Danielle Chua

daniellelcchua@yahoo.com.sg

Abstract

Forecast models on the sales performance of the retail supermarket stores built using SARIMA to aid in planning, performance monitoring and trigger timely actions.

Capstone project

Forecast Models

**Contents**

[**1** **A typical planning process of an organisation** 2](#_Toc53741217)

[**2** **Business Questions** 3](#_Toc53741218)

[**3** **Objective and Desired Outcome of the Capstone Project** 3](#_Toc53741219)

[**4** **Industry Domain** 4](#_Toc53741220)

[**5** **Stakeholders** 5](#_Toc53741221)

[**6** **Data** 5](#_Toc53741222)

[**7** **Data Analysis** 7](#_Toc53741223)

[**8** **Modelling & Validations** 11](#_Toc53741224)

[**9** **Conclusions** 26](#_Toc53741225)

[**10** **Future works** 26](#_Toc53741226)

**Appendix I - Details of 3 datasets** 27

**Appendix II - WI2 forecast model using full set of data** 29

# **A typical planning process[[1]](#footnote-1) of an organisation**

Every organisation, in particular, the large corporations which have numerous departments and offices spread across different geographical locations, has an overall annual operating plan / budget which all units of the organization have to align with. The overall annual operating plan, broken down into specific targets, are then cascaded down to the respective departments /offices to achieve. The achievement of the department/office’s specific targets would in turn contribute to the achievement of the overall annual operating plan of the organisation.

## **Top-Down Planning Approach**

The planning exercise is usually initiated by the management of the organisation. It typically starts off with looking at the past financial achievements, current operating environment and economic outlook of the country they operate in, to set the financial targets to be achieved next year. Competitors and shareholders’ expectations are also factored in setting the financial targets. Apart from the past financial achievements which are factual, the rest of the considerations such as the impact of competitors, shareholders’ expectations and economic outlook are largely assumption-based estimations. Thus, the financial targets tend to be expressed as a numerical growth projection such as 10% growth over prior year’s result. This financial target may be accompanied by a descriptive plan detailing the initiatives to achieve such target at the organizational level.

## **Bottom-Up Planning Approach**

In a bottom-up planning approach, the departments and offices would provide their initial plans and financial targets for next year to the management of the organisation. Such plans and targets are consolidated to become the initial organizational level annual plan. The initial organizational plan is then reviewed and revised after taken into consideration of the various economic, competitive and other external factors and the departments’ views, by top management level. The organizational plan is then agreed and finalized.

Some organizations may adopt a hybrid approach of top-down and bottom-up planning approaches depending on the organisation structure and management styles mainly.

## **Targets to be Achieved**

Regardless of the planning approaches applied by the organization, the finalized organizational plan is broken down into specific targets for the departments and offices to achieve. The cascaded targets are usually numerical growth projection or a set of numerical growth projections to be achieved next year. The immediate task for each department or office would be to prepare or revise the next year plan to align with the finalized target. How will the department or office go about with just the numerical growth target/s?

# **Business Questions**

## **How to assess the achievability of the financial target for next year by the department?**

Based on the business knowledge and experiences, the department head may have a gut feel about the achievability of the financial target and proceed to develop the plan on this basis to achieve the financial target for next year.

Alternatively, a common quantitative approach is to set the prior year’s financial results as a baseline for next year or derived based on the run-rate of the past months’ results or average of the past months’ results, and strip off any one-off events. With the baseline, the gap to the financial target is determined. The department’s plan is then developed to close the gap accordingly.

Although the quantitative approach is better than the gut-feel approach, it may not appropriately weigh the performance drivers especially if their impact is not material in the prior year’s results. In today’s ever-changing business environment, the drivers of the business are dynamic.

## **How to track the performance of the department on an ongoing basis against the financial target?**

The business plan is usually prepared annually and updated on a quarterly or half-yearly basis to incorporate any changes during the year. The department would use such plan to track its performance on a monthly basis and explain any variances, risks and opportunities to the plan. The department may also prepare a more updated forecast based on feedback from different sources, in the middle of every month to track the progress closely.

This seems like a full-time planning role. The department time would be better spent in running the business.

## **Is there a better way address the questions?**

What if there is a forecast model that is able to predict the next year outcome based on a set of known variables and historical data? Will this forecast model give a better baseline than the approach described in Section 2.1? Will the identified gap be closer to the financial target?

What if there is another forecast model that is able to predict the next 28 days of the business performance based on a set of known variables, historical data and the latest data? Will this data-driven forecast model give a more realistic prediction of the next 28 days performance than the assumption-based forecast?

# **Objective and Desired Outcome of the Capstone Project**

The objective of this Capstone project is to develop a credible set of forecast models for a Retail Store of Walmart which has a full year sales target of 5% growth over prior year’s sales volume. The set of forecast models comprised of:

* A full year (“FY”) forecast model of sales volume
* Next 28 days forecast of sales volume for ongoing performance monitoring.

Desired outcome of this project as follows:

1. Assess the given target using a robust full year forecast model rather than based on gut-feel or quantitative approach
2. Monitor daily sales performance against next 28 days daily forecast to identify any actions required, highlight changes to the current operating environment, etc
3. The forecast models are applicable for other stores
4. The forecast models are scalable.

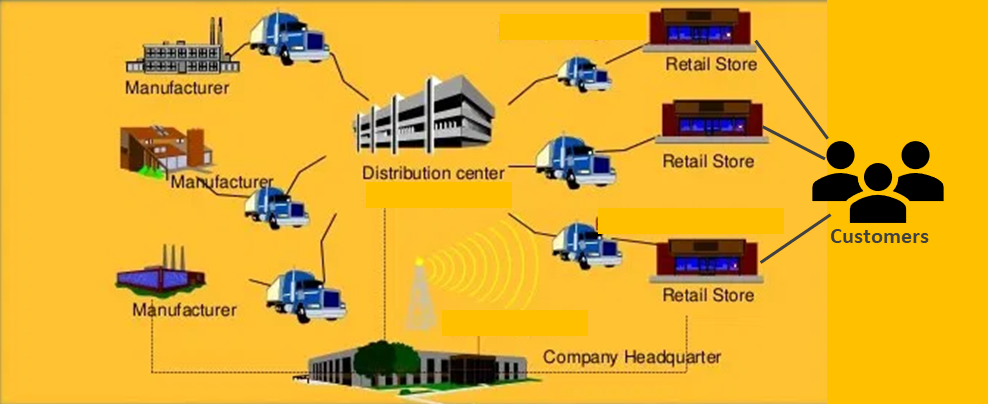
# **Industry Domain**

The business questions raised in Section 2 are relevant for any organisation, large or small. Also, they are applicable for all industries as well.

For this Capstone project, the industry domain selected is the retail supermarket based on the data availability.

## **Value chain of retail supermarket**

The value chain of the retail supermarket starts with the suppliers of the goods to be sold in the retail supermarket stores. The customers of the retail supermarket store are the end-point of the value chain. The goods move from the suppliers to the distribution center of the retail supermarket first and subsequently distributed to the retail supermarket stores to sell to the customers. The departments of the supermarket organisation such as the supply and logistic, procurement, transport drivers, inventory maintenance, etc are involved in the value chain.



**Figure Value Chain of Retail Supermarket Store**

The decisions made by the retail stores have a direct impact on the overall value chain. For example, the frequency of the request to replenish the food supply to the retail store would require the coordination of the warehouse and drivers to deliver the food supply to the store. The procurement team may need to arrange with the suppliers of the food supply to ensure there is sufficient stock in the inventory.

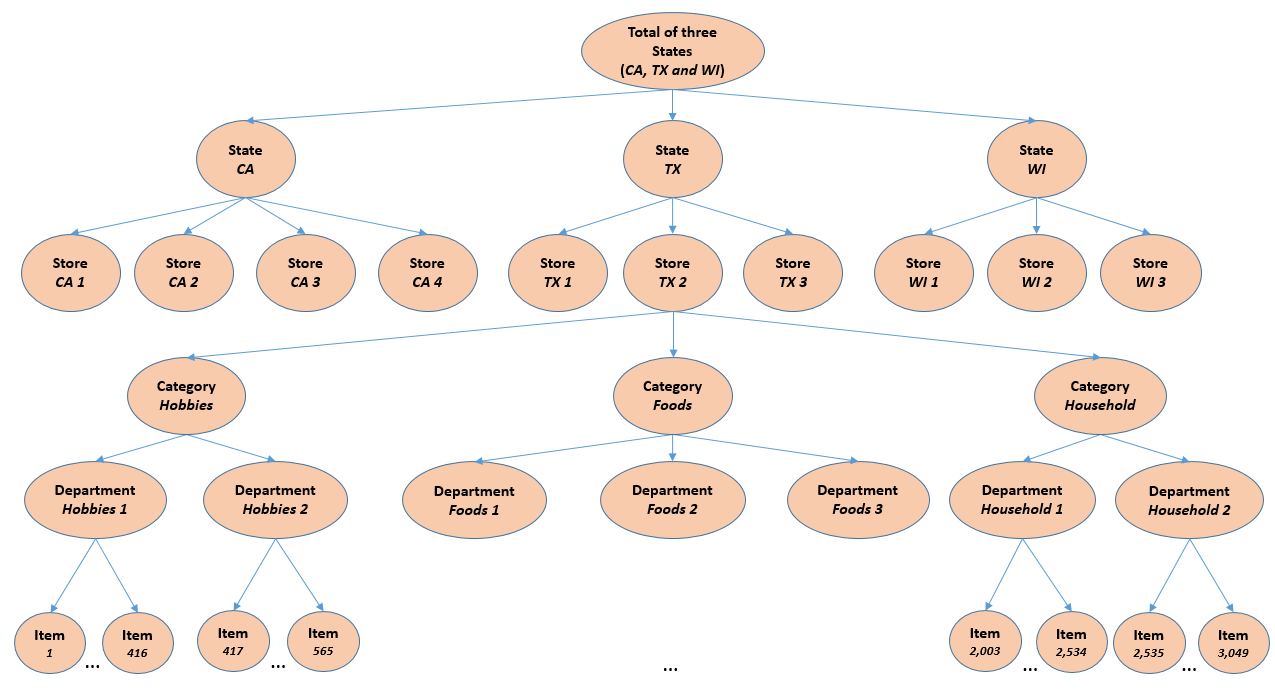
# **Stakeholders**

The main stakeholders are the retail store management. The departments involved in the value chain of the retail supermarket store are the stakeholders as well as they would be impacted by the action taken by the retail store management.

