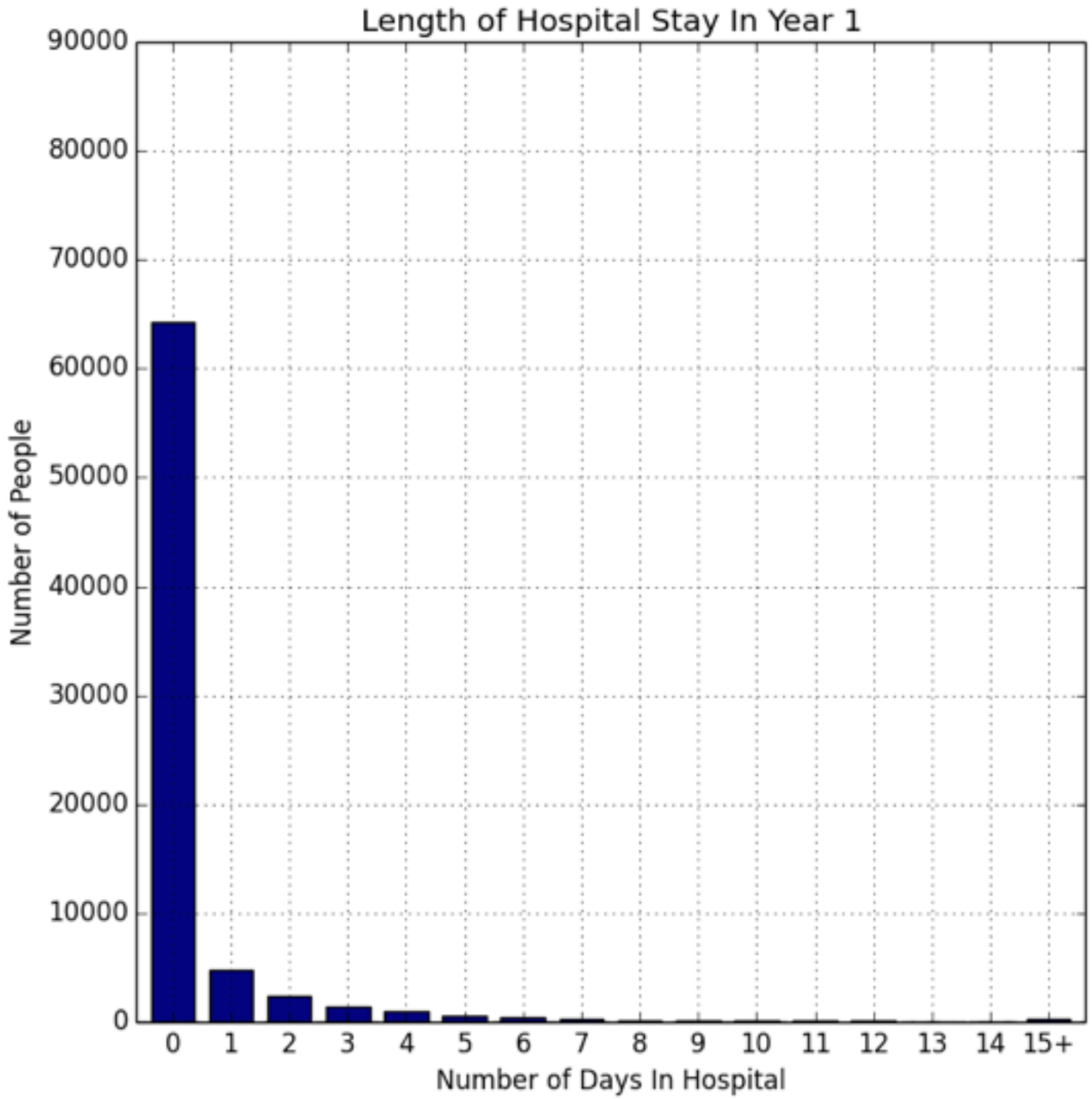
Health Net Data Project

Joe Rummel

Problem

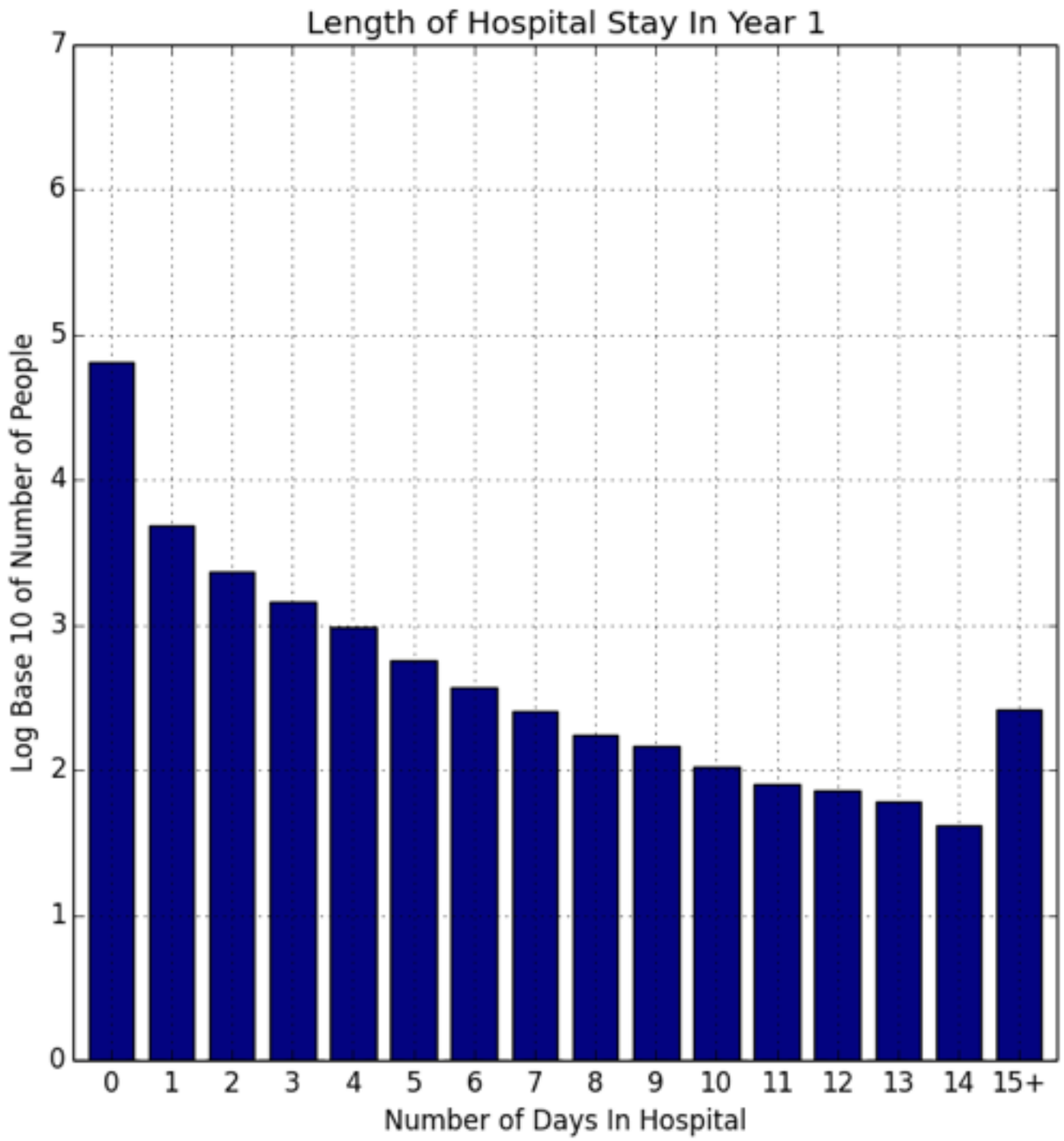
- > 71 million people admitted to hospitals in the U.S. every year
- > $30 billion spend on unnecessary hospital stays
- Use claims data from insurer to predict who will spend time in the hospital the following year

