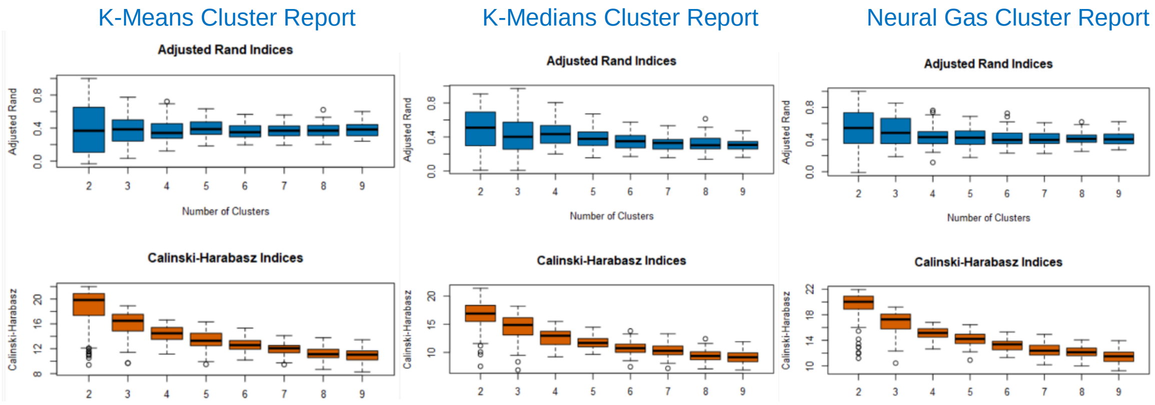
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

The optimal number of store formats is 3.

When looking at the 3 K-Centroids methods box-whisker plots, we can see that 2 and 3 clusters showed the highest median values on each index, meaning the closest similarity between store segments (AR Index) and the best compactness and distinctness of the clusters (CH Index).


K-Means Cluster Report

K-Medians Cluster Report

Neural Gas Cluster Report

When looking closer to the 2 and 3 Cluster Diagnostics, we see that Clustering into 3 store segments gives a better combination of low Interquartile range (IQR) and high Median values on K-Means method, which is why I selected this model and 3 as the number of clusters.

	K-Means			
	AR	CH	AR	CH
# of clusters	2	2	3	3
Median	0.36	19.83	0.38	16.5
IQR	0.53	3.34	0.25	2.69

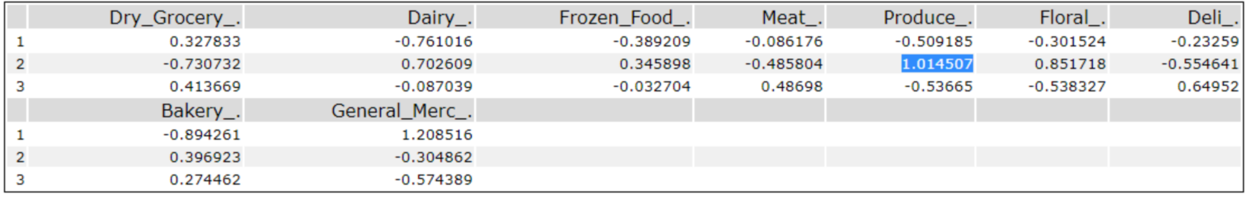
	K-Medians			
	AR	CH	AR	CH
# of clusters	2	2	3	3
Median	0.5	16.89	0.4	14.87
IQR	0.39	1.47	0.32	3.03

	Neural Gas Median			
	AR	CH	AR	CH
# of clusters	2	2	3	3
Median	0.54	20.03	0.48	17.27
IQR	0.37	2.01	0.32	2.27

I assigned the 85 stores into 3 clusters of 23, 29 and 33 stores, for Cluster 1, Cluster 2 and Cluster 3 respectively.