# **Data**

## **Data Structure**

The data used to develop the forecast models is the Walmart sales data of its retail supermarket stores. It is sourced from Kaggle M5 Forecasting - Accuracy Competition. It involves the unit sales of various products sold in the USA, organized in the form of grouped time series. The dataset involves the unit sales of 3,049 products, classified in 3 product categories (Hobbies, Foods, and Household) and 7 product departments, in which the above-mentioned categories are disaggregated. The products are sold across ten stores, located in three States (California (“CA”), Texas (“TX”), and Wisconsin (“WI”)). In this respect, the bottom-level of the hierarchy, i.e., product-store unit sales can be mapped across either product categories or geographical regions, as follows:



**Figure An overview of how the data is organized**

The historical data range from **2011-01-29** to **2016-06-19**. Thus, the products have a (maximum) selling history of 1,941 days / 5.4 years. A full year (“FY”) is assumed to run from current year, 23 May to next year, 22 May.

The dataset consists of three files as shown in the below schematic. The definition of the contents is available within Appendix I.

sales train.csv

Column:

* *item\_id*
* *dept\_id*
* *cat\_id*
* *store\_id*
* *state\_id*
* *d\_1*
* *d\_2*
* *d\_i,*
* *…*
* *d\_1941*

calendar.csv

Column:

* *date*
* *wm\_yr\_wk*
* *weekday*
* *wday*
* *month*
* *year*
* *d*
* *event\_name\_1*
* *event\_type\_1*
* *event\_name\_2*
* *event\_type\_2*
* *snap\_CA*
* *snap\_TX*
* *snap\_WI*

sell\_prices.csv

Column:

* *store\_id*
* *item\_id*
* *wm\_yr\_wk*
* *sell\_price*

**Figure Schematic diagram of 3 datasets.**

## **Data Wrangling**

Data cleaning and profiling are first performed on the 3 data files to make them “usable” for analytics and modelling.

### **File 1: “calendar.csv”**

* Replaced the nan values in columns “event\_name\_1”, “event\_type\_1”, “event\_name\_2” and “event\_type\_2” with “nil” values
* Created additional binary data columns for Christmas Day, New Year Day, Labour Day, SuperBowl event, Thanks Giving (including 2 days after Thanks Giving), as well as, Eve of holidays or events
* Converted to a timeseries dataset, indexed by column “date”

### **File 2: “sell\_prices.csv”**

* Merged with File 1 “calendar” to convert the weekly dataset to a daily timeseries dataset, indexed by column “date”
* Added a column for daily mean sell price by store\_id

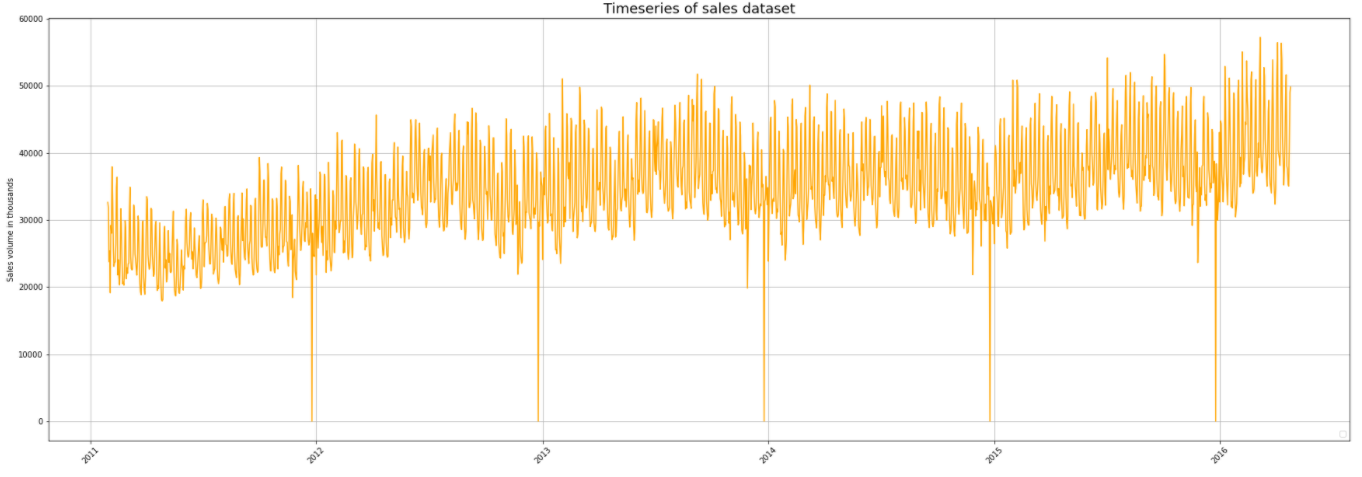
### **File 3: “sales\_train.csv”**

* Reshaped the dataset of 1,919 columns to 6 columns, ie the columns d\_1 … d\_1941 is consolidated under 1 column
* Merged with File 1 “calendar” to convert the reshaped dataset to a daily timeseries dataset, indexed by column “date”.

# **Data Analysis**

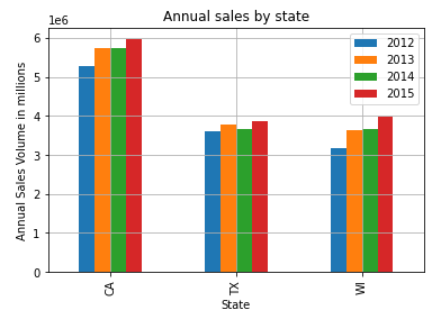
## **Overview of States, Stores and Products Sales Performances**

As shown in the chart below, Walmart sales volume is trending upwards with peaks and troughs, notably near year-end, generally across the years.



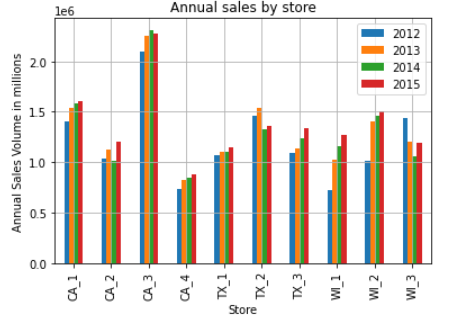
**Figure**  **Timeseries of sales transaction dataset**

State CA is the top state’s sales performer across the period 2012 – 2015 and growing annually. State TX sales performance is relatively stable with moderate growth. State WI’s sales jumped in 2013 and continued its growth trend in 2014 and 2015.

****

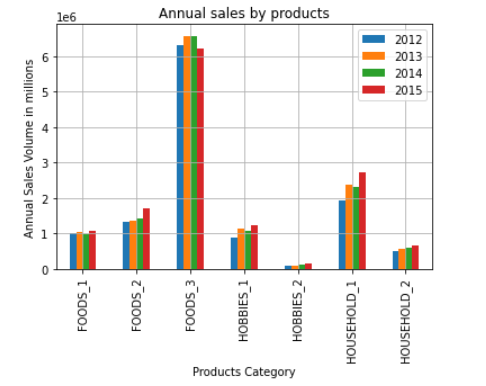
**Figure Annual sales by state**

Store CA3 is the top store’s sales performer across the period 2012 – 2015. Stores CA1, TX2, WI2 have comparable sales performance across the 4-year period.



**Figure Annual sales by store**

The best seller across the states and stores are the FOODS\_3 product category with significantly higher sales volumes than the rest of the products.

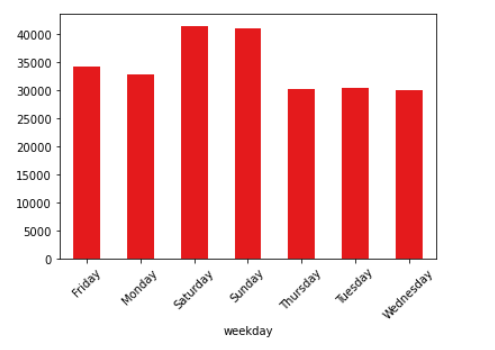


**Figure Annual sales by product category**

## **Features Analysis**

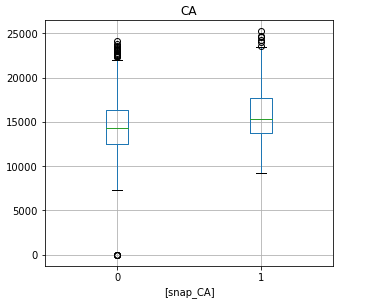
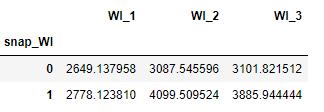
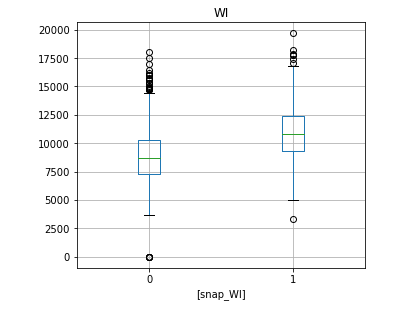
### **Sales volume by weekdays**

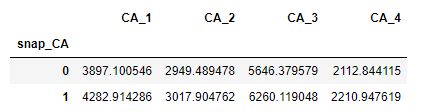
Saturday and Sunday have the highest sales in the week.

