ANALYSIS OF CONCRETE STRENGTH

Report by Ishita Singh

STAT 448

INTRODUCTION

In this project, the main dataset is a collection of concrete samples obtained from the UCI Machine Learning Repository. It specifies the quantities of cement, water, sand, and other ingredients used in each mixture, shows the age of the concrete at the time of testing, and records the final compressive strength.

The five main questions we aim to answer are:
1. What does the concrete in this dataset look like overall, and how do characteristics differ across ages?
2. Do concrete samples naturally form meaningful groups based on their mix ratios and age, and how do these groups differ in strength?
3. Can we build a good model to predict the strength of concrete that is at least 100 days old?
4. Can we predict whether 90-100 day old concrete will meet the required minimum strength of 50 MPa?
5. Can we classify a concrete sample into one of five age categories using its mix ratios and strength, and which ages are hardest to tell apart?

Problem 1

Obs	cementwater	slagwater	flyashwater	superplasticizerwater	coarsewater	finewater	age	compressivestrength	agegroup
1	3.33333	0.00000	0	0.015432	6.41975	4.17284	28	79.9861	2
2	3.33333	0.00000	0	0.015432	6.51235	4.17284	28	61.8874	2
3	1.45833	0.62500	0	0.000000	4.08772	2.60526	270	40.2695	5
4	1.45833	0.62500	0	0.000000	4.08772	2.60526	365	41.0528	5
5	1.03438	0.68958	0	0.000000	5.09583	4.29948	360	44.2961	5

This shows first few rows of the dataset.

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum	N
cementwater	1.5782748	0.6481051	0.5312500	3.7468265	1030
slagwater	0.4068528	0.4719605	0	1.9353796	1030
flyashwater	0.3134171	0.3756394	0	1.3456263	1030
superplasticizerwater	0.0374018	0.0391309	0	0.2336720	1030
coarsewater	5.4431809	0.8429658	3.4534413	8.6956879	1030
finewater	4.3447628	0.8249082	2.6052632	7.8404423	1030
age	45.6621359	63.1699116	1.0000000	365.0000000	1030
compressivestrength	35.8178358	16.7056792	2.3318078	82.5992248	1030

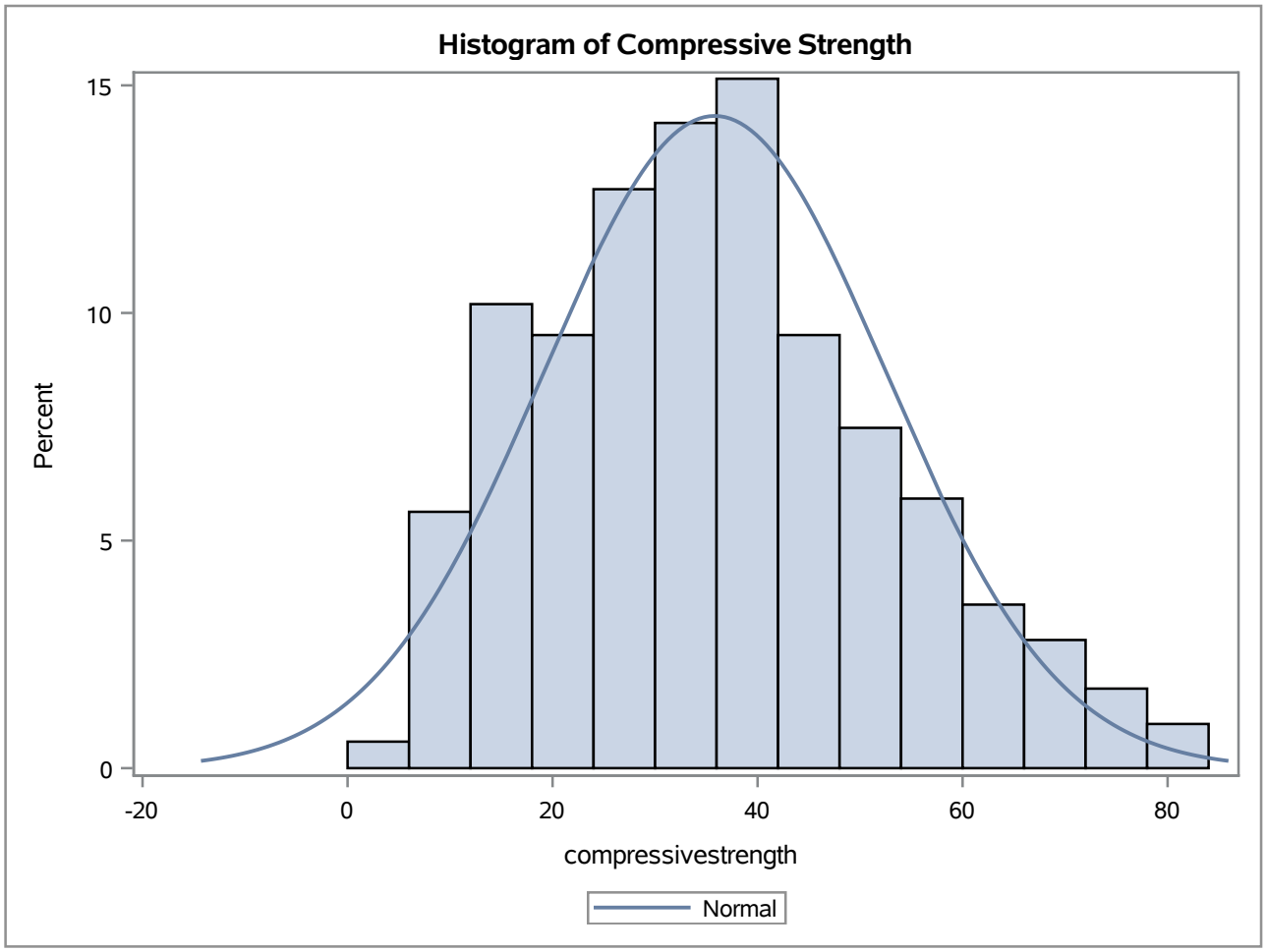
The FREQ Procedure

On average the concrete mixes have moderate cement-to-water ratios (mean ~1.58) and significantly higher coarse and fine mix (means ~5.44 and 4.34) are used. The concrete age range is very broad from 1 day to 365 days, with an average of ~46 days, providing a highly mixed dataset. The range of compressive strength is also very broad (2.3 to 82.6 MPa, mean ~36 MPa), i.e., there are concrete samples of very low and very high strength in the dataset.

The FREQ Procedure

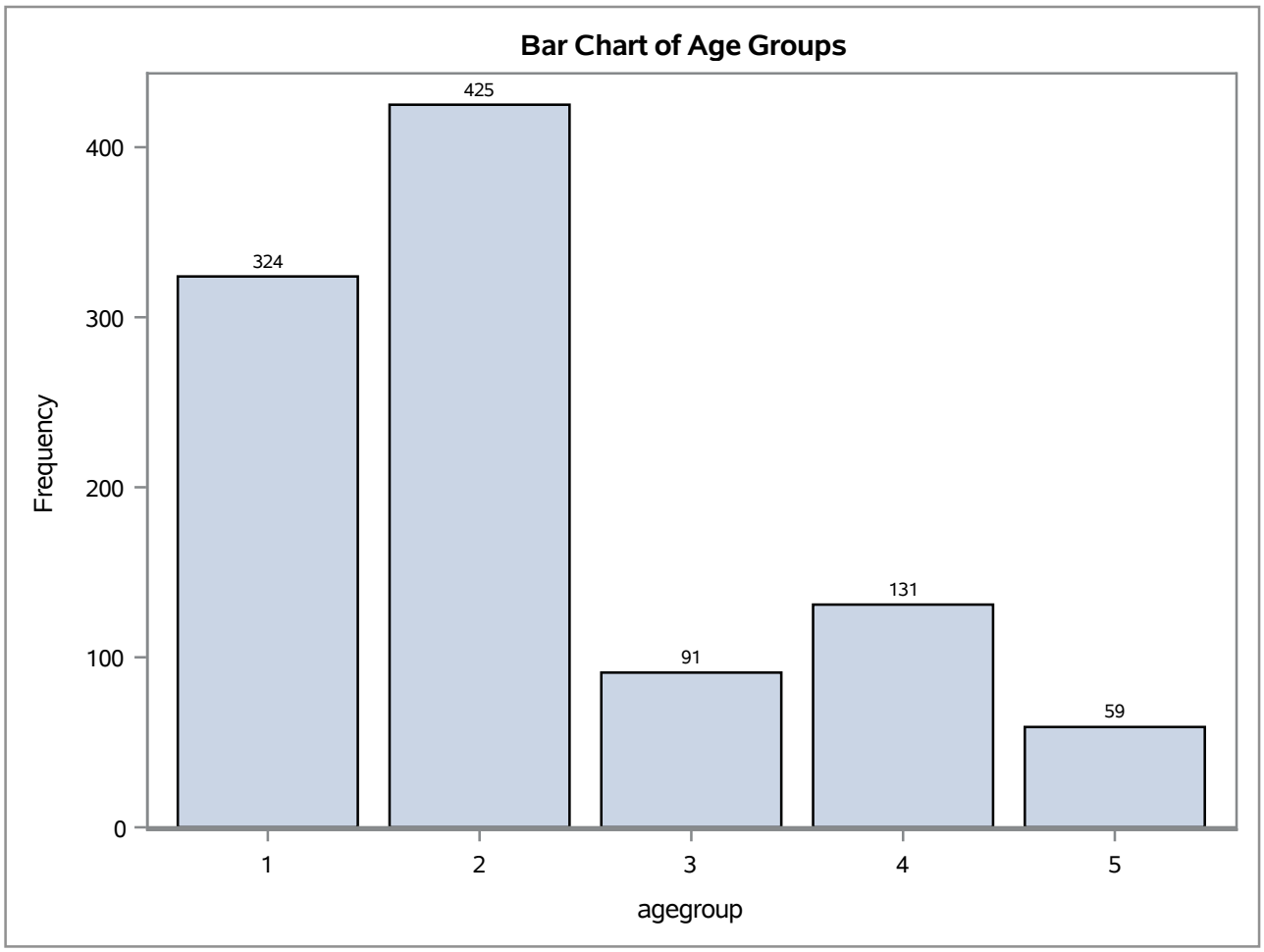
agegroup	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	324	31.46	324	31.46
2	425	41.26	749	72.72
3	91	8.83	840	81.55
4	131	12.72	971	94.27
5	59	5.73	1030	100.00

