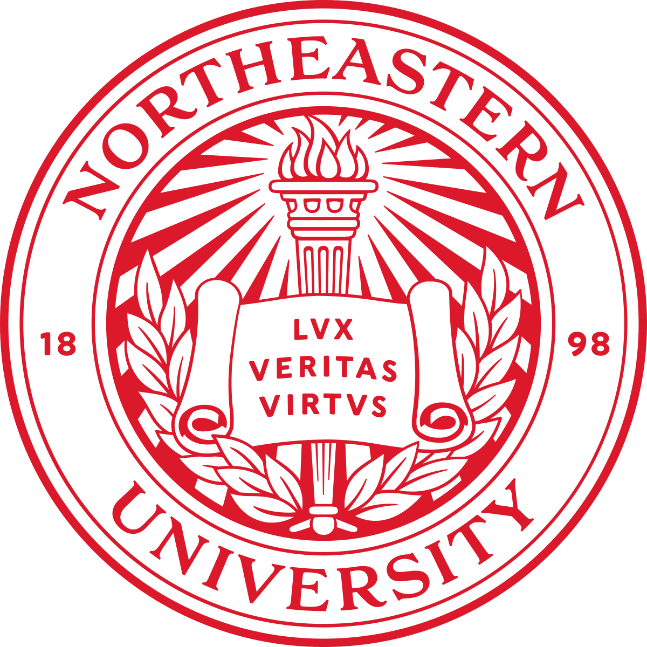
****

**ALY6110 71588 Data Management & Big Data**

**Module 6: Final Project Report - Group2**

**Instructor: Daya Rudhramoorthi**

**Date : 10/24/2020**

**Team Members**

**Priyanka Nagesh Adiga**

**Sindhusha Karra**

**Anubhav Saha**

**Roshni Vadiraja**

**Introduction**

In this report, we are exploring the Capital Bike sharing dataset, made available to us by the UC Irvine Machine learning repository [1]. We start by taking a look at the dataset, the variables involved and then clean the dataset for model building. The Capital bikeshare system collected the historical data logs in Washington DC over a period of two years, i.e. 2012 & 2012. With our analysis in this paper, we plan to determine the correlation of environmental factors like the weather conditions, season, time of the day, day of the week, etc. with the bike rental behavior of the public and then use the results from this analysis to come up with a strategy to promote the usage of bike sharing in both the casual and registered users, whilst also aim at converting the casual users into registered ones.

**Business Understanding**

Nowadays, more and more people are opting for environmentally friendly and healthy ways of transportation and riding a bike to the workplace or school is the best option for a lot of people to live healthy and eco-friendly. Modern day bike sharing system not only allows this to happen by renting bikes to people with the convenience of their mobile app and avoiding contact with people, but also serves as a cheap and fast way of getting around the city by avoiding traffic on the road. As more people are preferring this bike sharing option, there is an increase in the market space for such companies apart from the ones supported by the city administration and with greater market opportunity, it becomes important to look closely at the data and identify the renting habits of people in order to make the rental process more attractive to potential customers and increase the overall revenue. The mobile applications mentioned earlier, not only allows the users to identify nearby docking stations and pay for the rental, but also helps in tracking user locations throughout the ride and gather important data of movement of people around the city. This data can be crucial to identify the mobility of people within the city or a given area. We will begin this report by an exploratory data analysis of our two dataset, namely hour.csv and day.csv and present our analysis results. That will be followed by a dashboard presenting this result in the form of visualizations. Finally, we will build models to predict an outcome and test its accuracy and then conclude on our business problem with our findings.

**Data Understanding**

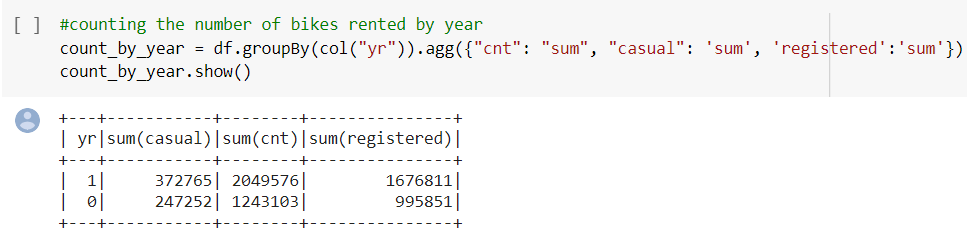
Our dataset consists of two different csv files, namely hour and day csv. These csv files are imported as data frames. The day\_dataframe set consists of 731 records with a total of 16 columns, whereas the hour\_dataframe set consists of 17379 records with a total of 17 columns, the additional column being the hour of the day. Most of the variables are named in a self-explanatory way but coded as numbers representing a categorical variable. The name of these variables and their datatypes are shown by the screenshot of the output of the info() method below:

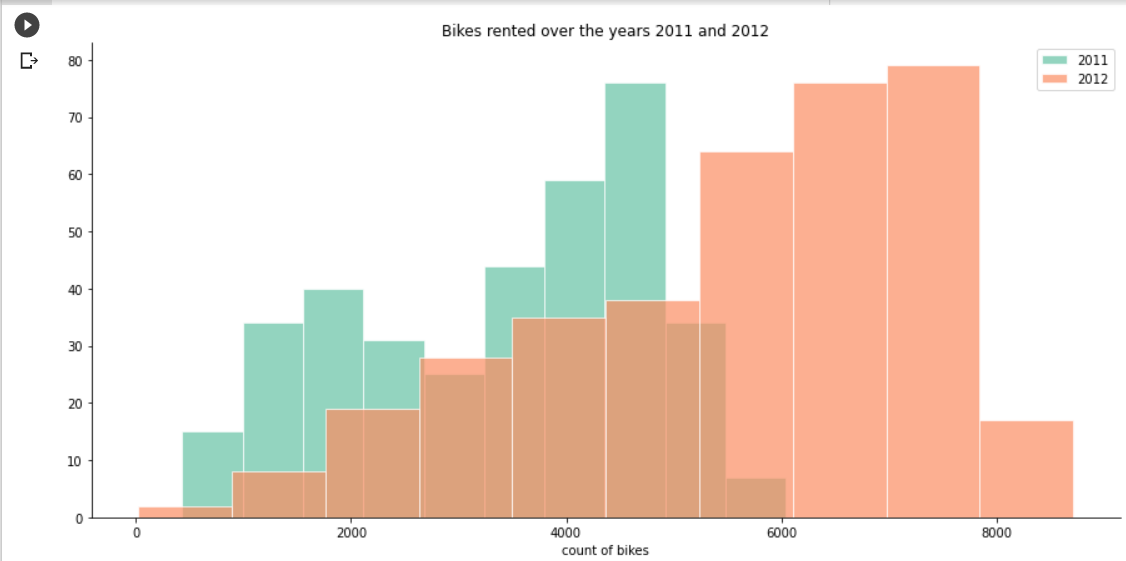
|  |  |
| --- | --- |
|  |  |

The variable season is of type int because it is coded as 1 being spring, 2 being summer, 3 being fall and 4 being winter. Likewise, yr 0 represents 2011 and yr 1 represents 2012. For holiday and working day variables, 0 and 1 act as False and True categories, respectively. Finally, the variable weathersit is coded as 1 being clear, 2 being misty, 3 being light showers and 4 representing heavy rainfall.

