Analysis of London Bike Shares

Muyang Shi, Shiyue Zhang, Yicheng Shen
Maths 245 - Applied Regression Analysis
Carleton College

Introduction:

Bike sharing programs have become increasingly popular during recent years. It serves as an alternative to public transportation or private vehicles and provides free or affordable access to bicycles for short-distance trips in many urban areas. Other benefits include transport flexibility and reduced noise and air pollution. In order to have a better understanding of how people are using bike sharing schema, we are interested in modeling the number of bike shares over some of its influential predictors such as temperature and holidays. There are some existing related studies on this topic, such as Yanyong Guo's research about bike sharing in Ningbo (Guo et al. 2017) and Dr. El-Geneidy's study about the relationship between the process of urbanization and bike sharing in Montreal (Faghih-Imani et al. 2014), but most of them focus on the influences of the location of bike sharing stations to the bike sharing usage. In this report, we will be exploring the relationship of bike-sharing numbers with time and weather or environment-related predictors.

Data Description:

The dataset we found includes 17414 observations of London shared bike from 2015 to 2017 (Mavrodiev, 2019). There are initially 10 variables in the dataset, and are included in List 1 below. The author of the dataset collected the bike share data from TfL's free transport data service[1], weather data from a website about weather reports[2], and the date information from the official U.K. bank holidays' website[3].

[1] https://cycling.data.tfl.gov.uk/ 'Contains OS data © Crown copyright and database rights 2016' and Geomni UK Map data © and database rights [2019] 'Powered by TfL Open Data'
[2] freemeteo.com
[3] https://www.gov.uk/bank-holidays

To capture the relationship between time and bike sharing usage, it is out of our capability to directly use timestamp variable in the dataset because there are simply too many levels. Thus, apart from the 10 variables included in the dataset, we have decided to add a categorical variable called datetime by dividing timestamp into 6 levels.

List 1 : Data Specification

Variable Name	Specification
timestamp	Timestamp field for grouping the data
cnt	The count of a new bike shares
t1	Real temperature in °C
t2	Temperature in °C "feels like"
hum	Humidity in percentage
wind_speed	Wind speed in km/h
weather_code	Category of weather 1 = Clear, mostly clear but have some values with haze/fog/patches of fog/ fog in the vicinity 2 = scattered clouds / few clouds 3 = Broken clouds 4 = Cloudy 7 = Rain/ light Rain shower/ Light rain 10 = rain with thunderstorm 26 = snowfall 94 = Freezing Fog
is_holiday	Boolean field: 1 if the day is a holiday
is_weekend	Boolean field: 1 if the day is a weekend
season	Category field meteorological seasons: 0 = spring ; 1 = summer; 2 = fall; 3 = winter.
daytime	New Divided Time levels: (24 hours) 1.Early Morning: 5:00 - 7:59 2.Morning: 8:00 -10:59 3.Day: 11:00-16:59 (baseline) 4.Early Evening: 17:00-20:59 5.Evening: 21:00-23:59 6.Late Night: 24:00-4:59

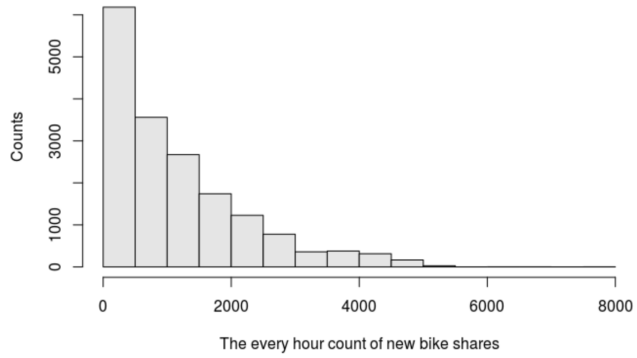
Results:

Exploratory Data Analysis

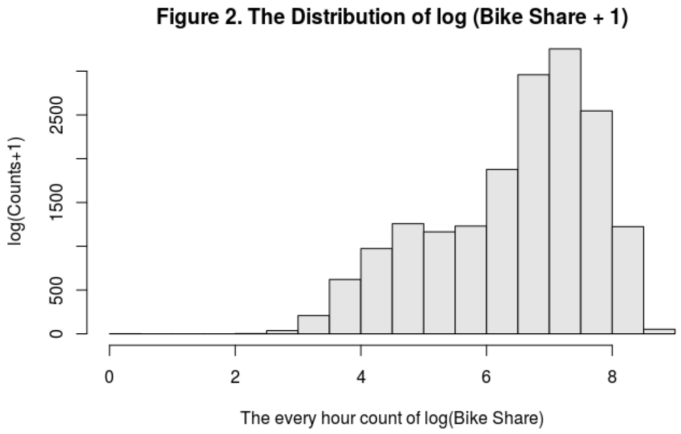
By conducting basic EDA on the dataset, we are able to visualize the distribution of the counts of hourly new bike shares (`cnt`). It displays a highly right-skewed distribution according to Figure 1. We can also observe this skewness through the five number summary in Table 1. The mean number of bike shares of 1143 is much greater than the median of 844. There is no missing data to be addressed. We have one potential outlier having zero number of bike share during one particular hour on March 29th, 2015, which is unusual compared with other cases. We will determine whether it should be removed through more analyses later.

	Five Number Summary of the count of new bike shares per hour				
Table 1.	Min	1st Quartile	Median	3rd Quartile	Max
Price (U.S. Dollars)	0	257	844	1672	7860

Figure 1. The Distribution of Bike Share



Our study first explores the scatterplot matrix of the response variable versus several other predictor variables, and we find that a natural log transformation on `cnt` will make a linear model more appropriate. Moreover, the distribution of `cnt` after a log transformation shows much less skewness and clusters of data. The log transformation version of `cnt` can be observed in Figure 2.



The Inappropriate Poisson Model:

We first consider using poisson distribution model since we want to predict the number of new bike shares in each hour. However, even though the p-value of all the predictors are statistically significant at 5% level, the goodness of fit test result is approximately 0. This means that the poisson model assumptions of independent trials and/or the constant probability of success across trials are not met. This indicates that there could be variations that our set of predictors are not capturing. For example, there might be also more people using bike sharing during rush hours and fewer people using bike sharing when late at night, which implies that the trial independence assumption is not met and we cannot use a poisson distribution to model our data. Since Poisson model does not fit in this situation, we resort back to multiple linear regression model to describe and predict the dependent variable `cnt` over the explanatory predictor variables we have in the dataset.

MLR model with a factor variable of four levels to represent daytime:

In order to apply the "`timestamp`" variable as an analyzable predictor in the model, we first divide one day into four consecutive time intervals, which are "Early Morning" ranging from 5 a.m. to 10 a.m., "Day" ranging from 10 a.m. to 4 p.m., "Early evening" ranging from 4 p.m. to 10 p.m., and "Night" ranging from 10 p.m. to 5 a.m.. We therefore constructed this new categorical variable called "`daytime`" with four levels.

In order to adjust for the "0" case in the `cnt` values, our initial MLR model is formed by constructing a multiple linear regression model of `log(cnt+1)` against nine available predictors in the dataset: `daytime, t1, t2, hum, wind_speed, weather_code, is_holiday, is_weekend` and `season`. Checking the case influence statistics shows that there is no unusual cases that have extremely high Cook's distance, leverages or standard residuals. Therefore, no extreme case is influential enough to be removed.

All the terms except `t1` show strong significance to the initial full model. Moreover, the high GVIF values (55.04, 52.69) indicate that `t1` and `t2` also exhibit strong collinearity with each other, which also makes sense intuitively: the real temperature is likely to have a linear relationship with the temperature that people feel like. By conducting an anova test of model having `t1` or `t2` with the reduced model, we find that `t2` is a much more significant predictor to the model than `t1` (p-value: 0.000215 < 0.09297). Therefore, we firstly remove `t1` from the model and the full model then possesses an adjusted R^2value of 58.03% with 17403 degrees of freedom.

