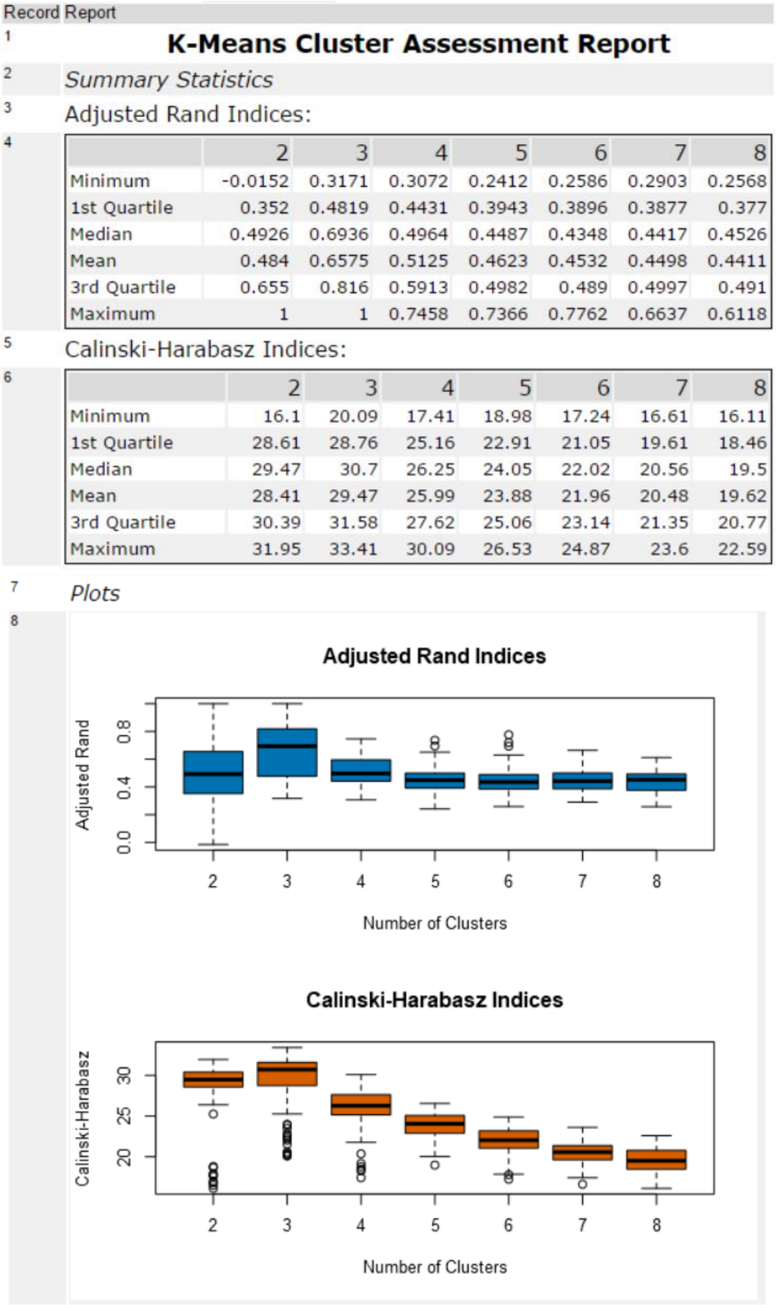
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

To determine the optimum number of store formats we need to run the K-Centroids Diagnostics on Alteryx to get AR and CH indices. As specified, K-means will be the selected methodology and results are shown below.



Record Report

1 **K-Means Cluster Assessment Report**

2 *Summary Statistics*

3 Adjusted Rand Indices:

4

	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

5 Calinski-Harabasz Indices:

6

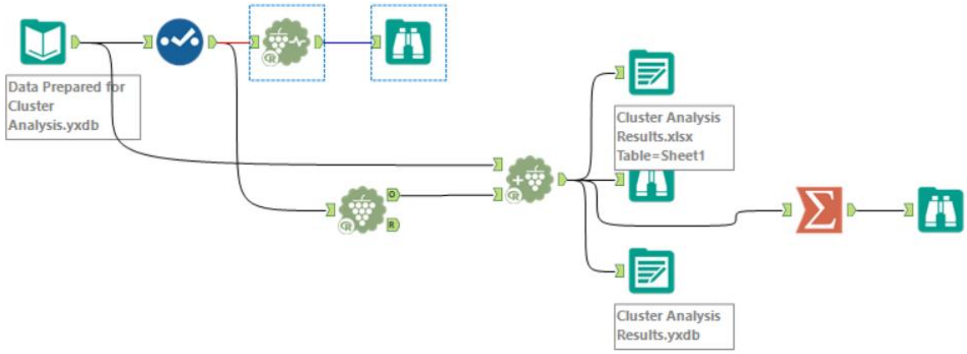
	2	3	4	5	6	7	8
Minimum	16.1	20.09	17.41	18.98	17.24	16.61	16.11
1st Quartile	28.61	28.76	25.16	22.91	21.05	19.61	18.46
Median	29.47	30.7	26.25	24.05	22.02	20.56	19.5
Mean	28.41	29.47	25.99	23.88	21.96	20.48	19.62
3rd Quartile	30.39	31.58	27.62	25.06	23.14	21.35	20.77
Maximum	31.95	33.41	30.09	26.53	24.87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

7 *Plots*

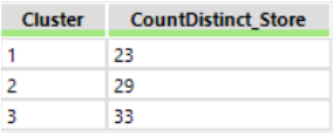
8

A three-cluster solution returns the highest AR (0.6936) and CH (29.47) indices and thus, it is recommended that all stores are configured into one of the three formats.

