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

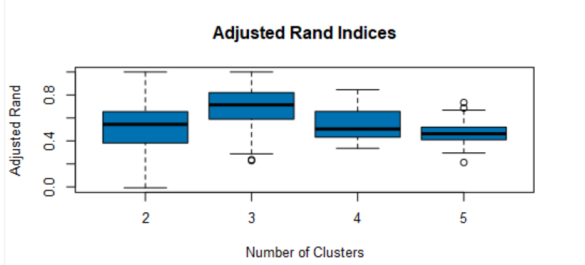
Task 1: Determine Store Formats for Existing Stores

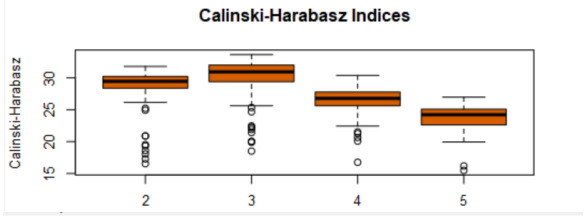
1. What is the optimal number of store formats? How did you arrive at that number?

   I worked out the percentage sales data per category per store and joined it with the rest of the store information. I then filtered for 2015 and used the K-centroids diagnostic tool to see how many clusters would lead to the optimal result, checking all three methods – K-mean, K-median, Neural Gas.

   The optimal number of store formats was 3 as shown by the K-means cluster report. Although the neural gas had a marginally higher mean for the 3 clusters adjusted rand indices (similarity) and CH indices (compactness and distinctness), the K-mean data was a lot more compact, making it the better model overall.

   K-means:





Adjusted Rand Indices:

	2	3
Minimum	-0.007972	0.228156
1st Quartile	0.381421	0.593906
Median	0.544002	0.713886
Mean	0.503774	0.688385
3rd Quartile	0.654956	0.820181
Maximum	1	1

Calinski-Harabasz Indices:

	2	3
Minimum	16.543	18.50776
1st Quartile	28.42185	29.41372
Median	29.45069	30.93662
Mean	28.56415	29.83325
3rd Quartile	30.21413	31.97449
Maximum	31.78345	33.63781

   Neural Gas:

Adjusted Rand Indices:

	2	3
Minimum	-0.007426	0.295256
1st Quartile	0.419539	0.546826
Median	0.544134	0.730805
Mean	0.535018	0.689074
3rd Quartile	0.724383	0.840406
Maximum	0.952941	0.958284

Calinski-Harabasz Indices:

	2	3
Minimum	17.16983	20.69073
1st Quartile	28.3229	28.23975
Median	29.5423	30.85988
Mean	28.60545	29.85938
3rd Quartile	30.48713	32.27007
Maximum	31.87048	33.62834

K-median:

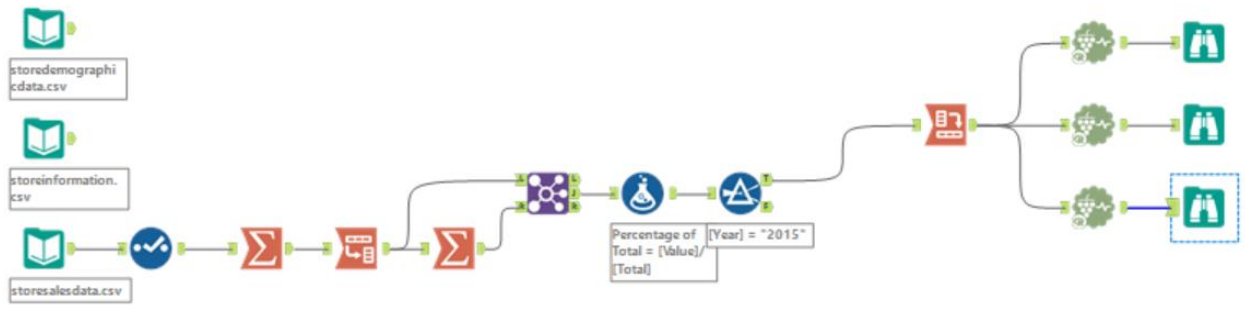
Adjusted Rand Indices:

	2	3
Minimum	-0.009376	0.266027
1st Quartile	0.484044	0.455147
Median	0.635849	0.637279
Mean	0.578859	0.632665
3rd Quartile	0.73445	0.78289
Maximum	0.907005	0.95859