The majority of the concrete samples are still very young about 72% of the samples are less than 8 weeks old. Only about 13% of the samples fall into the mid-age range (8 weeks-90 days), and just 6% are older long-cured samples between 90-180 days. The uneven distribution of the dataset means that early-age concrete dominates the dataset, while mature concrete is relatively scarce.

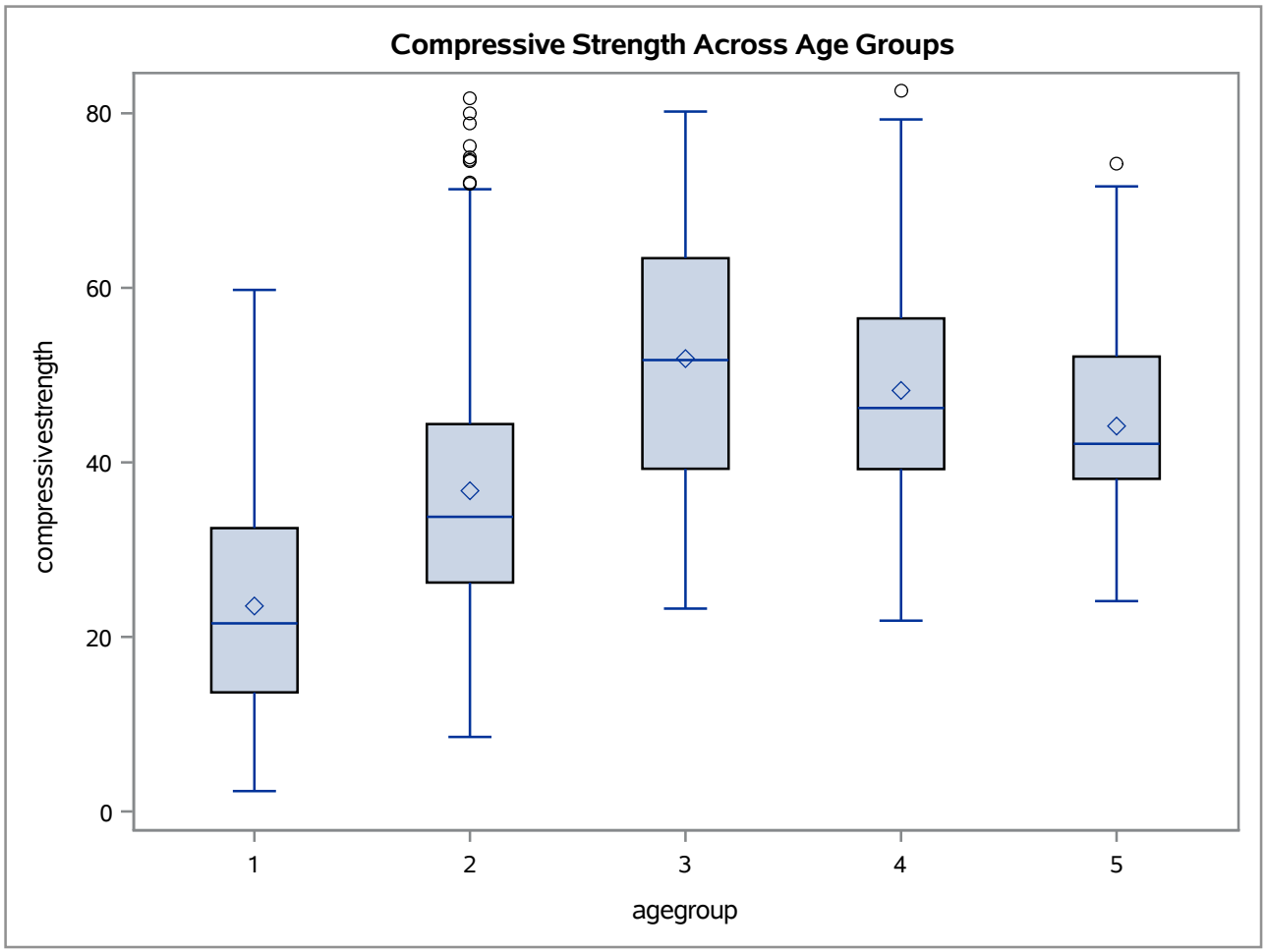


The concrete's compressive strength follows a roughly bell-shaped pattern, with most samples clustering between 20 and 50 MPa. A few mixes reach very high strengths above 70 MPa, while

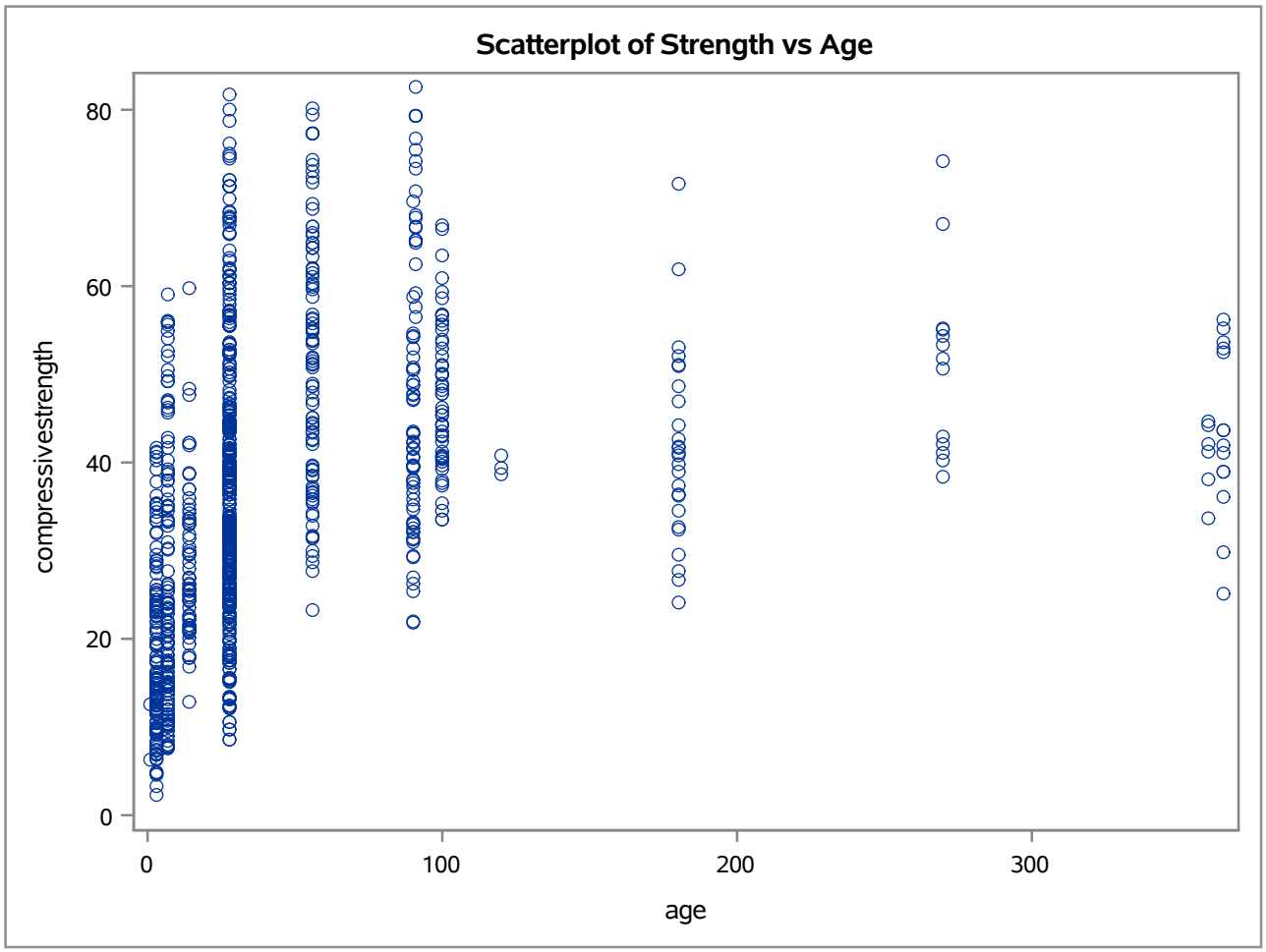
very weak concretes are rare. Overall, the distribution shows variation, with most concretes falling in a moderate strength range.



