Case Study Assignment

FARGO HEALTH

DATA ANALYTICAL APPROACH

Report By:

Raghavi Rajumohan

I. Business Overview

Fargo Health Group was established in 1916 by Dr. Nathanial Scott Fargo in Birmingham Alabama. The clinic, initially started off as a gastroenterological health clinic later transitioned into an extended healthcare service after its takeover by the Fargo Family, namely Dr. Nathaniel's sons in 1940. Now, they provide a wide array of medical services, which they achieved by expanding the number of doctors and researchers from a broad range of medical specialties. Currently, there are 34 successful clinics spread across the United States that provide quality healthcare services to the public.

One such service that the Fargo Health Group provides is the disability compensation benefits. They have established a Quality Assessment office that is in charge of collecting disability data of their patients from Fargo's 34 clinics. This service is essential for the improvement of lives of thousands of individuals as it helps the government suitably compensate them, and the examinations conducted by Frago will aid in this decision. This data also helps in increasing the quality of healthcare provided by hospitals all across the country and also customer satisfaction.If an individual wishes to receive disability compensation, they start the process by submitting a request to one of Fargo's 34 Local Offices (LO), which is then forwarded to one of their 34 Health Care centers (HC). Once such a request has been made, the HC must follow a 30-day mandate by which they have to revert back to the LO with the results of the examination.

II. Symptoms and Problem Statement

From the case study, we have identified some challenges and consequences faced by Fargo health centres with disability examination process. The following is a description of such challenges that have been brought up by the case study.

$200 fine for delayed reports:

Fargo's management requires that HCs perform disability examinations and return the results to the LO within 30 days of receiving the request. This means that the HC must complete the examination within the 30-day timeframe and report the results to the requesting LO. If the examination is not completed within the allocated time frame, Fargo will be fined $200 per day by the Regional Office of Health Oversight (ROHO). These overdue examinations have adverse effects on patients' health and well-being, as well as Fargo's reputation and financial situation. In reality, HCs frequently do not have enough examining physicians, making it difficult to meet the mandated 30-day timeframe. Consequently, they are unable to complete the requested examinations on time, resulting in delayed reports being sent back to the LOs, which often exceeds the 30-day deadline. As a part of a solution, the LO reroutes the request either to other Fargo HCs in the vicinity, which is not a common occurrence, or more commonly, to one of the neighbouring Outpatient Clinics (OCs) that are out-of-network, with the expectation that the OC will be able to find available staff and complete the examinations within the required timeframe.

Issues with rerouting:

The OCs are not part of Fargo's network, each request that is examined by an OC costs an average of $1,250 more than what Fargo would pay for an in-house examination if there were enough staff available. Moreover, while Fargo manages the HCs and enforces the 30-day deadline, there is no assurance that the OCs will meet this deadline since they are not part of Fargo's network. Therefore, redirecting requests from HCs to OCs represents another significant financial and reputational strain for the organization.

In cognizance with the above-mentioned issues, we have defined the **problem statement**:

There exists an acute shortage of examining physicians at the health centres for the disability examinations, so Fargo has to outsource their disability patients to OCs. These OCs are highly expensive for the company, and hence, Fargo requires a data-based approach to anticipate the incoming examination demand to effectively reallocate or schedule physicians at their HCs based on the predicted demand.

Solution

By looking at the past data, we will be able to identify patterns and trends ,and also predict future demand for disability examinations Frago health will receive. This will help Fargo Health Group allocate physicians to the right locations, so that they can meet the 30-day mandate. That is, based on the predictions, we can send physicians from clinics with low demand to those with high demand or hire more physicians at the Abbeville clinic accordingly using Time-series analysis.

