## Tallahassee Real Estate Market

## Analysis Report

## Ting Hu th19d

Department of Statistics, Florida State University

November 10, 2020

**Problem Background:**

Due to the Covid-19 pandemic, the US economy has suffered huge losses. The economy has experienced long-term stagnation and even recession. At the same time, the value of homes in the Tallahassee real estate market has been rising. Compared with the 2007 US financial crisis, this time the real estate market showed a different performance. This phenomenon surprised me.

**Purpose ：**

Analyze whether loan interest rates and raw material prices have a significant impact on real estate prices, build a model to estimate real estate prices.

**Clear raw data and prepare datasets**

**House sold data:** copy from website <https://www.manausa.com> , from 1-4-2017 to 10-30-2020, including 11 communities, has a total of 3837 observations.,the character Type of the data SoldPrice need to be transform to numerical.

|  |  |
| --- | --- |
|  |  |

The follow picture show that the Average house sold price for different livingSize on different Area and Year. Depend on the average house sold price, from lower to higher, each community get a number from 1 to 11. From this picture, there is no obviously trends that the sold house price increase year by year from 2027 till now.

|  |
| --- |
|  |

**Mortgage rate data:** download from <https://fred.stlouisfed.org/series/MORTGAGE30US>

from 1-5-2017 to 11-5-2020, has a total of 201 observations. I add one month for each date, because close date usually one month later than the date you get your Mortgage rate. For those days without mortgage rate, I give them a value equal to the previous one.

|  |  |
| --- | --- |
|  |  |

**Lumber trade price:** download from <https://www.investing.com/commodities/lumber-historical-data> , from 1-3-2017 to 11-3-2020, has a total of 975 obervations

|  |  |
| --- | --- |
|  |  |

**Merge data**: by CloseDate, the new dataset merged three raw data . because lumber date don’t match CloseDate completely, the new data loss 56 observations.

|  |  |
| --- | --- |
|  | |
|  |  |

**Clear data :** check missing data and correct wrong data.

on the BuiltYear row, Minimum column shows that the minimum BuiltYear is 0, this is a some wrong data.

One the Size row, N number shows that there are 3779 observations, less than 3781. There are 2 missing in the Size data.

|  |
| --- |
|  |

**Prepared data:** the prepared data has 3779 obervations,each row of the dataset contains the following 7 vavriables:

|  |  |
| --- | --- |
|  |  |
|  | |

**Date :** the date house closed , between 1-4-2017 and 10-30-2020(log)

**BuitYear :** the year house builted, between 1954 and 2020(log)

**Size:** house living size,between 977 and 9703(log)

**LumberPrice :** Lumber trade price, usd/1000 board feet,between 259.8 and 984.5

**Mortgage :** 30 years fixed interest rate, value between 2.8 and 4.94

**Sold\_Price:** the price house sold, between 50k and 1500K(log)

**Area:** the community house located, have 11 different categorical, depend on average sold price form low to high, set categorical number from 1 to 11.

**multiple linear regression**

**part1)Model Assumptions:**

Random errors

all pairs of random errors are independent.

The Sold\_Price is modeled as a function of Date, BuiltYear, LumberPrice, Size Area and Mortgage .

Response variable(Y) = Sold\_Price

Independent Variables(xi) = Date, BuiltYear, LumberPrice, Size Area and Mortgage.

|  |
| --- |
| Table  Description automatically generated |
| * A **Box Cox** transformation is a transformation of a non-normal dependent variables into a normal shapehere, the suggestion lambda for sold price is 0.35. others don’t need do transform. |

**Part2) scatter plot**

|  |  |
| --- | --- |
| A picture containing diagram  Description automatically generated | |
|  |  |
|  |  |
|  |  |

those plots show that sold price has a linear relation shop with other 6 variables. Sold price has stronger positive linear relationship with Size, has slightly negative linear relationship with Mortgage.

