Appendix

**Data Dictionary**

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| **VendorID** | A code indicating the TPEP provider that provided the record. |
| **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.** |
| **tpep\_pickup\_datetime** | The date and time when the meter was engaged. |
| **tpep\_dropoff\_datetime** | The date and time when the meter was disengaged. |
| **Passenger\_count** | The number of passengers in the vehicle. |
| This is a driver-entered value. |
| **Trip\_distance** | The elapsed trip distance in miles reported by the taximeter. |
| **Pickup\_longitude** | Longitude where the meter was engaged. |
| **Pickup\_latitude** | Latitude where the meter was engaged. |
| **RateCodeID** | The final rate code in effect at the end of the trip. |
| **1= Standard rate** |
| **2=JFK** |
| **3=Newark** |
| **4=Nassau or Westchester** |
| **5=Negotiated fare** |
| **6=Group ride** |
| **Store\_and\_fwd\_flag** | This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server. |
| **Y= store and forward trip** |
| **N= not a store and forward trip** |
| **Dropoff\_longitude** | Longitude where the meter was disengaged. |
| **Dropoff\_ latitude** | Latitude where the meter was disengaged. |
| **Payment\_type** | A numeric code signifying how the passenger paid for the trip. |
| **1= Credit card** |
| **2= Cash** |
| **3= No charge** |
| **4= Dispute** |
| **5= Unknown** |
| **6= Voided trip** |
| **Fare\_amount** | The time-and-distance fare calculated by the meter. |
| **Extra** | Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges. |
| **MTA\_tax** | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| **Improvement\_surcharge** | $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015. |
| **Tip\_amount** | Tip amount – This field is automatically populated for credit card tips. Cash tips are not included. |
| **Tolls\_amount** | Total amount of all tolls paid in trip. |
| **Total\_amount** | The total amount charged to passengers. Does not include cash tips. |

**Data Pre Processing:**

The data is available at nyc.gov website. The data also contains a few record of 2013 but to avoid the discrepancies as there’s huge seasonality in the data. We didn’t want to get biased in our way to approach the data so we took the entire 2014 and 2015 years’ data. Below is the link to refer the data.

Data Link: <http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>

Raw data had a period of second- interpolated to period of a day.

We observed few descripencies in the data like Trip amount was $3000 but the Trip distance and Number of Trips was zero. So, to elaborate on the discrepanies in the data set we’ve created an excel sheet for the same. Below is the excel:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **BOROUGH** | **FLAG\_ISSUE** | **NO\_TRIPS** | **PASSENGER\_COUNT** | **TRIP\_DISTANCE** | **FARE\_AMOUNT** |  |
| Brooklyn | N | 13567275 | 19349550 | 43909451.72 | 181,452,184 | 0.65% |
| Brooklyn | Y | 88587 | 117918 | 2402.05 | 2,786,914 |  |
| Manhattan | N | 21053064 | 29123128 | 57962698.64 | 243,530,727 |  |
| Manhattan | Y | 163200 | 210976 | 8922.32 | 3,563,816 | 0.78% |
| Queens | N | 19594 | 25328 | 98528.3 | 355,349 |  |
| Queens | Y | 5099 | 6649 | 71.36 | 152,282 | 26.02% |
| Staten Island | N | 9353 | 12345 | 53809.93 | 212,308 |  |
| Staten Island | Y | 2341 | 3437 | 62.62 | 158,248 | 25.03% |
| The Bronx | N | 90862 | 124271 | 323317.01 | 1,225,287 |  |
| The Bronx | Y | 3900 | 4710 | 96.46 | 71,705 | 4.29% |

Our data set on the website had been in csv format, we’ve converted it in bat file format.

Also, the data set available had been for entire New York city which we have distributed into dfferent boroughs based on their latitude and longitude.

**B**elow is the code used in SAS for the bat file conversion followed the boroughs distribution.

DM LOG 'CLEAR' LOG;  
  
%LET PATH C:\Project Files;  
LIBNAME SASDATA "&PATH.\SASDATA";  
  
OPTIONS COMPRESS=YES;  
  
%MACRO DATA\_READIN(YEAR,PER);  
  
DATA TD\_&YEAR.\_&PER.;  
INFILE "&PATH.\RAWDATA\green\_tripdata\_&YEAR.-&PER..csv" DELIMITER = ',' DSD FIRSTOBS=2 TRUNCOVER;  
   
INFORMAT VENDORID $100.;  
INFORMAT LPEP\_PICKUP\_DATETIME $50.;  
INFORMAT LPEP\_DROPOFF\_DATETIME ANYDTDTE.;  
INFORMAT STORE\_AND\_FWD\_FLAG $10.;  
INFORMAT RATECODEID $10.;  
INFORMAT PICKUP\_LONGITUDE BEST32.;  
INFORMAT PICKUP\_LATITUDE BEST32.;  
INFORMAT DROPOFF\_LONGITUDE BEST32.;  
INFORMAT DROPOFF\_LATITUDE BEST32.;  
INFORMAT PASSENGER\_COUNT BEST32.;  
INFORMAT TRIP\_DISTANCE BEST32.;  
INFORMAT FARE\_AMOUNT COMMA32.;  
INFORMAT EXTRA COMMA32.;  
INFORMAT MTA\_TAX COMMA32.;  
INFORMAT TIP\_AMOUNT COMMA32.;  
INFORMAT TOLLS\_AMOUNT COMMA32.;  
INFORMAT EHAIL\_FEE COMMA32.;  
INFORMAT IMPROVEMENT\_SURCHARGE COMMA32.;  
INFORMAT TOTAL\_AMOUNT COMMA32.;  
INFORMAT PAYMENT\_TYPE COMMA32.;  
INFORMAT TRIP\_TYPE $50.;  
  
FORMAT VENDORID $100.;  
FORMAT LPEP\_PICKUP\_DATETIME $50.;  
FORMAT LPEP\_DROPOFF\_DATETIME MMDDYY10.;  
FORMAT STORE\_AND\_FWD\_FLAG $10.;  
FORMAT RATECODEID $10.;  
FORMAT PICKUP\_LONGITUDE BEST32.;  
FORMAT PICKUP\_LATITUDE BEST32.;  
FORMAT DROPOFF\_LONGITUDE BEST32.;  
FORMAT DROPOFF\_LATITUDE BEST32.;  
FORMAT PASSENGER\_COUNT BEST32.;  
FORMAT TRIP\_DISTANCE BEST32.;  
FORMAT FARE\_AMOUNT COMMA32.;  
FORMAT EXTRA COMMA32.;  
FORMAT MTA\_TAX COMMA32.;  
FORMAT TIP\_AMOUNT COMMA32.;  
FORMAT TOLLS\_AMOUNT COMMA32.;  
FORMAT EHAIL\_FEE COMMA32.;  
FORMAT IMPROVEMENT\_SURCHARGE COMMA32.;  
FORMAT TOTAL\_AMOUNT COMMA32.;  
FORMAT PAYMENT\_TYPE COMMA32.;  
FORMAT TRIP\_TYPE $50.;  
  
INPUT  
  
VENDORID $  
LPEP\_PICKUP\_DATETIME $  
LPEP\_DROPOFF\_DATETIME  
STORE\_AND\_FWD\_FLAG $  
RATECODEID $  
PICKUP\_LONGITUDE  
PICKUP\_LATITUDE  
DROPOFF\_LONGITUDE  
DROPOFF\_LATITUDE  
PASSENGER\_COUNT  
TRIP\_DISTANCE  
FARE\_AMOUNT  
EXTRA  
MTA\_TAX  
TIP\_AMOUNT  
TOLLS\_AMOUNT  
EHAIL\_FEE  
IMPROVEMENT\_SURCHARGE  
TOTAL\_AMOUNT  
PAYMENT\_TYPE  
TRIP\_TYPE $  
  
;  
RUN;  
  
%MEND DATA\_READIN;  
  
%DATA\_READIN(2015,01);  
%DATA\_READIN(2015,02);  
%DATA\_READIN(2015,03);  
%DATA\_READIN(2015,04);  
%DATA\_READIN(2015,05);  
%DATA\_READIN(2015,06);  
%DATA\_READIN(2015,07);  
%DATA\_READIN(2015,08);  
%DATA\_READIN(2015,09);  
%DATA\_READIN(2015,10);  
%DATA\_READIN(2015,11);  
%DATA\_READIN(2015,12);  
  
  
DATA SASDATA.TD\_2015\_ALL;  
SET TD\_2015\_01 TD\_2015\_02 TD\_2015\_03 TD\_2015\_04 TD\_2015\_05 TD\_2015\_06 TD\_2015\_07 TD\_2015\_08 TD\_2015\_09 TD\_2015\_10 TD\_2015\_11 TD\_2015\_12;  
DATE\_P = INPUT(SUBSTR(COMPRESS(LPEP\_PICKUP\_DATETIME),1,10),YYMMDD10.);  
FORMAT DATE\_P DATE7.;  
TIME\_P = SUBSTR(COMPRESS(LPEP\_PICKUP\_DATETIME),11,8);  
DATETIME\_P = COMPRESS(PUT(DATE\_P,DATE7.)||":"||TIME\_P);  
DATETIME\_FINAL = INPUT(DATETIME\_P,DATETIME16.);  
FORMAT DATETIME\_FINAL DATETIME16.;  
HOUR = HOUR(DATETIME\_FINAL);  
DATE\_F = DATEPART(DATETIME\_FINAL);  
DATE\_NEW1 = PUT(DATE\_P,DATE7.)||":"||PUT(HOUR,2.)||":00:00";  
PICKUP\_DATE = INPUT(DATE\_NEW1,DATETIME16.);  
FORMAT PICKUP\_DATE DATETIME16.;  
\*KEEP PICKUP\_DATE PASSENGER\_COUNT TRIP\_DISTANCE FARE\_AMOUNT;  
RUN;  
  
DATA SASDATA.NYT\_DATA;  
SET SASDATA.TD\_2015\_ALL;  
LENGTH BOROUGH $50.;  
  
\*GET THE BOROUGH INFO BASED ON LAT/LONG APPROXIMATION;  
  
IF (PICKUP\_LONGITUDE GE -74.04 AND PICKUP\_LONGITUDE LE -73.72) AND (PICKUP\_LATITUDE GE 40.56 AND PICKUP\_LATITUDE LE 40.7281)   
THEN BOROUGH = "Brooklyn";  
ELSE IF (PICKUP\_LONGITUDE GE -73.97 AND PICKUP\_LONGITUDE LE -73.79) AND (PICKUP\_LATITUDE GE 40.68 AND PICKUP\_LATITUDE LE 40.88)   
THEN BOROUGH = "Manhattan";  
ELSE IF (PICKUP\_LONGITUDE GE -73.97 AND PICKUP\_LONGITUDE LE -73.86) AND (PICKUP\_LATITUDE GE 40.88 AND PICKUP\_LATITUDE LE 40.91)   
THEN BOROUGH = "The Bronx";  
ELSE IF (PICKUP\_LONGITUDE GE -74.26 AND PICKUP\_LONGITUDE LE -74) THEN BOROUGH = "Staten Island";  
ELSE IF (PICKUP\_LONGITUDE GE -73.94 AND PICKUP\_LONGITUDE LE -73.76) THEN BOROUGH = "Queens";  
ELSE IF (PICKUP\_LONGITUDE GE -75 AND PICKUP\_LONGITUDE LE -72.5) AND (PICKUP\_LATITUDE GE 40 AND PICKUP\_LATITUDE LE 41)   
THEN BOROUGH = "Other NY";  
ELSE BOROUGH = "Non-NYC";  
DATE = DATEPART(PICKUP\_DATE);  
FORMAT DATE MMDDYY10.;  
RUN;  
  
PROC SUMMARY DATA=SASDATA.NYT\_DATA NWAY MISSING;  
CLASS BOROUGH;  
VAR PASSENGER\_cOUNT TRIP\_DISTANCE FARE\_AMOUNT;  
OUTPUT OUT=SASDATA.NYT\_BOROUGH (DROP=\_TYPE\_ RENAME=\_FREQ\_=NO\_TRIPS) SUM=;  
RUN;  
  
PROC SQL;  
CREATE TABLE SASDATA.OTHER\_NYT AS SELECT DISTINCT(PICKUP\_LONGITUDE),PICKUP\_LATITUDE FROM SASDATA.NYT\_DATA WHERE BOROUGH = "Other NY";  
quit;  
  
PROC SUMMARY DATA=SASDATA.NYT\_DATA NWAY MISSING;  
CLASS BOROUGH PICKUP\_DATE;  
VAR PASSENGER\_COUNT TRIP\_DISTANCE FARE\_AMOUNT;  
OUTPUT OUT=SASDATA.DAY\_WISE\_B\_SUM(DROP=\_TYPE\_ RENAME=\_FREQ\_ = NO\_TRIPS\_) SUM= ;  
RUN;  
  
PROC SUMMARY DATA=SASDATA.NYT\_DATA NWAY MISSING;  
CLASS BOROUGH DATE;  
VAR PASSENGER\_COUNT TRIP\_DISTANCE FARE\_AMOUNT;  
OUTPUT OUT=SASDATA.DAY\_WISE\_B\_SUM\_2(DROP=\_TYPE\_ RENAME=\_FREQ\_ = NO\_TRIPS\_) SUM= ;  
RUN;  
  
PROC SUMMARY DATA=SASDATA.DAY\_WISE\_B\_SUM\_2 NWAY MISSING;  
CLASS BOROUGH ;  
VAR PASSENGER\_COUNT TRIP\_DISTANCE FARE\_AMOUNT NO\_TRIPS\_;  
OUTPUT OUT=SASDATA.NY\_BOROUGH\_SUM(DROP=\_TYPE\_ RENAME=\_FREQ\_ = NO\_DAYS) SUM= ;  
RUN;  
  
PROC SORT DATA=SASDATA.DAY\_WISE\_B\_SUM\_2 OUT=DAY\_WISE\_B\_SUM\_2;BY DATE;RUN;  
  
PROC TRANSPOSE DATA=DAY\_WISE\_B\_SUM\_2 OUT=SASDATA.TIME\_SERIES\_NYT;  
BY DATE;  
ID BOROUGH;  
VAR PASSENGER\_COUNT TRIP\_DISTANCE FARE\_AMOUNT NO\_TRIPS\_;  
RUN;  
  
%MACRO TIME\_SERIES\_SPLIT(BOR,VALUE);  
DATA SASDATA.TIME\_SER\_&BOR.;  
SET SASDATA.DAY\_WISE\_B\_SUM\_2;  
WHERE UPCASE(BOROUGH) = "&VALUE.";  
RUN;  
%MEND TIME\_SERIES\_SPLIT;  
  
%TIME\_SERIES\_SPLIT(BKLYN,BROOKLYN);  
%TIME\_SERIES\_SPLIT(QUEENS,QUEENS);  
%TIME\_SERIES\_SPLIT(MANHATTAN,MANHATTAN);  
%TIME\_SERIES\_SPLIT(BRONX,THE BRONX);  
%TIME\_SERIES\_SPLIT(STN\_ISLAND,STATEN ISLAND);  
%TIME\_SERIES\_SPLIT(OTHER\_NY,OTHER NY);  
%TIME\_SERIES\_SPLIT(NON\_NY,NON-NYC);

## Modeling Apporach

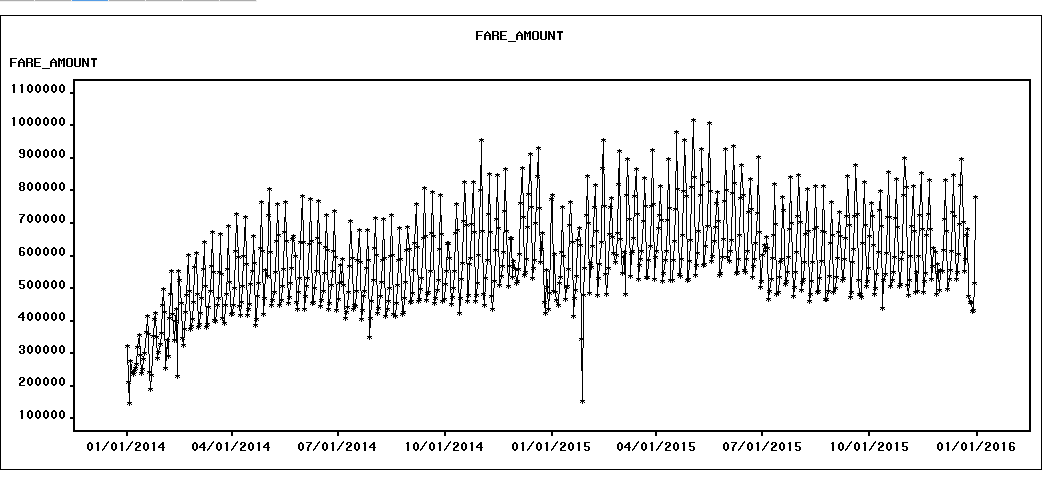
### Overall Green Taxi Data

1. **Forecasting Fare Amount for Overall data for green taxi without Uber data observations**

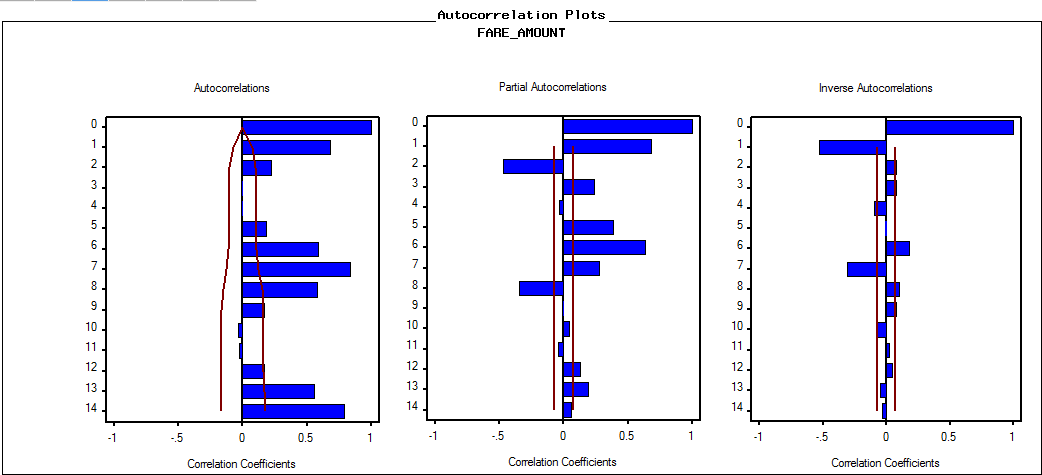
Data Range: 01-Jan-2014 to 31-Dec-2015

Period: Daily

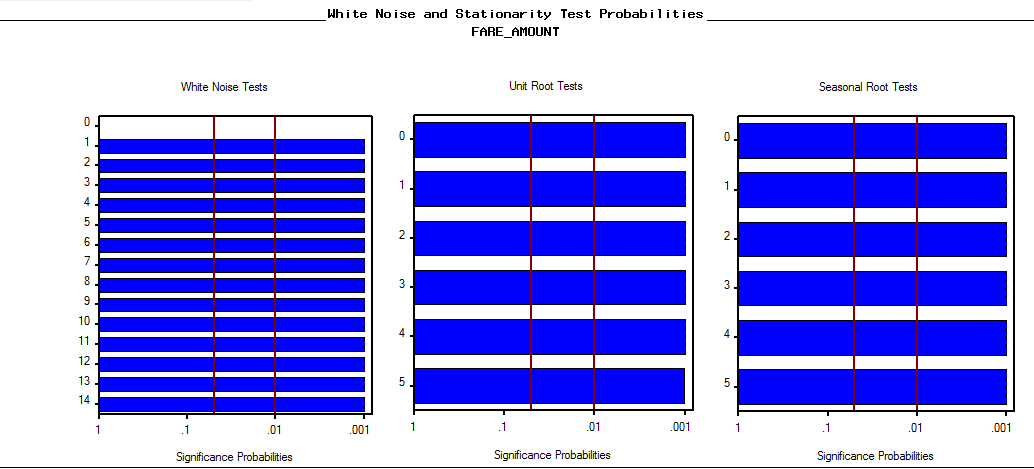
* Initial Data Exploration



We observe seasonality in the data and Point intervention at 01/27/2015



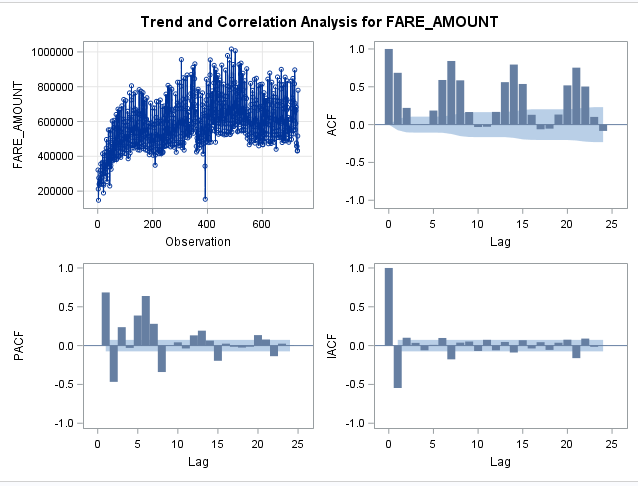
Autocorrelation plots depict seasonality in the data



White Noise test depicts that series is not white noise thus reject the Null Hypothesis that Series is white noise.

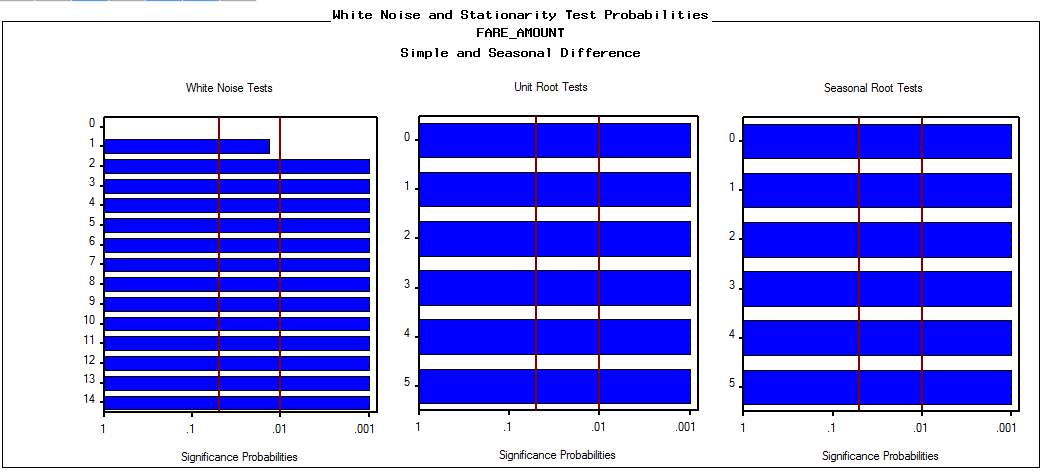
Unit root test shows that series is stationary.

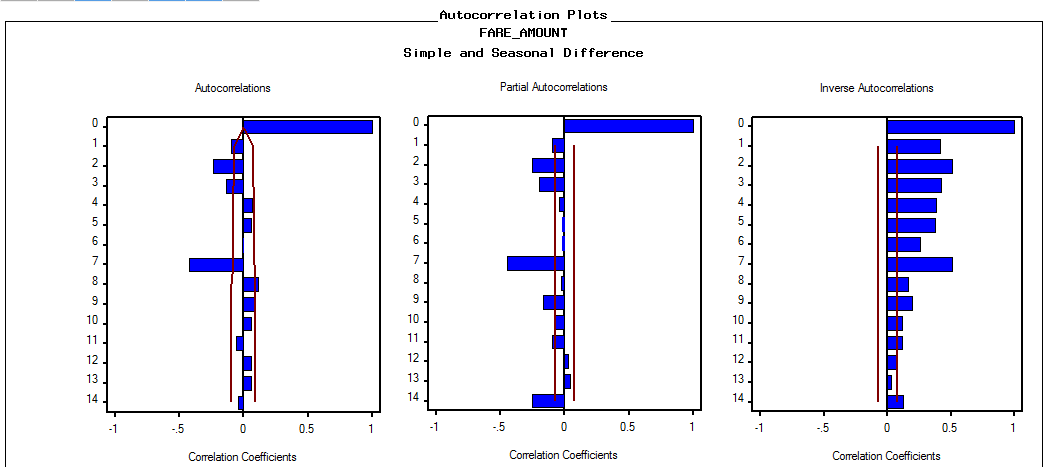
**SAS Base outputs**



* PERFORM A FIRST DIFFERENCING AND SEASONAL/WEEKLY DIFFERENCING

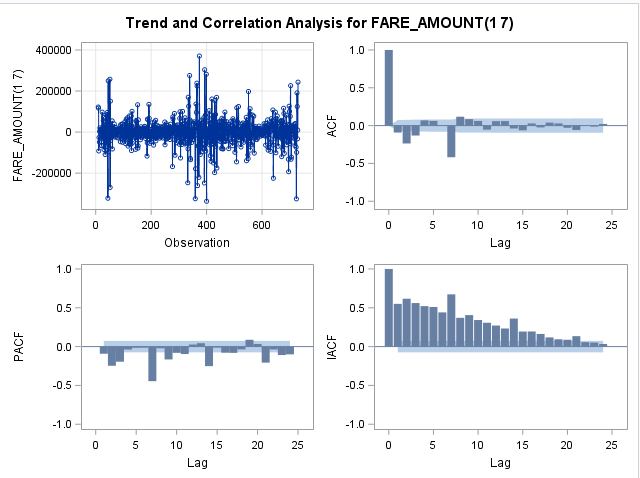






Plots show seasonality at 7th lag, thus series has weekly seasonality

SAS BASE Output

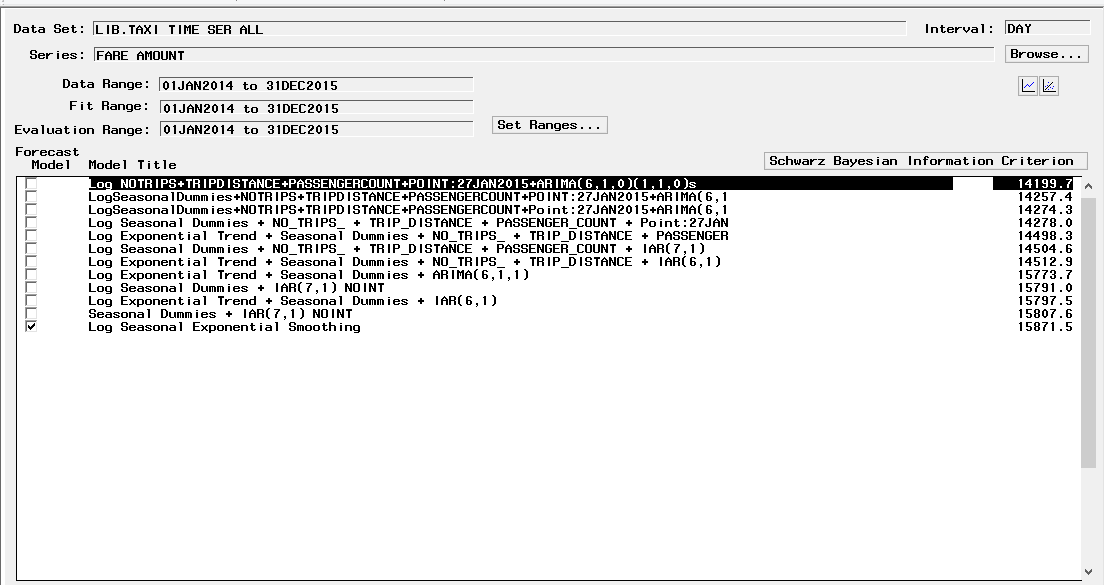


* Modeling Phase

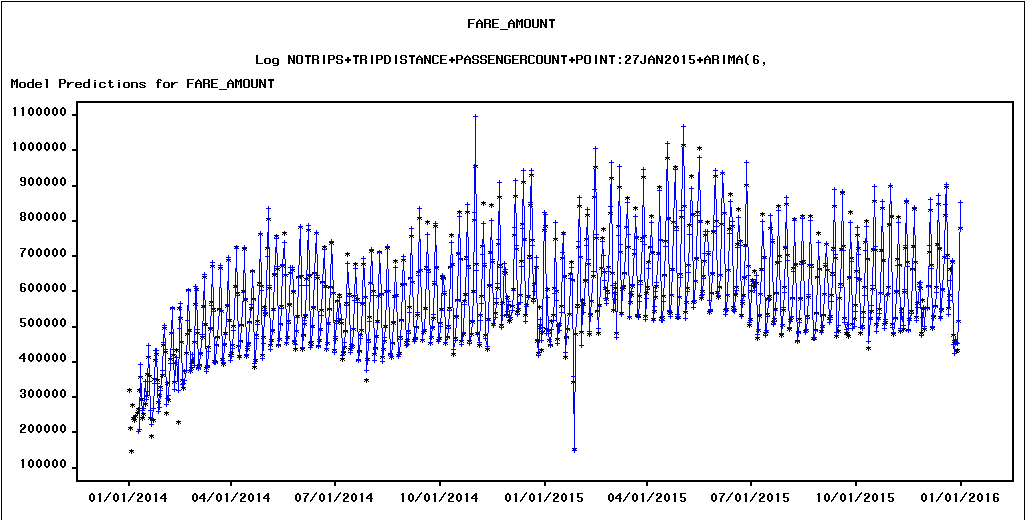
Below screenshot shows the models built in TSFS to forecast fare amount based on SBC and MAPE criterion

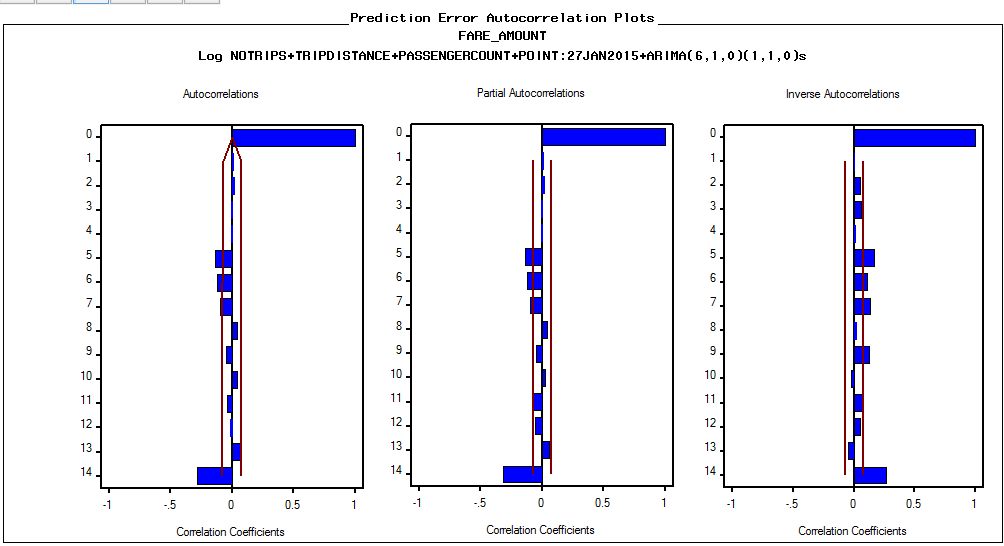
We are select best model based on following factors

* SBC value
* MAPE value
* Parameter estimates
* White noise and Unit root test
* Autocorrelation plots

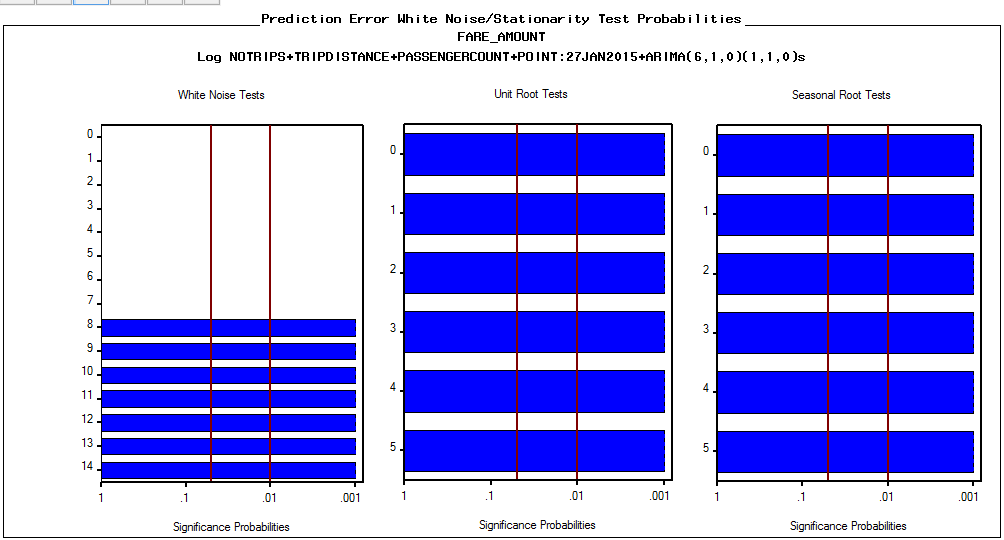


Log ARIMA (6,1,0) (1,1,0) s with Point intervention 01/27/2016 and Regressor is the best model with SBC values among the other models built in TSFS

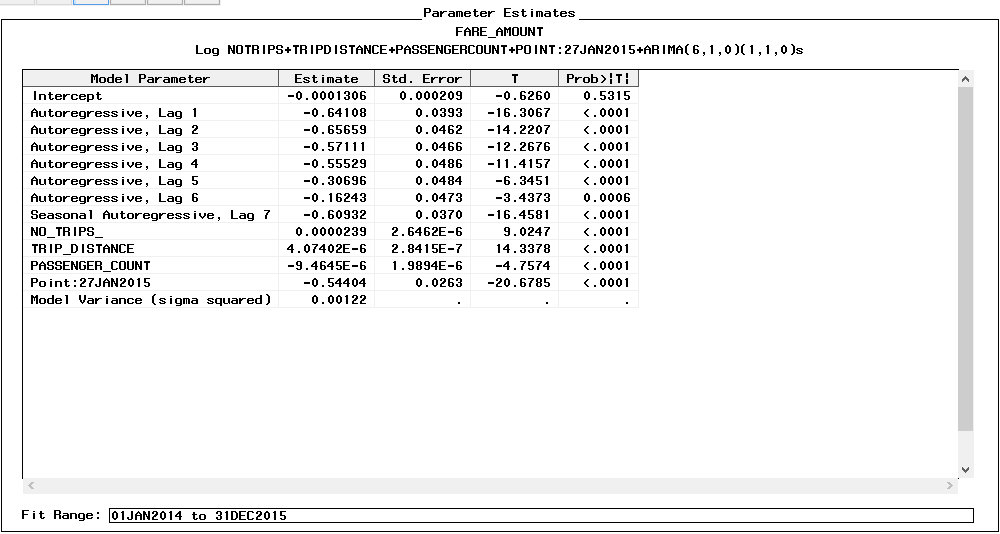




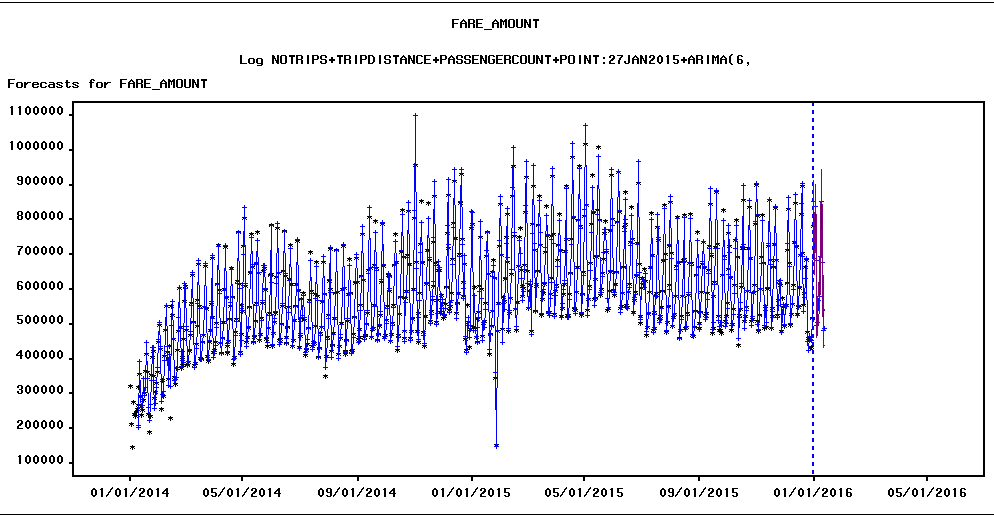
Auto correlation Plots shows that residuals are not correlated



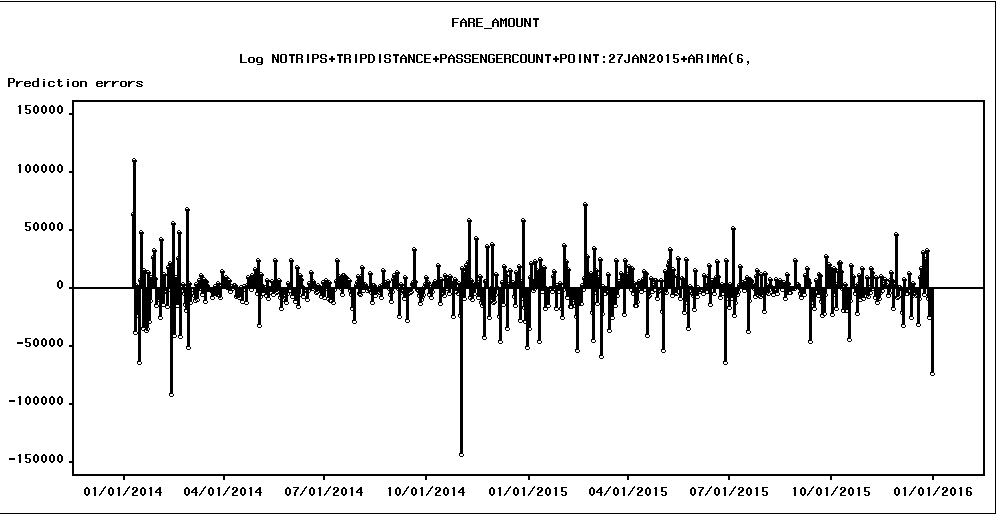
Series in white noise up to 8th lag



Parameter estimates is significant



Forecast plot seems to follow the seaonality



Residual plots seem fine except for few negative run overs

As the results were not satisfactory team proceeded to build better forecast model in SAS Base– using Cross Correlations and Pre-Whitening Techniques within PROC ARIMA Procedure:

**Prewhitening and CCF (Cross Correlation Function) analysis**

In order to estimate the CCF between the outcome series (Fare Amount) with a set of input/regressor series, we will need to prewhiten the input series, before applying the cross correlation function.

