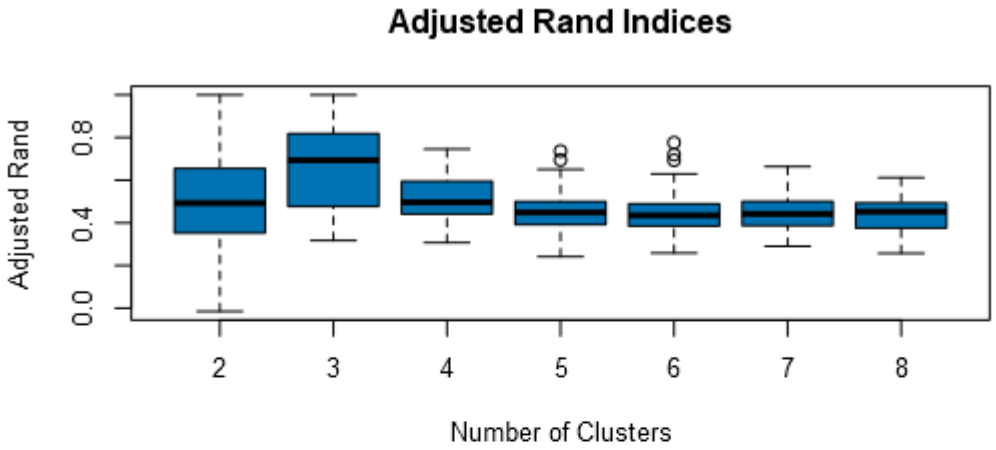
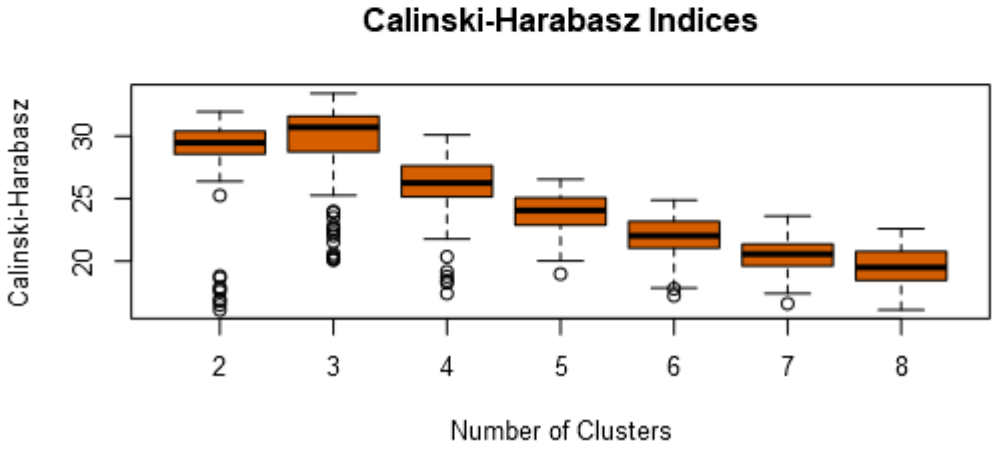
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

We determine the optimal number of store formats based on 2015 sales data. In particular, we use percentage sales per category per store for clustering (category sales as a percentage of total store sales). Furthermore, we chose to standardize the variables by using Z-Score. The K-Centroids Diagnostic tool allows making an assessment of the appropriate number of clusters. The selected clustering algorithm is K-Means. Two measures examined are the adjusted Rand index and the Calinski–Harabasz index. Below, we report the output of the tool, the distribution of the two statistics for differing numbers of clusters. The information is conveyed via two box and whisker plots (one each for the adjusted Rand index and the Calinski-Harabasz index).





The preferred number of clusters based on each measure corresponds to one with the highest mean and median of the solutions compared. Then, the optimal number of store formats is 3.

