Portfolio of Demonstrated Skills for the Certificate in SAS Programming and Data Analysis

Prepared by Zackary Baker
11.24.2020

Table of Contents

STA 5066

DATA MANAGEMENT & ANALYSIS WITH SAS

Homework 4: Reading Raw Data

Code

```
*Problem 6;
filename sales '/courses/d649d56dba27fe300/Data Sets/sales1.dat';
data work.sales;
 infile sales;
 input @1 id
         @8 FirstName $13.
         @21 LastName $18.
         @40 Gender $1.
         @43 JobTitle $20.
         @64 salary dollar7.
         @73 country $2.
         @76 BirthDate mmddyy10.
         @87 HireDate mmddyy10.;
 label id="Employee ID"
         FirstName="First Name"
         LastName="Last Name"
         Gender="Gender"
         JobTitle="Job Title"
         salary="Salary"
         country="Country"
         BirthDate="Birth Date"
         HireDate="Hire Date";
run;

proc contents data=work.sales;
run;

proc print data=work.sales (obs=6);
run;

proc print data=work.sales (obs=6);
format salary dollar7.
       BirthDate mmddyy10.
       HireDate mmddyy10.;
run;
```

```
*Problem 8;
filename sales '/courses/d649d56dba27fe300/Data Sets/sales3.dat';
data work.sales;
 infile sales;
   input @1 EmployeeID
         @8 FirstName $12.
         @21 LastName $18.
         @40 Gender $1.
         @43 JobTitle $20.;
   input @10 Country $2. @;
    if Country='AU' then
     input @1 Salary dollarx7.
           @13 BirthDate ddmmyy10.
          @24 HireDate ddmmyy10.;
   else if Country='US' then
     input @1 Salary dollar7.
           @13 BirthDate mmddyy10.
          @24 HireDate mmddyy10.;
run;

proc contents data=work.sales;
run;

proc print data=work.sales (obs=10);
where Country='AU';
run;

proc print data=work.sales (obs=15);
where Country='US';
run;
```

Selected Output

Problem 6:

Alphabetic List of Variables and Attributes				
#	**Variable**	**Type**	**Len**	**Label**
8	BirthDate	Num	8	Birth Date
2	FirstName	Char	13	First Name
4	Gender	Char	1	Gender
9	HireDate	Num	8	Hire Date
5	JobTitle	Char	20	Job Title
3	LastName	Char	18	Last Name
7	country	Char	2	Country
1	id	Num	8	Employee ID
6	salary	Num	8	Salary

Obs	id	FirstName	LastName	Gender	JobTitle	salary	country	BirthDate	HireDate
1	120102	Tom	Zhou	M	Sales Manager	$10,825	AU	08/11/1969	06/01/1989
2	120103	Wilson	Dawes	M	Sales Manager	$8,797	AU	01/22/1949	01/01/1974
3	120121	Irenie	Elvish	F	Sales Rep. II	$2,660	AU	08/02/1944	01/01/1974
4	120122	Christina	Ngan	F	Sales Rep. II	$2,747	AU	07/27/1954	07/01/1978
5	120123	Kimiko	Hotstone	F	Sales Rep. I	$2,619	AU	09/28/1964	10/01/2007

Problem 8:

Obs	EmployeeID	FirstName	LastName	Gender	JobTitle	Country	Salary	BirthDate	HireDate
1	120102	Tom	Zhou	M	Sales Manager	AU	10825	3510	10744
2	120103	Wilson	Dawes	M	Sales Manager	AU	8797	-3996	5114
3	120121	Irenie	Elvish	F	Sales Rep. II	AU	2660	-5630	5114
4	120122	Christina	Ngan	F	Sales Rep. II	AU	2747	-1984	6756
5	120123	Kimiko	Hotstone	F	Sales Rep. I	AU	2619	1732	17440
6	120124	Lucian	Daymond	M	Sales Rep. I	AU	2648	-233	17226
7	120125	Fong	Hofmeister	M	Sales Rep. IV	AU	3204	-1852	6999
8	120126	Satyakam	Denny	M	Sales Rep. II	AU	2678	10490	17014
9	120127	Sharryn	Clarkson	F	Sales Rep. II	AU	2810	6943	14184
10	120128	Monica	Kletschkus	F	Sales Rep. IV	AU	3089	9691	17106

Homework 8: Manipulating Data

Code

```
*Problem 4;
libname prg1 '/courses/d649d56dba27fe300/STA5066';

proc print data=prg1.customers_ex5 (obs=15);
run;

data work.names (keep=New_Name Name Gender);
   set prg1.customers_ex5;
   if Gender="M" then Title="Mr.";
   else if Gender="F" then Title="Ms.";
   if Customer_ID ne "platinum000-000-00-2806" then New_Name=
propcase(catx(" ", Title, scan(Name,2,' '), scan(Name,3, ' '),
scan(Name,1,', ')));
   else New_Name= propcase(catx(" ", Title, scan(Name,4,' '),
scan(Name,1,' '), scan(Name,2,' '), scan(Name,3,', ')));
run;

proc print data=work.names;
run;

*Problem 9;
libname prg2 '/courses/d649d56dba27fe300/STA5066';

proc print data=prg2.shipped;
run;

data shipping_notes(drop=Ship_Date1 Price1);
set prg2.shipped(rename=(Ship_Date=Ship_Date1) rename=(Price=Price1));
Ship_Date=put(Ship_Date1,date9.);
length Comment $ 21.;
Comment = cat("Shipped on ",Ship_Date);
Price=input(Price1, dollar7.2);
Total = Quantity * Price;
run;

proc print data=shipping_notes noobs;
format Total dollar7.2 Price dollar7.2;
run;
```

Selected Output

Problem 4:

Obs	Customer_ID	Name	Country	Gender	Birth_Date
1	000-000-00-0004	KVARNIQ, James	US	M	27JUN1974
2	Silver000-000-00-0005	STEPHANO, Sandrina	US	F	09JUL1979
3	000-000-00-0009	KRAHL, Cornelia	DE	F	27FEB1974
4	platinum000-000-00-0010	BALLINGER, Karen	US	F	18OCT1984
5	000-000-00-0011	WALLSTAB, Elke	DE	F	16AUG1974
6	Silver000-000-00-0012	BLACK, David	US	M	12APR1969
7	000-000-00-0013	SEPKE, Markus	DE	M	21JUL1988
8	000-000-00-0016	HEYDE, Ulrich	DE	M	16JAN1939
9	000-000-00-0017	EVANS, Jimmie	US	M	17AUG1954
10	000-000-00-0018	ASMUSSEN, Tonie	US	M	02FEB1954
11	000-000-00-0019	FÜßLING, Oliver S.	DE	M	23FEB1964
12	000-000-00-0020	DINELEY, Michael	US	M	17APR1959
13	000-000-00-0023	DEVEREAUX, Tulio	US	M	02DEC1949
14	Silver000-000-00-0024	KLEM, Robyn	US	F	02JUN1959
15	Gold000-000-00-0027	MCCLUNEY, Cynthia	US	F	15APR1969

Obs	Name	Gender	New_Name
1	KVARNIQ, James	M	Mr. James Kvarniq
2	STEPHANO, Sandrina	F	Ms. Sandrina Stephano
3	KRAHL, Cornelia	F	Ms. Cornelia Krahl
4	BALLINGER, Karen	F	Ms. Karen Ballinger
5	WALLSTAB, Elke	F	Ms. Elke Wallstab
6	BLACK, David	M	Mr. David Black
7	SEPKE, Markus	M	Mr. Markus Sepke
8	HEYDE, Ulrich	M	Mr. Ulrich Heyde
9	EVANS, Jimmie	M	Mr. Jimmie Evans
10	ASMUSSEN, Tonie	M	Mr. Tonie Asmussen
11	FÜßLING, Oliver S.	M	Mr. Oliver S. Füßling
12	DINELEY, Michael	M	Mr. Michael Dineley
13	DEVEREAUX, Tulio	M	Mr. Tulio Devereaux
14	KLEM, Robyn	F	Ms. Robyn Klem
15	MCCLUNEY, Cynthia	F	Ms. Cynthia Mccluney

