Project: Predictive Analytics Capstone
Project 6: Combining Predictive Techniques

PROJECT SUBMISSION WRITTEN BY CHIDIEBERE STEPHEN NWOSU

Task 1: Determine Store Formats for Existing Stores

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

Answer:

The optimal number of formats is 3. I aggregated the data, used the K-Centroid Diagnostic tool and used the K-means clustering model to choose the desirable store format of 3 and based on the Adjusted Rand Indices and the Calinski-Harabasz (CH) Indices shown below. The Adjusted Rand and Calinski-Harabasz (CH) indices in figure1 below, both prove highest index for 2 and 3 clusters. Obviously, Cluster3 being the best solution with that having a highest median and a fairly compact spread. So, I chose 3.

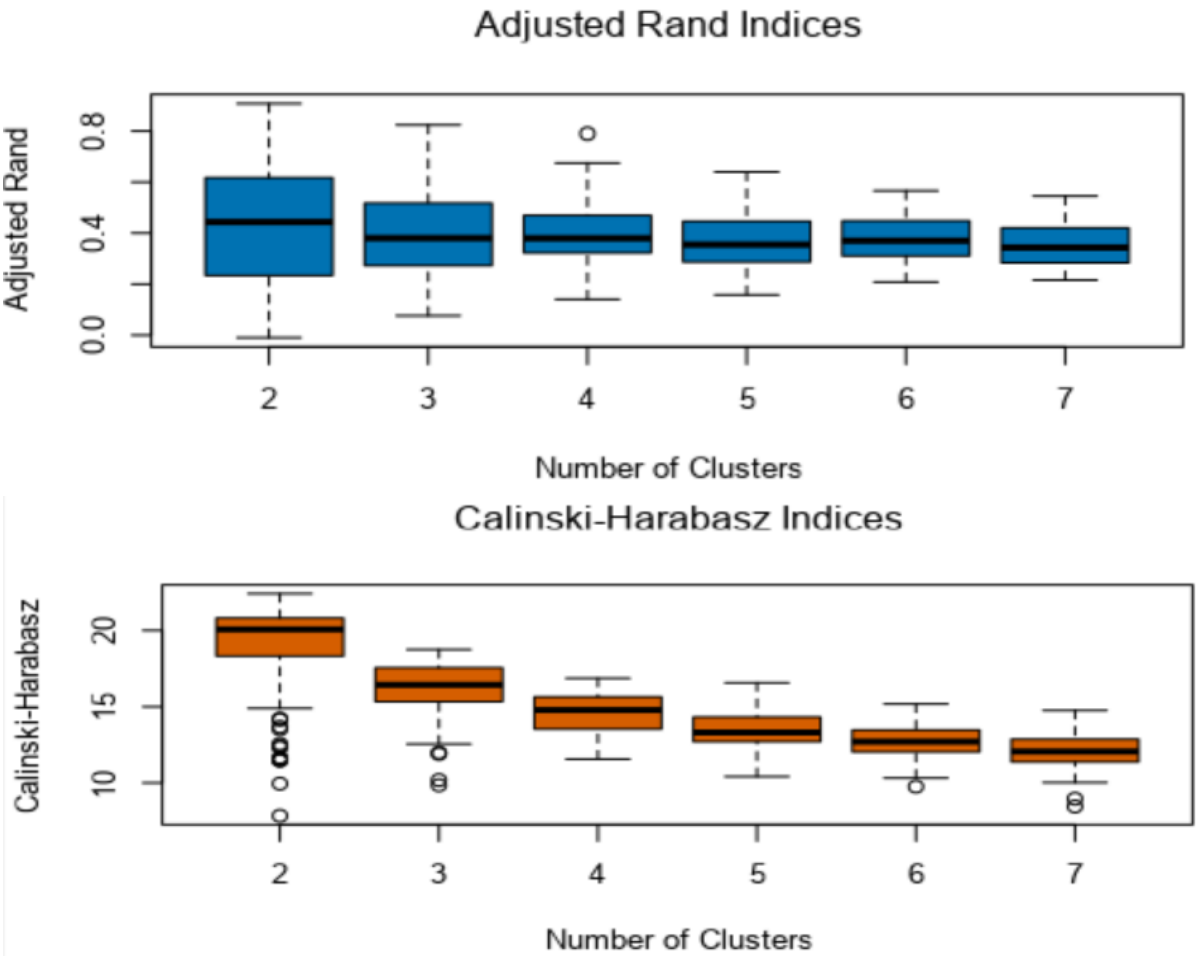


Figure1: Adjusted Rand and Calinski-Harabasz (CH) indices

2. How many stores fall into each store format?

Answer:

Cluster Information:

Cluster	Size
1	25
2	35
3	25

Figure2: Cluster Information

Cluster1 has 25 stores, Cluster2 has 35 stores and Cluster3 has 25 stores.

