

DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 18-785)

Names: Lucie Niyomutoni

AndrewID: lniyomut

**USED LIBRARY**

* **Pandas**
* **Matplotlib**
* **Scipy**
* **numpy**
* **statsmodels.tsa.stattools**

# QUESTION 1

In this question, the average daily energy intake of 11 women is compared to the recommended value of 7725 kJ through a t-test. The data is assumed to be normally distributed, and the sample mean, standard deviation, standard error of the mean (SEM), t-statistic, degrees of freedom, and p-value are required to be calculated. A decision is made about whether the null hypothesis—that the mean equals 7725 kJ—can be rejected, and the appropriate tail test (left-tailed, right-tailed, or two-tailed) is determined based on the results.

## STEPS

The following are the steps taken to solve the question

1. **Formulating Hypotheses**

The first step is to establish both the **null hypothesis and** the **alternative hypothesis**:

* The **null hypothesis (H₀)** assumes that there is no significant difference between the women’s energy intake and the recommended value of **7725 kJ**. In other words, the average energy intake is equal to 7725 kJ.

**H₀= 7725 kJ**

* The **alternative hypothesis (H₁)** assumes that the women’s energy intake does differ from the recommended value, meaning it could be either higher or lower.

**H₁: mu # 7725 kJ**

1. **The Significance Level**

A standard significance level of alpha = 0.05 was used, indicating a willingness to accept a 5% chance of rejecting the null hypothesis when it is true. If the p-value from the test is below 0.05, the null hypothesis is rejected; otherwise, it is not rejected.

1. **Determine the Type of Test**

Since the alternative hypothesis suggests that the energy intake could be either higher or lower than the recommended 7725 kJ, a **two-tailed** test was used. This test checks for any significant difference in either direction from the hypothesized mean

1. **Calculate the Sample Mean**

The sample mean was calculated to understand the average daily energy intake for the 11 women. The sample mean represents the central tendency of the given data.

1. **Identify the Population Mean**

The population mean is the recommended energy intake of 7725 kJ. This value was used to compare with the sample mean to test if the sample mean is statistically different.

1. **Calculate the Sample Standard Deviation and Standard Error of the Mean (SEM)**

The **sample standard deviation** measures how much the individual values deviate from the sample mean, reflecting the variability in the data. the **standard error of the mean (SEM)** wascalculated, which shows how much the sample mean is expected to vary from the true population mean. It is calculated by dividing the sample standard deviation by the square root of the sample size.

1. **Compute the t-Statistic**

The formula for the t-statistic is applied to calculate how many standard errors the sample mean deviates from the population mean. This calculation provides the test statistic, which is then compared to the critical value in a two-tailed test.

1. **Determining Degrees of Freedom**

The degrees of freedom (df) for this test are calculated by subtracting 1 from the sample size. This value is used to find the critical t-value from the t-distribution table.

1. **Calculate the p-Value**

The p-value indicates the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed, assuming that the null hypothesis is true. If the p-value is less than the significance level (alpha = 0.05), the null hypothesis is rejected.

## RESULT

1. **Sample Mean**: The average energy intake for the 11 women is **6753.636363636364 kJ**
2. **Sample Standard Deviation:** The variation in the energy intake values is **1142.1232221373727 kJ.**
3. **Standard Error of the Mean (SEM):** The expected variability of the sample mean is **344.3631083801271 kJ**.
4. **t-Statistic**: the calculated t-statistic is **-2.8207540608310198**, showing how far the sample mean is from the population mean in terms of standard errors.
5. **Degrees of Freedom:** The degrees of freedom for this test is **10**.
6. **p-Value**: The p-value obtained is **0.018137235176105812**, which is less than the significance level **(alpha = 0.05).**

**Appropriate test**

As it was said in the steps taken**,** Since the alternative hypothesis suggests that the energy intake could be either higher or lower than the recommended 7725 kJ, **a two-tailed** **test** is the appropriate test to use here. This test checks for any significant difference in either direction from the hypothesized mean

Since the p-value is lower than 0.05, the **null hypothesis is rejected.** This means the data provides sufficient evidence to suggest that the average energy intake of the women differs significantly from the recommended 7725 kJ.

