STAT 6800 HW4

Augustine Ennin

November 2025

Q1

```
data electric;
infile "/home/u63997979/sasuser.v94/Elliott and Morrell/Electric.dat";
input houseSize 1-3 familyIncome 6-11 airconCapacity 14-16 applianceIndex 19-23
    familyMembers 26-28 peakHourLoad 31-35;
run;
proc print data=electric(obs=10);
title "First 10 Observation of the Electric Data";
run;
```

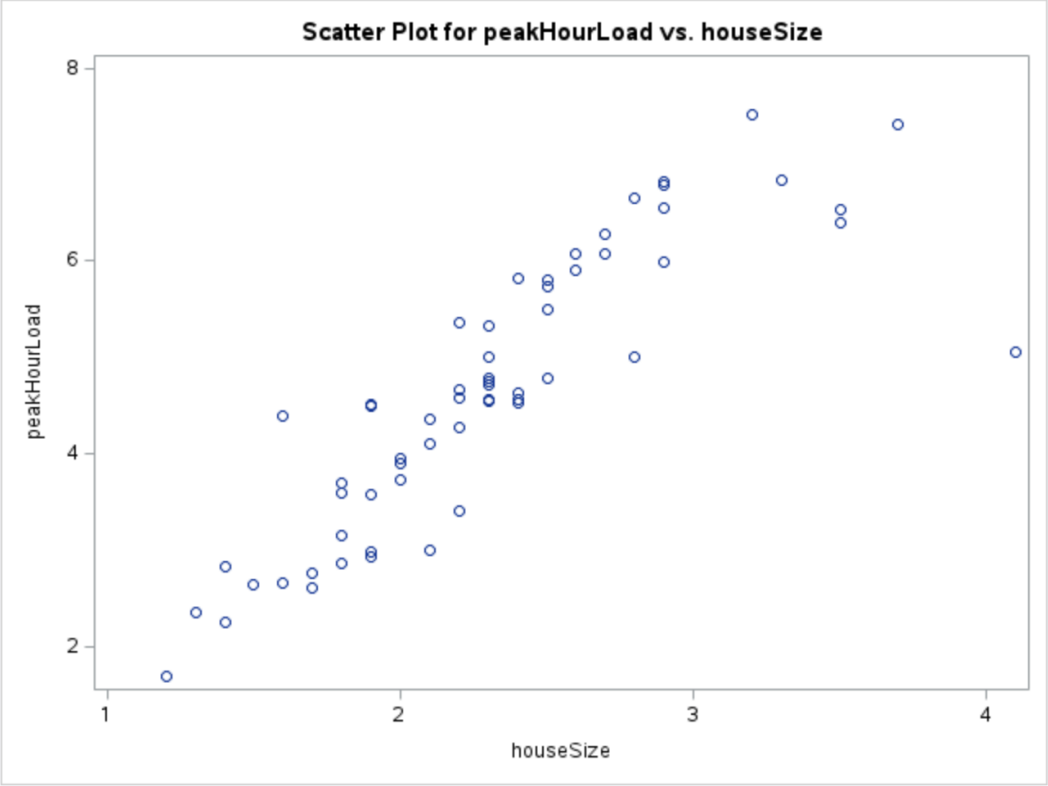
First 10 Observation of the Electric Data

Obs	houseSize	familyIncome	airconCapacity	applianceIndex	familyMembers	peakHourLoad
1	3.2	34.990	7.0	7.789	4.0	7.518
2	1.3	14.160	0.5	3.652	4.0	2.349
3	4.1	22.962	3.0	5.854	1.0	5.059
4	2.3	24.535	5.0	4.975	2.0	5.010
5	1.9	20.614	3.0	4.817	6.0	4.505
6	1.9	20.677	1.0	4.659	1.0	2.976
7	3.3	30.016	6.5	6.054	1.0	6.849
8	2.4	26.341	3.5	7.345	4.0	5.829
9	2.6	28.731	6.5	6.325	3.0	5.910
10	2.9	32.362	3.5	7.700	5.5	5.990

```
%macro reg_analysis(dsn, response, explanatory);
/* Q1(a);
proc sgplot data=&dsn;
scatter x=&explanatory y=&response;
Title "Scatter Plot for &response vs. &explanatory";
run;

/* Q1(b);
proc corr data=&dsn;
var &response &explanatory;
Title "Correlation between &response and &explanatory";
```

```
run;

/* Q1(c);
proc reg data=&dsn;
model &response = &explanatory;
ods exclude DiagnosticsPanel ResidualPlot FitPlot
Title "Simple linear regression of &response on &explanatory";
output out=reg_results p=yhat r=residual;
run;

/* Q1(d)(i);
proc sgplot data=reg_results;
scatter x=yhat y=&response;
lineparm x=0 y=0 slope=1 / lineattrs=(color=blue);
xaxis label="Predicted Values";
yaxis label="&response";
Title "&response vs. Predicted";
footnote "Blue line represent the best fit line";
run;
footnote;

/* Q1(d)(ii);
proc sgplot data=reg_results;
scatter x=yhat y=residual;
refline 0 /axis=y lineattrs=(color=black);
xaxis label="Predicted Values";
yaxis label="Residuals";
Title "Residuals vs. Predicted";
run;

/* Q1(d)(iii);

proc sgplot data=reg_results;
scatter x=&explanatory y=residual;
refline 0 /axis=y lineattrs=(color=black);
xaxis label="&explanatory";
yaxis label="Residuals";
Title "Residuals vs &explanatory";
run;

%mend reg_analysis;
```

```
/* 1.
%reg_analysis(electric, peakHourLoad, houseSize);
```



Correlation between peakHourLoad and houseSize

The CORR Procedure

2 Variables:	peakHourLoad houseSize

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
peakHourLoad	60	4.63792	1.42964	278.27500	1.68500	7.51800
houseSize	60	2.30333	0.59574	138.20000	1.20000	4.10000

Pearson Correlation Coefficients, N = 60 Prob > |r| under H0: Rho=0		
	peakHourLoad	houseSize
peakHourLoad	1.00000	0.85564 <.0001
houseSize	0.85564 <.0001	1.00000

