Project: Predictive Analytics Capstone

Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. You've been asked to provide analytical support to make decisions about store formats and inventory planning.

Data provided:

- StoreSalesData.csv - This file contains sales by product category for all existing stores for 2012, 2013, and 2014.
- StoreInformation.csv - This file contains location data for each of the stores.
- StoreDemographicData.csv - This file contains demographic data for the areas surrounding each of the existing stores and locations for new stores.

Task 1: Determine Store Formats for Existing Stores

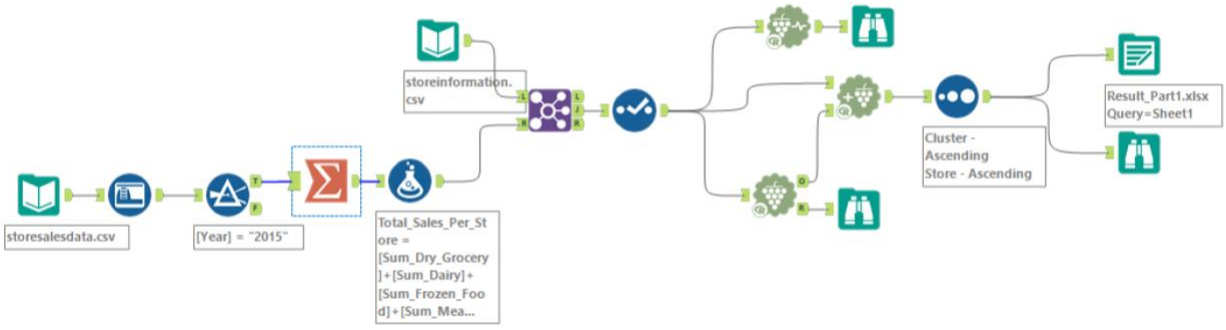


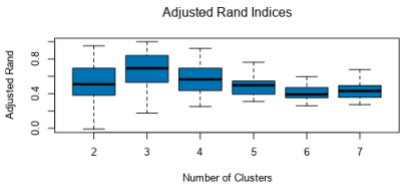
Figure 1. Model for Task 1 – K-Centroid Diagnostics and Cluster Analysis

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

The store sales dataset was filtered to year 2015 and the percentage sales per category per store was calculated for clustering purposes with K-means model (see figure 1 above). The K-centroid diagnostics shows that the three clusters is the most optimal method, as it has the highest median in both Adjusted Rand and Calinski-Harabasz Indices (see Figure 2 red boxes below). Therefore, I have chosen the three cluster method.

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	-0.008832	0.174002	0.251874	0.309521	0.260212	0.272005
1st Quartile	0.381396	0.545584	0.447519	0.393734	0.352747	0.358507
Median	0.50924	0.693307	0.564848	0.49 938	0.391986	0.430535
Mean	0.488908	0.67359	0.568496	0.489115	0.416063	0.426302
3rd Quartile	0.684345	0.836085	0.686873	0.545867	0.468145	0.492657
Maximum	0.952941	1	0.923028	0.763137	0.59593	0.678856



Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	17.09341	13.97288	18.34333	18.6761	16.47398	16.69673
1st Quartile	28.10155	28.49963	25.12113	22.8092	20.55176	19.61401
Median	29.61747	30.83944	26.46997	24.00 6	21.92928	20.47291
Mean	28.63788	29.34263	26.17876	23.73643	21.64855	20.42256
3rd Quartile	30.4766	31.91944	27.64665	24.95633	22.85341	21.30088
Maximum	32.07305	33.47176	29.76902	26.32012	24.73009	23.06992

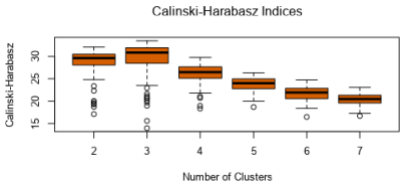


Figure 2. Adjused Rand and Calinski-Harabasz Indices

2. How many stores fall into each store format?

The K-Means cluster analysis shows that there are 25 stores in cluster 1, 35 stores in cluster 3 and 25 stores in cluster 3. The cluster results were saved for Task 2 of the project.

