

Jonathan Glacken – Predictive Analytics Sample (2020/11)

Please find attached a sample of predictive analytics work I've prepared. This was prompted by the Udacity Predictive Analytics for Business Nanodegree which I've earned.

I investigated a simple scenario that touches upon several concepts associated with predictive analytics. The goal here is to recommend opening several stores based on existing store information. The data I've collected and use here is store location information, store customer demographic information, and store sales information.

Jonathan Glacken – Predictive Analytics Sample (2020/11)

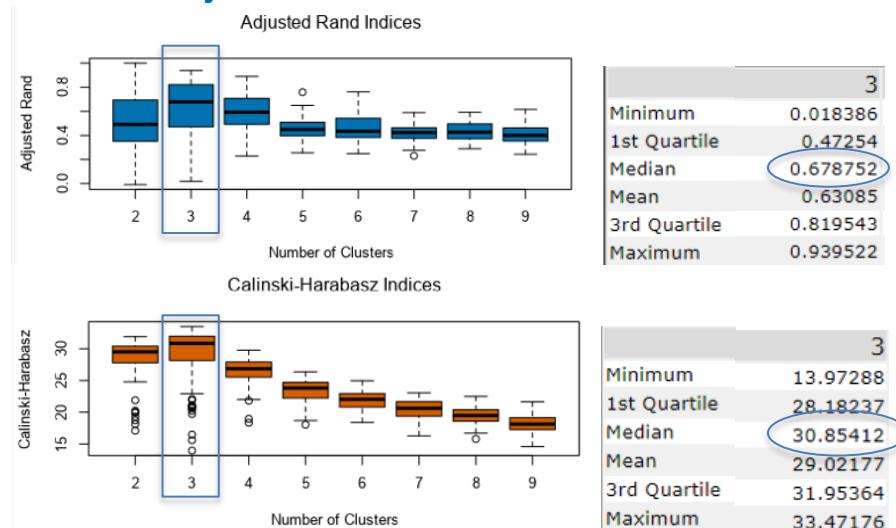
Task 1: Determine Store Formats for Existing Stores

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

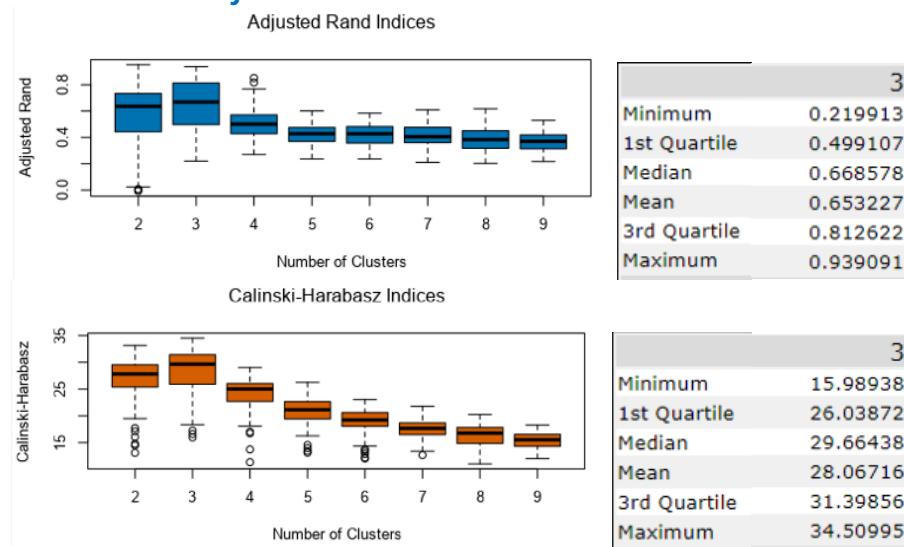
The optimal number of store formats is three based on the K-Means clustering method.

My analysis arrived at this result by relying upon the Adjusted Rand and Calinski-Harabasz Indices across K-Means, K-Median, and Neural Gas clustering methods. Specifically, I was looking for the highest median value for the indices across the number of clusters where spread was minimal and there were no outliers. The Adjusted Rand index is an indicator of cluster stability, how similar the objects within a cluster are. The Calinski-Harabasz index measures both the compactness and distinctiveness of the clusters. These formed the basis of my selection.

K-Means Analysis indicated:

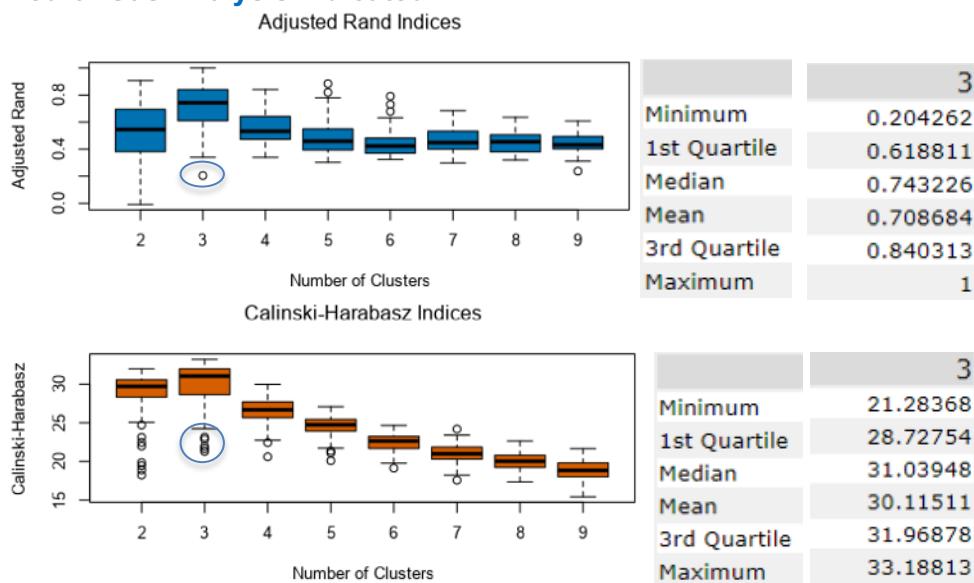


K-Medians Analysis indicated:



Jonathan Glacken – Predictive Analytics Sample (2020/11)

Neural Gas Analysis indicated:



- How many stores fall into each store format?

The number of stores that fall into each store format are as follows:

Cluster 1 = 25

Cluster 2 = 35

Cluster 3 = 25

This relies upon the percentage sales values of each food category, the fields were standardized with z-score, the k-means clustering method was used, and the seed starting number is 10.

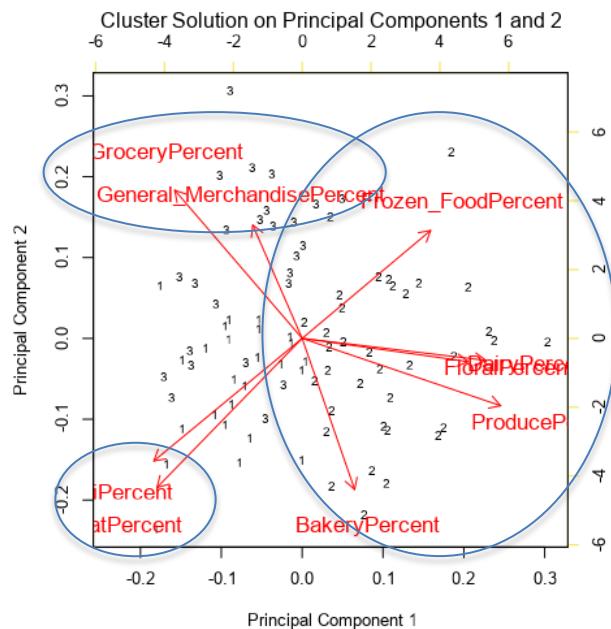
- Based on the results of the clustering model, what is one way that the clusters differ from one another?

