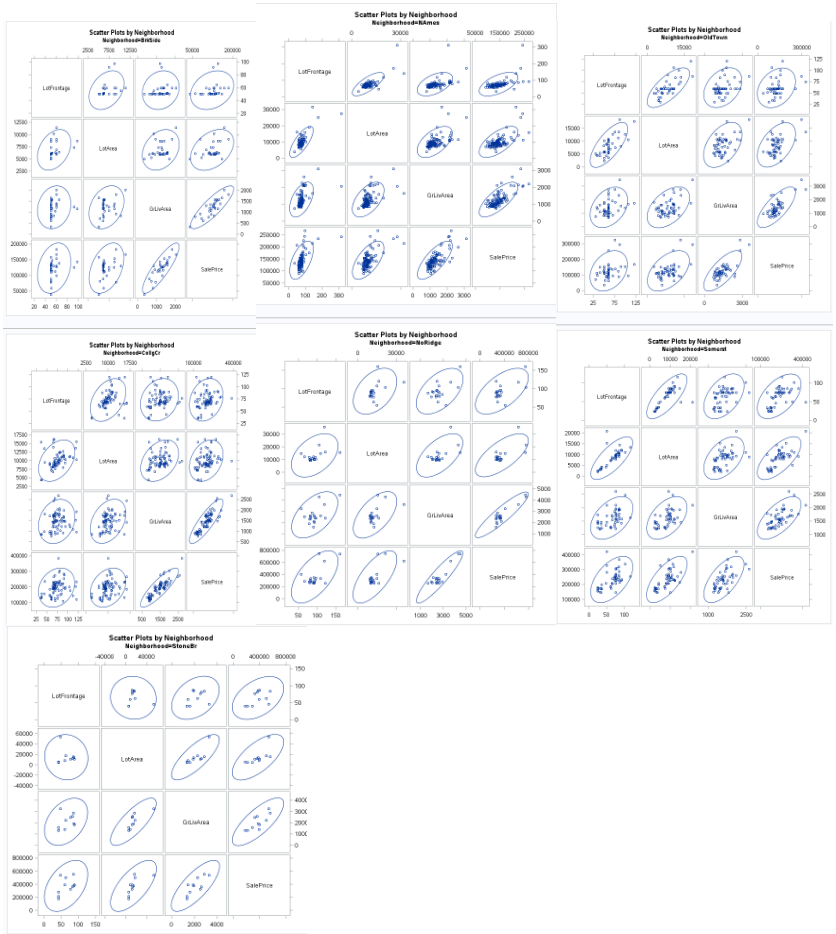
UNIT 8 HW: MANOVA and LDA

Consider the Ames housing data from a Kaggle project. I have adapted the data set (HW8AmesData.xlsx) so that there are only 7 neighborhoods represented. Using this data set we would like to perform the following analysis:

1. We would like to perform a MANOVA to test if there is evidence that any of the means of the of the response variables are different between any of the neighborhoods. In order to do this, we must first check the assumptions of the MANOVA:
    a. Note that for larger sample sizes, there is a multivariate normal central limit theorem and thus the MANOVA is robust to deviance from multivariate normality. We will assume that is the case here.
    b. Generate, copy and paste a matrix of scatter plots of the continuous variables for each of the neighborhoods. Make sure there is a fit ellipse on each. Make a judgement about the equality of the variance and covariances based on these plots. Please include your code as well. There is nothing to do or perform for this part.

```
title "Scatter Plots by Neighborhood";
proc sgscatter data = ames; by neighborhood;
matrix LotFrontage LotArea GrLivArea SalePrice / ellipse=(alpha = .05);
run;
title;
```



Assupmtions look good for the most part, you have but NAmes has the most variables outside the ellipses.

    c. It turns out that Bartlett's test is not robust to departures from the normality assumption and does not perform well in terms of type 1 error for tests with different sample sizes. Just for practice, run Bartlett's test and write a formal conclusion (one

sentence) for Bartlett's test. Note, for Barlett's test you will need to use proc discrim with pool=test option . Did it agree with your visual assessment above? Copy and paste your code and the table that shows the results of Bartlett's test.

```
proc discrim data=ames pool=test;
class neighborhood;
var LotFrontage LotArea GrLivArea SalePrice;
run;
```

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
547.932827	60	<.0001

d. From here you may proceed with the MANOVA assuming that the test is robust to any departures from multivariate normality and that the variance covariance matrices are equal (despite Bartlett's test). Simply proceed with caution with respect to the homoscedasticity assumption. You may also assume that observations for the same variables are independent both within and between neighborhoods. There is nothing to do or perform for this part.

2. Perform the MANOVA and copy and paste the MANOVA test result table below. This is the table with Wilk's Lambda, Hotelling-Lawley Trace, etc. Write a conclusion (one or two sentences) with respect to the test. Is there evidence that at least one mean between at least one pair of neighborhoods is different?

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Neighborhood Effect
H = Type III SSCP Matrix for Neighborhood
E = Error SSCP Matrix

S=4 M=0.5 N=170.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.26238653	23.33	24	1197.8	<.0001
Pillai's Trace	0.97613662	18.62	24	1384	<.0001
Hotelling-Lawley Trace	1.99569494	28.43	24	802.09	<.0001
Roy's Greatest Root	1.56838194	90.44	6	346	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

I don't see no evidence that one mean between the at least one pair is different in the MANVO. From the output of the MANOVA test we see a P-value of <.0001. We reject the null hypothesis that there is no difference in means for the neighborhoods.

