Assignment 1 for "Business Analytics I" course

The assignment should be done individually, the deadline is February 12. The exercises worth 25 points in total. You have to submit a Jupyter notebook file with the code that you used to solve the different tasks; use comments in the notebook to discuss the results and explain the output of your code.

1. (12 points) In this task, you have to work with a dataset in the file *rotten_tomatoes_top_movies.csv*. You can find more information about the data at `https://www.kaggle.com/datasets/thedevastator/rotten-tomatoes-top-movies-ratings-and-technical`. The data provides some information about movies and their rating on the Rotten Tomatoes site. By making use of the basic data analysis tools introduced in the course, answer the following questions:

   - Import the data and create a dataframe whcih consists of the following columns of interest: 'title', 'year', 'critic_score', 'people_score', 'total_reviews', 'total_ratings', 'rating', 'type', 'original_language', 'director', 'release_date_(theaters)', 'runtime'.
   - What is the average number of reviews for the movies included in the dataset (total_reviews column)? After calculating this value, filter your data for movies that have at least as many reviews as this average value.
   - How many rows do you have in this data in which the movie is more than 3 hours long? (Hint: what is the first character/digit of the runtime column if the movie is more than 3 hours long?)

   Before the following tasks, look at the type column, choose your favourite type, and filter your data to include only movies of that type. Perform the following tasks on this filtered data.

   - For which movie in this dataset you can find the largest difference between the critics and viewers evaluation (critic_score and people_score)?
   - How many directors do you have in the dataset with more than one movie?

2. (13 points) In this exercise, you will have to analyze a dataset about companies listed in Fortune 1000, the 1000 largest american companies by revenue (*fortune.csv*). You can find more information about the data at `https://www.kaggle.com/datasets/winston56/fortune-500-data-2021`. By making use of the basic data analysis tools introduced in the course, answer the following questions:

   - After importing the data, check how many new companies are in the data (companies that were not present the previous year, using the newcomer column). What is the rank of the best newcomer company (use the column rank)?
   - Is there any newcomer company in the top 100 companies based on profit? (Note: be careful here, this is not the same as column rank, as it is based on revenue; you have to determine the ranking based on profit first)
   - Check the correlation between revenue and number of employees. Based on the results, what is your opinion, does the size of the company in terms of employees impact revenue? If you calculate the correlation between profit and number of employees, would you make the same observation?
   - How many companies do you find on the list that made losses, even though based on revenue they are in the top 1000? What is the biggest loss (i.e., most negative profit) that you find in the data, and in what sector is the company that made that loss?
   - Create a categorical version of the revenue column with 4 groups: (i) values between 0 and 3500, (ii) values between 3500 and 6380, (iii) values between 6380 and 14620, and (iv) values between 14620 and 1000000. Label the groups ['Low', 'Medium', 'High', 'Very High']. In which category do you find the most companies?
   - Which sector has the highest average revenue?