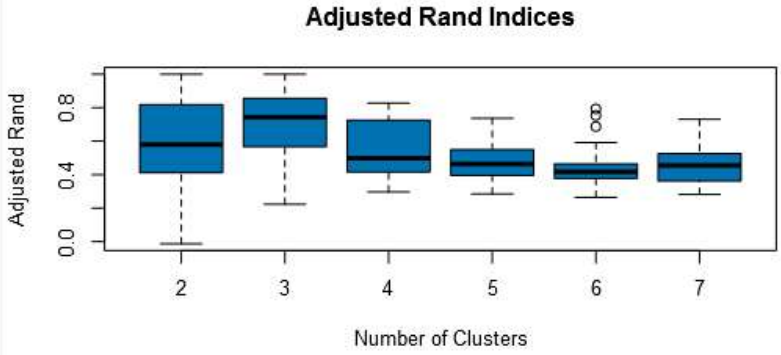
Kacper Ksieski

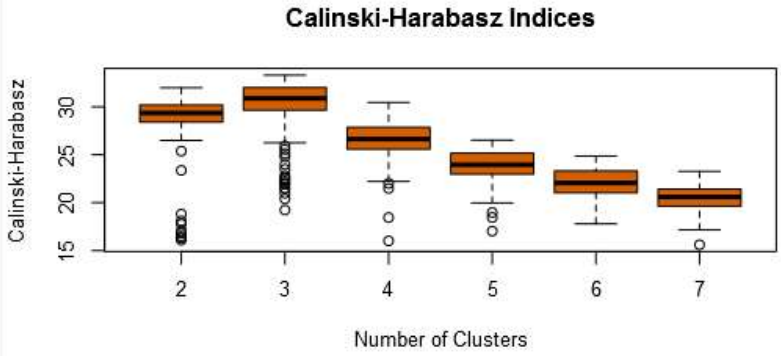
Project: Predictive Analytics Capstone

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 three. I used a K-Means Clustering method. The results, depicted below, show that three is the ideal number of clusters.





