THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF INDUSTRIAL AND MANUFACTURING ENGINEERING

HEALTHCARE ANALYSIS ON LENGTH OF STAY

MATTHEW DIGEL
FALL 2014

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Industrial Engineering
with honors in Industrial Engineering

Reviewed and approved* by the following:

Soundar Kumara
Allen, E., and Alen, M., Pearce Professor of Industrial Engineering and Computer Science
Thesis Supervisor

Paul Griffin
Del Pazzo Head of Industrial Engineering
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

This thesis is focused on how data analytics and big data are effecting organizations in all fields and then provides an example of how data analytics can provide insights to a hospital's operations and forecasting. Organization's decisions now have the chance to be much more informed with the enormous amount of data being collected constantly throughout our daily lives. From both a competitive and customer satisfaction perspective organizations desire to make the most informed decisions possible. This thesis discusses organizations who have embraced the big data movement to make more "data-driven" decisions that have created impressive results and helped achieve their goals. Hospitals are one type of organization that has the potential to greatly benefit from "data-driven" decision making. More specifically, if hospitals can better predict their patient's length of stay they can make more informed decisions regarding bed capacity and nurse staffing. In this thesis a data set with historical medical patient records is analyzed to find trends among the different fields in the data relative to a patient's length of stay. After an initial analysis of all the fields, ten clusters of patients with extreme length of stay tendencies were found. These clusters can serve to provide the hospital operations team with information about how long specific patients are likely to stay in the hospital to make more informed decisions that will reduce operating costs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction: Big Data

Since the dawn of the internet and the digitalization of many aspects of our lives, businesses have become increasingly interested in converting the enormous amount of data generated every day and insights taken from the data into better business outcomes. This strategic challenge arises from businesses' desire to make the most informed decisions possible. These decisions now have the chance to be more informed with data being collected constantly throughout our daily lives. From credit cards, to computers and smart phones; from the sensors on the trains, buses and bridges of cities; from the sensors on industrial equipment, automobiles, electrical meters and shipping boxes that can measure movement, vibration, temperature, and humidity(Shaw). With all data points considered it is easy to understand how, according to IBM, 90% of the data in the world has been created in the last two years alone (IBM). This influx of new information and the tools available to help draw insights gives businesses the potential to better understand their customers, industry and products.

Traditionally data analysis for businesses only focused on what is called structured data. Structured data includes financial information, transaction records, and interaction channels (call center or point-of-sale). Now due to smart phones, GPS, sensors and social networks, web traffic, location data and social media comments are probed through to make decisions, cut costs and increase sales. This new type of data content can be categorized into two groups. Unstructured data tends to be text heavy and uneasily organized usually coming from social

media comments. Multi-structured data is more complicated because it includes information from visual images along with texts and is usually derived from web applications. The data collected from sensors falls into the structured category because of its predictable nature. In addition, global analyst firm Gartner uses simpler terms to frame big data strategy. For example, "volume" (the amount of data), "velocity" (the speed of information generated) and "variety" (the kind of data available) are used to help Gartner analysts conceptually understand the data they are working with. It is this framework that many businesses are beginning to adapt themselves.

A recent study, cited on IBM's big data web page, found that 58% of executives found moving from data to insight to be a major business challenge. This is because some organizations lack an effective approach to capture data where others do not have the tools or expertise to apply it for useful insight. Investing in big data software platforms can help executives with this challenge. The investment in big data analytics proved to be effective according to a study done by Erik Brynjolfsson, an economist at Massachusetts Institute of Technology's Sloan School of Management. Professor Brynjolfsson studied 179 large companies and found that those adopting "data-driven decision making" achieved productivity gains that were 5 to 6 percent higher than other factors could explain (Lohr). Large tech companies like IBM hope to capitalize on this business need by providing big data software platforms to companies that don't have the means to do so themselves. Many tech savvy companies have figured out how to draw insights from big data on their own.

Across a variety of industries large companies with the ability to fund big data analytics projects themselves have been able to achieve impressive results. Google analyzed clusters of

searched words by geographical region to predict flu outbreaks. Uber is constantly tasked with analyzing urban traffic flows to become better at predicting where and when customer demand will arise. Target got a lot of media in the late 2000's for an algorithm it used to recognize when women had a high likelihood of being pregnant by tracking purchased items like unscented lotions and offering those customers discounts on baby related items. Business-analytic professionals at credit-card companies were able to recognize a pattern that people who buy anti-scuff pads for furniture are more likely to default on their payments. Shipping companies like FedEx, probe through truck delivery times and traffic patterns to optimize routing. Match.com constantly mines through their member's personal characteristics and reactions to improve on their already immensely successful algorithms for matching people on dates. Even Melissa Lora, president of Yum Brands Inc.'s Taco Bell International says "We find it invaluable to have people who can synthesize data" because she needs employees to sort data on service speed, product quality and social media. The impact of big data will generate much more than the insights for businesses. Gary King, director of Harvard's Institute for Quantitative Social Science says, "It's a revolution. We're really just getting under way. But the march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government. There is no area that is going to be untouched."

Virtually every area of work will be affected by big data analytics. Political campaigns have begun to hire political analytics firms like Deep Root Analytics or Blue State Digital to target key voters with machine like precision. During the 2008 Presidential Elections, Obama's election campaign team made extensive use of data mining and microtargeting to get the President reelected. Obama's embrace of big data lead to Washington starting Data.gov in 2009,

a site that opens the door to a variety of government data to the public. Since the launch of data.gov industries have benefited from and industrious people have done smart things with this open data. For example, agriculture is one industry that depends on public data for its predictions and products. The access to historical crime data has resulted in applications for police officers that predict where and when crimes are most likely to occur. Even the study of law has been influenced by the big data movement. For example, Kevin Quinn, previously an assistant professor of government at Harvard, found that his statistical model, based on six variables from past cases, could better predict the outcome of Supreme Court Cases than the judgments of 87 law professors in a year. Overall, it's fair to say that the practice of big data analytics is affecting any area with people that know how to harness its potential.

Graduate schools are being to take notice to this big data trend as well. In the last two years five business schools have rolled out business analytics programs to focus on the discipline of using data to solve problems and make decisions. According to the Wall Street Journal, the Massachusetts Institute of Technology is in the middle of creating a new program called a Masters in Analytics to be offered within the Operations Research Center. This fall the University of Southern California began an analytics program with 30 students and expects 50 to 60 students to enroll next year (Gellman). Northwestern University's Kellogg School of Management is taking a different approach. Instead of creating entirely new programs specifically for analytics, Kellogg offers several courses in analytics to a variety of graduate programs. Regardless of how universities choose to structure their analytic education, it is apparent that this skill is becoming more sought after and valuable to organizations.

It is important to note the drawbacks and potential risks that can result from relying too much on data. With large data sets and detailed measurement there is always a risk of generating false discoveries. Trevor Hastie, a statistics professor at Stanford, says "The trouble with seeking a meaningful needle in massive haystacks of data is that many bits of straw look like needles." These false discoveries can often result from the inaccurate assumption that correlation between variables means that one of those variables is effected by the other, in other words, one variable has causality to the other. An example of how correlation does not imply causation can be seen in Figure 1. how from 2000 – 2009, divorce rates in Maine and per capita consumption of margarine in the US had an almost perfect correlation of .9925. Two extremely unrelated variables have a strong correlation, but this does not mean that one effects the other. In addition, data is understood and then used through mathematical models. These models, regardless of the number of variables taken into consideration, are still extreme simplifications of the real world. Insights generated from these mathematical models are not always accurate and this possibility of error should always be considered when practicing "data driven" decision-making.