2. How many stores fall into each store format?

Below, the cluster information generated by the K-Centroids Cluster Analysis Tool.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

We derive that cluster 1 has 23 stores, cluster 2 has 29 stores and cluster 3 has 33 stores.

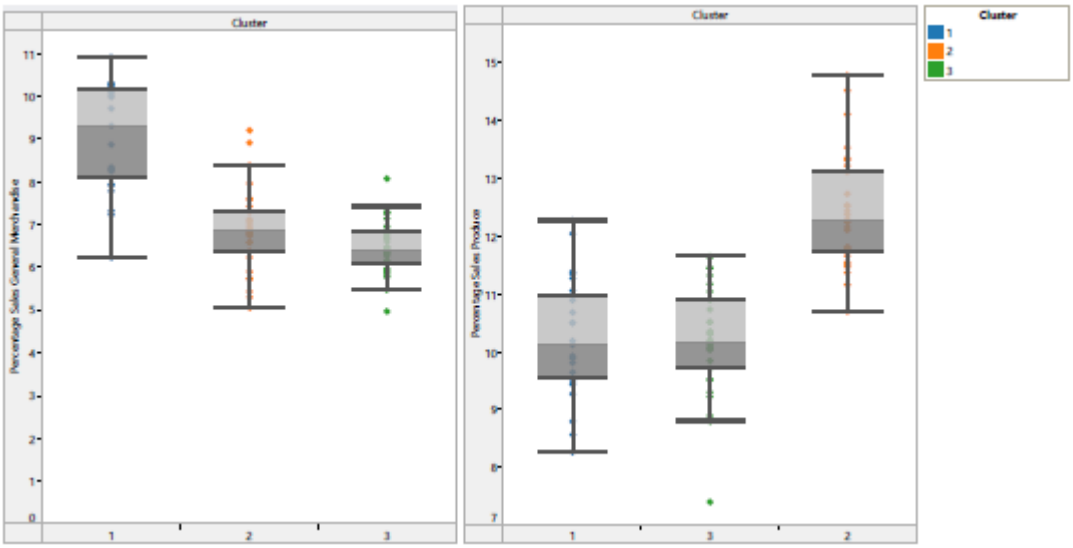
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Below, we report the final cluster centers generated by the K-Centroids Cluster Analysis Tool.

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

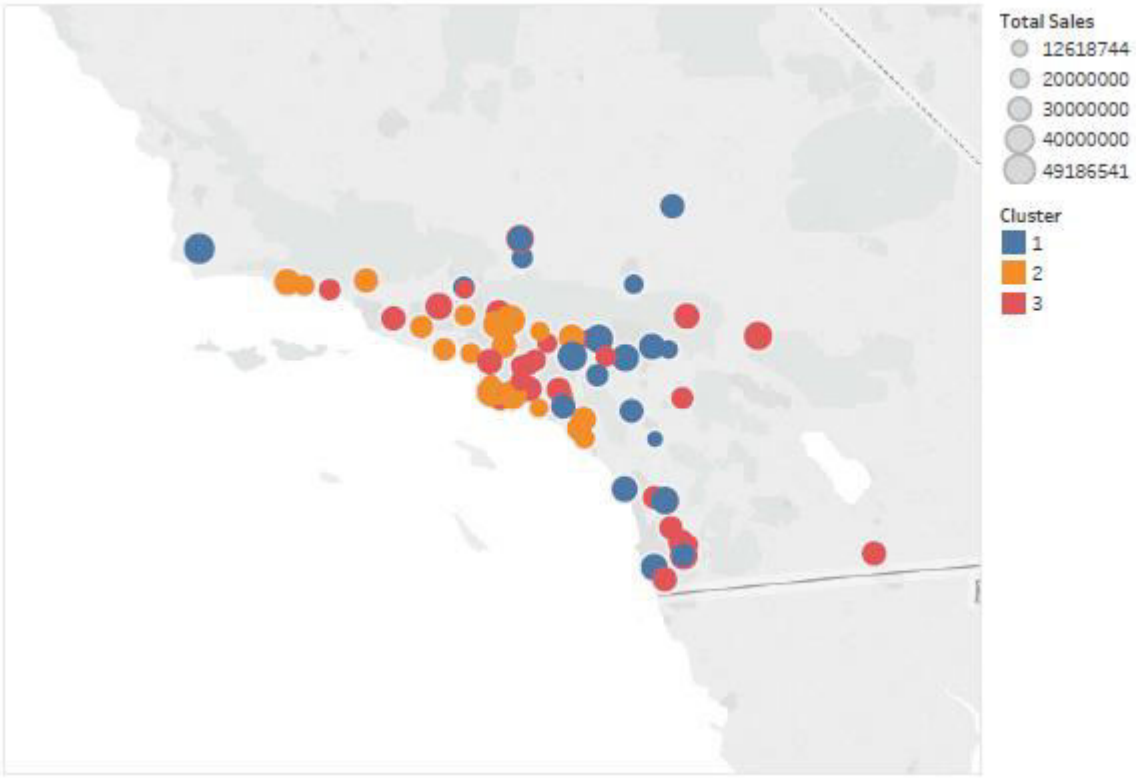
	Percent_Bakery	Percent_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

