~FINAL SAS CODE~

```
/*
'air_qualityants Are U.S.'
DSC 323 Final Project Notebook
Areeb Abubaker, Jordan Bickelhaupt, Julio Delgado, Jaime Moscoso,
Sasha Rukhina
Instructor: Nandhini Gulasingam
16 Nov '20
*/

*Import Data;
PROC IMPORT DATAFILE="pollution_us_2000_2016.csv" OUT = air_quality
replace;
DELIMITER = ',';
GETNAMES = yes;
RUN;
PROC PRINT DATA = air_quality(OBS = 5);
RUN;

*Create New Variable Overall AQI;
TITLE 'New variable = Overall AQI';
DATA air_quality;
SET air_quality;
overall_AQI = mean(CO_AQI, NO2_AQI, SO2_AQI);
log_AQI = log(overall_AQI);
RUN;

*Analyze New Variable Distribution Values;
TITLE 'Check Overall AQI Distribution';
PROC FREQ DATA = pollut;
TABLES overall_AQI log_AQI;
RUN;

*Visualize Distributions of Overall_AQI & log_AQI;
PROC UNIVARIATE NORMAL;
VAR overall_AQI log_AQI;
HISTOGRAM/NORMAL (MU=est SIGMA=est);
RUN;
*Scatterplot with other Vars;
PROC sgscatter DATA=pollut;
plot CO_Mean * log_AQI;
Run;
PROC sgscatter DATA=pollut;
plot NO2_Mean * log_AQI;
```

```
run;

*Boxplot for log_AQI by State;
PROC SORT;
BY State;
RUN;
TITLE "Boxplot for log_AQI by State";
PROC boxplot data = air_quality;
PLOT (log_AQI) * State;
insetgroup mean min max n
q1 q2 q3 range stddev;
RUN;


*Identify and Eliminate Collinearity;
*Result: All AQI and 1st_Max_Value Variables Removed;
PROC REG DATA = air_quality;
TITLE "FULL MODEL";
MODEL log_AQI = State_Code County_Code Site_Num Date_Local
NO2_1st_Max_Value NO2_Mean
NO2_1st_Max_Hour NO2_AQI O3_Mean O3_1st_Max_Value O3_1st_Max_Hour
O3_AQI SO2_Mean
SO2_1st_Max_Value SO2_1st_Max_Hour SO2_AQI CO_Mean CO_1st_Max_Value
CO_1st_Max_Hour CO_AQI /VIF RSQUARE;
RUN;

*Verify Result of Removing Variables;
PROC REG DATA = air_quality;
TITLE "Updated Full Model";
MODEL log_AQI = State_Code County_Code Site_Num Date_Local NO2_Mean
O3_Mean O3_1st_Max_Hour SO2_Mean
SO2_1st_Max_Hour CO_Mean  CO_1st_Max_Hour /VIF RSQUARE;
RUN;

*Modify Data and Overwrite air_quality;
TITLE 'Collinear and Descriptive Variables Removed';
DATA air_quality;
SET air_quality;
DROP VAR1 Address State County City Units_NO2 Units_O3 Units_SO2
Units_O3 NO2_1st_Max_Value NO2_AQI O3_1st_Max_Value O3_AQI
SO2_1st_Max_Value SO2_AQI CO_1st_Max_Value CO_AQI;
RUN;
```

```
*Discover Critical Correlations with Relevant Variables;
PROC CORR DATA = air_quality plots = matrix(histogram);
RUN;


*Updated Full Model;
PROC REG;
TITLE "Updated Full Model";
MODEL log_AQI = State_Code County_Code Site_Num Date_Local NO2_Mean
NO2_1st_Max_Hour O3_Mean O3_1st_Max_Hour SO2_Mean SO2_1st_Max_Hour
CO_Mean CO_1st_Max_Hour overall_AQI over_manual;
RUN;






*Final Model;
PROC REG DATA = air_quality;
TITLE "Fitted Model";
MODEL log_AQI = State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean /
SELECTION= ADJRSQ VIF TOL RSQUARE;
/*WARNING: LONG TIME TO PROCESS RESIDUAL AND CDF:
plot student.*(State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean
predicted.);
plot npp.*student.;
*/
RUN;

Cross Validation
Title "5-fold crossvalidation for Model 1";
proc glmselect data=air_quality
plots=(asePlot Criteria);
partition fraction(test=.25);
model log_AQI = State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean/
selection = stepwise(stop=cv) cvMethod=split(5) cvDetails=all;
run;

Title "5-fold crossvalidation for Model 2";
proc glmselect data=air_quality
plots=(asePlot Criteria);
partition fraction(test=.25);
```

```
model log_AQI = Date_Local State_Code NO2_Mean O3_Mean SO2_Mean
CO_Mean/ selection = backward(stop=cv) cvMethod=split(5)
cvDetails=all;
run;




*Final Model Prediction Process;

DATA new_poll;
INPUT State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean;
DATALINES;
2 11 16 21 26
;
PROC PRINT DATA= new_poll (OBS = 1);
RUN;

DATA prediction;
SET new_poll air_quality;
RUN;
PROC PRINT DATA = prediction (OBS = 5);
RUN;

PROC REG data = prediction;
TITLE "Fitted Model";
MODEL log_AQI(OBS = ‘1’) = State_Code NO2_Mean O3_Mean SO2_Mean
CO_Mean;
OUTPUT OUT = prediction P = phat /*lower = lcl upper = ucl*/;
RUN;
PROC PRINT DATA = prediction (OBS = 5);
RUN;


*Create New Variable Overall AQI;
TITLE 'New variable = Overall AQI';
DATA pollut;
SET pollut;
inv_log = 10**log_AQI;
RUN;
```