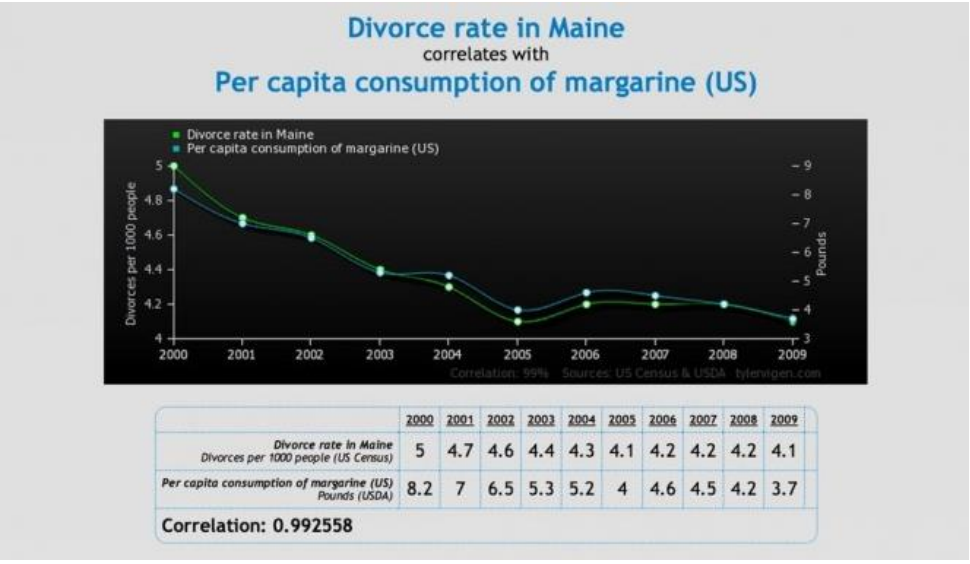

Figure 1. Divorce Rate vs. Consumption of Margarine (Abraham)

The increases in the amount of data available and the analytic tools have created an opportunity for organizations to capture value from their information assets through rigorous analysis enabling them to make more informed decisions. The "data driven" decision-making culture is quickly invading all areas of work because of its potential to improve performance. Though the practice is not perfect, basing decisions off data has begun and will continue to be a best practice in most disciplines.

1.1 Big Data Use in the Health Care Industry

To survive after the passing of The Patient Protection and Affordable Care Act (Obamacare) many healthcare providers are adopting radical process improvement and quality missions that can be aided by the use of data analytics. Their focus has shifted toward what the Institute of Healthcare Improvement calls the Triple Aim: "enhance the patient experience, improve the health of populations, and reduce the per capita cost of health care." The regulations within Obamacare will increase the data generation among all the healthcare players creating an opportunity for executives to make data-driven decisions.

The health care industry today is made up of a wide range of players all of whom are likely to benefit from the implementation of data-driven decision making. The key players in the industry are as follows: payers (insurance companies, employers and the government), providers (entities that provide health services, clinicians, hospitals and ambulatory centers), manufacturers (medical device, pharmaceutical, medical supply and technology companies), policy makers (federal and state government and advocacy groups) and finally the Patients. Obamacare has many implications on the listed players that overall will force the industry to increase access and

quality of care while also reducing costs. Over the next few years around 25 – 30 million new Americans will have health insurance. This increase in potential patients and the data about those patients along with their doctors has created specific trends within the industry. To start, patients are able to act more like costumers because of public access to plan/provider options taking into account quality scores and increased awareness of whether procedures or tests are medically necessary. This has made the industry much more transparent than it was in the past. In addition, large data-sets of patient information known as electronic medical records have become useful for improving clinical trials, forming standard operating procedures, creating algorithms for best practice and drug development/testing efficiency. Throughout America innovators are making good use out of data analytics to improve the health of others.

In June 2014, Fast Company magazine released an issue listing the 100 most creative people in business. Of those top 100, three of them worked in improving healthcare using data analytics. Joel Dudley, a biomedical informaticist at the Icahn School of Medicine at Mount Sinai, created models that could predict the most effective cancer therapies based on unique molecular patterns in tumors. Linda Avey, Cofounder of Curious Inc, created an online forum for patients to share their past health data and interpret it on an open forum. Linda believes that "the more [data] we have the more we're going to learn. The question is, How do we probe our bodies in interesting ways and quantify what's going on?" Anmol Madan, Cofounder of Ginger.io app, has created a new way to interact with your primary care physician. Through sensors and surveys on a patient's smart phone the app collects data. So now instead of a primary doctor sampling a patient twice a year, they can sample patients every ten minutes. These three creatives are examples of how people are beginning to use data toward improving people's health.

Hospitals can significantly benefit from data driven decision making as well. Obamacare installed a new rule into Medicare and Medicaid that penalizes hospitals for patients that are readmitted to the hospital within 30 days of being released. By analyzing a patient's health history through their electronic medical records (EMR), hospitals can better predict regression rates of patients with certain characteristics to avoid premature releases. In addition, this real-time patient data can be used to figure out which symptoms are signs of emerging conditions. The Washington Post gives an example, "a 77-year-old female recovering from a partial hip replacement has a high temperature, elevated respiratory rate, and dramatically low blood pressure (Scola). Subtle combined variations in these vitals can indicate the early signs of sepsis, a potentially deadly infection" (Washington Post). More specifically, Hospital operations can reduce waste and cost by taking advantage of these EMR analysis software and platforms that companies like IBM, Oracle and SAS provide. A patient's length of stay is important metric for hospitals when trying to forecast how many beds will be available in the coming weeks. Data analytics potential to help hospital operations predict a patient's length of stay will be the focus of the remaining of this thesis.

1.2 Organization of the Thesis

In chapter 2 undertakes an explanation of the length of stay problem. Chapter 3 discusses the methodology and in chapter 4 we discuss the results and visualization. Finally in chapter 5 recommendations and conclusions are discussed.

Chapter 2

Problem Definition and Description

The main responsibility of hospitals is to provide patients with the most effective treatment with the highest quality of care. The word hospital means an establishment where injured or sick people occupy space and are treated. The most commonly accepted criteria for classification of hospitals in America are the following: length of stay (LoS) of a patient, clinical basis, ownership, size, objectives, management and system of medicine. Some examples of different types of hospitals are Central Government Hospitals, Private Hospitals, Public Hospitals, Corporate Hospitals and hospitals based on service. At all these types of hospitals, LoS is an important metric when forming and improving financial and operational strategy. A patient might stay for a short period of time for treatments of disease or injury that is "acute" in nature, such as ulcers, peptic, pneumonia, or a broken bone. On the other hand a patient might stay for a while with treatment of disease or injury that is chronic in nature or severe such as cancer, tuberculosis, leprosy or the breaking of multiple bones. Knowing patients potential LoS is critical to reducing waste within a hospital.

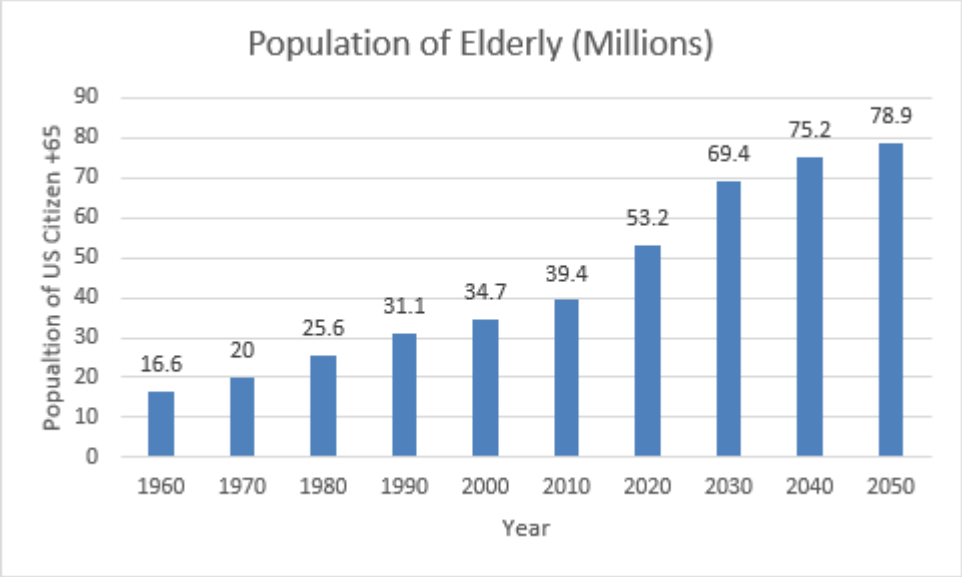

Figure 2. Population of Eldery (US Bureau of Census)

Hospitals face severe cost pressures due to Obamacare at a time when demographic shifts are increasing inpatient utilization. Figure 1 illustrates the increasing percent of Americans over the age of 50. This increase in age is estimated to increase inpatient utilization by 20% in the next 20 years (Milliman). Shortages of nurses, therapists and medical record personal will also add to the increased costs hospitals will face along with the necessity to upgrade management software. In the past hospitals operating margins averaged 3%, but now to finance their operations need 5% margins (Milliman). To increase margins, decision makers need to get more patients in and out of the hospital. The way to accomplish this is through the development of better patient flow strategies. The best way to increase patient flow is to decrease inpatient lengths of stay. In addition, by decreasing and controlling patient length of stay the hospital gains control over forecasting bed utilization. By being able to accurately predict how many beds are in use at a specific time in the future, hospitals can more efficiently staff their nurses, therapists and medical record personal and predict the usage of other important resources.

