Predicting Bike Share Ridership based on Climate Data in Seattle

Joey Rodriguez and Daniel Bhatti

2024-11-22

Introduction

Bike share has launched in many U.S. cities since its introduction in Washington, D.C. in 2010 (Biki). One iteration of bike share was *Pronto!* in downtown Seattle, Washington. From 2014 to 2017, 500 *Pronto!* bikes operated across 54 stations on the isthmus. The City of Seattle partnered with Socratica to collect system data during the operating window and made it publicly available via The City's open data platform. *Pronto!* fell short of the success realized by other similar schemes in the U.S. like Capital Bikeshare, Philly's Indego, and NYC's CitiBike. Researchers have used this system data to conduct a post-mortem analysis as dockless micromobility providers currently serve riders' demand in the region better than the *Pronto!* from before (University of Washington).

In this brief paper, we investigate the relationship between weather in the service area and daily ridership. In particular, we predict daily ridership based on weather data and time of year. After cleaning the data and analyzing candidates for the response variable, we took a stepwise approach to fitting a model. At first we considered only weather-related predictors, but we strengthened our model by adding temporal predictors. We considered various aspects when comparing models such as diagnostic plots, summary statistics, multicolinearity, systematic variable selection, and partial F-tests before validation. For our final model, we consider interpretations, limitations, and extensions worth more thorough investigation.

Exploratory Data Analysis

Data Cleaning

The data `trip.csv` and `weather.csv.xls` were downloaded from Kaggle (Pronto Cycle Share). The `trip` data frame contains 275,091 cases (or rides) and 12 variables describing each ride. These data were collected over 901 days from 13 October 2014 to 31 March 2017. The relevant variables from the original 12 in this dataset are `start_time` (day and time trip started, in PST) and `trip_duration` (time of trip in seconds). The `weather` data frame contains 689 cases (or days) and 21 variables describing the weather that day. These data were collected from 13 October 2014 to 31 August 2016, or 689 days. Notice that the dates covered by the `weather` data set are a proper subset of the dates covered by the `trip` data set.

We began by aggregating `trip` data for each of the 901 days. From the `trip` data frame, we created a new data frame called `ridership` that aggregates trips by day. At the end of this, `ridership` has 901 rows (days) and 3 columns (variables): `count`, `tripduration`, and `day_number`. Because the `trip` data covers 212 days after the last observation in the `weather` data, we want to keep only the observations in `trip` that match the observations in the smaller data frame, `weather`. We created our final data frame, `df`, by left-joining weather and ridership by `day_number`. The final data frame contains 689 rows (days) and 29 columns (variables). The variable names are listed in the table below with brief descriptions.

Table 1: Variable Descriptions (689 days, 29 variables)

Variable	Description
Max_Temperature_F	Maximum temperature (°F)
Mean_Temperature_F	Mean temperature (°F)
Min_TemperatureF	Minimum temperature (°F)
Max_Dew_Point_F	Maximum dew point (°F)
MeanDew_Point_F	Mean dew point (°F)
Min_Dewpoint_F	Minimum dew point (°F)
Max_Humidity	Maximum humidity (%)
Mean_Humidity	Mean humidity (%)
Min_Humidity	Minimum humidity (%)
Max_Sea_Level_Pressure_In	Maximum sea-level pressure (inches Hg)
Mean_Sea_Level_Pressure_In	Mean sea-level pressure (inches Hg)
Min_Sea_Level_Pressure_In	Minimum sea-level pressure (inches Hg)
Max_Visibility_Miles	Maximum visibility (miles)
Mean_Visibility_Miles	Mean visibility (miles)
Min_Visibility_Miles	Minimum visibility (miles)
Max_Wind_Speed_MPH	Maximum wind speed (MPH)
Mean_Wind_Speed_MPH	Mean wind speed (MPH)
Max_Gust_Speed_MPH	Maximum gust speed (MPH)
Precipitation_In	Precipitation (inches)
Events	Weather events (e.g., Rain, Snow)
temp_range	Temperature range (°F)
date	Date of the observation (%m/%d/%Y)
day_number	Days since 12 October 2014
total_trips	Count of total trips
total_durations	Sum of total duration for all trips (seconds)
average_durations	Average ride duration (seconds)
weekday_weekend	Encodes weekends: 1 if Saturday or Sunday, 0 otherwise
season	Encodes seasons: 0 if Spring, 1 if Summer, 2 if Fall, 3 if Winter
fall_winter	Encodes wet season: 1 if Fall or Winter, 0 if Summer or Spring