**Exploratory Data Analysis**

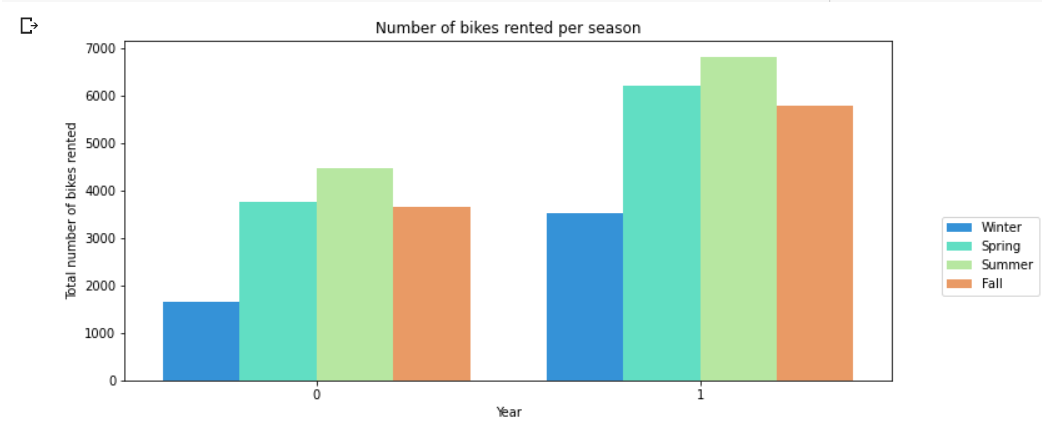
We continue our analysis by counting the number of registered and casual bikers each year as follows:





The graph above clearly shows us that the number of bike rentals have gone up significantly from the year 2011 to 2012.

It will be helpful to view the number of bikes rented across the two years segregated by different seasons to get a sense of the popularity of seasons. We use the following chart to achieve that:

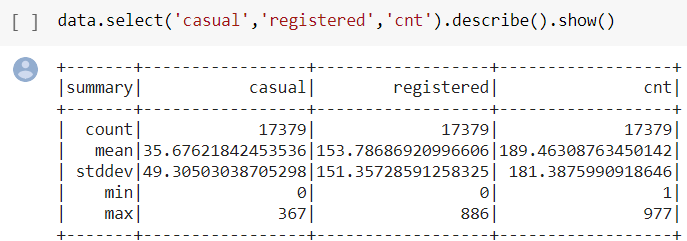


From the graph above, we can immediately notice that the number of bike rentals is the least during winter, which makes sense because of extreme cold winters across the Northeast.

Also as expected, Summer seems to be the most popular season in both 2011 and 2012 as the weather is usually nice and warm which is suitable for outdoor activities.

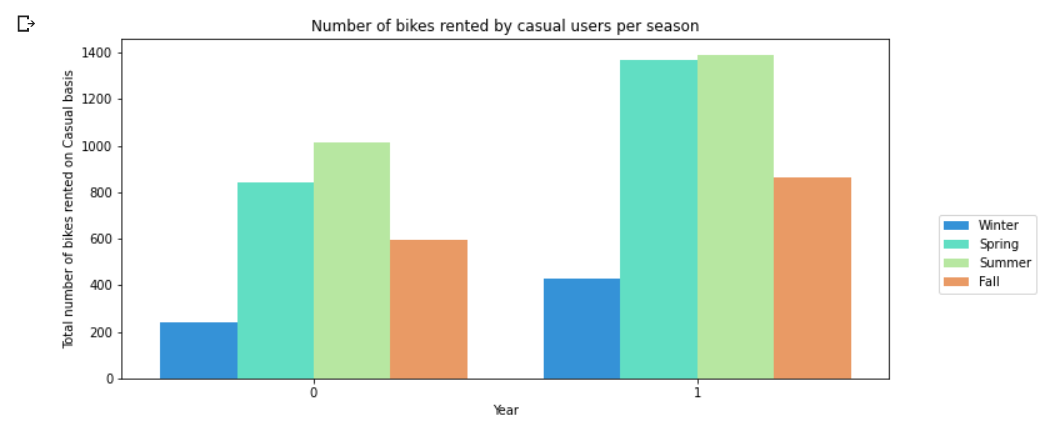
Fall and Spring are neck to neck in both these years, with Spring slightly having the edge.

It will be useful now to determine the count of the total number of casual bikers and total number registered bikers and obtain some statistics on those numbers as shown below.



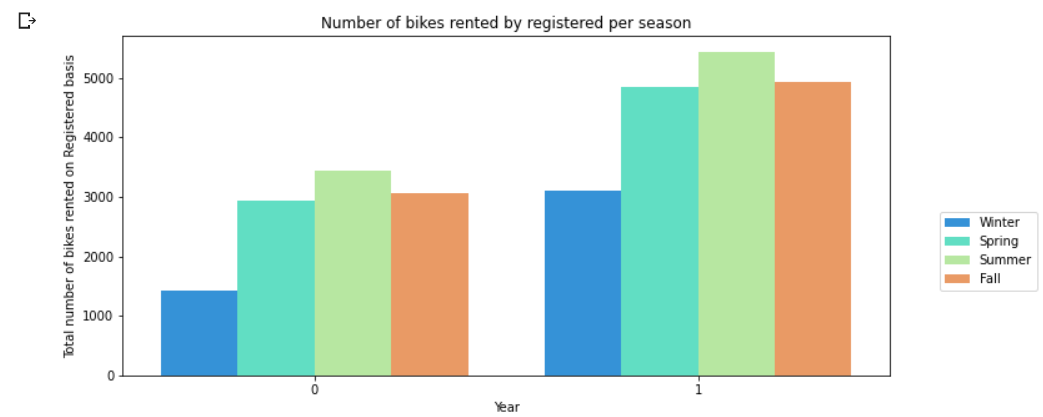
From the above stats, we can see that the maximum number of registered bikers was 886 and for casual bikers it was 367, almost less than a half of registered bikers. Registered bikers have an average count which is 5 times higher than casual bikers.

It will be helpful to visualize the results of casual users and segregate them over different seasons to understand the trend of rentals more deeply as shown below.



From the above graph, we can immediately notice a rising trend of rentals on the year 2012 compared to 2011, more significant rises occurring in the seasons of spring and summer.

Similar graph has been obtained for the registered users as shown below.



Similar to the case of casual bikers, here we can see an overall rising trend with the seasons of Spring and Summer seeing a steep rise. The major difference compared to the casual users being the significant updraft of the Fall bar as well, indicating the number of registered users increased in the Fall more proportionally than the casual users.

It will be helpful to take a look at what is happening within these seasons, which is visualized by the following graphs.

|  |  |
| --- | --- |
|  |  |

From the graph of the summer season above, we can see the rental trend of weekend/holiday falling till the month 8 and then rising again during after that, and the opposite trend on the weekday.

