DATA, INFERENCE

&

APPLIED MACHINE LEARNING

(COURSE 18-785)

ASSIGNMENT 3

Mark Iraguha

(miraguha)

Carnegie Mellon University Africa

Table of Contents

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].
- Statsmodels – a powerful python library that provides a wide range of statistical models from t-tests to chi-square and ANOVA tests useful in data analysis[5].

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.

```
Sample Mean: 6753.636363636364
Sample Standard Deviation: 1142.1232221373727
Standard error of mean: 344.3631083801271
T-statistic: -2.8207540608310193
Degree of freedom: 10
P-value: 0.018137235176105812
Conclusion: Null hypothesis (H0) rejected: The women's energy
 intake deviates systematically (either greater or less than the recommended 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

Statistical Test

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

Null hypothesis rejection

- The p-value obtained from the test is less than the significance level, so the null hypothesis was rejected.
- Rejecting the null hypothesis indicates significant evidence that supports the idea that women's energy intake systematically deviates from the recommended value of 7725kJ.

Insights

- 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.

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: 11.73775770205081
Degree of Freedom: 101
P-value (without rounding): 6.979768077580737e-21
P-value (rounded): 0.0000
Conclusion: Null hypothesis rejected - The GOES score in Ireland is
significantly higher than that in the Elsewhere group.
```

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

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 GOES score in Ireland is significantly higher than that in the Elsewhere group.

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.

