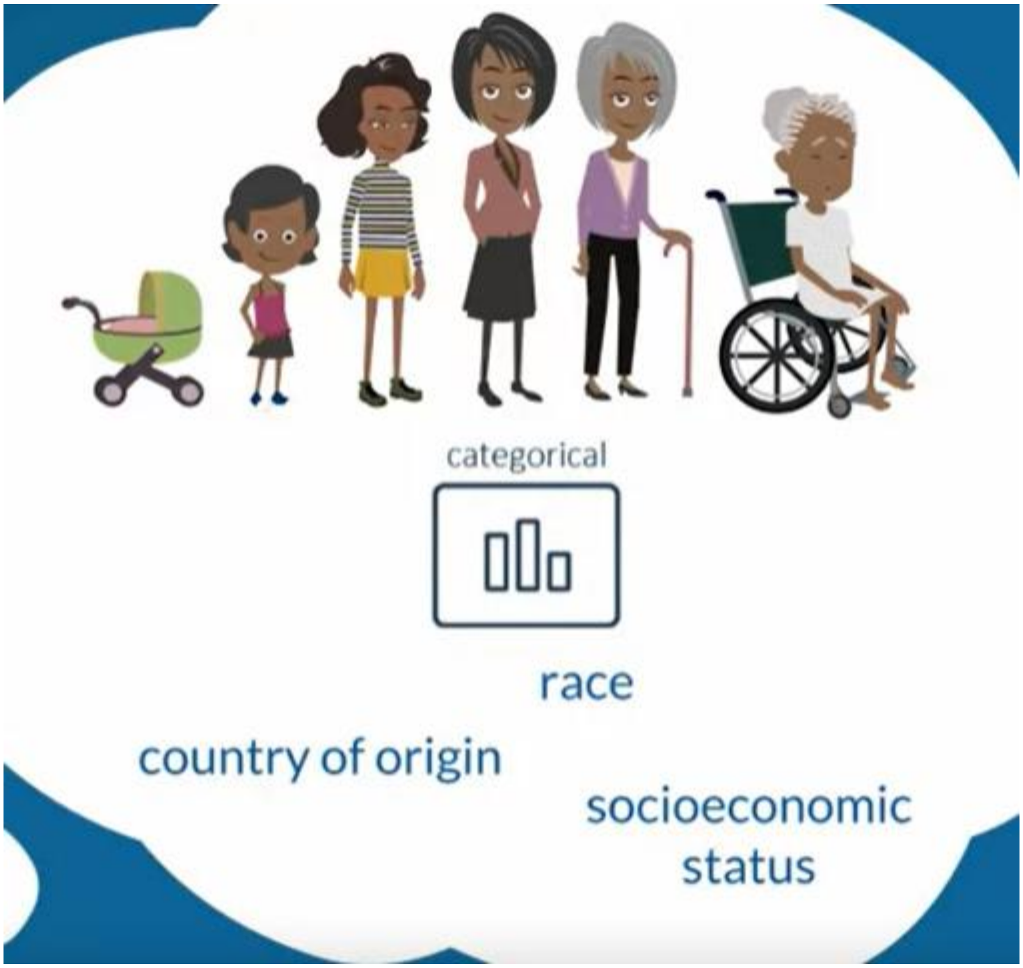
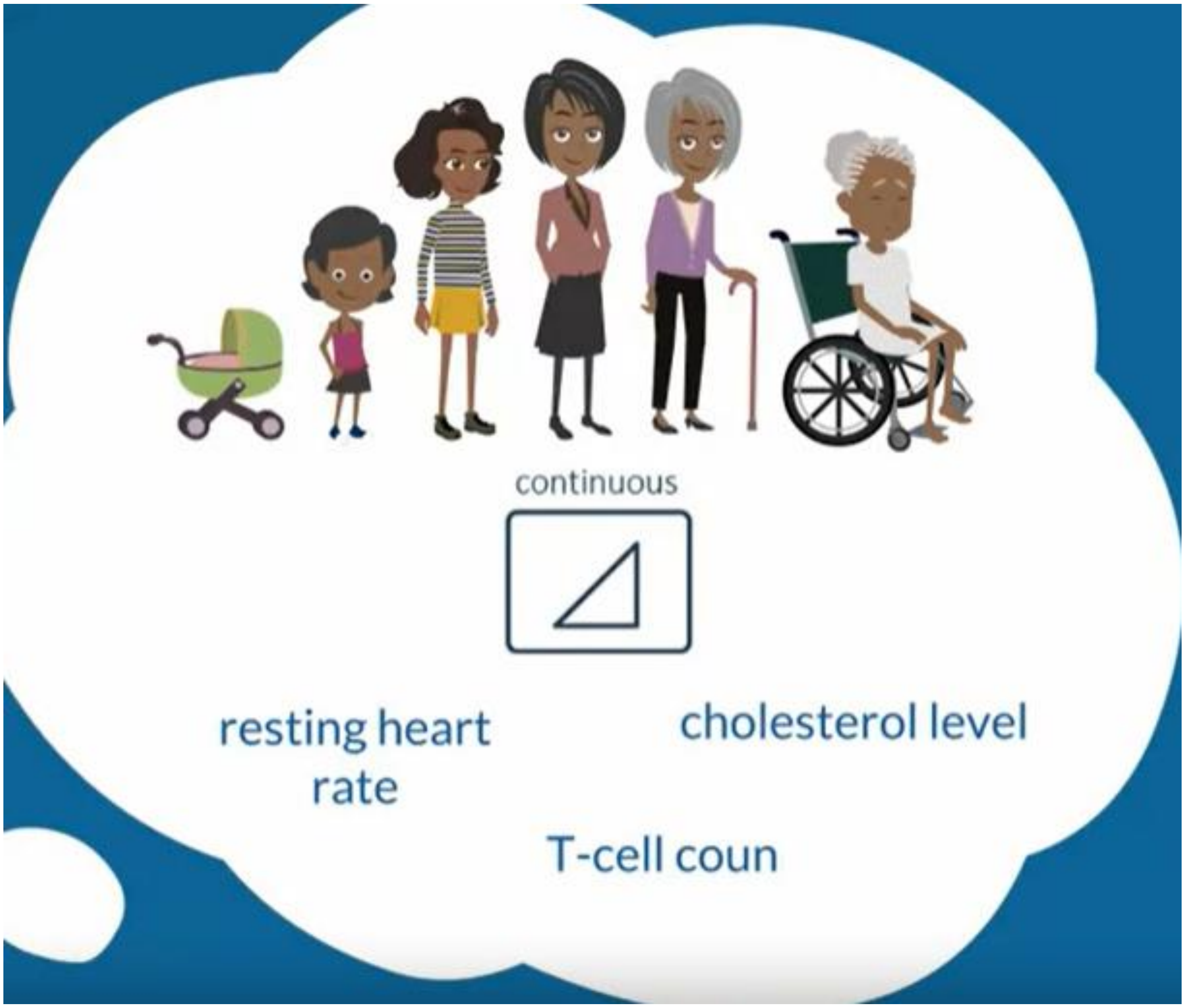
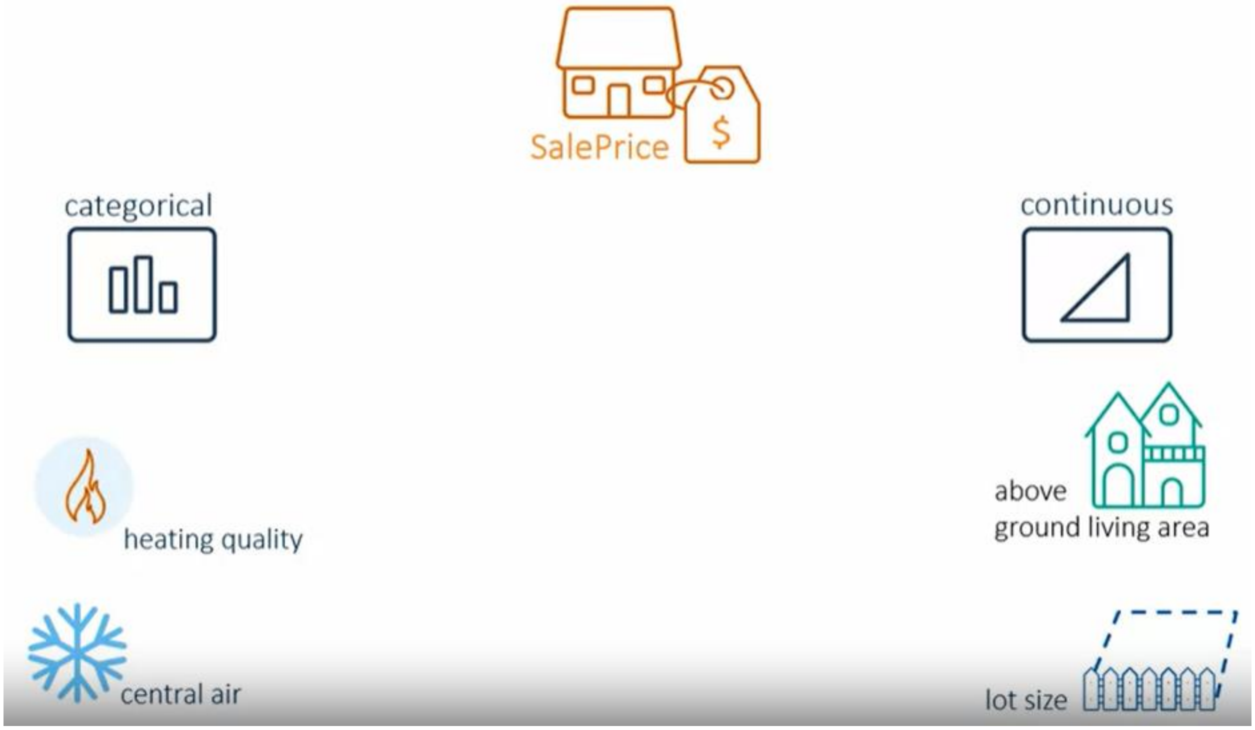
SBA: Statistical Business Analyst with SAS

SBA1: Introduction to Statistical Analysis: Hypothesis Testing

ANOVA Overview




continuous
resting heart rate
cholesterol level
T-cell coun

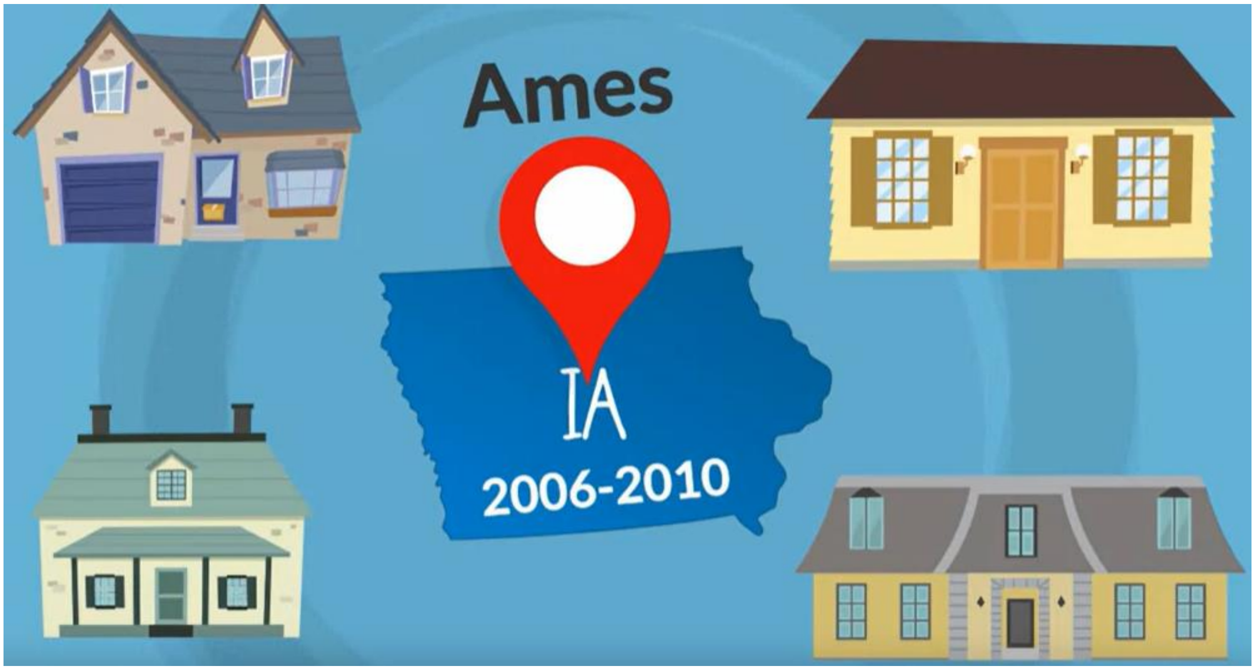

SalePrice
categorical
continuous
heating quality
above
ground living area
central air
lot size

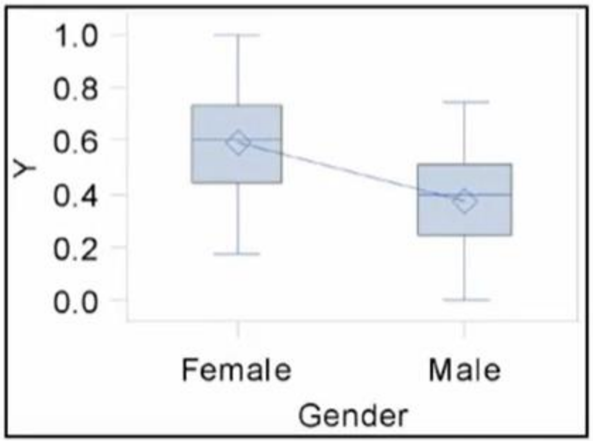

SalePrice

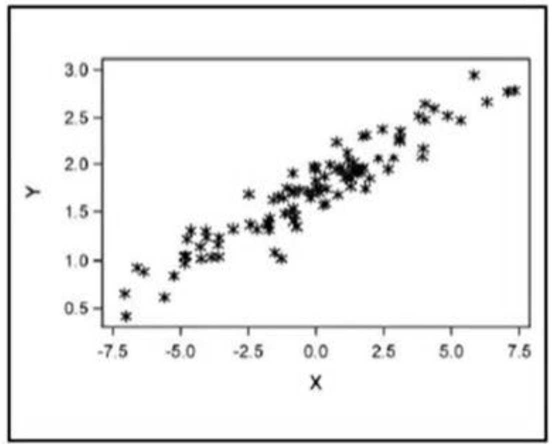

SalePrice
$

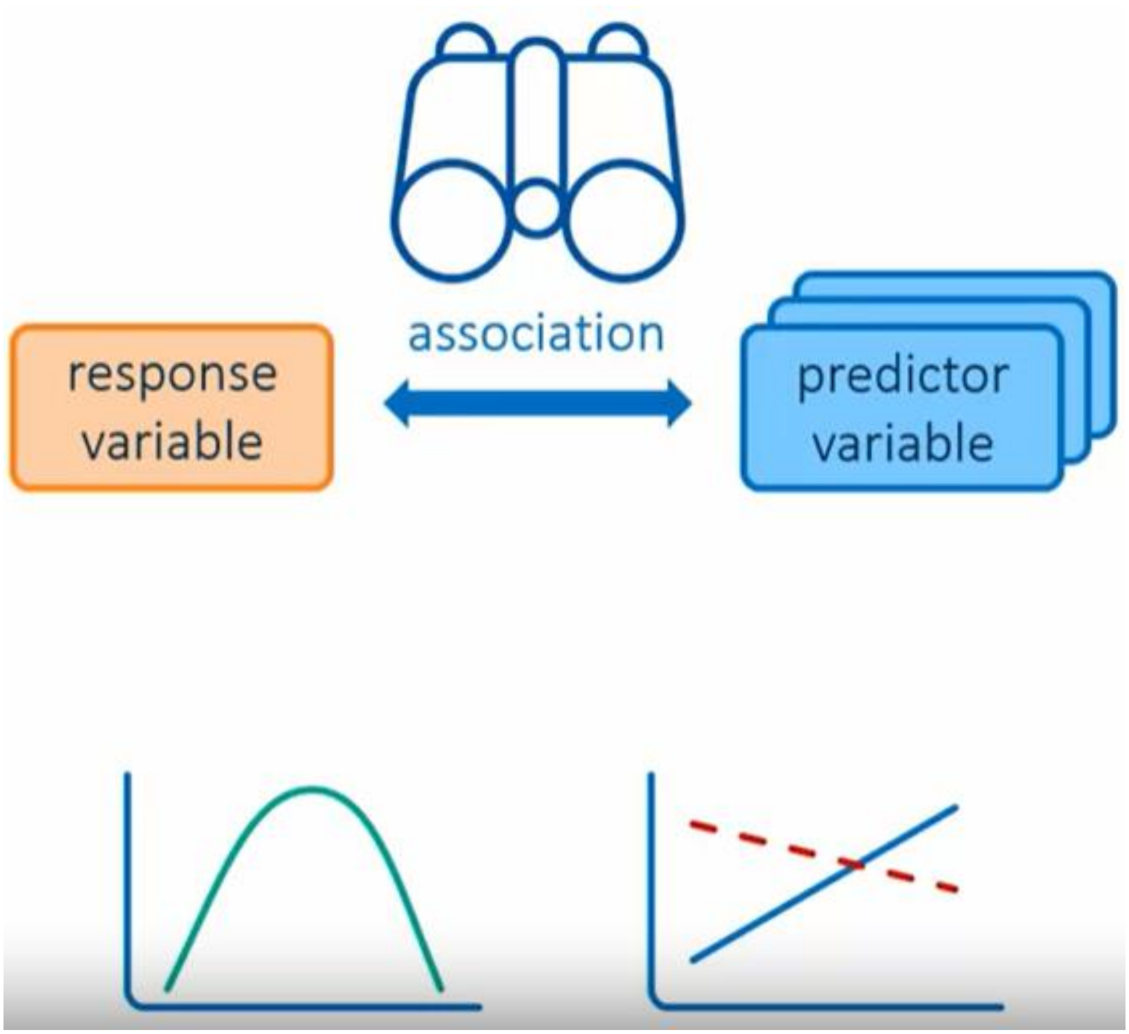

SalePrice

ANOVA Scenario

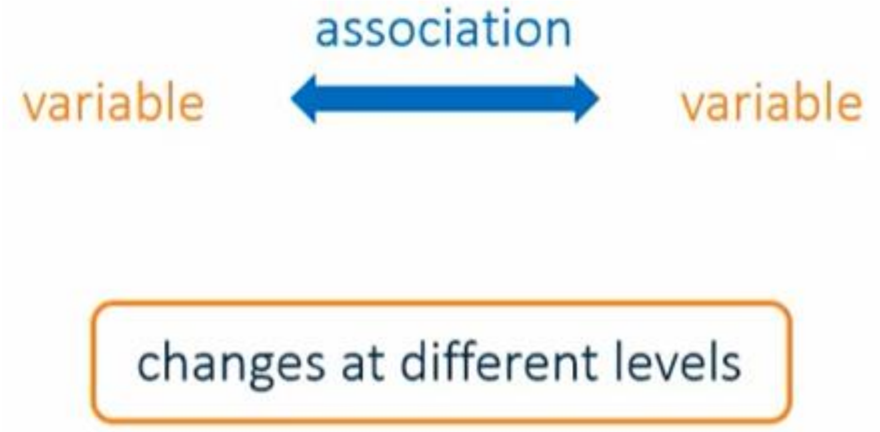


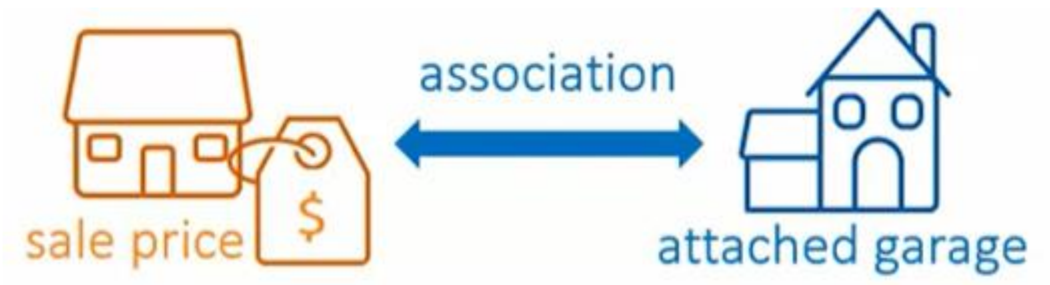

1.0
0.8
0.6
0.4
0.2
0.0
Y
Female
Male
Gender

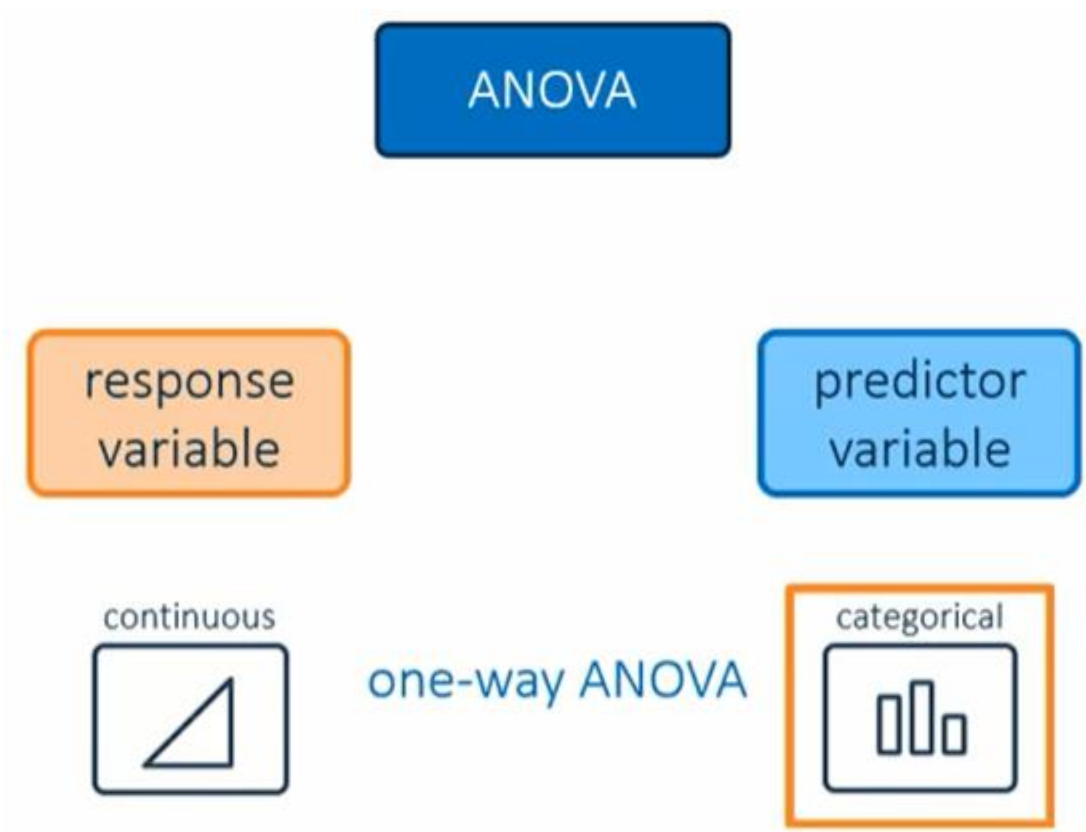

3.0
2.5
2.0
1.5
1.0
0.5
Y
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
X

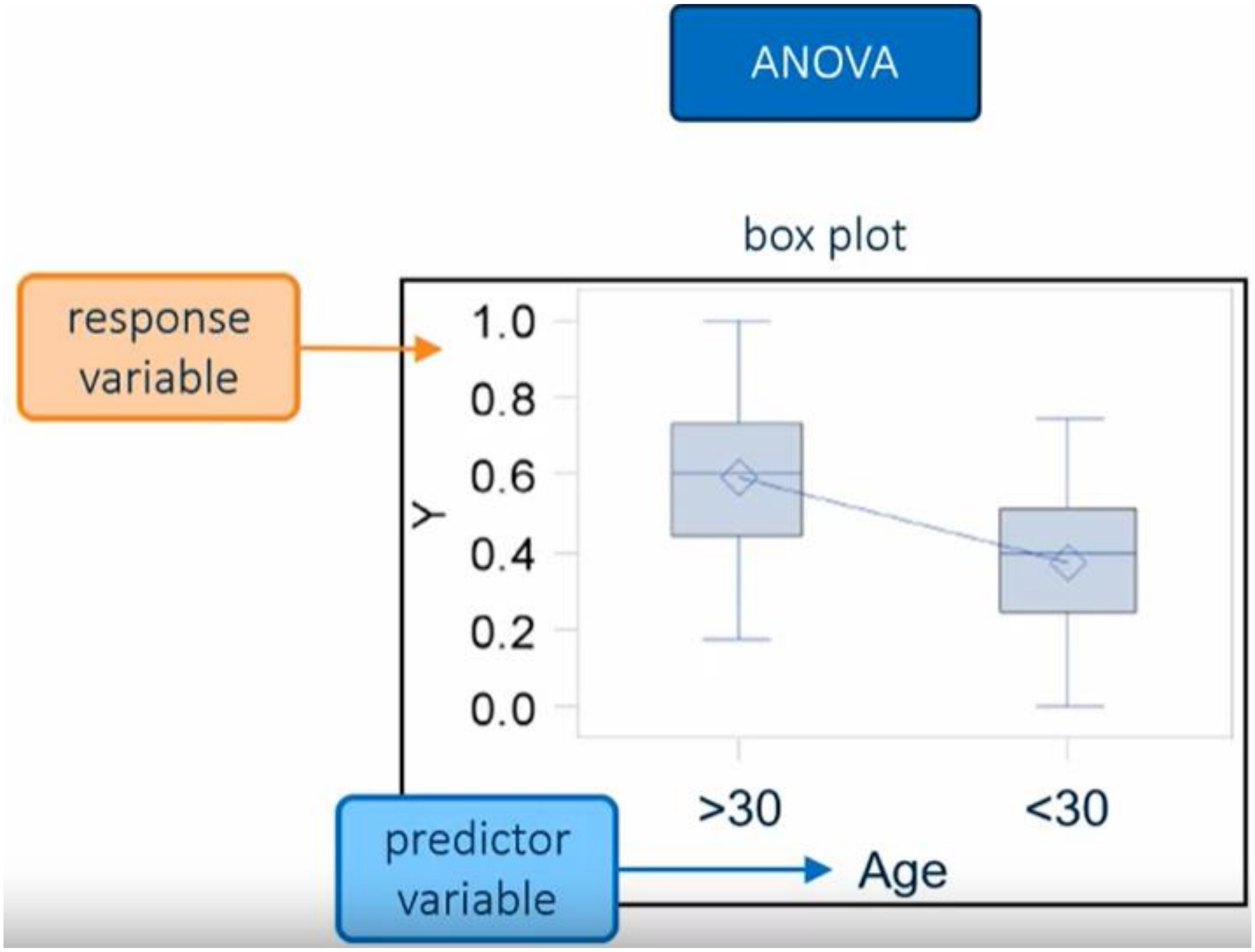

association
response variable
predictor variable

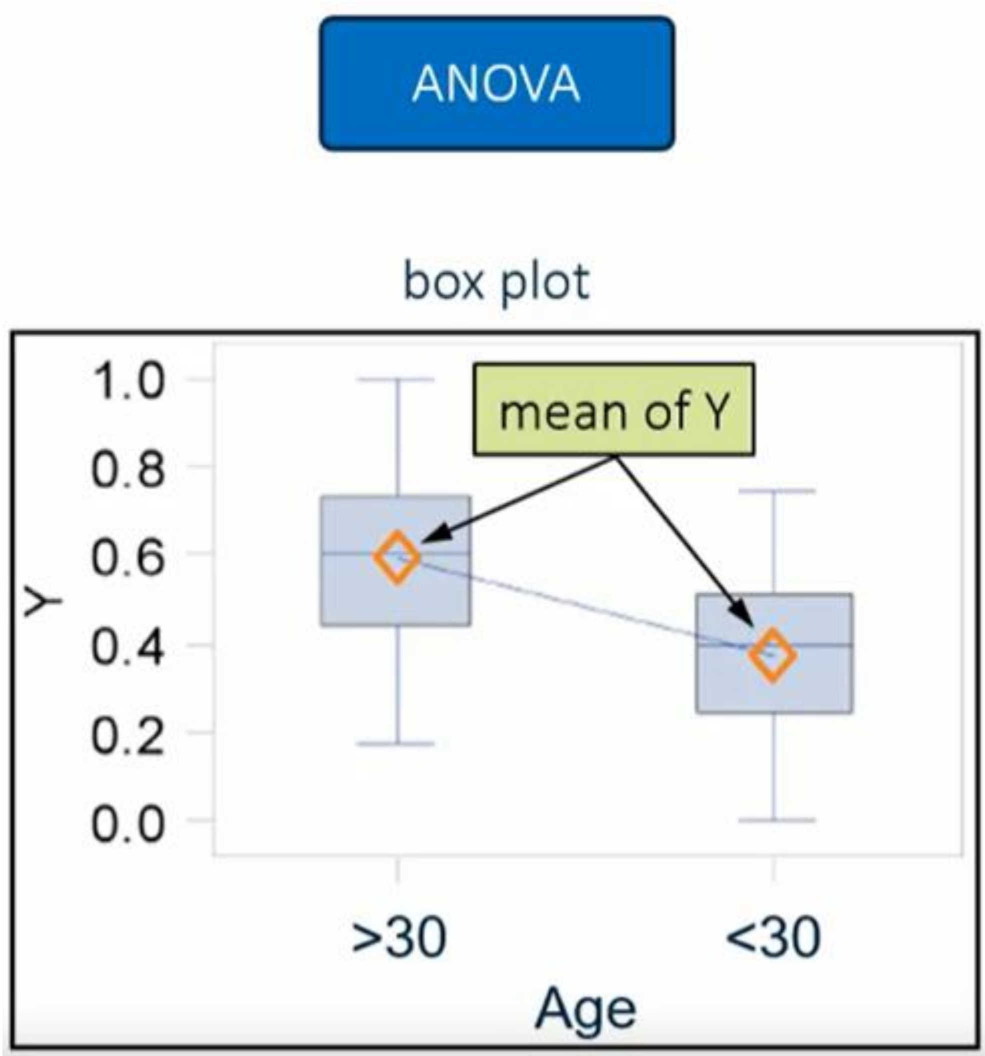
Identifying Associations in ANOVA with Box Plots

