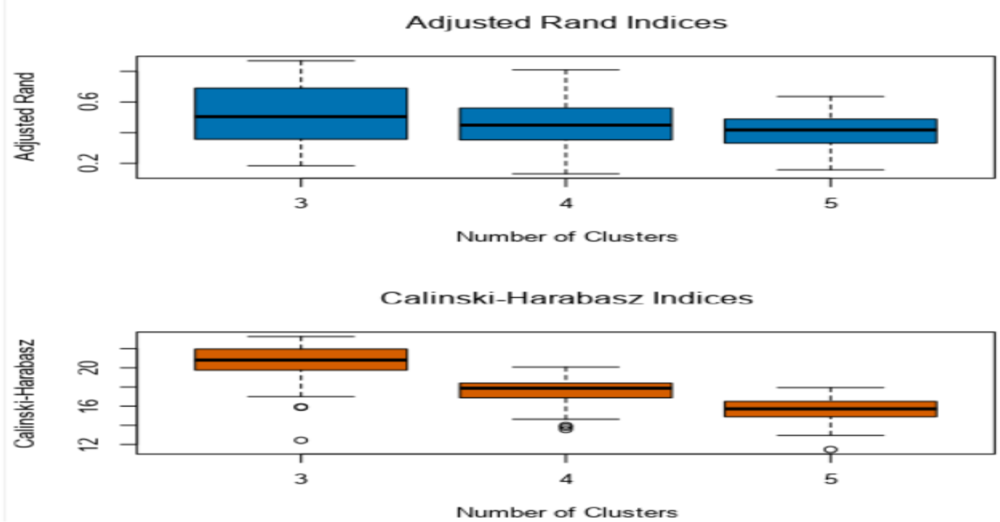
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. **What is the optimal number of store formats? How did you arrive at that number?**
   Based on the K-means report, Adjusted Rand and Calinski-Harabasz indices below, the optimal number of store formats is 2 when both the indices registered the highest median value.