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

Looking at Figure 4 below, higher number means that product category sells better in a cluster, e.g. Meat and Deli sells the best (see Figure 4 green boxes) in Cluster 1, but Floral, Produce and General Merchandise sells the worst (see Figure 4 red boxes) in the same cluster. These are the topmost driver products for each Segment: Cluster 1 – Meat and Deli, Cluster 2 – Dairy, Floral and Produce, Cluster 3 – General Merchandise

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.528249	-0.215879	-0.261597	0.61414	-0.655027	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178481

	Percent_Bakery	Percent_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Figure 4. K-Means Cluster Analysis

4. Tableau visualization that shows the location of the stores, uses color to show cluster, and size to show total sales. Tableau: https://public.tableau.com/app/profile/csaba.jakabos8229/viz/UD_proj/Sheet1

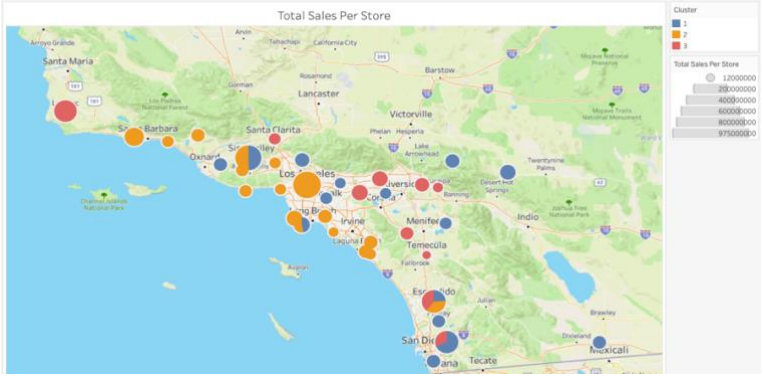


Figure 5. Tableau visualization of Clusters and Sales sizes, based on location

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?

Store demographic data together with the Cluster results from Task 1 have been used (Random seed 3 with 80% Estimation and 20% Validation samples) to setup different Prediction models (Decision Tree, Forest Model and Boosted model) and compare their results (see Figure 6). Both Forest Model and Boosted model has the same accuracy (see Figure 7), the latter has been chosen for further segment Scoring purposes of new Stores (see results in Table 1)

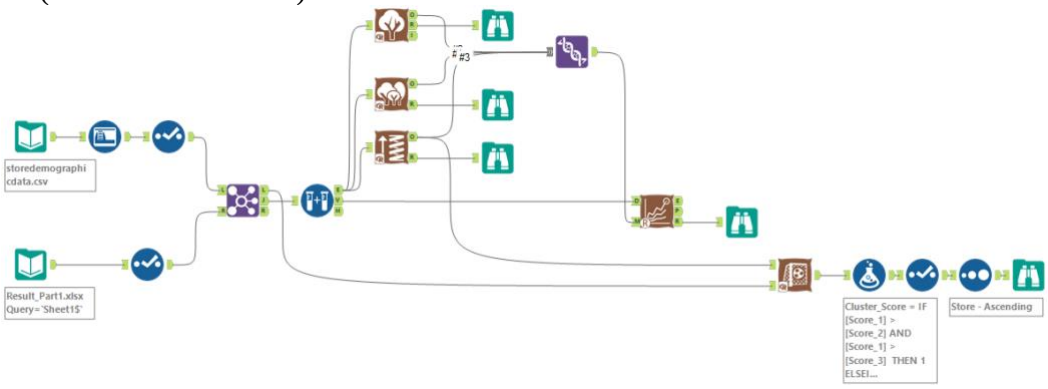


Figure 6. Task 2 – Predictive model comparison and Scoring

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

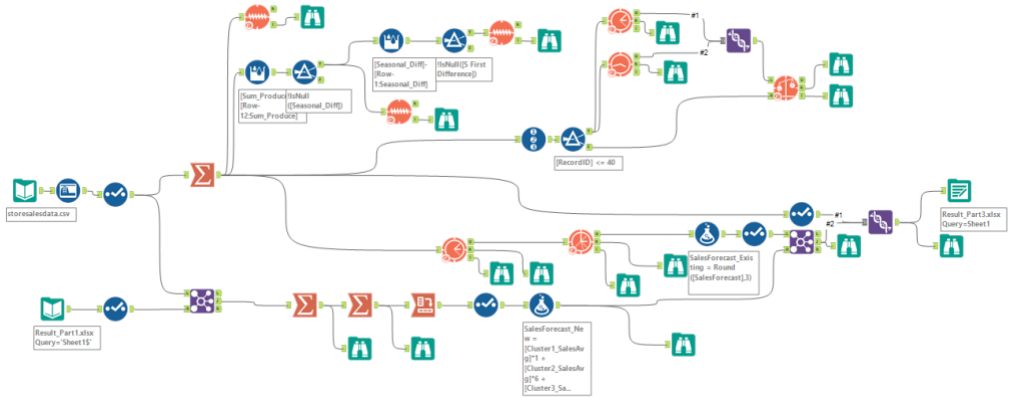
Figure 7. Accuracy comparison of different predictive models

