Predicting the Compressive Strength of Concrete by its Composition

Sam Wigley

December 02, 2015

Abstract

This report is an exportatory data analysis on a dataset provided by the University of California at Irvine's Machine Learning department. The dataset contains 1030 observations including 9 quantitative variables.

The UCI ML Cement dataset records compressive strengh of cement in MPa (Megapascals) given 8 other input variables, which are the amonts of the components in Kg, and it's age in days.

The readme for the dataset can be downloaded here: http://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/Concrete_Readme.txt

This study aims to define what a relatively strong concrete is in terms of compressive strength, then to see if we can predict the compressive strength of concrete based on the relative proportion of it's ingredients and it's age.

I selected this dataset not only becasue it met the requirements for the Udacity Exporatory Data Analysis course final project, but in a former job, I advised on computer datalogger setup for a chemical engineer who performed a similar experiment. This particular experiment was to test the compressive strength of concrete with varying quantities of synthetic and organic fiber additives. I did a little bit of research just now and found the results of that specific study here: http://www.solomoncolors.com/ResourceDownload.aspx?id=3085. I worked for a few months at Buckeye Cellulose during the product development cycle of what became known as the UltraFiber 500 concrete reinforcement product. This assigment was one of the specific moments in my IT career that led me to pursue a path more related to science and data analytics.

Univariate Plots Section

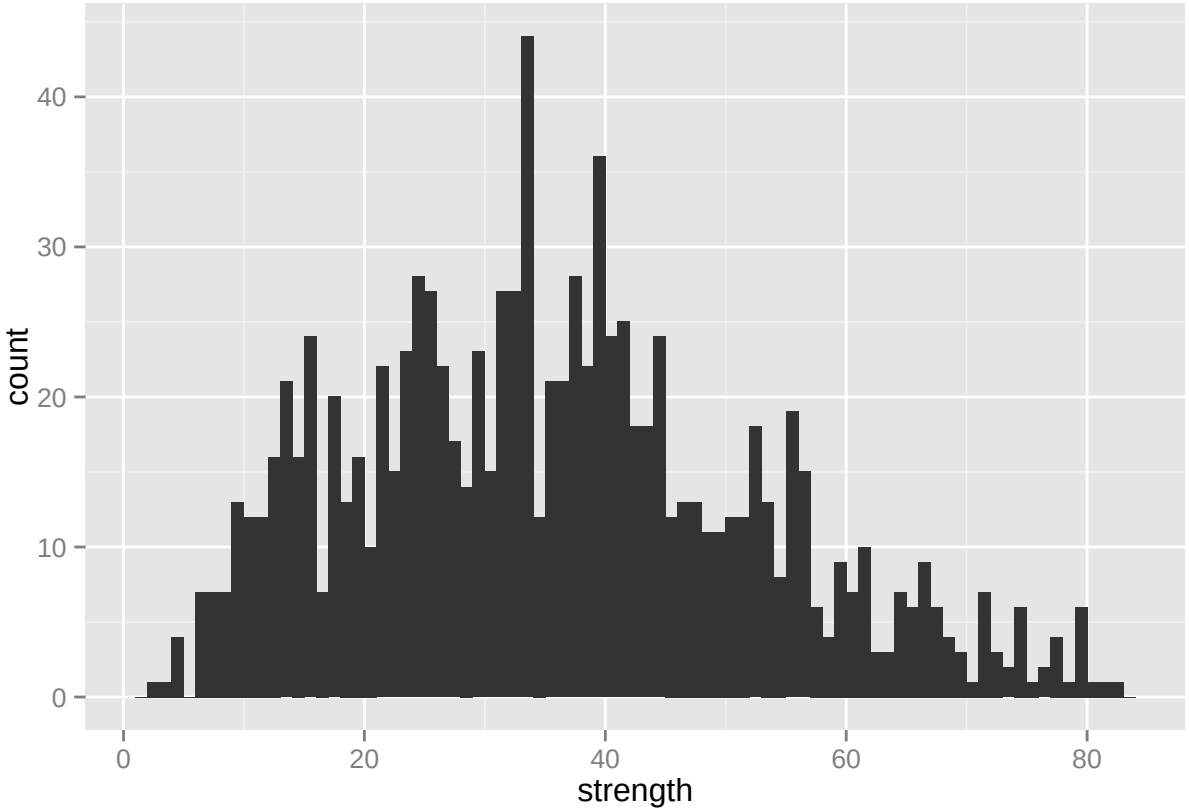
Input Data Summary

```
##      cement           slag           flyash           water      
##  Min.   :102.0   Min.   :  0.0   Min.   :  0.00   Min.   :121.8  
##  1st Qu.:192.4   1st Qu.:  0.0   1st Qu.:  0.00   1st Qu.:164.9  
##  Median :272.9   Median : 22.0   Median :  0.00   Median :185.0  
##  Mean   :281.2   Mean   : 73.9   Mean   : 54.19   Mean   :181.6  
##  3rd Qu.:350.0   3rd Qu.:142.9   3rd Qu.:118.30   3rd Qu.:192.0  
##  Max.   :540.0   Max.   :359.4   Max.   :200.10   Max.   :247.0  
##   plasticizer       coarse_agg       fine_agg          age        
##  Min.   : 0.000   Min.   : 801.0   Min.   :594.0   Min.   :  1.00  
##  1st Qu.: 0.000   1st Qu.: 932.0   1st Qu.:731.0   1st Qu.:  7.00  
##  Median : 6.400   Median : 968.0   Median :779.5   Median : 28.00  
##  Mean   : 6.205   Mean   : 972.9   Mean   :773.6   Mean   : 45.66  
##  3rd Qu.:10.200   3rd Qu.:1029.4   3rd Qu.:824.0   3rd Qu.: 56.00  
##  Max.   :32.200   Max.   :1145.0   Max.   :992.6   Max.   :365.00  
##     strength    
##  Min.   : 2.33  
```

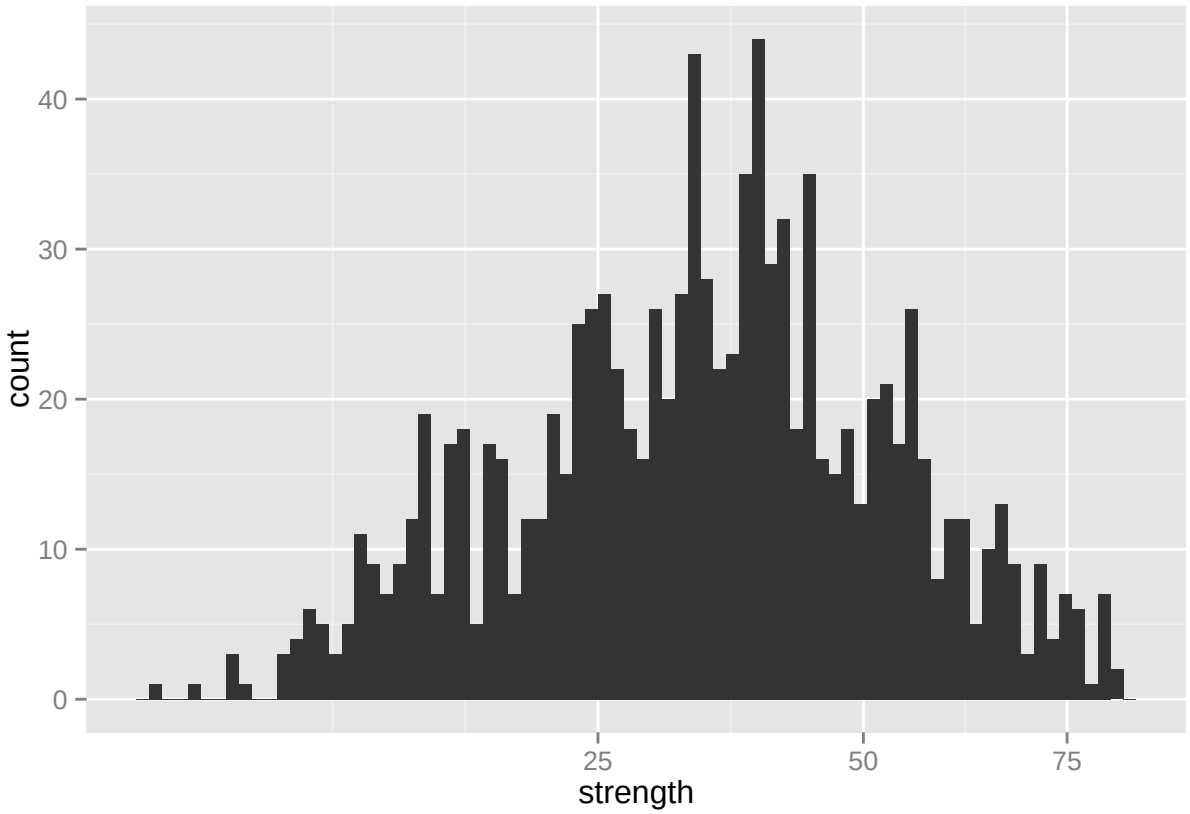
```
##  1st Qu.:23.71
##  Median :34.45
##  Mean   :35.82
##  3rd Qu.:46.13
##  Max.   :82.60
```

All of the variables are quantitive. Seven of the variables are weights in kg of components of a mixture. There is also an age variable which ranges from 1 to 365 in days, so these samples vary in cure time with the median cure time at 28 days and a maximum of 365 days or 1 year. That's a pretty big range, considering concrete that is only a few days old won't be expected to harden very much. Another interesting observation is that slag, flyash, and plasticizer aren't present in all of the samples. Interestingly, flyash has a median of zero, so most of the samples don't have any fly ash.

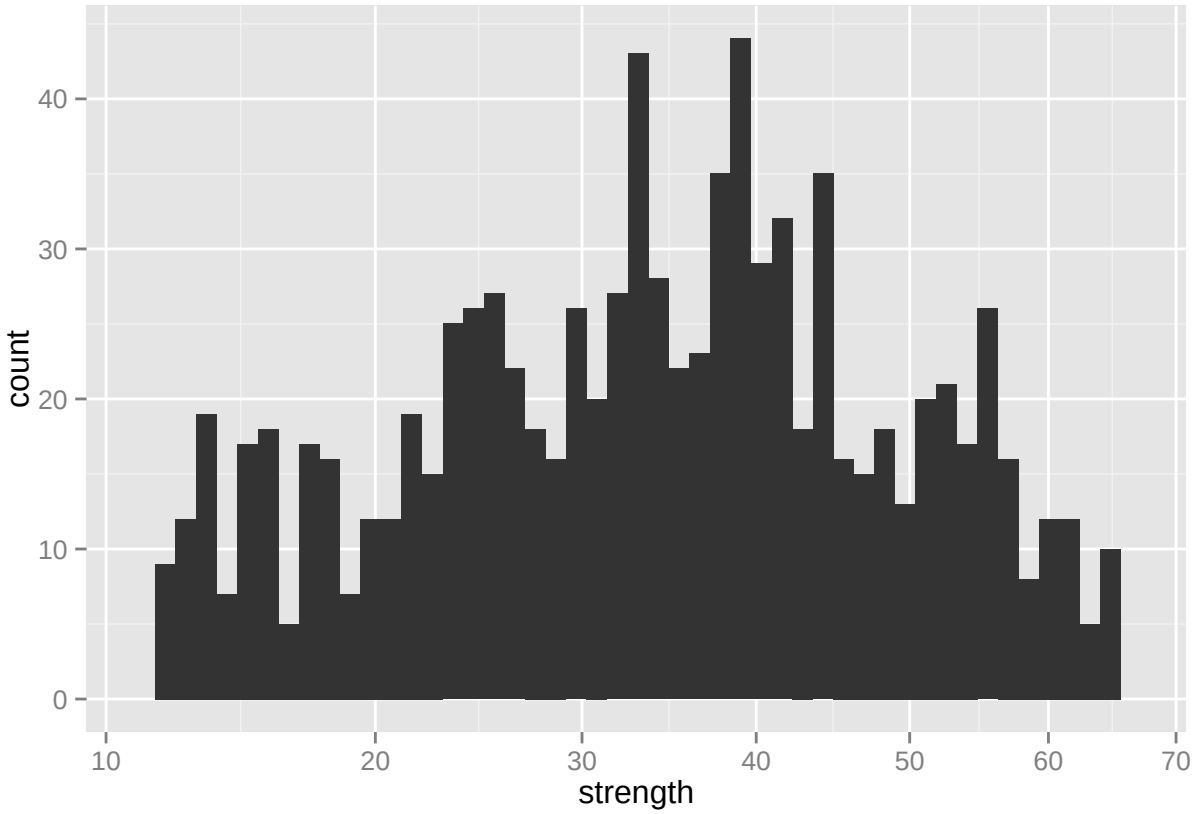
Histograms showing the distribution of values in the original set



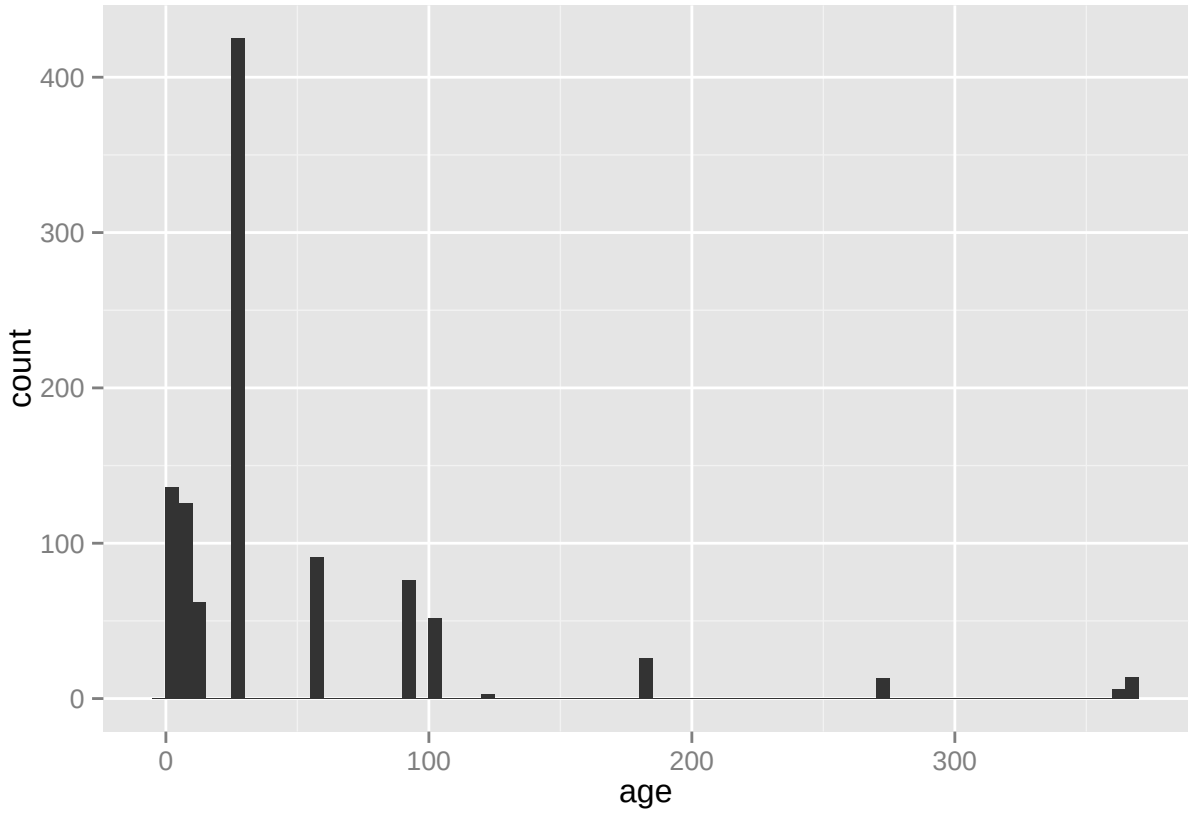
Strength of the samples appears to be almost normally distributed, with most samples being in the 50 MPa range, with the strongest samples in the 60-90 MPa range. However it's slightly right-skewed. So I'll try to improve it.



Now the transformed strength appears to be normally distributed. It looks like we could knock off a couple of outliers, and move it close to the center, though.

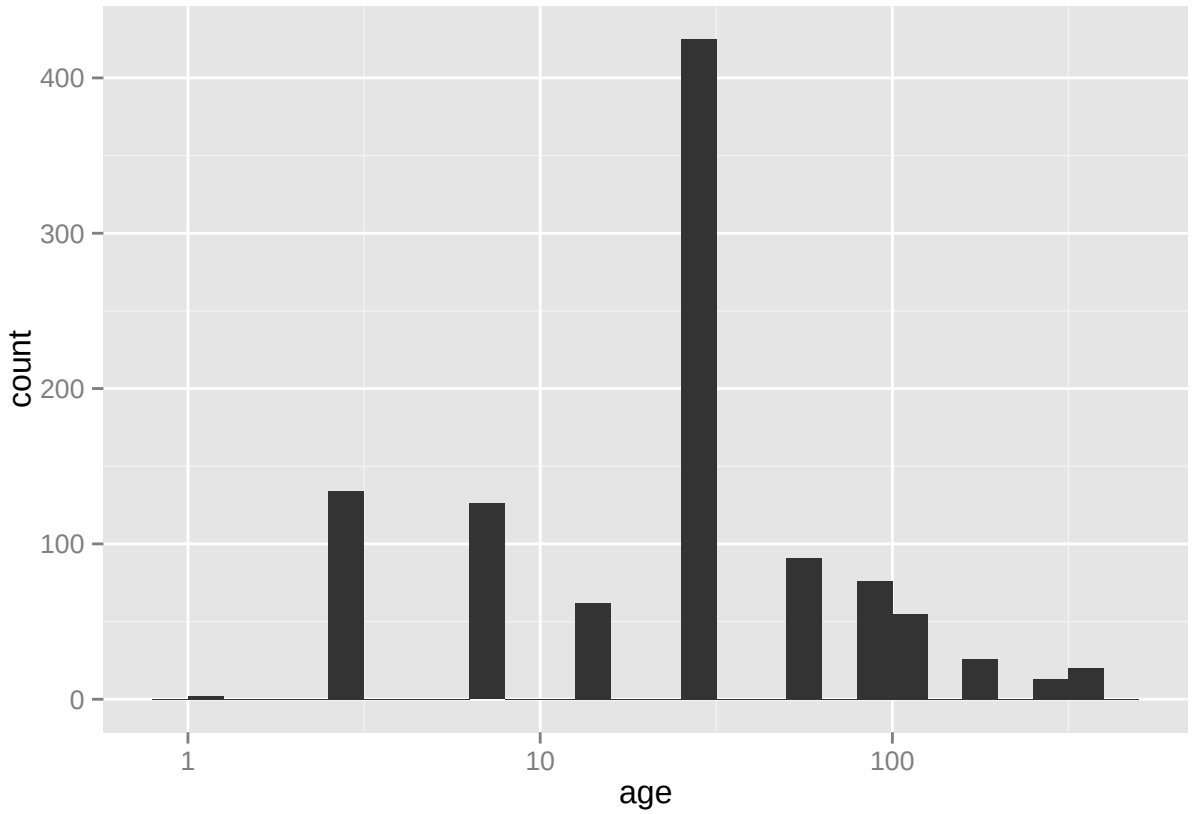


The final histogram appears to be close to a normal distribution.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   28.00   45.66   56.00  365.00
```

The age histogram shows most tests are taken from samples that have cured for 50 days or less. Very few samples were tested that were over 100 days old, however the oldest sample is 365 days, or 1 year old. The median age is 28 days. Age is also quite a bit right skewed, so I'll apply the log10 transformation there as well.



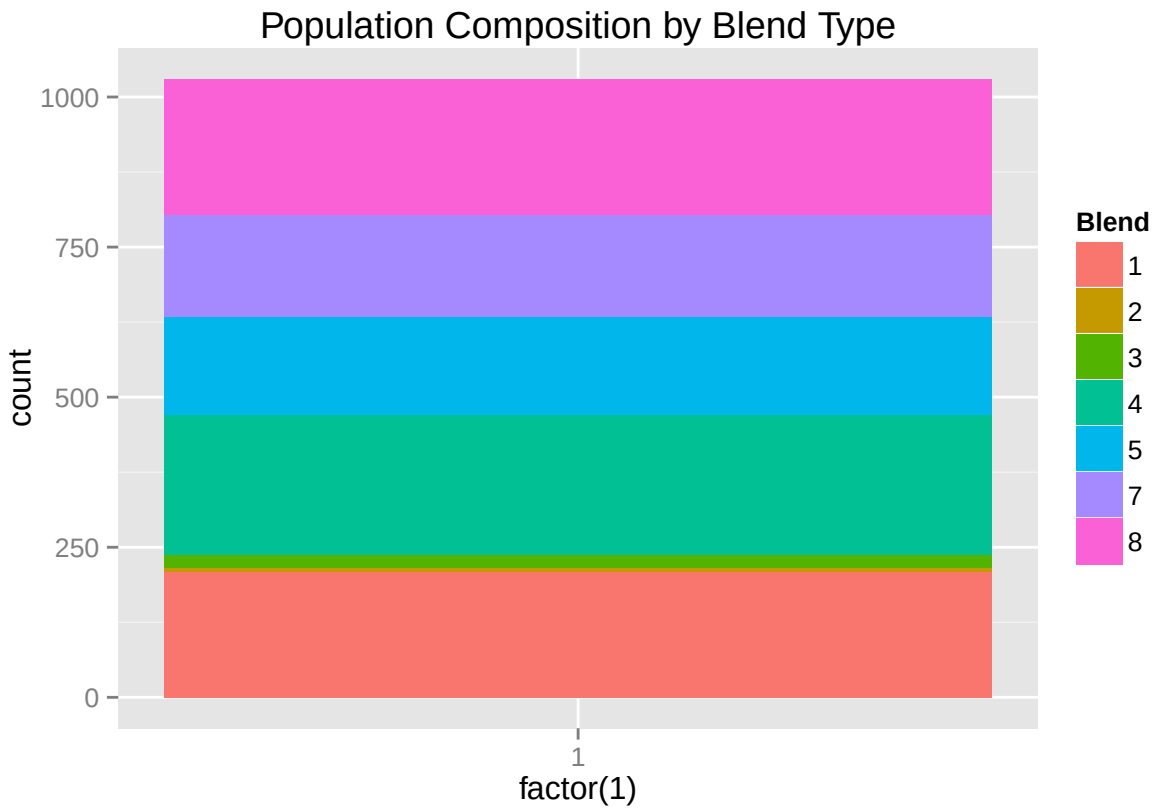
log10(age) looks more normally distributed with the mean age moving to the center of the distribution, so I'll keep this transformation for the model.

Check For the Absence of Components

From the data summary above we can see that several components are showing that they have zero presence in a lot of the observations. These must be optional additives, as water, cement, and aggregates are always ingredients. Here we will see how many samples omit additives such as slag, plasticizer, and flyash.

	Zeros
cement	0
slag	471
flyash	566
water	0
plasticizer	379
coarse_agg	0
fine_agg	0
age	0
strength	0

Of the 1030 samples more than half don't contain flyash, and nearly one third to one half of them are missing either plasticizer or slag.




