STAT 151A Final Project

Benjamin Lee, Stephanie Trinh, Zhi Long Yeo

2023-04-05

Model Question

Overview

We plan to analyze various datasets related to bike sharing data from different geographical locations (e.g. Seoul, Washington D.C., London etc) which contains information about the number of bikes rented on different days, along with weather conditions (temperature, humidity, etc) and miscellaneous information about the day these bikes were rented (weekends, holidays, etc).

Research question

The question we wish to answer using these datasets, is what various variables/factors can be used to predict the number of bikes that will be rented on a given day. We will also be testing the following hypothesis.

We also want to test the following hypotheses:

- B_temperature is different in different locations (We expect people in different areas to have different cold/heat tolerance)
- B_isHoliday in different locations
- Which is the most significant B? (We expect it to be B_temp, but expect B_windspeed to be significant too)

Practical Decisions

Our findings could help inform bike sharing companies into making better economic decisions.

i. Predictive maintenance: Our findings could help bike sharing companies predict the expected mileage (and thus wear-and-tear) on the bikes ahead of time. This could help them better schedule bike maintenance/replacements.

ii. Marketing decisions: A better understanding of the factors that affect number of users can provide valuable insights into marketing decisions, such as discounts and promotions. Companies can offer discounts during off-peak times or inclement weather to encourage greater usage.

Primary focus

Our primary focus is prediction accuracy. We will not be focusing on causal inference as we would have to control for confounding variables, and these confounding variables may not be captured in our datasets.

Data Overview

We are using the following datasets for data exploration and modeling:

- Bike Sharing in Washington D.C. Dataset (2011-2012)
- Seoul Bike Sharing Demand Data Set (2017-2018)
- London bike sharing dataset (2015-2017)

Each dataset contains the hourly count of rental bikes on each specific date, with additional information on weather and holiday schedules. Each observation corresponds to an hour of the day, resulting in observations being dependent on each other. Working with time series data will be one of the challenges of working with these datasets for linear modeling, but we plan to lessen the effects of dependence between observations by treating "hour" as a categorical variable.

It would probably be difficult for modeling on this data to be generalizable to a larger population and be applicable to other locations since the popularity of bike sharing and general trends varies across locations in a way that cannot be captured within the model. Working with these datasets will likely only provide us with a model appropriate specifically for Washington D.C, Seoul, and London.

Regarding additional features and data that we believe would be useful for analysis and modeling, we think that the inclusion of additional weather data like precipitation could be useful in improving the model. This is because it would make sense for the amount of rainfall to impact the number of people choosing to utilize rental bikes on a particular day. However, since not all the datasets we are using contain precipitation data, we are not currently planning to utilize rainfall as a feature in developing our model.

EDA

Data Cleaning

Before creating the following visualizations, we performed some preprocessing steps to clean up the data. Since we planned to use three different datasets for bike rentals in three locations, we first selected only columns in each dataset that was readily available in all other datasets and derived certain columns that weren't explicitly available (e.g. extracting hour data and workday data from timestamps). Some datasets required additional cleaning (e.g. un-normalizing temperature data in the Washington D.C. dataset, ensuring that all datasets used the same units of measurement for windspeed and defined the same months per season). We then standardized how categorical variables were encoded using logicals to make it easier in the future for modeling to treat those columns as categorical data. After ensuring all the datasets followed the same format, we merged the datasets together.

