EVALYNE AMBALWA LWOBA

DATA, INFERENCE & APPLIED MACHINE LEARNING

ASSIGNMENT 3

ANDREW_ID : elwoba

Libraries Used

```
from scipy import stats
import numpy as nupy
import pandas as pd
import matplotlib.pyplot as plt
import math
import statistics
import numpy as np
from numpy import cov
from statsmodels.graphics.tsaplots import acf
from scipy.stats import pearsonr
from scipy.special import stdtr
from tabulate import tabulate
```

Question 1

Higher values of the t-value, also called t-score, indicate that a large **difference** exists between the two sample sets. The smaller the t-value, the more similarity exists between the two sample sets. A large t-score indicates that the groups are different. A small t-score indicates that the groups are similar.

T tests are used to the similarity between two samples groups. A higher t -value indicates that there is a difference between the two samples while a lower t value indicates that there is a similarity between the two sample groups.

In this sample data of 11 women, we are required to test if the proposed mean is valid or true. My hypothesis statements are:

H_0 : mean = 7725

H_1 : mean != 7725

I used a two tailed test to check whether the sample mean is significantly greater or lower than the mean of the population. This will test both ends of the curve being that the samples are normally distributed

The statistics are calculated using inbuild libraries(SciPy, math and statistics).

The statistics are:

Degree of freedom	Mean	Standard Deviation	SEM	T_Statistic	p-Value
10	6753.64	1142.12	344.363	-2.82075	0.0181372

Findings

If a p-value is less than 0.05, then it is statistically significant. It indicates a strong evidence against the null hypothesis. In this case, our p-value is less the given alpha of 0.05. We therefore reject the null hypothesis and accept the alternative hypothesis that the sample mean is not equal to 7725.

Question 2

To test if the difference between the two means is significant, I first created two hypothesis:

H_0 : Ireland mean is equal to elsewhere mean

H_1 : Ireland mean is greater than elsewhere mean

Since this are 2 difference population samples, I used a 2 sample test to test their means. Since we are testing if the sample population mean is greater than the population mean, we shall do a right tailed t-test.

To test the two hypothesis, I used the below formula to calculate the t statistic and p-value which will be used to evaluate the two population mean differences.

t-value formula[1]

$$t=(x_1-x_2)\sqrt{S_1^2/n_1+s_2^2/n_2}$$

Where:

X_1 is the Ireland mean = 74 X_2 is elsewhere mean = 57

S_1 is Ireland standard deviation = 7.4 S_2 is elsewhere standard deviation = 7.1

N_1 is Ireland population = 42 N_2 is elsewhere population. = 61

I substituted the values provided into the formula and got my t statistic result.

To calculate the degree of freedom, I used the formula:

n1 – 1 where:

n1 is the size of the population sample.

To calculate the p-value, I used the formula

P_value = 2*stdtr(d_freedom, -np.abs(t_value))

The probability value(p value) is calculated by the degree of freedom of the sample population over the t-stistic value usint the stdtr() function from the pandas scipy library in python.

Results.

```
t_statistic =  11.647653131319812

p-value is 2.191034056528894e-20
```

Findings.

Since the p-value calculated above is less that the alpha value of 0.05, then the difference between the two means is significant.

We therefore, reject the null hypothesis.

