Contents

BIKE SHARING ANALYSIS

AUGUST 13TH, 2020

Introduction:

Bicycle sharing system is a service in which bicycles are available for shared use to individuals on a short-term basis. Many bike sharing systems allow people to borrow a bike from one dock and return it to another dock belonging to the same organization.

Bike sharing systems are the new generation of traditional bike rentals, where the whole process of membership, rental and return has changed automatically. With these settings, the user can easily rent a bike from one position to another. Currently, there are more than 500 bike sharing projects around the world, with over 500 thousand bicycles. Today, there is increasing interest in these organizations due to their key role in transportation, environmental and health issues.

Apart from the interesting real-world applications of bike sharing systems, the properties of the data generated by these systems make them attractive for research. In contrast to other transportation services such as bus or subway, travel time, departure and arrival status are explicitly recorded in these systems. This feature converts the bike sharing system into a virtual sensor network that can be used to sense movement in the city. Therefore, it is expected that by monitoring these data the most important events of the city can be detected

Capital Bikeshare is one such bike rental company that started in 2010 in Washington DC. The company's bike rental increased 2.6 times from 2010 to 2014 (LANG K., 2015). People use these bikes for work, occasional mistakes or for recreational purposes. The docks of these bikes collect data on bike use and are constantly monitored (Timo, p. 2009, p. 51).

Data collection and Data used for analysis

Data source: https://www.kaggle.com/marklvl/bike-sharing-dataset. This database includes data from Capital Bikeshare in Washington, D.C., and weather data for 2011 and 2012.

There are two different membership types in Capital Bikeshare: Annual and Daily. These are represented in the normal and registered data, the normal is related to the daily members and is recorded to the annual members. Each row in the database represents a day. Other Columns include week, season, month condition weather , temperature, humidity, wind, number of normal rides, recorded rides and total number of rides.

Capital Bikeshare collects data for each day and every hour for the years 2011 and 2012. The most important attribute of the data we work with is cnt, which is the number of bikes rented per hour. This characteristic will be the dependent variable of the study. Other characteristics include weather, season, temperature, humidity, wind speed and holiday information. In this article of analysis, we look at the dominant factors that can affect bike rentals.

The total number of observations in the dataset: 17379

The total number of attributes in the dataset: 16

Following are the details about every attribute of the dataset:

1.dteday: Date of registration

2.season: the data of season (1: spring, 2: summer, 3: fall, 4: winter)

3.yr: year information(0: 2011 and 1: 2012)

4.mnth: month information (from 1 to 12)

5.hr: hour information (from 0 to 23)

6.holiday: the field informs whether the day is a holiday or not.

7.weekday: information of day of the week

8.workingday: if the day is neither weekend nor holiday is 1, otherwise is 0.

9.weathersit: weather situation:-

-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

10.temp: Normalized temperature in Celsius.

11.atemp: Normalized feeling temperature in Celsius.

12. hum: Normalized humidity. The values are divided into 100 (max)

13.windspeed: Normalized wind speed. The values are divided into 67 (max)

14.casual: number of casual users.

15.registered: number of registered users.

16.cnt: count of total rental bikes including both casual and registered

The "cnt" attribute here is the most important attribute in this analysis, which gives the total number of bikes are rented for a day.

Descriptive Statistics / Exploratory data Analysis

Details about data:
Statistical details

	instant	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp
count	17379.0000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	8690.0000	2.501640	0.502561	6.537775	11.546752	0.028770	3.003683	0.682721	1.425283	0.496987
std	5017.0295	1.106918	0.500008	3.438776	6.914405	0.167165	2.005771	0.465431	0.639357	0.192556
min	1.0000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.020000
25%	4345.5000	2.000000	0.000000	4.000000	6.000000	0.000000	1.000000	0.000000	1.000000	0.340000
50%	8690.0000	3.000000	1.000000	7.000000	12.000000	0.000000	3.000000	1.000000	1.000000	0.500000
75%	13034.5000	3.000000	1.000000	10.000000	18.000000	0.000000	5.000000	1.000000	2.000000	0.660000
max	17379.0000	4.000000	1.000000	12.000000	23.000000	1.000000	6.000000	1.000000	4.000000	1.000000

	atemp	hum	windspeed	casual	registered	cnt
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	0.475775	0.627229	0.190098	35.676218	153.786869	189.463088
std	0.171850	0.192930	0.122340	49.305030	151.357286	181.387599
min	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	0.333300	0.480000	0.104500	4.000000	34.000000	40.000000
50%	0.484800	0.630000	0.194000	17.000000	115.000000	142.000000
75%	0.621200	0.780000	0.253700	48.000000	220.000000	281.000000
max	1.000000	1.000000	0.850700	367.000000	886.000000	977.000000

