Capstone 3 Ideas

Option 1

* Topic: Adult Income
* Source: https://www.kaggle.com/wenruliu/adult-income-dataset
* Background: An individual’s annual income results from various factors. Intuitively, it is influenced by the individual’s education level, age, gender, occupation, and etc..
* Data Description: The dataset contains 15 columns, 48842 rows.
* The problem statement here is to predict whether the income exceeds 50k a year or not based on the census data.

Option 2

* Topic: National Stock Exchange: Time Series
* Source: https://www.kaggle.com/atulanandjha/national-stock-exchange-time-series
* Background: The National Stock Exchange of India Ltd. (NSE) is an Indian stock exchange located at Mumbai, Maharashtra, India. National Stock Exchange (NSE) was established in 1992 as a demutualized electronic exchange. It was promoted by leading financial institutions on request of the Government of India. It is India’s largest exchange by turnover. In 1994, it launched electronic screen-based trading. Thereafter, it went on to launch index futures and internet trading in 2000, which were the first of its kind in the country.
* Data Description: The dataset contains three csv files. Each resembling to INFOSYS, NIFTY*IT*INDEX, and TCS, respectively. One can easily identify that by the name of CSV files.
  + Colum Descriptors:
  + Date: date on which data is recorded
  + Symbol: NSE symbol of the stock
  + Series: Series of that stock | EQ - Equity
  + Prev Close: Last day close point
  + Open: current day open point
  + High: current day highest point
  + Low: current day lowest point
  + Last: the final quoted trading price for a particular stock, or stock-market index, during the most recent day of trading.
  + Close: Closing point for the current day
  + VWAP: volume-weighted average price is the ratio of the value traded to total volume traded over a particular time horizon
  + Volume: the amount of a security that was traded during a given period of time. For every buyer, there is a seller, and each  
    transaction contributes to the count of total volume.
  + Turnover: Total Turnover of the stock till that day
  + Trades: Number of buy or Sell of the stock.
  + Deliverable: Volumethe quantity of shares which actually move from one set of people (who had those shares in their demat account before today and are selling today) to another set of people (who have purchased those shares and will get those shares by T+2 days in their demat account).
  + %Deliverble: percentage deliverables of that stock
* The problem statement here is to predict Indian IT companies stock prices.

Option 3

* Topic: Spotify dataset
* Source: <https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks>
* Background: This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.
* Data Description:

#### **Primary**:

* - **id** (Id of track generated by Spotify)

#### **Numerical**:

* - **acousticness** (Ranges from 0 to 1)
* - **danceability** (Ranges from 0 to 1)
* - **energy** (Ranges from 0 to 1)
* - **duration\_ms** (Integer typically ranging from 200k to 300k)
* - **instrumentalness** (Ranges from 0 to 1)
* - **valence** (Ranges from 0 to 1)
* - **popularity** (Ranges from 0 to 100)
* - **tempo** (Float typically ranging from 50 to 150)
* - **liveness** (Ranges from 0 to 1)
* - **loudness** (Float typically ranging from -60 to 0)
* - **speechiness** (Ranges from 0 to 1)
* On the basis of current data, try analyzing which features play the most important role in determining the popularity of the song. Divide the popularity into 10 categories. And determine the rating outcome based on the other features.