Forecasting on Capitol Bike sharing demand

Data Mining project 2016

Team 3

Abhishiktha Vanga, Mownika Chalichama, Seshi Saraswathi Harianathan, Samyuktha Anand

July 13, 2016

**TABLE OF CONTENTS**

**CHAPTER** **PAGE**

Executive Summary 2

Business Case 2

Problem Statement 3

Background 4

Methodology 5

* Data Exploration 6
* Pre-processing 9
* Post-processing 10

Conclusion 20

Business Value 20

References 21

Appendix 22

**Executive Summary**

Our current ways of living are implicating our quality of life, this has led to an understanding of the need for alternatives to our current urban development approaches among which higher priority needs to be placed on sustainable urban transportation systems, because urban transport systems represent the largest and greatest environmental and social opportunity to improving community quality of life.

One important feature of successful state-level growth management programs is a requirement for planning consistency. Bicycle sharing systems play an important part in increasing sustainable transport options in cities. An understanding of their potential use, and impact, across many diverse types of cities and multiple user types, is becoming increasingly important. With the creation of new systems and increased public availability of individual level origin–destination data for some systems (such as London, Washington DC, Minneapolis and Boston), the opportunities and applications of studying spatial, temporal and journey data associated with bike-sharing will continue to expand

Bikes receive increasing attention in city transportation, mainly because they “provide the missing link between existing points of public transportation and desired destinations” (Midgley, 2009). Bike-sharing can be described as a short-term bicycle rental service for inner-city transportation providing bikes at unattended stations. In recent years, bike-sharing systems (BSS) have rapidly emerged in major cities all over the world. Bike-sharing providers have to ensure high bike availability in order to satisfy customers.

**Business Case**

Forecast the bike rental demand in the Capital Bike Share Program in Washington, D.C. combining the historical usage patterns with weather data

Data set is obtained from the following source:

<https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

**Problem Statement**

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is also able to easily rent a bike from a particular place and return back at another place. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles. 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. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of the important events in the city could be detected via monitoring these data.

The Challenging task being, because movements of customers are highly dynamic and redistributing bikes is expensive we need to forecast the demand beforehand.

**Background**

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bike share system with the corresponding weather and seasonal information

Here are some of the observations which we thought could influence the demand of bikes:

* **Daily Trend**: Registered users demand more bike on weekdays as compared to weekend or holiday.
* **Hourly trend**: There is a high demand during office timings. Early morning and late evening can have different trend (cyclist) and low demand during 10:00 pm to 4:00 am.
* **Rain**: The demand of bikes will be lower on a rainy day as compared to a sunny day. Similarly, higher humidity will cause to lower the demand and vice versa.
* **Temperature**: In a country like India, temperature has negative correlation with bike demand. But, as per the Washington’s temperature graph, there is a positive correlation.
* **Pollution**: If the pollution level in a city starts soaring, people may start using Bike (it may be influenced by government / company policies or increased awareness).
* **Time**: Total demand should have higher contribution of registered user as compared to casual because registered user base would increase over time.
* **Traffic**: It can be positively correlated with Bike demand. Higher traffic may force people to use bike as compared to other road transport medium like car, taxi etc.

**Methodology**

**Data mining and knowledge Discovery:**

When forecasting a Bike Sharing demand location factors can be ascertained. To make the data suitable for DM it has to be properly prepped and selected. Also aggregation and normalization of data assures a solid basis for DM algorithms.



**Data Exploration:**

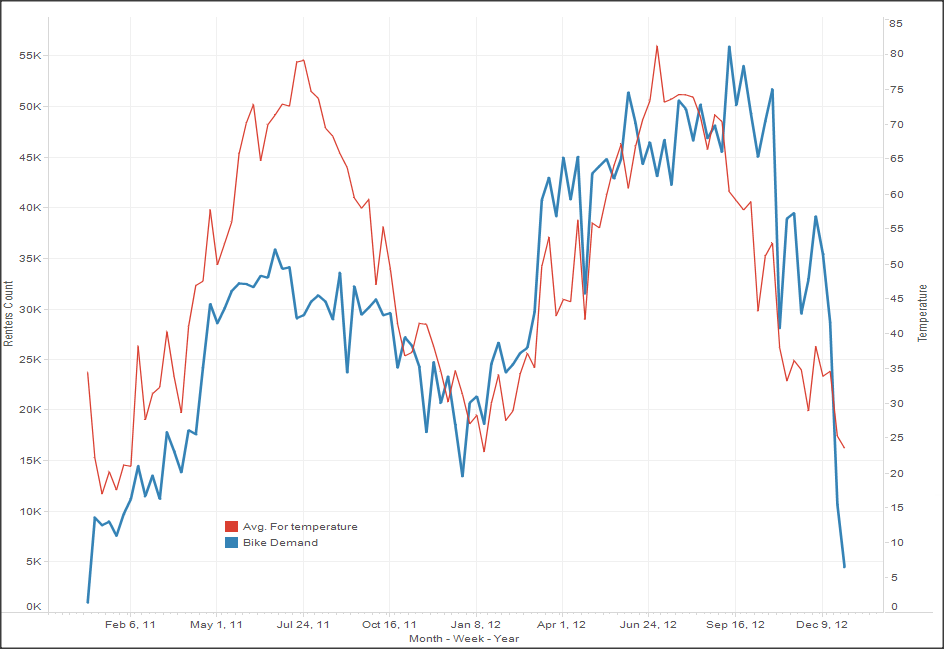
**Count of Bike share users over Time** – The graph below shows that there exists seasonality and trend in the number of bike share in the region over time.

Variation over every 3-4 months and variation in the number of users over the two years owing to many factors such as temperature, holiday, wind speed etc.

****

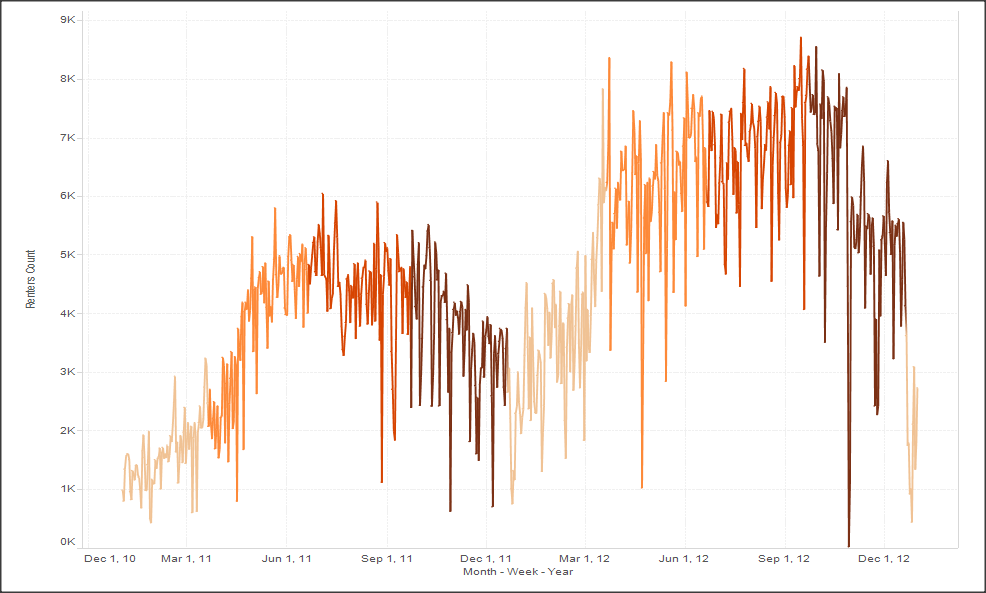
**Impact of temperature on bike demand:**

Temperature acts an influential factor in estimating the demand and shows a strong correlation as shown below.

