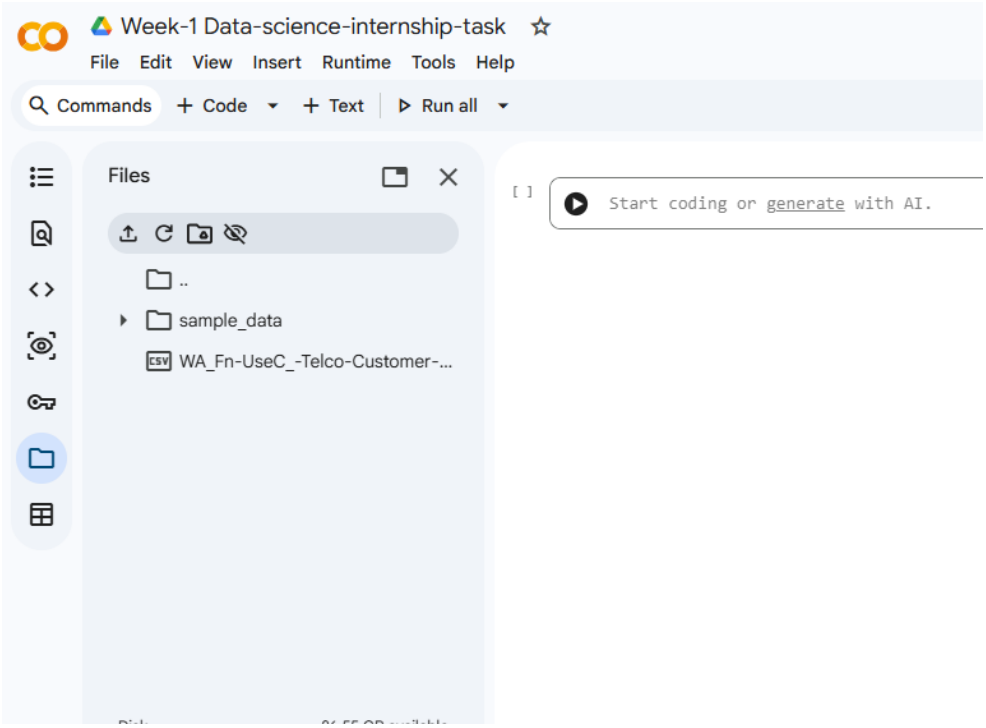
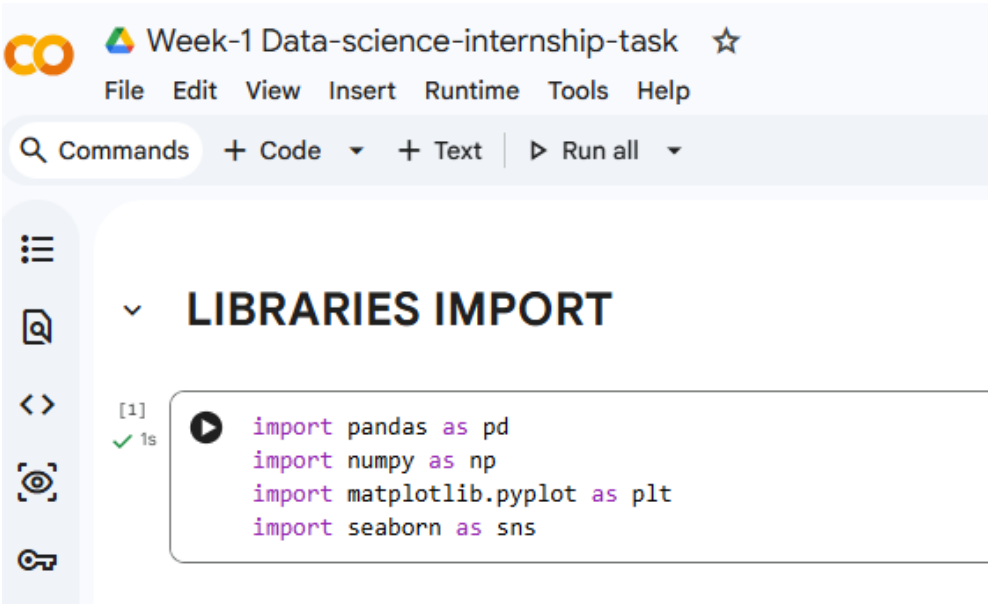
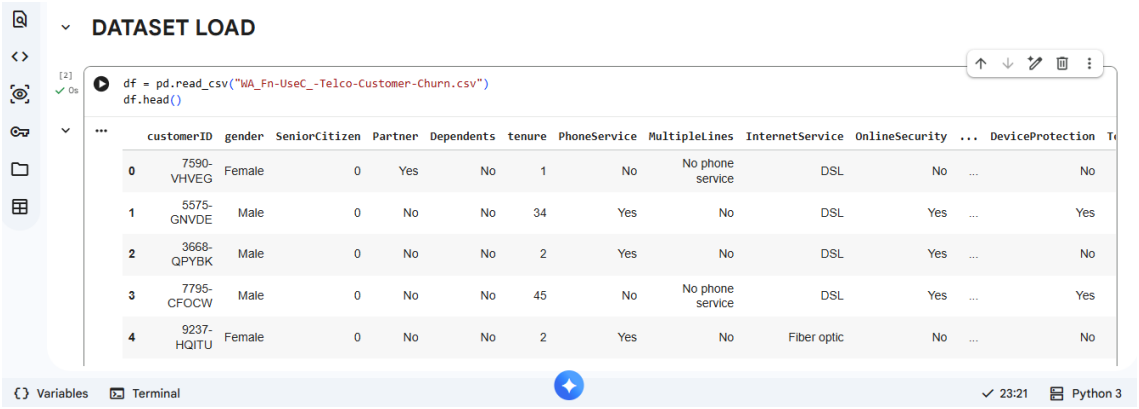
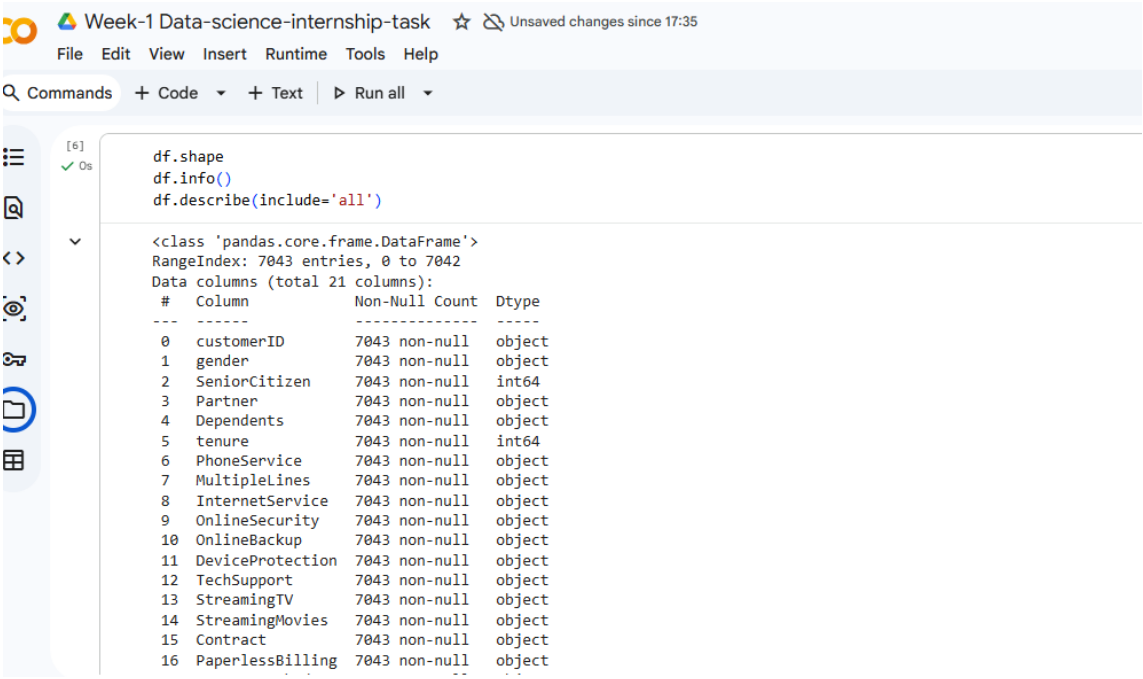
Introduction: This report documents the work completed for assignment in the Data Science Internship. The objective of this task was to perform advanced data cleaning and preprocessing on a real-world business dataset. The focus was on identifying data quality issues and applying professional techniques to make the dataset ready for analysis.