```
## discrete_scale(aesthetics = "colour", scale_name = "brewer",
##     palette = brewer_pal(type, palette))
```

Table 2: Table showing the various identified blends based on the presence of additives

Blend	Slag	Plasticizer	FlyAsh	Obs_Count
1	FALSE	FALSE	FALSE	209
2	FALSE	FALSE	TRUE	6
3	FALSE	TRUE	FALSE	23
4	FALSE	TRUE	TRUE	233
5	TRUE	FALSE	FALSE	164
6	TRUE	FALSE	TRUE	0
7	TRUE	TRUE	FALSE	170
8	TRUE	TRUE	TRUE	225

I decided to create classifications for blends based on the presence or absence of combinations of the optional additives, since we have three optional additives, we have 8 possible blend combinations. For this study, I label the blends 1 through 8, where blend 1 has none of the optional additives and blend 8 has all of them. Blends 1-4 have no slag. 5-8 have slag. Blends 3,4,7 and 8 contain plasticizer, and even numbered blends contain flyash.

This bar plot and table above show the makeup of the test poulation by blend classification. As you can see the blends are not evenly represented. In fact blends 2 and 3 have fewer than 30 observations between them, and there are no samples representing blend 6, so we'll have to work without it.

Mixture As a Ratio

Since all of the components in each sample that are ingredients in the mixture are measured in kg, I though it would only make sense to compare them as a ratio. That led me to the question: What ratio? I think there are a few valid choices which might have value. I tried a ratio of each ingredient to the water weight, a ratio to the cement weight, and a ratio to the total weight of the sum of the ingredients' weights.

I have a feeling there are some complex relationships going on here, and some components may have a stronger relationship with hydration, while others, may have a more interesting relationship with their proportion of the cement. For example, if there was some optimal blend between cement and one of the additives to produce a strong crystaline molecular structure, as tempered steel has with carbon, the cement ratio may have more value. Where another additive may nave a stronger association with it's concentration in relation to water.

To limit the scope of this project, I'm going to choose one and focus most of the work on it. I think a more complex model could be built that took advantange of multiple transformations of the dataset.

I finally decided to use the ratio of each components to the weight in cement. My my reasoning was that this is usually how, I think, cement is measured when mixed. But this was a slight source of struggle to pick the best ratio, and I examined all three before settling with the cement ratio. Later in the correlation plots was really when I decided. The correlations seemed to make the most sense like there may be a pattern there. So I selected it a little bit intuitively, without a quantitive basis, but I roughly checked all three and think there are some valuable properties in each ratio

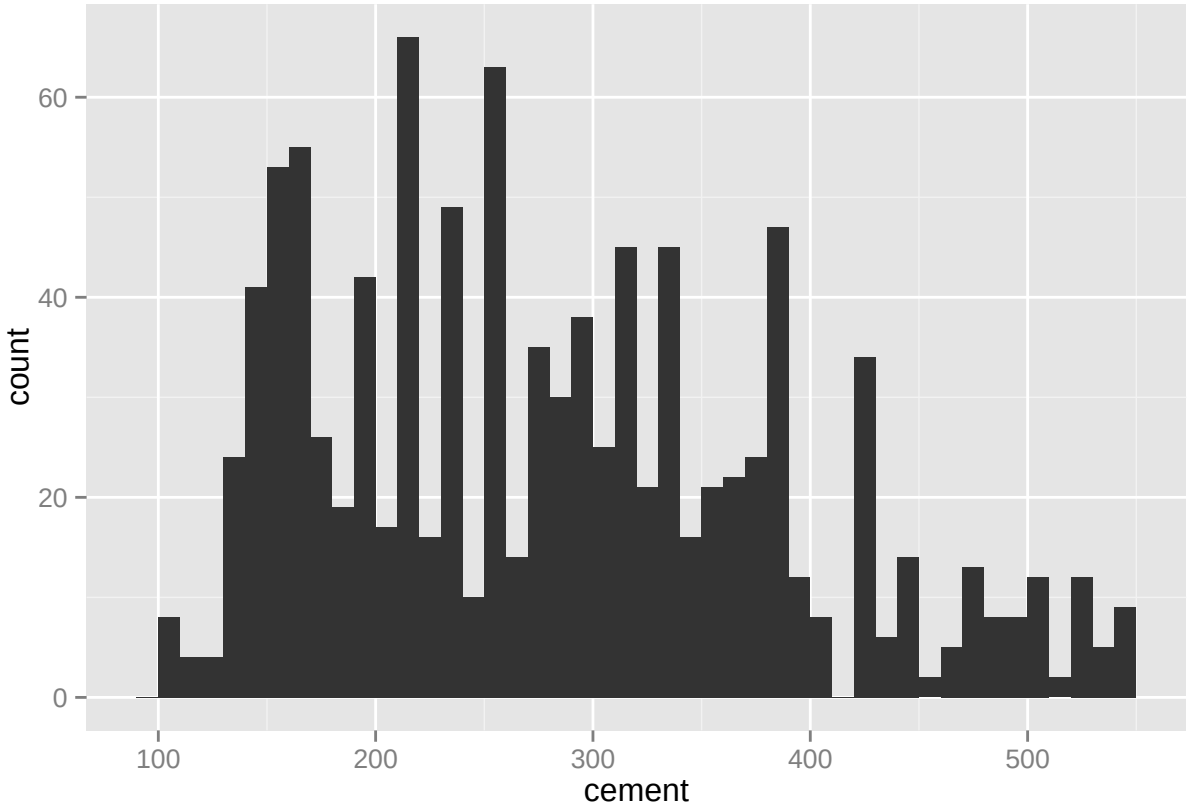
Cement Ratios

Ratio summary