Notice that nine variables in our data dictionary were created from other variables:

- `temp_range` is computed from the difference: `Max_Temperature_F - Min_TemperatureF`
- `date` strips the day/month/year `"%m/%d/%Y"` from the full `starttime "%m/%d/%Y %H:%M"`
- `day_number` are the days beginning 13 October 2014, the first day of observation
- `total_trips` are the total trips recorded for each day
- `total_durations` are the total durations for all trips each day
- `avg_durations` was computed from the quotient: `total_durations` / `total_trips`
- `weekday_weekend` = 1 if `date` was a Saturday or Sunday and 0 otherwise.
- `season` = 0 if `date` in Spring, 1 if Summer, 2 if Fall, 3 if Winter according the the summer and winter solstices and the spring and fall equinoxes in the Northern hemisphere.
- `fall_winter` was created based on whether the `season` was Fall or Winter, which roughly coincides with the wet season in the Puget Sound Region from October to April (Seattle Times).

Understanding Outliers

The figure below plots daily bike ridership in Seattle, with the total rides taken each day in blue circles and the sum of the durations of the rides taken each day in red triangles. This figure suggests that outliers in total riders tend to coincide with outliers in ride durations. For instance, the day with the highest bike riders – 941 on Monday, April 20, 2015 – was also the day with the second highest sum of ride durations (359.7 hours). This is day 190, the largest outlier in the data set. It's not clear from look-up what caused bike ridership to be so high on this day; like much of the data gathered from the real world, many factors likely contributed to high ridership on this day.

36 days earlier on Sunday, March 15, 2015 was the second-wettest March day on record in the Puget Sound Region (Komo News). The rain was so severe that a mudslide occurred in Western Seattle. Knowing this, you'd expect March 15 to have been a bad day for cycling, and you'd be right: only 34 trips took place on this day with a combined ride duration of just 6.3 hours. This was the second worst day for cycling behind Sunday, December 27, 2015 with just 30 trips totaling 4.5 hours. This also happens to be the greatest leverage point by far in our final model (0.144). The coincidence between trips and durations explains the flattening of the data – the decrease in variation from the mean – observed in `avg_durations`.

