REPORT OF ASSIGNMENT 3

Carnegie Mellon University Africa

Submitted By: Peace Ekundayo Bakare

Course: Data, Inference, and Applied Machine Learning

Libraries Used:

Pip Installed quandl, pandas, matplotlib, numpy

1. *Numpy as peacenp for calculations*.
2. *Import stats from scipy*.
3. *Tabulate from tabulate for tabular display or presentation*.
4. *Math for mathematical calculations*.
5. *Matplotlib.pyplot as peaceplt for scatter plot diagram*.
6. *Pandas as peacepd for dataframes and operations*
7. *Statsmodels.graphics.tsaplots to plot acf graph*
8. *Statsmodels.api as sm to plot the acf graph and annotation*.

QUESTION 1

Null hypothesis H_0 : μ = 7725 kJ

Alternative hypothesis H_1 : μ ≠ 7725 kJ

The Null hypothesis being 7725 kJ shows that the mean daily energy intake of the women is equal to the recommended value of 7725 kJ.

The Alternative hypothesis not equal to 7725 kJ shows that the mean daily energy of women is not the recommended value of 7725 kJ.

Therefore, I am conducting a **Two-tailed test** to determine if the mean energy intake of the women is either greater than or less than 7725 kJ, based on the alternative hypothesis not equal to 7725 kJ.

Using a significance level of α = 0.05

To Calculate the Sample Mean

The sample mean,

$$\bar{x} = \frac{\sum x}{n} = 6{,}753.64\ kJ$$

Where n = 11 and x is the daily energy intake for each of the women.

To Calculate the Standard Deviation of the Sample and the Standard Error of the Mean,

Standard Deviation,

$$s = \sqrt{\frac{\sum(xi\ -\ \bar{x})^2}{n-1}} = 1{,}142.12\ kJ$$

Where: xi is each observation of daily energy intake of the women,

$\bar{x}$ is the sample mean, and *n is the sample size, 11.*

Standard Error of the Mean (SEM),

$$SEM = \frac{s}{\sqrt{n}} = 344.363$$

Where s is the standard deviation of the sample, and n is the sample size, 11.

To Calculate the t-statistic,

$$t = \frac{\bar{x} - \mu}{SEM} = -2.82075$$

Where, $\bar{x}$ is the sample mean, μ is the mean of the population, and SEM is the standard error of the mean.

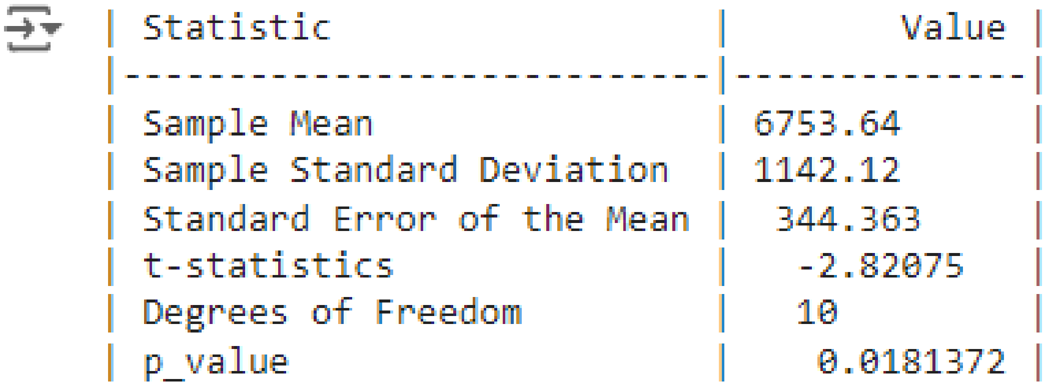
To Calculate the degrees of freedom and p-value

$$\begin{aligned} Degrees\ of\ Freedom &= Sample\ Size\ (n) - 1 \\ &= 11 - 1 \\ &= 10 \end{aligned}$$

$$p - value = 2\ x\ \Pr(T\ \geq |t|) = 0.0181372$$

Where, $\Pr\ (T\ \geq |t|)$ *is the probability that a t-distributed random variable with degrees of freedom is greater than or equal to the absolute value of the observed t-statistic.*

