Pilot Study: Forecasting Incoming Cardiovascular Examinations

Jerry Mischel

University of Wisconsin – Eau Claire

Data Science 700

**Situation Analysis & Business Problem**

As part of its current services, Fargo Health Group (referred to as Fargo moving forward) provides disability compensation benefits to thousands of patients every year. The current disability examination process starts with a patient submitting a request to one of the 34 local offices (LOs). A customer representative from the LO evaluates the request and either conducts the examination there or forwards the request to one of the 34 Health Centers (HCs), which is often the case, requesting that the HC perform the necessary disability examinations. The HC has 30 days, from the time of receiving the request, to conduct the examination and provide results, but there have been significant delays because of understaffing. These delays have resulted in significant fines paid to the Regional Office of Health Oversight (ROHO) and additional money paid to outsource the examinations to Outpatient Clinics (OCs).

The Quality Assessment Office (QAO) of Fargo, which is responsible for the collection of disability examination data from the 34 HCs and the subsequent analysis of that information, has asked for me to conduct a pilot study that could lead to the development of a predictive analytic product which could help Fargo accurately guess the incoming volume of cardiovascular requests into one of their HCs located in Abbeville, LA and thus use that information to improve the scheduling of physicians who specialize in those types of examinations. This predictive analysis would also lead to a reduction in fines to the ROHO and additional money paid to the OCs.

**Key Findings & Recommendations**

Through my data analysis, which is outlined in the Data Analytic Approach section, there were key findings and observations which could be integral in making potential business decisions at Fargo to ensure more efficient patient service and lessen the financial burden in the form of fines to the RHOC and outsourcing to OCs. Below you’ll find these key findings and observations, along with recommendations for the management team at Fargo.

* **Incoming examinations exponentially increase.** After cleaning the dataset and imputing missing values, the two graphs below show an exponential increase in the number of examinations between January 2006 – December 2013. To provide an idea of this increase in incoming examinations during the eight-year span, the number of yearly examinations increased from 5,699 in 2006 to 60,832 in 2013 (or a 967% increase). This increase demonstrates a possible need in the future for additional clinics, whether HCs or LOs, and additional medical staff in the Louisiana area to accommodate future influxes of patients. Exploring the efficiency of examination procedures might also be beneficial to see if there are ways to shore up examination timeframes. Future analysis of other HCs in the state could also determine priorities for location and staff.

|  |  |
| --- | --- |
|  |  |
| The number of cardiovascular examinations per year from 2006 to 2013. | Time series plot of the number of cardiovascular examinations per year from 2006 to 2013. |

* **Predicting incoming examinations over 2014.** The ARIMA model #1, chosen as the most accurate model to forecast incoming examinations for the next 12 months (Jan-Dec 2014), predicts approximately 78,608 cardiovascular examinations in 2014, a 29% increase from 2013. The predictive model, which is explained the Forecasting section, also highlights, with a 95% confidence level, the lower and upper bounds of projections to be 58,922 and 82,294, respectively. If the Abbeville HC wants to maintain an efficient service over the next year, Fargo management should implement short-term and long-term measures to service the highest level of incoming examinations. Short-term measures include hiring more cardiovascular staff, identifying any cardiovascular equipment gaps, and exploring examination procedures to shore up any inefficiencies and patient wait times. Long-term measures include building more clinics in Louisiana.

|  |
| --- |
|  |
| The number of cardiovascular examinations per year from 2006 to 2014 with forecasted values for 2014. |

|  |
| --- |
|  |
| A time series plot of the number of cardiovascular examinations per year from 2006 to 2014 with forecasted values for 2014. Blue line shows point values with shaded regions showing lower and upper bounds of projected values. |