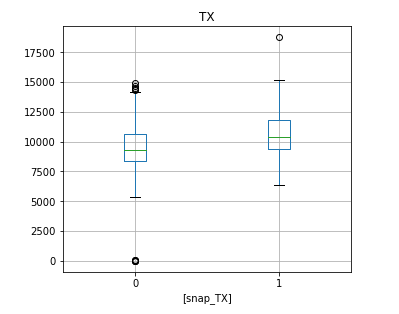


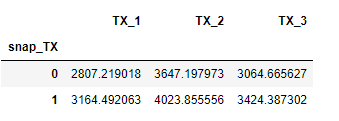
**Figure Sales by weekday**

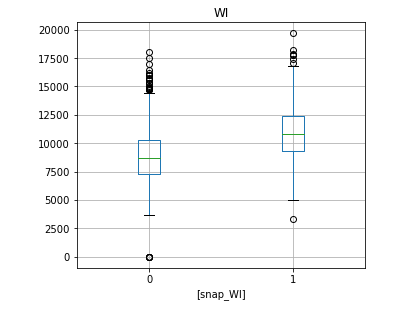
### **Sales volumes of SNAP event days vs Non-SNAP event days**

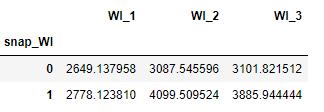
Higher sales during SNAP event across the 3 states and 10 stores than non-SNAP event days.







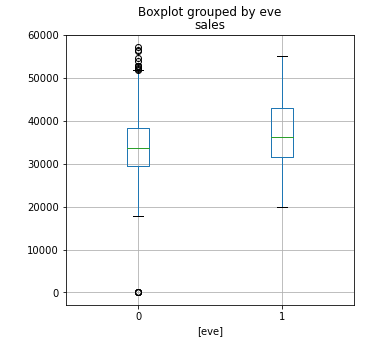




**Figure Sales of SNAP event vs non-SNAP event by state and stores**

### **Sales volumes of Eve of holiday/event vs Non-eve days**

Higher sales on eve of holiday/event vs non-eve day.

****

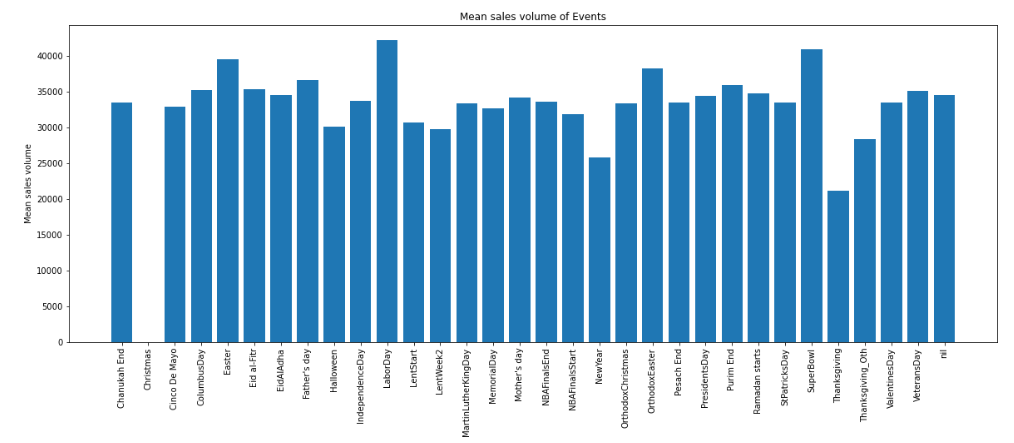
**Figure Sales on Eve of holiday/event vs non-eve day**

### **Sales volumes of holidays and events**

The sales volume is influenced by holiday and event.

1. Top sales volume on Labour Day, SuperBowl, and Easter
2. Near zero sales on Christmas day
3. Lowest sales volumes on Thanksgiving and New Year day

The ranking of the sales volumes of the above-mentioned holidays and events varied slightly across the different states and stores.



**Figure Sales of a holiday/event**

## **Features / Parameters Selection**

Based on the data analysis, the sales volume drivers as follow:

* SNAP event
* Eve of holiday / event
* Specific holidays / events namely, Christmas, Thanksgiving, New Year, Labour Day and SuperBowl
* Sell prices of the products

CA1 store is chosen for modelling as it has an average sales performance and is comparable with other stores.

# **Modelling & Validations**

## **SARIMA Modelling and Measurement of predictions**

The modelling is using the Econometric method with its basis on the statistical properties of the time series. As observed in the trend of the sales transaction dataset, there is peaks and troughs each year. This likely denotes seasonality in the sales transaction trend. Thus, this Capstone project applied the Seasonal Autoregression Moving Average model (“SARIMA”) from the statsmodels module to build the forecast models.

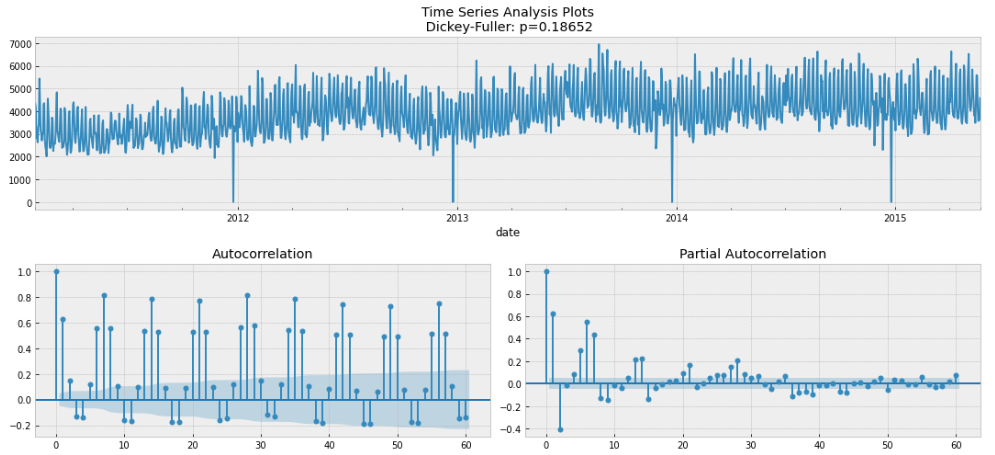
The accuracy of the forecast model is evaluated using the Mean Absolute Percentage Error (“MAPE”). The measure is calculated as follows:



## **Full Year Forecast Model – Store CA1**

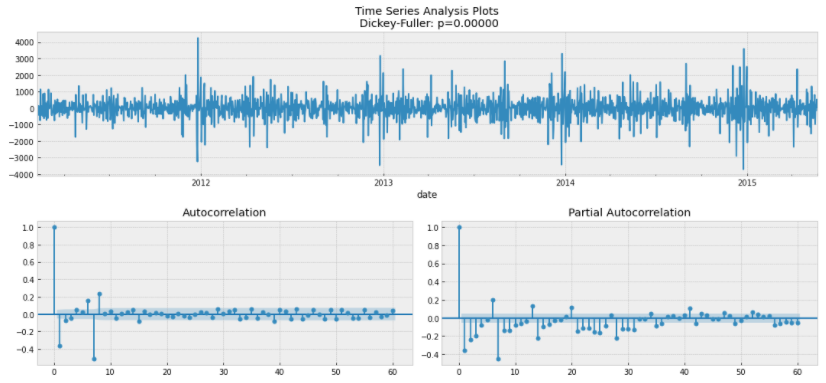
### **Built the model**

The store CA1 timeseries is not stationary based on the Augmented Dickey-Fuller (“ADF”) test as shown in Figure 12. The p-value of 0.187 which is greater than the threshold of 0.05, suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure. In order to make the store CA1 timeseries stationary, a differencing order of 1 is made to remove the trend and ‘flatten’ the time series. From the autocorrelation chart (“ACF”), every 7th lag peaks. From the Partial Autocorrelation chart (“PACF”), the 7th lag is the last significant peak. The differencing order of 1 ie difference of the timeseries with it shifted by 7 lags, is required to make the store CA1 timeseries stationary as well as to remove the additive seasonal effects. However, the autocorrelation function after differencing still has too many significant lags, a first difference of subtracting the series from itself with lag 1 is made.



**Figure Store CA1 Time Series, ACF and PACF**

The store CA1 time series after the differencing, shown below, looks like something indescribable, oscillating around zero. The ADF test indicates that it is stationary, and the number of significant peaks in ACF has dropped.