Details about each column:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 17 columns):
instant       17379 non-null int64
dteday        17379 non-null object
season        17379 non-null int64
yr            17379 non-null int64
mnth          17379 non-null int64
hr            17379 non-null int64
holiday       17379 non-null int64
weekday       17379 non-null int64
workingday    17379 non-null int64
weathersit    17379 non-null int64
temp          17379 non-null float64
atemp         17379 non-null float64
hum           17379 non-null float64
windspeed     17379 non-null float64
casual        17379 non-null int64
registered    17379 non-null int64
cnt           17379 non-null int64
dtypes: float64(4), int64(12), object(1)
memory usage: 2.3+ MB
```

Statistical Analysis

Analysis 1:

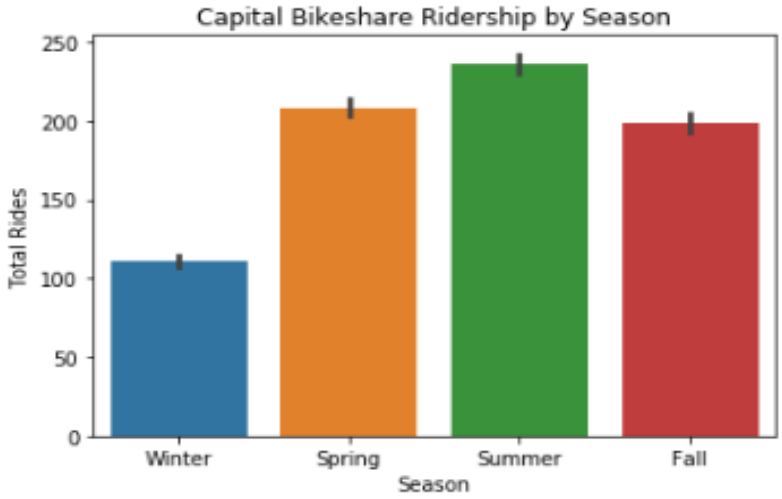
Showing The Capital Bike Sharing Ridership by Season:

Does season really affect the bikeshare usage?

Yes. The season has an impact on bikeshare usage. There are a limited number of rides in the winter and very large rides in the summer. A large T-value of -35.50 and a small B-value of 5.031e-258 suggest a significant difference between summer and winter travel. However, there are differences in ridership between spring and fall

The small T-value is not significant with 2.37 and the large B-value is 0.017.

