Predict the number of shared bikes by weather conditions

Zichu Chen, Kaiqi Yu, Ziya Zhao

1. Introduction

Bike-sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return has become automatic. The welcome public service which allows people to borrow a bike from a dock and return it in another place. Until 2017, there're over 500 bike-sharing systems with 500,000 bikes around the world. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real-world applications of bike-sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Different from other transport services, the duration of the trip, start and end time arrival positions are explicitly recorded in these systems. These features make the bike-sharing system a virtual sensor network that can be used for sensing mobility of a city. Through this research, we expect to find any factors that influence people's choice of using shared bicycles or not and predict how many times the bike-sharing system would be used in a period in Washington, DC.

2. Data Description

2.1 Data Source

There’re 3 data resources we would use:

a. 2011-2012 Bike Sharing in Washington D.C. Dataset from Kaggle: This dataset contains the hourly and daily count of rental bikes between the years 2011 and 2012 in the Capital bike-share system in Washington, DC with the corresponding weather and seasonal information. There're 17379 rows in this dataset and each row is a record for an hour in the years 2011 and 2012. The variables are shown below:

Name	**Description**
Instant	Record index
dteday	Date
Season	Season (1:springer, 2:summer, 3:fall, 4:winter)
yr	Year (0: 2011, 1:2012)
mnth	Month (1 to 12)
hr	Hour (0 to 23)
holiday	weather day is holiday or not (extracted from Holiday Schedule)
weekday	Day of the week
Workingday	If day is neither weekend nor holiday is 1, otherwise is 0
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
temp	The normalized temperature in Celsius. The values are derived via: (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
atemp	The normalized feeling temperature in Celsius. The values are derived via: (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
hum	Normalized humidity. The values are divided into 100 (max).
windspeed	Normalized wind speed. The values are divided into 67 (max).
casual	Count of casual users
registered	Count of registered users
cnt	Count of total rental bikes including both casual and registered

Link: https://www.kaggle.com/marklvl/bike-sharing-dataset#hour.csv

b. 2017 Bike Sharing in Washington D.C Dataset from capital bike-share: With the years 2011-2012's sharing bike dataset as the training set, we would use the year 2017's data as the test set. This dataset contains trip history data in the year 2017 from capital bike share. There're 8738 rows in this dataset and each row is a record for bike-sharing. The variables are shown below:

Name	**Description**
Duration	Duration of the trip
Start Date	Includes start date and time
End Date	Includes end date and time
Start Station	Includes start station name and number
End Station	Includes end station name and number
Bike Number	Includes ID number of bike used for the trip
Member Type	Indicates whether the user was a "registered" member or a "casual" rider

Link: https://www.capitalbikeshare.com/system-data

c. 2017 Daily Observations of Arlington County, VA Weather History from weather underground: The second dataset lack corresponding weather information so we use web scraping tools to build the third dataset as a supplement. This dataset contains weather information for each hour in 2017 and has these variables:

Name	**Description**
Date	Date
Hour	Hour
TF	Feeling temperature
Humidity	Humidity
WindSpeed	Windspeed

weather	Description of the weather

Link: https://www.wunderground.com/history/daily/us/va/arlington-county/KDCA/date/2017-1-1

The second and third datasets were combined to build a dataset with the same important variables as that of the first dataset.

3. Data pre-processing

The 2011-2012 bike-sharing in Washington DC dataset was provided and cleaned by Kaggle contributor Mark Kaghazgarian while the second data was a raw dataset with no weather information. Also, each row in the second dataset is a record of a specific bike sharing while in the first dataset, all the bike-sharing records were counted in hours. So we grouped the data in the second dataset according to the variable Start Date. We also dropped the variables which are not contained in the first dataset.

Then we used chrome driver and beautiful soup to scrape the observed weather information in 2017 from the website weather underground. It took about 6 hours to scrape all the information because loading the website takes a long time. The normal observe time of this observation station was on the 52nd minute in each hour so we chose all the normal observations and drop others to avoid repetition.

The second and third datasets were combined and converted into a new dataset named "hour2017.csv". This dataset has a structure the same as the first dataset and has only time, weather information and count of total rental bikes. Also, we normalized the feeling temperature, humidity, and wind speed in the same way as the description of the first dataset.

4. Exploratory Data Analysis

4.1 DC dataset

In the first part, we will explore the 'hour.csv' file. This data set records the hourly weather conditions and the number of shared bikes used between 2011 and 2012.

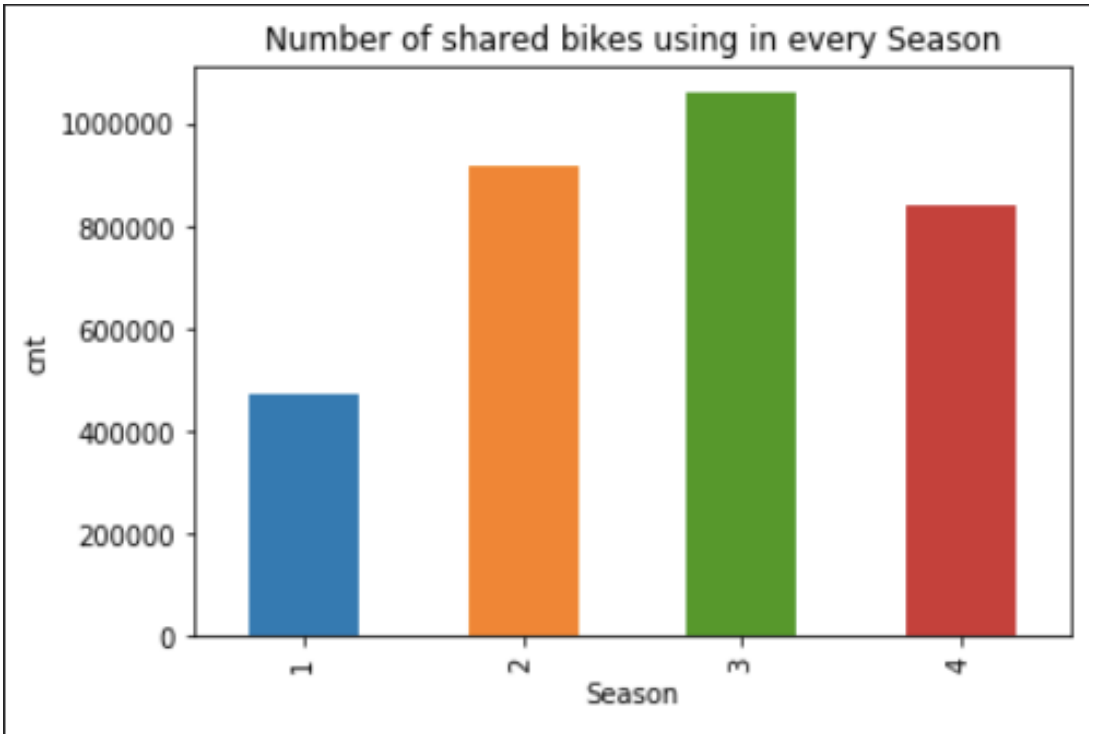
4.1.1 Basic analysis

We did not find NA values in the dataset, which means we can use the dataset directly for analysis.

