Fargo Health Group Forecasting Report

Business Problem

Fargo Health Group (Fargo) is a healthcare provider organization headquartered in Birmingham, AL. Fargo operates 34 clinics across the United States. As part of its service offerings, Fargo provides disability compensation benefits to thousands of patients every year through the Quality Assessment Office (QAO) of Fargo. Often requests for disability compensation must include an examination at one of Fargo's 34 Health Centers (HCs). Once an HC receives a request it has 30 days to complete the necessary examinations and return the results. Each day past the 30-day deadline incurs a $200 fine from the Regional Office of Health Oversight (ROHO).

Often HCs do not have the capacity to meet the 30-day timeframe, resulting in fines paid to the ROHO for delayed reports or examinations being rejected by the HCs due to known capacity constraints. Rejected exams are frequently routed to out-of-network Outpatient Clinics (OCs). OCs are not constrained by Fargo's 30-day deadline and also cost $1,250 more to Fargo to complete, resulting in financial and reputational risk to Fargo.

Fargo wishes to explore data analytics as a way to control costs to the OCs and the fees paid to ROHO. This pilot study project will use historical information provided by Fargo to forecast 12 months of demand for incoming cardiovascular examination volume at the Abbeville HC so that Fargo can appropriately staff the Abbeville HC to meet the goals of the pilot study.

Data-analytic Approach

Fargo Health Group provided me with a dataset include eight years (96 months) of past cardiovascular examination data. The steps taken to prepare the data for analysis and forecasting is found in the next section. Initial analysis of the cleaned dataset shows a clear positive trend in the number of cardiovascular examinations over the eight-year period. See Figure 1 where I used

several different periods of moving averages to smooth the data to help discern the pattern in the data. From this analysis I can see that forecasting models will need to include a trend component.
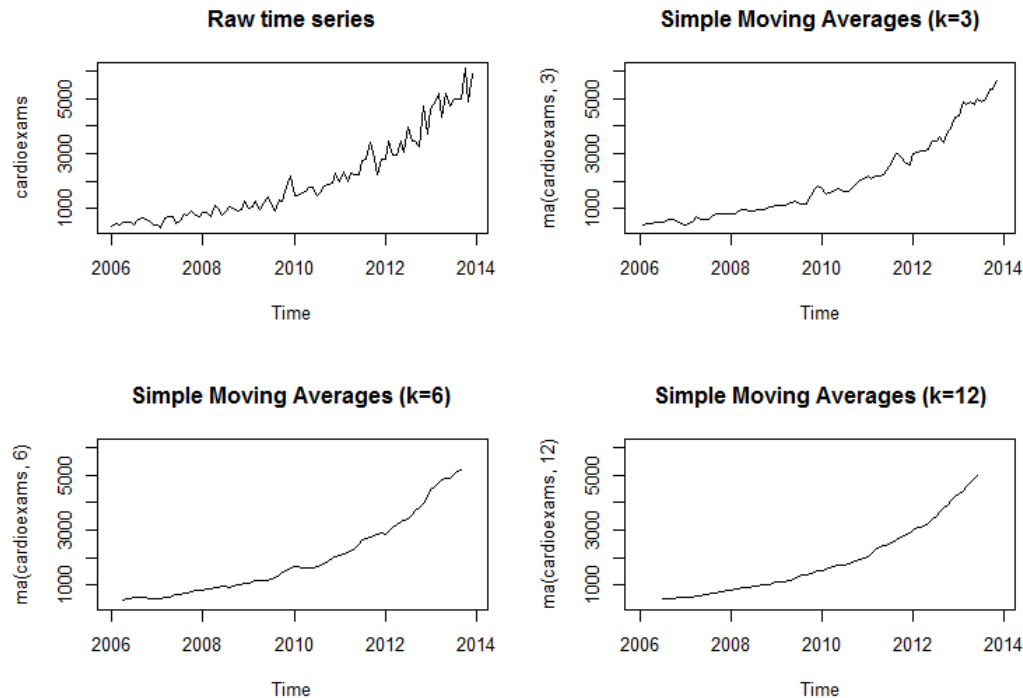

