DSC 324 - Advanced Data Analysis
Homework Module 1

Complete the following problems and submit your answers in a single pdf to the appropriate folder on d2l by **Sunday, January 16th at 11:59PM CST.**

Worth 25 points each:

1. Perform the following calculations by hand using the given matrices and vectors. Submit your answers as either scans or (clear!) photos.

$Z = \begin{bmatrix} 1 & 5 \\ 1 & -3 \\ 1 & 2 \\ 1 & 4 \end{bmatrix}$ $Y = \begin{bmatrix} 2 \\ 1 \\ -1 \\ 3 \end{bmatrix}$ $M = \begin{bmatrix} 20 & 15 & 0 \\ 5 & 25 & 10 \\ 0 & 20 & 5 \end{bmatrix}$ $N = \begin{bmatrix} -20 & 5 & 10 \\ 0 & -10 & 10 \\ 5 & 20 & -5 \end{bmatrix}$ $v = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$ $w = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$

a) v • w (dot product)
b) 3 * w
c) M*v
d) M+N
e) M-N
f) Z^T

g) Z^TZ
h) Z^{-1}
i) Z^TY
j) determinant of Z^TZ

1a) $v \cdot w$ (dot product)

Incompatible because both v and w are both 3×1 matric

$V = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$ 3×1 $\begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$ 3×1 Do Not Match

1b) $3 * w$

$3 \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \\ -3 \end{bmatrix}$

1c) M * V

$$\begin{bmatrix} 20 & 15 & 0 \\ 5 & 25 & 10 \\ 0 & 20 & 5 \end{bmatrix} * \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} = \begin{bmatrix} (20\times1)+(15\times-1)+(0\times3) = 5 \\ (5\times1)+(25\times-1)+(10\times3) = 10 \\ (0\times1)+(20\times-1)+(5\times3) = -5 \end{bmatrix} \begin{bmatrix} 5 \\ 10 \\ -5 \end{bmatrix}$$

1d) M + N

$$\begin{bmatrix} 20 & 15 & 0 \\ 5 & 25 & 10 \\ 0 & 20 & 5 \end{bmatrix} + \begin{bmatrix} -20 & 5 & 10 \\ 0 & -10 & 10 \\ 5 & 20 & -5 \end{bmatrix} = \begin{bmatrix} 0 & 20 & 10 \\ 5 & 15 & 20 \\ 5 & 40 & 0 \end{bmatrix}$$

20 + (-20) 15+5 0+10
5+0 25+(-10) 10+10
0+5 20+20 5+0

1e) M - N

$$\begin{bmatrix} 20 & 15 & 0 \\ 5 & 25 & 10 \\ 0 & 20 & 5 \end{bmatrix} - \begin{bmatrix} -20 & 5 & 10 \\ 0 & -10 & 10 \\ 5 & 20 & -5 \end{bmatrix} = \begin{bmatrix} 40 & 10 & -10 \\ 5 & 35 & 0 \\ -5 & 0 & 10 \end{bmatrix}$$

20-(-20) 15-5 0-10
5-0 25-(-10) 10-10
0-5 20-20 5-(-5)

1f) Z^T

$$Z = \begin{bmatrix} 1 & 5 \\ 1 & -3 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \quad Z^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & -3 & 2 & 4 \end{bmatrix}$$

1g) $Z^T Z$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & -3 & 2 & 4 \end{bmatrix} \times \begin{bmatrix} 1 & 5 \\ 1 & -3 \\ 1 & 2 \\ 1 & 4 \end{bmatrix}$$

