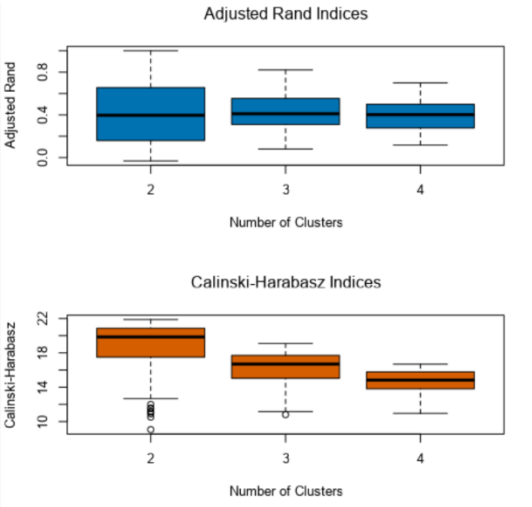
Project: Predictive Analytics Capstone

- I have provided screenshots of all the workflow diagrams in the end of the report.

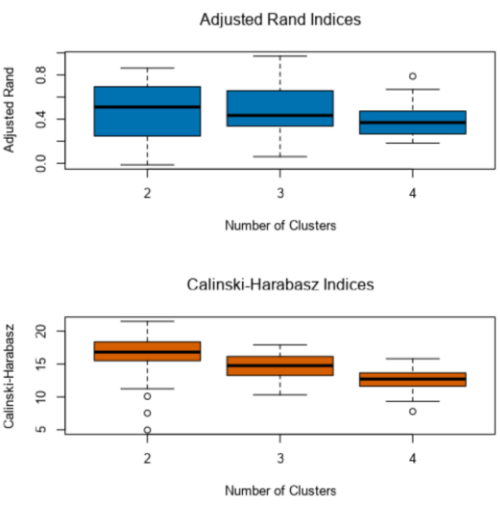
Task 1: Determine Store Formats for Existing Stores

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

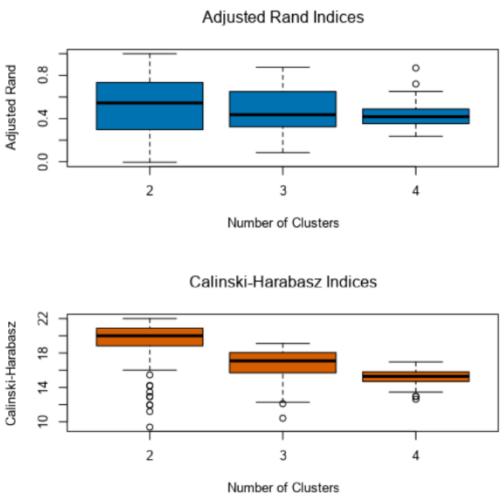
   3 is the optimal number of store formats. I ran the k-centroids diagnostic tool for the 3 different models considering the percentage of sales per category per store fields. AD indices and CH indices were used to compare the models for different cumber of cluster values.



K-Means



K-Medians



Neural Gas

We observe from the index plots that though 2 has higher median, the spread/variance is more. 4 cannot be considered because though in the AR indices it shows lesser variance and just a little less median than 3. In CH indices, 4 has very less median than 3. Thus, it is apt to select 3 as the K value since it has higher median and relatively less variance in both measurement indices.

Further analyzing the spread of the indices value among the different models for 3, we observe that k-Means is the best model

Models	K-Means		Neural Gas		K-Medians	
Indices	AR	CH	AR	CH	AR	CH
Minimum	0.08	10.8	0.08	10.41	0.06	10.32
Maximum	0.82	19.08	0.87	19.10	0.97	17.9

