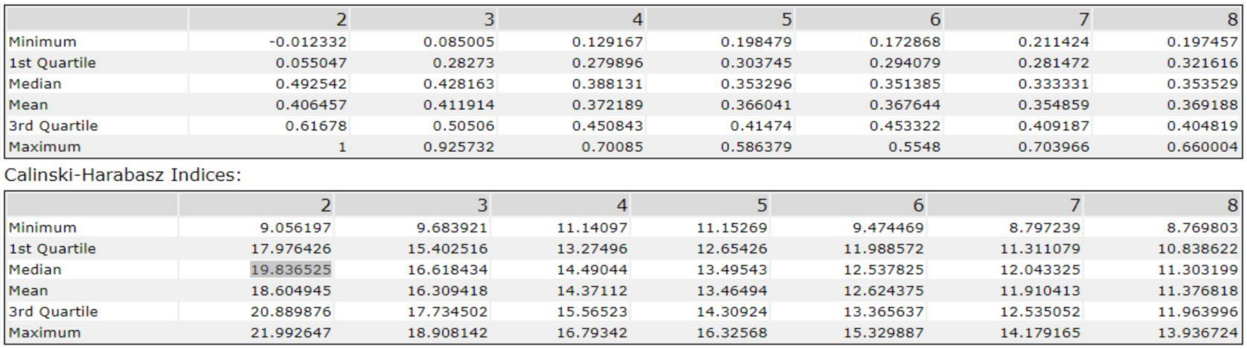
Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

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

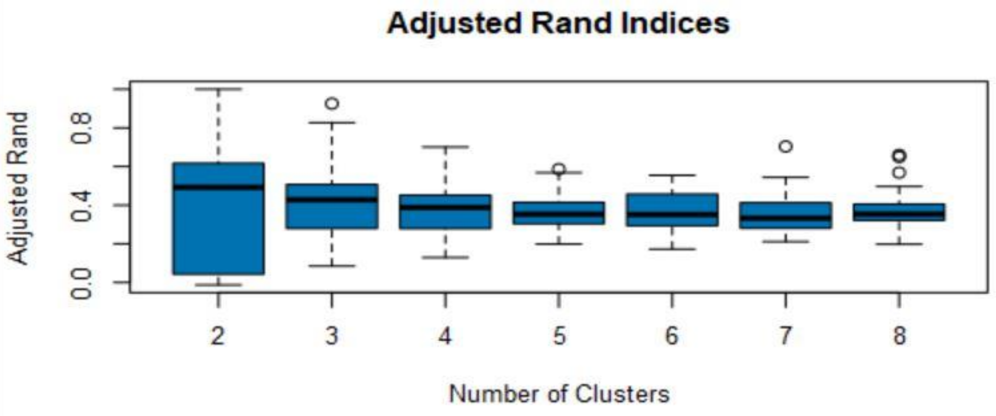
The optimal number of store formats is 3 as it
is stated in the supporting material that cluster must not have less than 20 and not over 40 stores.