Correlation between peakHourLoad and houseSize

The REG Procedure
Model: MODEL1
Dependent Variable: peakHourLoad

Number of Observations Read	60
Number of Observations Used	60

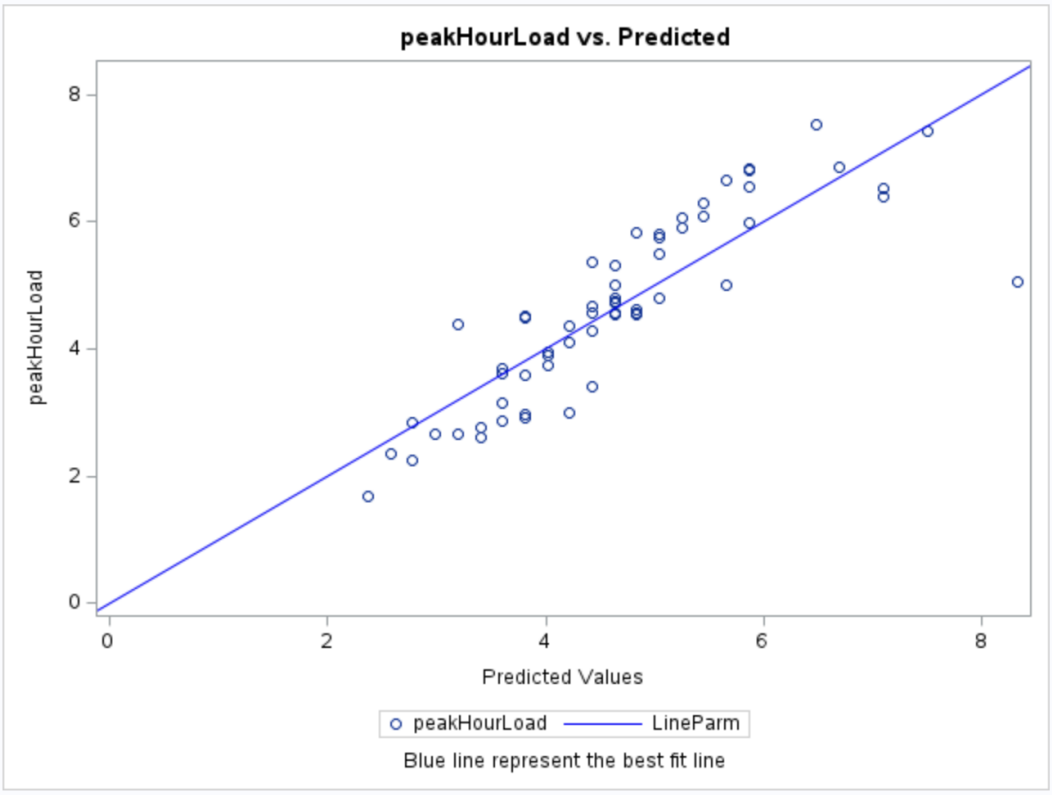
Analysis of Variance

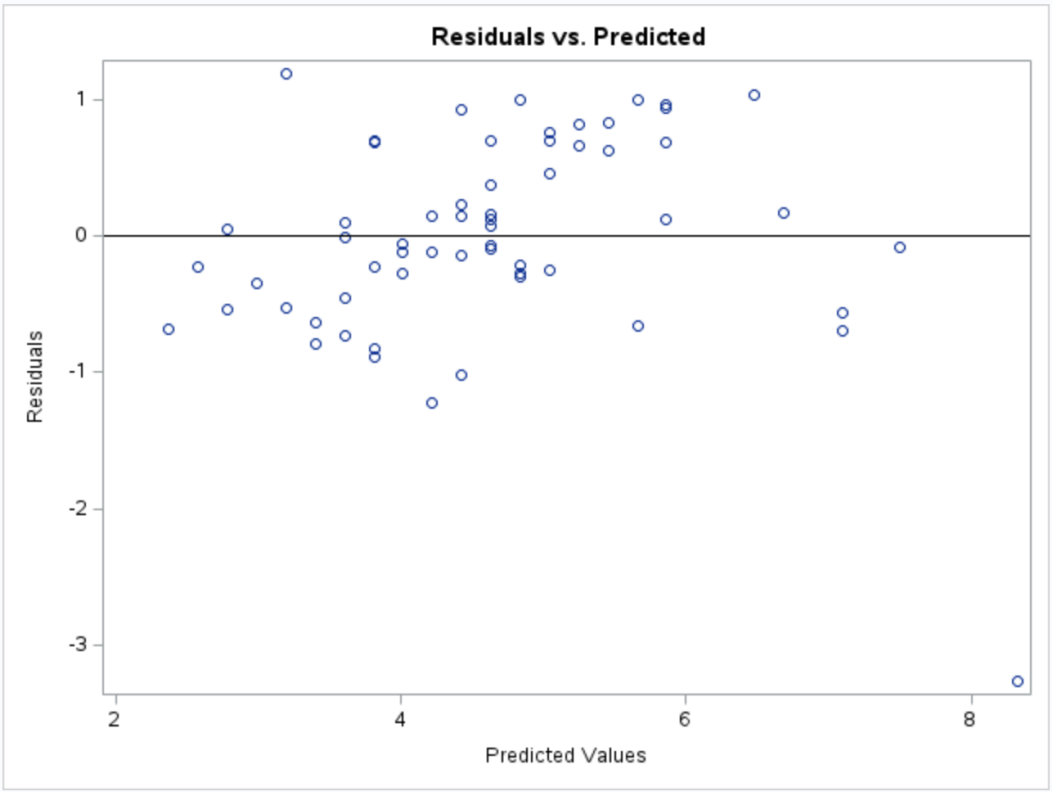
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	88.28471	88.28471	158.51	<.0001
Error	58	32.30444	0.55697		
Corrected Total	59	120.58915			

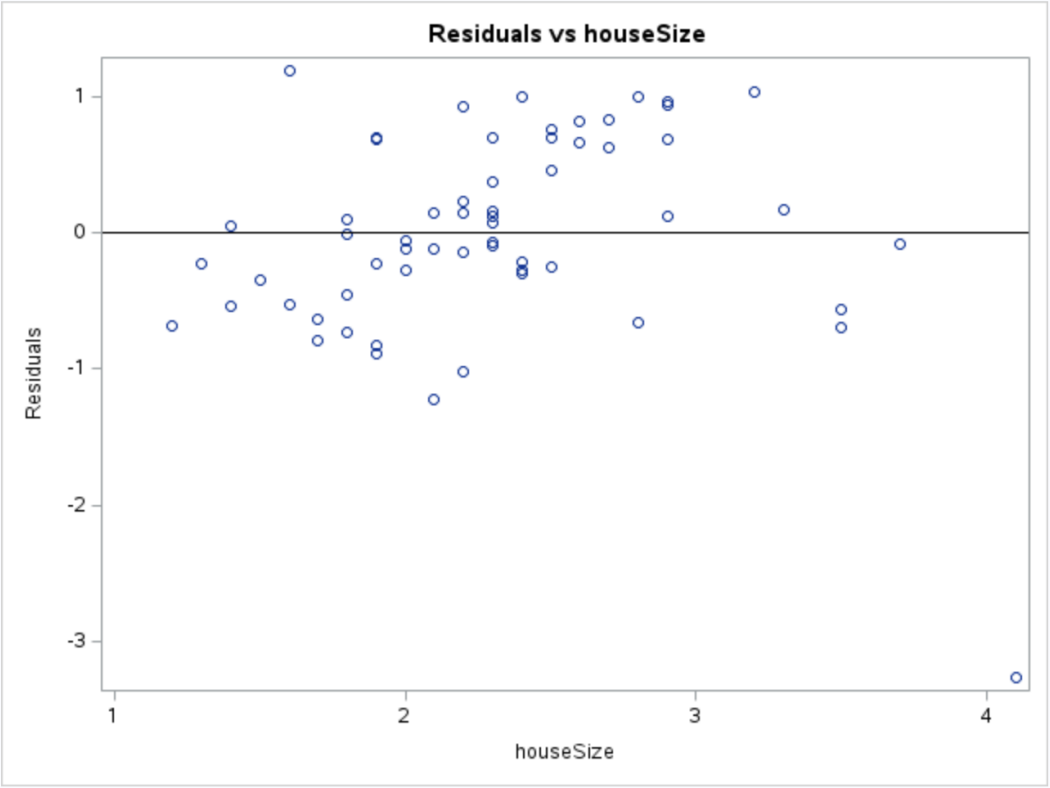
Root MSE	0.74631	R-Square	0.7321
Dependent Mean	4.63792	Adj R-Sq	0.7275
Coeff Var	16.09141		