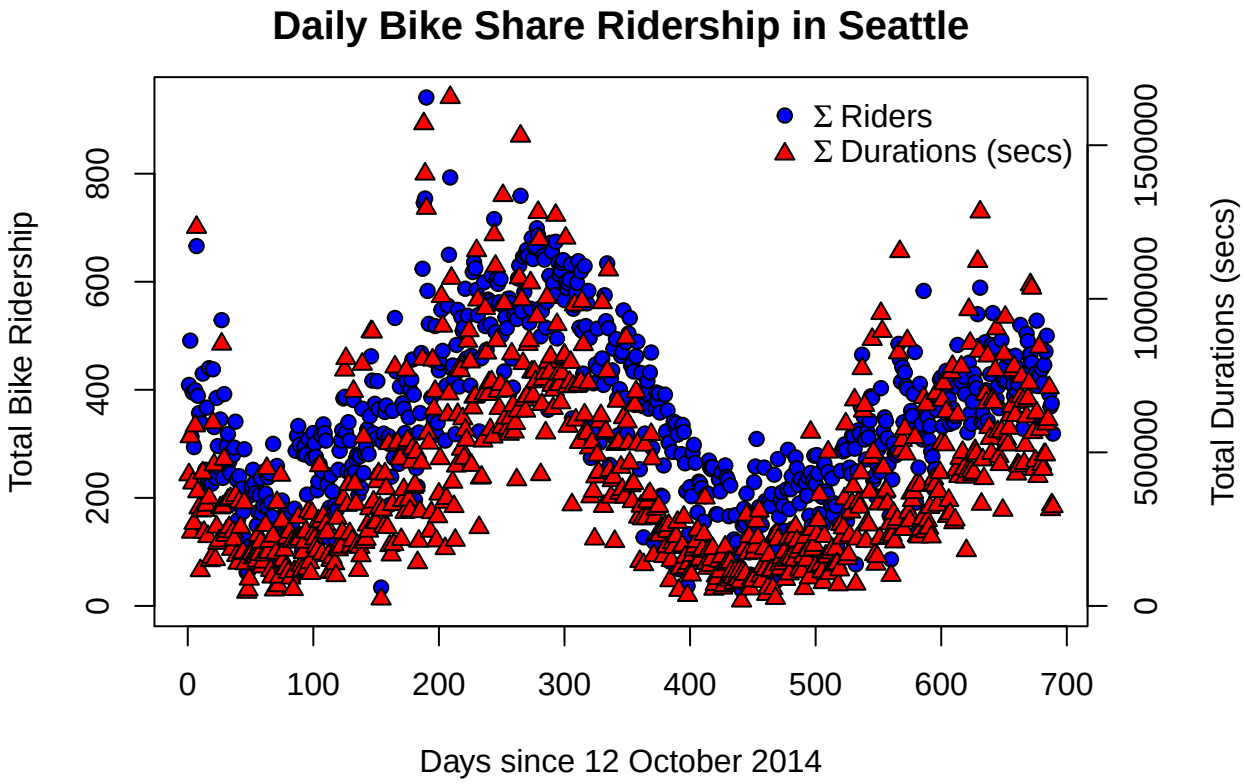

Figure 1: Daily Bike Share Ridership and Durations in Seattle.

Methods

Selecting the Response Variable

The three candidates for our response variable were created from the `trip.csv` data set, described in the `ridership` data frame, and merged into our final data frame: `total_trips`, `trip_durations`, and `avg_durations`. We briefly discuss the strengths of each response variable before a quantitative judgement:

- `total_trips` is the most intuitive measure for bike ridership on a given day. It directly answers the question "How many trips were there?" for a given day. It gives us a picture of how willing people in the service area were to pick a bike.
- `total_durations` gives a more complete picture for the ridership on a given day. Once a rider picked a bike, how long did they ride before docking it? This gives us a picture of how willing riders in the service area were to stay on their bikes once they mounted.
- `avg_durations` controls for the interaction between bike ridership and ridership durations. By dividing total ridership over total durations, we understand the willingness of those in the service area to picking up a bike *and* staying on it.

