Thank you for your interest in joining Zumr team. The following two challenges are designed for Data Scientists to be simple and fun to work while allowing for you to display your genius. We wish you all the best. For any help or clarification, please feel free to contact Omar either on LinkedIn (here), Email (omar@zumr.om) or phone number (71314156).

Submission: Please create a single repository on Github to use for the two challenges and share it with Omar when you have completed your submission via one of the above contact methods. Make sure to write clean and documented code and follow good commit practice. (Hint: uploading your work to github after you are done will not be considered good practice).

Challenge 1 (50 Points): In this challenge, you will build a predictive model that can accurately estimate the survival time of patients with advanced, inoperable lung cancer who were treated with chemotherapy. The dataset includes various variables that could potentially influence survival time.

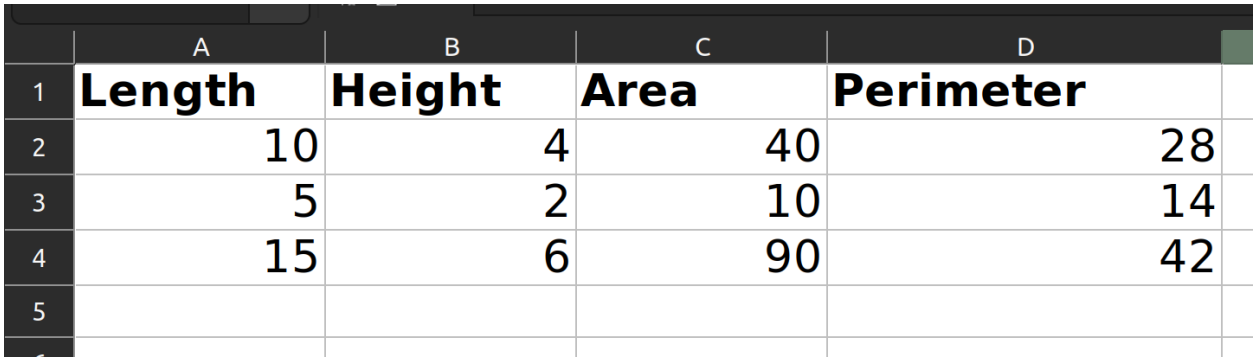
Requirements:

1. Load the dataset: Load the dataset from the .txt (attached) file to a pandas DataFrame.
2. Preprocess the dataset: Clean and prepare the data for training, handling missing values, and normalizing features if necessary.
3. Build a predictive model: Develop a model that can predict the survival time of patients based on the given features. You can use any machine learning algorithm of your choice.
4. Evaluation: Split the dataset into training and testing sets. Evaluate your model's performance using appropriate metrics such as mean squared error, mean absolute error, or others suitable for survival analysis.
5. Interpretation: Analyze the significance of different features in predicting survival time. Which variables seem to have the most impact on survival?
6. Utilize Git for version control throughout the development process. Maintain a Git repository with clear commit messages and a well-documented commit history.
7. Include all dependencies in a requirements.txt file
8. **Must use Python**

Challenge 2 (50 Points): Write a Python function that takes the name of an Excel file (.xlsx) and returns a dictionary of column and formula (as a string) pairs for each column computed using other columns.

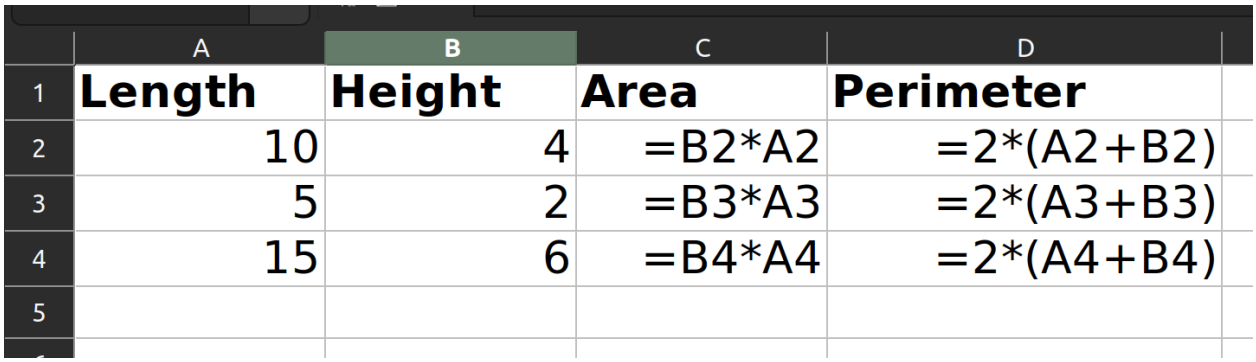
Example: If the following Excel file is provided as an input:

Picture of the content with the values shown:



	A	B	C	D
1	**Length**	**Height**	**Area**	**Perimeter**
2	10	4	40	28
3	5	2	10	14
4	15	6	90	42
5				

Picture of the same file with the formulas shown instead:



	A	B	C	D
1	**Length**	**Height**	**Area**	**Perimeter**
2	10	4	=B2*A2	=2*(A2+B2)
3	5	2	=B3*A3	=2*(A3+B3)
4	15	6	=B4*A4	=2*(A4+B4)
5				

The output will be: {“Area”: “Length*Height”, “Perimeter”: “2*(Length+Height”)”}

Requirements:

1. The function must take the name or path of the Excel file.
2. The output is a dictionary.
3. Utilize Git for version control throughout the development process. Maintain a Git repository with clear commit messages and a well-documented commit history.
4. Include all dependencies in a requirements.txt file.
5. **Using Python is recommended.**

Hints:

1. You can assume that the first row of the Excel file always contains headers
2. You can assume that in each column there can be either no formula or the same formula applied to all rows (except the headers)