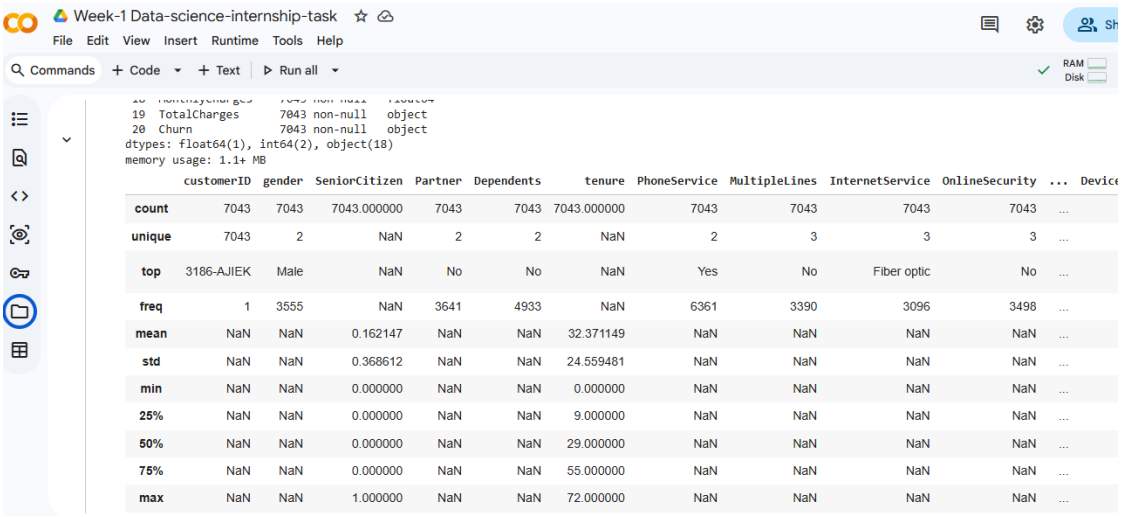






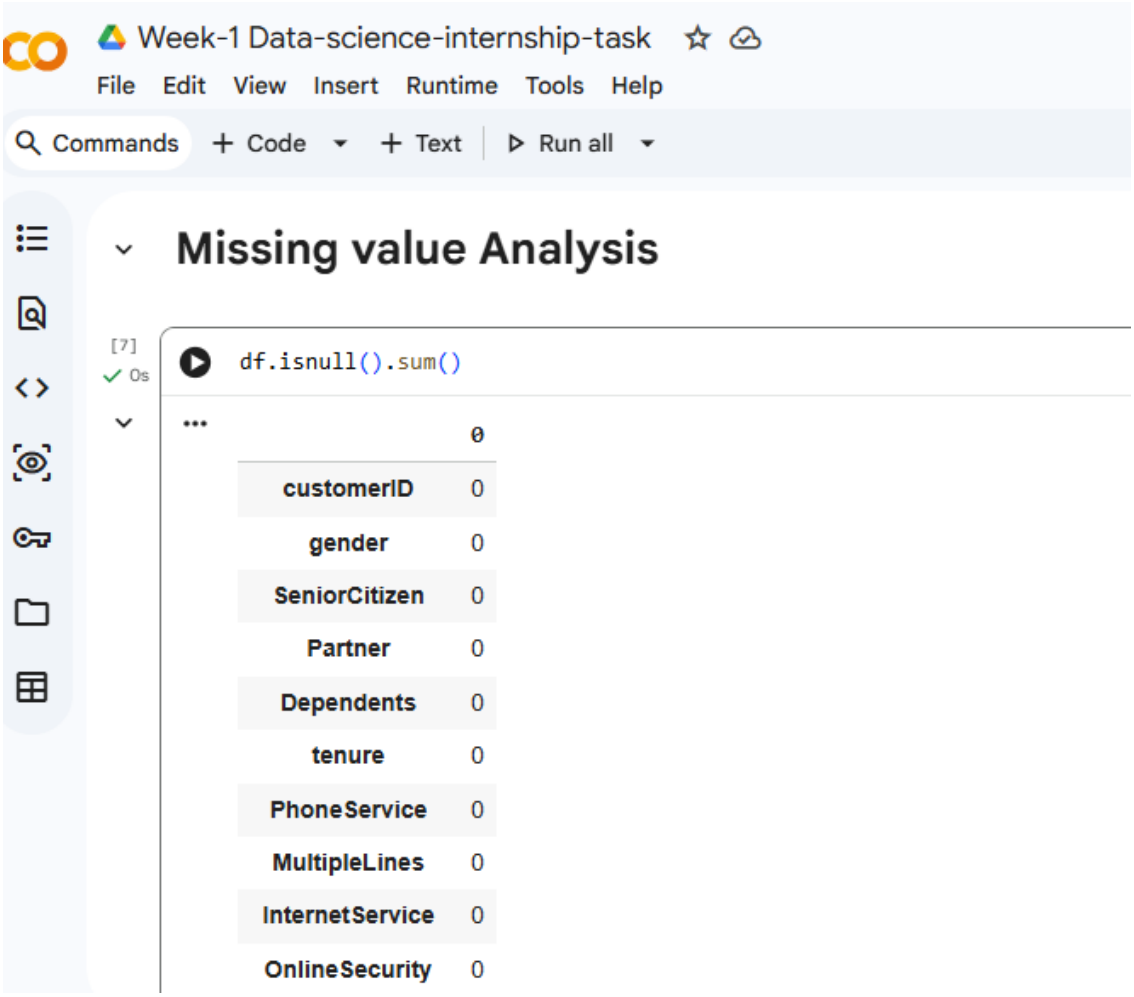
Dataset Description: The Telco Customer Churn dataset was selected for this assignment. It contains detailed information about telecom customers including demographic details, billing information, and service usage. The dataset included several missing values, inconsistent formats, and potential outliers, making it suitable for this cleaning task.