For the graph of the Winter season, the number of rentals is higher during weekends or holidays compared to the bike rentals weekdays.

Similar trends of the Spring and Fall seasons have been shown below:

|  |  |
| --- | --- |
|  |  |

From the two graphs above, we can see opposite trends, while the renting rises on the spring season on both weekends and weekdays, it falls at a similar pace during the Fall season.

This leads us to our next analysis on the impact of weather on these rental trends that we observed above as change of weather seems like a likely cause in the changing of trends in-between seasons.

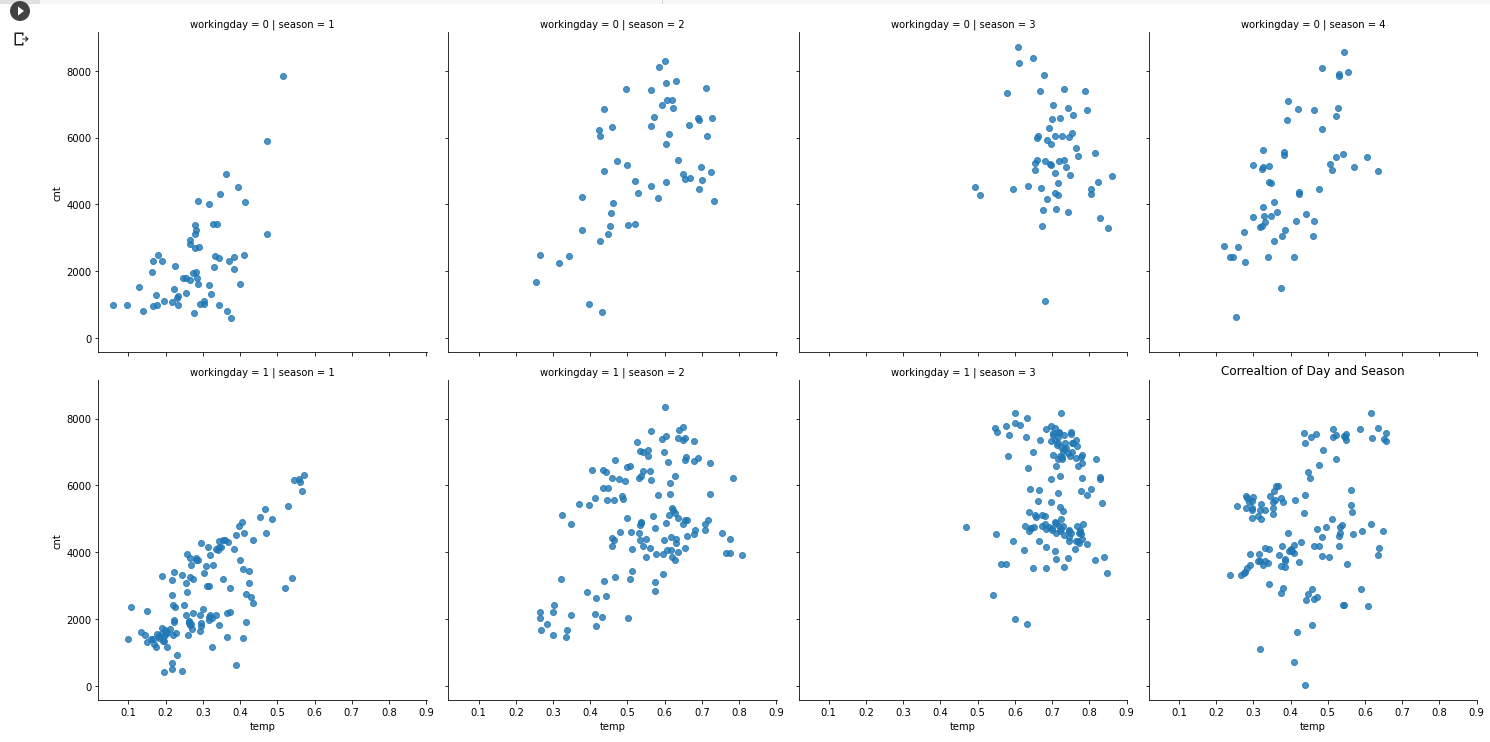
|  |  |
| --- | --- |
|  |  |
|  |  |

From the graphs above, we can see that the renting trends more strongly correlate to the weather than the seasons themselves as renting goes up during clear skies and falls during thunderstorms.

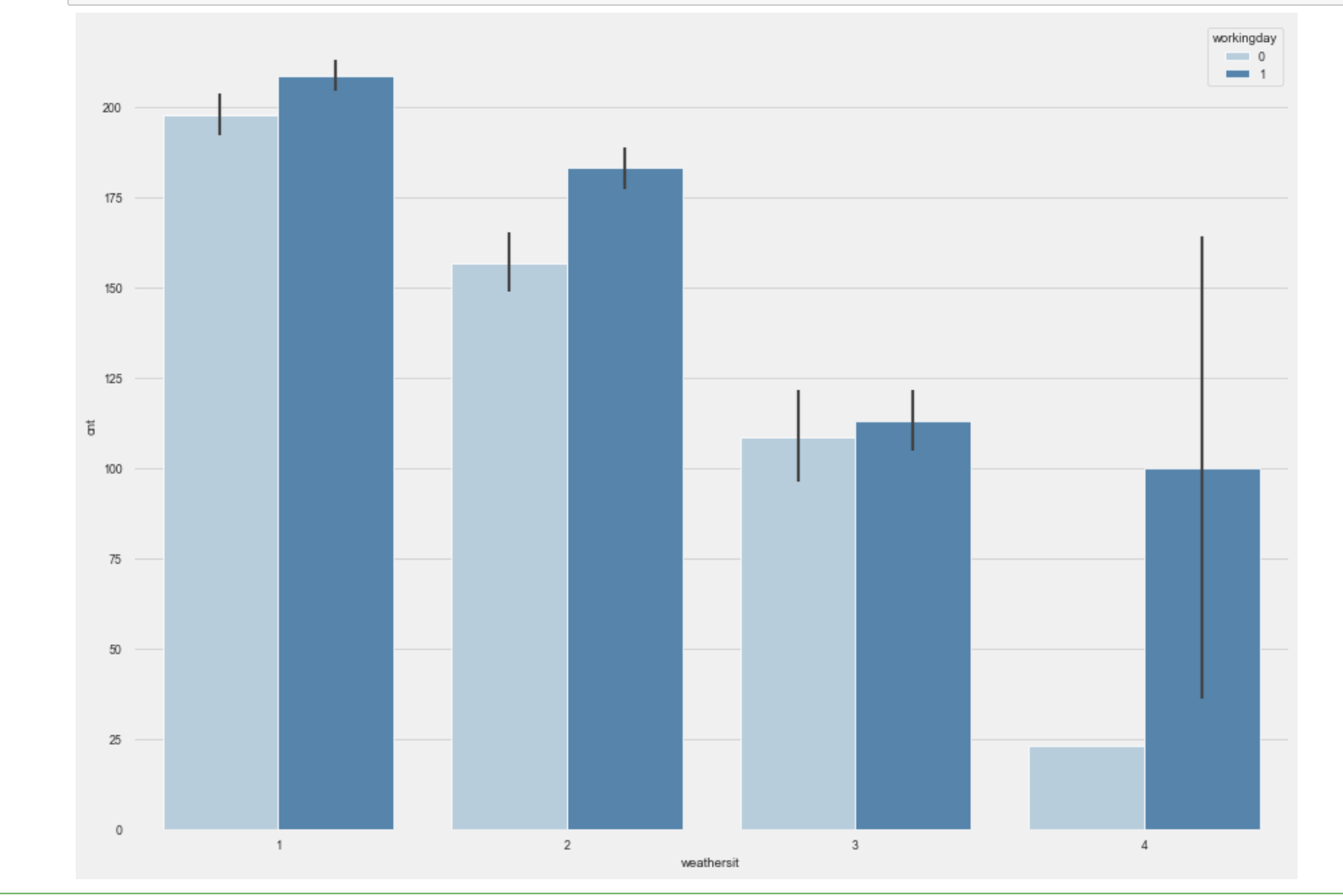
This supports our hypothesis that people prefer renting bikes more when there are clear skies, and they don’t prefer raining when there’s thunderstorms or heavy rains.