****

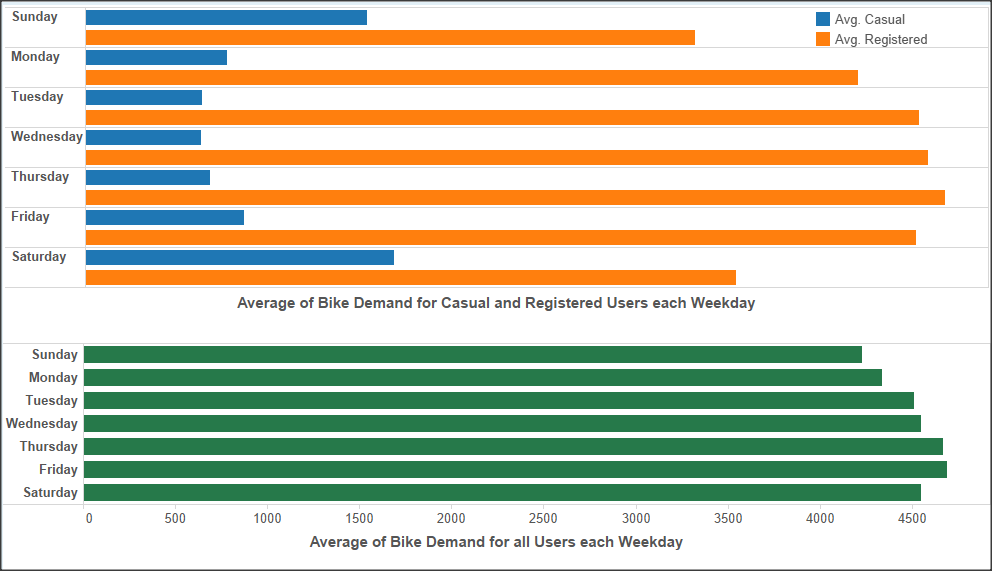
**Impact of Season on bike demand**

The below chart explains the 4 seasons with Dec-Mar being winter, Apr- Jun spring, July-Nov Fall.

****

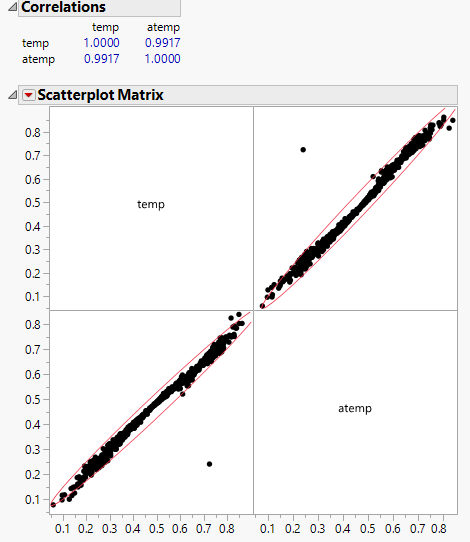
**Number of Bike share users across Week**

The below shows graph shows the number of users on a daily basis both as total count and separately as registered and casual users.

****

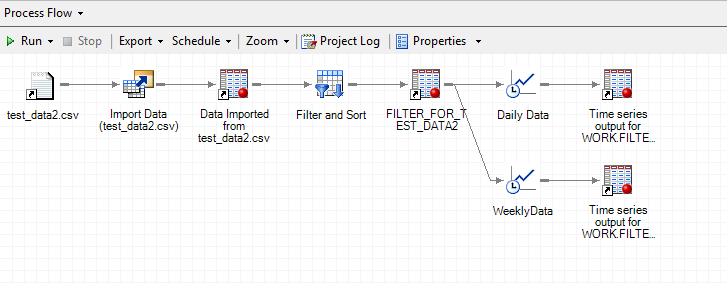
**Preprocessing**: Data pre-processing (cleaning) is done to make the data set suited for further processing

* Removal of insignificant columns (atemp as there was a high correlation between temp and atemp (feels like) and hence we considered only temp)



* Adjusted the seasons accordingly
* Modifying the format of the metadata to get daily data
* Getting the weekly data from daily data

The last two steps have been illustrated in the screenshot below:



**Series Analysis:**

Time series models adapts to the seasonality, trends in the past data and forecasts the future values. Generally, after modeling of the white noise tests are random, it can be said that the series has been pre-whitened. So, before building the AR and MA models, as the data contains seasonality and these seasonal variables changes over time, pre-whitening noise test is performed for these seasonal to determine the orders of each component. We have used the existing procedures in SAS to do this pre-whitening test and below is a piece of code depicting it.

/\* The below code is only for the seasonality part \*/

Ods graphics on/imagemap=on;

Proc X12 data=NEWPROJ.data date=date;

Var cnt;

x11;

outputout = out a1 d10 d11 d12 d13;

run;

ods graphics off;

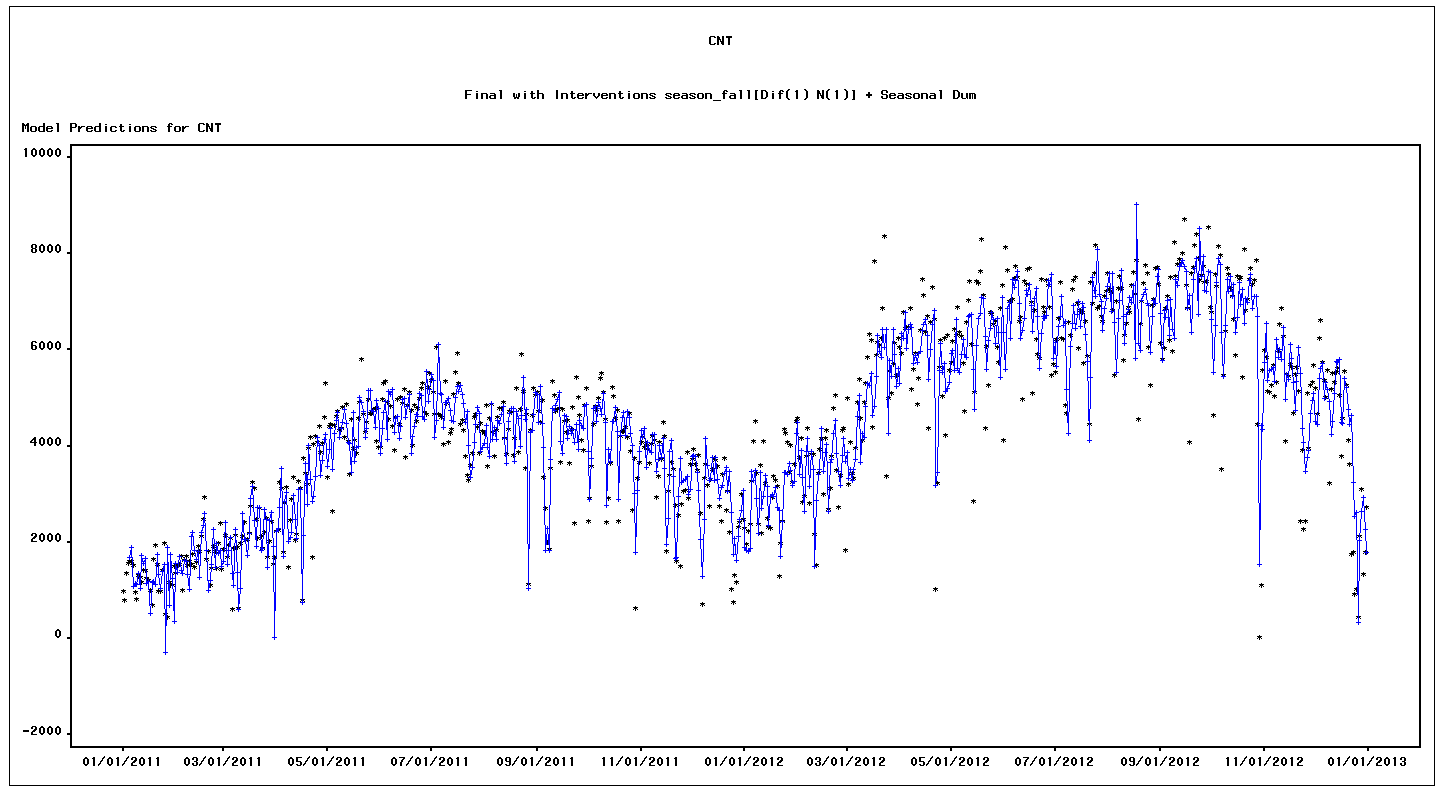
**Post processing**:

