Analysis of the 'London Bike Sharing Dataset'

David Escudero, Robert Beane, Kedrick Hill

Introduction:

For our project, we have chosen the 'London Bike Sharing Dataset', courtesy of user Hristo Mavrodiev on Kaggle[1]. The data was compiled from three sources, Transport for London for the daily numbers of new bike shares, freemeteo.com for the daily weather at the respective times, and gov.uk for identifying bank holidays. The original purpose for this data was for predicting future bike shares; we are also following suit and looking at which factors act as the best predictors for new bike shares. The data itself consists of ten factors:

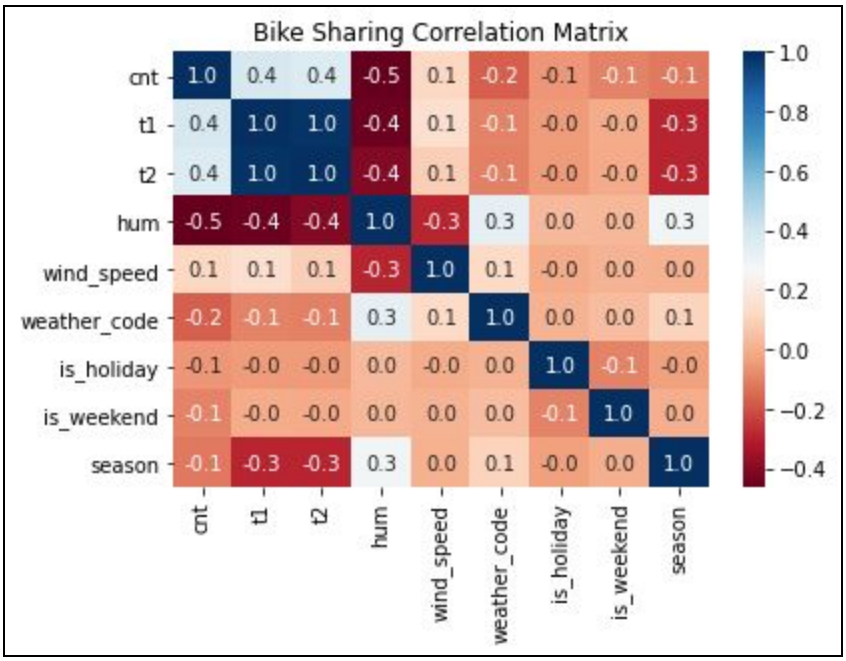
- timestamp - The time of a given record
- cnt - The count of new bike shares
- t1 - The temperature in Celsius
- t2 - The 'feels like' temperature, also in Celsius
- hum - percent humidity
- wind_speed - wind speed in km/h
- weather_code - weather category
    - 1 = clear
    - 2 = scattered clouds
    - 3 = broken clouds
    - 4 = cloudy
    - 7 = rain
    - 10 = thunderstorms
    - 26 = snow
    - 94 = freezing fog
- is_holiday - is it a holiday
- is_weekend - is it a weekend
- season - season coded 0-3
    - 0 = spring
    - 1 = summer
    - 2 = fall
    - 3 = winter

[1] London Bike Sharing Dataset

Thankfully, the data was already mostly curated for us before downloading, though we still made a couple of adjustments to the data prior to further exploration; notably, we adjusted wind_speed, t1, and t2 so that they would be in imperial units as opposed to metric. In addition, we split timestamp into two columns, date and time. For our hypothesis, we believed that bike sharing, based on season and weather (*t1*, *t2*, *hum*), will be lower when the weather is not good (cold or season is close to winter) and higher when the weather is good (warm or season is close to summer) which will be based on temperature.

Heatmap:

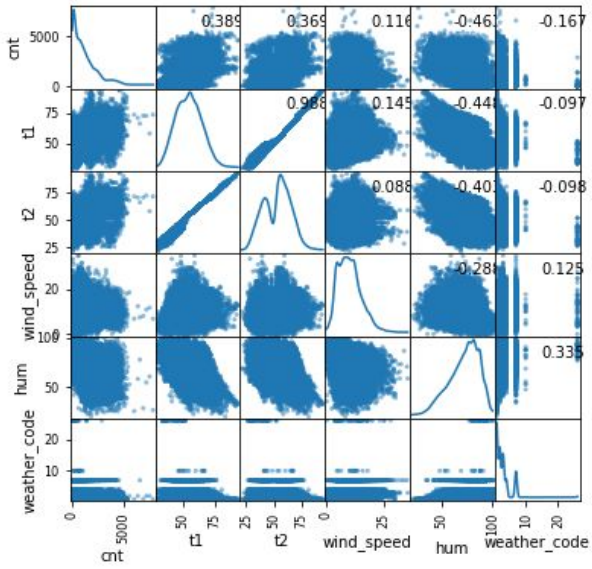
The first model that we decided to look at is the correlation matrix heatmap shown in the *bike sharing correlation matrix* below. As expected based on our hypothesis, we see that the humidity and temperature play a large role as predictors in the data. What we found surprising, however, was that weekend did not play as big of a role as we first anticipated nor did holiday. What the graph does show for us, is what data can be looked at closer as predictors. From the heatmap we can confirm that the outcome variable is *cnt* and the predictor variables are: *t1*, *t2*, *hum*, *wind_speed*, *weather_code*, *is_holiday*, *is_weekend*, and *season*. With our predictors