## INSIGHT

The results indicate that the **average energy intake of the 11 women is significantly lower than the recommended value of 7725 kJ**.

1. **Health Considerations**: A lower-than-recommended energy intake can lead to undernutrition, which may result in fatigue, weakened immunity, and reduced muscle mass. If this trend persists, it could contribute to long-term health problems.

2. The results suggest that the women may not be consuming enough energy to meet their daily requirements. Dietary adjustments might be necessary to increase caloric intake, which could involve more nutrient-dense foods such as whole grains, protein-rich foods, and healthy fats.

On a larger scale, if this pattern of insufficient energy intake is common among a wider population, it may indicate a need for nutritional education programs. Public health initiatives could focus on improving awareness about balanced diets and caloric needs, especially among women.

# QUESTION 2

The question examines whether Guinness served in Irish pubs is significantly better than Guinness served elsewhere, using a Guinness Overall Enjoyment Score (GOES). The mean GOES score for pints in Ireland is 74, compared to 57 elsewhere, based on sample sizes of 42 and 61, respectively. The difference is analyzed using a t-test to determine whether it is statistically significant or due to random variation. The type of test (one-sample, two-sample, or paired) and the direction of the tail for the test are considered. The t-statistic, p-value, and conclusions about statistical significance are also required.

## STEPS

1. **Hypothesis Formulation**

* Null Hypothesis( ): There is no significant difference between the mean GOES scores for Guinness served in Ireland and elsewhere.

: =

* **Alternative Hypothesis ():** The mean GOES score for Guinness served in Ireland is significantly higher than that for Guinness served elsewhere. :>

1. **Significance Level**

significance level of was used, which means that the null hypothesis will be rejected if the p-value is less than 0.05

1. **Appropriate test selection**

Since the means of two independent samples (Guinness served in Ireland vs. elsewhere) are being compared, a two-sample t-test is the appropriate test.

The samples are independent, as the pints consumed in Ireland and elsewhere are separate.

A **right-tailed** test is required because the objective is to determine whether Guinness served in Irish pubs is perceived to taste significantly better than pints served elsewhere. Since a specific interest is placed on whether the score in Irish pubs is greater, the test focuses on this one-sided direction, making a right-tailed test the appropriate choice

1. **Calculating degree of freedom**

The degree of freedom (df) for a two-sample t-test were calculated as follow:

Where

So ,

1. Calculating t-statistic

The formula for the t-statistic in a two-sample t-test is :

,

Where: is the mean GOES score for Ireland

is the mean GOES score for Elsewhere

is the standard deviation for Ireland

is the standard deviation for elsewhere

is the sample size for Ireland

is the sample size for elsewhere

Then the standard error (SE) and statistics were calculated by using and

1. **Determining the P-value**

**To find the p-value for the obtained t-statistic, python library was used**

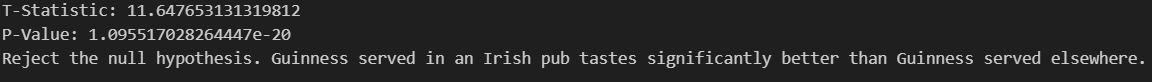
## RESULT

**Appropriate t-test**

Since the means of two independent samples (Guinness served in Ireland vs. elsewhere) are being compared, a **two-sample** t-test is the appropriate test. The samples are independent, as the pints consumed in Ireland and elsewhere are separate

A **right-tailed** test is required because the objective is to determine whether Guinness served in Irish pubs is perceived to taste significantly better than pints served elsewhere. Since a specific interest is placed on whether the score in Irish pubs is greater, the test focuses on this one-sided direction, making a right-tailed test the appropriate choice

After going through all the above steps and running the program the results were



The calculated t-statistic is approximately 11.64765. This high value indicates a strong difference between the means of the two groups being compared (Guinness served in Ireland vs. elsewhere).

