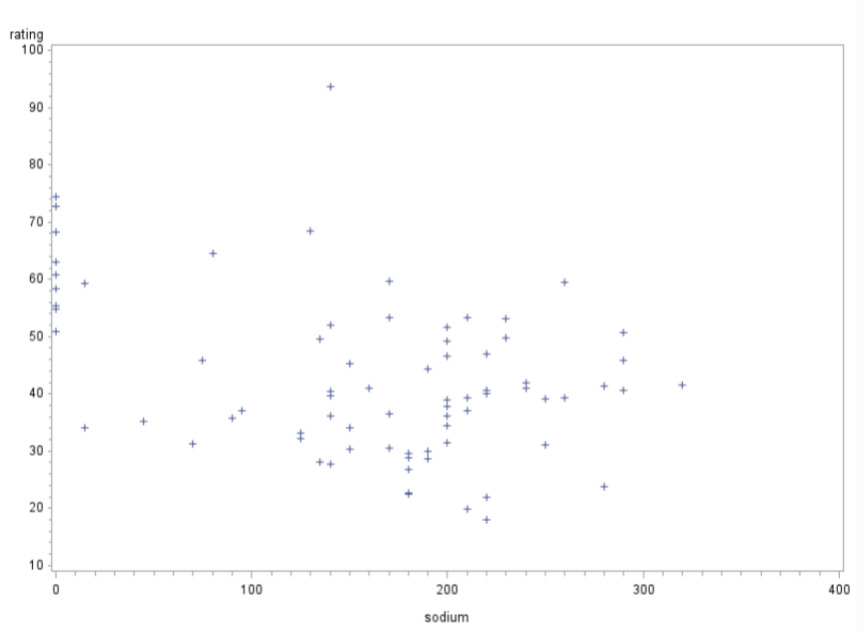
Q1:
According to the table below, seven components should be extracted. Because these seven components, which explain 95.73% of the variation, should be sufficient for almost any application.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	8.70254667	6.19412458	0.5802	0.5802
2	2.50842209	1.22997734	0.1672	0.7474
3	1.27844475	0.36819826	0.0852	0.8326
4	0.91024649	0.49023842	0.0607	0.8933
5	0.42000807	0.13596118	0.0280	0.9213
6	0.28404689	0.02892347	0.0189	0.9402
7	0.25512342	0.02662379	0.0170	0.9573
8	0.22849963	0.06453616	0.0152	0.9725
9	0.16396347	0.08690972	0.0109	0.9834
10	0.07705375	0.00884392	0.0051	0.9886
11	0.06820983	0.01558817	0.0045	0.9931
12	0.05262166	0.02506849	0.0035	0.9966
13	0.02755318	0.00929679	0.0018	0.9984
14	0.01825639	0.01325266	0.0012	0.9997
15	0.00500372		0.0003	1.0000

Q2:
a)



b)
Here is the result of regression. Because “Pr>F” of this model and “Pr>|t|” of variable s are less than 0.005, it is a reliable linear regression model. But the estimated parameter of sodium is -0.06724 which means it affects rating very slightly. Besides, Root MSE is up to 12.95 and Adj R-sq is only 0.1499. These analysis results show that using sodium to predict rating is unreliable.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: rating