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

Answer:

From figure3 below, the difference clearly indicates that, cluster2 has the highest volume of sales, next is cluster2 and cluster1 has the least volume of sales.

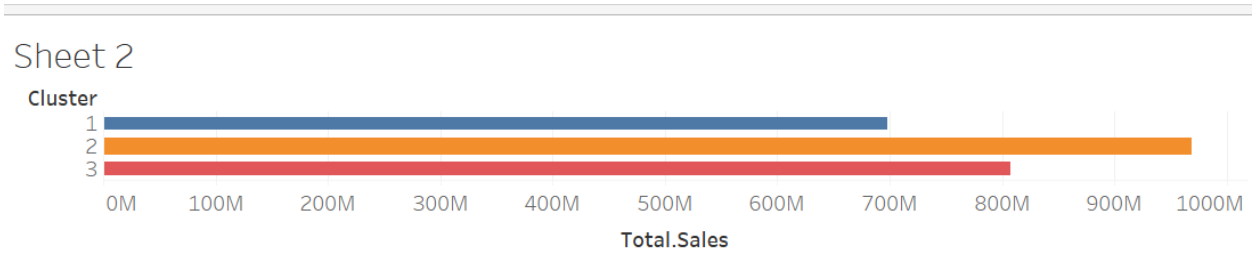


Figure3: Horizonal bar of Cluster versus Total Sales.

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/app/profile/chidiebere.nwosu/viz/Clusterlocationofstores/Sheet1

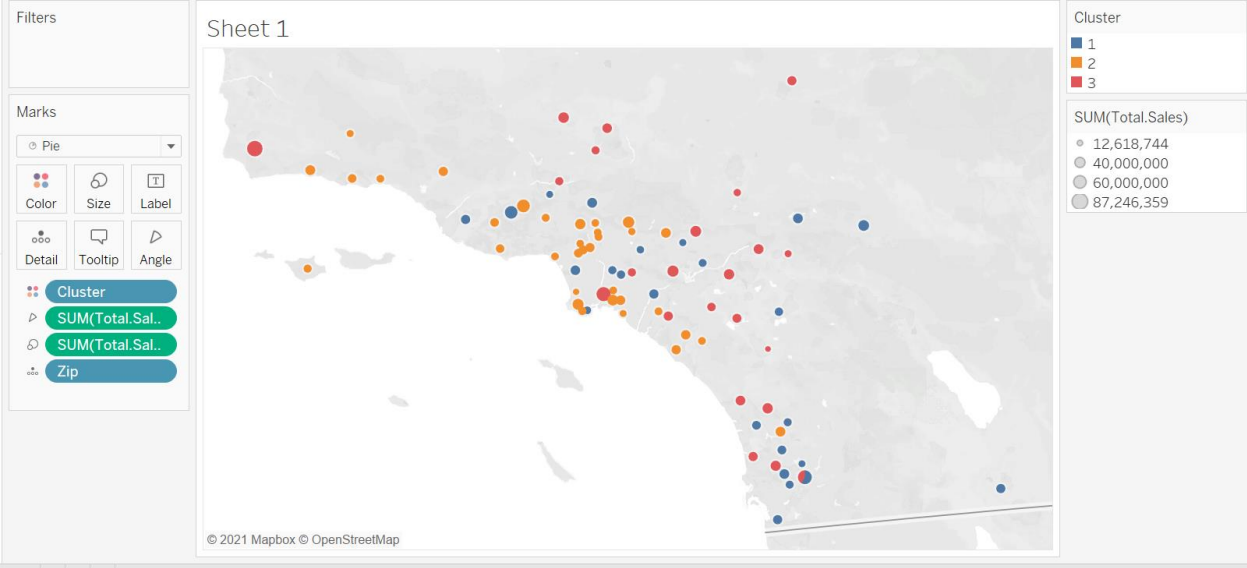


Figure4: Tabular Representation of the Existing Stores.

