# Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number? The optimal number of store formats is 3. This was arrived using the K-Centroid Diagnostics tool in Alteryx. AR and CH value for each of the cluster is compared, in which Median of 3 cluster is high. AR value 0.6936 and CH Value 30.7.

**K-Means Cluster Assessment Report**

*Summary Statistics*

Adjusted Rand Indices:

|  | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Minimum | -0.0152 | 0.3171 | 0.3072 | 0.2412 | 0.2586 | 0.2903 | 0.2568 |
| 1st Quartile | 0.352 | 0.4819 | 0.4431 | 0.3943 | 0.3896 | 0.3877 | 0.377 |
| Median | 0.4926 | 0.6936 | 0.4964 | 0.4487 | 0.4348 | 0.4417 | 0.4526 |
| Mean | 0.484 | 0.6575 | 0.5125 | 0.4623 | 0.4532 | 0.4498 | 0.4411 |
| 3rd Quartile | 0.655 | 0.816 | 0.5913 | 0.4982 | 0.489 | 0.4997 | 0.491 |
| Maximum | 1 | 1 | 0.7458 | 0.7366 | 0.7762 | 0.6637 | 0.6118 |

Calinski-Harabasz Indices:

|  | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Minimum | 16.1 | 20.09 | 17.41 | 18.98 | 17.24 | 16.61 | 16.11 |
| 1st Quartile | 28.61 | 28.76 | 25.16 | 22.91 | 21.05 | 19.61 | 18.46 |
| Median | 29.47 | 30.7 | 26.25 | 24.05 | 22.02 | 20.56 | 19.5 |
| Mean | 28.41 | 29.47 | 25.99 | 23.88 | 21.96 | 20.48 | 19.62 |
| 3rd Quartile | 30.39 | 31.58 | 27.62 | 25.06 | 23.14 | 21.35 | 20.77 |
| Maximum | 31.95 | 33.41 | 30.09 | 26.53 | 24.87 | 23.6 | 22.59 |

*Plots*
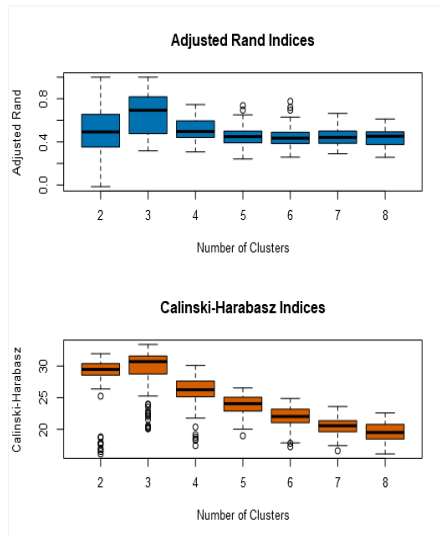


2. How many stores fall into each store format?

Using K-Means, cluster analysis is performed and the distribution is as shown below.