$(1\times1)+(1\times1)+(1\times1)+(1\times1) = 4$

$(1\times5)+(1\times-3)+(1\times2)+(1\times4) = 8$

$(5\times1)+(-3\times1)+(2\times1)+(4\times1) = 8$

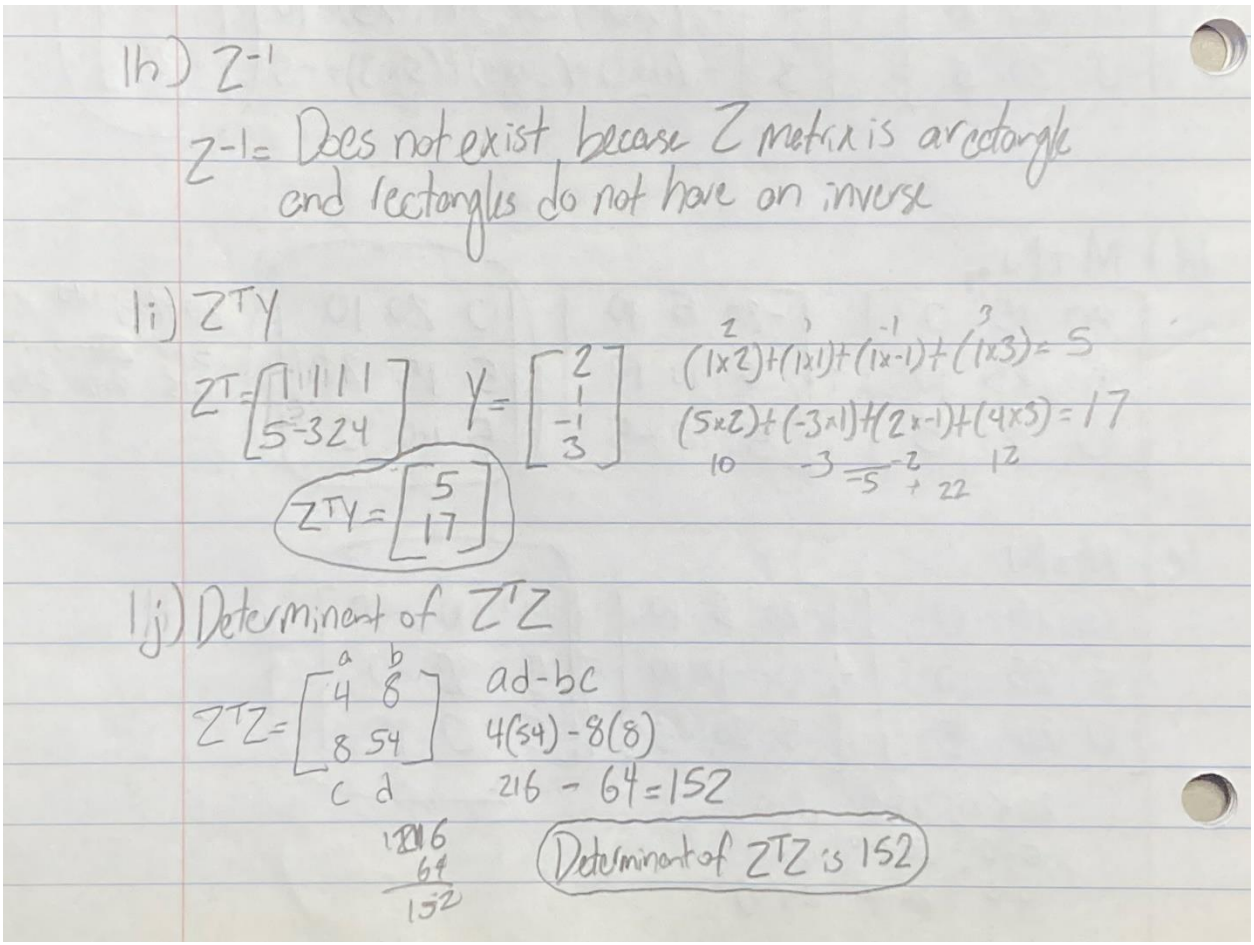
$(5\times5)+(-3\times-3)+(2\times2)+(4\times4) = 54$

25 9 4 16

13

29

$Z^T Z = \begin{bmatrix} 4 & 8 \\ 8 & 54 \end{bmatrix}$



2. The file "chicinsur.csv" contains insurance data collected from 47 zip codes in the Illinois area. The following is a description of the relevant variables:

NEWPOL - Number of new home insurance policies (minus cancelled policies) per 100 housing units
PCTMINOR - Percentage of minorities living in the area
FIRES - Number of fires per 1000 housing units
THEFTS - Number of thefts per 1000 people
PCTOLD - Percentage of housing units built before 1940
INCOME - Median income of the area

If our response variable (dependent variable) is NEWPOL, we want to see which of the other variables (independent variables) are significant predictors of NEWPOL. In other words, can the other five variables reliably predict the number of new home insurance policies?

a) Think about the potential relationships (correlations) between the independent variables and the response variable. List each independent variable and predict the sign of its coefficient in a multiple linear regression.

