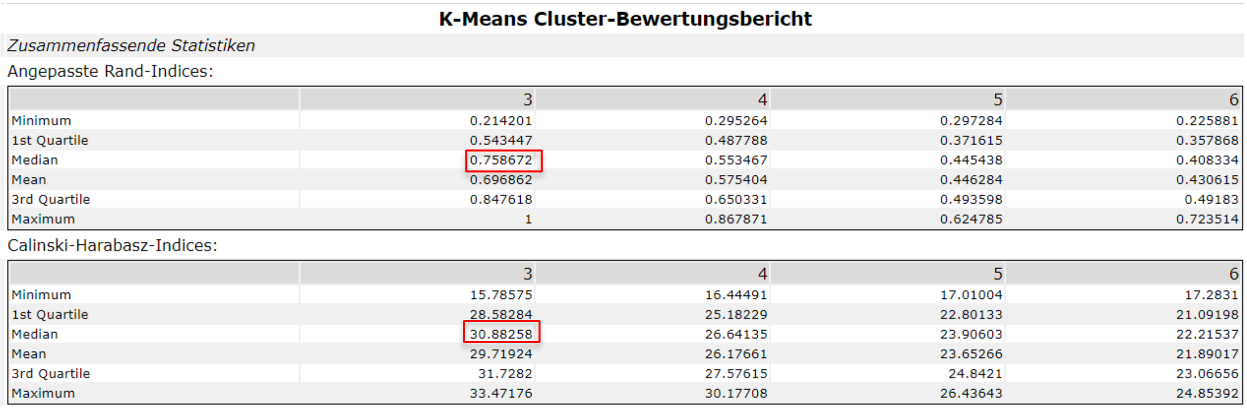
Project: Predictive Analytics Capstone

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?

Answer: Using the cluster analysis (K-means), the optimal number of store formats is 3. This can be seen from the high median value for adjusted Rand Indices (0.76) as well as the high median value for Calinski-Harabasz indices (30.88), which is both the highest value in comparison to the other values.



K-Means Cluster-Bewertungsbericht

Zusammenfassende Statistiken

Angepasste Rand-Indices:

	3	4	5	6
Minimum	0.214201	0.295264	0.297284	0.225881
1st Quartile	0.543447	0.487788	0.371615	0.357868
Median	0.758672	0.553467	0.445438	0.408334
Mean	0.696862	0.575404	0.446284	0.430615
3rd Quartile	0.847618	0.650331	0.493598	0.49183
Maximum	1	0.867871	0.624785	0.723514

Calinski-Harabasz-Indices:

	3	4	5	6
Minimum	15.78575	16.44491	17.01004	17.2831
1st Quartile	28.58284	25.18229	22.80133	21.09198
Median	30.88258	26.64135	23.90603	22.21537
Mean	29.71924	26.17661	23.65266	21.89017
3rd Quartile	31.7282	27.57615	24.8421	23.06656
Maximum	33.47176	30.17708	26.43643	24.85392