Based on the results of the clustering model, one way that the clusters differ from one another is the weighting of the categories profit percentages. No cluster can be unique to any other than for the percentage breakdowns (see below).

	Dry_GroceryPercent	DairyPercent	Frozen_FoodPercent	MeatPercent	ProducePercent	DeliPercent	FloralPercent
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	0.824834	-0.663872
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	-0.46168	0.71741
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.178482	-0.340502
	BakeryPercent	General_MerchandisePercent					
1	0.428226	-0.674769					
2	0.312878	-0.329045					
3	-0.866255	1.135432					

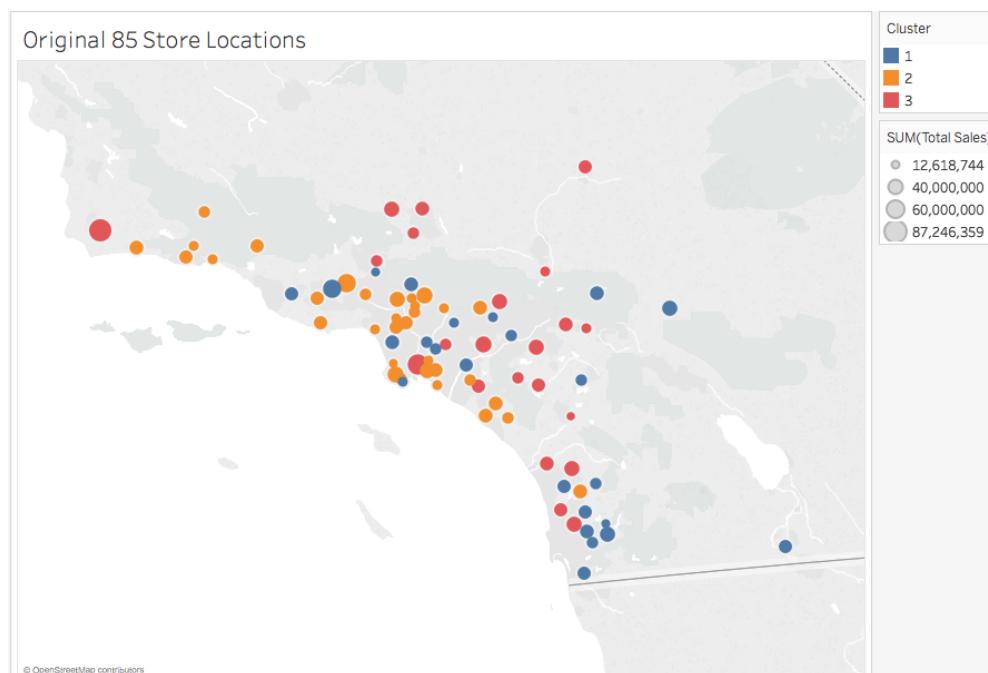
Another perspective on how the clusters differ from one another is shown below. Three categories of stores took form. Meat and Deli clustered as would be suspected as butchers and delis are comparable. Dry Grocery and General Merchandise clustered as these have similar offerings. Frozen Food, Floral, Produce, Bakery, and Dairy clustered. This makes sense as such things are often sold in dedicated stores.

Jonathan Glacken – Predictive Analytics Sample (2020/11)



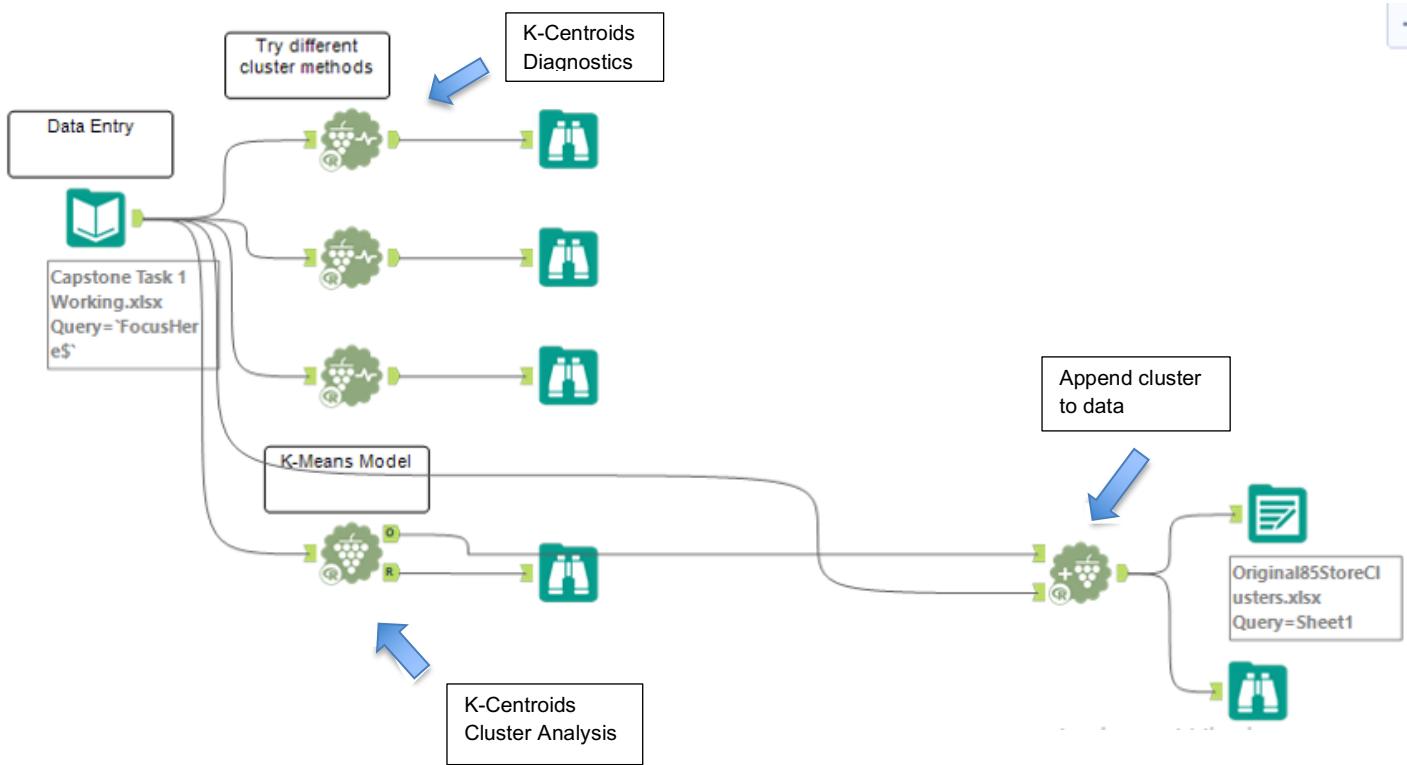
4. Please provide a Tableau visualization that shows the location of the stores, uses color to show cluster, and size to show total sales.

