Name: Niyomwungeri Parmenide ISHIMWE

Andrew-ID: parmenin

DATA, INFERENCE, AND APPLIED MACHINE LEARNING

18-785

ASSIGNMENT 3

03 OCTOBER 2022

I, the undersigned, have read the entire contents of the syllabus for course 18-785 (Data Inference and Applied Machine Learning) and agree with the terms and conditions of participating in this course, including adherence to CMU's AIV policy.

Signature: **Niyomwungeri Parmenide ISHIMWE**
Andrew ID: **parmenin**
Full Name: **Niyomwungeri Parmenide ISHIMWE**

The libraries used:

- numpy
- pandas
- matplotlib
- scipy
- statistics
- math
- statsmodels

QUESTION 1:

It was required to use a t-test to test whether women's energy intake deviates systematically from a recommended value of the population mean of 7725 kJ. This is done by first creating a list of sample elements, then using the list with **statistics.mean()** function to find the sample mean, **statistics.stdev()** function to find the sample's standard deviation, **stats.sem()** function to find the standard error of the mean, **stats.ttest_1samp()** function to find the t statistic, and p-value and calculate the degrees of freedom by subtracting one from the sample size.

Here, the Null hypothesis (Ho) can be "The women's energy intake has a mean of 7725 kJ" (μ = 7725). On the other hand, the Alternative hypothesis (H1) can be that "the women's energy intake deviates systematically from the mean of 7725 kJ" (μ ≠ 7725). Thus, as they were required, the sample mean is **6753.636363636364, t**he sample standard deviation is **1142.1232221373727, t**he standard error of the mean (SEM) is **344.3631083801271, t**he degrees of freedom are **10,** the t statistic is **-2.8207540608310193,** and the p-value is **0.018137235176105812.**

From that, we can infer that a **two-tailed hypothesis** is an appropriate test because the alternative hypothesis contains the not equal **"≠"** sign. In other words, it doesn't specify if the mean is greater or less than the population mean. In addition, because the p-value **(0.018137235176105812)** is less than (alpha level) **α = 0.05**, the null hypothesis can be rejected, and the alternative hypothesis can be kept. Hence, this confirms that the women's energy intake deviates from a recommended value of 7725 kJ.

QUESTION 2:

Here, it was asked to examine two samples using a t-test to test whether Guinness served in an Irish pub tastes better than pints served elsewhere around the globe. This means that the Null hypothesis can be "Guinness served in an Irish pub is better than pints served elsewhere around the globe" and the alternative can be "Guinness served in an Irish pub is not better than pints served elsewhere around the globe". Hence, this implies that a **two-sample test** is appropriate as it analyses the difference between the means of two independent samples. Finding the p-value and t-statistic required using both means, standard deviations, sample sizes, and the **stats.ttest_ind_from_stats()** functions.

Using the t-test, the p-value of **2.3158901628742276e-19** and a **t statistic** of **11.647653131319812** were found. Therefore, the difference between 74 and 57 is significant, because the p-value is lower than the significance level (**p <= α)**, hence the null hypothesis can be ruled out and the alternative hypothesis can be considered. This is a **left-tailed** test because the alternative hypothesis tends to be at the left since it is formulated as Guinness served elsewhere tastes less good than that served in Ireland.

QUESTION 3:

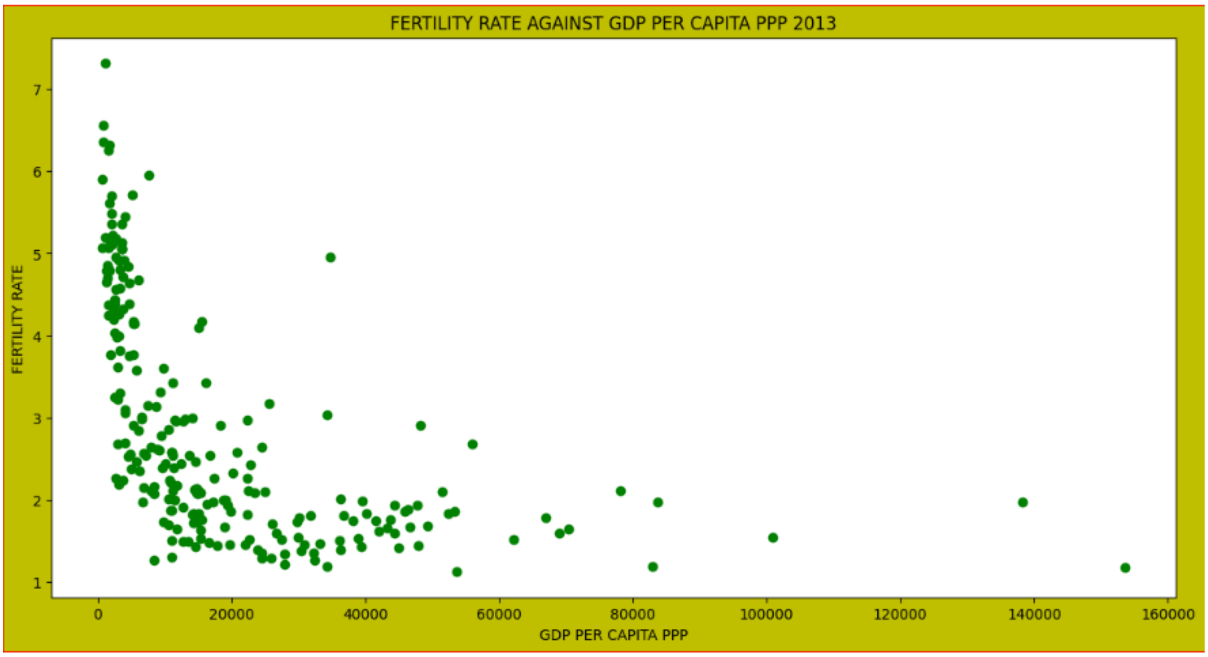

Figure 1: Fertility rate, total versus GDP per capita PPP

As it was required to graph the World Bank Indicators for 2013 to study the relationship between Fertility rate, total (births per woman) versus GDP per capita PPP (current international $), and estimate the correlation coefficient, the first thing to do is to import and extract 'Country Name' and '2013' columns from the two datasets and then plot them using the 2013 year only.

The plotted graph illustrates that the fertility rate decreases significantly as the GDP increases, this means that countries with low GDP have a tremendous number of births per woman.

Furthermore, calculating the coefficient of correlation gives **-0.525513538022326**, this shows a negative relationship between the fertility rate and GDP as the correlation coefficient value is less than zero. Obtaining a Negative coefficient of correlation tells that when the fertility rate increases, the value of the GDP tends to decrease. i.e., when one increases the other tends to decrease to the same degree and vice versa.

QUESTION 4:

It was asked to plot the time series and monthly average house price data in pounds sterling (£) from Jan 1991 to Dec 2016 and construct the autocorrelation function (ACF) of the monthly returns, and then show the values for lags of one up to 20 using a bar-graph. This is done by reading Monthly Average House Price from the excel file, naming the date column, and selecting the data between 1991-01-01 and 2016-12-31 to give the following plot. It can be inferred from the graph that the monthly average increases as time pass.