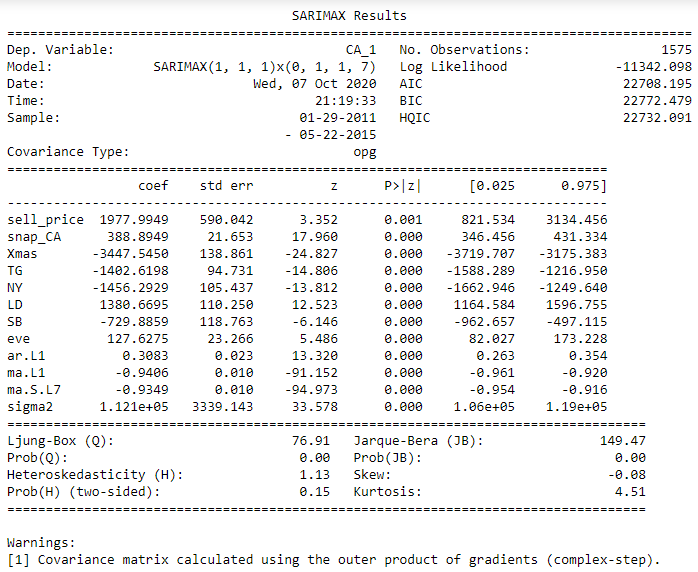
**Figure Store CA1 Time Series, ACF and PACF after Differencing**

The initial parameters for the FY Forecast model SARIMA (p, d, q) x (P, D, Q, s) identified as follows:

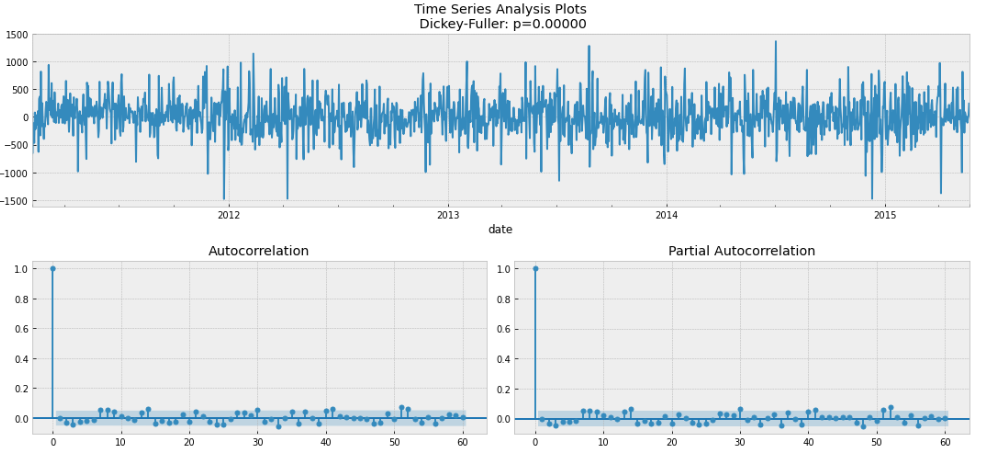
* p is likely to fall in the range of 1-7 since the 7th lag in PACF chart is the last significant lag after which **most** other lags become insignificant
* q is likely to fall in the range of 1-7 as similarly observed in the ACF chart
* d and D are 1 as first differences is made for both trend and seasonality
* s is 7, the season period length of the series
* P is likely to fall in the range of 0-2 since 7th and 14th lags are significant in the PACF chart
* Q is likely to fall in the range of 0-2 since 7th lag is significant in the ACF chart

The set of the initial parameters is fed into the “Optimizer” (a defined function) to find the best combination of the parameters that generates the lowest Akaike Information Criterion (“AIC”) score. AIC is to estimate the likelihood of a model to predict/estimate the future values. Apart from AIC score, the p-value of the exogenous variables of SARIMA model results is another important factor to determine the best fit model. The exogenous variables are the identified drivers of the sales volumes in Section 7.3.

The FY forecast model for store CA1 is SARIMA (1, 1, 1) x (0, 1, 1, 7). The SARIMAX results summary of the FY forecast model and the residuals of the model are shown below. The residuals of the model is stationary, and there are no apparent autocorrelations. The model is ready to make predictions.



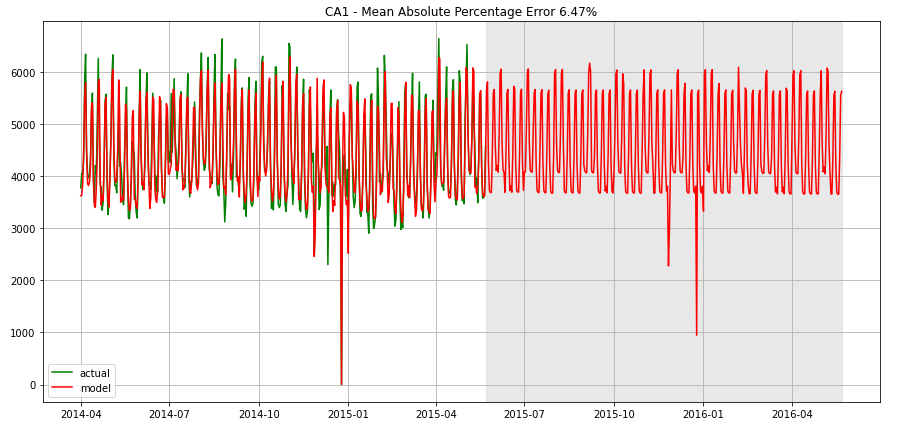
**Figure Store CA1 FY Forecast model results summary**



**Figure Residuals of Store CA1 FY Forecast model**

### **Accuracy of the FY Forecast model**

The FY forecast of store CA1 made by the model has a MAPE of 6.47%.



**Figure Predictions of Store CA1 FY Forecast model**

### **Evaluation of Store CA1 Full Year Model**

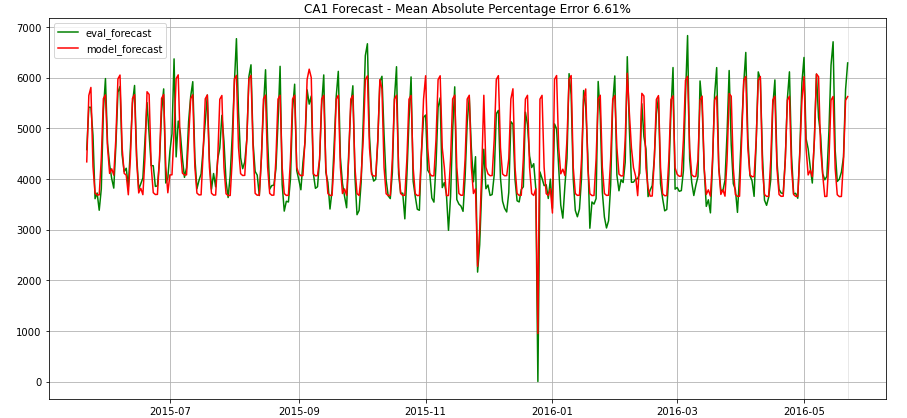
|  |  |
| --- | --- |
| Actual sales FY 2014/2015 (2014-05-23 to 2015-05-22) | 1,600,222 |
| Forecast sales FY 2014/15 | 1,595,730 |
| **Actual sales vs Forecast sales** | **0.28%** |
| Target Growth for next FY 2015/16 (2015-05-23 to 2016-05-22) | 5% |
| ie Target sales | 1,680,233 |
| Forecast sales for next FY 2015/16 | 1,659,149 |
| **Gap** | **1.25%** |
| Actual sales for next FY 2015/16 | 1,633,096 |
| **Actual sales for FY 2015/16 vs Forecast sales** | **-26,053** |
| **Behind Target** | **-2.81%** |

The FY Forecast model has predicted FY 2014/15 sales of 1.596m which is 0.3% marginally lower than the actual sales in FY 2014/15 of 1.600m.

With the target growth of 5% over prior year, the sales target for FY 2015/16 is 1.680m. It appears to be 1.25% higher than the forecast sales for FY 2015/16 of 1.659m. The gap to the sales target looks achievable for store CA1 if no significant changes in the business environment.

With the MAPE of forecast model being 6.47%, the actual sales would be different from the prediction by the FY Forecast model. With hindsight, the actual sales for FY 2015/16 was lower than forecast sales by 26K / 1.60%, ie lower than the sales target by 2.81%. The MAPE of FY forecast model for FY 2015/16 versus actual sales turns out to be 6.61%.

The store CA1 would need to monitor the sales performance for next year closely.

