# CSE 351 - Introduction to Data Science (Summer 2024)

## Course Project

**Due Date: Course Project is due by 11:59 PM New York Time on Monday July 1st**

This homework will reinforce the concepts taught in the class in all lectures, familiarizing you with the real-world examples of these concepts.

### Kaggle and Github Resources

There are plenty of resources available on the internet such as Kaggle and Github. You are encouraged to take advantage of such resources in order to supplement your learning experience and gain new skills. However, the direct use (copying) of implemented codes are **prohibited** unless specified otherwise in the problem description.

### Use of Online Resources

You are allowed to use online resources, including **Generative AI (GAI)** tools such as Chat **GPT** on the conditions that

1. you will list the used sources and the exact prompt in the case of GAI.
2. you will **not** use the **whole or any parts of question** or **exercise** as the prompt (you can paraphrase and break down the questions into parts).
3. you will **not** directly copy the GAI answer.
4. you will **double check** GAI answers and list a **reference** (other than the GAI) for that.

Report the used resources at the last page of your assignment, under a section called **Used Resources** and use in-text citation for referencing.

### Project Report

You need to report your project results in the following ways

1. **Code** in *pdf* and *ipynb* format
2. **Project Presentation** in Powerpoint format

There is no need for a separate written report, but you are required to document your work in the ipynb notebook. If you decide not to use Ipython notebooks, you need to submit a py file and an accompanying PDF of the py file, plus an additional PDF file to make up for written text and figures, etc. Please simply merge the PDF that contains your text and figures with the PDF of your python script.

### Submission Guidelines

1. This assignment must be done by the students in each group. Your answers and codes will be checked thoroughly to detect copying/plagiarism. Do your own work!
2. If you do not have much experience with Python and the associated tools, we suggest familiarizing yourself using resources like: Python.org, Python REU Tutorial by Prof. Michael Zingale
3. Please use **Piazza** to ask any questions.
4. Submit everything through **BrightSpace**. One person from your group will need to upload:
   1. a **PPT** of your presentation (the cover page should include **all the group members' full legal names - as stated in SBU ID**).
   2. **Python** file (**ipynb**)
   3. **PDF** (your **Python** file)

   These files should be named with the following format:
   1. cse351_project_Group_number.**pdf**
   2. cse351_project_**code**_Group_number.**py**
   3. cse351_project_**demo**_Group_number.**ppt**

**Please keep in mind that:**

**(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.**

**(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.**

## Project #1: Movie Revenue Prediction (2-3 People)

Film industry is booming, the revenues are growing. There are many factors which affect the revenue of a film. In this project, you will explore what features can help to predict the revenue.

**Datasets:**

The TMDB dataset is to be used for this project. You will need to read the dataset description on *Kaggle* which describes the dataset metadata, its curation and various recorded features/columns. Then through preprocessing steps of your choice (data cleaning may be needed), gain additional insight into the dataset. You must split the dataset into training and testing.

**EDA (10 points):**

Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Clean the dataset, remove the outliers, before any data analysis. Explain what you did and provide justification for them.
- Some of the columns contain lists and dictionaries. Extract information you need and reformat them.
- Count the number of movies released by day of week, month and year, are there any patterns that you observe?
- What are the movie genre trend shifting patterns that you can observe from the dataset?
- What are the strongest and weakest features correlated with movie revenue?
- You can also use some external datasets to integrate into your revenue prediction analysis to make it better.

**Modeling and Question Answering (10 points):**

1. Extract the features you think are necessary in predicting the movie revenue.
2. Build three models, train them on the training set, and predict the revenue on the test set (after dropping the revenue column in the test set).
3. Explain how each model works (briefly introduce the machine learning algorithms behind them).
4. Evaluate the performance of each model based on the original outcome in the test set.
5. If your predictions are not so accurate, what do you think is the reason?
6. Report your accuracy using metrics such as Residual Standard Error (RSE).
7. Split the data further to include a cross validation set. Did this improve your model's performance on the test set?

**Project Report (10 points):**

The suggested developing environment would be *Ipython notebooks* such as *Jupytor* and *Google Colab*, as they allow simultaneous online editing and access to cloud accelerators such as GPU and TPU, and inclusion of text, figures and code script within the same environment in the notebook.

