Data Science Group Project

1. Group Formation

- Form groups of **3–4 members**. Working alone is allowed but **not recommended**.
- Each group is responsible for **task coordination, meeting deadlines, submitting all deliverables, and giving the stage-wised & final presentation**.

---

2. Topic Selection

- You may choose a topic from the **provided project list** or propose your own.
- If proposing your own topic, make sure:
    - The dataset is **open-source** and **publicly accessible**.
    - The dataset is **large enough** and has **meaningful features** for analysis and modeling.
    - It is suitable for **all four project stages**.

Think Before You Act!

- What project will your team work on?
- Is your dataset appropriate and fair for analysis (avoiding bias)?
- Changing topics later in the semester will be difficult — **plan carefully from the start**.

Team Formation & Proposal Submission

- **Form your teams** this week and begin in-depth discussions about your chosen project.
- **Create a team GitHub repository** and add all members **and me** as collaborators.
- **Submit a one-page project proposal** including:
    - **Team name**
    - **Team members** (full names, emails, and GitHub usernames)
    - **Project title**
    - **Dataset source** (website link or resource)
    - **Brief project description** – how you understand the problem and what aspects you plan to study.

Titanic Survival Prediction

Dataset: https://www.kaggle.com/c/titanic

Description: Predict survival based on passenger demographics and ticket info.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Predict survival based on passenger demographics and ticket info.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Movie Revenue Forecasting

Dataset: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

Description: Predict movie revenue or rating using metadata.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Predict movie revenue or rating using metadata.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Fake News Detection

Dataset: https://www.kaggle.com/c/fake-news

Description: Classify articles as real or fake using NLP + ML.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Classify articles as real or fake using NLP + ML.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Global Suicide Rates Analysis

Dataset: https://www.kaggle.com/szamil/who-suicide-statistics

Description: Analyze and model suicide rates worldwide.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Analyze and model suicide rates worldwide.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

COVID-19 Impact on Air Travel

Dataset: https://ourworldindata.org/covid-deaths

Description: Analyze pandemic effects on air travel demand.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Analyze pandemic effects on air travel demand.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Spotify Song Popularity Prediction

Dataset: https://www.kaggle.com/datasets/maharshipandya/-spotify-datasets

Description: Predict song popularity using audio and metadata.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Predict song popularity using audio and metadata.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Flight Delay Prediction

Dataset: https://www.kaggle.com/datasets/usdot/flight-delays

Description: Predict flight arrival delays using airline and weather data.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Predict flight arrival delays using airline and weather data.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Crop Yield Prediction

Dataset: https://www.kaggle.com/datasets/faoallfoodagriculture/crop-production

Description: Forecast crop yields using weather and soil data.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Forecast crop yields using weather and soil data.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

YouTube Trending Video Analysis

Dataset: https://www.kaggle.com/datasets/datasnaek/youtube-new

Description: Analyze factors contributing to trending status of videos.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Analyze factors contributing to trending status of videos.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

AI Tools Popularity & Sentiment

Dataset: https://huggingface.co/

Description: Analyze adoption trends and sentiment of AI tools.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Analyze adoption trends and sentiment of AI tools.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Housing Price Prediction

Dataset: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Description: Predict house prices using structural and location features.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Predict house prices using structural and location features.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Global Renewable Energy Trends

Dataset: https://ourworldindata.org/renewable-energy

Description: Analyze renewable energy growth and forecast adoption.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Analyze renewable energy growth and forecast adoption.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

E-Commerce Product Review Analysis

Dataset: https://www.kaggle.com/datasets/bittlingmayer/amazonreviews

Description: Sentiment classification and trend analysis of reviews.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Sentiment classification and trend analysis of reviews.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Climate Change & Extreme Weather Events

Dataset: https://www.ncdc.noaa.gov/cdo-web/

Description: Predict occurrence/severity of extreme weather events.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Predict occurrence/severity of extreme weather events.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables

Traffic Accident Severity Prediction

Dataset: https://www.kaggle.com/sobhanmoosavi/us-accidents

Description: Predict severity of accidents using weather, road, and time data.

1. Problem Statement

Describe the main problem the project aims to solve based on the dataset: Predict severity of accidents using weather, road, and time data.

2. EDA Checklist

- Explore key variables and their distributions
- Check for missing values and outliers
- Examine relationships between main features and target variable
- Group data by relevant categories and compare statistics
- Visualize correlations with heatmaps, boxplots, histograms

3. Hypothesis Testing Ideas

- Apply t-tests, ANOVA, or Chi-square tests depending on feature types
- Test whether differences in target variable across categories are significant
- Correlation tests for numerical features

4. ML/DL Model Suggestions

- Baseline: Logistic Regression, Decision Tree, Linear Regression
- Advanced: Random Forest, XGBoost, LightGBM, Neural Networks
- For time-series datasets: LSTM, GRU, Prophet

5. Visualization Suggestions

- Time-series line charts for temporal data
- Heatmaps for correlation or category combinations
- Boxplots for category vs numerical variable comparison
- Bar charts for frequency counts
- Scatter plots for two continuous variables