By using forward selection, we try to add some significant interaction terms to the model. Everytime we compare the full model having one more interaction term at a time with the reduced model and conduct anova test to determine the significance of that interaction. In summary, ten interaction terms are significant enough to be added into the full model, which are `t2 * season, t2 * is_weekend, t2 * is_holiday, t2 * weather_code, t2 * daytime, weather_code * season, wind_speed * season, daytime * is_weekend, daytime * is_holiday, daytime * season`, most of which involve interactions with `t2` and `daytime`.

The full model with daytime as 4 levels is shown below in Table 2. This model has an adjusted R^2value of 67.7% with 17385 degrees of freedom.

Table 2.	**Terms**	**Estimates**	**SE**	**p-value**
(Intercept)	beta_0	7.6648243	0.0660398	0.0000000
daytimeEvening	beta_1	0.3859505	0.0497571	0.0000000
daytimeMorning	beta_2	0.2495973	0.0498311	0.0000006
daytimeNight	beta_3	-2.0162343	0.0469891	0.0000000
t2	beta_4	0.0259905	0.0026317	0.0000000

hum	beta_5	-0.0176034	0.0005838	0.0000000
wind_speed	beta_6	0.0050538	0.0012884	0.0000880
weather_code	beta_7	-0.0116945	0.0066914	0.0805319
is_holiday	beta_8	-0.1499184	0.1112040	0.1776313
is_weekend	beta_9	0.3053034	0.0353224	0.0000000
season	beta_10	0.2058118	0.0190587	0.0000000
t2:season	beta_11	-0.0024747	0.0009754	0.0111830
t2:is_weekend	beta_12	0.0122174	0.0019108	0.0000000
t2:is_holiday	beta_13	0.0375171	0.0069697	0.0000001
t2:weather_code	beta_14	-0.0013041	0.0004030	0.0012145
daytimeEvening:t2	beta_15	0.0132609	0.0024916	0.0000001
daytimeMorning:t2	beta_16	0.0106569	0.0026965	0.0000778
daytimeNight:t2	beta_17	0.0332428	0.0025044	0.0000000
weather_code:season	beta_18	-0.0066841	0.0022729	0.0032780
wind_speed:season	beta_19	-0.0056877	0.0006350	0.0000000
daytimeEvening:is_weekend	beta_20	-0.8656177	0.0347161	0.0000000
daytimeMorning:is_weekend	beta_21	-1.8861264	0.0370449	0.0000000
daytimeNight:is_weekend	beta_22	0.3238117	0.0341009	0.0000000
daytimeEvening:is_holiday	beta_23	-0.7953769	0.1073415	0.0000000
daytimeMorning:is_holiday	beta_24	-1.8581104	0.1154893	0.0000000
daytimeNight:is_holiday	beta_25	0.1820553	0.1055084	0.0844544
daytimeEvening:season	beta_26	-0.0504417	0.0149958	0.0007707
daytimeMorning:season	beta_27	-0.0737964	0.0159589	0.0000038
daytimeNight:season	beta_28	-0.0509503	0.0148827	0.0006197

The linearity and constant variance assumptions are well met by observing the residual plot of the fitted values. The points (residuals) are scattered evenly above and below the 0-reference line. The variation of residuals' magnitudes above and below the 0-reference line are also generally constant. However, several predictor variables like `hum` and `wind_speed` show a fan-shaped distribution of residuals, violating the constant variance assumption.

The normality assumption is not also perfectly met according to the normal Q-Q plot. The concave down shape and the left long tail in the plot indicate the nonnormality of the distribution. It is not suggested to do prediction with this model, but for description it is good enough for our analysis.