Number of Observations Read	77
Number of Observations Used	77

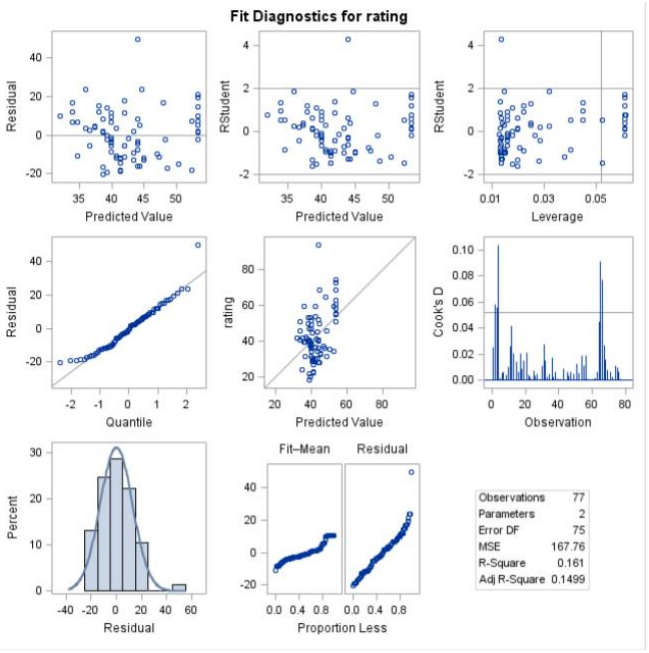
Analysis of Variance

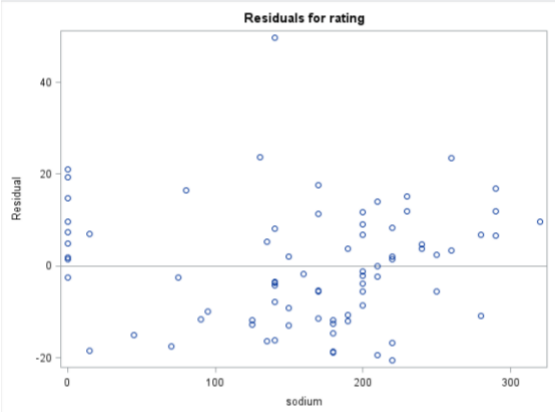
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2415.05236	2415.05236	14.40	0.0003
Error	75	12582	167.75664		
Corrected Total	76	14997			

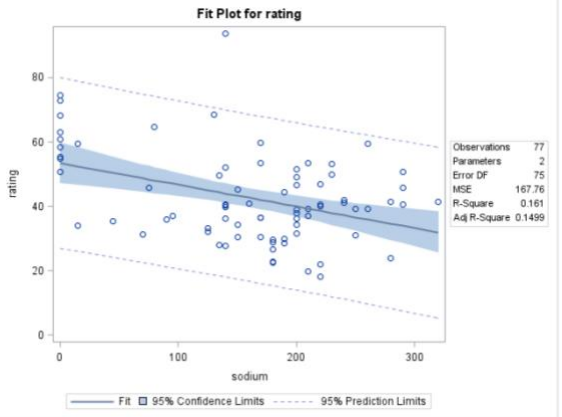
Root MSE	12.95209	R-Square	0.1610
Dependent Mean	42.66570	Adj R-Sq	0.1499
Coeff Var	30.35715		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	53.40270	3.19164	16.73	<.0001
sodium	1	-0.06724	0.01772	-3.79	0.0003







Q3:
a)
Here is the result of regression.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Cat_total

Number of Observations Read	294
Number of Observations Used	294

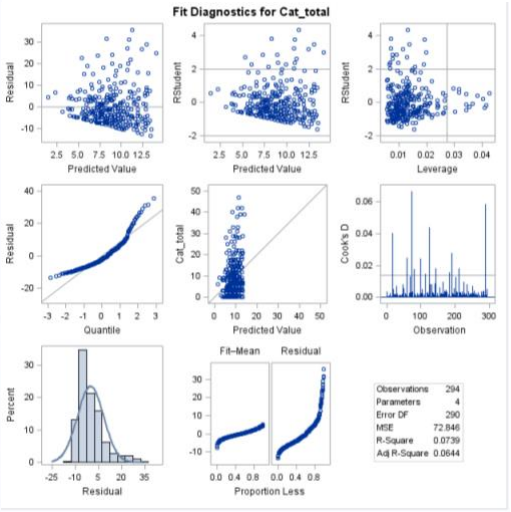
Analysis of Variance

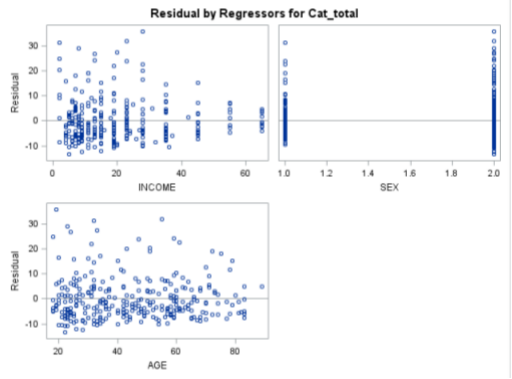
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1686.62825	562.20942	7.72	<.0001
Error	290	21125	72.84634		
Corrected Total	293	22812			

Root MSE	8.53501	R-Square	0.0739
Dependent Mean	8.88435	Adj R-Sq	0.0644
Coeff Var	96.06784		