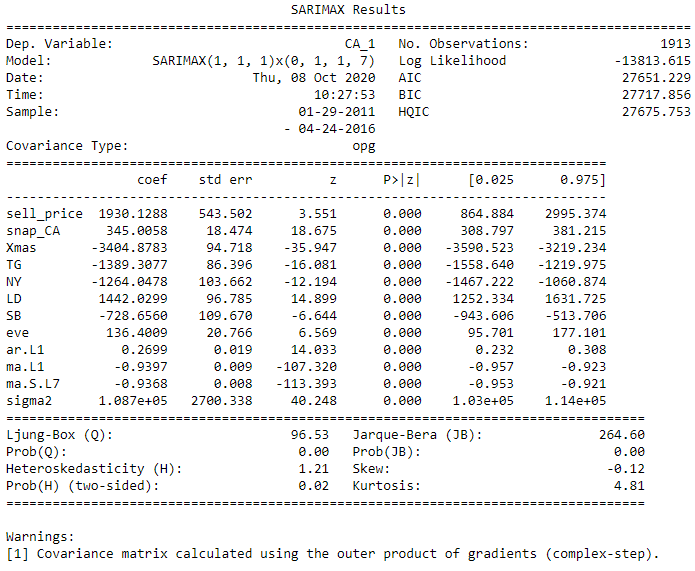


**Figure** **FY2015/16 Predictions of Store CA1 FY Forecast model**

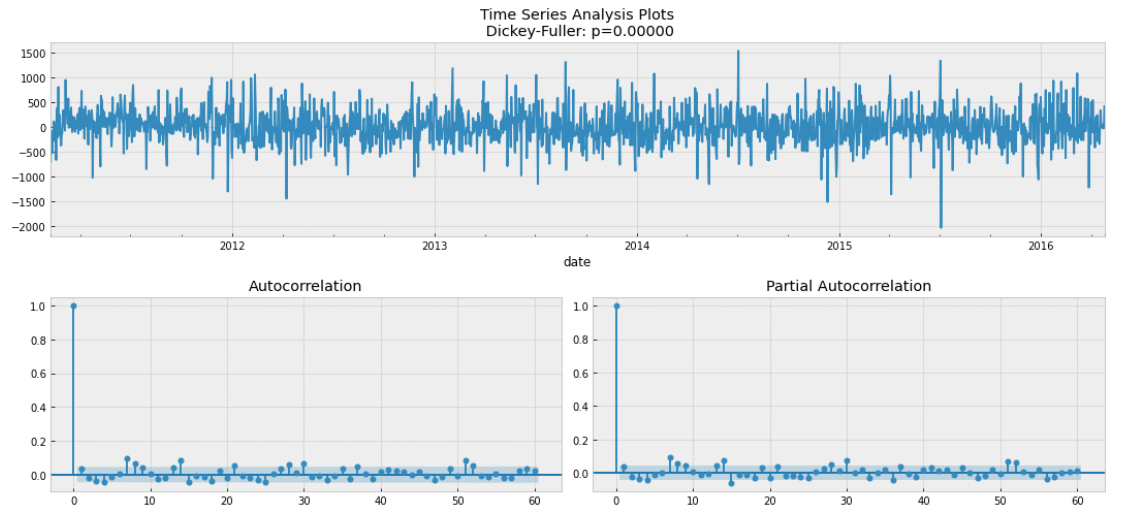
## **Store CA1 Next 28 days Forecast Model**

Store CA1 monitored the daily sales performance by comparing the actual daily sales with the daily sales forecast generated by the Next 28 days Forecast model. The Next 28 days forecast model for store CA1 is SARIMA (1, 1, 1) x (0, 1, 1, 7) with similar set of exogenous variables as per the FY model. The data applied included the latest data ie up to day 1913 or 2016-04-24.

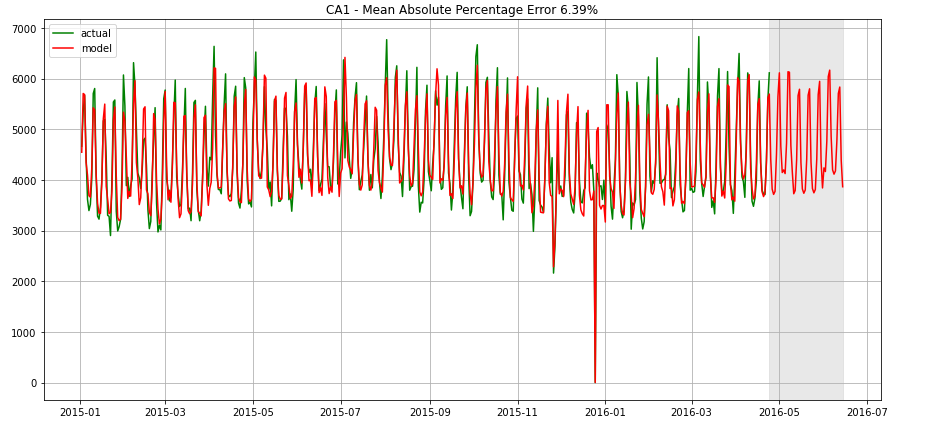
The SARIMAX results summary of the forecast model, the residuals of the model and the predictions made by the model are shown below. The MAPE of the Next 28 days forecast model for store CA1 is 6.39%.



**Figure Store CA1 Next 28 days Forecast model results summary**



**Figure Residuals of store CA1 Next 28 days Forecast model**



**Figure Predictions of Next 28 days Forecast model for store CA1**

The predicted sales forecast for next 28 days ie period 2016-04-25 to 2016-05-22 generated by the model is shown below.



**Figure 2016-04-25 to 2016-05-22 Predictions of next 28 days Forecast model**

The actual sales was trending closely to the forecast sales until 2016-05-11 where the difference started to widen with actual sales higher than forecast sales. This could trigger the store CA1 management to check on the stock availability to cater for a possible uptick in sales and additional staff resources to cope with the higher sales traffic. With hindsight, actual sales was higher than forecast sales. The MAPE of forecast sales and actual sales for next 28 days is 6.38%.

If the actual sales took a downturn instead, ie trending below the forecast, the store CA1 management should closely monitor for the causes for the downturn and assess the possibility of external competition or economic related changes which might require a longer-term solution. In the immediate term, the store CA1 management should consider ways to reduce potential wastage of perishable goods by having flash sales for example.

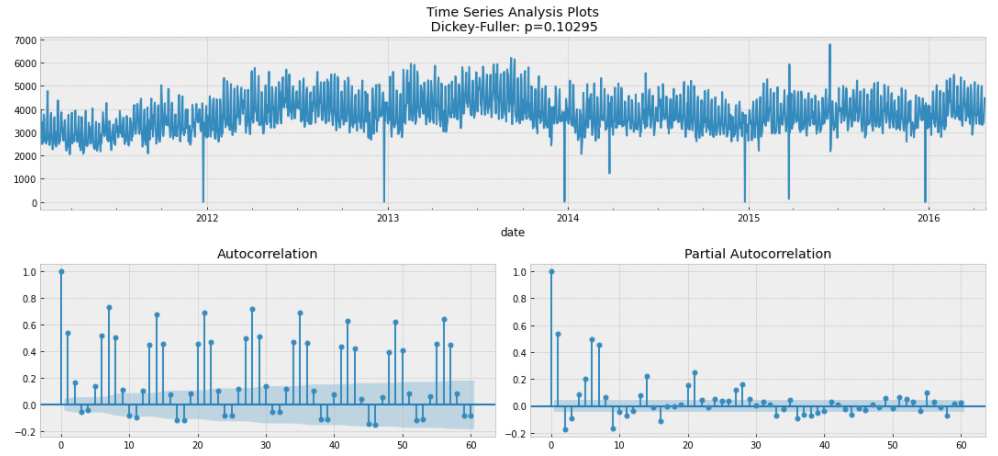
## **Desired Outcome #3**

The 2 forecast models for store CA1 have demonstrated to meet the desired outcome #1 and #2. In order to assess whether the 2 forecast models are able to meet the desired outcome #3 ie being applicable for other stores, the Next 28 days Forecast model for store CA1 is applied to store TX2 and store WI2.

The Next 28 days Forecast models for both stores TX2 and WI2 are built using the similar set of exogenous variables as store CA1 and the latest sales data of TX2 and WI2.

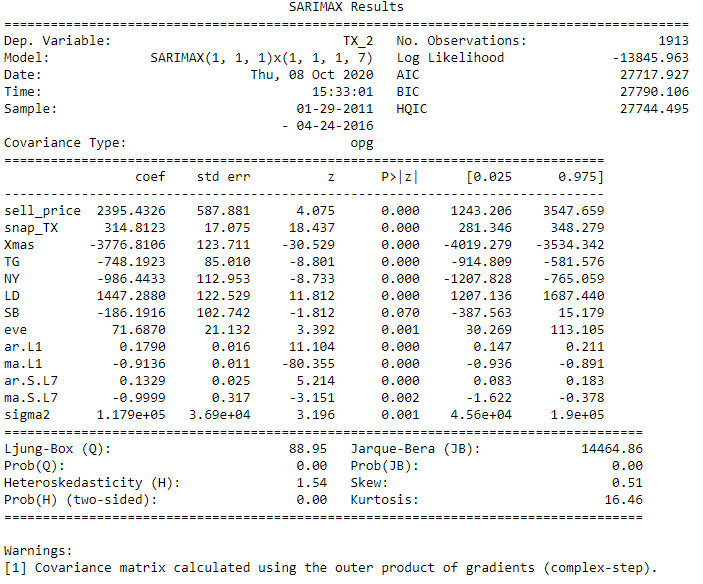
### **Store TX2**

Store TX2 timeseries is not stationary with p-value of 0.103 derived by the ADF test. Similar differencing treatment as per store CA1 model is applied to remove the trend and seasonality.