```
Winter vs Spring
Ttest_indResult(statistic=-28.56228809916416, pvalue=1.5115313254768806e-171)
Winter vs Summer
Ttest_indResult(statistic=-35.503258721186356, pvalue=5.031871922307298e-258)
Winter vs Fall
Ttest_indResult(statistic=-26.162526539941013, pvalue=3.691977909937702e-145)
Spring vs Fall
Ttest_indResult(statistic=2.370523439417162, pvalue=0.01778474157284087)
Spring vs Summer
Ttest_indResult(statistic=-6.759876431441242, pvalue=1.467921954992745e-11)
Summer vs Fall
Ttest_indResult(statistic=9.094821678300208, pvalue=1.156579487676187e-19)
```



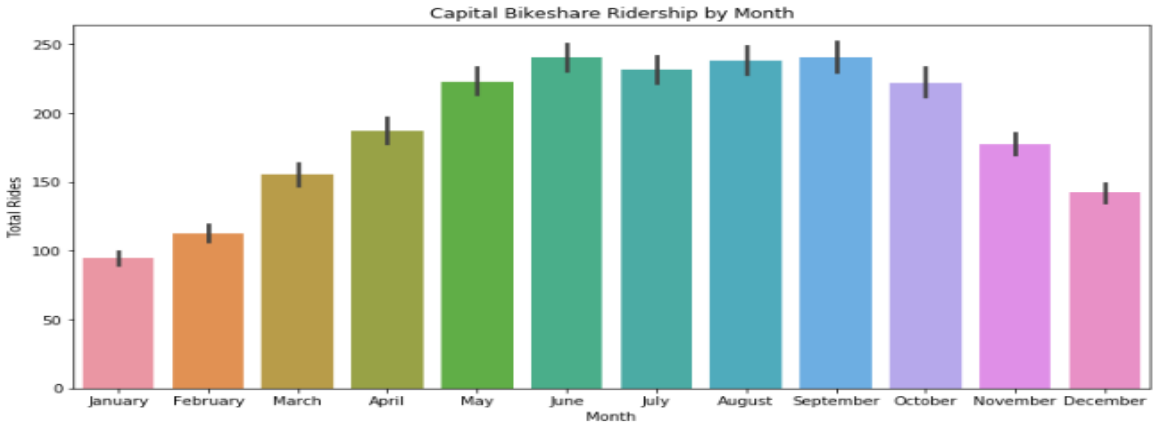
So, by above picture one can see the difference between ridership in different seasons, so season is really affecting ridership.

Analysis 2:

Showing Capital Bike Sharing Ridership by Month:

Does really season affects the bikeshare usage?

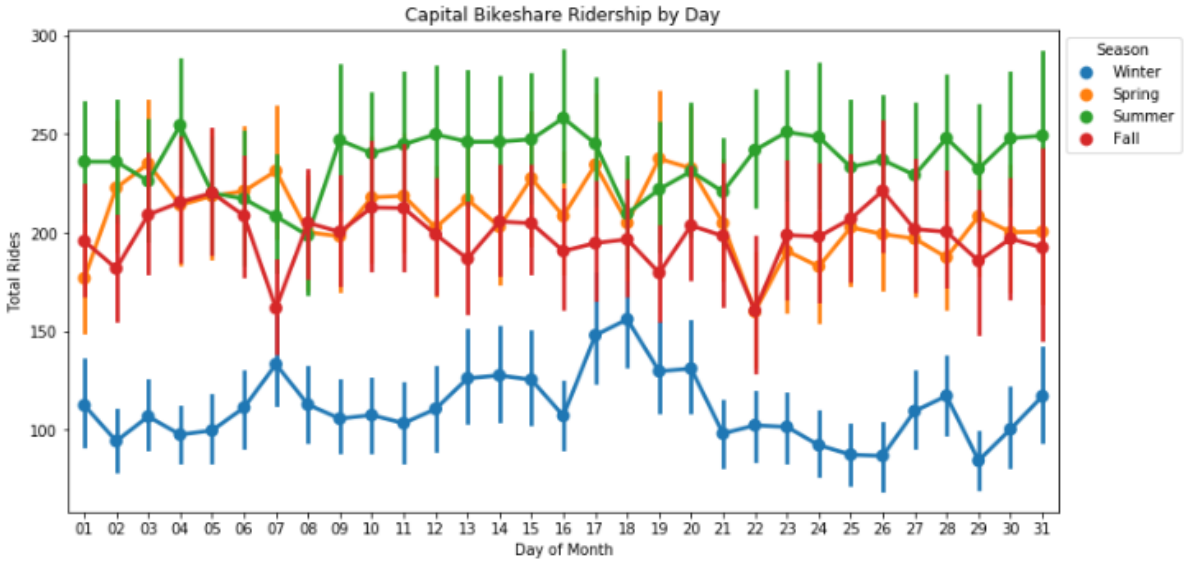
Yes, by seeing below picture, its easy to understand , some months are good with ridership normally from May to October, there are increasing the number of ridership.



Analysis 3:

Showing Capital Bike Sharing Ridership by Day:

Time _series Analysis:



So same as above two examples and graphs, here also showing the days of winter provides less ridership as compare with summer and in very first and last days of month , there are more chances of ridership.

Analysis 4:

Capital Bike Sharing Ridership by Membership:

Do ridership trends vary on membership type?

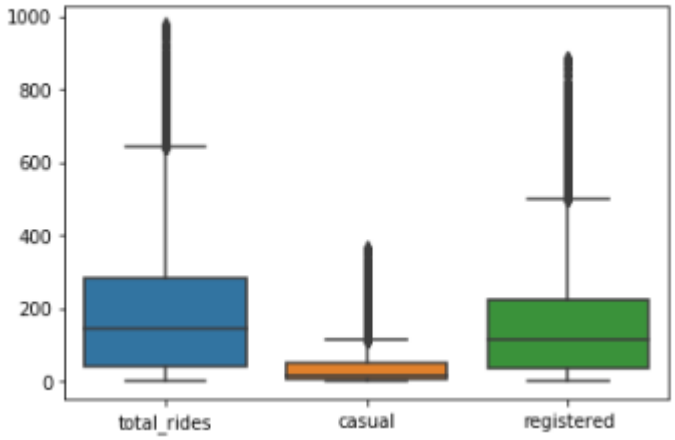
Yes. We can see below that registered riders make more trips than casual riders. The large T-value of 44.97.81 suggests a significant difference between the two groups. A very small p-value of 0.0 indicates that this is unlikely to have happened by accident

