Projeto: Capstone de análise preditiva

Tarefa 1: Determine formatos de loja para as lojas existentes

1. Qual é o número ideal de formatos de loja? Como você chegou a esse número?

O número ideal de formatos de loja foi baseado nos dados de 2015 e utilizamos a média de vendas por categoria e loja para cada agrupamento. A algoritmo de clusterização utilizado foi o K-Means e os gráficos e tabela abaixo mostram que o número ideal de clusters é de **3**, isso por conta que ambos gráficos (Adjusted Rand Indices e Calinski-Harabasz Indices) mostram uma mediana maior dentro do cluster 3.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.27351	0.31976	0.274316	0.235718
1st Quartile	0.35943	0.594017	0.46406	0.39294	0.377774
Median	0.544023	0.705326	0.53195	0.456588	0.421798
Mean	0.524263	0.69161	0.548167	0.470346	0.435429
3rd Quartile	0.694147	0.800179	0.635682	0.520656	0.493589
Maximum	0.952939	0.969034	0.942222	0.841981	0.677532

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.281	17.38103	18.89398	16.69676	15.71092
1st Quartile	28.22121	29.21236	25.03471	22.86498	21.10249
Median	29.4157	31.14178	26.33467	24.22188	21.96958
Mean	28.56936	30.07118	26.18037	23.72205	21.92474
3rd Quartile	30.21867	32.17467	27.4999	25.09459	22.95561
Maximum	31.71569	33.63781	30.1583	26.63063	24.72038

Figura 1: K-Means Cluster Assessment Report

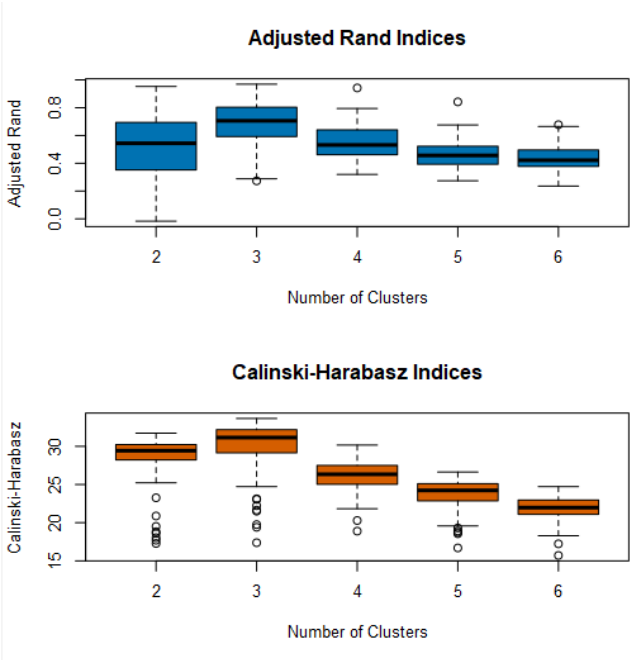

Figura 2: Adjusted Rand Indices e Calinski-Harabasz Indices

2. Quantas lojas enquadram-se em cada formato?

Com base nos números gerados pela ferramenta K-Centroids Cluster Analysis, temos um total de **23 lojas** para o **Cluster 1**, **29 lojas** para o **Cluster 2** e **33 lojas** para o **Cluster 3**.

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

Figura 3: Cluster Information

3. Com base nos resultados do modelo de agrupamento, de que forma os *clusters* diferem um do outro?

Analisando os resultados do total de vendas e das categorias com maior representatividade dentro dos nossos dados (Grocery, Merchandise & Produce) percebemos, conforme gráficos abaixo, que o Cluster 1 tem maior representatividade no Total Sales, Grocery & Merchandise, enquanto o Cluster 2 se mostra maior para a categoria Produce. O Cluster 3 se mostra mais flat sem muita variação dentro do próprio range, ou seja, as lojas são muito parecidas em questões de vendas.

