STAT 448: Final Project
Analysis of Concrete Compressive Strength Data

1. **Introduction**
   Concrete is the most important and most common material in civil engineering due to its strength. The concrete compressive strength is a highly nonlinear function of age and ingredients. And its properties are affected by those ingredients. So, we now want to understand how components of concrete, age of the concrete and its compressive strength are related. We know that the strength of concrete is known to be related to the ratio of cement and water, and that more cement for the same amount of water should make for stronger cement. For this project we will be attempting regression techniques to predict the concrete strength based on other independent variables and discriminant analysis methods to classify concrete samples into six different age groups. We also will look at the common characteristics by considering clustering technique to observe any differences in compressive strength across the identified groups.
2. **Data Description**
   The dataset that is chosen for the project describes the concrete compressive strength. It was obtained from UCI Machine Learning repository:
   https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
   There are 1030 observations, and 9 attributes is (8 quantitative input variables, and 1 quantitative output variable). But we will be defining an additional variable (agegroup) that describes the groups by age variable. Below is the description of the variables that will be used in this project:
   - cementwater = Cement/Water ratio -- quantitative – Input Variable
   - slagwater = Blast Furnace Slag/Water ratio -- quantitative -- Input Variable
   - flyashwater = Fly Ash/Water ratio -- quantitative -- Input Variable
   - superplasticizerwater = Superplasticizer/Water ratio -- quantitative -- Input Variable
   - coarsewater = Coarse Aggregate/Water ratio -- quantitative -- Input Variable
   - finewater = Fine Aggregate/Water ratio -- quantitative -- Input Variable
   - age (days)-- quantitative -- Input Variable
   - compressivestrength -- quantitative -- MPa -- Output Variable
   - agegroup – qualitative (ordinal) – Output Variable

   As can be seen, each of the concrete component measurements in the data provided is given as a ratio of the densities of the concrete component and water. We can classify the following groups based on age.
   - age < 1 week -> group 1
   - 7 days <= age < 28 days -> group 2
   - 28 days <= age < 56 days -> group 3
   - 56 days <= age < 90 days -> group 4
   - 90 days <= age < 180 days -> group 5
   - 180 days <= age -> group 6
3. **Project Goal**
   The goal of the project is to determine the approximate age if the concrete composition and strength are known. We are also interested in differences in strength for different ages of concrete, differences in strength for different compositions of concrete, and predictability of strength.
4. **Topics**
   Below is the detailed topics(parts) that will be discussed in the project:
   1) General descriptive overview of characteristics of concrete in the data and description of differences across concrete ages.
   2) Identifying any groupings of concrete samples based on component to water ratios and age. Observing any differences in the concrete characteristics, in compressive strength across the identified groups.

3) Building the model that can predict the compressive strength of concrete that is at least 90 days old.
4) Building the model for predicting if concrete that has cured for 90 to 100 days will have a strength of at least 50 MPa.
5) Identifying the approximate age of concrete based on composition and compressive strength. Building the classification model for those age groups based on component to water ratios and compressive strength.

5. Part I

The UNIVARIATE Procedure
Variable: age

Moments			
N	1030	Sum Weights	1030
Mean	45.6621359	Sum Observations	47032
Std Deviation	63.1699116	Variance	3990.43773
Skewness	3.2691774	Kurtosis	12.168989
Uncorrected SS	6253742	Corrected SS	4106160.42
Coeff Variation	138.341999	Std Error Mean	1.96830165

Basic Statistical Measures			
Location		Variability	
Mean	45.66214	Std Deviation	63.16991
Median	28.00000	Variance	3990
Mode	28.00000	Range	364.00000
		Interquartile Range	49.00000

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.590706	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.337291	Pr > D	<0.0100
Cramer-von Mises	W-Sq	24.34705	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	128.9549	Pr > A-Sq	<0.0050

We want to see the general descriptive statistics for variables age and compressivestrength. For age variable, the mean is 45.66, the median is 28, and standard deviation is 63.17. For compressivestrength variables, the mean is 35.82, the median is 34.44, and the standard deviation is 16.70. The skewness is 0.42, it means that it is slightly skewed to the right.

The distribution and probability plots for compressivestrength are also shown below. It can be seen from Shapiro-Wilk test that the distribution is not normal. However, the plots look good, with little deviation.

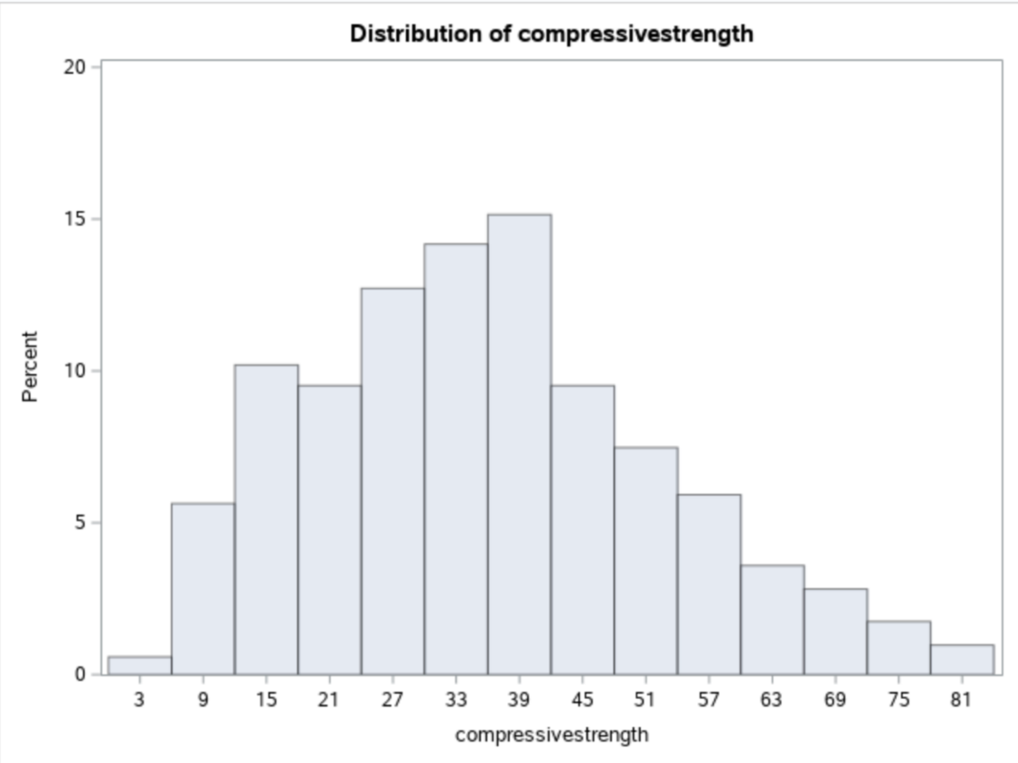
The UNIVARIATE Procedure
Variable: compressivestrength

