Homework 8

You will be working with a dataset that explores the **compressive strength of concrete mixtures**, originally published by Yeh (2006) in a civil engineering journal and later featured in *Applied Predictive Modeling* by Kuhn and Johnson (2013). This dataset comes from a series of experiments designed to understand how different components in concrete—such as cement, slag, fly ash, water, and aggregates—affect its overall strength.

Each row in the dataset represents a unique concrete mixture used in an experimental test. In addition to the quantities of the various mix components, the dataset includes the **age of the mixture in days**, which reflects the curing time before testing. The main outcome of interest is the **compressive strength** of the concrete, a key measure of its durability and performance in structural applications.

This real-world dataset provides a meaningful context to explore predictive modeling in regression. It captures the complexity of physical materials, the interaction between ingredients, and the effects of aging—offering a rich setting for applying and interpreting linear models.

library(tidymodels) ## Data Extraction --- E

library(dplyr) ## Data Transformation --- T  
library(ggplot2) ## Data Visualization --- V  
library(GGally) ## Data Visualization --- V

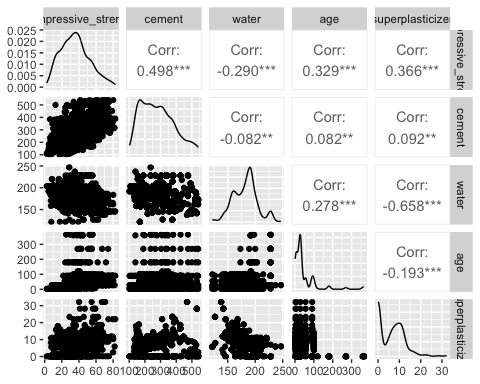
In this project, we aim to predict **compressive strength**, a measure of how much pressure a concrete mixture can withstand. To build a meaningful model, it’s important to choose explanatory variables that have a logical and scientific connection to the strength of the concrete. The four variables we selected—**cement**, **water**, **age**, and **superplasticizer**—are not only commonly discussed in concrete design literature, but they also reflect fundamental aspects of how concrete hardens and gains strength.

We included **cement** because it acts as the primary binding agent in concrete. A higher cement content typically results in stronger concrete, as more cement leads to a denser and more cohesive mixture. **Water** is also critical—concrete needs water to undergo hydration, the chemical reaction that causes hardening. However, excess water can dilute the mixture and create pores, reducing strength. The **age** of the concrete is another essential factor because compressive strength develops over time. Most concrete gains the majority of its strength within the first 28 days, so age allows us to capture how curing time affects performance. Finally, **superplasticizer** is a chemical additive that allows for easier mixing and flow without increasing water content. This can improve strength by enabling workability while preserving a low water-to-cement ratio.

Together, these variables capture a balance of chemical composition, material proportions, and time—three critical elements that influence how concrete sets and performs. By including them, we are making informed hypotheses grounded in engineering principles, which will help us build a more accurate and interpretable regression model.

# Create a new dataframe with selected variables  
concrete\_df <- concrete %>%  
 select(compressive\_strength, cement, water, age, superplasticizer)

# Create a pairs plot  
ggpairs(concrete\_df)



## Dissecting the **age** variable

Based on the density plot of **age** in your pairs plot, we observe that the variable is highly right-skewed, with distinct clusters around low integer values. Most observations fall at or near specific values such as **1, 7, 14, 28, 56, and 90 days**, which are standard curing periods in concrete testing. These points likely correspond to meaningful time-based testing intervals used in industry and research.

To simplify modeling and improve interpretability, it would make sense to **group age into ordered categorical bins** based on these observed concentrations. Here’s a good choice of breaks:

**Suggested Age Breaks**:

* **1–7 days** → “Very Early”
* **8–14 days** → “Early”
* **15–28 days** → “Standard”
* **29–56 days** → “Late”
* **57+ days** → “Long-term”

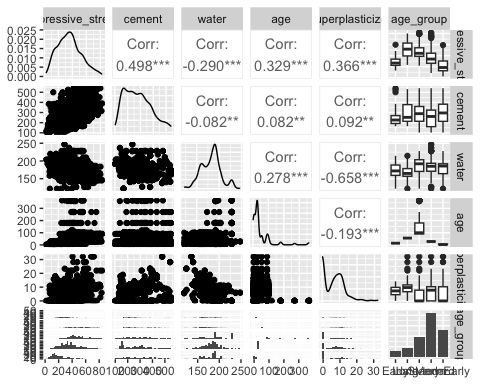
These breaks align with traditional testing windows in concrete design (especially the 28-day benchmark), balance sample sizes across bins, and retain meaningful interpretation.