* **Hurricanes don’t happen frequently, but they are not abnormal in Louisiana.** One core outlier, which represented the influx of examinations routed to Abbeville during the Oct. 2008 hurricane in New Orleans where the HC was closed. According to Wikipedia (2017), hurricanes are not an abnormal event in New Orleans, as approximately 28 tropical or subtropical cyclones affected the state of Louisiana since 2000, so this anomaly could happen again to any of the HCs or LOs in Louisiana. A crisis response plan should be in place, if there already isn’t one, to ensure patients are rerouted accordingly and additional on-call staff are deployed in a timely manner. All clinics in Louisiana and surrounding states should be familiar with this response plan to respond quickly without confusion.
* **No Seasonality of Incoming Examinations.** Seasonality is a trend in a dataset when the same pattern appears every year during a specific timeframe. For example, airline travel reaches a peak during Thanksgiving and Christmas as people travel to visit family. Understanding seasonality with cardiovascular examinations in Abbeville could further help Fargo determine staffing requirements during these times of year. An initial seasonal plot of the dataset (without imputed values), see graph below, revealed a lack of seasonality month to month in cardiovascular examinations. This outcome provided a proof point that incoming cardiovascular examinations happen at random each month, most likely based on people’s health status, which means staffing by time of year probably isn’t necessary.

|  |
| --- |
|  |
| Seasonal plot of initial dataset (without imputed values). Each line represents a year, plotted against each month. |

* **Data Integrity Needs a Boost.** Data integrity was less than ideal with missing values for 10 months, inconsistencies when collecting data, duplicate values using the Request ID and data existing in multiple worksheets. This revealed a need for the Fargo to reassess its data collection techniques and possibly implement rules and regulations across all clinics in Louisiana when inputting data. Additional analysis into the data collection techniques across the entire network of Fargo would also be beneficial, as well as considering a master database implementation across the network and assigning appropriate staff to maintain data integrity.

**Social, Legal and Ethical Considerations**

With making forecasts using this healthcare data, there are social, legal and ethical considerations, which have been outlined below and may impact current recommendations.

* **Privacy.** First and foremost, there are privacy considerations to ensure the anonymity of people’s health conditions, as protected by HIPPA (Health Insurance Portability and Accountability Act of 1996). There can’t be any indirect identifiers, which in combination with other information can determine a patient’s identity. Fortunately, I didn’t come across any specific identifiers which would expose someone’s identity, but precautions should be considered.
* **Cardiovascular-focused and Abbeville-Focused.** This pilot study only focuses on the prediction of one kind of examination from one of Fargo’s HCs. While this is a good sample, given the number of examinations analyzed, focusing on one type of examination at one HC doesn’t quite capture the holistic nature of the services offered by Fargo HCs and the rest of its network. Additional investigations should take place to predict the volume of different examination types at the Abbeville HC and other clinics in the network.
* **Financial and Timing Delay Data Missing.** Two key parts to the current business problem are the timing delays on exams results and the exorbitated amount of money being paid as part of fines to the ROHO and outsourcing examinations to OCs. However, these two pieces of data were not provided. Further investigation of this data could reveal additional insights into staffing issues, examination efficiencies and equipment gaps, and possible correlations with exponential volume incurred over the eight-year period.
* **Consent.** None of the materials provided to conduct this pilot study highlight what kind of consent was gathered from patients to use their data. Since we are analyzing the total number of cardiovascular examinations by month without any distinguishable identifiers, I assume consent was provided.
* **Ownership.** Fargo ultimately owns the dataset and the results of this pilot study to predict future cardiovascular examinations. The business decision lies with Fargo to implement this predictive model and make the appropriate changes accordingly to ensure timely service for patients and the reduction of fines to ROHO and payments to OCs.
* **Fairness.** The fairness of the predictive results will be determined by appropriate analysis of other clinics, rather than applying this predictive model blindly to all clinics. The data focuses on only cardiovascular-related examinations coming into one HC in Abbeville, LA. However, the same analytical methods can be applied to build a predictive model for each of the other clinics.

**Data-Analytic Approach**

To provide Fargo with the appropriate predictive analysis, my data-analytic approach happened in multiple stages. First, I examined the dataset for quality issues, including outliers, missing values, inconsistent data and duplicated data. Second, I needed to address those quality issues by cleaning the data, which included removing duplicates, imputing missing values, and aggregating the monthly cardiovascular data for Abbeville HC into one spreadsheet. The last stage, once we had a clean dataset, required the building of multiple forecasting models and determine their accuracies. The most accurate model would be used as the final predictive analytic product for Fargo’s pilot study.

