Carnegie Mellon University Africa

DATA, INFERENCE

&

APPLIED MACHINE LEARNING

(COURSE 18-785)

ASSIGNMENT 1

Mark Iraguha

(miraguha)

1. **Libraries Used:**
   Matplotlib – a python plotting library used to create animated, interactive and static visualizations.[1]
   Pandas – another Python library used that provides data structures and functions used to carry out data analysis.[2]
   Math – mathematical python module that provides useful math functions in python i.e. exponential function.[3]
   Numpy – a simple yet powerful data structure provided in python.[4]

2. **Introduction:**

   This report details the completion of Assignment 1. Assignment 1 requests answers to 10 critical thinking and data analytical questions.

3. **Objective**
   - The objective is to demonstrate, with full evidence, that I completed this assignment and demonstrate a full understanding of each step in the process including textual descriptions of each result and insights that can be gained.

Question 1 Report:

Question: A piece of paper is 1mm thick. Assuming you can fold it as many times as you want, how many folds would it take to exceed the height of Mount Everest at 8,848 m?

4. **Methodology**

a. Finding the number of folds, it would take to exceed a given height.
   - Approach:
     - Defines key variables i.e paper thickness, everestHeight.
     - Converted the thickness of paper to meters so it matches the unit of the height of Mount Everest.
     - Exponentially increase the paper thickness up to a point where any other fold would exceed the height of Everest.
     - Display the number of folds up to this point.

5. Results:

a. Finding number of folds

Through the use of a loop, I found the folds to get close to the height of Mount Everest. I then added 1 extra fold to my folds counter because that extra fold would mean exceeding the height of Mount Everest. The total number of folds are then displayed.

```
Total number of folds required to exceed mount everest:  24
```

Question 1: Snapshot showing code output

6. Analysis and Insights

a. Exponential Behavior of paper folds
   - It was learnt that papers once folded double in thickness each time. With this new knowledge, it would then be observed that this behavior can be called exponential growth.[5]

Question 2 Report:

Question: The volume of water in a reservoir decreases at an exponential

rate, following v(t) = v(0)exp(-at) with a=0.1. How much time, t, does it take for the volume to decrease to less than one half of its initial volume, v(0)?

7. **Methodology**
   a. Finding the time at which volume of water is half its initial volume
      - Approach:
        - We assume that a is 0.1. a represents the decay constant in the exponential equation.
        - We also recognize that time t will be equal to v(0)/2. v(0) being the initial volume.
        - We can then re-arrange the formula using the half-life formula from the exponential equation. Resulting into t = ln(2)/a.
        - ln(2) is the Natural Logarithm of 2 and in our code we use the log function to represent this.[6]
        - This way, we can now assume any volume of water i.e. 22,777 to give us time t which is approximately 6.931

8. **Results:**
   a. Finding time t
      Upon following the approach outlined above, we now know that any volume of water assumed will give use the approximate units of time it would take for that volume of water to reach half of its initial volume.

```
It will take approximately 6.93 units of time.
```

Question 2: Snapshot showing code output

9. **Analysis and Insights**
   a. Exponential decay

- The equation used represents exponential decay and it is surprising that any volume assumed produces the same half-life. [6] For example, v(0) as 100 or 10 or 22,000 all produce the same units of time, approximately 6.931.

Question 3 Report:

Question: If you deposit $100 in a bank account that offers an annualized interest rate of 5% (compounded annually), how much money will you have (round to the $) after one, two, three, four and five years?

10. Methodology

a. Utilizing the compound interest formula
   - Approach:
     - We can use the compound interest formula $a=p(1+r/n)^{nt}$
     - 'a' is the annual compounding, 'p' is the principal, 'r' is the annualized interest rate (as a decimal), 'n' is the number of months in a year, 't' is the number of years.
     - With this formula, we can also use a loop to get the annual compounding at the end of each year of the 5 year period.