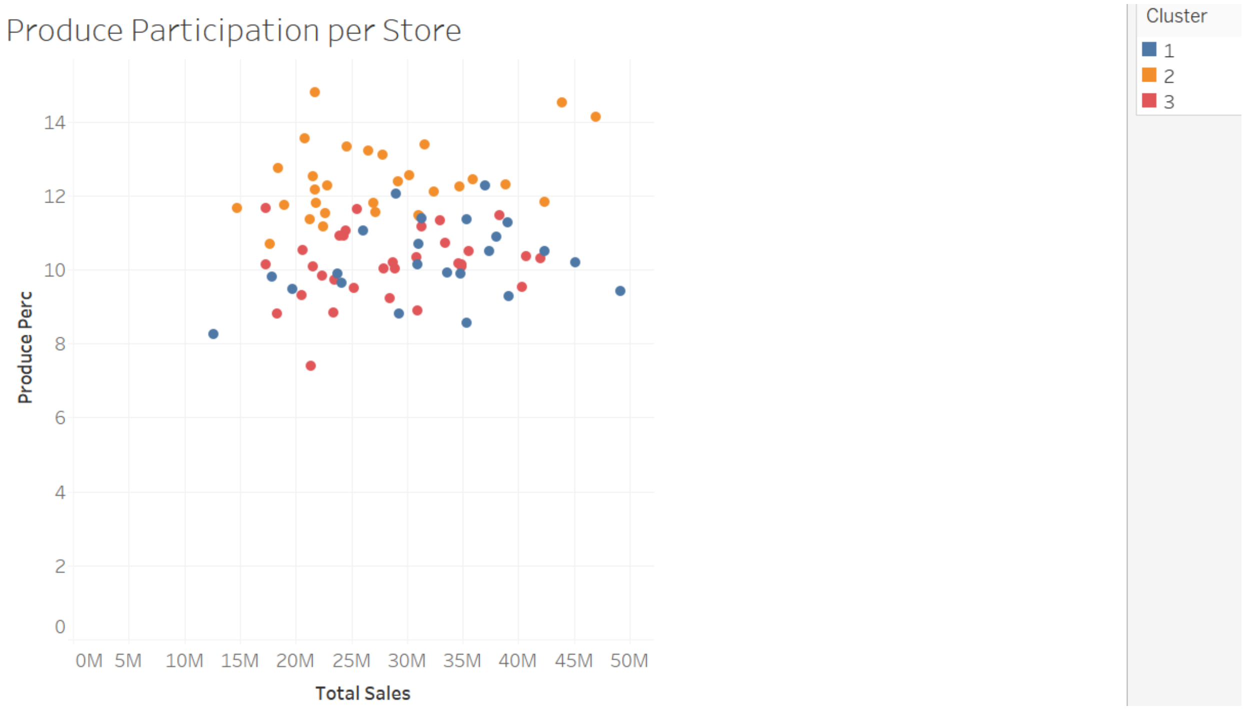
The Produce sales percentage is a great distinguisher of clusters on this data (the greatest distance between clusters: -0.509185, 1.014507 and -0.5366.

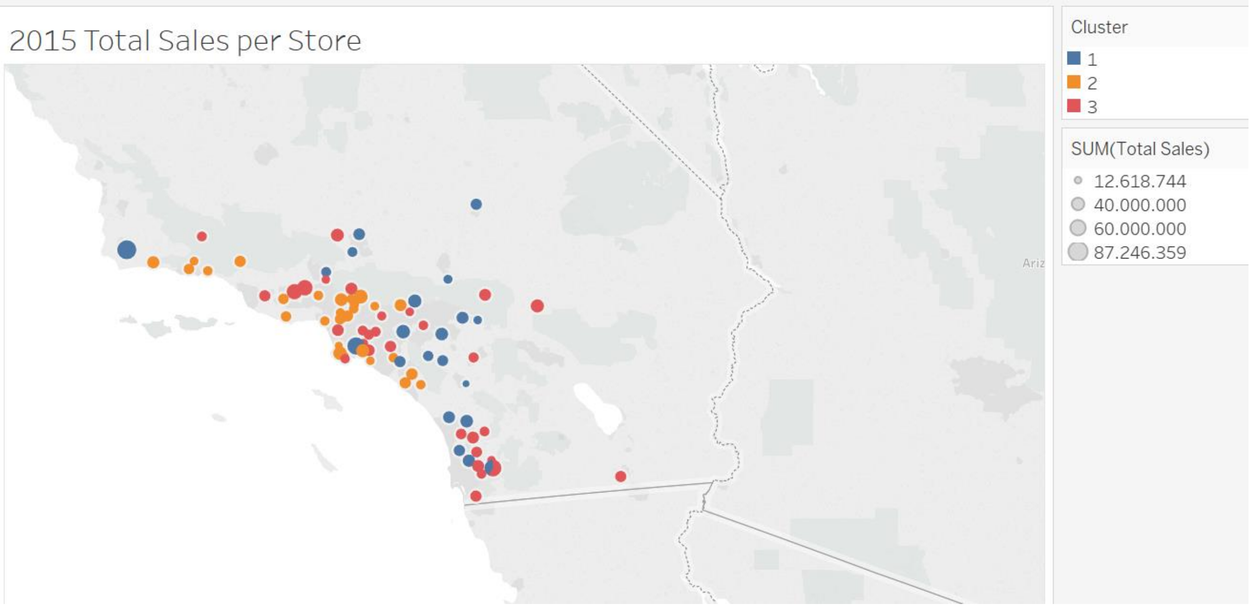
Summary Report of the K-Means Clustering Solution Store_Formats



