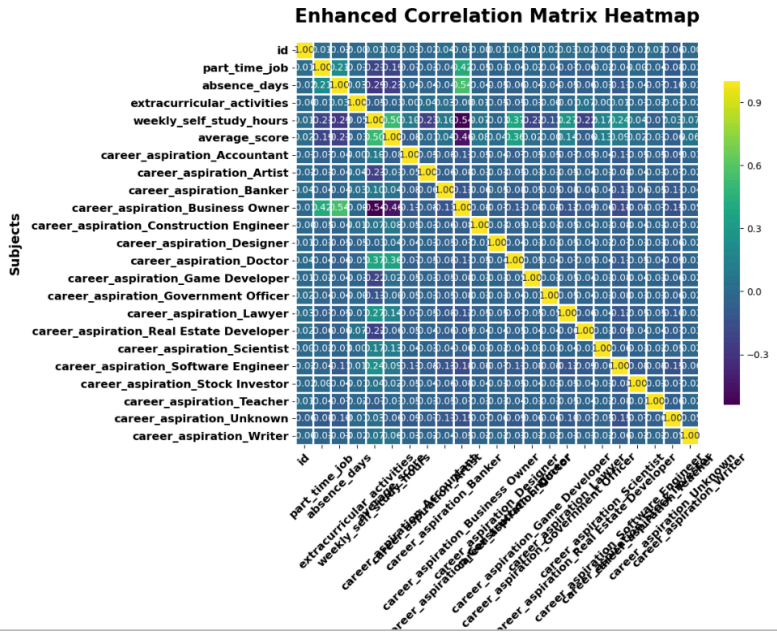
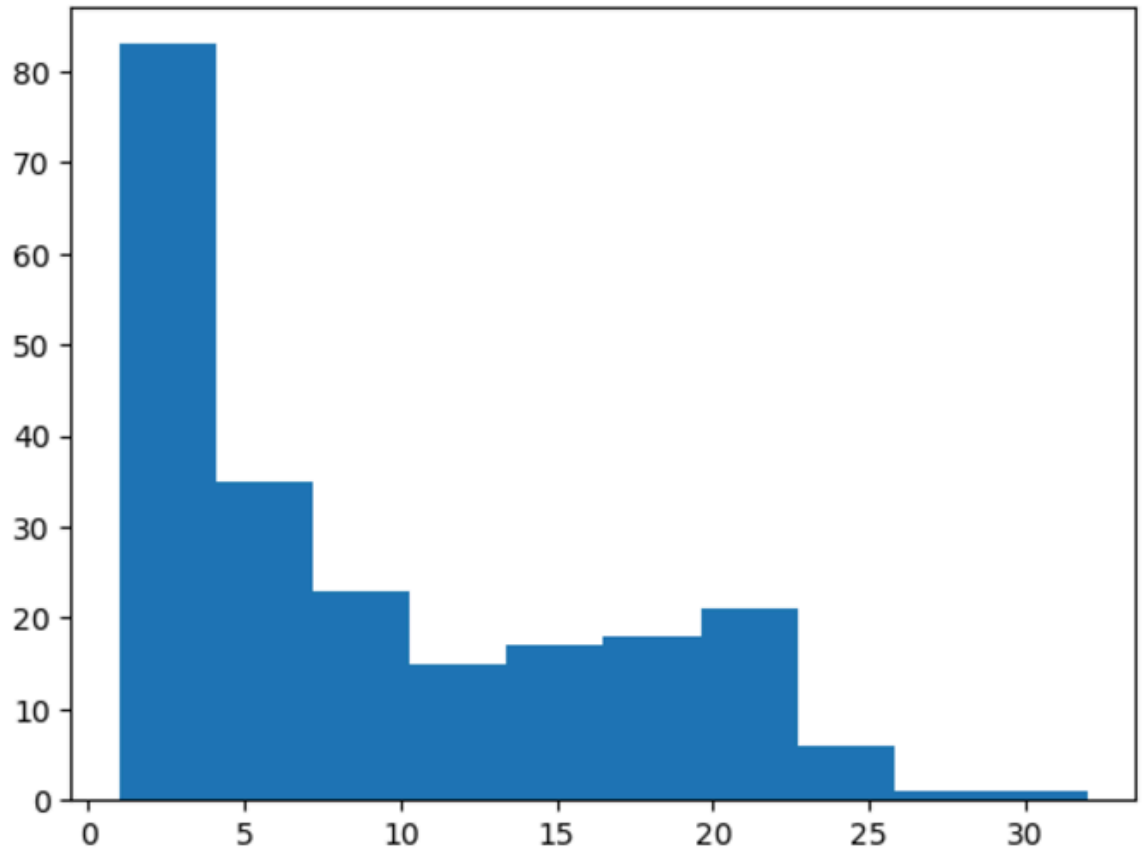
Data Summary