11. Results:

a. Finding compounded interest after each year of 5 year period
   Upon applying the formula, we get the annualized interest added to the principal from the previous year. This is printed out for each year of the 5 year period.

```
Accumulated compound interest after year 1: 105
Accumulated compound interest after year 2: 110
Accumulated compound interest after year 3: 116
Accumulated compound interest after year 4: 122
Accumulated compound interest after year 5: 128
```

Question 3: Snapshot showing code output

12. Analysis and Insights

a. Annualized Compound Interest

- With a simple formula you can get to know your return for a 5 year period without adding the interest of year manually after every end of the year.
- Noticeably, the interest return for a principal of $100 is approximately $5 dollars using an annualized compound interest of 5%.

Question 4 Report:

Question: Suppose you want to buy a car worth $20,000. A financial institution can provide a loan with a monthly interest rate of 1%.

What is the monthly payment to pay off the debt in one, two and three years (rounded to the nearest $)?

13. Methodology

a. Using the loan repayment formula

- Approach:
  - We utilize the loan repayment formula: $p=r(pv)/1-(1+R)^{-n}$
  - Where ‘p’ is the payment, ‘PV’ is the present value, ‘r’ is the rate per period, ‘n’ number of periods

14. Results:

a. Finding monthly contribution if the payment term was 'n' years. Using the loan repayment formula, we found the monthly contribution to the car loan if the loan repayment period was 1 or 2 or 3 years.

```
Monthly loan contribution for 1 year loan term: 1776
Monthly loan contribution for 2 year loan term: 941
Monthly loan contribution for 3 year loan term: 664
```

Question 4: Snapshot from code output

15. Analysis and Insights

a. Loan Repayment

- Having utilized the formula for the repayment period, we observe that I would pay more monthly if the payment term if shorter as opposed to a longer term.

Question 5 Report:

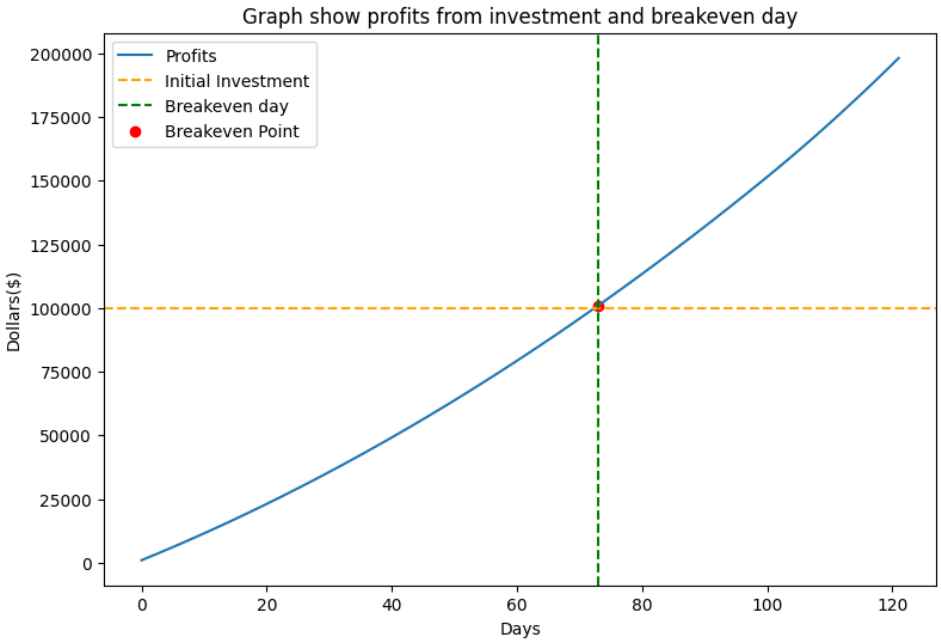
Question: You are about to set up a new business and will invest $100,000. On day one you expect to have 100 customers and the number of customers will grow at a rate of 1% per day. If each customer provides profits of $10, how many days will it take to repay your initial investment based on cumulated profits? Plot cumulated profits per day, show initial investment and mark breakeven day.

