**DATA SCIENCE MINOR PROJECT REPORT**

**Project Title: Glioma Grading Clinical and Mutation Features**

**Submitted by: Sangam Srivastav**

**Submitted to: Mr. Manpreet Singh Sehgal**

**Registration No: 12323431**

**Section & Roll No. : K23SG , 18**

**Course Code: INT 375**

**Under the Guidance of: Mr. Manpreet Singh Sehgal**

**Discipline of CSE/IT**

**Lovely Professional University, Phagwara**

**DECLARATION**

I am, **Sangam Srivastav**, a student of CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

**Date:** 12-04-2025  
**Signature:** *Sangam Srivastav*

**CERTIFICATE**

This is to certify that **Sangam Srivastav** has completed the project titled, “**Glioma Grading Clinical and Mutation Features**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort, and study.

**Signature and Name of the Supervisor:**  
**Mr. Manpreet Singh Sehgal**  
**School of Engineering**  
Lovely Professional University  
Phagwara, Punjab  
**Date:** 12-04-2025

**ACKNOWLEDGEMENT**

I sincerely thank my mentor for their valuable support and guidance throughout the course of this project. I also acknowledge the university and faculty for providing access to resources and a conducive learning environment. Finally, I extend my gratitude to my peers and family for their encouragement.

**TABLE OF CONTENTS**

1. Introduction
2. Source of Dataset
3. EDA Process
4. Analysis on Dataset  
   i. Introduction  
   ii. General Description  
   iii. Specific Requirements, Functions and Formulas  
   iv. Analysis Results  
   v. Visualization
5. Conclusion
6. Future Scope
7. References

**1. Introduction**

This project analyzes clinical and genetic mutation data from glioma patients (GBM and LGG) using data science techniques. The main aim is to study how different features such as age, gender, race, and mutation profiles correlate with glioma grading. The findings may help in better understanding the disease progression and treatment planning.

**2. Source of Dataset**

The dataset is obtained from the UCI Machine Learning Repository:  
[**Glioma Grading Clinical and Mutation Features Dataset**](https://archive.ics.uci.edu/dataset/759/glioma+grading+clinical+and+mutation+features+dataset)

**3. EDA Process**

* Missing values were identified and handled appropriately.
* Columns like age at diagnosis were converted from days to years for clarity.
* Data types were standardized for processing.
* Categorical variables such as gender, race, and grade were labeled for clarity in visualization.
* Mutation frequency was assessed using the 20 most commonly affected genes.

**4. Analysis on Dataset**

**Objective 1: Grade Distribution(Count of GBM vs LGG)**

**General Description:** Compare LGG vs GBM count

**Functions/Tools Used:** Pandas value\_counts(), barplot

**Result:** LGG cases outnumber GBM cases

**Visualization**: Bar chart showing grade counts

grade\_counts = df['Grade'].value\_counts()

plt.figure(figsize=(6, 4))

grade\_counts.plot(kind='bar', color=['skyblue', 'salmon'])

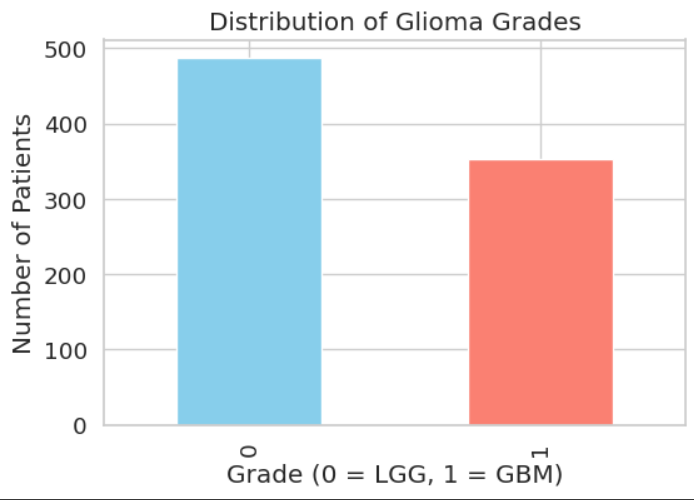
plt.title('Distribution of Glioma Grades')

plt.xlabel('Grade (0 = LGG, 1 = GBM)')

plt.ylabel('Number of Patients')

plt.grid(True)

plt.show()