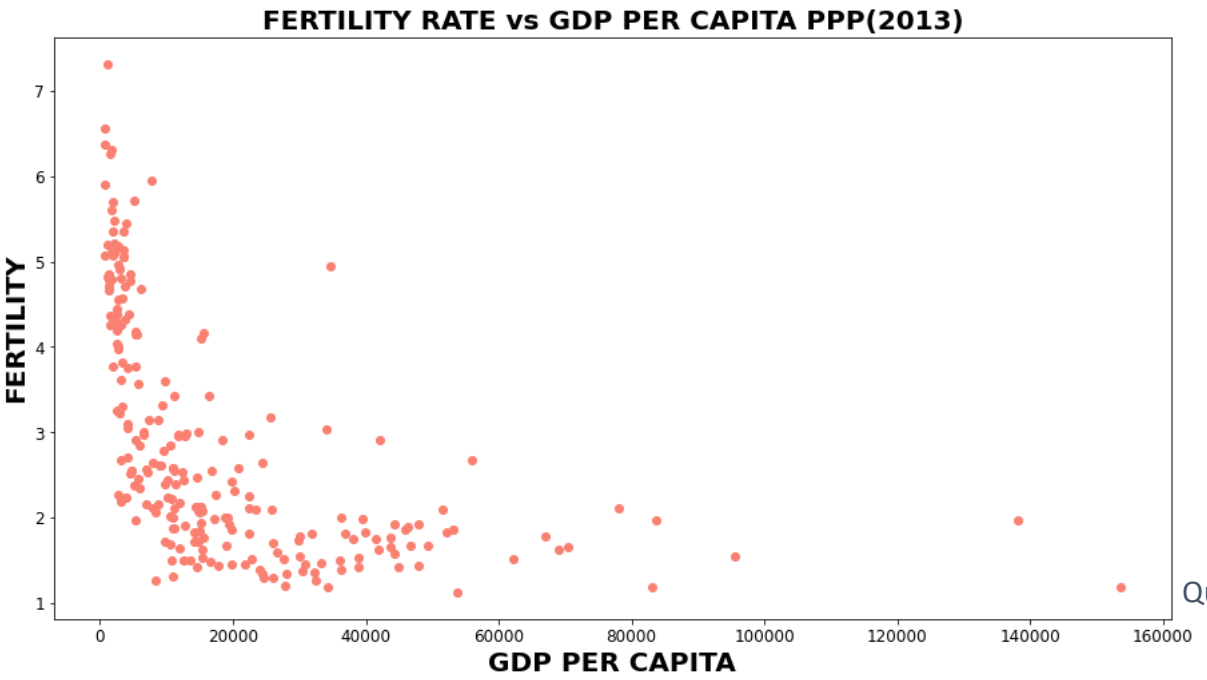
Question 3

I downloaded and uploaded the 2 data sets into my working environment. The two data sets (GDP per capita PPP and Fertility rate, total births per woman) have extra rows which I dropped by skipping the first three rows when reading the data files. I extracted the required columns (country name and 2013) for each data frame and formed 2 sub data sets which I then merged on the common column country name to form one data frame to use when plotting. I renamed the columns in my data frame to for ease in plotting then dropped all null values.

Part a

I did the plotting of the graph of Fertility rate against GDP using matplot library.

Results-Graph3a



Question 3a

Findings.

This graph shows a negative nonlinear relationship between GDP and fertility rate. As GDP grows, the fertility rate decreases. There are a few outlier countries in the graph.

Part b

According to Professor Baur, autocorrelation is used to check the strength of a linear relationship between two variables, in our case, GDP and Fertility rate. A coefficient value that is greater than 0 indicates a positive relationship while a value that is less than 0 indicates a negative relationship.

To compute the autocorrelation for 2013 fertility rate, I used the inbuild Pearsons function imported from scipy.stats library.I passed in the columns from my data set as the variables and got a correlation value.

Result:

```
Pearsons correlation: -0.527
```

Findings:

This results show a negative correlation meaning, the two variables are changing in the opposite directions. From our data set, as GDP grows, fertility rate decreases.