Cluster Analysis

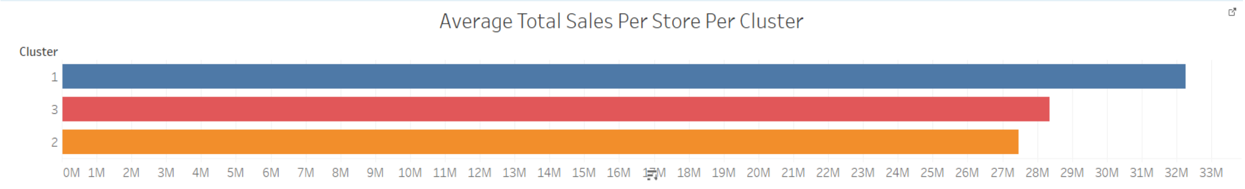

Data Prepared for Cluster Analysis.yxdb
Cluster Analysis Results.xlsx Table=Sheet1
Cluster Analysis Results.yxdb

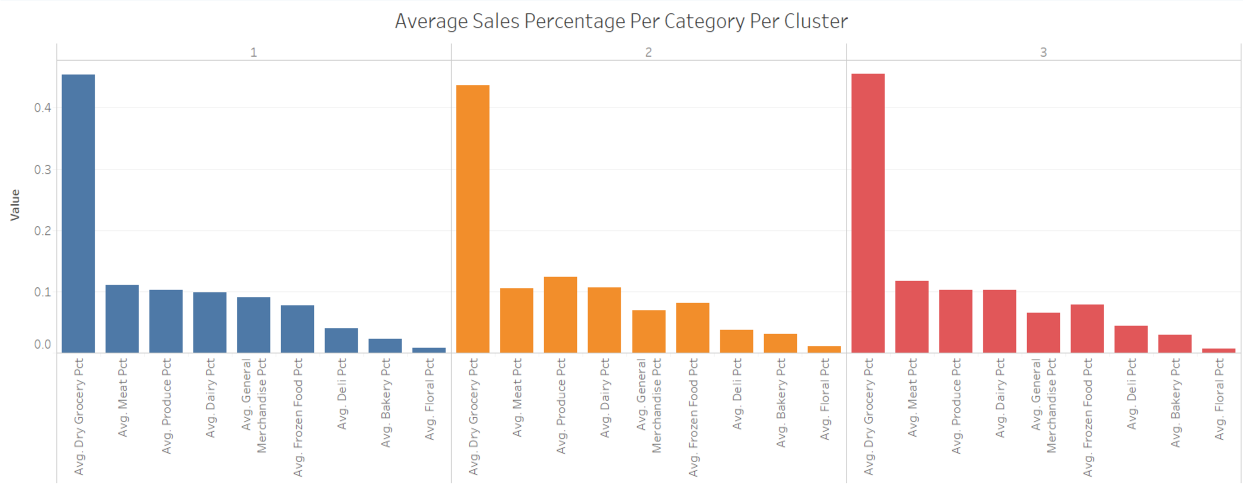
Using the k-means method, a cluster analysis is performed and all existing stores were allocated to one of the three clusters. Distribution of the total no. of stores per cluster is:



Cluster	CountDistinct_Store
1	23
2	29
3	33

Cluster Differences



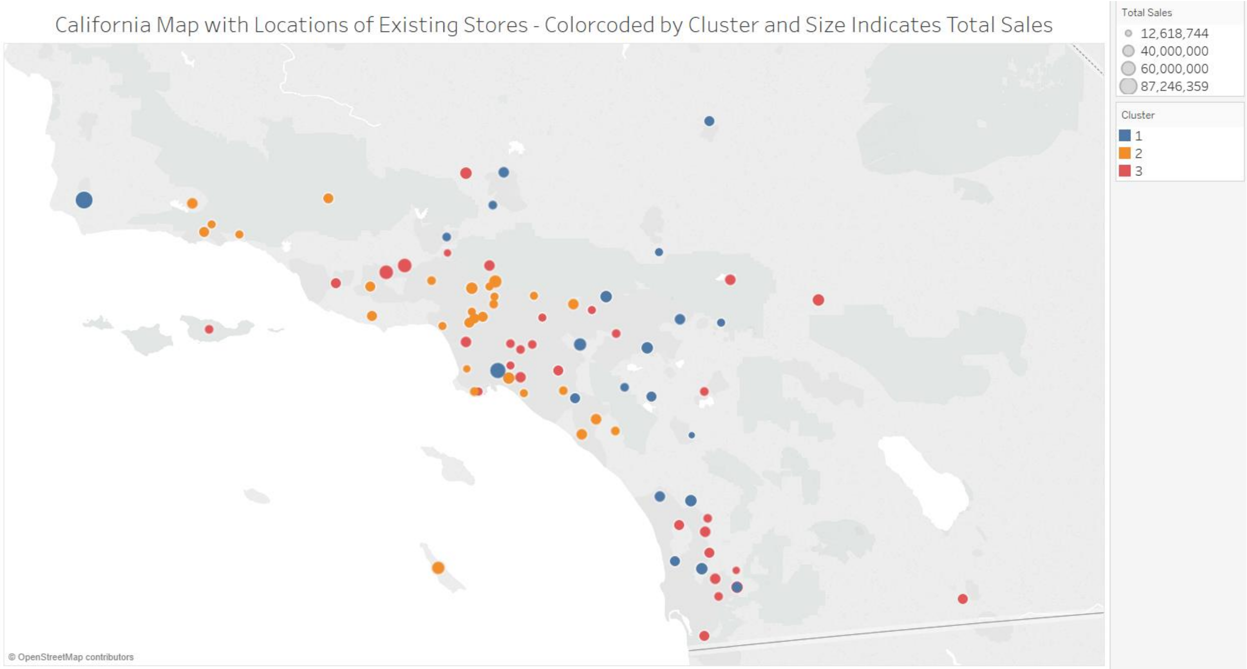


Comparing the three clusters / formats, cluster 1 has the highest average sales indicating that it may have a better mix of products offered than the other clusters. For the percentage sales graphs, we see that cluster 2 did not sell as much meat as cluster 1 as a percentage and Cluster

3 did not sell as much general merchandise. Taking into account that meat is generally more expensive than other offerings, it could be one of the causes of lower sales.