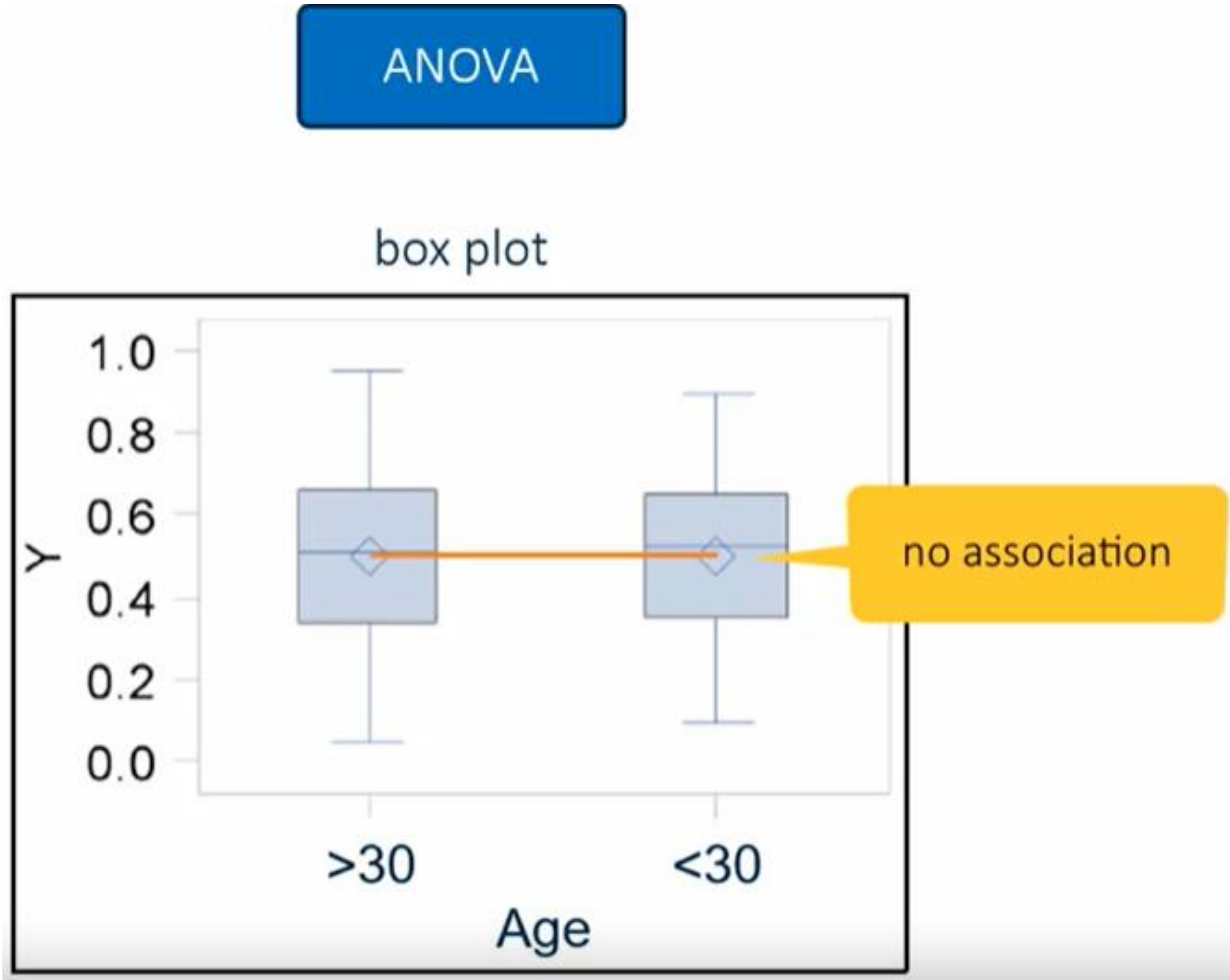


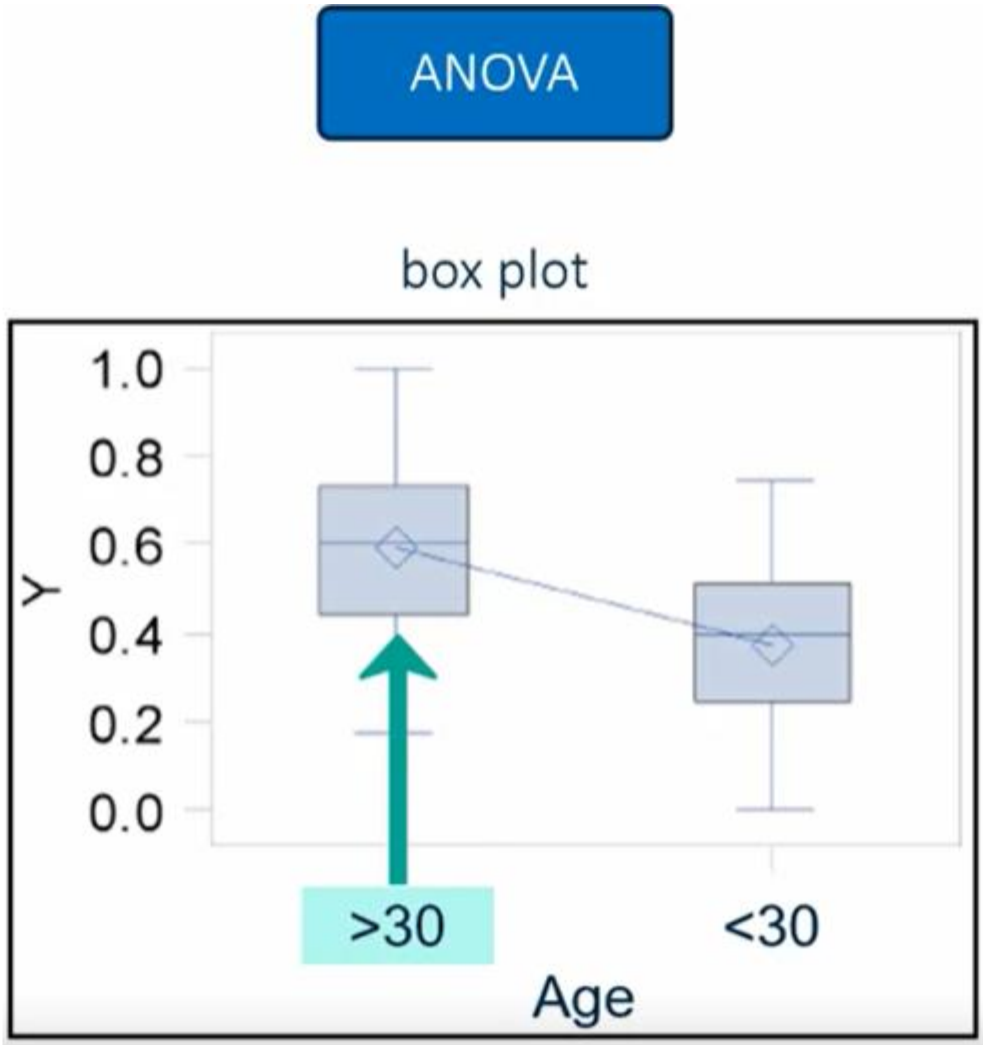




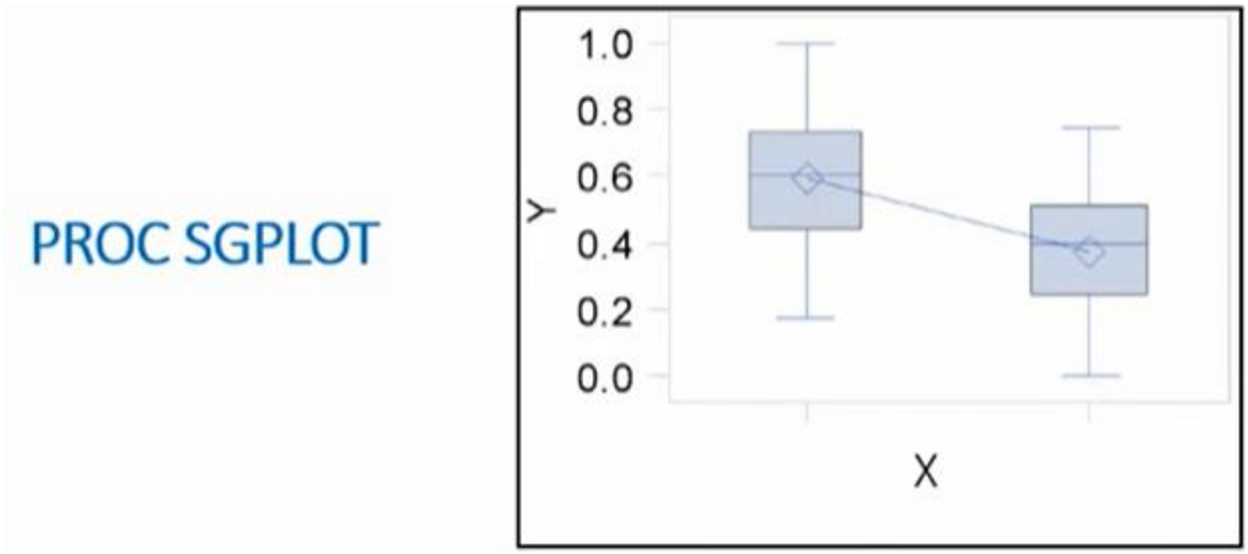

ANOVA
box plot
response variable
1.0
0.8
0.6
0.4
0.2
0.0
Y
>30
<30
predictor variable
Age

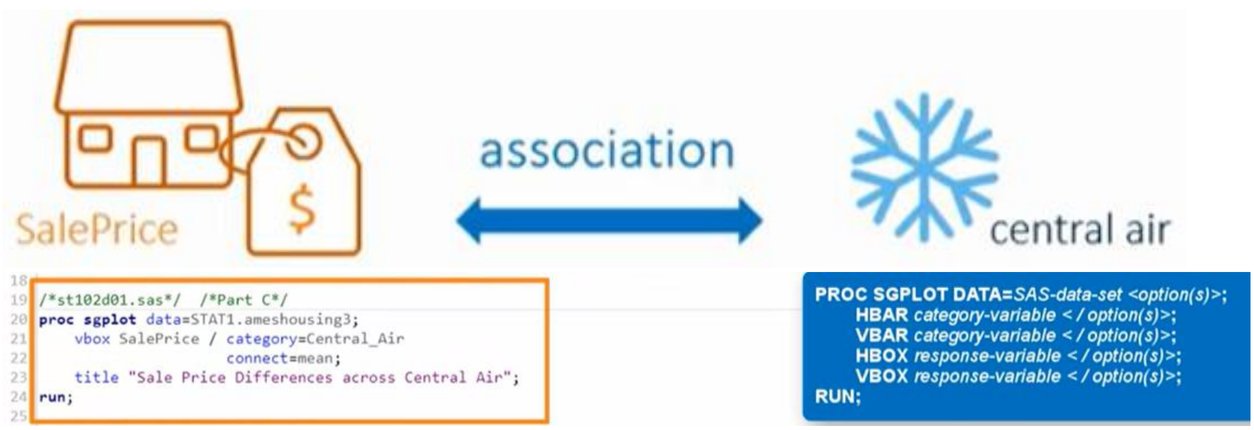

ANOVA
box plot
mean of Y
1.0
0.8
0.6
0.4
0.2
0.0
Y
>30
<30
Age

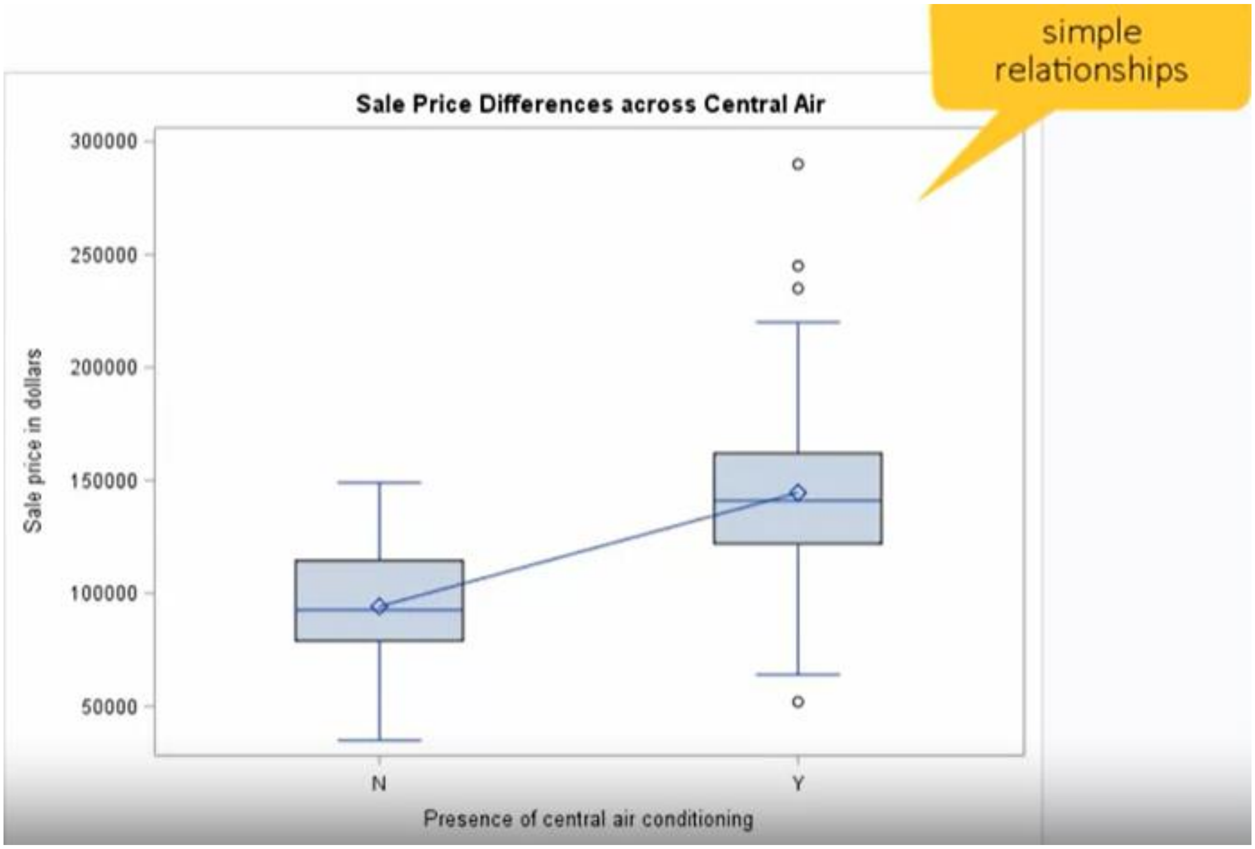

ANOVA
box plot
1.0
0.8
0.6
0.4
0.2
0.0
Y
>30
<30
Age
no association


ANOVA
box plot
1.0
0.8
0.6
0.4
0.2
0.0
Y
>30
<30
Age

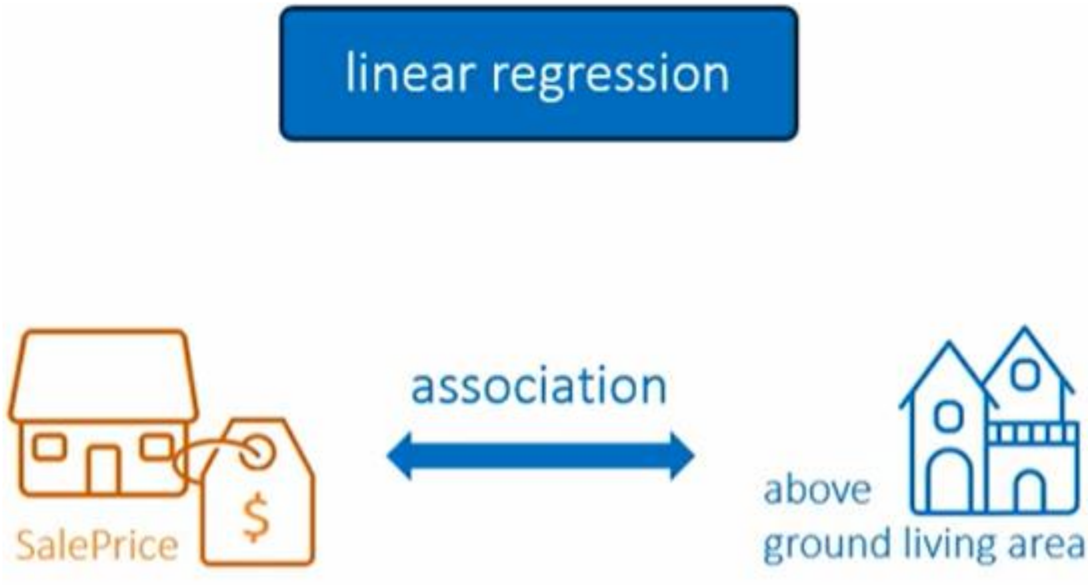
Demo Exploring Associations Using PROC SGPLOT

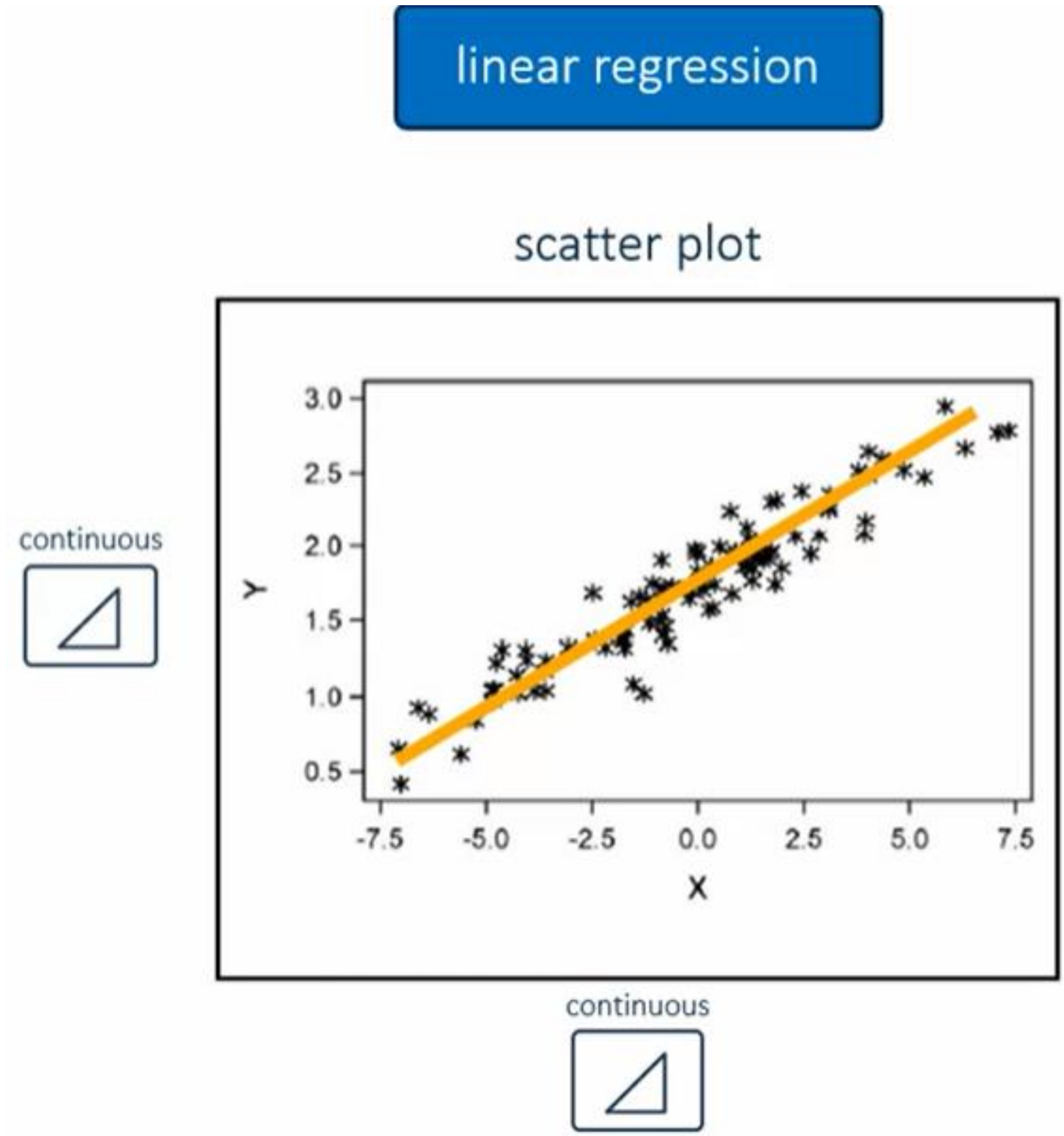


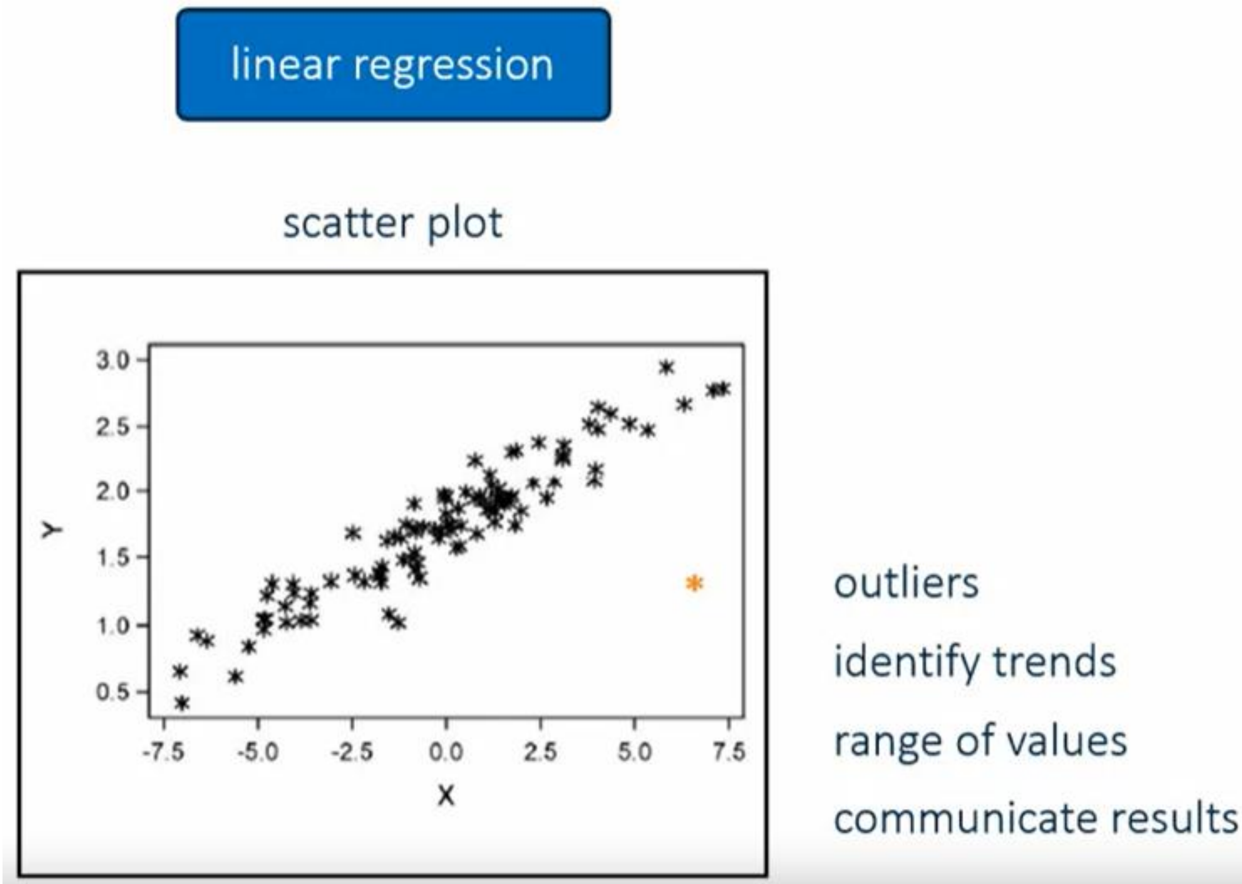


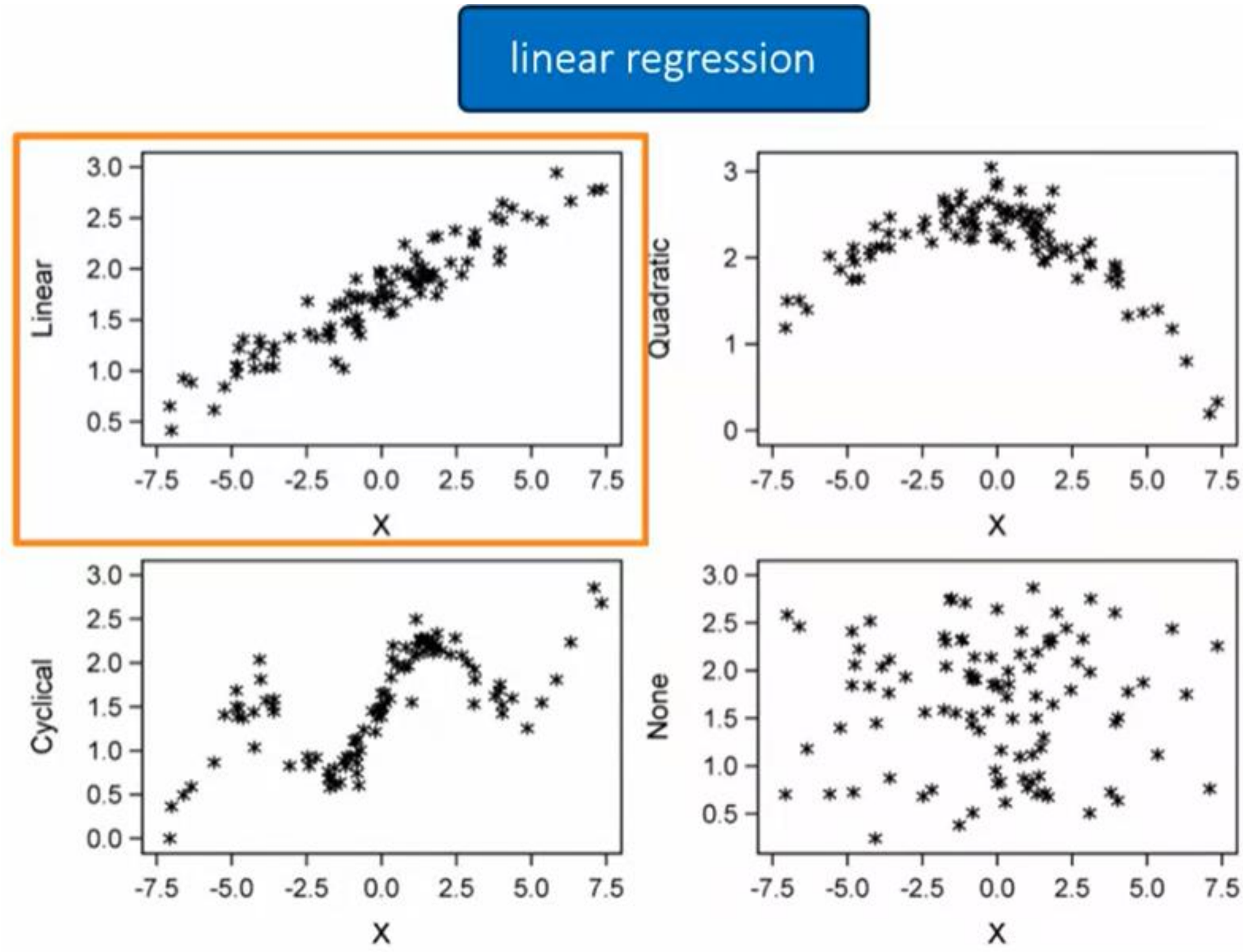

simple
relationships
Sale Price Differences across Central Air
300000
250000
200000
150000
100000
50000
Sale price in dollars
N
Y
Presence of central air conditioning