**Part 3) correlation check**

|  |
| --- |
| **Table  Description automatically generated** |

From chart we get, the result show us that response Sold\_Price has linear relationship with all 6 variables, and Mortgage have a negative influence in Sold\_Price. Rate have a very weak correlation with YearBuilt and LivingSize.

**Part4) separate data**

|  |
| --- |
| * **/\* random splitted data \*/ proc surveyselect data=housedata out=want method=srs samprate=0.80 outall seed=12345 noprint; samplingunit Date; run;  data traindata testdata; set work.want; if Selected = 1 then output traindata; else output testdata;  run;** |

The data was be random separated to the train date with 80 percent, and test data with 20 percent.

**Part5) Forward stepwise to choose variables**

|  |  |  |
| --- | --- | --- |
| **Table  Description automatically generated** | | |
| Criterion– represents the fraction of the sample variation of the y values that is explained by the independent variables. One drawback of is adding more independent variables in the model will increase eventually to 1. • Adjusted or MSE Criterion – takes into account the sample size and the number of parameters in the model. increases only if MSE decreases. The largest or smallest MSE indicates the best fit of the model.  Criterion – A small value of indicates that the total mean square error and the regression bias are minimized.  Depending on Adjusted R-Square, AIC ,BIC ,and C(p), the full model and the model without LumberPrice are possible choice. Let’s try full model first. | | |
| 1. **try the full model:** | | |
|  | | |
| Adj R-Sq is 0.8544 indicate that 0.8544 percent of response value SoldPrice is explained by the this full model. But p-value for Lumber Price is 0.1424 ,which indicate that Lumber Price is not significance, the LumberPrice should to be removed from the fitted model. | | |
| 1. **try the model without LumberPrice:** | | |
| Table  Description automatically generated | | |
|  | | Graphical user interface, application  Description automatically generated |
| * The residuals plotted against the predicted values (Row 1, Col 1) show no trends or patterns. If there are any patterns such as the “cone” or “sphere” shapes, this indicates the lack of model fit and unequal variances. * Of all the assumptions, the normality assumption is the least restrictive. The Q-Q plot (Row2, Col 1) shows a linear trend with a slight deviation at the tail, which suggests that the normality assumption is satisfied. * The histogram (Row 3, Col 1) shows the distribution is mound-shaped with a slightly skewed right tail. Studentized Residual vs. * Leverage graph (Row 1, Col 3) shows some potential outliers and influential observations outside of the reference lines. * Cook’s distance shows the influence of each observation on the fitted response values. An observation with Cook’s distance larger than three times the mean Cook’s distance or threshold 4/3079=0.0013 might be an outlier. * After checked the data in outlier (0.08 percent), there is no extreme value in variables, such as SoldPrice, Size, Mortgage rate, and so on. All value in outlier data have the same span as regular sample. | | |
| **try the model after delete the outlier** | | |
| Table  Description automatically generated | | |
| After delete outlier, the adjust R Square is 0.8989. which indicate the model without LumberPrice is better than the full model. | | |
| A picture containing diagram  Description automatically generated | Graphical user interface, diagram, application  Description automatically generated | |
| * The points on the plot of the dependent variable (Sold\_Price) versus the predicted values lie along a 45-degree line, indicating that the model successfully predicts the behavior of the dependent variable. * The normal quantile plot of the residuals and the residual histogram are consistent with the assumption of Gaussian errors. * The plot of the dependent variable versus the predicted value exhibits a quadratic form around the 45-degree line that represents a perfect fit.   The “Residual-Fit” (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals shows that the spread in the residuals is no greater than the spread in the centered fit. For inappropriate models, the spread of the residuals in such a plot is often greater than the spread of the centered fit. In this case, the RF plot shows that the linear model does indeed capture the increasing trend in the data, and hence accounts for much of the variation in the response. | | |