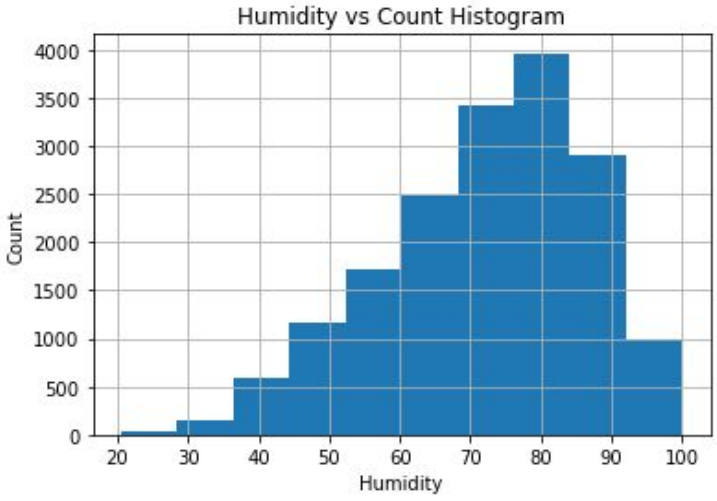


and outcome variables set, we decided to look more into the data by comparing and contrasting the predictors with the outcomes variable and predictors to predictors.

Data Visualization:

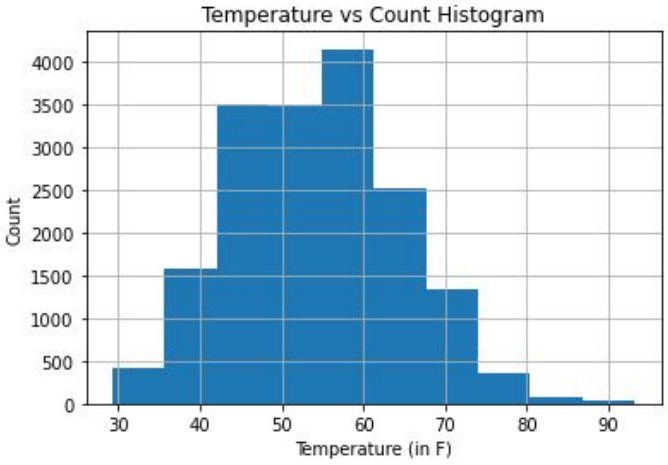
We used an assortment of visualization techniques throughout this project, one of which was the above heatmap. The first thing we did so we could find out which predictors had correlations with each other. Though most of our techniques worked, there were some that we found that did not really help us at all. One of the examples of something not working was the scatter matrix we attempted to create. We realized that our dataset contained too much data for a scatter matrix to be useful so we relied more on the heatmap. From the heatmap, we decided we wanted to take a look at humidity, temperature, and wind speed compared to the count. We decided a simple histogram would be the best choice for viewing the correlation of these predictors. From the humidity histogram, we can see that there was an obvious drop in count when the humidity hit 90 to 100 percent most likely due to