Identifying Associations in Linear Regression with Scatter Plots

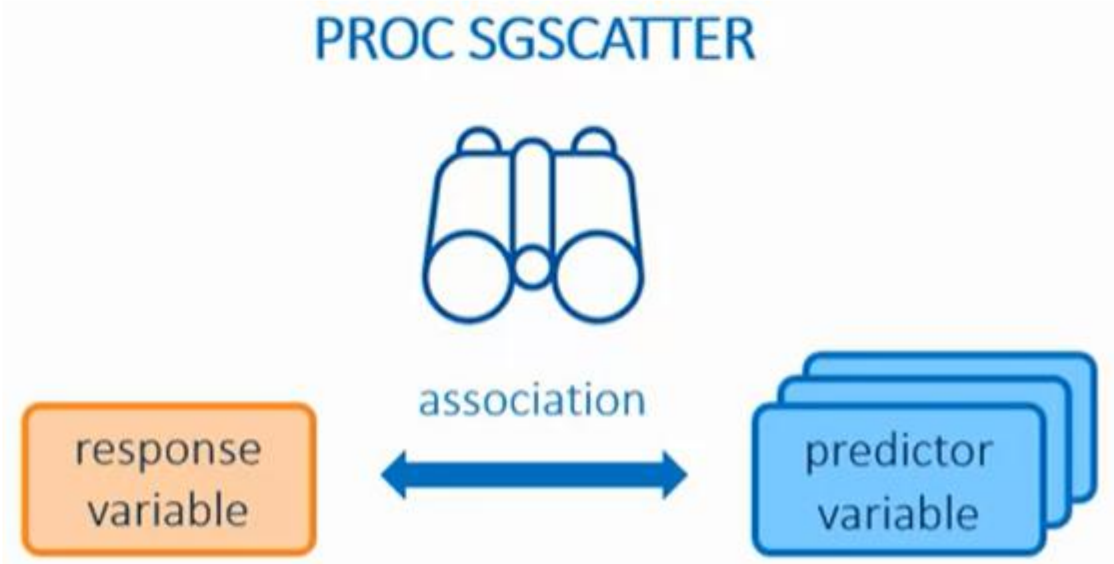


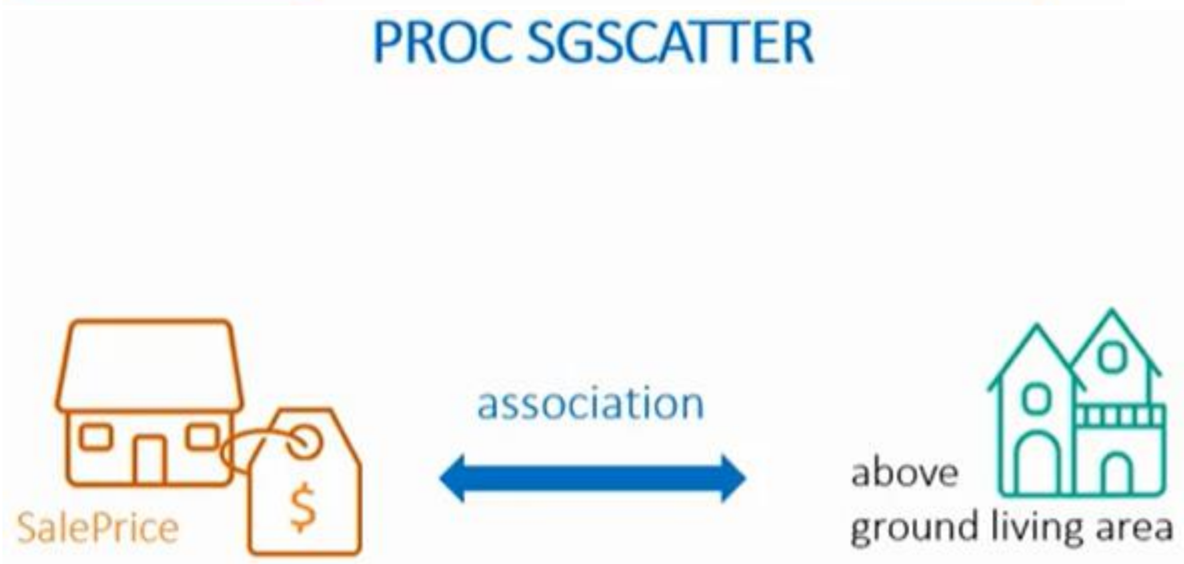

linear regression
scatter plot
continuous
Y
3.0
2.5
2.0
1.5
1.0
0.5
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
X
continuous


linear regression
scatter plot
Y
3.0
2.5
2.0
1.5
1.0
0.5
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
X
outliers
identify trends
range of values
communicate results


linear regression
Linear
Quadratic
Cyclical
None
X

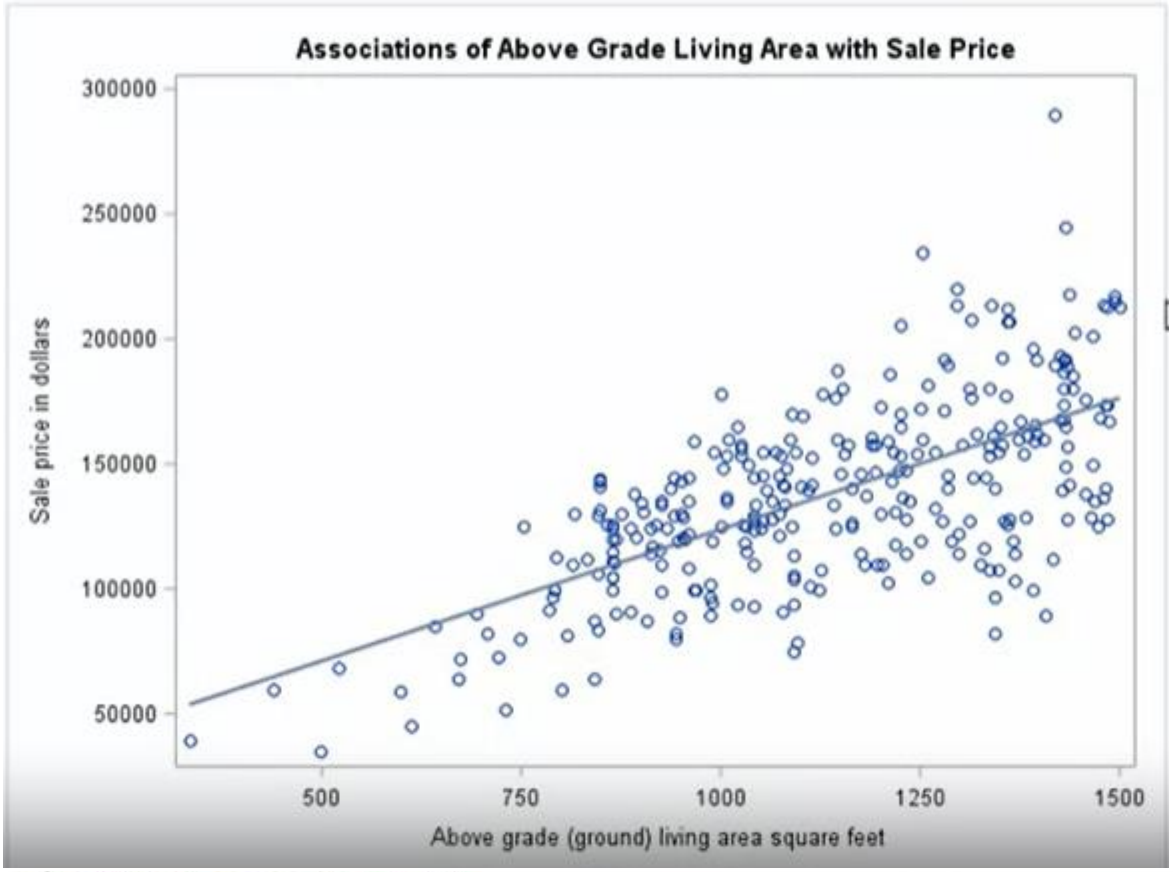
Demo Exploring Associations Using PROC SGSCATTER



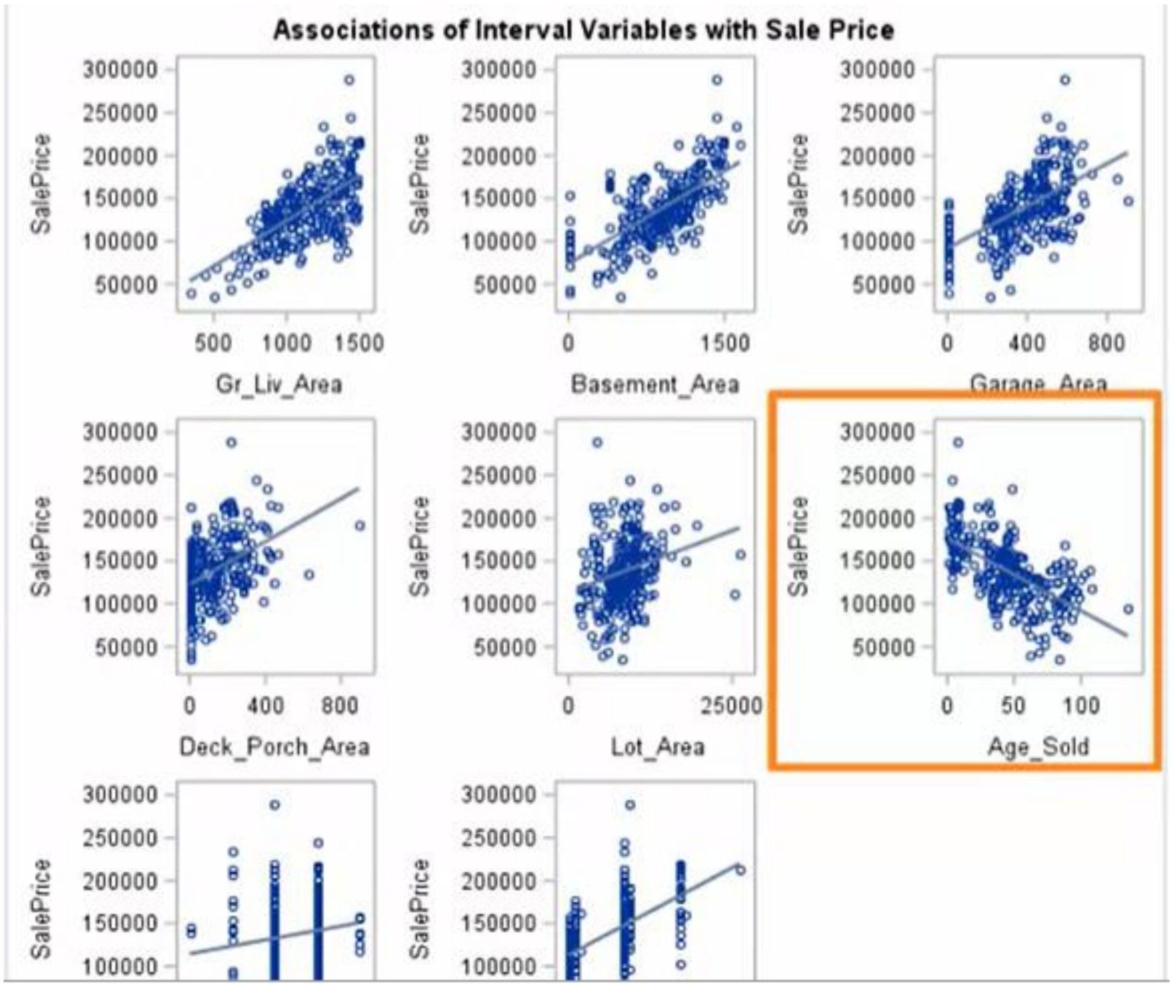



```
/*st102d01.sas*/  /*Part A*/
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*Gr_Liv_Area / reg;
    title "Associations of Above Grade Living Area with Sale Price";
run;

/*st102d01.sas*/  /*Part B*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
```

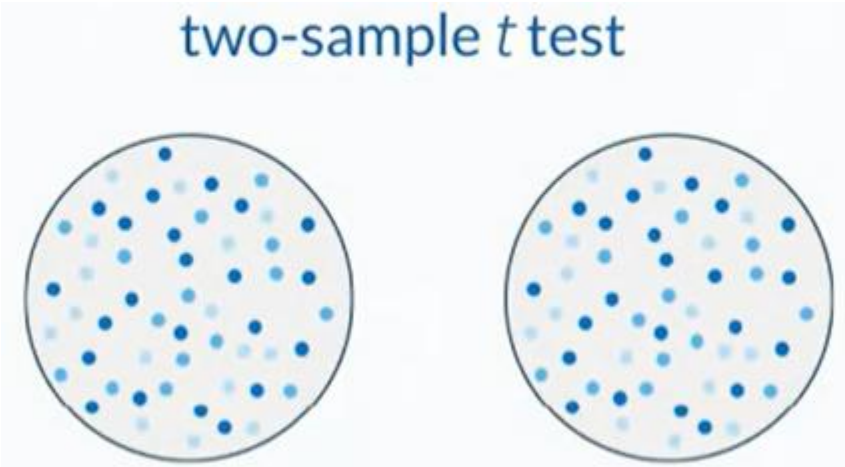
PROC SGSCATTER *<options>*;
    PLOT *plot-request(s) < / options>*;
RUN;

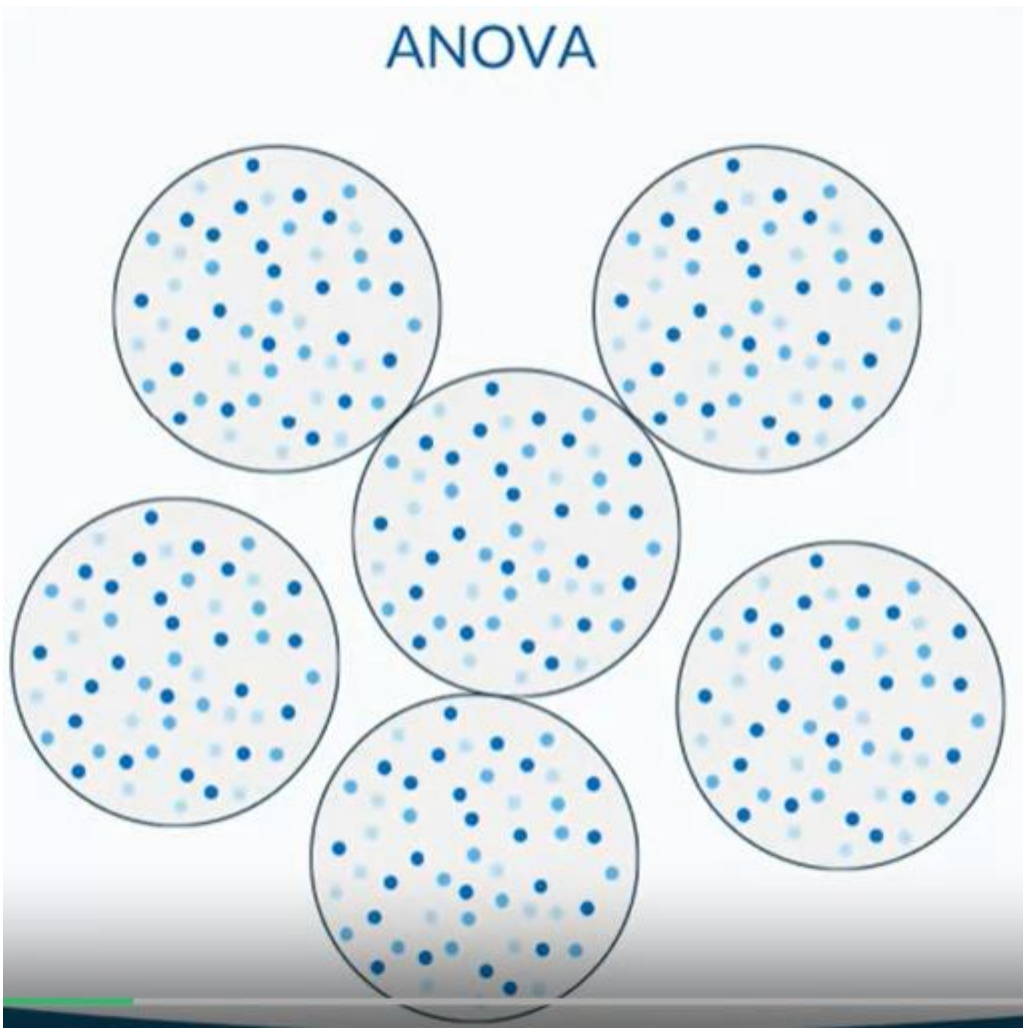



```
/*st102d01.sas*/  /*Part B*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

/*PROC SGSCATTER is used to explore relationships among continuous variables*/
/*using scatter plots*/
options nolabel;
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*(&interval) / reg;
    title "Associations of Interval Variables with Sale Price";
run;
```


Associations of Interval Variables with Sale Price
SalePrice
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Age_Sold

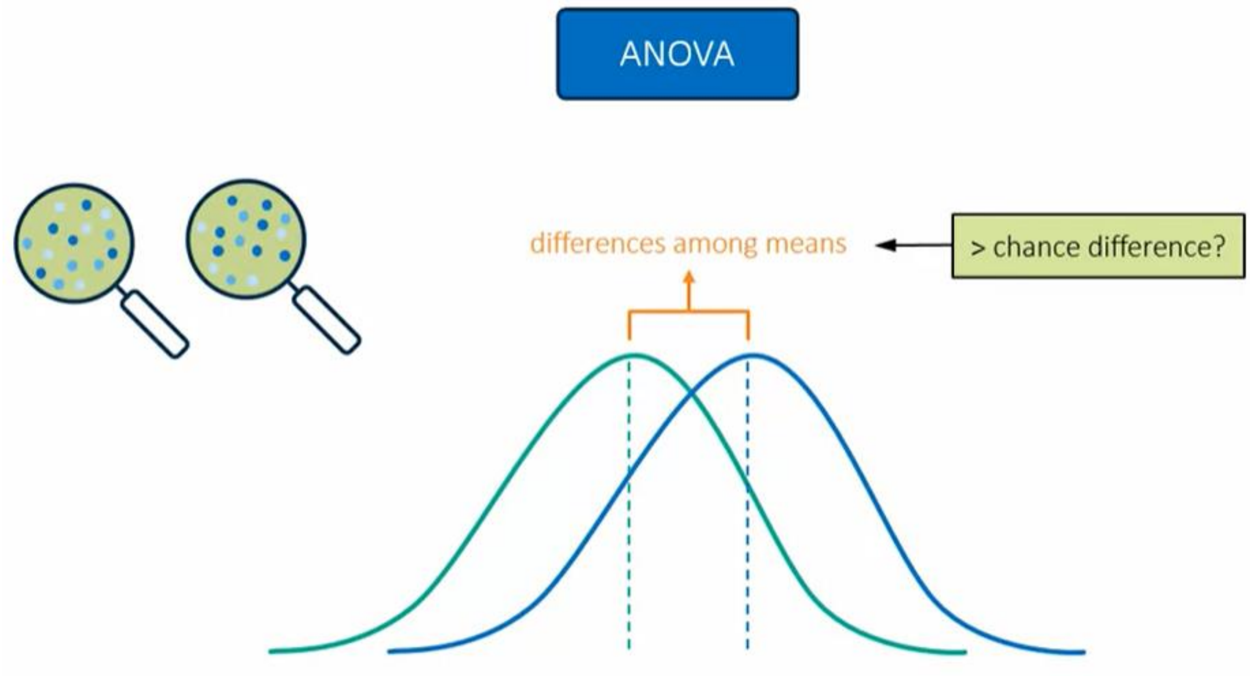
One-Way ANOVA

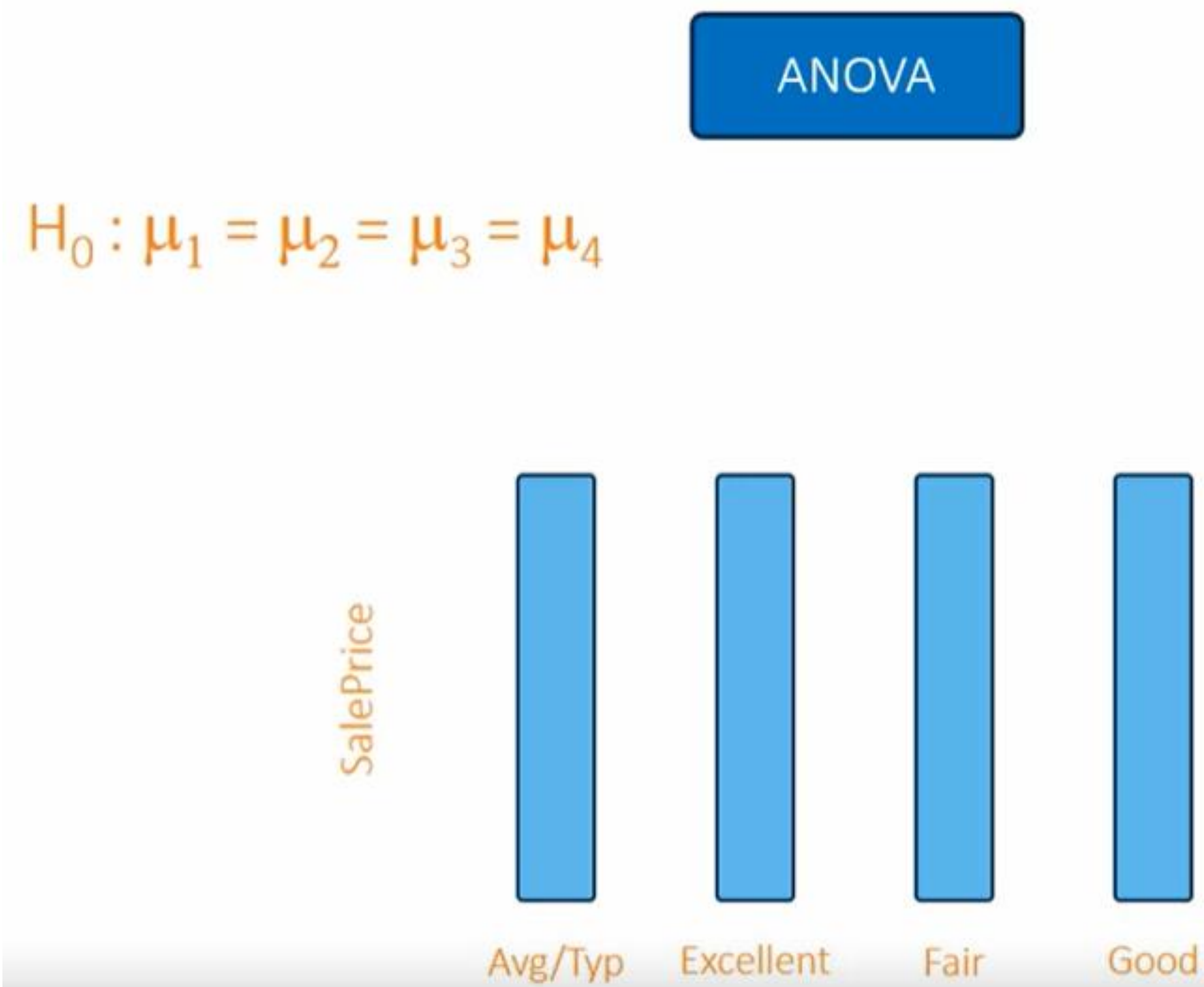


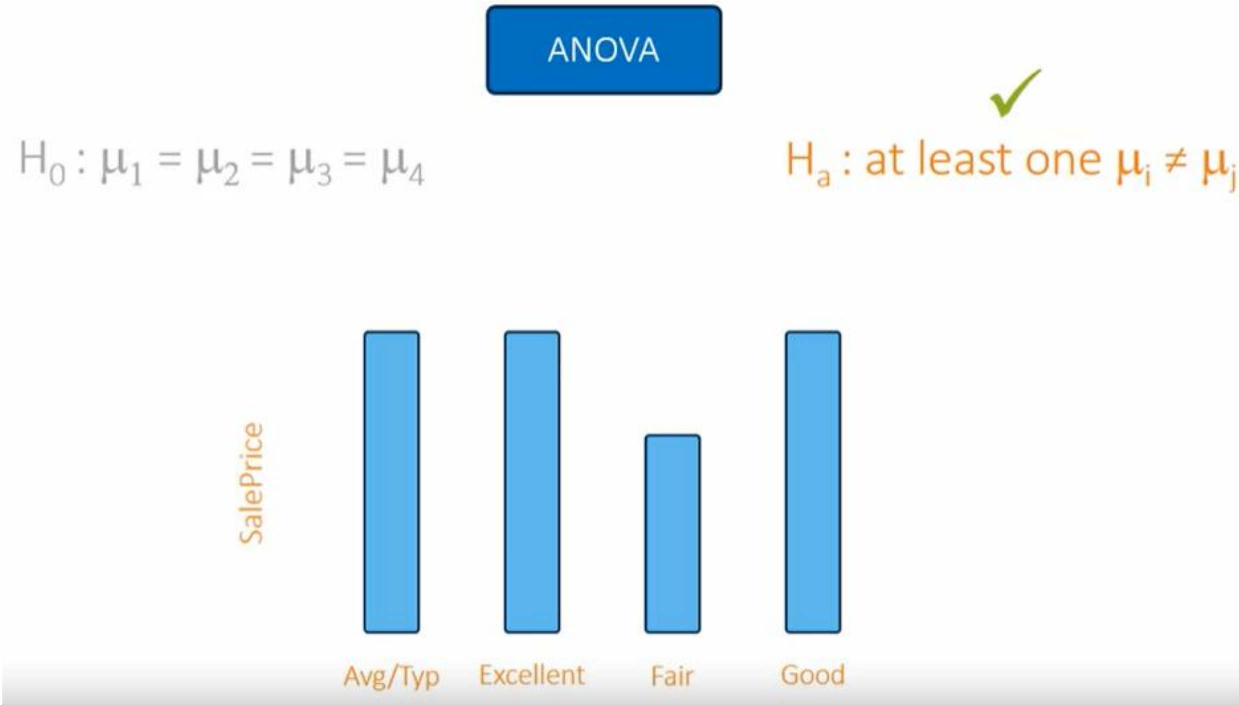



ANOVA
$$
SOLD

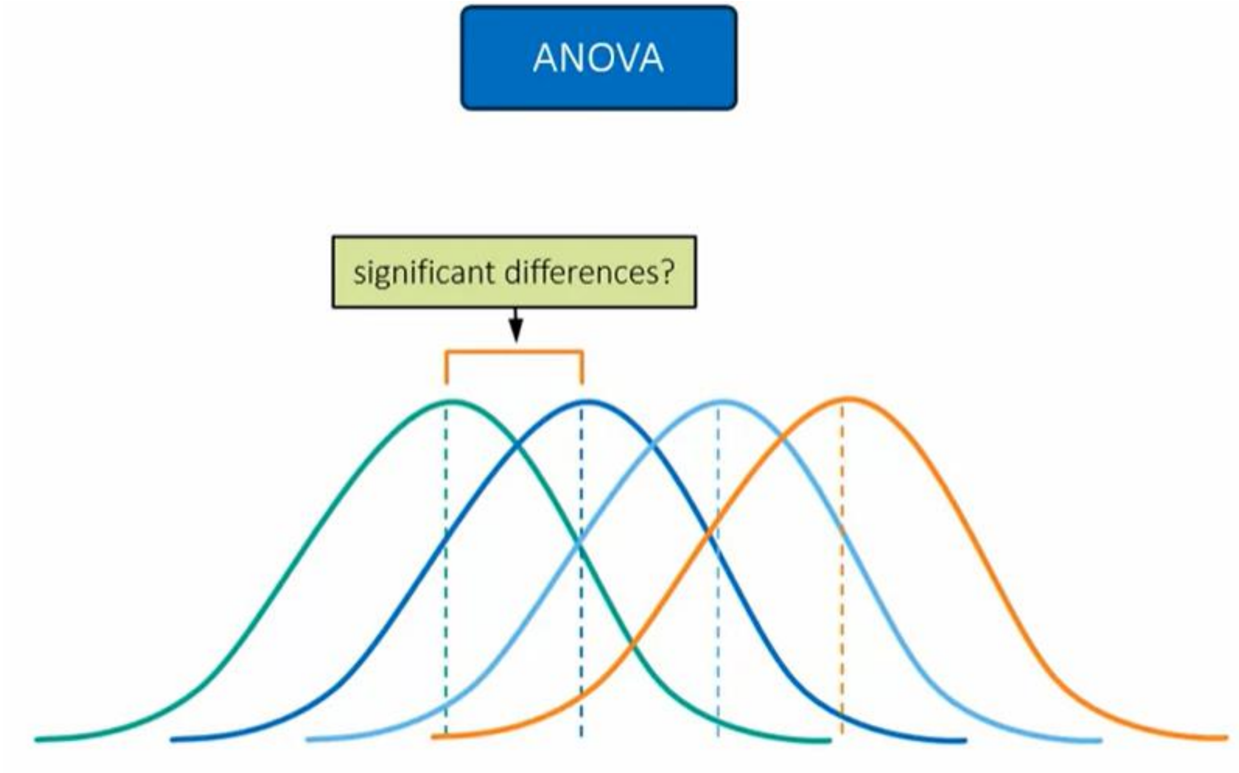
The ANOVA Hypothesis

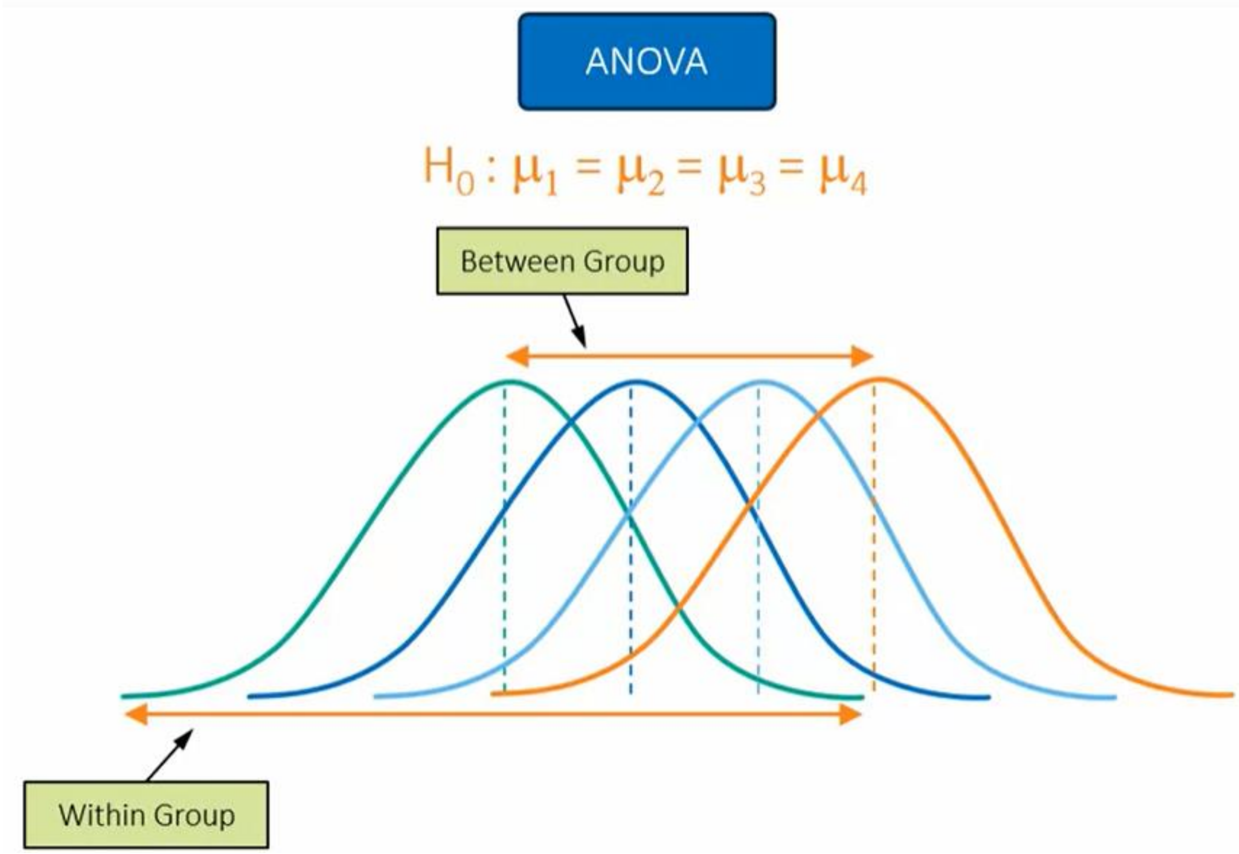

ANOVA
differences among means
> chance difference?