**Cluster Information:**

| Cluster | Size |
|---|---|
| 1 | 23 |
| 2 | 29 |
| 3 | 33 |

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

## Average Sales per Cluster
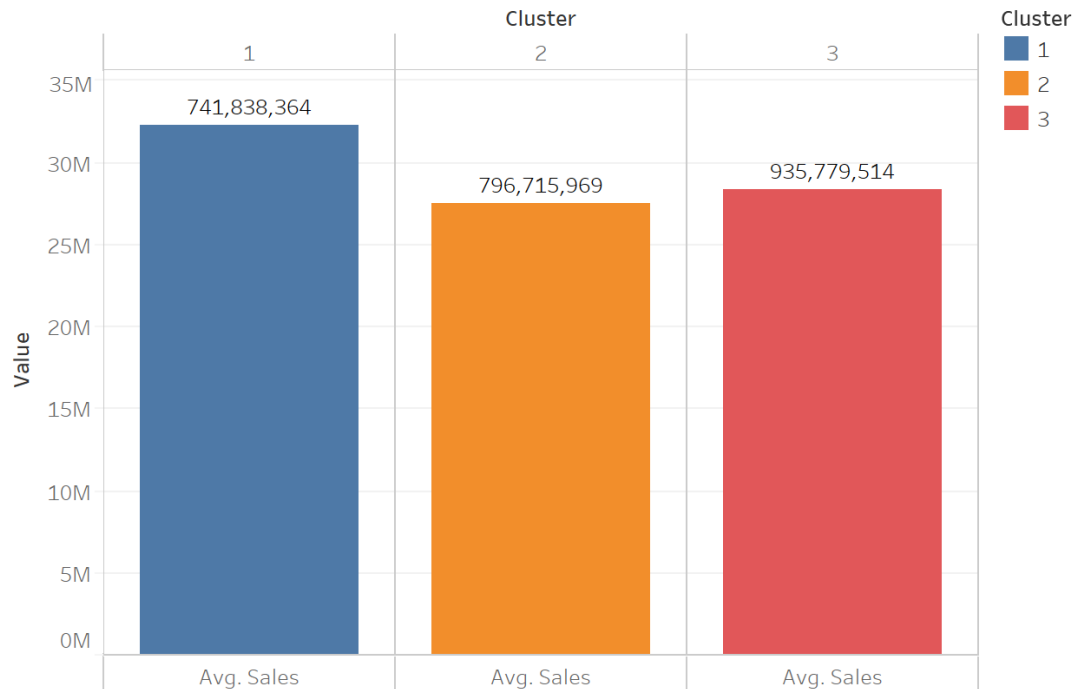


With the above results, we can notice that Stores in Cluster 1 sell more on an average when compared to the stores in the other two clusters.

4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

Store Cluster by Total Sales
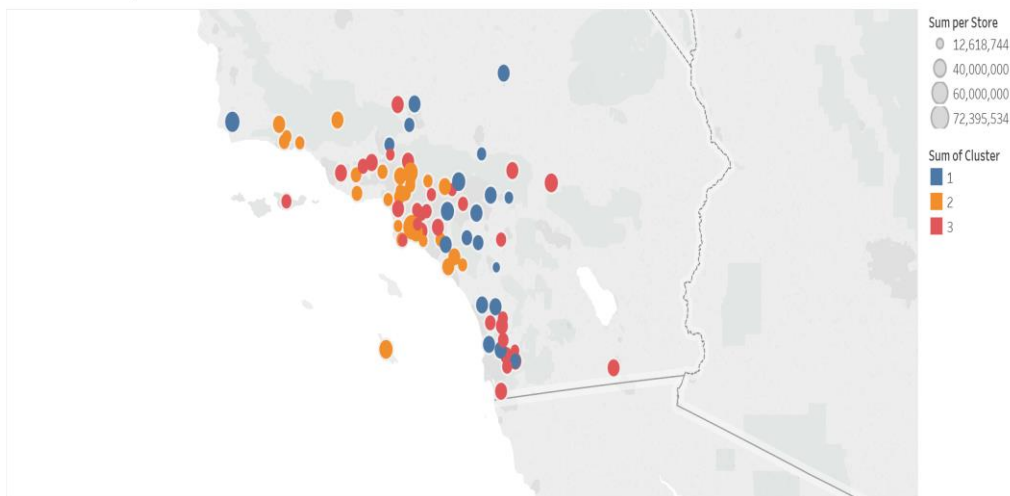


https://public.tableau.com/views/ProjectP8/StoreClusterbyTotalSales?:embed=y&:display_count=yes

# Task 2: Formats for New Stores

1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

   Constructed and compared the three models Decision Tree, Boosted Model and Forest Model and the Model comparison report is shown below.

**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| Decision_Tree | 0.7059 | 0.7327 | 0.6000 | 0.6667 | 0.8333 |
| Boosted_Model | 0.8235 | 0.8543 | 0.8000 | 0.6667 | 1.0000 |
| Forest_Model | 0.8235 | 0.8251 | 0.7500 | 0.8000 | 0.8750 |

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

**Confusion matrix of Boosted_Model**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 4 | 0 | 1 |
| Predicted_2 | 0 | 4 | 2 |
| Predicted_3 | 0 | 0 | 6 |

**Confusion matrix of Decision_Tree**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 3 | 0 | 2 |
| Predicted_2 | 0 | 4 | 2 |
| Predicted_3 | 1 | 0 | 5 |

**Confusion matrix of Forest_Model**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 3 | 0 | 1 |
| Predicted_2 | 0 | 4 | 1 |
| Predicted_3 | 1 | 0 | 7 |

From the above report, we can observe that the overall accuracy of two of the models (Boosted Model and Forest Model) are the same. This is not uncommon when the datasets are small as in this case. Therefore, to pick the best model, I chose the F1 score. Based on the F1 score, I decided to use Boosted Model. To summarize the Boosted Model Overall accuracy – 82.35%, F1 score: 85.43%

