Matt Allen

DS700 Spring 2018

Final Project Data Cleaning

I combined tabs “May-2007 Violet, LA”, “May-2007 New Orleans, LA”, “May-2007 Lafayette, LA”, and “May-2007 Baton Rouge, LA” into one tab called “ReroutesCombined” by appending records in each tab in Excel. I added a column called “RoutedTo” to indicate the routed to hospital. The routed to hospital name is captured in the tab name. For the four tabs I gave the values Violet\_LA, New\_Orleans\_LA, Lafayette\_LA, and Baton\_Rouge\_LA. Appending the tabs together gave an initial 7407 records. I also took the spaces out of the header names, so that the columns are now: OriginalHospitalLocation, Examination, Date, RequestID, and RoutedTo.

The tabs have all reroutes from all locations. This needed to be filtered down to just reroutes from Abbeville. I used Excel filter function to filter the column OriginalHospitalLocation in the tab ReroutesCombined on just the value Abbeville. I copied these results to a new tab called ReroutesAbbeville. The tab ReroutesAbbeville now contains 1590 records.

The tab ReroutesAbbeville needed to be filtered down further to heart related exams. I created a tab called ExaminationDesc with all the unique values of the Examination column in the ReroutesAbbeville, and added a column indicating if it was heart related. I got the unique values by pasting the Examination column from the ReroutesAbbeville tab into the ExaminationDesc tab, and then used the Excel function remove duplicates. This reduced the original 1590 values to 28 unique values. I then went through all the values with the help of the internet to decide if they were heart related. The result is shown in Table 1.

I then used the Excel VLOOKUP function to look up values from the ExaminationDesc tab to classify the exams in ReroutesAbbeville into heart related exams. A column called IsHeartRelated was added to the ReroutesAbbeville tab. Here is an example of the VLOOKUP function:

=VLOOKUP(B2,ExaminationDesc!$A$2:$B$30,2,FALSE)

I then filtered the data by IsHeartRelated = TRUE. This left 1054 records. I moved this to a new tab called ReroutesAbbevilleHeart.

The date formats in the column Date in ReroutesAbbevilleHeart are inconsistent. Excel was able recognize and automatically parse most of them, but the ones that are of the format ‘14 May, 2007’ did not parse. I found all the dates of this format and manually corrected through a VLOOKUP, and the function IFNA. Once all the dates were set as Excel dates, I proceeded to convert them using Excel date functions and concatenation to the format YYYYMM.



**Table 1: Classification of Heart Related Exams by Examination Description**

I created another tab ‘HeartReroutesAbbevilleByMonth’ to store the reroutes with Year Month format YYYYMM. The other column is the reroute hospital. This will be a final step before doing a pivot table to summarize counts for reroutes by Year Month. Before this step the December 2013 data will need to be added.

The Routing SYSID column in December 2013 Data was parsed into three columns. Using the Excel functions Left and Right, I parsed out the first four characters and the last four characters. I parsed out the condition code by using the Excel function Mid. By inspection, it appears the condition code starts at the eighth character in the SYSID and is six characters long. I used this to lookup in the ‘Heart-related Condition Codes’ tab to see indicate if the exam was heart related. I determined if each record was routed from Abbeville by checking the first and last four characters for the combinations SYSID starts with “L839” and end in either “TGU3” or “ROV8”. An example function for determining whether it was routed from Abbeville is below.

=IF(AND(C2="L839",OR(D3="TGU3",D3="ROV8")),TRUE,FALSE)

I then filtered down by Heart Related Condition Codes and if it was routed from Abbeville. The total number of records after filtering were 5933. This indicates that there were 5933 heart related reroutes from Abbevile. I added these records to the data set in the tab ‘HeartReroutesAbbevilleByMonth’ with YearMo = 201312. RoutedTo is unknown.

I used a pivot table to summarize the data in the HeartReroutesAbbevilleByMonth tab. The aggregation in the pivot table was a count on the record this should represent the number of heart related exam reroutes by month and year. My results are shown in Table 2.



**Table 2: Heart Related Exam Reroutes from Abbeville, LA by Year Month**

My final table for analysis combines the total cardiac exams given in the tab ‘Abbeville, LA’ and the results from Table 2. I formatted the year and month to be in the same form in Table 2, and did a VLOOKUP to match up the total exams and the number of reroutes by year month. I then sorted this data by year month. The table has three columns YearMonth, HeartExamReroutes, and TotalHeartExams.