The initial dataset, culled from Fargo’s data repositories, provided insight on the historical monthly examination volume of cardiovascular examinations from the HC located in Abbeville, Louisiana. The dataset included multiple worksheets, including:

* Abbeville, LA: Contains the incoming monthly aggregate medical examination volume at the Abbeville, LA HC for cardiovascular exams from January 2006 to December 2013
* May-2007 Violet, LA: Includes all the medical exams that were rerouted to the Violet, LA HC during May 2007
* May-2007 New Orleans, LA: Includes all the medical exams that were rerouted to the New Orleans, LA HC during May 2007
* May-2007 Lafayette, LA: Includes all the medical exams that were rerouted to the Lafayette, LA HC during May 2007
* May-2007 Baton Rouge, LA: Includes all the medical exams that were rerouted to the Baton Rouge, LA HC during May 2007
* December 2013 Data: Contains the exams that were rerouted from Abbeville HC to nearby HCs during December 2013. The worksheet is at the examination level, which includes all exams, not just cardiovascular exams.
* Heart-Related Condition Codes: Provides the six-digit heart condition codes with December 2013 data
* Condition Code Map: All six-digit condition codes

As part of the cleaning stage, the first step was to comb through these worksheets to find any quality issues with the data. But before exploring each worksheet, an examination of the current Abbeville, LA worksheet using R and running the data as a time series was valuable to uncover an initial exponential trend increase in the data, as seen by the following diagrams. To provide an idea of this increase in incoming examinations during the eight-year span, the number of yearly examinations increased from 5,699 in 2006 to 60,832 in 2013 (or a 967% increase). Also, these initial graphs uncovered several outliers in Oct 2008 and Dec 2011, which required additional investigation. In the graph on the right, you can also see several gaps in the exponential curve, which most likely meant the existence of missing data, which also required additional investigation to uncover the source.

|  |  |
| --- | --- |
|  |  |
| Initial dataset scatterplot by year and incoming examinations. | Initial dataset time series plot by year and incoming examinations. |

From this initial analysis, I applied my data analytics approach by first identifying the missing values, the outliers, duplicates and any other data inconsistencies. Through this investigation, which was aided by the Explanation of the Data Set, I found the following data quality issues and adjusted values appropriately, as outlined further in the Appendix Figure 2:

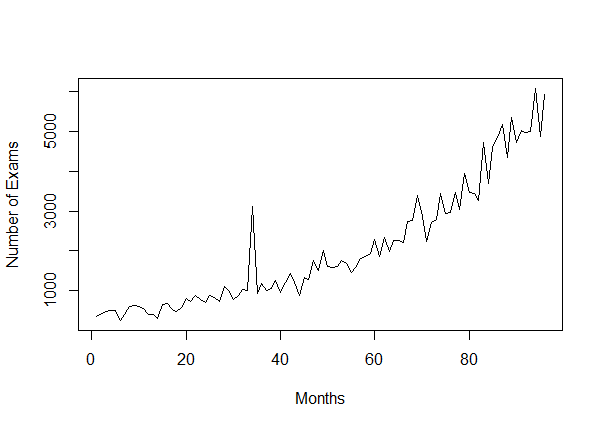
* Missing values in the Abbeville, LA worksheet for rows 4 (March 2006), 7 (June 2006), 49-51 (Dec 2009-Feb 2010) and Dec 2013.
* Incorrectly entered values in the Abbeville, LA worksheet for rows 37 (Dec 2008), 42 (May 2009), 55 (June 2010) and 62 (Jan 2011).
* Row 73 (Dec 2011) in the Abbeville, LA worksheet said, “Closed for holidays.” Since patients came in for exams in December for 2006, 2007 and 2010, this appears to be a clerical error inputting the data. Plus, the holidays in December usually only mean Dec. 24 and 25.
* While row 25 (Oct 2008) looked like an abnormality, which was confirmed by the Explanation of the Dataset, I didn’t feel like removing it would be beneficial to my analysis because the numbers were recorded accurately. There was just a surge in the data because of the hurricane for that month, as the New Orleans HC was closed and rerouted all examinations to other HCs in Louisiana.
* Within each of the May-2007 worksheets for Violet, New Orleans, Baton Rouge and Lafayette, all of the examinations were included (even those that were not cardiovascular-related nor those rerouted only from Abbeville). Cleaning of these worksheets required the removal of all examinations that were rerouted from other places besides Abbeville and all examinations which were non-cardiovascular. (This was determined by additional research to determine appropriate cardiovascular examinations, as outlined in the Appendix Figure 1.)
* Within these May-2007 worksheets, I also discovered a combination of May 2007 data and 2013 data (for the months of May, June and July) and duplicates of the Request ID which, according to the Explanation of the Dataset, is “a unique identifier for each request.” I interpreted this to mean there cannot be any duplicates of the same Request ID, so I used Microsoft Excel to remove these duplicates from the May 2007 data for Violet and from the 2013 data for New Orleans, Lafayette and Baton Rouge. Once I had my totals for each month/year after removing duplicates, I added those totals back into the Abbeville worksheet under each appropriate month, as these examinations were not accounted for originally. I also found two examinations for August 2013 within this data, but did not add back into Abbeville worksheet because the Explanation of the Data Set did not determine those exams were missing.
* In the December 2013 tab, I used multiple methods in Microsoft Excel to reveal the condition codes for each of the examinations, as this information was embedded within each 17-digit SYSID. Once I had a list of condition codes, I was able to remove all non-cardiovascular condition codes (as determined by additional research highlighted in the Appendix Figure 1 and referenced in the Condition Code Map worksheet) and ended up with 5,933 examinations for December 2013, which were input into the Abbeville worksheet as the missing value.
* Finally, I adjusted the format of all examination dates to mm/dd/yyyy for consistency during analysis.