2.1 Description of Data Set

With an anonymized data set from a hospital in Pennsylvania the goal of this analysis is to find patterns in patient attributes that can be used to predict patients LoS upon diagnosis. With an improved ability to forecast LoS this hospital will be able to make more informed decisions about how to staff their employees and how many new patients they can sign in. The given data set has a variety of different patient attributes that all might affect LoS. To start the data provides basic patient demographic information such as gender, age, ethnic group and county. This demographic information will be helpful with finding more obvious patterns, for example, perhaps the older someone is the more likely they are to have a longer LoS. Admission time into the hospital is included as well and can be useful for picking up on trends of illness over a specific period of time. The most important attribute when trying to predict LoS is always the diagnosis of the patient. The data set provides a vast amount of information regarding this topic. These diagnosis related attributes include complaints, severity level, primary ICD9 (diagnosis code), admission diagnosis and diagnosis group (higher level diagnosis code), days since last admission, Emergency Department duration, patient service line and surgery name. In addition, the data set includes insurance information and historical medical records. The following analysis focuses on the patient demographics and initial diagnosis for pattern discovery.

Chapter 3

Analysis Methodology

The hospital provided a year's worth of medical data with a variety of patient characteristics to search for patterns to predict the LoS of patients. The majority of the analysis was done in Microsoft Excel with a few graphics generated in the statistical software R. The goal of the analysis was to find ten clusters of patients that could be flagged as patients with a high chance of having either a very large LoS or very small LoS. In this analysis a patient cluster is defined as patients with more than two similar characteristics. For example, patients might be in a cluster together because they gave the same initial compliant, are the same gender and will be receiving the same surgery. By identify clusters of patients that have extremely long or short LoS the hospital can have a better grasp on how long their patients will stay in the hospital.

The first step in the analysis was to compare each field in the data set to LoS. In this description of the analysis methodology, fields are defined as the columns in the data set. The fields that were analyzed relative to LoS were age, BMI, gender, acuity level, ICD9 codes, initial chief complaint, hospital service line, time spent in the emergency department, surgery received, a flag indicating whether the patient had come from the emergency department, and insurance company. To analyze a specific field within Excel, all the rows for the field were copy and pasted into a separate sheet along with the respective LoS metric and the Encounter ID. From this point the analysis was handled differently according to nature of the field. For continuous data like age, BMI and emergency duration the data was rounded to integers so the average LoS could be found for each integer. These averages were then graphed on a scatter plot to be examined for general trends. For the fields that only had a few different characteristics like gender, acuity level, and the emergency department flag the averages, standard

deviations and patient count was determined using Excel's pivot tables and then displayed on a bar graph. This bar graph was then examined for noticeable differences in LoS between the few characteristics. Finally, the last field type had a high number of characteristics. These fields included ICD9 codes, initial chief complaint, hospital service line, surgery received and insurance company. As was done with the other field types, the field being analyzed, LoS and Encounter ID columns were copy and pasted into a separate sheet from the base data. Again, the average LoS, standard deviation of LoS and patient count was determined using Excels pivot table feature. From there the characteristics were ranked from most occurrences to least occurrences. For example, the most frequent chief complaint was "Short of Breath" which occurred 1598 times. All characteristics that had a significant sample size of 30 occurrences were extracted and made into a new list. If there were less than 50 characteristics with at least 30 occurrences then the characteristics with the top 50 occurrences were included in this new list. Using the list of characteristics with at least 30 occurrences or top 50 most occurrences, the characteristics were then sorted in descending order by LoS. Then a new worksheet was created and two tables were made for each field. One table displays the top ten characteristics with the longest average LoS and the other table displays the top ten characteristics with the shortest LoS. After completing the initial analysis the next step was to find clusters using the trends discovered about each field.

Finding clusters of patients with extreme LoS tendencies involved a variety of methods. In some instances, guidance from the basic analysis lead to discovery of extreme clusters through filtering the three different fields in the base data. For example, the ICD9 code 995.91 ranked within the top ten longest LoS for all ICD9 codes with at least 30 occurrences, so the ICD9 field was filtered and for 995.91. Since males and patients with an acuity level of L1 tend to have a longer LoS these characteristics were filtered for in their respective fields as well. The result was the discovery of a cluster of 52 patients with a long LoS of 7.52 days. This process was repeated for an assortment of fields and characteristics to find more clusters. The other method involved a similar approach, but instead of using filters this method used

pivot tables to discover extreme LoS clusters. With the described methods ten clusters were discovered, five clusters tending to have extremely short LoS and five clusters tending to have extremely long LoS.

Chapter 4

Visualization and Results

The data analysis compared twelve fields in the data set to the LoS field. The following visualizations are a result of the initial analysis that looks for correlations between each field and LoS.

Patient Class:

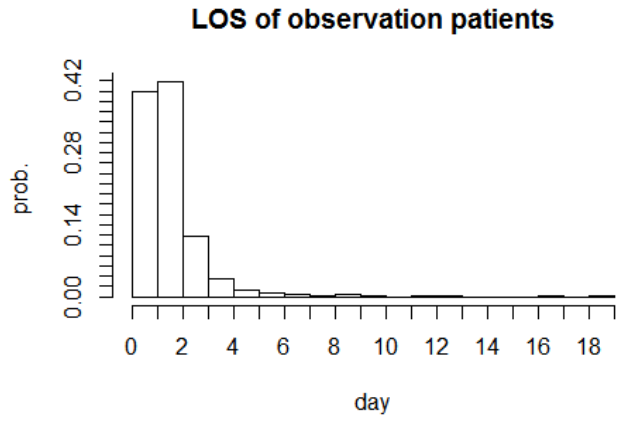

Figure 3. LoS of Observation Patients

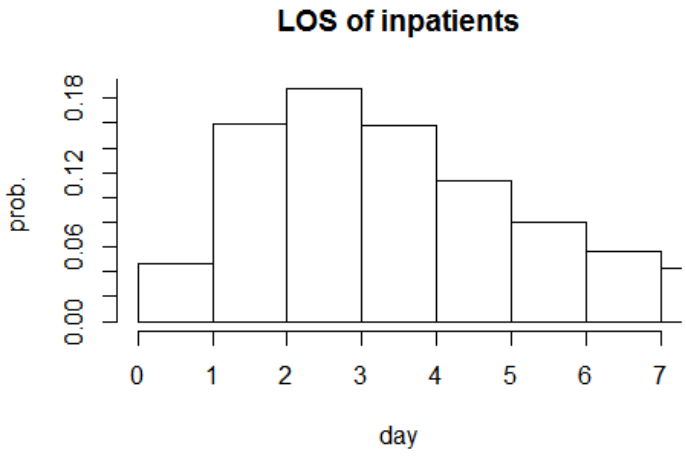

Figure 4. LoS of Inpatients

The analysis of patient class was the only field analyzed in R. The code used to generate these graphs can be found in the Appendix. This analysis verified that a majority of the time observation patients stay less than 2 days. This is convenient to know because of a recent healthcare policy passed called the Two – Midnight rule. This new policy is an attempt to provide hospitals more guidance with patient status determination issues which many hospitals struggle with daily. A patient is considered an outpatient or observation because they are expected to stay for less than two midnights as opposed to an inpatient that is expected to stay for more than two midnights. The class of a patient determines billing for Medicare so knowing this ahead of time can be useful in many ways financially. The histograms shows that 18.49% of patients placed in the observation class ended up staying more than two days as opposed to inpatients with only 20.54% that stayed less than 2 days. This analysis reaffirmed the expectation that inpatients stay more than two days and observations stay less than 2 days.

Age:

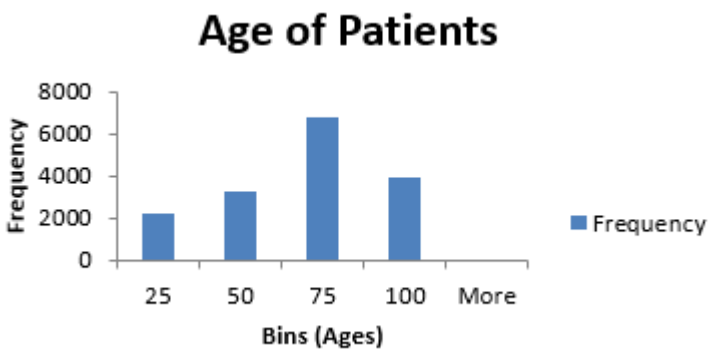

Figure 5. Age of Patients Histogram

This histogram shows the age distribution of the 1629 encounters with patients at the hospital. The age group that comes to this hospital the most are patients between the ages of 50-75. This could be

the result of the elderly being more likely to get sick or that the community surrounding the hospital tends to be older.

