Kuangyou Chen
MA 678
Dec 8, 2021

Bike Sharing Dataset

Background

Bike sharing system is a new version of traditional bike rentals that is updated with automated mechanisms on tracking rental, return and membership. The system changes the way people commute. All users are able to get access to the bikes conveniently in particular regions.In current society, the bike sharing system is prevalent for its value on healthcare, environment and efficiency.

Introduction

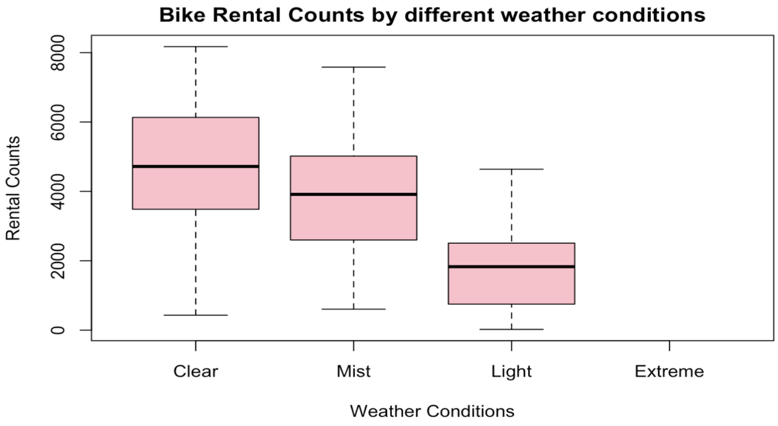
In addition to the large-scale application in the city, the characteristics of data are attractive and valuable to explore. The bike sharing dataset records features based on time pattern. The data considers the majority of external factors that may affect the rental bikes. The project will mainly focus on the EDA and fitted model.

Method

The dataset is published on UCI Machine Learning Repository and Kaggle. There is no null or duplicated value according to the instruction of Kaggle and function check by R programming. When I created boxplots in order to detect outliers of each numerical variable in the dataset. I removed those values in the column casual and hum, which are separately the total counts of casual users and normalized humidity. As for those normalized variables, I created new columns about their original values, which showed as “original_”. In addition, those categorical variables are dummied and re-labeled based on real-world circumstances.

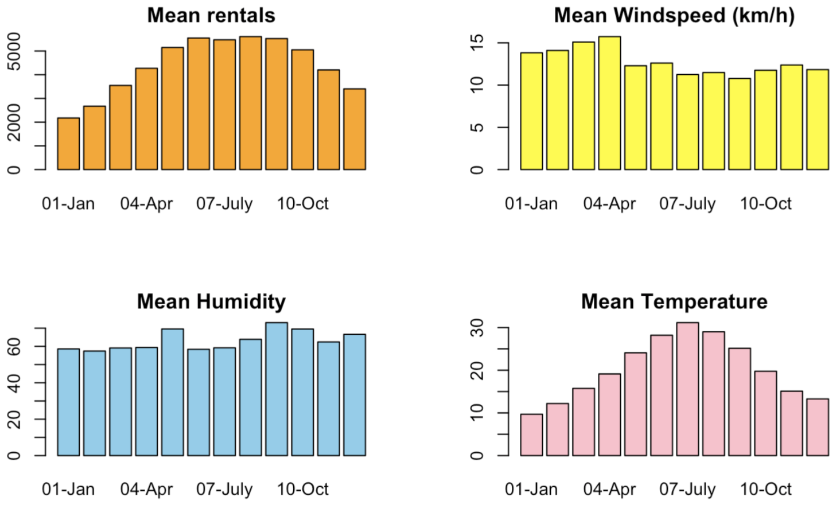
Exploratory Data Analysis

For the market of bike sharing systems, seasonal patterns are a significant factor in visualizing or predicting the trend. Here I grouped by the seasons and created some plots. From the summary analysis, the lowest extreme and average temperature are in the spring (2.42°C,12.11°C) while the highest extreme and average temperature are in the fall (35.32°C,28.97°C).The boxplot shows that the lowest temperatures are at spring season and followed by winter, while the highest temperatures at fall and followed by summer. Those outliers in the boxplot might be season shifts and extreme weather circumstances.