Fig: Statistics and their values.



```
| Statistic                     |          Value |
|-------------------------------|----------------|
| Sample Mean                   | 6753.64        |
| Sample Standard Deviation     | 1142.12        |
| Standard Error of the Mean    |  344.363       |
| t-statistics                  |   -2.82075     |
| Degrees of Freedom            |   10           |
| p_value                       |    0.0181372   |
```

Conclusion:

If the p-value is less than 0.05 $(p < 0.05) => (0.01814 < 0.05)$, we reject the null hypothesis and conclude that there is significant evidence that the women's daily energy intake is different from the recommended 7725 kJ.

If the p-value were greater than 0.05 $(p > 0.05)$, we would fail to reject the null hypothesis implying that the data does not provide sufficient evidence to suggest that the daily energy intake differs from 7725 kJ.

Since the p-value is less than 0.05, we reject the null hypothesis.

Insight:

In the real world, studies like this are essential for evaluating people's eating habits. For instance, understanding if a group consistently consumes more or less energy than recommended can help health professionals tailor public health policies, create better meal plans, and develop interventions that improve overall diet and nutrition. These insights are especially valuable when addressing issues like malnutrition and obesity.

QUESTION 2

- **Come up with your null and alternative hypothesis**

Null hypothesis H_0: $\mu_{Ireland} = \mu_{Elsewhere}$

Alternative hypothesis H_1: $\mu_{Ireland} \neq \mu_{Elsewhere}$

The Null hypothesis H_0 being $\mu_{Ireland} = \mu_{Elsewhere}$ shows that there is no significant difference between the mean of Guinness Overall Enjoyment Score served in Ireland and the mean of Guinness Overall Enjoyment Score served elsewhere, as they are equal.

The Alternative Hypothesis H1 being $\mu_{Ireland} \neq \mu_{Elsewhere}$ shows that there is difference between the mean of Guinness Overall Enjoyment Score served in Ireland and the mean of Guinness Overall Enjoyment Score served elsewhere.

- **Use level of significance (alpha level) $\boldsymbol{\alpha = 0.05}$**

I am using a significance level $\alpha = 0.05$

- **Determine whether a one-sample, two-sample or paired test is appropriate**

In this situation, a paired test is not appropriate because there are two independent groups, and there is no pairing between them. I am using a **two-sample t-test** because I am comparing the mean from the Guinness Overall Enjoyment Score served in Ireland to the mean from the Guinness Overall Enjoyment Score served elsewhere.

- **Calculate the Degrees of Freedom**

Two sample t-test, unequal variances [1]

$$Degrees\ of\ Freedom = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

Where:

$\boldsymbol{s_1} = \mathbf{7.4}\ is\ the\ standard\ deviation\ of\ the\ sample\ from\ Ireland$
$\boldsymbol{s_2} = \mathbf{7.1}\ is\ the\ standard\ deviation\ of\ the\ sample\ from\ elsewhere$
$\boldsymbol{n_1} = \mathbf{42}\ is\ the\ sample\ size\ for\ Ireland$
$\boldsymbol{n_2} = \mathbf{61}\ is\ the\ sample\ size\ for\ elsewhere$

$$\boldsymbol{Degrees\ of\ Freedom = 85.87}$$

- **Calculate the t-statistic and p-value**

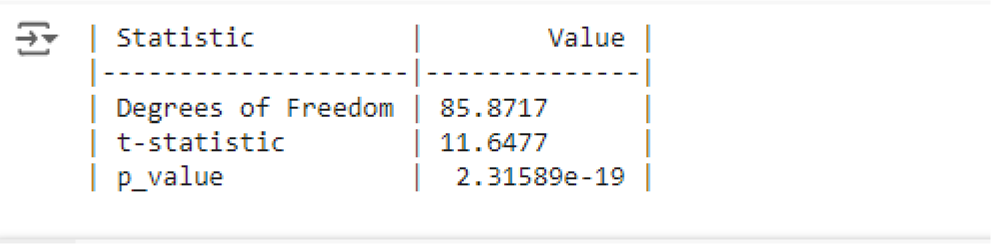
$$t - statistic = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