```
ttest_ind(df0['registered'], df0['casual'])
```

```
Ttest_indResult(statistic=97.81332643791566, pvalue=0.0)
```

Membership type by Box Plot

```
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x1ef3e4eb148>
```

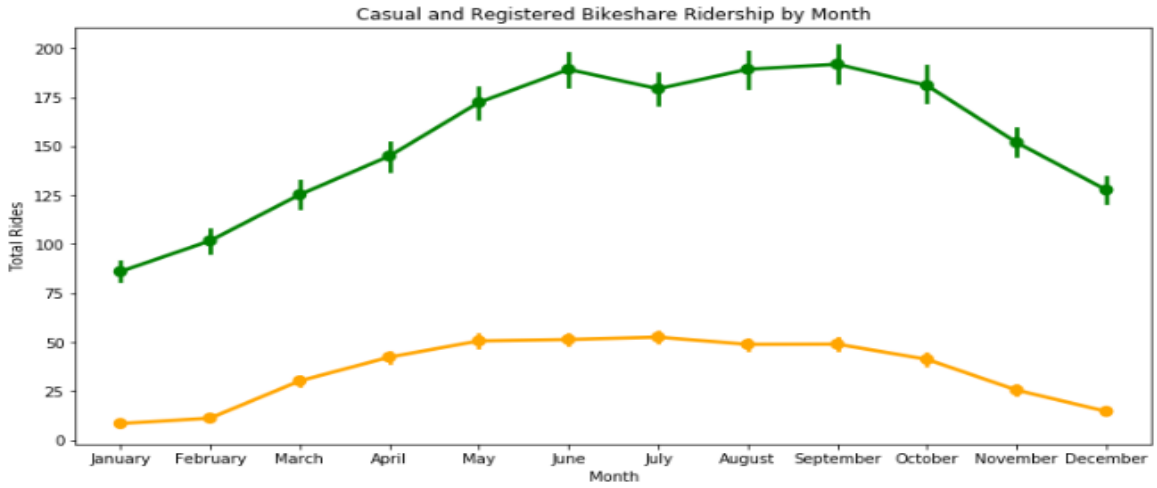


Analysis 5

Capital Bike Sharing Ridership By Weather:

Does really weather affects the bikeshare usage?

Yes. As we can see below, types of weather have a large impact on ridership. There are significantly less rides during snow and thunderstorms than during periods of nicer weather.



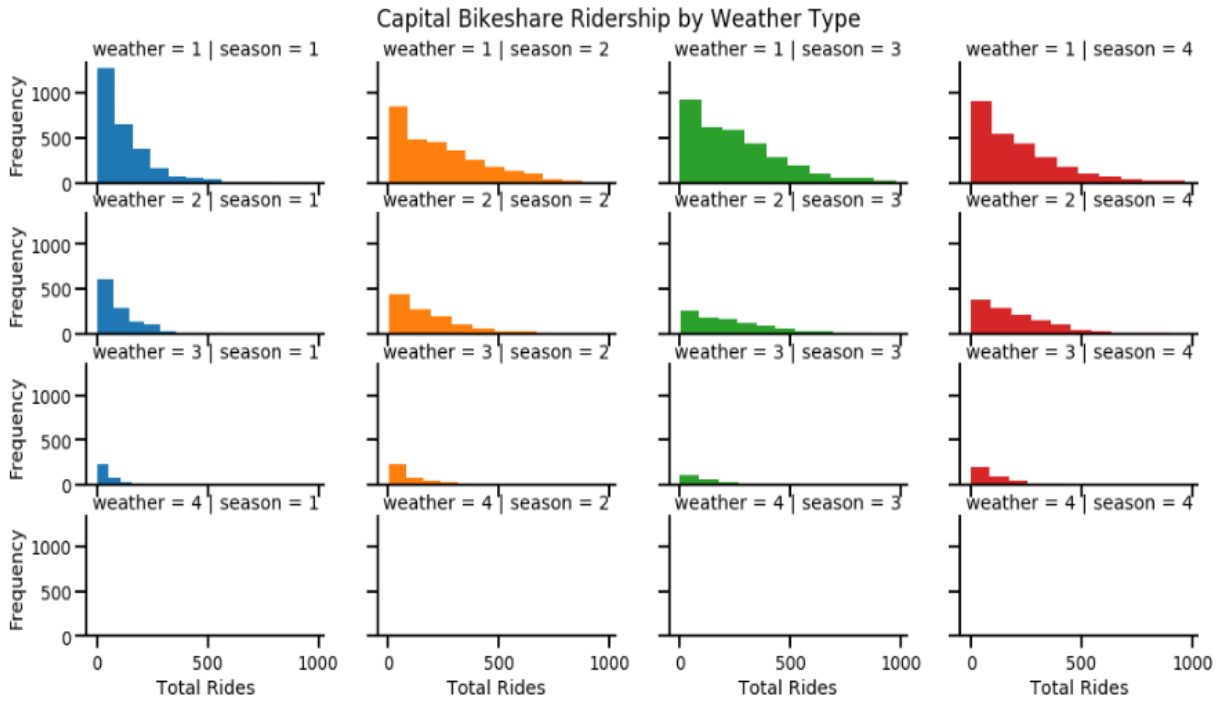
We can also see that this trend holds up across all seasons.

Types of Weather

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



Analysis with dependent variable

Our analysis focuses on finding the impact of various conditions on the number of bikes rented. Therefore, the most important attribute, which is the dependent variable for this analysis "Cnt". The “cnt” attribute is the number of bikes rented per hour per day.

There are 13 important independent variables that can affect the cnt function variable. Excluding the attributes from the regression analysis, the following attributes are removed to find the correlation between the important independent columns with the per column.

1.dteday: This attribute has been removed because it is an everyday date attribute. The date attribute does not support regression analysis. Therefore, the specific derived attribute is generated month (month), which will meet the requirements of the analysis.

