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**.*

   *This was arrived at by using the K-Centroids Diagnostics and the K-Centroids Cluster Analysis tools in Alteryx. The number of formats was taken from the **K-means** results of the K-Centroids Diagnostics tool. As may be seen below, despite having higher medians, 2 clusters did not present a fair image of the data when visualized in all the configurations, and since 3 clusters from K-Medians and Neural Gas configurations did not meet the specifications, 3 clusters from K-Means was tested and used.*

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.017586	0.208197	0.181585	0.133772	0.158757	0.222502	0.21093
1st Quartile	0.352613	0.377392	0.302314	0.331809	0.314419	0.299658	0.322749
Median	0.509257	0.466169	0.398104	0.380556	0.387434	0.366279	0.375409
Mean	0.494056	0.479493	0.404888	0.388834	0.39306	0.381404	0.384298
3rd Quartile	0.693746	0.58771	0.481097	0.454895	0.46369	0.447859	0.436717
Maximum	0.952939	0.788895	0.661744	0.614672	0.64242	0.62851	0.720498

	9	10
Minimum	0.244439	0.212783
1st Quartile	0.325103	0.315087
Median	0.386151	0.380127
Mean	0.390303	0.379638
3rd Quartile	0.457811	0.442954
Maximum	0.538277	0.604545

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	10.38298	10.31461	11.34984	10.77356	9.80353	9.577281	9.253901
1st Quartile	18.69647	16.03968	14.46704	12.9405	12.24542	11.378557	11.166056
Median	20.07012	17.00754	15.19152	13.65142	12.83476	12.07357	11.697797
Mean	19.08577	16.73685	14.98778	13.68998	12.83426	12.156743	11.681178
3rd Quartile	20.87407	17.78773	15.74729	14.53404	13.67175	12.859807	12.311206
Maximum	22.41555	18.73715	16.93911	16.10526	15.30862	14.460893	13.955665

	9	10
Minimum	8.822973	8.153824
1st Quartile	10.648806	10.002731
Median	11.287124	10.760594
Mean	11.359959	10.745482
3rd Quartile	11.937564	11.429852
Maximum	13.731897	13.433832

K-Medians Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.011625	0.023243	0.165826	0.166425	0.212192	0.175595	0.232058
1st Quartile	0.291722	0.367372	0.320798	0.278806	0.279906	0.27514	0.291267
Median	0.509863	0.557525	0.425894	0.350641	0.315885	0.324318	0.351894
Mean	0.490912	0.51263	0.445272	0.37181	0.345952	0.339311	0.353458
3rd Quartile	0.655035	0.638373	0.515183	0.459978	0.387725	0.399183	0.401008
Maximum	0.952937	0.846231	0.760995	0.648317	0.636514	0.57597	0.526216

	9	10
Minimum	0.119022	0.177088
1st Quartile	0.262001	0.264437
Median	0.318669	0.316678
Mean	0.325854	0.324894
3rd Quartile	0.386093	0.364165
Maximum	0.537179	0.500337

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	3.903988	9.545742	9.136534	7.824218	8.203725	8.048619	7.748428
1st Quartile	15.299826	14.048997	12.180618	11.1566	10.167925	9.414703	9.179474
Median	16.889811	15.29056	13.042937	11.729624	10.933304	10.163199	9.860843
Mean	16.325924	14.952343	12.947509	11.77561	10.928587	10.273794	9.846821
3rd Quartile	18.188167	16.095023	13.777309	12.535918	11.759507	11.047944	10.571864
Maximum	21.23701	18.275767	16.024179	14.308403	13.725384	13.848898	12.087412

	9	10
Minimum	6.202633	6.846759
1st Quartile	8.548525	7.951996
Median	9.190434	8.66602
Mean	9.136729	8.669178
3rd Quartile	9.838418	9.325841
Maximum	11.200767	11.424016

Kwame Odoi Otchere

Neural Gas Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	0.002397	0.068269	0.154807	0.192092	0.242117	0.263842	0.220207
1st Quartile	0.451232	0.326333	0.341585	0.337371	0.319306	0.344956	0.365118
Median	0.579991	0.471298	0.420662	0.400811	0.405259	0.436643	0.415016
Mean	0.563813	0.458652	0.441426	0.423175	0.408265	0.420112	0.421285
3rd Quartile	0.734318	0.580912	0.513595	0.486694	0.490069	0.486272	0.466776
Maximum	0.952939	0.861507	0.804965	0.798678	0.609444	0.56248	0.666219

	9	10
Minimum	0.313812	0.284086
1st Quartile	0.37902	0.368301
Median	0.419101	0.398248
Mean	0.434627	0.41099
3rd Quartile	0.490915	0.4454
Maximum	0.667531	0.597406

