DIAML: Assignment 3

by Anthony Rizkallah

Table of contents

1 Python Packages

Numpy

Scipy

Pandas

Unidecode

Matplotlib

Statsmodels

2 Question 1

2.1 Steps

To determine whether the women's energy intake deviates systematically from the recommended value of 7,725 kJ, it is appropriate to first define the hypotheses of this problem such that

H_0: The mean daily energy intake is equal to the recommended value ($\mu = 7,725\text{kJ}$)

H_1: The mean daily energy intake is NOT equal to the recommended value ($\mu \neq 7,725\text{kJ}$)

Next, a two-tailed test is appropriate because the goal is whether the value IS 7,725 kJ or IS NOT 7,725 kJ and this could happen on either side ($\pm$Z).

However, the number of sample sizes is less than 30 (n<30), so the appropriate method to use is the *t-test*, such that $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$.

Moreover, for a two-tailed test, there are twice as many critical regions because it is in both directions. This means that $\frac{\alpha}{2}$ should be considered rather than α.

Therefore, to prove/disprove the Null Hypothesis, the idea is

$$\begin{aligned} p &< \frac{\alpha}{2} \\ 2p &< \alpha \\ 2P(\bar{x} > \mu_0) &< \alpha \end{aligned}$$

2.2 Results

Energy Intakes = [5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770]

$n = 11$

$\alpha = 0.05$

$\mu_0 = 7,725$

$\bar{\text{X}} = \frac{\sum x_i}{n} \approx 6,753.64$

$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \approx 1,142.12$

$\text{SEM} = \frac{\sigma}{\sqrt{n}} \approx 344.36$

$Z = \frac{\bar{X} - \mu_0}{\text{SEM}} \approx -2.82 \ (t_{\text{statistic}})$

$\text{df} = n - 1 = 10$

$p = 2 \times P(Z \geq |-2.82|) = 2 \times (1 - \text{CDF}(Z)) \approx 0.0181$

Finally, $p = 0.0181 < \alpha = 0.05$, so the Null Hypothesis (H_0) is rejected, and in fact the smaller the p-value is from α, the more FALSE the Null Hypothesis is. Therefore, using the mean energy intake value of 7,725kJ is not accurate at all. For better visualization, here is the final table:

$\bar{X}$	σ	SEM	T-statistic	Degrees of Freedom	P-value
6,753.64 kJ	1,142.12 kJ	344.36 kJ	-2.82	10	0.0181

Table 1. Q1 Relevant Values

3 Question 2

3.1 Steps

To determine whether Guiness served in Irish pubs tastes better than pints served elsewhere in the globe, it is appropriate to first define the hypotheses of this problem such that

H_0: Guiness scores in Ireland are simply due to natural, random variation, and there is no difference in taste between Ireland and Elsewhere

H_1: Guiness scores in Ireland are not due to natural, random variation, and there is a difference in taste between Ireland and Elsewhere

Next, a two-tailed test is appropriate because the goal is whether Guiness IS better in Ireland or IS NOT better in Ireland, and not trying to find a specifically higher score, but rather whether the difference in taste is the same ($\mu_d = 0$) or different ($\mu_d \neq 0$).

So, the appropriate method to use is the *t-test*, such that $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$.

Moreover, for a two-tailed test, there are twice as many critical regions because it is in both directions. This means that $\frac{\alpha}{2}$ should be considered rather than α.

Therefore, to prove/disprove the Null Hypothesis, the idea is

$$\begin{aligned} p &< \frac{\alpha}{2} \\ 2p &< \alpha \\ 2P(\bar{x} > \mu_0) &< \alpha \end{aligned}$$

Finally, a two-sample test is appropriate because the comparison is being done between two different samples that are supposedly not subsets of each other or the same sample before or after an intervention.

To find the t statistic for a two-sampled tests with different variances is

$$t-\text{statistic} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

and the degree of freedom for the two-sample test is

$$\text{df} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\frac{\sigma_1^2}{n_1}}{(n_1-1)} + \frac{\frac{\sigma_2^2}{n_2}}{(n_2-1)}}$$

3.2 Results

$n_1 = 42$, $n_2 = 61$

$\overline{X_1} = 74$, $\overline{X_2} = 57$

$\sigma_1 = 7.4$, $\sigma_2 = 7.1$

$\alpha = 0.05$

$t-\text{statistic} \approx 11.65$ ($t_{\text{statistic}}$)

$\text{df} \approx 99.57$

$p = 2 \times P(Z \geq 11.65) = 2 \times (1 - \text{CDF}(t)) \approx 2.69 \times 10^{-20}$

Finally, $p = 2.69 \times 10^{-20} < \alpha = 0.05$, so the Null Hypothesis (H_0) is rejected, and in fact the smaller the p-value is from α, the more FALSE the Null Hypothesis is. Therefore, Guiness tastes better in Ireland than in other countries.