Figure 1

My next step in the analysis was to detect if there was a seasonal component to the time series data. Figure 2 uses the monthplot() function to show the subseries for each month (all January values connected, all February values connected, etc.) along with the average of the subqueries. Figure 3 uses the seasonplot() function from the forecast package in R by showing each month stacked by year, indicating if a similar pattern is present each year during the same month. From these two graphs I do not detect a seasonal component to the time series data.
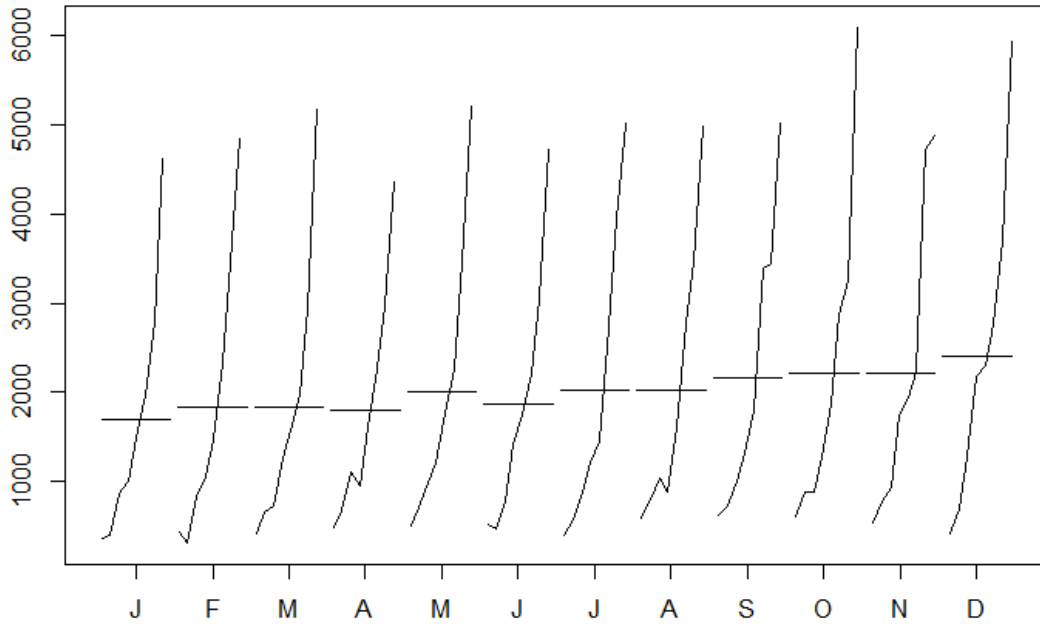
Fargo Health Group Forecasting Report

