The George Washington University

Bike Sharing Forecasting Report

Submitted by

Kewei Chen
Peijia Wu
Vi Pham
Yixuan Yang

Time Series Forcasting_DNSC 6219
Professor: Refik Soyer

Table of Contents

1. Introduction and Overview

1.1 Background Introduction

Climate change is the hottest debated issue recently due to human activities. To reduce the consequences of this problem in the future, bike sharing has been known as a new mode of transportation in high density areas such as Washington DC or New York. However, the demand will depend on the day in a week or special occasions, so forecasting techniques will be able to provide the estimated demand in bike sharing.

The data is about the Daily bike sharing customers and weather in the metro DC area (January 1, 2011 - December 31, 2018). It can be used to forecast demand to avoid oversupply and shortages. We can also use it to find out other things we are interested in. For example, we want to figure out the factors contributing to the total customers which include the non-registered and registered customers, and then make forecasts about future customers to know the demand of bikes in order to avoid oversupply or shortages.

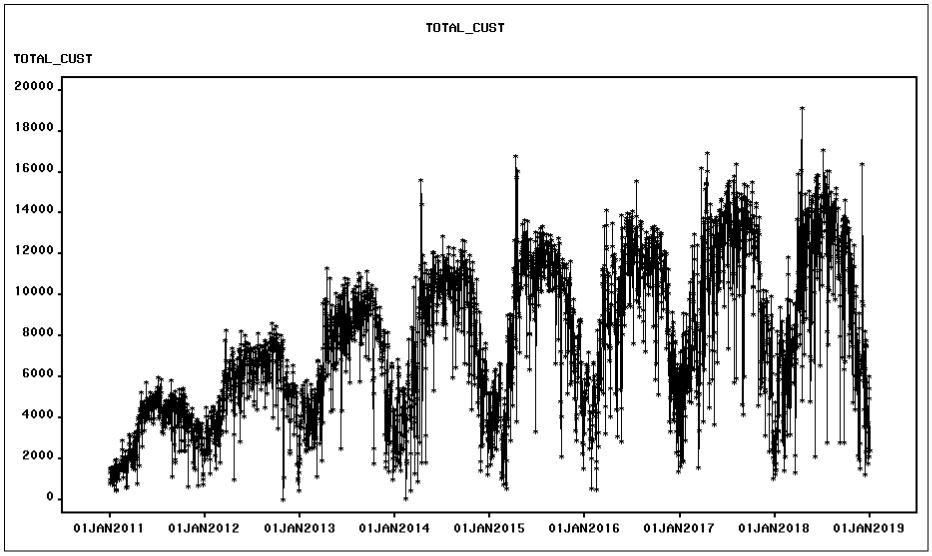
1.2 Dataset Introduction

Our data was obtained from the Kaggle. The link is https://www.kaggle.com/juliajemals/bike-sharing-washington-dc. There are 29 columns and 2922 rows in the dataset. Our goal is looking for which factors would obviously affect the total customer. In order to work on this goal, we keep the 13 variables that may relate to total customers and drop other 16 columns that are not useful for our project and lack data. There are also 4 null values in the current dataset. We interpolated our 4 missing values by using the previous numbers. In our cleaned data, there are 5 categorical variables, wt_fog, wt_thunder, wt_rain, wt_snow, holiday, and the rest are numeric variables. Holiday indicated the day is holiday or not rather than weekend. For 5 categorical variables, we recode it to 0 and 1 as binary variables. Finally we will use 365 observations(data of year 2018) as hold-out samples.

1.3 Overview of the series

1.3.1 Studies on original time series

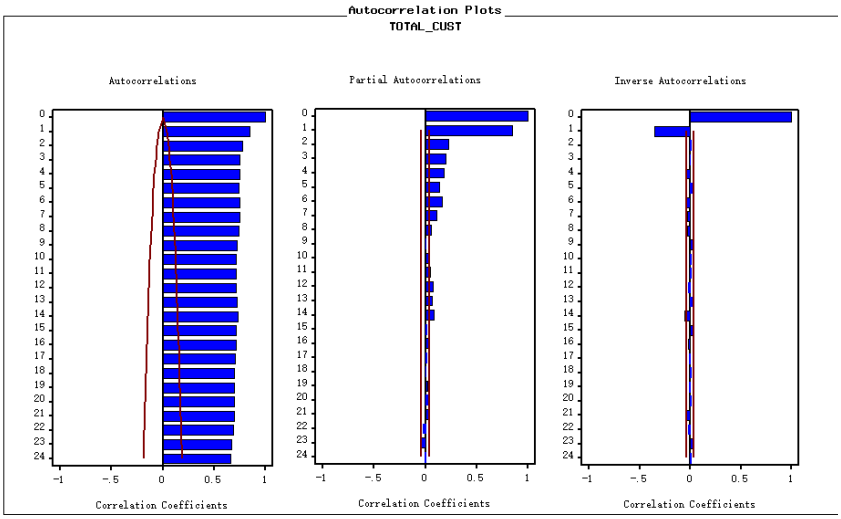
- Time series plot



1.3a

There is a seasonal behavior (shown in the picture 1.3a above) since every quarter 2 and 3 has an obvious higher level than other seasons and the data in January is lower than other months in each year. There is also an upward trend in this time series and we can see that the variance of the series is changing with time from the plot. Therefore, intuitively there is a small number of total customers of shared bikes in DC in January of each year and the number of total customers is increasing gradually.

- ACF Plot:

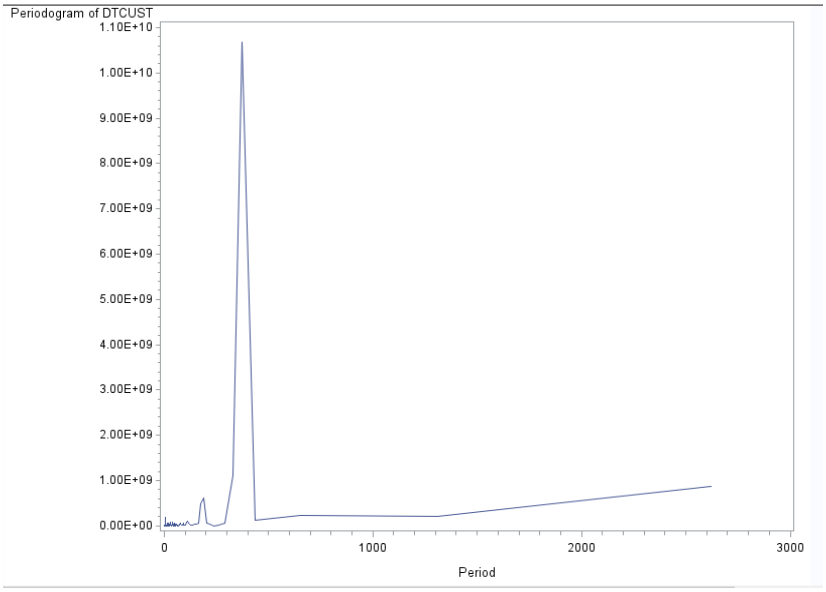


1.3b

From the above plot (shown in the picture 1.3b above), we can see that the autocorrelation is decaying slowly as the lag increases, so this time series is not stationary. We should take other considerations to the first difference of the series.

1.3.2 Studies on other plots

- Periodogram:

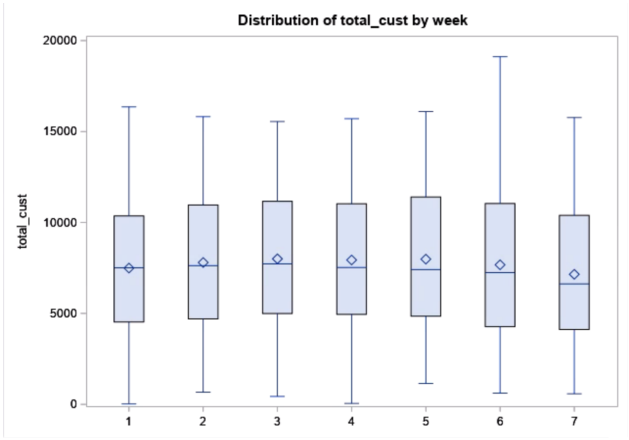


1.3c

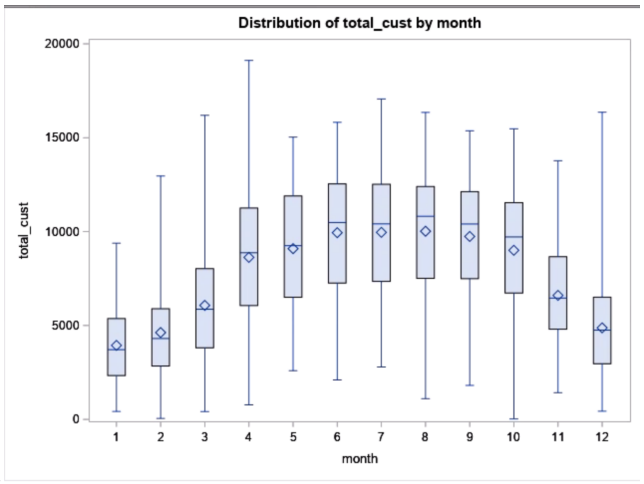
From this plot(shown in the picture 3.2c above), it may be hard to identify the period contributed mostly. But according to the data we can see the periods 2622, 374.57, 655.5, 327.75, 187.29, 174.8, 874,1311, 6.99, 524.4, 437, 114 respectively have larger contributions than the others. Therefore, in the latter chapter we did deeper analysis focusing on these periods.

