CARNEGIE MELLON UNIVERSITY - AFRICA

# DATA, INFERENCE & APPLIED MACHINE LEARNING

(COURSE 18-785)

Professor Patrick McSharry

**Assignment 3**

MUNYANEZA Kenny Roger

2nd October 2023

# Introduction

This is a final report designed as a fulfillment of the requirements to complete assignment 3 DIAML.

## Tools used

- Python
- Vs Code

# Question 1

## Third party libraries used

- **pandas**: It was used to formulate a dataframe out of the 11 observations of Daily energy intake in kJ measured for 11 women.
- **scipy**: Used the stats submodule to compute the p-value of the t-distribution.
- **numpy:** Used to compute mathematical operations such square root.
- **prettytable:** Used to display all the result of the hypothesis test in a table.

## Implementation process used

I designed hypothesis test for the actual value of the distribution mean of the 11 observations of Daily energy intake in kJ measured for 11 women. I set the **null hypothesis to be that** "*The mean of the distribution is 7725 and there's no systematic deviation from the recommended value*".
I set the **alternative hypothesis to be that** "There's a systematic deviation from the recommended value and the mean of the distribution is not 7725". Since the alternative hypothesis uses the **not equal** term to compare the recommended value of the mean and the actual mean, the **two tailed hypothesis** was used in this case. This means that the sample mean may be greater than or less than the recommended value. Then, I created a dataframe from the 11 provided observations using the pandas library series object and I computed the summary statistics of the sample, namely the sample mean using mean() function of pandas which is equal to 6753.63, the Standard deviation using the std() of pandas which is equal to 1142.12, the Standard Error of the mean by dividing the standard deviation of the distribution by the square root of the sample and I obtained the standard error of the mean equal to 344.36. Next, I calculated the degree of freedom which is given by subtracting one from the sample size and I calculated the t-statistic by dividing the difference between the recommended mean and the actual mean of the distribution by the standard error of the mean. Lastly, I calculated the p-value to determine the feasibility of the null hypothesis. I obtained the p-value using the " **stats.t.cdf "** function from the stats sub-module of scipy which takes the t-statistics and the degree of freedom. This is a function that calculates the Cumulative distribution function of the student's t distribution. It depends on the degree of freedom and the t-statistic value. [1]

## Results

```
Null hypothesis: The mean of the distribution is 7725 and there's no systematic deviation from the recommended mean


Alternative hypothesis: There's a systematic deviation from the recommended mean and the mean of the distribution is not 7725.


Form of Alternative test: A two tailed test

+-----------------------------------------+
| Hypothesis testing of the t-distribution |
+--------------------+--------------------+
|       Metric       |        Value       |
+--------------------+--------------------+
|        Mean        |  6753.636363636364  |
| Standard Deviation |  1142.1232221373727 |
|   Standard Error   |  344.3631083801271  |
|    t-statistic     | -2.8207540608310198 |
| Degree of Freedom  |          10         |
|      P-value       | 0.01813723517610577 |
+--------------------+--------------------+

Since the p-value is less than the significance level (0.05), the probability of the mean of the distribution being 7725 is close to 0 hence the null hypothesis can be rejected
```

## Conclusion

There is less than 5% probability that the mean of the distribution is 7725 kJ since the p-value 0.018 is less that the significance level 0.05. Therefore, the null hypothesis is rejected in this case.

# Question 2

## Third party libraries used

- **scipy**: Used the stats submodule to compute the p-value of the t-distribution.
- **math:** Used to compute mathematical operations such square root and exponents.

## Implementation Process

I conducted a hypothesis test to determine whether the Guiness served in Irish pubs tastes significantly better than pints served elsewhere around the globe. I set the **null hypothesis** to be that the pints in Irish pubs have similar test to pints served elsewhere and the two population mean values are equal, and the **alternative hypothesis** to be that Guiness served in Irish pubs taste differently from pints served elsewhere around the world and the two population mean values are not equal. Since we are testing whether the means of the two unknown populations of the given two samples are equal or not, the appropriate test used in this case is **two-sample t-test**. Also, since “not equal” is being used to compare the two population means the test required in this case is a **two-tailed** test whereby the first mean value is greater than the second mean value or vice-versa.[2] Next, I proceeded to getting the evidence that allows or rejects the null hypothesis. Since the standard deviations of the two samples provided are different, the variances are different as well. Therefore, I used the **Welch-Satterthwaite equation** to calculate the degree of freedom for the two samples with unequal variance. [3]