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.)

Answer:

Overall, I used Decision tree, Boosted and Forest model to train the demographic data, and to predict 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 compared the overall accuracy of the Decision tree, Boosted and Forest model by using the Model comparison tool. I observed below from comparison report that the Boosted model performed better than the other two models. Therefore, boosted model should be used to predict the 10 new stores.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_	0.6471	0.6667	0.5000	1.0000	0.5000
Forest_model	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_model	0.7647	0.8333	0.5000	1.0000	1.0000

Confusion matrix of Boosted_model

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

Confusion matrix of Decision_Tree_

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

Confusion matrix of Forest_model

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

Figure6: Model Comparison Report

2. 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	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	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?

Answer:

The model to use here lies in the scenario of Time Series which involves, linear or exponential trend and constant or increasing seasonality components. From the time series diagram Figure7 below, it clearly shows that there is increasing error similar to an increasing seasonality and there is increasing seasonality as well, no existing trend.

I used ETS(MNM).

Figure7: Time Series

Below is the order to make the time series stationary.

Figure8 below shows, a time series with seasonality, it must be differenced to make it stationary.

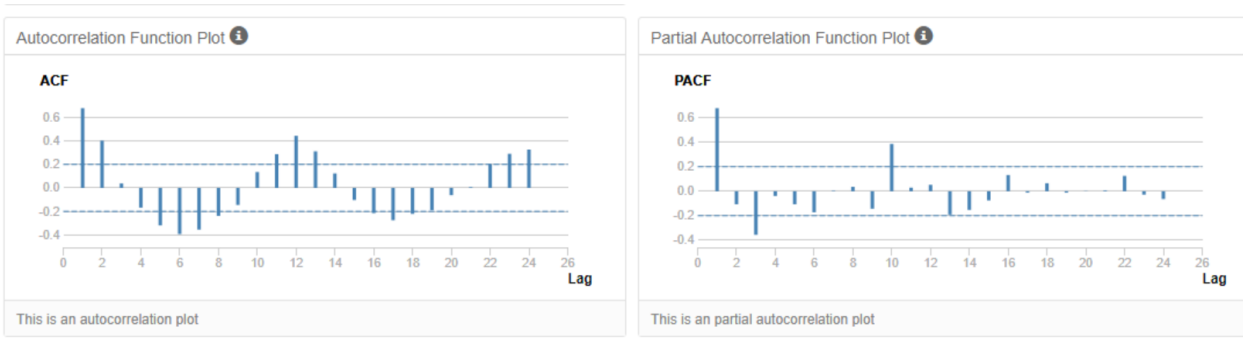


Figure8: Time Series with Seasonality

From Figure9 below, shows that the difference of a given period and 12 period earlier (1year) has been accounted for. But the time series is not completely stationary.

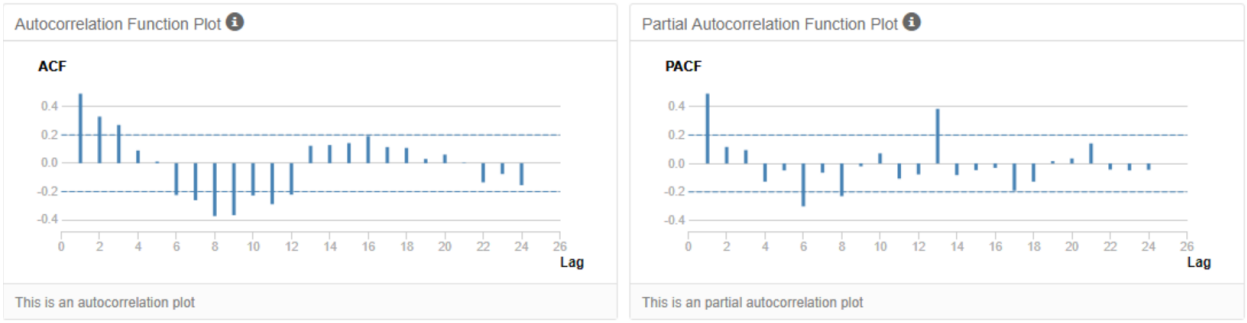


Figure9: Difference of a given period and 12 period.

Figure10 below, depict that the time series is completely stationary, by applying first differencing.