Problem 9:

Product_ID	Quantity	Ship_Date	Comment	Price	Total
240800200021	2	05JAN2007	Shipped on 05JAN2007	$42.45	$84.90
240800200035	6	04JAN2007	Shipped on 04JAN2007	$12.15	$72.90
240200100225	2	04JAN2007	Shipped on 04JAN2007	$77.85	$155.70
210200500002	3	09JAN2007	Shipped on 09JAN2007	$5.70	$17.10

STA 5067

ADVANCED DATA MANAGEMENT & ANALYSIS WITH SAS

Homework 7: Complex Queries

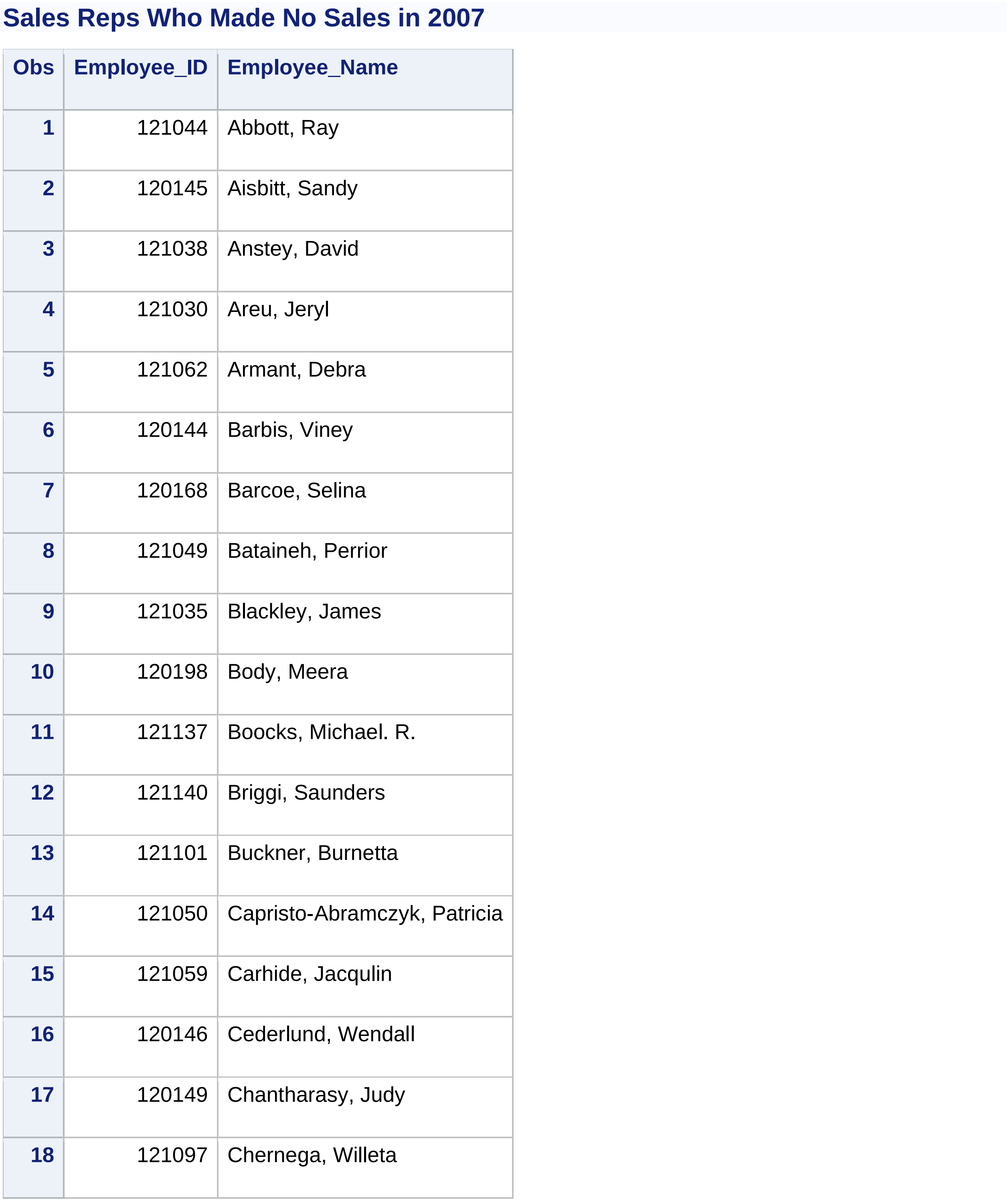
Code

```
*Problem 4;
libname orion '/courses/d649d56dba27fe300/STA5067/SAS Data/orion';
proc sql;
create table tmp4 as
select Employee_ID
from orion.sales
where scan(Job_Title,-1,' ') in ('I','II','III','IV')
     except corr
select Employee_ID
from (select *
      from orion.order_fact
      where year(Order_Date)=2007)
;
create table tmp5 as
select t.Employee_ID, a.Employee_Name
from tmp4 as t
    inner join
    orion.employee_addresses as a
    on t.Employee_ID=a.Employee_ID
;
quit;

proc print data=tmp5;
title1 "Sales Reps Who Made No Sales in 2007";
run;

*Problem 5;
libname orion '/courses/d649d56dba27fe300/STA5067/SAS Data/orion';
proc sql;
create table tmp6 as
select a.Customer_ID, Customer_Name
from (select Customer_ID
      from orion.order_fact
      intersect corr
      select Customer_ID
      from orion.customer) as a
inner join
orion.customer as c
on a.Customer_ID=c.Customer_ID
;
quit;
```

```
proc print data=tmp6 noobs;
title1 "Customers Who Placed Orders";
run;

*Problem 6;
libname orion '/courses/d649d56dba27fe300/STA5067/SAS Data/orion';
proc sql;
title1 "Payroll Report for Sales Representatives";
create table tmp10 as
select "Total Paid to ALL Female Sales Representatives" as Gender,
sum(salary) label="Total Payroll" format=dollar12., count(*)
label= "# of Employees"
from orion.sales
where Gender='F' and scan(Job_Title,-1,'') contains "Rep"
union
select "Total Paid to ALL Male Sales Representatives" as Gender,
sum(salary) label="Total Payroll" format=dollar12., count(*)
label= "# of Employees"
from orion.sales
where Gender='M' and scan(Job_Title,-1,'') contains "Rep"
;
quit;

proc print data=tmp10 noobs label;
run;
```

Selected Output

Problem 4:



Sales Reps Who Made No Sales in 2007

Obs	Employee_ID	Employee_Name
1	121044	Abbott, Ray
2	120145	Aisbitt, Sandy
3	121038	Anstey, David
4	121030	Areu, Jeryl
5	121062	Armant, Debra
6	120144	Barbis, Viney
7	120168	Barcoe, Selina
8	121049	Bataineh, Perrior
9	121035	Blackley, James
10	120198	Body, Meera
11	121137	Boocks, Michael. R.
12	121140	Briggi, Saunders
13	121101	Buckner, Burnetta
14	121050	Capristo-Abramczyk, Patricia
15	121059	Carhide, Jacqulin
16	120146	Cederlund, Wendall
17	120149	Chantharasy, Judy
18	121097	Chernega, Willeta

Problem 5:

Customers Who Placed Orders

Customer_ID	Customer_Name
4	James Kvarniq
5	Sandrina Stephano
9	Cornelia Krahl
10	Karen Ballinger
11	Elke Wallstab
12	David Black
13	Markus Sepke
16	Ulrich Heyde
17	Jimmie Evans
18	Tonie Asmussen
19	Oliver S. Füßling
20	Michael Dineley
23	Tulio Devereaux

Problem 6:

Payroll Report for Sales Representatives

Gender	Total Payroll	# of Employees
Total Paid to ALL Female Sales Representatives	$1,872,360	67
Total Paid to ALL Male Sales Representatives	$2,566,785	92