- Boxplot

Weekly: Monthly:



1.3d



1.3e

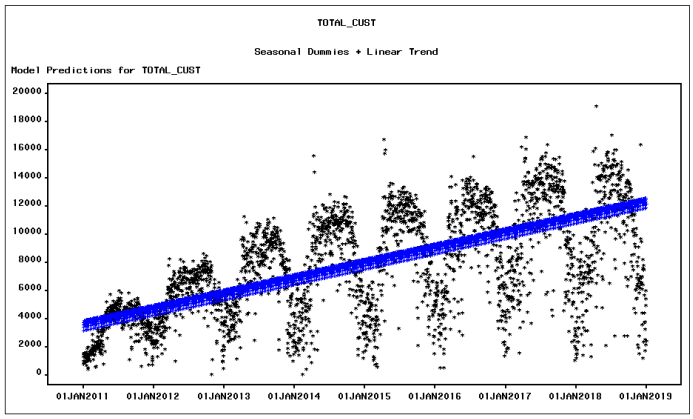
Based on the weekly box plot (shown in the picture 1.3d above), we can see there are similar mean and median on each day, which means weekday or weekend might not obviously affect customers using shared bikes.

Based on the monthly box plot (shown in the picture 1.3e above), we can see there is an obvious change for total customers from January to December. January, February and December show a very low mean value and other months except March and November show much higher values. We can conclude that there is a seasonality, and month would obviously affect customers using shared bikes.

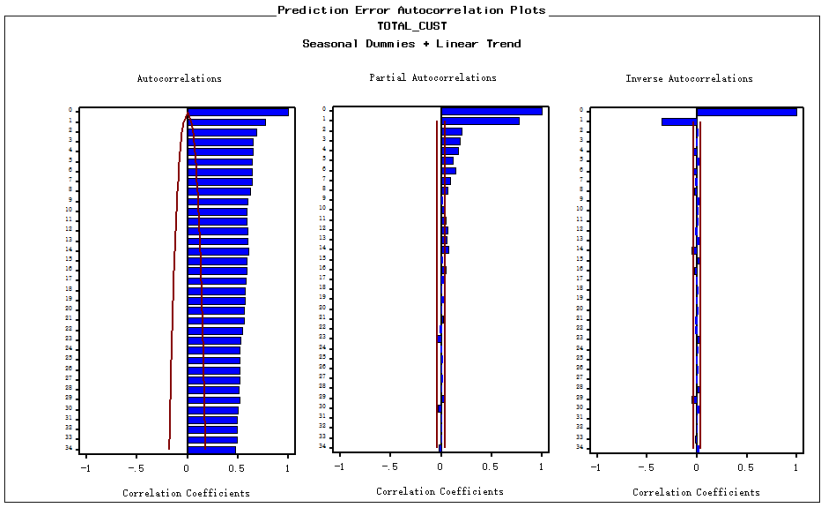
2. Univariate Time-series models

2.1 Deterministic Time Series Models

2.1.1 seasonal dummies+linear trend



2.1a



2.1b

Parameter Estimates

TOTAL_CUST

Seasonal Dummies + Linear Trend

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	3576	179.9813	19.8688	<.0001
Seasonal Dummy 1	-417.64114	212.8942	-1.9617	0.0506
Seasonal Dummy 2	-159.47294	213.0400	-0.7486	0.4546
Seasonal Dummy 3	119.79918	213.0399	0.5623	0.5742
Seasonal Dummy 4	284.61103	213.0399	1.3360	0.1824
Seasonal Dummy 5	207.89411	213.0400	0.9758	0.3298
Seasonal Dummy 6	282.45391	213.0400	1.3258	0.1857
Linear Trend	2.98267	0.0772	38.6564	<.0001
Model Variance (sigma squared)	8294278	.	.	.

2.1c

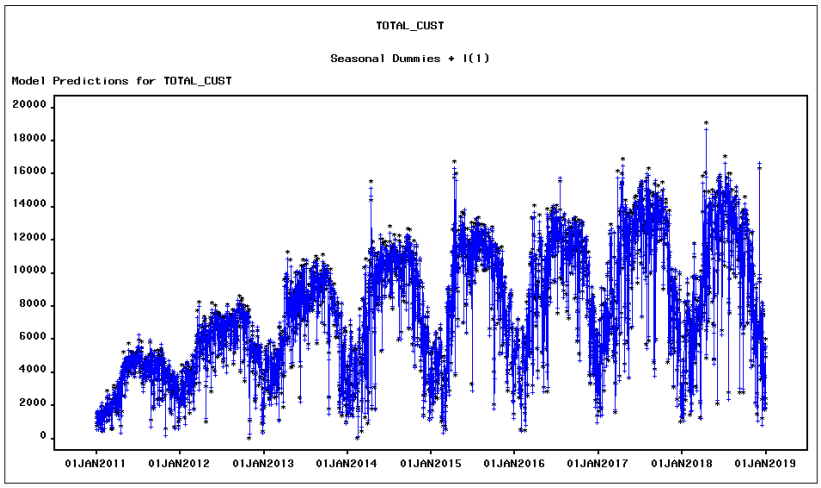
As with our previous studies on the original time series, we thought there may exist seasonality and trend components so we fit a seasonal dummies and linear trend model. As a result, we found the blue line can not fit most of the data points(shown in the picture 2.1a above) and the MAPE is 74.99 which is quite high. And the P-value of coefficients are large which is not statistically significant(shown in the picture 2.1c above) . Through the ACF of prediction error plot, we can see that the autocorrelation decays slowly which means it is

nonstationary(shown in the picture 2.1b above). So, we decided to add and explore error models by differencing the original series.

2.1.2 Log seasonal dummy + First difference + MA(2)

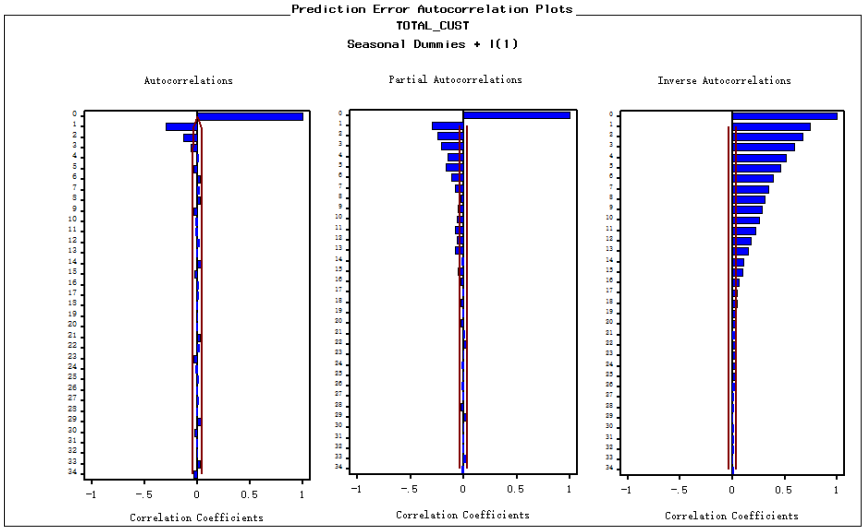
(2.1.2-1)

First, let's explore the Log seasonal dummy+First difference model. Obversely, this model has a lower MAPE than the first model according to the above picture(shown in the picture 2.1.2-1 above).