~END OF SAS CODE~

```
* sig var;
proc reg data=air_quality;
model log_AQI = State_Code County_Code Site_Num Date_Local NO2_Mean
O3_Mean O3_1st_Max_Hour SO2_Mean
SO2_1st_Max_Hour CO_Mean  CO_1st_Max_Hour/influence r;
plot student.*(State_Code County_Code Site_Num Date_Local NO2_Mean
O3_Mean O3_1st_Max_Hour SO2_Mean
SO2_1st_Max_Hour CO_Mean  CO_1st_Max_Hour pred.);
plot npp.*student.;
run;

title "Remove Influencial Points and Outliers";
data new_poll;
set air_quality;
if _n_ = -- then delete;
run;

proc print;
run;
```

*TO DO

***Plot residual and remove relevant outliers to fix low_AQI skew

Finetune any visualizations, analysis, modeling

Resolve VAR1 - variable created because dataset (open in notepad) begins with ‘,’?

***Review ‘Modify Data’ highlighted step

Complete Presentation & Recordings

*CURRENT CODE ISSUES

```
*Note: overall_AQI mean() defaults to skip missing values;
* e.g. CO_AQI entry no.1 = NA,;
* output: overall_AQI = 29.5 & mean_manual = NA;
```

```
////////////
```

The error is the result of having the incorrect PLOTS= value specified in pam_vcapca_runPrincipalComponentAnalysis.sas and then running an analysis.

```
*Visualize Correlations ERROR: Java virtual machine exception.
java.lang.OutOfMemoryError: GC overhead limit exceeded
;
proc sgscatter;
Matrix log_AQI NO2_Mean CO_Mean;
Run;
////////////
```

```
////////////
*ERROR: The number of panels needed is 5850 which exceeds the maximum
of 20.
PROC BOXPLOT;
plot log_AQI*Date_Local;
RUN;
//////////////
```

```
//////////////
*NOTE: Invalid argument to function LOG(0) at line 953 column 11
//////////////
```

```
*Correlation matrix reference
proc corr data=sashelp.iris plots=matrix(histogram);
Run;
```

```
//////////////
*G-Plots irrelevant and provide no analysis;
proc gplot;
plot log_AQI*(NO2_Mean CO_Mean);
run;
```

```
//////////////
```

