Analytics Sample

Walk-through of an agile end-to-end analytics process

Prepared by:

Rick Pflasterer

[linkedin.com/in/rickjfp/](https://www.linkedin.com/in/rickjfp/)

**Table of Contents**

[Introduction 3](#_Toc50973710)

[1. Excel (Appendix A) 4](#_Toc50973711)

[VLOOKUP 4](#_Toc50973712)

[Pivot Table 4](#_Toc50973713)

[Data Preparation 5](#_Toc50973714)

[2. SQL/Relational Databases (Appendix B) 5](#_Toc50973715)

[Microsoft (MS) SQL Server Management Studio 5](#_Toc50973716)

[3. Tableau Visualization (Appendix C) 5](#_Toc50973717)

[User-Centric Design 5](#_Toc50973718)

[Business Intelligence / Dashboard Development 6](#_Toc50973719)

[4. Data Mining / Machine Learning (Also in Appendix A) 9](#_Toc50973720)

[Data Profiling 9](#_Toc50973721)

[XL Miner 10](#_Toc50973722)

[5. Innovation Philosophy 12](#_Toc50973723)

# Introduction

Modern organizations use a collection of best-in-class electronic systems to record critical business transactions. These systems typically provide basic reporting functions to summarize the values captured but more advanced analysis can lead to additional value for organizational improvement.

As technology evolves, there are constantly new ways to access and analyze business data, which requires analytics professionals to continuously learn and adopt new skills. This skill toolbox might include experience with a variety of software applications, programming languages and advanced modeling concepts to access, transform, and explore many different data sources.

The projects that I have completed as a professional in the health analytics industry commonly use protected data under HIPAA regulations. Instead of using an example from a past project, I prepared a sample case study using open source deidentified healthcare claims data from the public data enthusiast community, Kaggle.com\*.

This demo is designed to showcase an agile approach to data problem-solving by leveraging a handful of analytical skills.

Because I am using an open-source data set, I want to mention that the analysis outcomes will solely represent the results of the process and will not represent any “real” clinical or financial healthcare insights.

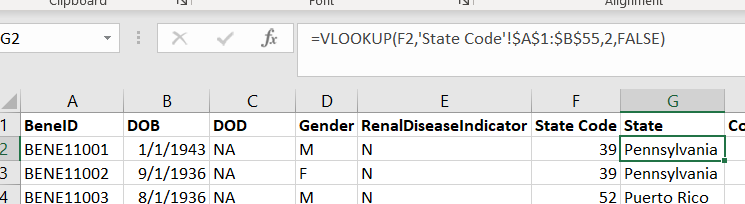
\*\* Appendices can be provided by email, or accessed in GitHub, upon request due to file sizes and upload restrictions.

\* Kaggle Data Set Reference: <https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis/data?select=Train_Outpatientdata-1542865627584.csv>

# 1. Excel (Appendix A)

### VLOOKUP

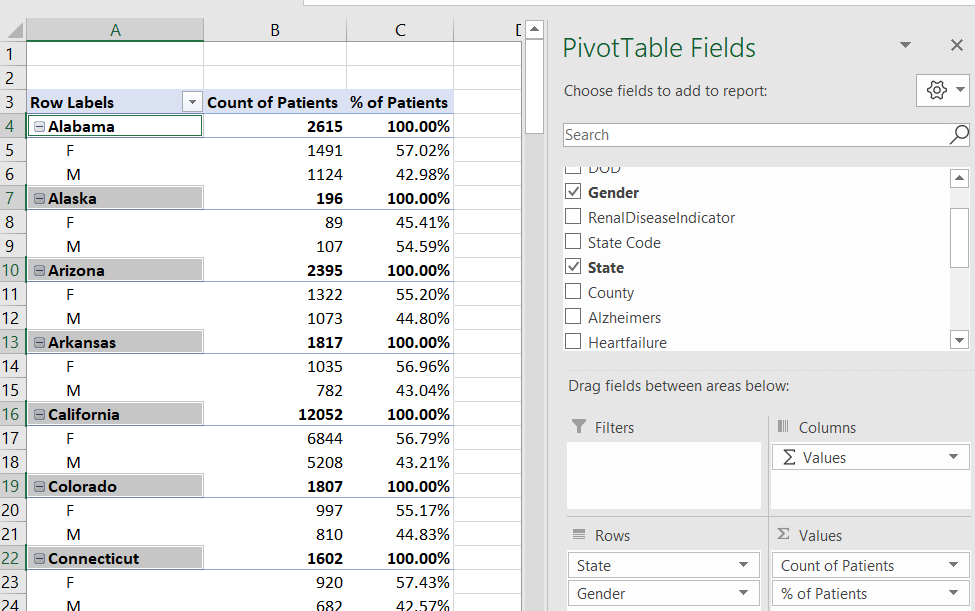
* The original patient data only includes a code field for US states, so a simple VLOOKUP was used to match the state with the code for more usable analysis (e.g. Pivot Table and Tableau)