$$\bar{x}_1 = 74 \text{ is the mean score in Ireland}$$
$$\bar{x}_2 = 57 \text{ is the mean score elsewhere}$$

$$\boldsymbol{t - statistic = 11.65}$$
$$\boldsymbol{p - value = 2.32\ x\ 10^{-19}}$$

Fig: Table showing the Degrees of Freedom, t-statistic and p-value



```
| Statistic          |         Value |
|--------------------|--------------|
| Degrees of Freedom | 85.8717      |
| t-statistic        | 11.6477      |
| p_value            |  2.31589e-19 |
```

- **Give Explanation based on the results obtained**

If p *is less than* 0.05 *i.e.* $p < 0.05$, **we reject the null hypothesis** and conclude that there is a significant difference between the Guinness Overall Enjoyment Scores in Ireland and elsewhere.

If p *is greater than* 0.05 *i.e.* $p > 0.05$, **we fail to reject the null hypothesis** which shows that the observed difference between the Guinness Overall Enjoyment Score of Ireland and Elsewhere could be due to a random variation or a natural occurrence.

Since the p-value is less than 0.05, we reject the null hypothesis.

QUESTION 3

- **Download the required datasets and find ways to deal with the first rows which are not required**

I downloaded the 2 datasets required namely, Fertility rate, total (births per woman) and the GDP per capita PPP (current international $) from the World Bank Indicators.

To deal with the first rows, I utilized the ‘skiprows’ parameter while reading the csv file into the dataframe I created and named as *df_GDP_data* and *df_fertility_data* for the GDP per capita dataset and Fertility rate dataset, respectively.

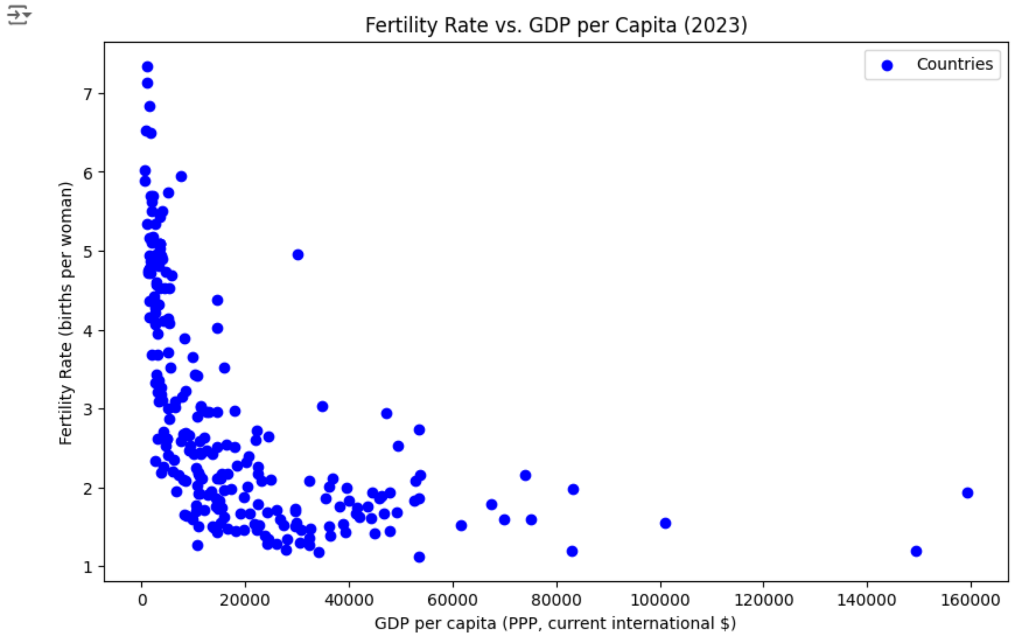
In addition, I used the pandas function *head()* to view the first 5 rows in each dataframe.