Parameter Estimates

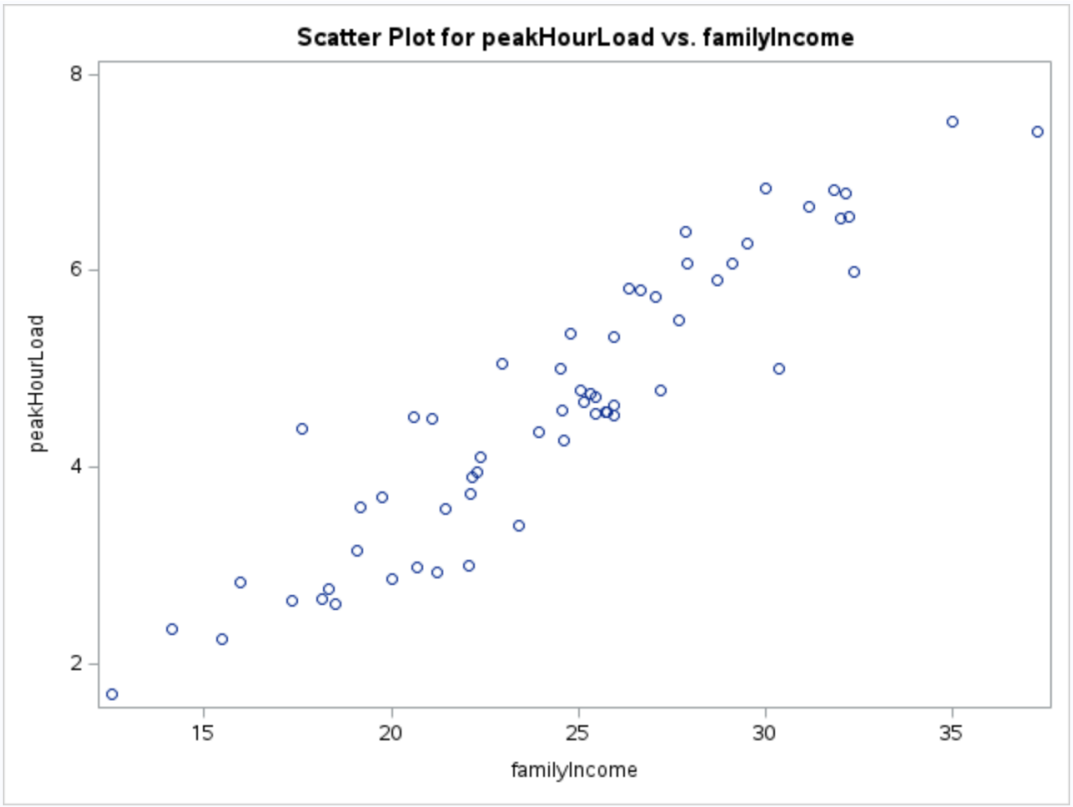
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.09161	0.38782	-0.24	0.8141
houseSize	1	2.05334	0.16309	12.59	<.0001




Residuals vs. Predicted
Residuals
Predicted Values
1
0
-1
-2
-3
2
4
6
8


Residuals vs houseSize
Residuals
houseSize
1
0
-1
-2
-3
1
2
3
4

```
/* 2.
%reg_analysis(electric, peakHourLoad, familyIncome);
```



Correlation between peakHourLoad and familyIncome

The CORR Procedure

2 Variables:	peakHourLoad familyIncome

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
peakHourLoad	60	4.63792	1.42964	278.27500	1.68500	7.51800
familyIncome	60	24.48660	5.23006	1469	12.53100	37.26700

Pearson Correlation Coefficients, N = 60 Prob > |r| under H0: Rho=0		
	peakHourLoad	familyIncome
peakHourLoad	1.00000	0.93003 <.0001
familyIncome	0.93003 <.0001	1.00000

Correlation between peakHourLoad and familyIncome

The REG Procedure
Model: MODEL1
Dependent Variable: peakHourLoad

Number of Observations Read	60
Number of Observations Used	60

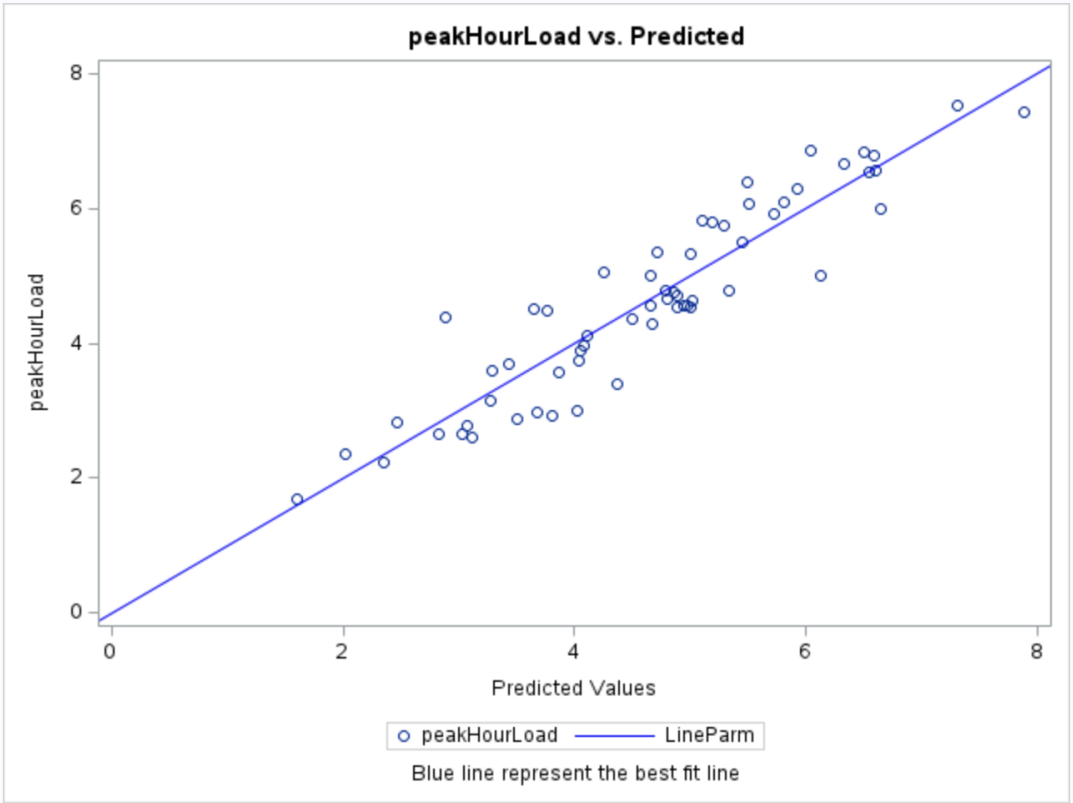
Analysis of Variance

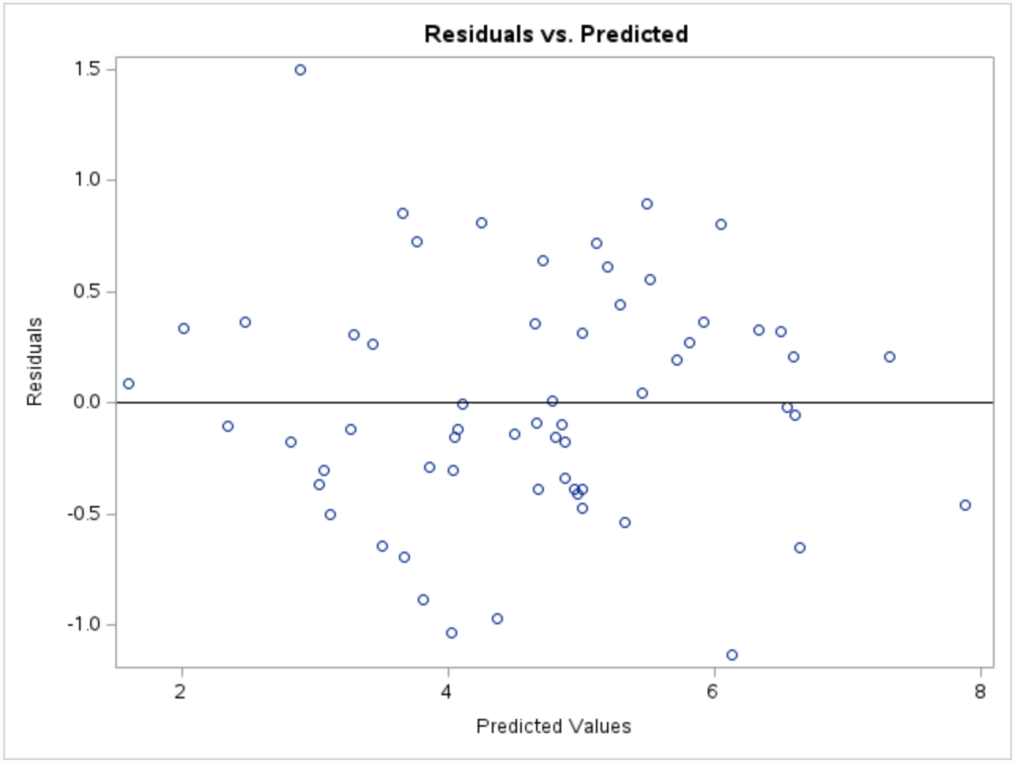
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	104.30541	104.30541	371.52	<.0001
Error	58	16.28375	0.28075		
Corrected Total	59	120.58915			

