# Final Project Report

## Summary

This analysis aims to determine what are the main factors that affect bike sharing's demand using the data from UCI Machine Learning Repository. The analysis applies methods including but not limited to: preliminary screening using exploratory data analysis (EDA), model fitting using linear regression with multiple predictors, model selection using BIC and ANOVA test, and model validation using assumptions assessment,Cook's distance, and multicollinearity (VIF). Our final model shows that there is evidence suggesting that bike sharing's demand is the highest in autumn. It also shows that bike sharing's demand is significantly higher during morning and evening compared to other time periods. Additionally, non-holiday's bike sharing demand is 14% higher than a holiday. When it comes to weather, in general, we have higher demand when the weather is good. More specifically, holding all other variable constant, when it rains, we expect a 82% decrease in the bike sharing's demand; When the humidity increases by 1%, we expect the bike sharing's demand to decrease by 1.3%; When the solar radiation increases by 1 $MJ/m^2$ , the bike sharing's demand is expected to decrease by 7%.

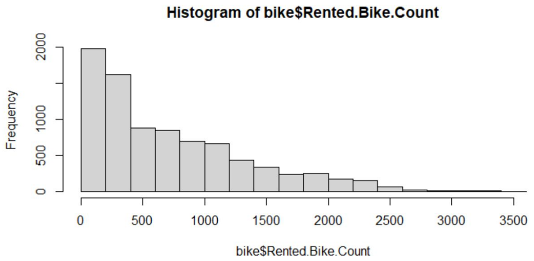
## Introduction

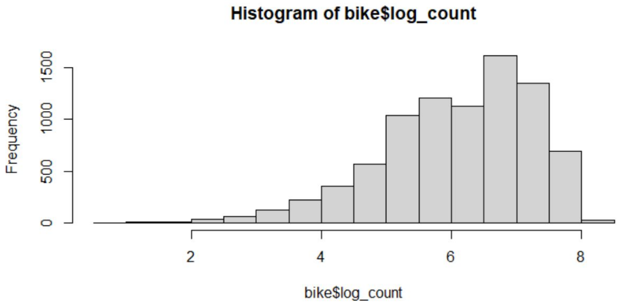
Bike sharing has now become the new generation of public transportation where the whole process from membership, rental and return are completed using your mobile app. Currently, there are about over 500 bike-sharing programs around the world. In order to provide a stable supply for the shared bikes, we want to know what are the main factors that affect the demand of bike sharing. More specifically, we developed the following research questions:

- Does bike sharing have higher demand during the holidays?
- Which season has the highest demand for bike sharing?
- What is a likely range for the difference in demand for bike sharing between holidays and non-holidays?
- Is there any evidence that the association between time of the day and bike sharing demand differs by holiday?

## Data

We obtained our data from UCI Machine Learning Repository. It contains every hour's sharing bike rented count from December 2017 to D ecember 2018 in Seoul, Korea. It contains 8760 observations and 14 variables with no missing value. Since *Temperature* and *DewpointTemp* contain repeated information, we will exclude *DewpointTemp* from this analysis. Moreover, to make the variable *hour* easier to interpret we regrouped and assigned a new variable *time* that contains 4 levels: early morning(0am - 6am), morning(6am - 12am), afternoon(12pm - 6pm), evening(6pm - 12pm). Additionally, to make the intercept interpretation more meaningful, we mean centered all continuous variables. Finally, the histogram of our response variable *rented bike count* is not normal, and we applied a log transformation and created a new variable *log_count*. Although the log transformed rented bike count is not perfectly normal, it is still better than before.





Below is a list of variables we use in this analysis:

- Continuous:
    - Log_count: log transformed count of bikes rented at each hour.
    - Temperature: Temperature in Celsius.
    - Humidity: Humidity in percent.
    - Windspeed:  Units Meters per second.
    - Visibility: measured in 10 meters.
    - Solar radiation: measured in $MJ/m^2$

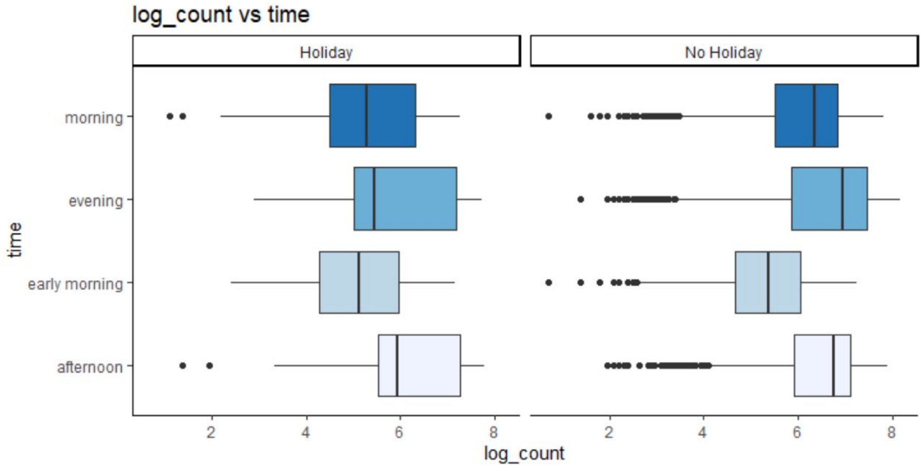
- Discrete:
    - Rain: rain indicator, 1 = rain; 0 = no rain.
    - Snow: snow indicator, 1 = snow ; 0 = no snow.
    - Season: contains four levels, spring, summer, autumn, winter.
    - Time: contains four levels, early morning, morning, afternoon, evening.
    - Holiday: indicate if that day is a Korean public holiday, 1 = holiday, 0 = not holiday.

