Analysis and Forecasting of Pharmacy Drug Sales Data

Introduction

The data set for this project contains sales data of 8 drugs for a pharmacy from 2014-2019. The data set has 4 CSV files which contain hourly, daily, weekly and monthly sales data respectively. An accurate description of the data as well as links to the dataset have been provided in the appendix. The aim of this project is to make forecasts for future drug sales and provide insights that can help predict future sales and understand sales patterns of different drugs. The analysis and forecasting for this project will be done in R and the Rmarkdown file will also be attached in the submission.

Exploratory Data Analysis

Most of the data had already been processed but when exploring the data, we noticed issues that needed fixing. The data ends on 10/08/2019 which is the first day of the second week of October 2019. This means that we only we had 8 days of data for the month of October and because this would have a negative impact on our forecasting, we decided to remove the month of October for the year 2019 from the datasets. Additionally, in the monthly sales dataset, the month of January for the year 2017 had 0 for all the drug sales. This wasn't right so we summed up the weekly sales data for that period and inputted it into the monthly sales data.

We then computed basic statistics for the datasets such as the mean, median and standard deviation to gain some initial insights on the data. From this analysis, we noticed that in the hourly sales data most of the drug sales are 0. This makes sense as we don't expect the pharmacy to be making sales every single hour of the day. Therefore, we can conclude that it will be unnecessary to use the hourly data for forecasting. However, we may still be able to use it to gain an understanding of what hour of the day people are buying drugs.

Timeseries Analysis

We created a timeseries plot for each drug category.

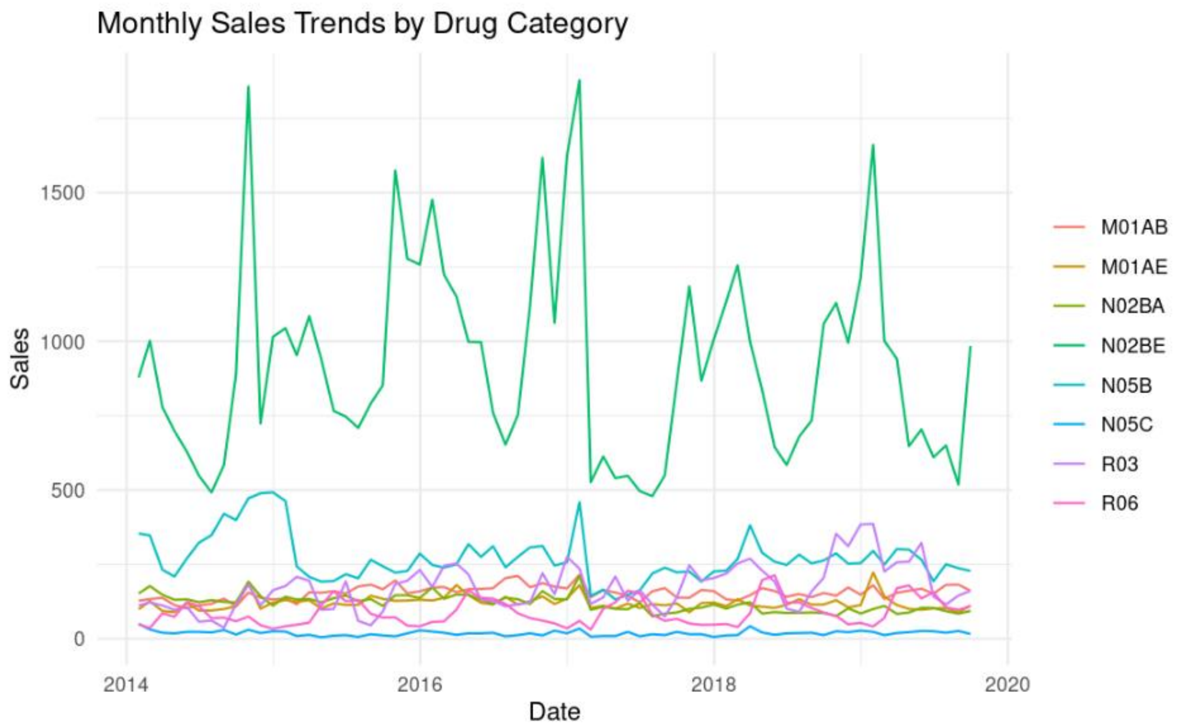


Figure 1: Monthly sales trends of each drug category

From figure 1 above, we can see that sales of N02BA are much greater than all other drugs. Additionally, N05C sales are very low when compared to other drugs. We will carry out seasonality analysis to gain more insights.

Seasonality Analysis

We investigated how the time of year affects sales of different drugs.

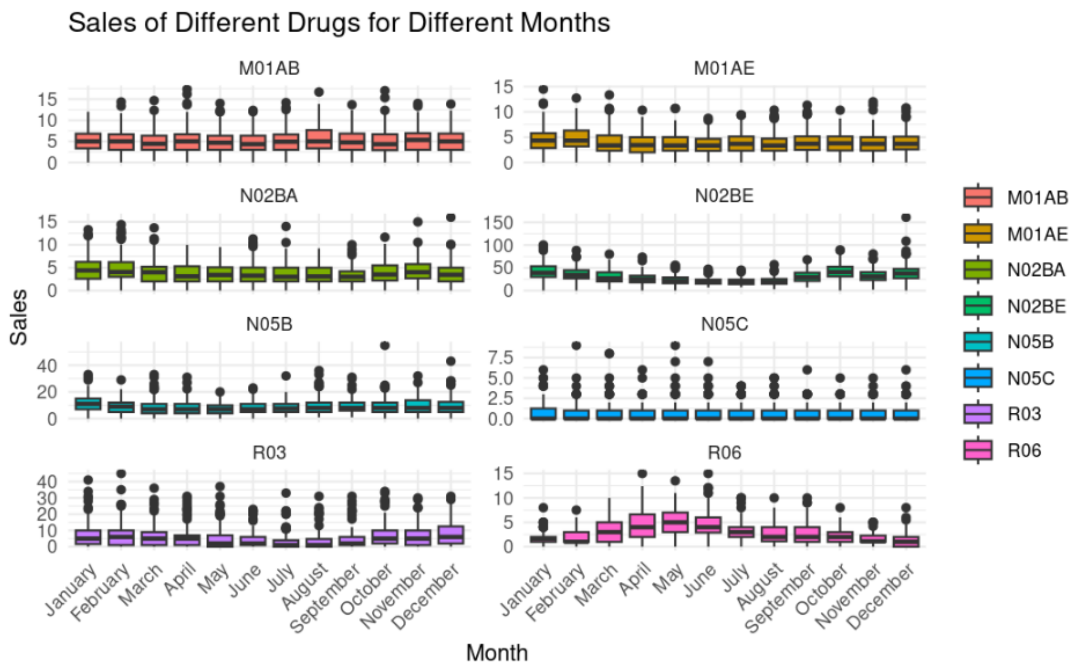


Figure 2: Box plots showing the sales of different drugs for different months

From figure 2, we get an idea of what drugs are popular in various months and this can aid the sales team when running adverts for drugs in different seasons. We notice that drug sales for most drugs are similar through the year. However, we see that N02BE, R03 and R06 have clear seasonal patterns. R06 for example is an antihistamine drug and it makes sense for it to be more popular in the spring months when more people are suffering from allergy reactions.

We also investigated the weekly seasonality but as we can see in figure 3 below, even though there is some seasonality for specific drugs, it is not very strong.

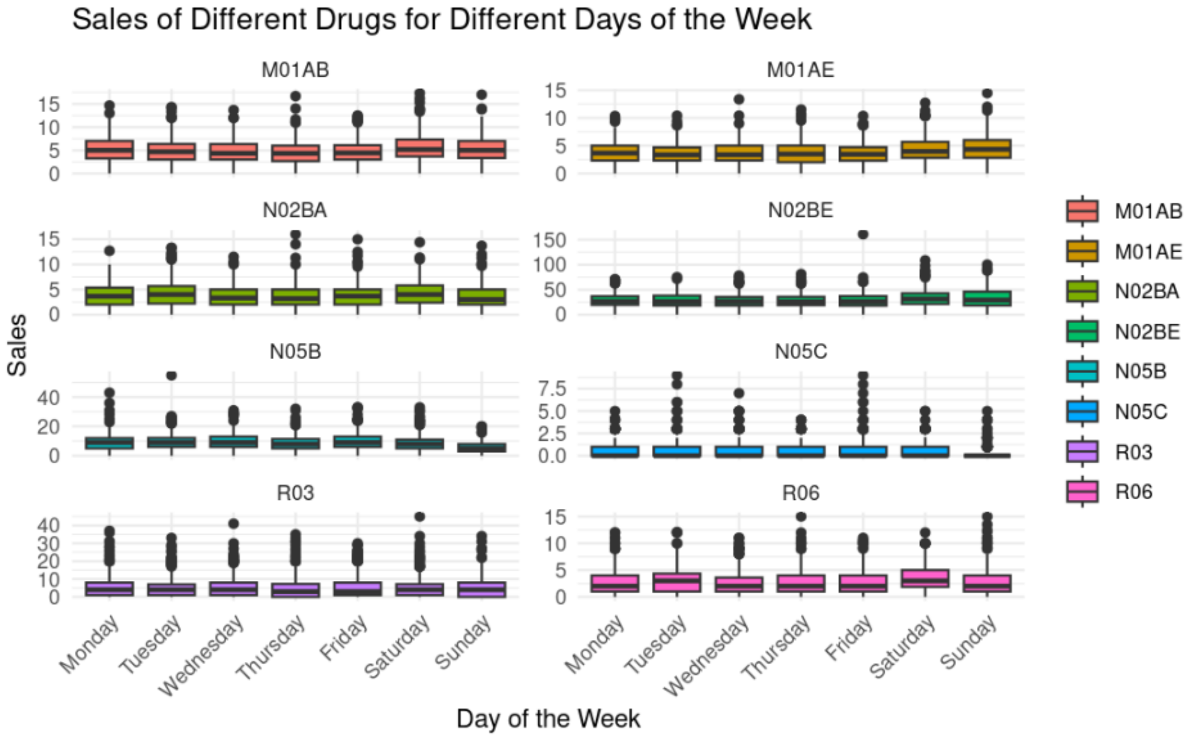


Figure 3: Box plots showing the sales of different drugs for different days of the week

Finally, we decided to look at what time of the day people are buying drugs. We originally looked at a boxplot but due to the number of outliers, we decided bar plots would represent the data better.

