Forecasting Future Dental Customers

Julie Bazalewski

Background

The file BestSmileDental.csv contains the number of patient/customer visits for a dental clinic "Best Smile Dental" for the past seven years. The data is rolled up monthly by year. Using forecasting techniques, the customer/patient count for the 12 months of 2008 can be predicted.

Data Cleansing

The first step is to clean the data. First, I used the as.integer() function to convert all numeric customer counts to integers, and transform entries which include non-numeric characters to NA.

I then called the summary() function again to determine if there were still incorrect numeric values. This returned a minimum of -999999 in the customer field, indicating incorrect values. An examination of the data also revealed several zero values, which appear to be outliers. Therefore, I converted all negative or zero values to NA as well. At this point, the summary() function returns values that appear to be reasonable and also indicates 10 NA values.

Next, I plotted the data using the aggr() function and the missmap() function from the Amelia package, in Figures 5 and 6, respectively. These figures confirm a total of 10 missing values. Figure 6 displays the positions of the missing values and indicates the values are missing at random.

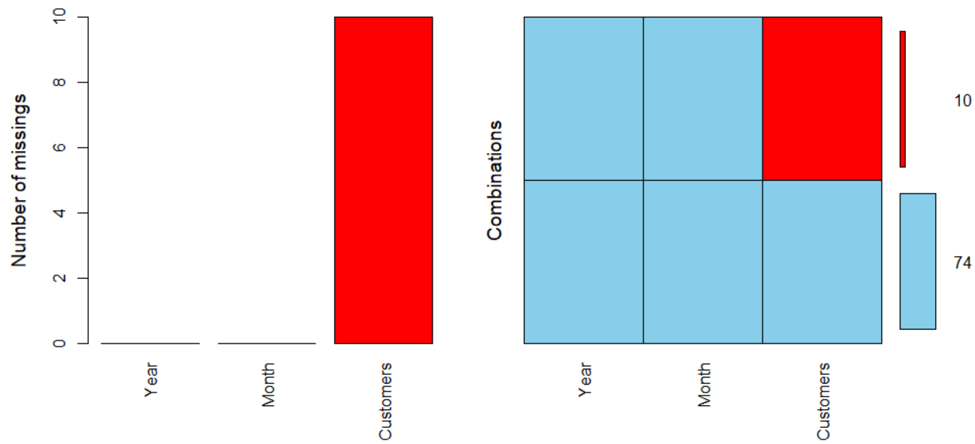

Figure 5: Missing Values in Dataset "dental"

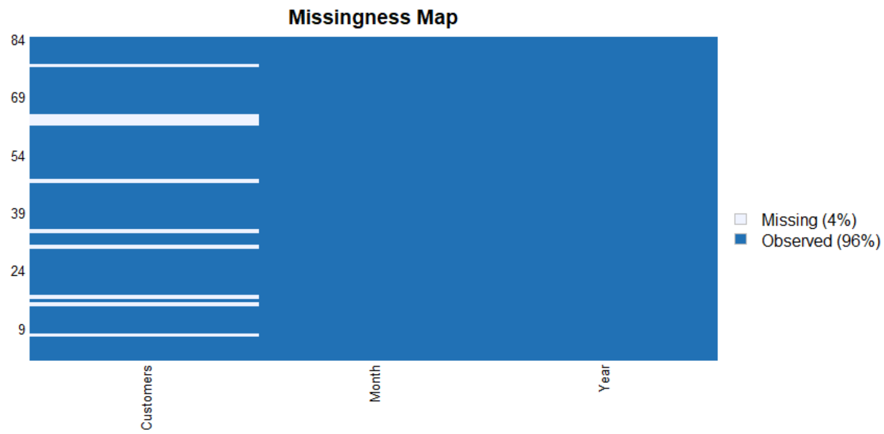

Figure 6: Missing Values in Dataset "dental"

I also plotted the data as a time series to inspect the data further in Figure 7. This figure confirms that the remaining data is reasonable. The missing values can be seen as well.