```
##      cement       slag             flyash           water
##  Min.   :1   Min.   :0.00000   Min.   :0.0000   Min.   :0.2669
##  1st Qu.:1   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.5333
##  Median :1   Median :0.05352   Median :0.0000   Median :0.6753
##  Mean   :1   Mean   :0.34751   Mean   :0.2591   Mean   :0.7483
##  3rd Qu.:1   3rd Qu.:0.53740   3rd Qu.:0.4706   3rd Qu.:0.9352
##  Max.   :1   Max.   :1.58389   Max.   :1.4296   Max.   :1.8824
##   plasticizer        coarse_agg       fine_agg          age
##  Min.   :0.00000   Min.   :1.552   Min.   :1.135   Min.   :  1.00
##  1st Qu.:0.00000   1st Qu.:2.724   1st Qu.:2.183   1st Qu.:  7.00
##  Median :0.02528   Median :3.600   Median :2.892   Median : 28.00
##  Mean   :0.02472   Mean   :4.000   Mean   :3.199   Mean   : 45.66
##  3rd Qu.:0.04038   3rd Qu.:5.144   3rd Qu.:4.157   3rd Qu.: 56.00
##  Max.   :0.12500   Max.   :8.696   Max.   :9.235   Max.   :365.00
##     strength     additive_coef
##  Min.   : 2.33   Min.   :1.000
##  1st Qu.:23.71   1st Qu.:4.000
##  Median :34.45   Median :5.000
##  Mean   :35.82   Mean   :4.885
##  3rd Qu.:46.13   3rd Qu.:7.000
##  Max.   :82.60   Max.   :8.000
```

Above is a summary of the data in terms of each component's ratio to the level of cement. The units for each component weight in this ratio dataset are in kg per 1kg of cement.

Cement Plots

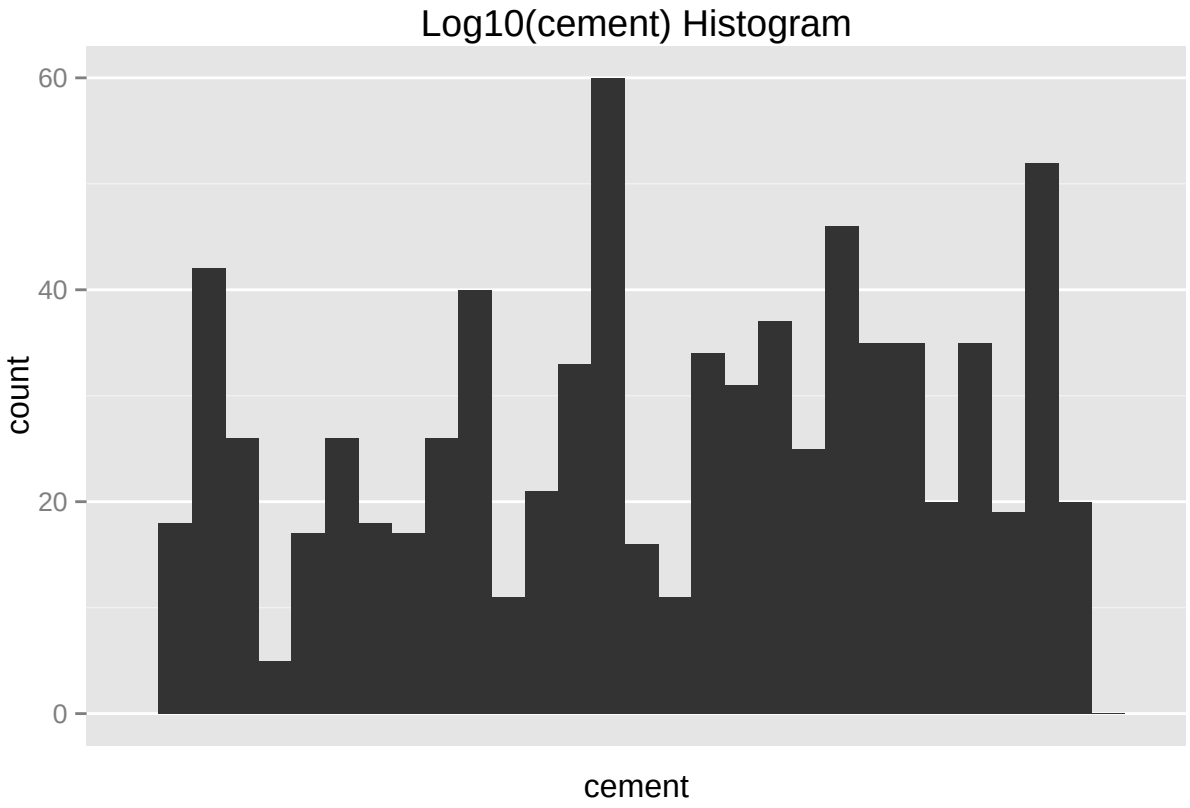


After the ratio, everything will be compared against cement, which will equal 1. I wanted to summarize the cement levels in the study before making ratios and flattening the cement weights.

The histogram showing the cement weights appears to be approximately normally distributed, however slightly right-skewed.

The wide range seems to say this mixtures are going to vary a great deal in total mixture weight. It seems like this might introduce other factors like the size and shape of the test sample. The amount of time the concrete mixture was allowed to set before each sample was poured. The amount of surface area that was exposed to air before it was poured. Factors like that might be introduced, which aren't accounted for in the data.

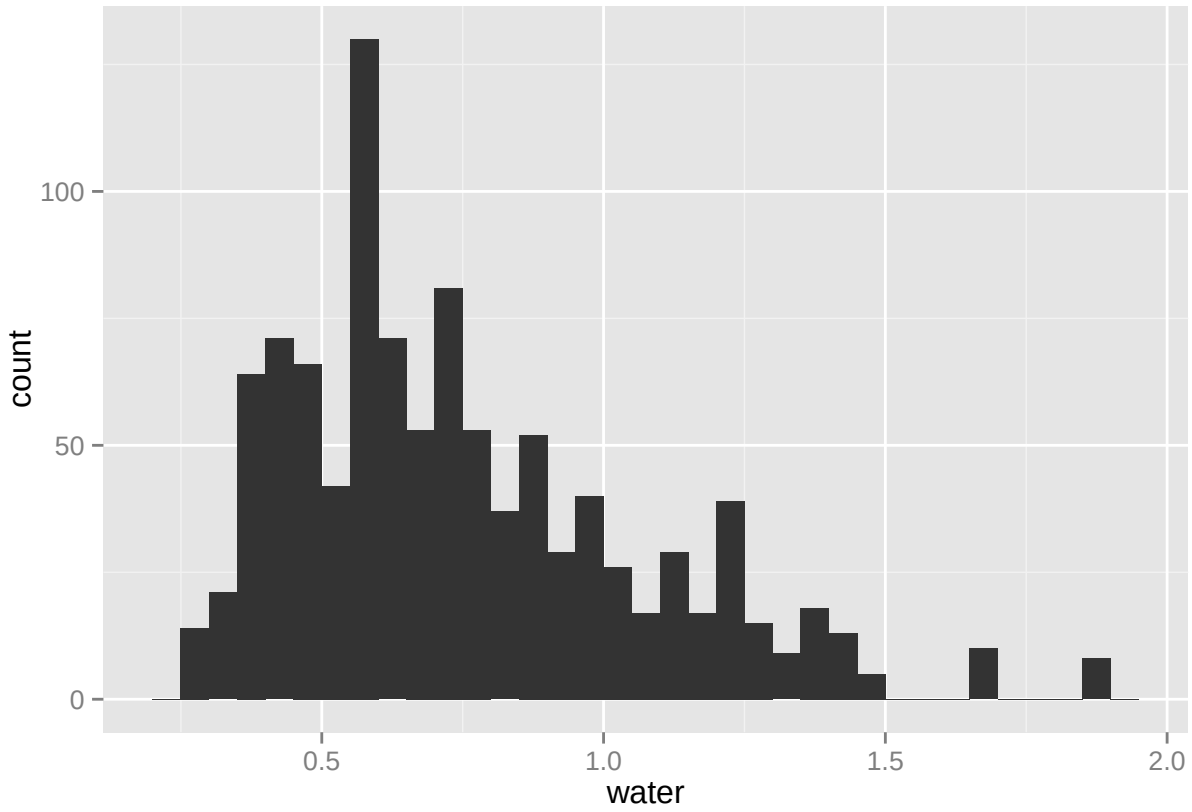
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



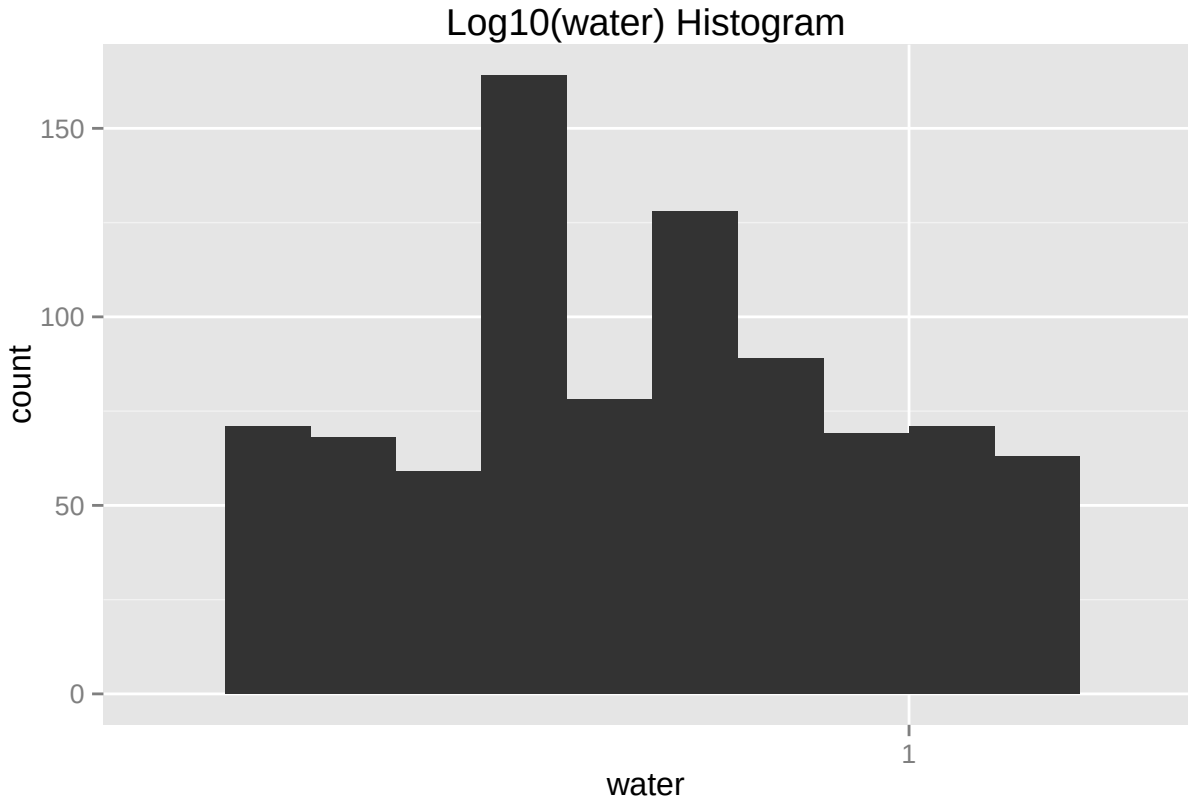
Adding log10 transformation to the cement plot and it appears to bring it closer to a normal distribution.

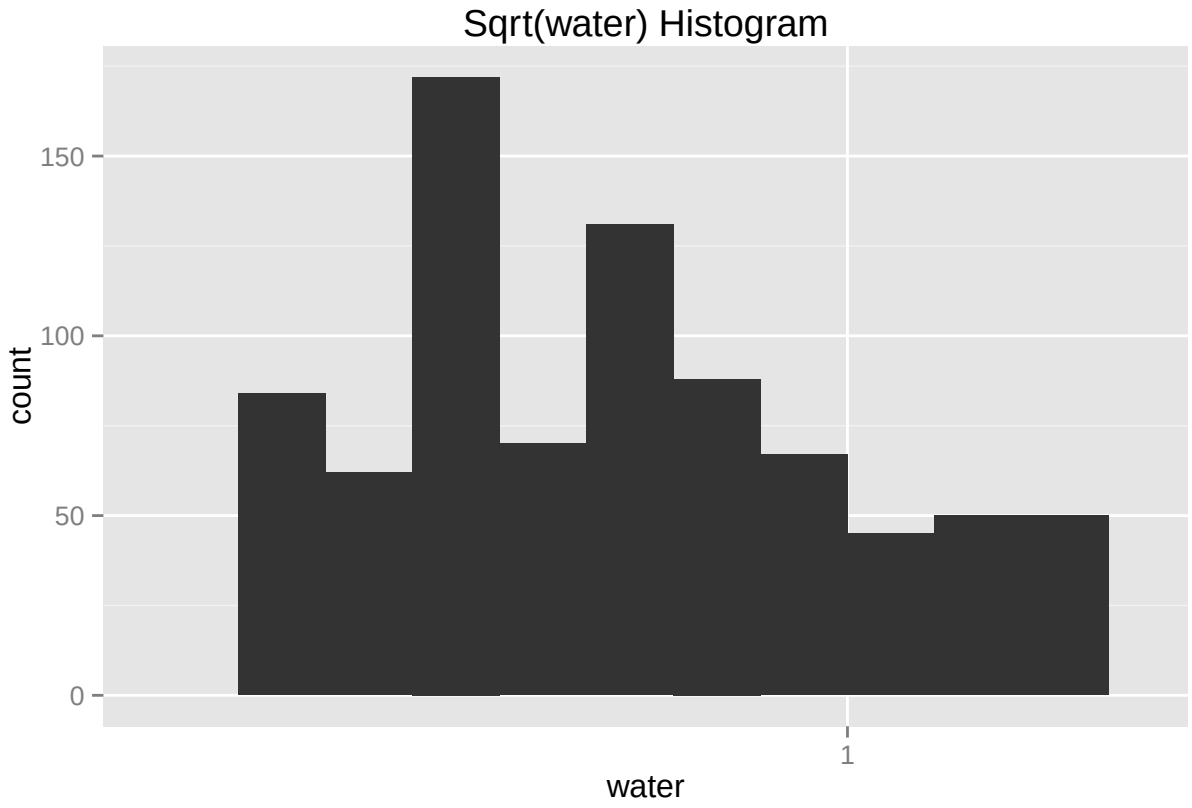
Histograms Showing the Makeup of the Observations in Relation to Cement Weight

In this section I addded several histograms to detail the counts of occourences of levels of each component in relation to the weight of cement in each observation of the study.

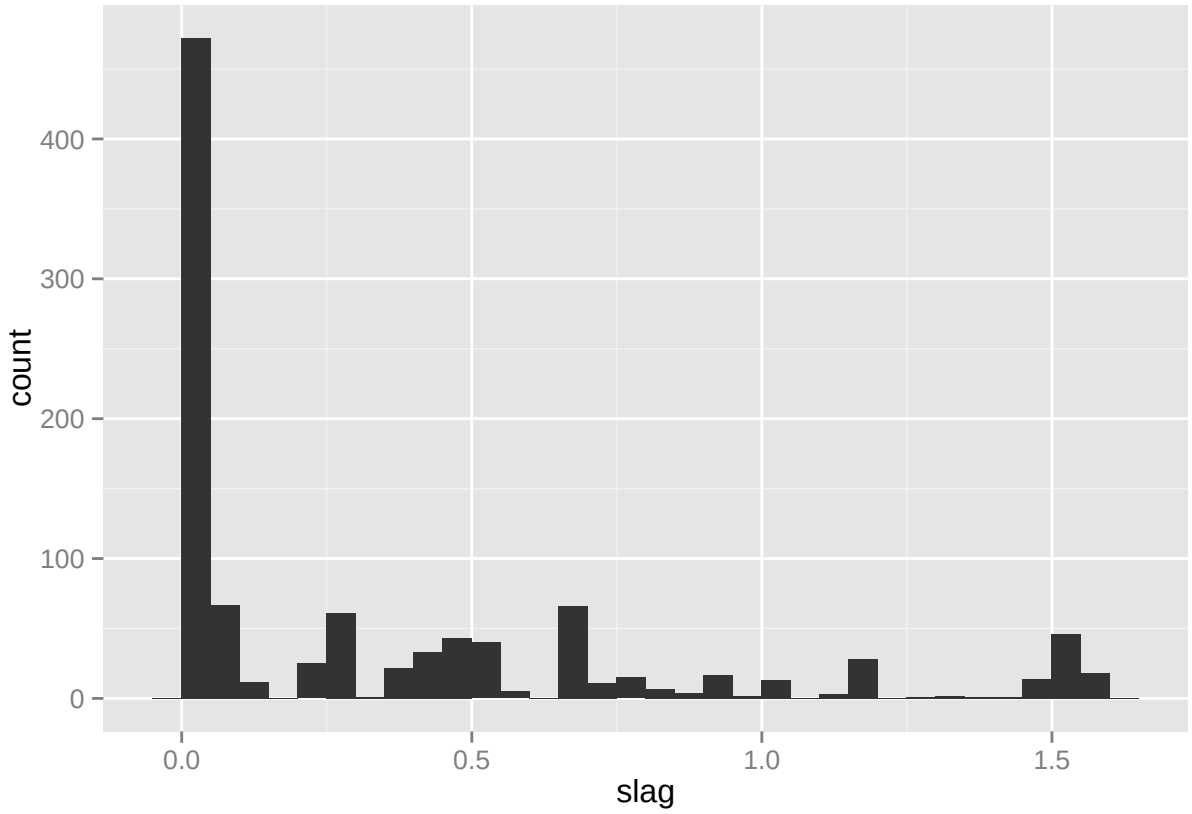


Water looks a little left skewed, and there are a few outliers. It could be improved, I think.

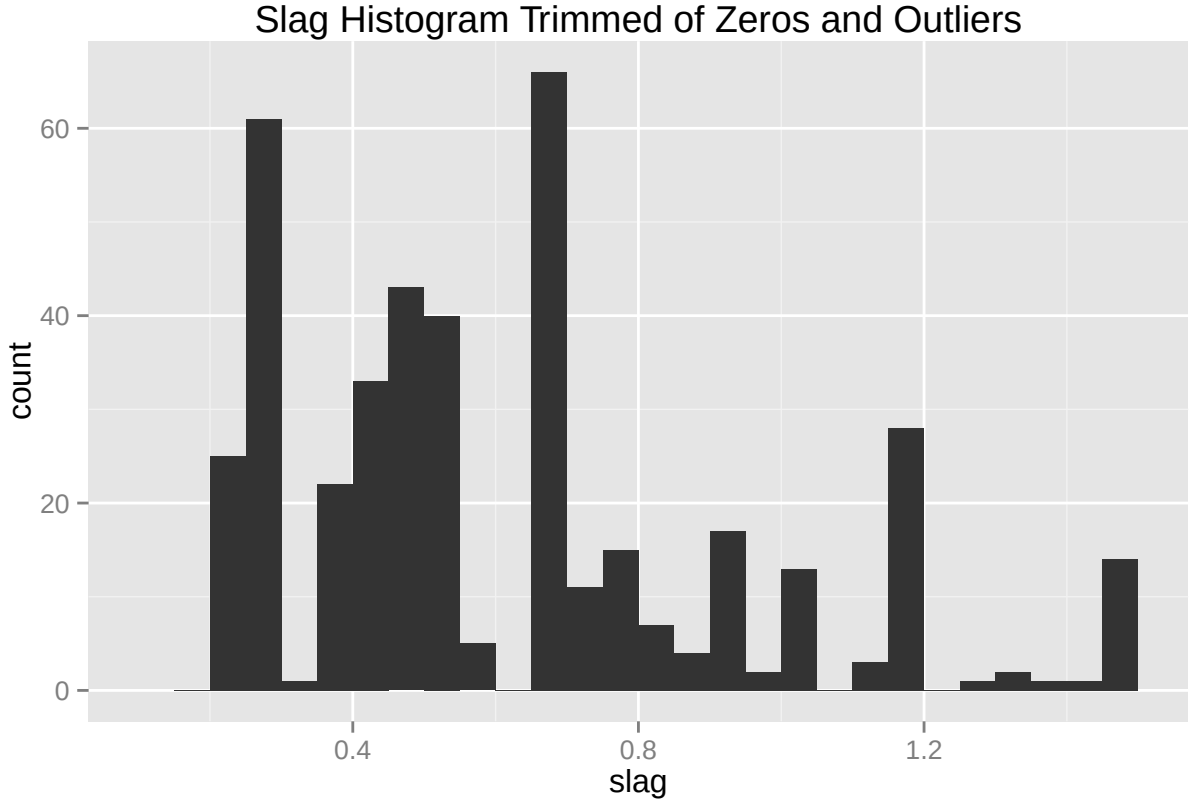


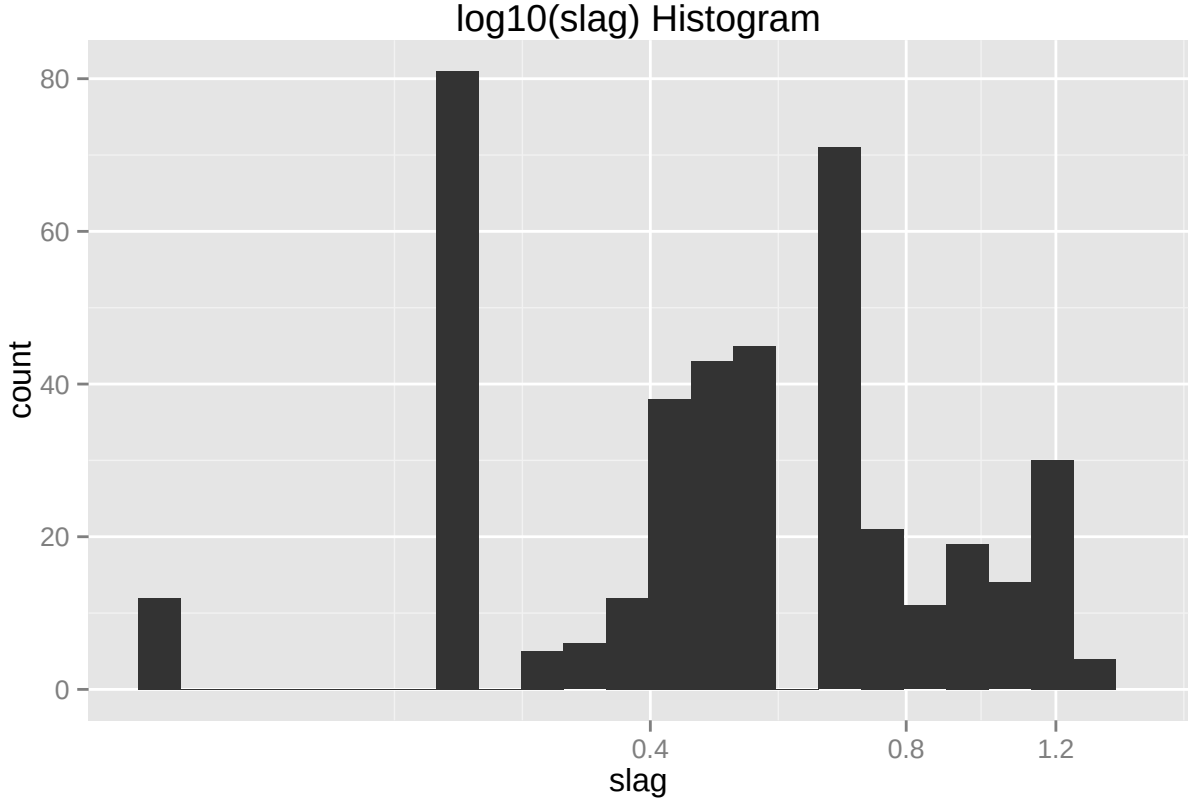


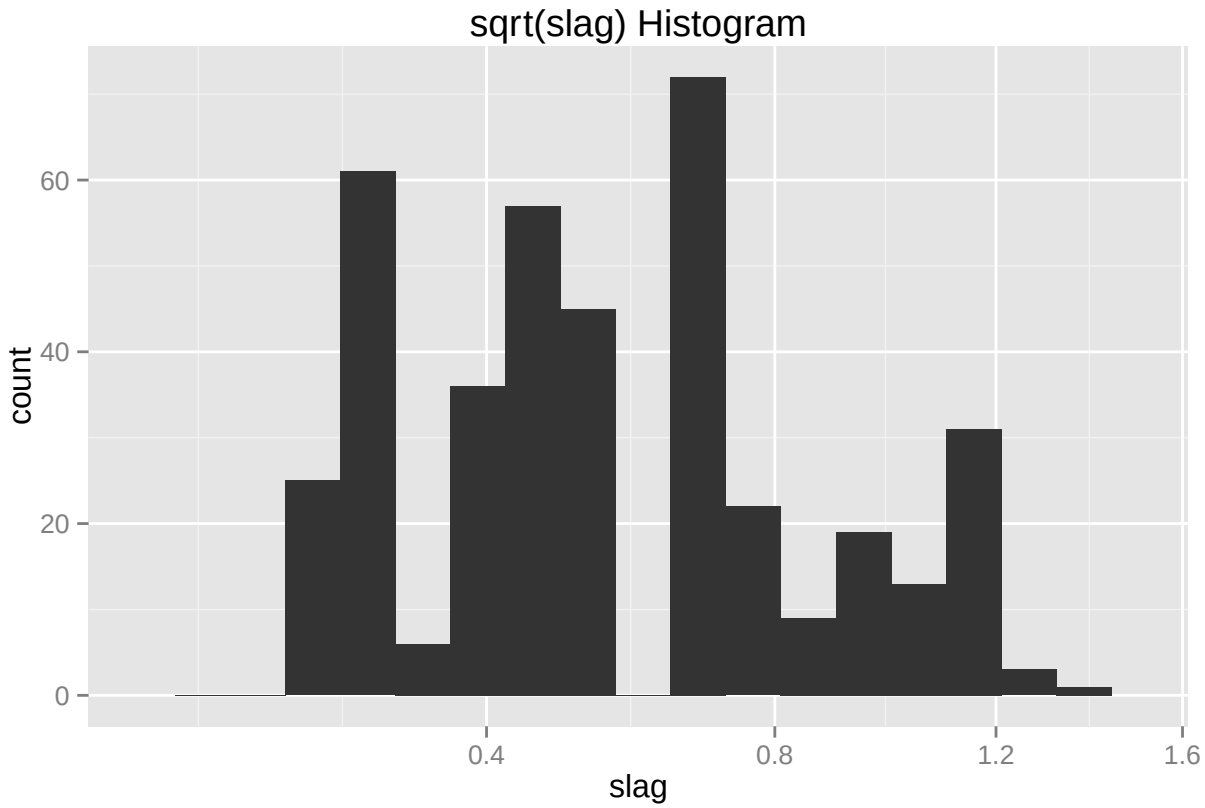
The log10 looks to put more of the weight in the chart in the center. The sqrt transformation leaves a lot of the distribution on the left side. I'm going to consider the log10 transformation when building the linear model.



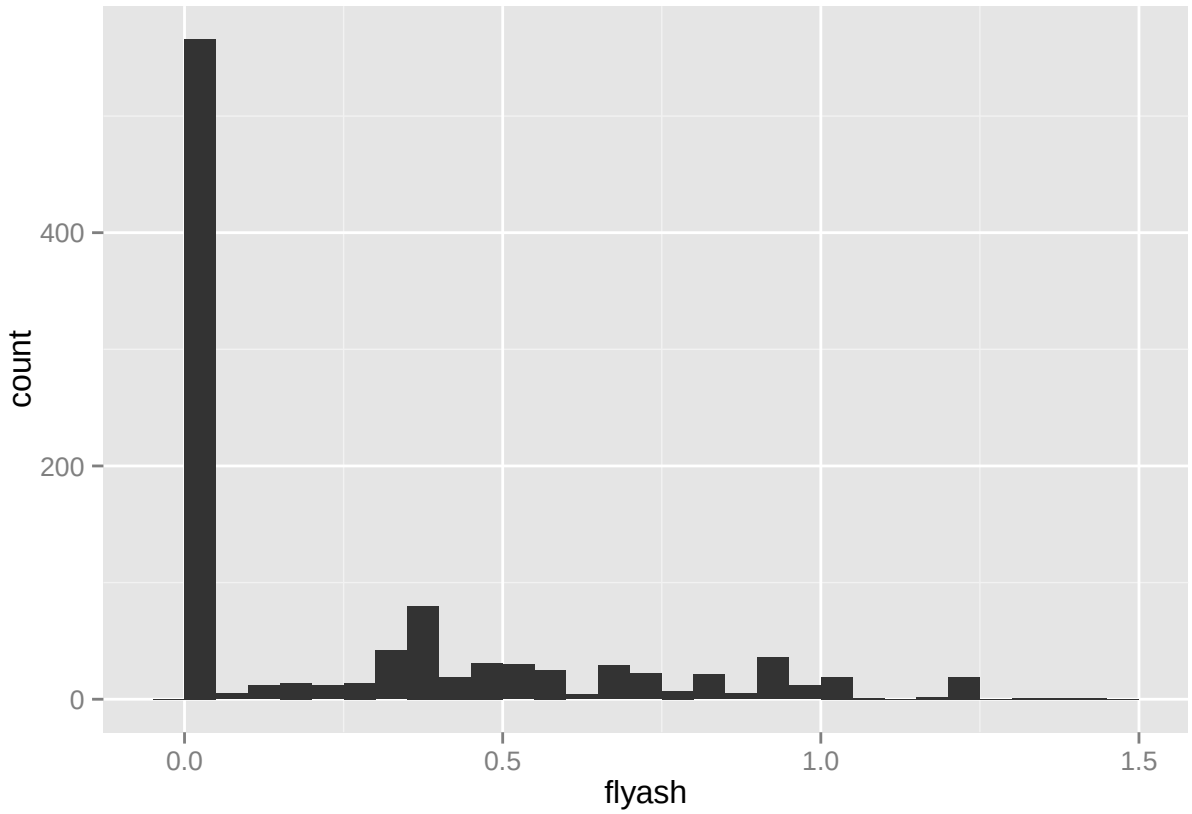
Slag is not always present, but it looks like many of the samples are in the 0.5 - 1.5kg per kg of cement range. Because of the large number of zero values, I'm not sure the non-zero ones on the low-end are outliers, though. It's just that a lot of samples don't have any slag.



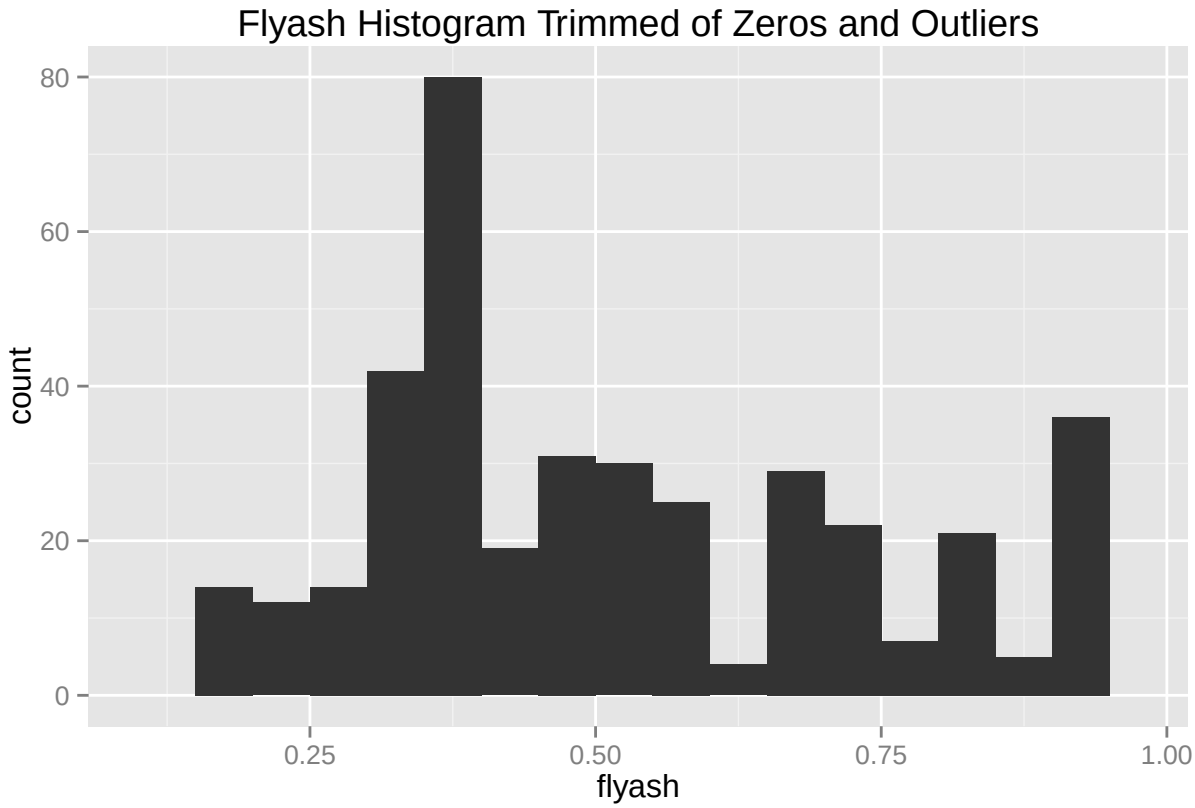


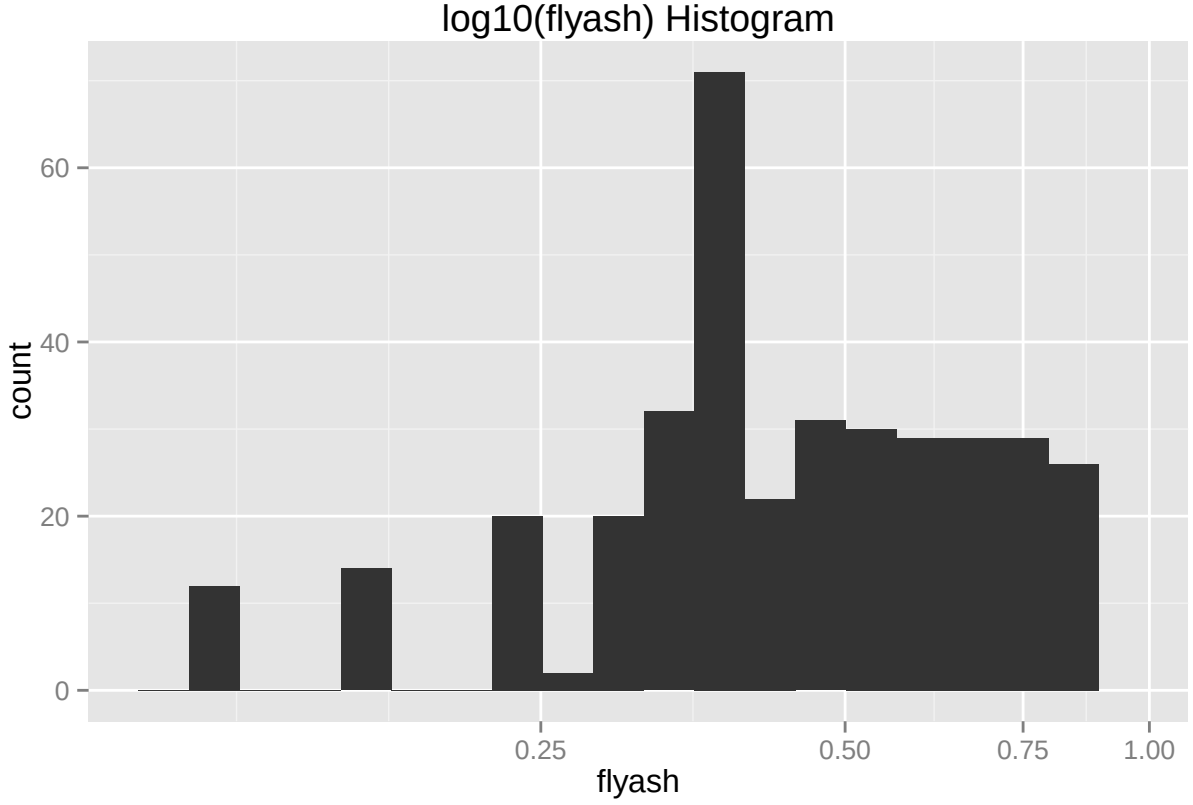


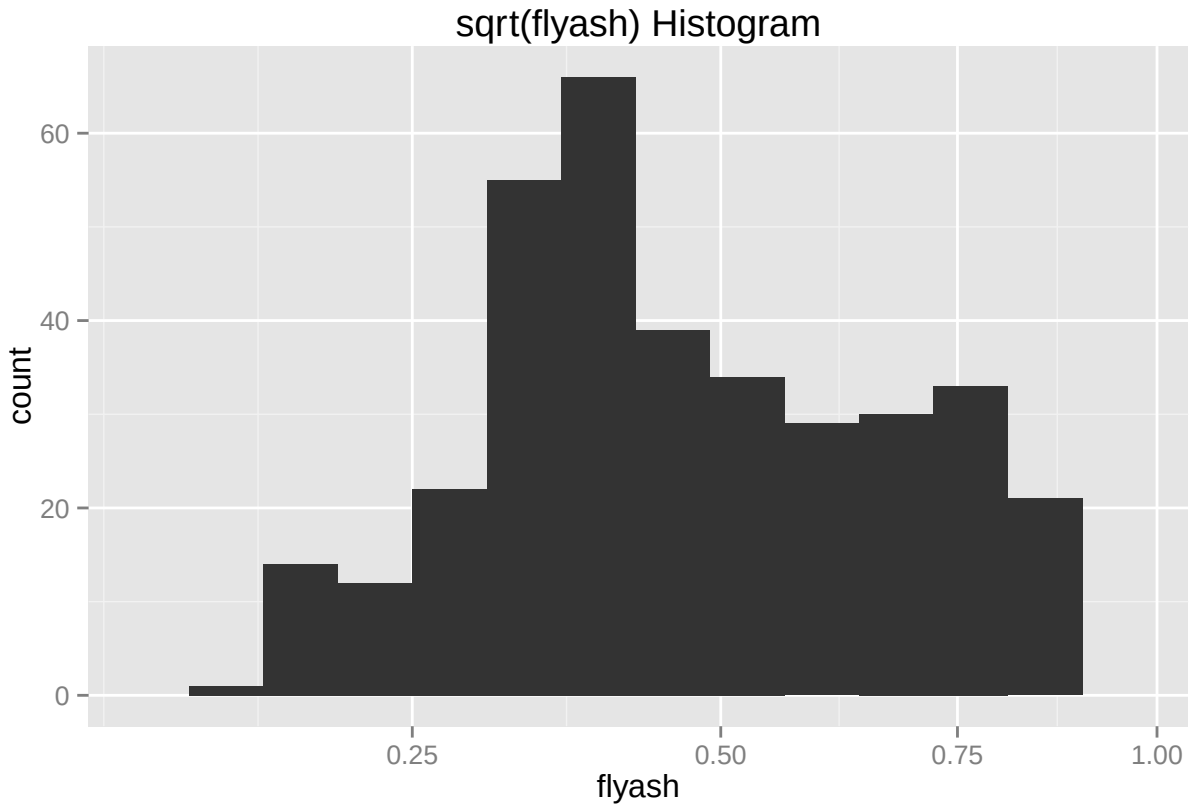
I trimmed the outliers and tried log10 and sqrt transformations above. Log10 created a left-skewed distribution, but sqrt(slag) looked more normally distributed, so I'm going to make a note of that. None of the distributions looked all that great without trimming a lot of the lower outliers to focus on the group above zero.



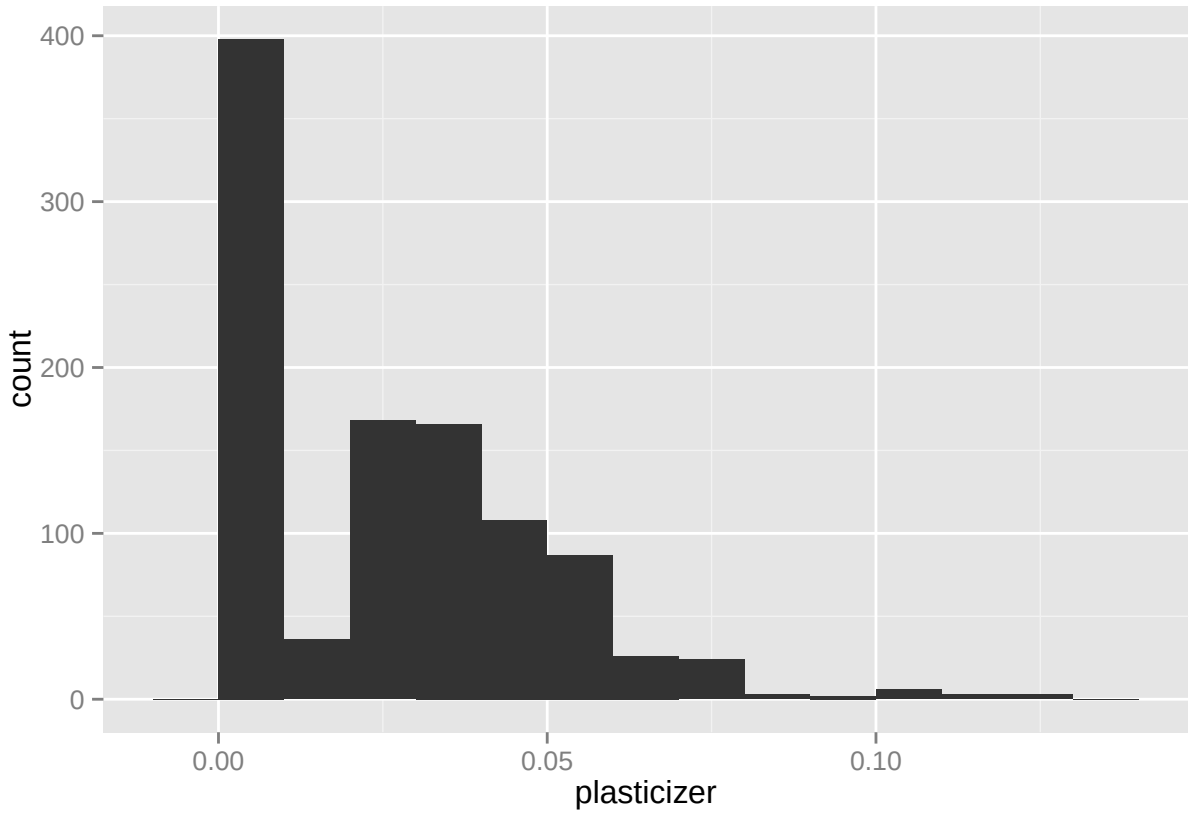
Flyash is not present in many of the samples, but like slag, many of the samples that do have it are in the 0.5 - 1kg per kg of cement range. With a lot of zeros the distribution looks similar to slag. I'll try a similar treatment and try to shape it up a bit.


Flyash Histogram Trimmed of Zeros and Outliers
count
flyash
0
20
40
60
80
0.25
0.50
0.75
1.00

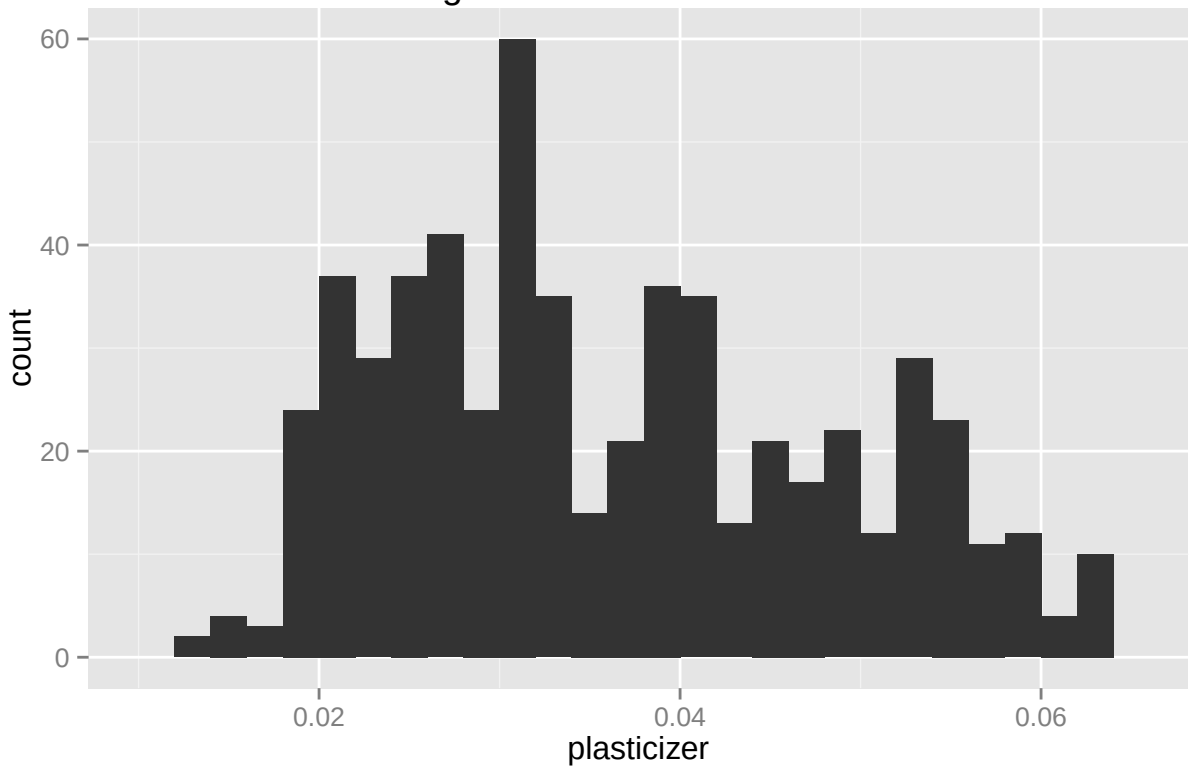

log10(flyash) Histogram
count
flyash
0
20
40
60
0.25
0.50
0.75
1.00

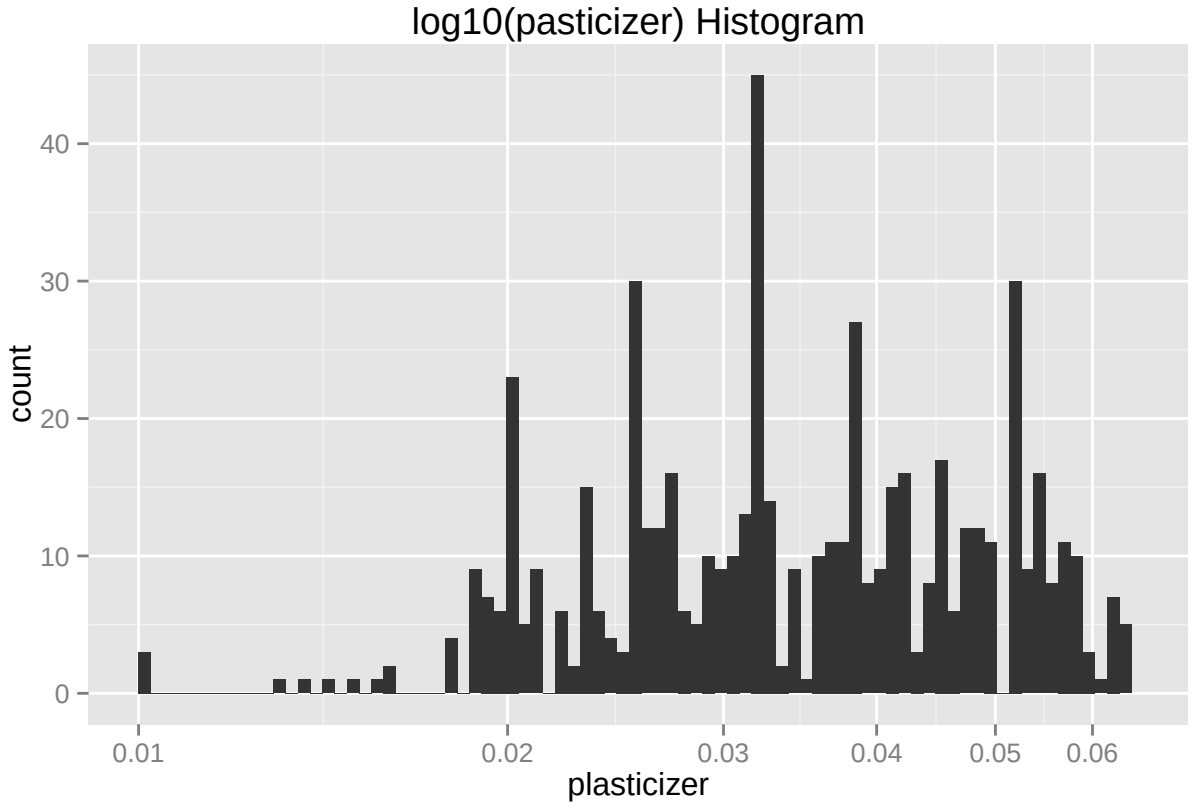


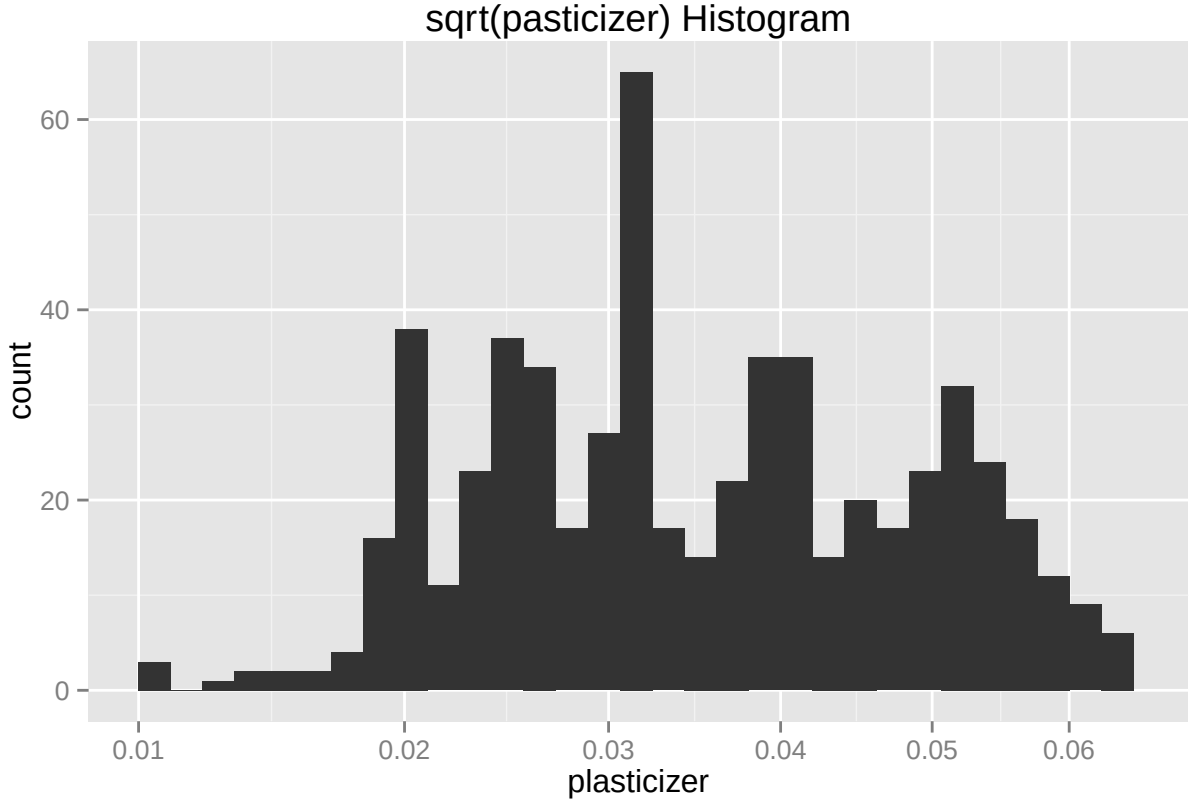
Both sqrt(flyash) and log10(flyash) transformations look better than the non-transformed value, with more normal distributions, however sqrt(flyash)$is more normally distributed, so I'm going to start with that in our linear model. In observations that contain flyash it looks like the mean is around 0.38kg per kg of cement.



