Alex Song

DATA 5421 – 01

Project Phase 1

January 25, 2026

Introduction

This project uses a bike sharing usage data from Kaggle (Bike Share Daily Count). The dataset contains daily observations of total bike rental counts and a breakdown into casual and registered users, along with weather-related factors such as weather situation, temperature, humidity, and windspeed; and calendar-related factors such as season, month, weekday, holiday, and working day status. The primary goal of Phase 1 of the project is to explore the 2011 data, conduct exploratory data analysis (EDA) to understand how these factors affect daily bike rental counts, and develop regression models.

The analysis includes EDA of numerical and categorical variables, single-predictor regression models, interaction-effect models, backward selection to determine a final model, and evaluation of the final model on test data. Note that the project uses data from 2011 as the training dataset to build predictive models and data from 2012 as the test dataset to evaluate model performance.

Data Methodology

The raw dataset was imported and separated into training (year 2011) and test (year 2012) datasets based on the yr variable. The dataset contains two types of variables: categorical and numerical variables. Categorical variables—season, month (mnth), holiday, weekday, workingday, and weather situation (weathersit)—were converted into factor variables. Numerical variables—temperature (temp), apparent temperature (atemp), humidity (hum), and windspeed—as well as response variables such as casual, registered, and total count (cnt) were retained in their normalized format. The year variable (yr) and record index (instant) were removed since the dataset was already split by year and R provides an index automatically.:

1. Exploratory data analysis (EDA) of numerical and categorical variables, including statistical summaries, frequency tables, correlation matrix, boxplots, and histograms.
2. Single predictor regression models to examine each predictor's relationship with daily rental counts.
3. Additive and Interaction models with three hypothesized interaction effects.
4. Full model including all predictors and interactions, then it was reduced to a final model using the backward selection method.
5. Evaluation of the final model on 2011 and 2012 dataset.

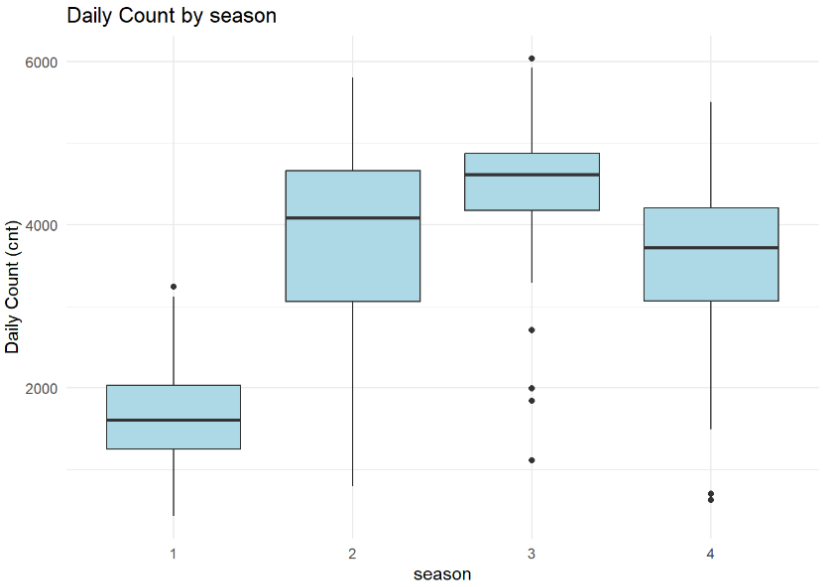
Result

a. Exploratory Data Analysis (EDA)

Statistical summaries of numerical variables show that rental counts range from 431 to 6,043, with a mean of 3,410 and a median of 3,740. Temperature shows moderate variability, while humidity shows a narrower range.

	temp	atemp	hum	windspeed	cnt
temp	1.000	0.996	0.146	-0.114	0.771
atemp	0.996	1.000	0.156	-0.137	0.775
hum	0.146	0.156	1.000	-0.216	0.002
windspeed	-0.114	-0.137	-0.216	1.000	-0.278
cnt	0.771	0.775	0.002	-0.278	1.000