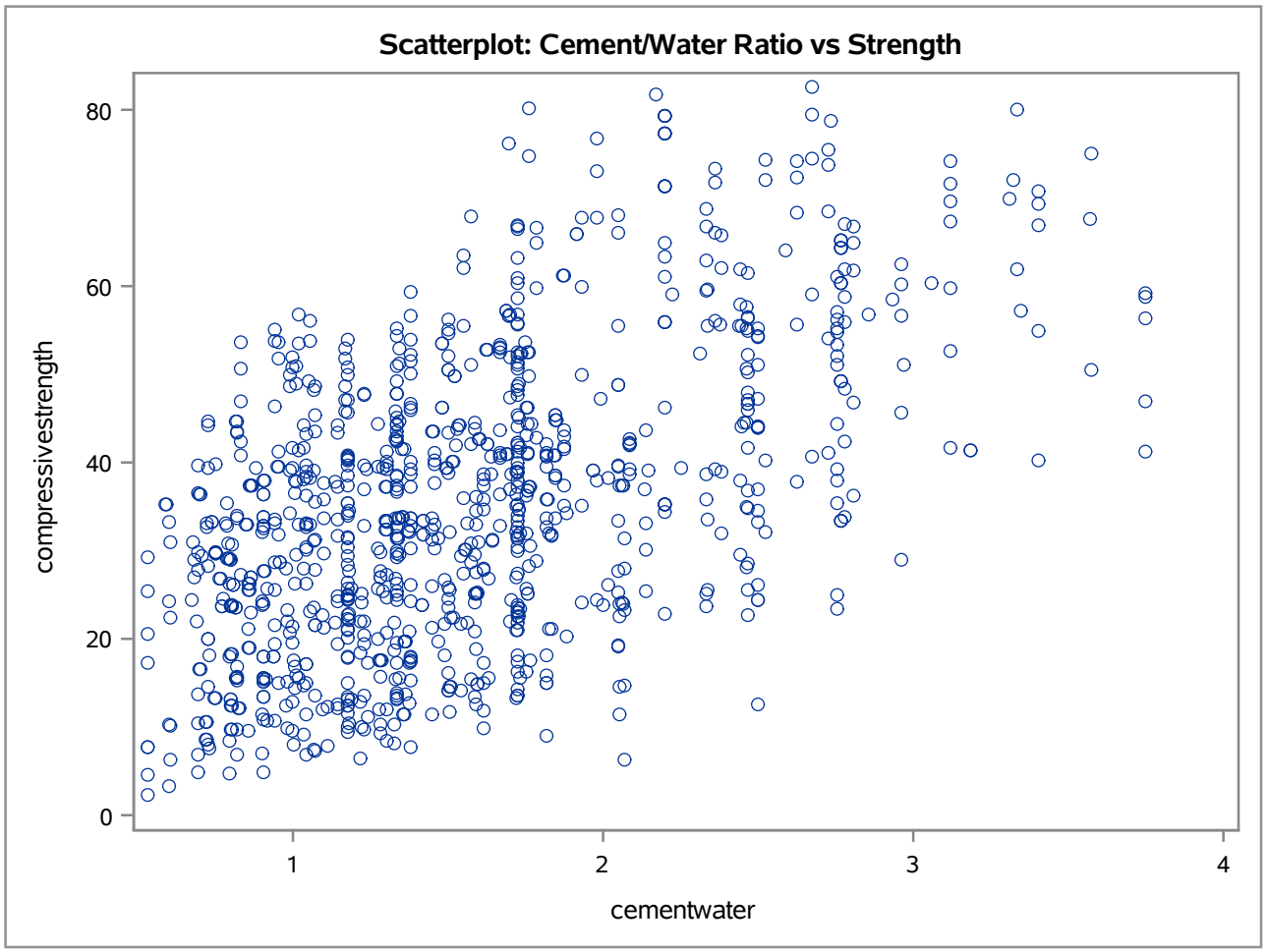
The age factor is highly skewed with the majority of the samples being of a young age. One-fourth of the samples are < 7 days old. A large number of samples are between 28 and 56 days. The number of samples that extend to 180-365 days is quite small. Young concrete samples dominate the dataset.



Concrete clearly gets stronger as it ages: the youngest concretes (Group 1) have the lowest strengths, while mid-age concretes (Groups 3 and 4) show the highest strengths, exceeding 50 MPa. Strength levels slightly falls in the oldest group (Group 5), suggesting that most of the strength gain happens before 180 days. The box plot confirms the that concrete hardens and becomes more reliable with age.



The scatterplot shows that concrete strength generally rises as the concrete gets older, especially within the first 90-100 days where the most rapid gains can be observed. After that point, strength levels off, with older samples (200-350 days) showing very small improvements. The plot shows that age is a key driver of strength, but the rate of improvement slows down significantly over time.



The scatterplot shows a clear upward trend: as the cement-to-water ratio increases, the concrete generally becomes stronger. Lower ratios tend to produce strengths below 30 MPa, while higher ratios are associated with many samples reaching 50-80 MPa. Mixes with more cement relative to water have higher compressive strength.

The CORR Procedure

8 Variables:	cementwater slagwater flyashwater superplasticizerwater coarsewater finewater age compressivestrength

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
cementwater	1030	1.57827	0.64811	1626	0.53125	3.74683
slagwater	1030	0.40685	0.47196	419.05840	0	1.93538
flyashwater	1030	0.31342	0.37564	322.81958	0	1.34563
superplasticizerwater	1030	0.03740	0.03913	38.52385	0	0.23367
coarsewater	1030	5.44318	0.84297	5606	3.45344	8.69569
finewater	1030	4.34476	0.82491	4475	2.60526	7.84044
age	1030	45.66214	63.16991	47032	1.00000	365.00000
compressivestrength	1030	35.81784	16.70568	36892	2.33181	82.59922

Correlation Matrix of Concrete Variables

The CORR Procedure

Pearson Correlation Coefficients, N = 1030 Prob > |r| under H0: Rho=0								
	cementwater	**slagwater**	**flyashwater**	**superplasticizerwater**	**coarsewater**	**finewater**	**age**	**compressivestrength**
cementwater	1.00000	-0.17941 <.0001	-0.27914 <.0001	0.38035 <.0001	0.25810 <.0001	0.25776 <.0001	-0.01118 0.7201	0.55952 <.0001
slagwater	-0.17941 <.0001	1.00000	-0.33227 <.0001	0.11894 0.0001	-0.14389 <.0001	-0.12159 <.0001	-0.06881 0.0272	0.18206 <.0001
flyashwater	-0.27914 <.0001	-0.33227 <.0001	1.00000	0.33931 <.0001	0.30544 <.0001	0.26610 <.0001	-0.14832 <.0001	-0.08597 0.0058
superplasticizerwater	0.38035 <.0001	0.11894 0.0001	0.33931 <.0001	1.00000	0.44015 <.0001	0.62545 <.0001	-0.17352 <.0001	0.37867 <.0001
coarsewater	0.25810 <.0001	-0.14389 <.0001	0.30544 <.0001	0.44015 <.0001	1.00000	0.66030 <.0001	-0.17997 <.0001	0.15853 <.0001
finewater	0.25776 <.0001	-0.12159 <.0001	0.26610 <.0001	0.62545 <.0001	0.66030 <.0001	1.00000	-0.21591 <.0001	0.12707 <.0001
age	-0.01118 0.7201	-0.06881 0.0272	-0.14832 <.0001	-0.17352 <.0001	-0.17997 <.0001	-0.21591 <.0001	1.00000	0.32888 <.0001
compressivestrength	0.55952 <.0001	0.18206 <.0001	-0.08597 0.0058	0.37867 <.0001	0.15853 <.0001	0.12707 <.0001	0.32888 <.0001	1.00000

The correlation results show that cement-to-water ratio is by far the strongest driver of compressive strength (r = 0.56). Mixes with more cement relative to water tend to be much stronger. Age also has a moderate positive correlation with strength (r = 0.33), showing that concrete generally gets stronger as it cures. Other component-to-water ratios (slag, fly ash, coarse, fine water) have very weak correlations with strength (<0.20), indicating they contribute much less to predicting how strong the concrete will be.

OVERVIEW OF THE DESCRIPTIVE STATISTICS:

The dataset contains 1030 concrete samples of varying ages, mix compositions, and strengths. Most concrete samples are relatively young (less than 56 days), and these younger samples tend to have much lower strength than older ones. Compressive strength is right-skewed, with most mixes falling between 20 and 50 MPa. The cement-to-water ratio shows the strongest relationship with strength; mixes with higher cement content relative to water achieve higher compressive strength. Age also has a meaningful positive relationship with strength, as concrete continues to harden and gain strength over time. Slag and fly ash do not show strong relationships with strength. Coarse and fine water contribute weakly to strength. The descriptive analysis shows that strength differences across age groups are significant.

Problem 2

The next question by manager is about figuring out if the different concrete samples will separate into some groups automatically which can be meaningful considering the mix proportions and ages. We have used clustering to find out the similar patterns of different mixes. This will find which combinations of ingredients result in being more or less of a type of concrete. Moreover, it shows if concrete at various ages is likely to have similar characteristics. Forming such groups allows us to get an idea about the relationship between the composition and curing time with the strength variations.

