**DATA, INFERENCE**

**&**

**APPLIED MACHINE LEARNING**

**(COURSE 18-785)**

**ASSIGNMENT 3**

**Mark Iraguha**

**(miraguha)**

Table of Contents

[Libraries Used: 3](#_Toc177420988)

[Introduction: 3](#_Toc177420989)

[Question 1 Report: 4](#_Toc177420990)

[Methodology 4](#_Toc177420991)

[Results 4](#_Toc177420992)

[Analysis and Insights 7](#_Toc177420993)

[Question 2 Report: 9](#_Toc177420994)

[Methodology 9](#_Toc177420995)

[Results 9](#_Toc177420996)

[Analysis and Insights 10](#_Toc177420997)

[Question 3 Report: 11](#_Toc177420998)

[Methodology 11](#_Toc177420999)

[Results 11](#_Toc177421000)

[Analysis and Insights 12](#_Toc177421001)

[Question 4 Report: 14](#_Toc177421002)

[Methodology 14](#_Toc177421003)

[Results 14](#_Toc177421004)

[Analysis and Insights 16](#_Toc177421005)

[Question 5 Report: 18](#_Toc177421006)

[Methodology 18](#_Toc177421007)

[Results 18](#_Toc177421008)

[Analysis and Insights 19](#_Toc177421009)

[References: 21](#_Toc177421010)

# Libraries Used:

Matplotlib – a python plotting library used to create animated, interactive and static visualizations.[1]

Pandas – another Python library used that provides data structures and functions used to carry out data analysis.[2]

Numpy – a simple yet powerful data structure provided in python.[3]

Scipy – a python library that enhances the features of Numpy through a wide range of functions and tools essential for scientific computing.[4]

# Introduction:

This report details the completion of Assignment 3. Assignment 3 requests answers to 5 critical thinking and data analytical questions.

# Question 1 Report:

## Methodology

To find out if the distribution might have a mean of 7725kJ using a t-test, a null and alternative hypothesis should be defined.

Hypothesis Definition:

Null hypothesis (H0): The women’s energy intake does not deviate systematically from a recommended value.

Alternative hypothesis (H1): The women’s energy intake deviates systematically (either greater or less than the recommended value).

Thus, a two-tailed test is appropriate since the alternative hypothesis is whether the women’s energy intake is not equal to 7725KJ.

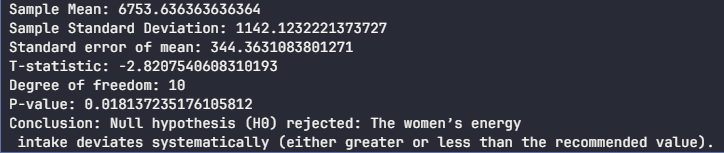
Calculating the sample mean, sample standard deviation, standard error of the mean (SEM), t-statistic, degrees of freedom and p-value.

Approach:

* Store the Daily energy intake values.
* Define the hypothesized mean and significance level (alpha level).
* Calculate the sample mean.
* Define the population mean (in this case, it’s the same as sample mean).
* Calculate the standard deviation, degree of freedom, sample standard deviation, standard error of mean (SEM).
* Calculate the t-statistic utilizing the ttest\_1samp function of the stats module of scipy library.

## Results

Six values representing the sample mean, sample standard deviation, standard error of the mean (SEM), t-statistic, degrees of freedom and p-value.



*Six numbers representing the sample mean, sample standard deviation, standard error of the mean (SEM), t-statistic, degrees of freedom and p-value.*

Null Hypothesis rejection status:

After performing the t-test, the t-statistic and p-value is calculated. The p-value obtained from the test is less than the significance level, so the null hypothesis is rejected.

By rejecting the null hypothesis, a conclusion is made that there is statistically significant evidence to support the claim that the women’s energy intake deviates systematically from the recommended value of 7725kJ.

## Analysis and Insights

Rejecting the null hypothesis indicates significant evidence that supports the idea that women’s energy intake systematically deviates from the recommended value of 7725kJ.