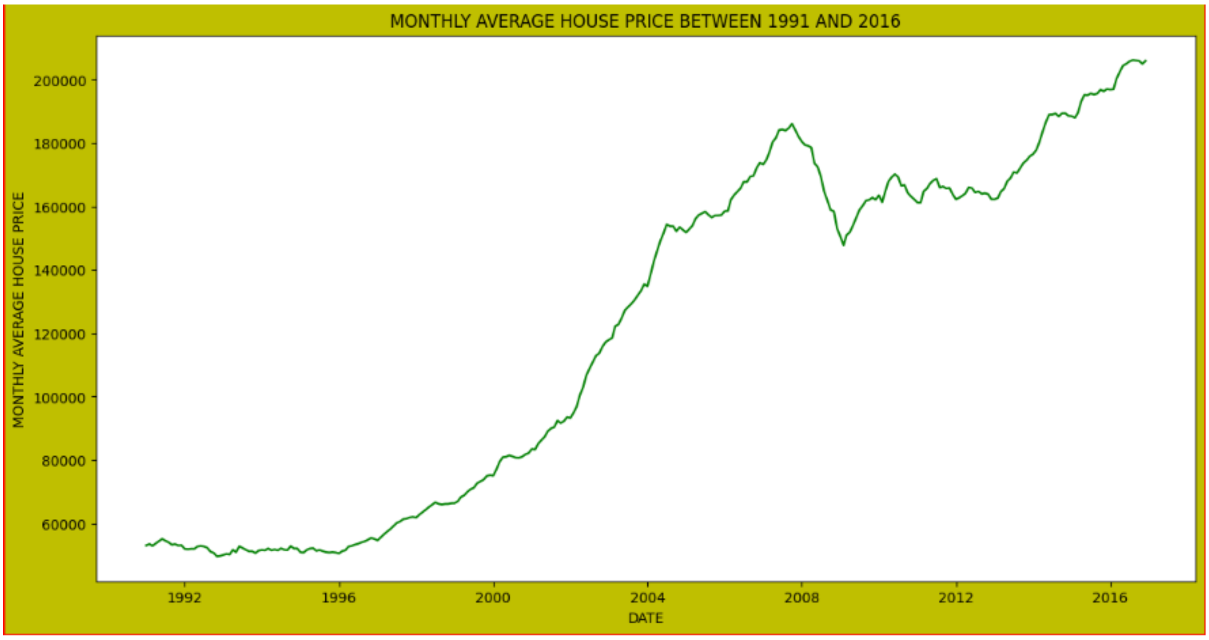

Figure 2: Monthly average house price 1991 – 2016

To construct the autocorrelation function (ACF) of the monthly returns, the first thing to do is to calculate and store the returns in the list. Secondly, use **tsa.acf()** function from the **statsmodels** library to calculate autocorrelation values, and then select 20 of them, and plot them by showing their values for lags of one up to 20 using a bar graph. Moreover, the values of the ACF that correspond to a statistically significant result at p<0.05 are calculated using the ((+ or -) 1.96/(**n**)**0.5) formula where n is the number of returns and shown using horizontal lines.

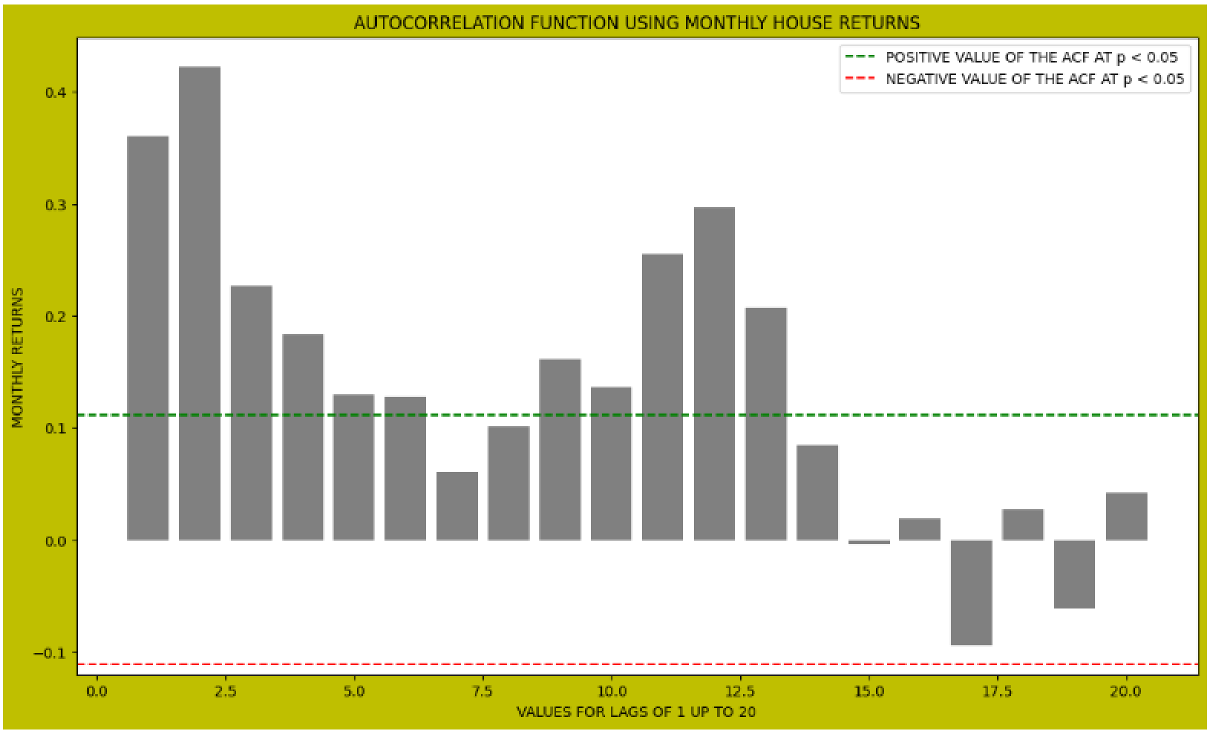

Figure 3: Autocorrelation function of the house monthly returns

From the monthly average house price plot, there is no seasonality because the graph patterns have unpredictable intervals.