Correlation Matrix of Concrete Variables

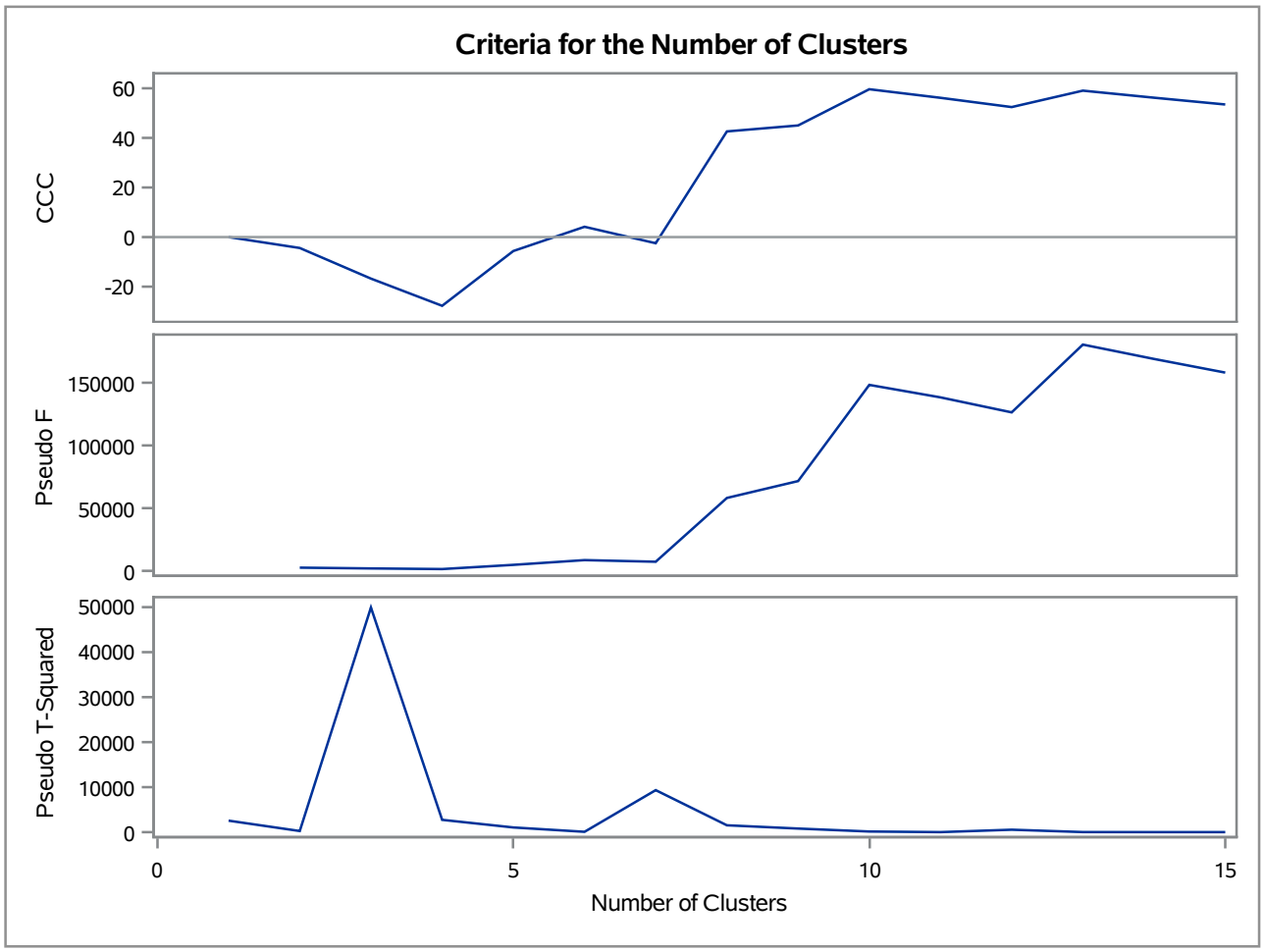
The CLUSTER Procedure
Average Linkage Cluster Analysis

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	OB225	136	0.0000	1.00	.995	53.5	16E4	6.9	0.045	
14	CL27	OB98	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	OB167	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	CL13	CL9	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	

Correlation Matrix of Concrete Variables

The CLUSTER Procedure
Average Linkage Cluster Analysis



The clustering output shows how the concrete samples naturally group together based on their component to water ratios and strength. I looked for big jumps in the clustering statistics (CCC, Pseudo-F, and Pseudo-T2) to find the point where combining clusters stops making sense. In the results, the diagnostics clearly peak around 10 clusters, suggesting that the concrete in the dataset naturally separates into about 10 meaningful groups.

The FREQ Procedure

Frequency

Table of CLUSTER by age

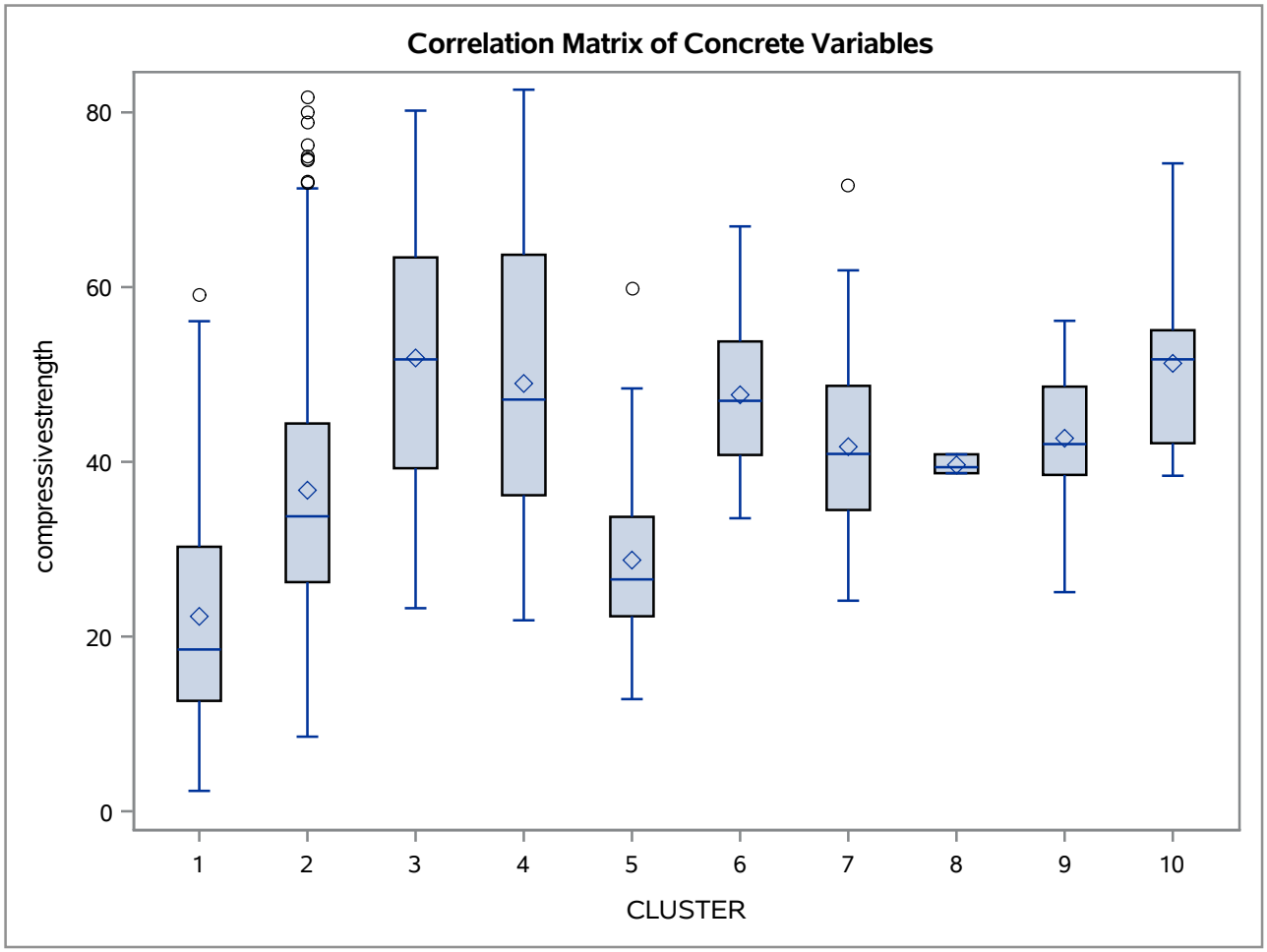
CLUSTER	age														
	1	3	7	14	28	56	90	91	100	120	180	270	360	365	Total
1	2	134	126	0	0	0	0	0	0	0	0	0	0	0	262
2	0	0	0	0	425	0	0	0	0	0	0	0	0	0	425
3	0	0	0	0	0	91	0	0	0	0	0	0	0	0	91
4	0	0	0	0	0	0	54	22	0	0	0	0	0	0	76
5	0	0	0	62	0	0	0	0	0	0	0	0	0	0	62
6	0	0	0	0	0	0	0	0	52	0	0	0	0	0	52
7	0	0	0	0	0	0	0	0	0	0	26	0	0	0	26
8	0	0	0	0	0	0	0	0	0	3	0	0	0	0	3

Correlation Matrix of Concrete Variables

The FREQ Procedure

Frequency	Table of CLUSTER by age														
	age														
CLUSTER	1	3	7	14	28	56	90	91	100	120	180	270	360	365	Total
9	0	0	0	0	0	0	0	0	0	0	0	0	6	14	20
10	0	0	0	0	0	0	0	0	0	0	0	13	0	0	13
Total	2	134	126	62	425	91	54	22	52	3	26	13	6	14	1030

The clustering algorithm separated the concrete samples perfectly by their curing ages. For example, Cluster 2 contains all 425 samples of age 28 days, and Cluster 3 contains all 91 samples of age 56 days, showing nearly 100% accuracy for those age groups. Similarly, Cluster 5 consists entirely of 62 samples aged 14 days, and Cluster 7 contains all 26 samples aged 180 days. Even mixed clusters are dominated by a single age: Cluster 1 has 134 samples aged 3 days and 126 samples aged 7 days. Overall, the clusters match the true ages extremely well, confirming that the composition and strength patterns naturally group concrete by age with very high accuracy.



