ASSIGNMENT SUBMISSION FORM

Treat this as the first page of your assignment

Course Name: *BSFC*

Assignment Title: *Tourism Demand Forecasting - Sikkim*

Submitted by:

(Student name or group name)

Group Member Name	PG ID
Palash Borah	*61210086*
Saurabh Agarwal	*61210054*
Varun Sayal	*61210006*
Dipayan Dey	*61210091*
Abhishek Kumar	*61210131*

(Let us not waste paper, please continue writing your assignment from below)

Contents

Contents 2
Executive Summary 3
Data 4
Stakeholders 6
Goal 6
Naïve Forecast 7
Visualization 8
Methods 10
Choice and Performance 11
Final forecast and prediction intervals 13
Key learning and observations from the Project 14
Exhibits 15

Executive Summary

Problem Description – The objective of the forecasting is to enable Sikkim Government (and other stakeholders) to do forecasts for the next 12 months for state of Sikkim, month after month.

The data source for the analysis was the official website of department of tourism, Govt of Sikkim. We got monthly tourist visits from Jan 2005 to May 2011. The data was available in form of two time series one for domestic tourist visiting Sikkim and other for foreign tourists visiting Sikkim. The domestic time series had an upward trend with yearly seasonality. The foreign time series did not have a trend but there was six month seasonality.

Model Description - Final model is Multiple Linear Regression (MLR) for both domestic and foreign time-series, which is widely used in prediction modeling and statistics. We have used a multiplicative version of this model i.e. Demand = Fac1 * Fac2 * Fac3 * Fac4

Model Performance - Our model performs much better than the Naïve forecasts, i.e. accepting previous K months forecast as next months forecast. This value K was 12 in case of domestic naïve and 6 in case of foreign naïve. Looking at the graphs of actual vs predicted forecasts we saw that predictions from our model fitted very well with the actual values and captured any important changes.

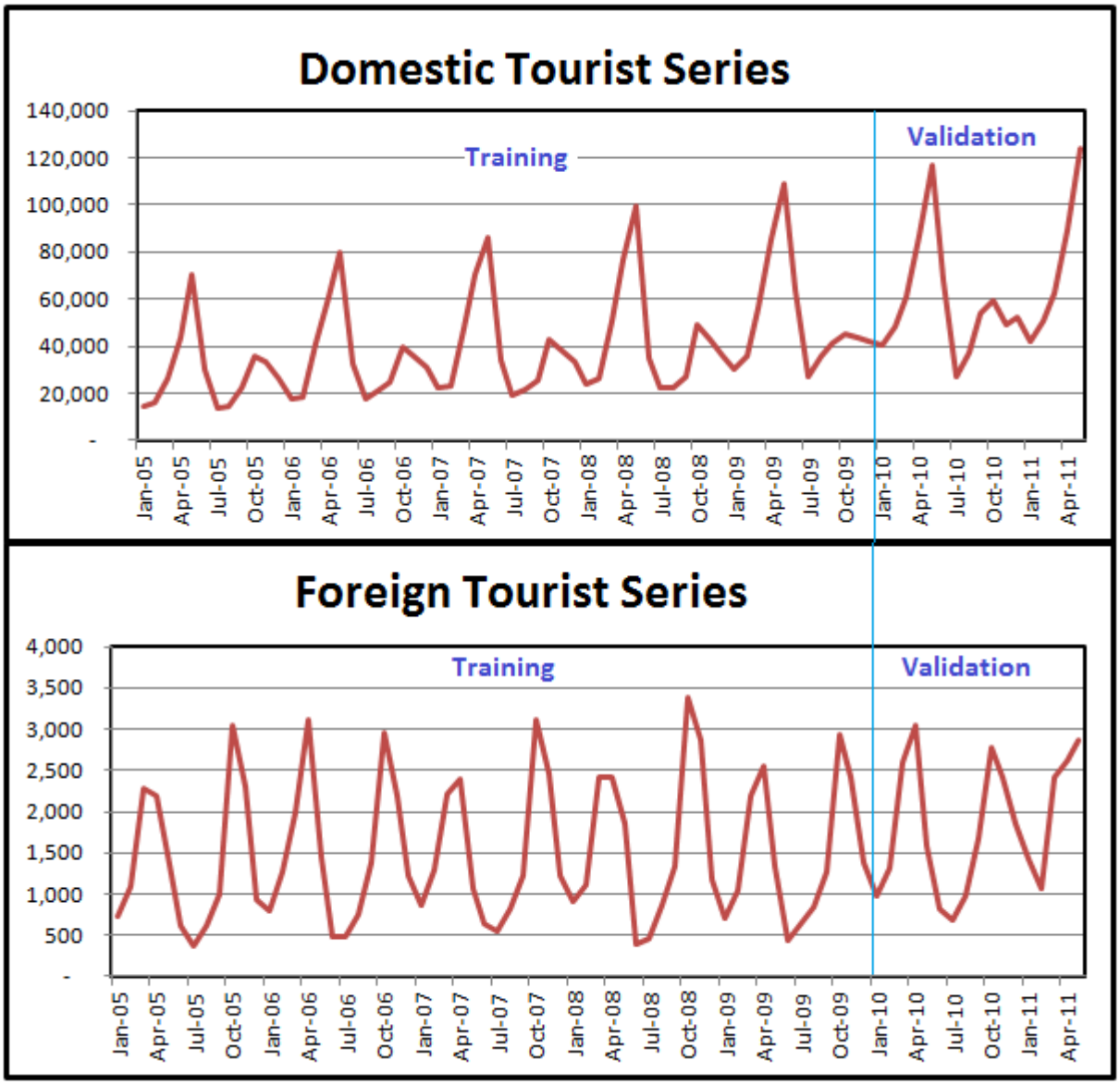
Forecasts and their assumptions - We generated 17 months forecast in future along with their confidence intervals, i.e. the interval between which the forecast could vary. Some key assumptions for our forecasts are, Firstly data for at-least 12 months back is available for forecasting**,** secondly there won't be any huge macroeconomic changes in the world economy.