4.1.2 How many people use bike-sharing in each season?

The count of season bike using shows below:

Season Id	Season Name	Number of using
1	Spring	471,348
2	Summer	918,589
3	Autumn	1,061,129
4	Winter	841,613

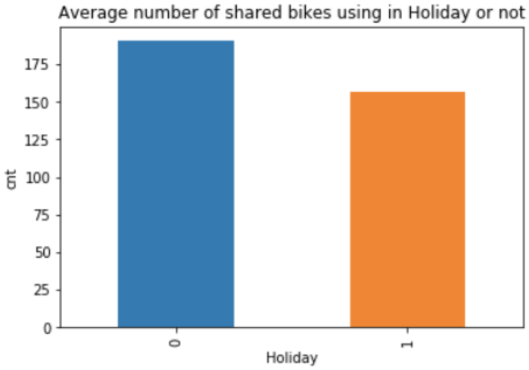


As shown in the figure, most people using shared bicycles in autumn, followed by summer, winter, and spring.

4.1.3 How many people use bike sharing on holidays and non-holidays?

We use averages for comparison because the number of holidays and non-holidays is different. The count of holidays and non-holidays bike using shows below:

Holiday variable	Whether it is a holiday	Number of using
0	Non-holiday	190.42858
1	Holiday	156.87000

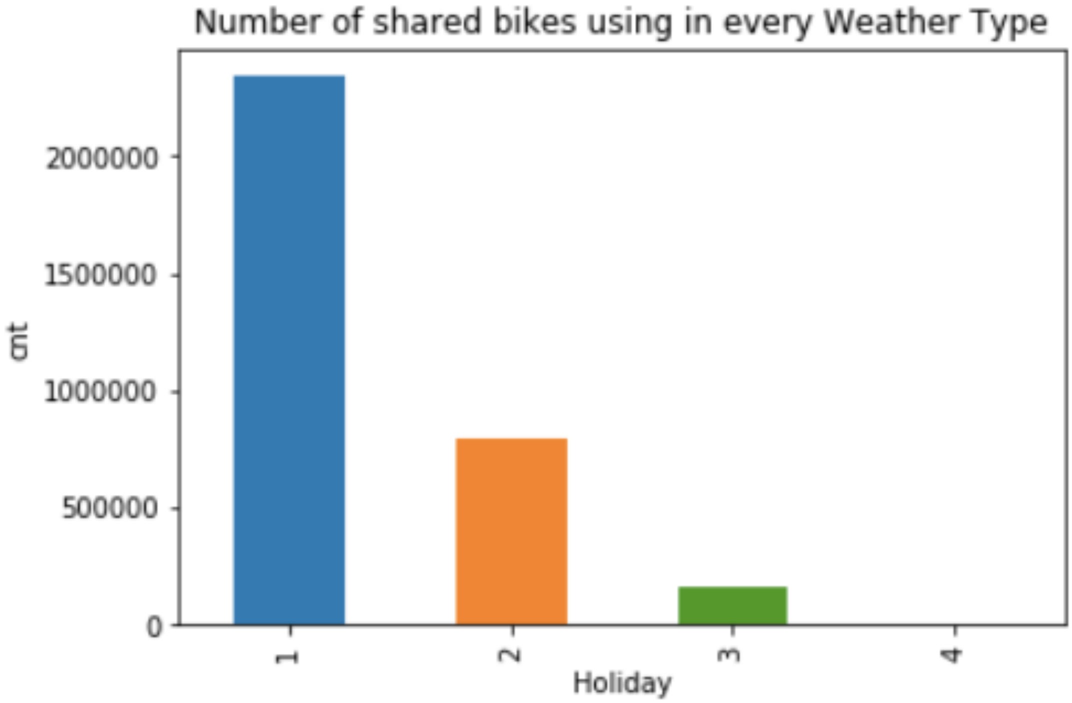


As shown in the figure, the average number of bikes used by people during holidays and non-holidays is very close, indicating that people will often use bikes during holidays or non-holidays.

4.1.4 How many people use bike sharing on different weather types?

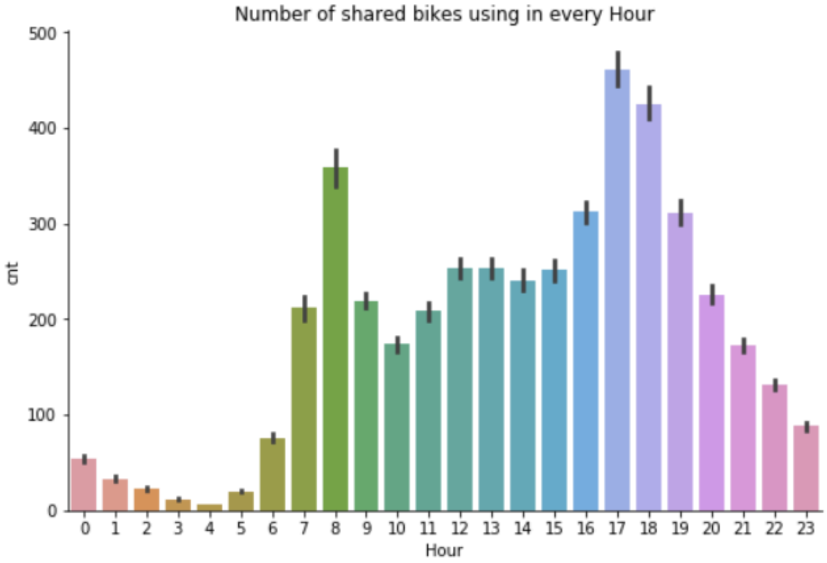
The count of different weather bike using shows below:

Weather id	Weather type	Number of using
1	Clear, Few clouds, Partly cloudy, Partly cloudy	2,338,173
2	Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist	795,952
3	Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds	158,331
4	Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog	223

