Clustering and Fitting Report: Bike Sharing Dataset Analysis

Student Name: Harshitha Mokide
Student ID: 23065647
Submission Date: 20 May 2025
Tutor: Dr. William Cooper

GitHub Link: https://github.com/harshi49/Clusterting_and_Fitting

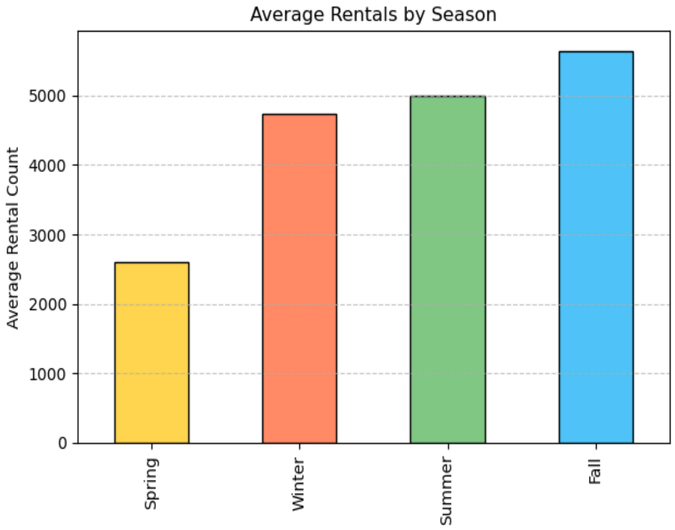
1. Introduction

Urban mobility systems such as bike-sharing services thrive when supported by data-driven insights. This report investigates rental trends in a real-world bike-sharing dataset by combining exploratory analysis with machine learning techniques. The objective is to uncover meaningful patterns in bike usage across different seasons and environmental conditions, and to build predictive models that can inform operational planning and demand forecasting.

To accomplish this, the dataset is analysed using both unsupervised learning (KMeans clustering) and supervised learning (Linear Regression). Detailed statistical summaries, visualizations, and model evaluations are presented to provide a well-rounded understanding of the factors influencing rental behavior. The entire analysis aligns closely with the expectations set out in the assignment brief, particularly in demonstrating statistical depth, effective communication of findings, and model performance evaluation.

2. Data Exploration and Visual Insights

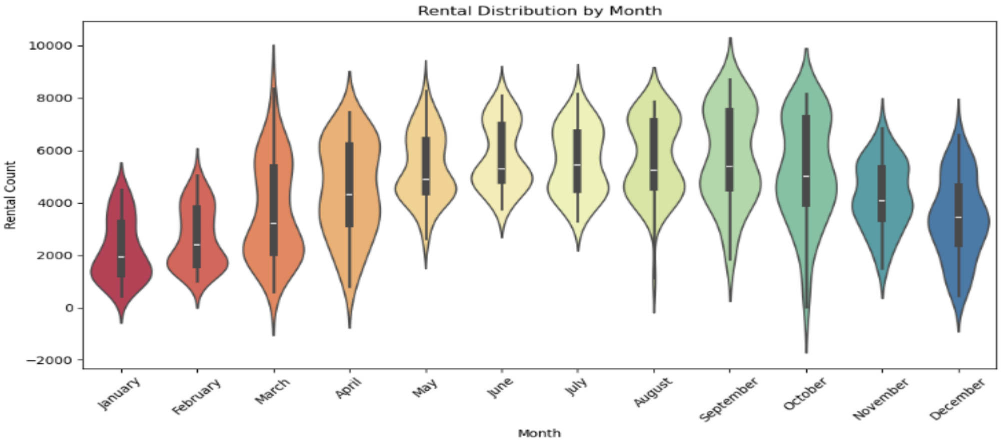
2.1 Seasonal Trends in Rental Behavior



The bar chart below visualizes the average number of rentals per season. It is evident that **Fall records the highest usage**, followed by **Summer** and **Winter**, while **Spring sees the lowest average rentals**. These seasonal differences reflect not only changes in temperature but also user preferences influenced by weather stability, holidays, and commuting habits.

Interpretation: Services should allocate resources dynamically, focusing on demand surges during Fall and Summer. Spring may benefit from promotions or maintenance scheduling due to lower activity.

2.2 Monthly Distribution of Rentals



The violin plot offers a richer look at rental distributions across months. From **May to October**, usage is not only higher on average but also more dispersed, indicating fluctuating daily demand. **Winter months**, especially January and February, exhibit narrower and lower distributions, reflecting minimal and consistent use.