2.yr: This is year attribute which tells if the date is of 2010 or 2011 year. This analysis is not looking for year on year growth. We are looking for natural and work schedule attributes and its effect on the bike rentals. So, this attribute is removed from the analysis.

Correlation

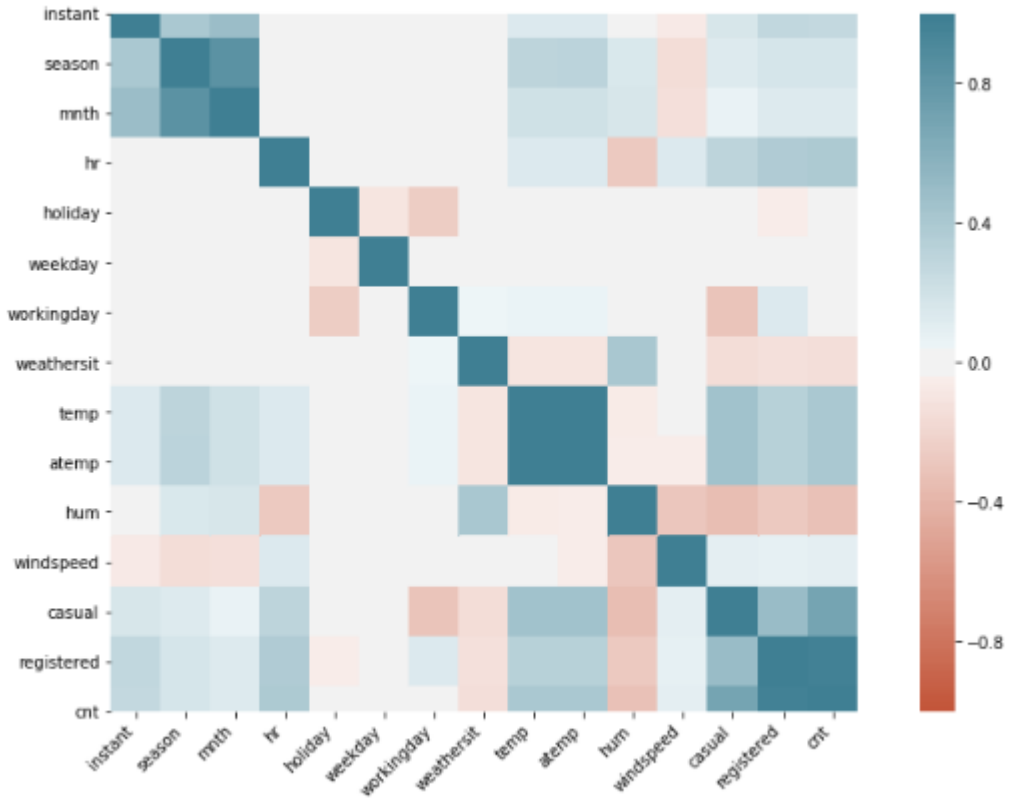
	season	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
season	1.000000	0.830386	-0.006117	-0.009585	-0.002335	0.013743	-0.014524	0.312025	0.319380	0.150625	-0.149773	0.120206	0.174226	0.178056
mnth	0.830386	1.000000	-0.005772	0.018430	0.010400	-0.003477	0.005400	0.201691	0.208096	0.164411	-0.135386	0.068457	0.122273	0.120638
hr	-0.006117	-0.005772	1.000000	0.000479	-0.003498	0.002285	-0.020203	0.137603	0.133750	-0.276498	0.137252	0.301202	0.374141	0.394071
holiday	-0.009585	0.018430	0.000479	1.000000	-0.102088	-0.252471	-0.017036	-0.027340	-0.030973	-0.010588	0.003988	0.031564	-0.047345	-0.030927
weekday	-0.002335	0.010400	-0.003498	-0.102088	1.000000	0.035955	0.003311	-0.001795	-0.008821	-0.037158	0.011502	0.032721	0.021578	0.026900
workingday	0.013743	-0.003477	0.002285	-0.252471	0.035955	1.000000	0.044672	0.055390	0.054667	0.015688	-0.011830	-0.300942	0.134326	0.030284
weathersit	-0.014524	0.005400	-0.020203	-0.017036	0.003311	0.044672	1.000000	-0.102640	-0.105563	0.418130	0.026226	-0.152628	-0.120966	-0.142426
