1. Income and Debt Relationship:
    a. Project Description: This project investigates the relationship between income and debt in the United States from 2003 to 2016, utilizing data from various Federal Reserve sources and SSTI. Analysis and data visualization were conducted in R to explore potential correlations.

    b. Summary/Synopsis: The project aimed to determine the correlation between median household income and debt in the US, covering the years 2003 to 2016. Data from multiple reputable sources were imported, cleaned, and analyzed using R, focusing on average income and debt at the national level. Visual analysis, represented by combined line and point graphs, revealed a surprising positive correlation—contrary to initial expectations—that as income rose, so did debt levels. The analysis suggested that increased confidence from higher incomes may lead households to take on more debt. Limitations included the scope of the analysis and my level of proficiency with R, which influenced the project approach and simplification of certain aspects.

2. NBA Player performance:
    a. Project Description: In this project, I examined how the number of minutes played in an NBA game impacts player performance, focusing particularly on points scored. Using Python, I conducted exploratory data analysis and regression analysis to answer the statistical question.

    b. Summary/Synopsis: I set out to investigate the impact of minutes played on NBA player performance, using 'Minutes Played' as the primary independent variable and assessing metrics such as field goals made, field goals attempted, field goal percentage, and points scored. Through Python-based exploratory data analysis (EDA), including cumulative distribution functions, probability mass functions, scatterplots, and Pearson correlations, I gained deeper insights into variable relationships. The regression analysis concluded with an R-squared value of 0.731, indicating that 73.1% of the variability in points scored is explained by minutes played. This result confirmed my initial assumptions. While the analysis was comprehensive, I discovered a need to better understand analytical distribution modeling for future projects.

3. Correlation Between Crime and Incarceration:
    a. Project Description: I explored trends and connections between crime and incarceration in the U.S. over recent decades, using Python to analyze data from multiple sources, including Wikipedia, Kaggle, and the FBI Crime Data API. This project involved merging datasets to draw meaningful insights while addressing ethical concerns tied to data representation and transparency.

    b. Summary/Synopsis: This project aimed to analyze the relationship between crime rates and incarceration trends in the U.S., using three datasets spanning from 2001 to 2020. I utilized Python for data wrangling, exploration, and merging, which involved challenges, particularly in converting JSON data to Pandas DataFrames and integrating the tables. Despite the datasets being comprehensive and published by the FBI, ensuring the accuracy and ethical integrity of the analysis was vital. Ethical considerations included the risk of bias, data manipulation, and impacts on public trust. The project highlighted the importance of careful data handling and transparent methodologies to support responsible crime and incarceration analyses.

4. Bank Credit Card Churn:
    a. Project Description: For my project on bank credit card churn, I aimed to predict which customers were likely to stop using the bank's credit card services, based on customer data. Using Python code, I analyzed and prepared the data, then developed and evaluated machine learning models to identify churn indicators. The findings can assist banks in taking proactive measures to retain customers and minimize churn.

    b. Summary/Synopsis: The "Bank Credit Card Churn" project used a dataset of 10,000 customers with 18 features sourced from Kaggle to identify customers at risk of churning. Python code was employed to clean and preprocess the data, generate exploratory visualizations, and build predictive models. A Gradient Boosting Classifier initially showed suspiciously perfect results due to over-reliance on a single feature, so I proceeded with a logistic regression model. This model achieved an accuracy of 87.17%, effectively identifying non-churn cases but requiring further refinement for better churn prediction. Feature importance analysis highlighted key factors influencing churn, such as the number of customer contacts and months of inactivity. The project underscored the potential and challenges in using machine learning for churn prediction, suggesting further model testing and data balancing to enhance performance.

5. Key Predictors of Heart Disease:
    a. Project Description: For my Predictive Analytics project, I focused on heart disease prediction, using Python code to analyze a comprehensive dataset from IEEE.org. The goal was to identify the main predictors of heart disease and evaluate the performance of predictive models like Logistic Regression and Random Forests.

    b. Summary/Synopsis: Heart disease is a significant public health issue in the United States, responsible for one in every five deaths annually. To address this, I conducted an analysis using a heart disease dataset that merged information from multiple international sources, consisting of 1,190 instances and 11 features. Leveraging Python, I explored data relationships through visualizations and built predictive models. The Random Forest model outperformed Logistic Regression, achieving a higher accuracy and precision rate. Key predictors identified included ST slope, chest pain type, oldpeak, and maximum heart rate. While further validation is needed for deployment, these findings highlight the potential for machine learning to aid early heart disease detection and prevention.