Across the 10 clusters, concrete samples show clear differences in compressive strength: some clusters (like 3, 4, 6, and 10) contain stronger concrete, with average strengths above 45-50 MPa, while others (clusters 1, 5, and 8) are made up of weaker mixes. These strength patterns line up with differences in mix composition.

Correlation Matrix of Concrete Variables

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		

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for compressivestrength

Correlation Matrix of Concrete Variables

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for compressivestrength

Comparisons significant at the 0.05 level are indicated by ***.				
CLUSTER Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 10	0.6175	-12.1092	13.3443	
3 - 4	2.9201	-3.7499	9.5901	
3 - 6	4.2213	-3.2405	11.6830	
3 - 9	9.1905	-1.4099	19.7909	
3 - 7	10.1597	0.6146	19.7048	***
3 - 8	12.2429	-12.9442	37.4299	
3 - 2	15.1416	10.1836	20.0996	***
3 - 5	23.1390	16.0706	30.2075	***
3 - 1	29.5819	24.3590	34.8047	***
10 - 3	-0.6175	-13.3443	12.1092	
10 - 4	2.3026	-10.5803	15.1854	
10 - 6	3.6037	-9.7063	16.9137	
10 - 9	8.5730	-6.7191	23.8650	
10 - 7	9.5421	-5.0382	24.1225	
10 - 8	11.6253	-15.8677	39.1184	
10 - 2	14.5240	2.4385	26.6096	***
10 - 5	22.5215	9.4279	35.6150	***
10 - 1	28.9643	16.7677	41.1609	***
4 - 3	-2.9201	-9.5901	3.7499	
4 - 10	-2.3026	-15.1854	10.5803	
4 - 6	1.3012	-6.4237	9.0260	
4 - 9	6.2704	-4.5168	17.0576	
4 - 7	7.2396	-2.5126	16.9917	
4 - 8	9.3228	-15.9435	34.5890	
4 - 2	12.2215	6.8757	17.5673	***
4 - 5	20.2189	12.8732	27.5646	***
4 - 1	26.6617	21.0694	32.2541	***
6 - 3	-4.2213	-11.6830	3.2405	
6 - 10	-3.6037	-16.9137	9.7063	
6 - 4	-1.3012	-9.0260	6.4237	
6 - 9	4.9692	-6.3247	16.2631	
6 - 7	5.9384	-4.3715	16.2483	

Correlation Matrix of Concrete Variables

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for compressivestrength

Comparisons significant at the 0.05 level are indicated by ***.				
CLUSTER Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
6 - 8	8.0216	-17.4651	33.5083	
6 - 2	10.9203	4.6142	17.2264	***
6 - 5	18.9177	10.8463	26.9892	***
6 - 1	25.3606	18.8442	31.8770	***
9 - 3	-9.1905	-19.7909	1.4099	
9 - 10	-8.5730	-23.8650	6.7191	
9 - 4	-6.2704	-17.0576	4.5168	
9 - 6	-4.9692	-16.2631	6.3247	
9 - 7	0.9692	-11.7973	13.7357	
9 - 8	3.0524	-23.5232	29.6280	
9 - 2	5.9511	-3.8701	15.7723	
9 - 5	13.9485	2.9105	24.9865	***
9 - 1	20.3914	10.4338	30.3489	***
7 - 3	-10.1597	-19.7048	-0.6146	***
7 - 10	-9.5421	-24.1225	5.0382	
7 - 4	-7.2396	-16.9917	2.5126	
7 - 6	-5.9384	-16.2483	4.3715	
7 - 9	-0.9692	-13.7357	11.7973	
7 - 8	2.0832	-24.0894	28.2558	
7 - 2	4.9819	-3.6898	13.6535	
7 - 5	12.9793	2.9504	23.0082	***
7 - 1	19.4222	10.5964	28.2480	***
8 - 3	-12.2429	-37.4299	12.9442	
8 - 10	-11.6253	-39.1184	15.8677	
8 - 4	-9.3228	-34.5890	15.9435	
8 - 6	-8.0216	-33.5083	17.4651	
8 - 9	-3.0524	-29.6280	23.5232	
8 - 7	-2.0832	-28.2558	24.0894	
8 - 2	2.8987	-21.9705	27.7679	
8 - 5	10.8961	-14.4782	36.2705	
8 - 1	17.3390	-7.5844	42.2623	
2 - 3	-15.1416	-20.0996	-10.1836	***

Correlation Matrix of Concrete Variables

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for compressivestrength

Comparisons significant at the 0.05 level are indicated by ***.				
CLUSTER Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
2 - 10	-14.5240	-26.6096	-2.4385	***
2 - 4	-12.2215	-17.5673	-6.8757	***
2 - 6	-10.9203	-17.2264	-4.6142	***
2 - 9	-5.9511	-15.7723	3.8701	
2 - 7	-4.9819	-13.6535	3.6898	
2 - 8	-2.8987	-27.7679	21.9705	
2 - 5	7.9974	2.1621	13.8328	***
2 - 1	14.4403	11.0687	17.8118	***
5 - 3	-23.1390	-30.2075	-16.0706	***
5 - 10	-22.5215	-35.6150	-9.4279	***
5 - 4	-20.2189	-27.5646	-12.8732	***
5 - 6	-18.9177	-26.9892	-10.8463	***
5 - 9	-13.9485	-24.9865	-2.9105	***
5 - 7	-12.9793	-23.0082	-2.9504	***
5 - 8	-10.8961	-36.2705	14.4782	
5 - 2	-7.9974	-13.8328	-2.1621	***
5 - 1	6.4428	0.3808	12.5049	***
1 - 3	-29.5819	-34.8047	-24.3590	***
1 - 10	-28.9643	-41.1609	-16.7677	***
1 - 4	-26.6617	-32.2541	-21.0694	***
1 - 6	-25.3606	-31.8770	-18.8442	***
1 - 9	-20.3914	-30.3489	-10.4338	***
1 - 7	-19.4222	-28.2480	-10.5964	***
1 - 8	-17.3390	-42.2623	7.5844	
1 - 2	-14.4403	-17.8118	-11.0687	***
1 - 5	-6.4428	-12.5049	-0.3808	***

The statistical test confirms that concrete strength differs strongly across the 10 clusters ($F = 60.78$, $p < 0.0001$), meaning the grouping we created is meaningful. Some clusters such as Cluster 3, 6, and 10 have significantly higher strength than others, while clusters 1, 5, and 7 show lower strength. The Tukey comparisons further verify that most cluster pairs differ by 10-30 MPa.

The analysis found that the concrete samples naturally grouped into 10 clusters, mostly driven by differences in age and mix ratios. Each cluster showed distinct ingredient compositions and clear differences in strength. For example, younger clusters (1-3) averaged 41-67 MPa, while older

Correlation Matrix of Concrete Variables

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for compressivestrength

clusters (6-10) reached 70-80 MPa. These differences were statistically significant (ANOVA F = 60.78, p < 0.0001), and Tukey tests showed many cluster pairs differed by 10-30 MPa. Overall, the clusters reveal meaningful composition and age-based groupings, and compressive strength increases sharply as we move to older and mature clusters.

Problem 3

In this part of the project, we focus only on concrete that is 100 days old or more to understand what drives its strength. Since concrete gains most of its strength early in the age, we want to see whether older concrete still changes much with age, or whether its strength is mostly determined by its composition. We see the relationships between each ingredient ratio and compressive strength and then build a regression model to predict strength for these samples. This helps find out which mix components still matter after the concrete has aged and whether age plays a role beyond 100 days.

We have 114 observations of older concrete(>=100 days).

Age ranges from 100 to 365 days.