Pasticizer Histogram with Zeros and Outliers Removed

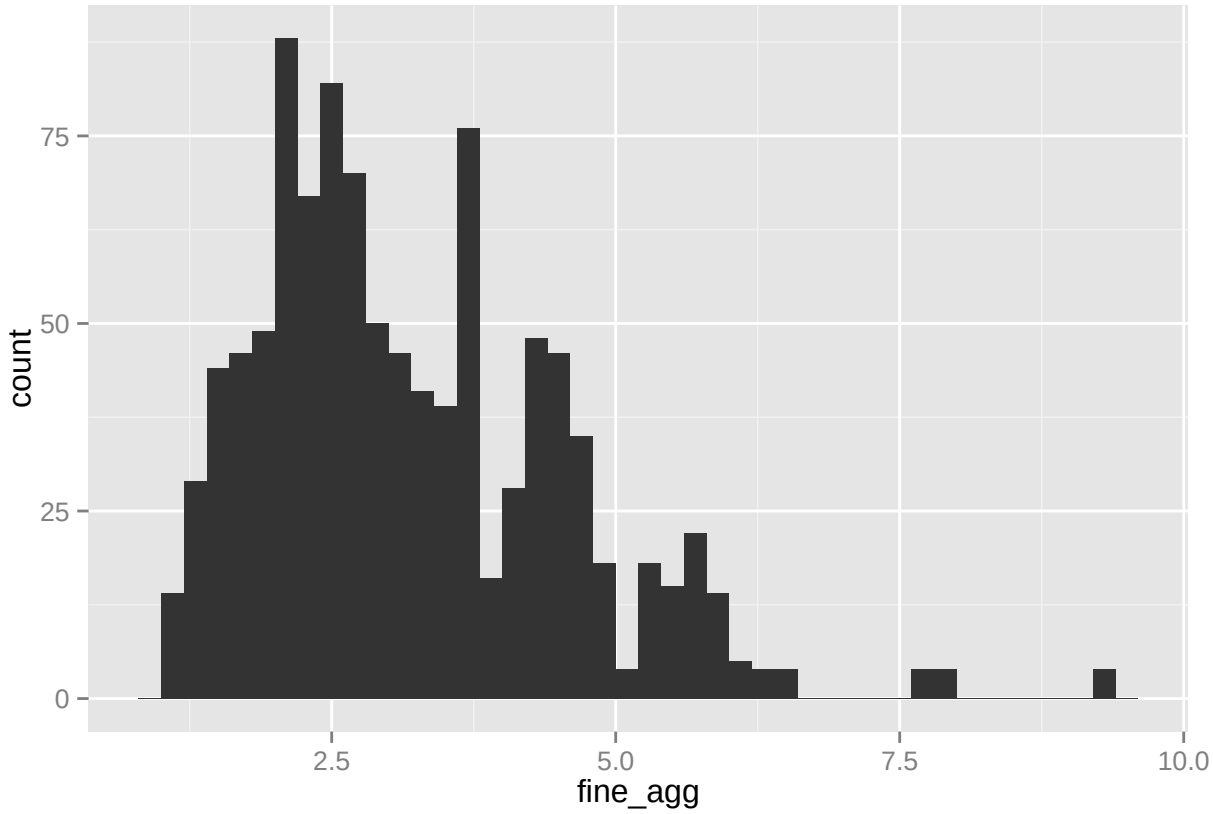


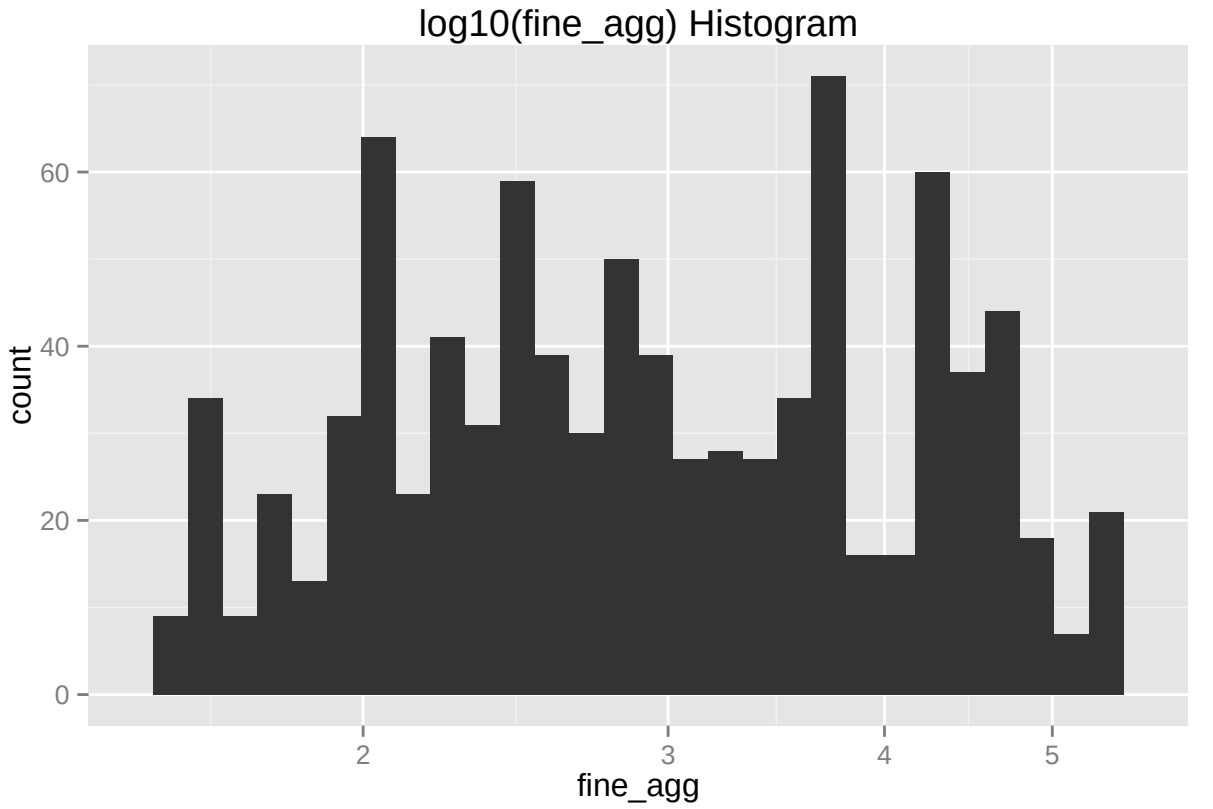


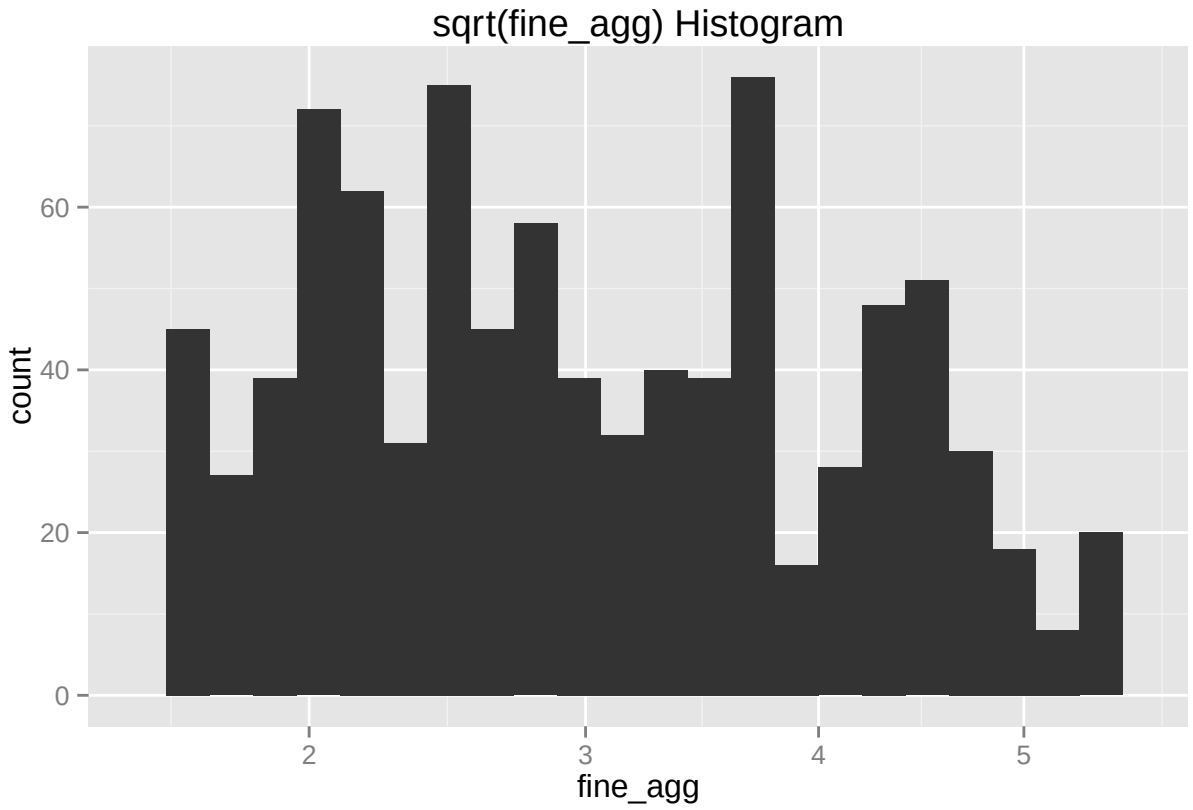


Many samples don't contain plasticizer, but there is a grouping of samples that do form a skewed distribution with a few outliers. The plot using sqrt(plasticizer) seems to show the most normalized distribution, so I'm going to try using that transformation. The mean level of plasticizer in samples that do have plasticizer

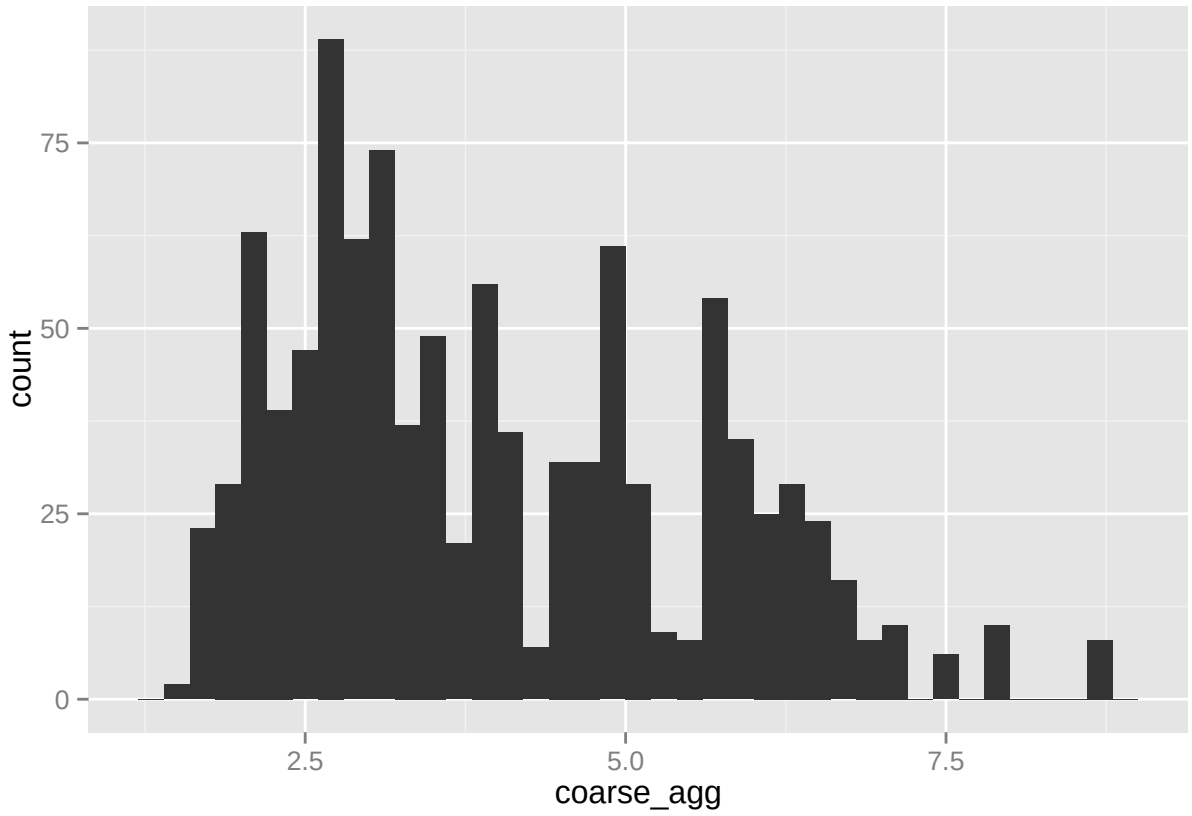
appears to be a little over 0.03kg per kg of cement.

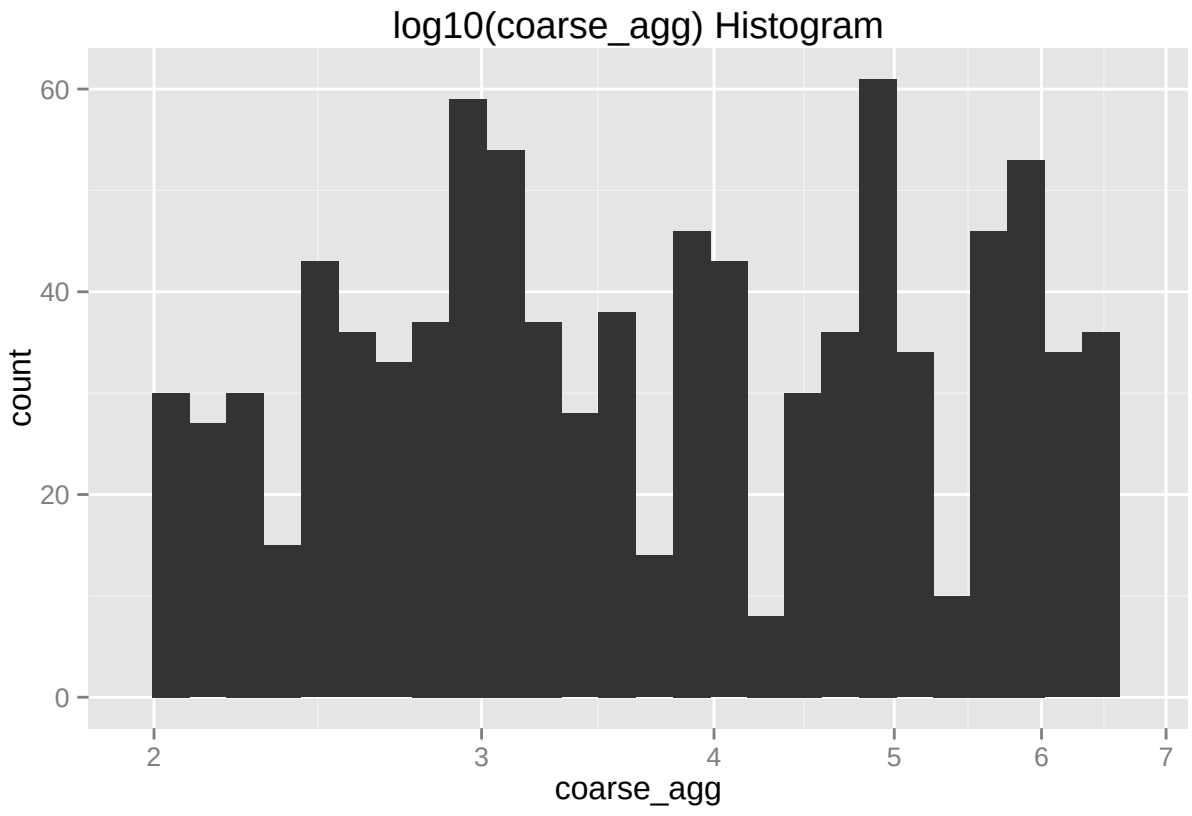


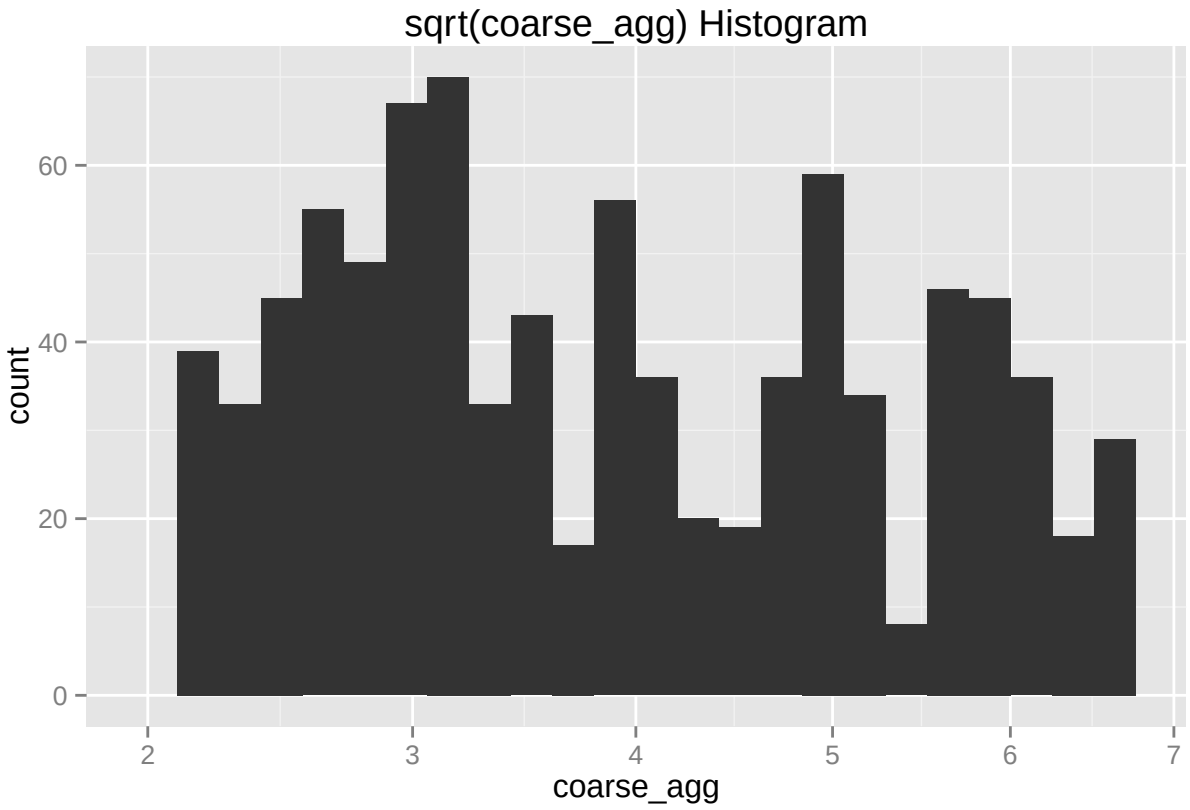




The amount of fine aggregate appears to be somewhat normally distributed. Unlike some of the additives fine aggregate is present in every sample. Log10(fine_agg) seems to bring together the values around the median a little better. The mean level of fine aggregate appears to be around 3kg per kg of cement.


75
50
25
0
count
2.5
5.0
7.5
coarse_agg


log10(coarse_agg) Histogram
60
40
20
0
count
2
3
4
5
6
7
coarse_agg



The amount of coarse aggregate also appears to be somewhat normally distributed. It's also present in every sample. The Log10(course_agg), like the fine aggregate seems to bring the distribution together a little bit. The mean ratio of coarse aggregate weight to cement is around 4:1.

Univariate Analysis

What is the structure of your dataset?

Number of observations: **1030** Number of input Attributes: **8** Number of output Variables: **1**

What is/are the main feature(s) of interest in your dataset?

The original dataset before transformation:

1. **Cement** kg in mixture – Input Variable
2. **Blast Furnace Slag** kg in mixture – Input Variable
3. **Fly Ash** kg in mixture – Input Variable
4. **Water** kg in mixture – Input Variable
5. **Superplasticizer** kg in mixture – Input Variable
6. **Coarse Aggregate** kg in mixture – Input Variable
7. **Fine Aggregate** kg in mixture – Input Variable
8. **Age** Day (1~365) – Input Variable

Output: **Concrete** MPa (Megapascals) – Output Variable

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

I think the relationships between the components will make a difference. The cement to other ingredients, for example, and the presence or absence of optional additives and the roles they play in improving the strength of a concrete mixture.

Did you create any new variables from existing variables in the dataset?

I created binary variables to denote the presence or absence of an optional additive, and I also added another variable to classify blends into groups based on the presence of different combinations of the optional additives.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

All of the optional additives had a lot of zero values in the observations. I classified observations into 8 different blends based on the presence or absense of each of the optional additives.

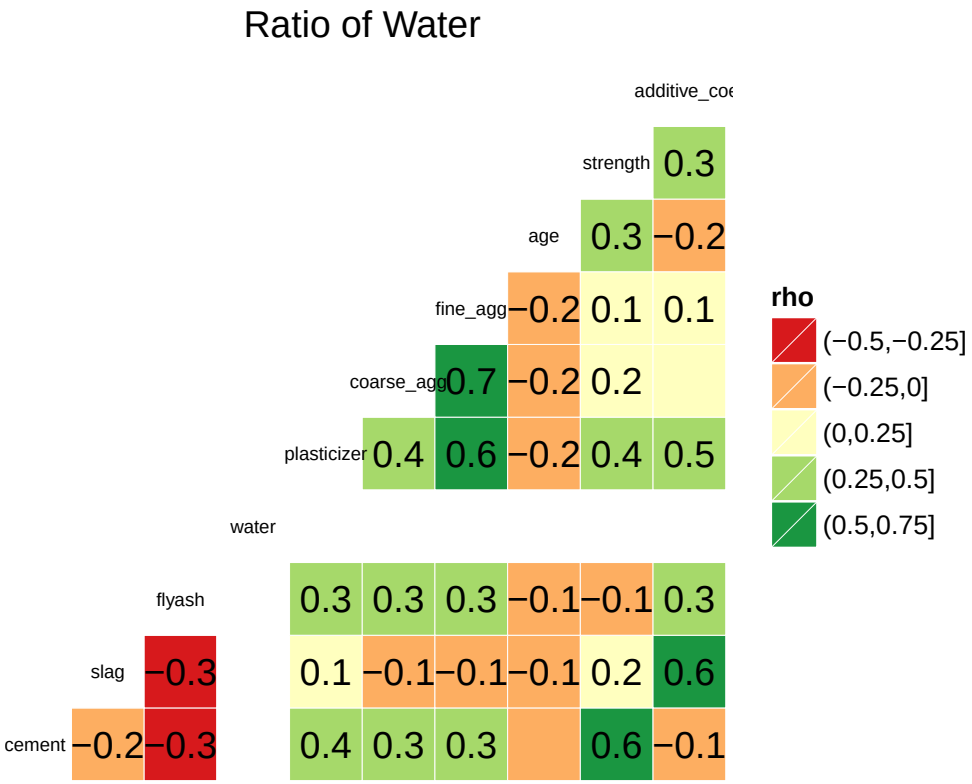
I like to bake bread, and it's all about the ratio of dry ingredients to wet ingredients, so I wanted to experement with ratios. I took ratios of the components since they were all measured in weights.

I first tried a ratio based on the amount of water in the mixture. I liked it, but I wasn't seeng really close-looking relationships using a scatterplot matrix, so I wanted to explore a little more. I also tried calculating each ingredient as a portion of the total weight of the entire mixture, but it didn't look very strong on the scatterplot matrix either. The one that showed the best looking correlations to me, was to base the ratio of each ingredient to the weight of the cement in the observation. So I converted the weitghts to Kg per Kg of cement.

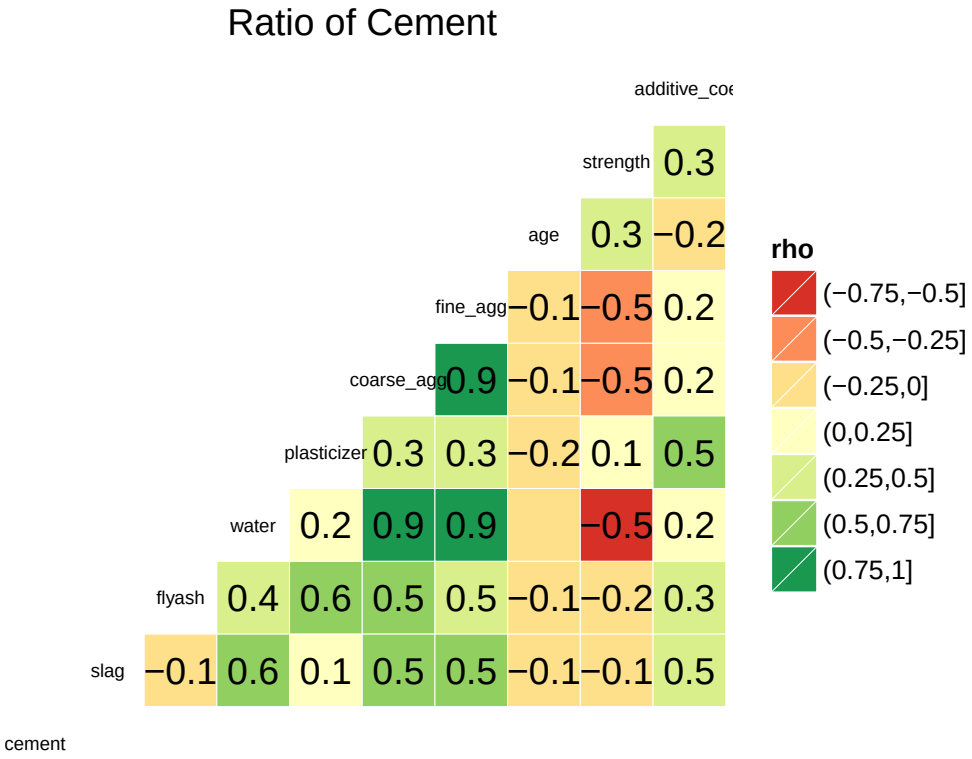
Bivariate Plots Section

Using the three ratios I wanted to look at the correlation matrices for each, and check the correlations of the seven variables that make up the concrete mixtures with the strength variable. Then choose a ratio that will give us the best results to use in a simple multiple linear regression model.

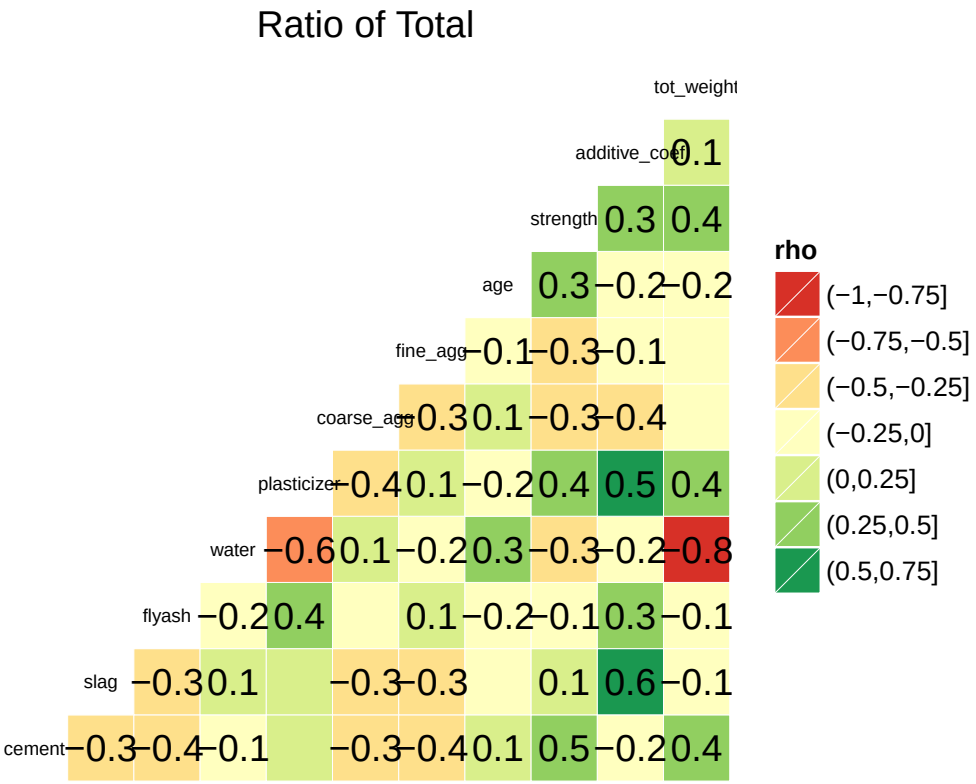
Correlation Matrix



In the above correlation matrix, I'm looking for correlations based on the ratio of each ingredient to the weight of the **water** in the mixture. Interestingly besides the ratio of cement and age, plasticizer shows the largest correlation with strength here.



In the above correlation matrix, I'm looking for correlations of strength based on the ratio of each ingredient to the weight of the **cement** in the mixture. This matrix shows several strong correlations with strength, including water and aggregates with slightly weaker correlation to additives.

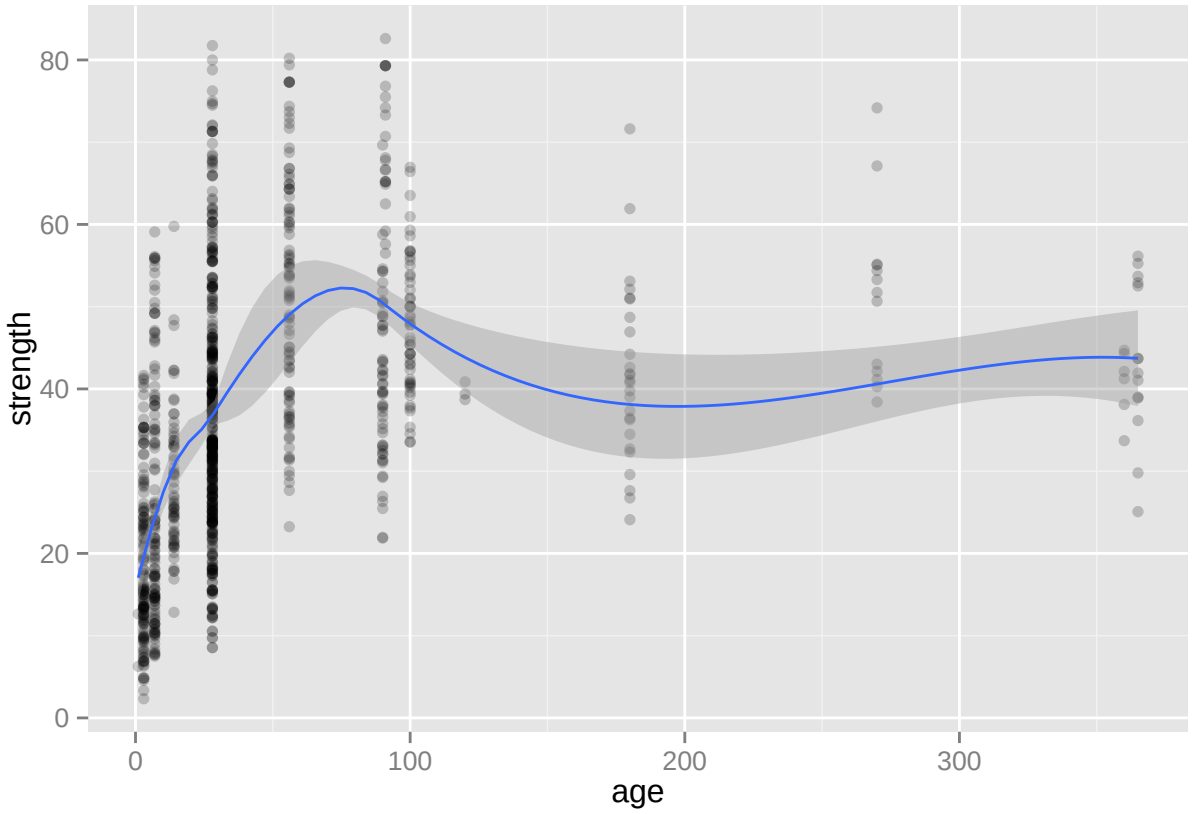


Summary of Total Weights of the Observations.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2195    2291    2349    2344    2390    2551
```