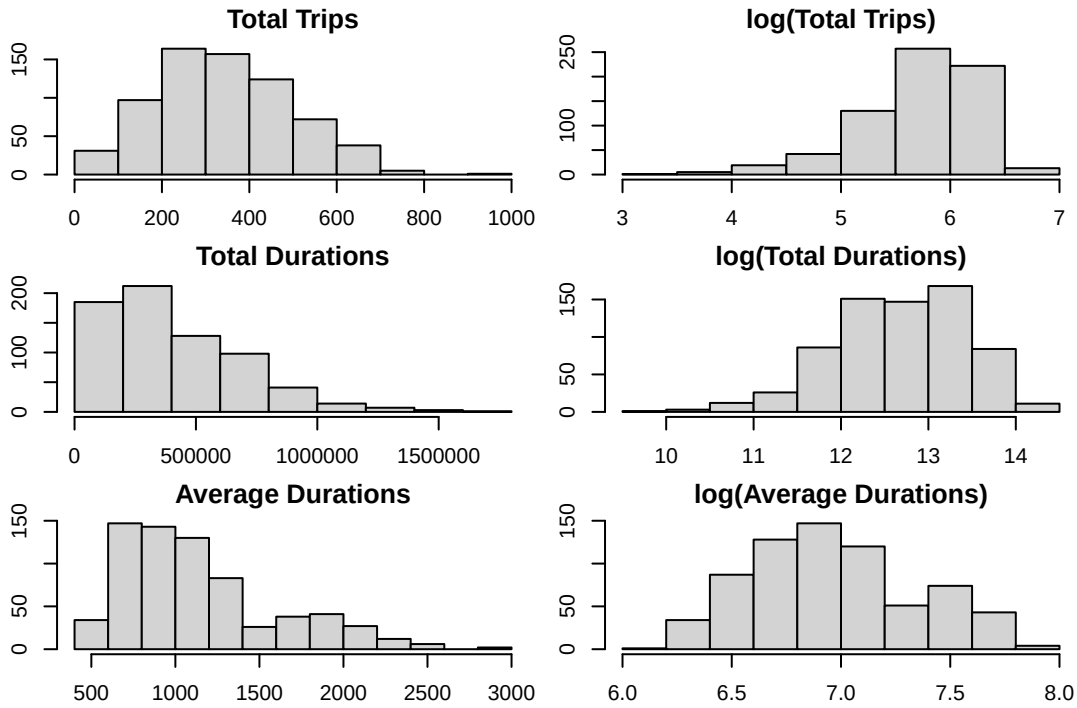

Figure 2: Variables Measuring Bike Share Ridership in Seattle

Note that `total_trips` and `total_durations` are highly correlated (>0.82). This is consistent with our discussion on outliers. Considering the shape and spread of the distributions, note that Average Durations is bimodal, Total Durations is right-skewed, and Total Trips is roughly normal with various units in use. It's

clear that Total Durations makes for the easiest interpretation without being transformed. Based on visual inspection, its the most normal appearing of the six distributions. We therefore use `total_trips` as our response variable going forward.

Selecting the Predictors

Table 2: Correlation between Weather Features and Total Trips

Feature	Min	Mean	Max
Temperature (°F)	0.640	0.750	0.786
Visibility (miles)	0.470	0.364	0.058
Dew Point (°F)	0.396	0.452	0.433
Humidity (%)	-0.648	-0.680	-0.579
Sea Level Pressure (in)	0.180	0.079	-0.065