There is a bit of a trend in the time series because the graph shows that from 1991 until around 2007, the monthly house price tends to increase steadily for this period of around 17 years.

Finally, to calculate the annualized return as a percentage, the formula used is “Annualized Rate of Return Formula = ((Ending Value / Beginning Value) ** 12/n) – 1” where n is the number of months that resulted in **5.35423853535919%** as the annualized return over the given period as a percentage.

QUESTION 5:

The task was to plot cumulative returns from the House market and the FTSE100 index on the same graph with the time series normalized such that each starts at 100 in Jan-1991 and the average annualized return from the FTSE100. This is done by first bringing in both house prices and FTSE100 data from excel files, naming the date column for the house prices dataset, and sorting the FTSE100 data set on the date column to align with the house prices dataset, selecting the data between 1991-01-01 and 2016-12-31 for the house prices dataset. After that, the returns are calculated and stored in corresponding lists, next is to calculate the cumulative sums using the calculated returns, normalizing them to start at 100, and plotting the normalized values with time series to produce the following plot. It can be inferred from the graph that the monthly average increases as time pass.

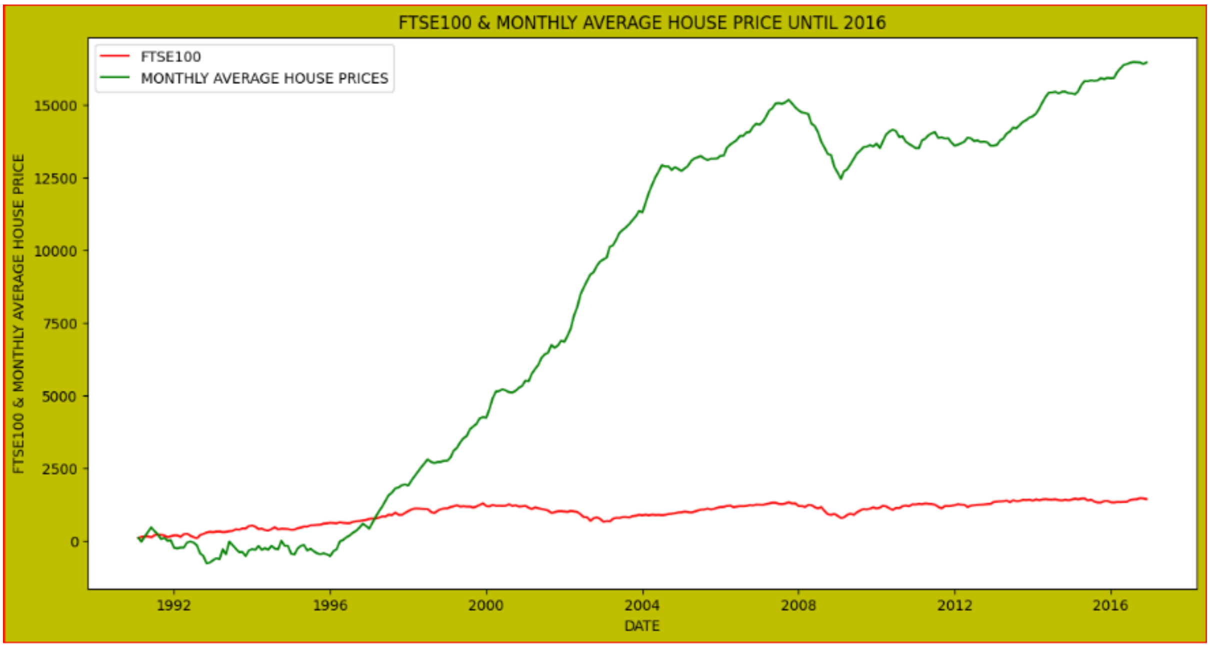

Figure 4: FTSE100 and monthly average house price between 1991 and 2016

The annualized return for the FTSE100 is **4.462515478640672%.** This implies that it is better to invest in the UK house than to invest in the UK stock market because house prices have a higher percentage of annualized returns of **5.35423853535919%** than that of the stock market which is **4.462515478640672%.** In addition, the graph of the monthly average house price increases sharply compared to that of the FTSE100 stock market, which can also lead to investing in UK houses rather than the UK stock market.