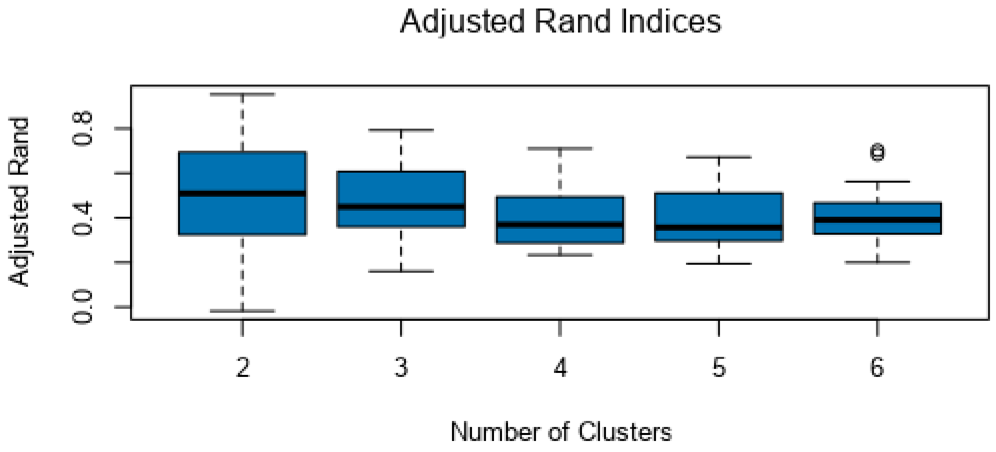
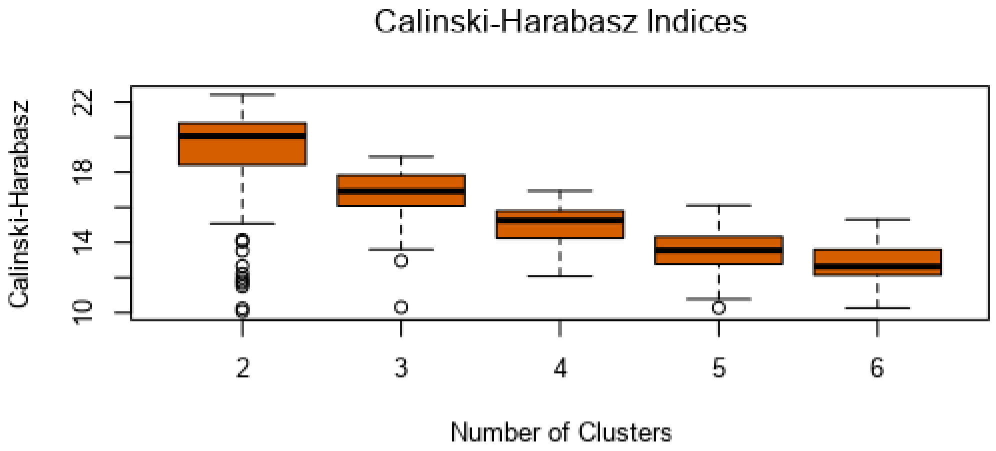
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

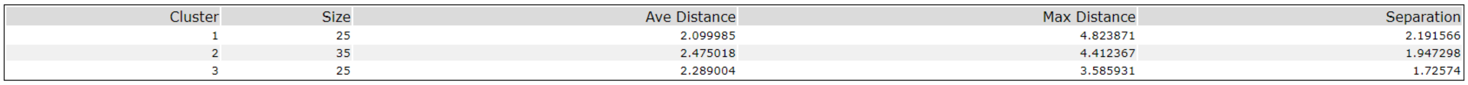
The optimal number of store formats is 3. I came to that conclusion after running a K-Means Cluster Assessment and checking the Adjusted Rand and CH Indeces, using 2015 data and each category percent on total sales as inputs. The indices indicated that 2 and 3 clusters would give me the higher median among the tested options, but the option for 3 clusters was made since there is a smaller variance. This is an indicator that this option gives both the smaller distance between stores in the same cluster and the maximum separation between clusters.





2. How many stores fall into each store format?

Cluster 1 has 25 stores, cluster 2 has 35 stores and cluster 3 has 25 stores.