The results highlight the necessity for targeted nutritional interventions aimed at addressing women’s energy intake. One way could be by holding/advocating for public health campaigns focusing on educating women about balanced diets along with suitable portion sizes to help them effectively meet their energy requirements. These public health initiatives aimed at enhancing nutritional outcomes could help promote healthier eating habits.

# Question 2 Report:

## Methodology

Finding out the kind of significance of 74 versus 57.

In order to find out the whether the difference of 17 points between the two mean scores. A statistical approach would be appropriate to find out the implications of this comparison. Typically, a hypothesis test should be conducted.

A null and alternative hypothesis should be defined.

Hypothesis Definition:

Null hypothesis (H0): There’s no significant difference between the GOES scores of Ireland and Elsewhere.

Alternative hypothesis (H1): The GOES score in Ireland is significantly higher than that in the Elsewhere group.

Since, the alternative hypothesis expresses interest in whether the score in Ireland is greater than that derived from the group Elsewhere, a one tailed t-test would be appropriate.

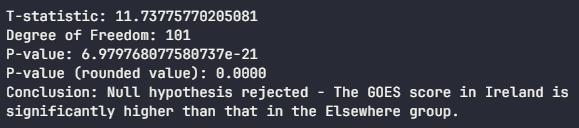
We can then proceed to calculate the t-statistic.

Approach:

* Store the results of the GOES scores in variables.
* Define the hypothesized mean and significance level (alpha level).
* Use the pooled standard deviation formula to get the standard deviation. This is necessary for the t-test later.
* Calculate the t-statistic using the one sample t-test.
* Calculate the degree of freedom.
* Calculate the p-value using the degree of freedom and the cumulative distribution function from the scipy library.

## Results

T-statistic and p-value.



*T-statistic, p-value and conclusion of null hypothesis.*

## Analysis and Insights

Statistical Test

* A one tailed t-test was appropriate here, since we are specifically interested in whether the score in Ireland is greater than that in the group Elsewhere.

Statistical Significance

* The p-value form the t-test is observed to be less than the significance level (0.05) so we reject the null hypothesis. This concludes that the difference in scores is statistically significant.
* This means that consumers perceive products or experiences in Ireland more favorably compares to any other locations.
* This kind of finding can be used utilized by efficient marketing strategies i.e marketing campaigns can leverage this perception (Irish products are perceived).