**Part 6) model test:**

|  |
| --- |
| **True value VS Forecast Forecast = 80.86793 + 1.15099\*Date+2.43519\*BuiltYear+10.64085\*Size  +0.48883\*Area-0.22751\*Mortgage;** |
| **Chart, line chart  Description automatically generated** |
| * The result can be accepted, but not reasonable. The Area should be considered as a categorical number then as a numerical number. * Let’s do the model with the categorical Area. |

**Part7). Model Adequacy: try the model with category Area:**

|  |  |  |
| --- | --- | --- |
|  | |  |
|  | |  |
|  |  | |
|  |  | |
|  |  | |
|  |  | |

**Part8) The result of the model 1 VS model 2**

|  |  |
| --- | --- |
| **Model1: Area as a class gategory**  data testF; set testdata; if Area = 1 then  Forecast = 80.64266 + 1.83433\*Date+2.65367\*BuiltYear+8.31403\*Size+0.18671\*Mortgage; if Area = 2 then  Forecast = 79.01669 + 1.29438\*Date+5.76099\*BuiltYear+10.24671\*Size-0.18215\*Mortgage; if Area = 3 then  Forecast = 81.56427 + 1.22014\*Date+2.23623\*BuiltYear+10.09379\*Size-0.56806\*Mortgage; if Area = 4 then  Forecast = 82.3774 + 1.10362\*Date+1.95486\*BuiltYear+10.67578\*Size-0.08247\*Mortgage; if Area = 5 then  Forecast = 74.65604 + 0.02874\*Date-2.97073\*BuiltYear+4.49431\*Size-2.39495\*Mortgage; if Area = 6 then  Forecast = 79.65507 + 0.67039\*Date+6.01361\*BuiltYear+10.14203\*Size-0.27673\*Mortgage; if Area = 7 then  Forecast = 83.75358 + 1.02303\*Date+3.94321\*BuiltYear+10.71405\*Size-0.4387\*Mortgage; if Area = 8 then  Forecast = 86.1288 + 1.03477\*Date+2.2486\*BuiltYear+9.00189\*Size-0.18412\*Mortgage; if Area = 9 then  Forecast = 87.33043 + 0.91533\*Date+2.32301\*BuiltYear+6.04445\*Size-0.48667\*Mortgage; if Area = 10 then  Forecast = 83.54671 + 0.98859\*Date+4.43417\*BuiltYear+8.87047\*Size-0.47849\*Mortgage; if Area = 11 then  Forecast = 85.73354 + 1.70971\*Date+12.98992\*BuiltYear+4.56618\*Size+0.82731\*Mortga;  i = \_n\_ ;  Truevalue = Sold\_Price;  per = int(Forecast/Truevalue\*100);  Run; | |
| **Model2:Area as a continue number**  Forecast = 80.86793 + 1.15099\*Date+2.43519\*BuiltYear+10.64085\*Size  +0.48883\*Area-0.22751\*Mortgage; | |
| Model1 | Model 2 |
|  |  |

|  |  |
| --- | --- |
|  |  |

**Per = Forecast / SoldPrice \* 100.**

**For the model 1): the percent that Per locate in the 100% + - 10% span is 96.7%-0.86% = 95.84%**

**For the model 2): the percent that Per locate in the 100% + - 10% span is 95.84%-1.58% = 94.26%**

**Part9) summary:**

* **House sold price don’t response to lumber price**
* **Mortgage Rate have negative contributed to house sold price**
* **The model looks good but not perfect to estimate the house value**
* **The model built with categorical Area is better than the model built with numerical Area.**