The independence assumption of the model is satisfied through the way that the dataset was collected. There is no spatial correlation between each case since the response variable measures the number of bike shares of the entire city. The time correlation is also not strong since we have already considered time as a very significant categorical variable in the data.

The adjusted R^2 of the first model is not very large and the model does not satisfy all of the assumptions of a multiple linear regression model. In order to explain more variation in our response variable `cnt`, we continue to refine our predictor variables and possible interactions in the model.

MLR model with a factor variable of six levels to represent daytime:

The variable `daytime` in our previous model only has 4 levels to represent a day (variable `timestamp`), which gives us an adjusted-R^2 of 0.67. We would like to improve our model fit and in order to do so, we divide the `timestamp` variable into 6 levels as the following sequence. We divide one day into six consecutive time intervals, which are "`Early Morning`": from 5 AM to 8 AM, "`Morning`": from 8 AM to 11 AM, "`Day`": from 11 AM to 5 PM, "`Early Evening`": from 5 PM to 9 PM, "`Evening`": from 9 PM to midnight, "`Late Night`": from midnight to 5 AM.

We construct this new categorical variable called "`daytime`" with six levels. Our initial model is formed by constructing a multiple linear regression model of `log(cnt+1)` against nine predictors: `daytime, t1, t2, hum, wind_speed, weather_code, is_holiday, is_weekend` and `season`. SImilarly as shown above, `t1` and `t2` are high collinear since they have high GVIF values. Therefore, we only keep `t2`, which shows what the temperature feels like, in our model. This model possesses an adjusted R^2value of 71.4% with 17400 degrees of freedom.

Same as the previous model selection process, there is no outlier that needs to be addressed. By using forward and backward selections, we again try to add significant interaction terms and predictors to the model, and removes the insignificant variable (for example, `season` in this model). We also add the term hum^2 because in reality, there would be an "optimal" humidity where people would like to ride a bike. The scatterplot matrix also shows that the relationship between hum^2 and $log(cnt+1)$ appears to be more linear. This time we also compare the full model with one interaction term at a time with the reduced model and conduct anova test to determine the significance of the interaction. In summary, eight interaction terms are significant

enough to be added into the full model, which are `t2:is_weekend, t2:is_holiday, t2:weather_code, daytime:t2, daytime:is_weekend, daytime:is_holiday, daytime:hum, daytime:wind_speed`. Again, most of the interactions are related to `t2` and `daytime`, meaning that their effects on the response `cnt` can be influenced by other predictor variables.

The full model with daytime as 6 levels is shown below in Table 3. This model has an adjusted R^2value of 81.3% with 17373 degrees of freedom.

