Predicting Concrete Compressive Strength from Mix Ingredients using Regression Analysis

Group D: Fahim Shahriar, Ayesha Noshin, Sifat Redwan Wahid

November 3, 2025

Abstract

Concrete is a unique building material that has high compressive strength and relatively low tensile strength. Several studies independently have shown that concrete strength development is determined not only by the water-to-cement ratio, but is also influenced by the content of other concrete ingredients. High-performance concrete is a highly complex material, which makes modeling its behavior a very difficult task. In this project we have aimed at developing a robust multiple linear regression model capable of modeling the behavior of concrete strength. According to our analysis, the model does not break any model assumption and provides an R^2 value of 0.83 and and RMSE value of 1.54.

Contents

1 Introduction

Concrete is a common material used in construction, and its strength is important for building safety and durability. Predicting its strength from its ingredients has significant impact on optimizing cost and ensuring long-term reliability. In this project, we used a dataset from the UCI Machine Learning Repository to predict concrete compressive strength based on its ingredients, such as cement, water, and other additives. Our main goal was to apply multiple linear regression (MLR) to build a robust model that can explain how these ingredients affect the concrete strength. We followed all key steps of regression analysis, including data exploration, model selection, selected model investigation, data transformation, checking model assumptions and then finally evaluating the predictive performance of the model. While our model showed satisfactory predictive power, we also discussed its limitations and considered ways to improve it in the future.

1.1 Research Question and Motivation

This project primarily focuses on modeling a dataset using multiple linear regression. We applied key techniques and methods learned in this course to a real-world problem to gain hands-on experience in handling complex datasets.

The dataset we chose addresses a problem from the civil engineering domain. However, domain knowledge is not critical here, as our focus is on applying techniques that demonstrate our understanding of regression analysis. We also formulated two research questions for this project:

RQ1: How accurately can concrete compressive strength be predicted using multiple linear regression based on the mix ingredients?

RQ2: How can the model assumptions be fixed and to what extent it affects the predictive performance of the model.

1.2 Hypotheses

The primary goal of this project is to predict concrete compressive strength based on various mixed ingredients and age. Based on previous research and exploratory analysis, we propose the following hypotheses:

Null Hypothesis, H_0 : There is no significant linear relationship between the predictor variables and concrete compressive strength.

Alternative Hypothesis, H_a : There is a significant linear relationship between at least one predictor variable and concrete compressive strength

1.3 Dataset Description

The data set is collected from UCI Machine Learning Repository. It is named as the "Concrete Compressive Strength" dataset. The target variable of the dataset is concrete compressive strength. The dataset consists of 1030 instances and 8 predictor variables and 1 response variable. The predictor and response variables are briefly described in the following table.

Variable	**Description**	**Unit**	**Variable Name**
Cement	Amount of cement in the mixture	kg/m^3	X1
Blast Furnace Slag	Quantity of slag used as partial cement replacement	kg/m^3	X2
Fly Ash	Quantity of fly ash in the mixture	kg/m^3	X3
Water	Water content in the mixture	kg/m^3	X4
Superplasticizer	Chemical additive to improve workability	kg/m^3	X5
Coarse Aggregate	Amount of coarse aggregate	kg/m^3	X6
Fine Aggregate	Amount of fine aggregate	kg/m^3	X7
Age	Curing period of the concrete	Days	X8
Compressive Strength	Target variable to predict	MPa	Y

Table 1: Dataset overview with descriptions and types

1.4 Exploratory Data Analysis

We generated density plots to perform a univariate analysis of the predictor variables and observed several key patterns in their distributions. The variables Age, blast furnace slag, fly ash, superplasticizer, and water show high right-skewness, indicating that most observations are concentrated at lower values with a few large values. Coarse aggregate and fine aggregate show multi-modal distributions, suggesting the presence of different subgroups or mixtures within the dataset. Concrete compressive strength (the response variable) follows a roughly symmetric distribution, slightly skewed to the right.