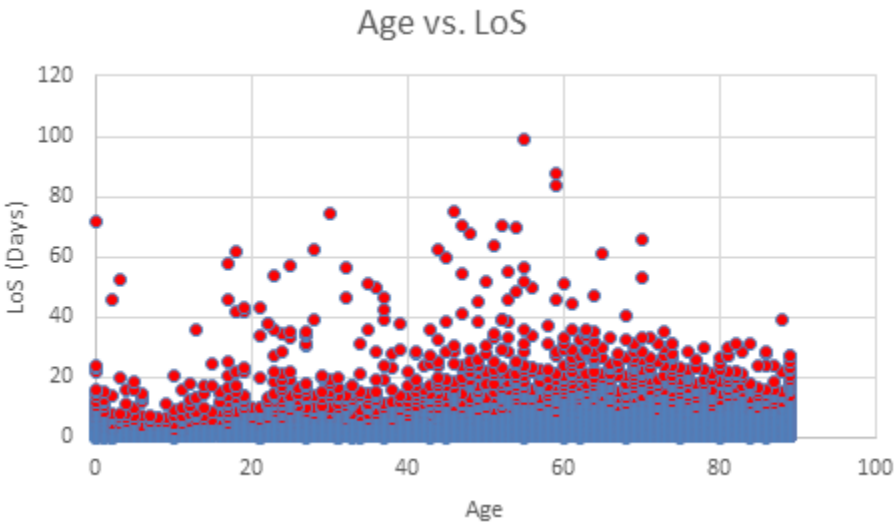

Figure 6. Age vs LoS Scatter Plot

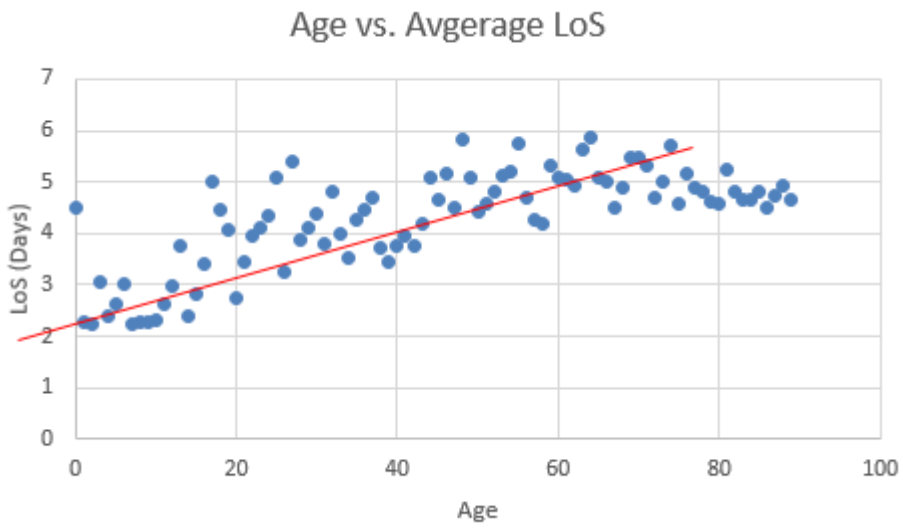

Figure 7. Age vs. Average LoS Scatter Plot

The analysis of LoS relative to different ages reveled two key takeaways. In Figure 6 it can be seen that the most amount of outliners occurred between the ages of 40 – 60. Figure 7 displays how on average the older someone is the longer their LoS is likely to be.

Gender:

Table 1. LoS Metrics by Gender

Gender	Average LoS	Std Dev	Count of Patients
Female	4.214	3.516	7745
Male	4.165	3.585	8228

An initial look at the average LoS and standard deviation of female verses male did not lead to any meaningful takeaways besides that gender doesn't seem to matter. Thus, further analysis was done to see if either gender has higher probability of staying more than 2 days. Below, in Figure 8 it can be seen that males have a slightly higher chance of staying for more than 2 days.

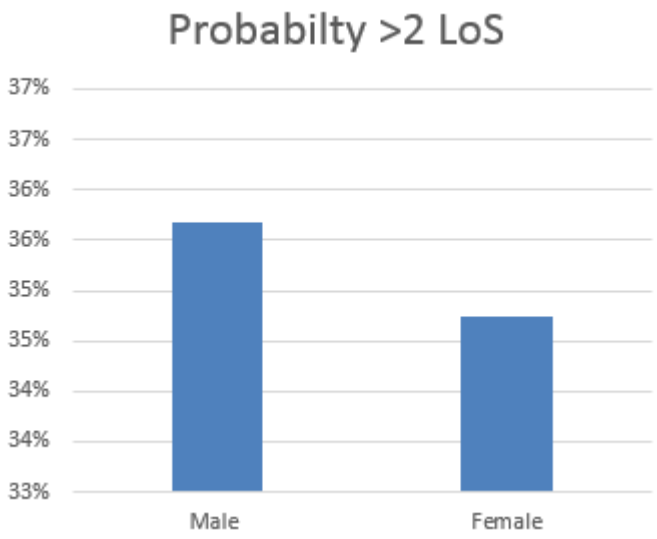

Figure 8. Probability LoS is Less Than Two For Gender

BMI:

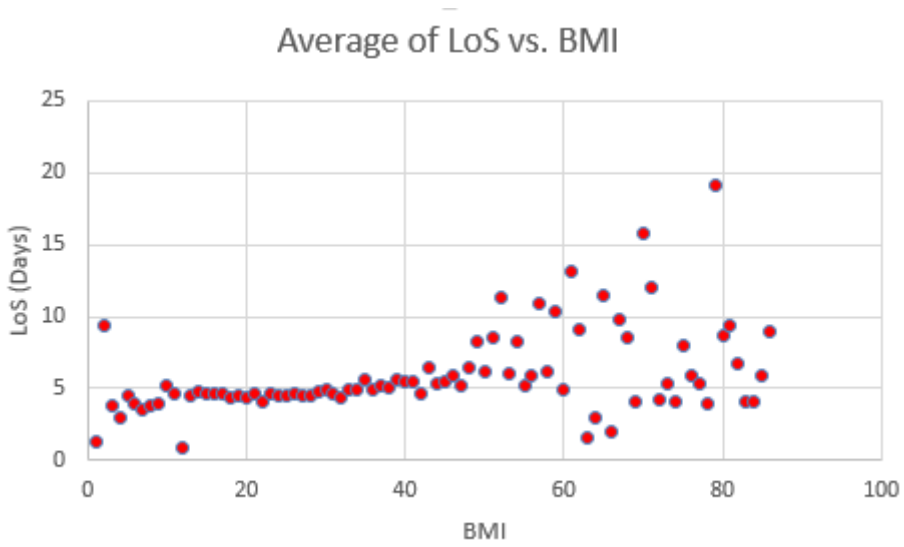

Figure 9. LoS vs. BMI

Figure 9 displays how as patient's BMI goes up the LoS becomes more sporadic while also increasing the average LoS. It should also be noted that entries with over 100 BMI and over 20 LoS were deleted to form a graph without extreme outliers.

Acuity Level:

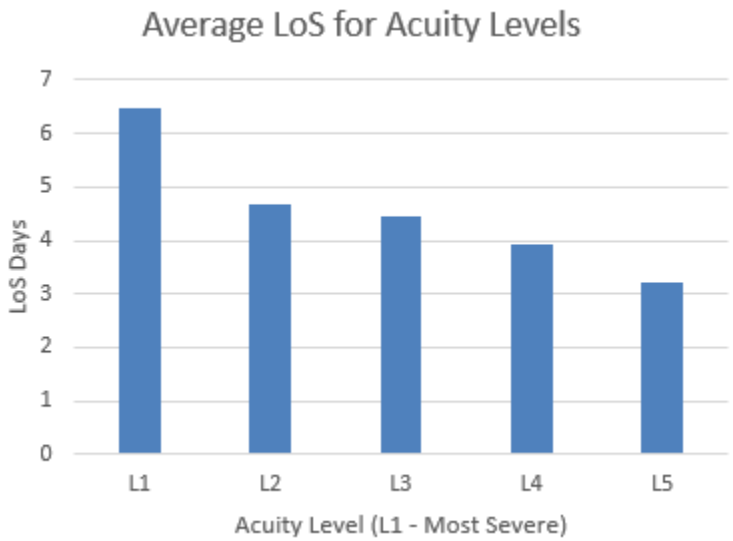

Figure 10. Average LoS for Acuity Levels

The analysis of acuity level proved to justify common sense. Acuity level is the rating of severity of illness or injury that a patient receives upon their diagnosis. Since L1 is the most severe rating it makes sense that it would have the highest average LoS. In addition, the rest of the acuity levels average LoS descends as the severity of illness decreases. Thus, acuity level and LoS are positively correlated.

