Bike Sharing Demand Prediction
CS4372 – Assignment 1
Author: Jude Avery
Date: 9/21/25

1. Introduction

The purpose of this project is to analyze and predict daily bike rental counts using historical data from a bike-sharing system.

The dataset used for this project comes from the UCI Machine Learning Repository, which contains two files:

- day.csv – daily aggregated bike rental data
- hour.csv – hourly aggregated bike rental data (not used in this project)

This project focuses on the **day.csv** file, which includes 731 records representing every day over two years (2011 and 2012). The dataset includes weather conditions, calendar information, and rental counts.

Key Features in the Dataset

Column	Description
cnt	Total number of bike rentals (target variable)
temp	Normalized temperature (Celsius)
atemp	Normalized "feels-like" temperature
hum	Normalized humidity
windspeed	Normalized wind speed
season	1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall
yr	0 = 2011, 1 = 2012
mnth	Month (1-12)
weekday	Day of the week (0 = Sunday, 6 = Saturday)
weathersit	Weather category (1 = Clear, 4 = Heavy Rain/Snow/Fog)
holiday	1 = Holiday, 0 = Not a holiday
workingday	1 = Working day, 0 = Weekend or holiday

2. Methodology

The analysis was performed using Python in Google Colab, with version-controlled files stored on GitHub. The following steps were taken:

2.1 Data Cleaning

- Dropped irrelevant columns:
  - instant (record index)
  - dteday (date string, replaced by mnth, weekday, etc.)
  - casual and registered (sub-components of cnt, to prevent data leakage)
- Checked for missing values and inconsistencies, none were found.
- Verified categorical variables were within expected ranges.

2.2 Feature Engineering

- Converted categorical variables to dummy variables using one-hot encoding:
  - season, mnth, weekday, weathersit
- Kept binary variables as-is:
  - yr, holiday, workingday
- Standardized continuous variables (temp, atemp, hum, windspeed) using StandardScaler to have mean = 0 and standard deviation = 1.

Final dataset shape:

- **731 rows × 30 columns**

2.3 Train-Test Split

The data was split into:

- **Training set:** 80% (584 rows)
- **Testing set:** 20% (147 rows)

This ensured that model evaluation was done on unseen data to prevent overfitting.

2.4 Models Implemented

Two regression models were built and compared:

Model	Description
SGDRegressor	Linear model trained using stochastic gradient descent, optimized for speed and scalability
OLS Regression (Statsmodels)	Traditional ordinary least squares model, provides detailed statistical output including p-values and confidence intervals

