**Question 1**

Use data from **Returns.csv** to answer this question. The dataset contains 5 daily stock returns (in terms of percentage) from 2016 to 2017, including Apple (AAPL), Amazon (AMZN), Facebook (FB), Google (GOOG), and Microsoft (MSFT). For example, on 05-January-2016, Apple's stock dropped by 2.51% from the previous trading day.

1. Obtain point and conventional 95% interval estimations of the mean return of each stock. Present the formulae used for the interval estimation and comment on the results.

2. Obtain point and conventional 95% interval estimations of the proportion of each stock having daily positive returns. Present the formulae used for the interval estimation and comment on the results.

3. Plot a heatmap with hierarchical clustering based on the pairwise correlation among the five stock returns. Comment on the results.

**Question 2**

Use the file **movies.xlsx** to answer this question. The file contains information of 190 movies, including two variables: Gross (Gross Revenue) and Budget (Spending), both are in millions (USD).

1. Classify movies into two groups: money-making (Gross >= Budget) and money-losing (Gross < Budget). Test if the following claim is true: "more movies are money-losing than money-making." Use the 6 steps of hypothesis testing to report the testing outcomes.

2. Estimate a simple linear regression model of 'Gross' on 'Budget'. Interpret the estimation outcomes (in terms of parameter significance, fitness, and the meaning of the regression slope & intercept).

3. Make a scatterplot of Gross on Budget and identify a clear outlier from the plot (find the outlier movie). Estimate a simple regression as in Part 2 but without the outlier. Compare and comment on the regression results with and without the outlier.

**Question 3**

The file **Party.csv** contains data on a sample of 250 voters with tracked variables, including party preference (Party=1 or 0), Age, Female (gender), Married (marital status), Income (in thousands), Education (schooling years), and Religion (Religion=1: religious, and 0: non-religious).

1. Estimate a logistic regression of Party on Age, Female, Married, Income, Education, and Religion with statsmodels. Discuss the significance of each coefficient & model fitness.

2. Based on the results of Part 1, build the confusion matrix with (in-sample) prediction. Compute and discuss the predication accuracy, precision, and recall.

3. Based on the results of Part 1, construct two groups of voters: Group A is formed by voters with over **75%** of predicted probability to vote for Part 1 and Group B is formed by voters with over **75%** of predicted probability to vote for Part 0. How many voters are in Group A? How many voters are in Group B? Find the 90% confidence interval of the mean income of these two group of voters. Comment on the results.

4. Perform KMeans clustering with Age, Female, Income, Education, and Religion and use the Elbow curve to justify the optimal number of clusters is 3. Form 3 clusters and use the crosstab to check if the clustering outcome reflects party preference. Comment on the results.

**Question 4**

Use the dataset **billionaire.xlsx** to answer the following questions. The dataset contains the following variables: Nation (country name), Number (number of billionaire), GDP (in billions USD), and Population (in millions).

1. Obtain the median of GDP per capita: GDP\_pc = GDP/Population, and use the median of GDP\_pc to make countries into two groups: "Rich" (countries with GDP per capita above the median) and "Not\_Rich" (countries with GDP per capita equal to or below the median). Find the mean number of billionaires for each group & use the barplot of Seaborn to plot these two means. Comment on the results and discuss if the two CI's are "symmetric around the mean value".

2. Let the mean number of billionaires of the Rich group as "meanN-Rich" and the mean number of the Not\_Rich group as "meanN-Not\_Rich". Show that none of the following two null hypotheses (i) meanN-Rich = meanN-Not\_Rich (ii) meanN-Rich = 2 \* meanN-Not\_Rich can be rejected at 10% significance level. Discuss what might contribute to the non-rejections.

3. Obtain two scatter plots with fitted lines: Number vs. GDP and Number vs. GDP per capita. Test if (i) Number and GDP are correlated and (ii) Number and GDP per capita are correlated. Comment on the results.

4. Estimate the following two multiple regressions: (i) Number on GDP and Population (ii) Number on GDP per capita and Population. Compare and comment on the results of the two regressions.

5. Use the scatterplot to visualize the fitness of two regression predictions from Part 4 (as in Topic 8). Discuss the role of the United States in the fitness.