Problem Set #4

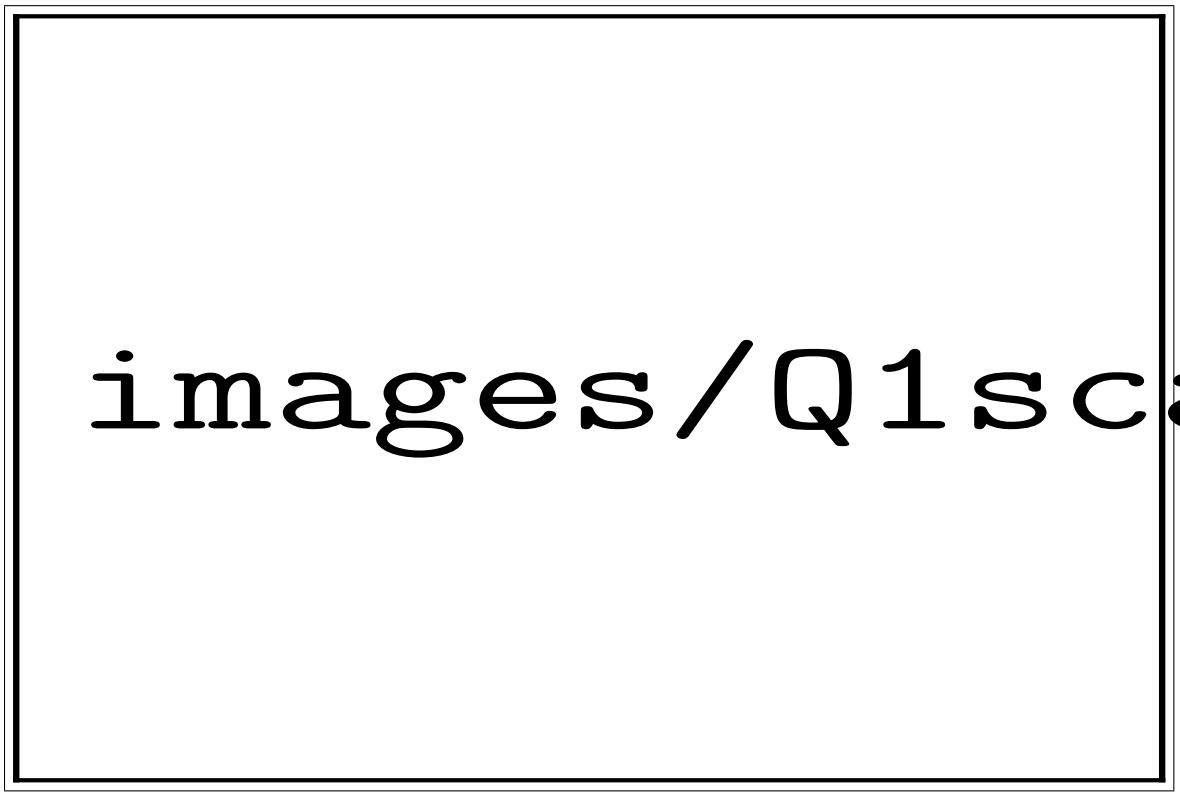
MACS 30000, Dr. Evans
Cheng Yee Lim

Problem 1

Part (a). I created a Kaggle account, username chengyeelim.

Part (b). Describe a Kaggle Competition of interest to you. Outbrain is holding the Outbrain Click Prediction Competition to search for algorithms that can better predict recommended content each user will click. With the slew of information available on the web, Outbrain aims to serve as the internets leading content discovery platform and pairs readers with relevant content. Every month, Outbrain has provided 250 billion personalized recommendations from thousands of sites for their users. Thus, the recommendation algorithms are at the heart of their business, which they hope to augment with the brilliant minds on Kaggle.

Figure 1: Scatter plot of lifetime temperature of Ricardo



Part (a).

Using the datetime function, I created a new index that corresponds to the month and day of a date separately for leap years and non-leap years i.e. the index starts at 265 and ends at 630 for leap years and starts at 264, and ends at 628 for non-leap years. This would allow us to plot the x-axis as one complete year, with several years of months and dates.

Part (b).

I then merged the high and low temperature of cities by appending the axes to form a dataframe of all temperatures for its corresponding monthday. This dataframe was used as the y variable of the scatter plot.

Part (c).

I set my marker color as black and size as s=0.01 for the general markers.

Part (d).

I set my marker color as maroon and size as s=0.01 for the Ricardo's Illinois markers.

