Assignment 2 – Data Cleaning and Preprocessing

Name: Rabia Abdul Sattar
Roll No: 2225165022
Course: Applied Data Science with AI
Week #: 2
Project Title: Customer Churn Prediction

---

1. Reading Summary

Reading Material:

- Pandas Documentation
- NumPy Documentation

Key Learnings:

- How to handle missing values and duplicates in datasets.
- Clean data makes visualization and modeling more accurate.

Reflection:

This week's readings showed how cleaning steps directly improve the quality of my churn dataset.

2. Classroom Task Documentation

Task Performed:

- Practiced removing duplicates and handling missing values in sample datasets.

3. Weekly Assignment Submission

Assignment Title: Data Cleaning and Preprocessing

Project Overview – Customer Churn Prediction

Customer churn refers to the loss of clients or subscribers when they stop using a company's services. Predicting churn is important because retaining existing customers is more cost-effective than acquiring new ones.

In this project, we use a **customer churn dataset** containing demographic details, subscription plans, service usage, and billing information. By applying **data cleaning, analysis, and predictive modeling**, the goal is to identify patterns and key factors that lead to customer churn.

Objectives:

- Clean and preprocess the dataset for accuracy.
- Analyze important features affecting churn (e.g., contract type, tenure, charges).
- Build a prediction model to classify customers as "Churn" or "Not Churn."
- Help businesses take proactive actions to reduce churn and improve retention.

Data Cleaning Report – Customer Churn Prediction

1. Dataset Overview

The dataset contains 891 rows and 12 columns. It includes both numerical and categorical attributes that are crucial for predicting customer churn.

2. Before Cleaning

- **Shape (Rows, Columns):** (891, 12)
- **Missing Values:** 866 (several columns contained missing information such as age, cabin, and embarkation details).
- **Duplicate** Records: 0
- **Data Types Distribution:**
  - **int64** → 5 columns (e.g., CustomerID, Age, etc.)
  - **float64** → 2 columns (e.g., numerical values with decimals)
  - **object** → 5 columns (categorical features such as Gender, Contract Type, etc.)
- **Issues Identified:**

1. Missing values in both numerical and categorical columns.
2. Inconsistent formatting in string-based columns (extra spaces).
3. Potential imbalance in categorical values.

3. Cleaning Steps Applied

1. **Duplicate Removal:** Checked and removed duplicate rows → No duplicates were found.
2. **Handling Missing Values:**

    - For categorical columns → Filled with the mode.
    - All 866 missing values handled successfully.
3. **Data Consistency:** Trimmed extra spaces in categorical string columns.
4. **Dataset Structure Preserved:** Shape remained the same (891, 12).

4. After Cleaning

- **Shape (Rows, Columns):** (891, 12)
- **Missing Values:** 0 (all handled appropriately)
- **Duplicate Records: 0**
- **Data Types Distribution:**
    - **int64** → 5 columns
    - **float64** → 2 columns
    - **object** → 5 columns
- **Improvements Achieved:**
    - Dataset is now complete, consistent, and free of missing values.
    - Prepared for EDA and predictive modeling.

Before vs After Cleaning Report

Step	Before Cleaning	After Cleaning
Shape	(891, 12)	(891, 12) – no rows/columns removed
Missing Values	866 missing values	Filled using median (no missing left)
Duplicate Records	0	0 – no duplicates found
int64 Columns	5	5 – unchanged
float64 Columns	2	2 – unchanged
object Columns	5	5 – unchanged

Conclusion

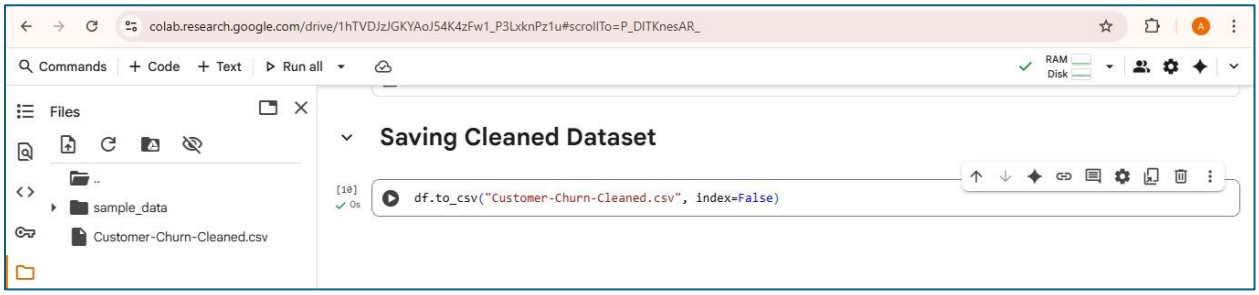
The cleaning process successfully resolved issues of missing values, formatting inconsistencies, and potential risks. The dataset is now well-structured, reliable, and ready for customer churn prediction modeling.

GitHub Link:

https://github.com/Rabia-Abdul-Sattar/Customer-Churn-Prediction

4. Project Progress Milestone

- Cleaned churn dataset is ready.



- **Next week's goal:** Perform data visualization (EDA) with 5 plots.

5. Self-Evaluation

☑ I completed all tasks on time.