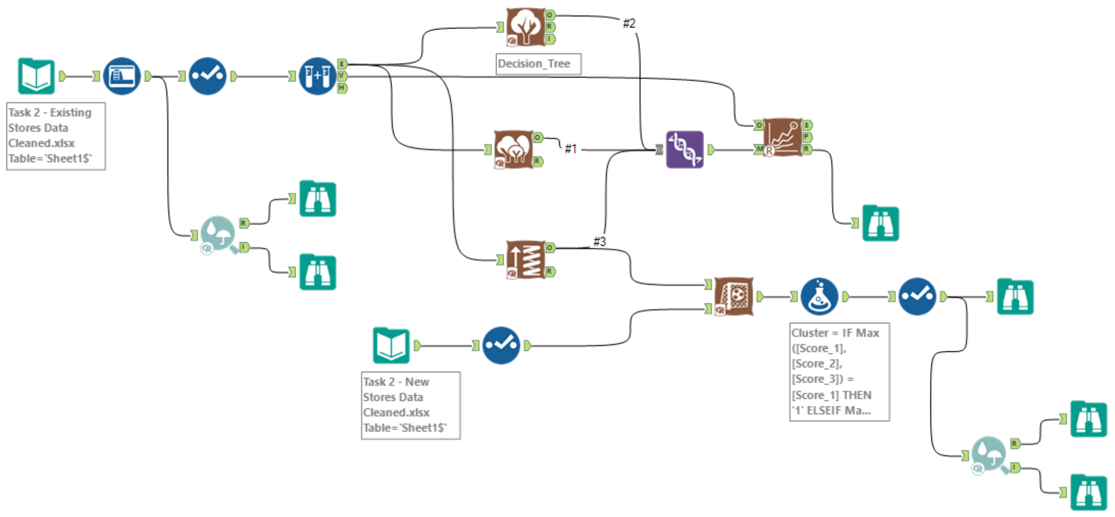
Cluster Visualization

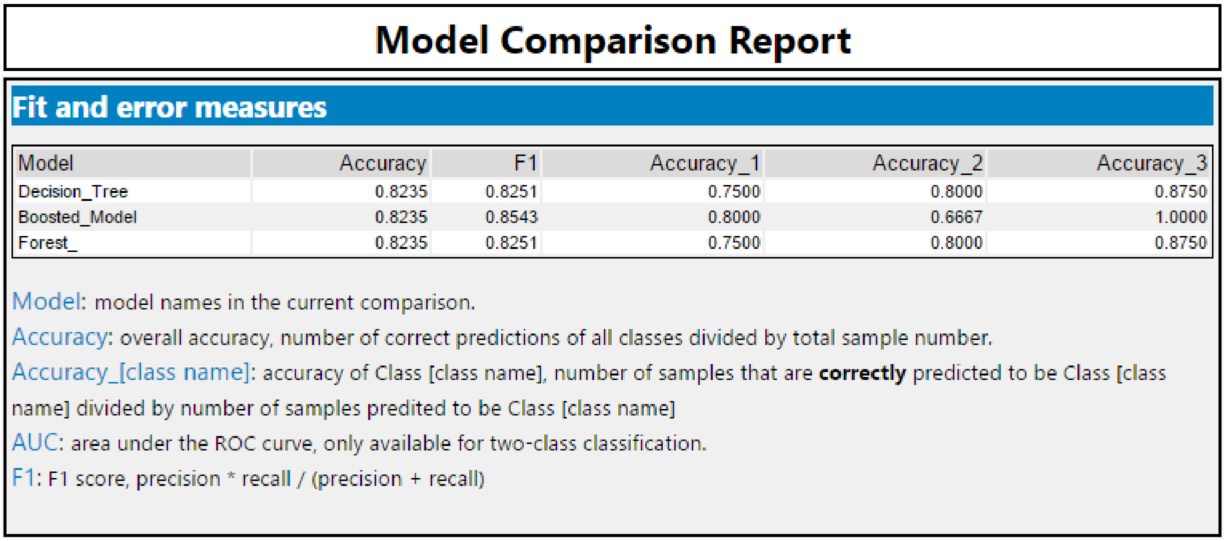
Tableau Public link: California Map with Locations of Existing Stores - Colorcoded by Cluster and Size Indicates Total Sales


California Map with Locations of Existing Stores - Colorcoded by Cluster and Size Indicates Total Sales
Total Sales
12,618,744
40,000,000
60,000,000
87,246,359
Cluster
1
2
3
© OpenStreetMap contributors

Task 2: Formats for New Stores

This is a non-binary classification problem and as such there are three different methodologies to make a format prediction. After preparing two datasets, existing and new stores, by including all the demographic data of the area, the following models were constructed: a) decision tree model, b) forest model, c) boosted model. For this purpose, a validation sample of 20% was withheld so to check and compare accuracy of the models.


Task 2 - Existing Stores Data Cleaned.xlsx Table=`Sheet1$`
Decision_Tree
#2
#1
#3
Cluster = IF Max ([Score_1], [Score_2], [Score_3]) = [Score_1] THEN '1' ELSEIF Ma...
Task 2 - New Stores Data Cleaned.xlsx Table=`Sheet1$`



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Forest_	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)

The models performed equally well against the validation sample, each one achieving a 0.8235 accuracy. For the purposes of the project, the Boosted_Model was selected as it exhibits a higher F1 score than the other two.

Subsequently, the new stores dataset was imported and the stores were allocated to one of the three clusters as shown below.