**Modelling**

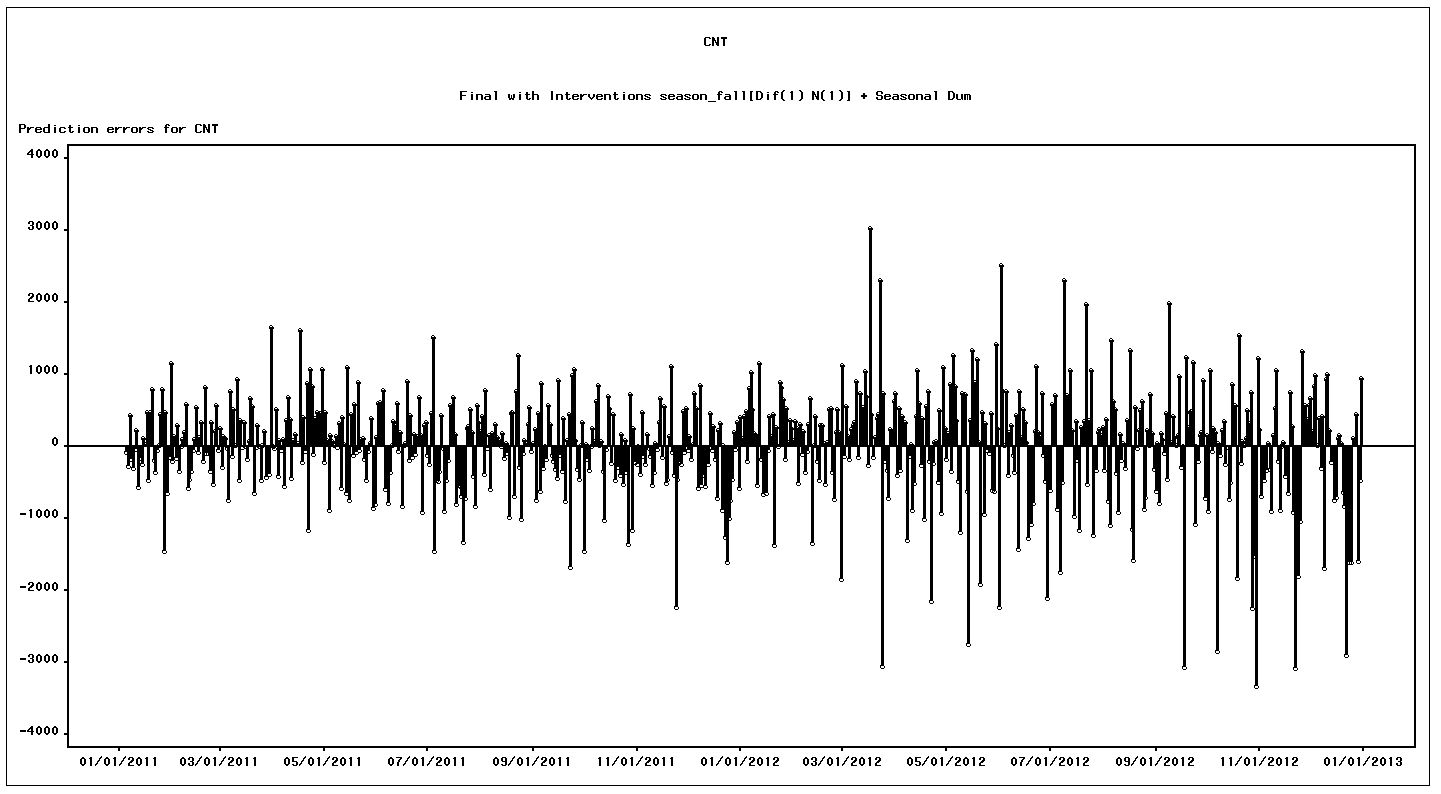
Modelling was done on the data for two time durations as shown below:

1. Daily basis
2. Weekly basis

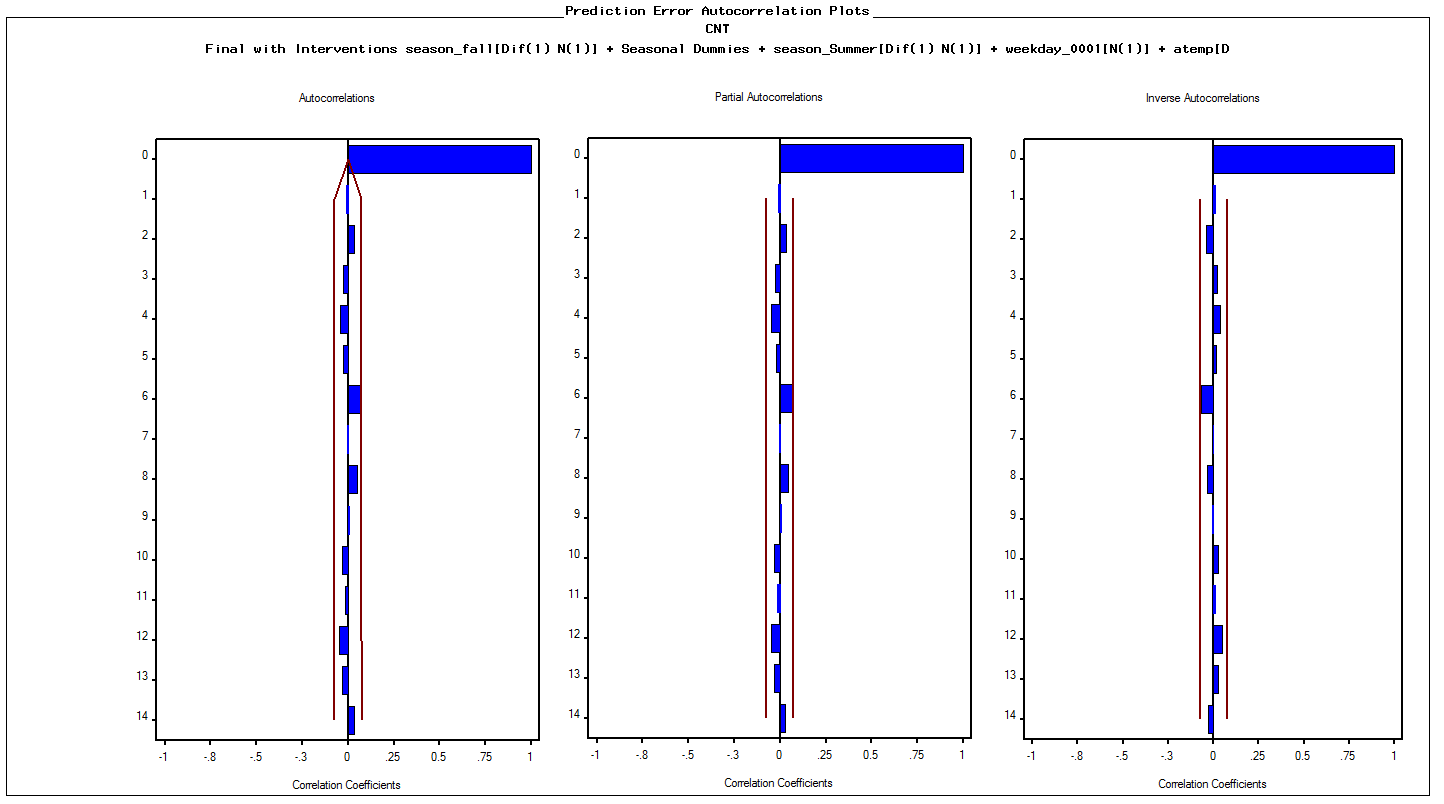
Daily forecast to determine the number of registered users and casual users. Also, the total count to determine the demand on a daily basis.



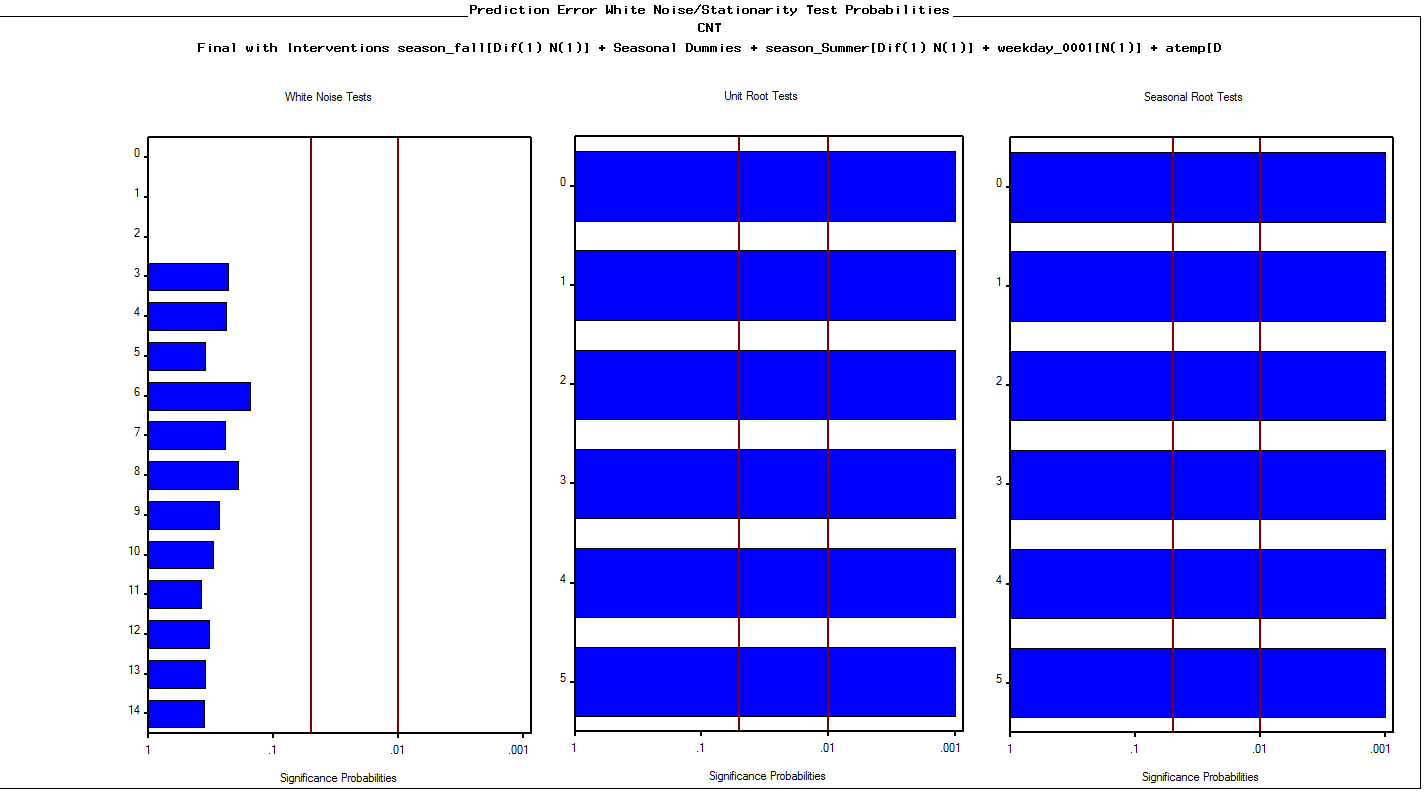
The model is performing better in predicting the demand of the bike share on daily basis.