Month Plot

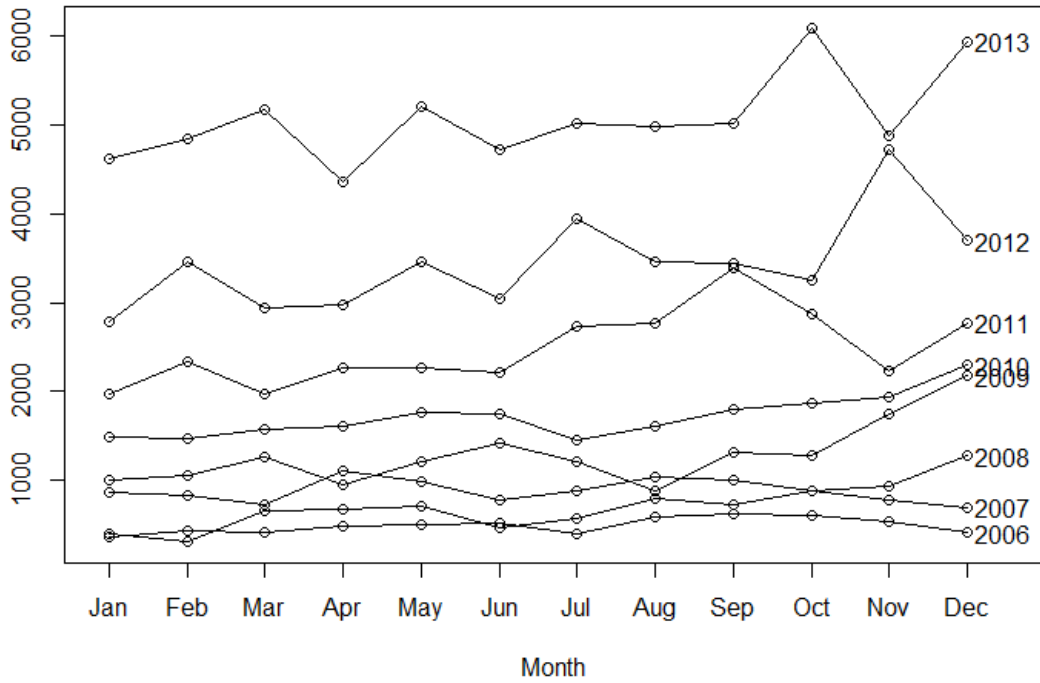*Figure 2*

Season Plot

Figure 3

Fargo Health Group Forecasting Report

Once I determined the time series data shows a trend but does not show a seasonal component I was able to pick my models and begin forecasting. For my first model I selected the Holt exponential smoothing approach because it fit well with the above characteristics of the Fargo data. For the parameters, I chose to use multiplicative (M) for the error type because the variability of the observations increase as the volume of examination increase. I used additive (A) for the trend parameter and none (N) for the seasonality parameter. I confirmed my choice of parameters by using the ets() function to automatically select the best-fitting model for the data. Using Holt's approach I forecast 12 months of cardiovascular exams for the Abbeville HC. My results are shown in Figure 4.
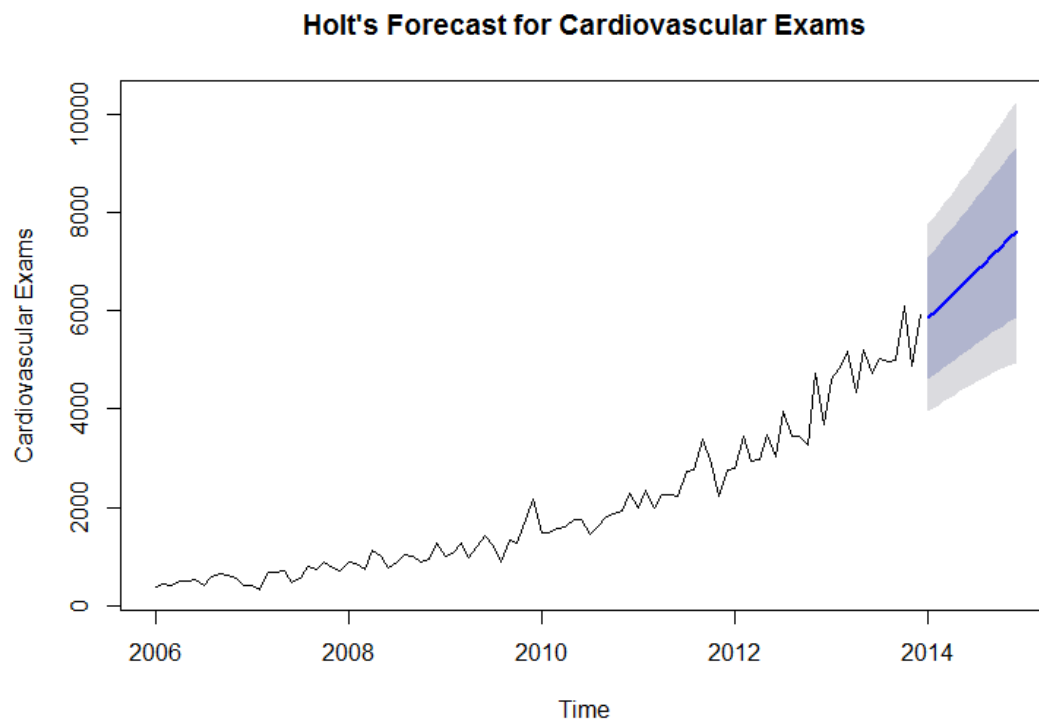