**code**

|  |
| --- |
| Part1) multiple regression analyze:  FILENAME data5939 '/home/u45032752/myfolder/data5939.csv'; proc import datafile=data5939  DBMS=csv  out= test;  Getnames= YES; quit;  proc transreg details data=work.test detail pbo;  model BoxCox(Sold\_Price / convenient lambda=-2 to 2 by 0.05) =  identity(Date BuiltYear LumberPrice Size Area Mortgage); run; /\*normalize transfer \*/ proc sql ; create table house1 as select   Sold\_Price\*\*0.35 as SoldPrice ,  (Date - mean(Date))/std(Date) as CloseDate ,  (BuiltYear - mean(BuiltYear))/std(BuiltYear) as YearBuilt ,  (LumberPrice - mean(LumberPrice))/std(LumberPrice) as Lumber ,  (Size- mean(Size))/std(Size) as LivingSize,  (Mortgage - mean(Mortgage))/std(Mortgage) as Rate,  Area from work.test ; run;  proc sql ; create table housedata as select   SoldPrice as Sold\_Price ,  CloseDate as Date ,  YearBuilt as BuiltYear,  Lumber as LumberPrice ,  LivingSize as Size,  Rate as Mortgage,  Area from work.house1 ; run; proc transreg details data=work.housedata detail pbo;  model BoxCox(Sold\_Price / convenient lambda=-2 to 2 by 0.05) =  identity(Date BuiltYear LumberPrice Size Area Mortgage); run;  /\* scatter plot \*/  PROC SGSCATTER DATA=work.housedata;   MATRIX Sold\_Price Date BuiltYear LumberPrice Size Area Mortgage/diagonal=(histogram normal) ; RUN;  /\* correlaion \*/ PROC CORR DATA = housedata ; VAR Sold\_Price Date BuiltYear LumberPrice Size Area Mortgage; RUN ;  /\* random splitted data \*/ proc surveyselect data=housedata out=want method=srs samprate=0.80  outall seed=12345 noprint;  samplingunit Date; run;  data traindata testdata; set work.want; if Selected = 1 then output traindata;  else output testdata;  run;  /\* model building\*/  /\*stepwise to choose variable\*/ proc reg data=work.traindata; model Sold\_Price = Date BuiltYear LumberPrice Size Area Mortgage / selection= rsquare adjrsq AIc BIC cp mse best=3; quit;  /\*built the model depend on the result from stepwise\*/ /\* try the full model \*/ proc reg data=work.traindata; ALL\_REG:  model Sold\_Price = Date BuiltYear LumberPrice Size Area Mortgage; quit; /\* model without LumberPrice\*/ proc reg data=work.traindata; ALL\_REG:  model Sold\_Price = Date BuiltYear Size Area Mortgage; quit;  /\* delete outlier \*/ proc reg data=work.traindata noprint;  PREDICT:   model Sold\_Price = Date BuiltYear Size Area Mortgage/ r influence;  output out=outliers   rstudent=rstud dffits=dfits cookd=cooksd;  title; run; quit; /\* set the values of these macro variables, \*/ /\* based on your data and model. \*/ %let numparms = 6; /\* # of predictor variables + 1 \*/  %let numobs = 3079; /\* # of observations \*/ /\*EM peocedure\*/  data influential;  set work.outliers;   cutdifts =2\*(sqrt(&numparms/&numobs));  cutcookd = 4/&numobs;   rstud\_i = (abs(rstud)>3);  dfits\_i = (abs(dfits)>cutdifts);  cookd\_i = (cooksd>cutcookd);  sum\_i = rstud\_i + dfits\_i + cookd\_i;  if sum\_i > 0; run; /\* delete outlier data from dataset \*/  proc sql ; create table train\_data as select Sold\_Price,Date, BuiltYear, Size, Area,LumberPrice, Mortgage from work.traindata  except select Sold\_Price,Date, BuiltYear, Size, Area,LumberPrice, Mortgage from influential ; quit;  /\* model after outlier deleted\*/ proc reg data=work.train\_data; ALL\_REG:  model