Carnegie
Mellon
University
Africa

COURSE 18-785: DATA, INFERENCE & APPLIED MACHINE LEARNING

ASSIGNMENT 3

Nchofon Tagha Ghogomu

ntaghagh

September 30, 2024

LIBRARIES:

The following libraries were used:

- Math
- Pandas[1]
- Numpy[2]
- Pyplot from Matplotlib [3]
- Matplotlib.ticker from MultipleLocator - for defining axis spacing
- Quandl: The quandl library
- Tabulate: Library for table presentation
- Scipy: A library for scientific computing
- Statsmodels.graphics.tsaplots

Programming Language:

- Python

INTRODUCTION

This assignment was made of 5 practical questions that portray real application of data analytics on world data, from a variety of sources. For every question, we had to come up with strategies to analyse and draw the insights required. In summary, throughout this assignment, we get to:

- Explore and analyse world data centred around development.
- Explore using statistic concepts including hypothesis testing.
- Decision making based on these values.

Python was the programming language used and Jupyter notebook was the programming environment.

SOLUTIONS

Question 1:

1) Goal:

Determine if the women energy intake deviate systematically from recommended value. We will do this by coming up with a null and alternative hypothesis, setting up a significance level and computing statistical values.

2) Steps:

To arrive at the plots and insights, the following steps were used:

- The required data was imported into and array
- The null and alternative hypothesis was made
- Determine whether to use a left-tail, right-tail or two-tailed test.
- Calculate the sample mean, then proceed to calculate the degree of freedom, sample standard deviation and the standard error of mean (SEM).
- The t statistic and p value is calculated with the help of the imported Stats module.
- Make a conclusion based on these calculated values.

3) Results and observation:

The null hypothesis, $\mu0$:

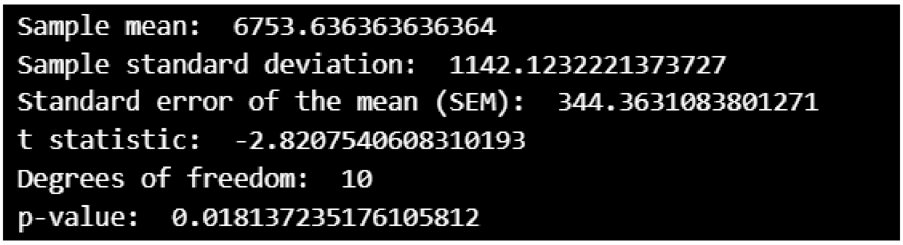
- The average energy intake of women in kJ is equal to the recommended value: 7725 kJ

Alternative hypothesis $\mu1$:

- The average energy intake of women in kJ is greater than or less than to the recommended value: 7725 kJ

In this experiment we will be using the two-tailed test because we want to check that the average energy intake is greater than or less than the set value. We test if $\mu < \mu0$ or $\mu > \mu0 =>$ $\mu \neq \mu0$. The significance level was set to $\alpha = 0.05$.

The Numpy function was used to calculate the sample mean and sample standard deviation, the standard error of the mean was calculated using its formular. With the help of the **Stats** module, we were able to calculate the t statistic and the p value as seen bellow: Based on the p value < $\alpha = 0.05$, we **reject** the null hypotheses.



```
Sample mean:  6753.636363636364
Sample standard deviation:  1142.1232221373727
Standard error of the mean (SEM):  344.3631083801271
t statistic:  -2.8207540608310193
Degrees of freedom:  10
p-value:  0.018137235176105812
```

Question 2:

1) Goal:

Determine if the difference in score value of pints consumed in Ireland and elsewhere is significantly different. In this process, we come up with a null and alternative hypothesis, setting up a significance level and computing statistical values.

1) Steps:

To arrive at the plots and insights, the following steps were used:

- The required data was imported as variables
- The null and alternative hypothesis was made
- We determine whether to use a left-tail, right-tail or two-tailed test.
- The t value is calculated using the
- The degree of freedom is calculated
- The p value is calculated.
- Make a conclusion based on these calculated values.

2) Results and observation:

The null hypothesis, µ0:

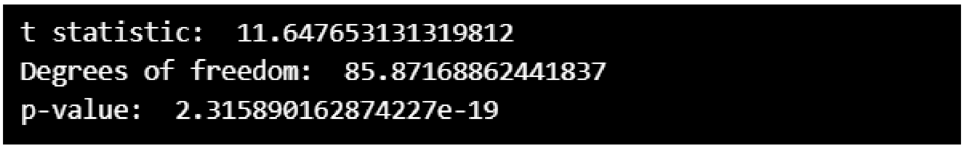
- The GOES score is the same in Ireland as it is elsewhere.

Alternative hypothesis, µ1:

- The GOES elsewhere is lower than it is in Ireland.

