Project Team 4
Members:
Chang Zhou
Harsha Jakkappanavar
Savan Patel
Jenny Johns

1 November 2017

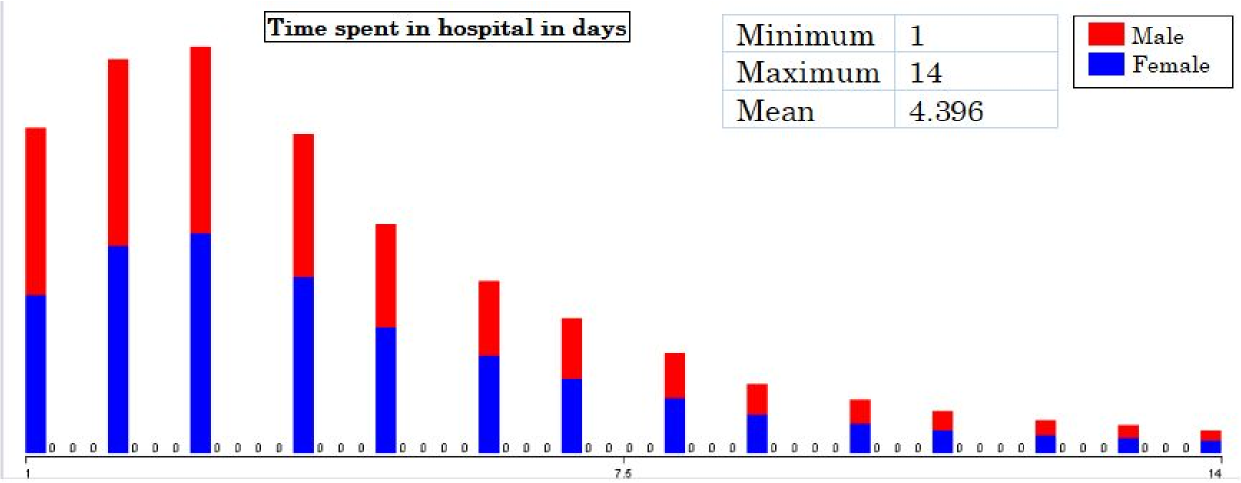
Update 1

Project Title:
Study of dataset on Diabetes from 130 US hospitals over 10 years

Exploratory Analysis:
During the exploratory analysis of this dataset, we found that patients were categorized in one of five possible races, including, African American, Asian, Caucasian, Hispanic, or Other. A majority of the dataset is seen to be of Caucasian race with 76,099 out of 101,767 patient records. The second largest group were African Americans, third Hispanics, fourth Others, and finally Asians. There were some missing values from the dataset.

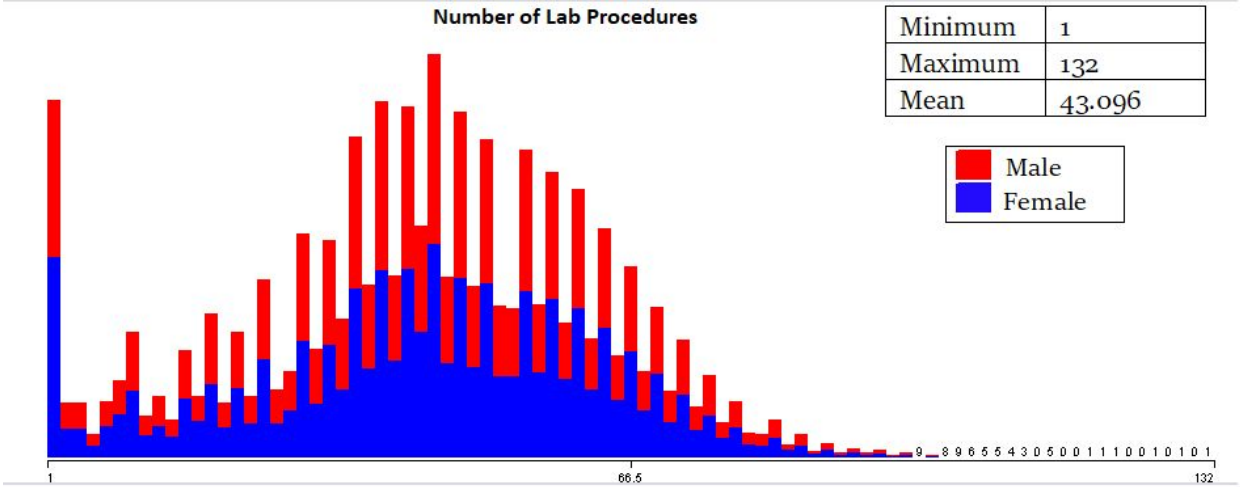
Gender was also broken up into three categories: Male, Female, and Unknown/Invalid. The majority of patients were female with 54,708 females versus 47,055 males. Of these patient records, there was an age range of 0 - 100, broken down into 10 year intervals. The largest subset of these patients came from the range of [70 - 80) with 26,068 patients.