Also, from both dataframes, I selected the data from the “Country Name” and “2013” columns and renamed the columns to “Country” and “2013 GDP per capita” for the GDP dataframe, and “2013 Fertility rate” for the Fertility rate dataframe.

Furthermore, I merged the two dataframes into one dataframe named *df_merged_data* on the “Country” column, using inner merge. To the newly created dataframe, I dropped the NaN values using the *dropna()* function of pandas.

- **Make a scatter plot for Fertility Rate against GDP per capita PPP using the year 2013 for each dataset and label the graph**

I plotted the graph of Fertility Rate against GDP per capita PPP using the year 2013 for each dataset. The x and y axis properly labelled, the right title given and a legend showing the countries plotted.

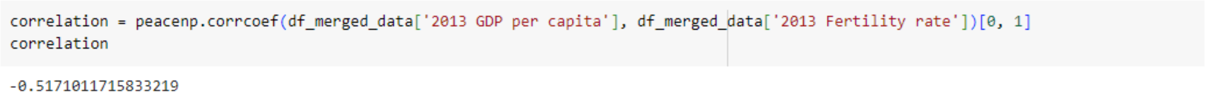


- **Give your interpretation of the graph**

The Scatter plot shows that countries with higher GDP per capita have lower fertility rates than countries with lower GDP per capita. This trend suggests that rising income tends to lower fertility rates. Several reasons could be considered as to why countries with higher GDP per capita have lower fertility rates. Some of which are,

1. Countries with higher GDP per capita often have better access to healthcare services and family planning options, including contraception and reproductive health education, which allows people to make more informed decisions about the size of their families.
2. The cost of raising children in countries with higher GDP per capita can be significantly higher, including healthcare, education etc., when compared to countries with lower GDP per capita. Hence, families in these higher GDP per capita countries may opt in for fewer children to maintain a higher standard of living.
3. Due to professional and personal goals, countries with higher GDP per capita tend to delay their childbearing. Delay in childbearing for women could affect the total number of children they eventually have.

- **Calculate the correlation coefficient using the same columns**



```
correlation = peacenp.corrcoef(df_merged_data['2013 GDP per capita'], df_merged_data['2013 Fertility rate'])[0, 1]
correlation
```

```
-0.5171011715833219
```

I used the *corrcoef()* function in the numpy library to calculate the correlation coefficient.

The value gotten is $\mathbf{-0.517101}$

- **Give your interpretation of the estimated correlation coefficient**

The correlation coefficient gotten is not close to -1 and not close to 0. It indicates a **moderate negative correlation**, implying that the higher GDP per capita is strongly correlated with lower fertility rates.