(2.1.2-2)

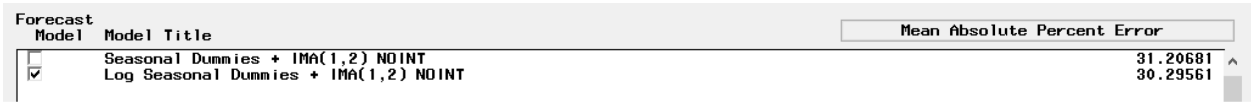
By looking at the model prediction plot, we can see that it may not perfectly match but it is also better fit than the first model(shown in the picture 2.1.2-2 above).



(2.1.2-3)

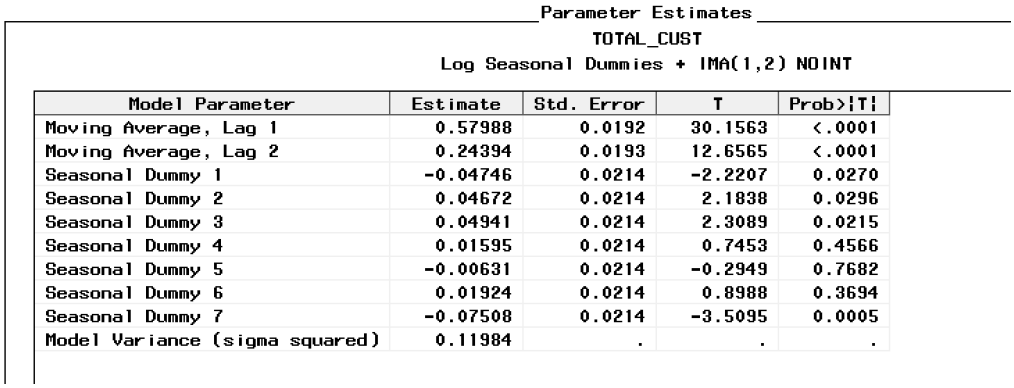
Through the ACF plot of residuals(shown in the picture 2.1.2-3 above), we can see that the autocorrelation is chopped off after lag 2, while the partial autocorrelations and inverse autocorrelation are decaying exponentially. Thus, we concluded that the error model is MA(2).

Now, let's use the MA(2) process to fit the error model and also use log transformation.



Forecast Model	Model Title	Mean Absolute Percent Error
☐	Seasonal Dummies + IMA(1,2) NOINT	31.20681
☑	Log Seasonal Dummies + IMA(1,2) NOINT	30.29561

(2.1.2-4)



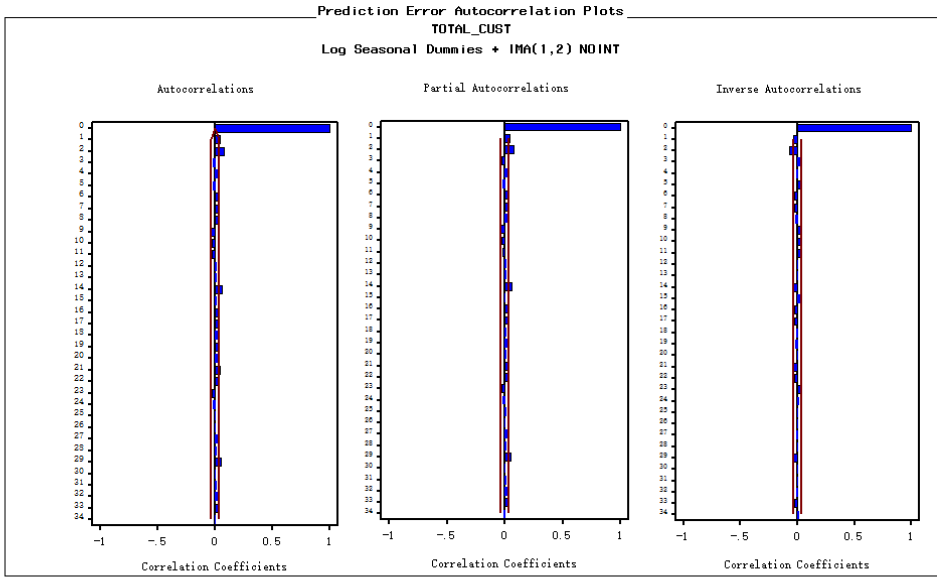
Parameter Estimates

TOTAL_CUST

Log Seasonal Dummies + IMA(1,2) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.57988	0.0192	30.1563	<.0001
Moving Average, Lag 2	0.24394	0.0193	12.6565	<.0001
Seasonal Dummy 1	-0.04746	0.0214	-2.2207	0.0270
Seasonal Dummy 2	0.04672	0.0214	2.1838	0.0296
Seasonal Dummy 3	0.04941	0.0214	2.3089	0.0215
Seasonal Dummy 4	0.01595	0.0214	0.7453	0.4566
Seasonal Dummy 5	-0.00631	0.0214	-0.2949	0.7682
Seasonal Dummy 6	0.01924	0.0214	0.8988	0.3694
Seasonal Dummy 7	-0.07508	0.0214	-3.5095	0.0005
Model Variance (sigma squared)	0.11984	.	.	.

(2.1.2-5)



(2.1.2-6)

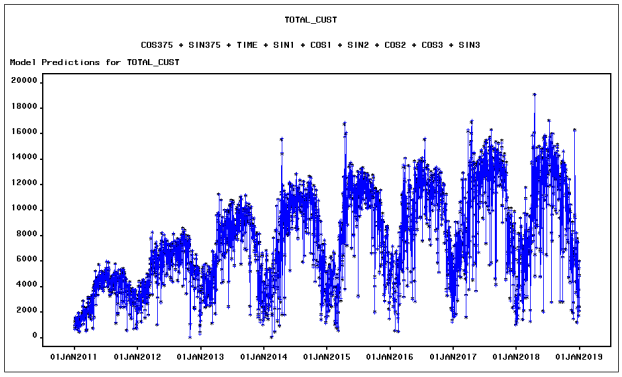
According to the above pictures(shown in the picture 2.1.2-4~6 above), we can see that the MAPE decreases after we use MA(2) to fit the error model. And most dummy variables are statistically significant except for the dummy variable 4, 5 and 6.

2.1.3 Cyclical trend model

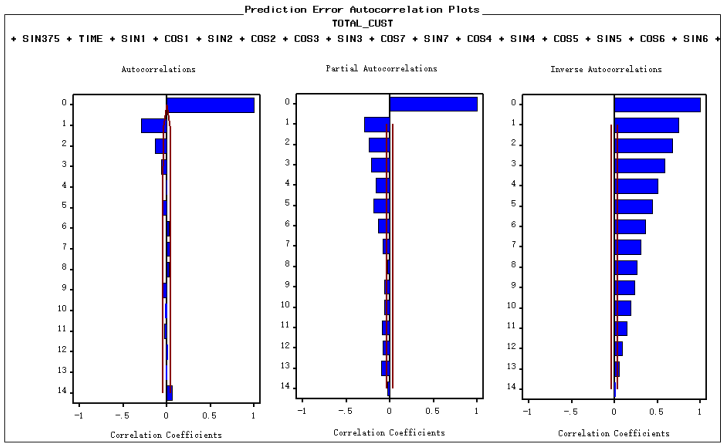
We also tried to fit the cyclical trend model. As we can see we chose 12 harmonics. Their periods are 2622, 374.57, 655.5, 327.75, 187.29, 174.8, 874,1311, 6.99, 524.4, 437, 114 respectively which have larger contributions than the others(shown in the picture 2.1.3a below). We put them as regressors and also added the first difference to decrease the model error. However, even though most variables are significant due to small P-value, after we fit the model, we noticed that the MAPE is 33.35 which is very high and the errors are not white noise(shown in the picture 2.1.3b~d below).