Table 3.	estimate	std.error	statistic	p.value
(Intercept)	6.9729404	0.1045205	66.7136226	0.0000000
daytimeEarly Evening	0.4652465	0.0917057	5.0732566	0.0000004
daytimeEarly Morning	-0.8376233	0.1545350	-5.4202839	0.0000001
daytimeEvening	-1.1334805	0.1183069	-9.5808493	0.0000000
daytimeLate Night	-2.9065040	0.1217854	-23.8657870	0.0000000
daytimeMorning	-0.0616086	0.1174143	-0.5247114	0.5997905
t2	0.0278037	0.0016290	17.0680463	0.0000000
hum	0.0118782	0.0029188	4.0695846	0.0000473
I(hum^2)	-0.0001709	0.0000223	-7.6544349	0.0000000
wind_speed	-0.0088581	0.0010646	-8.3203754	0.0000000
weather_code	-0.0217557	0.0028596	-7.6078335	0.0000000
is_holiday	-0.1616908	0.0862512	-1.8746492	0.0608578
is_weekend	0.2104341	0.0272491	7.7225987	0.0000000
t2:is_weekend	0.0166581	0.0014595	11.4138833	0.0000000
t2:is_holiday	0.0361848	0.0053154	6.8075530	0.0000000
t2:weather_code	-0.0023533	0.0002813	-8.3646260	0.0000000
daytimeEarly Evening:t2	0.0105409	0.0021749	4.8465875	0.0000013
daytimeEarly Morning:t2	0.0023010	0.0024316	0.9463145	0.3440013
daytimeEvening:t2	0.0187753	0.0024171	7.7675931	0.0000000
daytimeLate Night:t2	0.0228891	0.0020903	10.9501789	0.0000000
daytimeMorning:t2	-0.0122110	0.0024219	-5.0418632	0.0000005
daytimeEarly Evening:is_weekend	-0.9837973	0.0295742	-33.2654346	0.0000000
daytimeEarly Morning:is_weekend	-1.7374502	0.0330315	-52.5998374	0.0000000
daytimeEvening:is_weekend	-0.5774087	0.0326225	-17.6996908	0.0000000
daytimeLate Night:is_weekend	0.7158614	0.0283570	25.2446451	0.0000000
daytimeMorning:is_weekend	-1.4231156	0.0326810	-43.5456824	0.0000000
daytimeEarly Evening:is_holiday	-0.9476753	0.0914254	-10.3655559	0.0000000
daytimeEarly Morning:is_holiday	-1.6500688	0.1025318	-16.0932405	0.0000000
daytimeEvening:is_holiday	-0.5384243	0.1009360	-5.3343150	0.0000001
daytimeLate Night:is_holiday	0.4897636	0.0876155	5.5899179	0.0000000
daytimeMorning:is_holiday	-1.5499088	0.1013314	-15.2954517	0.0000000
daytimeEarly Evening:hum	-0.0020084	0.0009974	-2.0135321	0.0440741
daytimeEarly Morning:hum	0.0021986	0.0017429	1.2614454	0.2071654
daytimeEvening:hum	0.0040484	0.0013321	3.0390073	0.0023771
daytimeLate Night:hum	0.0032344	0.0013918	2.3239754	0.0201383
daytimeMorning:hum	0.0114681	0.0012983	8.8330098	0.0000000

```
|daytimeEarly Evening:wind_speed |  0.0056516| 0.0017418|   3.2447460| 0.0011778|
|daytimeEarly Morning:wind_speed |  0.0040047| 0.0020591|   1.9448710| 0.0518066|
|daytimeEvening:wind_speed       |  0.0008227| 0.0019438|   0.4232584| 0.6721120|
|daytimeLate Night:wind_speed    | -0.0016001| 0.0016919|  -0.9457295| 0.3442997|
|daytimeMorning:wind_speed       | -0.0008794| 0.0019087|  -0.4607566| 0.6449790|
```

As with the previous model, both the linearity and constant variance assumptions are met by observing the residual plot and the fitted values. Although we do see that for predictor `hum`, there appear to be greater variation at larger predictor values and smaller variance at smaller predictor values, generally speaking most of the residual plots have their residuals scatter evenly above and below the 0-reference line, and the variation around that line appear to be random.

The normality assumption is not perfectly met according to the normal Q-Q plot. We do see a heavier left tail and a heavier right tail. But generally, the middle part fits the normal QQ line well. It is not suggested to do dedicate predictions with this model, but for description it should be enough.

The independence assumption of the model is satisfied through the way that the dataset was collected. There is no spatial correlation between each case since the response variable measures the number of bike shares of the entire city. The time correlation is also not strong since we have already accounted for time as a significant categorical variable in the data.

The adjusted-R^2 of this second model is improved by more than 10% compared to our first model.It generally satisfies the assumptions for a multiple linear regression analysis. Therefore, our study uses this multiple linear regression model for describing/modeling the median count of bike shares.

Discussion:

Big Picture and Model Interpretation

The analysis can be helpful to the U.K. bike shares companies in that the company can make their market strategies through the inferences of the number of bike

shares active in the market according to the explanatory variables such as weather projection, date of the day, day time, etc. They could then prepare the appropriate amount of shared bikes into the society/market. Also, they could make inferences about the number of bike shares based on daytime, in order to circulate new bikes and recall and repair the damaged ones during the most proper time intervals.