Predicting sign predictions of coefficients:

PCTMINOR- Negative I feel it is a negative correlation the more minorities in an area I feel that they would not have the funds to have money to buy insurance for a new home insurance granted not many new homes would be made around mainly minorities too.

FIRES- Positive I think it would be positive because if there are fires the more house insurance would be a common buy because of that safety net they spend into.

THEFTS- Positive I feel in any area when a new home is being made the more willing a buyer would want to have house insurance for their house for thefts because a new house appeals more that it looks "newer".

PCTOLD- Negative this is for new home insurance so I feel that if a house would have insurance, it would have something already and these houses are old enough where it could be bought out one day, and a new house will be built where it stood.

INCOME- Positive since the more money a person is making the more willing, they would be to buy new home insurance for their house to protect it.

b) Generate either a correlation matrix or scatterplot matrix for this dataset. Compare the direction of the correlations with your predictions. Note any that are different from what you expected. Include your graph in your answer.

Correlation Matrix-

Pearson Correlation Coefficients, N = 47
Prob > |r| under H0: Rho=0

	NEWPOL	PCTMINOR	FIRES	THEFTS	PCTOLD	INCOME
NEWPOL	1.00000	-0.75942 <.0001	-0.68648 <.0001	-0.31162 0.0330	-0.60574 <.0001	0.75098 <.0001
PCTMINOR	-0.75942 <.0001	1.00000	0.59280 <.0001	0.25506 0.0836	0.25051 0.0894	-0.70373 <.0001
FIRES	-0.68648 <.0001	0.59280 <.0001	1.00000	0.55621 <.0001	0.41222 0.0040	-0.61045 <.0001
THEFTS	-0.31162 0.0330	0.25506 0.0836	0.55621 <.0001	1.00000	0.31763 0.0296	-0.17292 0.2451
PCTOLD	-0.60574 <.0001	0.25051 0.0894	0.41222 0.0040	0.31763 0.0296	1.00000	-0.52867 0.0001
INCOME	0.75098 <.0001	-0.70373 <.0001	-0.61045 <.0001	-0.17292 0.2451	-0.52867 0.0001	1.00000

Comparison- From looking at the correlation matrix and comparing it to my predictions they mostly are correct when it comes to the sign of the coefficients. The ones I predicated right were PCTMINOR at negative, PCTOLD at negative, and INCOME at positive.

The ones I got wrong were FIRES which I said was positive, but it is negative and THEFTS which I said was positive but is negative.

c) Run a multiple regression of NEWPOL on the independent variables.

i) Evaluate the overall fit of the regression. Include and interpret any relevant statistics.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	574.55657	114.91131	31.59	<.0001
Error	41	149.16173	3.63809		
Corrected Total	46	723.71830			

F = MSR/MSE

= 114.91131/3.63809 (Values taken in Mean Square)

=31.59 (F Value)

F Value=31.59 and with a p-value at <.0001 less than 0.05 (at alpha=0.05)

The goodness of fit test is testing the overall fit for the regression. Based on the data the F-Value when checking if it is right checks out by my math and the significance backs up this data fits being under the .05 threshold at <.0001.

ii) List the significant predictors given a threshold value of 0.05. Include the values of any relevant statistics.

Initial results parameter estimates:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.06107	2.81878	4.28	0.0001
PCTMINOR	1	-0.05947	0.01318	-4.51	<.0001
FIRES	1	-0.10185	0.04801	-2.12	0.0400
THEFTS	1	0.01356	0.01624	0.84	0.4084
PCTOLD	1	-0.06437	0.01583	-4.07	0.0002
INCOME	1	0.00011636	0.00018040	0.64	0.5225

Based on the threshold of 0.05 looking in the Pr>|t| section for significant predictors is PCTMINOR <.0001, FIRES 0.0400, PCTOLD 0.0002. When checking for the threshold, there are two non-significant predictors INCOME which is above the threshold of 0.05 at 0.5225 it will be removed and run again. For THEFTS while over .05 threshold I want to remove INCOME first since it has a higher value then recheck if THEFTS is still insignificant after first removal.