ICD9:

Rank	ICD9 Code	Avg LoS	StdDev	Count
1	298.9	10.206	17.539	66
2	854.00	10.123	12.909	37
3	518.81	8.953	8.789	124
4	507.0	8.170	7.459	30
5	805.4	8.099	11.488	46
6	995.91	6.441	4.774	140
7	431	6.318	6.208	88
8	430	6.317	6.027	89
9	486	6.205	6.033	350
10	038.9	6.154	4.650	400

Figure 11. Top Ten Longest LoS of ICD9 Codes

Rank	ICD9 Code	Avg LoS	StdDev	Count
1	493.92	2.806	2.571	45
2	574.20	2.802	2.117	63
3	780.4	2.573	2.340	38
4	427.32	2.458	1.339	31
5	435.9	2.223	1.580	167
6	977.9	2.099	2.048	126
7	541	2.088	1.802	67
8	540.9	2.078	2.064	51
9	786.5	1.911	2.104	673
10	850.9	1.622	2.429	100

Figure 12. Top Ten Shortest LoS of ICD9 Codes

Within the patient data set there were 1157 different ICD9 codes. Upon finding the LoS metrics for each code the above two figures were taken from a list that only includes ICD9 codes that have a sample size of at least 30. There were 96 IDC9 codes with statistically significant samples which

accounted for 70% of all encounters. With this list the ICD9 codes with the longest and shortest LoS were determined. These lists can be used as guidance to the hospital operations team for knowing which IDC9 codes are associated with extreme LoS. In addition these tables were used to help find patient clusters.

ED – Duration:

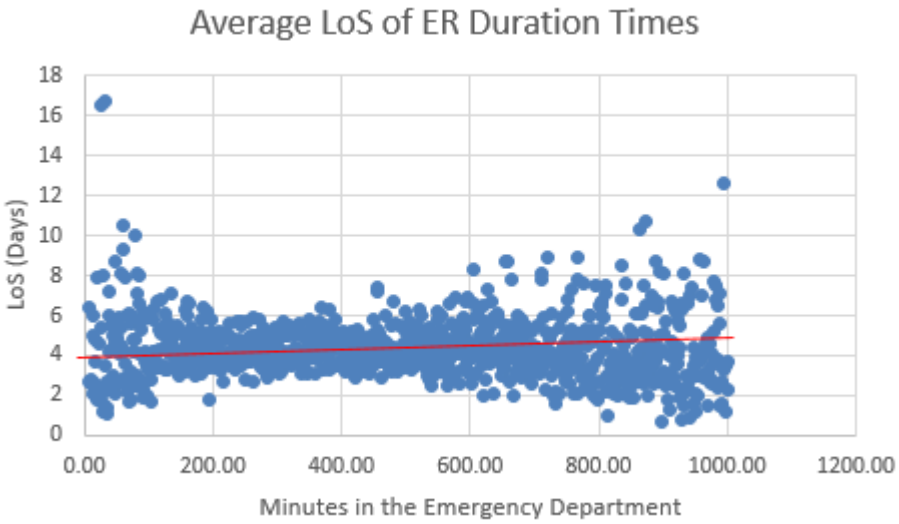

Figure 13. Average LoS of ER Duration Times

Figure 13 displays the average LoS relative to the amount of minutes a patient stayed in the emergency department. From this initial analysis it appears that the amount of time a patient stays in the emergency room has little effect on a patients LoS. This finding contradicts common sense so further analysis should be conducted to see if ED duration truly does not affect LoS.

ER Flag:

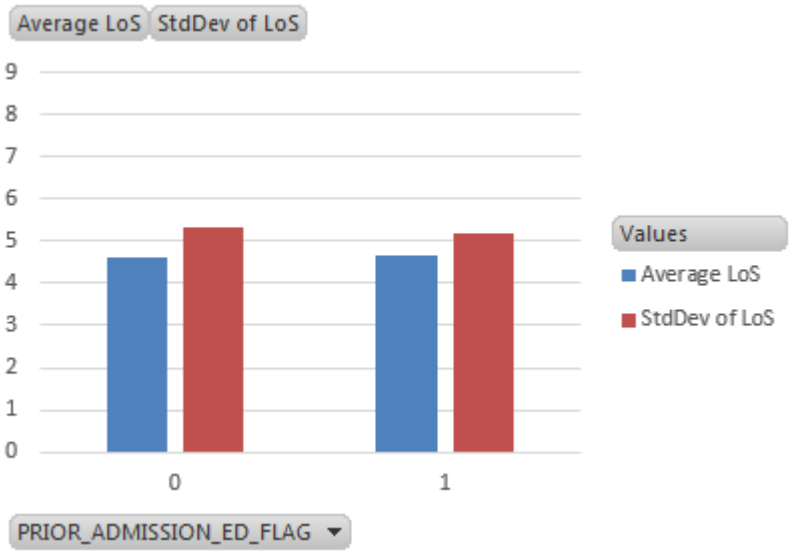

Figure 14. Emergency Department Flag Metrics

Similar to the ER – Duration, whether a patient had an emergency flag or not seemed to have little effect on LoS. In Figure 14 the "0" represents patients that did not come from the emergency room and "1" represents patients that did come from the emergency room.

Line of Service:

Rank	Service Lines	Avg LoS	StdDev	Count
1	Cardiac Surgery (GMCCSG)	11.593	8.450	51
2	Cardiovascular Surgery (GMCCVS)	9.353	4.392	39
3	Critical Care Medicine (GMCCCM)	8.338	10.694	261
4	Hematology (GMCHMA)	5.902	6.061	214
5	Neurosurgery (GMCNES)	5.764	5.543	188
6	Psychiatry (GMCPSY)	5.624	8.914	929
7	General Internal Medicine (GMCGIM)	5.559	5.042	6153
8	Oncology/Gynecology (GMCGYO)	5.346	3.995	46
9	General Surgery (GMCGLS)	5.143	4.669	473
10	Medicine GSACH (GSACHKIM)	4.911	3.249	203

Figure 15. Top Ten Longest LoS of Service Lines

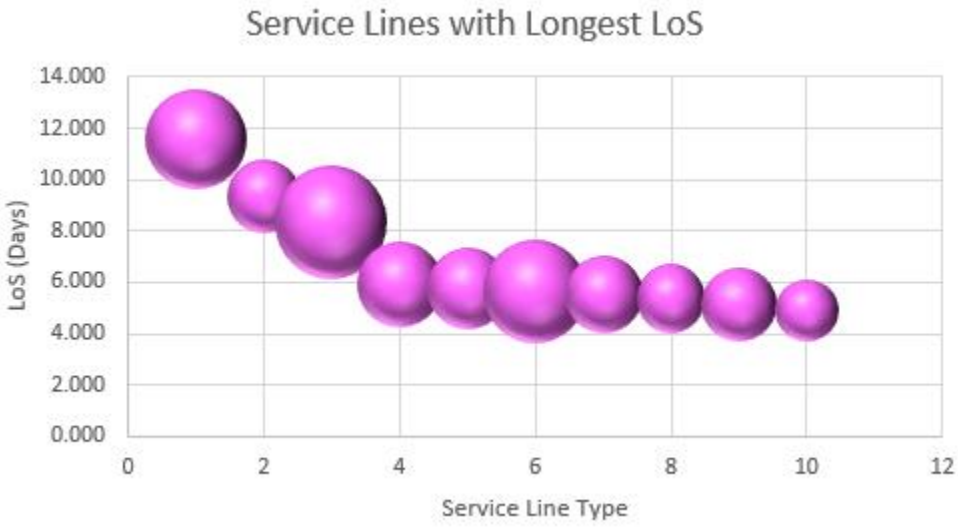

Figure 16. Top Ten Longest LoS of Service Lines Bubble Graph

The service line analysis was similar to the analysis done on the ICD9 codes. A service line is the part of the hospital that is responsible for the patient. Of the 90 service lines the top 50 most common accounted for 97% of all encounters. Figures 15 and 17 show the service lines with the top ten longest and shortest average LoS and can be used by the hospital operations team to recognize service lines with extreme LoS averages. The bubble graphs display the average LoS for the respective top ten services lines. The size of the bubble represents the size of the standard deviation.