Figure 1 K-Means Cluster Assessment Report

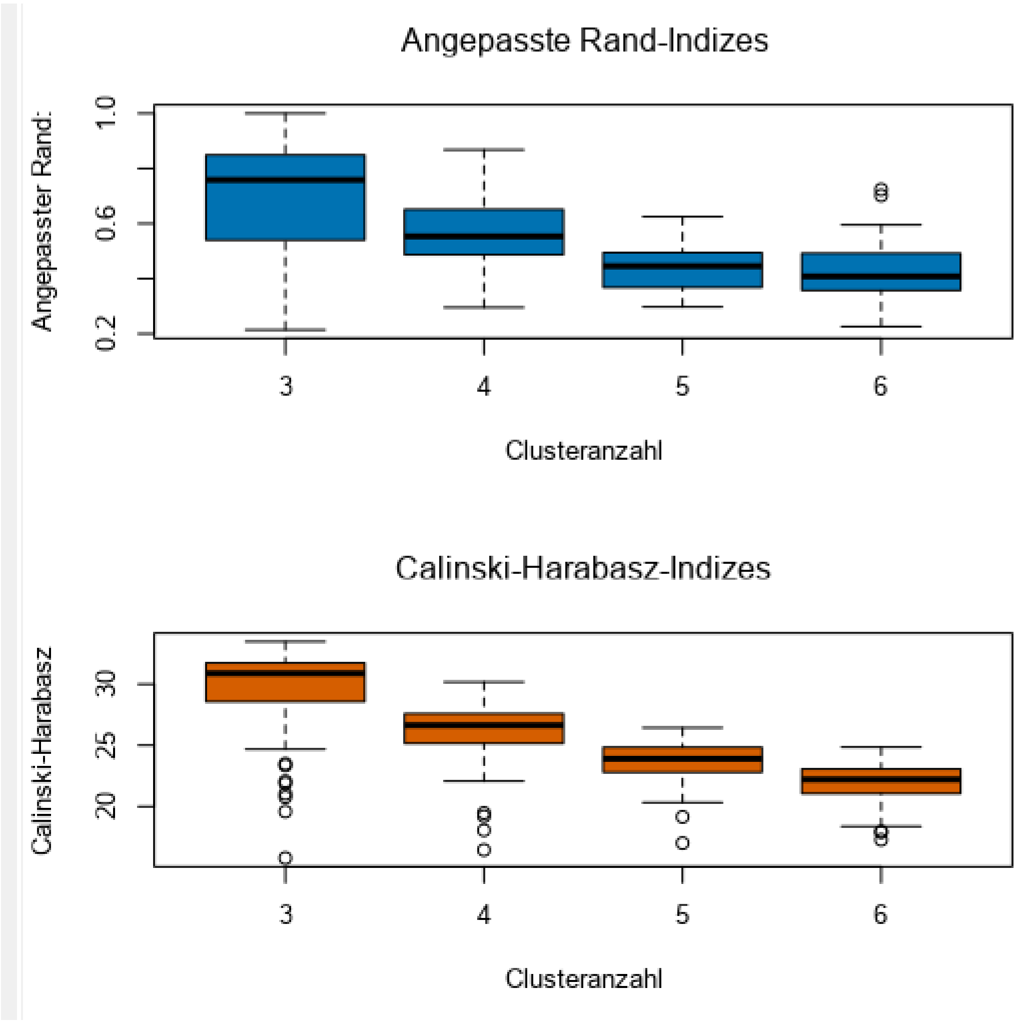
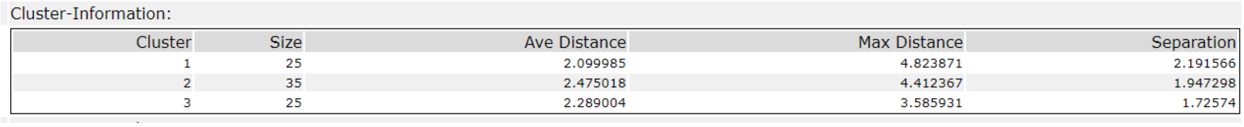


Figure 2 Adjusted Rand Indices - Calinski-Harabasz-Indices

2. How many stores fall into each store format?

Answer: Cluster 1 contains 25 stores, cluster 2 has 35 stores and cluster 3 has 25 stores.



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

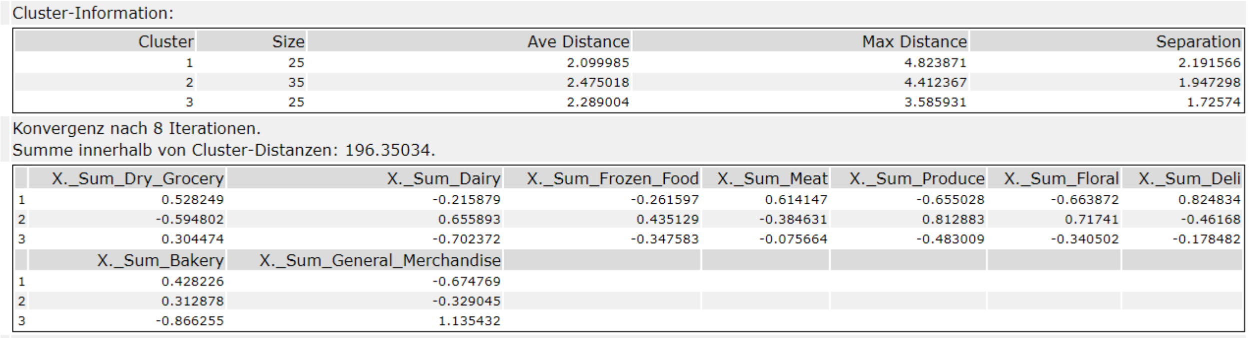
Figure 3 Cluster information

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

Answer: If you compare the numbers for each product to each cluster you can make up some differences. The larger the number, the higher the sale in this category

