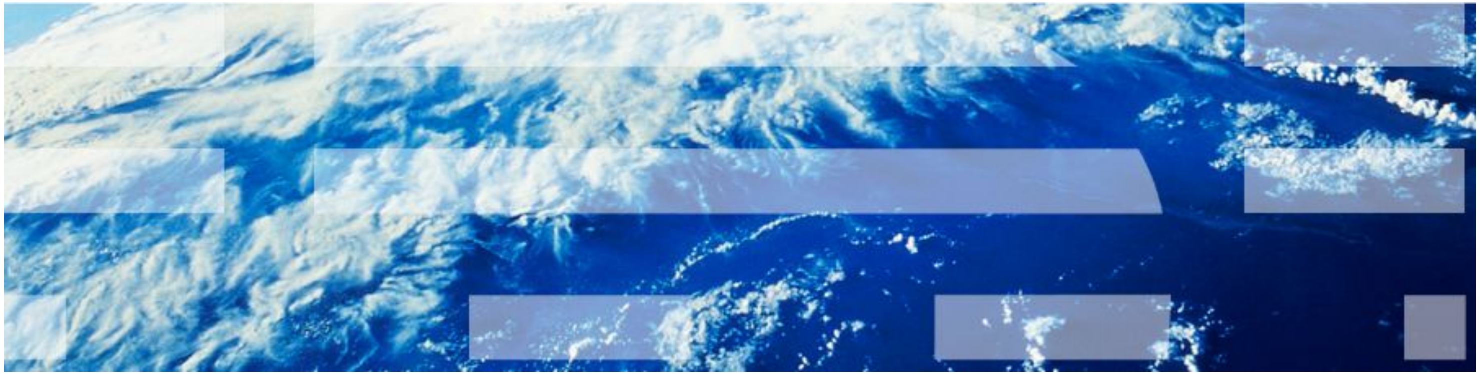
E6893 Big Data Analytics

NY State Inpatient Healthcare Analysis

Project ID: 201912-28
Team Members (with UNI):

Justin Cho (yc3522), Yuan-Fang Lin (yl4042), Jing Qian (jq2282)



Intro and Goal

Introduction:

Healthcare is becoming ever more expensive in the modern era yet access to care and supply of services is not always optimized for the demands of growing population. We will aim to explore healthcare trends in the New York State geography and uncover underlying patterns in healthcare trends across various cohorts based on age, gender, race, location, and diagnoses to drive insights that can lead to better healthcare service and access.

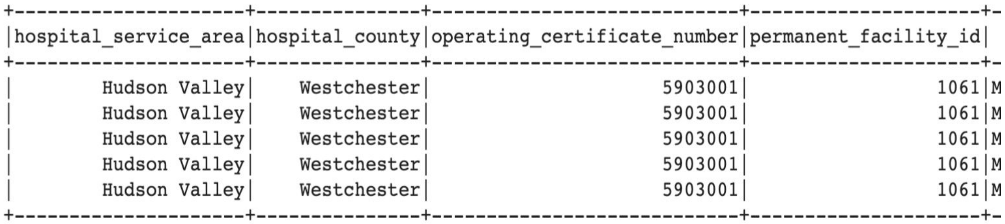
Goal:

To develop a better understanding of the healthcare trends and patterns in NY State and also translate findings into a visual interface where users can derive actionable insights to inform decisions and policy-making.

Dataset

The dataset used is the New York State Hospital Inpatient Discharges: 2017 which includes 2.3 million inpatient records with 34 features including age, gender, zip code, ethnicity, race, and diagnosis.

Data source: https://on.ny.gov/2qa8Qlm



```
+---------------------+---------------+---------------------------+---------------------+-
|hospital_service_area|hospital_county|operating_certificate_number|permanent_facility_id|
+---------------------+---------------+---------------------------+---------------------+-
|        Hudson Valley|    Westchester|                    5903001|                 1061|M
|        Hudson Valley|    Westchester|                    5903001|                 1061|M
|        Hudson Valley|    Westchester|                    5903001|                 1061|M
|        Hudson Valley|    Westchester|                    5903001|                 1061|M
|        Hudson Valley|    Westchester|                    5903001|                 1061|M
+---------------------+---------------+---------------------------+---------------------+-
```

