10/27/2025

Data Mining Project Report: Uncovering Hidden Demand Shift in Bike Sharing

Vishnu Priya Parambathu Manoharan
Tresa Nancy Thareparambil Paul
Ridha Mariam Rajan
GROUP: 14

1. Introduction and Project Goal

This project analyzes the UCI Bike-Sharing dataset to uncover and characterize shifts in bicycle rental demand. By combining **Time Series Analysis**, **Clustering**, and **Anomaly Detection**, the study aims to move beyond simple aggregated statistics to reveal both stable user patterns and unusual demand deviations. The ultimate goal is to generate **actionable insights** for improving operational efficiency, such as optimizing bike availability and redistribution schedules.

This analysis is fundamentally based on treating the data as a **Time Series**: "A sequence of observations taken sequentially in time" (**Time Series Analysis – Slide 8**), where adjacent hourly observations are dependent. The methodology adheres to the core course objectives, focusing on data mining techniques, data visualization, and acquiring **hands-on experience** (**Introduction – Slide 3**).

2. Research Question and Descriptive Analysis

2.1 Research Question

The core objective of this study is framed by the specific **research question** required to "discover hidden patterns and relationships" (**Lecture interestingness – Slide 4**): ***"What are the deviations in the rush hour?"***

This targeted question focuses the entire subsequent analysis on finding and explaining unusual behaviour (anomalies) specifically during the daily peak demand periods, where operational decisions regarding bike placement are most critical.

2.2 Descriptive Statistics and Feature Engineering

Initial analysis began with **descriptive statistics** (**Introduction – Slide 43**), summarizing the key numerical features (e.g., mean and standard deviation of total counts, temperature, and humidity).

Feature Engineering was crucial for addressing the research question. The hourly data was first transformed into **daily demand profiles** (24-dimensional vectors) to capture the overall trend. Crucially, the analysis then **filtered and focused** on the identified rush hour periods (e.g., 7:00-9:00 and 16:00-18:00) to isolate the data relevant to the research question. Features like workingday and weathersit were leveraged as external covariates to enrich the dataset and explain future anomalies.

3. Methodology and Pattern Analysis

3.1 Data Preparation and Visualization

The data was treated as a **Time Series** (**Time Series Analysis – Slide 10**) and underwent thorough **Data Preprocessing** (**Introduction – Slide 33**). This involved cleaning the dataset and then **normalizing** the 24 hourly features. Normalization was critical because the **Euclidean distance** used in K-Means (**Lecture 3 Clustering – Slide 6-23**) could otherwise be disproportionately influenced by high-volume hours, thus ensuring a fair measure of pattern similarity.

Data Visualization was employed from the outset for **sense-making** and to "see patterns, trends or anomalies" (**Data Visualization – Slide 8**). Initial time series plots showed seasonality, and various charts were used to visualize correlations between demand and weather variables.

3.2 Demand Pattern Clustering

Clustering was applied to the daily profiles to identify **stable demand patterns** (**Lecture 3 – Slide 4**). The **K-Means algorithm**, a **Partition-Based** method (**Lecture 3 Clustering– Slide 6**), was used to categorize days based on their similarity in hourly usage.

- **Cluster Characterization (Hidden Demand Shifts):** The resulting cluster centroids, visualized using **cluster-profile plots**, revealed the two **hidden demand shifts**: the **Commuter Profile** (double-peaked rush hour usage) and the **Leisure Profile** (single, gradual midday peak). These clusters establish the "expected behavior" for the anomaly detection phase.
- **Cluster Characterization:** The resulting clusters represent **stable, hidden demand patterns** or "demand shifts."
    - **Cluster 1 (Commuter Profile):** Characterized by high rental counts during typical morning (7:00-9:00) and afternoon (16:00-18:00) **rush hours**, indicating a strong weekday, work-related usage pattern.
    - **Cluster 2 (Leisure Profile):** Showed a single, gradual peak between noon and late afternoon (12:00-17:00), reflecting non-work, recreational usage typically seen on weekends or holidays.

These profiles were visually confirmed using **cluster-profile plots**, which employ **Data Visualization** for **sense-making and communication**, helping viewers "see patterns, trends or anomalies" (**Lecture 2: Data Visualization – Slide 3**).

---

4. Anomaly Detection and Insight

3.1 Identifying Unusual Behaviour

With the typical profiles established, **Anomaly Detection** was performed. An **anomaly** (or outlier) is defined as "a pattern in the data that does not conform to the expected behaviour" (**Anomaly Detection (1) – Slide 3**). The analysis used a **clustering-based** approach (**Anomaly Detection (2) – Slide 19**), where days exhibiting a significantly large distance from their assigned cluster's centroid were flagged as demand outliers. The focus was specifically on deviations during peak commute hours, as these pose the greatest operational challenge.

Anomaly Type	Operational Implication
Demand Spikes	Unexpectedly high rentals (potential for bike shortages).
Demand Drops	Unusually low rentals (potential for bike oversupply/wasted redistribution efforts).

4.1 Identifying and Justifying Rush Hour Deviations

Anomaly Detection was performed to find deviations from the stable cluster patterns (Anomaly Detection (1) – Slide 3), directly addressing the research question. The system used a clustering-based

approach (Anomaly Detection (2) – Slide 19), where the distance from a day's hourly profile to its assigned cluster centroid was used as the anomaly score.

The analysis was filtered to isolate and classify anomalous rush hour demand (spikes and drops), and these deviations were justified using external information:

1. Weather Extremes: Extreme temperatures or poor weather conditions were strongly correlated with significant demand drops during rush hours. These conditions caused a day classified as a "Commuter Profile" to register as an anomaly, as the extreme weather condition overrode the expected behavioural pattern.

2. Working Day/Holiday Shift: Anomalous days were often working days adjacent to holidays. Here, the demand pattern dramatically shifted from the high-volume Commuter expectation to the low-volume Leisure Profile, demonstrating a clear and predictable deviation in rush hour demand that the model uncovered.

5. Conclusion and Actionable Insights

This combined approach successfully segmented typical usage into two major regimes Commuter and Leisure and accurately highlighted deviations caused by predictable external factors. The **uncovered hidden demand shifts** provide direct, **actionable insights** for the bike-sharing operator:

- **Dynamic Bike Redistribution:** Operational planning should not solely rely on the working day flag. **Monitor weather forecasts** for extreme conditions, as these are a stronger predictor of imminent **demand drops** than the day-of-the-week alone. Redistribution efforts should be scaled down on forecasted bad weather days.

- **Holiday Staffing:** Staffing and bike maintenance/redistribution efforts should reflect the lower-volume 'Leisure Profile' on all observed holidays, not the high-volume weekday Commuter routine.

- **Optimizing Rush Hour Planning:** Integrating weather and holiday data into predictive models based on these clusters is critical for avoiding over- or under-stocking during critical rush hours, thereby improving user experience and cost efficiency.