*Figure 1.1 VLOOKUP demo for data aggregation and analysis*

### Pivot Table

* By using the VLOOKUP fields in our Pivot Table, we can look at different population distributions to summarize our data. Pivot tables can be useful for ad hoc analysis but excel has limitation when it comes to distribution and sharing insights (e.g. Problems with version control, locked files in shared network drives, and file corruption that could lead to loss of data).
* The brief example below shows patient counts and gender distributions for each state. I often use Excel to build Proof of Concept demos that can then be released to a wider audience through a Business Intelligence (BI) tool or data storage system.



*Figure 1.2 Pivot table example that can be used for exploration*

### Data Preparation

* Since the claims and client data came in separate files and were not formatted to be immediately presentable, Excel can be used to clean-up entry errors and reformat the data so it can be more easily analyzed in a BI tool or uploaded into a database.
* In this set, many of the blank fields came in as “NA” strings which works fine for many text data types but would produce an error while loading our “date” type fields if I failed to modify those entries.
* By reviewing the data in Excel, I can also start to plan out my relational model architecture by identifying the keys that can be used to connect the tables for more robust analysis.

# 2. SQL/Relational Databases (Appendix B)

### Microsoft (MS) SQL Server Management Studio

* I am using MS SQL Server Management Studio (SSMS) to connect to an MS SQL database (db) that I created as a local resource rather than using a cloud server.
* I imported my Excel extract files to create two new extract tables in the db.
* Using my extract tables as my data sources, I created SQL stored procedures that transform the raw data into the dimensional model for the business intelligence visuals.
* In the Tableau Visualization section, I talk about user requirements. I use those requirements to design the data architecture to ensure I meet the needs of the end users.