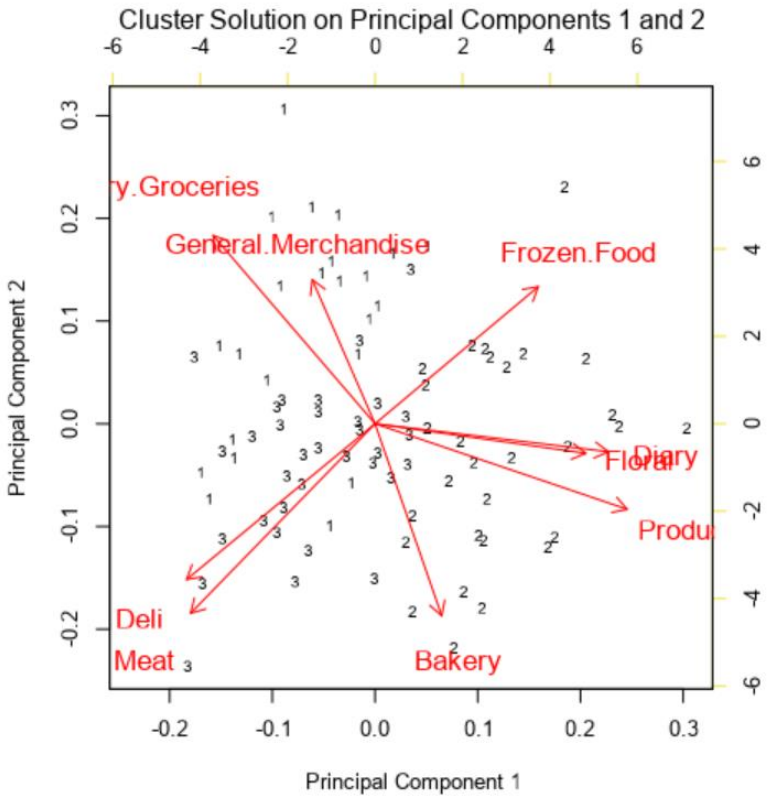
2. How many stores fall into each store format?

Cluster Information:

Cluster	Size
1	23
2	29
3	33

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

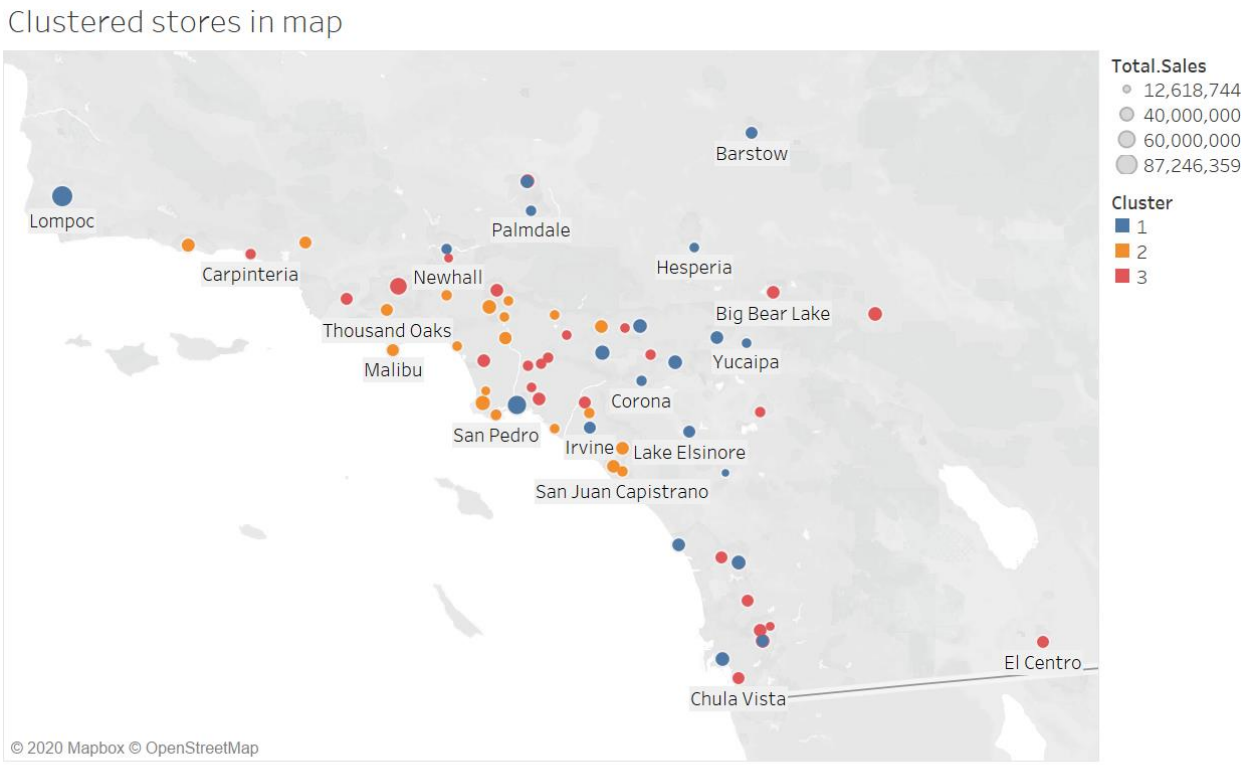
Looking at the plot and Cluster centroid values we observe that 1st Cluster has more effect of General Merchandise and Dry Grocery selling stores. 2nd Cluster has more effect of produce, diary and floral selling stores. 3rd Cluster has more effect of deli and meat stores.