Homework 15: Macro Programs

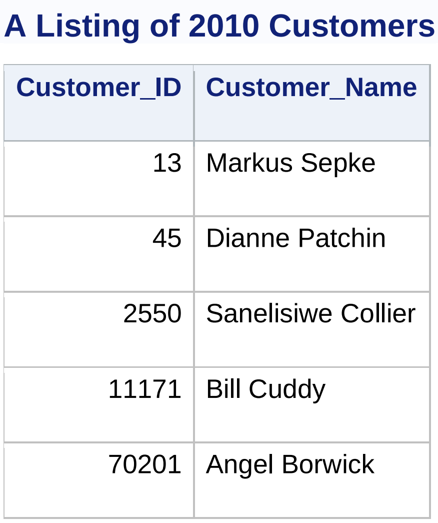
Code

```
*Problem 2;
%macro listing(custtype)/minoperator;
  proc sql noprint;
  select distinct Customer_Type_ID
     into :idlist separated by ' '
  from orion.customer_type
  ;
  quit;
  %if &custtype in &idlist . %then %do;
     %let flag = 0;
  %end;
  %else %do;
     %let flag = 1;
  %end;
  %if &flag = 0 %then %do;
       %if &custtype=. %then %do;
       proc print data=orion.customer noobs;
          var Customer_ID Customer_Name Customer_Type_ID;
          title "A Listing of All Customers";
       run;
       %end;
       %else %if &custtype in &idlist %then %do;
       proc print data=orion.customer noobs;
          where Customer_Type_ID =%eval(&custtype);
          var Customer_ID Customer_Name;
          title "A Listing of &custtype Customers";
       run;
       %end;
  %end;
  %else %if &flag = 1 %then %do;
     %put ERROR: Value for CUSTTYPE is invalid;
     %put Valid values are &idlist;
  %end;
%mend listing;

%listing(2010);

*Problem 3;
%macro generatecode(bartype=VBAR, dims=3D,
                    var=Customer_Age_Group, color=pink,
                    surface=S) / minoperator;
```

```
    %let numerrors=0;
    %if not(&bartype in VBAR HBAR) %then %do; %let
numerrors=%eval(&numerrors+1); %end;
    %if not(&dims in 3D null) %then %do; %let
numerrors=%eval(&numerrors+1); %end;
    %if not(&surface in S X1 X2 X3 X4 X5) %then %do; %let
numerrors=%eval(&numerrors+1); %end;
    %if &numerrors=0 %then %do;
       proc gchart data=orion.customer_dim;
         &bartype&dims &var;
         pattern color=&color value=&surface;
    %end;
       run;
       quit;
%mend generatecode;
%generatecode();

*Problem 5;
%macro tops(obs=3);
   proc means data=orion.order_fact sum nway noprint;
     var Total_Retail_Price;
     class Customer_ID;
     output out=customer_freq sum=sum;
   run;

   proc sort data=customer_freq;
     by descending sum;
   run;

   data _null_;
     set customer_freq(obs=&obs);
     call symputx('top'||left(_n_), Customer_ID);
   run;

   proc print data=orion.customer_dim noobs;
     where Customer_ID in (%do num=1 %to &obs; &&top&num %end;);
     var Customer_ID Customer_Name Customer_Type;
     title "Top &obs Customers";
   run;
%mend tops;

%tops()
%tops(obs=5)
```

Selected Output

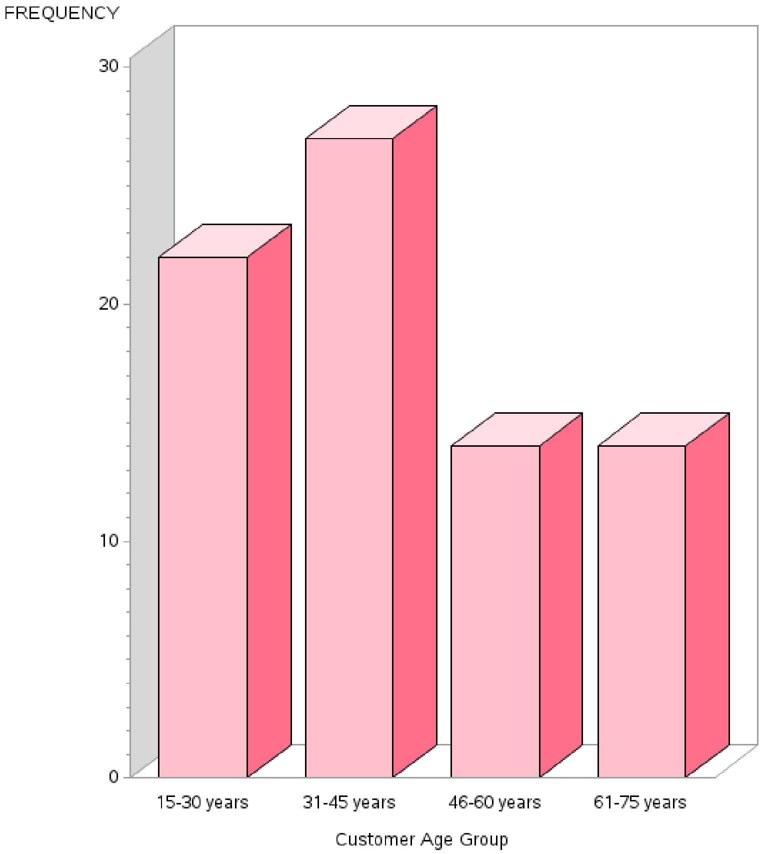
Problem 2:



A Listing of 2010 Customers

Customer_ID	Customer_Name
13	Markus Sepke
45	Dianne Patchin
2550	Sanelisiwe Collier
11171	Bill Cuddy
70201	Angel Borwick

Problem 3:



Problem 5:

Top 3 Customers

Customer_ID	Customer_Name	Customer_Type
10	Karen Ballinger	Orion Club members high activity
16	Ulrich Heyde	Internet/Catalog Customers
45	Dianne Patchin	Orion Club Gold members low activity

Top 5 Customers

Customer_ID	Customer_Name	Customer_Type
10	Karen Ballinger	Orion Club members high activity
16	Ulrich Heyde	Internet/Catalog Customers
45	Dianne Patchin	Orion Club Gold members low activity
195	Cosi Rimmington	Orion Club members low activity
2806	Raedene Van Den Berg	Orion Club members medium activity