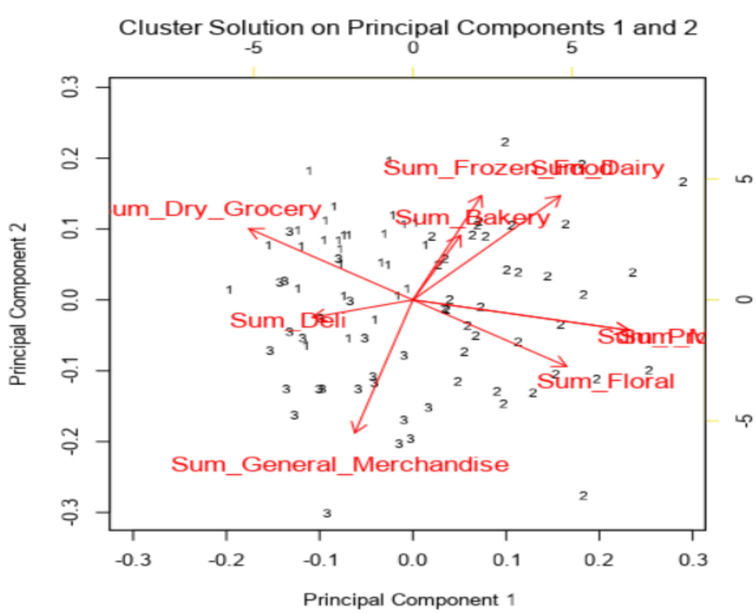


2. **How many stores fall into each store format?**

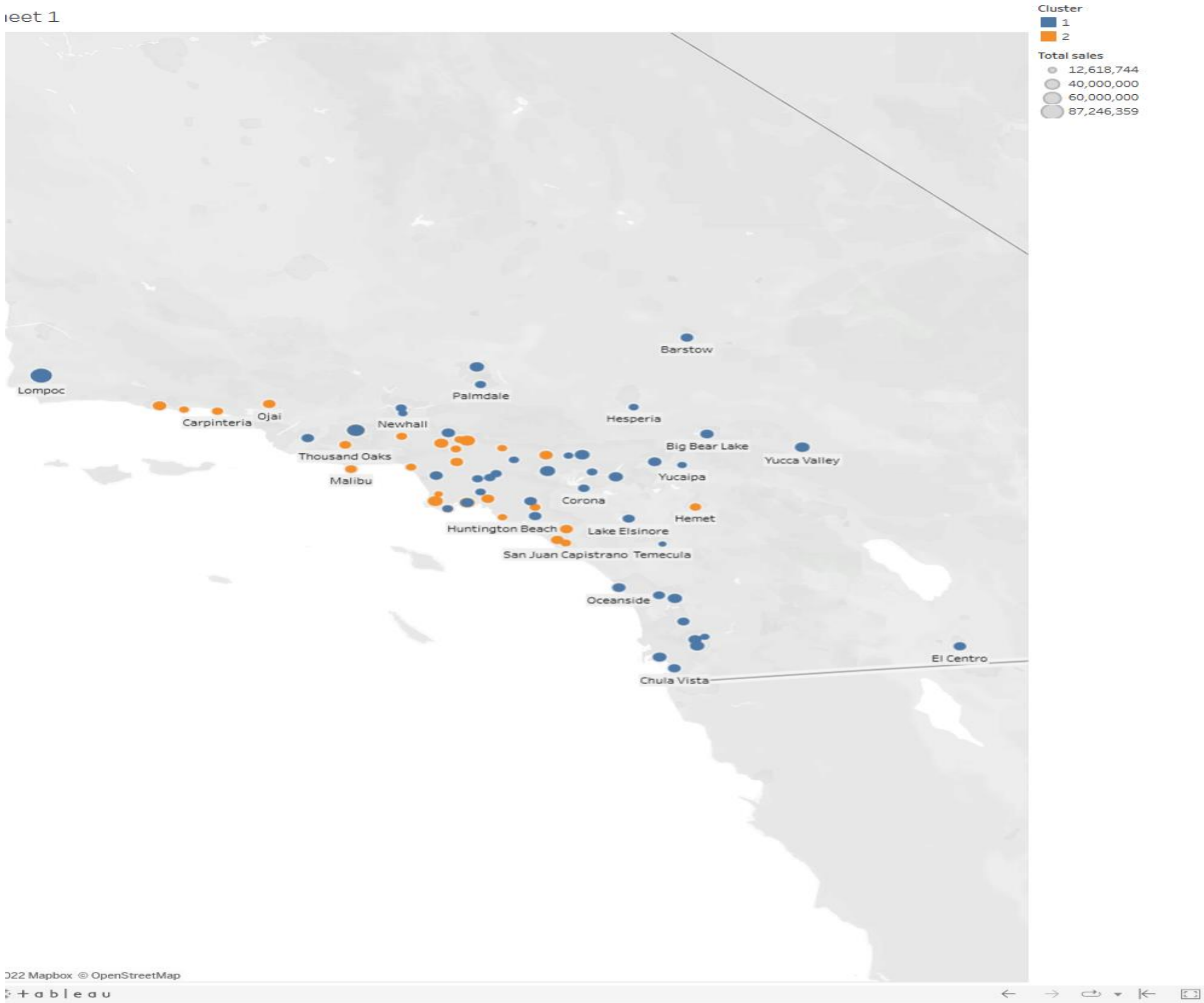
Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	28	1.986754	4.343274	2.108582
2	34	2.479944	4.275177	1.901292
3	23	2.173117	3.487179	1.701299

3. **Based on the results of the clustering model, what is one way that the clusters differ from one another?**
   The differences between the clusters. The third cluster has the smallest Average distance, being the most compact and stable among the three. Meanwhile, the first cluster has the highest Maximum length of 2.1 from the centroid.



