# Machine Learning Courses

## Advanced Data Analysis and GitHub Workflow

Please work on the following questions below and write a detailed report to be summited and make available online.

## Hour 1: Deep Data Cleaning & GitHub Workflow

### Task 1: Data Quality Assessment (30 mins)

**Objective:** Ensure the dataset is clean and structured properly.

**Questions for Students**

- Are there any duplicate rows? If yes, why might duplicates exist in this dataset?
- Are there inconsistencies in categorical values (e.g., spelling variations, case sensitivity)?
- What should we do if a column has **too many missing values**?
- Are there outliers in numerical columns? If so, should we remove or adjust them?

**Activities:**

- Identify missing values & decide whether to **drop, fill, or flag** them.
- Check for duplicates and remove them if necessary.
- Standardize categorical values (e.g., ensuring "Male" and "male" are the same).
- Detect and handle **outliers** using box plots and the IQR method.

**GitHub Integration:**

- Students should commit their cleaned dataset as **version 1** in GitHub with a commit message:

  > "Cleaned dataset: removed duplicates, handled missing values, and standardized categories.

### Task 2: Dataset Documentation & Collaboration (30 mins)

**Objective:** Improve dataset readability for team collaboration.

**Questions for Students**

- If you were a **new data scientist** joining this project, what would you need to understand this dataset?
- How do you write effective **column descriptions**?
- What's the best way to track changes in a dataset over time?

**Activities:**

- Write a **README.md** file explaining the dataset.
- Create a **data dictionary** (column names, types, and descriptions).
- Document all cleaning steps taken so far.

**GitHub Integration:**

- Push the **README.md** file and commit changes with:

  "Added README and data dictionary for dataset clarity."

**Hour 2: Deep Exploratory Data Analysis (EDA) & GitHub Branching**

**Task 3: Relationship Analysis (30 mins)**

**Objective:** Move beyond individual variables and analyze relationships.

**Questions for Students**

- Does gender influence dating app preferences?
- Are people using multiple dating apps simultaneously?
- Which age group is most active on dating apps?
- Do people in urban vs. rural areas show different usage patterns?

**Activities:**

- **Correlation Matrix & Heatmap**: Identify relationships between numerical variables.
- **Pivot Tables & Groupby Aggregations**: Summarize data based on gender, age, and location.
- **Stacked Bar Charts**: Visualize categorical comparisons (e.g., dating app usage by gender).

**GitHub Integration:**

- Have students **create a new branch** (feature-EDA) and push their exploratory work separately.
- Encourage them to **open a pull request (PR)** for feedback.
- Discuss the importance of **branching in collaborative ML projects**.

**Task 4: Detecting Patterns & Bias in Data (30 mins)**

**Objective:** Identify potential biases or unexpected trends in the dataset.

**Questions for Students**

- Is there **gender bias** in dating app usage?
- Are certain demographics overrepresented in the dataset?
- Is there **a missing group of users** that should have been included?
- Can we trust this dataset for making general claims about Gen-Z dating in India?

**Activities:**

- Use **histograms** and **density plots** to compare distributions across demographic groups.
- Identify **overrepresented** and **underrepresented** groups.
- Discuss **ethical considerations** in data collection.

**GitHub Integration:**

- Students should update their PR with a new commit message:

  "Added bias detection analysis and demographic trends."

**Hour 3: Advanced Data Visualization & Feature Engineering**

**Task 5: Temporal & Regional Trends (30 mins)**

**Objective:** Analyze how dating app usage changes over time and location.

**Questions for Students**

- Are younger or older Gen-Z users more active on dating apps?
- Does dating app preference change over time?
- Do **metro city users** behave differently from users in smaller towns?

**Activities:**

- **Line Charts:** Track trends over time.
- **Geospatial Visualizations:** Show differences in app usage across cities/states.
- **Bubble Charts:** Represent app popularity across age groups.

**GitHub Integration**

- Students should merge their **feature-EDA** branch into main.
- Encourage **peer reviews** before merging.

**Task 6: Feature Engineering for Future Modeling (30 mins)**

**Objective:** Prepare data for potential machine learning models.

**Questions for Students**

- How can we convert categorical variables into numbers for ML models?
- Should we **normalize** numerical data? Why or why not?
- What new features could we create to enhance predictive modeling?

**Activities**

- Encode categorical variables using **One-Hot Encoding** or **Label Encoding**.
- Normalize numerical variables using **MinMaxScaler** or **StandardScaler**.
- Create a new feature, **"active app count"**, by summing the number of apps used per user.

**GitHub Integration:**

- Commit the engineered dataset with a message:

  "Feature engineering: Encoded categorical variables & added new features."

**Hour 4: Reflection, Documentation & GitHub Best Practices**

**Task 7: Collaborative Review & Documentation (30 mins)**

**Objective:** Strengthen teamwork and documentation skills.

**Questions for Students**

- What were the most surprising insights from this dataset?
- What would you do differently if given more time?
- How can we improve dataset documentation for others?

**Activities**

- Write a final summary in the **README.md** file.
- Discuss learnings in small groups.
- Review and merge all pull requests.

---

**Task 8: The Final Challenge (30 mins) – Interactive Q&A**

**The Challenge**

Students must answer these questions **without coding** – just by looking at the dataset & visualizations.

1. If a dating app wanted to expand into **rural India**, which insights from this dataset would be most valuable?
2. If you were designing a new dating app based on this data, what **two features** would you add?
3. What were the **biggest data cleaning challenges** in this dataset?

**Final Outcomes**

By the end of this session, students will have cleaned, explored, visualized, and documented real-world data while mastering GitHub collaboration.