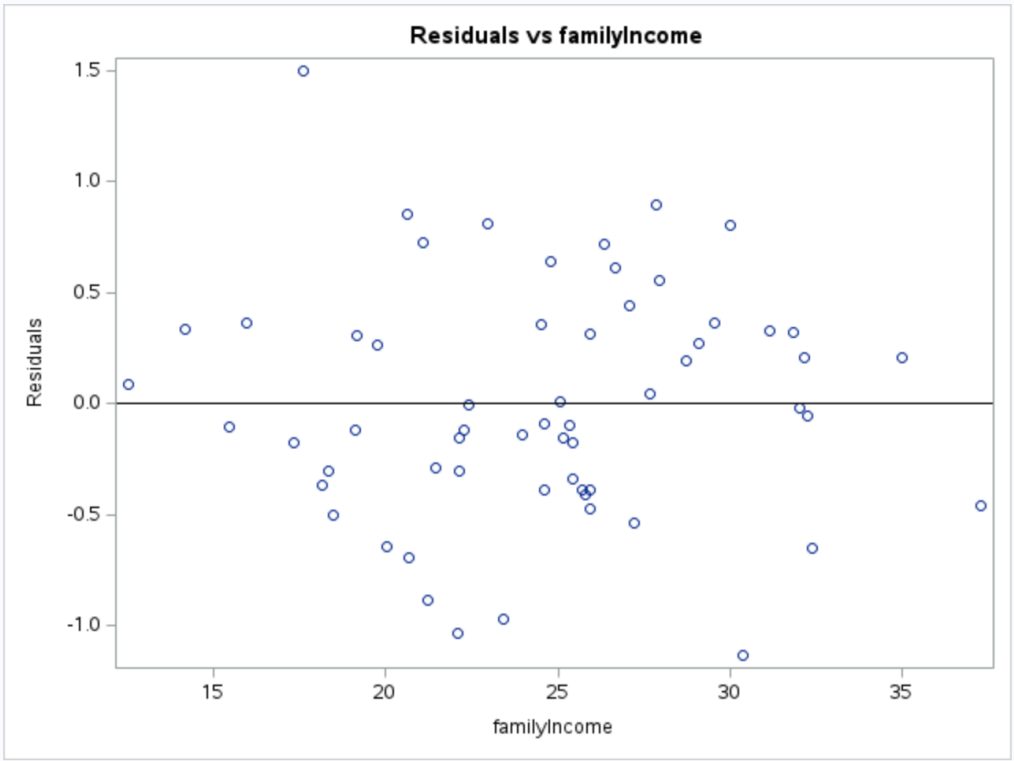
Root MSE	0.52986	R-Square	0.8650
Dependent Mean	4.63792	Adj R-Sq	0.8626
Coeff Var	11.42458		

Parameter Estimates

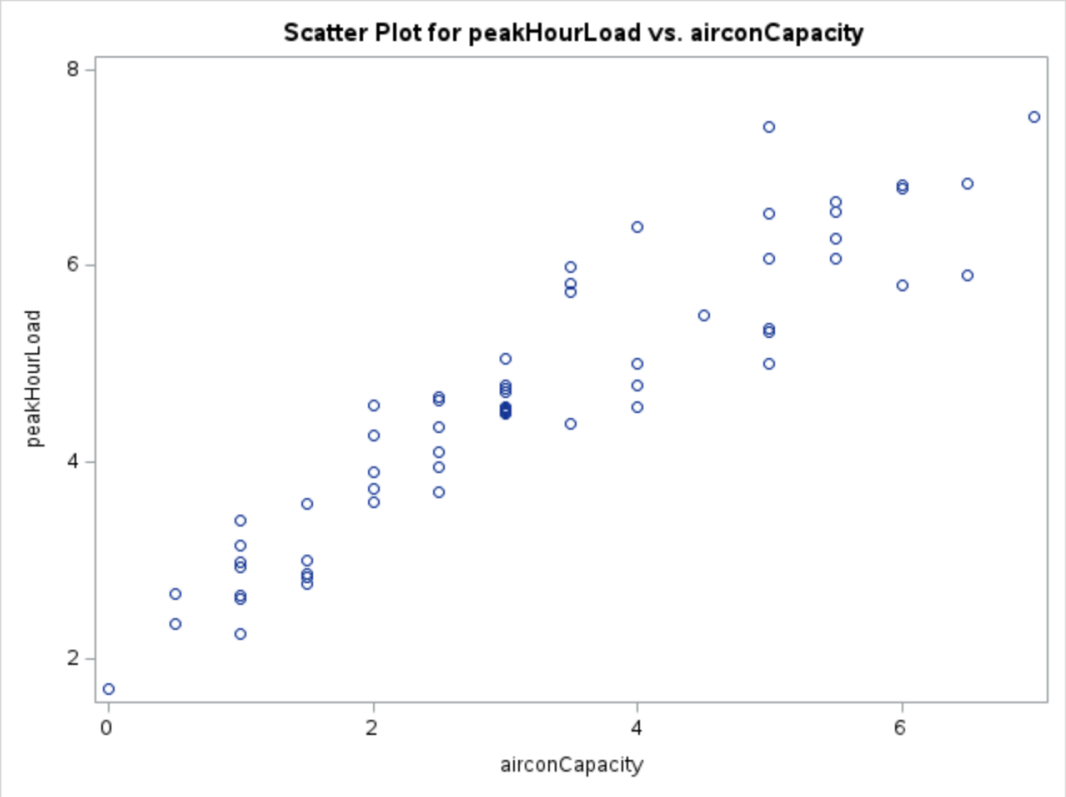
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.58722	0.33013	-4.81	<.0001
familyIncome	1	0.25423	0.01319	19.27	<.0001




Residuals vs. Predicted
Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
2
4
6
8
Predicted Values


Residuals vs familyIncome
Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
15
20
25
30
35
familyIncome

```
1 /* 3.
2 %reg_analysis(electric, peakHourLoad, airconCapacity);
```



Correlation between peakHourLoad and airconCapacity

The CORR Procedure

2 Variables:	peakHourLoad airconCapacity

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
peakHourLoad	60	4.63792	1.42964	278.27500	1.68500	7.51800
airconCapacity	60	3.20000	1.78791	192.00000	0	7.00000

Pearson Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

	peakHourLoad	airconCapacity
peakHourLoad	1.00000	0.92727 <.0001
airconCapacity	0.92727 <.0001	1.00000