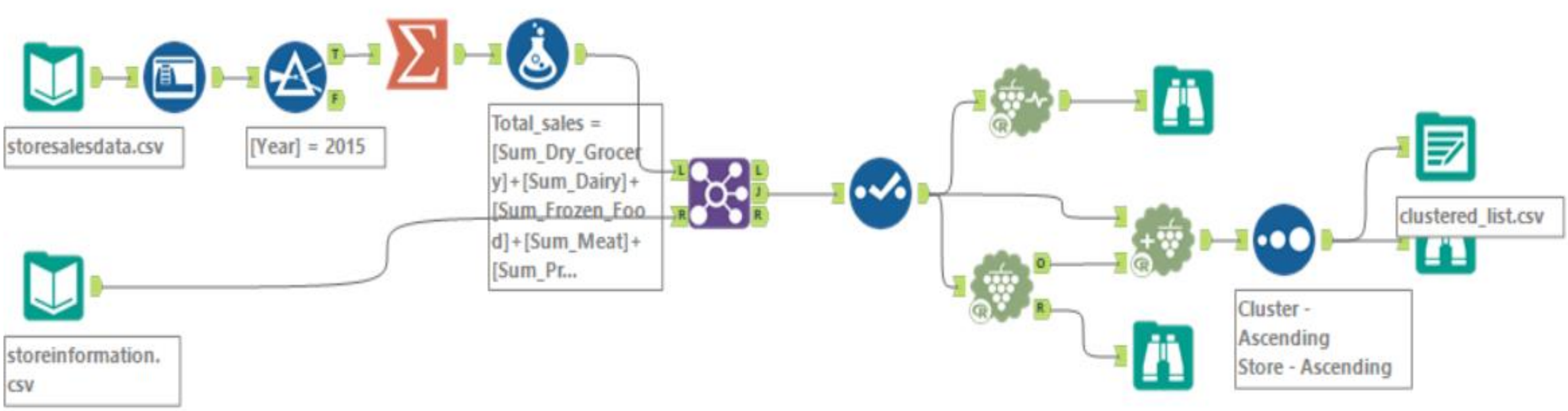
4. **Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses co**



lor to show cluster, and size to show total sales.

Figure 1- https://public.tableau.com/shared/QD99WT358?:display_count=n&:origin=viz_share_link

Alteryx Workflow: Task1



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.)**

The model comparison report below shows comparison matrix of Decision Tree, Forest Model and Boosted Model.
Forest Model is chosen despite having same accuracy as Decision Tree due to higher F1 value.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8611	0.7500	0.8333	1.0000
Decision_Tree	0.5294	0.5833	0.2500	0.8333	0.6667
Boosted_Model	0.7059	0.7639	0.6250	0.6667	1.0000

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

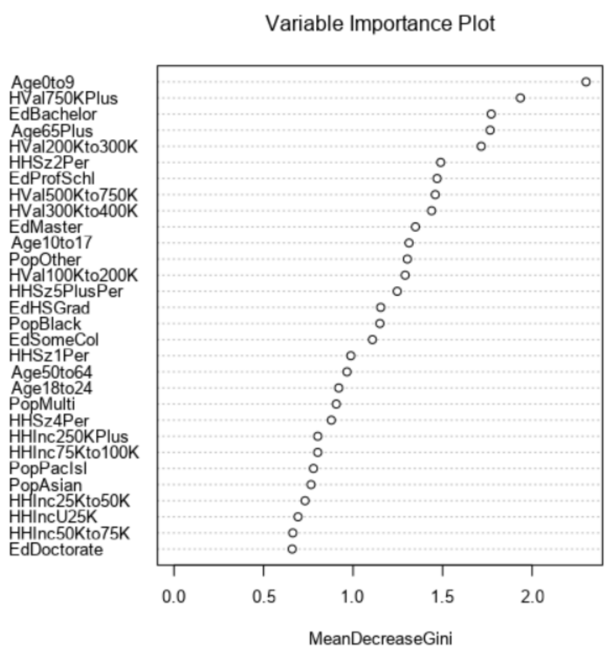
	Actual_1	Actual_2	Actual_3
Predicted_1	2	0	1
Predicted_2	3	5	0
Predicted_3	3	1	2

Confusion matrix of Forest_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	6	0	0
Predicted_2	1	5	0
Predicted_3	1	1	3

