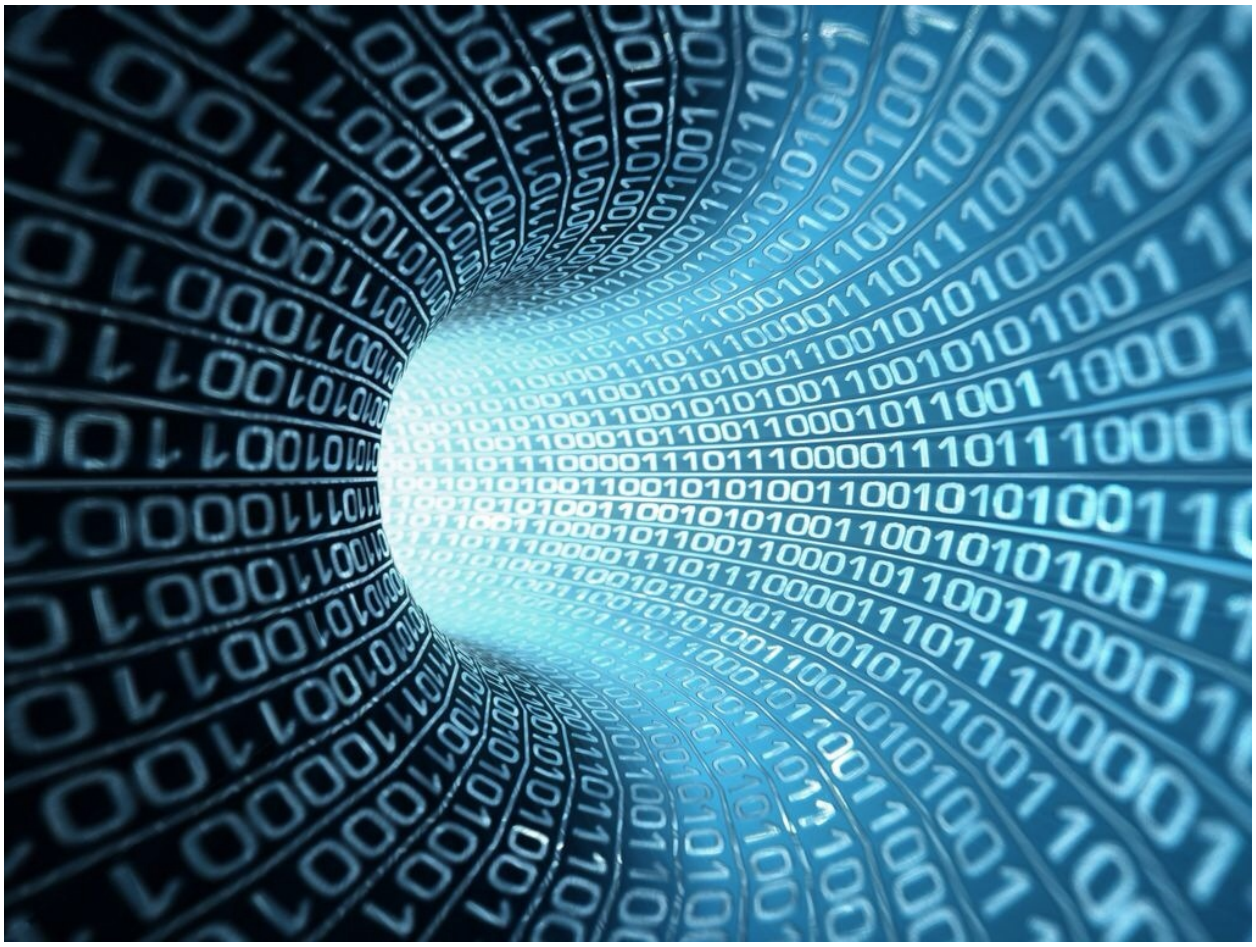
Report: DIAML Assignment 3



I, the undersigned, have read the entire contents of the syllabus for course 18-785 (Data Inference and Applied Machine Learning) and agree with the terms and conditions of participating in this course, including adherence to CMU's AIV policy.