**R Code to Create Age Groups**:

mod\_1\_concrete\_df <- concrete\_df %>%  
 mutate(age\_group = case\_when(  
 age <= 7 ~ "Very Early",  
 age <= 14 ~ "Early",  
 age <= 28 ~ "Standard",  
 age <= 56 ~ "Late",  
 TRUE ~ "Long-term"  
 ))

We chose these age breaks to reflect commonly used industry curing times while reducing the skewness of the original variable. Grouping age in this way helps us capture non-linear changes in strength over time without forcing a linear relationship. It also improves interpretability, as students can connect strength differences to concrete’s development stages (e.g., initial set vs. full cure).

# Create a pairs plot  
ggpairs(mod\_1\_concrete\_df)



Upon closer examination of the pairs plot, we observe that **compressive strength peaks in the “Early” age group** before declining in later stages. This non-monotonic trend suggests that a simple linear relationship with age may not fully capture the underlying dynamics—opening the door to a more nuanced model that treats age as a categorical or non-linear predictor. Additionally, we notice that **water content is negatively correlated with both compressive strength and superplasticizer**, which aligns with the known trade-off between workability and strength. Another key observation is that **superplasticizer use appears skewed and concentrated at lower values**, potentially indicating sporadic or specialized usage in certain mixes rather than being uniformly applied across all age groups.

These five models were selected to reflect patterns observed in the exploratory data visualizations while exploring a range of modeling strategies. **Model 1** serves as a baseline, incorporating three continuous predictors—cement, water, and age—based on their linear relationships with compressive strength. **Model 2** builds on this by introducing an interaction between cement and water, motivated by the negative correlation between the two and the engineering relevance of the water-to-cement ratio. From the plots, we saw that compressive strength varied across the age\_group categories in a non-linear pattern, with strength peaking in the “Early” group. This observation motivated **Model 3**, which replaces age with age\_group to capture non-linear effects.

**Model 4** takes this further by allowing the effect of cement on strength to vary across age groups (via interaction), acknowledging that the role of cement may differ depending on how long the concrete has cured. Similarly, **Model 5** tests whether the relationship between water content and compressive strength changes across age groups, based on the patterns in the boxplots. Together, these models provide a structured way to compare additive and interaction effects while incorporating both quantitative and categorical predictors.

# Model 1:   
model\_1 <- lm(compressive\_strength ~ cement + water + age, data = mod\_1\_concrete\_df)  
# Model 2:   
model\_2 <- lm(compressive\_strength ~ cement \* water, data = mod\_1\_concrete\_df)  
# Model 3:   
model\_3 <- lm(compressive\_strength ~ cement + water + age\_group, data = mod\_1\_concrete\_df)  
# Model 4:   
model\_4 <- lm(compressive\_strength ~ age\_group \* cement , data = mod\_1\_concrete\_df)  
# Model 5:   
model\_5 <- lm(compressive\_strength ~ age\_group \* water, data = mod\_1\_concrete\_df)

## Model 1

summary(model\_1)

Call:  
lm(formula = compressive\_strength ~ cement + water + age, data = mod\_1\_concrete\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-37.619 -9.381 -0.592 8.310 40.976   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 63.096395 3.646504 17.30 <2e-16 \*\*\*  
cement 0.069678 0.003729 18.68 <2e-16 \*\*\*  
water -0.284347 0.018935 -15.02 <2e-16 \*\*\*  
age 0.104212 0.006401 16.28 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 12.39 on 1026 degrees of freedom  
Multiple R-squared: 0.4519, Adjusted R-squared: 0.4503   
F-statistic: 281.9 on 3 and 1026 DF, p-value: < 2.2e-16

