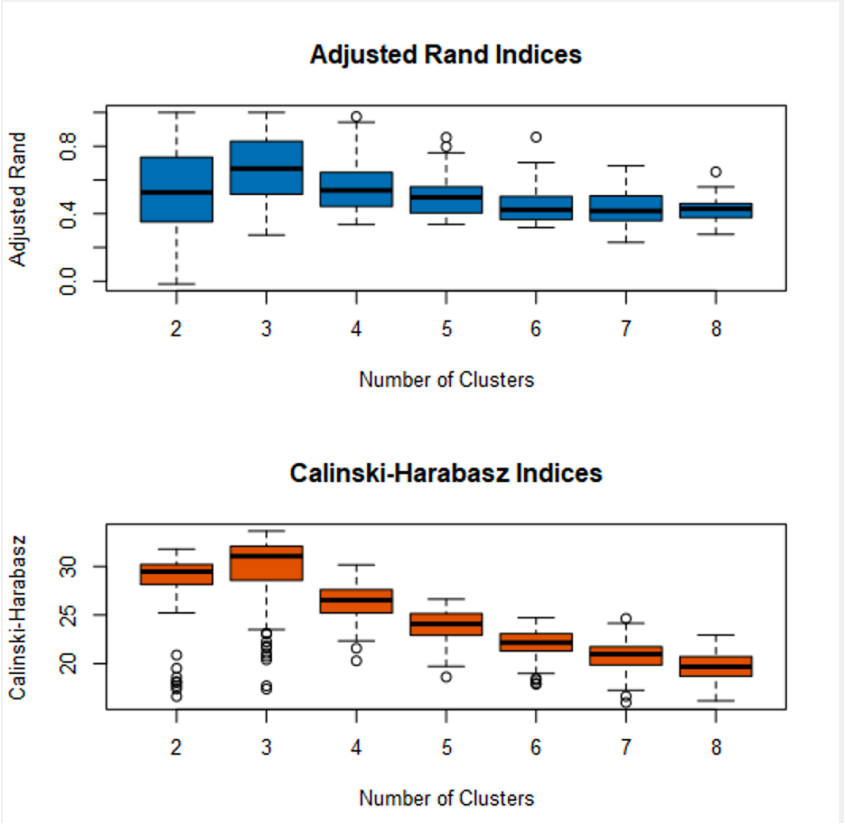
Projeto: Capstone de análise preditiva

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?

   Determinamos o número ideal de formatos de lojas com base nos dados de vendas de 2015. Em particular, usamos o percentual de vendas por categoria por loja para efetuar o agrupamento (vendas de cada categoria como porcentagem do total de vendas das lojas). Padronizaremos as variáveis usando o Z-Score. A ferramenta K-Centroids Diagnostic permite fazer uma avaliação do número apropriado de clusters. O algoritmo de clusterização selecionado é K-Means. Duas medidas examinadas são o índice de Rand ajustado e o índice de Calinski-Harabasz. No seguinte gráfico chegamos a conclusão de qual o numero ideal de formatos de loja:



   O número ideal de clusters com base em cada medida corresponde a um com a média, sendo a mediana mais altas das soluções comparadas. Então, o número ideal de formatos de armazenamento é 3.

2. Quantas lojas enquadram-se em cada formato?

   A seguir as informações do cluster geradas pela Ferramenta de Análise de Cluster K-Centroids.

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

   Podemos dizer então que o cluster 1 tem 23 lojas, o cluster 2 tem 29 lojas e o cluster 3 tem 33 lojas.