Signature: Tunga Tessema

Andrew ID: tchamiss

Full Name: Tunga Tessema

Librairies

- **Matplotlib.pyplot:** to plot the graphs
- **Pandas:** to read excel and csv files, convert them to dataframes and perform operations on them
- **Numpy:** to create an array, calculate square roots and get the size of an array
- **scipy.stats:** to calculate p-values
- **Datetime**: to filter the column by a specific date
- **statsmodels.api**: to calculate the ACF function

Question 1

```
sample mean: 6753.636363636364
sample standard deviation: 1142.1232221373727
sample standard error of the mean (SEM): 344.3631083801271
t: -2.8207540608310198
degrees of freedom: 10
p-value: 0.018137235176105812
a two-tailed test should be used because the alternative hypothesis checks if the mean daily intake is not equal to 7725 kJ
the null hypothesis is rejected because the p-value is less than alpha
```

Steps: I first identified the null and alternative hypothesis. The null hypothesis is that a women's recommended daily energy intake is 7725 kJ. The alternative hypothesis is that a woman's daily energy intake is not equal to 7725 kJ. I then created a numpy array containing the 11 womens energy intake. I calculated the mean using the mean function. I calculated the standard deviation using the std function and gave it a ddof of 1 because I wanted to calculate the sample standard deviation. I then calculated the standard error of mean by dividing the sample's standard deviation by the square root of the size of the sample. I calculated t by subtracting 7725 from the sample mean and dividing it by the standard error of mean. I calculated the degrees of freedom by subtracting 1 from the size of the sample. Using the values of t and degrees of freedom, I calculated the p-value. For that I used scipy.stats.t.sf function and gave it

the absolute value of t as the first parameter and the degrees of freedom as the second parameter. I multiplied the result by 2 because we are doing a two-tailed test.

Insights: As described in the steps above, since our alternative hypothesis is that a woman's daily energy intake is not equal to 7725, we will use a two-tailed test.
The sample mean shows that on average the 11 women have a daily energy intake of 6753 kJ. The standard deviation indicates the variability of the sample. On average a woman's daily energy intake differs from the mean by 1142 kJ. We can see that the sample has a high variability. The standard error of the mean is 344 and it shows us how far the sample mean of the data is likely to be from the true population mean.
A t-test is a statistics that checks if two means are reliably different from each other. For that we first calculate the t value by dividing the mean difference by the standard error of the mean. We found a t-value of -2.82.
The degrees of freedom is 10 which is the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.
Using the t-value and the degrees of freedom, we got a p-value of 0.0181. The p-value is the probability that the pattern produced by our data could be produced by chance. The p-value is less than 0.05, so we reject the null hypothesis. Hence, we are 95% sure that a woman's daily energy intake differs from 7725 kJ.

Question 2

```
test should be two-sample test because we are trying to compare the mean of two samples: the ireland and elsewhere
the test is a right-tailed test because the alternative hypothesis is that ireland's guinness is greater than elsewhere
t: 11.73775770205081
degrees of freedom: 101
p-value: 6.979768077580737e-21
The difference is significant because p-value is less than alpha
```

Test statistic: $(\bar{x}_1 - \bar{x}_2) \ / \ s_p(\sqrt{1/n_1 + 1/n_2})$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, n_1 and n_2 are the sample sizes, and where s_p is calculated as:

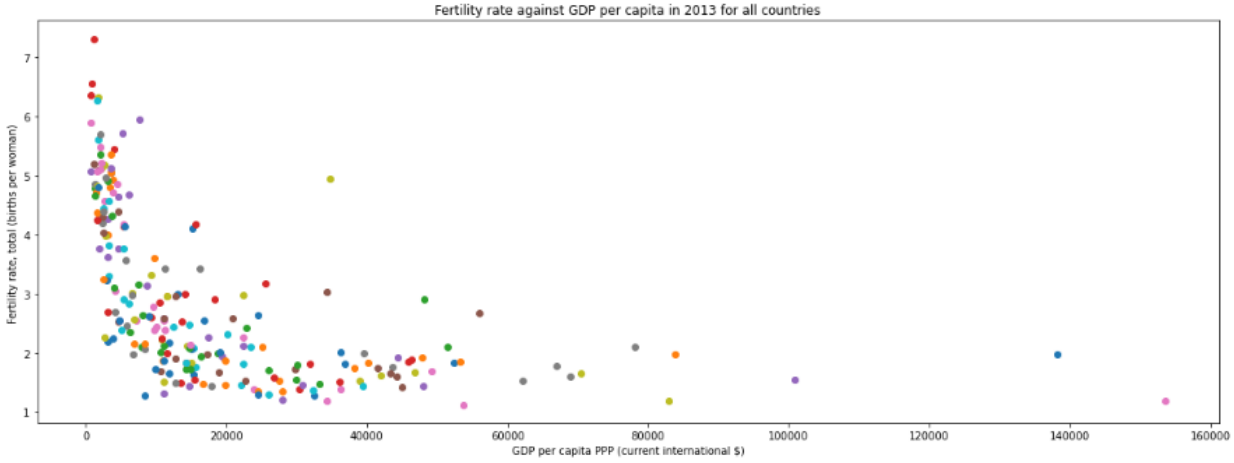
$$\mathbf{s_p} = \sqrt{(n_1\text{-}1)s_1^2 + (n_2\text{-}1)s_2^2 \ / \ (n_1+n_2\text{-}2)}$$