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

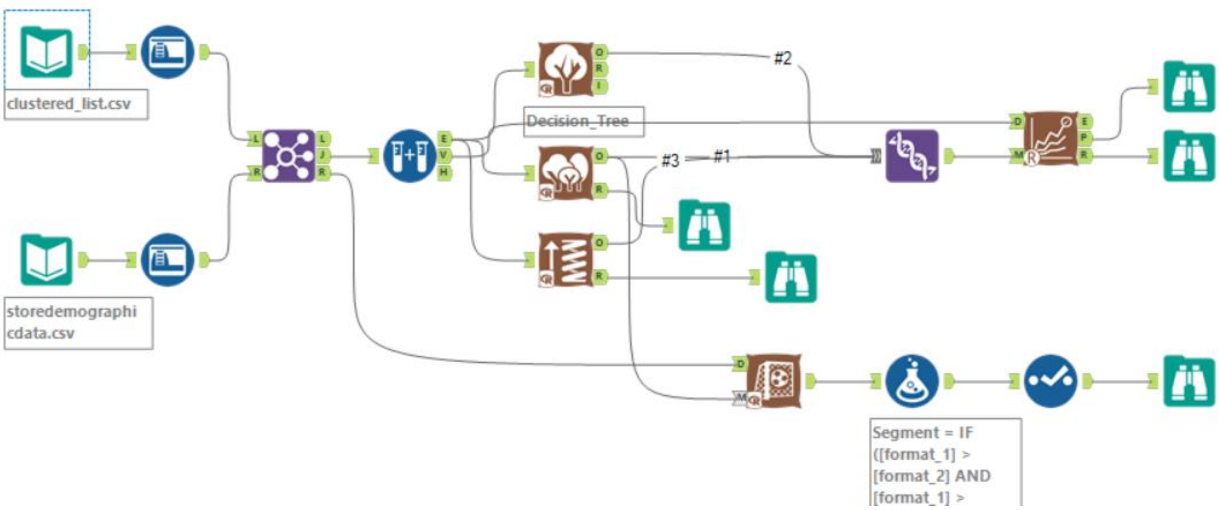
Ave0to9, HVal750KPlus and EdBachelor are the three most important variables.