**Prewhiten ‘No-Trips’:**

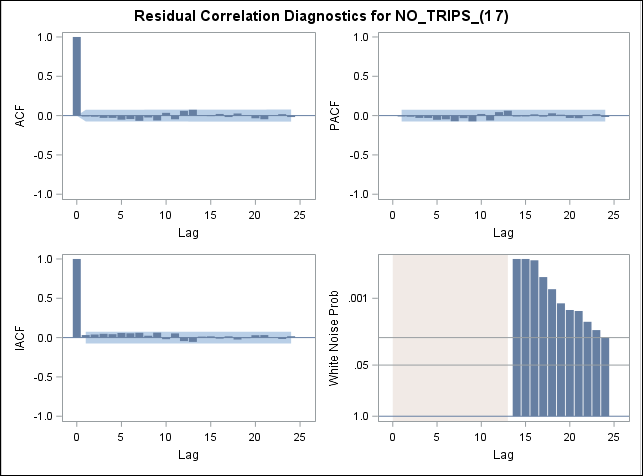
The code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

IDENTIFY VAR=NO\_TRIPS\_(**1**,**7**) NLAG=**30**;

**RUN**;

ESTIMATE P=**6** Q=**7**;**RUN**;

Application of this prewhitening model has removed all the auto-correlations from the input series – Number of Trips



**Prewhiten ‘Mean Temperature’:**

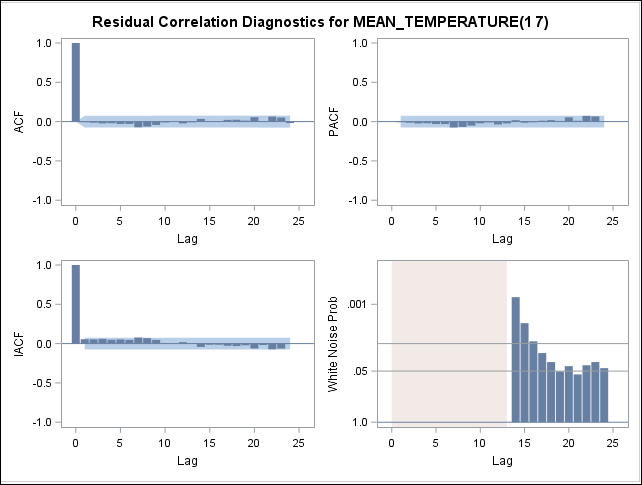
The code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

IDENTIFY VAR=MEAN\_TEMPERATURE(**1**,**7**) ;

**RUN**;

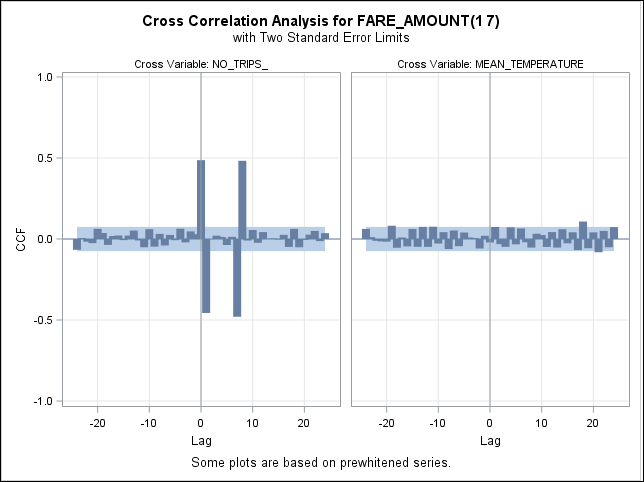
ESTIMATE P=**6** Q=**7**;**RUN**;

The Application of this prewhitening model has removed all the auto-correlations from the input series – Mean Temperature. The ACF Plots and white noise test results are below



**Cross-Correlation Analysis:**

Once the input series are pre-whitened, we perform a cross correlation between the outcome series (Fare Amount) and the input series (No-Trips and Mean Temperature)



As evident from the analysis, there are some significant lags which impact Fare Amount from both the ‘No-Trips’ as well as the ‘Mean Temperature’ time series.

The Lags 0, 1, 7 and 8 for ‘No-Trips’ impact Fare Amount

The Lag 18 for ‘Mean Temperature’ impacts Fare Amount.

**Model Construction:**

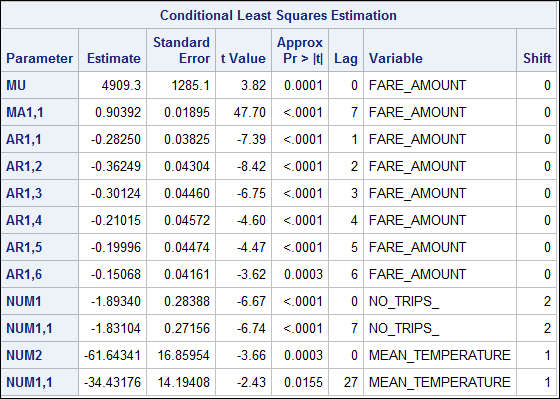
Using the above results, we built a number of models using the transfer function models. The results of the final model are documented below.

**Transfer Function model for Fare-Amount:**

ESTIMATE P=**6** Q= (**7**) INPUT=( **2** $ (**7**)/ NO\_TRIPS\_ **1** $ (**27**)/ MEAN\_TEMPERATURE);**RUN**;

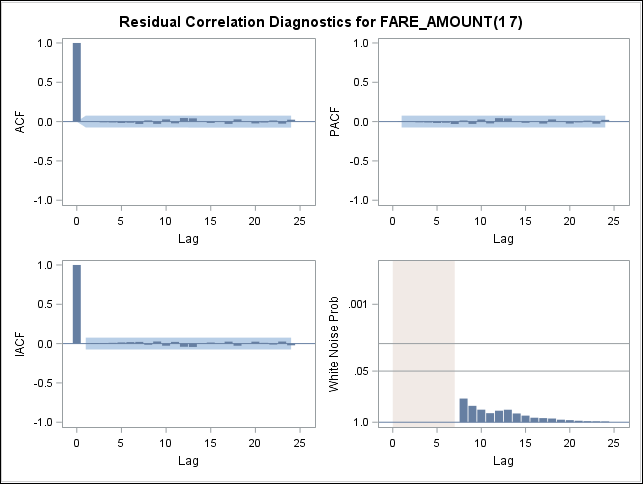
**Model Definition**:

Parameter Estimates



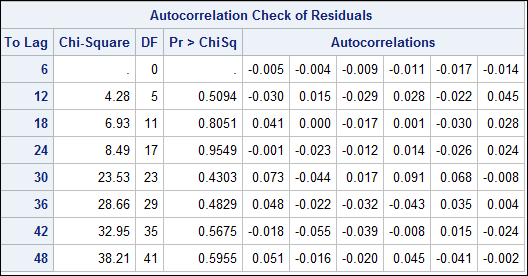
As observed from the p-values of the parameter estimates, all the chosen parameters are statistically significantly different than zero

ACF Plots and White Noise test:

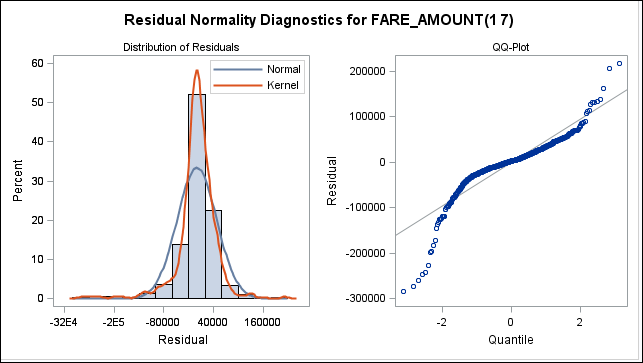


As observed, there is no autocorrelation among various lags in the forecasted time series residuals after applying this model.

The White noise probability test also shows insignificant results, showing that the corresponding residuals obtained after obtaining this model resembles white-noise and there is no more information that we can extract from the residuals.



Also, as evident in the above table, there are no autocorrelations in any lags (as high as lag-48) within the residuals



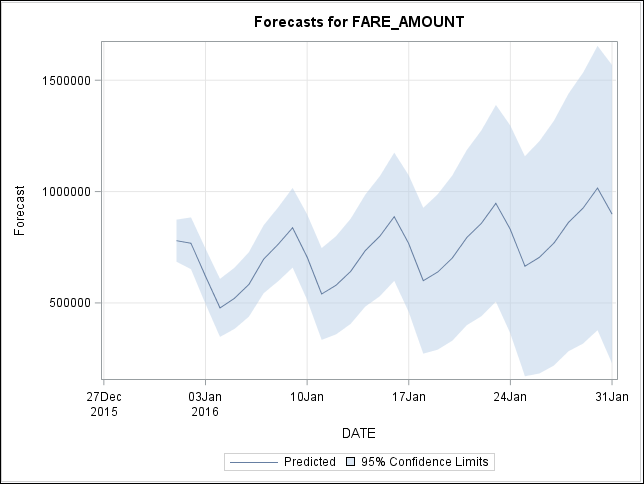
The residuals are normally distributed

**Model Definition:**





**Forecasted Values within 95% confidence interval:**



1. **Forecasting Fare Amount for Overall data for green taxi without uber data observations**

Data Range: 01-Jan-2014 to 31-Dec-2015

Period: Daily

**Prewhitening and CCF (Cross Correlation Function) analysis**

In order to estimate the CCF between the outcome series (Number of Trips) with a set of input/regressor series, we will need to prewhiten the input series, before applying the cross correlation function.

**Prewhiten ‘Fare Amount’:**

The code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

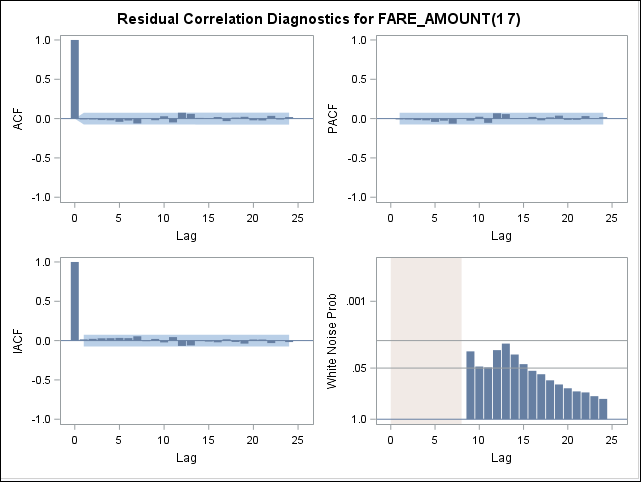
\*PRE-WHITENING FARE\_AMOUNT;

IDENTIFY VAR=FARE\_AMOUNT(**1**,**7**) ;

**RUN**;

ESTIMATE P=**6** Q= (**3**,**7**);

Application of this prewhitening model has removed all the auto-correlations from the input series – Number of Trips



**Prewhiten ‘Mean Temperature’, ‘Passenger Count’ and ‘Trip Distance’:**

The code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

\*PRE-WHITENING MEAN TEMPERATURE;

IDENTIFY VAR=MEAN\_TEMPERATURE(**1**,**7**) ;

**RUN**;

ESTIMATE P=**6** Q=**7**; **RUN**;

\*PRE-WHITENING PASSENGER\_COUNT;

IDENTIFY VAR=PASSENGER\_COUNT (**1**,**7**) ;

**RUN**;

ESTIMATE P=**6** Q= (**3**,**5**,**6**,**7**);

**RUN**;

\*PRE-WHITENING TRIP-DISTANCE;

IDENTIFY VAR=TRIP\_DISTANCE(**1**,**7**) ;

**RUN**;

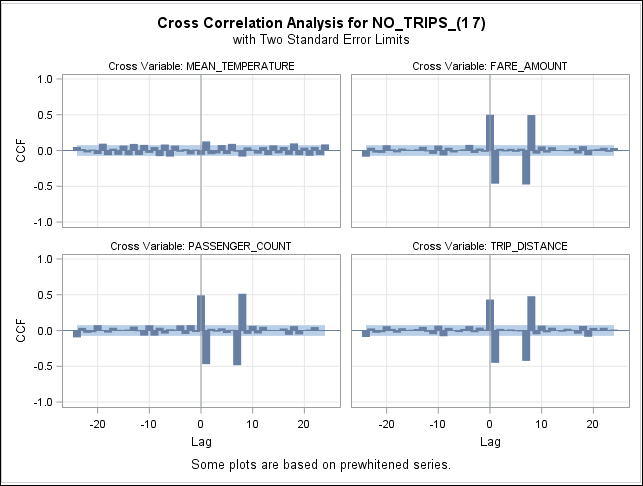
ESTIMATE P=**7** Q=(**3**,**6**,**7**);

**RUN**;

The Application of this prewhitening model has removed all the auto-correlations from the input series – Mean Temperature.

**Cross-Correlation Analysis:**

Once the input series are pre-whitened, we perform a cross correlation between the outcome series (No.Trips) and the input series (Fare Amount, Mean Temperature, Passenger Count and Trip Distance)



As evident from the analysis, there are some significant lags which impact Fare Amount from both the ‘No-Trips’ as well as the ‘Mean Temperature’ time series.

The Lags 0, 1, 7 and 8 for ‘Fare Amount’ impact ‘Number of Trips’

The Lag 1 for ‘Mean Temperature’ impacts ‘Number of Trips’.

The Lags 0, 1, 7 and 8 for ‘Passenger count’ impact ‘Number of Trips’

The Lags 0, 1, 7 and 8 for ‘Trip Distance’ impact ‘Number of Trips’

**Model Construction:**

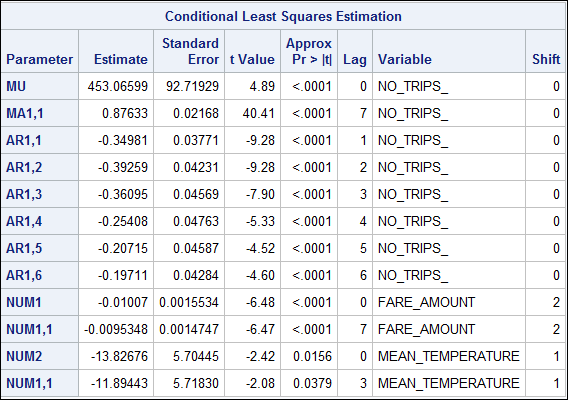
Using the above results, we built a number of models using the transfer function models. The results of the final model are documented below.

**Transfer Function model for ‘Number of Trips’:**

ESTIMATE P=**6** Q= (**7**) INPUT=( **2** $ (**7**)/ FARE\_AMOUNT **1** $ (**3**)/ MEAN\_TEMPERATURE);**RUN**;

**Model Definition:**

Parameter Estimates



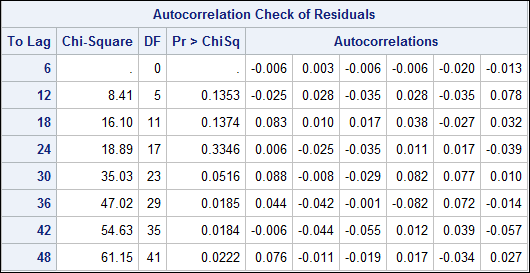
As observed from the p-values of the parameter estimates, all the chosen parameters are statistically significantly different than zero

ACF Plots and White Noise test:



As observed, there is no autocorrelation among various lags in the forecasted time series residuals after applying this model.

The White noise probability test also shows insignificant results, showing that the corresponding residuals obtained after obtaining this model resembles white-noise and there is no more information that we can extract from the residuals.



Also, as evident in the above table, there are no autocorrelations in any lags (as high as lag-48) within the residuals



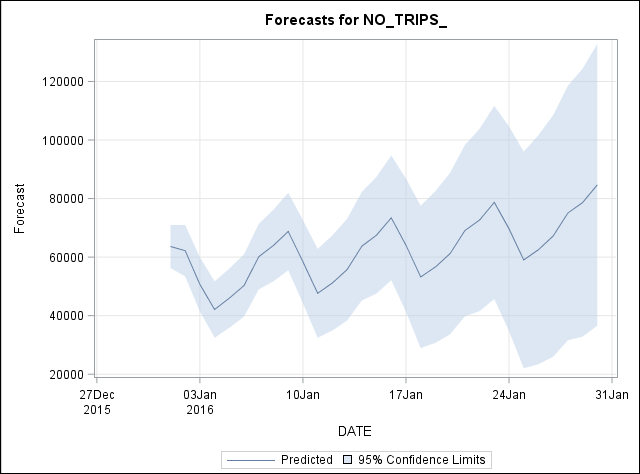
The residuals are normally distributed

**Model Definition:**





**Forecasted Values**:



1. **Forecasting Fare Amount for Overall data for green taxi with Uber data observations**

Data Range: 01-Jan-2015 to 30-June-2015

Period: Daily

**Prewhitening and CCF (Cross Correlation Function) analysis**

In order to estimate the CCF between the outcome series (Fare Amount) with a set of input/regressor series, we will need to prewhiten the input series, before applying the cross correlation function.

**Prewhiten ‘Number of Trips’:**

The code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

\*PREWHITEN NO\_TRIPS;

IDENTIFY VAR=NO\_TRIPS\_(**1**,**7**) ;

**RUN**;

ESTIMATE P=**6** Q=**7**;**RUN**;

Application of this prewhitening model has removed all the auto-correlations from the input series – Number of Trips



**Prewhiten ‘Mean Temperature’, ‘Number of Uber Trips’:**

Similar to the above, the code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

\*PRE-WHITENING MEAN TEMPERATURE;

IDENTIFY VAR=MEAN\_TEMPERATURE(**1**,**7**) ;

**RUN**;

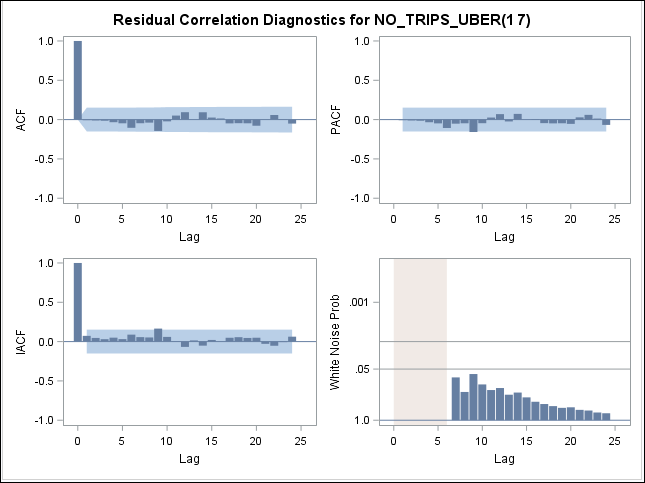
ESTIMATE P=**6** Q=**7**;**RUN**;

\*PRE-WHITENING NO\_UBER\_TRIPS;

IDENTIFY VAR=NO\_TRIPS\_UBER(**1**,**7**) ;

**RUN**;

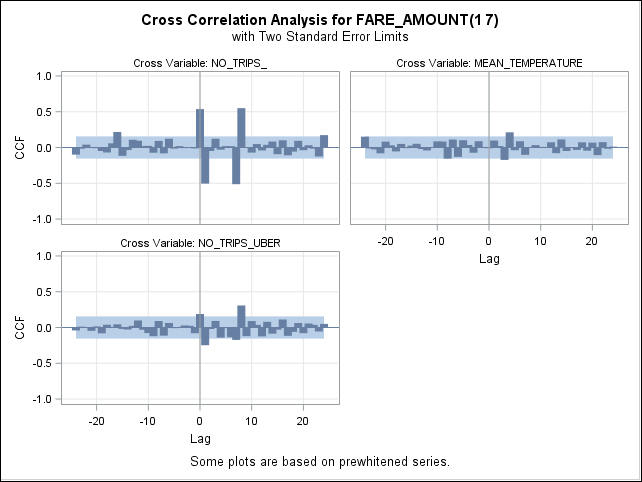
ESTIMATE P=**5** Q=(**7**);**RUN**;



The Application of this prewhitening model has removed all the auto-correlations from the input series – Mean Temperature.

**Cross-Correlation Analysis:**

Once the input series are pre-whitened, we perform a cross correlation between the outcome series (Fare Amount) and the input series (Number of Trips, Mean Temperature and Number of Uber Trips)



As evident from the analysis, there are some significant lags which impact Fare Amount from both the ‘No-Trips’ as well as the ‘Mean Temperature’ time series.

**Model Construction:**

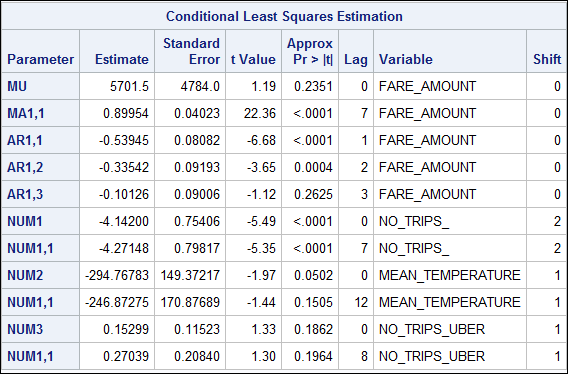
Using the above results, we built a number of models using the transfer function models. The results of the final model are documented below.

**Transfer Function model for ‘Number of Trips’:**

ESTIMATE P=**3** Q= (**7**) INPUT=( **2** $ (**7**)/ NO\_TRIPS\_ **1** $ (**12**)/ MEAN\_TEMPERATURE **1** $ (**8**)/ NO\_TRIPS\_UBER);**RUN**;

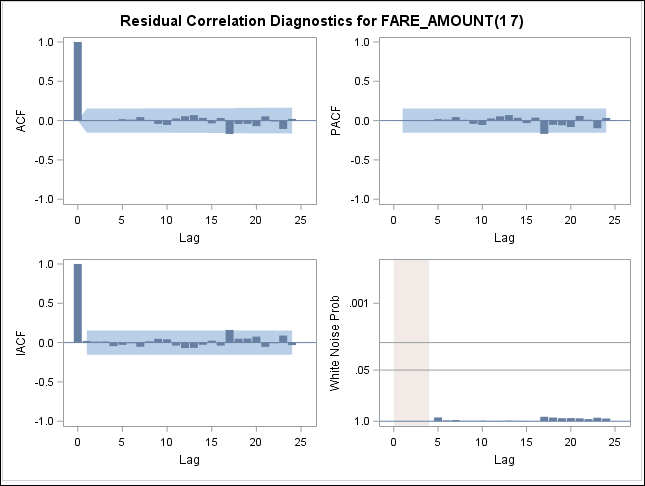
**Model Definition:**

Parameter Estimates



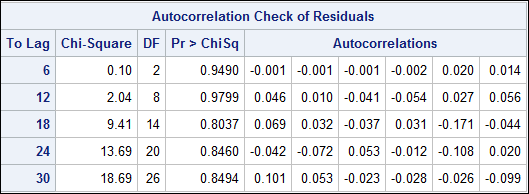
As observed from the p-values of the parameter estimates, all the chosen parameters are statistically significantly different than zero

ACF Plots and White Noise test:

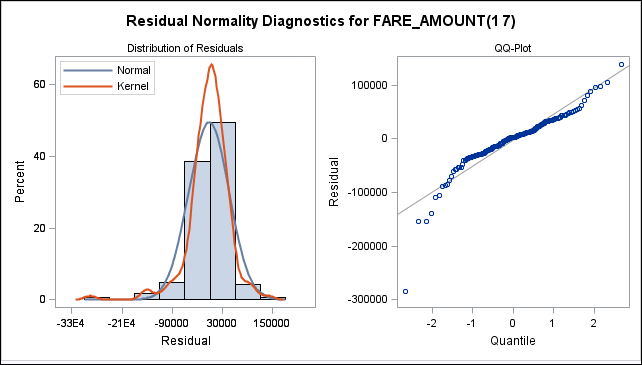


As observed, there is no autocorrelation among various lags in the forecasted time series residuals after applying this model.

The White noise probability test also shows insignificant results, showing that the corresponding residuals obtained after obtaining this model resembles white-noise and there is no more information that we can extract from the residuals.

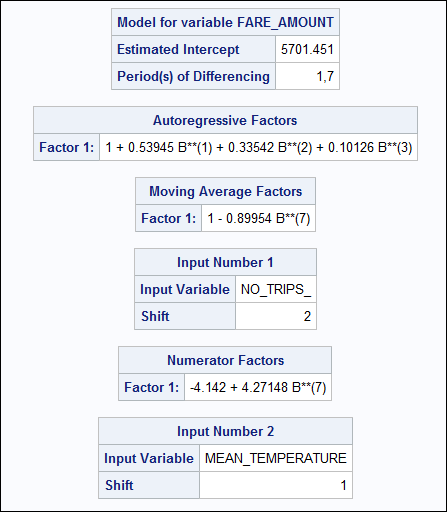


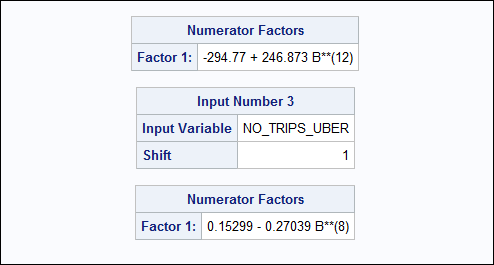
Also, as evident in the above table, there are no autocorrelations in any lags (as high as lag-48) within the residuals



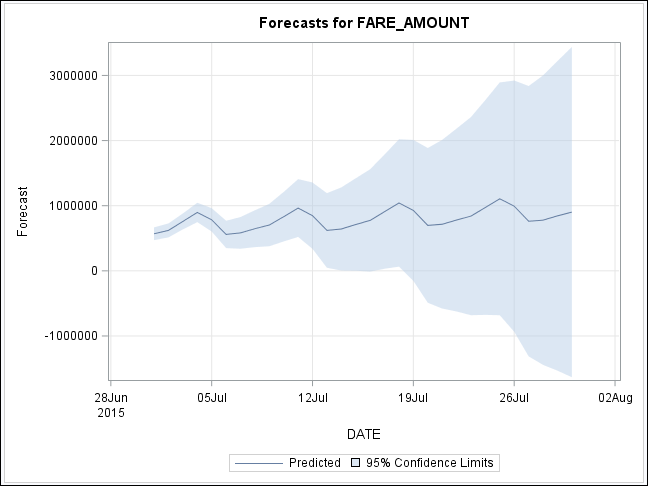
The residuals are normally distributed

**Model Definition:**





**Forecasted Values:**



1. **Forecasting Number of Trips for Overall data for green taxi with Uber data observations**

Data Range: 01-Jan-2015 to 30-June-2015

Period: Daily

**Prewhitening and CCF (Cross Correlation Function) analysis**

In order to estimate the CCF between the outcome series (Number of Trips) with a set of input/regressor series, we will need to prewhiten the input series, before applying the cross correlation function.

**Prewhiten ‘Fare Amount’:**

The code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

\*PRE-WHITENING FARE\_AMOUNT;

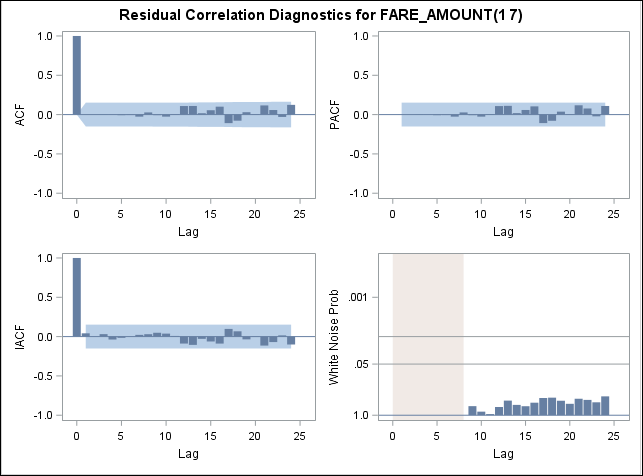
IDENTIFY VAR=FARE\_AMOUNT(**1**,**7**) ;

**RUN**;

ESTIMATE P=**6** Q=(**3**,**7**);

**RUN**;

Application of this prewhitening model has removed all the auto-correlations from the input series – Number of Trips



**Prewhiten ‘Mean Temperature’, ‘Number of Uber Trips’:**

Similar to the above, the code below will perform the prewhitening of ‘No-Trips’ series (within a PROC ARIMA Procedure)

\*PRE-WHITENING MEAN TEMPERATURE;

IDENTIFY VAR=MEAN\_TEMPERATURE(**1**,**7**) ;

**RUN**;

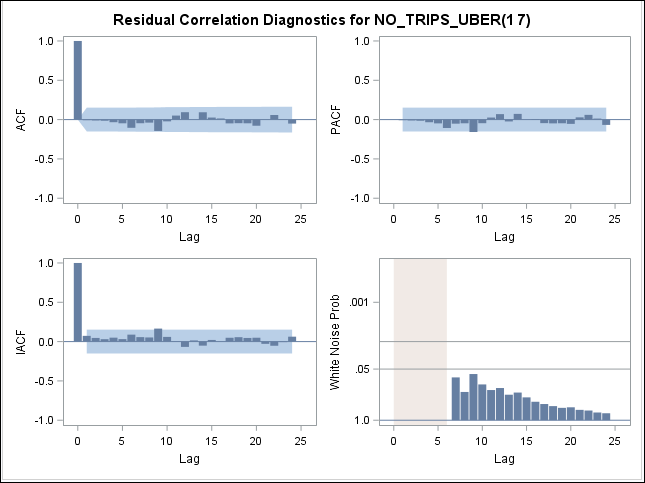
ESTIMATE P=**6** Q=**7**;**RUN**;

\*PRE-WHITENING NO\_UBER\_TRIPS;

IDENTIFY VAR=NO\_TRIPS\_UBER(**1**,**7**) ;

**RUN**;

ESTIMATE P=**5** Q=(**7**);**RUN**;



The Application of this prewhitening model has removed all the auto-correlations from the input series – Mean Temperature.

**Cross-Correlation Analysis:**

Once the input series are pre-whitened, we perform a cross correlation between the outcome series (Number of Trips) and the input series (Fare Amount, Mean Temperature and Number of Uber Trips)



As evident from the analysis, there are some significant lags which impact Fare Amount from both the ‘No-Trips’ as well as the ‘Mean Temperature’ time series.

**Model Construction:**

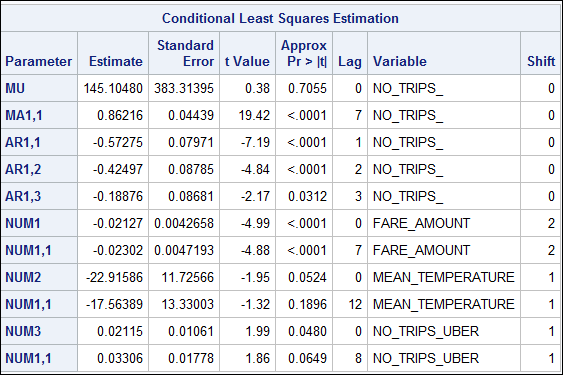
Using the above results, we built a number of models using the transfer function models. The results of the final model are documented below.

**Transfer Function model for ‘Number of Trips’:**

ESTIMATE P=**3** Q= (**7**) INPUT=( **2** $ (**7**)/ FARE\_AMOUNT **1** $ (**12**)/ MEAN\_TEMPERATURE **1** $ (**8**)/ NO\_TRIPS\_UBER);**RUN**;

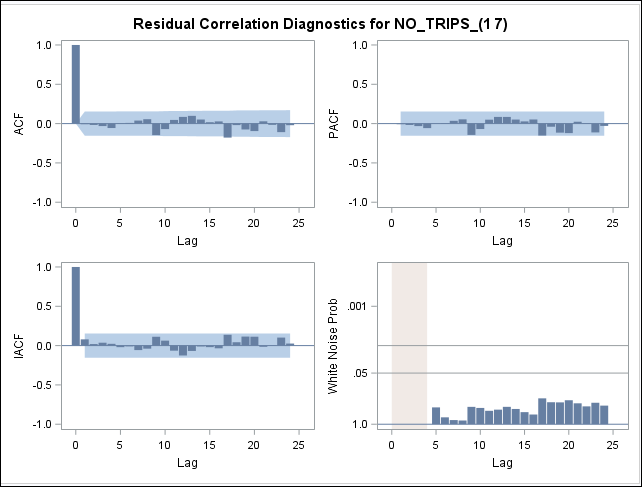
**Model Definition:**

Parameter Estimates



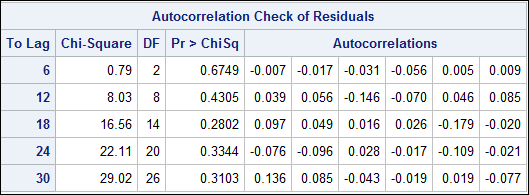
As observed from the p-values of the parameter estimates, all the chosen parameters are statistically significantly different than zero

ACF Plots and White Noise test:

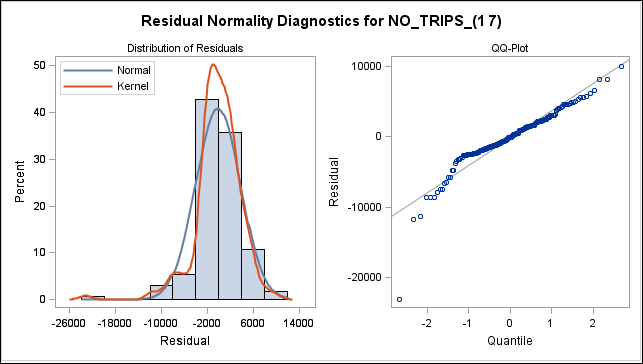


As observed, there is no autocorrelation among various lags in the forecasted time series residuals after applying this model.

The White noise probability test also shows insignificant results, showing that the corresponding residuals obtained after obtaining this model resembles white-noise and there is no more information that we can extract from the residuals.

