Albert Wiryawan
Avw2@illinois.edu
673431511
MSE 404 MAC

# Matlab Data Analysis

## PART A: Analysis of strength-moisture data for ASTMC TYPE I Portland cement

### A1:

```
%import data using rules declared above
ASTMCTypeI = readtable("/home/avw2/MSE404-MAC/ASTM_C_TypeI.csv")
%turn table into array using table2array function
ASTMCTypeI_array = table2array(ASTMCTypeI)

%set variables for col
% array slicing (row, column)
moisture_I = ASTMCTypeI_array(:,1);
strength_I = ASTMCTypeI_array(:,2);
```

### A2:

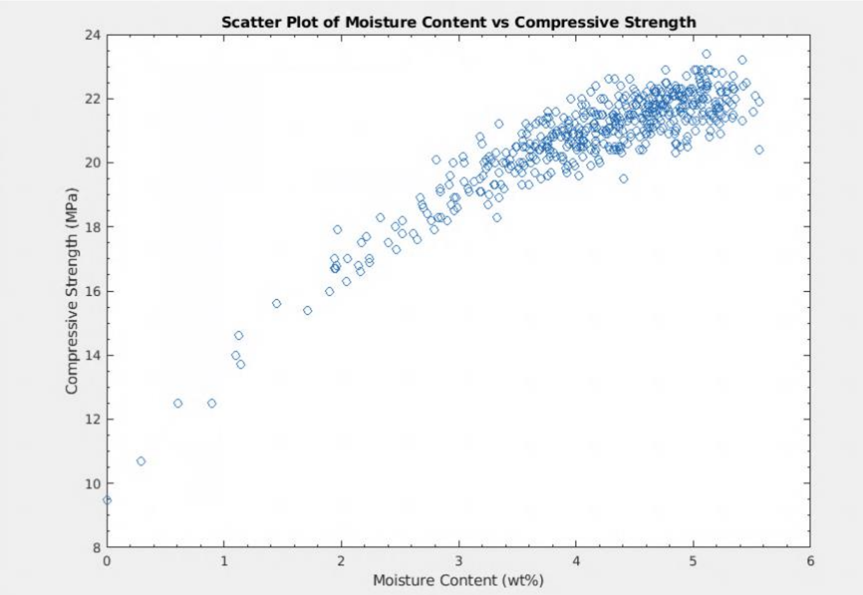

Figure 1. Evident positive direct correlation between moisture content and compressive strength

**A3:**

i. one-dimensional histogram of the observed moisture contents

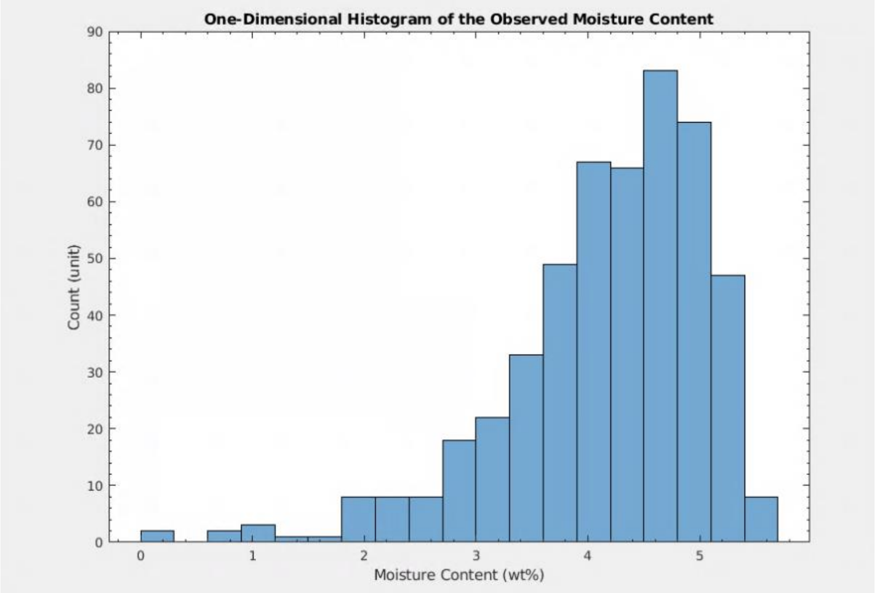

Figure 2. One Dimensional Histogram for observed moisture content. Common moistures contents lie within 4-5%.