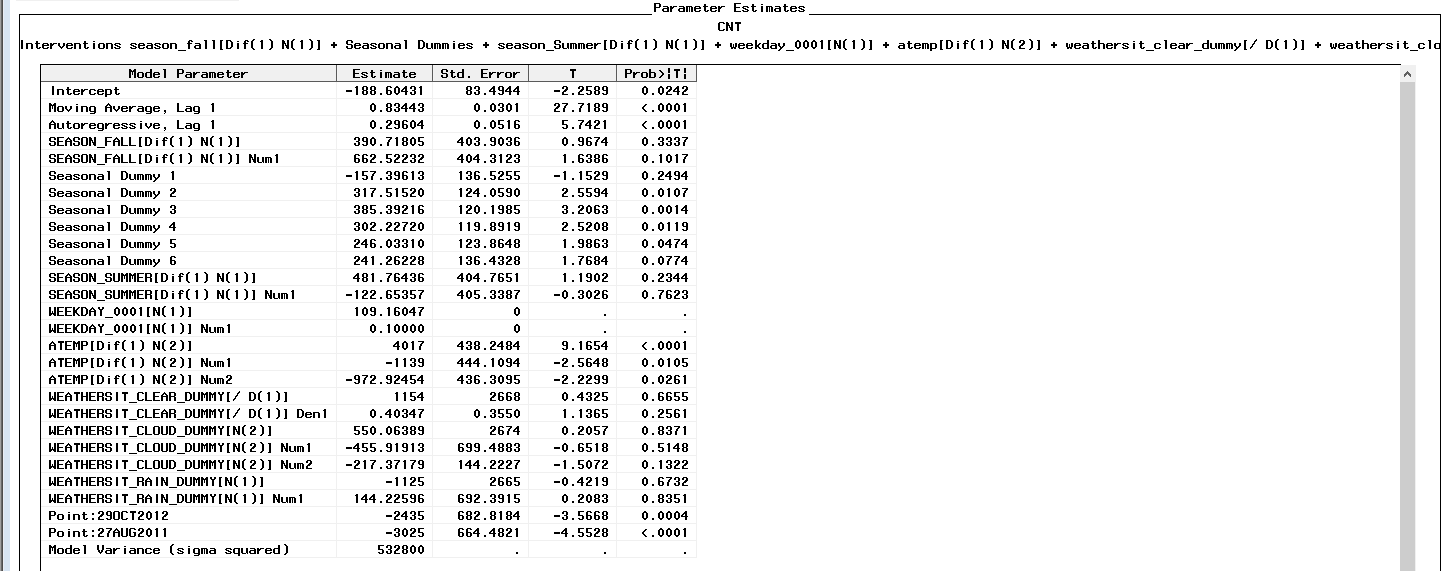
The residual errors have been distributed quite randomly except some negative bias in the recent periods.



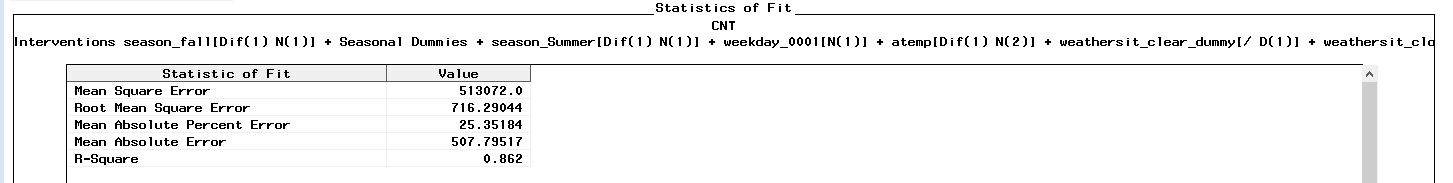
From the auto correlation plots we could see that there is no significant correlation between the errors.



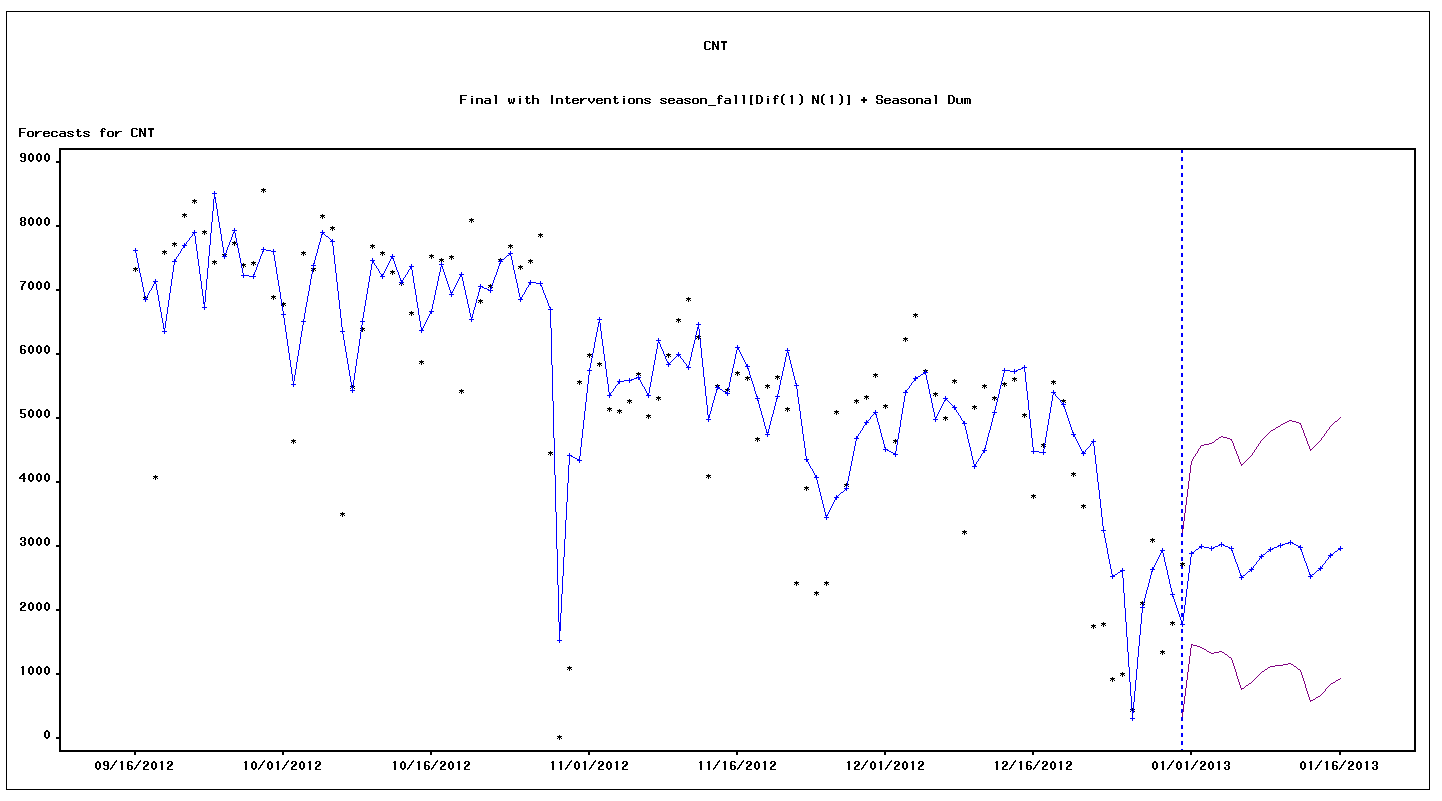
From the white noise test we fail to reject the null hypothesis that the errors are randomly distributed. So there is no significant trend in the errors. The seasonal root test and the unit root test indicates that the model is stationary.