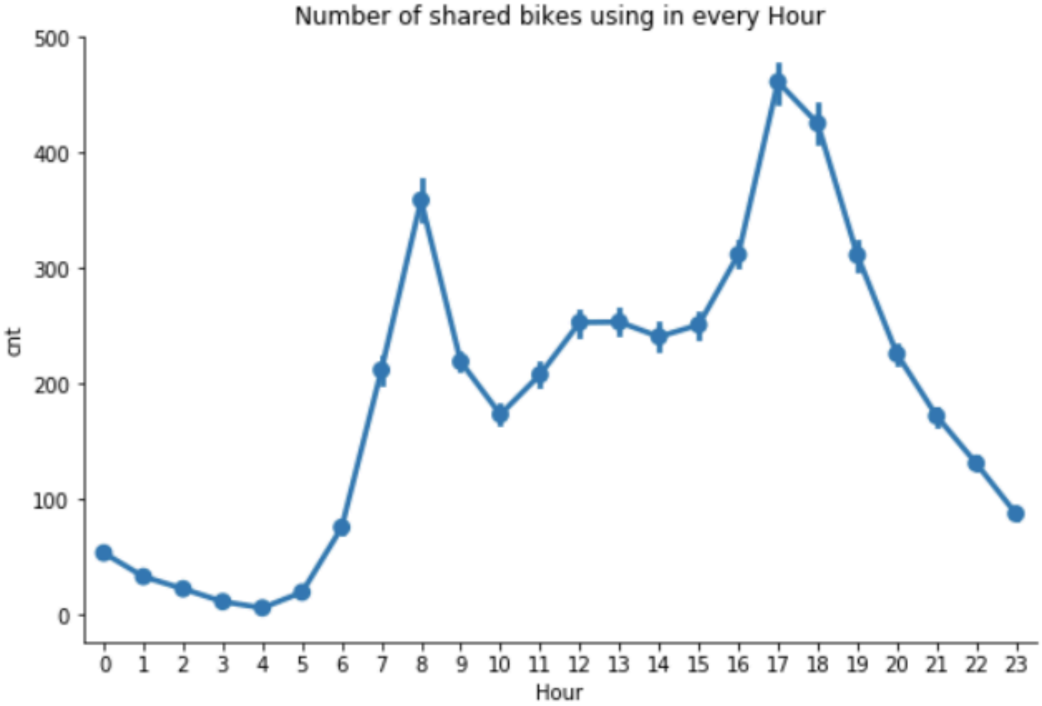


As shown in the figure, the better the weather, the more people use it.

4.1.5 How many people use bike-sharing in each hour?

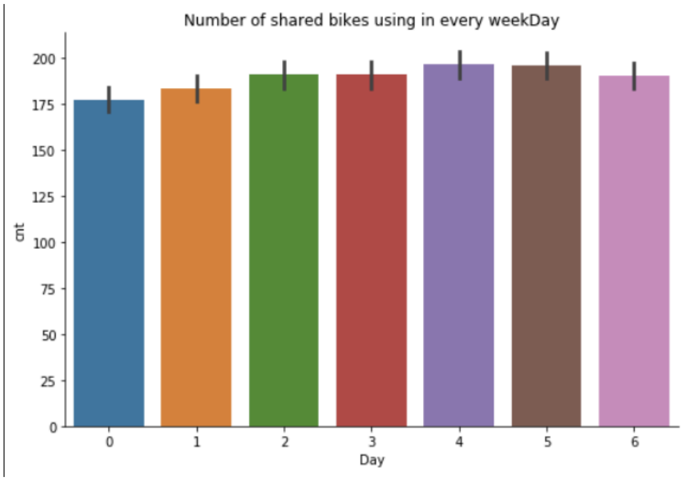


As shown in the figure, in daily life, the peaks of people's use of shared bicycles are 8 AM and 5-6 PM, with 261,001 and 336,860 to 309,772 people, respectively.



We can better observe this phenomenon in the line chart.

4.1.6 How many people use bike-sharing in each weekday?

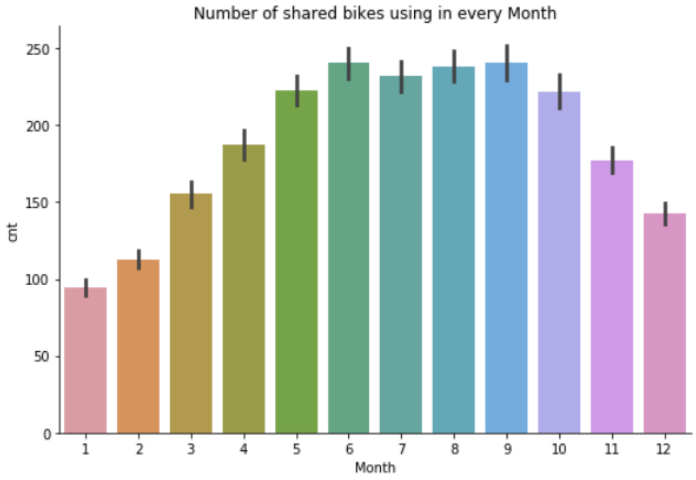


As shown in the figure, the number of shared bicycles used by people is very close every day, indicating that in daily life, shared bicycles are used every day.