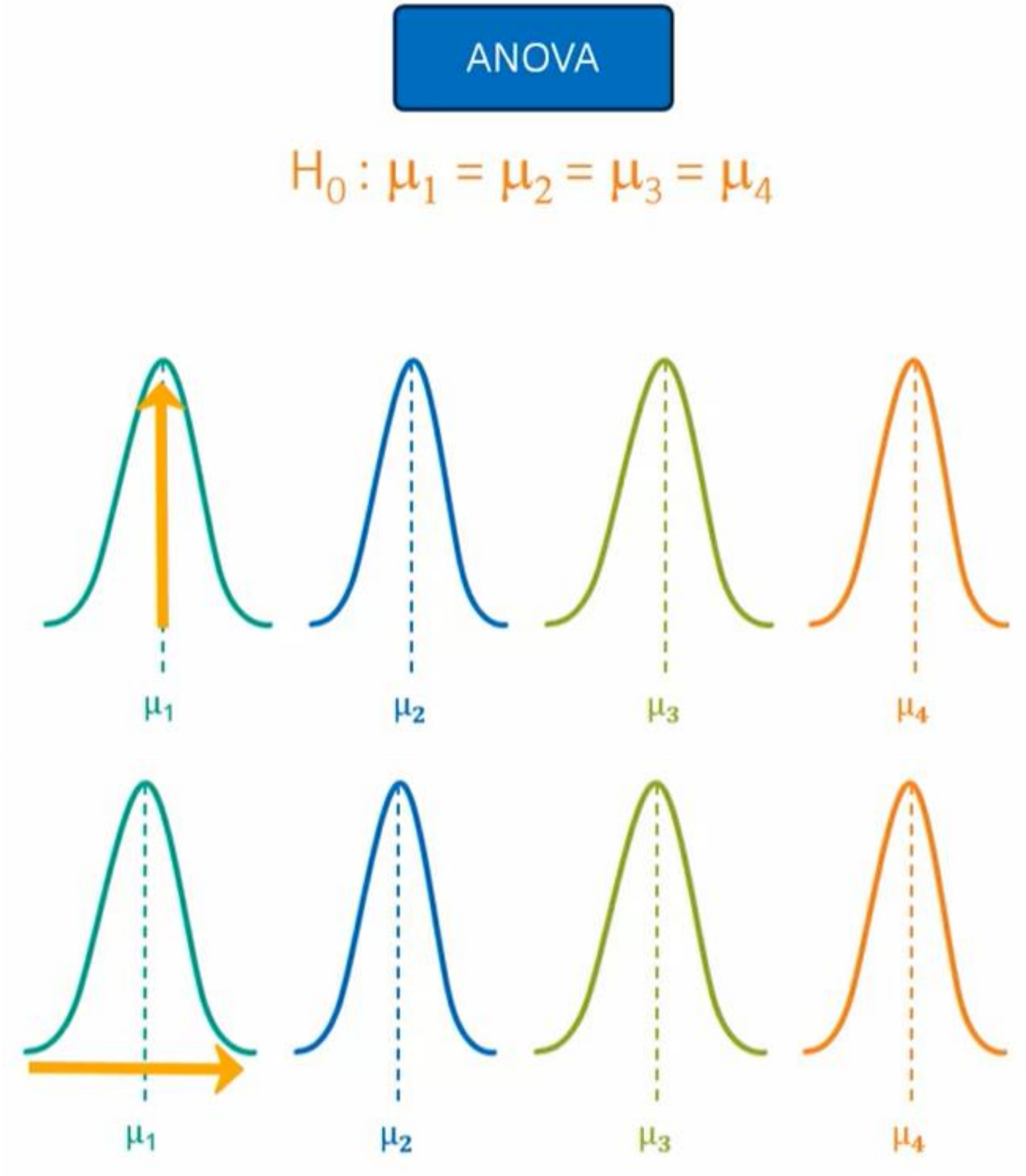

ANOVA
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
SalePrice
Avg/Typ
Excellent
Fair
Good

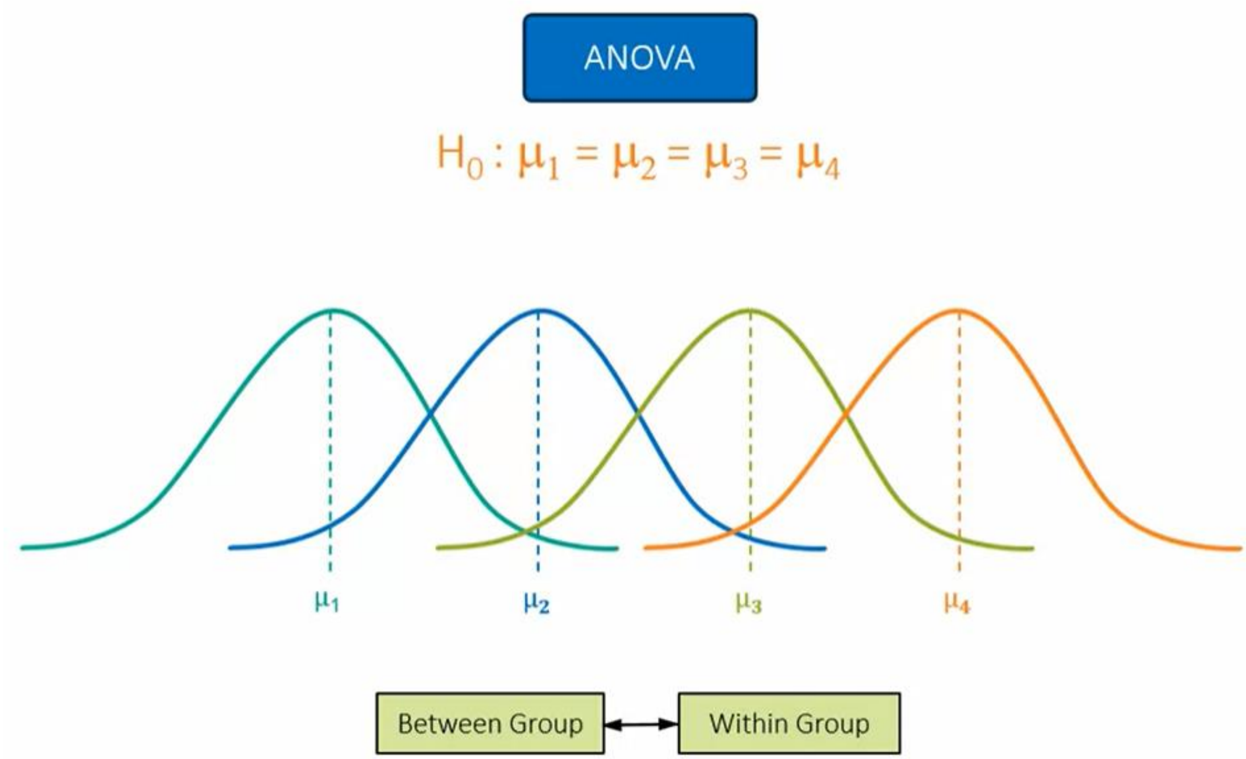


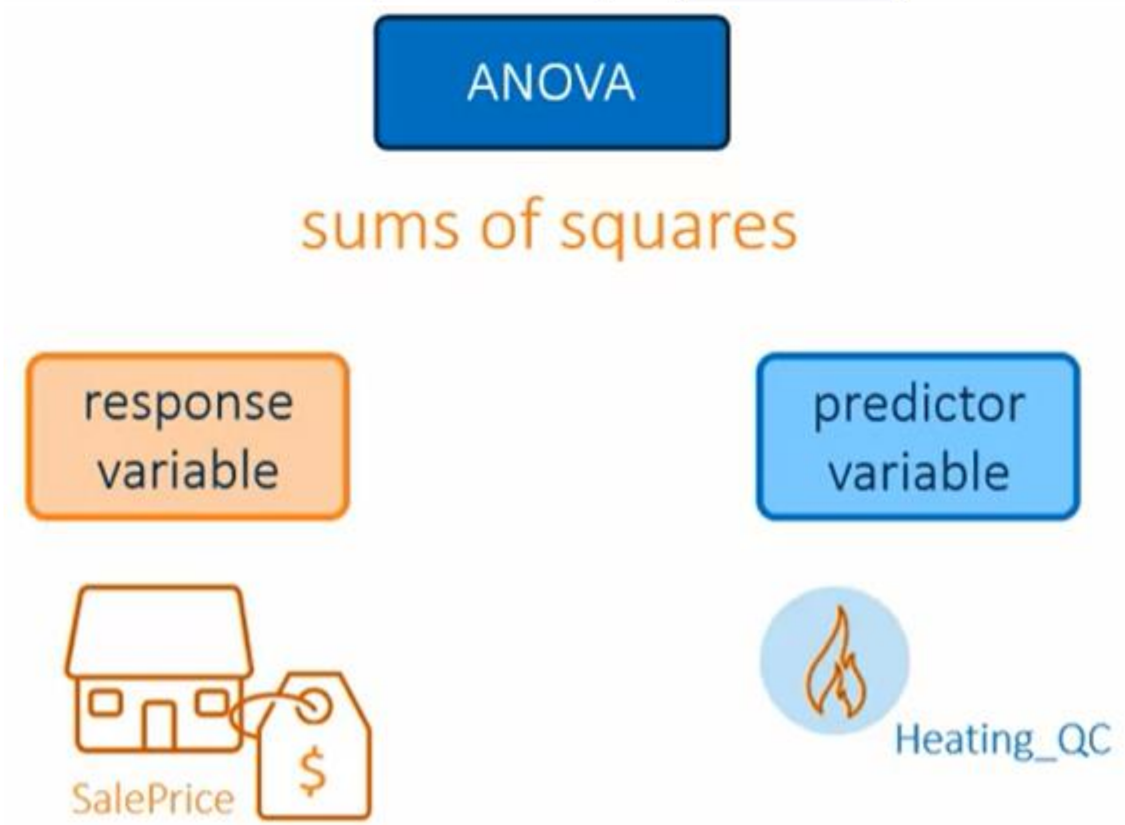
Partitioning Variability in ANOVA


ANOVA
significant differences?


ANOVA
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
Between Group
Within Group

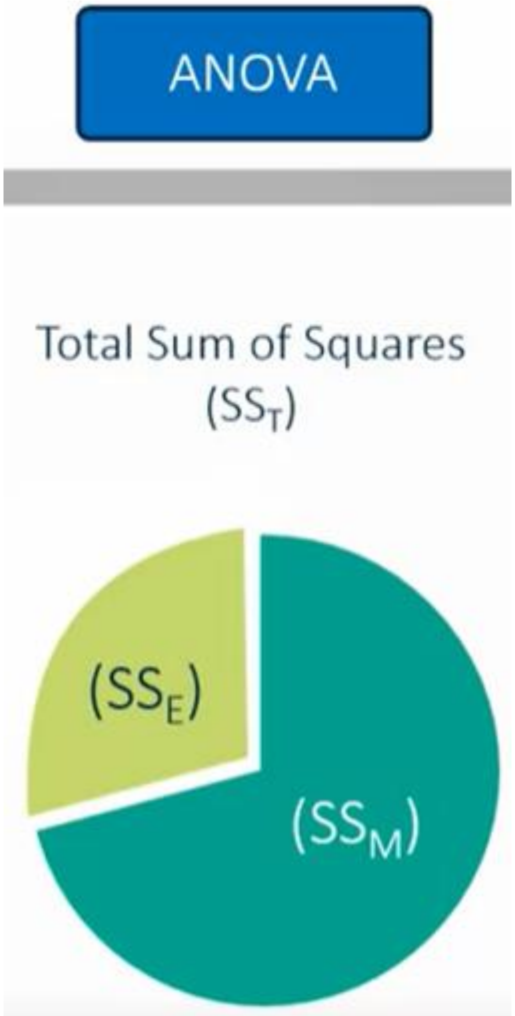

ANOVA
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
μ_1
μ_2
μ_3
μ_4
μ_1
μ_2
μ_3
μ_4


ANOVA
H0 : μ1 = μ2 = μ3 = μ4
μ1
μ2
μ3
μ4
Between Group
Within Group

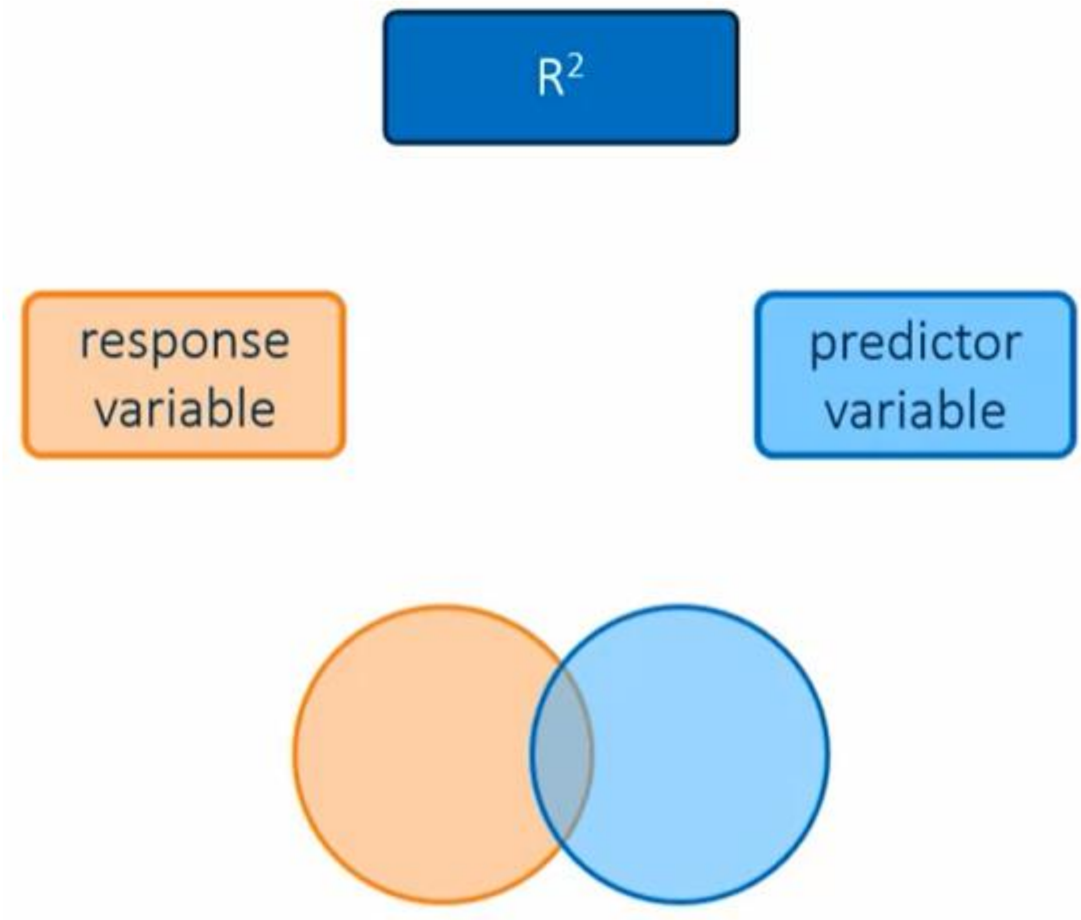

ANOVA
sums of squares
response variable
predictor variable
SalePrice
Heating_QC

ANOVA

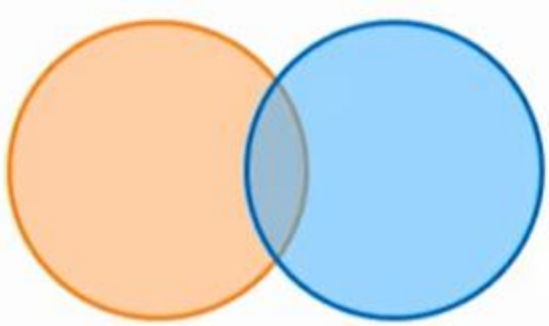
Total Variation	Total Sum of Squares (SS_T)	$\sum\sum\left(Y_{ij}-\bar{\bar{Y}}\right)^2$
Between Group Variation	Model Sum of Squares (SS_M)	$\sum n_i\left(\bar{Y}_i-\bar{\bar{Y}}\right)^2$
Within Group Variation	Error Sum of Squares (SS_E)	$\sum\sum\left(Y_{ij}-\bar{Y}_i\right)^2$



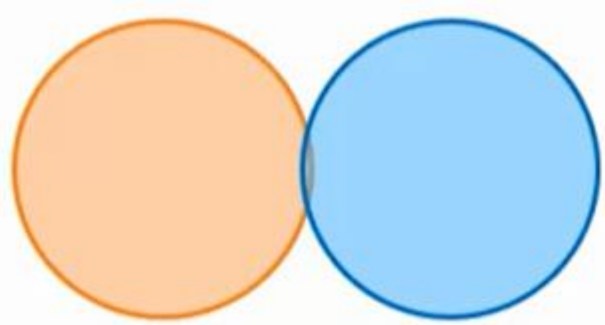
Coefficient of Determination



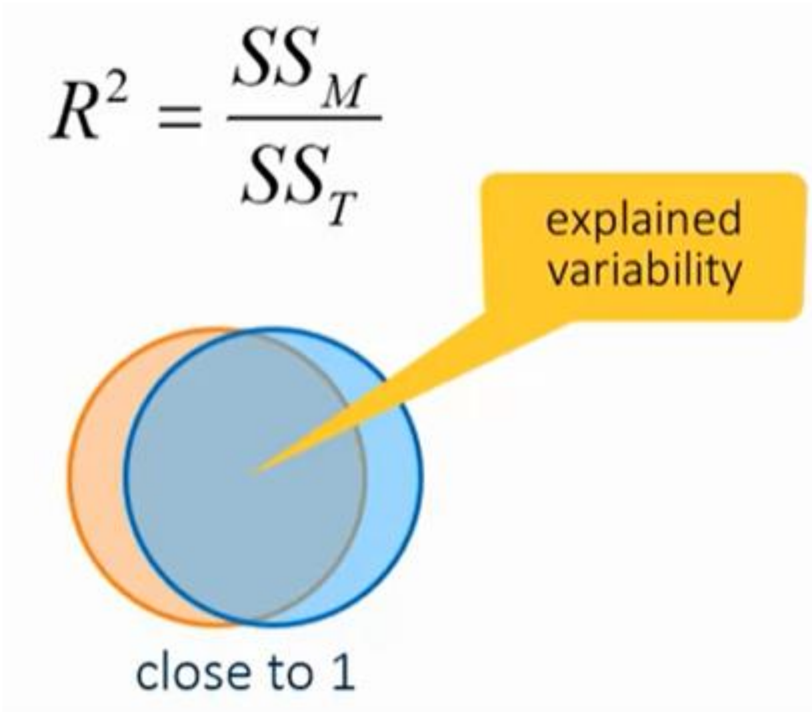
$$R^2 = \frac{SS_M}{SS_T}$$

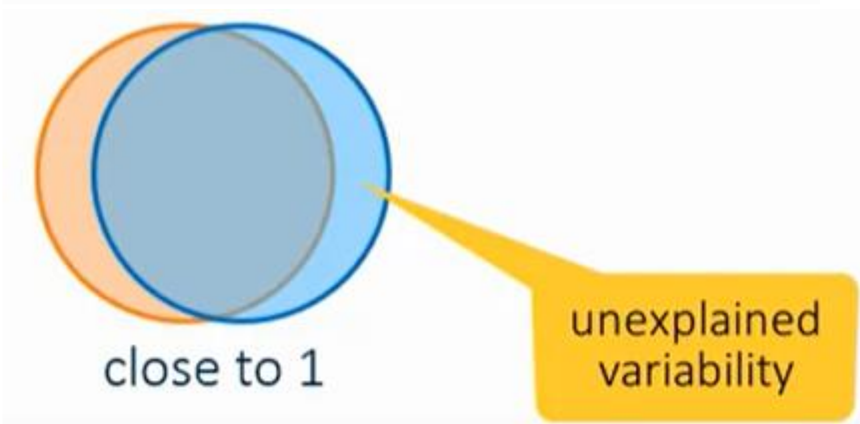


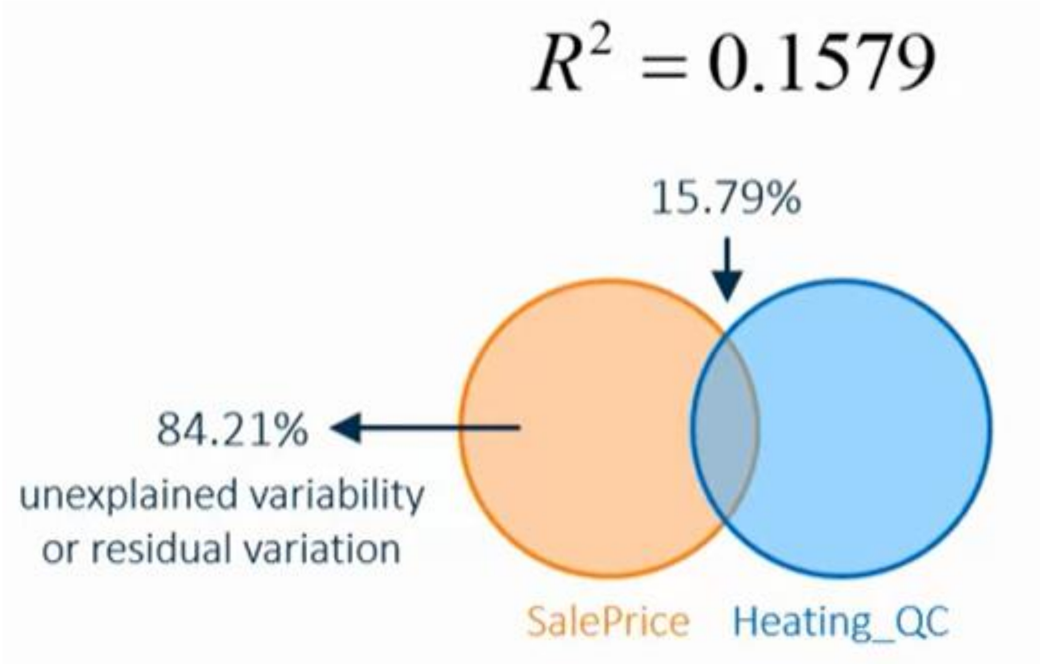
between 0 and 1



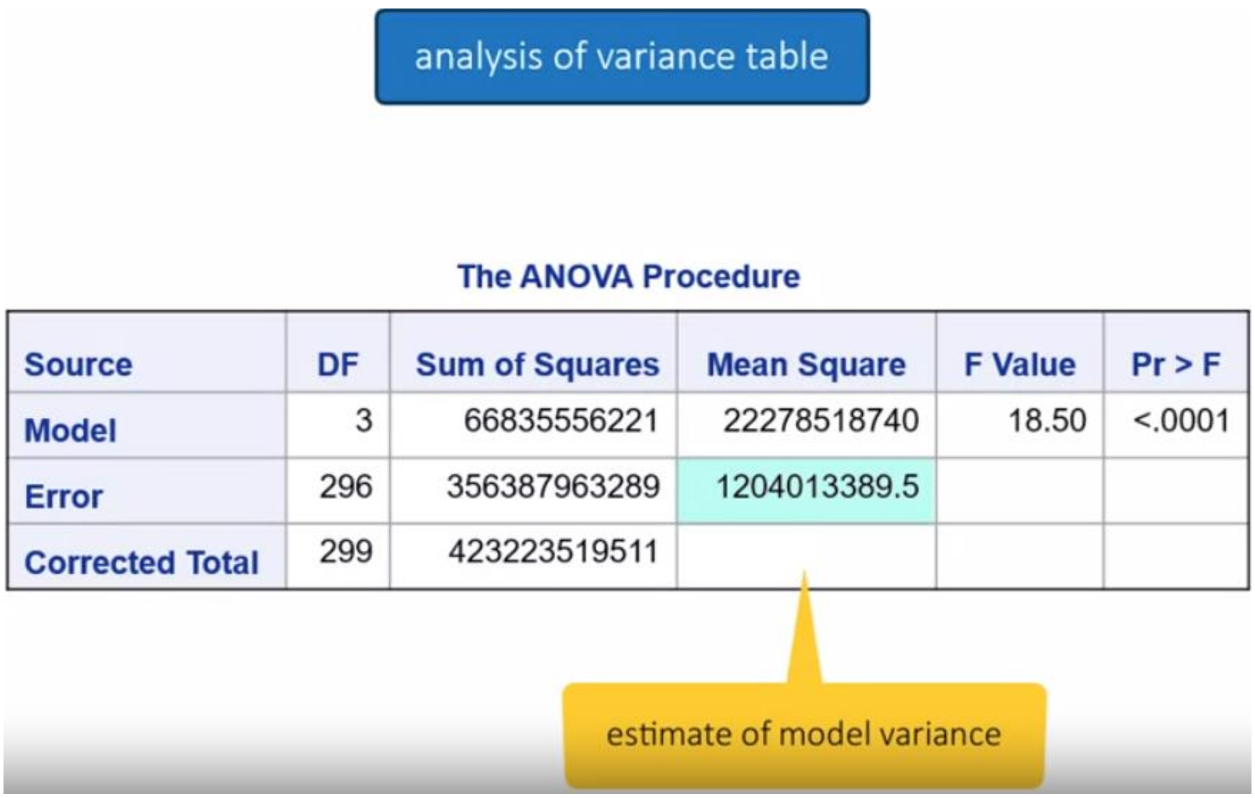
close to 0


$$R^2 = \frac{SS_M}{SS_T}$$
explained variability
close to 1


close to 1
unexplained variability



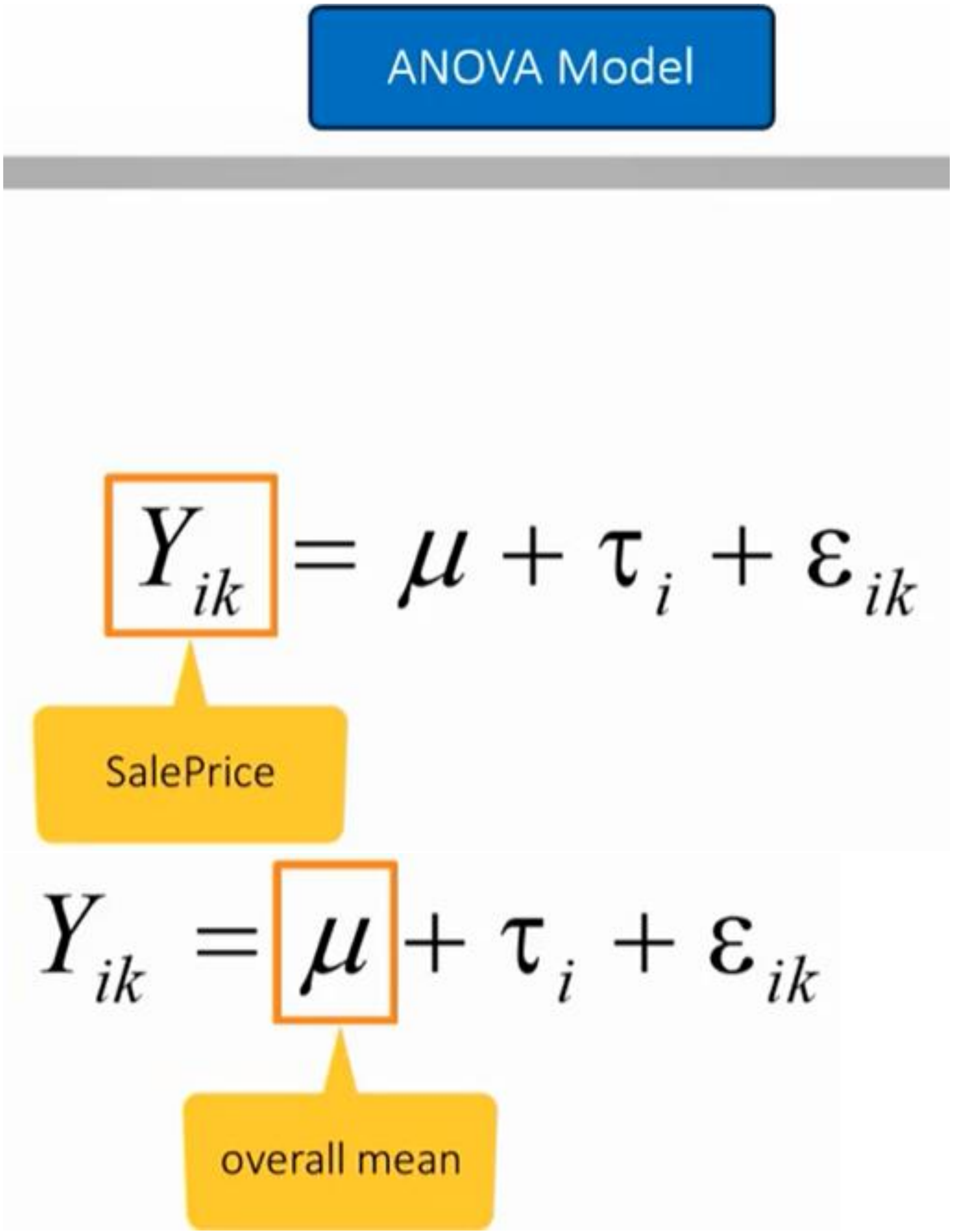
F Statistic and Critical Values



The ANOVA Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

The ANOVA Model



levels 1-4

$$Y_{ik} = \mu + \boxed{\tau_i} + \varepsilon_{ik}$$

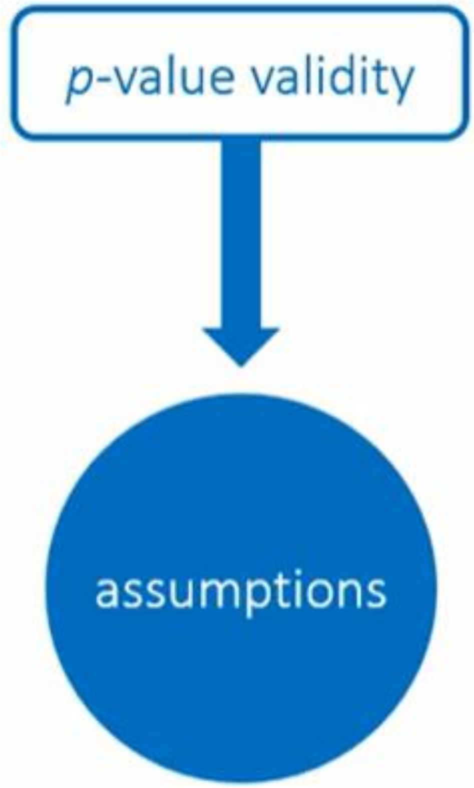
Heating_QC

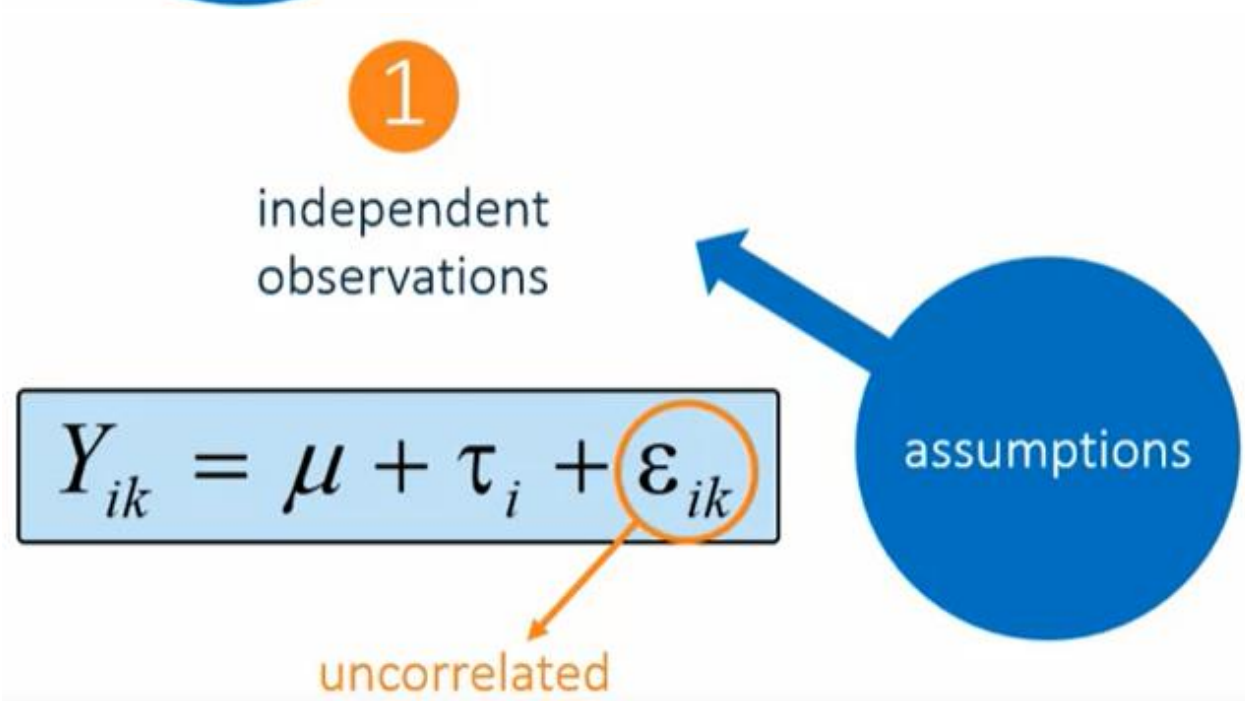
$$Y_{ik} = \mu + \tau_i + \boxed{\varepsilon_{ik}}$$