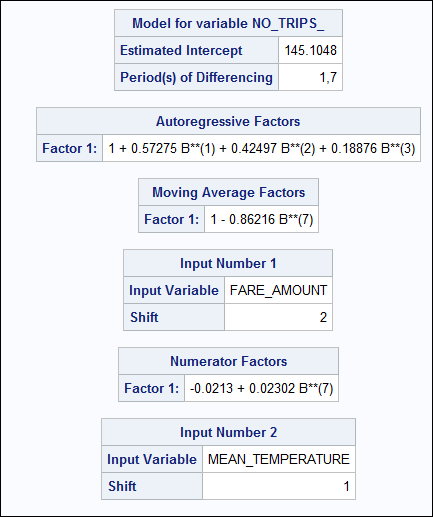


Also, as evident in the above table, there are no autocorrelations in any lags (as high as lag-48) within the residuals

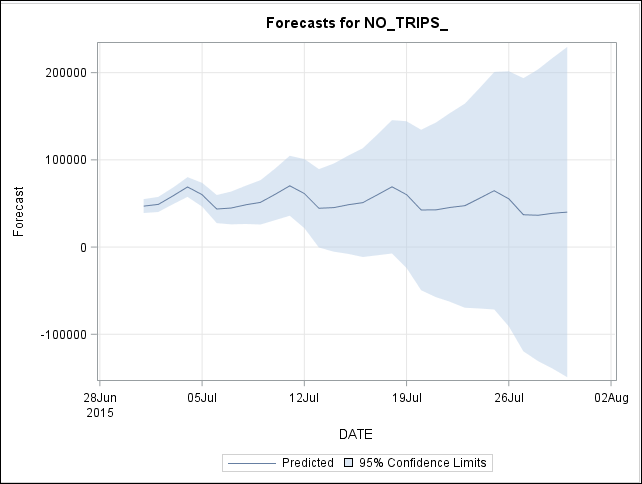


The residuals are normally distributed

**Model Definition:**

**Forecasted Values:**

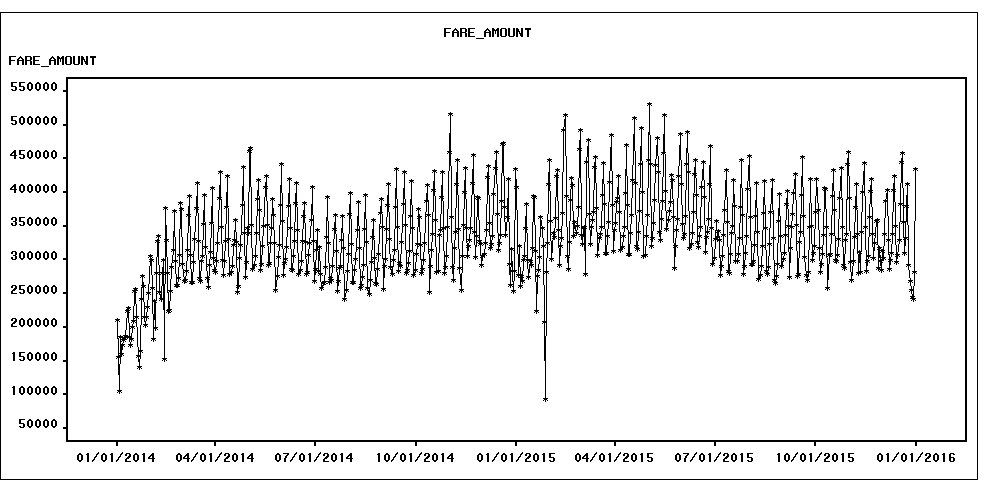


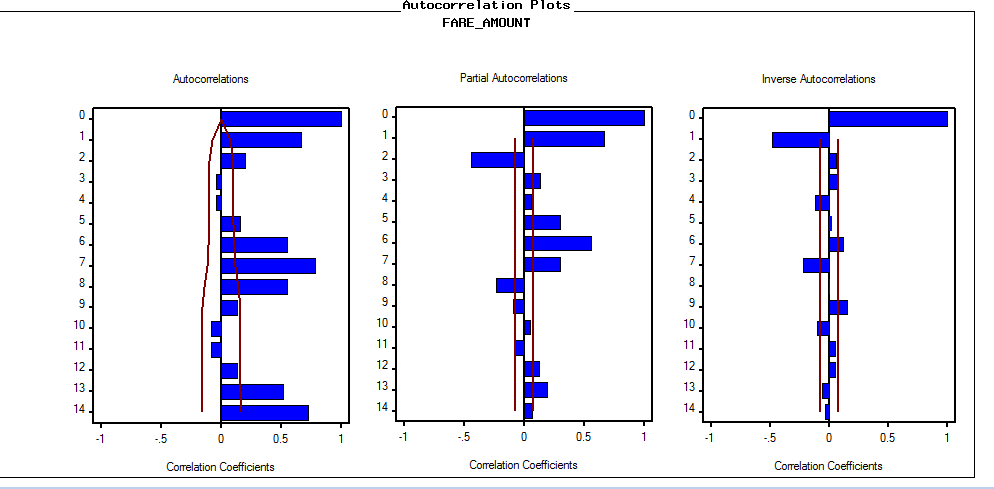
Manhattan Time Series for Green Taxi data

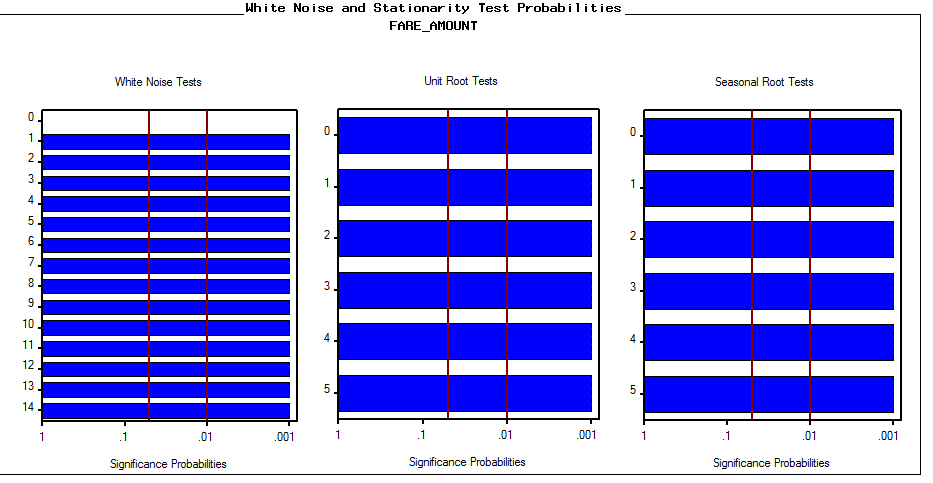
**Without Hold out**

1. **Forecasting Fare Amount for Manhattan Series**

* Initial data exploration

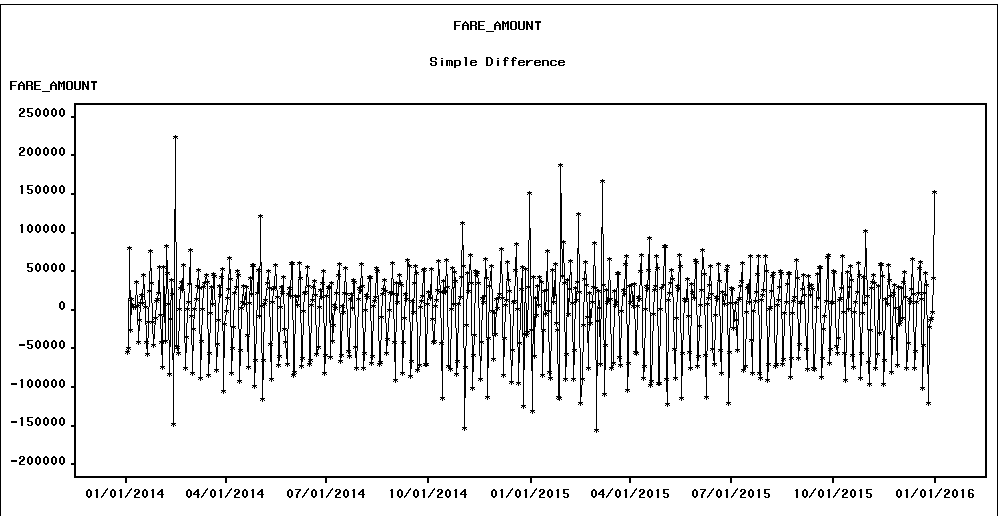


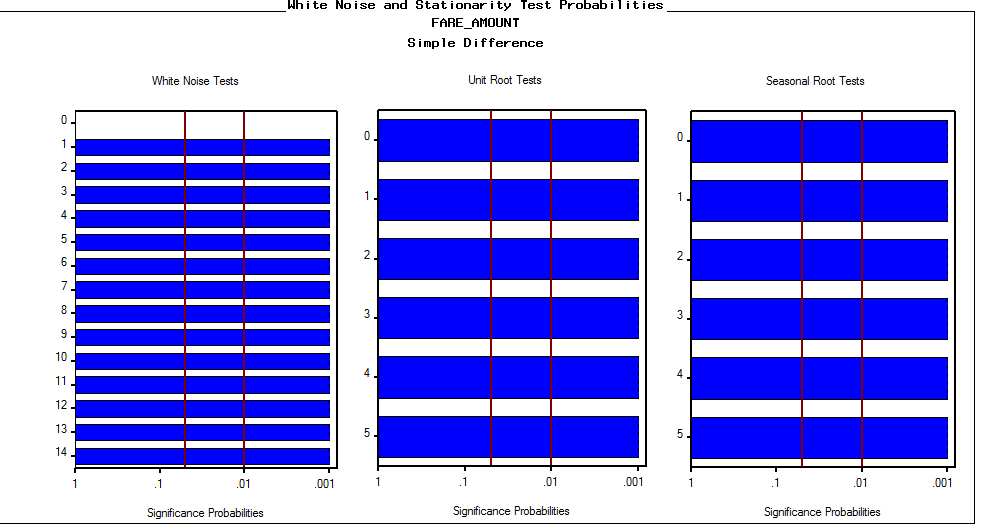




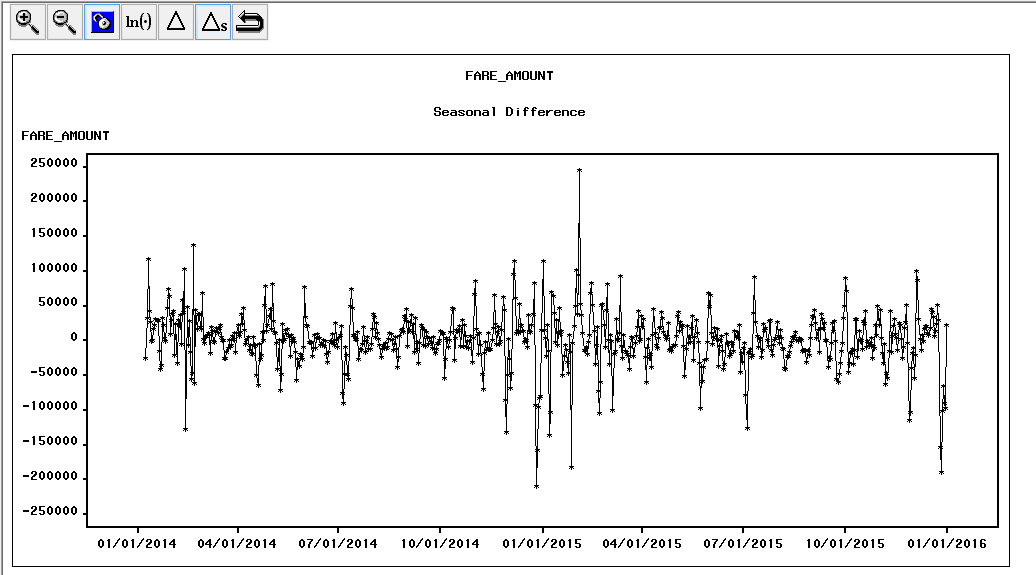
Findings

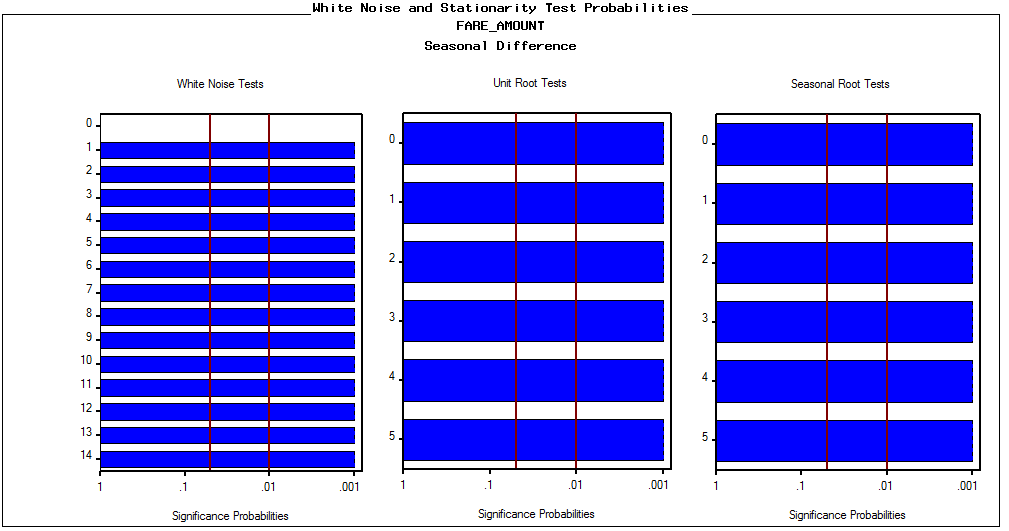
* From ACF plot, series fails the white noise Test. (Null hypothesis that series is a white noise is rejected).
* From the raw graph, there are few point interventions (high and low) that could be potentially useful source of information
  + 1st Dec, 2014 🡪 516,21
  + 27th Jan, 2015 🡪 92,996
  + 14th Feb, 2015 🡪 515,095 **(valentine’s day special** ☺ **)**
  + 2nd May, 2015 🡪 530,994
* Looks like there was a **trend** in the raw series in the **first two quarters** of the year 2014 (**Jan – May**).
* Most importantly, Data is **heavily seasonal.**
* Simple Difference



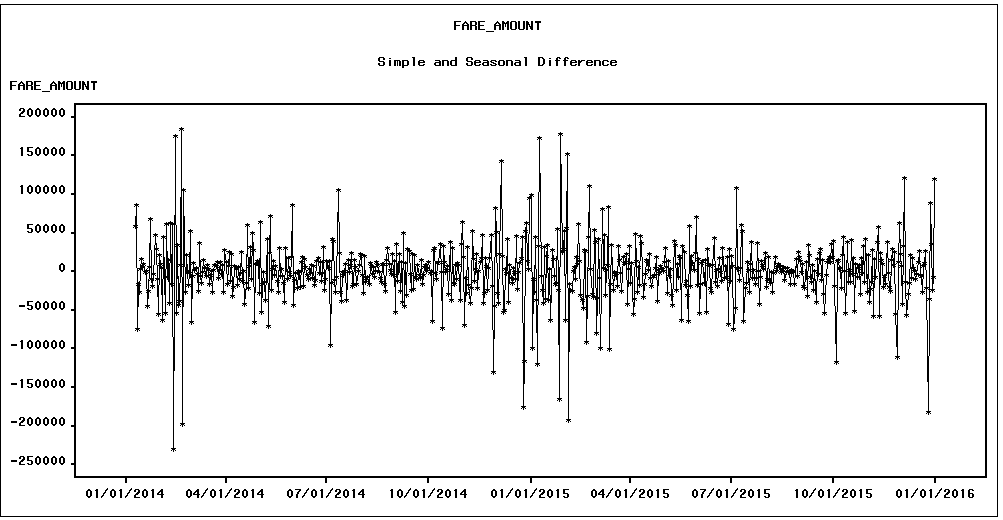


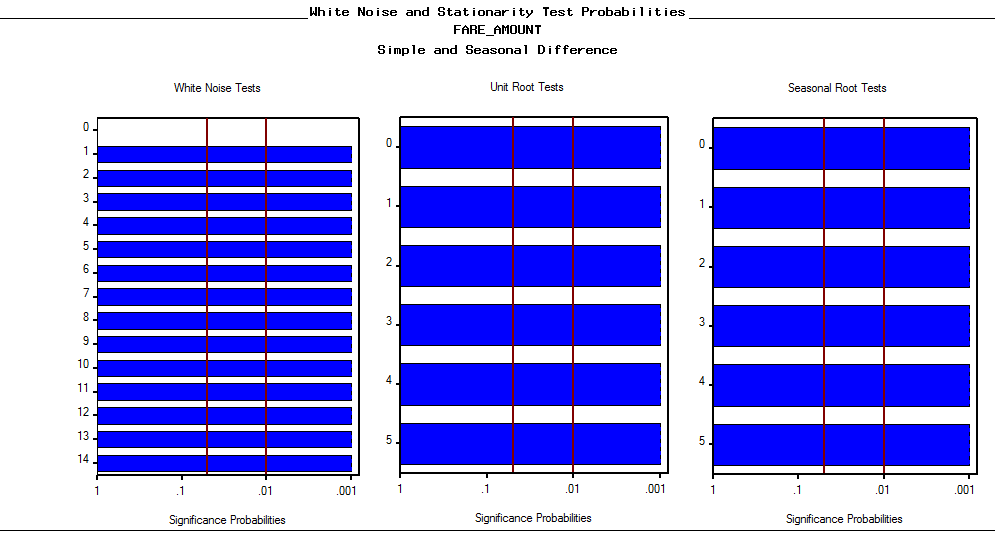
* Seasonal Difference





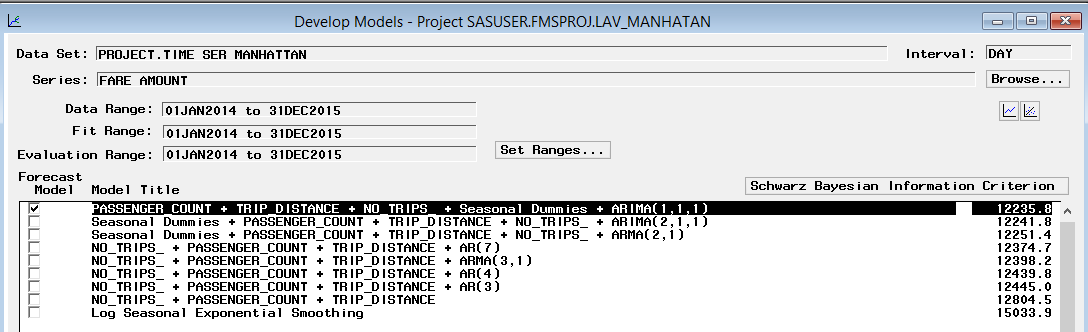
* First and Seasonal Difference:





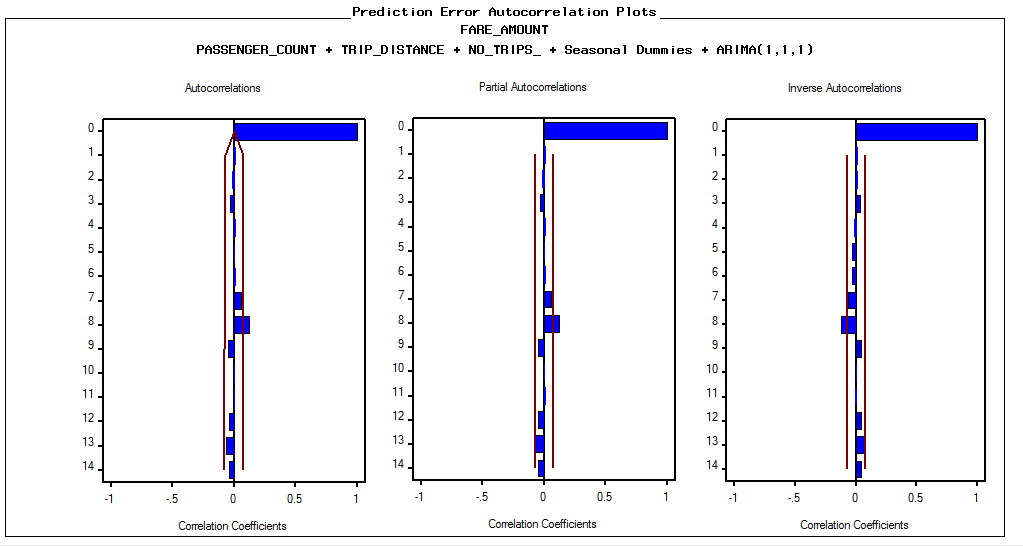
* Modeling Phase

Following Models were built in TSFS (please note that **log Seasonal Exponential Smoothing** was fit automatically).



Regressors + Seasonal Dummies + ARIMA (1, 1, 1) was the best model as per the factors as discussed earlier

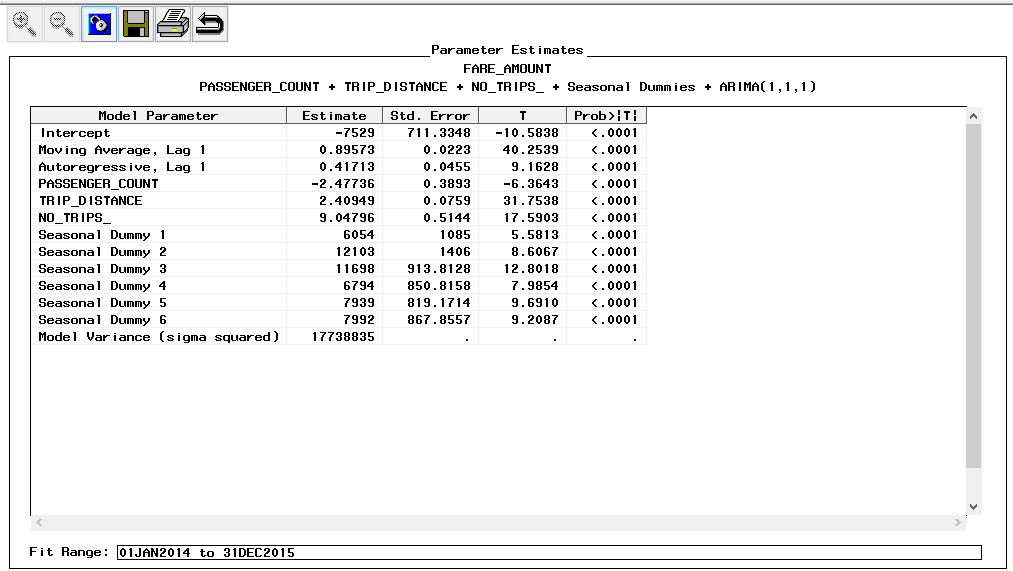
ACF, PACF and IACF plot:



Series is almost stationary after applying the best model so far. At lag 8 and 9, p-value is slightly insignificant in comparison to α-value



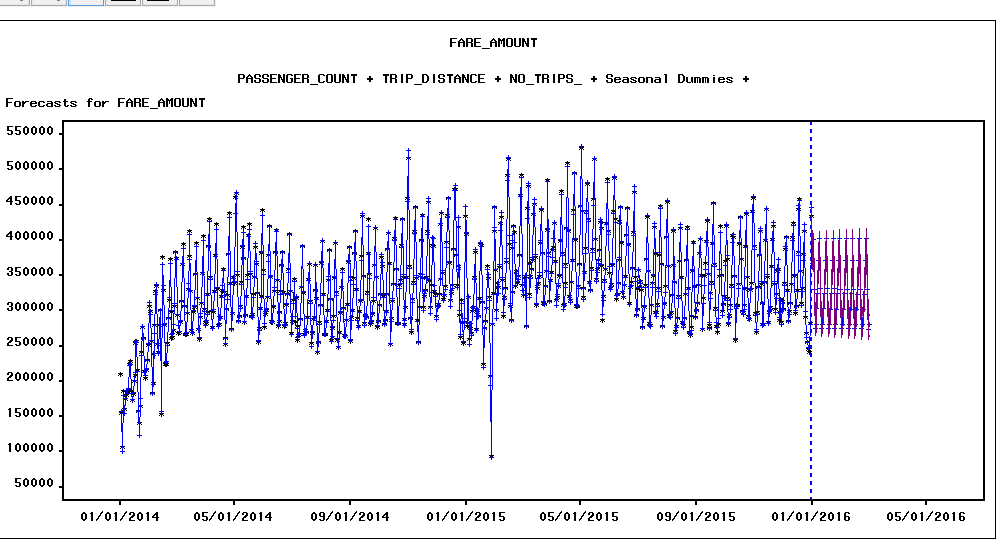
All the parameters are signifcant even at 1% significantly different from zero.



Based on MAPE, model accuracy is around 99.043%



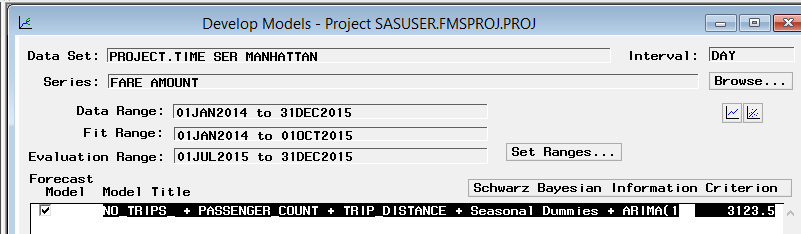
Forecasting is shown below. I have set the Forecast range is from Jan 2016 – March 2016

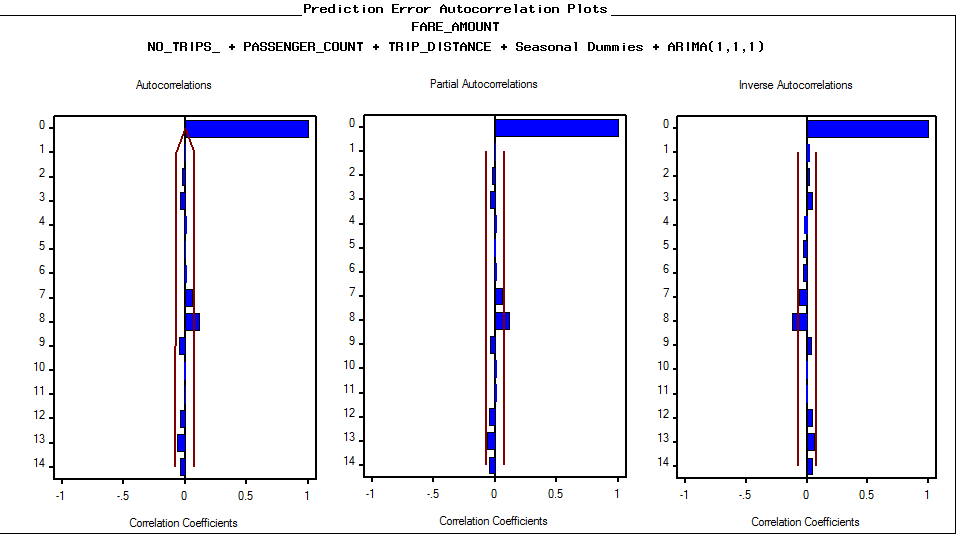


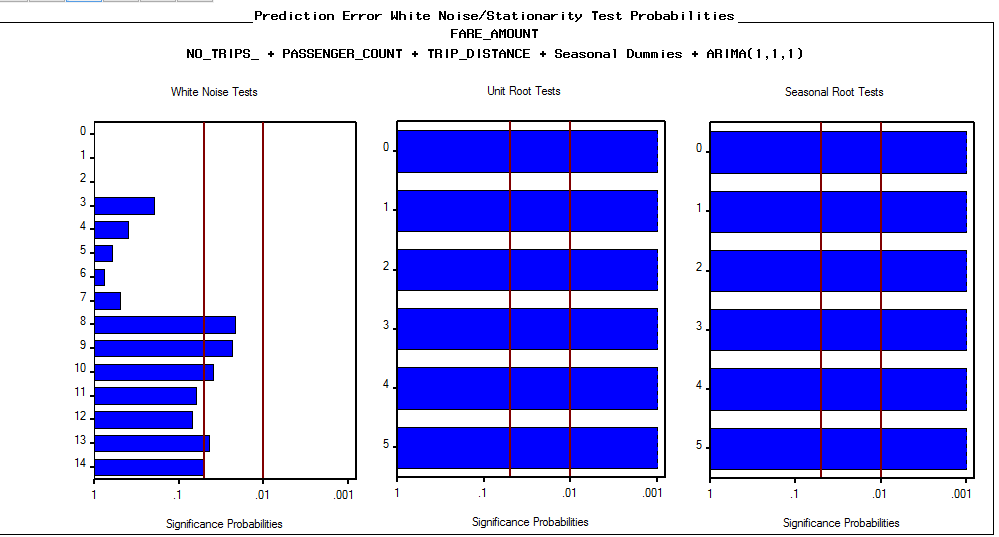
**Observation:**

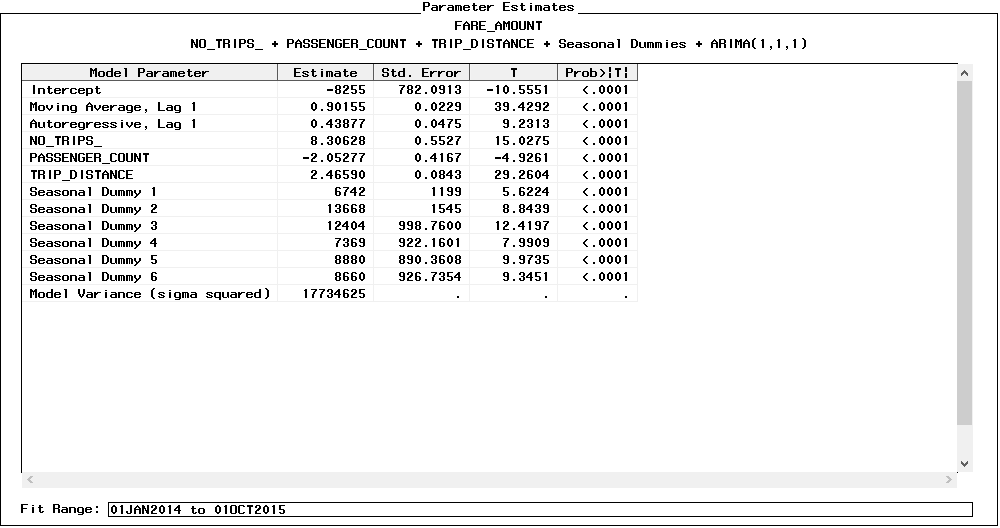
I tried using point interventions mentioned previously in the section Findings from Initial Data Exploration“ but there wasn’t any significant impact on the model statistical criteria (SBIC, RMSE, AIC). On the contrary, statistical criteria were shooting up if using point interventions.

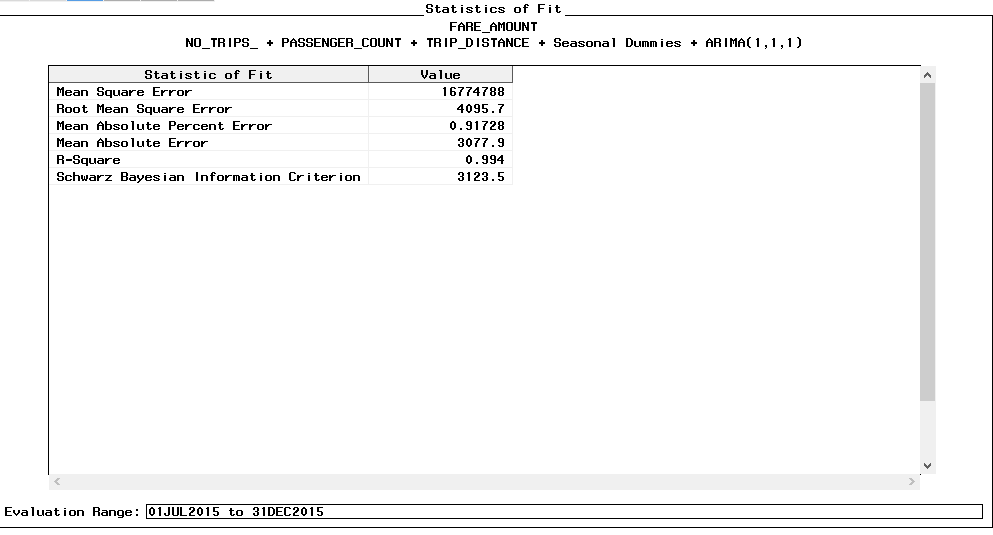
**With holdout sample**











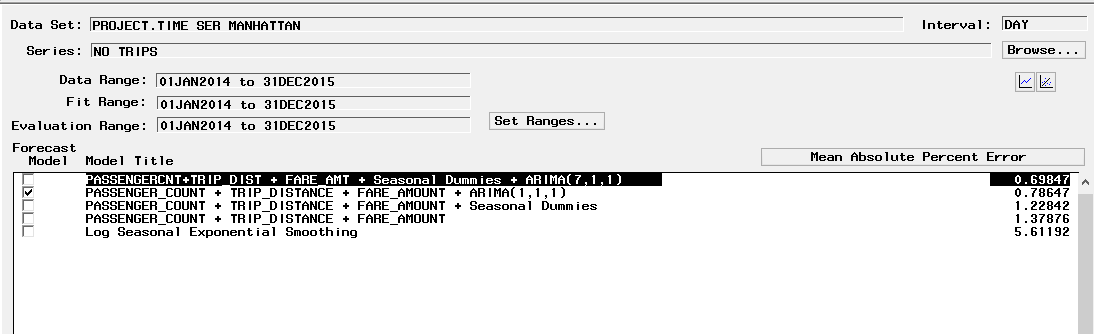
**Inference**: Manhattan series with and without holdout sample is behaving in almost similar fashion using the model- **Regressor + seasonal dummies + ARIMA (1, 1, 1)** and hence there isn’t much difference in the statistical criteria.

1. **Forecasting Number of Trips for Manhattan Series**

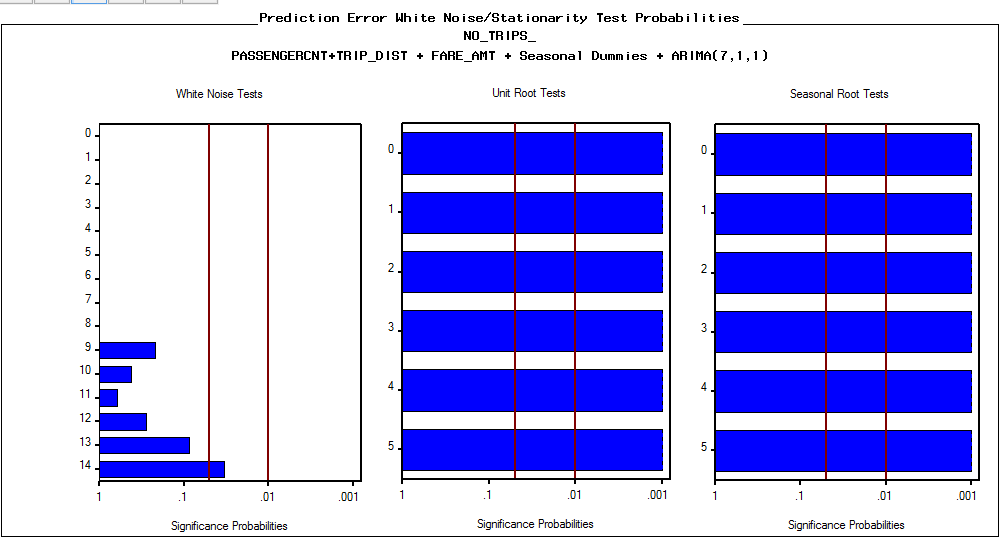
**Without Hold out sample**

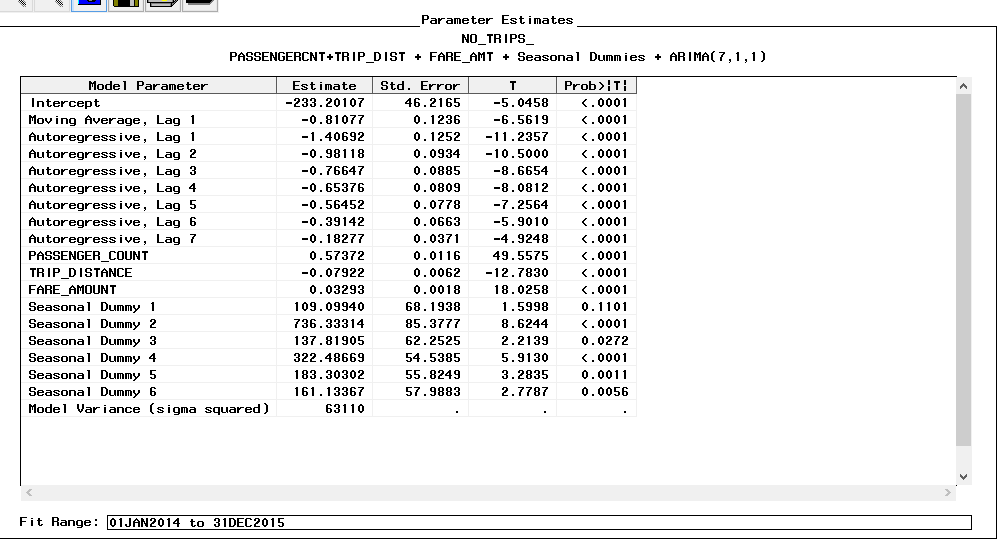
Following screenshot shows the models built and the best model was selected based on the factors as discussed above

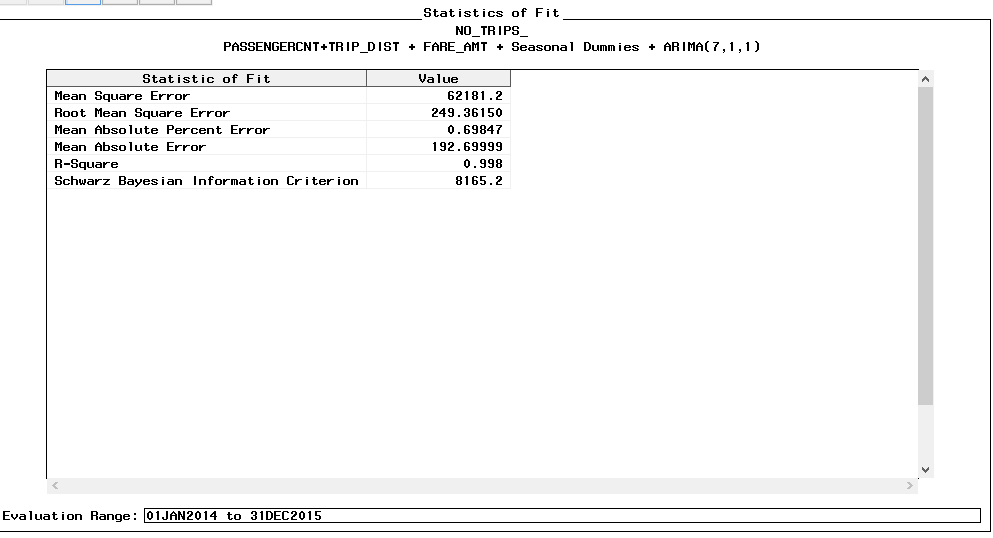
Best Model 🡪 **Passenger Count + Trip Dist + Seasonal dummies +ARIMA (7, 1, 1)**