**Temperature impacts on Bike rentals**

Now to identify the ideal temperature for bike rentals, irrespective of the seasons or the weather conditions, we can take a look at the scatterplots below:



**Impact of Holidays and Weather**

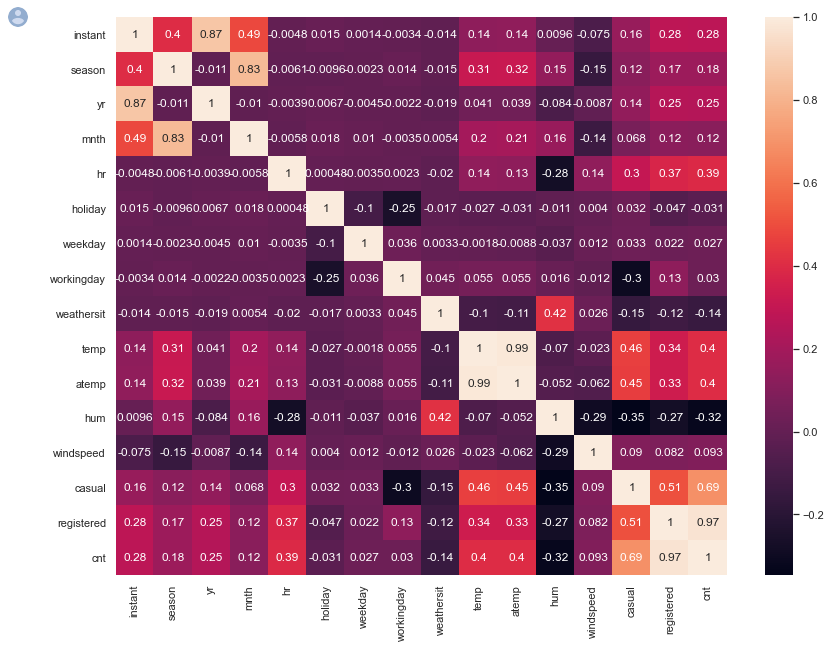


Now to look deeply on the impact of weather during the regular commute on working days and weekends, we will make use of the above bar plot.

The darker blue bar indicates the working days, and the lighter blue showcases the holidays. The variable weathersit is coded as 1 being clear, 2 being misty, 3 being light showers and 4 representing heavy rainfall. The stark difference appears for the weather conditions light showers and heavy rainfall where the biking rentals drop significantly for the holidays, more compared to the working days.

This gives us a very important insight that people who rent during the working days must use those bikes to commute to schools or offices, hence little showers don’t bring that rate down drastically compared to what happens when it’s a holiday or a weekend.

**Correlation Matrix Breakdown.**



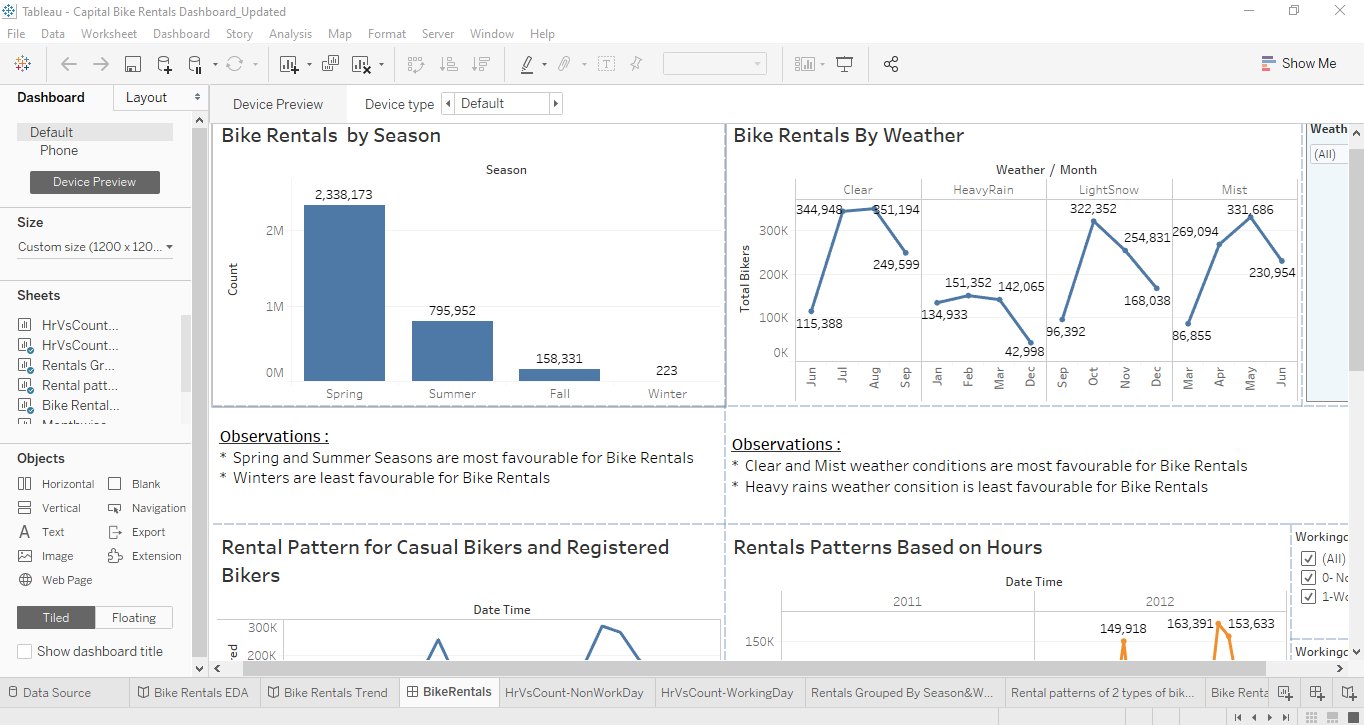
We can use the correlation matrix above to see the correlation among different variables in our dataset. The lighter the color, the stronger is the correlation.

