***Project:*** Analysis of Disease data set

***Author:***Sekhar Mekala

In this project I will be analyzing diseases data set present at (which is available at data.gov - <http://catalog.data.gov/dataset/infectious-disease-cases-by-county-year-and-sex>). This data set contains data about various diseases occurred in the state of California from 2001 to 2014. This data set has the following variables:

*Disease –* Name of the disease

*County –* County name

*Year* –Calendar Year

*Sex –* Gender (Male/Female)

*Count –* Number of people effected by the disease

*Population –* Total population

*Rate –* Computed as *(count/population multiplied by 105)*

The complete data analysis will be presented as IPython Notebook.

**Here are the project requirements:**

Ignore the county variable, and perform the following analysis (to get an understanding about the diseases trend). We will confine our analysis to the diseases that have effected more than a specific number (for example, 500 people in the state):

Most of our analysis is driven by graphical plots, and drawing conclusions about the diseases trend and any other hidden patterns.

NOTE: In the following requirements, whenever I mention the diseases, I am referring to the diseases that have effected at least a predefined number of people in a year. This predefined number will be determined later, during the implementation.

**Requirement 1:**

**Year/Disease analysis:**

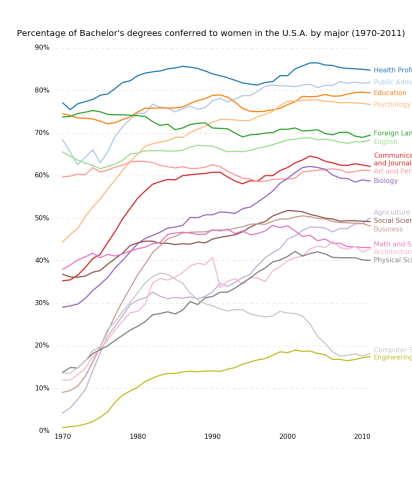
**Plot 1:** Draw a trend line of all the diseases occurred every year. In a single graph, we should plot all the diseases trends over years. Using this graph, we can get the following information:

1. Which diseases have occurred the most among the population? Is that trend persistent over the years?

2. Would the percentage of people effected by the diseases change each year?

3. Is there any pattern in the diseases occurring together/not together?

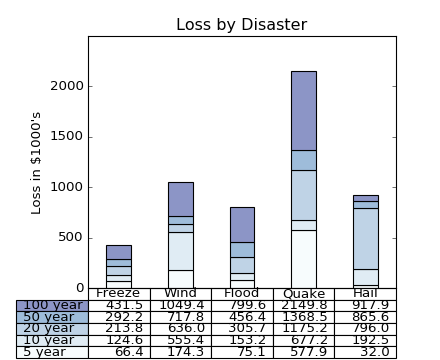
The graph should look like something similar to:



**Requirement 2:**

**Plot 2:** Draw a segmented bar plot (standardized and non-standardized) for each year, in such a way that the bars are broken down by disease. These plots will help us to identify which years have the most/least number of diseases registered. Is there any decrease/increase in the overall disease rate over the years? Do you see any trend in specific diseases occurring over the years?

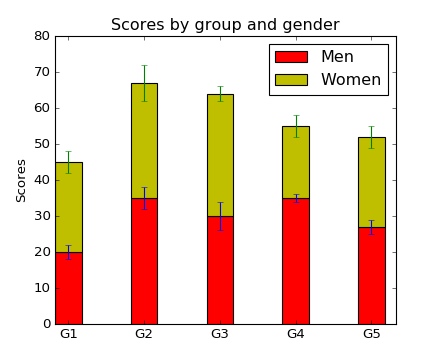
Something similar to:



**Requirement 3:**

**Gender/Disease analysis:**

Plot 3A: Draw a segmented bar plot (non-standardized), for each disease, in such a way that the bars are further broken down by gender (stacked bar plot). This will help us to identify which disease(s) have effected the most population over the years, and which have effected the least.



Plot 3B: Draw a segmented bar plot (standardized), for each disease, in such a way that the bars are further broken down by gender (stacked bar plot). This plot will help us to see if there is any trend in diseases effecting a specific gender of the population.

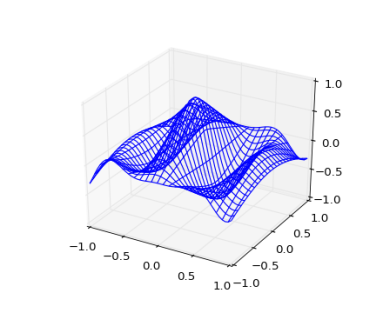
*Same as above plot, but all bars will be of the same length.*

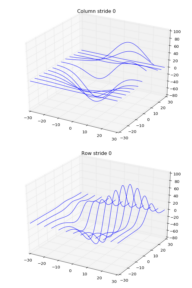
**Requirement 4:**

**County/Disease/Year analysis:**

Draw a 3D plot with County and Year variables on the X and Y axis, and Z axis representing the disease count. This plot will help us to visually see if diseases effect more percentage of people in some counties, and less at other counties.

Plot something similar to:

or

[](http://matplotlib.org/examples/mplot3d/wire3d_zero_stride.html)

Based on the number of diseases we study (based on a predefined number of people who are effected), draw bar charts, if it helps to find association between counties and diseases, in each year.

**Requirement 5:**

Use Monte Carlo simulation to predict the maximum percentage of people effected by various diseases, in each county, over the next 3 years. Print the data something similar to the following. If it is too complex to run the simulation for each county and for each disease, at least get the predictions at the state level.:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year: 2015 | | | | | | | | | |
| County | Disease1 | Disease2 | Disease3 | . | . | . | . | . | . |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year: 2016 | | | | | | | | | |
| County | Disease1 | Disease2 | Disease3 | . | . | . | . | . | . |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year: 2017 | | | | | | | | | |
| County | Disease1 | Disease2 | Disease3 | . | . | . | . | . | . |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |