Calculating Regression using R

Angus Tsui, Grant Merkel, Chris Collins

Our Specified Task

We took the SPARCS dataset of 2.5 million people and broke it up into subsets that were easier to work with, such as county, age, gender, etc.

We found relationships between these subsets and were able to use R to easily visualize them.

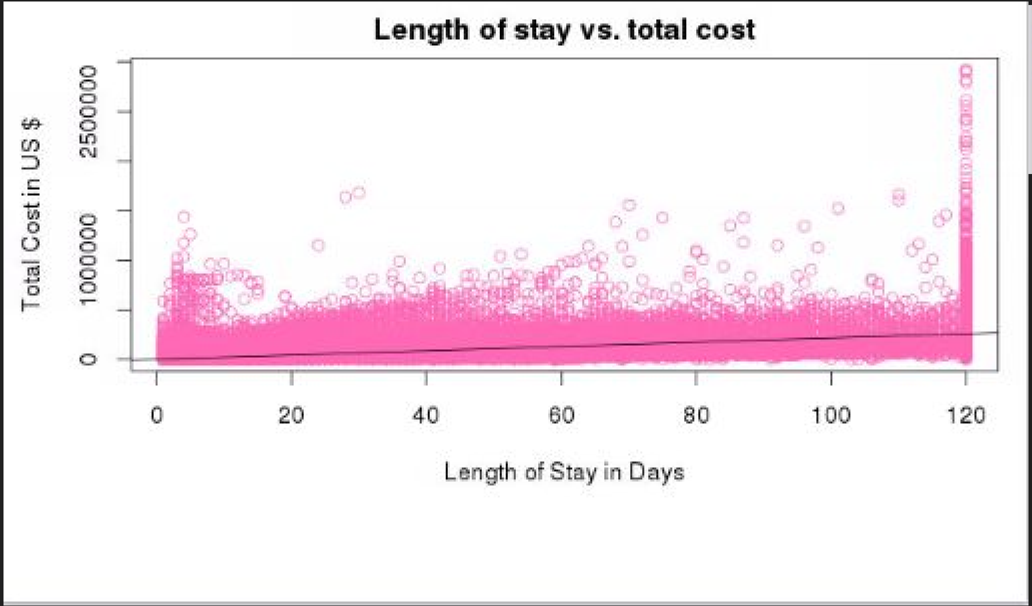
Basics of Linear Regression

What is Linear Regression? Linear Regression is using a line to fit to a set of data points on a graph to show a particular relationship between two variables.

What is correlation? Correlation is a way to model the linear relationship between two variables.

Correlation Coefficient is the measure of how well the linear regression model fits to the data points.

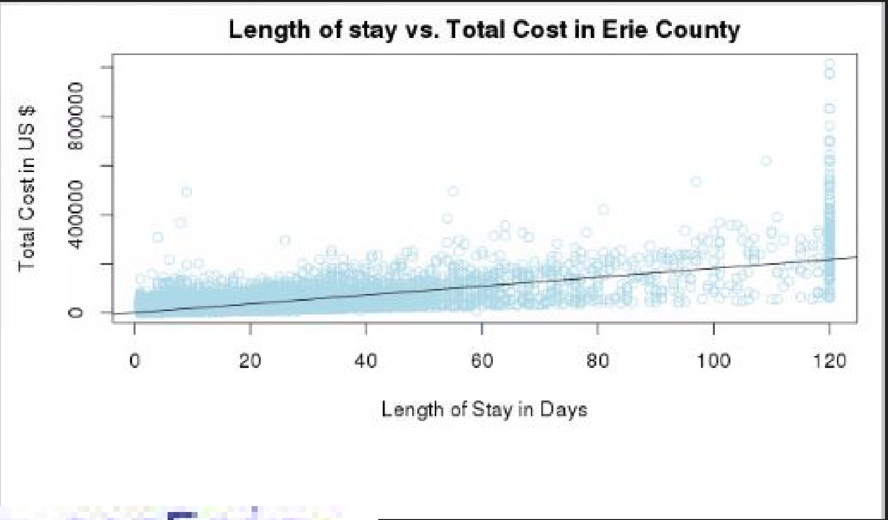
Length of Stay and Total Cost

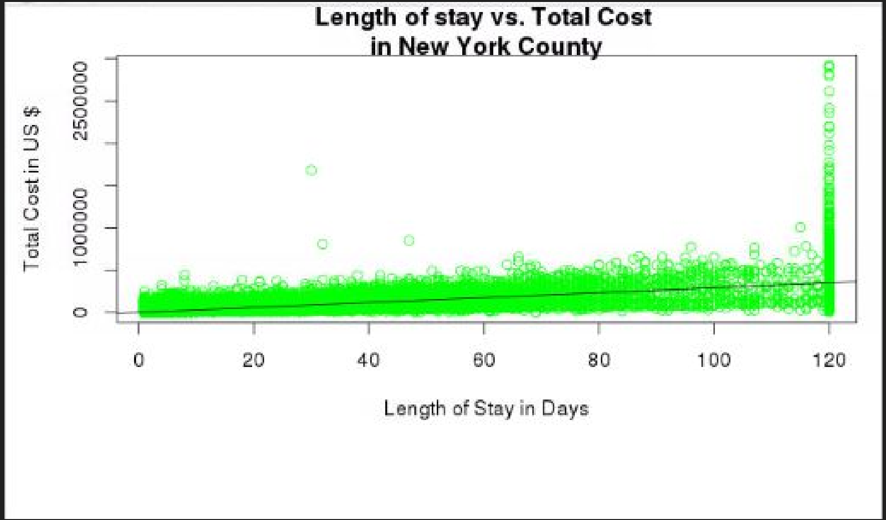



```
> print(correlation)
[1] 0.6630077
```

- Poor fit for data
- The correlation coefficient points to a positive and moderate correlation between length of stay and total costs.
- What can we do to better represent the data with regression?

Subsetting the Data by County






```
> corErie
[1] 0.7588
> corNY
[1] 0.7389
```

- Compared with correlation coefficient of entire dataset, subsetting by county increases correlation coefficient.