Obs	FREQ	PERIOD	P_01	i
8	0.01677	374.57	10671409977	7
9	0.01917	327.75	1112456073	8
2	0.0024	2622	864563388.1	1
15	0.03355	187.29	620352481.4	14
16	0.03594	174.8	479175432.1	15
5	0.00959	655.5	240672435.1	4
4	0.00719	874	223681600.9	3
3	0.00479	1311	201618353.4	2
376	0.89862	6.99	182483062.4	375
6	0.01198	524.4	166117946.5	5
7	0.01438	437	118203080.1	6
24	0.05512	114	108478002.4	23

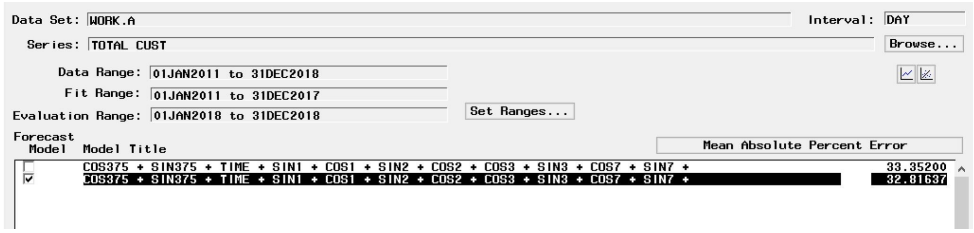
2.1.3a



2.1.3b

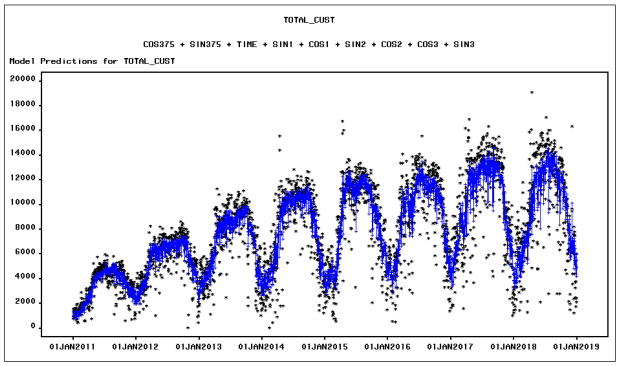


2.1.3c

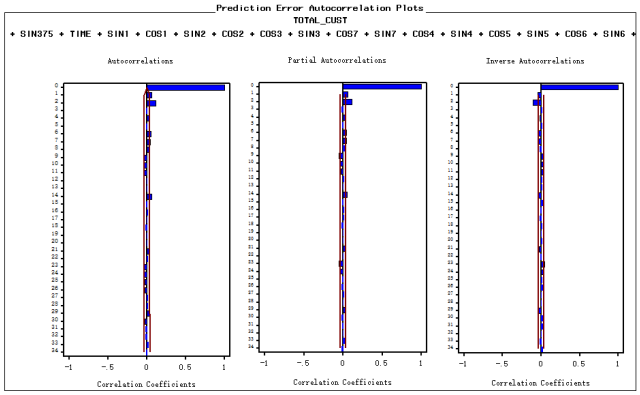


2.1.3d

Based on the plot 2.1.3c, we decided that we added the error model MA(2), and then we got the time series plot(shown in the 2.1.3e below). And its error is white noise finally (shown in the 2.1.3f below) and it is improved which has less MAPE 32.82.



2.1.3e



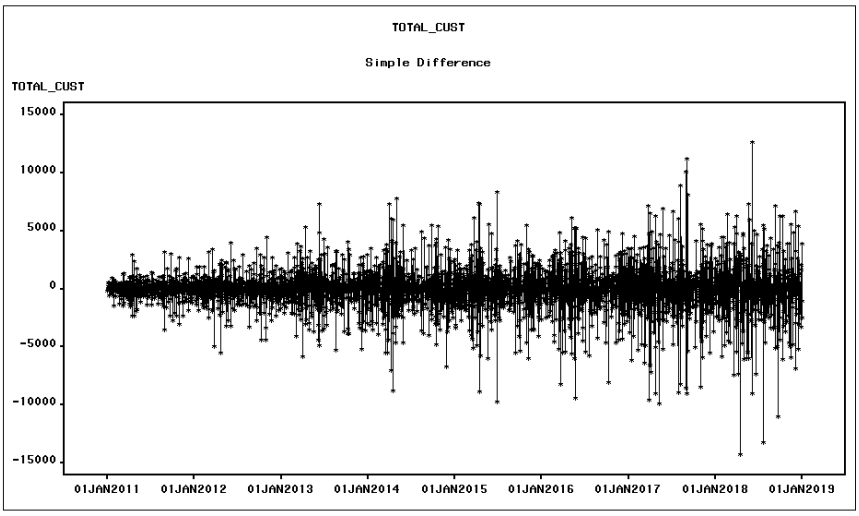
2.1.3f

2.2 ARIMA models

2.2.1 Studies on the differenced series

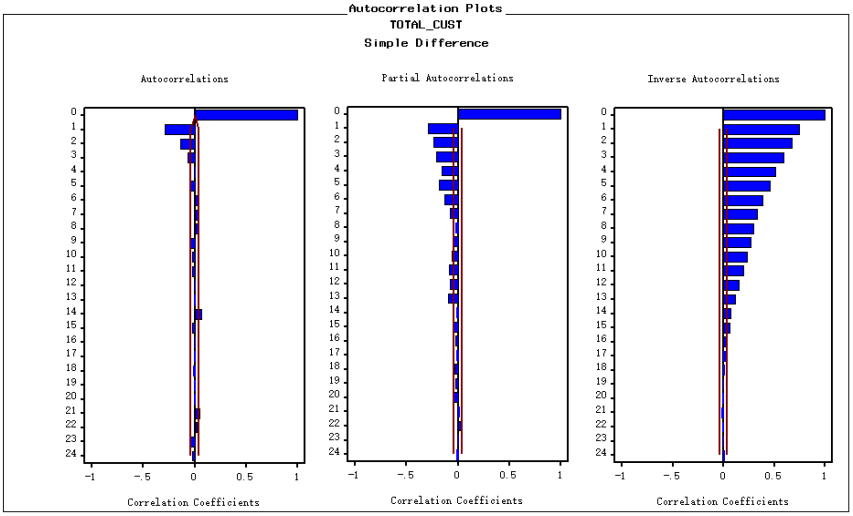
According to 1.1, note that the slowly decaying ACF and the level change in the time series plot imply that it is not stationary. Thus, we first convert it to a stationary series by differencing.

- Time Series Plot:



After we did the differencing on the original time series. The difference of the time series does not show any trend or seasonality. Also, it shows constant mean and variance.

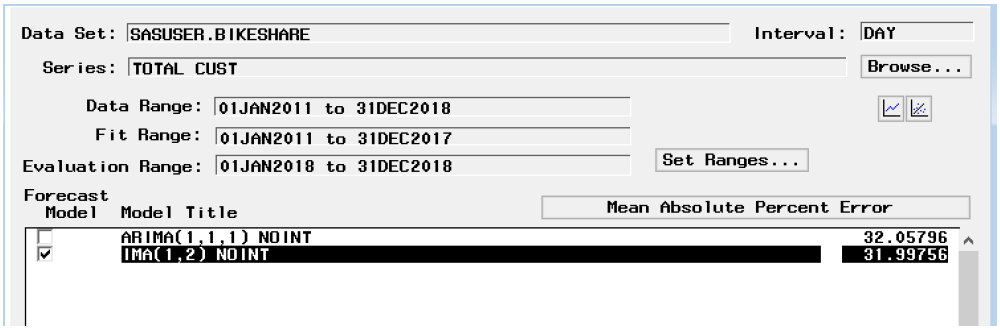
- ACF Plot of first difference:



The autocorrelation is decaying quickly as the lag increases, so this differenced time series can be considered as stationary. Since the ACF is chopped off after lag 2 and PACF is decaying exponentially, we consider it may be the MA(2) process or ARIMA(1,1) process.

2.2.2 ARIMA(1,1,1), ARIMA(0,1,2)

Thus, we decided to try an ARIMA(1,1,1) and ARIMA(0,1,2) process to fit the model first.

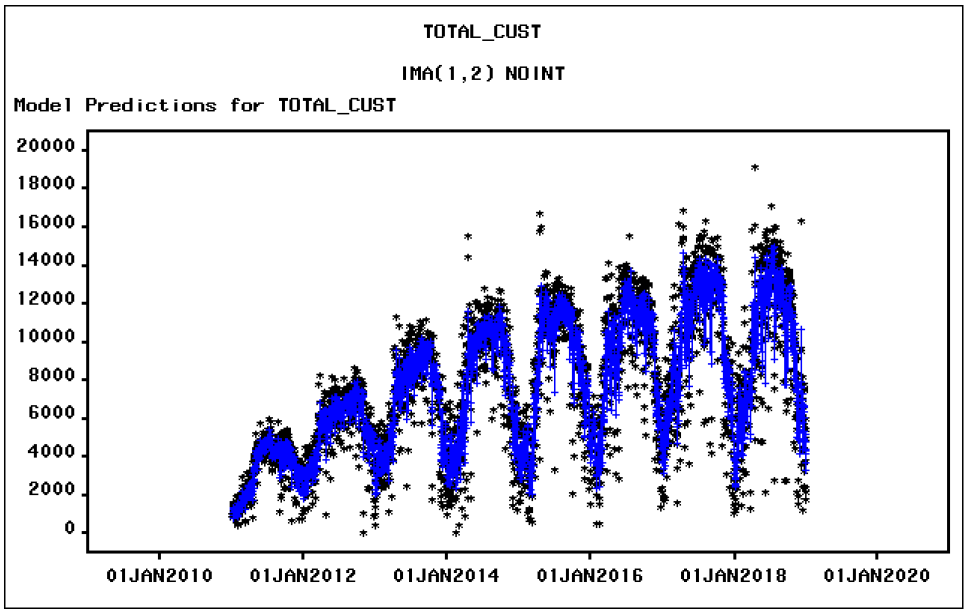


(2.2-1)