	Dry_Grocery_.	Dairy_.	Frozen_Food_.	Meat_.	Produce_.	Floral_.	Deli_.
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Bakery_.	General_Merc_.
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389





Task 2: Formats for New Stores

I used the Random Forest Model because it shows a higher Accuracy (82%) and a lower bias compared to the other Models. As showed on the tables below, the Random Forest has the closest values between Positive Predictive Values (0.75) and Negative Predictive Values (0.80), which is the lowest Bias of our models.

Boosted Model	**Actual_1⍰**	**Actual_2**	**Actual_3**
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_2	0	0	6
Total	17		
Positive Predictive Values	**0.8000**		
Negative Predictive Values	**0.6667**		
Accuracy	**0.8235**		

Decision Tree Model	**Actual_1⍰**	**Actual_2**	**Actual_3**
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_2	1	0	5
Total	17		
Positive Predictive Values	**0.6000**		
Negative Predictive Values	**0.6667**		
Accuracy	**0.7059**		

Random Forest Model	**Actual_1⍰**	**Actual_2**	**Actual_3**
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_2	1	0	7
Total	17		
Positive Predictive Values	**0.7500**		
Negative Predictive Values	**0.8000**		
Accuracy	**0.8235**		

Predicted Cluster of the 10 new stores

Store Number	Segment
S0086	3
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

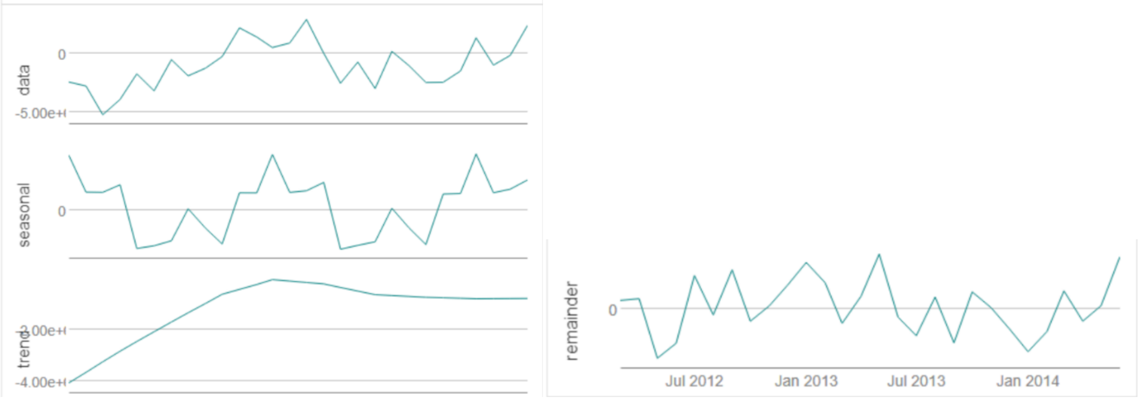
Before selecting the model to predict Produce Sales, I explained what components each model would use.