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

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

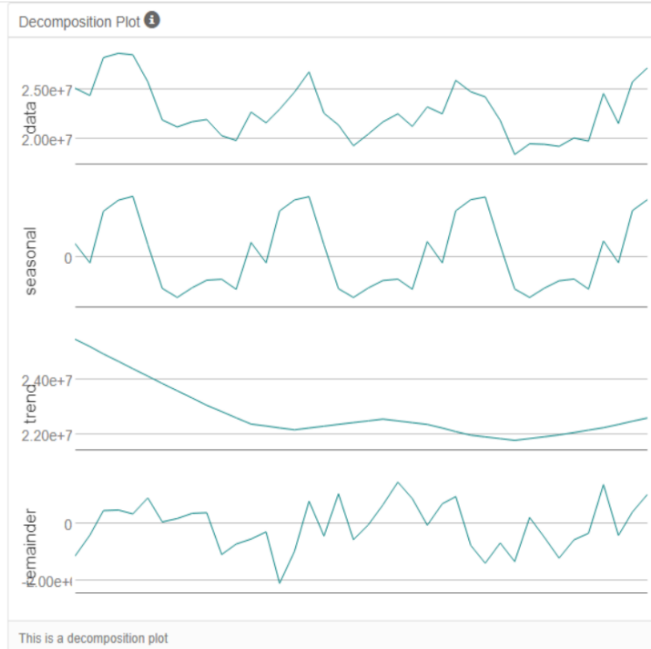
Alteryx Workflow: Task2

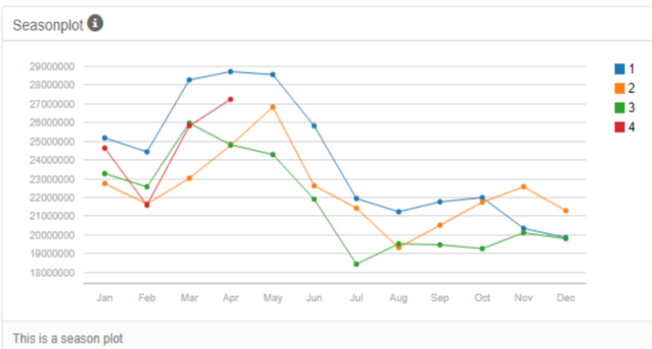


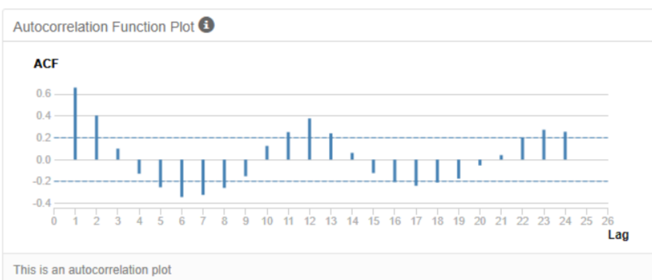
Task 3: Predicting Produce Sales

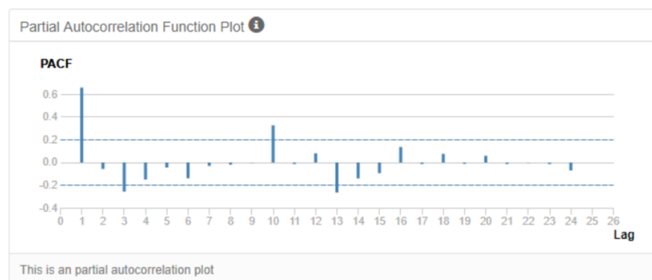
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

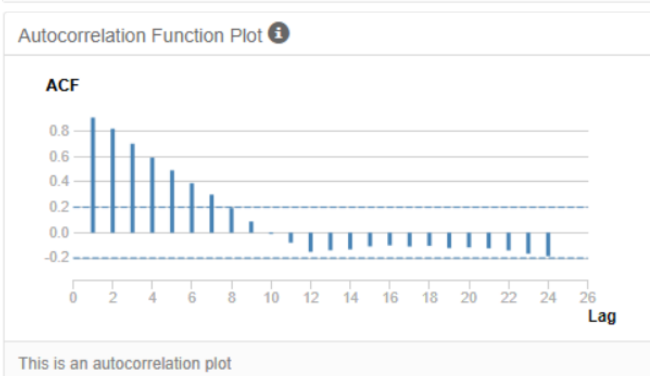
ETS(M,N,M) with no dampening is used for ETS model.