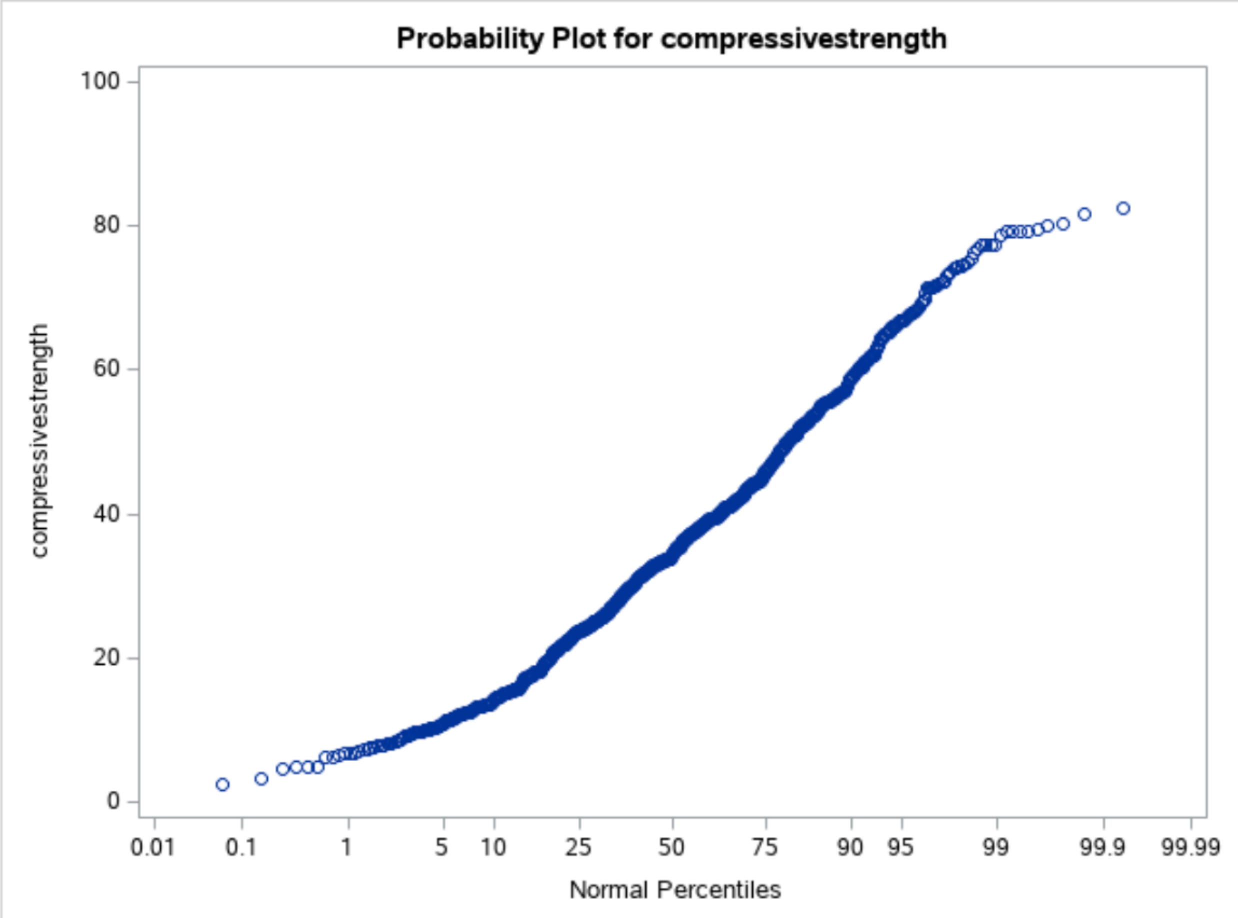
Moments			
N	1030	Sum Weights	1030
Mean	35.8178358	Sum Observations	36892.3709
Std Deviation	16.7056792	Variance	279.079717
Skewness	0.41692228	Kurtosis	-0.3138437
Uncorrected SS	1608577.91	Corrected SS	287173.028
Coeff Variation	46.6406716	Std Error Mean	0.52052971

Basic Statistical Measures			
Location		Variability	
Mean	35.81784	Std Deviation	16.70568
Median	34.44277	Variance	279.07972
Mode	33.39822	Range	80.26742
		Interquartile Range	22.50519

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.979793	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.041303	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.520665	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.068344	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure





The frequency table of different agegroups is shown below. It would be helpful for us in the future parts of the project to compare number of observations in newly identified groups.