Subsetting Further

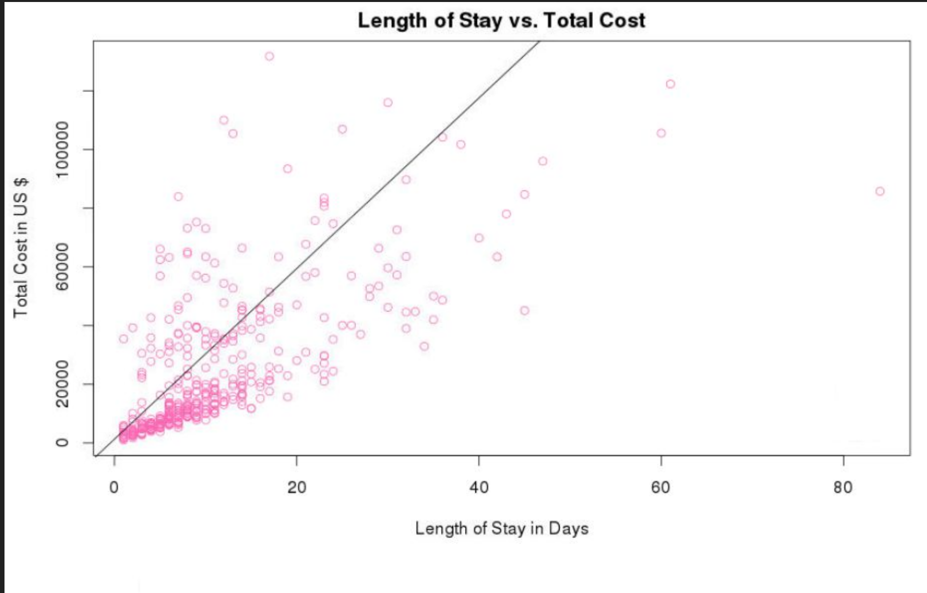
County: Erie County

Diagnosis: Circulatory disease

Severity of Illness: Extreme

Risk of death: Major

Subsetting the Data Even Further by Severe Conditions

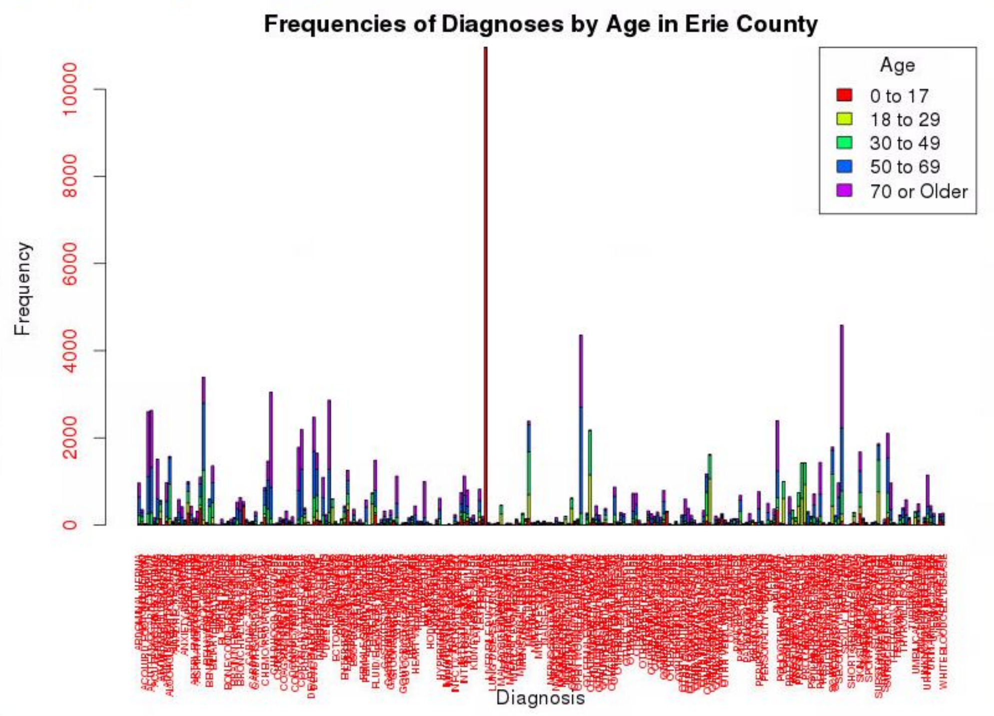



```
> print(corSevereCirculatory)
[1] 0.6813755
```

- Regression is Better Fit than the Regression for entire SPARCS dataset, but still not a good fit
- Correlation between length of stay and total cost is positive, moderate linear association
- Homoscedasticity is not satisfied as residuals increase as length of stay increases

