Take-Home Final Project

Yijin Wang

Introduction

This project adopts a modified Concrete Compressive Strength Data Set which was originally obtained from UCI Machine Learning Repository. The modified dataset includes nine variables: six of them are the component/water ratios obtained from dividing the concrete components by the water content - cement/water ratio, blast furnace slag/water ratio, fly ash/water ratio, superplasticizer/water ratio, coarse aggregate/water ratio, and fine aggregate/water ratio; one of them is the age of the concrete in days; one of them is the measure of concrete compressive strength in MPa – megapascals; and the one left is the age group that each individual observation in. We define the age group as: group 1 if age < 7; group 2 if 7 <= age < 28; group 3 if 28 <= age < 56; group 4 if 56 <= age < 90; group 5 if 90 <= age < 180; and group 6 otherwise (i.e., if age >= 180).

This report will address five topics, including but not limited to: provide general overview of the concrete's characteristics; identifying groupings of concrete samples based on all the information and the dataset; predicting compressive strength of concrete under specific age and other conditions; and classifying concrete samples into some known age groups. I will now address each individual question through the SAS programming software below, and then provide a general conclusion on the dataset at the end of the report.

Client Questions:

1. **Provide a general descriptive overview of characteristics of concrete in the data and make sure to describe differences across concrete ages.**

I sorted the data by age group first so we can do further analysis later. The diagnostic tables (**Figure 1)** I created below contain the basic statistical summary for seven variables – six component-water ratios and one for compressive strength, per each age group. The table included the number of observations, mean value, standard deviation, minimum, and maximum, for each of the seven variables per age group.

As we can see, cement-water ratio has highest mean in group 4 - 1.7822699; slag-water ratio has highest mean in group 3 - 0.4700745; fly ash-water has highest mean in group 4 - 0.5098146; superplasticizer-water has highest mean in group 4 - 0.0617891; coarse-water has highest mean in group 4 - 5.9157502; and fine-water has highest mean in group 4 - 4.8370210. In summary, the observations in age group 4 almost have all the highest mean component-water ratio compared to other age groups. Moreover, the compressive strength has the highest mean value in age group 4 as well, which could be explained by the known relation between the strength of concrete and the ratio of cement and water.

Finally, since the age group variable is defined by concrete age in ascending order, there seems to be an increasing trend for the variable compressive strength as the age increase. Although there is a decrease in age group 5 and 6, the first four age groups perfectly show the trend.

The MEANS Procedure

agegroup=1

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	136	1.6670945	0.6682539	0.5312500	3.7468265
slagwater	136	0.3688792	0.4635181	0	1.5386289
flyashwater	136	0.3375402	0.3861533	0	1.3456263
superplasticizerwater	136	0.0408767	0.0428485	0	0.2336720
coarsewater	136	5.6178401	0.7899494	4.0895522	8.6956879
finewater	136	4.5750836	0.7866970	3.0650000	7.8404423
compressivestrength	136	18.8409587	9.8644153	2.3318078	41.6374556

agegroup=2

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	188	1.6409523	0.6314903	0.5312500	3.7468265
slagwater	188	0.3792917	0.5005638	0	1.9353796
flyashwater	188	0.2441780	0.3613424	0	1.3456263
superplasticizerwater	188	0.0295704	0.0407623	0	0.2336720
coarsewater	188	5.6026605	0.7368399	4.0877193	8.6956879
finewater	188	4.3969597	0.7897588	2.6052632	7.8404423
compressivestrength	188	26.9411856	12.9660966	7.5070147	59.7637797

agegroup=3

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	425	1.4802639	0.6538234	0.5312500	3.7468265
slagwater	425	0.4700745	0.4784474	0	1.9353796
flyashwater	425	0.3482293	0.3715628	0	1.3456263
superplasticizerwater	425	0.0401605	0.0337756	0	0.2336720
coarsewater	425	5.2958251	0.8298615	3.4534413	8.6956879
finewater	425	4.2395479	0.7221993	2.6052632	7.8404423
compressivestrength	425	36.7484803	14.7112108	8.5357129	81.7511693

agegroup=4

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
compressivestrength	91	51.8900612	14.3084945	23.2451909	80.1998483

agegroup=5

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
compressivestrength	131	48.2399568	13.4960605	21.8591471	82.5992248

agegroup=6

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	59	1.5905343	0.5320103	0.7270833	3.1213873
slagwater	59	0.2700918	0.3560820	0	1.0906250
flyashwater	59	0	0	0	0
superplasticizerwater	59	0	0	0	0
coarsewater	59	4.6658565	0.6762949	4.0877193	6.5028902
finewater	59	3.4014829	0.7271430	2.6052632	4.5854922
compressivestrength	59	44.1614169	10.6060816	24.1040810	74.1669333

Figure 1

Next, I made six Histogram plots (**Figure 2)** with Normal Distribution test below, each per age group, to see the data distributions visually. We see that age group 1, 2, 5 have left toward/positive skewness, which means mean > median > mode in these age groups. On the other hand, age group 3, 4, 6 have well-shaped distributions, which means mean, median, and mode have similar values.