Welch's equation:

$$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

Lastly, I calculated the t-static and the p-value using a function given by stats module called **stats.ttest_ind_from_stats** that takes the means, standard deviations, and sample sizes of the two samples as parameter and returns the t-statistic and the p-value. [4]

## Results observed

```
Null Hypothesis: The difference between the two distribution means is due to random variation
                and is not significant hence the pints in Irish pubs have similar test to pints served elsewhere.


Alternative Hypothesis: The difference between the two distribution means is significant hence the pints in Irish pubs taste differently from pints served elsewhere.


Type of t-test used: Two-sample t-test


Form of alternative test: Two-tailed test


Degree of freedom = 85.87168862441837


t-statistic = 11.73775770205081

p-value =  0.0

Based on the p-value obtained less than 0.05, we reject the hypothesis that The difference between the two distribution means is due to random variation
and is not significant. Therefore the means of the two populations are different.
```

## Conclusion / Interpretation

There p-value obtained is 0.0 which is below the significance level 0.05, hence the hypothesis that the mean values of the two populations are equal is rejected. This also means that the taste of Guiness in Irish pubs differs from the pint served elsewhere around the globe.

# Question 3

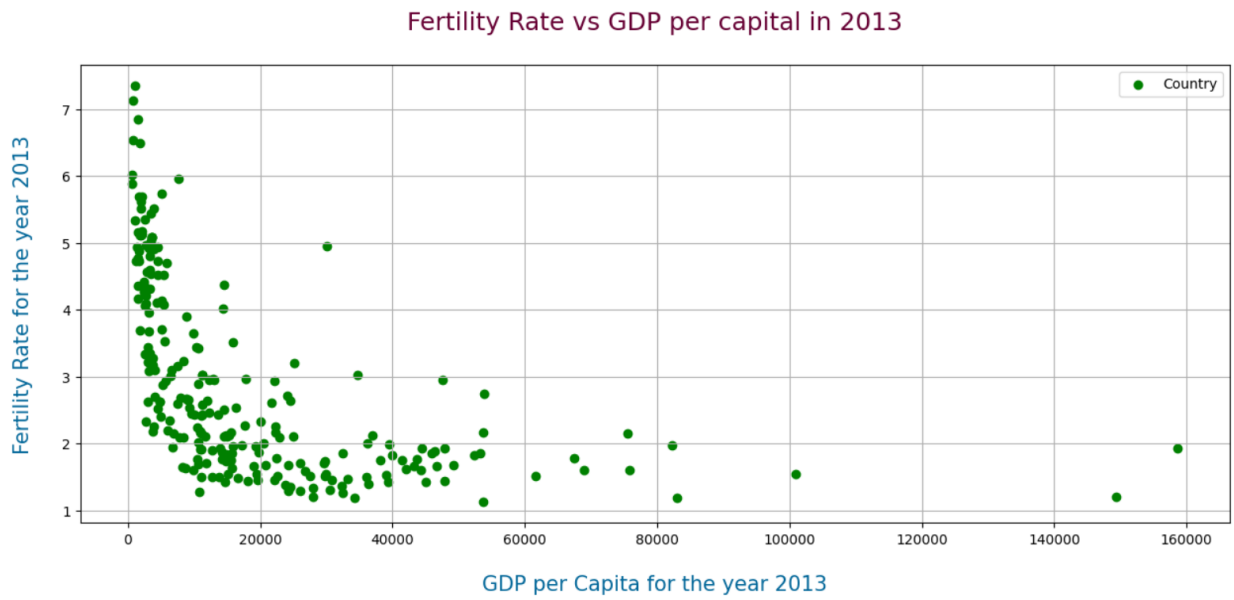
## Third party libraries used

- **Matplotlib**: Used the Pyplot submodule of Matplotlib to plot the graphs required in the question.
- **pandas**: It was used to manipulate the data-frame provided which involved reading the dataset in csv format and calculating the correlation coefficient.

## Implementation Process

To solve this problem, I downloaded two datasets, namely GDP per capita and Fertility rate from the world bank Indicators websites [Indicators | Data (worldbank.org)] and the used the pandas framework's function called read_csv() to read them. Then I extracted the records for the year 2013 on both GDP per capita and Fertility rate variables in their respective datasets. Lastly, I used the Pyplot submodule of matplotlib scatter plot the fertility rate for the year 2013 against the GDP per capita for the year 2013. Lastly, I calculated the correlation coefficient between the two variables using the .corr() function of pandas framework. Then, I printed the results to the console.

## Results




## Correlation Coefficient: -0.515

## Interpretation of the results

- Since correlation between GDP per capita and Fertility rate is negative, then the two variables are negatively associated with each other, and the Fertility rate decreases as GDP per capita increases and vice versa.
- From the graph we can affirm the above assertions since the fertility rate values are decreasing for higher values of GDP per Capita for the year 2013.

# Question 4

## Third party libraries used

- **pandas**: It was used to manipulate the data-frame provided which involved reading the dataset in csv format.
- **Matplotlib**: Used the Pyplot submodule of Matplotlib to plot the time series required in the question
- **NumPy:** Used to create 1D arrays in this case.
- **Statsmodels:** Used to calculate the autocorrelation for every time lag in a time series.

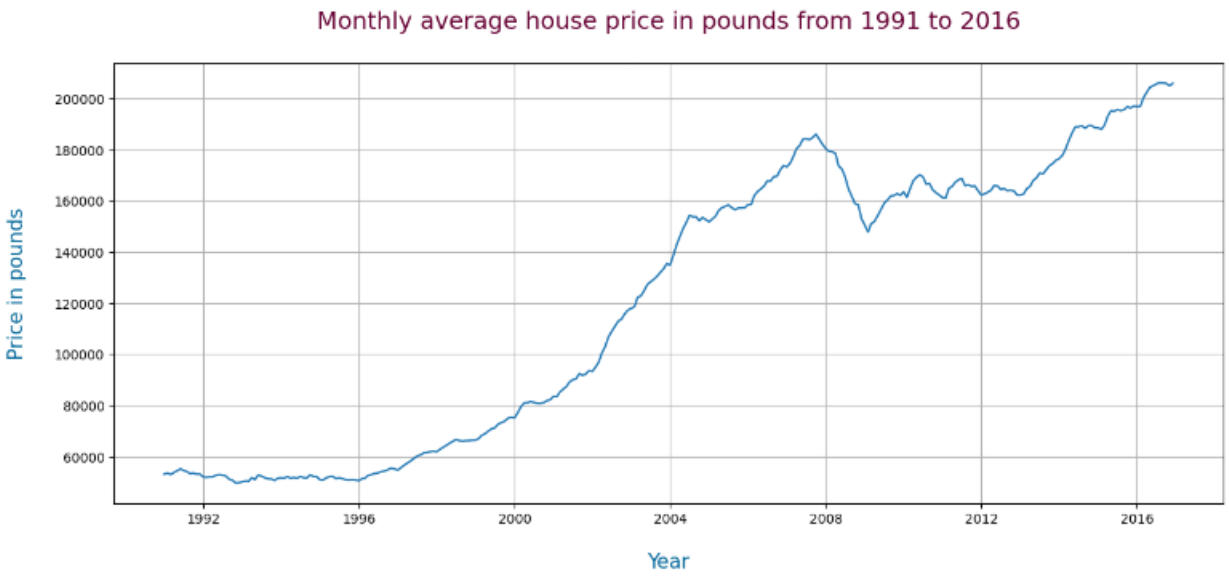
## Implementation Process

I read the dataset of monthly average house prices into a dataframe using pandas library's read_excel() function setting the date column as timestamp index of the dataframe, and I filtered the dataframe to exclude all records dated before 1991 and after 2016. Then, I assigned the timestamp index date column and the "Average House Price" to a pandas series object each and I plotted the time series line graph of Average house prices against years. Then I calculated the monthly returns for the average house prices and I used these returns to calculate the autocorrelation for every lag in the time series' monthly returns using the statsmodel library. [5] Using the autocorrelation values computed and lags ranging from 1 to 20, I plotted a bar graph of autocorrelation values against the twenty lags and labelled the graph using the Pyplot submodule of matplotlib. To determine the values of the ACF that would correspond to a statistically significant result at $p < 0.05$, I calculated the 95% confidence intervals using the formula $\pm 1.96/\sqrt{n}$. This gave me two values of the confidence interval and I display them by plotting two horizontal lines on the bar graph. The next process was to calculate the annualized return as a percentage, which I computed using the following formula.

