Overall process:

**Suggestions for improvement:**

While our project achieved a root-mean-squared logarithmic error of 0.44, there are several improvements that could have been made to our model:

1. **Further feature engineering:** Our feature engineering involved modifying existing features to make them more meaningful. However, we could have taken this a step further by excluding certain features from our model and evaluating the impact of such a change. This could help redistribute the importance of features from the less important ones, which will now be excluded, to the more important ones.
2. **Developing entirely different models for registered and casual:** The same model was trained and tested separately for registered and casual users. A possible improvement to this is to run completely different models for registered and casual users. The two user segments demonstrate fairly distinct behaviors as seen in figure X. Casual users show a spike in usage on weekends while registered users demonstrate a more uniform pattern with slightly elevated usage on weekdays.

Binarized features

1. **Binarizing all features:** We binarized a subset of the features from the raw data but could have also experimented with binarizing all of them. Continuous features like actual temperature and feels-like temperature would first have to be made discrete. Making continuous values discrete by putting them in ranges could also improve variability in the dataset by clumping together sparse values.
2. **Accounting for correlation between features:** Features like actual temperature and feels-like temperature are highly correlated and this correlation is not accounted for in our model. It is unclear whether accounting for this will improve our model but it is something we could have experimented with. In order to correct for the correlations, we could have conducted a principle component analysis or PCA that transforms correlated features into linearly uncorrelated ones.[[1]](#footnote-1)
3. **Correcting for heteroscedasticity:** There is a lot of variability in the variance of our model. This is problematic as it may result in a higher variance. In order to deal with heteroscedasticity, we could have conducted a logarithmic transformation of the variables and weighted our regressor.

1. http://ordination.okstate.edu/PCA.htm [↑](#footnote-ref-1)