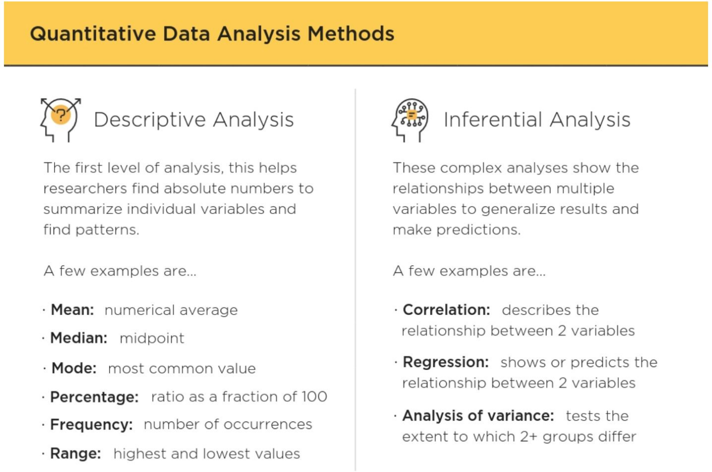
Note about Compliance and HIPAA:

The data set contains basic record level details and does not include protected health information (PHI) under HIPAA. Data records are not individually identifiable, and data that are considered identifiable have been redacted. De-identified data is public under Freedom of Information Law (FOIL).

Feature Descriptions:
https://on.ny.gov/2JCA9vF

Methods

- Descriptive analysis used for analyzing single variable, showed as bar graph, pie chart, etc.
- Inferential analysis used for looking for relationships between multiple features
  - Correlation matrix for summarizing data and representing relevance
  - Regression Model for potential output prediction, e.g. total cost/charge, diagnoses
    - Potential Models: Random Forest, Decision Tree, Gradient Boosting
- K-means or Hierarchical Clustering for data analysis
- Data visualization in high dimension, e.g. t-SNE



Quantitative Data Analysis Methods

Descriptive Analysis

The first level of analysis, this helps researchers find absolute numbers to summarize individual variables and find patterns.

A few examples are...

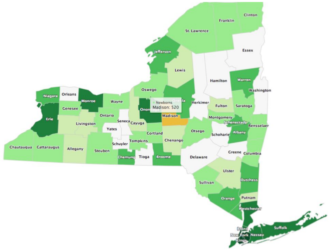
- **Mean:** numerical average
- **Median:** midpoint
- **Mode:** most common value
- **Percentage:** ratio as a fraction of 100
- **Frequency:** number of occurrences
- **Range:** highest and lowest values

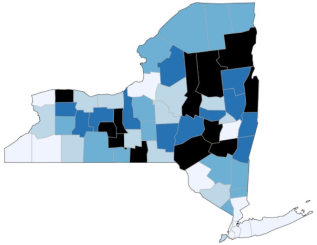
Inferential Analysis

These complex analyses show the relationships between multiple variables to generalize results and make predictions.