	Dry.Groceries	Diary	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.Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

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.



Map based on Longitude (generated) and Latitude (generated). Color shows details about Cluster. Size shows sum of Total.Sales. The marks are labeled by City. Details are shown for Zip.

Tableau Public Link:
https://public.tableau.com/profile/mohan.raj8847#!/vizhome/Udacity_PAFB_nanodegree_FinalProject_Task1/Sheet1?publish=yes

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.)
    a. Used a test split to get 20% validation sample.
    b. Trained 3 models each of Decision Tree, Forest model and boosted model with the train samples.
    c. Performed union of the 3 models to compare the accuracy of the models using model comparison tool. From which, Boosted Model was observed to have better accuracy, F1 values and less confusion in the confusion matrix
    d. Hence, Boosted Model was used to predict the best store format for the new stores

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_6	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667
Forest_model	0.8235	0.8426	0.7500	1.0000	0.7778

Confusion matrix of Boosted_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of Decision_Tree_6

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

Confusion matrix of Forest_model

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
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?

I used ETS(M,N,M) model for forecasting. Because:

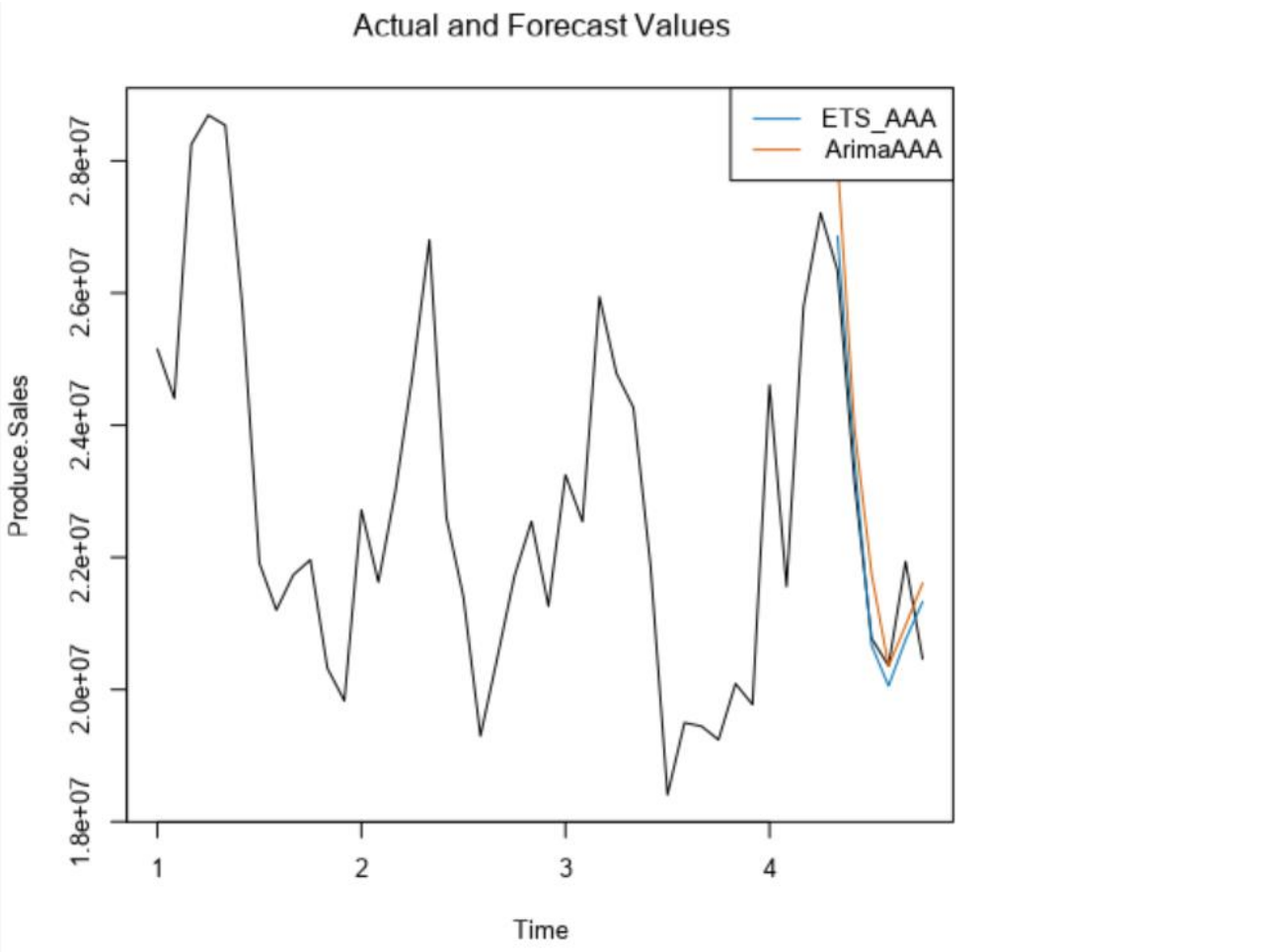
- I used auto selection of type for error, trend and seasonality. Running the model gave output with MNM configuration for ETS model
- I used auto selection of type for AR, I and MA values in a ARIMA model. Running the model gave outputs in (1,0,0)(1,1,0) configuration
- Combining the results of the 2 models using union and running it through a TS compare tool showed ETS model to be more accurate in predicting the hold off samples.