After removal of INCOME:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.79865	0.82377	16.75	<.0001
PCTMINOR	1	-0.06436	0.01071	-6.01	<.0001
FIRES	1	-0.11216	0.04496	-2.49	0.0166
THEFTS	1	0.01691	0.01528	1.11	0.2746
PCTOLD	1	-0.06941	0.01367	-5.08	<.0001

The now updated significant variables are: PCTMINOR <.0001, FIRES 0.0166, PTOLD <.0001 which are at 0.05 or below. After reexamining THEFTS 0.2746 is still above the .05 threshold which means, it needs to be removed.

After INCOME and THEFTS have been removed:

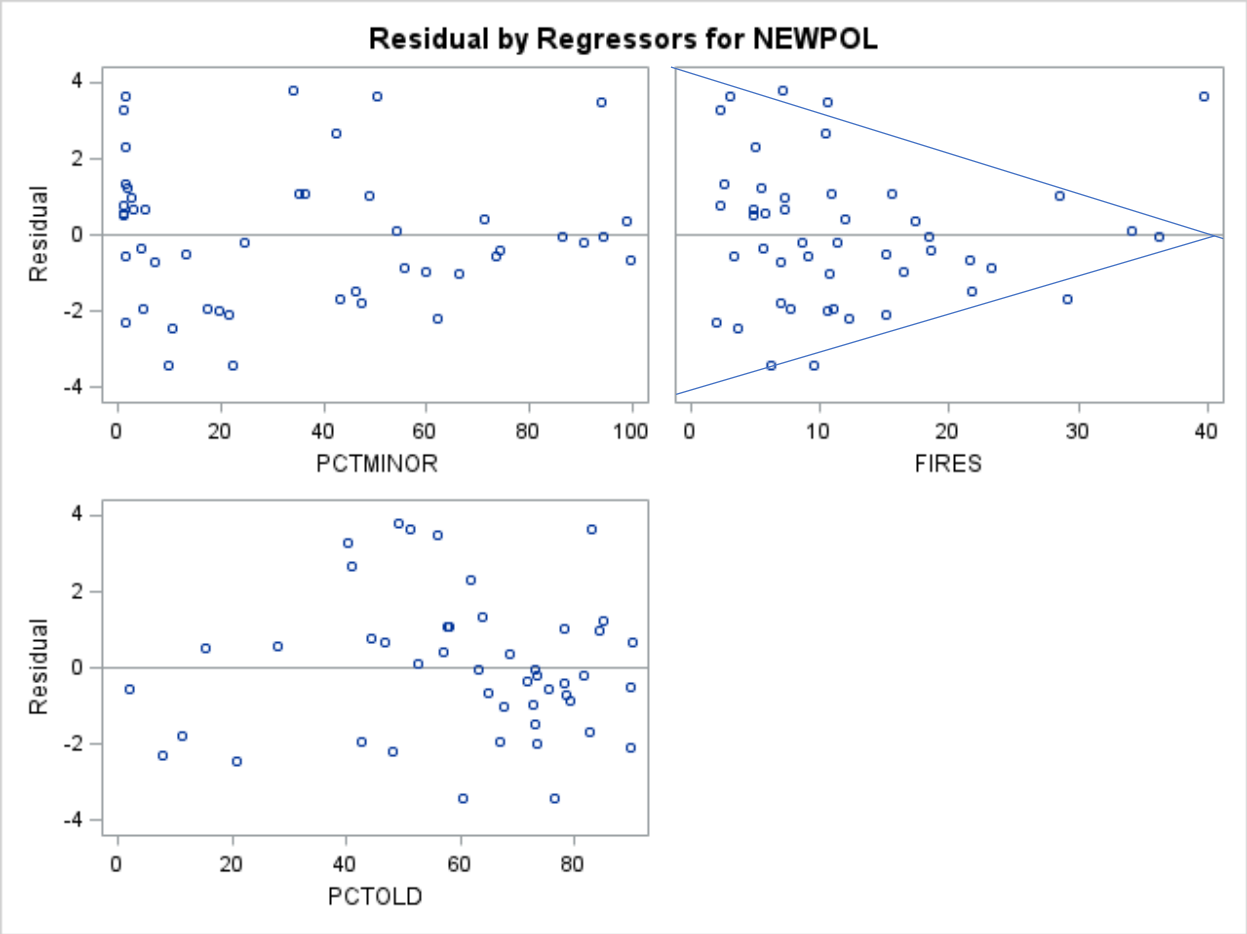
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.99581	0.80639	17.36	<.0001
PCTMINOR	1	-0.06570	0.01067	-6.16	<.0001
FIRES	1	-0.08862	0.03972	-2.23	0.0309
PCTOLD	1	-0.06762	0.01361	-4.97	<.0001