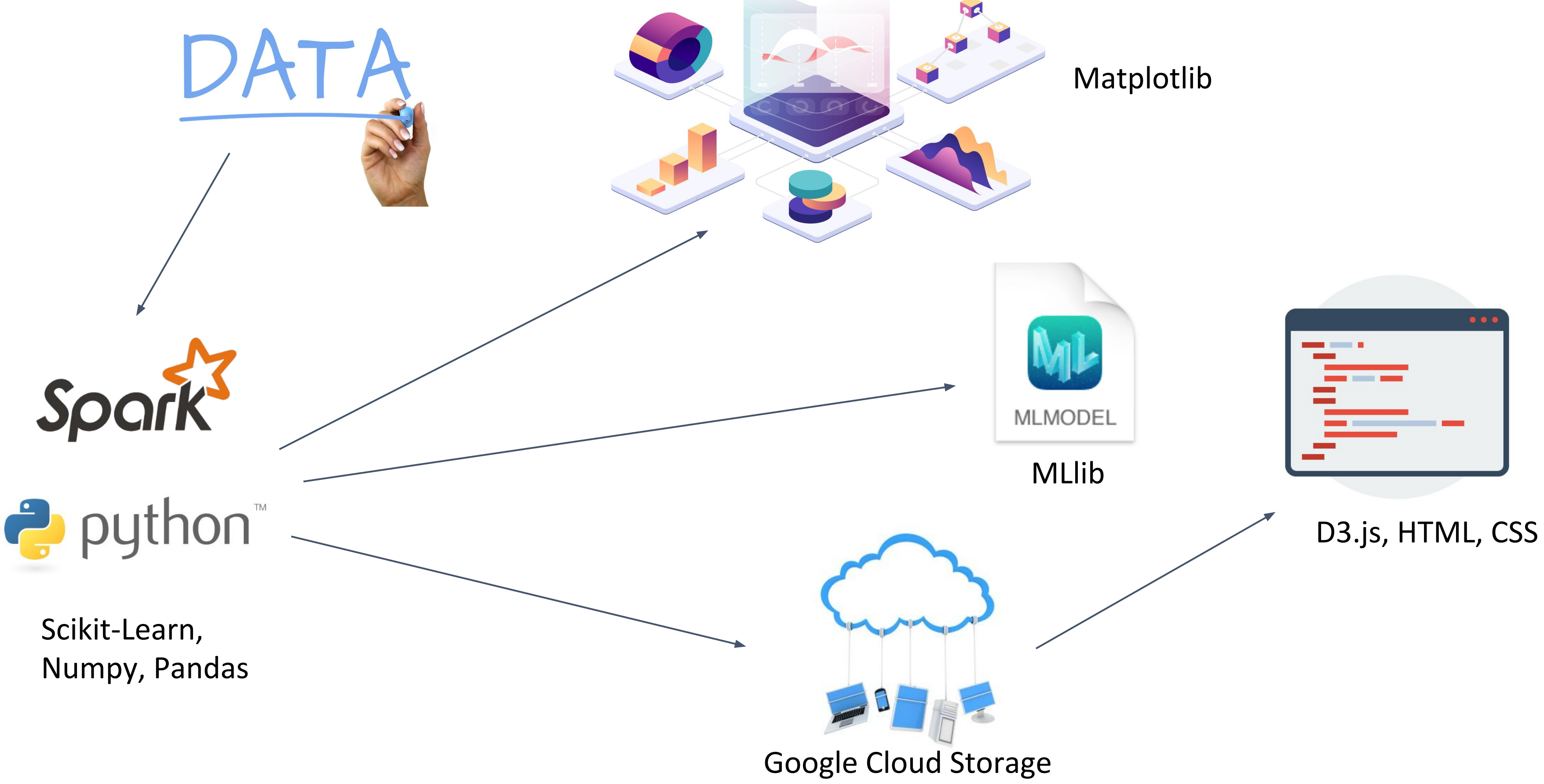
A few examples are...

- **Correlation:** describes the relationship between 2 variables
- **Regression:** shows or predicts the relationship between 2 variables
- **Analysis of variance:** tests the extent to which 2+ groups differ





System



Matplotlib

MLlib

D3.js, HTML, CSS

Scikit-Learn,
Numpy, Pandas

Google Cloud Storage

Expected Outcome and Schedule

Expected Outcome:

Through data exploration and analysis, we will generate a better understanding of NY state healthcare service. We will model trends (e.g. total cost or diagnoses) and also refine data visualization for various features (e.g. age, race, type of admission, or location) which can help users gain insight into underlying patterns in healthcare services consumption. We could also help potential patients with hospital recommendation and healthcare tips through front-end applications.

Schedule:

11/4 - 11/10 : Data cleaning, exploratory data analysis and data summary
11/11 - 11/17 : Begin analytics, test hypothesis and visualization examples
11/18 - 11/24 : Draft/Sketch front-end and begin implementation of pipeline
11/25 - 12/2 : Experiments and refine visualization
12/3 - 12/13 : Finalize reports and presentation video