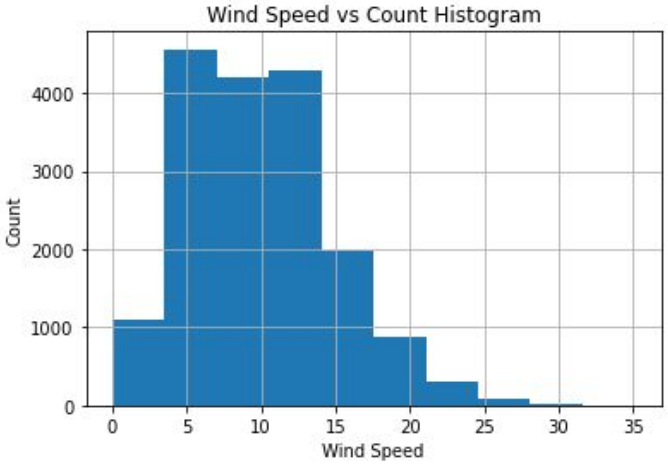


the fact that it was probably raining. When we initially created the histogram for temperature, we noticed that the temperature was in celsius and because most of us are used to using fahrenheit, we decided to change both of our temperature predictors (*t1* and *t2*) to fahrenheit. To do this, we created a simple function that is applied to every row in the *t1* and *t2* columns. After the function was applied, we got the following histogram for temperature related to the count. We found that most of the temperatures were between 45 and 60 degrees. From this we started noticing that temperature was a relatively big factor for usage. Sticking with weather, we wanted to look at how wind speed affected the usage of bikes. Similar to the situation with the temperatures, the dataset was using kilometers per hour instead of miles per hour, so we created another simple function. After it was applied, we were able to create another histogram for wind speed vs count; from this graph we can see that for most days, wind speed stayed at or below 15 mph. We were satisfied with these three histograms and decided we wanted to look at the average usage by hour to see what hours were the busiest. To do this, we created a separate dataset

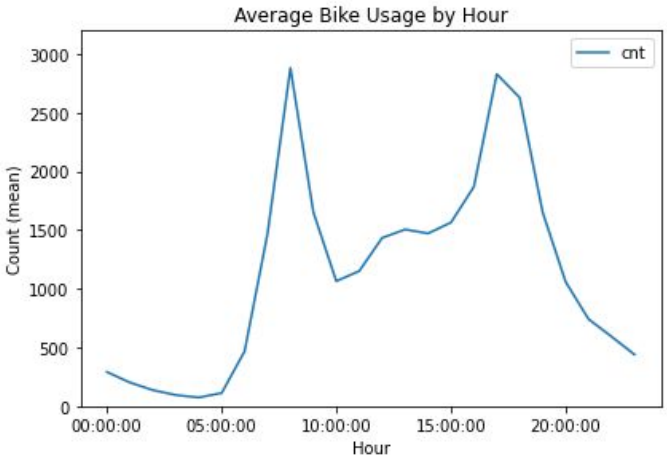
```
# Changing celsius to fahrenheit
def to_fahrenheit(x):
    return x * (9 / 5) + 32
```

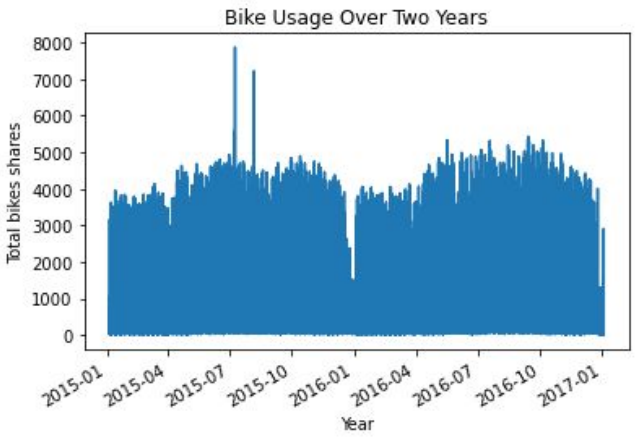



```
# changing km/h to mph
def to_mph(x):
    return x * 0.621371
```



that contained only *cnt* and *Time* making it easier to handle the data. After that we just took the mean count and plotted that per hour getting us a pretty nice looking graph. Looking at this graph we can see that there are obvious times that have spikes of usage. Those two spikes seem to be at 8 in the morning and 8 in the afternoon probably due to people going to and coming from work. We can also see that there is no point in time where the average is 0 either meaning people forgot to return their bikes or there are bikes missing. We also tried to add more x-ticks to show each hour but we could not figure out how to do it, we tried multiple methods, none of which worked. After doing this line graph, we decided we wanted to look at bike usage over the 2 years that the dataset supplies. To do this we used *pd.to_datetime* and *pd.Series* in which we got the following graph. Looking at this we can see that there were two obvious random spikes which could be due to an event (which we googled to see if there were any obvious events that took place at about that timeframe but couldn't find any obvious results) or just a random surge in usage. We also noticed that there was an obvious decrease in usage as the winter months approached, we can see this in the middle and at the end of the graph. From all the graphs that we created, we can see that there is indeed a





correlation to the weather (temperature, wind speed, humidity, etc.) and the usage of bikes. We can see that there were obvious decreases in the usage of bikes during low and high temperature days, days with high speed winds and days with high humidity.