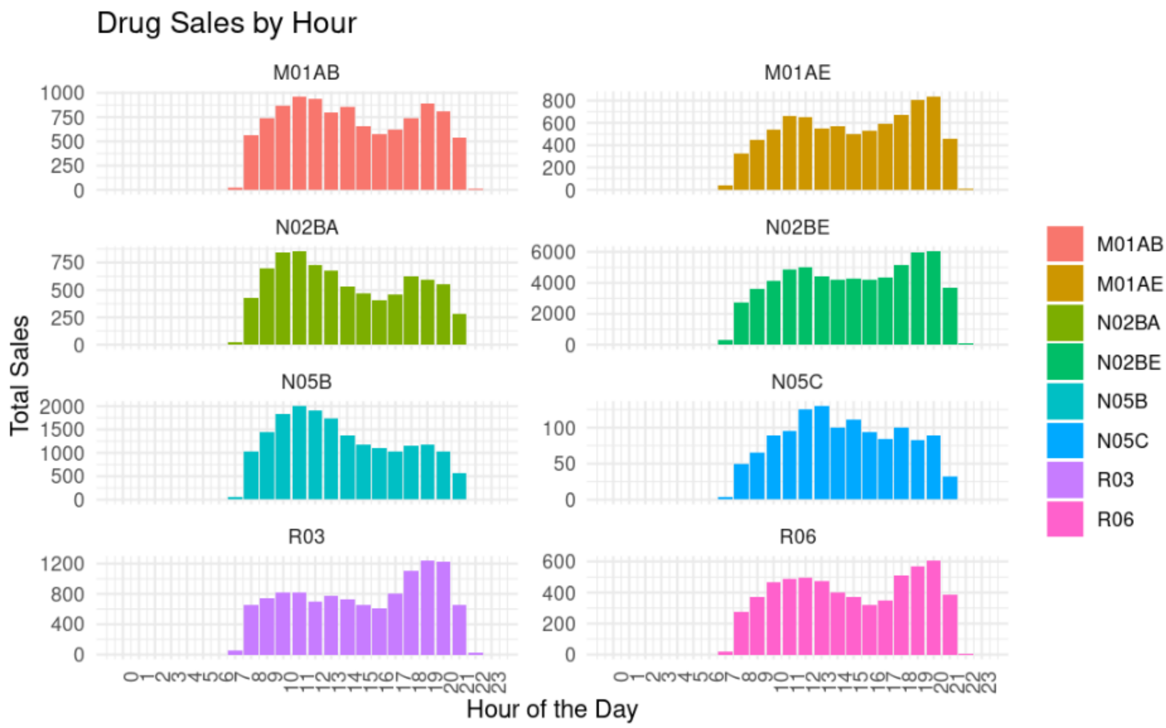
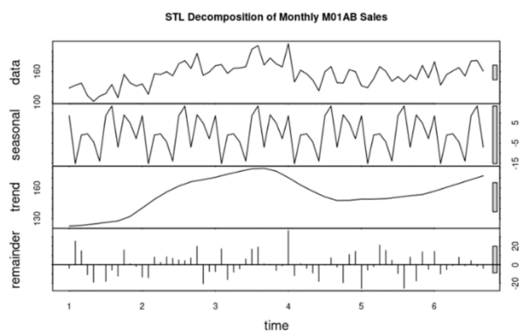
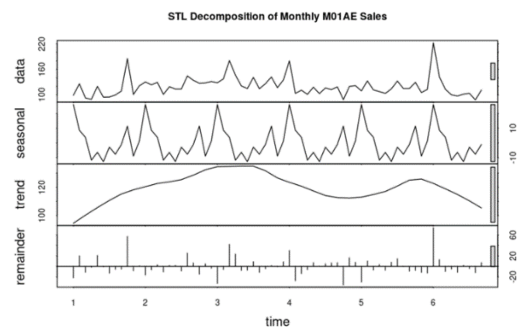


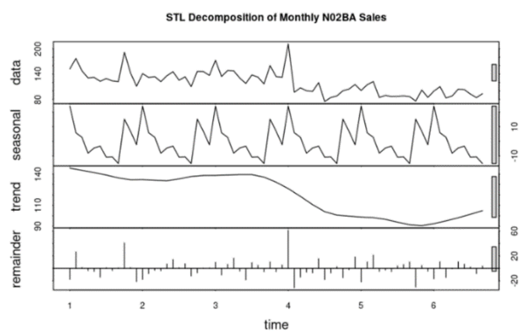
Figure 4: Bar plots showing the sales of drugs at different hours of the day

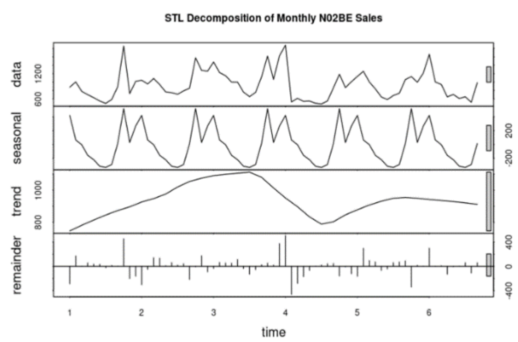
We can see strong seasonality in this data. A lot of drugs have 2 peak sales periods. The first between 9-11am and the second between 6-8pm so it is important that shelves are adequately restocked before these peak periods.

To get a better understanding of the seasonality and trend components, we carried out STL decomposition of the timeseries data.









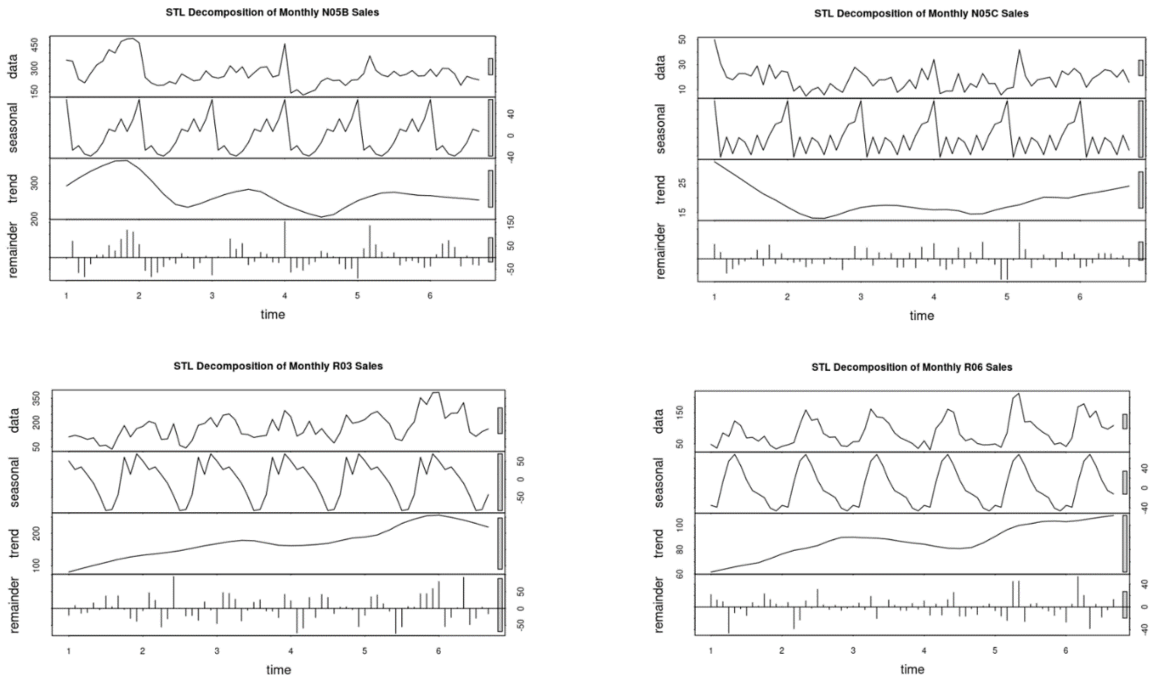


Figure 5: STL decomposition of sales data for each drug

From figure 5 above, we can see that N02BE, R03 and R06 have very strong seasonal components, while N05C, M01AB and N02BA have strong trend components. R03 and R06 also have strong trends but they aren't as relevant as the seasonality components for those drugs. We will take all these into account when forecasting.