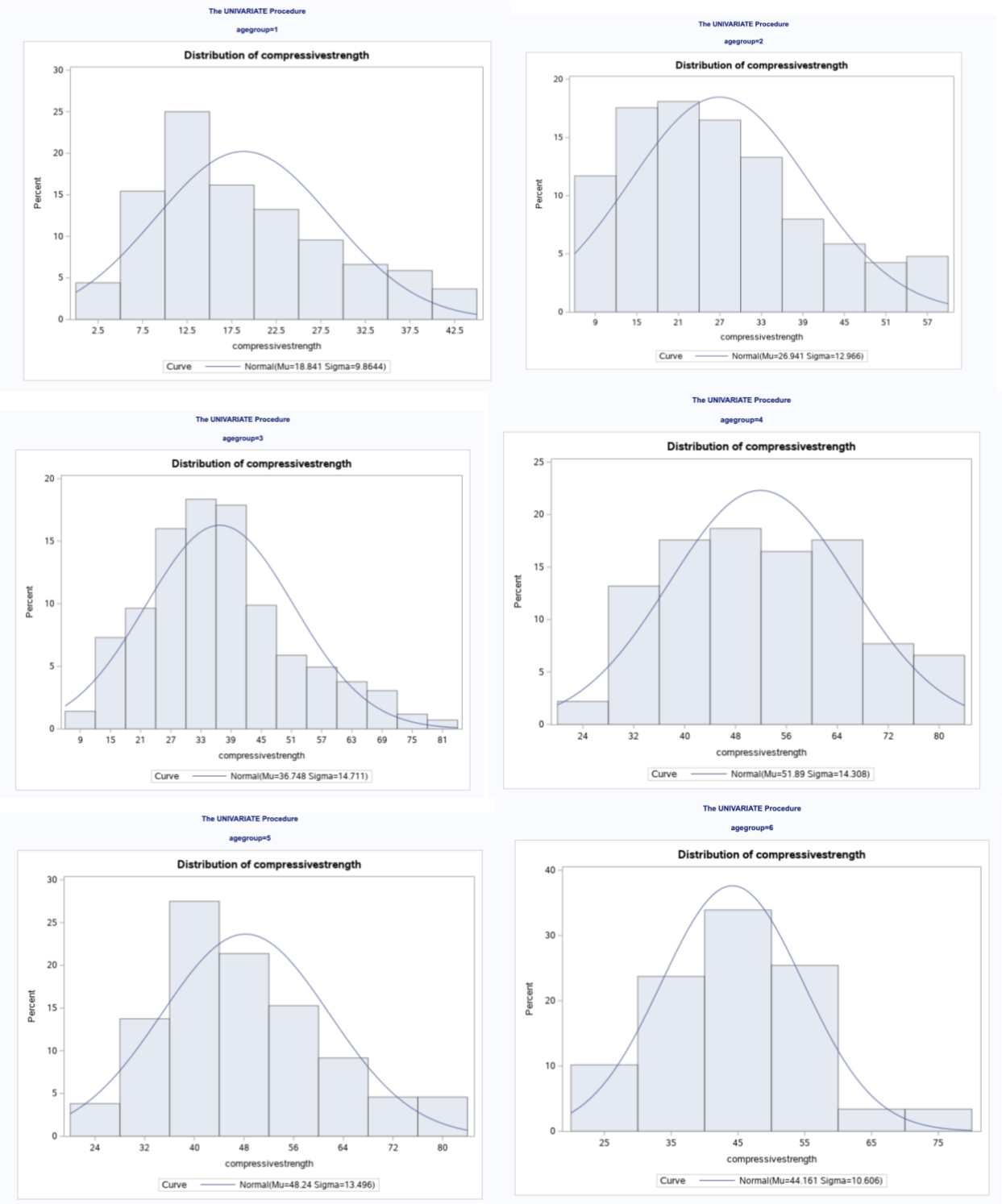


Figure 2

Finally, I made a scatter plot with a linear regression line (**Figure 3)** for us to see the relationship between the compressive strength and the concrete age clearly. Obviously from the plot, there is a positive relationship between the age and compressive strength. (i.e., as age increase, the compressive strength goes larger and larger.)

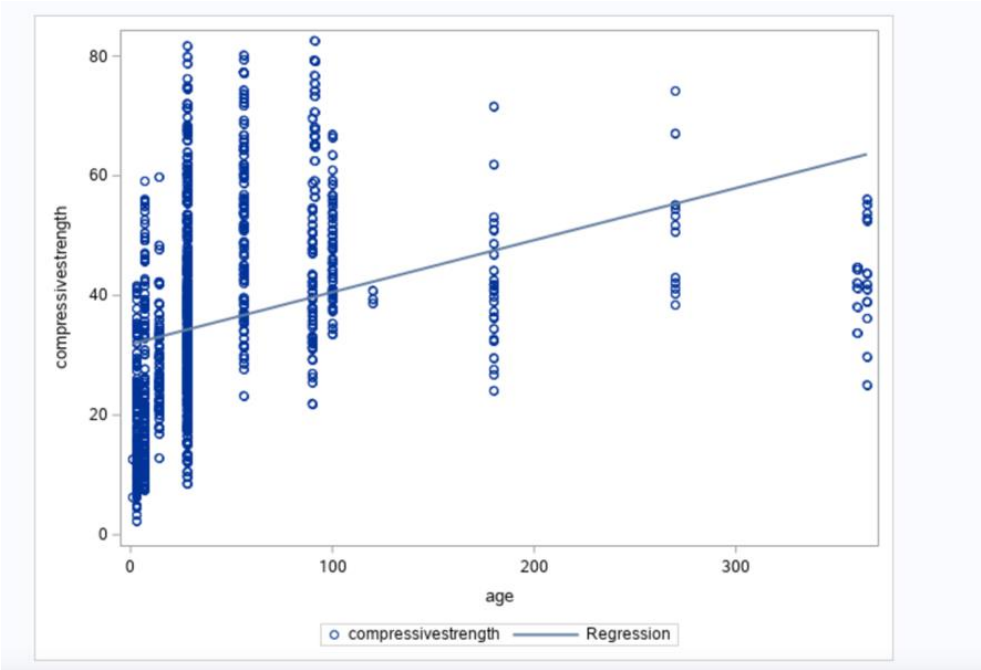


Figure 3

Summary:
Based on all the tables and plots above, we conclude that there is a positive relationship between age and compressive strength. Moreover, the group 4 has the highest compressive strength and that could strongly be explained by the high component-water ratios of the observations in that group.