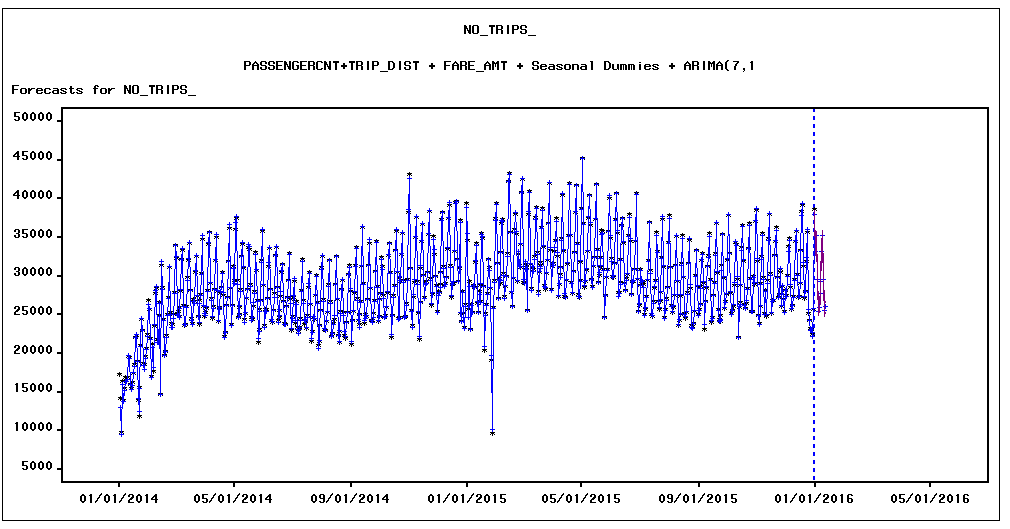






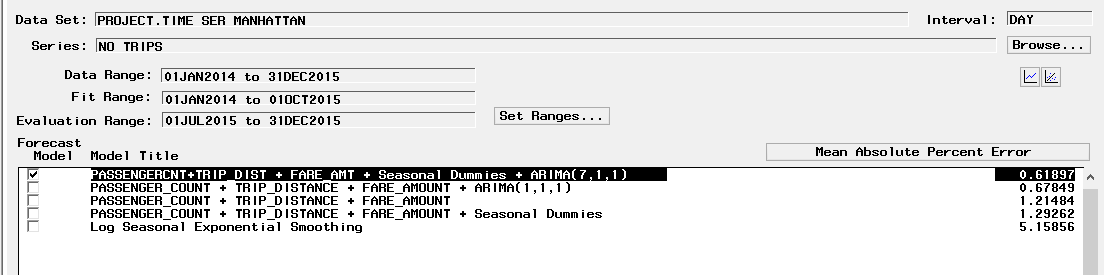


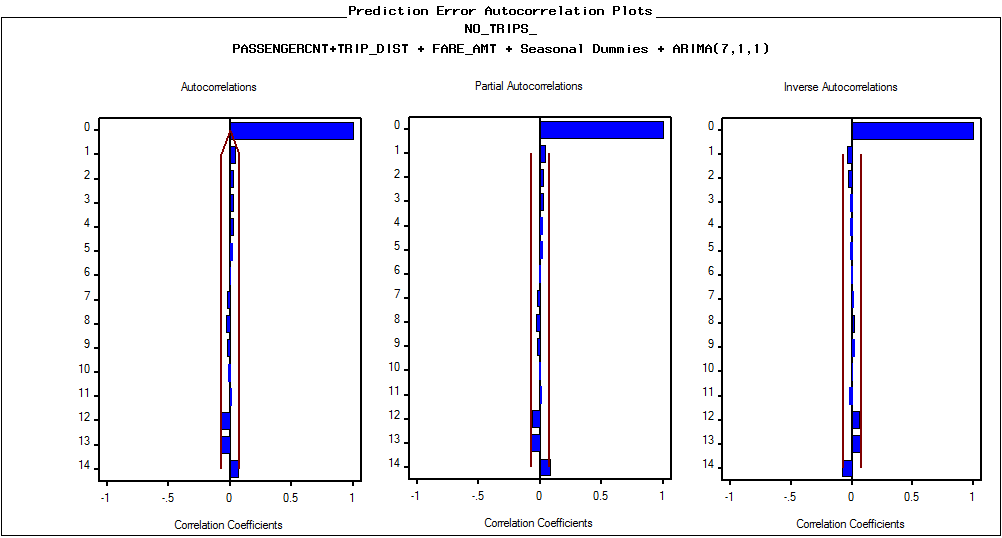


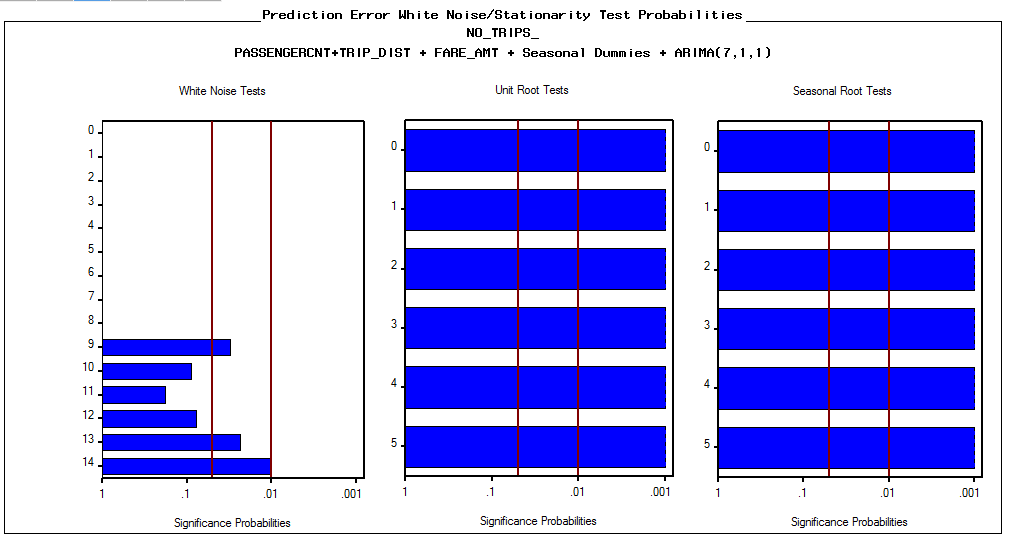


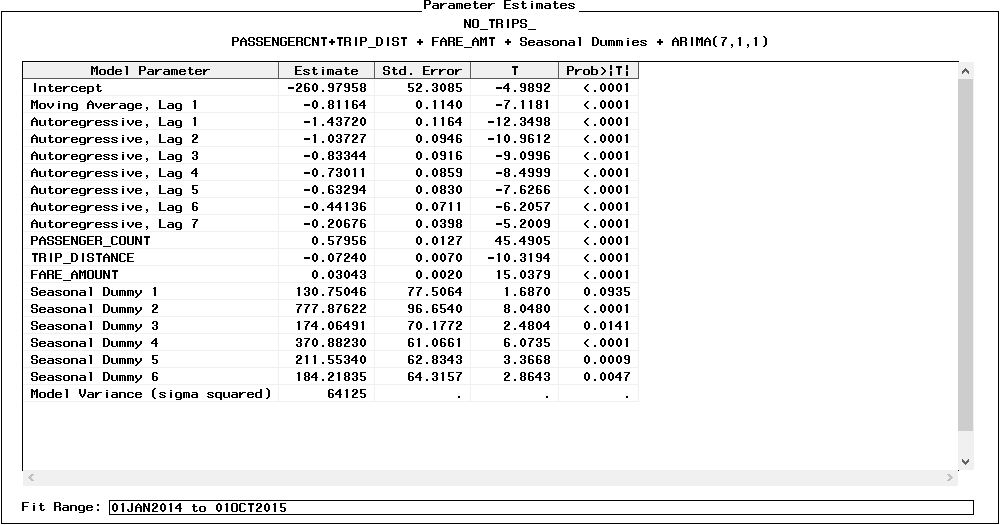
**With holdout sample**

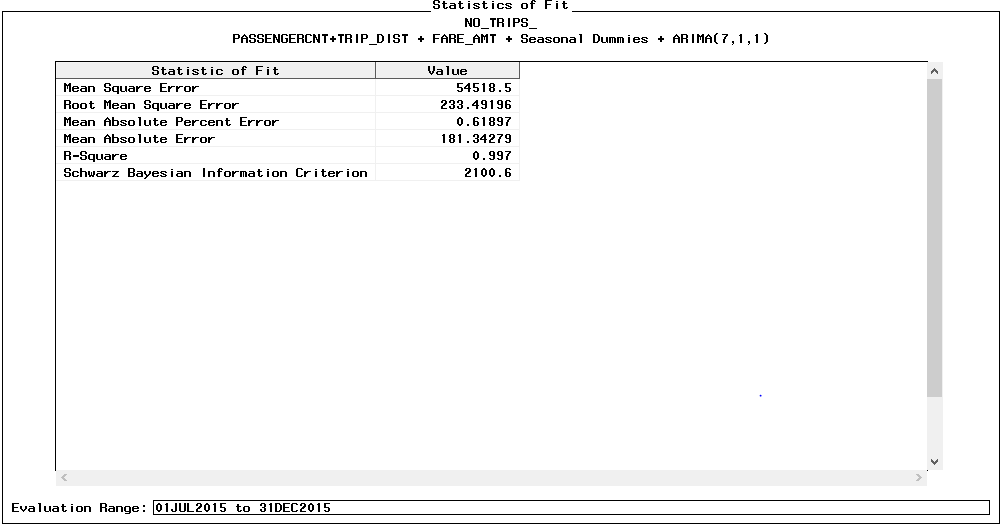
Best Model 🡪 **Passenger Count + Trip Dist + Seasonal dummies +ARIMA (7, 1, 1)**

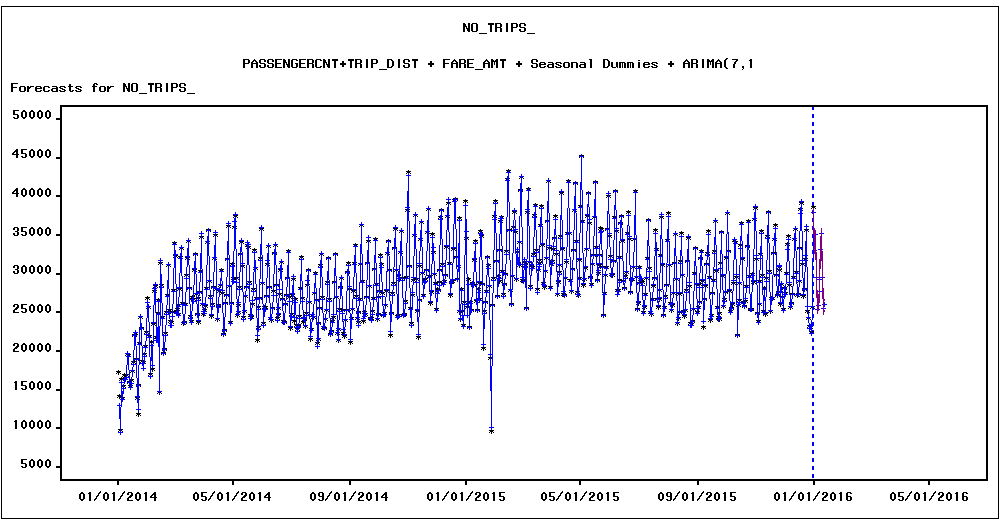










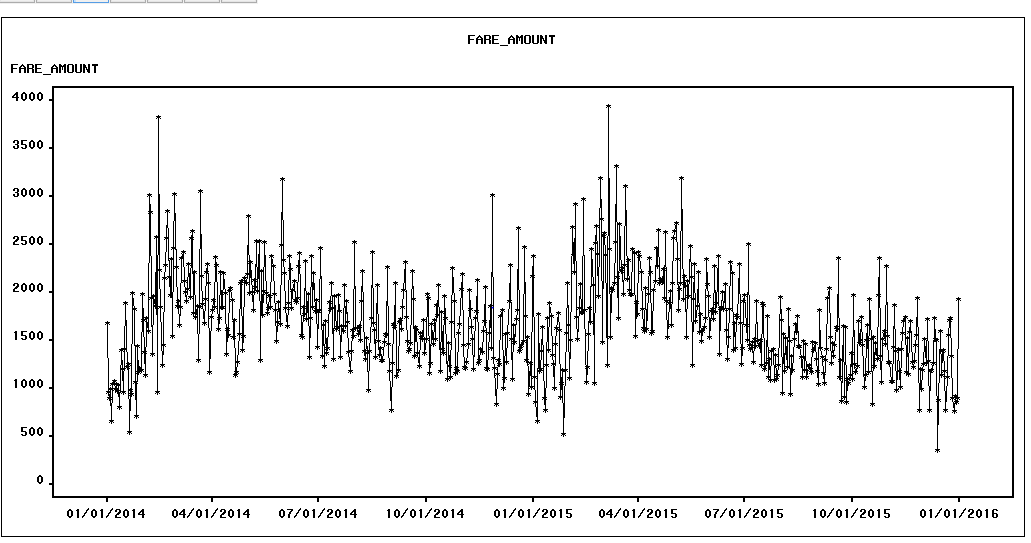


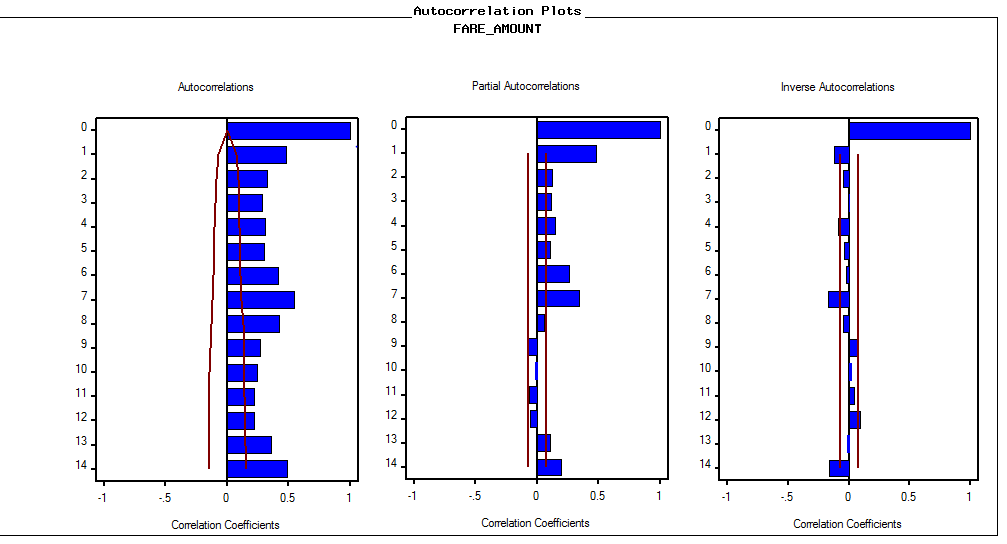
### Bronx Time Series

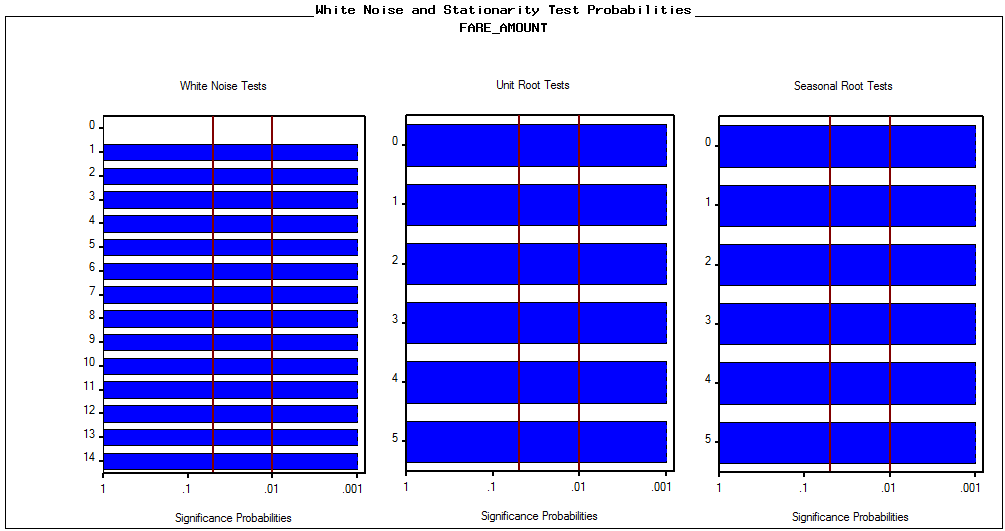
1. **Forecasting Fare Amount for Bronx Time Series**

**WITHOUT HOLD OUT SAMPLE**

* Initial Data exploration



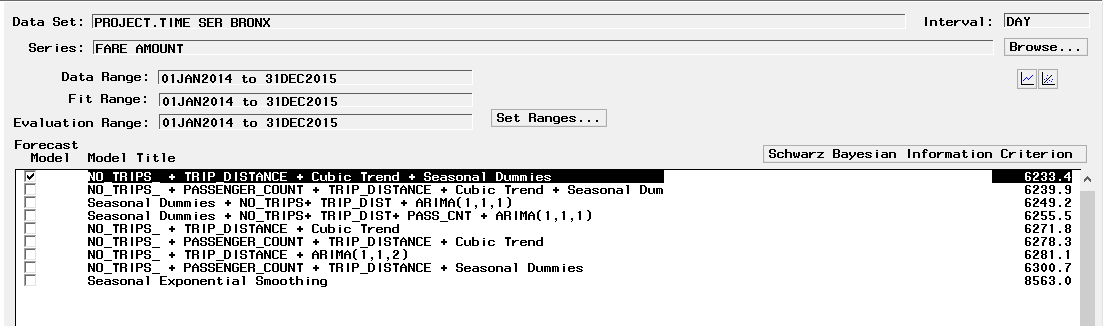


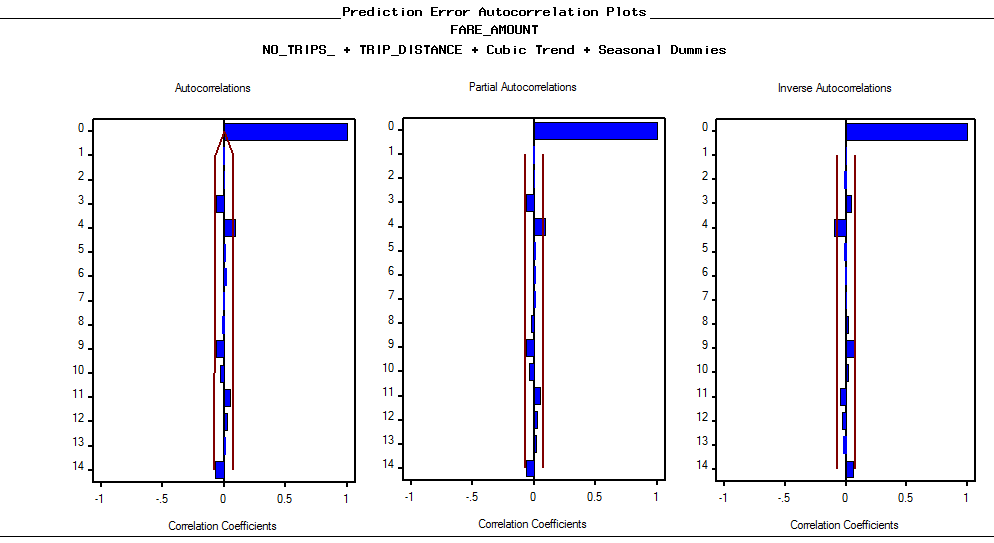


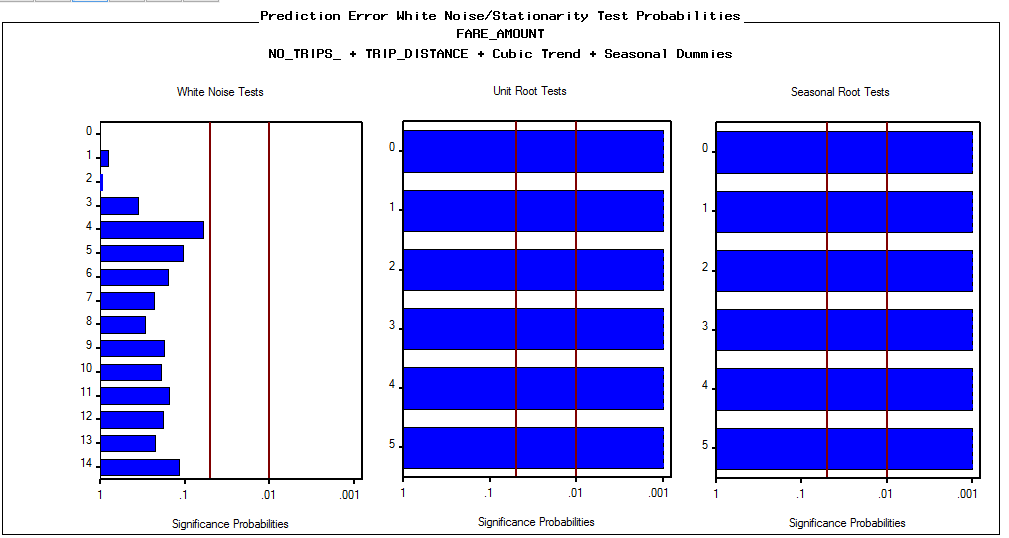
* Modeling

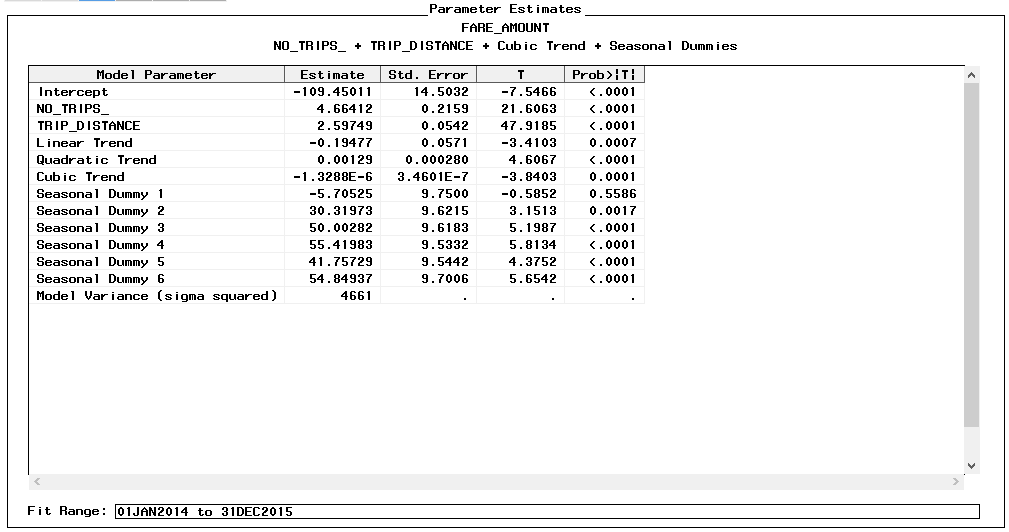
Below screenshots show the models built in TSFS

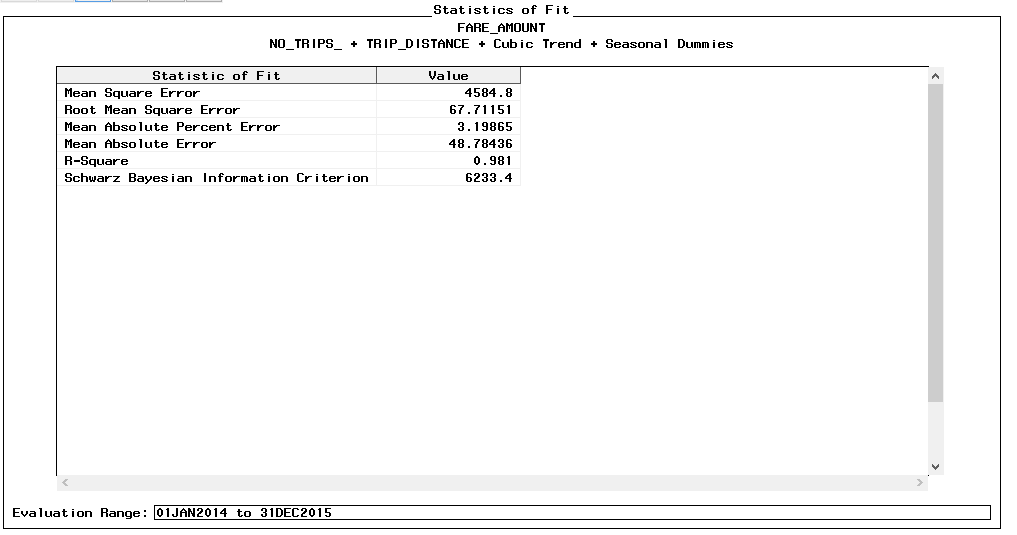
Best model is Regressor with Cubic Trend and Seasonal Dummies as per the factors discussed above.









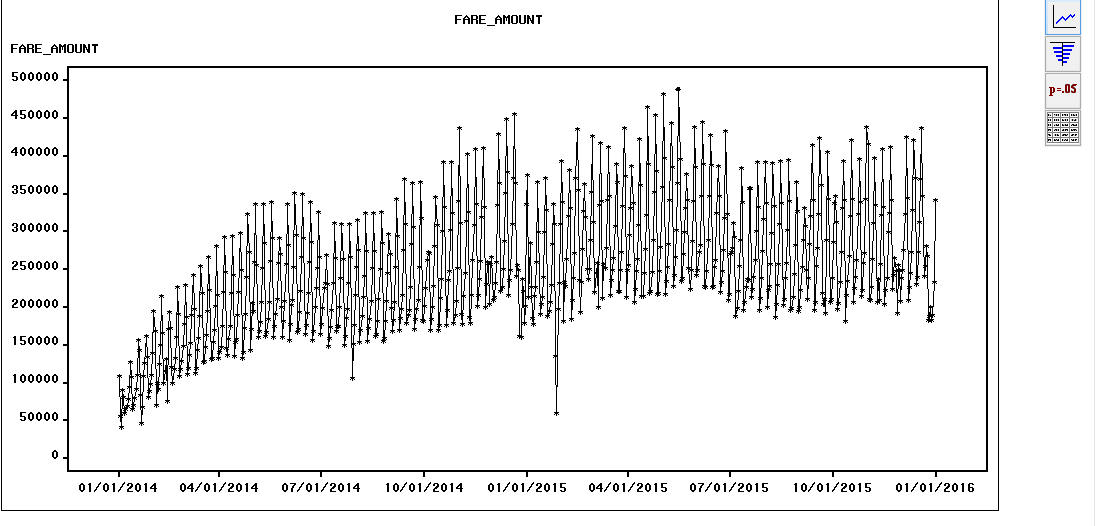


## **Brooklyn Dataset**

**Without Holdout Sample**

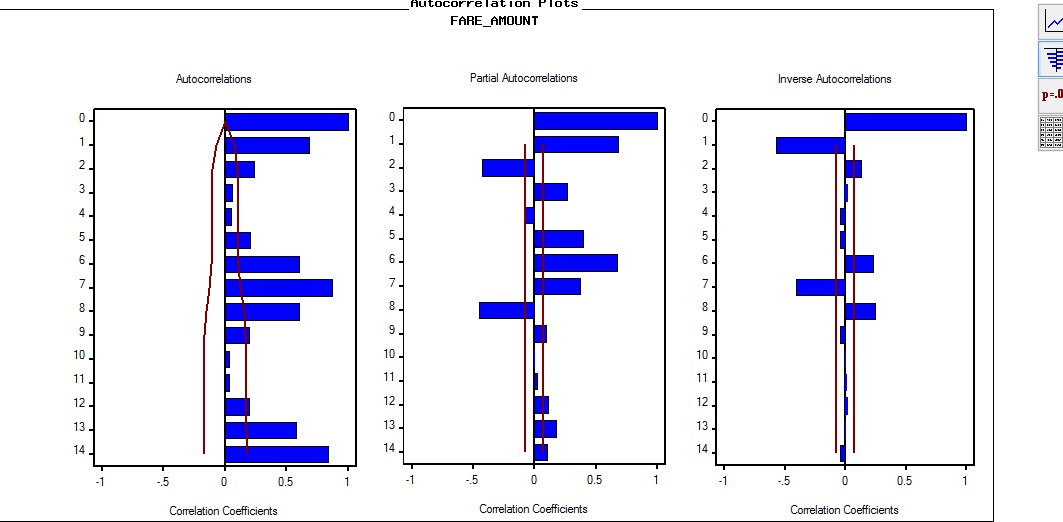
1. **Forecasting Fare Amount for Brooklyn Series**

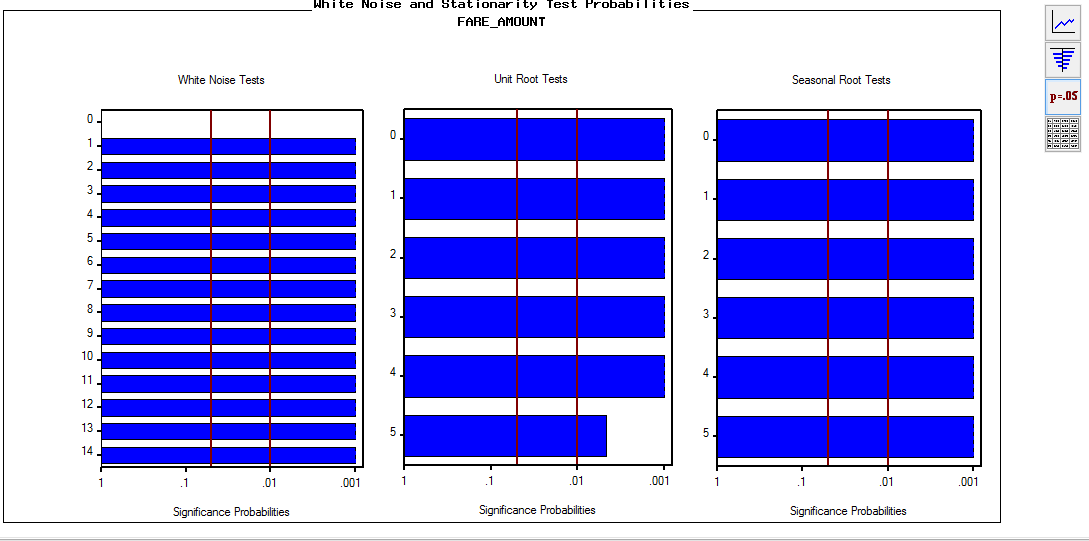
* Initial data exploration



**Findings:**

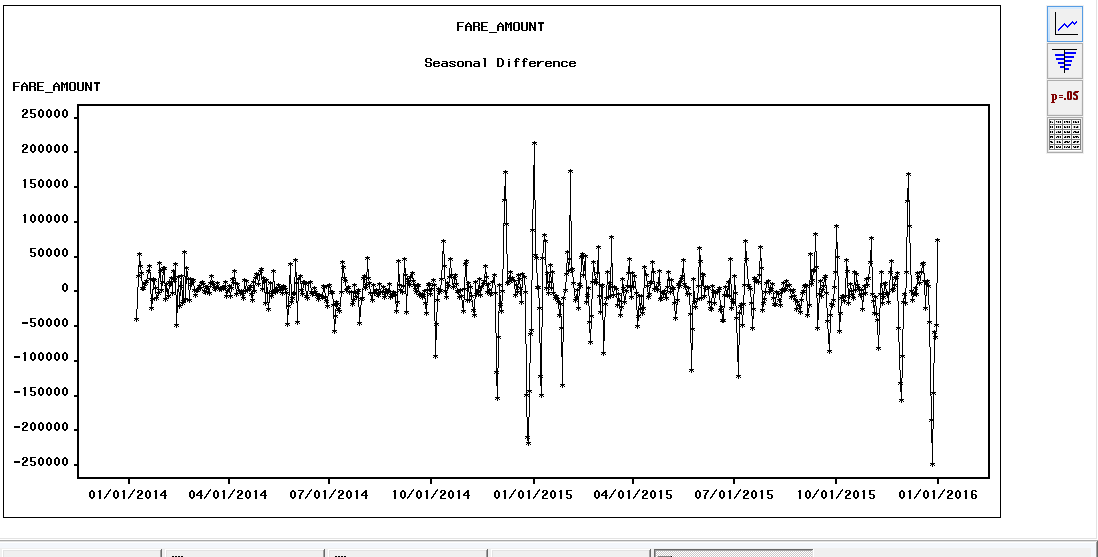
* Initial inferences of raw data sets:
* The data seems to follow an upward trend.
* There are less of single peaks, except one-two.
* Seasonality seems to be an important component here
* White noise is quite significant.

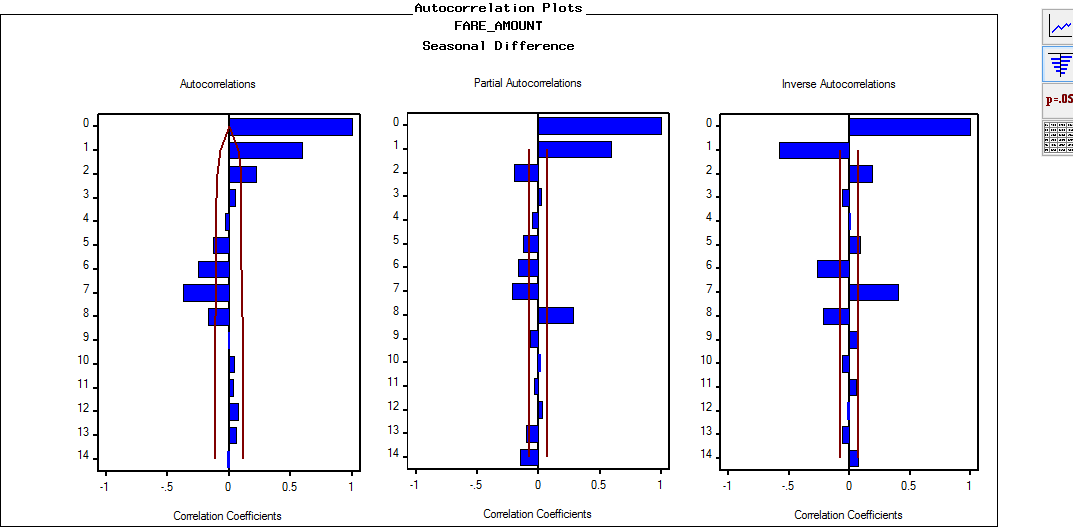


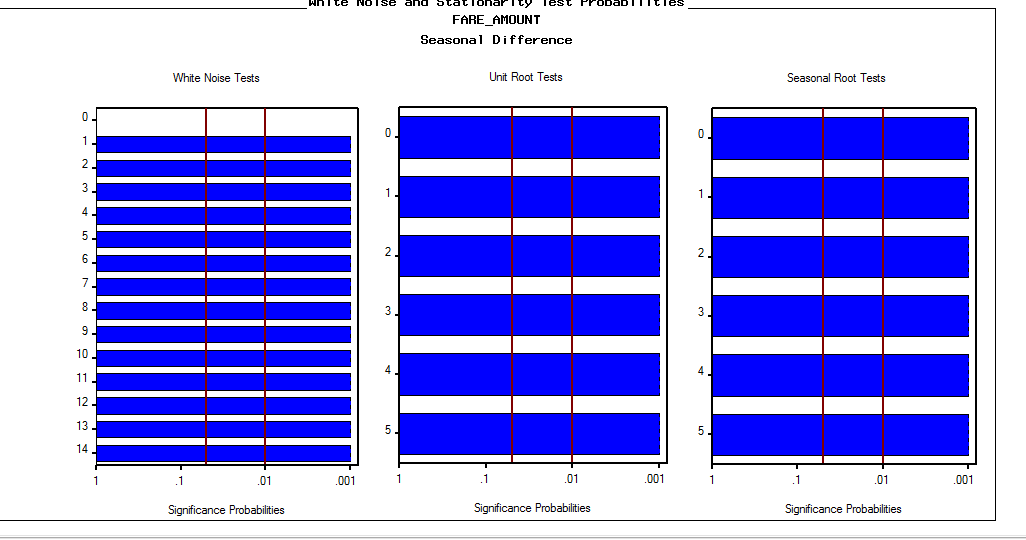


Data set is stationary but fails the white noise test.

**Seasonal Diff:** On applying seasonal difference we see a few prominent peaks but white noise is still quite significant.

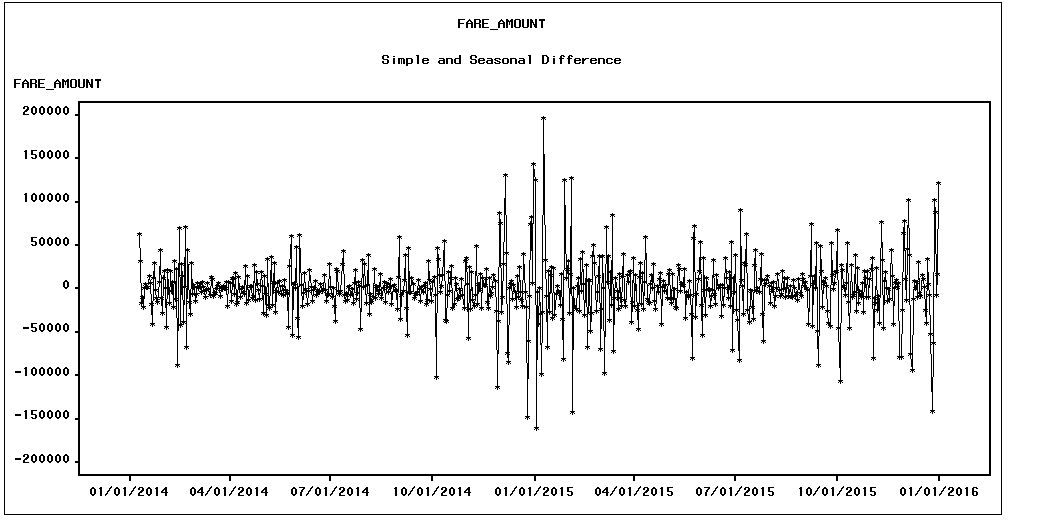


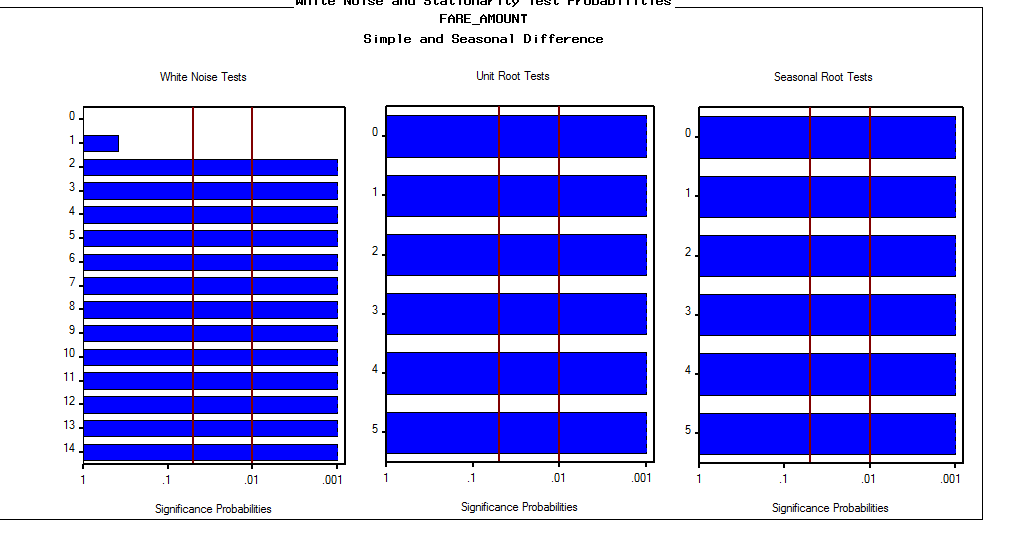




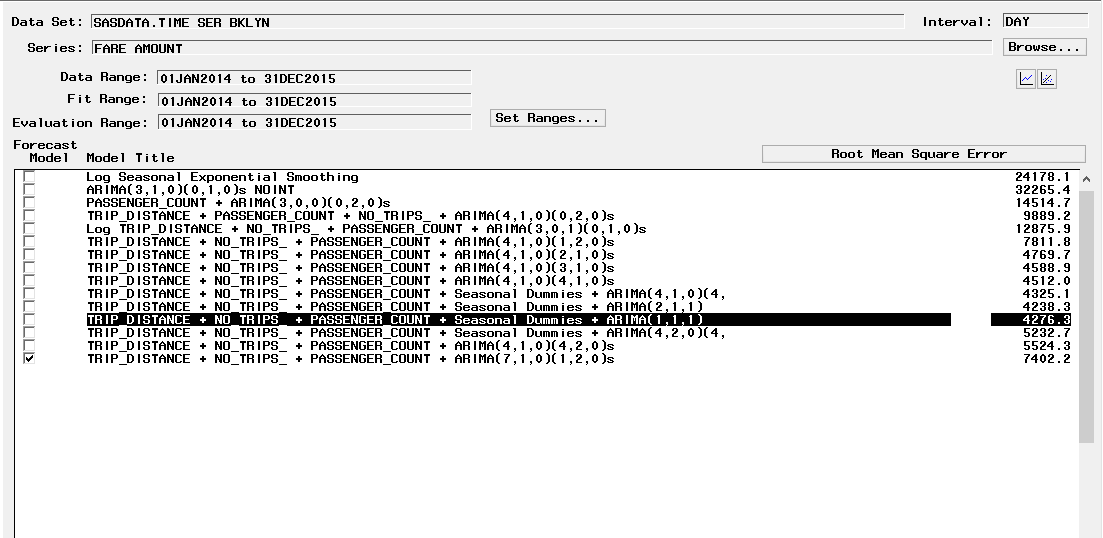
**Simple & Seasonal Difference:** Simple difference couldn’t bring much diff, so tried simple and seasonal diff.