Actual and Forecast Values:

Actual	ETS_AAA	ArimaAAA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

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

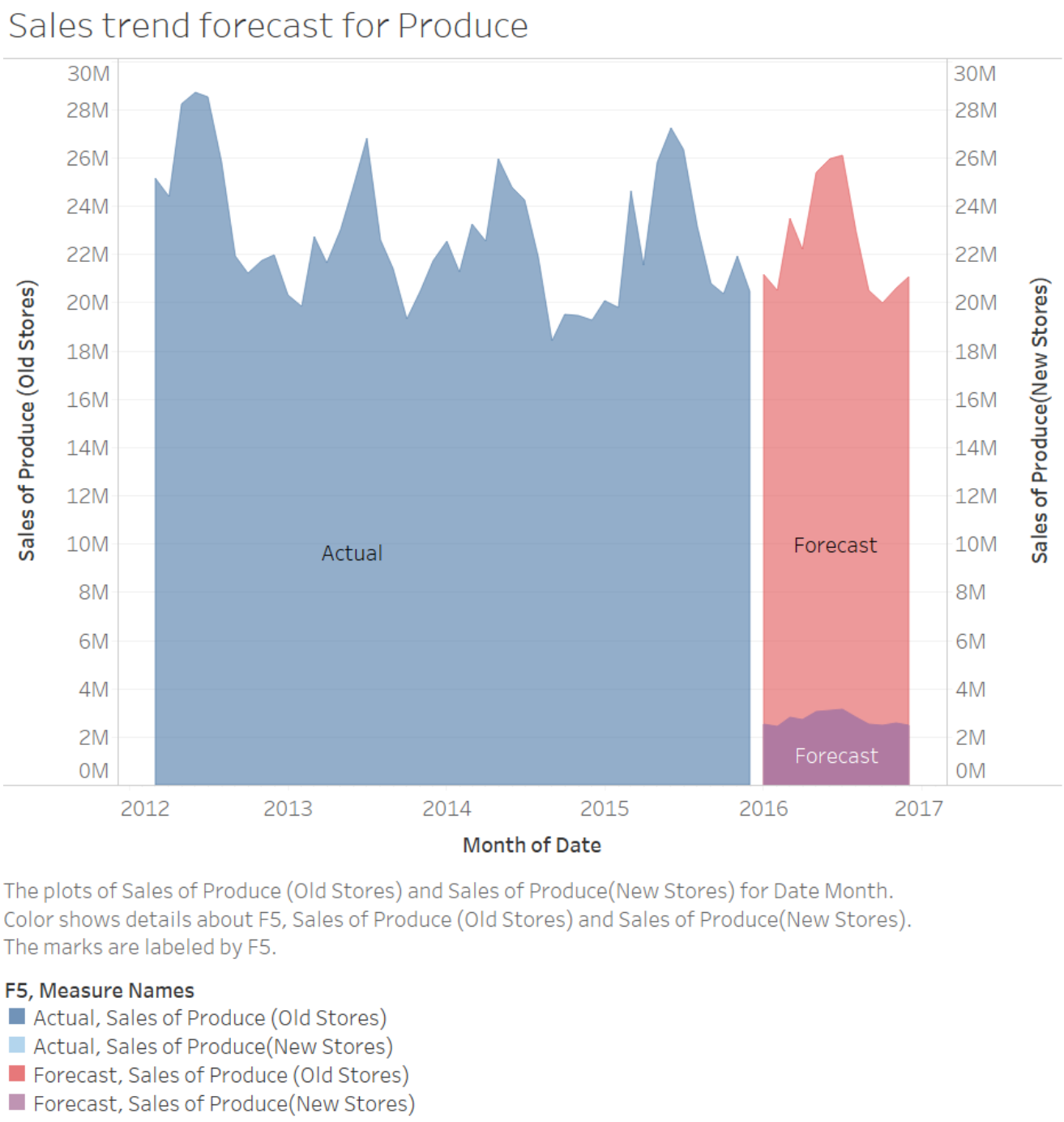


3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Year	Month	New Stores sales	Existing store sales
2016	1	2527763	21136642
2016	2	2418610	20507039
2016	3	2812488	23506566
2016	4	2696077	22208406
2016	5	3042604	25380148
2016	6	3094376	25966799
2016	7	3141200	26113793
2016	8	2796594	22899286
2016	9	2498076	20499584
2016	10	2466227	19971243
2016	11	2553585	20602666
2016	12	2478434	21073222

Visualization in next page.

Historical data - Blue, existing stores forecasts- Pink, and new stores forecasts-Purple.



https://public.tableau.com/profile/mohan.raj8847#!/vizhome/Salestrendforecast/Sheet1?publish=yes