ii. one-dimensional histogram of the observed compressive strengths

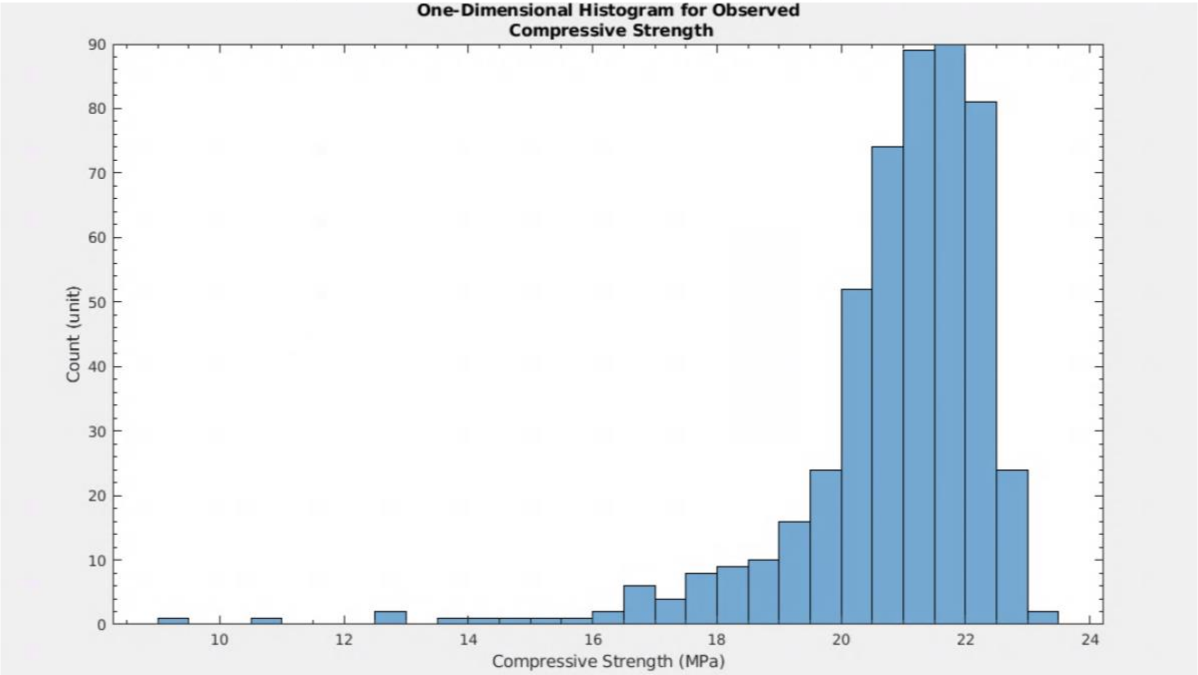

Figure 3. One dimensional histogram for compressive strength. Common compressive strength lies within 20-22 MPa

iii. two-dimensional histogram of the moisture contents and compressive strengths.