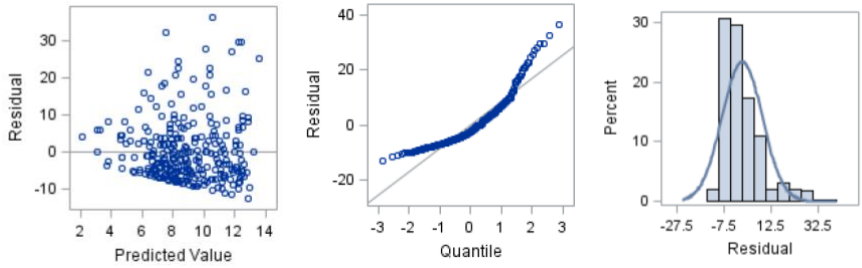
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	12.44897	2.41928	5.15	<.0001	0
INCOME	1	-0.10321	0.03375	-3.06	0.0024	1.07108
SEX	1	1.82034	1.04396	1.74	0.0823	1.03368
AGE	1	-0.09894	0.02809	-3.52	0.0005	1.03824





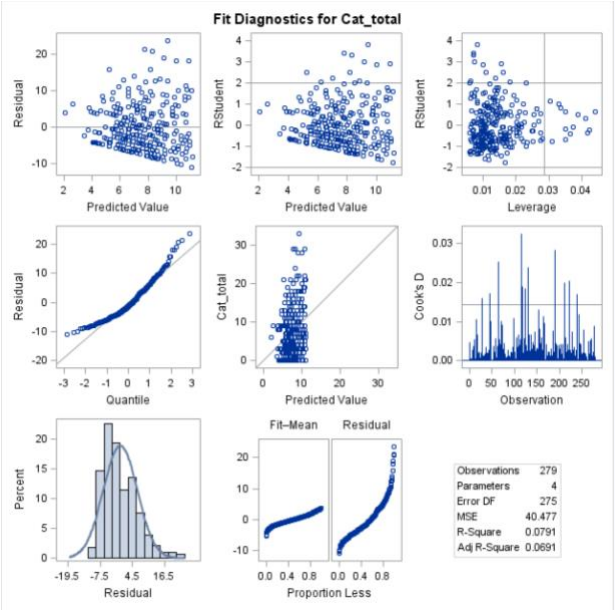
These three graphs show that residuals are not follow a normal distribute.



b)
Determined influence of observations by cook's distance.

	ID	Cook's D Influence Statistic	SEX	AGE	MARITAL	EDUCAT	EMPLOY	INCOME	RELIGION	Cat_01	Cat_02	Cat_03	Cat_04	Cat_05	Cat_06
1	73	0.0665827285	1	32	3	2	3	2	4	2	2	3	2	2	2
2	289	0.0584056734	2	19	4	2	5	28	1	3	3	3	3	3	3
3	124	0.0437553447	2	18	2	2	3	2	3	3	3	2	3	3	1
4	17	0.0401962438	2	23	1	4	1	8	1	3	3	2	3	3	3
5	189	0.028055287	2	24	1	3	1	13	4	3	3	3	3	3	0
6	58	0.0248424486	2	55	2	3	5	23	1	3	3	3	1	1	0
7	99	0.0240708902	1	72	5	2	4	11	1	0	2	3	0	2	0
8	288	0.0233833512	1	61	2	3	1	28	4	2	3	0	0	3	0
9	211	0.0180460232	1	47	2	2	7	23	2	0	3	3	0	3	3
10	144	0.0180203356	2	33	2	2	5	20	4	3	3	3	0	3	3
11	80	0.0179728357	2	75	5	2	5	7	1	1	1	2	1	1	1
12	182	0.0155149029	2	59	5	3	1	13	1	2	3	2	2	3	3
13	112	0.0145887851	2	79	2	3	4	15	1	3	2	0	0	2	0
14	60	0.0143557847	1	20	1	3	1	28	4	2	2	3	2	2	2
15	114	0.0136471014	2	51	1	2	5	2	4	3	2	2	0	1	1

Here is the result of new model. Compared to old model, the distribution of "Cook's D" is evener. And residuals are more likely to a normal distribution. This new model is more reliable because "Pr<|t|" of sex is less than 0.05 and Root MSE decreased.