Drug Sales Forecasting

For the sales forecasts, we have used the ETS and ARIMA forecasting models and compared both models for each drug to get the best forecasts. We have evaluated the best models between ETS and ARIMA based on their Root Mean Squared Errors (RMSE). The data set we used for most of our forecasts was the monthly data as this enabled us to do long term forecasts and gave the most accurate forecasts. To be sure monthly sales was the best data set to use, we also carried out forecasts for M01AB and R06 with the weekly sales data and compared the results to those obtained from using monthly data. To compare the results, we used the Mean Absolute Percentage Error (MAPE) instead of RMSE because of the different scales of monthly and weekly sales. When we made these comparisons, we discovered that forecasts with monthly data were more accurate. We will now discuss the forecast results for each drug category and the performance of the models.

Methodology For Conducting Forecasts

Step 1: Split data into training and test set. An 80:20 split was used in all cases as we found this to be the most ideal split.

Step 2: Specify the appropriate parameters for error, trend and seasonality for the ETS model and train it on the training set.

Step 3: Evaluate the performance of the ETS model based on the test set.

Step 4: Carry out stationary tests such as adf and kpss on the training set to ensure the time series is stationary.

Step 5: Use the auto.arima function to find the ARIMA model with the best parameters and train the training set with that model.

Step 6: Carry out residual analysis to ensure the residuals have an average close to 0 and a constant variance. We will also check the ACF plot to ensure that the residuals are all within the 95% confidence interval. This indicates that there is no residual autocorrelation in the model. We will also check the histogram plot to ensure that the residuals are around a mean of 0 and follow a bell-shaped curve which signifies a white noise process.

Step 7: Evaluate the performance of the ARIMA model based on the test set.

Step 8: Compare the RMSE of both models to choose the superior model.

Step 9: Recalibrate the entire sample with the superior model and produce the forecast results.

M01AB Results

We will go into detail when discussing the results for this drug to give a better understanding of the methodology used and for the other drugs, we will give more of a summary of our findings and results.

ETS model

For the ETS model, we specified Multiplicative, Additive and None (MAN) for the Error, Trend and Seasonality respectively. This was due to our analysis of the different components from the STL decomposition we carried out earlier. We then trained our model and evaluated its performance on the test. Plots of the model performance can be seen in figures 6 and 7 below.

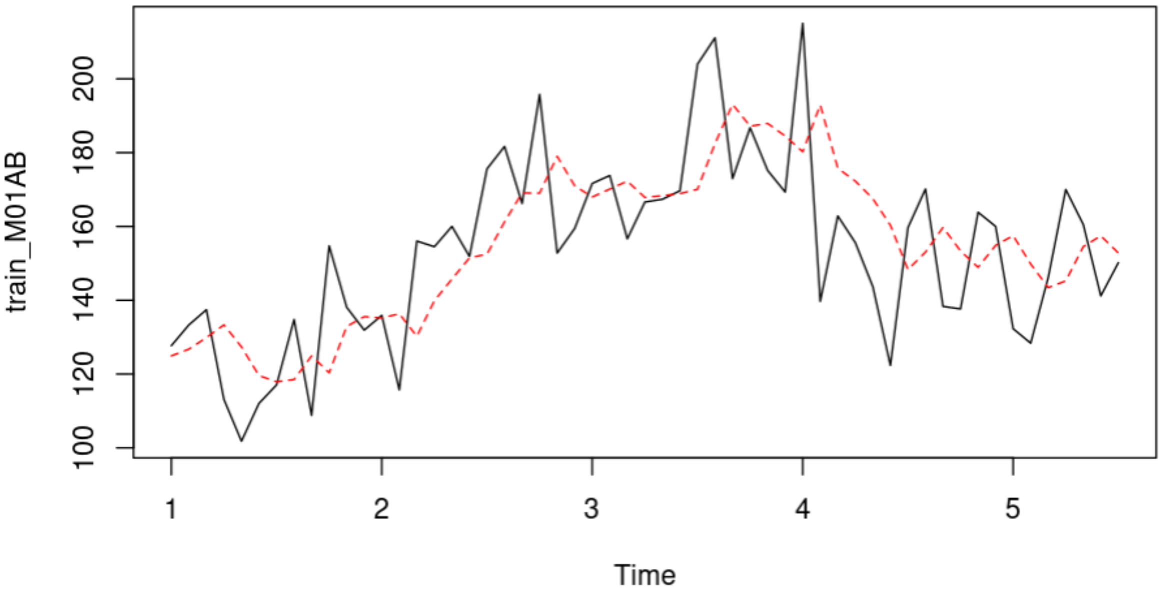


Figure 6: Performance of the ETS model on the training set

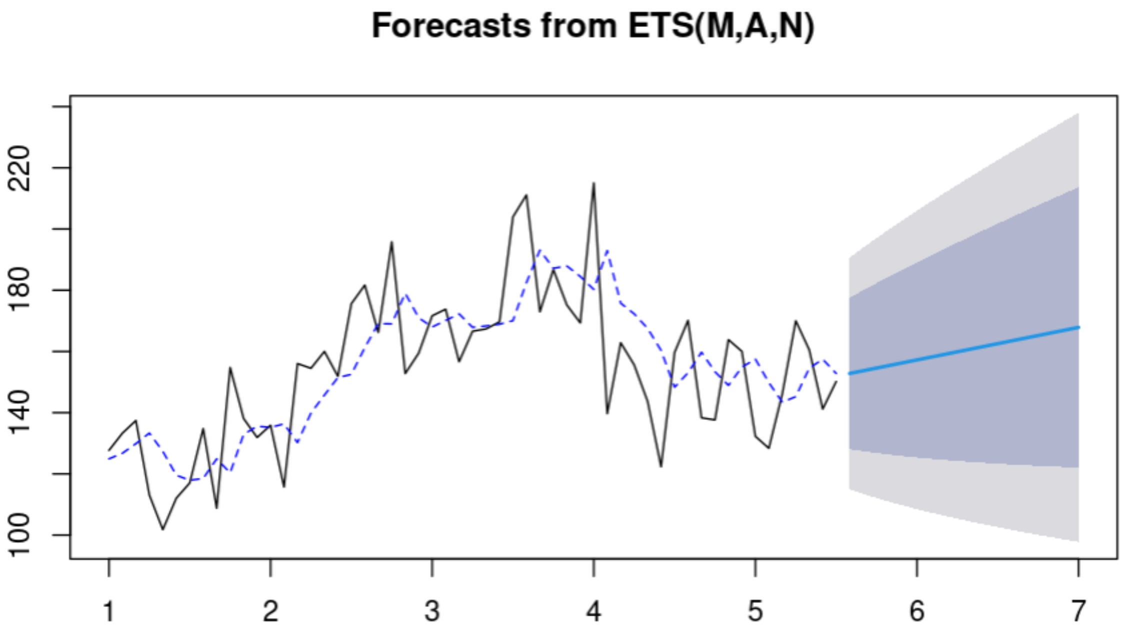


Figure 7: Forecasts made from the trained ETS model

ARIMA Model

We carried stationary tests and realized that the time series was not stationary and we would need to take first order differencing when specifying our model parameters. Using the auto.arima function, we got 1, 1, 0 for the parameters p, d, q respectively. We then trained the ARIMA model based on these parameters.

Next, we carried out residual analysis and saw that the model had no autocorrelation as it had constant variance and a mean around 0. The ACF plot also has lags within the 95% confidence interval. We can see the Plots from the residual analysis in figure 8 below.

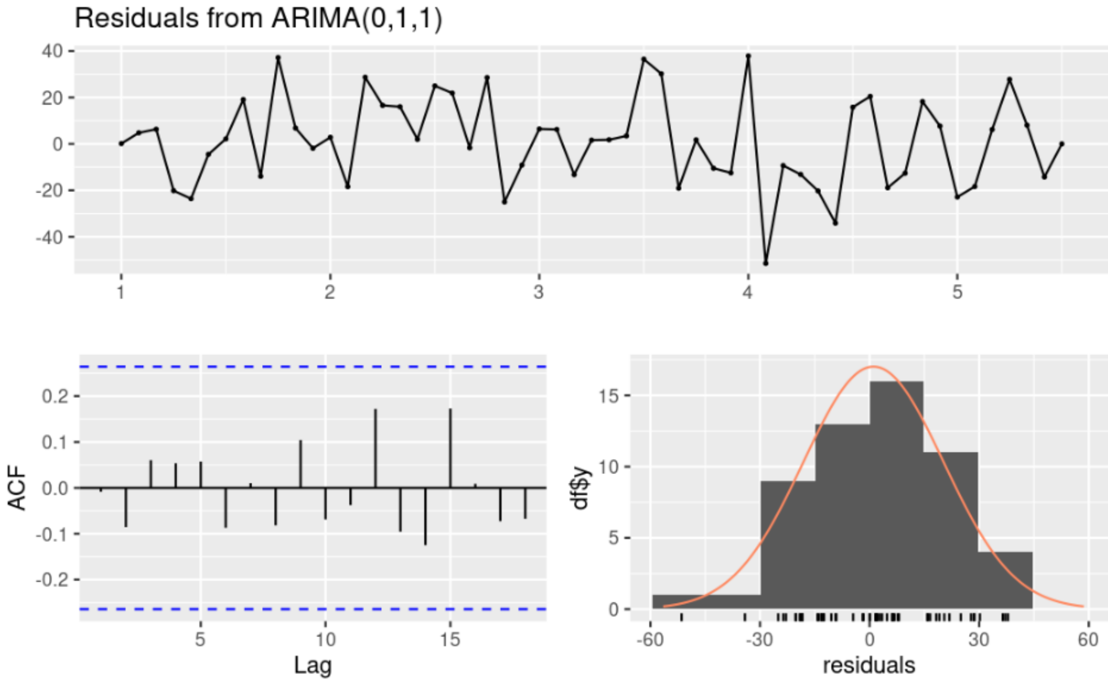


Figure 8: Residual, histogram and ACF plots showing the analysis of the residuals

We can now check the performance of the trained ARIMA model and use it to make forecasts. In figure 9 and 10 below we can see the plots for evaluation.