Correlation between peakHourLoad and airconCapacity

The REG Procedure
Model: MODEL1
Dependent Variable: peakHourLoad

Number of Observations Read	60
Number of Observations Used	60

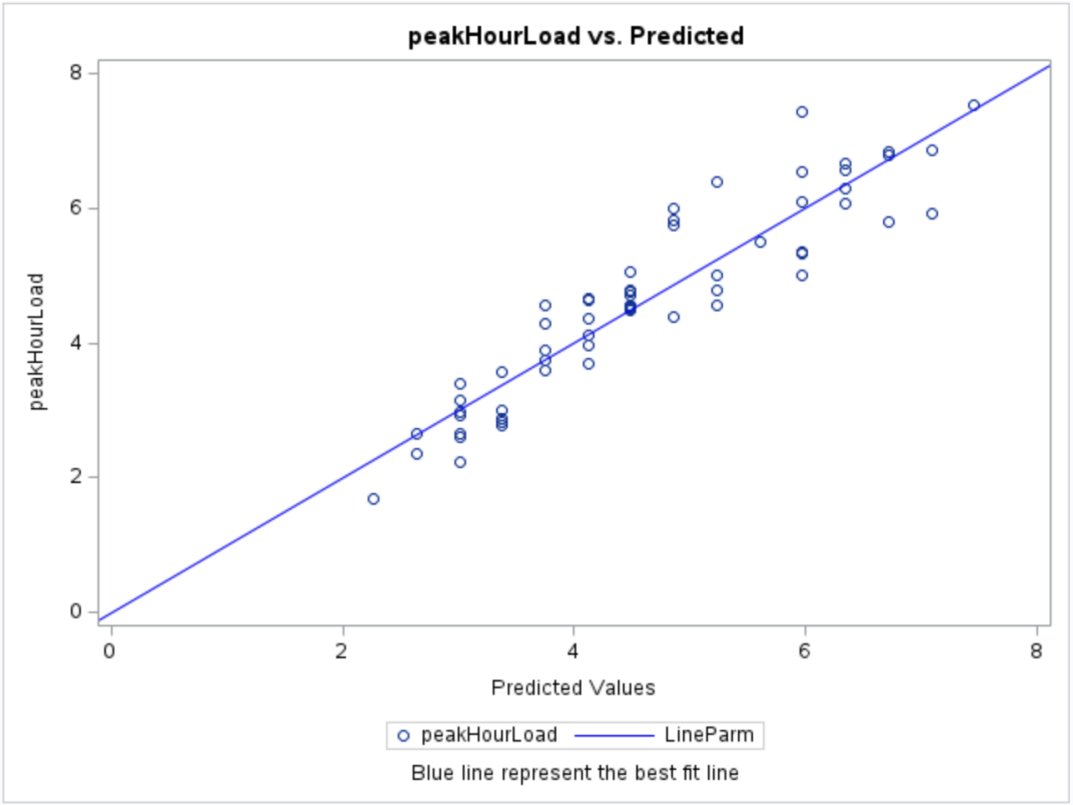
Analysis of Variance

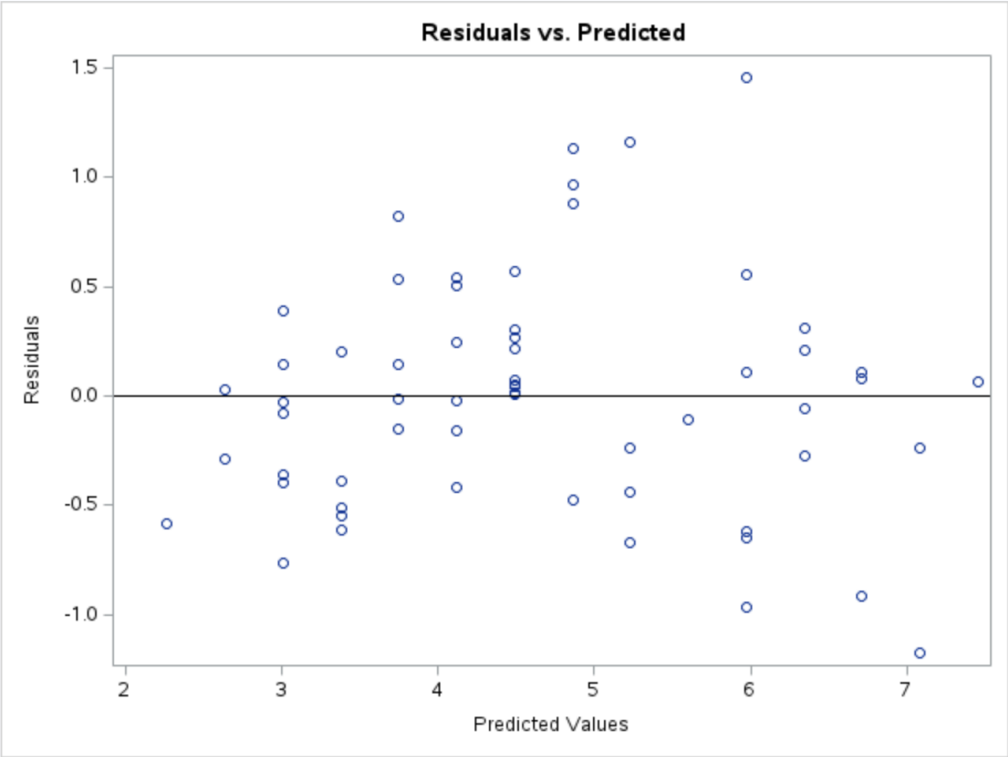
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	103.68699	103.68699	355.80	<.0001
Error	58	16.90217	0.29142		
Corrected Total	59	120.58915			

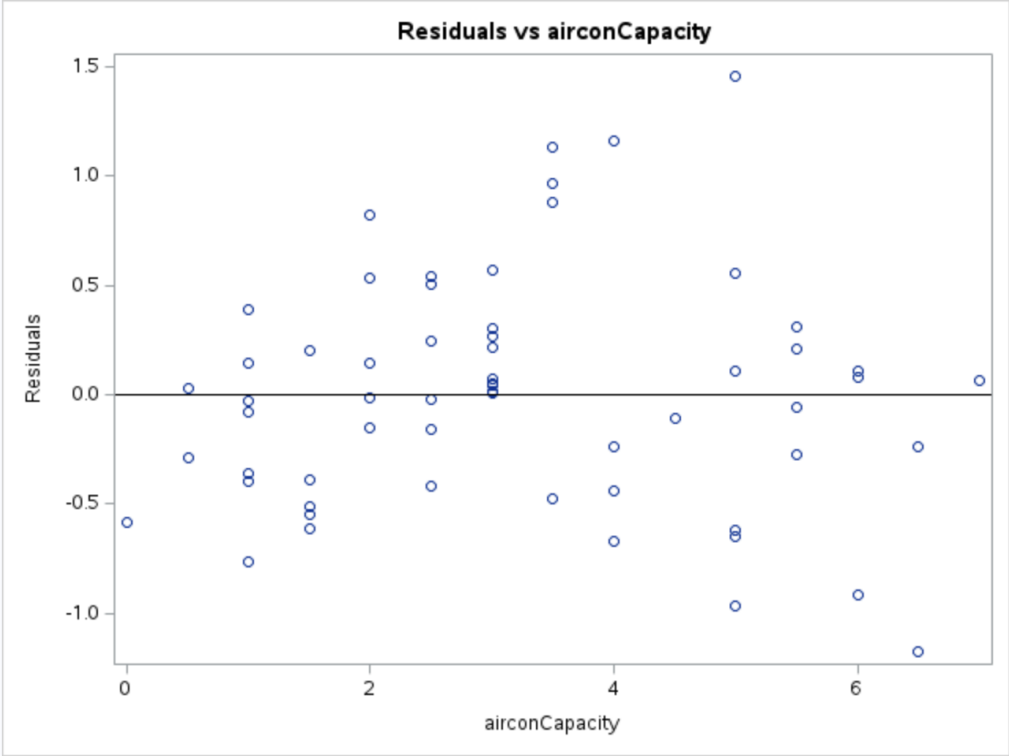
Root MSE	0.53983	R-Square	0.8598
Dependent Mean	4.63792	Adj R-Sq	0.8574
Coeff Var	11.63950		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.26523	0.14380	15.75	<.0001
airconCapacity	1	0.74147	0.03931	18.86	<.0001







Q2

```
1 data golf;
```

```
infile "/home/u63997979/sasuser.v94/Elliott and Morrell/golf.dat.txt";
input golfer compression material distance;
run;
proc print data=golf(obs=10);
title "First 10 Observation of the Golf Data";
run;
```

First 10 Observation of the Golf Data

Obs	golfer	compression	material	distance
1	1	90	1	241.00
2	1	90	1	218.00
3	1	90	1	200.25
4	2	90	1	219.50
5	2	90	1	283.42
6	2	90	1	242.75
7	3	90	1	245.83
8	3	90	1	273.17
9	3	90	1	251.58
10	4	90	1	309.00

```
%macro golf_analysis(start, stop);
%do loop=&start %to &stop;

/* Descriptive statistics;
proc means data=golf n mean std min q1 median q3 max range skewness;
where material = &loop;
var distance;
Title "Descriptive statistics for materials type &loop";
run;

/* Histogram with normal curve;
proc sgplot data=golf;
where material = &loop;
histogram distance / transparency=0.5;
```