temp	0.312025	0.201691	0.137603	-0.027340	-0.001795	0.055390	-0.102640	1.000000	0.987672	-0.069881	-0.023125	0.459616	0.335361	0.404772
atemp	0.319380	0.208096	0.133750	-0.030973	-0.008821	0.054667	-0.105563	0.987672	1.000000	-0.051918	-0.062336	0.454080	0.332559	0.400929
hum	0.150625	0.164411	-0.276498	-0.010588	-0.037158	0.015688	0.418130	-0.069881	-0.051918	1.000000	-0.290105	-0.347028	-0.273933	-0.322911
windspeed	-0.149773	-0.135386	0.137252	0.003988	0.011502	-0.011830	0.026226	-0.023125	-0.062336	-0.290105	1.000000	0.090287	0.082321	0.093234
casual	0.120206	0.068457	0.301202	0.031564	0.032721	-0.300942	-0.152628	0.459616	0.454080	-0.347028	0.090287	1.000000	0.506618	0.694564
registered	0.174226	0.122273	0.374141	-0.047345	0.021578	0.134326	-0.120966	0.335361	0.332559	-0.273933	0.082321	0.506618	1.000000	0.972151
cnt	0.178056	0.120638	0.394071	-0.030927	0.026900	0.030284	-0.142426	0.404772	0.400929	-0.322911	0.093234	0.694564	0.972151	1.000000

Above is correlation of cnt dependent variable with every independent variable numerically. There is a strong relationship with casual , registered, atempt, temp and hr.

Next, we will see this result by graph.

Correlation by Heatmap

Where the color is brighter there is more strong relationship in heatmap, so same as above casual, registered , atemp, temp and hr. are looking good against dependent variable regarding strong relationship.



To see more clear relation of cnt with every independent variable, let's see another method.