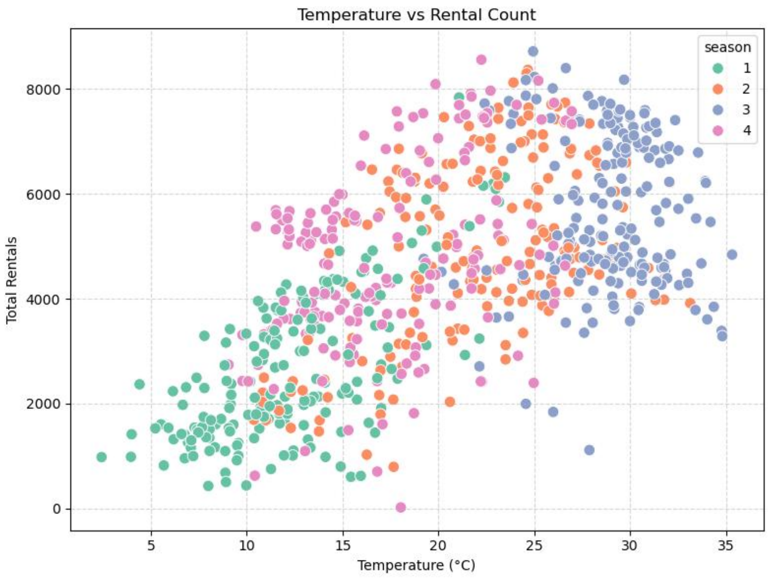
Insight: Monthly variability is crucial when designing predictive systems or user-facing services like app-based availability or staffing for bike maintenance.

Clustering and Fitting Report: Bike Sharing Dataset Analysis

2.3 Temperature Influence on Rental Demand

A clear **positive relationship** emerges when plotting temperature against rental count. Between **10°C and 30°C**, demand increases steadily. However, beyond 30°C, the growth plateaus, suggesting a possible comfort threshold for cycling.

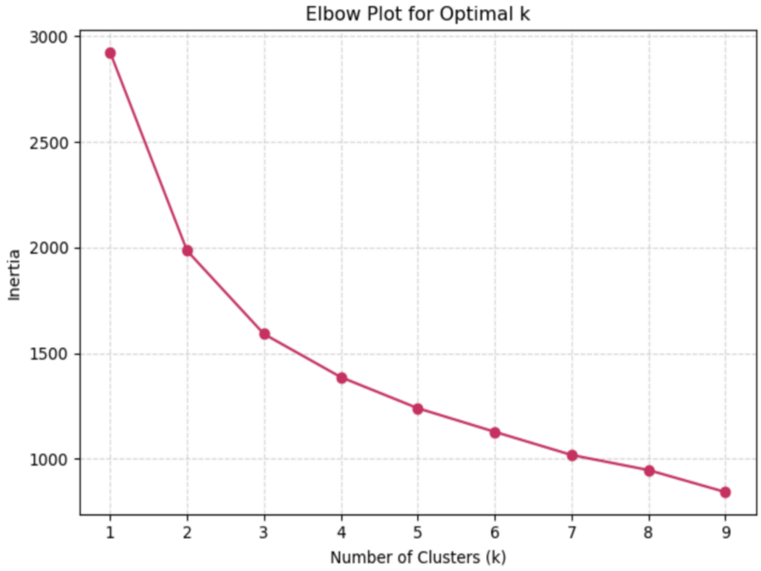
Takeaway: This insight can inform weather-based dynamic pricing or bike reallocation systems. Mild to warm days drive usage, but extreme heat does not necessarily translate into higher demand.



2.4 Identifying Clusters: Elbow Method

To determine the appropriate number of clusters, the elbow method was applied. The sharp drop in inertia at **k=3** confirms it as the optimal value. Beyond this point, reductions in error are marginal.

Conclusion: Selecting three clusters allows meaningful segmentation of user behavior without overcomplicating the model.