2. How many stores fall into each store format?
   The number of stores per format are displayed below.

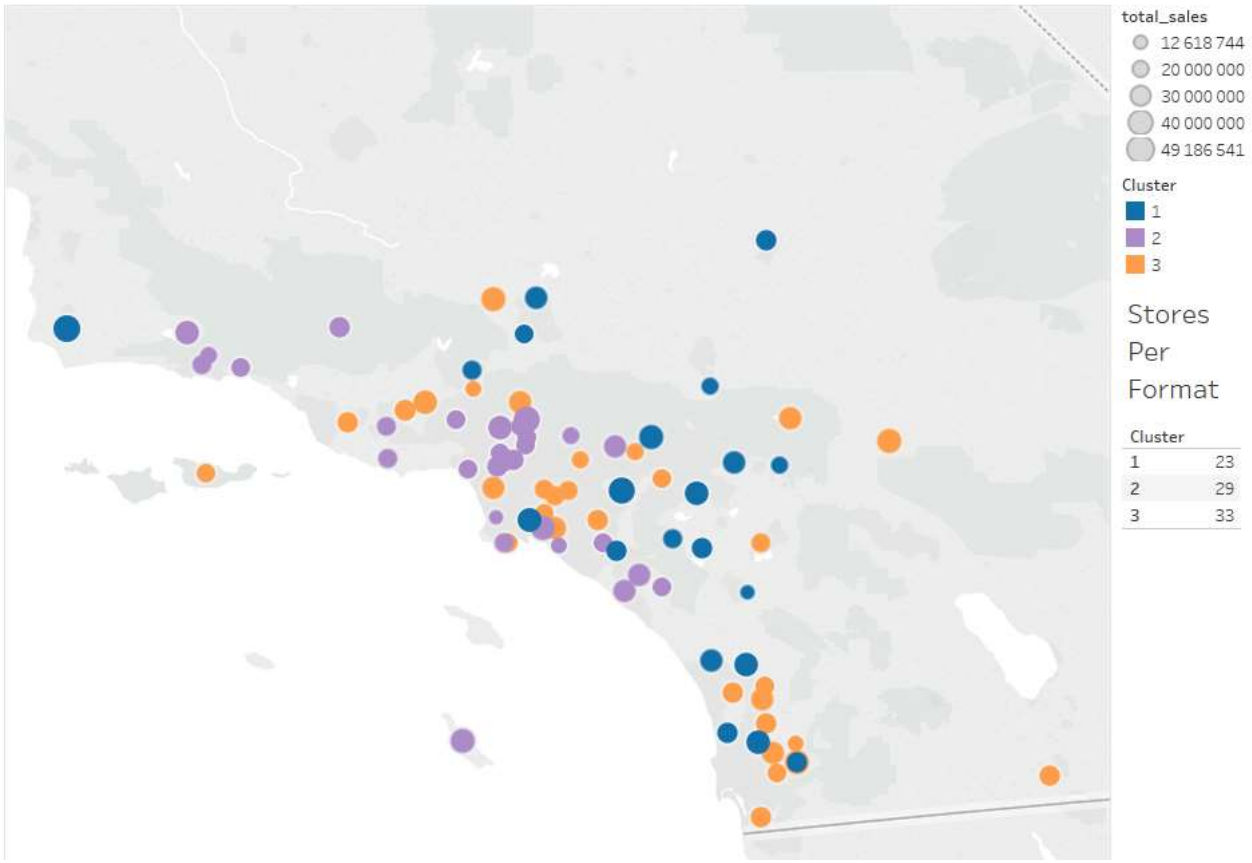
Cluster	Count
1	23
2	29
3	33

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   As depicted below, the clusters differ in Floral sales percent of total. After multiplying by the average store sales, the Floral dollar sales differ widely among the clusters.

Cluster	Floral_pct_of_total	Avg_store_sales	Avg_floral_sales
1	0.007573	32253841.90	244272.97
2	0.010486	2747964.45	288072.19
3	0.006941	28356954.96	196828.43

Kacper Ksieski

4. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template.



Kacper Ksieski

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 Boosted Model. As depicted in the screenshot below, while the accuracy of the three models is equivalent, the F1 measure of the Boosted Model is superior.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
TREE	0.8235	0.8251	0.7500	0.8000	0.8750
FOREST	0.8235	0.8251	0.7500	0.8000	0.8750
BOOST	0.8235	0.8543	0.8000	0.6667	1.0000

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	1
S0092	2
S0093	1
S0094	2
S0095	2

Kacper Ksieski

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 chose the ETS(M,N,M) model because most of the in-sample error measurements, depicted below, are superior to the ARIMA model. The more important measures, RMSE and MASE, are smaller and therefore better in the ETS model.

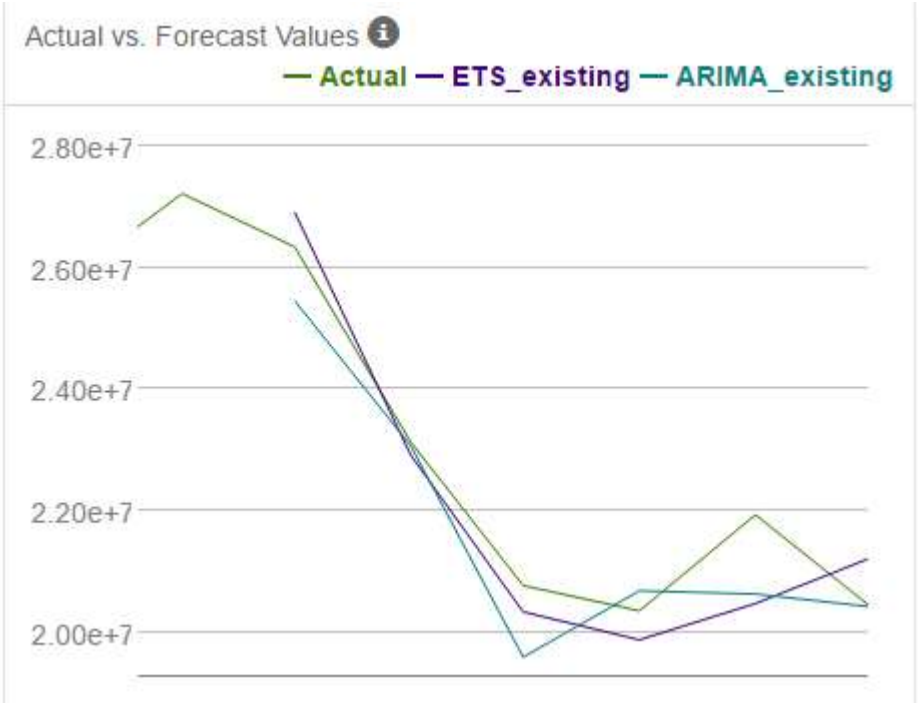
Name	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	AIC
ARIMA	166650.6641392	1430610.6110795	934256.056469	0.6051296	4.1163464	0.5215287	-0.0019723	856.8308
ETS	-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788	1283.1197
betterModel	ETS	ETS	ETS	ETS	ETS	ETS	ARIMA	ARIMA

   The accuracy measures back up the decision. Where the ETS model is better, it is much better, especially at the RMSE measurement. Where ARIMA is better, it is only slightly better, especially at the MASE measurement.

Name	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	520597	813457.4	630163.2	2.2909	2.8291	0.3708
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
betterModel	ETS	ETS	ARIMA	ETS	ARIMA	ARIMA

   Finally, the forecasts show, against the holdout sample of six months, the ETS model makes extremely accurate predictions. This is shown in both the table and the chart below.