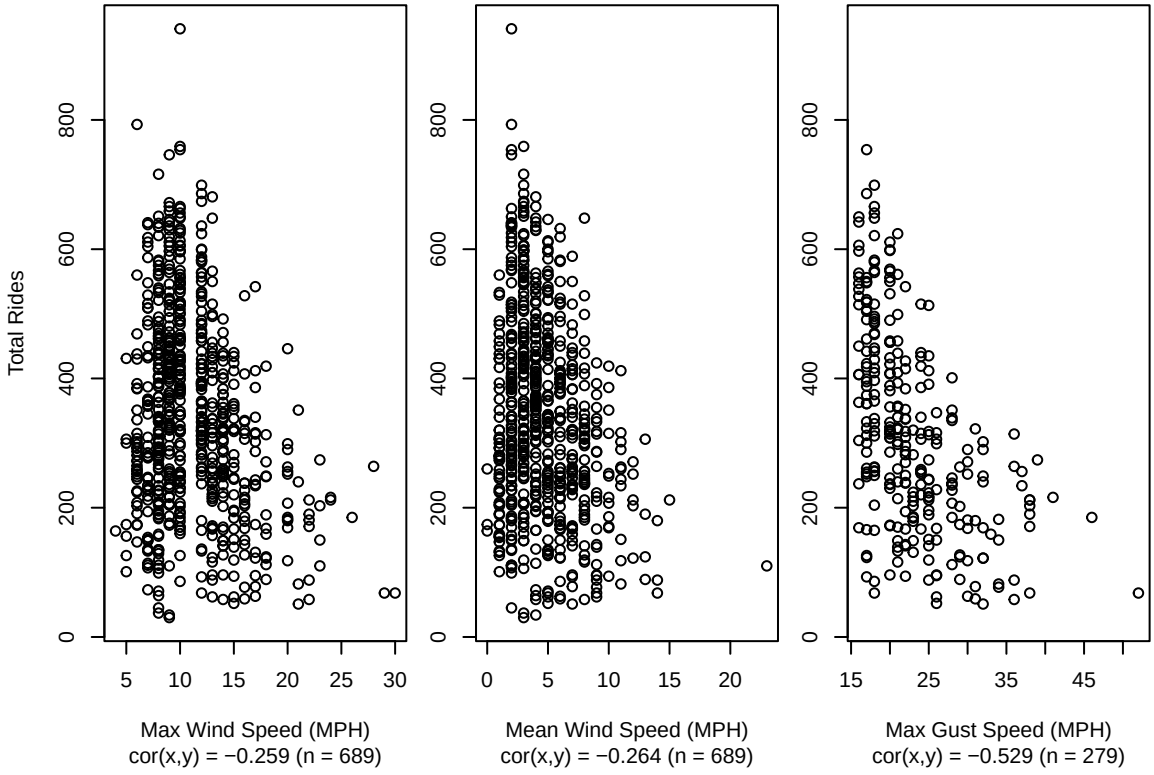

Figure 3: Effect of Wind Speed on Bike Share Ridership in Seattle.

For each of the continuous variables with recorded min, mean, and max — Visibility, Temperature, Dew Point, Humidity, and Sea Level Pressure – we calculate their correlations with the response Total Trips. By feature, it turns out that Max Temperature, Min Visibility, Mean Dew Point, Mean Humidity, and Min Sea Level Pressure have the highest correlation with Total Trips (Table 2). Total Trips' high correlation with Max Temperature may be explained by the temperature at midday, when the temperature is usually

highest and when people are prone to be out biking between commutes and leisure. We added each of these highest-correlation variables with the response to our baseline model.

Wind speed is a special case with variables Max Wind Speed, Mean Wind Speed, and Max Gust Speed. Though `Max_Gust_Speed_MPH` had higher correlation with Total Trips than either `Max_Wind_Speed_MPH` or `Mean_Wind_Speed_MPH`, it also had 410 missing values. We opted to add `Mean_Wind_Speed_MPH` which had the next-highest correlation with Total Trips (Figure 3). We also added `Precipitation_In` (a zero-inflated continuous variable) and the descriptive variable Events. We utilized `Events` in two ways. First, `Events` was converted to a dummy variable indicating whether there was an event for a given day day. There were 328 days with weather events and 361 without. Second, `Events` was converted to a dummy variable based only on whether there was rain or snow (precipitation) on that day.

Model Selection

Our baseline model based on initial findings from exploratory data analysis is described below:

$$\begin{aligned}\texttt{total\_trips}_i^{base} &= \beta_{i0} + \beta_{i1}\texttt{Mean\_Humidity}_i + \beta_{i2}\texttt{MeanDew\_Point\_F}_i \\ &+ \beta_{i3}\texttt{Mean\_Wind\_Speed\_MPH}_i + \beta_{i4}\texttt{Max\_Temperature\_F}_i + \beta_{i5}\texttt{Min\_Visibility\_Miles}_i \\ &+ \beta_{i6}\texttt{Min\_Sea\_Level\_Pressure\_In}_i + \beta_{i7}\texttt{Precipitation\_In}_i + \beta_{i8}\texttt{Events}_i\end{aligned}$$

We removed `Min_Visibility_Miles`, `Min_Sea_Level_Pressure_In` and `Events` because they were insignificant predictors for Total Trips. Our full model at this point ($RSE = 83.09$, $R^2 = 0.7141$, $R^2_{adj} = 0.712$, *all terms significant to* 0.01) is:

$$\begin{aligned}\texttt{total\_trips}_i^{full} &= \beta_{i0} + \beta_{i1}\texttt{Max\_Temperature\_F}_i + \beta_{i2}\texttt{MeanDew\_Point\_F}_i \\ &+ \beta_{i3}\texttt{Mean\_Wind\_Speed\_MPH}_i + \beta_{i4}\texttt{Mean\_Humidity}_i + \beta_{i5}\texttt{Precipitation\_In}_i\end{aligned}$$