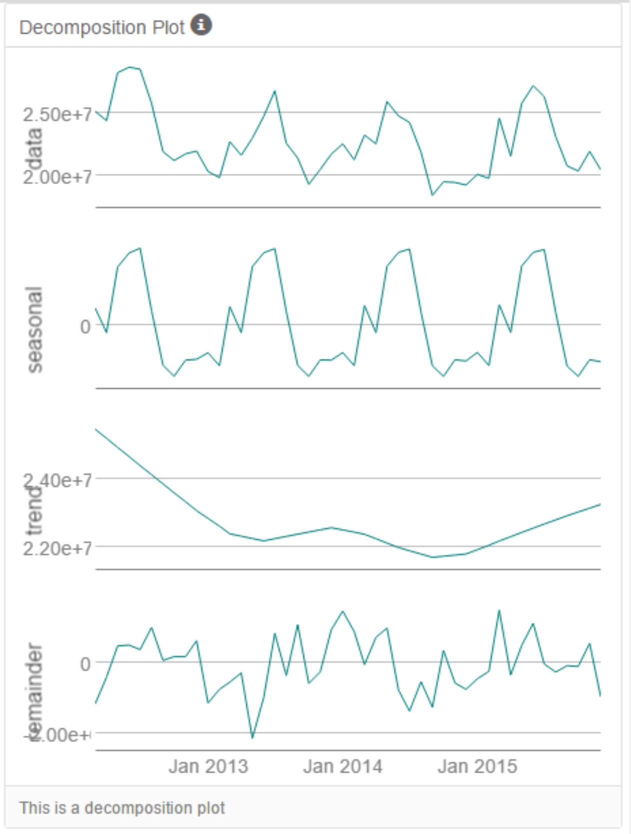
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1

S0094	2
S0095	2

Task 3: Predicting Produce Sales

To forecast the produce sales for existing and new stores we must investigate whether ETS or ARIMA methodology would yield better results. However, it is first necessary to configure the respective models in terms of the ETS(e,t,s) and ARIMA(p,d,q)(P,D,Q)m parameters.

Existing ETS

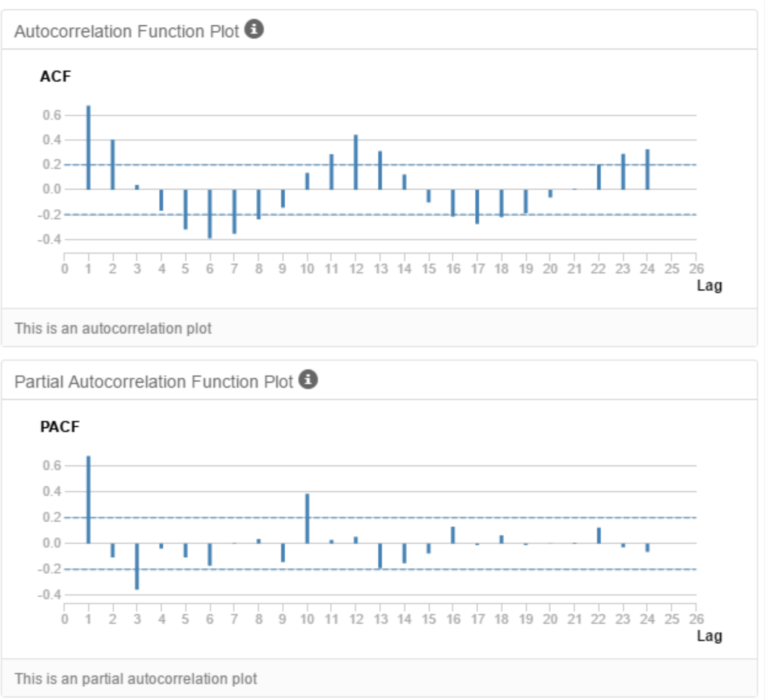

Decomposition Plot

This is a decomposition plot

In this case, the error is basically constant, so the error will be applied additively (a). Apparently, there is no clear trend, which means that t = None (n). The seasonal component is also constant and will be applied additively(a). So we have an ETS(a,n,a) model.

Existing ARIMA

Our time series needs differencing as can be seen by the ACF and PACF plots below.



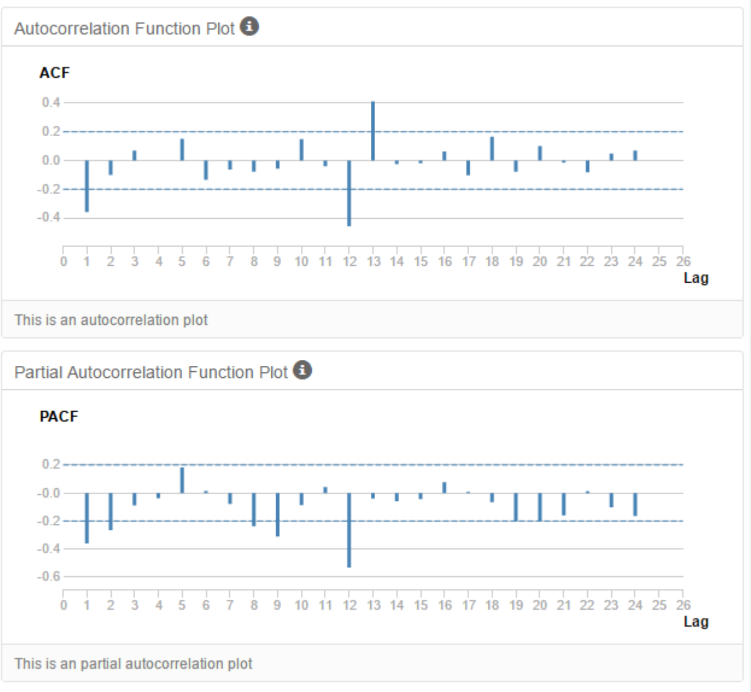
Autocorrelation Function Plot

This is an autocorrelation plot

Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

After differencing twice (one non-seasonal and one seasonal) the ACF and PACF plots are:



Autocorrelation Function Plot

This is an autocorrelation plot

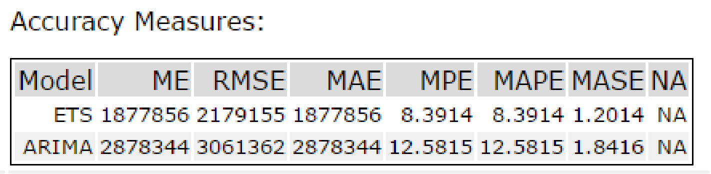
Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

- 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.