STA 5238

APPLIED LOGISTIC REGRESSION

Homework 4: Logistic Regression Diagnostics

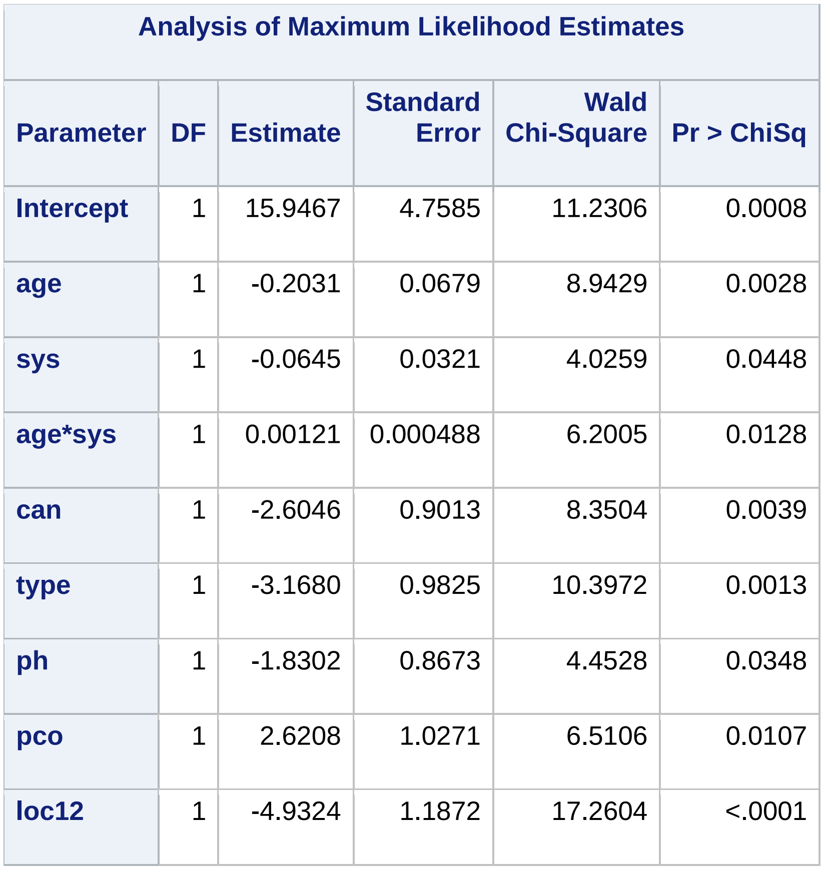
Code

```
*Load the icu data;
data icu;
infile "/home/u42193532/my_courses/huffer/5238/icu.txt";
input id sta age gender race ser can crn inf cpr sys hra pre type fra
po2 ph pco bic cre loc;
run;

*Create loc12 variable;
data icu_loc12 ;
set icu;
loc12 = 0;
if loc = 1 then loc12 = 1;
if loc = 2 then loc12 = 1;
run;

*Baseline model;
proc logistic data=icu_loc12
plots(unpack label) = (influence dfbetas phat dpc leverage);
model sta(event="1") = age sys age*sys can type ph pco loc12;
output out=add_c_difchisq c=c difchisq=difchisq;
run;

*Remove 4 largest chi-squared deletion differences;
data delete_large_difchisq;
set add_c_difchisq;
if difchisq > 15 then delete;
run;

*Model without largest chi-squared deletion differences;
proc logistic data=delete_large_difchisq;
model sta(event="1") = age sys age*sys can type ph pco loc12;
run;

*Remove 3 largest C values;
data delete_large_c;
set add_c_difchisq;
if c > 0.9 then delete;
run;
```

```
*Model without largest c values;
proc logistic data=delete_large_c;
model sta(event="1") = age sys age*sys can type ph pco loc12;
run;

*Remove largest chi-squared deletion differences and c values;
data delete_large_c_difchisq;
set add_c_difchisq;
if c > 0.9 then delete;
if difchisq > 15 then delete;
run;

*Model without largest chi-squared deletion differences and c values;
proc logistic data=delete_large_c_difchisq;
model sta(event="1") = age sys age*sys can type ph pco loc12;
run;
```

Selected Output

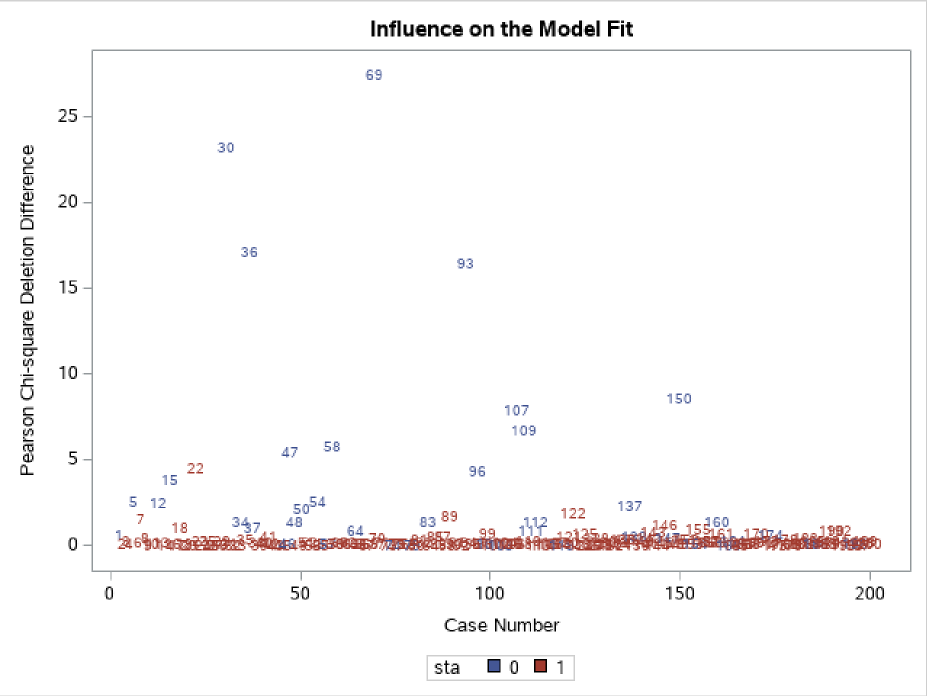
The table of maximum likelihood estimates for the specified model is shown below. Each of the coefficients is statistically significant at the 5% level. AGE, SYS, CAN, TYPE, PH, and LOC12 each have negative coefficient estimates, while AGE*SYS and PCO have positive coefficient estimates.