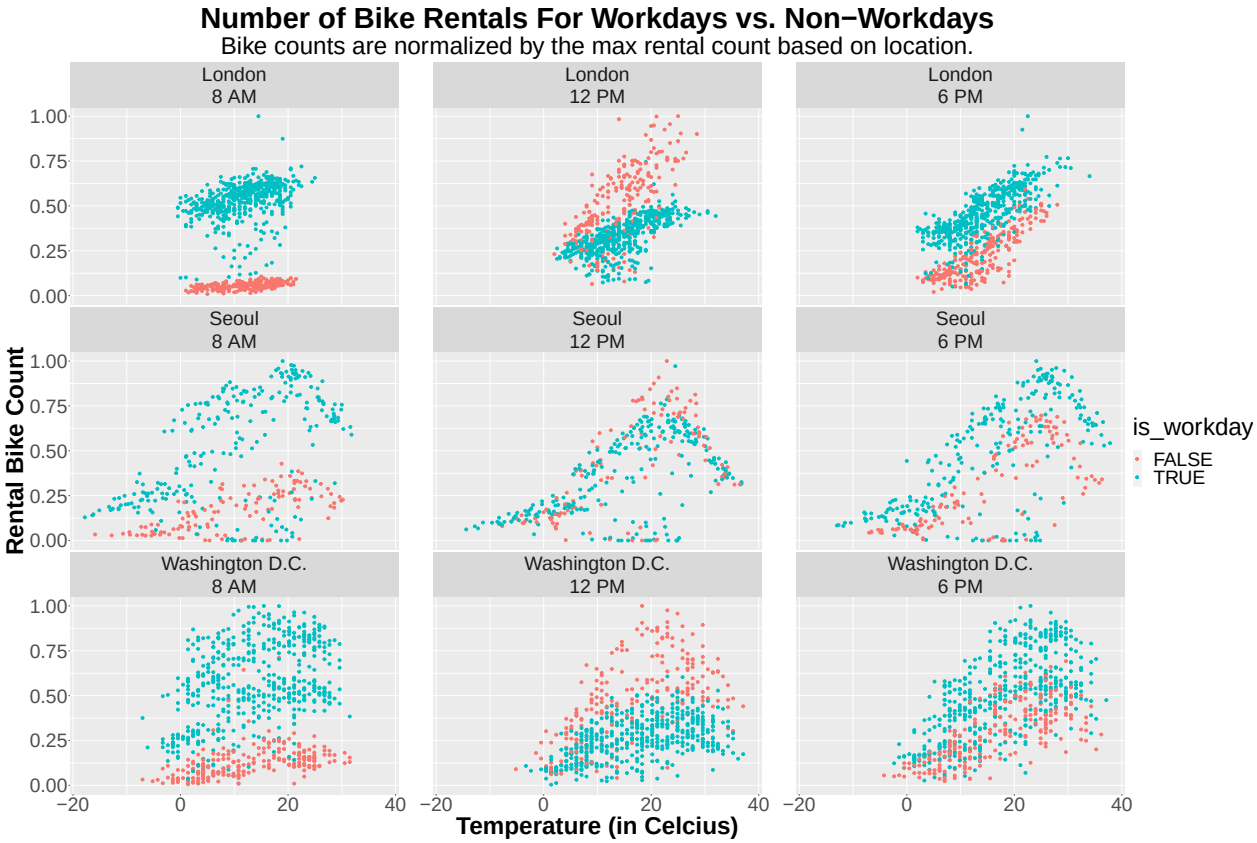
Visualization 1

cnt is the number of registered users normalized by the maximum of each location.

We see that in Seoul, during different seasons, there are different trends of count against temp. In winter we observe that the gradient of count with respect to temp is small and positive, in spring and fall, it is moderately positive and strongly negative in summer. This trend is not seen in London and much less obvious in Washington D.C. This gives us justification to use the interaction term $temp * location * season$ and all other interaction terms implied by the principle of marginality with the inclusion of this term.

We posit that this is because the annual temperature ranges in London and Washington DC is similar to that of Seoul's temperature range during Spring/Fall, because we see that the gradient of count against temp seems to be negative when temperature approaches the typical summer temperature of Seoul. However, it is difficult for us to determine a proper temperature cutoff to determine the 3 specific temperature regions, hence we think that using the season as a proxy for these different temperature categories is reasonable.

