Data Science Course: Capstone Project 1

Milestone Report

Capstone Project : Statistical Data Analysis

Name: Xingkai Wu

Problem

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world. One challenge that bike- sharing program needs to solve is the demand prediction based on the weather situation, temperature, wind speed, humidity, holiday, day of the year and the time of the day in order to allocate the bikes at different kiosk locations. As for now, predicting the overall bike-sharing demand is this project's goal based on the Kaggle competition's requirement, the data is provided by Capital Bike- sharing program in Washington, D.C.

Client

As mentioned in the previous session, by predicting the overall bike demand, it can be used for allocate bikes at different kiosk locations to meet the demand and therefore promote memberships to maximize revenue.

Data Set

The hourly and daily logged data is found on http://capitalbikeshare.com/system-data, and the weather data is available on http://www.freemeteo.com, the aggregated data from 2011 to 2012 is available on https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset, all the data wrangling and storytelling and statistical analysis are based on the aggregated data.

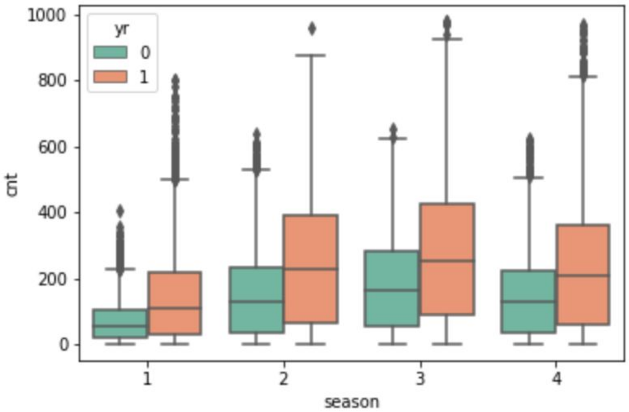
Data Wrangling

- Inspect the data type for each column, change the data type into categorical data for season, workingday and holiday columns.
- In windspeed column, the mode value is 0, zero wind speed is almost impossible, therefore they are missing values, by using Random Forest in sklearn package, based on the existing values of season, weathersit, hum, mnth, temp, yr, atemp from non- zero wind speed to predict and replace the zero valued windspeed.
- Plot scatter plot of each feature vs. target, i.e. cnt, and drop the columns that do not seem to affect cnt. Atemp seems correlated with cnt, but it is almost the same as temp, as it is feeling temperature, therefore we should drop the atemp column.

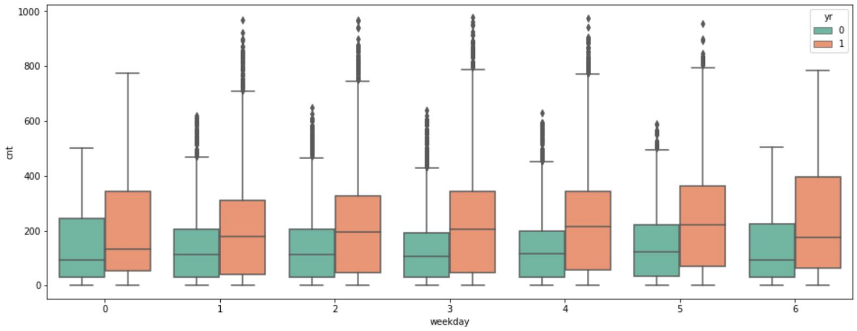
Initial Finding

As expected, the bike sharing demand is correlated with temperature, humidity and windspeed. The demand is also depend on the season of the year, day of the week and the time of the day.

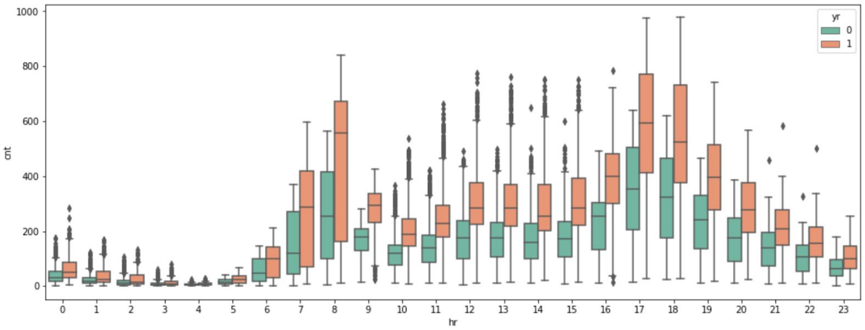
Interestingly looks like Summer and Fall has much higher demand than that in Spring and Winter.