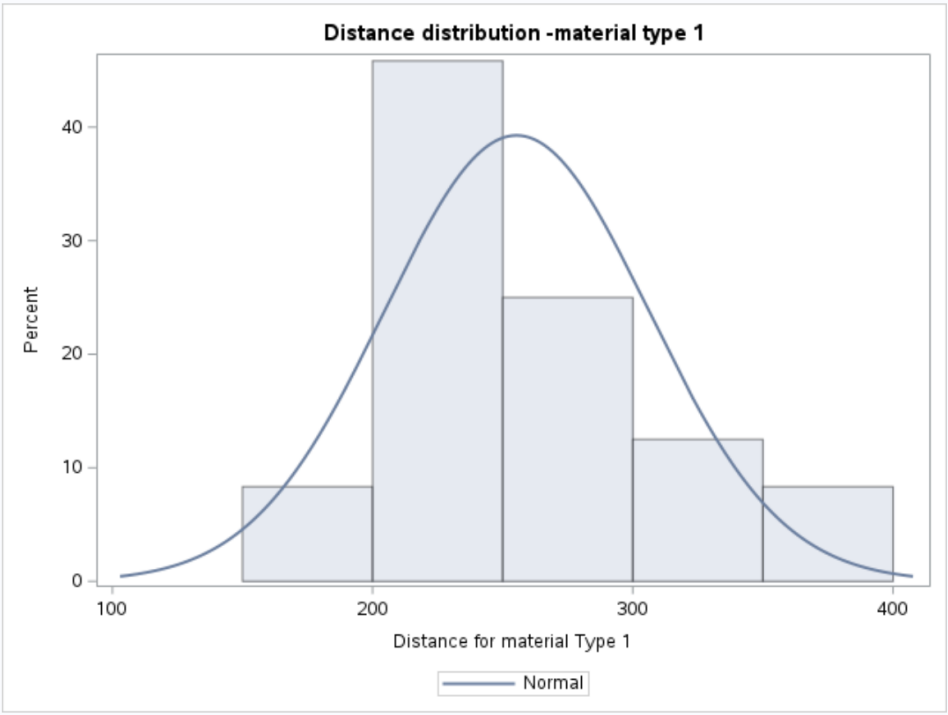
```
density distance / type=normal;
xaxis label="Distance for material Type &loop";
title "Distance distribution -material type &loop";
run;
title;
%end;
%mend golf_analysis;
```

```
golf_analysis(1, 3);
```

Descriptive statistics for materials type 1

The MEANS Procedure

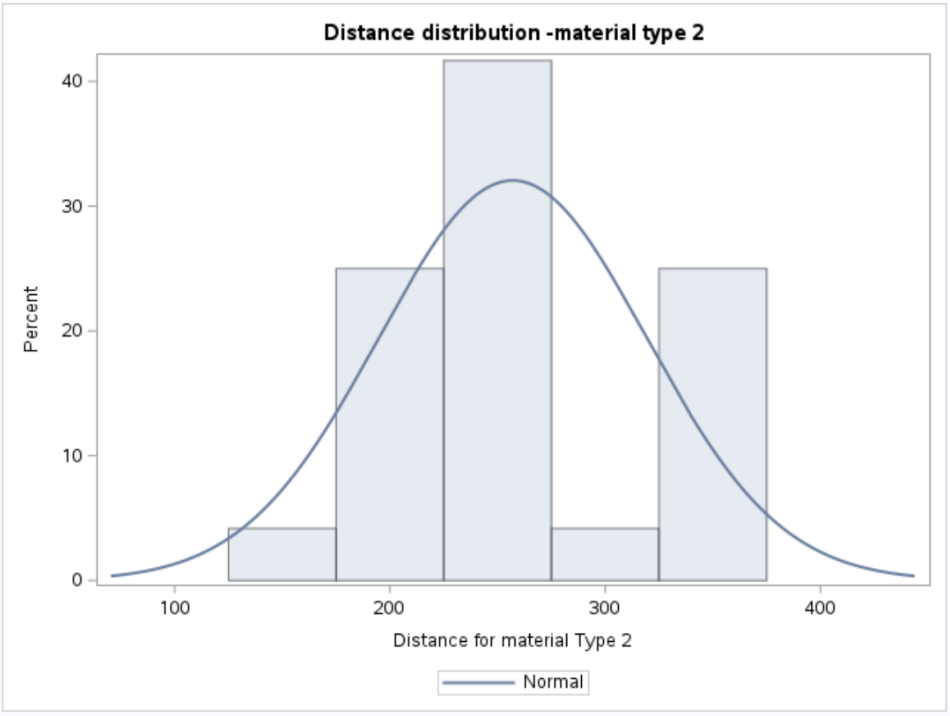
Analysis Variable : distance									
N	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Skewness
24	255.3333333	50.7955032	162.0000000	219.2500000	244.2900000	283.5000000	370.6700000	208.6700000	0.6166245



Descriptive statistics for materials type 2

The MEANS Procedure

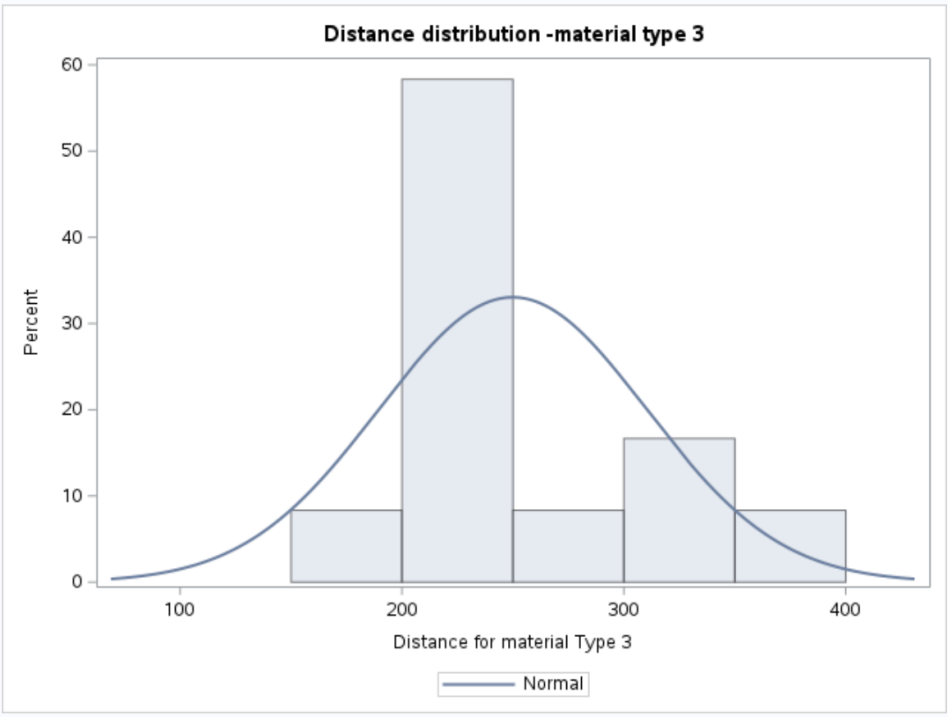
Analysis Variable : distance									
N	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Skewness
24	257.0408333	62.2243643	147.5000000	208.7500000	241.5850000	320.9150000	365.2500000	217.7500000	0.4445904



Descriptive statistics for materials type 3

The MEANS Procedure

Analysis Variable : distance									
N	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Skewness
24	249.8879167	60.3614560	164.0000000	208.9600000	225.4150000	297.7900000	385.7500000	221.7500000	0.9280819



The mean distances are very similar across material types, ranging from about 250 to 257, indicating that, on average, materials perform similarly. However, they differ in variability and distribution shape. Material Type 1 is the most consistent since it has the smallest standard deviation and range of distance, while Types 2 and 3 show greater variability in distance. Although all the materials have a slightly right-skewed distribution, Type 3 shows the strongest. Overall, Material Type 1 is considered more stable and reliable, while Types 2 and 3 exhibit greater fluctuations in their results, especially Type 3, despite the average distances being similar among these materials.

