Time Series Analysis of Pharma Sales

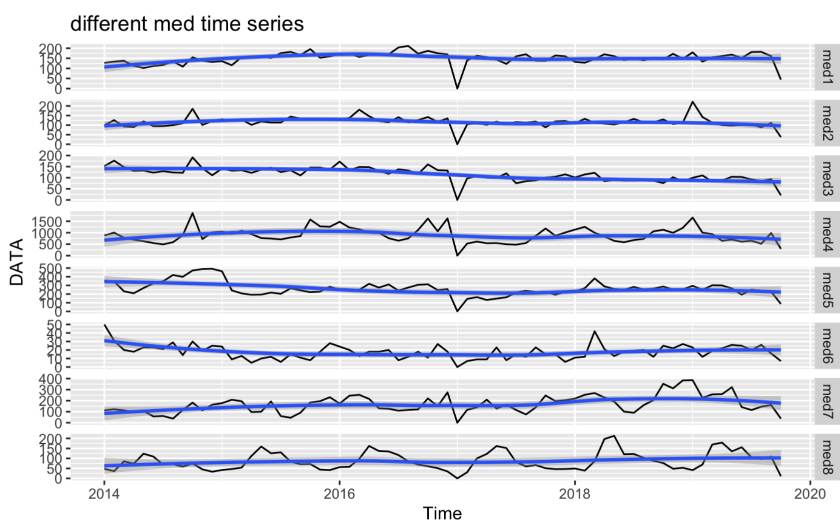
1 Introduction

The objective of this project is to analyze a dataset about pharmaceutical sales that focuses on eight different drugs. There are 70 specific data points for each of the 8 drugs (same dates used for each data point for each drug). The motivation behind studying this dataset is to understand how specific pharmaceutical drugs have changed overtime to result in a growth or decline of their sales.

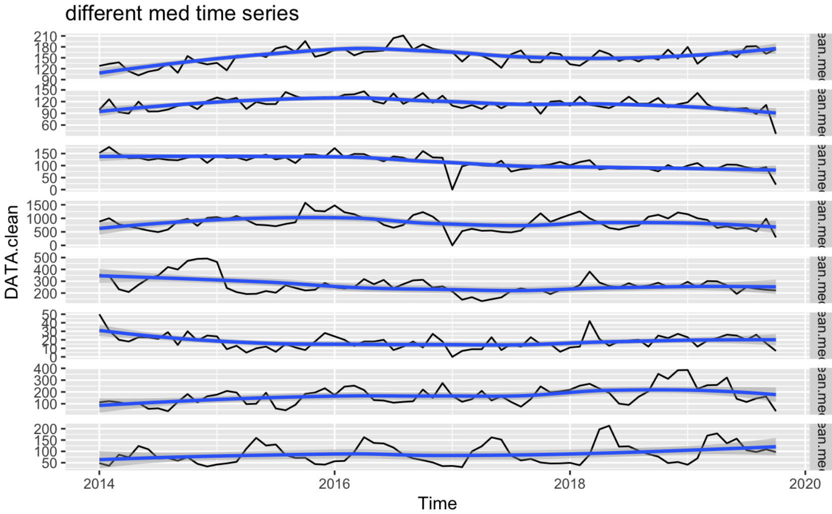
Code	Variable	Description
Med 1	M01AB	Anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances
Med 2	M01AE	Anti-inflammatory and antirheumatic products, non-steroids, Propionic acid derivatives
Med 3	N02BA	Anti-inflammatory and antirheumatic products, non-steroids, Propionic acid derivatives
Med 4	N02BE	Other analgesics and antipyretics, Pyrazolones and Anilides
Med 5	N05B	Psycholeptics drugs, Anxiolytic drugs
Med 6	N05C	Psycholeptics drugs, Hypnotics and sedatives drugs
Med 7	R03	Drugs for obstructive airway diseases
Med 8	R06	Antihistamines for systemic use

2 Data Cleaning

Data cleaning is conducted by replacing the missing values using tsoutliers() and tsclean().