Exploring the Data Using Bar Plots

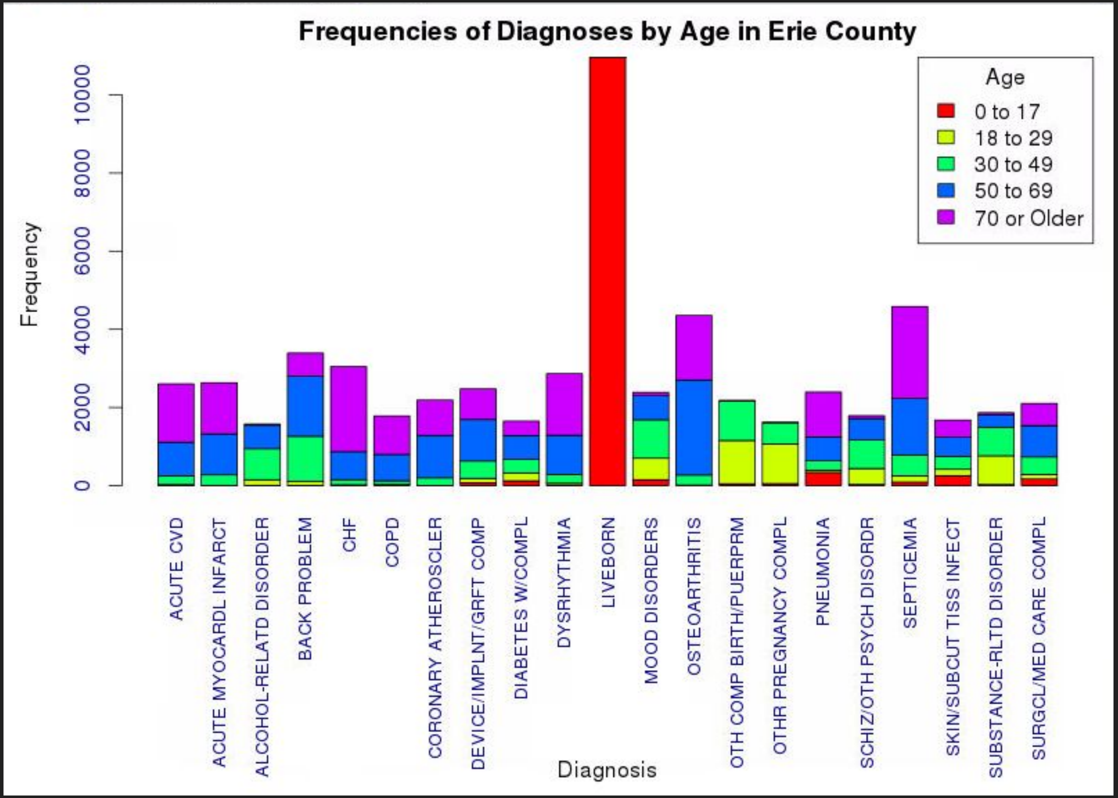
Barplot of Diagnosis in Subset by Age in Erie County (first attempt)



In all of Erie County, there are way too many different diagnoses to make sense of

What is that massive red bar?

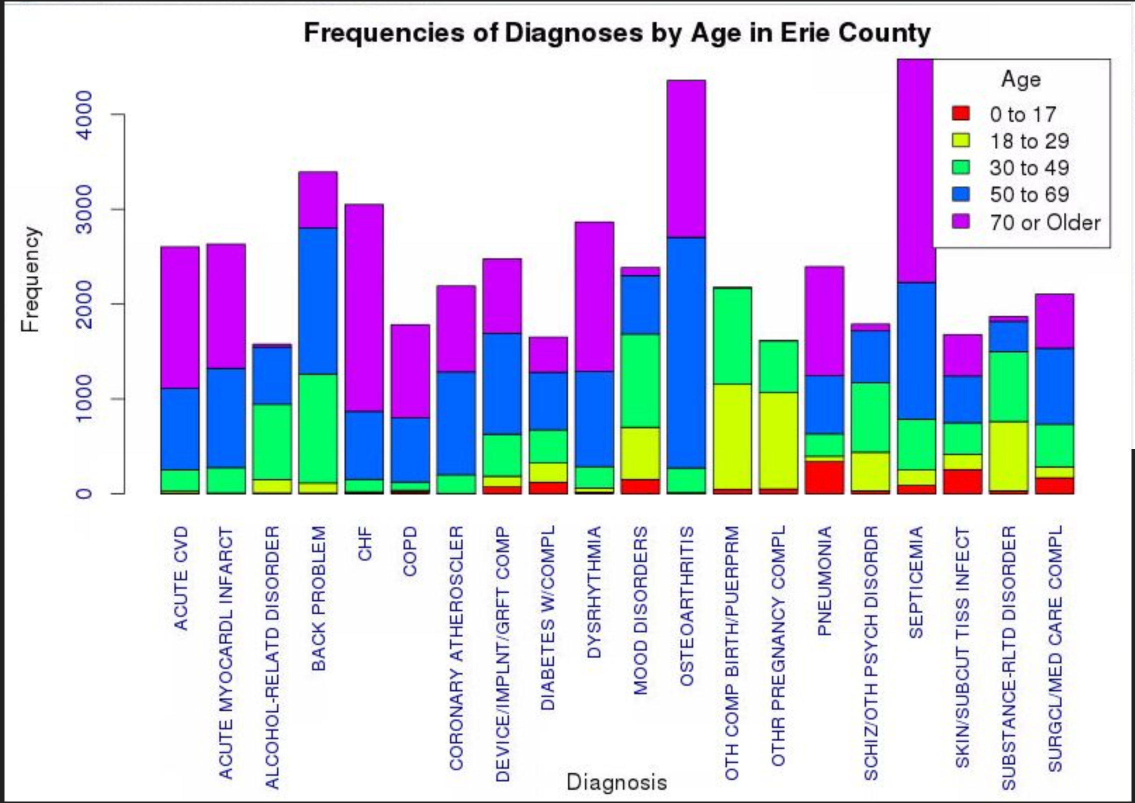
Barplot of Diagnosis in Subset by Age in Erie County



