Patterns in Blue Bike Usage: analysis at a per-bike level

Miraya Gupta '25 Data Science Major Capstone

WELLESLEY

Research Question

(Q1) What is the usage of a typical bike? **(Q2)** How is the number of bikes in use at a given hour correlated with season and day of week? **(Q3)** How well can we predict bike usage at a given hour?

Background

Blue bike is a company and a public service. It must therefore both stay profitable and serve customers for the best outcome. As a result, demand forecasting is important.

Data

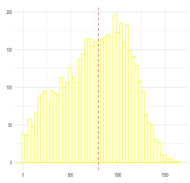
Source: Blue bikes rides data
Population: rides from April 2022-April 2023
Variables of interest: bikeid, starttime, stoptime

Manipulation

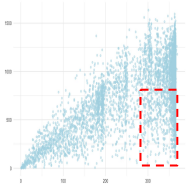
· data grouped by bikeid to obtain bike-level metrics
· month, day_of_week, season extracted from starttime
· data grouped by these variables and filtered by hour to get bikes_in_use at a given hour

(Q1) Bike usage

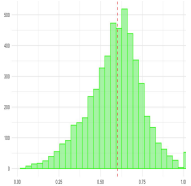
· A typical bike completes about 760 rides in a year
· A bike is utilised on ~60% of the days that it is available
· Number of rides varies positively with days a bike is available, however lots of bikes are underutilised



distribution of rides per bike



no. rides vs days bike is available

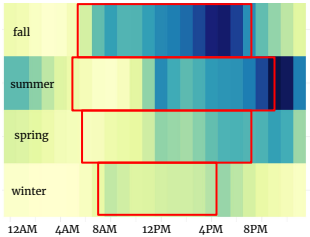


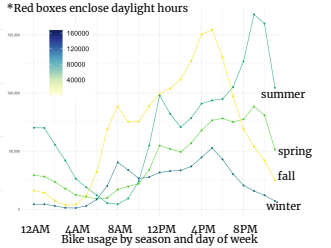
distribution of bike utilisation

(Q2) Variable correlations

A. season and time of day

· Bikes are most used in the summer, then fall, spring and winter
· Times of day that have maximum bike usage correlate strongly with sunrise and sunset times in that time of year
· There could be some linear association between the 2 variables

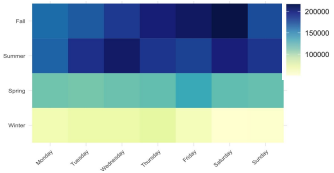




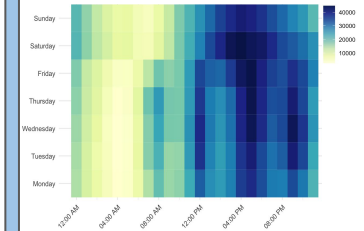
Bike usage by season and day of week

B. season and day of week

· In winter, weekday usage is higher than weekend indicating use for non-leisure commuting
· Fall Saturdays and summer Wednesdays have higher bike usage, potential interaction terms



Bike usage by time of day and day of week



C. day of week and time of day

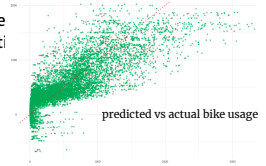
· Weekday mornings tend to be busier than weekend mornings
· More people seem to use bikes to commute back from work than to it
· On weekdays, usage peaks at 12PM, 5PM and 9PM
· Weekend usage peaks between 12PM and 8PM

(Q3) Making predictions

Model selection: Given observed trends, a linear model with interaction terms was fit to forecast how many bikes will be in use at a given hour. Since prediction accuracy is the goal over interpretability, the quadratic model is chosen.

model <chr>	coeffs <chr>	rmse <dbl>	Rsq <dbl>	AIC <dbl>
interaction1	hour,month,day,hour:month,month:day,hour:day	394	0.470	129765
interaction2	hour,month,day,hour:month	399	0.464	129837
quadratic	hour*month,month*day,hour*day,hour^2*month,hour^2*day	397	0.471	126620

Predictions: About 400 fitted values we < 0, which may be a quirk of the quadrati model. The model captures a decent amount of variation of the response.



Diagnostics: The model assumes that errors are uncorrelated, given the predictors, however in reality they are likely related to one another. The residual plot indicates non-constant variance and non-normality towards the tails.

Discussion & Conclusion

· Given low bike utilisation, it is likely that there are always more total bikes available than what is required. Therefore an analysis of the spatial distribution of bike usage would be a good next step to find opportunities for things like dynamic pricing or adding new stations
· The exploration of variables as well as the model results indicate that time information and weather information (like season) are important predictors for bike usage
· Given the assumptions violated in this linear model, it may be better to fit a non-parametric model like RF, or even a time series model

Limitations

Predictive value: Since the relevant variable **bikeid** is no longer collected post April 2023, this method may not be useful in making predictions beyond a few years from now.