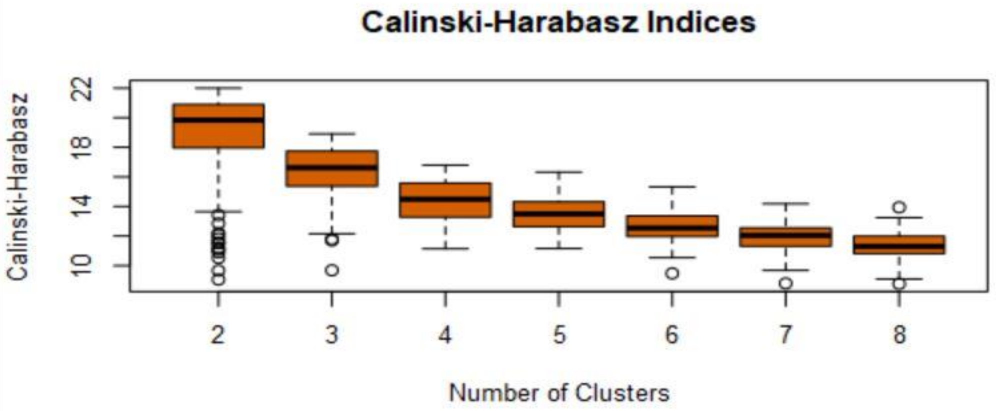


	2	3	4	5	6	7	8
Minimum	-0.012332	0.085005	0.129167	0.198479	0.172868	0.211424	0.197457
1st Quartile	0.055047	0.28273	0.279896	0.303745	0.294079	0.281472	0.321616
Median	0.492542	0.428163	0.388131	0.353296	0.351385	0.333331	0.353529
Mean	0.406457	0.411914	0.372189	0.366041	0.367644	0.354859	0.369188
3rd Quartile	0.61678	0.50506	0.450843	0.41474	0.453322	0.409187	0.404819
Maximum	1	0.925732	0.70085	0.586379	0.5548	0.703966	0.660004

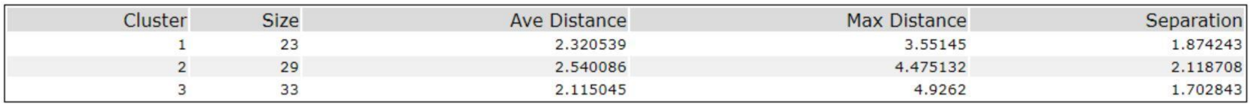
Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	9.056197	9.683921	11.14097	11.15269	9.474469	8.797239	8.769803
1st Quartile	17.976426	15.402516	13.27496	12.65426	11.988572	11.311079	10.838622
Median	19.836525	16.618434	14.49044	13.49543	12.537825	12.043325	11.303199
Mean	18.604945	16.309418	14.37112	13.46494	12.624375	11.910413	11.376818
3rd Quartile	20.889876	17.734502	15.56523	14.30924	13.365637	12.535052	11.963996
Maximum	21.992647	18.908142	16.79342	16.32568	15.329887	14.179165	13.936724



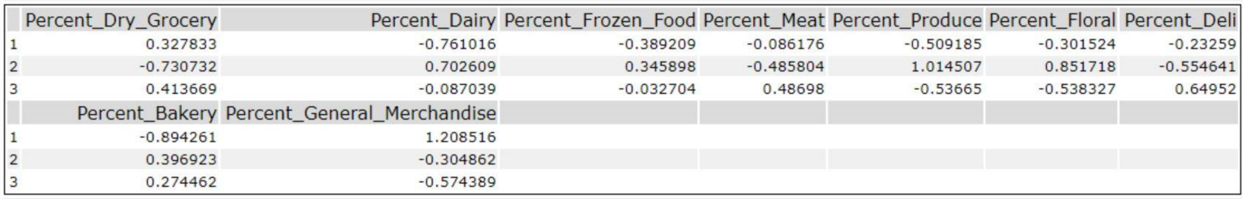


2. How many stores fall into each store format?
   Cluster 1 has 23 stores, cluster 2 has 29 stores and cluster 3 has 33 stores.



Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Based on the result shown below, cluster 1 sells more of General Merchandise compared to the others. Cluster 2 sells more of Produce compared to the others.