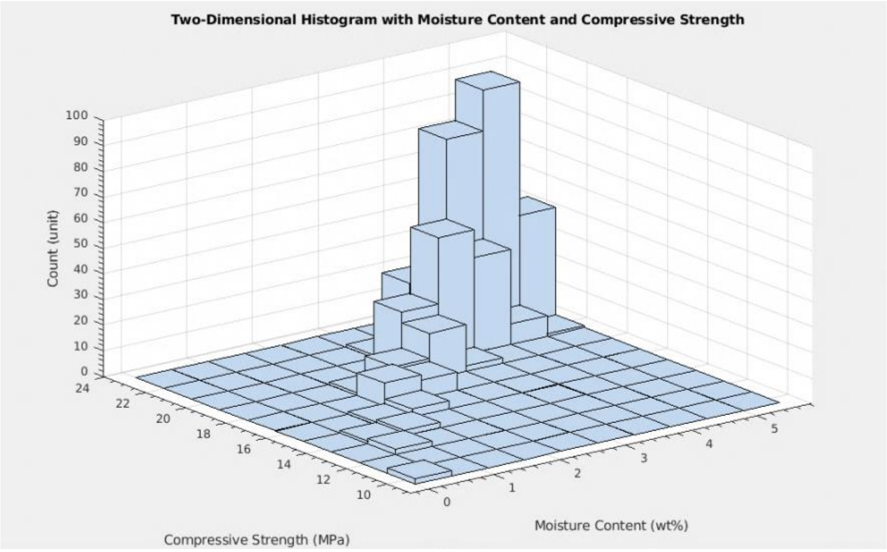

Figure 4. Two-dimensional histogram of moisture content and compressive strength with count on the z-axis. From the figure it looks like most samples are have a large moisture content and compressive strength.

**A4:**

**Code:**

```
[r,p,r_lower, r_upper] = corrcoef(moisture_I, strength_I);
```

(a) The Pearson correlation coefficient was determined to be 0.8909.

(b) This number shows that there is a strong positive correlation between moisture content and compressive strength. The p-value obtained for this correlation was 0.00. A small p-value like this indicates that the correlation is statistically significant.

**A5:**

**Code:**

```
a = moisture_I.';
b = strength_I.';
[RHO,PVAL] = corr(a',b','Type','Spearman');
```

(a) The spearmen correlation coefficient was found to be 0.8019.

(b) The value found states that the correlation between the two are statistically significant. The p-value assocatied with this statistic was found to be 1.720204406102098e-113 which is extremely close to 0 indicating that the calculated spearmen coefficient is statistically significant. The use of the spearmen coefficient is better suited in this scenario as the data is not highly linear. As a result, spearmen coefficient would be better for a monotonic assessment.

**A6:**

**Code:**

```
model1 = fitlm(moisture_I, strength_I);
Model1
Anova(model1)
```

(i) No, there are no physical grounds to set C = 0. This is because compressive strength will not be initially be set to zero as there might be external interactions as well as bonding factors that can prevent compressive strength from being initially set to 0 for its intercept.

(ii) The $R^2$ value was determined to be 0.794 indicating that 79.4% of the variation of compressive strength is explained by the moisture content regressor. As a result, this indicates a good fit.

(iii) The p-value was determined to be 8.26e-173 and the F-value was found to be 1.92e+03. The small value obtained for the p-value indicates that the modeled relationship is statistically significant. In addition, the large F-value obtained indicates that the variable of the group means is large relative to within the group variability. These factors indicate a good model fit.

(iv) The independent variable is moisture content while the dependent variable is compressive strength. This makes sense as moisture content can be controlled and the compressive stress can be evaluated by its effects inside the lattice due to things like bonding.

**A7:**

**Code:**

```
%create script for AIC
y = strength_I;
L = length(strength_I);
%check polynomials fits up to degree 10
k = linspace(2,11,10);
%initialize an array of zeros that will be manipulated lated
RSS = zeros(1,10);
%apply power over all data points and obtain their SSR to calculatw AIC
for i = 1:10
   X = ones(L, i+1);
   for j = 1 : i
      X(:,j) = moisture_I.^j;
   end
   [b, bint, r, rint,stats] = regress(y,X);
   RSS(i) = r'*r;
end
AIC = zeros(1,10);
for x = 1:10
   AIC(x) = 2*k(x) + L*log(RSS(x)/L);
End
%store value and index to determine the power at which the lowest value of AIC is
[val, idx] = min(AIC);
```