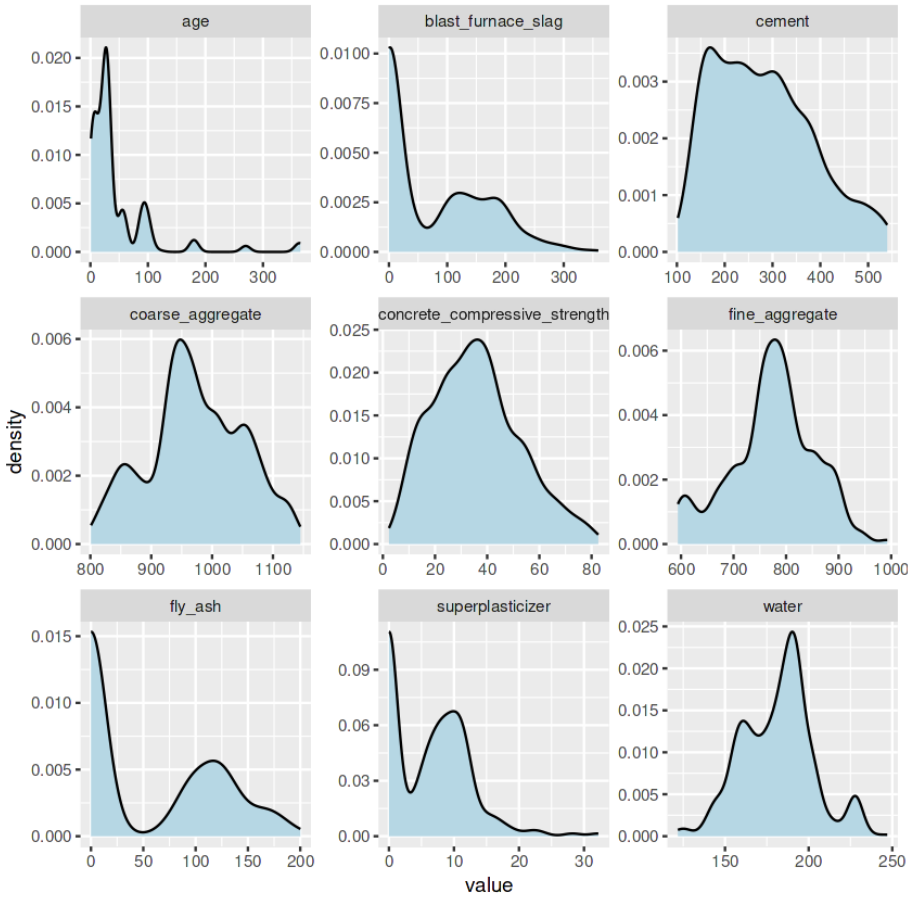

Figure 1: Density plot of the Predictor variables

To analyze the bivariate relationships (the relationship between the predictor and response variable we determined the correlation matrix. From figure 2 it can be seen that no extreme correlation exists between the predictor variables themselves. The highest correlation between predictors is between water and superplasticizer (-0.6575), which is noticeable but not alarming yet. Additionally, Cement has the strongest positive correlation with concrete compressive strength (0.4978), which indicates strong cement is closely associated with concrete strength.

A matrix: 9 × 9 of type dbl

	cement	blast_furnace_slag	fly_ash	water	superplasticizer	coarse_aggregate	fine_aggregate	age	concrete_compressive_strength
cement	1.0000	-0.2752	-0.3975	-0.0816	0.0924	-0.1093	-0.2227	0.0819	0.4978
blast_furnace_slag	-0.2752	1.0000	-0.3236	0.1073	0.0433	-0.2840	-0.2816	-0.0442	0.1348
fly_ash	-0.3975	-0.3236	1.0000	-0.2570	0.3775	-0.0100	0.0791	-0.1544	-0.1058
water	-0.0816	0.1073	-0.2570	1.0000	-0.6575	-0.1823	-0.4507	0.2776	-0.2896
superplasticizer	0.0924	0.0433	0.3775	-0.6575	1.0000	-0.2660	0.2227	-0.1927	0.3661
coarse_aggregate	-0.1093	-0.2840	-0.0100	-0.1823	-0.2660	1.0000	-0.1785	-0.0030	-0.1649
fine_aggregate	-0.2227	-0.2816	0.0791	-0.4507	0.2227	-0.1785	1.0000	-0.1561	-0.1672
age	0.0819	-0.0442	-0.1544	0.2776	-0.1927	-0.0030	-0.1561	1.0000	0.3289
concrete_compressive_strength	0.4978	0.1348	-0.1058	-0.2896	0.3661	-0.1649	-0.1672	0.3289	1.0000

Figure 2: Correlation Matrix of the dataset

2 Regression Analysis

2.1 Model Selection

2.1.1 Insights from Full model

At first, we fit the model with all predictor variables. The summary from full model is presented in Table 2. The Residual vs Fitted and QQ-normal plot are presented in Figure 3 and 4.