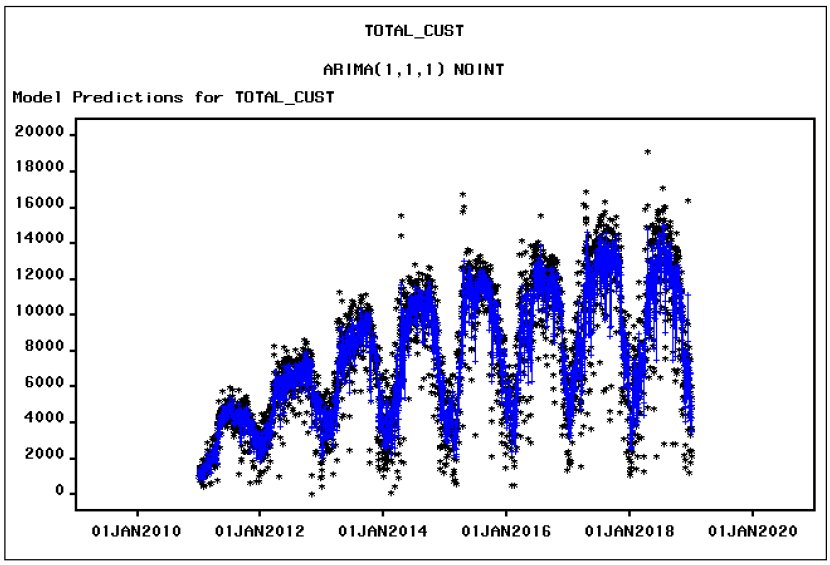
From plot (2.2-1), we can see that the MAPE of IMA(1,2) and ARIMA(1,1,1) are almost the same with no big difference.

So, we will compare the model fit and variance of the two processes.

Model Fit:

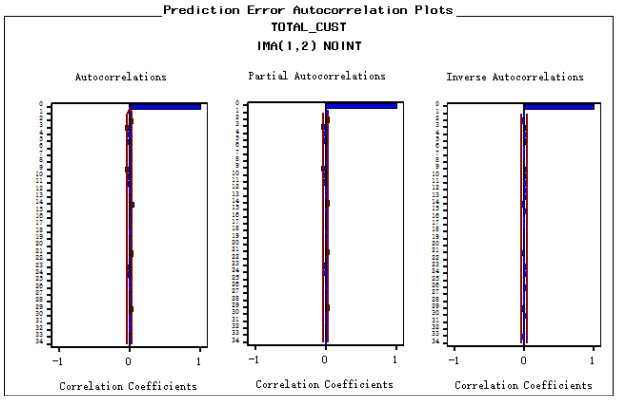


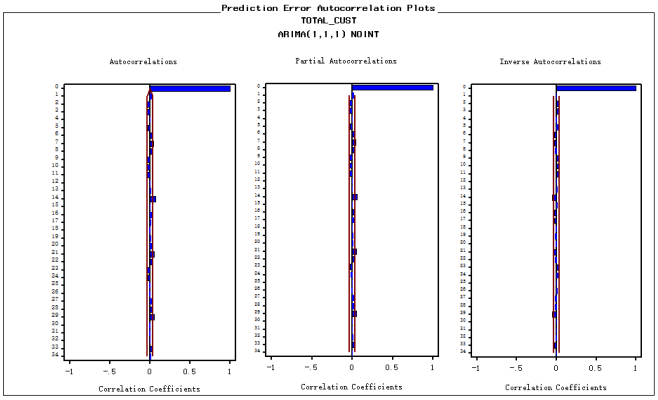
2.2-2



2.2-3

Error-ACF & PACF:





Based on the model prediction and ACF plot, we cannot distinguish between these two processes. Both model predictions exist outliers and the ACF shows that the prediction error is White Noise.

Parameter Estimates Comparison:

ARIMA(1,1,1)

Parameter Estimates
TOTAL_CUST
ARIMA(1,1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.88541	0.0120	73.7741	<.0001
Autoregressive, Lag 1	0.34822	0.0240	14.5240	<.0001
Model Variance (sigma squared)	2823398	.	.	.

ARIMA(0,1,2)

Parameter Estimates
TOTAL_CUST
IMA(1,2) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.55146	0.0191	28.8222	<.0001
Moving Average, Lag 2	0.25592	0.0192	13.3543	<.0001
Model Variance (sigma squared)	2838646	.	.	.

From the above results, we can see that ARIMA(1,1,1) has smaller model variance, while two processes have the same Mean Absolute Percent Error.

2.2.3 Add Regressor -- Holiday on the previous models

a) From the plot 2.2-2 and 2.2-3, we can see that both processes exist outliers. We assume that it may be caused by the holiday effect. So we decided to explore it by adding Holiday as a regressor in our model.

From the above results, we can see that the MAPE is slightly decreasing in both models after adding the holiday regressor. Also, both models have similar predictions plot after adding holiday.

Parameter Estimates:

Parameter Estimates
TOTAL_CUST
holiday + IMA(1,2) NOINT

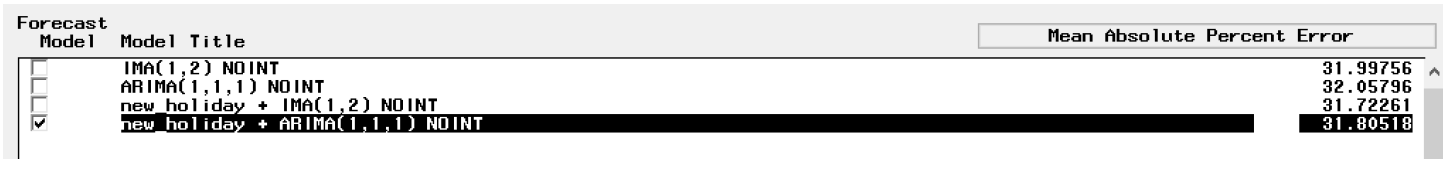
Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.55223	0.0191	28.8550	<.0001
Moving Average, Lag 2	0.25600	0.0192	13.3554	<.0001
holiday	-614.56009	172.9115	-3.5542	0.0004
Model Variance (sigma squared)	2825779	.	.	.

Parameter Estimates
TOTAL_CUST
holiday + ARIMA(1,1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.88487	0.0120	73.5834	<.0001
Autoregressive, Lag 1	0.34506	0.0240	14.3690	<.0001
holiday	-553.65801	172.4905	-3.2098	0.0014
Model Variance (sigma squared)	2813158	.	.	.

By comparing the model variance, we can see that holiday + ARIMA(1,1,1) has smaller variance.

b) Since the model performance was not greatly improved after adding holiday as regressor in the ARIMA model, we decided to add the summer holiday and winter holiday in the holiday variable, which means by changing May 15 to Aug 15 and Dec 15 to Jan 15 from “0” into "1" in the dataset and rename it as “new_holiday”. We will try to explore whether the new_holiday variable would have any impact on the model performance.



2.2.3-b

From picture 2.2.3-b, we can see that the MAPE has slightly decreased after adding the new_holiday for each model. However, it was not a big change.

Parameter Estimates
TOTAL_CUST
new_holiday + IMA(1,2) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.55095	0.0191	28.7969	<.0001
Moving Average, Lag 2	0.25737	0.0192	13.4341	<.0001
new_holiday	-375.31790	151.9794	-2.4695	0.0140
Model Variance (sigma squared)	2832994	.	.	.

Parameter Estimates
TOTAL_CUST
new_holiday + ARIMA(1,1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.88613	0.0119	74.2174	<.0001
Autoregressive, Lag 1	0.34875	0.0239	14.5718	<.0001
new_holiday	-359.62649	152.2682	-2.3618	0.0187
Model Variance (sigma squared)	2818349	.	.	.

Also, the model variance is both very similar to the previous one for each model.