The final updated significant variables are: PCTMINOR <.0001, FIRES 0.0309, and PCTOLD <.0001 which are under 0.05 threshold meaning they are all significant.

iii) Do any of the predictors have signs that are different than the correlation plot suggested? If so, list them and describe why the signs are unexpected.

Out of the remaining significant predictors there are no signs that are different from the correlation plot. Correlation matrix had PCTMINOR, FIRES, and PCTOLD as negative and the final estimates had all three negatives too.

iv) Create a residual plot and examine it. Describe any problems you see and what they might indicate.



When looking at the residual plots of the significant variables the only problem I can tell would be bias in the residual plots. When looking at FIRES it resembles a cone shaped plot where at the beginning all the data is mainly within 20 and the further out it goes very few data points. I was able to draw out a shape on the FIRES which means the data is not evenly spread-out creating bias.

3. The file "maple.csv" includes data from a study of the growth of sugar maple seeds that were collected from various parts of the USA and Canada, then planted in a nursery in Ohio. This dataset contains three variables:

LeafIndex - A measure of how much leafing out has occurred for a seed. A high value indicates advanced leafing out.
JulyTemp - The average temperature in July of the seed's origin
Latitude - The latitude of the seed's origin

a) Find the regression of LeafIndex on Latitude. Is Latitude a significant predictor of leaf index? Why or why not? Comment on the significance of the model as a whole.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.66716	3.05202	-0.55	0.5889
Latitude	1	0.45369	0.07427	6.11	<.0001

Latitude is a significant predictor to LeadIndex based on the threshold set at 0.05. Latitude's significance is <.001 as an individual model.

b) Find the regression of LeafIndex on JulyTemp. Is Latitude a significant predictor of leaf index? Why or why not? Comment on the significance of the model as a whole.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	40.74297	4.45498	9.15	<.0001
JulyTemp	1	-0.33318	0.06206	-5.37	<.0001

JulyTemp is a significant predictor to LeadIndex based on the threshold set at 0.05. JulyTemp's significance is <.001 as an individual model.

c) Find the regression of LeafIndex on both Latitude and JulyTemp. Compare your results to the results in a) and b). What happened to the statistical significance of each independent variable? What about the model overall? How different are the coefficients?

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.73184	11.42026	1.20	0.2389
Latitude	1	0.31393	0.12388	2.53	0.0169
JulyTemp	1	-0.13524	0.09676	-1.40	0.1728

When looking at the significance value of each independent variable when they are all together the significance values of both Latitude and JulyTemp go up from <.0001 to 0.0169 for Latitude and JulyTemp is at 0.1728 instead of <.0001 in their respective individual models.

When it comes to the coefficients value from individual models in a) Latitude was a higher value than in the combined model in c). In a) Latitude's coefficient is 0.45369 and in c) is 0.31393. For JulyTemp individually in a) the coefficient value is -0.3318 and in c) it is higher value at -0.13524.

d) Provide an explanation for what may be causing these results. Name the problem here and why it's having these effects.

The problem with this could be multicollinearity because of the fact the independent variables have high correlation to each other which can make their values inaccurate.

4. Looking beyond the surface

In past courses, you've been tasked with building models around a given parameter of interest (dependent variable). However, often the most interesting trend or variables are not the most obvious. This problem asks you to look below the surface to find a story in the data that is more interesting than the obvious.

The data in "olympics.csv" contains information on the performance of various countries in the 2012 London Summer Olympic Games. Each country is represented with its medal counts, number of athletes by gender, national population figures, and national gross domestic product (GDP).

