Data engineering project

Project Instructions and Evaluation Criteria:

You are individually expected to take up one of the two problem statements discussed below. We shall evaluate your submissions based on

1. Solution approach: scoping the problem, identifying data requirements, understanding set of tasks for the problem, defining metric for analysis (10 Marks)
2. Technical Implementation: your approach to coding, handling, storing as well as efficiently using data/ data structures (20 Marks)
3. Analysis: implementing relevant descriptive analysis with Python or SQL (should you choose to do it with MySQL tables), providing thought and implementation for answering specific business insight questions (20 Marks)

Problem Statements

1. National Cancer Institute maintains statistics of cancer – type of cancer, rate of new diagnosis, death rate, survival statistics by various factors such as stage of detection, age, race etc. The data is available over last 30 years. Can you web-scrape the data, organize the data in relevant format (files or tables) and provide analysis for the following questions –
   1. Types of cancer that will be research priority given adversarial impact it can have on society
   2. Cancer treatments that have seen significant improvements either in detection or treatments over last many years
   3. Cancer incidences by differences races – suggest what specific targeted education or awareness may be required among different communities.

The data is available to scrape at <https://seer.cancer.gov/statfacts/>

1. Travel and hospitality business has been badly affected last two years due to Covid-19 pandemic. We want to create an index for attractiveness of tourist places in post Covid world – we want to understand if certain cities or tourist attractions more badly impacted on account of Covid-19 pandemic? We do not have survey data; the idea is to scrape TripAdvisor website – assuming travellers are no less enthusiastic in writing reviews in post-covid world; write a code deciding
   1. Relevant pages to scrape from TripAdvisor
   2. How do you arrive at some index for attractiveness of a tourist spot – remember we have to examine it with respect to pre-covid world
   3. Which national attractions have remained equally attractive in post covid world? Which religious public places see definitive decline in tourist interests?
   4. Are specific cities or states badly affected in terms of tourist attractiveness post Covid?
   5. Make relevant choices of which data to scrape, how to save it and use it for the analysis.