3. You should find that there is evidence of a difference in means for at least one variable between at least one pair of neighborhoods! Now we need to investigate which variables and which pairs of neighborhoods. There are lots of combination to look at here; therefore, simply use a contrast to test for which variables have evidence of different means between the North Ames and Brook Side neighborhoods. Copy and paste the relevant tables as well as the code and be sure and provide at least a one or two sentence conclusion / summary (similar to what is found in the powerpoints).

```
proc glm data=ames;
class neighborhood;
model LotFrontage LotArea GrLivArea SalePrice = neighborhood;
contrast 'Brook Side vs North Ames' neighborhood -1 0 1 0 0 0 0;
estimate 'Brook Side vs North Ames' neighborhood -1 0 1 0 0 0 0;
manova h=neighborhood / printe printh;
lsmeans neighborhood / pdiff tdiff adjust=bon;
run
```

Dependent Variable: GrLivArea

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	40426369.3	6737728.2	32.11	<.0001
Error	346	72594254.8	209810.0		
Corrected Total	352	113020624.1			

R-Square	Coeff Var	Root MSE	GrLivArea Mean
0.357690	30.71033	458.0502	1491.518

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Neighborhood	6	40426369.34	6737728.22	32.11	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Neighborhood	6	40426369.34	6737728.22	32.11	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Brook Side vs North Ames	1	67743.47494	67743.47494	0.32	0.5703

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Brook Side vs North Ames	52.9789879	93.2358637	0.57	0.5703

Dependent Variable: SalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1.643207E12	273867835926	76.05	<.0001
Error	346	1.2460323E12	3601249316.9		
Corrected Total	352	2.8892393E12			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.568733	32.58458	60010.41	184168.1

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Neighborhood	6	1.643207E12	273867835926	76.05	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Neighborhood	6	1.643207E12	273867835926	76.05	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Brook Side vs North Ames	1	11107138910	11107138910	3.08	0.0799

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Brook Side vs North Ames	21452.1598	12215.0856	1.76	0.0799

The GrLivArea and SalePrice shows evidence of significant different means between North Ames and Brook Side. The p- values are 0.5703 and 0.0799.

4. Next let's perform a discriminant analysis (LDA or QDA) to help classify (discriminate between) neighborhoods using the variables above as explanatory variables (features).
Evaluate your model (LDA or QDA) with respect to mis-classification rate from a cross validation rather than a resubstitution. Copy and paste the cross validated confusion matrix and mis-classification table as well as your code. Again, summarize your findings in at least one or two sentences. (Assume equal priors for the neighborhoods.)

```
proc discrim data=ames pool=test crossvalidate;
class neighborhood;
var LotFrontage LotArea GrLivArea SalePrice;
run;
```

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.AMES
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Neighborhood								
From Neighborhood	BrkSide	CollgCr	NAmes	NoRidge	OldTown	Somerst	StoneBr	Total
BrkSide	22 70.97	0 0.00	3 9.68	0 0.00	6 19.35	0 0.00	0 0.00	31 100.00
CollgCr	5 6.33	42 53.16	17 21.52	3 3.80	2 2.53	9 11.39	1 1.27	79 100.00
NAmes	25 22.94	11 10.09	56 51.38	2 1.83	13 11.93	1 0.92	1 0.92	109 100.00
NoRidge	0 0.00	2 10.00	0 0.00	13 65.00	0 0.00	3 15.00	2 10.00	20 100.00
OldTown	18 34.62	2 3.85	7 13.46	4 7.69	20 38.46	1 1.92	0 0.00	52 100.00
Somerst	0 0.00	18 35.29	0 0.00	2 3.92	1 1.96	26 50.98	4 7.84	51 100.00
StoneBr	0 0.00	0 0.00	0 0.00	4 36.36	0 0.00	4 36.36	3 27.27	11 100.00
Total	70 19.83	75 21.25	83 23.51	28 7.93	42 11.90	44 12.46	11 3.12	353 100.00
Priors	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	

Error Count Estimates for Neighborhood								
	BrkSide	CollgCr	NAmes	NoRidge	OldTown	Somerst	StoneBr	Total
Rate	0.2903	0.4684	0.4862	0.3500	0.6154	0.4902	0.7273	0.4897
Priors	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	

BrkSide is the best at predicting since it has a error count of around 29% and Stone Br the worst at 72%.

5. Finally, we would like to predict / classify / impute a house that we do not know what neighborhood in which it belongs. This house has a lot frontage of 52 ft, a lot area of 6000 sqft, an above ground living area of 1,400 sqft and a sale price of $110,000. Copy and paste the relevant table, your code and please include a short summary of your findings.

```
data newhouse;
input LotFrontage LotArea GrLivArea SalePrice;
cards;
52  6000  1400  110000
;

proc discrim data=ames pool=test crossvalidate testdata=newhouse
testout=newhouseclassify;
class neighborhood;
var LotFrontage LotArea GrLivArea SalePrice;
priors "BrkSide"=.1429 "CollgCr"=.1429 "NAmes"=.1429 "NoRidge"=.1429 "OldTown"=.1429
"Somerst"=.1429 "StoneBr"=.1429;
run;

proc print data = newhouseclassify;
run;
```

Obs	LotFrontage	LotArea	GrLivArea	SalePrice	BrkSide	CollgCr	NAmes	NoRidge	OldTown	Somerst	StoneBr	_INTO_
1	52	6000	1400	110000	0.65889	.001099465	0.060232	.007616294	0.24888	0.021996	.001281314	BrkSide

It was predicted the house to be into BrkSide with the highest probability of 65%