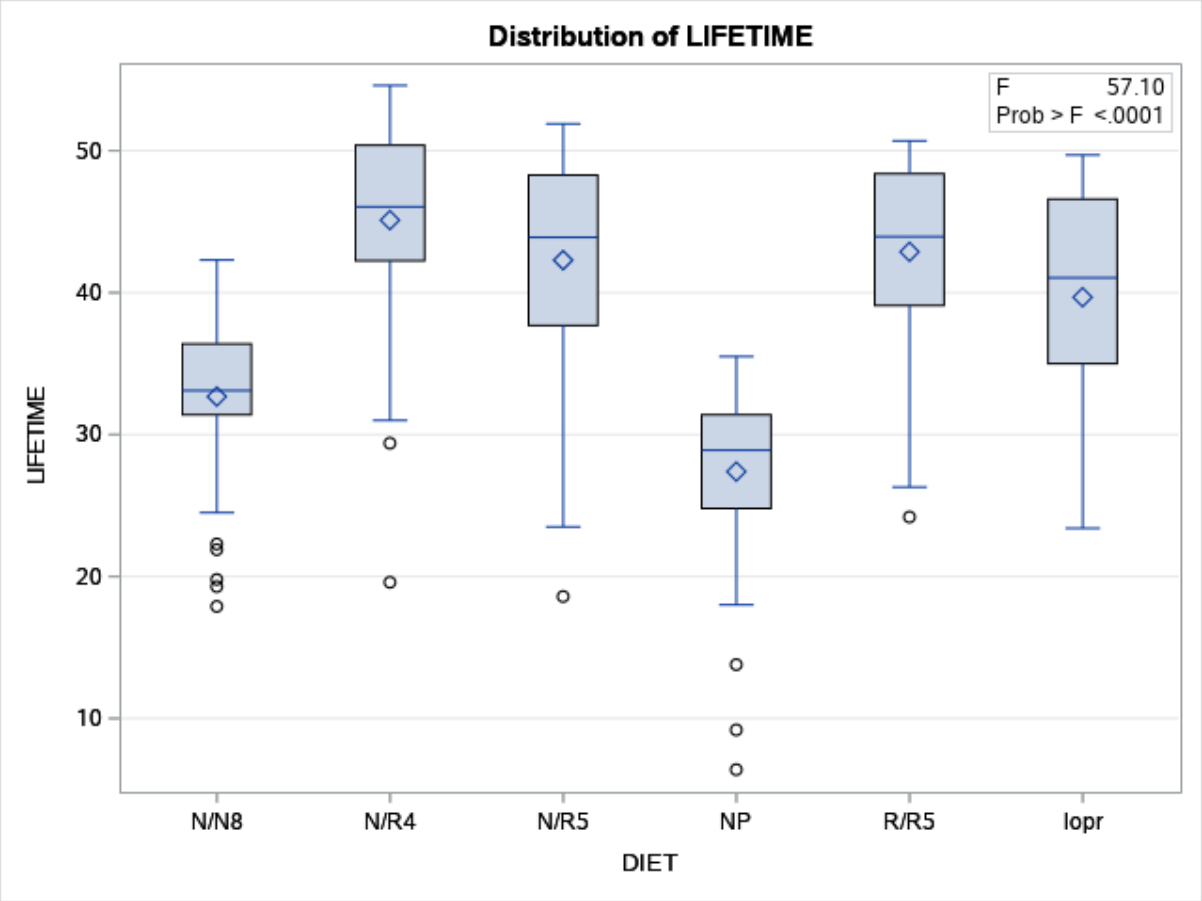
GROUP B
Adrian Harrop, Nikhil Jagatia, Thomas Zee

ASSIGNMENT 6

1. DIET

1. Using the appropriate numerical and visual summaries comment on your data.



DIET	N	Mean	95% Confidence Limits	
N/R4	60	45.1167	43.4209	46.8124
R/R5	56	42.8857	41.1304	44.6410
N/R5	71	42.2972	40.7383	43.8561
lopr	56	39.6857	37.9304	41.4410
N/N8	57	32.6912	30.9514	34.4311
NP	49	27.4020	25.5255	29.2785

The summary of the data of different diets show the variation of the number of entries for each diet and their respective means. Interestingly we can see an overlap in the confidence limits of the mean between N/R40, R/R50 , N/R50 and lopro. By contrast N/N85 and NP do not share any overlaps.
Using the visual representation of the box-plots shows how the median values are similar for the overlapping mean diets, in addition they are positively skewed and have seemingly similar interquartile ranges, with N/R50 having the greatest range identified by its whiskers. By contrast the boxplots for N/N85 and NP are more symmetrical. The only diet with no outliers is the lopro, yet NP and N/N85 have many more in comparison, therefore as the shortest overall boxplot the diet which data has the greatest accuracy is NP diet.
The boxplot suggests the greatest diet fluctuations were from N/R50 and N/R40 which could suggests that intermittent calorie cycling causes a consequential weight change.