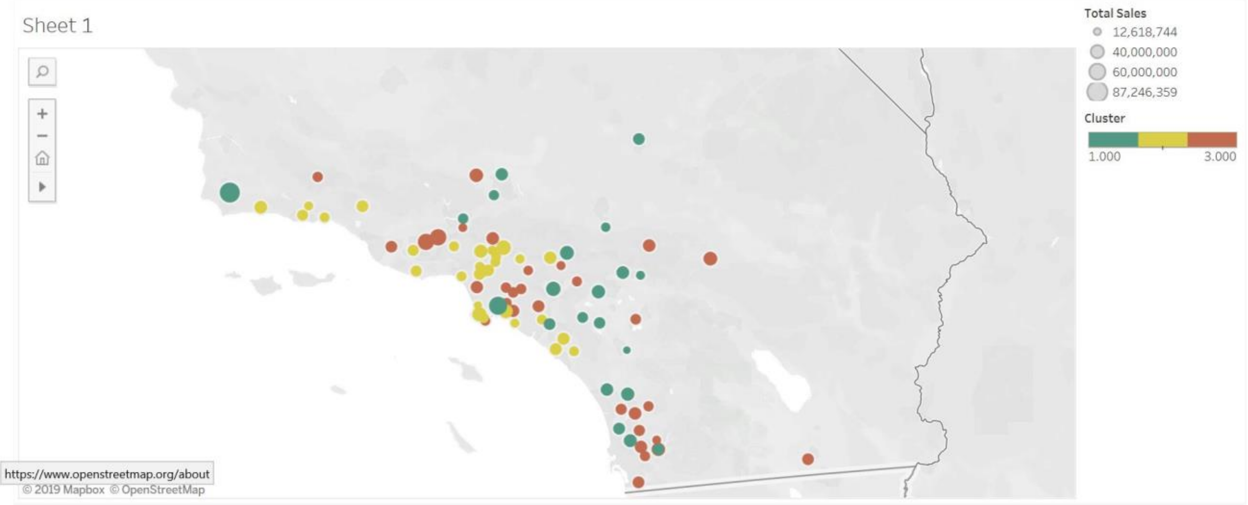


	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Percent_Bakery	Percent_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.

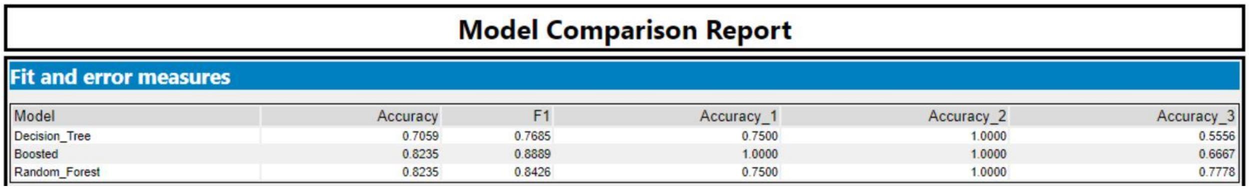
   https://public.tableau.com/views/Task1_15578447757820/Sheet1?:embed=y&:display_count=yes&:origin=viz_share_link



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.)

   Boosted and Random_Forest have the same accuracy, but Boosted Model is chosen due to higher F1 value of 0.8889.