Q3

```
data grade;
infile "/home/u63997979/sasuser.v94/Elliott and Morrell/Grades.dat.txt";
input id $ 1-3 gender $ 5 class 7 quiz_grade 9-10 first_grade 12-14 second_grade
    16-18 lab_grade 20-22 final_grade 25-27;
run;
proc print data=grade(obs=10);
title "First 10 Observation of the Grade Data";
run;
```

First 10 Observation of the Grade Data

Obs	id	gender	class	quiz_grade	first_grade	second_grade	lab_grade	final_grade
1	air	f	4	50	93	93	98	162
2	aln	m	4	49	95	98	97	175
3	bam	m	4	39	63	84	95	95
4	bag	f	3	46	92	96	88	150
5	bes	f	4	45	100	98	96	191
6	bec	f	3	44	98	100	85	175
7	bej	m	3	41	86	86	94	138
8	bis	f	4	50	100	100	99	166
9	blc	m	4	50	95	97	96	162
10	boc	f	4	48	71	100	97	143

```
%macro grade_analysis(class_value);
%if &class_value = 2 %then %do;

/* Sophomore: Descriptive statistics for first grade and second grade;
proc means data=grade n mean std min q1 median q3 max range skewness;
where class = &class_value;
var first_grade second_grade;
title "Descriptive Statistics for Sophomore First and Second Exams Grade";
run;
%end;

%else %if &class_value = 3 %then %do;

/* Junior: Plots for quiz grade and lab grade;
proc sgplot data= grade;
where class=&class_value;
histogram quiz_grade;
title "Distribution for Quiz Grade with Normality Curve for Junior Class";
density quiz_grade / type=normal;
run;

proc sgplot data=grade;
where class = &class_value;
histogram lab_grade;
title "Distribution for Lab Grade with Normality Curve for Junior Class";
density lab_grade / type=normal;
run;
%end;

%else %if &class_value = 4 %then %do;

/* Senior: Normality test for final grade;
proc univariate data=grade normal;
where class = &class_value;
var final_grade;
probplot final_grade / normal(mu=est sigma=est);
run;
%end;

%mend grade_analysis;
```

```
1 %grade_analysis(2);
```

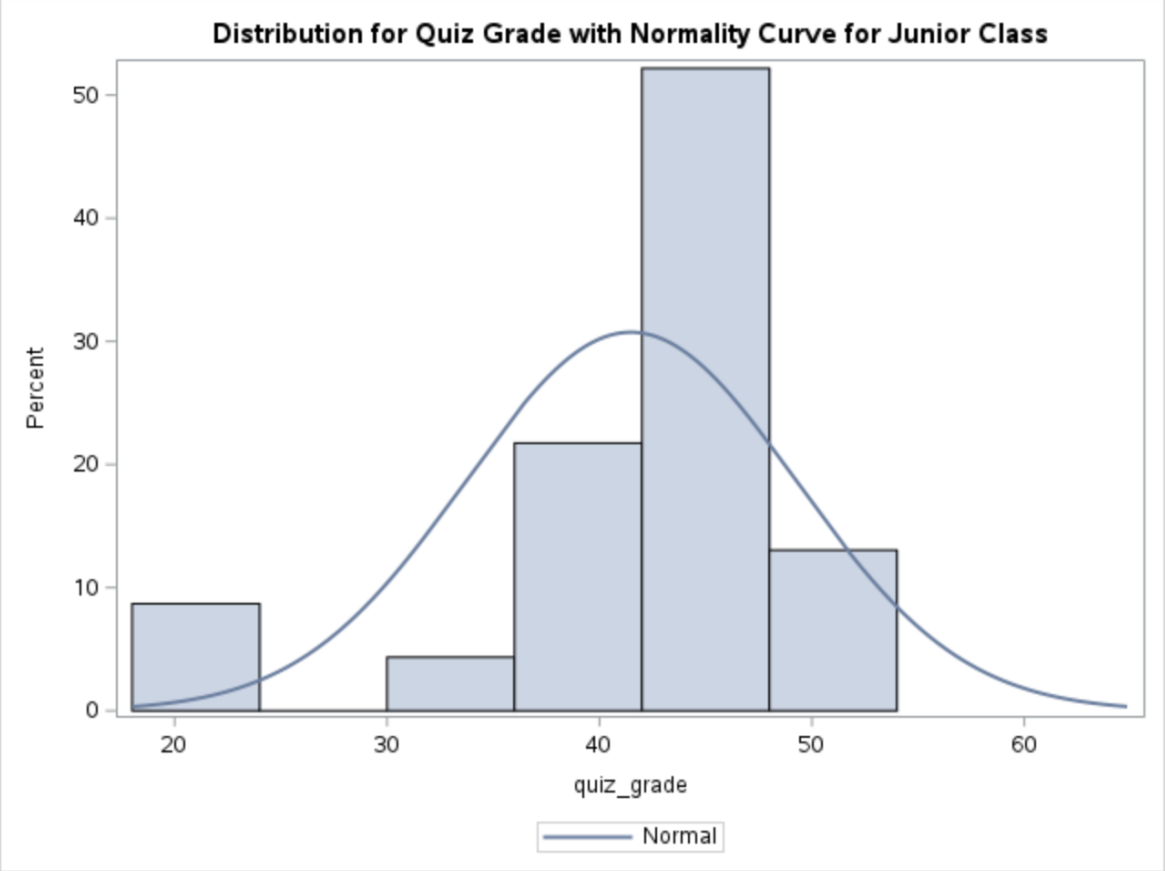
Descriptive Statistics for Sophomore First and Second Exams Grade