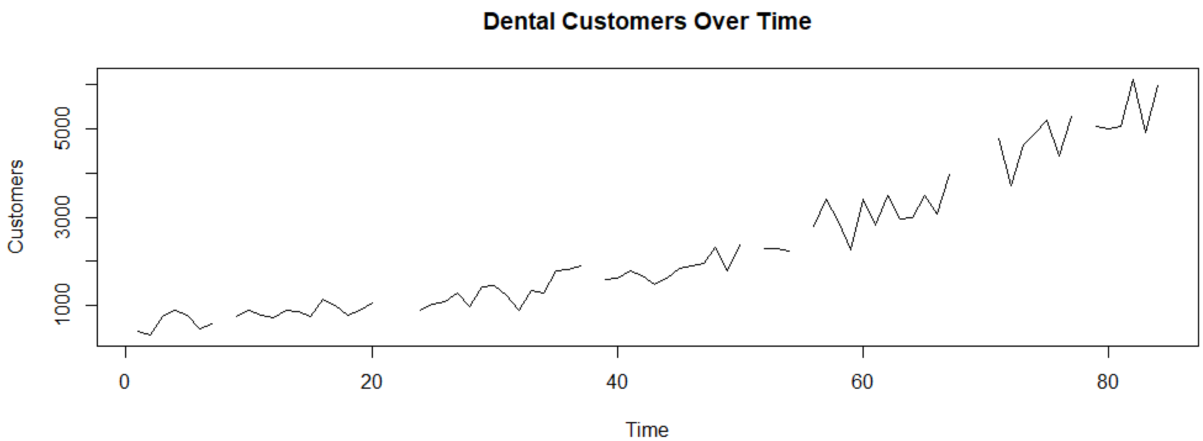

Figure 7: Missing Values in Dataset "dental"

Next, imputation is performed on the missing values. I used the MICE package (Multivariate Imputation by Chained Equations) to create three separate datasets with imputed models. I set a seed of 1234 so that my results are reproducible. I plotted each of the datasets with the remaining original time series data in Figure 8 to determine an appropriate set. The three sets are colored red, blue, and green, respectively. I chose the blue set, set 2, for the analysis because it tends to fall between the other two sets for most points. I appended the "where" column from the variable containing the MICE call (imp) to the variables containing the imputed data sets (created by using the complete() function on imp) to designate rows which were imputed. The imputed rows have a where value of TRUE.

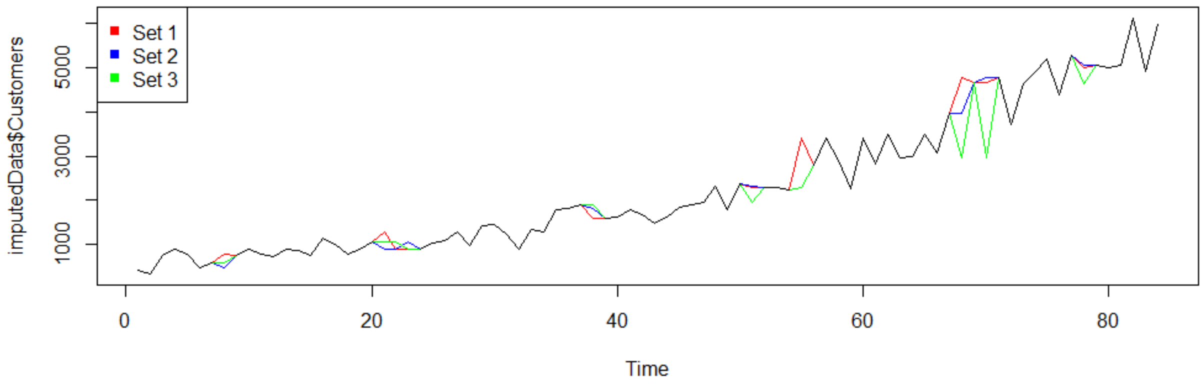

Figure 8: Imputed Data Using MICE

The imputed dataset from set 2 is saved in variable imputed2 and reproduced below:

	Year	Month	Customers	imputed		Year	Month	Customers	imputed
1	2001	1	416	FALSE	43	2004	7	1484	FALSE
2	2001	2	329	FALSE	44	2004	8	1634	FALSE
3	2001	3	750	FALSE	45	2004	9	1835	FALSE
4	2001	4	904	FALSE	46	2004	10	1893	FALSE
5	2001	5	794	FALSE	47	2004	11	1961	FALSE
6	2001	6	485	FALSE	48	2004	12	2321	FALSE
7	2001	7	584	FALSE	49	2005	1	1790	FALSE
8	2001	8	485	TRUE	50	2005	2	2361	FALSE
9	2001	9	750	FALSE	51	2005	3	2321	TRUE
10	2001	10	904	FALSE	52	2005	4	2289	FALSE
11	2001	11	794	FALSE	53	2005	5	2286	FALSE
12	2001	12	716	FALSE	54	2005	6	2244	FALSE
13	2002	1	893	FALSE	55	2005	7	2289	TRUE
14	2002	2	858	FALSE	56	2005	8	2799	FALSE
15	2002	3	742	FALSE	57	2005	9	3410	FALSE
16	2002	4	1133	FALSE	58	2005	10	2896	FALSE
17	2002	5	1015	FALSE	59	2005	11	2266	FALSE
18	2002	6	793	FALSE	60	2005	12	3420	FALSE
19	2002	7	904	FALSE	61	2006	1	2816	FALSE
20	2002	8	1059	FALSE	62	2006	2	3482	FALSE
21	2002	9	904	TRUE	63	2006	3	2967	FALSE
22	2002	10	904	TRUE	64	2006	4	2995	FALSE
23	2002	11	1059	TRUE	65	2006	5	3498	FALSE
24	2002	12	893	FALSE	66	2006	6	3069	FALSE
25	2003	1	1022	FALSE	67	2006	7	3978	FALSE
26	2003	2	1083	FALSE	68	2006	8	3978	TRUE
27	2003	3	1281	FALSE	69	2006	9	4651	FALSE
28	2003	4	980	FALSE	70	2006	10	4761	TRUE
29	2003	5	1431	FALSE	71	2006	11	4761	FALSE
30	2003	6	1447	FALSE	72	2006	12	3726	FALSE
31	2003	7	1223	FALSE	73	2007	1	4642	FALSE
32	2003	8	908	FALSE	74	2007	2	4873	FALSE
33	2003	9	1338	FALSE	75	2007	3	5204	FALSE
34	2003	10	1294	FALSE	76	2007	4	4383	FALSE
35	2003	11	1775	FALSE	77	2007	5	5271	FALSE
36	2003	12	1809	FALSE	78	2007	6	5046	TRUE
37	2004	1	1908	FALSE	79	2007	7	5046	FALSE
38	2004	2	1809	TRUE	80	2007	8	5010	FALSE
39	2004	3	1596	FALSE	81	2007	9	5040	FALSE
40	2004	4	1622	FALSE	82	2007	10	6126	FALSE
41	2004	5	1776	FALSE	83	2007	11	4906	FALSE
42	2004	6	1682	FALSE	84	2007	12	5965	FALSE

Forecasting Analysis

Now that I have a complete data set, I can perform the forecasting analysis with Holt Winters and ARIMA models. First, I try three Holt Winters models using the HoltWinters() function. The three models contain the following parameters for alpha (smoothing constant), beta (trend), and gamma (seasonality):

1. alpha = 0.2, beta= FALSE, gamma=FALSE
2. alpha = 0.3, beta= TRUE, gamma=FALSE
3. alpha is determined by R, beta is determined by R, gamma=FALSE

For each of the three models, I create the model with the HoltWinters() function, predict the next 12 steps with predict(), and save the accuracy measures to a variable with accuracy().

Next, I print out and examine the predictions with upper and lower bounds, the fitted means, and smoothing parameters and coefficients for each model.

For model 3, R computed the following parameters:

```
Smoothing parameters:
 alpha: 0.2616823
 beta : 0.1529528
 gamma: FALSE
```

Finally, I plotted the fitted models and forecasts for each model, included in Figures 9, 10, and 11. The fitted model is plotted in green along with the original data in black. The red lines represent the upper and lower bounds of the next twelve predictions, and the blue lines are the predicted values. From these plots, Model 3 appears to be the best fit.