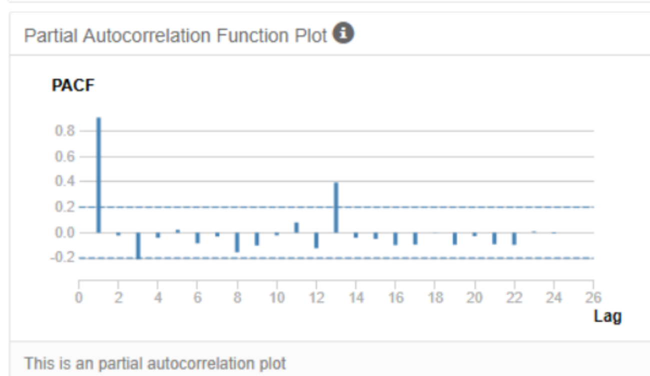


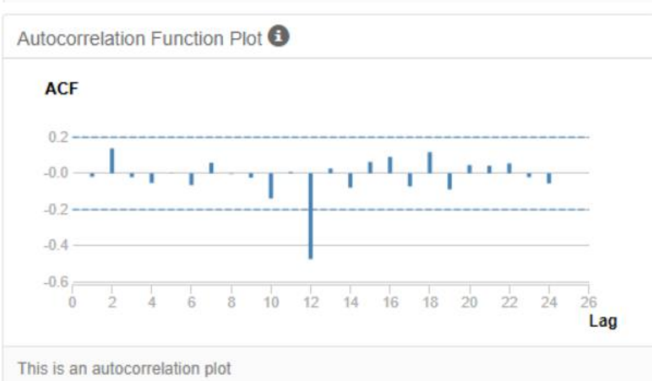


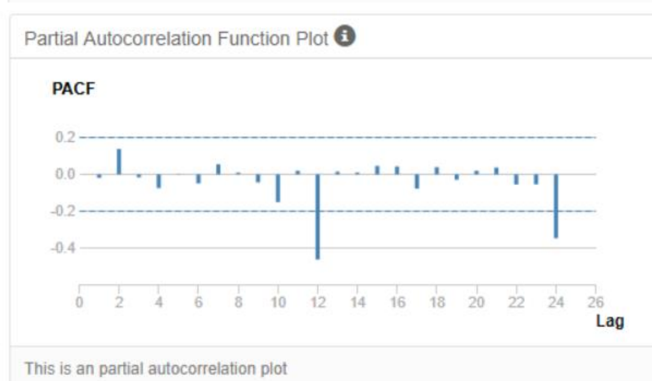












ETS model's accuracy is higher when compared to ARIMA model. A holdout sample of 6 months data is used. Its RMSE of 663707.2 is lower than ARIMA's 1050239.2 while its MASE is 0.33 compared to ARIMA's 0.55. ETS also has a higher AIC at 1,279 while ARIMA's AIC is 880.

Method:

ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

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

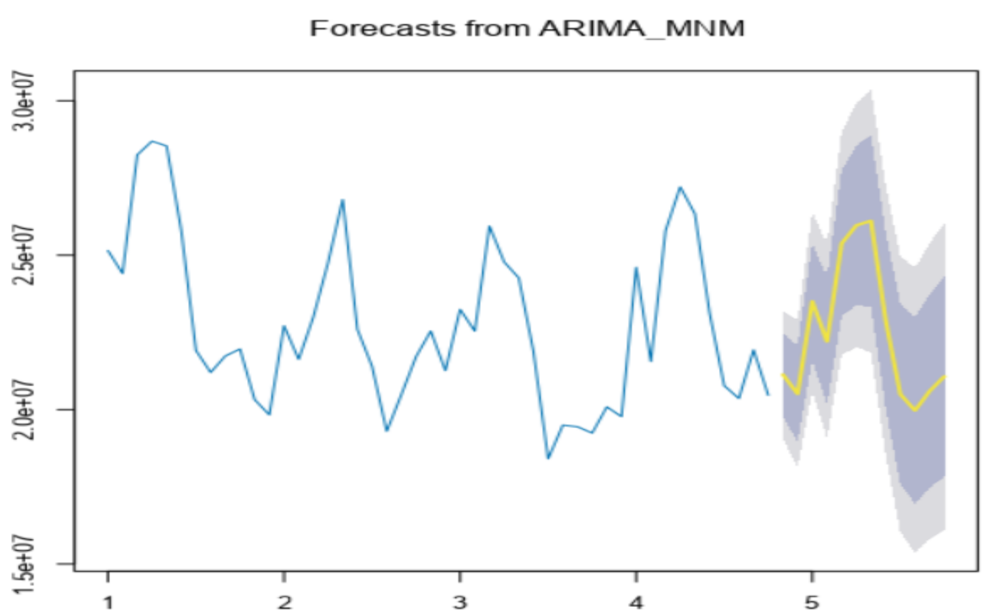
Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

The graph and table below shows actual and forecast value with 80% & 95% confidence level interval



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	Existing Stores Sales	New Stores Sales
2016	1	21,136,641.78	2,519,802.04
2016	2	20,507,039.12	2,437,777.35
2016	3	23,506,565.98	2,865,644.76
2016	4	22,208,405.76	2,707,852.80
2016	5	25,380,147.77	3,067,216.48
2016	6	25,966,799.47	3,101,250.12
2016	7	26,113,792.57	3,106,010.40
2016	8	22,899,285.77	2,765,066.18
2016	9	20,499,583.91	2,454,988.28
2016	10	19,971,242.82	2,401,303.34
2016	11	20,602,665.92	2,504,805.05
2016	12	21,073,222.08	2,492,008.11

The chart Below shows the historical and forecast sales for existing stores and new stores.