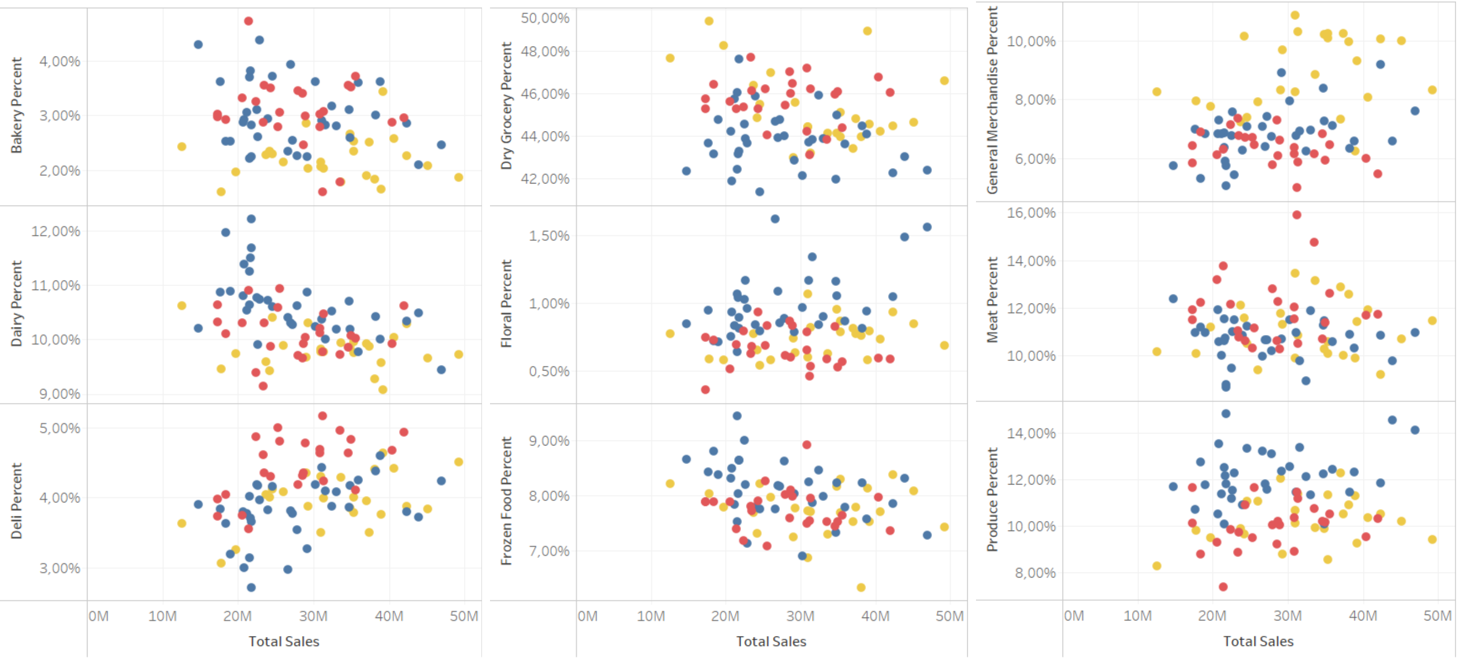


Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

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

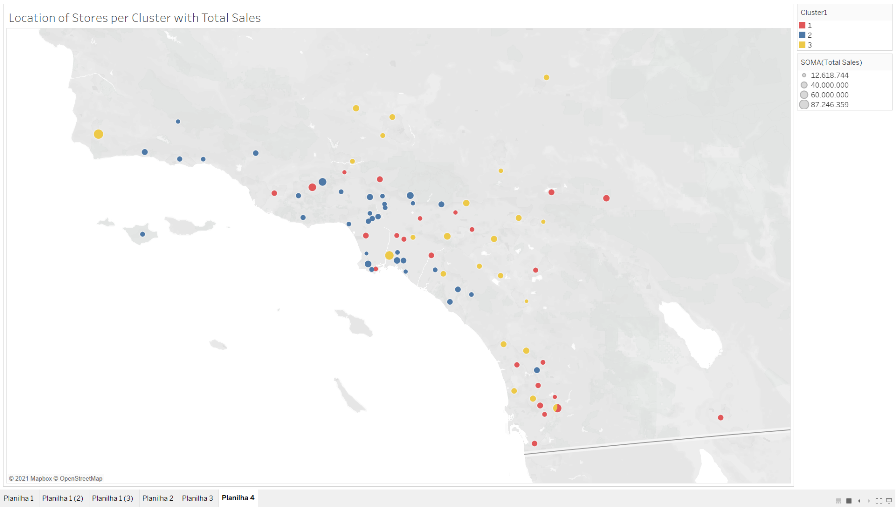
Cluster
🟥 1
🟦 2
🟨 3

Cluster 1 has a higher share of sales in General Merchandise, Cluster 2 has higher shares of Dry Groceries and Deli than the other segments, and Cluster 3 has a higher share of Floral and Produce.



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.

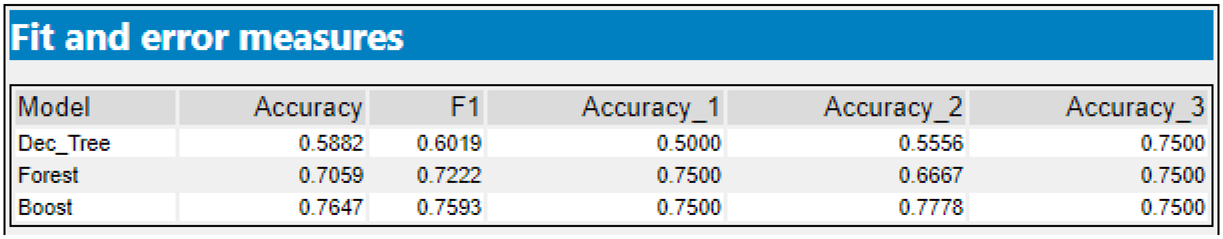
https://public.tableau.com/profile/felipe.soares.pereira#!/vizhome/data_viz_16174883057430/Planilha4