Here is a visualization, created in Tableau, that shows the location of the stores, uses color to show cluster, and size to show total sales:



Jonathan Glacken – Predictive Analytics Sample (2020/11)

Snapshot Alteryx Visual Coding for Task 1



Jonathan Glacken – Predictive Analytics Sample (2020/11)

Task 2: Formats for New Stores

- What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?

I used the Boosted Model, which is an ensemble of decision trees, to predict the best store format for new stores. I developed a Boosted Model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store. I used a 20% validation sample with Random Seed = 3 when creating samples with which to compare the accuracy of the models.

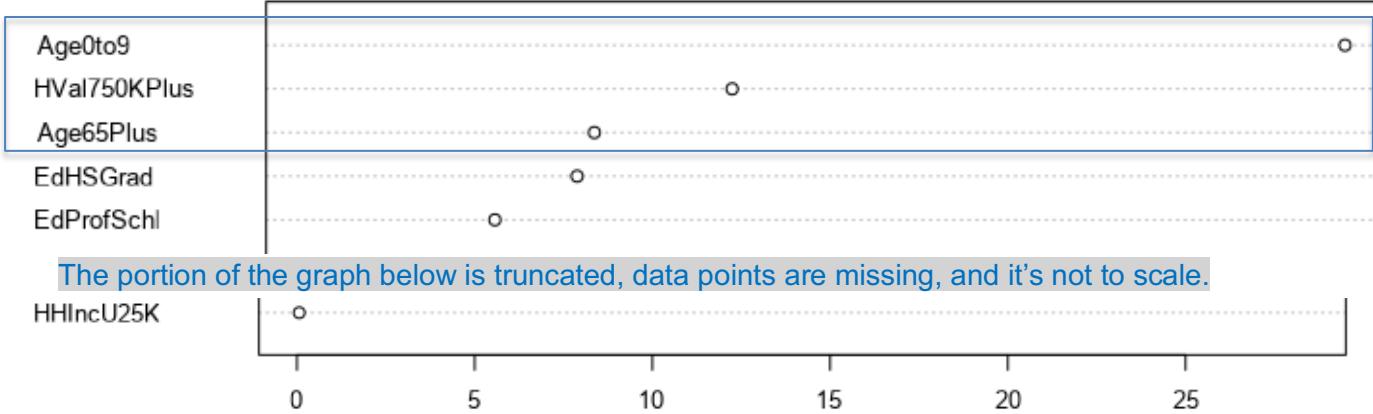
The Boosted Model performed similar to the Random Forest model when I trained and validated the model. The accuracy, how often the classifier is correct, was 0.7647. The F1, a weighted average of recall and precision, was 0.8333. Boosting helps reduce variance and bias.

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_36	0.6471	0.6667	0.5000	1.0000	0.5000
Boosted	0.7647	0.8333	0.5000	1.0000	1.0000
randomforest	0.7059	0.7500	0.5000	1.0000	0.7500

The three most important variables that help explain the relationship between demographic indicators and store formats are: Age0to9, HVal750kPlus, and Age65Plus (see below).

Variable Importance Plot

The user options for graphics width and height has been overridden for readability of y axis labels

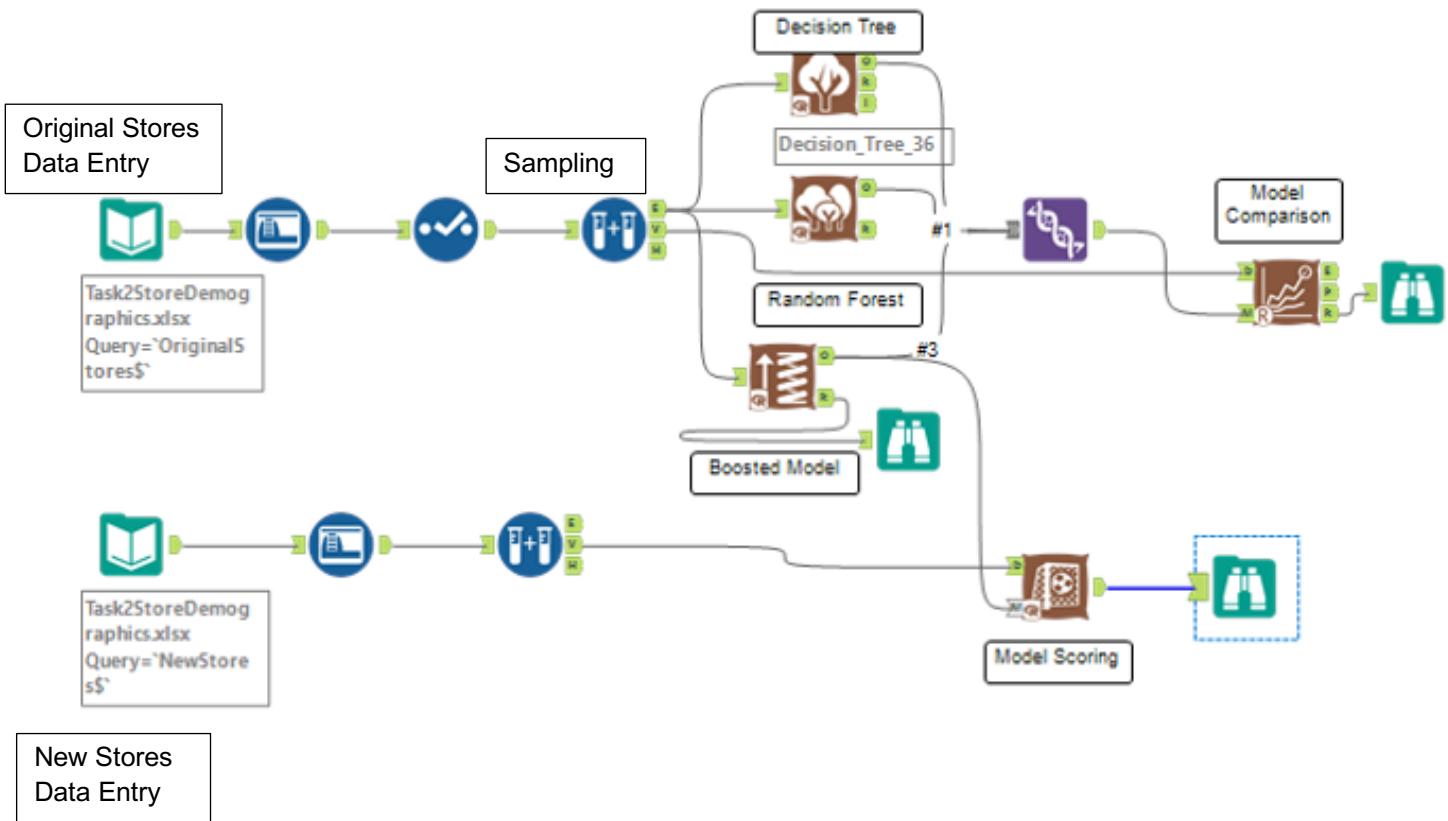


- What format do each of the 10 new stores fall into? Please fill in the table below.