## Model 2

summary(model\_2)

Call:  
lm(formula = compressive\_strength ~ cement \* water, data = mod\_1\_concrete\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-36.321 -10.806 0.279 9.106 41.691   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 25.7654305 11.2189895 2.297 0.0218 \*   
cement 0.1553303 0.0344956 4.503 7.47e-06 \*\*\*  
water -0.0625532 0.0613391 -1.020 0.3081   
cement:water -0.0004377 0.0001897 -2.307 0.0212 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 13.86 on 1026 degrees of freedom  
Multiple R-squared: 0.3138, Adjusted R-squared: 0.3118   
F-statistic: 156.4 on 3 and 1026 DF, p-value: < 2.2e-16

## Model 3

summary(model\_3)

Call:  
lm(formula = compressive\_strength ~ cement + water + age\_group,   
 data = mod\_1\_concrete\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-22.801 -6.575 -0.790 6.059 33.876   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 49.348371 2.912710 16.942 < 2e-16 \*\*\*  
cement 0.078925 0.002864 27.555 < 2e-16 \*\*\*  
water -0.230851 0.014410 -16.020 < 2e-16 \*\*\*  
age\_groupLate 17.979820 1.559510 11.529 < 2e-16 \*\*\*  
age\_groupLong-term 18.252231 1.408409 12.959 < 2e-16 \*\*\*  
age\_groupStandard 8.709520 1.290314 6.750 2.47e-11 \*\*\*  
age\_groupVery Early -9.261139 1.343907 -6.891 9.66e-12 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 9.423 on 1023 degrees of freedom  
Multiple R-squared: 0.6837, Adjusted R-squared: 0.6818   
F-statistic: 368.5 on 6 and 1023 DF, p-value: < 2.2e-16

## Model 4

summary(model\_4)

Call:  
lm(formula = compressive\_strength ~ age\_group \* cement, data = mod\_1\_concrete\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-28.865 -7.716 -0.570 6.501 40.362   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 16.52017 4.23703 3.899 0.000103 \*\*\*  
age\_groupLate 8.10213 5.43238 1.491 0.136152   
age\_groupLong-term 15.01507 4.81936 3.116 0.001887 \*\*   
age\_groupStandard -4.62049 4.45557 -1.037 0.299974   
age\_groupVery Early -20.85373 4.65690 -4.478 8.38e-06 \*\*\*  
cement 0.04968 0.01635 3.038 0.002439 \*\*   
age\_groupLate:cement 0.04301 0.01968 2.186 0.029047 \*   
age\_groupLong-term:cement 0.00266 0.01793 0.148 0.882078   
age\_groupStandard:cement 0.04393 0.01705 2.577 0.010118 \*   
age\_groupVery Early:cement 0.03900 0.01744 2.236 0.025571 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 10.41 on 1020 degrees of freedom  
Multiple R-squared: 0.6149, Adjusted R-squared: 0.6115   
F-statistic: 181 on 9 and 1020 DF, p-value: < 2.2e-16

## Model 5

summary(model\_5)

Call:  
lm(formula = compressive\_strength ~ age\_group \* water, data = mod\_1\_concrete\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-32.705 -8.527 -1.476 7.293 38.176   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 53.47198 15.13039 3.534 0.000428 \*\*\*  
age\_groupLate 48.54316 19.56879 2.481 0.013275 \*   
age\_groupLong-term 31.69435 16.55417 1.915 0.055825 .   
age\_groupStandard 38.80747 16.18194 2.398 0.016655 \*   
age\_groupVery Early 23.67525 16.73906 1.414 0.157558   
water -0.14257 0.08679 -1.643 0.100733   
age\_groupLate:water -0.15677 0.11386 -1.377 0.168858   
age\_groupLong-term:water -0.05824 0.09358 -0.622 0.533856   
age\_groupStandard:water -0.16077 0.09222 -1.743 0.081557 .   
age\_groupVery Early:water -0.16254 0.09540 -1.704 0.088721 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 12.41 on 1020 degrees of freedom  
Multiple R-squared: 0.4533, Adjusted R-squared: 0.4484   
F-statistic: 93.96 on 9 and 1020 DF, p-value: < 2.2e-16