Rank	Insurance Company	Avg LoS	StdDev	Count
1	TRICARE-HEALTH NET FEDERAL	3.176	3.033	58
2	WORKERS COMP	3.307	3.454	105
3	TPA - GEISINGER HEALTH PLAN	3.351	3.218	200
4	GHP HMO COMMERCIAL	3.417	4.539	1194
5	AMISH & MENNONITE	3.445	3.624	59
6	BLUE CROSS	3.727	4.312	268
7	OTHER MEDICAID HMOS	3.770	6.736	54
8	CIGNA	3.816	3.708	46
9	KEYSTONE HEALTH PLAN CENTRAL	3.860	4.246	75
10	HIGHMARK	3.878	4.403	1190

Figure 17. Top Ten Shortest LoS of Service Lines

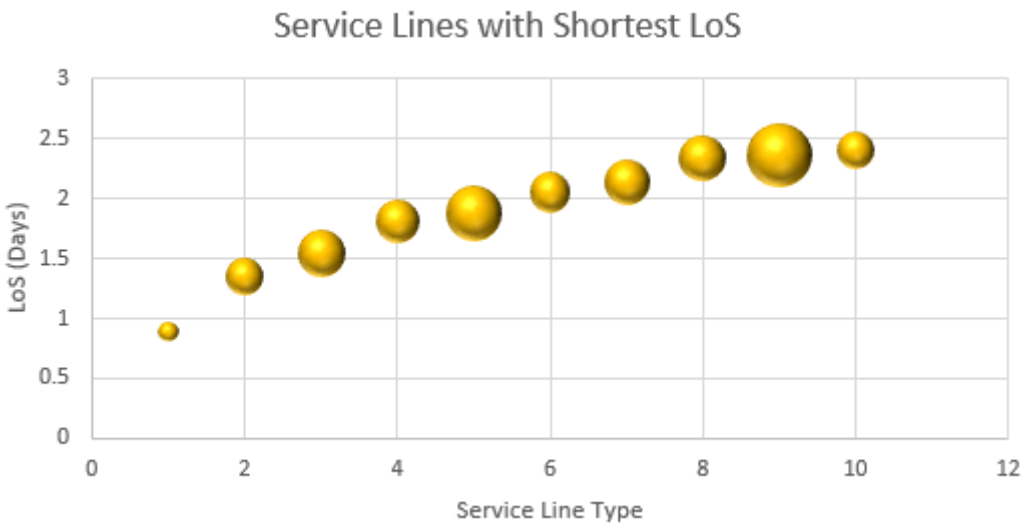

Figure 18. Top Ten Shortest LoS of Service Lines Bubble Graph

Chief Complaint:

Rank	Chief Complaint	Avg LoS	StdDev	Count
1	PSYCHOLOGICAL EVALUATION	6.376	11.073	382
2	HYPOTENSION	6.296	5.806	57
3	RESPIRATORY DISTRESS	6.194	8.157	117
4	CARDIAC ARREST	6.148	6.742	62
5	EDEMA	6.041	8.460	88
6	ALTERED MENTAL STATUS	5.815	6.033	736
7	GI BLEED	5.638	5.954	109
8	COUGH	5.546	6.734	113
9	SHORT OF BREATH	5.441	4.816	1598
10	BACK PAIN	5.364	4.049	154

Figure 19. Top Ten Longest LoS of Chief Complaints

Of the 332 chief complaints, 70 of them had a sample size larger than 30 which accounted for 89% of all encounters with chief complaints. Similar to the figures like this in previous analyses it can be used by the hospital operations team to identify chief complaints and to find clusters.

Rank	Chief Complaint	Avg LoS	StdDev	Count
1	CHEST PRESSURE	2.383	2.737	126
2	OVERDOSE	2.667	3.019	181
3	CHEST PAIN	2.825	3.164	1307
4	SEIZURE	3.167	3.566	154
5	SLURRED SPEECH	3.485	3.058	56
6	FLANK PAIN	3.509	3.370	87
7	ATV CRASH	3.610	5.301	49
8	EPIGASTRIC PAIN	3.657	3.683	58
9	DIZZINESS	3.787	4.922	157
10	SYNCOPE	3.930	4.288	168

Figure 20. Top Ten Shortest LoS of Chief Complaints

Surgery Type:

Of the 453 surgery types, 15 of them had a sample size larger than 30 which accounted for 5% of all encounters with a surgery. The figures below can be viewed in a similar matter as in previous analyses.

Rank	Surgery Type	Avg LoS	StdDev	Count
1	TRACHEOSTOMY PLANNED	22.435	14.017	46
2	EXPLORATORY LAPAROTOMY	11.827	8.316	179
3	CORONARY ARTERY BYPASS GRAFT USING ARTERY 1 GRAFT	10.162	4.705	37
4	CATHETER PLACEMENT, BRACHIOCEPHALIC, THIRD ORDER BRANCH	7.828	5.835	64
5	DEBRIDEMENT SKIN SUBCUTANEOUS TISSUE MUSCLE AND BONE	7.353	3.984	34

Figure 21. Top Five Longest LoS of Surgery Type

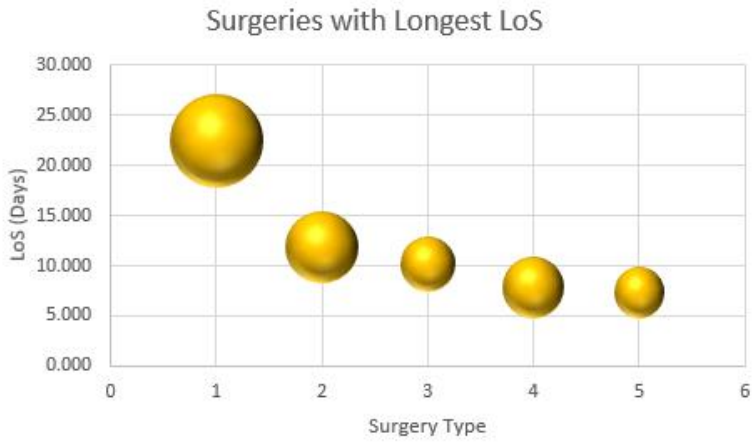

Figure 22. Top Five Longest LoS of Surgery Type Bubble Graph

Rank	Surgery Type	Avg LoS	StdDev	Count
1	LAPAROSCOPIC APPENDECTOMY	2.250	1.915	108
2	INCISION AND DRAINAGE FOREARM WRIST DEEP	2.257	1.704	35
3	LAPAROSCOPIC CHOLECYSTECTOMY WITH CHOL	2.387	1.283	31
4	LAPAROSCOPY DIAGNOSTIC	3.787	3.712	47
5	LAPAROSCOPIC CHOLECYSTECTOMY	3.820	3.477	167

Figure 23. Top Five Shortest LoS of Surgery Type

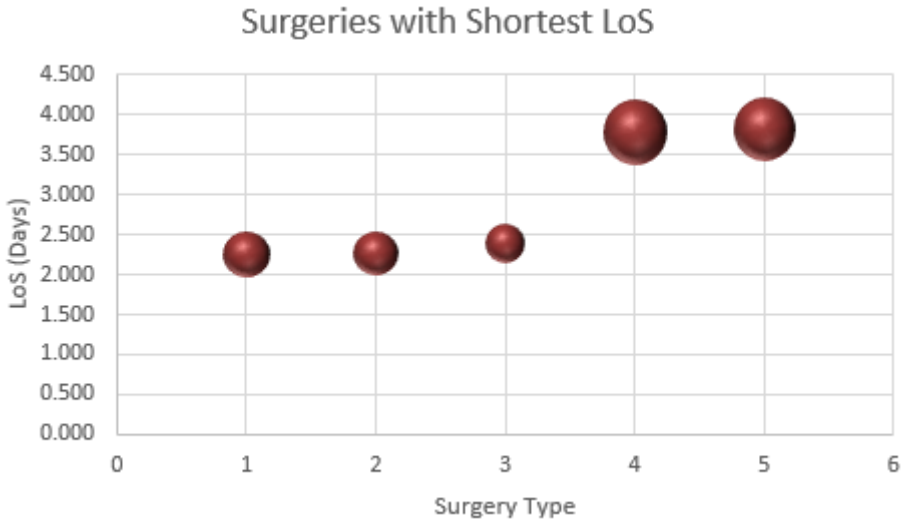

Figure 24. Top Five Shortest LoS of Surgery Type

Insurance Company:

Of the 33 insurance payers, 27 of them had a sample size larger than 30 which accounted for 98% of all encounters. The figures below can be viewed in a similar matter as in previous analyses.