Variable Selection:

Visualization of the data showed us how the data was distributed using a wide variety of models ranging from histograms to line graphs. However, being able to confirm our predictors and outcome variables is another way to initiate dimension reduction. The first model that we

```
Variables: t1, t2, hum, wind_speed, weather_code, is_holiday, is_weekend, season
Start: score=172387.67
Step: score=172387.67, remove None
```

applied to the predictors was the backward elimination, shown above, to check to make sure that all of our variables mattered and that there were not excess variables that did not contribute enough or assist enough with prediction. The results proved to us that the predictors that we had were all important and that we did not have an excess amount. Therefore, predictor reduction does not need to be applied to the data.

```
Variables: t1, t2, hum, wind_speed, weather_code, is_holiday, is_weekend, season
Start: score=175607.39, constant
Step: score=173095.20, add hum
Step: score=172553.67, add t1
Step: score=172450.73, add is_weekend
Step: score=172416.00, add season
Step: score=172400.03, add is_holiday
Step: score=172392.38, add wind_speed
Step: score=172388.97, add weather_code
Step: score=172387.67, add t2
Step: score=172387.67, unchanged None
```

We also applied forward and stepwise selection to check the most useful predictors by scoring them using the *AIC_score* method, shown above. As we hypothesized, humidity played an important role in predicting and influencing the other predictors by obtaining the highest score which was followed by: *t1*, *is_weekend*, and *season* as the top four predictors.

We wanted to see what effect regularization had on our data so we compared it to our other models that use linear regression. We found that using *Bayesian Ridge* resulted in most of our regression summary error calculations being higher than that of our linear regression. The only value that was lower in bayesian was Mean Absolute Percentage Error (MAPE) which only was by a small margin. With the results in mind, we chose that our base linear models were applicable and were not needed for regularization since we want lower error rates in the summary.

```
Regression statistics

                    Mean Error (ME) : 1.5866
     Root Mean Squared Error (RMSE) : 932.6830
          Mean Absolute Error (MAE) : 694.1534
        Mean Percentage Error (MPE) : -233.6171
Mean Absolute Percentage Error (MAPE) : 263.1245
```

```
Regression statistics

                    Mean Error (ME) : 1.6152
     Root Mean Squared Error (RMSE) : 932.7283
          Mean Absolute Error (MAE) : 694.1543
        Mean Percentage Error (MPE) : -233.6457
Mean Absolute Percentage Error (MAPE) : 263.1051
Bayesian ridge chosen Regularization:  0.0022576629683602784
```

Results:

In our exploration step, we found that there was a surprisingly high correlation between *t1* and *cnt* as well as *hum* and *cnt* through the use of a correlation matrix. Specifically, we found that we had correlation values of .4 and -0.5 respectively. This tells us that it appears as though higher temperatures and lower humidity will lead to more bike shares. From there, we decided to explore further and see the total usefulness of the variables through backward elimination as well as forward and stepwise selection. This analysis helped to confirm our hypothesis in that our highest two scoring predictors were *t1* and humidity. One of the other findings that came out of this was that among the lowest scoring groups we found *t2*, indicating that bike shares will stay relatively unaffected by the ‘feels like’ temperature. Upon first glance, it appears as though this would contradict the correlation table we made earlier. We account for this by knowing that there

is definite collinearity between itself and *t1*, an important predictor. Once we finished selecting our variables, we actually proceeded with creating a linear model of the number of new bike shares, taking into account the rest of the predictors.

Conclusion:

We set out to locate evidence of the influence that weather and seasons had on the total number of bike shares and if the variables in the data played a significant role. As was talked about in our results section, we learned a lot about how the data works and what predictors have a high influence on our outcome. We learned the distribution of data among top predictors, that holidays and weekends do not halt bike sharing usage and season is not as strong of a predictor as we initially thought.

We explored our data beyond what was already talked about in Data Visualization and have learned that there were things that we could have done differently and how we could improve that. One such way was using more predicting models and testing our trained data more. We spent much of our time exploring and using various models to show how the correlation plays an important role but did not do as much model prediction as we would have liked. To improve on that, we would have liked to train our data more using training and test sets to find an accurate model to predict what the number of bike shares could potentially be under certain conditions. Though as we mentioned, we did explore our data and verified our variables using variable selection and dimension reduction more than model prediction but this would not change what we have done so far as we believe it is an important part of the project and would do the same again.