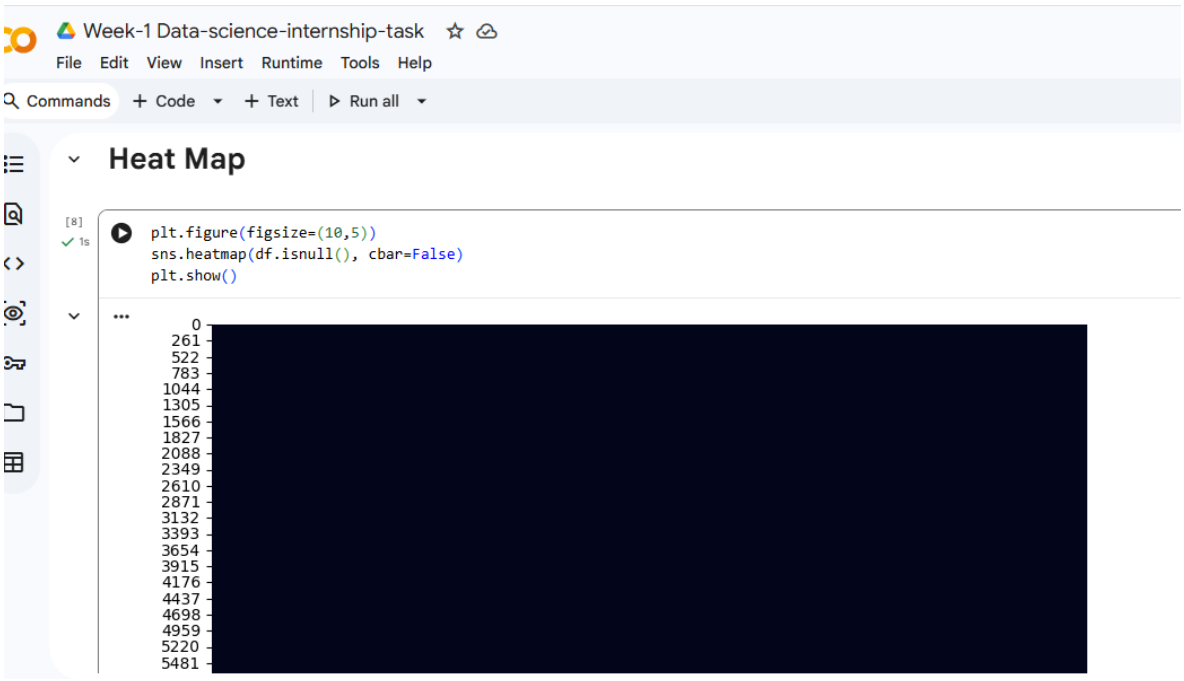


	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	Device
count	7043	7043	7043.000000	7043	7043	7043.000000	7043	7043	7043	7043	...	
unique	7043	2	NaN	2	2	NaN	2	3	3	3	...	
top	3186-AJIEK	Male	NaN	No	No	NaN	Yes	No	Fiber optic	No	...	
freq	1	3555	NaN	3641	4933	NaN	6361	3390	3096	3498	...	
mean	NaN	NaN	0.162147	NaN	NaN	32.371149	NaN	NaN	NaN	NaN	...	
std	NaN	NaN	0.368612	NaN	NaN	24.559481	NaN	NaN	NaN	NaN	...	
min	NaN	NaN	0.000000	NaN	NaN	0.000000	NaN	NaN	NaN	NaN	...	