# 3. Tableau Visualization (Appendix C)

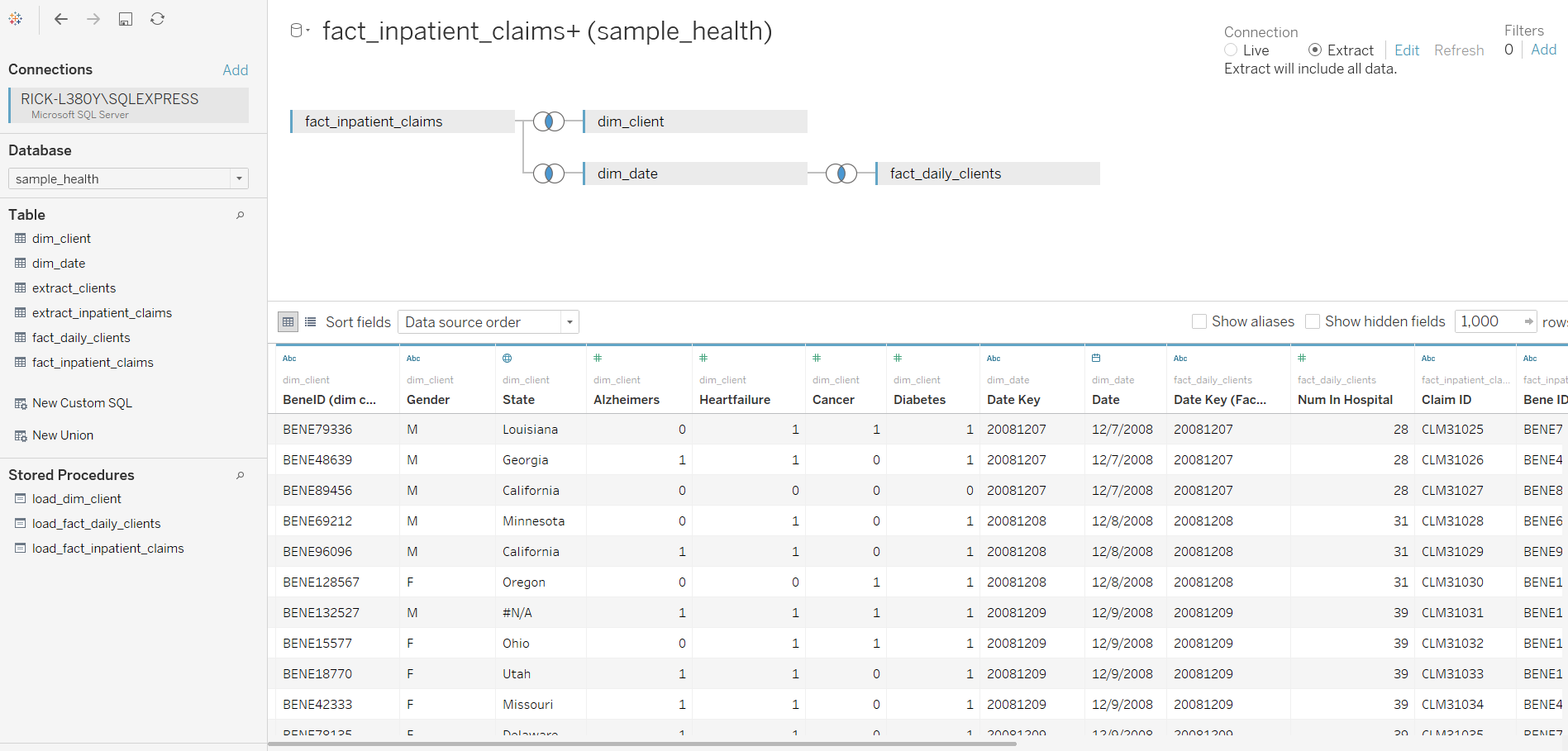
### User-Centric Design

* In this sample case, I designed two user profiles that each require different dashboards. The data is coming from the same dimensional model but is setup to meet their unique needs. The dashboards are designed to help turn insights into action.
  + **Department Manager**
    - Background: The Department Manager wants to understand the staffing capacity required for patient intakes during the day, and they want to understand how many clients are typically in the hospital each day.
    - Questions to Answer:
      * What is our average number of admissions per day?
        + Total, per month
      * How many people are in the hospital at a time?
        + Filter by date
    - Actions: By understanding the average admissions per day and the total volume of patients, the department manager can use the information to modify staffing and can make strategic decisions about capacity.

* + **Clinical Director**
    - Background: The Clinical Director is responsible for understanding the mix of services provided and they want to use Length of Stay as an indicator of improving performance.
    - Questions to Answer:
      * What are the total Number of Services provided?
        + Annual, monthly
        + Breakdown by service type
      * What is the total number of unique patients served?
        + Annual, monthly
      * What is the average length of stay?
        + Ability to adjust by month
        + Ability to adjust by procedure
    - Actions: By using the trends, the director wants to perform additional research in target areas to help design program bundles that integrate services and can lead to better value-based care outcomes. The director has also implemented new policies and wants to see a decrease in patient length of stay.
* Working collaboratively with the users is an important fusion in development. The users need to understand how the tool works and they need to understand how to use it to progress. A constant feedback cycle leads to iterative improvements and better adoption.