The FREQ Procedure

agegroup	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	136	13.20	136	13.20
2	188	18.25	324	31.46
3	425	41.26	749	72.72
4	91	8.83	840	81.55
5	131	12.72	971	94.27
6	59	5.73	1030	100.00

	compressivestrength		
	Mean	Std	N
agegroup			
1	18.84	9.86	136
2	26.94	12.97	188
3	36.75	14.71	425
4	51.89	14.31	91
5	48.24	13.50	131
6	44.16	10.61	59

We can see that all of the three group 3 contains the highest number of samples (425) and group 6 has only 59 samples.

Table above shows the average value of compressivestrength among different age groups. This would be useful to compare in the next parts.

For all pairwise comparisons, we can use Tukey's test and the results are shown in the following table. We can conclude that there are significant differences in average value of compressivestrength between groups 1 and 4, 2 and 4, 3 and 4, 4 and 6, 3 and 5, 2 and 5, 1 and 5, 3 and 6, 2 and 6, 1 and 6, 1 and 2, 1 and 3, 2 and 3, 6 and 3.

Comparisons significant at the 0.05 level are indicated by ***.				
agegroup Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
4 - 5	3.6501	-1.5864	8.8866	
4 - 6	7.7286	1.3148	14.1425	***
4 - 3	15.1416	10.7093	19.5739	***
4 - 2	24.9489	20.0486	29.8492	***
4 - 1	33.0491	27.8522	38.2460	***
5 - 4	-3.6501	-8.8866	1.5864	
5 - 6	4.0785	-1.9378	10.0949	
5 - 3	11.4915	7.6568	15.3261	***
5 - 2	21.2988	16.9316	25.6659	***
5 - 1	29.3990	24.7015	34.0965	***
6 - 4	-7.7286	-14.1425	-1.3148	***
6 - 5	-4.0785	-10.0949	1.9378	
6 - 3	7.4129	2.0818	12.7441	***
6 - 2	17.2202	11.4941	22.9464	***
6 - 1	25.3205	19.3385	31.3024	***
3 - 4	-15.1416	-19.5739	-10.7093	***
3 - 5	-11.4915	-15.3261	-7.6568	***
3 - 6	-7.4129	-12.7441	-2.0818	***
3 - 2	9.8073	6.4462	13.1684	***
3 - 1	17.9075	14.1271	21.6879	***
2 - 4	-24.9489	-29.8492	-20.0486	***
2 - 5	-21.2988	-25.6659	-16.9316	***
2 - 6	-17.2202	-22.9464	-11.4941	***
2 - 3	-9.8073	-13.1684	-6.4462	***
2 - 1	8.1002	3.7806	12.4198	***
1 - 4	-33.0491	-38.2460	-27.8522	***
1 - 5	-29.3990	-34.0965	-24.7015	***
1 - 6	-25.3205	-31.3024	-19.3385	***
1 - 3	-17.9075	-21.6879	-14.1271	***
1 - 2	-8.1002	-12.4198	-3.7806	***

6. **Part II**

We want to apply 'proc cluster' procedure to group the concrete samples based on component to water ratios and age. We chose 6 clusters to see if they have some similarities with previous agegroup variables. (Since it was quite impossible to identify the number of clusters from dendrogram and plots). Below is a table of cluster by agegroup. As can be seen from the table, Cluster 1 contains agegroup = 1,2,3 observations. Cluster 2 contains only agegroup = 4 values, Cluster 3 contains only agegroup = 5 values, Cluster 4, Cluster 5, and Cluster 6 contains samples from agegroup = 6.

The FREQ Procedure

Frequency

Table of CLUSTER by agegroup

CLUSTER	agegroup 1	2	3	4	5	6	Total
1	136	188	425	0	0	0	749
2	0	0	0	91	0	0	91
3	0	0	0	0	131	0	131
4	0	0	0	0	0	26	26
5	0	0	0	0	0	20	20
6	0	0	0	0	0	13	13
Total	136	188	425	91	131	59	1030

It is possible to observe the mean values for each concrete characteristics, age and compressive strength values across the identified groups.

The MEANS Procedure

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	749	1.5545208	0.6556858	0.5312500	3.7468265
slagwater	749	0.4289133	0.4831269	0	1.9353796
flyashwater	749	0.3201714	0.3738530	0	1.3456263
superplasticizerwater	749	0.0376324	0.0376203	0	0.2336720
coarsewater	749	5.4313112	0.8141631	3.4534413	8.6956879
finewater	749	4.3399835	0.7614955	2.6052632	7.8404423
age	749	18.7636849	10.9097643	1.0000000	28.0000000
compressivestrength	749	31.0352702	15.2336352	2.3318078	81.7511693

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	91	1.7822699	0.6615712	0.9412331	3.7468265
slagwater	91	0.3388049	0.4411470	0	1.5386289
flyashwater	91	0.5098146	0.3683491	0	1.3456263
superplasticizerwater	91	0.0617891	0.0382558	0	0.2336720
coarsewater	91	5.9157502	0.7768427	4.0895522	8.6956879
finewater	91	4.8370210	0.7753905	3.3285714	7.8404423
age	91	56.0000000	0	56.0000000	56.0000000
compressivestrength	91	51.8900612	14.3084945	23.2451909	80.1998483