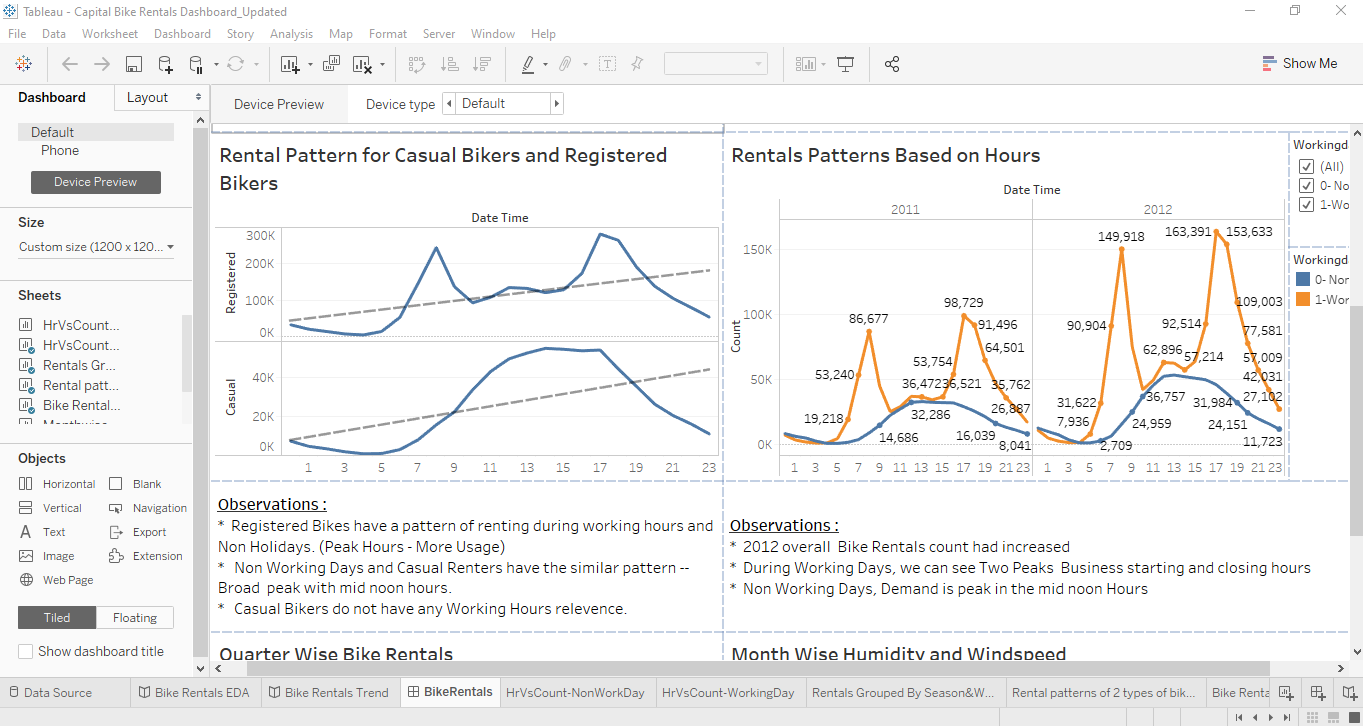
Some of the highest correlations can be found between the registered users and their high rentals during the weekdays compared to the casual users who are highly correlated with the weekends, supporting the results of our previous analyses.

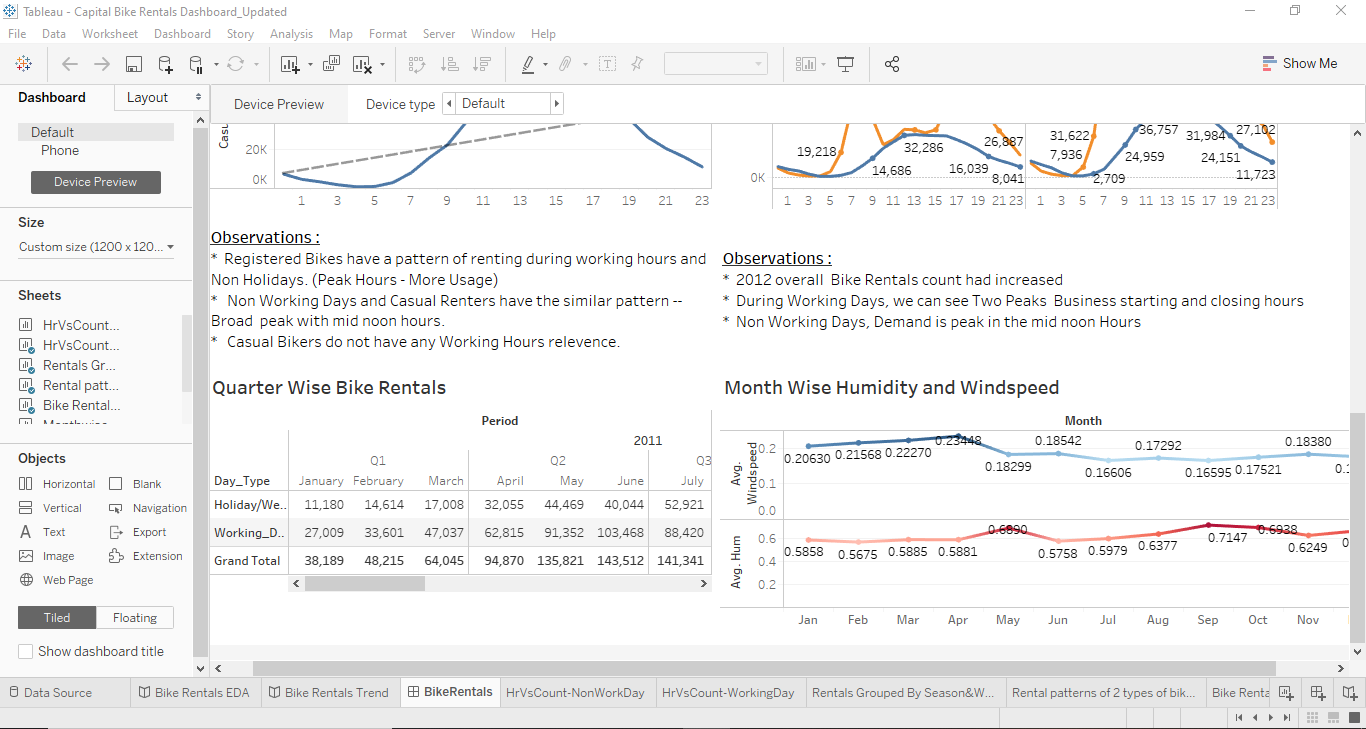
**Dashboard**

For our Dashboard, we are using the BI tool Tableau and creating 6 different charts that highlight on the bike rental trends as per the weather conditions, the seasons, patterns based on hours of the day, patterns based on the status of registration and finally the humidity and windspeed impacts closely.

The screenshots of our dashboard have been shown below by scrolling the screen downwards using the toggle bar on the far right.





