Stats2\_Proj

Pankaj Kumar , Karl Jurek, Shawn Jung

6/15/2019

Table of Contents

[Introduction 1](#_Toc11539441)

[Data Description 1](#_Toc11539442)

[Data Cleaning / Wrangling 2](#_Toc11539443)

[Build\_year 2](#_Toc11539444)

[Kitch\_sq 2](#_Toc11539445)

[State 2](#_Toc11539446)

[EDA 2](#_Toc11539447)

[Outliers: 2](#_Toc11539448)

[Multicollinearity: 2](#_Toc11539449)

[Checking Assumptions: 3](#_Toc11539450)

[Variable Selection: 4](#_Toc11539451)

[Modeling 4](#_Toc11539452)

[Final Prediction 4](#_Toc11539453)

[Appendix 1 5](#_Toc11539454)

[Appendix 2 5](#_Toc11539455)

[SAS code to check Assumptions: 5](#_Toc11539456)

[SAS Code to check Multi collinearity: 6](#_Toc11539457)

[SAS Code for Data Wrangling : 6](#_Toc11539458)

## 

## Introduction

Using the Modeling.csv file we created the following models to predict the cost of housing in the predictionData.csv file.

## Data Description

1. The modelingData.csv file contains properties sold between August 20, 2011 and June 30, 2015.
2. 25471 observations / rows
3. 292 variables / columns
   * + - 1. The predictionData.csv file contains properties sold between July 2015 and May 2016.
         2. 5000 observations / rows
         3. 291 variables / columns

## Data Cleaning / Wrangling

### Build\_year

### Kitch\_sq

### State

## EDA

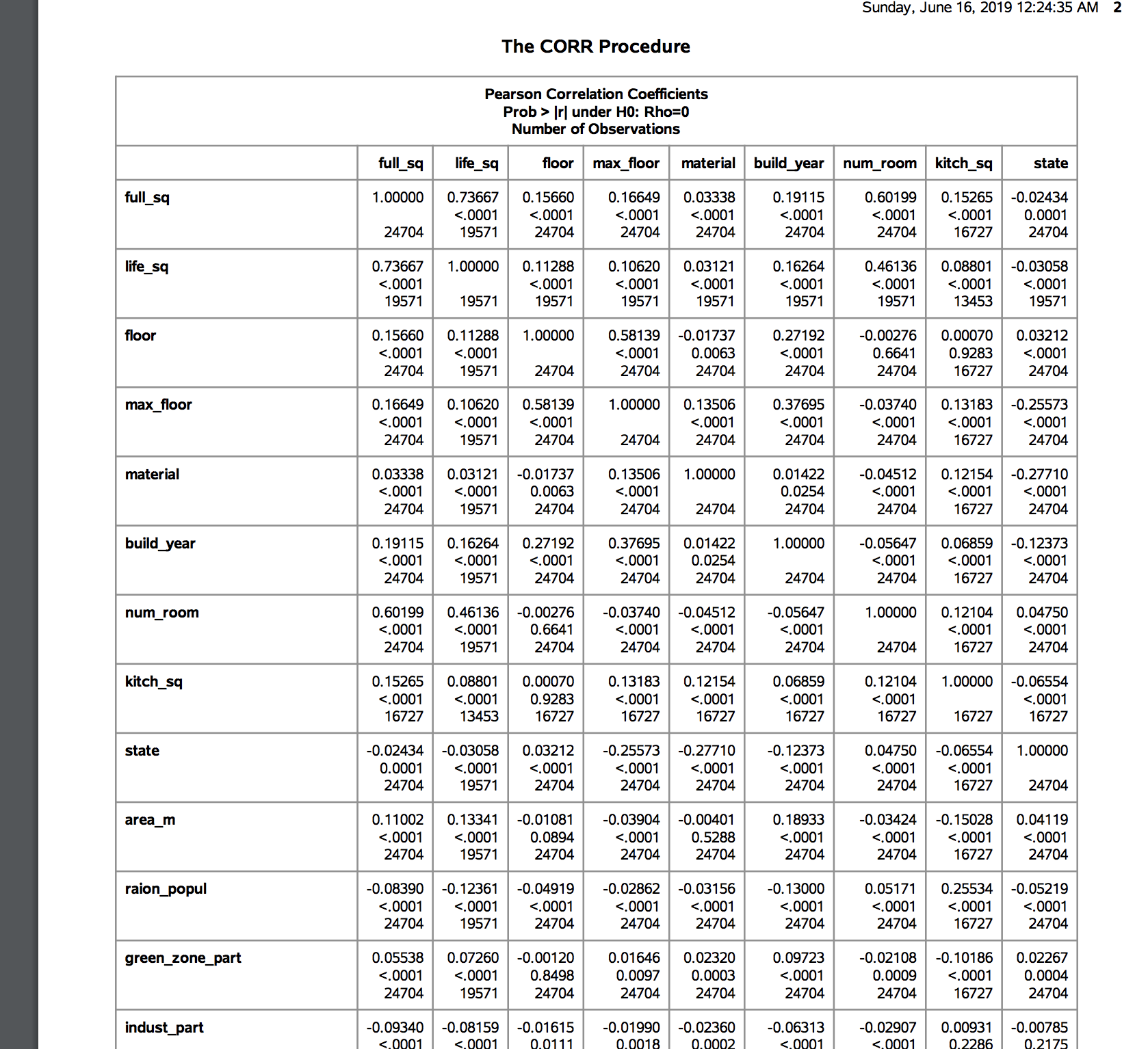
### Outliers:

SAS CODE

### Multicollinearity:

We can see some of the independent variables are highly correlated to each other and they might impact coefficients for linear regressions.

Check appendix 2 for code :



As we can see in above output correlation between life\_sq and full\_sq and num\_room and full\_sq are sufficiently high to produce effects due to multi collinearity.

### Checking Assumptions:

We didn’t find any variable that needed transformation.

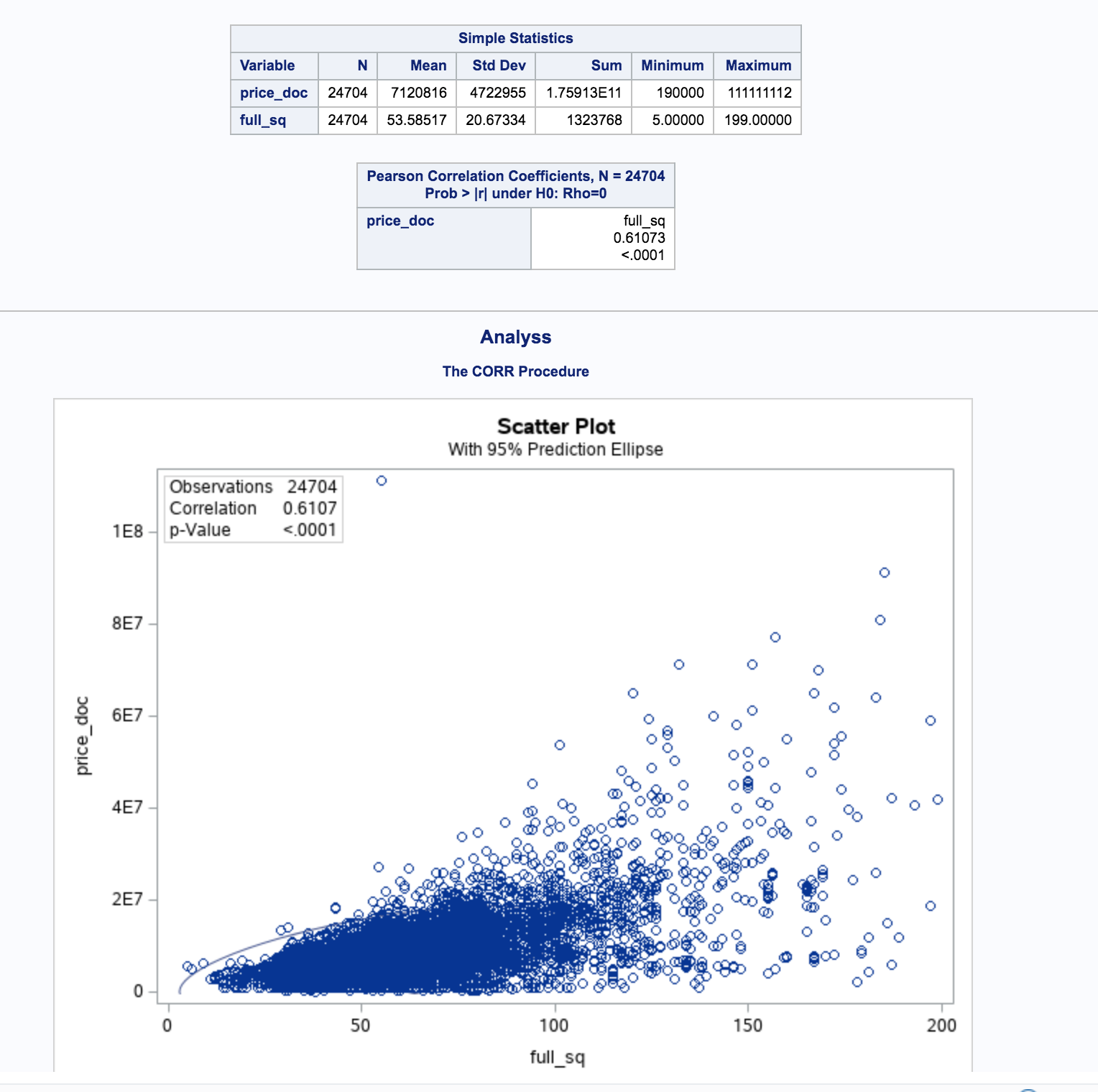
We checked area parameters and they seem to be linearly correlated to our dependent variable : price doc.