MISC

```
*HEAD DATA;
proc print data=air_quality(obs=10);
Run;

*Fitted Model for Selection Method;
PROC REG data = air_quality;
TITLE "Fitted Model";
Model  AQI2 = State_Code
NO2_Mean
  /* NO2_AQI  */       O3_Mean
       /*O3_AQI*/ SO2_Mean
      /* SO2_AQI    */
CO_Mean      /* CO_AQI*/ / selection = stepwise VIF TOL rsquare ;
RUN;
----------------------------------------
“ADJ RSQ”
PROC REG data = air_quality;
TITLE "Fitted Model";
Model  AQI2 = State_Code
NO2_Mean
O3_Mean
SO2_Mean
CO_Mean / selection = ADJRSQ VIF TOL rsquare ;
RUN;
```

```
PROC IMPORT datafile = 'air_qualityion_us_2000_2016.csv' out = air_qualityion replace;
delimiter = ',';
getnames = yes;
datarow=2;
RUN;
```

```
proc print data=air_qualityion (obs=10);
Run;

title 'OVerall AQI';
data air_quality;
   set air_qualityion;
   *overall_AQI=mean(CO_AQI, NO2_AQI, SO2_AQI, O3_AQI);
   AQI2 = (CO_AQI + NO2_AQI + SO2_AQI + O3_AQI) / 4;
   log_AQI = log(AQI2);
   run;

title 'check AQI';
proc freq data = air_quality;
tables log_AQI;
run;

PROC UNIVARIATE normal;
var log_AQI;
histogram/normal (mu=est sigma=est);
run;


PROC PRINT data = air_quality ( obs=20);
RUN;

PROC REG data = air_quality;
TITLE "FULL MODEL";
Model  AQI2 = State_Code
County_Code
Site_Num
Date_Local
NO2_1st_Max_Value
NO2_Mean
NO2_1st_Max_Hour    NO2_AQI           O3_Mean
O3_1st_Max_Value    O3_1st_Max_Hour    O3_AQI SO2_Mean
SO2_1st_Max_Value    SO2_1st_Max_Hour   /* SO2_AQI    */
CO_Mean    CO_1st_Max_Value    CO_1st_Max_Hour    CO_AQI /VIF rsquare ;
RUN;


PROC REG data = air_quality;
```

```
TITLE "Full Model with Selection Method with Values deleted";
Model  AQI2 = State_Code
County_Code
Site_Num
Date_Local
NO2_Mean
NO2_1st_Max_Hour   /* NO2_AQI  */    O3_Mean
   O3_1st_Max_Hour    /*O3_AQI*/ SO2_Mean
   SO2_1st_Max_Hour   /* SO2_AQI    */
CO_Mean     CO_1st_Max_Hour   /* CO_AQI*/ / selection = stepwise VIF TOL rsquare ;

RUN;

PROC REG data = air_quality;
TITLE "Full Model with both Max hours and Max values deleted";
Model  AQI2 = State_Code
County_Code
Site_Num
Date_Local
NO2_Mean
  /* NO2_AQI  */            O3_Mean
      /*O3_AQI*/ SO2_Mean
     /* SO2_AQI    */
CO_Mean      /* CO_AQI*/ / selection = stepwise VIF TOL rsquare ;

RUN;


PROC REG data = air_quality;
TITLE "Fitted Model";
Model  AQI2 = State_Code
NO2_Mean
  /* NO2_AQI  */            O3_Mean
      /*O3_AQI*/ SO2_Mean
     /* SO2_AQI    */
CO_Mean      /* CO_AQI*/ / VIF TOL rsquare ;
plot student.*(State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean predicted.);
plot npp.*student.;

RUN;

PROC REG data = air_quality;
TITLE " Full Model with ADJ R2";
Model  AQI2 = State_Code    County_Code    Site_Num Date_Local    NO2_1st_Max_Value
```

```
NO2_Mean
NO2_1st_Max_Hour   /* NO2_AQI */    O3_Mean    O3_1st_Max_Value
O3_1st_Max_Hour   /* O3_AQI*/  SO2_Mean    SO2_1st_Max_Value
SO2_1st_Max_Hour   /* SO2_AQI */   CO_Mean     CO_1st_Max_Hour  /* CO_AQI*/ / selection =
ADJRSQ VIF TOL rsquare;

RUN;


*With forward;
PROC REG data = air_quality;
TITLE " Full Model with ADJ R2";
Model  AQI2 = State_Code    County_Code    Site_Num Date_Local    NO2_1st_Max_Value