2. **Identify any groupings of concrete samples based on component to water ratios and age. Explain any interesting differences in the concrete characteristics across the groups and determine if there are any noteworthy differences in compressive strength across the identified groups.**

I first cluster the dataset by the CLUSTER procedure in SAS and only print the first 15 rows for us to visually see the clustering and related information (**Figure 1)**. After tried three different linkages: average, single, and complete, I found that the average linkage has overall better Criteria for the Number of Clusters plot (**Figure 2)**, which is attached below. From the plot, we see that there is a highest CCC (Cubic Clustering Criterion) value with 10 clusters, and a high Pseudo F statistics value for 10 clusters with a relatively large decline in F for 11 which therefore suggests 10 as a possible choice of the number of clusters. Also, there is a relatively low Pseudo t^2 value at 10. Therefore, based on all three criteria, I chose 10 as the number of clusters, and did some analysis on these 10 clusters.

The CLUSTER Procedure
Average Linkage Cluster Analysis

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
15	CL21	OB36	136	0.0000	1.00	.995	53.5	16E4	6.9	0.045	
14	CL27	OB148	126	0.0000	1.00	.995	56.2	17E4	9.3	0.0476	
13	CL24	CL20	425	0.0000	1.00	.994	59.1	18E4	15.3	0.0493	
12	CL15	CL14	262	0.0003	.999	.993	52.4	13E4	541	0.0506	
11	CL22	OB857	76	0.0000	.999	.991	56.2	14E4	7.9	0.0507	
10	CL49	CL96	20	0.0000	.999	.990	59.6	15E4	141	0.0595	
9	CL12	CL17	324	0.0010	.998	.987	45.0	72E3	798	0.1066	
8	CL11	CL18	128	0.0007	.997	.984	42.6	58E3	1526	0.112	
7	CL9	CL13	749	0.0204	.977	.979	-2.5	7270	9336	0.2442	
6	CL8	CL643	131	0.0005	.977	.972	4.15	8551	77.8	0.2938	
5	CL7	CL16	840	0.0274	.949	.960	-5.6	4790	1039	0.4349	
4	CL5	CL6	971	0.1431	.806	.937	-28	1422	2733	0.8276	
3	CL25	CL35	39	0.0171	.789	.889	-17	1920	5E4	1.0074	
2	CL3	CL10	59	0.0758	.713	.750	-4.4	2556	252	1.7825	
1	CL4	CL2	1030	0.7132	.000	.000	0.00	.	2556	2.7405	

Figure 1

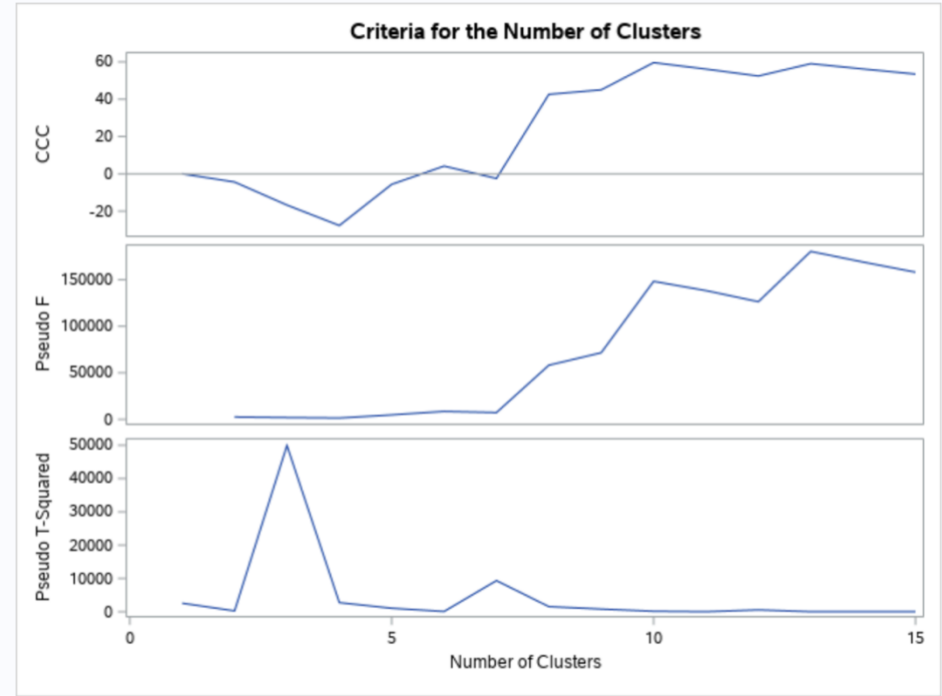


Figure 2

The first analysis diagnostic I did is the ANOVA test (**Figure 3)** with the compressive strength as the dependent variable. The test gives us a large F value (60.78) with small p-value (<.0001), which means there is strong evidence to reject the null hypothesis and there are significant differences of the compressive strength values among the ten clusters. Moreover, I used the Levene's Homogeneity Test for the response variable variance and concluded that there are significant differences of the compressive strength values between the group means as well.

Next, I created a cross-frequency table for us to see how many observations in each cluster per age groups (**Figure 4)**. The table shows cluster 1 has about half age group 1 (136) and half age group 2 (126); cluster 2 only contain age group 3 (425); cluster 3 only contain age group 4 (91); cluster 4 only contain age group 5 (76); cluster 5 also only contain age group 2 (62); cluster 6 also only contain age group 5 (52); cluster 7 only contain age group 6 (26); cluster 8 also only contain age group 5 (3); cluster 9 only contain age group 6 as well(20); and cluster 10 only contain 13 of the age group 6. By the total number of each cluster, we see that they are well-separated.