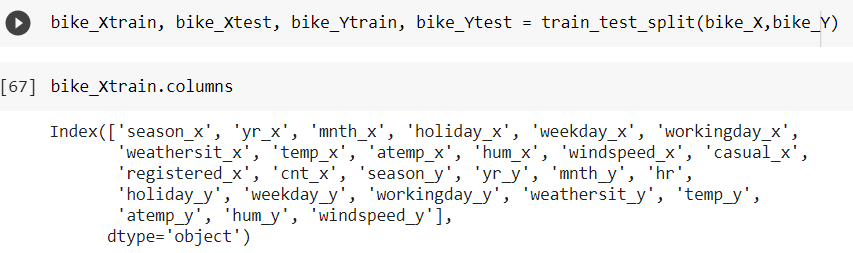


Our dashboard solidifies the results obtained earlier on our EDA and confirms the same observations that we obtained previously using Python generated charts.

**Data Preparation**

We have created a regression model in order to predict the bike rentals on an hourly basis. Prior to constructing the model, the data is pre-processed in order to remove some of the attributes such as date and count of registered rentals and count of hourly rentals in order to determine all the other attributes that contribute in predicting the count of the bike rentals.

As a part of data preprocessing step attributes such as date, instant, number of casual bike renters, number of registered bike renters are removed from the dataset under analysis. This step is followed by splitting the data set into training and test sets and further grouped into dependent and independent variables. A glimpse of the training data after preprocessing can be seen below.

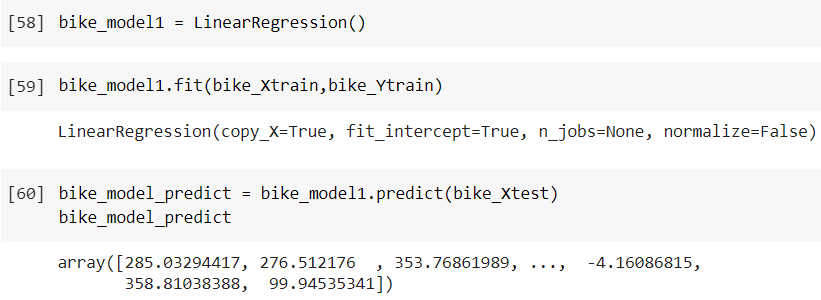


This is followed by model construction.

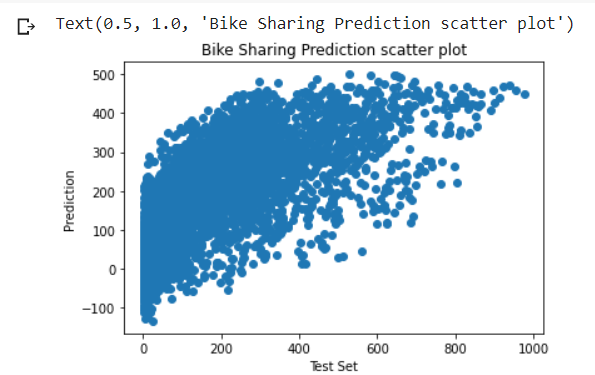
**Modelling**

**Linear Regression Model**

We are implementing a linear regression model. And the model is trained using the train datasets as shown below. This is followed by predicting the number of bike rentals using the variables available in the test dataset. The steps performed in building the model, training the model prediction can be seen below.



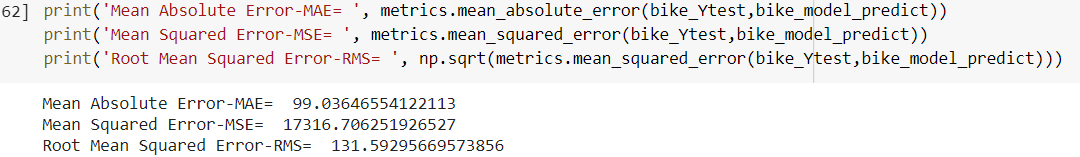
The prediction results are plotted on a scatter plot for better understanding of the model outcome.



From the above plot we can see that the model has predicted a numerically higher number of bike rentals compared to the test set.

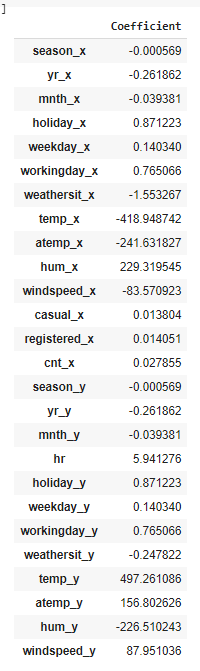
**Evaluation**

The next in this process is to determine the Mean Absolute Error, Mean Squared Error and Root Mean Squared error as shown below. This will help us validate our results.

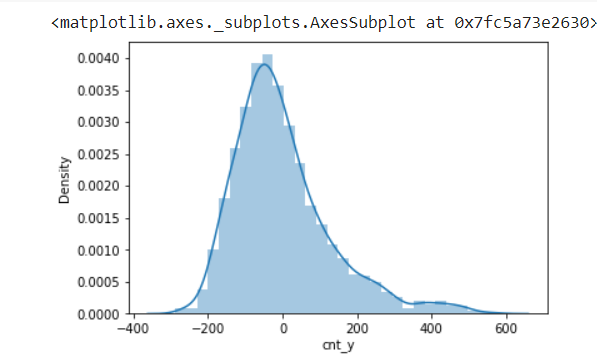


The smaller the root mean squared error, the closer we are to finding the line of best fit. However, the value of these metrics mainly depends upon the dataset under consideration. If the data are widely scattered as shown in the previous plot, the chances of root mean squared error being higher are more.

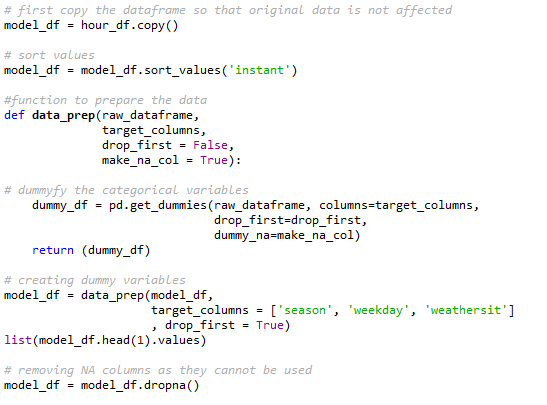
Below is the list of coefficients assigned to different attributes by the model.

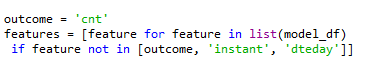


Following is the density plot of the results of our analysis. Density plot provides a representation of the distribution of the count of bike rentals. The following density plot uses a kernel density estimate to show the probability density function of the bike rentals count variable.



**Data Preparation for Non-Linear Modelling**

****

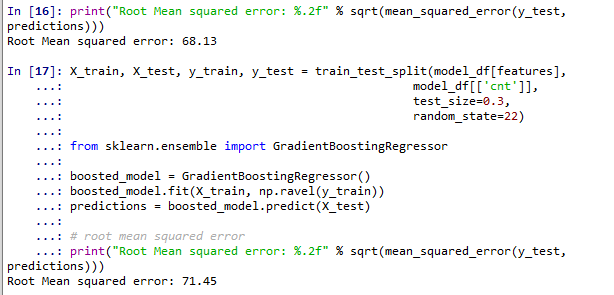
****

**Modelling and Evaluation**

**Non-Linear modeling using Gradient Boosting:**

We can further improve the RMSE score by opting for a gradient boosting regressor after using dummy variables for our categorical ones.