In the above correlation matrix, I'm looking for correlations of strength based on the ratio of each ingredient to the **total weight** of the all ingredients in the mixture. Interestingly this shows a stronger correlation of strength to coarse aggregate, pasticizer, and slag.

Age vs Strength

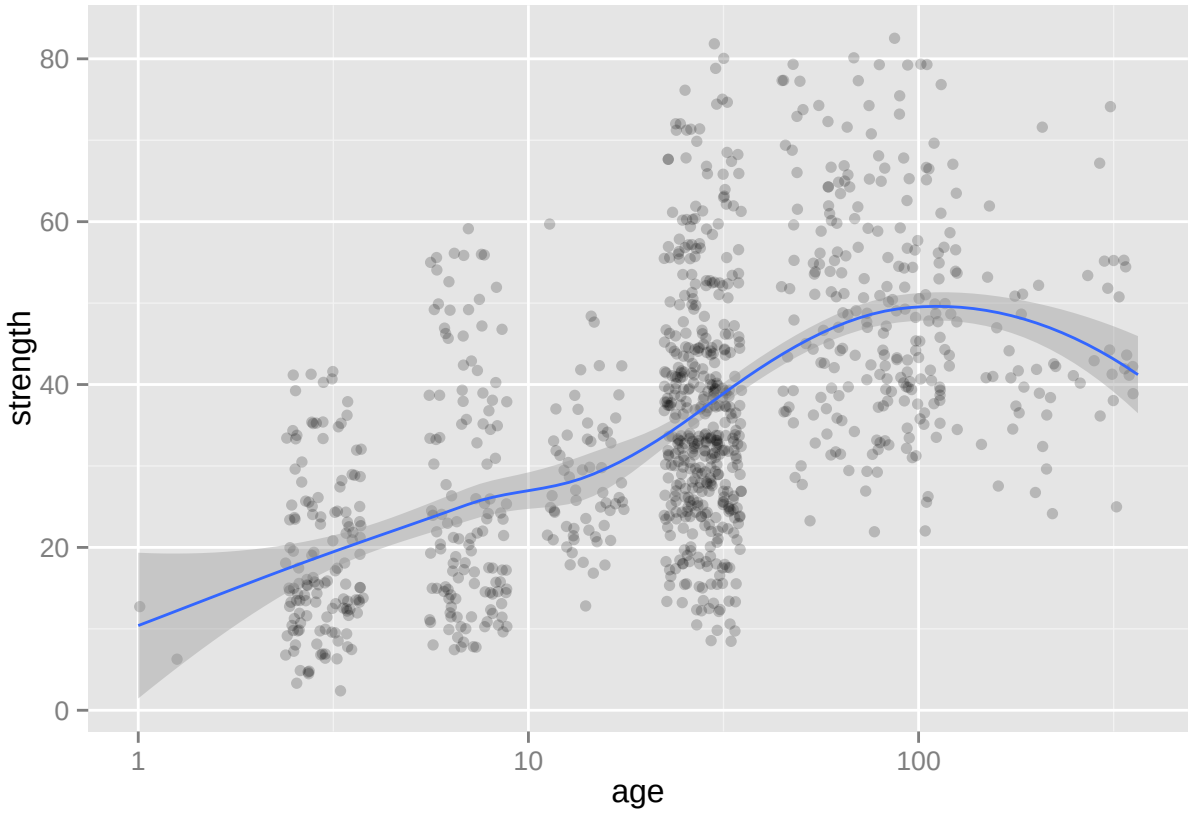



```
##
## Call:
## lm(formula = strength ~ age, data = c_ratios, rm.na = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.512 -11.290  -1.517   9.424  47.468
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.846595   0.606952   52.47   <2e-16 ***
## age          0.086973   0.007789   11.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.78 on 1028 degrees of freedom
## Multiple R-squared:  0.1082, Adjusted R-squared:  0.1073
## F-statistic: 124.7 on 1 and 1028 DF,  p-value: < 2.2e-16
```

Interestingly, as the histogram tests showed, it looks like there is some kind of logarithmic relationship between age and strength. We can probably straighten this some by using transforms like we did with the univariate plots.

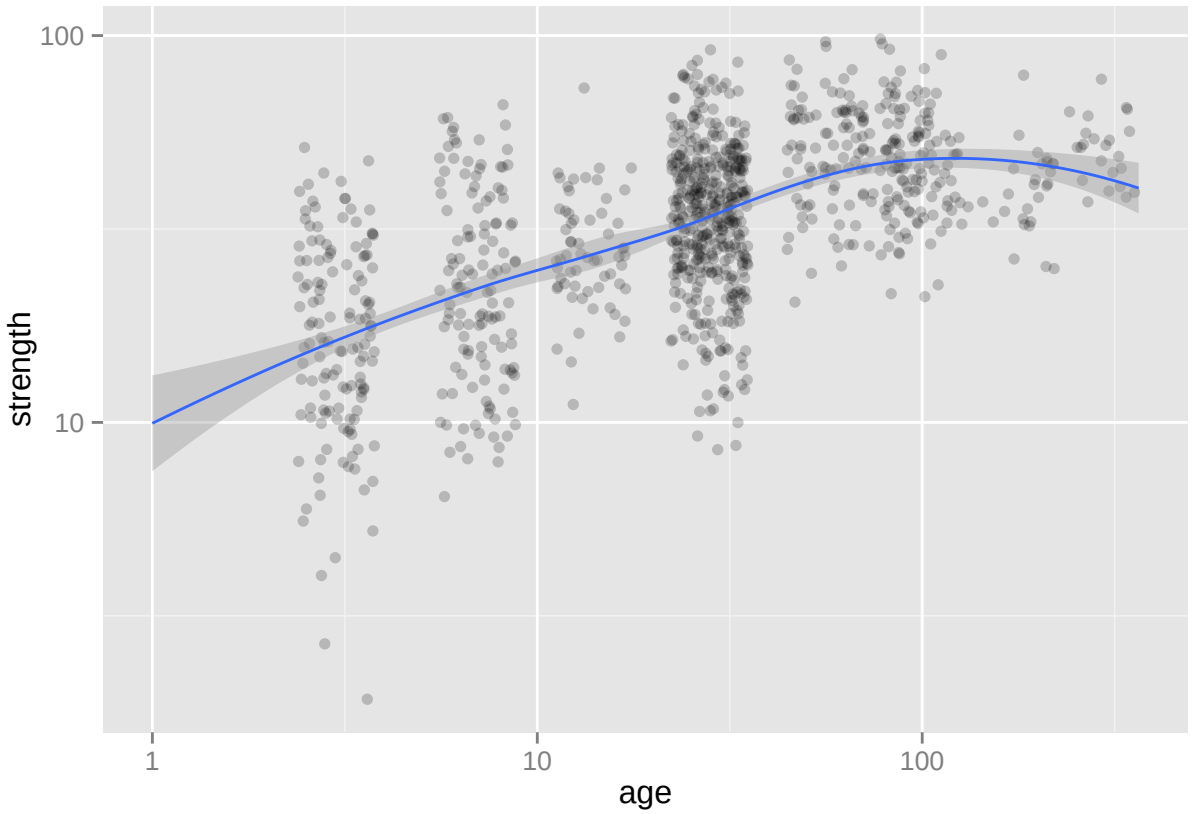
Since age has have nothing to do with the mixture, I decided to check it out on the linear model with only strength to see what happens.

With no transformations we show an R^2 of 0.11. Based on the curve in that plot, and the transformations on the age and strength histograms, I think we can do better.




```
##
## Call:
## lm(formula = log10(strength) ~ age, data = c_ratios, rm.na = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07467 -0.12233  0.03553  0.15583  0.43801
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.4381279  0.0086536  166.19   <2e-16 ***
## age         0.0012983  0.0001111   11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.225 on 1028 degrees of freedom
## Multiple R-squared:  0.1173, Adjusted R-squared:  0.1165
## F-statistic: 136.7 on 1 and 1028 DF,  p-value: < 2.2e-16
```

Transforming age using log10 straightens out the plot a little bit. Trimming the outliers clears up the overplotting from the first chart. Also the linear model gives a slightly better R^2 score.



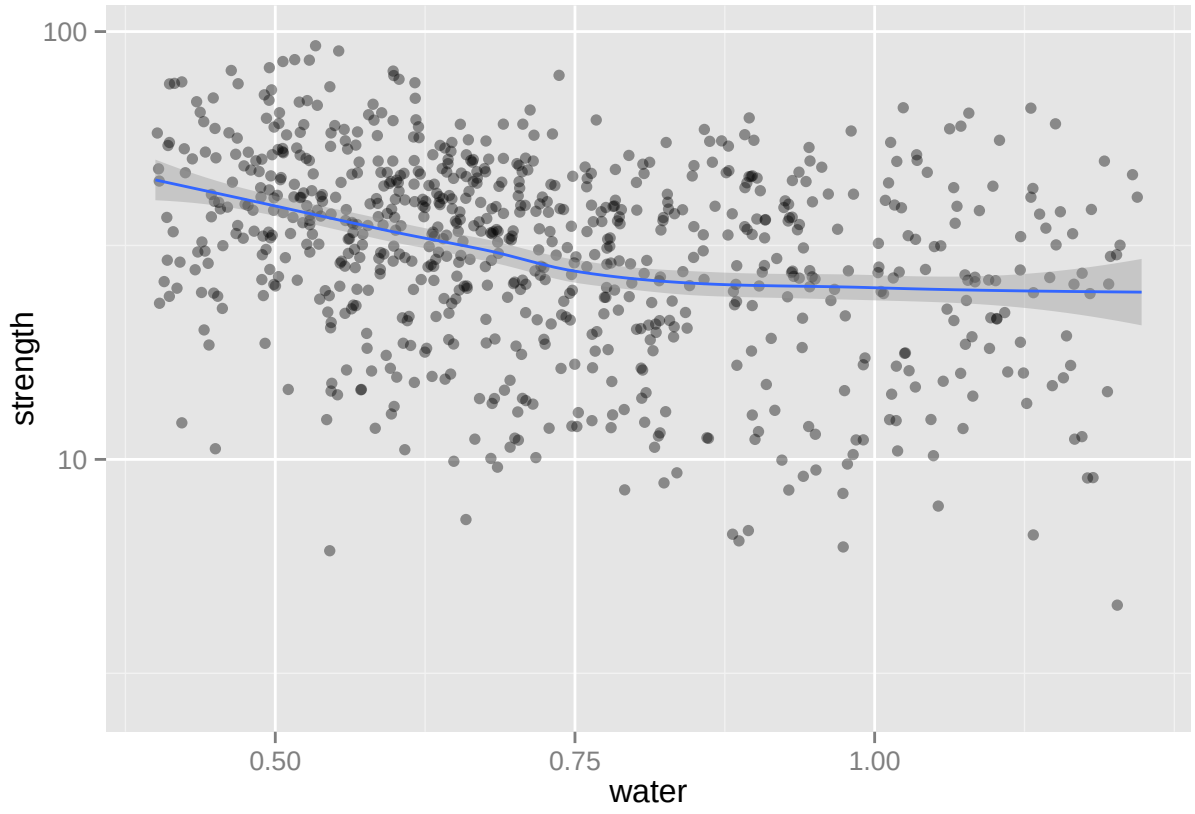

```
##
## Call:
## lm(formula = log10(strength) ~ log10(age), data = c_ratios, rm.na = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88289 -0.11751  0.00174  0.13180  0.41994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.11886    0.01703   65.69   <2e-16 ***
## log10(age)   0.27537    0.01160   23.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1925 on 1028 degrees of freedom
## Multiple R-squared:  0.3542, Adjusted R-squared:  0.3536
## F-statistic: 563.8 on 1 and 1028 DF,  p-value: < 2.2e-16
```

Transforming both age and strength using log10 straightens out the plot a little bit more. Also the linear model gives a much better R^2 score of 0.35 when transforming both axes with log10().

Cement ratio plots

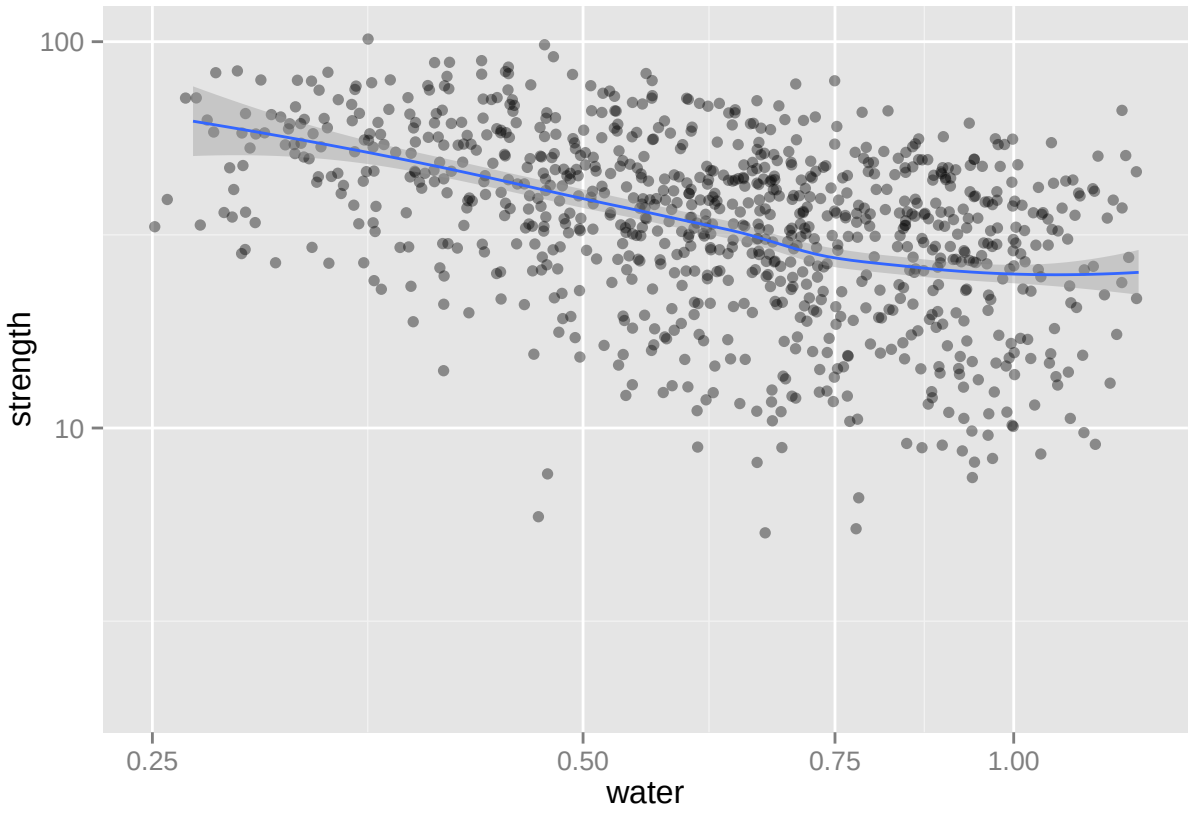
For the following plots I chose to use the cement ratio since it shows some good strong correlations on water and aggregates, and I thing concrete is usually mixed per the amount of cement.

Hydration vs Strength



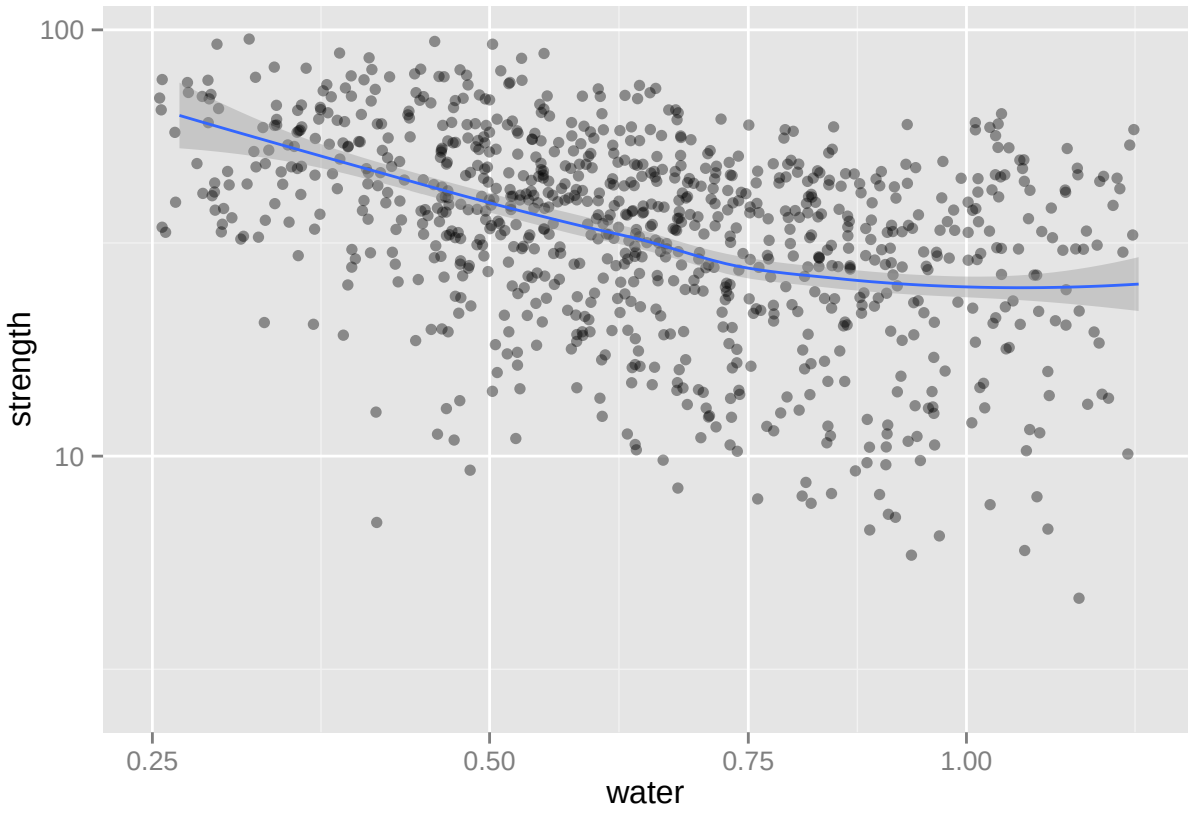

```
##
## Call:
## lm(formula = log10(strength) ~ water, data = c_ratios, rm.na = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80004 -0.11950  0.03677  0.15679  0.40594
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.77932    0.01678  106.06   <2e-16 ***
## water       -0.37675    0.02067  -18.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2082 on 1028 degrees of freedom
## Multiple R-squared:  0.2442, Adjusted R-squared:  0.2434
## F-statistic: 332.1 on 1 and 1028 DF,  p-value: < 2.2e-16
```

First, I wanted to check the linearity of the relationship between log10(strength), which plots nicely with age and water. Without transformation these produce a distinctly curved relationship, so I want to check a couple more transformations to compare. The linear model shows an R^2 of 0.24 with no transformation on water.




```
##
## Call:
## lm(formula = log10(strength) ~ log10(water), data = c_ratios,
##     rm.na = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82823 -0.11653  0.04191  0.15246  0.39957
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.385110   0.008667  159.81   <2e-16 ***
## log10(water) -0.691867   0.035998  -19.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2054 on 1028 degrees of freedom
## Multiple R-squared:  0.2643, Adjusted R-squared:  0.2636
## F-statistic: 369.4 on 1 and 1028 DF,  p-value: < 2.2e-16
```