# Question 3 Report:

## Methodology

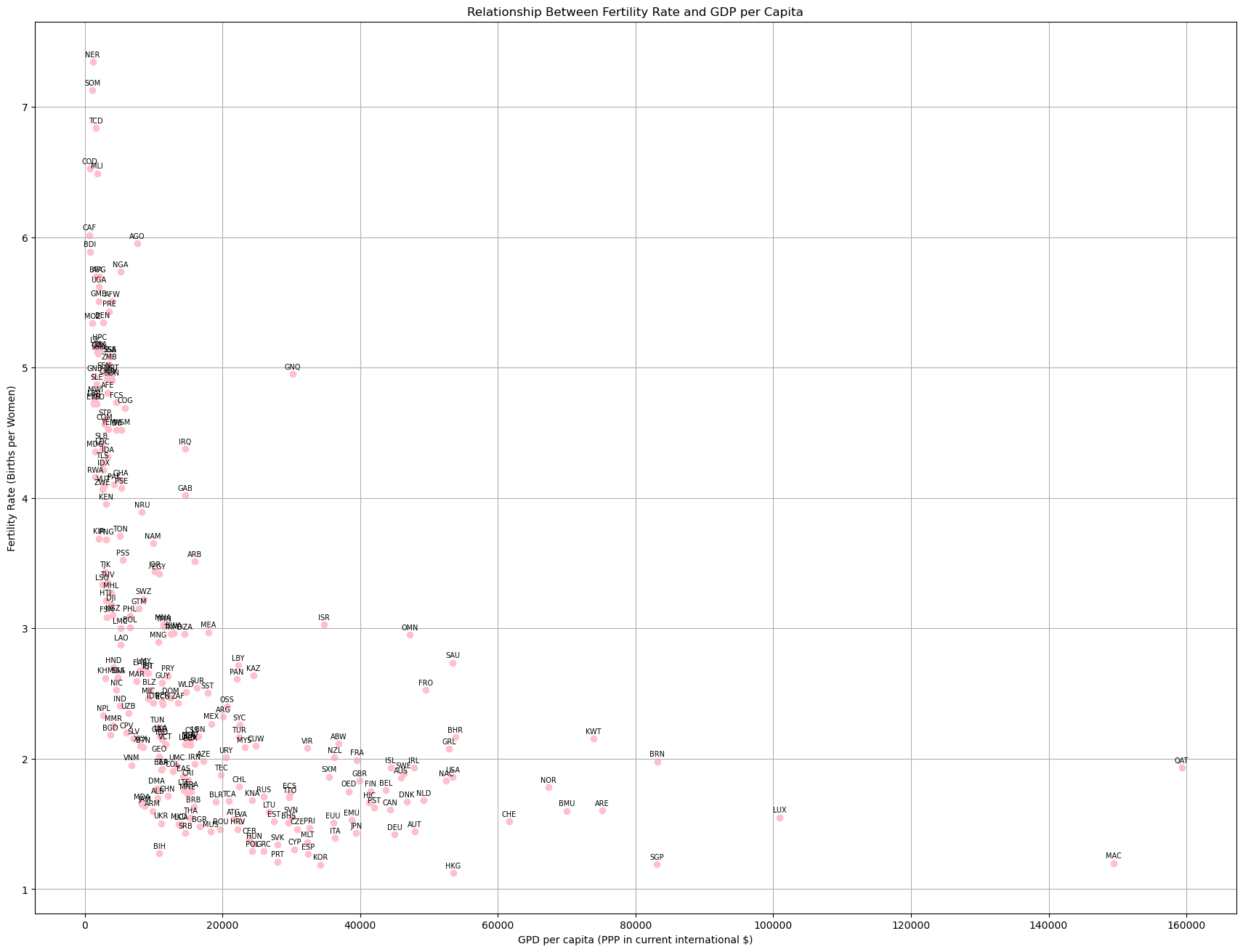
Plot graph to aid study the relationship between Fertility Rate, total (births per woman) versus GDP per capita PPP (current international $)

Approach:

* Download data for “Fertility Rate, total (births per women)” and “GDP per capita PPP (current international $)”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as a single column).
* Filter data to year 2013.
* Reset the index.
* Plot a graph with data (scatter plot).

## Results

A L-shaped curve of a scatter plot, as GDP rises, fertility rates tend to decline.



*Figure showing the relationship between fertility rates and GDP per capita*

**

*Correlation Coefficient using fertility rates and GDP per capita*

## Analysis and Insights

Coefficient Correlation Interpretation

* The resulting Correlation Coefficient of approximately -0.517 indicates that as one variable increases the other tends to decrease. In this case, there’s a strong negative correlation, suggesting that as GDP per capita increases, the fertility rate tends to decrease significantly.
* This trend is well-documented in demographic studies; poorer countries often have higher fertility rates compared to wealthier countries[5].
* This could signify that as countries develop economically, they may implement policies that promote smaller family sizes through education and increased access to contraception.

Interpretation of graph

* The graph plotted is a scatter plot that takes on the L-curve shaped graph.
* The l-shaped graph, kind of like a inverted j-shaped graph, can be divided into two main phases: first phase and second phase.
* In the first phase, it’s observed that as GDP increases, there’s a great decrease in fertility rates. This signifies how economic development often correlates with lower fertility rates due to factors like increased education.
* In the second phase, once the GDP reaches a certain threshold, the fertility rate stabilizes. Whereas the GDP continues to increase, the fertility rate does not increase significantly. This suggests that once peoples’ basic needs are met, any further economic growth does not lead to further reductions in fertility rates.

# Question 4 Report:

## Methodology

Utilizing numpy to produce cumulative distribution functions.

