Final Project Overview and Rubric:

Objective: The goal of this project is to apply exploratory data analysis and basic statistical techniques to real-world data within the topic each team choses.

Description: Students will work with publicly available data from a reputable source that contains relevant data to their topic. From this, each group will create a well rounded analysis on their topic of choice and present their findings to the class. Analysis should be done in Python or R, **not in excel**. You may not use a dataset/project route with solutions already available online. You also may not use chatgpt or other such AI generative tools for your complete solution. However, you can use them as an aid just as you would any other resource as long as you provide appropriate attribution. Follow the guidance in the syllabus and/or synopsis for such use.

Tasks:

- **Designate Team Member Roles:** Teams should designate their team lead and team scribe clearly in their presentation.
- **Data Preprocessing:** Students will clean the data and perform necessary preprocessing steps. This may include handling missing values and removing outliers.
- **Exploratory Data Analysis:** Students will explore the data to understand the distribution of different features, identify any outliers, and formulate hypotheses about the properties of their data.
- **Statistical Analysis:** Students will apply basic statistical techniques to test their hypotheses. This might include t-tests, chi-square tests, correlation analysis, or regression analysis.
- **Presentation:** Each team will prepare a presentation explaining their approach, their findings, and any challenges they faced during the project. They should also discuss the implications of their work for the field.

Deliverables:

- A Jupyter notebook (or equivalent) containing all code used for data preprocessing, exploratory data analysis, and statistical analysis
- A final report detailing the team's approach, findings, and a discussion of the results (5-8 pages minimum)
- A presentation summarizing the project

Grading Rubric:

Final Report (50% of grade)

- Clarity and organization of the report: 10%
- Quality of data preprocessing: 10%
- Depth and quality of exploratory data analysis: 15%
- Appropriateness and execution of statistical analysis: 10%
- Quality of conclusions and discussion: 5%

Presentation (50% of grade)

- Clarity and organization of the presentation: 5%
- Quality of presentation / presentation skills demonstrated: 5%
- Quality of visualizations: 10%
- Depth of understanding demonstrated: 15%
- Quality of conclusions and discussion: 10%
- Teamwork and collaboration: 5%

Topics:

Astronomy: Students can work with datasets from NASA or other space agencies to analyze celestial bodies or phenomena. For example, they could use machine learning to classify galaxies based on their shapes, or analyze the light curves of stars to detect exoplanets

- Descriptive Statistics and Visualization
    - Calculate summary statistics for key variables (e.g., star brightness, distance).
    - Create visualizations:
        - Scatter plots: Explore relationships between star properties.
        - Histograms: Understand distributions of star characteristics.
        - Sky maps: Plot star positions in the Milky Way.
- Exploring Populations
    - Investigate different types of stars:
        - Categorize stars based on spectral types (O, B, A, etc.).
        - Analyze luminosity and temperature variations.
    - Compare properties of stars in globular clusters and open clusters.
- Analysis
    - Time Series Analysis
        - Plot light curves over time.
        - Identify periodic patterns.
    - Spatial Analysis
        - Formation Regions
- Machine Learning Applications (if applicable)
    - Regression:
        - Predict star properties (e.g., luminosity) based on other features.
    - Classification:
        - Classify stars into different categories (e.g., main sequence, giants).
    - Clustering:
        - Group stars based on similarity (e.g., spectral characteristics).
- Example Datasets:
    - https://data.nasa.gov/browse?limitTo=datasets
    - https://catalog.data.gov/dataset/national-space-science-data-center-master-catalog
    - https://astrostatistics.psu.edu/MSMA/datasets/

Forensic Accounting: Students can analyze financial data to detect fraudulent activities. This could involve looking for patterns or anomalies in large datasets, or building predictive models to identify high-risk transactions.

- Descriptive Statistics and Summary Metrics
    - Calculate basic summary statistics (mean, median, standard deviation, etc.) for key variables.
    - Visualize data distributions using histograms, box plots, and scatter plots.
- Exploring Data Patterns
    - Investigate relationships between variables:
        - Correlation analysis: Identify correlations between financial variables.
        - Scatter plots: Visualize pairwise relationships.
    - Explore time-series patterns (if applicable):
        - Plot financial data over time (e.g., revenue, expenses, cash flow).
- Apply EDA techniques specifically for fraud detection:
    - Benford’s Law analysis: Check if numerical data follows expected distribution patterns.
    - Anomaly detection: Identify unusual transactions or outliers.
    - Segmentation analysis: Group data based on relevant criteria (e.g., customer segments, transaction types)
- Create informative visualizations:

    - Heatmaps, bar charts, and pie charts to highlight patterns.
    - Geographic maps (if location data is available).
    - Time-series plots for trend analysis.
- Reporting and Insights
    - Summarize key findings from the EDA.
    - Highlight potential areas of concern or suspicious patterns.
    - Provide actionable recommendations for further investigation.
- Example Datasets:
    - https://www.sec.gov/dera/data/financial-statement-data-sets
    - https://www.mergentonline.com/companyfinancials.php?compnumber=8865 (Must sign in using UARK credentials to access company data)

Bioinformatics: Students can analyze DNA sequence data to identify genes associated with certain traits or diseases. This could involve data preprocessing, exploratory data analysis, and implementing machine learning algorithms.