### Business Intelligence / Dashboard Development

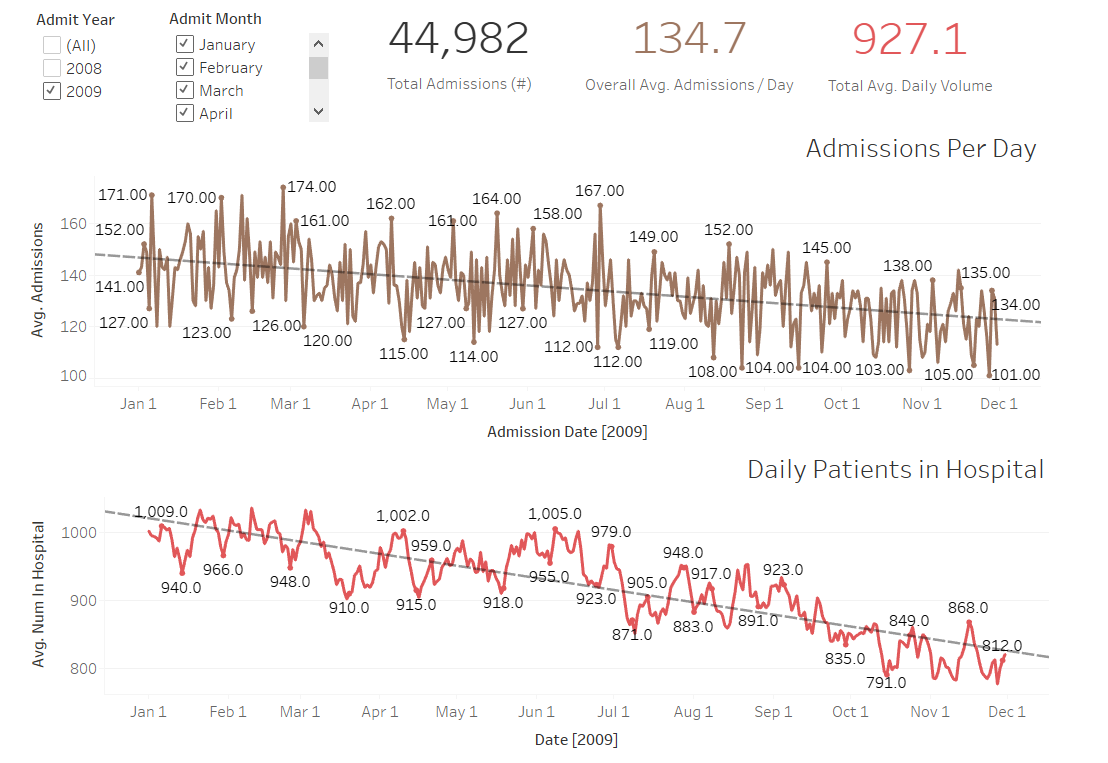
* I used Microsoft SQL Server (review SQL section for details) to create the right relational data model to answer stakeholder questions.



*Figure 3.1 Screenshot of the analytics dimensional model in Tableau.*

* I focus on creating clear and concise visuals that communicate the information without adding additional noise to distract from the purpose.
  + The dashboard uses appropriate spacing, sizing and colors to communicate relationships and important features.
  + The granularity of the data flows from top to bottom, with the highest-level metrics on top, and more details about those metric trends below.
  + I also aggregate filters in one area and use them to refresh all metric so users can see how each metric is related.

