**Group Project**

Your group (max size is 3) will pick a real-world dataset from annexure 1 and apply the concepts learn in Data preparation and transformation for modeling. You are required to use Jupyter notebook which contain coding and assignment report. So, **you are required to make sure to include detailed explanations wherever possible using Markdown cells**.

**Evaluation Criteria**

Your submission will be evaluated using the following criteria:

• Dataset must contain at least 4 columns and 200 rows of data

• You must ask and answer at least 4 questions about the dataset

• Your submission must include at least 5 visualizations graphs

• Your submission must include explanations using markdown cells, apart from the code.

• Your work must not be plagiarized.

**Follow this step-by-step guide to work on your project (Total 30 marks)**

**Step 1: Select a real-world dataset**

* Find a one of the interesting datasets from Annexure 1
* The data should be in CSV format, and should contain at least 4 columns and 200 rows

**Step 2: Perform data preparation & cleaning (7 marks)**

* Load the dataset into a data frame using Pandas
* Explore the number of rows & columns, ranges of values etc.
* Handle missing, incorrect and invalid data
* Perform any additional steps (parsing dates, creating additional columns, merging multiple datasets etc.)

**Step 3: Perform exploratory analysis & visualization (7 marks)**

* Compute the mean, sum, range and other interesting statistics for numeric columns
* Explore distributions of numeric columns using histograms etc.
* Explore relationship between columns using scatter plots, bar charts etc.
* Make a note of interesting insights from the exploratory analysis

**Step 4: Ask & answer questions about the data (11 marks)**

* Ask at least 4 interesting questions about your dataset
* Answer the questions either by computing the results using Numpy/Pandas or by plotting graphs using Matplotlib/Seaborn
* Create new columns, merge multiple datasets and perform grouping/aggregation wherever necessary
* Wherever you're using a library function from Pandas/Numpy/Matplotlib etc. explain briefly what it does

**Step 5: Summarize your inferences & write a conclusion (5 marks)**

* Write a summary of what you have learned from the analysis
* Include interesting insights and graphs from previous sections

**Step 6: Make a submission & share your work (due date 17th April 2023)**

* Upload your notebook in project section of Brightspace.

**Annexure 1**

* Video Games sales: <https://www.kaggle.com/gregorut/videogamesales>
* World University Rankings: <https://www.kaggle.com/mylesoneill/world-university-rankings>
* Netflix Tv shows and Movies: <https://www.kaggle.com/shivamb/netflix-shows/notebooks>
* StackOverflow Developer Survey: <https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey>
* Google Play Store Android Apps Data: <https://www.kaggle.com/lava18/google-play-store-apps>
* Indian Stock Market Data: <https://www.kaggle.com/rohanrao/nifty50-stock-market-data>
* Indian Air Quality: <https://www.kaggle.com/rohanrao/air-quality-data-in-india>
* Worldwide Covid-19 Cases: <https://www.kaggle.com/imdevskp/corona-virus-report>
* USA Covid-19 Cases: <https://www.kaggle.com/sudalairajkumar/covid19-in-usa>
* US Election Results (2012): <https://www.kaggle.com/tunguz/us-elections-dataset>
* US Stock Market: <https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs/>
* Crop production in India: <https://www.kaggle.com/srinivas1/agricuture-crops-production-in-india>
* Agricultural raw material prices: <https://www.kaggle.com/kianwee/agricultural-raw-material-prices-19902020>
* Agricultural land values: <https://www.kaggle.com/jmullan/agricultural-land-values-19972017>
* Digital payments in India: <https://www.kaggle.com/lazycipher/upi-usage-statistics-aug16-to-feb20>
* US Unemployment Rate Data: <https://www.kaggle.com/jayrav13/unemployment-by-county-us>
* India Road accident Data: <https://community.data.gov.in/statistics-of-road-accidents-in-india/>
* Data Science Jobs Data:
  + <https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us>
  + <https://www.kaggle.com/jonatancr/data-science-jobs-around-the-world>
  + <https://www.kaggle.com/rkb0023/glassdoor-data-science-jobs>
* Youtube Trending Videos: <https://www.kaggle.com/datasnaek/youtube-new>
* Asteroid Dataset: [https://www.kaggle.com/sakhawat18/asteroid-dataset](https://www.kaggle.com/sakhawat18/asteroid-dataset )
* Solar flares Data: [https://www.kaggle.com/khsamaha/solar-flares-rhessi](https://www.kaggle.com/khsamaha/solar-flares-rhessi )
* F-1 Race Data: <https://www.kaggle.com/cjgdev/formula-1-race-data-19502017>
* Automobile Insurance: [https://www.kaggle.com/aashishjhamtani/automobile-insurance](https://www.kaggle.com/aashishjhamtani/automobile-insurance )
* PUBG video game matches: [https://www.kaggle.com/skihikingkevin/pubg-match-deaths](https://www.kaggle.com/skihikingkevin/pubg-match-deaths )
* CounterStrike GO (video game)
  + <https://www.kaggle.com/mateusdmachado/csgo-professional-matches>
  + <https://www.kaggle.com/skihikingkevin/csgo-matchmaking-damage>
* Dota 2 (video game): <https://www.kaggle.com/devinanzelmo/dota-2-matches>
* Cricket One-Day Internationals Data: [https://www.kaggle.com/jaykay12/odi-cricket-matches-19712017](https://www.kaggle.com/jaykay12/odi-cricket-matches-19712017 )
* Cricket Indian Premier League Data: <https://www.kaggle.com/nowke9/ipldata>
* Basketball (NCAA): [https://www.kaggle.com/ncaa/ncaa-basketball](https://www.kaggle.com/ncaa/ncaa-basketball )
* Basketball NBA Players Stats: <https://www.kaggle.com/ncaa/ncaa-basketball>
* Football datasets:
  + [https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017](https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017 )
  + <https://www.kaggle.com/abecklas/fifa-world-cup>
  + [https://www.kaggle.com/egadharmawan/uefa-champion-league-final-all-season-19552019](https://www.kaggle.com/egadharmawan/uefa-champion-league-final-all-season-19552019 )
* Hotel Booking Demand: [https://www.kaggle.com/jessemostipak/hotel-booking-demand](https://www.kaggle.com/jessemostipak/hotel-booking-demand )
* New York Airbnb listings: [https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data )