Explorations of data

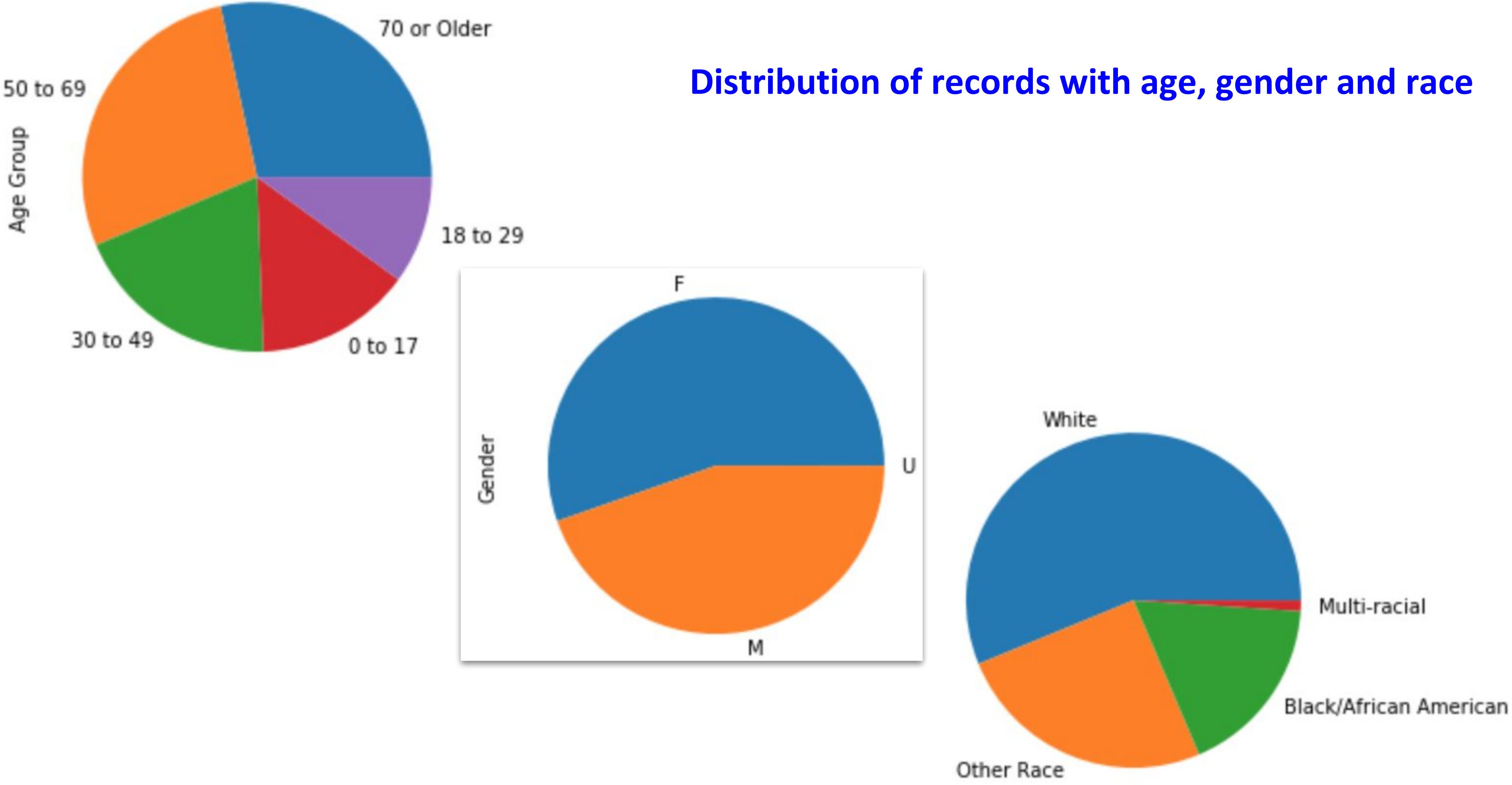
Taking the dataset from 2017 as example, it has 2,343,569 rows of records and 34 columns (features). Some inputs are missing so we need to do some data cleaning before further analysis.

Some results of preliminary data analysis are shown in the next three pages.