They computed as the mean for each variable within each final cluster. The final cluster centers reflect the characteristics of the typical case for each cluster. In particular, we derive that the Stores in cluster 1 characterized by high percentage sales of general merchandise. The stores in cluster 2 by high percentage sales of Produce. Below, we provide a Tableau visualization that shows box plots of the percentage sales of general merchandise and percentage sales of Produce and they confirm the previous conclusions.



https://public.tableau.com/profile/santino1079

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/profile/santino1079

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.)

To predict which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store, we use a decision tree, forest, and boosted model. The Model Comparison tool compares the performance of different predictive models. Below, the report generated.

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted	0.8235	0.8543	0.8000	0.6667	1.0000
Random_Forest	0.8235	0.8251	0.7500	0.8000	0.8750

Boosted Model is chosen. It has same overall accuracy as Forest Model but higher F1 value.

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

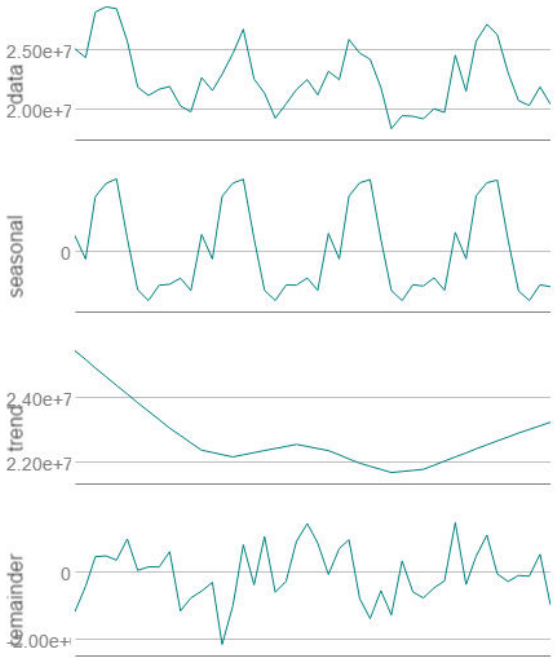
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2