The dataset, **Comprehensive Student Scores Dataset: Insights into Academic Performance and Student Behavior** from Kaggle, includes various demographic and academic features. It provides data on gender, addresses, names, part-time jobs, career aspirations, and weekly study hours. Test scores in subjects such as Math, History, Physics, Chemistry, Biology, English, and Geography are also included.

The target variable, **average test score**, was calculated as the sum of various scores like math, history, geography, etc, divided by 7. The **career_aspirations feature** (with nine unique values) was one-hot encoded.



Average Scores



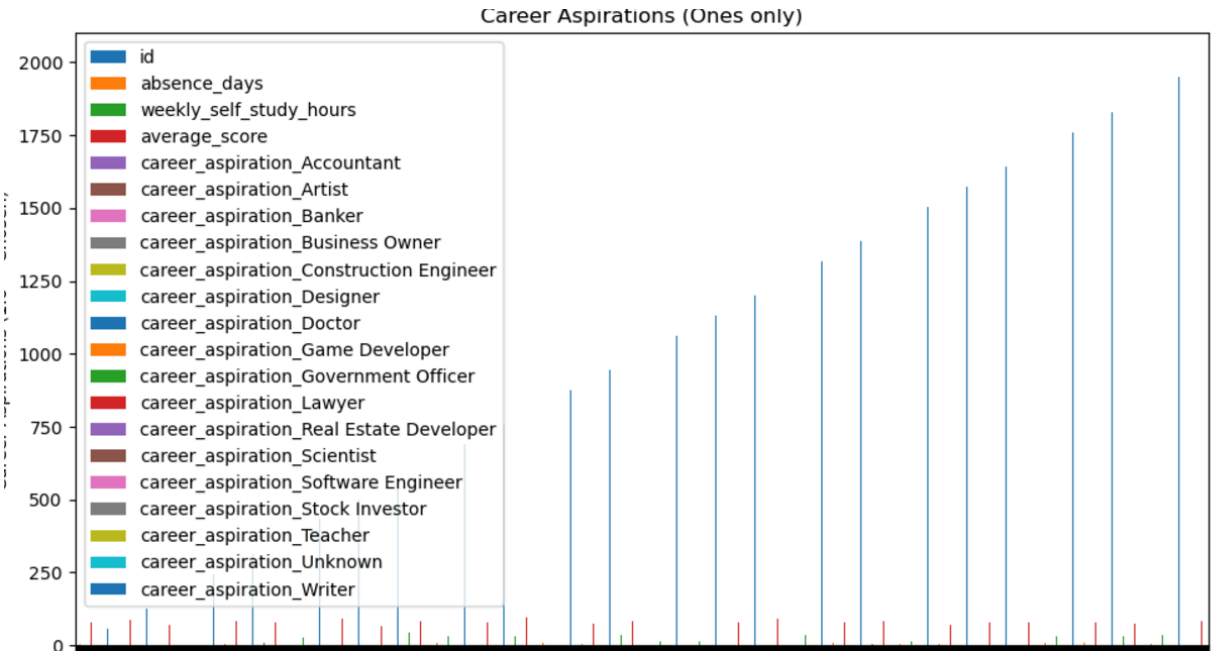
Analysis Objective

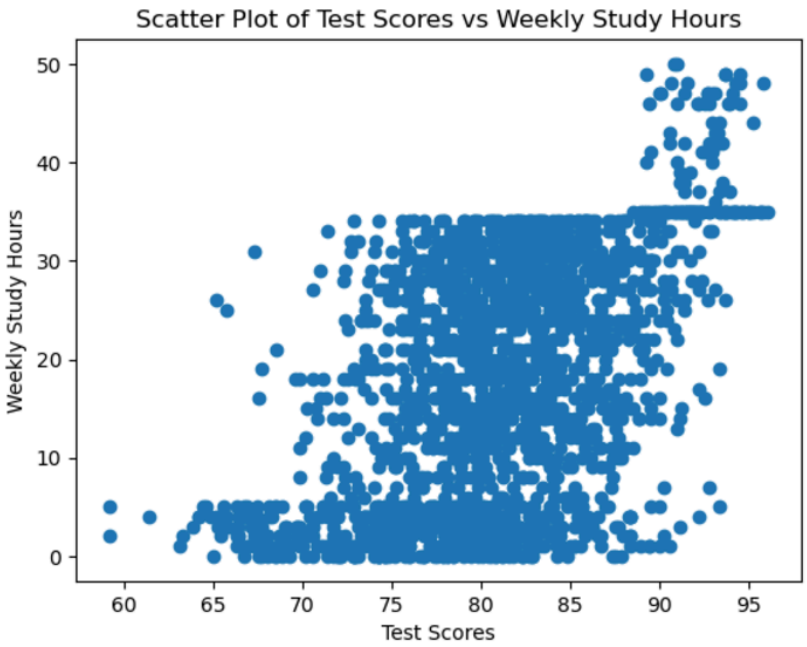
The main objective of this analysis is to predict **average test scores** and determine which features most influence academic performance. After preprocessing, the dataset was split into features (**X**) and target (**Y**), and a standard scaler was applied to normalize the data.

Models tested included **Random Forest Regressor, XGBoost Regressor, and Linear Regression**, with **XGBoost achieving the best Mean Squared Error (MSE) of 0.42**. The Random Forest Regressor had an **MSE of 0.43**, slightly underperforming, while Linear Regression performed less effectively overall.

Due to the limited feature diversity the model is heavily under fitted. I only had 5 features contributing to the prediction. I will later engineer more features to increase the mean squared error. There are also Non-linear relationships between study hours and performance indicating that it will be hard for the model to detect the pattern. There is also an imbalance distribution in career aspiration