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	3
S0092	2
S0093	3
S0094	2
S0095	2

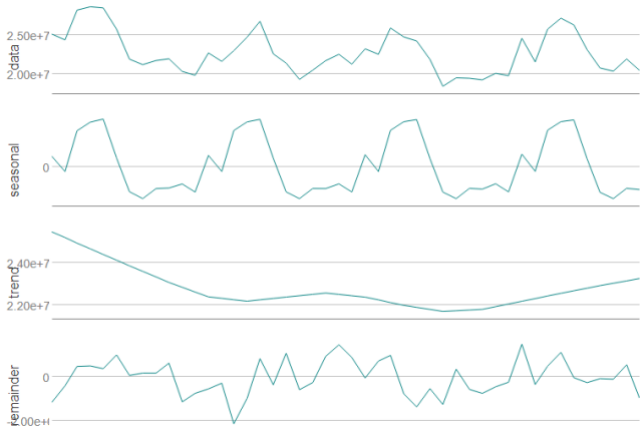
Table 1. Segment scoring of new Stores.

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?

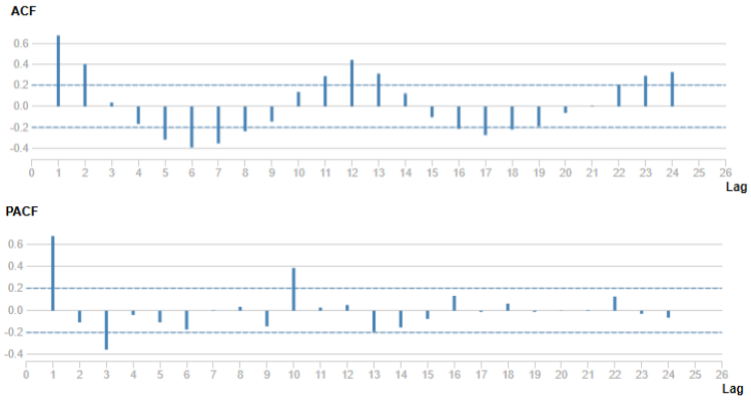
For deciding the setup of the forecasting models, a TS Plot (Timeseries analysis function) module's Decomposition plot has been used, which shows three components: seasonal, trend and remainder (error).



ETS model: Since the magnitude of seasonality does change over time, m – multiplicative is chosen. Since the trend line is almost linear, n – none is chosen. Since the error changing in magnitude over time, m – multiplicative is chosen. The overall ETS model is **ETS(M, N, M)**.

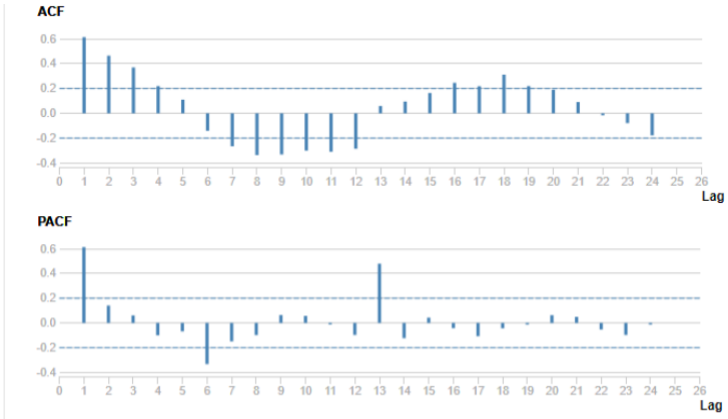
ARIMA model: Since the seasonal plot shows seasonality (it is periodic), therefore ARIMA (p, d, q) (P, D, Q)S model is used.

Time Series ACF and PACF



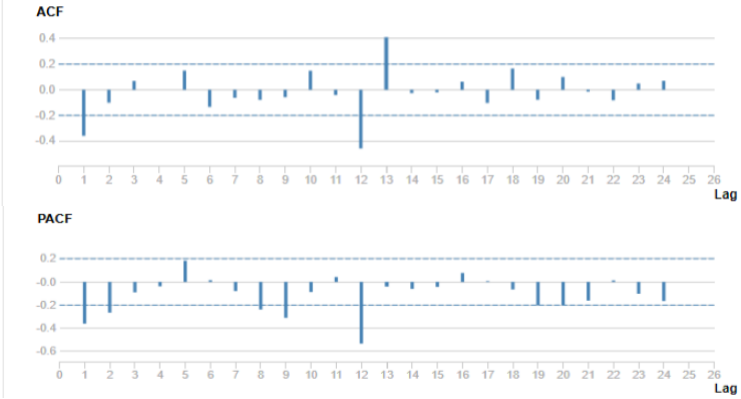
ACF plot shows decaying serial correlation towards 0 with increases at seasonal lags, since the serial correlation is high, seasonal difference is needed to be applied on the time series.