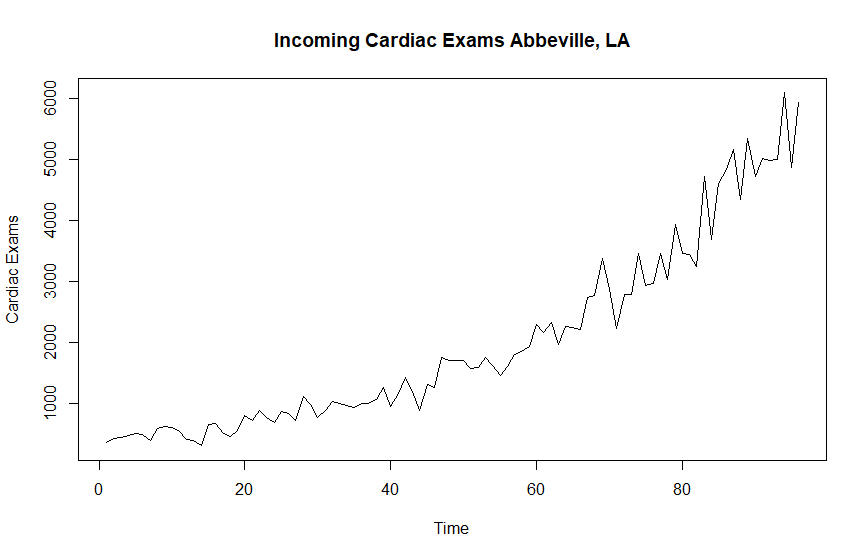
During December 2009 to February 12, there were no incoming examinations recorded. From Explanation of the dataset, there were a total of 5129 incoming requests during this time. I spread the 5129 across the three months. 1710 incoming examinations were put into December 2009 and January 2010, and 1709 were put into February 2010. All incoming requests were rerouted out of Abbeville in December 2013. I changed the \* in total requests to 5933, which is the number of rerouted requests. December 2011 has the value “Closed for holidays”. I set this to blank. January 2011 and December 2008 has the value 999999999. This was set to blank, because it is an unrealistic number. March and June of 2006 have the value \*. These were also set to blank. May 2009 has the value “entered by J.f. williams”. This was set to blank. June 2010 has the value “xx?\*&?/..”. This was set to blank.

May 2007 has a total of 400 reroutes but shows only 107 total incoming exams. From the explanation of the dataset, the 107 exams from May 2017 did not include the rerouted exams. I added the 400 reroutes shown in Table 2 to the 107 exams to make the total incoming exams for May 2007 total 507 exams. Due to New Orleans HC being closed because of hurricane, Abbeville had an unusually high number of incoming exams in October 2008. October 2008 was set to blank, because it is recognized as an outlier. I will impute its value. From the explanation of the dataset the months of May, June, and July in 2013 are incomplete, I added the rerouted values from Table 2 to the exams already present in the Abbeville tab. Finally, I saved the file as a separate csv. The dataset has the columns YearMonths, and TotalHeartExams. There are 96 records spanning 200601 to 201312.

All the blank values will be handled by imputation methods. The R package imputeTS was used to handle imputation. I used imputeTS to fill in NA values by Exponential Weighted Moving Average with four points on either side. The missing year months were 200603, 200606, 200810, 200812, 200905, 201006, 201101, and 201112. The results of imputation are show in Table 3.



**Table 3: Imputed Exams by Year Month**



**Figure 1: Initial Plot of Abbeville, LA Incoming Cardiac Exams**

Figure 1 shows the first plot of the cleaned Abbeville data. It looks like there is a clear upward trend over time, but I do not notice seasonality. For my two models, I will use Holt-Winters and an ARIMA model.

I created a Holt-Winters model in R. Below are the predicted results. Figure 2 shows the plot along with the predicted values and its upper and lower bounds.

fit upr lwr

97 5811.824 6420.220 5203.428

98 5937.570 6562.289 5312.851

99 6063.316 6708.285 5418.346

100 6189.061 6858.361 5519.761

101 6314.807 7012.581 5617.033

102 6440.552 7170.926 5710.179

103 6566.298 7333.315 5799.280

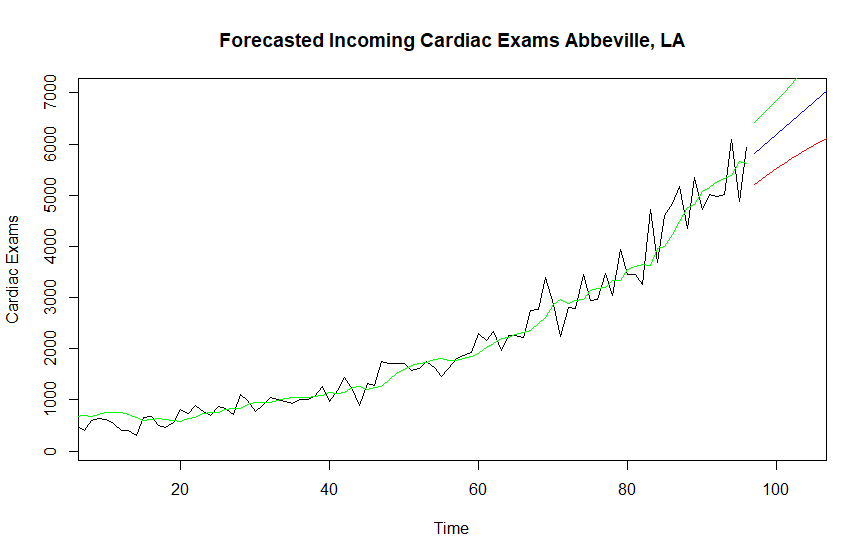
104 6692.043 7499.621 5884.465

105 6817.789 7669.686 5965.893

106 6943.535 7843.331 6043.738

107 7069.280 8020.377 6118.183

108 7195.026 8200.641 6189.411



**Figure 2: Holt-Winters Forecast of Abbeville, LA Incoming Cardiac Exams**

For the second model, I created an ARIMA model. The results are below.

Series:

ARIMA(1,1,1) with drift

Coefficients:

ar1 ma1 drift

-0.3687 -0.4412 54.7285

s.e. 0.1410 0.1364 13.0553

sigma^2 estimated as 98048: log likelihood=-679.53

AIC=1367.07 AICc=1367.51 BIC=1377.28