Conclusions & Recommendation

The final forecasting model recommended is the multiple linear regression model mentioned above. Secondly, we need to ensure that we have the latest data available while generating the forecast. This is based on the assumption that the govt. agencies and other stake holders preparing this forecast will have access to latest data which may not be published on the website. In case the data is not available then appropriate amount of error buffer should be built in while planning.

Data

- **Source:** Website of Department of Tourism, Govt. of Sikkim
- **Period:** 77 months data from Jan 2005 to May 2011
- The data was available in for two time series as can be seen from the graphs below:
    - *Domestic Tourist Visiting Sikkim every month*
    - *Foreign Tourist Visiting Sikkim every month*
- **Data Availability Assumption** – The assumption here is that these stake holders will have access to latest demand data. In case the latest data is not available then the forecasts might have more errors and should be factored in while planning.
- **Data Partitioning** – As shown below, data partitions were made after December 2009. So training set had 60 records and validation set had 17 records for first analysis.


Domestic Tourist Series
Training
Validation
140,000
120,000
100,000
80,000
60,000
40,000
20,000
-
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11
Apr-11
Foreign Tourist Series
Training
Validation
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
-
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11
Apr-11

Stakeholders

Government of Sikkim	• Capacity Planning • Tourism Advisory
Hotel Owners	• Capacity Planning • Pricing
Tourist Service Providers	• Capacity Planning • Pricing

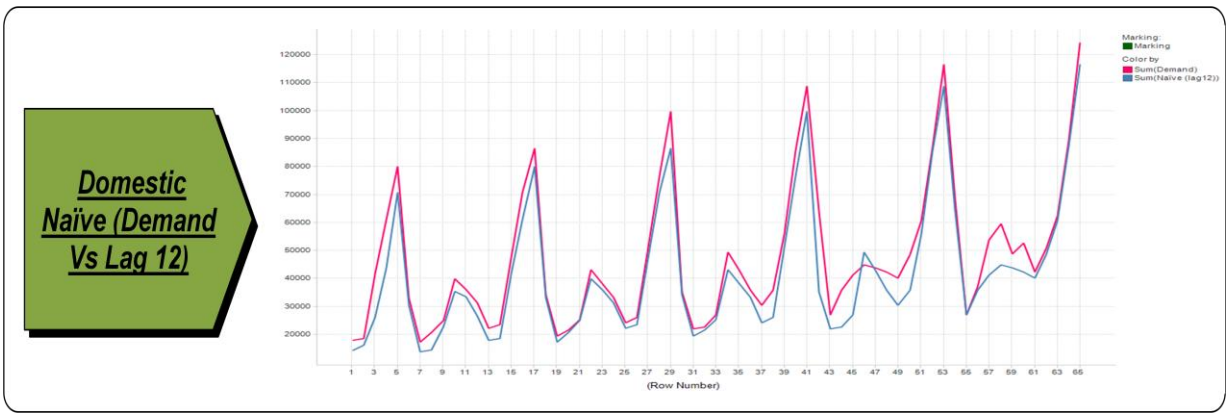
Goal

The objective of the forecasting is to enable Sikkim Government (and other stakeholders) to do monthly rollover forecasts, so that they can predict monthly k-step tourist visit forecasts (both domestic and international) for the next 12 months for state of Sikkim.

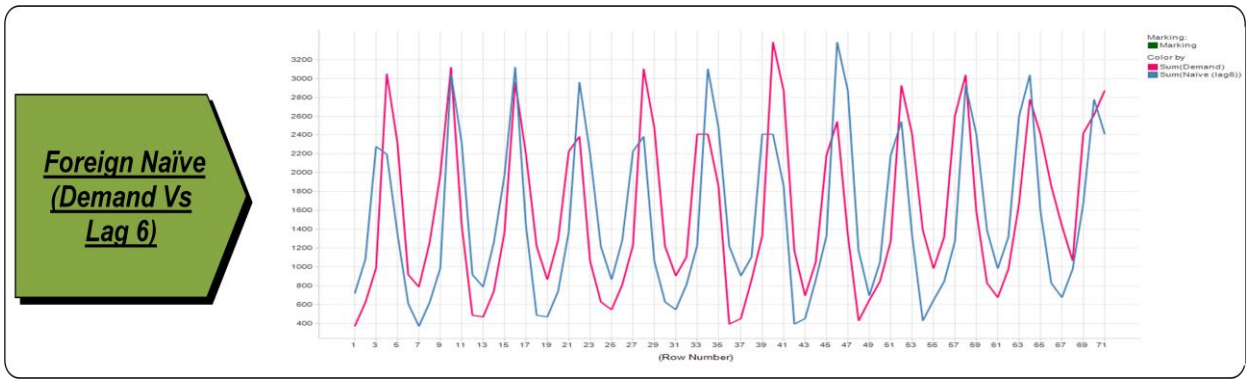
Another alternative was forecasting peak-period tourism demand only, but we decided that a k-step forecast would be better since the monthly data is being tracked and k-step covers all periods.