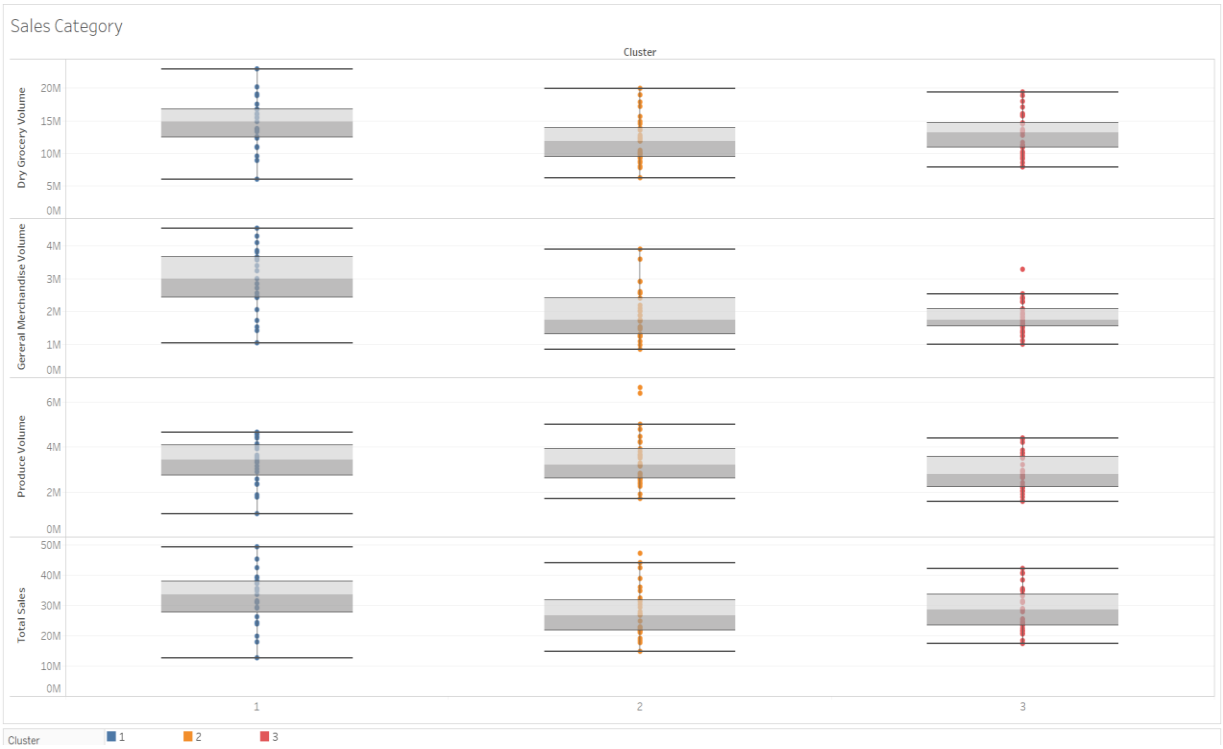

Figura 4: Sales Category Plot in Tableau

4. Envie um dashboard do Tableau (salvo como um arquivo público do Tableau) que mostre a localização das lojas e utilize cores para mostrar os *clusters* e tamanhos para mostrar as vendas totais.