- weathersit :
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

There are no type 4 conditions in our dataset. The boxplot demonstrates that the highest mean value of rentals have days with the 1st weather type (clear, partly cloudy etc.),while the lowest number of rentals happened at the type 3 (light snow, light rain+thunderstorm+scattered clouds, light rain).

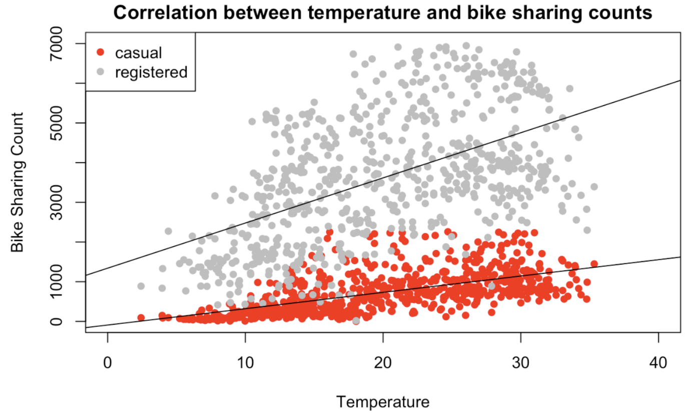


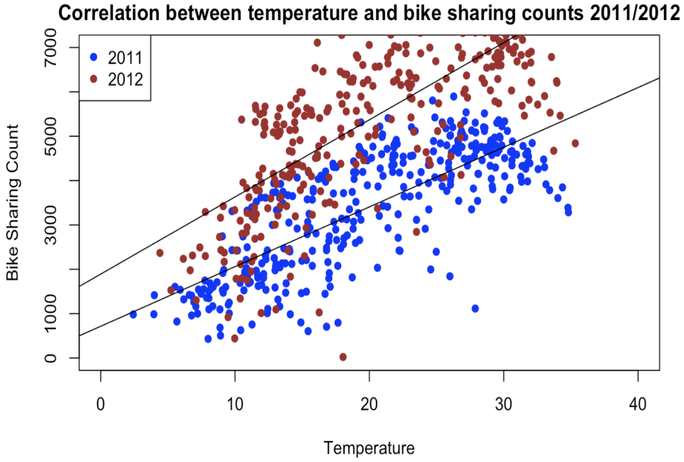
The combination of above four histograms shows the seasonal trend variety of rentals, windspeed, humidity and temperature. The highest frequency of rental counting shows in the summer, with the relatively low wind speed, low humidity and highest temperature. While the lowest rental counting shows in the winter, with the relatively high wind speed, relatively low humidity and low temperature. In order to figure out the dependency and correlation among rentals and independent variables like wind speed, humidity, and temperature, we would further conduct ANOVA Table and modeling.

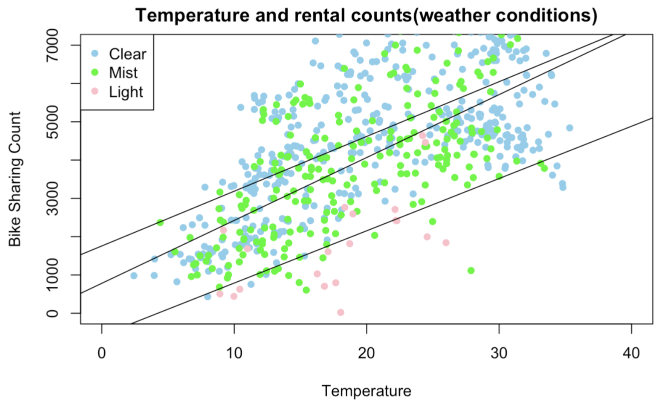
ANOVA Table

According to the result of Tukey's Honest Significant Difference test, the most significant difference in total number of bike sharing counts are between Fall and Spring. While the least significant difference is between Winter and Summer. Additionally for weather conditions, the most significant difference would be between light weather to clear, while the least significant difference is between mist and clear.