Sold\_Price = Date BuiltYear Size Area Mortgage; quit;   /\* testdata test \*/ data housetest; set testdata; Forecast = 80.86793 + 1.15099\*Date+2.43519\*BuiltYear+10.64085\*Size+0.48883\*Area-0.22751\*Mortgage; i = \_n\_ ; Truevalue = Sold\_Price; per = int(Forecast/Truevalue\*100); run;  proc sgplot data=housetest; \*band x=i lower=L95 upper=U95; series x=i y=Truevalue; series x=i y=Forecast; where i <= 50; run;  /\* data sort by asending Area\*/ proc sort data= train\_data; by Area; run; /\* model building with category Area\*/ proc reg data=work.traindata; ALL\_REG:  model Sold\_Price = Date BuiltYear Size Area Mortgage; by Area; quit;  /\* test the class Area model\*/ data testF; set testdata; if Area = 1 then Forecast = 80.64266 + 1.83433\*Date+2.65367\*BuiltYear+8.31403\*Size+0.18671\*Mortgage; if Area = 2 then Forecast = 79.01669 + 1.29438\*Date+5.76099\*BuiltYear+10.24671\*Size-0.18215\*Mortgage; if Area = 3 then Forecast = 81.56427 + 1.22014\*Date+2.23623\*BuiltYear+10.09379\*Size-0.56806\*Mortgage; if Area = 4 then Forecast = 82.3774 + 1.10362\*Date+1.95486\*BuiltYear+10.67578\*Size-0.08247\*Mortgage; if Area = 5 then Forecast = 74.65604 + 0.02874\*Date-2.97073\*BuiltYear+4.49431\*Size-2.39495\*Mortgage; if Area = 6 then Forecast = 79.65507 + 0.67039\*Date+6.01361\*BuiltYear+10.14203\*Size-0.27673\*Mortgage; if Area = 7 then Forecast = 83.75358 + 1.02303\*Date+3.94321\*BuiltYear+10.71405\*Size-0.4387\*Mortgage; if Area = 8 then Forecast = 86.1288 + 1.03477\*Date+2.2486\*BuiltYear+9.00189\*Size-0.18412\*Mortgage; if Area = 9 then Forecast = 87.33043 + 0.91533\*Date+2.32301\*BuiltYear+6.04445\*Size-0.48667\*Mortgage; if Area = 10 then Forecast = 83.54671 + 0.98859\*Date+4.43417\*BuiltYear+8.87047\*Size-0.47849\*Mortgage; if Area = 11 then Forecast = 85.73354 + 1.70971\*Date+12.98992\*BuiltYear+4.56618\*Size+0.82731\*Mortgage; i = \_n\_ ; Truevalue = Sold\_Price; per = int(Forecast/Truevalue\*100); \*keep Truevalue Forecast i; run;  proc sgplot data=testF; \*band x=i lower=L95 upper=U95; series x=i y=Truevalue; series x=i y=Forecast; where i <= 50; run;  proc freq data= testF nlevels; table per; run;  proc freq data= housetest nlevels; table per; run; |
| Part2) dataset prepare:  FILENAME data '/home/u45032752/myfolder/5939data.csv'; /\* read the tallahassee real estate market data\*/ PROC IMPORT DATAFILE=data  DBMS=CSV  OUT=WORK.realmarket;  GETNAMES=YES; RUN;  /\* check the data set\*/ proc print data= work.realmarket(obs=10);run;  /\*check the data format \*/  proc contents data=work.realmarket;run;  /\*format date\*/ data tallahassee; set work.realmarket(rename=(CloseDate = Date )); format Date mmddyy10.; run;  /\* proc print data = tallahassee(obs=10);run; \*/  FILENAME lumber '/home/u45032752/myfolder/5939lumber.csv'; /\* import lumber history price\*/  PROC IMPORT DATAFILE=lumber  DBMS=CSV  OUT=WORK.lumberprice;  GETNAMES=YES; RUN;  /\* check the data set \*/ proc print data = work.lumberprice(obs=10);run;  /\*check the date format\*/ proc contents data= work.lumberprice;run;   data lumber;  set work.lumberprice(rename=(Price = LumberPrice));  Date = datepart(Date);  format Date mmddyy10.;  run;    /\* merge the house data with lumber data\*/ proc sql; create table new1 as select \*   from work.tallahassee as t,  work.lumber as l  