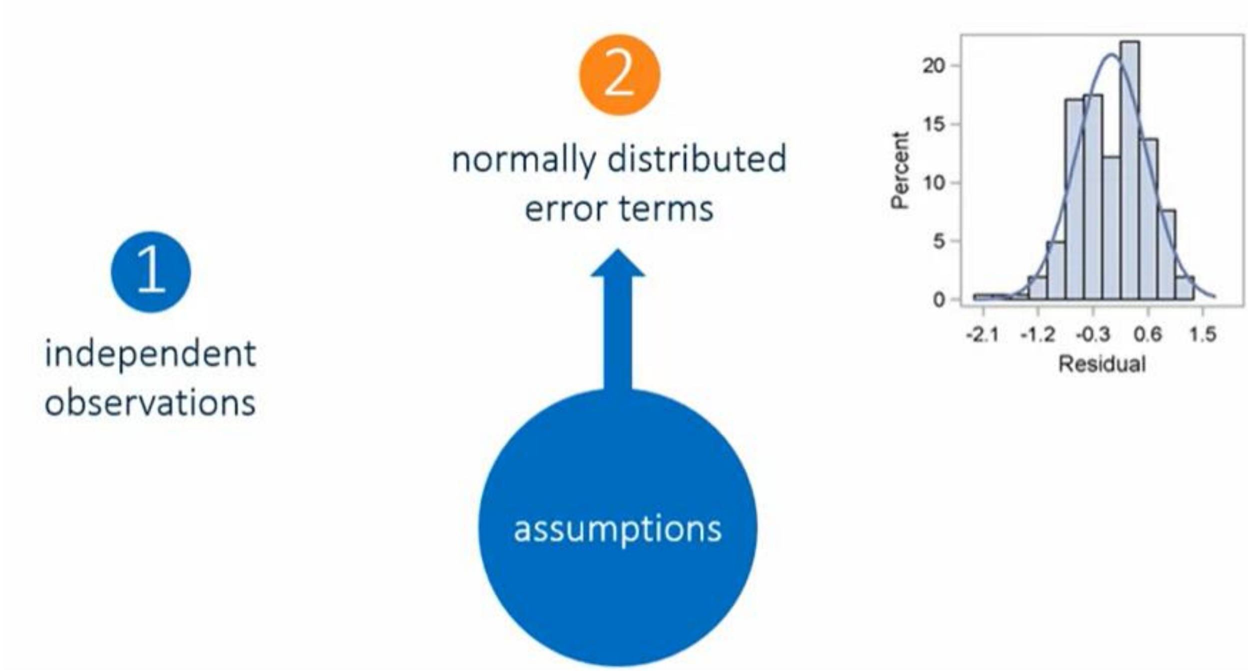
unaccounted for variation

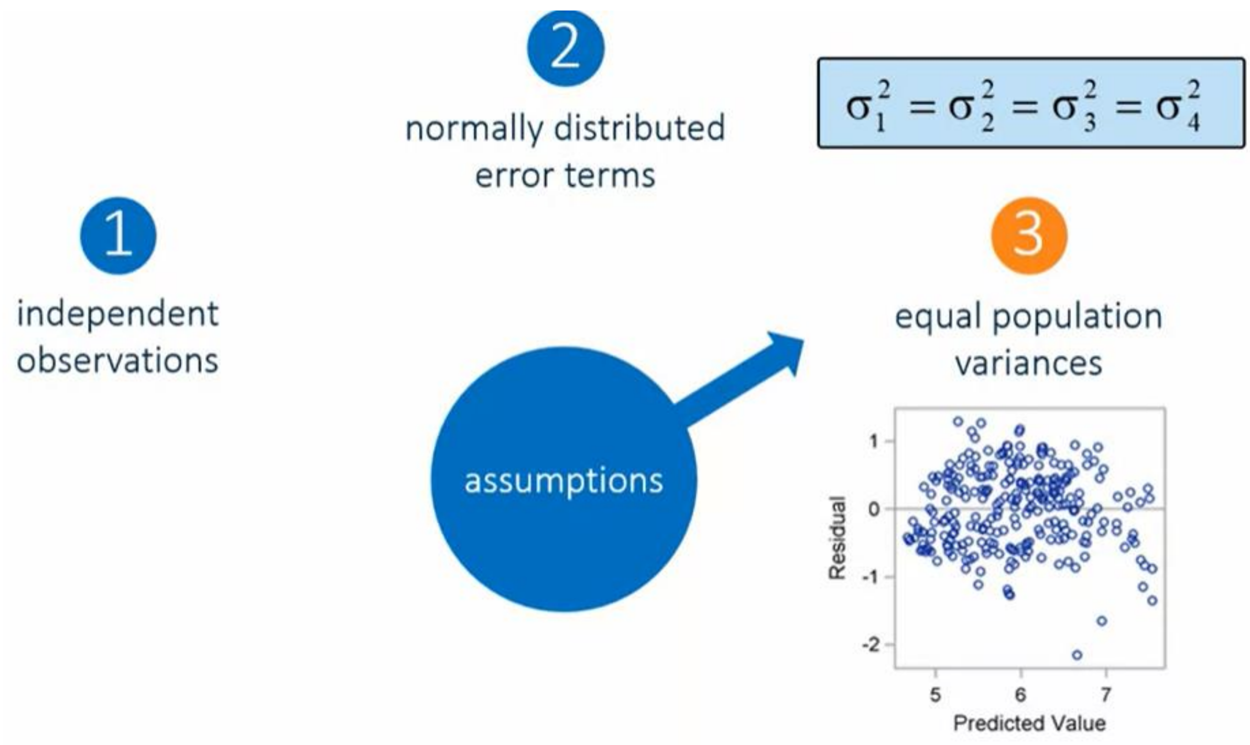
Observed – Predicted = Residual

Observation	Heating_QC	Observed	Predicted	Residual
1	Ex	213500.0000	154919.1869	58580.8131
2	Ex	191500.0000	154919.1869	36580.8131
3	TA	115000.0000	130573.5294	-15573.5294
4	Ex	160000.0000	154919.1869	5080.8131
5	Ex	180000.0000	154919.1869	25080.8131
6	TA	125000.0000	130573.5294	-5573.5294
7	TA	206000.0000	130573.5294	75426.4706
8	Gd	159000.0000	130844.0862	28155.9138
9	TA	180500.0000	130573.5294	49926.4706
10	Gd	142125.0000	130844.0862	11280.9138

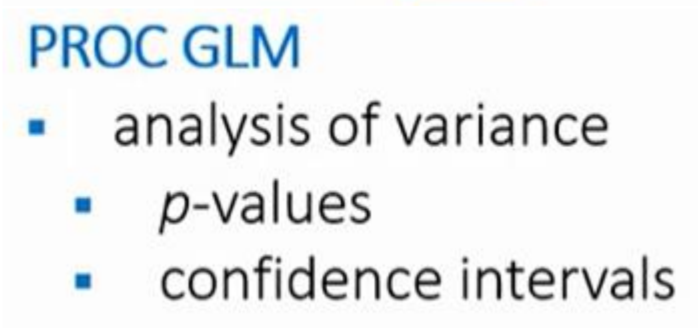

p-value validity
assumptions

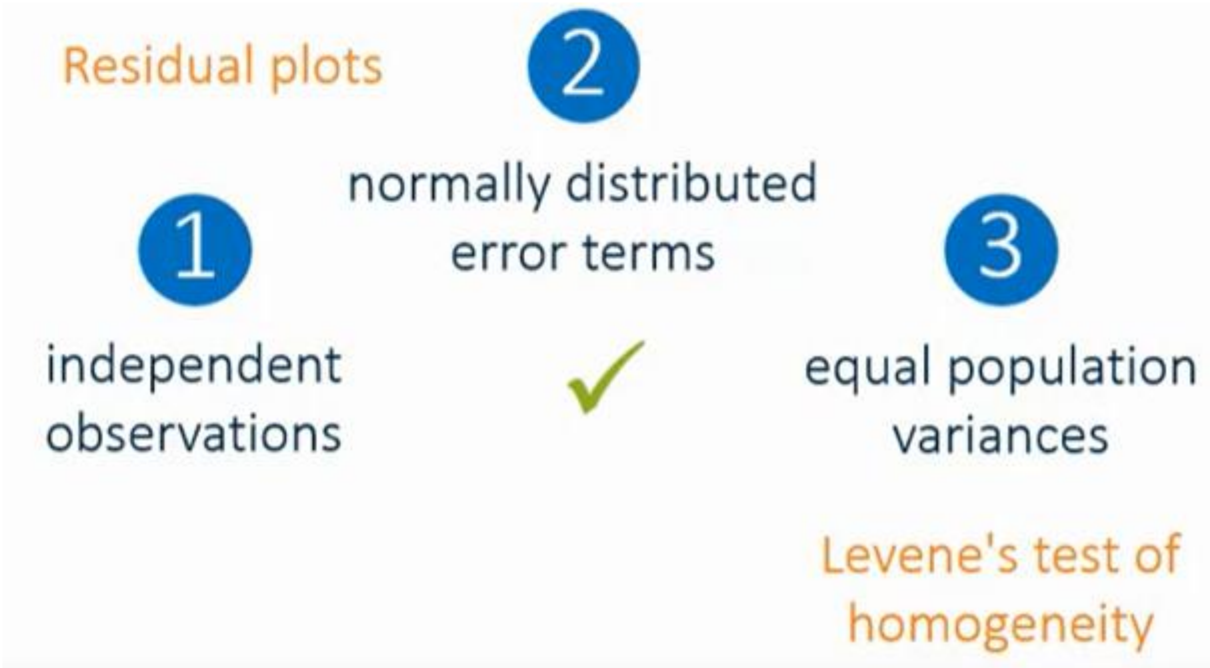

1
independent
observations
$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$
uncorrelated
assumptions


2
normally distributed error terms
1
independent observations
assumptions
Percent
Residual
20
15
10
5
0
-2.1
-1.2
-0.3
0.6
1.5


2
normally distributed error terms
$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
1
independent observations
3
equal population variances
assumptions
Residual
Predicted Value
1
0
-1
-2
5
6
7

Demo Performing a One-Way ANOVA Using PROC GLM






```
/*st102d02.sas*/
ods graphics;

proc glm data=STAT1.ameshousing3 plots=diagnostics;
    class Heating_QC;
    model SalePrice=Heating_QC;
    means Heating_QC / hovtest=levene;
    format Heating_QC $Heating_QC.;
    title "One-Way ANOVA with Heating Quality as Predictor";
run;
quit;

title;
```

```
PROC GLM DATA=SAS-data-set <options>;
    CLASS variable(s);
    MODEL dependent-variables=independent-effects < / options>;
    MEANS effects < / options>;
RUN;
```

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

proportion of variance accounted for by the model

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.157920	25.23100	34698.90	137524.9

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.157920	25.23100	34698.90	137524.9

Root MSE/SalePrice Mean * 100

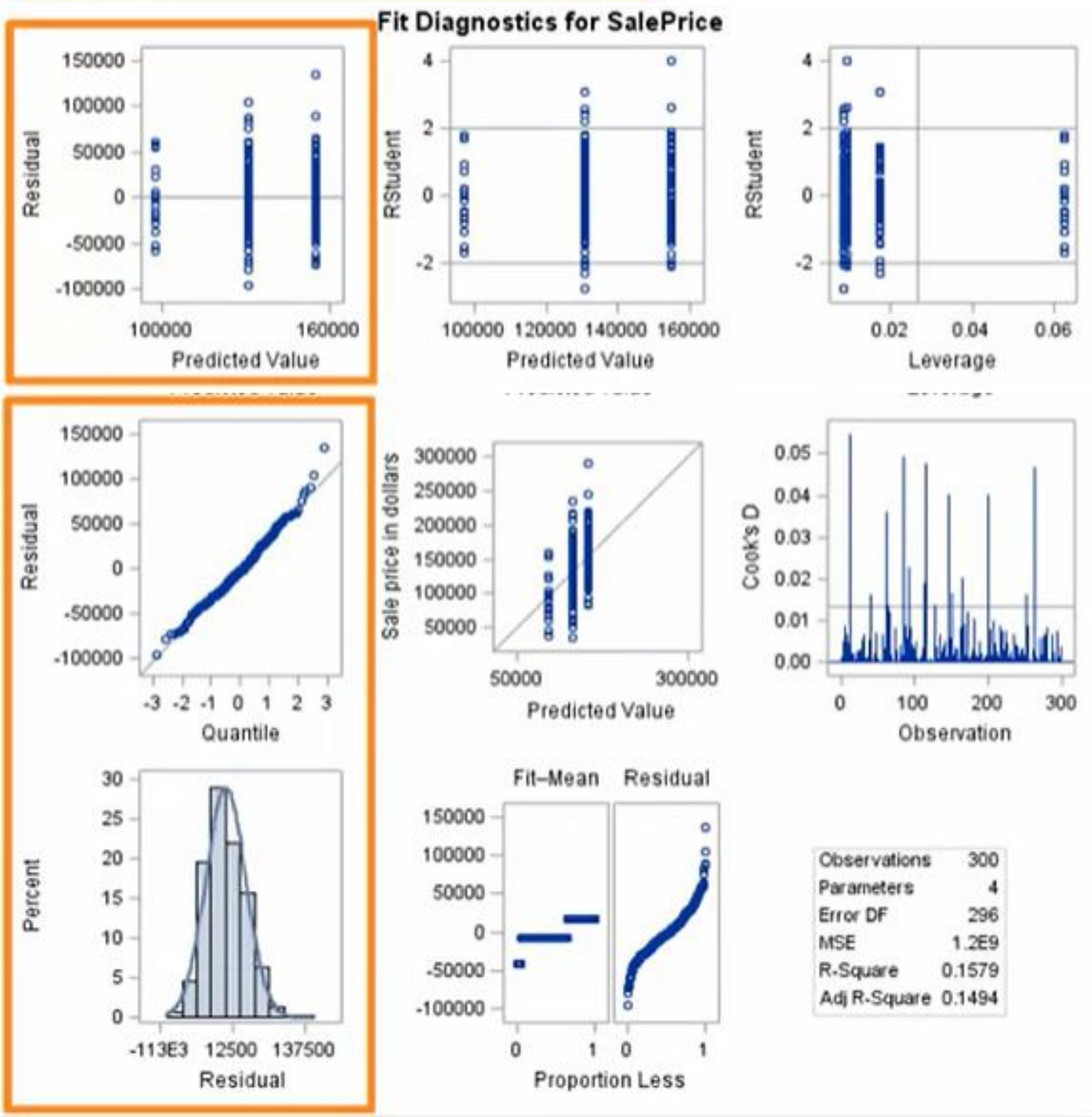
One-Way ANOVA with Heating Quality as Predictor

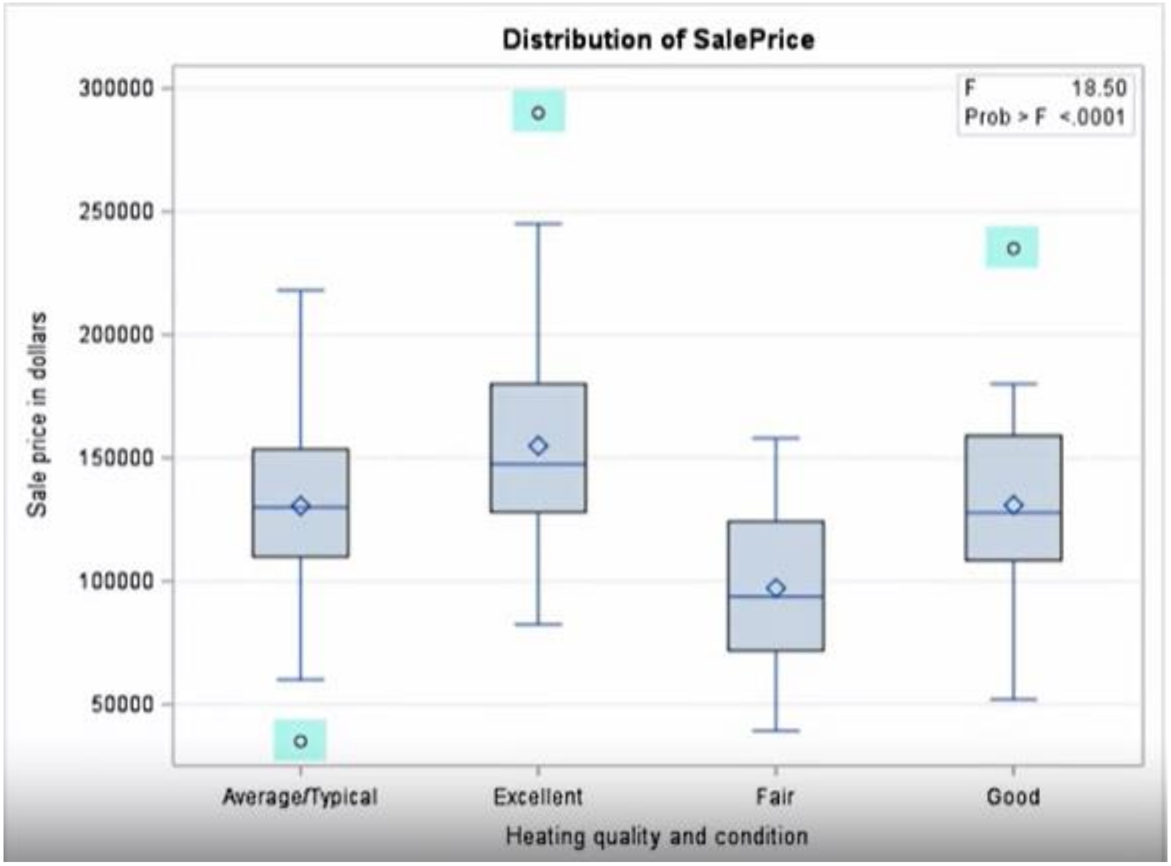
The GLM Procedure
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.157920	25.23100	34698.90	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001



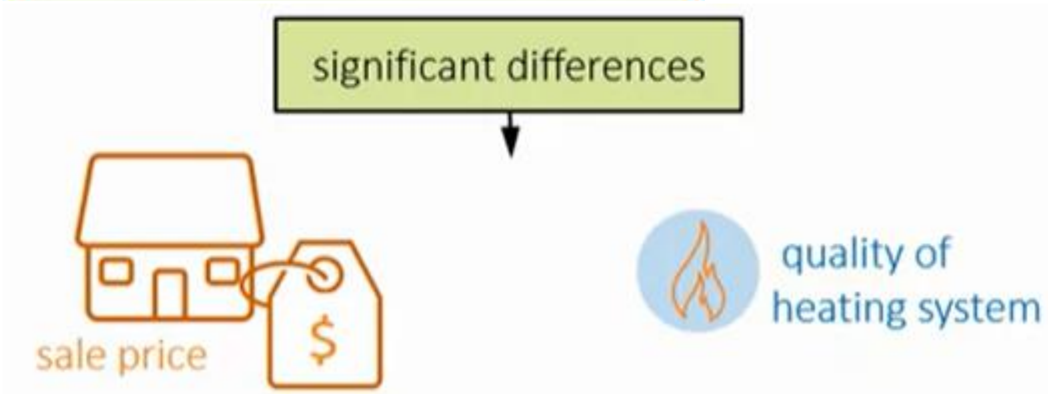

Distribution of SalePrice
F 18.50
Prob > F <.0001
300000
250000
200000
150000
100000
50000
Sale price in dollars
Average/Typical
Excellent
Fair
Good
Heating quality and condition

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

Levene's Test for Homogeneity of SalePrice Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Heating_QC	3	5.931E18	1.977E18	0.58	0.6305
Error	296	1.014E21	3.426E18		




```
/*st102d02.sas*/
ods graphics;


proc glm data=STAT1.ameshousing3 plots=diagnostics;
```

```
   class Heating_QC;
   model SalePrice=Heating_QC;
         means Heating_QC / hovtest=levene;
   format Heating_QC $Heating_QC.;
   title "One-Way ANOVA with Heating Quality as Predictor";
run;
quit;

title;
```

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

Class Level Information		
Class	Levels	Values
Heating_QC	4	Average/Typical Excellent Fair Good

Number of Observations Read	300
Number of Observations Used	300

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

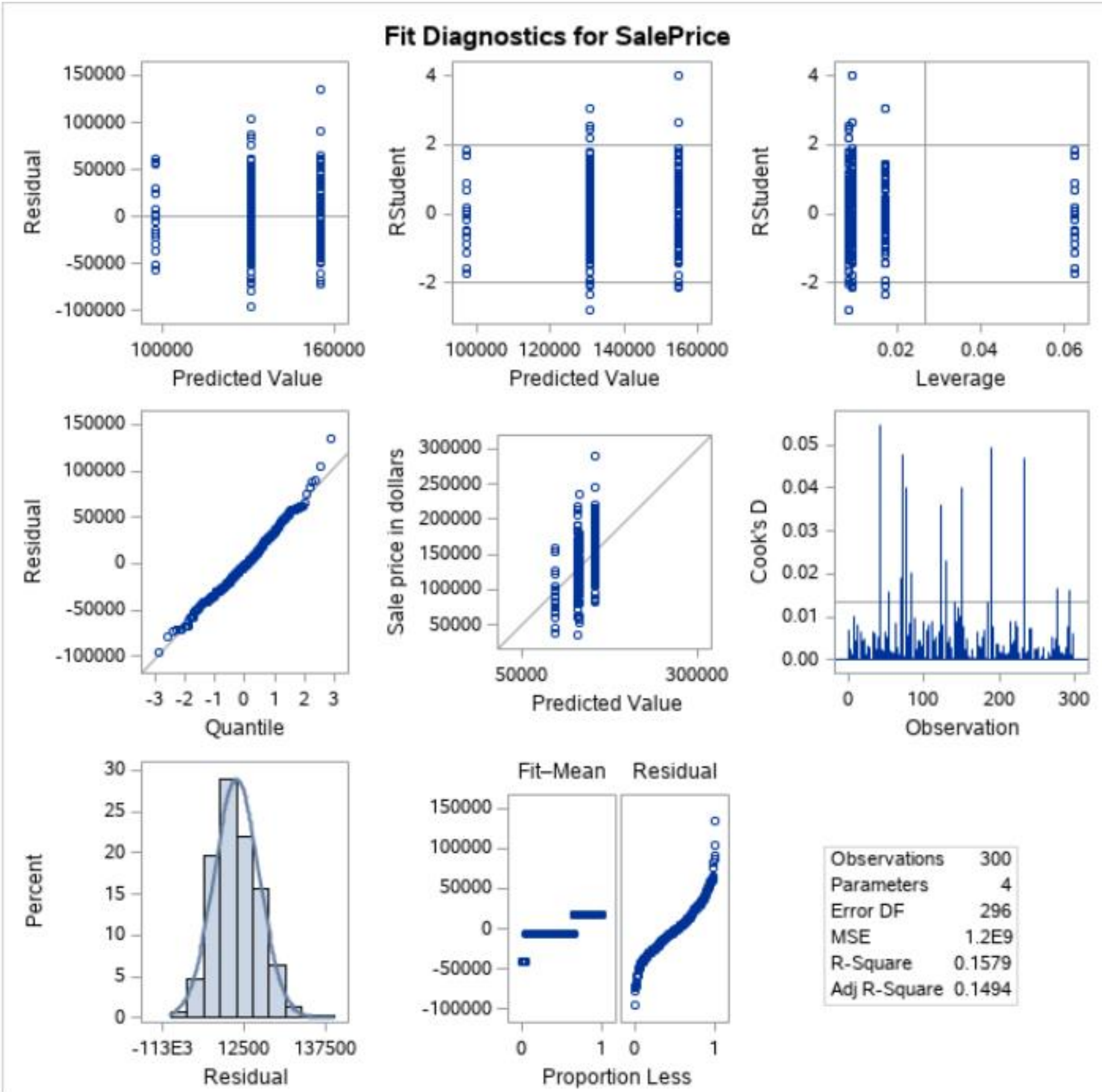
Dependent Variable: SalePrice Sale price in dollars

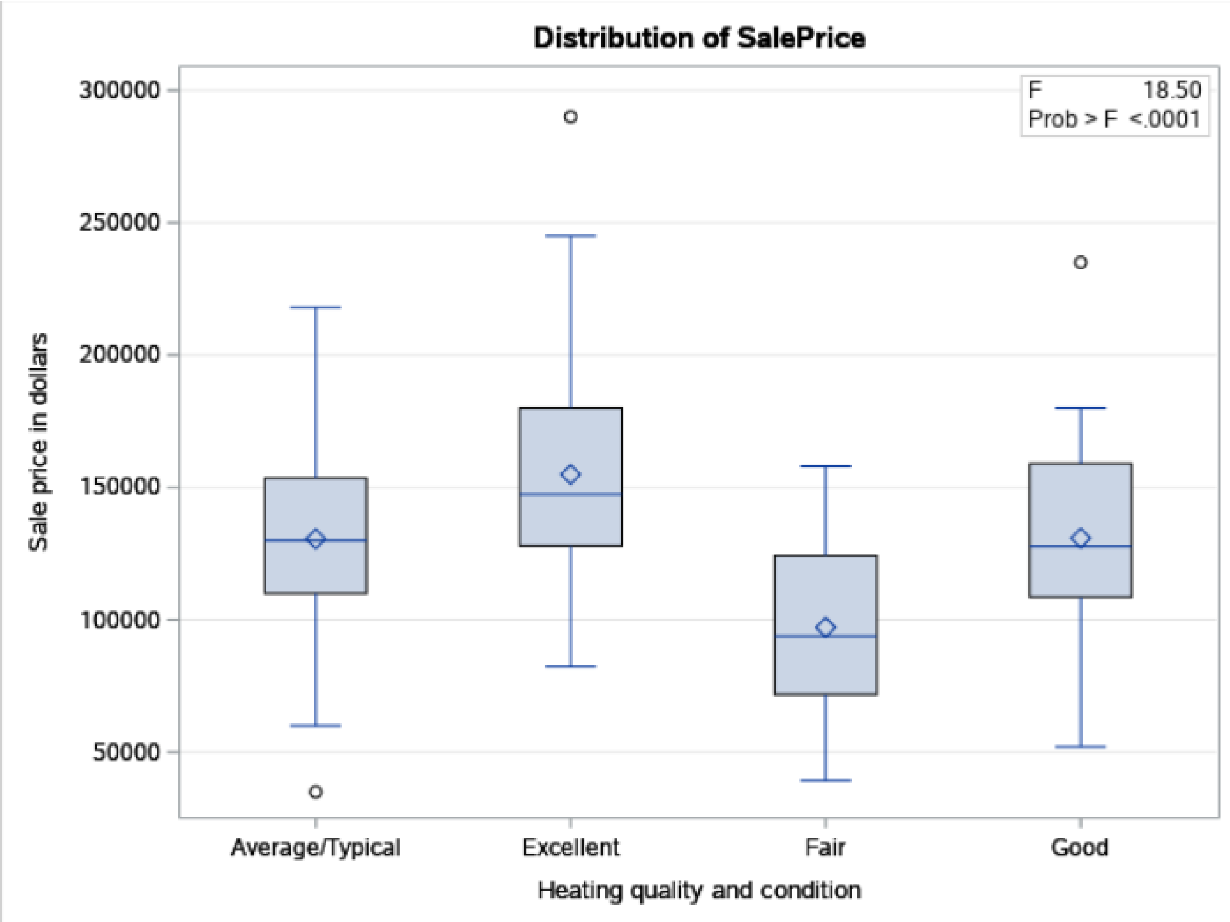
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.157920	25.23100	34698.90	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001