The resulting p-value is approximately 0.00000. This p-value is significantly lower than the chosen significance level of 𝛼=0.05

Based on these results, the null hypothesis is rejected, concluding that there is a statistically **significant** difference in the mean GOES scores between Guinness served in Irish pubs and those served elsewhere. Specifically, the mean score for Guinness served in Ireland (74) is considerably higher than the average score for Guinness served elsewhere (57).

Since the p-value is significantly less than the alpha level of 0.05, **the null hypothesis is rejected**. This indicates that there is a statistically significant difference in the GOES scores, supporting the claim that Guinness served in an Irish pub is perceived to taste significantly better than Guinness served elsewhere.

## INSIGHT

1. **Statistical Significance:**

* The extremely low p-value (essentially 0) strongly supports the alternative hypothesis that the mean GOES score for Guinness served in Ireland is higher than that served elsewhere.
* With a t-statistic of 11.64, there is a clear and **significant** difference between the two groups, far beyond random variation.



The mean GOES score in Ireland is 74, which is substantially higher than the score of 57 for Guinness served elsewhere. This suggests a noticeable difference in the perceived quality or taste of Guinness depending on location.

The strong deviation between these two means (17 points) indicates that this perception is not marginal but a meaningful difference

1. **Practical implications**

* The significant difference in enjoyment scores suggests that the cultural context and ambiance of Irish pubs play a crucial role in the experience of drinking Guinness. Factors such as traditional music, the presence of knowledgeable staff, and a sense of community may enhance the overall enjoyment of the beverage
* The results may influence consumer preferences, suggesting that people may be willing to travel to Ireland for a more enjoyable Guinness experience. This insight can encourage travel and tourism sectors to promote trips centred around authentic culinary experiences.

# QUESTION 3

In this Question, data from the World Bank Indicators for 2013 is used to examine the relationship between the total fertility rate (births per woman) and GDP per capita (PPP) in current international dollars. A graph will be created, with each country represented by a single data point and properly labeled. Additionally, the correlation coefficient will be calculated, and an interpretation of its meaning will be provided, focusing on how these two factors are related.

## STEPS

1. Firstly, data was downloaded from the World Bank Indicators. The data included two key indicators:

**Fertility rate, total (births per woman**) and **GDP per capita (PPP) (current international $)**

1. Ensured that the necessary Python libraries were imported. Specifically, **pandas** for data manipulation and **matplotlib** for data visualization. Additionally, **xlrd** to was installed to facilitate the reading of Excel files
2. The panda’s library was utilized to read the Excel files containing the fertility and GDP data.
3. Relevant data for the year 2013 was extracted from both datasets.
4. A scatter plot was generated to visually represent the relationship between GDP per capita and fertility rates.
5. The correlation coefficient was computed to assess the strength and direction of the relationship between fertility rates and GDP per capita.

## RESULT

After going through all the above steps the results were a scatter plot illustrating the relationship between GDP per capita and fertility rates for various countries in 2013, and The correlation coefficient.

**Correlation Coefficient**: After performing the calculations, the correlation coefficient was found to be **-0.5171011715833227**. This indicates a moderate to strong negative correlation between GDP per capita and fertility rates. A correlation coefficient of -0.517 indicates a moderate strength of the relationship. While it is not a strong correlation, it is significant enough to suggest that changes in one variable are associated with changes in the other