**Figure Store TX2 Time Series, ACF and PACF**

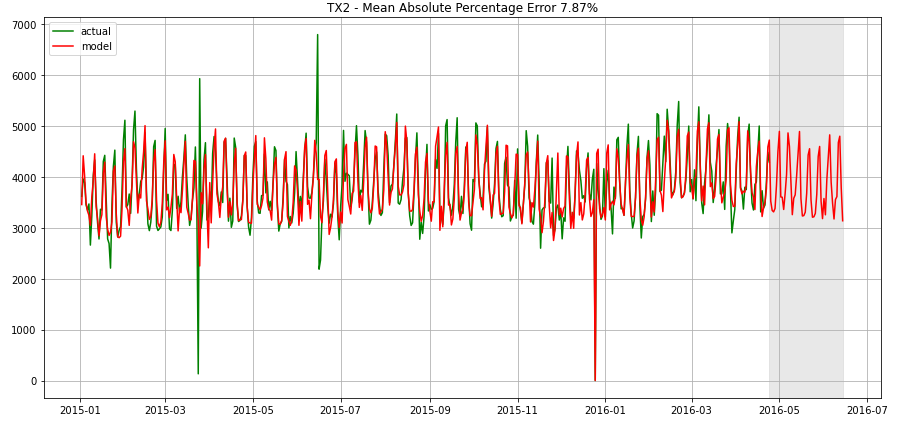
The SARIMAX results summary of the forecast model, the residuals of the model and the predictions made by the model are shown below.



**Figure Store TX2 Next 28 days Forecast model results summary**

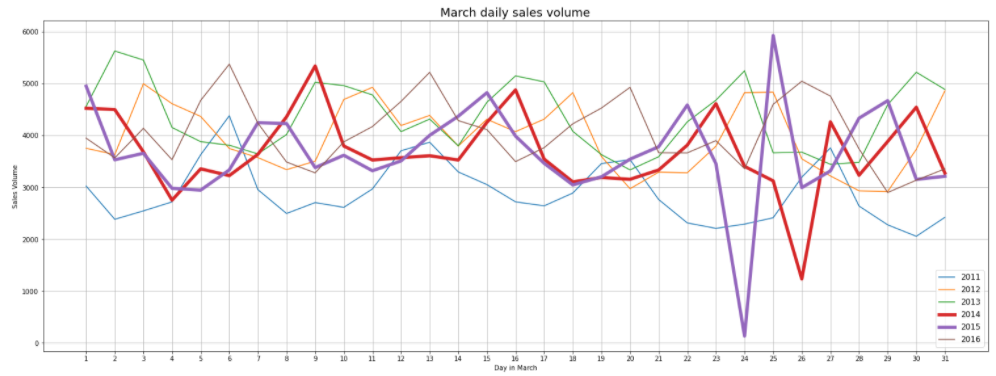


**Figure Residuals of store TX2 Next 28 days Forecast model**



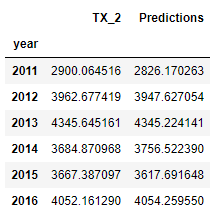
**Figure Predictions of Next 28 days Forecast model for store TX2**

The MAPE of the Next 28 days Forecast model for store TX2 is 7.87%. This is likely due to the several outliers in TX2 sales data, notably 3 outliers observed on 26 March 2014, 24 and 25 March 2015 as follow:



**Figure Sales outliers for store TX2**

Excluding the mentioned 3 outliers, the MAPE drops to 6.54%. The average sales volume in March across the years as follows:



**Figure Mean sales volume in March for Store TX2 where TX\_2 is Actual**

In conclusion, store CA1’s forecast model is applicable for store TX2.

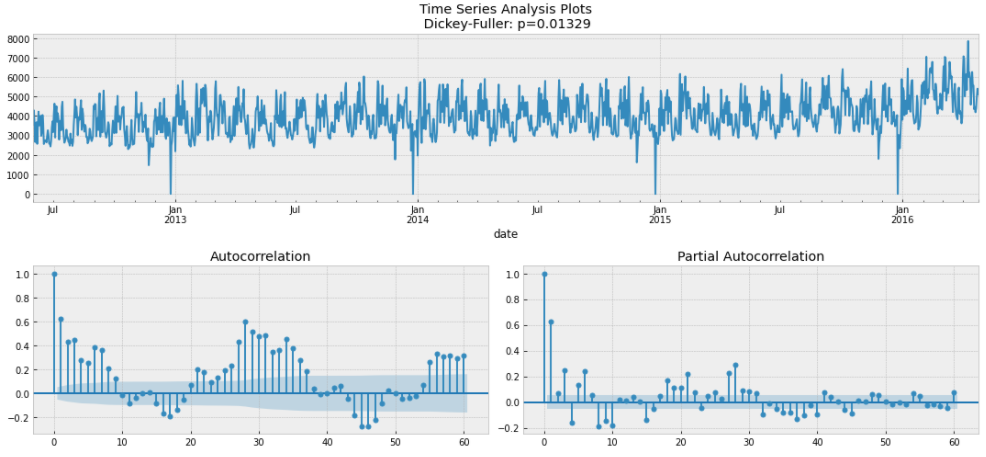
### **Store WI2**

Store WI2 timeseries shows a spike in the sales volume in 2012 2nd half and sales volume has maintained without dropping back to pre-2012 2nd half level in the subsequent years. Thus, the pre-2012 2nd half sales data is discarded as it would distort the model predictability.



**Figure Store WI2 Time Series, ACF and PACF**

The trimmed timeseries of store WI2, excluding the sales data of 2011-2012 1st half has a p-value of 0.013 derived by the ADF test. Although the ADP test shows that the trimmed timeseries is stationary, differencing treatment as per store CA1 model is applied to remove the trend and seasonality.

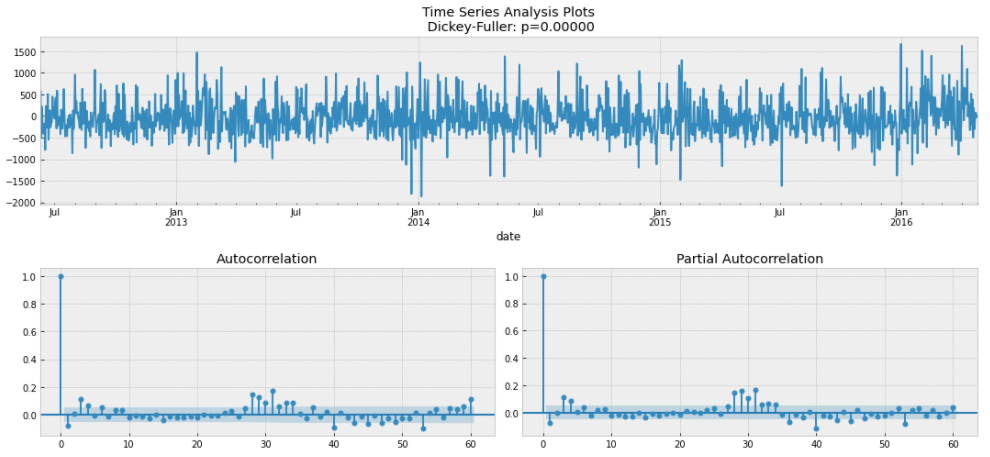


**Figure Store WI2 Time Series, ACF and PACF, excluding 2011-2012 1st half data**

The SARIMAX results summary of the forecast model, the residuals of the model and the predictions made by the model are shown below.

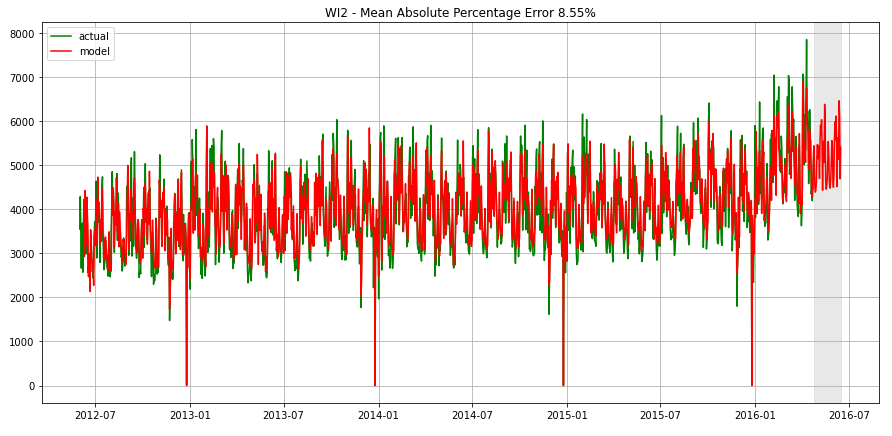


**Figure Store WI2 Next 28 days Forecast model results summary**



**Figure Residuals of store WI2 Next 28 days Forecast model**

The MAPE of the store WI2 model is 8.55%.

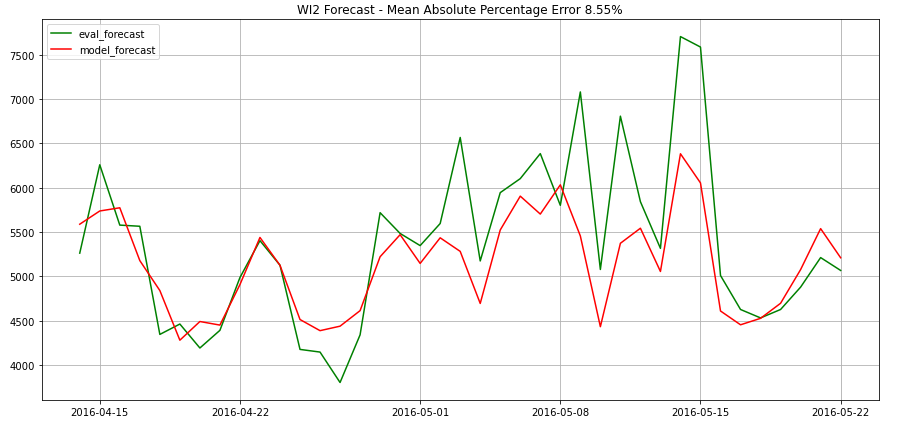


**Figure Predictions of Next 28 days Forecast model for store WI2**

The store WI2 Next 28 days forecast model built using the full set of data is available within Appendix II for comparison purpose.