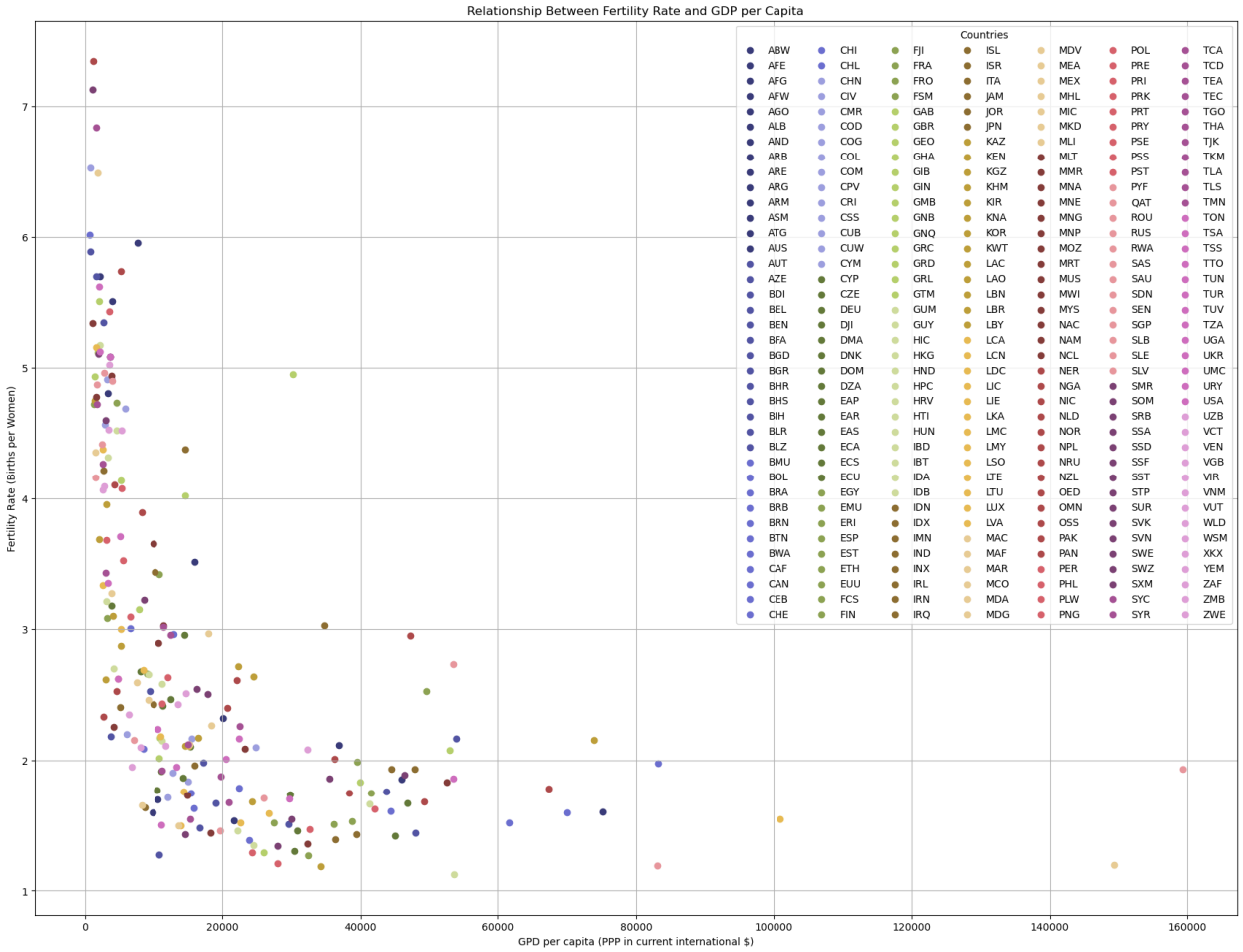

Figure showing the relationship between fertility rates and GDP per capita

```
Correlation Coefficient: -0.5171011715833221
```

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[6].
- 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 an 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 statsmodels to produce autocorrelation function (ACF) of the monthly returns.

Approach:

- Download data for "UK Monthly indices (Post '91)".
- Clean the data and filter the data frame.
- Configure and plot a time series graph for data up to 2016.
- Calculate the average monthly returns.
- Calculate the Autocorrelation Function (ACF) using the stattools module from the statsmodels library.
- Configure and plot bar graph.
- Calculate the annualized return as a percentage.

Results

Time series graph showing monthly UK house prices for the period 1991 - 2016.

Bar graph showing the Autocorrelation Function (ACF) of House Price Monthly Returns while showing the values of the ACF for lags of one up to 20.