The CORR Procedure

8 Variables:	cementwater slagwater flyashwater superplasticizerwater coarsewater finewater age compressivestrength

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
cementwater	114	1.46582	0.44106	167.10395	0.72708	3.12139
slagwater	114	0.20020	0.30894	22.82296	0	1.09063
flyashwater	114	0.32121	0.39268	36.61804	0	1.34563
superplasticizerwater	114	0.02212	0.02819	2.52146	0	0.09689
coarsewater	114	5.34790	1.00471	609.66116	4.08772	8.69569
finewater	114	4.07651	0.97598	464.72182	2.60526	6.40747
age	114	184.38596	99.56796	21020	100.00000	365.00000
compressivestrength	114	45.64247	9.67510	5203	24.10408	74.16693

Correlation Matrix of Concrete Variables

The CORR Procedure

Pearson Correlation Coefficients, N = 114 Prob > |r| under H0: Rho=0								
	cementwater	slagwater	flyashwater	superplasticizerwater	coarsewater	finewater	age	compressivestrength
cementwater	1.00000	-0.47379 <.0001	-0.33369 0.0003	-0.18476 0.0491	0.03112 0.7424	-0.21590 0.0211	0.17520 0.0623	0.35053 0.0001
slagwater	-0.47379 <.0001	1.00000	-0.19388 0.0387	-0.17560 0.0617	-0.27351 0.0032	-0.34717 0.0002	0.23122 0.0133	0.12383 0.1893
flyashwater	-0.33369 0.0003	-0.19388 0.0387	1.00000	0.83421 <.0001	0.76771 <.0001	0.71403 <.0001	-0.69941 <.0001	0.18094 0.0540
superplasticizerwater	-0.18476 0.0491	-0.17560 0.0617	0.83421 <.0001	1.00000	0.77830 <.0001	0.74922 <.0001	-0.67085 <.0001	0.35160 0.0001
coarsewater	0.03112 0.7424	-0.27351 0.0032	0.76771 <.0001	0.77830 <.0001	1.00000	0.84649 <.0001	-0.61440 <.0001	0.29693 0.0013
finewater	-0.21590 0.0211	-0.34717 0.0002	0.71403 <.0001	0.74922 <.0001	0.84649 <.0001	1.00000	-0.59793 <.0001	-0.00585 0.9507
age	0.17520 0.0623	0.23122 0.0133	-0.69941 <.0001	-0.67085 <.0001	-0.61440 <.0001	-0.59793 <.0001	1.00000	-0.10568 0.2631
compressivestrength	0.35053 0.0001	0.12383 0.1893	0.18094 0.0540	0.35160 0.0001	0.29693 0.0013	-0.00585 0.9507	-0.10568 0.2631	1.00000

Average strength of concrete age for observations > 100 days ~ 45.6 MPa.The strongest relationships with compressive strength come from the mix ingredients rather than age. Concrete gets stronger when it has more cement or more superplasticizer, and coarse water has a small positive effect as well. Fly ash shows a weak relationship, and fine water doesn't really correlate with strength at all. Age has almost no relationship at this stage, showing that once concrete is fully aged, its strength depends entirely on how it was mixed, not how long it has been sitting.

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

Number of Observations Read	114
Number of Observations Used	114

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	7	6073.86341	867.69477	20.42	<.0001
Error	106	4503.79135	42.48860		
Corrected Total	113	10578			

Root MSE	6.51833	**R-Square**	0.5742
Dependent Mean	45.64247	**Adj R-Sq**	0.5461
Coeff Var	14.28128		

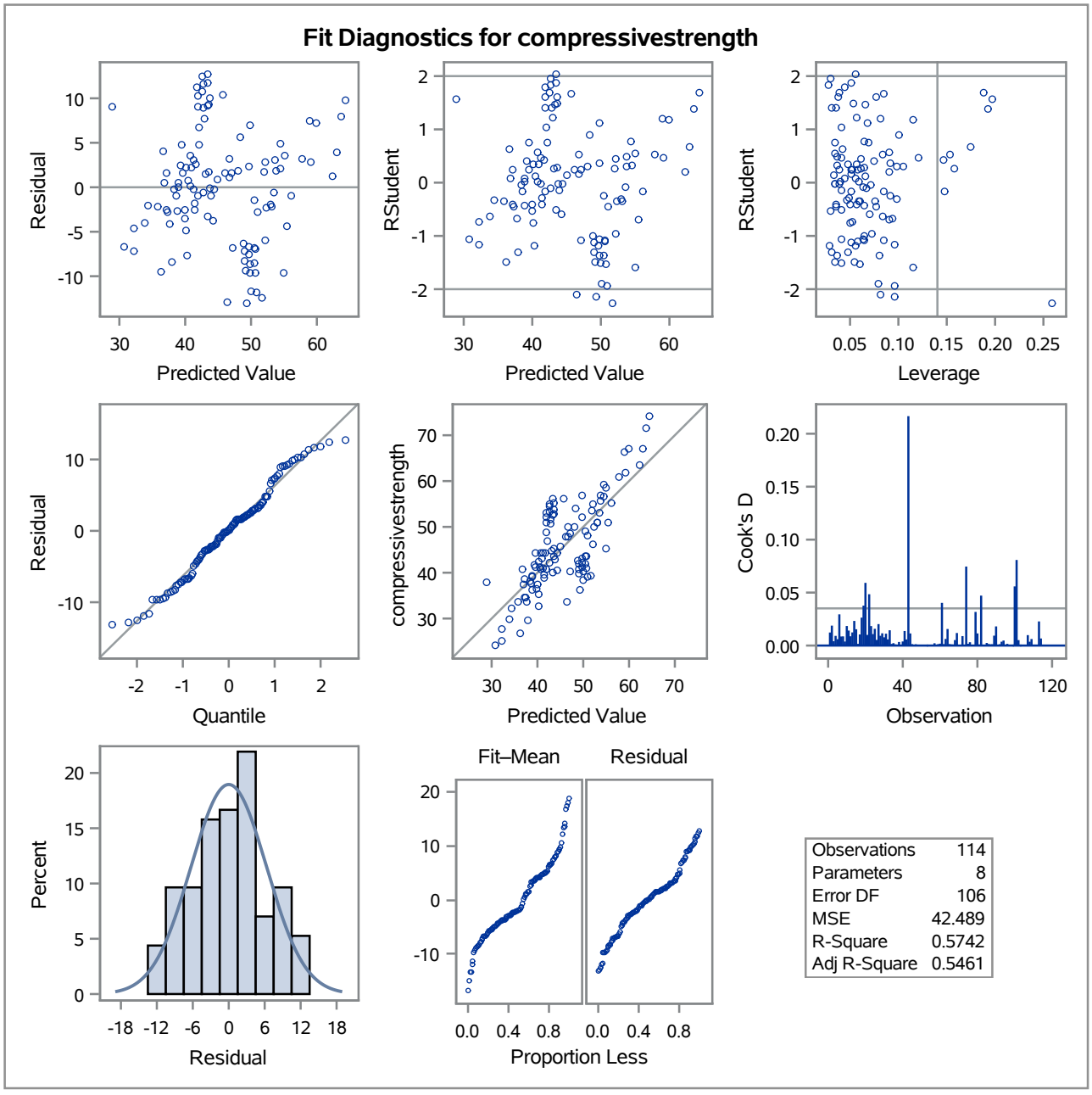
Correlation Matrix of Concrete Variables

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	21.99493	6.46747	3.40	0.0009	0
cementwater	1	13.83188	2.96880	4.66	<.0001	4.56003
slagwater	1	14.24446	3.41561	4.17	<.0001	2.96130
flyashwater	1	3.98048	4.19714	0.95	0.3451	7.22409
superplasticizerwater	1	217.22744	46.35140	4.69	<.0001	4.54073
coarsewater	1	1.67997	1.98355	0.85	0.3989	10.56259
finewater	1	-3.94446	1.84767	-2.13	0.0351	8.64839
age	1	0.00831	0.00896	0.93	0.3557	2.11687

Correlation Matrix of Concrete Variables

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength



I applied a full regression model with all the variables. The model explains about 58% of variation in the model. It shows that strength mainly depends on the mix itself: more cement or more slag makes concrete much stronger, and superplasticizer has a powerful effect. Fly ash and coarse water don't really matter at this age, and too much fine water slightly weakens the mix. Age also stops mattering after 100 days. The full model had some overlapping information between variables, so stepwise selection was used to keep only the ingredients that truly add value. The diagnostic plots indicate that the regression model satisfies all major linear regression assumptions. The residuals are randomly scattered around zero with no noticeable patterns, showing that linearity and constant variance assumption hold. The QQ plot shows residuals that follow a straight line, indicating normality. Most standardized residuals fall within +2 & -2, and the Cook's distance values remain below 1, showing that no influential observations affect the model. Overall, the diagnostics support that the regression model is valid for predicting compressive

Correlation Matrix of Concrete Variables

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

strength of concrete aged at least 100 days.