Figure 4

For a second forecasting model I chose the autoregressive integrated moving average (ARIMA) model, again due to its suitability to the Fargo time series data of a trend without seasonality. I used the auto.arima() function to choose the best ARIMA model, which resulted in parameters of

Fargo Health Group Forecasting Report

p = 1, d = 1, and q = 1, with drift. I validated these parameters by running the model several more times with different parameters and I was unable to produce more accurate results based on the model quality measures such as Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). Using the ARIMA model I again forecast 12 months of cardiovascular exams for the Abbeville HC. My results are in Figure 5.
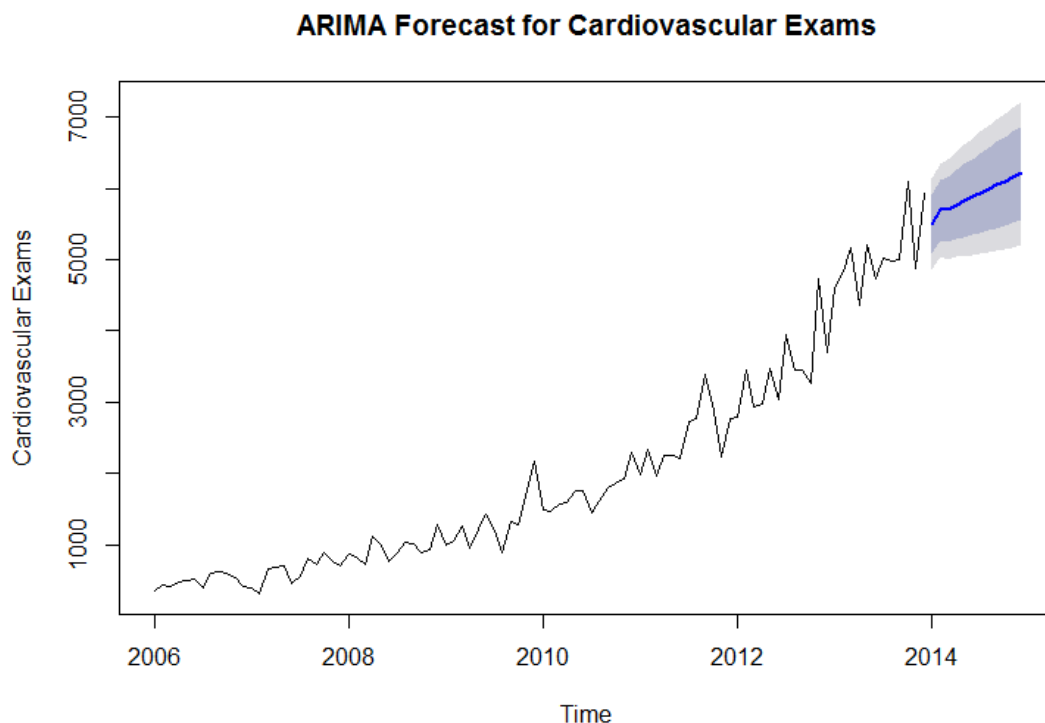

ARIMA Forecast for Cardiovascular Exams

Figure 5

Model Comparison

Table 1 below shows a comparison between the metrics for the Holt and ARIMA models. We see that the Holt method has produced a more accurate model than ARIMA when reviewing each of the quality measures.

Fargo Health Group Forecasting Report

	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)	Akaike's Information Criterion (AIC)	Bayesian Information Criterion (BIC)
Holt (MAN)	301.6435	197.9526	11.97469	1498.29	1511.11
ARIMA (1,1,1) w/drift	317.3250	220.5245	15.17759	1373.65	1383.86

Table 1

Assumptions

- This report is being completed in a time that makes sense for forecasting the next twelve months of exams (perhaps early January 2014)

- The reader has an understanding of statistics, time series data, forecasting models, statistical model quality measures, etc.