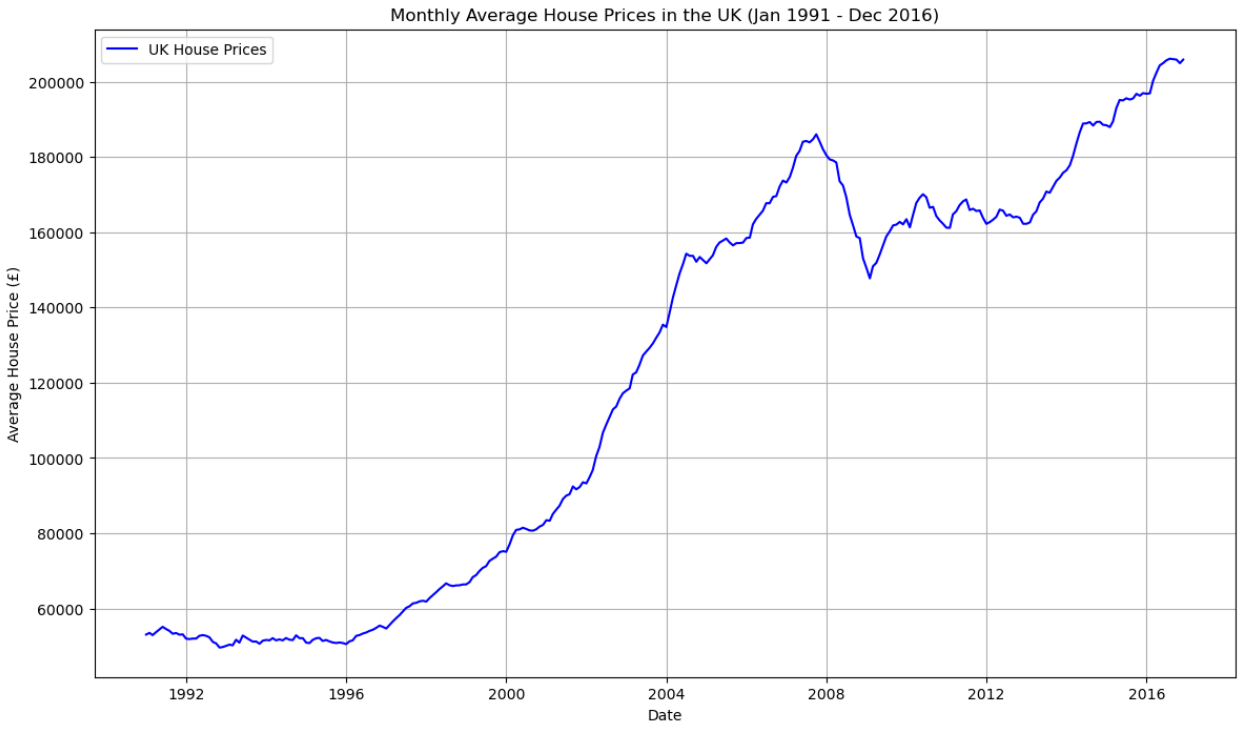

Figure showing the time series trend of monthly average house price in the UK

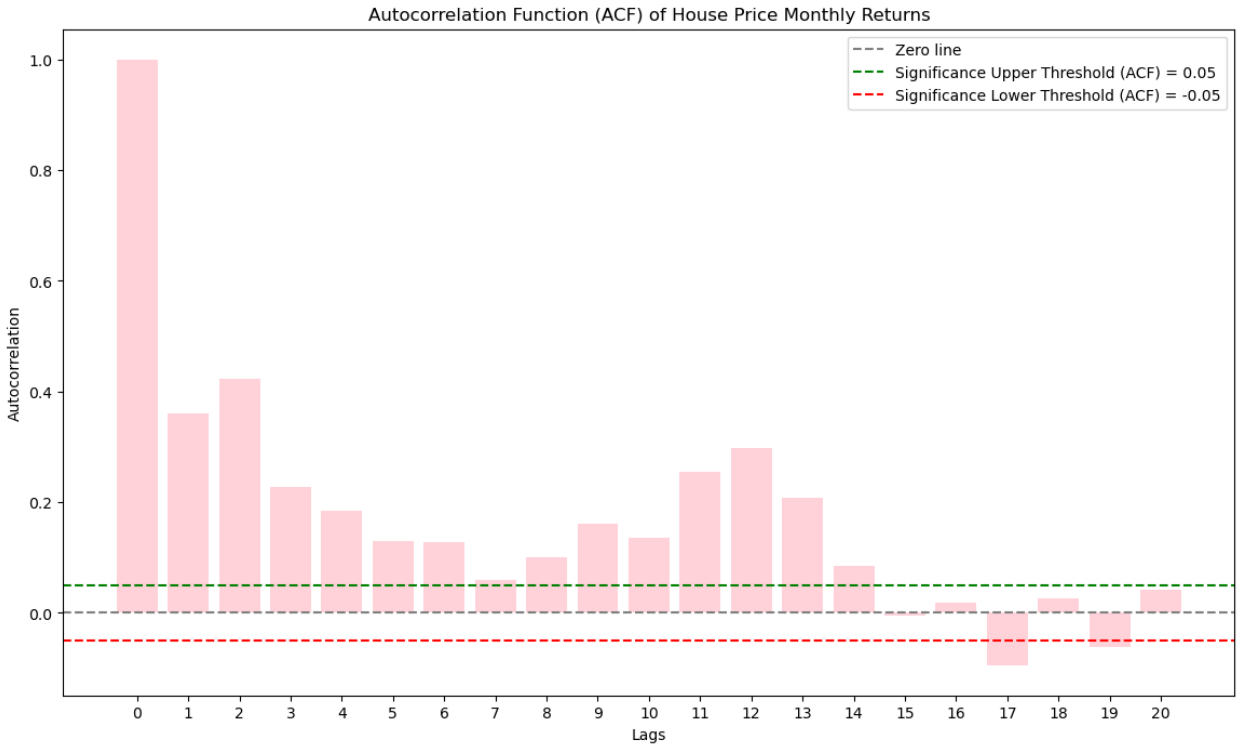

Figure showing Autocorrelation Function (ACF) of House Price Monthly Returns and the values of the ACF using horizontal lines that would correspond to a statistically significant result at p<0.05.

```
Annualized Return (as a percentage): 5.453020256940899%
```

Annualized return over the period of 1991 – 2016 expressed as a percentage.