All patients were brought into the hospital with a coded admission type, ranging from 1 to 8, each representing a unique admission type. The most common admission type, was code 1, which represented an emergency admission to the facility. The time patients spent in the hospital ranged from a minimum of 1 day to a maximum of 14 days. The average time spent in the hospital was around 4 days.

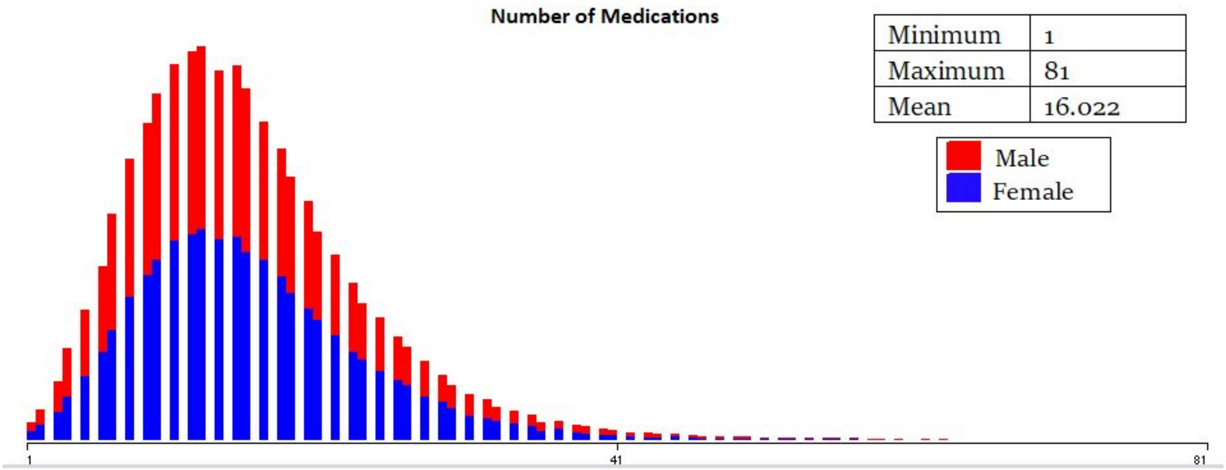


Minimum	1
Maximum	14
Mean	4.396

All patients had at least one lab procedure completed during their duration at the hospital. The maximum number of lab procedures was 132 and the average amount was around 43.

