Assignment 4 – Correlation Analysis

Name: Aman Nadeem
Roll No: 2225165002
Course: Applied Data Science with AI
Week #: 4
Project Title: Customer Churn Prediction

1. Reading Summary

Reading Material:

- Khan Academy – Statistics & Probability
- Introductory Stats for Data Science Notes

Key Learnings:

- Mean, median, and mode describe the center of data.
- Variance shows how spread out the data is.
- Correlation explains how two variables are related, either positively or negatively.
- Understanding these measures helps to find which features matter most for predictions.

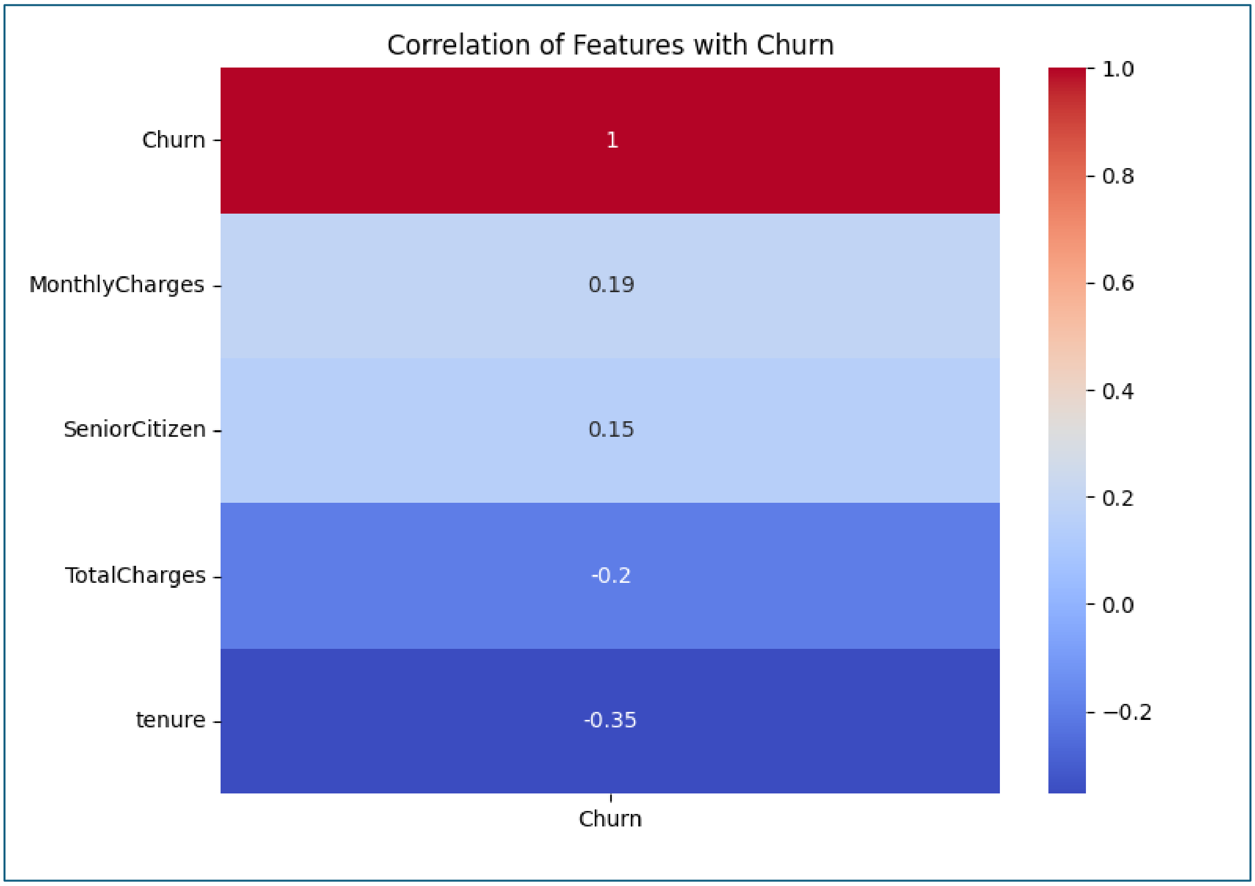
Reflection:

This week's readings helped me understand how statistical measures guide us in identifying key factors that affect customer churn.

2. Classroom Task Documentation

Task Performed:

- Practiced calculating mean, median, mode, variance, and correlation in datasets.
- Learned how to build correlation heatmaps using Seaborn.