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

Jonathan Glacken – Predictive Analytics Sample (2020/11)

Snapshot Alteryx Visual Coding for Task 2



Jonathan Glacken – Predictive Analytics Sample (2020/11)

Task 3: Predicting Produce Sales

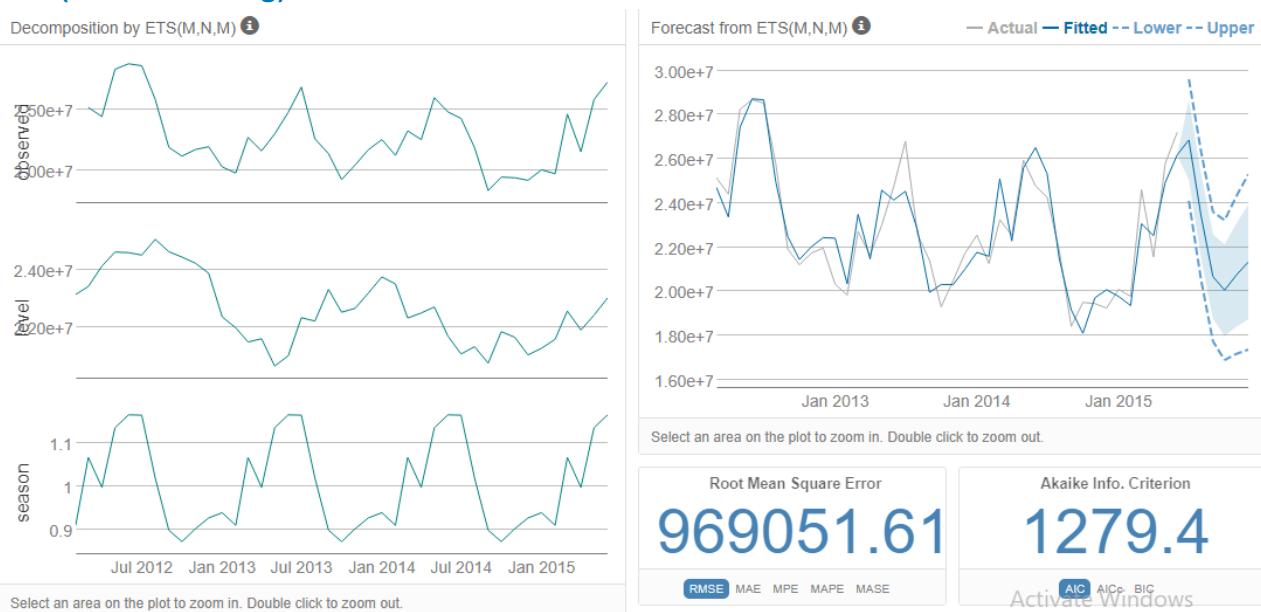
1. What type of ETS (error, trend, seasonal) model or ARIMA (Auto Regressive Integrated Moving Average) model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

I used the ETS (M,N,M) exponential smoothing model (the “ETS model”) for forecasting. This was based on an iterative approach that analyzed the time series data as well as how well the forecasting was for a known set of data used for validation. The ETS model had a **Root Mean Square Error (RMSE) of 663707.2** with the validation set. Alternatively, I investigated the ARIMA(0,1,1)(1,1,0)[12] model (the “ARIMA model”) which had a **RMSE of 1710251** with the validation set. I chose the ETS model as it has a lower RMSE value.

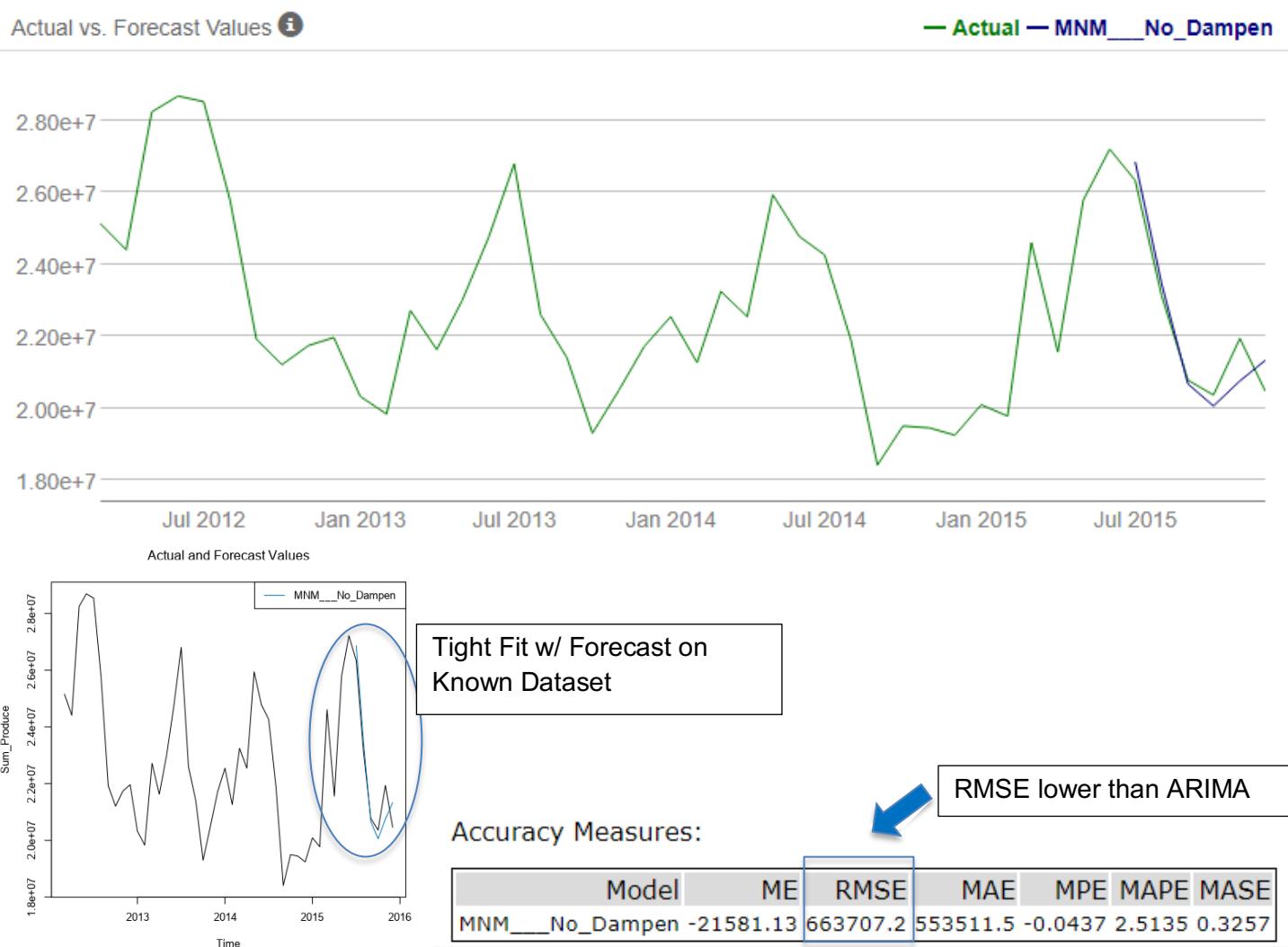
Jonathan Glacken – Predictive Analytics Sample (2020/11)

ETS versus ARIMA (High Level Comparison When Testing Which Model is Better)

ETS (Model Training)