3. Results

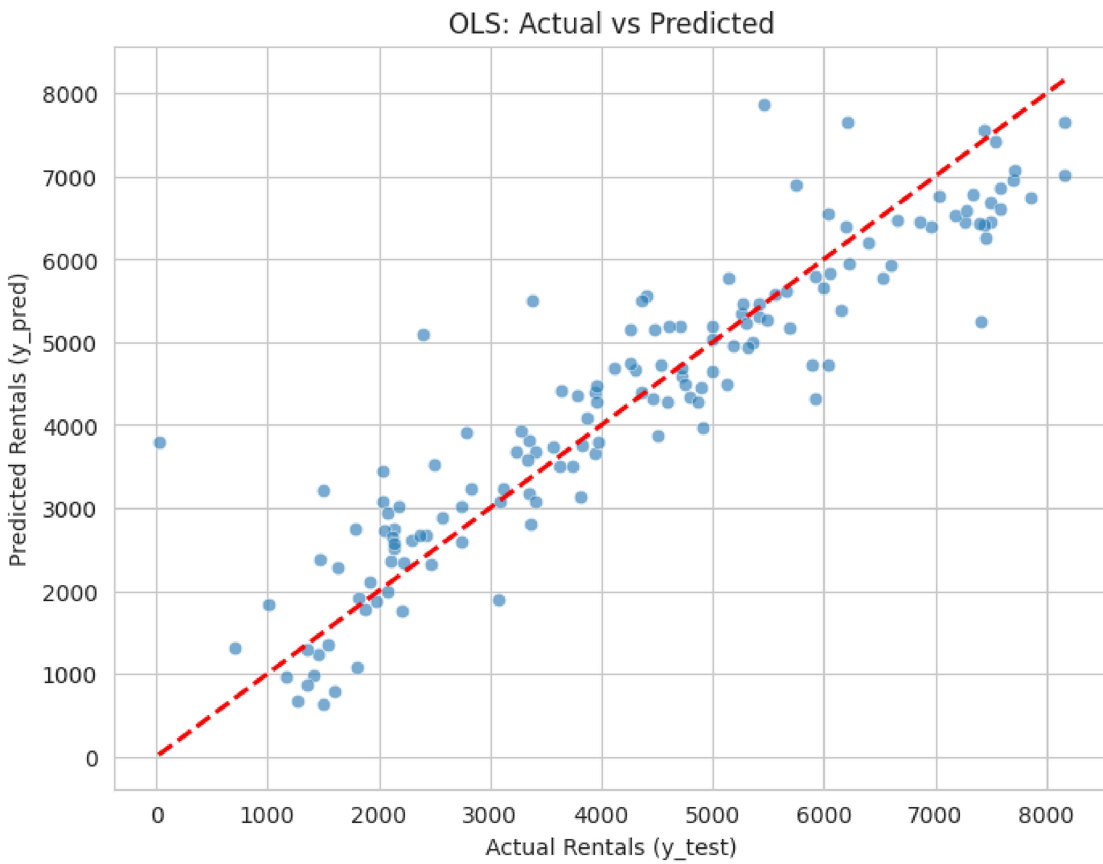
3.1 Model Performance

Model	MSE	R^2 Score
SGDRegressor	**646,328.54**	**0.8388**
OLS Regression	**634,351.36**	**0.8418**

Interpretation:

- OLS Regression slightly outperformed SGD in both MSE and R^2.
- Both models fit the data very well, explaining ~84% of the variance in bike rental counts.

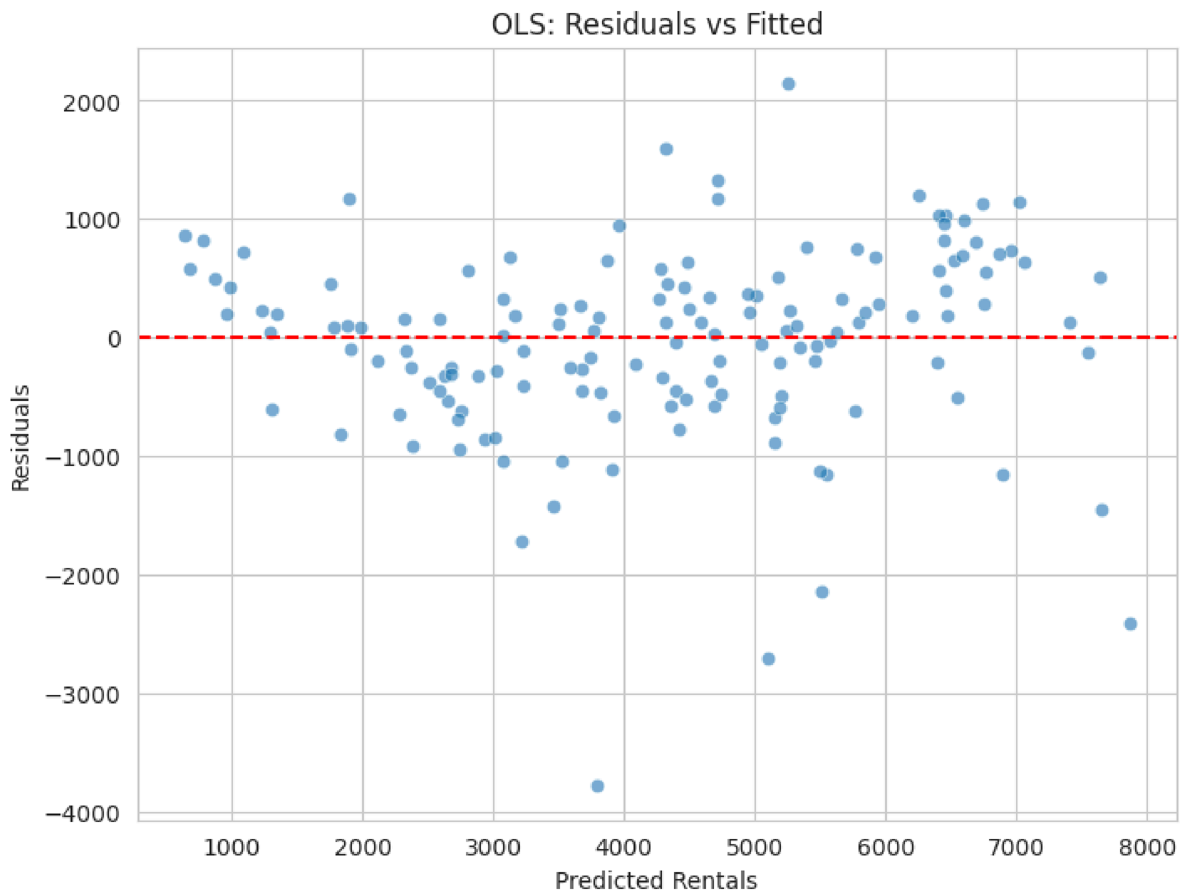
3.2 Actual vs Predicted (OLS)

