Department of Statistics, The Chinese University of Hong Kong

STAT 5106, Programming Techniques for Data Science (Term 1, 2024-25)

ASSIGNMENT 4

(Deadline: 5 Dec 2024, 2359)

*[Note: Although the assignment is about Python script, we still welcome to use R script for handling those non-Datacamp questions. But please make sure the scripts are runnable and give the correct answers.]*

Q1. (35%)

[Big Mac Index]

The [Big Mac Index](https://www.economist.com/big-mac-index) is published by The Economist as an informal way of measuring the purchasing power parity (PPP) between two currencies and provides a test of the extent to which market exchange rates result in goods costing the same in different countries. It "seeks to make exchange-rate theory a bit more digestible."



The full algorithm any interested can refer to [github](https://github.com/TheEconomist/big-mac-data).

In this assignment we aim to apply the graphic package with utilizing the data.

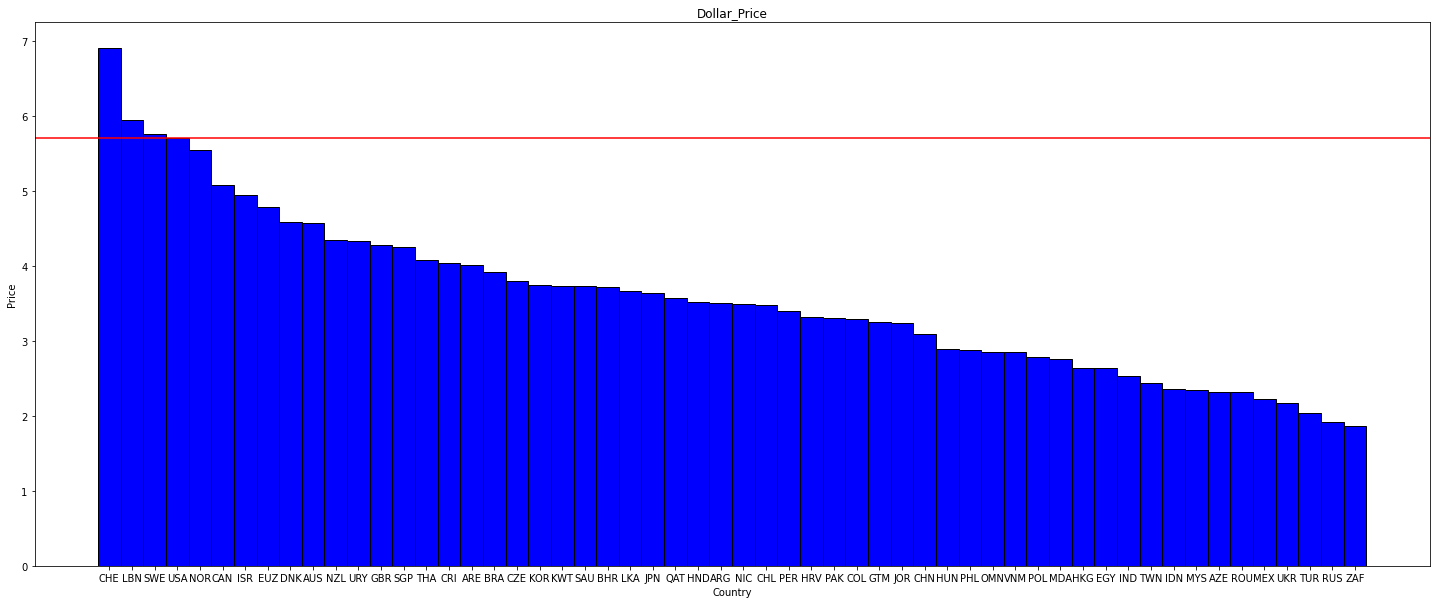
The zipped dataset can be obtained from the following URL:

<https://drive.google.com/uc?id=1LpH6AGIZn_h3eX_up_RwSC8Wc3BDTg2->

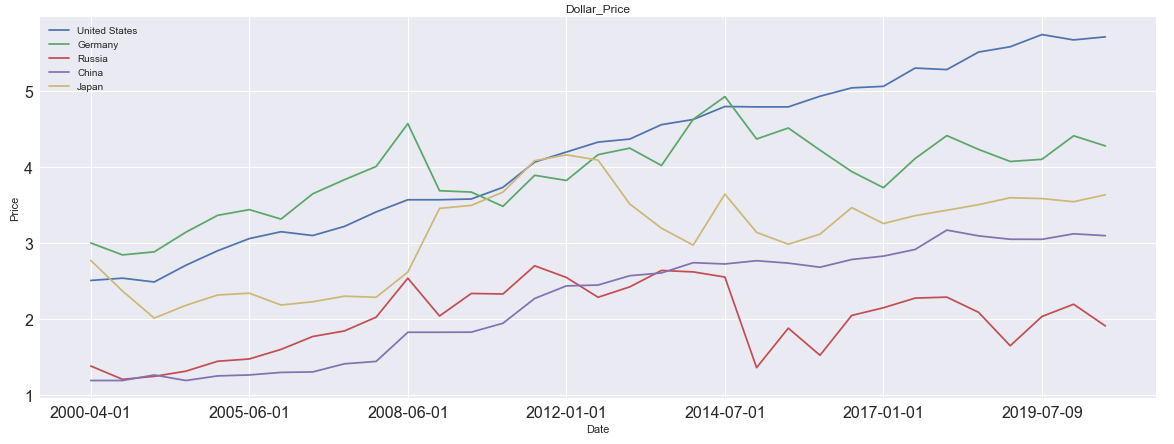
1. Use python code to unzip and import the dataset.

Please work the following statistical charts, using matplotlib or seaborn.  
(of course you may need to pre-process the data before plotting !)

1. draw a bar chart of dollar prices against different countries, of the **most recent date - 1 Jul 2020**, with
   1. ordering the bars with descending order, and
   2. draw a horizontal reference line of USA - how much for a big mac in USA.



1. draw a time-series line plot of dollar prices against various dates
   1. with 5 countries only: United States, Euro Area, Russia, China, Japan
   2. create a legend showing the colors for 5 countries.

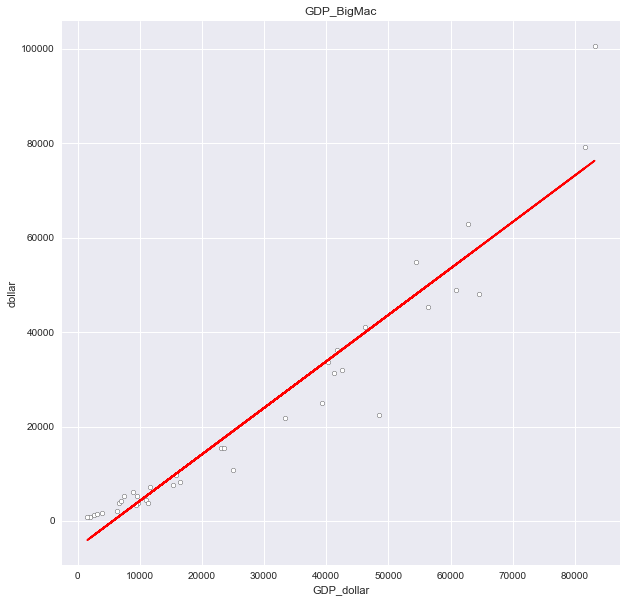


1. draw a scatter plot of GDP big mac adjusted, i.e.

where dollarPriceUSA = A big mac price in USA, in US dollar  
against dollar prices, at **1 Jul 2020** , with

1. with a regression line over the scatter plot.

Note the big mac index = the prediction of the regression line.