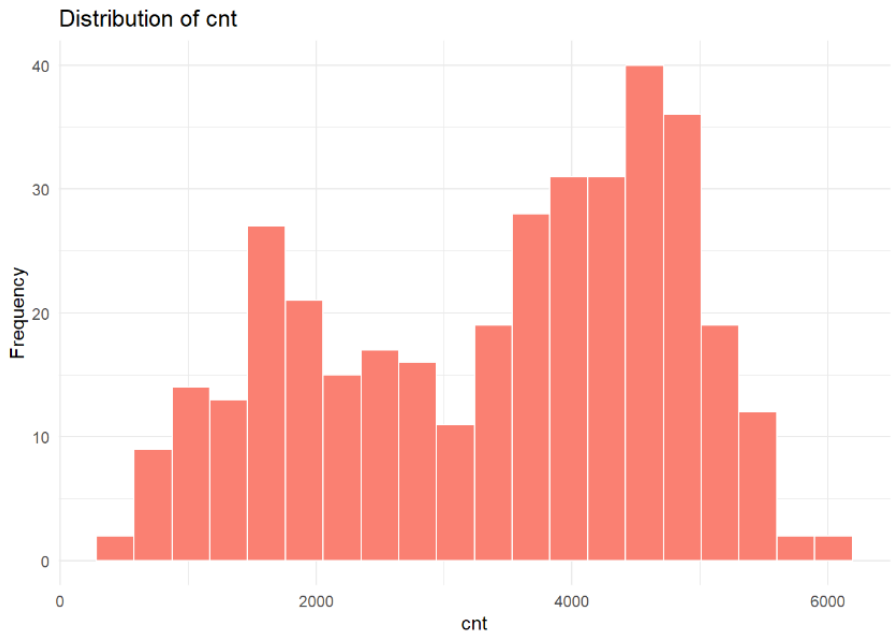
Correlation matrix shows a strong positive relationship between temp and rental count (0.771) and a strong relationship between atemp (feeling temperature) and rental count (0.775). Humidity shows almost no linear relationship with rental count (.002). Windspeed shows a negative correlation (-0.278).



Daily Count by mnth

Boxplot shows clear patterns in season and month related to rental count (cnt), showing highest ride count in warmer seasons (2 = Summer, 3 = Fall) compared to Spring (season = 1) and the same pattern is observed in Month boxplot. Weather situations also show strong relationships, with significant higher rental count in clear (weathersit = 1) and misty weather (weathersit = 2). However,

rainy weather (weathersit = 3) occurs infrequently (about 4% of the total observation). Holiday and weekday show minimal variation.



Histograms show that most numeric variables have roughly unimodal distributions, whereas rental counts show a mild right-skew, as daily rental count raging near 431 and near 6,000. Based on these result, it appears that temperature, season, and month appears to be the strongest predictors, while humidity shows little predictive value.

b. Single-Predictor Models

Individual linear models were fitted for each of the ten predictors.

Numeric predictors: temp, atemp (feeling temp), hum (humidity), windspeed.

Categorical predictors: season, mnth (month), holiday, weekday, workingday, weathersit (weather situations).

The strongest predictors were season, month, and temperature (temp, atemp). Season shows an adjusted R-squared of 0.569, month shows an adjusted R-squared of 0.691, and temperature shows an adjusted R-squared of 0.594. The weakest predictors were workingday and humidity (both have r-squared of -0.002), and weekday also showed an adjusted r-squared near 0. They have a low r-squared and high p-value indicating no significant linear relationship with daily rental count. This result aligns with the previous finding based on EDA.

Based on the above results, month (mnth) showed the best fit followed by temperature (temp) and season. Model for working day and humidity probably have no predictive power. And the residual analysis for the strongest models showed high variability:

- Season: RSE = 905.4 (min –3349.4, max 2029.8)
- Month: RSE = 766.4 (min –3357.2, max 2149.7)
- Temperature: RSE = 878.9 (min –3375, max 1985.7)

These results indicate that individual predictors alone cannot explain the data's variability.

c. Interaction-effect Models

Since no individual predictor can explain the substantial variability in daily bike rental counts on its own, three interaction models were explored.