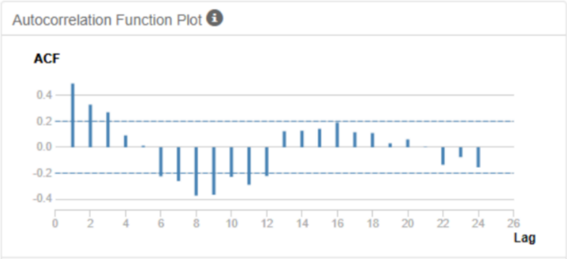
ARIMA (1,0,0) (1,1,0) [12]

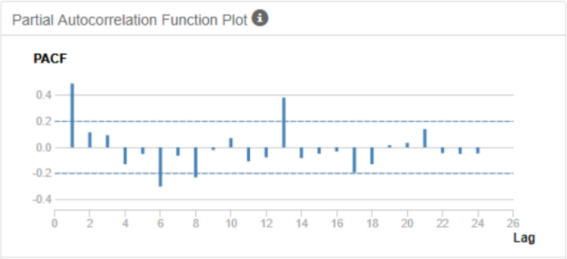
I used **(1,0,0)** on the Non-Seasonal part;
1 ➔ 1 autoregressive component was enough to make the time series stationary
0 ➔ as I just used a Seasonal Difference to make the time series stationary, I did not use any First Difference term on the model. Also, there is no trend on the data, so we should not apply non-seasonal differencing
0 ➔ I did not use moving average on the model because seasonal autocorrelation is positive

I used **(1,1,0)** on the Seasonal part;
1 ➔ as 1 seasonal autoregressive component was enough to make the time series stationary
1 ➔ as I used a Seasonal Difference to make the time series stationary
0 ➔ I did not use moving average on the model because seasonal autocorrelation is positive, ACF slowly decays and PACF drops off substantially at first lag

[12] refers to the 12 months forecast



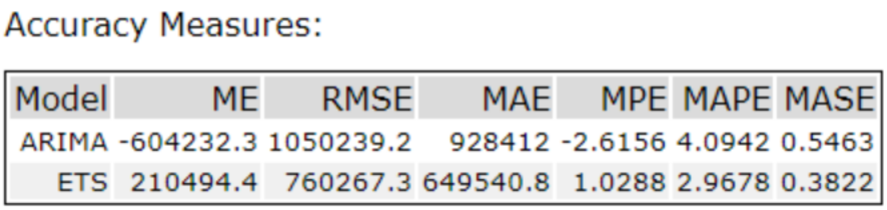

Autocorrelation Function Plot


Partial Autocorrelation Function Plot

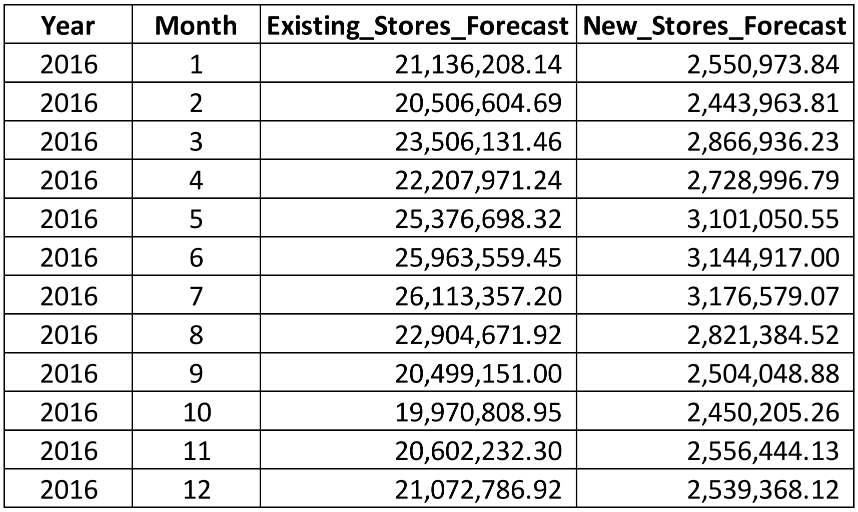
ETS (M,N,M)

As the time series model has an increasing error, no trend and a growing/shrinking seasonality overtime, I assigned a Multiplicative(M), None(N) and Multiplicative(M) components to the ETS model.