Model Comparison Report

Fit and error measures

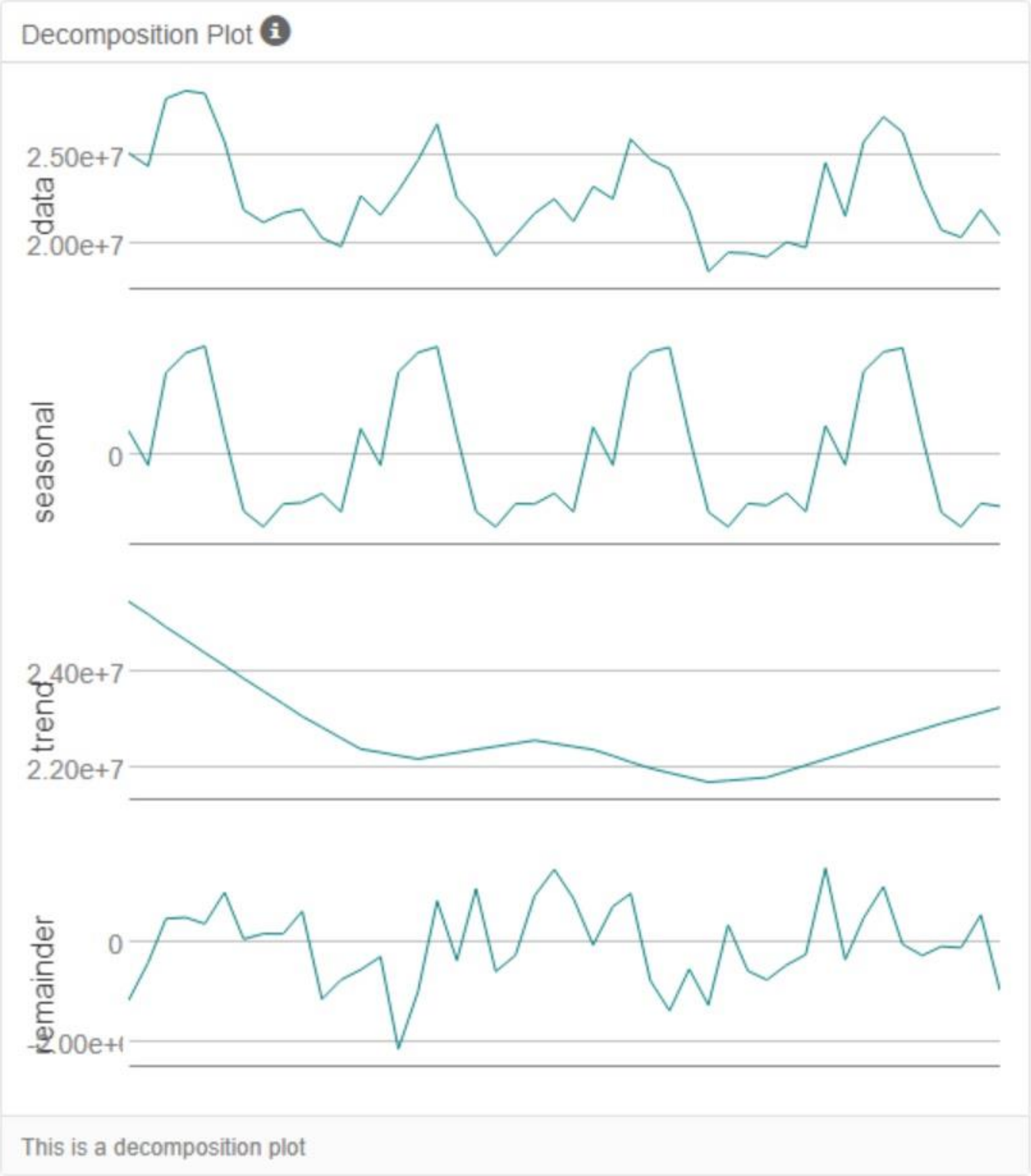
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667
Random_Forest	0.8235	0.8426	0.7500	1.0000	0.7778

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

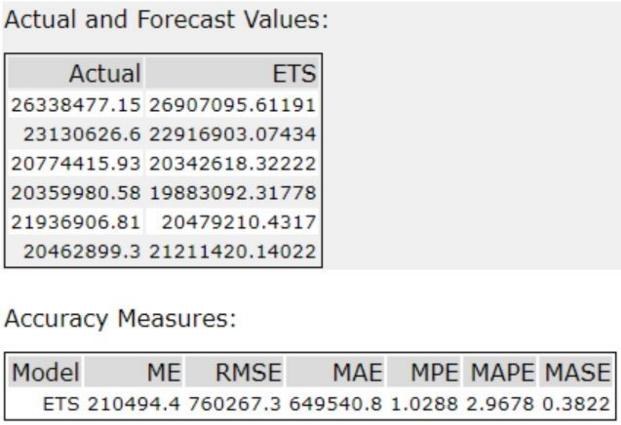
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?



Decomposition Plot

This is a decomposition plot

The time series decomposition plot shown above allows us to observe the seasonality, trend and error terms of a time series. There is no clear trend, so no trend component is included (N). The size of the seasonal fluctuations tends to increase or decrease with the level of time series, so we apply it multiplicatively (M). The error plot is fluctuating between large and small errors over time, we apply it multiplicatively (M).