The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Cat_total

Number of Observations Read	279
Number of Observations Used	279

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	956.16443	318.72148	7.87	<.0001
Error	275	11131	40.47706		
Corrected Total	278	12087			

Root MSE	6.36216	R-Square	0.0791
Dependent Mean	7.53763	Adj R-Sq	0.0691
Coeff Var	84.40524		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.03810	1.85092	5.42	<.0001	0
INCOME	1	-0.06888	0.02552	-2.70	0.0074	1.07255
SEX	1	1.58830	0.79743	1.99	0.0474	1.03258
AGE	1	-0.08186	0.02174	-3.77	0.0002	1.04157

c)

Determined leverage of observations by leverage.

	ID	Leverage	SEX	AGE	MARITAL	EDUCAT	EMPLOY	INCOME	RELIGION	Cat_01	Cat_02	Cat_03	Cat_04	Cat_05	Cat_06
1	121	0.0426035719	1	62	2	3	1	65	4	0	0	0	0	0	0
2	222	0.0414811712	2	52	2	4	5	65	1	0	1	1	0	0	0
3	234	0.0408098259	2	19	1	4	2	65	1	0	0	0	0	1	0
4	95	0.0401788342	2	21	1	5	2	65	3	1	1	0	0	1	0
5	223	0.0382509592	2	37	2	4	2	65	4	0	0	0	0	0	0
6	77	0.0364290448	1	20	1	3	1	65	3	1	0	0	0	0	0
7	162	0.0355001053	1	45	2	5	1	65	1	0	0	0	0	1	0
8	103	0.0343240049	1	37	3	6	1	65	4	1	1	2	0	1	0
9	166	0.0342744942	1	36	2	7	1	65	3	0	0	0	0	0	0
10	270	0.0342744942	1	36	3	7	1	65	4	0	0	0	0	0	0
11	292	0.0316238026	1	64	2	4	1	55	3	0	0	0	0	0	1
12	227	0.0315733891	2	89	5	7	2	31	1	0	2	2	0	2	0
13	168	0.0300588262	2	57	2	5	2	55	1	0	0	0	0	0	0
14	49	0.0296321707	2	56	2	6	1	55	4	0	1	0	0	0	0
15	43	0.0267672895	2	47	2	6	1	55	1	0	0	0	0	0	0

Here is the result of new model. The model only changed slightly by deleting high leverage observations which means that these high leverage observations affect model slightly.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Cat_total

Number of Observations Read	279
Number of Observations Used	279

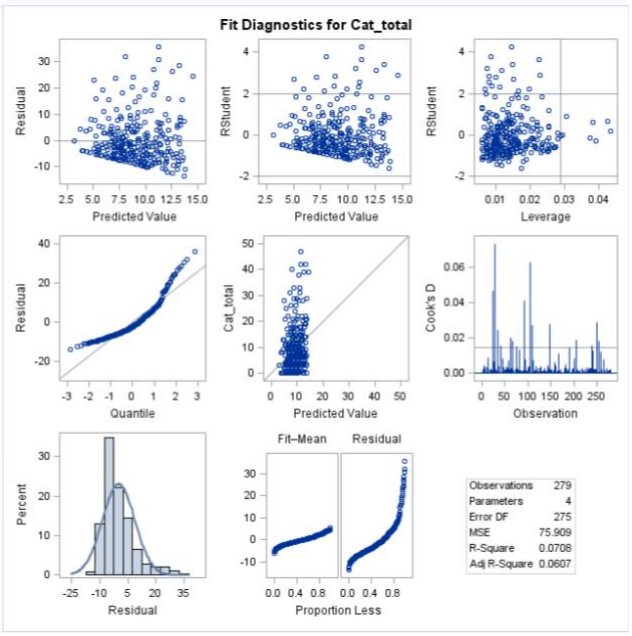
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1591.45191	530.48397	6.99	0.0002
Error	275	20875	75.90888		
Corrected Total	278	22466			

Root MSE	8.71257	R-Square	0.0708
Dependent Mean	9.04659	Adj R-Sq	0.0607
Coeff Var	96.30773		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.15999	2.62384	5.02	<.0001	0
INCOME	1	-0.12075	0.04405	-2.74	0.0065	1.09064
SEX	1	1.75538	1.09812	1.60	0.1111	1.03627
AGE	1	-0.10686	0.02969	-3.60	0.0004	1.05414



Q4:

I distributed all data to 4 tasks by a prime number and let each task to calculate the sum of customer's age and the number of customers. Here are the results form 4 remote tasks.

Task	Sum of age	Number of customers
A	1416932	25199
B	1408472	25061
C	1405243	25024
D	1389963	24716

Then collect results from tasks and calculate average age. It's 56.2061.