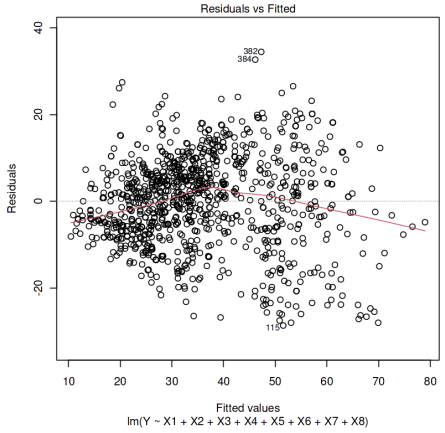

Figure 3: Residual vs Fitted plot

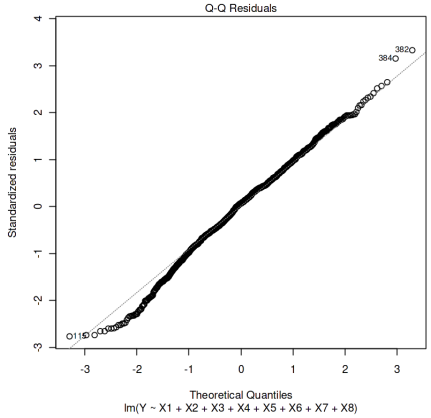

Figure 4: QQ plot

Comparing the p-value of individual predictors at a 95% significance level, we can see that X6 (coarse_aggregate) and X7 (fine_aggregate) can

Table 2: Full Model Summary

Predictor	**Estimate**	**Std. Error**	**Pr(>|t|)**	**Significance**
Intercept	-23.331214	26.585504	0.380372	
X1	0.119804	0.008489	< 2e-16	***
X2	0.103866	0.010136	< 2e-16	***
X3	0.087934	0.012583	5.02e-12	***
X4	-0.149918	0.040177	0.000201	***
X5	0.292225	0.093424	0.001810	**
X6	0.018086	0.009392	0.054425	.
X7	0.020190	0.010702	0.059491	.
X8	0.114222	0.005427	< 2e-16	***

be dropped from the model since they have p-values higher than 0.05. Because our goal was a parsimonious linear model that balances a good fit with model complexity, we employed forward + backward stepwise selection using Akaike's Information Criterion (AIC). AIC penalises additional parameters, so the model with the lowest AIC is preferred. Result of that is presented in Table 3.

Table 3: Forward–backward stepwise selection path based on AIC

Step	Model formula	RSS	AIC
0	$Y \sim 1$	287175	5801.45
1	$Y \sim X_1$	216003	5510.10
2	$Y \sim X_1 + X_5$	186327	5359.88
3	$Y \sim X_1 + X_5 + X_8$	148827	5130.43
4	$Y \sim X_1 + X_5 + X_8 + X_2$	128905	4984.40
5	$Y \sim X_1 + X_5 + X_8 + X_2 + X_4$	119381	4907.34
6	$Y \sim X_1 + X_5 + X_8 + X_2 + X_4 + X_3$ (final)	110843	4832.91

We also employed Forward and Backward selection separately, both of them yielded the same model as the stepwise function. Which is written below:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_5 + \beta_3 X_8 + \beta_4 X_2 + \beta_5 X_4 + \beta_6 X_3$$

2.2 Train-Test Split

As we wanted to evaluate the model's predictive performance we split our dataset into train-test sets using the 80-20 approach. This resulted in 824 and 206 data points for train and test sets respectively.

2.3 Model Assumptions and Diagnostics

In Figure 5 and 6, we can see the residual vs fitted and QQ normal plot of the selected model. In the residual vs fitted plot, we can see a fanning effect, indicating that our model is breaking linearity assumption.