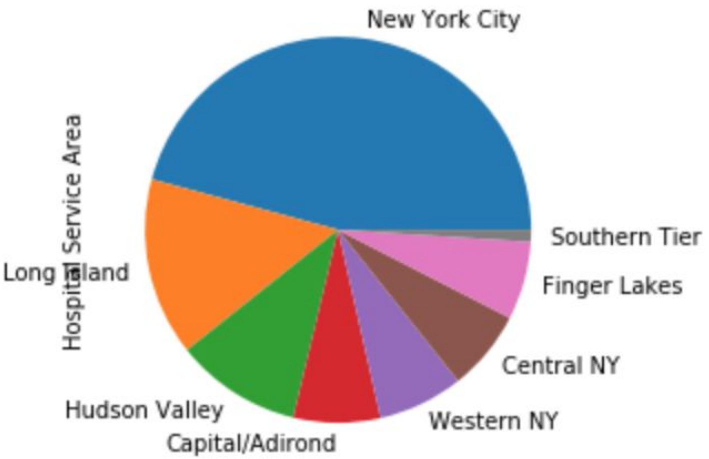
Some potential idea about our project:

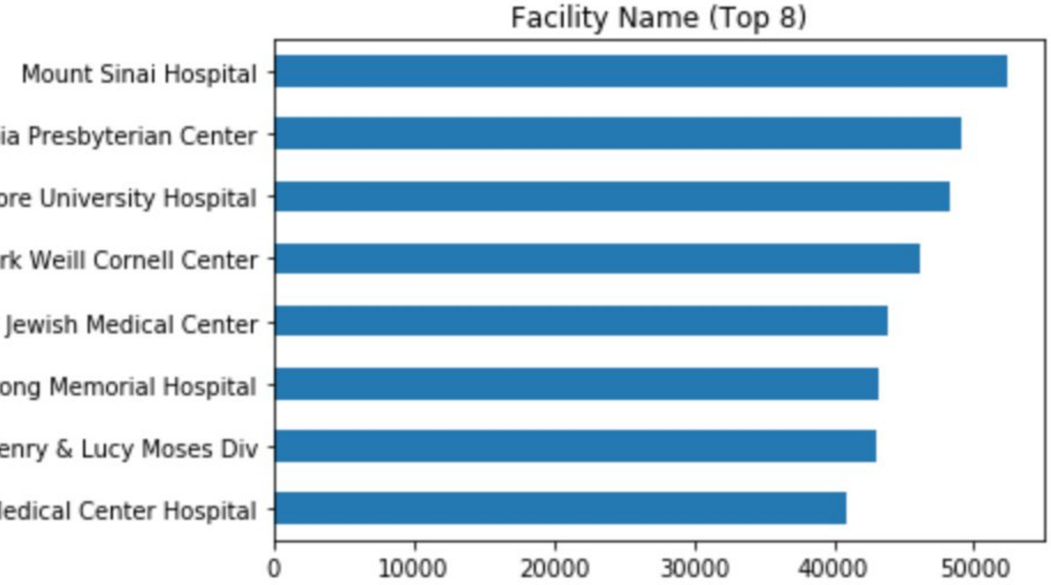
1. Test for systemic bias in healthcare cost based on features (e.g. race, gender) and controlled variables
2. Recommendation system with collaborative filtering, recommending patients with proper facilities (like based on geographical positions, costs, emergence and severity)
3. In idea 2, we may have an advanced version with extra data sources, like specialities, reviews and capacity of hospitals
4. We could do clustering of the cases, especially focusing on geographical locations. Do people in different areas of NY state have similar records? If not, what may be the hidden reasons?
5. Since we have data from 2 years (maybe more?) Could we find some discrepancy between two-year records? If so, why? (The EDA of two years seem similar. How about other years?)

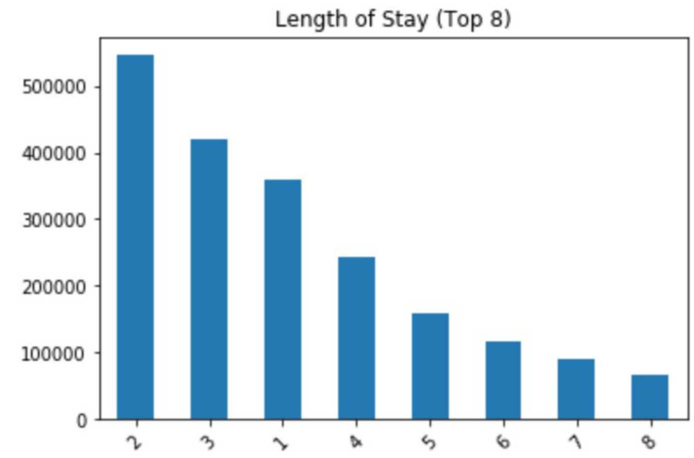
Distribution of records with age, gender and race