Data Clean-up Approach

1. Filter the results in the Violet, NOLA, Lafayette, and Baton Rouge tabs for Original Hospital Location = Abbeville and for heart-related exams (see assumptions)

2. Combined the data filtered data from the previous step into a single tab

3. Removed duplicate exams based on the Request ID field, which removed 185 observations, leaving 869 observations

4. Manually cleaned up remaining date formatting issues (e.g. 16 May, 2007) to produce consistent date data.

5. Due to the confusing nature of the Explanation of Dataset instructions for second of May 2007 and the low number (107) of exams in the Abbeville tab compared to the number of

rerouted exams (349), I set this value to NA so that it could be imputed using the Amelia II package

6. Determined if the exams from the other HCs needed to be added or subtracted from the aggregated values for Abbeville based on the Explanation of Dataset (see assumptions)

7. Parsed out the December 2013 tab data according to the Explanation of Dataset for that data, resulting in 5,933 of the 10,481 observations to update the corresponding row on the Abbeville, LA tab

8. Set the large, obvious outliers (99,999,999; 999,999,999) to NA and also changed the value for October 2008 to NA as the Explanation of Dataset noted that this month was an outlier due to a hurricane for imputing with Amelia

9. Estimated the values for December 2009 – February 2010 using the given number of 5,129 exams during this time by multiplying the given value by the percentage of total exams in those same months from 2006 and 2007 since those years had numbers for the same set of months.

10. Added the May – July 2013 exams to the incomplete data for those rows in the Abbeville, LA tab.

11. Summarized the number of exams by month and year in a summary tab, which was then imported into RStudio for multiple imputation.

12. Used the Amelia package to create five imputations for missing values, which were then written to CSV files and averaged to create values for the missing time periods. Before and after multiple imputation plots are seen below in Figure 6.

    a. I used the polytime argument for the multiple imputation since the data is a time series and is showing an upward trend in the number of cardiovascular exams over time. I ran

the imputation with polytime set to 1, 2, and 3. Polytime = 3 resulted in the smoothest estimates for missing values. See Figure 7 below for a comparison of the polynomials.
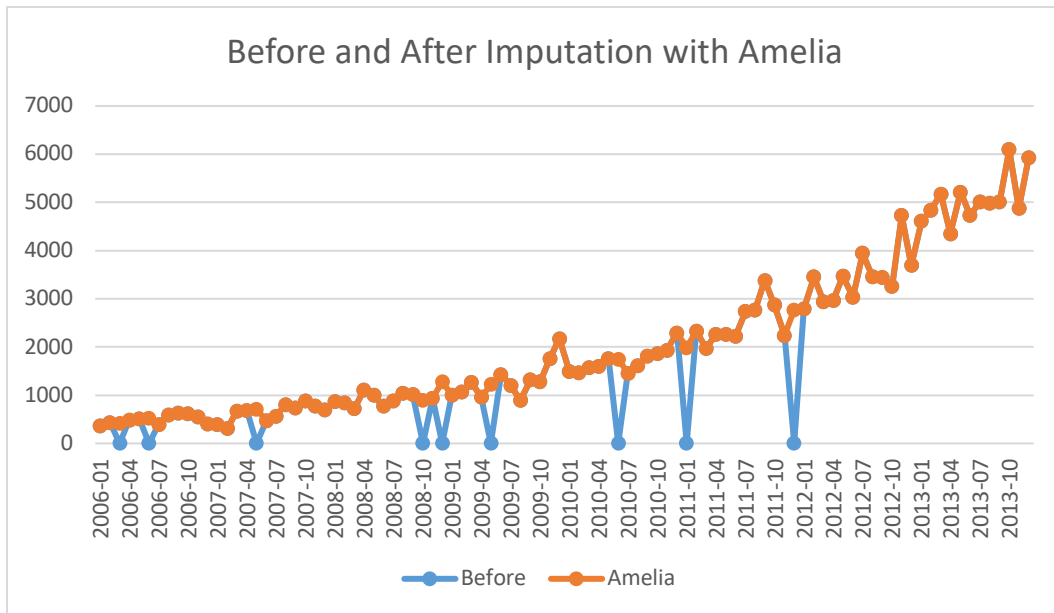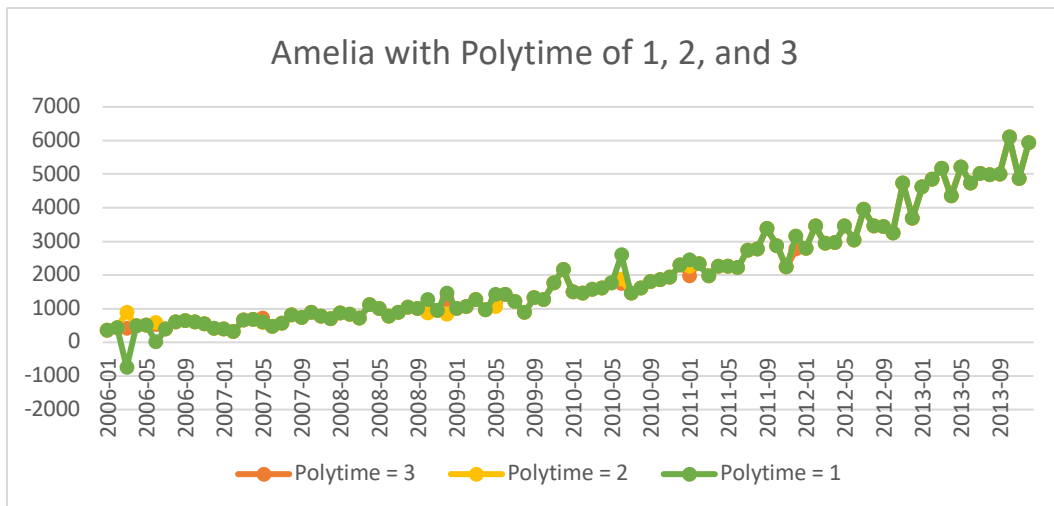