Naïve Forecast

Domestic Naive Forecast – The following series seems to have an upward trend with yearly seasonality. Therefore the naive forecast method uses last year demand to forecast the next years demand.



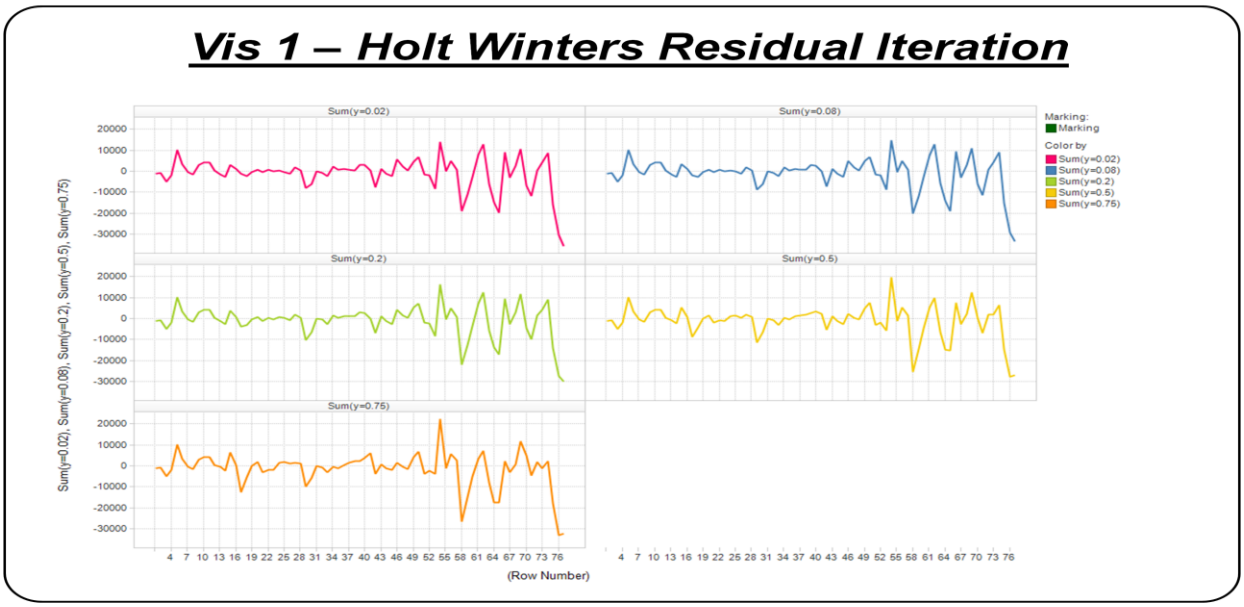
Foreign Naive Forecast – While visualizing the Foreign Tourist series it appeared to be following 6 month seasonality without any trend. Naive Demand Forecast with a lag of 6 months doesn't seem to give very accurate forecasts and the Error metrics (MSE & MAPE) also support this fact.



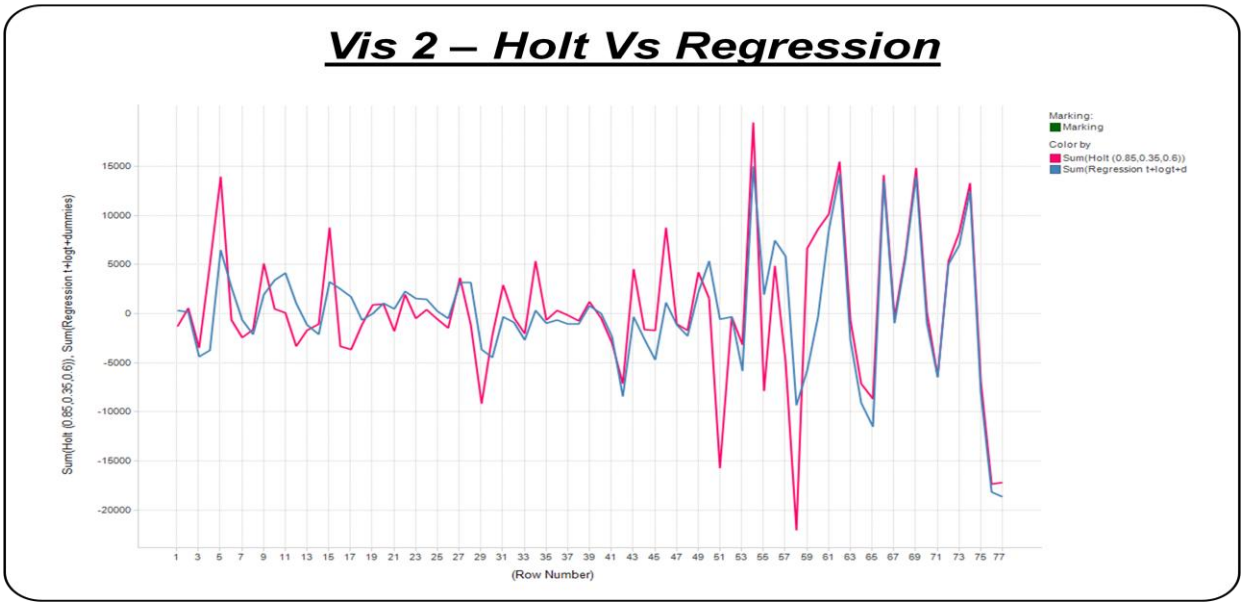
	MSE	MAPE
Domestic Naiive	59476273.37	12.92
Foreign Naiive	527468.55	51.05