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

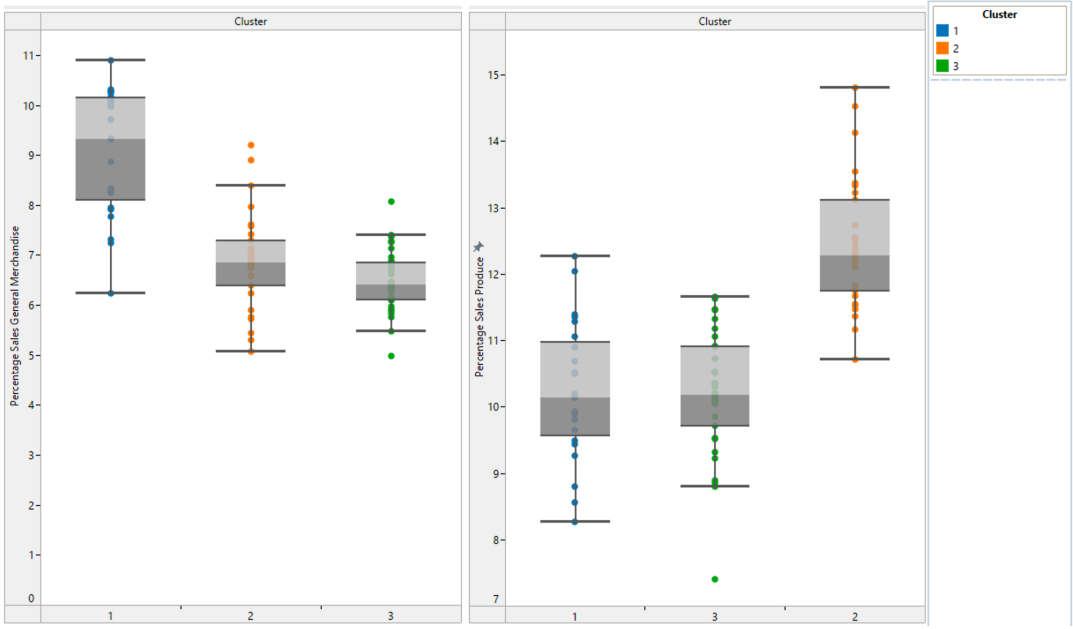
   A seguir as informações do cluster geradas pela Ferramenta de Análise de Cluster K-Centroids.

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

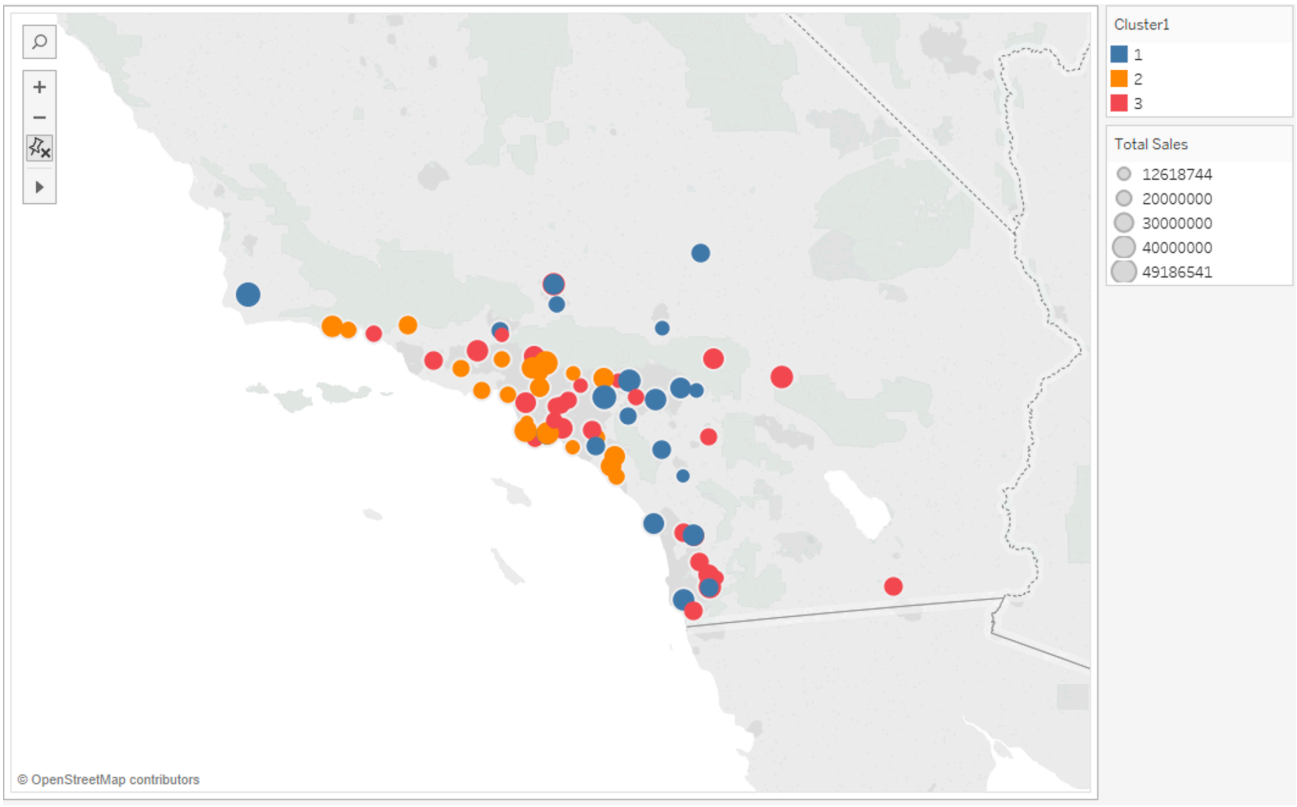
	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327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