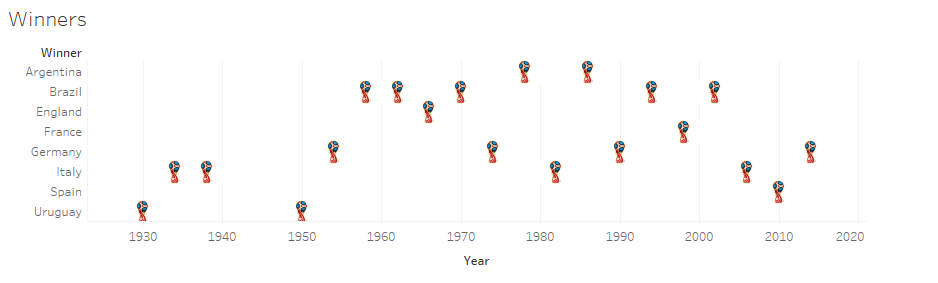
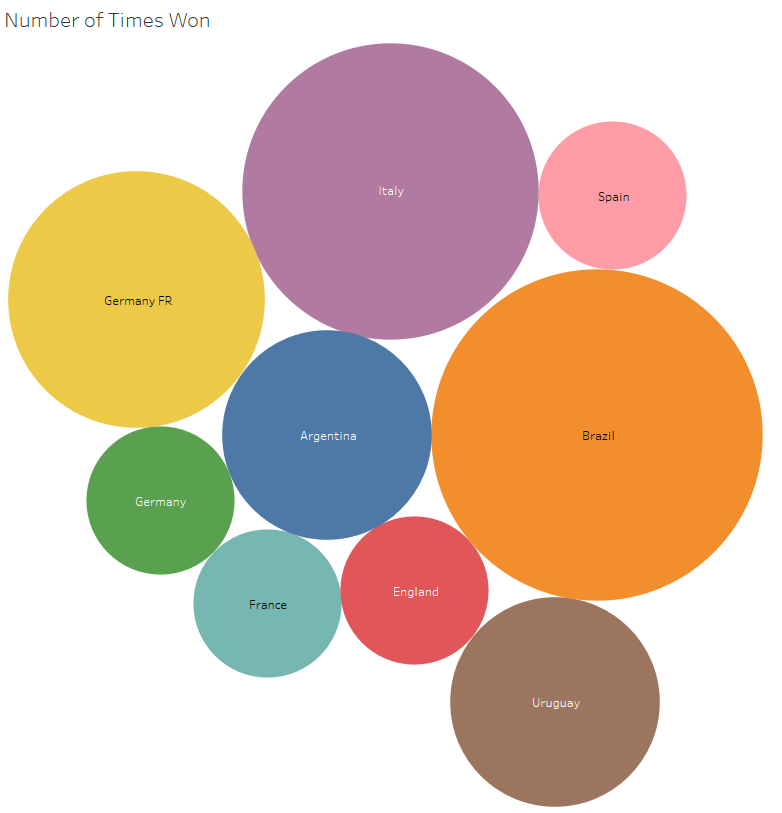
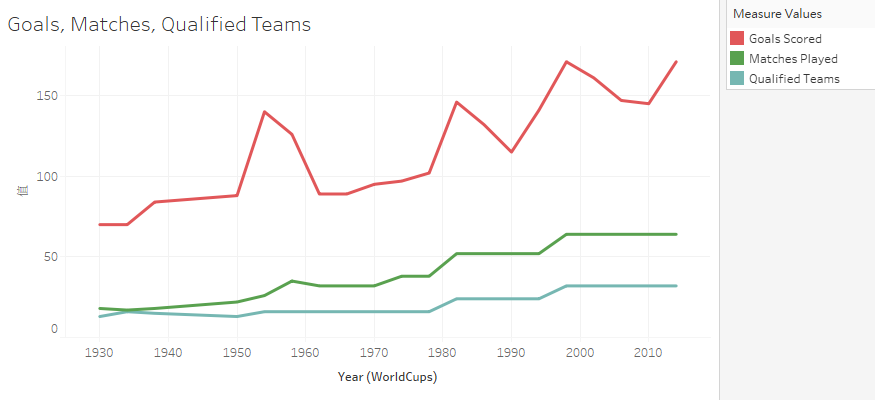
Q2. (25%)  
(Please use **Tableau Public** for completing this question)

[Fifa World Cup]



The [FIFA World Cup](https://www.fifa.com/tournaments/mens/worldcup/qatar2022) ([wiki](https://en.wikipedia.org/wiki/FIFA_World_Cup)) is the biggest competition in the world, played every four years by the top 32 national teams in a month-long tournament. It’s a grand congregation and celebration of people of various countries, cultures, tribes, races. It’s a time to cry over a loss, cheer for a win and hope for a draw for your team. This question provides a great opportunity to use Tableau to explore the history of the World Cup.