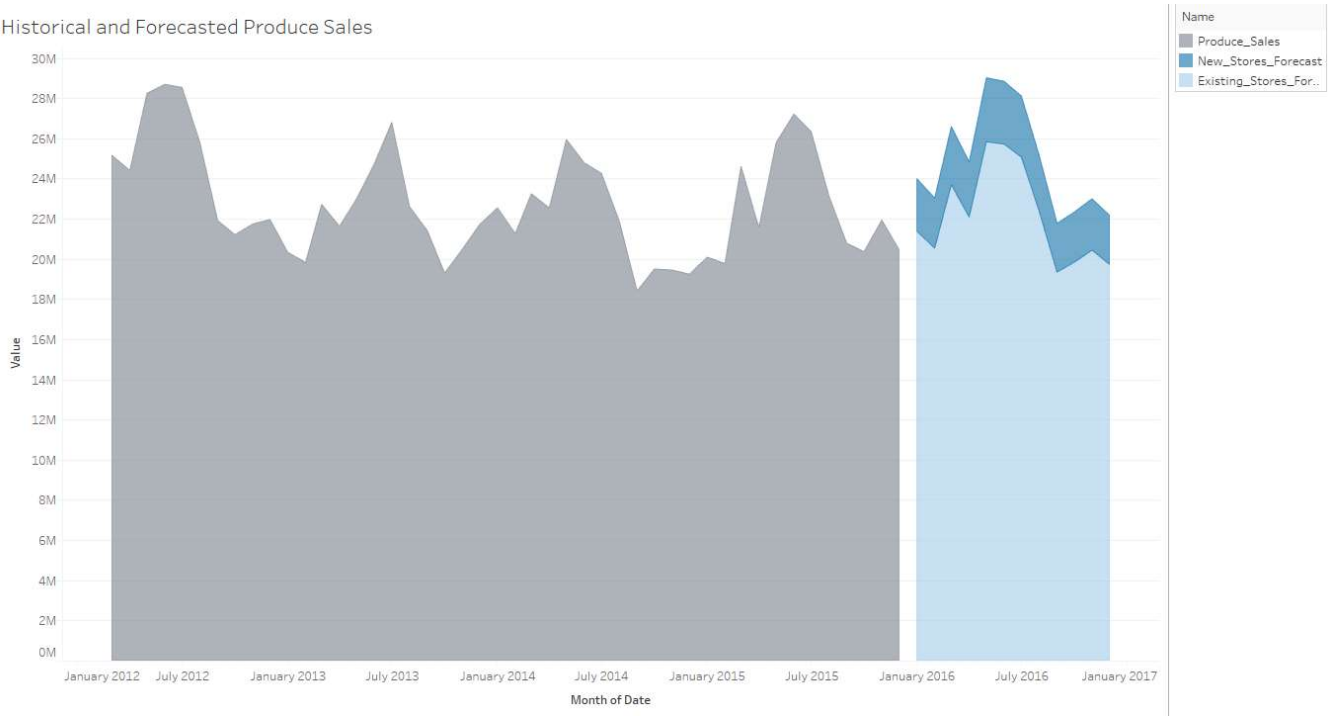
Actual	ETS_existing	ARIMA_existing	ETS_Abs_Diff	ETS_Rel_Diff	ARIMA_Abs_Diff	ARIMA_Rel_Diff	Better_Model
26338477.15	26907095.61191	25454225.03787	-568618.46191	-0.021589	884252.11213	0.033573	ETS
23130626.6	22916903.07434	23071096.30787	213723.52566	0.00924	59530.29213	0.002574	ARIMA
20774415.93	20342618.32222	19598371.02787	431797.60778	0.020785	1176044.90213	0.05661	ETS
20359980.58	19883092.31778	20688679.39787	476888.26222	0.023423	-328698.81787	-0.016144	ARIMA
21936906.81	20479210.4317	20635860.61787	1457696.3783	0.066449	1301046.19213	0.059309	ARIMA
20462899.3	21211420.14022	20431492.19787	-748520.84022	-0.036579	31407.10213	0.001535	ARIMA



Kacper Ksieski

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.
   I have depicted below a table of produce forecasts for new stores, existing stores, and the total for each month in 2016. I have produced a stacked area chart below which shows historical produce sales with forecasts for existing and new store produce sales.

Year_Month	New_Stores_Forecast	Existing_Stores_Forecast	Total_Produce_Forecast
2016-01	2 623 914	21 370 818	23 994 731
2016-02	2 509 815	20 525 731	23 035 546
2016-03	2 904 798	23 684 288	26 589 086
2016-04	2 756 237	22 073 944	24 830 181
2016-05	3 192 380	25 826 610	29 018 990
2016-06	3 130 219	25 708 731	28 838 950
2016-07	3 047 901	25 059 365	28 107 266
2016-08	2 785 938	22 355 893	25 141 831
2016-09	2 436 863	19 333 714	21 770 577
2016-10	2 504 158	19 829 131	22 333 288
2016-11	2 567 039	20 428 496	22 995 534
2016-12	2 459 791	19 720 851	22 180 642



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.