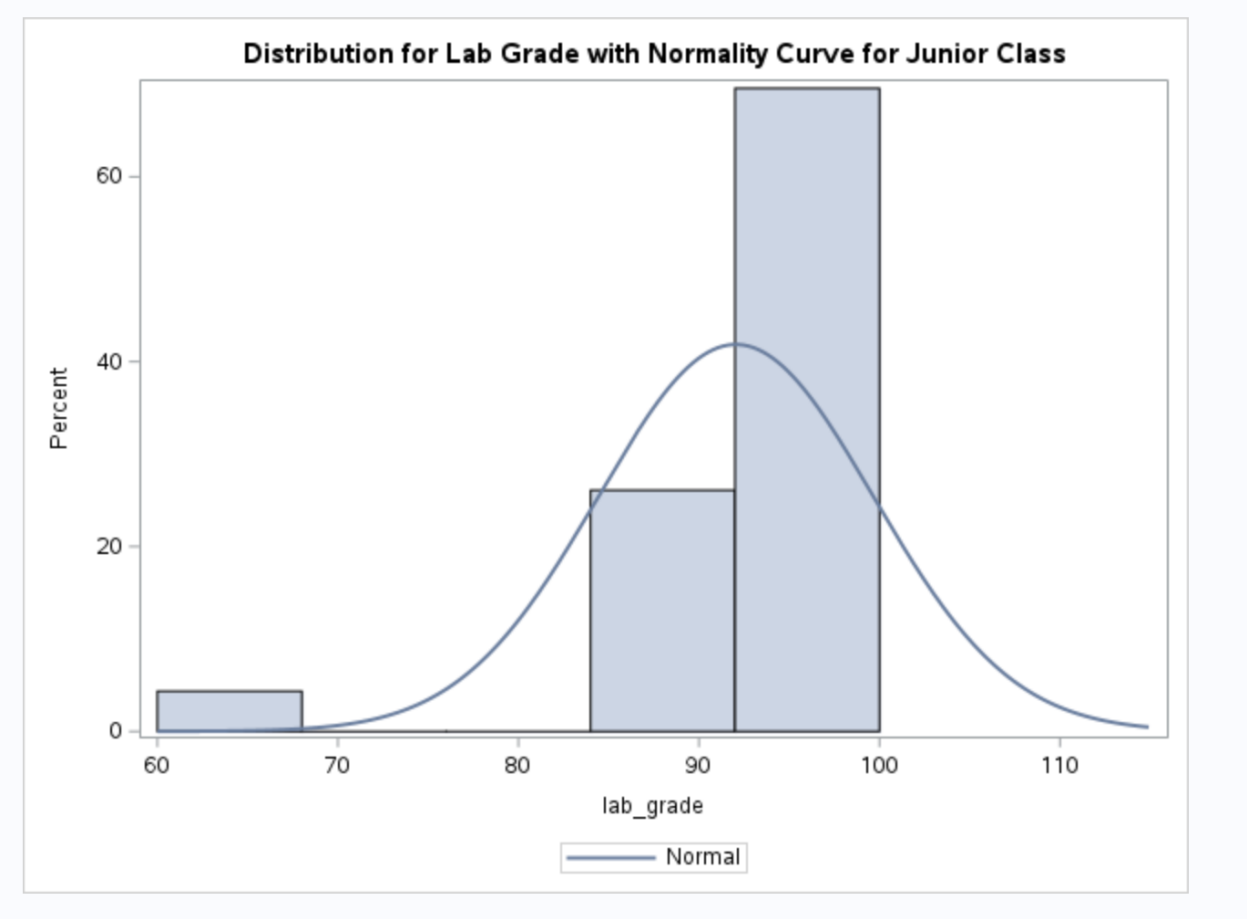
The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Skewness
first_grade	2	79.0000000	2.8284271	77.0000000	77.0000000	79.0000000	81.0000000	81.0000000	4.0000000	.
second_grade	2	92.5000000	2.1213203	91.0000000	91.0000000	92.5000000	94.0000000	94.0000000	3.0000000	.

There are only two students in the sophomore class, with average exam scores of 79 and 92.50 in exam 1 and exam 2, respectively.

```
1 %grade_analysis(3);
```





Quiz grades are normally distributed with a central value around 40 and a good spread lab grades are bimodal with peaks around 70 and around 90–100, indicating that there might well be two kinds of performers among juniors, and greater variability in lab performance compared to quizzes.

```
1 %grade_analysis(4);
```

The UNIVARIATE Procedure
Variable: final_grade

Moments			
N	24	Sum Weights	24
Mean	149.041667	Sum Observations	3577
Std Deviation	21.6081288	Variance	466.911232
Skewness	-0.6330708	Kurtosis	0.89085974
Uncorrected SS	543861	Corrected SS	10738.9583
Coeff Variation	14.4980456	Std Error Mean	4.41074083

Basic Statistical Measures			
Location		Variability	
Mean	149.0417	Std Deviation	21.60813
Median	150.5000	Variance	466.91123
Mode	147.0000	Range	96.00000
		Interquartile Range	24.00000

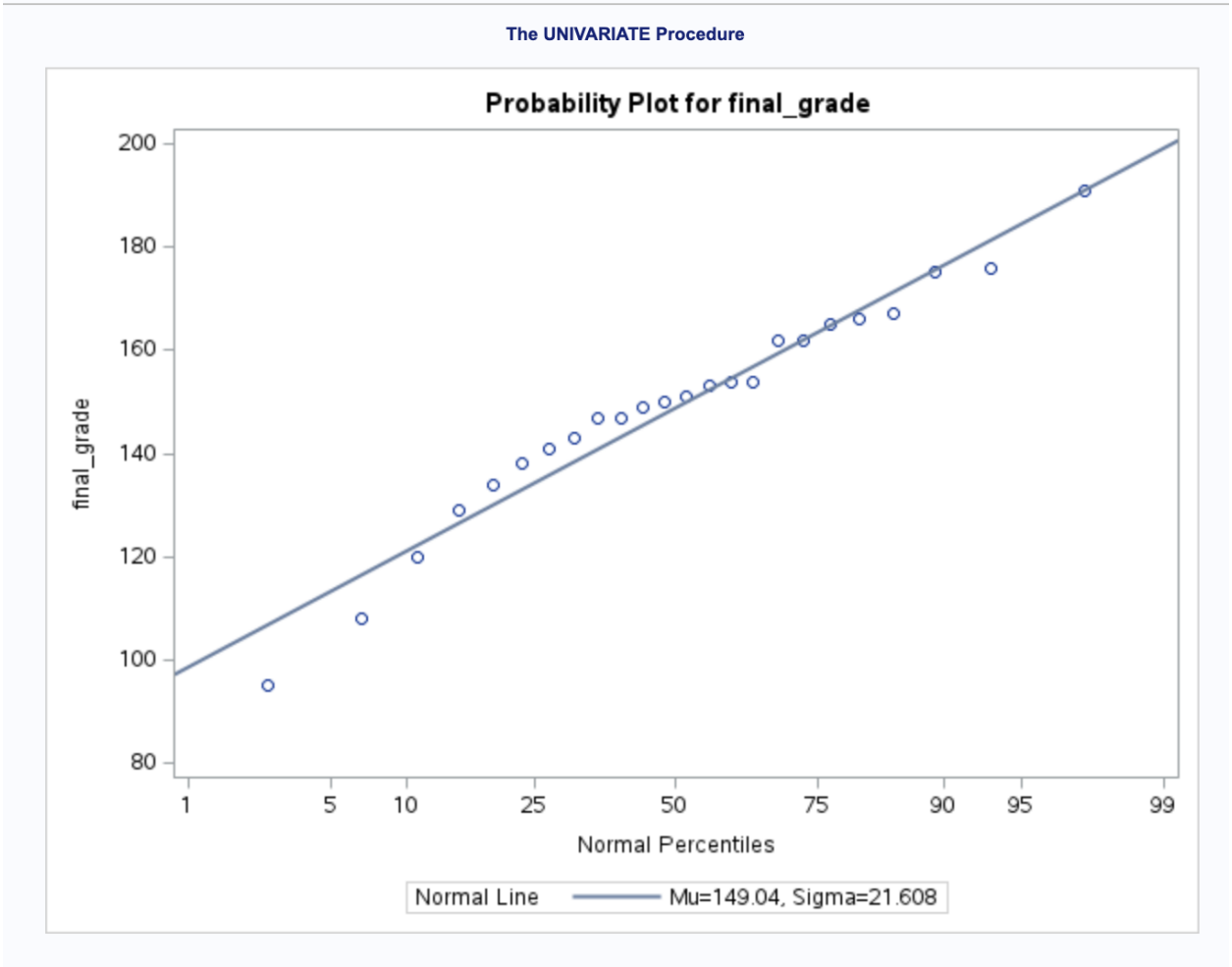
Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	33.79062	Pr > |t|	<.0001
Sign	M	12	Pr >= |M|	<.0001
Signed Rank	S	150	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.965711	Pr < W	0.5632
Kolmogorov-Smirnov	D	0.129028	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.060438	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.365642	Pr > A-Sq	>0.2500

Quantiles (Definition 5)	
Level	Quantile
100% Max	191.0
99%	191.0
95%	176.0
90%	175.0
75% Q3	163.5
50% Median	150.5
25% Q1	139.5
10%	120.0
5%	108.0
1%	95.0
0% Min	95.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
95	3	166	8
108	24	167	48
120	41	175	2
129	13	176	46
134	19	191	5



Q3.(b)

H_0 : Final exam grade for seniors is normally distributed

H_1 : Final exam grade for seniors is not normally distributed

Because the p-value (>.2500) is greater than the significance level $\alpha = 0.05$, we fail to reject the null hypothesis. This indicates that there is insufficient evidence to conclude that the final exam grades for seniors deviate from a normal distribution. Therefore, based on this test, the data do not provide a reason to doubt the assumption of normality for the senior final exam grades.