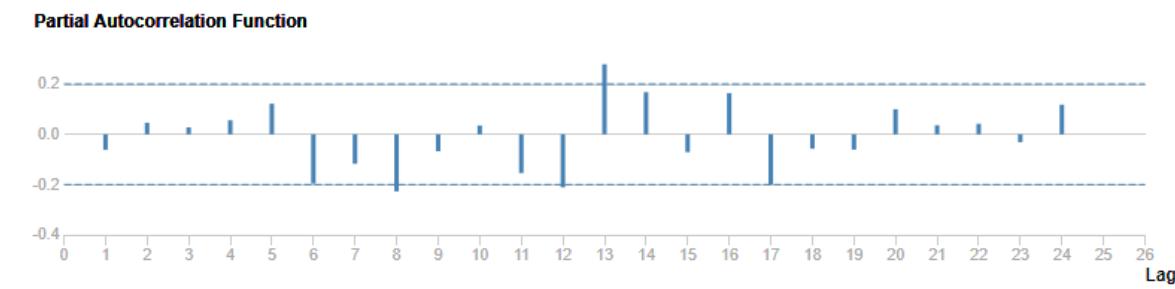
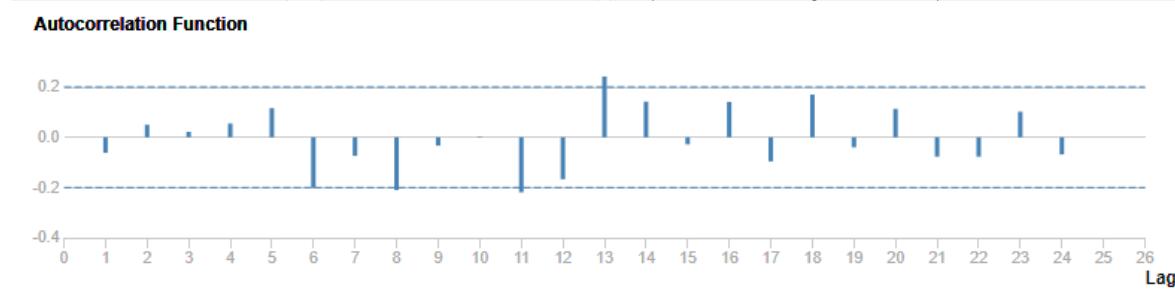
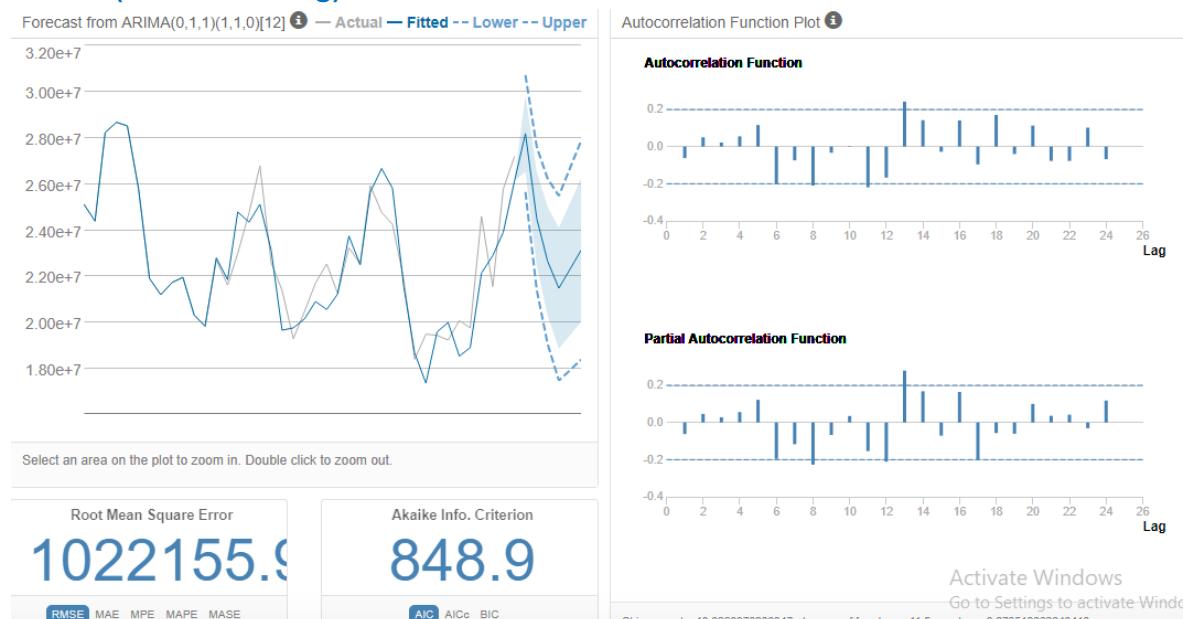


ETS (Model Testing with Validation Set)



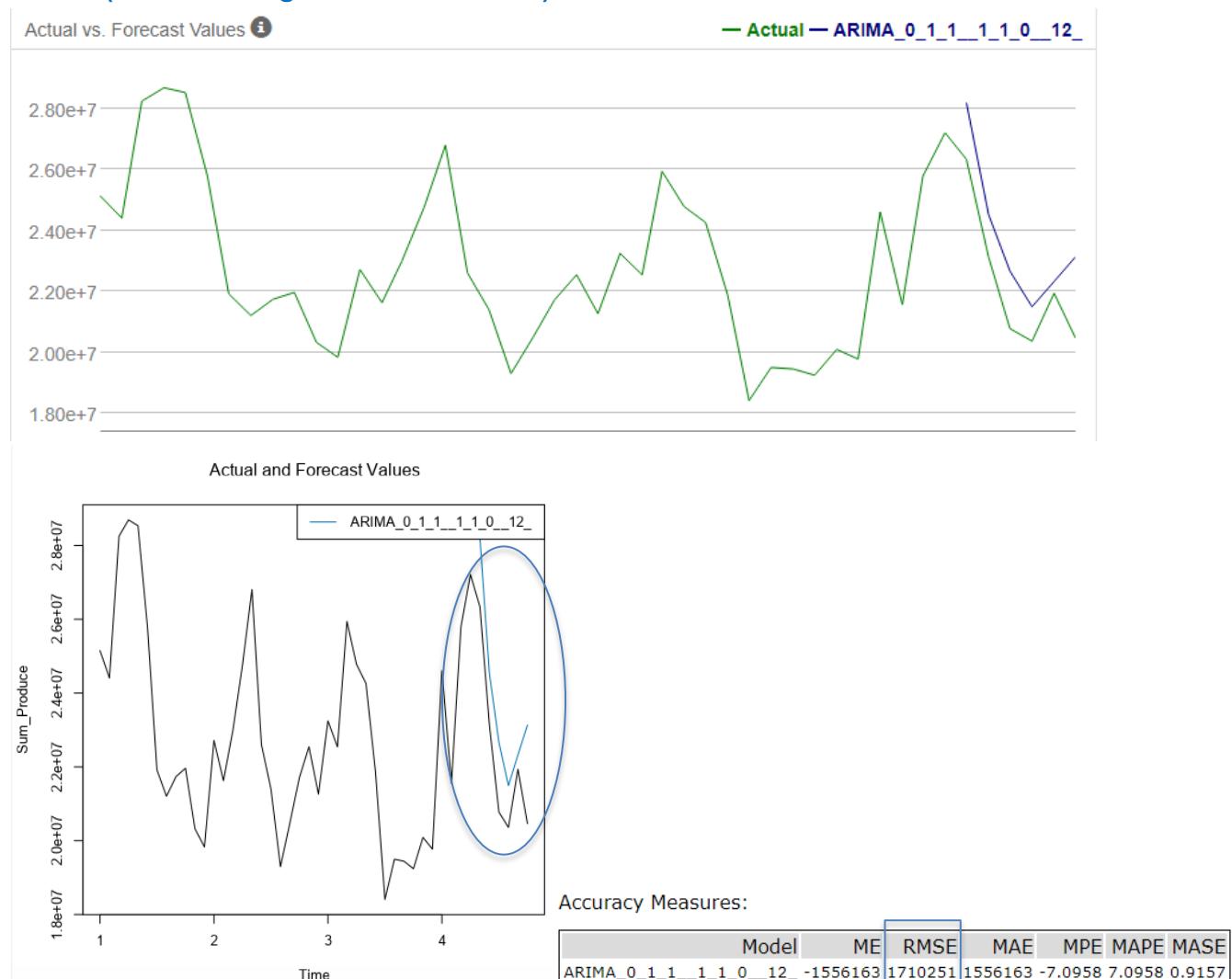
Jonathan Glacken – Predictive Analytics Sample (2020/11)