Data Modeling





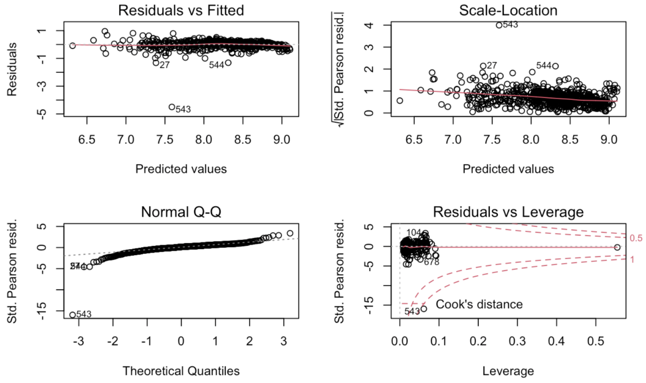


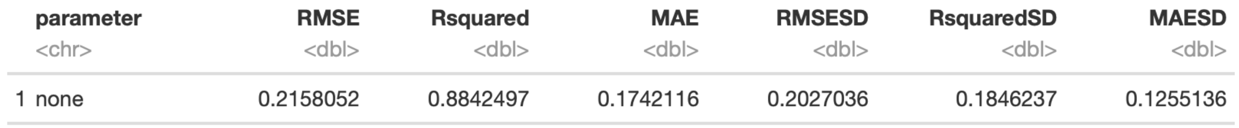
Those three visualizations are the analysis of temperature metric. I separately grouped by casual-register, 2011-2012 and various weather conditions. For the correlation among temperature and casual-registered rental counts, the registered group has higher coefficient of variation than the casual group. The correlation between bike sharing counts and temperature in 2011 is higher than the one in 2012. Additionally, the mist (mist+cloudy, mist+broken clouds, mist+few clouds) weather declares the highest correlation among rental counts and temperature rather than any other weather conditions.

```
Call:
glm(formula = log(cnt) ~ season + yr + weathersit + holiday + 
    workingday + mnth + log(temp) + log(atemp) + log(hum) + log(windspeed), 
    data = bikeday)

Deviance Residuals: 
    Min       1Q    Median       3Q      Max  
-4.4980  -0.0884   0.0425   0.1447   0.9521  