1. temp * season:

   This interaction was chosen because both variables were strong predictors, and temperature may affect rental counts differently across seasons.

   It was highly significant for Season 2 and Season 3, Summer and Fall respectively. Looking at their small p-values, temperature has the strongest effect in Summer (p = 0.012) and somewhat weaker effect in fall (p = 0.009), suggesting that warmer temperature increases the number to daily rental in Summer and little less during Fall.

   The interaction effect model improved the r-squared valued from 0.691 (that of additive, i.e., temp + season) to 0.713 and lowered RSE from 770.5 (additive) to 745.7, note that it's the strongest improvement among the three interactions. This provides strong evidence of a temperature-season interaction, making it a meaningful improvement over the additive model.

2. temp * weathersit:

   This interaction was picked because temp in one of the strongest predictors and weather situations (weathersit) is one of the weakest predictors. This interaction tested whether temperature's effect varies across weather conditions. It appears that the interaction was not significant, and both temp:wethersit2 and temp:weathersit3 show large p-value (0.216 and 0.578), indicating temperature affects rental count consistently regardless of weather situations. Also, the interaction model has slightly higher r-squared value (0.668) compared to the additive model (0.666), suggesting that the interaction provides no meaningful improvement in explanatory power. In other words, there is no statistical evidence that temperature interacts with weather situations.

3. workingday * weathersit:

   This interaction was picked because they were the weakest predictors and based on hypothesis that working day affects rental count differently across weather situations, if people commute to their work regardless of the weather situations. This interaction shows very small t-value and large p-values—- workingday:weathersit2 (t = 0.183, p = 0.855) and workingday:weathersit3 (t =

0.934, p = 0.351). Also, r-square value of 0.116 indicates that model explains only about 11% of the variation in terms of daily rental count. This can be interpreted as there is no statistical evidence that commuters on working days react differently to weather than non-working days, and neither workingday nor the interaction with weather situations help explain bike rental pattern.

d. Backward Selection and Final Model

A full model including all predictors and interaction terms were reduced using backward selection to create a final model. The final model includes the following predictors: season, month (mnth), weather situation (weathersit), temperature (temp), humidity (hum), windspeed, interaction effects—season*temp, weathersit*temp. The final model demonstrates a strong fit to the 2011 dataset (adjusted R-squared = 0.856) and highlights the season–temperature interaction as a key driver of daily bike rental patterns.