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

Below is the variable importance plot for the Boosted model.
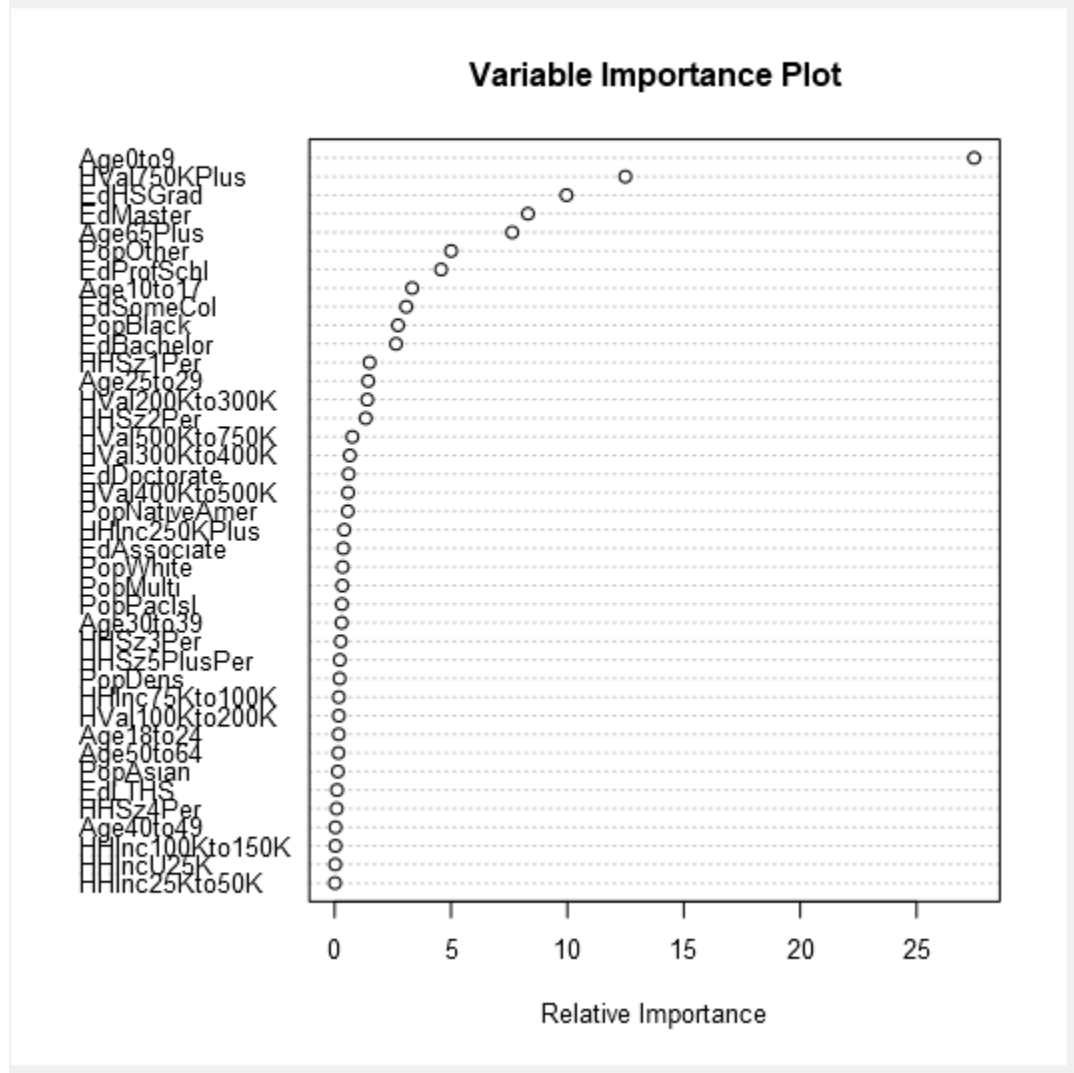


Based on the above plot, the 3 most important variables for the Boosted model are:

| Variable |
| --- |
| Age0to9 |
| HVal750KPlus |
| EdHSGrad |

3. What format do each of the 10 new stores fall into? Please fill in the table below.

| Record # | Store | Cluster |
|----------|-------|---------|
| 1 | S0086 | 1 |
| 2 | S0087 | 2 |
| 3 | S0088 | 3 |
| 4 | S0089 | 2 |
| 5 | S0090 | 2 |
| 6 | S0091 | 1 |
| 7 | S0092 | 2 |
| 8 | S0093 | 1 |
| 9 | S0094 | 2 |
| 10 | S0095 | 2 |

# Task 3: Predicting Produce Sales

1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

   To forecast sales for existing stores, Decomposition plot with the Produce data was generated with the TS-Plot tool.