6. Time Series, Clustering, etc…:
    a. Summary/Synopsis: This folder contains a collection of assignments showcasing my proficiency in time series analysis, clustering, developing a recommender system, and leveraging data-driven insights to improve MLB team attendance. For instance, in the MLB attendance project, key strategies were identified using OLS regression to enhance crowd numbers, such as the significant impact of promotional events like shirt and bobblehead giveaways. Additionally, I developed a recommender system capable of suggesting movies based on user input, demonstrating my capability in building personalized recommendation solutions.

7. Airline travelling:
    a. Project Description: For this project, I demonstrated my ability to turn data insights into visualizations that effectively communicate with different audiences, including managers, colleagues, and the public. The goal was to show, using various data tools, that air travel is still safe despite recent airplane crashes.

    b. Summary/Synopsis: This project addressed the challenge of countering negative media claims about the safety of air travel due to recent crashes. As a data scientist at an airline, I created a series of data-driven visualizations and presentations to reassure my colleagues, management, and the public that we are currently living in the safest era for air travel. The tasks included developing a dashboard for coworkers, a PowerPoint for managers, a blog for public dissemination, an infographic for public display, and a video for media broadcast. Tools such as WIX, Canva, Power BI, PowerPoint, Jupyter Notebook, and Python were utilized to create these engaging visualizations.

8. Bankruptcy Predictions:
    a. Project Description: For this project, I used Python code to predict bankruptcy for public companies listed on the NYSE and NASDAQ, identifying the key financial and operational features that signal potential bankruptcy. My aim was to enhance risk assessment models for stakeholders such as investors and financial institutions.

    b. Summary/Synopsis: The project "Bankruptcy Prediction for Public Companies" focused on addressing the risk posed by potential bankruptcies in public companies. By using Python, I analyzed a dataset of accounting data from 8,262 companies between 1999 and 2018, splitting it into training and testing sets to train a LightGBMClassifier. While the model achieved an overall accuracy of 93.44%, it struggled with detecting bankrupt companies, as shown by a low recall of 0.04 for the "failed" class. This highlighted the challenge of class imbalance and suggested a bias toward predicting companies as "alive." Feature importance analysis indicated that retained earnings, inventory, and market value were significant predictors. The project concluded with recommendations for improving model performance, future applications for financial forecasting, and ethical considerations to ensure responsible usage.

9. Predicting Rainfall:
    a. Project Description: I completed a project that focused on predicting rainfall at London-Heathrow Airport using machine learning and Python. By analyzing meteorological data from 2000-2010, I developed models to help improve rain forecasting accuracy, crucial for sectors like aviation and logistics.

    b. Summary/Synopsis: This project aimed to address the business problem of enhancing rainfall prediction at London-Heathrow Airport, a critical hub where weather miscalculations can lead to costly flight delays and operational disruptions. Using a dataset sourced from Kaggle, originally provided by the European Climate Assessment & Dataset (ECA&D), I analyzed daily weather data spanning 2000-2010 with Python. After preparing the data, I applied logistic regression and random forest models to classify rainfall based on meteorological features. The random forest model performed best, achieving an accuracy of 78%, and identified 'pressure' and 'humidity' as the most influential factors. This project demonstrated that machine learning can significantly enhance weather prediction, supporting better decision-making and minimizing disruptions in weather-dependent operations.

10. Female Malnutrition in Asia and Sub Saharan Africa:
    a. Project Description: I worked on a project titled Predicting and Analyzing Key Factors Influencing Underweight Women to identify which factors contribute most to underweight women aged 15 to 49 using Python. The project used machine learning techniques to analyze a dataset with various environmental, geographical, and socio-economic features, providing insights into public health strategies.

    b. Summary/Synopsis: This project aimed to understand and predict the factors influencing underweight women (BMI < 18.5) aged 15 to 49, particularly in 11 priority countries as identified by USAID's Feed the Future initiative. Leveraging Python code, I processed the dataset and employed a random forest regressor, optimizing it with GridSearchCV to uncover critical features such as longitude, latitude, travel time to urban centers, and socio-economic indicators. Visual tools like SHAP values and bar charts were used to interpret feature importance. This analysis provides actionable insights to help guide public health policies and targeted interventions for malnutrition reduction and improved well-being of women in these regions.