****

**Objective 2: Gender Distribution**

**General Description:** Analyze gender-wise representation  
**Functions/Tools Used:** value\_counts(), seaborn countplot  
**Result:** Slightly more males than females  
**Visualization:** Bar chart by gender

gender\_counts = df['Gender'].value\_counts()

plt.figure(figsize=(6, 4))

gender\_counts.plot(kind='bar', color=['lightgreen', 'lightcoral'])

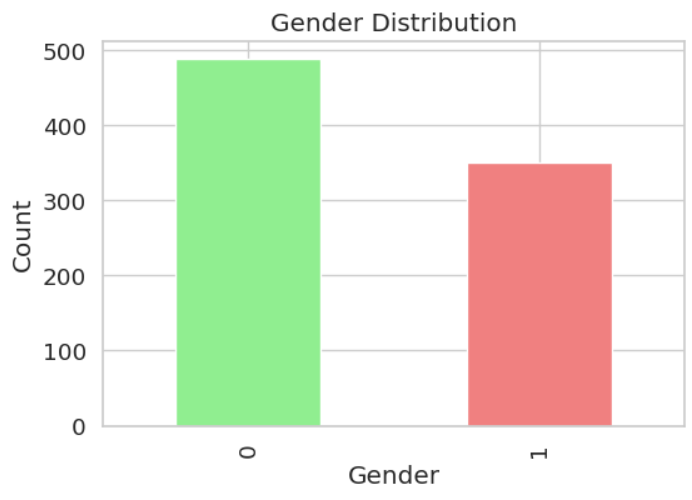
plt.title('Gender Distribution')

plt.xlabel('Gender')  # Avoid hardcoding labels until we know encoding

plt.ylabel('Count')

plt.grid(True)

plt.show()



**Objective 3: Race Distribution**

**General Description:** Show how race varies among patients  
**Functions/Tools Used:** value\_counts(), countplot  
**Result:** Most patients were White; other races less frequent  
**Visualization:** Bar chart by race

race\_counts = df['Race'].value\_counts()

plt.figure(figsize=(8, 5))

race\_counts.plot(kind='bar', color='purple')

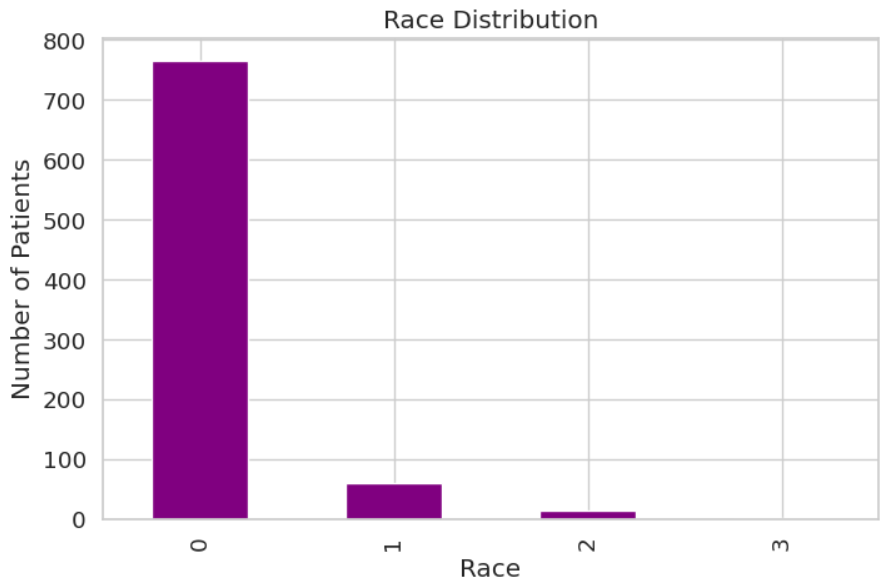
plt.title('Race Distribution')

plt.xlabel('Race')

plt.ylabel('Number of Patients')

plt.grid(True)

plt.show()



**Objective 4: IDH1 Mutation by Grade**

**General Description:** Compare IDH1 mutation presence across grades  
**Functions/Tools Used:** crosstab, stacked barplot  
**Result:** IDH1 mutations are more common in LGG  
**Visualization:** Stacked bar chart

idh1\_dist = df.groupby(['Grade', 'IDH1']).size().unstack(fill\_value=0)

plt.figure(figsize=(6, 4))

idh1\_dist.plot(kind='bar', stacked=True, color=['grey', 'blue'])

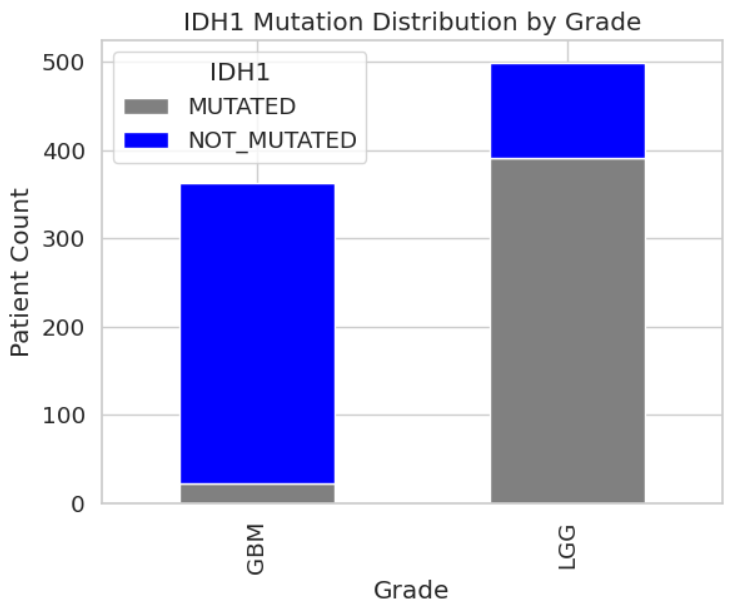
plt.title('IDH1 Mutation Distribution by Grade')

plt.xlabel('Grade')

plt.ylabel('Patient Count')

plt.grid(True)

plt.show()



**Objective 5: TP53 Mutation Status**

**General Description:** Pie chart of mutation status  
**Functions/Tools Used:** value\_counts(), plt.pie  
**Result:** Balanced proportion of mutated and non-mutated  
**Visualization:** Pie chart

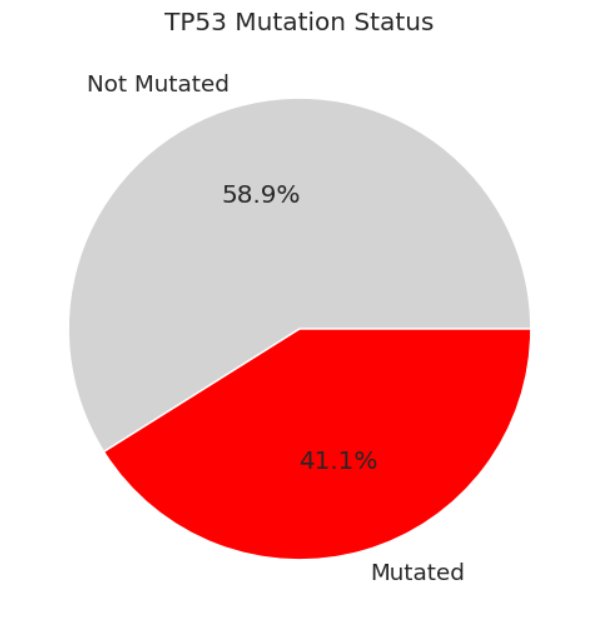
tp53\_counts = df['TP53'].value\_counts()

plt.figure(figsize=(6, 6))

plt.pie(tp53\_counts, labels=['Not Mutated', 'Mutated'], autopct='%1.1f%%', colors=['lightgray', 'red'])

plt.title('TP53 Mutation Status')

plt.show()



**Objective 6: Race-wise Grade Distribution**

**General Description:** Cross-tabulation by race and grade  
**Functions/Tools Used:** pd.crosstab(), stacked bar chart  
**Result:** White patients dominate both grades  
**Visualization:** Bar chart by race and grade

race\_grade = pd.crosstab(df['Race'], df['Grade'])

plt.figure(figsize=(8, 6))

race\_grade.plot(kind='bar', stacked=True, colormap='viridis')

plt.title('Grade Distribution across Races')

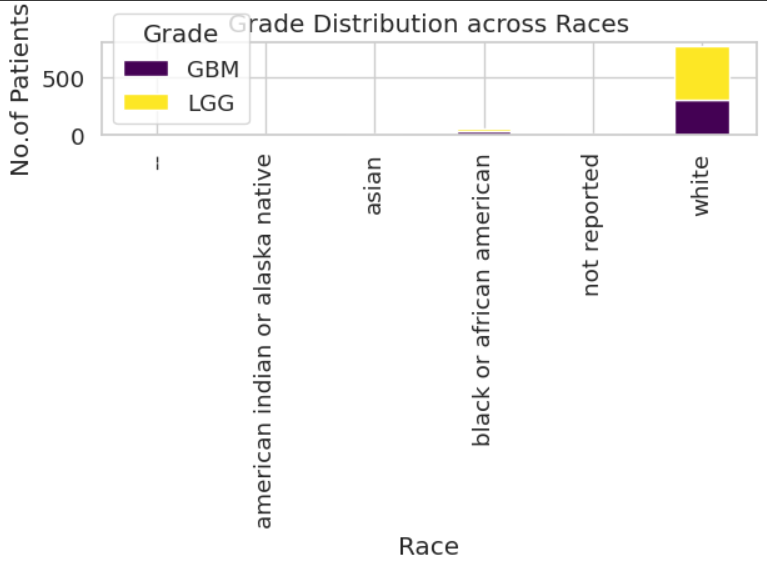
plt.xlabel('Race')

plt.ylabel('No.of Patients')

plt.grid(True)

plt.tight\_layout()

plt.show()



**Objective 7: Boxplot of Age by Grade**

**General Description:** Show age distribution by grade  
**Functions/Tools Used:** boxplot  
**Result:** GBM age range is slightly older  
**Visualization:** Boxplot

plt.figure(figsize=(8, 6))

sns.boxplot(x='Grade', y='Age\_at\_diagnosis', data=df, palette=['#1f77b4', '#ff7f0e'])

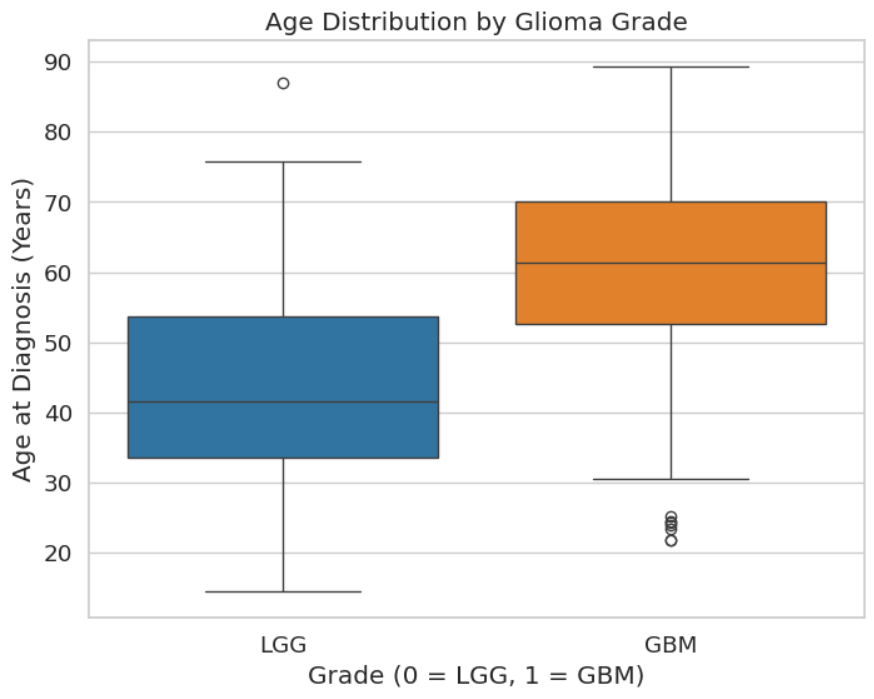
plt.title('Age Distribution by Glioma Grade')

plt.xlabel('Grade (0 = LGG, 1 = GBM)')

plt.ylabel('Age at Diagnosis (Years)')

plt.xticks([0, 1], ['LGG', 'GBM'])

plt.show()



**Objective 8: Violin Plot – Mutation Count per Patient**

**General Description:** Compare mutation burden by grade  
**Functions/Tools Used:** sum(axis=1), violinplot  
**Result:** Mutation burden differs between LGG and GBM  
**Visualization:** Violin plot

df['Mutation\_Count'] = df[mutation\_cols].sum(axis=1)

plt.figure(figsize=(8, 6))

sns.violinplot(x='Grade', y='Mutation\_Count', data=df, palette=['#1f77b4', '#ff7f0e'])

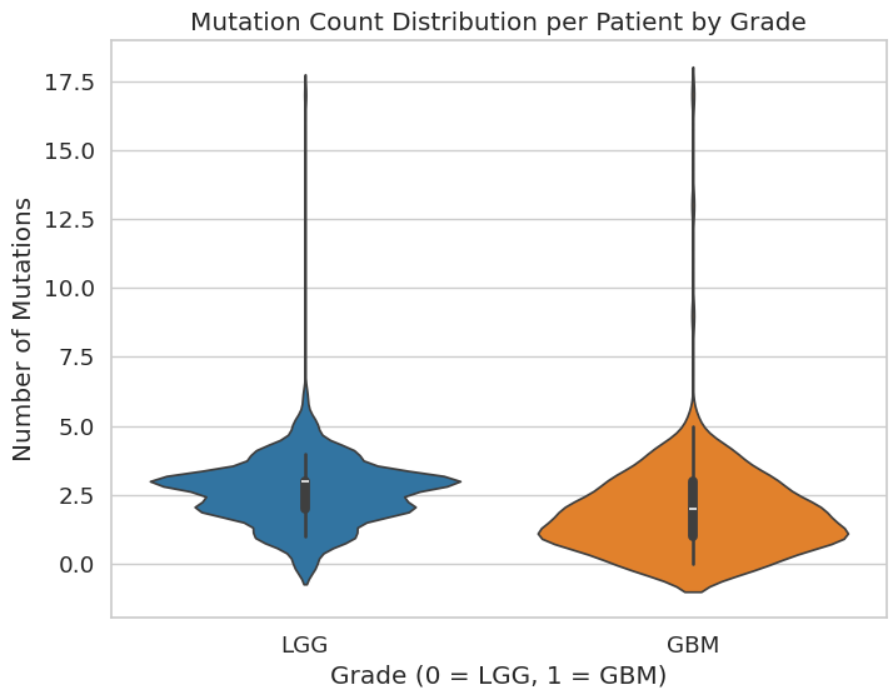
plt.title('Mutation Count Distribution per Patient by Grade')

plt.xlabel('Grade (0 = LGG, 1 = GBM)')

plt.ylabel('Number of Mutations')

plt.xticks([0, 1], ['LGG', 'GBM'])

plt.show()



**Objective 9: Feature Importance (Random Forest)**

**General Description:** Feature contribution to grade prediction  
**Functions/Tools Used:** RandomForestClassifier, feature\_importances\_  
**Result:** Age and IDH1 were most important  
**Visualization:** Bar chart of feature importance

X = df[['Age\_at\_diagnosis', 'Gender', 'Race'] + mutation\_cols]

y = df['Grade']

clf = RandomForestClassifier(n\_estimators=100, random\_state=42)

clf.fit(X, y)

feature\_importance = pd.Series(clf.feature\_importances\_, index=X.columns).sort\_values(ascending=False)

plt.figure(figsize=(12, 6))

sns.barplot(x=feature\_importance.index, y=feature\_importance.values, palette='magma')

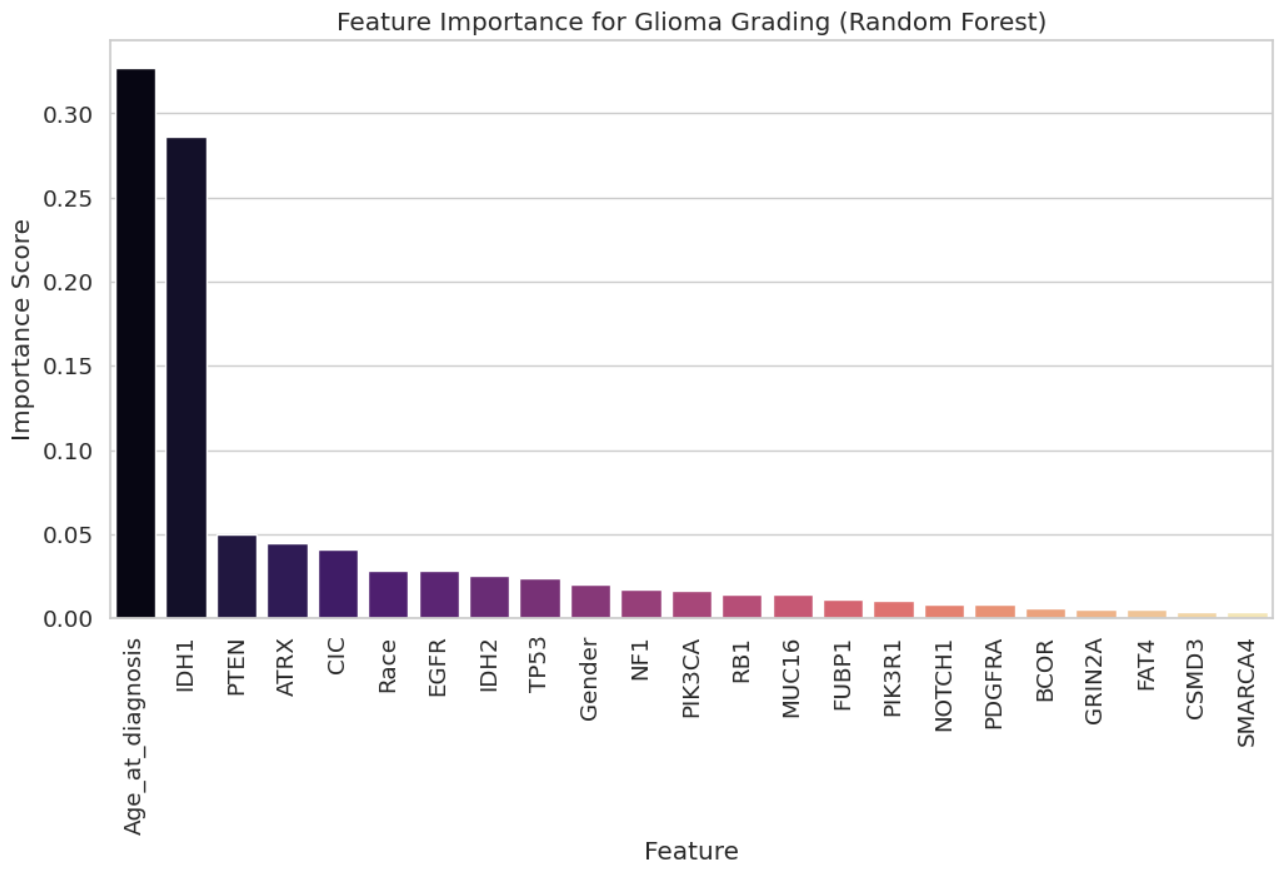
plt.title('Feature Importance for Glioma Grading (Random Forest)')

plt.xlabel('Feature')

plt.ylabel('Importance Score')

plt.xticks(rotation=90)

plt.show()



**Objective 10: Gene Mutation Comparison between GBM & LGG**

**Result: Distinct mutation profiles observed between the two grades**

**Visualization: Line chart**

gbm = df[df['Grade'] == 1][mutation\_cols].mean()

lgg = df[df['Grade'] == 0][mutation\_cols].mean()

plt.figure(figsize=(14, 6))

plt.plot(gbm.index, gbm.values, label='GBM', color='#ff7f0e', marker='o')

plt.plot(lgg.index, lgg.values, label='LGG', color='#1f77b4', marker='s')

plt.title('Gene Mutation Frequency: GBM vs LGG')

plt.xlabel('Gene')

plt.ylabel('Mutation Frequency (Proportion)')

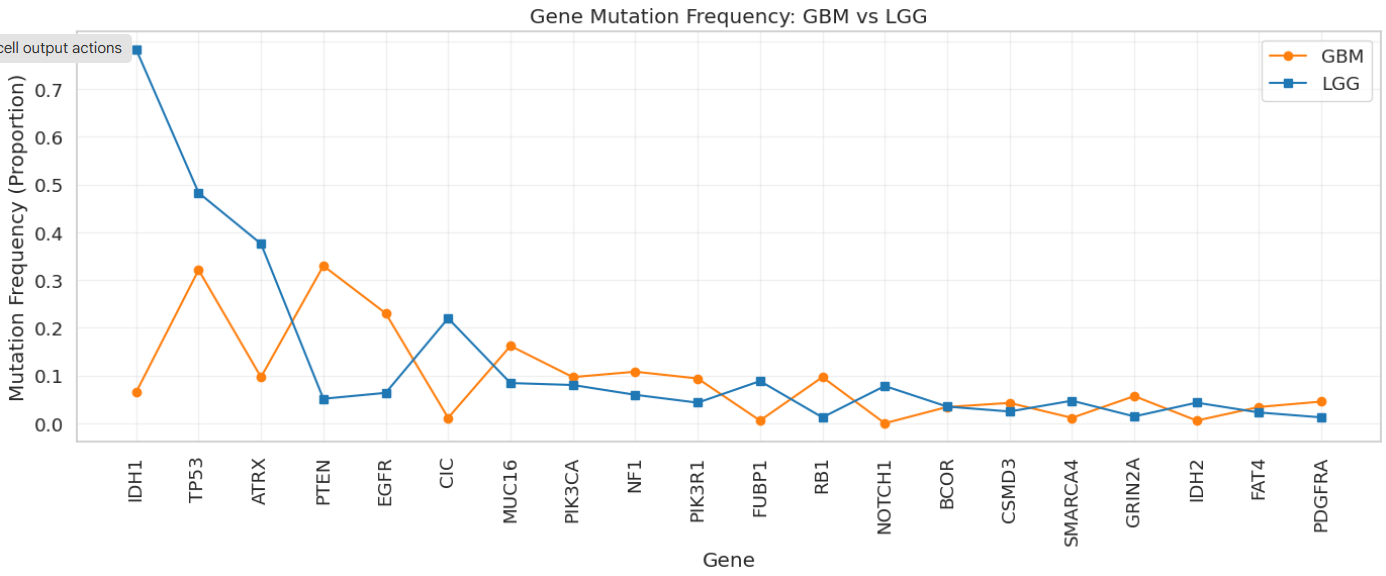
plt.xticks(rotation=90)

plt.legend()

plt.grid(True, alpha=0.3)

plt.tight\_layout()

plt.show()



**5. Conclusion**

This project explored the relationship between clinical features and gene mutations in glioma patients. The analysis showed meaningful patterns between age, gender, mutation burden, and glioma grade. Visualizations effectively communicated trends and comparisons across groups.

**6. Future Scope**

* Incorporate survival prediction models using machine learning
* Extend to multi-class glioma grading
* Include treatment and outcome data for prognosis analysis
* Develop a web dashboard for real-time clinical decision support

**7. References**

1. UCI Machine Learning Repository – Glioma Dataset
2. Scikit-learn, Pandas, Seaborn, Matplotlib libraries
3. Lovely Professional University – Course Resources
4. Research articles on Glioma and Tumor Mutations