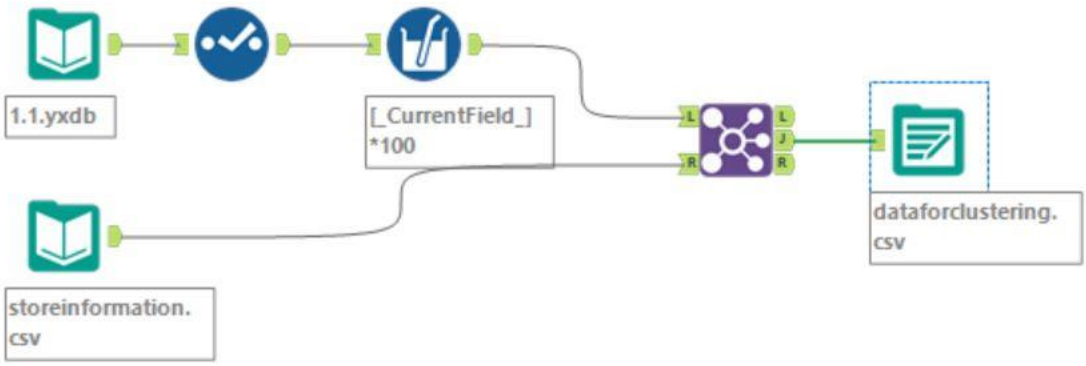


Figure 1: 1.1 Initial Data Prep

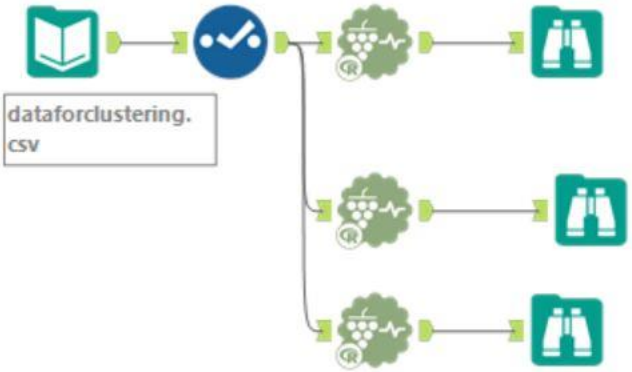


Figure 2: 1.2 Cluster model Assesments

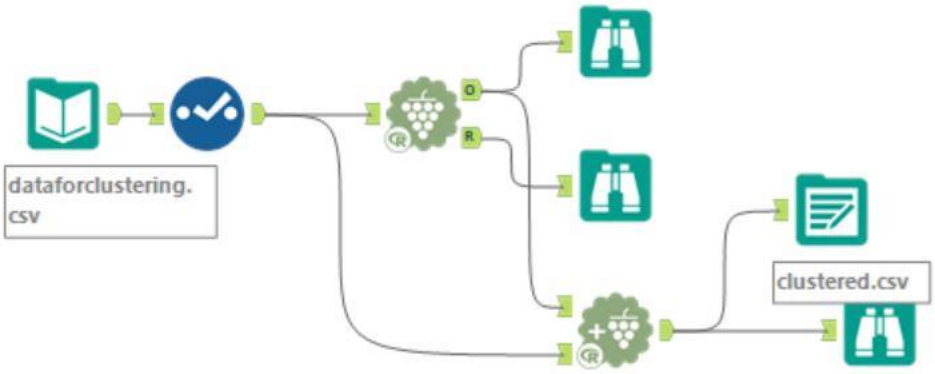