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	131	1.5668620	0.6238590	0.5312500	3.7468265
slagwater	131	0.3895853	0.4625762	0	1.5002457
flyashwater	131	0.2795270	0.3818017	0	1.3456263
superplasticizerwater	131	0.0359877	0.0437053	0	0.2336720
coarsewater	131	5.5328662	0.8656055	4.0877193	8.6956879
finewater	131	4.4549742	0.9009696	2.6052632	7.8404423
age	131	94.8244275	6.1224441	90.0000000	120.0000000
compressivestrength	131	48.2399568	13.4960605	21.8591471	82.5992248

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	26	1.6156645	0.5464462	0.7270833	3.1213873
slagwater	26	0.2271234	0.3521260	0	1.0906250
flyashwater	26	0	0	0	0
superplasticizerwater	26	0	0	0	0
coarsewater	26	4.7846621	0.6708764	4.0877193	6.5028902
finewater	26	3.5529123	0.7222981	2.6052632	4.5854922
age	26	180.0000000	0	180.0000000	180.0000000
compressivestrength	26	41.7303758	10.9297302	24.1040810	71.6227669

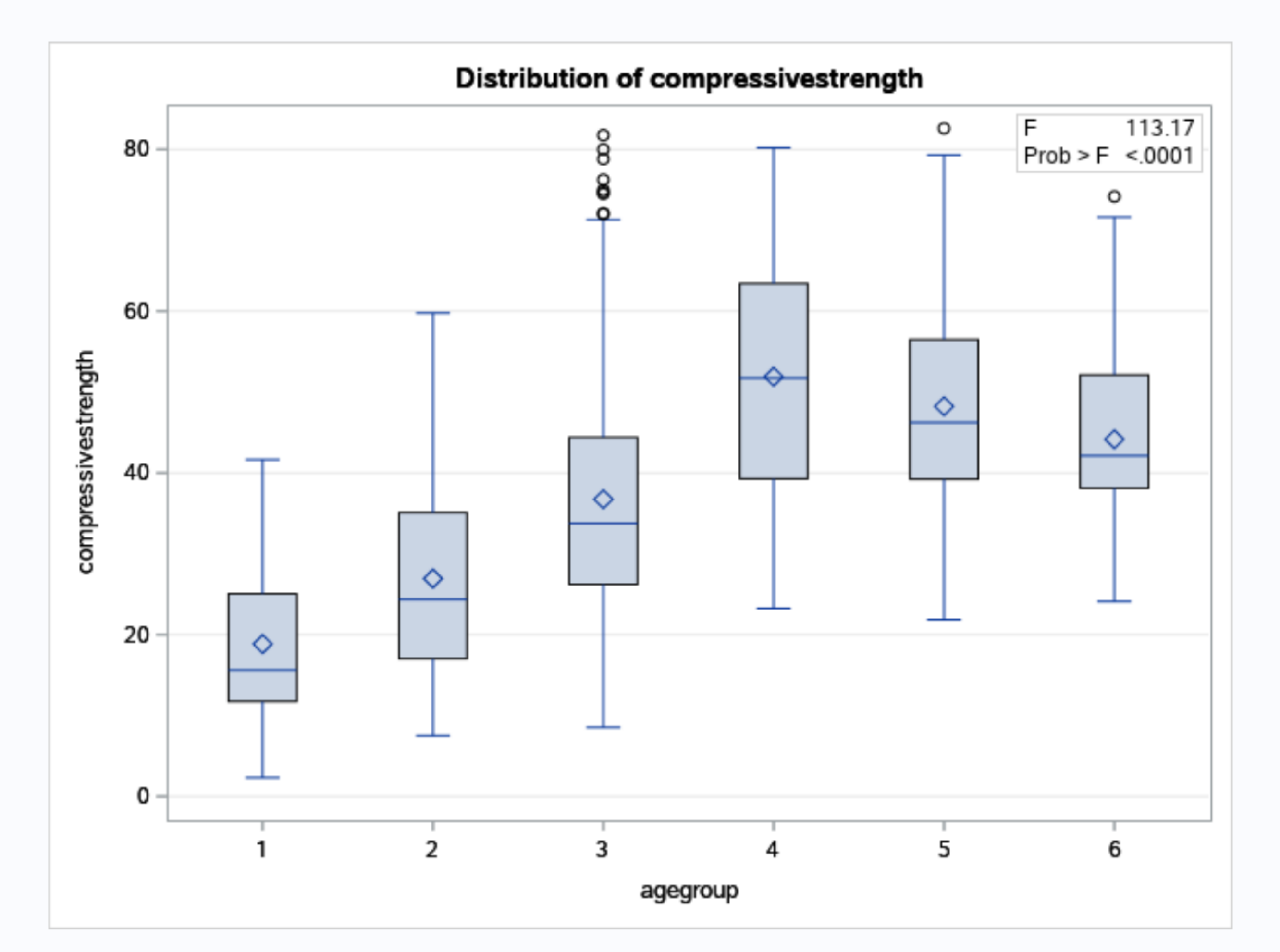
CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	20	1.4405041	0.3570856	0.7270833	2.0833333
slagwater	20	0.2952604	0.3767908	0	1.0906250
flyashwater	20	0	0	0	0
superplasticizerwater	20	0	0	0	0
coarsewater	20	4.5963428	0.5390057	4.0877193	5.4531250
finewater	20	3.5370365	0.8033054	2.6052632	4.5854922
age	20	363.5000000	2.3508117	360.0000000	365.0000000
compressivestrength	20	42.6995589	8.3489640	25.0831369	56.1419623