This is a decomposition plot
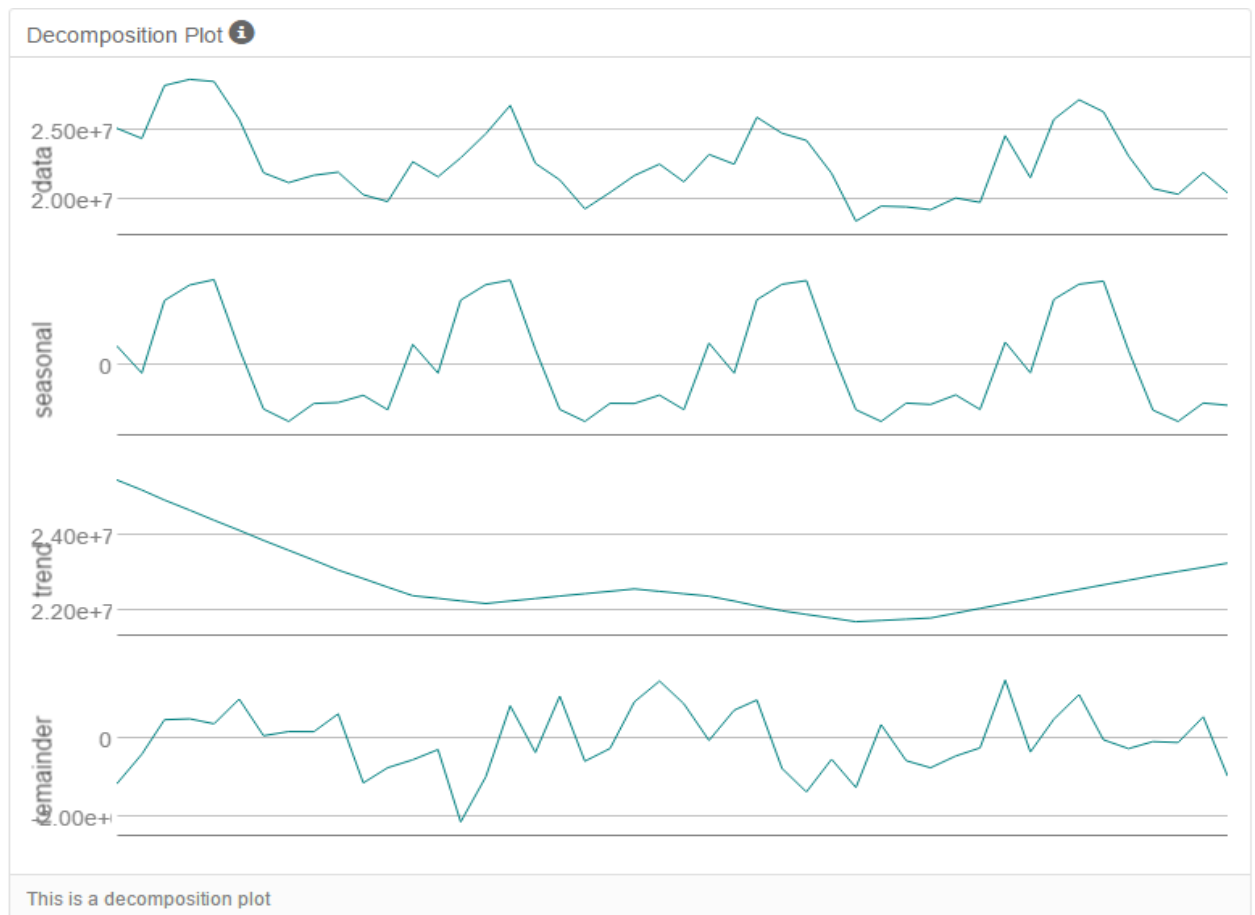
ETS:

From the above Decomposition plot, we can derive the ETS Model.