Data Overview



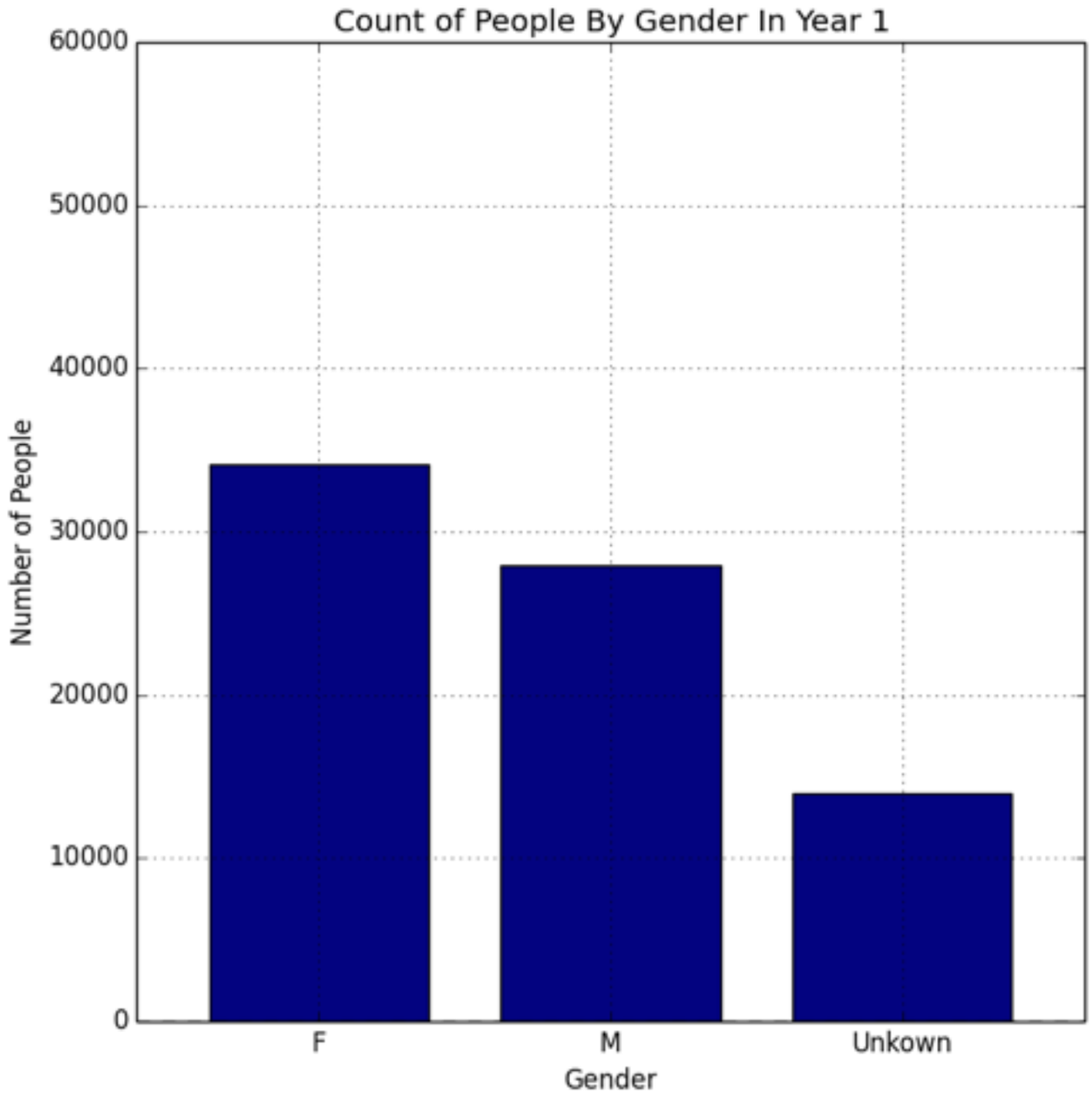
Data is very skewed toward zero days in the hospital