- Investigate specific features relevant to bioinformatics:
    - Genomic Variation: Analyze mutations, SNPs, and indels.
    - Gene Expression: Explore differential expression patterns.
    - Protein Domains: Identify conserved regions.
    - Clinical Markers: Investigate associations with diseases.
- Look for outliers or unusual patterns:
    - Gene Outliers: Identify genes with extreme expression values.
    - Structural Aberrations: Detect anomalies in protein structures.
    - Clinical Outliers: Investigate unusual patient profiles.
- Interpret findings:
    - Biological Significance: Relate EDA results to existing knowledge.
    - Hypothesis Generation: Formulate hypotheses for further analysis.
    - Data Limitations: Discuss any limitations or biases.
- Summarize your EDA process and key observations:
    - Visual Reports: Include plots and figures.
    - Narrative Summary: Explain insights in plain language.
- Example Datasets
    - https://bioinformatics.mdanderson.org/public-datasets/
    - https://data.jrc.ec.europa.eu/collection/id-0052
    - https://catalog.data.gov/dataset/forensic-dna-open-dataset-a26bc

Customer Segmentation for Marketing Optimization: Customer segmentation is a powerful technique in data science and marketing that involves dividing a customer base into distinct groups based on shared characteristics. These segments allow businesses to tailor their strategies, personalize marketing efforts, and optimize resource allocation

- Explore the dataset to gain insights:
    - Calculate summary statistics for key variables (e.g., average purchase amount, frequency of interactions).
    - Visualize customer distributions using histograms, scatter plots, or bar charts.
    - Identify any patterns or outliers.
- Choose relevant features for segmentation:
    - Behavioral Features: Purchase frequency, recency, average transaction value.
    - Demographic Features: Age, gender, location.
    - Psychographic Features: Interests, preferences, lifestyle.
- Develop personalized approaches for each segment:
    - Targeted Campaigns: Create specific promotions or offers based on segment preferences.
    - Channel Optimization: Allocate resources effectively (e.g., focus on social media for one segment, email for another).

    - Customer Lifetime Value (CLV): Predict future value and prioritize high-value segments.
- Present your findings visually:
    - Segment profiles (bar charts, pie charts)
    - Heatmaps showing feature correlations
    - Customer journey maps
- Example Datasets
    - https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset
    - https://data.world/carmichael/consumer-spending

Predictive Sales Analysis: Many retail companies work to improve their sales forecasting accuracy by collecting historical sales data, including transaction dates, product IDs, quantities sold, and other relevant features. Using a relevant dataset, students can develop a predictive model that forecasts future sales based on historical data (EC) and/or use basic EDA techniques to explore the dataset, identify patterns, and gain insights to provide a thorough sales analysis.

- Explore the data:
    - Load the data into a suitable tool (e.g., pandas, NumPy).
    - Summarize the data: check for missing values, data types, and basic statistics.
    - Explore the distribution of sales, customer demographics, product categories, etc.
    - Visualize key features using histograms, scatter plots, box plots, etc
- Feature Engineering:
    - Create relevant features for analysis (e.g., total sales per customer, average sales per product).
    - Consider time-based features (e.g., day of the week, month, season) for seasonality analysis.
- Time Series Analysis
    - Investigate sales trends over time (daily, weekly, monthly).
    - Apply techniques like moving averages, exponential smoothing, or decomposition.
    - Identify any seasonality or cyclic patterns.
- Correlation Analysis
    - Explore relationships between sales and other variables (e.g., weather, holidays, promotions).
    - Calculate correlation coefficients and visualize
- Predictions:
    - Use your findings to form conclusions around how sales will be influenced by the various factors you have identified
    - OR you can build a ML model to predict sales for you (make sure to evaluate model performance and interpret feature importance)
- Example Datasets
    - https://data.cityofnewyork.us/Housing-Development/DOF-Summary-of-Neighborhood-Sales-Citywide-for-Cla/hdu7-ujt4/about_data
    - https://data.world/dataman-udit/us-regional-sales-data
    - https://www.mergentonline.com/companyfinancials.php?compnumber=8865 (Must sign in using UARK credentials to access company data)

Housing Price Predictions: Many real estate agencies work to enhance their property valuation accuracy by collecting historical housing sales data, including features like square footage, location, number of bedrooms, and other relevant variables. Using a relevant dataset, students can develop a predictive model that estimates housing prices based on historical data (EC) and/or use basic EDA techniques to explore the dataset, identify patterns, and gain insights to provide a thorough analysis of what features impact a housing price.

- EDA:
    - Load the data into a suitable tool (e.g., pandas, Jupyter Notebook)

    - Summarize key statistics, explore distributions, and visualize features (histograms, scatter plots, etc.).
    - Investigate correlations between features and the target variable (price).
- Feature Engineering
    - Create relevant features (e.g., total square footage, neighborhood characteristics).
    - Handle categorical variables (one-hot encoding, label encoding).
    - Consider time-based features (e.g., year built, recent renovations).
- Analysis
    - Determine which features are most impactful on housing prices
    - Use this to make conclusions on housing prices
        - Check against housing records
    - OR you can build a ML model to predict prices for you (make sure to evaluate model performance and interpret feature importance)
- Example Datasets:
    - https://www.nar.realtor/research-and-statistics/housing-statistics/existing-home-sales
    - https://data.ct.gov/Housing-and-Development/Real-Estate-Sales-2001-2021-GL/5mzw-sjtu/about_data
    - https://www.zenrows.com/datasets/us-real-estate