4.1.7 How many people use bike-sharing in each month?

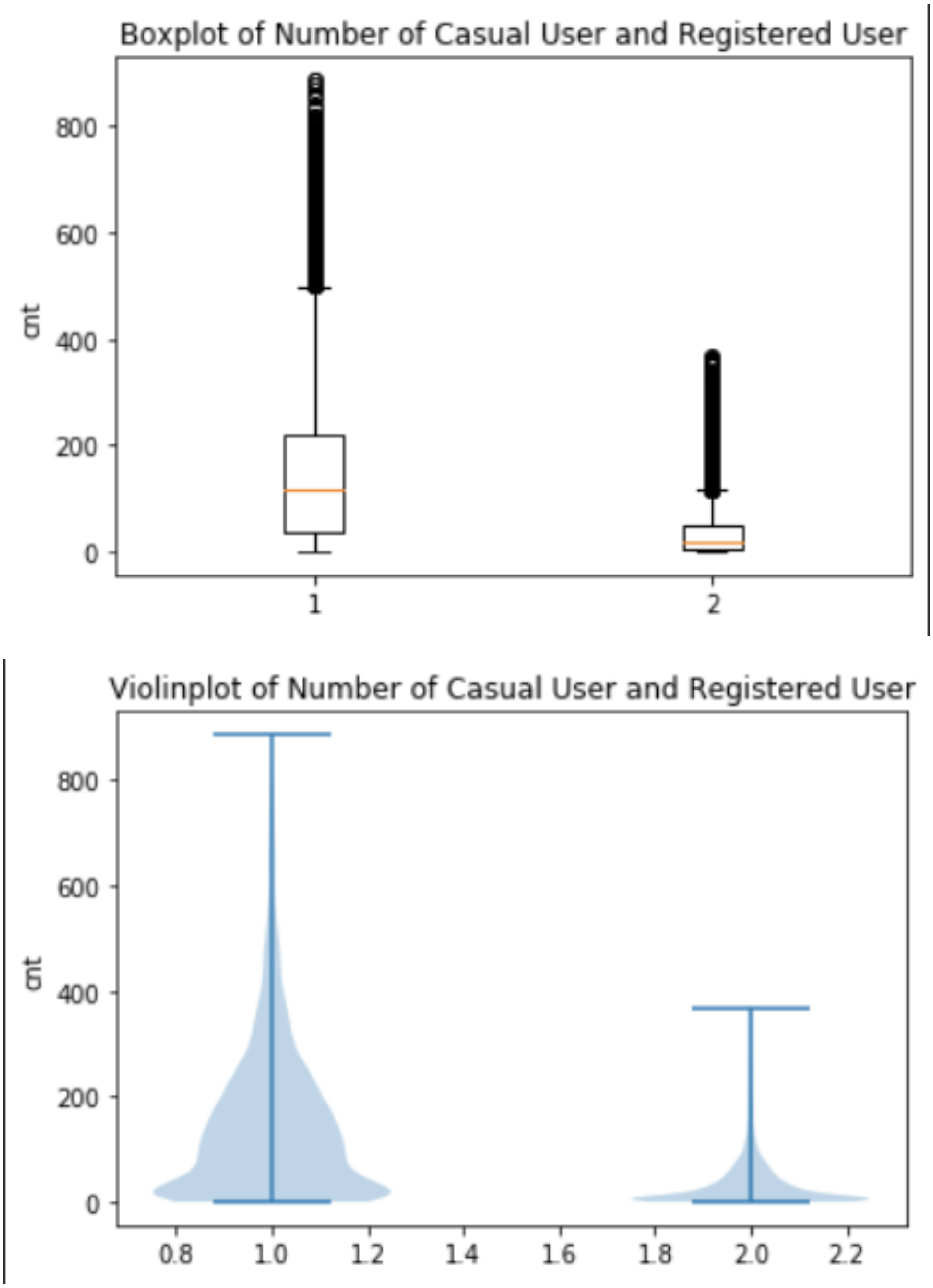
The count of month bike using shows below:

Month	Count	Month	Count
1	134,933	7	344,948
2	151,352	8	351,194
3	228,920	9	345,991
4	269,094	10	322,352
5	331,686	11	254,831
6	346,342	12	211,036



As shown in the figure, the peak period of bicycle sharing is concentrated from June to September.

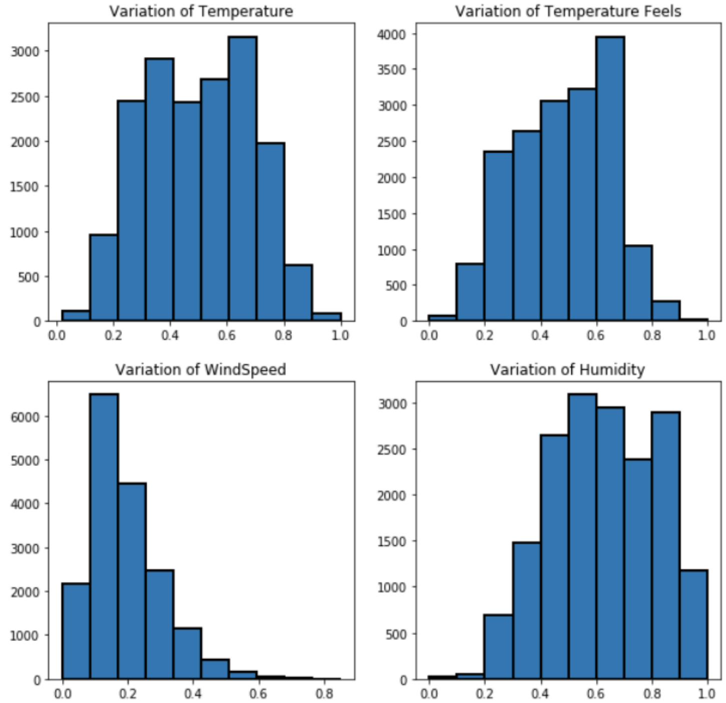
4.1.8 Observe registered users and occasional registered users



1 for registered users, 2 for accidentally registered users

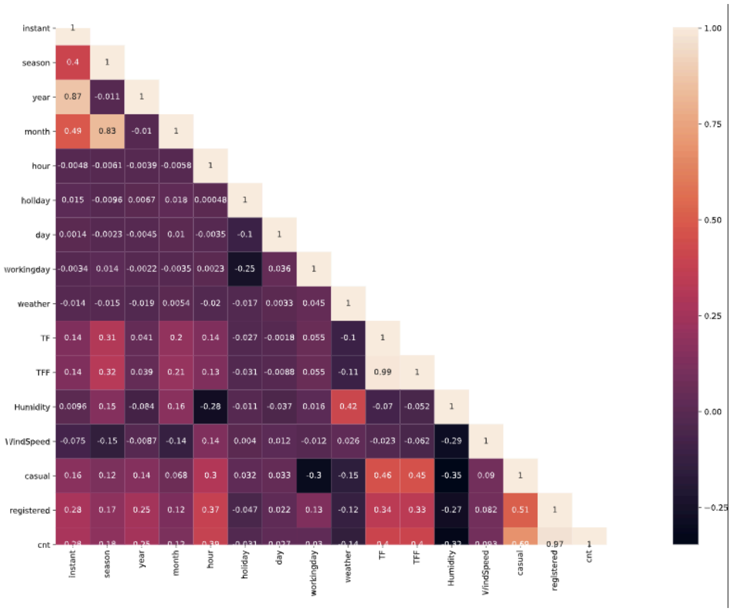
It can be seen from the box and violin charts that the number of registered users is far more than the number of accidental registrations, indicating that many people often use shared bicycles.

4.1.9 How many people use bike-sharing in different weather conditions?



After comprehensively comparing the analysis results of weather conditions, we can conclude that most people use bicycle sharing when the temperature is high, the more people use it when the wind speed is low, and most people use it when the humidity is high. But when these factors reach extreme values, the number of sharing bicycles people used is the least.