Data Overview



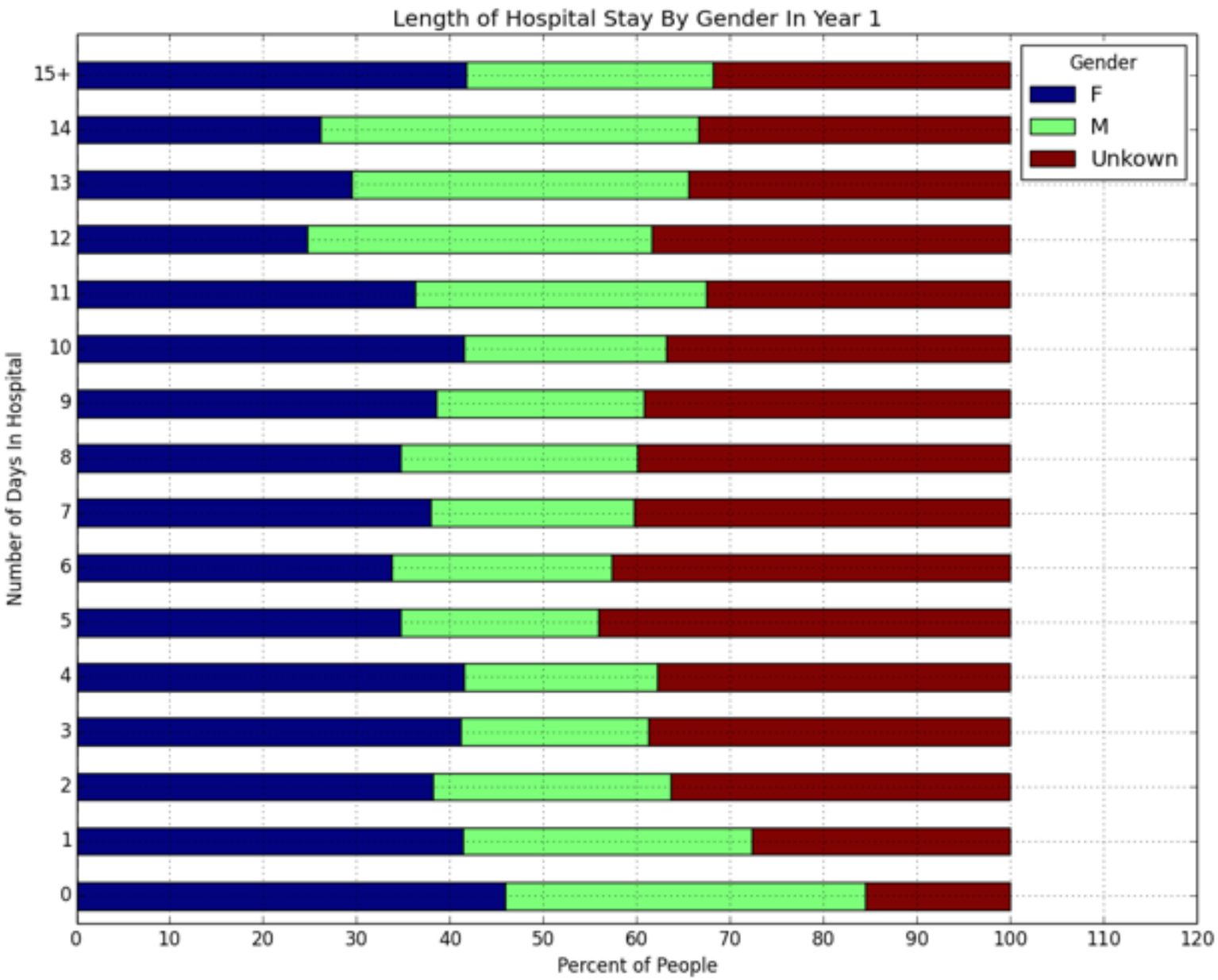
Downward trend for number of people vs. days in hospital