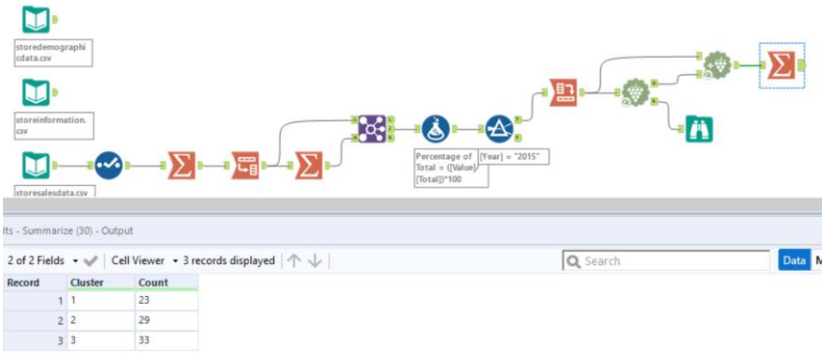
Calinski-Harabasz Indices:

	2	3
Minimum	14.38085	15.34501
1st Quartile	26.00385	25.73566
Median	27.67012	29.33659
Mean	27.26555	27.82628
3rd Quartile	29.47139	31.07688
Maximum	32.30654	33.56494

Workflow



2. How many stores fall into each store format?
   Cluster 1 – 23
   Cluster 2 – 29
   Cluster 3 – 33

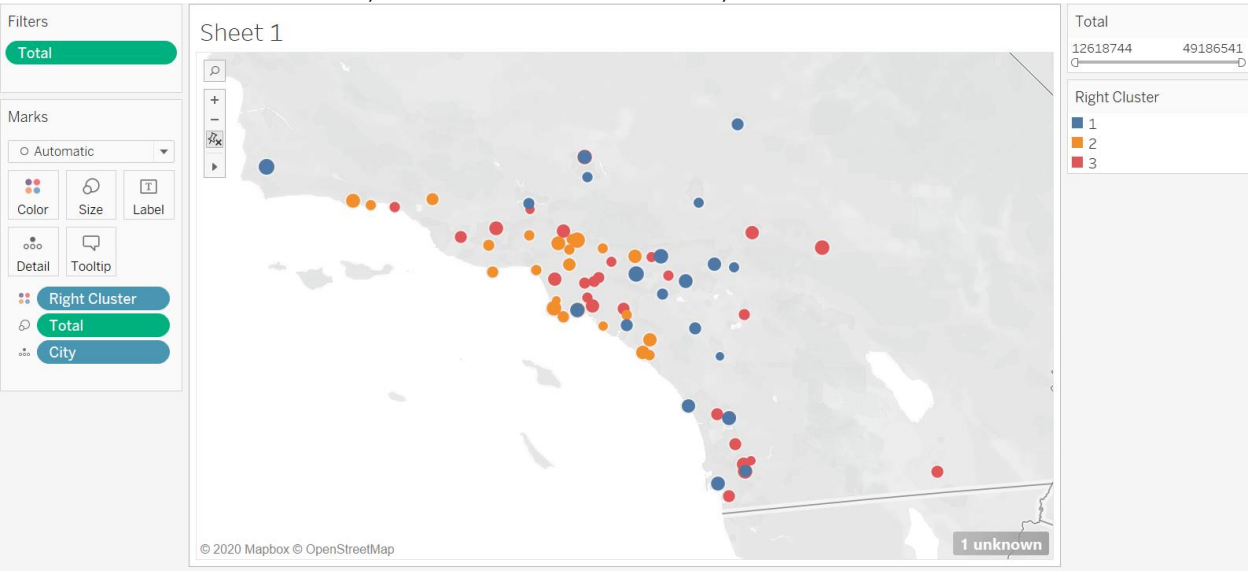


3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   The average total sales between the clusters differ. Cluster 1 has the highest average sales, followed by cluster 3 then cluster 2. This could indicate three different sized stores, on average.