Massive red bar is live borns

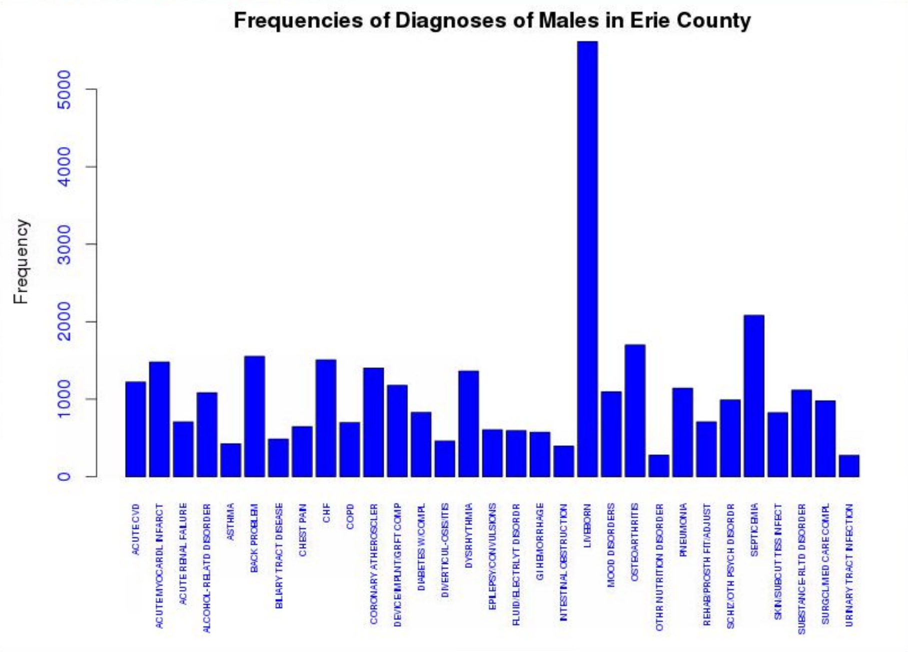
We can get a better picture still

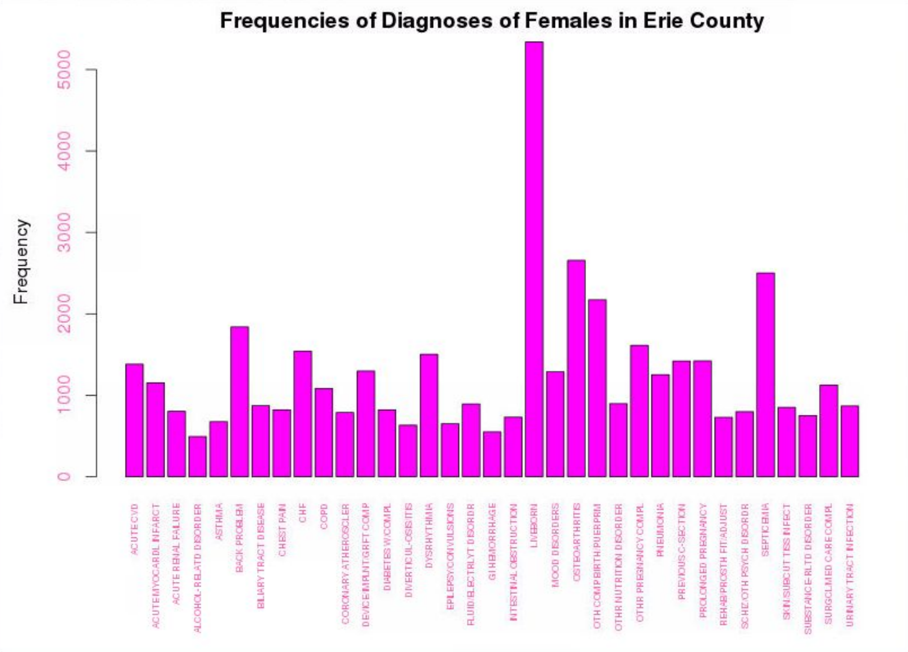
Barplot of Diagnosis in Subset by Age in Erie County



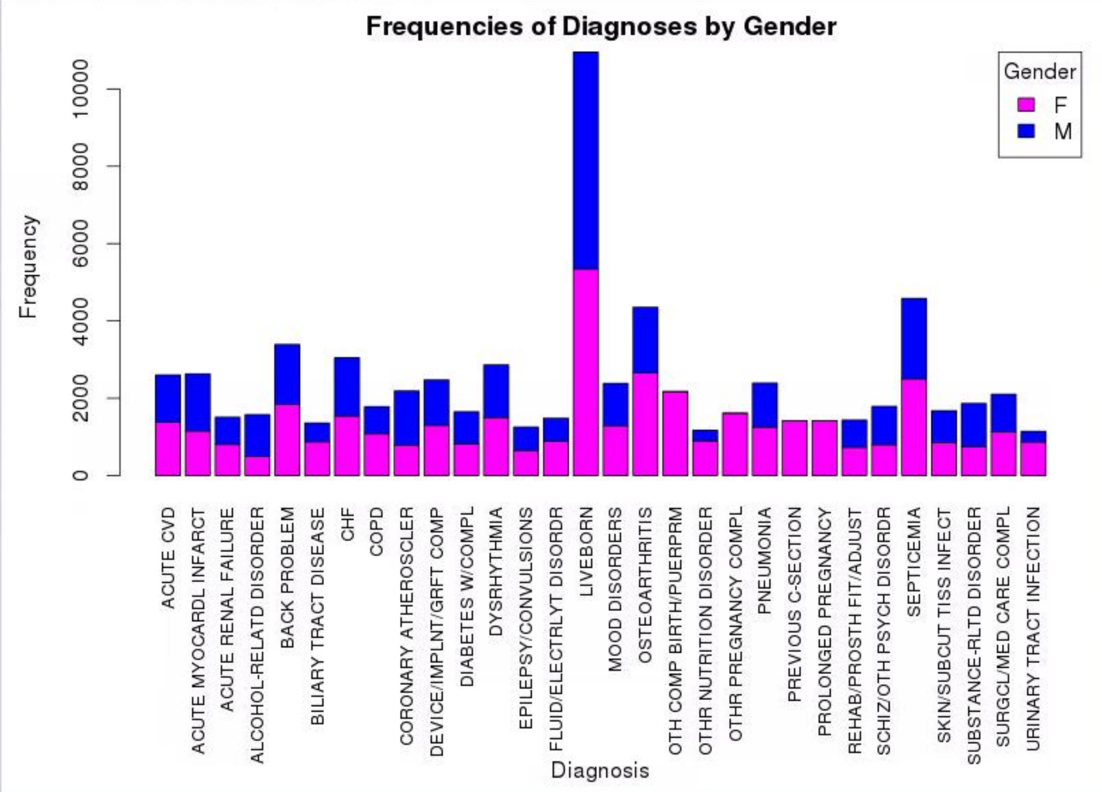

```
c = droplevels(Erie2)
counts=table(c$age, c$ccs_diagnosis)

par(mar=c(12,4.5,2,1))
barplot (counts,
         main="Frequencies of Diagnoses by Age in Erie County",
         ylab="Frequency",
         las=3,
         cex.names=0.8,
         col.axis="darkblue",
         col=colorsAge1,
)

mtext("Diagnosis", side = 1, line =10)

legend(x="topright",                          # location for legend
       title="Age",              # title for legend
       levels(factor(Erie2$age)),      # names in legend
       fill=rainbow(length(table(Erie2$age))))
```