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.)

I designed 3 models to predict store formats using Decision Three, Random Forest and Boosted methods. The last one had the best overall accuracy and F1 score, so I chose this methodology to predict the best store format for the new stores.

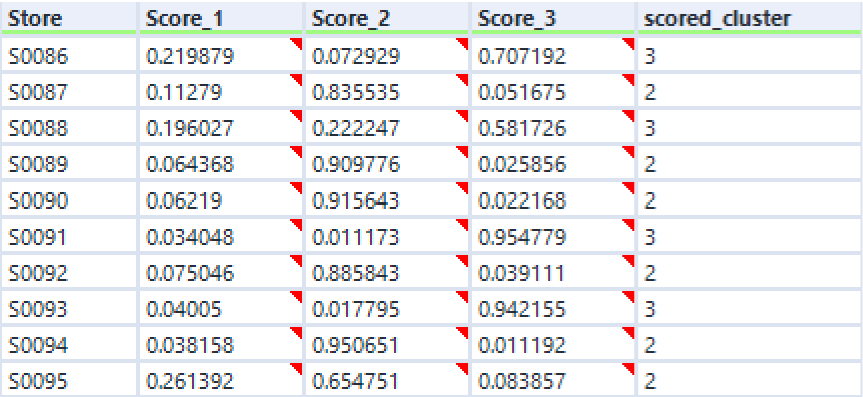


Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Dec_Tree	0.5882	0.6019	0.5000	0.5556	0.7500
Forest	0.7059	0.7222	0.7500	0.6667	0.7500
Boost	0.7647	0.7593	0.7500	0.7778	0.7500