For example …

- *Cluster 1 has a high number at the product category "Deli" (0.82) but has a low number at the product category "General Merchandise" (-0.67)*
- *Cluster 2 has a high number at the product category "Produce" (0.81) but has a low number at the product category "Dry Grocery" (-0.59)*
- *Cluster 3 has a high number at the product category "General Merchandise" (1.14) but has a low number at the product category "Bakery" (-0.87)*



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

Konvergenz nach 8 Iterationen.
Summe innerhalb von Cluster-Distanzen: 196.35034.

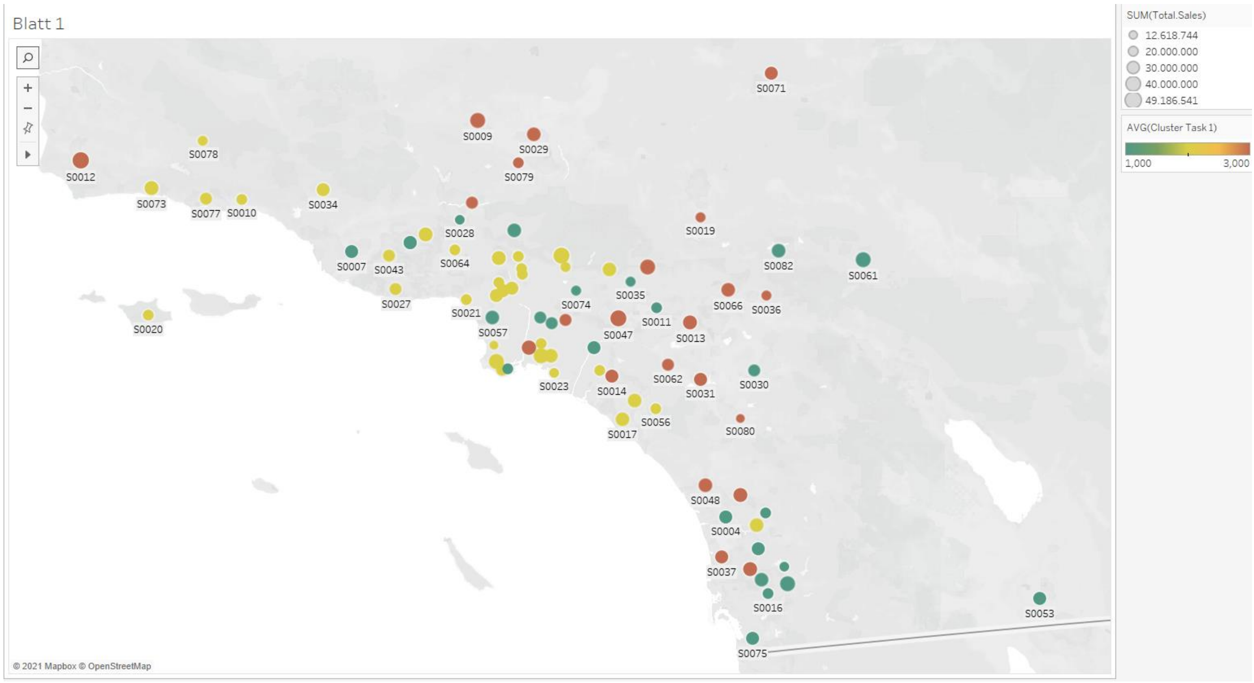
	X._Sum_Dry_Grocery	X._Sum_Dairy	X._Sum_Frozen_Food	X._Sum_Meat	X._Sum_Produce	X._Sum_Floral	X._Sum_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	X._Sum_Bakery	X._Sum_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Figure 4 Cluster Analysis

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.

Answer: In this generated Tableau map you can see all 85 stores. There are 3 clusters. First Cluster is green, second cluster is yellow and the third cluster is red. The size of the circle should indicate the total sales.



