DS8003 Final Project Description

Goal: To perform some data analysis using the different big data tools and report results

Requirements:

1. To use the tools learnt in at least 4/10 lectures in this course in the project
2. Find a problem definition that makes sense for the dataset you have chosen
3. Produce 5 insights or findings from the data that relate to the problem definition. You are welcome to use visualizations or statistics or other forms of reporting to showcase your results
4. Each of the 5 insights should have used one or more of the tools that were taught in this course.
5. You are welcome to use additional tools that are outside of the scope of the course, but it does not count towards the 4/10 requirement.
6. We have provided some sample datasets with some ideas, but you are welcome to choose another dataset entirely (provided it is in a similar scale to what's been provided as samples).
7. You are expected to work in groups of 4 people in the project.
8. Put your group information as well as the selected topic on the spreadsheet on this link: https://docs.google.com/spreadsheets/d/1O2H7Zn-daL_Tokh9QNGX7Q0CpOwzAyjYQqrz2Eg2_rc/edit?usp=sharing

Rubric:

1. Project Proposal (1-2 pages) (5 points)
    a. Problem Definition - 2 points
    b. Suggested solution (focus on how it will satisfy the requirements) - 2 points
    c. Dataset descriptions - describe some of the key attributes you want to potentially use in your analysis and open source link to download the dataset - 1 point
2. **Project Document (at most 20 pages; 12 point and 1.5 line spacing; Calibri)** (25 points)
    a. Problem Definition - 1 points
    b. Data Description - 2.5 points
        i. Attributes description - 0.5 point
        ii. Statistics of the data (use the tools learnt in this course to generate the data statistics) - 2 points
    c. Work distribution
    d. Solution description - 13 points (3.25 points per tool usage)
        i. Tools used
        ii. How they were used

iii. Why they were chosen for that particular problem space
iv. Code snippets and explaining the logic behind the code snippets

e. Describing the 5 Insights gathered from data - 7.5 points (1.5 per each insight)
f. Future Work (what are the next steps for this work) - 1
g. References (-1 if not present)

3. **Final Presentation (~6-8 slides) (10 points)**
   a. Each team will have 15+5 min
   b. Presentation should contain following information
      i. Description of the problem
      ii. Work Distribution among team members
      iii. Data sources with information
      iv. Description of the processing/analysis and the tools that were used
      v. Walk-through your analysis and your insights
      vi. Lessons Learned
   c. I will time and stop you at 15 min on the dot.
   d. At the end of 15 min, I will open it out to the class for 5 min Q & A.
4. Voting (5 points bonus)
   a. At the end of both the presentation sessions, each class member will vote for their favorite presentation
   b. Most votes get 5 points, 2nd place gets 4, 3rd most votes get 3 points, rest get 2 points each
   c. Do not vote for your own projects :)

Project Topics and Datasets (examples)

Note: Don't forget, you can propose your dataset as well. You need to get approval from the instructor for your dataset.

Water Quality and Potability

This dataset contains water quality measurements and assessments related to potability, which is the suitability of water for human consumption. The dataset's primary objective is to provide insights into water quality parameters and assist in determining whether the water is potable or not. Each row in the dataset represents a water sample with specific attributes, and the "Potability" column indicates whether the water is suitable for consumption.
https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability

Spotify Top Hit Playlist (2010-2022)

The data is extracted through Spotify API directly. The dataset includes information on songs/tracks (100 per year) from Top Hit playlists from 2010 to 2022 created by Spotify.
https://www.kaggle.com/datasets/josephinelsy/spotify-top-hit-playlist-2010-2022

Used Cars Price Prediction

Used Car Price Prediction Dataset is a comprehensive collection of automotive information extracted from the popular automotive marketplace website, https://www.cars.com. This dataset comprises 4,009 data points, each representing a unique vehicle listing, and includes nine distinct features providing valuable insights into the world of automobiles.
https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset

Airline Dataset

It encompasses flight details, passenger profiles, and preferences, aiding airlines in optimizing services. Analyzing delay and cancellation data helps enhance punctuality and passenger experiences. Additionally, regulatory authorities depend on this data to ensure safety standards and formulate aviation policies. Researchers use it to study market trends, environmental impacts, and sustainable growth strategies. In summary, airline data is the cornerstone of informed decision-making, operational excellence, and the progression of the aviation sector. This dataset comprises diverse parameters relating to airline operations on a global scale. These columns collectively provide comprehensive insights into passenger demographics, travel details, flight routes, crew information, and flight statuses.
https://www.kaggle.com/datasets/iamsouravbanerjee/airline-dataset

Cancer Database

Cancer DB is a public domain database and website containing structured data on key concepts in cancer integrated into one model. Someone's new treatment could by just 2 insights away from being the next great cure. Cancer DB can help them figure out what those 2 missing insights are.
https://www.kaggle.com/datasets/sujaykapadnis/cancer-database

Smoking and Drinking Dataset with Body Signal

This dataset is collected from the National Health Insurance Service in Korea. It contains the body signals and people’s health status and information of being a drinker or smoker or not.
https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset

E-commerce Customer Data For Behavior Analysis

The "E-commerce Customer Behavior and Purchase Dataset" is a synthetic dataset generated using the Faker Python library. It simulates a comprehensive e-commerce environment, capturing various aspects of customer behavior and purchase history within a digital marketplace.
https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis

Work University Ranking 2023

The World University Rankings 2023 dataset include 1,799 universities across 104 countries and regions, making them the largest and most diverse university rankings to date. The table is based on 13 carefully calibrated performance indicators that measure an institution's performance across four areas: teaching, research, knowledge transfer and international outlook. This year's ranking analyzed over 121 million citations across more than 15.5 million research publications and included survey responses from 40,000 scholars globally. Overall, we collected over 680,000 datapoints from more than 2,500 institutions that submitted data.
https://www.kaggle.com/datasets/alitaqi000/world-university-rankings-2023

National Health and Nutrition Examination Survey

NHANES, a program by the National Center for Health Statistics, assesses the health and nutrition status of U.S. adults and children through interviews and physical examinations. Starting in the 1960s, it transitioned to a continuous program in 1999. Each year, NHANES examines a nationally representative sample of approximately 5,000 individuals across 15 counties. The survey covers demographics, socioeconomic factors, dietary habits, and health-related questions in interviews. Additionally, it includes medical, dental, physiological measurements, and lab tests conducted by trained medical staff. NHANES plays a vital role in tracking and addressing evolving health and nutrition priorities in the United States.
https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

Amazon Sales Dataset

The Dataset Contains 1K+ Amazon Product's Ratings and Reviews
https://www.kaggle.com/datasets/ahmedsayed564/amazon-sales-dataset

Life expectancy & Socio-Economic (world bank)

Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. It is a key metric for assessing population health.
Life expectancy has burgeoned since the advent of industrialization in the early 1900s and the world average has now more than doubled to 70 years. Yet, we still see inequality in life expectancy across and within countries.
https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank

FitBit Fitness Tracker Data

This dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.

https://www.kaggle.com/datasets/nurudeenabdulsalaam/fitbit-fitness-tracker-data

LinkedIn Job Postings Dataset

This dataset contains information about job postings on LinkedIn. The data is divided into several files, each containing different aspects of the job postings
https://www.kaggle.com/datasets/rajatraj0502/linkedin-job-2023

World Energy Consumption

This dataset is a collection of key metrics maintained by Our World In Data. It includes data on energy consumption (primary energy, per capita, and growth rates), energy mix, electricity mix and other relevant metrics.
https://www.kaggle.com/datasets/pralabhpoudel/world-energy-consumption

Netflix Movies and TV Shows

They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.
https://www.kaggle.com/datasets/shivamb/netflix-shows

Apple (AAPL) Stock Data

This dataset offers a granular look at the daily Open, High, Low, Close, and Volume (OHLCV) data for some of the Apple Inc. stock (AAPL) Updated every trading day, it aims to serve as a resource for researchers, traders, and finance enthusiasts
https://www.kaggle.com/datasets/guillemservera/aapl-stock-data