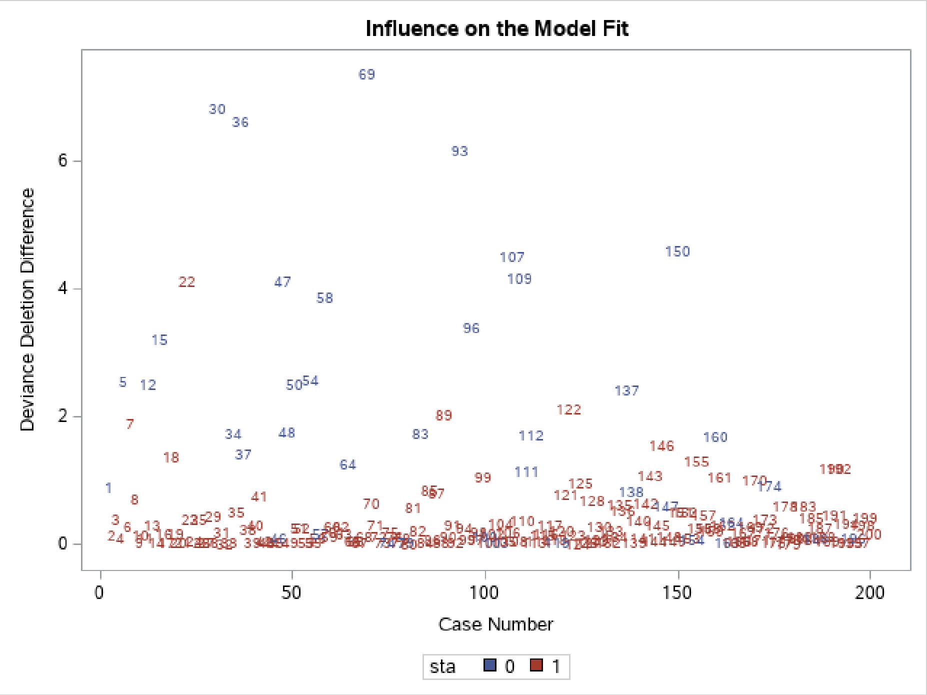


Analysis of Maximum Likelihood Estimates

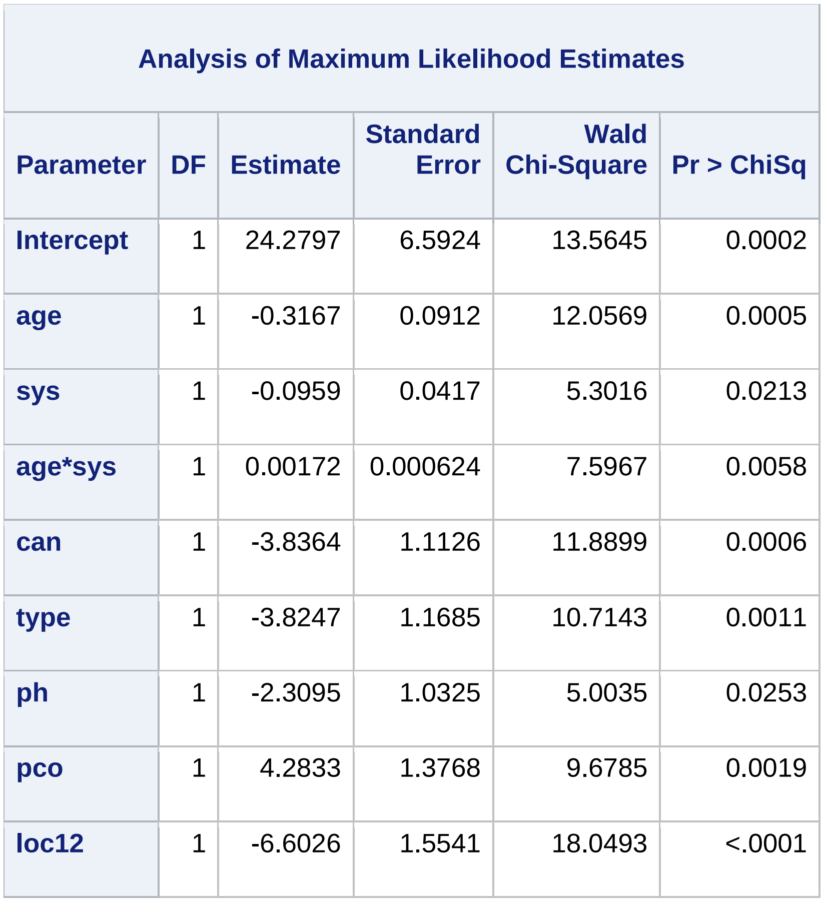
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.9467	4.7585	11.2306	0.0008
age	1	-0.2031	0.0679	8.9429	0.0028
sys	1	-0.0645	0.0321	4.0259	0.0448
age*sys	1	0.00121	0.000488	6.2005	0.0128
can	1	-2.6046	0.9013	8.3504	0.0039
type	1	-3.1680	0.9825	10.3972	0.0013
ph	1	-1.8302	0.8673	4.4528	0.0348
pco	1	2.6208	1.0271	6.5106	0.0107
loc12	1	-4.9324	1.1872	17.2604	<.0001

To identify cases that the model fits poorly, I focus on the Pearson Chi-Square Deletion Differences and Deviance Deletion Differences plots. As shown in the plots below, cases 69, 30, 36, and 93 have Pearson Chi-Square Deletion Differences greater than 15 and Deviance Deletion Differences greater than 6. These values appear to be set apart from the rest of the points in both plots.





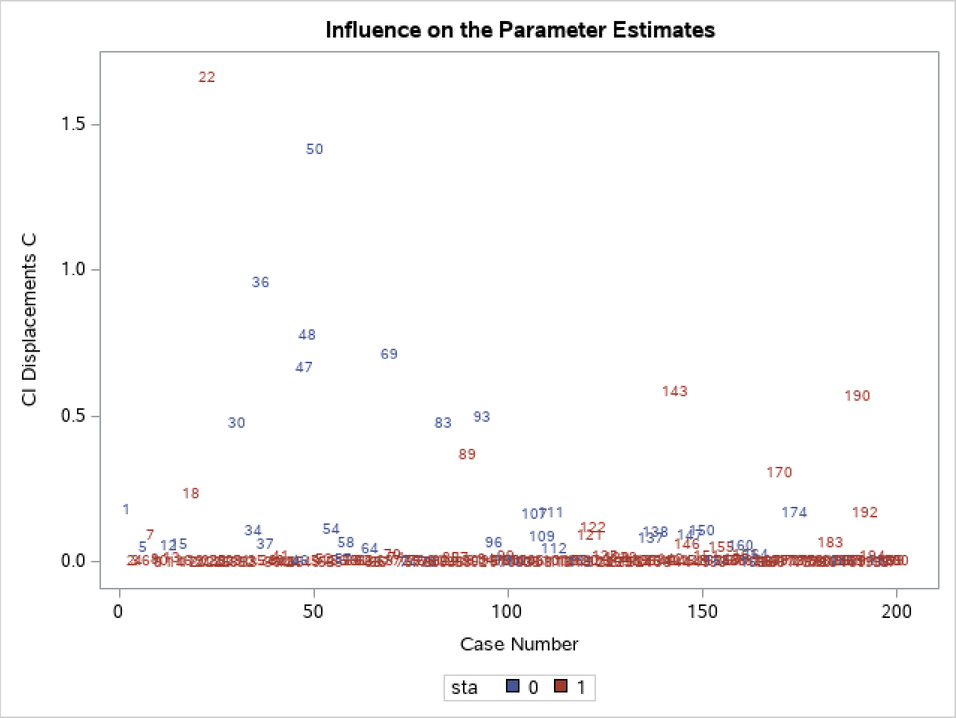
To investigate the effect these four observations have on the fitted model, I removed the cases and re-fit the initial model. The table of maximum likelihood estimates for the resulting model is shown below. All of the coefficient values are still statistically significant and have the same signs as before. Several of the coefficient estimates changed substantially: the estimate for CAN fell from -2.6046 to -3.8364, the estimate for PCO increased from 2.6208 to 4.2833, and the estimate for LOC12 decreased from -4.9324 to -6.6026.