4 Question 3

4.1 Steps

From the World Bank, the data for GDP per capita and Fertility Rate (births per woman) averages are downloaded, and will be used to graph Fertility Rate vs GDP Per capita on a graph.

Afterwards, finding the correlation coefficient between the data is

$$\text{COEF} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

and can be estimated in Python using a function in *stats*.

A 0 correlation means that the relationship between the two datasets (GDP and Fertility Rate) is totally random (or does not exist). A non-zero positive correlation coefficient means that the relationship exists and the farther away from zero and closer to 1 is a stronger relationship. A non-zero negative correlation coefficient means that the relationship exists and the farther away from

zero and closer to -1 is a stronger relationship.

4.2 Results

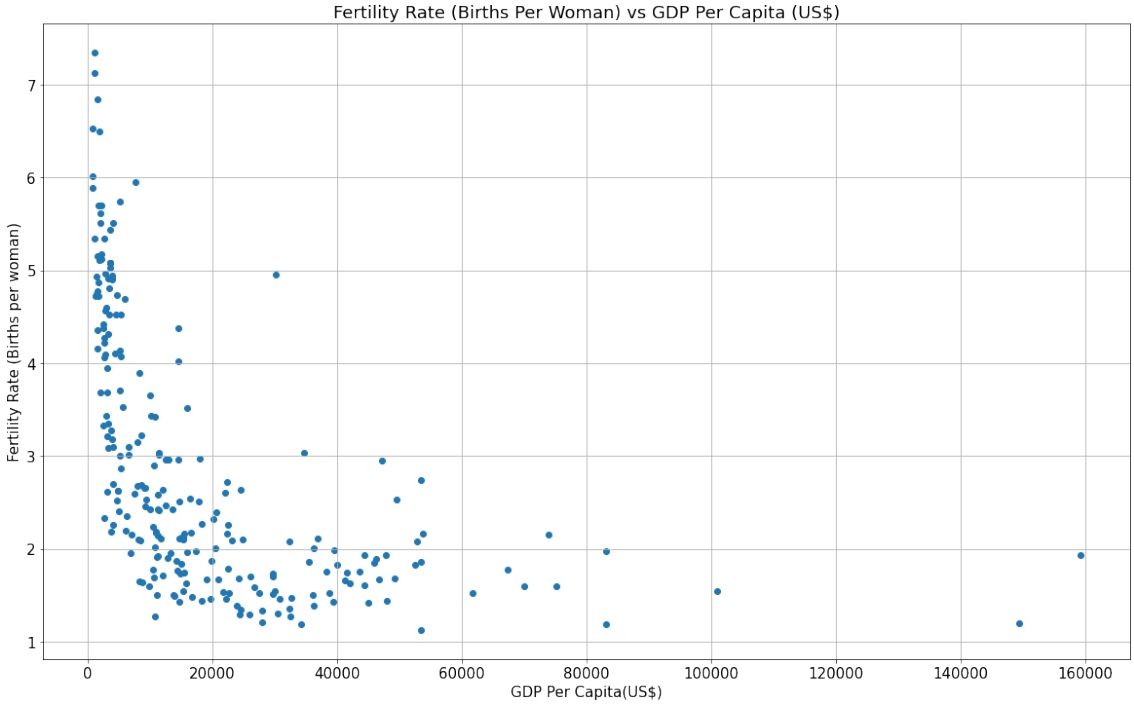

Figure 1. Fertility Rate (Births per woman) vs GDP Per Capita (USD)

The correlation coefficient is approximately -0.517 which means that the data is moderately correlated, and not strong. This means that the higher the GDP Per Capita in a country the lower the fertility rate (could be due to higher standards of living leading to people wanting to get less kids, but a more detaield analysis would be required for that assumption).

5 Question 4

5.1 Steps

First, the data of Housing Average Price in the United Kingdom from January of 1991 to August 2021 has a daily return rate of $r(t) = \frac{p(t)}{p(t-1)} - 1$, where $p(t)$ is the housing price at time t, and $p(t-1)$ is the housing price at the previous time.

Then, the ACF measures how values at time t in a time series are correlated with previous values (*lags*). To interpret ACF, looking at spikes at regular intervals is *seasonality*.