CLUSTER=6

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	13	1.7710896	0.6854345	0.8333333	3.1213873
slagwater	13	0.3173077	0.3496602	0	1.0416667
flyashwater	13	0	0	0	0
superplasticizerwater	13	0	0	0	0
coarsewater	13	4.5351895	0.8693992	4.0877193	6.5028902
finewater	13	2.8900798	0.2886116	2.6052632	3.5433526
age	13	270.0000000	0	270.0000000	270.0000000
compressivestrength	13	51.2725115	10.6446660	38.4079501	74.1669333

Cluster 1 has the smallest number of days = 19, Cluster 2 has 56 days, Cluster 3 has 95 days, Cluster 4 180 days, Cluster 5 has 363.5 days, and Cluster 6 has 270 days. It seems that Cluster 1 matches with agegroup = 2, Cluster 2 the same as agegroup = 3, Cluster 3 is very close to agegroup = 4, Cluster 4 the same as agegroup = 5, Clusters 5 and 6 match agegroup = 6. However, our clusters did not include samples with age<7 days on average. It might be possible that we may have very small number of observations. Composition ratio to water values have similar values across the groups.
Similar results can be observed from the box plot below.



7. Part III

The CORR Procedure

Pearson Correlation Coefficients, N = 190
Prob > |r| under H0: Rho=0

	cementwater	slagwater	flyashwater	superplasticizerwater	coarsewater	finewater	age	compressivestrength
cementwater	1.00000	-0.15506 0.0327	-0.27261 0.0001	0.43420 <.0001	0.23269 0.0012	0.13751 0.0585	-0.02733 0.7082	0.59261 <.0001
slagwater	-0.15506 0.0327	1.00000	-0.29273 <.0001	0.16811 0.0204	-0.10551 0.1474	-0.06500 0.3729	-0.11136 0.1261	0.35336 <.0001
flyashwater	-0.27261 0.0001	-0.29273 <.0001	1.00000	0.36609 <.0001	0.63809 <.0001	0.45847 <.0001	-0.29477 <.0001	0.03501 0.6315
superplasticizerwater	0.43420 <.0001	0.16811 0.0204	0.36609 <.0001	1.00000	0.63325 <.0001	0.70693 <.0001	-0.35453 <.0001	0.57475 <.0001
coarsewater	0.23269 0.0012	-0.10551 0.1474	0.63809 <.0001	0.63325 <.0001	1.00000	0.75117 <.0001	-0.39265 <.0001	0.36485 <.0001
finewater	0.13751 0.0585	-0.06500 0.3729	0.45847 <.0001	0.70693 <.0001	0.75117 <.0001	1.00000	-0.42681 <.0001	0.16526 0.0227
age	-0.02733 0.7082	-0.11136 0.1261	-0.29477 <.0001	-0.35453 <.0001	-0.39265 <.0001	-0.42681 <.0001	1.00000	-0.11673 0.1087
compressivestrength	0.59261 <.0001	0.35336 <.0001	0.03501 0.6315	0.57475 <.0001	0.36485 <.0001	0.16526 0.0227	-0.11673 0.1087	1.00000

Before we build the linear regression model, we could observe the relationship between independent and dependent(strength) variables by applying ‘proc corr’. There is a positive relationship between cementwater and strength (0.59), superplasticizerwater and strength (0.57), and slightly small values for slagwater and coarsewater.

We want to predict the compressive strength of concrete that is at least 90 days old, so we will consider a subset our dataset. We got 190 observations out of 1030. We will apply ‘proc reg’ model and use stepwise selection to find out the best predictors. In the output of stepwise selection, it is clear that four variables cementwater, slagwater, flyashwater and finewater were selected into our final model, and they are all significant.

Bounds on condition number: 1.2448, 10.722

Stepwise Selection: Step 4

Variable finewater Entered: R-Square = 0.7419 and C(p) = 4.5611

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	22755	5688.63807	131.51	<.0001
Error	183	7915.65659	43.25495		
Corrected Total	187	30670			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.93518	2.47714	1368.85903	31.65	<.0001
cementwater	18.85077	0.93842	17454	403.52	<.0001
slagwater	18.72679	1.21286	10312	238.40	<.0001
flyashwater	20.56733	1.88654	5141.11644	118.86	<.0001
finewater	-1.69531	0.59238	354.26579	8.19	0.0047

Bounds on condition number: 1.708, 22.307

All variables left in the model are significant at the 0.0500 level.