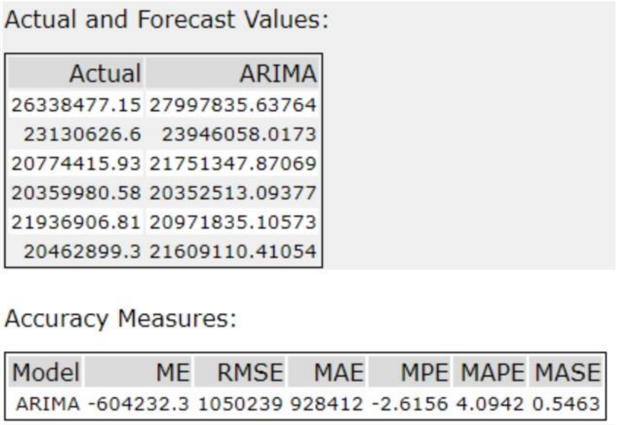


Actual and Forecast Values:

Actual	ETS
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822



Actual and Forecast Values:

Actual	ARIMA
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

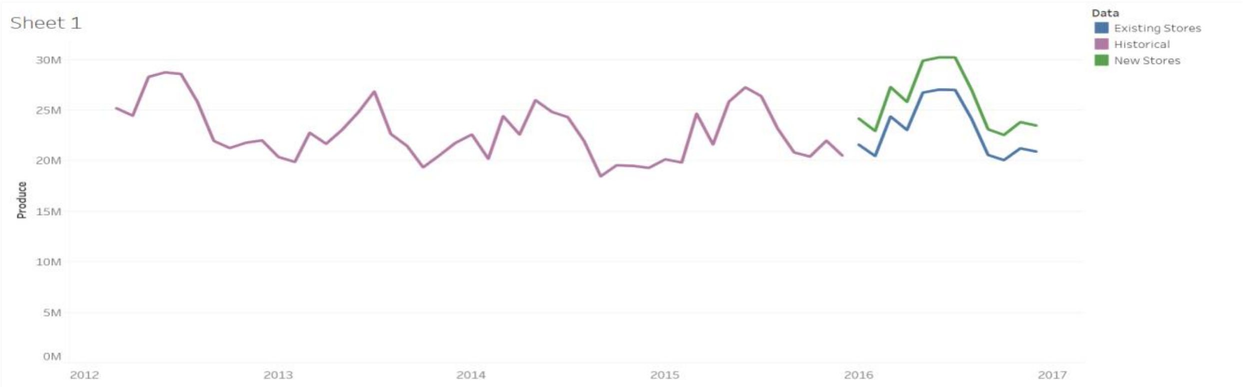
By comparing the forecast and actual results, we can see that ETS model's accuracy is higher with overall lower errors across all variable. The ETS model's RMSE (760,267.3) and MASE (0.3822) are lower.

Based on the above, ETS(M,N,M) is chosen as our forecasting mode.