Visualization

Visualization 1:

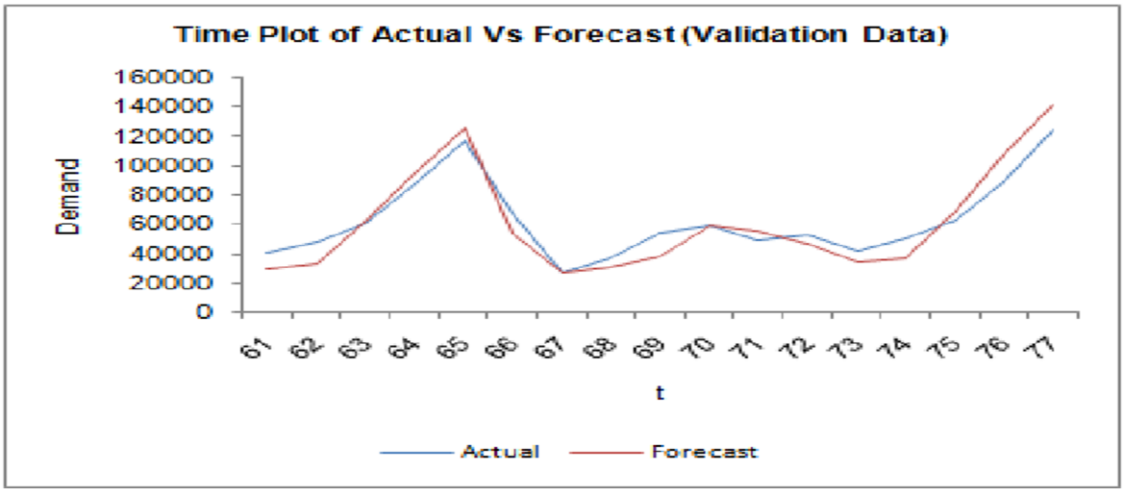


Visualization 2:



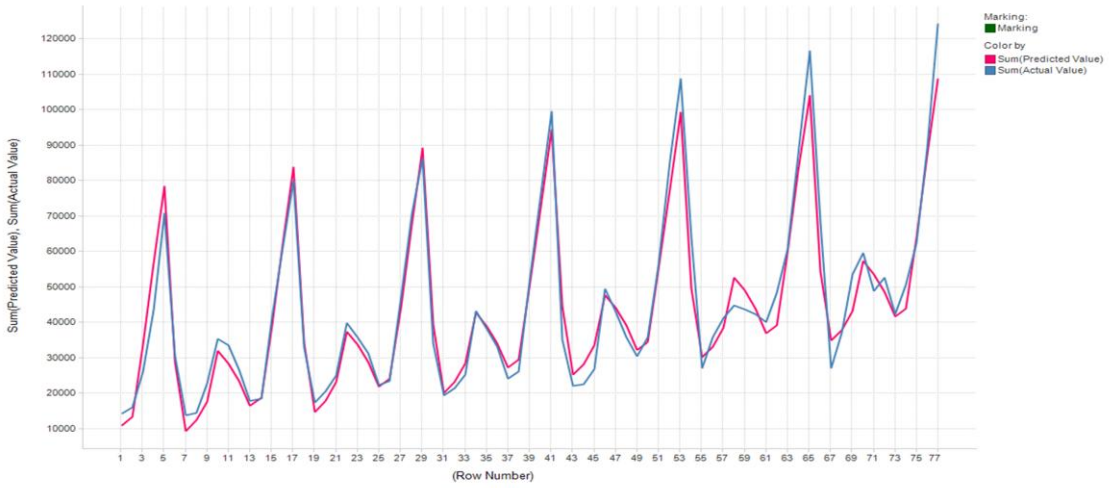
Visualization 3:

Vis 3 – Holt Winters on Validation Set



Visualization 4

Vis 4 – Holt Winter Actual Vs Predicted



Methods

Linear Regression

- We carried out a linear regression of **Demand** Vs **t, t^2, lag12, monthly dummies**
- We tried different combinations, rejected this method, due to a very clear seasonality in residuals

Linear Regression (Multiplicative)