The ANOVA Procedure

Dependent Variable: compressivestrength

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	100243.3659	11138.1518	60.78	<.0001
Error	1020	186929.6625	183.2644		
Corrected Total	1029	287173.0285			

The ANOVA Procedure

Levene's Test for Homogeneity of compressivestrength Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
CLUSTER	9	3028281	336476	4.97	<.0001
Error	1020	69100475	67745.6		

Figure 3

The FREQ Procedure

Frequency

Table of CLUSTER by agegroup

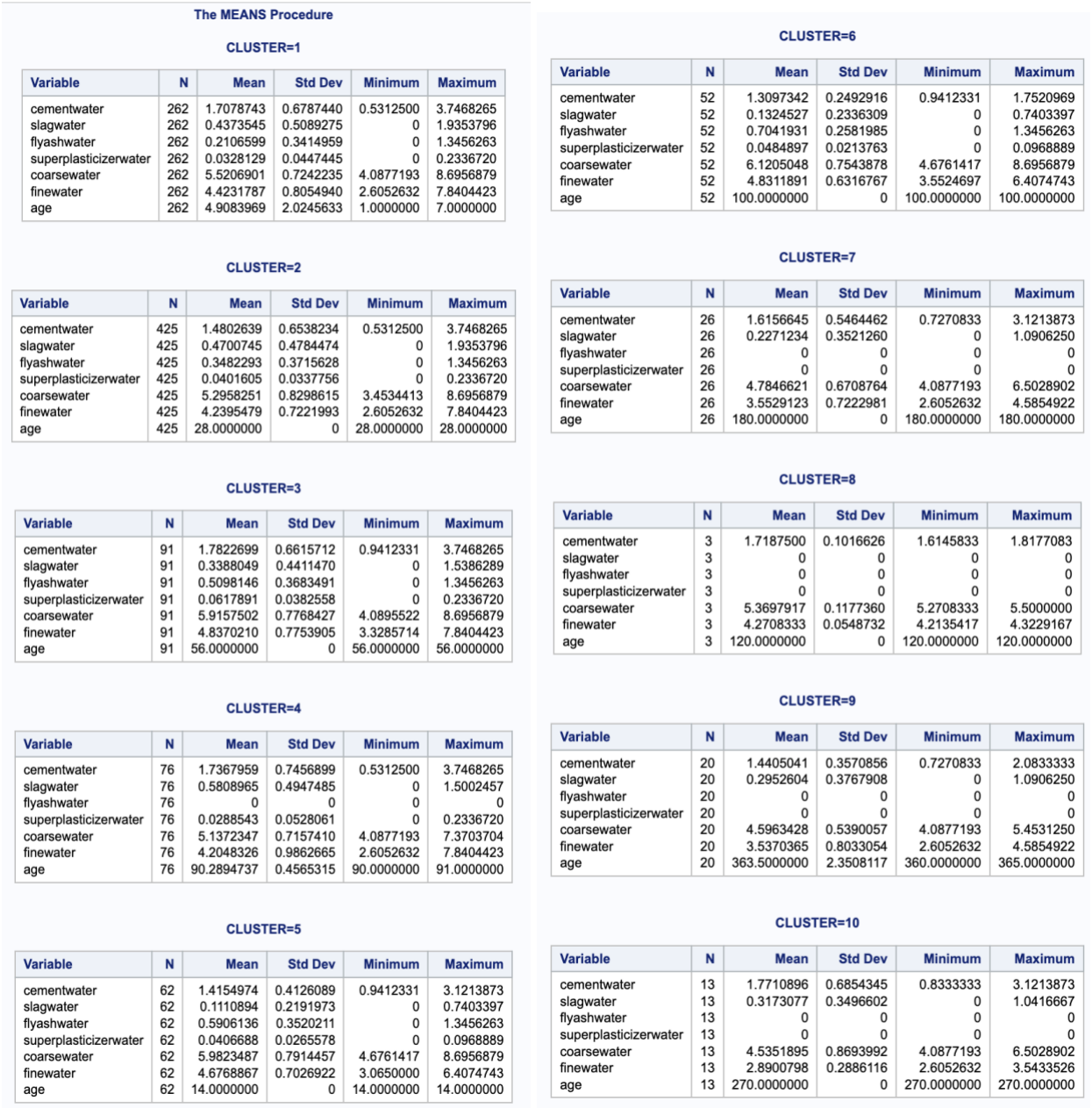
CLUSTER	agegroup 1	2	3	4	5	6	Total
1	136	126	0	0	0	0	262
2	0	0	425	0	0	0	425
3	0	0	0	91	0	0	91
4	0	0	0	0	76	0	76
5	0	62	0	0	0	0	62
6	0	0	0	0	52	0	52
7	0	0	0	0	0	26	26
8	0	0	0	0	3	0	3
9	0	0	0	0	0	20	20
10	0	0	0	0	0	13	13
Total	136	188	425	91	131	59	1030

Figure 4

Finally, I adopt the MEANS procedure to find out the general statistical summaries of each cluster based on the six component-water ratios and concrete age (**Figure 5**).

Here are some noteworthy differences from the figure: cluster 3 has the highest mean cement-water ratio (1.7822699); cluster 4 has the highest mean slag-water ratio (0.5808965); cluster 6 has the highest mean fly ash-water ratio (0.7041931); cluster 3 has the highest mean superplasticizer-water ratio (0.0617891); cluster 6 has the highest mean coarse-water ratio (6.1205048); cluster 3 has the highest mean fine-water ratio (4.8370210); and the cluster 9 has the highest mean age (363.5000000). From the frequency table (**Figure 4)** earlier, we have noticed that three clusters 7, 9, 10 only contain the observations in age group 6, which is the age group that has the highest concrete age (>=180 days). In this test, we see that clusters 7, 9, 10 truly have the highest age mean value among the ten clusters. Therefore, the results we got here are consistent with the results we got before.

Given all the results from the frequency table and the means test here, we can see that cluster 3 only contain age group 4 but with the highest mean cement-water, superplasticizer-water, and fine-water ratios. Therefore, we can conclude that there are at least some relationships between the age group 4 and the cement-water ratio, superplasticizer-water ratio, and fine-water ratio values. Moreover, cluster 4 and 6 only contain age group 5 but with the highest mean slag-water ratio, fly ash-water ratio, and the highest mean coarse-water ratio, we say there probably some relationships between the age group 5 and the three ratios - slag-water ratio, fly ash-water ratio, and coarse-water ratio.