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

   Foi calculado como a média para cada variável dentro de cada cluster final. Os centros de clusters finais refletem as características do caso típico de cada cluster. Em particular, deduzimos que as lojas do cluster 1 caracterizam-se pela alta porcentagem de vendas de mercadorias em geral. As lojas no cluster 2 por alta porcentagem de vendas de produtos, a seguir o gráfico que mostra o percentual de vendas por mercadoria e o percentual produzida por cluster: https://public.tableau.com/profile/jose.carlos.soto.morales#!/



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.



https://public.tableau.com/profile/jose.carlos.soto.morales#!/

Tarefa 2: Formato das lojas novas

1. Qual metodologia você usou para prever o melhor formato para as lojas novas?
   Para prever em qual segmento cada loja se enquadra, com base nas características demográficas e socioeconômicas da população que reside na área em torno de cada nova loja, usamos uma árvore de decisão, floresta e modelo "Boosted". A ferramenta de comparação de modelos, compara o desempenho dos diferentes modelos preditivos, segue os resultados:

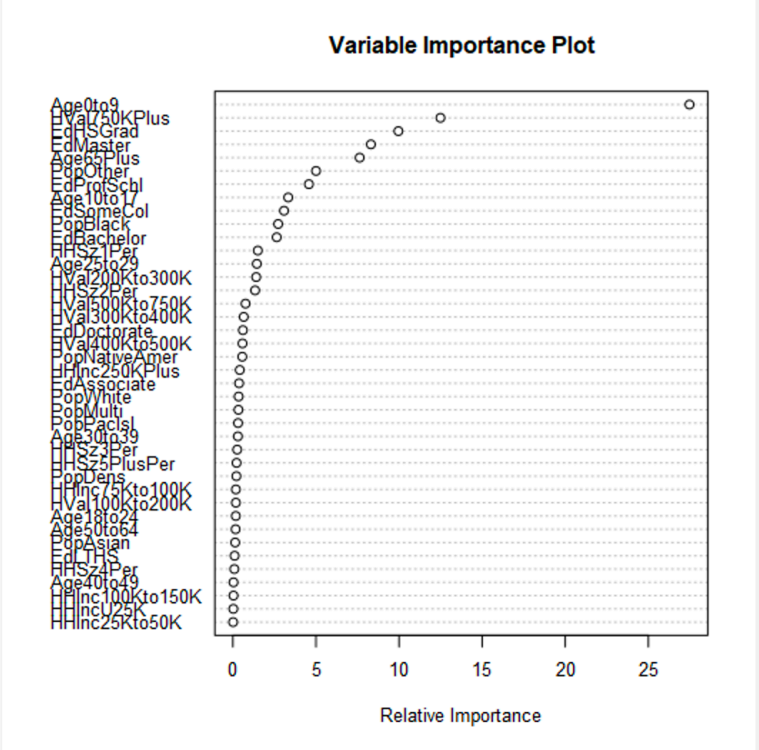
Model Comparison Report

Fit and error measures

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

   Logo escolhemos o modelo “Boosted” ja que a pesar de ter a mesma acurácia (0.8235) do “Decision Tree”, o valor F1 (0.8543) e um pouco maior.