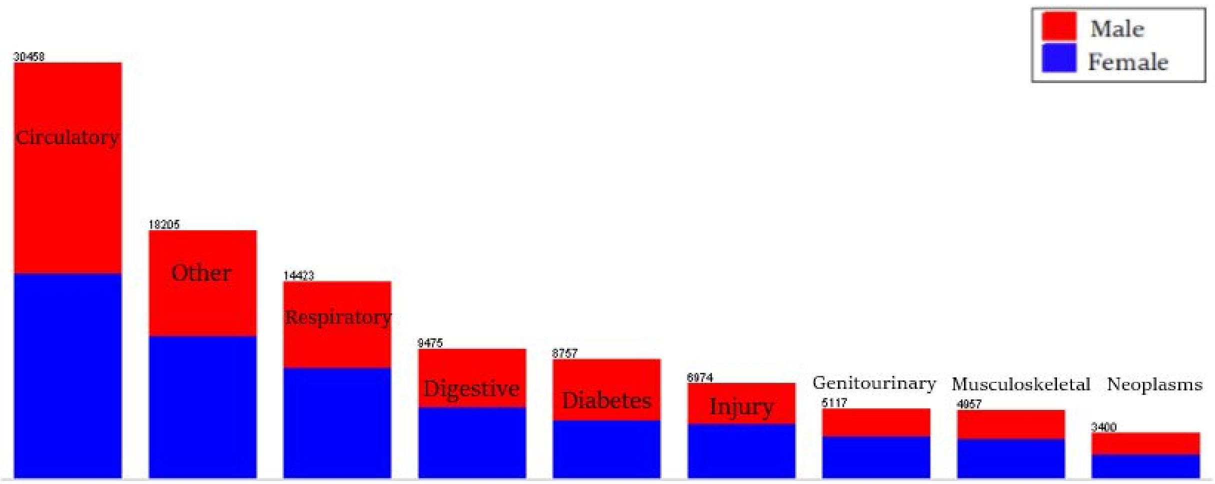


Also during their hospital stay, medications were administered to the patient, ranging from a minimum of one medication to a maximum of 81 medications, the average being around 16 medications per encounter.