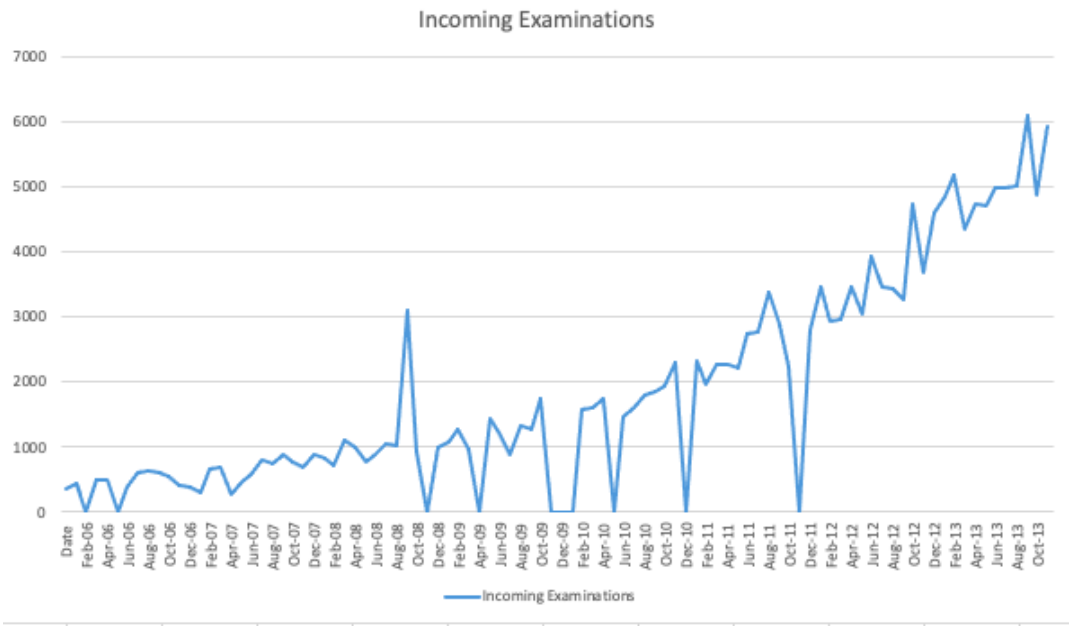
III. Data Cleaning

The first step for data analysis is cleaning the data. Through this step, our aim is to remove any data points that might skew our results, ensuring that our predictions are accurate.

1. Looking at general trend

- First, a time-series graph was plotted to understand the general trend of the data points, this was a simple line graph with dates in the x-axis and total number of examinations on the y-axis.
- From Graph1 we see that there is a general upward trend in the total number of cardiovascular examinations over the eight-year period.
- There are also a lot of data points that have taken the value 0, these would need to predicted using mathematical modeling.
- Before predicting the NaN values, we would first have to further clean the data set.

Graph1: Trend of Cardiovascular Examinations Over Time (Uncleaned Data)



2. Identifying Missing Values

- At first glance, we notice that there are certain non-numerical values that need to be identified as missing values. This is necessary to produce an accurate data set for future predictions.

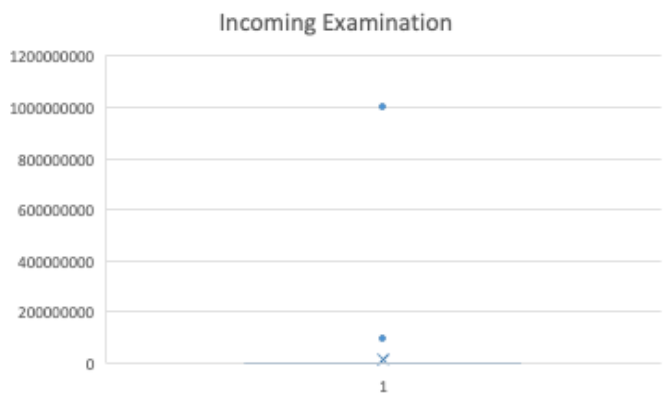
  The following were the changes made:

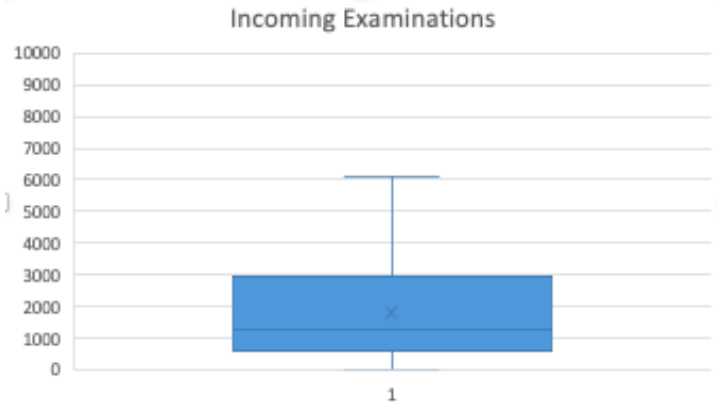
Original Value	Changed Value
xx?*&?/..	NaN
entered by J.f. williams	NaN
*	NaN
Closed for Holidays	0

3. Removing Outliers

- We then identified any large values in the dataset to prevent any predictive bias due to a skewed data set.
- This was done so by plotting a box and whisker plot to identify the max-min range. Any data points beyond this range would be changed to a missing value.
- From Graph 2 we see that there two such values that are extremely high, adjusting the axis (Graph 3) we can conclude that any value above 6094 will be considered as an outlier.

Graph2: Box Plot of Total Examinations (Adjusted Axis) *Graph3: Box Plot of Total Examinations*





- Based on the graphs, the following changes were made:

Original Value	Changed Value
99999999	NaN
999999999	NaN

4. May 2007

- Looking at the values individually, we noticed that total incoming examinations in the month of May for the year 2007 (**107**) is very low, compared to the rest of the year. So we assume that due to certain external factors, such as renovations, a few examinations were transferred over to other healthcare centres. Since the transferred was not done due to a lack of resources such as physicians, we would need to add the rerouted number of examinations to Abbeville demand for the month of May, to make accurate predictions.

- To identify the total number of examinations that were transferred out, we analysed the individual city-wise data set for incoming examinations for the 4 other HCs in LA.
- For each of the transferred HCs – Baton Rouge, Violet, Lafayette and New Orleans – all cardiovascular exams from Abbeville were filtered out, resulting in **336** data entries. We incorporated all cardiovascular and cardiac tests that either matched or were similar to the codes associated with heart-related conditions, as well as other heart-related exams
- Using the Request ID, 51 duplicate examinations were removed, leaving behind a total of **285** cardiovascular examinations that were transferred out of Abbeville in May 2007 due to renovations.
- This value was then added to the original value of 107, to give us **392** total cardiovascular examinations in Abbeville for May 2007.

5. December 2013

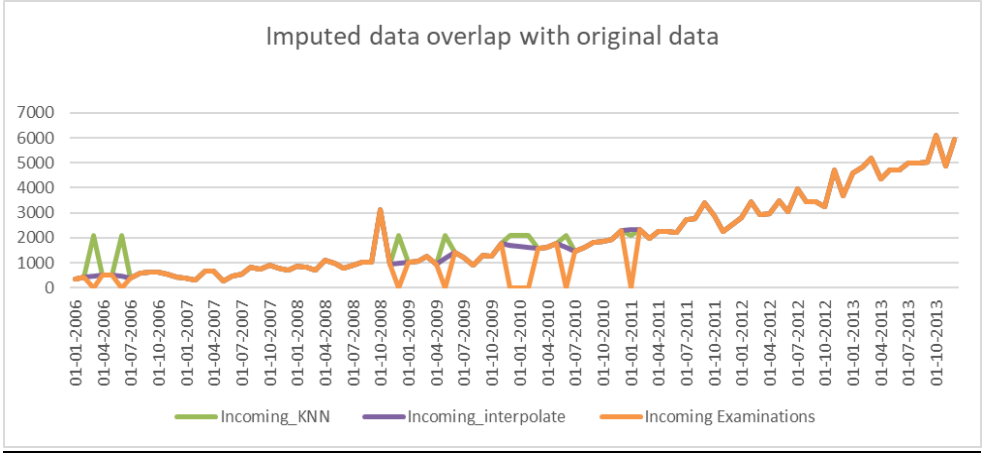
- In the data set, there was a list of all the rerouting SYSID for the month of December 2013 along with codes for heart related conditions.
- For each of the codes, that is for each of the type of cardiovascular diseases, the total number of SYSIDs were calculated.
- The total count, 5933 was added to the total incoming examinations for Dec 2013.