ARIMA (Model Training)



Jonathan Glacken – Predictive Analytics Sample (2020/11)

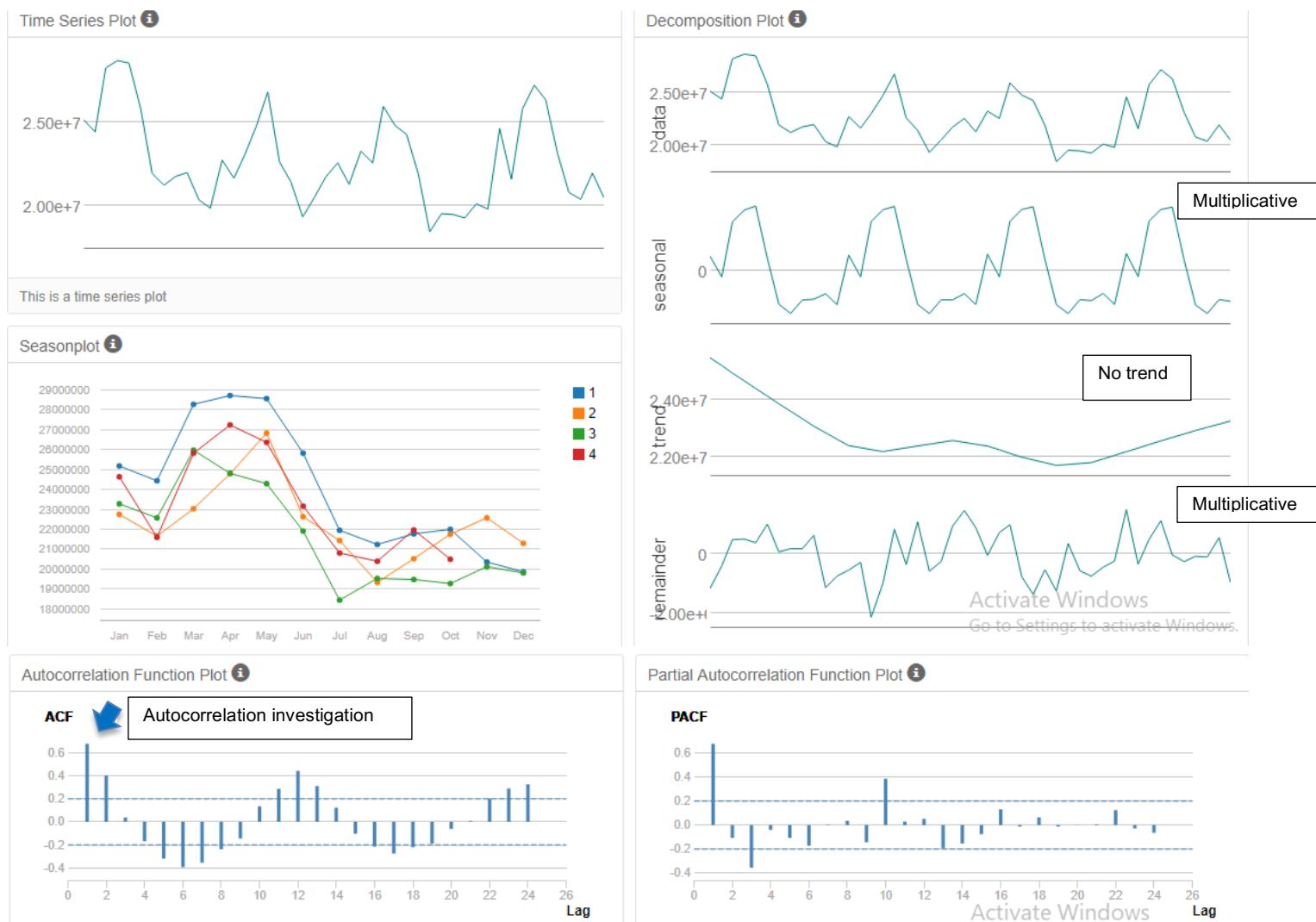
ARIMA (Model Testing with Validation Set)



Jonathan Glacken – Predictive Analytics Sample (2020/11)

The information below, “Original Information” and “First Difference – Original Information,” served as the basis for the specifications in the ETS and ARIMA models I tested. The Decomposition Plot associated with the “Original Information” informed seasonal (M), trend (N), and remainder (M) specifications. Additionally, the ACF and PACF plots informed me about correlation and bias potential. The first difference of the original information helped make the dataset stationary.

Original Information



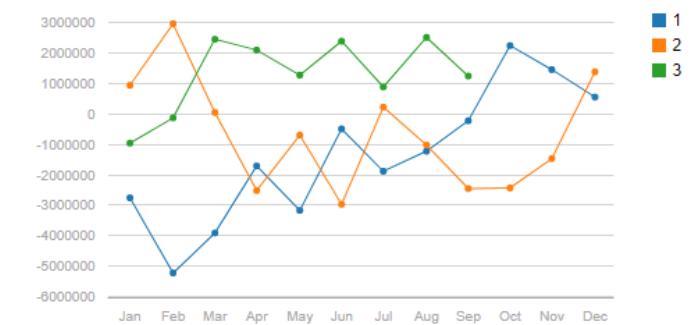
Jonathan Glacken – Predictive Analytics Sample (2020/11)