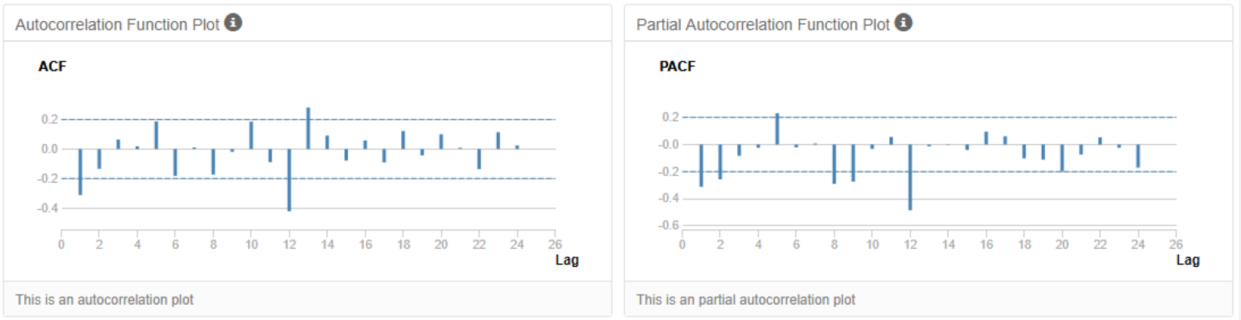


Figure10: Completely Stationariz Time Series.

With the diagrams in Figure9 and 10 above, I used the ARIMA(0,1,0)(0,1,2)[12] model, because I used seasonal difference and first seasonal difference and there is a Lag-2.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA_MA2	-1280274.16	1523735.3	1332388.1	-5.7356	5.9731	0.784

Figure11.

Overall, hold sample of 6month data from the time series was ran and compared between the ETS and ARIMA model. From the table above in Figure11, it proves that the **ETS model** has a better accuracy measure, because it has a lower RMSE and MASE value compare to the **ARIM model** that has higher values. I'll choose **ETS model**.

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,

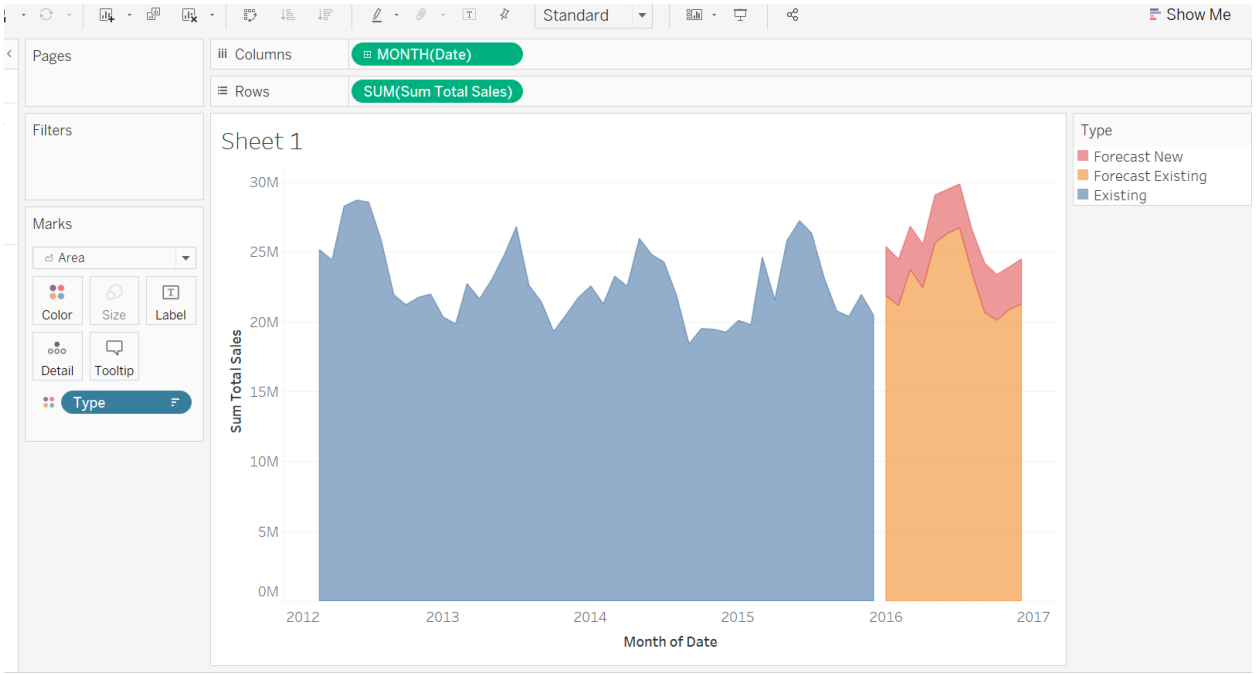
and new store forecast.