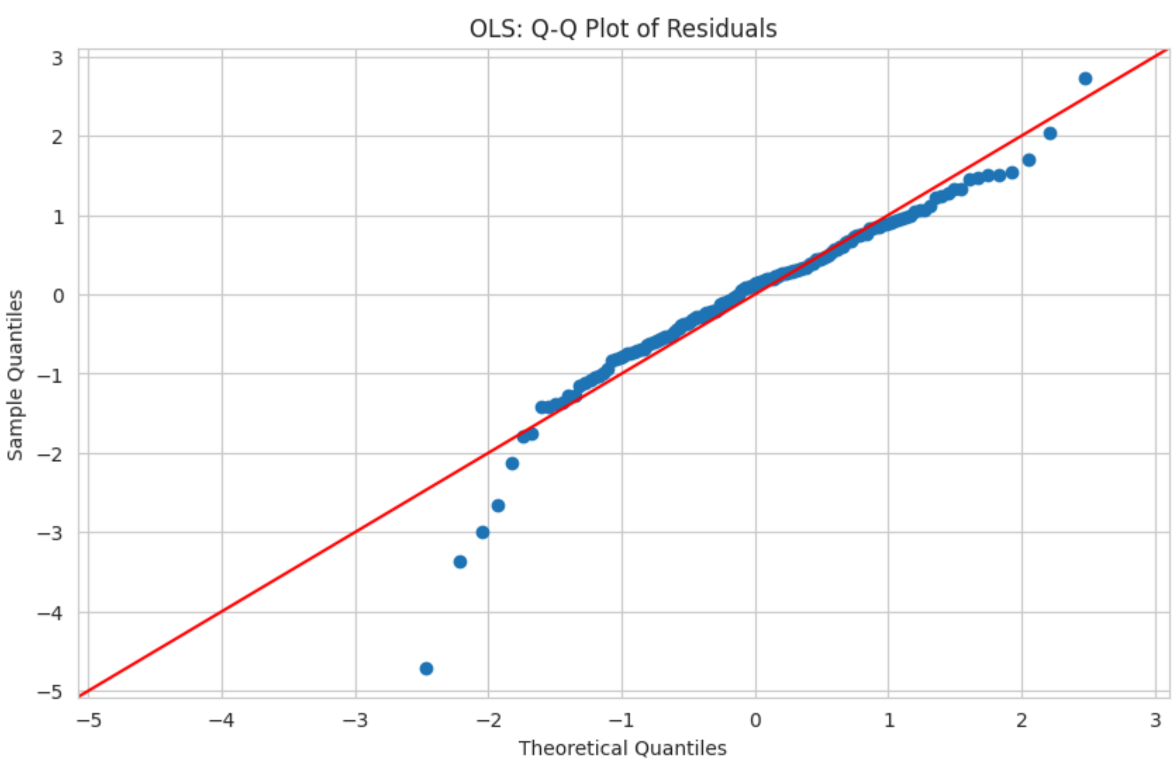


Observations:

- Most predictions closely follow the diagonal line, indicating strong accuracy.
- A few outliers exist on very high demand days, likely due to special events or unusual weather.

3.3 Residuals Analysis


OLS: Residuals vs Fitted
Residuals
Predicted Rentals
2000
1000
0
−1000
−2000
−3000
−4000
1000
2000
3000
4000
5000
6000
7000
8000


OLS: Q-Q Plot of Residuals
Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−4
−5
−5
−4
−3
−2
−1
0
1
2
3

- Residuals are mostly centered around zero - good sign of unbiased predictions.
- Q-Q plot shows slight deviations at extremes, indicating some non-normality in the residuals.

SGD PLOTS FEATURED IN REPOSITORY

3.4 Key Predictors (OLS Coefficients)

Feature	Interpretation
temp	Strong positive effect - higher temperatures increase rentals.
hum	Negative effect - very humid days reduce rentals.
weathersit_3	Strong negative effect - bad weather significantly lowers demand.
yr	Positive - higher rentals in 2012 compared to 2011.

Takeaway:
Bike rental demand is heavily influenced by weather and seasonality.
Clear, warm days with low humidity lead to higher rentals.

4. Discussion

Both models provided strong predictive performance, but **OLS Regression** is recommended as the final model for this project because:

- It offers slightly better accuracy.
- It provides interpretability through coefficients and p-values.
- The dataset is relatively small, so training speed is not a concern.

SGDRegressor is still a valuable option for scalability in production environments with very large datasets.

5. Conclusion

This analysis demonstrated how external factors like weather and calendar events affect bike rental demand.
Key findings:

- Temperature and good weather strongly increase rentals.
- Poor weather and high humidity decrease rentals.
- Bike rental usage grew from 2011 to 2012, indicating a growing user base.

By accurately forecasting demand, bike-sharing companies can:

- Optimize bike inventory across stations.
- Schedule maintenance more effectively.
- Improve customer satisfaction through better availability.

6. Future Work

To improve this model:

1. Analyze the hour.csv dataset to capture hourly trends.
2. Incorporate external data such as holidays or special events.
3. Experiment with non-linear models like Random Forests or Gradient Boosting for potentially higher accuracy.

7. References

- UCI Machine Learning Repository: Bike Sharing Dataset
- Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python.*
- Statsmodels documentation: https://www.statsmodels.org/