In this experiment, we use the one tailed test with α = 0.5 as given. We calculate the dof using the formular: $df = \frac{(\frac{s_1^2}{n1}+\frac{s_2^2}{n2})}{\frac{1}{n1-1}(\frac{s_1^2}{n1})^2+\frac{1}{n2-1}(\frac{s_2^2}{n2})^2}$

We observe that the p value is by far less than the level of statistical significance. Therefore, the null hypothesis is rejected.



```
t statistic:  11.647653131319812
Degrees of freedom:  85.87168862441837
p-value:  2.315890162874227e-19
```

Question 3:

1) Goal:

The goal is to study the relationship between Fertility rate, total (births per woman) versus GDP. per capita PPP (current international $), in 2013.

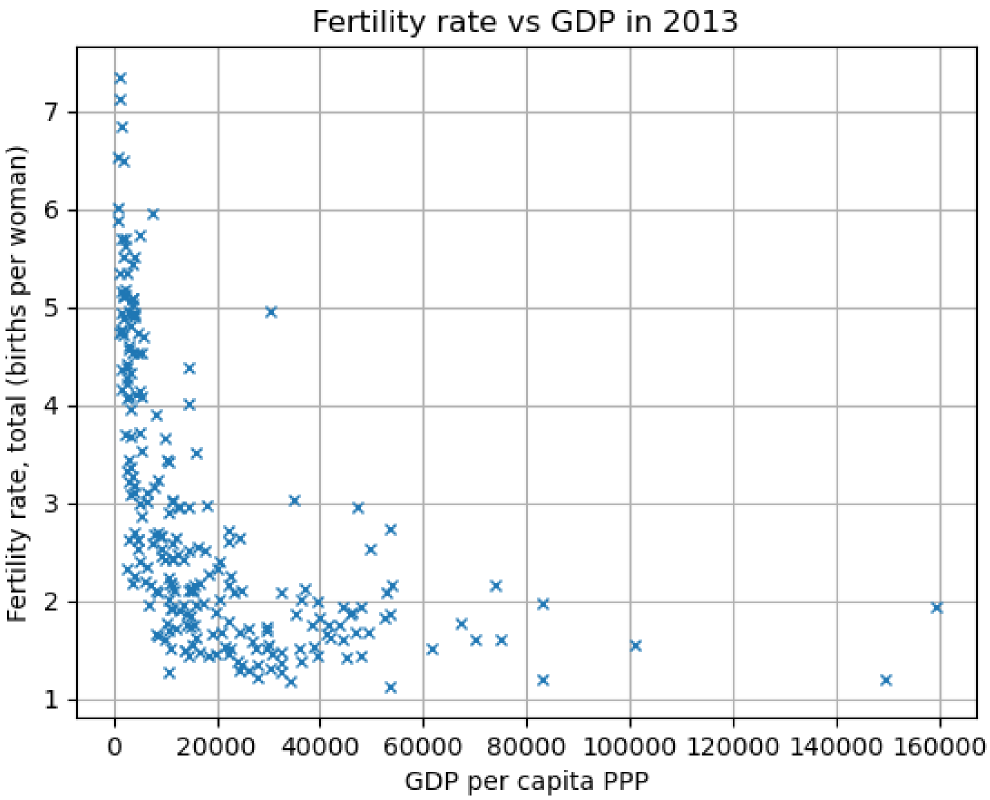
2) Steps:

To arrive at the plots and insights, the following steps were used:

- Both data were extracted into a data frame from the World Bank Indicator.
- The required year (2013) is extracted from the data frame.
- Predict relationship.
- We plot the GDP against fertility rate and observe the trend.
- Calculate the correlation coefficient using inbuilt corr function
- Make our conclusion.

3) Results and observation:

At the start, a negative correlation between both values was predicted. The plot obtained is seen bellow. The fertility rate and GDP per capita was observed to vary inversely and the calculated correlation value indicated a negative correlation between them. This means an increase in one is reflected by a decrease in the other.



Question 4:

1) Goal:

Plot the time series of average housing prices, calculate monthly and cumulative monthly return, the autocorrelation function, annualised return and using the values of lags, show the correlation between the actual value and that of previous years.