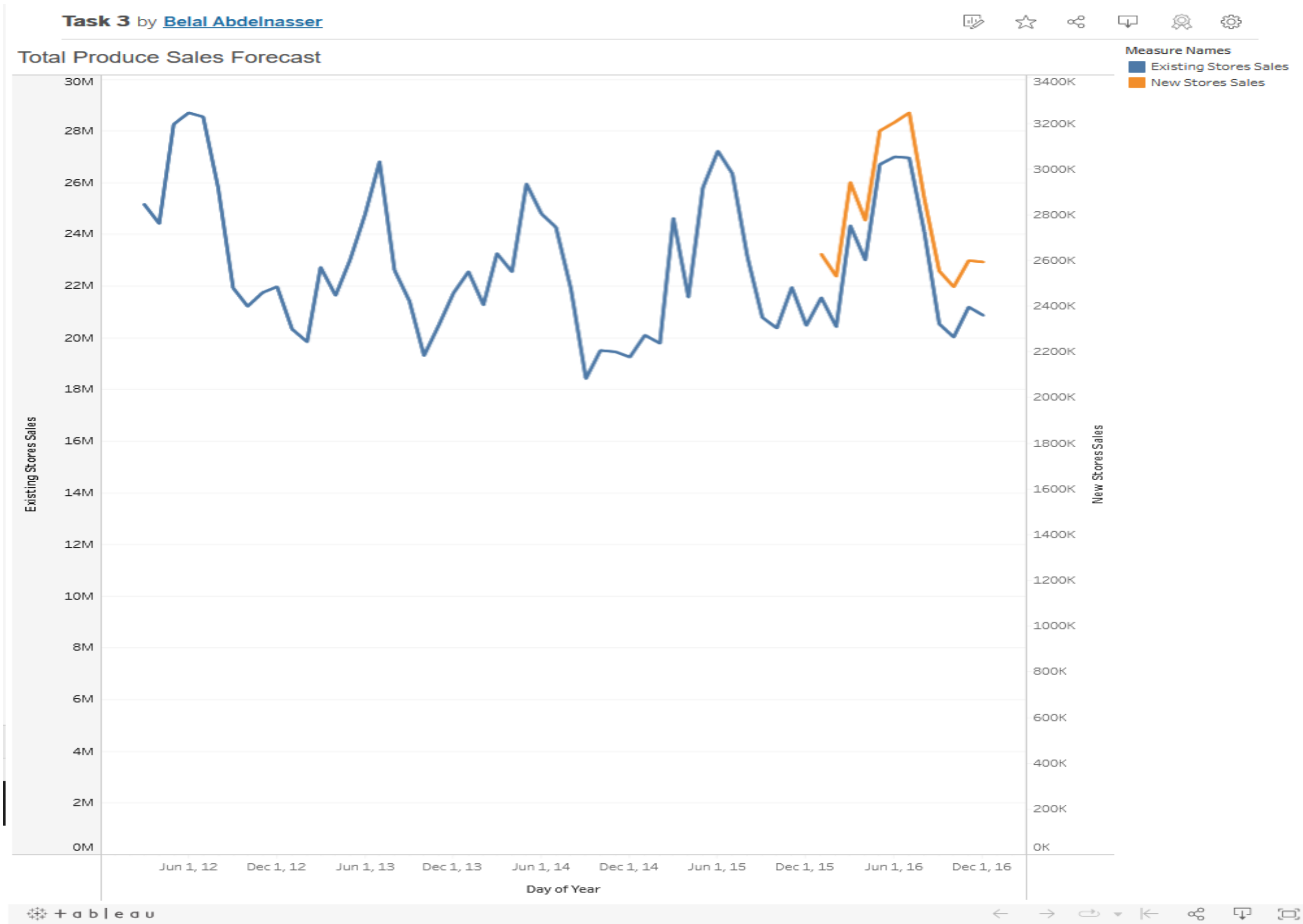


Figure 2 - https://public.tableau.com/app/profile/r221609/viz/Task3_53/Task3