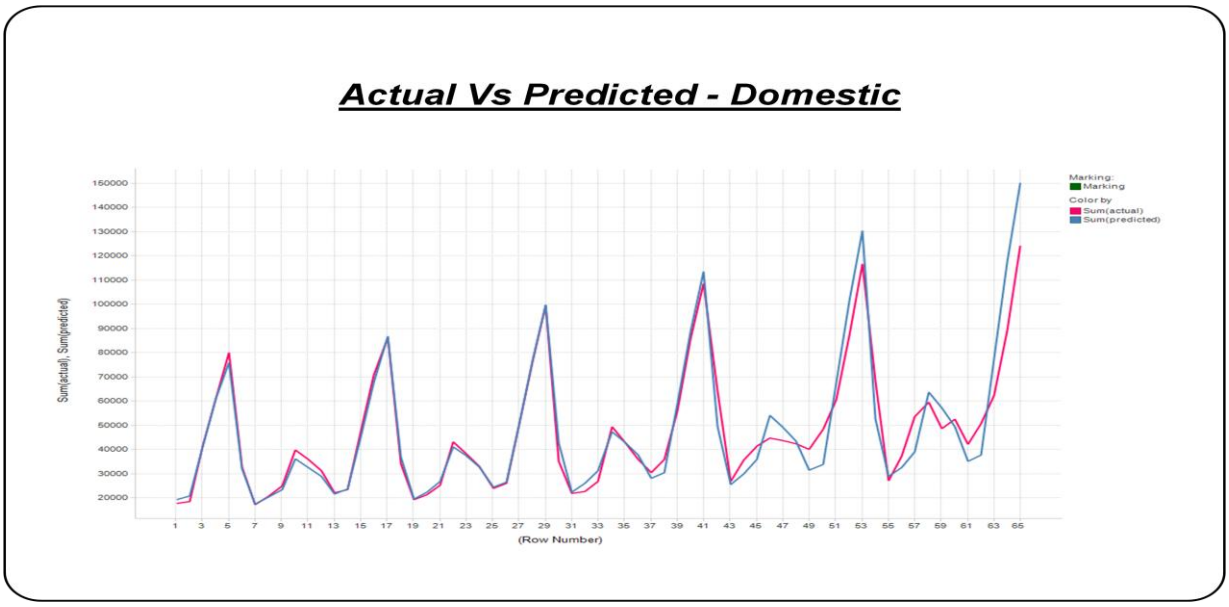
- We regressed **log(demand)** Vs **t, t^2, log(lag12), monthly dummies**
- We again tried different combinations, stuck to taking **t, log(lag12) and monthly dummies** for domestic and **t and monthly dummies** for foreign

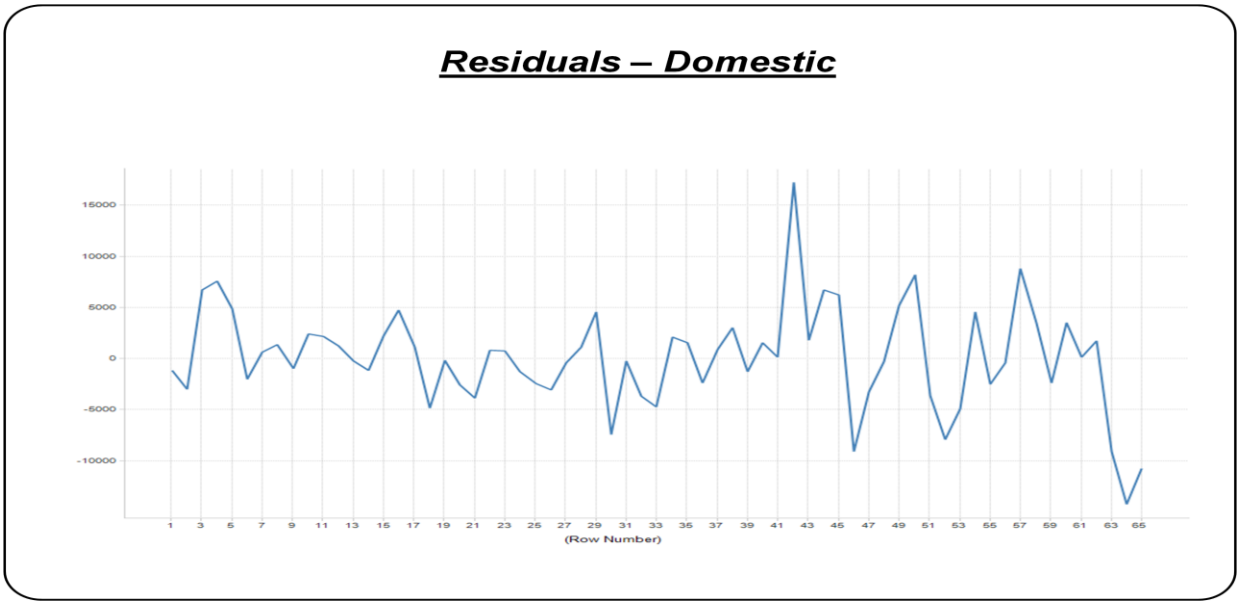
Holt Winter's Method

- For domestic series we tried around 20-30 combinations and finally decided upon; **α = 0.85, β = 0.35, ϒ = 0.6** for domestic series as a good candidate.
- For foreign series initial results with **α = 0.2, β = 0.15, ϒ = 0.05** were not very promising so it was rejected outright

Choice and Performance

Domestic: log Demand = β0 + β1 * t + β2 * log (lag12) + β3 * D1 + β4 * D2 + β5 * D3 . . . . . + β13 * D11