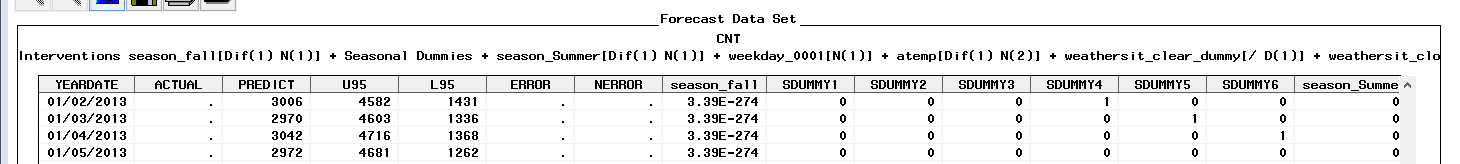


From the above parameter estimates we could see that the seasonal dummies are not significant whereas the temperature and the weather situations are significant in determining the demand.



The above statistical parameters show that the model is performing better in terms of RMSE but not on MAPE. The y forecast graph shows that the daily demand is not following the trend exactly. The trend has been nullified and only the seasonality is being captured.

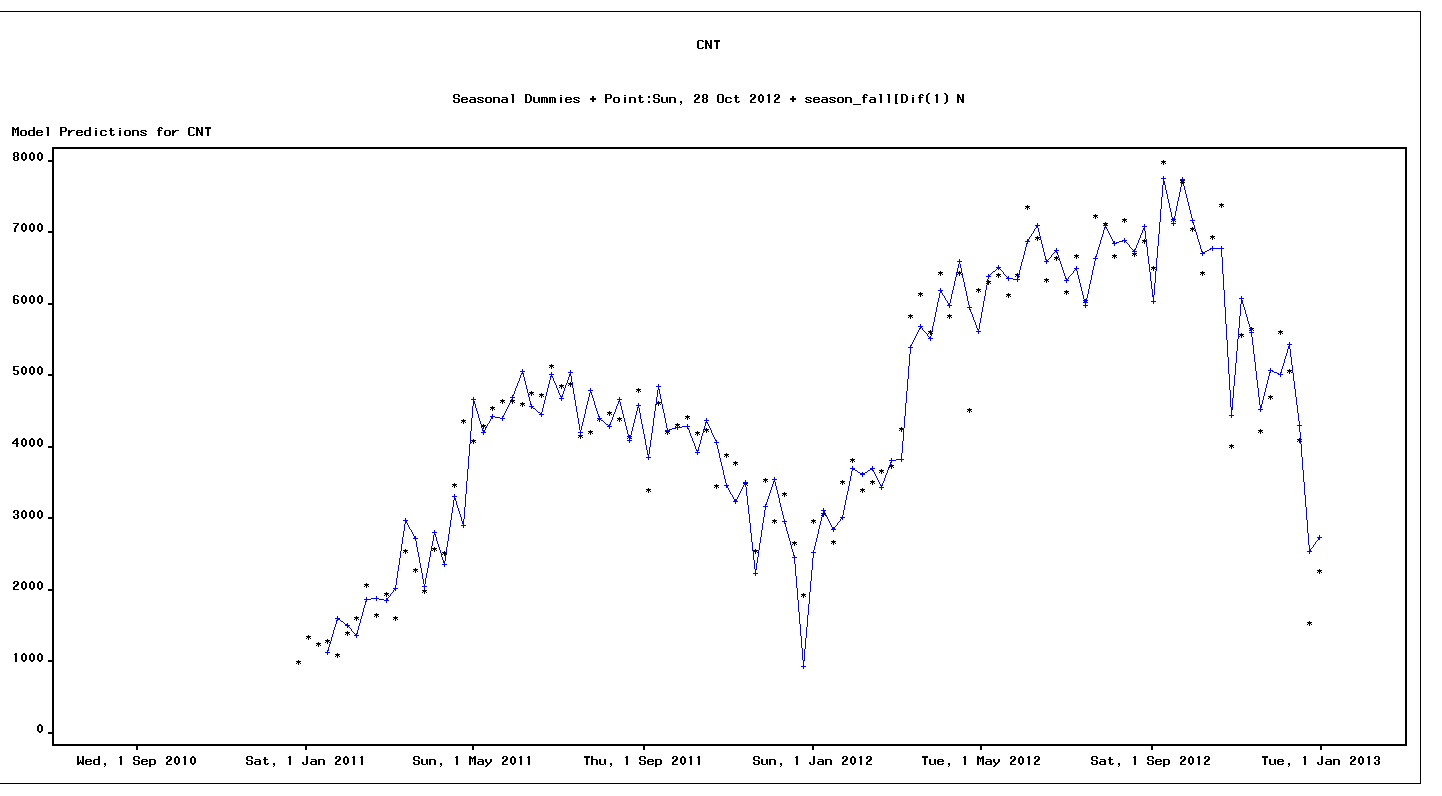
The forecasted values of the model are as follows.



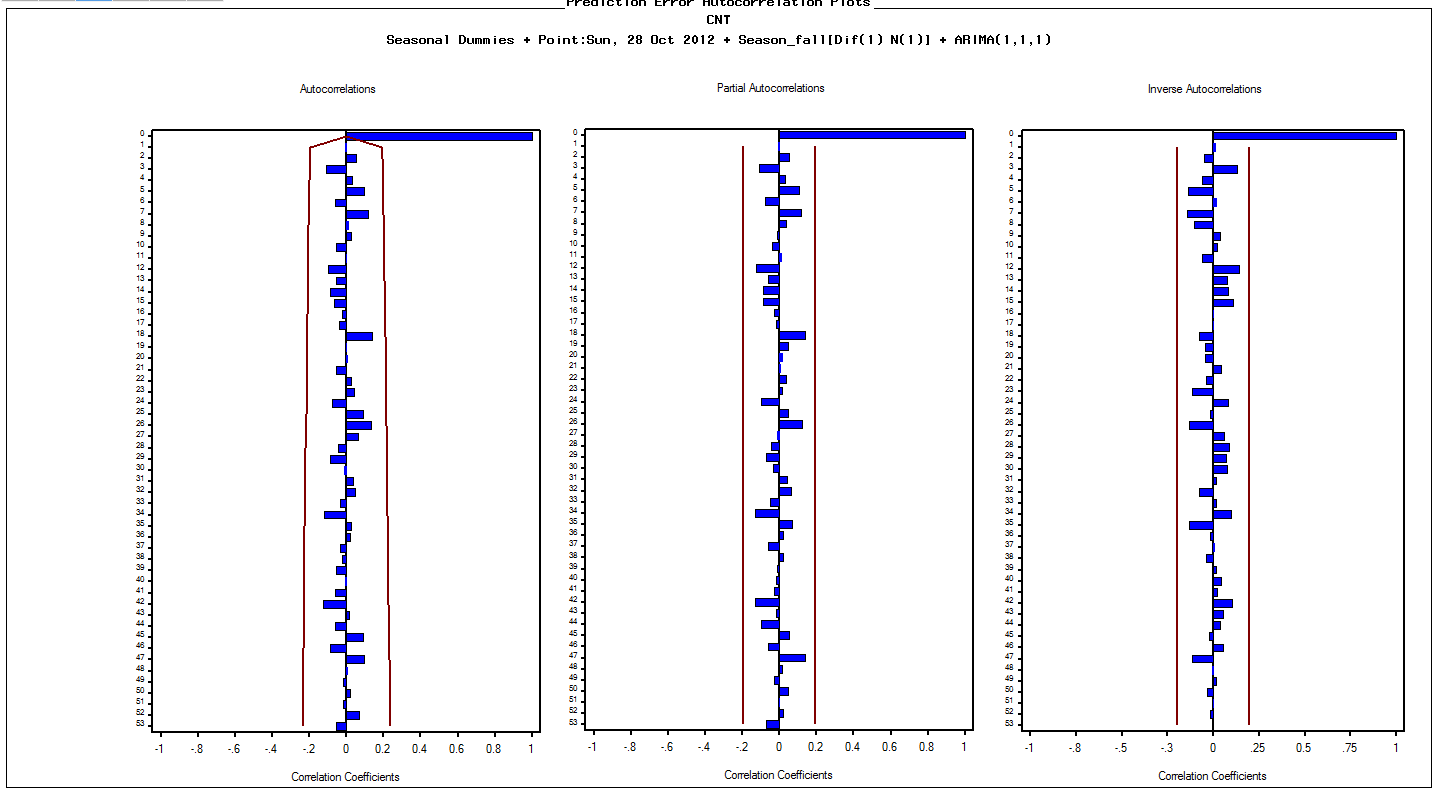
**Weekly Screen shot:**

As we could see that the model is not performing better on daily demand forecasting and even the total demand is remaining almost consistent all over the week. So we moved the weekly forecasting.

