# Bike Sahring Analysis

Qinmiao Deng Kira Shen Xinbei Yu Kaishuo Zhang Qi Zhao

Brown University

## Introduction

Bike sharing systems offer a seamless, automated process from membership to rental and return. By analyzing data on user behavior, we can gain insights into usage patterns and user preferences. This project utilizes the UCI Machine Learning Repository's bike sharing dataset, consisting of 17 features and 17379 data points, to predict the total number of rental bikes (cnt in the dataset). Our analysis aims to provide valuable insights for bike sharing providers and urban planners, helping to optimize the system for users and promote sustainable transportation.

## Exploratory Data Analysis

The variable instant, which track the position of each data point in the document, is dropped as it records the same information as default index. The variable dteday, which measures date of each data point documented is also dropped as our analysis does not have a time component. Based on the correlation heatmap, we dropped feature atemp (apparent temperature) from the model as it's 0.99 correlated with temp (temperature).


Figure 1. Fig. 1 Correlation between features

## Methods

1. **Linear Regression**, We use linear regression as our first method because the high interpretability of the linear regression model. To be specific, unlike other complex model or dimension reduction method, the simple linear regression model provides information on the important features that contribute to our prediction.
2. **Principal Component Analysis**. PCA can be used as a preprocessing step to improve the performance of machine learning algorithms. By reducing the dimensionality of the data, it can help reduce noise and improve the signal-to-noise ratio.

## Linear Regression

Target variable cnt is log-transformed because qq plot suggests a positive-skewed distribution, and a skewed distribution will violate the normality assumption of our linear regression model.

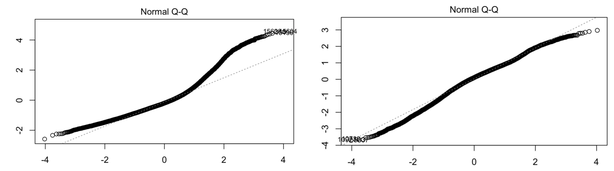


Figure 2. Fig. 1 qq plot before and after log-transformation

Forward feature selection is used because the number of features are relatively small. The final model is consist of season(1 for spring, 2 for summer, 3 for fall, and 4 for winter), yr(year of collection), hr(hour of the day from 0-23), holiday(whether the day recorded is a holiday or not), workingday(whether the day recorded is a holiday or not, neither holiday nor weekend), temp(normalized temperature of the day), hum(normalized humanity), windspeed(normalized wind speed), and whethersit(whether situation from 1 as clear to 3 as light snow, light rian...).

This project uses $R^2$ as our evaluation metrics. The second method used is principal component analysis(PCA), and proportion of variance explained is a key metrics in PCA, making the comparison between the two model easier

The model is also evaluated using cross validation with dataset being split in 80-20.

## Principal Component Analysis

PCA is an unsupervised learning approach to reduce the dimensionality of a dataset. In our case, we wish to transform the original covariates into abstract components of less dimensions but retain as much information as possible. We first perform PCA on the input matrix and then perform linear regression on the new matrix from PCA. We hope the new model will have higher predictive power as it addresses some issues regarding multi-collinearity and over-fitting

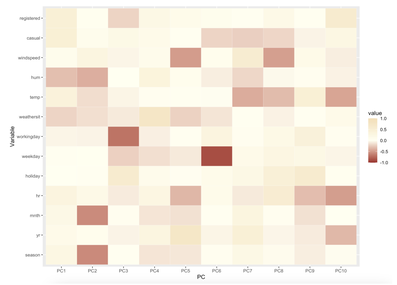


Figure 3. Fig. 3 heatmap for the top 10 PCs

## Results

| Model/Metrics | Simple Linear Model | Linear Model with PCA |
|---|---|---|
| $R^2$ | 0.471 | 0.591 $\alpha$ |
| MSE | 1.157 | 0.899 $\beta$ |
| MAE | 0.843 | 0.715 $\delta$ |

Table 1. A table caption.

- Linear model with PCA(the most predictive model): $R^2$ = 0.591, MSE = 0.899
- Three Different Global Feature Importances: AIC, BIC, Permutation Feature Importance
- People are less willing to ride a bike if the weather or temperature is not comfortable enough. But for other factors, such as working days, it somehow does not affect the bike rental

## Contribution

Qinmiao Deng

Kira Shen

Xinbei Yu

Kaishuo Zhang

Qi Zhao

## References

[1] J. Cai. *humidity: Calculate Water Vapor Measures from Temperature and Dew Point*, 2019. R package version 0.1.5.

[2] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[3] National Centers for Environmental Information. U.s. local climatological data, 2021.