White noise wipes off but only from 1st lag.

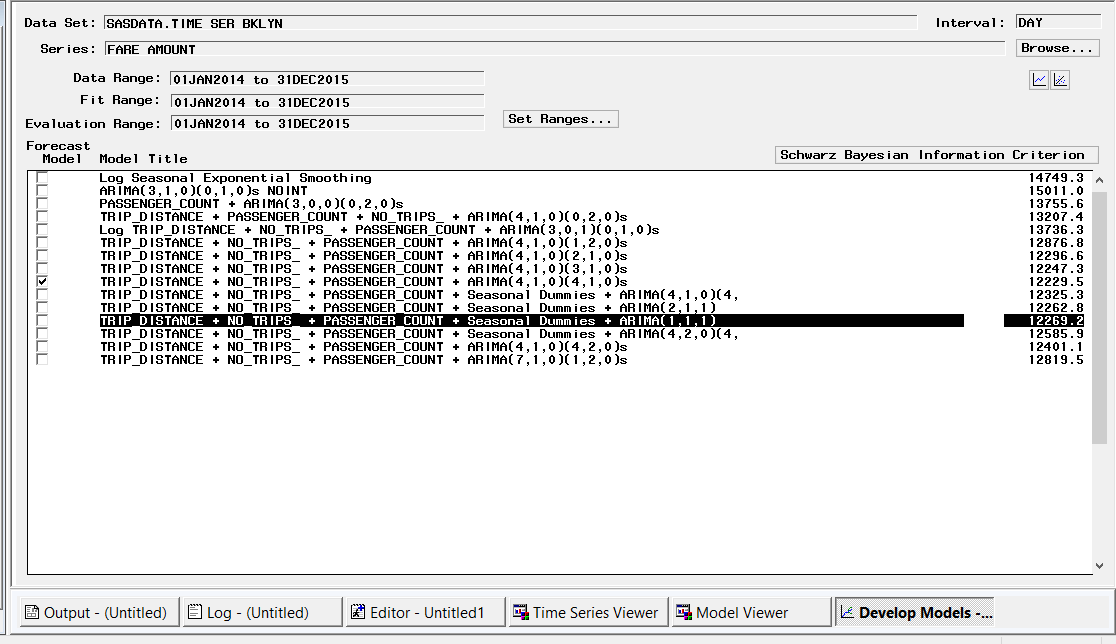




**Models based on RMSE:**

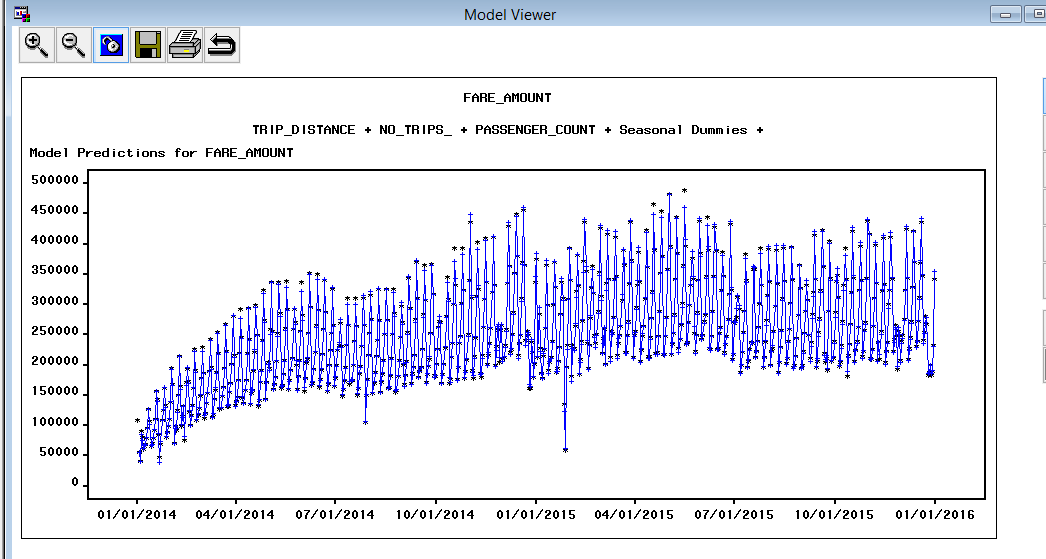


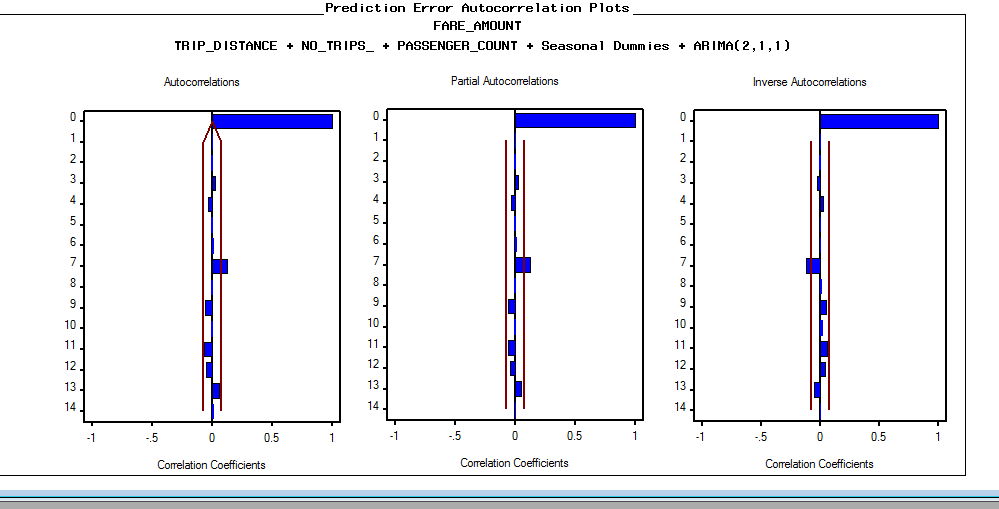
**SBC values of the tried models:**

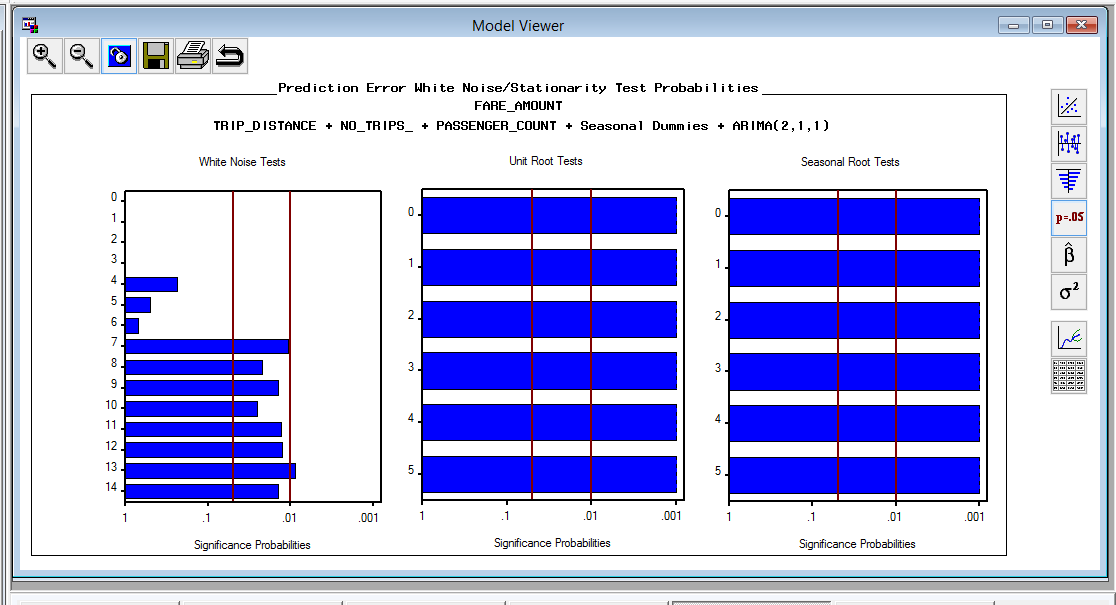


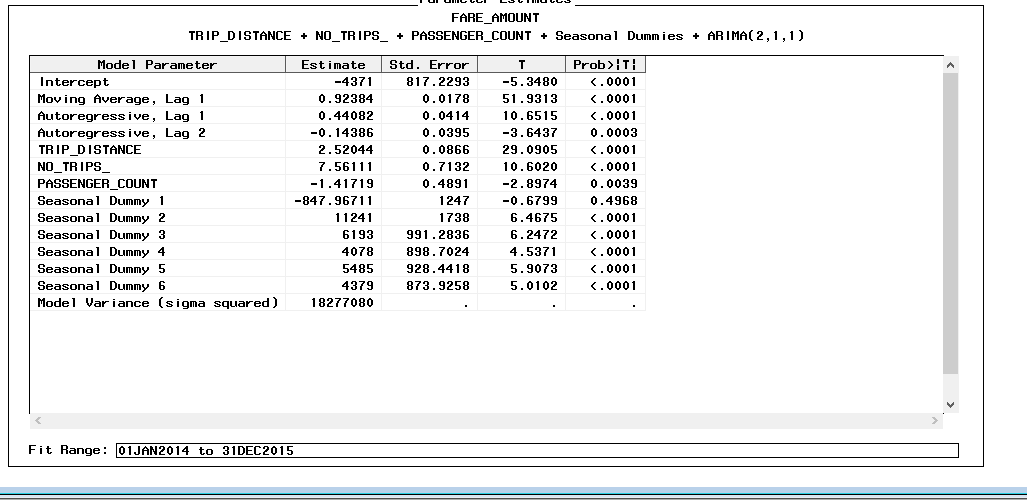
Using the regressors, ARIMA model, the best model derived is:

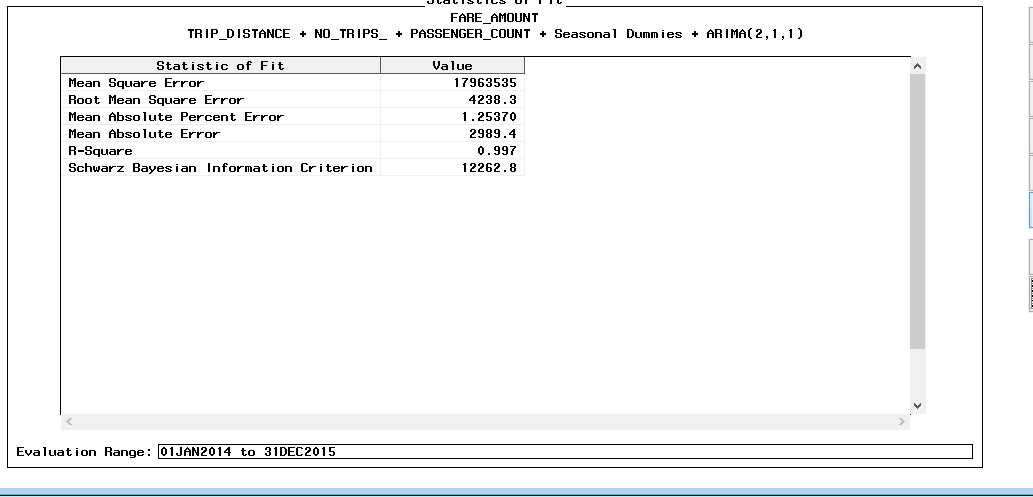
* With and without holdout sample this model seems to be the best.
* Inferences of the derived model:
* White noise is quite insignificant except a few lags.
* Model is stationary.
* Its highly seasonal, adding regressors greatly reduces white noise.
* Interventions adding couldn’t bring much difference to the model.