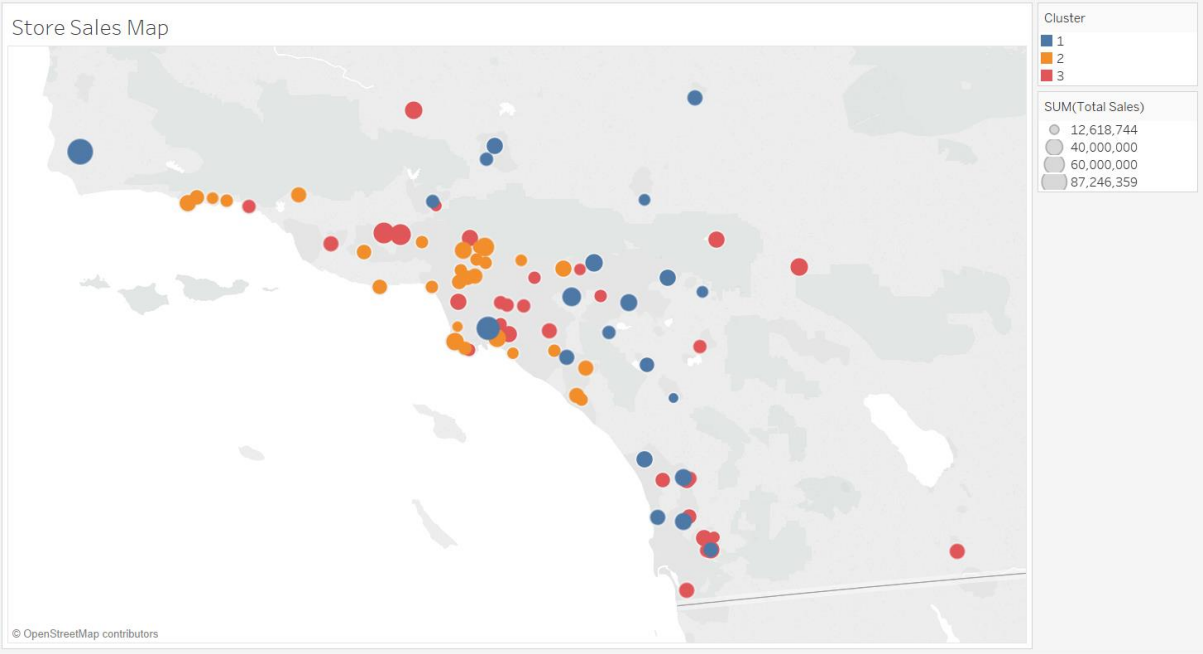

Figura 5: Sales Map by Store & Cluster in Tableau

https://public.tableau.com/profile/jose.cypriano.de.oliveira.junior#!/

Tarefa 2: Formato das lojas novas

1. Qual metodologia você usou para prever o melhor formato para as lojas novas?

Após aplicar os modelos de *Decision Tree, Random Forest* e *Boosted* e analisar os resultados de acurácia de cada metodologia, percebemos que o modelo ***Boosted*** tem mais aderência aos nossos dados. Por mais que o modelo Boosted tenha a mesma acurácia do modelo Random Forest, a variável F1 tem mais precisão no modelo escolhido, conforme podemos ver na tabela abaixo.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_New_Stores	0.7059	0.7685	0.7500	1.0000	0.5556
RF_New_Stores	0.8235	0.8426	0.7500	1.0000	0.7778
BM_New_Stores	0.8235	0.8889	1.0000	1.0000	0.6667

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.

Figura 6: Model Comparison Report

2. Quais são as três variáveis mais importantes que ajudam a explicar a relação entre os indicadores demográficos e o formato das lojas?

De acordo com o gráfico de importância das variáveis, percebemos que as variáveis que mais explicam a relação entre os indicadores demográficos e o formato das lojas são: Age0to9, HVal750kPlus e EdHsGrad.

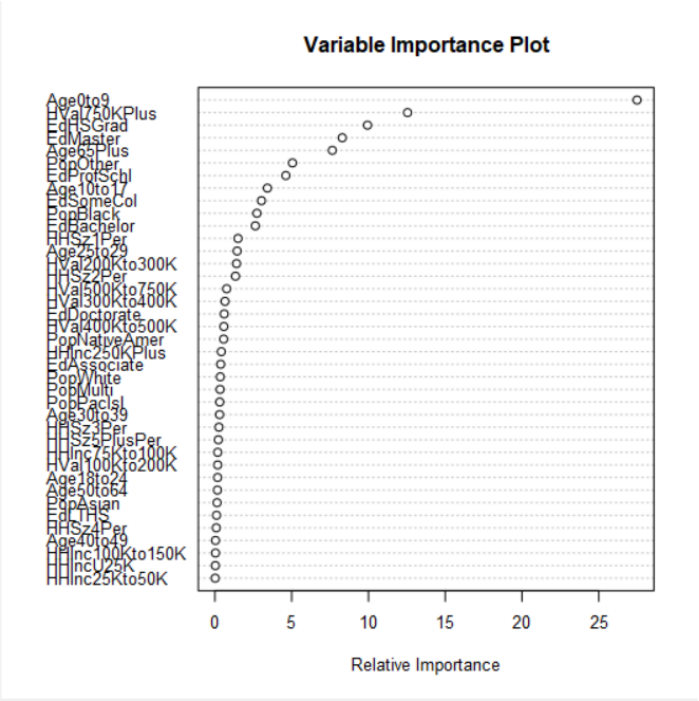

Figura 7: Importance Plot

3. Em que formato cada uma das 10 lojas novas se enquadra? Preencha a tabela abaixo:

Número da loja	Segmento
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Tarefa 3: Prevendo a vendas de produtos

1. Qual tipo de modelo, ETS ou ARIMA, você usou para cada previsão? Use a notação ETS (a, m, n) ou ARIMA (ar, i, ma). Como você chegou a essa decisão?

Ambos os modelos ETS não amortecido e ARIMA foram utilizados para a previsão do nosso problema.