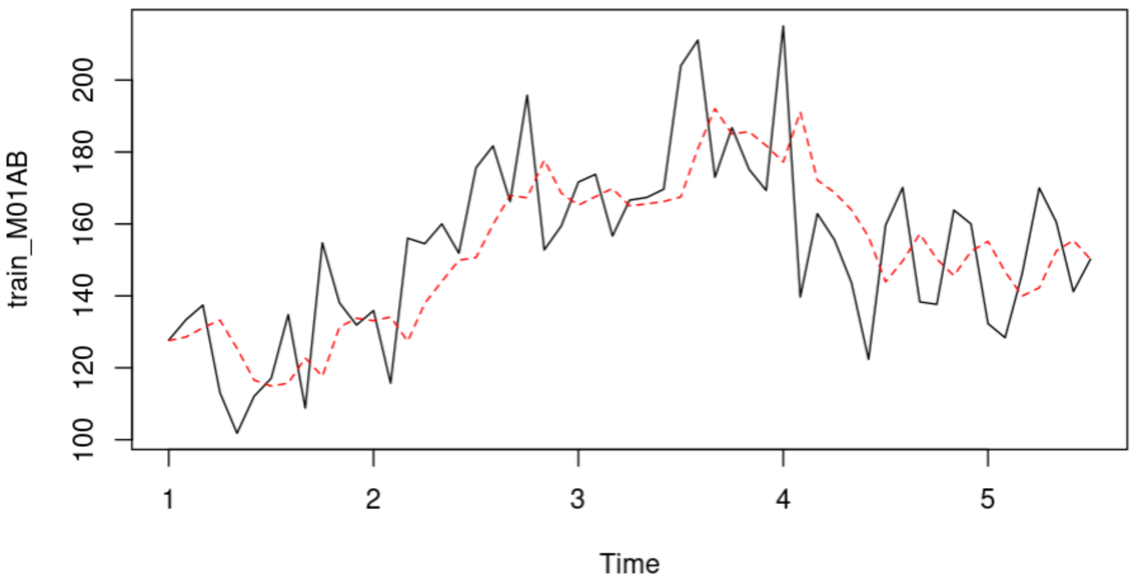


Figure 9: Performance of the ARIMA model on the training set

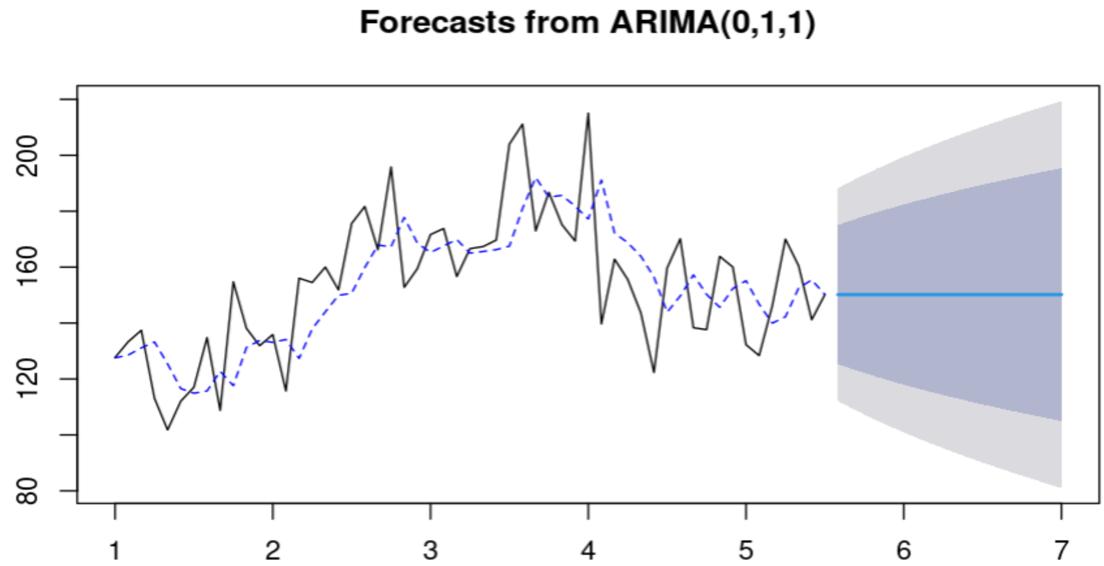


Figure 10: Forecasts made from the trained ARIMA model

We compare the accuracy and performance of both models in the table below.

		RMSE	MAPE
ARIMA Model	**Training Set**	19.04506	10.012625
	Test Set	17.56851	8.497004
ETS Model	**Training Set**	19.04243	10.128920
	Test Set	13.59778	7.117033

Table 1: Performance of both models

Model performance is evaluated on the RMSE. Note that we are comparing the RMSE on the test set as we are more concerned with the performance of our model against future data. From table, the ETS model is superior. Now, we will re-calibrate the ETS model with the whole data set and make forecasts for the next 18 months.

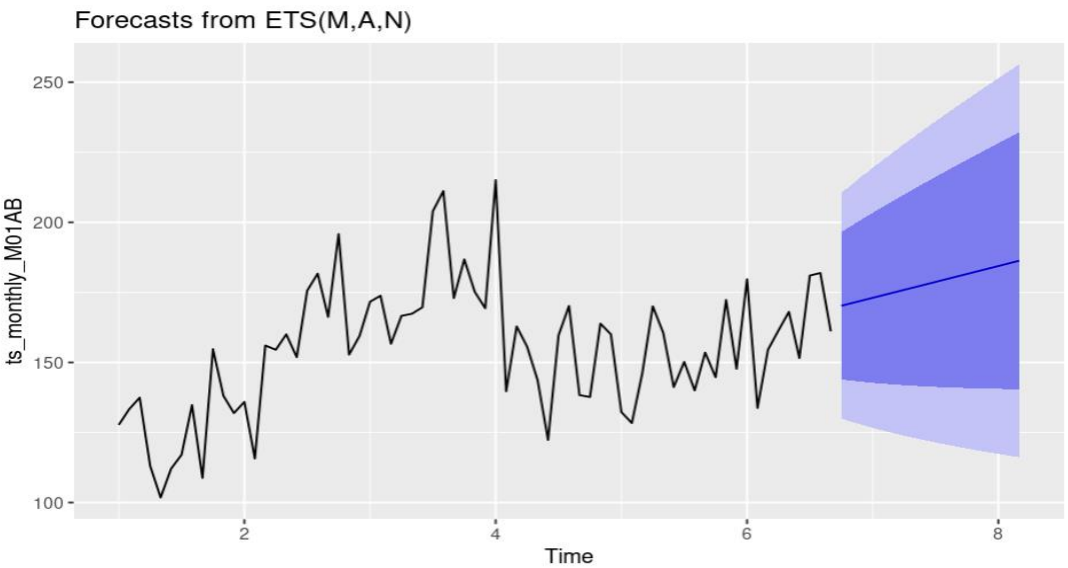


Figure 11: Forecasts made from the chosen model

```
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Oct 6       170.2129 143.9197 196.5061 130.0009 210.4249
Nov 6       171.1583 143.4707 198.8458 128.8138 213.5027
Dec 6       172.1036 143.0726 201.1346 127.7046 216.5027
Jan 7       173.0490 142.7185 203.3794 126.6626 219.4354
Feb 7       173.9943 142.4029 205.5858 125.6794 222.3093
Mar 7       174.9397 142.1212 207.7582 124.7481 225.1312
Apr 7       175.8850 141.8697 209.9004 123.8630 227.9071
May 7       176.8304 141.6452 212.0156 123.0192 230.6416
Jun 7       177.7758 141.4450 214.1065 122.2126 233.3389
Jul 7       178.7211 141.2668 216.1754 121.4397 236.0025
Aug 7       179.6665 141.1086 218.2243 120.6974 238.6356
Sep 7       180.6118 140.9687 220.2549 119.9829 241.2407
Oct 7       181.5572 140.8455 222.2689 119.2941 243.8203
Nov 7       182.5026 140.7377 224.2674 118.6288 246.3764
Dec 7       183.4479 140.6441 226.2518 117.9851 248.9108
Jan 8       184.3933 140.5635 228.2231 117.3614 251.4252
Feb 8       185.3386 140.4950 230.1822 116.7563 253.9210
Mar 8       186.2840 140.4378 232.1302 116.1683 256.3997
```

Figure 12: Forecast results for the chosen model

In figure 11 and 12 above, we can see the forecasted results from our models for the next 18 months. We have included the point forecasts and the 80 and 95% error bands.