3. Weekly Assignment Submission

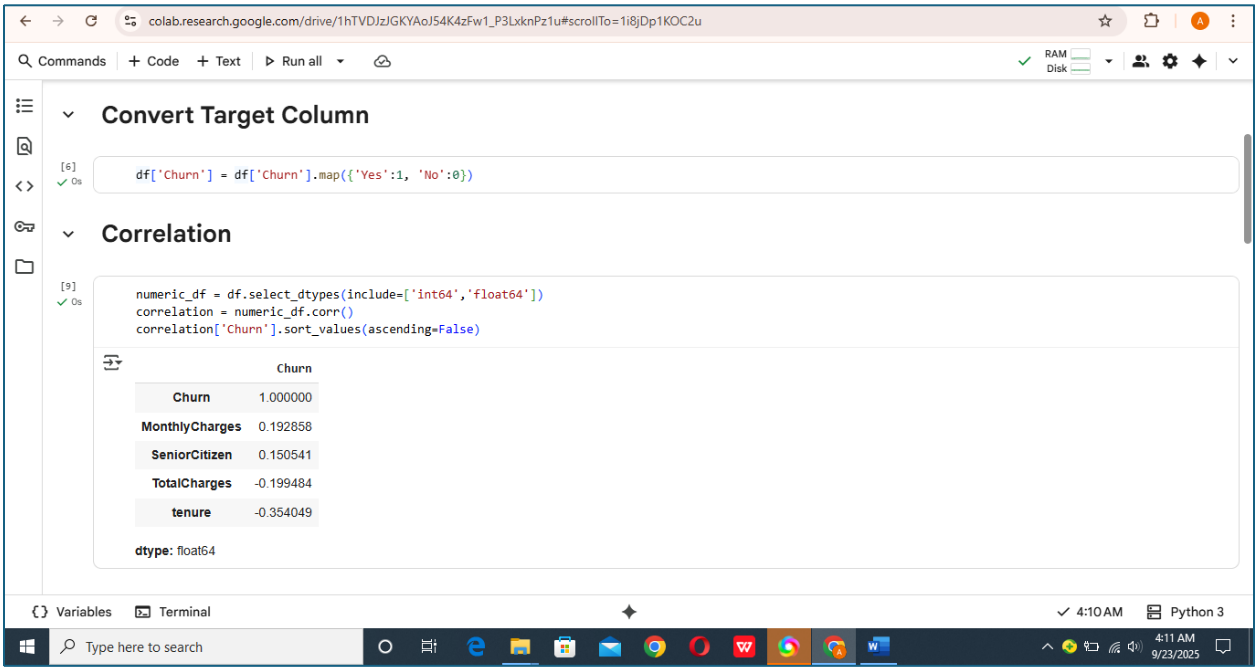
Assignment Title: Correlation Analysis

Steps Taken:

1. Loaded cleaned Churn dataset.

2. Converted Churn column to numeric (Yes = 1, No = 0).

3. Selected numeric features and computed correlation matrix.



Convert Target Column

```
df['Churn'] = df['Churn'].map({'Yes':1, 'No':0})
```

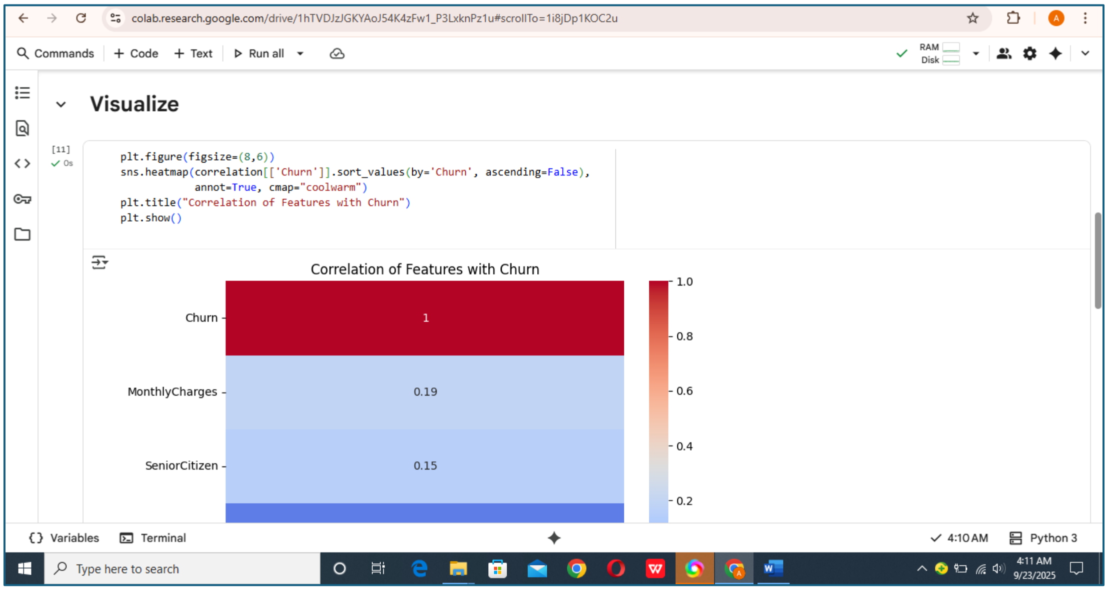
Correlation

```
numeric_df = df.select_dtypes(include=['int64','float64'])
correlation = numeric_df.corr()
correlation['Churn'].sort_values(ascending=False)
```

```
                   Churn
Churn           1.000000
MonthlyCharges  0.192858
SeniorCitizen   0.150541
TotalCharges   -0.199484
tenure         -0.354049

dtype: float64
```