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

Number of Observations Read	114
Number of Observations Used	114

Stepwise Selection: Step 1

Variable superplasticizerwater Entered: R-Square = 0.1236 and C(p) = 108.1763

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1307.65068	1307.65068	15.80	0.0001
Error	112	9270.00408	82.76789		
Corrected Total	113	10578			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	42.97342	1.08487	129870	1569.08	<.0001
superplasticizerwater	120.67287	30.35951	1307.65068	15.80	0.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable cementwater Entered: R-Square = 0.3024 and C(p) = 65.6789

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	3198.28094	1599.14047	24.05	<.0001
Error	111	7379.37382	66.48085		
Corrected Total	113	10578			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	28.53797	2.87624	6544.74352	98.45	<.0001
cementwater	9.43639	1.76950	1890.63026	28.44	<.0001
superplasticizerwater	147.95166	27.68563	1898.57363	28.56	<.0001

Correlation Matrix of Concrete Variables

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

Stepwise Selection: Step 2

Bounds on condition number: 1.0353, 4.1414

Stepwise Selection: Step 3

Variable slagwater Entered: R-Square = 0.5335 and C(p) = 10.1366

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5643.17400	1881.05800	41.93	<.0001
Error	110	4934.48076	44.85892		
Corrected Total	113	10578			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	14.10223	3.06687	948.48952	21.14	<.0001
cementwater	16.02492	1.70565	3959.68470	88.27	<.0001
slagwater	17.94672	2.43097	2444.89307	54.50	<.0001
superplasticizerwater	201.53475	23.87222	3197.15057	71.27	<.0001

Bounds on condition number: 1.4256, 11.962

Stepwise Selection: Step 4

Variable finewater Entered: R-Square = 0.5569 and C(p) = 6.3051

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5890.94727	1472.73682	34.25	<.0001
Error	109	4686.70749	42.99732		
Corrected Total	113	10578			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	26.34445	5.91806	852.04208	19.82	<.0001
cementwater	14.46274	1.79221	2800.05032	65.12	<.0001
slagwater	14.95091	2.68735	1330.85151	30.95	<.0001
superplasticizerwater	258.82795	33.40454	2581.38385	60.04	<.0001
finewater	-2.60512	1.08523	247.77327	5.76	0.0181

Correlation Matrix of Concrete Variables

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

Stepwise Selection: Step 4

Bounds on condition number: 2.9482, 34.929

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	superplasticizerwater		1	0.1236	0.1236	108.176	15.80	0.0001
2	cementwater		2	0.1787	0.3024	65.6789	28.44	<.0001
3	slagwater		3	0.2311	0.5335	10.1366	54.50	<.0001
4	finewater		4	0.0234	0.5569	6.3051	5.76	0.0181

The variable selection method had the following result:
For concrete that is already 100+ days old, strength is no longer driven by age but by the other ingredients. The stepwise model showed that four factors matter most: cement, slag, superplasticizer, and fine water. Cement, slag, and superplasticizer all increase strength, with superplasticizer having the biggest impact, while fine water slightly reduces strength. Overall, the best model shows that aged concrete's strength depends on how it was mixed rather than how long it has aged.

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

Number of Observations Read	114
Number of Observations Used	114

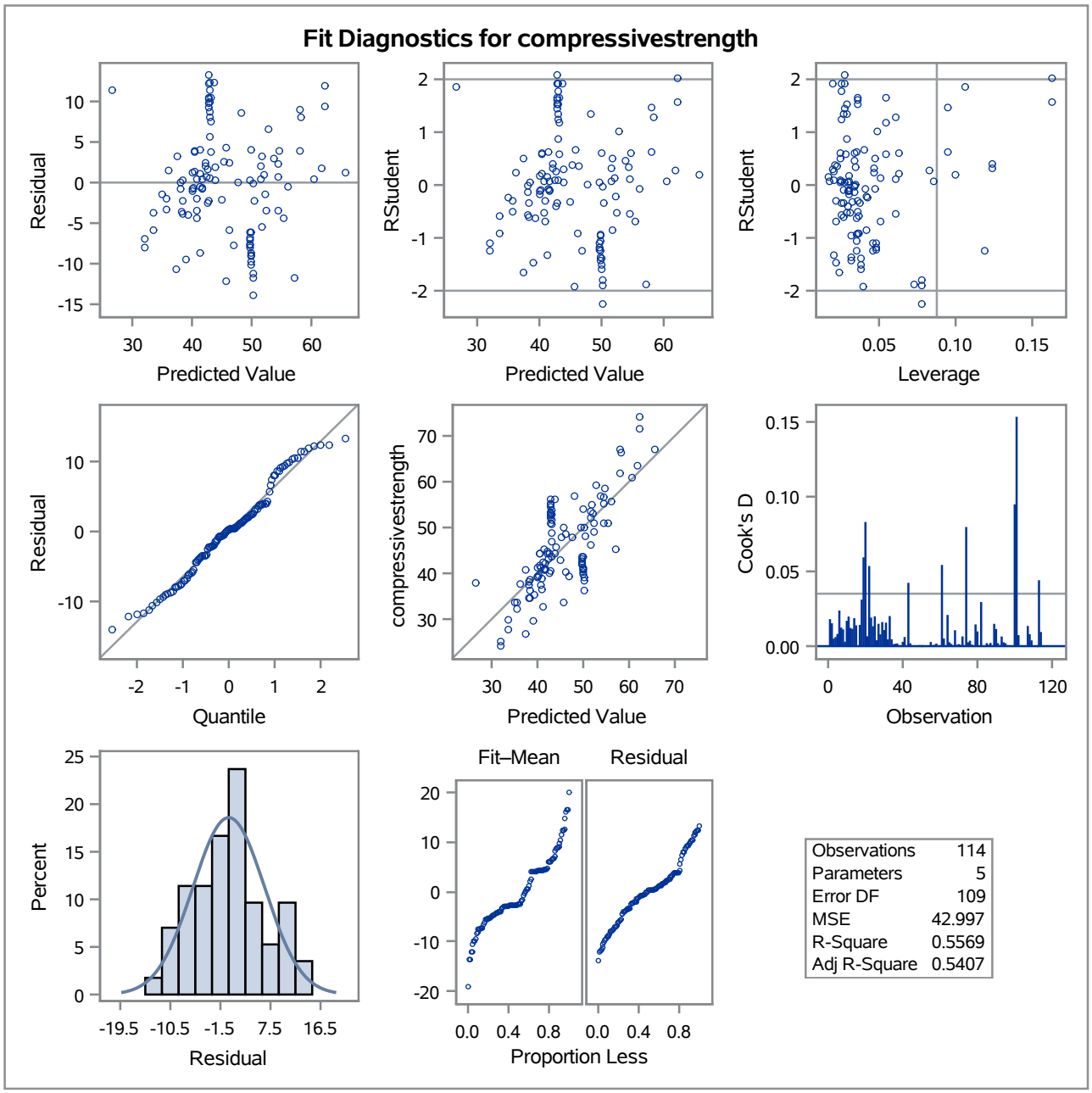
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5890.94727	1472.73682	34.25	<.0001
Error	109	4686.70749	42.99732		
Corrected Total	113	10578			

Root MSE	6.55723	**R-Square**	0.5569
Dependent Mean	45.64247	**Adj R-Sq**	0.5407
Coeff Var	14.36652		

Correlation Matrix of Concrete Variables

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	26.34445	5.91806	4.45	<.0001	0
cementwater	1	14.46274	1.79221	8.07	<.0001	1.64216
finewater	1	-2.60512	1.08523	-2.40	0.0181	2.94821
superplasticizerwater	1	258.82795	33.40454	7.75	<.0001	2.33046
slagwater	1	14.95091	2.68735	5.56	<.0001	1.81144



The final regression model explains about 56% of the strength variation in concrete that is over 100

Correlation Matrix of Concrete Variables

The REG Procedure
Model: MODEL1
Dependent Variable: compressivestrength

days old. It shows that mixes with more cement, more slag, and more superplasticizer are much stronger, while mixes with more fine water tend to be slightly weaker. All selected parameters are statistically significant. The final model included cementwater, slagwater, superplasticizerwater, and finewater, all of which had VIF values below 3, indicating that multicollinearity was no longer a concern. The diagnostic plots look good, residuals are random, mostly centered, and normally distributed, with no outliers.

Problem 4

The aim for this problem is to figure out if the concrete that has been curing for 90–100 days will be strong enough to be moved to the next stage of the construction. The minimum strength needed is 47 MPa, but to be really safe, the manager sets a stricter cutoff of 50 MPa. We construct a logistic regression model that outputs the probability of a sample in this age range achieving at least 50 MPa. This helps us understand which components become the biggest contributors to the concrete passing the strength requirement.

First we filter data to age between 90 and 100 days to predict whether concrete that has cured for 90-100 days will be strong enough (>=50 MPa) before the next construction stage.

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CONC90100
Response Variable (Events)	strong
Response Variable (Trials)	total
Model	binary logit
Optimization Technique	Fisher's scoring

Model Fit Statistics			
		Intercept and Covariates	
Criterion	**Intercept Only**	**Log Likelihood**	**Full Log Likelihood**
AIC	174.127	68.302	68.302
SC	176.980	91.119	91.119
-2 Log L	172.127	52.302	52.302

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	119.8252	7	<.0001
Score	74.5323	7	<.0001
Wald	25.5062	7	0.0006

Correlation Matrix of Concrete Variables

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-13.3338	22.0231	0.3666	0.5449
cementwater	1	7.8442	1.8342	18.2898	<.0001
slagwater	1	8.1291	1.9298	17.7447	<.0001
flyashwater	1	6.3587	2.6200	5.8901	0.0152
superplasticizerwate	1	84.9396	30.5314	7.7397	0.0054
coarsewater	1	-1.8824	0.9703	3.7640	0.0524
finewater	1	-0.7222	0.9446	0.5845	0.4445
age	1	0.0713	0.2265	0.0991	0.7529

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
cementwater	>999.999	70.053	>999.999
slagwater	>999.999	77.233	>999.999
flyashwater	577.495	3.399	>999.999
superplasticizerwate	>999.999	>999.999	>999.999
coarsewater	0.152	0.023	1.019
finewater	0.486	0.076	3.093
age	1.074	0.689	1.674

This logistic regression model predicts whether a 90 to 100-day concrete sample will be strong enough to reach the 50 MPa requirement based on its mix ingredients. The global tests at the top show that the overall model is highly significant ($p < 0.0001$), meaning at least one ingredient is useful for predicting strength. The parameter estimates and p-values tell us which ingredients matter: cement, slag, fly ash, and superplasticizer all have strong positive effects, making the concrete much more likely to pass the 50 MPa cutoff. In contrast, coarse water, fine water, and age are not significant predictors, they do not change the odds of meeting the strength requirement. The odds ratios show that cement, slag, and superplasticizer dramatically increase the chances of strong concrete. This model shows that mixture proportions determine whether 90-day concrete will be strong enough.