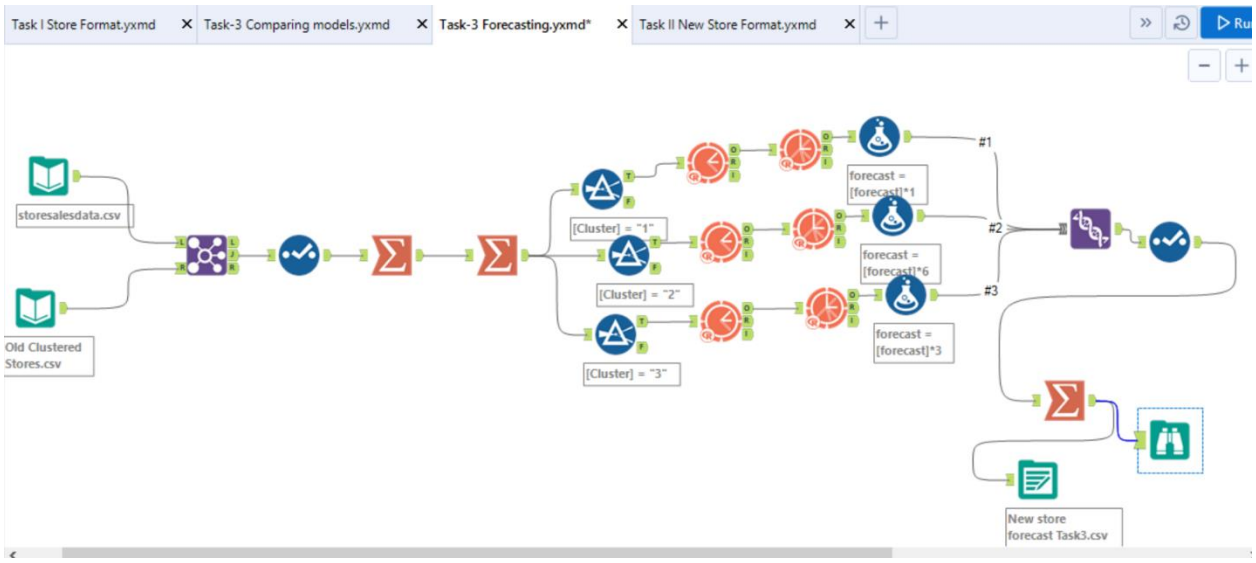
Serial Number	Month	Existing Stores	New Stores
1	January 2016	21,829,060	2,634,749
2	February 2016	21,146,329	2,490,130
3	March 2016	23,735,686	2,437,934
4	April 2016	22,409,515	2,455,686
5	May 2016	25,621,828	2,580,550
6	June 2016	26,307,858	2,460,094
7	July 2016	26,705,092	2,461,334
8	August 2016	23,440,761	2,482,275
9	September 2016	20,640,047	2,640,179
10	October 2016	20,086,270	2,527,917
11	November 2016	20,858,119	2,472,497
12	December 2016	21,255,190	2,53,1870

Here is the Visualization of my forecast that includes historical data, existing stores forecasts, and new stores forecasts.