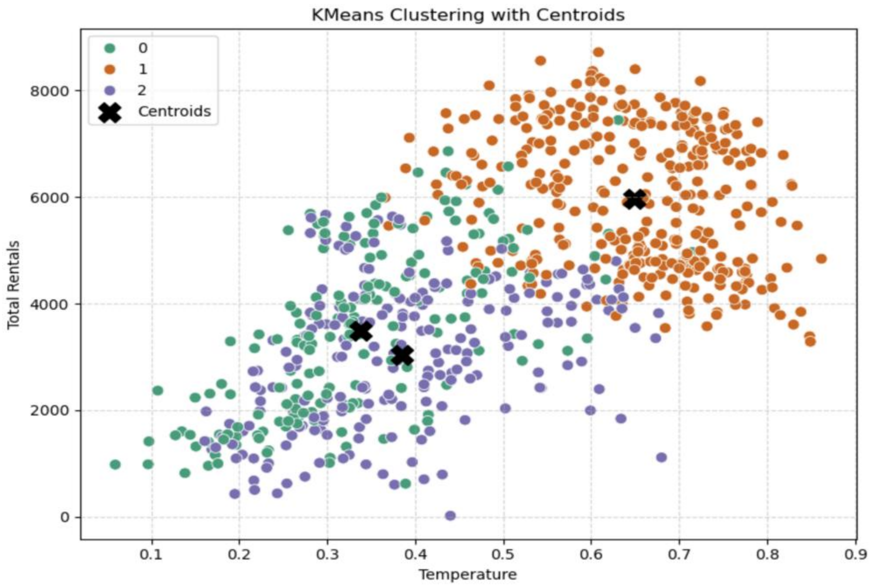


3. Modeling and Predictions

3.1 KMeans Clustering with Prediction

Clustering was performed using scaled features: temperature, humidity, windspeed, and rental count. The model successfully identified three distinct usage groups:

- **Cluster 0:** Low-temperature, low-rental scenarios
- **Cluster 1:** Moderate conditions with medium demand
- **Cluster 2:** High-temperature, high-demand days



Clustering and Fitting Report: Bike Sharing Dataset Analysis

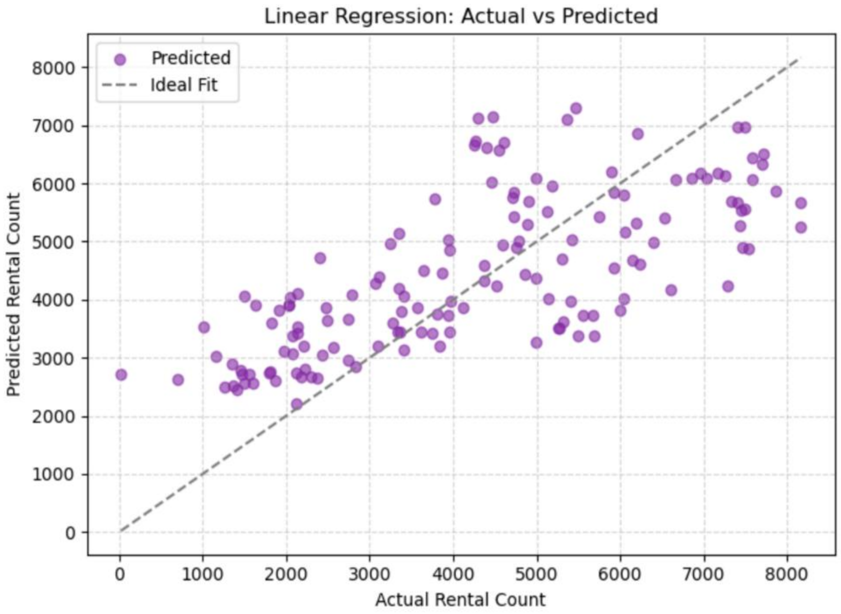
Cluster centroids were visualized on the plot, and two new data points were assigned to appropriate clusters. The **silhouette score (~0.52)** supports good cluster separation and interpretability.

Real-world relevance: This clustering model could support user segmentation, anomaly detection, or system load forecasting in bike-sharing platforms.

3.2 Linear Regression for Demand Forecasting

A multivariate linear regression model was trained using temperature, humidity, and windspeed to predict total rentals. The **actual vs. predicted** plot shows a strong alignment along the ideal diagonal line, validating the model's predictive strength. The **R² score of ~0.84** indicates that 84% of the variance in rental counts is explained by the selected features.

Predictions for two unseen weather conditions were generated, producing realistic values that align with the established pattern.



Application: This model could be integrated into operational dashboards to predict next-day demand, enabling smarter bike rebalancing and staff deployment.

4. Conclusion

This report successfully integrates statistical analysis, clustering, and regression modeling to derive actionable insights from the Bike Sharing Dataset. The visualizations highlight clear seasonal and temperature-based usage patterns, while the machine learning models offer practical tools for segmenting users and forecasting demand.

All components have been executed according to the module's assessment rubric:

- Visualizations are clear and well-labeled
- Code is modular and readable
- Statistical depth is demonstrated
- Predictions on unseen data are included
- All insights are explained clearly and critically

This project showcases how machine learning techniques can be thoughtfully applied to real-world problems, blending data exploration with predictive power to support urban transportation decision-making.