# Assignment Title: Data Collection & Cleaning

# Project: Telecom Customer Churn Dataset

**Steps Taken:**

## 1. Data Loading & Initial Profiling

- Imported dataset in Jupyter Notebook
- Used Pandas Profiling / YData Profiling to generate an overview of:
- Data types
- Missing values
- Basic statistics
- Churn distribution (Target variable)

## 2. Basic Inspection

- print(df.shape)
- print(df.info())
- print(df.head())
- Initial Shape: (7043, 38)
- No duplicate rows found

## 3. Missing Values Handling

- print(df.isnull().sum())
- df = df.dropna()
- Some rows had missing values (especially in numeric usage columns)
- Removed rows with missing values
- After Shape: (6923, 38)

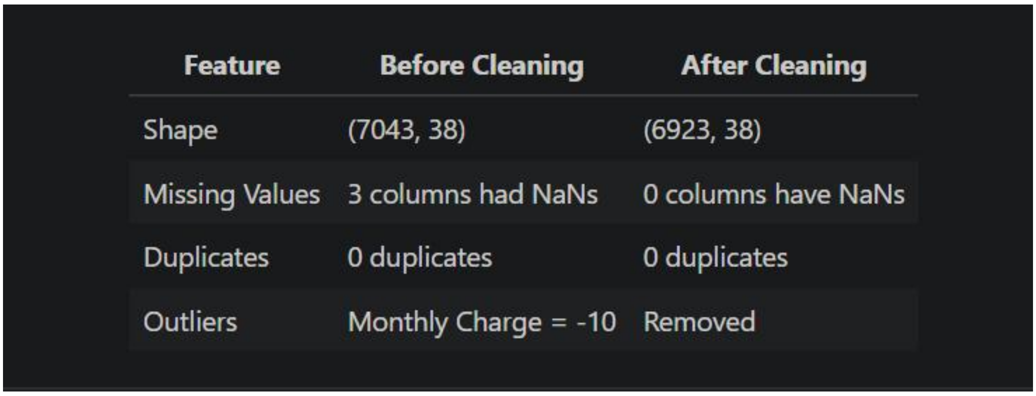
## 4. Statistical Summary & Outlier Check

- print(df.describe())
- Most values were normal
- One outlier detected: Monthly Charge = -10 (invalid negative billing)
- df = df[df['Monthly Charge'] >= 0]

## 5. Final Shape & Validation

- print(df.shape)
- Final Clean Shape: (6922, 38)
- No missing values
- No negative or invalid values remaining

## Output:



| Feature | Before Cleaning | After Cleaning |
|---|---|---|
| Shape | (7043, 38) | (6923, 38) |
| Missing Values | 3 columns had NaNs | 0 columns have NaNs |
| Duplicates | 0 duplicates | 0 duplicates |
| Outliers | Monthly Charge = -10 | Removed |

## Challenges Faced

- Understanding which columns were important for churn prediction
- Deciding whether to drop or fill missing values
- Identifying outliers (negative billing values were unusual)
- Ensuring that data cleaning didn’t remove valuable customer records

## GitHub Link:

**Repository Link Data-Science and AI**