**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]

Tabulate – a python library that tabulates data to an output[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

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

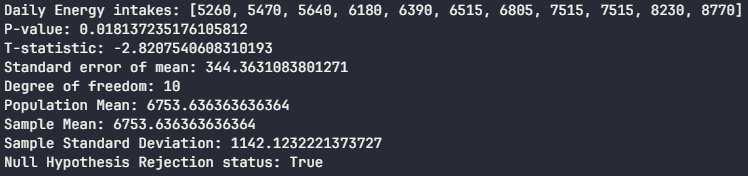
Making a scatter plot of malnutrition prevalence against GDP per capita.

Approach:

* Download data for “GDP per capita” and “Prevalence of underweight, weight for age”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as single column).
* Drop the NaN (Not a number) values.
* Configure and plot the graph.

## Results

Three scatter plot graphs showing Malnutrition prevalence against GDP per capita.



## Analysis and Insights

Expected Relationship

* The kind of relationship we expect is a negative correlation. This is where we should observe that as GDP per capita increases, prevalence of malnutrition should decrease[5].

Malnutrition prevalence against GDP per capita analysis

* It’s observed that it’s indeed true that countries with a higher GDP per capita are associated with a lower Malnutrition prevalence. Hence forming a negative relationship. However, studies have shown that a higher GDP per capita might not be the sole reason for a low malnutrition prevalence as other contributing factors might play a larger part in lowering malnutrition in affected countries[5]. One of these factors include education[5].
* When a country falls into a lower middle-income group, it’s observed that the country potentially has a higher Malnutrition Prevalence. This is seen with countries in the South Asian region.
* Majority of the countries with a high-income level are seen with high levels of GDP per capita and low levels of Malnutrition Prevalence. This backs the idea that a high GDP per capita will positively affect the Malnutrition level of a country.

# Question 2 Report:

## Methodology

Making a line graph showing the maximum and minimum prices of Wheat, Crude Oil and Gold.

Approach:

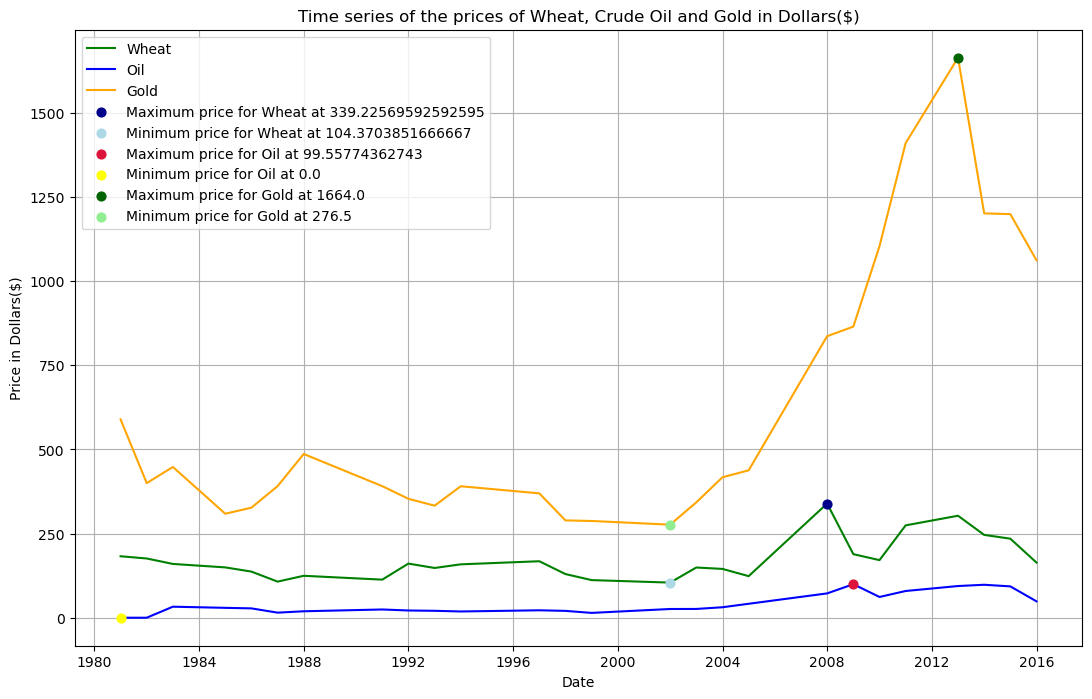
* Utilized Quandl API to download data for Wheat, Crude Oil and

Gold prices in dollars ($).

* Synchronized the time stamps.
* Formatted the data in the data frames.
* Set unique colors for each variable to be used in the graph.
* Configured and plotted graph while indicating the highest and lowest price for Wheat, Crude Oil and Gold.