The lowest value of AIC is  -5.508584506294627e+02 which is at the power of 2

**A8:**

```
X = [1,2,3,4,5,6,7,8,9,10]
Plot(X, AIC)
```

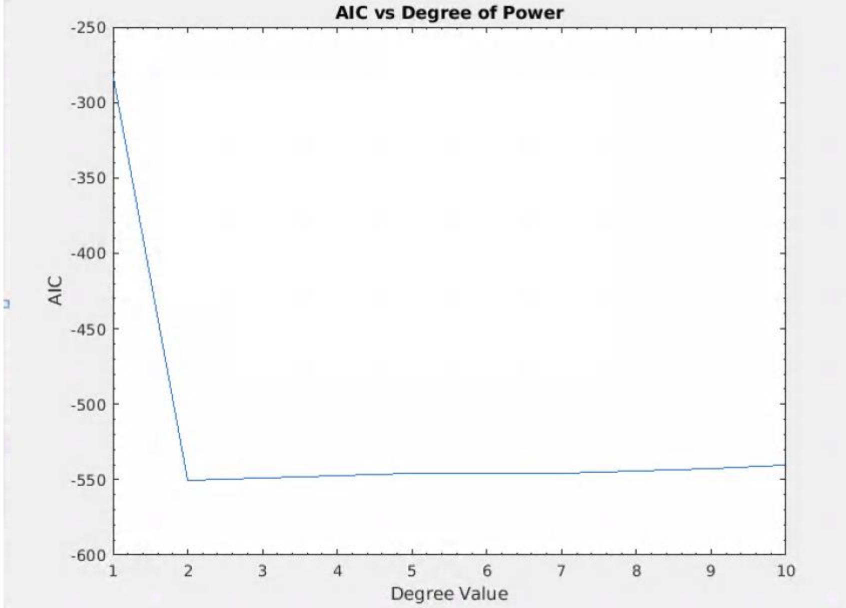

Figure 5. Overfitting is seen when degree of the power Is greater than 2

**A9:**

**Code:**

```
%create script for AIC
y = strength_l;
L = length(strength_l);
CV_MSE_arr = nan(L,1);
MSE = zeros(1,10);
for i = 1:10
  X = ones(L, i+1);
  for j = 1 : i
    X(:,j) = moisture_l.^j;
  end
  [b, bint, r, rint,stats] = regress(y,X);
  MSE(i) = mean(r.^2);
end

for a = 1:10
  for x = 1:L

    idx = setdiff(1:L, x);
    X = ones(L-1,i+1);
    X_solo = ones(1, i+1);
    for j = 1:10
      X(:,j) = moisture_I(idx).^j;
      X_solo(:,j) = moisture_I(x).^j;
    end
    [b, ~,r,~,stats]= regress(y(idx),X);
    CV_MSE_array(x) = (y(x)-X_solo*b).^2;

  end
  CV_MSE(i) = mean(CV_MSE_arr);
End
plot(x, MSE)
```

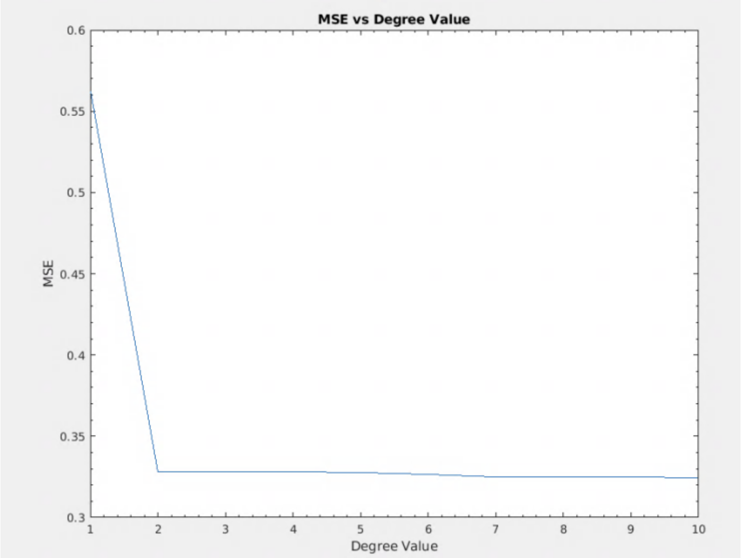

Figure 6. MSE vs Degree value. There is a sharp decrease at n =2 and the graph continues to decrease.