Pearson correlation

Filter Method:

As the name implies, in this method, you only take a subset of the filter and related features. The model is built after selecting the features. Here filtering is done using the correlation matrix, and this is done using Pearson correlation. Here we will first plot the Pearson correlation heatmap and look at the interactions of independent variables with the output variable cnt. We will only select features that are greater than 0.3 (taking the absolute value) associated with the output variable.

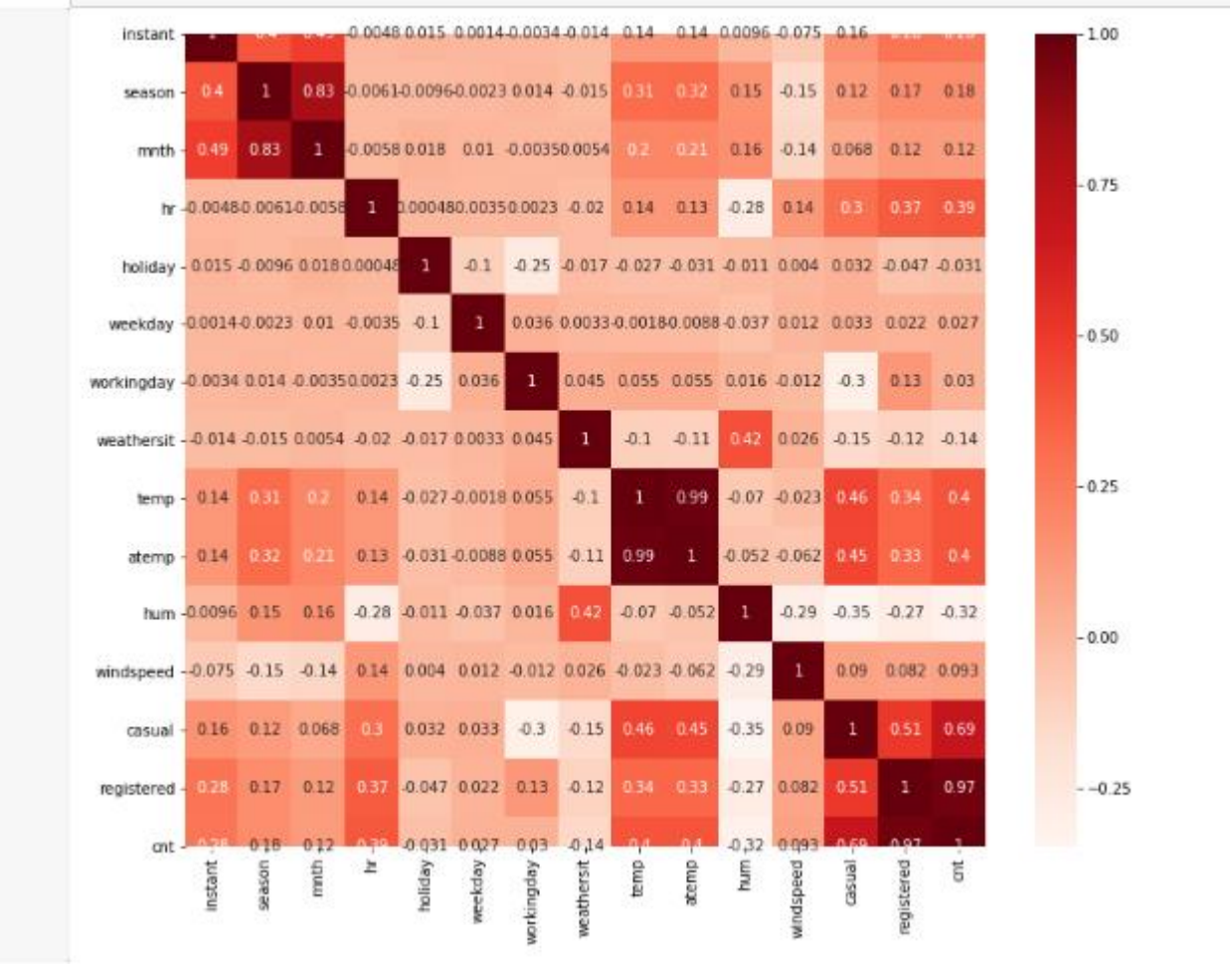
The correlation coefficient has values from -1 to 1

- A value close to 0 indicates a weak correlation (a correct 0 indicates no correlation)

- A value close to 1 indicates a strong positive correlation

- A value close to -1 indicates a strong negative correlation

— A value closer to -1 implies stronger negative correlation



Independent Variable	Dependent Variable	Coefficient Correlation
Casual	Cnt	0.69
Registered	Cnt	0.97
Temp	Cnt	0.4
Atemp	Cnt	0.4
Holiday	Cnt	0.4
Hr	Cnt	0.39

The above table shows the best coefficient correlation, and these are the best independent features.

Regression Analysis

We will run a regression analysis on the independent variables to generate a multilinear regression model. This multiple linear regression model will help s formulate the relation between depended(cnt) and independent variables.

We find this by calculating R square.

R Square