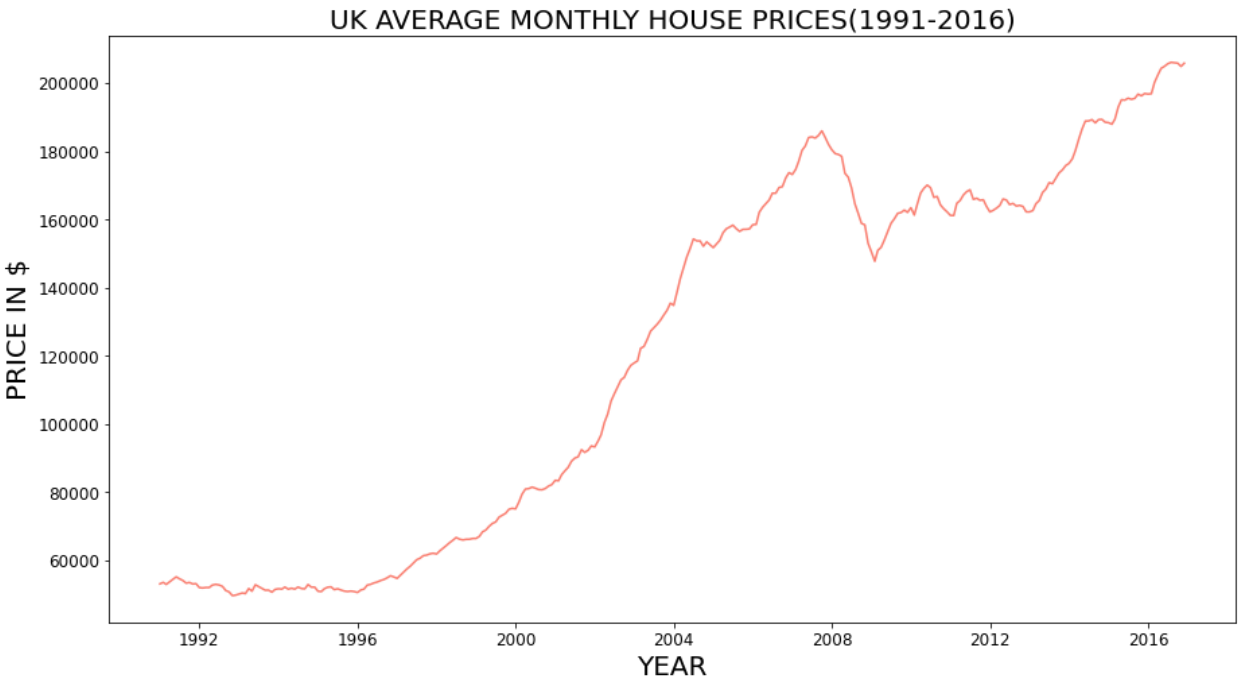
Question 4

I downloaded and saved my data files into my working folder.

To plot a graph of 1991 to 2016, I extracted the data of the required time frame through row index and saved it in a new data frame. Then I plotted the UK monthly prices time series for 1991 to 2016.

Part 4a

Results-Graph4a



Question 4a

Findings

The graph shows a positive trend in the house prices from 19991 to 2016. The prices was maintained at a steady rate of below $60000 from 1991 to 1996 where we see a gradual increase in prices to a high of

$375 in 2008. A sudden dip in the prices is noted in 2008 to a low of $300 in 2010 after which the prices became to increase again though with bits shakes.

Part 4b

To calculate monthly returns I used the provided formula below. In python this is made possible by using the shift () method. From statsmodels.graphics library, I imported acf function which I passed in the housing monthly returns data and specified number of lags required to generate the acf array values to plot the graph.

$$r(t) = [p(t)/p(t-1)]-1$$

To calculate annualized returns I created a function which took the data frame as arguments. The actual calculations is done used the following formula[3].

```
annual_return=series.add(1).prod() ** (12 /count_months) – 1
```

To plot a bar graph, I used 2 variables; the acf array calculated using the above formula and range (the length of the acf array then specified the color.