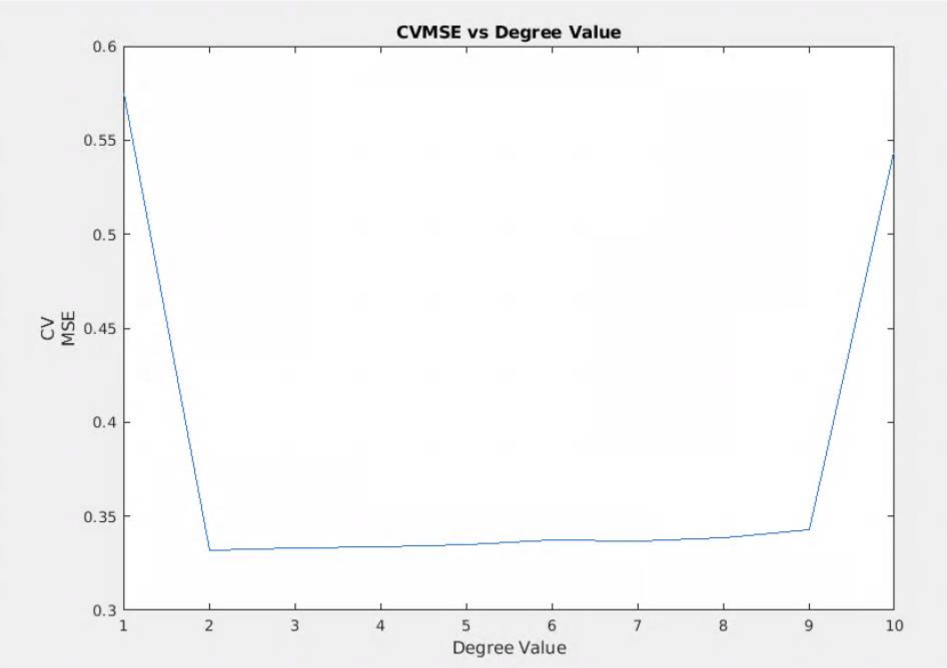

Figure 7. CV_MSE vs Degree Value. There is a sharp decrease at n=2 and increase at n=9

From the CV perspective, the minimum CV-MSE is located at a degree value of 2. As such, this degree of polynomial would be the best model fit for the modeled data.

**Part B:**

**B1:**

**Code:**

```
%create arrays for 5% of both
strength_I_5p = [];
strength_III_5p = [];
x =1
for i = 1:size(strength_I)
   if(moisture_I(i) >=4 && moisture_I(i) <= 6)
      strength_I_5p(x) = strength_I(x);
      x = x+1
   end
end
y = 1
for i = 1:size(strength_III)
   if(moisture_III(i) >=4 && moisture_III(i) <= 6)
      strength_III_5p(y) = strength_III(i);
      x = x+1
   end
end
```