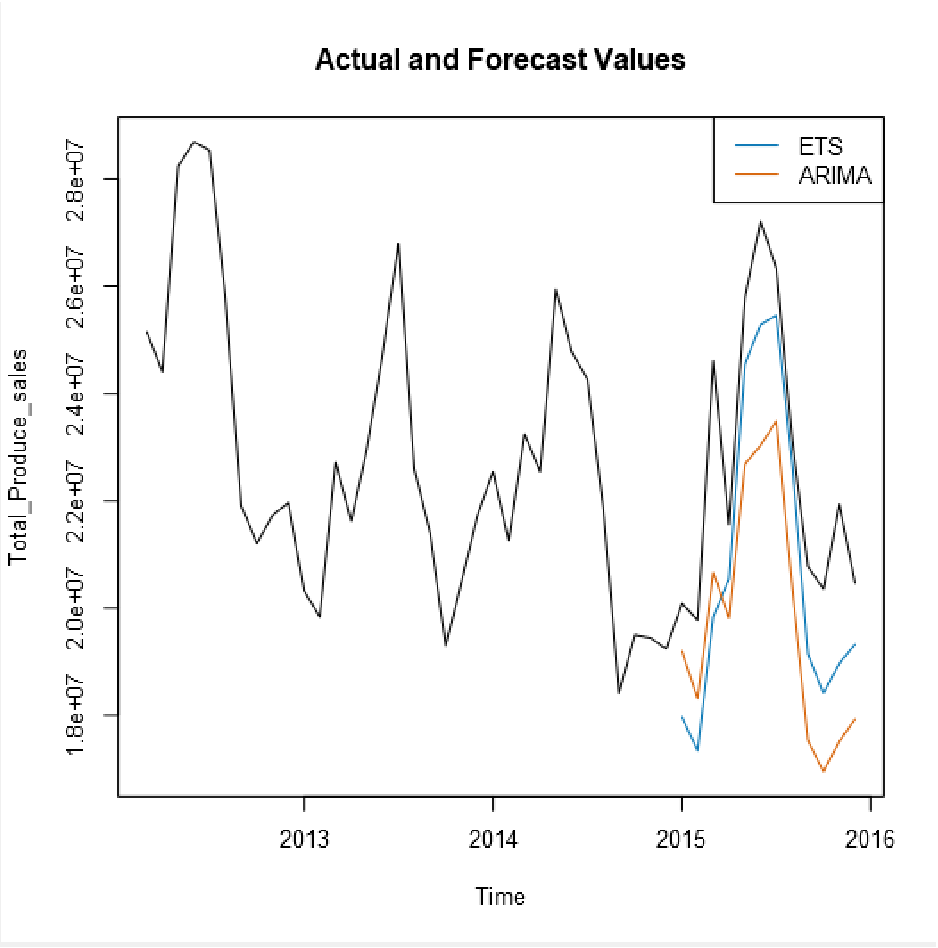
Model Comparison



Accuracy Measures:

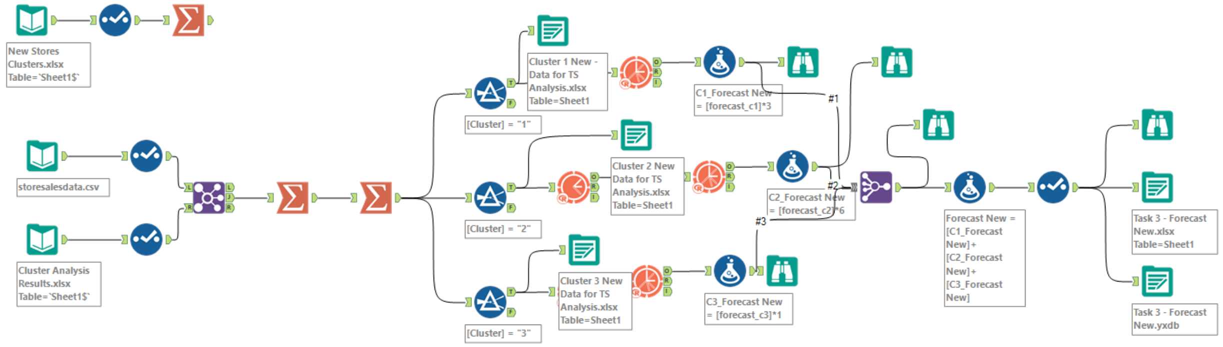
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	1877856	2179155	1877856	8.3914	8.3914	1.2014	NA
ARIMA	2878344	3061362	2878344	12.5815	12.5815	1.8416	NA

ETS performs better than the ARIMA against the holdout sample in every aspect of the analysis by looking at the errors above and can also be verified visually. Hence, this will be an ETS forecast.