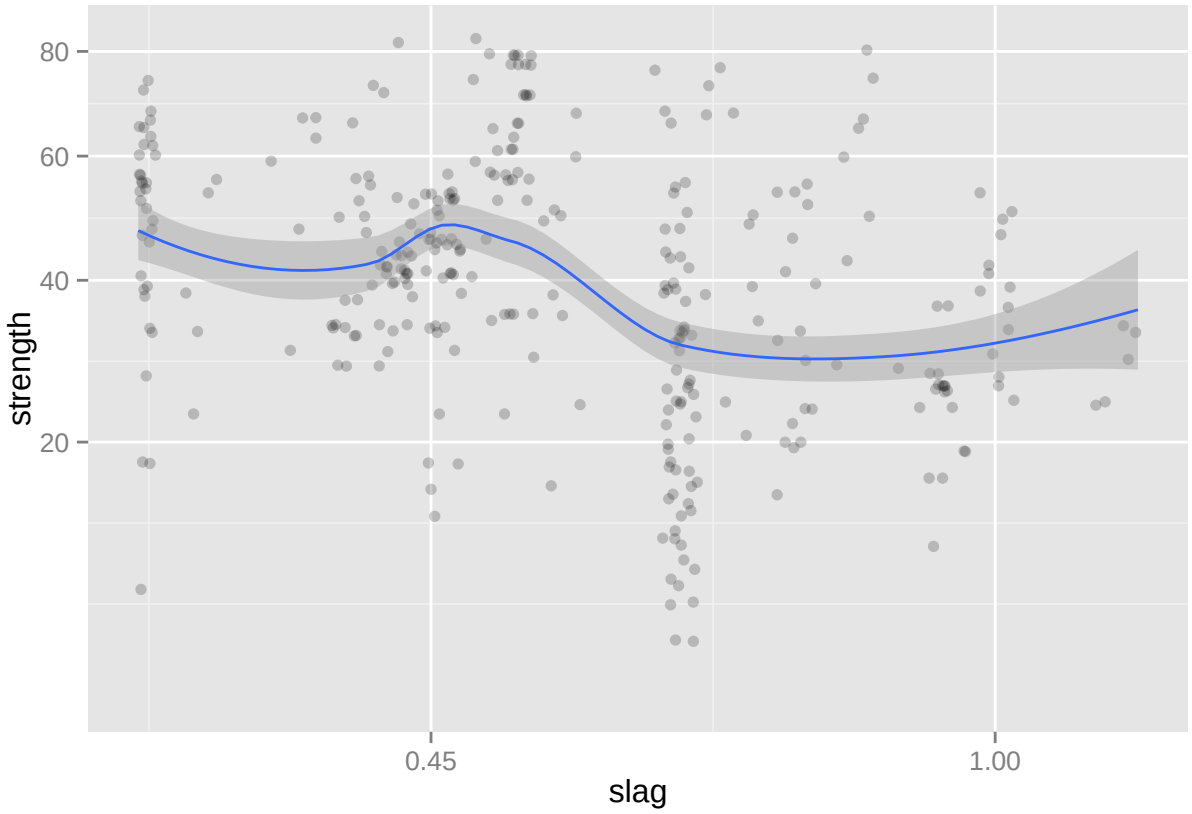
Adding log10 to the water axis straightens out the plot and improves the R^2 score of the linear model slightly to 0.26, so I think water may benefit from this transformation.




```
##
## Call:
## lm(formula = log10(strength) ~ sqrt(water), data = c_ratios,
##     rm.na = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80563 -0.11768  0.03527  0.15277  0.40471
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.08502    0.03180   65.56   <2e-16 ***
## sqrt(water) -0.69364    0.03677  -18.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2064 on 1028 degrees of freedom
## Multiple R-squared:  0.2572, Adjusted R-squared:  0.2565
## F-statistic: 355.9 on 1 and 1028 DF,  p-value: < 2.2e-16
```

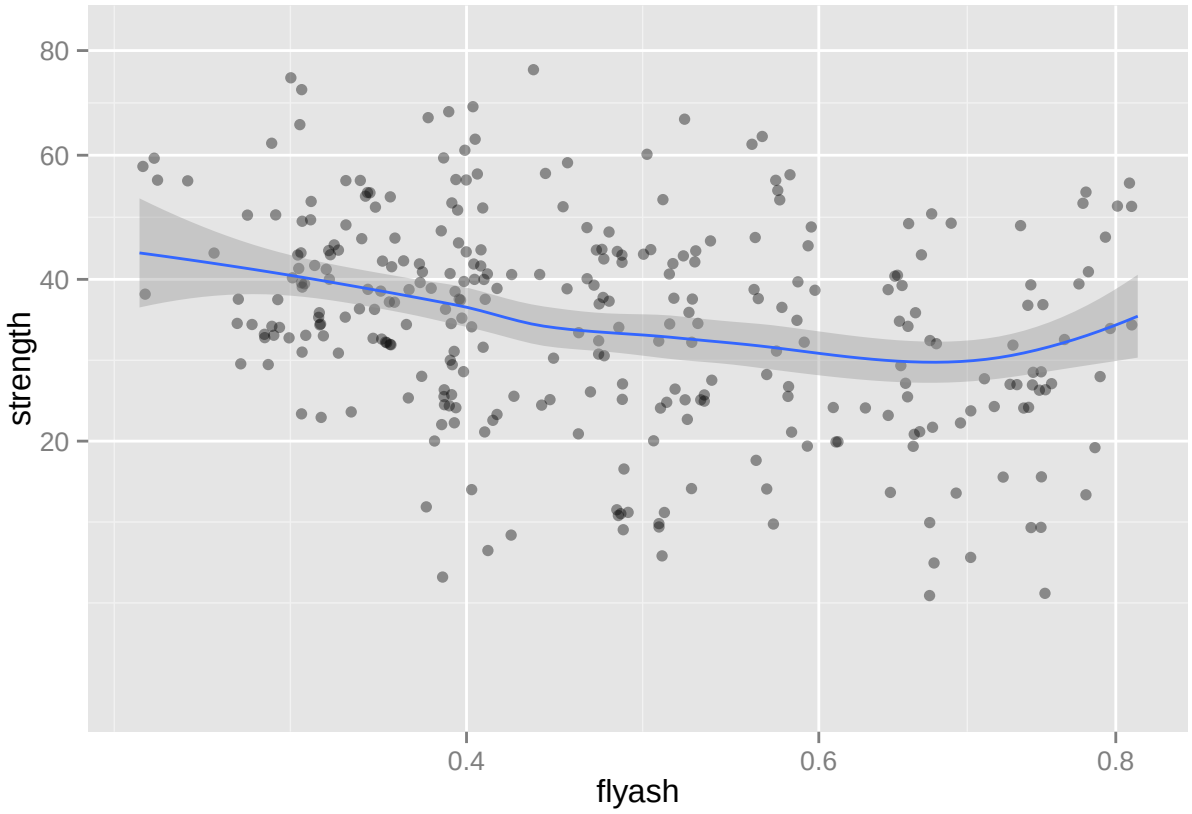
Transforming water with sqrt also seems to straighten out the plot somewhat, when compared with using water with no transformation, however it produces a more non-linear relationship than log10(water) does. Also the R^2 score is slightly lower when using sqrt(water) at 0.25, meaning the linear model doesn't have quite as tight of a correlation as it does when using log10(water). It's close, but I'm going to stick with log10(water) with a 0.26 R^2 for the linear model.

Slag vs Strength



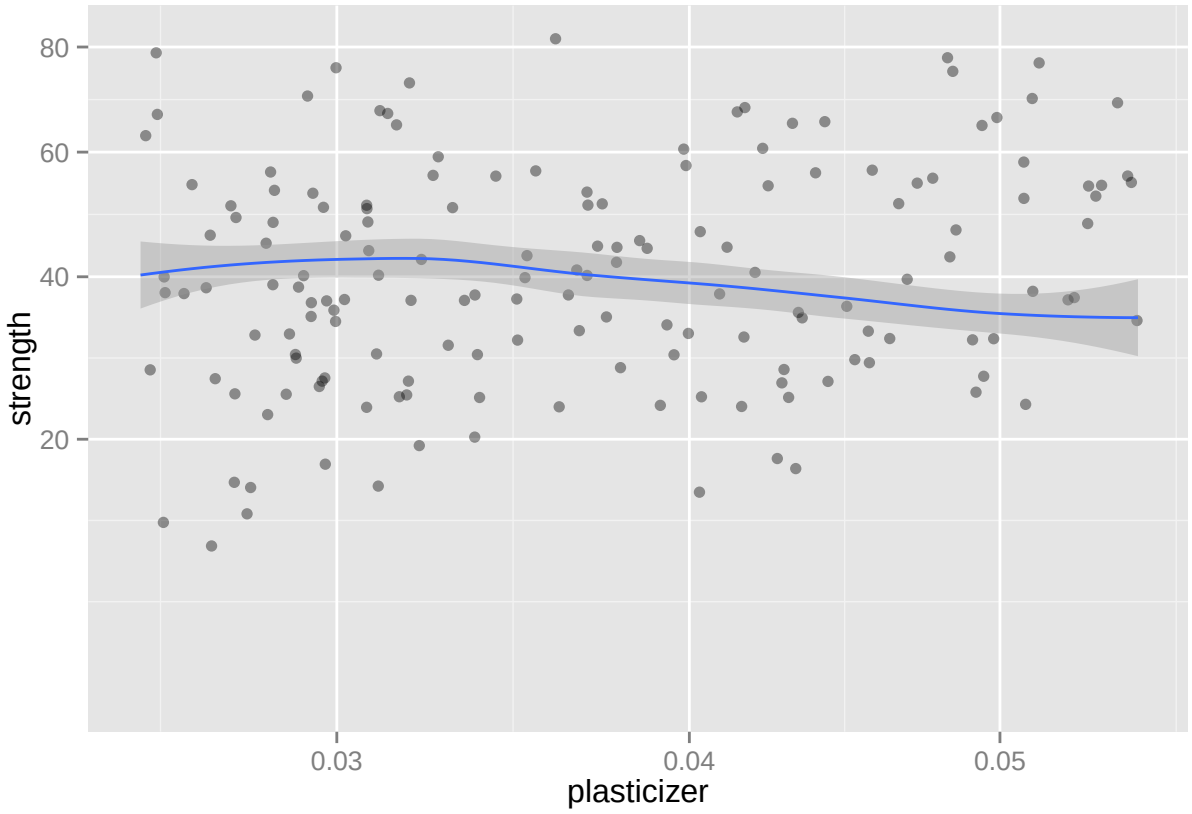
This plot shows the relationship of slag vs strength. It has already been transformed based on the tests in the Univariate Plots section. Interestingly it shows that there seems to be a 'sweet spot' for slag around 0.46kg per kg of cement. The relationship is not quite linear, but looks straighter than when the variables aren't transformed. This model would probably be improved if we could introduce a function that more closely matched this curve for slag.

Flyash vs Strength



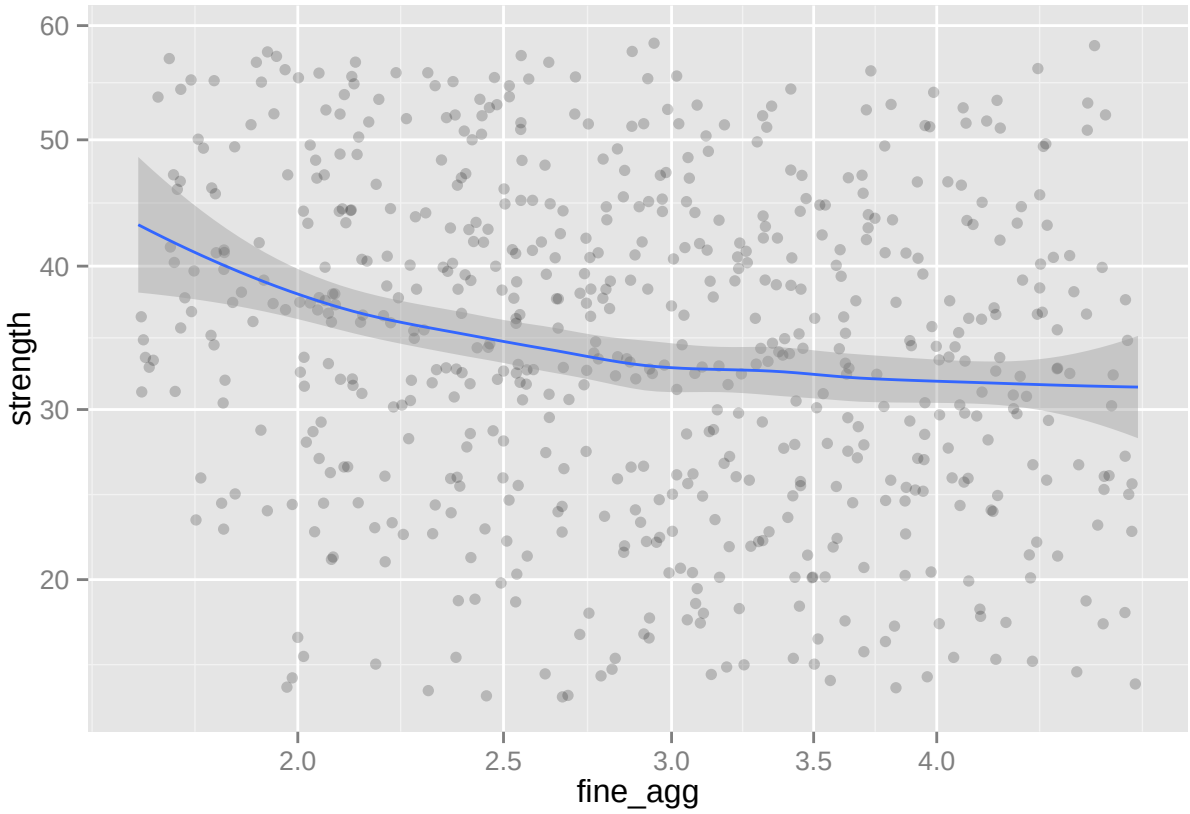
Fly ash seems to negatively affect strength slightly. The strongest samples with flyash are the ones with the least in relation to the proportion of cement.

Plasticizer vs Strength



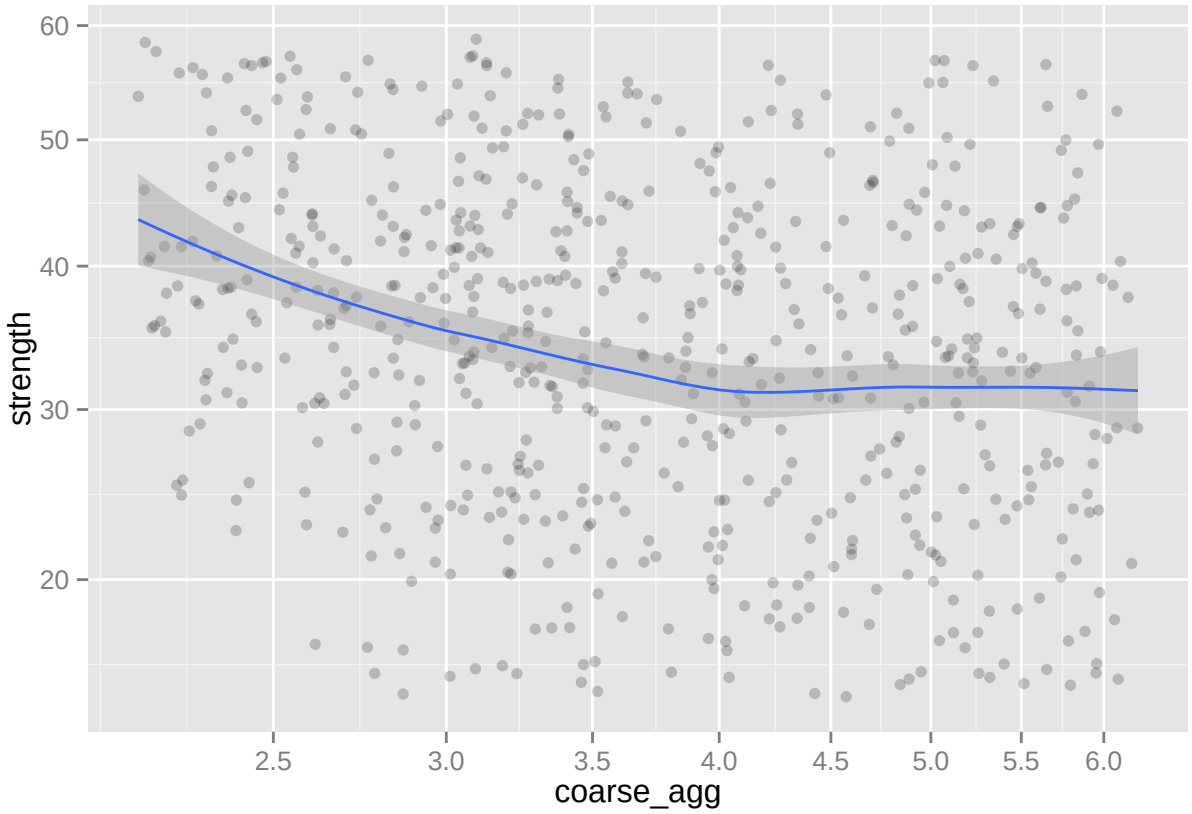
The plasticizer vs strengh chart looks almost flat, however the strongest blend appears to be around 0.32 kg per kg of cement. More plasticizer than that seems to slightly weaken the blend.

Fine Aggregate vs Strength



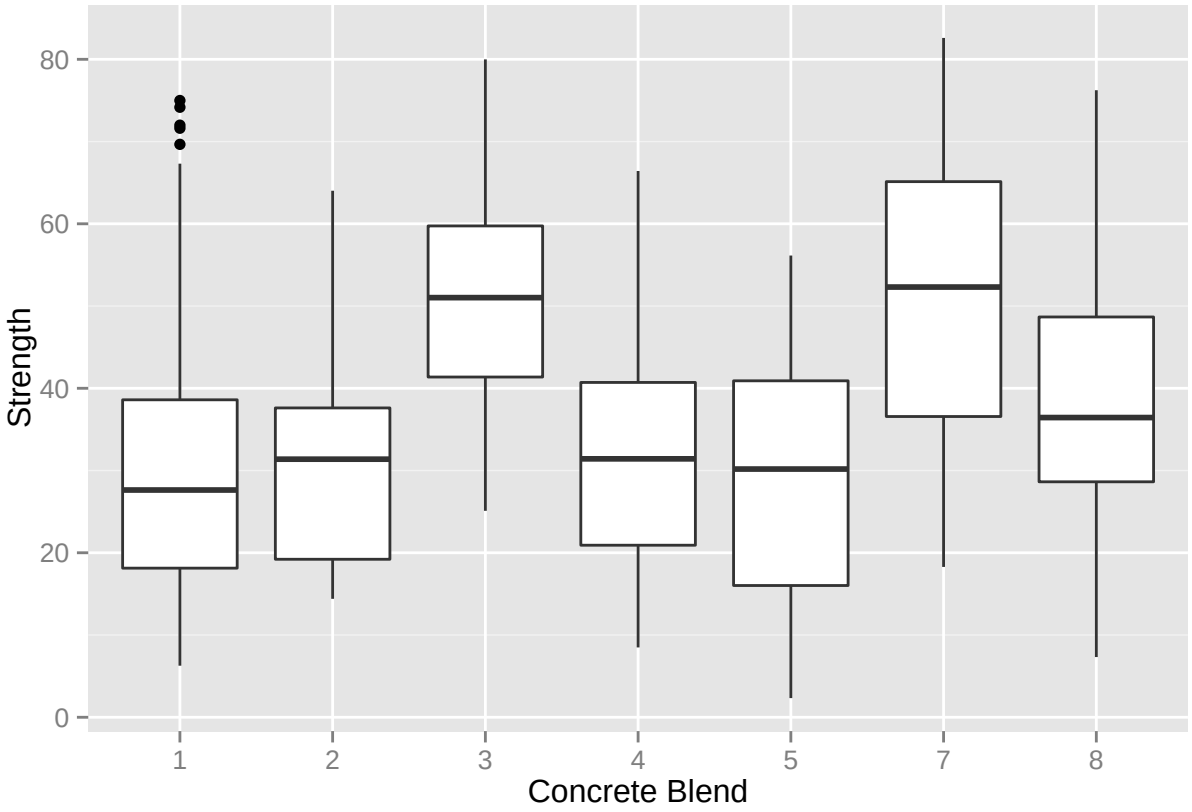
Like water and flyash, adding more than enough fine aggregate can weaken a concrete blend.

Coarse Aggregate vs Strength



Like fine aggregate, adding more than enough coarse aggregate can also weaken a concrete blend. It seems that some aggregate is needed, but too much in relation to the cement will weaken the blend.

Compressive Strength of Each Blend



Following the trends in the previous charts, it looks like blends with no flyash, but some plasticizer and slag tend to be stronger than blends with no additives or blends that include fly ash.

Blend #7 appears to be the strongest blend, which includes plasticizer and slag, but no fly ash. Blend #8 is similar to #7 but contains flyash and is weaker. Similar to #7 and #8, #3 contains no flyash and #4 contains flyash.

Interestingly when only fly ash is present, but no plasticizer or slag as in blend #2, the flyash blends appear to be slightly stronger. Is there some incompatibility between plasticizer and flyash?

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

- The water to cement ratio is inversely related to strength.
- The strongest mixtures have the least amounts of water.
- The strengthening of concrete is logarithmic over time and appears to completely harden in about 100 days.
- A little bit of plasticizer helps make strong concrete more consistently, but too much will weaken the mixture
- The mixtures with the least fly ash were strongest. It appears to impede strength.
- The sweet spot for strength with slag is slighly less than a 1:2 ratio with cement
- Plasticizer creates it's strongest blend at about 30g per Kg of cement.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Blends 3,4,7,8 seem to indicate that when plasticizer is present, the mixture is weakened with the addition of flyash, but 1 & 2 show that when no plasticizer is present, fly ash doesn't impair the strength. I think they may be incompatible additives.

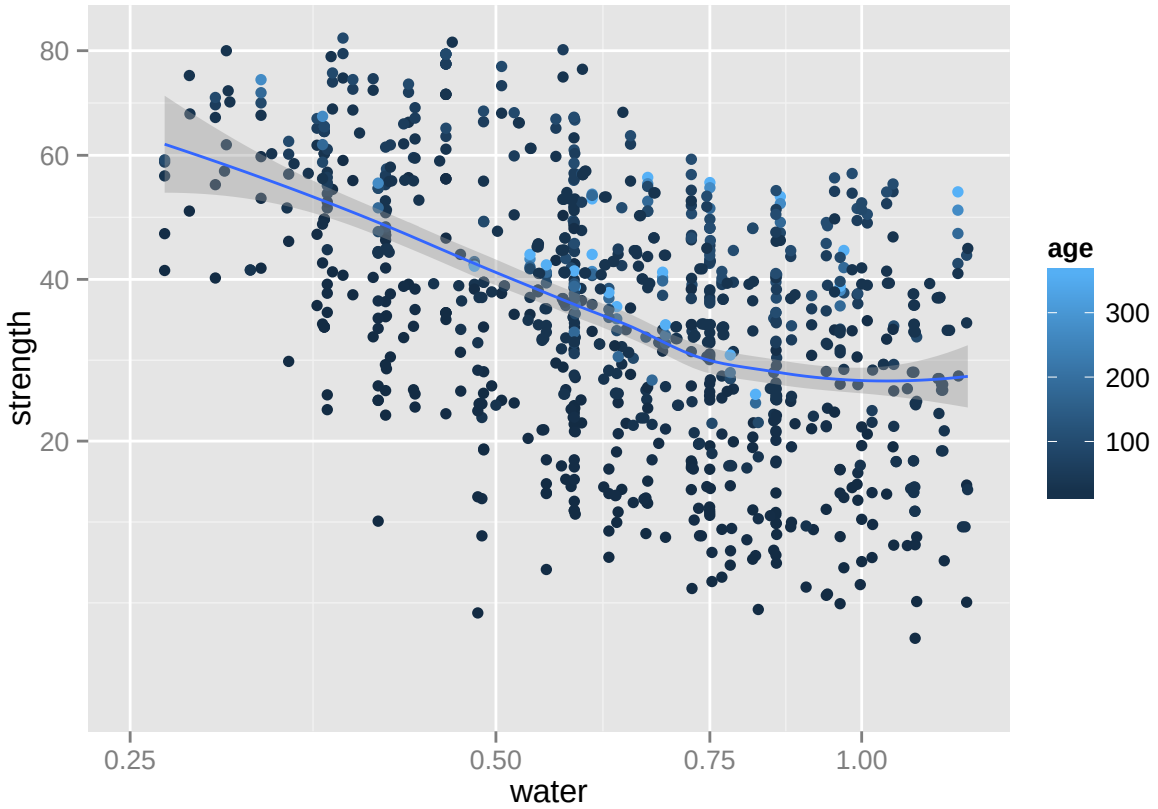
Particular blends result in having different compressive strentghs, and are also more consistently strong dispite a wetter mixture. The additives and plasticizer and slag in particular, seem to make blends more consistently strong when they harden. They also cure harder when mixed at a higher hydration when plasticizer is present.

What was the strongest relationship you found?

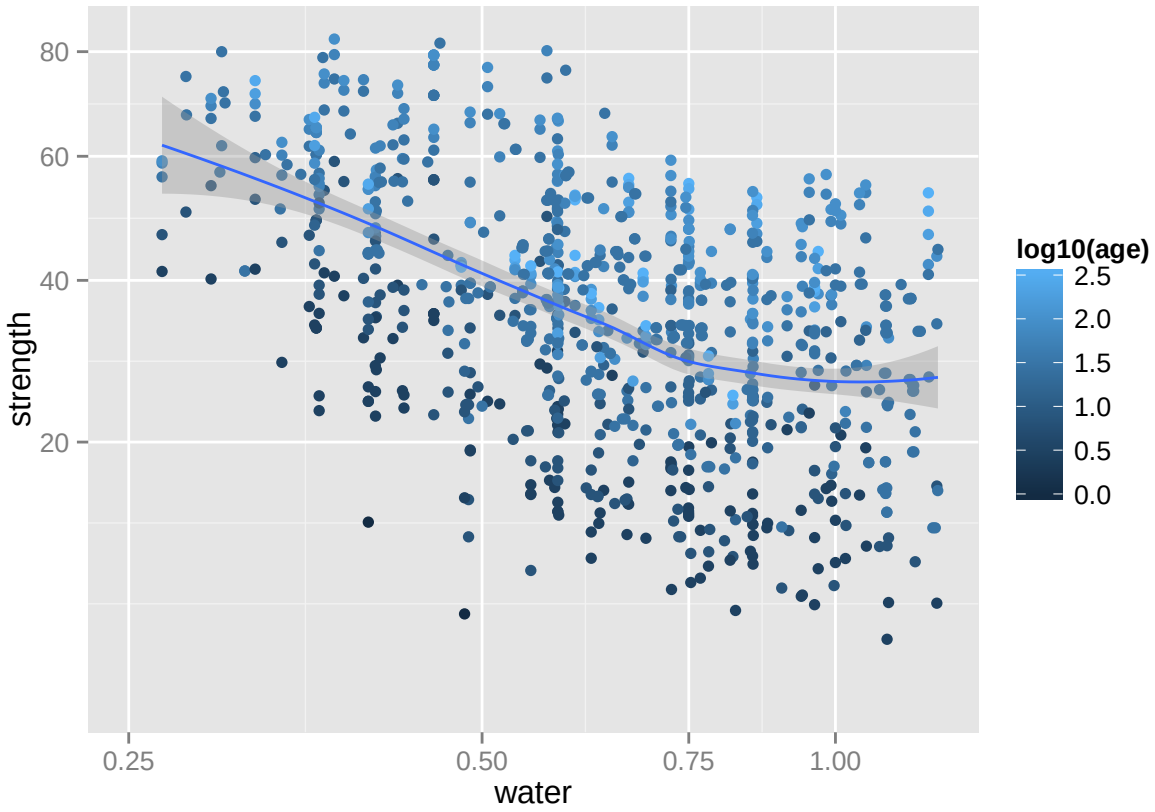
The strongest factor in the strength of a cement mixture is to keep the water to cement ratio on the low side. Even without any additives many dry concrete mixtures are among the strongest mixtures in the population.