Also, Since the residuals are not forming a horizontal band around zero line,

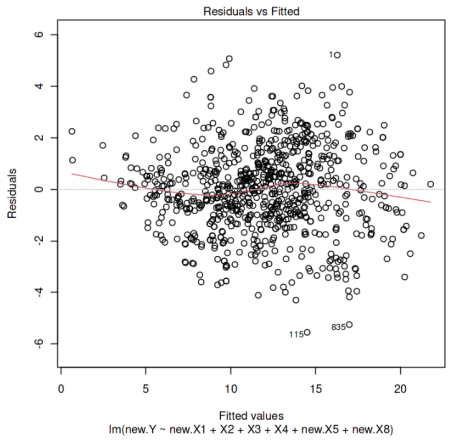

Figure 5: Residual vs Fitted plot

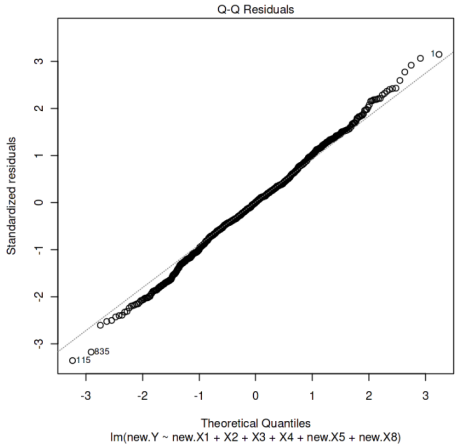

Figure 6: QQ plot

we conduct the Breusch-Pagan test. The test yielded a p-value of 2.2e-16. So, we concluded that the errors do not have a constant variance.

Moreover, upon looking at the QQ plot and box plot of residuals 7, we find some variation from normal line, so we conducted shapiro wilk normality test and found out that error is not normal since the normality test has a p-value of 0.006998.

In order to correct these violations in the assumptions and better fit our model, we tried the following techniques. These techniques and their outcomes are listed below in order:

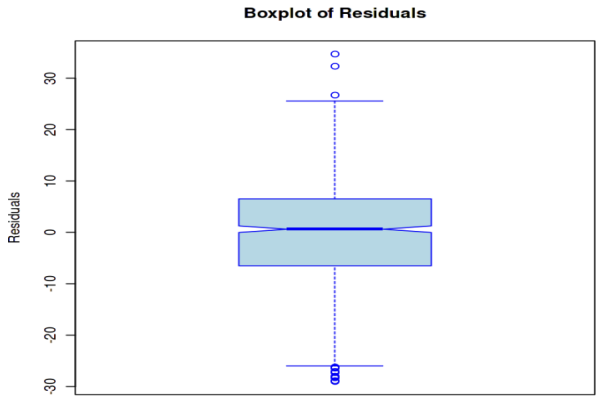

Figure 7: Box Plot of Residuals

2.3.1 Transformations

Y Transformation: To fix non constant error variance and non normal error distribution, we applied a transformation on Y. We looked at the box cox plot to find the value of λ which gives $\lambda = 0.7$.

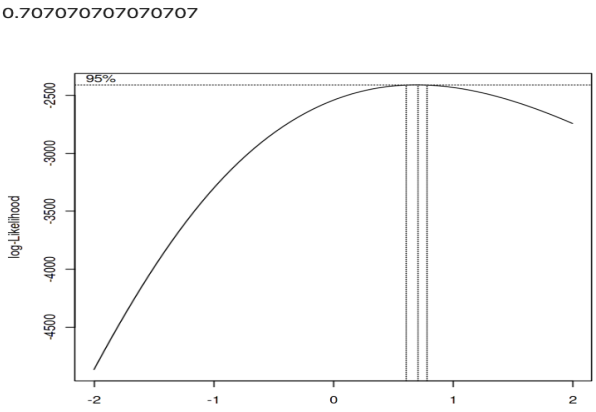

Figure 8: Box Cox Plot

X Transformation: To fix non linearity, we looked at Y vs predictor plots. Based on the patterns on these plots, shown in Figure 9, we decided to apply

log transformation on X5 and X8 and square root transformation on X5.