Methodology

a. Find profits by exponential growth of customers
    - Approach:
        - We make sure to calculate the exponential growth of customers every day.
        - All daily profits are then calculated according to the new number of customers.
        - We store the results in a collection.
        - We then find the point (day) when the initial investment is realized.
        - We use the data above to configure a line graph plotting the profits against the days.
        - We mark the initial investment and breakeven day.

16. Results:

a. Plotting cumulated profits:
   Utilizing the matplotlib[1] and pandas[2], we use the profits per day data to plot a line graph marking the initial investment and breakeven day.



Question 5: Figure 1 showing Graph

17. Analysis and Insights

a. Exponential decay

- We observe a linear trend in the growth of profits. Noticing that getting double the initial investment takes a shorter time compared to the time it took to realize the initial investment. We could predict that profits get larger as the exponential growth of customers gets bigger.

Question 6 Report:

Question: Using data from http://bit.ly/1JJyf29 and linear interpolation, estimate the dates when the number of cases and deaths due to Ebola exceeded 100, 500, 1000, 2000 and 5000. Graph the cases and deaths (observations and interpolations) and mark the dates when thresholds were exceeded with a circle.

18. Methodology

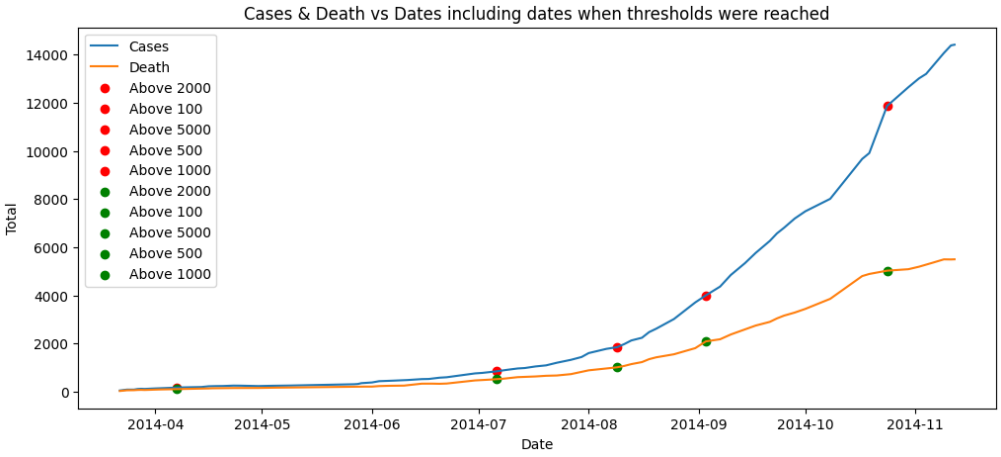
a. Using linear interpolation and graphing the cases and deaths

- Approach:
  - We create a dataframe by reading from a data file using matplotlib[1].
  - We use the minimum date and maximum date to come up with a data range. This will be our new index.
  - We reindex the dataframe using the new index.
  - Apply linear interpolation and save the outcome to a new dataframe.
  - Using the thresholds stated in the question, we iterate through the data series and store the first occurrences of points where any of our threshold is realized.
  - We use the store these points in a collection and plot a graph showing the same.

19. Results:

a. Plotting cases and deaths:

Utilizing the matplotlib[1], we plotted the cases and deaths while marking the first occurs.



Question 6: Figure showing line graph

20. Analysis and Insights

a. Linear interpolation for cases and deaths
    - Utilizing the matplotlib[1], we observe a linear trend in the growth of profits. So, we use linear interpolation to make it easier to compare the two series in the data frame.
    b. Cases and Deaths
    - It can be observed that the time at which series 'ts' reaches the threshold at approximately the same time.

Question 7 Report:

Question: Using data from 2014, downloaded in the previous question, what is the average growth rate per day, as a percentage, in the number of Ebola cases and deaths?