Figure 3: 1.3 Clustering

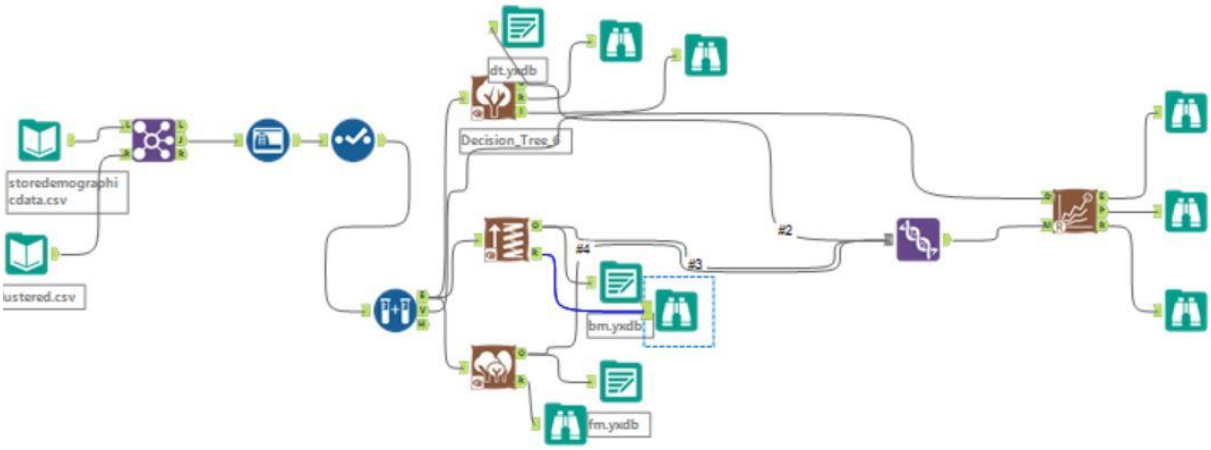


Figure 4: 2.1 Analysing Prediction models

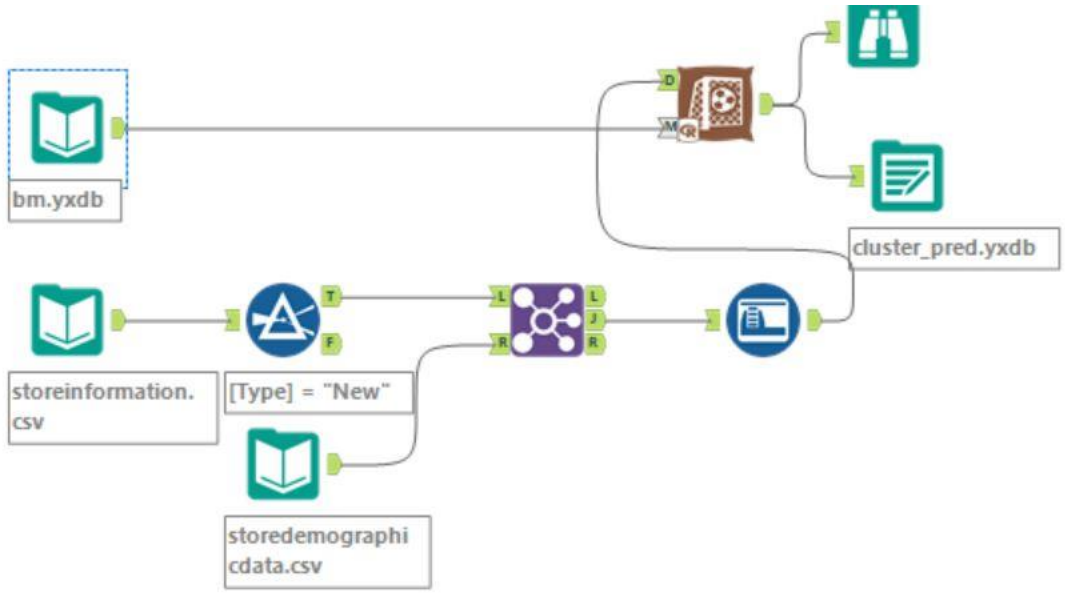