GROUP B
Adrian Harrop, Nikhil Jagatia, Thomas Zee

```
PROC   ANOVA  DATA=DIET;
    CLASS     DIET;
    MODEL     LIFETIME=DIET;
    MEANS   DIET   /   HOVTEST=BF   HOVTEST=BARTLETT   CLDIFF   TUKEY;
    MEANS     DIET/CLM      T;
RUN;
```

Equality of Variances
H_0 : Variances of each diet are the same:
var NP = var N/N85 = var N/R50 = var R/R50 = var lopro = var N/R40
H_1: at least one is different

The ANOVA Procedure

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
DIET	5	39307.0	7861.4	1.41	0.2216
Error	343	1918361	5592.9		

Brown and Forsythe's Test for Homogeneity of LIFETIME Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
DIET	5	252.8	50.5557	2.72	0.0199
Error	343	6372.3	18.5781		

Bartlett's Test for Homogeneity of LIFETIME Variance

Source	DF	Chi-Square	Pr > ChiSq
DIET	5	10.9962	0.0515

Running an ANOVA procedure for the three tests, Levene, Brown and Forsythe and Bartlett.

```
1  FILENAME REFFILE '/folders/myshortcuts/MyFolders/diet.csv';
2
3  PROC IMPORT DATAFILE=REFFILE
4          DBMS=CSV
5          OUT=MICE;
6          GETNAMES=YES;
7  RUN;
8
9  PROC CONTENTS DATA=MICE;
10 RUN;
11
12 PROC ANOVA DATA= MICE;
13 CLASS DIET;
14 MODEL LIFETIME = DIET;
15 MEANS DIET/ HOVTEST = LEVENE HOVTEST = BF HOVTEST = BARTLETT;
16 RUN;
17
```

GROUP B
Adrian Harrop, Nikhil Jagatia, Thomas Zee

Looking at the box plots produced in Table 1, two diets (NP and N/N8 seem to have different variances (size of box) than the others. There are a number of outliers that we need to consider.
In deciding which test to use, we have chosen to look at the p value for Brown and Forsythe. This uses the median when applying the tests and performs best when the data is skewed. As the data for some of the most of the diets has outliers, it has been shown that using the median is recommended against non-normal data and can be used with confidence if you are unsure of the underlying distribution of the data. The p value of 0.0199, which is less than 0.05. There is evidence to reject the null hypothesis of equal variances, and this suggest that the variances of the diets differ for at least one of them.

2. Conduct a test for the overall equality of the means. State your hypothesis, conduct the test and comment on the results.

   $H_0 : \mu_{indifferent}$ = The means of each diet are equal = 0
   $H_1 : \mu_{different}$ = The means are not equal for at least 1 diet ≠ 0

Comparisons significant at the 0.05 level are indicated by *.**				
DIET Comparison	**Difference Between Means**	**95% Confidence Limits**		
N/R4 - R/R5	2.231	-0.210	4.672	
N/R4 - N/R5	2.819	0.516	5.123	***
N/R4 - lopr	5.431	2.990	7.872	***
N/R4 - N/N8	12.425	9.996	14.855	***
N/R4 - NP	17.715	15.185	20.244	***
R/R5 - N/R4	-2.231	-4.672	0.210	
R/R5 - N/R5	0.589	-1.759	2.936	
R/R5 - lopr	3.200	0.718	5.682	***
R/R5 - N/N8	10.194	7.723	12.666	***
R/R5 - NP	15.484	12.914	18.053	***
N/R5 - N/R4	-2.819	-5.123	-0.516	***
N/R5 - R/R5	-0.589	-2.936	1.759	
N/R5 - lopr	2.611	0.264	4.959	***
N/R5 - N/N8	9.606	7.270	11.942	***
N/R5 - NP	14.895	12.456	17.335	***
lopr - N/R4	-5.431	-7.872	-2.990	***
lopr - R/R5	-3.200	-5.682	-0.718	***
lopr - N/R5	-2.611	-4.959	-0.264	***

lopr - N/N8	6.994	4.523	9.466	***
lopr - NP	12.284	9.714	14.853	***
N/N8 - N/R4	-12.425	-14.855	-9.996	***
N/N8 - R/R5	-10.194	-12.666	-7.723	***
N/N8 - N/R5	-9.606	-11.942	-7.270	***
N/N8 - lopr	-6.994	-9.466	-4.523	***
N/N8 - NP	5.289	2.730	7.848	***
NP - N/R4	-17.715	-20.244	-15.185	***
NP - R/R5	-15.484	-18.053	-12.914	***
NP - N/R5	-14.895	-17.335	-12.456	***
NP - lopr	-12.284	-14.853	-9.714	***
NP - N/N8	-5.289	-7.848	-2.730	***

```
PROC GLM DATA = DIET;
CLASS DIET;
MODEL LIFETIME = DIET;
MEANS DIET/ CLDIFF T;
RUN;
```