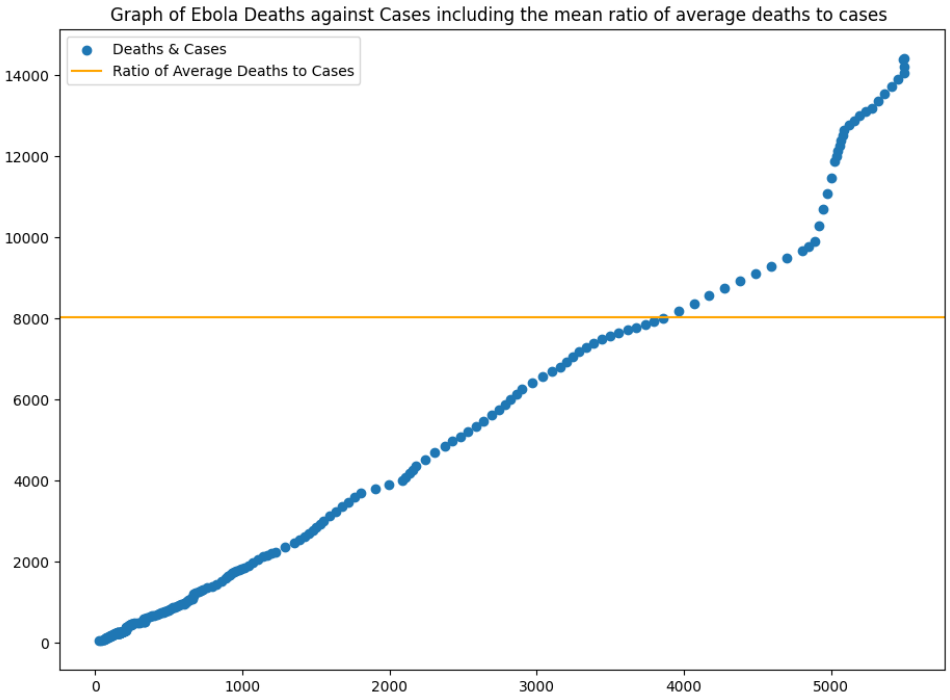
21. Methodology

a. Calculate the average growth rate per day using compounding formula
    - Approach:
        - We find the initial and final values from our series. This applies for both the Cases and Death.
        - Using the Annual growth rate formula[7], we get the average annual growth rate for each series (cases and death).
        - We convert them to daily growth rates and express them as a percentage.
        - We finally display our results.

22. Results:

a. Average Daily Growth Rate:

Having utilized the average annual growth rate formula[7], we got the daily growth rate (through conversion) and expressed it as a percentage.



Question 7: Figure showing scatter graph

23. Analysis and Insights

a. Average Daily Growth Rate

- The relationship between the cases and deaths are so similar which perhaps suggests that there a little to no survivors out of the cases considered.

Question 8 Report:

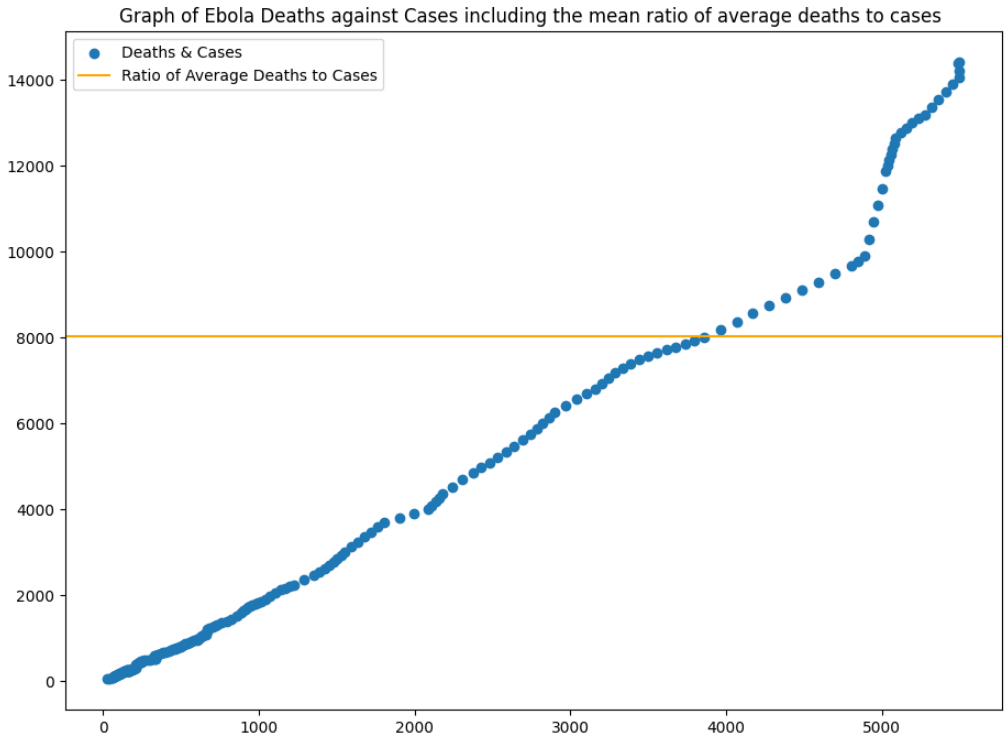
Question: Using the same date, plot the number of deaths versus the number of cases and estimate the average ratio of Ebola deaths to cases.

24. Methodology

a. Get the ratio of deaths vs cases as a dataframe series
   - Approach:
     - We use the death and cases series from our data frame from question 6 and get the ratio as a series.
     - We utilize the mean function to get the average ratio against the ratio series.
     - We use the resulting series to plot the average ratio of Ebola deaths to cases utilizing the matplotlib library.[1]

25. Results:

a. Plotting the average ratio of ebola deaths to cases:
   Utilizing the dataframe from question 6, we use the deaths and cases to plot a scatter graph and represent the estimate of average ratio of Ebola deaths to cases.



Question 8: Figure showing scatter graph

26. Analysis and Insights

a. Average ratio of ebola deaths to cases
   - It appears that the ratio falls just after the mid mark of the deaths and cases and just before the sharp rise in the ebola cases and deaths.

Question 9 Report:

Question: Obtain daily prices for two ETFs called SPY and TLT which track the S&P500 index and long-term Treasury Bond. Select the adjusted closing prices. Plot the two-time series during 12/31/2013 – 08/31/2015 and make them comparable by starting from prices of $100 on the first day in 12/31/2013 – 08/31/2015.