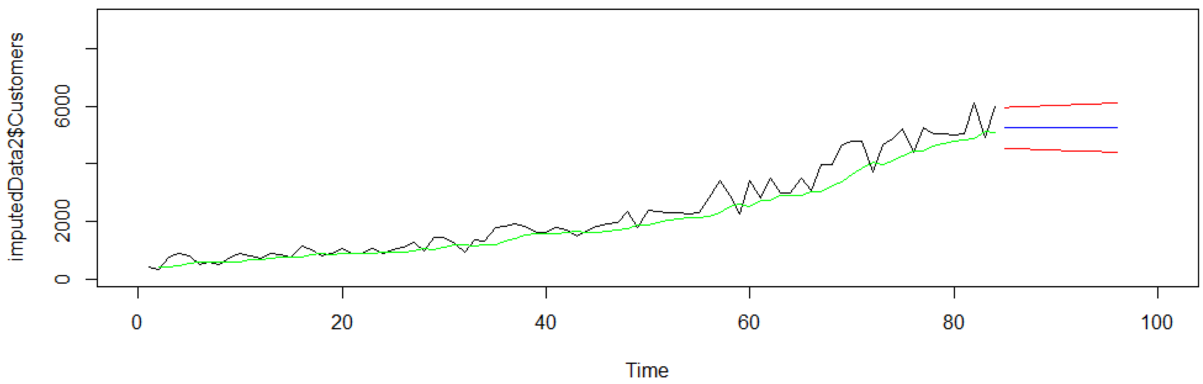

Figure 9: Holt Winters Forecast for alpha: 0.2, beta: FALSE, gamma: FALSE

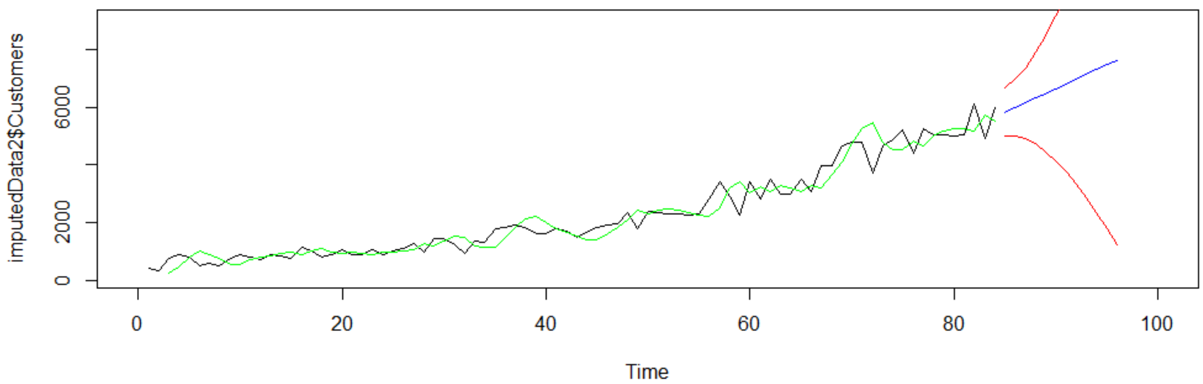

Figure 10: Holt Winters Forecast for alpha: 0.3, beta: TRUE, gamma: FALSE

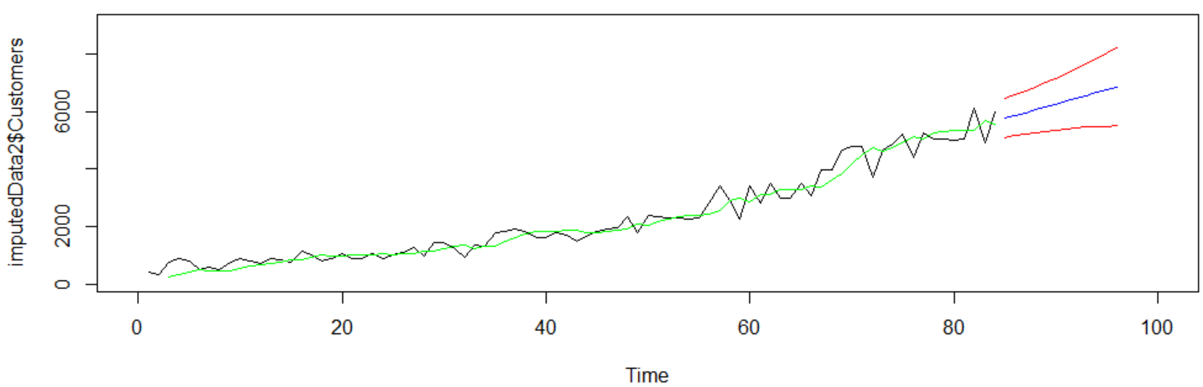

Figure 11: Holt Winters Forecast for alpha: 0.26, beta: 0.15, gamma: FALSE

Next, I try three iterations of ARIMA (autoregressive integrated moving average) models. The three models contain the following parameters:

1. p,d,q each chosen by R with auto.arima()
2. pdq = (0,2,1)
3. pdq = (1,1,2)

For each case, I create an ARIMA model using either the arima() or auto.arima() function. I then print and inspect the results for the arima() function and the acf, pcf, and coefficients. The acf for all models is outside the bounds at the beginning of the time series. The pcf is generally within the bounds for all three models except at one time for models 2 and 3. This indicates that model 1 may be the best of the models.

For model 1, R chose ARIMA(0,1,1) with drift. This accounts for the upward trend in the data.

Model 1 has the lowest AIC value, which is another indicator that this may be the best model. The AIC for each model is as follows:

1. 1211.28
2. 1235.43
3. 1223.98

I then use the forecast() function to determine the next 12 customer predictions. Finally, I plot the forecasts and save the accuracy data to variables. Forecast plots for each of the three ARIMA models are included in Figures 12, 13, and 14.

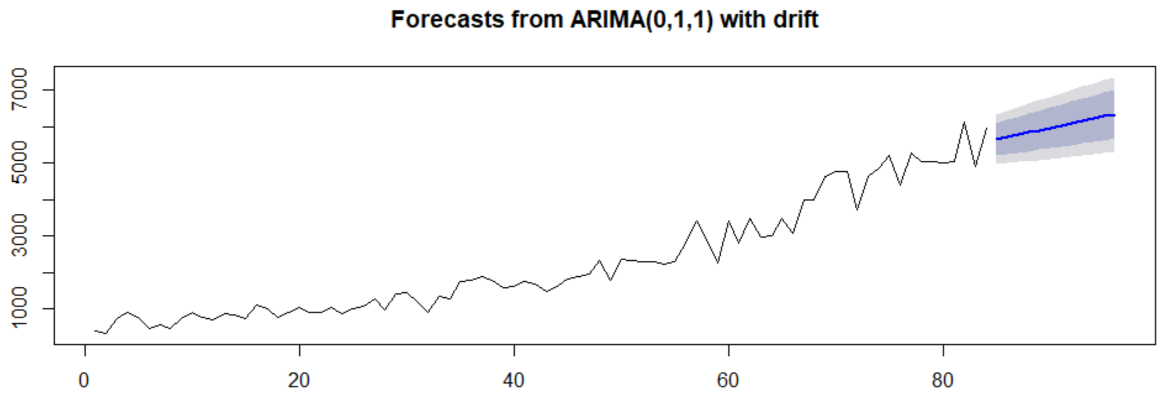

Figure 14: ARIMA Model 1 Forecast

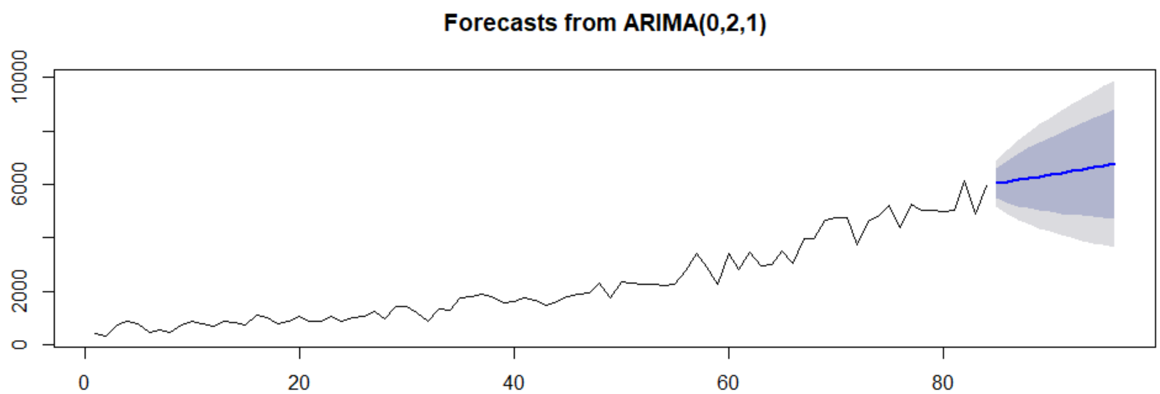

Figure 14: ARIMA Model 2 Forecast

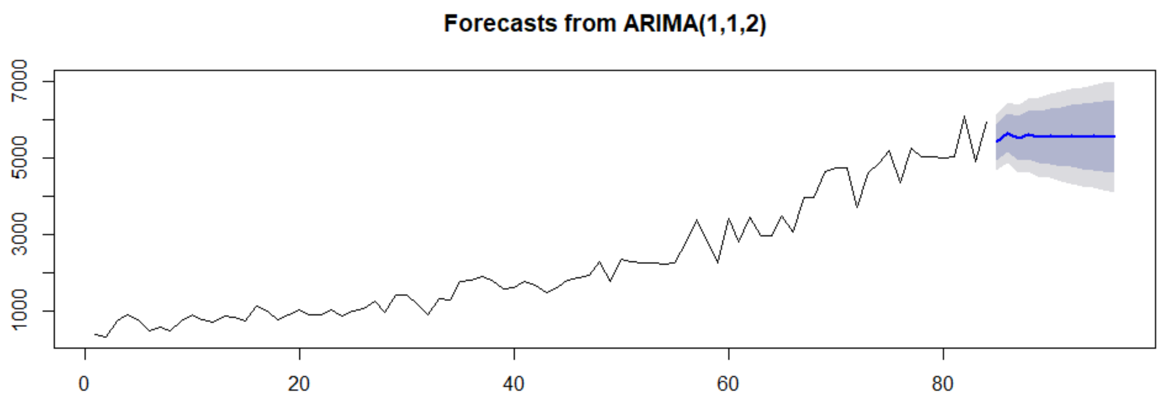

Figure 14: ARIMA Model 3 Forecast

Finally, I compare the accuracy measures of each of the Holt Winters and each of the ARIMA models to determine the overall best model. The R output is shown below:

```
> hwAccuracy
                  ME     RMSE      MAE      MPE     MAPE     MASE      ACF1
Training set 291.165 462.7624 337.1007 11.22032 15.21418 1.115204 0.1231819
> hwAccuracy2
                   ME     RMSE      MAE      MPE     MAPE     MASE       ACF1
Training set 10.22394 421.7589 309.9154 -1.54596 17.21147 1.025269 0.03697826
> hwAccuracy3