Fit Diagnostics for SalePrice
Residual
RStudent
Predicted Value
Leverage
Quantile
Sale price in dollars
Cook's D
Observation
Percent
Fit–Mean
Residual
Proportion Less
Observations 300
Parameters 4
Error DF 296
MSE 1.2E9
R-Square 0.1579
Adj R-Square 0.1494



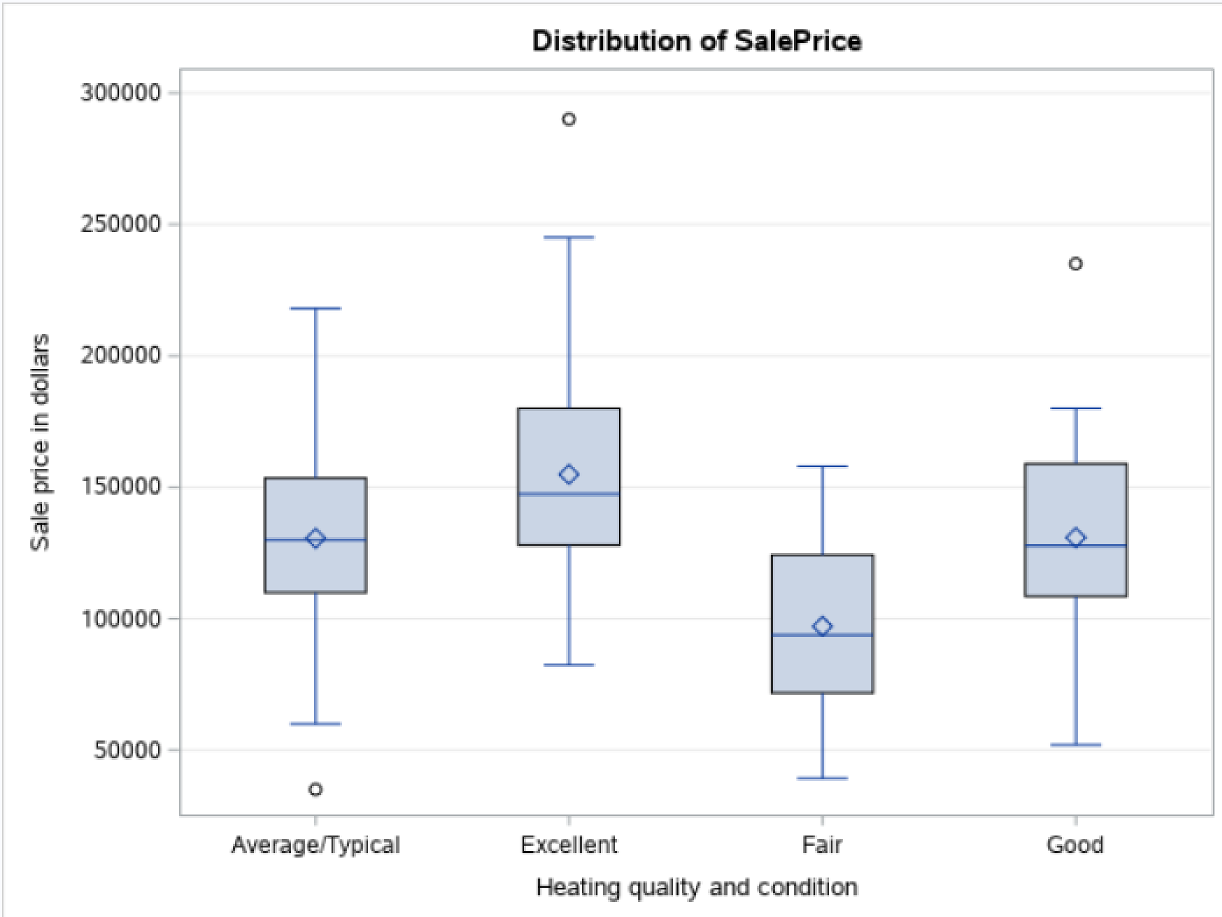
One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

Levene's Test for Homogeneity of SalePrice Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Heating_QC	3	5.931E18	1.977E18	0.58	0.6305
Error	296	1.014E21	3.426E18		

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure



Level of Heating_QC	N	SalePrice	
		Mean	Std Dev
Average/Typical	119	130573.529	32177.4508
Excellent	107	154919.187	36822.8795
Fair	16	97118.750	37423.5437
Good	58	130844.086	34912.5027

Practice - Performing a One-Way ANOVA

TOTAL POINTS 2

1.
Question 1
Montana Gourmet Garlic is a company that uses organic methods to grow garlic. It specializes in hardneck varieties. Knowing a little about experimental methods, the owners design an experiment to test whether growth of the garlic is affected by the type of fertilizer. They limit the experimentation

to a Rocambole variety named Spanish Roja, and test three organic fertilizers and one chemical fertilizer (as a control). They "blind" themselves to the fertilizer by using containers with numbers 1 through 4. (In other words, they design the experiment in such a way that they do not know which fertilizer is in which container.) One acre of farmland is set aside for the experiment. The land is divided into 32 beds, and they randomly assign fertilizers to the beds. At harvest, they calculate the average weight of garlic bulbs in each of the beds. The data are in the **stat1.garlic** data set.

Consider an experiment to study four types of fertilizer, labeled 1, 2, 3, and 4. One fertilizer is chemical and the rest are organic. You want to see whether the average weights of the garlic bulbs are significantly different for plants in beds that use different fertilizers.

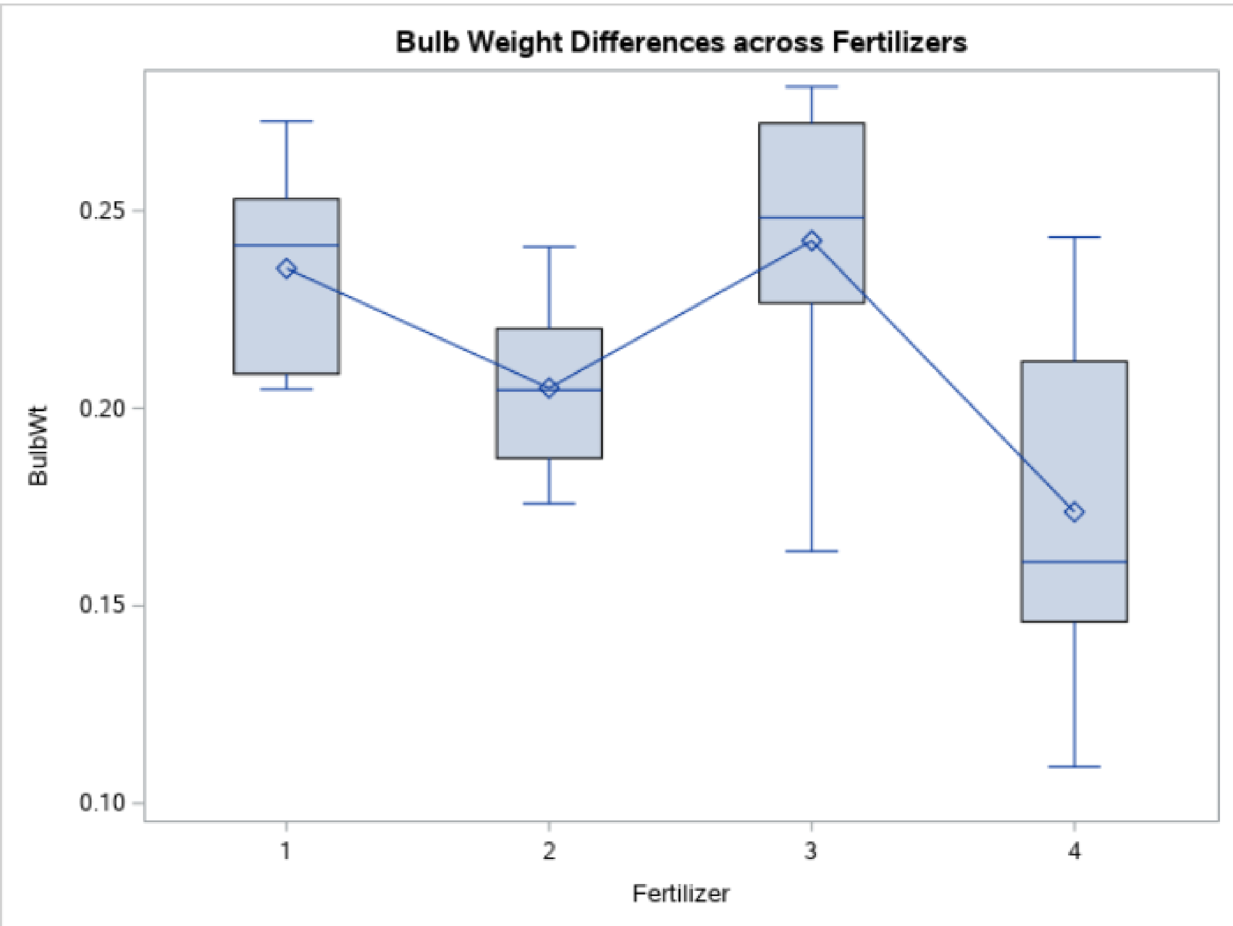
1. Test the hypothesis that the means are equal. Use PROC MEANS to generate descriptive statistics for the four groups, and use PROC SGPLOT to produce box plots of bulb weight for the four groups. Submit the code and view the results.
2. Which fertilizer has the highest mean? **Note**: for your answer, type the fertilizer number (1, 2, 3, or 4).

```
/*st102s01.sas*/  /*Part A*/

proc means data=STAT1.Garlic;

  var BulbWt;

  class Fertilizer;

  title 'Descriptive Statistics of BulbWt by Fertilizer';

run;


proc sgplot data=STAT1.Garlic;

   vbox BulbWt / category=Fertilizer

              connect=mean;

   title "Bulb Weight Differences across Fertilizers";

run;


title;
```

Descriptive Statistics of BulbWt by Fertilizer

The MEANS Procedure

Analysis Variable : BulbWt						
Fertilizer	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	8	8	0.2353998	0.0254092	0.2047856	0.2726395
2	8	8	0.2051141	0.0222098	0.1758361	0.2408676
3	8	8	0.2424075	0.0386855	0.1638284	0.2813780
4	8	8	0.1737649	0.0444702	0.1092144	0.2433058



Perform a one-way ANOVA using PROC GLM. Be sure to check that the assumptions of the analysis method that you choose are met. Submit the code and view the results.

What conclusions can you reach at this point in your analysis?

Fertilizer 3 has the highest mean, 0.2424075, although its mean is quite close to fertilizers 1 and 2.

```
/*st102s01.sas*/  /*Part B*/
ods graphics;

proc glm data=STAT1.Garlic plots=diagnostics;
   class Fertilizer;
   model BulbWt=Fertilizer;
   means Fertilizer / hovtest=levene;
   title "One-Way ANOVA with Fertilizer as Predictor";
run;
quit;

title;
```

One-Way ANOVA with Fertilizer as Predictor

The GLM Procedure

Class Level Information		
Class	Levels	Values
Fertilizer	4	1 2 3 4

Number of Observations Read	32
Number of Observations Used	32

One-Way ANOVA with Fertilizer as Predictor

The GLM Procedure

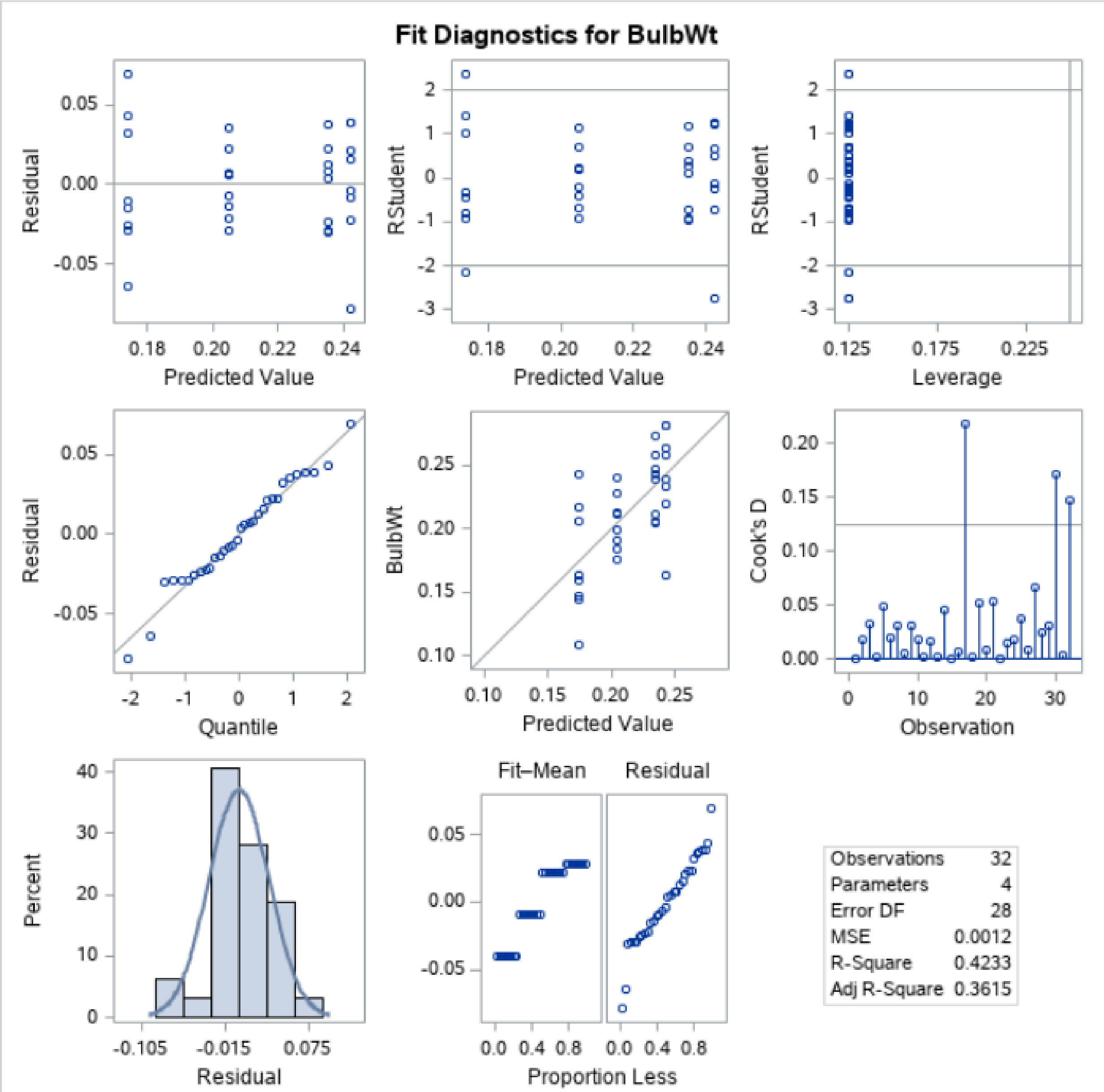
Dependent Variable: BulbWt

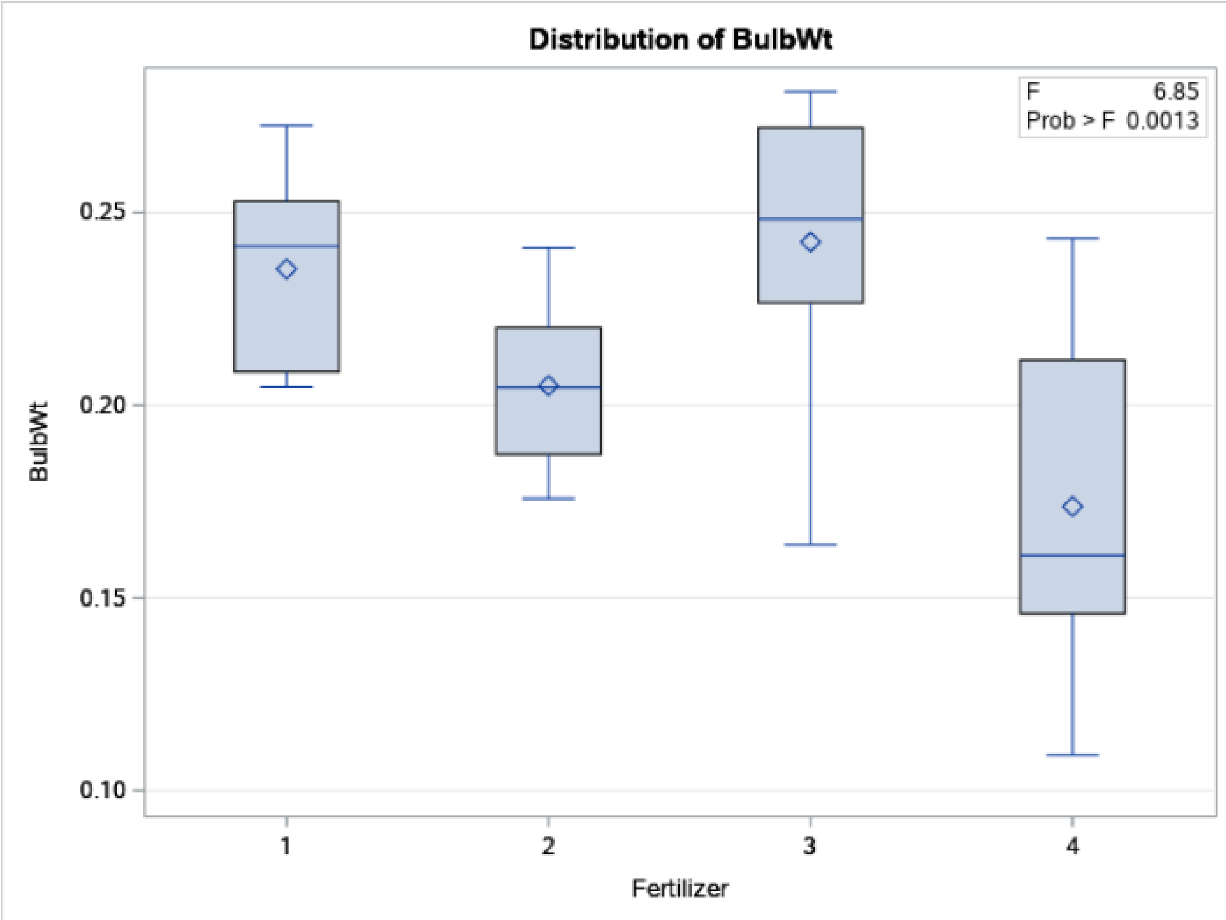
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.02370114	0.00790038	6.85	0.0013
Error	28	0.03229141	0.00115326		
Corrected Total	31	0.05599255			

R-Square	Coeff Var	Root MSE	BulbWt Mean
0.423291	15.85633	0.033960	0.214172

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Fertilizer	3	0.02370114	0.00790038	6.85	0.0013

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fertilizer	3	0.02370114	0.00790038	6.85	0.0013


Fit Diagnostics for BulbWt
Residual
Predicted Value
RStudent
Leverage
Quantile
BulbWt
Cook's D
Observation
Percent
Fit–Mean
Proportion Less
Observations 32
Parameters 4
Error DF 28
MSE 0.0012
R-Square 0.4233
Adj R-Square 0.3615



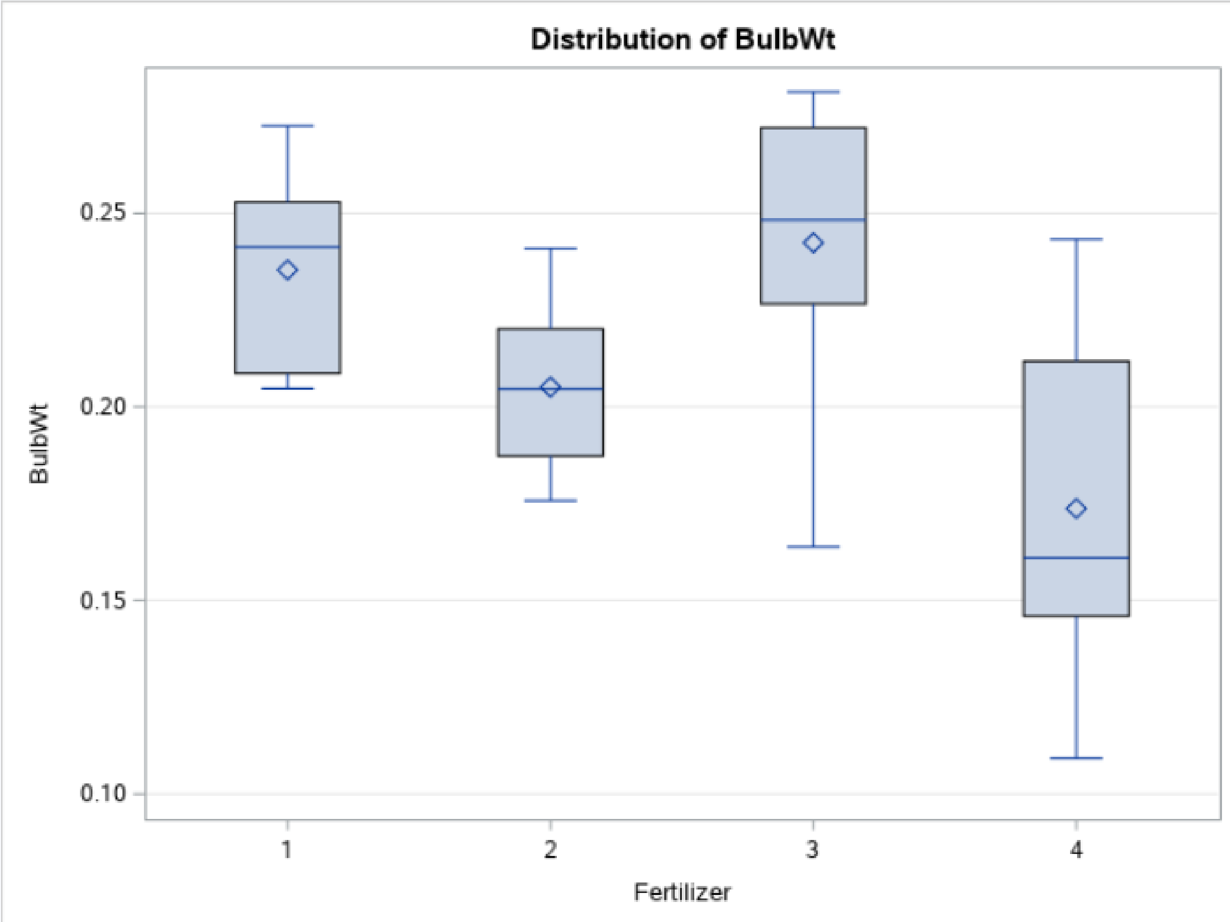
One-Way ANOVA with Fertilizer as Predictor

The GLM Procedure

Levene's Test for Homogeneity of BulbWt Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Fertilizer	3	9.13E-6	3.043E-6	1.54	0.2257
Error	28	0.000055	1.974E-6		

One-Way ANOVA with Fertilizer as Predictor

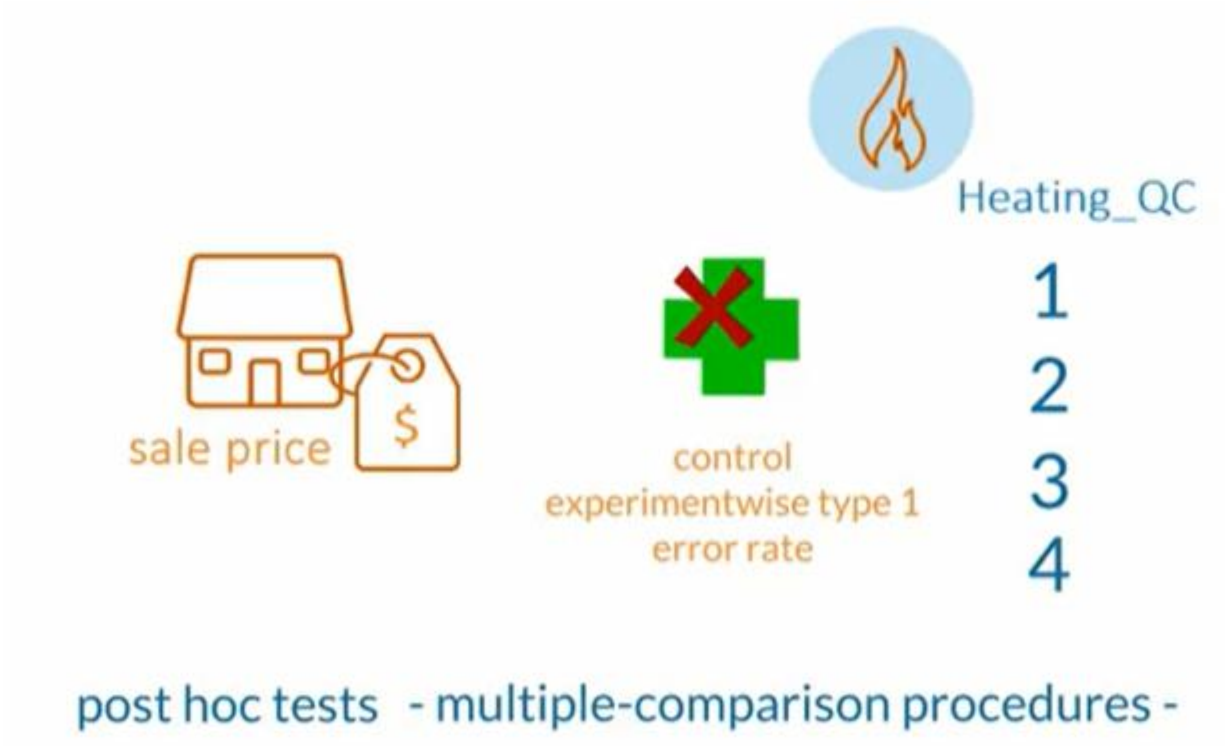
The GLM Procedure

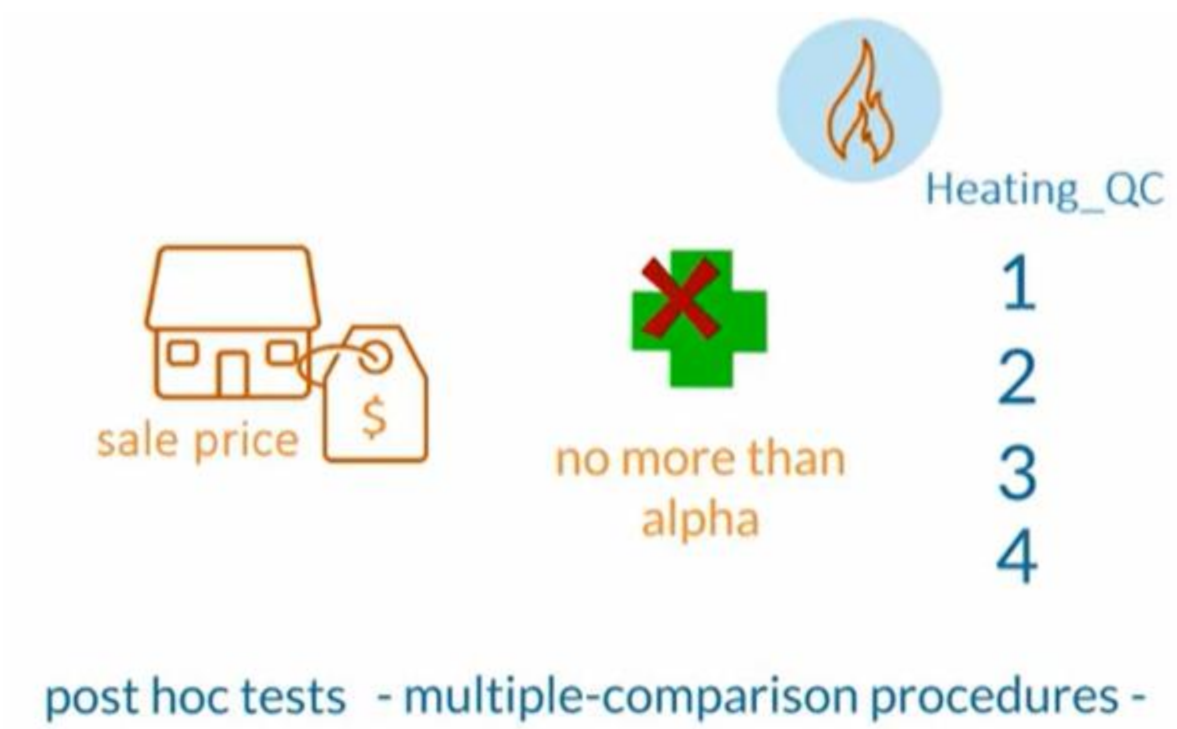


Level of Fertilizer	N	BulbWt	
		Mean	Std Dev
1	8	0.23539981	0.02540915
2	8	0.20511406	0.02220977
3	8	0.24240747	0.03868547
4	8	0.17376488	0.04447015

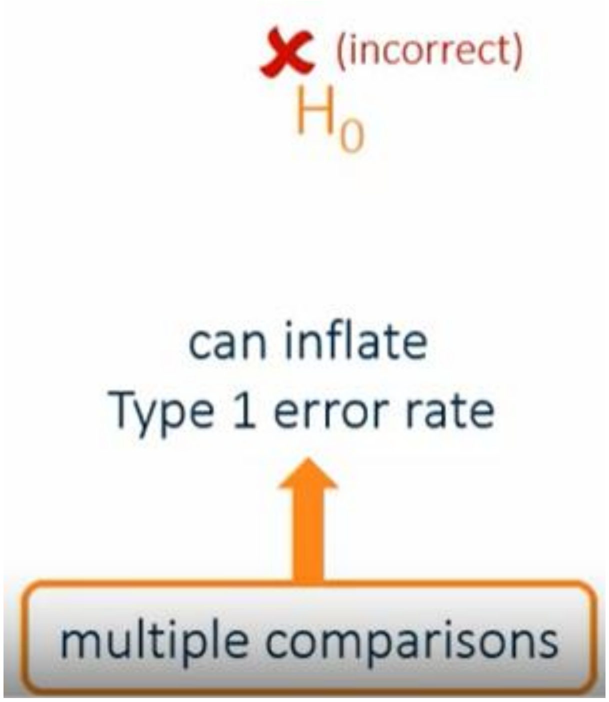
The overall *F* value from the analysis of variance table is associated with a *p*-value of 0.0013. Presuming that all assumptions of the model are valid, you know that at least one treatment mean is different from one other treatment mean. At this point, you don't know which means are significantly different from one another. Both the histogram and Q-Q plot show that the residuals seem relatively normally distributed (one assumption for ANOVA).The Levene's Test for Homogeneity of Variance table shows a *p*-value greater than alpha. Therefore, do not reject the hypothesis of homogeneity of variances (equal variances across fertilizer types). This assumption for ANOVA is met.

