DATA
MACHINE LEARNING
VISUALIZATION

Data Scientist Challenge

Damavis 2022

Data Scientist Challenge

Hello, future Damavis teammate! First of all thanks for accepting our challenge.

Your solution's code must be written in the Python programming language. We encourage you to use third-party libraries to ease your analysis. All the comments and any documentation must be written in English.

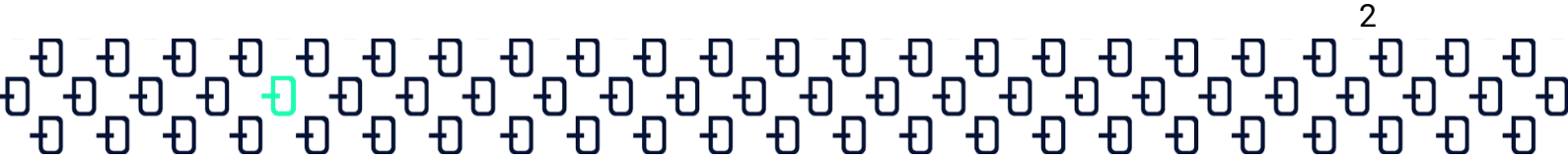
We value strong statistical reasoning, good communication skills, readable code, structure, good code principles and best practices.

If you cannot finish the challenge, please send us your partial solution anyway.

You have to submit two files:

- A jupyter notebook with all the code you have developed and the necessary comments.
- A requirements.txt with all the dependencies to be able to run your code.

Please send us your own code, thought and written by yourself without any help from anyone. **BEST OF LUCK**!



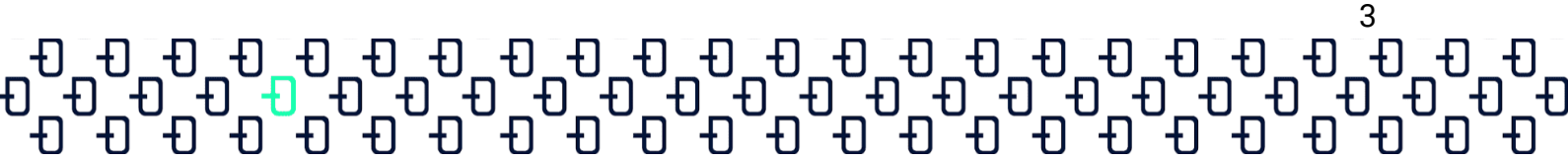
Problem 1: Programming

Consider a box with three balls, each with a different value: ball A is worth 2 points, ball B is worth 3 points and ball C is worth 4 points. Write a function that calculates the number of different ways of reaching a total sum of N points by extracting from the box one ball at a time, taking into account the order in which the balls are drawn.

For example, for N=6 you have 4 different possibilities:

- Taking ball B twice in a row → 3 + 3 = 6 points
- Taking ball A three times in a row → 2 + 2 + 2 = 6 points
- Taking ball A, then ball C → 2 + 4 = 6 points
- Taking ball C, then ball A → 4 + 2 = 6 points

Your function must receive the target number of points N as a parameter and return the number of different ways of getting to that number. Make the function as efficient as possible and discuss its computational complexity.



Problem 2: Demand Estimation

The sales of two companies, Company 1 and Company 2, in two regions, Region 1 and Region 2, are provided here. The description of the only three columns are:

- Sales_U → Sales in equivalent units (lbs)
- Sales_USD → Sales in $
- date → starting date for the week

Tasks

- Construct time-series plots of sales and prices for Company 1 in Region 1 and 2. Repeat the exercise for Company 2. Describe the differences or similarities between Company 1 and 2 pricing policies.
- Construct scatter-plots of sales versus prices for Company 1 in Region 1 and Region 2 separately. Repeat the exercise for Company 2. Is there evidence for a negatively sloped demand-curve in the data? Eye-balling these plots, does demand appear more elastic in Region 1 or 2?
- Estimate the **price elasticity of demand** for Company 1 and 2 at Region 1 and 2 (four different demand models). Is the demand elasticity higher (in absolute magnitude) in Region 1 or 2?
- Compute the % change in unit sales for a 10% increase in the price of Company 1 at Region 1.
- You may be called upon to report to your manager whether your brand is vulnerable to a competitor's pricing policies. That is, to what extent does the demand for your product depend on (or is affected by) your competitors' pricing policy? Which brand is more "vulnerable"? Be specific as to why.
- While making a crucial presentation of the above results in front of your team, your analyst colleague questions your results as follows: "This is all fine. But, you know, you're missing a lot of variables in your so-called regression model. For instance, the sales of Company 1 at Region 2 are clearly affected by store traffic. When it snows, less people visit Region 2, and you don't have such factors -- the weather, temperature, traffic congestions, etc. So aren't your cross-price effects all wrong?" Is your colleague right or wrong?