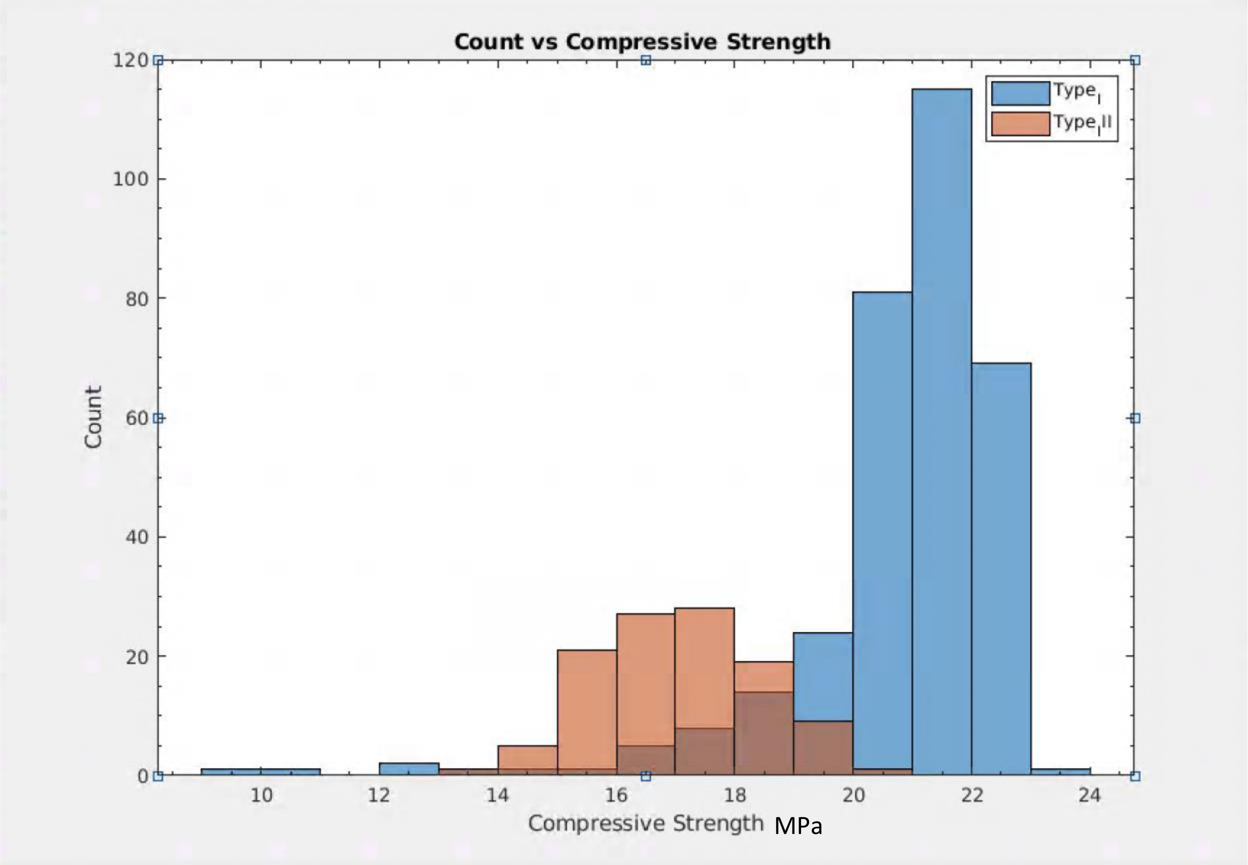

Figure 8. The distributions show by type I and type III have different distributions under the same range of moisture. Type I is able to handle more compressive strength

**B2:**

**Code:**

```
[h,p,ci,stats] = ttest2(strength_III_5p,strength_I_5p,0.05, 'both','unequal')
```

Utilizing a t-test the value of h is given as 1 which means that the null hypothesis is rejected at a 5% significance level interval. As a result, the sets of data belong to are significantly different and do not belong to the same population distribution.

**B3:**

**Code:**

```
for i = 1:size(strength_I)
   if(moisture_I(i) >=2 && moisture_I(i) <= 3)
      strength_I_5p(x) = strength_I(x);
      x = x+1
   end
end
y = 1
for i = 1:size(strength_III)
   if(moisture_III(i) >=2 && moisture_III(i) <= 3)
      strength_III_5p(y) = strength_III(i);
```

```
    y = y+1
  end
end

histogram(strength_I_5p)
hold on
histogram(strength_III_5p)
hold off
[h,p,ci,stats] = ttest2(strength_III_5p,strength_I_5p,0.05, 'both','unequal')
[h,p,ci,stats] = ttest2(strength_III_5p,strength_I_5p,0.01, 'both','unequal')
```