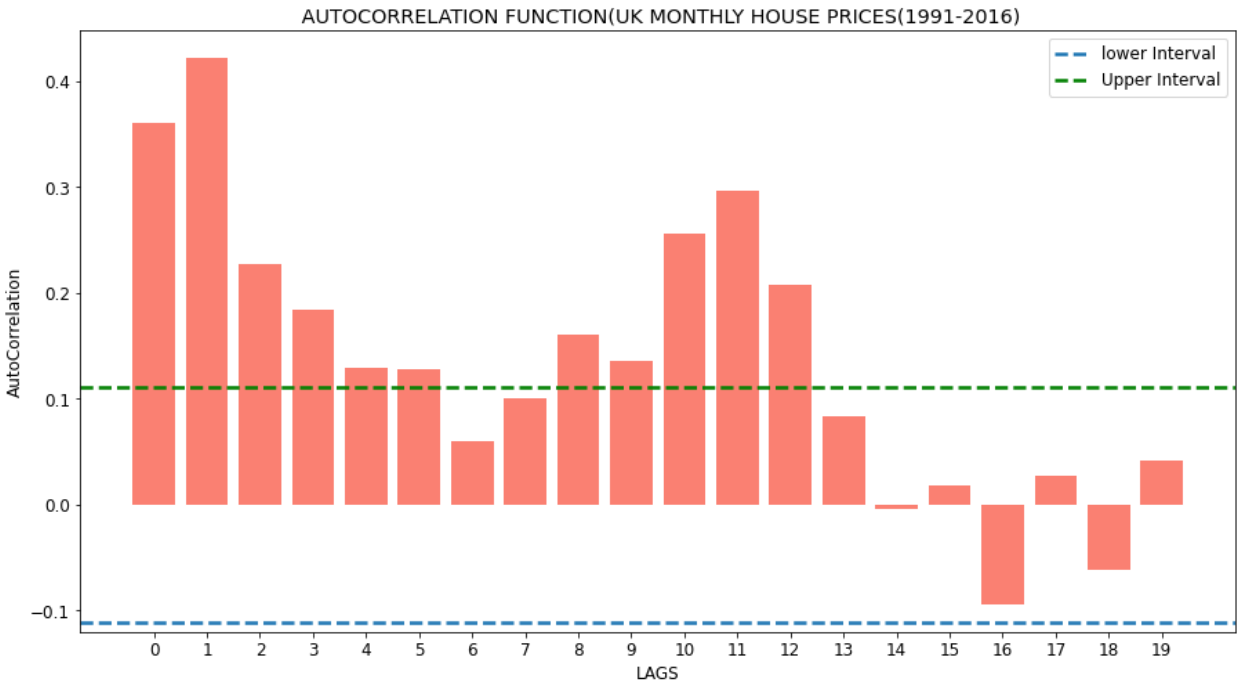
To indicate the ACF at $p<0.05$, I calculated the confidence intervals by using the provided formula below:

$$\pm 1.96 / \sqrt{n}$$

Where n is the size/ length of the data., the negative results will be the lower interval while the positive will be the upper interval for 0.95% confidence.

The resulting intervals(0.111, -0.111) were then used to plot the horizontal line using the axhline() method.

Results Graph 4b:



Question 4b

Findings.

From the autocorrelation graph above, we can see high coefficienct in lags number 2, followed by lags number 1,12. Lags number 17 had the lowest coefficient. Lags 5 and 6 have partial correlation.

From the graph I observe that there is seasonality where the prices peak on lag 1, then dops until lag 12 where it peaks again then starts to drop again.