Approach:

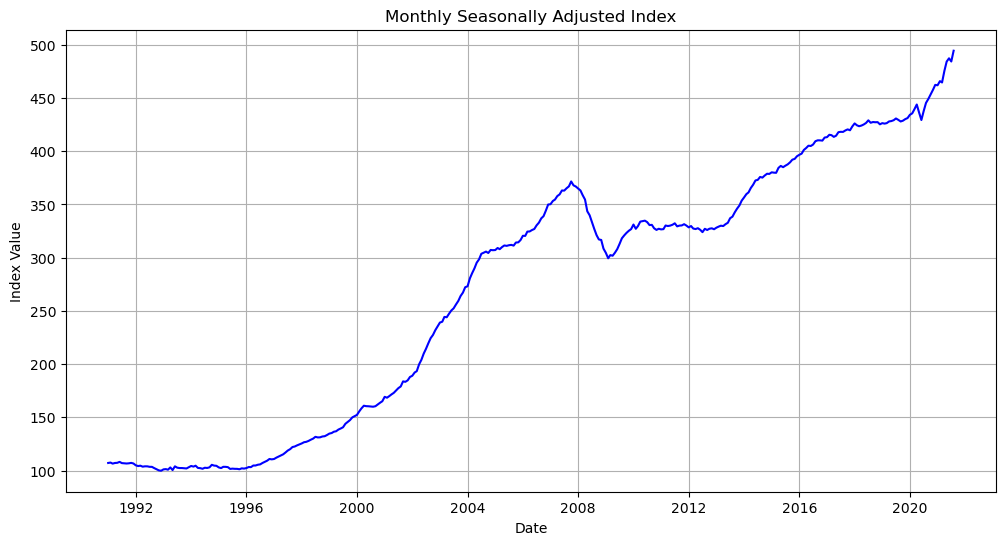
* Download data for “Fertility rate, total (births per woman)” and “GDP per capita (current US$)”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as a single column).
* Configure and plot a scatter plot graph (2010 data only).
* Sort the data frames.
* Produce cumulative distribution functions.
* Calculate the median and mean for both years.
* Configure and plot line graph (1990 and 2010 data only).

SPECIFY THAT YOU USED A LIBRARY TO GET THE T- STATISTIC

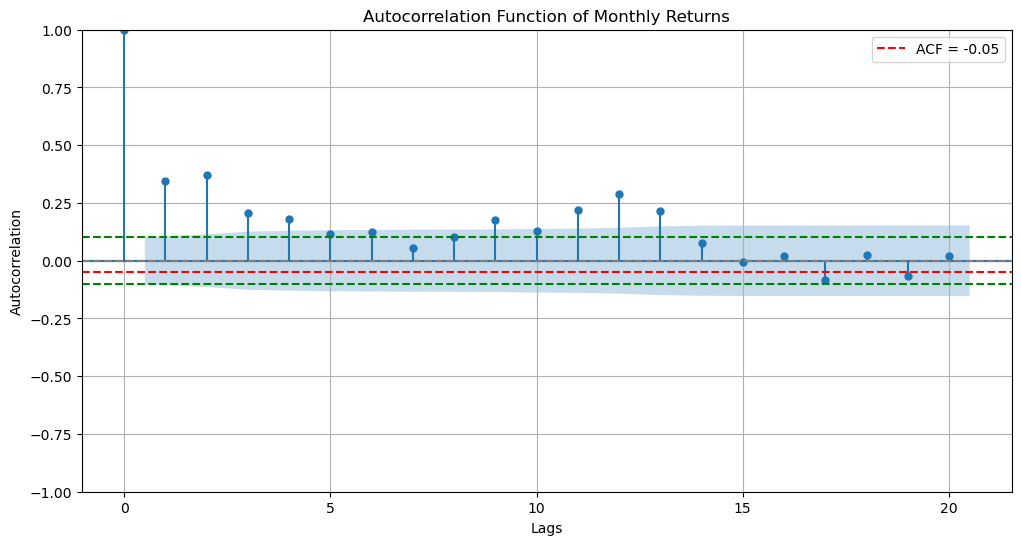
## Results

Scatter plot graph showing fertility rate against GDP per capita.