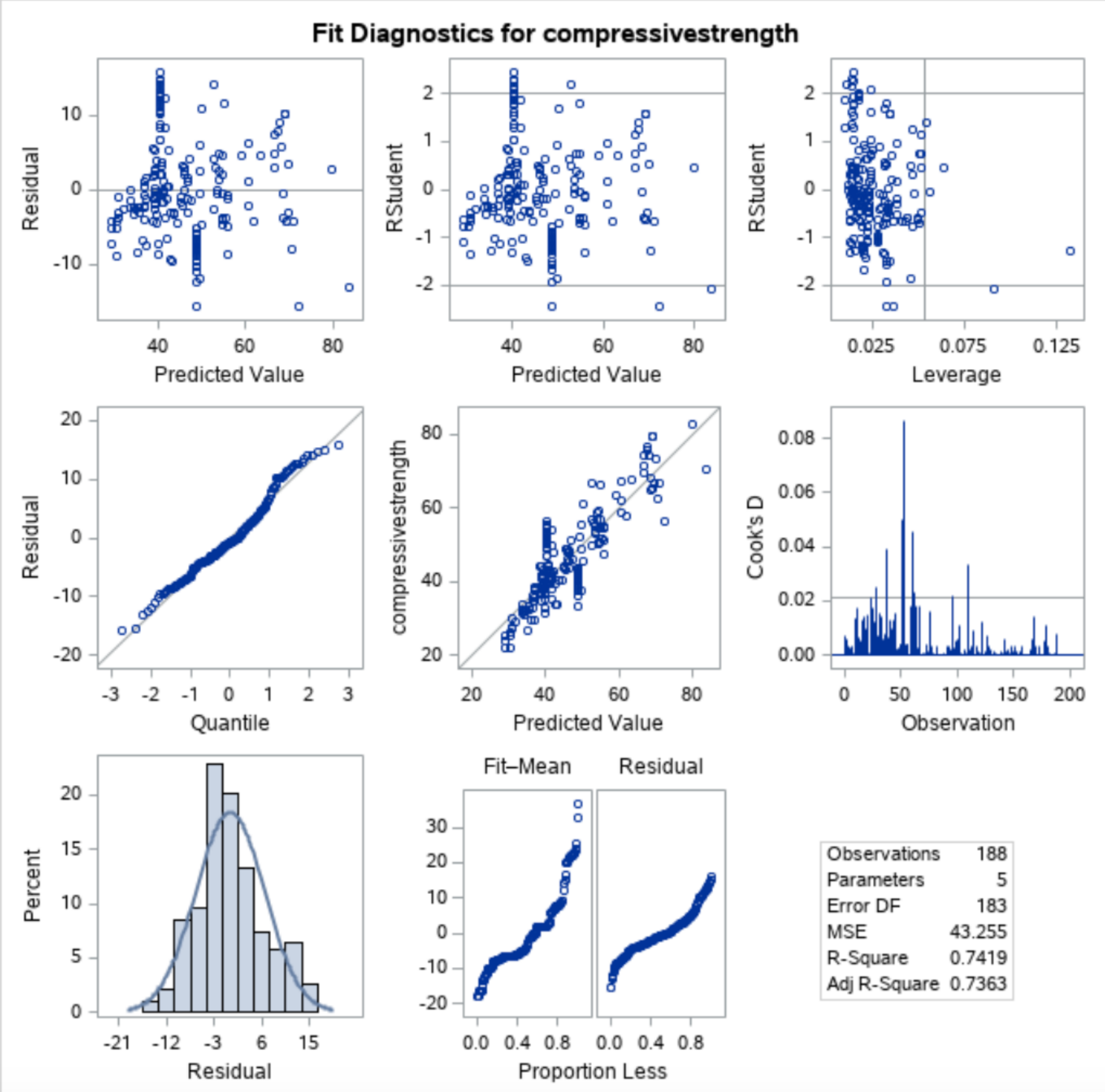
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	cementwater		1	0.3560	0.3560	271.514	102.83	<.0001
2	slagwater		2	0.2029	0.5589	129.979	85.11	<.0001
3	flyashwater		3	0.1714	0.7304	10.7316	116.97	<.0001
4	finewater		4	0.0116	0.7419	4.5611	8.19	0.0047

As a first step, we observed diagnostics and found out that we can remove unduly influential points based on Cook's Distance that is higher than 0.08. Hence, our data was reduced to 188 observations. Above is the output for the final model of linear regression.
Our final model gave us R-squared of 74.19%, which is the percentage of variation explained by our model. It is not so bad, hence can be considered as a good model. Our model is statistically significant in general. This model has expected increases of 18.85, 18.73, 20.57 and -1.69 in compressivestrength for one unit increases in cementwater, slagwater, flyashwater and finewater, respectively.

Below is the diagnostics for our model. The residual plots look fine, QQ plot and histogram also perfectly aligned with our assumptions.

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength



8. **Part IV**
Before we build the linear regression model, we similarly could observe the relationship between independent and dependent(strength) variables by applying ‘proc corr’. There is a positive relationship between cementwater and strength (0.53), superplasticizerwater and strength (0.39), slagwater and strength (0.47).

The CORR Procedure

Pearson Correlation Coefficients, N = 51
Prob > |r| under H0: Rho=0

	cementwater	slagwater	flyashwater	superplasticizerwater	coarsewater	finewater	age	compressivestrength
cementwater	1.00000	0.01555 0.9138	-0.61897 <.0001	0.54668 <.0001	0.04327 0.7631	0.36465 0.0085	-0.61430 <.0001	0.53076 <.0001
slagwater	0.01555 0.9138	1.00000	-0.53092 <.0001	0.08304 0.5624	-0.19794 0.1638	-0.09247 0.5187	-0.51348 0.0001	0.47337 0.0005
flyashwater	-0.61897 <.0001	-0.53092 <.0001	1.00000	-0.03454 0.8099	0.48353 0.0003	0.16151 0.2575	0.90719 <.0001	-0.42506 0.0019
superplasticizerwater	0.54668 <.0001	0.08304 0.5624	-0.03454 0.8099	1.00000	0.48763 0.0003	0.82425 <.0001	0.01075 0.9403	0.39457 0.0042
coarsewater	0.04327 0.7631	-0.19794 0.1638	0.48353 0.0003	0.48763 0.0003	1.00000	0.55980 <.0001	0.50547 0.0002	0.19013 0.1814
finewater	0.36465 0.0085	-0.09247 0.5187	0.16151 0.2575	0.82425 <.0001	0.55980 <.0001	1.00000	0.21938 0.1219	0.26351 0.0617
age	-0.61430 <.0001	-0.51348 0.0001	0.90719 <.0001	0.01075 0.9403	0.50547 0.0002	0.21938 0.1219	1.00000	-0.38964 0.0047
compressivestrength	0.53076 <.0001	0.47337 0.0005	-0.42506 0.0019	0.39457 0.0042	0.19013 0.1814	0.26351 0.0617	-0.38964 0.0047	1.00000