```
: from sklearn import linear_model

  lr_model=linear_model.LinearRegression()
  lr_model

: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

: X=df0[['hr','holiday','atemp','temp','casual','registered']]
  lr_model.fit(X, df0['cnt'])

: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

: print('The R-square is: ', lr_model.score(X, df0['cnt']))

  The R-square is:  1.0
```

R square is equal to 1 means our multiple linear regression is very good and having no anomalies or strong relationship between dependent and independent variables and our multiple linear regression model will be perfect in that case.

Anova Analysis

Next important thing is ANOVA (Analysis of Variance). ANOVA issued to compare variances between two groups or datasets. Here the ANOVA is used to comparing the dataset that was provided as input after best feature selection . As we can see that F-test is very large P values is

very small , this proves that our regression model is a good model for predicting the dependent variable.

	sum_sq	df	F	PR(>F)
C(cnt)	1.712755e+03	868.0	9.109441e+24	0.000000
holiday	1.793398e-26	1.0	8.279288e-02	0.773551
atemp	1.470289e-26	1.0	6.787642e-02	0.794457
registered	4.576372e+07	1.0	2.112699e+32	0.000000
casual	2.882180e-10	1.0	1.330569e+15	0.000000
temp	4.223059e-25	1.0	1.949591e+00	0.162650
Residual	3.575408e-21	16506.0	NaN	NaN

Summary and evaluation of the Null Hypothesis

Using the regression model and the Pearson correlation analysis, anova analysis, and heatmap (correlation) in the previous sections we can see that bike rentals depend on weather conditions , seasons , days ,member type and work schedule of people. We have also proved that the regression model is statistically significant by analyzing the R square, ANOVA and descriptive statistics. So, we can safely reject the Null hypothesis and accept the alternate hypothesis mentioned below.

Null Hypothesis:

Bike rental usage does not depend on weather conditions, season member type.

Alternate Hypothesis:

Comfortable weather conditions help increase the number of bike rentals .This is the initial hypothesis, and there will be other factors that will be evaluated, like time of the day, holiday or workday, to get insight into how these factors impact the bike rentals. Months starts from April to October helps to increase bike rentals.

Also, winter season has a smaller number of bike rentals however their numbers are increasing in summer.

Conclusion

I started to see first relevant examples of bike sharing over internet, through these systems, user can easily rent a bike from a position and return at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues. Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors .

Once I got some of knowledge about dataset, I downloaded it, explore it, during exploring found that this was very clean dataset, some of the analysis have done so far, by which one can say that many people mostly use the service in summer, starting days of month and from April to October.

References:

Bhushan Yadav Capital Bike sharing usage Analysis [Retrieved from https://www.academia.edu/37848901/Capital_Bikeshare_usage_analysis]

Long, K.. (2015, May 7). Which capital bikeshare stations see the highest traffic? [Blog post]Retrieved from https://districtmeasured.com/2015/05/07/which-capital-bikeshare-stations-see-the-most-traffic.]

Tara Boyle (2018) Capital Bike share exploratory analysis [Retrieved from https://www.kaggle.com/tboyle10/capital-bikeshare-exploratory-analysis]

Buck, D., Beauhler, R., Hop, P., Rawls, P., Chung, P., & Borecki, n. (2013). Are bikeshare users different from regular cyclists? First look at short-term users, annual members and area cyclists in the Washington, D.C. region. Traffic Research Record: Journal of the Transport Research Board, (2387), 112-119.

DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. Journal of public transportation, 12(4), 3.

El-Assi, W., Mahmoud, M. S., & Habib, K. N. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. Transportation, 44(3), 589-613.

Yang, S., & Huang, M. (2017). OFO Bike Sharing: Riding on a flat road. University of Hong Kong.

Verhof, b. C., Reynolds, Wm. J., & Kraft, M. (2010). A new perspective on customer management is customer involvement. Service Research Journal, 13 (3), 247-252.