Cursory transformations failed to produce superior models. Applying power transformations produced coefficients that were mostly close to one. Both forwards and backwards selection failed to produce a superior model, although they selected models similar to ours. We discovered a partial model ($RSE = 83.45$, $R^2 = 0.7112$, $R^2_{adj} = 0.7095$, *all terms infinitesimal*) with satisfactory summary statistics and diagnostic plots. The partial model is just the full model with the `Max_Temperature_F` term dropped.

$$\begin{aligned}\texttt{total\_trips}_i^{part} &= \beta_{i0} + \beta_{i1}\texttt{MeanDew\_Point\_F}_i \\ &+ \beta_{i2}\texttt{Mean\_Wind\_Speed\_MPH}_i + \beta_{i3}\texttt{Mean\_Humidity}_i + \beta_{i4}\texttt{Precipitation\_In}_i\end{aligned}$$

Based on the Partial F-test, the full model provides a better fit to the observed data over the partial model ($F = 6.9579$, $p = 0.0085$). However, Variance Inflation Factor (VIF) – measuring multicolinearity between the predictors and Total Trips – is abnormally high for `Mean_Humidity` ($VIF = 9.05$), `MeanDew_Point_F` ($VIF = 10.68$), and `Max_Temperature_F` ($VIF = 19.09$) in the full model.

It turns out that `Max_Temperature_F` has high correlation with `MeanDew_Point_F` (0.72) and `Mean_Humidity` (-0.67). The results from VIF analysis makes clear the confounding results in step-wise selection; In the forwards case, `Max_Temperature_F` makes the best 1-predictor model and it is never eliminated in successive models. The partial model has all VIF values near one. Thus, the partial model is the better model.

Finally, we strengthen our model by considering temporal and seasonal variation. Adding the dummy variable `weekday_weekend` accounts for weekly changes in ridership based on commute pattern changes in the downtown for leisure, work and school. Adding the variable `season` accounts for changes in ridership based on the solstices and equinoxes while adding `fall_winter` accounts for changes in ridership based only

on the solstices:

$$\begin{aligned}\texttt{total\_trips}_i^{temp} &= \beta_{i0} + \beta_{i1}\texttt{MeanDew\_Point\_F}_i \\ &\quad + \beta_{i2}\texttt{Mean\_Wind\_Speed\_MPH}_i + \beta_{i3}\texttt{Mean\_Humidity}_i + \beta_{i4}\texttt{Precipitation\_In}_i \\ &\quad + \beta_{i5}\texttt{weekday\_weekend}_i + \beta_{i6}\texttt{season}_i + \beta_{i7}\texttt{fall\_winter}_i\end{aligned}$$

Considering both the high correlation between variables `season` and `fall_winter` and the insignificance of the `season` variable, we opt to drop this variable from our final model:

$$\begin{aligned}\texttt{total\_trips}_i^{fin} &= 485.40 + 8.10\texttt{MeanDew\_Point\_F}_i \\ &\quad - 8.85\texttt{Mean\_Wind\_Speed\_MPH}_i - 6.20\texttt{Mean\_Humidity}_i - 103.49\texttt{Precipitation\_In}_i \\ &\quad - 50.23\texttt{weekday\_weekend}_i - 34.11\texttt{fall\_winter}_i\end{aligned}$$

Each regressor is significant to at least the 0.01 level, and the diagnostic plots are satisfactory. The residual plot moving-average looks flat and the data appear randomly scattered about the line excepting a few outliers, signaling linearity is satisfied. The normal quantile plot has few departures from the line, signaling normality is satisfied. The scale location plot is relatively flat showing constant variance across fitted values, signaling that homogeneity is satisfied.