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	10.77461	12.86636	12.20183	11.83885	10.48533	10.66598	10.14757
1st Quartile	19.31314	16.05372	14.60668	13.5645	12.65724	12.07733	11.48257
Median	20.21008	17.09902	15.21939	14.05685	13.25259	12.53605	12.15169
Mean	19.66541	16.82225	15.14546	14.12861	13.29958	12.65569	12.1028
3rd Quartile	21.01604	17.81209	15.81965	14.69094	13.94259	13.41143	12.72572
Maximum	22.42053	18.90104	16.95503	16.2352	15.31723	14.52259	14.37977

	9	10
Minimum	9.596019	8.800118
1st Quartile	11.017055	10.588682
Median	11.548929	11.157616
Mean	11.682011	11.176003
3rd Quartile	12.317831	11.861388
Maximum	13.669426	13.367846

2. How many stores fall into each store format?
   The distribution of stores per format is summarized in the image below, with formats 1 and 3 having 25 stores each while format 2 had 35 stores;

Record	Cluster	Count
1	1	25
2	2	35
3	3	25

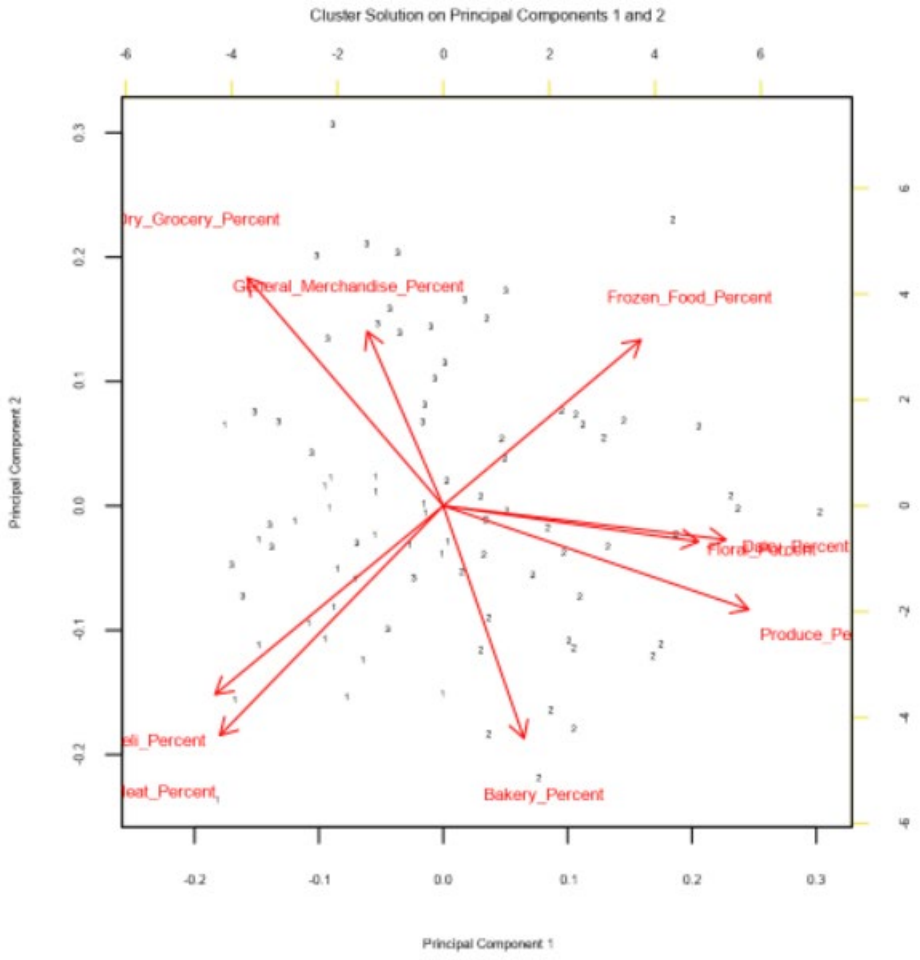
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Based on the summary report below, it is seen that Cluster 1 has the smallest average distance, making it the most compact of all. Additionally, having the highest separation means that it is farthest from the other clusters. Cluster 3 has the least max distance, implying that the objects are not as far apart as in the other clusters as may be supported by its fairly low average distance.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