1. You are required to document your project. This includes usage of written answers, plots and figures. If you decide not to use Ipython notebooks, you need to submit a py and an additional PDF file to make up for written text and figures, etc. (see Project Report section on the first page)
2. Make your code readable through comments, modulations etc.
3. Don't forget to include the team members' contribution information in the documentation.
4. Include visualizations to explain your results and prove your point in your conclusion.
5. You should prepare a powerpoint presentation for your demo. You should address all tasks and questions asked in previous sections in your presentation.

**Demo (5 points):**

- All the team members should be present during the demo.
- Be prepared to answer questions related to your work.
- You should present your findings for the project.
- You should also be able to run your code.

## Project #2: Titanic - Who will survive? (1-2 People)

Titanic, a British passenger liner, sank in the North Atlantic Ocean on 15 April 1912, after striking an iceberg during her maiden voyage from Southampton to New York City. Of the estimated 2,224 passengers and crew aboard, more than 1,500 died, making the sinking at the time one of the deadliest of a single ship and the deadliest peacetime sinking of a superliner or cruise ship to date. This project allows us to gain insight into how to survive from such a catastrophe, is it pure luck or is it something else?

**Datasets:**
The Titanic dataset is to be used for this project. You will need to read the dataset description on *Kaggle* which describes the dataset metadata, its curation and various recorded features/columns. The dataset is already splitted into training and testing sets.

**EDA (10 points):**
Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Clean the dataset, remove the outliers, before any data analysis. Explain what you did.
- Explore the socio-economic status of the passenger, is there any relationship between socio-economic status with other features, such as age, gender, number of family members on board, etc.
- Explore the distribution of survival victims in relation to age, gender, socioeconomic class, etc.
- What features seem to be the most important ones? Perform a correlation analysis before your prediction task.
- How can you extract information from the non-numerical features?

**Modeling and Question Answering (10 points):**

1. Build three models, train them on the training set, and predict the outcome on the test set (after dropping the survival column in the test set).
2. Explain how each model works (briefly introduce the machine learning algorithms behind them).
3. Evaluate the performance of each model based on the original outcome in the test set.
4. If your predictions are not so accurate, what do you think is the reason?
5. Use other evaluation metrics to evaluate your models (Precision, Recall, F-score).
6. Split the data further to include a cross validation set. Did this improve your model's performance on the test set?

## Project #3: Fatal Force in the US (2-3 people)

In the United States, use of deadly force by police has been a high-profile and contentious issue. 1000 people are shot and killed by US cops each year. The ever-growing argument is that the US has a flawed Law Enforcement system that costs too many innocent civilians their lives. In this project, we will analyze one of America's hottest political topics, which encompasses issues ranging from institutional racism to the role of Law Enforcement personnel in society.

**Datasets:**
The Fatal Police Shootings in the US (2015-2020) dataset is to be used for this project. You will need to read the dataset description on *Kaggle* which describes the dataset metadata, its curation and various recorded features/columns. Then through preprocessing steps of your choice (data cleaning may be needed), gain additional insight into the dataset. You must split the dataset into training and testing.

**EDA (10 points):**
Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Clean and merge the datasets, explain what you did.
- Which state has the most fatal police shootings? Which city is the most dangerous?
- What is the most common way of being armed?
- What is the age distribution of the victims? Compare age distribution of different races.
- Compare the total number of people killed per race. Compare the number of people killed per race as a proportion of respective races. What difference do you observe?

**Modeling and Question Answering (10 points):**

1. Apply three machine learning algorithms to explore whether it is possible to predict the race of a victim based on other features.
2. Train your models on the training set, and make predictions for the test set with the "race" column dropped.
3. Evaluate the accuracy of your predictions.
4. If your predictions are not very accurate, what do you think is the reason?

## Project #4: What makes people in a country happy? (2-3 people)

The World Happiness Report is a landmark survey of the state of global happiness that ranks countries by how happy their citizens perceive themselves to be. The report gains global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. This project allows us to gain insight into the state of happiness in the world today.