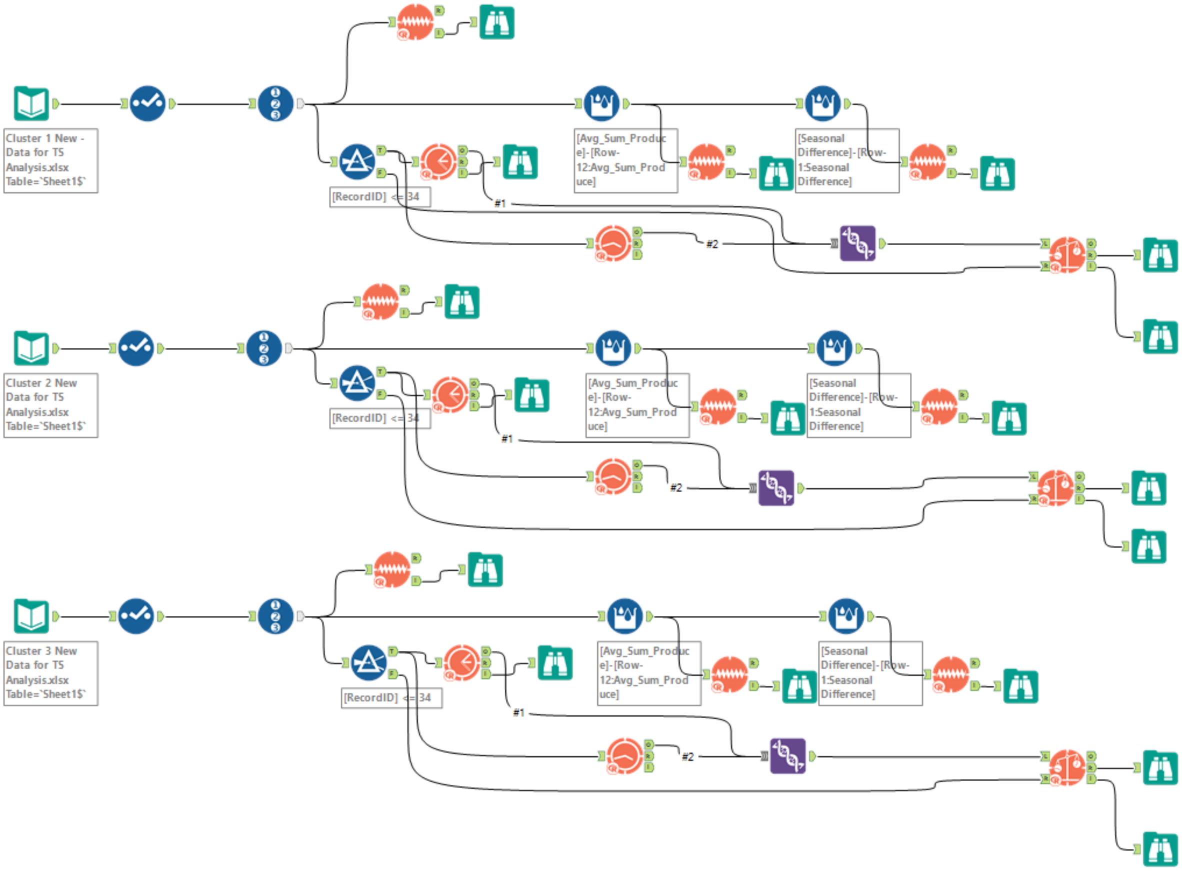


New Stores

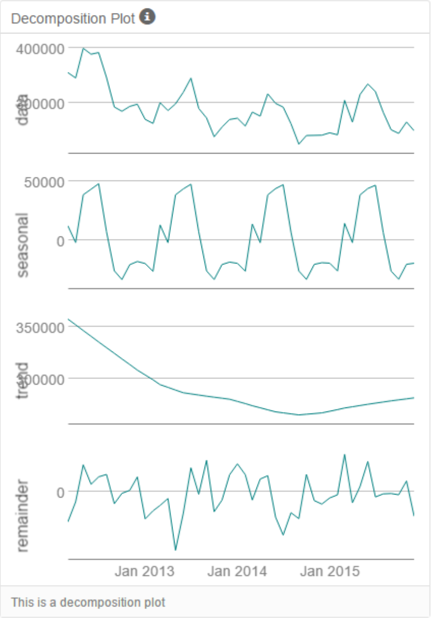
First, the dataset was prepared for analysis by preparing the average produce sales per cluster per month.


New Stores Clusters.xlsx Table=`Sheet1$`
storesalesdata.csv
Cluster Analysis Results.xlsx Table=`Sheet1$`
[Cluster] = "1"
[Cluster] = "2"
[Cluster] = "3"
Cluster 1 New - Data for TS Analysis.xlsx Table=Sheet1
Cluster 2 New Data for TS Analysis.xlsx Table=Sheet1
Cluster 3 New Data for TS Analysis.xlsx Table=Sheet1
C1_Forecast New = [forecast_c1]*3
C2_Forecast New = [forecast_c2]*6
C3_Forecast New = [forecast_c3]*1
#1
#2
#3
Forecast New = [C1_Forecast New]+ [C2_Forecast New]+ [C3_Forecast New]
Task 3 - Forecast New.xlsx Table=Sheet1
Task 3 - Forecast New.yxdb

Consequently, analyses were performed using both methodologies for each cluster.


Cluster 1 New - Data for TS Analysis.xlsx Table=`Sheet1$`
[RecordID] <= 34
[Avg_Sum_Produce]-[Row-12:Avg_Sum_Produce]
[Seasonal Difference]-[Row-1:Seasonal Difference]
#1
#2
Cluster 2 New Data for TS Analysis.xlsx Table=`Sheet1$`
[RecordID] <= 34
[Avg_Sum_Produce]-[Row-12:Avg_Sum_Produce]
[Seasonal Difference]-[Row-1:Seasonal Difference]
#1
#2
Cluster 3 New Data for TS Analysis.xlsx Table=`Sheet1$`
[RecordID] <= 34
[Avg_Sum_Produce]-[Row-12:Avg_Sum_Produce]
[Seasonal Difference]-[Row-1:Seasonal Difference]
#1
#2

Cluster 1


Decomposition Plot

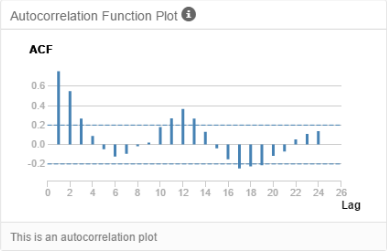
This is a decomposition plot

ETS_C1

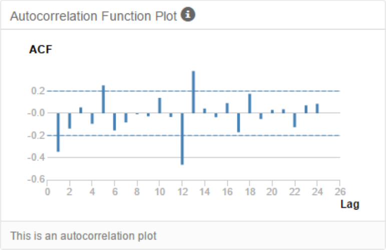
Error constant = additively (a), No trend (n), constant seasonality = additively (a)

- ➔ ETS_c1(A,N,A)