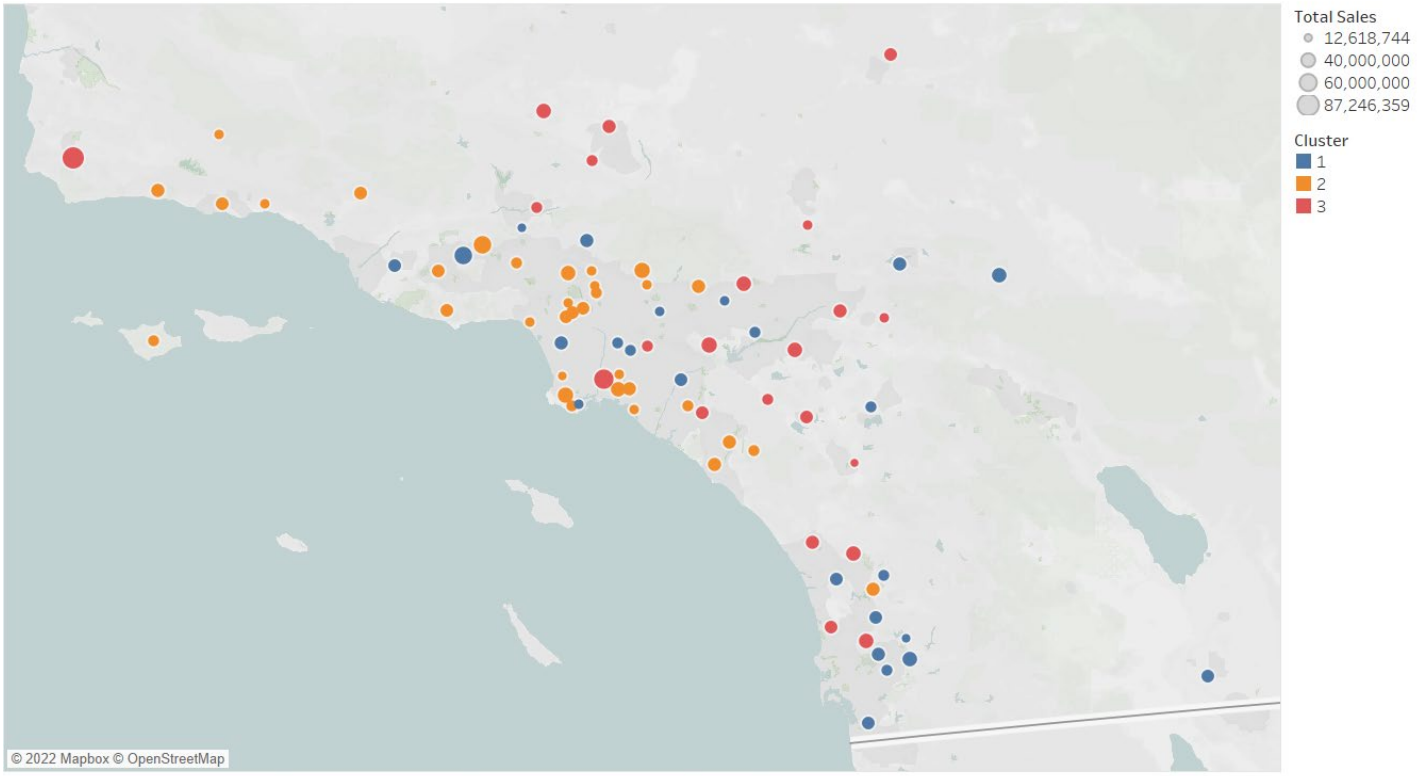
The plot attached also shows that Cluster 2 purchase more of Frozen Food, Dairy, Floral, Produce, and Bakery products. Cluster 3 is more tuned towards Dry Grocery and General Merchandise while Cluster 1 is tuned towards Meat and Deli.

Kwame Odoi Otchere



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.
   The visualization can be accessed here:
   https://public.tableau.com/views/UdacityPANDCapstoneProjectTask1-ClustertheStores/Sheet1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Location Distribution of Clustered Stores



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

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.)
   *I used the **Forest Model** to predict the best store format for the new stores. This is because, per the captures below, despite tying on overall accuracy with the Boosted Model, it had a higher F1 score than the Boosted Model.*

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Forest_Model	0.7059	0.7917	0.3750	1.0000	1.0000