The MEANS Procedure

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	262	1.7078743	0.6787440	0.5312500	3.7468265
slagwater	262	0.4373545	0.5089275	0	1.9353796
flyashwater	262	0.2106599	0.3414959	0	1.3456263
superplasticizerwater	262	0.0328129	0.0447445	0	0.2336720
coarsewater	262	5.5206901	0.7242235	4.0877193	8.6956879
finewater	262	4.4231787	0.8054940	2.6052632	7.8404423
age	262	4.9083969	2.0245633	1.0000000	7.0000000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	425	1.4802639	0.6538234	0.5312500	3.7468265
slagwater	425	0.4700745	0.4784474	0	1.9353796
flyashwater	425	0.3482293	0.3715628	0	1.3456263
superplasticizerwater	425	0.0401605	0.0337756	0	0.2336720
coarsewater	425	5.2958251	0.8298615	3.4534413	8.6956879
finewater	425	4.2395479	0.7221993	2.6052632	7.8404423
age	425	28.0000000	0	28.0000000	28.0000000

CLUSTER=3

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

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	76	1.7367959	0.7456899	0.5312500	3.7468265
slagwater	76	0.5808965	0.4947485	0	1.5002457
flyashwater	76	0	0	0	0
superplasticizerwater	76	0.0288543	0.0528061	0	0.2336720
coarsewater	76	5.1372347	0.7157410	4.0877193	7.3703704
finewater	76	4.2048326	0.9862665	2.6052632	7.8404423
age	76	90.2894737	0.4565315	90.0000000	91.0000000

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	62	1.4154974	0.4126089	0.9412331	3.1213873
slagwater	62	0.1110894	0.2191973	0	0.7403397
flyashwater	62	0.5906136	0.3520211	0	1.3456263
superplasticizerwater	62	0.0406688	0.0265578	0	0.0968889
coarsewater	62	5.9823487	0.7914457	4.6761417	8.6956879
finewater	62	4.6768867	0.7026922	3.0650000	6.4074743
age	62	14.0000000	0	14.0000000	14.0000000

CLUSTER=6

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	52	1.3097342	0.2492916	0.9412331	1.7520969
slagwater	52	0.1324527	0.2336309	0	0.7403397
flyashwater	52	0.7041931	0.2581985	0	1.3456263
superplasticizerwater	52	0.0484897	0.0213763	0	0.0968889
coarsewater	52	6.1205048	0.7543878	4.6761417	8.6956879
finewater	52	4.8311891	0.6316767	3.5524697	6.4074743
age	52	100.0000000	0	100.0000000	100.0000000

CLUSTER=7

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

CLUSTER=8

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	3	1.7187500	0.1016626	1.6145833	1.8177083
slagwater	3	0	0	0	0
flyashwater	3	0	0	0	0
superplasticizerwater	3	0	0	0	0
coarsewater	3	5.3697917	0.1177360	5.2708333	5.5000000
finewater	3	4.2708333	0.0548732	4.2135417	4.3229167
age	3	120.0000000	0	120.0000000	120.0000000

CLUSTER=9

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

CLUSTER=10

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

Figure 5

Summary:

In a nutshell, we can conclude that the 10 number of clusters are good enough for us to do analysis for all the observations. From all the diagnostics and tests above, we conclude that there are significant differences in compressive strength across the ten identified groups. Also, there are some differences and characteristics in the component-water ratios across the 10 clusters and different age groups. For example, cluster 3, which only contain age group 4, has the highest mean cement-water, superplasticizer-water, and fine-water ratios. From there, we can conclude that these three component-water ratios are more likely to be high when in the middle age of concrete.

3. **The manager has a specific interest in strength of concrete that is at least 90 days old. With the data available, build the best model you can for predicting compressive strength of concrete that is at least 90 days old and explain what that model tells us.**

Because of the continuous data type of our response variable compressive strength and the six predictors (component-water ratio), I decided to use linear regression model to address this topic. I first made another dataset that only contain the observations we interested in – concretes that at least 90 days old. Then I used stepwise variable selection to find the most appropriate predictors for predicting our response compressive strength with 190 out of the total 1030 observations. The summary of our model selection is attached below (**Figure 1)**:

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	0.3512	0.3512	242.597	101.76	<.0001
2	slagwater		2	0.2031	0.5543	110.409	85.23	<.0001
3	flyashwater		3	0.1506	0.7049	12.9307	94.92	<.0001
4	finewater		4	0.0155	0.7204	4.7119	10.23	0.0016

Figure 1

We see that there are four predictors (cement-water ration, slag-water ratio, fly ash-water ratio, and fine-water ratio) left with small p-value, which means they are the statistically significant predictors in this linear regression model. Therefore, our best model contains only these four predictors and the response variable compressive strength. After that I made some diagnostics and analysis based on our best model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	22245	5561.28909	119.15	<.0001
Error	185	8634.55161	46.67325		
Corrected Total	189	30880			

Root MSE	6.83178	R-Square	0.7204
Dependent Mean	46.97346	Adj R-Sq	0.7143
Coeff Var	14.54392		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	15.98000	2.49284	6.41	<.0001
cementwater	1	18.28954	0.95387	19.17	<.0001
slagwater	1	18.45323	1.25495	14.70	<.0001
flyashwater	1	19.40942	1.91617	10.13	<.0001
finewater	1	-1.94854	0.60908	-3.20	0.0016

Figure 2

I first did ANOVA test (**Figure 2**) to see the variance of model. The result - large F value (119.15) and less p-value (<.0001) mean that the differences between group means are statistically significant in this model. Moreover, the R-square is 0.7204, which means this model has explained about 72.04% of the overall observation variances. Finally, the Parameter Estimates test shows us that the intercept and the four parameters of the predictors in our model are all statistically significant, which provides us a double-check for the variables we selected.