The table 4 below has the exponentiated coefficients and respective confidence intervals of every variable in our full model:

Table 4.	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1067.3565827	0.1045205	66.7136226	0.0000000	869.6316422	1310.0375141
daytimeEarly Evening	1.5924067	0.0917057	5.0732566	0.0000004	1.3304193	1.9059849
daytimeEarly Morning	0.4327378	0.1545350	-5.4202839	0.0000001	0.3196504	0.5858337
daytimeEvening	0.3219109	0.1183069	-9.5808493	0.0000000	0.2552852	0.4059249
daytimeLate Night	0.0546665	0.1217854	-23.8657870	0.0000000	0.0430576	0.0694053
daytimeMorning	0.9402508	0.1174143	-0.5247114	0.5997905	0.7469535	1.1835698
t2	1.0281938	0.0016290	17.0680463	0.0000000	1.0249160	1.0314820
hum	1.0119490	0.0029188	4.0695846	0.0000473	1.0061761	1.0177550
I(hum^2)	0.9998291	0.0000223	-7.6544349	0.0000000	0.9997854	0.9998729
wind_speed	0.9911810	0.0010646	-8.3203754	0.0000000	0.9891148	0.9932516
weather_code	0.9784793	0.0028596	-7.6078335	0.0000000	0.9730101	0.9839792
is_holiday	0.8507042	0.0862512	-1.8746492	0.0608578	0.7183834	1.0073975
is_weekend	1.2342137	0.0272491	7.7225987	0.0000000	1.1700226	1.3019265
t2:is_weekend	1.0167976	0.0014595	11.4138833	0.0000000	1.0138930	1.0197105
t2:is_holiday	1.0368475	0.0053154	6.8075530	0.0000000	1.0261009	1.0477065
t2:weather_code	0.9976494	0.0002813	-8.3646260	0.0000000	0.9970994	0.9981998
daytimeEarly Evening:t2	1.0105967	0.0021749	4.8465875	0.0000013	1.0062976	1.0149141
daytimeEarly Morning:t2	1.0023037	0.0024316	0.9463145	0.3440013	0.9975379	1.0070922
daytimeEvening:t2	1.0189526	0.0024171	7.7675931	0.0000000	1.0141365	1.0237917
daytimeLate Night:t2	1.0231531	0.0020903	10.9501789	0.0000000	1.0189696	1.0273538
daytimeMorning:t2	0.9878632	0.0024219	-5.0418632	0.0000005	0.9831847	0.9925640
daytimeEarly Evening:is_weekend	0.3738886	0.0295742	-33.2654346	0.0000000	0.3528311	0.3962028
daytimeEarly Morning:is_weekend	0.1759685	0.0330315	-52.5998374	0.0000000	0.1649364	0.1877385
daytimeEvening:is_weekend	0.5613511	0.0326225	-17.6996908	0.0000000	0.5265799	0.5984183
daytimeLate Night:is_weekend	2.0459483	0.0283570	25.2446451	0.0000000	1.9353320	2.1628870
daytimeMorning:is_weekend	0.2409621	0.0326810	-43.5456824	0.0000000	0.2260105	0.2569028
daytimeEarly Evening:is_holiday	0.3876411	0.0914254	-10.3655559	0.0000000	0.3240432	0.4637210
daytimeEarly Morning:is_holiday	0.1920367	0.1025318	-16.0932405	0.0000000	0.1570735	0.2347824
daytimeEvening:is_holiday	0.5836672	0.1009360	-5.3343150	0.0000001	0.4788974	0.7113579
daytimeLate Night:is_holiday	1.6319304	0.0876155	5.5899179	0.0000000	1.3744152	1.9376944
daytimeMorning:is_holiday	0.2122673	0.1013314	-15.2954517	0.0000000	0.1740299	0.2589063
daytimeEarly Evening:hum	0.9979936	0.0009974	-2.0135321	0.0440741	0.9960444	0.9999467
daytimeEarly Morning:hum	1.0022010	0.0017429	1.2614454	0.2071654	0.9987830	1.0056307
daytimeEvening:hum	1.0040566	0.0013321	3.0390073	0.0023771	1.0014383	1.0066818
daytimeLate Night:hum	1.0032397	0.0013918	2.3239754	0.0201383	1.0005066	1.0059802
daytimeMorning:hum	1.0115341	0.0012983	8.8330098	0.0000000	1.0089632	1.0141116
daytimeEarly Evening:wind_speed	1.0056676	0.0017418	3.2447460	0.0011778	1.0022401	1.0091068