Confusion matrix of Boosted_Model

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

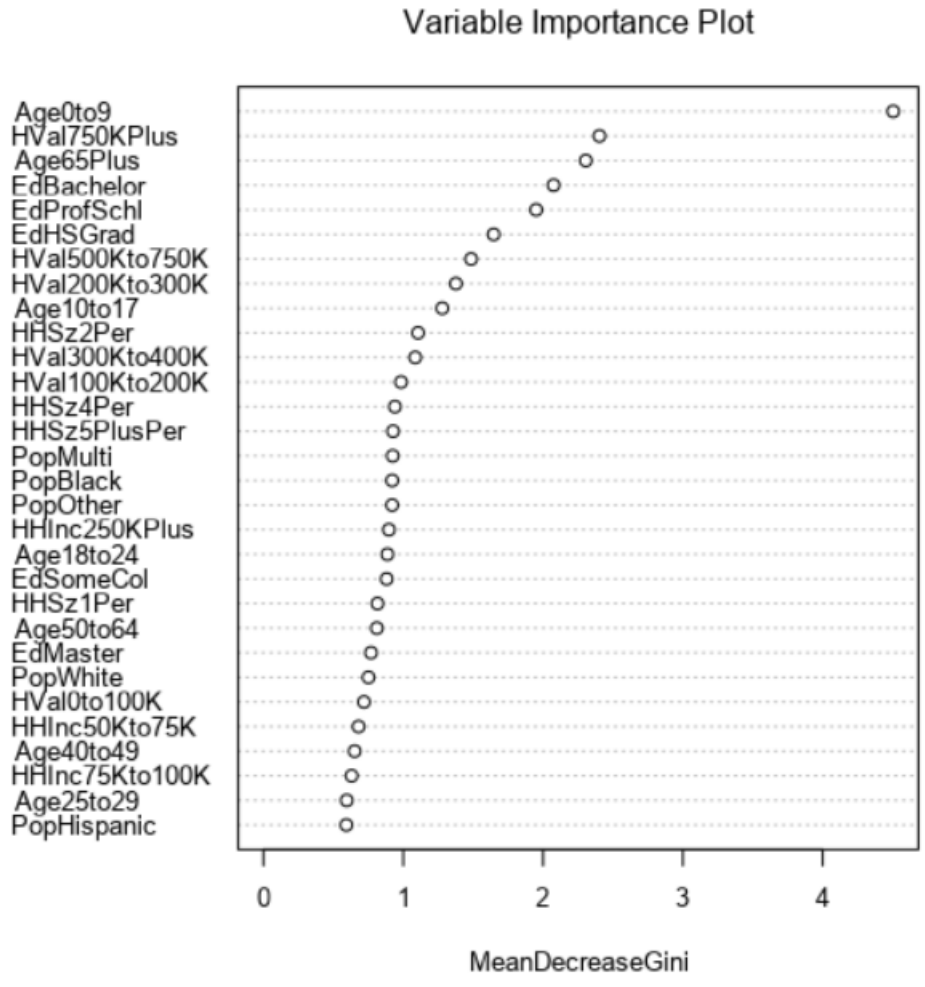
Confusion matrix of Decision_Tree

	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	3	0	0
Predicted_2	3	5	0
Predicted_3	2	0	4

*The most important variables that help explain the relationship between demographic indicators and store formats are **Age, HV,** and **Education**, in that order as visualized in the Variable Importance Plot below:*



Kwame Odoi Otchere

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	1
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