Question 5

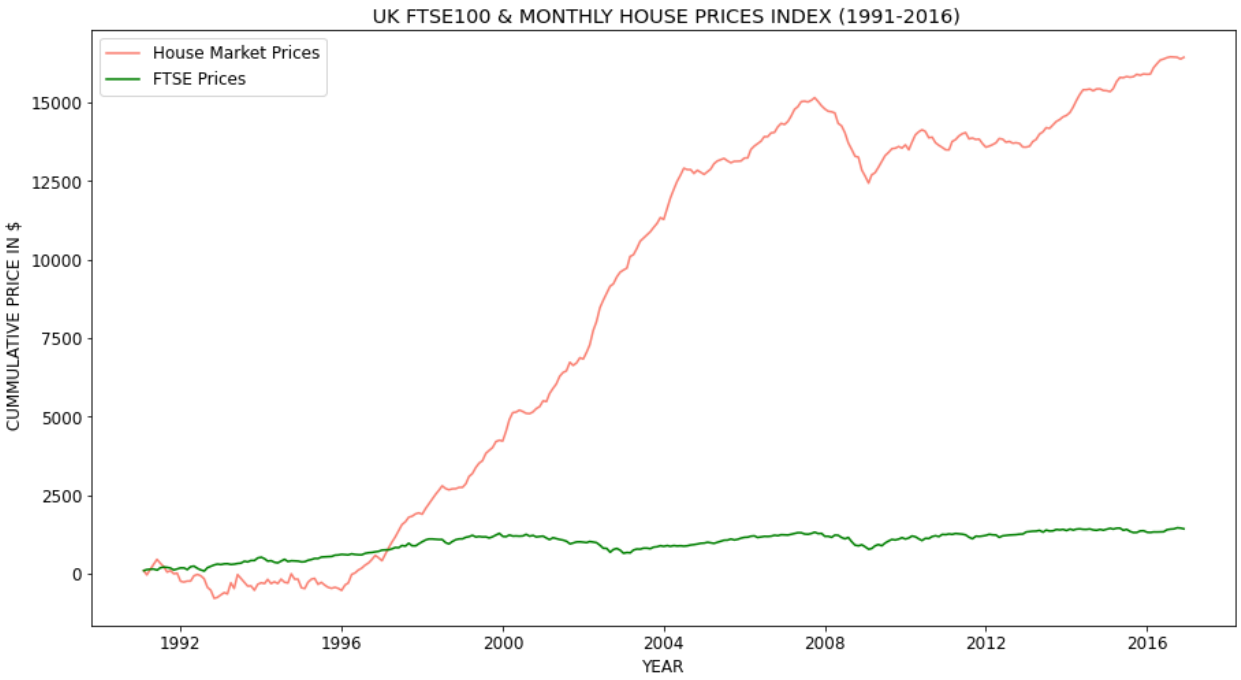
I imported The FTSE data file into my working and accessed the data. I changed the Date column value s from string to datetime then set it as the index in ascending order. This was to have the oldest date as the first row. I then extracted the required column (Adj Close) and created a subset data frame. I renamed the column for ease in plotting.

I also imported and extracted the required data from the HPI data frame.

I calculated the monthly returns on both data sets and stored in new columns, appended on the original data frames. Next, I computed the cumulative sum using the cumsum() passed on the monthly returns columns of both data sets. Finally, I normalized the data frames to start with 100 by multiplying all the values in the column and dividing by the first value in the column. I rounded the values to get whole figures.

I plotted the graph of UK FTSE 100 prices and houses monthly prices on the same graph using different colors.

Results: Graph 5a



Question 5a

Finding:

The graph shows indices of FTSE 100 and UK monthly house prices from 1991 to 2016. From the graph we can see that the House market prices is a positive slope with a gradual increase of prices from 1991 to 2016. The prices was almost steady from 1991 to 1996 where it increased steadily until 2008 where there was a dip until 2010 where the was some gain in prices. On the other hand FTSE prices remain in a plateau from 1991 to 2016. No much gains were made from the prices. No major deeps are noted for FTSE100 prices.

To calculate annualized return. I used the same formula as question 4The average annualized return for FTSE100 is calculated by

I would invest in the FTSE 100 stock market since there was a lot of gain in the prices with a little dips between 2008 and 2010 that didn't really affect the returns made. The trend shows the prices are highly likely to increase that drop.

Results 5b:

```
The annualized return rate is 4.463 %
```

References

[1] "statistics how to," 2021. [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/t-distribution/t-score-formula/. [Accessed 29 09 2021].

[2] "Kent State University," 2021. [Online]. Available: https://libguides.library.kent.edu/spss/independentttest. [Accessed 30 09 2021].

[3] [Online]. Available: https://www.youtube.com/watch?v=LGxgIjztWuU. [Accessed 30 09 2021].