It looks like an outlier usage only happens on Monday through Friday.

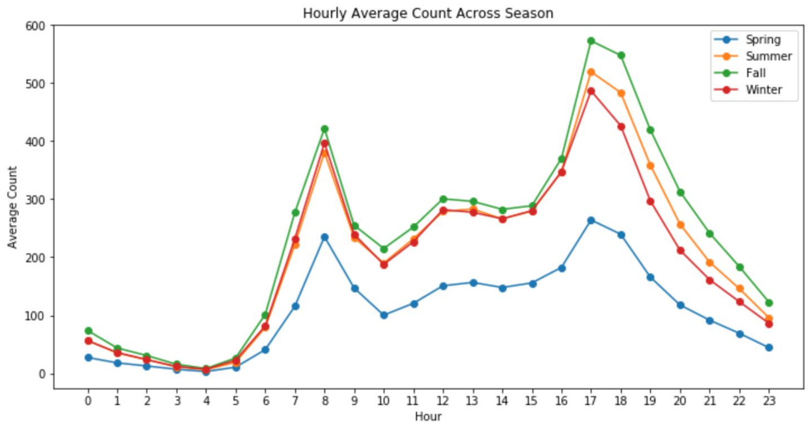


In this next box, it clearly explained the outlier from Monday to Friday, the demand is extremely high during rush hours 7a.m. to 9a.m. and 4p.m to 7p.m.

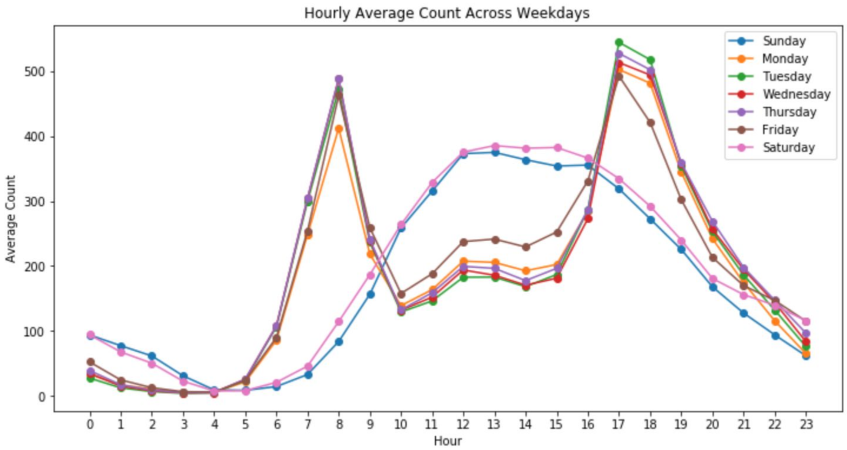


In the 3 plots above, it is clear that year 2012 has a lot more demand than that in the year 2011.

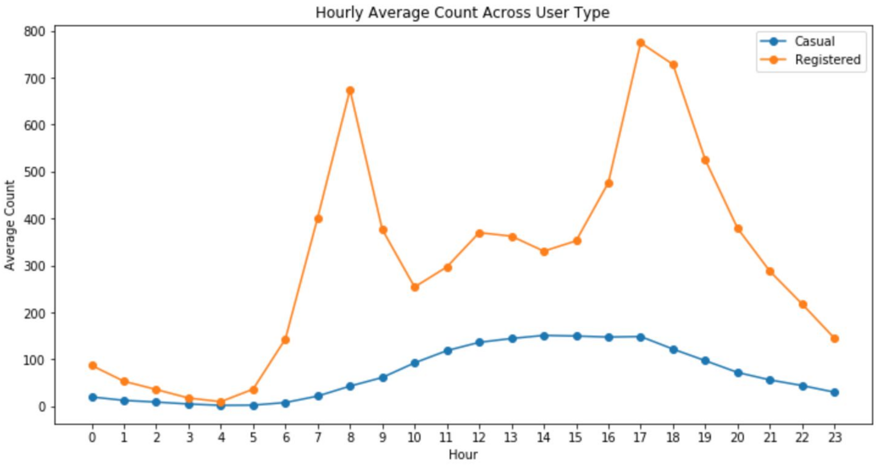
In the following hourly trend in different seasons, the demand is always relatively higher in rush hours.