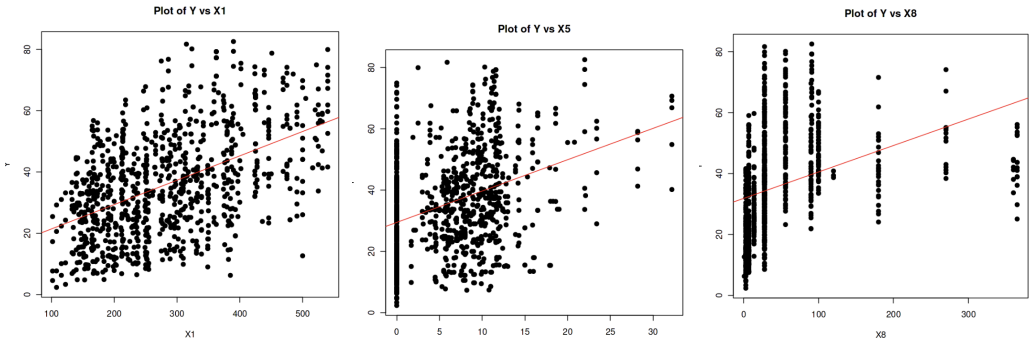

Figure 9: Scatterplots between Y and predictors X1, X5, and X8.

2.3.2 Non Normality fixed

After performing these transformations, we conducted Breusch-Pagan and shapiro wilk normality test again. These tests gave p-value of 3.365e-09 and 0.3375 respectively. These results indicate that error distribution is normal while non-constant error variance persists.

2.3.3 Testing for Multicolinearity:

To test for multicollinearity, we looked at VIF values which are presented in Table 4. All VIFs are well below the common rules-of-thumb of 4 (moderate) and 10 (serious), indicating that multicollinearity is not a practical concern in the chosen model.

2.3.4 Outlier Detection:

To ensure the validity and stability of our regression estimates, we performed two complementary diagnostic checks: one for response outliers (studentized residuals) and one for predictor outliers (high-leverage points).

Response Outliers (Studentized Residuals) We computed the externally studentized residuals

$$t_i \;=\; e_i \sqrt{\frac{n-p-1}{\big(\mathrm{SSE}(1-h_{ii})-e_i^2\big)}}\,,$$

where e_i is the raw residual for observation i, h_{ii} its leverage, and SSE the full-model sum of squared errors. Applying a Bonferroni-adjusted significance level of $\alpha^* = 0.1/(2n)$ yields a critical value

$$t_{\mathrm{crit}} = qt\big(1-\alpha^*,\, n-p-1\big) \approx 3.91.$$

We found no $|t_i|$ exceeding 3.91, so there are no statistically significant outliers with respect to the response.

Predictor Outliers (High Leverage Points) Next, we examined the hat-values h_{ii}. Observations with leverage above twice the average leverage,

$$h_{ii} > 2\,\overline{h}, \quad \text{where} \quad \overline{h} = \frac{p+1}{n},$$

were flagged as high-leverage. Using the rule

$$\mathrm{cutoff} = 2\,\mathrm{mean}(h_{ii}),$$

we identified 45 such points (rows 17, 59, 94, ..., 823). These observations were removed from the training set via negative indexing.

Results of Data Reduction After excluding those 45 high-leverage observations, the cleaned training set contained

$$n_{\mathrm{clean}} \;=\; n - 45$$

observations. This reduction step mitigated undue influence from extreme X-values and leading to more reliable coefficient estimates in our final regression model. It also improved p-value of Breusch-Pagan test to 1.923e-07.

Variable	VIF
X_1	1.87
X_2	1.75
X_3	2.33
X_4	1.93
X_5	2.42
X_8	1.11

Table 4: Variance–Inflation Factors (VIFs) for the final six-predictor model

2.4 Model Fit and Evaluation

We presented the evaluation values from our final model in Table 5. The transformed model explains 83 % of the variation in the response ($R^2 = 0.832$) variable, indicating a strong fit given the complexity of the data. The mean squared error (MSE = 2.36) represents the average squared prediction error on the transformed scale; taking its square root yields the root mean squared error (RMSE = 1.54), which can be interpreted as the typical magnitude of a prediction error. Because RMSE penalises large residuals more than the mean absolute error (MAE = 1.24), the close proximity of RMSE and MAE suggests that extreme prediction errors are limited. That is, no substantial outliers dominate the residual distribution. All four metrics are computed on the transformed outcome.