2. Quais são as três variáveis mais importantes que ajudam a explicar a relação entre os indicadores demográficos e o formato das lojas?
   "Ave0to9", "HVal750KPlus" e “EdHSGrad" são as variáveis mais importantes em relação aos indicadores demográficos e o formato das lojas



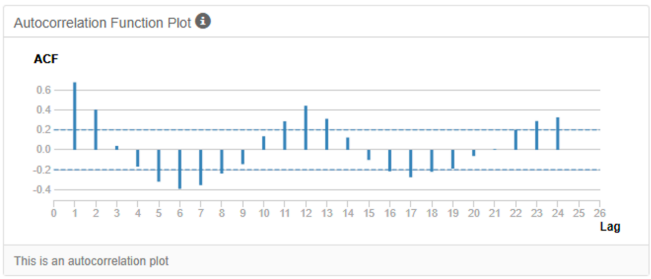
3. Em que formato cada uma das 10 lojas novas se enquadra? Preencha a tabela abaixo:

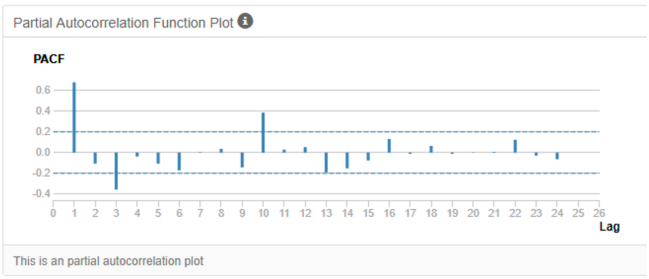
Número da loja	Segmento
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?

   Preparamos uma previsão com granuralidade mensal para vendas de produtos para o ano 2016 para as lojas existentes e as novas. Para prever vendas para lojas existentes, agregamos as vendas em todas as lojas por mês e produzimos uma previsão. A série temporal é decomposta em três sub-séries temporais que é o componente sazonal, o componente de tendência e o resto. Abaixo, relatamos o gráfico de decomposição:



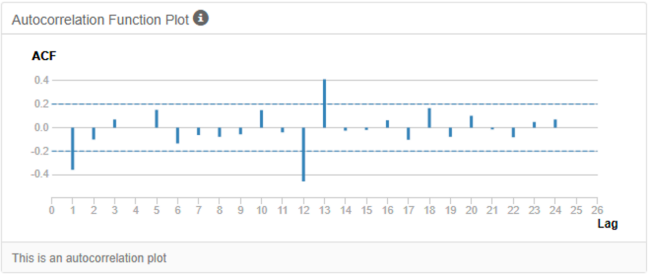


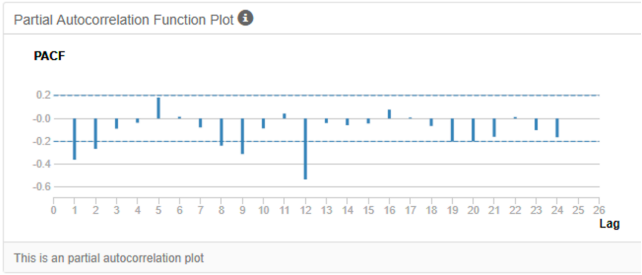
   Construímos o modelo ETS, examinando o componente sazonal, componente de tendência e componente de resto no gráfico de decomposição de série temporal. A sazonalidade está crescendo ligeiramente ao longo do tempo (os picos estão aumentando muito lentamente), então aplicamos isso multiplicativamente, a série

não apresenta tendência, o erro está aumentando ou diminuindo ao longo do tempo, então aplicamos o erro multiplicativamente, portanto escolhemos o ETS (M, N, M).

A montagem de um modelo ARIMA exige que a série seja estacionária. Os gráficos de autocorrelação (ACF) ou de autocorrelação parcial (PACF) nos ajudam a determinar a existência de autocorrelação:

Notamos que o ACF mostra uma oscilação, indicando uma série sazonal, nos “lags" podemos observar vários períodos sazonais. Nos dados mensais, podemos observar que nas defasagens 12, 24, os picos ocorrem em intervalos de 12 meses e 24 meses, além disso, observamos que um pico no atraso 1 em um gráfico de ACF indica uma forte correlação entre cada valor da série e o valor anterior. Em seguida, ajustamos a série com o modelo sazonal ARIMA. Séries não estacionárias podem ser corrigidas por uma transformação como a diferenciação. Aplicando a primeira diferença sazonal, podemos observar no gráfico abaixo que a série temporal foi estacionada. Observando os gráficos de autocorrelação ACF e PACF da primeira diferença sazonal, podemos identificar os números de termos ARIMA necessários.





Observando os dois picos negativos na FAC no desfasamento 1, o que indica termos de AM não sazonais. Para os termos sazonais, notamos que há um pico negativo nos intervalos de 12 meses. Isso indica termos sazonais de MA. Então, o modelo que se ajusta é ARIMA (0, 1, 1) (0, 1, 1) 12.