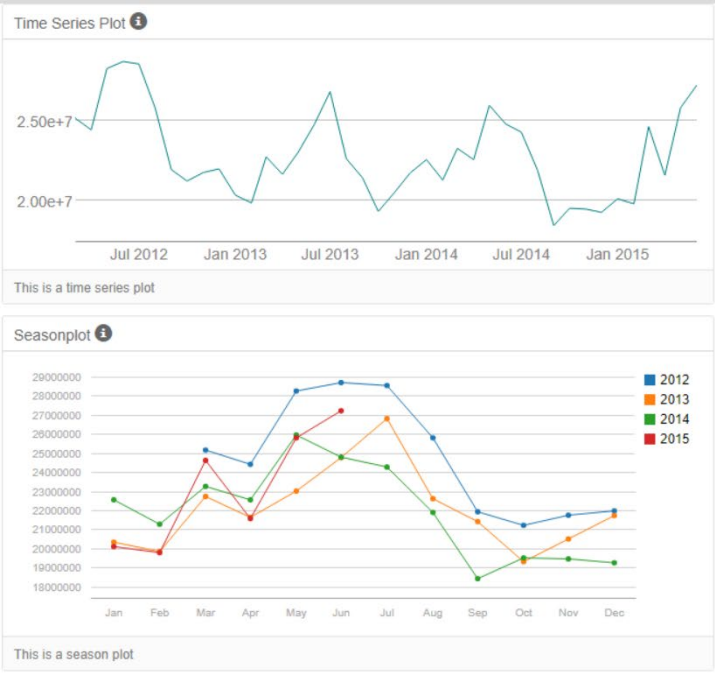
Task 3: Predicting Produce Sales

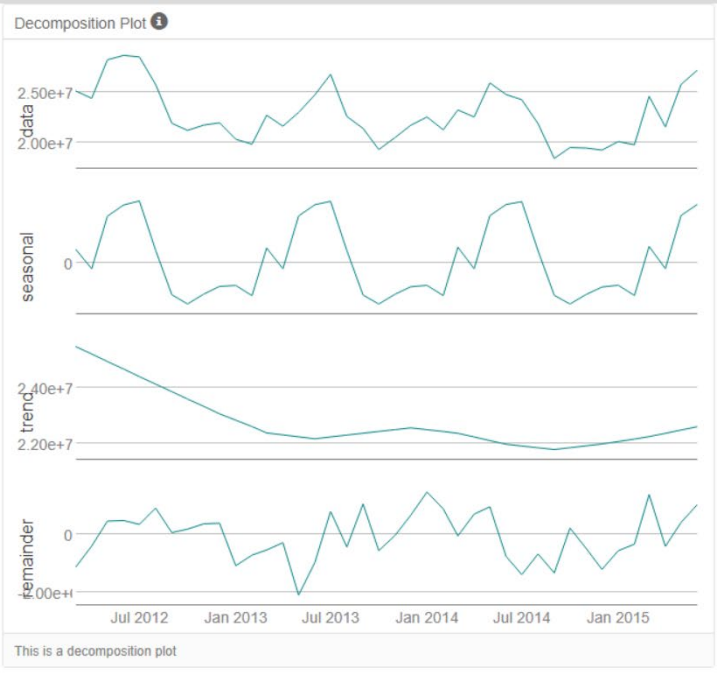
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?
I used an ETS (M,N,M) model was used since it had lower RMSE and MASE values versus the ARIMA model as seen below.

Accuracy Measures:

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

Also, per the TS Plot image extract below, it may be seen that the error is fluctuating (indicating a multiplicative E-component), the trend over the period is unclear (indicating an N T-component), and the peaks and valleys (seasonality) seem to flow with time (indicating an M S-component). These components were used accordingly in the forecasting.