Para o modelo ETS (Erro, Trend & Seasonality) percebemos uma **sazonalidade** dentro dos anos, portanto utilizamos o modelo *multiplicativamente.* Em relação à **tendência**, não foi observado nenhum tipo de tendência nos dados, portanto não utilizamos essa variável. O indicador de **erro**, demonstrado no gráfico Remainder, demonstra claramente uma variação ao longo do tempo, portanto consideramos o modelo como *multiplicativamente.* Podemos ver todos os pontos descritos nos gráficos abaixo.

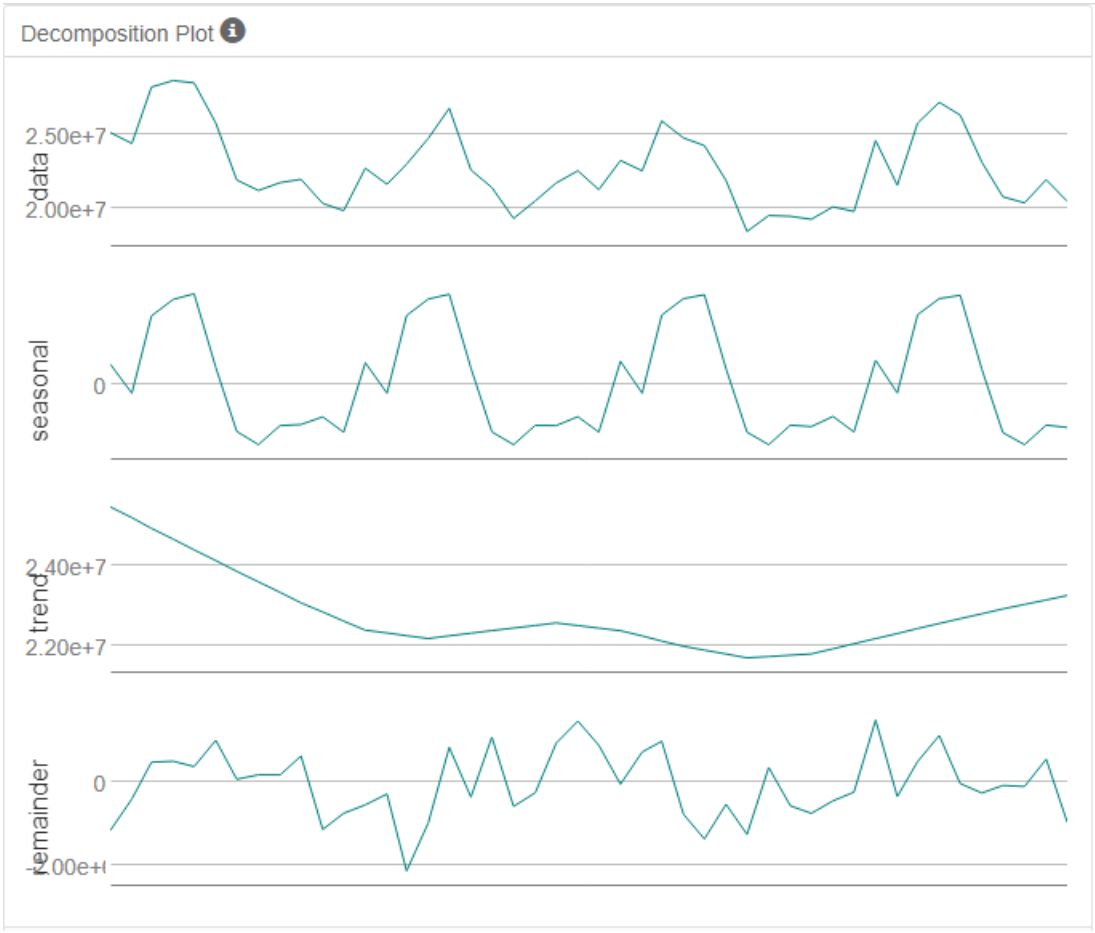

Figura 8: Error, Trend & Seasonality

Para o modelo ARIMA, por conta dos nossos dados conterem sazonalidade e o modelo de trabalhar com dados estacionários, precisamos ajustar a séria temporal para estacionária, utilizando a metodologia de diferença sazonal. Nos gráficos abaixo vemos uma alta variabilidade dos dados e alta correlação nos números, devido a sazonalidade.