Despite the mixture, you want the cement to fully cure for at least 100 days before putting it under a full compressive test.

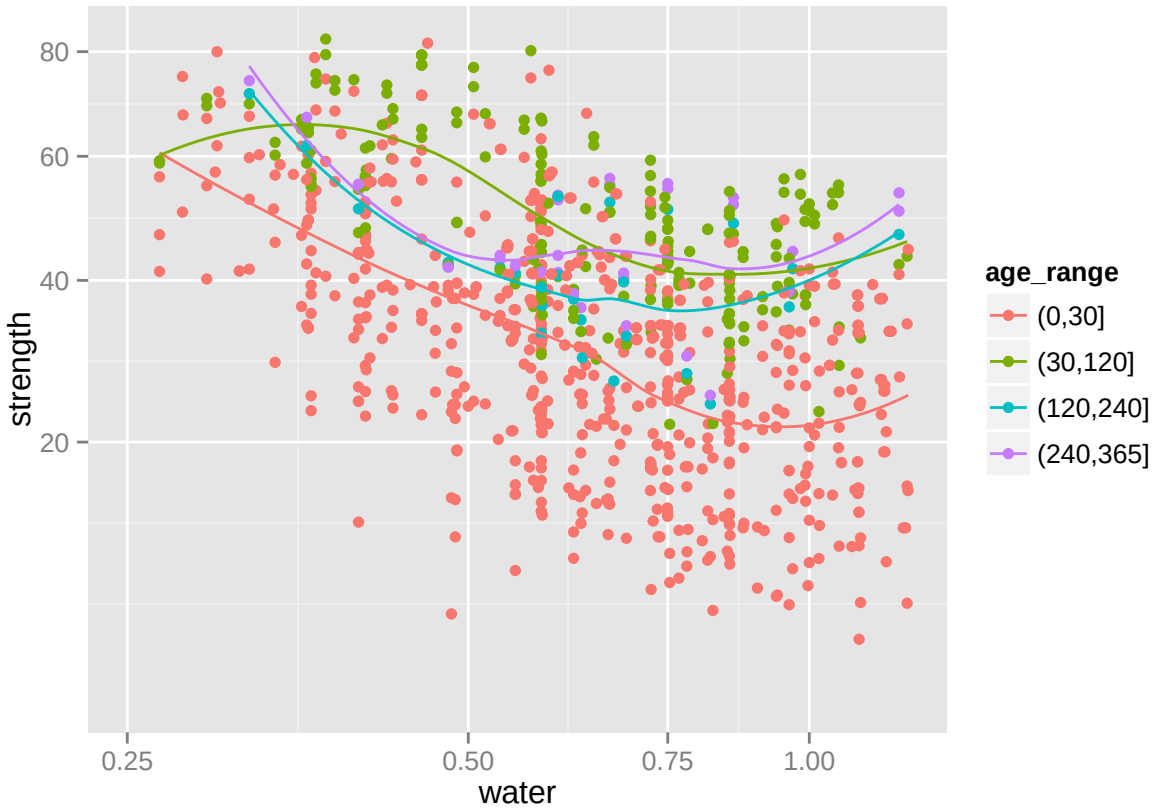
Multivariate Plots Section



I wanted to try to show the importance of cure time to strength, but so many of the samples are under 50 days old, it makes it a little unclear.

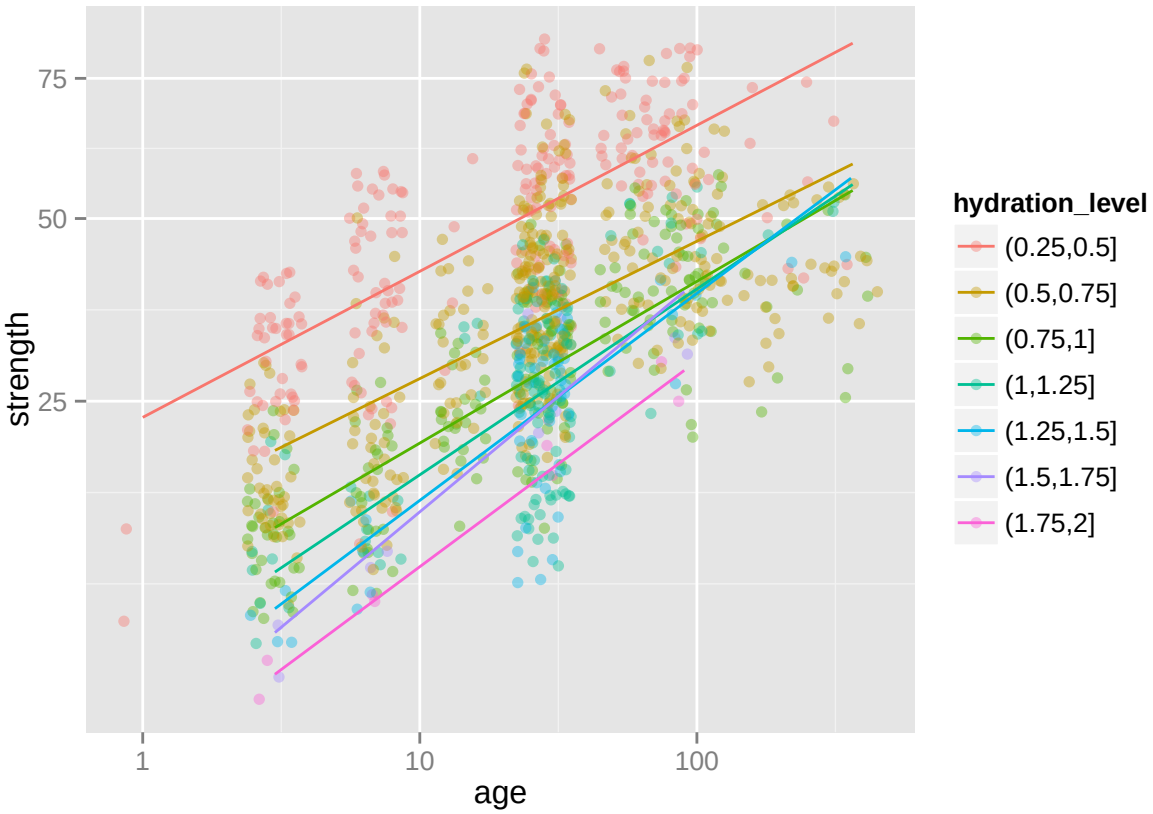


Taking the log(10) of age makes a nice gradient distribution showing the oldest blends, but the units are really unclear. It does however highlight the older blends are among the strongest. I think it might be best to break age down into buckets to separate older samples from the majority, which are under 30 days of cure time.



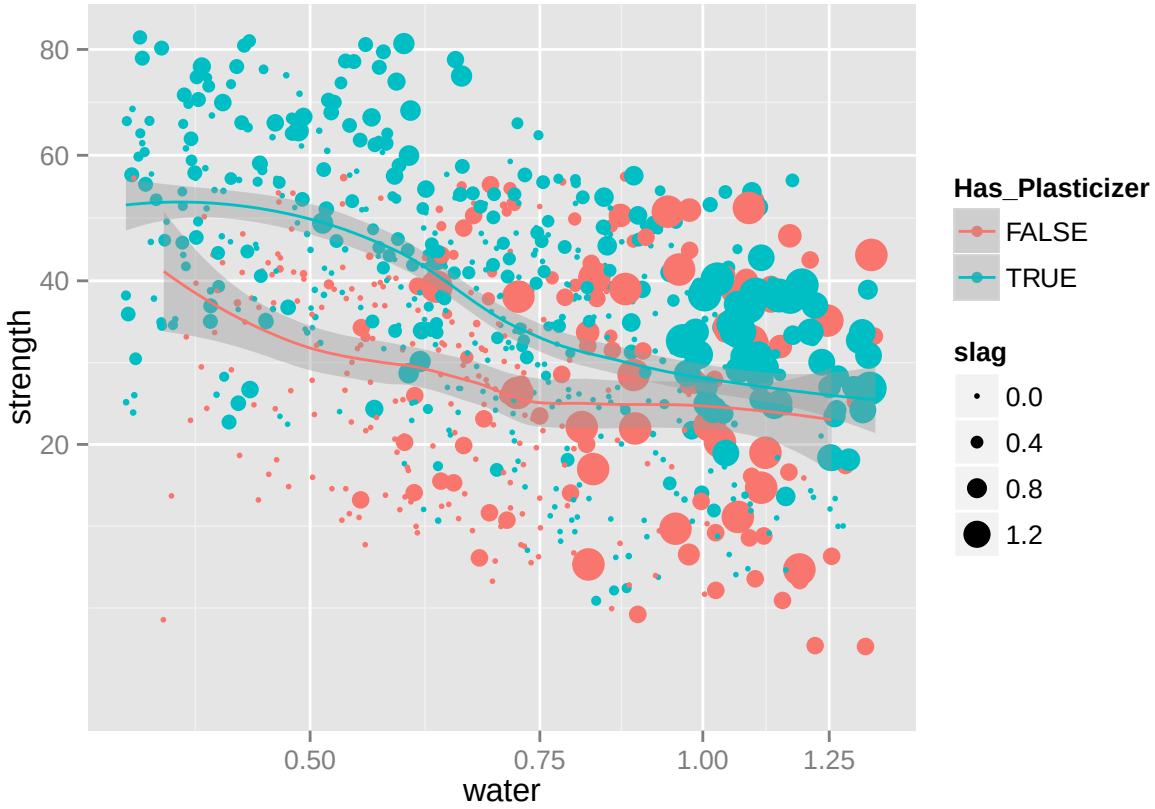
This makes the cure-time ranges more distinct, for sure. It also hightlights the issue with the distribution of cure times in the study. Most of the samples are all under 30 days, however many of the strongest blends have cure times in the neighborhood of 100 days or higher. There aren't really enough samples over 100 days to draw any accurate conclusions about what happens to compressive strength after 100 days, but most of these charts suggest it begins to weaken somewhat.

The cluster of 240-365 day old samples that are relatively strong in the high-hydration side of the chart may suggest that it takes more time for wetter samples to reach the hardness of dryer samples, but I don't think we have enough timespan in the data, or enough samples from that age group to be sure.

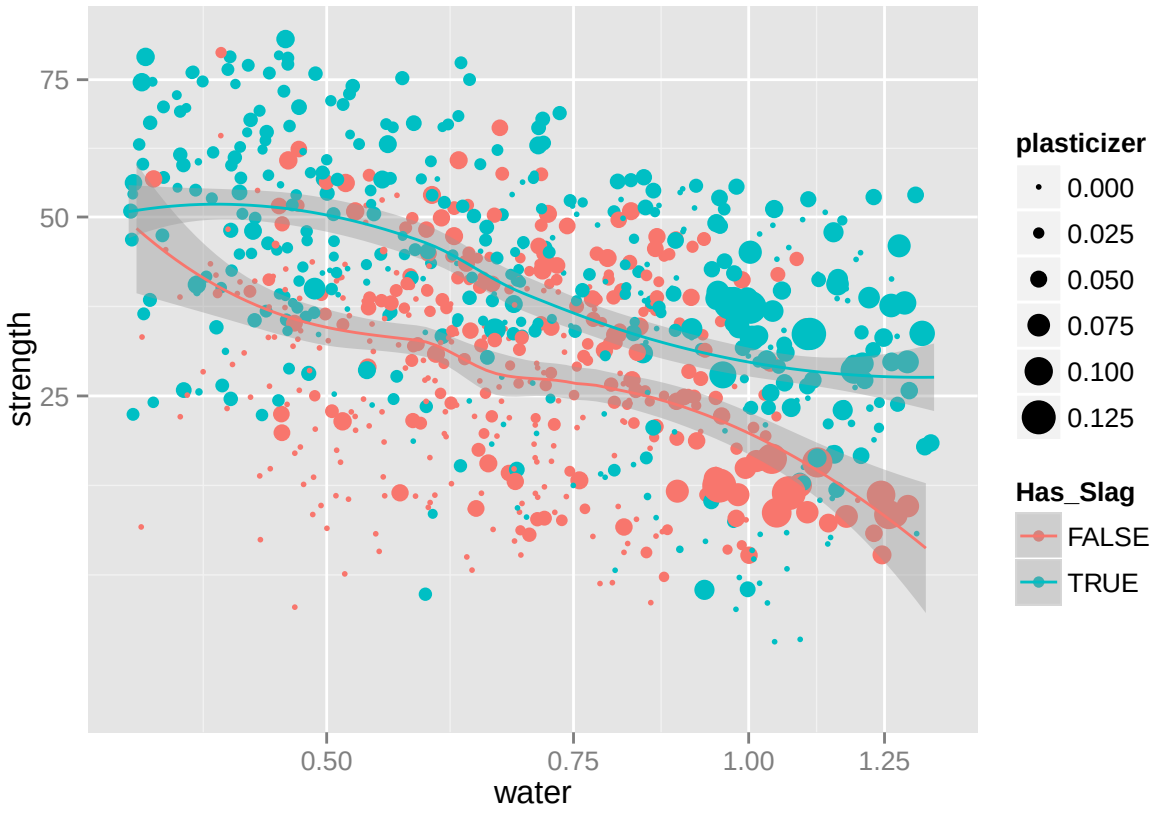


I wanted to compare hydration levels, cure times, and strength. I think it's interesting how wetter blends have a steeper slope in dry time. I think the more water in the mixture the higher the hardening rate because there is more water to evaporate out of the curing concrete. This is probably why the log10(age) and log10(water) transformations help build a better linear relationship with strength. After 100 days the data gets sparse and messy. I would have like to see if wet mixtures would eventually harden as much as dryer mixtures given enough time.

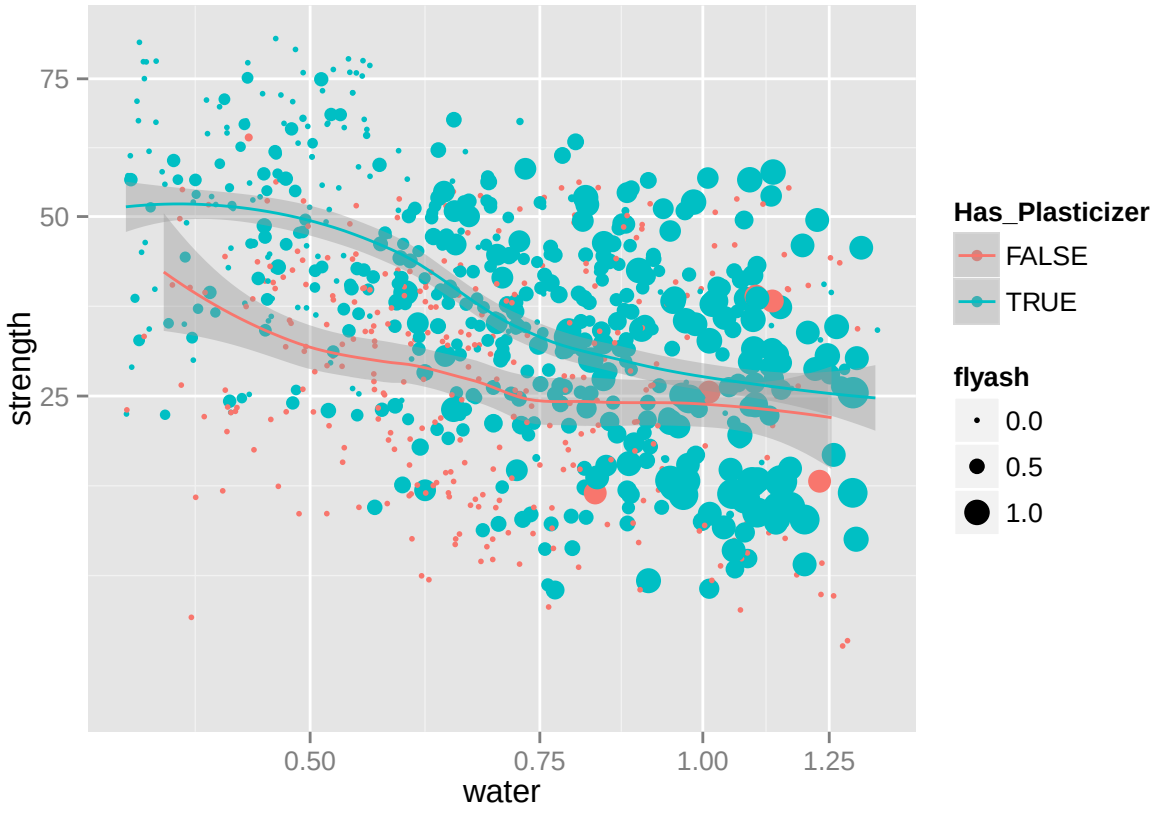
Additive Interactions



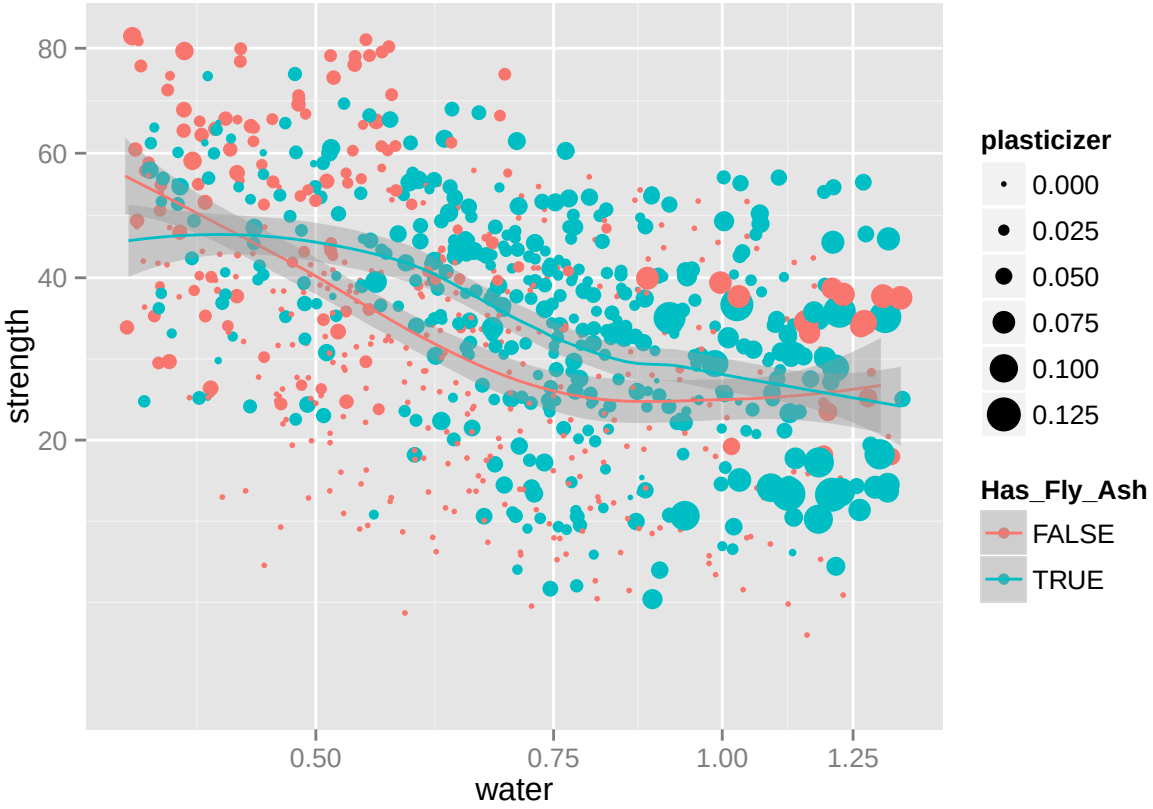
This chart shows the presence of plasticizer in blue breaks the apparent ceiling for compressive strength above 60 MPa. With plasticizer samples are able to reach above 80 MPa. These high limits can't be broken without any slag, though. The size of the dots shows that just the right amount of slag is required to get the most strength. None of the strongest samples are especially heavy or light with slag.



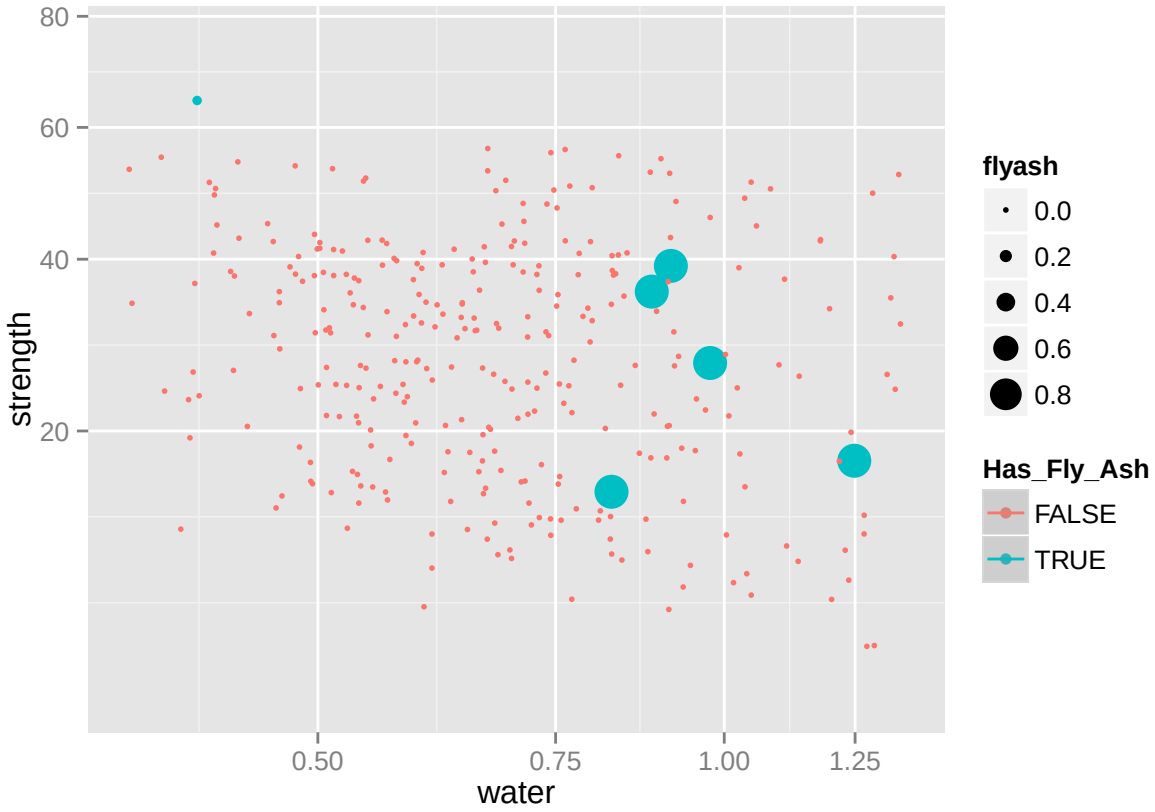
This chart shows that too much plasticizer can also weaken a concrete mixture, but slag is also required for the mixture to be in the top-tier of compressive strength.



This chart shows the interactions of flyash in mixtures with and without plasticizer. Some of the strongest blends don't have any fly ash, so it isn't really needed in a plasticizer blend for strength, however some really strong blends have a little bit of fly ash. Too much fly ash seems to weaken a sample.

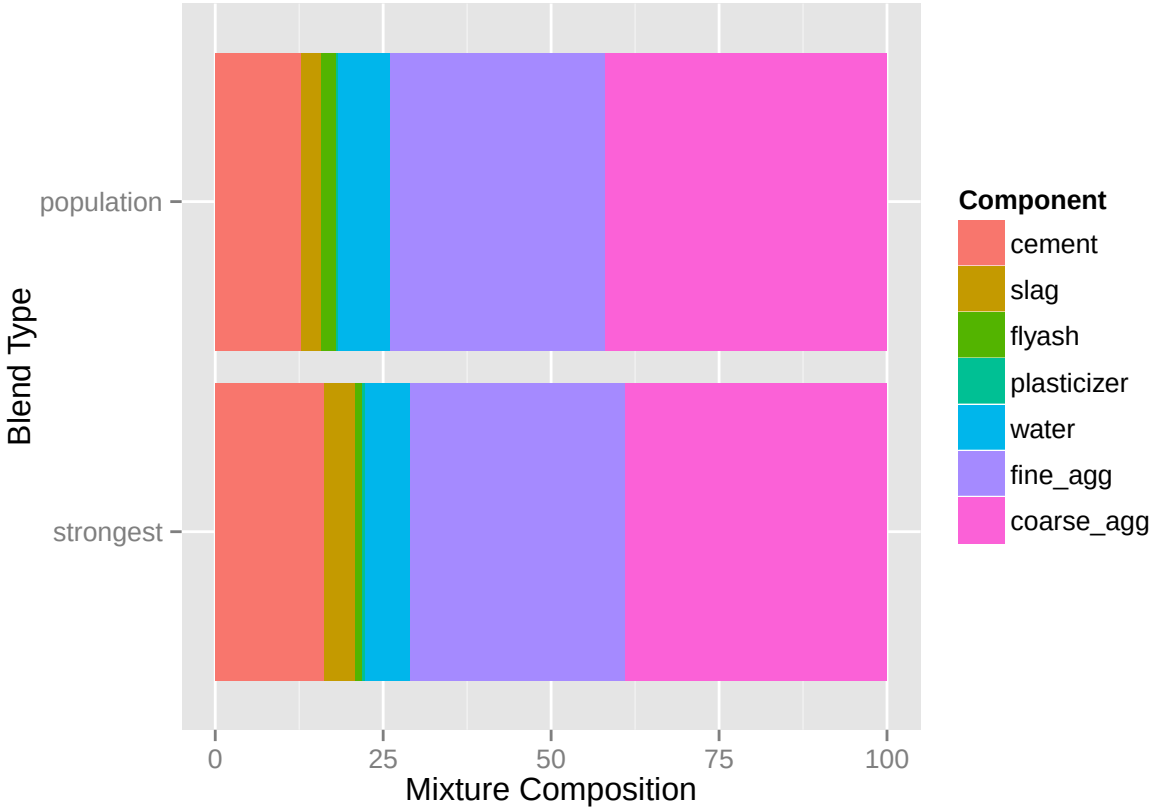


This chart shows that fly ash is not required for a mixture to be in the strongest category, although some very strong blends have a little bit of fly ash. As with any additive, adding too much can really weaken a concrete mixture.



I wanted to examine the behavior of fly ash in blends that do not have plasticizer. Unfortunately there are only six samples that have fly ash, but no plasticizer, so we won't do that comparison with this data.