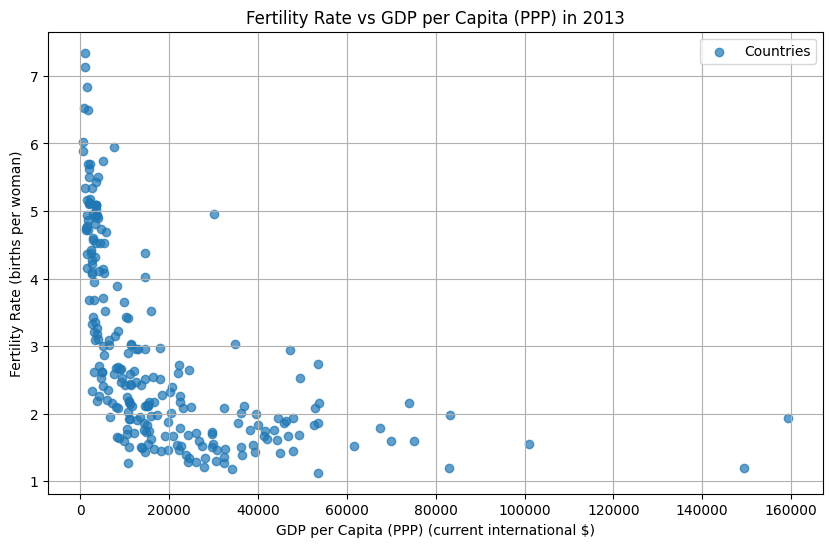


Figure Fertility rate vs DGP per capita in 2013

The graph above illustrates the relationship between fertility rate (total births per woman) and GDP per capita (current international $), with fertility rate represented on the vertical (Y) axis and GDP per capita on the horizontal (X) axis.

Notably, within the GDP per capita range of $0 to $60,000, many countries exhibit fertility rates between 1 and 3 births per woman. In contrast, countries with GDP per capita between $0 and $20,000 display a wider range of fertility rates, spanning from 1 to 7 births per woman.

This trend clearly indicates that as GDP per capita increases, fertility rates tend to decrease. In other words, wealthier countries generally have fewer children per woman. This observation suggests a negative correlation between GDP per capita and fertility rate: as nations become richer, the average number of children born to women decreases.

## INSIGHT

Countries with higher GDP per capita tend to have better access to education, healthcare, and family planning services. As economies grow, people are more likely to have access to contraception and reproductive health care, enabling them to make informed decisions about family size. For instance, nations like Germany and Japan, which have high GDP per capita, also have low fertility rates, as families often prioritize education, careers, and a higher quality of life over having many children.

One of the key factors contributing to lower fertility rates in wealthier countries is increased access to education, especially for women. Educated women often marry later, have fewer children, and invest more in each child's development and education.

Higher GDP per capita also leads to improved healthcare systems, which reduce infant mortality rates. As a result, families no longer feel the need to have many children to ensure that some survive. This phenomenon is more common in lower-income countries, such as Sub-Saharan Africa.

In high-income countries, people often prioritize their careers and economic opportunities over having large families. The desire to maintain a certain lifestyle and higher living standards can lead couples to delay having children or limit family size

# QUESTION 4

The question asks to analyse monthly average house price data in pounds (£) from January 1991 to December 2016. The data should be downloaded from a specified source, and a time series graph must be created and labelled. An autocorrelation function (ACF) of the monthly returns, should be constructed, displaying values for lags from one to twenty using a bar graph. Statistically significant results need to be marked with horizontal lines. The analysis should check for seasonality and trends in the data, and the annualized return over this period should be calculated as a percentage.

## STEPS

1. The process began with the downloading of data on monthly average house prices
2. Next, all necessary libraries were imported including acf for autocorrelation and o other essential libraries such as numpy, pandas, matplotlib, and stats.
3. The data was loaded using the pandas function to read the Excel file.
4. The unnamed date column was converted to a proper datetime format and set as the index for time-series analysis, restricting the data to the period from January 1991 to December 2016.
5. The average house price column was extracted for analysis.
6. The time series of house prices for the period of January 1991 to December 2016 was plotted.
7. Monthly returns were calculated using the **pct**\_**change** function, which works like the formula , where represents the average house price at time . These returns help assess short-term price changes and volatility.
8. The ACF of monthly returns for lags 1 to 20 was computed and plotted to examine correlations between past and future price movements, as well as to check for any seasonal patterns. Statistical significance was tested using horizontal lines representing a 95% confidence interval.
9. Finally, the annualized return over the period to quantify the long-term growth of house prices as a percentage was calculated.