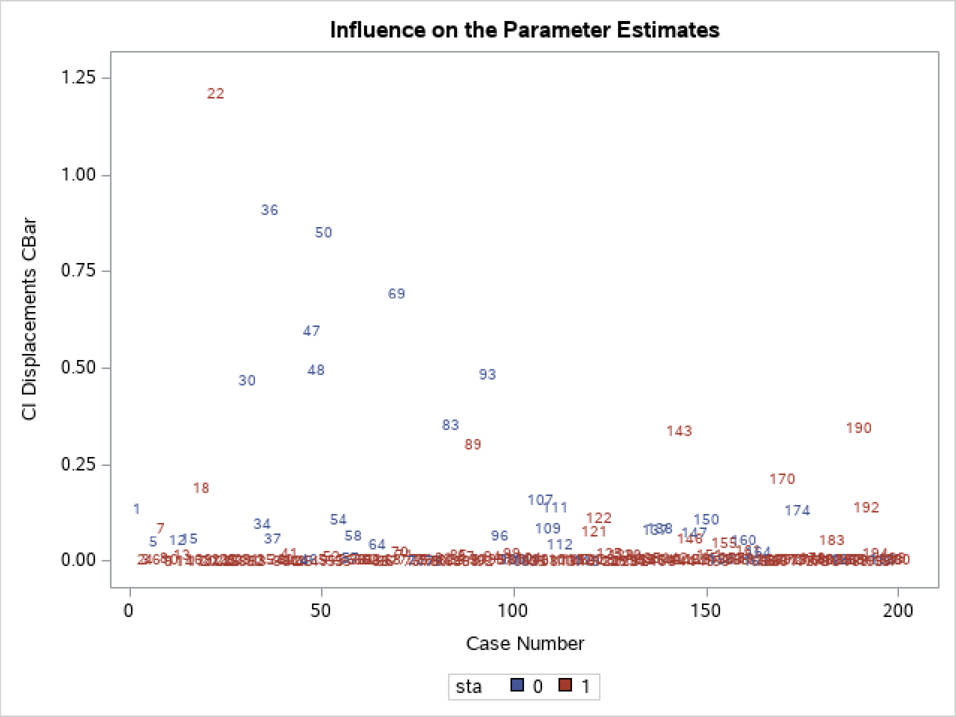


Analysis of Maximum Likelihood Estimates

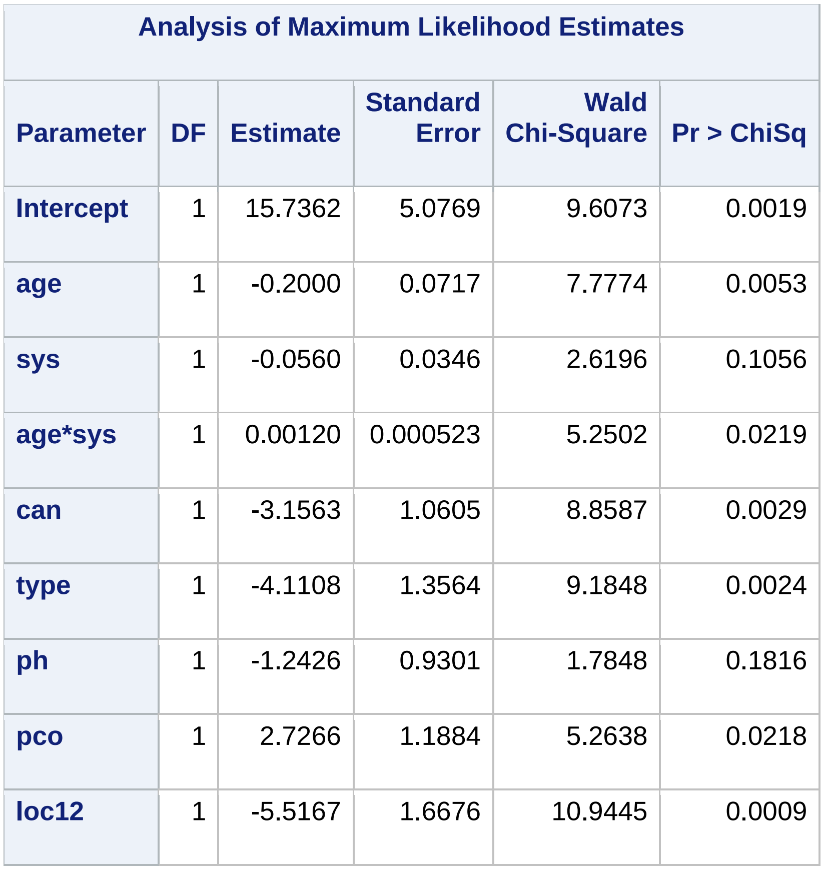
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	24.2797	6.5924	13.5645	0.0002
age	1	-0.3167	0.0912	12.0569	0.0005
sys	1	-0.0959	0.0417	5.3016	0.0213
age*sys	1	0.00172	0.000624	7.5967	0.0058
can	1	-3.8364	1.1126	11.8899	0.0006
type	1	-3.8247	1.1685	10.7143	0.0011
ph	1	-2.3095	1.0325	5.0035	0.0253
pco	1	4.2833	1.3768	9.6785	0.0019
loc12	1	-6.6026	1.5541	18.0493	<.0001

Next, I used the C and CBAR plots to identify cases with the largest influence on the parameter estimates. As shown in the two plots below, cases 22, 50, and 36 each have C values greater than 0.9 and CBAR values greater than 0.75. These points are clearly separated from the rest of the points in these two plots.





To learn more about the influence of these cases on the model, I deleted the three most influential cases and re-fit the initial model. The table of maximum likelihood estimates for the resulting model is shown below. Interestingly, the coefficient estimates for SYS and PH are no longer statistically significant. The coefficient estimate for TYPE experienced the most significant change from the initial model, falling from -3.1680 to -4.1108, while the other coefficients changed slightly. None of the coefficients switched signs after the three cases were removed.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.7362	5.0769	9.6073	0.0019
age	1	-0.2000	0.0717	7.7774	0.0053
sys	1	-0.0560	0.0346	2.6196	0.1056
age*sys	1	0.00120	0.000523	5.2502	0.0219
can	1	-3.1563	1.0605	8.8587	0.0029
type	1	-4.1108	1.3564	9.1848	0.0024
ph	1	-1.2426	0.9301	1.7848	0.1816
pco	1	2.7266	1.1884	5.2638	0.0218
loc12	1	-5.5167	1.6676	10.9445	0.0009

Lastly, I removed all of the cases identified in the two earlier steps (cases 22, 30, 36, 50, 69, and 93) at the same time. The table of maximum likelihood estimates is shown below. As in the previous model, the coefficient estimates for SYS and PH are no longer statistically significant on the 5% level. Several coefficients changed significantly from the original model: the coefficient estimate for CAN fell from -2.6046 to -4.0388, the estimate for TYPE decreased from -3.1680 to -4.4532, the estimate for PCO increased from 2.6208 to 3.4580, and the estimate for LOC12 fell from -4.9324 to -6.2880.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	21.6183	6.7857	10.1496	0.0014
age	1	-0.2771	0.0938	8.7290	0.0031
sys	1	-0.0747	0.0451	2.7412	0.0978
age*sys	1	0.00147	0.000659	4.9991	0.0254
can	1	-4.0388	1.1877	11.5643	0.0007
type	1	-4.4532	1.4480	9.4579	0.0021
ph	1	-1.9627	1.0522	3.4795	0.0621
pco	1	3.4580	1.3626	6.4407	0.0112
loc12	1	-6.2880	1.7347	13.1389	0.0003

STA 5939

INTRODUCTION TO STATISTICAL CONSULTING

Final Project Report: Economic Growth & Subjective Well-Being

Summary

In this project, I used lasso regression and backward stepwise variable elimination to model country “happiness scores” reported by the United Nations using economic development indicators. I began with a dataset of 88 indicators and ran lasso regression to determine the most important predictors. I then used a power transformation on the GNI per capita variable to ensure a linear relationship with the happiness score variable. Lastly, I used backward stepwise variable elimination to choose the final model using the variables previously specified by the lasso model.

Code

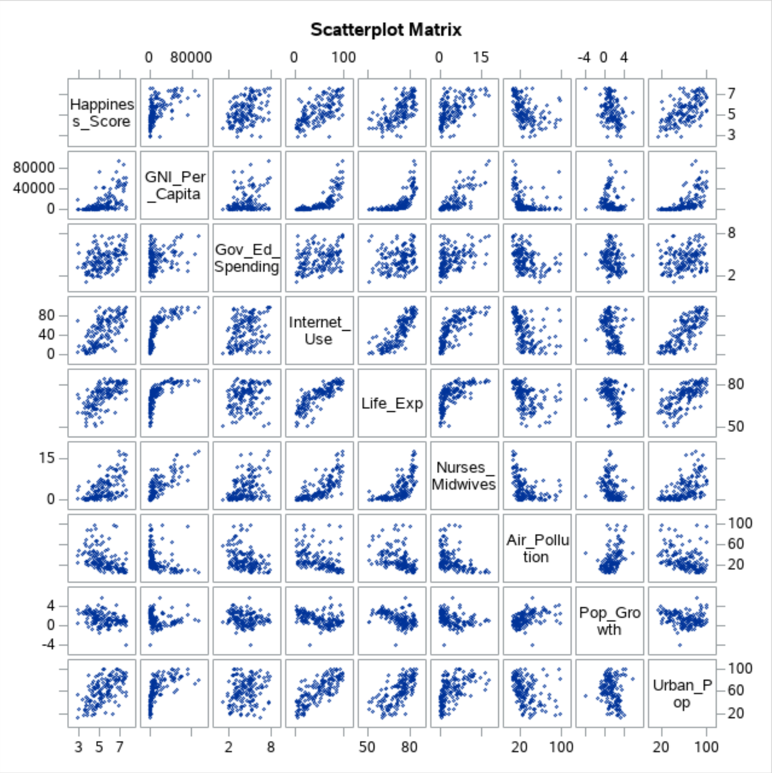
```
*Load the sub_data

*Ran the lasso regression model in R, left with these candidate
variables for further analysis in SAS.;

%web_drop_table(WORK.sub_data);
FILENAME REFFILE '/folders/myfolders/sasuser.v94/sub_data.csv';
PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=WORK.sub_data;
     GETNAMES=YES;
RUN;
%web_open_table(WORK.sub_data);

*Rename variables;
data work.sub_data(drop=Var1);
set work.sub_data;
rename GNI_per_capita__Atlas_method__c=GNI_Per_Capita
       Government_expenditure_on_educa=Gov_Ed_Spending
       Individuals_using_the_Internet=Internet_Use
       Life_expectancy_at_birth__total=Life_Exp
       Nurses_and_midwives__per_1_000=Nurses_Midwives
       PM2_5_air_pollution__mean_annua=Air_Pollution
       Population_growth__annual___=Pop_Growth
       Urban_population____of_total_po=Urban_Pop
;
```

```
run;

*Add dummy variable for income level (high/low) - Using WB category;
data work.sub_data;
set work.sub_data;
if GNI_Per_Capita > 12475 then High_Income = 1;
else High_Income=0;
run;

*Create variable scatterplot for the data;
proc sgscatter data=work.sub_data;
  title "Scatterplot Matrix";
  matrix Happiness_Score GNI_Per_Capita Gov_Ed_Spending Internet_Use
  Life_Exp Nurses_Midwives Air_Pollution Pop_Growth Urban_Pop
;
RUN;
title;

*Most of these variables seem to have a linear relationship with the
happiness score,
but GNI per capita seems to be an issue. Let's plot GNI vs. happiness
score to get a closer
look;
proc sgscatter data=work.sub_data;
  title "Scatterplot Matrix";
  matrix Happiness_Score GNI_Per_Capita
;
RUN;
title;

*Sqrt is the best transformation for GNI, so we can use that.

*Add sqrt(GNI_PC) to the data set, as well as interaction term;
DATA work.sub_data;
    SET work.sub_data;
    sqrt_GNI_PC = (GNI_Per_Capita)**(1/2);
    High_INC_GNI_PC=High_Income*sqrt_GNI_PC;
    High_INC_Pop_Growth=High_Income*Pop_Growth;
RUN;

*Plot the new variable vs. happiness score;
proc sgscatter data=work.sub_data;
```