We also carried out forecasts using the weekly sales data to compare performance. However, 2 issues arose. Firstly, the frequency of our weekly data is 52 and R has issues computing the seasonality component when the data has a frequency above 24. Therefore, seasonality was ignored when building models. Secondly, when we compare the Mean Absolute Percentage Errors (MAPE) we can see that the monthly sales forecasts perform much better than the weekly sales forecasts. In table 2 below we can see a comparison of the ETS models of both.

		RMSE	MAPE
Weekly ETS Model	**Training Set**	7.810193	19.52706
	Test Set	6.925324	14.32051
Monthly ETS Model	**Training Set**	19.04243	10.128920
	Test Set	13.59778	7.117033

Table 2: Performance of weekly and monthly ETS models

From table 2 above, the RMSE of the weekly model is smaller but this is not a good indication of model performance because monthly sales are much higher than weekly sales so will have a higher RMSE due to scale. Therefore, MAPE is a better judge for the model performance. As we can see, the MAPE of the weekly model is double that of the monthly so the monthly forecasts are better.

M01AE Result

After performing analysis on this drug, we concluded that the superior model was the ETS model with parameters “ANN”. In the table below we can see the comparison of their performances.

		RMSE	MAPE
ARIMA Model	**Training Set**	20.36952	12.24512
	Test Set	31.68497	16.50110
ETS Model	**Training Set**	20.15961	11.75428
	Test Set	31.64920	13.96700

Table 3: Performance of both models

In the figures below we can see the forecast results for the chosen model.

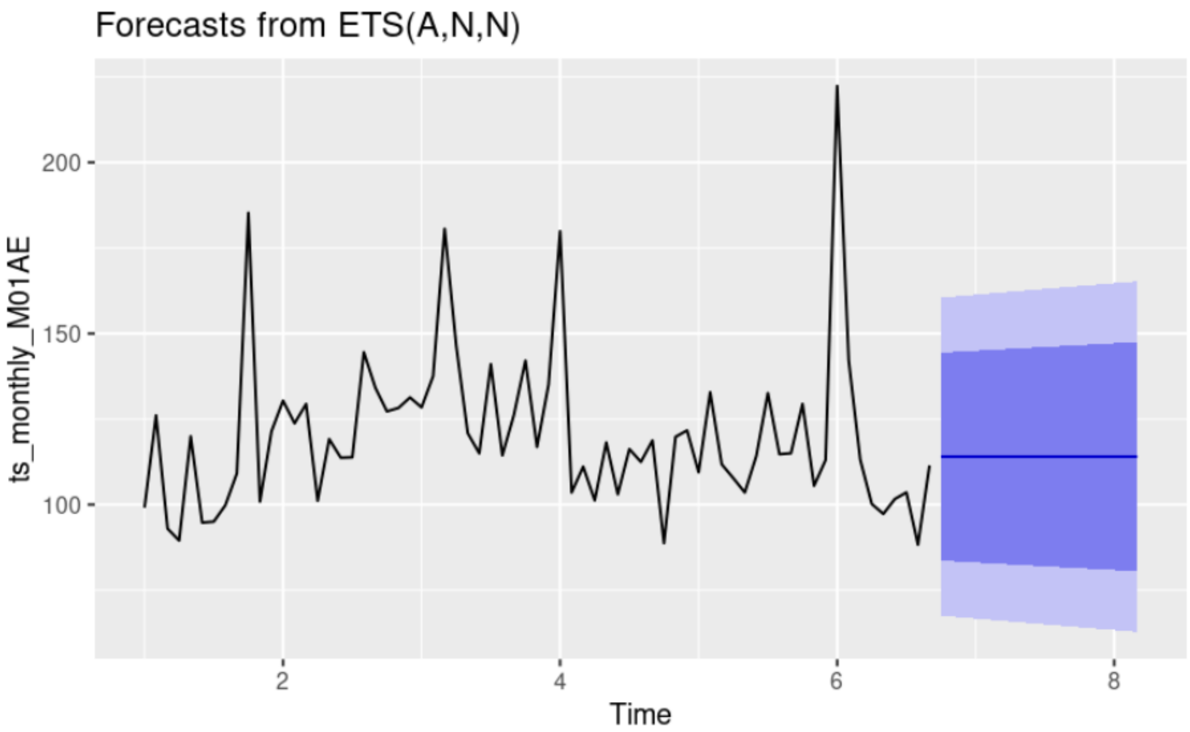


Figure 13: Forecasts made from the chosen model

```
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Oct 6       114.0009 83.62463 144.3771 67.54441 160.4573
Nov 6       114.0009 83.43268 144.5691 67.25085 160.7509
Dec 6       114.0009 83.24192 144.7598 66.95912 161.0426
Jan 7       114.0009 83.05235 144.9494 66.66919 161.3326
Feb 7       114.0009 82.86393 145.1378 66.38102 161.6207
Mar 7       114.0009 82.67664 145.3251 66.09459 161.9072
Apr 7       114.0009 82.49046 145.5113 65.80986 162.1919
May 7       114.0009 82.30538 145.6964 65.52680 162.4750
Jun 7       114.0009 82.12137 145.8804 65.24538 162.7564
Jul 7       114.0009 81.93842 146.0633 64.96558 163.0362
Aug 7       114.0009 81.75651 146.2452 64.68737 163.3144
Sep 7       114.0009 81.57561 146.4261 64.41072 163.5910
Oct 7       114.0009 81.39573 146.6060 64.13560 163.8662
Nov 7       114.0009 81.21682 146.7849 63.86199 164.1398
Dec 7       114.0009 81.03889 146.9629 63.58987 164.4119
Jan 8       114.0009 80.86192 147.1398 63.31921 164.6825
Feb 8       114.0009 80.68588 147.3159 63.04999 164.9518
Mar 8       114.0009 80.51077 147.4910 62.78218 165.2196
```

Figure 14: Forecast results for the chosen model

N02BA Result

After performing analysis on this drug, we concluded that the superior model was the ETS model with parameters “MNN”. In the table below we can see the comparison of their performances.

		RMSE	MAPE
ARIMA Model	**Training Set**	22.144213	13.921165
	Test Set	9.921671	9.081432
ETS Model	**Training Set**	22.986424	13.705283
	Test Set	9.530664	8.802342

Table 4: Table showing performance of both models

In the figures below we can see the forecast results for the chosen model.

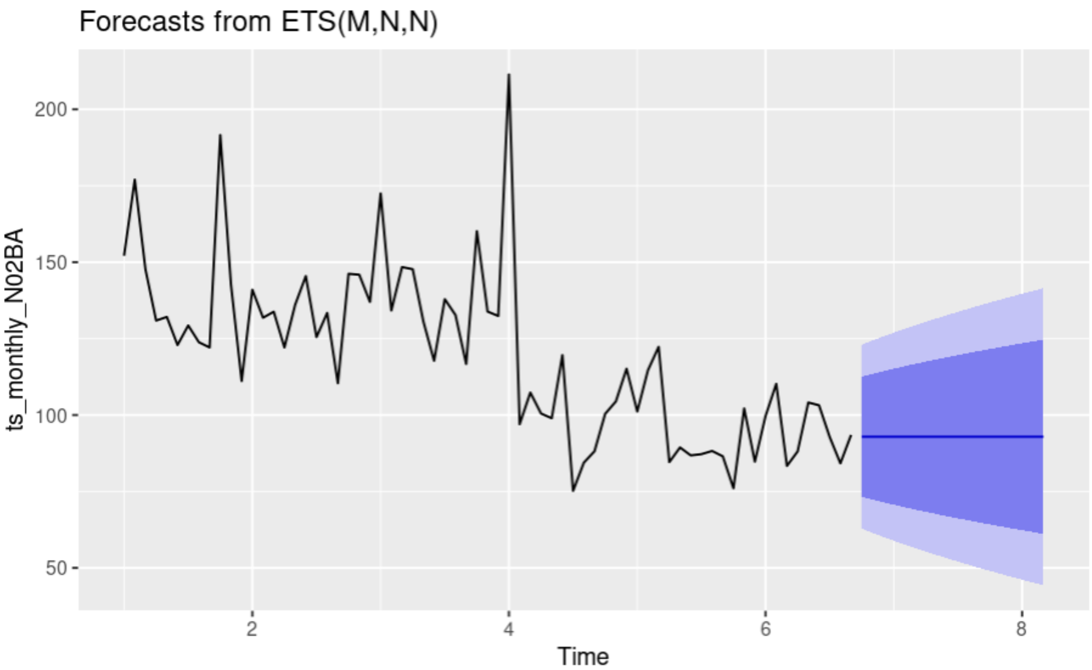


Figure 15: Forecasts made from the chosen model

```
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Oct 6       92.90884 73.22991 112.5878 62.81251 123.0052
Nov 6       92.90884 72.33806 113.4796 61.44855 124.3691
Dec 6       92.90884 71.48124 114.3364 60.13815 125.6795
Jan 7       92.90884 70.65540 115.1623 58.87513 126.9425
Feb 7       92.90884 69.85719 115.9605 57.65439 128.1633
Mar 7       92.90884 69.08385 116.7338 56.47166 129.3460
Apr 7       92.90884 68.33301 117.4847 55.32335 130.4943
May 7       92.90884 67.60267 118.2150 54.20640 131.6113
Jun 7       92.90884 66.89110 118.9266 53.11814 132.6995
Jul 7       92.90884 66.19680 119.6209 52.05630 133.7614
Aug 7       92.90884 65.51844 120.2992 51.01884 134.7988
Sep 7       92.90884 64.85487 120.9628 50.00400 135.8137
Oct 7       92.90884 64.20506 121.6126 49.01019 136.8075
Nov 7       92.90884 63.56809 122.2496 48.03603 137.7817
Dec 7       92.90884 62.94313 122.8746 47.08024 138.7374
Jan 8       92.90884 62.32945 123.4882 46.14170 139.6760
Feb 8       92.90884 61.72638 124.0913 45.21938 140.5983
Mar 8       92.90884 61.13331 124.6844 44.31235 141.5053
```

Figure 16: Forecast results for the chosen model

N02BE Result

After performing analysis on this drug, we concluded that the superior model was the ARIMA(1,0,0)(1,1,0)[12] model. In the table below we can see the comparison of their performances.