**SAS Code :**

proc corr data=housing.clean\_modeling rank PEARSON  
plots (maxpoints = 40000)= all;  
var full\_sq;  
with price\_doc ;  
title Analyss ;  
run;

Check appendix 2 for the code

**Output** :

****

As we can see the price increases linearly as full\_sq values increase. We can assume this data to be linearly related to Price doc and add this to the model without transformation.

### Variable Selection:

We used forward selection, backward selection , stepwise selection and Lasso for variable selections.

Check appendix 2 for code :

## Modeling

## Final Prediction

We created different predictions based on selection models.

## Appendix 1

## Appendix 2

### SAS code to check Assumptions:

**proc** **corr** data=housing.clean\_modeling rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var full\_sq life\_sq floor max\_floor material;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data=housing.clean\_modeling rank PEARSON

plots (maxpoints = **40000**)= all;

var full\_sq;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data= housing.clean\_modeling rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var build\_year num\_room

kitch\_sq state area\_m

;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data=work.refined\_data rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school

;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data=work.refined\_data rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion

### SAS Code to check Multi collinearity:

**proc** **corr** data= housing.clean\_modeling;

var full\_sq life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion

culture\_objects\_top\_25\_raion shopping\_centers\_raion ;

**run**;

**proc** **glm** data=housing.clean\_modeling;

class preschool\_quota children\_school school\_quota;

model price\_doc = full\_sq raion\_popul life\_sq preschool\_quota children\_school school\_quota;

**run**;

### SAS Code for Data Wrangling :

**data** housing.clean\_modeling ;

set housing.modeling;

region=compress(sub\_area,"'");

if full\_sq = **5326** then delete;

if missing(full\_sq) then full\_sq = **54.3**;

if full\_sq >**200** then delete;

if full\_sq <**2** then full\_sq = **54.3**;

if missing(build\_year) then build\_year = **1985**;

if build\_year = **4965** then build\_year = **1965**;

if build\_year = **20052009** then build\_year = **2007**;

if build\_year = **71** then build\_year = **1971**;

if build\_year < **220** then delete;

if state = **33** then state = **3**;

if missing(state) then state = **5**;

if kitch\_sq > **500** then delete;

if missing(num\_room) then num\_room=**2**;

if missing(floor) then floor=**8**;

if floor = **0** then floor =**1**;

if missing(max\_floor) then max\_floor=floor;

if max\_floor = **0** then max\_floor =floor;

if missing(material) then material =**1**;

log\_full\_sq= log(full\_sq);

idd = input(VAR1, **8.**);

**run**;

**data** housing.clean\_projection ;

set housing.projection;

if missing(state) then state = **5**;

if missing(material) then material =**1**;

if build\_year = **71** then build\_year = **1971**;

idd = input(VAR1, **8.**);

if missing(num\_room) then num\_room=**2**;

if missing(state) then state = **5**;

if kitch\_sq > **500** then delete;

if missing(num\_room) then num\_room=**2**;

if missing(floor) then floor=**8**;

if floor = **0** then floor =**1**;

if missing(max\_floor) then max\_floor=floor;

**data** housing.merged\_data;

set housing.clean\_modeling housing.clean\_projection;

### SAS Code For Models:

/\* Backward selection. \*/

**proc** **glmselect** data=housing.merged\_data;

class life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion;

model price\_doc = full\_sq life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion

/ selection=Backward(stop=CV) cvmethod=random(**5**) stats=adjrsq;

output out = results p = predict ;

**run**;

**data** result\_backward\_selection ;

set results;

price\_doc = predict;

if missing(predict) then price\_doc = **140000**;

keep id price\_doc;

where idd > **28000**;

**proc** **export** data=result\_backward\_selection dbms=csv

outfile="/folders/myfolders/stats2\_hw/term\_proj1/output/result\_backward\_selection.csv"

replace;

**run**;

**proc** **print** data = result\_Backward\_selection;

**run**;

/\* Stepwise. \*/

**proc** **glmselect** data=housing.merged\_data;

class life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion;

model price\_doc = full\_sq life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion

/ selection=Stepwise(stop=CV) cvmethod=random(**5**) stats=adjrsq;

output out = results p = predict ;

**run**;