**Datasets:**
The World Happiness Report dataset for different countries from year 2015 to year 2019. We will treat data of year 2015 to year 2018 as the training set, and year 2019 data as the test set. Description of the data fields can be found on the FAQ page of World Happiness Report: https://worldhappiness.report/faq/

**EDA (10 points):**
Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Merge and clean the data. Explain what you did.
- What are the central tendencies of happiness score over the years? Did they increase or decrease?
- Which countries have stable rankings over the years? Which countries improved their rankings?
- Visualize the relationship between happiness score and other features such as GDP, social support, freedom, etc.
- Find out what features contribute to happiness. If you are the president of a country, what would you do to make citizens happier?

**Modeling and Question Answering (10 points):**
The happiness rankings in the datasets are determined by happiness scores only. Now we want to predict the ranking using a machine learning approach.

1. Build three models based on data from year 2015 to year 2018.
2. Explain how each model works (briefly introduce the machine learning algorithms behind them).
3. Predict the happiness ranking for the year 2019 (drop the “overall rank” and “score” columns first).
4. Compare your rankings to the original rankings in “2019.csv”. How does each model perform?
5. Invent your own formula to calculate happiness score using features of your choice.

## Project #5: Can we predict whether a Hotel Booking will be canceled?

When it comes to hotel bookings, customers have a variety of options and deals and are sometimes often canceling certain bookings for several reasons. Given hotel booking data for two major hotels, can we predict whether a customer will cancel the booking or not? We will explore the main concepts of EDA and modeling classification algorithms in this project.

**Datasets:**
The Hotel Bookings dataset is to be used for this project. You will need to read the dataset description on *Kaggle* which describes the dataset metadata, its curation and various recorded features/columns. Then through preprocessing steps of your choice (data cleaning may be needed), gain additional insight into the dataset. You must split the dataset into training and testing.

**EDA (10 points):**
Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Which country saw the most hotel bookings according to the data?
- What is the distribution like for both hotels with respect to the price of a room per night?
- Which months are the most busy for both hotels? Which months see the most expensive per night costs?
- Which months see the most cancellations for both hotels?
- Examine distributions of bookings vs market segment.
- Which room type was most commonly booked? Most commonly canceled?
- What percentage of the data recorded cancellations for each hotel?

**Modeling and Question Answering (10 points):**

1. Apply three machine learning algorithms to predict whether or not a customer will cancel a booking.
2. Train your models on the training set, and make predictions for the test set with the "is_canceled" and "reservation_status" columns dropped.
3. Evaluate the accuracy of your predictions.
4. If your predictions are not so accurate, what do you think is the reason? Use other evaluation metrics to evaluate your models (Precision, Recall, F-score).
5. Split the data further to include a cross validation set. Did this improve your model's performance on the test set?

## Project #6: Covid-19 in Germany Analysis (2-3)

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease has since spread worldwide, leading to an ongoing pandemic. In this project, you will explore the Covid cases in Germany.

**Datasets:**
The COVID-19 Tracking Germany dataset is to be used for this project. YYou will need to read the dataset description on *Kaggle* which describes the dataset metadata, its curation and various recorded features/columns. Then through preprocessing steps of your choice (data cleaning may be needed), gain additional insight into the dataset. You must split the dataset into training and testing.
Here is the additional data that might be helpful for this project:
https://github.com/GoogleCloudPlatform/covid-19-open-data

**EDA (10 points):**
Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Clean the dataset, remove the outliers, before any data merging and analysis. Explain what you did.
- What is the covid case trend in Germany, and how is it different from each state/county? Which state/county has the highest/lowest increasing rate?
- What is the covid death rate trend in Germany, and how is it different from each state/county? Which state/county has the highest/lowest increasing rate?
- Which age/gender group has the highest covid positive cases?
- Which age/gender group has the highest covid death cases?
- What contributes to the spreading of the covid cases in Germany? (Additional datasets probably will be helpful)

**Modeling and Question Answering (10 points):**

1. Apply three machine learning algorithms to explore whether it is possible to predict whether the covid patient would survive .
2. Train your models on the training set, and make predictions for the test set with the “death” column dropped.
3. Evaluate the accuracy of your predictions.
4. If your predictions are not very accurate, what do you think is the reason? Use other evaluation metrics to evaluate your models (Precision, Recall, F-score).
5. Split the data further to include a cross validation set. Did this improve your model’s performance on the test set?