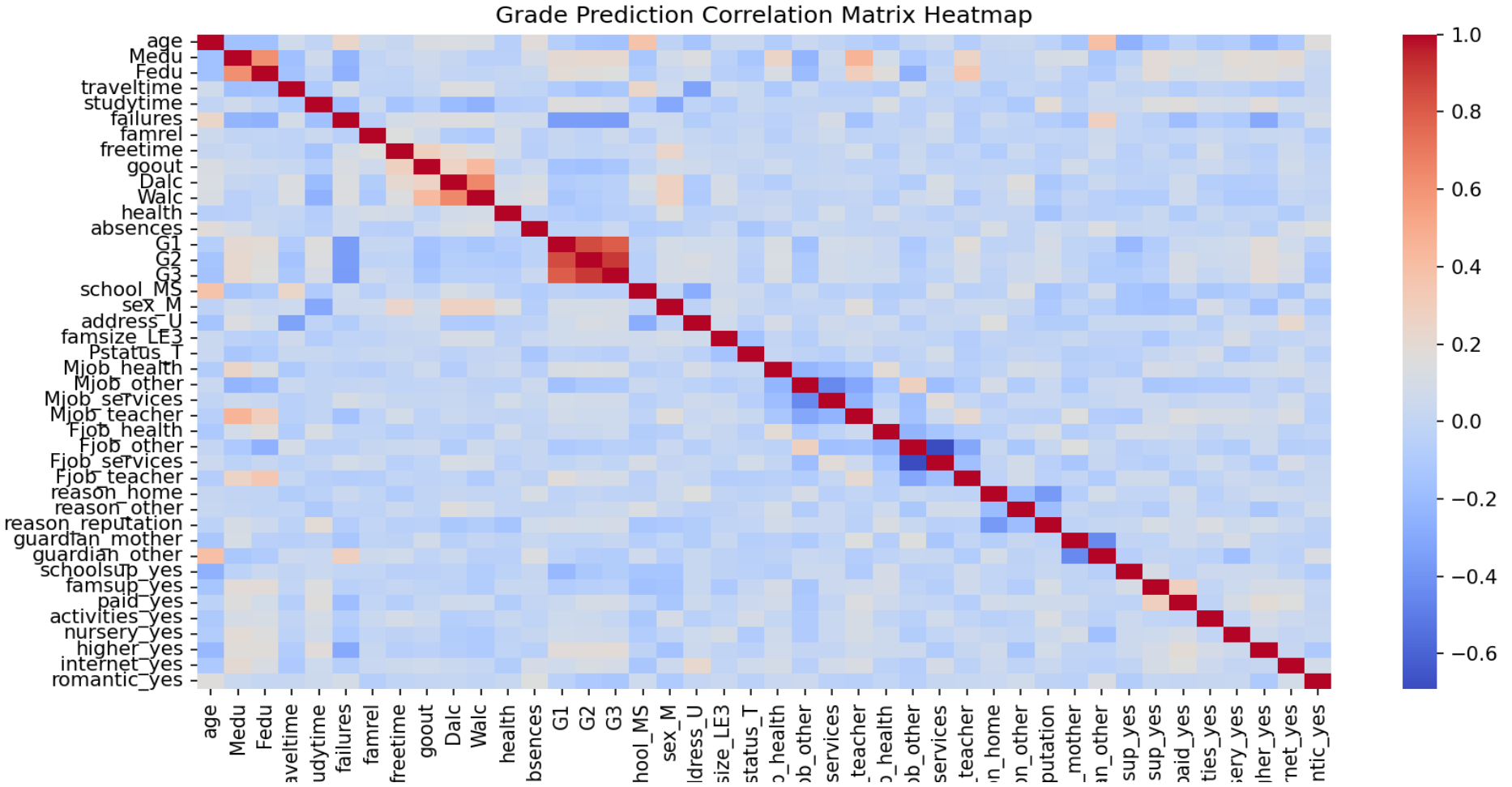
STUDENT GRADE PREDICTION USING LINEAR REGRESSION

SYED MUHAMMAD HAYYAN HASAN

LIBRARIES USED & DATA SET

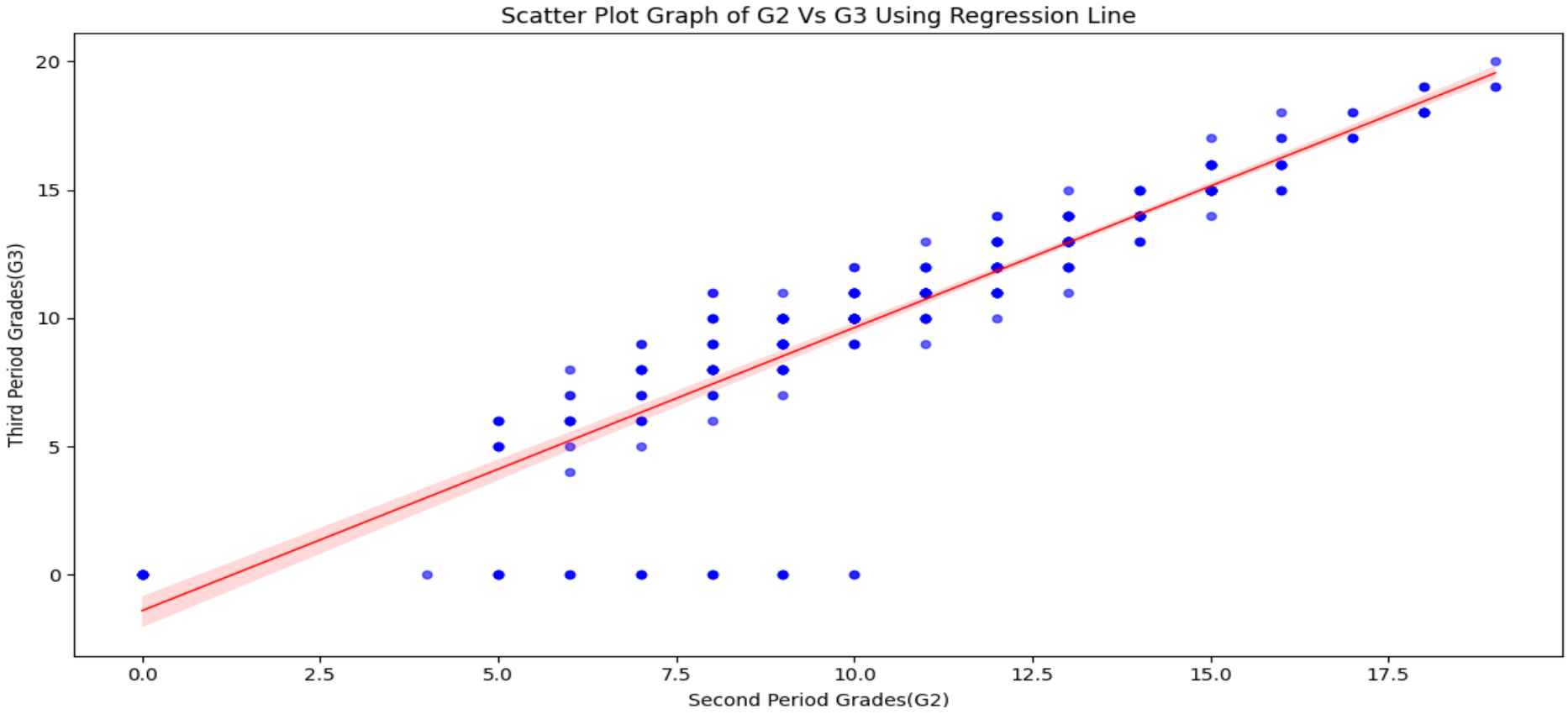
- **Libraries:** Pandas, matplotlib.pyplot, seaborn, sklearn.model_selection, sklearn.linear_model, sklearn.metrics
- **Dataset Name:** Student Performance Data Set
- **Provided by:** UCI Machine Learning Repository
- **Main Goal:** Predicting Final Grades(G3) using Linear Regression and Machine Learning
- **Important features used:** First Period Grades(G1), (Second Period Grades)G2, failures, absences

CORRELATION HEATMAP



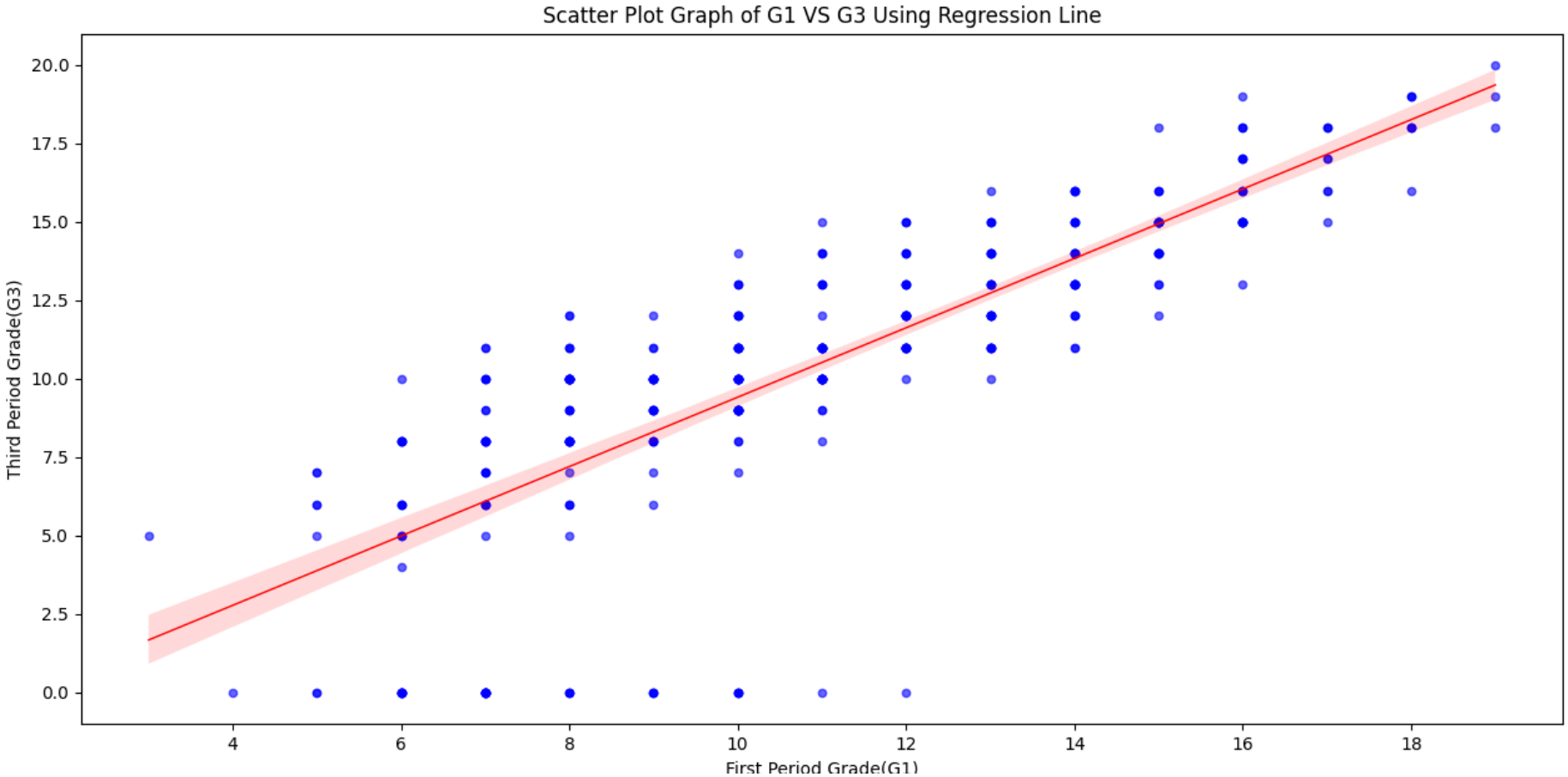
- There seems to be a correlation between several features and the final grades (G3).

VISUAL ANALYSIS (G2 VS G3)



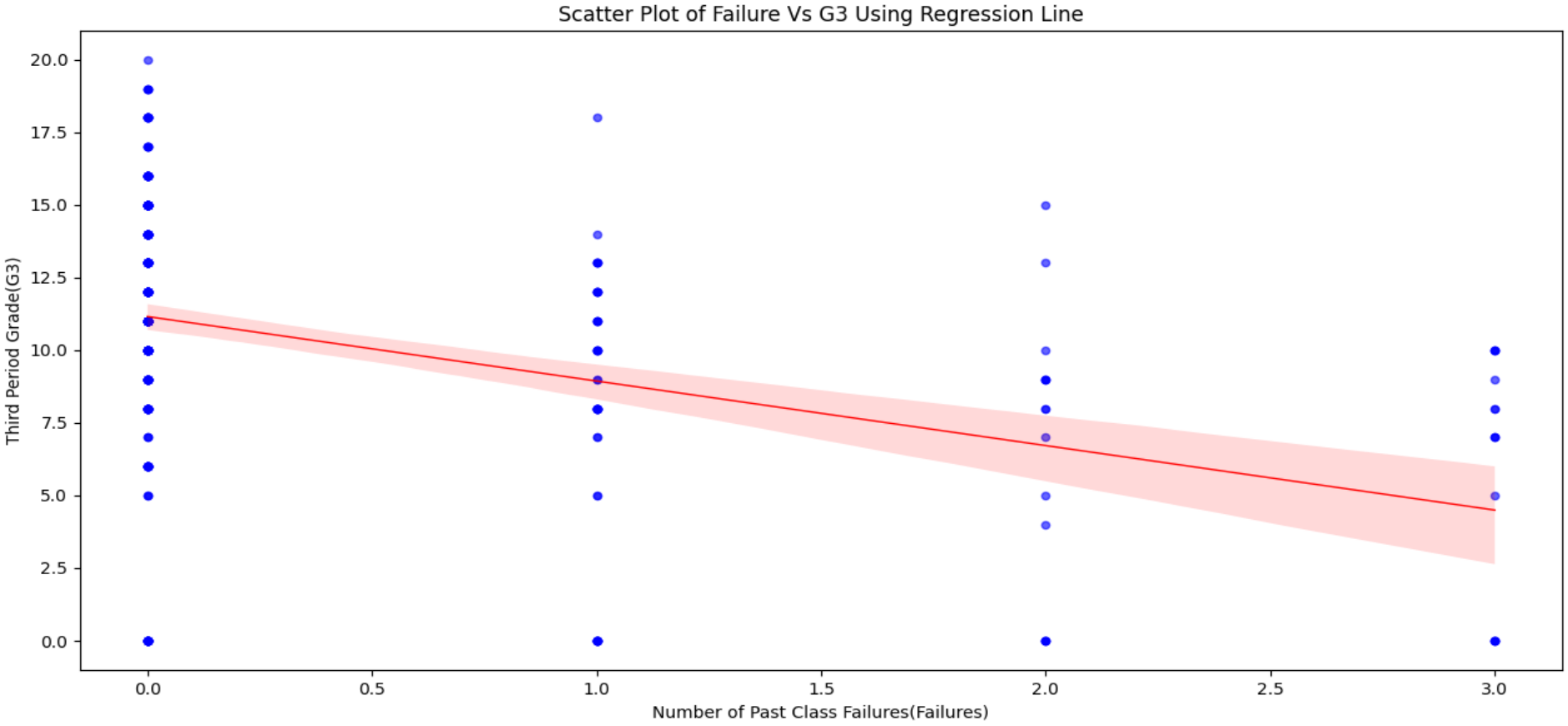
- The scatter plot shows a strong positive correlation between second period grades (G2) and final grades (G3).

VISUAL ANALYSIS (G1 VS G3)



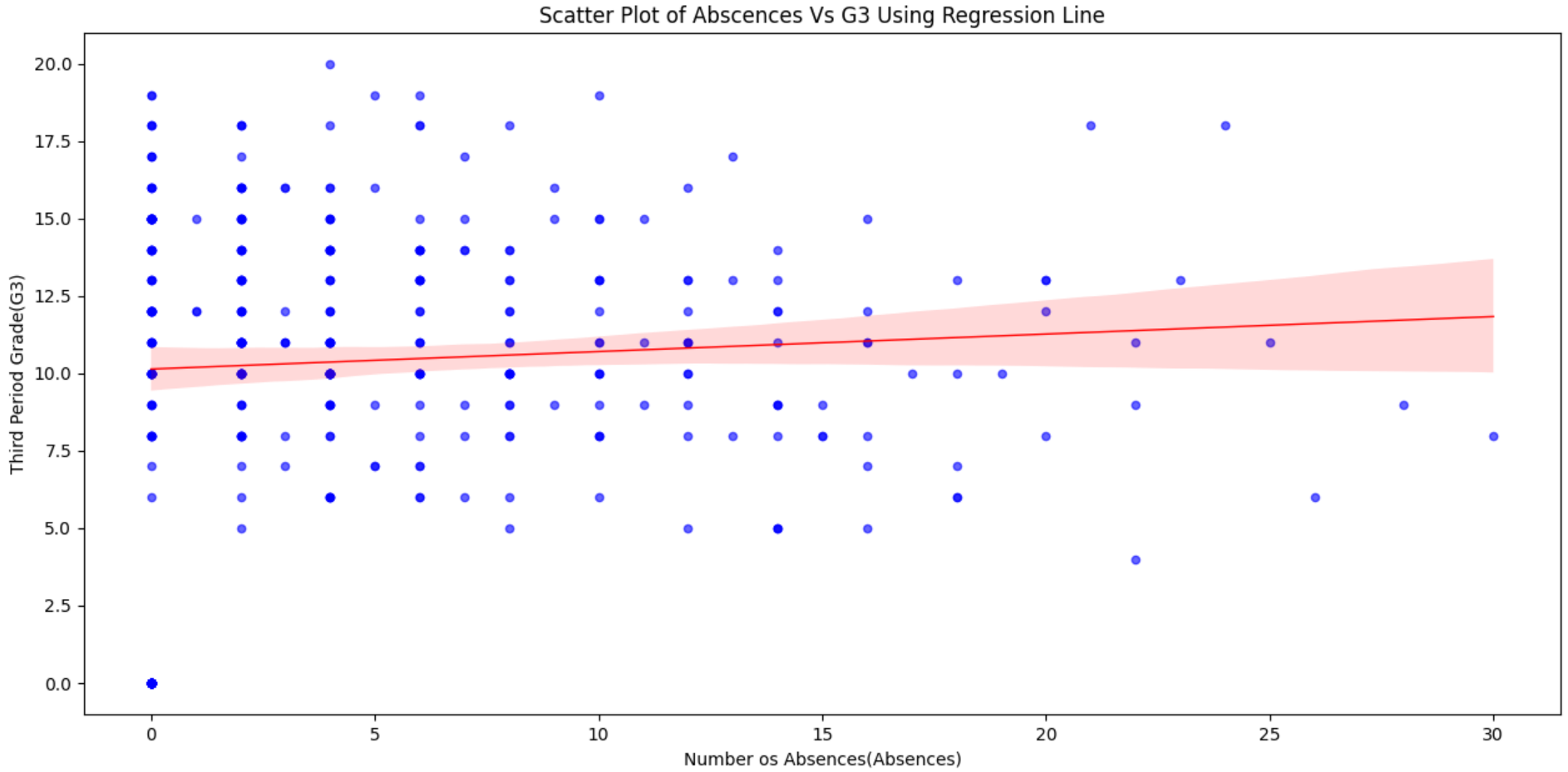
- The scatter plot indicates a moderate positive correlation between first period grades (G1) and final grades (G3).

VISUAL ANALYSIS (FAILURES VS G3)



- Failures vs G3: Slight negative correlation
- The Data was cleaner than anticipated, hence it seems inorganic

VISUAL ANALYSIS (SCATTER PLOTS)

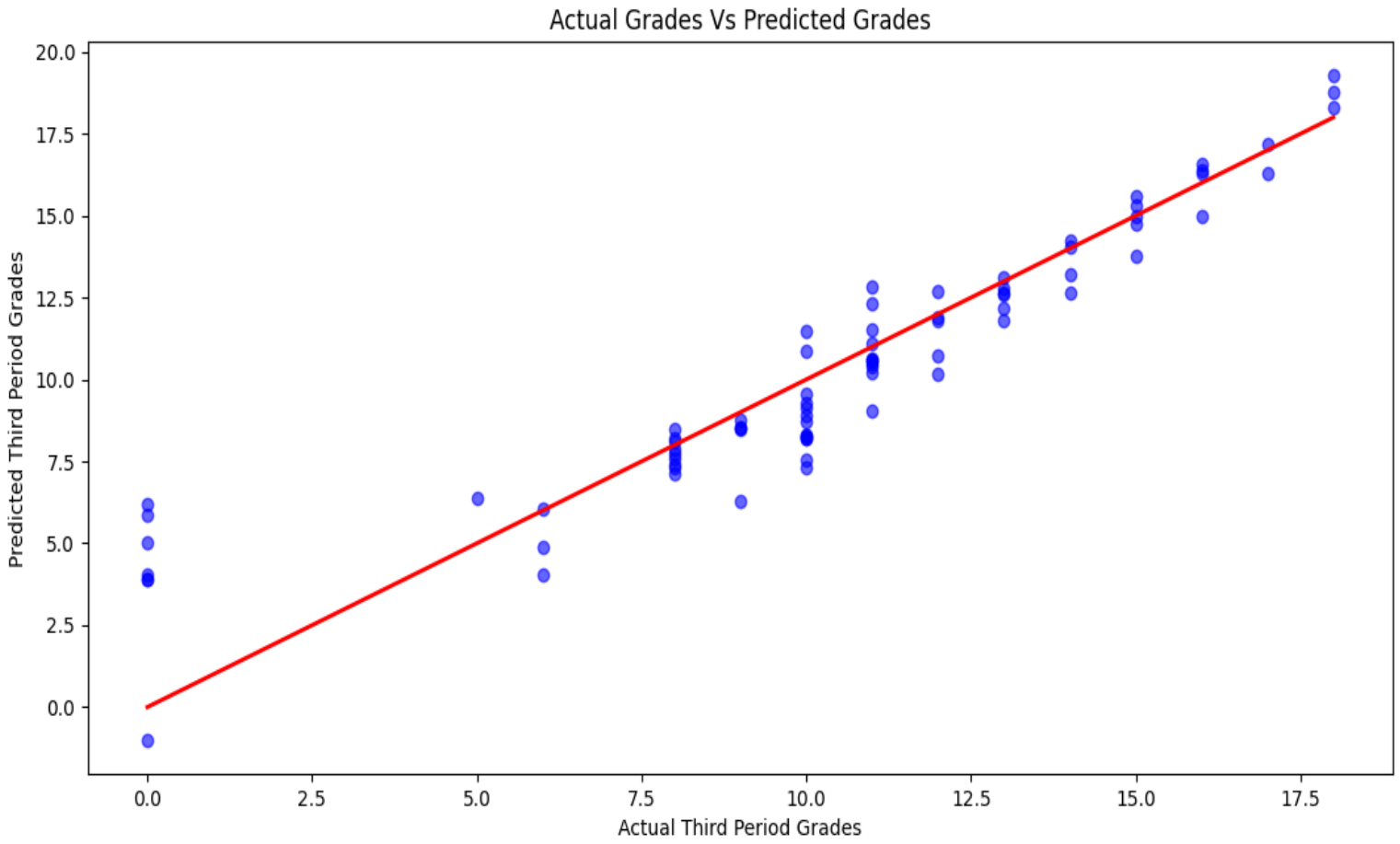


- Had to limit the analysis to 30 as it was too inconclusive after it

MODEL BUILDING & RESULTS

- Used Linear Regression model
- Train-test split: 80/20 with random_state = 2
- Evaluation metrics:
- - Mean Squared Error (MSE): 2.87
- - R^2 Score: 0.849

ACTUAL VS PREDICTED G3



PROJECT PROBLEMS & FIXES

- String columns caused correlation issue → Fixed using One-Hot Encoding
- Heatmap used for overwhelming correlation data
- Failure vs G3 looked artificial → due to clean data
- Absences vs G3 too scattered → limited plot to 30 absences

CONCLUSION

- Successfully built a Linear Regression model to predict final grades (G3) using G1, G2, absences, and failures.
- Identified G1 and G2 as the most impactful predictors through correlation analysis and visualizations.
- Some features (absences, failures) had weaker impact
- Achieved a strong R^2 score of 0.85, indicating high model accuracy and reliability.
- Gained hands-on experience in data preprocessing, exploratory analysis, model training, and performance evaluation using Python and relevant libraries.