e. Model Performance on test dataset—2012

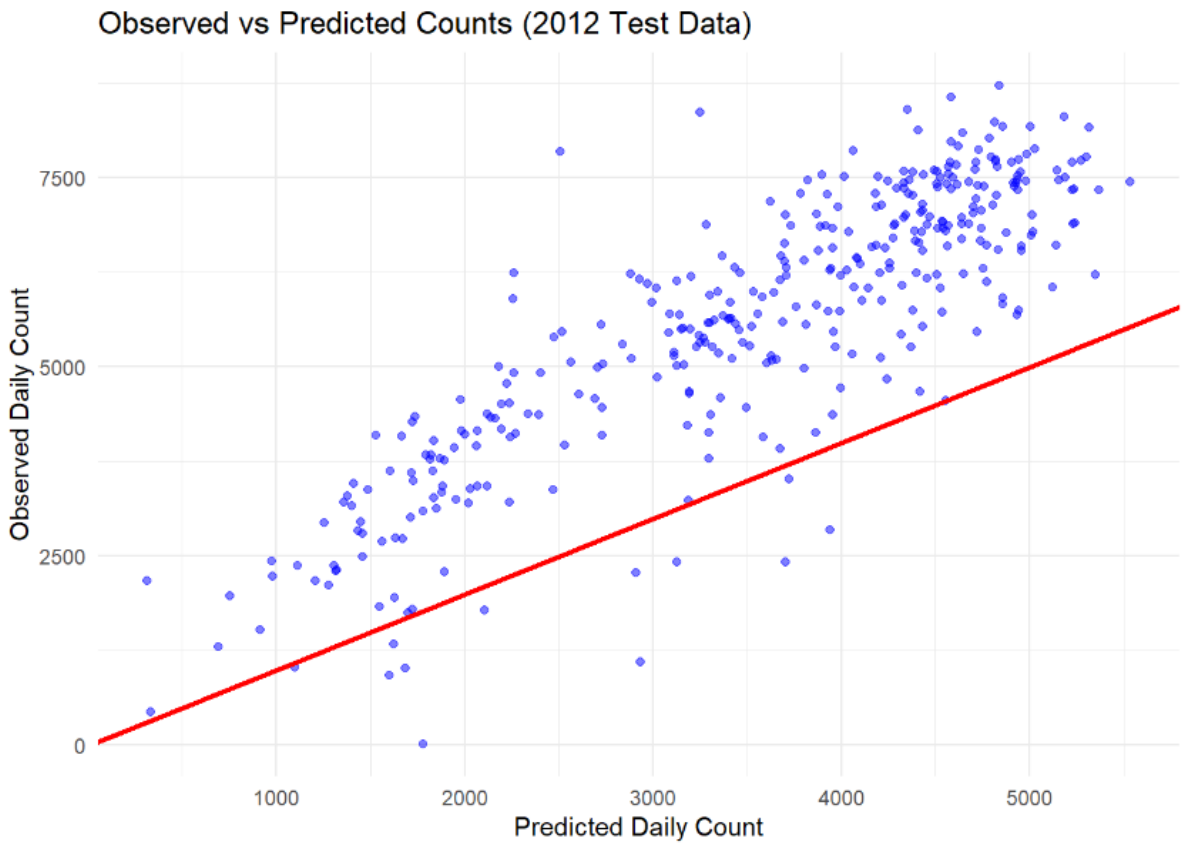
However, the final model showed a weaker fit on the 2012 test dataset, demonstrating that its predictive performance does not generalize beyond the 2011 dataset.

```
Model Performance on 2012: Test Data
RMSE : 2283.22
MAE  : 2114.76

Comparison with 2011 v 2012 Data
Training RMSE (2011) : 505.37
Test RMSE (2012)     : 2283.22
Difference           : 1777.85

Training MAE (2011) : 372.23
Test MAE (2012)     : 2114.76
Difference          : 1742.53
```

The RMSE on the test data (2283.22) is more than four times larger than the training error (505.37). This dramatic increase implies that the model does not generalize to new data. The MAE also increased dramatically from 372.23 to 2,114.76, suggesting that the model's predictions degrade severely when applied to 2012 dataset. The large increases in both RMSE (+1777.85) and MAE (+1742.53) indicate that the model does not generalize to 2012 data. This suggests that the final model is overfitting to 2011 patterns that fail to hold in 2012.



Additionally, the Observed vs. Predicted scatterplot proves the same, most observed data points (blue) lie above the prediction line (red).

Discussion

The analysis identifies clear relationships between daily bike rental counts and key predictors such as temperature, season, and weather conditions, whereas variables such as working day, humidity, and some weather indicators offer little to no explanatory power. The analysis also successfully identified interaction effects, particularly between temperature and season, which are important for capturing seasonal variation in bike rental behavior. Interaction models indicate that the effect of temperature on bike rental differs across seasons.

However, the final model's poor predictive performance on the 2012 test data indicates substantial overfitting to the 2011 training dataset. This suggests that simple linear models may not adequately capture nonlinear or complex relationships. Additional regularization or dimensionality reduction techniques may be necessary to reduce overfitting and improve generalization. Potential improvements include using more flexible modeling approaches that better capture nonlinear patterns and yield stronger predictive performance on the 2012 dataset.