Steps: I first used variables to store the means, standard deviation and sample size of both samples. I then calculated t using the formula above. I also calculated the degrees of freedom using the n1 + n2 - 2 formula. I used this formula because the two samples have approximately the same variance. The ratio of their standard deviations is greater than 0.5 and less than 2. I

then used the scipy.stats.t.sf function to calculate the p-value by giving it the t-value and the degrees of freedom. I calculated the p-value for a one-tailed test

Insights: The null hypothesis is that the mean in Ireland and elsewhere is equal. The alternative hypothesis is that the mean in Ireland is greater than the mean elsewhere. The test should be a two-sampled test because we are comparing the means of two different samples. The test is a right-tailed test because our alternative hypothesis is that Ireland's sample mean is greater than elsewhere's sample mean. We found a t-value of 11.73 and 101 degrees of freedom. We got a p-value of approximately 6.9797e-21. The p-value is the probability that the pattern produced by our data could be produced by chance. The p-value is less than 0.05, so we reject the null hypothesis. Hence, we are 95% sure that Ireland's Guinness tastes better than anywhere else in the world.

Question 3



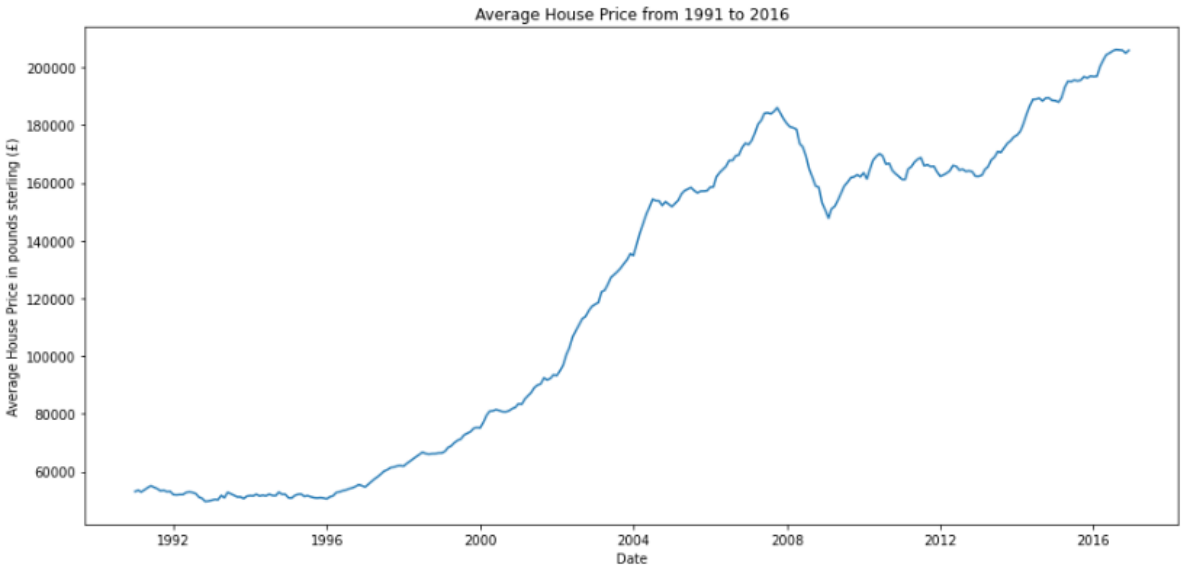

```
correlation coefficient: -0.5255135380223263
The correlation coefficient of -0.52 shows us that the fertility rate and GDP per capita have a negative correlation.      Howe
ver their negative correlation is not perfect since their coefficient is not -1.
```

Steps: I first read the excel files of GDP and fertility rate. I removed the first two unnecessary rows and selected only the needed column: Country Name, Country Code and 2013. I looped over every row and plotted the Fertility rate vs the GDP for each country in 2013. I calculated the correlation coefficient by using the corr function of pandas.