Original Value	Code	Total Count
Cardiac arrest	VVN284	1704
Acute myocarditis NEC	PJU008	791
Angina pectoris NEC/NOS	ABN441	750
Heart disease NOS	UMP621	613
Heart failure NOS	UMX710	556
Myocardial degeneration	TUX333	476
Cardiac dysrhythmias NEC	KPN015	325
Cardiac dysrhythmia NOS	RLX001	252
Endocarditis NEC	WPC608	197
Left heart failure	LLA092	149
Acute cor pulmonale	LLN112	64
Cardiomegaly	TNR628	38
Takotsubo syndrome	LOR159	6
Hyperkinetic heart dis	KON421	5
Atrial fibrillation	KOZ198	4
Chordae tendinae rupture	ROB001	3
	TOTAL	**5933**

6. Prediction of Missing Values

- We then had to decide whether to use KNN or Interpolate to impute the missing values that were identified (NaN). Various Imputation methods have been studied and used that are specific to Time-series.
- On comparing the plot of the original Incoming Examinations to the Incoming Interpolate and the Incoming KNN, we obtain this chart:

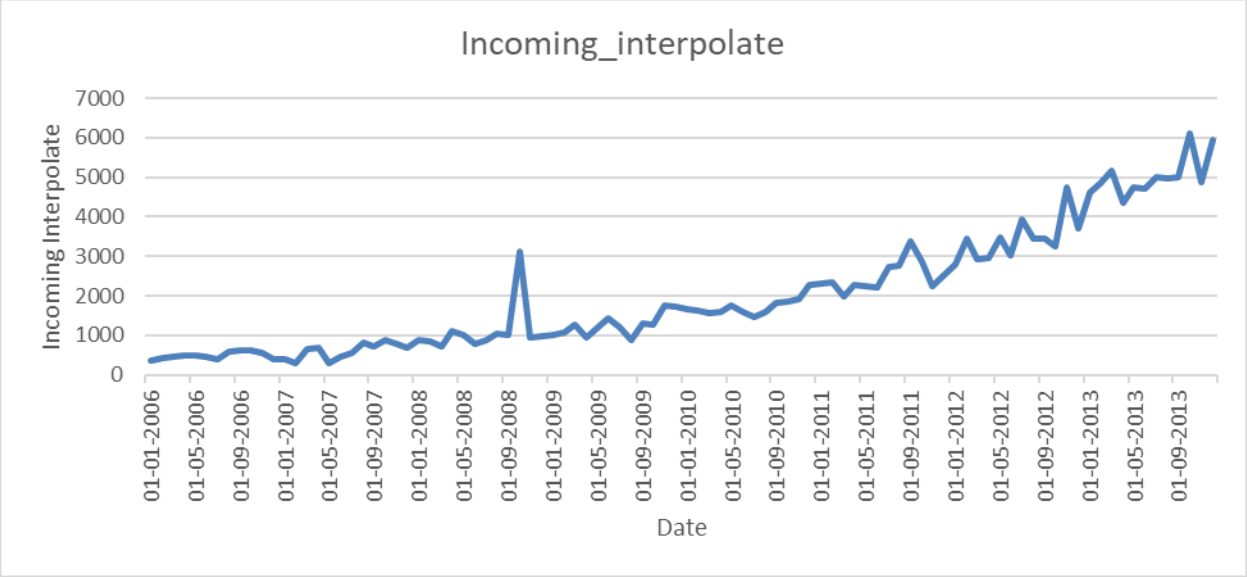
Graph 4



- As you can clearly see, Incoming KNN seems to be having quite a few upward spikes, while Incoming Interpolate has almost none. This was because KNN was filling in each missing value with the same numerical value of 2083. However, we have seen earlier that there is a gradual increase in the total number of examinations over the years. Hence it would be inaccurate to predict that the total number of examinations in the earlier years would be the same as in the later.