4.1.10 correlation matrix of DC dataset

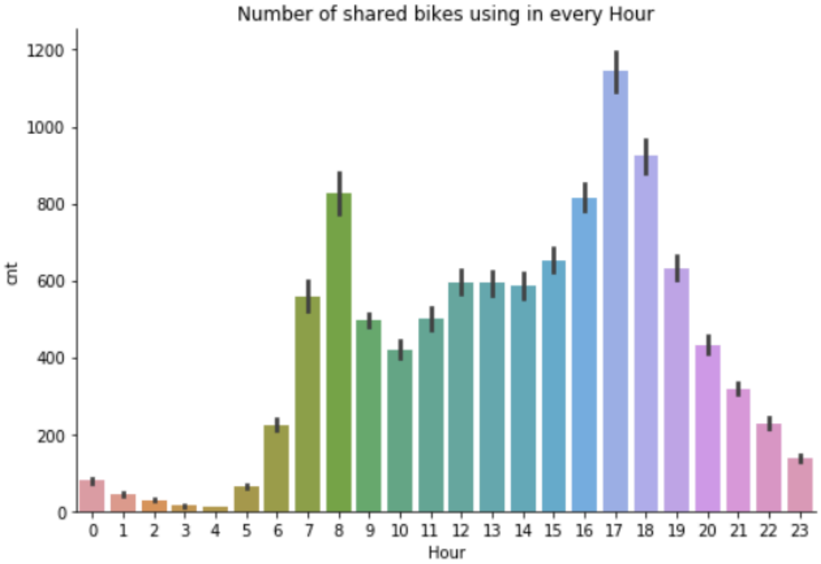


Through this figure, we use appropriate features for modeling

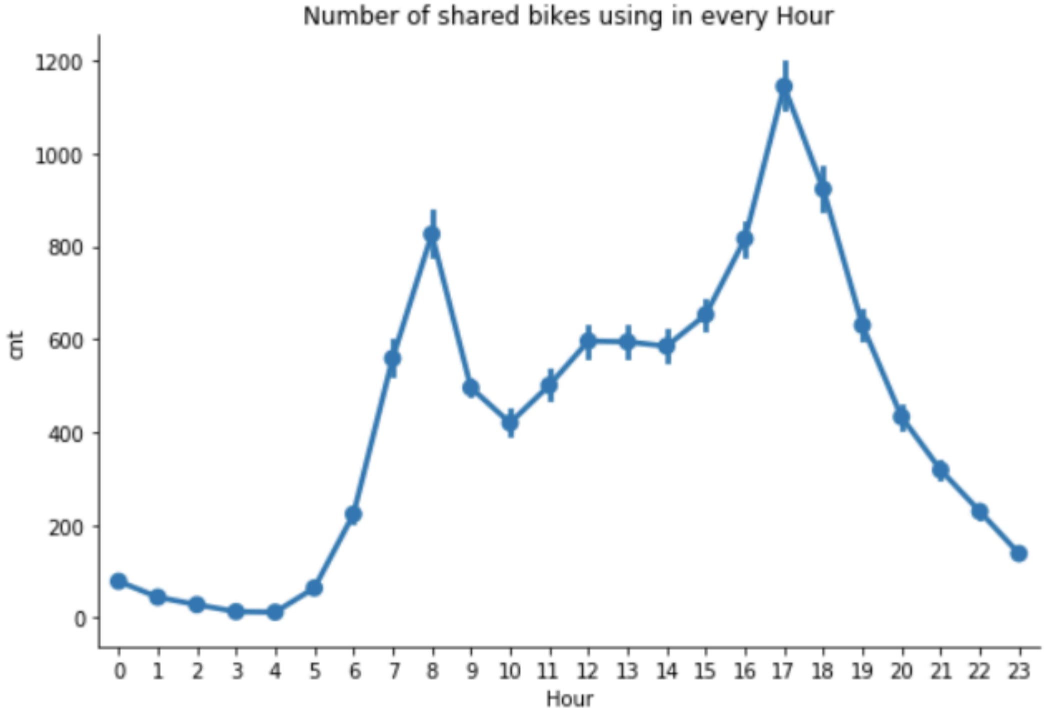
4.2 DC2017 dataset

In the first part, we will explore the 'hour2017.csv' file. This data set records the hourly weather conditions and the number of shared bikes used in 2017.

4.2.1 How many people use bike-sharing in each hour?

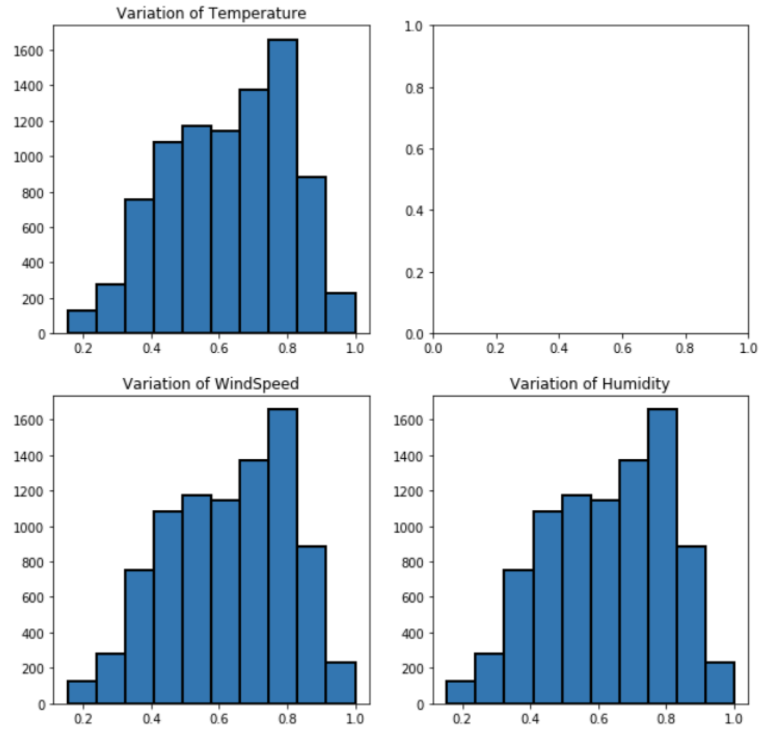


As shown in the figure, in daily life, the peaks of people's use of shared bicycles are 8 AM and 5 PM, with 299,108and 415,574 people, which is very similar to data recorded in 2011 and 2012.



We can better observe this phenomenon in the line chart.

4.2.2 How many people use bike-sharing in different weather conditions?



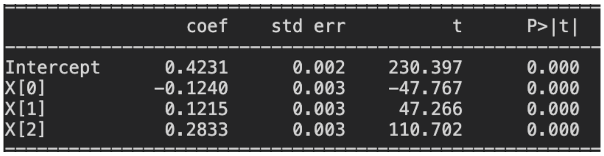
In the 2017 record, the impact of weather conditions on the use of shared bicycles has not changed much. Compared with the changes in 2011-2012, people will use shared bicycles more when the wind speed is moderate.

5. Modeling

5.1 Regression Part I: Linear regression model

After we import the data from 'dc' bike-sharing from 2011 to 2012, we will create dummy variables for categorical type variable 'weather' and 'season'. For the categorical type variable, the symbolized number does not represent a higher value from 1 to 4. Also for the 'hour', we will use the C() function in the formula to automatically create the dummy instead of input x one by one.