Cluster	Avg_Total	Count
1	32253841.90087	23
2	27472964.449655	29
3	28356954.955758	33

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/Tableaumap_finalproject/Sheet1?:language=en-GB&:display_count=y&publish=yes&:origin=viz_share_link

Task 2: Formats for New Stores

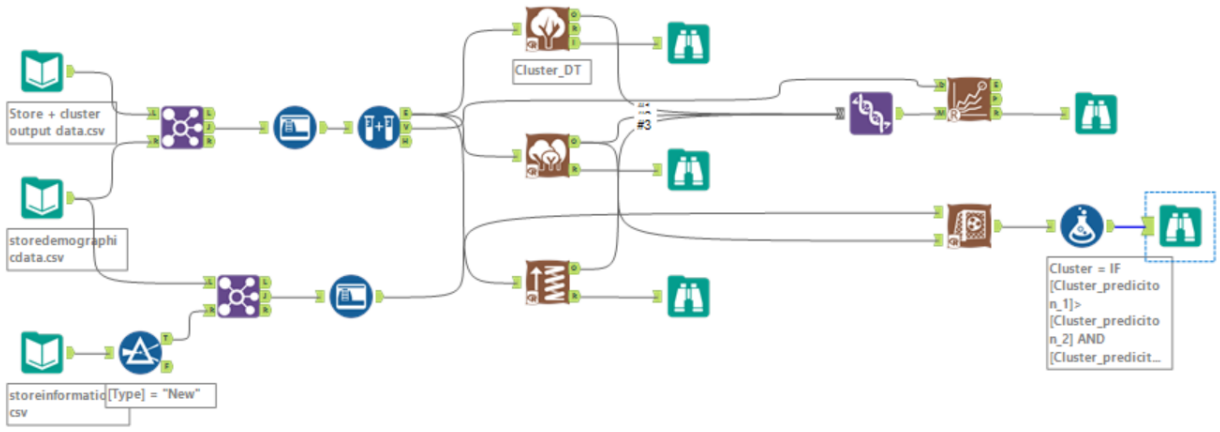
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Cluster_forest	0.8235	0.8426	0.7500	1.0000	0.7778
Cluster_boosted	0.8235	0.8889	1.0000	1.0000	0.6667
Cluster_DT	0.8235	0.8426	0.7500	1.0000	0.7778

I compared the decision tree, random forest model and boosted models.
I went with the boosted model as it had the highest F1 score, which shows the precision of the model. Also, the accuracy for each cluster segment was higher (overall) than the other two models.

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

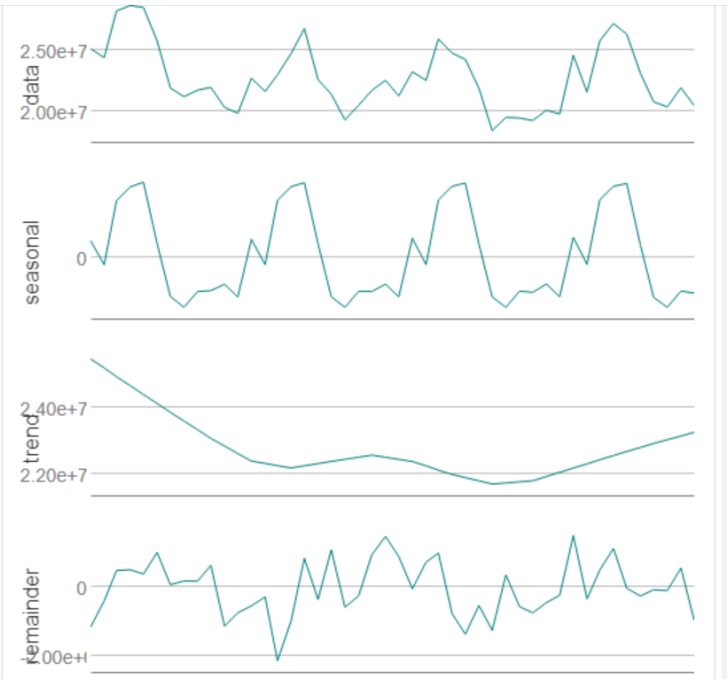
Store Number	Segment
S0086	3
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2



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) as we can see from the decomposition graphs that there is no trend, and seasonality and error are multiplicative.
I set the ARIMA model to auto.