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CONC90100
Response Variable (Events)	strong
Response Variable (Trials)	total

Correlation Matrix of Concrete Variables

The LOGISTIC Procedure

Model Information	
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	128
Number of Observations Used	128
Sum of Frequencies Read	128
Sum of Frequencies Used	128

Response Profile		
Ordered Value	**Binary Outcome**	**Total Frequency**
1	Event	51
2	Nonevent	77

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics			
		Intercept and Covariates	
Criterion	**Intercept Only**	**Log Likelihood**	**Full Log Likelihood**
AIC	174.127	73.731	73.731
SC	176.980	85.139	85.139
-2 Log L	172.127	65.731	65.731

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	106.3964	3	<.0001
Score	61.1653	3	<.0001
Wald	25.2923	3	<.0001

Correlation Matrix of Concrete Variables

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-12.4253	2.4447	25.8331	<.0001
cementwater	1	5.1968	1.1403	20.7703	<.0001
slagwater	1	5.9010	1.3536	19.0052	<.0001
superplasticizerwate	1	76.9592	16.6005	21.4921	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cementwater	180.696	19.334	>999.999
slagwater	365.415	25.739	>999.999
superplasticizerwate	>999.999	>999.999	>999.999

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	95.1	**Somers' D**	0.903
Percent Discordant	4.8	**Gamma**	0.903
Percent Tied	0.1	**Tau-a**	0.436
Pairs	3927	**c**	0.951

The stepwise logistic regression identifies which ingredients best predict whether concrete aged 90-100 days will reach the required 50 MPa strength. The procedure starts with no predictors and then adds variables only if they improve the model. Superplasticizer is selected first, followed by cement and slag, showing that these three ingredients play the strongest role in determining whether the concrete will be strong enough. Fine water enters the model but is removed because it does not add meaningful predictive power. The model perfoms well with a c-statistic of 0.951, meaning it can very accurately distinguish strong vs. weak concrete. It is statistically significant and explains the outcome well, indicating that concrete strength is driven by the proportions of cement, slag, and superplasticizer.

Problem 5

The goal of this analysis is to build a classification model that can estimate which age range a concrete sample belongs to based only on its composition and compressive strength. By grouping concrete into age categories, we can give the manager a way to estimate the age of unknown samples and understand which age groups are difficult to distinguish based on their properties.

Correlation Matrix of Concrete Variables

The FREQ Procedure

agegrp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	324	31.46	324	31.46
B	425	41.26	749	72.72
C	91	8.83	840	81.55
D	131	12.72	971	94.27
E	59	5.73	1030	100.00

After dividing the concrete samples into five age groups, we can see that most of the data comes from younger concrete. Groups A (less than 4 weeks old) and B (4-8 weeks old) together make more than 70% of all samples. The middle-aged concrete (Group C) and the 90-180 day range (Group D) are much smaller portions of the dataset, and the oldest group (Group E, 180+ days) is the smallest overall. This tells us that the dataset is heavily weighted toward younger concrete, which may make older age groups a bit harder for the model to classify accurately.

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		0.3379	130.76	<.0001	0.66212192	<.0001	0.08446952	<.0001
2	2	cementwater		0.2753	97.23	<.0001	0.47987193	<.0001	0.13277208	<.0001
3	3	slagwater		0.1539	46.51	<.0001	0.40603300	<.0001	0.15331089	<.0001
4	4	flyashwater		0.1888	59.47	<.0001	0.32936483	<.0001	0.19038078	<.0001
5	5	finewater		0.0731	20.14	<.0001	0.30528034	<.0001	0.20754751	<.0001
6	6	superplasticizerwater		0.0323	8.52	<.0001	0.29540660	<.0001	0.21310871	<.0001
7	7	coarsewater		0.0186	4.84	0.0007	0.28990186	<.0001	0.21640247	<.0001

To improve the age-classification model, I applied STEPDISC to identify the predictors that most strongly separate the five age groups. All seven variables were statistically significant contributors.

The DISCRIM Procedure

Total Sample Size	1030	DF Total	1029
Variables	7	DF Within Classes	1025
Classes	5	DF Between Classes	4

Number of Observations Read	1030
Number of Observations Used	1030

Correlation Matrix of Concrete Variables

The DISCRIM Procedure

Class Level Information					
agegrp	Variable Name	Frequency	Weight	Proportion	Prior Probability
A	A	324	324.0000	0.314563	0.200000
B	B	425	425.0000	0.412621	0.200000
C	C	91	91.0000	0.088350	0.200000
D	D	131	131.0000	0.127184	0.200000
E	E	59	59.0000	0.057282	0.200000

Within Covariance Matrix Information		
agegrp	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
A	7	-10.30850
B	7	-9.99257
C	7	-11.58309
D	7	-11.24646
E	5	-51.09543
Pooled	7	-9.98953

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
2725.661599	112	<.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.

The DISCRIM Procedure

Generalized Squared Distance to agegrp					
From agegrp	A	B	C	D	E
A	-10.30850	-5.60308	3.92801	6.87772	218161983
B	-5.88638	-9.99257	-6.86260	-6.08877	311840244
C	1.11968	-7.16998	-11.58309	-9.99132	706816600
D	2.17521	-7.20291	-9.89684	-11.24646	228176541
E	3.62571	-4.96030	-2.20187	-9.35542	-51.09543

Correlation Matrix of Concrete Variables

The DISCRIM Procedure

Multivariate Statistics and F Approximations					
S=4 M=1 N=508.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.28990186	53.79	28	3675.5	<.0001
Pillai's Trace	0.86560989	40.32	28	4088	<.0001
Hotelling-Lawley Trace	1.94265734	70.62	28	2537.8	<.0001
Roy's Greatest Root	1.66654387	243.32	7	1022	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CONCRETE5
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into agegrp						
From agegrp	**A**	**B**	**C**	**D**	**E**	**Total**
A	162 50.00	27 8.33	18 5.56	0 0.00	117 36.11	324 100.00
B	27 6.35	223 52.47	74 17.41	24 5.65	77 18.12	425 100.00
C	1 1.10	4 4.40	68 74.73	16 17.58	2 2.20	91 100.00
D	0 0.00	3 2.29	45 34.35	30 22.90	53 40.46	131 100.00
E	0 0.00	0 0.00	0 0.00	0 0.00	59 100.00	59 100.00
Total	190 18.45	257 24.95	205 19.90	70 6.80	308 29.90	1030 100.00
Priors	0.2	0.2	0.2	0.2	0.2	

Error Count Estimates for agegrp						
	A	**B**	**C**	**D**	**E**	**Total**
Rate	0.5000	0.4753	0.2527	0.7710	0.0000	0.3998
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

The discriminant analysis shows that concrete mix ratios and compressive strength differ enough across age groups to allow classification (all MANOVA tests $p < 0.0001$), but several groups overlap heavily, especially in the middle ranges. The model correctly classifies very young (50%) and very old concrete (100%) with reasonable accuracy, but performance drops for the mid-age groups, where mixture ratios and strengths are more similar. Overall resubstitution accuracy is about 60%, indicating that the model can identify very old concrete but finds it difficult to distinguish between the middle age categories. Because the age groups differ in their covariance patterns, QDA is more appropriate than LDA, allowing each age group to have its own variability structure.

Correlation Matrix of Concrete Variables

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CONCRETE5
Resubstitution Summary using Quadratic Discriminant Function

CONCLUSION

1. Concrete strength varies widely, from 2 MPa to 83 MPa, and increases strongly with age. Over 70% of samples are younger than 8 weeks, and these younger samples are weaker in strength.
2. Using clustering, the concrete naturally separated into 10 clear groups, and these groups matched curing ages well (some with 100% accuracy, like all 425 samples at 28 days). Strength differences across these clusters were highly significant (ANOVA $F = 60.78$, $p < 0.0001$).
3. For concrete older than 100 days, strength no longer depends on age but on composition. The regression model explained 56% of strength variation, showing cement, slag, and superplasticizer make concrete much stronger, while excess fine water slightly weakens it.
4. Concrete behavior is driven by two forces early curing age and composition. Early age strength rises rapidly with time, but after ~100 days, strength depends on the ingredient ratios (cement, slag, additives) rather than age.
5. For concrete cured 90-100 days, a logistic model predicts whether it reaches the required 50 MPa strength with 95.1% accuracy (c-statistic = 0.951). The strongest predictors are cement, slag, and superplasticizer.
6. When classifying concrete into the 5 age groups, Quadratic Discriminant Analysis achieved 60% accuracy. The youngest and oldest concrete were predicted well, but mid-age groups overlapped and were hardest to classify.