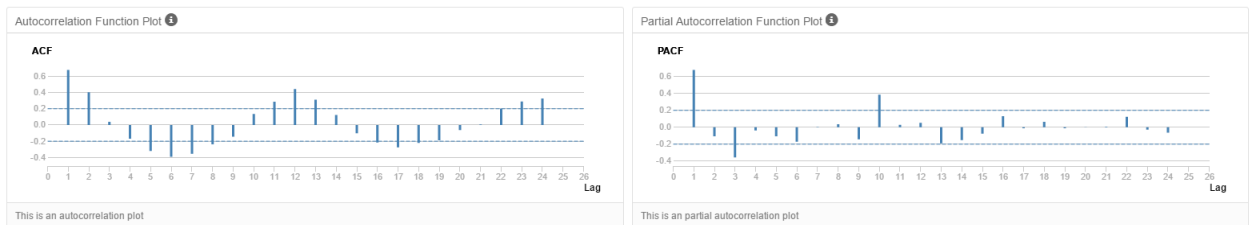Error: The Remainder/Error is increasing in variance and hence it would be Multiplicative.

Trend: There is no clear trend, hence it would be None.

Seasonality: The sales fluctuate in similar intervals, thereby indicating the presence of Seasonality. It can also be observed that the Sales is also growing, hence it would be Multiplicative.

So, based on the above observations, the ETS Model would be ETS(M,N,M).

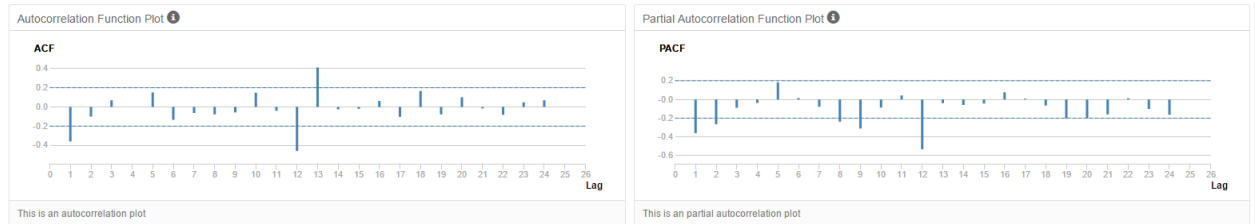| Root Mean Square Error | Akaike Info. Criterion |
|---|---|
| 1015013 | 1089.7 |
| RMSE MAE MPE MAPE MASE | AIC AICc BIC |

ARIMA:



The ACF graph shows a gradual decline to zero with seasonal lag but there is high serial correlation, therefore it is required to be seasonally differenced. The PACF graph also shows high serial correlation after lag1, confirming the need to difference the dataset.

Seasonal Difference ACF & PACF:
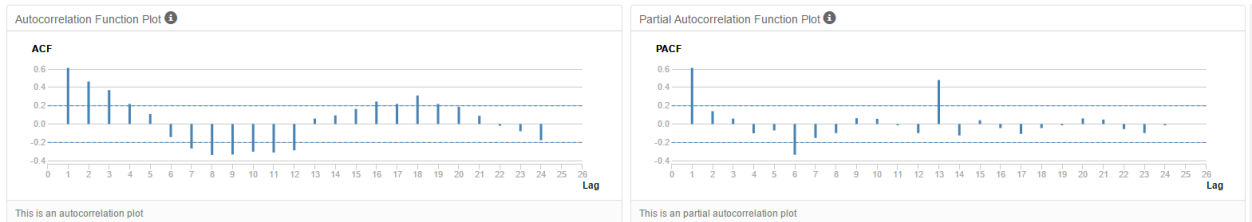


The ACF graph shows positive co-relation at lag 1 and slow decay towards 0 with seasonal lags. However, the serial correlation of is still high. Therefore, we will need to seasonally difference it, to stationarize the time-series

First Seasonal Difference:

Autocorrelation Function Plot / Partial Autocorrelation Function Plot

This is an autocorrelation plot / This is an partial autocorrelation plot

We can observe from the above chart the serial correlation has now subsidized and the time series is now stationary.

- Non-seasonal component: p=0 and q=1 as ACF negative and cuts off sharply, d=1 as mentioned above
- Seasonal component: P=0 and Q=1 as ACF negative at lags 1, 12 and D=1 as mentioned above.
- m=12, because holdout sample = 12

So, we have an ARIMA (0,1,1) (0,1,1)12 model.