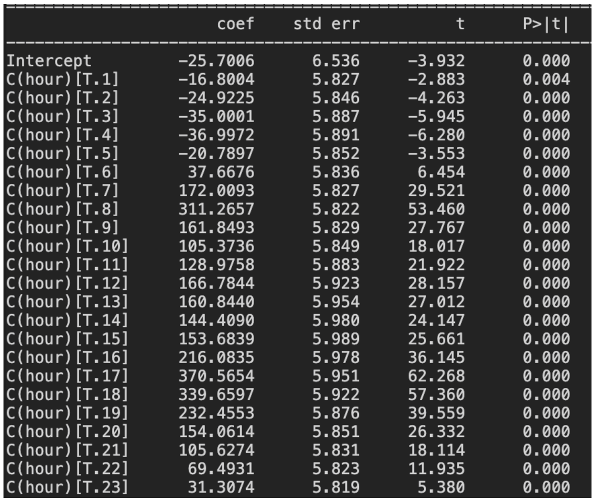
For the 'season' variable, we will use temperate as the dependent variable to compare temperature for each season using s_4 as the benchmark. And the result is the rank of the average temperature for four seasons from low to high is: s_1, s_4, s_2, s_3.



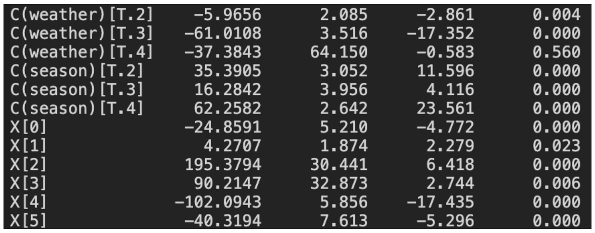
```
==============================================================
                 coef    std err          t      P>|t|
--------------------------------------------------------------
Intercept      0.4231      0.002    230.397      0.000
X[0]          -0.1240      0.003    -47.767      0.000
X[1]           0.1215      0.003     47.266      0.000
X[2]           0.2833      0.003    110.702      0.000
==============================================================
```

Then for the linear regression model, we will use all variables showing in the data set and 'cnt' (the total number of people rent the sharing bike) as the dependent variable and it will be the same for the regression tree in the later part. The result is showing as below:

We can see that the highest number of people using share bikes is around 5:00 pm which is the time that most people will go home from work. And for X0 to X5, it separately represents 'holiday', 'workingday', 'TF', 'TFF', 'Humidity', 'WindSpeed'. We can see that more people choose to use sharing bikes during the workday than a holiday and the higher the temperate with lower humidity and wind speed, the more people would like to use the shared bike as a commuting method.

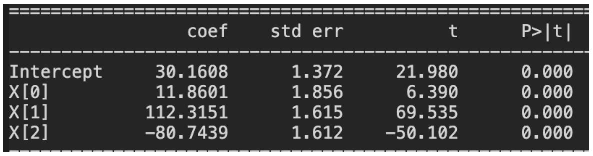


```
======================================================================
                     coef    std err          t      P>|t|
----------------------------------------------------------------------
Intercept        -25.7006      6.536     -3.932      0.000
C(hour)[T.1]     -16.8004      5.827     -2.883      0.004
C(hour)[T.2]     -24.9225      5.846     -4.263      0.000
C(hour)[T.3]     -35.0001      5.887     -5.945      0.000
C(hour)[T.4]     -36.9972      5.891     -6.280      0.000
C(hour)[T.5]     -20.7897      5.852     -3.553      0.000
C(hour)[T.6]      37.6676      5.836      6.454      0.000
C(hour)[T.7]     172.0093      5.827     29.521      0.000
C(hour)[T.8]     311.2657      5.822     53.460      0.000
C(hour)[T.9]     161.8493      5.829     27.767      0.000
C(hour)[T.10]    105.3736      5.849     18.017      0.000
C(hour)[T.11]    128.9758      5.883     21.922      0.000
C(hour)[T.12]    166.7844      5.923     28.157      0.000
C(hour)[T.13]    160.8440      5.954     27.012      0.000
C(hour)[T.14]    144.4090      5.980     24.147      0.000
C(hour)[T.15]    153.6839      5.989     25.661      0.000
C(hour)[T.16]    216.0835      5.978     36.145      0.000
C(hour)[T.17]    370.5654      5.951     62.268      0.000
C(hour)[T.18]    339.6597      5.922     57.360      0.000
C(hour)[T.19]    232.4553      5.876     39.559      0.000
C(hour)[T.20]    154.0614      5.851     26.332      0.000
C(hour)[T.21]    105.6274      5.831     18.114      0.000
C(hour)[T.22]     69.4931      5.823     11.935      0.000
C(hour)[T.23]     31.3074      5.819      5.380      0.000
```



```
C(weather)[T.2]    -5.9656      2.085     -2.861      0.004
C(weather)[T.3]   -61.0108      3.516    -17.352      0.000
C(weather)[T.4]   -37.3843     64.150     -0.583      0.560
C(season)[T.2]     35.3905      3.052     11.596      0.000
C(season)[T.3]     16.2842      3.956      4.116      0.000
C(season)[T.4]     62.2582      2.642     23.561      0.000
X[0]              -24.8591      5.210     -4.772      0.000
X[1]                4.2707      1.874      2.279      0.023
X[2]              195.3794     30.441      6.418      0.000
X[3]               90.2147     32.873      2.744      0.006
X[4]             -102.0943      5.856    -17.435      0.000
X[5]              -40.3194      7.613     -5.296      0.000
```

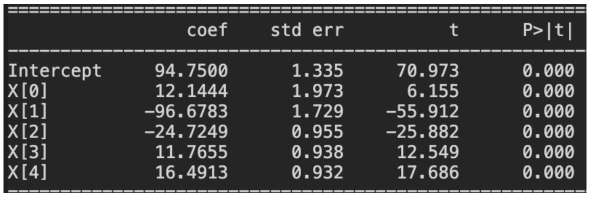
Also, to have a deeper understanding of our data for the casual user and registered user we have separated the season and temperature effect, working day and holiday to avoid multicollinearity. The following are the four models we build:



```
==============================================================
                 coef    std err          t      P>|t|
--------------------------------------------------------------
Intercept     30.1608      1.372     21.980      0.000
X[0]          11.8601      1.856      6.390      0.000
X[1]         112.3151      1.615     69.535      0.000
X[2]         -80.7439      1.612    -50.102      0.000
```