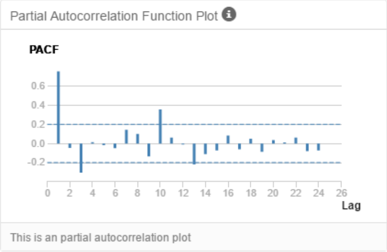
ARIMA_C1


Autocorrelation Function Plot

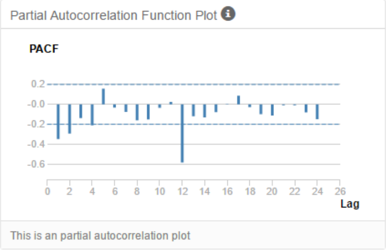
This is an autocorrelation plot


Autocorrelation Function Plot

This is an autocorrelation plot

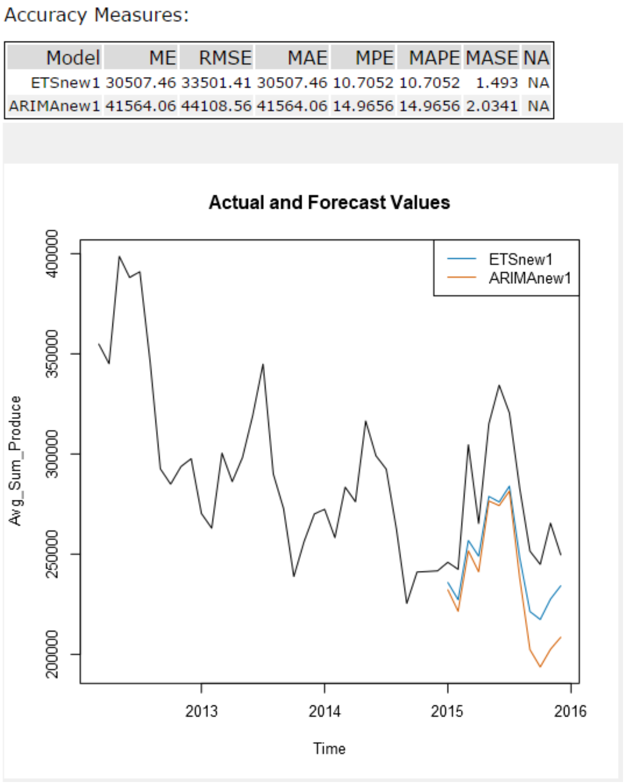

Partial Autocorrelation Function Plot

This is an partial autocorrelation plot


Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

The ACF/PACF plots on the left dictate differencing the series twice (seasonal d=1 and non-seasonal D=1). The differenced series ACF/PACF plots on the right suggest an ARIMA(0,1,1)(0,1,1)12 modelling configuration. This is because the ACF of the differenced series is negative (p=0,q=1) and is also negative at the next seasonal lag = 12 (P=0,Q=1). No. od periods is 12, so m=12.

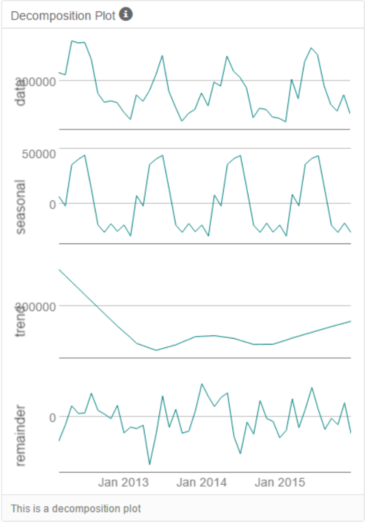


Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETSnew1	30507.46	33501.41	30507.46	10.7052	10.7052	1.493	NA
ARIMAnew1	41564.06	44108.56	41564.06	14.9656	14.9656	2.0341	NA

The ETS model outperforms the ARIMA model in every aspect of the analysis ETS_C1 will be used for forecasting.

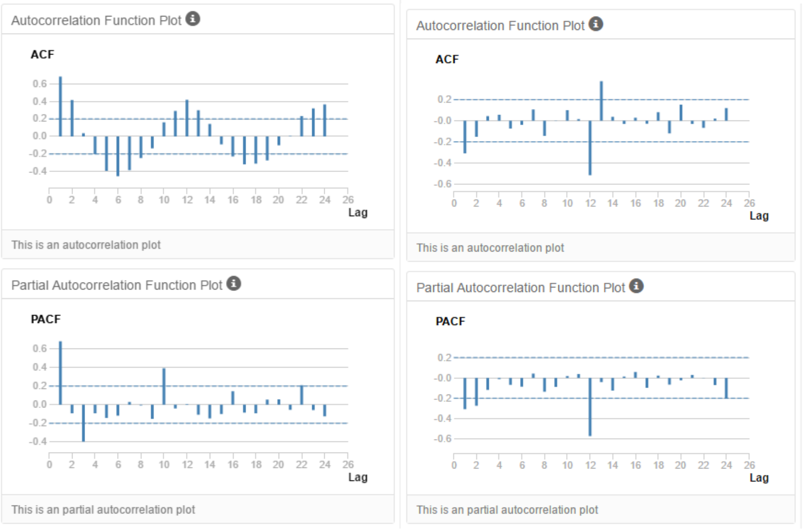
Cluster 2



ETS_C2

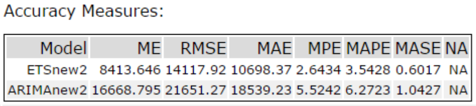
The error is constant, there is no trend and seasonality is constant → ETS_C2(A,N,A).