## Results

Line plot graph showing the prices of Wheat, Crude Oil and Gold highlighting the lowest and highest prices.



*Line graph showing time series of the prices of Wheat, Crude Oil and Gold in dollars ($)*

## Analysis and Insights

Dramatic Gold Price Increase

* The graph shows that once Gold reached a low price of $276.5 (lowest between a 1980 and 2016), it begun to steadily rise, and a sharp increase is observed around 2009. This rise would be attributed to the fact that investors at that time thought it would be the safest option to keep their money safe during the ongoing financial crisis[6].
* During the same time Gold prices were increasing, Oil and Wheat were facing the opposite behavior. The demand for Oil and Wheat sharply reduced and this was primarily driven by the ongoing financial crisis, which led to decreased consumer spending and industrial activity.

# Question 3 Report:

## Methodology

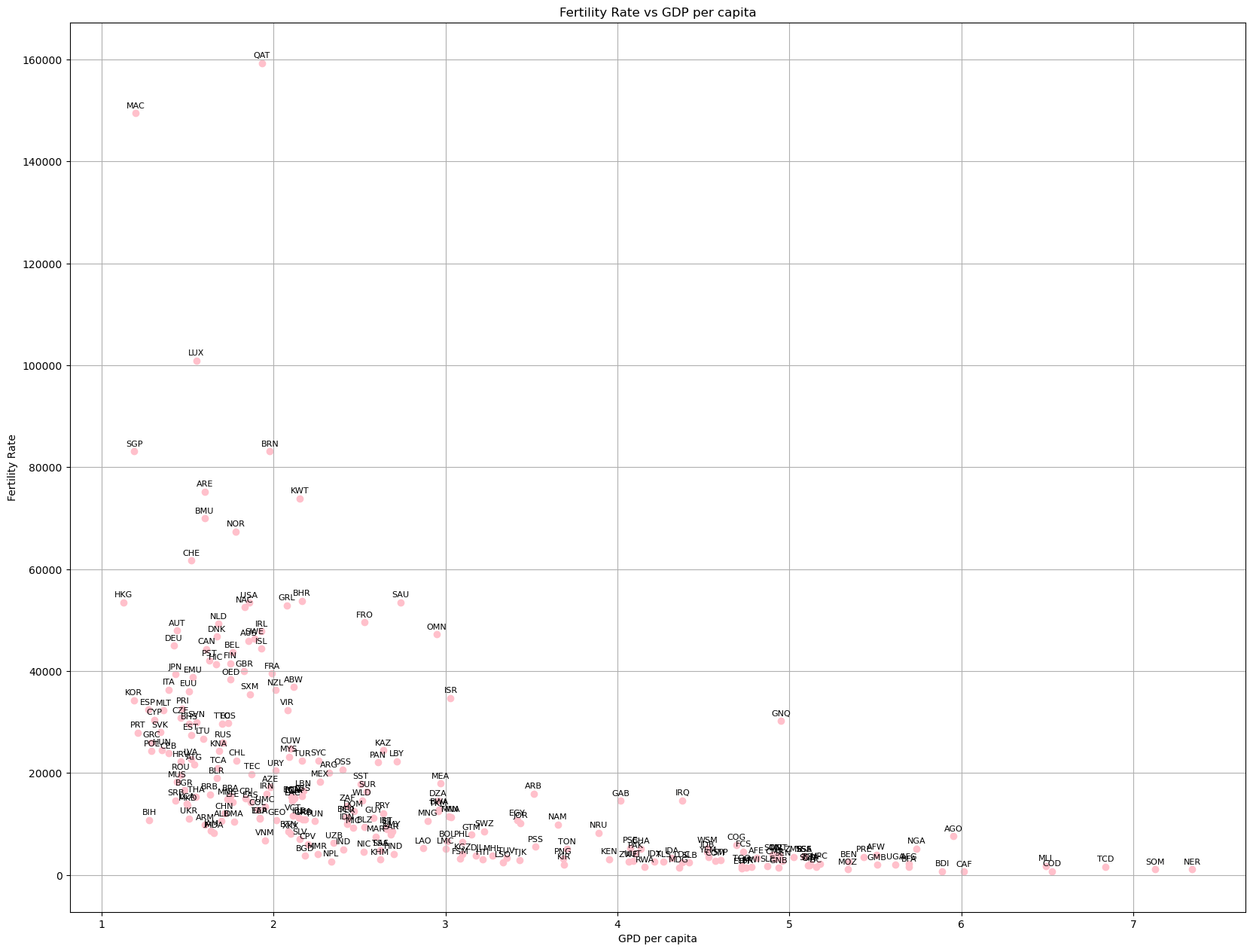
Using pandas library to calculate summary statistics[2]

Approach:

* Download data for “C02 emissions (metric tons per capita)” and “School enrolment, primary (% net)”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as a single column).
* Drop the NaN (Not a number) values.
* Calculate summary statistics.
* Display statistics in tabular format.

