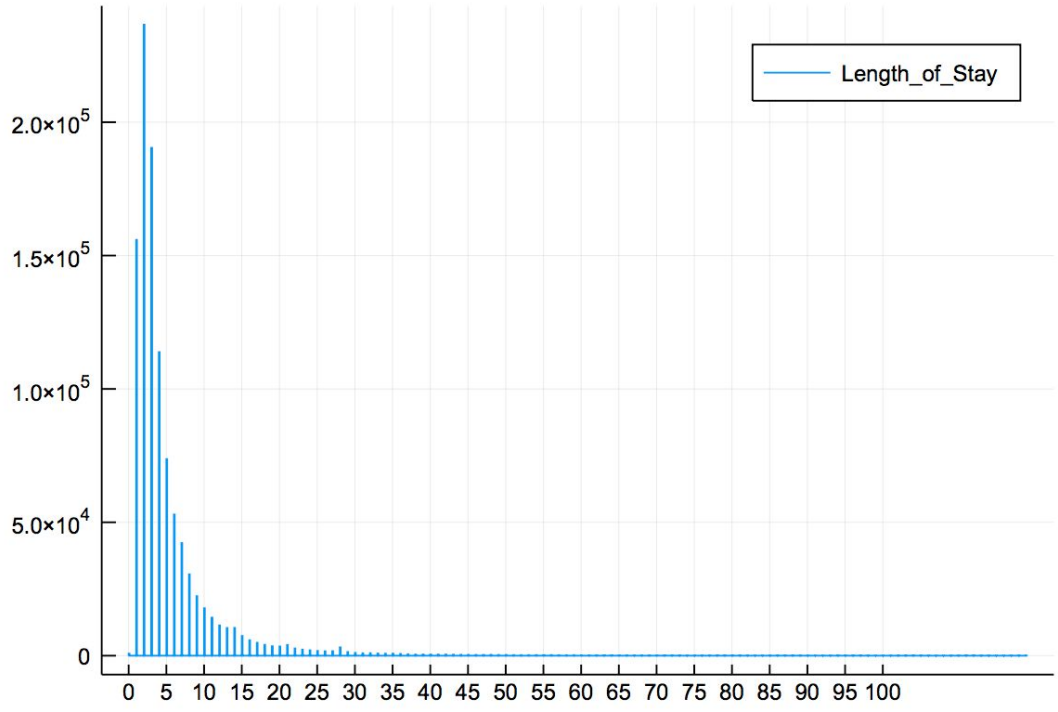
Problem Statement

Our project aims at giving the patients helpful information on choosing a hospital. Given their personal information and medical history, we hope to estimate their length of stay and the cost for staying at particular hospitals. We hope that based on our model, the patients are able to choose the hospital more scientifically based on their symptoms and budgets.

About the Dataset

We use the Statewide Planning and Research Cooperative System's (SPARCS) Hospital Inpatient Discharges dataset for hospitals in New York State. This dataset contains 2,544,543 examples and 35 different features.

```
maximum: 119.0
minimum: 0.0
mean:    5.324166607061965
std:     7.29668777777166
```



According to the histogram of length of stay feature shown above, the average length of stay is approximately 5.32 days with maximum days of 119 days and minimum days of 0 days. Since the average length of stay in a hospital is used to gauge the efficiency of a healthcare facility and the national average for a hospital stay is 4.5 days, less than 50% of the hospitals in New York States are able to operate in an efficient manner. Also, there are many outliers that are far from the mean length of stay days.

According to the Agency for Healthcare Research and Quality, the average cost is $10400 per day, which exceeds the maximum cost in New York State, which is 999.98.

Data Cleaning

We first filter our dataset and only look at examples for which the information we care about is known (APR_Risk_of_Mortality, Age_Group, Race, Ethnicity, Length_of_Stay, Total_Charges, Emergency_Department_Indicator). After applying the filter, we have 1048533 examples. We consider it enough for our project.

Then, for the convenience of doing regression, we converted some nominal variables into numerical values. For example, we used one-hot encoding to convert Race column which contains the patient's racial information into a four-column vector: [1(race = "White"), 1(race = "Other Race"), 1(race = "Black/African American"), 1("Unknown")]. In a similar manner, we also converted the Age_Group column into ["70 or Older", "50 to 69", "30 to 49", "0 to 17", "18 to 29"], the Ethnicity column into ["Not Span/Hispanic", "Unknown", "Spanish/Hispanic"], and Payment_Typology_1 into ["Medicare", "Medicaid", "Blue Cross/Blue Shield", "Private Health Insurance", "Managed Care, Unspecified", "Self-Pay", "Department of Corrections", "Miscellaneous/Other", "Federal/State/Local/VA", "Unknown"].

Preliminary Model

Our first trial is to predict the patient's risk of mortality. We picked the features that we think might be influential (Age_Group, Ethnicity, Race, Length_of_Stay, Gender), and run linear regression on each of these features. We also run linear regression on different combinations of these features. After comparing the MSE of different combinations, we decide that we will fit a model against Age_Group, Ethnicity, Race and Length_of_Stay to predict the patient's risk of mortality.

Future Work

While we have our preliminary model, we find out that it is too general and does not give much useful information that helps hospital selection. We would like to narrow our target and focus on a particular case (for example, patients with heart attack), and examine the risk of mortality of that case. We would also like to take the cost into consideration. We came up with two possible ways to improve our model:

1. Instead of using APR_Risk_of_Mortality as our target, we will come up with a "score" that take into consideration more aspects that will affect the choice of hospital.

2. We will fit one model that predicts APR_Risk_of_Mortality, and another model that predicts costs.
3. We will focus on a particular disease so that our model is more precise. In the future, we can also develop similar models for other diseases. In this way, patients with different diseases can refer to our models accordingly.