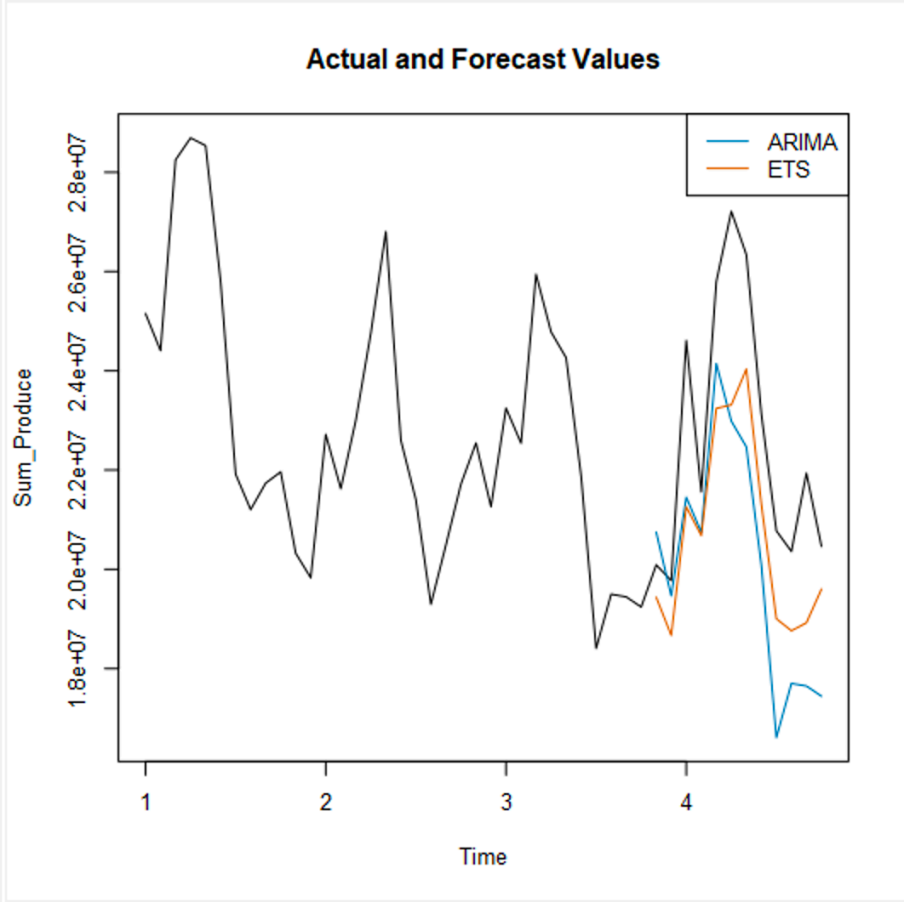
Ao escolher modelos, usamos uma parte dos dados disponíveis para teste, como amostra de validação e usamos o restante dos dados para verificar o modelo. O tamanho da amostra do holdout deve ser o número de períodos que queremos prever e, dado que, o objetivo é fornecer uma previsão para os próximos 12 meses de vendas. Os pontos de dados de 2015-01 a 2015-12 foram removidos da série de dados, podemos observar a tabela das estatísticas de precisão para cada modelo.