daytimeEarly Morning:wind_speed	1.0040128	0.0020591	1.9448710	0.0518066	0.9999686	1.0080733
daytimeEvening:wind_speed	1.0008231	0.0019438	0.4232584	0.6721120	0.9970172	1.0046434
daytimeLate Night:wind_speed	0.9984012	0.0016919	-0.9457295	0.3442997	0.9950956	1.0017177
daytimeMorning:wind_speed	0.9991210	0.0019087	-0.4607566	0.6449790	0.9953901	1.0028658

As we can see from the exponentiated coefficients table, different time intervals like `daytimeEarly Evening, Early Morning, Evening, Late Night, Morning,` have various effects on the estimated median number of bike shares with respect to daytime `day`, which is the baseline level here. While holding other variables constant, the median number of bike shares during early morning is only 43.4% of the median bike shares during the day. Late night has very dramatic effects on the number of bike shares, as we can see from the model that the median number of bike shares decreases by 94.5% compared with that during the day.

Holidays and weekends can also have interaction effects with daytime on the estimated median number of bike shares. For example, if it is during late night on a Saturday or Sunday, the estimated median of bike shares will increase by ((1.2342137*2.0459483)-1)*100% = 152.5% compared to the median number of bike shares at late night during weekdays. Similarly, if it is a holiday on a weekday, then during early morning, the estimated median of bike shares will decrease by (1-(0.8507042*0.1920367))*100% = 83.6% compared to the median number of bike shares at early morning on a normal working weekday.

Extreme temperatures can also discourage people from getting outside and using bike shares. We have found that people are more likely to use bike shares during warmer days. While holding other variables constant during the day, every one degree increase in the temperature that people feel like will increase the median number of bike shares by 2.83%. A 95% confidence interval of the effect of one degree increase in `t2` on the median number of bike shares is from 2.49% to 3.15%.