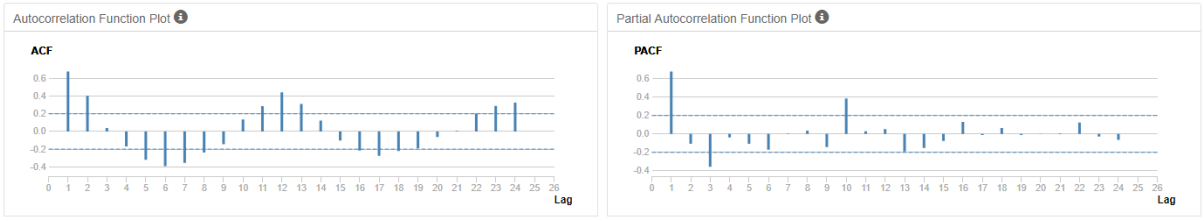

Figura 9: ACF & PACF sem diferenciação

Após fazermos a primeira transformação pela diferenciação, percebemos que os dados ficaram estacionários, ou seja, podemos considerar que a primeira diferença sazonal ajusta os nossos dados para a aplicação do modelo.

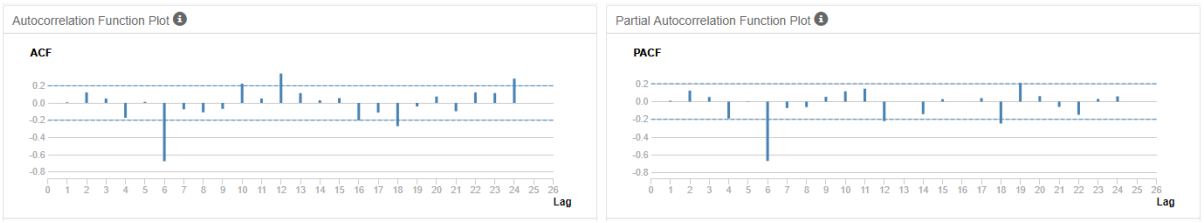

Figura 10: ACF & PACF primeira diferença sazonal

Como os dados estão estacionários, podemos fazer a definição dos termos de AR e/ou MA para aplicação do modelo. Para as séries não sazonais, como lag-1 é positivo e pouco correlacionado, podemos considerar os termos: AR = 1, I = 0 e MA = 1. Enquanto as séries sazonais, percebemos que há mais picos nos intervalos à cada 6 meses, portanto podemos assumir que os termos são: AR = 0, I = 1 e MA = 0. Por fim, nosso modelo fica da seguinte forma: ARIMA (1,0,1) (0,1,0) [12].

Por fim, ao analisar os resultados de ambos os modelos, percebemos que o ETS tem melhor aderência aos nossos dados, portanto é o que devemos usar para realizar nossa previsão. Percebemos isso, pois os indicadores de RMSE e MASE são menores para ETS (RMSE: 1,020,596 | MASE: 0.45), em relação ao modelo de ARIMA (RMSE: 1,352,529 | MASE: 4.28), conforme podemos ver na tabela abaixo.

Method:
ETS(M,N,M)

In-sample error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

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

In-sample error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	-11381.9055206	1352529.2975154	962914.7018434	-0.2027857	4.2852656	0.5417792	-0.0156691

Information Criteria:

AIC	AICc	BIC
1073.4312	1074.2312	1078.0103

Figura 11 - Model Comparison

Além dos indicadores de erro mostrarem que o modelo ETS tem mais acurácia nos nossos dados, o indicador AIC é maior para ETS (1,283) em relação ao ARIMA (1,073).

Abaixo segue conceitos dos indicadores de RMSE, MASE e AIC:

Root Mean Squared Error (RMSE): representa o desvio padrão das diferenças entre os valores previstos e realizados. Esta é uma ótima medida para ser usado quando se deseja comparar modelos porque ela mostra quantos desvios padrão o valor previsto está da média.
Mean Absolute Scaled Error (MASE): é outra medida relativa de erro que é aplicável somente a dados de séries temporais. É definida como o erro médio absoluto do modelo dividido pelo valor médio absoluto da primeira diferença da série. Como a medida de erro é relativa e pode ser aplicado entre os vários modelos, é considerado uma das melhores métricas para a medição do erro.
Akaike Information Criterion (AIC): Mede a qualidade relativa de um modelo estatístico. Esse cálculo equilibra a qualidade do fit do modelo e a complexidade do modelo. Podemos falar que esse é o melhor indicador para definir se um modelo é melhor que outro.