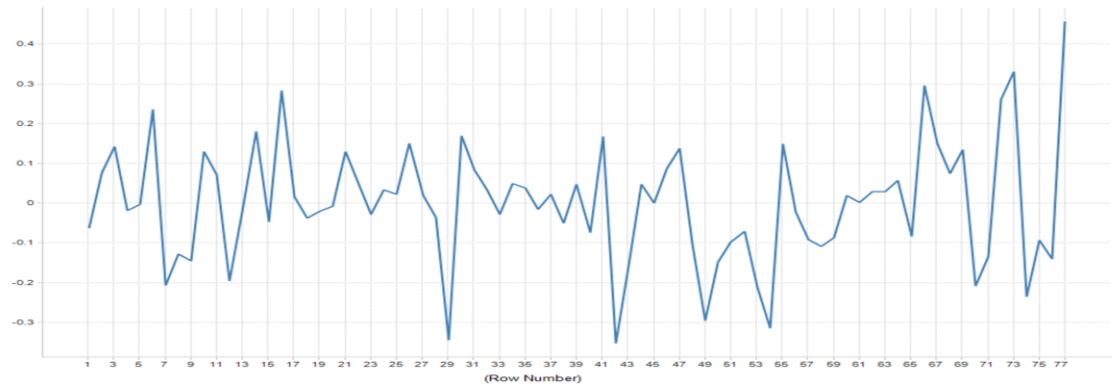




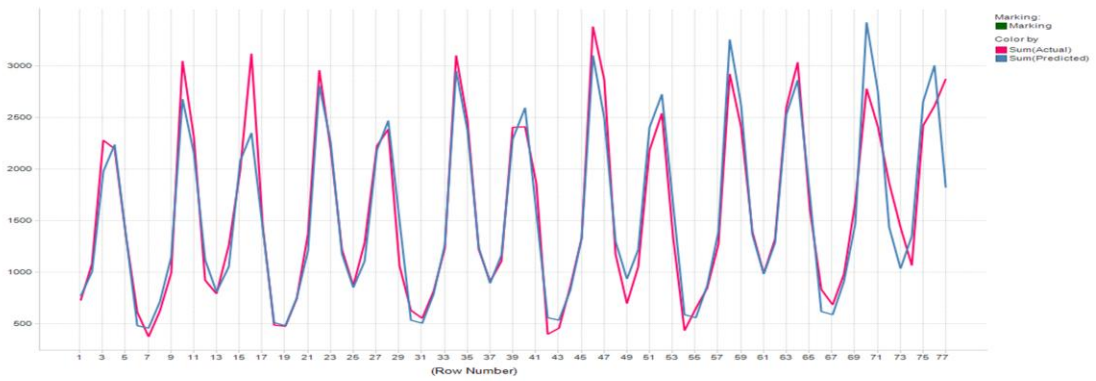
Final Model: MSE: 24628680.97 MAPE: 7.94

Foreign: log (Demand) = β0 + β1 * t + β2 * D1 + β3 * D2 + β4 * D3 . . . . . + β12 * D11