S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

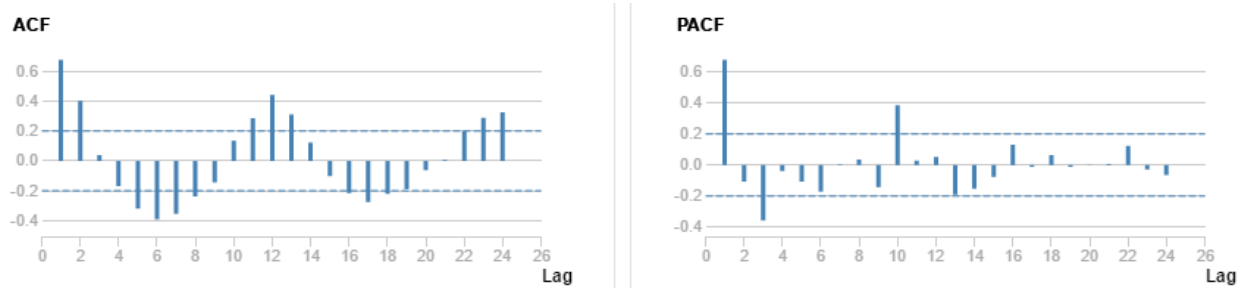
Task 3: Predicting Produce Sales

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?

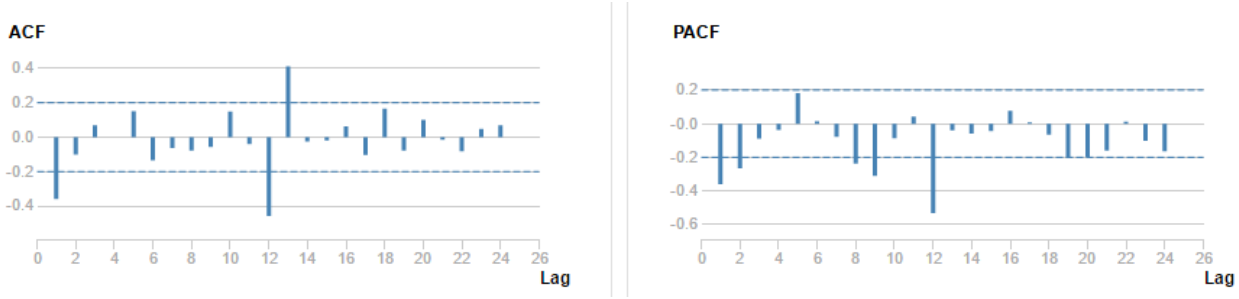
We prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To forecast sales for existing stores we aggregate sales across all stores by month and produce a forecast. The time series decomposed into three sub-time series that is the seasonal component, the trend component, and the remainder. Below, we report the time series decomposition plot.



We build the ETS model by looking the seasonal component, trend component and noise/remainder component in the time series decomposition plot. Seasonality is growing slightly overtime (the peaks are increasing ever so slowly), so we apply it multiplicatively. The series does not present a trend. The error is growing or shrinking over time, we apply the error multiplicatively. Then, we run ETS (M, N, M).
Fitting an ARIMA model requires the series to be stationary. Autocorrelation plots (ACF) or partial autocorrelation (PACF) plots are a visual tool in determining the existence of autocorrelation for any particular lag.



In particular, we note that the ACF shows an oscillation, indicative of a seasonal series. Examine the patterns across lags that are multiple seasonal periods. For monthly data, look at lags 12, 24, we mote the peaks occur at lags of 12 months and 24 months. Furthermore, we observe a spike at lag 1 in an ACF plot indicates a strong correlation between each series value and the preceding value. Then, we fit the times series with seasonal ARIMA model. Non-stationary series can be corrected by a simple transformation such as difference. We consider seasonal first difference, we can observe in the plot below and the time series has been stationarized. By looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the seasonal first difference, we can identify the numbers of AR and/or MA terms that are needed.

