Data Science Intern Assignment

Objective

Your assignment explores the attached "the_office_lines_scripts" dataset. Your task is to create a report on the dataset using Python scripts or optionally Jupyter.

Brief

Everyone loves "The Office," a popular show that aired from 2005 to 2013. While doing research, you stumbled across this dataset, with the lines of all the episodes. You decide to explore the dataset and answer some questions.

Tasks

Your first task is to create a Python script or Jupyter notebook and add the `.py` or `. ipynb` files to the repository. You will work on this script/notebook for the rest of the challenge.

We want you to answer at least four of the following questions:

- How many characters are there? What are their names?
- For each character, find out who has the most lines across all episodes
- What is the average of words per line for each character?
- What is the most common word per character?
- Number of episodes where the character does not have a line, for each character
- Number of times "That's what she said" joke comes up
    - Include five examples of the joke
- The average percent of lines each character contributed each episode per season
- Come up with 2-3 interesting questions yourself surrounding the dataset

Evaluation Criteria

- The report is easy to read, and its sections are well divided
- The data exploration is done in an easy-to-understand way

- Any graph or figure created is clear and easy to read (creating graphs or figures in Jupyter is completely optional)
- At least four of the questions are answered. Bonus points for each additional answer. Even if the answer is incorrect, the arguments and how you are presenting them are valuable

Submitting

Please organise, design, test and document your code as if it were going into production - then push your changes to an online repository and share the URL with us.

All the best and happy coding,
The Forwardize Team

the_office_lines_scripts.csv