Residuals – Foreign

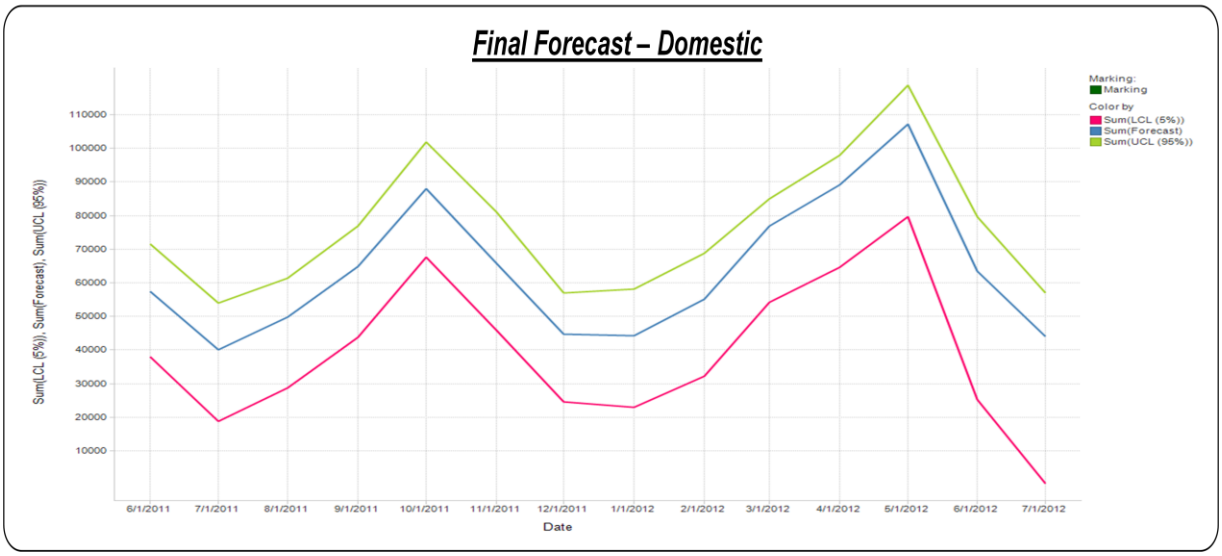


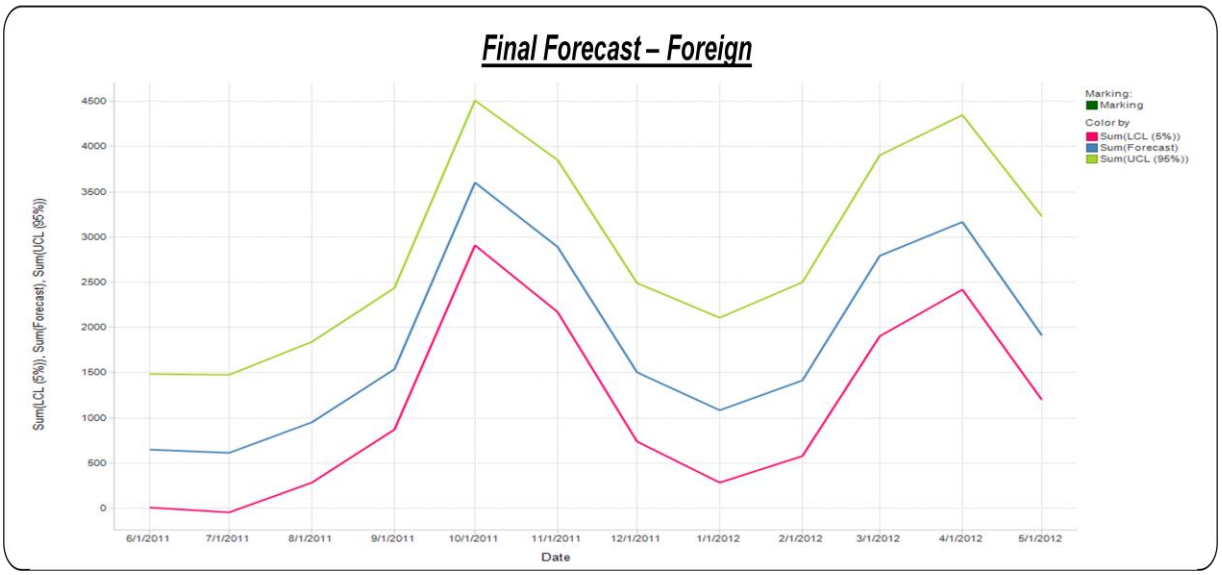
Actual Vs Predicted - Foreign



Final Model: MSE: 60667.99 MAPE: 11.56

Final forecast and prediction intervals

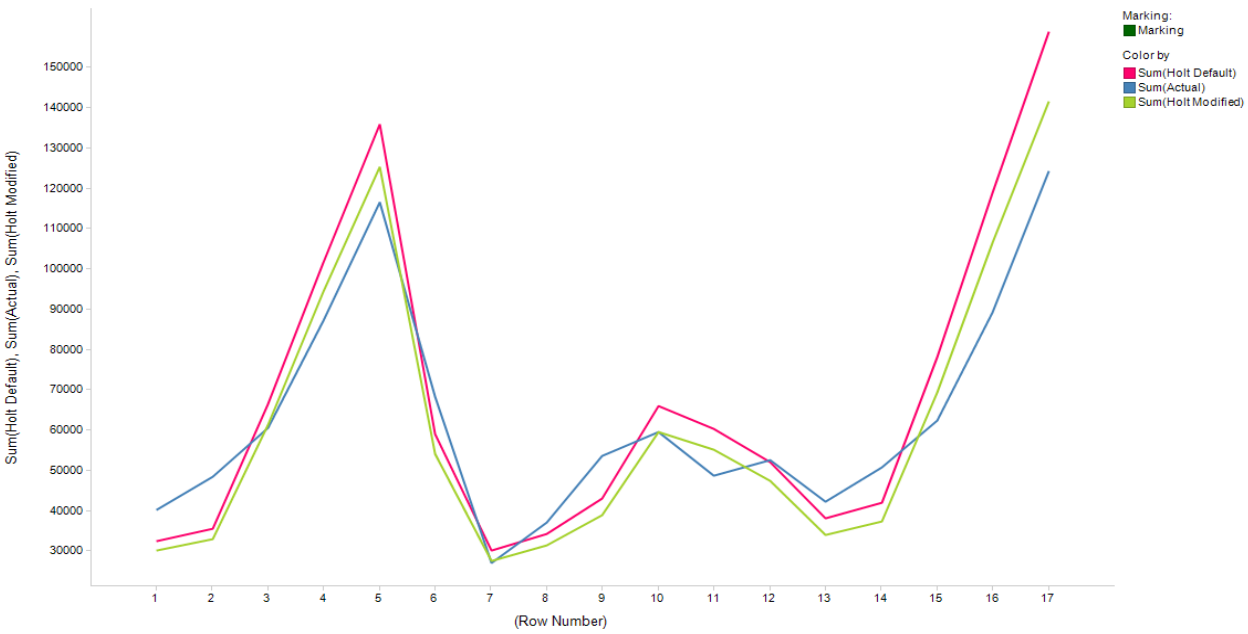




Key learning and observations from the Project

Final Model chosen: Regression was our final model for both Domestic and Foreign Series, as described above in detail.

Possible Alternate model: Holt winter's was a possible alternate with significantly high values for alpha, beta and gamma. This was done because we wanted the model to learn quickly and not to fit the actual vs predicted very closely. There is indeed a global pattern but towards the end there are data points that defy the global pattern, this is where Holt winter's method seems very promising, as it can learn quickly and take into account the sudden variations, if any.



Above chart is for validation set from domestic series and comparison is between Actual values (blue), Holt Default (pink) and Holt modified (green). As you can see from the overlaid chart at the point 10 the modified Holt quickly learns of a dip and captures the dual local peak very well, but the Holt with default values fails to capture that peak.

So over-fitting will not be an issue here as these parameters will not be updated all the time to suit data, but will help the model to learn quickly and grasp localized patterns. In any case this was not the final model we chose, but just an after-thought of our analysis.

Comparison between Domestic and Foreign series

We created a overlaid MA(12) trend-line chart of domestic vs foreign series, and tried to compare them on multiple scales in one chart as below:

It is evident from the above chart that internally both series display a rather mixed correlation at different times, for some time periods they moved together and for some they are totally opposite.