Here are the list of instruction completing this:

1. Please download and install Tableau Public 2024.3 with the URL: <https://public.tableau.com/en-us/s/>   
   (Acceptable if the version is workable. But I highly encourage you download the most updated version)
2. Simultaneously sign-up and create your own account for Tableau Public.
3. The dataset can be accessed through here - [a google spreadsheet](https://docs.google.com/spreadsheets/d/1NijzluWXuBUu9qtN7qDzYPMZDTtLoNJXWX-QpKDOaas/edit?usp=sharing) .  
   Please download and place in your google drive or local environment.
4. Open using Tableau Public 2024.3, and make your dataset downloaded in google spreadsheet, so this can be connected to Tableau Public.
5. Include the following plots in your Tableau Public workbook:
   1. Winners against year.  
        
      (Note: Welcome to try any custom icons. Not restricted to the world cup.   
      How to make this ? Please check the [tableau site](https://www.tableau.com/drive/custom-shapes).)
   2. No. of times won.  
      
   3. Goals, Matches, and Qualified Teams  
        
      (Hint: How to put measure values to the same chart ? Please refer to this [tableau site](https://help.tableau.com/current/pro/desktop/en-us/datafields_understanddatawindow_meavalues.htm))
6. Save to Tableau Public, using your account created in Step 2.
7. In your setting page, please set as thisGraphical user interface, text, application

   Description automatically generated
8. Finally please include the **URL published in your submission**.

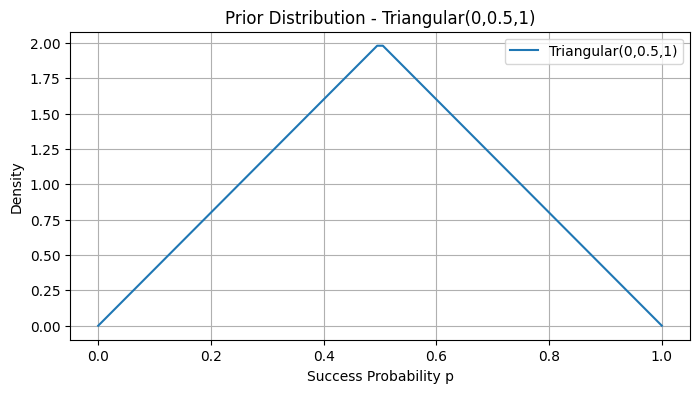
Q3. (40%)

[Medical Application of Bayesian Inference]