Finally, to calculate the annualized rate of return, using the formula

$$a_{\text{return}} = \left(\frac{\text{Final Price}}{\text{Initial Price}} \right)^{\frac{1}{\text{years}}} - 1$$

5.2 Results

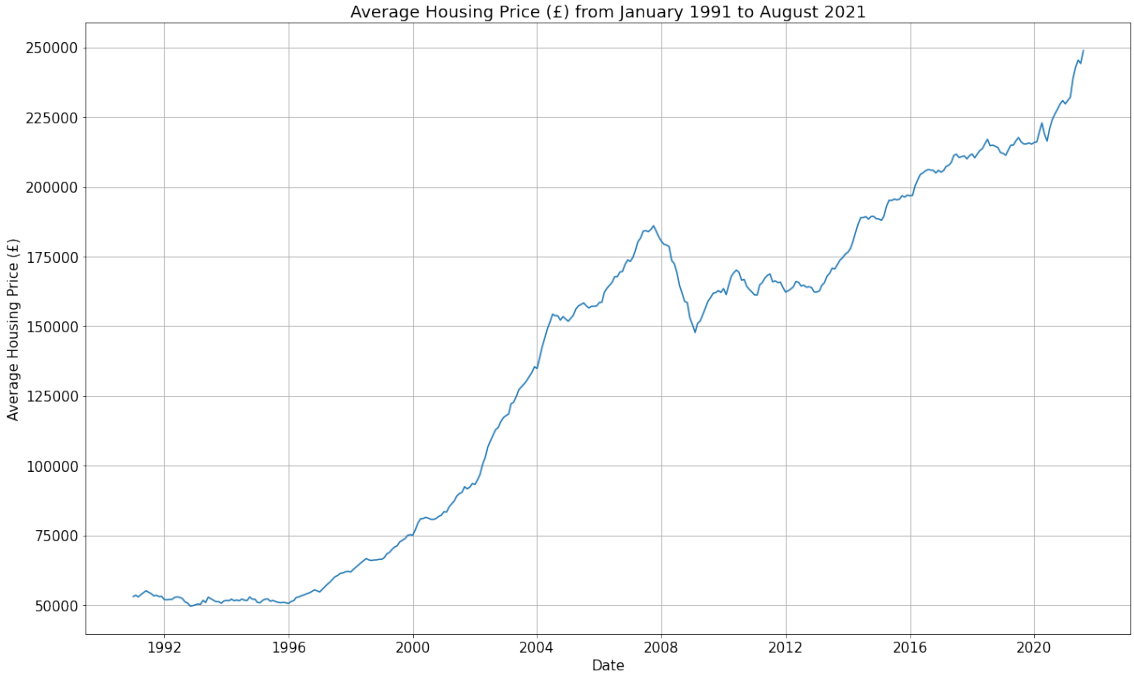

Figure 2. Average Housing Price over time (January 1991 to August 2021)

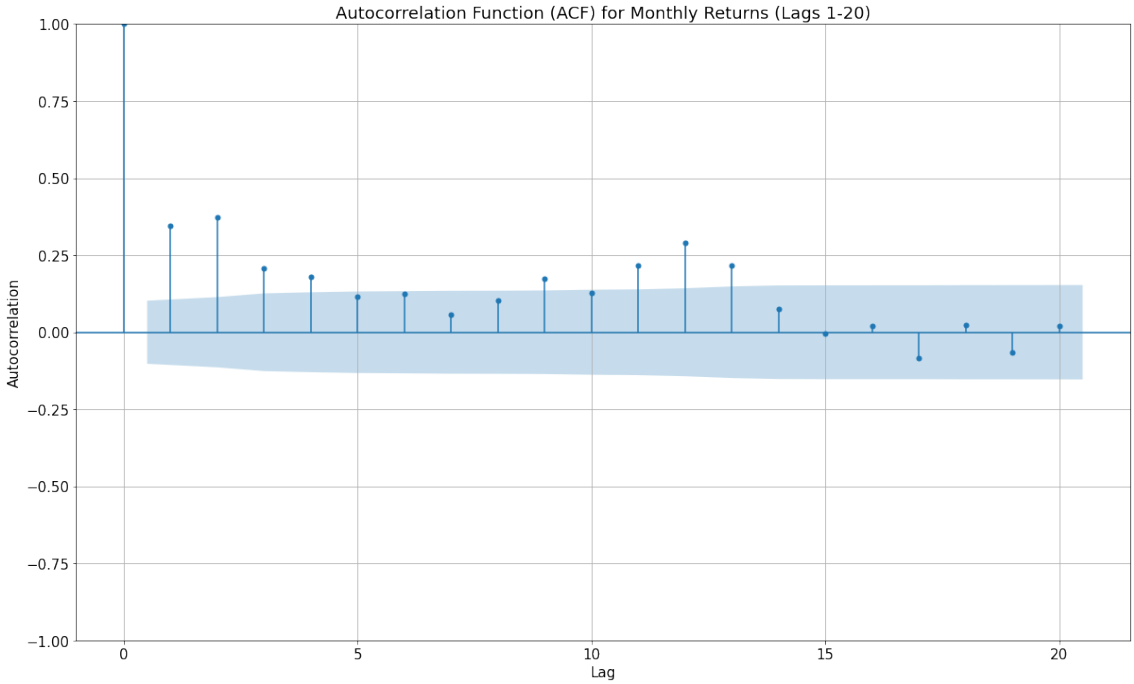

Figure 3. Autocorrelation Function for Monthly Returns (20 lags)

From the ACF graph, there is a trend of seasonality, where the blue shaded area represents the threshold of $p<0.05$. Any bars extending beyond these lines suggests strong autocorrelation at that lag.