25%	NaN	NaN	0.000000	NaN	NaN	9.000000	NaN	NaN	NaN	NaN	...	
50%	NaN	NaN	0.000000	NaN	NaN	29.000000	NaN	NaN	NaN	NaN	...	
75%	NaN	NaN	0.000000	NaN	NaN	55.000000	NaN	NaN	NaN	NaN	...	
max	NaN	NaN	1.000000	NaN	NaN	72.000000	NaN	NaN	NaN	NaN	...	

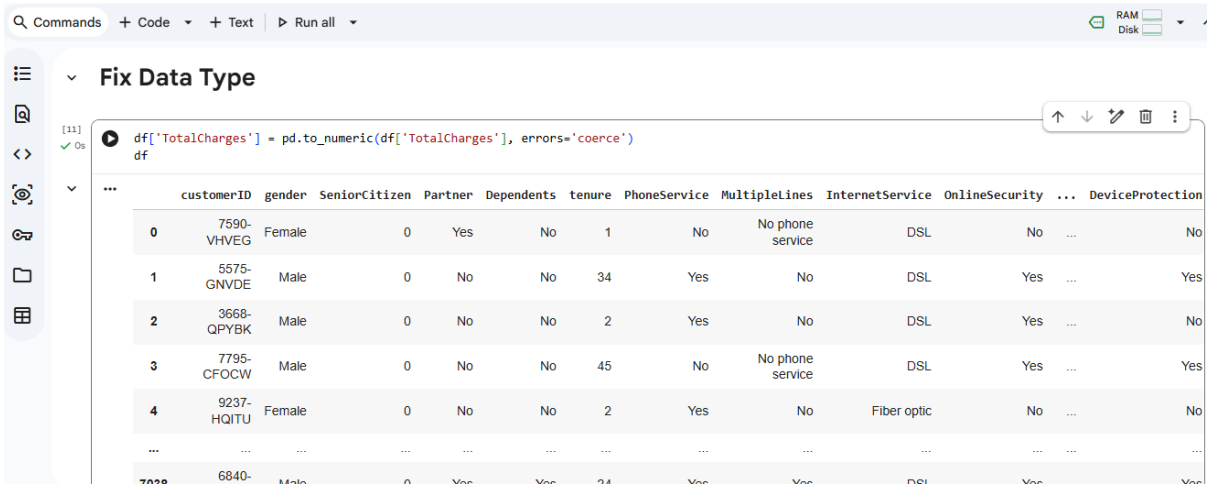
Missing Value Analysis: An in-depth missing value analysis was conducted to understand which columns required attention. Both tabular summaries and visualization techniques were used to analyze the distribution of missing data.