Figure 5: 2.2 Predicting the clusters

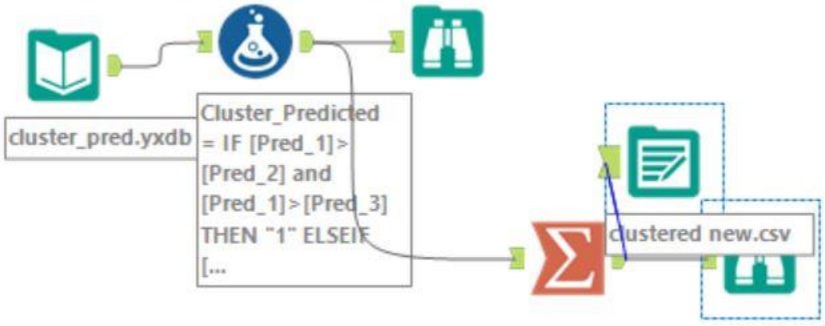


Figure 6: 2.3 Appending cluster values

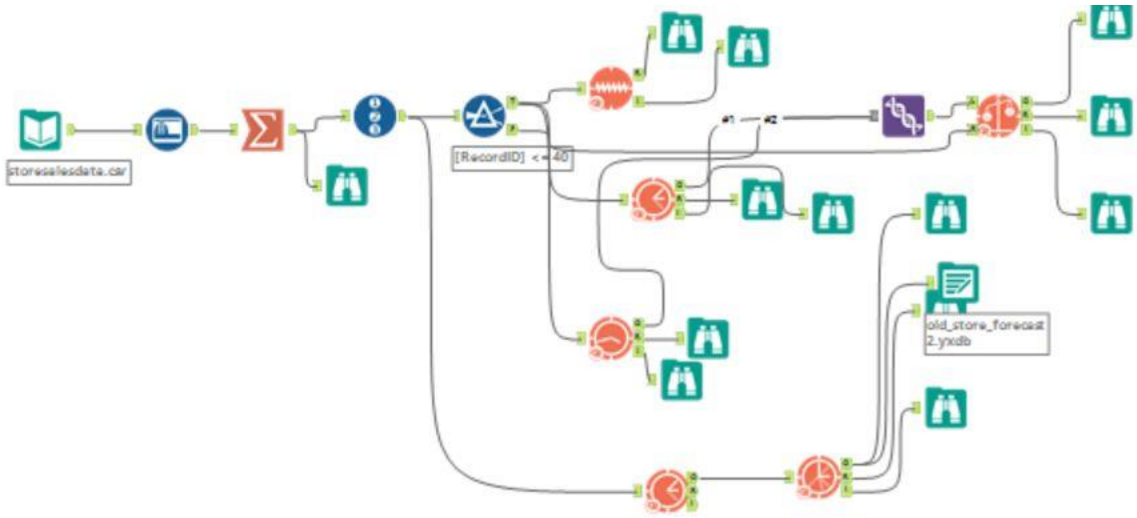


Figure 7: 3.1 Forecasting existing store data