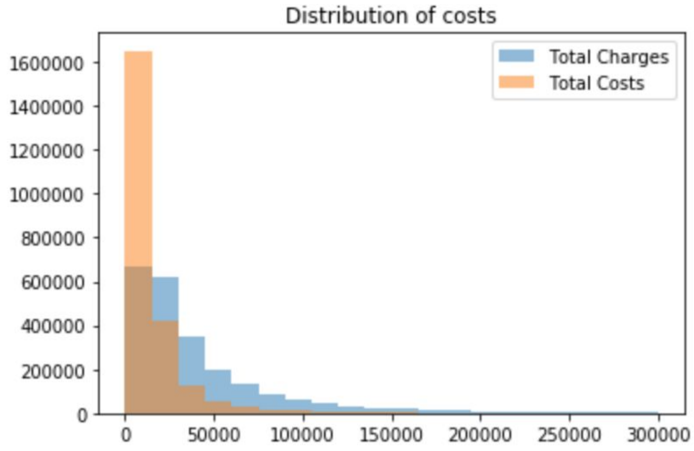
Geographical distribution of records



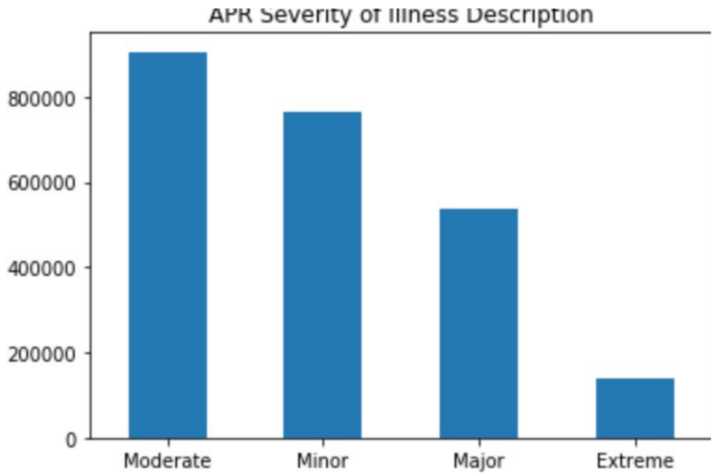




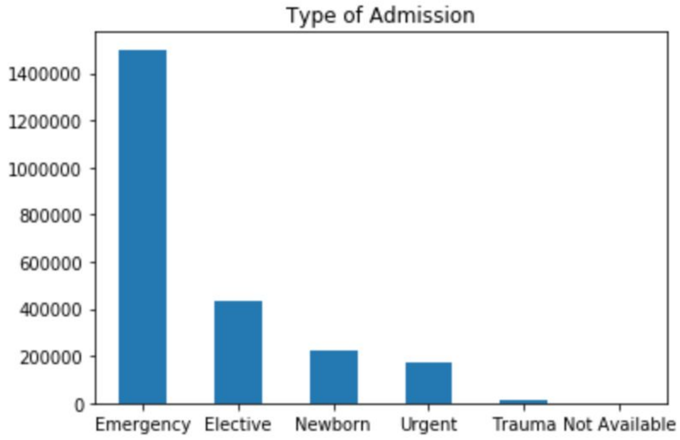
Length of Stay (Top 8)



Distribution of costs



APR Severity of Illness Description



Type of Admission

Data
https://health.data.ny.gov/dataset/Hospital-Inpatient-Discharges-SPARCS-De-Identified/22g3-z7e7/data?fbclid=IwAR2MLdD9ODpn8Kl2evSjoOOhoskv4SNRUNmfX0LRltNu7NaMjc4gVIbC5OM

Overview:
https://health.data.ny.gov/api/views/22g3-z7e7/files/45cf1421-3f5b-4319-8492-d209a7a70e44?download=true&filename=NYSDOH_SPARCS_De-Identified_Overview_2017.pdf

Feature description:
https://health.data.ny.gov/api/views/22g3-z7e7/files/bc44a808-8bab-47d0-ba05-cb7303b78179?download=true&filename=NYSDOH_SPARCS_De-Identified_Data_Dictionary_2017.pdf

https://en.wikipedia.org/wiki/Diagnosis-related_group
http://health.utah.gov/opha/IBIShelp/codes/CCS.htm