## EDA

We used box plots to explore the relationship between discrete variables and the response variable. In general, the median log count of rented shared bikes is less when it is rainy compared to no rain. Similarly, the median log count of rented shared bikes is less when it is snowing compared to no snow. The box plot also shows summer has the highest median log count of rented shared bikes, autumn the second, spring the third and winter the fourth. Additionally, evening has the highest median log count of rented shared bikes according to the box plot and early morning is the least demanded time period. Finally, compared to holidays, non-holidays have a higher median log count of rented shared bikes.

When it comes to continuous variables, log rented bike count has an obvious increasing trend as temperature increases. Also, as humidity increases, the log rented bike count has a decreasing trend. However, according to the point plots, no clear trend or relationship can be identified for *Windspeed, solar radiation,* and *Visibility*.

The most interesting interaction we found is the interaction between time and holiday. More specifically, for a holiday, the afternoon has the highest median log count of rented shared bikes, whereas for a non-holiday, the evening is the most demanded time period.



All of our EDA findings are based on visual inspections, and we will test their statistical significance in the model section.

## Model

Model selection

*noInterModel:*

$$ln(rented.count) = \beta_0 + \beta_1 Temperature_i + \beta_2 Humidity_i + \beta_3 Windspeed_i + \beta_4 Visibility_i + \beta_5 Solar.radiation_i$$
$$+ \beta_6 Rain_i + \beta_7 Snow_i + \beta_8 Season_i + \beta_9 Time_i + \beta_{10} Holiday_i + \beta_{11} Time : Holiday_i + \epsilon_i; \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$$

We started with a multiple linear regression model *noInterModel* that contains all variables but no interactions. Based on the model summary, all predictors are significant except *Windspeed, Visibility,* and *Snow*. However, in this analysis, we want to understand if there is any evidence that the association between time of the day and bike sharing demand differs by holiday. Therefore we built another model *InterModel* that includes all predictors *noInterModel* had and in addition, also included an interaction term *time:holiday.* To test whether the new interaction term is statistically significant, we performed an F-test and the P value we got is 3.147e-14 which means predictor *time:holiday* is very significant. We then used stepwise BIC as our selection criterion to drop insignificant variables. The resulting BIC model suggested we can drop *Windspeed, Visibility* and *Snow*. We then performed F-tests to test whether these dropped variables are indeed statistically insignificant. According to the F-test result, *Windspeed* has a P

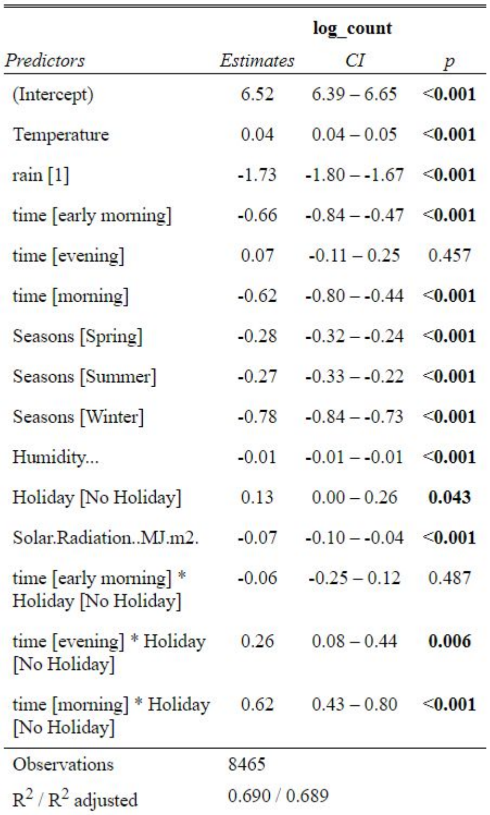
value of 0.75; *Visibility* is not significant with a P value of 0.99, and finally *Snow* is not significant with a P value of 0.86. To conclude, all three variables have large P values which means they are not significant and we decided to remove them from our model. After we dropped the insignificant variables, the model equation was:
*BIC_Model:*

$$ln(rented.count) = \beta_0 + \beta_1 Temperature_i + \beta_2 Humidity_i + \beta_3 Solar.radiation_i + \beta_4 Rain_i + \beta_5 Season_i + \beta_6 Time_i + \beta_7 Holiday_i + \beta_8 Time : Holiday_i + \epsilon_i; \ \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$$