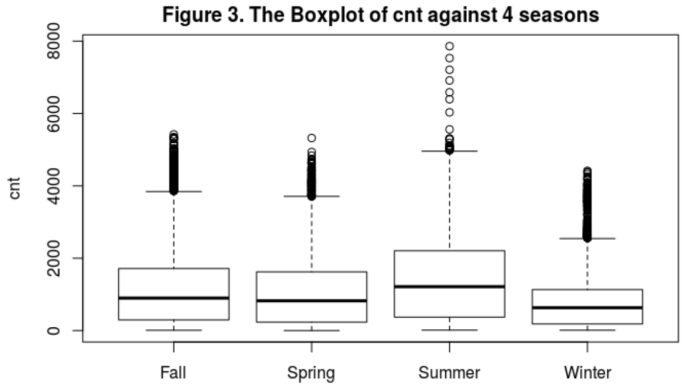
Careful when Making Predictions

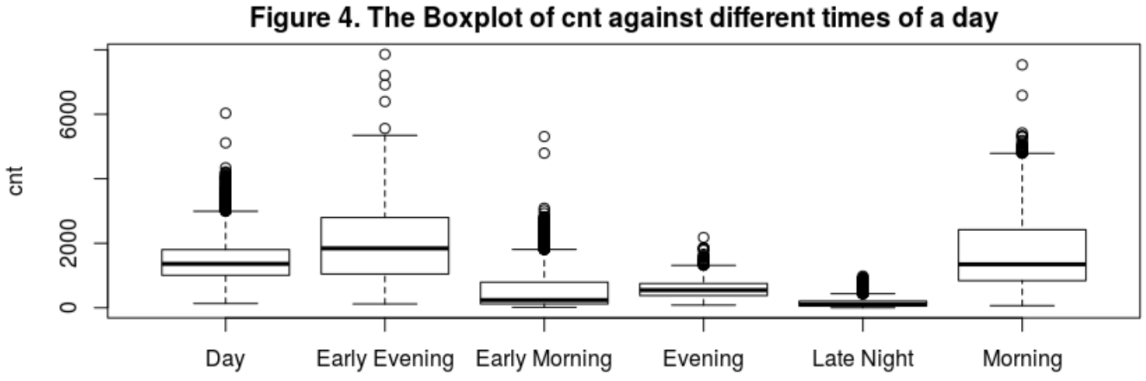
According to the normal Q-Q plot, the central part of our final model fits along the line, but two tails are heavier. So, for “average cases” (where average cases means that the temperature, wind speed, humidity etc. are around the average), we might be

considering using it for prediction, but for extreme cases where any variable has a closer to extreme value, the Q-Q plot tells us that the residuals does not follow a normal distribution, so we should not trust the model for making any predictions in such cases.

Intuitive Results

We did not find previous study on this topic, but the results coming out of this study is intuitive. The results of our study have now shown that weather and different times of a day can significantly impact the likelihood of using bike sharing. For example, people use bike sharing more during warm seasons than cold winter (Figure 3) and during rush hours than during late night (Figure 4), which are both reasonable phenomenons that confirm our previous belief. We also see that the effects of daytime can vary during the weekends or holidays since people have largely different schedules during these particular times than weekdays.





Problems and Further Improvements

The biggest problem for us in this study is that the data points are correlated in time. In the scope of our capabilities, we can only split the `timestamp` variable into

categorical variables with some levels to explain the relationship between the number of bike shares and time. We wish we could know more about time series analysis, in order to fully analyze the time correlation nature of this dataset.

Moreover, most of the variables in our dataset focus primarily on time and weather as predictors of the number of bike sharing in London. The variations of our response variable may also be explained by other potential factors, for example the number of bikes available for people to use around the city, the changes in the population of the city, and the location where the bikes are offered. All these factors can possibly exert some significant influences on our response and are not taken into consideration in our study due to the lack of data.

In order to refine our model, we would like to get the number of people living in the area that the data is collected from. Therefore, we can make more inferences using general linear model (logistical models) about the odds of a citizen using bike shares under different conditions. Had we given more detailed personal information about each individual, we could be able to make inferences about the individuals' age and social status etc. to explain if he/she uses bike shares.

Conclusion:

This study explores how to explain the median number of bike shares by using different times of a day, the date of a day (whether it is a weekend or holiday) and a series of predictors related to environment/weather conditions such as temperature, humidity, wind speed, season and weather code. The model shows that the time of the day has a great influence on the median number of bike shares. Whether a day is a weekend or a holiday also has a significant interaction with day time on the median number of bike shares. People also tend to rent more bike on better weather conditions. Overall, this model explains a large amount of variation of median number of bike shares; to further improve the model, further study should look into time series analysis to better explore the timely correlation nature of the dataset.

Reference:

Guo Y., Zhou J., Wu Y., & Li Z. (2017). Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China. *PloS one*, 12(9), from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5608320/

Mavrodiev, H. (2019, October 10). London bike sharing dataset. Retrieved November 25, 2019, from https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset.

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., & Haq, U. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. Journal of Transport Geography, 41, 306–314. doi: 10.1016/j.jtrangeo.2014.01.013