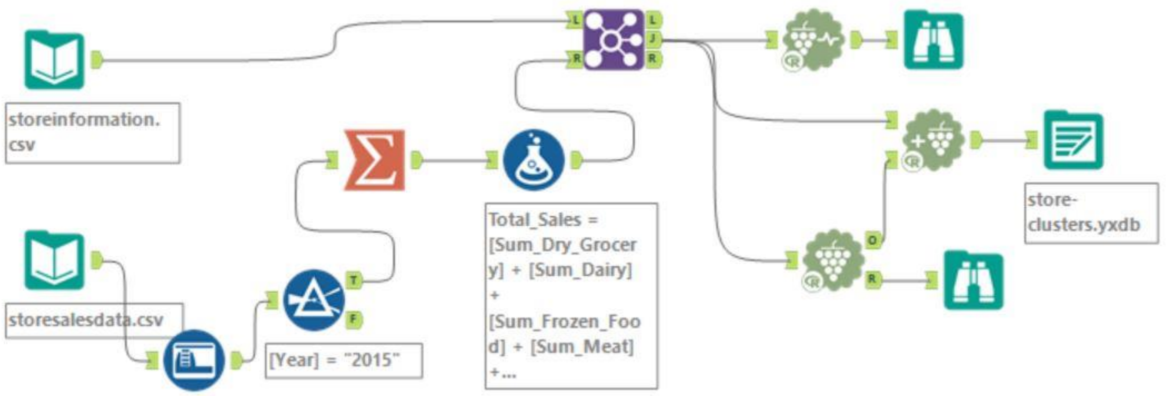
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Month	Existing Stores	New Stores
Jan 2016	21,539,936	2,587,451
Feb 2016	20,413,771	2,477,353
Mar 2016	24,325,953	2,913,185
Apr 2016	22,993,466	2,775,746
May 2016	26,691,951	3,150,867
Jun 2016	26,989,964	3,188,922
Jul 2016	26,948,631	3,214,746
Aug 2016	24,091,579	2,866,349
Sep 2016	20,523,492	2,538,727
Oct 2016	20,011,749	2,488,148
Nov 2016	21,177,435	2,595,270
Dec 2016	20,855,799	2,573,397

https://public.tableau.com/views/TotalSalesForecast_15593713505240/Sheet1?:embed=y&:display_count=yes&publish=yes&:origin=viz_share_link