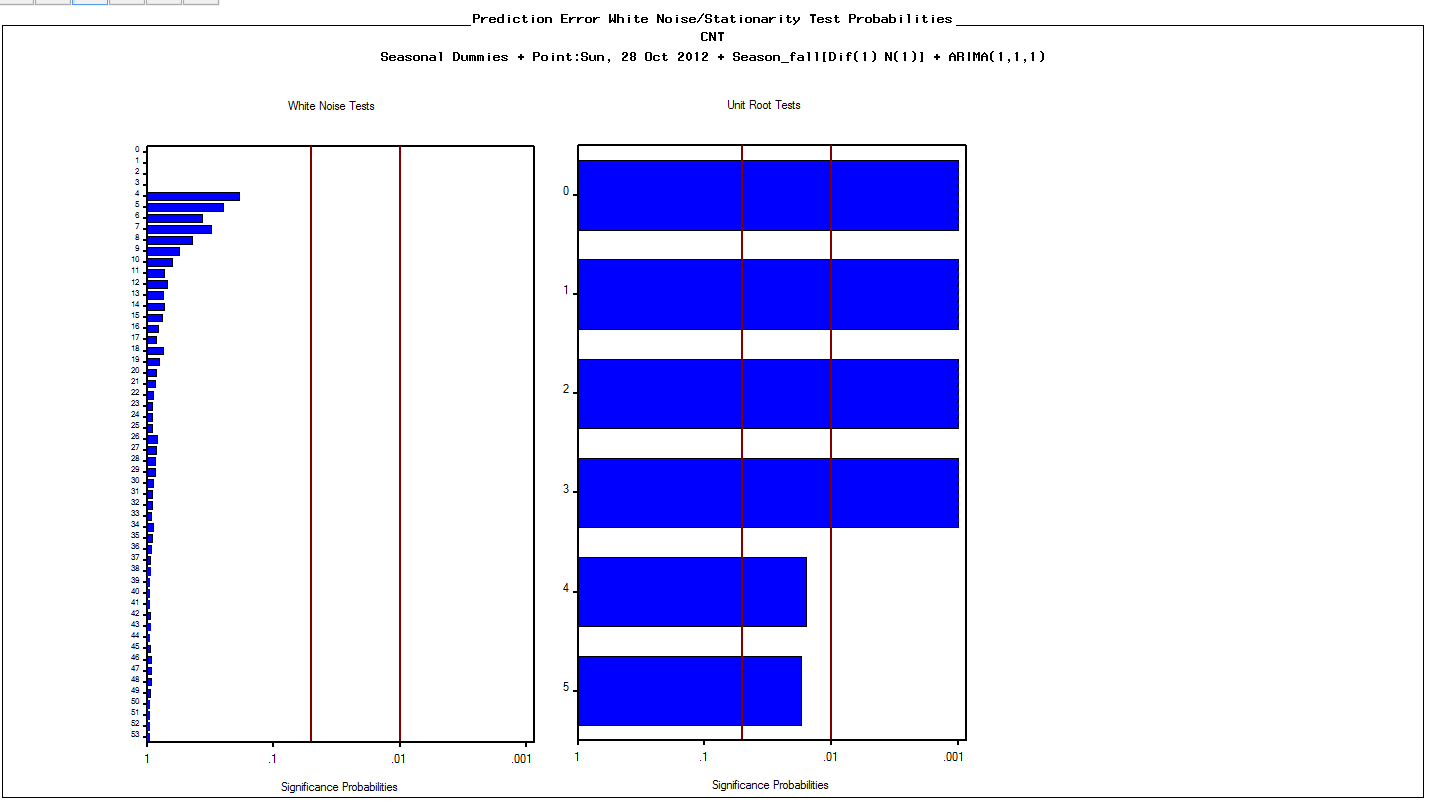
Daily forecast to determine the number of registered users and casual users. Also, the total count to determine the demand on a daily basis.



On the weekly data we could that the model is performing good in predicting the demand of the bikes.



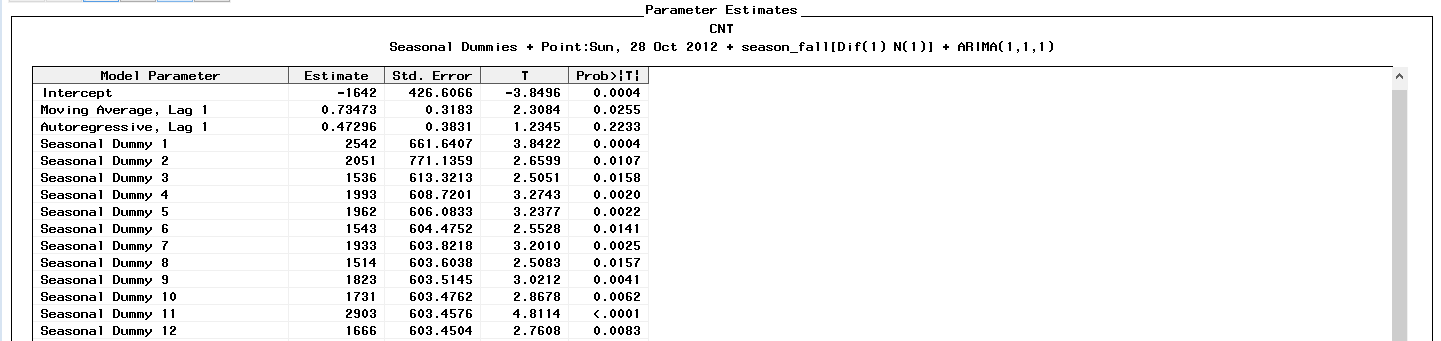
The auto correlation graphs show that there is no significant correlation between the errors in the model

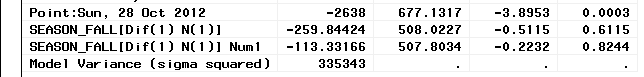


From the above white noise test we fail to reject the null hypothesis that errors are white noise and we could see that the model is stationary in terms of unit root test.

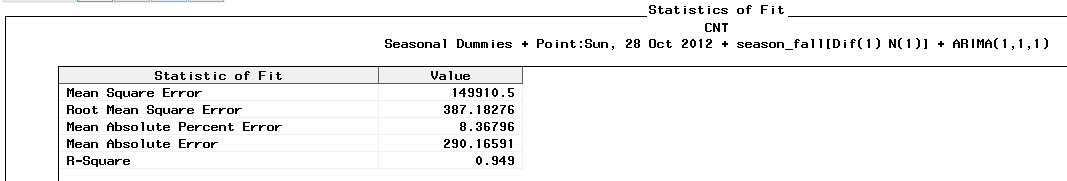
.