Here choosing the GLM method because there is more than 2 variables to consider for the comparison of the overall equality of the means. From the table generated above it shows that the majority of means in comparison to each respective diet are unequal at the 0.05 significance level. The N/R40 against R/R50 and R/R50 against N/R50 were the only diets that has significantly equal means. In addition, found by the ANOVA procedure the P value overall found was <0.0001 which therefore means we must reject the null hypothesis of overall equal means. There's evidence to suggest that the mean longevity gains from each diet are significantly different.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	12733.94181	2546.78836	57.10	<.0001

```
PROC ANOVA DATA = DIET;
CLASS DIET;
MODEL LIFETIME = DIET;
MEANS DIET/ CLDIFF T;
RUN;
```

3. Using the appropriate comparisons method, use the NP group as your reference point (control) and check if compared to NP:
   a) Reducing the diet to 85 kcal has an effect

GROUP B
Adrian Harrop, Nikhil Jagatia, Thomas Zee

Comparisons significant at the 0.05 level are indicated by *.**

Comparison	Difference Between Means	95% Confidence Limits		
NP - N/N8	-5.289	-7.848	-2.730	***

This portion of the table shows that reducing the diet to 85kcal did have a significant effect at the 0.05 level where 0 is not contained in the confidence interval limits, showing a significant difference.

b) Reducing the diet to 40 kcal has an effect.

Comparisons significant at the 0.05 level are indicated by *.**

Comparison	Difference Between Means	95% Confidence Limits		
NP - N/R4	-17.715	-20.244	-15.185	***

Similarly to 85kcal, reducing the diet to 40kcal also has a significant effect compared to the NP group where 0 is also not contained within the limits indicating a significant effect in longevity due to reducing the diet to 40kcal.

```
PROC ANOVA DATA=DIET;
   CLASS DIET;
   MODEL LIFETIME = DIET;
   MEANS DIET / CLDIFF T;
Run;
```

4. Select the appropriate groups (hint: do not use NP) to check if:

a) Pre-weaning dietary restrictions have an effect.

Because we are only interested in Pre-weaning being used as a reference point of comparison we need to choose the R/R50 as the base.

Comparisons significant at the 0.05 level are indicated by *.**

Comparison	Difference Between Means	95% Confidence Limits		
R/R5 - N/R4	-2.231	-5.787	1.325	
R/R5 - N/R5	0.589	-2.832	4.009	
R/R5 - lopr	3.200	-0.417	6.817	
R/R5 - N/N8	10.194	6.593	13.796	***

R/R5 - NP	15.484	11.740	19.228	***

Here we can see that the three diets that do not have a significant effect and contain 0 within the 95% confidence intervals are N/R50 (mice were fed normally before weaning and were given 50 kcal per week after weaning.) and R/R50 (mice were restricted to 50kcal before weaning; and after weaning, their caloric intake was also 50 kcal per week.) .

b) Restricting the protein intake has an effect.

From the table below we can see that by restricting the diet so that lopro is compared to N/R50 which does not have a difference where lopro is chosen initially as the control. Therefore the confidence limit includes 0, this suggests there's no evidence that longevity is increased as they get older.

Comparisons significant at the 0.05 level are indicated by *.**

Comparison	Difference Between Means	95% Confidence Limits		
lopr - N/R4	-5.431	-8.987	-1.875	***
lopr - R/R5	-3.200	-6.817	0.417	
lopr - N/R5	-2.611	-6.032	0.809	
lopr - N/N8	6.994	3.393	10.596	***
lopr - NP	12.284	8.540	16.028	***

```
PROC ANOVA DATA=DIET;
   CLASS DIET;
   MODEL LIFETIME= DIET;
   MEANS DIET / HOVTEST=BF HOVTEST=BARTLETT CLDIFF TUKEY;
   MEANS DIET / CLM T;
RUN;
```

GROUP B
Adrian Harrop, Nikhil Jagatia, Thomas Zee

Q2 GOLF.