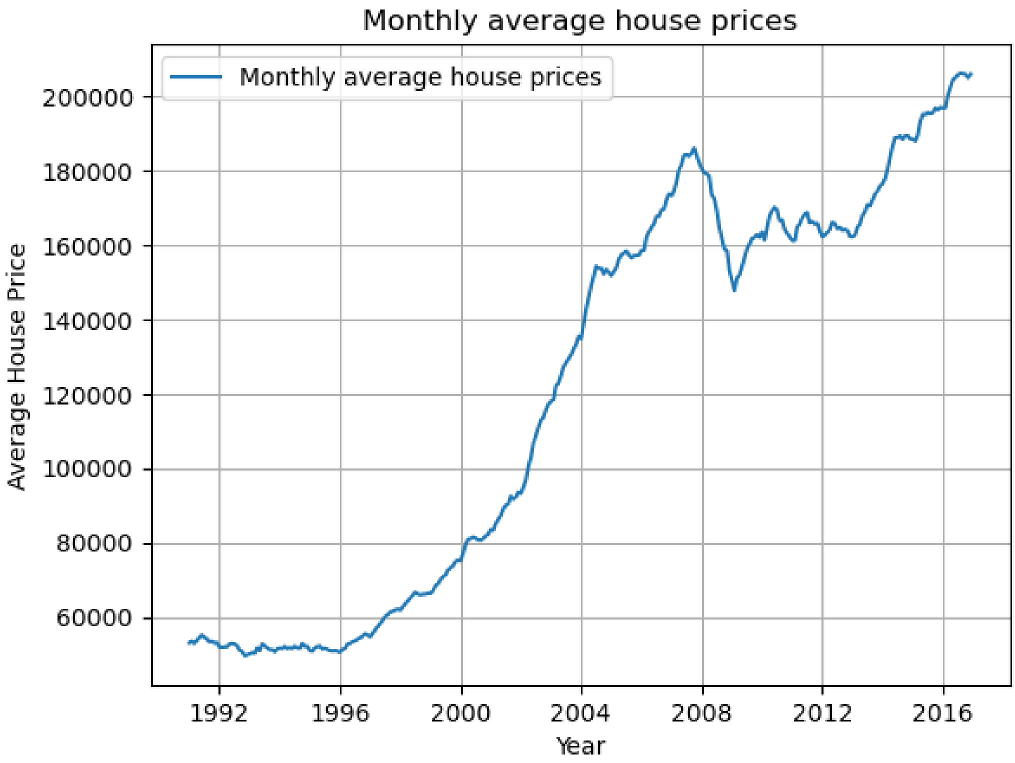
2) Steps:

To arrive at the plots and insights, the following steps were used:

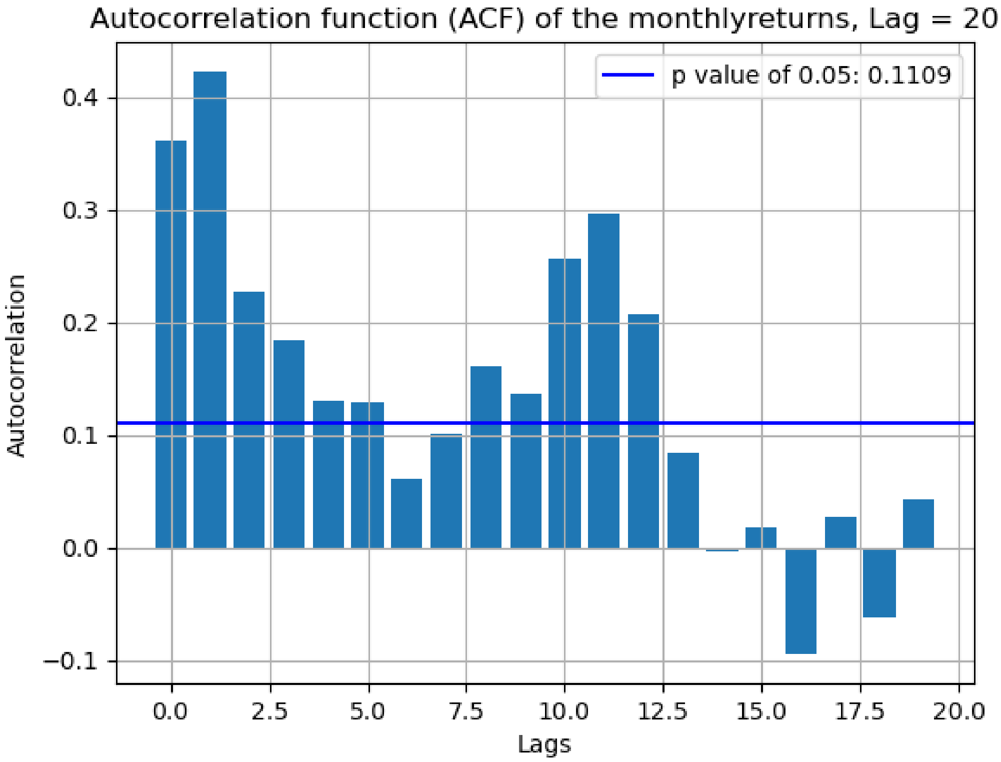
- Dataset is downloaded and extracted into data frame.
- A time series of the monthly average house prices is plotted.
- The trend is observed and commented (In this case an upward trend is identified)
- Monthly return is calculated using provided formular
- The ACF function using monthly return is plotted and the p value is indicated.
- The observation is explained
- The annualised return for this period is calculated.