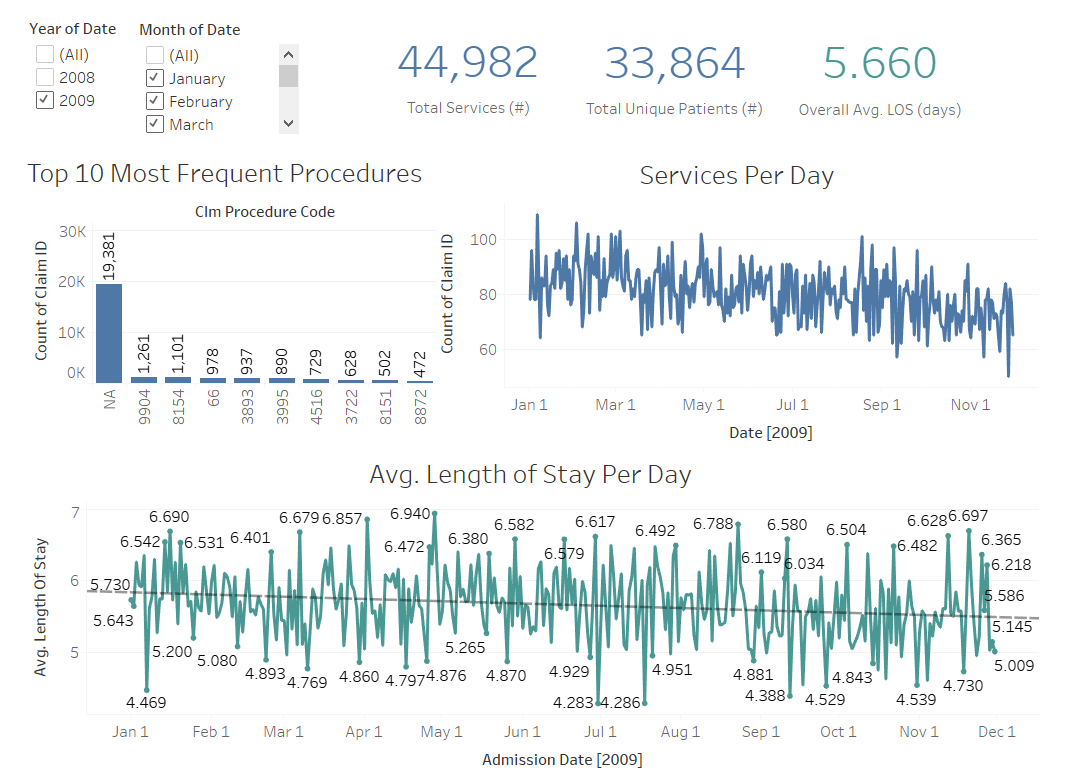
**Department Manager Dashboard**



*Figure 3.2 Department Manager dashboard deriving from the sample\_health database*

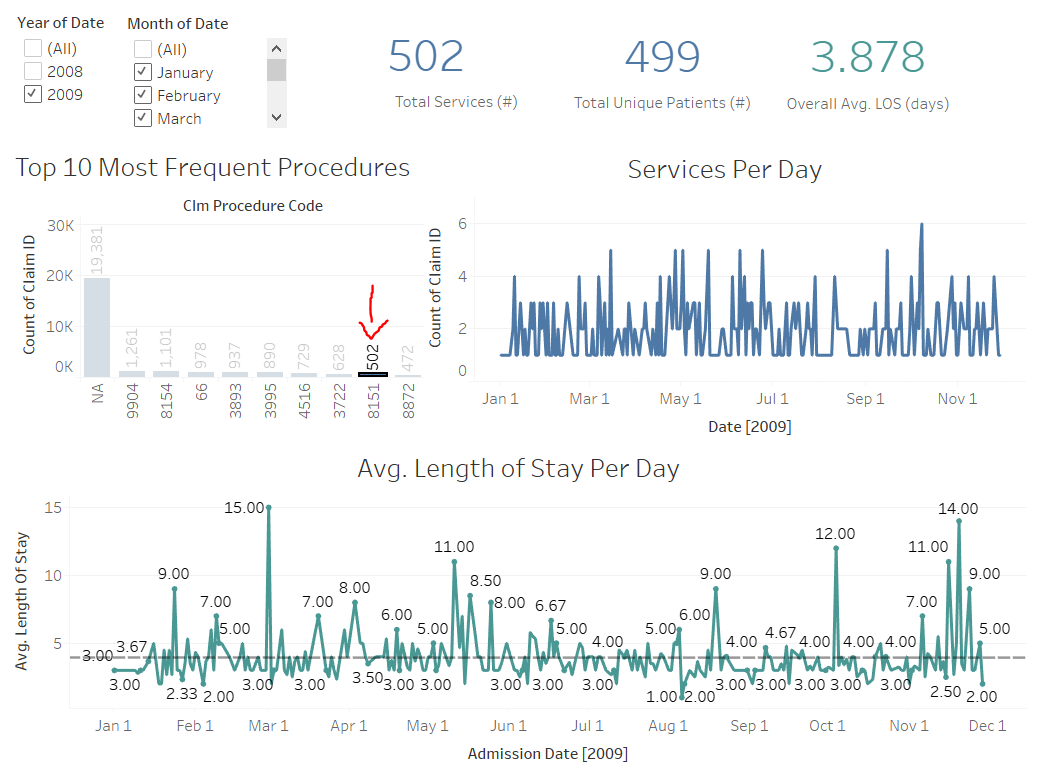
* The dashboard shows that both average admissions per day, and the total volume of patients in the hospital are decreasing. If this trend continues, I would be interested in working with the department managers to further investigate if there are opportunities to adjust staff hours and department bed capacity levels which could reduce/repurpose resources to improve efficiency without sacrificing patient care quality.