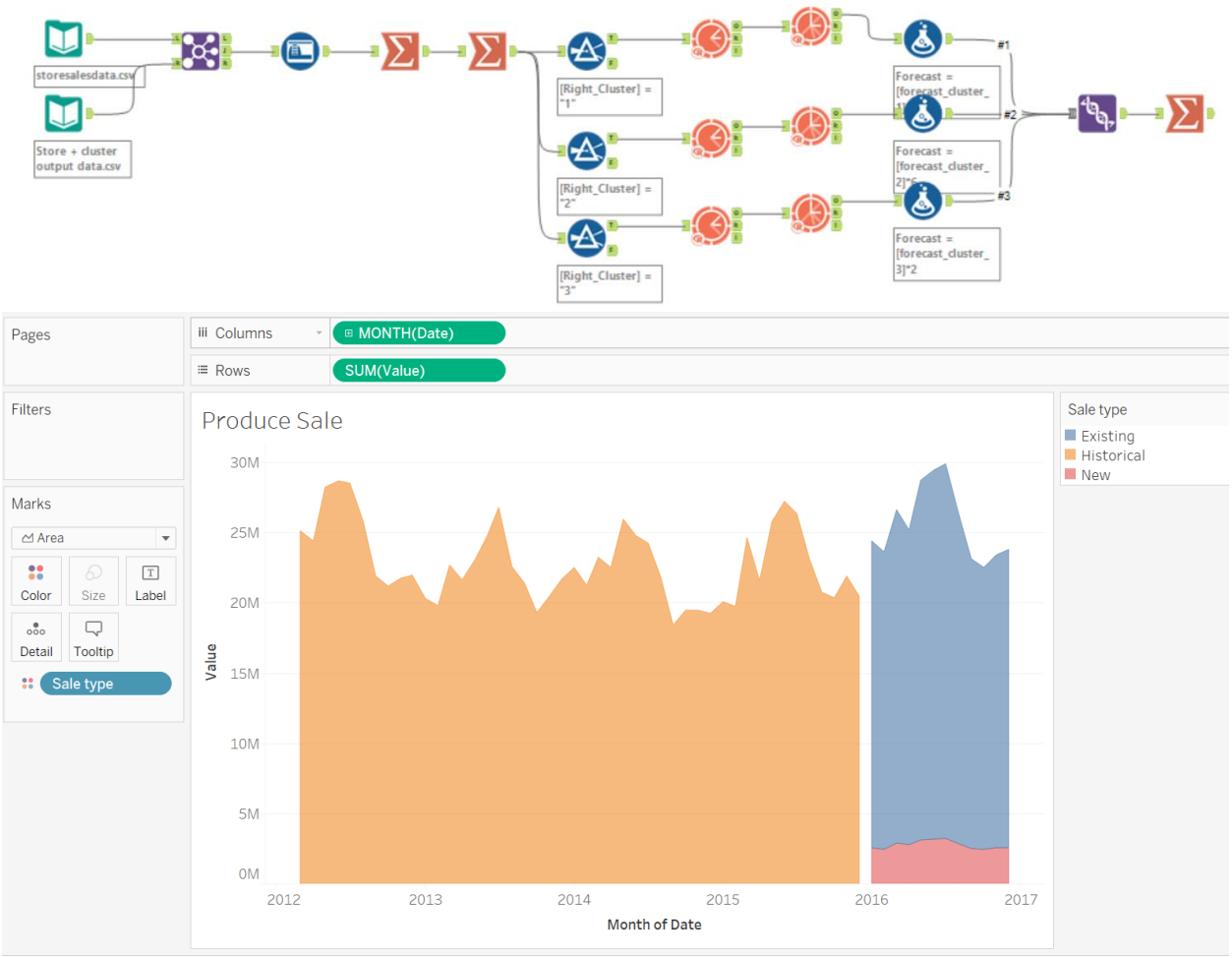


Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_forecast	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463
ETS_forecast	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

As we can see from the accuracy measures, RSMA, MAPE and MASE are lower for ETS than for ARIMA, indicating is it a stronger 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 stores forecasts.

Year	Month	New	Existing
2016	1	2558242	21829060
2016	2	2468197	21146330
2016	3	2883620	23735687
2016	4	2762836	22409515
2016	5	3129542	25621829
2016	6	3170534	26307858
2016	7	3199198	26705093
2016	8	2842411	23440761
2016	9	2512051	20640047
2016	10	2460445	20086270
2016	11	2555392	20858120
2016	12	2534378	21255190



https://public.tableau.com/views/Producesalegraphs/Sheet2?:language=en-GB&:display_count=y&publish=yes&:origin=viz_share_link

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.