First Difference - Original Information

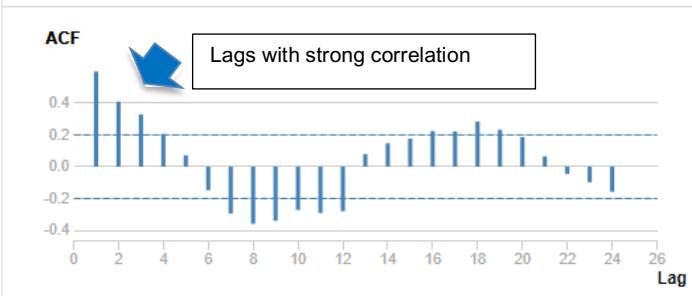
Time Series Plot



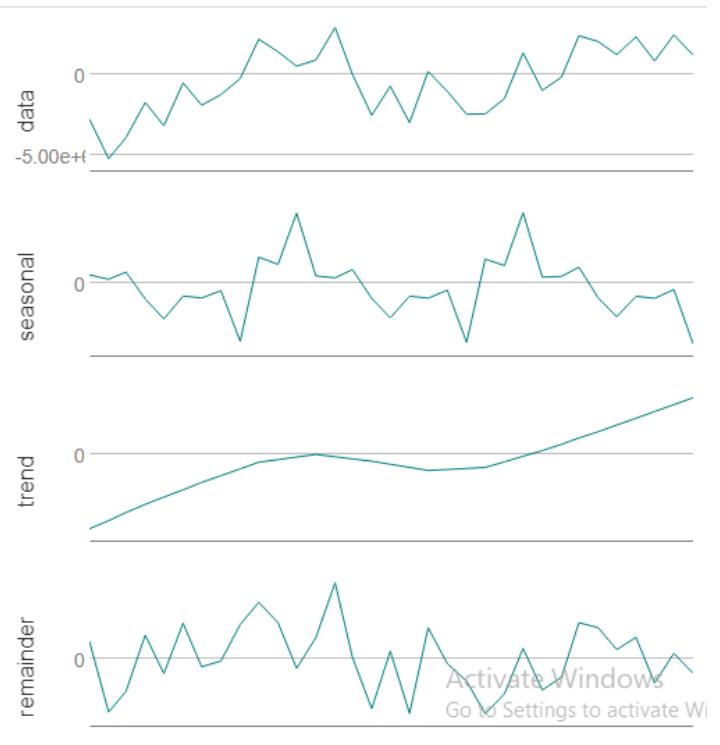
Seasonplot



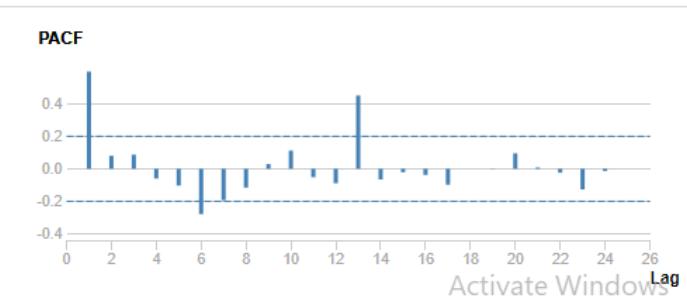
Autocorrelation Function Plot



Decomposition Plot



Partial Autocorrelation Function Plot



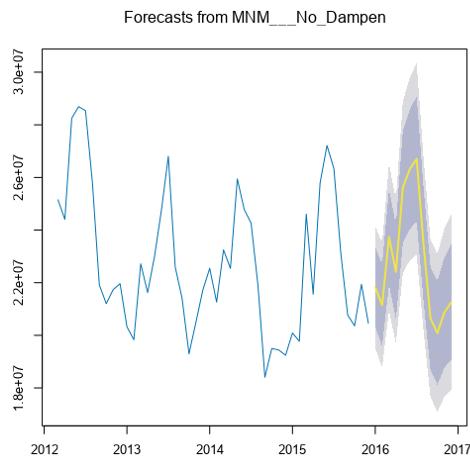
Jonathan Glacken – Predictive Analytics Sample (2020/11)

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.

This chart below focuses on produce sales.

Month	New Stores	(All) Existing Stores [ETS]
January 2016	\$ 2,386,310.53	\$ 21,829,060.03
February 2016	\$ 2,316,117.18	\$ 21,146,329.63
March 2016	\$ 2,706,999.46	\$ 23,735,686.94
April 2016	\$ 2,567,294.00	\$ 22,409,515.28
May 2016	\$ 2,912,347.27	\$ 25,621,828.73
June 2016	\$ 2,966,212.12	\$ 26,307,858.04
July 2016	\$ 2,990,737.44	\$ 26,705,092.56
August 2016	\$ 2,643,838.56	\$ 23,440,761.33
September 2016	\$ 2,314,571.37	\$ 20,640,047.32
October 2016	\$ 2,275,523.88	\$ 20,086,270.46
November 2016	\$ 2,356,886.66	\$ 20,858,119.96
December 2016	\$ 2,394,478.99	\$ 21,255,190.24

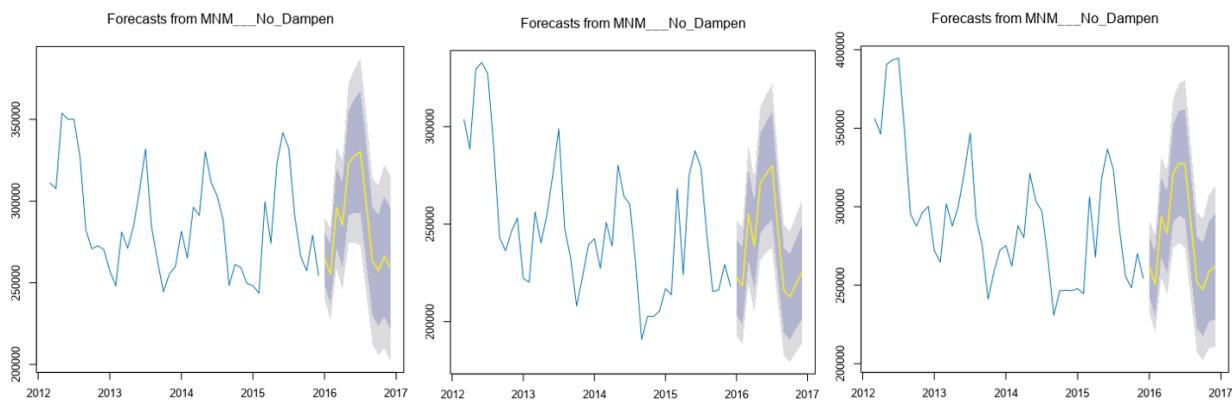
Existing Stores



Period	Sub_Period	ForecastETSExisting	ForecastETSExisting_high_95	ForecastETSExisting_high_80	ForecastETSExisting_low_80	ForecastETSExisting_low_95
2016	1	21829060.031666	24149899.115321	23346575.14138	20311544.921952	19508220.948011
2016	2	21146329.631982	23512577.365832	22693535.862148	19599123.401815	18780081.898131
2016	3	23735686.93879	26517865.796798	25554855.912929	21916517.964651	20953508.080782
2016	4	22409515.284474	25150243.401256	24201581.075733	20617449.493214	19668787.167691
2016	5	25621828.725097	28880596.484529	27752622.431914	23491035.018279	22363060.965665
2016	6	26307858.040046	29777680.067343	28576652.715009	24039063.365084	22838036.01275
2016	7	26705092.556349	30348682.320364	29087507.847195	24322677.265503	23061502.792334
2016	8	23440761.329527	26742106.733295	25599395.061562	21282127.597491	20139415.925758
2016	9	20640047.319971	23635033.372194	22598363.439189	18681731.200753	17645061.267747
2016	10	20086270.462075	23084199.797487	22046511.090727	18126029.833423	17088341.126662
2016	11	20858119.95754	24055437.105831	22948733.269445	18767506.645635	17660802.809249
2016	12	21255190.244976	24596988.126893	23440274.43075	19070106.059202	17913392.363058

Jonathan Glacken – Predictive Analytics Sample (2020/11)

New Stores (Format One, Two, Three)