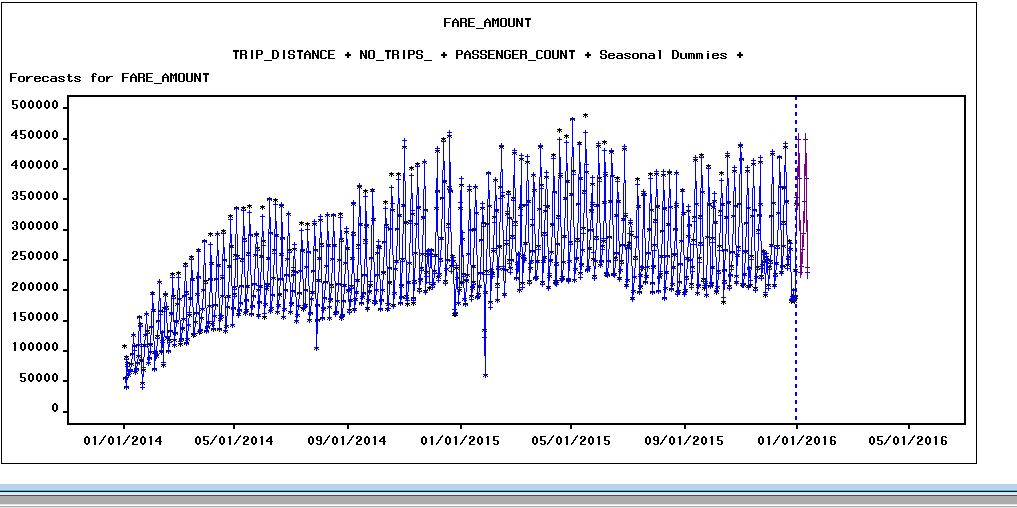




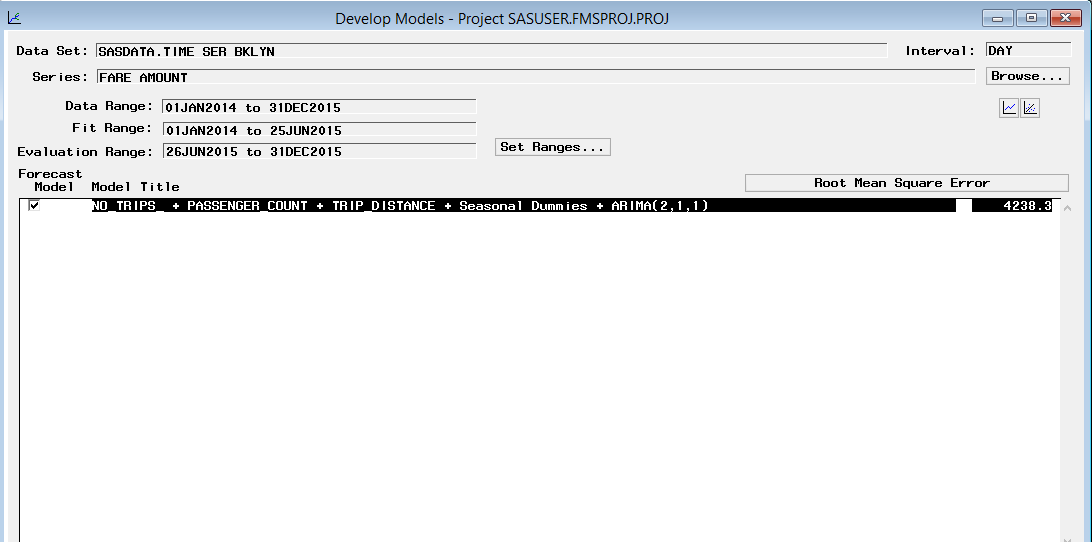


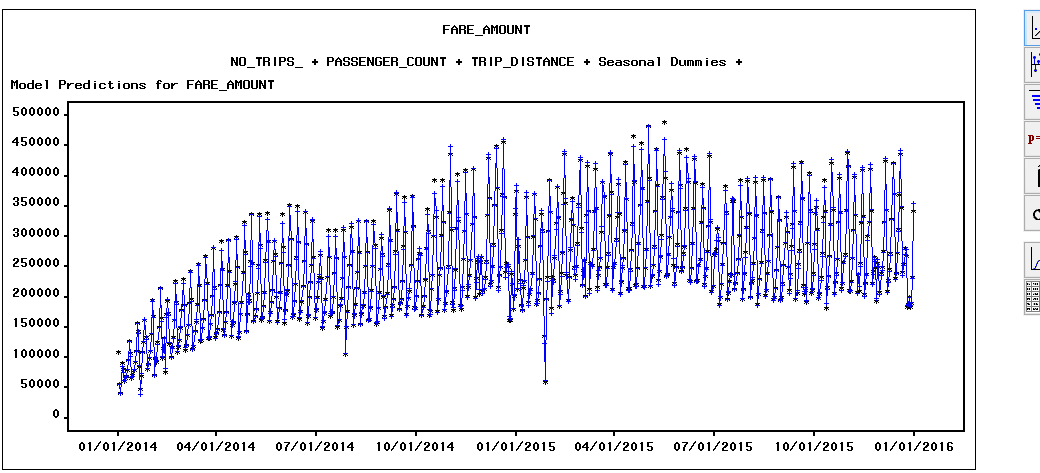


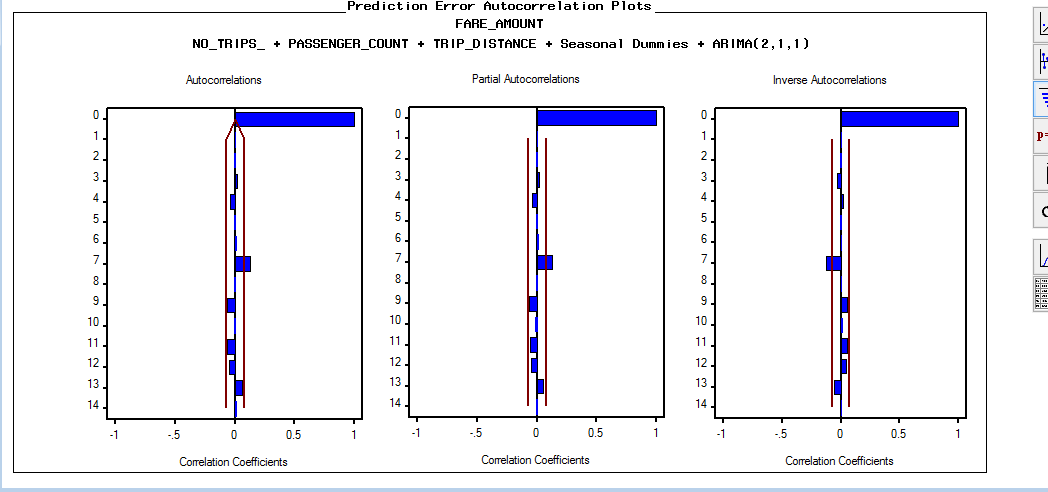


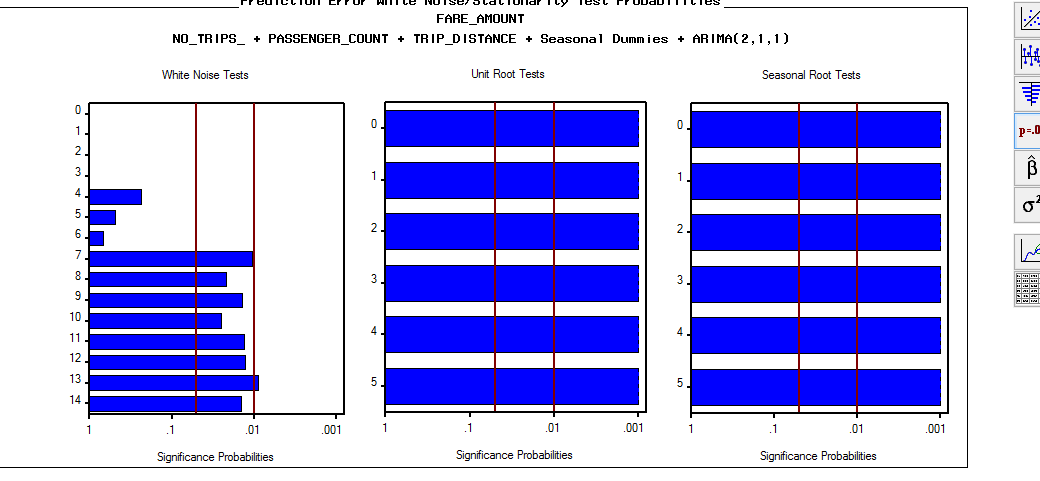


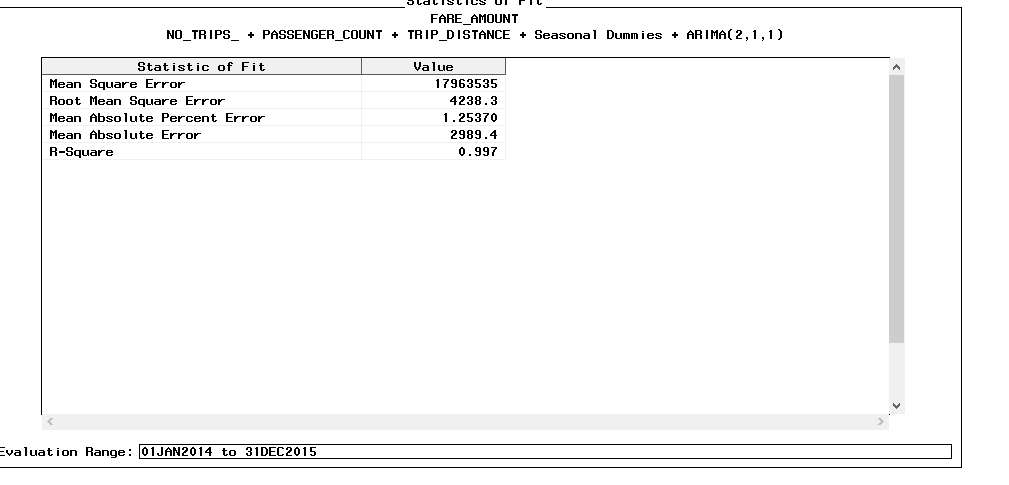
**With holdout sample:**

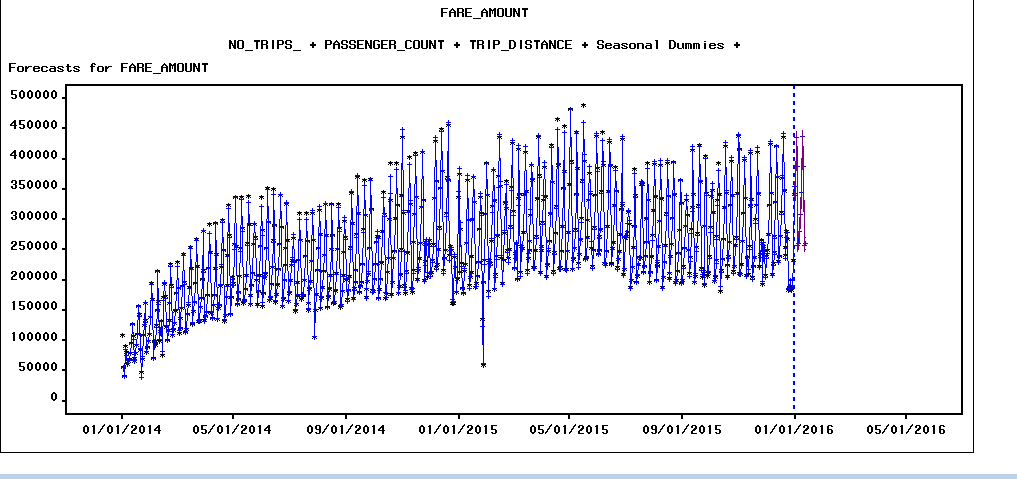




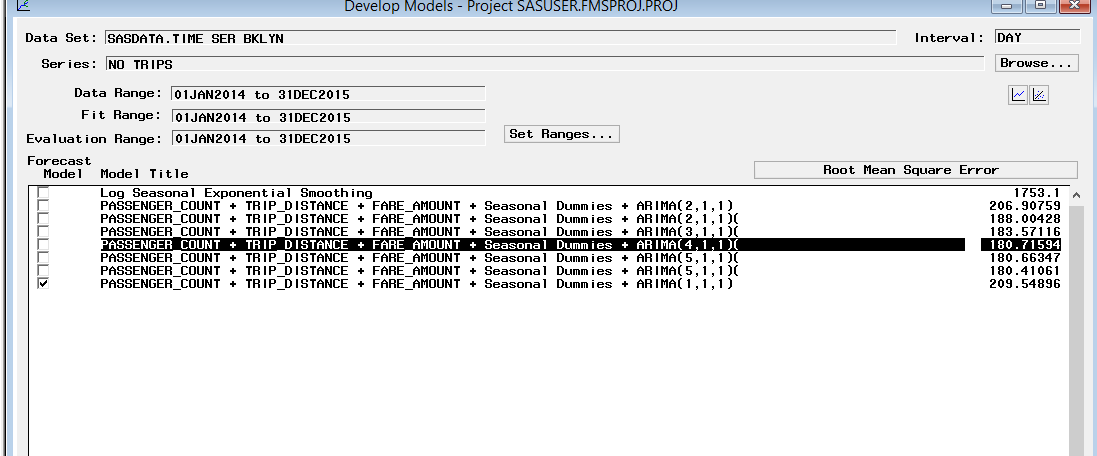


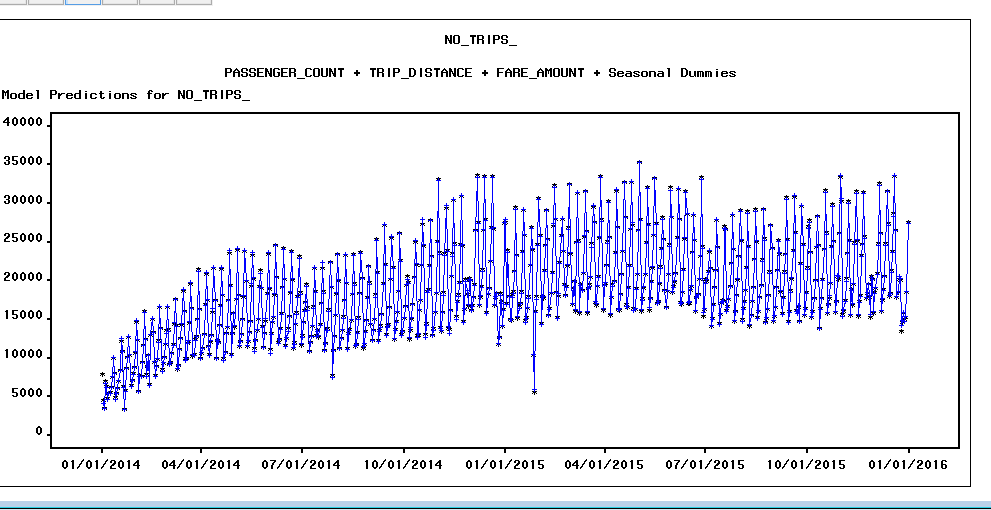


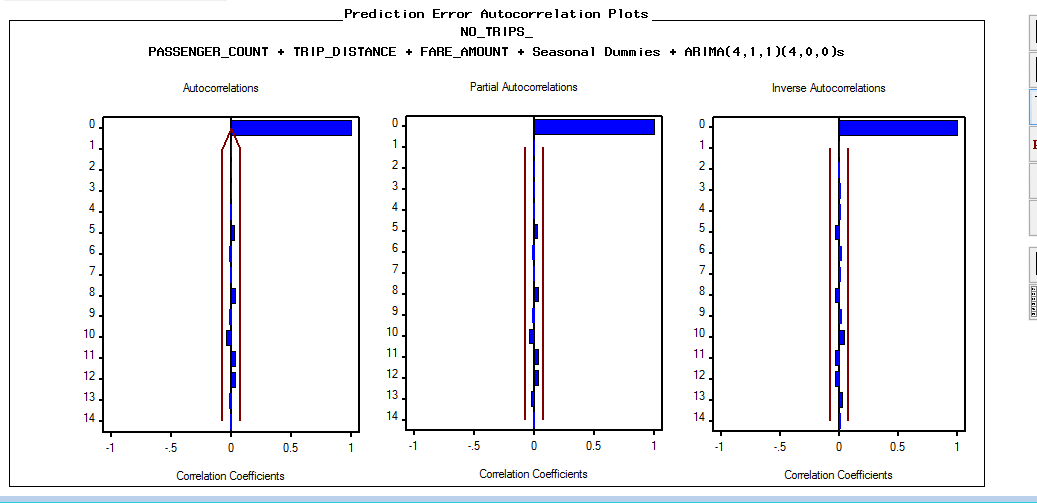


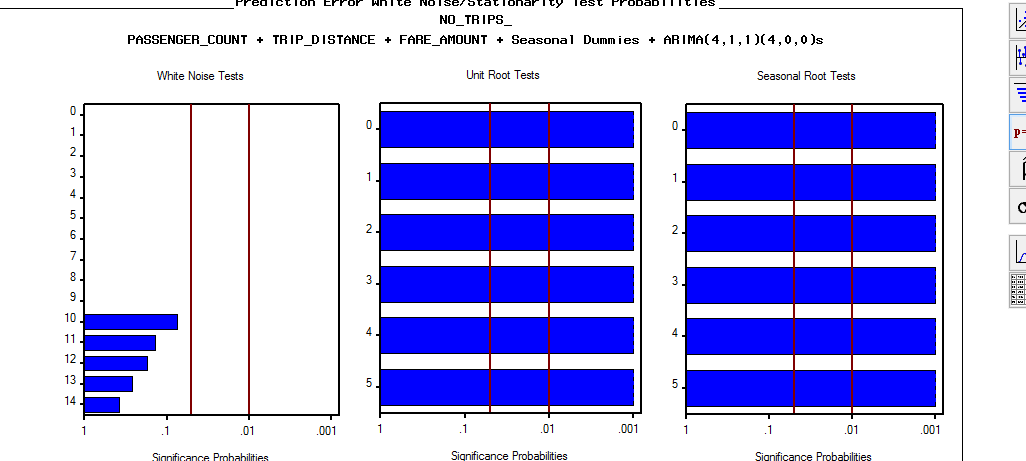


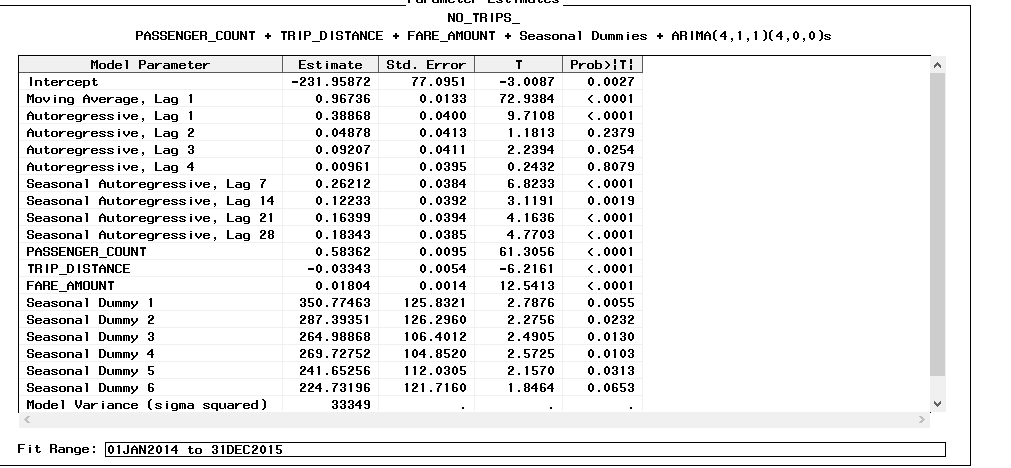
**No. of Trips**

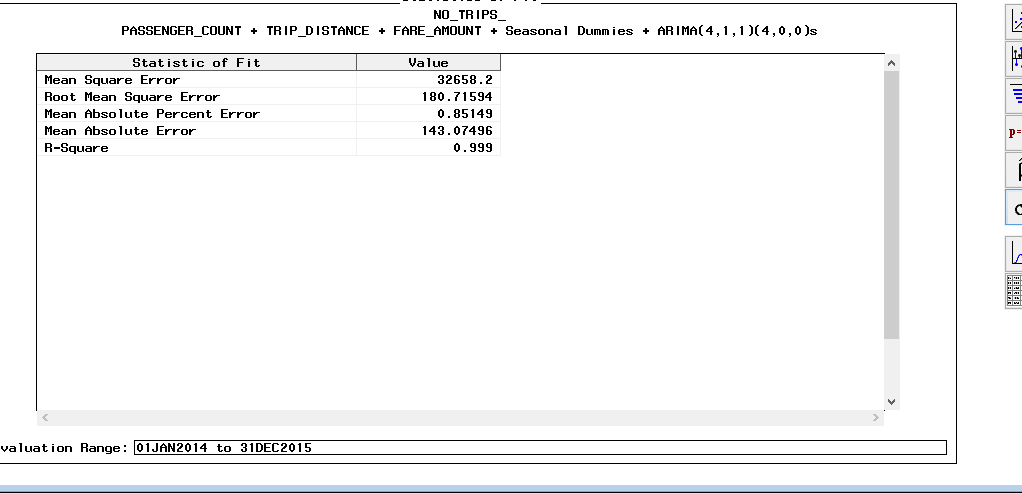


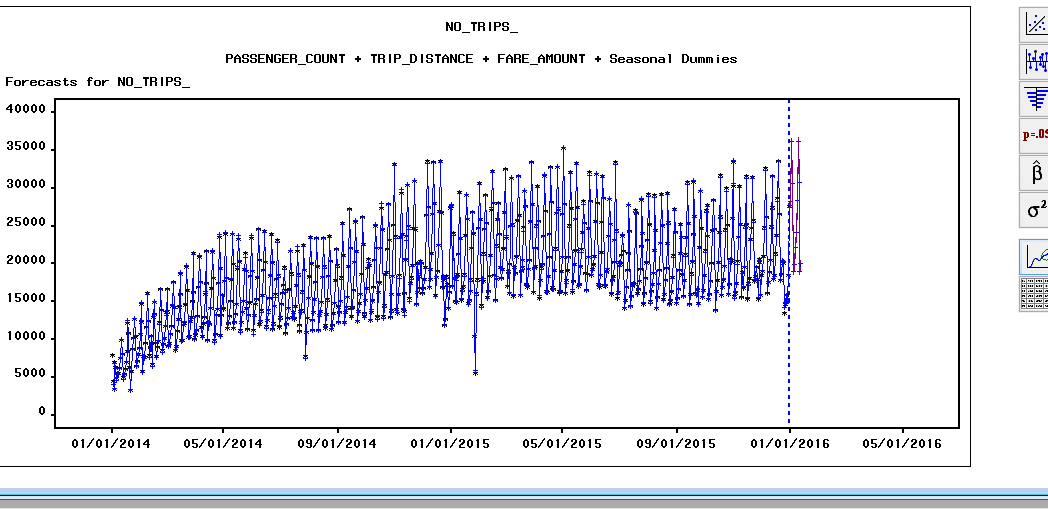




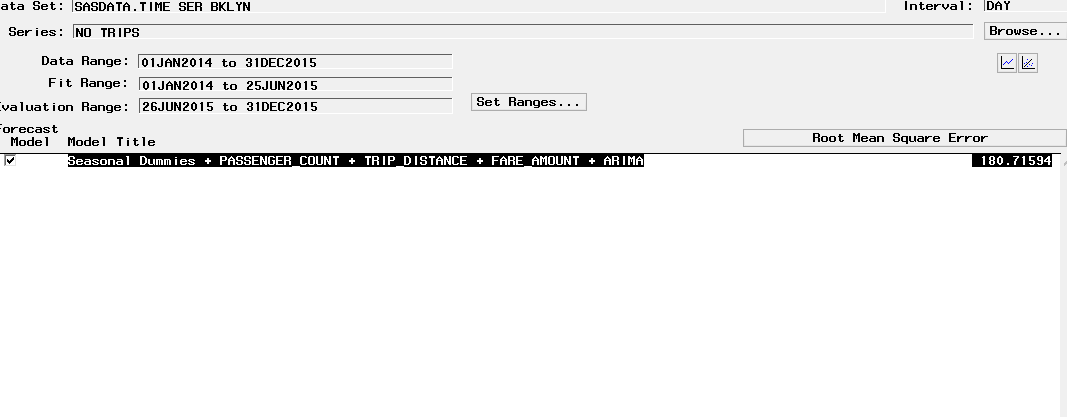


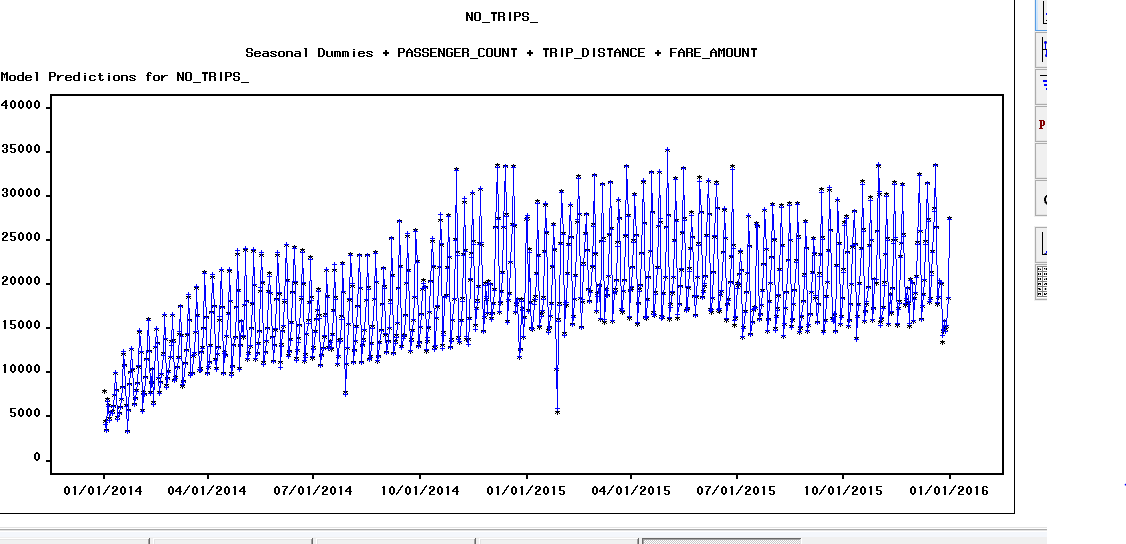


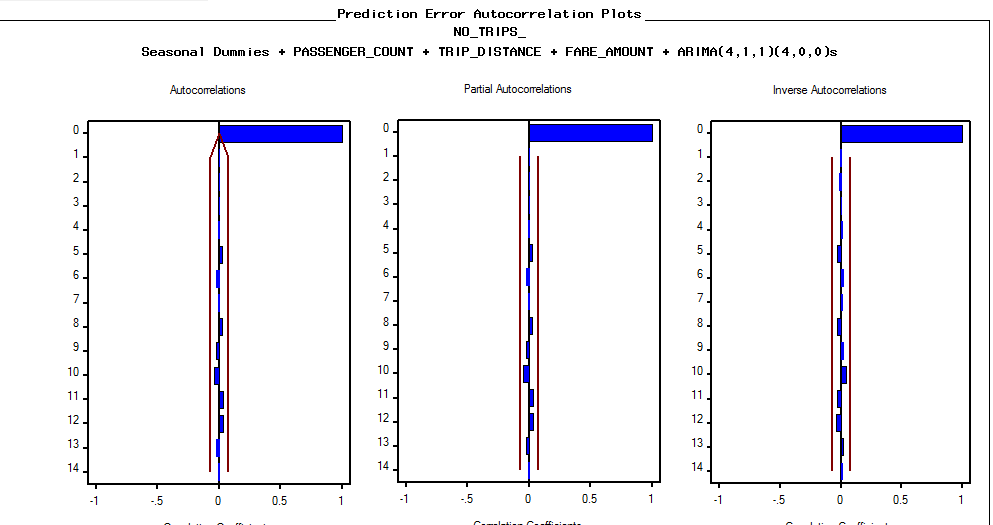


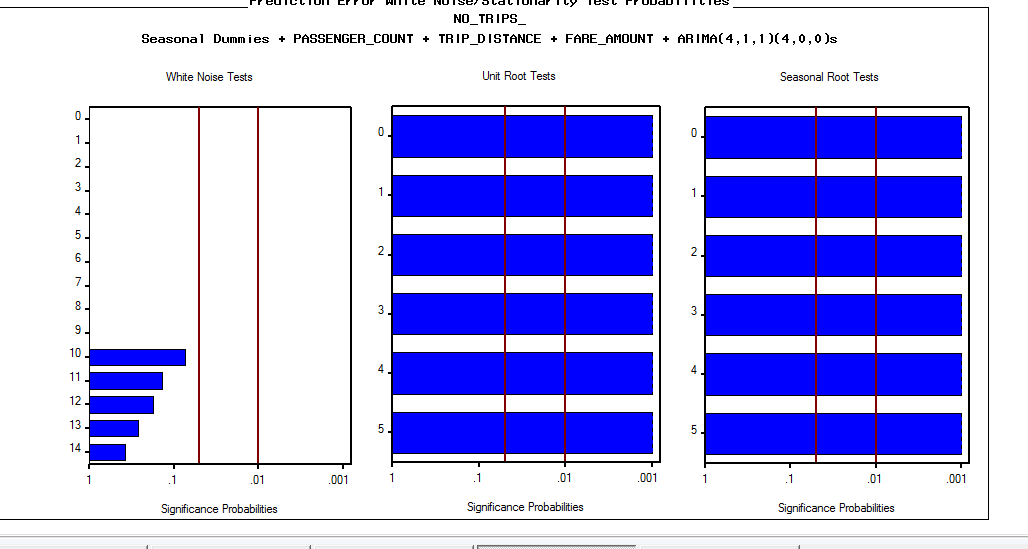


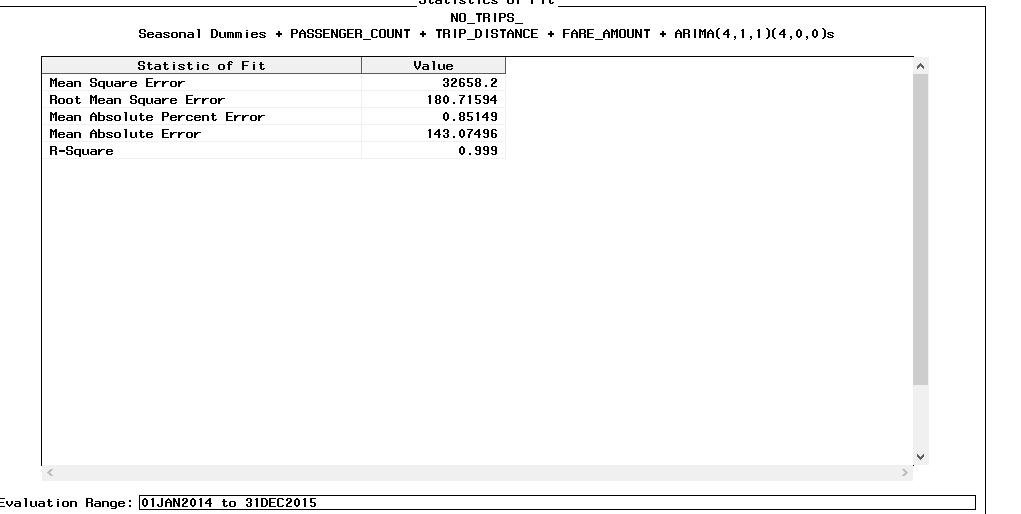
**Without holdout Sample:**

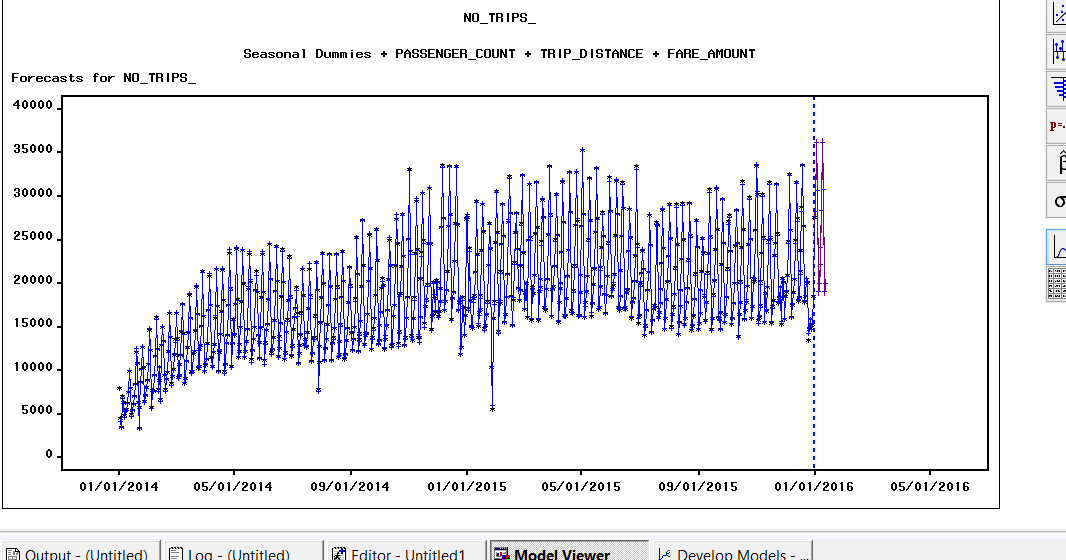






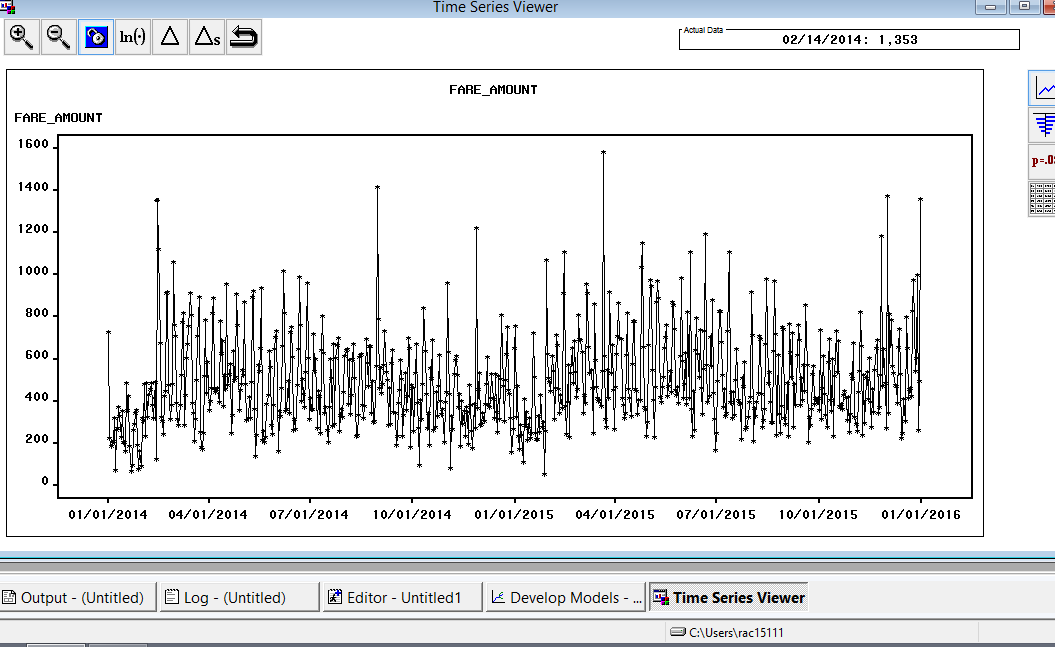






## **Queens Dataset**

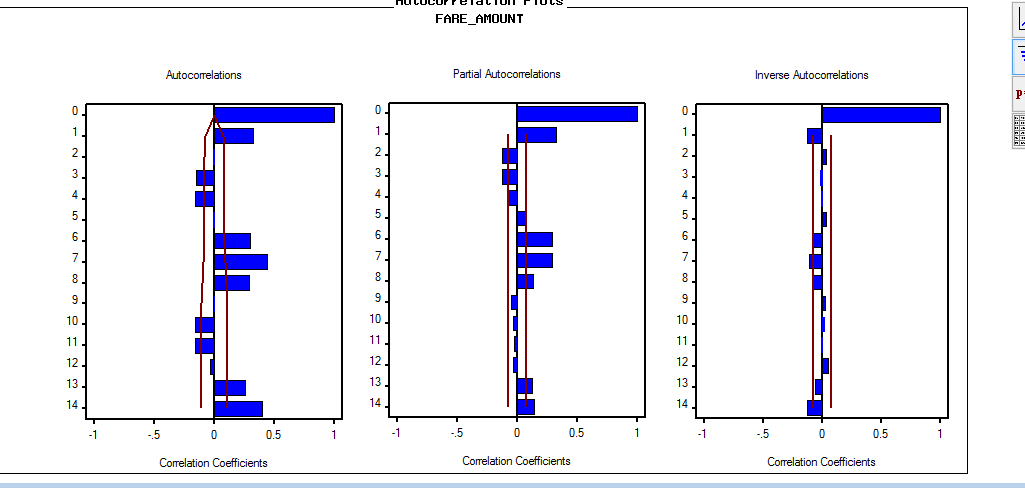
Queens Raw Dataset:

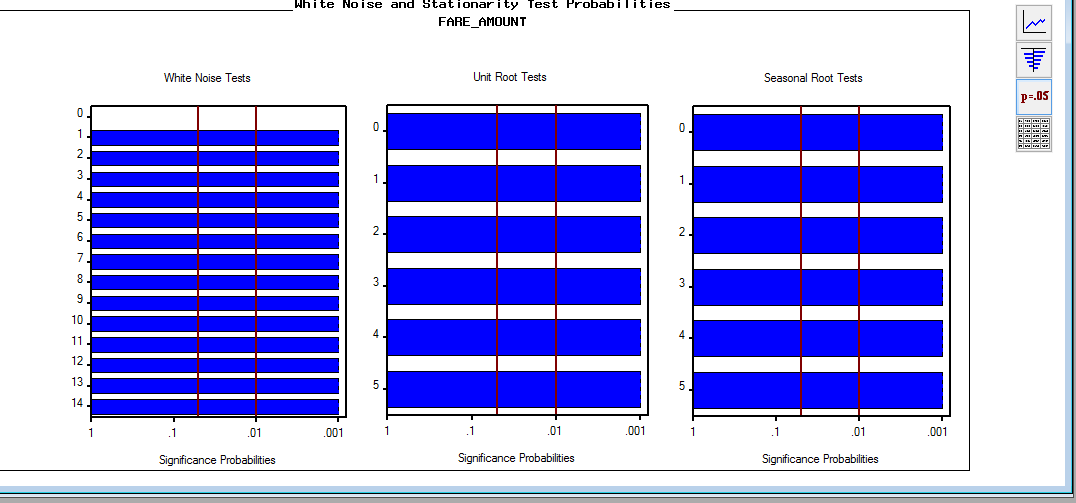


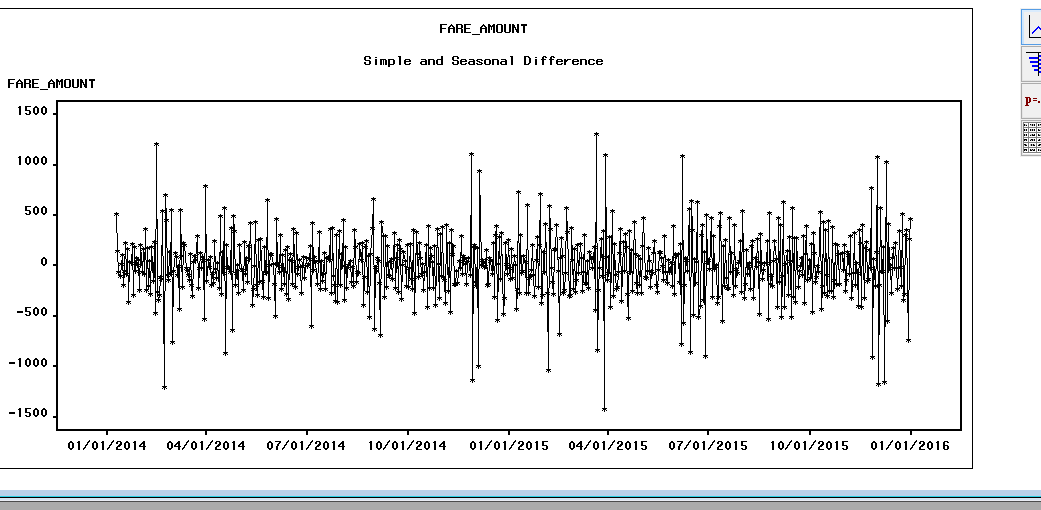
**Insights:**

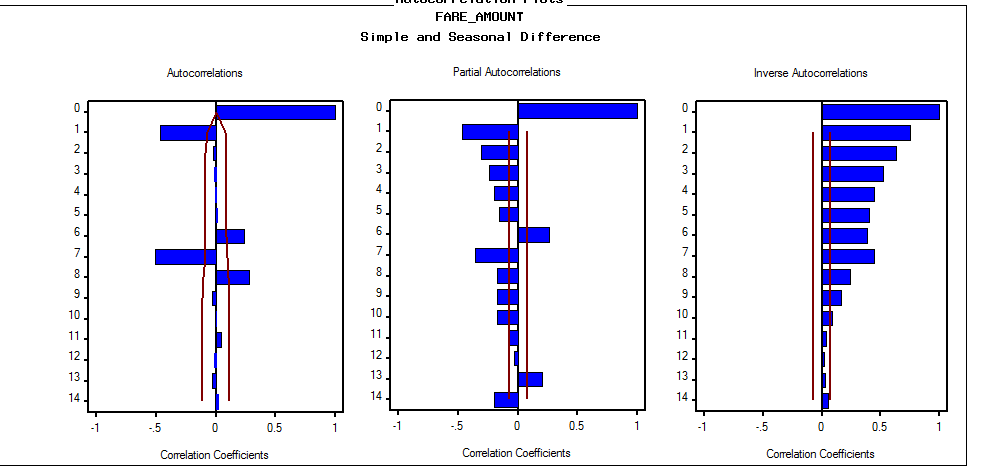
Peaks are generally found at:

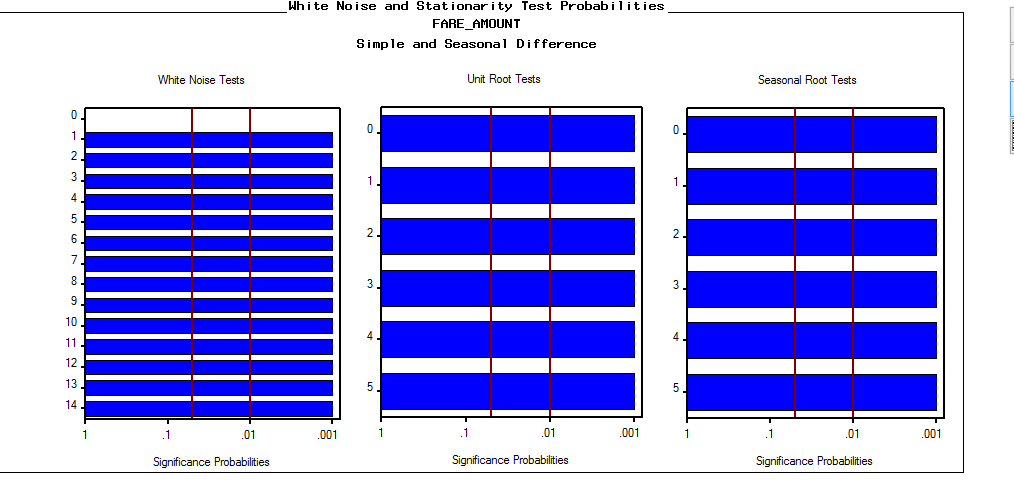
* 1st week of December, after second week there seems to be a drop which is after Christmas.
* On 14 Feb there’s a peak – Valentine’s Day
* Near November end, 27 Nov is a peak – Thanks Giving Day
* In March – Highest peak – spring begins and long weekend; may be spurious.
* Significant white noise
* Seasonality an important component.





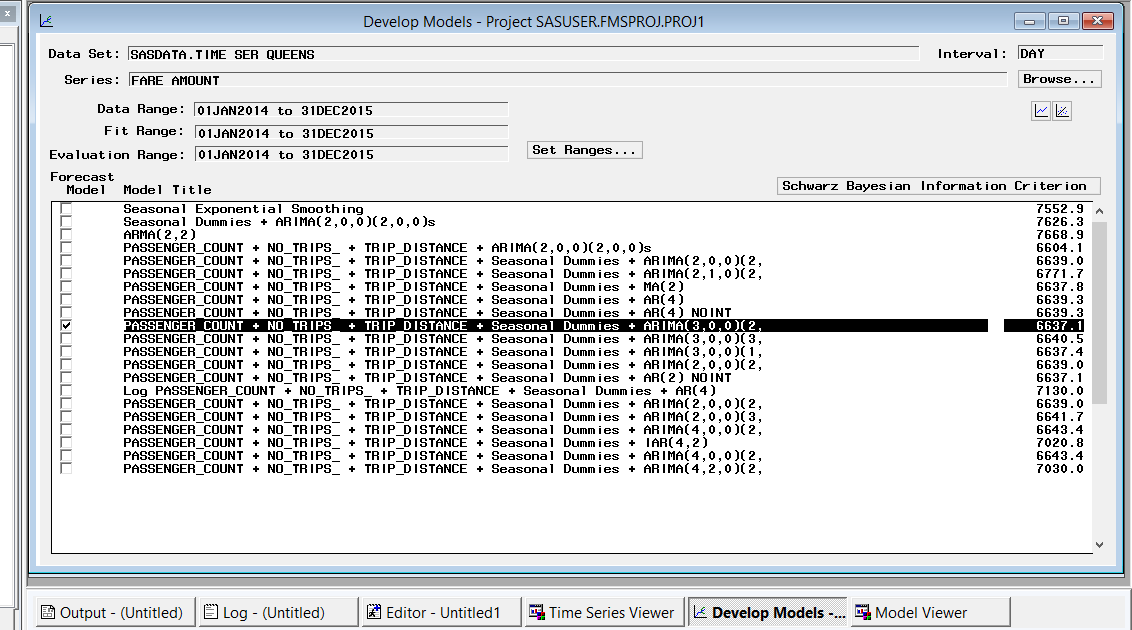


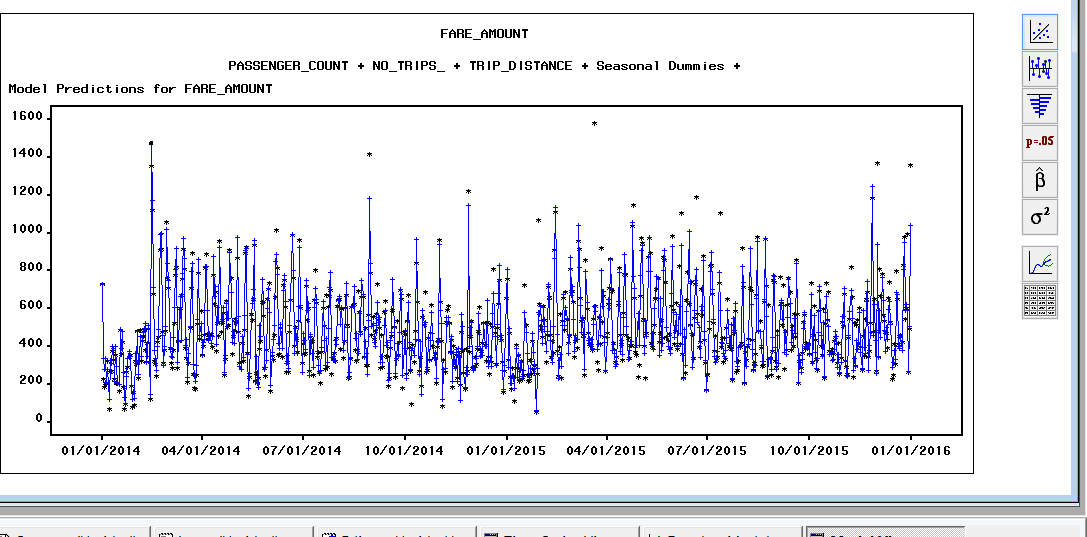




After applying regressors and ARIMA model:

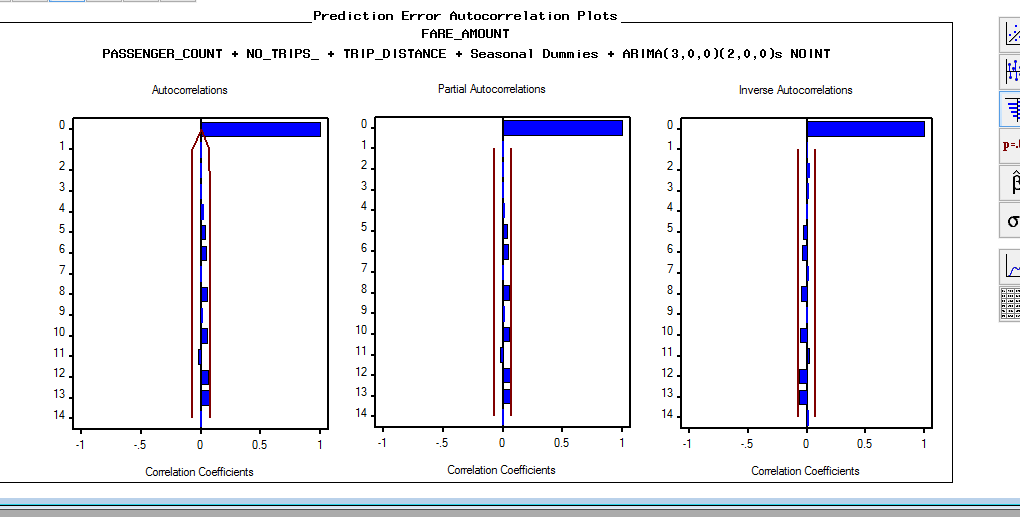
**Best model=> Regressor + Seasonal Dummies +ARIMA (3,0,0)(2,0,0)**

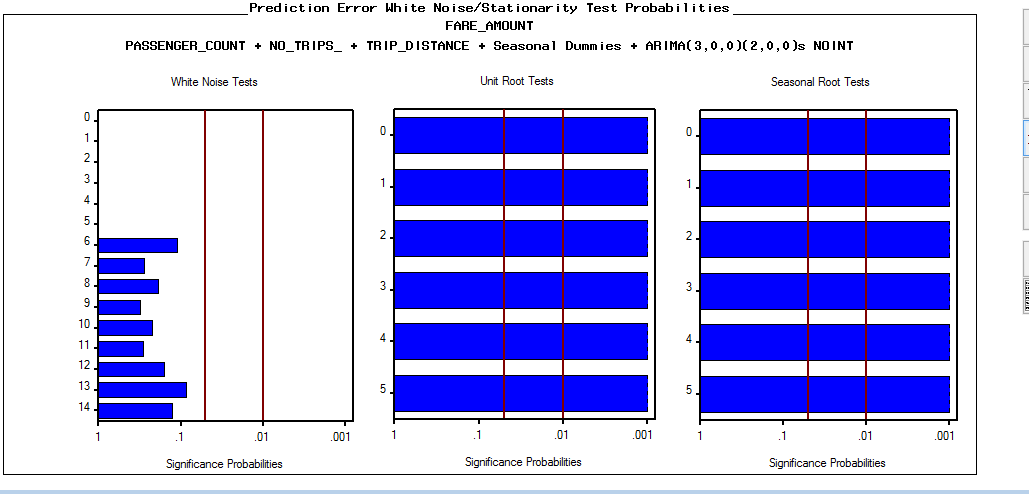


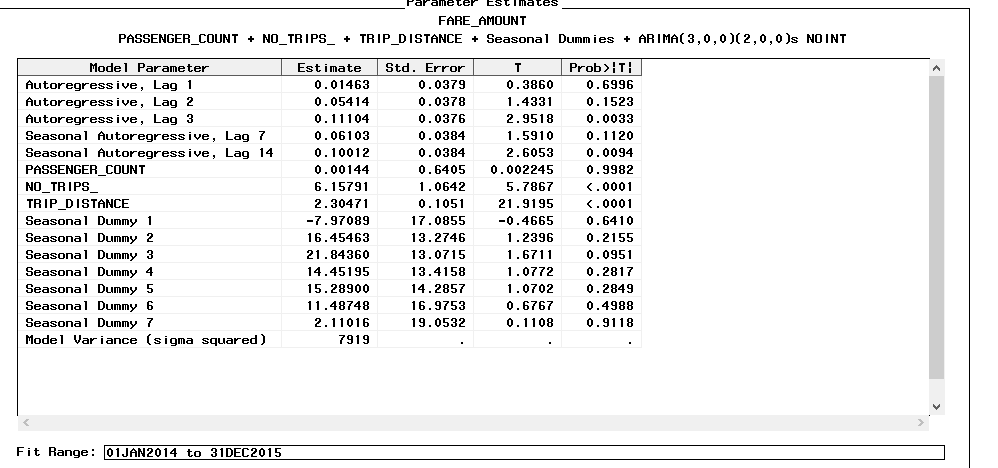


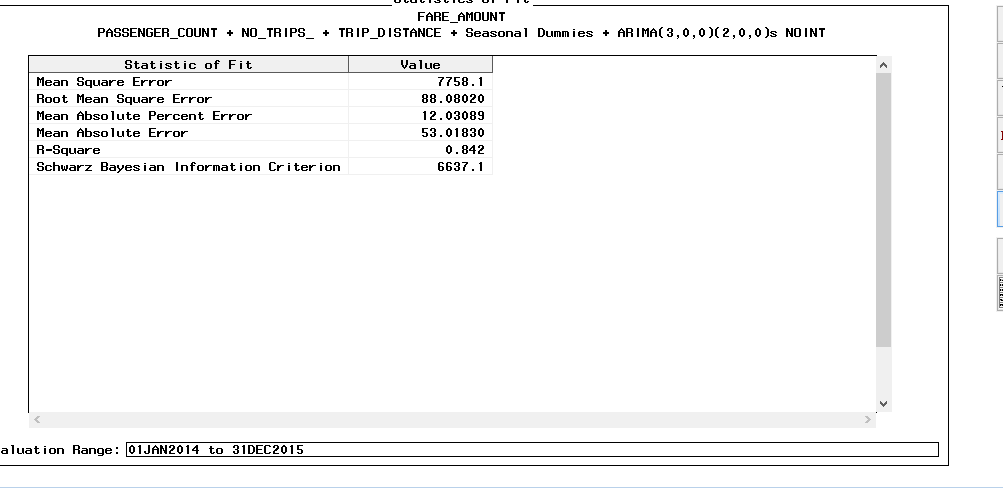
**Inferences after modeling:**

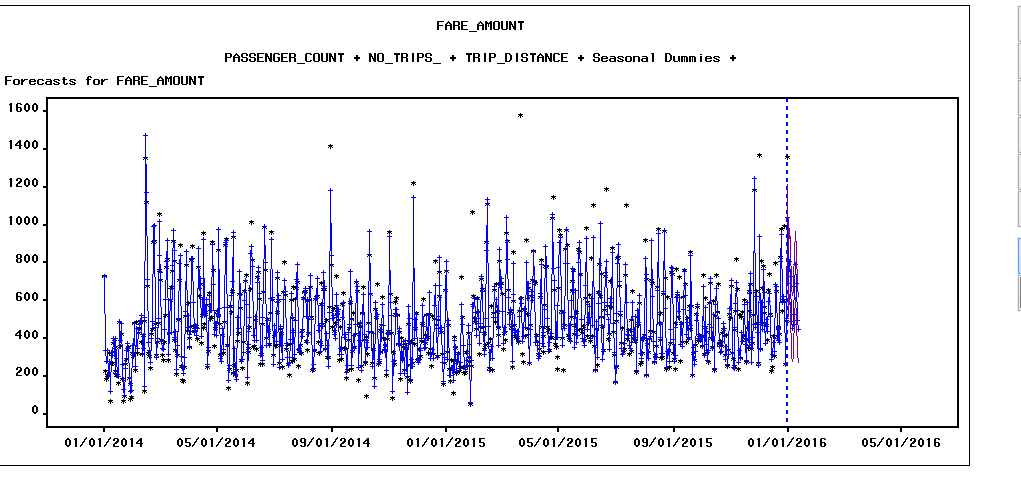
* Highest Peak: 14 Feb 2014
* White noise insignificant
* Seasonality prominent