Exhibits

Exhibit 1 - Domestic Tourism Forecast – Iterations for generating k step Error Residual

Residuals	Iter1	Iter2	Iter3	Iter4	Iter5	Iter6	Iter7	Iter8	Iter9	Iter10	Iter11	Iter12	Iter13	Iter14	Iter15	Iter16	Iter17
step 1	8529.9	13192	-9735	-17010	-13816	9846.8	-3387	713	11347	1213.8	-5477.3	3187.7	53.54	2398	-13592	-19963	-14093
step 2	14495	-8342	-18060	-15346	9851.5	-2965	370.48	11433	958.2	-5324.2	2868.1	138.27	2406.6	-13531	-21562	-16328	
step 3	-7091.8	-15977	-16727	9318.1	-3117	1545.4	11094	871.8	-5338	2945.4	183.42	2480.2	-13528	-21589	-18387		
step 4	-14107	-15504	8837.1	-3441	1512.3	12039	716.94	-5383	3294.7	152.14	2670.4	-13260	-21587	-18423			
step 5	-13852	13082	-3733	1175.2	11909	-347.2	-5580	3312	1129.2	2593.7	-13876	-21162	-18420				
step 6	15614	-2711	871.08	11485	-1078	-6149	3043.8	1244	3937.5	-13725	-22268	-17873					
step 7	-1959	3196.6	11103	-1865	-6726	3188	825.41	4104	-13132	-21986	-19297						
step 8	4619.9	13127	-2576	-7418	2816.1	2084.6	3573.5	-13108	-21471	-18934							
step 9	14463	-3961	-8041	2242	1967.1	5409.6	-13589	-21492	-18270								
step 10	-4213.6	-8549	1724.1	1486.1	5367.7	-13411	-22111	-18298									
step 11	-8378.4	2455.2	1051.4	4856.2	-14116	-22496	-19094										
step 12	3220.7	3656.9	4394.1	-15182	-23680	-19590											
step 13	7039.4	10620	-18523	-29455	-24407												
step 14	12938	-16974	-31212	-26891													
step 15	-15322	-30113	-29141														
step 16	-28261	-28068															
step 17	-25873																

Exhibit 2 – Foreign Tourism Forecast – Iterations for generating k step Error Residual

Residuals	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	Iter 6	Iter 7	Iter 8	Iter 9	Iter 10	Iter 11	Iter 12	Iter 13	Iter 14	Iter 15	Iter 16	Iter 17
Step 1	134.2	71.6	205.6	295.3	74.1	279.3	132.2	106.9	280.1	-743.2	-355.0	539.7	478.1	-337.9	-244.6	-418.6	1226.4
Step 2	82.3	213.0	307.1	81.9	280.5	142.3	115.1	286.9	-717.5	-380.3	532.9	491.7	-311.1	-278.3	-431.4	1216.1	
Step 3	233.6	315.5	88.4	283.3	143.5	131.0	299.9	-700.7	-360.0	520.7	486.1	-291.6	-226.7	-469.8	1209.2		
Step 4	338.9	93.0	285.7	146.3	133.0	325.3	-668.1	-346.7	530.5	475.9	-299.7	-189.2	-410.8	1188.2			
Step 5	105.9	287.4	148.6	137.3	328.4	-604.4	-320.9	536.9	484.1	-314.3	-204.7	-368.0	1220.4				
Step 6	292.1	150.2	140.9	335.3	-596.8	-270.5	549.3	489.4	-302.6	-232.8	-385.8	1243.8					
Step 7	154.7	143.5	341.1	-579.5	-264.4	573.6	499.7	-294.9	-210.3	-417.8	1234.1						
Step 8	150.7	345.2	-565.0	-250.7	576.5	519.9	-280.1	-195.5	-392.1	1216.6							
Step 9	356.7	-554.7	-239.3	583.0	522.3	-251.2	-166.9	-375.2	1230.6								
Step 10	-526.0	-231.1	588.6	527.7	-247.7	-111.2	-342.6	1239.9									
Step 11	-208.4	592.5	532.3	-239.9	-104.5	-279.0	1257.7										
Step 12	603.4	535.6	-233.3	-89.5	-271.3	1292.5											
Step 13	566.1	-216.6	-42.8	-205.7	1309.2												
Step 14	-202.0	-32.6	-189.4	1319.8													
Step 15	-4.6	-177.9	1328.8														
Step 16	-145.9	1335.1															
Step 17	1352.7																

Exhibit 3 – Domestic Series, Forecast with prediction intervals

Date	Forecast	LCL (5%)	UCL (95%)
Jun-11	57450.96	37969.16	71451.57
Jul-11	40095.03	18873.95	53949.01
Aug-11	49909.65	28681.71	61361.96
Sep-11	64846.90	43780.90	76986.97
Oct-11	88113.69	67661.07	101832.52
Nov-11	65762.50	45763.43	81160.14
Dec-11	44620.12	24473.43	57030.72
Jan-12	44247.55	23013.30	58024.62
Feb-12	55025.03	32100.21	68778.49
Mar-12	76880.20	54157.80	85101.51
Apr-12	89245.70	64519.46	98051.44
May-12	107121.07	79713.99	118801.48
Jun-12	63401.07	25231.28	79680.40
Jul-12	44123.94	248.24	56930.68

Exhibit 4 – Foreign Series, Forecast with Prediction Intervals

Date	Forecast	LCL (5%)	UCL (95%)
Jun-11	648.04	11.56	1487.36
Jul-11	614.51	-43.96	1476.88
Aug-11	954.48	287.21	1843.41
Sep-11	1540.92	872.17	2438.57
Oct-11	3599.46	2909.74	4510.19
Nov-11	2894.03	2169.22	3854.27
Dec-11	1505.60	735.11	2490.60
Jan-12	1083.31	286.55	2105.14
Feb-12	1416.06	580.59	2497.16

Mar-12	2792.36	1905.43	3903.66
Apr-12	3163.14	2420.33	4349.03
May-12	1911.17	1204.81	3229.98