**Clinical Director Dashboard**



*Figure 3.3 Clinical Director dashboard deriving from the sample\_health database*

* The Clinical Director dashboard shows that the most common procedure code on our claims file is “NA”, with close to 20,000 incorrectly marked. This could be a critical problem if it were a real-world scenario. We would need to track down why these records are not being labeled correctly, as this could be a significant issue that impacts billing and reimbursement.
* The Length of Stay (LOS) at the hospital is decreasing, which aligns with the Director’s goal to reduce that metric. Since these visuals are connected, when clicking on a smaller subset of the data, the visuals around it adjust to show the values pertaining to that subset. When we do this, we notice that procedure code ‘8151’ is actually experiencing a slight increase in Length of Stay. This could lead to a more in-depth review of the operations for that procedure to identify opportunities for improvements (see below for filtered view).



*Figure 3.4 Filtered view of the Clinical Director Dashboard, showing a slight increasing trend for one procedure code.*

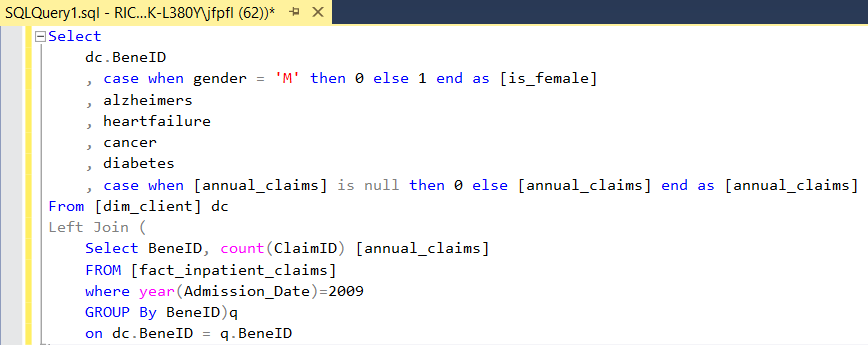
# 4. Data Mining / Machine Learning (Also in Appendix A)

### Data Profiling

* The goal of data mining, i.e. Machine Learning (ML), is to go from descriptive analytics, which helps us understand past trends, to predictive and prescriptive analytics approaches where we want to use past trends to predict outcomes as well as create algorithms that help us make decisions.
* Storing data in a tabular form is cost efficient to preserve resources, analytics models help with processing data more quickly to make resource-intensive calculations more readily available to users, and data mining / machine learning (ML) requires aggregated data sets that are typically consolidated into a single table before the model is run.
* The profiling stage is an opportunity to take our aggregated data and get to better understand how the data is related.

### XL Miner

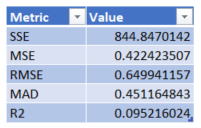
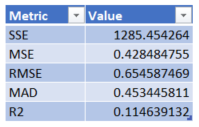
* XL Miner is an Excel add-on that allows you to run predictive modeling from a spreadsheet. For this example, I ran a SQL query to pull training data on client demographics and claims data for year 2009.



*Figure 4.1 SQL code to pull the data set needed for Linear Regression*

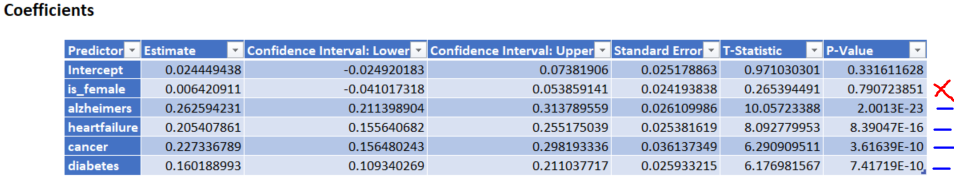
* I want to use gender and the list of underlying chronic conditions to see if I can create an algorithm that predicts the number of annual inpatient services a client will have (data set and work is included in Appendix A excel file). I am using Multiple Linear Regression analysis.
  + ml\_raw = the initial dataset pulled from the database. Cut data set down to only 5000 entries for validation and processing efficiency.
  + A value of “1” means “yes” for all fields, 0 means “no”.