To validate the model, we completed a preliminary cross validation with an 80%/20% train/test split ($RMSE = 84.80$, $MAE = 63.55$). We completed a more rigorous Leave One Out Cross Validation (LOOCV) ($RMSE = 80.00$, $MAE = 60.56$, $k = 689$). Notice that $RMSE$ from LOOCV roughly matches the RSE from our final model (79.64), suggesting that our model performs well on unseen data (Appendix).

Conclusion

Interpretations

For a given day with other variables held fixed, our model dictates ...

- For a 1% increase in Mean Relative Humidity, we expect 6.2 fewer trips.
- For a 1°F increase in Mean Dew Point, we expect 8.1 more trips.
- For a 1 inch increase in Precipitation, we expect 103.5 fewer trips.
- For a 1 MPH increase in the Mean Wind Speed, we expect 8.8 fewer trips.
- For a day falling in between (including) 22 September and 19 March, we expect 34.1 fewer trips than outside of those days.
- For a weekend day, we expect 50.2 fewer trips than on a weekday.

It makes sense for trips to drop as relative humidity rises, because higher relative humidity is uncomfortable to riders. Because the air is already saturated under high humidity, the perspiration on your skin is slower to evaporate. It makes sense for trips to increase as the Mean Dew Point increases, as a higher Dew Point generally signals warmer weather. It makes sense that any precipitation would flatten ridership, as biking in rain or snow could be more risky and uncomfortable without weather-proof infrastructure investments.

It makes sense that an increase in Mean Wind Speed would decrease ridership because faster winds increase wind chill. It makes sense that ridership decreases in the fall and winter, as this is when days are shorter, when Seattle's rainy season starts, and when the leaves fall, potentially making it so that biking is slippery and less scenic without weather-proof infrastructure investments. Lastly, an explanation for why ridership drops on the weekend is that commute patterns shift from school and work to primarily leisure.

Limitations and Extensions

Given that the climate of Seattle is unique from other major U.S. cities, our results are not readily generalizable beyond the Puget Sound Region. For example, the blistering heat in Houston, Texas during the summer must have a non-linear relationship with ridership. In Houston heat, a higher wind speed should raise rather than lower ridership. Additionally, it's not clear where the weather data was recorded from. The best information we have to go off of is from Kaggle, which reports that the weather was recorded in the service area. It's not clear whether time of day or location were random or scheduled.

There is a non-linear relationship between Mean Humidity and Mean Dew Point that might be exposed in a polynomial model. A more careful consideration of definitions and relations in meteorology could inform better variable selection. Because our data are recorded on successive days, the weather from the day before predicts the weather of the current day and thus the data are loosely dependent. Our data might be better modeled in the time-series setting. Another potential extension would be to use interaction terms between Temperature and Wind Speed, for instance (Kim 2016, An 2018). Overall, the model is impressive given the unpredictability of human behavior.

References

Biki, 2024. "History of Bikeshare," Available online at: https://gobiki.org/biki-leaks/history-of-bikeshare/.

University of Washington, 2019. "Bike-sharing in Seattle," UW News, October 7, 2019. Available online.

Pronto Cycle Share, 2016. "Pronto Cycle Share Dataset," Kaggle. Available online at: https://www.kaggle.com/datasets/pronto/cycle-share-dataset.

KOMO News, 2015. "Puget Sound region slogs through 2nd wettest March day on record," November 20, 2015. Available online at: https://komonews.com/news/local/puget-sound-region-slogs-through-2nd-wettest-march-day-on-record-11-20-2015.

Seattle Times, 2015. "Seattle-specific seasons and the different types of rain," Available online at: https://www.seattletimes.com/seattle-news/weather/seattle-specific-seasons-and-the-different-types-of-rain/.

Kim, Kyoungok, 2018. "Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations," Journal of Transport Geography, Elsevier, vol. 66(C), pages 309-320.