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       7.982176   0.096967  82.318  < 2e-16 ***
seasonSummer      0.290686   0.042193   6.889 1.29e-11 ***
seasonFall        0.257499   0.056132   4.587 5.36e-06 ***
seasonwinter      0.475212   0.055874   8.505  < 2e-16 ***
yr                0.434169   0.022734  19.098  < 2e-16 ***
weathersitMist   -0.100580   0.029647  -3.393 0.000733 ***
weathersitLight -0.943272   0.075799 -12.444  < 2e-16 ***
holiday          -0.199521   0.072419  -2.755 0.006027 ** 
workingday        0.082912   0.026233   3.161 0.001645 ** 
mnth             -0.008052   0.005859  -1.374 0.169812    
log(temp)         0.732180   0.205170   3.569 0.000384 ***
log(atemp)       -0.056706   0.215156  -0.264 0.792200    
log(hum)         -0.295488   0.067367  -4.386 1.34e-05 ***
log(windspeed)  -0.151045   0.029330  -5.150 3.43e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For this dataset, GLM is a good option because it shows balanced performance for both numerical and categorical variables. I applied log ratio for all the numerical normalized variables and combined them with other categorical variables. The image above shows the final result with a linear regression model. All the variables except for atemp (feeling temperatures) and mnth(month from Jan to Dec) are significant (P-value is extremely low).

Model Validation





	parameter <chr>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	none	0.2158052	0.8842497	0.1742116	0.2027036	0.1846237	0.1255136

The accuracy score is 88% when applying the K-Fold Cross Validation.

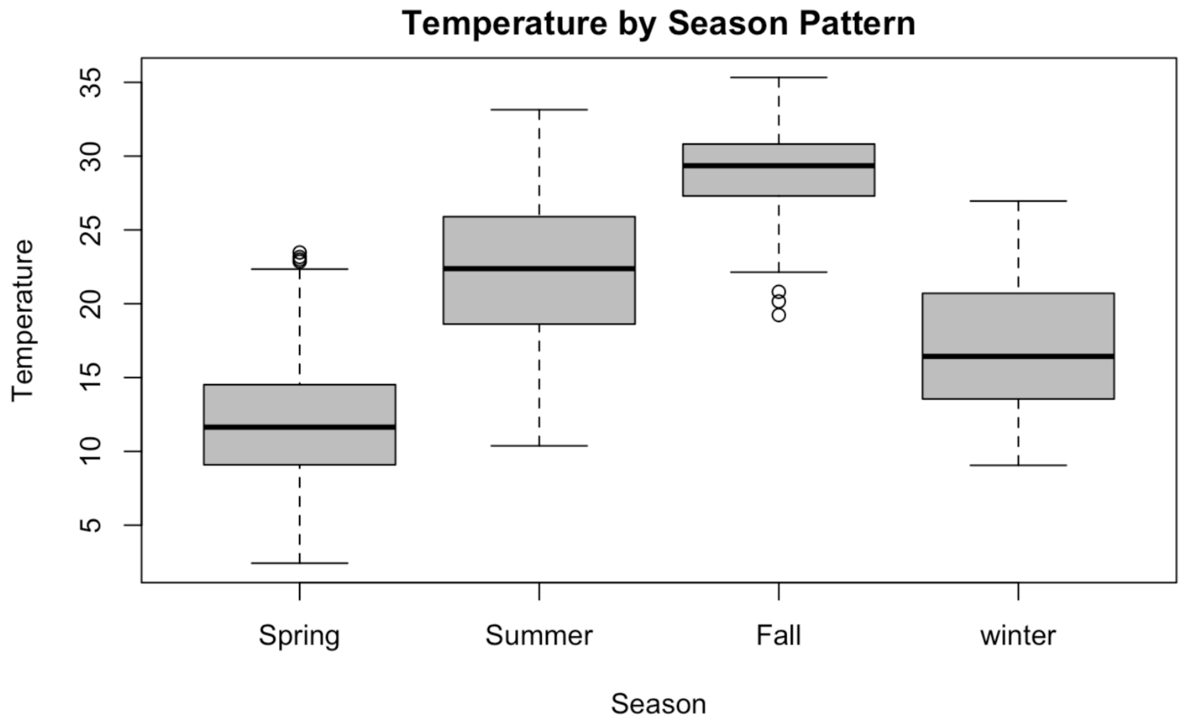
Discussion

The estimates are reasonable and reliable based on the expectation and visualizations in the EDA part. However, what I am confused about is the statistical insignificance of month value in model fitting. Based on the EDA, it is explicitly stated that the mean rental shows a large fluctuation in the seasonal pattern. While the result of modeling questions about the sufficiency of evidence to prove its correlation. For further improvement, I would like to explore more on feeling temperature and seasonal patterns.