# 1. ('holiday', 'TF', 'Humidity')

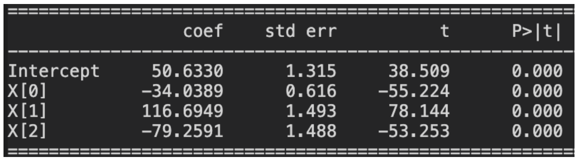
For casual users, We can see from the result that on holiday there are more casual users renting a bike. Also, the higher the temperature and lower humidity, the more casual users will rent the bike.



```
==============================================================
                 coef    std err          t      P>|t|
--------------------------------------------------------------
Intercept     94.7500      1.335     70.973      0.000
X[0]          12.1444      1.973      6.155      0.000
X[1]         -96.6783      1.729    -55.912      0.000
X[2]         -24.7249      0.955    -25.882      0.000
X[3]          11.7655      0.938     12.549      0.000
X[4]          16.4913      0.932     17.686      0.000
==============================================================
```

# 2. ('holiday', 'Humidity', 's_1', 's_2', 's_3')

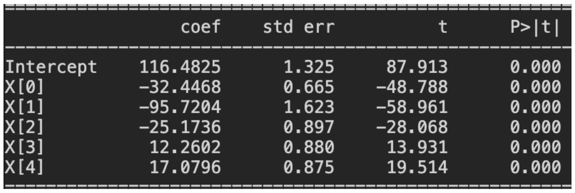
More casual users rent a bike on holidays when humidity is low and in the season that has a higher rank in temperature



```
==================================================
                 coef    std err          t      P>|t|
--------------------------------------------------
Intercept     50.6330      1.315     38.509      0.000
X[0]         -34.0389      0.616    -55.224      0.000
X[1]         116.6949      1.493     78.144      0.000
X[2]         -79.2591      1.488    -53.253      0.000
==================================================
```

# 3. ('workingday', 'TF', 'Humidity')

More casual users are less likely to rent a bike on workingday and if they do more casual users rent a bike when the temperature is high and humidity is low

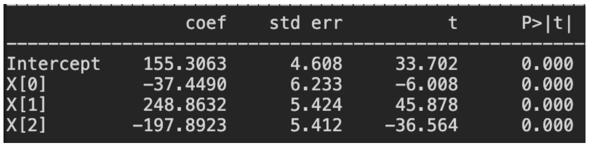


```
==================================================
                 coef    std err          t      P>|t|
--------------------------------------------------
Intercept    116.4825      1.325     87.913      0.000
X[0]         -32.4468      0.665    -48.788      0.000
X[1]         -95.7204      1.623    -58.961      0.000
X[2]         -25.1736      0.897    -28.068      0.000
X[3]          12.2602      0.880     13.931      0.000
X[4]          17.0796      0.875     19.514      0.000
==================================================
```

# 4. ('workingday', 'Humidity', 's_1', 's_2', 's_3')

More casual users are less likely to rent a bike on workingday and if they do more casual users rent a bike when humidity is low and season have a higher rank in temperature

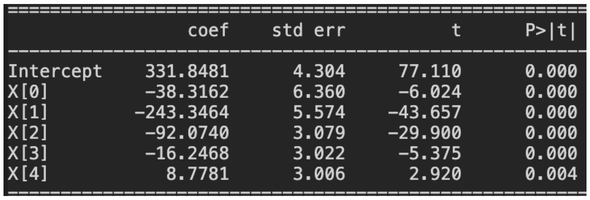
For registered users, We can see from the result that on holiday there are fewer registered users renting a bike. For the registered user, the main reason they register is that they will use a lot and we can imagine those people are because of work commute need. In this case, when it comes to holidays, those people will just rest and not using bikes. Also, the higher the temperature, the more casual users will rent the bike.



```
                 coef    std err          t      P>|t|
--------------------------------------------------
Intercept    155.3063      4.608     33.702      0.000
X[0]         -37.4490      6.233     -6.008      0.000
X[1]         248.8632      5.424     45.878      0.000
X[2]        -197.8923      5.412    -36.564      0.000
```

# 1. ('holiday', 'TF', 'Humidity')

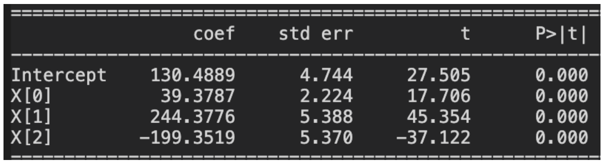
Less registered users rent a bike on holidays and if they do, more people will rent a bike when the temperature is high and humidity is low



```
==================================================
                 coef    std err          t      P>|t|
--------------------------------------------------
Intercept    331.8481      4.304     77.110      0.000
X[0]         -38.3162      6.360     -6.024      0.000
X[1]        -243.3464      5.574    -43.657      0.000
X[2]         -92.0740      3.079    -29.900      0.000
X[3]         -16.2468      3.022     -5.375      0.000
X[4]           8.7781      3.006      2.920      0.004
==================================================
```

# 2. ('holiday', 'Humidity', 's_1', 's_2', 's_3')

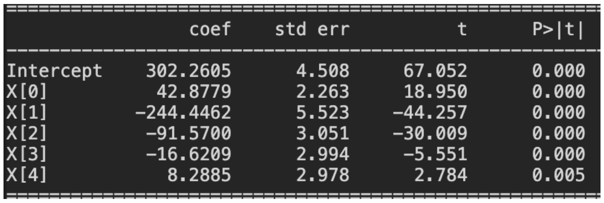
Fewer registered users rent a bike on holidays and if they do, more people will rent a bike when humidity is low. the rank of the number of people renting bikes for each season from low to high is s_1, s_2, s_4, s_3. This is different from the result we get from 'TF' (previously we have the temperature order from low to high as s_1, s_4, s_2, s_3)



```
==================================================
                 coef    std err          t      P>|t|
--------------------------------------------------
Intercept    130.4889      4.744     27.505      0.000
X[0]          39.3787      2.224     17.706      0.000
X[1]         244.3776      5.388     45.354      0.000
X[2]        -199.3519      5.370    -37.122      0.000
==================================================
```

# 3. ('workingday', 'TF', 'Humidity')

More registered users are renting a bike on workingday when humidity is low and seasons have a higher rank in temperature



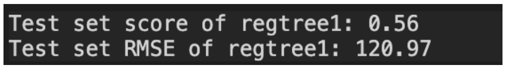
```
==============================================================
                 coef    std err          t      P>|t|
--------------------------------------------------------------
Intercept    302.2605      4.508     67.052      0.000
X[0]          42.8779      2.263     18.950      0.000
X[1]        -244.4462      5.523    -44.257      0.000
X[2]         -91.5700      3.051    -30.009      0.000
X[3]         -16.6209      2.994     -5.551      0.000
X[4]           8.2885      2.978      2.784      0.005
==============================================================
```

# 4. ('workingday', 'Humidity', 's_1', 's_2', 's_3')

More registered users are renting a bike on workingday when humidity is low. And the rank of the number of people renting bikes for each season from low to high is s_1, s_2, s_4, s_3. This is different from the result we get from 'TF' (previously we have the temperature order from low to high as s_1, s_4, s_2, s_3)

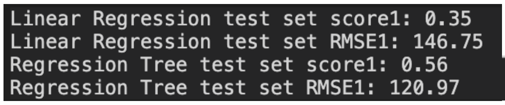
5.2 Regression Part II: Regression tree model

We will use the same variables in the OLS model for comparison and split the dc dataset into 80% train, 20% test. The following is the result we have for instantiating a Decision Tree Regressor.