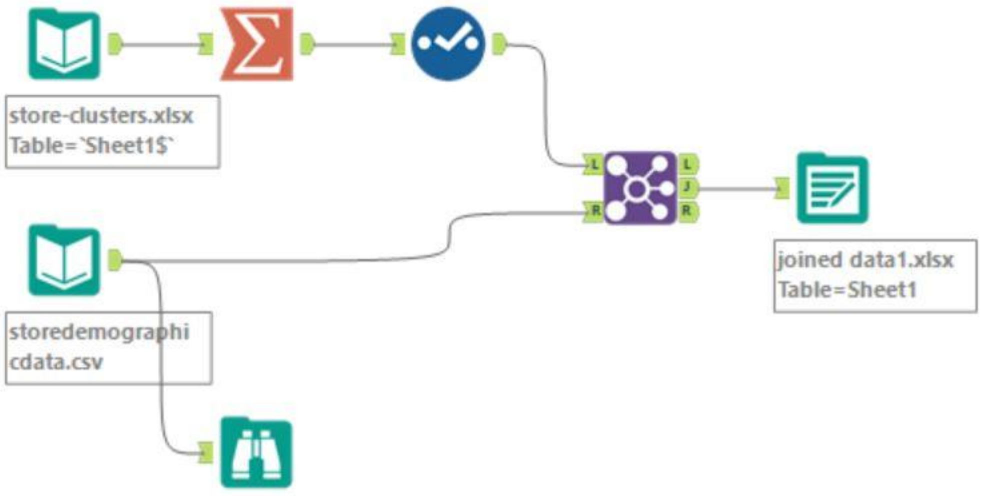


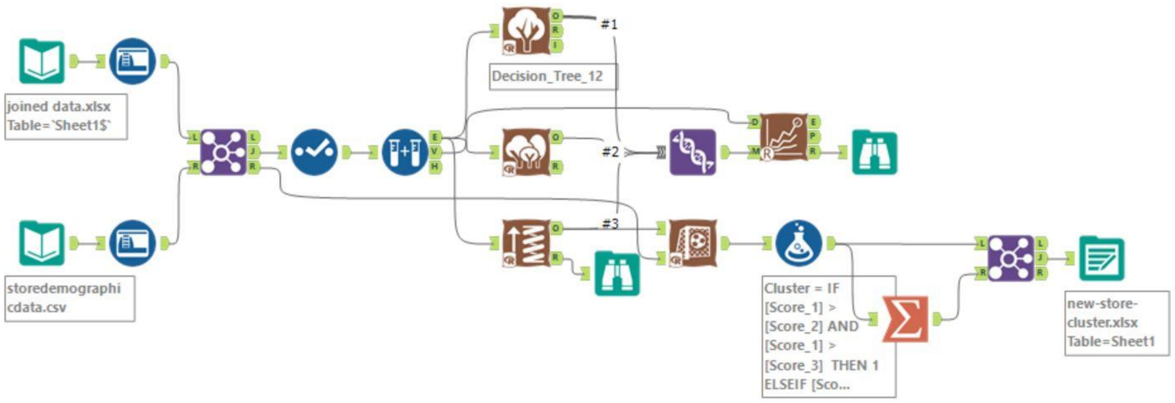
Alteryx Data Flow



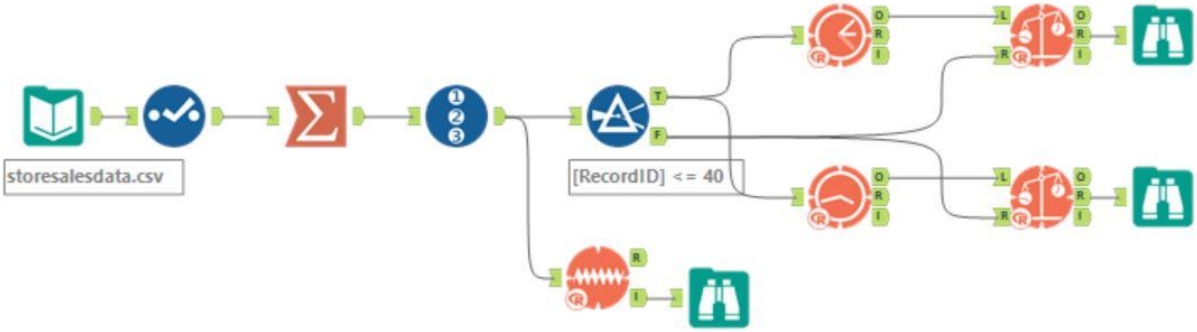
```
[Year] = "2015"
```

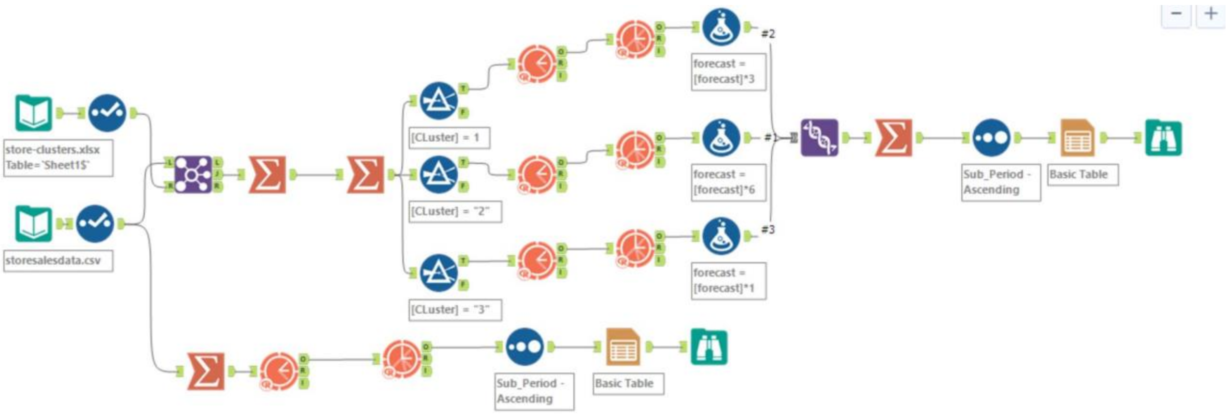
```
Total_Sales =
[Sum_Dry_Grocer
y] + [Sum_Dairy]
+
[Sum_Frozen_Foo
d] + [Sum_Meat]
+...
```





```
Cluster = IF
[Score_1] >
[Score_2] AND
[Score_1] >
[Score_3] THEN 1
ELSEIF [Sco...
```





Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.