Data Overview



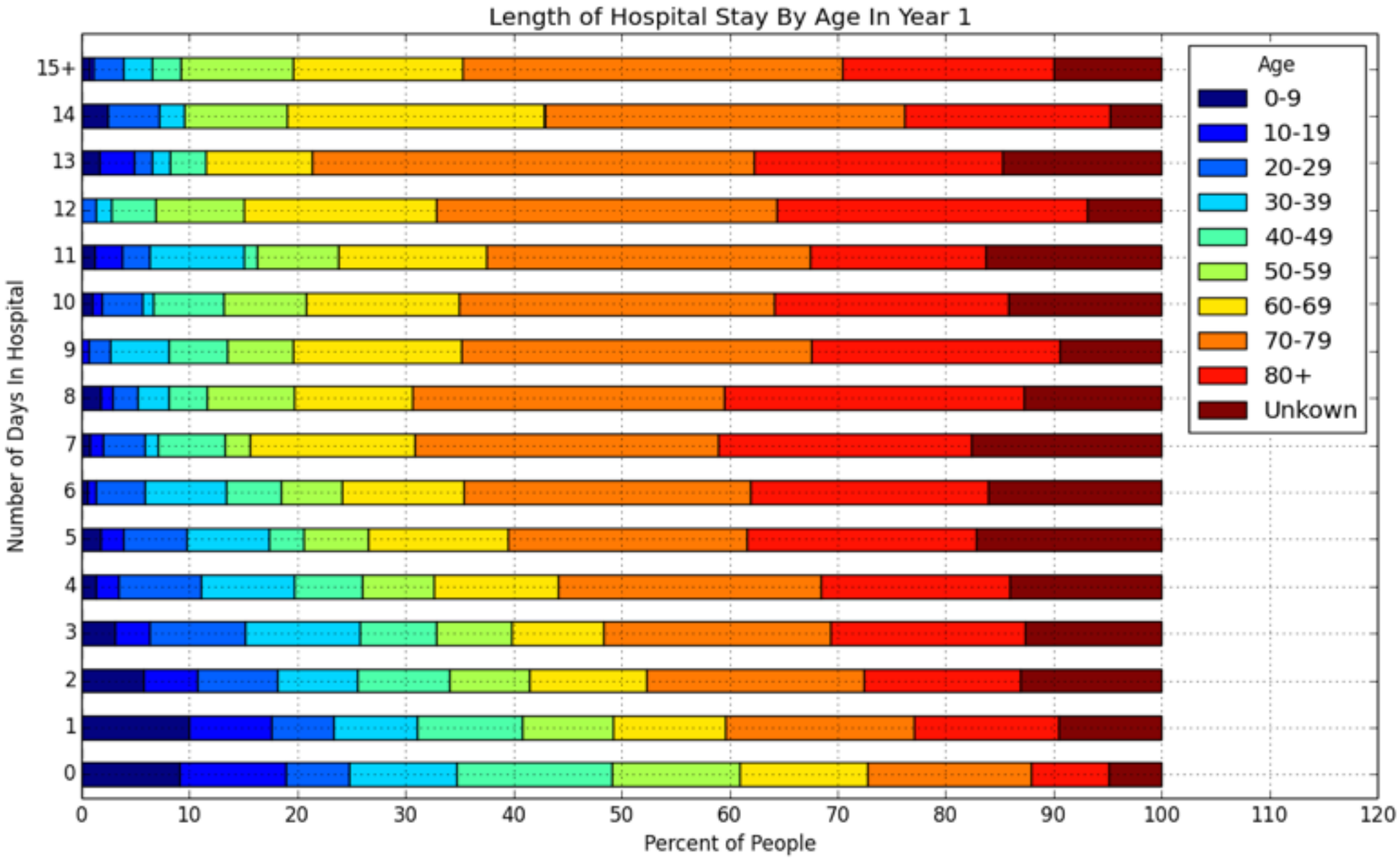
- More females than males, but not many
- Significant number of people with unknown gender