Based on the next 28 days forecast for period 2016-04-25 to 2016-05-22, store WI2 seems to be experiencing a possible uptick in sales volumes with the actual sales higher than forecast sales since 2016-05-01. Similar to store CA1, store WI2 management should check on the stock and staff resources availability to meet the growing demand. Further finetuning of store WI2 forecast model is required if the actual sales is consistently higher than the predictions. With hindsight, the total actual sales for period 2016-04-25 to 2016-05-22 is 154K, 9K/6% higher than the total forecast sales of 145K.

With both stores CA1 and WI2 showing a possible uptick in sales volume in 2016-05, the inventory and procurement departments would need to review the stock inventory and earlier replenishment of goods than expected.

****

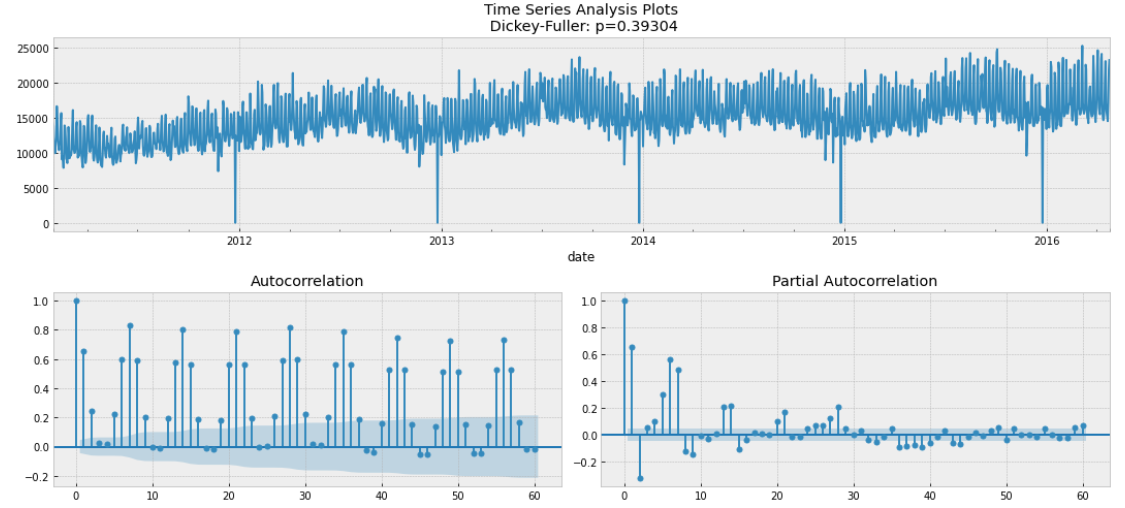
**Figure 2016-04-25 to 2016-05-22 Predictions of next 28 days Forecast model for store WI2**

## **Desired Outcome #4**

The desired outcome #4 is the scalability of forecast models. The State CA forecast model is built as per store CA1 Next 28 days Forecast model to demonstrate the scalability of such model. It used similar set of exogenous variables and the latest sales data of State CA.

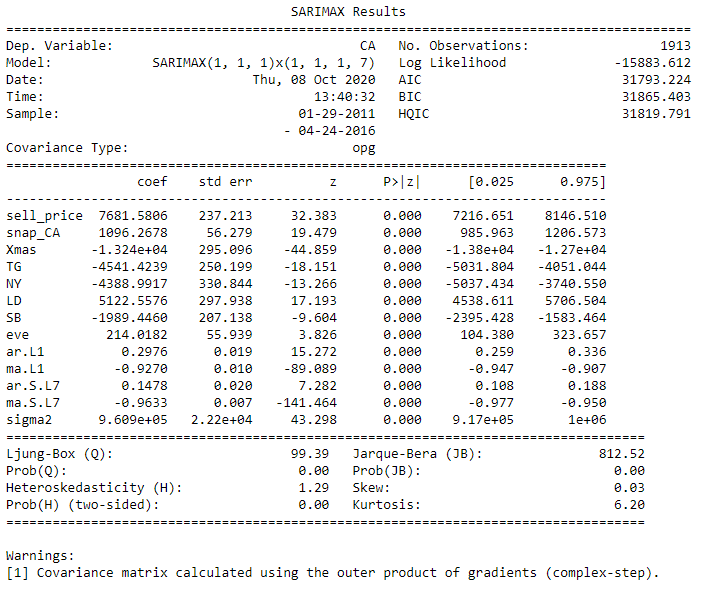
### **State CA Next 28 days Forecast Model**

The State CA Daily Timeseries is not stationary, having p-value of 0.393 per ADF test. The ACF of State CA timeseries is similar to store CA1 with peaks on every 7th lag. Whilst the 7th lag of the PACF chart is the last significant lag after which the rest of the lags oscillating around zero.

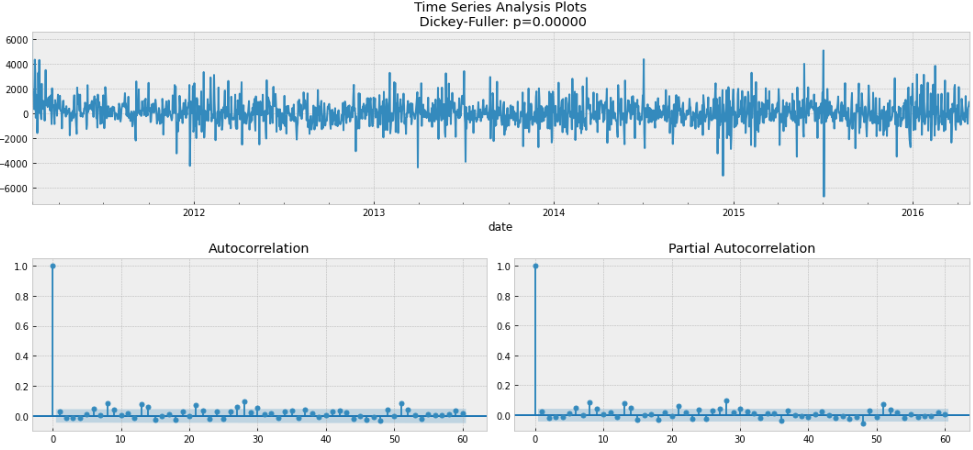


**Figure State CA Time Series, ACF and PACF**

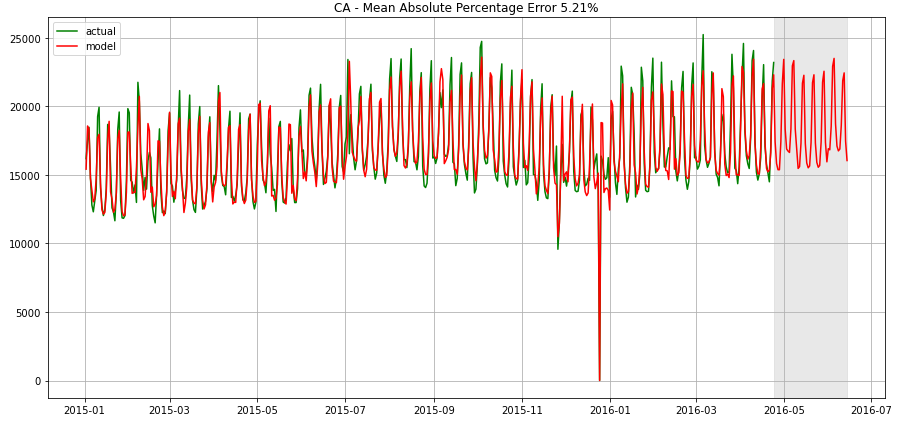
The SARIMAX results summary of the State CA forecast model, the residuals of the model and the predictions made by the model are shown below. The MAPE of the Next 28 days forecast model for State CA is 5.21%.



