex modeling in the absence of labels

Notably, these insights were derived without knowing what the data represented. They emerged from disciplined exploration and validation rather than assumptions.

Business Implications

For a product-focused organization these findings translate into several practical implications:

- **System-level changes can be detected early** by monitoring shifts in variability and anomaly frequency rather than relying only on averages
- **Correlated metrics should not be treated independently**, as doing so overstates confidence and increases noise in decision-making
- **Regime-aware monitoring is critical**, since benchmarks and expectations may become invalid after structural changes
- **Stable, stationary metrics are ideal for health monitoring**, alerting, and early warning systems
- **Predictive models are most useful for short-term guidance**, not long-term automation, in evolving systems

These insights support a cautious, adaptive approach to decision-making in environments where product behavior can change abruptly.

Limitations

The primary limitations stem from missing information rather than analytical choices:

- No feature definitions or units
- No timestamps or external event markers
- No downstream objective or target metric

These constraints informed a focus on structural inference rather than optimization.

Improvements and Next Steps

With additional time or context, the following extensions would add value:

- Expanding analysis to additional columns beyond the initial focus set
- Investigating whether other latent groups or structural changes exist
- Linking detected regime shifts to external events once context is available
- Refining predictive models for operational deployment

Closing Note

This project was approached as an ambiguity-first analysis. With no labels, no objectives, and no context, the priority was to build understanding incrementally and let the data guide conclusions.

The accompanying notebook serves as a technical appendix containing the full empirical evidence behind the insights summarized here. I would welcome discussion around alternative interpretations, modeling choices, or how these findings could translate into concrete product or operational decisions.