Figure 6

Figure 7

Assumptions

- Abbeville would like to complete as many of the requested exams as possible, up to 100%

- Examinations of type Cardiovascular, Cardiac, matching or similar to the Heart-related condition codes, or other heart-related exams not included in the previous categories (e.g. chest pain, heart) are in-scope for this pilot study

- Did not include Examinations of type MRI because it is too generic to assume it will be specific to Cardiovascular examinations

- Request ID is a unique identifier across the Fargo organization enterprise and can be used to detect and clean duplicate records in the dataset

Ethical Implications

Context: The original purpose of the data collection is not specified in the documentation but it is reasonable to assume it is part of normal operational record keeping to track requests, completions, re-routings, and file for reimbursement. The new use of this data is in line with how it was originally collected since it is common to use aggregated operational data for business reporting and planning purposes.

Consent: Informed consent is required by law in the United States via the Health Insurance Portability and Accountability Act (HIPAA), usually by a HIPAA form that is given to the patient for his/her review and signature. Patients are allowed to decline HIPAA form signature and the provider, i.e. Fargo Health Group, cannot deny care for lack of signature. The dataset provided to me does not contain any protected health information (PHI). From this dataset I could not track an examination back to a specific individual.

Reasonability: The depth and breadth of the dataset is reasonable for the forecast because it goes back 96 months (8 years) and shows a clear growth trend for cardiovascular exams. Enough valid values for the Abbeville HC were included to reasonable impute any missing or outlier values in the series.

Fargo Health Group Forecasting Report

Fairness: I believe the results of the forecast deployment will be equitable to all parties. If the Abbeville HC is staffed more appropriately that means that patients will receive their exams in a timely manner, Fargo Health will be able to better control their reputation and costs for the disability benefits compensation program, public health agencies such as ROHO will receive their results in a more timely fashion, and employees will benefit from better staffing of their health center.

Ownership: The dataset, analysis, and insights from the data analysis belong to Fargo Health Group since they collect and hold the data, and have paid for the analysis. I believe there is a moral obligation for Fargo to act given the benefits outlined above, especially for patients requesting disability benefits to have their adjudications done in a timelier manner.

Accountability: Ultimately Fargo Health Group is responsible for the mistakes and unintended consequences in data collection and analysis. Fargo has provided the data and set the parameters for the analysis. Affected parties could check the results that affect them by reviewing this report, the code used to impute the missing data and forecast the future demand for cardiovascular exams, the clean dataset, and the original dataset if those materials are made available by Fargo.