Here, we list the results for the ARIMA(1,1,1) model as this model+regressor has a better model fit.

	holiday + ARIMA(1,1,1)	new_holiday + ARIMA(1,1,1)
MAPE	31.75	31.81
Model Variance	2813158	2818349

In conclusion, we think that adding the regressor holiday may have some effect on the model performance improvement but it was not the main effect which can greatly improve the model performance.

2.3 Comparison of models

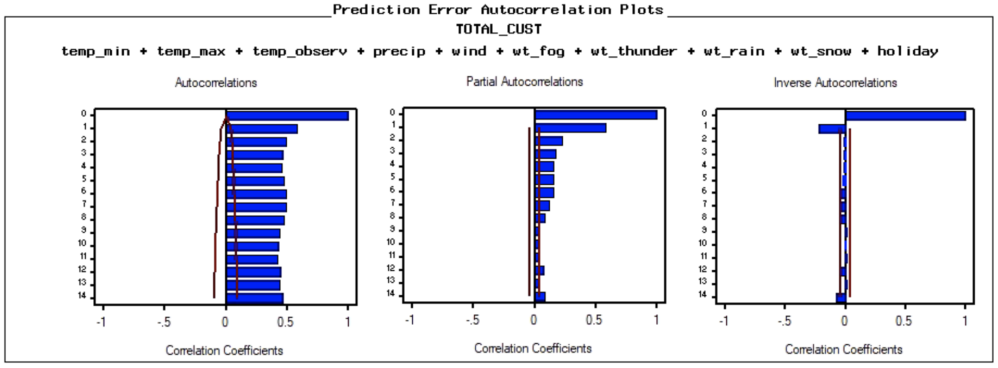
Model	MAPE	Model Variance
seasonal dummies+linear trend	74.98789	8294278
Log seasonal dummy + First difference + MA(2)	30.29561	0.11984
Cyclical trend model+ MA(2)	32.82	3213355
ARIMA(1,1,1)	32.06	2823398
ARIMA(0,1,2)	31.99756	2838646
holiday+ARIMA(1,1,1)	31.75055	2825779
holiday+ARIMA(0,1,2)	31.63643	2813158

Among the whole models, we mainly compared their Mean Absolute Percent Error and model variance. For the Deterministic Time Series Models, we found the model called Log seasonal dummy + First difference + MA(2) has the smallest MAPE(30.296). For the ARIMA models, we found holiday+ARIMA(0,1,2) is a better model with a small MAPE(31.636) and small model variance(2813158).

3. Multivariate Time Series Models

3.1 Regression model and analysis of regression residuals

3.1.1 Regression model with all predictors



(3.1a)

Parameter Estimates
TOTAL_CUST
temp_min + temp_max + temp_observ + precip + wind + wt_fog + wt_thunder + wt_rain + wt_snow + holiday

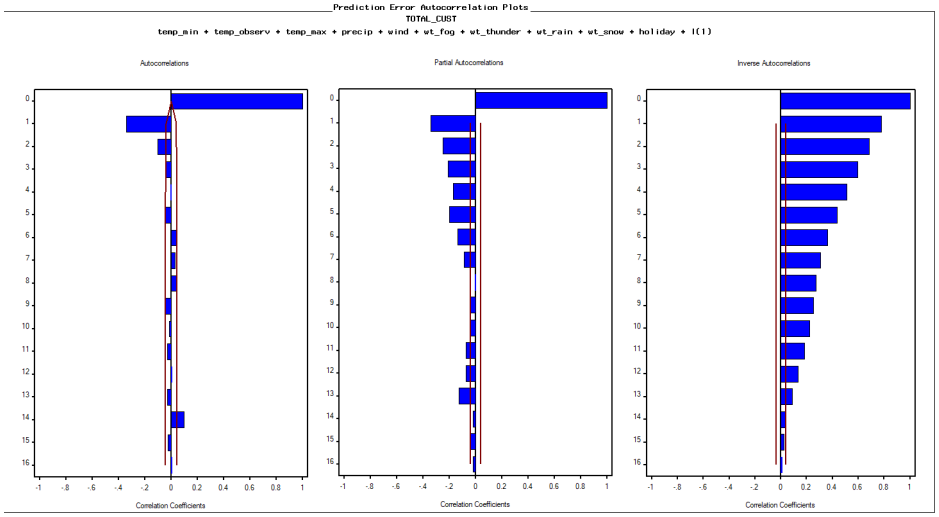
Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	4464	259.0970	17.2306	<.0001
temp_min	41.67714	30.3712	1.3723	0.1709
temp_max	233.75476	19.3429	12.0848	<.0001
temp_observ	-41.03328	32.3822	-1.2672	0.2059
precip	-23.26286	7.1165	-3.2689	0.0012
wind	-120.15714	38.1360	-3.1508	0.0018
wt_fog	-304.53220	108.2488	-2.8133	0.0052
wt_thunder	-553.31404	141.6272	-3.9068	0.0001
wt_rain	-3812	142.9994	-26.6592	<.0001
wt_snow	-340.95146	294.0481	-1.1595	0.2470
holiday	-1177	286.9407	-4.1034	<.0001
Model Variance (sigma squared)	6206425	.	.	.

Fit Range: 01JAN2011 to 31DEC2017

(3.1b)

Based on the table 3.1b, MAPE is 33.223, model variance is 6206423.From the ACF plot (3.1a), we can see that the prediction error is not decaying slowly which means it is not stationary. So, in order to explore the residuals and find the optimal model to fit the error, we decided to convert it to a stationary series by differencing.

3.1.2. The first difference of the regression model with all predictors.



(3.1c)

Looking at the ACF, PACF and Inverse correlation at 3.1c figure above, we can see that PACF and Inverse correlation decay exponentially while the ACF is chopped off after lag 2. So, we consider the model with the first difference and MA(1), MA(2).

Parameter Estimates

TOTAL_CUST

temp_min + temp_observ + temp_max + precip + wind + wt_fog + wt_thunder + wt_rain + wt_snow + holiday + I(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	0.59977	35.8331	0.0167	0.9867
temp_min	-128.27730	23.4920	-5.4605	<.0001
temp_observ	67.42115	17.2370	3.9114	0.0001
temp_max	132.36679	14.9336	8.8637	<.0001
precip	17.07048	3.9785	4.2907	<.0001
wind	-190.70868	24.7971	-7.6908	<.0001
wt_fog	-737.22077	63.8478	-11.5465	<.0001
wt_thunder	-374.65273	87.3508	-4.2891	<.0001
wt_rain	-303.92515	106.0016	-2.8672	0.0044
wt_snow	-220.28856	181.4580	-1.2140	0.2256
holiday	-438.17744	146.1307	-2.9985	0.0029
Model Variance (sigma squared)	3281918	.	.	.

Fit Range: 01JAN2011 to 31DEC2017

(3.1d)

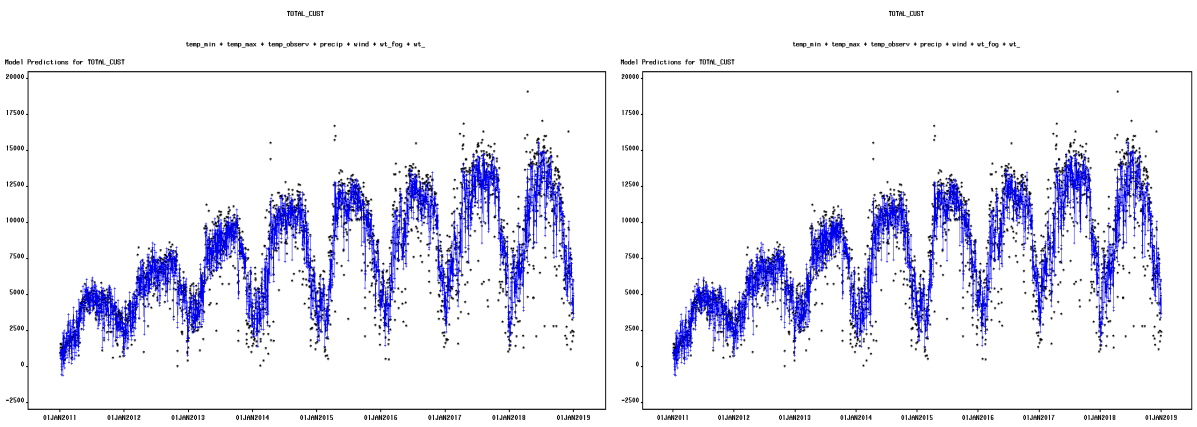
Based on the table 3.1d, the intercept and wt_snow variable have p-value higher than 0.05, it is insignificant. We can remove these options out of the model.

