**TABLE OF CONTENTS**

[Survey of healthcare analytics area](#h.8xiyatjl7aov)

[Healthcare delivery problem](#h.qhxmfkcnh6n4)

[Data mining problem](#h.vxhb87pc41n1)

[Exploration of data](#h.vnk5x0dkcjbs)

[References](#h.na3svjayerss)

# Survey of healthcare analytics area

The healthcare industry has modernized its operations, and is increasingly adopting electronic health records (EHRs). This has lead to the deployment of new health information technology systems that constantly create, collect, and manage their information. Therefore, the amount of data available to clinicians and hospital administrators in the healthcare domain is growing at an exponential rate. However, despite such advances in technology and proven success in making data driven decisions in other domains such as retail,marketing etc., healthcare providers often report only minor improvements in decision making capability through the use of existing data.[1]

Costs and risk are not spread evenly across a population in many healthcare systems. Therefore, relatively small number of patients who are classified as high-risk patients tend to consume or utilize more medical resources than their peers. Also, studies are showing that deficits in managing care for these patients could lead to higher expenses. These findings necessitate and warrant systematic efforts that focus on identifying high risk patients to ensure that they receive the most efficient and effective care possible. [2]

The digitization of healthcare domain stand to realize significant benefits. Some of the potential benefits include detecting diseases at earlier stages when they can be treated more effectively; detecting health care fraud by individuals claiming insurance more quickly and efficiently. Many outcomes could be predicted and/or estimated by making use of the vast amounts of historical data, such as length of stay (LOS); patients who will choose surgery; patients who are not likely to benefit from surgery; patients at risk for medical complications, etc. McKinsey has recently stated that big data analytics has the potential to enable savings of more than $300 billion per year in U.S. healthcare. Also, McKinsey believes big data could help minimize inefficiencies in the following three areas[3]:

1. *Clinical operations*
2. *Research & development*
3. *Public health*

# Healthcare delivery problem

Our world has a population of more than 7 billion people today. This growing trend of global population has triggered some of the biggest healthcare challenges that the healthcare industry is facing. One of the major business problems in healthcare domain is the assessment and management of clinical problems in the hospital. Governments and patients evaluate a hospital's quality of care by looking at performance data. In many countries, the data used to compare and evaluate outcomes is frequently based on Diagnosis Related Groups (DRGs).[5] This involves classifying a patient’s severity of illness or his mortality rate, or reducing the length of stay (LoS) of patients in the hospital based on the initial diagnosis.

For example, by classifying the patients based on the severity of the illness, the doctors can plan a treatment schedule well ahead of time. The hospital can also plan the logistics for the proposed treatment which can help in the efficient treatment of the patient. Similarly, by predicting the length of stay of a patient, hospital can solve the problem of manpower allocation. This can help the hospital in effective scheduling for admission of elective patients.[3]

# Data mining problem

## Problem 1: Classify severity of illness

Severity of illness is defined as the extent of physiologic decompensation or organ system loss of function. The four severity of illness subclasses are numbered sequentially from 1 to 4 indicating respectively, minor, moderate, major, and extreme severity of illness. Based on our preliminary research we found that the variables listed below impact the severity of illness:[4]

1. Principal Diagnosis coded in ICD-9-CM (PRINC\_DIAG\_CODE)
2. Secondary Diagnoses coded in ICD-9-CM (OTH\_DIAG\_CODE)
3. Procedures Coded in ICD-9-CM (PRINC\_ICD9\_CODE,OTH\_ICD9\_CODE)
4. Age (PAT\_AGE)
5. Sex (SEX\_CODE)
6. Discharge Disposition (PAT\_STATUS)

## Problem 2: Classify risk of mortality

Risk of mortality is defined as the likelihood of dying. The four risk of mortality subclasses are numbered sequentially from 1 to 4 indicating respectively, minor, moderate, major, and extreme severity of illness. Based on our preliminary research we found that the variables listed below impact the risk of mortality:[4]

1. Principal Diagnosis coded in ICD-9-CM (PRINC\_DIAG\_CODE)
2. Secondary Diagnoses coded in ICD-9-CM (OTH\_DIAG\_CODE)
3. Procedures Coded in ICD-9-CM (PRINC\_ICD9\_CODE,OTH\_ICD9\_CODE)
4. Age (PAT\_AGE)
5. Sex (SEX\_CODE)
6. Discharge Disposition (PAT\_STATUS)