H_0: The null hypothesis is the median of the differences of the average scores between round 2 and round 3 is 0 $\mu=0$

H_1: The alternative hypothesis is the median of the differences of the average scores between round 2 and round 3 is bigger than 0 >0

We will be using a one tailed t-test because we are only interested if the difference is at least 3 or higher, so we will be testing if there is a significant difference greater than 0.

Basic Statistical Measures			
Location		**Variability**	
Mean	5.600000	**Std Deviation**	3.77712
Median	5.000000	**Variance**	14.26667
Mode	5.000000	**Range**	13.00000
		Interquartile Range	1.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	4.688423	**Pr > |t|**	0.0011
Sign	**M**	4	**Pr >= |M|**	0.0215
Signed Rank	**S**	26.5	**Pr >= |S|**	0.0039

Using the SAS code below, we have the 2 tables above. If we look at the signed ranked test, the two tailed p value is 0.0039, therefore the one tailed p value is 0.0039/2 = 0.00195. Since 0.00195 < 0.05, there is sufficient evidence to reject the null hypothesis. This means that the differences between the two rounds (round 3 – round 2) is bigger than 0, which is the alternative hypothesis. Now that we have confirmed that the difference is bigger than 0, we look at the statistical measures table. The median is 5, which is bigger than 3, so the data given does support the commentators claim, that the average scores of players were likely to be at least three higher than those for the second round.

GROUP B
Adrian Harrop, Nikhil Jagatia, Thomas Zee

```
DATA GOLF;
INPUT PLAYER$ ROUND2 ROUND3;
DATALINES;
A 73 72
B 73 79
C 74 79
D 66 77
E 71 83
F 73 78
G 68 70
H 72 78
I 73 78
J 72 77
;
RUN;
PROC PRINT DATA=GOLF;
RUN;
DATA GOLF2;
SET GOLF;
DIFF=ROUND3 - ROUND2;
RUN;
PROC PRINT DATA=GOLF2;
RUN;
/* H_0: MEDIAN OF ROUND3 - ROUND2 = 0
   H_1: MEDIAN OF ROUND3 - ROUND2 > 0
*/
ODS RTF FILE = "ASSIGNMENTQ2.RTF";
PROC UNIVARIATE DATA=GOLF2 LOCCOUNT;
VAR DIFF;
RUN;
ODS RTF CLOSE;
```

Q3 BIRTHDAYS.

1)

H_0 = no difference in the birth rate between days of the week that shows consistency.
H_1 = there is a difference in the birth rate between days of the week, so it is not consistent.

Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)				
Statistic	**Alternative Hypothesis**	**DF**	**Value**	**Prob**
1	**Nonzero Correlation**	1	2.5247	0.1121
2	**Row Mean Scores Differ**	6	7.6909	0.2616

The p value is 0.2616 > 0.05 so we do not reject the null hypothesis. This means there is sufficient evidence to suggest there is no difference in birth rate between days of the week that shows consistency.

GROUP B
Adrian Harrop, Nikhil Jagatia, Thomas Zee

```
FILENAME                                                    REFFILE
'H:\Documents\MASTERS\STATS\COURSEWORK\Assignment6\Q3.CSV';

PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=Q3;
    GETNAMES=YES;
     RUN;
RUN;
ODS RTF FILE = "ASSIGNMENTQ3.RTF";
PROC FREQ DATA=Q3;
     TABLES WEEK*DAY*BIRTH / CMH2 SCORES=RANK NOPRINT;
     RUN;
ODS RTF CLOSE;
```

2)

H_0 = no difference in the birth rate between the 10th, 20th, 30th and 40th weeks.
H_1 = there is a difference in the birth rate between the 10th, 20th, 30th and 40th weeks.

Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)				
Statistic	**Alternative Hypothesis**	**DF**	**Value**	**Prob**
1	**Nonzero Correlation**	1	7.0588	0.0079
2	**Row Mean Scores Differ**	3	9.8382	0.0200

Looking at the p value of 0.02, which is < 0.05, we have enough evidence to reject the null hypothesis. Therefore, we conclude a difference in the birth rate between the 10th, 20th, 30th and 40th weeks showing significant inconsistencies.

```
FILENAME                                                    REFFILE
'H:\Documents\MASTERS\STATS\COURSEWORK\Assignment6\Q3.CSV';

PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=Q3;
    GETNAMES=YES;
     RUN;
RUN;
ODS RTF FILE = "ASSIGNMENTQ3.RTF";
PROC FREQ DATA=Q3;
     TABLES DAY*WEEK*BIRTH / CMH2 SCORES=RANK NOPRINT;
     RUN;
ODS RTF CLOSE;
```