The RMSE that we obtain using GBR is significantly lesser than the linear model, as it varies from 68 to 72 based on different proportions of the testing set used.

****

**Time Series Analysis**

We performed time series analysis using three different techniques.

1. Simple moving average
2. Weighted moving average and
3. Exponential smoothing.

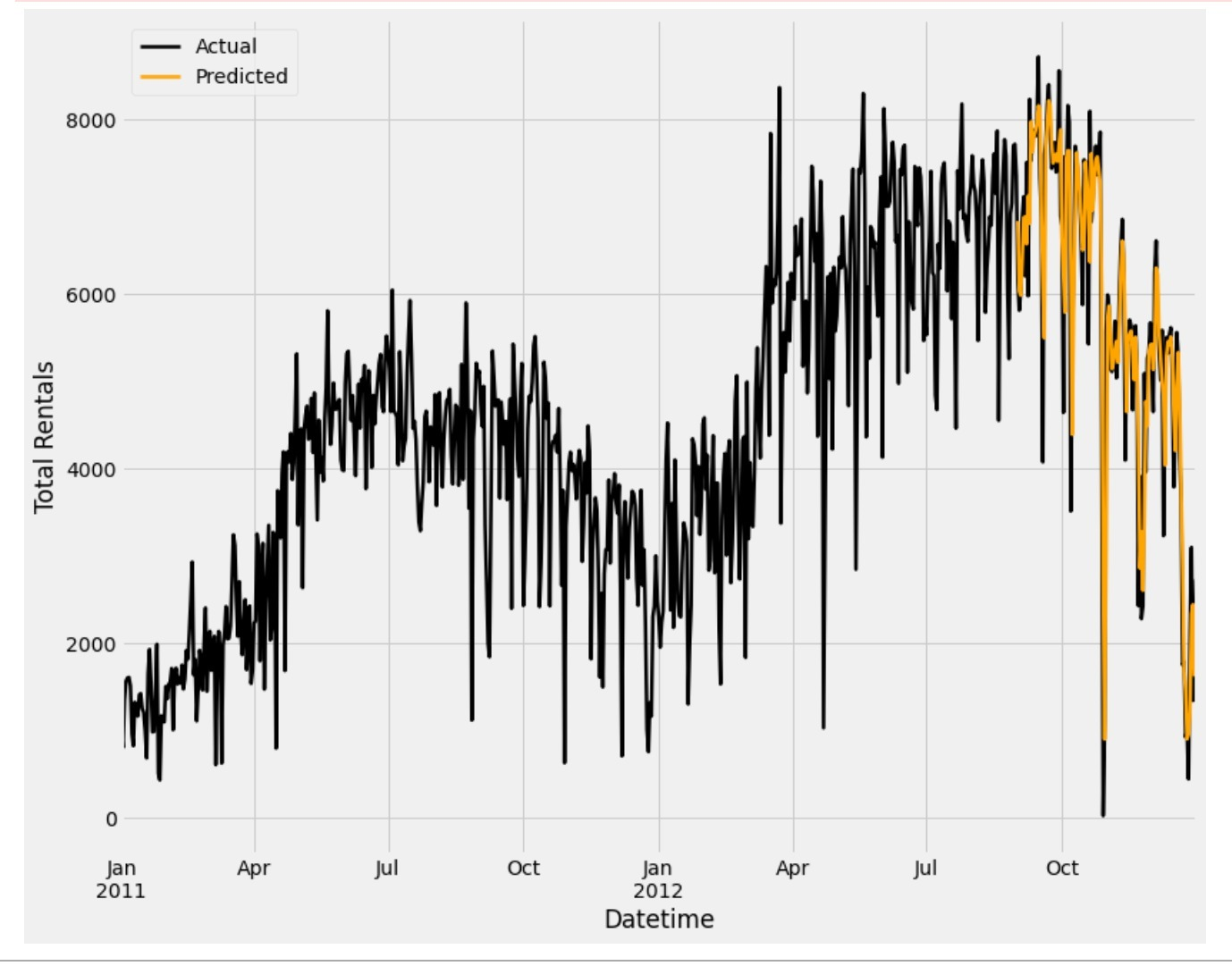
The predictions were made of the following columns using the above methods

1. Number of casual users renting bikes per day
2. Number of registered users renting bikes per day and
3. Number of bike rentals per day

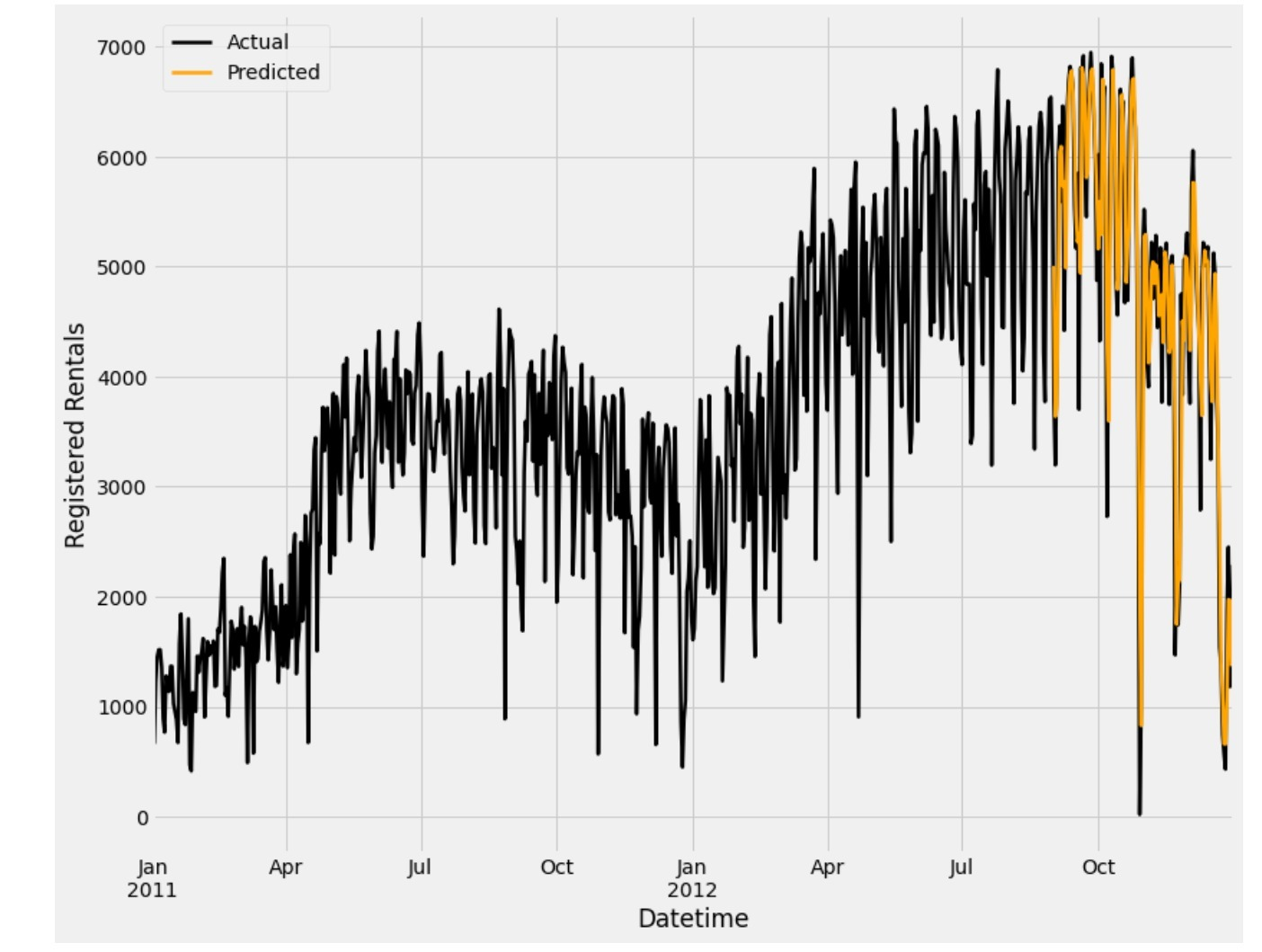
We have 2 years’ worth of data of which 18 months was used as training data and 6 months as testing data. When we ran the 3 months on the available data to predict the aforementioned columns, the most accurate method to predict turned out to be the Weighted moving average method. The prediction results when using the weighted moving average method is as follows