Format One

Period	Sub_Period	ForecastETSNew	ForecastETSNew_high_95	ForecastETSNew_high_80	ForecastETSNew_low_80	ForecastETSNew_low_95
2016	1	265048.250562	289637.489162	281126.289481	248970.211644	240459.011963
2016	2	255209.550891	283340.66053	273603.494801	236815.606981	227078.441253
2016	3	296060.071531	333152.382665	320313.430153	271806.712908	258967.760396
2016	4	285343.982717	324930.300889	311228.084888	259459.880546	245757.664546
2016	5	323190.902935	371991.30219	355099.768893	291282.036976	274390.503679
2016	6	327777.162664	380990.306444	362571.367178	292982.958151	274564.018885
2016	7	330149.437749	387252.619534	367487.201397	292811.6741	273046.255963
2016	8	297726.544925	352202.092953	333346.191209	262106.898641	243250.996897
2016	9	263375.751672	314067.806602	296521.504792	230229.998551	212683.696742
2016	10	257575.268156	309484.439333	291516.850814	223633.685498	205666.096979
2016	11	266162.340209	322111.646047	302745.624929	229579.055488	210213.03437
2016	12	259111.202342	315739.447499	296138.421621	222083.983063	20482.957185

Format Two

Period	Sub_Period	ForecastETSNew	ForecastETSNew_high_95	ForecastETSNew_high_80	ForecastETSNew_low_80	ForecastETSNew_low_95
2016	1	222560.647536	252301.985309	242007.463031	203113.832042	192819.309764
2016	2	218283.939939	248067.9784	237758.675913	198809.203965	188499.901478
2016	3	255102.064391	290613.237512	278321.572436	231882.556345	219590.89127
2016	4	238756.908112	272638.474943	260910.873878	216602.942346	204875.341281
2016	5	271034.477121	310215.920023	296653.845583	245415.10866	231853.03422
2016	6	275836.017356	316430.689131	302379.446509	249292.588203	235241.345581
2016	7	279905.989933	321816.790797	307309.989641	252501.990224	237995.189068
2016	8	246518.025325	284050.692146	271059.317054	221976.733596	208985.358505
2016	9	216006.36965	249429.072616	237860.300461	194152.438839	182583.666685
2016	10	212701.776731	246132.225918	234560.772524	190842.780937	179271.327543
2016	11	219103.765442	254067.034953	241965.018573	196242.512311	184140.495931
2016	12	225020.843159	261461.240743	248847.938423	201193.747895	188580.445574

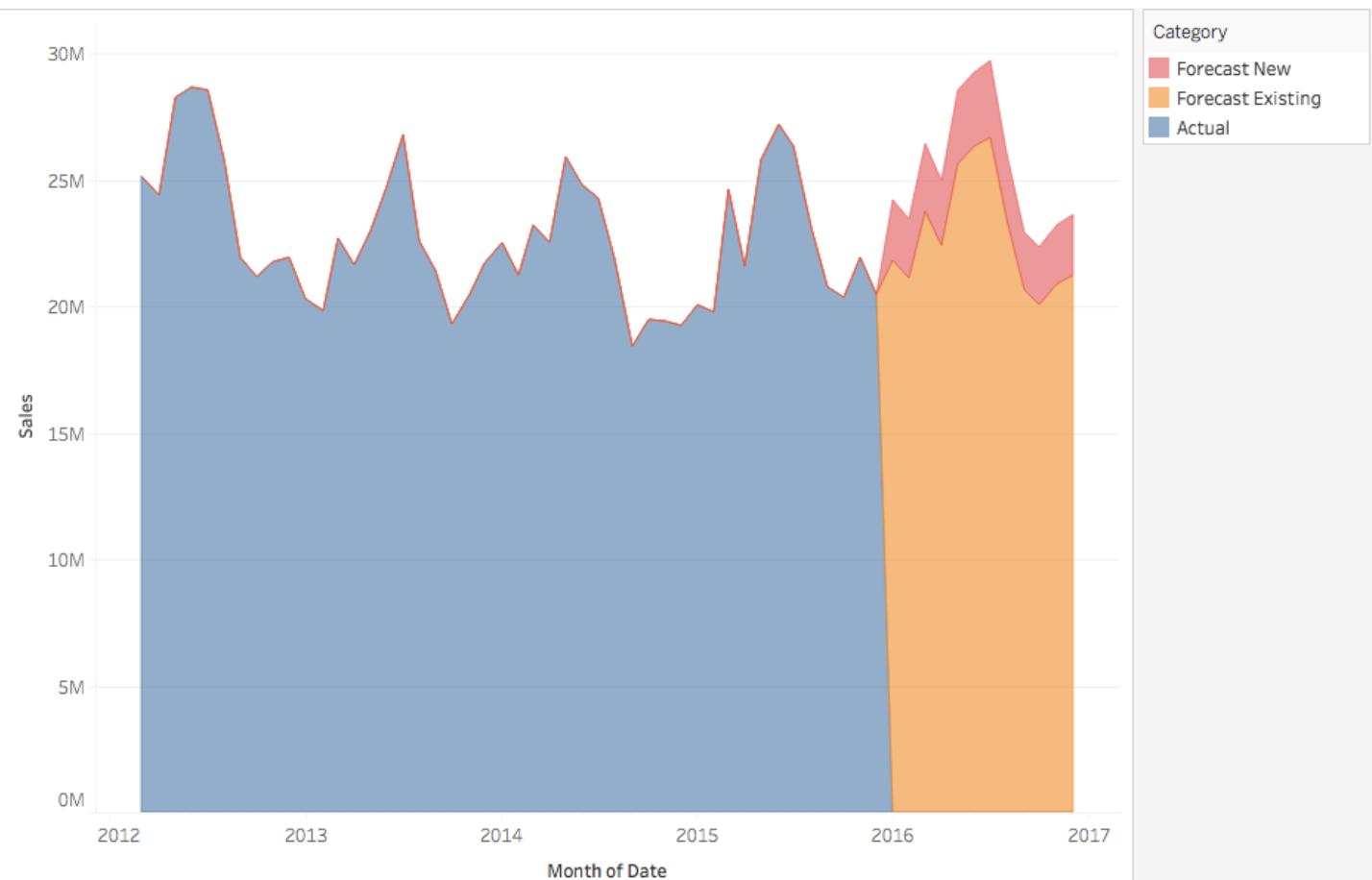
Format Three

Period	Sub_Period	ForecastETSNew	ForecastETSNew_high_95	ForecastETSNew_high_80	ForecastETSNew_low_80	ForecastETSNew_low_95
2016	1	261966.130313	291334.973806	281169.38489	242762.875736	232597.28682
2016	2	250401.331481	280934.994767	270366.220782	230436.44218	219867.668196
2016	3	293442.333068	331894.551183	318584.88707	268299.779065	254990.114953
2016	4	283136.522406	322648.435277	308971.973566	257301.071247	243624.609536
2016	5	320983.169593	368353.075964	351956.686529	290009.652657	273613.263222
2016	6	327806.285088	378679.90414	361070.756606	294541.813569	276932.666035
2016	7	327050.688759	380182.785654	361791.899564	292309.477954	273918.591864
2016	8	289001.287074	337962.529048	321015.322433	256987.251715	240040.0451
2016	9	251719.133388	296047.275838	280703.747801	222734.518976	207390.990939
2016	10	247245.984337	292379.978938	276757.517149	217734.451525	202111.989736
2016	11	258700.577123	307537.02953	290633.016969	226768.137277	209864.124717
2016	12	261747.577124	312738.486837	295088.740883	228406.413365	210756.667411

Here is a visualization, created in Tableau, which explores produce sales from March 2012 - December 2016.

Jonathan Glacken – Predictive Analytics Sample (2020/11)

Produce Sales Forecast (2012-2016) (Stacked)



Jonathan Glacken – Predictive Analytics Sample (2020/11)

Snapshot Alteryx Visual Coding for Task 3