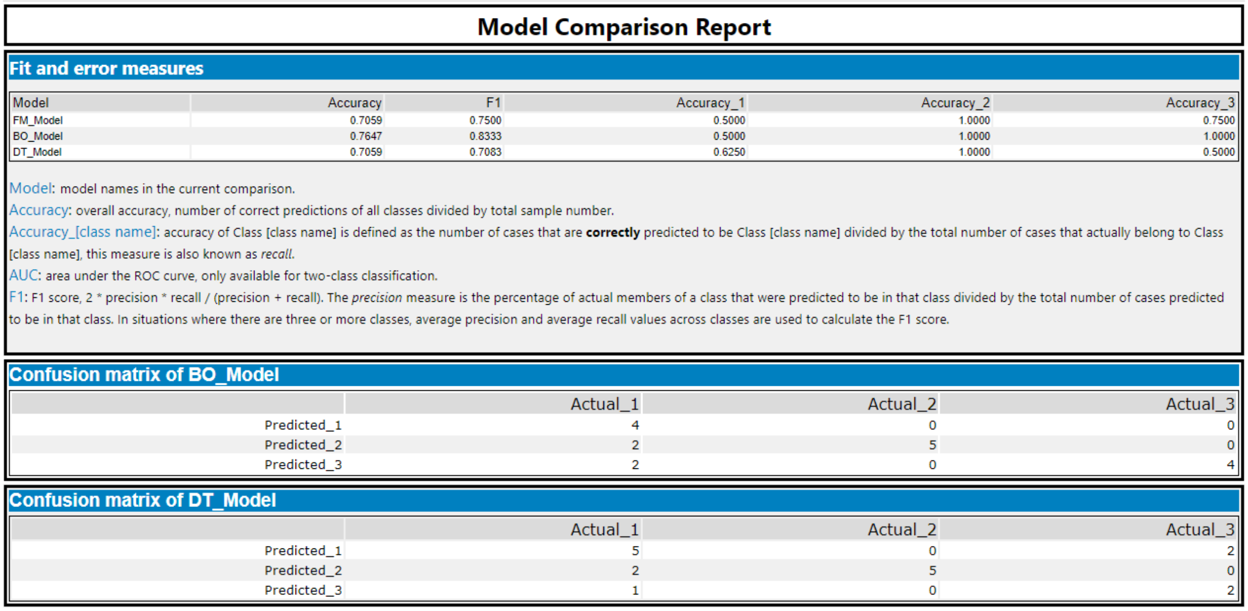
Map 1 Tableau - Store Location - Cluster - Total Sale

Link: https://public.tableau.com/profile/fabian2976#!/vizhome/Task_1_predicitive_analysis_for_business/Blatt1?publish=yes

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.)

Answer: After using the Model Comparison for the Decision Tree, Random Forest and the Boosted Model you can see, that the accuracy for the Boosted Model is the highest, that's why I am follow up with this Model.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
FM_Model	0.7059	0.7500	0.5000	1.0000	0.7500
BO_Model	0.7647	0.8333	0.5000	1.0000	1.0000
DT_Model	0.7059	0.7083	0.6250	1.0000	0.5000

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BO_Model

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

Confusion matrix of DT_Model

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

Figure 5 Model Comparison Report

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?

Answer: After compared the performance of an ETS- and an ARIMA Model I have chosen the ETS Model (M, N, M).

- *The error is irregular -> multiplicatively*
- *The seasonality shows a slide increase trend -> multiplicatively*
- *The trend is not clear -> none*