Insights: The above graph shows the fertility rate vs GDP per capita in 2013 for all countries. We can see that most countries who have a GDP greater than 20 000 dollars have a fertility rate of 2 or less. We can also see that countries who have a high fertility rate (greater than 3) are countries who have a GDP of approximately 2500 or less. However, we can see a set of countries whose GDP is between approximately 2500 and 22500 but they have a fertility rate of 3 or less. The correlation coefficient of -0.52 shows us that there is a negative correlation between the fertility rate and GDP. The more a country's GDP increases, the more the fertility

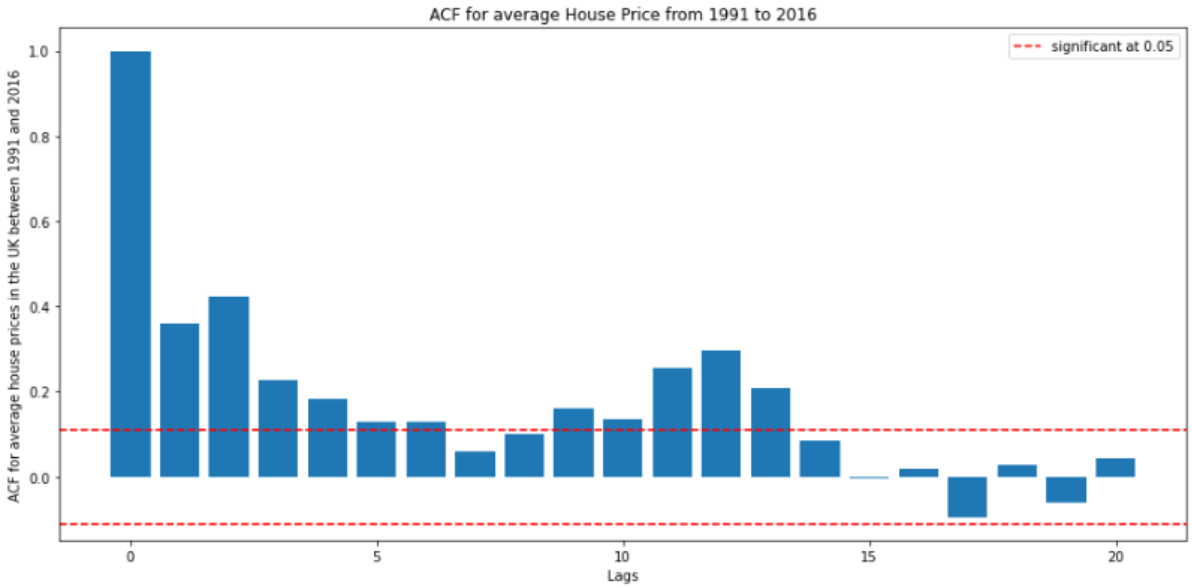
rate decreases. However, this negative correlation is not perfect since the coefficient's value is not -1.

Question 4



Steps: I first read the average housing price excel file using pandas and loaded it into a dataframe. I then renamed the unnamed 0 column to date. I selected only the rows that had a date between 1991 and 2016 and plotted them using matplotlib.

Insights: The above graph shows us that between 1991 and 1996, house prices in the UK have decreased by a small amount. From 1996 to approximately 2008, we can see that there is a sharp increase in the price of houses. From 2008 to approximately 2009, there is a sharp decrease: the average house price goes from approximately 185 000 to 150 000. After 2009, the average house price starts to increase again although there are some small decreases along the way, it generally kept increasing until 2016.



Significant values can be identified using the 95% confidence intervals (corresponding to a normal distribution) at $\pm 1.96/\sqrt{n}$ where n samples are employed

Steps: I first calculated the monthly return of the average house price by using the pct_change function. I then used the sm.tsa.acf function to calculate the ACF function with a lag of 20. I plotted it using a bar graph. I then calculated the values of ACF that are significant at $p < 0.05$. To calculate that I used the above equation. I then plotted the corresponding positive and negative values as horizontal red lines.

Insights: Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals. Autocorrelation can be used to measure how much influence past monthly returns have on future monthly returns.
The following graph shows that there is some element of seasonality. The ACF value reaches a peak for lag 2 and 12. It also shows that after reaching that peak value the trend is for the monthly return to decrease again. However, that trend and seasonality is not seen after the 14th lag.