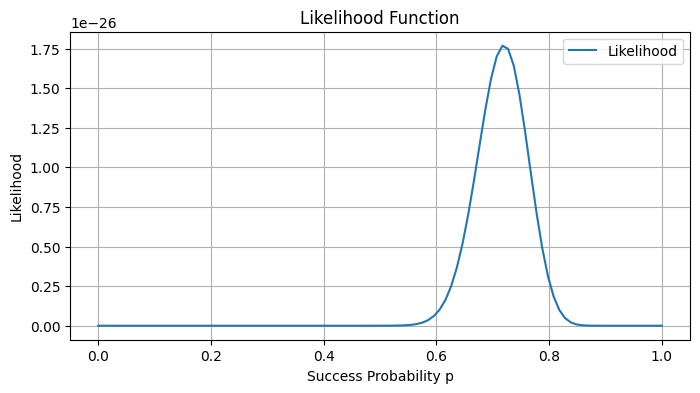


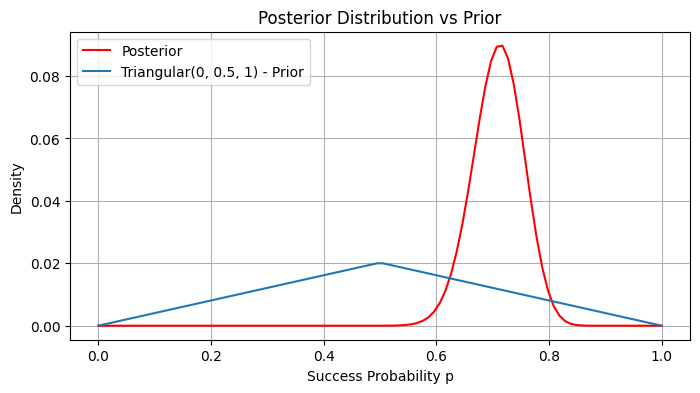
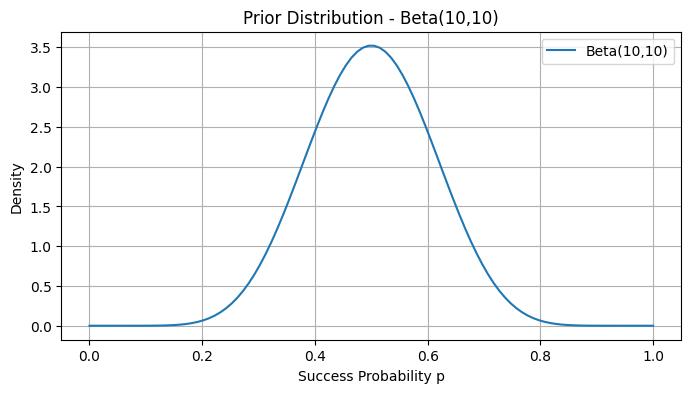
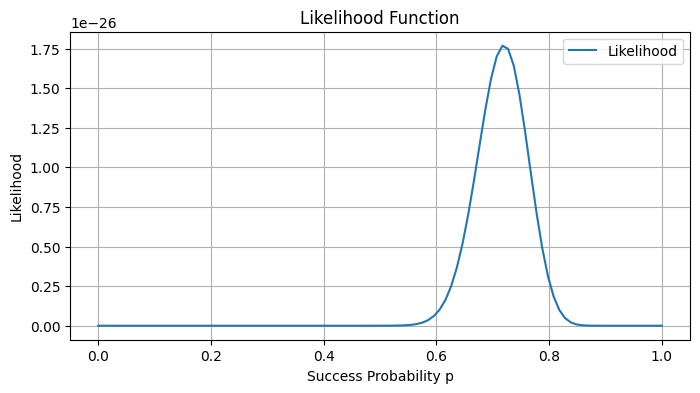
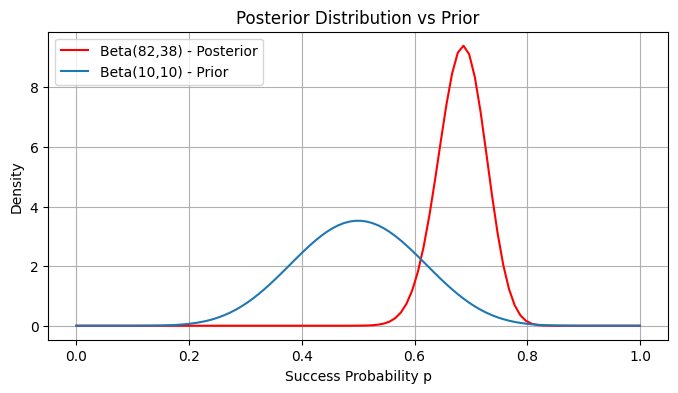
In this question, you will apply Bayesian statistical inference to analyze a dataset concerning the success rates of a new treatment. The goal is to estimate the posterior distribution of the success probability of the treatment given the observed data and prior beliefs, even if we have only a single data.

1. **Define the Prior Distribution:**Assume that the prior distribution for the success probability of the treatment is a Triangular distribution, . Choose , and as the shape parameters for the prior distribution.  
     
   Note: [Numpy Triangular Random Number](https://numpy.org/doc/stable/reference/random/generated/numpy.random.triangular.html) and [Scipy Triangular Distribution](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.triang.html) are different sets of parameters. Hints to transform by [this](https://stackoverflow.com/questions/31745934/drawing-from-a-continuous-triangular-distribution-in-python-scipy).

Plot the prior distribution using Python, as the following  


1. **Compute the likelihood Function:**Given that the data shows 72 successes out of 100 trials, assume a binomial likelihood function.

Write a Python function to compute the likelihood of observing the data for different values of ranging from 0 to 1. Plot this likelihood function as the following.  
  


1. **Compute and Plot the Posterior Distribution:**Update the parameters of the prior to get the parameters of the posterior distribution. Plot the posterior distribution and compare it with the prior.  
     
   
2. **Analyze the Posterior:**
   1. Calculate and plot the cumulative distribution function (CDF) of the posterior to find the probability that the success rate is greater than 0.5.
   2. Calculate the 95% credible interval for **.**
3. **Bayesian Influence with Conjugate Prior**
   1. Repeat tasks in a, b, c, but assume the prior to be . Then you can simplify the analysis with conjugate prior property.  
        
        
      
   2. Discuss how the results would change if the prior parameters were different(e.g., and )