## Problem 3: Predict patient’s Length of Stay (LoS)

Length of stay (LoS) is a term to describe the duration of a single episode of hospitalization. Inpatient days are calculated by subtracting day of admission from day of discharge.[7]

Based on our preliminary research, we found that the variables listed below impact the LoS:

1. Principal Diagnosis coded in ICD-9-CM (PRINC\_DIAG\_CODE)
2. Secondary Diagnoses coded in ICD-9-CM (OTH\_DIAG\_CODE)
3. Age (PAT\_AGE)

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

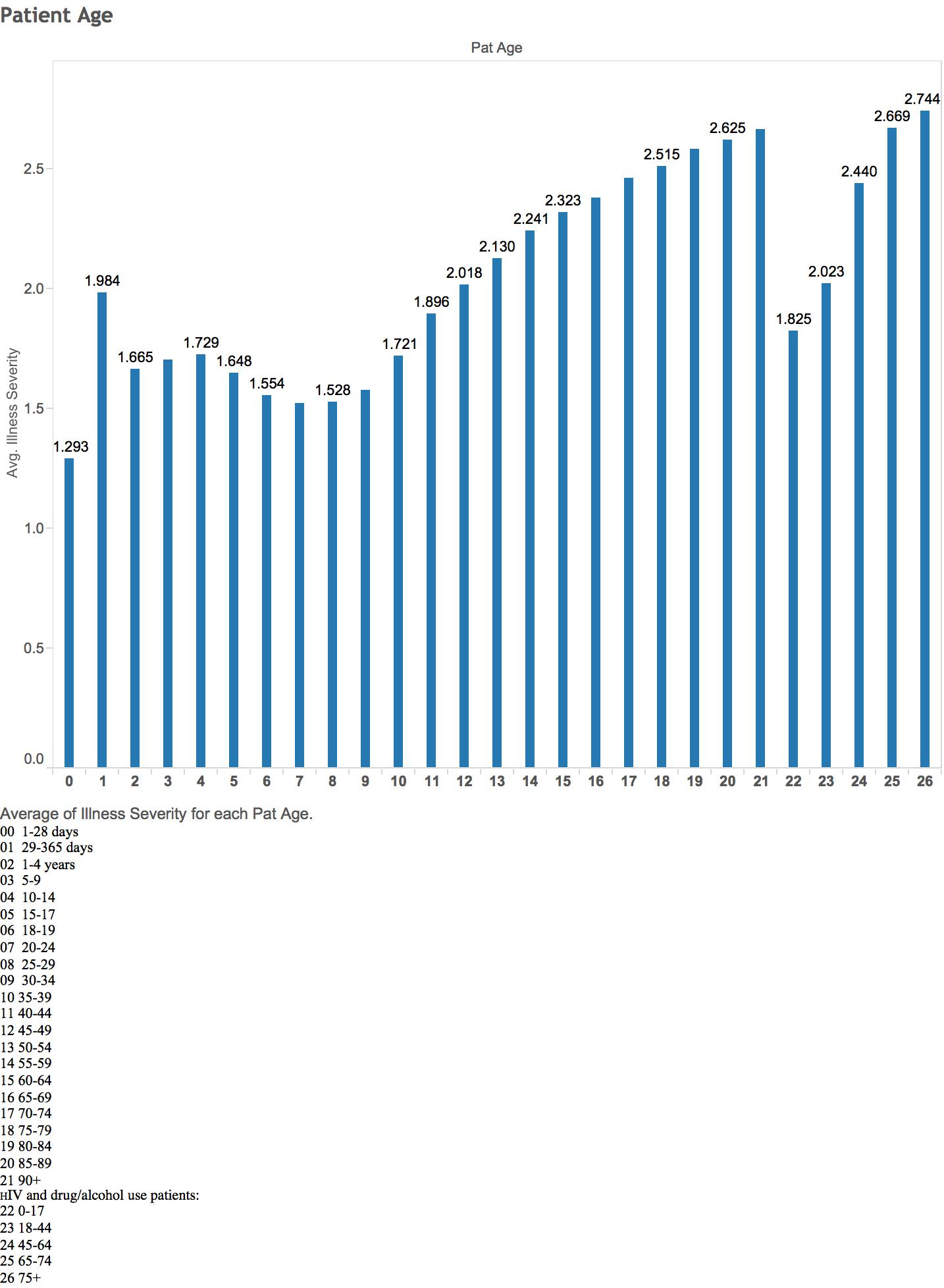
# 

# 

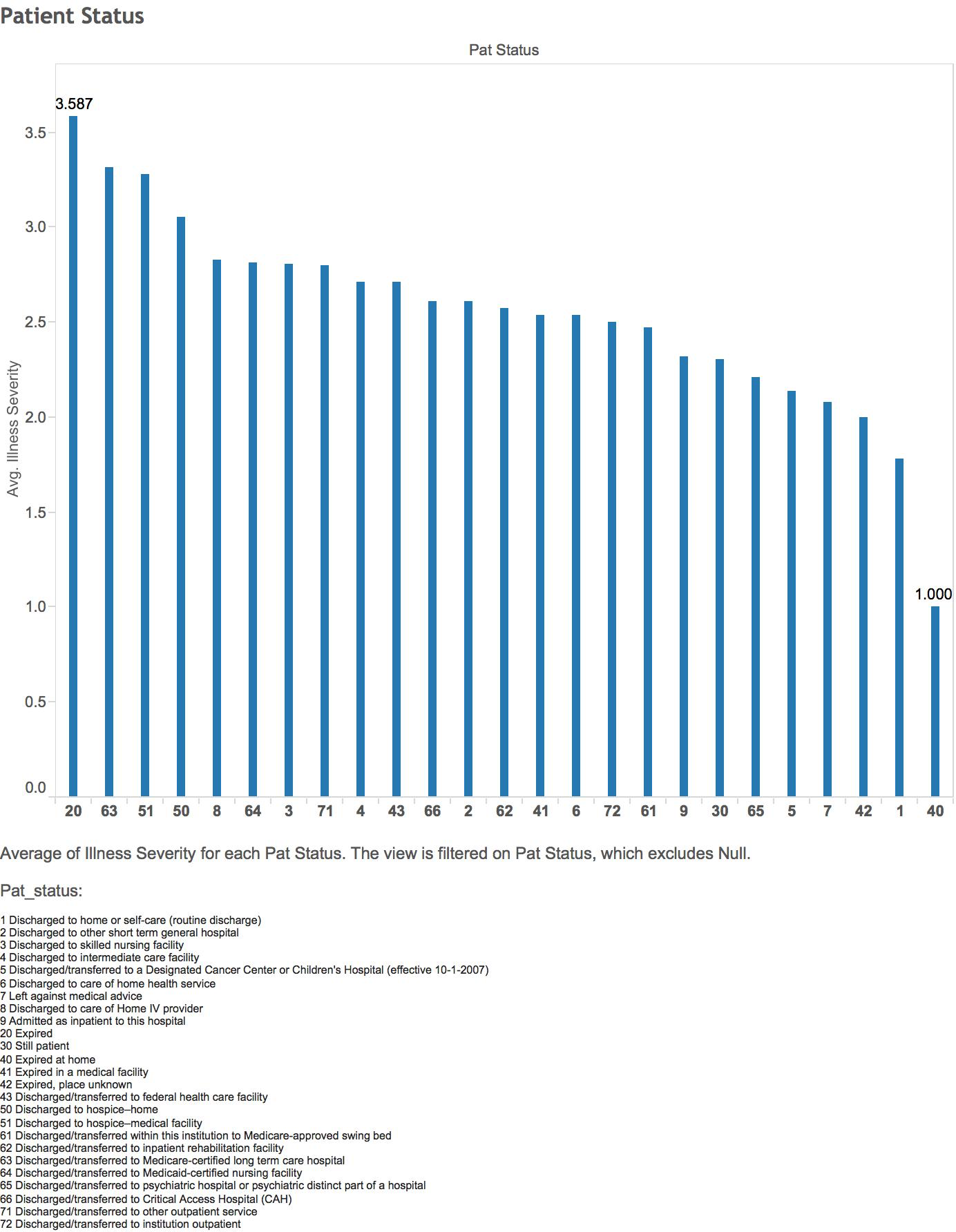
# 

# Exploration of data

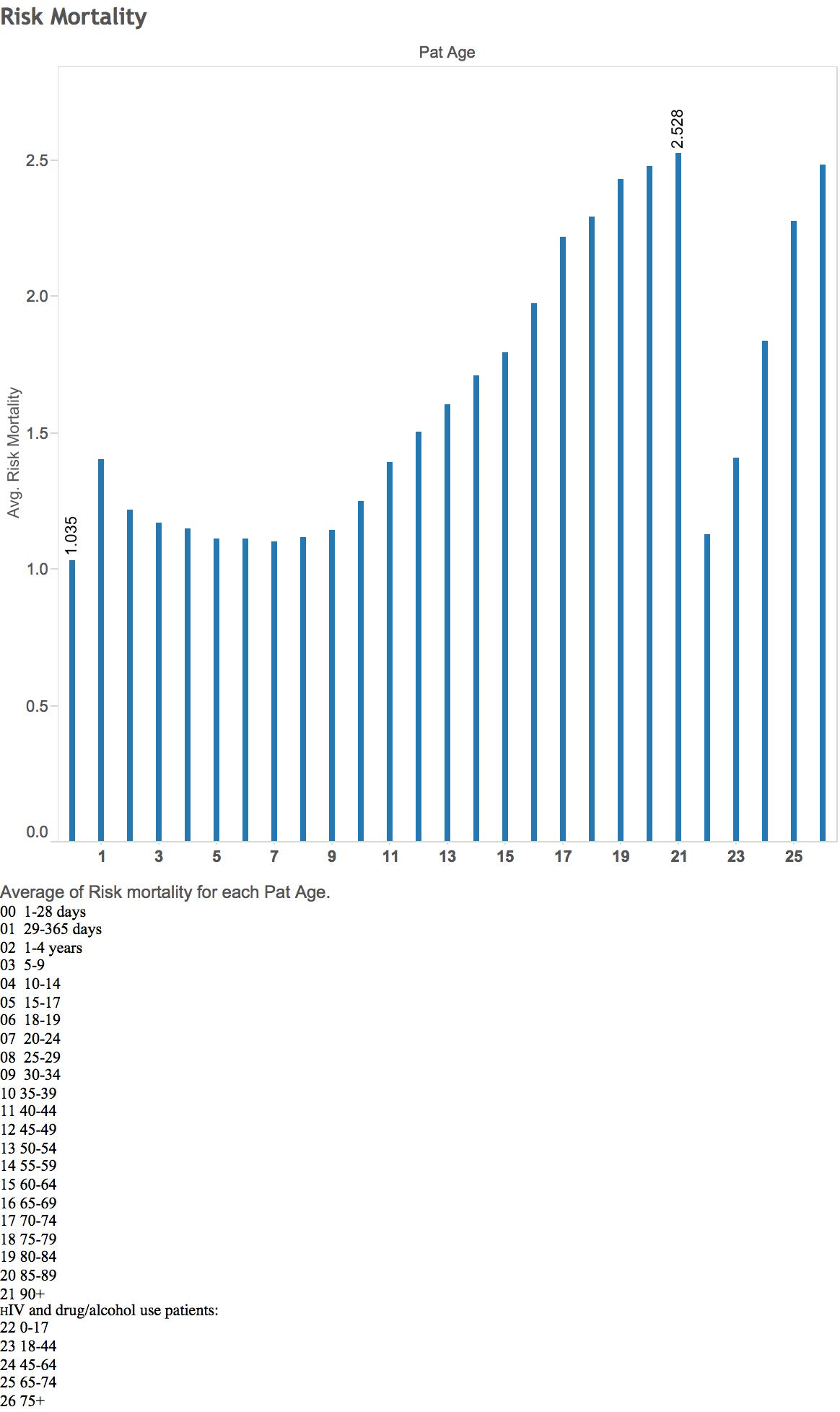
Based on our initial exploration of the datasets using JMP and tableau, we see that there is strong correlation between the focus variables and other variables within the dataset. The following visualizations show a relationship between variables but not a conclusive enough evidence to prove the same.