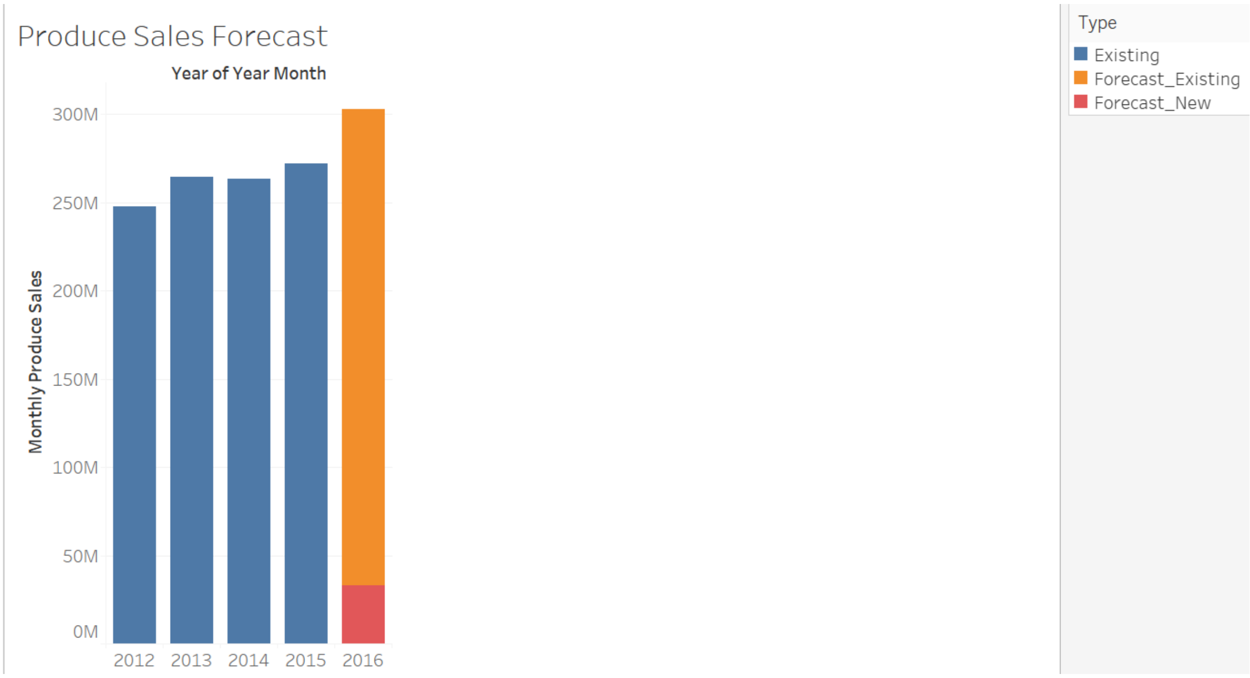
I decide to use the ETS model because it showed a lower average of the difference between actual and forecasted values (ME), a lower sample standard deviation of the differences between predicted values and observed values (RMSE), a lower Mean Absolute Error on the forecast (MAE), a lower average of the percent difference between actual and forecasted values (MPE), a lower Mean Absolute Percentage Error (MAPE) and a lower Mean Absolute Scaled Error (MASE).


Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822



Year	Month	Existing_Stores_Forecast	New_Stores_Forecast
2016	1	21,136,208.14	2,550,973.84
2016	2	20,506,604.69	2,443,963.81
2016	3	23,506,131.46	2,866,936.23
2016	4	22,207,971.24	2,728,996.79
2016	5	25,376,698.32	3,101,050.55
2016	6	25,963,559.45	3,144,917.00
2016	7	26,113,357.20	3,176,579.07
2016	8	22,904,671.92	2,821,384.52
2016	9	20,499,151.00	2,504,048.88
2016	10	19,970,808.95	2,450,205.26
2016	11	20,602,232.30	2,556,444.13
2016	12	21,072,786.92	2,539,368.12



Here is the shared folder with all the tables, Alteryx and Tableau files used on the Project:

https://drive.google.com/open?id=1T3AnbQebJFNZms3olzk4f0BETADN4Zcr