- This is why, it was deduced that the Incoming Interpolate can be used to best fill in the missing values of the Incoming Examinations. In other words, Using Interpolate method, the line graph 'Incoming_interpolate' overlaps much better with the original data(Incoming examinations) than the KNN-imputed data (Incoming_KNN), implying that imputed data using interpolation coincides with the original data mostly.

Graph 5 : Line Graph of Cardiovascular Examination Demand with Interpolated NaN Values



- Compared to the orange line of the Incoming Examinations in the previous chart, there is little to no difference between the Incoming Interpolate and the Incoming Examinations which further proves that using the Incoming Interpolate is best for filling-in the missing values.

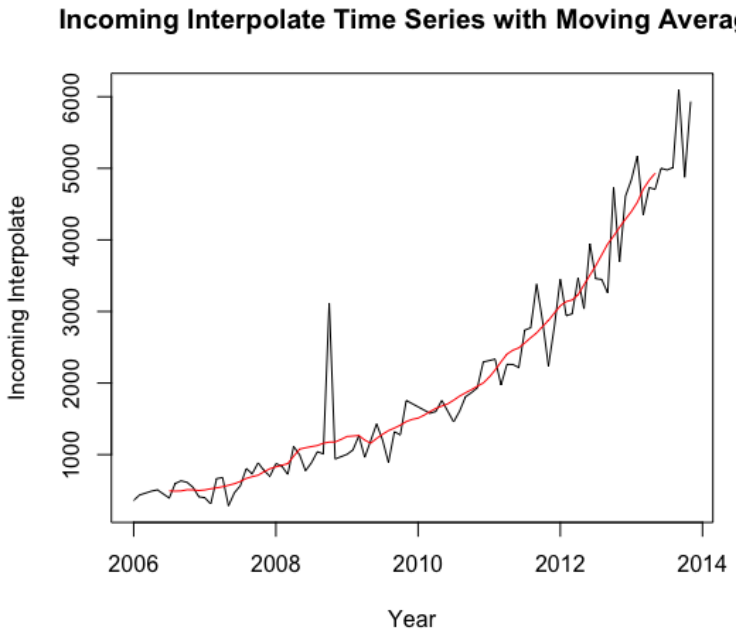
With this we had our final cleaned data set, ready for our prediction model

IV. Primary Data Analysis (Characteristics)

Any time series data that needs to be modelled should be stationary. Due to this, we needed to check the nature of the data to be able to understand whether the data is stationary or not. The properties of stationary data are Constant Mean, Constant Variance, and No Seasonality. To check for Stationarity, we first once again looked at the Incoming_interpolate plot with the Date as the x-axis.

