**Stat291 Lab 4**

**The box office earning data (movie\_boxoffice.csv) contains the following variables for movies released from 1980 to spring 2018:**

* Move: name of the movie
* Month, Day, Year: the releasing date of the movie
* Budget: budget of the movie (in Millions)
* Domestic\_Gross: the domestic box office earning (in Millions)
* Worldwide\_Gross: the worldwide (or global) box office earning (in Millions)

**Use set.seed() and distinct() to set random seed and get rid of duplicates before analysis**

1. How many duplicated records were removed from boxoffice data? What is your random seed number?

**Part I: Sampling**

1. Treat the box office data as your population. Have a histogram of the global box office earning. Describe the shape of the distribution
2. In the population, what is the average global box office earning? What is the standard deviation? What is the proportion of movies whose global box office earning exceeds budget?
3. Take a random sample of 200 movies from the population, get the histogram of global box office earning, and describe the shape of the distribution. You are going to use this sample in part II and future homework.
4. In your sample, what is the average global box office earning? What is the standard deviation? What is the proportion of movies whose global box office earning exceeds budget? Are these summary statistics from the sample close to those population parameters?
5. Take a random sample of n movies from the population, calculate the average global box office earning, and the proportion of movies whose global box office earning exceeds budget. Repeat 500 times. (Do this step for n=20, 50, 100, 200 respectively)
6. For each n=20, 50, 100 and 200, get the histogram of the average global box office earning. Have histograms using facet\_wrap() so the four histograms are in the same picture for easy comparison. Also get the mean and standard error of the average global box office earning.
7. Compare the distributions among different n values, and also compare them to the distribution from the population in Q1.
8. For each n=20, 50, 100 and 200, get the histogram of the proportion of movies whose global box office earning exceeds budget. Have histograms using facet\_wrap() so the four histograms are in the same picture for easy comparison. Also get the mean and standard error of the proportion of movies whose global box office earnings exceed budget.
9. Compare the distributions among different n values

**Part II: Bootstrapping**

1. Use data from Q3 as the initial sample, use bootstrapping method to resample once with 200 movies. Are there any duplicated movies in your bootstrap sample? Is this expected or something is wrong?
2. Get the histogram of global box office earing in the bootstrap sample. Describe the shape of the distribution, and compare it to Q3
3. In the bootstrap sample in Q10, what is the average global box office earning? What is the standard deviation of it? What is the proportion of movies whose global box office earning exceeds budget? Are they close enough to those in the initial sample?
4. Get the bootstrapping distribution of the average global box office earning, and the proportion of movies whose global box office earning exceeds budget, by resampling 500 times with bootstrapping method. Get the mean and standard error of the average global box office earning. Get the mean and standard error of the proportion of movies whose global box office earning exceeds budget
5. Have the histogram of the sample mean and sample proportion from Q13. Have histogram using facet\_wrap() to make comparison of the bootstrap distribution with those the in Q6 and Q8 when n=200. Describe the comparison.