Assignment 2 – Data Cleaning and Preprocessing

Name: Aman Nadeem
Roll No: 2225165002
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:

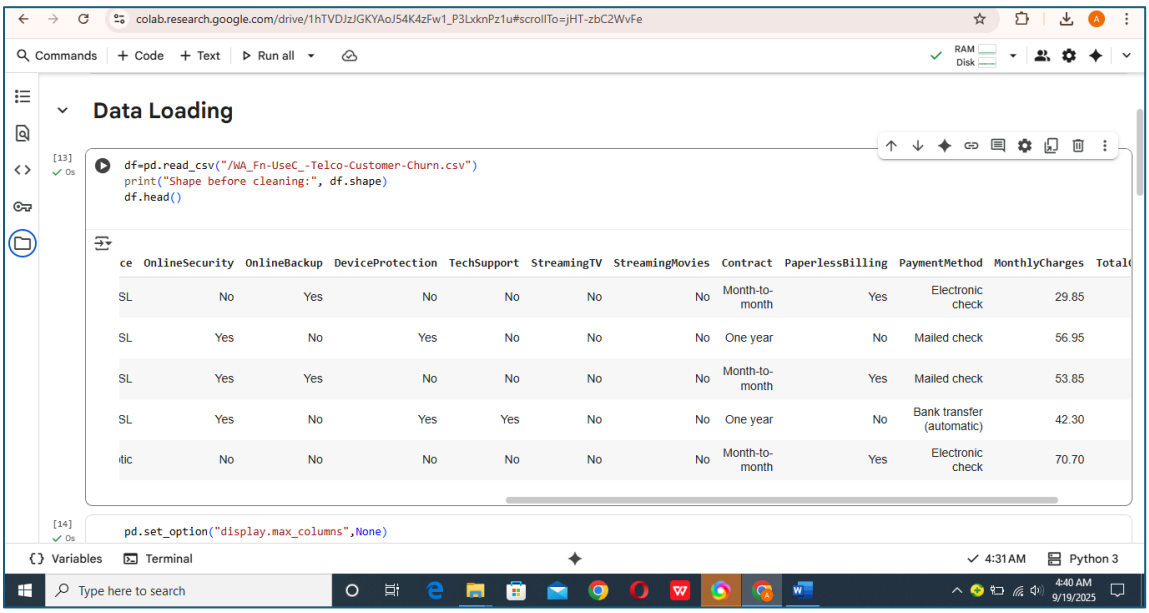
- Practiced removing duplicates and handling missing values in sample datasets.

3. Weekly Assignment Submission

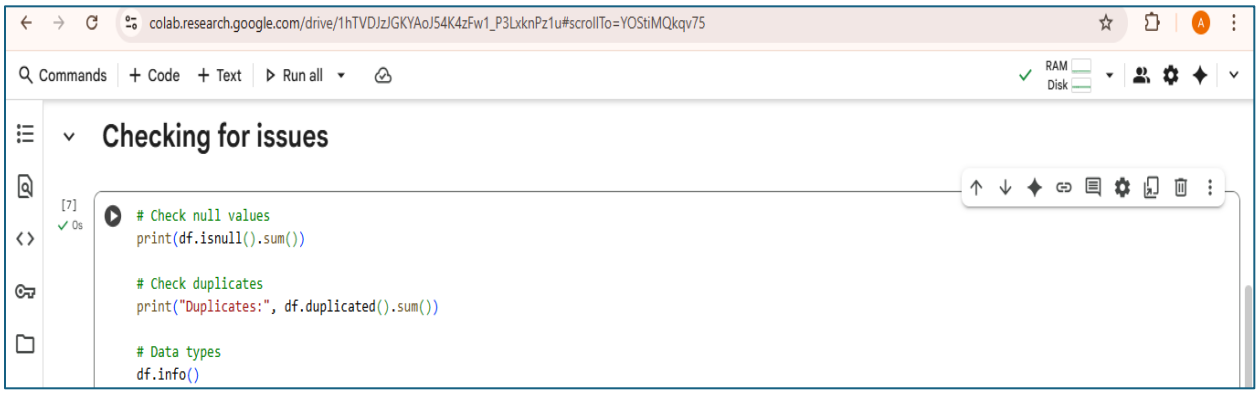
Assignment Title: Data Cleaning and Preprocessing

Steps Taken:

1. Loaded Telco Customer Churn dataset.



2. Checked for nulls, duplicates, and wrong data types.



3. Fixed TotalCharges column, removed NaNs and duplicates, dropped customerID.

Cleaning of the code

[8]

```python
# Convert 'TotalCharges' to numeric (some blank values cause issues)
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')

# Drop rows with missing values (only a few in this dataset)
df.dropna(inplace=True)

# Drop duplicate rows if any
df.drop_duplicates(inplace=True)

# Drop 'customerID' column (not useful for prediction)
df.drop(columns=['customerID'], inplace=True)

print("Shape after cleaning:", df.shape)
```

```
Shape after cleaning: (7032, 20)
```

Output:

- Before cleaning shape: ***7043 rows, 21 columns.***

```
Shape before cleaning: (7043, 21)
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	Strea
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	

ce	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
SL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
SL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
SL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
SL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
tic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

- After cleaning shape: ***7032 rows, 20 columns.***

Challenges Faced:

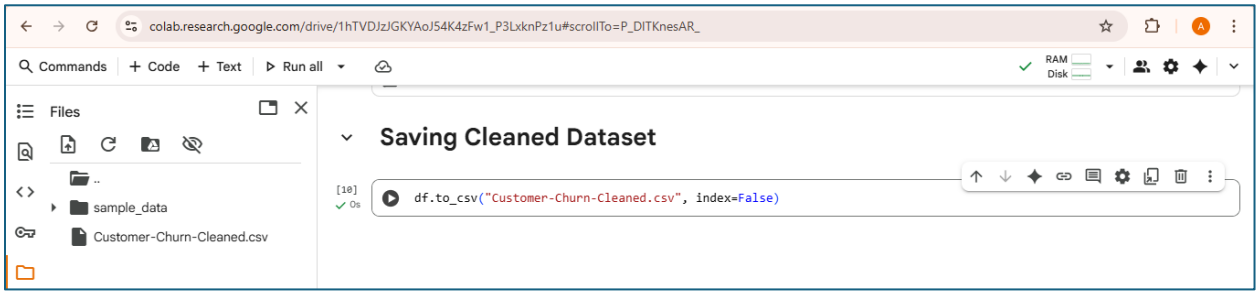
At first, TotalCharges was stored as string due to blank values. Solved it by converting with pd.to_numeric(errors="coerce").

GitHub Link:

https://github.com/amannadeem126/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.