1. Moving Average

Graph 6: Moving Average of Cardiovascular Examinations to Test for Stationarity

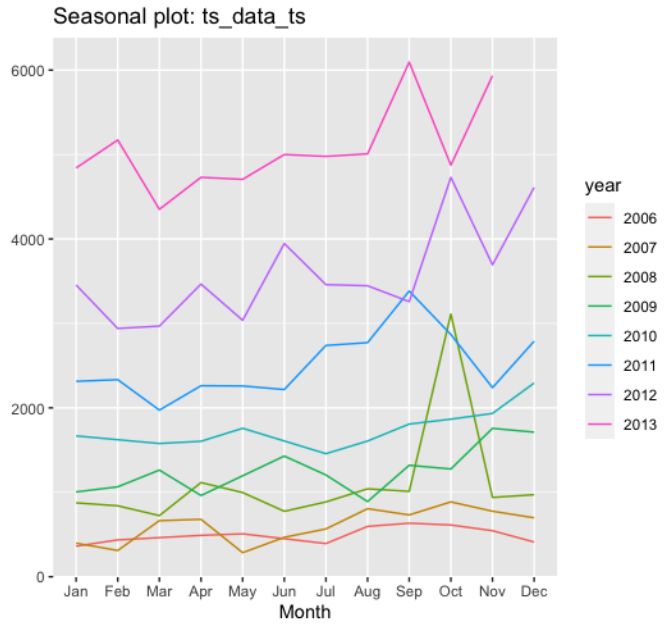


- After viewing the plot, it was very clear that the data had no fixed constant mean so it was wise to consider that the data is not stationary. But to further prove that it is not stationary, we did an Augmented Dickey-Fuller(ADF) regression Stationarity Test. This test reveals to us,

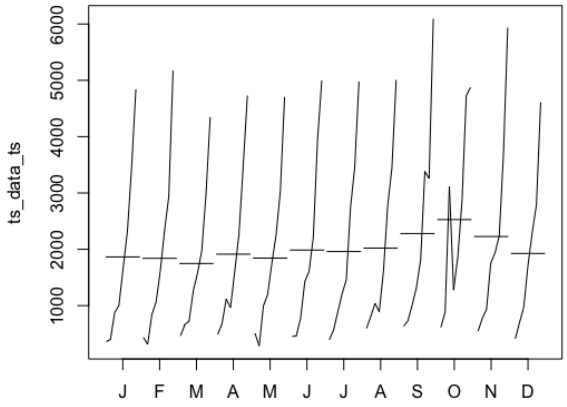
the p-value of the chosen column (in our case, it being Incoming_interploate). If the p-value is less than 0.05, the data is Stationary. If it is greater than 0.05, the data is not Stationary. After performing the test, our data's p-value came out to be 0.99, which is clearly greater than 0.05, proving that the data is not stationary.

2. Seasonality

Graph 7: Yearly Distribution



Graph 8: Monthly Trends: Check for Seasonality



- To test for seasonality, we plot the total incoming examinations for each month classified by the year and a monthly plot of total examinations.
- In graph 7, we can see that over the years there has been an increase in the number of patients/tests at an increasing rate with the gap between consecutive years (specifically 2013 and 2012) increasing. However, the changes throughout the year do not seem to differ across the months.

- Similarly in Graph 8, we see even graphs for most of the months, with all the points connecting. The averages for each month also lie very close to each other.
- Based on this observation, we can conclude that there is no seasonality in the data.

V. Choice of Prediction Model

Using an ARIMA model seemed to be the best method because an ARIMA model integrates the data by performing a difference operation that converts non-stationary data into stationary data.

ARIMA stands for Auto-regression, Integration and Moving Average. These three parts of ARMIA are denoted by (p,d,q) respectively. The main job is then to decide what order these processes should take place in with regards to our data. By using the auto_arima function that finds the ARIMA model with the minimum AIC score, we were able to find out that the best ARIMA model had an AIC value of 1398.291 and was of the order (0,1,1). Auto Arima function in R does not require the user to perform any diagnostic tests, including autocorrelation tests. Now that we were aware of this, we could ultimately train and fit the model to begin making predictions.

VI. Results

Table 1: Summary of ARIMA Model (0,1,1) Coefficients and Error Measures

Coefficients		
	MA1	Drift
	-0.6942	55.1885
Standard error	0.0617	12.8395
Sigma^2	161068	
Log likelihood and Information Criteria		
Log likelihood	-696.21	
AIC	1398.42	
AICc	1398.68	
BIC	1406.05	
Training set error measures		
ME	-1.58	
RMSE	394.94	
MAE	255.33	
MPE	-10.5	
MAPE	18.46	
MASE	0.355	
ACF1	-0.093	