Blend Composition



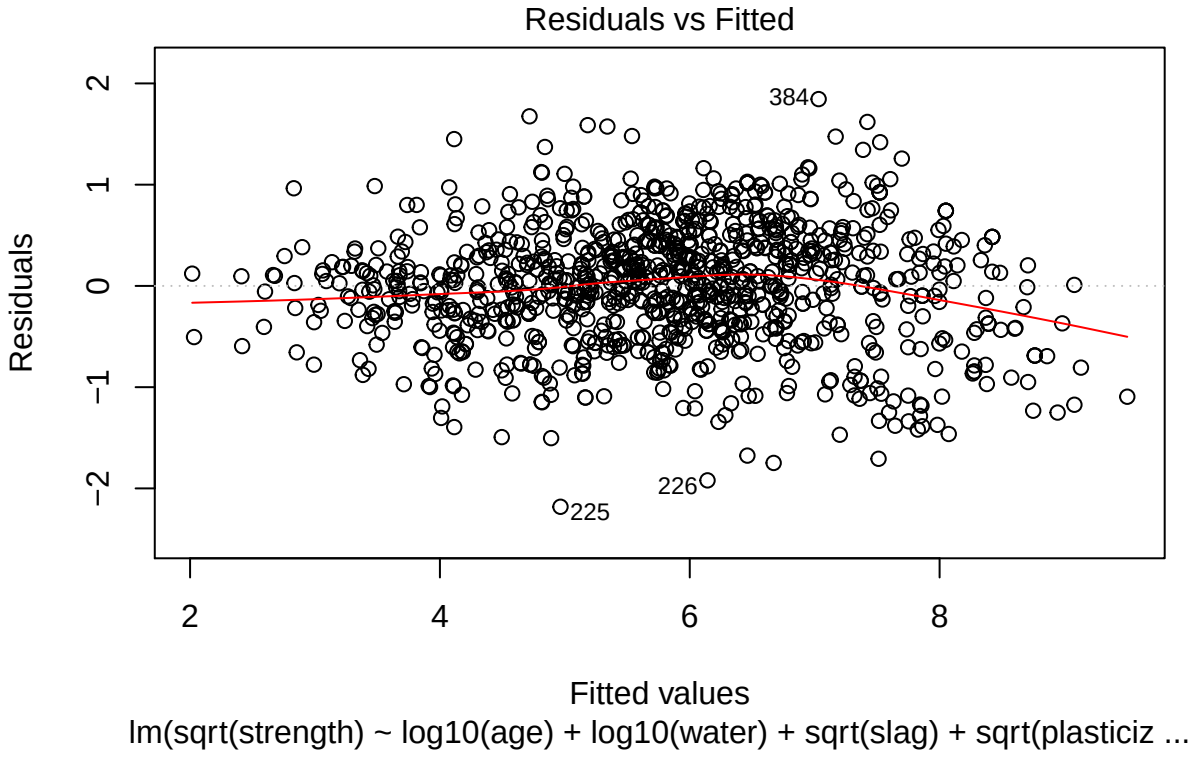
The bar blot above shows a comparison of the composition of a sample of the strongest blends with strengths greater than 60 MPa and the population blends, which are taken from a random sample of the population. The strongest blends show they are sparing on the water, aggregates and additives and proportionally heavier on the cement.

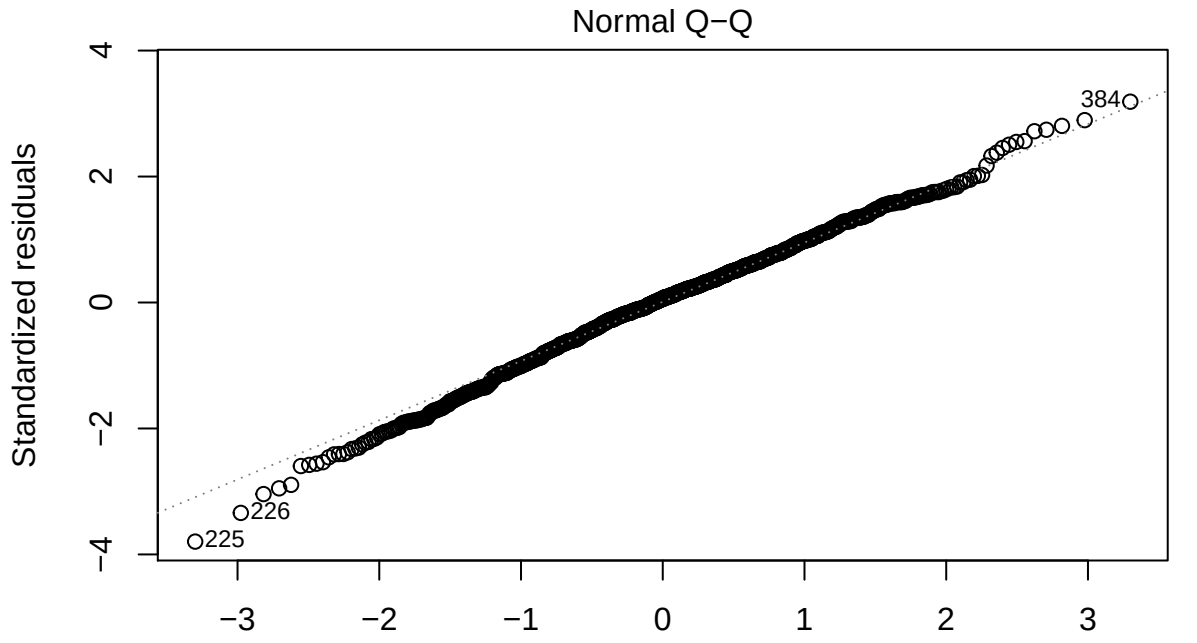
Linear Regression

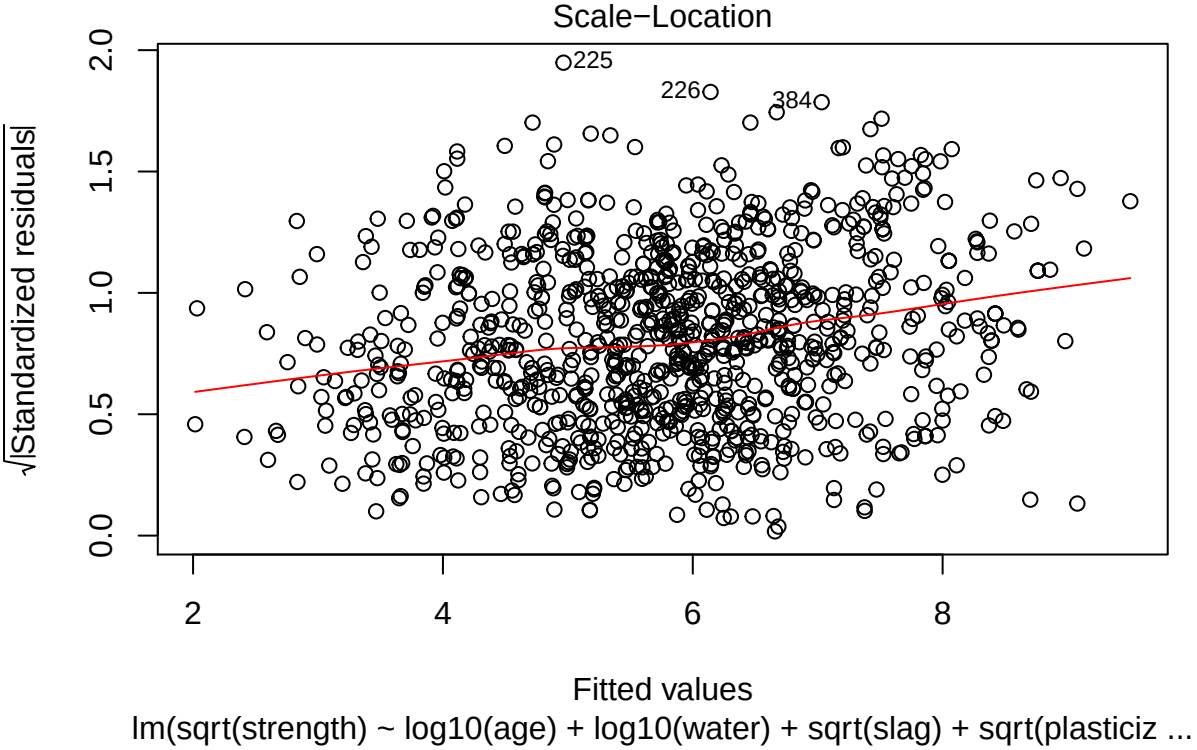
```
##
## Call:
## lm(formula = sqrt(strength) ~ log10(age) + log10(water) + sqrt(slag) +
##     sqrt(plasticizer) + sqrt(flyash) + log10(fine_agg) + log10(coarse_agg),
##     data = c_ratios, rm.na = FALSE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18126 -0.35907  0.03286  0.37365  1.84514
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.74287    0.24030   3.091  0.00205 **
## log10(age)         1.75925    0.03568  49.306  < 2e-16 ***
## log10(water)      -7.54533    0.34754 -21.711  < 2e-16 ***
## sqrt(slag)         1.58794    0.06130  25.905  < 2e-16 ***
## sqrt(plasticizer)  1.33086    0.32758   4.063 5.22e-05 ***
```

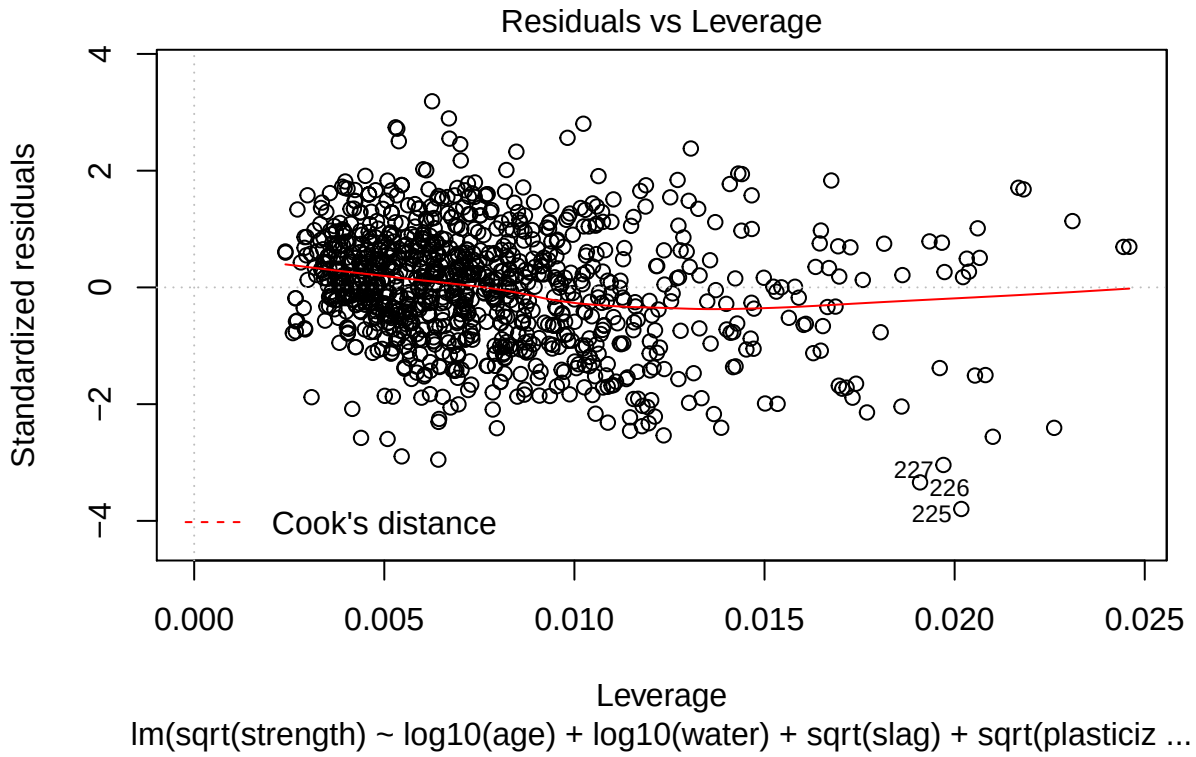
```
## sqrt(flyash)       0.79718    0.09054   8.804  < 2e-16 ***
## log10(fine_agg)   -0.15028    0.32337  -0.465  0.64222
## log10(coarse_agg)  0.75316    0.37819   1.992  0.04669 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5805 on 1022 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8394
## F-statistic: 769.4 on 7 and 1022 DF,  p-value: < 2.2e-16
```

```
##       (Intercept)        log10(age)      log10(water)        sqrt(slag)
##         0.7428654         1.7592455        -7.5453333         1.5879361
## sqrt(plasticizer)      sqrt(flyash)   log10(fine_agg) log10(coarse_agg)
##         1.3308558         0.7971776        -0.1502816         0.7531594
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(strength) ~ log10(age) + log10(water) + sqrt(slag) + sqrt(plasticiz ...
384
226
225
4
2
0
−2
−4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
lm(sqrt(strength) ~ log10(age) + log10(water) + sqrt(slag) + sqrt(plasticiz ...
225
226
384
2.0
1.5
1.0
0.5
0.0
2
4
6
8



Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

There were a lot of interesting relationships. Each of the optional additives seems to play a role in determining the strength of the resultant concrete. Plasticizer and Slag seem to improve the strength in the right concentrations, but if too much is added, they impair the strength of the mixture. Fly ash seems to weaken the mixture if any is added in a plasticizer mixture.

Were there any interesting or surprising interactions between features?

The charts show the affects of fly ash and slag on mixtures containing plasticizer. They confirm the sweet spot of 30g of plasticizer per kg of cement. They also show the effects of the presence of fly ash and slag to a mixture containing plasticizer. They show that increasing the wetness of the mixture above 1:2 ratio with cement will impair the strenth of concrete, and possibly increase the time it takes to fully harden.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I created a linear regression model starting the transformations that I found made the best normal distributions and actually got a fairly good R^2 score of 0.84. I changed the transformations some, but wasn't able to signficantly improve the R^2 beyond 0.84 for the model.

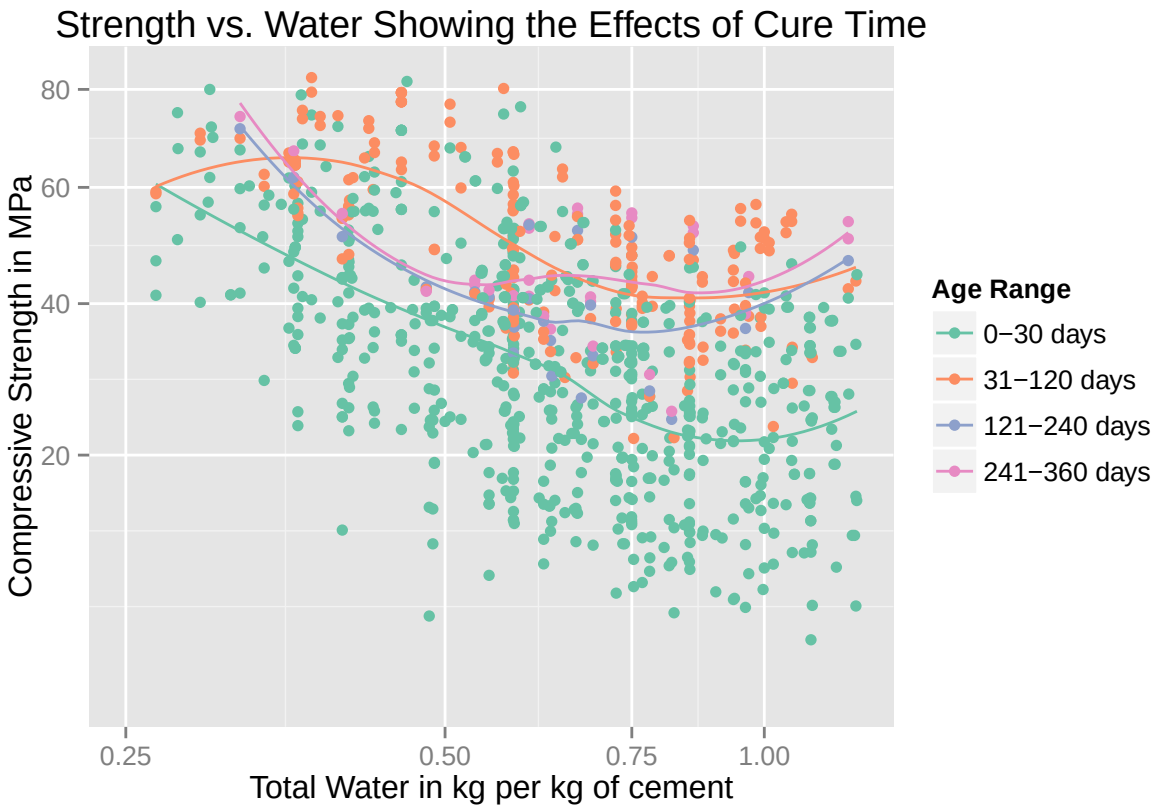
I removed and re-added variables and transformations, but wasn't able to improve the score much. The residuals plot seems to have a little bit of a curve, and looks fairly evenly distributed and a little bit curved which seems like there must be some unaccounted for variable hiding in there.

I actually thought when I started I wouldn't get a model quite this good with this data, so it did exceed my expectations, but I think there are limitations. For example the data doesn't take into account the environmental condions when the cement was curing. The plots also don't fit an exact linear model after the transformation. I think some non-linear regression model might get a closer fit. The slag vs strength curve probably deserves some more complex function to match the sweet spot for strength in it's ratio to cement.

A more complex model that considers multiple different transformations, like the ratio of ingredients to the total weight, might add additional benefits in this model. I think the ratio of some ingredients to cement is important, but there may other interactions based on an ingredient's exposure to water that could be important to compressive strengh.

Final Plots and Summary

Plot One

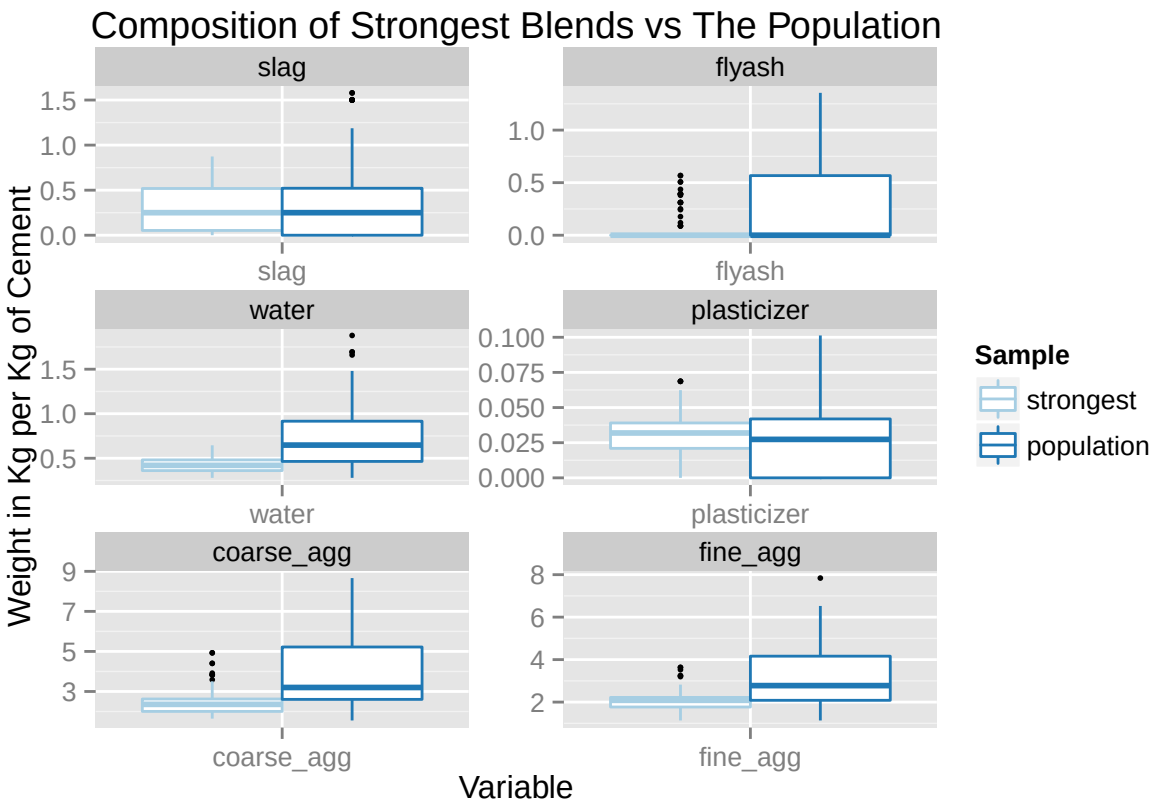


Description One

This plot shows strength vs water content of each sample, but additionally shows the age range by color. This plot demonstrates the effects of age on hardness. Notice how all the samples over 30 days are in the high range for compressive strength. Also this plot shows the disproportionately high number of newer samples under 30 days old in our study.

Plot Two

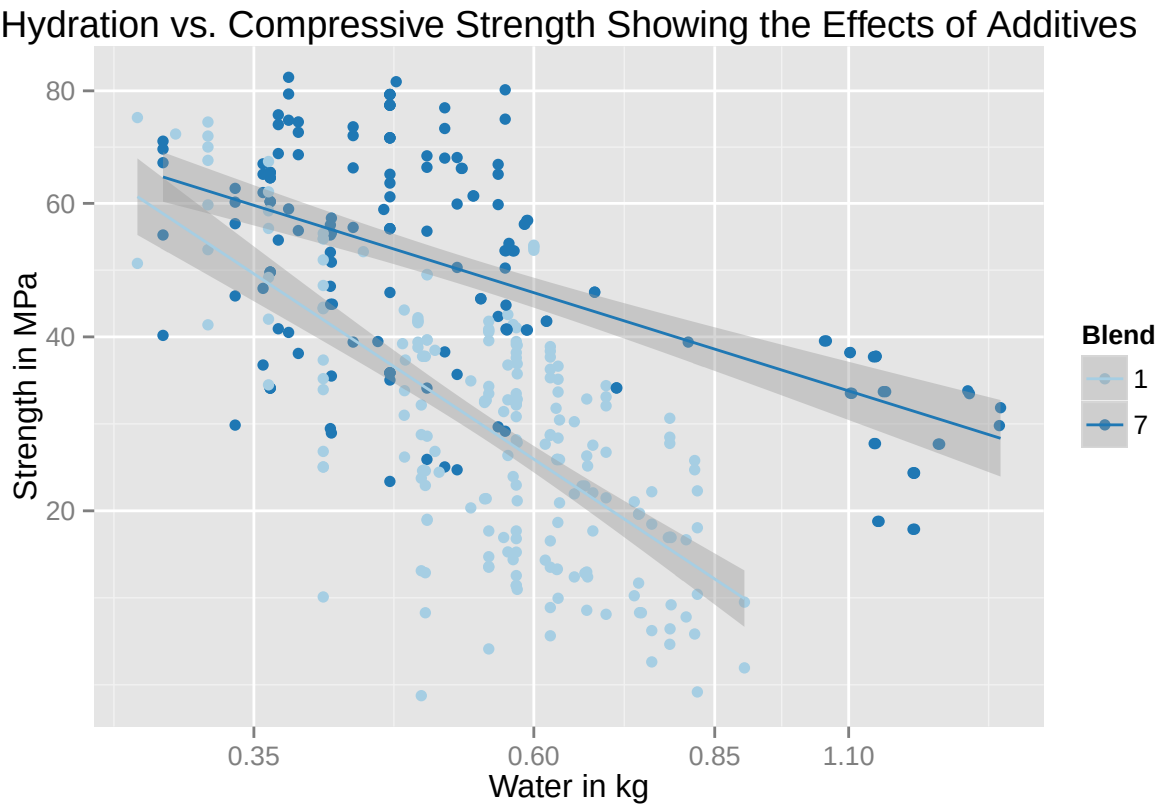
Boxplot showing difference in means between strong and population samples



Description Two

This plot shows the proportionate makeup of the strongest blends in comparison with the same size sample of the general population. The strongest blends have a lower level of water and more cement. A little bit of plasticizer and slag are needed, but often too much are added and this contributes to weakness. Try to keep fly ash out of the mixture if your goal is a strong concrete mixture.

Plot Three



Description Three

This plot shows the difference water vs. strength in the strongest blend, blend 7, and the blend with no optional additives, blend 1. It shows that the concrete is able to better maintain it's strength dispite a higher level of hydration. I think the addition of the additives probably aids in pouring and forming the cement while maintaining it's hardness.

Reflection

In this study first we identified the distribution of attributes in the dataset. We then found correlations between the ratio of the various components relative to the weight of cement in the mixture.

Once these relationships were identified, and transformed to more approximate linear relationships, a linear regression model was built. The model was tuned to get the best R^2 score using transformations that proved benificial in the univariate analysis.

Overall I'm surprised about how much can be inferred about the complex interactions of the components used in concrete mixtures with only a little over 1000 rows of data. Before starting this project I knew very little about what made a good concrete mixture. Now without studying how to blend concrete, or the purpose of the various components, I think I have a good idea how the components are used, and the ideal composition if your goal is strong concrete. It's also interesting to see how the interactions can be plotted and tried to fit to a linear relationship in order to predict a result, then test the quality of that estimation.

Early in the study I found ratios to be the best way to analyze the observations in terms of compressive strength. The issue I encountered was what ratio to choose. I still don't think I've explored all this data, and I think the remainder of the variables to predict compressive strengh may lie in the interactions between components that I left unexplored.

I added some categorical variables and plotted some apparent relationsips between components and their effects on compressive strengh. Overall there were some good looking correlations there, once they were cropped and shaped up, but there were some issues with the scope and composition of the data, that left me wishing I had a little bit of an improved dataset.

I wanted to see a more balanced range of cure times. I'm still not sure there are enough older samples to make accurate statements about the strength of samples over 100 days old, or with higher hydration levels. I have a suspicion that if we saw a study that went on for two to three years, there would be more strong samples in the longer cure time ranges with wetter blends.

I also wanted to expore what seemed to be a conflicting ratio between fly ash and plasticizer, but when I got to the point of comparing blends, there weren't enough observations with fly ash an no plasticizer to make any meaningful comparisons.

I think the multiple linear model I used here would make a good starting point for a more complex and more accurate model that took more factors into account, like the environmental conditions in cure time, and to compare more samples at the same points in age, such as to have a higher quantity of samples at the 100 day cure time mark. Using more complex functions to match the non-linear behaviors of certain components, like slag, would also likely yield a stronger predictive score.

I also think chosing only one ratio may be limiting the predicive accuracy of this model. In the water ratio correlation matrix it shows plasticizer has a fairly strong predictive influence on strength. I expect this might be an indication that the ratio of water to plasticizer is also important, and might be a good improvement to this regression model. I think there may be a few unexplored interactions like this.

If I had more data and spent more time, I would like show more comparisons with hydration, cure time and strength. I would think more observations at various cure time intervals, and possibly additional variables like average temperature and humidity where the samples cured might help build a better predictive model.