Below are the parameter estimates of the seasonal dummies and regressors

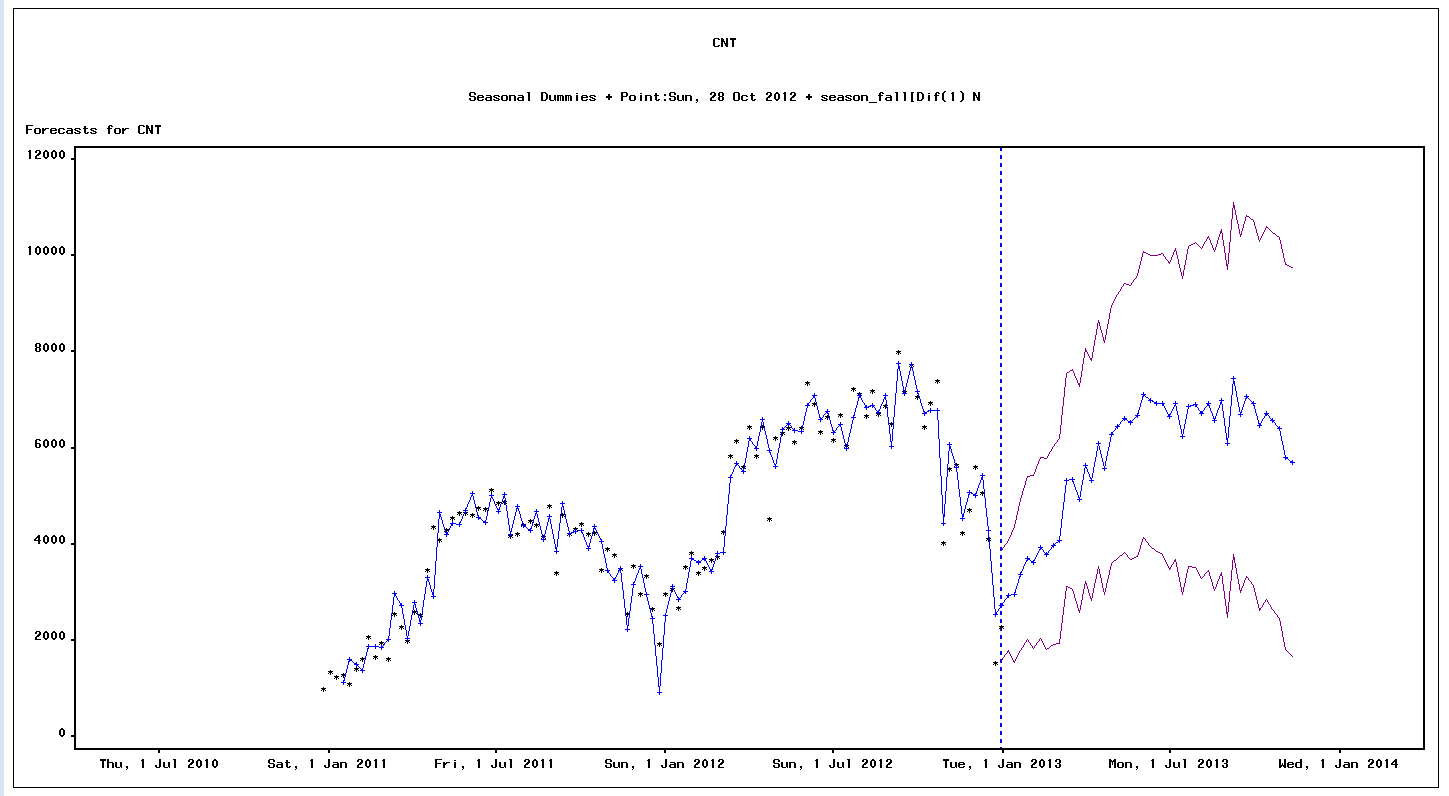




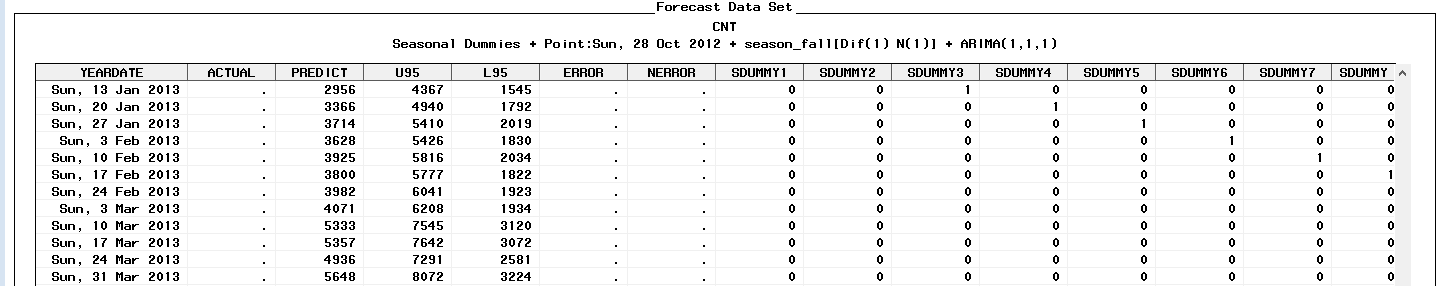
The statistical parameters indicate that the model is doing good in terms of R-square and the MAPE values.



Below are the forecasted output details:



From the above forecasted outputs, we can see that the model is carrying out the seasonality and trend far better and the confidence intervals are narrow indicates that the model is able to forecast accurately.