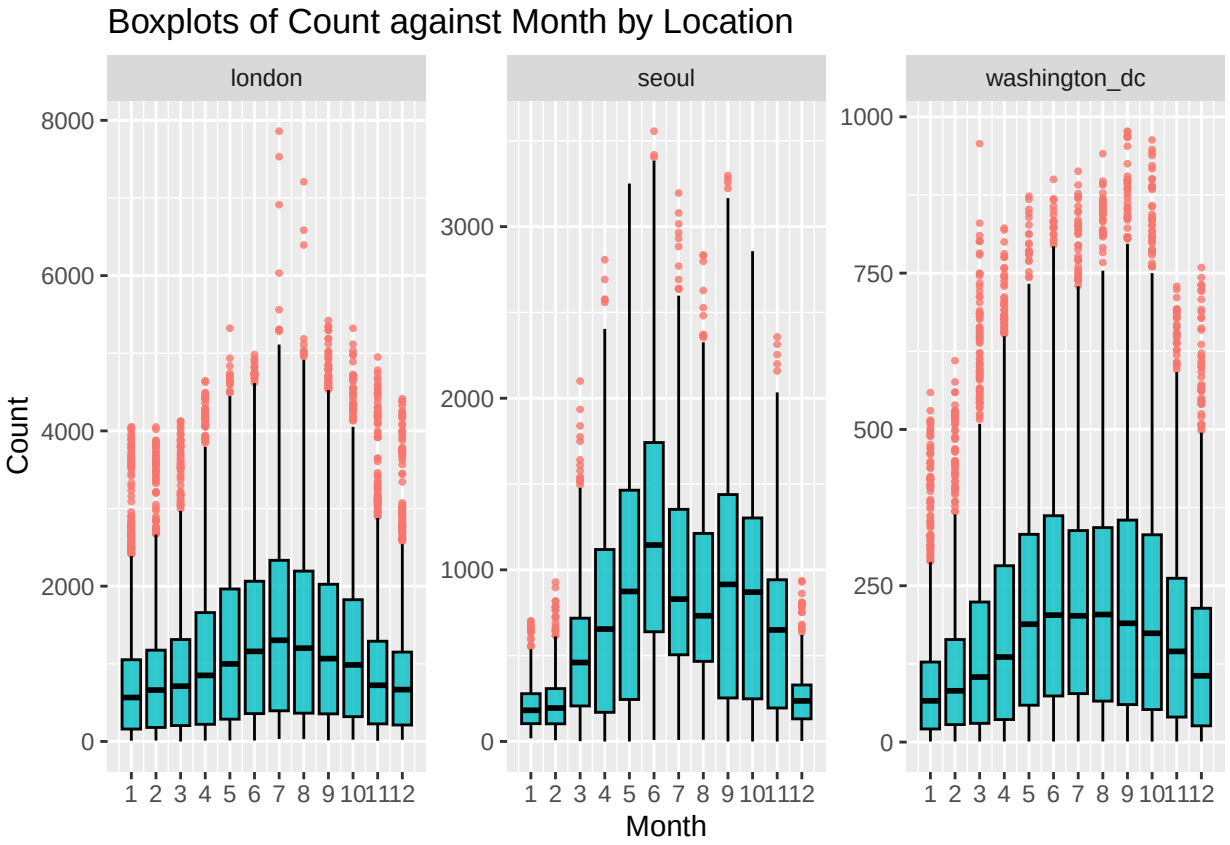
Visualization 2



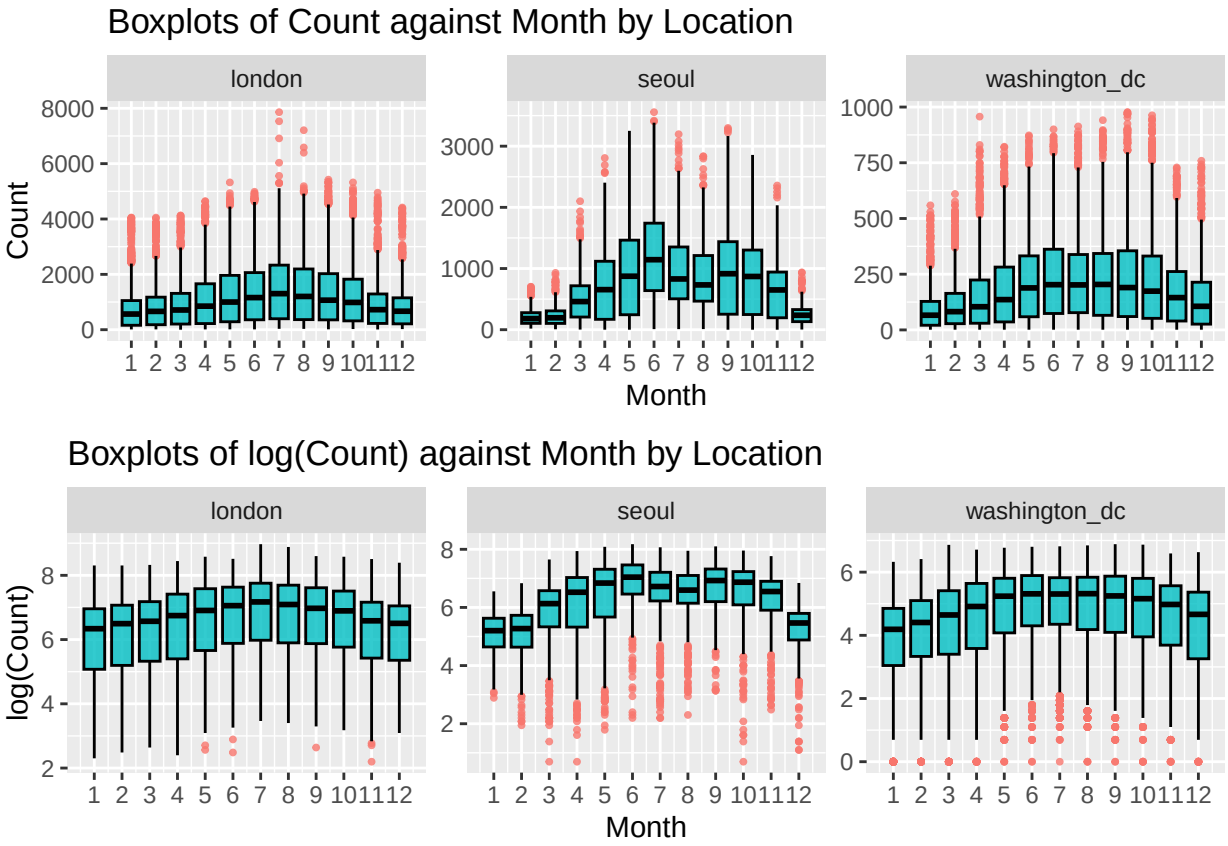
From this visualization, we can see there is a clear separation/grouping for the number of bike rentals between working days vs non-working days during hours of the day when we can expect many people to be commuting to and from work (e.g at 8 AM and 6 PM). In all three locations, we can see that there are more bike rentals during peak work commute hours on working days compared to non-working days. However, during other times of the day when we don't expect people to be commuting to and from work (e.g at 12 PM), there isn't as clear of a distinction between the effects of it being a workday vs non-workday. For instance, we can observe that Seoul at 12 PM follows very similar trends for the number of bike rentals against the temperature regardless of whether it is a workday or non-workday.

This visualizaton gives us some motivation to potentially use interaction term $temp * hr * is_workday$ along with all other relevant interaction terms implied by the principle of marginality. However, we'd like to note that while actually testing the model, we would like to analyze whether or not we actually need the triple interaction term, since we noticed that some of the graphs in the extended visualization (looking at all 24 hours of the day) appear to have the same gradient between workdays and non-workdays with a potential offset depending on the hour, while others appear to have differing gradients.

Visualization 3




```
## Warning: Removed 296 rows containing non-finite values (`stat_boxplot()`).
```



From the top row of boxplots we observe that the average count and, in particular, variance of bike users count differs across the months. The non-uniform variance violates the canonical assumption of homoscedasticity. Furthermore, we also observe that the distribution of count in each month does not appear to be normal, as evident from the presence of numerous outliers at high values of count. In the second row, we make an attempt to normalize the data using a log-transformation in an attempt to pull in high values of count, but our resulting boxplots still show that the distribution is not normal.

These observations motivate treating the data as count data and working from a Poisson GLM framework. A Poisson framework would, in particular, address the non-uniform variance of our data. We observe that variance of count in each month increases as mean increases, which is characteristic of Poisson processes.