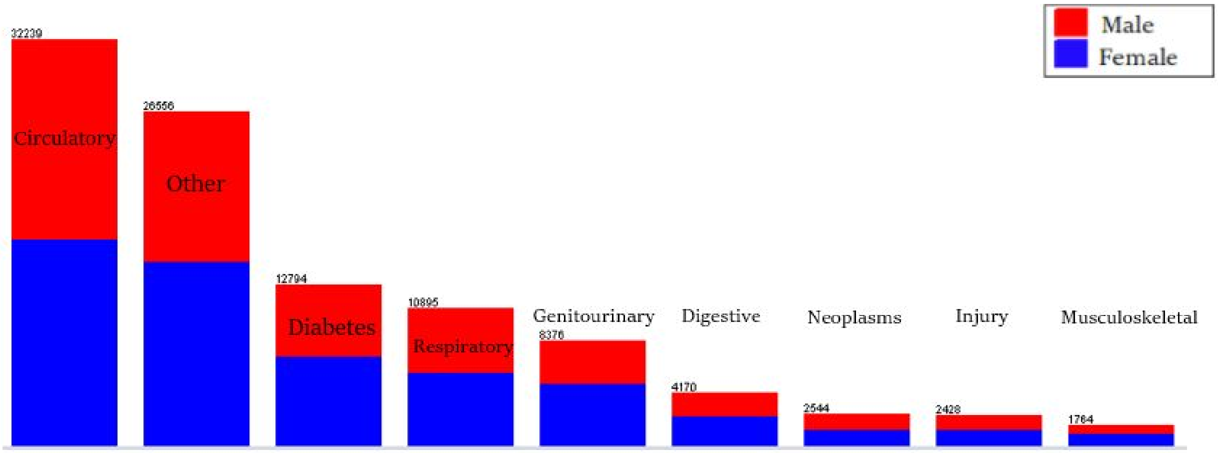


There were three different columns for diagnoses within this dataset. The first, ‘diag_1’ being the primary diagnosis, the second, ‘diag_2’ being the secondary diagnosis, and the third, ‘diag_3’ being an additional secondary diagnosis. Each diagnosis is a code from the icd9, representing a specific illness. Some things we noticed while working with the dataset was that a majority of the primary diagnoses were from ‘Heart Failure’ (icd9 code: 428) affecting 6,862 patients, from the second diagnosis, the majority of diagnoses were ‘Disorders of fluid, electrolyte, and acid-base balance’ (icd9 code: 276) affecting 6,752, and from the third diagnosis, the majority of diagnoses were ‘Diabetes mellitus’ (icd9 code: 250), affecting 11,555 patients. These diagnoses were a nominal value, and had many levels (each different icd9 code), so we grouped the codes based on which system of the body they affected, code 428, would be grouped as circulatory. Circulatory seemed to be the most highly affected system for all three diagnoses, followed by Other. The different categories were: Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Other.

Diagnosis 1



Diagnosis 2



Diagnosis 3

There are 24 different features for medications that can be administered during a patient's hospital stay. There are four possible values for the medications within the dataset, ‘up’, ‘down’, ‘steady’, or ‘no’. For all the medications, the value with the highest percentage was ‘no’ meaning that the medication was never prescribed. Our dataset also included whether there was a change in the medications administered to the patient. The possible values were ‘ch’ for change for ‘no’ for if there was no change. A majority of the patients didn’t have a change in their medications, with a count of 54,755.

There was a column within the dataset that specified whether any diabetic medication was prescribed during that patient’s stay in the hospital, possible values being ‘no’ and ‘yes’. With a count of 78,363 for ‘yes’, meaning that bulk of patients were prescribed some sort of diabetic medication during their encounter. Readmission was another feature for which a record was kept. Possible values were ‘<30’ for if the patient was readmitted less than 30 days after they were discharged, ‘>30’ for if the patient was readmitted more than 30 days after they were discharged, and ‘no’ for if there was no record of readmission for that patient.

Work Completed:

Dataset

The dataset was converted from CSV to ARFF for processing in weka.
The dataset contains many nominal features, and each feature has many levels. These nominal features have been broken down by their level. For example, the columns that contain the diagnoses have values of their icd9 codes, these have been broken down into which system of the body is affected in that diagnosis. If the code is 396, (diseases of mitral and aortic valves), it will be returned as ‘Circulatory’.

FP Growth

Within our dataset, there are 24 features of medications which all have values of either, ‘up’, ‘down, ‘steady’ or ‘no’. The value was transcribed as ‘up’ if the dosage of that particular medication was increased during that patient’s hospital stay, ‘down’ if it was decreased, ‘steady’ of the dosage wasn’t changed, and ‘no’ if it was never prescribed. We specifically only looked at the first 18 features of medications because the last six are combinations of those medication features. After converting this data to the ARFF format, it was run through weka’s FP Growth algorithm, which was able to produce two association rules:

1. [i1=1]: 19988 ==> [i18=1]: 10012 <conf:(0.5)> lift:(0.72) lev:(-0.05) conv:(0.61)
2. [i18=1]: 54383 ==> [i1=1]: 10012 <conf:(0.18)> lift:(0.72) lev:(-0.05) conv:(0.91)

The rules are showing that the medication ‘metformin’, (i1), and ‘insulin’, (i18), have an association between them. From the first rule, we can see that the confidence of ‘insulin’ occurring given ‘metformin’ is 0.5 and from rule two, we can see that the confidence of ‘metformin’ occurring given ‘insulin’ is 0.18. These confidences are not high, which means that there is not really a strong association between these medications. And we can see that the lift is 0.72, and since in lift a ‘1’ means they are independent, this shows that the two medications, metformin and insulin, are pretty close to occurring independently from each other.

Next Steps:

We would like to perform a dimensionality reduction on our data using principal component analysis (PCA). We would also like to experiment with box-cox normality plots to find the transformations that can normalize our data, as well as continue to apply other algorithms. Further data cleaning is needed using dummy coding for all nominal variables within the dataset. During our exploratory analysis, we realized that the dataset can be mostly used for supervised learning. Since our course is mainly exploring unsupervised learning algorithms, we are looking into adding a second dataset to our project and find good clustering mechanisms.