After cleaning the dataset and inputting values from other worksheets into the Abbeville worksheet, I had 10 missing values in the Abbeville worksheet for rows 4 (March 2006), 7 (June 2006), 37 (Dec 2008), 42 (May 2009), 49-51 (Dec 2009, Jan-Feb 2010), 55 (June 2010), 62 (Jan 2011), and 73 (Dec 2011). To ensure I had a complete dataset from which to formulate predictive models, I imputed the missing values using the Amelia package in R which essentially guesses the missing values based on previous values using statistical techniques. Choosing a method to impute values depends on the number of imputations and subjectivity among data scientists, but the results are similar. In this case, a multiple imputation method, like Amelia or MICE, was needed. For my needs, the Amelia package in R is a more streamlined way with a beneficial user interface to impute missing values (compared to MICE), especially when there’s a group of consecutive missing values, like in rows 49-51. For these values, I merely had to add some prior values to help with estimating values to be imputed.

For rows 49-51, I needed to add some prior values to help with imputation, so I leveraged the argument in the PowerPoint posted on Piazza by Dr. Gaurav Bansal from the University of Wisconsin-Green Bay that the standard deviation for these three months (Dec-Feb) across the eight-year span were approximately 3.5% of the sum of these months. Using this argument and the fact that the three values need to add up to 5,129, I used 1,710 (5,129/3) for the mean and 180 for the standard deviation (3.5% \* 5,129) when adding parameters around previous values for exams in these three months.

After running Amelia, I was presented with five different spreadsheets with imputed values for the missing data points. According to Dr. Bansal, the appropriate next step includes a random pick of a spreadsheet with the imputed values. As such, I randomly picked the 3rd sheet with imputed values for all missing rows. I rounded all missing values to the nearest integer, as you cannot have a fraction of an exam. Then all values were added to respective rows in Abbeville sheet.

The result of cleaning the dataset and plotting the time series data can be seen in the graph below.



**Forecasting**

To forecast the next 12 months (Jan-Dec 2014) from my full, clean dataset, I used three different ARIMA model iterations and a Holt-Winters model in R. After evaluating these models and their forecasting errors—Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Scaled Error (MASE)—I found the ARIMA #1 model to display the most accuracy (as you can see by the forecasting error breakdown below). That is, ARIMA #1 has the least amount of errors in three out of four categories and the second lowest AIC value, though R didn’t allow me to find the AIC value for Holt-Winters. ARIMA #1 model was generated by R, which specifically picked out my parameters (p, d and q values). More details below about my forecasting approach are included below for reference.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | RMSE | MAE | MAPE | MASE | AIC |
| ARIMA #1 | 390.6268 | 257.0962 | 18.5919 | 0.7654 | 1413.22 |
| ARIMA #2 | 412.2307 | 264.9014 | 16.7252 | 0.7887 | 1423.22 |
| ARIMA #3 | 413.2958 | 268.1228 | 17.2380 | 0.7983 | 1412.34 |
| Holt-Winters | 393.4109 | 267.2245 | 18.9895 | 0.7956 | N/A |

According to Kabacoff (2015), both ARIMA and Holt Winters are perfect forecasting models for time series data which include a trend in the data. In this case, there’s an exponential increase of incoming examinations over time with one main outlier (the 2008 Hurricane). Before diving into the forecasting models, I also had to look at seasonal components—essentially, is there a seasonal trend in the data like more examinations in the summer months or winter months? After running the seasonplot() function in R, I determine there’s no seasonality to this dataset as the troughs and peaks are more random as it exponentially increases.