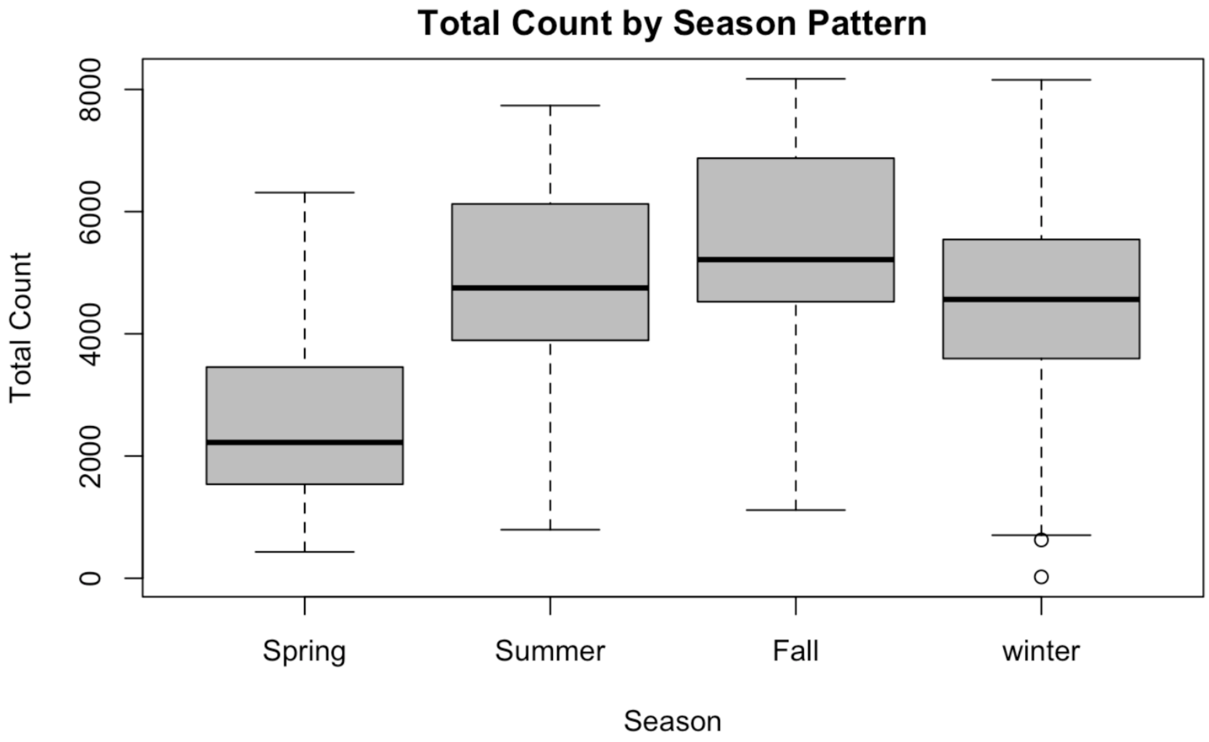
Appendix

Here is the attribute information:

instant	Record index	weathersit	1: clear, 2:mist,3:light snow,rain, 4:heavy rain, ice pallets
dterday	date	temp:	Normalized temperature in Celsius
season	1: spring, 2:summer,3:fall,4:winter	atemp	Normalized feeling temperature in Celsius
yr	year(0:2011,1:2012)	hum	Normalized humidity
mnth	month(1-12)	windspeed	Normalized wind speed
hr	hour(0-23)	Casual	Count of casual users
holiday	Whether the day is holiday or not	registered	Count of registered users
weekday	Day of the week	cnt	Count of total rental bikes including both casual and registered
Working day	1 if day is neither weekend or holiday, 0 otherwise		

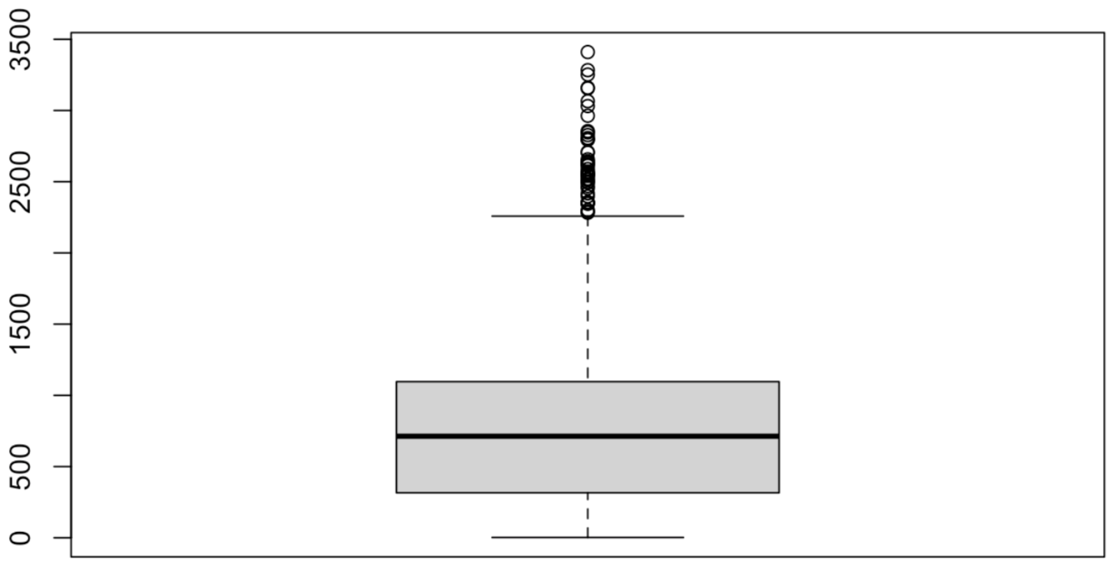


Temperature by Season Pattern

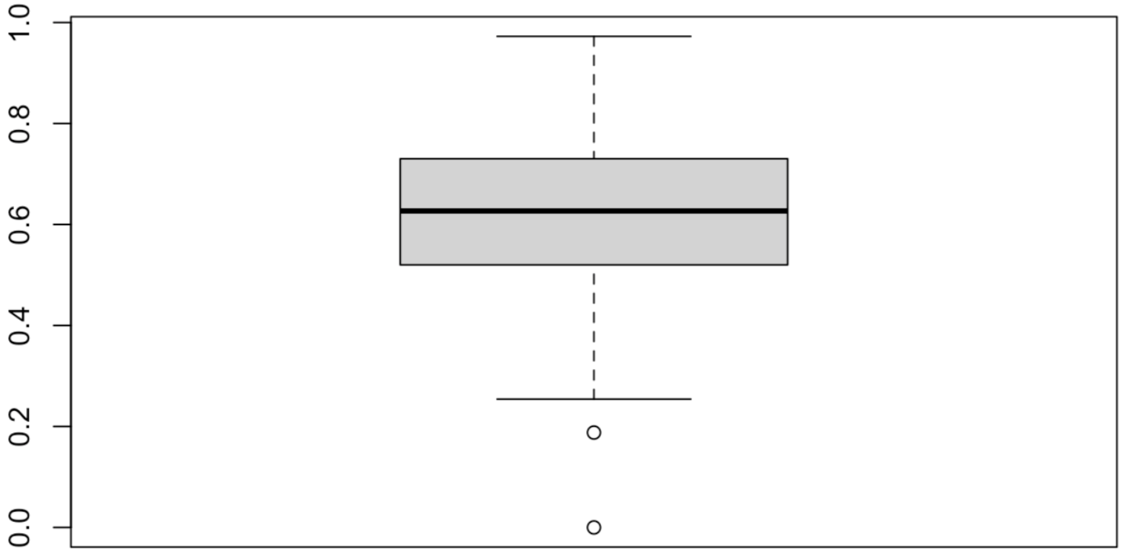


Total Count by Season Pattern

```
boxplot(bikesharing_day$casual)
```



```
boxplot(bikesharing_day$hum)
```


1.0
0.8
0.6
0.4
0.2
0.0