From the results, we know that our linear regression model is (P_strength = "predicted value of compressive strength"):

P_strength = 15.98 + 18.29 * cement-water + 18.45 * slag-water + 19.41 * fly ash-water - 1.95 * fine-water

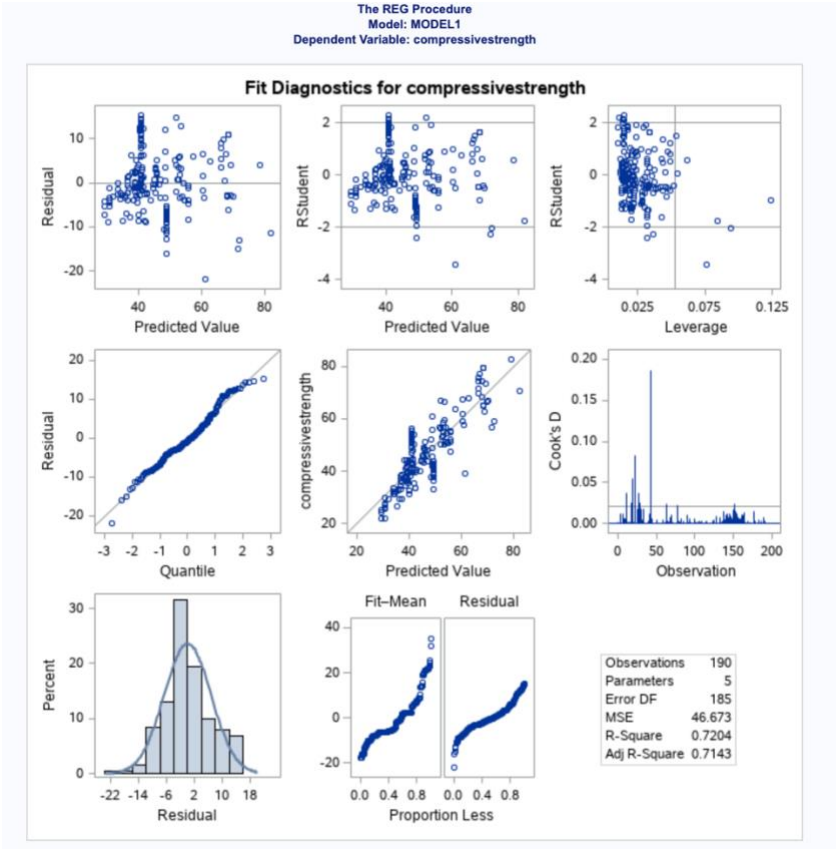


Figure 3

Next, I created a Fit Diagnostics plot (**Figure 3**) for our response variable compressive strength. The plot contains some useful sub-plots: the Residual plot, Studentized-Residual plot, Leverage plot, Quantile plot, Cook's Distance plot, etc. From the residual plot, we see that the residuals are almost all around the horizontal line (i.e., y=0 line), and are randomly spread without any noticeable trends and shape. Therefore, we say our model has relatively less differences between the observed and predicted responses and our model have constant variances as predicted value increase (Homoskedasticity). Next, from the Leverage plot, we know that there is extremeness of only few data points with respect to the remaining observations, and most of the data points are well lying in the appropriate leverage range. Furthermore, from the Cook's Distance plot, we see that the influence of data points on our model fitting are overall small, with less than 1 for all the observations. There is only one observation greater than 0.15 but rest of them are all less than 0.10. This suggests that the observations are not highly influential at all.

Finally, we can see the Residual vs Regressor / Predictors plot (**Figure 4)**. The big plot contains four small residual vs predictor plots, one for each predictor – cement-water ratio, slag-water ratio, fly ash-water ratio, and fine-water ratio. We can see that the upper-left residual plot has the best residual shape. All the residuals are randomly lying around the horizontal 0 line, which means the differences between observed and predicted response are small with the cement-water ratio

predictor. (i.e., cement-water ratio could predict the compressive strength the best among the four components.) Similarly, slag-water ratio and fly ash-water ratio have less ability to predict the compressive strength, or there is more error when predicting the compressive strength with these two variables.

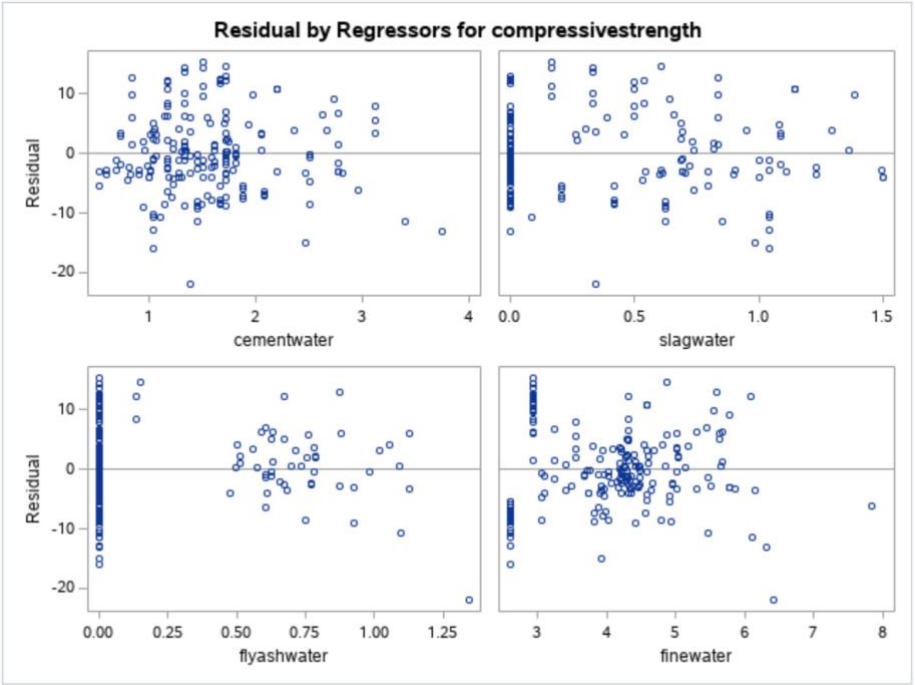


Figure 4

Summary:

From all the results above, we can conclude that our simple linear regression model (P_strength = 15.98 + 18.29 * cement-water + 18.45 * slag-water + 19.41 * fly ash-water - 1.95 * fine-water) is appropriate to predict the compressive strength. All the parameters are statistically significant and provide valuable information, such as the positive/negative relationship between the specific predictor and the response, for the prediction. Moreover, all the four components are good for predicting the compressive strength with relatively small errors/residuals when we see the residual plots.