Rank	Insurance Company	Avg LoS	StdDev	Count
1	FIRST PRIORITY HEALTH	5.75078405	7.418038	31
2	PSYCH HMO	5.550200321	10.55559	156
3	OTHER MEDICARE HMOS	5.462768251	4.77731	554
4	NO FAULT AUTO	5.306115363	7.159171	294
5	MEDICARE-FEE FOR SERVICE	5.219532407	5.471962	5100
6	MEDICAL ASSISTANCE	4.96067811	8.593955	1437
7	GHP GOLD	4.728774189	4.143136	2670
8	GOLD CHOICE	4.708818134	3.739686	53
9	OTHER FFS	4.685610178	6.340128	506
10	HEALTH AMERICA	4.419224684	5.237331	158

Figure 25. Top Ten Longest LoS of Insurance Companies

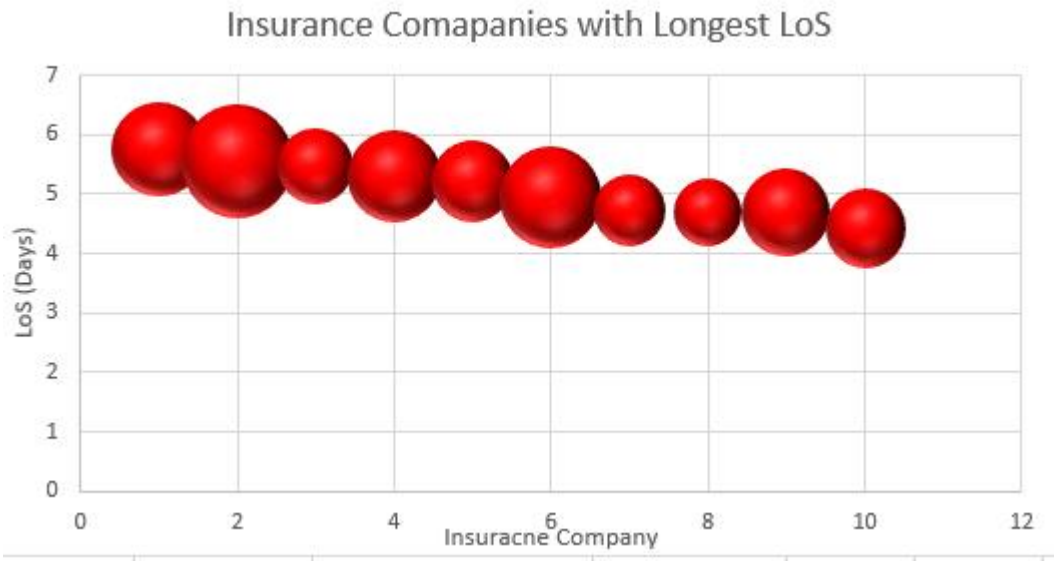

Figure 26. Top Ten Longest LoS of Insurance Companies Bubble Chart

Rank	Insurance Company	Avg LoS	StdDev	Count
1	TRICARE-HEALTH NET FEDERAL	3.176	3.033	58
2	WORKERS COMP	3.307	3.454	105
3	TPA - GEISINGER HEALTH PLAN	3.351	3.218	200
4	GHP HMO COMMERCIAL	3.417	4.539	1194
5	AMISH & MENNONITE	3.445	3.624	59
6	BLUE CROSS	3.727	4.312	268
7	OTHER MEDICAID HMOS	3.770	6.736	54
8	CIGNA	3.816	3.708	46
9	KEYSTONE HEALTH PLAN CENTRAL	3.860	4.246	75
10	HIGHMARK	3.878	4.403	1190

Figure 27. Top Ten Shortest LoS of Insurance Companies

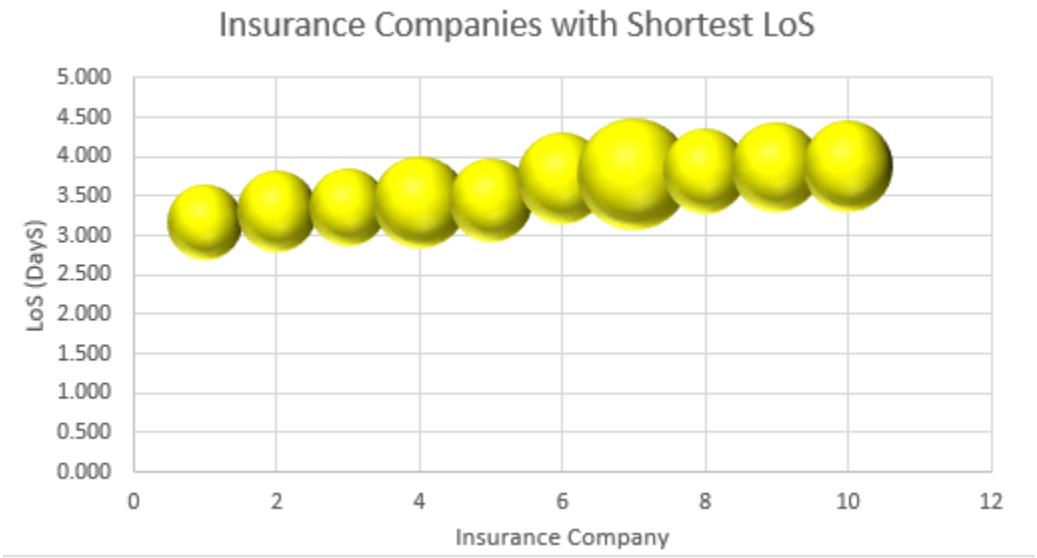

Figure 28. Top Ten Shortest LoS of Insurance Companies Bubble Chart

Clusters:

With the initial analysis complete the next step was to find clusters of patients with extreme LoS metrics. In this analysis a cluster is defined as a group of at least 30 patients that share at least 2 characteristics. These clusters were found using filters and pivot tables in Excel. These clusters do not include age and BMI as characteristics because of their continuous nature. Instead these fields can be used as an addition resource for prediction. For example, if a new patient is both old and fits the characteristics

of cluster 1 from Table 2 (ICD9: 995.91, Acuity Level: L1, Gender: Male) this increases the chances that patient will have a long LoS even more.

Table 2. Clusters with Long LoS

Clusters with Long LOS					
No.	Characteristic 1	Characteristic 2	Characteristic 3	Avg LoS	Count
1	ICD9: 995.91	Acuity: L1	Gender: Male	7.52	52
2	Chief Compliant: Altered Mental Status	ICD9: 780.97	Acuity Level: L2	6.03	118
3	Chief Compliant: Short of Breath	ICD9: 518.81	Gender: Male	9.86	33
4	Surgery: Exploratory Laparotomy	Chief Compliant: Abdominal Pain	ICD9: 560.9	9.12	32
5	Chief Compliant: Psychological Evaluation	Insurance: Medicare Fee for Service	None	10.98	82

Table 3. Clusters with Short LoS

Clusters with Short LOS					
No.	Characteristic 1	Characteristic 2	Characteristic 3	Avg LoS	Count
1	Chief Compliant: Chest Pain	Acuity: L3	Gender: Female	2.73	171
2	Surgery: Laparoscope Appendectomy	ICD9: 540.9	Acuity: L3	1.74	30
3	County: Northumberland	ICD9: 435.9	Gender: Female	2.78	33
4	Chief Compliant: Abdominal Pain	Service Line: Ped Surg	Acuity: L2 Gender: Male	1.72	46
5	Chief Compliant: Seizure	Acuity: L3	Male	2.68	39

Chapter 5

Conclusions and Recommendations

The primary goal of this thesis was to discuss the use of data analytics in the world and more specifically in healthcare and then to conduct an analysis of a real life data set from a Pennsylvania hospital. The goal of the data analysis was to examine how each field in the data set effects the LoS of a patient and to use these trends to find ten clusters with extreme LoS tendencies.

The initial analysis lead to many key findings about how each field correlates with LoS. Age positively correlated with LoS. As BMI increased average LoS increased and became more sporadic. Males tended to have a slightly higher chance of staying for more than two days. The acuity level or severity of the diagnosis was positively correlated with LoS. How long a patient was in the emergency department and whether or not the patient had been in the emergency department before showed no signs of correlation with LoS. Lastly, for the fields ICD9, Service Line, Chief Compliant, Surgery Type and Insurance Company lists of the characteristics with the longest and shortest LoS were created to help guide the discovery of clusters and to be used by the hospital operations team to see which characteristics have extreme tendencies when making bed capacity and nurse staffing decisions.

The initial analysis served as a guide to finding the clusters with Excels filters and pivot tables. Awareness of these clusters extreme LoS tendencies will help the hospital better predict the LoS for future patients. Under the assumption that the data set is an accurate portrayal of the type of patients that come to this hospital from year to year, the hospital would have a strong grasp on how long patients within these clusters will be staying at the hospital. Ideally, a mathematical model would be created for the hospital that would be able to generate an exact LoS prediction based on all the patient characteristics

upon diagnosis, but that is beyond the scope of this thesis. In further studies similar to this thesis one might generate more clusters to create a denser database of clusters with extreme LoS tendencies.