**ARIMA Models**

For the ARIMA model, I had to first evaluate the data through the autocorrelation function (ACF) and partial autocorrelation function (PACF), which are plots that graphically summarize the strength of a relationship with an observation in a time series with observations at prior time steps (or lags), according to Brownlee (2017). Plotting these graphs revealed a tendency toward auto regression – indicating a moving average process, as indicated by the ACF graph. In the PACF graph below, you can see a couple of large spikes in the first couple of lags then smoothing off afterward.

|  |  |
| --- | --- |
|  |  |
| ACF plot of full, clean dataset | PACF plot of full, clean dataset |

Because there was an outlier (the Oct. 2008 Hurricane) and the data has an exponential trend, I had to make the data stationary, or without a trend, one main component of using ARIMA. This resulted in a data transformation called differencing, which addresses the trend data and the anomaly (the outlier) by comparing the difference between consecutive observations. R determined one order of diffing would be sufficient, which essentially transformed this data that has a trend and drift – both of which are present in this data. The resulting diffing graph is shown below, which now shows a stationary representation of the data.

|  |
| --- |
|  |
| Differencing (or diffing plot) after one order of diffing |

The next progression of analysis included experimentation with the ARIMA package to find the best forecasting model. The first model (ARIMA #1) used R to determine the parameters (p, d, q values). The parameters for the second model (ARIMA #2) and the third model (ARIMA #3) were determined by me, as the data scientist, which included p,d,q values of (2,1,1) and (1,2,1), respectively. For each of the models, I plotted against the residuals in ACF and PACF graphs which would show me the potential variance (or accuracies) in these predictive models. As you can see from the graphs below, the ARIMA #1 and ARIMA #3 models demonstrate the most likely predictions for future incoming examinations, as ARIMA #2 flattens out in 2014, which doesn’t seem likely according to the current trend.

|  |
| --- |
|  |
| ARIMA #1 model forecasting the next 12 months (Jan-Dec 2014) |

|  |  |
| --- | --- |
|  |  |
| ARIMA #2 model forecasting the next 12 months (Jan-Dec 2014) | ARIMA #3 model forecasting the next 12 months (Jan-Dec 2014) |

**Holt-Winters Forecasting Model**

Choosing Holt-Winters was a bit simpler, as Kabacoff (2015) states “a triple exponential model (also called a Holt-Winters exponential smoothing) fits a time series with both a level, trend and seasonal components” and includes an irregular component at time i. While I determined no seasonal components to this data set, I still used the Holt-Winters exponential smoothing technique as an increasing trend and an outlier at Oct 2008 existed in this data.

The result of this analysis can be seen below in a graphical representation of the Holt-Winters exponential model. The black jagged line is the original time series data. The green line is the Holt-Winter smoothing line which continues with the predictive values for Jan-Dec 2014, as represented by the blue line. The two red lines represents the lower- and upper-bounds of the Holt-Winters predictive model with an 80% confidence level.

|  |
| --- |
|  |
| Holt-Winters exponential smoothing model predicts the next 12 months (Jan-Dec 2014). |

**Appendix**

**Appendix Figure 1**

The following conditions were determined to be cardiovascular or non-cardiovascular, based on searches on WebMD and other well-known health oriented websites.

|  |  |
| --- | --- |
| Cardiovascular conditions or tests   * Aortic valve stenosis * Angina * Arrhythmia * CAD (Acronym for coronery Artery Disease) * Cardiac * Cardiovascular * Chest Pain * coronery Artery Disease * Cor Pulmonale * Endocarditis * Heart * Heart Palpitations * Ischemic Heart Disease * MRI (This is a test to identify potential heart conditions, so I kept) * Myocardial Infarction * Myocarditis * Premature Ventricular Contraction * Stress Test (This is a test to identify potential heart conditions, so I kept) * Ventricular Septal Defect (VSD) * VSD (This is the same as Ventricular Septal Defect) | Non-Cardiovascular Conditions or Tests   * ENT (Ear Nose Throat) * Gastro * Hearing * OBGYN * Ophthalmology * PTSD |

**Appendix Figure 2**

The following shows the clean dataset, organized by year (from 2006-2014) and month, of the incoming examinations to the Abbeville, LA HC. To understand my final counts for May 2007, May/June/July 2013, December 2013 and imputed values for the 10 missing values, I’ve outlined my approach below.