We then started to perform model assessments. The normality assumption holds because most points on the Normal Q-Q plot still seem to fall on the 45 degree angle line. The independence and equal variance assumption hold as the points do not seem to follow an obvious pattern and the spread looks equal. The linearity assumption still holds as there is no obvious pattern for our continuous variables. In terms of potential outliers, all points are within the 0.5 cook's distance range and are below 0.012 leverage score on the standardized residuals vs. leverage plot. In the end, we check for multicollinearity. The interaction term seems to have inflated our VIF values. Predictor *Time* and *Time:Holiday* has VIF values around 30. All other predictors have VIF values under 5. Since we did not see clear violations during the model assessment and model validation, we decided to use *BIC_Model* as our final model.
*FinalModel:*

$$ln(rented.count) = \beta_0 + \beta_1 Temperature_i + \beta_2 Humidity_i + \beta_3 Solar.radiation_i + \beta_4 Rain_i + \beta_5 Season_i + \beta_6 Time_i + \beta_7 Holiday_i + \beta_8 Time : Holiday_i + \epsilon_i; \ \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$$

Model summary and CI:



| | | **log_count** | |
|---|---|---|---|
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | 6.52 | 6.39 – 6.65 | **<0.001** |
| Temperature | 0.04 | 0.04 – 0.05 | **<0.001** |
| rain [1] | -1.73 | -1.80 – -1.67 | **<0.001** |
| time [early morning] | -0.66 | -0.84 – -0.47 | **<0.001** |
| time [evening] | 0.07 | -0.11 – 0.25 | 0.457 |
| time [morning] | -0.62 | -0.80 – -0.44 | **<0.001** |
| Seasons [Spring] | -0.28 | -0.32 – -0.24 | **<0.001** |
| Seasons [Summer] | -0.27 | -0.33 – -0.22 | **<0.001** |
| Seasons [Winter] | -0.78 | -0.84 – -0.73 | **<0.001** |
| Humidity... | -0.01 | -0.01 – -0.01 | **<0.001** |
| Holiday [No Holiday] | 0.13 | 0.00 – 0.26 | **0.043** |
| Solar.Radiation..MJ.m2. | -0.07 | -0.10 – -0.04 | **<0.001** |
| time [early morning] * Holiday [No Holiday] | -0.06 | -0.25 – 0.12 | 0.487 |
| time [evening] * Holiday [No Holiday] | 0.26 | 0.08 – 0.44 | **0.006** |
| time [morning] * Holiday [No Holiday] | 0.62 | 0.43 – 0.80 | **<0.001** |
| Observations | 8465 | | |
| $R^2$ / $R^2$ adjusted | 0.690 / 0.689 | | |

Answer to research questions:

- According to our model summary, holidays in general have lower bike sharing demand. More specifically, compared to a holiday, the count of bikes rented on a non-holiday is expected to increase by 14% holding all other variables constant.
- Autumn has the highest demand for bike sharing. Based on the model summary, compared to in autumn, the counts of bikes rented in spring, summer, and winter are expected to decrease by 25%, 24%, and 54% respectively, holding all other variables constant.
- The confidence interval gives an answer to the third question we had. On a non-holiday, the count of bikes rented is expected to have an up to 30% increase. More scientifically speaking, the confidence interval indicates we expect the count of bikes rented to increase on a holiday with 95% CI (0%,30%)
- We do find a significant interaction between time and holiday. According to the model summary, compared to afternoon on a holiday, the count of bikes rented on a non-holiday evening is expected to increase by 29% holding all other variables constant; Compared to afternoon on a holiday, the count of bikes rented on a non-holiday morning is expected to increase by 85% holding all other variables constant.

## Conclusion

This analysis finds out important factors that affect bike sharing's demand and therefore it provides solutions to provide a stable supply for the shared bikes. When supplying shared bikes, bike sharing companies and organizations should consider weather information like rain, temperature, humidity and solar radiation. They should also consider information such as seasons, public holidays and the time of supply. Admittedly, this analysis is also limited. Rush hours and weather patterns can vary from city to city. The collected data is from Seoul, Korea. It cannot represent the bike sharing's demand in other cities. For example, compared to Seoul, Boston has a longer and colder winter. Therefore our conclusion might not be the case in Boston. Another limitation is we did not take weekends into consideration, because on weekends, people might use shared bikes on different time periods and for different purposes. To improve this study in the future, we can consider collecting data from other bike sharing organizations from various climate zones. Also, besides the holiday variable, we should also add an indicator for weekends.