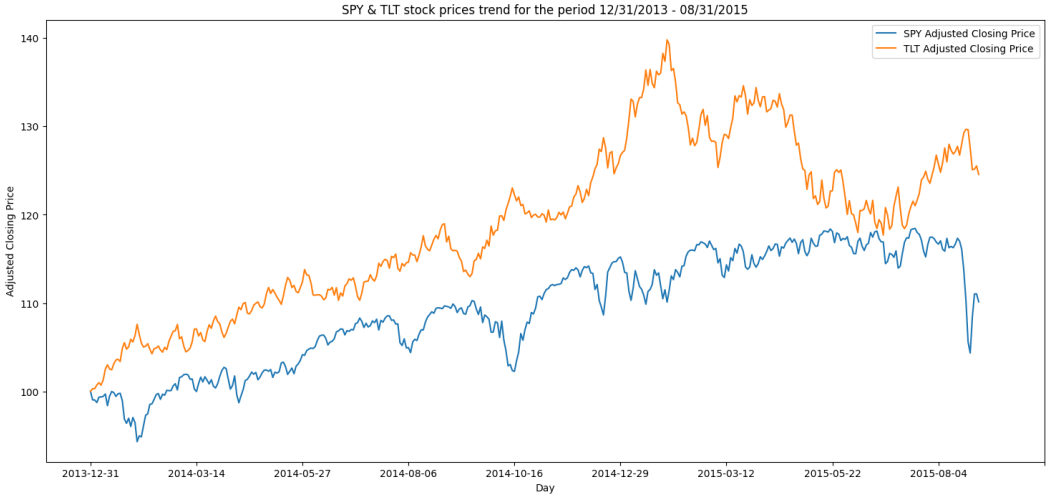
20. Methodology

a. Using normalization to make the two ETFs comparable
   - Approach:
     - We read the ETFs from the csv files provided using pandas read_csv function.[2]
     - We get the first value of each of the ETFs and normalize them.
     - We create a unified dataframe holding the normalized data.
     - We use the normalized data frame to plot a two-time series during the period 12/31/2013– 08/31/2015.
     - We then display the line graph.

21. Results:

a. Plotted two-time series graph using normalized ETF data:

Having normalized the two ETFs (SPY and TLT), we plot the two-time series graph for the period 12/31/2013– 08/31/2015.



Question 9: Figure showing two-series line graph

22. Analysis and Insights

a. Normalization

- Normalization has helped us to compare two datasets that strongly differ so we can easily carryout analysis on them to better inform our thoughts and decisions about the two ETFs.

Question 10 Report:

Question: For the ETFs on the previous question, calculate daily returns, r(t) = p(t)/p(t-1)-1, for each trading day in the same time period as above. Calculate the average, min and max daily return for each of the two ETFs during the time period and express these as percentages.

23. Methodology

a. Using daily returns to calculate average, min and max daily returns for SPY and TLT
   - Approach:
     - To get the daily returns, we take advantage of the pct_change function of the pandas software library. [8]
     - We then use the result series to compute the mean, minimum and maximum daily return.
     - We follow the same steps for each of the ETFs.
     - Finally, we print the results.

24. Results:

a. Average, minimum and maximum daily returns from the two ETFs: Once we have got the daily returns for the two ETFs, we then compute the average daily returns, find the minimum and maximum daily return and display the results for each of the ETFs.

```
SPY Average Percentage: 0.03%
SPY Min Daily Return Percentage: -4.21%
SPY Max Daily Return Percentage: 3.84%
TLT Average Percentage: 0.06%
TLT Min Daily Return Percentage: -2.43%
TLT Max Daily Return Percentage: 2.65%
```

Question 10: Snapshot showing code output

25. Analysis and Insights

a. Daily Return Analysis
   - Using the resulting data, we observe the while the SPY ETF has the higher daily return, it also has the lowest daily return.
   - The SPY ETF also has the higher Average Daily return.

- With this knowledge the SPY seems to be a high risk and high reward kind of ETF.

References:

[1] “Matplotlib,” *Wikipedia*. Aug. 30, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Matplotlib&oldid=1243075914
[2] “pandas (software),” *Wikipedia*. Jul. 15, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pandas_(software)&oldid=1234683004
[3] “math — Mathematical functions,” Python documentation. Accessed: Sep. 01, 2024. [Online]. Available: https://docs.python.org/3/library/math.html
[4] R. Python, “NumPy Tutorial: Your First Steps Into Data Science in Python – Real Python.” Accessed: Sep. 02, 2024. [Online]. Available: https://realpython.com/numpy-tutorial/
[5] K. S. Kruszelnicki, “Folding paper.” Accessed: Sep. 01, 2024. [Online]. Available: https://www.abc.net.au/science/articles/2005/12/21/1523497.htm
[6] “Exponential decay,” *Wikipedia*. Jun. 07, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Exponential_decay&oldid=1227744550
[7] “Growth Rates: Formula, How to Calculate, and Definition,” Investopedia. Accessed: Sep. 02, 2024. [Online]. Available: https://www.investopedia.com/terms/g/growthrates.asp
[8] “Python | Pandas dataframe.pct_change(),” GeeksforGeeks. Accessed: Sep. 02, 2024. [Online]. Available: https://www.geeksforgeeks.org/python-pandas-dataframe-pct_change/

Carnegie
Mellon
University
Africa