* For the May 2007 data in the Violet, Baton Rouge and Lafayette tabs, I added those together (after removing 51 duplicates from Violet) with the current total of 107 (in the Abbeville tab) to get a new value for May 2007 of 472 exams.
  + May 2007 Total from Violet, Baton Rouge, Lafayette = 230 + 55 + 80 = 365. Added to current total of 107 = 365 + 107 = 472.
* For the 2013 data, I added those numbers together of each appropriate month (after removing 48 duplicates for Baton Rouge, 39 duplicates for New Orleans and 47 duplicates for Lafayette in May 2013 data), and added the totals for each of those months in 2013 in the Abbeville tab. I also found 2 missing exams for August, but those weren’t determined to be missing so I did not add to original total for August 2013.
  + May 2013 Total from New Orleans, Baton Rouge, Lafayette = 258 + 105 + 127 = 490. Add to current total of 4730 = 4730 + 490 = 5220.
  + June 2013 Total from New Orleans, Baton Rouge, Lafayette = 10 + 19 = 29. Added to current total of 4706 = 4706 + 29 = 4735.
  + July 2013 Total from New Orleans, Baton Rouge, Lafayette = 12 + 4 = 16. Add to current total of 5000 = 5000 + 16 = 5016
  + August 2013 Total from Lafayette = 2. Did not add back into Abbeville data tab.
* For the December 2013 data, I ended up with 5,933 exams after finding embedded cardiovascular condition codes.

**Imputing Values**

* Row 4 (March 2006) -- 467.6535 (rounded to 468)
* Row 7 (June 2006) -- 243.2199 (rounded to 243)
* Row 37 (Dec 2008) -- 1191.494 (rounded to 1191)
* Row 42 (May 2009) -- 1205.128 (rounded to 1205)
* Row 55 (June 2010) -- 1694.403 (rounded to 1694)
* Row 62 (Jan 2011) -- 1870.909 (rounded to 1871)
* Row 73 (Dec 2011) -- 2719.713 (rounded to 2720)
* Rows 49-51 (Dec 2009, Jan-Feb 2010) add up to 4550.434.
  + To impute these values, I determined the ratio of data points and used that ratio to convert three data points appropriately to add up to 5129. Here’s how I determined ratio:
    - Row 49 - 1434.621/4916.503 = 0.2918
    - Row 50 - 1938.639/4916.503 = 0.3943
    - Row 51 - 1543.243/4916.503 = 0.3139
  + Then I used that ratio to convert three data points appropriately to add up to 5129
    - 0.2918 \* 5129 = 1496.6422 ~ 1497
    - 0.3943 \* 5129 = 2022.3647 ~ 2022
    - 0.3139 \* 5129 = 1609.9931 ~ 1610

**Forecasted Values**

* After choosing ARIMA #1 as the most accurate forecasting model, I added projections for Jan-Dec 2014 at the bottom of the dataset and rounded to the nearest integer.