where t.Date = l.Date;  quit;  proc contents data=new1;run;  FILENAME interest '/home/u45032752/myfolder/5939mortgage.csv'; /\*import interest data\*/ proc import datafile=interest  DBMS=csv  out= mortgage;  Getnames= YES; run;  /\*check the data set\*/ proc print data= work.mortgage(obs=50);run;  /\* check the data format\*/ proc contents data=work.mortgage;run;  data interest; set work.mortgage(rename=(DATE = Date)); format Date mmddyy10.; run;  proc sql; create table new2 as  select \*  from work.new1 as n  left join work.interest as i  on n.Date = i.Date; run;  proc print data=new2(obs=30);run;  proc contents data= new2;run;  /\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/ FILENAME data1 '/home/u45032752/myfolder/data.csv'; proc import datafile=data1  DBMS=csv  out= data2;  Getnames= YES; run; /\*check the data \*/ proc contents data=data2;run;   /\* prepare data set : correct wrong data and delete missing rows\*/ data data ;  set data2(rename=(MORTGAGE30US = Mortgage Price = LumberPrice));  SoldPrice = substr(SoldPrice,2,length(SoldPrice)-1);  SoldPrice=compress(scan(SoldPrice, 1, ",")||scan(SoldPrice, 2, ","));  Sold\_Price = input(SoldPrice,8.);  if BuiltYear = 0 then do;  BuiltYear = 2020;  end;  if Size eq . then delete;  drop Address Price\_sqf SoldPrice ; run;    title "check whether some wrong or missing in the row data"; proc means data=data; var Sold Date BuiltYear Size Community LumberPrice Mortgage; run; title ;    /\* get average value for deffirent community\*/ proc sql; create table community as   select mean(Sold\_Price) "Average" format = Best12., Community  from work.data  group by Community  ; quit;  /\*let area get the numeric depend on average value\*/ data new3; set work.data;  select(Community);  when(6) do;  Area = 1;  end;  when(7) do;  Area = 2;  end;  when(11) do;  Area = 3;  end;   when(10) do;  Area = 4;  end;   when(8) do;  Area = 5;  end;   when(2) do;  Area = 6;  end;   when(3) do;  Area = 7;  end;   when(5) do;  Area = 8;  end;   when(4) do;  Area = 9;  end;   when(1) do;  Area = 10;  end;   otherwise do;  Area = 11;  end;   end; drop Community; run; run; /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/ FILENAME data5939 '/home/u45032752/myfolder/5939data1.csv'; proc import datafile=data5939  DBMS=csv  out= newdata;  Getnames= YES; /\*check dataset\*/ title"dataset after clearing"; proc means data=newdata;run; title;  title "prepared dataset"; proc print data=work.newdata(obs=5);run; title; proc contents data= work.newdata;run; /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/  /\*create new dataset \*/ data analyze1; length LivingSize $8 Community $30; set work.newdata; Quarter = qtr(Date); Month = month(Date); Year = year(Date); select(Year);  when(2017) do; Sold2017 = Sold\_Price;end;  when(2018) do; Sold2018 = Sold\_Price;end;  when(2019) do; Sold2019 = Sold\_Price;end;  otherwise do; Sold2020 = Sold\_Price;end;   end; if Size<1600 then   LivingSize = "low"; if Size >2400 then   LivingSize = "high"; if 1600 <= Size <= 2400 then   LivingSize = "middle"; select (Area);  when(1) do;Community = 'kln-acres'; end;  when(2) do;Community = 'piney-z'; end;   when(3) do;Community = 'kln-estates'; end;  when(4) do;Community = 'kln-lakes'; end;  when(5) do;Community = 'waverly'; end;  when(6) do;Community = 'southwood'; end;  when(7) do;Community = 'bullrun'; end;  when(8) do;Community = 'xbottom'; end;  when(9) do;Community = 'sumbrook'; end;  when(10) do;Community = 'Goldeneagle'; end;  otherwise do;Community = 'centerville'; end;  end; run; /\*create new dataset \*/ data analyze1; length LivingSize $8 Community $30; set work.newdata; Quarter = qtr(Date); Month = month(Date); Year = year(Date); select(Year);  when(2017) do; Sold2017 = Sold\_Price;end;  when(2018) do; Sold2018 = Sold\_Price;end;  when(2019) do; Sold2019 = Sold\_Price;end;  otherwise do; Sold2020 = Sold\_Price;end;   end; if Size<1600 then   LivingSize = "Small"; if Size >2400 then   LivingSize = "Large"; if 1600 <= Size <= 2400 then   LivingSize = "middle"; select (Area);  when(1) do;Community = 'kln-acres'; end;  when(2) do;Community = 'piney-z'; end;   when(3) do;Community = 'kln-estates'; end;  when(4) do;Community = 'kln-lakes'; end;  when(5) do;Community = 'waverly'; end;  when(6) do;Community = 'southwood'; end;  when(7) do;Community = 'bullrun'; end;  when(8) do;Community = 'xbottom'; end;  when(9) do;Community = 'sumbrook'; end;  when(10) do;Community = 'Goldeneagle'; end;  otherwise do;Community = 'centerville'; end;  end; run;  /\* check the sold price distribute by year\*/ proc sql; create table trend as select Month,   avg(Sold2017) as y2017,   avg(Sold2018) as y2018,   avg(Sold2019) as y2019,   avg(Sold2020) as y2020 from work.analyze1 group by Month; run; quit;  symbol1 value=dot cv=red interpol=join ci=red line=4; symbol2 value=+ cv=green interpol=join ci=green line=4; symbol3 value=# cv=yellow interpol=join ci=yellow line=4; symbol4 value=\* cv=blue interpol=join ci=blue line=4; proc gplot data=work.trend; title 'Yearly mean Sold price Series by month'; plot y2017\*MOnth y2018\*Month y2019\*Month y2020\*Month/overlay legend; run; quit;  /\* show the average sold price by month\*/ data yprice ; set analyze1; where LivingSize = "middle"; select(Year);   when(2017) do; Month = Month;end;  when(2018) do; Month = Month+12;end;  when(2019) do; Month = Month+24;end;  otherwise do; Month = Month+36;end;   end; keep Sold\_Price Month; run;  proc sql; create table YearPrice as select Month, avg(Sold\_Price) as Price from work.yprice group by Month ; run; quit;  symbol1 value=dot cv=red interpol=join ci=red line=4; proc gplot data=work.YearPrice; title "Average Sales Price trend by year on middle size house"; plot Price\*Month; run ;  /\* analyze the mean sold price on different living size by community and year\*/  title "Average Sales Price by LivingSize on diffirent Area and year"; legend1 cborder=black label=("Year:") position=(bottom right outside) mode=protect across=1; proc gchart data=analyze1;  block LivingSize /sumvar=Sold\_Price   type= mean  group= Area  subgroup= Year  legend=legend1  noheading; run; quit; /\* list Area -- Community\*/ proc sql; select mean(Area) as Area,Community from work.analyze1 group by Community order by Area; run;  title "Average Sales Price by LivingSize on diffirent Community and year"; legend1 cborder=black label=("Year:") position=(bottom right outside) mode=protect across=1; proc gchart data=analyze1;  block LivingSize /sumvar=Sold\_Price    type= mean  group= Community  subgroup= Year  legend=legend1  noheading; run; quit; |