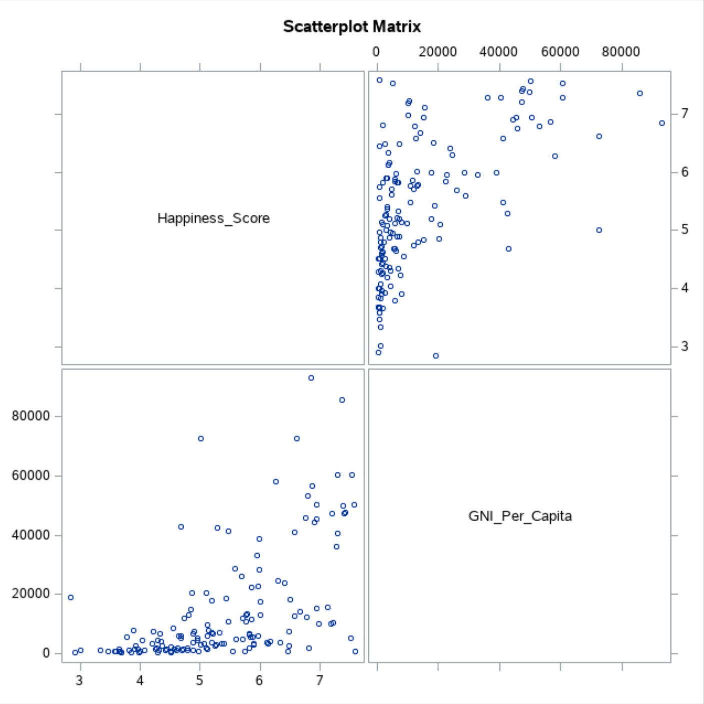
```
  title "Scatterplot Matrix";
  matrix Happiness_Score sqrt_GNI_PC
;
RUN;
title;

*Run the initial model selected by the lasso regression in R, plus
interaction term;
proc reg data=work.sub_data;
model Happiness_Score = sqrt_GNI_PC High_INC_GNI_PC Gov_Ed_Spending
Internet_Use Life_Exp
Nurses_Midwives Air_Pollution Pop_Growth Urban_Pop High_INC_Pop_Growth
/ vif
;
run;

*Some of the variances are inflated, seems like there is still
multicollinearity. Let's try stepwise selection to see if we can
reduce the number of variables without reducing the amount of
information produced by the model.;
proc reg data=work.sub_data;
model Happiness_Score = sqrt_GNI_PC High_INC_GNI_PC Gov_Ed_Spending
Internet_Use Life_Exp
Nurses_Midwives Air_Pollution Pop_Growth Urban_Pop High_INC_Pop_Growth
/ influence vif slstay=0.15 slentry=0.15
selection=backward ss2 sse aic;
  ;
run;
```

Selected Output

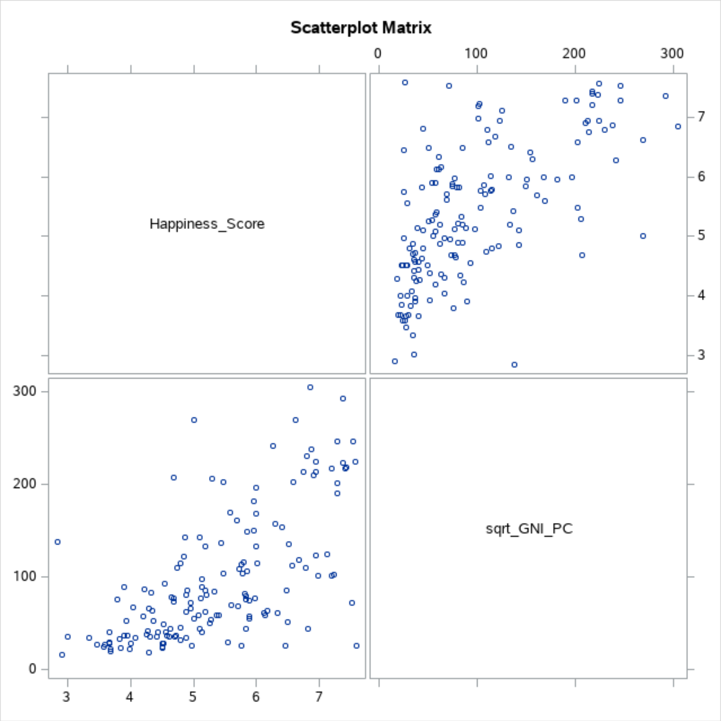
Initial variable scatterplot matrix:



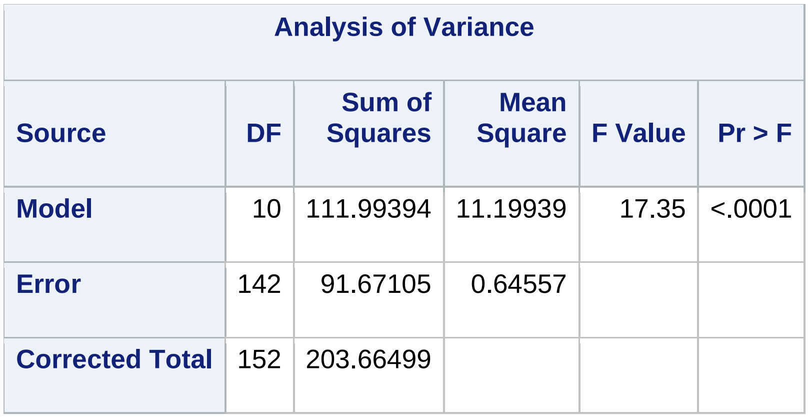
Scatterplot matrix for GNI per capita before transformation:



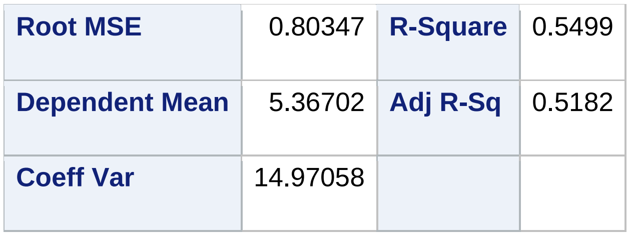
Scatterplot matrix for GNI per capita after transformation:



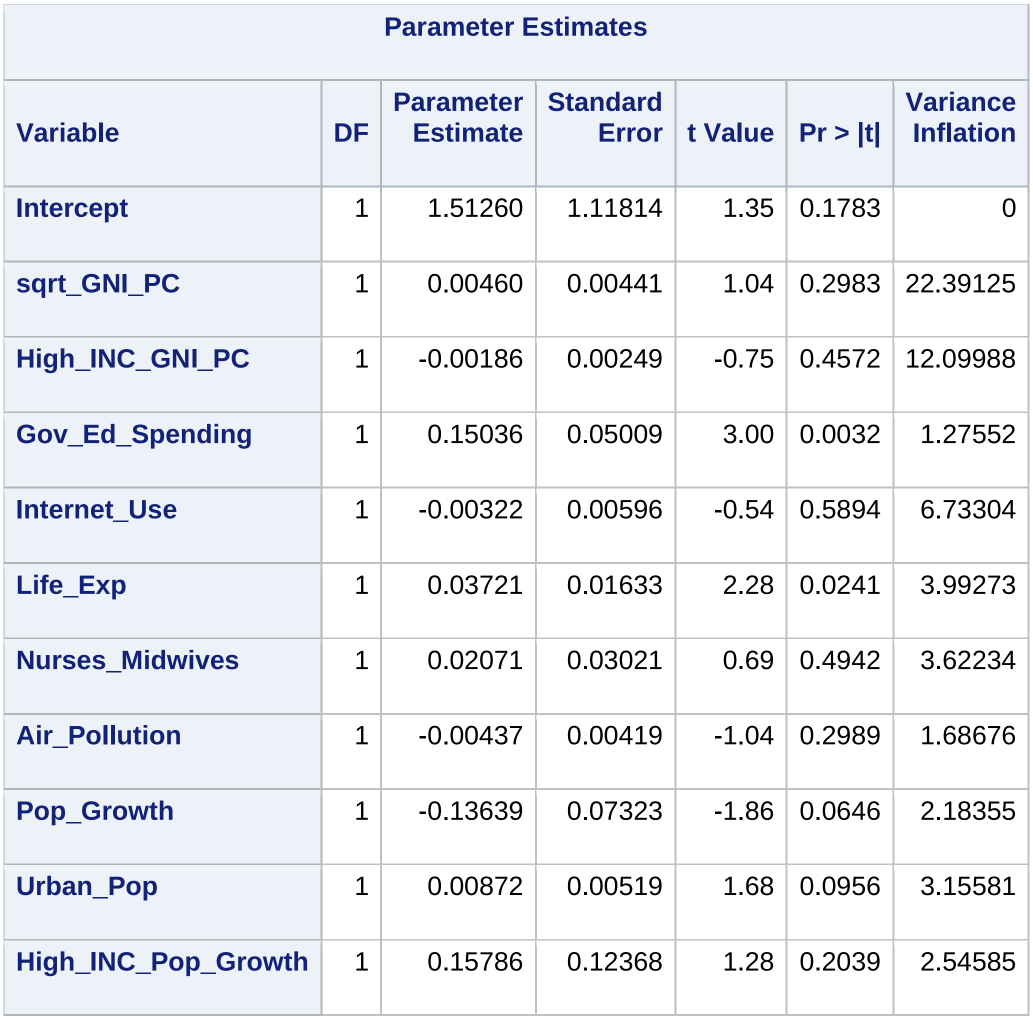
Initial model specified by lasso regression:



Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	10	111.99394	11.19939	17.35	<.0001
Error	142	91.67105	0.64557		
Corrected Total	152	203.66499			

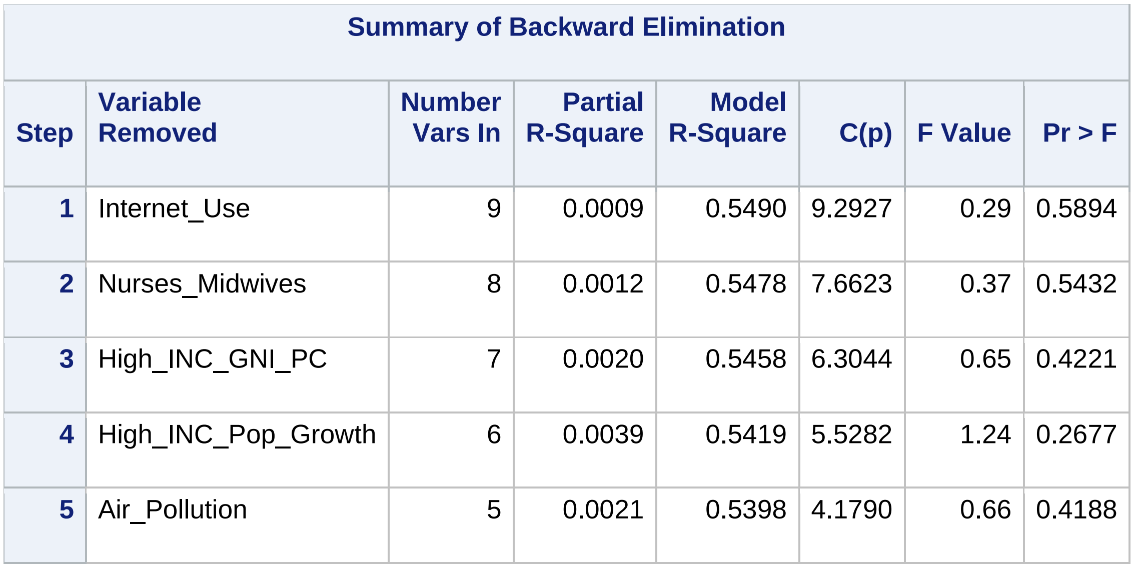


Root MSE	0.80347	**R-Square**	0.5499
Dependent Mean	5.36702	**Adj R-Sq**	0.5182
Coeff Var	14.97058		



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.51260	1.11814	1.35	0.1783	0
sqrt_GNI_PC	1	0.00460	0.00441	1.04	0.2983	22.39125
High_INC_GNI_PC	1	-0.00186	0.00249	-0.75	0.4572	12.09988
Gov_Ed_Spending	1	0.15036	0.05009	3.00	0.0032	1.27552
Internet_Use	1	-0.00322	0.00596	-0.54	0.5894	6.73304
Life_Exp	1	0.03721	0.01633	2.28	0.0241	3.99273
Nurses_Midwives	1	0.02071	0.03021	0.69	0.4942	3.62234
Air_Pollution	1	-0.00437	0.00419	-1.04	0.2989	1.68676
Pop_Growth	1	-0.13639	0.07323	-1.86	0.0646	2.18355
Urban_Pop	1	0.00872	0.00519	1.68	0.0956	3.15581
High_INC_Pop_Growth	1	0.15786	0.12368	1.28	0.2039	2.54585

Variables eliminated from the initial model by backward stepwise elimination:



Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Internet_Use	9	0.0009	0.5490	9.2927	0.29	0.5894
2	Nurses_Midwives	8	0.0012	0.5478	7.6623	0.37	0.5432
3	High_INC_GNI_PC	7	0.0020	0.5458	6.3044	0.65	0.4221
4	High_INC_Pop_Growth	6	0.0039	0.5419	5.5282	1.24	0.2677
5	Air_Pollution	5	0.0021	0.5398	4.1790	0.66	0.4188

Final model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	109.94169	21.98834	34.49	<.0001
Error	147	93.72330	0.63757		
Corrected Total	152	203.66499			

Root MSE	0.79848	R-Square	0.5398
Dependent Mean	5.36702	Adj R-Sq	0.5242
Coeff Var	14.87757		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS	Variance Inflation
Intercept	1	1.01008	0.98339	1.03	0.3060	0.67266	0
sqrt_GNI_PC	1	0.00334	0.00166	2.01	0.0461	2.57860	3.21942
Gov_Ed_Spending	1	0.16770	0.04625	3.63	0.0004	8.38273	1.10096
Life_Exp	1	0.04018	0.01482	2.71	0.0075	4.68820	3.33095
Pop_Growth	1	-0.10567	0.05674	-1.86	0.0645	2.21154	1.32704
Urban_Pop	1	0.00916	0.00451	2.03	0.0438	2.63730	2.40374

END