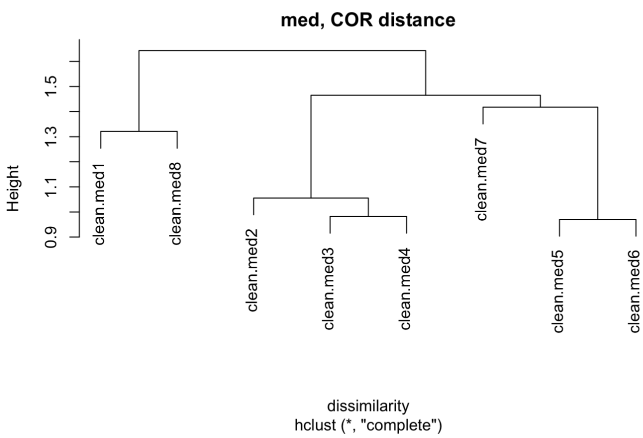


Before: Missing values & Outliers



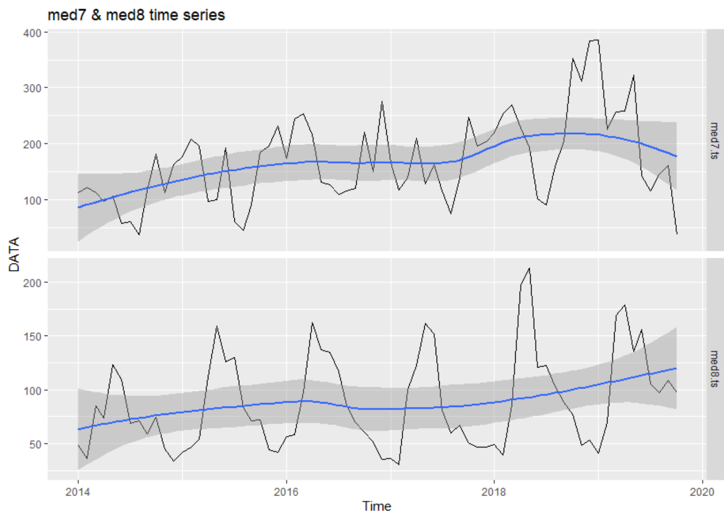
Cleaned Data Time Series

3 Similarity Check Through Correlation Distance

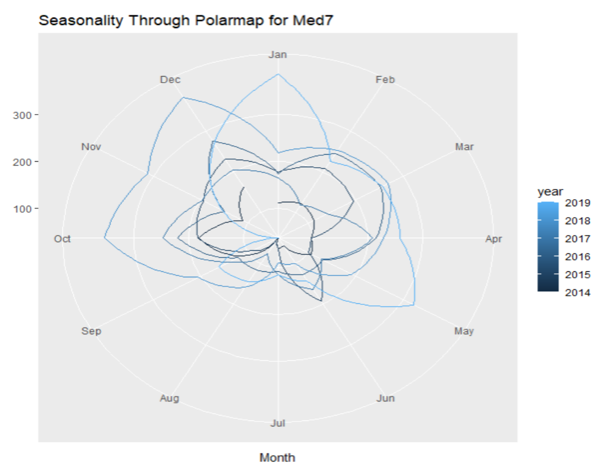


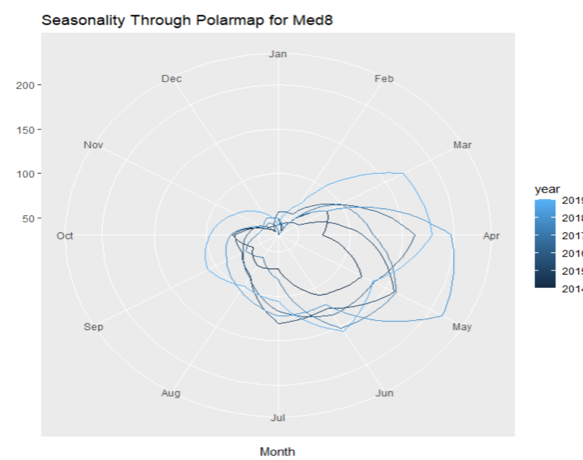
4 Data Exploration

The figure below shows an upward trend for both Medicine 7 and Medicine 8. The waves are getting bigger indicating a multiplicative model should be considered. Both of them show seasonality with their synonymous cycles being about one year. However, Medicine 8 shows stronger seasonality than Medicine 7. The rises and falls of Medicine 7 and 8 are roughly opposite with each other.