Abaixo os dados gerados para a previsão dos nossos números, utilizando a metodologia de ETS non dempened.

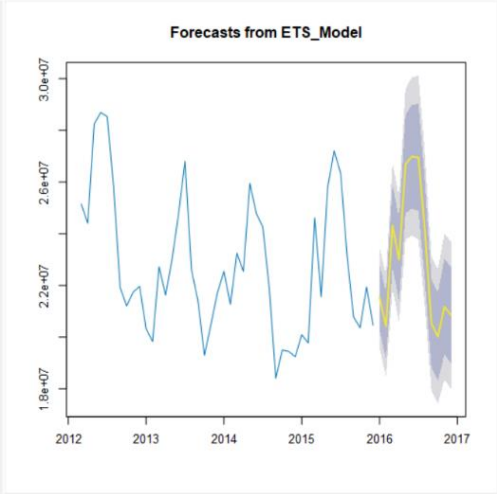

Figura 12 – Forecast Results using ETS Model

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

Figura 13 – Forecast Results Table

2. Envie um dashboard do Tableau (salvo como um arquivo público do Tableau) que inclua uma tabela e um gráfico das três previsões mensais; um para as existentes, um para as novas e um para todas as lojas. Nomeie a aba no arquivo "Tarefa 3" do Tableau.

Abaixo tabela de previsão de vendas para as novas e existentes lojas.

Month of Year-M..	Forecast Existing	Forecast New	Forecast Existing + New
January 2016	21,539,936	2,626,198	24,166,134
February 2016	20,413,771	2,529,186	22,942,956
March 2016	24,325,953	2,940,264	27,266,217
April 2016	22,993,466	2,774,135	25,767,601
May 2016	26,691,951	3,165,320	29,857,272
June 2016	26,989,964	3,203,286	30,193,250
July 2016	26,948,631	3,244,464	30,193,095
August 2016	24,091,579	2,871,488	26,963,067
September 2016	20,523,492	2,552,418	23,075,910
October 2016	20,011,749	2,482,837	22,494,586
November 2016	21,177,435	2,597,780	23,775,215
December 2016	20,855,799	2,591,815	23,447,614

Figura 14: Tabela Forecast de Vendas para Novas e Existentes lojas

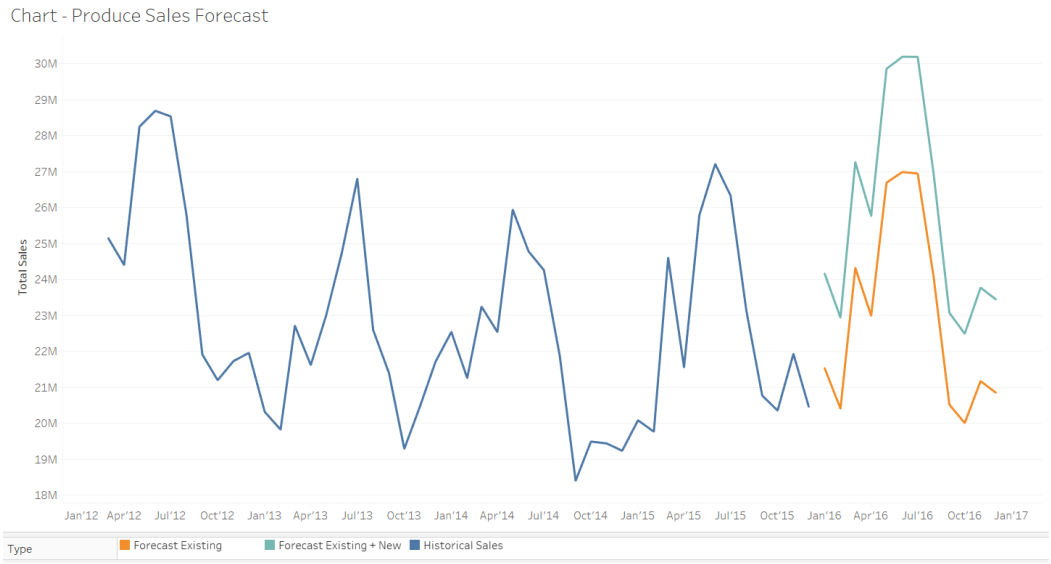

Figura 15: Dados históricos e Forecast de Vendas para Novas e Existentes lojas

https://public.tableau.com/profile/jose.cypriano.de.oliveira.junior#!/