Bounds on condition number: 1.0447, 4.1789

Stepwise Selection: Step 3

Variable coarsewater Entered: R-Square = 0.6955 and C(p) = 5.0304

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2847.06067	949.02022	32.73	<.0001
Error	43	1246.61533	28.99105		
Corrected Total	46	4093.67600			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	21.52283	6.78969	291.31486	10.05	0.0028
cementwater	7.04607	1.27346	887.54459	30.61	<.0001
slagwater	11.51195	1.71235	1310.31673	45.20	<.0001
coarsewater	3.51588	1.01992	344.51057	11.88	0.0013

Bounds on condition number: 1.0643, 9.4116

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

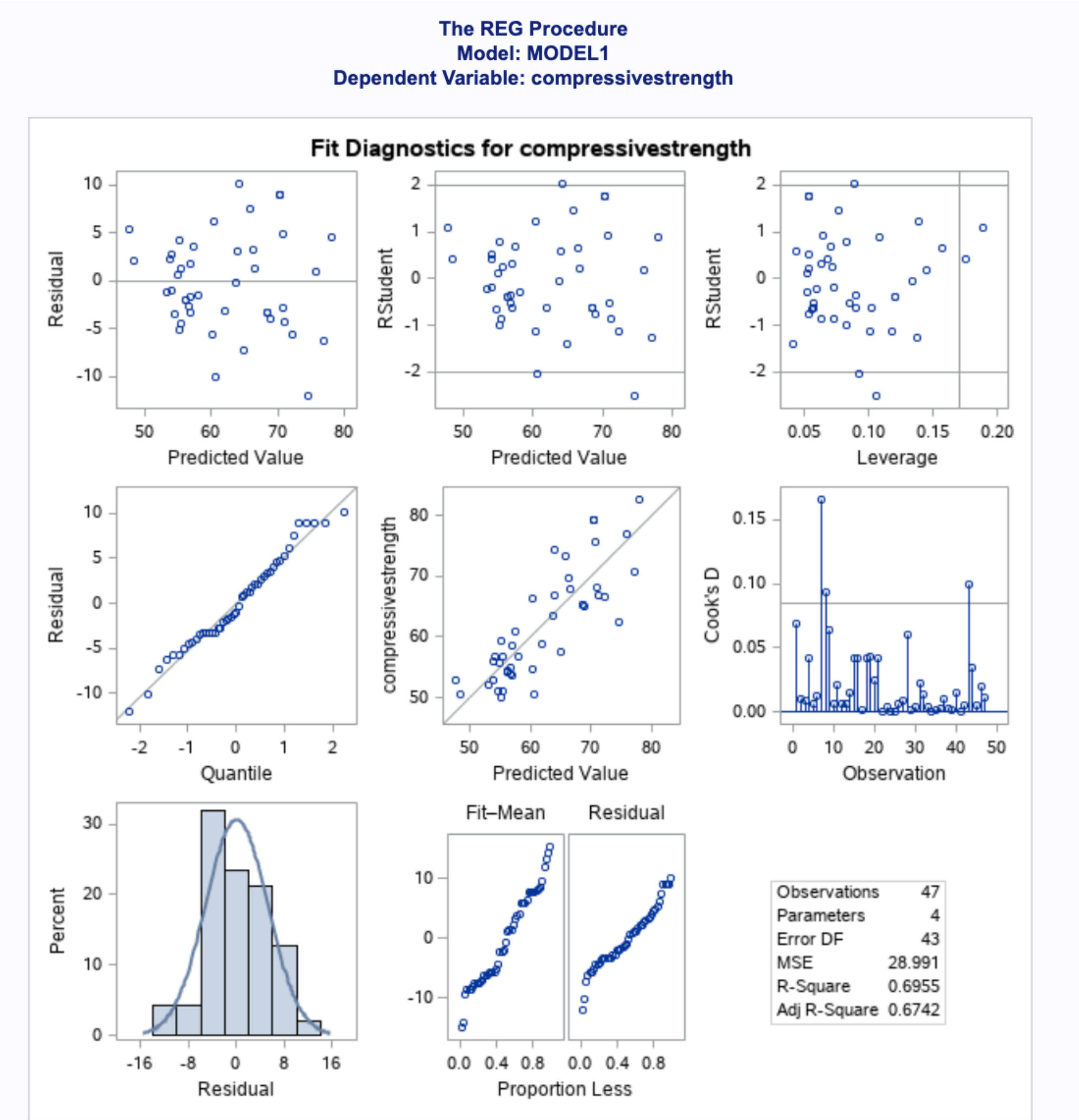
Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	slagwater		1	0.4077	0.4077	42.6377	30.98	<.0001
2	cementwater		2	0.2036	0.6113	15.1986	23.05	<.0001
3	coarsewater		3	0.0842	0.6955	5.0304	11.88	0.0013

Next, we want to predict if concrete that has cured for 90 to 100 days will have a strength of at least 50 MPa. In a similar fashion, we will consider a subset of data. We got 51 observations out of 1030. We will apply ‘proc reg’ model and use stepwise selection to find out the best predictors. In the output of stepwise selection, it is clear that four variables slagwater, cementwater, coarsewater were selected into our final model, and they are all significant.

As a first step, we observed diagnostics and found out that we can remove unduly influential points based on Cook’s Distance that is higher than 0.08. Hence, our data was reduced to 47 observations. Above is the output for the final model of linear regression.
Our final model gave us R-squared of 69.55%, which is the percentage of variation explained by our model. It is slightly smaller than our previous model. Our model is statistically significant in general. This model has expected increases of 7.05, 11.51, and 3.56 in compressivestrength for one unit increases in cementwater, slagwater, and coarsewater, respectively.