Analysis and Insights

1. Is there evidence of seasonality?
   Yes, there is evidence of seasonality in the housing market. The ACF plot 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")[7]. This aligns with observation 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 period 1991-2016. The graph signifies 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 and increasing demand for housing.
   The consistent rise in house prices reflects broader economic conditions and demographic trends, such as urbanization and population growth. As demand continues to outstrip supply in many areas, particularly in urban centers, house prices are likely to maintain their upward trajectory[8].

3. What is the annualized return over this period as a percentage?
   Based on the earlier calculations, the annualized return for UK house prices over this period can be approximated at around 5.453%. This figure represents the average annual growth rate of house prices from 1991 to 2016.
   This annualized return is significant as it not only reflects the appreciation of property values but also serves as a benchmark for comparing real estate investments against other asset classes such as equities or bonds.

Question 5 Report:

Methodology

Plot time series graph showing cumulative returns from FTSE100 index and House market.

Approach:

- Download data for "FTSE100 Index" and "House Price Index".
- Clean the data and filter the data frames to the period 1991 – 2016.
- Sort the FTSE index data frame.
- Normalize the average house prices and FTSE100 index to start each at 100 and calculate the cumulative returns for each.
- Configure and plot a time series graph.
- Calculate the annualized return for FTSE index.

Results

Time series graph showing cumulative returns from the House market as well as the FTSE index with each starting at 100 after normalization.