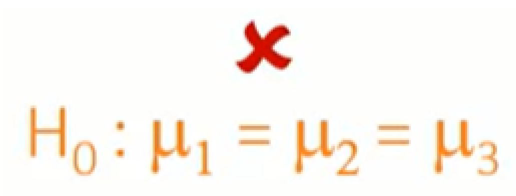
ANOVA Post Hoc Tests

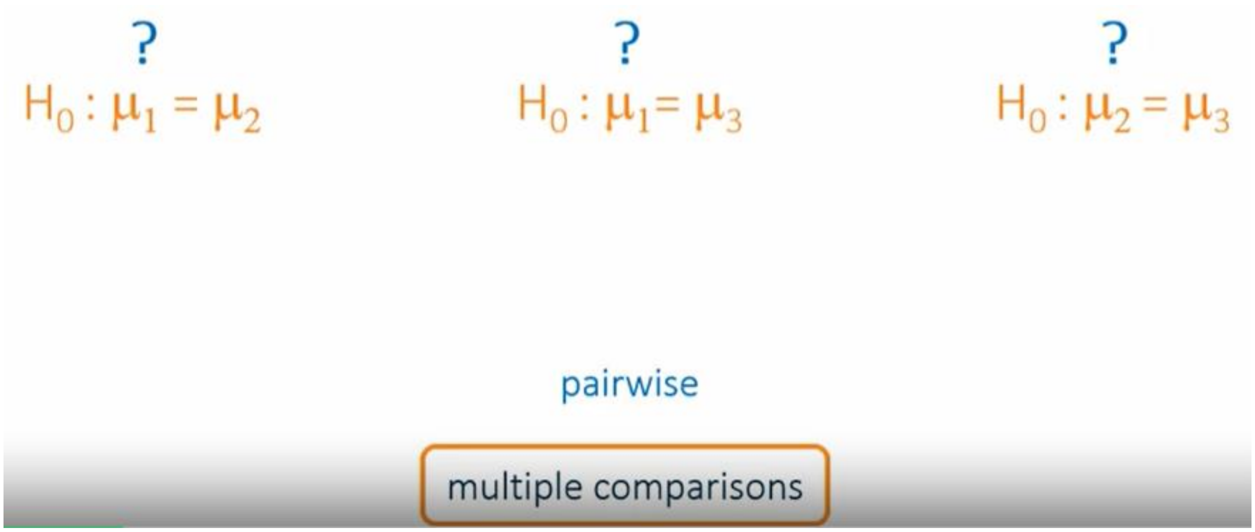


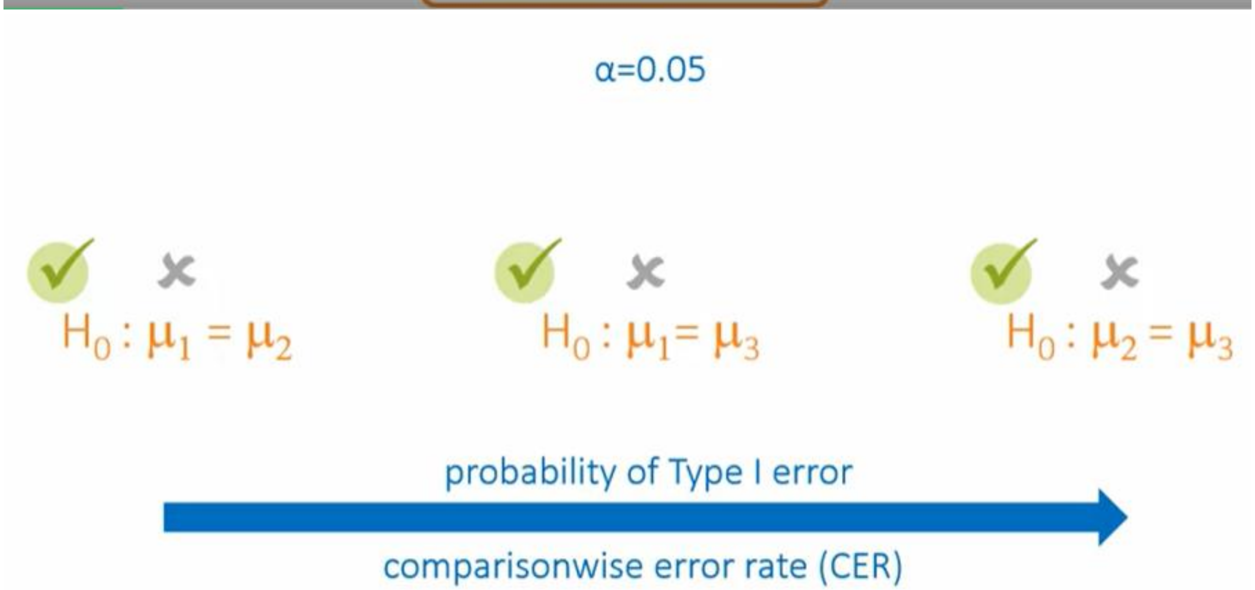


Multiple Comparison Methods


(incorrect)
H_0
can inflate
Type 1 error rate
multiple comparisons


$H_0 : \mu_1 = \mu_2 = \mu_3$


?
$H_0 : \mu_1 = \mu_2$
?
$H_0 : \mu_1 = \mu_3$
?
$H_0 : \mu_2 = \mu_3$
pairwise
multiple comparisons


α=0.05
$H_0 : \mu_1 = \mu_2$
$H_0 : \mu_1 = \mu_3$
$H_0 : \mu_2 = \mu_3$
probability of Type I error
comparisonwise error rate (CER)

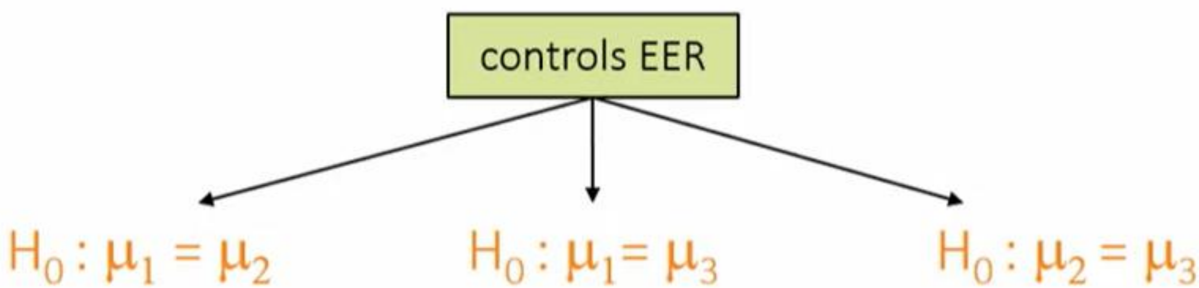
$$\text{EER} = 1-(1-\alpha)^{nc}$$

Comparisonwise Error Rate	Number of Comparisons	Experimentwise Error Rate
0.05	1	0.05
0.05	3	0.14
0.05	6	0.26
0.05	10	0.40

Tukey's and Dunnett's Multiple Comparison Methods

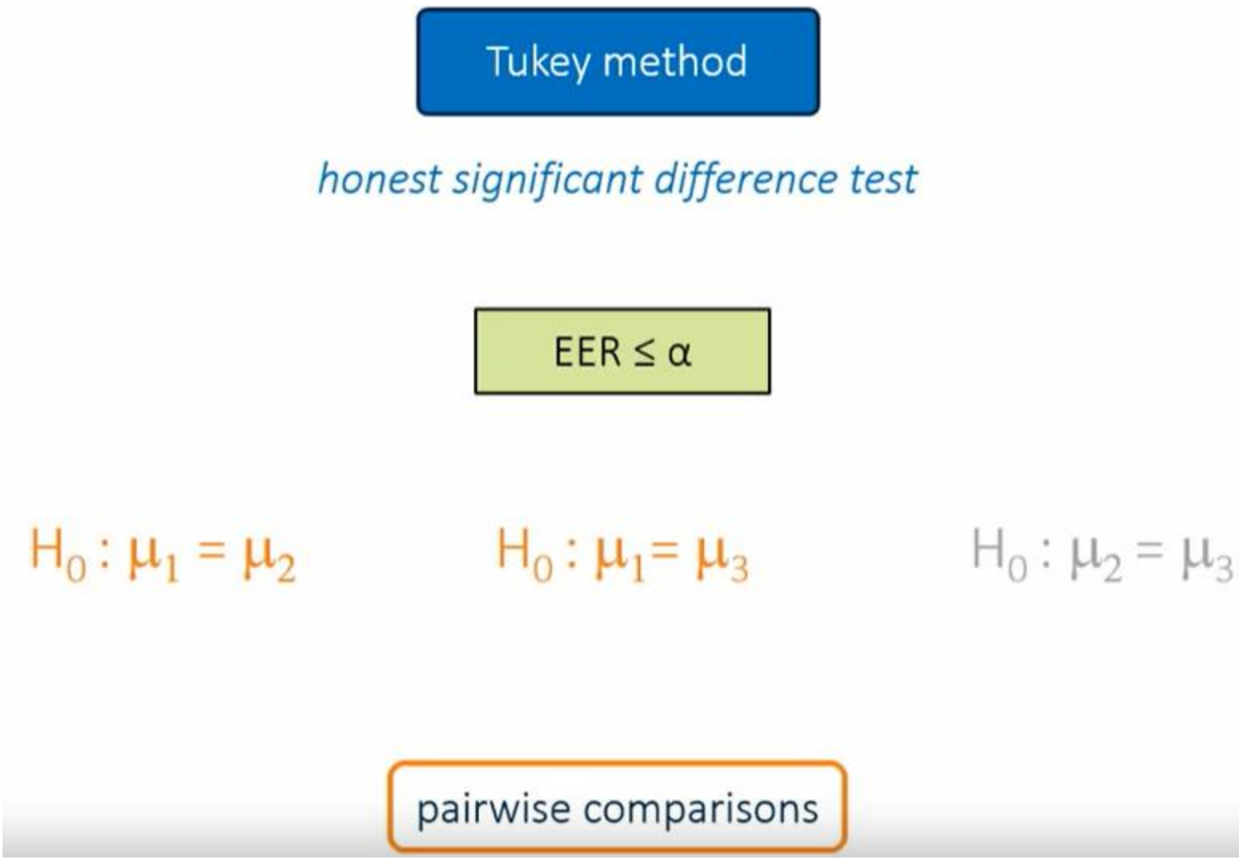
Tukey method

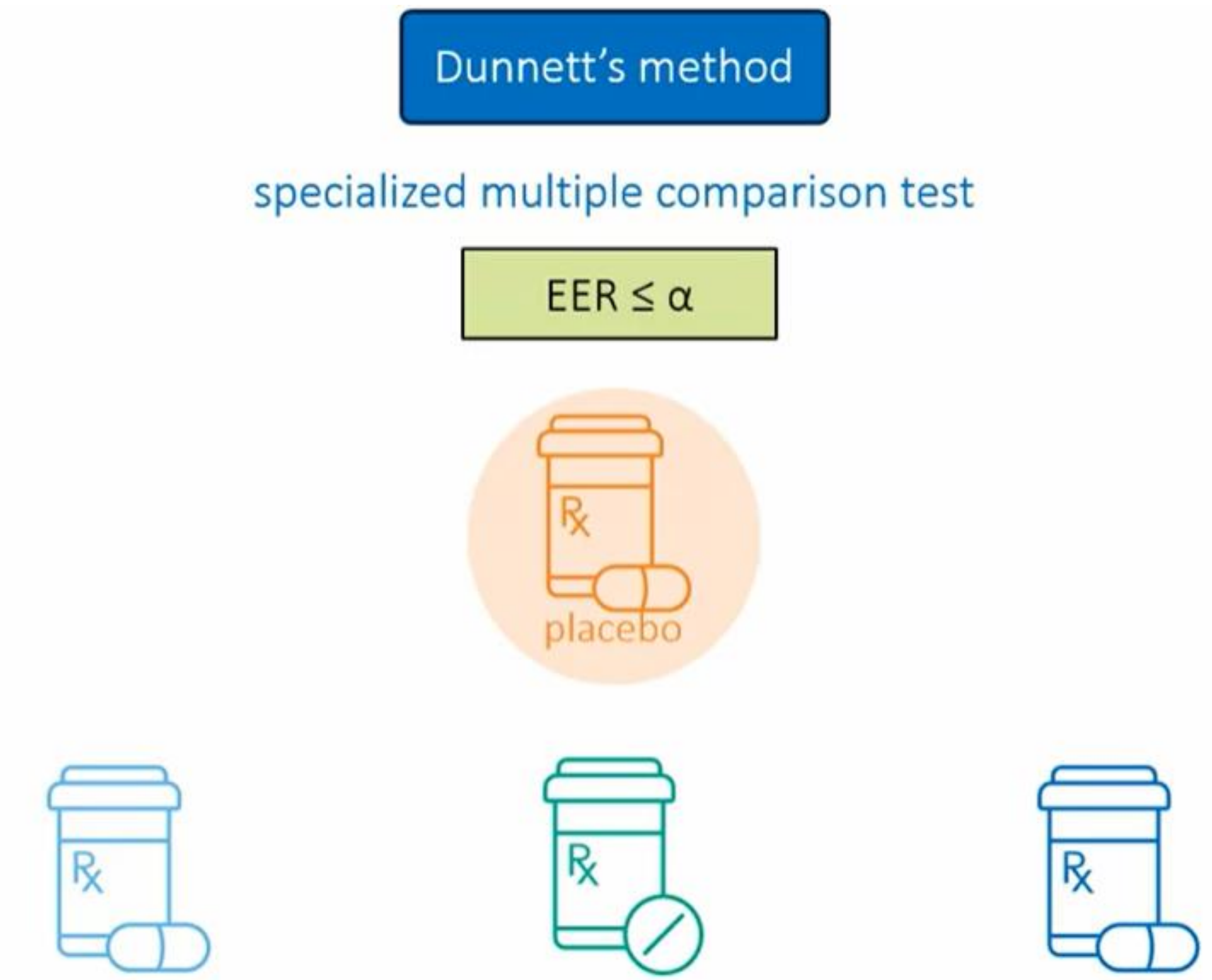
honest significant difference test



pairwise comparisons

overcorrects


Tukey method
honest significant difference test
$EER \leq \alpha$
$H_0 : \mu_1 = \mu_2$
$H_0 : \mu_1 = \mu_3$
$H_0 : \mu_2 = \mu_3$
pairwise comparisons


Dunnett's method
specialized multiple comparison test
EER ≤ α
Rx
placebo
Rx
Rx
Rx

other methods

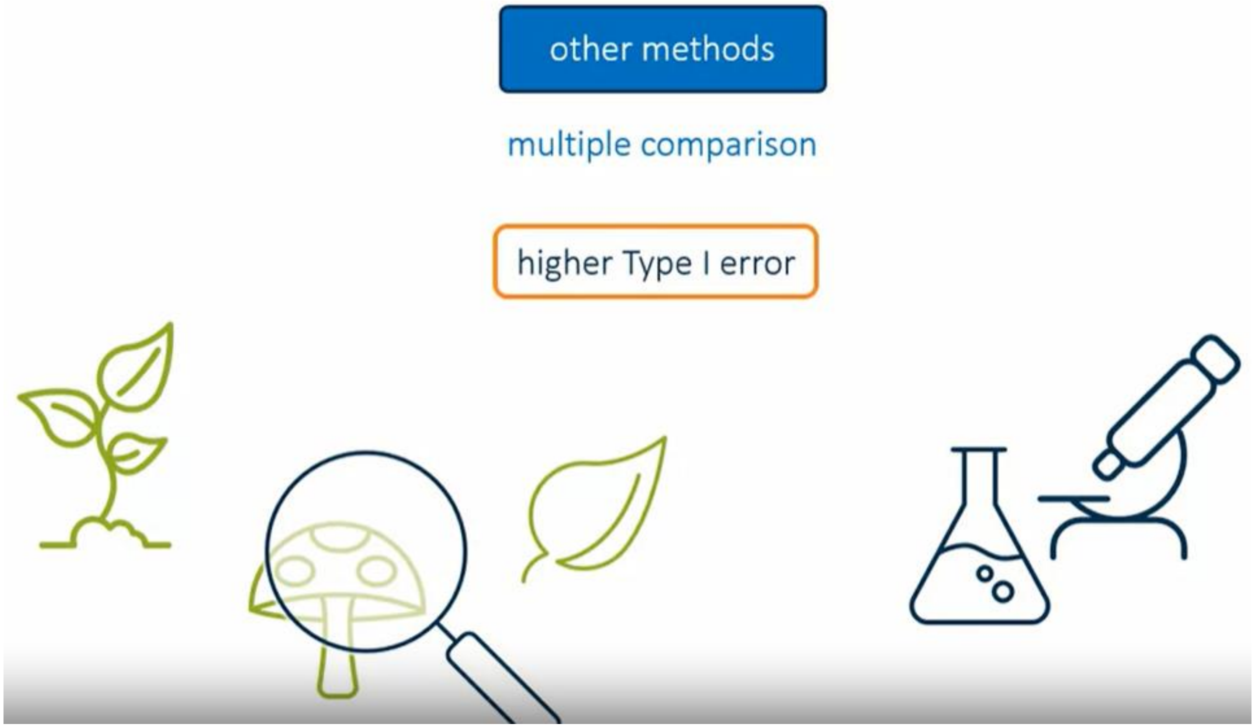
multiple comparison

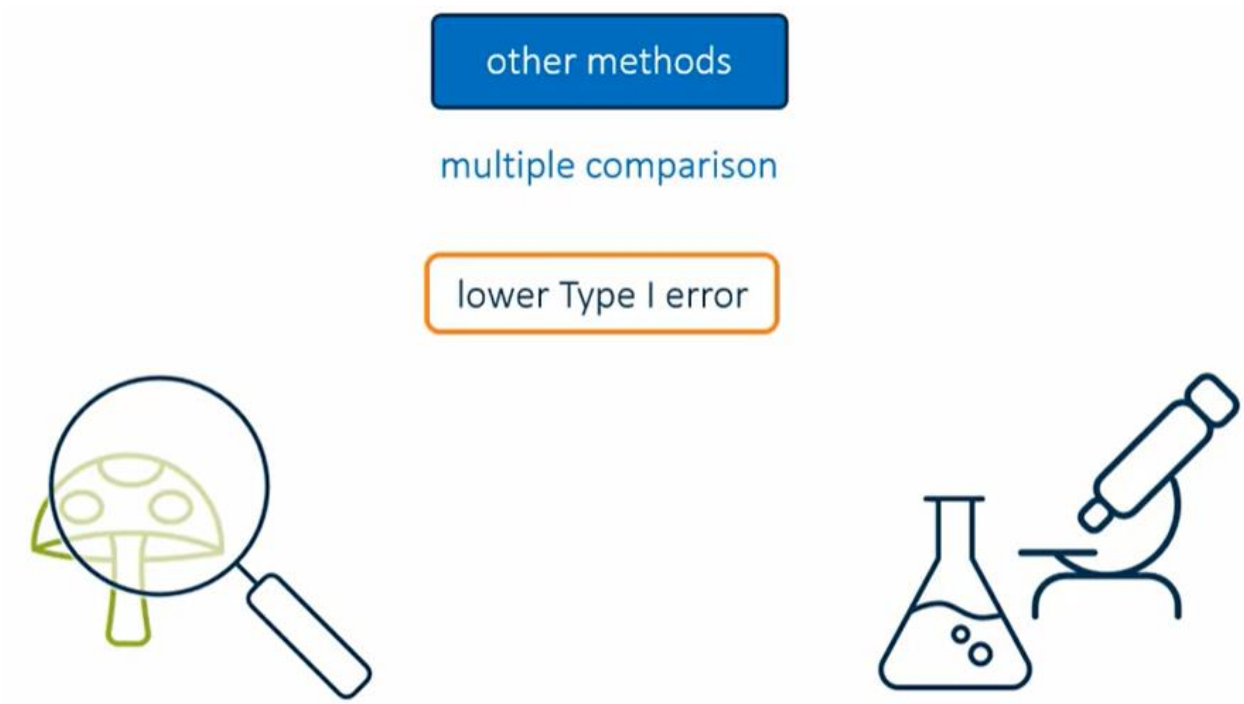
EER control varies

other methods

multiple comparison

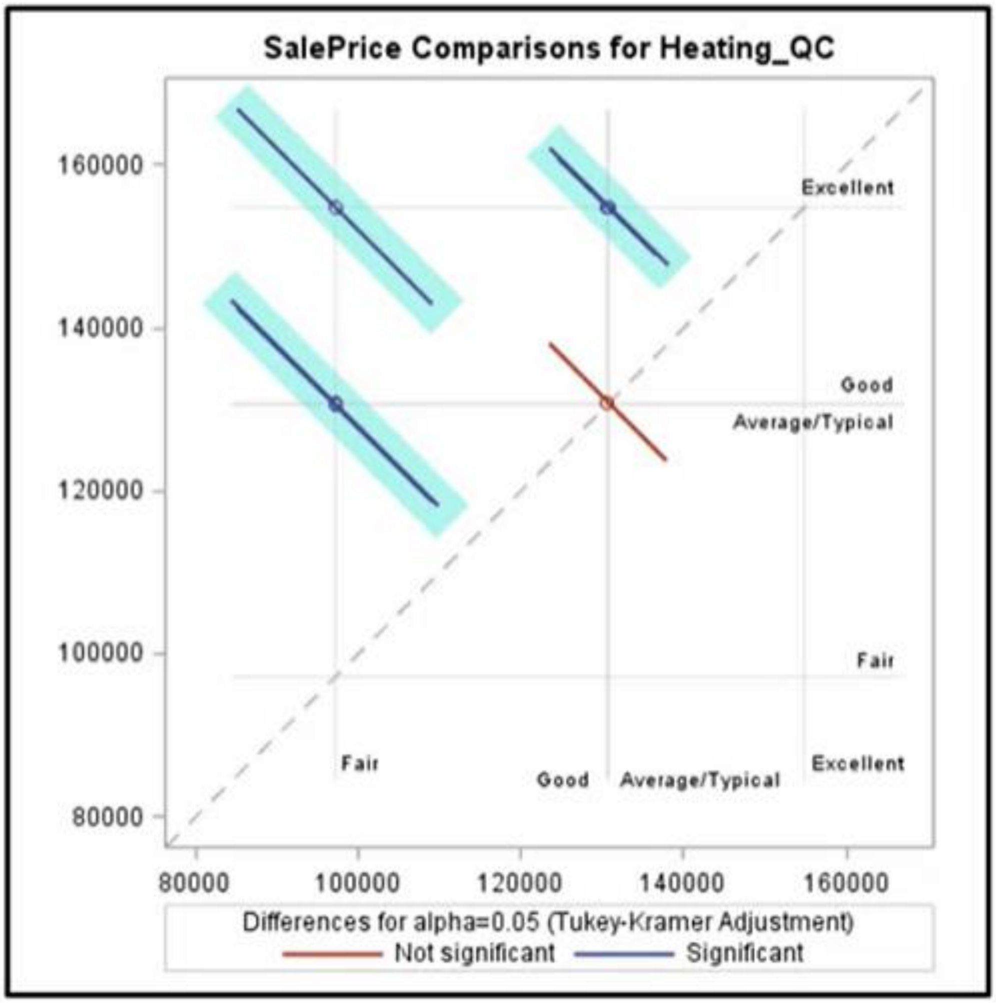
Decision \ Actual	H_0 is true	H_0 is false
Fail to reject H_0	Correct	Type II error p(Type II | H_a)=β
Reject H_0	Type I error p(Type I | H_0)=α	Correct $(1-\beta)$=Power

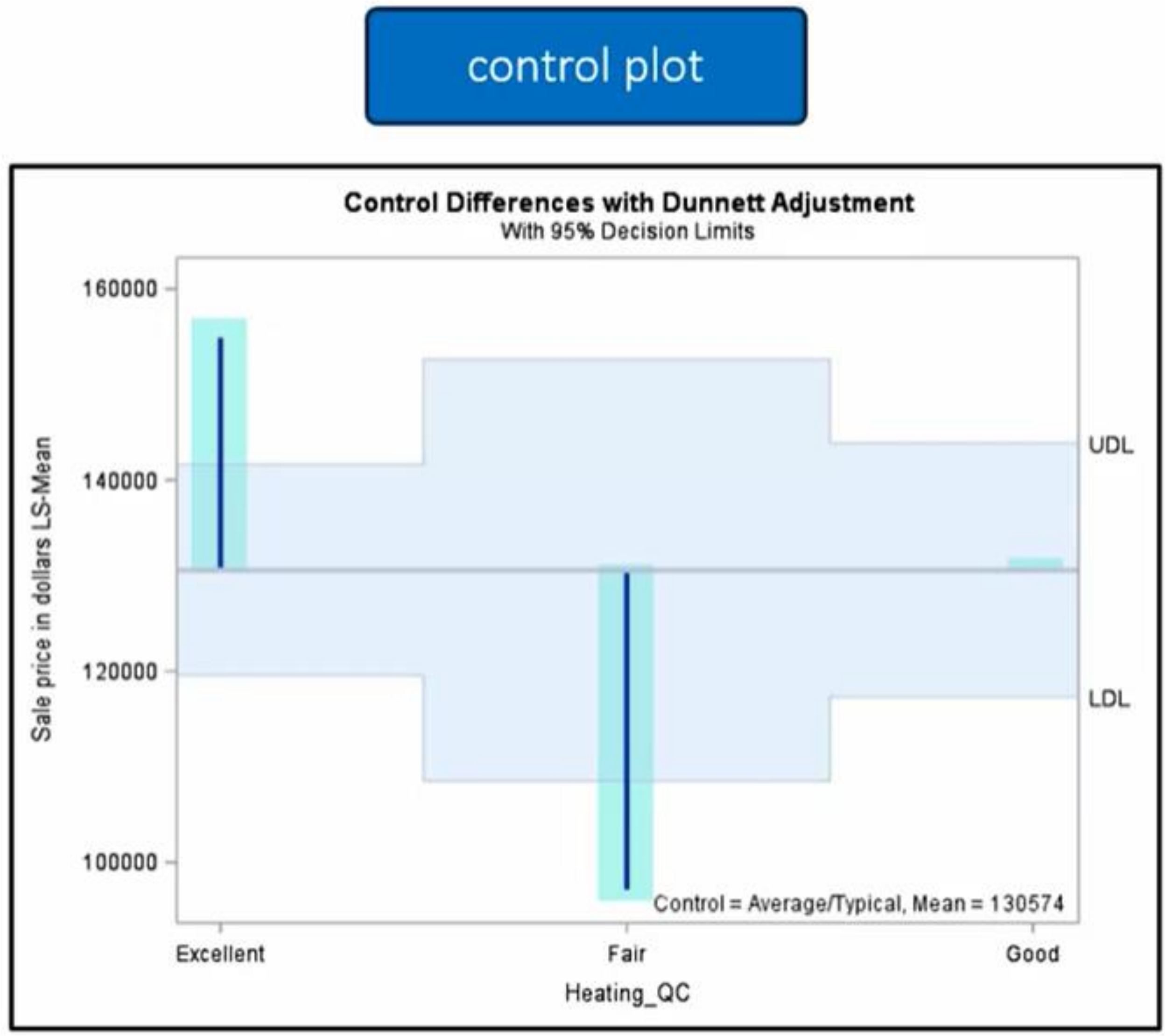




Diffograms and Control Plots

diffogram


SalePrice Comparisons for Heating_QC
160000
140000
120000
100000
80000
Excellent
Good
Average/Typical
Fair
Fair
Good
Average/Typical
Excellent
80000
100000
120000
140000
160000
Differences for alpha=0.05 (Tukey-Kramer Adjustment)
Not significant
Significant



Demo Performing a Post Hoc Pairwise Comparing Using PROC GLM


Excellent
Good
Average/Typical
Fair
quality of
heating system


PROC GLM
Excellent
Good
Good
Fair
Excellent
Average/Typical
Good
Average/Typical
Excellent
Fair
Fair
Average/Typical
sale price
$

```
/*st102d03.sas*/
ods graphics;

ods select lsmeans diff diffplot controlplot;
proc glm data=STAT1.ameshousing3
        plots(only)=(diffplot(center) controlplot);
    class Heating_QC;
    model SalePrice=Heating_QC;
    lsmeans Heating_QC / pdiff=all
                         adjust=tukey;
    lsmeans Heating_QC / pdiff=control('Average/Typical')
                         adjust=dunnett;
    format Heating_QC $Heating_QC.;
    title "Post-Hoc Analysis of ANOVA - Heating Quality as Predictor";
run;
quit;

title;
```