Root Mean Square Error
**862271.08**
RMSE MAE MPE MAPE MASE

Akaike Info. Criterion
**661.8**
AIC AICc BIC

Now, comparing the two Model with the below RMSE, MAE, MPE and MASE values, we can observe that ETS performs better than ARIMA. Also, the AIC value of ETS model is higher than the ARIMA Model. So, we will be using ETS model for further forecasting.

| Record # | Model | ME | RMSE | MAE | MPE | MAPE | MASE | NA |
|----------|-------|-----|------|-----|-----|------|------|-----|
| 1 | ETS | 1983592.6926 | 2226512.5538 | 1983592.6926 | 8.4729 | 8.4729 | 1.2691 | [Null] |
| 2 | ARIMA | 2878344.1382 | 3061362.1418 | 2878344.1382 | 12.5815 | 12.5815 | 1.8416 | [Null] |

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three-monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

| Record # | Period | Sub_Period | Forecast_for_NewStores | Forecast_for_ExistingStores | Total_Forecast |
|---|---|---|---|---|---|
| 1 | 2016 | 1 | 2587450.851495 | 21539936.007498 | 24127386.858993 |
| 2 | 2016 | 2 | 2477352.892393 | 20413770.601348 | 22891123.493741 |
| 3 | 2016 | 3 | 2913185.23625 | 24325953.097608 | 27239138.333858 |
| 4 | 2016 | 4 | 2775745.609767 | 22993466.348591 | 25769211.958358 |
| 5 | 2016 | 5 | 3150866.835326 | 26691951.419141 | 29842818.254466 |
| 6 | 2016 | 6 | 3188922.00336 | 26989964.010544 | 30178886.013904 |
| 7 | 2016 | 7 | 3214745.646251 | 26948630.764769 | 30163376.41102 |
| 8 | 2016 | 8 | 2866348.663392 | 24091579.349105 | 26957928.012497 |
| 9 | 2016 | 9 | 2538726.84886 | 20523492.408639 | 23062219.257499 |
| 10 | 2016 | 10 | 2488148.287462 | 20011748.668594 | 22499896.956055 |
| 11 | 2016 | 11 | 2595270.386448 | 21177435.485843 | 23772705.872291 |
| 12 | 2016 | 12 | 2573396.62905 | 20855799.109612 | 23429195.738661 |

## Forecast for Produce

| Month of Date | ExistingStores | NewStores |
|---|---|---|
| January 2016 | 21,539,936 | 2,587,451 |
| February 2016 | 20,413,771 | 2,477,353 |
| March 2016 | 24,325,953 | 2,913,185 |
| April 2016 | 22,993,466 | 2,775,746 |
| May 2016 | 26,691,951 | 3,150,867 |
| June 2016 | 26,989,964 | 3,188,922 |
| July 2016 | 26,948,631 | 3,214,746 |
| August 2016 | 24,091,579 | 2,866,349 |
| September 2016 | 20,523,492 | 2,538,727 |
| October 2016 | 20,011,749 | 2,488,148 |
| November 2016 | 21,177,435 | 2,595,270 |
| December 2016 | 20,855,799 | 2,573,397 |

Actual and Forecast over Time
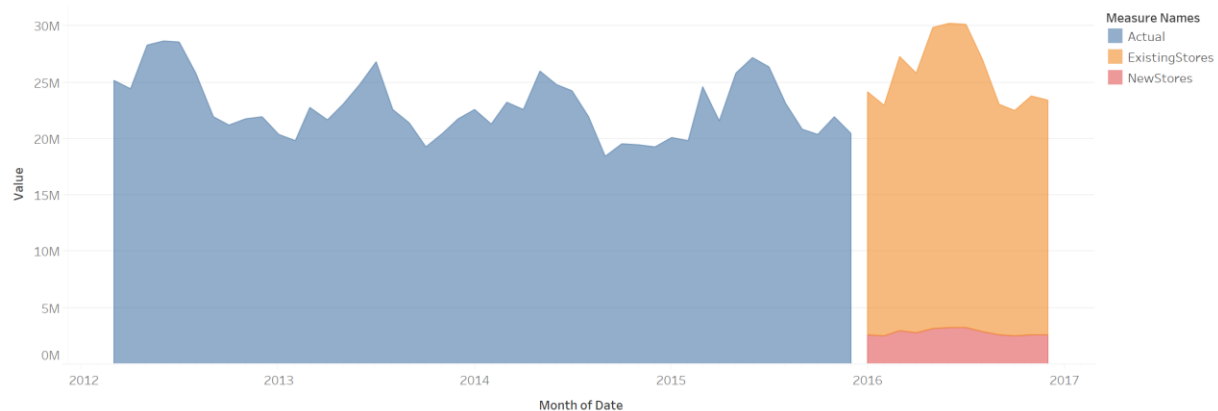


https://public.tableau.com/views/CapstoneProject_14/ForecastTable?:embed=y&:display_count=yes