Below is the diagnostics for our model. The residual plots and QQ plot look fine, however histogram plot is not perfect, due to the small number of observations, possibly.



9. **Part V**

In this part, we will apply 'proc discrim' procedure to classify our observations by age groups. First, 'proc stepdisc' will be applied to determine the best predictors that will be used in the model. Below is the output of the selection summary. We can see that all the composition (ratio to water) variables should be used in the classification model.

Next, we will perform test of Homogeneity of within Covariance Matrix to identify which kind of discriminant function should be applied. Since the Chi-Square value (2855.01) is significant at the 0.05 level, so Quadratic Discriminant analysis should be used.

The STEPDISC Procedure

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	compressivestrength		0.3559	113.17	<.0001	0.64409164	<.0001	0.07118167	<.0001
2	2	cementwater		0.2988	87.19	<.0001	0.45162892	<.0001	0.11199218	<.0001
3	3	slagwater		0.1699	41.85	<.0001	0.37487606	<.0001	0.12909978	<.0001
4	4	flyashwater		0.2271	59.99	<.0001	0.28974765	<.0001	0.16227206	<.0001
5	5	finewater		0.0734	16.17	<.0001	0.26847098	<.0001	0.17602881	<.0001
6	6	superplasticizerwater		0.0324	6.82	<.0001	0.25977622	<.0001	0.18077211	<.0001
7	7	coarsewater		0.0200	4.16	0.0009	0.25456903	<.0001	0.18407300	<.0001

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
2855.013554	140	<.0001

The DISCRIM Procedure

Multivariate Statistics and F Approximations					
S=5 M=0.5 N=508					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.25456903	47.06	35	4284.8	<.0001
Pillai's Trace	0.92036501	32.94	35	5110	<.0001
Hotelling-Lawley Trace	2.28333577	66.33	35	2884.8	<.0001
Roy's Greatest Root	1.99063872	290.63	7	1022	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

Next, we can read from the above MANOVA that all tests are highly statistically significant, indicating that there are significant differences in the means of at least some of these measurements across at least some of the species, and so discriminant analysis may be able to do a good job of classifying at least some species.

We will be applying cross-validation and 'priors proportional' in the analysis.

Below is the result of our classification model with cross-validation. We can see that overall error rate is estimated to be around 50%. The misclassification error rates for Gr.1, Gr.2, Gr.3, Gr.4, Gr.5 and Gr.6 are 46%, 70%, 37.41, 65%, 76.34% and 0%, respectively.

We can see that 54.41% of Gr.1, 29.26% of Gr.2, 63% of Gr.3, 35.16% of Gr.4, 24% of Gr.5 and 100% of Gr.6 were correctly classified by the model. However, one can note that Gr6. contains some other observations from all groups, even though it has 0% of error rate. It may be observed that agegroups 2, 3, 4 and 6 might be very difficult to classify.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CONCRETERATS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into agegroup							
From agegroup	1	2	3	4	5	6	Total
1	74 54.41	16 11.76	3 2.21	0 0.00	0 0.00	43 31.62	136 100.00
2	11 5.85	55 29.26	49 26.06	4 2.13	0 0.00	69 36.70	188 100.00
3	0 0.00	33 7.76	266 62.59	32 7.53	18 4.24	76 17.88	425 100.00
4	0 0.00	2 2.20	23 25.27	32 35.16	32 35.16	2 2.20	91 100.00
5	0 0.00	0 0.00	11 8.40	36 27.48	31 23.66	53 40.46	131 100.00
6	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	59 100.00	59 100.00
Total	85 8.25	106 10.29	352 34.17	104 10.10	81 7.86	302 29.32	1030 100.00
Priors	0.13204	0.18252	0.41262	0.08835	0.12718	0.05728	

Error Count Estimates for agegroup							
	1	2	3	4	5	6	Total
Rate	0.4559	0.7074	0.3741	0.6484	0.7634	0.0000	0.4981
Priors	0.1320	0.1825	0.4126	0.0883	0.1272	0.0573	

10. Conclusion

In this project, we described the general characteristics of some of the necessary variables for further analysis. Secondly, we considered clustering procedure to group concrete samples based on age and compressivestrength. Moreover, we had a chance to build linear regression models to predict the compressive strength for concrete. Finally, we constructed classification model to predict age groups for the samples based on compositions and strength.

From the first part we concluded that there are significant differences in average value of compressivestrength between groups 1 and 4, 2 and 4, 3 and 4, 4 and 6, 3 and 5, 2 and 5, 1 and 5, 3 and 6, 2 and 6, 1 and 6, 1 and 2, 1 and 3, 2 and 3, 6 and 3. The second part suggests us that Cluster 1 has compressive strength of 31.03 MPa, Cluster 2 has 51.89 MPa, Cluster 3 has 48.24 MPa, Cluster 4 has 41.73 MPa, Cluster 5 has 42.7 MPa and Cluster 6 has 51.27 MPa. Parts 3 and 4 involved two linear regression models for two different subsets of data. First model had R-2 of around 74.19% and the second model had 69.55%. The difference is not large, but both can be improved by applying some transformation methods on variables or considering other factors/variables. The last classification model did not give a satisfactory result in general, with overall misclassification error 49.81%. As can be seen from the tables, all the groups contained samples from other groups, and all of the group's values got confused and can be easily misclassified. Mainly Group 6 contained the observations from all the groups. Hence, quadratic discriminant analysis gave us confusing model and is not really useful to classify the samples based on composition and strength.

Finally, in order to achieve better results, one could collect more data and do further analysis to be able to correctly classify data and effectively predict compressive strength.