	0
customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0

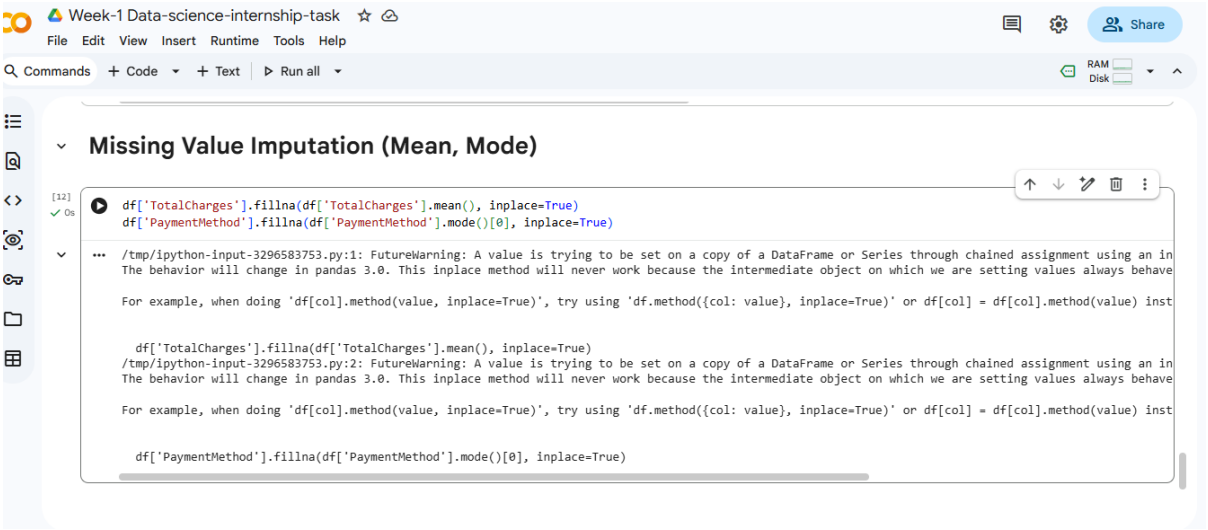


Data Type Correction: During preprocessing, the TotalCharges column was found to contain non-numeric values stored as text. This issue was corrected by converting the column into proper numeric format to ensure consistency and accuracy.

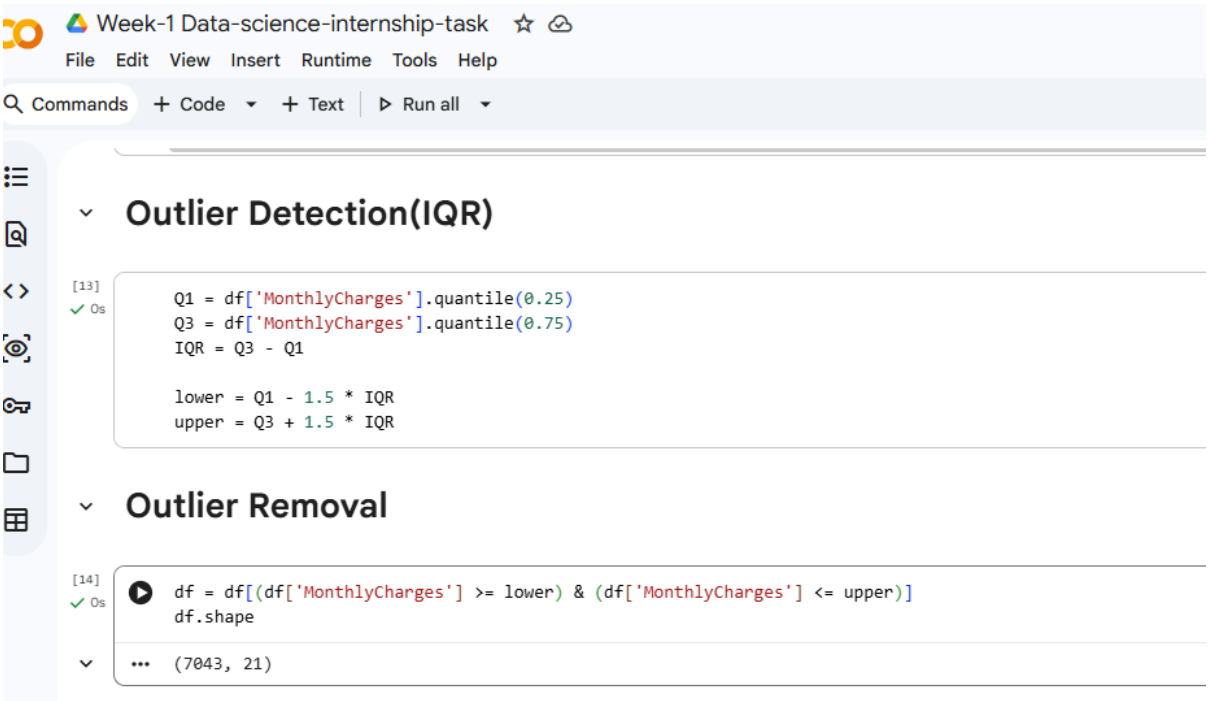


	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No
...	...	...	...	...	...	...	...	...	...	...	...	...
7038	6840-	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes		Yes