An, Ran & Zahnow, Renee & Pojani, Dorina & Corcoran, Jonathan, 2019. "Weather and cycling in New York: The case of Citibike," Journal of Transport Geography, Elsevier, vol. 77(C), pages 97-112.

Appendix

Code Used for Data Cleaning

```
trip = read_csv('pronto-cycle-share-trip-data.csv')
# map unique dates to integers starting at 1
# strips the date from its current format
trip$date <- as.Date(trip$starttime, format = "%m/%d/%Y %H:%M")
unique_dates <- sort(unique(trip$date)) # this collects unique dates
# this maps unique date to the integers, starting at 1
date_to_number <- setNames(seq_along(unique_dates), as.character(unique_dates))
# this adds the integer mapping as a column, day_number
trip$day_number = date_to_number[as.character(trip$date)]
```

```r
trip$count = 1 # this adds a one to each obs; useful for add
trip = dplyr::select(trip, count, tripduration, day_number)
```

```r
# construct new df, ridership, that aggregates trips by day
ridership = trip %>% group_by(day_number) %>%
  summarise(total_trips = sum(count),
            total_durations = round(sum(tripduration), 1),
            .groups = 'drop'); dim(ridership)
```

```r
weather = read_csv('weather.csv.xls')
# calculates temperature range for each day
weather$temp_range = weather$Max_Temperature_F - weather$Min_TemperatureF
# strips the date from its current format
weather$date <- as.Date(weather$Date, format = "%m/%d/%Y")
# maps unique date to the integers, like the chunk above
date_to_number <- setNames(seq_along(unique_dates), as.character(unique_dates))
weather$day_number = date_to_number[as.character(weather$date)]
weather = weather[,-1] # remove the old date
# this will be our data frame going forward
df = left_join(weather,ridership, by='day_number'); dim(df)
df$avg_durations = round(df$total_durations / df$total_trips, 1)
```

```r
df$weekday_weekend <- ifelse(weekdays(df$date) %in% c("Saturday", "Sunday"),1,0)
# Define a function to classify seasons based on actual start dates
get_season <- function(date) {
  year <- lubridate::year(date)
  spring_start <- as.Date(paste0(year, "-03-20"))
  summer_start <- as.Date(paste0(year, "-06-21"))
  fall_start   <- as.Date(paste0(year, "-09-22"))
  winter_start <- as.Date(paste0(year, "-12-21"))
  ifelse(date >= spring_start & date < summer_start, 0,  # Spring
  ifelse(date >= summer_start & date < fall_start,   1,  # Summer
  ifelse(date >= fall_start & date < winter_start,   2,  # Fall
  3)))  # Winter
}
# Apply the function to the 'date' variable
df$season <- sapply(df$date, get_season)
# Create fall/winter dummy: 1 if Fall or Winter, 0 otherwise
df$fall_winter <- ifelse(df$season %in% c(2, 3), 1, 0)
```

Summary Statistics for Final Model

```
##
## Call:
## lm(formula = total_trips ~ Mean_Humidity + MeanDew_Point_F +
##     Precipitation_In + Mean_Wind_Speed_MPH + fall_winter + weekday_weekend,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -221.24  -51.52   -3.87   45.97  442.98
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           485.3972    25.9669  18.693  < 2e-16 ***
## Mean_Humidity          -6.1999     0.3101 -19.995  < 2e-16 ***
## MeanDew_Point_F         8.1029     0.4673  17.338  < 2e-16 ***
## Precipitation_In     -103.4870    14.9002  -6.945 8.82e-12 ***
## Mean_Wind_Speed_MPH    -8.8492     1.1689  -7.570 1.21e-13 ***
## fall_winter           -34.1058     8.7685  -3.890  0.00011 ***
## weekday_weekend       -50.2268     6.7387  -7.453 2.76e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.64 on 682 degrees of freedom
## Multiple R-squared:  0.7378, Adjusted R-squared:  0.7354
## F-statistic: 319.8 on 6 and 682 DF,  p-value: < 2.2e-16
```

Diagnostic Plots for Final Model