## Results

Finding Two tables showing summary statistics for C02 emissions and School enrollment for the year 2010.



*Figure showing C02 emissions (metric tons per capita) summary statistics for all countries for the year 2010*

## Analysis and Insights

C02 emissions

* The low mean signifies countries have a low C02 emission.
* The median is significantly lower than the mean which strongly suggests the presence of outlier countries with very high emissions that are pulling the mean upward.
* The relatively high Standard Deviation is likely due to the differences among countries in areas such as energy sources, industrialization and population.
* The table indicates low values for the 5th, 25th and 75th percentiles, which signifies that a large portion of countries have relatively low emissions. However, the 95th percentile value of 15.17 signifies that many countries have high C02 emissions which could call for immediate intervention for the affected countries.

School Enrollments

* The high mean value signifies that, on average, many countries had a lot of school enrollments for the year 2010. This signifies a positive trend in educational participation across the countries in the data.
* The median value suggests that majority of the countries had a high number of enrollments, so a possibility of an outlier is low. Although it doesn’t mean that there aren’t any outliers completely, there could still be some outliers but are not heavily influencing the mean.
* The low relative standard deviation indicates that there isn’t a huge variation from one country to another in terms of number of school enrollments. In other words, this means that most countries have similar enrollment figures.
* Countries in the 5th percentile have relatively low school enrollments whereas countries in higher percentiles (25th, 75th, 95th) signify that they are experiencing a high number of enrollments. This indicates a trend towards higher enrollments as you move up through the percentiles.

# Question 4 Report:

## Methodology

Utilizing numpy to produce cumulative distribution functions.

Approach:

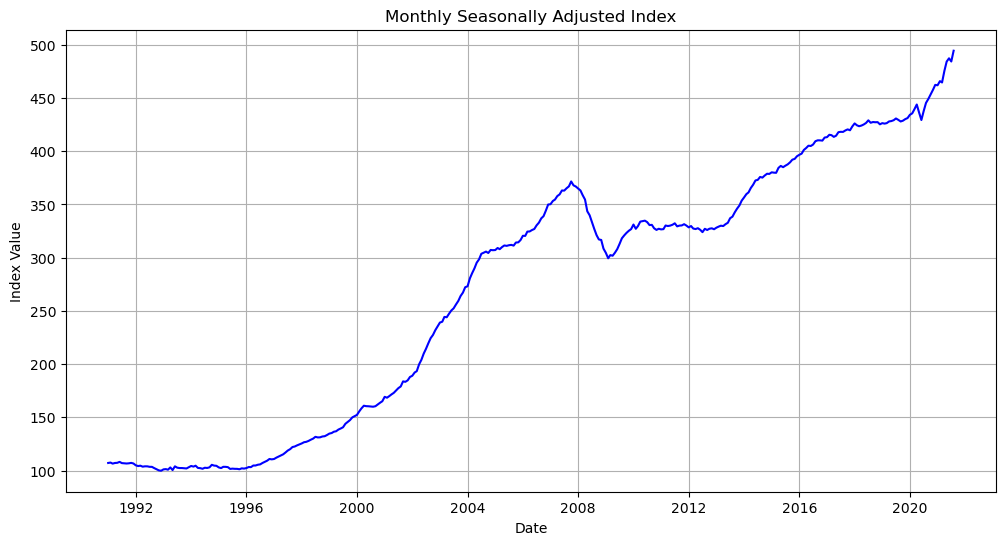
* Download data for “Fertility rate, total (births per woman)” and “GDP per capita (current US$)”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as a single column).
* Configure and plot a scatter plot graph (2010 data only).
* Sort the data frames.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