Barplot of Diagnosis by Gender in Erie County (separate)

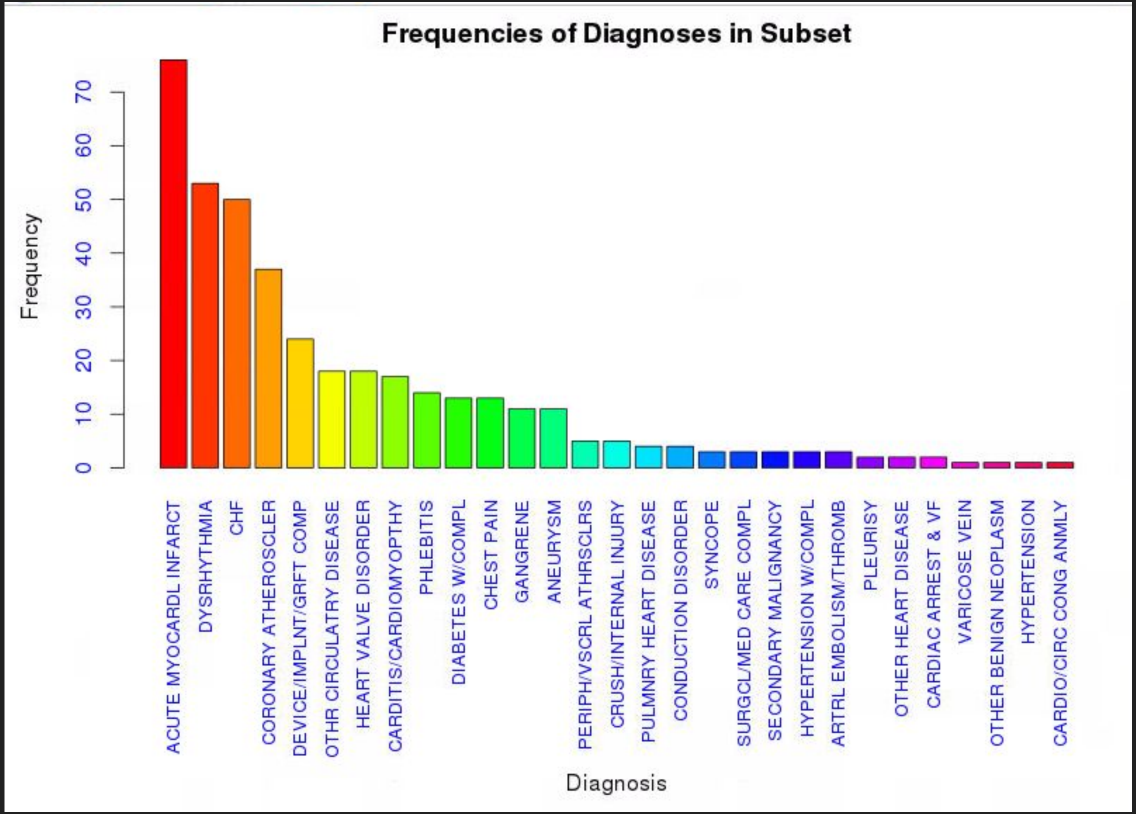




Barplot of Diagnosis by Gender in Erie County



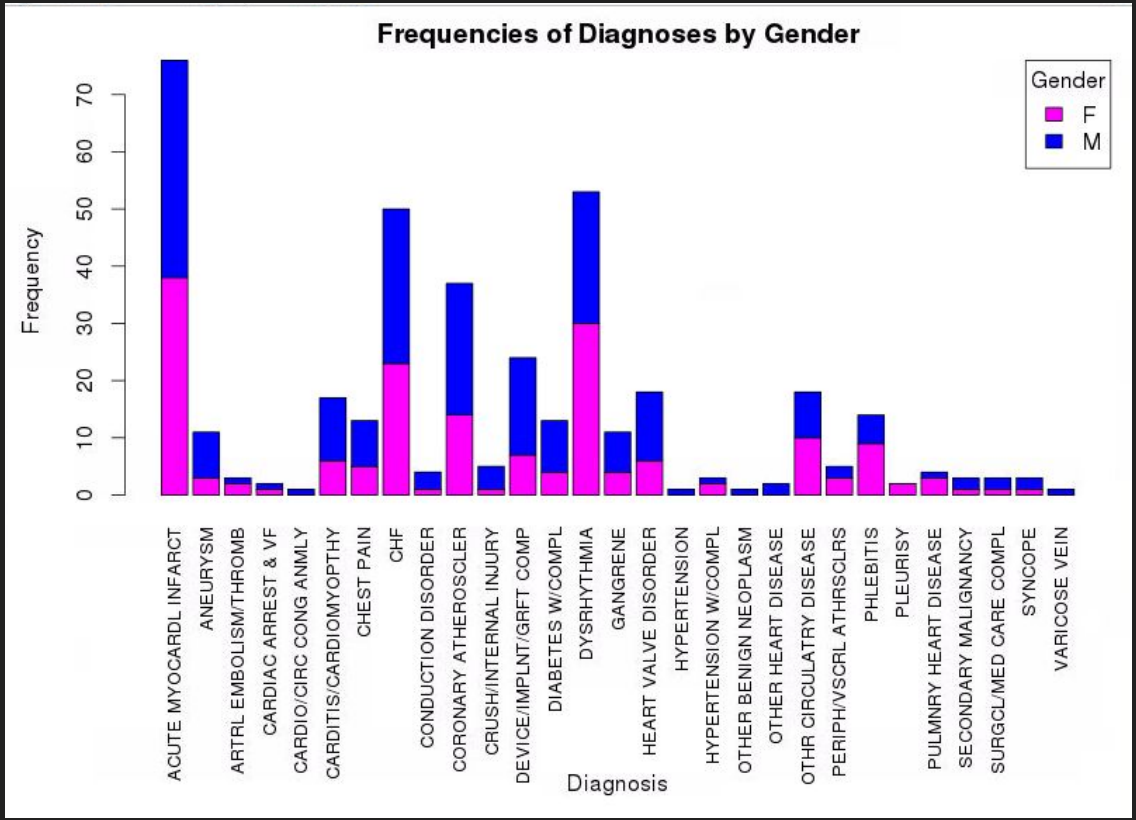
Barplot of Diagnoses in Subset



Some of highest are:

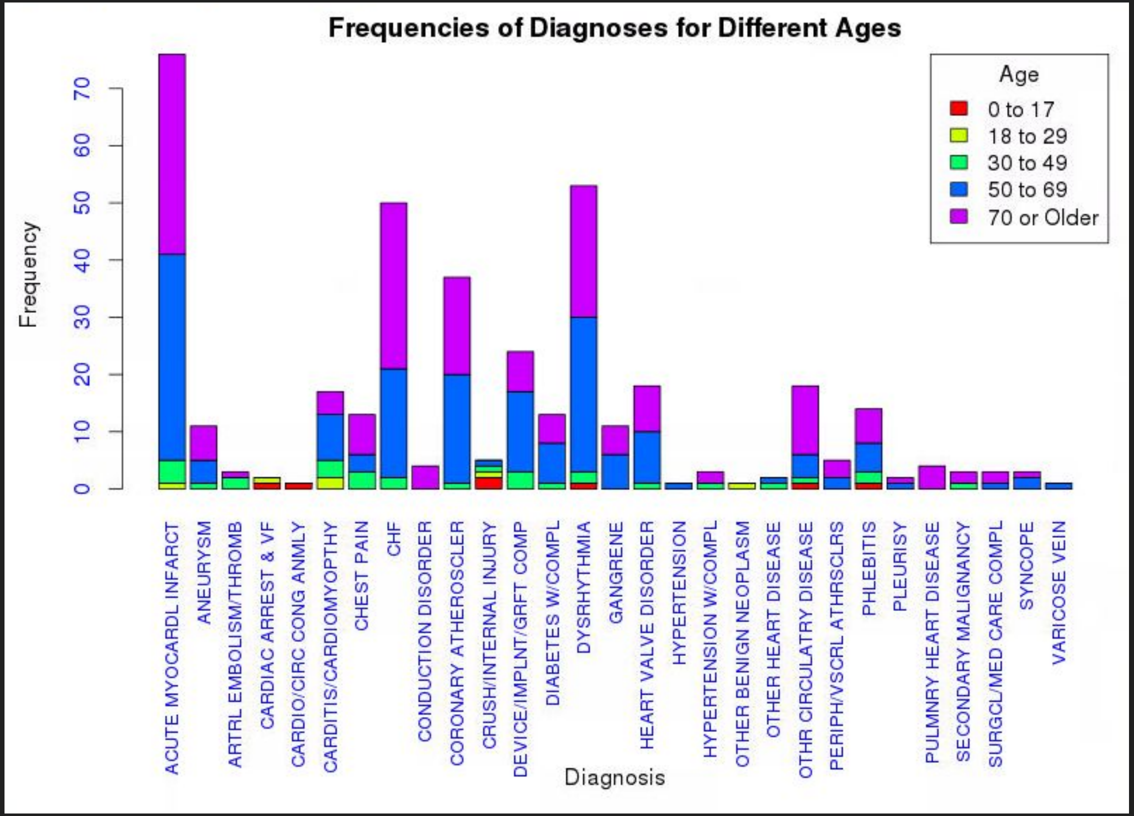
1. Acute Myocardial Infarct (Heart attack)
2. Dysrhythmia (Irregular heartbeat)
3. CHF (Congestive Heart Failure
4. Coronary Atherosclerosis

Barplot of Diagnosis by Gender in Subset



Most of diagnoses are split approximately 50/50

Barplot of Diagnosis by Age in Subset



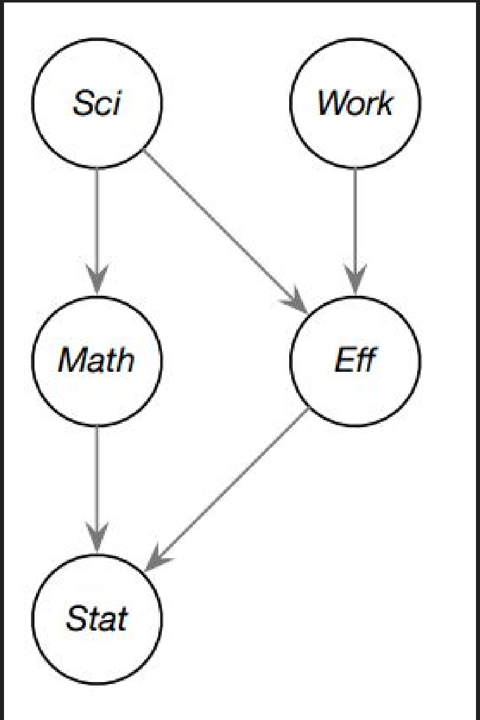
About equal numbers of diagnoses for the 50-69 and 70+ age groups

“Other benign neoplasm”

Using this data

“Using Past to Predict Future – Bayesian Networks and Medical Data”