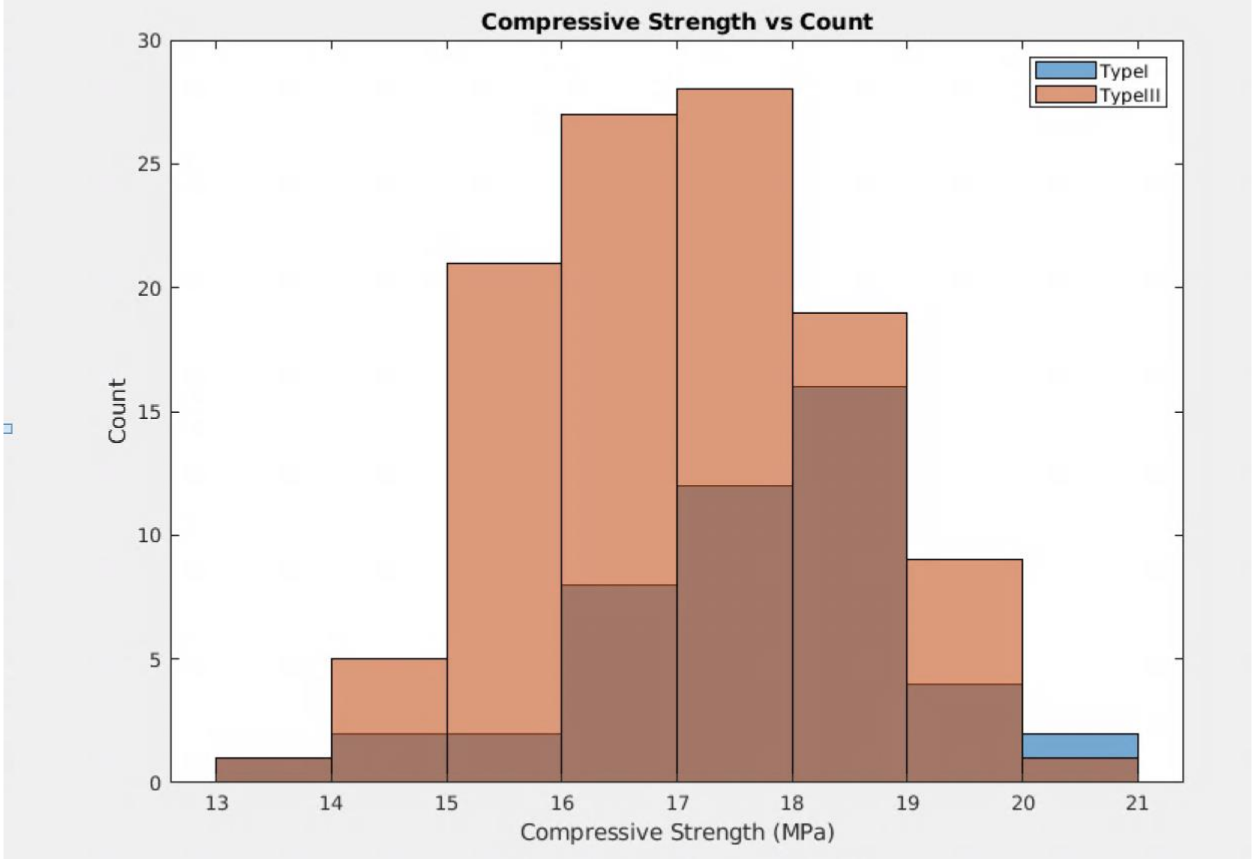

Figure 9. Histogram of compressive strength Type I (2%) and Type III(3%)

In the t-test for a 5% level of significance, the obtained value for h was 1. As a result, the null hypothesis is rejected indicating that there is a significant difference between Type I (2∓1) and Type III (5∓1).

In the t-test for a 1% level of significance, the obtained p-value is 0.0117 which is greater than 0.01 which indicates that there is no significant difference between the two data sets at this level of significance.