		RMSE	**MAPE**
ARIMA Model	**Training Set**	226.8824	15.77439
	Test Set	126.4114	11.06551
ETS Model	**Training Set**	203.4455	14.25648
	Test Set	180.0802	15.05260

Table 5: Table showing performance of both models

In the figures below we can see the forecast results for the chosen model.

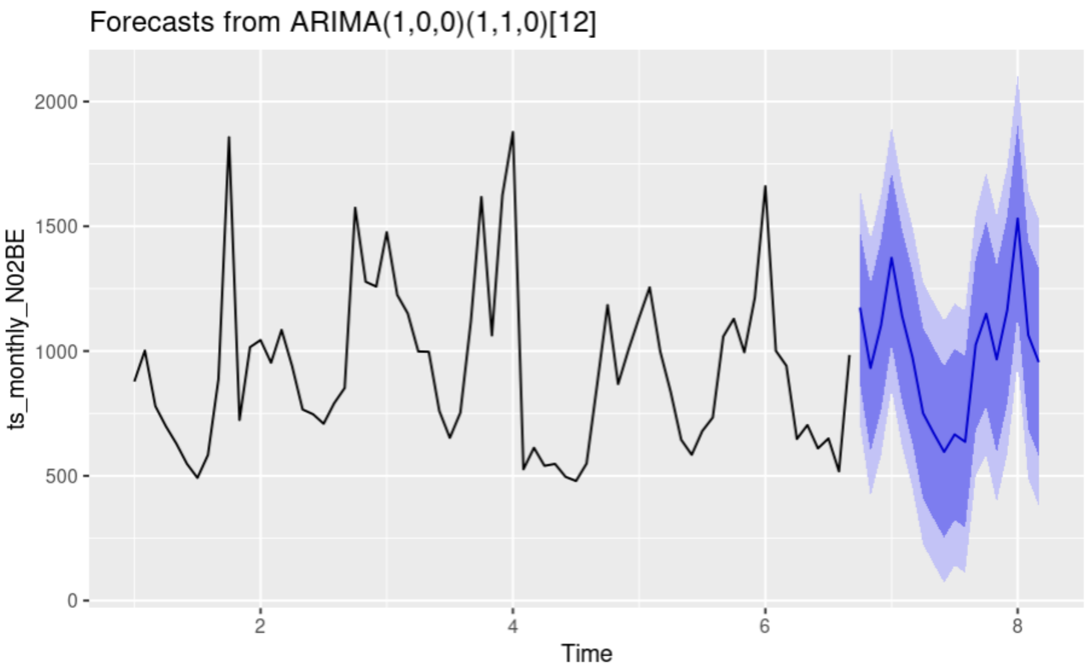


Figure 17: Forecasts made from the chosen model

```
       Point Forecast     Lo 80     Hi 80     Lo 95    Hi 95
Oct 6       1174.3256  869.1453 1479.5060 707.59258 1641.059
Nov 6        932.2102  597.0133 1267.4071 419.57080 1444.850
Dec 6       1103.7641  762.7009 1444.8273 582.15295 1625.375
Jan 7       1373.7354 1031.4739 1715.9968 850.29169 1897.179
Feb 7       1141.0624  798.5542 1483.5707 617.24135 1664.884
Mar 7        973.1473  630.5882 1315.7064 449.24837 1497.046
Apr 7        750.9693  408.3996 1093.5389 227.05423 1274.884
May 7        671.3515  328.7797 1013.9233 147.43314 1195.270
Jun 7        595.9735  253.4012  938.5457  72.05444 1119.892
Jul 7        665.9982  323.4259 1008.5705 142.07903 1189.917
Aug 7        636.2754  293.7031  978.8478 112.35622 1160.195
Sep 7       1024.8964  682.3240 1367.4688 500.97720 1548.816
Oct 7       1149.6504  780.3112 1518.9897 584.79476 1714.506
Nov 7        966.6858  592.0602 1341.3114 393.74537 1539.626
Dec 7       1164.1183  788.4109 1539.8257 589.52347 1738.713
Jan 8       1530.8713 1154.9410 1906.8016 955.93556 2105.807
Feb 8       1064.4598  688.4836 1440.4361 489.45378 1639.466
Mar 8        955.5661  579.5804 1331.5519 380.54557 1530.587
```

Figure 18: Forecast results for the chosen model

N05B Result

After performing analysis on this drug, we concluded that the superior model was the ETS model with parameters "MNN". In the table below we can see the comparison of their performances.

		RMSE	MAPE
ARIMA Model	**Training Set**	72.05985	18.45279
	Test Set	37.41619	13.09619
ETS Model	**Training Set**	71.77178	18.44728
	Test Set	36.93779	12.94604

Table 6: Table showing performance of both models

In the figures below we can see the forecast results for the chosen model.

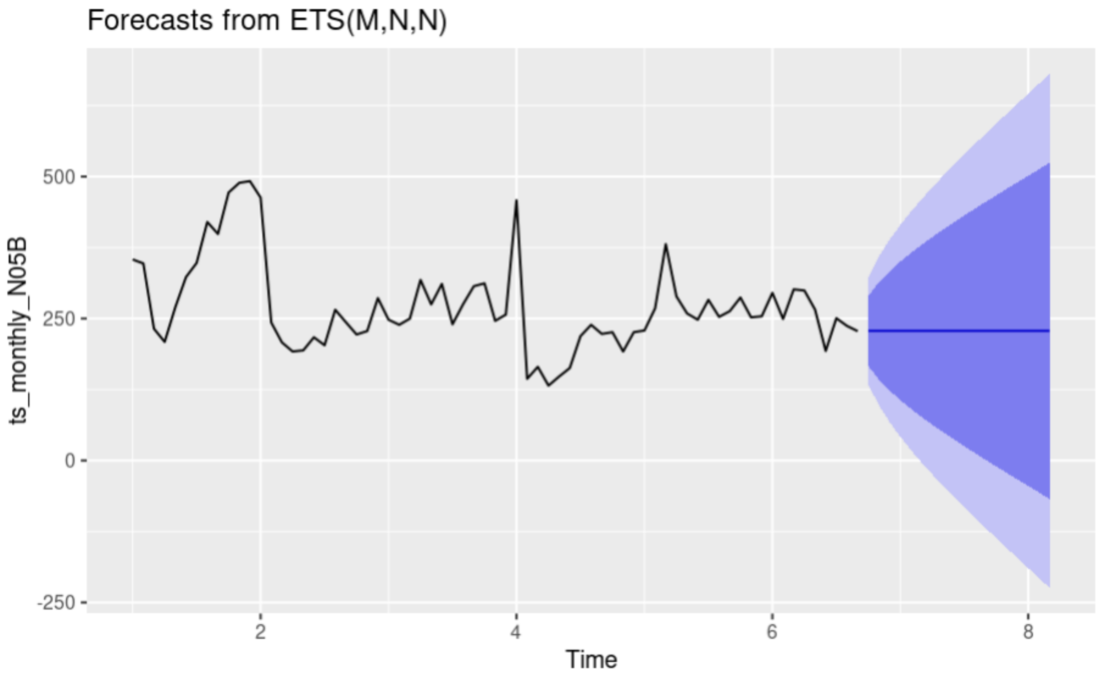


Figure 19: Forecasts made from the chosen model

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Oct 6	228.3472	166.644976	290.0494	133.981799	322.7126
Nov 6	228.3472	142.605533	314.0888	97.216643	359.4777
Dec 6	228.3472	123.292162	333.4022	67.679391	389.0150
Jan 7	228.3472	106.414788	350.2796	41.867675	414.8267
Feb 7	228.3472	91.050892	365.6435	18.370623	438.3238
Mar 7	228.3472	76.719806	379.9746	-3.546881	460.2413
Apr 7	228.3472	63.133586	393.5608	-24.325212	481.0196
May 7	228.3472	50.103565	406.5908	-44.252909	500.9473
Jun 7	228.3472	37.498066	419.1963	-63.531356	520.2257
Jul 7	228.3472	25.220733	431.4736	-82.307917	539.0023
Aug 7	228.3472	13.198395	443.4960	-100.694496	557.3889
Sep 7	228.3472	1.373792	455.3206	-118.778666	575.4730
Oct 7	228.3472	-10.299018	466.9934	-136.630688	593.3251
Nov 7	228.3472	-21.857688	478.5521	-154.308148	611.0025
Dec 7	228.3472	-33.333652	490.0280	-171.859121	628.5535
Jan 8	228.3472	-44.753585	501.4480	-189.324400	646.0188
Feb 8	228.3472	-56.140446	512.8348	-206.739102	663.4335
Mar 8	228.3472	-67.514263	524.2086	-224.133854	680.8282

Figure 20: Forecast results for the chosen model

N05C Result

After performing analysis on this drug, we concluded that the superior model was the ARIMA(1,0,0) with non-zero mean model. In the table below we can see the comparison of their performances.