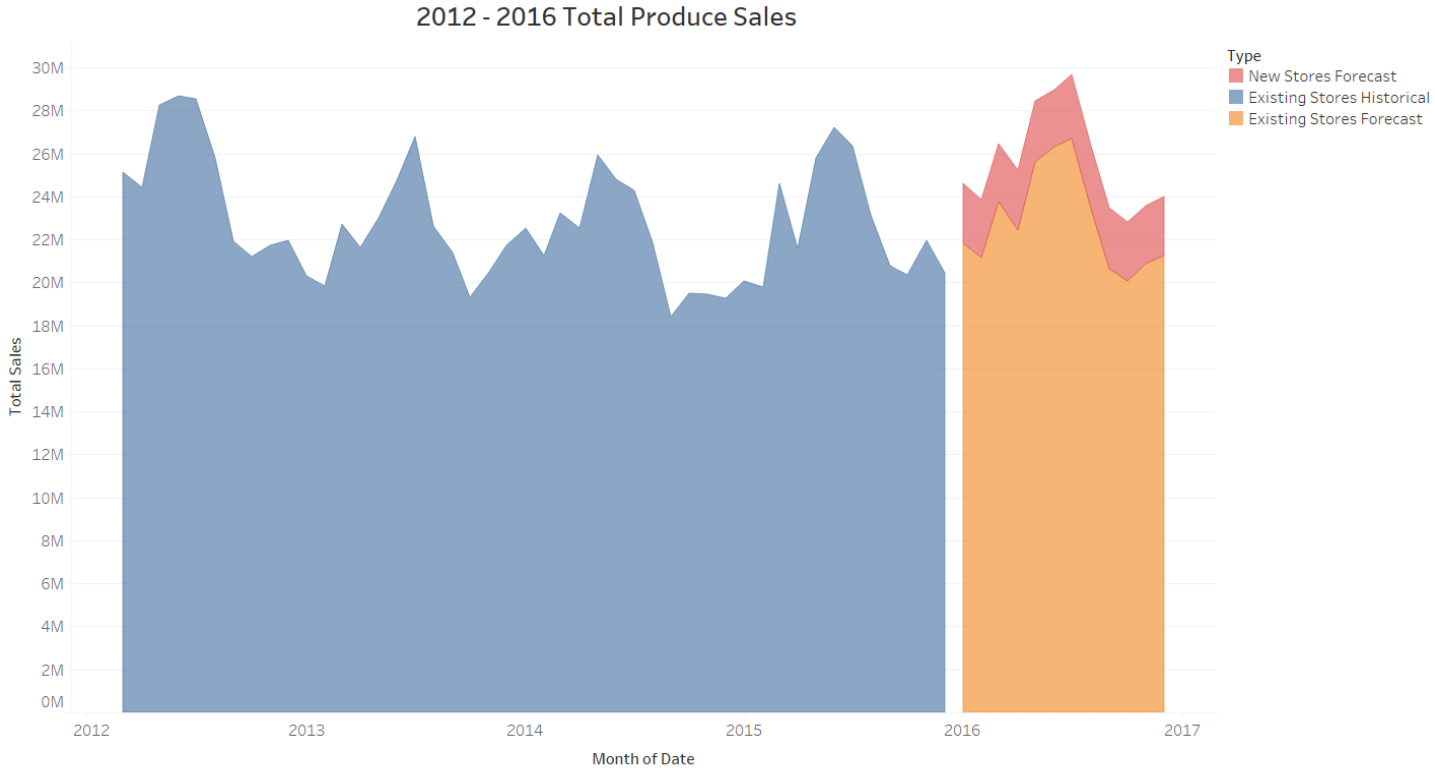




3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Month	Existing Stores	New Stores
Jan-16	21,829,060.03	2,786,991.73
Feb-16	21,146,329.63	2,696,833.29
Mar-16	23,735,686.94	2,719,678.80
Apr-16	22,409,515.28	2,829,740.79
May-16	25,621,828.73	2,832,644.31
Jun-16	26,307,858.04	2,669,943.17
Jul-16	26,705,092.56	2,953,317.28
Aug-16	23,440,761.33	2,915,367.23
Sep-16	20,640,047.32	2,821,053.00
Oct-16	20,086,270.46	2,727,588.82
Nov-16	20,858,119.96	2,708,276.94
Dec-16	21,255,190.24	2,718,828.40

The visualized historical data and forecasted data can be seen here: https://public.tableau.com/views/UdacityPANDCapstoneTask3-VisualizetheHistoricalProduceSalesandForecasts/TotalProduceSalesForecasts?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link



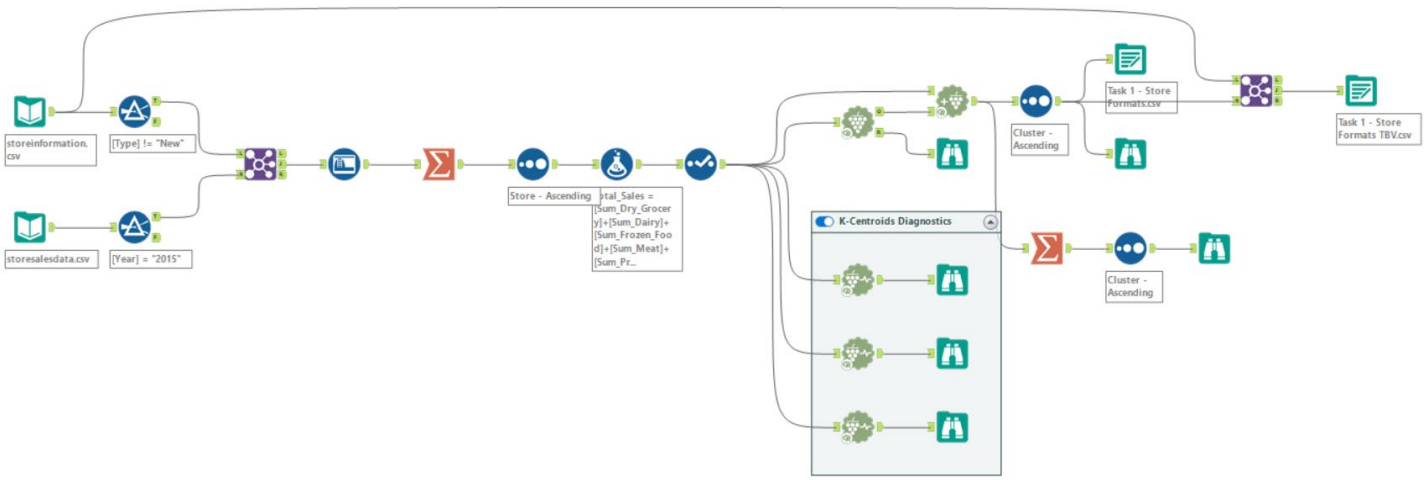
The plot of sum of Total Sales for Date Month. Color shows details about Type.