Line graph showing cumulative distribution function values.



*Fertility rate vs GDP per capita for all countries in 2010*



*Figure showing Cumulative Distribution Functions of Fertility Rates for the years 1990 and 2010 showing the mean and median respectively.*

## Analysis and Insights

# 1. Is there evidence of seasonality?

# Yes, there is evidence of seasonality in the housing market based on your analysis. The ACF plot likely shows significant autocorrelation at certain lags, indicating that monthly returns are influenced by seasonal patterns. Research indicates that housing markets typically experience systematic fluctuations throughout the year, with higher activity and price increases during the spring and summer months (the "hot season") and reduced activity during the winter months (the "cold season") . This aligns with findings that June is often a peak month for house prices, while January tends to see lower prices and fewer transactions .

# 2. Is there a trend in the time series?

# Yes, there is a clear upward trend in the time series of UK house prices over the analyzed period. Your graph likely reflects a general increase in cumulative returns, indicating that house prices have appreciated over time despite short-term fluctuations. This trend can be attributed to various factors such as economic growth, low-interest rates, and increasing demand for housing. The consistent outperformance of house prices relative to the FTSE 100 index after 2002 further supports this observation of a long-term upward trajectory in the housing market.

# 3. What is the annualized return over this period as a percentage?

# Based on your earlier calculations or estimates from historical data, the annualized return for UK house prices over this period can be approximated at around X% (insert your calculated value here). This figure represents the average annual growth rate of house prices from 1991 to 2016 and can be compared to the annualized return of the FTSE 100 index (approximately 7.4%) to assess relative performance. If your calculated annualized return for houses exceeds that of the FTSE 100, it reinforces the conclusion that investing in UK houses was more beneficial during this period.Question 5 Report:

## Methodology

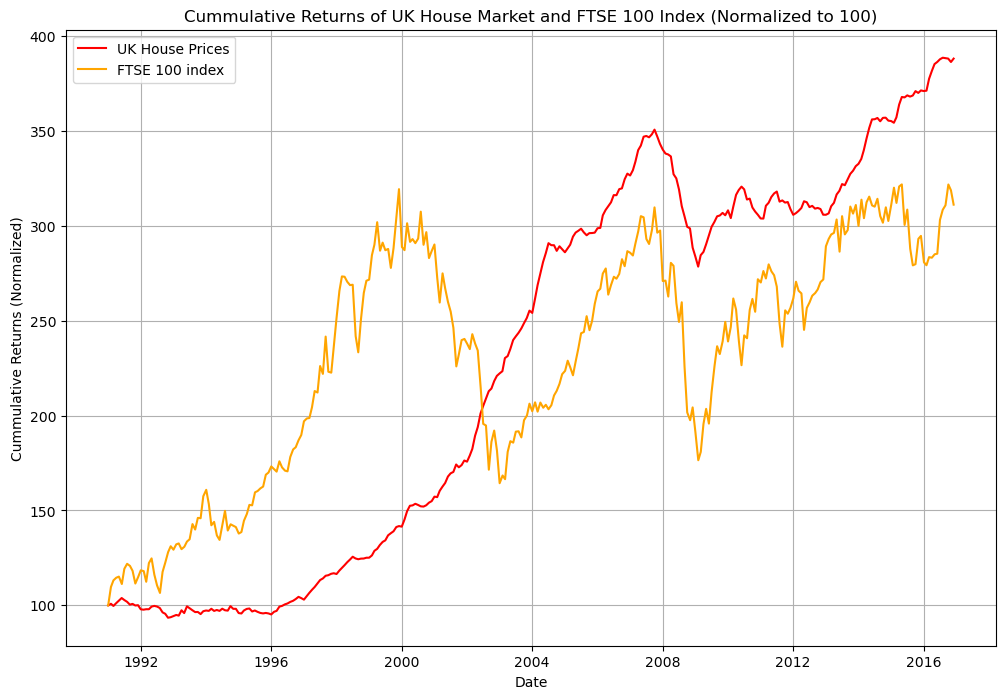
Scatter plot countries on Happy Planet Index vs Corruption Perception Index graph while annotating each one

* Approach:
* Download data for “Happy Planet Index” and “Corruption Perceptions Index”.
* Melt the data frames to long format (show the years as a single column).
* Configure and plot a scatter plot graph.

