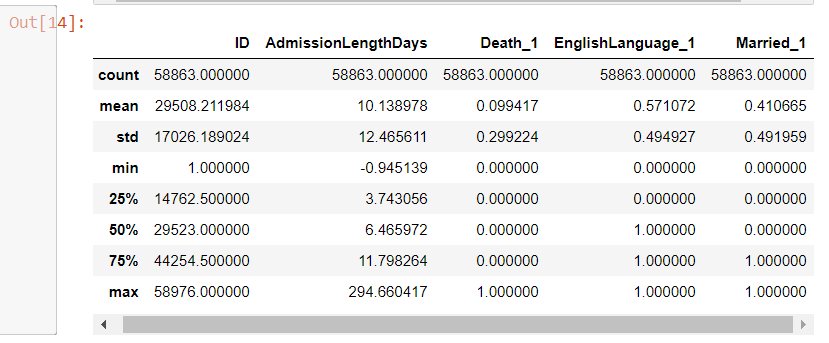
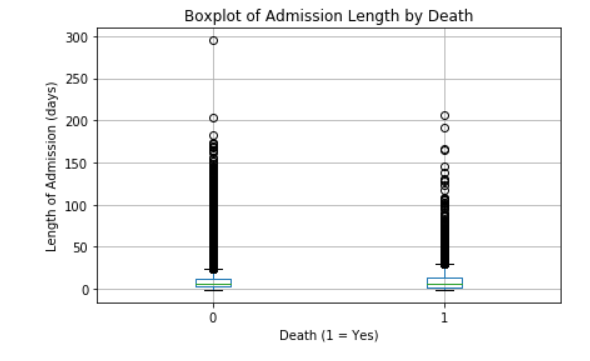
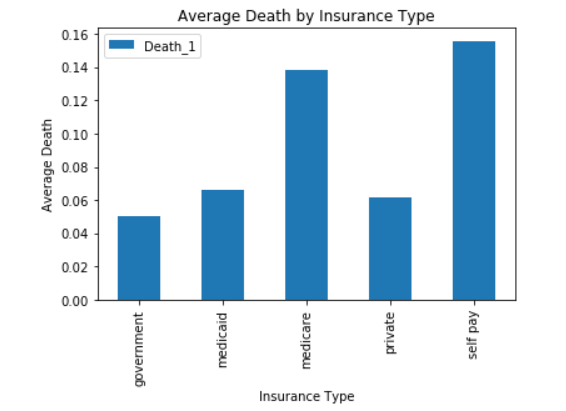
# Exploratory Data Analysis

## Task for evaluating EDA skills

* Import pandas, matplotlib.pyplot, stats from scipy
* Put HospitalAdmissionsData into a dataframe:<https://umich.instructure.com/courses/38100/files/folder/Case_Studies/18_HospitalAdmissions?preview=5005926>
* Answer the following:
  + What are the column headers or features in this data set?
    - ['ID', 'AdmissionLengthDays', 'Death\_1', 'Admission\_Type', 'Insurance\_Type', 'EnglishLanguage\_1', 'Religion\_Type', 'Married\_1', 'Race', 'Dx']
  + How many features are floats? how many are integers?
    - Floats: 1, Int: 4
  + Which features are objects (non-numeric variables)?
    - Admission\_Type, Insurance\_Type, Religion\_Type, Race, Dx
  + What are the possible 'values' for features that are objects?
    - Respectively: elective, medicare, catholic, white, newborn
  + Show the summary statistics for the numerical data
    - 
  + What is the most common admission type, insurance type, religion type, race, and diagnosis (Dx)?
    - Emergency, medicare catholic, white, newborn
  + Show a histogram for admission days (with appropriate labels, titles, etc), and also show a histogram for admission days on a log scale. Describe what you see.
    - The normal histogram shows a unimodal bell curve that is relatively flat centering around 5. The log historgram has a high peak reaching around 12000 and is steep centering around 3.
  + Compare the average length of admission between those who died versus those who did not die. Show a visualization, with appropriate labels, titles, etc.
    - 
  + Describe the association between death and insurance type and show a visualization or test.
    - Government insurance type has the lowest death rate while self pay had the highest.
    - 

Recommended HW (not required):

[More on Categorical Variables](https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9)