4. **In some applications, the manager must allow the concrete to dry for at least 90 days and no more than 100 days before the next stage of the project, and the concrete must have a compressive strength of at least 47 MPa at that time. To be safe, the manager uses 50 MPa as the cutoff instead. Determine the best model for predicting if concrete that has cured for 90 to 100 days will have a strength of at least 50 MPa and interpret what that model tells us.**

I first created a new dataset satisfied the condition - concrete has age greater than or equal to 90 days and less than or equal to 100 days. Then, I modified the dataset by adding a binary term GoodStrength (0|1) – the concrete that meet the age condition having a strength of greater than or equal to 50 MPa (1), or the concrete that meet the age condition does not have a strength of greater than or equal to 50 MPa (0).

In order to address this topic with probabilities, I created the Logistic Regression model with the Binary response GoodStrength and six continuous predictors – cement-water ratio, slag-water ratio, fly ash-water ratio, superplasticizer-water ratio, coarse-water ratio, and fine-water ratio to begin with. Here's the model information (**Figure 1**):

Model Information	
Data Set	WORK.INTERESTDATA3
Response Variable	GoodStrength
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	128
Number of Observations Used	128

Response Profile		
Ordered Value	GoodStrength	Total Frequency
1	1	51
2	0	77

Figure 1

We see that the model is called Binary Logit, and we use Fisher's scoring as optimizing Technique. The number of observations we have is 128, which is the total number of interested concrete with age between 90 and 100 days. From the frequency table, we see that there are 51 observations of the total 128 interested observations having good strength (i.e., having compressive strength >= 50); and there are 77 observations not identifies as good strength.

After that, I did the Backward Elimination process to find the best model with the most appropriate predictors. Here's the summary: (**Figure 2)**

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	finewater	1	5	0.5211	0.4704

Figure 2

We see that only fine-water ratio are eliminated after the selection.

After fitting the observations based on the remaining five predictors to our logistics regression model, we got the following results (**Figure 3)**:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.3006	3.1820	5.2640	0.0218
cementwater	1	7.9146	1.6476	23.0768	<.0001
slagwater	1	8.1185	1.7508	21.5019	<.0001
flyashwater	1	6.8592	2.4441	7.8760	0.0050
superplasticizerwate	1	79.9769	24.8791	10.3338	0.0013
coarsewater	1	-2.3090	0.8019	8.2918	0.0040

Figure 3

We can see that each predictors left having statistically significance with Chi-Square test, and all having meaningful information to our model with positive or negative parameter estimates with the Maximum Likelihood Analysis. Therefore, we got the best model after variable/model selection, which is:

log (pi_i / (1 - pi_i)) = -7.3006 + 7.9146 * cement-water$_i$ + 8.1185 * slag-water$_i$ + 6.8592 * fly ash-water + 79.9769 * superplasticizer-water - 2.3090 * coarse-water$_i$

(Where pi_i is the probability a concrete that cured for 90 to 100 days have a strength of at least 50 MPa, and i represents all the individual trails.)

I did some diagnostics and analysis based on the final model chosen. From the Odds Ratio Estimates table below (**Figure 4**), we can obtain the information of each parameter's odds ratio – the ratio of probability that the event happens to not happen. In this case, we have big estimates for cement-water ratio, slag-water ratio, fly ash-water ratio, and superplasticizer-water ratio, which means these four main effects are the predictors more likely to have GoodStrength; on the other hand, there is a very small estimate value for coarse-water ratio (0.099), which means it is more likely having low good-strength rate or low probability that the concrete has compressive strength greater than or equal to 50 based on coarse-water ratio. Finally, we see that all parameters having significant odds ratios because 1 is not in their confidence intervals for the odds ratios. This somewhat indicate the predicted probability result we got in this model are good and significant.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
cementwater	>999.999	108.358	>999.999
slagwater	>999.999	108.526	>999.999
flyashwater	952.570	7.915	>999.999
superplasticizerwate	>999.999	>999.999	>999.999
coarsewater	0.099	0.021	0.478

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.4	Somers' D	0.927
Percent Discordant	3.6	Gamma	0.927
Percent Tied	0.0	Tau-a	0.448
Pairs	3927	c	0.964

Figure 4

Next, I did the Association analysis between the predicted probabilities and the observed response (**Figure 5**). There are 96.4% of Concordant and only 3.6% of discordant, which also indicate our model has good fitting for the observations and are well modeled.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.4	Somers' D	0.927
Percent Discordant	3.6	Gamma	0.927
Percent Tied	0.0	Tau-a	0.448
Pairs	3927	c	0.964

Figure 5

Summary:

In a nutshell, the above logistics regression model created (log (pi_i / (1 - pi_i)) = -7.3006 + 7.9146 * cement-water$_i$ + 8.1185 * slag-water$_i$ + 6.8592 * fly ash-water + 79.9769 * superplasticizer-water - 2.3090 * coarse-water$_i$) with binary response variable GoodStrength and five continuous component-water ratio variables as predictors works well with our dataset with 96.4% concordant. Based on the model, we could know that given a concrete having age between 90 and 100 (inclusive), what is the ratio of the probability that the concrete will have a compressive strength greater than or equal to 50 to the probability that the concrete does not have a compressive strength.