		RMSE	MAPE
ARIMA Model	Training Set	8.606824	47.35359
	Test Set	5.626004	24.59408
ETS Model	Training Set	7.713228	44.68575
	Test Set	6.768376	27.90100

Table 7: Table showing performance of both models

In the figures below we can see the forecast results for the chosen model.

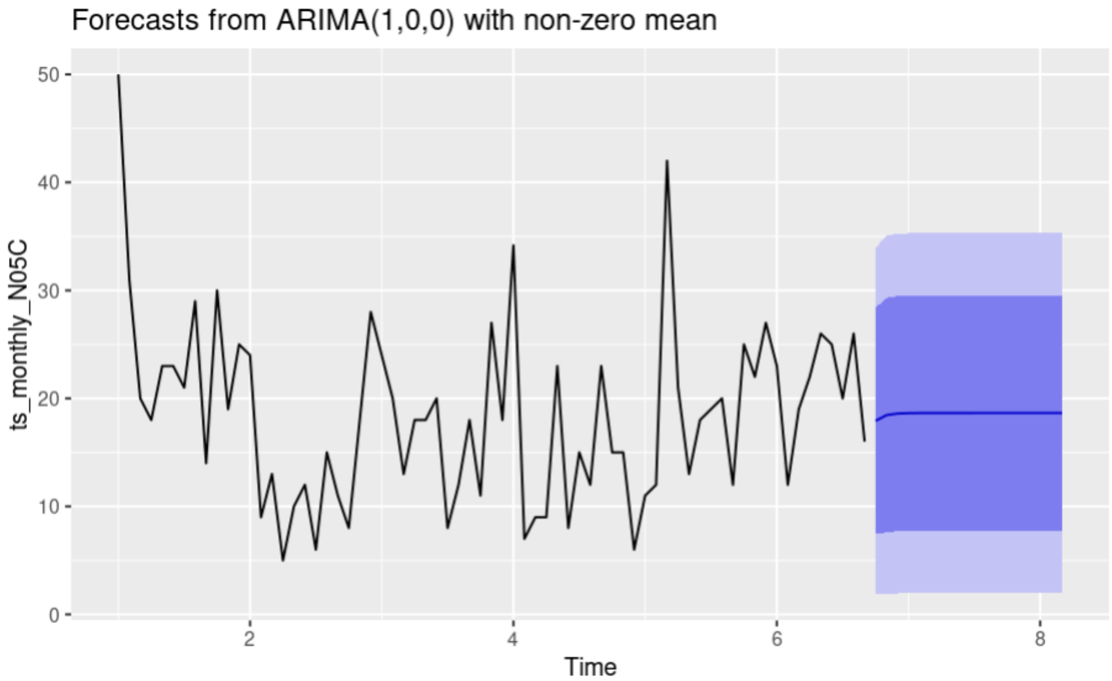


Figure 21: Forecasts made from the chosen model

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Oct 6	17.92488	7.455287	28.39448	1.913017	33.93675
Nov 6	18.45286	7.596564	29.30915	1.849589	35.05613
Dec 6	18.59768	7.712845	29.48251	1.950763	35.24459
Jan 7	18.63740	7.750424	29.52438	1.987206	35.28759
Feb 7	18.64830	7.761158	29.53543	1.997855	35.29874
Mar 7	18.65128	7.764134	29.53843	2.000825	35.30174
Apr 7	18.65210	7.764953	29.53925	2.001643	35.30256
May 7	18.65233	7.765178	29.53948	2.001868	35.30279
Jun 7	18.65239	7.765239	29.53954	2.001930	35.30285
Jul 7	18.65241	7.765256	29.53956	2.001947	35.30287
Aug 7	18.65241	7.765261	29.53956	2.001951	35.30287
Sep 7	18.65241	7.765262	29.53956	2.001953	35.30287
Oct 7	18.65241	7.765263	29.53956	2.001953	35.30287
Nov 7	18.65241	7.765263	29.53956	2.001953	35.30287
Dec 7	18.65241	7.765263	29.53956	2.001953	35.30287
Jan 8	18.65241	7.765263	29.53956	2.001953	35.30287
Feb 8	18.65241	7.765263	29.53956	2.001953	35.30287
Mar 8	18.65241	7.765263	29.53956	2.001953	35.30287

Figure 22: Forecast results for the chosen model

R03 Result

After performing analysis on this drug, we concluded that the superior model was the ARIMA(2,0,0)(1,1,0)[12] with drift model. In the table below we can see the comparison of their performances.

		RMSE	MAPE
ARIMA Model	**Training Set**	39.30147	19.32243
	Test Set	71.97493	20.50164
ETS Model	**Training Set**	39.08761	22.17073
	Test Set	97.74442	28.91406

Table 8: Table showing performance of both models

In the figures below we can see the forecast results for the chosen model.

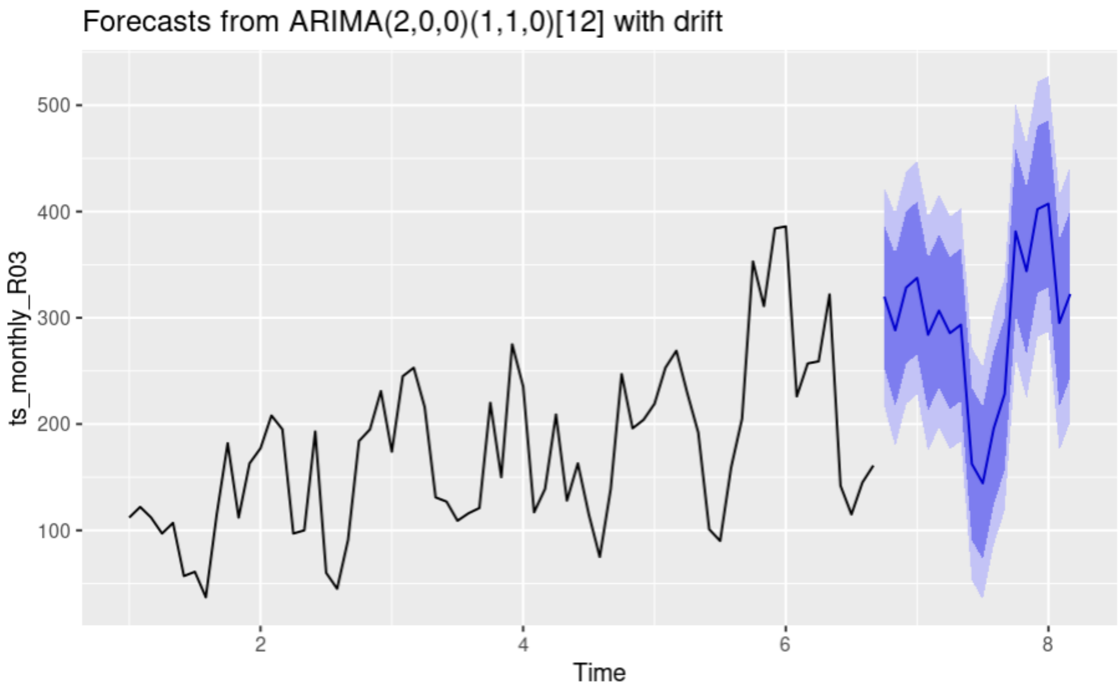


Figure 23: Forecasts made from the chosen model

```
      Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
Oct 6       319.9058 252.81775 386.9939 217.30348 422.5081
Nov 6       288.2598 216.77097 359.7487 178.92706 397.5926
Dec 6       328.4659 256.87964 400.0522 218.98416 437.9477
Jan 7       337.5099 265.92115 409.0987 228.02435 446.9955
Feb 7       284.0763 212.48570 355.6670 174.58791 393.5648
Mar 7       306.7200 235.12922 378.3108 197.23135 416.2086
Apr 7       285.5754 213.98460 357.1662 176.08674 395.0640
May 7       293.5015 221.91076 365.0923 184.01289 402.9902
Jun 7       162.5221  90.93128 234.1128  53.03341 272.0107
Jul 7       144.3349  72.74411 215.9257  34.84624 253.8235
Aug 7       195.8149 124.22412 267.4057  86.32625 305.3036
Sep 7       228.3379 156.74716 299.9287 118.84929 337.8266
Oct 7       381.2372 303.56128 458.9130 262.44216 500.0321
Nov 7       343.8885 265.42440 422.3526 223.88803 463.8890
Dec 7       402.1564 323.67444 480.6385 282.12857 522.1843
Jan 8       407.3208 328.83836 485.8033 287.29225 527.3494
Feb 8       295.1938 216.71098 373.6766 175.16469 415.2229
Mar 8       322.4399 243.95703 400.9227 202.41072 442.4690
```

Figure 24: Forecast results for the chosen model

R06 Result

After performing analysis on this drug, we concluded that the superior model was the ETS model with parameters “MAM”. In the table below we can see the comparison of their performances.