Polar maps below show clear color gradients, darker wires on the inside and lighter color ones on the outside, indicating a strong upward trend.





5 Data modeling and comparison

Below Table shows all 8 models created for Medicine 7 and some accuracy measures. The recommended models are the STL random walk and Best Arima. They both have tight forecast intervals and small MAPE/MASE values and there is no model violation. Also, they both have good performance in immediate forecasting which could help to predict the sales for the near future.

Model	MAPE	MASE	Ljung-Box Test	AIC	Forecast Interval
SNaive	Training: 30.86 Testing: 94.78	Training: 1 Testing: 1.76	p = 0.017 dependent residuals	N/A	256.92
Transformed SNaive	Training: 96.68 Testing: 94.77	Training: 3.18 Testing: 1.75	p = 0.022 dependent residuals	N/A	385.02
STL Random walk	Training: 22.11 Testing: 69.75	Training: 0.57 Testing:1.45	N/A	N/A	187.35
Transformed STL Random walk	Training: 96.24 Testing: 79.86	Training: 2.92 Testing: 1.47	N/A	N/A	286.39

Best ETS (MAM)	Training: 23.14 Testing: 94.55	Training: 0.56 Testing:1.61	p < 0.001 dependent residuals	682.28	239.51
Transformed (ETS) (ANA)	Training: 20.42 Testing: 83.62	Training: 0.58 Testing: 1.44	p < 0.001 dependent residuals	110.35	291.64
Best Arima ARIMA(0,0,1)(1,1,0)[12] with drift	Training: 20.29 Testing: 83.85	Training: 0.58 Testing: 1.47	p = 0.342 Independent	496.17	185.88
Transformed (Arima) ARIMA(0,0,1)(1,1,0)[12] with drift	Training: 20.95 Testing: 93.14	Training: 0.61 Testing: 1.50	p = 0.353 Independent	39.56	312.06

Below table shows all 10 models created for medicine 8 and some accuracy measures. The recommended model is the transformed ARIMA. The forecast intervals for the transformed ARIMA are tight and more meaningful for the purpose of this data set.

Model	MAPE	MASE	Ljung-Box Test	AIC	Forecast Intervals
SNaive	Training: 24.39 Testing: 23.35	Training: 1.00 Testing: 1.28	p = 0.23 independent residuals	N/A	106.95
Transformed SNaive	Training: 94.08 Testing: 23.35	Training: 3.99 Testing: 1.27	p = 0.005 dependent	N/A	1.18
STL Random walk	Training: 15.15 Testing: 22.77	Training: 0.60 Testing: 1.10	N/A	N/A	90.69
Transformed STL Random walk	Training: 93.92 Testing: 21.92	Training: 3.79 Testing: 1.12	N/A	N/A	0.825
Best ETS (M,N,M)	Training: 15.17 Testing: 20.03	Training: 0.59 Testing: 0.99	p = 0.007 dependent residuals	568.85	43.22
Transformed ETS (A,N,A)	Training: 14.24 Testing: 20.99	Training: 0.58 Testing: 1.04	p = 0.005 dependent residuals	67.77	47.07
Best Arima (0,0,0)(0,1,1) with drift	Training: 17.08 Testing: 19.82	Training: 0.64 Testing: 0.96	p = 0.57 independent residuals	427.8	90.69
Transformed Arima (1,0,0)(0,1,1)	Training: 13.99 Testing: 23.42	Training: 0.61 Testing: 1.27	p = 0.87 independent residuals	9.43	47.93
Neural	Training: 17.25 Testing: 22.36	Training: 0.69 Testing: 1.22	N/A	N/A	N/A
Bagging	Training: 13.41 Testing: 21.50	Training: 0.54 Testing: 1.07	N/A	N/A	N/A

6 Conclusion

This study has shown importance as it has provided insight as to where (in terms of the year) there's a higher patient demand for drugs. This can in turn allow pharmaceutical companies to have a better idea of how much inventory on hand they will need to stock and the amount of pharmaceuticals they will need to provide during different parts of the year.