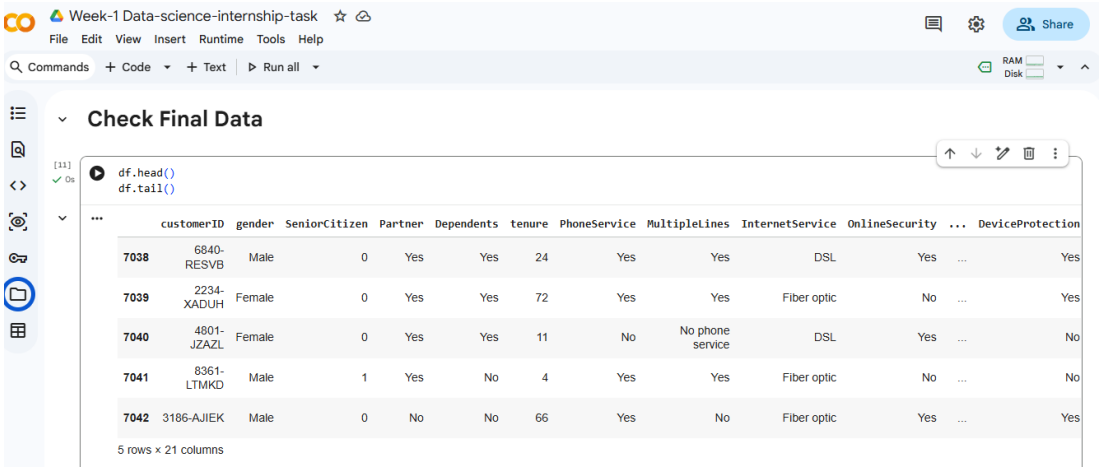
Missing Value Imputation: Two different imputation techniques were applied. First, statistical imputation using mean and mode was used for appropriate columns. Then, KNN Imputation was implemented as an advanced technique. After imputation, the dataset was verified to ensure no missing values remained.



Outlier Detection & Treatment: Outliers in the MonthlyCharges column were detected using the IQR method. Identified extreme values were treated through documented filtering to improve the reliability of the dataset.

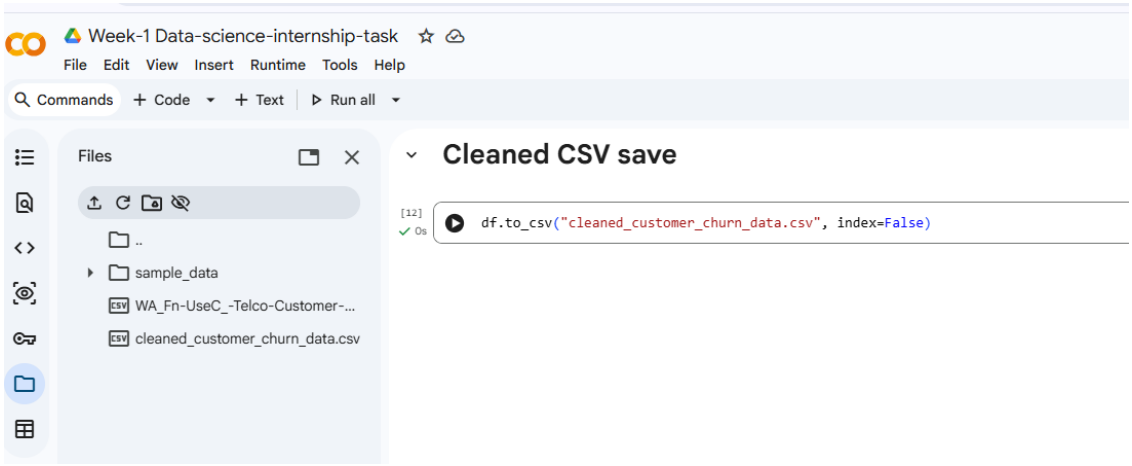


Final Dataset Export:



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes	...	Yes
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	No	...	Yes
7040	4801-JZAZL	Female	0	Yes	Yes	11	No	No phone service	DSL	Yes	...	No
7041	8361-LTMKD	Male	1	Yes	No	4	Yes	Yes	Fiber optic	No	...	No
7042	3186-AJIEK	Male	0	No	No	66	Yes	No	Fiber optic	Yes	...	Yes

5 rows × 21 columns



Conclusion: This assignment helped me understand how real-world business data often contains many quality issues. By applying Python-based preprocessing techniques, I was able to convert messy data into a clean and structured format. The final dataset is now consistent and ready for further analytics and machine learning tasks.