3) Results and observation:

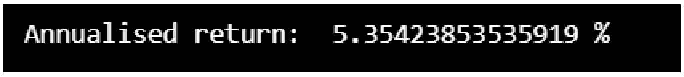
As observed from the time series, an upward trend is identified. There has been a steady increase since 1996 till about 2008, when it peaked and dropped significantly. There was fluctuation after the fall till 2012, after which there was a steady increase till 2016. This drop could have been caused by the great economic recession that happened during that period.



The ACF function plotted with lags tell us about the relationship between actual values and previous values. As observed from the autocorrelation function of monthly return for a lag of 20, it was observed that is the strongest correlation between value and two lags back. This correlation fades down to the 7 lag and rises to a value of about 0.3 in the12th lag. From the 13th lag, this correlation stays bellow the p value and even becomes negative.
The negative value corresponds to the point when prices experience a significant fall.



The annualised house price return as a percentage was calculated to 5.3% as seen the screen capture bellow. There has been a net positive return but not as significant.



```
Annualised return:  5.35423853535919 %
```

Question 5:

1) Goal:

Observe evolution and compare the cumulative returns of FTSE100 and house prices over a given period and state which of these would have been a better investment.

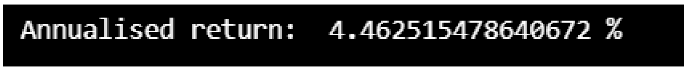
2) Steps:

To arrive at the plots and insights, the following steps were used:

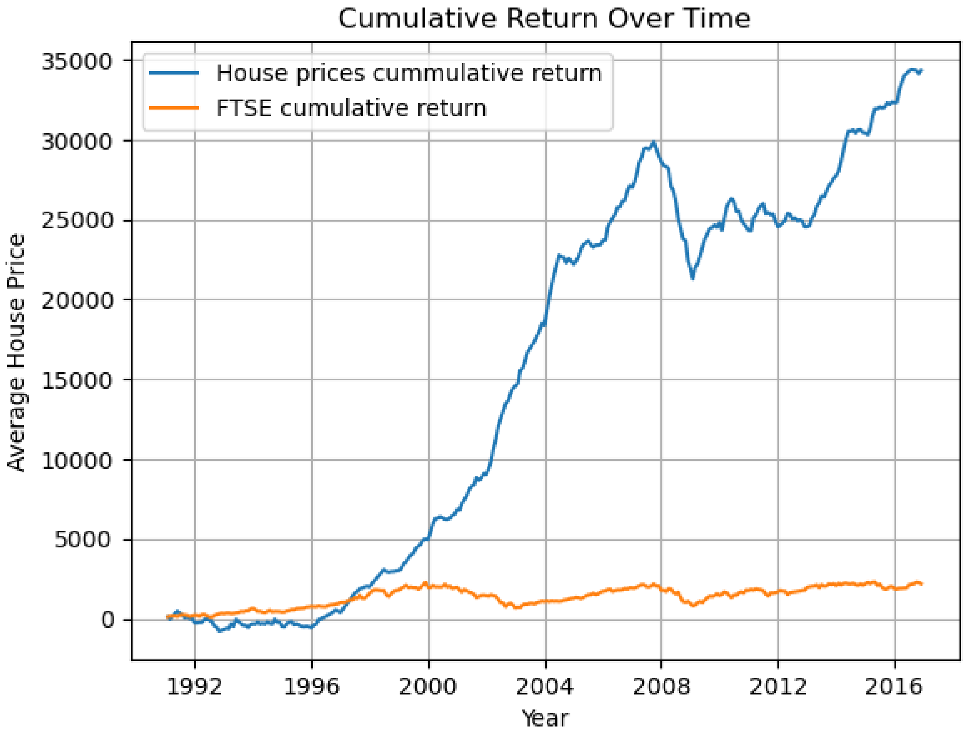
- Dataset is downloaded and extracted into data frame.
- The FTSE100, and house prices cumulative return is simple calculated.
- Both values are normalised to start at 100.
- A plot of both normalised values on the same graph is done
- The annualise return is calculated
- Based on observation, we make our comment.

3) Results and observation:

The normalised cumulative return for house prices portrays a significant increase in value within the given time frame compared to the FTSE. The annualised return for FTSE100 was calculated to be 4.46 %.



```
Annualised return:  4.462515478640672 %
```



Given the results, over this period, it was more profitable to invest in the UK house prices. Over this period, the house prices has observed a significant growth in returns.

REFERENCES

[1] “pandas documentation — pandas 2.2.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/index.html

[2] “NumPy -.” Accessed: Sep. 02, 2024. [Online]. Available: https://numpy.org/

[3] “Matplotlib documentation — Matplotlib 3.9.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://matplotlib.org/stable/