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

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

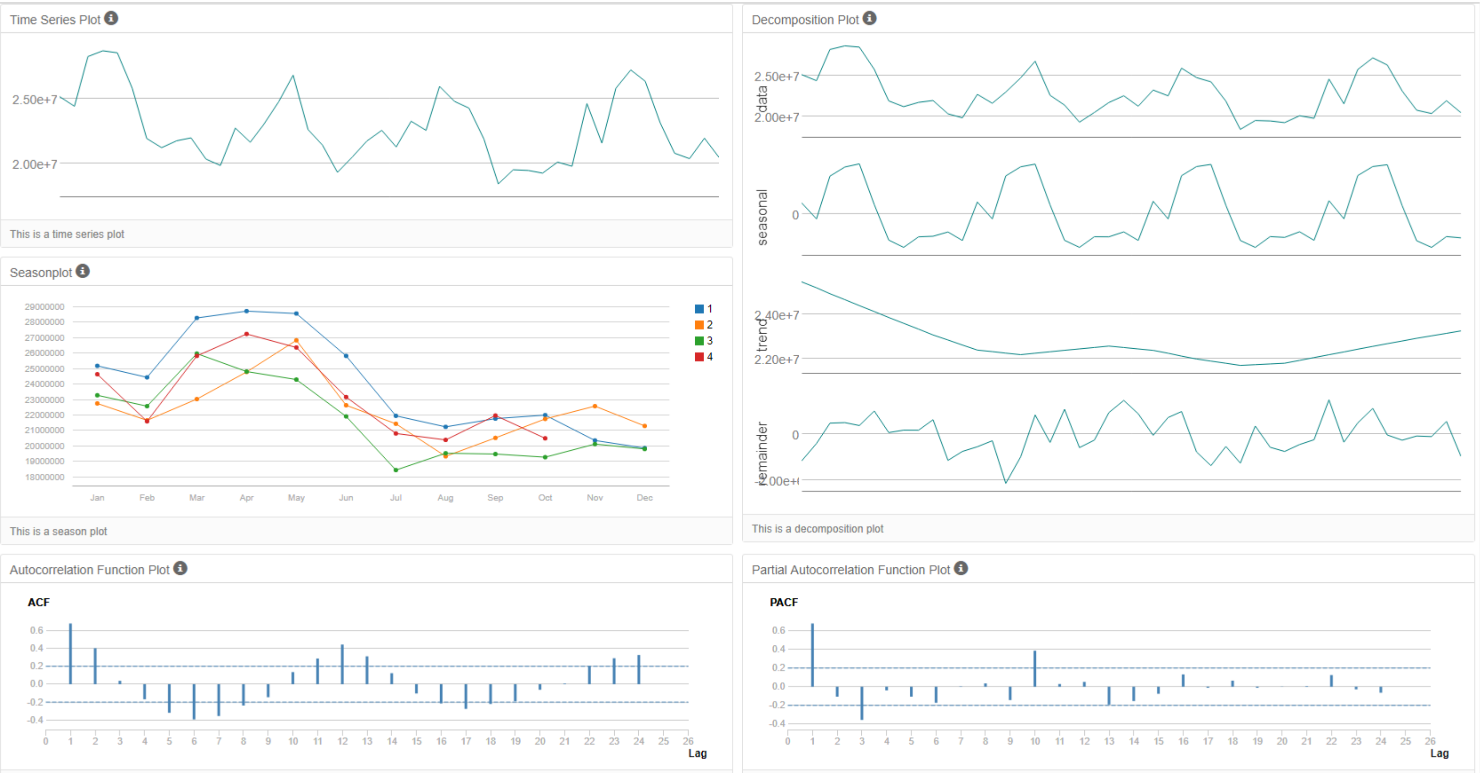


Store	Score_1	Score_2	Score_3	scored_cluster
S0086	0.219879	0.072929	0.707192	3
S0087	0.11279	0.835535	0.051675	2
S0088	0.196027	0.222247	0.581726	3
S0089	0.064368	0.909776	0.025856	2
S0090	0.06219	0.915643	0.022168	2
S0091	0.034048	0.011173	0.954779	3
S0092	0.075046	0.885843	0.039111	2
S0093	0.04005	0.017795	0.942155	3
S0094	0.038158	0.950651	0.011192	2
S0095	0.261392	0.654751	0.083857	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?