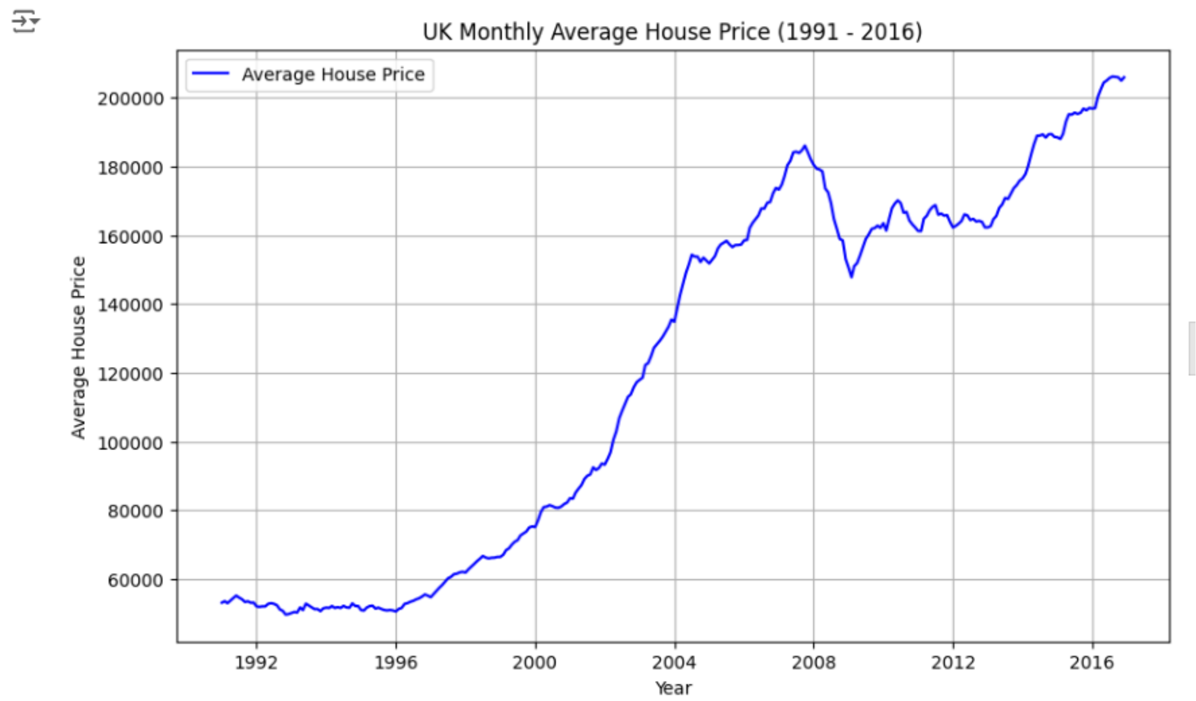
QUESTION 4

- **Download the required datasets**

I downloaded the dataset using *read_excel*() function in the pandas library, after which I checked the head and tail of the dataframe using *head()* and *tail()* respectively, to understand the data. Also, I renamed the date column from "Unnamed: 0" to "Date."

- **Plot a carefully labelled time series graph of UK monthly indices (Post '91)**

To plot the graph of Average House Price against Year, I had to filter the data with the date column, selecting only dates less than or equal to 2016, as required by the question.



The graph shows that the years are plotted on the x-axis, calibrated from 1992 to 2016, while the y-axis contains data for the average house prices, ranging from 60,000 pounds to 200,000 pounds.

- **Give your interpretation of the graph**

From the graph, it can deduced that the house prices remained relatively steady but there was steady rise in prices from mid 90s to early 2000. From 2000 to 2007, the most noticeable growth was observed, reaching a price of 200,000 pounds. This reflects a housing boom during this period, where property values surged dramatically. Between

2007 and 2009, there was a sharp decline which aligns with the Global Financial Crisis, bringing the average house price to about 160,000 pounds. We could say that there is a recovery and continued growth from 2009 because there is continued demand for property.