Part (e).

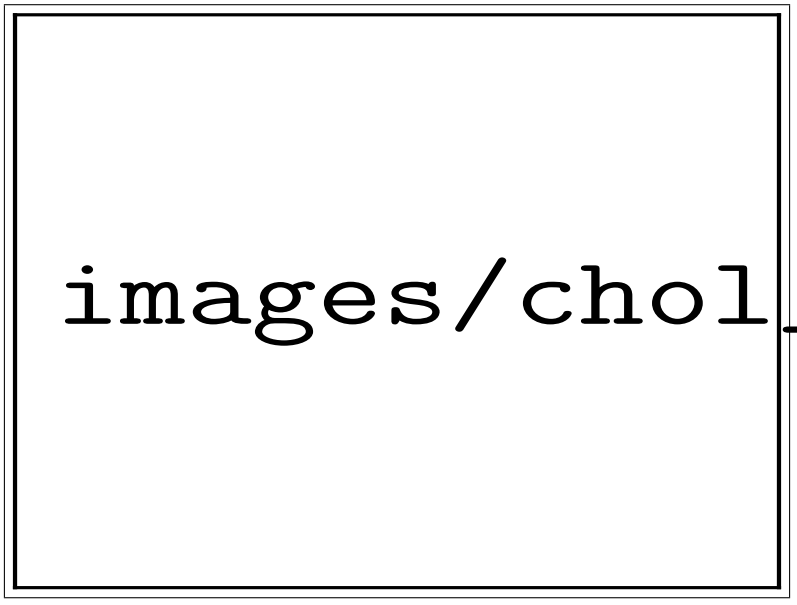
I highlighted life events of "born" and "little league all-star team wins regional championship" by setting the marker color as yellow, and adding a black bold border around the marker. Using the annotate function, I then labeled these datapoints with descriptive text.

Part (f).

I then labeled the x-axis evenly spaced as 'Autumn', 'Winter', 'Spring', 'Summer'. The y-axis was labeled as temperature.

Problem 2
Part (a).

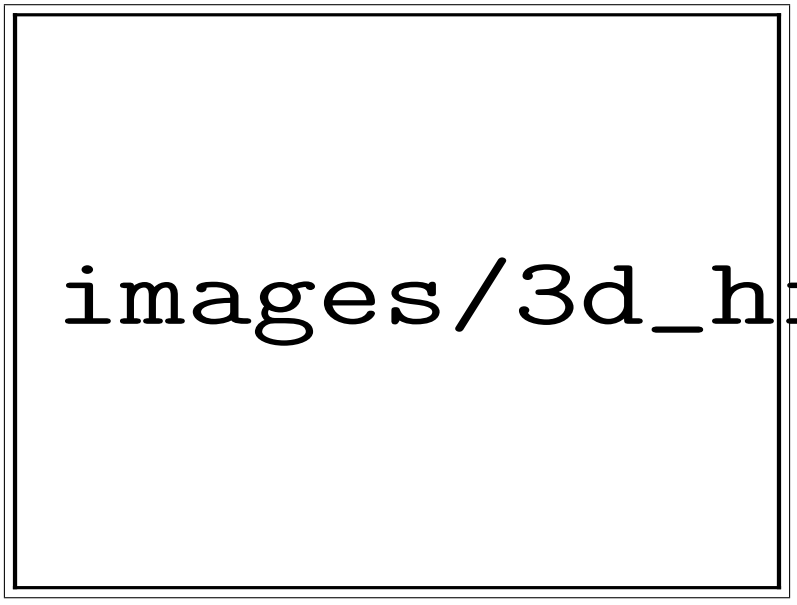
Figure 2: 2D frequency histogram of concentration of plasma cholesterol (mg/dl)



I extracted the information on concentration of plasma cholesterol from the lipids.csv file and plotted a 2D frequency histogram with 25 equally spaced bins. The midpoint of the bin with the highest frequency is 237.50.

Part (b).