---





Model Comparisons

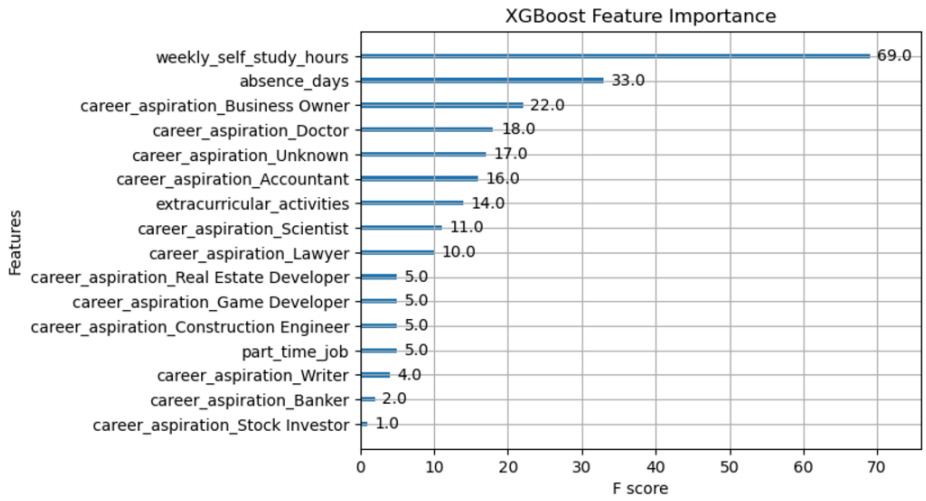
The models tested were designed to predict the average test scores based on study hours, career aspirations, and part-time jobs. The **study_hours feature was identified as the most important predictor** of academic performance. The optimal time to study per week (total for all subjects) is around 35 hours for the best grades(note this varies person to person). . Insights like **optimal study schedules** will guide personalized learning strategies. **GridSearchCV was applied** to optimize the XGBoost model using the following parameters:

python

Copy code

```
param_grid = {
  'learning_rate': [0.01, 0.05, 0.1],
  'n_estimators': [50, 100, 200],
  'max_depth': [3, 5, 7],
  'subsample': [0.7, 0.8, 1.0],
  'colsample_bytree': [0.7, 0.8, 1.0]
}
```

<Figure size 1000x800 with 0 Axes>



Next Steps

The current model's performance could be improved with additional features, such as extracurricular involvement or parental education levels. The small number of features may have led to underfitting in some models. Stacking models could address this issue and boost predictive accuracy. Stacking models typically overfit meaning it will reach the overfit and underfitting balance. I will use XGboost, linear regression, random classifier for voting stack, and use logistic regression as the final voter. I can expand the features by gathering career exploration data or data on extracurricular time.