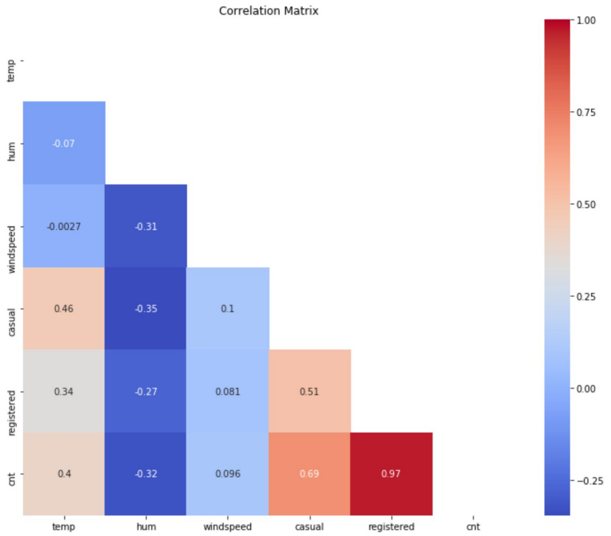
The weekday hourly trend across the year looks similar to in different seasons but Saturdays and Sundays, the weekend trends are more close to the normal distribution except the ending tails are on the beginning of the next day.



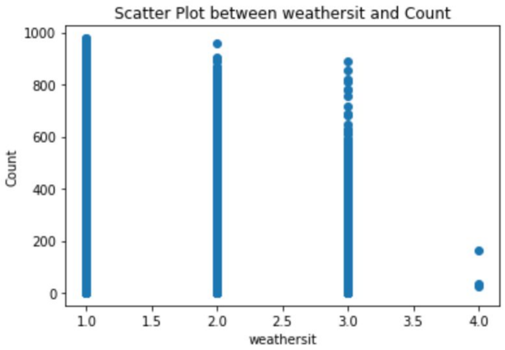
The registered user make the major contribution to the overall usage count.

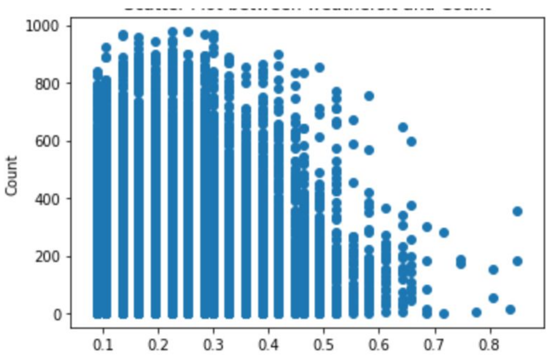


In the following Correlation matrix, it tells all correlations between variables. Interestingly, the temperature correlates with casual users than registered users, therefore it is a good idea to promote membership in warm days, i.e. Summer and Fall.



Humidity has a negative correlation to the count as rainy days have much higher humidity than sunny/windy days, The following scatter plot illustrates that rainy days have lower demand than sunny and windy days. The windspeed has lower influence to the cnt, but it is indeed a factor as the cnt is getting lower when the windspeed is higher





Overall, there is a 99% confidence that there will be 186 user count any hour of the day in a year.

And, in Spring, there is a 99% confidence that there will be 106 user count any hour of the day.

In Summer, there is a 99% confidence that there will be 201 user count any hour of the day.

In Fall, there is a 99% confidence that there will be 229 user count any hour of the day.

In Winter, there is a 99% confidence that there will be 192 user count any hour of the day.

The t value between registered and casual cnt is 4.69 and p value is 3.4e-6. The low p value means the casual users and registered users are not using a similar amount of times of bike. Therefore, from now on, we should focus on the prediction in overall using count, as the casual users does not affect the overall using count much, and they are the potential registered users in the future when promotion is up in the Summer and Fall.