In general, the MAPE of this model is 27.653 and model variance is 3281918, so it is better than the one without the first difference.

In this step, we compare IMA(1,1) and IMA(1,2) without intercept to check which model is better. The table below shows the MAPE and variance between the two models in which IMA(1,2) performs better than IMA(1,1) because it has lower MAPE (28.30) and variance(2257081).

Also, the figure 3.1e shows there is no significant difference between IMA(1,1) and IMA(1,2). In conclusion, we decide to use IMA(1,2) NOINT to check further.

Model	All regressors + MA(1) NOINT	All regressors + MA(2) NOINT
MAPE	28.30	27.64
Model Variance	2305059	2257081

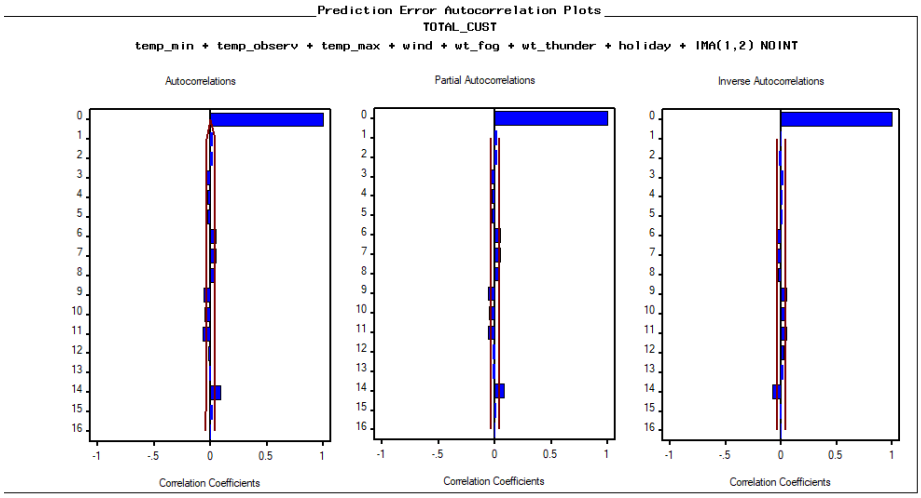


3.1e

3.1.3 Add Error model on the previous model

In this part, we try to fit the regression model with IMA(1,2) without wt_snow variable and intercept first. Then, we continue to remove wt_rain and precip variables because of its insignificance. Hence, we have the model shown below.

- Regression model with IMA(1,2) and without wt_snow, wt_rain, precip and Intercept (figure 3.1f)
  MAPE: 27.69



Parameter Estimates
TOTAL_CUST
temp_min + temp_observ + temp_max + wind + wt_fog + wt_thunder + holiday + IMA(1,2) NOINT

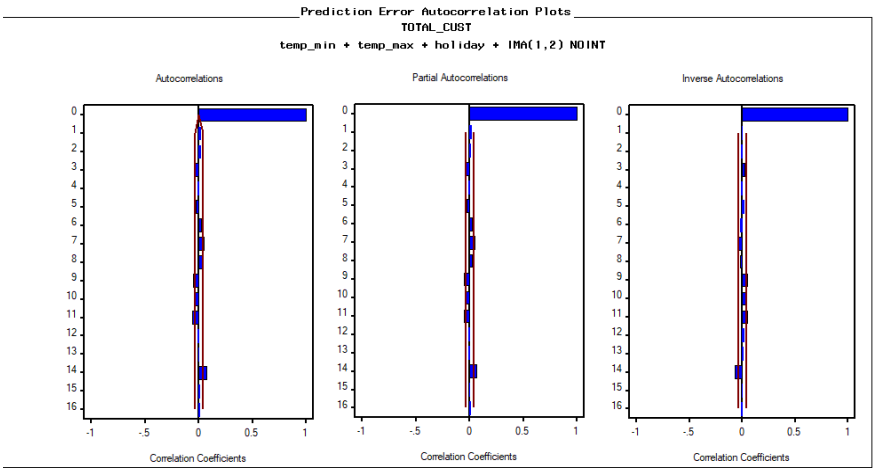
Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.69458	0.0204	33.9817	<.0001
Moving Average, Lag 2	0.15675	0.0202	7.7697	<.0001
temp_min	-113.16541	20.8292	-5.4330	<.0001
temp_observ	49.47041	18.7934	2.6323	0.0088
temp_max	190.31092	13.0169	14.6203	<.0001
wind	-250.34268	23.7890	-10.5235	<.0001
wt_fog	-738.77666	65.5138	-11.2767	<.0001
wt_thunder	-507.00391	86.1223	-5.8870	<.0001
holiday	-729.18579	161.7985	-4.5068	<.0001
Model Variance (sigma squared)	2260948	.	.	.

Fit Range: 01JAN2011 to 31DEC2017 (3.1f)

- Regression model with temperature and holiday variables (figure 3.1g)

Furthermore, we also want to see how the combination of the temperature and the holiday affect the number of the customers. So, we fit the regression model with IMA(1,2) and temperature, holiday variables.

MAPE: 29.91



Parameter Estimates

TOTAL_CUST

temp_min + temp_max + holiday + IMA(1,2) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.66357	0.0203	32.6658	<.0001
Moving Average, Lag 2	0.18629	0.0199	9.3518	<.0001
temp_min	-155.32299	14.4090	-10.7796	<.0001
temp_max	232.64370	13.0225	17.8648	<.0001
holiday	-656.61734	169.1390	-3.8821	0.0001
Model Variance (sigma squared)	2510664	.	.	.

Fit Range: 01JAN2011 to 31DEC2017 (3.1g)

The MAPE(29.91) and model variance(2510664) are a little bit higher compared to others.

In conclusion, we decide to choose the IMA(1,2) model without wt_snow, wt_rain, precip and intercept for this part because it provides the lowest MAPE(27.69) and model variance(2260948).

3.2 Cross correlation analysis to identify lagged values of predictors and use them as predictors in your model.

Because there are many independent variables in our dataset, we chose temp_min, temp_max and precip these continuous variables as predictors to do the cross correlation analysis. We firstly found that the predictors temp_min and temp_max are non-stationary because their autocorrelations deacy very slowly. Then we will do the differencing on them in the following, however, the precip shows it's stationary which means it will not need differencing in the following steps.

When we did cross correlation between the total_cust and temp_min, we found there is no correlation outside the bound (shown in the picture 3.2a below) so there is no significant relationship at specific lag. The correlation of all the lags are not significant. So we can not use useful lag values as our predictors here.

Variable temp_min has been differenced.

Correlation of total_cust and temp_min	
Period(s) of Differencing	1
Variance of input =	8.02862
Number of Observations	2921
Observation(s) eliminated by differencing	1

Crosscorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
-24	-72.044286	-.01217	. .
-23	-24.215580	-.00409	. .
-22	-163.397	-.02759	* .
-21	-30.603803	-.00517	. .
-20	215.329	0.03636	. *
-19	28.187685	0.00476	. .
-18	8.721831	0.00147	. .
-17	134.962	0.02279	. .
-16	-254.004	-.04289	* .

3.2a

The below table3.2b shows cross correlation between the total_cust and temp_max. (It's interesting that most of lags just show one star in the autocorrelation, so we have to choose lag 1, 2 which have one more stars in the autocorrelation.) As we can see from the picture below when the lag=0, it's a significant positive relationship between the total_cust and temp_max. When the lag=1,lag=2, it's a significant negative relationship between the total_cust and temp_max. Based on the current period, they are related to each other positively but if it increases to the last period, it will have a negative effect on the total_cust. As a result, we choose temp_max at lag1 and 2 as our predictors.

-1	2126.731	0.31277	. |******
0	890.003	0.13089	. |***
1	-682.212	-.10033	**| .
2	-604.157	-.08885	**| .

3.2b

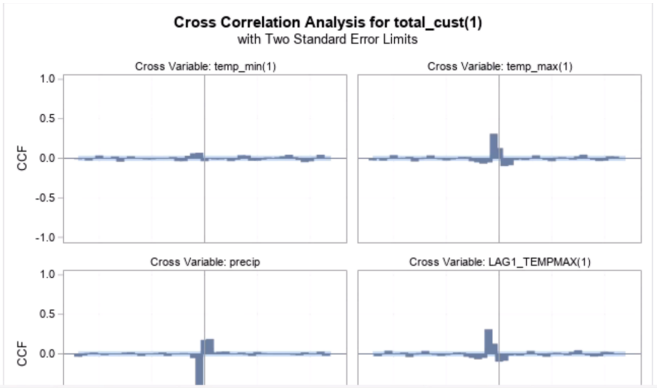
The below table3.2c shows cross correlation between the total_cust and precip. When the lag=-2, it's a significant positive relationship between the total_cust and precip. When the lag=-1, it's a significant negative relationship between the total_cust and precip. When the lag=1, it's a significant positive relationship between the total_cust and precip. Based on the given period, if it increases to the last period(lag=1), it will have a positive effect on the total_cust. As a result, we choose precip at lag 1 as our predictors.