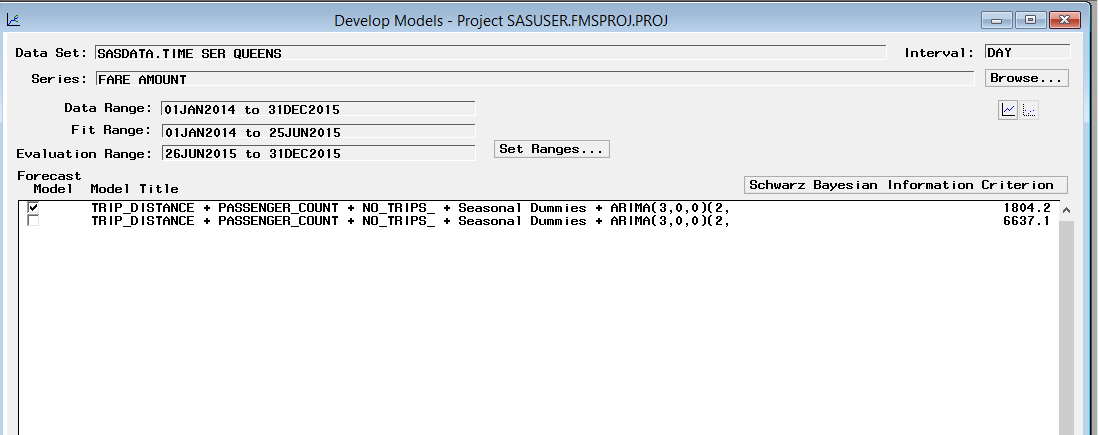


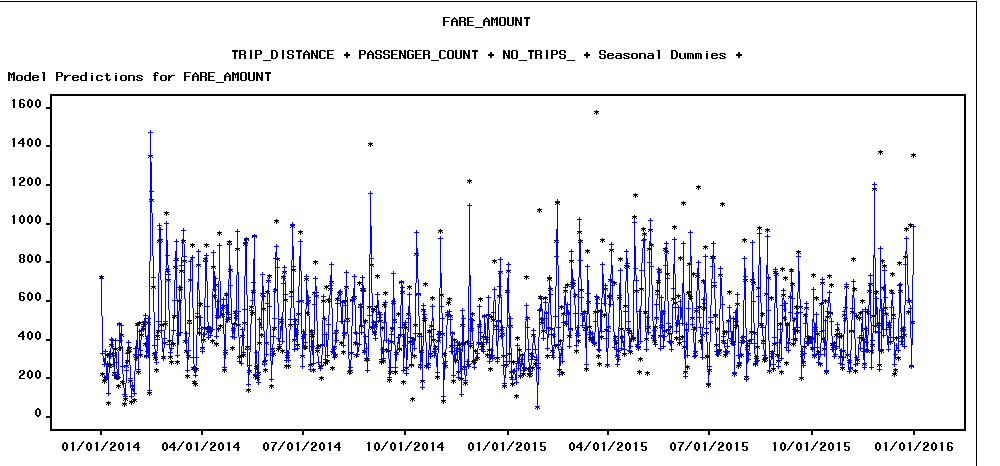


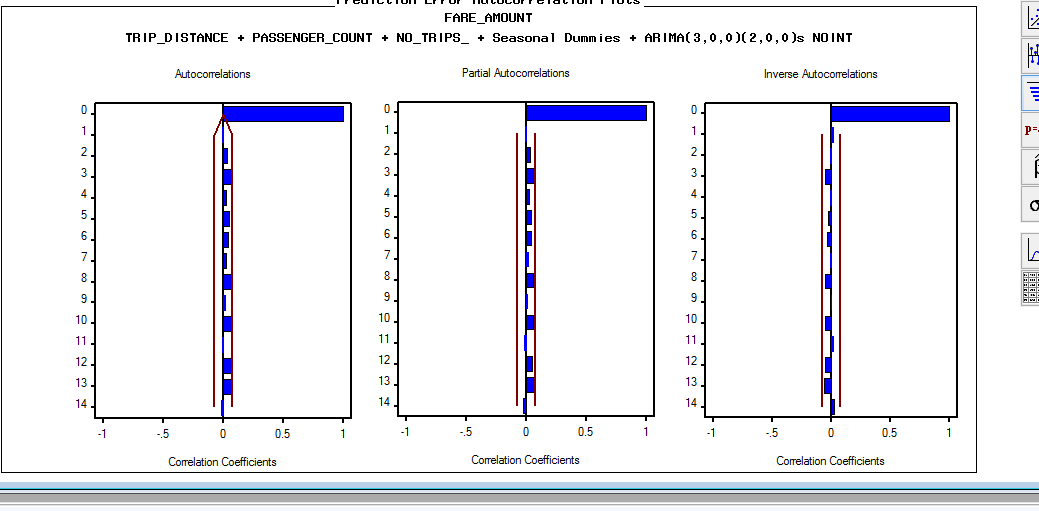




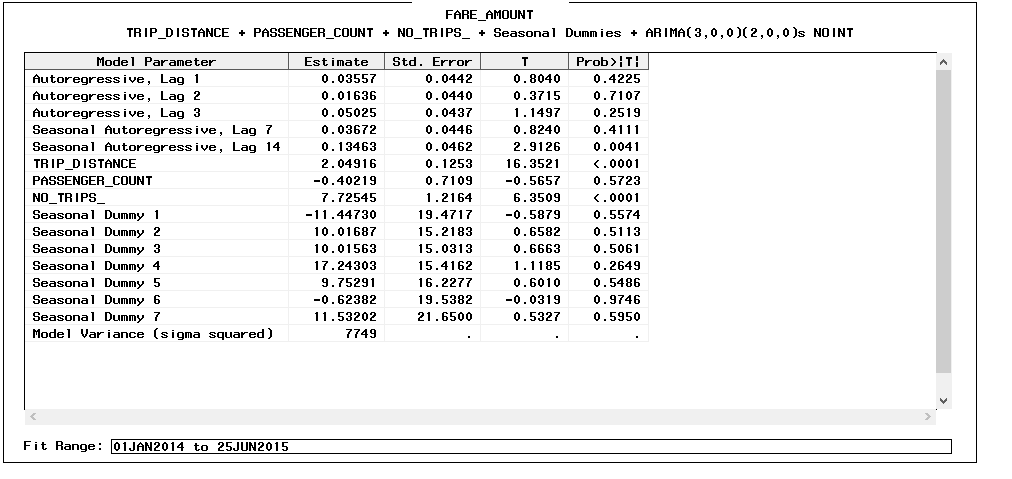
**With Holdout Sample:**

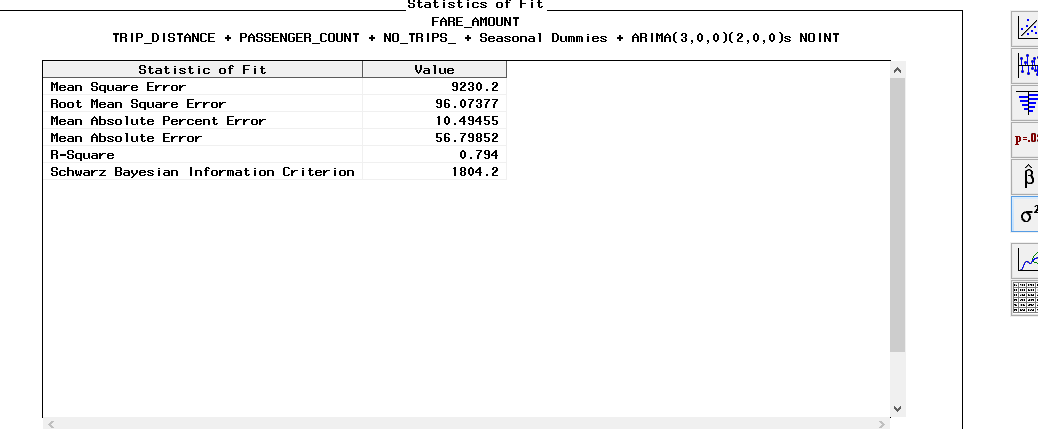


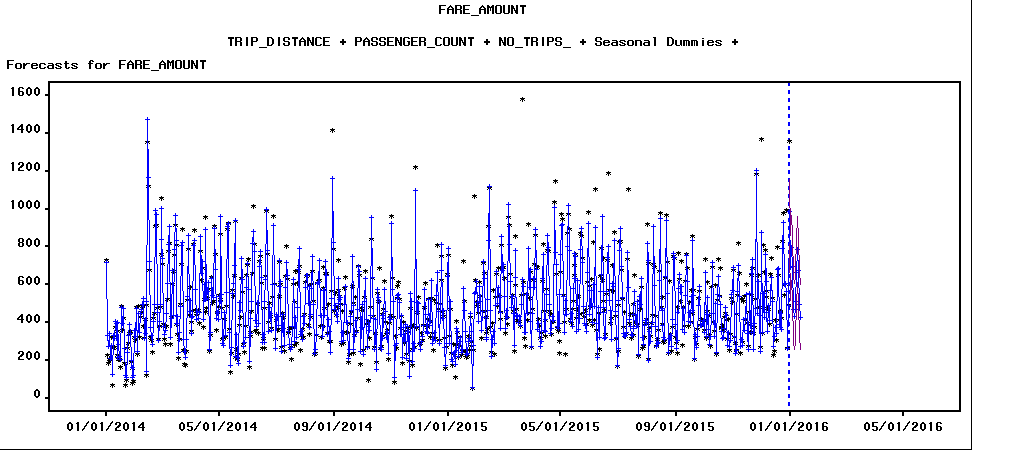




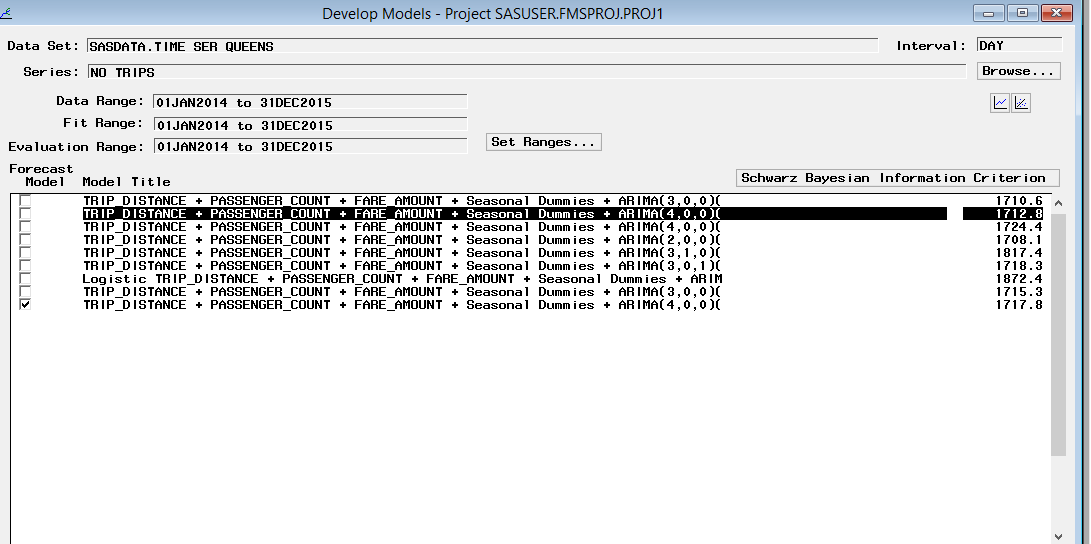


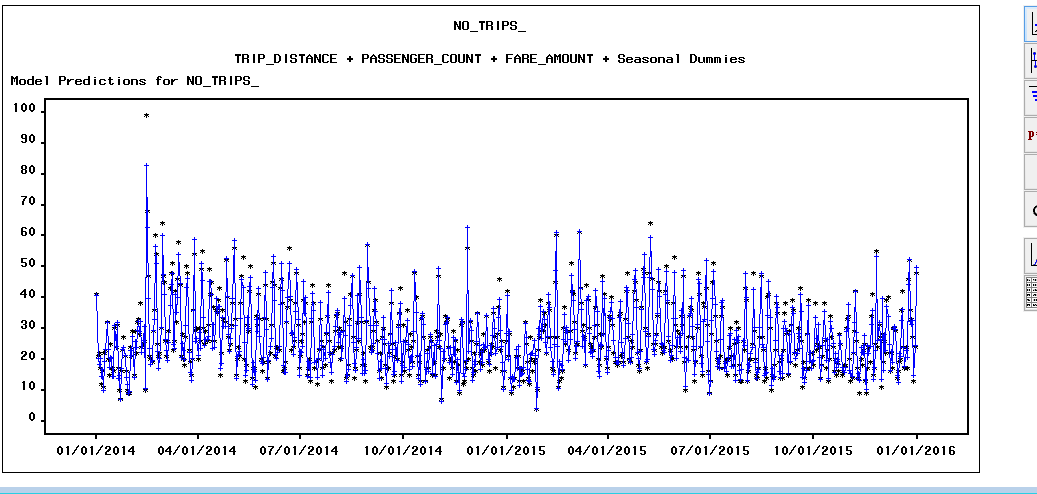




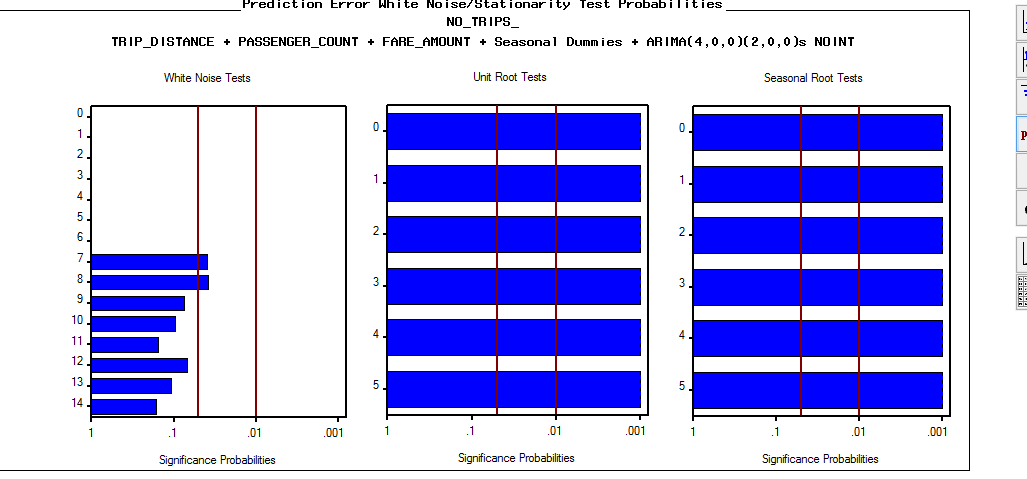


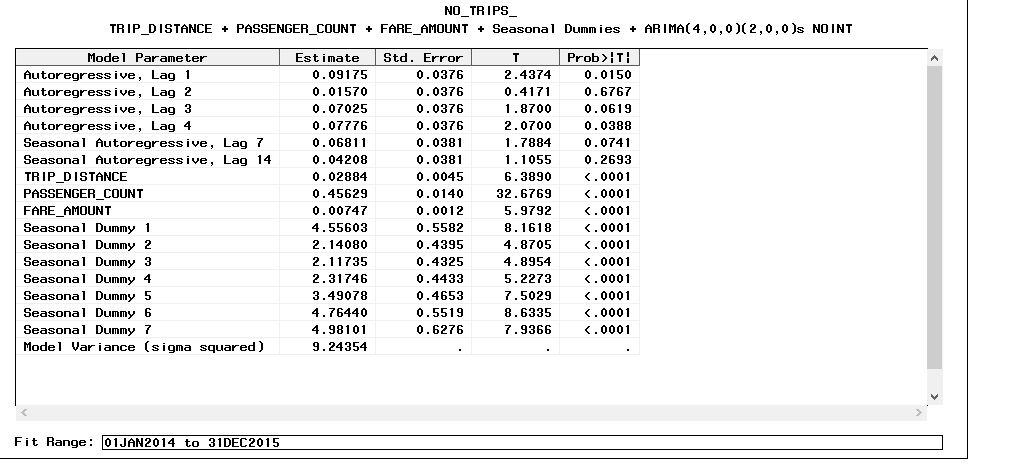
**No of trips:**

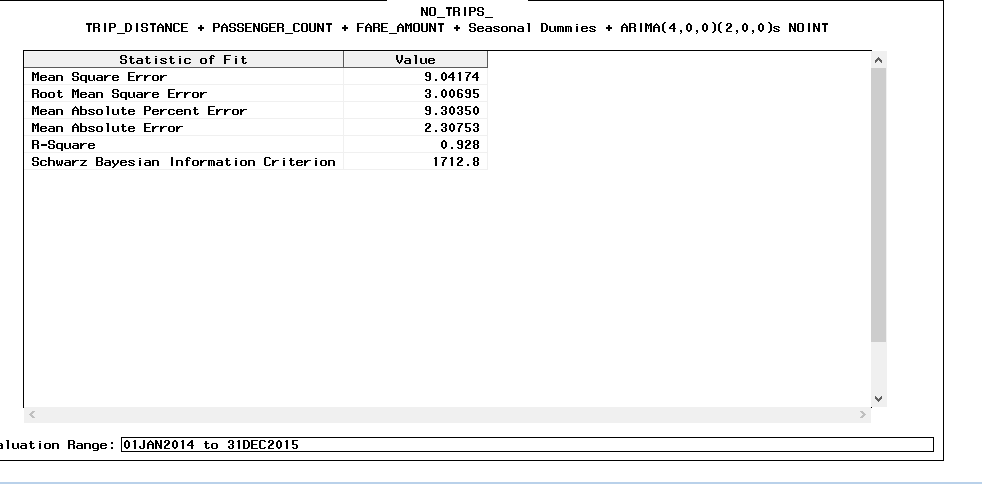


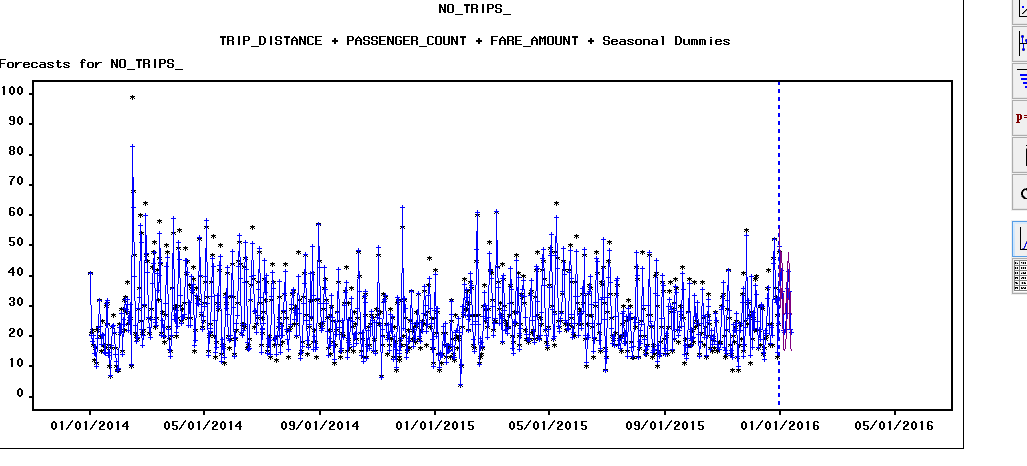




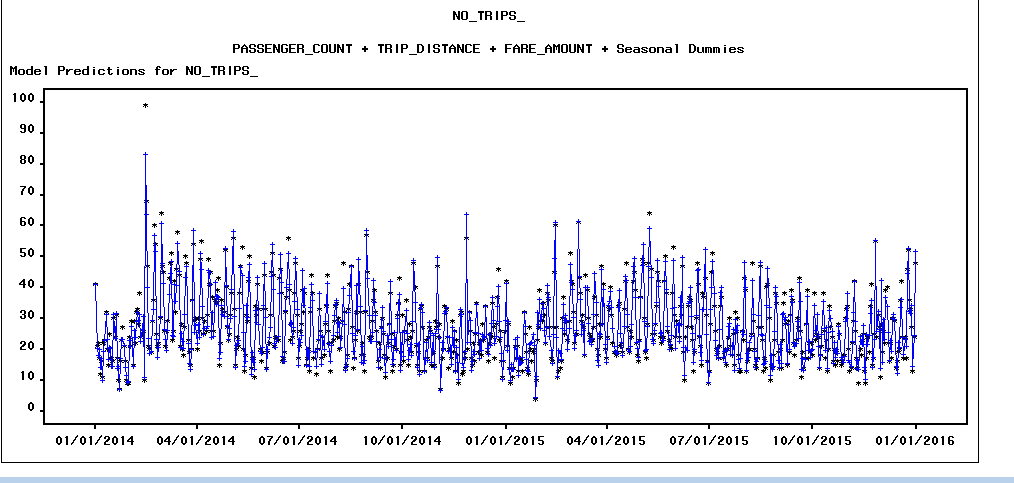


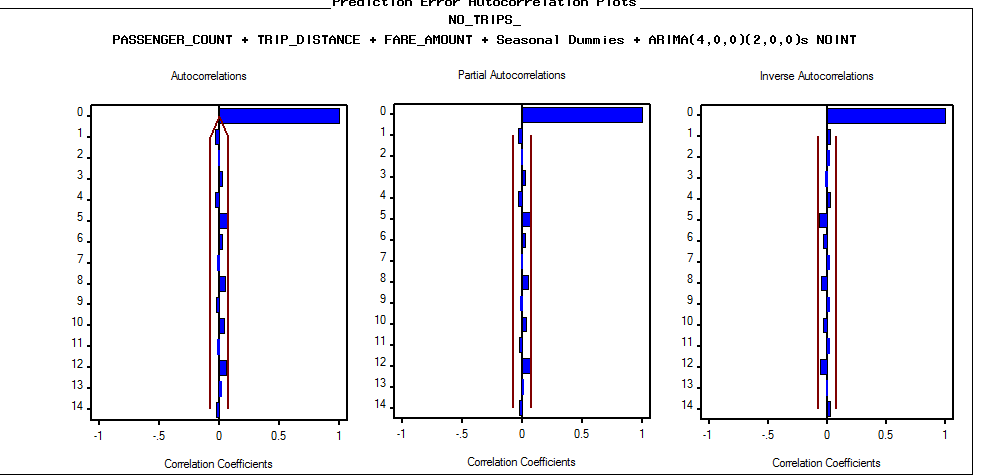


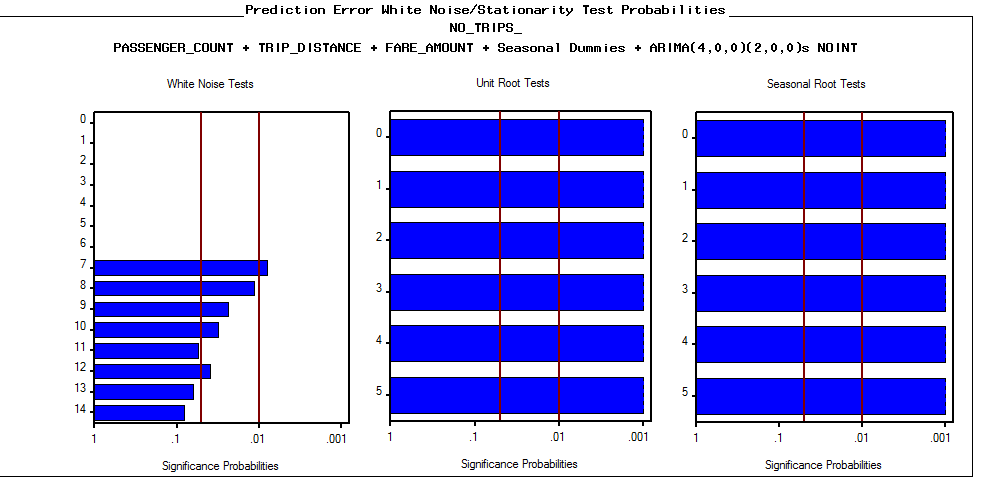


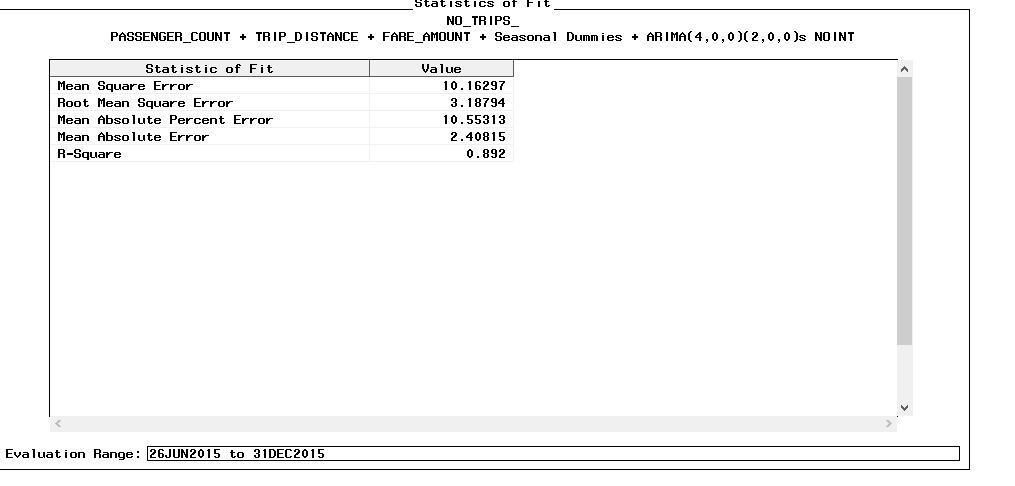


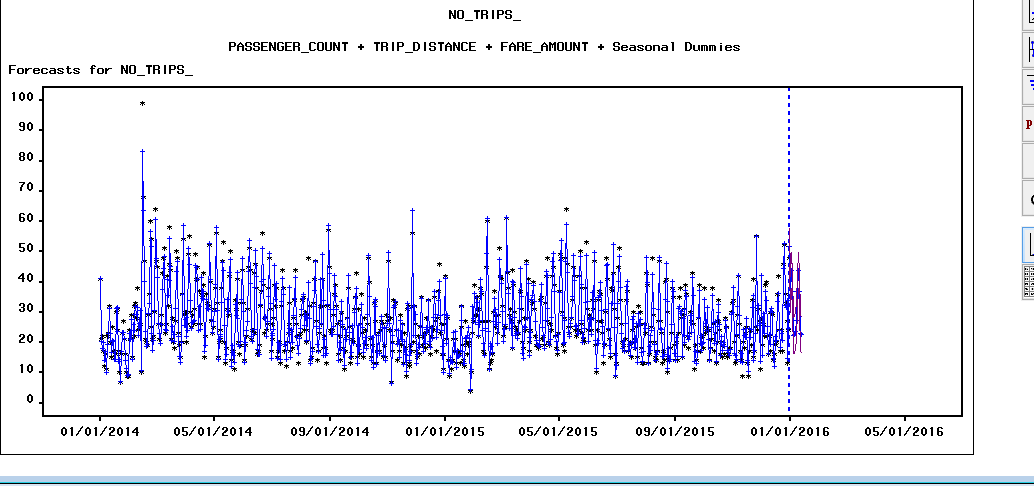
**With holdout Sample:**











Findings:

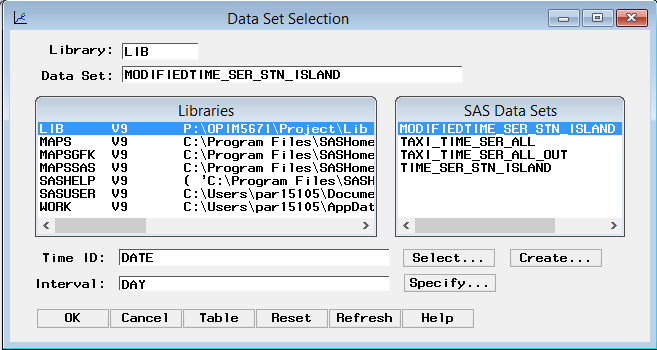
No. Of trips model:

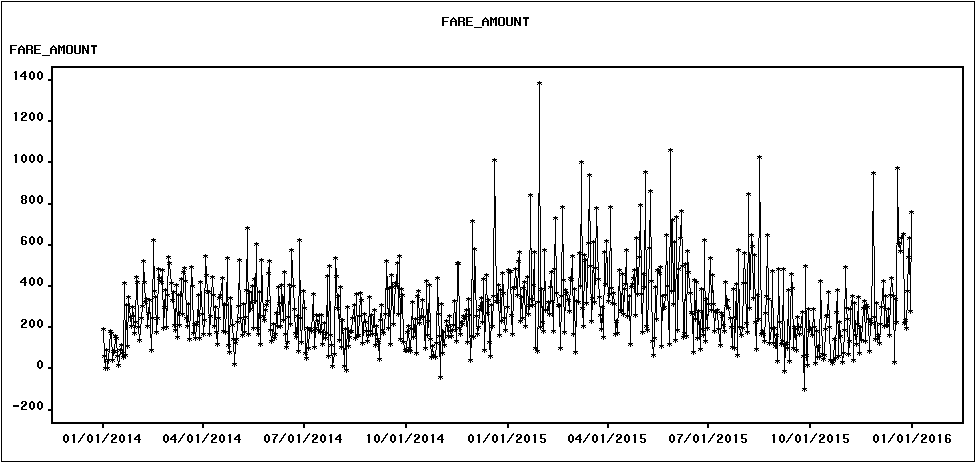
* The best model for No of trips and fare amount are different.
* The white noise and accuracy for the model of no of trips is better than that for fare amount for the boroughs.

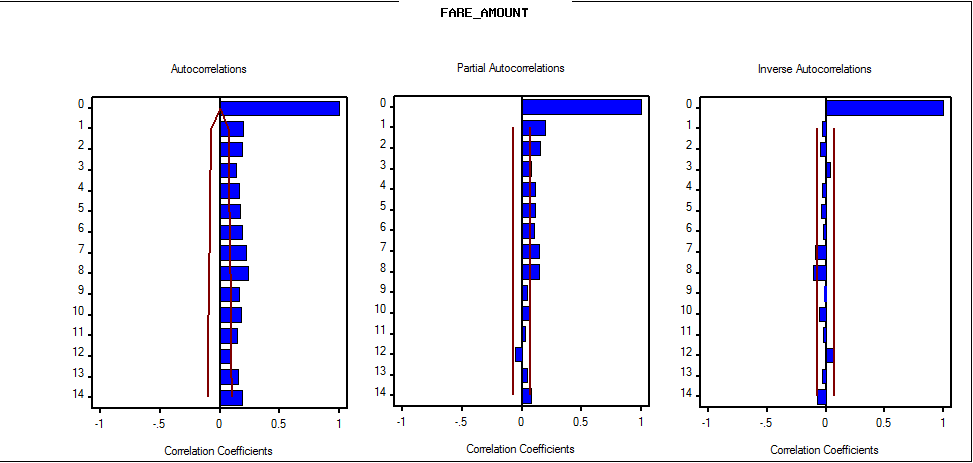
With & W/o Holdout Sample:

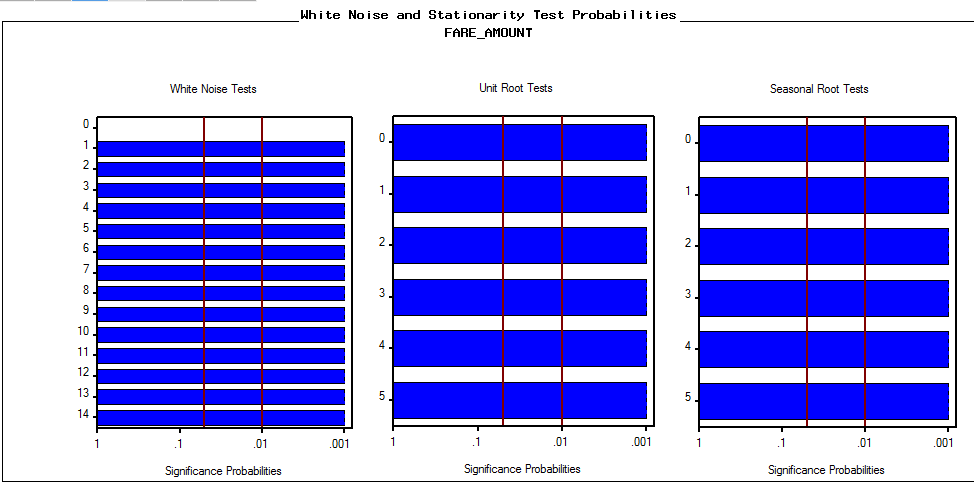
* The data accuracy increases for holdout sample.
* The value of SBC and RMSE also reduces.
* The white noise gets little more significant than for w/o holdout.