                   ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set 56.75932 350.8077 270.7408 3.893735 14.11802 0.8956708 -0.03712948
```

> arimaAccuracy

```
                    ME    RMSE     MAE       MPE     MAPE      MASE       ACF1
Training set 0.2555796 341.134 261.151 -5.629266 14.68872 0.8639458 -0.0903424
```

```
> arimaAccuracy2
                  ME    RMSE      MAE       MPE    MAPE      MASE       ACF1
Training set 25.4505 424.442 296.2716 -1.109132 14.7668 0.9801326 -0.5278212
> arimaAccuracy3
                   ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 115.1606 364.3218 268.9935 3.327019 13.85472 0.8898903 -0.1148581
```

Conclusion and Results

From these results, I determine that the ARIMA model 1, ARIMA(0,1,1) with drift, is the overall best model for this dataset because it has the lowest errors for most measures, including RMSE (Root mean squared error) and MAPE (Mean absolute percentage error). This case is bolded in the output above.

The forecasted customers for the next 12 months from this model are reproduced below:

```
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
85       5654.339 5209.136 6099.542 4973.460 6335.219
86       5716.655 5247.020 6186.289 4998.411 6434.898
87       5778.970 5286.114 6271.826 5025.212 6532.728
88       5841.285 5326.254 6356.316 5053.613 6628.957
89       5903.601 5367.310 6439.891 5083.415 6723.786
90       5965.916 5409.178 6522.654 5114.458 6817.374
91       6028.231 5451.770 6604.693 5146.609 6909.853
92       6090.547 5495.015 6686.079 5179.759 7001.334
93       6152.862 5538.852 6766.872 5213.814 7091.910
94       6215.177 5583.229 6847.126 5248.695 7181.659
95       6277.493 5628.101 6926.884 5284.334 7270.651
96       6339.808 5673.430 7006.186 5320.671 7358.945
```