Evaluation Metric	Transformed model
R^2	0.83
RMSE	1.54
MAE	1.24

Table 5: Evaluation metrics for the transformed model

3 Discussion and Limitations

The dataset included 8 quantitative variables. During model selection, we applied forward, backward, and stepwise regression using AIC. All methods converged on the same model, excluding "Coarse Aggregate" and "Fine Aggregate," so no further model comparison was needed.

We then assessed model assumptions using residual vs. fitted plots, QQ plots, and statistical tests. Initial diagnostics revealed violations of both the normality (Shapiro-Wilk) and constant variance (Breusch-Pagan) assumptions. A Box-Cox transformation of the response variable addressed the normality issue, but heteroscedasticity remained despite additional transformations on selected predictors.

We also checked for multicollinearity using VIF, with all values well below concerning thresholds. Outlier analysis using studentized residuals and leverage values identified 45 high-leverage points, which were removed to enhance model stability.

Finally, we evaluated the model's predictive performance using an 80-20 train-test split. The model achieved an R^2 of 0.83, RMSE of 1.53, and MAE of 1.24 on the test set, indicating strong predictive capability despite model assumption violations.

3.1 Key Findings

- The final model included six predictors: Cement (X1), Blast Furnace Slag (X2), Fly Ash (X3), Water (X4), Superplasticizer (X5), and Age (X8).
- Two most significant predictors of Concrete compressive strength are Superplasticizer and Cement content.
- Three most influencial predictors of Concrete compressive strength are Cement, Age and Superplasticizer.
- Outlier detection methods identified and removed 45 high-leverage points, improving model assumptions.
- The final model achieved an R^2 of 0.83, an RMSE of 1.54, and an MAE of 1.24, indicating good predictive performance.

3.2 Limitations

Multiple Linear Regression (MLR) was used to model this dataset. Although this approach provided good predictive performance this approach assumes a linear relationship between predictors and the response variable, which may oversimplify complex real-world datasets. MLR also assumes independence,

constant error variance, and normality of residuals which is difficult to gurantee for real-world examples. Even after doing several transformations of the dataset it still failed the Breusch-Pagan test. After doing several literature review of this dataset we found that Artificial Neural Network (ANN) could be a better approach to model this large dataset.

3.3 Future Work and Improvements

We performed extensive analysis of the selected model and transformations to fix the model assumptions however our model still breaks the constant error variance assumption. For future work, different data preprocessing techniques and transformations can be done to check for further improvement. For example, the detected outliers can be analyzed to see their influence on the model. We performed extensive analysis of the selected model and transformations to fix the model assumptions however our model still breaks the constant error variance assumption. For future work, different data preprocessing techniques and transformations can be done to check for further improvement. For example, the detected outliers can be analyzed to see their influence on the model. Additionally, applying Weighted Least Squares (WLS) may help account for constant-error variance and lead to a more robust model.

4 Conclusion

In this project, we successfully applied multiple linear regression to model the compressive strength of concrete based on its mix components. Through extensive model selection to model diagnostics we built a model that explains a large portion of the variability in strength ($R^2 = 0.83$) with an RMSE of 1.54. While we addressed issues like non-normality and multicollinearity, our model still violated the constant variance assumption, even after transformations. Despite this, the model performed reasonably well. Overall, this project helped us apply key regression techniques in a real-world dataset and understand both the strengths and limitations of multiple linear regression models.

5 Additional Work

5.1 Alternative Models Considered

We first standardized each predictor to zero mean and unit variance, aiming to mitigate scale differences across variables, but observed no meaningful change in coefficient estimates or overall model fit (R^2, RMSE, MAE). Next, we applied min–max normalization to rescale variables to the [0,1] range, yet diagnostic plots still showed heteroscedasticity and slight deviations from normality in the residuals. To explore potential synergistic effects, we introduced interaction terms (e.g., cement $\times$ age and water $\times$ superplasticizer) into the regression model, but it also failed to improve performance.

None of these adjustments yielded improvements in either assumption diagnostics or predictive performance, so they were excluded to maintain a less complex, interpretable model.

6 References

- Yeh, I-Cheng. "Modeling of strength of high performance concrete using artificial neural networks." *Cement and Concrete Research*, vol. 28, no. 12, 1998, pp. 1797–1808.
- UCI Machine Learning Repository. `https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength`

A R Code

Link to the R notebook we used can be found here: Github Repository