## RESULT

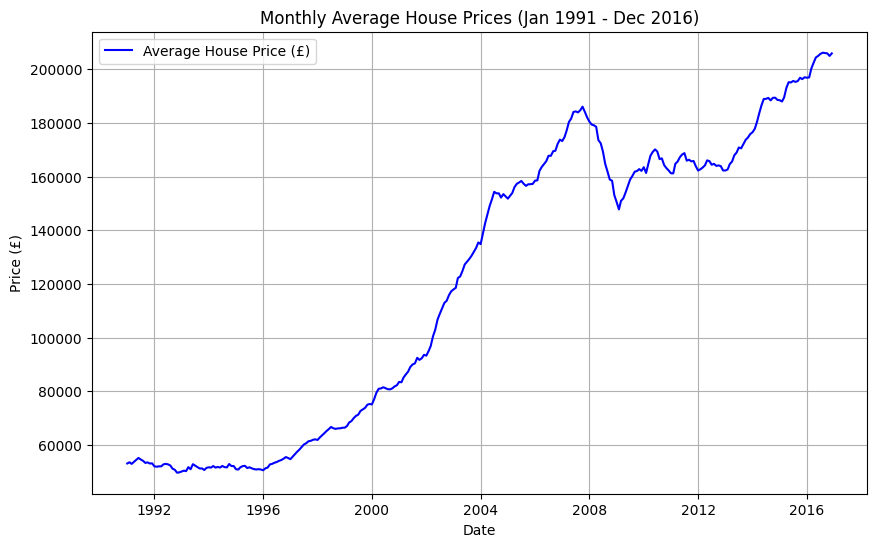


Figure Monthly Average House Prices

The time series plot displays the monthly average house prices from January 1991 to December 2016. The X-axis represents time, specifically the years between 1991 and 2016, while the Y-axis shows the average house price in British pounds (£). The graph reveals a steady upward trend, indicating that house prices generally increased over this 26 years period. However, there are moments of fluctuation, such as around 2008, which likely reflects the impact of the global financial crisis. Despite these occasional drops, the overall trend points to long-term growth in the housing market, meaning that houses became more expensive over time.

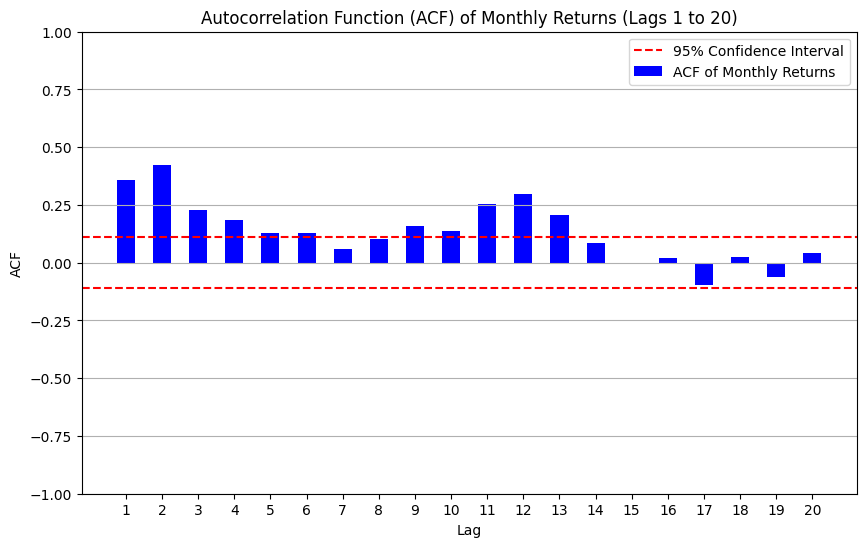


Figure ACF of Monthly Returns

The bar chart of the Autocorrelation Function (ACF) shows how much past monthly returns are related to future returns over different time periods, called "lags." The X-axis represents these lags, which go from 1 to 20 months, while the Y-axis shows how strong the relationship (autocorrelation) is between returns at each lag. Most of the bars fall within the red dashed lines, which represent the 95% confidence interval. This means the relationships between past and future returns are not strong enough to be statistically significant. In simple terms, the chart suggests that previous changes in house prices don’t have a strong influence on predicting future changes for lags up to 20 months. There’s also no clear seasonal pattern, and no strong correlation is observed.