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

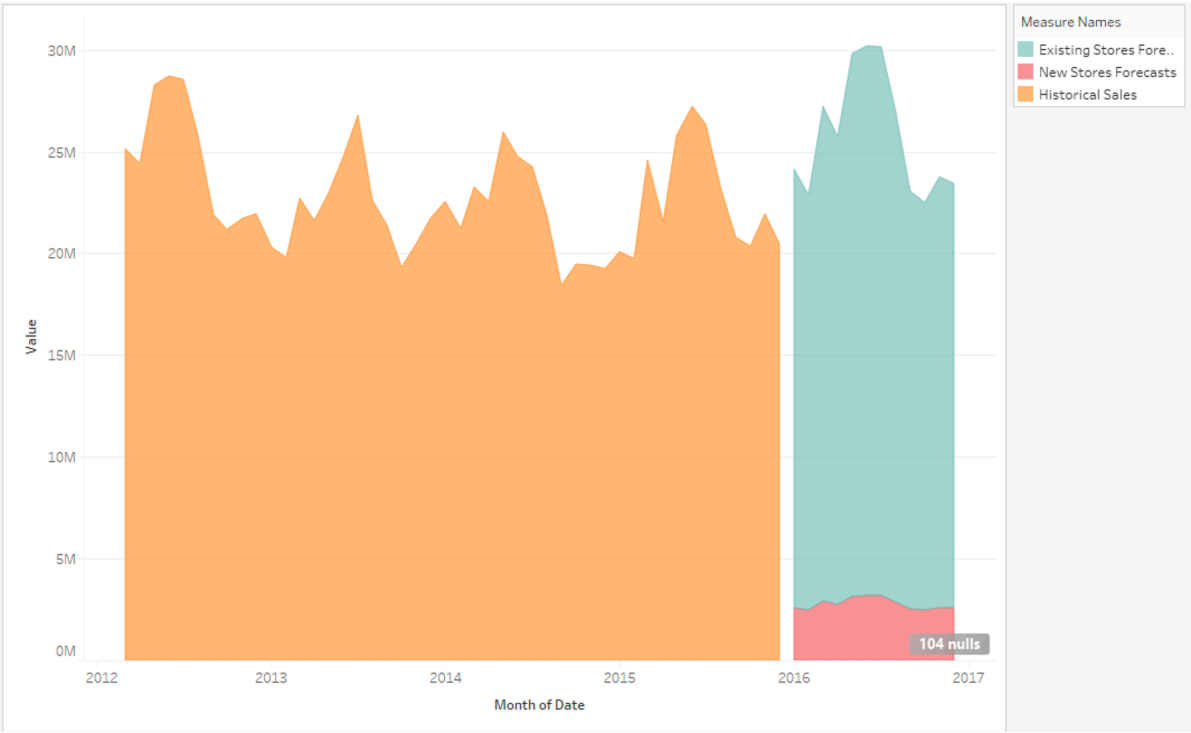
A partir dos valores da tabela, podemos concluir que o modelo ETS é melhor que o modelo ARIMA para este problema, dado que, RMSE e MASE do modelo ETS são inferiores ao modelo ARIMA. Para o modelo ETS, RMSE é 1983593 e MASE 1.2691 e para o modelo ARIMA, RMSE é 2999244 e MASE é 1.6988, segue o gráfico que mostra todos os valores das séries temporais e valores de previsão para todos os modelos comparados.

No teste podemos observar como o modelo ETS tem um comportamento mais exato que o modelo ARIMA para este conjunto de dados, reafirmando a utilização do ETS para o nosso problema.



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.

   https://public.tableau.com/profile/jose.carlos.soto.morales#!/



https://public.tableau.com/profile/jose.carlos.soto.morales#!/

Year of Date	Month of D..	Historical Sales	Existing Stores F..	New Stores Forec..
2012	March	25,151,526		
	April	24,406,048		
	May	28,249,539		
	June	28,691,364		
	July	28,535,707		
	August	25,793,521		
	September	21,915,642		
	October	21,203,563		
	November	21,736,159		
	December	21,962,977		
2013	January	20,322,684		
	February	19,829,621		
	March	22,717,070		
	April	21,625,385		
	May	23,000,152		
	June	24,755,406		
	July	26,803,106		
	August	22,600,217		
	September	21,401,266		
	October	19,296,578		
	November	20,489,773		
	December	21,715,707		
2014	January	22,544,458		
	February	21,262,413		
	March	23,247,169		
	April	22,541,988		
	May	25,943,047		
	June	24,782,178		
	July	24,263,118		
	August	21,879,989		
	September	18,407,264		
	October	19,497,572		
	November	19,444,753		
	December	19,240,385		
2015	January	20,088,529		
	February	19,772,333		
	March	24,608,407		
	April	21,559,729		
	May	25,792,075		
	June	27,212,464		
	July	26,338,477		
	August	23,130,627		
	September	20,774,416		
	October	20,359,981		
	November	21,936,907		
	December	20,462,899		
2016	January		21,539,936	2,587,451
	February		20,413,771	2,477,353
	March		24,325,953	2,913,185
	April		22,993,466	2,775,746
	May		26,691,951	3,150,867
	June		26,989,964	3,188,922
	July		26,948,631	3,214,746
	August		24,091,579	2,866,349
	September		20,523,492	2,538,727
	October		20,011,749	2,488,148
	November		21,177,435	2,595,270
	December		20,855,799	2,573,397