$$\text{Annual Return} = \left( \left( (1+ R_1)x(1+R_2)x(1+R_3).....x(1+R_n) \right)^{[1/n]} - 1 \right) \text{ x } 100$$

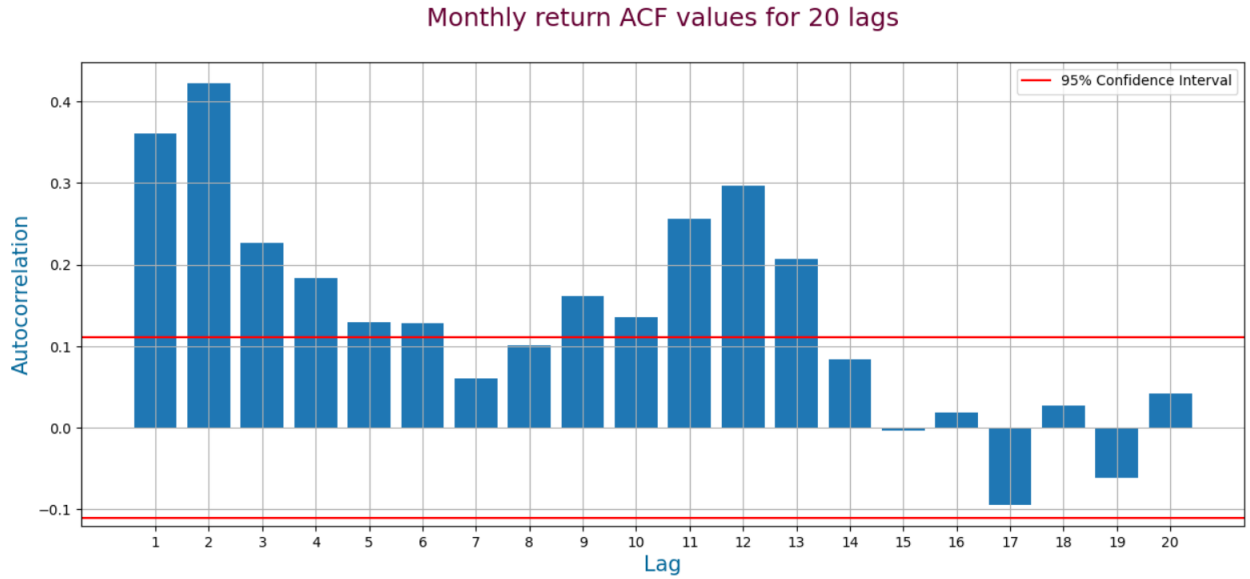
Whereby Rn represents the monthly return and n represents the number of years. [6]

## Results




## Interpretation of the house price time series line graph

The time series has a positive trend in a way that even though there is no consistent increase in the house prices, but the overall direction of the trend is positive and house prices in early years like 1991 were significantly lower than house prices in 2016.




## Interpretation of the monthly ACF values

It can be inferred from the graph that there is evidence of seasonality because the “gradual decrease immediately after a peak value is reached” -pattern is repeated at the 2$^{nd}$, 9$^{th,}$ and 12$^{th}$ lags.

```
The annualized return from 1991 to 2016 = 5.35 %
```