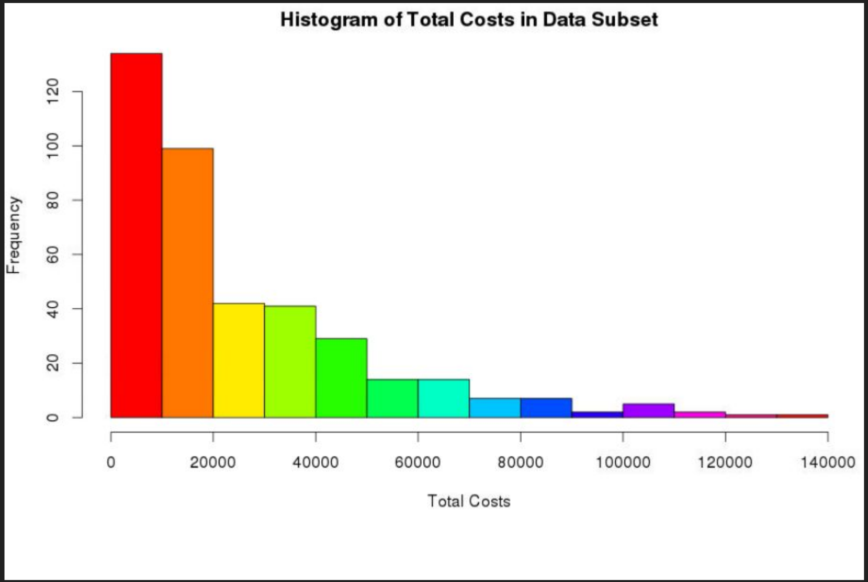
-Dr. Jaroslaw Zola

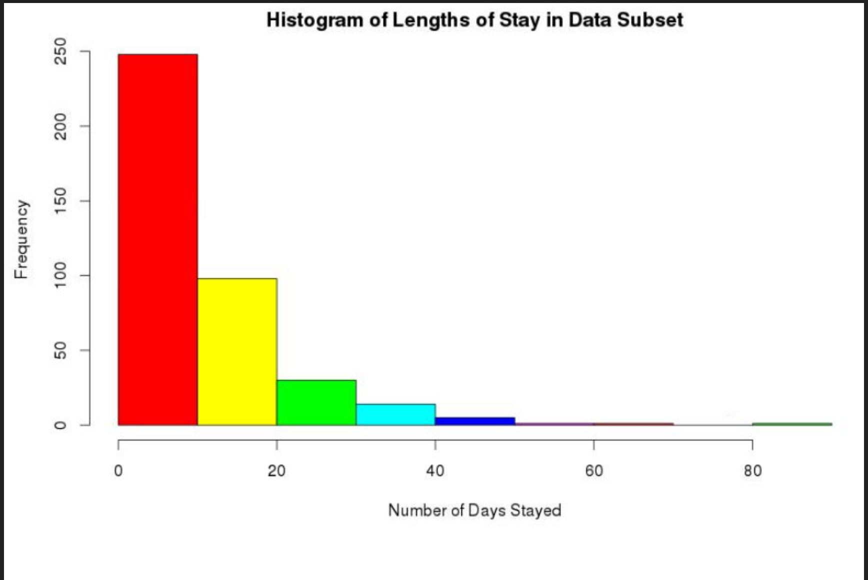


Diagnoses are given when patients enter the hospitals

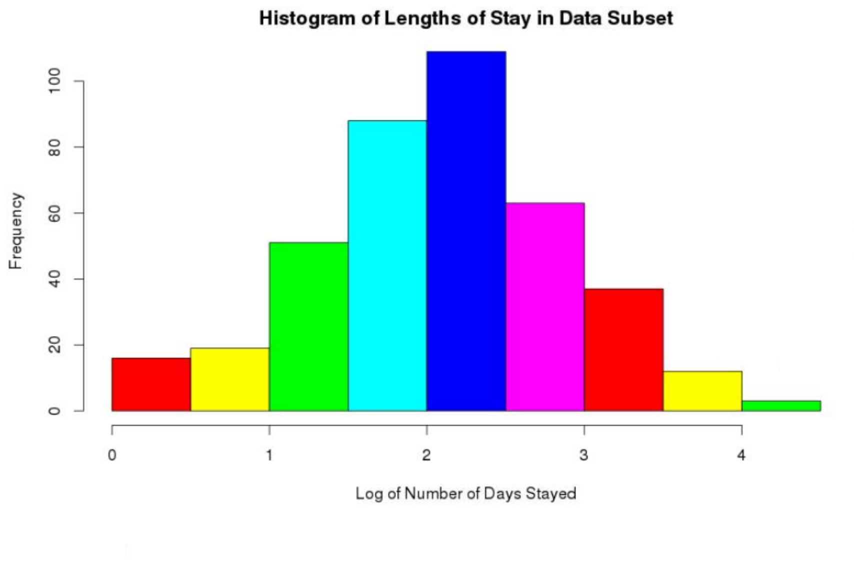
Can be used to predict who will develop critical conditions