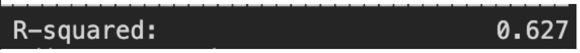


```
Test set score of regtree1: 0.56
Test set RMSE of regtree1: 120.97
```

Then we will compare the performance with OLS, and we will use Root Mean Square Error to see which model has a better fit. we can see from the result that the regression tree has lower RMSE and higher score which means that it has a better model fit. The unit of RMSE is the same as the dependent variable. If your data has a range of 0 to 100000 then the RMSE value of 3000 is small, but if the range goes from 0 to 1, it is pretty huge.



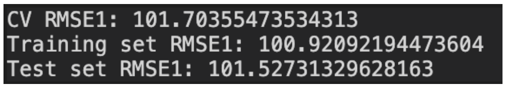
```
Linear Regression test set score1: 0.35
Linear Regression test set RMSE1: 146.75
Regression Tree test set score1: 0.56
Regression Tree test set RMSE1: 120.97
```



```
R-squared:                      0.627
```

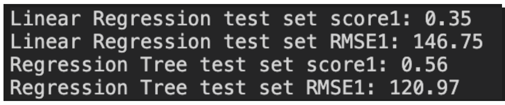
In this case, we have RMSE around 130 and the range of casual users is from 0 to 977, it is about 0.1 of the largest number and the R-squared is 0.6, which shows that it is also a good fit for the total user

Also, we will evaluate the list of Mean Square Error obtained by 10-fold CV, we can see from the result that the training set and test set have a similar amount of RMSE which means that the model we built is a good one. And the regression tree is better than the linear regression.

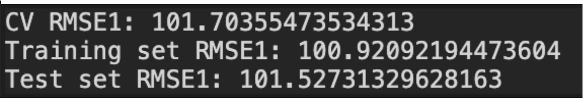


```
CV RMSE1: 101.70355473534313
Training set RMSE1: 100.92092194473604
Test set RMSE1: 101.52731329628163
```

To have deeper valuation of our model, we will separate the situation for working days and holidays as we did in the linear regression model, and the dependent variable will be 'cnt'(the total number of people rent the sharing bike) The result is the similar as shown above for the day only in working day.

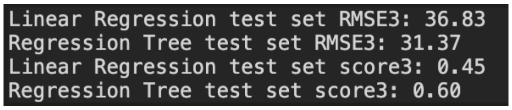


```
Linear Regression test set score1: 0.35
Linear Regression test set RMSE1: 146.75
Regression Tree test set score1: 0.56
Regression Tree test set RMSE1: 120.97
```

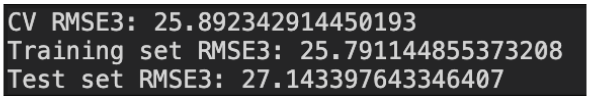


```
CV RMSE1: 101.70355473534313
Training set RMSE1: 100.92092194473604
Test set RMSE1: 101.52731329628163
```

Then we will use 'casual' as the dependent variable for another valuation. And the result is showing below:

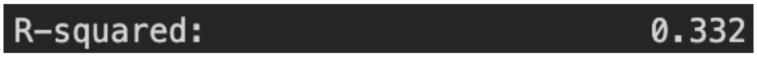


```
Linear Regression test set RMSE3: 36.83
Regression Tree test set RMSE3: 31.37
Linear Regression test set score3: 0.45
Regression Tree test set score3: 0.60
```



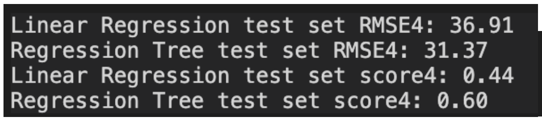
```
CV RMSE3: 25.892342914450193
Training set RMSE3: 25.791144855373208
Test set RMSE3: 27.143397643346407
```

In this case, we have RMSE around 30 and the range of casual users is from 0 to 367, it is about 0.1 of the largest number and the R-squared is 0.3, which shows that it is also a good fit for the casual user. And the regression tree is better than the linear regression.

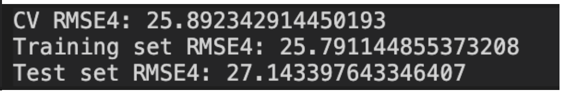


```
R-squared:                       0.332
```

To have a deeper valuation of our model, we will separate the situation for working days and holidays as we did in the linear regression model, and the dependent variable will be 'casual' The result is similar as shown above for the day only in a working day.



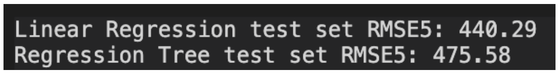
```
Linear Regression test set RMSE4: 36.91
Regression Tree test set RMSE4: 31.37
Linear Regression test set score4: 0.44
Regression Tree test set score4: 0.60
```



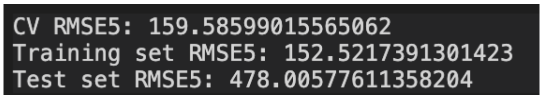
```
CV RMSE4: 25.892342914450193
Training set RMSE4: 25.791144855373208
Test set RMSE4: 27.143397643346407
```

5.3 Prediction

After we compared our models with two regression, we will use the web scrapping data from 2017 weather and bike-sharing to see how our model performs. And we use original dc data as a train set and dc 2017 data as the test set. The variables will be 'TF', 'humidity, 'windSpeed', and 'hour', and the dependent variable is 'cnt'. We can see from the result that the linear regression model is better than the regression tree model instead and the RMSE has increased a lot from around 120 to around 440. And the training set RMSE is much lower than the test set RMSE which means that we have over-fitting the data. And the reason is that that we do not have enough variables as we did in the previous data set of dc (which contains over 10 different variables). If we do not have the limitation of online sourcing data set of web scrapping and have more variables, we can have better prediction



```
Linear Regression test set RMSE5: 440.29
Regression Tree test set RMSE5: 475.58
```



```
CV RMSE5: 159.58599015565062
Training set RMSE5: 152.5217391301423
Test set RMSE5: 478.00577611358204
```

Github link:
https://github.com/Kelv1nYu/ItDMProj