The annualized return as percentage from 1991 to 2016 is 5.35%

# Question 5

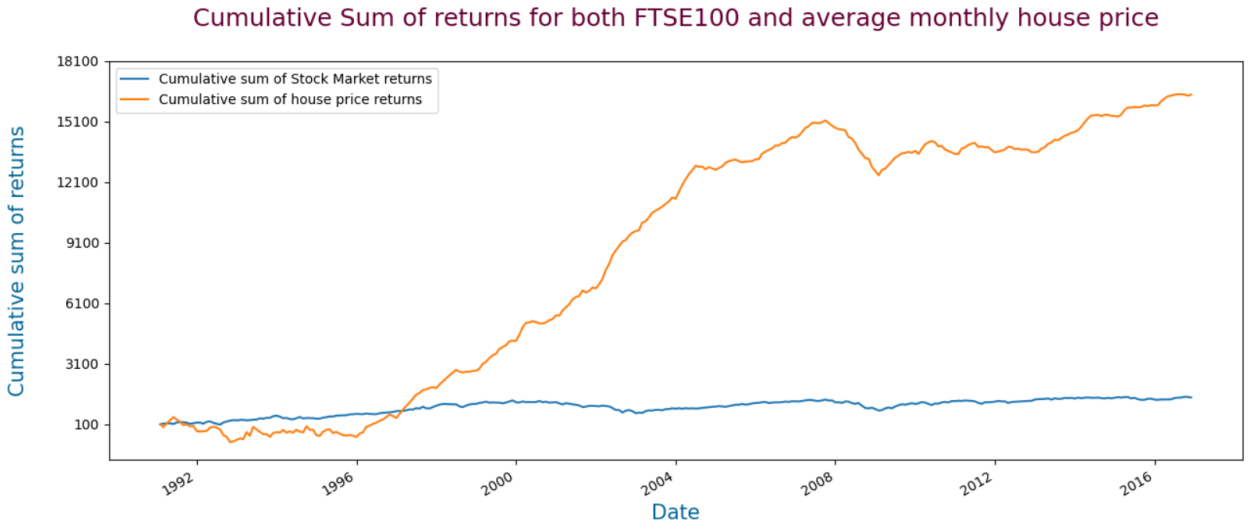
## Third party libraries used

- **pandas**: It was used to manipulate the data-frame provided which involved reading the dataset in csv format.
- **Matplotlib**: Used the Pyplot submodule of Matplotlib to plot the time series required in the question.

## Implementation Process

To solve this problem, I read the two datasets, one for Stock markets and the second one for monthly house pricing. The stock market column was ordered in descending order of dates hence I had to sort it to be ordered in ascending order. Next, I computed the monthly returns for both datasets, synchronized the two datasets to start from 100, and I computed the cumulative sums of the monthly returns for both datasets using pandas function called cumsum(). Later, I plotted the synchronized cumulative sums of house price returns, and the cumulative sums of the stock market returns as time series. Lastly, I calculated the annualized returns for house prices and stock market.

## Results



## Insights

The cumulative Sum of returns for house prices have increased at a higher growth rate from 1991 to 2016 compared to the growth rate of the cumulative returns for the stock market.

```
The annualized return for stock market from 1991 to 2016 = 4.46251547864067 2
The annualized return for average house prices from 1991 to 2016 = 5.35423853535919
```

Conclusion: Given that the annualized return percentage for house price is greater than the annualized return percentage for stock market, it would have been better to invest in a UK house than investing in a UK stock market.

[1] ‘scipy.stats.t — SciPy v1.11.3 Manual’. Accessed: Oct. 01, 2023. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html
[2] Zach, ‘Two Sample t-test: Definition, Formula, and Example’, Statology. Accessed: Oct. 01, 2023. [Online]. Available: https://www.statology.org/two-sample-t-test/
[3] Zach, ‘Welch’s t-test: When to Use it + Examples’, Statology. Accessed: Oct. 01, 2023. [Online]. Available: https://www.statology.org/welchs-t-test/
[4] ‘scipy.stats.ttest_ind_from_stats — SciPy v1.7.1 Manual’. Accessed: Oct. 01, 2023. [Online]. Available: https://docs.scipy.org/doc/scipy-1.7.1/reference/reference/generated/scipy.stats.ttest_ind_from_stats.html
[5] Zach, ‘How to Calculate Autocorrelation in Python’, Statology. Accessed: Oct. 02, 2023. [Online]. Available: https://www.statology.org/autocorrelation-python/
[6] ‘Annual Return’, Corporate Finance Institute. Accessed: Oct. 02, 2023. [Online]. Available: https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/annual-return/