The fitted model equation for an ARIMA(0,1,1) model with drift is:

$$y_t = y_{\{t-1\}} - 0.6942\ e_{\{t-1\}} + 55.1885 + e_t,$$

where y_t is the value of the time series at time **t**, e_t is the error term at time **t**, and $e_{\{t-1\}}$ is the error term at time **t-1**. ARIMA model with drift is an extension of the standard ARIMA(p,d,q) model which includes a constant term('drift term') in the model equation. The drift term represents a time trend in the data that is not accounted for by the autoregressive and moving average components of the model. In other words, the autoregressive and moving average components of the model capture the short-term dynamics of the time series, while the drift term captures the long-term trend or level shift in the data

This equation can be used to make predictions for future values of the time series, where the predicted value at time t+1 would be:

$$y_{\{t+1\}} = y_t - 0.6942\ e_t + 55.1885 + e_{\{t+1\}}$$

Note that the initial value y_0 is required to make predictions for the first value of the time series, since the model assumes that the first difference is stationary.

The ARIMA(0,1,1) model means that the time series has been differenced once (d=1) to make it stationary and has a moving average (MA) component of order 1 (q=1). There is no autoregressive (AR) component in the model because the order of differencing is 1 (d=1). The MA1 coefficient of -0.6942 means that the current value of the time series is negatively influenced by the previous error term by a factor of 0.6942. This coefficient is statistically significant because its standard error is much smaller than its value.

The drift coefficient of 55.1885 represents a constant trend in the data. This means that the time series is expected to increase by 55.1885 units over time, regardless of its previous values. This coefficient is also statistically significant because its standard error is much smaller than its value.

The variance of the error term (sigma^2) is 161068, which represents the amount of variation in the data that is not explained by the model.

The log likelihood of the model is -696.21, which measures the goodness of fit of the model. Higher values of the log likelihood indicate a better fit.

The Akaike information criterion (AIC) is 1398.42, which is a measure of the quality of the model while taking into account the complexity of the model. Lower values of AIC indicate a better trade-off between model fit and complexity.

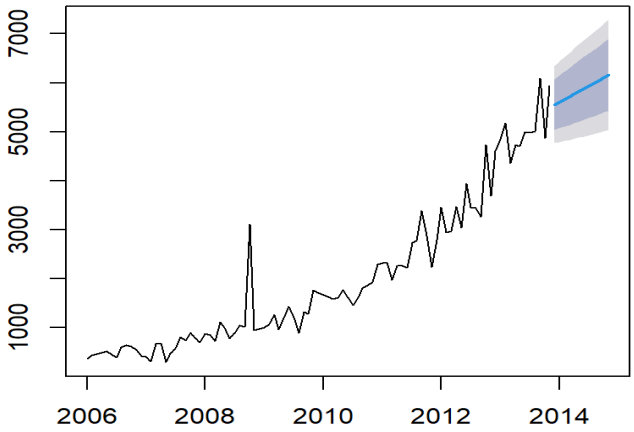
The corrected AIC (AICc) is 1398.68, which is a version of AIC that is adjusted for the sample size. It is similar to AIC, but it is preferred when the sample size is relatively small.

The Bayesian information criterion (BIC) is 1406.05, which is another measure of the quality of the model while taking into account the complexity of the model. Like AIC, lower values of BIC indicate a better trade-off between model fit and complexity.

The training set error measures provide information on the accuracy of the model's predictions on the training set. The ME (mean error) is -1.576232, indicating that the model tends to slightly overestimate the true values. The RMSE (root mean squared error) of 394.9454 is the standard deviation of the errors and indicates that the model's predictions have a large amount of variability. The MAE (mean absolute error) of 255.3338 measures the average magnitude of the errors. The MPE (mean percentage error) of -10.50525 and the MAPE (mean absolute percentage error) of 18.45821 indicate that the model's predictions tend to be biased and have a large percentage error. The MASE (mean absolute scaled error) of 0.355478 measures the accuracy of the model relative to a naive forecast that predicts the next value to be equal to the previous value. Finally, the ACF1 (autocorrelation of the residuals at lag 1) of -0.09353905 measures the degree of correlation between the errors of adjacent observations. A value close to 0 indicates that there is no significant correlation between the errors, which is a desirable property for a well-fitting model.