**Training Results vs. Validation Results**



*Figure 4.2 Training results shown on left, validation shown on the right*

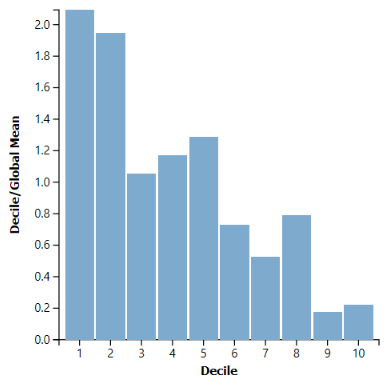
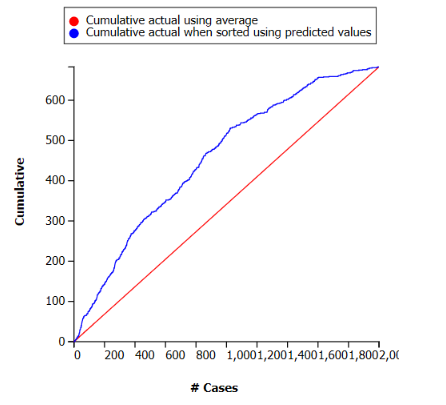
* The R2 value shows the percentage of the output that can be attributed to the regression equation. 0.115 for training and then 0.095 for validation partitions means that it is not able to predict a significant amount of the outcomes with just these variables. Target range for prediction is typically 0.50 – 0.99.



*Figure 4.3 Coefficient table uses the statistical p-value to evaluate if a value is significant*

* When running the model, I use p-value = 0.05 as a cutoff to indicate if the variable is significant.
  + Alzheimers, heartfailure, cancer, and diabetes are all significant but gender is reportedly not a variable to consider in the equation.

**Validation Partition Performance**



*Figure 4.4 Gain Chart shown on the left, the Decile Chart on the right*

* The Gain Chart shows that the algorithm (blue line) performs better than a completely random guess (red line). This means the algorithm does provide some predictive power over a random assignment.
* The Decile Chart shows a pattern that the top two deciles (most probable) are over 2x more likely to be correct using the algorithm than using a random guess. A downward staircase pattern is ideal because it means the algorithm can predict the most predictable entries relatively well and as the entries become more difficult to determine, the prediction accuracy decreases. The prediction power tapers off at around the third decile indicates poorer performance.

**Conclusion**

* I removed the [is\_female] field because the model found that it was not a significant predictor and reran the regression to discover this algorithm:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Estimate** | **P-value (significance)** |
| **Intercept** | 0.027961 | 0.191837413 |
| **alzheimers** | 0.262633 | 1.93916E-23 |
| **heartfailure** | 0.205567 | 7.75401E-16 |
| **cancer** | 0.226777 | 3.70465E-10 |
| **diabetes** | 0.160354 | 7.0077E-10 |

**# of Annual Inpatient Admissions** =

0.028

+ 0.263 [Has Alzheimers]

+ 0.206 [Has Heart Failure]

+ 0.227 [Has Cancer]

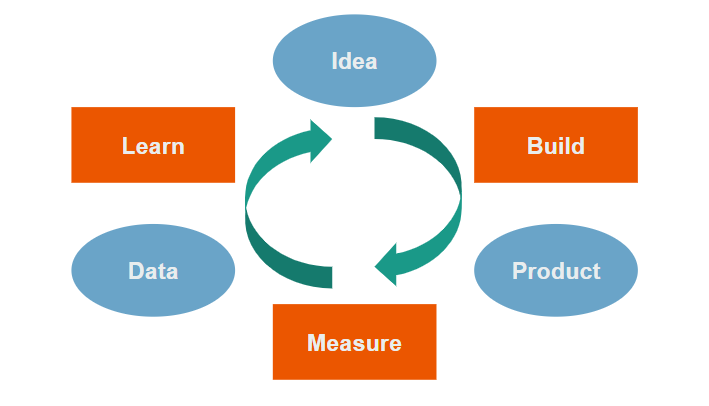
+ 0.16 [Has Diabetes]

*Figure 4.5 Results of the Linear Regression using XL Miner*

# 5. Innovation Philosophy

The Lean Start-Up by Eric Ries presents concepts that translate well into analytics product. Ries discusses the Build Measure Learn (BML) Feedback Loop as a framework that can fuel continuous creativity. The cycle relies on taking ideas, testing them, getting feedback, and then applying lessons learned to generate new ideas.

Rather than innovation being an unplanned spontaneous event, we can create environments where we make innovation repeatable and practiced. The BML framework can be applied to small scope projects, like developing a dashboard, or even larger undertakings, like developing a new product line. Having the willingness to learn and make continuous improvements is a key philosophy for personal and professional evolution.



*Figure 5.1 Diagram of the Build Measure Learn Feedback Loop from The Lean Start-Up by Eric Ries*