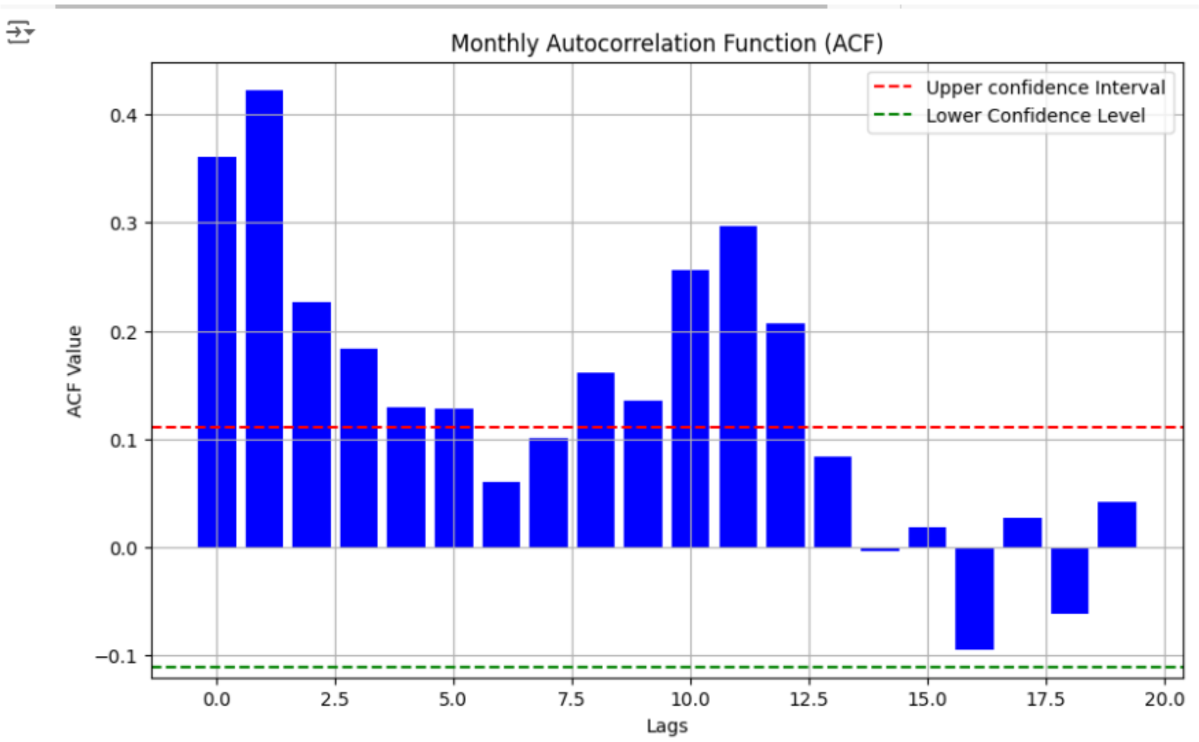
Summarily, the graph reflects the cyclical nature of the housing market, with periods of growth, a major downturn, and recovery. Investing in the UK housing market during this period (1991 – 2016) would have yielded substantial gains in the long term, despite temporary setback during the financial crisis.

- **Use the formula provided to calculate the monthly returns**

Given the formula $r(t) = \left[\frac{p(t)}{p(t-1)}\right] - 1$, I calculated the monthly returns of the house prices by dividing the current house price by the previous month's price in the "Average House Price" column, and subtracting 1 from the value returned. These values are stored in a column called, "Monthly Return".

- **Plot the ACF function using the monthly returns**

I used the values in the "Monthly Return" column to calculate values stored in a *acf_data* which I plotted in bars. I imported the statsmodels.api library to make this calculation. The legend of the graph is also included. The y-axis is called "ACF Value" while the x-axis is called "Lags".



I have also carefully labelled the graph and indicated the values of the ACF at p < 0.05 using horizontal lines red to indicate the upper confidence level and the green dotted line to indicate the lower confidence level.

- **Give your explanation on whether there is evidence of seasonality**

There is some evidence of seasonality, especially around lag 10 to 12, which may suggest that the returns could have a yearly cyclic pattern. The significant autocorrelation at low lags also indicates that past returns impact future returns over short periods.

- **Calculate the annualized return for this period as a percentage**

I calculated the annualized return by selecting the first house price in 1991, and the last house price in 2016. I subtracted the dates between these two prices and converted it to years by dividing by 365.25 i.e. 365 one-quarter days. I calculated the annualized return using the formula below:

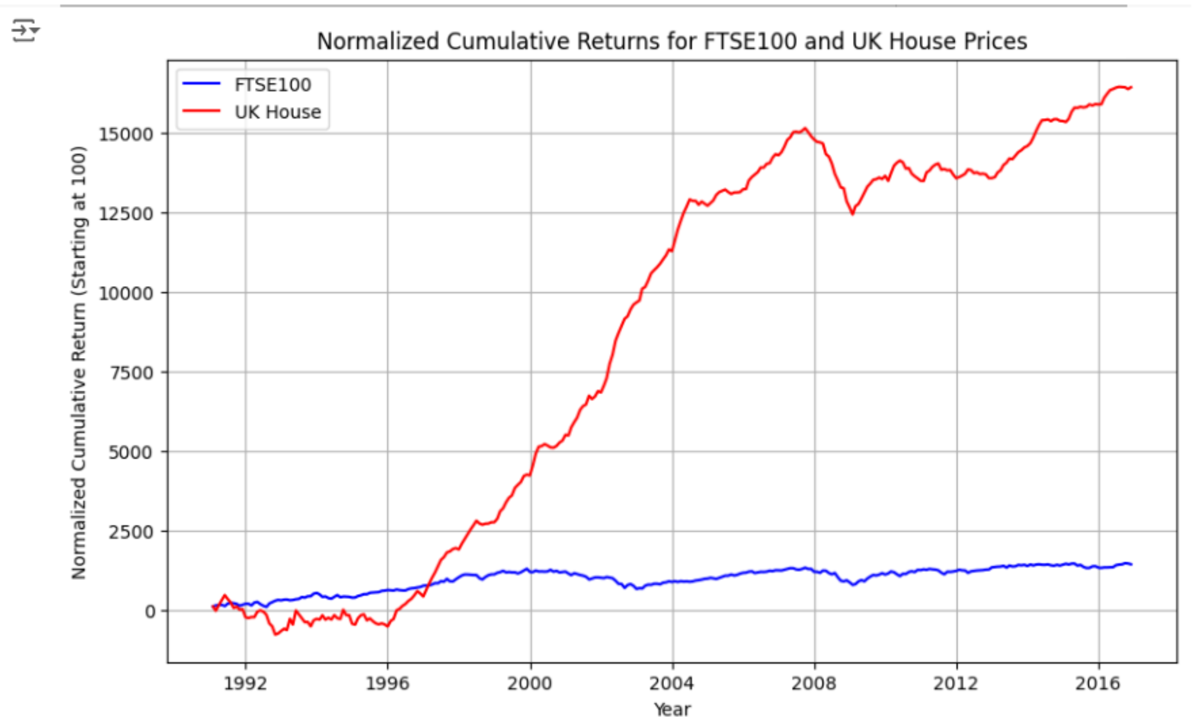
$$Annualized\ Return = (\frac{Final\ Price}{Initial\ Price})^{\frac{1}{Years}} - 1$$

QUESTION 5

I downloaded the FTSE data using *read_csv()* function from the pandas library. I used the *head()* and *tail()* to understand the data, i.e. the start year and end year. I converted the Date column to datetime to enable filtering without error and data type conversion error. After that, I sorted the data in the dataframe using the Date column in ascending order.

I calculated the cumulative of returns for FTSE after filtering the sorted data with the dates column, selecting data less than or equal to 2016.

I plotted the graph of Normalized Cumulative Returns for both FTSE100 and UK House Prices, having the normalized cumulative return, starting at 100, on the y-axis, and the years on the x-axis.



The Annualized return for FTSE100 was gotten as **4.477211147844482** as previously calculated for UK House prices in Question 4 above.

I have also printed the 2 values together to make a comparison and make a decision if it would have been better to invest in a UK house or the UK Stock market over this period.

Obviously, it would be better to have invested in the UK House than in the UK stock market over this period considering the annualized return for both commodities, the UK House price returning **5.3719455** which is above the **4.47721114** of the FTSE.

References

[1] “people.umass.edu/bwdillon/files/linguist-609-2020/Notes/TwoSampleT-Test.html.” Accessed: Sep. 28, 2024. [Online]. Available: https://people.umass.edu/bwdillon/files/linguist-609-2020/Notes/TwoSampleT-Test.html