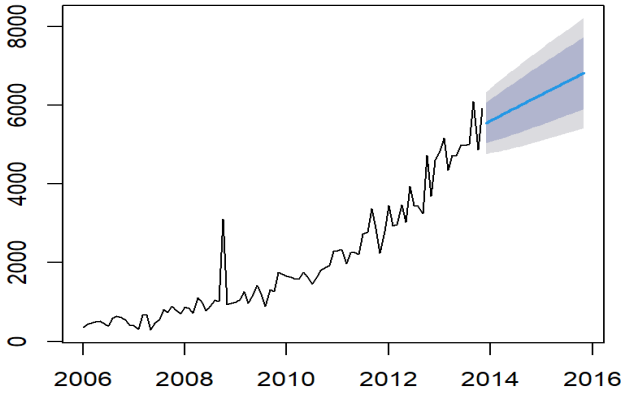
Forecasting graphs:

Graph-9:Forecast of cardiovascular examination demand for an year after 2013 along with 95% confidence intervals



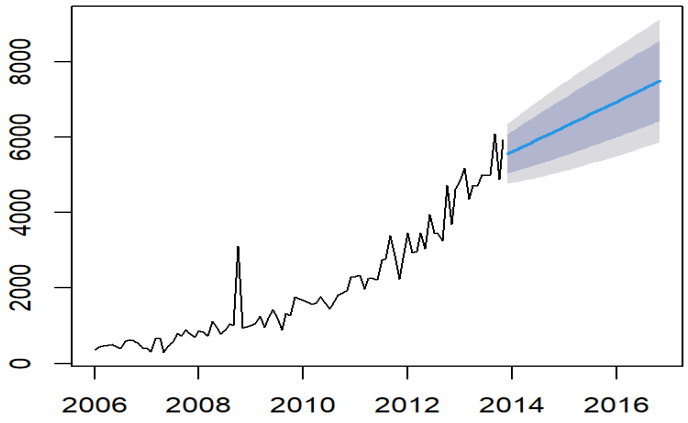
The above graph illustrates the forecasted values for the next 12 months (from Dec 2013 to Nov 2014) along with the 95% prediction intervals. The point forecast (or the expected value) for December 2013 is 5555.219, with aa 95% prediction interval of 4768.620 to 6341.817. The point forecasts for the following months gradually increase, reflecting an increasing trend in the cardiovascular examination demand. The prediction intervals also widen as the forecast horizon increases, reflecting increasing uncertainty as we move further away from the observed data.

Graph-10:Forecast of cardiovascular examination demand for two years after 2013 along with 95% confidence intervals



The above output extends the forecast horizon by another 24 months (from Dec 2013 to Nov 2015). The point forecast for December 2014 is 6217.480, with a 95% prediction interval of 5071.643 to 7363.317. Similarly, the point forecasts for the remaining months gradually increase, reflecting an increasing demand for cardiovascular examinations, while the prediction intervals widen as we move further away from the observed data.

Graph-11:Forecast of cardiovascular examination demand for three years after 2013 along with 95% confidence intervals



Similarly, this graph shows examination demand forecasted for three years from December 2013. As observed earlier, the point forecasts gradually increase over time, while the prediction intervals widen, reflecting increasing uncertainty as we move further away from the observed data.

VII. Conclusion

The ARIMA model with the order (0,1,1) predicts increasing demand for cardiovascular examinations in the future(from 2014) for Abbeville HC. However, with the current model employed for analysis, prediction intervals widen as the forecast horizon increases, indicating increasing uncertainty further from the observed data. The width of the prediction interval is influenced by several factors, including the level of confidence desired, the sample size of the observed data, the variability in the data, and the complexity of the model used for the prediction. In general, the wider the prediction interval, the less certain we are about the predicted outcome. However, it’s important to note that while widening prediction intervals can be an indication of increased uncertainty, it doesn't necessarily mean that the predictions are less accurate. A wider prediction interval simply means that there is more potential for variation in the predicted outcome.