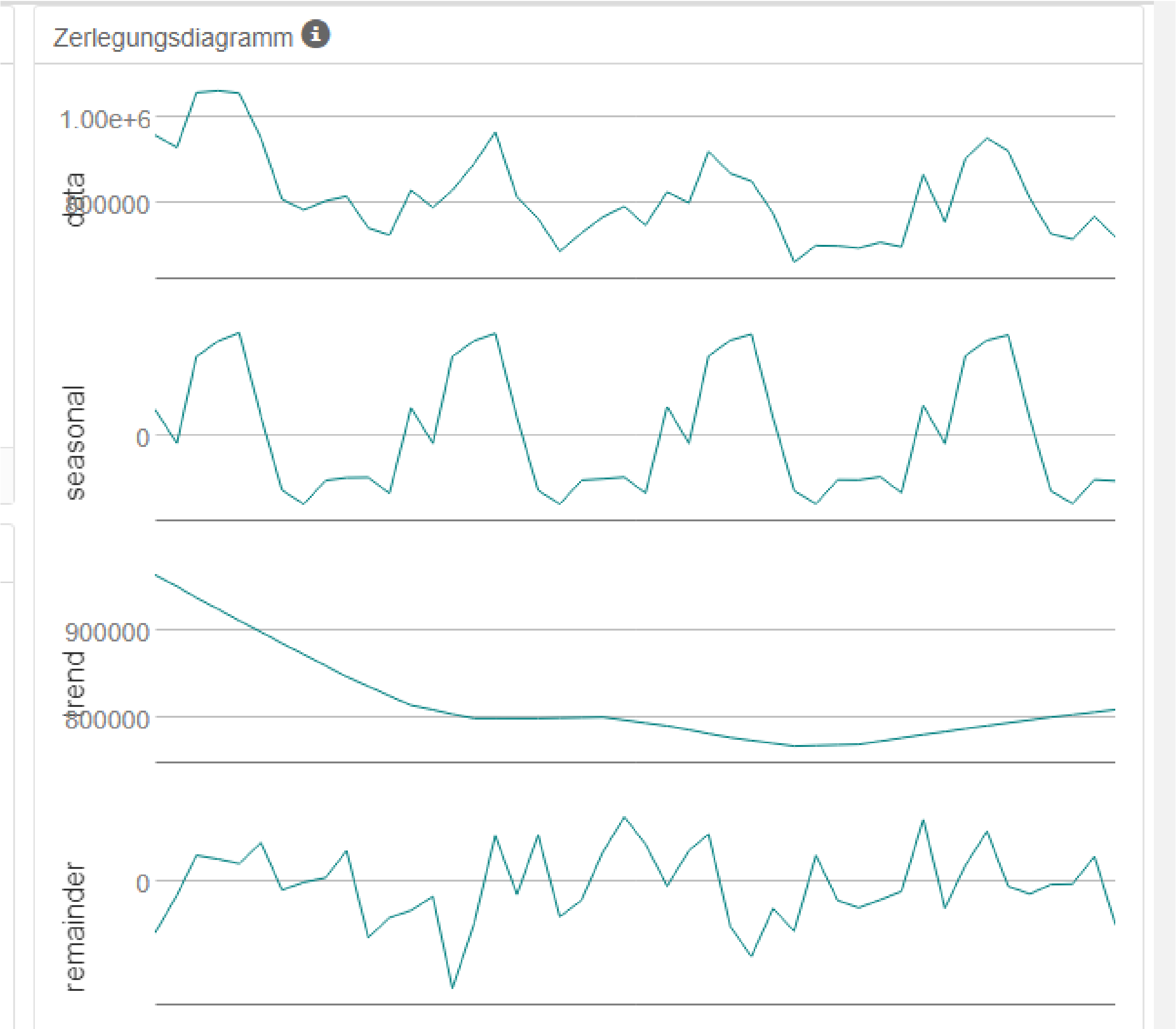


Figure 6 Decomposition diagram – data graph – seasonal graph – trend graph – remainder graph

Answer: If you look at the ACF and PACF the parameters (0,1,2) and (0,1,0) have to been chosen.