|  |  |  |
| --- | --- | --- |
| Year | Month | Incoming  Exams |
| 2006 | 1 | 362 |
| 2006 | 2 | 436 |
| 2006 | 3 | 468 |
| 2006 | 4 | 490 |
| 2006 | 5 | 508 |
| 2006 | 6 | 243 |
| 2006 | 7 | 393 |
| 2006 | 8 | 596 |
| 2006 | 9 | 634 |
| 2006 | 10 | 613 |
| 2006 | 11 | 545 |
| 2006 | 12 | 411 |
| 2007 | 1 | 398 |
| 2007 | 2 | 311 |
| 2007 | 3 | 664 |
| 2007 | 4 | 680 |
| 2007 | 5 | 472 |
| 2007 | 6 | 467 |
| 2007 | 7 | 566 |
| 2007 | 8 | 806 |
| 2007 | 9 | 732 |
| 2007 | 10 | 886 |
| 2007 | 11 | 776 |
| 2007 | 12 | 698 |
| 2008 | 1 | 875 |
| 2008 | 2 | 840 |
| 2008 | 3 | 724 |
| 2008 | 4 | 1115 |
| 2008 | 5 | 997 |
| 2008 | 6 | 775 |
| 2008 | 7 | 886 |
| 2008 | 8 | 1041 |
| 2008 | 9 | 1011 |
| 2008 | 10 | 3110 |
| 2008 | 11 | 939 |
| 2008 | 12 | 1191 |
| 2009 | 1 | 1004 |
| 2009 | 2 | 1065 |
| 2009 | 3 | 1263 |
| 2009 | 4 | 962 |
| 2009 | 5 | 1205 |
| 2009 | 6 | 1429 |
| 2009 | 7 | 1205 |
| 2009 | 8 | 890 |
| 2009 | 9 | 1320 |
| 2009 | 10 | 1276 |
| 2009 | 11 | 1757 |
| 2009 | 12 | 1497 |
| 2010 | 1 | 2022 |
| 2010 | 2 | 1610 |
| 2010 | 3 | 1578 |
| 2010 | 4 | 1604 |
| 2010 | 5 | 1758 |
| 2010 | 6 | 1694 |
| 2010 | 7 | 1457 |
| 2010 | 8 | 1607 |
| 2010 | 9 | 1808 |
| 2010 | 10 | 1866 |
| 2010 | 11 | 1934 |
| 2010 | 12 | 2294 |
| 2011 | 1 | 1871 |
| 2011 | 2 | 2334 |
| 2011 | 3 | 1973 |
| 2011 | 4 | 2262 |
| 2011 | 5 | 2259 |
| 2011 | 6 | 2217 |
| 2011 | 7 | 2739 |
| 2011 | 8 | 2772 |
| 2011 | 9 | 3383 |
| 2011 | 10 | 2869 |
| 2011 | 11 | 2239 |
| 2011 | 12 | 2720 |
| 2012 | 1 | 2789 |
| 2012 | 2 | 3455 |
| 2012 | 3 | 2940 |
| 2012 | 4 | 2968 |
| 2012 | 5 | 3466 |
| 2012 | 6 | 3037 |
| 2012 | 7 | 3946 |
| 2012 | 8 | 3459 |
| 2012 | 9 | 3446 |
| 2012 | 10 | 3258 |
| 2012 | 11 | 4729 |
| 2012 | 12 | 3694 |
| 2013 | 1 | 4610 |
| 2013 | 2 | 4841 |
| 2013 | 3 | 5172 |
| 2013 | 4 | 4351 |
| 2013 | 5 | 5220 |
| 2013 | 6 | 4735 |
| 2013 | 7 | 5016 |
| 2013 | 8 | 4978 |
| 2013 | 9 | 5008 |
| 2013 | 10 | 6094 |
| 2013 | 11 | 4874 |
| 2013 | 12 | 5933 |
| 2014 | 1 | 5482 |
| 2014 | 2 | 5674 |
| 2014 | 3 | 5692 |
| 2014 | 4 | 5757 |
| 2014 | 5 | 5808 |
| 2014 | 6 | 5864 |
| 2014 | 7 | 5918 |
| 2014 | 8 | 5973 |
| 2014 | 9 | 6028 |
| 2014 | 10 | 6083 |
| 2014 | 11 | 6137 |
| 2014 | 12 | 6192 |

**References**

Bansal, Dr. Gaurav. (2017). *Missing Values with Amelia.* Retrieved from [Guideline\_for\_imputing\_missing\_values\_with\_Ameliagb11292016.pptx.](Guidelines_for_imputing_missing_values_with_Ameliagb11292016.pptx.)

Kabacoff, Robert I. (2015). *R in Action: Data analysis and graphics with R.* Shelter Island, NY: Manning Publication, Inc.

Khachatryan, Davit. (October 2014). Fargo Health Group: Managing the Demand for Medical Examinations Using Predictive Analytics. Boston, MA: Harvard Business Publishing.

Khachatryan, Davit. (October 2014). Explanation of the Dataset.

Wikipedia. (2017). List of Louisiana Hurricanes (2000-present). Retrieved from <https://en.wikipedia.org/wiki/List_of_Louisiana_hurricanes_(2000%E2%80%93present)>.