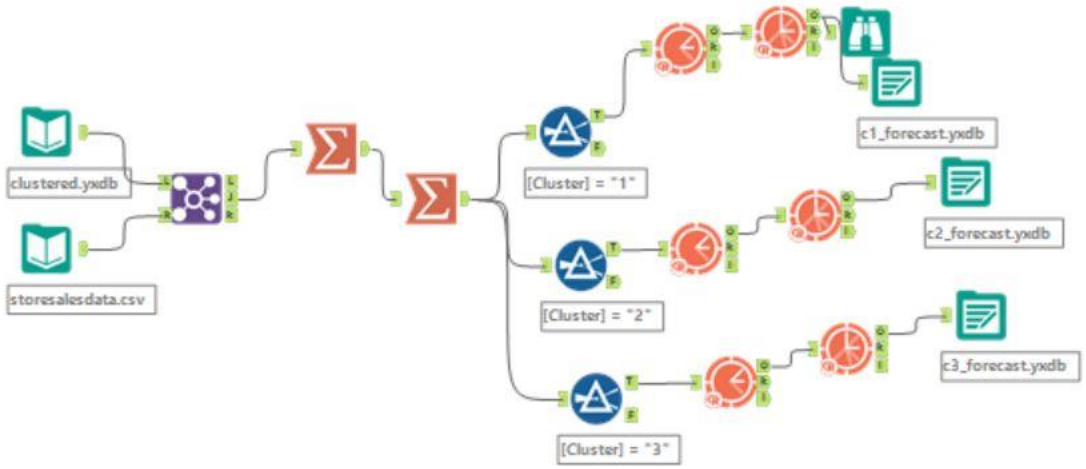


Figure 8: 3.2 Cluster specific Forecasting

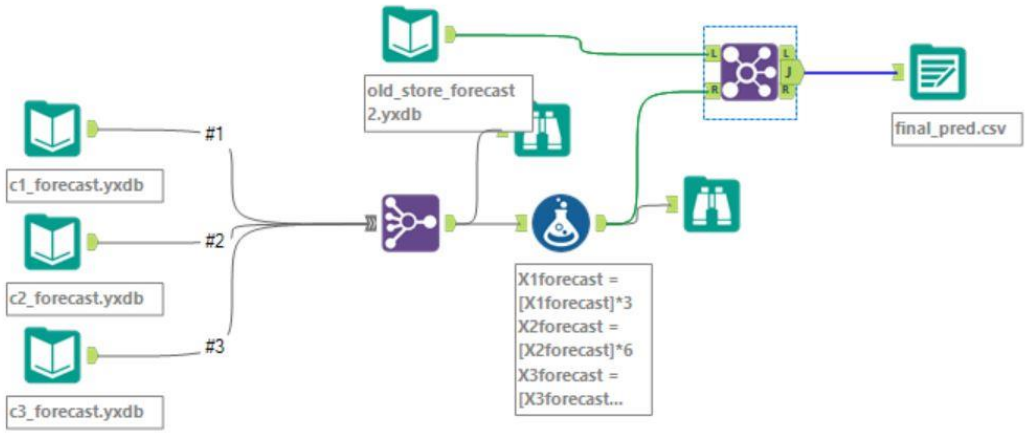


Figure 9: Combining all forecasts