**Figure State CA Next 28 days Forecast model results summary**



**Figure Residuals of State CA Next 28 days Forecast model**

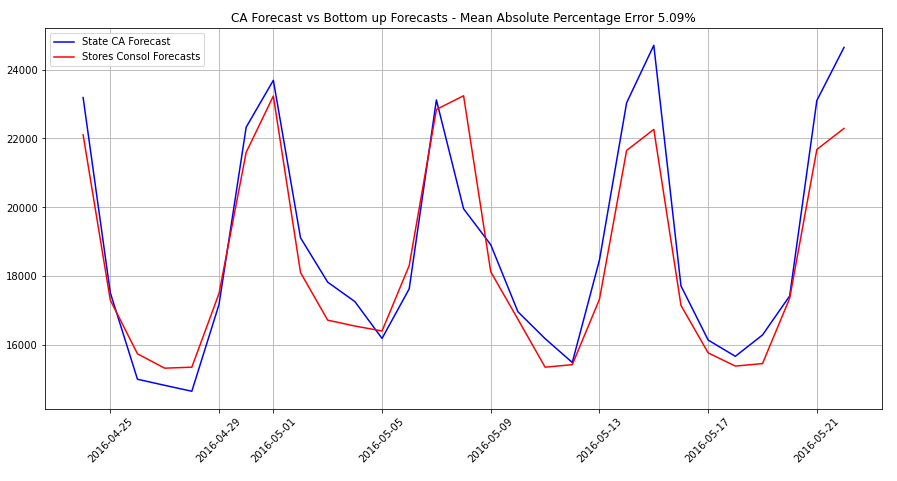


**Figure Predictions of State CA Next 28 days Forecast model**

### **Reconciliation of State CA and Consolidated Stores Models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **State/ Store** | **MAPE of Model** | **MAPE of 28 days forecast sales** | **Actual Sales of next 28 days** | **Forecast Sales of next 28 days** |
| Store CA1 | 6.39% | 6.38% | 134,032 | 129,859 |
| Store CA2 | 7.18% | 7.16% | 132,920 | 126,717 |
| Store CA3 | 5.96% | 5.95% | 175,360 | 178,987 |
| Store CA4 | 6.04% | 6.06% | 78,858 | 74,784 |
| **Consolidation** | **-** | **-** | **521,170** | **510,347** |
| **State CA** | **5.21%** | **5.19%** | **521,170** | **512,843** |

The State CA Forecast model has predicted the sales for period 2016-04-25 to 2016-05-22 of 513K, 2.5K/0.5% higher than the summation of the 4 individual stores prediction of 510K. The predictability of State CA Next 28 days Forecast model is comparable with the summation of the 4 individual stores prediction. In fact, the State CA model gives better prediction than the summation of the 4 individual stores. Its sales prediction compared to actual sales is 8K lower versus the summation of the 4 individual stores difference of 11K.



**Figure Predictions of State CA vs Summation of individual Stores**

# **Conclusions**

The objective of this Capstone project is achieved with the Full Year and Next 28 days Forecast models built with prediction vs actual error rates of less than 10%.

The FY forecast model allows the store CA1 to assess its financial target with a baseline built with explainable features and not assumptions.

The Next 28 days forecast model allows the store CA1 to monitor the sales performance closely and trigger timely actions.

Also, the 2 models are applicable to other stores, TX2 and WI2, as well as, being scalable to the State level.

These 2 models, in general, are useful for organisations which have financial targets to meet and require sales performance monitoring tool.

# **Future works**

Although the 2 forecast models have achieved the objective of this Capstone project, the models can be further enhanced as follows:

* Include macro-economics variables such as GDP, consumer price index in the modelling to improve the robustness of the predictability
* Use machine learning to build the models which may give a better prediction result
* Expand the forecast scope to include products which would be very useful for inventory control purpose.

**Appendix I - Details of 3 datasets**

The dataset consists of the following three files:

**File 1: “*calendar.csv”***

Contains information about the dates the products are sold. It has 14 columns by 1,969 rows. Columns are:

* *date*: The date in a “y-m-d” format.
* *wm\_yr\_wk*: The id of the week the date belongs to.
* *weekday*: The type of the day (Saturday, Sunday, …, Friday).
* *wday*: The id of the weekday, starting from Saturday.
* *month*: The month of the date.
* *year*: The year of the date.
* *event\_name\_1*: If the date includes an event, the name of this event.
* *event\_type\_1*: If the date includes an event, the type of this event.
* *event\_name\_2*: If the date includes a second event, the name of this event.
* *event\_type\_2*: If the date includes a second event, the type of this event.
* *snap\_CA*, *snap\_TX*, and *snap\_WI*: A binary variable (0 or 1) indicating whether the stores of CA, TX or WI allow SNAP[[2]](#footnote-2) purchases on the examined date. 1 indicates that SNAP purchases are allowed.

**File 2: *“sell\_prices.csv”***

Contains information about the price of the products sold per store and date. It has 4 columns by 6,841,121 rows. Columns are:

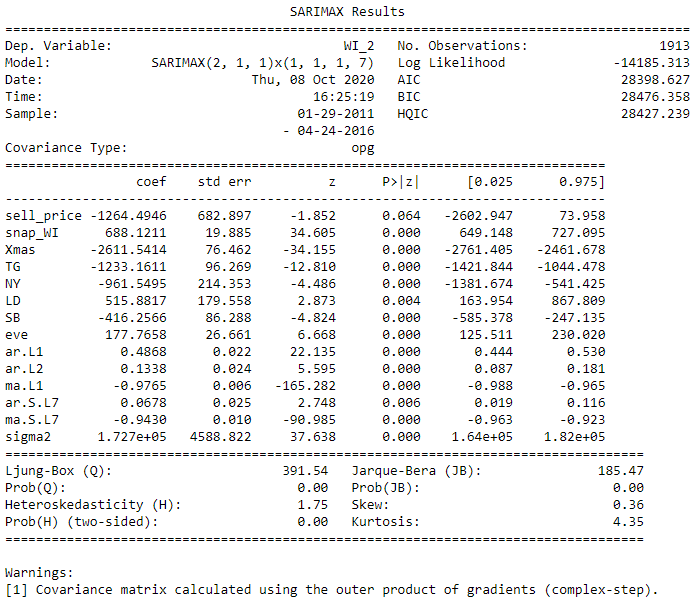
* *store\_id*: The id of the store where the product is sold.
* *item\_id*: The id of the product.
* *wm\_yr\_wk*: The id of the week.
* *sell\_price*: The price of the product for the given week/store. The price is provided per week (average across seven days). If not available, this means that the product was not sold during the examined week. Note that although prices are constant at weekly basis, they may change through time (both training and test set).

**File 3: “*sales\_train.csv*”**

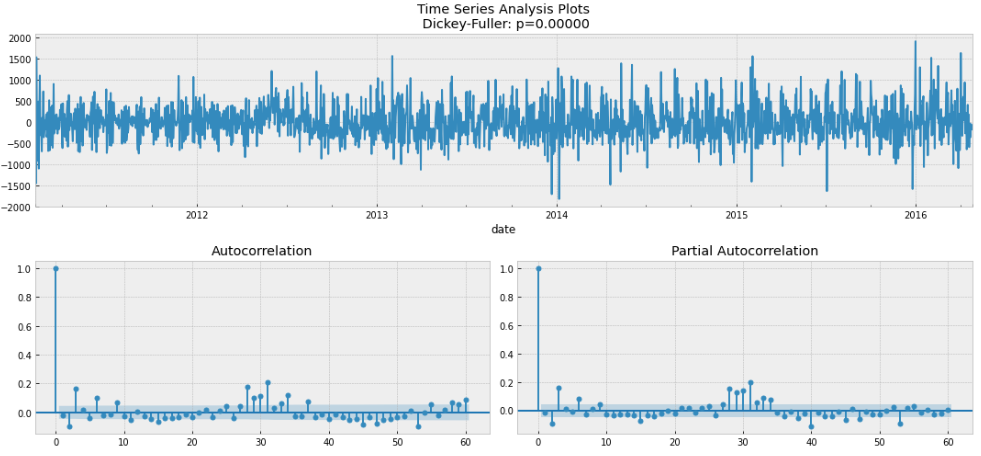
Contains the historical daily unit sales data per product and store. It has 1,919 columns by 30,490 rows. Columns are:

* *item\_id*: The id of the product.
* *dept\_id*: The id of the department the product belongs to.
* *cat\_id*: The id of the category the product belongs to.
* *store\_id*: The id of the store where the product is sold.
* *state\_id*: The State where the store is located.
* *d\_1, d\_2, …, d\_i, … d\_1941*: The number of units sold at day *i*, starting from 2011-01-29.

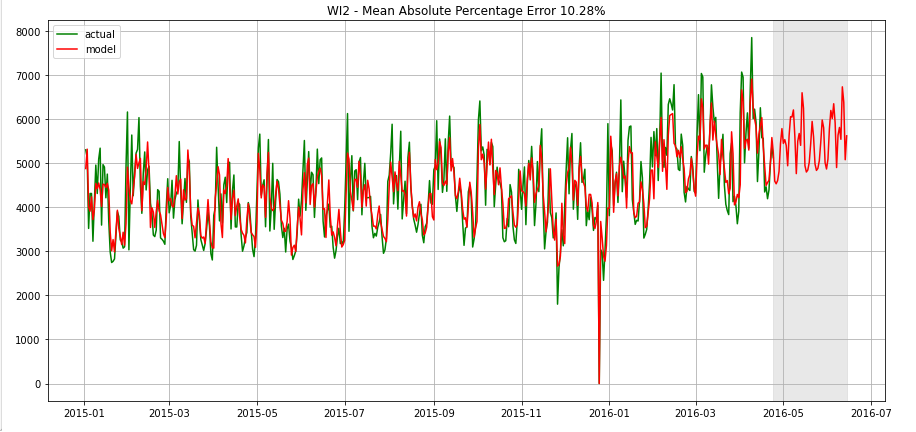
**Appendix II – WI2 forecast model using full set of data**



**Figure Store WI2 Next 28 days Forecast model results summary**



**Figure Residuals of Store WI2 Next 28 days Forecast model**



**Figure Predictions of Store WI2 Next 28 days Forecast model**

1. This is a simple illustration of a planning process only. It does not refer to the approach undertaken by the mentioned organization. [↑](#footnote-ref-1)
2. The United States federal government provides a nutrition assistance benefit called the Supplement Nutrition Assistance Program (SNAP). SNAP provides low income families and individuals with an Electronic Benefits Transfer debit card to purchase food products. In many states, the monetary benefits are dispersed to people across 10 days of the month and on each of these days 1/10 of the people will receive the benefit on their card. More information about the SNAP program can be found here: <https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program> [↑](#footnote-ref-2)