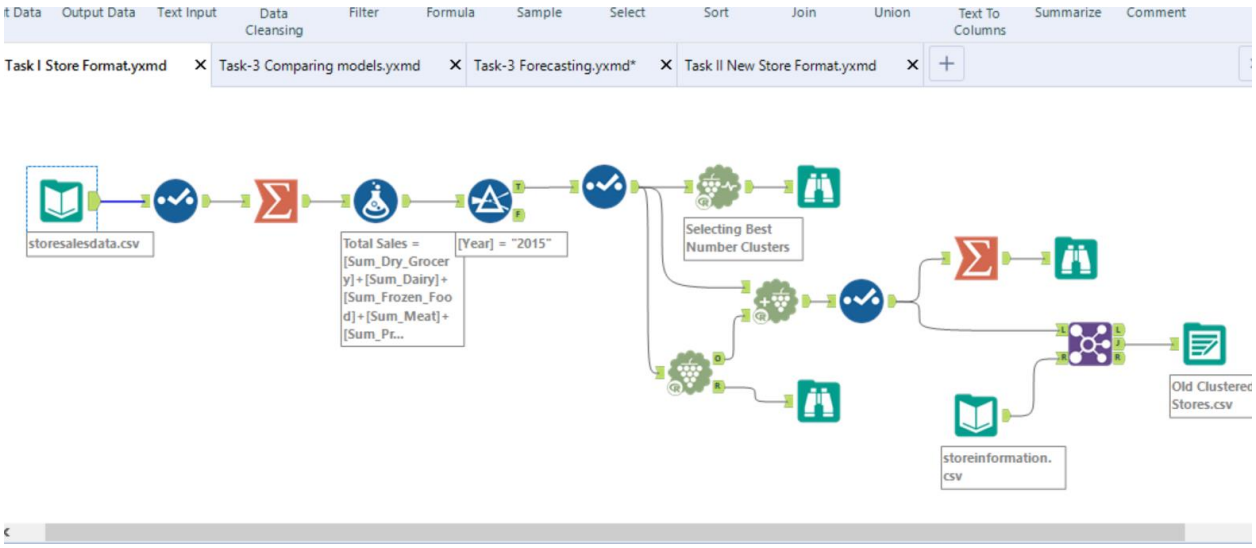
The link to my Tabula public Visualization below.

https://public.tableau.com/app/profile/chidiebere.nwosu/viz/ForecastedSalesValue/Sheet1

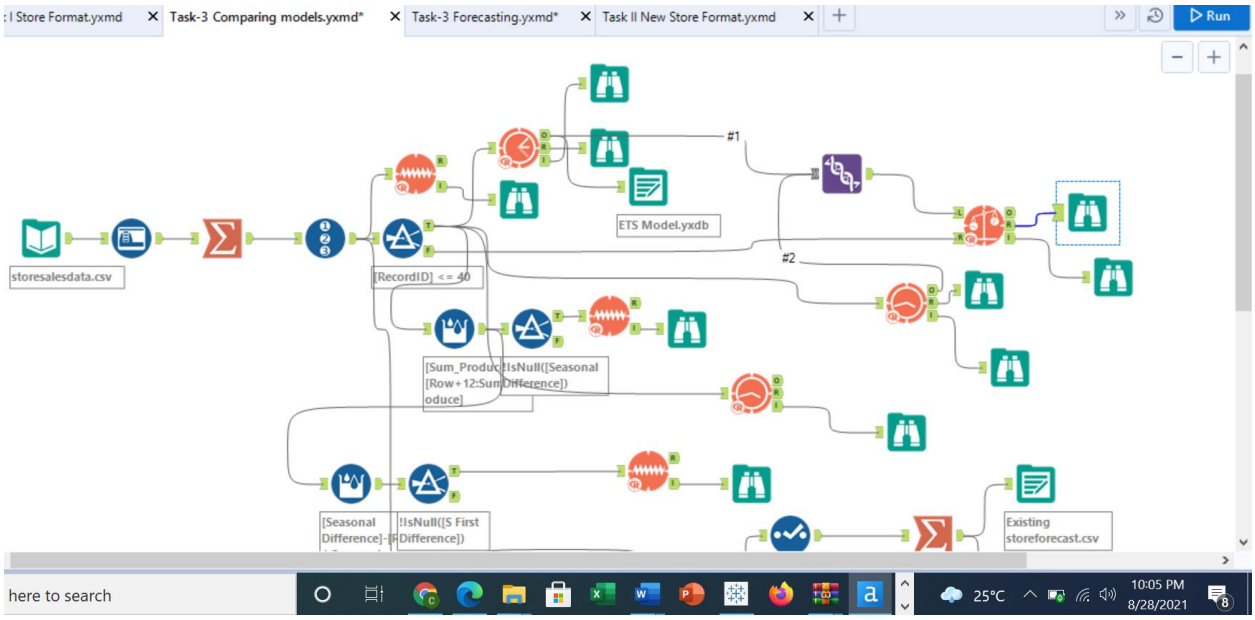




Task3 Forecasting



Store Format



Comparing model