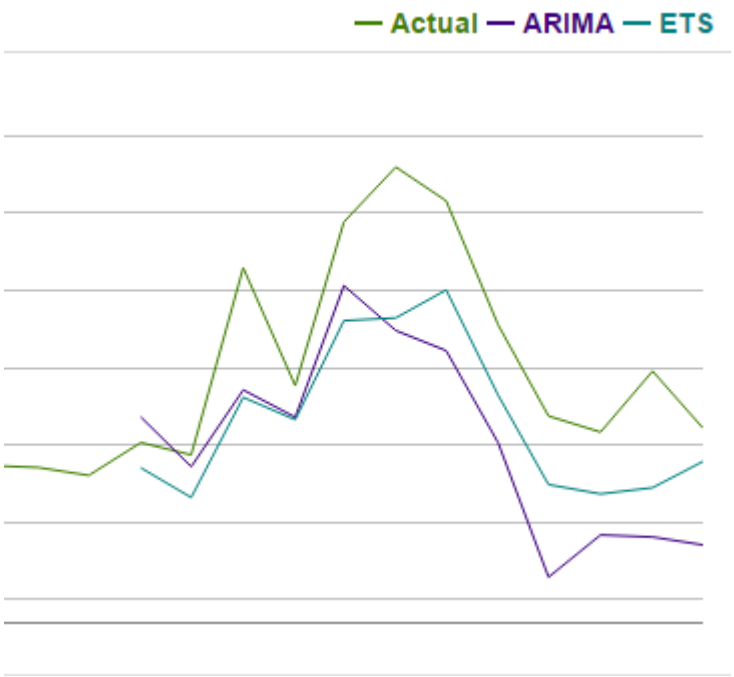


For non-seasonal terms, we examine the early lags and we observe two negative spikes in the ACF at lag 1, this indicates non-seasonal MA terms. For seasonal terms, we note that there is a negative peak occur at lags of 12 months. This indicates seasonal MA terms. Then, the model that fits is ARIMA (0, 1, 1) (0, 1, 1) 12.
When choosing models, we use a portion of the available data for testing, holdout sample, and use the rest of the data for estimating the model. The size of the holdout sample should be the number of periods we want to forecast, and, given that, the goal is to provide a forecast for the next 12 months of sales. The data points from 2015-01 to 2015-12 removed from the data series. Below, we report with the table of the accuracy statistics for each model.

Accuracy Measures:

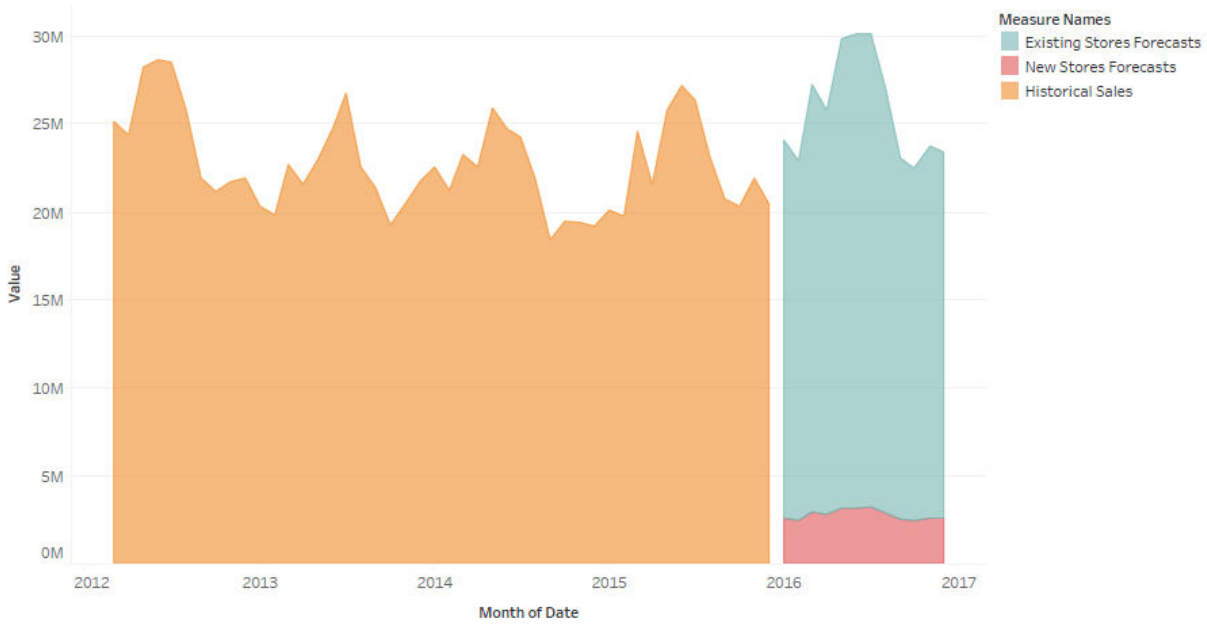
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988	NA
ETS	1983593	2226513	1983593	8.4729	8.4729	1.2691	NA