ARIMA_C2



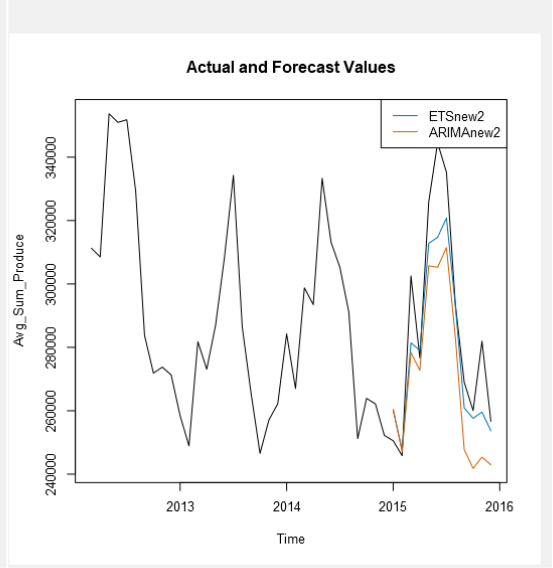
ACF of the differenced series on the right is negative both at lag1 and lag 12 and the time series was differenced twice (seasonal, non-seasonal) and the no. of periods is 12. So, for the same reasons as cluster 1 this is an ARIMA(0,1,1)(0,1,1)12 model.

Model Comparison



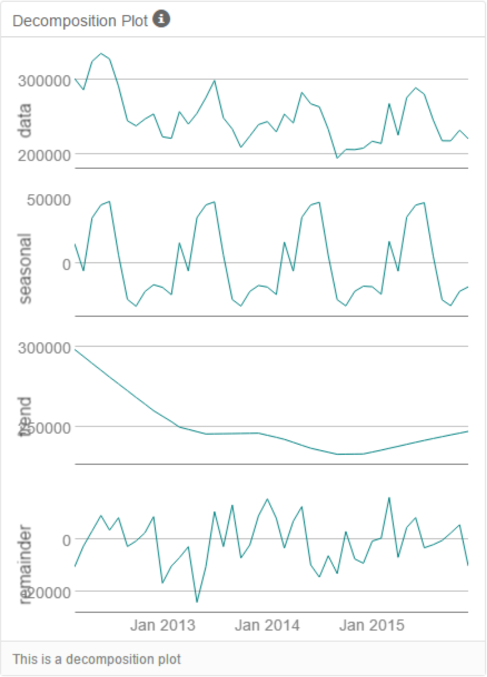
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETSnew2	8413.646	14117.92	10698.37	2.6434	3.5428	0.6017	NA
ARIMAnew2	16668.795	21651.27	18539.23	5.5242	6.2723	1.0427	NA



The ETS model outperforms the ARIMA model in every aspect and can be verified visually, so the ETS_C2 will be used to forecast produce sales.

Cluster 3

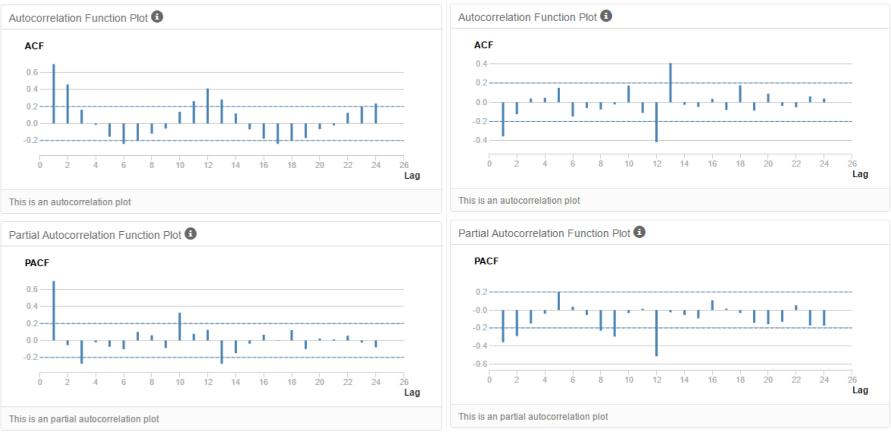


This is a decomposition plot

ETS_C3: Error is constant (a), no trend (n), seasonality is constant (a) -> ETS_C3(A, N, A).

ARIMA_C3

First, we need to investigate if the time series needs differencing by looking at the ACF/PACF plots.



This is an autocorrelation plot

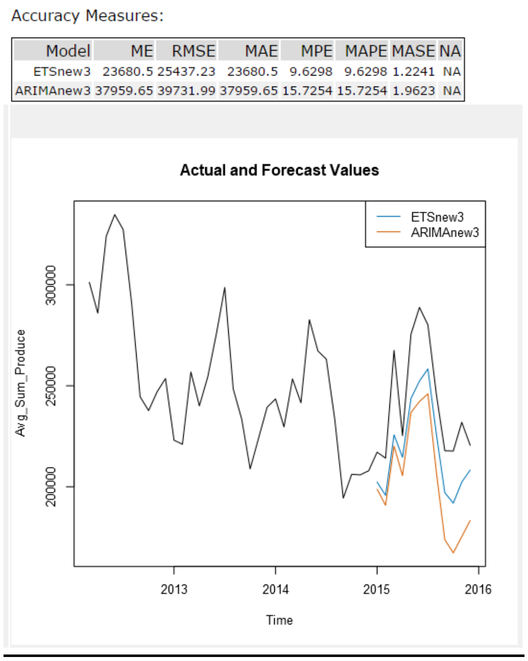
This is an autocorrelation plot

This is an partial autocorrelation plot

This is an partial autocorrelation plot

The ACF of the twice differenced series is negative in lag 1 and lag12, so we would consider an ARIMA(0,1,1)(0,1,1)12 model, where 12 is the no. of periods.

Model comparison



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETSnew3	23680.5	25437.23	23680.5	9.6298	9.6298	1.2241	NA
ARIMAnew3	37959.65	39731.99	37959.65	15.7254	15.7254	1.9623	NA

The ETS_C3(A, N, A) model has a better performance, as can be seen by the accuracy measurements and the comparison graph so it will be used for the forecast.

Subsequently, all new stores TS forecasting results were added together to obtain a 12 month forecasting table for produce sales in 2016.

Forecasted Produce Sales

Link to Tableau Public Visualization – Task 3