The annualized return for UK house prices from January 1991 to December 2016 was **5.35423853535919%**. This means that, on average, house prices in the UK increased by **5.35423853535919%** each year over the 26 years period. Instead of looking at price changes for individual months or years, the annualized return gives us a single percentage that reflects the average yearly growth across the entire period. This shows a steady increase in house prices over time, even though there may have been ups and downs in certain years.

**Is there evidence of seasonality?**

From the ACF plot, there is no clear evidence of seasonality in the monthly returns of house prices. Seasonality would show up as a repeating pattern in the autocorrelation values at specific time lags (for example, every 12 months if there was a yearly cycle). However, in this case, the autocorrelation values for lags 1 to 20 are not statistically significant, meaning that past monthly price changes do not consistently predict future changes. This suggests that house prices do not follow a predictable seasonal pattern each month.

**Is there a trend in the time series?**

Yes, the time series plot of house prices from 1991 to 2016 shows a clear upward trend. Over this 26 years period, house prices steadily increased, despite some short-term fluctuations like during the 2008 financial crisis. This positive trend indicates long-term growth in the housing market, which can be attributed to factors like increased demand, inflation, or economic expansion over time.

**What is the annualized return over this period?**

The annualized return for house prices from 1991 to 2016 is approximately **5.35%.** This means that, on average, house prices increased by **5.35423853535919%** each year over the entire 26 years period. The annualized return gives a clear picture of the overall growth rate, smoothing out the short-term fluctuations and showing the consistent long-term growth in the housing market.

## INSIGHT

Housing Market Stability and Long-Term Investment:

The sustained growth in house prices between 1991 and 2016 suggests that the UK housing market has historically been a stable long-term investment. For investors looking to build wealth over time, real estate could continue to be a solid asset class. However, this growth was likely supported by favourable economic conditions, such as low interest rates and government incentives. Investors today need to be mindful of potential changes in these factors, such as rising interest rates or economic slowdowns, which could affect future market performance.

Since there are no significant seasonal patterns, housing market prices seem to be more influenced by major economic factors like inflation, employment rates, and fiscal policy changes. Policymakers and developers should prioritize these broader economic trends when making decisions about housing supply, affordability measures, or regulations. Moreover, prospective homeowners and investors should monitor government housing policies, interest rates, and inflation forecasts, as these are likely to impact house prices more than seasonal variations.

# QUESTION 5

In the question, the FTSE100 index data from January 1, 1991, to December 31, 2016, is to be loaded, and cumulative returns for both the housing market and the FTSE100 are plotted on the same graph, starting at a normalized value of 100 in January 1991. The average annualized return for the FTSE100 is then calculated, and a comparison is made to determine whether investing in UK housing or the UK stock market would have been better decision during this period

## STEPS

In this analysis, the performance of UK house prices was compared to the FTSE100 index over the period from January 1, 1991, to December 31, 2016. The following steps were undertaken:

1. **Data Loading**: Monthly average house price data was imported from an Excel file, while FTSE100 data was loaded from a CSV file. The house price data's column was renamed for clarity.
2. **Data Filtering**: Both datasets were filtered to include only the dates within the specified range. The average house price and adjusted closing prices for the FTSE100 were retained.
3. **Cumulative Return Calculation**: Monthly returns were calculated for both datasets using the percentage change method. These returns were then cumulatively summed to obtain cumulative returns, which were normalized to start at 100 in January 1991 for comparative purposes.
4. **Plotting Results**: A graph was created to visualize the normalized cumulative returns of house prices against the FTSE100 index over the given period. The results were clearly labelled, and a horizontal line at the 100 mark was added for reference.
5. **Average Annualized Return Calculation**: The average annualized return was calculated for both the FTSE100 and house prices. The formula used was:

1. **Comparison and Conclusion**: Finally, the average annualized returns for the FTSE100 and the average house prices were compared. Based on this analysis, it was determined whether it would have been better to invest in the UK stock market or the housing market during the specified period

## RESULT

A graph with blue line and orange line

Description automatically generated

Figure House prices vs FTSE100

The graph compares the cumulative returns of the UK housing market and the FTSE100 index from January 1, 1991, to December 31, 2016. The x-axis represents the time, spanning from 1991 to 2016, while the y-axis indicates the cumulative return percentage, starting at 100 for both investments. The blue line represents the returns from the housing market (house prices), while the orange line shows the returns from the FTSE100 index. Both lines start at the same point, meaning that they were worth the same at the beginning.

Over the years House prices, represented by the blue line, increased significantly, especially between 1996 and 2007. This rise was likely due to strong demand for housing, economic growth, and low-interest rates during that time. However, after reaching a peak just before the financial crisis, house prices dropped sharply around 2008. This decline lasted until about 2009, reflecting the broader economic struggles. Following this downturn, house prices began to recover slowly, continuing to rise and reaching almost 15,000 by 2016.

On the other hand, the FTSE 100 index, exhibited a much steadier and slower increase throughout the same period. It remained below 2,500, indicating that stock market returns were not as high as those of the housing market during these years. While there were some fluctuations in the index, it did not experience the same dramatic growth that house prices did, especially leading up to the financial crisis. Overall, this graph highlights how the housing market showed stronger performance and resilience compared to the stock market during the period from 1991 to 2016.

**Average Anualized Return for FTSE100 = 4.46%**

The average annualized return for the FTSE100 from January 1, 1991, to December 31, 2016, was about 4.46% per year. This means that if someone invested in the FTSE100 at the beginning of this period and kept their money invested until the end, their investment would grow by around 4.46% each year on average. For example, if they invested £100, after one year, they would have £104.46. This average return shows that even though the stock market exhibited a much steadier and slower increase during these years, the overall trend was still positive, suggesting that investing in the FTSE100 could be a reasonable choice one can make.

After analysing and comparing the average annualized returns for both FTSE100 and UK Monthly Average House prices, it can be seen **that it would have been better to invest in a UK house rather than the UK stock market over the period from January 1, 1991, to December 31, 2016**. The analysis shows that house prices steadily increased, while the FTSE100 index had slower increase. By the end of 2016, the cumulative return from the housing market was significantly higher than that of the FTSE100, indicating that investments in UK houses provided a more stable and profitable option during this timeframe.

## INSIGHT

The housing market bounced back quickly after the 2008 financial crisis, showing that investing in property can be safer when the economy is struggling. This makes real estate a good choice for people worried about losing money in stocks.

Housing prices usually go up over time, which helps keep your money's value. As prices for everyday things increase, home values tend to rise too, making real estate a smart way to protect your savings.

The FTSE 100 index has shown steady growth, which means that while investing in stocks might not give you high returns quickly, it can lead to more reliable growth over the long term. This steady growth is attractive for cautious investors who want to take fewer risks. The data suggests that people should think about spreading their investments across both real estate and stocks. This mix can help balance out the risks and rewards they face. Overall, looking at these trends can help people make smarter choices about where to put their money and plan.

# REFERENCES

1. <https://www.indeed.com/career-advice/career-development/how-to-calculate-annualized-return>
2. <https://www.open.edu/openlearn/science-maths-technology/data-analysis-hypothesis-testing/content-section-6>
3. <https://www.jmp.com/en_us/statistics-knowledge-portal/t-test.html>
4. <https://ncbi.nlm.nih.gov/pmc/articles/PMC5579465/>
5. <https://www.statology.org/calculate-degrees-of-freedom-for-t-test/#:~:text=Here%20is%20how%20to%20calculate,the%20total%20number%20of%20pairs>.
6. <https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/annual-return/>