Figure 3: 3D frequency histogram of concentration of plasma triglycerides (mg/dl) and cholesterol (mg/dl)



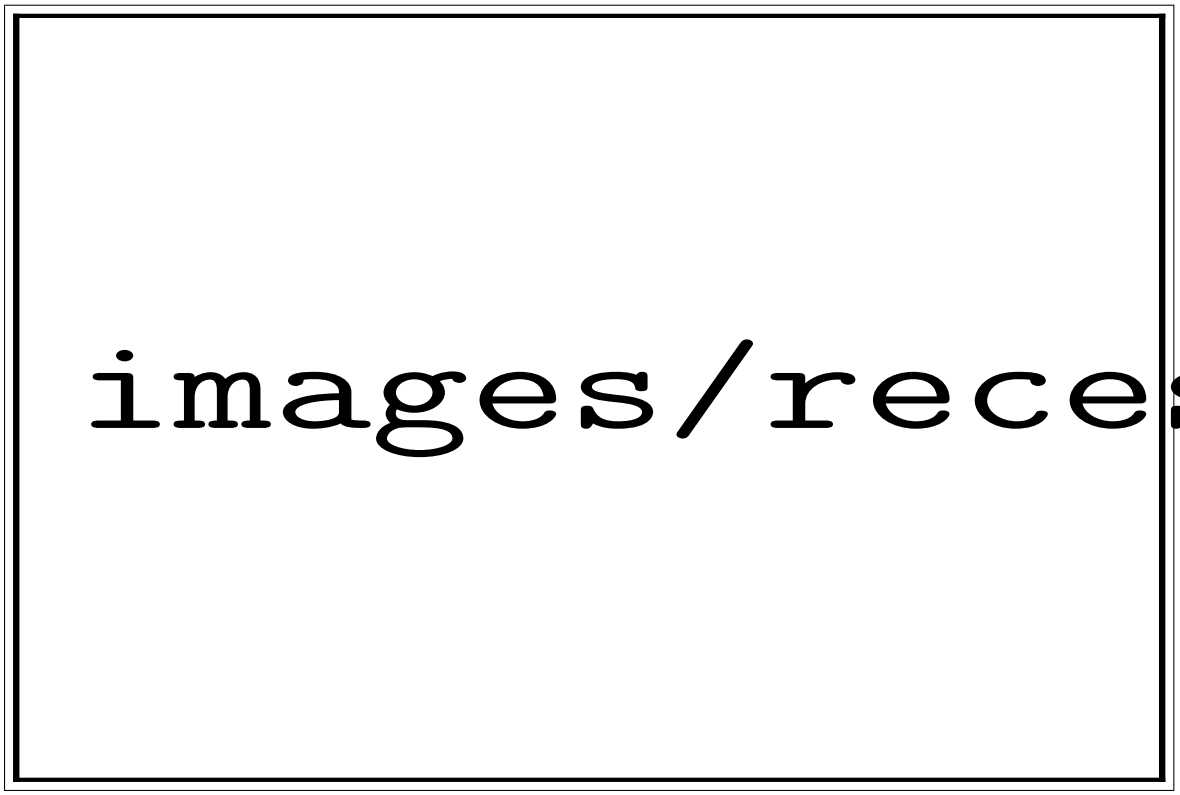
The key new finding that emerged from the data is that individuals who have low concentration of plasma triglycerides are more likely to have low concentration of plasma cholesterol. However, individuals generally have higher concentration of plasma cholesterol than plasma triglycerides.

Part (c).

Individuals with concentration of plasma triglycerides and plasma cholesterol of higher than 300mg/dl have the highest risk for heart disease.

Problem 3

Figure 4: Normalized peak plot of job growth in the past 14 recessions in the US



Part (a).

I created 14 segments of job growth data series by selecting the dates in the indexes for 14 recessions since 1929.

Part (b).

I normalized each of the 14 series such that the jobs level at the peak date equals to 1.

Part (c)(d)(e)(f)(g)

The 14 series were plotted as a line graph with normalized jobs against time from the peak to show the percent change from the peak jobs level. Each line plot has a different combination of linestyle and color and a legend is made outside of the axis on the right that states the beginning date of each recession. The y-axis is then labelled as "Jobs/Peak" and the x-axis is labelled as "Time from Peak", with 9 markers: -1yr, Peak, +1yr, +2yr, +3yr, +4yr, +5yr, +6yr, +7yr.

Part (h).

A dashed grey horizontal line at Jobs/Peak =1 and a dashed grey vertical line at Peak were drawn, to act as a reference point in interpreting the graph.

Part (i).

The line plot for the Great Depression is black and solid and the line plot for the Great Recession is red and solid. The line plots are also thicker than the other 12 plots.

Part (j).

The recession in 1945 February was worse than the Great Recession in terms of the percent of jobs loss, but the recession in 1945 February experienced a much faster job recovery rate.

Part (k).

There are no ways in which the Great Recession has been worse than the Great Depression in the United States, in terms of jobs losses.