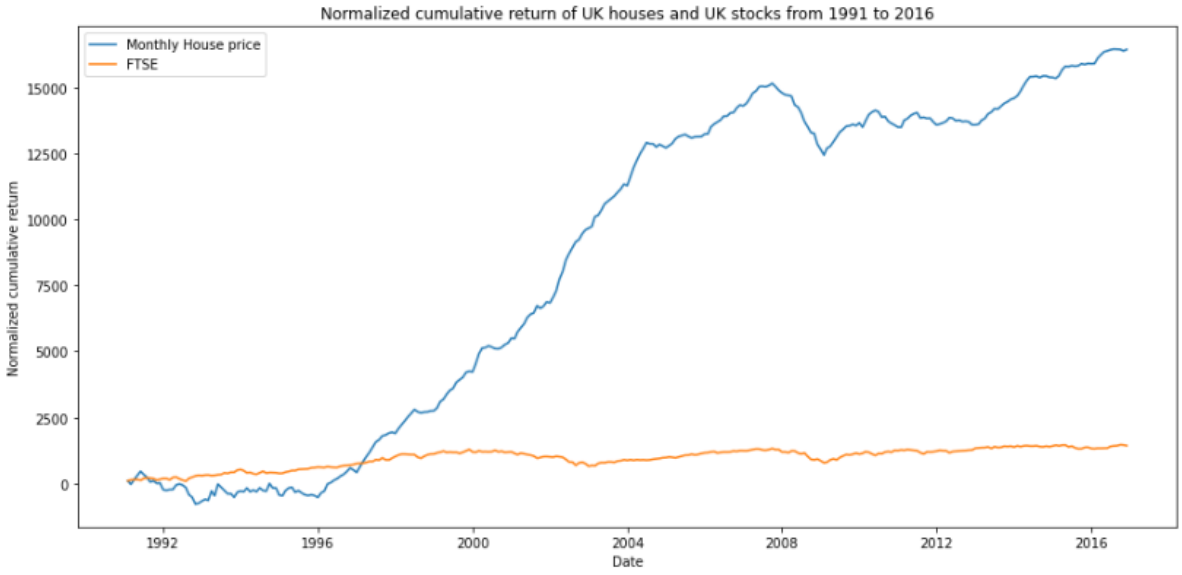
```
annualized return of Houses in %: 5.35423853535919
There seems to be a trend where the price reaches a peak and decreases again. But the peak and the rate it decreases is not the
same
There seems to be that there are seasons where the price increases and after reaching a certain point it decreases. But that do
esn't continue after the 14th lag
```

- **N in years**:
$rate = (1 + Return)^{1/N} - 1$
- **N in months**:
$rate = (1 + Return)^{12/N} - 1$
- Convert any time length to an annual rate:
- **Return** is the total return you want to annualize.
- **N** is number of periods so far.

Steps: To calculate the annualized return I used the above formula. The return is the final value minus the initial value over the initial value. I got the total return by using the final month of 2016 and the first month of 1991. I then used 26 as the N value, since this was over the period of 26 years.

Insights: The annualized return tells us how much return we can expect yearly over the period of 26 years. If we invest in houses we can expect an average return of 5.35% per year.

Question 5



Steps: I first read the FTSE csv file using pandas, filtered the rows to get only those whose dates are between 1991 and 2016. I then reversed the rows because the values were ordered starting from 2016. I used the pct_change function to calculate the monthly return and the cumsum function to calculate the cumulative return for the FTSE and average house prices. I then normalized the cumulative returns to start with a value of 100. I plotted the graphs using matplotlib.

Insights: The above graph shows us the cumulative return of stocks and house prices in the UK from 1991 to 2016. If a person decided to hold his investment for that long, it is better to invest in a house. The cumulative return of a house got from 100 to approximately 16400 whereas that of the stocks got from 100 to approximately 1400. However, if wanted to take out his investment between 1991 and 1997, investing in stocks was a better option because during that period the cumulative returns of stocks was greater than houses.

```
annualized return of FTSE in %: 4.462515478640672
```

- **N in years**:

$$rate = (1 + Return)^{1/N} - 1$$

- **N in months**:

$$rate = (1 + Return)^{12/N} - 1$$

- Convert any time length to an annual rate:
- **Return** is the total return you want to annualize.
- **N** is number of periods so far.

Steps: To calculate the annualized return I used the above formula. The return is the final value minus the initial value over the initial value. I got the total return by using the final month of 2016 and the first month of 1991. I then used 26 as the N value, since this was over the period of 26 years.

Insights: The annualized return tells us how much return we can expect yearly over the period of 26 years. If we invest in stocks we can expect an average return of 4.46% per year.
When we compare the FTSE and House annualized return we can again see that houses yield on average 1% more every year.