		RMSE	MAPE
ARIMA Model	**Training Set**	20.50096	16.51494
	Test Set	28.01761	20.54529
ETS Model	**Training Set**	16.66536	14.52099
	Test Set	27.90611	20.07944

Table 9: Table showing performance of both models

In the figures below we can see the forecast results for the chosen model.

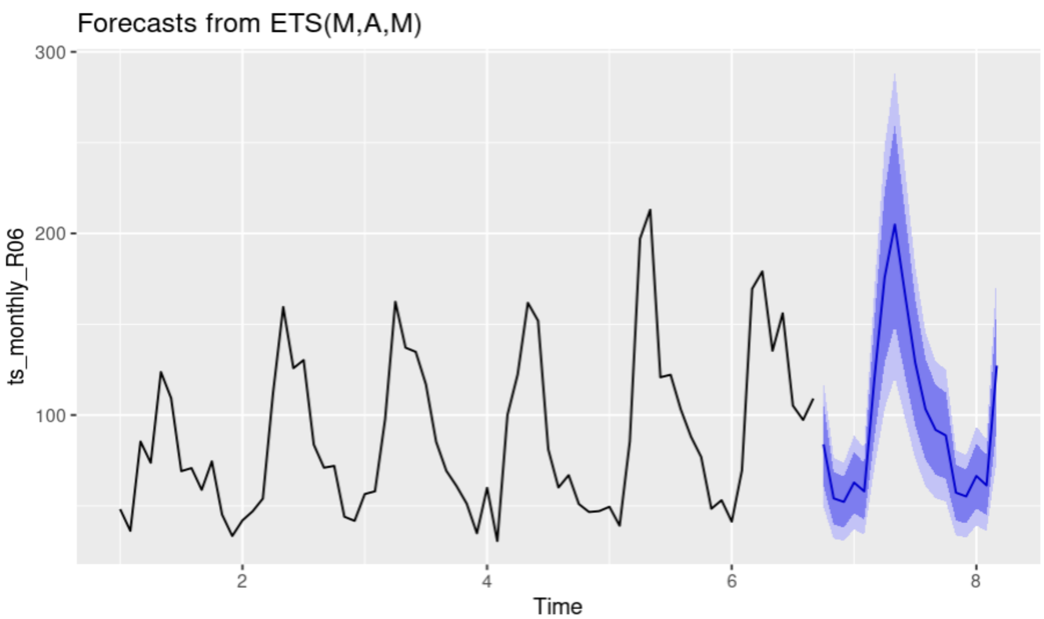


Figure 25: Forecasts made from the chosen model

```
      Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Oct 6       83.86868  61.38624 106.35111  49.48476 118.25259
Nov 6       54.12498  39.61585  68.63410  31.93518  76.31477
Dec 6       52.17722  38.19022  66.16422  30.78595  73.56849
Jan 7       62.84219  45.99626  79.68812  37.07856  88.60582
Feb 7       57.97110  42.43095  73.51126  34.20448  81.73772
Mar 7      120.39403  88.12032 152.66774  71.03565 169.75241
Apr 7      175.79329 128.66883 222.91776 103.72264 247.86394
May 7      204.90629 149.97755 259.83504 120.90003 288.91256
Jun 7      166.60724 121.94520 211.26927  98.30255 234.91193
Jul 7      128.93125  94.36891 163.49359  76.07271 181.78979
Aug 7      103.37272  75.66178 131.08367  60.99249 145.75295
Sep 7       91.83311  67.21555 116.45068  54.18380 129.48243
Oct 7       88.75425  64.96200 112.54650  52.36714 125.14136
Nov 7       57.26267  41.91231  72.61303  33.78632  80.73902
Dec 7       55.18745  40.39337  69.98153  32.56186  77.81305
Jan 8       66.45037  48.63701  84.26372  39.20719  93.69354
Feb 8       61.28375  44.85538  77.71211  36.15873  86.40877
Mar 8      127.24111  93.13145 161.35077  75.07489 179.40734
```

Figure 26: Forecast results for the chosen model

Conclusion

From our analysis, we conclude that daily and annual seasonality are important to determine what time of the day shelves should be stocked (before 9-11am and 6-8pm) and what time of the year certain drugs are popular (Winter months for N02BE and Spring months for R06). We tried to conduct hierarchical clustering to understand if the sales of certain drugs followed similar patterns but our findings from this were inconclusive. Additionally, we plotted the correlation matrix of the drug sales but from our findings, none of the drug sales were strongly correlated with one another.

Our forecasting gave us a good estimate of what future sales would look like. We could further improve our forecasting by building a multivariate model and including more explanatory variables in the model such as price of drugs or weather information which would have an impact on antirheumatic and antihistamine drugs.

Appendix

About the Dataset:

The dataset is built from the initial dataset consisted of 600000 transactional data collected in 6 years (period 2014-2019), indicating date and time of sale, pharmaceutical drug brand name and sold quantity, exported from Point-of-Sale system in the individual pharmacy. Selected group of drugs from the dataset (57 drugs) is classified to the following Anatomical Therapeutic Chemical (ATC) Classification System categories:

- M01AB - Anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances
- M01AE - Anti-inflammatory and antirheumatic products, non-steroids, Propionic acid derivatives
- N02BA - Other analgesics and antipyretics, Salicylic acid and derivatives
- N02BE/B - Other analgesics and antipyretics, Pyrazolones and Anilides
- N05B - Psycholeptics drugs, Anxiolytic drugs
- N05C - Psycholeptics drugs, Hypnotics and sedatives drugs
- R03 - Drugs for obstructive airway diseases
- R06 - Antihistamines for systemic use

  Sales data are resampled to the hourly, daily, weekly and monthly periods. Data is already pre-processed, where processing included outlier detection and treatment and missing data imputation.

Dataset and the about section were obtained from the following link on kaggle: https://www.kaggle.com/datasets/milanzdravkovic/pharma-sales-data

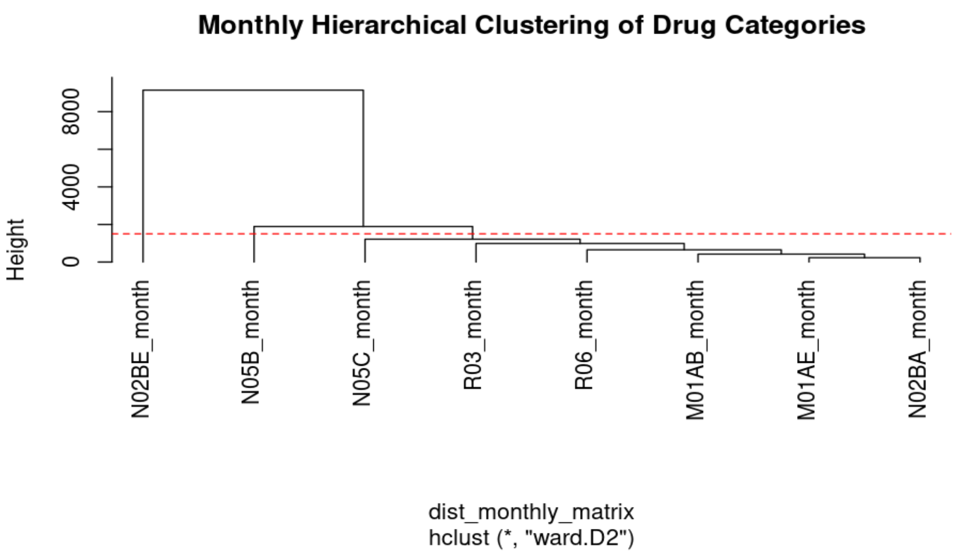


Figure 27: Hierarchical clustering of drug categories

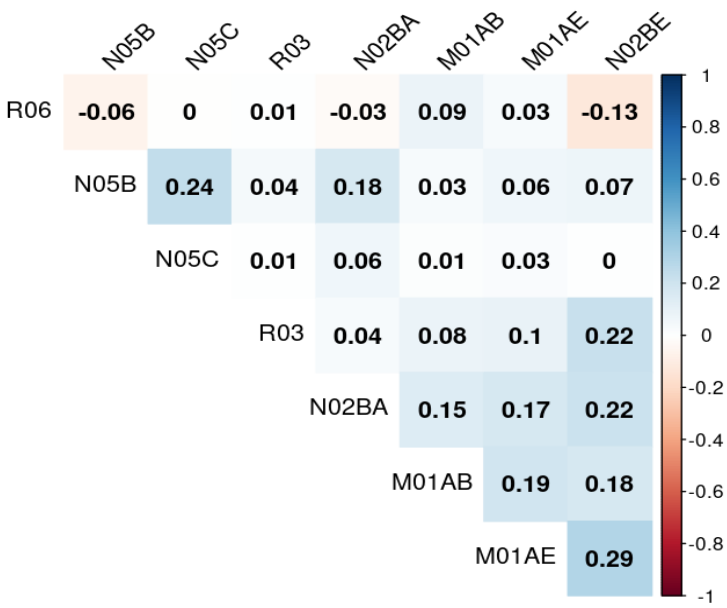


Figure 28: Correlation matrix of Drug Categories