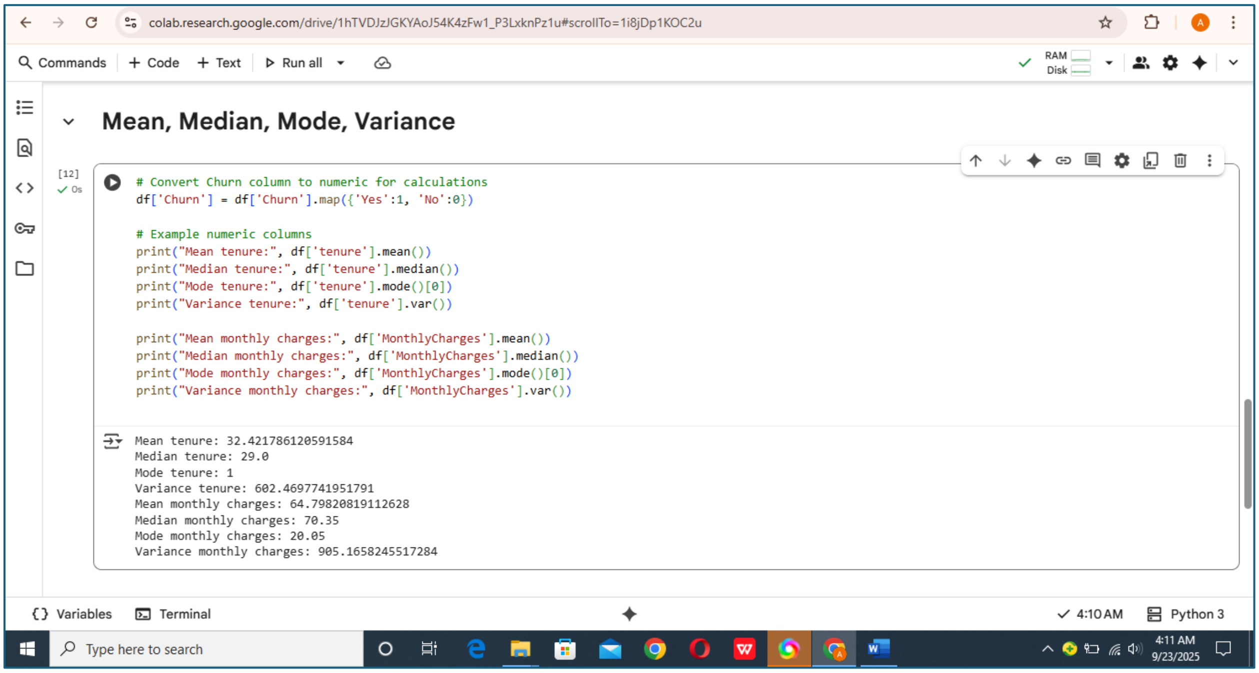
4. Created heatmap to visualize correlations.



Visualize

```
plt.figure(figsize=(8,6))
sns.heatmap(correlation[['Churn']].sort_values(by='Churn', ascending=False),
            annot=True, cmap="coolwarm")
plt.title("Correlation of Features with Churn")
plt.show()
```

5. Calculated mean, median, mode, and variance for numeric columns.



Mean, Median, Mode, Variance

[12]

```python
# Convert Churn column to numeric for calculations
df['Churn'] = df['Churn'].map({'Yes':1, 'No':0})

# Example numeric columns
print("Mean tenure:", df['tenure'].mean())
print("Median tenure:", df['tenure'].median())
print("Mode tenure:", df['tenure'].mode()[0])
print("Variance tenure:", df['tenure'].var())

print("Mean monthly charges:", df['MonthlyCharges'].mean())
print("Median monthly charges:", df['MonthlyCharges'].median())
print("Mode monthly charges:", df['MonthlyCharges'].mode()[0])
print("Variance monthly charges:", df['MonthlyCharges'].var())
```

```
Mean tenure: 32.421786120591584
Median tenure: 29.0
Mode tenure: 1
Variance tenure: 602.4697741951791
Mean monthly charges: 64.79820819112628
Median monthly charges: 70.35
Mode monthly charges: 20.05
Variance monthly charges: 905.1658245517284
```

Statistical Summary:

- Tenure → Mean ≈ 32, Median ≈ 29, Mode = 1, Variance ≈ 600
- MonthlyCharges → Mean ≈ 64, Median ≈ 70, Mode = 20, Variance ≈ 900

Output & Key Findings:

- **Tenure (-0.35):** Strong negative correlation with churn. Short-tenure customers are more likely to leave.
- **MonthlyCharges (+0.19):** Higher monthly bills increase the chance of churn.
- **TotalCharges (-0.20):** Customers who paid more overall are less likely to churn.

- **SeniorCitizen (+0.15):** Slightly higher churn tendency in senior citizens.

Challenges Faced:

At first, I tried correlation on all columns, but it failed due to categorical data. Fixed it by selecting only numeric features.

GitHub Link:

https://github.com/amannadeem126/Customer-Churn-Prediction

4. Project Progress Milestone

- Identified three most important features for churn: **Tenure, MonthlyCharges, TotalCharges**.
- Next week's goal: Build a baseline regression model.

5. Self-Evaluation

I completed all tasks on time.