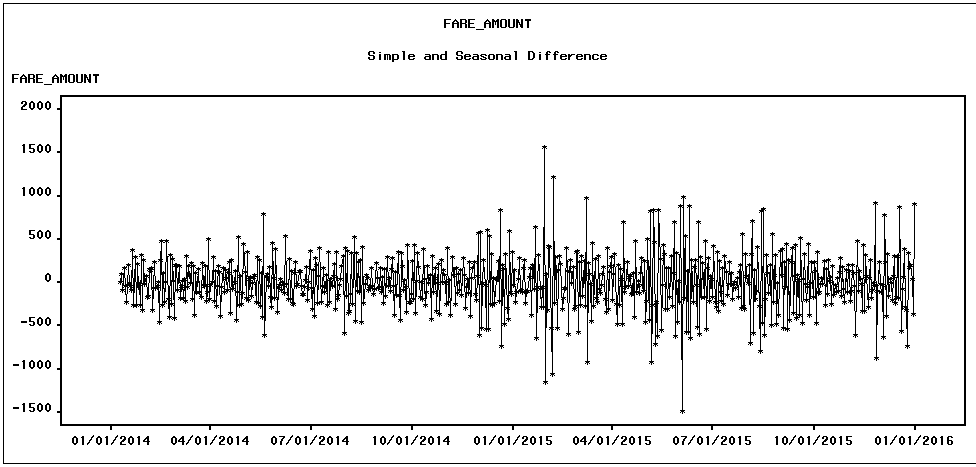
## **Staten Island Dataset**

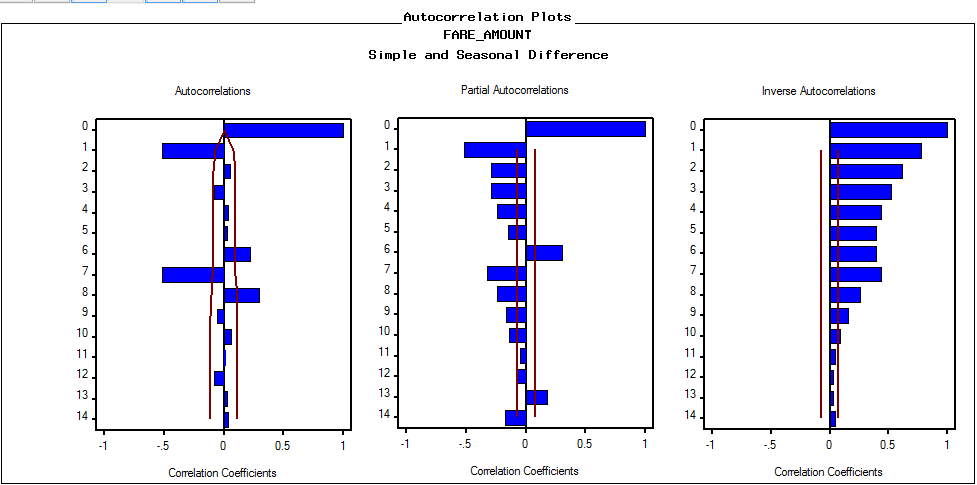


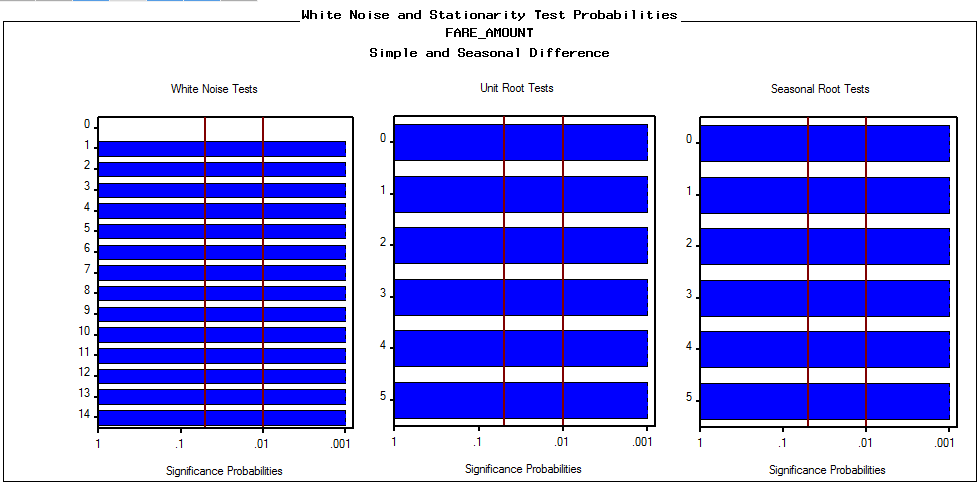


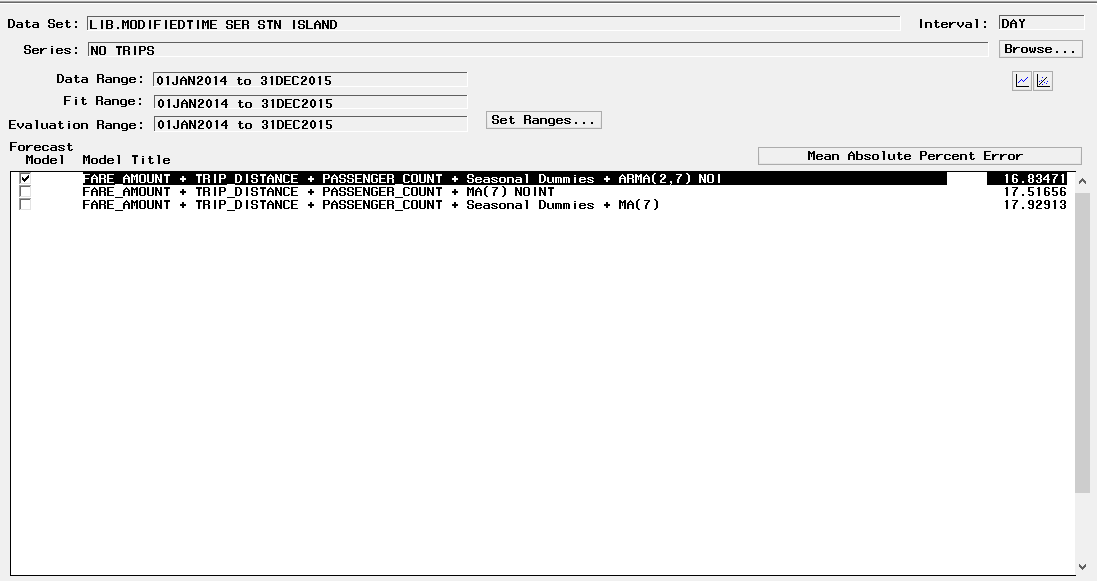


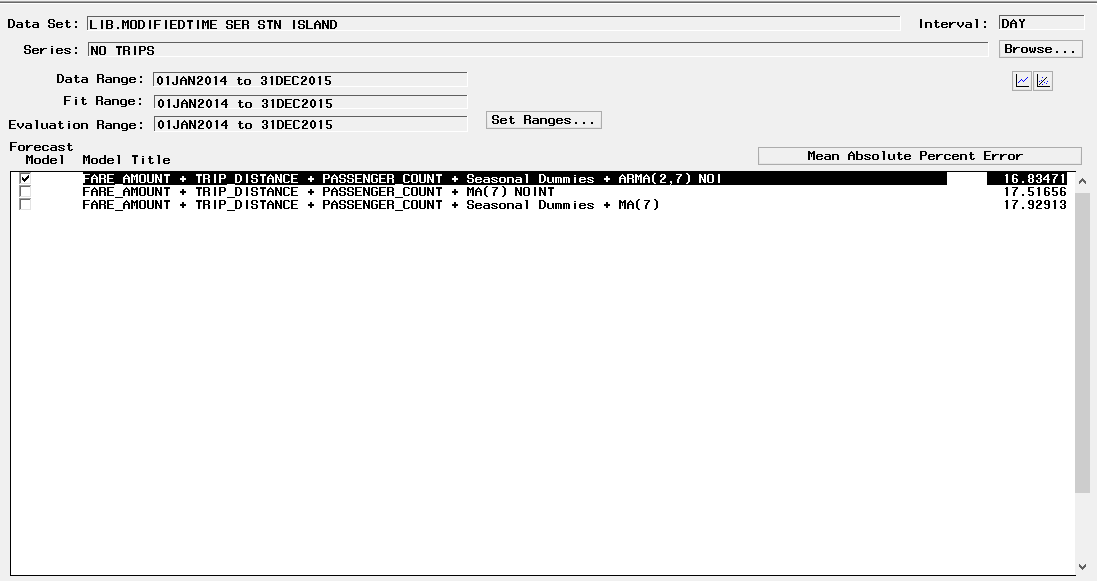


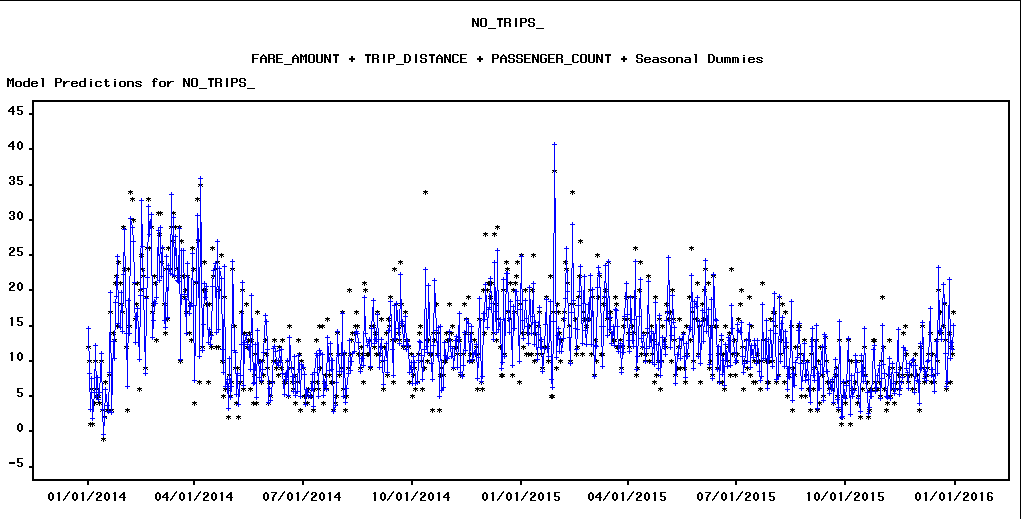


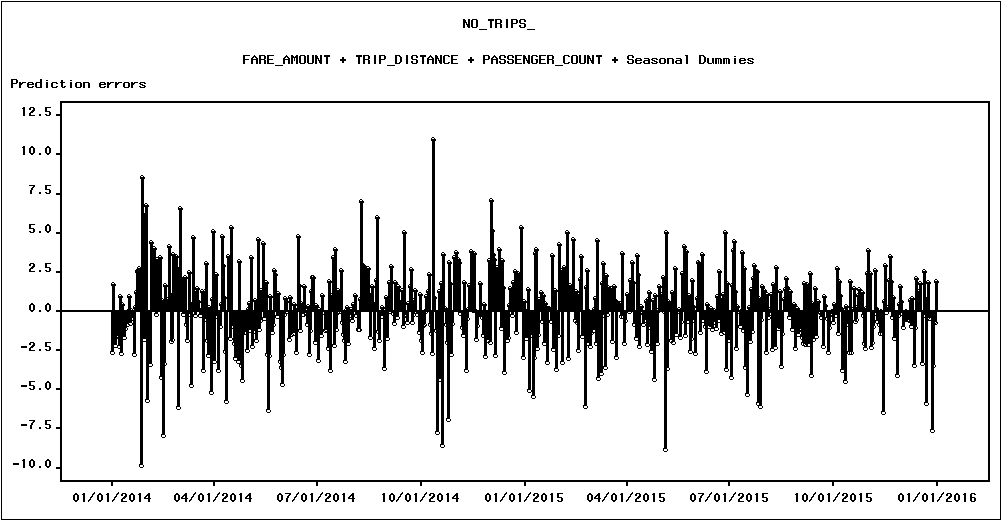


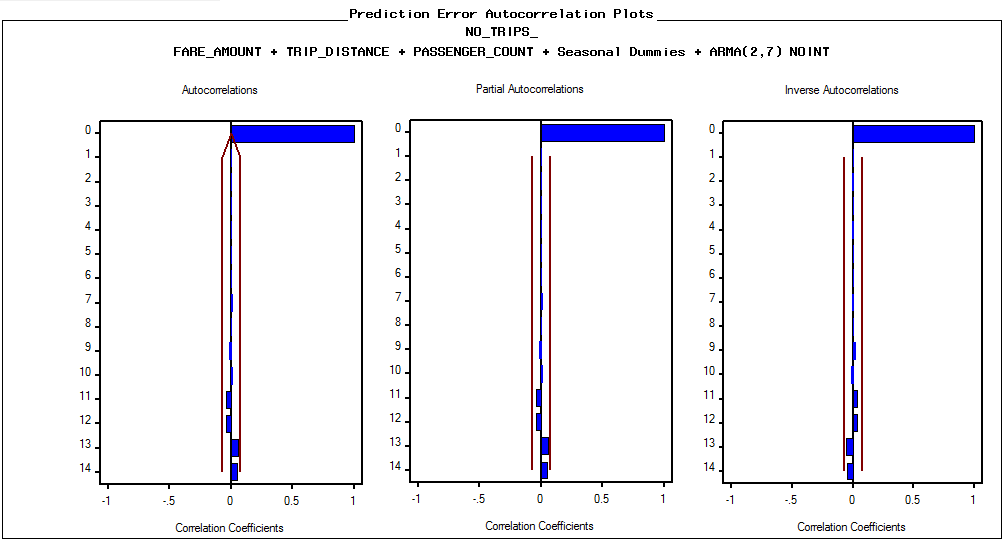


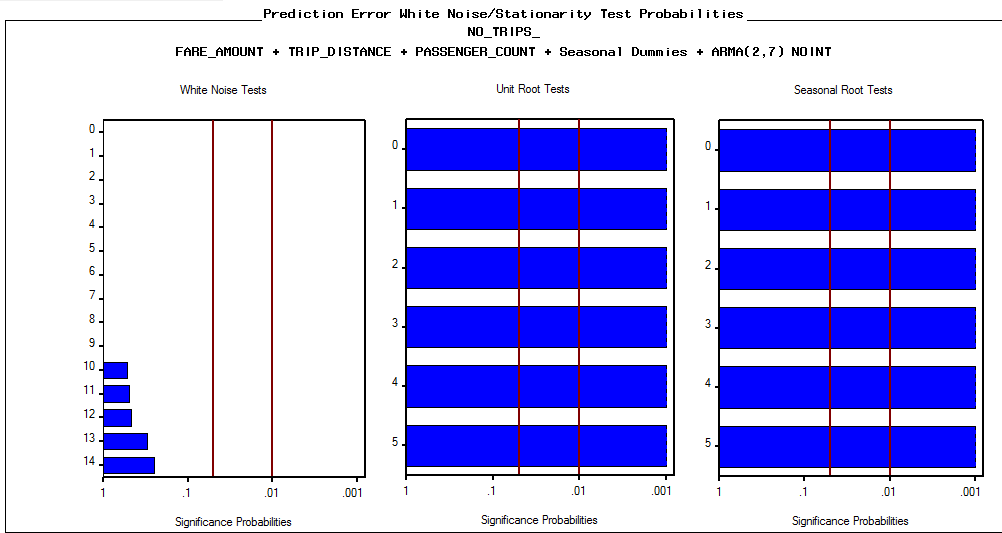


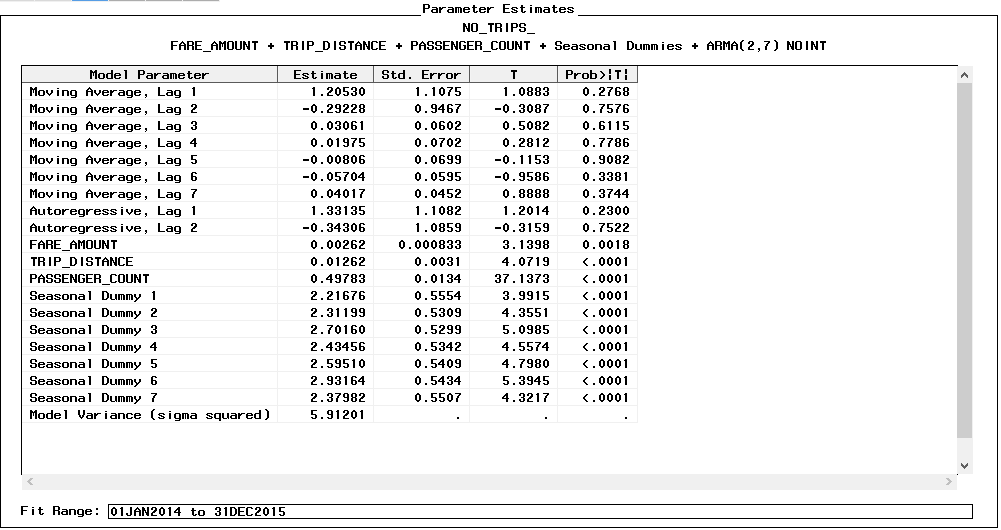


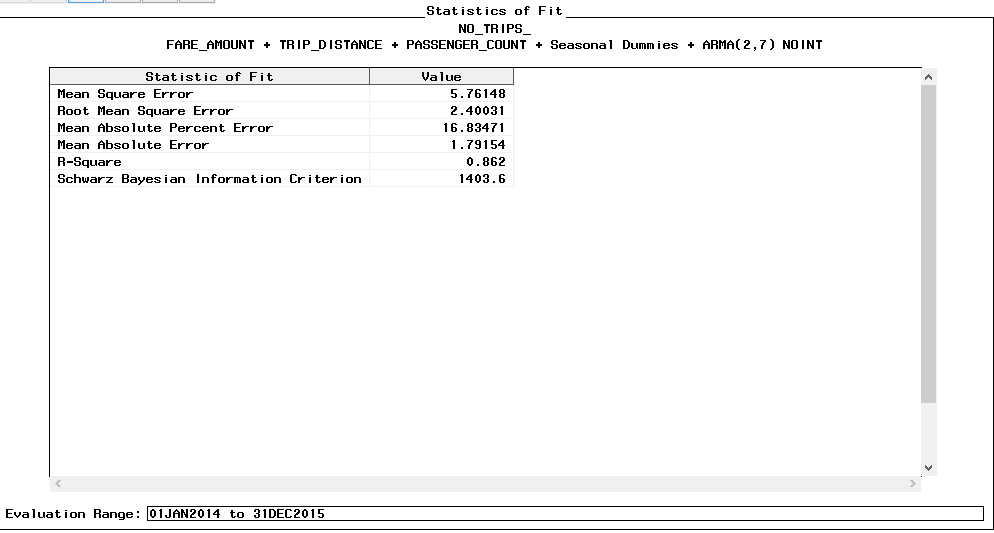


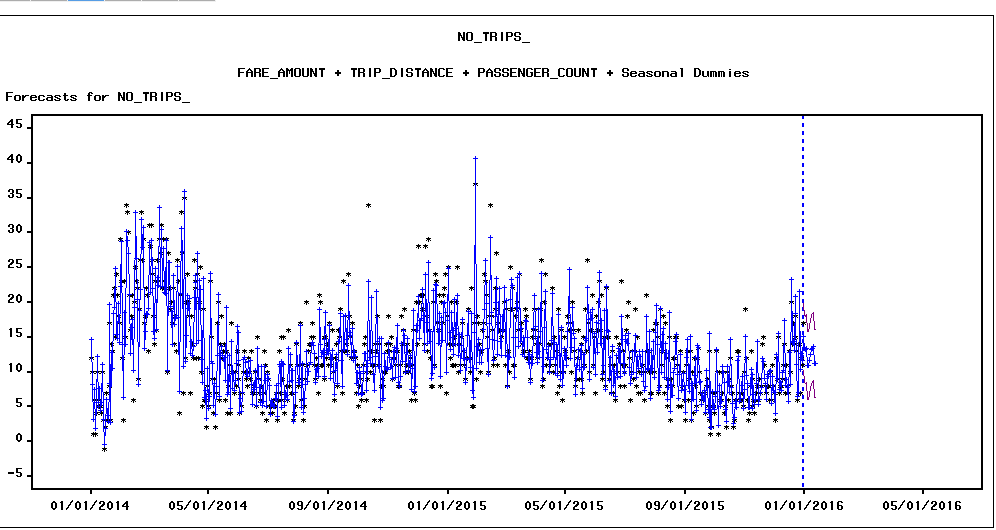




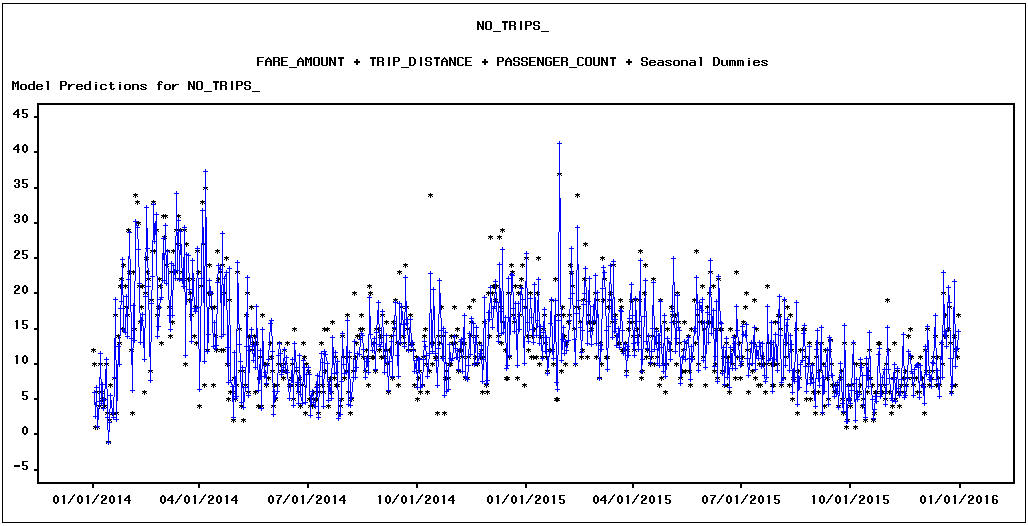


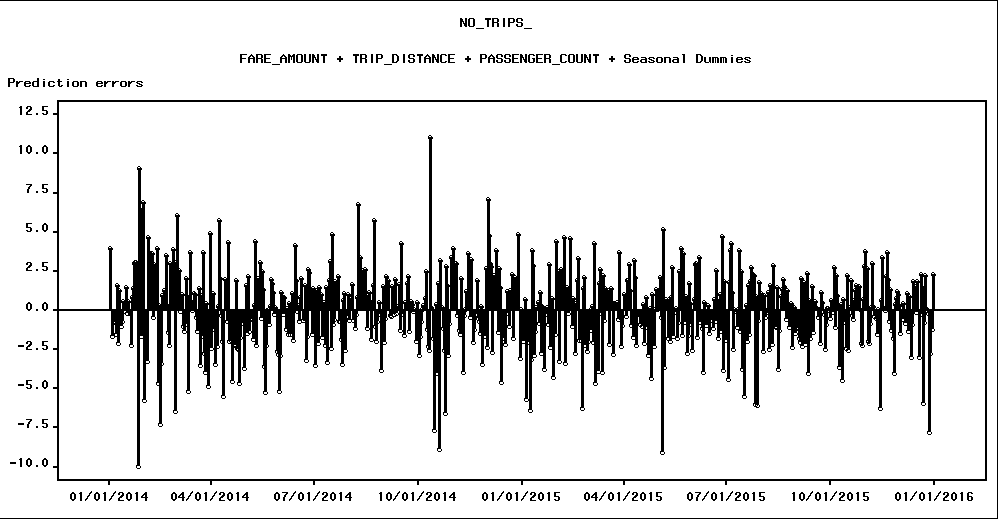


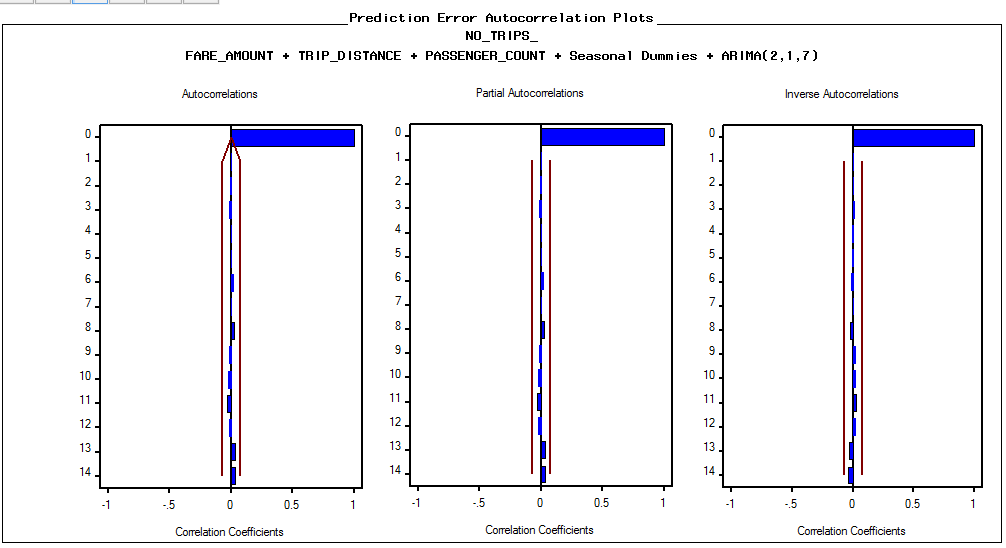


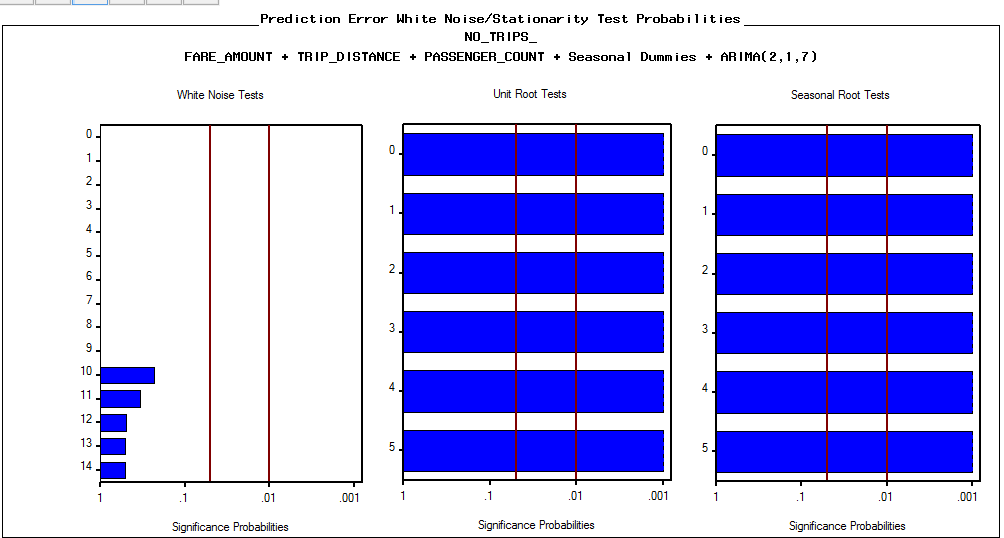


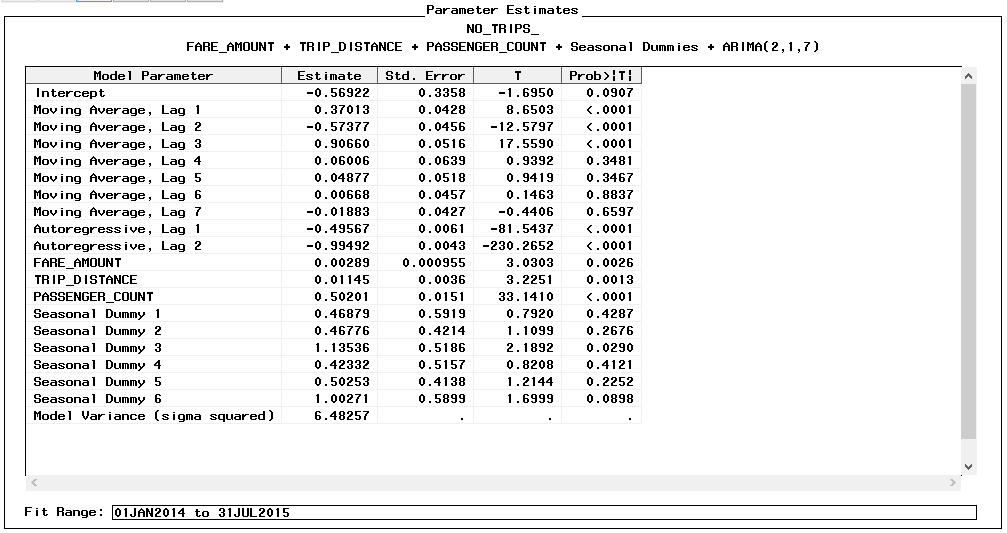
**WITH HOLDOUT SAMPLE:**

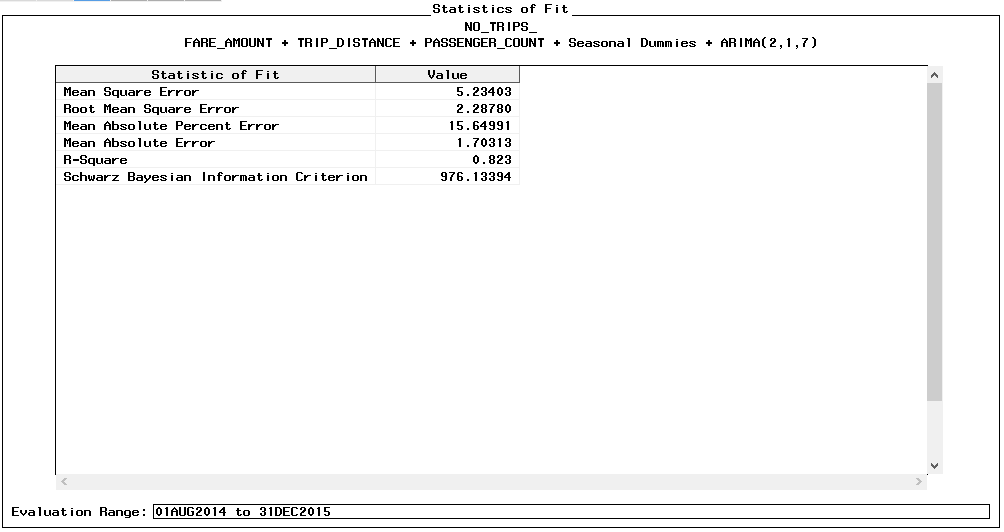


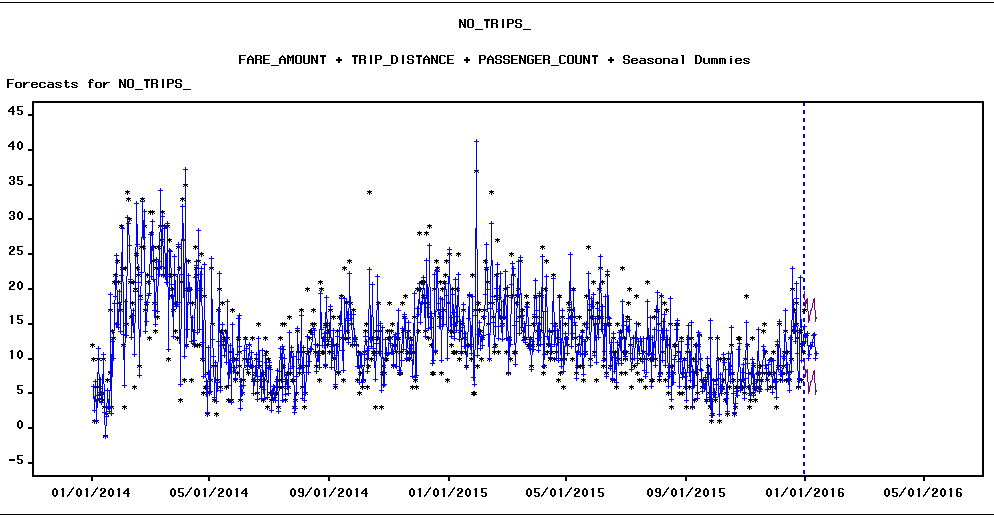




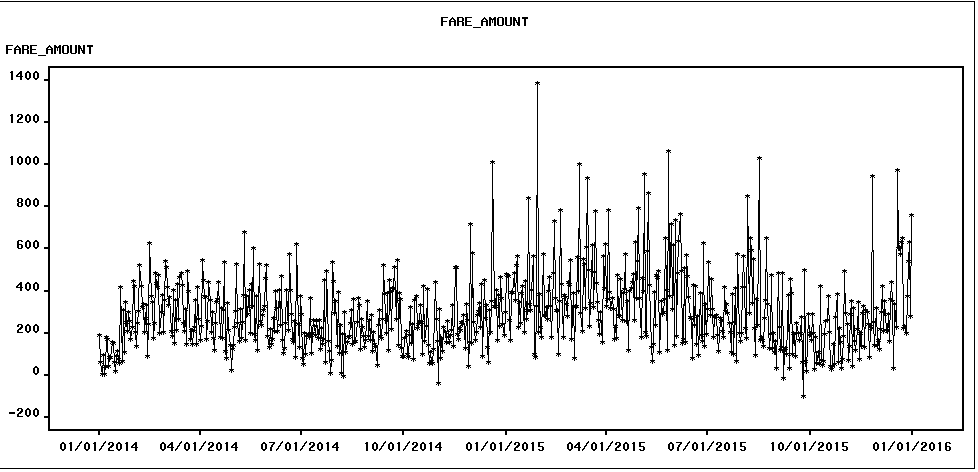


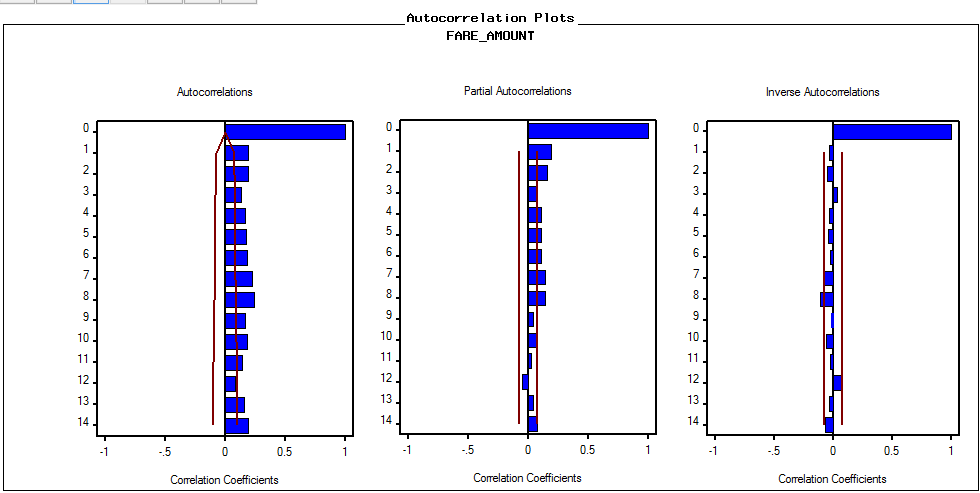


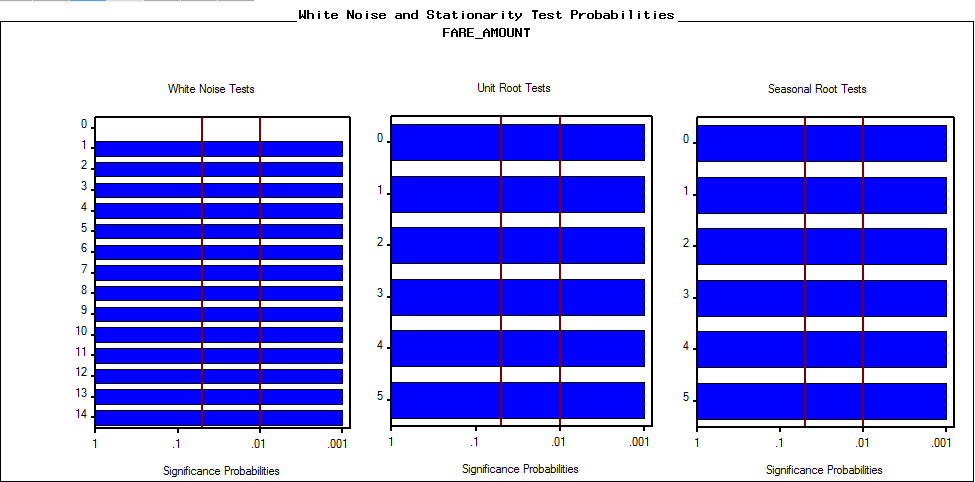


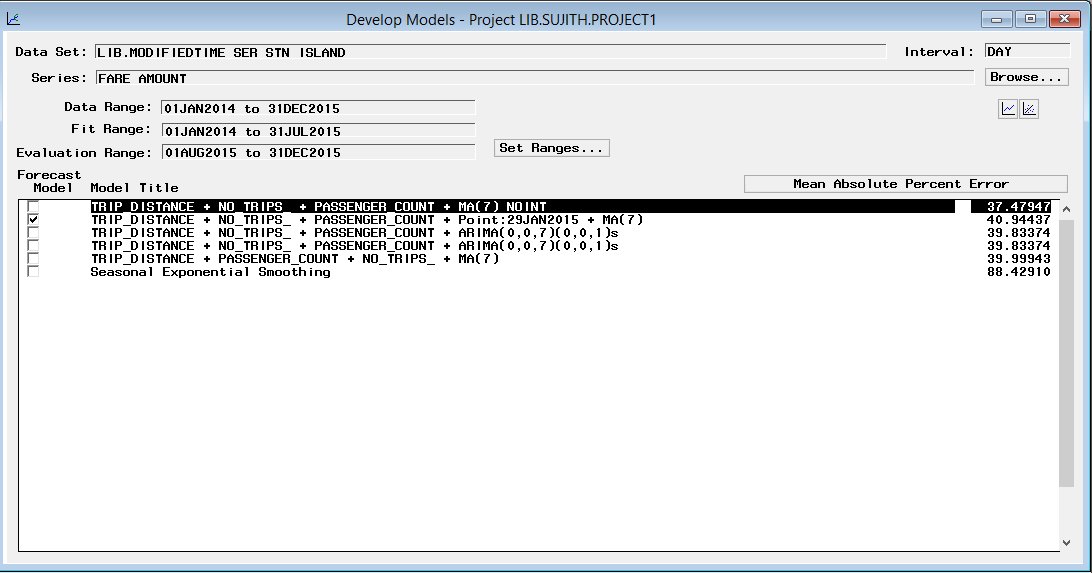


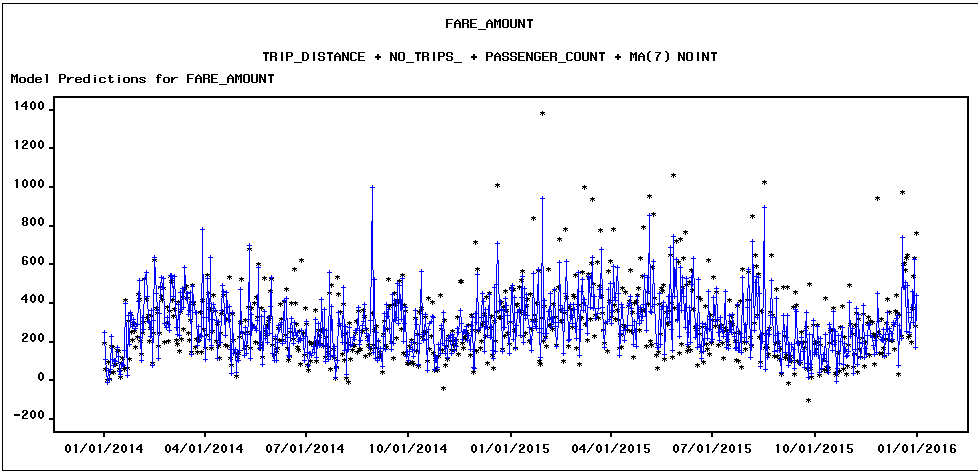
**FARE AMOUNT**:

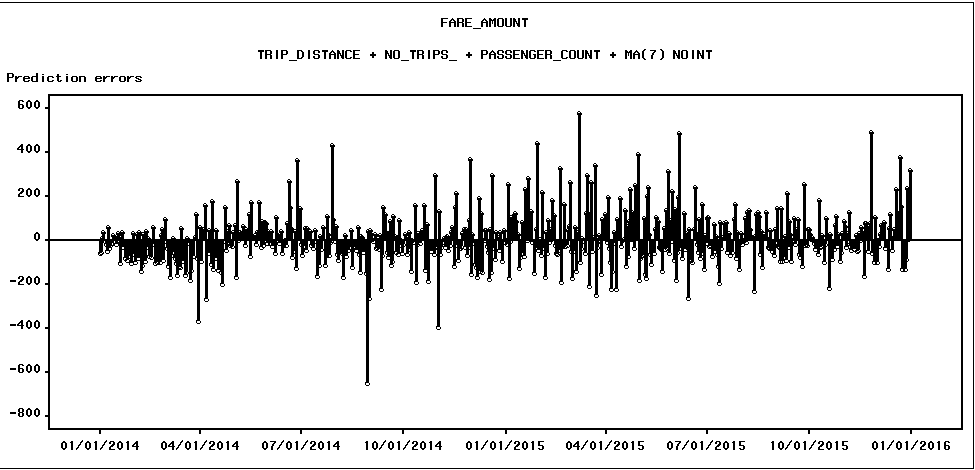


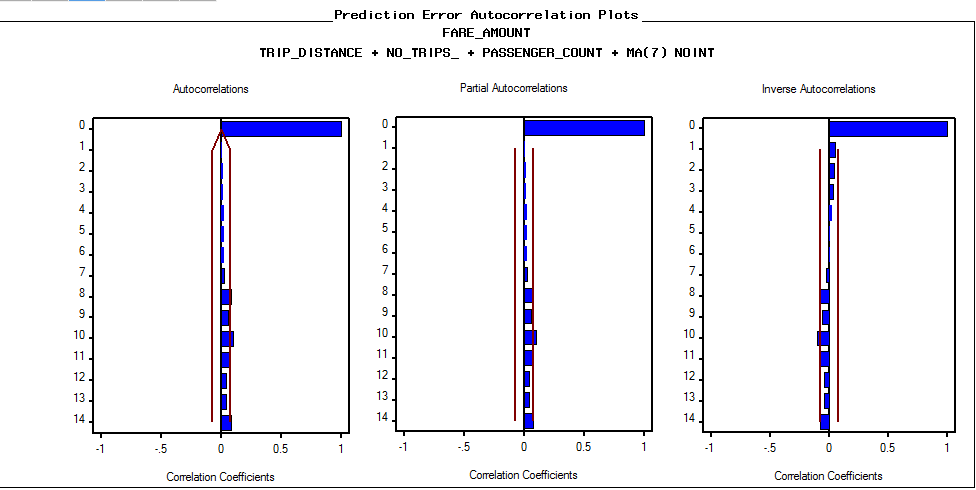


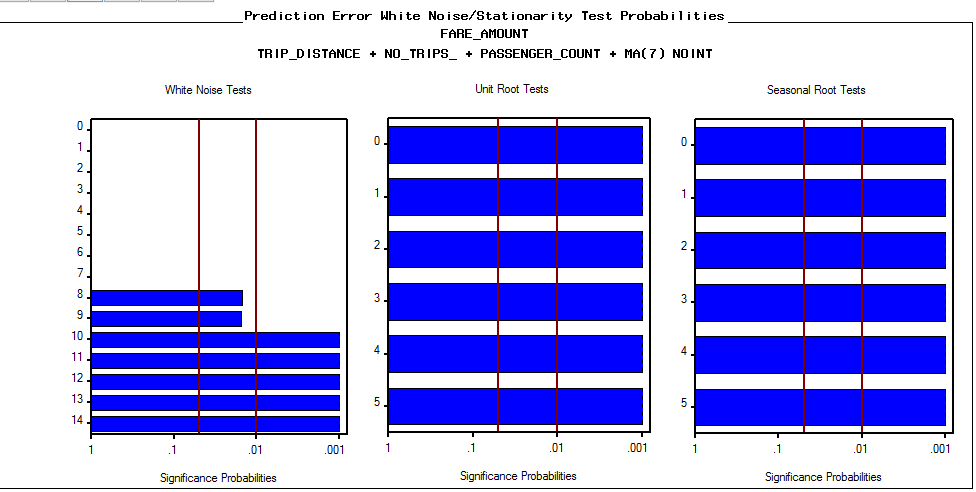




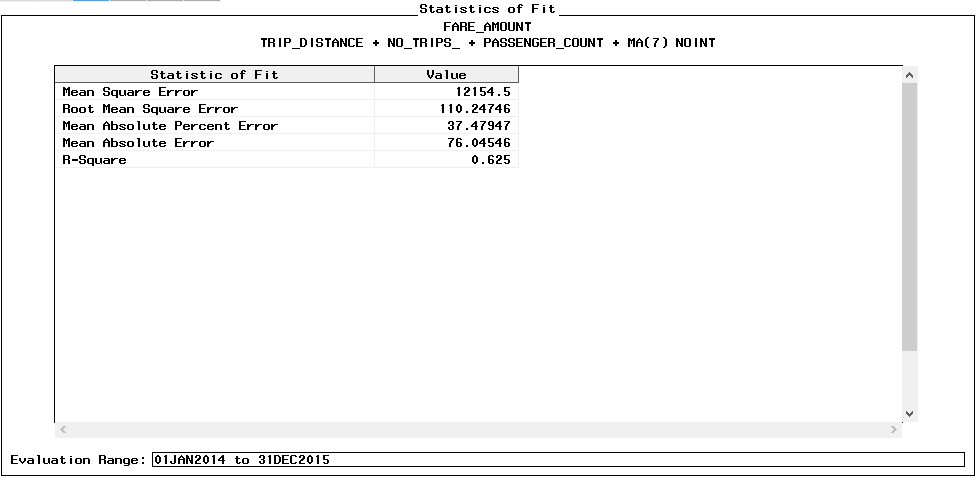


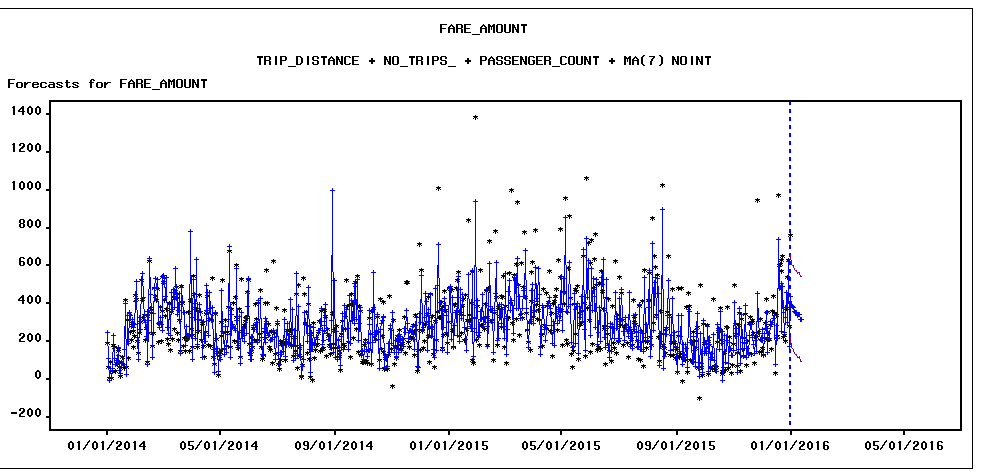












**WITH HOLDOUT SAMPLE**