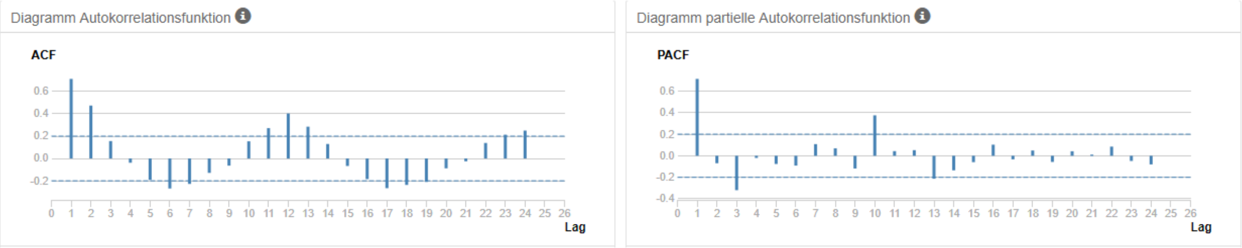
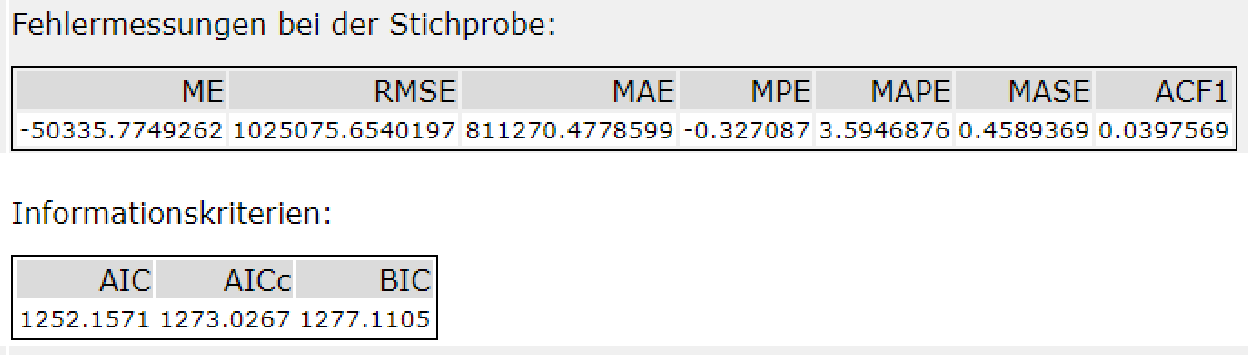


Figure 7 Autocorrelation – ETS

Answer: If you compare the ETS model and ARIMA model regarding the accuracy you can see that the accuracy of the ETS Model is higher. After a holdout of 6 months data is used the RMSE at the ETS model lower with 1,025,075.7 in comparison to 1,071,563 of the ARIMA Model. Also, the MASE of the ETS model is a bit higher 0.46 in comparison to 0.43 of the ARIMA Model. If you look at the AIC you can see that the ETS model has a higher AIC with 1,273 in comparison to the ARIMA model with 851.Based on all these facts the ETS model is the best model for the forecast.



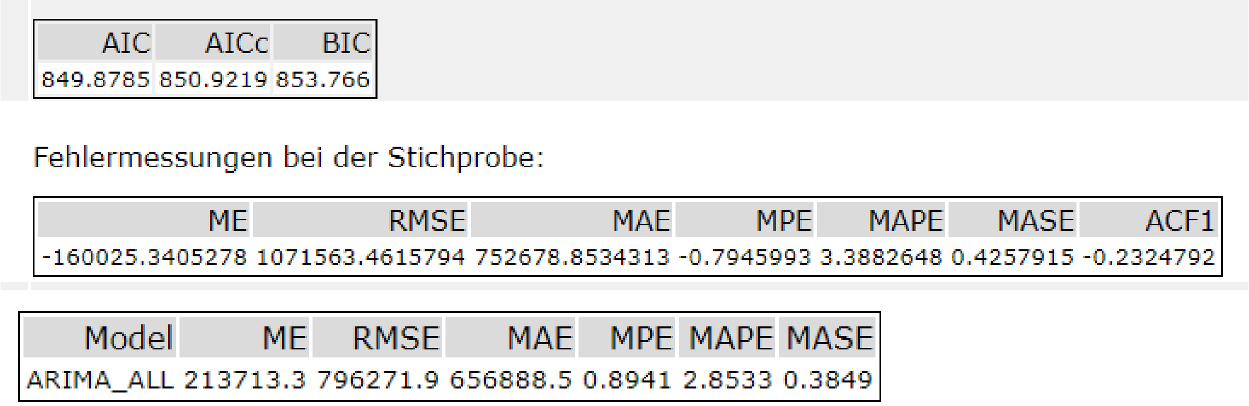
Fehlermessungen bei der Stichprobe:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-50335.7749262	1025075.6540197	811270.4778599	-0.327087	3.5946876	0.4589369	0.0397569

Informationskriterien:

AIC	AICc	BIC
1252.1571	1273.0267	1277.1105

Figure 8 Comparison Tool – ETS



AIC	AICc	BIC
849.8785	850.9219	853.766

Fehlermessungen bei der Stichprobe:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-160025.3405278	1071563.4615794	752678.8534313	-0.7945993	3.3882648	0.4257915	-0.2324792

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_ALL	213713.3	796271.9	656888.5	0.8941	2.8533	0.3849

Figure 9 Comparison Tool - ARIMA

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	Sum new Stores	Sum existing stores and new stores
Jan-16	2.563.357	24.392.418
Feb-16	2.483.924	23.630.254
Mar-16	2.910.944	26.646.630
Apr-16	2.764.881	25.174.397
May-16	3.141.305	28.763.134
Jun-16	3.195.054	29.502.912
Jul-16	3.212.390	29.917.483
Aug-16	2.852.385	26.293.146
Sep-16	2.521.697	23.161.744
Oct-16	2.466.750	22.553.020
Nov-16	2.557.744	23.415.863
Dec-16	2.530.510	23.785.700

Table 1 existing stores and new stores

Answer: In this illustration, you can see that through the new stores, sales are increased. Here visible through the orange area.

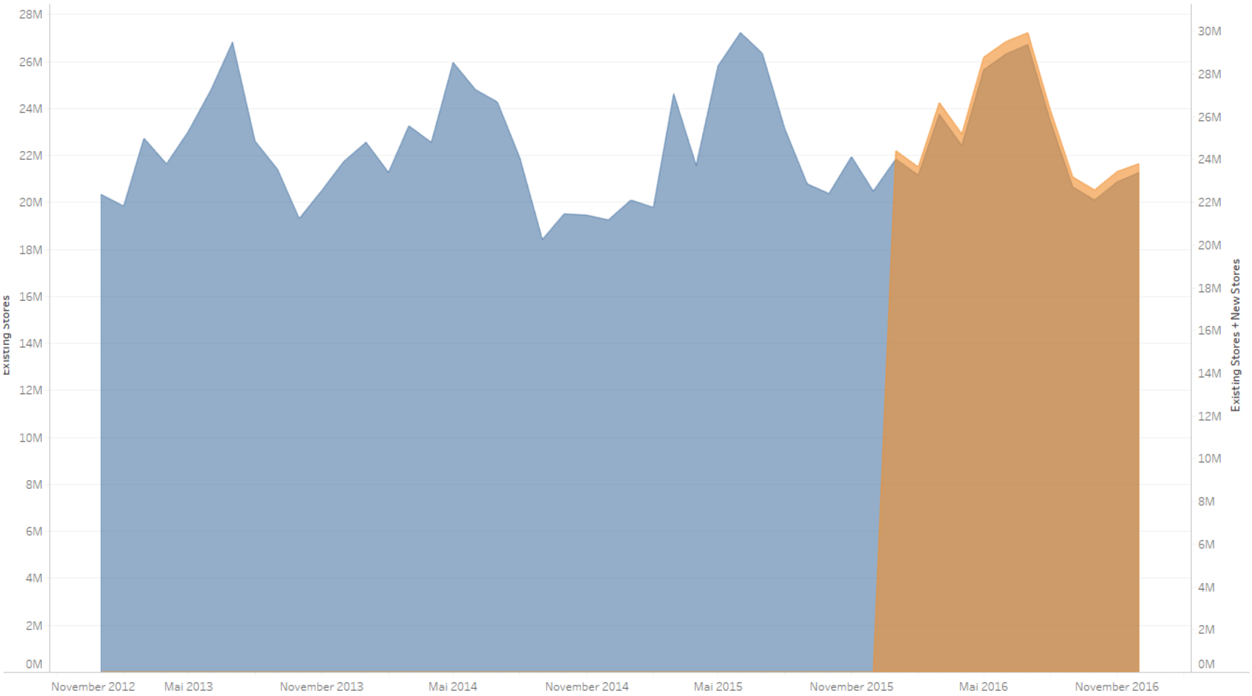


Figure 10 History Data excisting stores + Forecast new stores

Link: https://public.tableau.com/views/Forecastexstingstoresnewstores/Blatt1?:language=de&:display_count=y&publish=yes&:origin=viz_share_link