5. **The manager would also like to be able to determine the approximate age of concrete based on composition (the composition will be fixed in many applications) and compressive strength. In particular, the manager is interested in classifying concrete samples into one of the following age ranges： age<1week; 1 week<=age<4weeks; 4 weeks <= age < 8 weeks; 8 weeks <= age < 90 days; 90days<=age<180days; 180days<=age. Explain to the manager how well this model works for classifying age groups and identify any age groups that are difficult to tell apart.**

First, I did the stepwise selection to find out the appropriate predictors in our discrimination model, by the STEPDISC procedure in SAS. The stepwise selection summary was attached below (**Figure 1**). As we see, all the predictors compressive strength, cement/water ratio, blast furnace slag/water ratio, fly ash/water ratio, superplasticizer/water ratio, coarse aggregate/water ratio, and fine aggregate/water ratio are statistically significant with small p-values (almost all of them are <.0001), and they will all be included in our discrimination model.

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

Figure 1

Next, I performed Chi-Square test (**Figure 2)** to test the homogeneity or equality of within-group covariance matrices. Since the results contain a large Chi-Square value with small p-value (<0.0001), we reject the null hypothesis that there are equal within-group covariances, and thus conclude that there are differences and thus we need to use the Quadratic Discriminate Analysis (QDA) instead of the Linear Discriminate Analysis (LDA).

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

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

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Figure 2

Finally, I did the Cross Validation Classification test with Error Rate Estimates, and here's the summary (**Figure 3**):

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	

Figure 3

From the summary, we could know the correct and incorrect observation classifications:

- For age group 1, we see that there are 54.41% correct classification with 11.76% misclassified to age group 2, 2.21% misclassified to age group 3, and 31.62% misclassified to age group 6.
- For age group 2, there are 29.26% correct classification with 5.85% misclassified to age group 1, 26.06% misclassified to age group 3, 2.13% misclassified to age group 4, and 36.70 % misclassified to age group 6.
- For age group 3, there are 62.59% correct classification with 7.76% misclassified to age group 2, 7.53% misclassified to age group 4, 4.24% misclassified to age group 5, and 17.88% misclassified to age group 6.
- For age group 4, there are 35.16% correct classification with 2.20% misclassified to age group 2, 25.27% misclassified to age group 3, 35.16% misclassified to age group 5, and 2.20% misclassified to age group 6.
- For age group 5, there are 23.66% correct classification with 8.40% misclassified to age group 3, 27.48% misclassified to age group 4, and 40.46% misclassified to age group 6.
- For age group 6, there are 100% correct classification.

Finally, from the Error Count Estimates table, we see that the total estimate error rate is 49.81% (0.4981) about 50% percent, which is made up from about 45.59% of age group 1, 70.74% of age group 2, 37.41% of age group 3, 64.84% of age group 4, and 76.34% of age group 5.

Generally speaking, this model does not work well for classifying age groups, since it has a really high error rate - 49.81%, which means there is one misclassification for every two classifications. Although age group 6 can be classified perfectly, all other age groups have confusion of classifying and some of them with large error rate are difficult to tell apart from other age groups. For example, age group 2 has only 29.26% correct classification with 26.06% misclassified to age group 3 and

36.70 % misclassified to age group 6. It can be inferred that some concrete's characteristics based on component-water ratio and compressive strength of age group 2 can be easily confused with age group 3 and 6. Moreover, age group 5 has only 23.66% correct classification with 27.48% misclassified to age group 4 and 40.46% misclassified to age group 6. It can also be inferred that some concrete's characteristics based on component-water ratio and compressive strength of age group 5 can be easily confused with age group 4 and 6. Similar to age group 4, which cannot tell apart from age group 3 and 5.

Summary:
To sum up, I would like to say that the model generally does not work well for classifying age groups with a large error rate 49.81%. Only one the age groups – age group 6 – has been classified correctly with 100%, all other age groups are misclassified with small or large error rate. There are 3 of the remaining 5 age groups have error rates larger than 50%, which means they are difficult to be identified apart from other age groups based on only the composition and compressive strength.

Overall Conclusion:

To sum up, based on all the plots, graphs, analysis, and diagnostics above, we have a good understanding the dataset and the relationship between features now.

From the first two topics, we learned that there are differences in compressive strength for different ages and component-water ratios of concrete. From topic 1 we concluded that there is a positive relationship between age and compressive strength. Moreover, the age group 4 (56 <= age < 90) has the highest compressive strength and that could strongly be explained by the high component-water ratios of the observations in that group. From topic 2, we clustered the data to ten groups and concluded that three component-water ratios are more likely to be high when in age group 4. From the next two topics, we learned the predictability of strength. From topic 3, we constructed a linear model (P_strength = 15.98 + 18.29 * cement-water + 18.45 * slag-water + 19.41 * fly ash-water - 1.95 * fine-water) for predicting the compressive strength based on the four component-water ratios. Also from topic 4, we constructed a logistics regression model (log (pi_i / (1 - pi_i)) = -7.3006 + 7.9146 * cement-water$_i$ + 8.1185 * slag-water$_i$ + 6.8592 * fly ash-water + 79.9769 * superplasticizer-water - 2.3090 * coarse-water$_i$) with binary response variable GoodStrength and five continuous component-water ratio variables as predictors. The model works well with 96.4% concordant. From the topic 5, we learned how well the age can be determined if the concrete composition and strength are known. We constructed model classifying the age groups. Although the Cross-Validation model does not work very well for the classification with a large error rate 49.81%, we got some meaningful information from the model as well. For example, some age groups such as age group 2, 4, 5 are difficult to be identified apart from other age groups based on only the composition and compressive strength.

Overall, the dataset provides some real-world relevance on the positive relationship between the concrete compressive strength and age, and how well we could predict the strength under specific conditions, which all useful in the construction work in our society. In the future, statistical modeling and data analysis will become more and more necessary in a variety of work type, and we could adopt more advanced technical tools based on different needs and requirements.