As we can see from the below 3 graphs, the predicted rental totals is very close to the actual rentals.

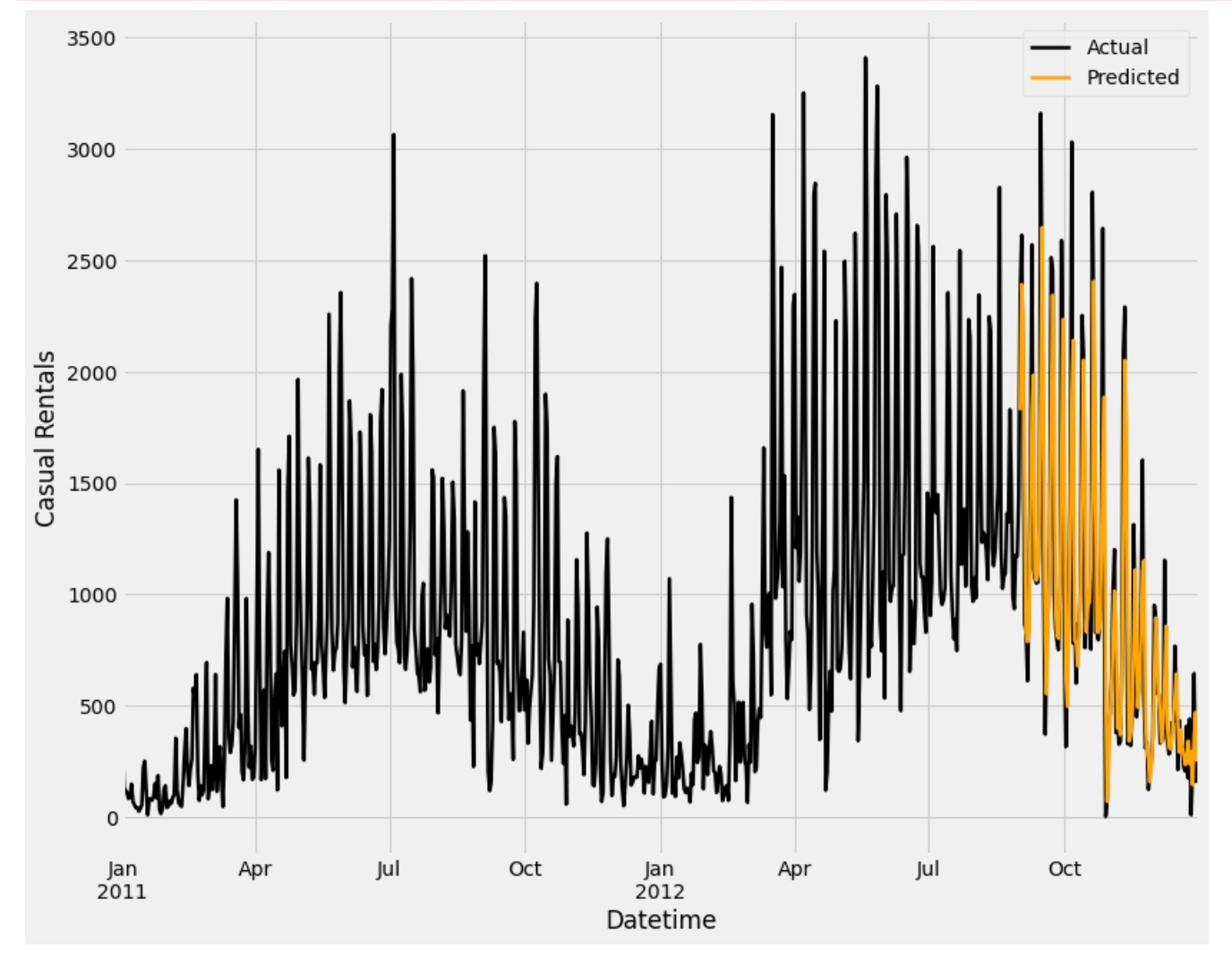
Prediction of total rentals per day



Prediction of Registered rental users per day



Prediction of Casual rental users per day



**Conclusion / Answers to the business questions**

For the business questions raised earlier in this paper, we can answer them using our analysis results in the above sections as follows:

1. **Should more money be invested in the bike sharing system with a good ROI chance?**

**Ans**: Yes. The results above indicate that the trend in the bike sharing system is likely to go up and hence market for such system will go up too.

1. **How should the maintenance be scheduled looking at the availability?**

**Ans**: As per the results above, it would be advisable to consider winter season for all the maintenance work as other seasons are heavily preferred for bike renting.

1. **Can the casual users be turned registered?**

**Ans.** Marketing can be done based on the results above to attract specific users, be it casual or registered. However, the casual users can be given incentive to become registered based on their rental activity.

1. **Can more people be encouraged and attracted to this bike sharing system?**

**Ans.** The results indicate that if the weather conditions are favorable, more people can be attracted to try out the bike sharing systems, and they are likely to do so increasingly in the future. By specific marketing based on the data available, this can be sped up significantly.

**Reference**

Hadi Fanaee-T, UCI. (n.d.). Center for Machine Learning and Intelligent Systems. Bike Sharing Dataset Data Set. Retrieved from [https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset#](https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset)

Stellar. (22 August 2017). CRISP-DM Process Diagram. After 20 years, CRISP-DM still a leader in data mining models. Retrieved from <https://www.stellarconsulting.co.nz/data/crisp-dm-still-a-leader/>