**Packages used**

1. SAS JMP
2. Tableau
3. SAS Miner
4. SAS Enterprise guide

**Scenario Analysis on Predicted Count Vs Temperature**

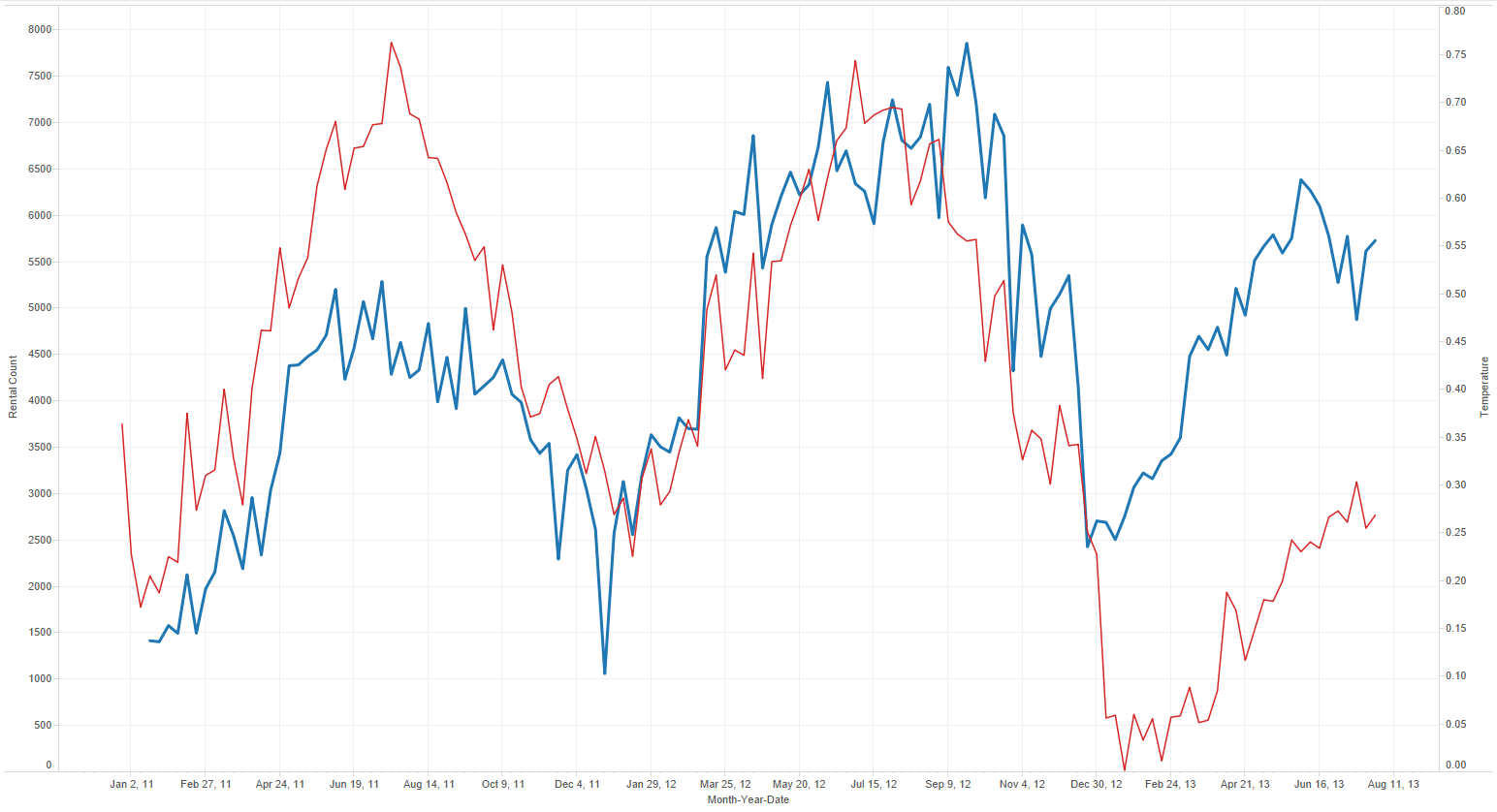
**Scenario:**

For a Temperature change of every **5**deg Celsius

**Impact:**

The Bike rental demand goes up by **200**

Note: Till threshold of 60 Celsius



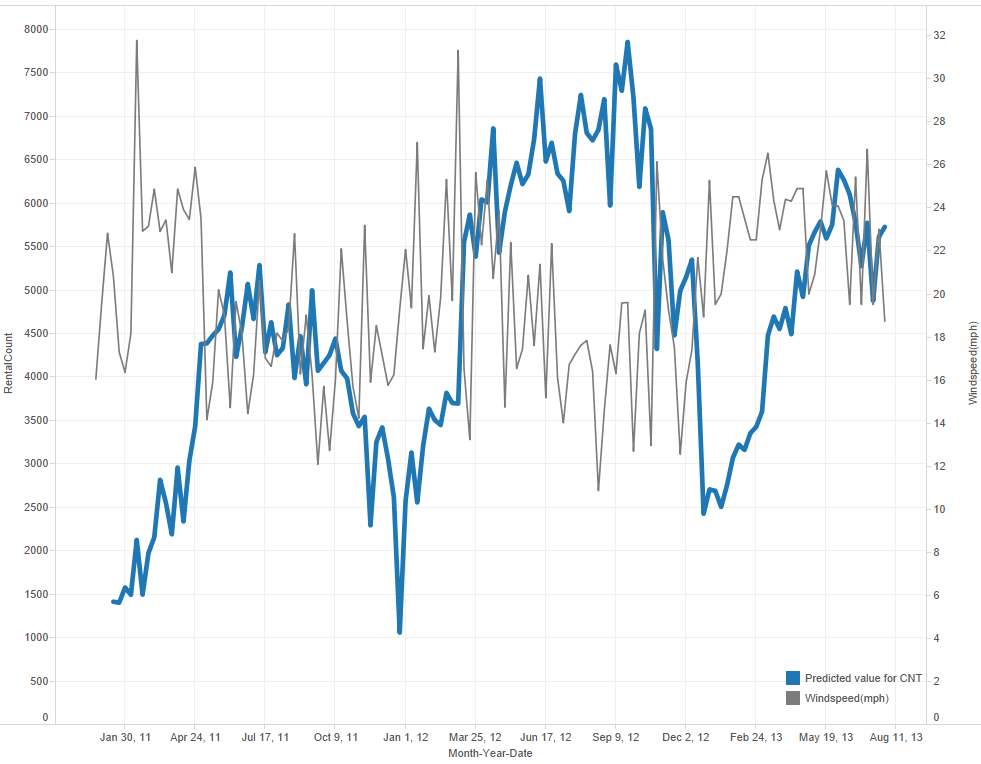
**Scenario Analysis on Predicted Count Vs Wind Speed**

**Scenario**:

Wind speed increased from 19mph to 26mph

**Impact:**

The Bike rental demand goes down by 300

****

**Conclusions**

Using different combinations of models and features, we found that weeks are by far the most predictive feature for our problem, whereas weather variables play a much smaller, but still noticeable effect in contributing to accurate predictions. In addition, ARIMAX models – in particular, ARIMA (1,1,1) with point interventions, seasonal dummies and regressors were able to best capture the relationships within the data.

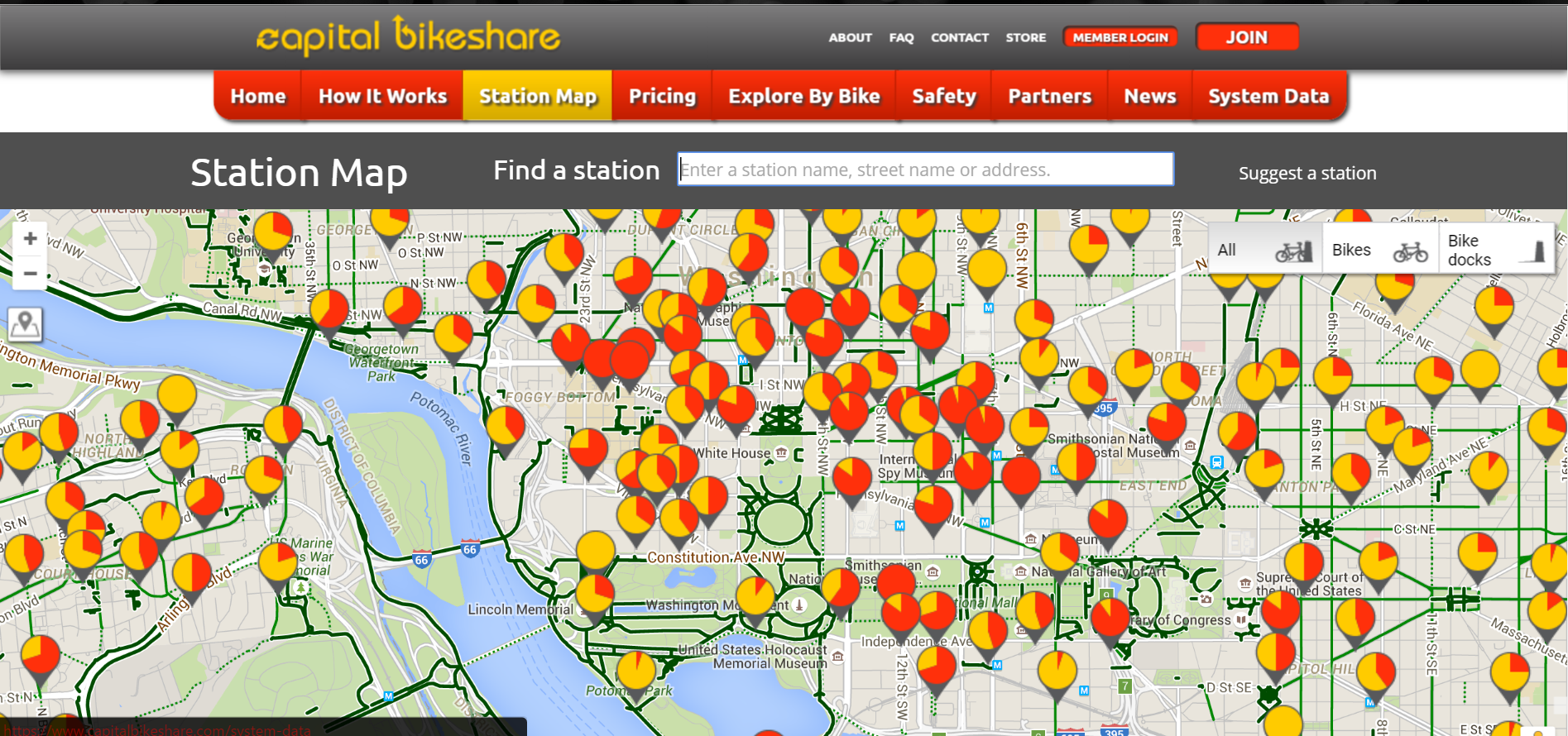
**Business value:**

Demand forecasting is an area of predictive analytics dedicated to understanding consumer demand for goods or services. That understanding is harnessed and used to forecast consumer demand. Knowledge of how demand will fluctuate enables the supplier to keep the right amount of stock on hand. If demand is underestimated, sales can be lost due to the lack of supply of bikes. If demand is overestimated, then they may incur excess expenditure for surplus supply which can also be a financial draining.

The implications of empty bicycle stations i.e. when a user goes to a bicycle station and finds no bikes can be solved by forecasting the number of bikes needed on a particular day.

Also, the implications of full bicycle stations i.e. when a user goes to bicycle station and finds difficulty in returning the bicycle due to lack of preparedness can be solved as with the forecasting the need for each day on seasonal and trend basis. This will result in a beneficial business decision in a booming industry as this by helping bicycle redistribution accordingly.

Last but not least, forecasting systems can we used in understanding the effect of catastrophes (such as the two hurricanes that hit the Washington area during the duration of two years under consideration) on the business sector, giving the key decision makers some slack mitigate the impact.



**References**

1. https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
2. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns Patrick Vogela,\*, Torsten Greisera , Dirk Christian Mattfelda
3. DeMaio, P. (2009). Bike-sharing: History. Impacts. Models of Provision. and Future. In: Journal of Public Transportation. Vol. 12. No. 4. 2009
4. Mining bicycle sharing data for generating insights into sustainable transport systems Oliver O’Brien ⇑ , James Cheshire, Michael Batty
5. Forecasting Enrollment in Higher Education using SAS® Forecast Studio® by Erik Bowe, Steven Merritt (Kennesaw State University)

**Appendix:**

**Data Dictionary:**

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

* 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 [Web Link])
* 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 : 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: 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 to 100 (max)
* windspeed: Normalized wind speed. The values are divided to 67 (max)
* casual: count of casual users
* registered: count of registered users
* cnt: count of total rental bikes including both casual and registered