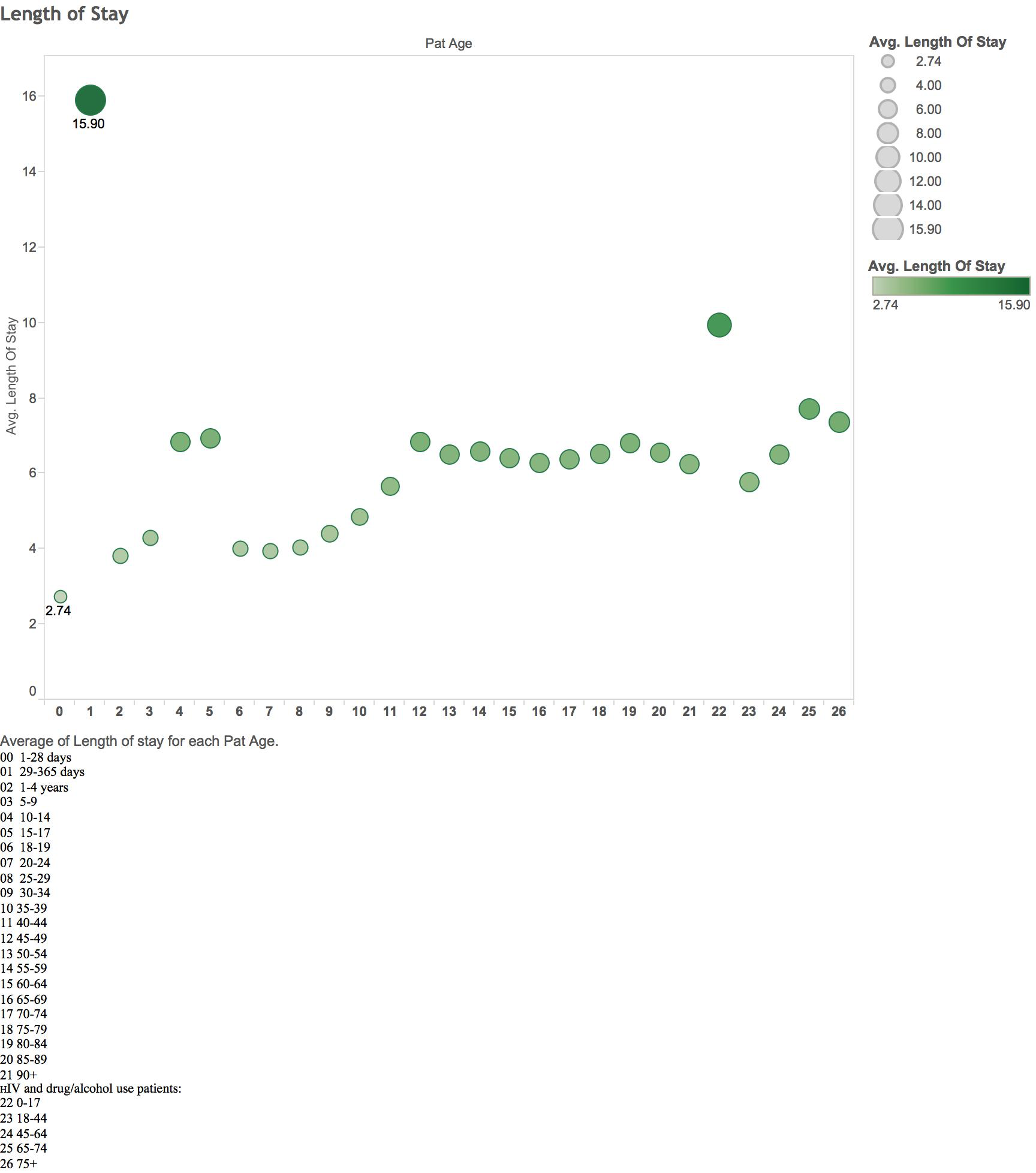


Inference: The severity of illness is measured on a scale of 1-4 where 1 is the lowest severity and 4 denotes highest severity of illness. The above graph shows that there is a strong correlation between illness severity and age of the patients. This denotes that the patient's age is considered to be a strong factor that influences the severity of illness of patients during diagnosis. Though the severity data is discrete in the original data, for analysis purposes we have considered the average illness severity across various age groups in consideration.



Inference: The severity of illness is measured on a scale of 1-4 where 1 is the lowest severity and 4 denotes highest severity of illness. The above graph shows that there is a strong correlation between illness severity and patient status. This denotes that the patient's status is considered to be a strong factor that denotes the severity of illness of patients during diagnosis. Though the severity data is discrete in the original data, for analysis purposes we have considered the average illness severity across various age groups in consideration.





# References

1. Jesus J Caban and David Gotz.Visual analytics in healthcare – opportunities and research challenges. *Journal of the American Medical Informatics Association.*
2. David Gotz, Harry Stavropoulos and Jimeng Sun. ICDA: A Platform for Intelligent Care Delivery Analytics. *US National Library of Medicine National Institutes of Health*
3. Gordon H. Robinson, Louis E. Davis, and Richard P. Leifer. Prediction of Hospital Length of Stay. *Health Services Research*
4. 3M Health Information Systems. APRTM-DRG Classification Software -Overview
5. P. Fontaine.Using Severity Adjustment Classification for Hospital Internal and External Benchmarking. *2004 IFHRO Congress & AHIMA Convention Proceedings, October 2004*
6. Length of Stay (LoS). Wikipedia. [*https://en.wikipedia.org/wiki/Length\_of\_stay*](https://en.wikipedia.org/wiki/Length_of_stay)
7. Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems.*