Your job is to find some sort of interesting trend or message in the data that you could communicate to the general public. You must look beyond the obvious surface messages that larger countries with higher GDPs generally win more medals, and that countries who won more of each type of medal won more medals overall.

It will take some investigation to find that message. You should look at several relationships, and perhaps transform or create new variables based on the original variables before settling on one. Note that there are several opportunities for interesting analyses in this data. Think about whether there is an important trend or lesson that you would like the public to understand? Below are some things to consider. You do not have to investigate all of these. They are provided to help you think about the data.

- Do any surprises emerge? Often, the most interesting results are surprising ones because they tell us things we didn't expect.
- For example, are there ways to evaluate a country's "performance" beyond raw medal counts?
- Remember that transformations don't just remove non-linearities, they can also even-out or account for certain effects in the data. Are there any transformations or ways of combining variables that can reveal more subtle patterns than simply overall population/GDP?
- Is there any relationship between participation of certain demographics and the country's performance?
- Are there any relationships that have nothing do to with medal performance that are interesting or impactful?

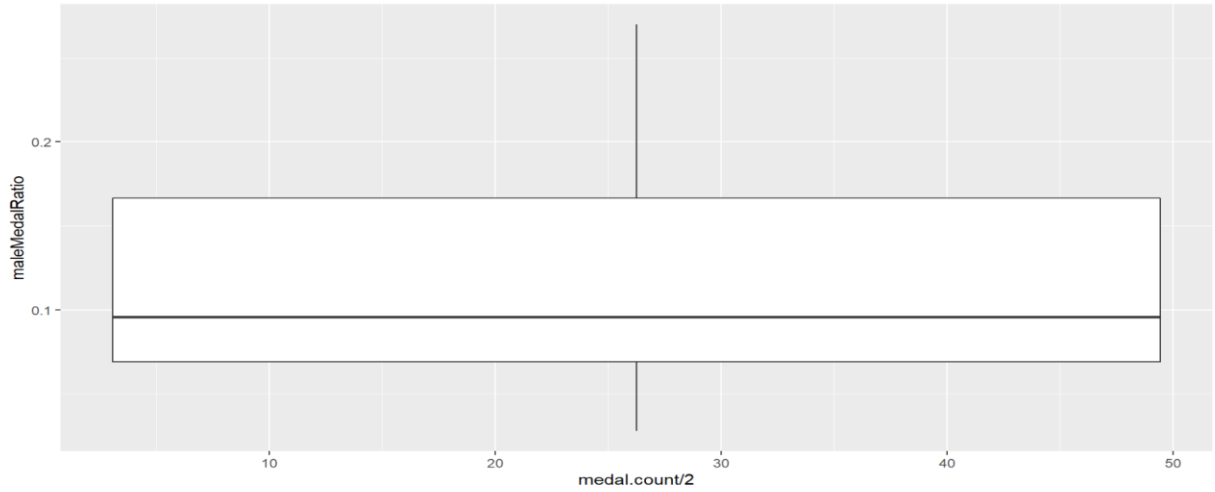
You may dry different multiple-regressions and plots, but your writeup should only include your best, most interesting analysis. Be thorough and be sure to include any graph(s) and statistics you are using to see the relationships. Clearly indicate the message you're trying to tell and why it's significant. There are many possible relationships to consider but you will be graded on the clarity and the thoroughness of your graphs and written analysis. You should be able to fill at least a page.