Data Overview



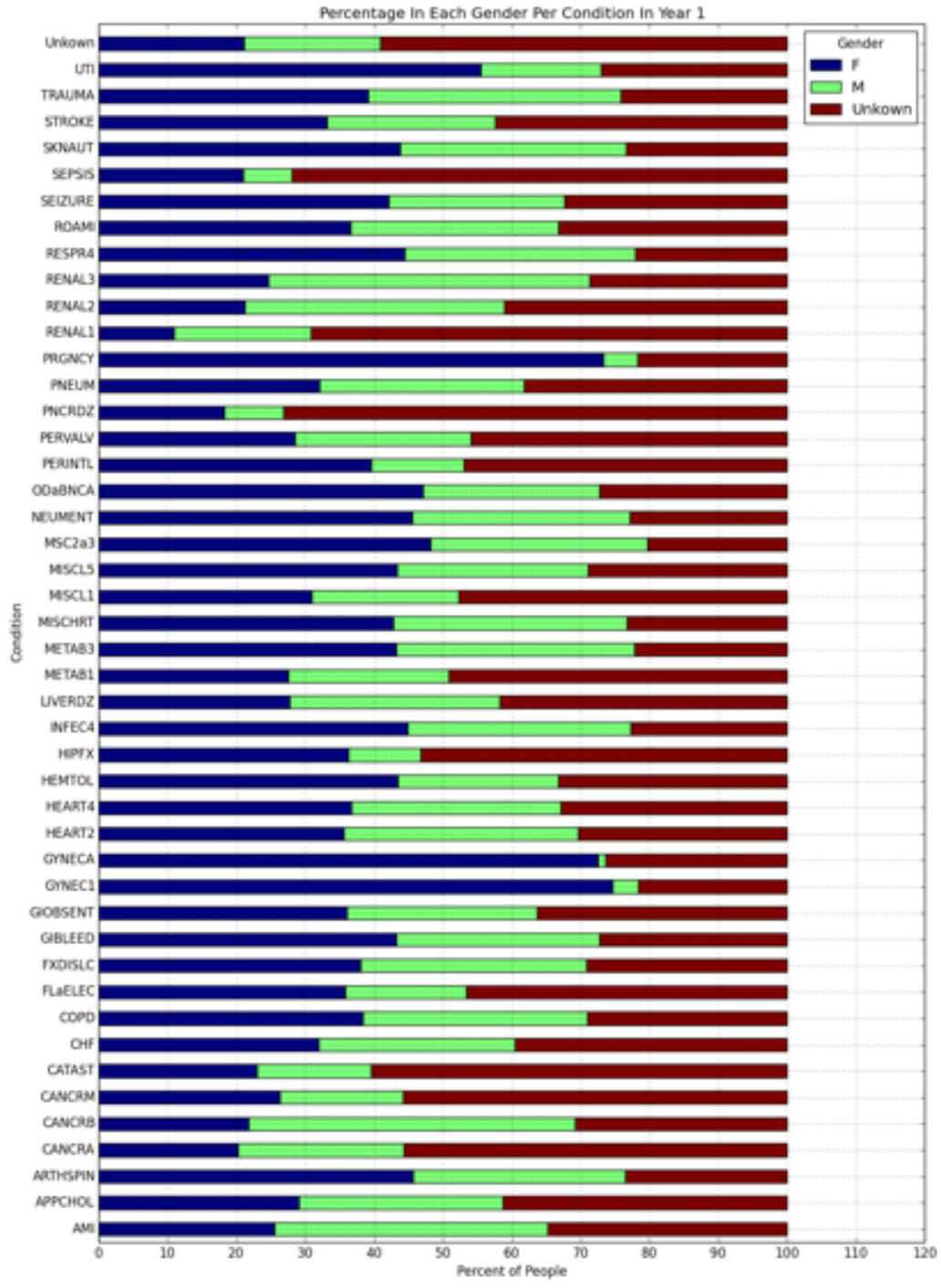
People with unknown gender account for significant percentage of people who spent time in hospital

