**TEAM # 4 – CAPSTONE PROJECT**

Veena Puvvada

**Four stages of the Project:**

1. **First Segment: Sketch It Out:** Decide on your overall project, select your question, and build a simple model. You'll connect the model to a fabricated database, using comma-separated values (CSV) or JavaScript Object Notation (JSON) files, to prototype your idea.
2. **Second Segment: Build the Pieces:** Train your model and build out the database you'll use for your final presentation.
3. **Third Segment: Plug It In:** Connect your final database to your model, continue to train your model, and create your dashboard and presentation.
4. **Fourth Segment: Put It All Together:** Put the final touches on your model, database, and dashboard. Lastly, create and deliver your final presentation to your class.

#### Technologies Used

#### Data format:

#### CSV

#### SQL

#### Data Cleaning and Analysis

#### Python will be used to clean, prepare, and explore the data, as well as to complete initial analysis.

#### Pandas will be used to clean the data and perform an exploratory analysis. Further analysis will be completed using Python.

#### Database Storage

#### Use database integration (Postgres, MongoDB, or SQLite) to store your cleaned data.

#### Mongo is the database we intend to use, and we will integrate Flask to display the data.

#### Postgres

#### If you use a SQL database, you must provide your Entity Relationship Diagram (ERD) with relationships.

#### Machine Learning

#### SciKitLearn is the ML library we'll be using to create a classifier. Our training and testing setup is \_\_\_. Extra ML verbiage here.

#### Dashboard/Visualization

#### Python libraries – Numpy, Matplotlib, Pandas

#### R – Ggplot2

#### JavaScript libraries (such as Data-Driven Documents, or D3, and Plotly), and Tableau can be used to create visuals

#### Tableau

#### In addition to using a Flask template, we will also integrate D3.js for a fully functioning and interactive dashboard. It will be hosted on \_\_\_.

#### Programs/Tools used:

#### GitHub

#### Images/Screenshots

#### Data Sets

<https://www.springboard.com/blog/data-science/15-fun-datasets-to-analyze/>

1. **Trending Shows on Streaming Platforms**

With so many streaming platforms to choose from, viewers have plenty of choices. From new releases to enduring favorites, the most-streamed shows make for an ever-changing dataset, and are often reflective of the current cultural zeitgeist (remember when Tiger King inspired all those [pandemic-related memes?](https://time.com/5810608/tiger-king-memes/)). Using [this dataset](https://www.kaggle.com/prasertk/netflix-daily-top-10-in-us) on Netflix’s top 10 shows from March 2020 to March 2022, you can analyze what people were binge-watching throughout the COVID-19 pandemic.

<https://www.kaggle.com/datasets/prasertk/netflix-daily-top-10-in-us>

10 columns X 7101 rows

1. **Students Performance in Exams**

This data set consists of the marks secured by the students in various subjects.

Example Research Questions:

* How effective is the test preparation course?
* Which major factors contribute to test outcomes?
* What would be the best way to improve student scores on each test?
* What patterns and interactions in the data can you find? Let me know in the comments section below.

<https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams>

#### 8 columns X 1001 rows

#### Amazon Data Science Books Dataset

* The dataset contains 946 books obtained from scraping [Amazon](https://www.amazon.com/) books related to data science, statistics, data analysis, Python, deep learning, and machine learning.  
  There are 18 columns:
* title: title of the book
* author: author (or the authors) of the book
* price: price (in dollars)
* pages: number of pages
* avg\_reviews: average reviews (out of 5)
* n\_reviews: reviews done for each book
* star5: percentage of 5 star reviews
* star4: percentage of 4 star reviews
* star3: percentage of 3 star reviews
* star2: percentage of 2 star reviews
* star1: percentage of 1 star reviews
* dimensions: size of the book (in inches)
* weight: weight (in pounds or ounces)
* language: language of the book
* publisher: publisher
* ISBN-13: ISBN\_13 code
* link: link of the Amazon book
* complete\_link: complete link of the Amazon book (including the domain https://amazon.com)

<https://www.kaggle.com/datasets/die9origephit/amazon-data-science-books>

18 columns X 947 rows

1. **Superstore Sales Dataset**

Retail dataset of a global superstore for 4 years.  
Perform EDA and Predict the sales of the next 7 days from the last date of the Training dataset!

<https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting?select=train.csv>

18 rows X 9801 rows

1. **YouTube Data Science Video Trends**

<https://www.kaggle.com/datasets/sandhyakrishnan02/youtube-datascience-video-views>

7 columns X 1559 columns