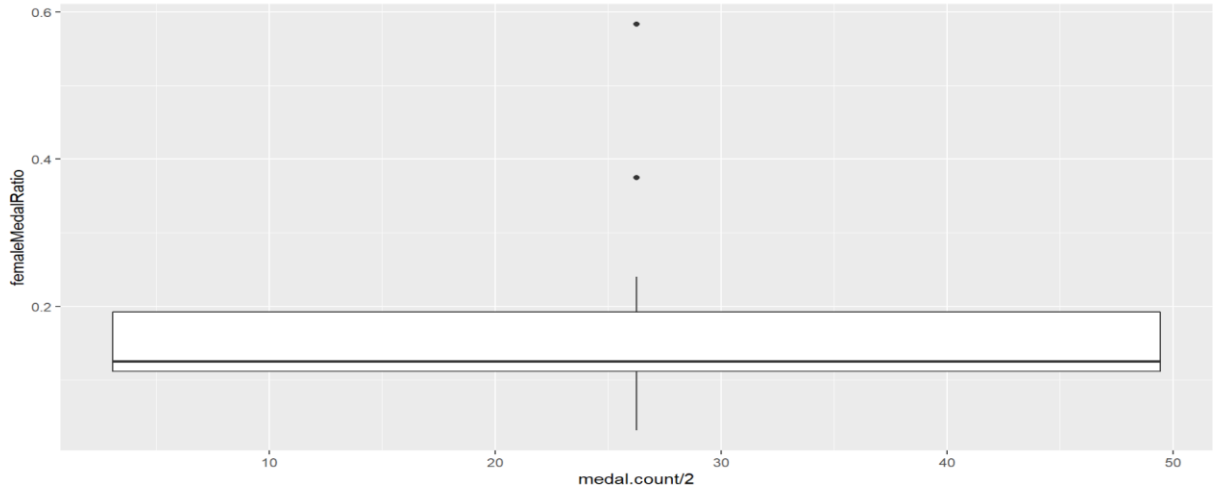
Female.count <int>	Male.count <int>	Gold.medals <int>	Silver.medals <int>	Bronze.medals <int>	medal.count <int>	femaleMedalRatio <dbl>	maleMedalRatio <dbl>
271	260	46	29	29	104	0.19188192	0.20000000
208	163	38	27	23	88	0.21153846	0.26993865
162	141	7	14	17	38	0.11728395	0.13475177
176	219	11	19	14	44	0.12500000	0.10045662
148	187	11	11	12	34	0.11486486	0.09090909
128	138	3	5	9	17	0.06640625	0.06159420
269	287	29	17	19	65	0.12081784	0.11324042
122	159	8	9	11	28	0.11475410	0.08805031
227	208	24	26	32	82	0.18061674	0.19711538
23	60	0	2	4	6	0.13043478	0.05000000
25	25	4	4	4	12	0.24000000	0.24000000
6	29	1	3	3	7	0.58333333	0.12068966
40	15	4	0	2	6	0.07500000	0.20000000
4	21	0	1	2	3	0.37500000	0.07142857
13	16	0	2	3	5	0.19230769	0.15625000

1-15 of 20 rows | 5-12 of 12 columns Previous 1 2 Next

Female.count <int>	Male.count <int>	Gold.medals <int>	Silver.medals <int>	Bronze.medals <int>	medal.count <int>	femaleMedalRatio <dbl>	maleMedalRatio <dbl>
11	15	1	0	0	1	0.04545455	0.03333333
10	12	0	0	2	2	0.10000000	0.08333333
3	13	0	0	1	1	0.16666667	0.03846154
16	18	0	1	0	1	0.03125000	0.02777778
4	6	1	0	0	1	0.12500000	0.08333333

When looking at the olympics.csv file I wanted to look at if I took a specific gender and halved the medals totals how much gender had influence per a countries medal count. The first thing I did was create female and male ratios but would half it assuming males and females evenly split their total medal counts. I also totaled per each country their respective gold, silver, and bronze medals into a count. From there I would divide the medal count by two and per gender (male and female). A few interesting things I found from the data just at a human eye that only five of the twenty countries had a higher male ratio in medals than females. To show as an example how more frequent females won medals look at the bottom 7 countries, they all had ratios higher than males. Countries that stuck out in female's ratio was Georgia at .5833 compared to their males at .1206. Another country's female medal ratio that had more than males that stuck out was Armenia at .375 compared to their males .0714. From there I decided to best show my results in one organized way was to make boxplots.





In my boxplots I found for male medal ratio their Q1 was ~.75, the median was right under 1, ~.9, and their Q3 was ~1.25. In the female medal ratio boxplot, the first thing that stuck out was the outliers I found by eye. When looking at the data it can be assumed it may make the boxplot look not normal instead, they became outliers. What I also noticed was that the females Q1 was about ~1.1, median was ~1.2, and Q3 was ~1.9. What this means and tells me is that females had a much closer range of data where they had a much higher minimum value range than the males.