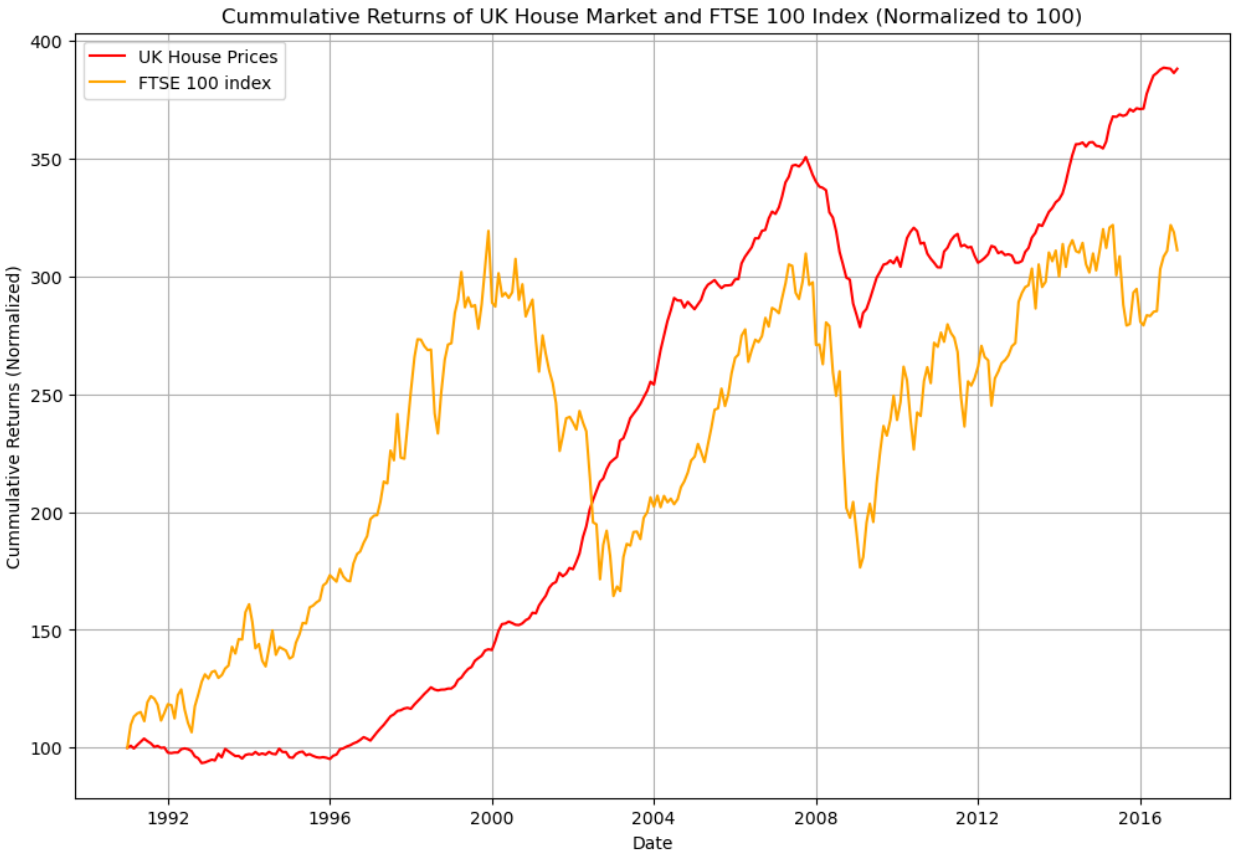

Figure showing scatter plot graph of Happy Planet Index (HPI) against Corruption Perceptions Index (CPI)

```
Annualized Return (as a percentage):4.645100329517793%
```

Annualized Return from the FTSE100 expressed as a percentage

Analysis and Insights

Annualized returns from the FTSE100

Based on the calculations performed, the FTSE 100 index had an annualized return of around 4.6451% over the period 1991 - 2016.

Would it have been better to invest in a UK house or the UK stock market over this period?

- To determine whether it was better to invest in a UK house or the UK Stock market, analysis can be performed on historical performance data while considering other factors that might have influences both markets.
- From 1991 to 2016, the cumulative returns of UK house prices and the FTSE 100 index showed distinct trends that could have been influenced by various economic factors. Initially, from 1991 to 2001, the FTSE outperformed house prices, particularly observed during the tech boom period that peaked around 1998.
- However, a significant crossover occurred around 2002 when UK house prices surpassed the FTSE 100 index. UK house prices reached a cumulative return value of approximately 200. This crossover was influenced by bursting of the dot-com bubble and a growing investor preference for real estate as a stable investment amid market uncertainties[9].
- After the crossover, UK house prices consistently remained higher than the FTSE 100 index. Even during economic fluctuations, such as the global financial crisis of 2007-2008, house prices maintained a steady upward trend due to persistent demand and limited supply[10]. House prices never fell below the FTSE 100.
- On the statistical side, annualized returns of FTSE 100 index were approximately 4.6451% during this period. The UK House prices had an annualized return of around 5.453%. Since the annualized return for UK houses exceeds FTSE 100, the conclusion would be that investing in a UK house was better during this period.

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] P. Mania, "What Is Statsmodels in Python? The Ultimate Guide," Python Mania. Accessed: Sep. 30, 2024. [Online]. Available: https://pythonmania.org/what-is-statsmodels-in-python-the-ultimate-guide/

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

[7] L. R. Ngai and S. Tenreyro, "Hot and Cold Seasons in the Housing Market," *Am. Econ. Rev.*, vol. 104, no. 12, pp. 3991–4026, Dec. 2014, doi: 10.1257/aer.104.12.3991.

[8] "Navigating the Housing Market: A Seasonal Perspective." Accessed: Sep. 30, 2024. [Online]. Available: https://www.nar.realtor/blogs/economists-outlook/navigating-the-housing-market-a-seasonal-perspective

[9] "What the Tech Bubble in 2000 May Tell Us About the Stock Market Today," Investopedia. Accessed: Sep. 30, 2024. [Online]. Available: https://www.investopedia.com/what-2000-tech-bubble-tells-us-about-the-stock-market-today-8684561

[10] "The 2008 Financial Crisis Explained," Investopedia. Accessed: Sep. 30, 2024. [Online]. Available: https://www.investopedia.com/articles/economics/09/financial-crisis-review.asp