PROC GLM DATA=*SAS-data-set <options>***;**
CLASS *variable(s)*;
MODEL *dependent-variables=independent-effects < / options>*;
MEANS *effects < / options>*;
RUN;

```
/*st102d03.sas*/
ods graphics;

ods select lsmeans diff diffplot controlplot;
proc glm data=STAT1.ameshousing3
        plots(only)=(diffplot(center) controlplot);
    class Heating_QC;
    model SalePrice=Heating_QC;
    lsmeans Heating_QC / pdiff=all
                         adjust=tukey;
    lsmeans Heating_QC / pdiff=control('Average/Typical')
                         adjust=dunnett;
    format Heating_QC $Heating_QC.;
    title "Post-Hoc Analysis of ANOVA - Heating Quality as Predictor";
run;
quit;

title;
```

PROC GLM DATA=*SAS-data-set <options>***;**
CLASS *variable(s)*;
MODEL *dependent-variables=independent-effects < / options>*;
LSMEANS *effects < / options>*;
RUN;

Post-Hoc Analysis of ANOVA - Heating Quality as Predictor

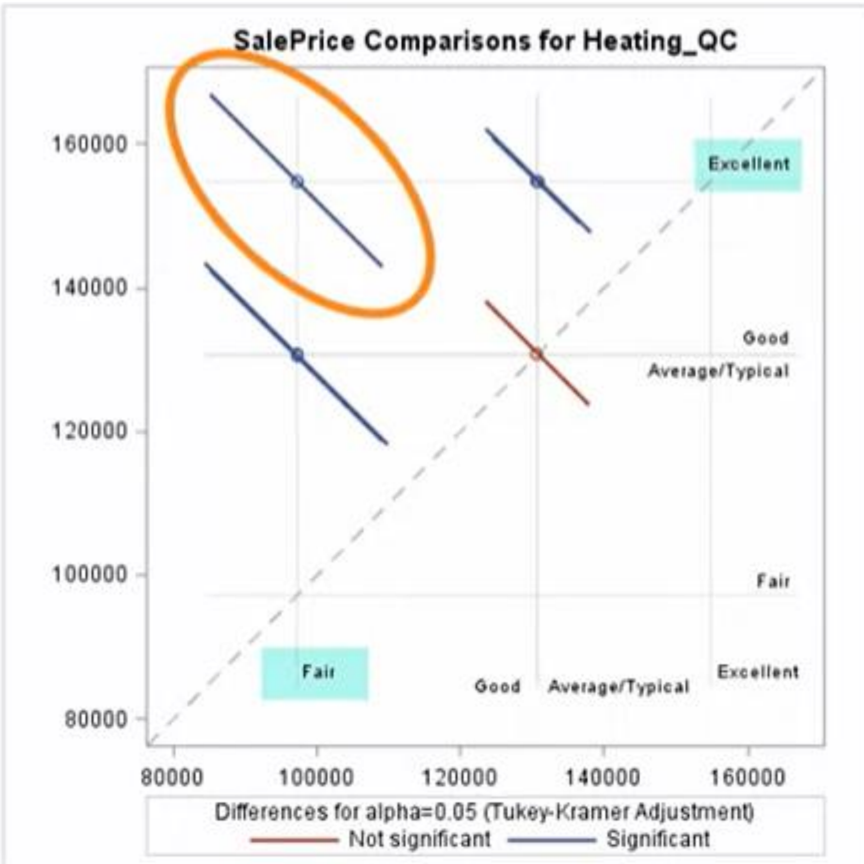
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Heating_QC	SalePrice LSMEAN	LSMEAN Number
Average/Typical	130573.529	1
Excellent	154919.187	2
Fair	97118.750	3
Good	130844.086	4

Least Squares Means for effect Heating_QC
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SalePrice

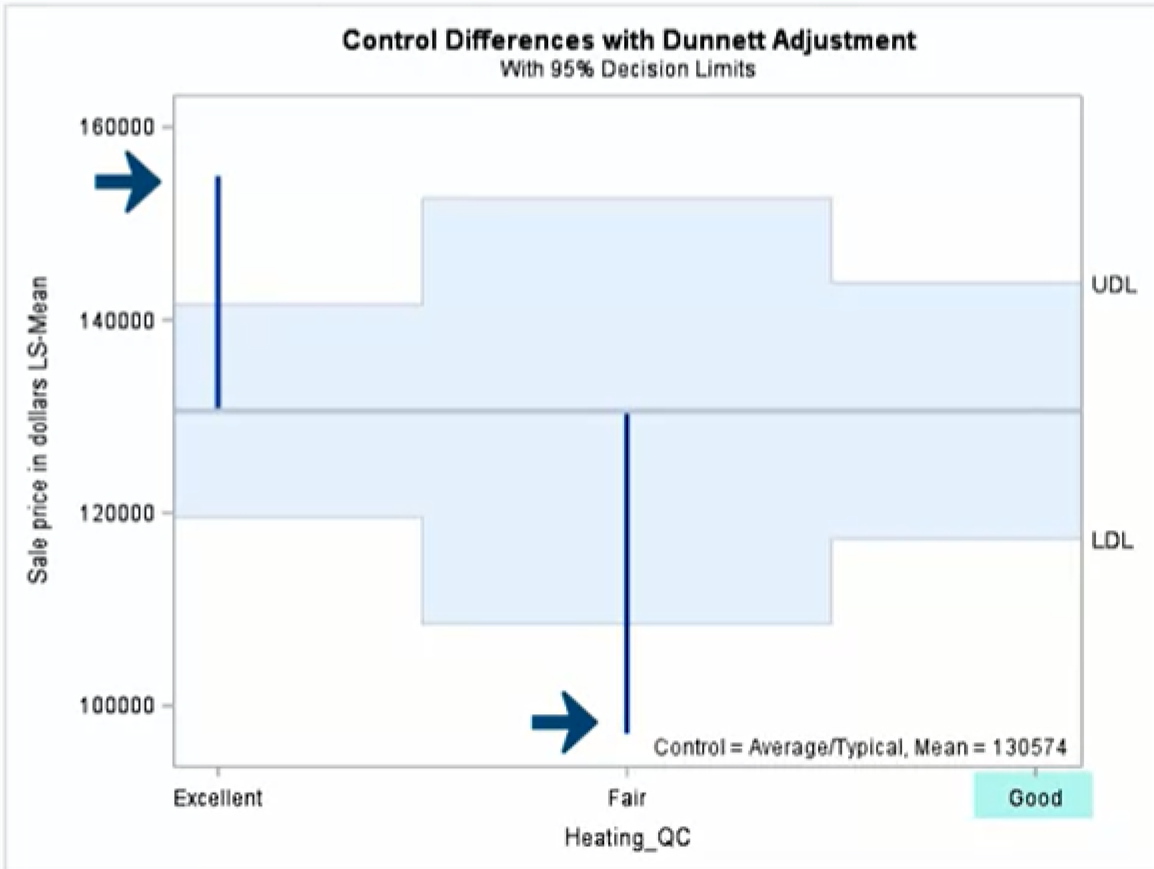
i/j	1	2	3	4
1		<.0001	0.0020	1.0000
2	<.0001		<.0001	0.0002
3	0.0020	<.0001		0.0037
4	1.0000	0.0002	0.0037	



Post-Hoc Analysis of ANOVA - Heating Quality as Predictor

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

Heating_QC	SalePrice LSMEAN	H0:LSMean=Control Pr > |t|
Average/Typical	130573.529	
Excellent	154919.187	<.0001
Fair	97118.750	0.0010
Good	130844.086	0.9999




```
/*st102d02.sas*/
ods graphics;


proc glm data=STAT1.ameshousing3 plots=diagnostics;
   class Heating_QC;
   model SalePrice=Heating_QC;
         means Heating_QC / hovtest=levene;
   format Heating_QC $Heating_QC.;
   title "One-Way ANOVA with Heating Quality as Predictor";
run;
quit;


title;
```

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

Class Level Information		
Class	Levels	Values
Heating_QC	4	Average/Typical Excellent Fair Good

Number of Observations Read	300
Number of Observations Used	300

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

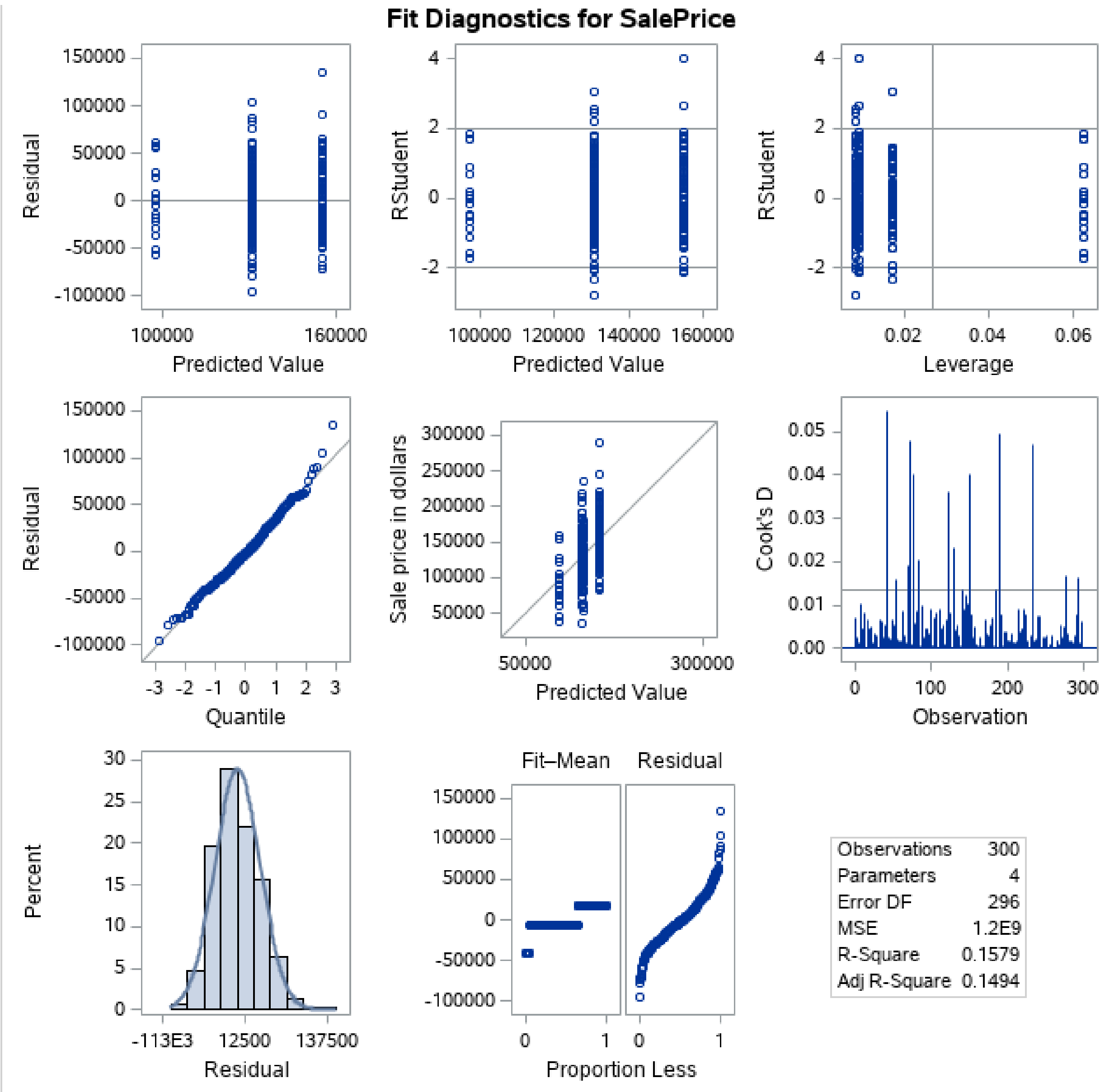
Dependent Variable: SalePrice Sale price in dollars

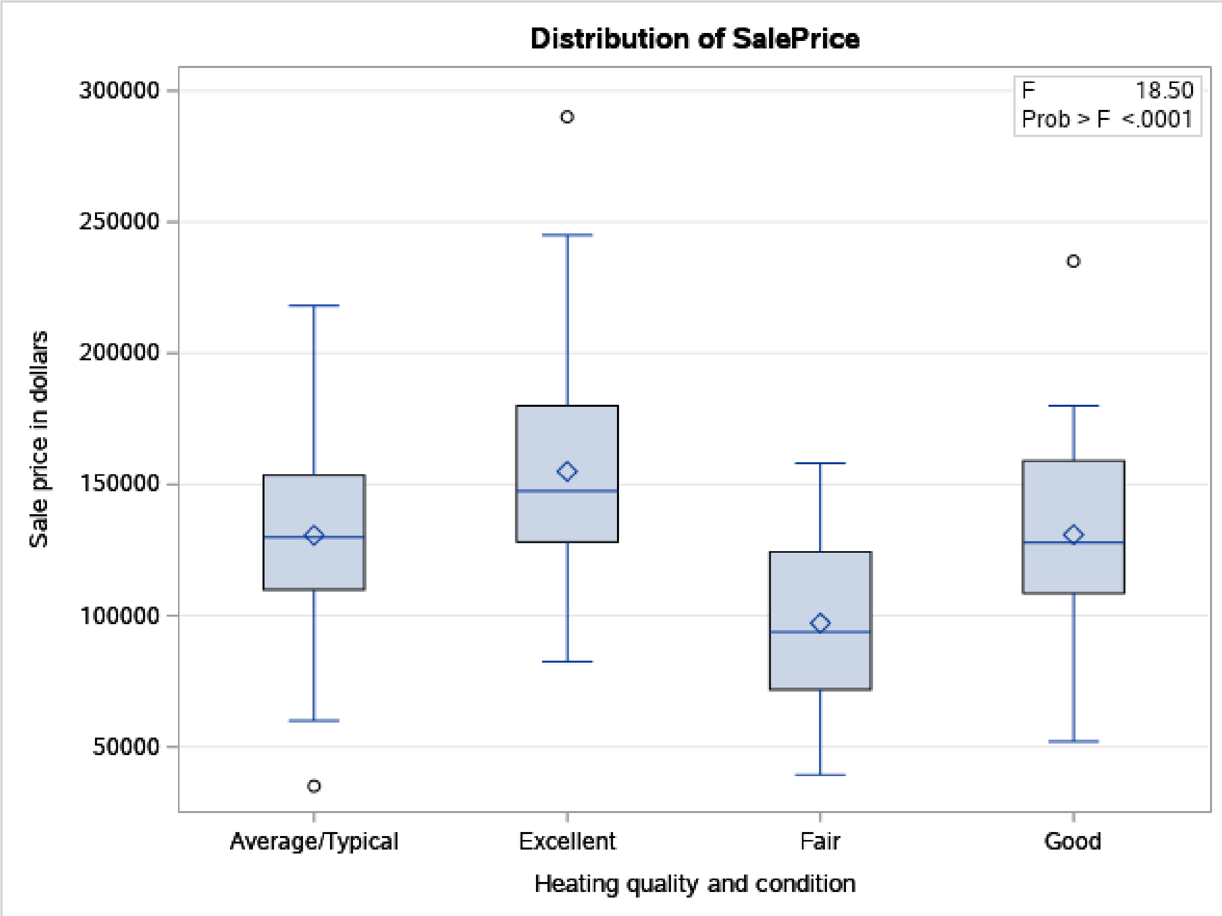
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.157920	25.23100	34698.90	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Sale price in dollars
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 300
Parameters 4
Error DF 296
MSE 1.2E9
R-Square 0.1579
Adj R-Square 0.1494



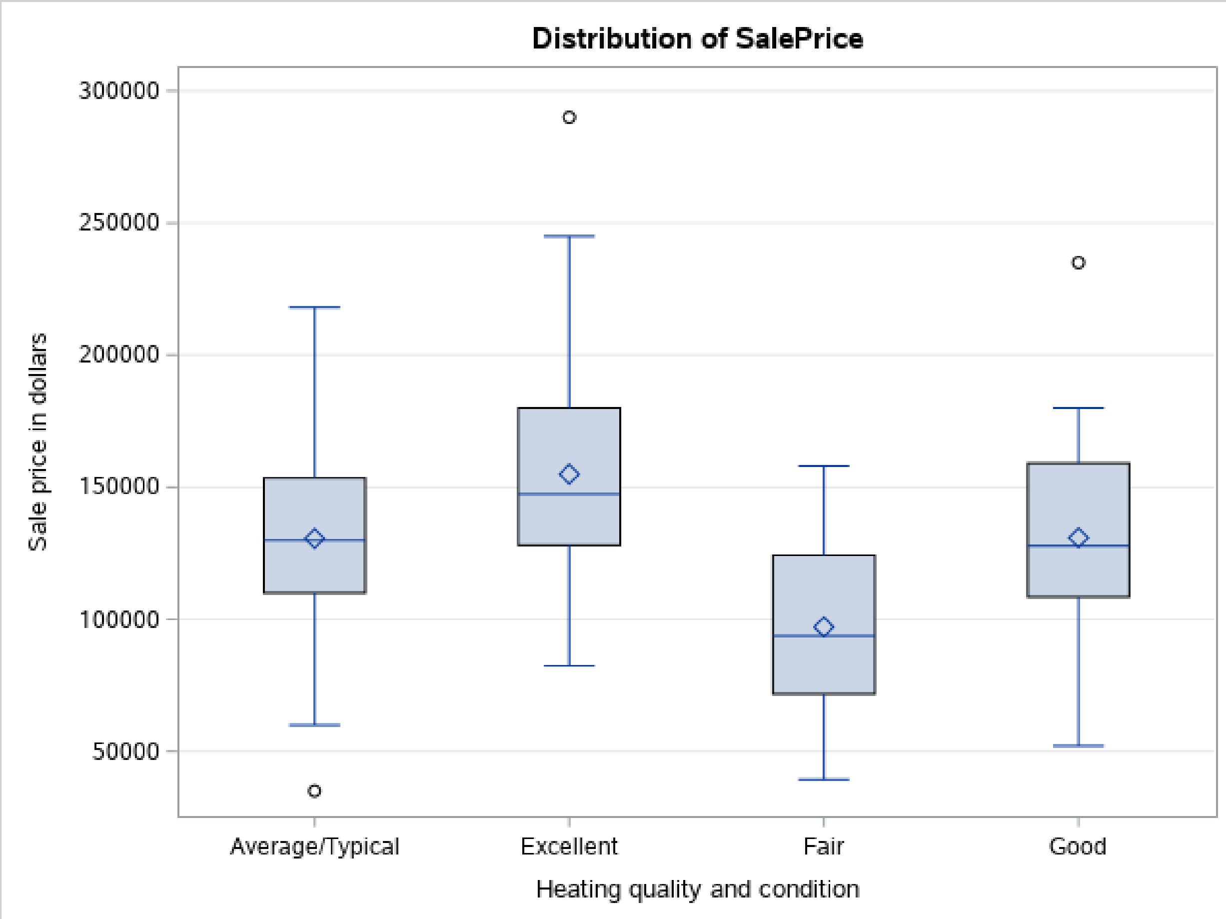
One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

Levene's Test for Homogeneity of SalePrice Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Heating_QC	3	5.931E18	1.977E18	0.58	0.6305
Error	296	1.014E21	3.426E18		

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure



Level of Heating_QC	N	SalePrice	
		Mean	Std Dev
Average/Typical	119	130573.529	32177.4508
Excellent	107	154919.187	36822.8795
Fair	16	97118.750	37423.5437
Good	58	130844.086	34912.5027

A statistically conservative multiple comparison test tends to do which of the following?

With a statistically conservative multiple comparison method, such as the Tukey or Dunnett method, you control for the EER, so there's a tendency to find fewer significant differences than might otherwise be found. When you make no adjustments for multiple comparisons, you are likely to find more significant differences than might otherwise be found.

Compared to a multiple comparisons test that controls the experimentwise error rate, what characteristics will a multiple comparisons test that only controls the comparisonwise error rate tend to have?

If only the comparisonwise error rate is controlled, the overall risk of a Type I error across all the comparisons is increased (and therefore the risk of Type II error is decreased), so the test might find more significant differences than would otherwise be found.

Practice - Using PROC GLM to Perform Post Hoc Parwise Comparisons

Question 1

Consider the analysis of the **garlic** data set. In the previous exercise, you used PROC GLM to perform one-way ANOVA, and found that there was a statistically significant difference among mean garlic bulb weights for the different fertilizers. Now, perform a post hoc test to look at the individual differences among means.

1. Use PROC GLM to conduct pairwise comparisons with an experimentwise error rate of α=0.05. (Use the Tukey adjustment.) Submit the code and view the results.
2. Which types of fertilizer are significantly different?

   The Tukey comparisons show significant differences between fertilizers 3 and 4 (p=0.0020) and 1 and 4 (p=0.0058).

Question 2

1. Use level 4 (the chemical fertilizer) as the control group and perform a Dunnett's comparison with the organic fertilizers to see whether they affected the average weights of garlic bulbs differently from the control fertilizer.
2. Which types of fertilizer are significantly different?

   The Dunnett comparisons show the same pairs as significantly different, but with smaller *p*-values than with the Tukey comparisons (3 versus 4 p=0.0011, 1 versus 4 p=0.0031). This is due to the fact that the Tukey adjustment is for more pairwise comparisons than the Dunnett adjustment.

```
/*st102s02.sas*/

ods graphics;


ods select lsmeans diff diffplot controlplot;

proc glm data=STAT1.Garlic

       plots(only)=(diffplot(center) controlplot);

   class Fertilizer;

   model BulbWt=Fertilizer;

   Tukey: lsmeans Fertilizer / pdiff=all
```

```
                    adjust=tukey;
    Dunnett:lsmeans Fertilizer / pdiff=control('4')
                    adjust=dunnett;
    No_Adjust: lsmeans Fertilizer / pdiff=all adjust=t;
    title "Post-Hoc Analysis of ANOVA - Fertilizer as Predictor";
run;
quit;

title;
```

Post-Hoc Analysis of ANOVA - Fertilizer as Predictor

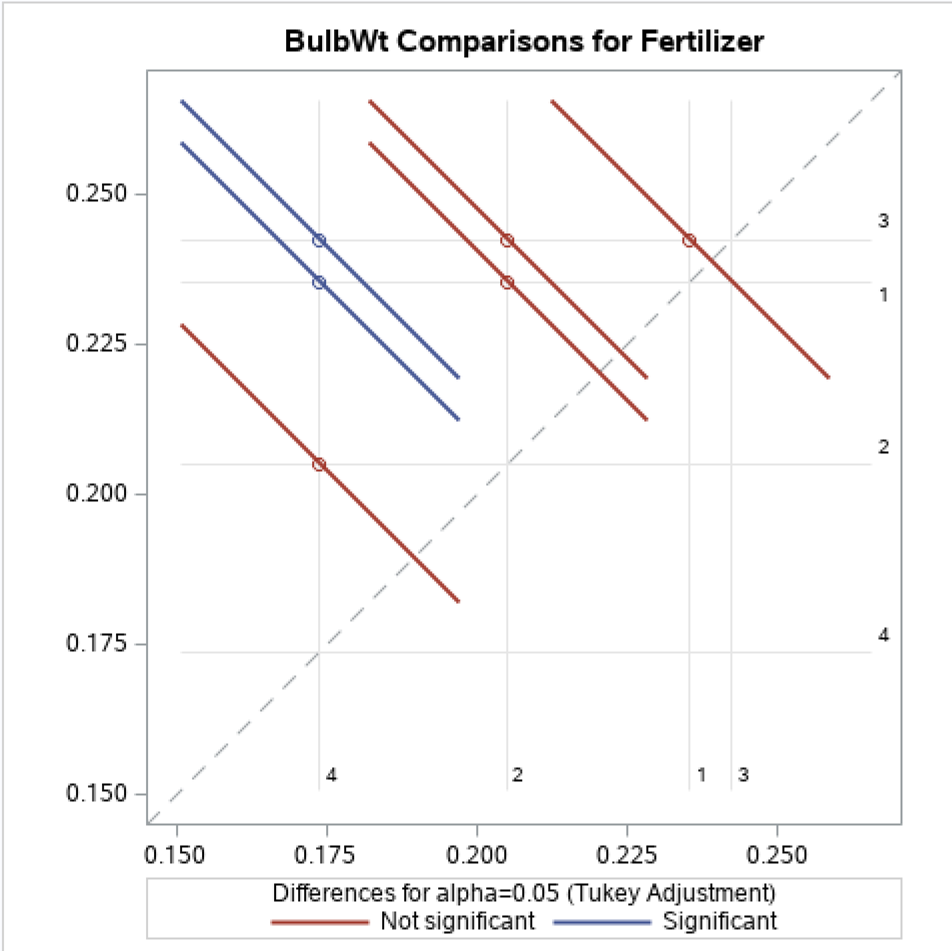
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Fertilizer	BulbWt LSMEAN	LSMEAN Number
1	0.23539981	1
2	0.20511406	2
3	0.24240747	3
4	0.17376488	4

Least Squares Means for effect Fertilizer
Pr > \|t\| for H0: LSMean(i)=LSMean(j)

Dependent Variable: BulbWt

i/j	1	2	3	4
1		0.3021	0.9758	0.0058
2	0.3021		0.1490	0.2738
3	0.9758	0.1490		0.0020
4	0.0058	0.2738	0.0020	

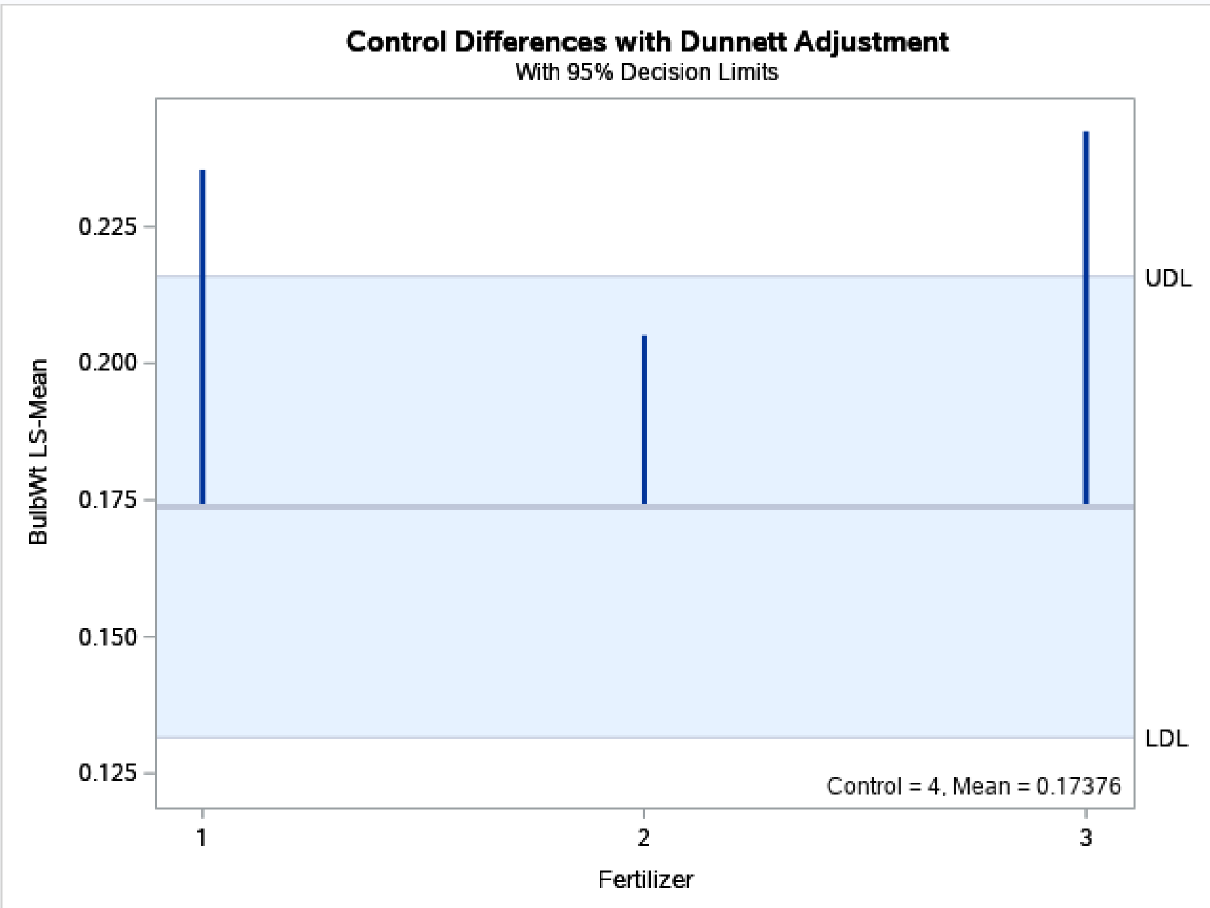


BulbWt Comparisons for Fertilizer
0.250
0.225
0.200
0.175
0.150
0.150
0.175
0.200
0.225
0.250
3
1
2
4
4
2
1
3
Differences for alpha=0.05 (Tukey Adjustment)
Not significant
Significant

Post-Hoc Analysis of ANOVA - Fertilizer as Predictor

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

Fertilizer	BulbWt LSMEAN	H0:LSMean=Control Pr > |t|
1	0.23539981	0.0031
2	0.20511406	0.1801
3	0.24240747	0.0011
4	0.17376488	



Post-Hoc Analysis of ANOVA - Fertilizer as Predictor

The GLM Procedure
Least Squares Means

Fertilizer	BulbWt LSMEAN	LSMEAN Number
1	0.23539981	1
2	0.20511406	2
3	0.24240747	3
4	0.17376488	4

Least Squares Means for effect Fertilizer
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: BulbWt

i/j	1	2	3	4
1		0.0853	0.6830	0.0011
2	0.0853		0.0365	0.0755
3	0.6830	0.0365		0.0004
4	0.0011	0.0755	0.0004