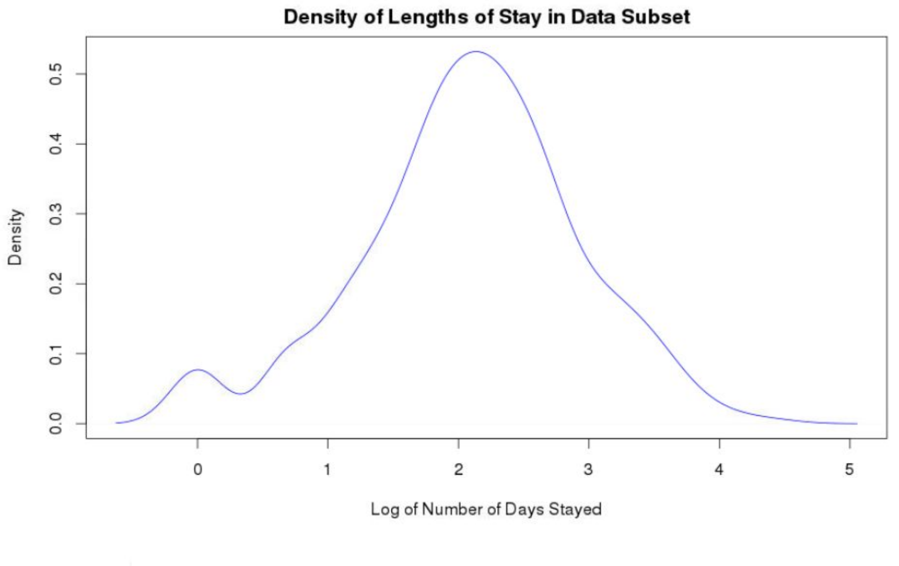
Length of Stay and Total Cost





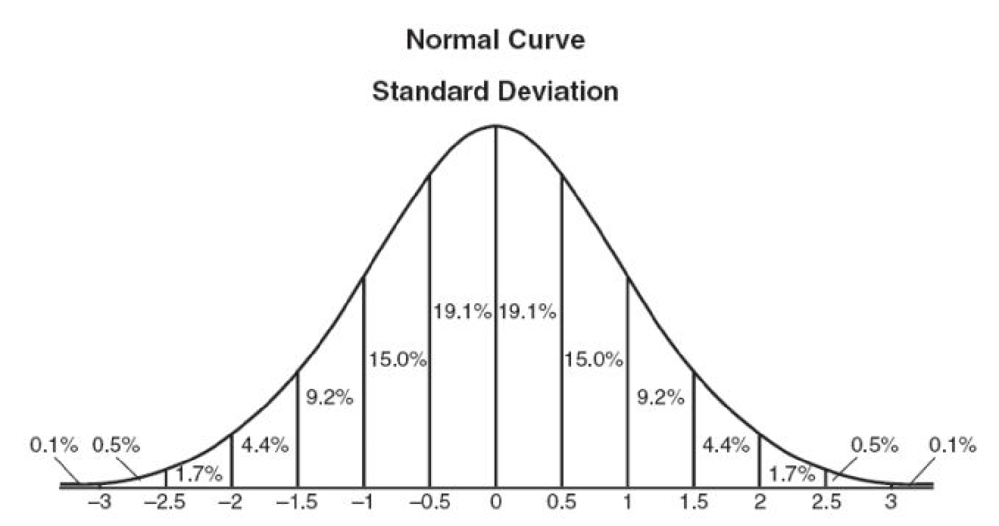
The Logarithm and its relation to distribution



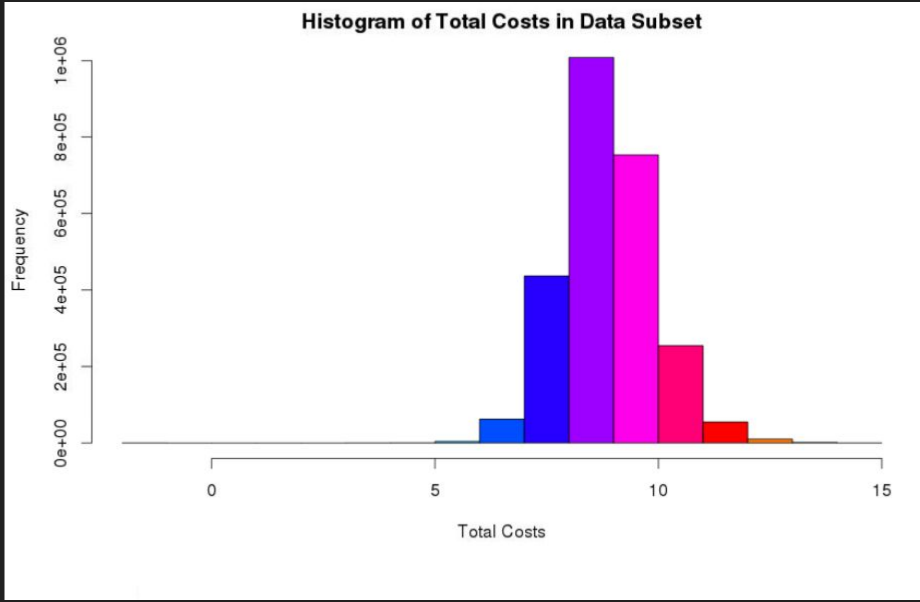


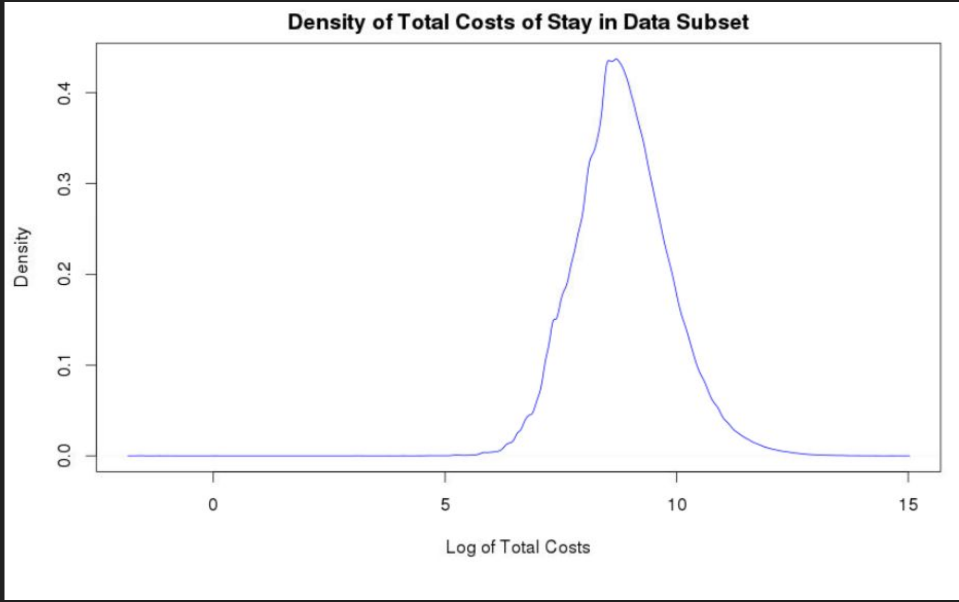
Normal Distribution

Dr. Winklestein spoke to us about how this distribution is similar to a normal bell curve. The cause of this occurrence is still unknown by experienced statisticians



We continue to see this trend





Our experience in this workshop introduced us to the world of computational science and gave us a taste of what type of work can be found in a real world environment.