Seasonal Difference ACF and PACF



ACF plot shows again decaying serial correlation towards 0 with increases at seasonal lags, since the serial correlation is still moderate, further difference is needed to be applied on the seasonal difference series.

Seasonal First Difference ACF and PACF



The seasonal first differences of the series removed most of the significant lags, thus there is no need for further differencing. Since both seasonal difference and seasonal first difference needed to be applied, the differencing terms will be D = 1 (for seasonal difference) and d = 1 (for seasonal first difference).

The ACF plot at lag 1 shows strong negative correlation (same for lag 12, but that can be ignored due to seasonality), which is also confirmed by PACF plot. Since it is negative correlation at lag 1 and it decays to zero, so MA part is 1 and the AR part is 0 (Q = 1 and P = 0), to explain the seasonal autocorrelation.

Both ACF and PACF shows negative autocorrelations at lag 1 and lag 12, therefore q=1 is used to explain the non-seasonal autocorrection, thus MA part is 1 and AR part is 0 (q = 1 and p = 0),.

ARIMA(0, 1, 1)(0, 1, 1)[12]

Comparison of Arima and ETS errors
The ETS model has lower errors in all categories, thus better accuracy, so that model was chosen for further prediction of future sales.

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

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.

2016	Existing stores	New stores
January	21 829 059	2 574 250
February	21 146 331	2 483 373
March	23 735 688	2 972 113
April	22 409 514	2 823 371
May	25 621 830	3 197 153
June	26 307 858	3 249 052
July	26 705 094	3 259 756
August	23 440 761	2 904 986
September	20 640 048	2 574 827
October	20 086 269	2 506 604
November	20 858 121	2 601 483
December	21 255 189	2 559 789

Table 2. Sales prediction of existing and new stores.

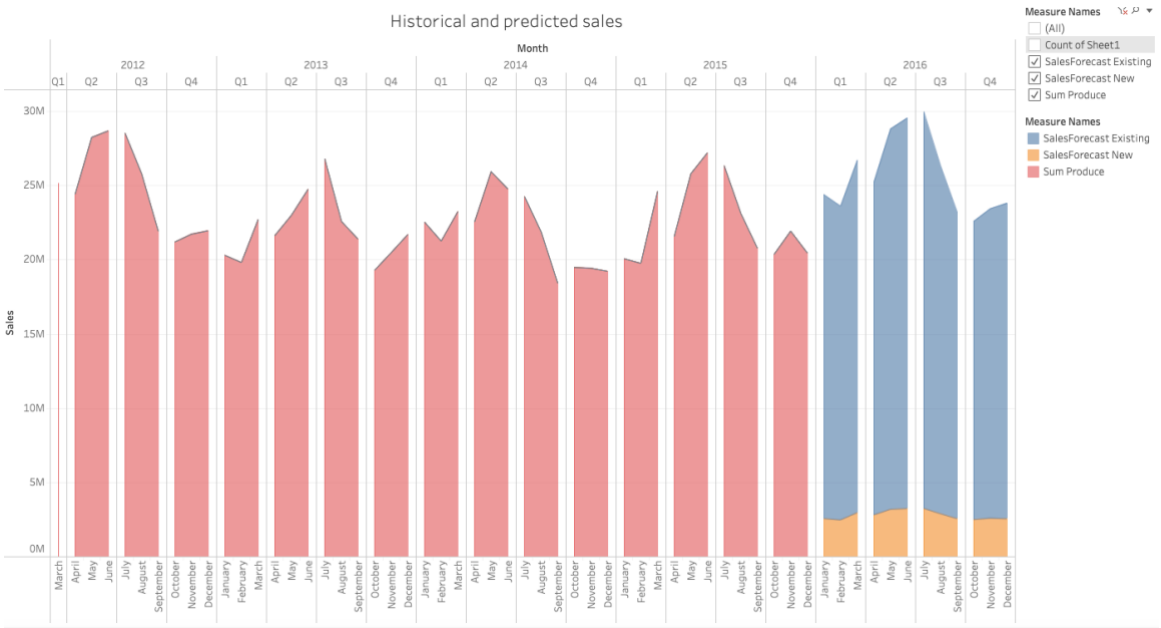


Tableau: https://public.tableau.com/app/profile/csaba.jakabos8229/viz/UD_proj_forecast/Sheet1