Moreover, from the ACF graph, there is a trend of seasonality at lags almost every 12 lags, where the bars are higher than 0.05 which represent strong correlation.

The annualized rate of return (a_{return}) is approximately 5.17% over this time period, and this was calculated using the formula provided in the previous section of this question.

6 Question 5

6.1 Steps

To normalize a set of values, first find the scaling ratio, which in this case it is dependent onthe first value being 100. After doing that, the ratio is the one to scale everything to. It takes the form of

$$\text{Scaled Number} = \frac{\text{Target Number}}{\text{Original Number}} \times \text{Original Number}$$

Applying this to all values in the dataset to get the desired scale. For example, [150, 200, 250] scaled to start with 100 becomes:

$$\begin{aligned}
\text{Scaled Number}_1 &= \frac{100}{150}150 = 100; \left(\text{Scaling Factor is } \frac{100}{150}\right) \\
\text{Scaled Number}_2 &= \frac{100}{150}200 = 133.33 \\
\text{Scaled Number}_3 &= \frac{100}{150}250 = 166.67
\end{aligned}$$

Then, if the original numbers are required at anytime in the analysis, multiplying by the inverse of the scaling factor gives them back.

6.2 Results

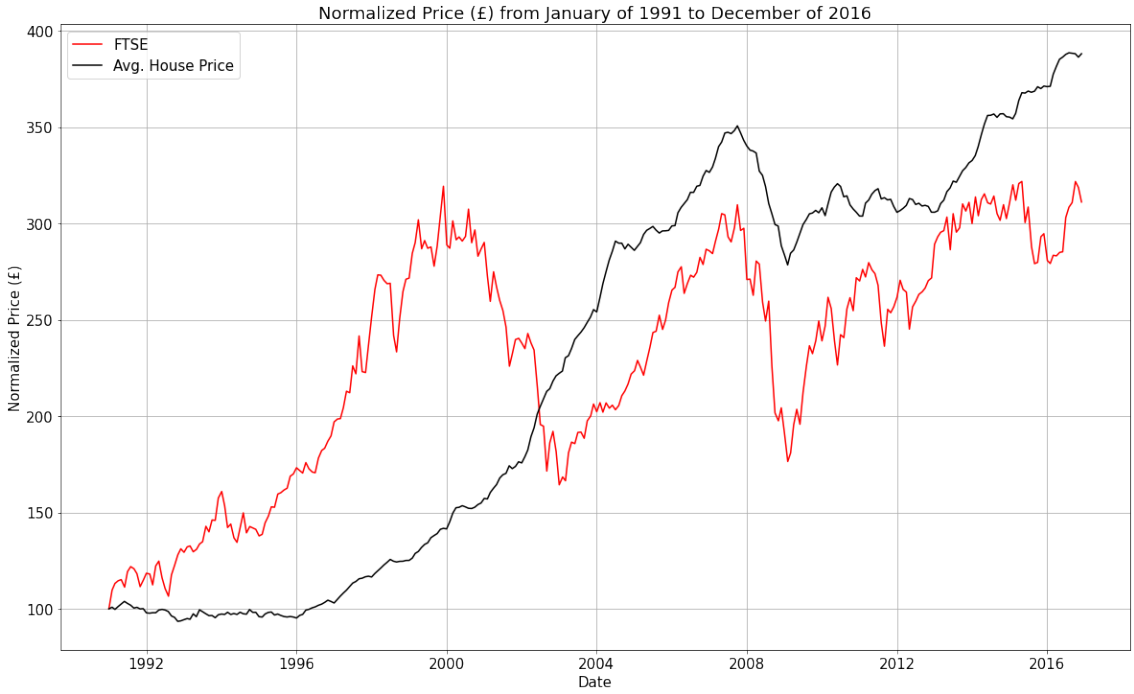

Figure 4. Normalized Price (Sterling) over time January 1991 to December of 2016

The annualized rate of return is approximately 4.46% for the FTSE100 over this time period, and this was calculated using the formula provided in the Steps section of the previous question.

It would have been better to invest in a house rather than than the FTSE100 because first, the annualized return rate of a house is 5.17% vs that of the FTSE100 which is 4.46%. Second, from the graph, the average price of a house is almost 60 normalized sterlings greater than that of the FTSE100. This is a little odd judging that real estate is usually considered a safe investment that grows steadily over time, whereas the FTSE100 is the riskier investment which is supposed to yield higher returns.