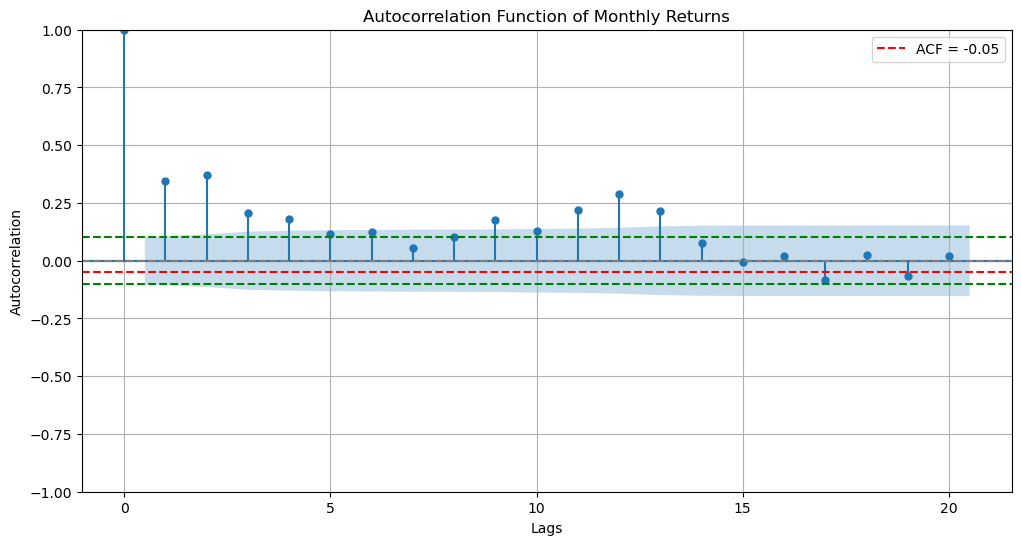
## Results

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

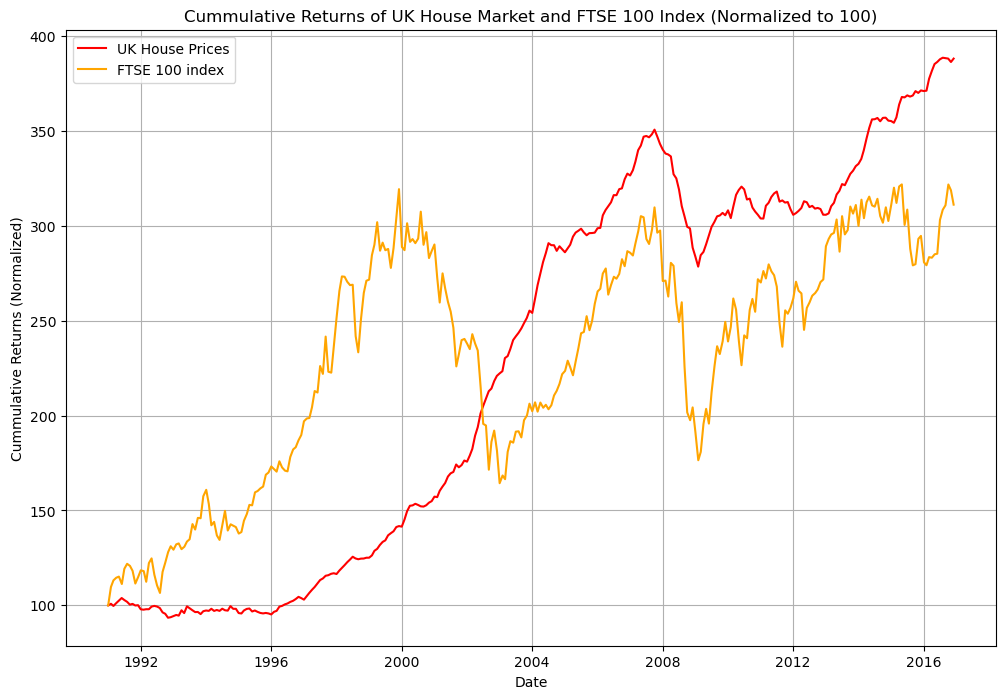
Scatter plot countries on Happy Planet Index vs Corruption Perception Index graph while annotating each one

* 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] *tabulate: Pretty-print tabular data*. Python. Accessed: Sep. 15, 2024. [OS Independent]. Available: https://github.com/astanin/python-tabulate

[5] N. Büttner, M. Heemann, J.-W. De Neve, S. Verguet, S. Vollmer, and K. Harttgen, “Economic Growth and Childhood Malnutrition in Low- and Middle-Income Countries,” *JAMA Netw. Open*, vol. 6, no. 11, p. e2342654, Nov. 2023, doi: 10.1001/jamanetworkopen.2023.42654.

[6] B. H. Program Producer Price Index, “Gold prices during and after the Great Recession,” Bureau of Labor Statistics. Accessed: Sep. 15, 2024. [Online]. Available: https://www.bls.gov/opub/btn/volume-2/gold-prices-during-and-after-the-great-recession.htm

[7] J. Lynch, R. Carver, and J. M. Virgo, “Quadrant Analysis as a Strategic Planning Technique in Curriculum Development and Program Marketing,” *J. Mark. High. Educ.*, vol. 7, no. 2, pp. 17–32, 1996.