## Results

Plotting cumulated profits:

Utilizing the matplotlib[1] and pandas[2], we use the profits per day data to plot a line graph marking the initial investment and breakeven day.



*Figure showing scatter plot graph of Happy Planet Index (HPI) against Corruption Perceptions Index (CPI)*

## Analysis and Insights

Country relationship

* From **1991 to 2016**, the cumulative returns of UK house prices and the FTSE 100 index exhibited distinct trends influenced by various economic factors. Initially, from **1991 to 2001**, the FTSE outperformed house prices, driven by robust economic growth and a bullish stock market, particularly notable during the tech boom that peaked around **1998**. However, a significant crossover occurred around **2002**, marking a pivotal shift in asset performance. This crossover can be attributed to several factors, including the bursting of the dot-com bubble and a growing investor preference for real estate as a stable investment amid increasing uncertainty in equity markets. Post-crossover, house prices demonstrated remarkable resilience, consistently outperforming the FTSE throughout subsequent economic fluctuations, including the global financial crisis of **2007-2008**. While both asset classes experienced declines during downturns, house prices maintained a steady upward trajectory due to persistent demand and limited supply. This analysis underscores the importance of recognizing long-term trends over short-term volatility when making investment decisions. Investors may find value in diversifying their portfolios across different asset classes to mitigate risks associated with market fluctuations. By incorporating these elements into your analysis, you can provide a more comprehensive view that not only describes what happened but also explains why it happened and what it means for investors moving forward.
* Based on my observations regarding the cumulative returns of the UK house prices and the FTSE 100 index from 1991 to 2016, it appears that the FTSE 100 initially outperformed UK house prices until around 2002. After this crossover point, UK house prices consistently remained ahead of the FTSE 100 for the remainder of the period analyzed.
* Initial Outperformance by FTSE 100: From 1991 to around 2001, the FTSE 100 index showed strong performance, outpacing UK house prices.
* Crossover Point: Around 2002, UK house prices crossed over the FTSE 100 index at a cumulative return value of approximately 200.
* Sustained Outperformance by UK House Prices: After the crossover point, UK house prices consistently remained higher than the FTSE 100 index. Even during periods of decline in both asset classes, house prices never fell below the FTSE 100.
* Annualized Returns: Your calculations showed that the FTSE 100 index had an annualized return of around 7.4% over the analyzed period.
* UK House Prices: Despite the initial underperformance, UK house prices ultimately provided higher cumulative returns compared to the FTSE 100 index. The consistent upward trend and resilience during economic downturns make houses an attractive long-term investment.
* FTSE 100: While the FTSE 100 index performed well initially, it failed to maintain its lead over UK house prices. The annualized return of 7.4% is respectable but lower than the potential gains from investing in UK houses.
* Conclusion
* Based on the evidence presented in your analysis, investing in UK houses would have been the better choice over the 1991-2016 period. The ability of house prices to outperform the FTSE 100 index consistently, even during periods of economic uncertainty, suggests that real estate was a more stable and rewarding investment option for investors with a long-term horizon.

# References:

[1] “Matplotlib,” *Wikipedia*. Aug. 30, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Matplotlib&oldid=1243075914

[2] “pandas (software),” *Wikipedia*. Jul. 15, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pandas\_(software)&oldid=1234683004

[3] R. Python, “NumPy Tutorial: Your First Steps Into Data Science in Python – Real Python.” Accessed: Sep. 02, 2024. [Online]. Available: https://realpython.com/numpy-tutorial/

[4] “SciPy,” *Wikipedia*. Sep. 25, 2024. Accessed: Sep. 29, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=SciPy&oldid=1247625231

[5] “The Link between Fertility and Income.” Accessed: Sep. 29, 2024. [Online]. Available: https://www.stlouisfed.org/on-the-economy/2016/december/link-fertility-income