Kwame Odoi Otchere

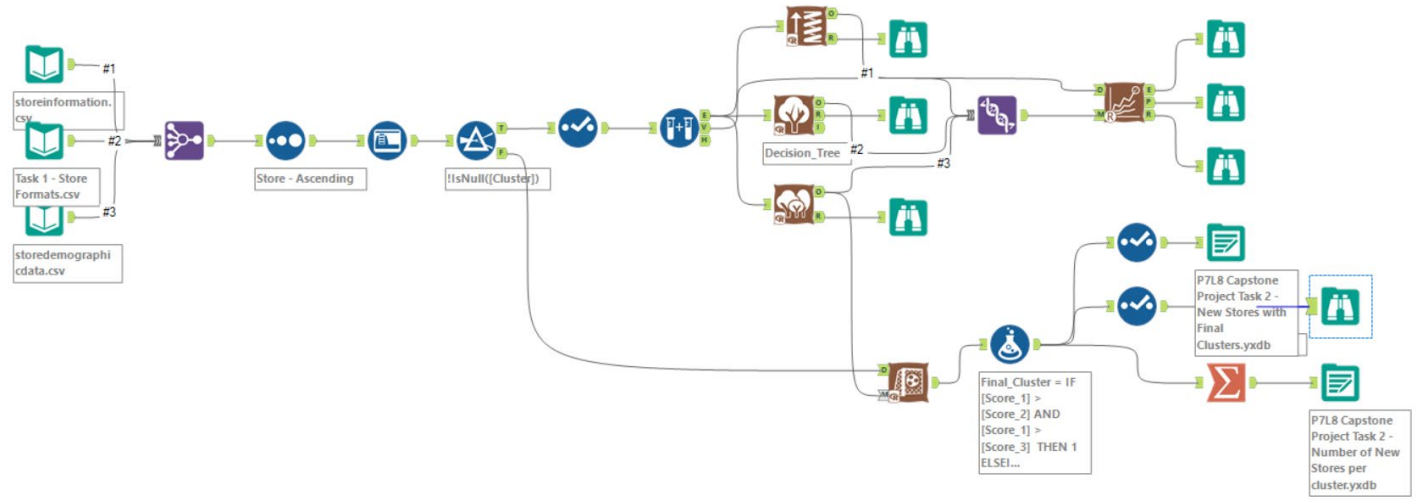
Workflows

Below are visual representations of the workflows for the various tasks:

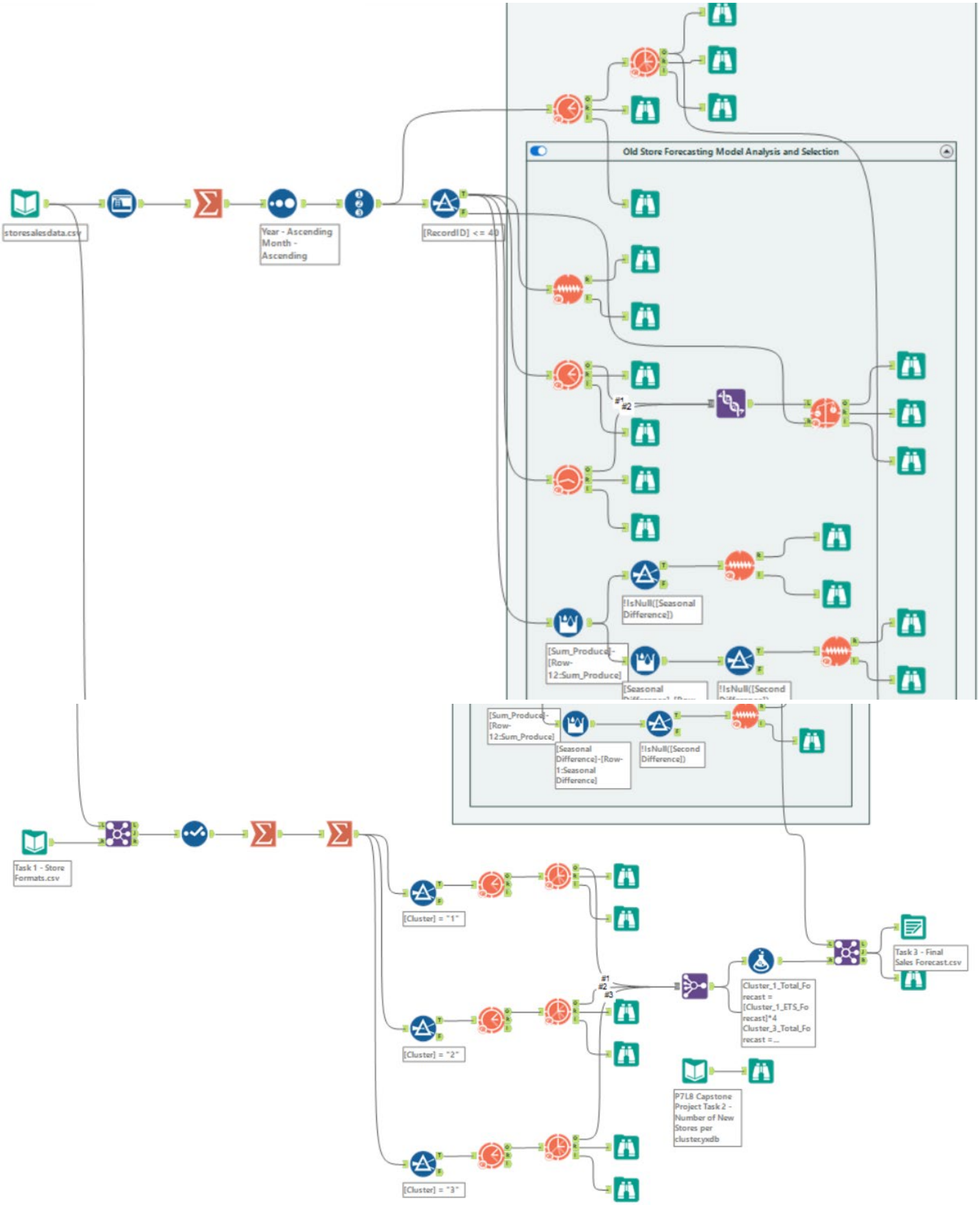
1. Task 1: Cluster the Existing Stores

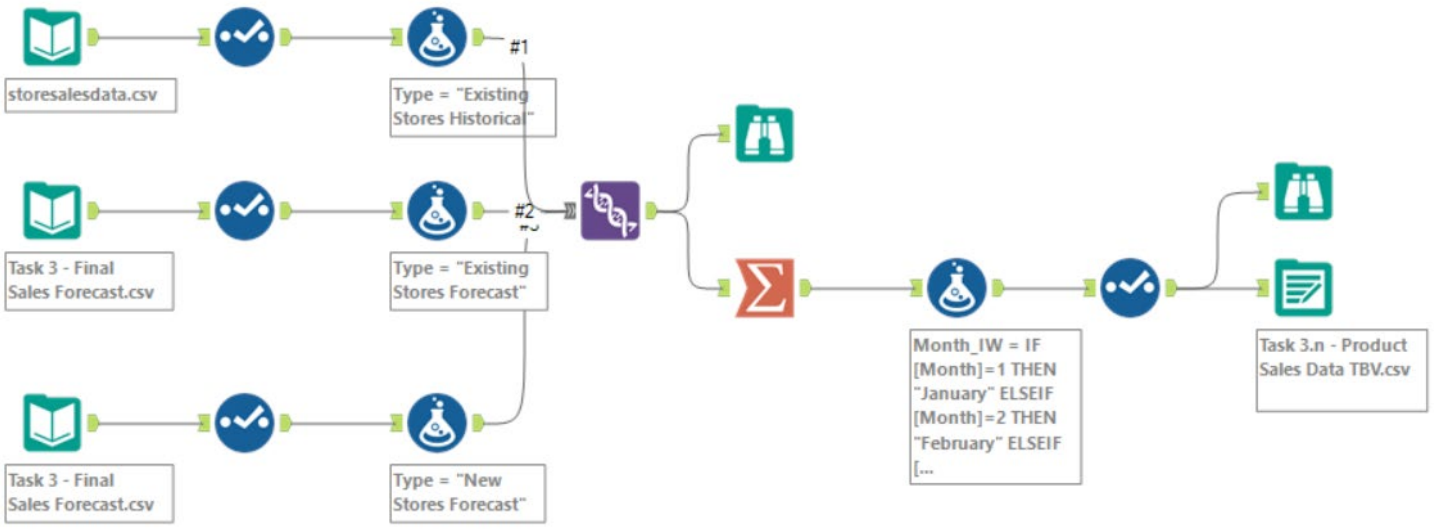


2. Task 2: Cluster the New Stores



3. Task 3: Forecasting




storesalesdata.csv
Type = "Existing Stores Historical"
#1
Task 3 - Final Sales Forecast.csv
Type = "Existing Stores Forecast"
#2
Task 3 - Final Sales Forecast.csv
Type = "New Stores Forecast"
Month_IW = IF [Month]=1 THEN "January" ELSEIF [Month]=2 THEN "February" ELSEIF [...
Task 3.n - Product Sales Data TBV.csv

References

1. Reviewer’s Notes and Recommendations

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.