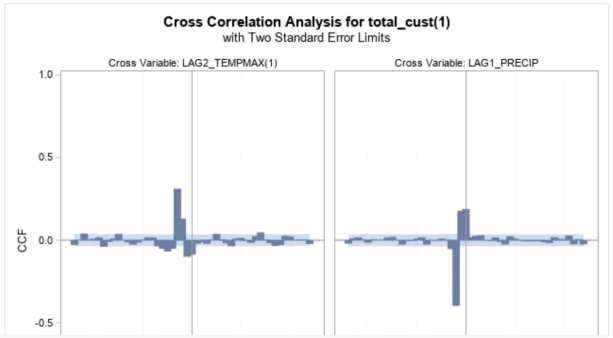
-2	5881.441	0.28333	. |******
-1	-9870.966	-.47553	**********| .
0	-200.755	-.00967	. | .
1	2972.243	0.14319	. |***

3.2c

After choosing temp_max at lag1 and 2 and precip at lag 1 as our predictors to add into our model through the SAS code. The 2 pictures below show the CCF of all variables after differencing. And they show the CCF of predictors that temp_max at lag1 and 2 and precip at lag 1(3.2d).

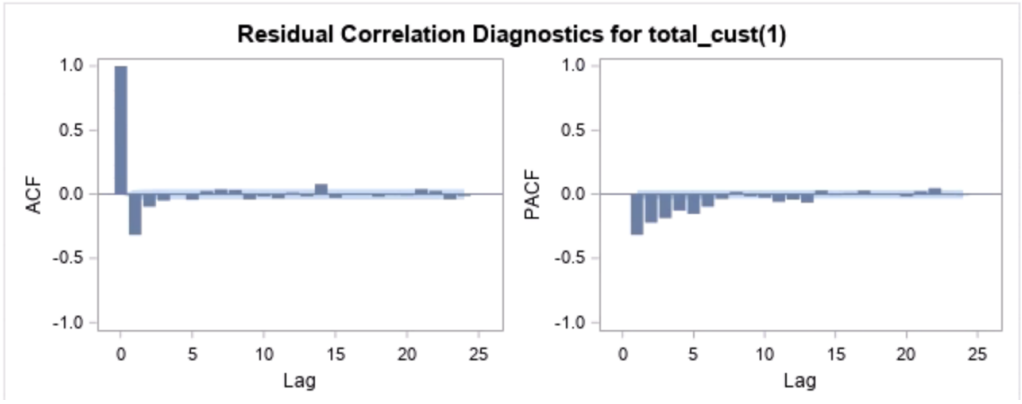
.





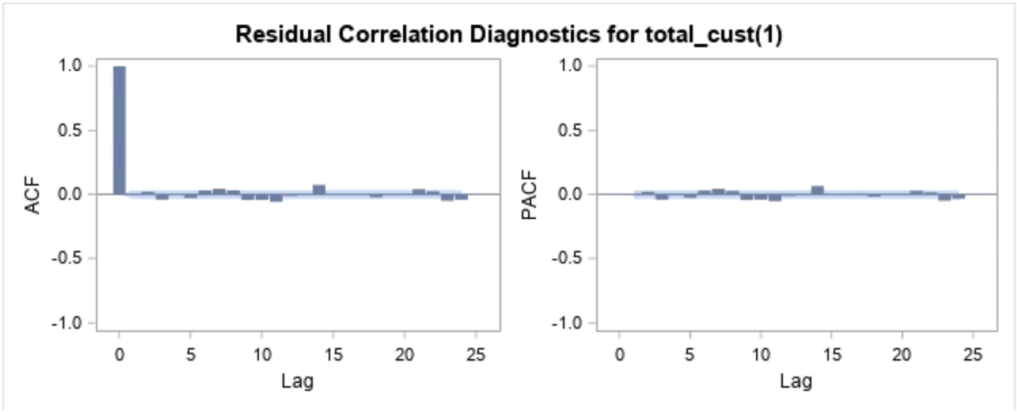
3.2d

The picture below 3.2c shows that the errors are not white noise after we added the significant lags into the model so it means we should add the error model MA(2).



3.2c

Then we added error model MA(2) into the previous model. As we can see from the plot below 3.2d, we got the white-noise residuals. And this model variance is 2907258. And our final model is temp_min+temp_max+precip+lag1_tempmax+lag2_tempmax+lag1_precip+MA(2)



3.2d

4. Conclusion

Model	MAPE	Model Variance
seasonal dummies+linear trend	74.98789	8294278

Log seasonal dummy + First difference + MA(2)	30.29561	0.11984
Cyclical trend model	43.15	3213355
ARIMA(1,1,1)	32.06	2823398
ARIMA(0,1,2)	31.99756	2838646
holiday+ARIMA(1,1,1)	31.75055	2825779
holiday+ARIMA(0,1,2)	31.63643	2813158
Regression model with all predictors	33.223	6206420
Regression model with all predictors + first difference	27.653	3281918
Regression model + IMA(1,2) + no wt_snow, wt_rain, precip+ no Intercept	27.69	2260940
Regression model + IMA(1,2)+Temperature, Precip + No intercept	29.91	2510664

Among the whole models, we mainly compared their Mean Absolute Percent Error and model variance. For the Deterministic Time Series Models, we found the model called Log seasonal dummy + First difference + MA(2) has the smallest MAPE(30.296). For the ARIMA models, we found holiday+ARIMA(0,1,2) is a better model with small MAPE(31.636) and small model variance(2813158).

After we did multivariate models in the following, we also included them in this part and made comparisons. We found that there is no obvious change between Regression model with all predictors + first difference model and Regression model + IMA(1,2) + no wt_snow, wt_rain, precip+ no Intercept which we deleted some insignificant variables and added error model. But the model is the best model with the smallest MAPE among all the models.

Based on the best model we chose, we checked its forecast data set and we got some business findings. The demands of shared bike are mainly influenced by temperature, wind, thunder, fog, holiday, etc. And from the left picture 4a, we found in December of 2018 the demand for bikes is changing obviously from 1208 bikes to 7466 bikes. However, at the beginning of January, the demand is predicted that it will decrease to around 4000 bikes per day(4b), therefore, the company can decrease supply in January to guarantee a enough demand so that the company can avoid the loss in the oversupply or over shortage.

DATE	ACTUAL	PREDICT
15DEC2018	1208	6460
16DEC2018	2363	4544
17DEC2018	7772	5602
18DEC2018	7514	6778
19DEC2018	7466	7200
20DEC2018	4399	7165
21DEC2018	5696	6538
22DEC2018	3710	6202
23DEC2018	3270	6045
24DEC2018	2492	4505
25DEC2018	1744	3933
26DEC2018	3752	4716

4a

Forecast Data Set

TOTAL_CUST

temp_min + temp_max + temp_observ + wind + wt_fog + wt_thunder + holiday + IMA(1,2) NOINT

DATE	ACTUAL	PREDICT	U95	L95	ERROR	NERROR	temp_min	temp_max
30DEC2018	4929	5957	8904	3010	-1028	-0.6837	0.8167	11.0667
31DEC2018	2401	3614	6561	667.1366	-1213	-0.8069	0.2500	8.5167
01JAN2019	.	3895	6842	948.1407	.	.	0.5949	8.5192
02JAN2019	.	4064	7145	982.1459	.	.	0.7873	8.5192
03JAN2019	.	4069	7182	956.5515	.	.	0.6469	8.5192
04JAN2019	.	4101	7245	958.2272	.	.	0.3918	8.5192
05JAN2019	.	4166	7339	991.9998	.	.	-0.0207	8.5192
06JAN2019	.	4168	7371	963.8883	.	.	-0.0537	8.5192
07JAN2019	.	4041	7274	807.1414	.	.	0.2509	8.5192
08JAN2019	.	4085	7348	822.3702	.	.	0.5949	8.5192
09JAN2019	.	4064	7356	771.3093	.	.	0.7873	8.5192
10JAN2019	.	4069	7390	747.6817	.	.	0.6469	8.5192
11JAN2019	.	4101	7452	751.2702	.	.	0.3918	8.5192

4b