NO2_Mean
NO2_1st_Max_Hour   /* NO2_AQI */    O3_Mean    O3_1st_Max_Value
O3_1st_Max_Hour   /* O3_AQI*/  SO2_Mean    SO2_1st_Max_Value
SO2_1st_Max_Hour   /* SO2_AQI */   CO_Mean     CO_1st_Max_Hour  /* CO_AQI*/ / selection =
forward VIF TOL rsquare;

RUN;

*Fitted Model and MC Problems;
PROC REG;
Model =

/*title 'new variable';
data air_quality;
   set air_quality;
   overall_AQI=mean(CO_AQI, NO2_AQI, SO2_AQI);
run;*/





/*

LOG TRANSFOREMD Y




*/
```

```
PROC REG data = air_quality;
TITLE "Fitted Model";
Model  AQI2 = State_Code
NO2_Mean
  /* NO2_AQI  */            O3_Mean
      /*O3_AQI*/ SO2_Mean
     /* SO2_AQI    */
CO_Mean      /* CO_AQI*/ / VIF TOL rsquare ;
plot student.*(State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean predicted.);
plot npp.*student.;

RUN;

*Transformed;
PROC REG data = air_quality;
TITLE "Fitted Model";
Model  log_AQI = State_Code
NO2_Mean
  /* NO2_AQI  */            O3_Mean
      /*O3_AQI*/ SO2_Mean
     /* SO2_AQI    */
CO_Mean      /* CO_AQI*/ / VIF TOL rsquare ;
plot student.*(State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean predicted.);
plot npp.*student.;

RUN;


PROC PRINT;
RUN;
-----------------------------------------------------------------
-------------------------


PROC IMPORT DATAFILE="pollution_us_2000_2016.csv" OUT = pollut replace;
DELIMITER = ',';
GETNAMES = yes;
RUN;
PROC PRINT DATA = pollut(OBS = 5);
RUN;

*Create New Variable Overall AQI;
TITLE 'New variable = Overall AQI';
DATA pollut;
```

```
SET pollut;
overall_AQI = mean(CO_AQI, NO2_AQI, SO2_AQI);
log_AQI = log(overall_AQI);
RUN;
PROC REG DATA = pollut;
TITLE "Fitted Model";
MODEL log_AQI = State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean / SELECTION = ADJRSQ
VIF TOL RSQUARE;
plot student.(State_Code NO2_Mean O3_Mean SO2_Mean CO_Mean predicted.);
plot npp.student.;
RUN;

PROC REG data = pollut;
TITLE " Full Model with ADJ R2";
Model  AQI2 = State_Code        County_Code    Site_Num Date_Local    NO2_1st_Max_Value
NO2_Mean
NO2_1st_Max_Hour   /* NO2_AQI */     O3_Mean        O3_1st_Max_Value
O3_1st_Max_Hour   /* O3_AQI*/  SO2_Mean   SO2_1st_Max_Value
SO2_1st_Max_Hour   /* SO2_AQI */   CO_Mean          CO_1st_Max_Hour  /* CO_AQI*/ / selection =
ADJRSQ VIF TOL rsquare;

RUN;


PROC REG data = pollut;
TITLE " Full Model with ADJ R2";
Model  AQI2 = State_Code        County_Code    Site_Num Date_Local    NO2_1st_Max_Value
NO2_Mean
NO2_1st_Max_Hour   /* NO2_AQI */     O3_Mean        O3_1st_Max_Value
O3_1st_Max_Hour   /* O3_AQI*/  SO2_Mean   SO2_1st_Max_Value
SO2_1st_Max_Hour   /* SO2_AQI */   CO_Mean          CO_1st_Max_Hour  /* CO_AQI*/
/ selection = ADJRSQ VIF TOL rsquare;

RUN;


*With forward;
PROC REG data = pollut;
TITLE " Full Model with Forward";
Model  AQI2 = State_Code        County_Code    Site_Num Date_Local    NO2_1st_Max_Value
NO2_Mean
NO2_1st_Max_Hour   /* NO2_AQI */     O3_Mean        O3_1st_Max_Value
O3_1st_Max_Hour   /* O3_AQI*/  SO2_Mean   SO2_1st_Max_Value
SO2_1st_Max_Hour   /* SO2_AQI */   CO_Mean          CO_1st_Max_Hour  /* CO_AQI*/ /
```

```
selection = forward VIF TOL rsquare;

RUN;
```