Data Overview



- Many people in the hospital have unknown age
- Older people spent more time in hospital than younger people

Data Overview



There is some bad data (i.e. pregnant males, males with gynecological issues)

Parameters

- Used gender, age, conditions, number of prescriptions, number of lab tests, number of claims
  - Gender: “M”, “F”, “Unknown”
  - Age: “0-9”, “10-19” … “80+”, “Unknown”
  - Conditions: Binary number
  - Prescription and Lab Count: Average per visit
  - Number of claims: Total number of claims the patient made that year

Modeling Technique

- The data is very skewed, so linear methods will not work. Need proof?

Logistic Regression Accuracy:
0.849807517323

Metrics Report for Logistic Regression:

	precision	recall	f1-score	support
0	0.85	1.00	0.92	60706
1	0.00	0.00	0.00	4464
2	0.00	0.00	0.00	2182
3	0.00	0.00	0.00	1429
4	0.00	0.00	0.00	842
5	0.00	0.00	0.00	528
6	0.00	0.00	0.00	287
7	0.00	0.00	0.00	218
8	0.00	0.00	0.00	143
9	0.00	0.00	0.00	115
10	0.00	0.00	0.00	103
11	0.00	0.00	0.00	65
12	0.00	0.00	0.00	62
13	0.00	0.00	0.00	50
14	0.00	0.00	0.00	23
15+	0.00	0.00	0.00	218
avg / total	0.72	0.85	0.78	71435

Modeling Technique

- Decision Tree responds better to skewed data
- Used 10 tree Random Forest, which increased accuracy over a single decision tree by 10%
- Random Forest in sklearn had an option to weight the target data, which helps increase accuracy when dealing with skewed data

Modeling Technique

- Results: Not great, but it did not predict everything to be zero

Metrics Report For Random Forest:

	precision	recall	f1-score	support
0	0.85	0.93	0.89	60706
1	0.06	0.03	0.05	4464
2	0.05	0.02	0.03	2182
3	0.02	0.01	0.01	1429
4	0.03	0.01	0.02	842
5	0.05	0.01	0.02	528
6	0.01	0.00	0.00	287
7	0.01	0.00	0.01	218
8	0.01	0.01	0.01	143
9	0.00	0.00	0.00	115
10	0.00	0.00	0.00	103
11	0.00	0.00	0.00	65
12	0.00	0.00	0.00	62
13	0.00	0.00	0.00	50
14	0.00	0.00	0.00	23
15+	0.02	0.00	0.01	218
avg / total	0.73	0.80	0.76	71435

Modeling Technique

- Dimension Reduction did not help, which was no surprise since there were not many parameters to begin with

Random Forest 10-Fold Cross Validation:
[0.82496055 0.8229879   0.81917412 0.81772751 0.82114676 0.81693845
  0.81772751 0.81917412 0.81967644 0.81573063]

Random Forest With SVD Dimension Reduction 10-Fold Cross Validation:
[0.80865334 0.80733824 0.80746975 0.80628617 0.80746975 0.81128353
  0.80996844 0.80576013 0.80704985 0.80704985]

Implementation Plan

- Get the data into a SQL database
- Run the data munging script to create a clean data file
- Run the modeling script

Conclusions

- Domain expertise is essential
- Further refining of the Decision Tree is needed
- Handling skewed data is a big challenge
- Better data would probably get better results