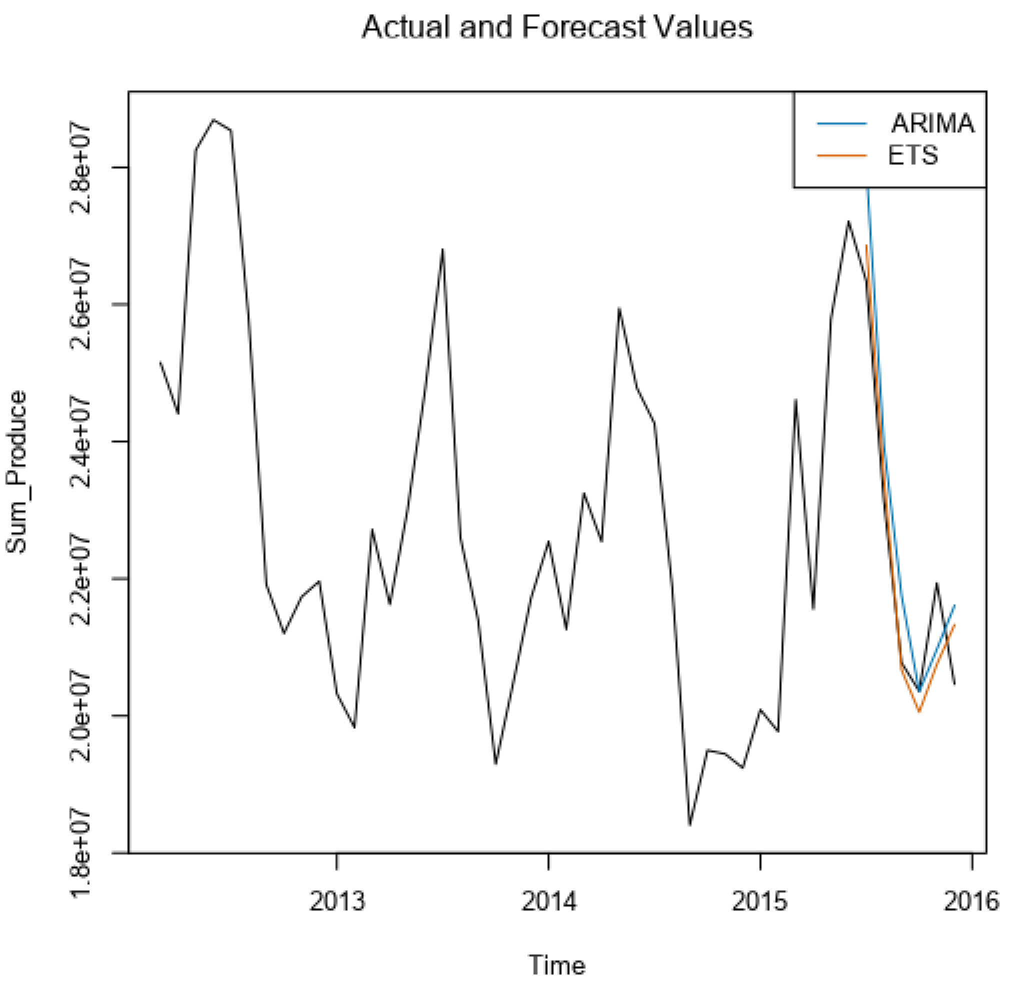
I evaluated the Time Series Decomposition and the Auto-correlation and Partial Auto-correlation plots of the time series and its differentiations (both simple and seasonal). This is the TS Plot of the original data:



There is a strong indication of seasonality on the time series, with no constant trend and a remainder that shows a lot of variation, which calls for a ETS(M,N,M). The ACF and PACF plots of the difference and seasonal difference indicate that there should be an autoregressive component in the ARIMA model. For the forecast, I used an ARIMA(1,0,0)(1,1,0)[12]. Comparing both models, the ETS showed less error (ME, RMSE) for the 6 months holdout:

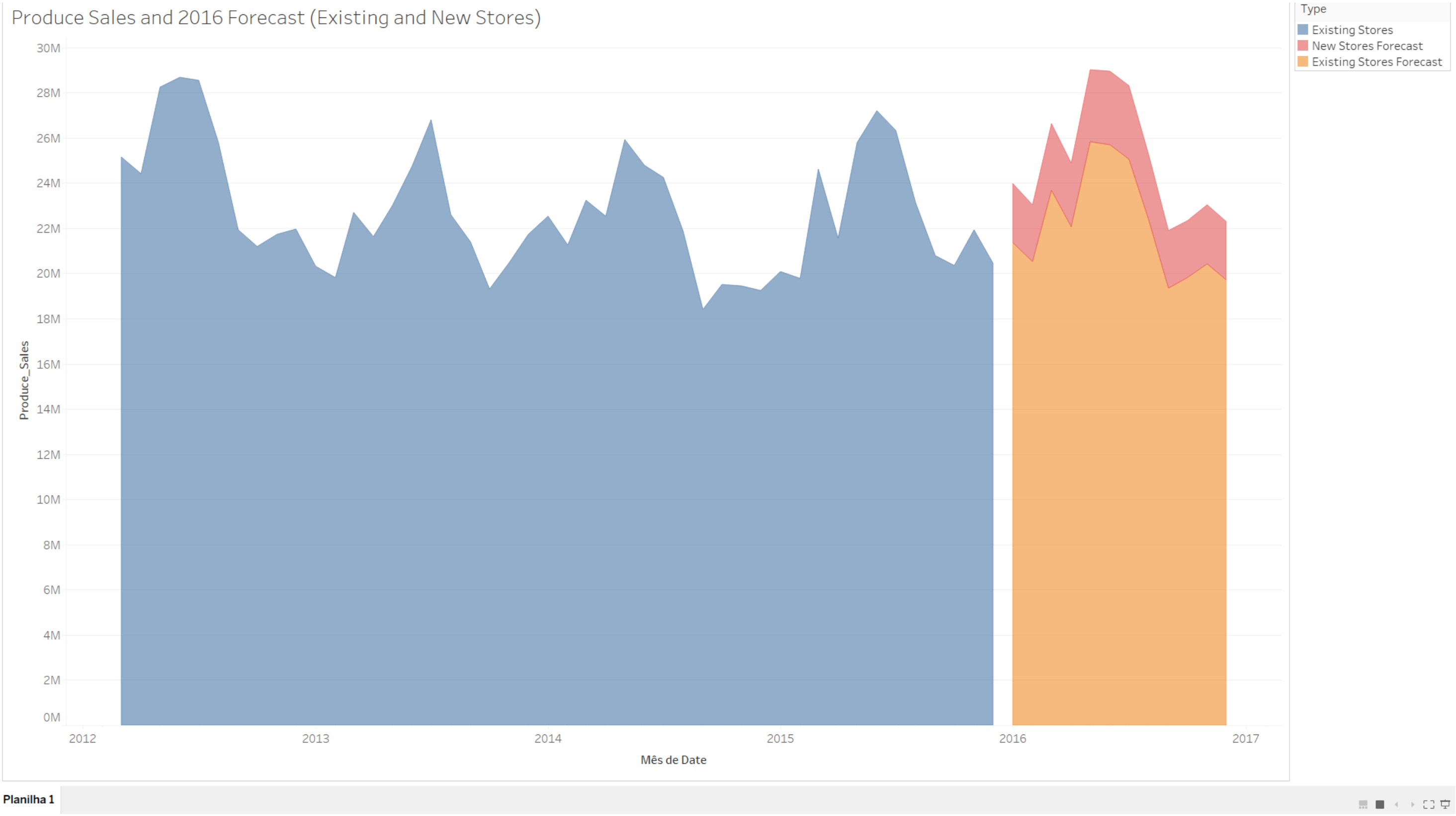
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257



2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Year	Month	Date	Existing_stores	New_stores	Total
2016	1	01/01/2016	$21.370.817,54	$2.599.377,32	$23.970.194,86
2016	2	01/02/2016	$20.525.730,69	$2.521.694,70	$23.047.425,39
2016	3	01/03/2016	$23.684.288,43	$2.949.837,56	$26.634.125,99
2016	4	01/04/2016	$22.073.943,97	$2.807.606,12	$24.881.550,09
2016	5	01/05/2016	$25.826.609,82	$3.184.515,86	$29.011.125,68
2016	6	01/06/2016	$25.708.731,31	$3.239.505,42	$28.948.236,73
2016	7	01/07/2016	$25.059.364,56	$3.252.236,86	$28.311.601,42
2016	8	01/08/2016	$22.355.892,96	$2.890.501,56	$25.246.394,51
2016	9	01/09/2016	$19.333.713,88	$2.556.762,81	$21.890.476,69
2016	10	01/10/2016	$19.829.130,62	$2.499.240,55	$22.328.371,18
2016	11	01/11/2016	$20.428.495,86	$2.597.965,92	$23.026.461,78
2016	12	01/12/2016	$19.720.850,69	$2.569.681,17	$22.290.531,87


Produce Sales and 2016 Forecast (Existing and New Stores)
Type
Existing Stores
New Stores Forecast
Existing Stores Forecast
30M
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
Produce_Sales
Mês de Date
Planilha 1