From the values of the table, the ETS model is better than the ARIMA model. Given that, RMSE and MASE of the ETS model are lower than the ARIMA model. For the ETS model, RMSE is 1983593 and MASE 1.2691 and for the ARIMA model, RMSE is 2999244 and MASE is 1.6988.
Furthermore, below we report the plot that shows all the values of the time series and forecast values for all the models compared.



The plot shows two forecast methods applied to the monthly produces sales using data only to the end of 2014. The actual values for the period 2015 also shown. The graph confirms that the ETS method is best for these data. Therefore, we chose the ETS model. This model used including as regards to forecast sales for new stores. In particular, we forecast produce sales for the average store (rather than the aggregate) for each cluster. Then, we multiply the average store sales forecast by the number of new stores in that cluster. Finally, to get the forecast for all new stores we sum the new stores' sales forecasts for each of the segments.

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".



Year of Date	Month of D..	Historical Sales	Existing Stores F..	New Stores Forec..
2012	Marzo	25.151.526		
	Aprile	24.406.048		
	Maggio	28.249.539		
	Giugno	28.691.364		
	Luglio	28.535.707		
	Agosto	25.793.521		
	Settembre	21.915.642		
	Ottobre	21.203.563		
	Novembre	21.736.159		
	Dicembre	21.962.977		
2013	Gennaio	20.322.684		
	Febbraio	19.829.621		
	Marzo	22.717.070		
	Aprile	21.625.385		
	Maggio	23.000.152		
	Giugno	24.755.406		
	Luglio	26.803.106		
	Agosto	22.600.217		
	Settembre	21.401.266		
	Ottobre	19.296.578		
	Novembre	20.489.773		
	Dicembre	21.715.707		
2014	Gennaio	22.544.458		
	Febbraio	21.262.413		
	Marzo	23.247.169		
	Aprile	22.541.988		
	Maggio	25.943.047		
	Giugno	24.782.178		
	Luglio	24.263.118		

Year of Date	Month of D..	Historical Sales	Existing Stores F..	New Stores Forec..
2014	Agosto	21.879.989		
	Settembre	18.407.264		
	Ottobre	19.497.572		
	Novembre	19.444.753		
	Dicembre	19.240.385		
2015	Gennaio	20.088.529		
	Febbraio	19.772.333		
	Marzo	24.608.407		
	Aprile	21.559.729		
	Maggio	25.792.075		
	Giugno	27.212.464		
	Luglio	26.338.477		
	Agosto	23.130.627		
	Settembre	20.774.416		
	Ottobre	20.359.981		
	Novembre	21.936.907		
	Dicembre	20.462.899		
2016	Gennaio		21.539.936	2.587.451
	Febbraio		20.413.771	2.477.353
	Marzo		24.325.953	2.913.185
	Aprile		22.993.466	2.775.746
	Maggio		26.691.951	3.150.867
	Giugno		26.989.964	3.188.922
	Luglio		26.948.631	3.214.746
	Agosto		24.091.579	2.866.349
	Settembre		20.523.492	2.538.727
	Ottobre		20.011.749	2.488.148
	Novembre		21.177.435	2.595.270
	Dicembre		20.855.799	2.573.397

https://public.tableau.com/profile/santino1079