Data analytics is becoming one of today's most important tools for discovery. As more and more organizations are adopting the "data-driven" decision making mind set data analytic skills will become more valuable and a driving force behind our society's advancement into the future. This thesis as served as an example of what can be discovered through data analysis and the decisions that can be made as a result.

Appendix

Data Set and R Code

Headline of Hospital Data Set:

	A	B	C	D	E	F	G	H	I	J
1	ENCOUNT	PAT_ID	ADMISSIO	ADMIT_YE	ADMIT_M	ADMIT_W	ADMIT_TI	ADMIT_HI	ADMIT_M	DISCHARG
2	26524	5567	EMERGEN	2013	JUNE	SATURDA	15:09:00	15	9	2013
3	28049	6409	EMERGEN	2013	JUNE	WEDNESD	16:12:00	16	12	2013
4	12939	4003	EMERGEN	2013	NOVEMBE	SATURDA	21:12:00	21	12	2013
5	12716	5889	EMERGEN	2013	NOVEMBE	TUESDAY	18:54:00	18	54	2013
6	22067	7692	EMERGEN	2013	AUGUST	THURSDA	15:46:00	15	46	2013

	K	L	M	N	O	P	Q	R	S	T	U
	DISCHARG	DISCHARG	DISCH_TI	DISCH_HR	DISCH_MI	ETHNIC_G	DEATH_IN	GENDER	COUNTY	WEIGHT	HEIGHT
	JUNE	SATURDA	15:45:00	15	45	Not Hispa	Deceased	Female	Lycoming	104	5' 5"
	JUNE	WEDNESD	17:44:00	17	44	Not Hispa	Alive	Male			
	NOVEMBE	SATURDA	23:19:00	23	19	Not Hispa	Deceased	Male	Northumberland		
	NOVEMBE	TUESDAY	21:07:00	21	7	Not Hispa	Alive	Male	Union		

	V	W	X	Y	Z	AA	AB	AC	AD	AE
	BMI	LOS_IN_D	PRECISE_LOS	los_final	ED_DURA	ED_TRIAG	ED_TRIAG	ACUITY_LE	PAT_AGE	AGE
	17.30633	1	0 00:36	0.025	46	15:09:51	15:12:22	L1	60	60
	0	1	0 01:32	0.063888889	91			L2	45	45
	0	1	0 02:07	0.088194444	27				68	68
	0	1	0 02:13	0.092361111	145	18:43:19	18:45:40	L2	21	21
	0	1	0 02:36	0.108333333	160	15:46:22		L1	2	2
	0	1	0 02:37	0.109027778	175			L2	34	34
	21.92687	1	0 02:38	0.109722222	167	22:40:02	22:46:47	L2	33	33

	AF	AG	AH	AI	AJ	AK	AL	AM	AN
	PAT_CLAS	PAT_SVC	PRIMARY_	ADMITTING_D	BILLING_DIAGNOSIS	DX_GROU	ADMIT_CE	PRE_6_CE	PRE_12_C
	Inpatient	Cardiology (GMCCVM)	427.5	Cardiac arrest	AMI NOS-INITIAL EPI	CIRCULAT	364	411	395
	Observati	Trauma Surgery (GMCT	994.8	Electrocution,	EFFECTS ELECTRIC CU	INJURY AN	405	468	377
	Inpatient	Critical Care Medicine	427.5	Cardiac arrest	VENTRICULAR FIBRILL	CIRCULAT	376	380	427
	Observati	Trauma Surgery (GMCT	850.9	Concussion	OPEN WOUND OF HE	INJURY AN	402	486	384
	Observati	Trauma Surgery (GWV	959.9	Injury	SKULLL VLT FX W/O C	INJURY AN	497	548	450

Figure 29. Data Set Part 1

AO	AP	AQ	AR	AS	AT	AU	AV
PRE_18_CI	ENC_REASON_NAME_1	ENC_REAS	ENC_REAS	ENC_REAS	ENC_REAS	ENC_REAS	ENC_REAS E
392	CARDIAC ARREST	TRANSFER OF RECORDS					
386							
423							
373	ASSAULT						
450	FALL						

AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE
ENC_REAS	ENC_REAS	ENC_REAS	ENC_REAS	MS_DRG	BILLING_PAYC	PRIOR_EN	PRIOR_AD	PRIOR_ADMISSION_ED_FLAG	NUM_LAB
				238	GHP HMO COI	70881	1124	1	29
					GHP FAMILY	60444	659	1	9
				310	MEDICARE-FEI	47638	424		12
					SELF	82036	1891	1	6
					MEDICAL ASSI	57323	512		11
						22536	80	1	15

BF	BG	BH	BI	BJ
NUM_MRI	NUM_CTS	NUM_MEI	SURGERY_	RECNO
				33927
	1			39887
		1		22848
	1			36366
	1	2		45892

Figure 30. Data Set Part2

R Code for Figures 3 and 4:

```
#LOS of inpatients#

#total#

mydata1<-read.csv('d:/temp/LOS_inpatient.csv', header=T, sep=",")

hist(mydata1$los_day, freq=F, xlab="day",ylab="prob.",xlim=c(0,210),
```

```
 breaks=seq(0,209,by=1), main="LOS of inpatients in days", axes=F)
axis(side=1, at=seq(0,210,10))
axis(side=2,at=seq(0,0.20,0.02))
#only 7 days#
mydata1<-read.csv('d:/temp/LOS_inpatient.csv', header=T, sep=",")
hist(mydata1$los_day, freq=F, xlab="day",ylab="prob.",xlim=c(0,7),
 breaks=seq(0,209,by=1), main="LOS of inpatients in days", axes=F)
axis(side=1, at=seq(0,7,1))
axis(side=2,at=seq(0,0.20,0.02))
################################
#LOS of observation patients#
mydata2<-read.csv('d:/temp/LOS_observation.csv', header=T, sep=",")
hist(mydata2$los_day, freq=F, xlab="day",ylab="prob.",xlim=c(0,19),
 breaks=seq(0,19,by=1), main="LOS of observation patients in days",
 axes=F)
axis(side=1, at=seq(0,19,1))
axis(side=2,at=seq(0,0.44,0.02))
```

BIBLIOGRAPHY

"100 Most Creative People in Business." Fast Company June 2014: 103. Web.

Abraham. "Funny Graphs Show Correlation between Completely Unrelated Stats [9 Pictures]." 22 Words: Funny Graphs Show Correlation between Completely Unrelated Stats 9 Pictures. N.p., 9 May 2014. Web. 21 Nov. 2014.

"Bringing Big Data to the Enterprise." IBM. N.p., n.d. Web. 20 Nov. 2014. <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>.

Gellman, Lindsay. "Big Data Gets Master Treatment at B-Schools." The Wall Street Journal. Dow Jones & Company, 5 Nov. 2014. Web. 20 Nov. 2014.

Herrle, Greg. "Reducing Inpatient Length of Stay: The Time Has Come To Revisit This Discarded Strategy." Consultant's Corner (n.d.): n. pag. Publications.milliman.com. Web.

Liebowitz, Jay. Big Data and Business Analytics. Boca Raton: Taylor Francis Group, 2013. Print.

Lohr, Steve. "The Age of Big Data." The New York Times. The New York Times, 11 Feb. 2012. Web. 20 Nov. 2014.

Parker, Ashley, and Nick Corasaniti. "Data-Driven Campaigns Zero In on Voters, but Messages Are Lacking." The New York Times. The New York Times, 30 Oct. 2014. Web. 20 Nov. 2014.

Scola, Nancy. "Obama, the 'big Data' President." Washington Post. The Washington Post, 12 June 2013. Web. 20 Nov. 2014.

Shaw, Johnathan. "Why “Big Data” Is a Big Deal." Harvard Magazine Apr.-May 2014: n. pag. Web.

Xerox. "Streamlining Big Data to Help Hospitals Save Lives." Washington Post. The Washington Post, 03 Nov. 2014. Web. 20 Nov. 2014.

ACADEMIC VITA

Matthew Digel
Med5342@psu.edu

Education:
The Pennsylvania State University - University Park, PA
Bachelor of Science in Industrial and Manufacturing Engineering
Class of December 2014

Honors and Awards:
Industrial Engineering Department Scholarship 2012, 2013, 2014
Archie Griffin Sportsman of the Year Award (2010)

Association Memberships:
Sigma Chi Fraternity
Presidential Leadership Academy

Professional Experience:

Assembly Line Intern - General Electric Transportation (Summer 2012)

Application Sales Engineering Intern – General Electric Energy (Summer 2013)

Management Consulting Intern – Pricewaterhouse Coopers (Summer 2014)