STAT 448 Homework1

Name: Zixin Ouyang

Exercise1

(a)

Variable: Weight

Moments			
N	110	**Sum Weights**	110
Mean	415.730909	**Sum Observations**	45730.4
Std Deviation	319.994944	**Variance**	102396.764
Skewness	0.5688731	**Kurtosis**	-1.0364528
Uncorrected SS	30172788.1	**Corrected SS**	11161247.3
Coeff Variation	76.9716509	**Std Error Mean**	30.5103208

Basic Statistical Measures			
Location		**Variability**	
Mean	415.7309	**Std Deviation**	319.99494
Median	295.0000	**Variance**	102397
Mode	700.0000	**Range**	1100
		Interquartile Range	555.00000

The mean and median of weight is 415.73 and 295, with standard deviation 319.99. Since mean is greater than median, the distribution is likely to have a long tail to the right. The negative skewness 0.57 supports it. Range is summarized as 1100.

(b)

1. Species=Bream

Moments			
N	34	**Sum Weights**	34
Mean	626	**Sum Observations**	21284
Std Deviation	206.604585	**Variance**	42685.4545
Skewness	0.19191733	**Kurtosis**	-0.7105166
Uncorrected SS	14732404	**Corrected SS**	1408620
Coeff Variation	33.0039273	**Std Error Mean**	35.432394

Basic Statistical Measures			
Location		**Variability**	
Mean	626.0000	**Std Deviation**	206.60459
Median	615.0000	**Variance**	42685
Mode	500.0000	**Range**	758.00000
		Interquartile Range	245.00000

2. Species=Perch

Moments			
N	56	**Sum Weights**	56
Mean	382.239286	**Sum Observations**	21405.4
Std Deviation	347.617717	**Variance**	120838.077
Skewness	0.84436769	**Kurtosis**	-0.869673
Uncorrected SS	14828079.1	**Corrected SS**	6646094.25
Coeff Variation	90.9424359	**Std Error Mean**	46.4523714

Basic Statistical Measures			
Location		**Variability**	
Mean	382.2393	**Std Deviation**	347.61772
Median	207.5000	**Variance**	120838
Mode	130.0000	**Range**	1094
		Interquartile Range	575.00000

3. Species=Roach

Moments			
N	20	**Sum Weights**	20
Mean	152.05	**Sum Observations**	3041
Std Deviation	88.828916	**Variance**	7890.57632
Skewness	0.97149967	**Kurtosis**	1.76776364
Uncorrected SS	612305	**Corrected SS**	149920.95
Coeff Variation	58.4208589	**Std Error Mean**	19.8627495

Basic Statistical Measures			
Location		**Variability**	
Mean	152.0500	**Std Deviation**	88.82892
Median	147.5000	**Variance**	7891
Mode	120.0000	**Range**	390.00000
		Interquartile Range	76.00000

Now we repeat the analysis by species. Firstly mean values for Bream, Perch and Roach are 626, 382.24 and 152.05, and the median values are 615, 347.62 and 147.5, respectively. The Bream has heavier weight than Perch, while Perch has heavier weight than Roach. The standard deviation and range for Bream, Perch and Roach are (206.6, 758), (347.62, 1094) and (88.83, 390), respectively. It can be seen that fish species with larger weight has larger standard deviation and range. There seems to be greater variation within Bream and Perch as compared to Roach.

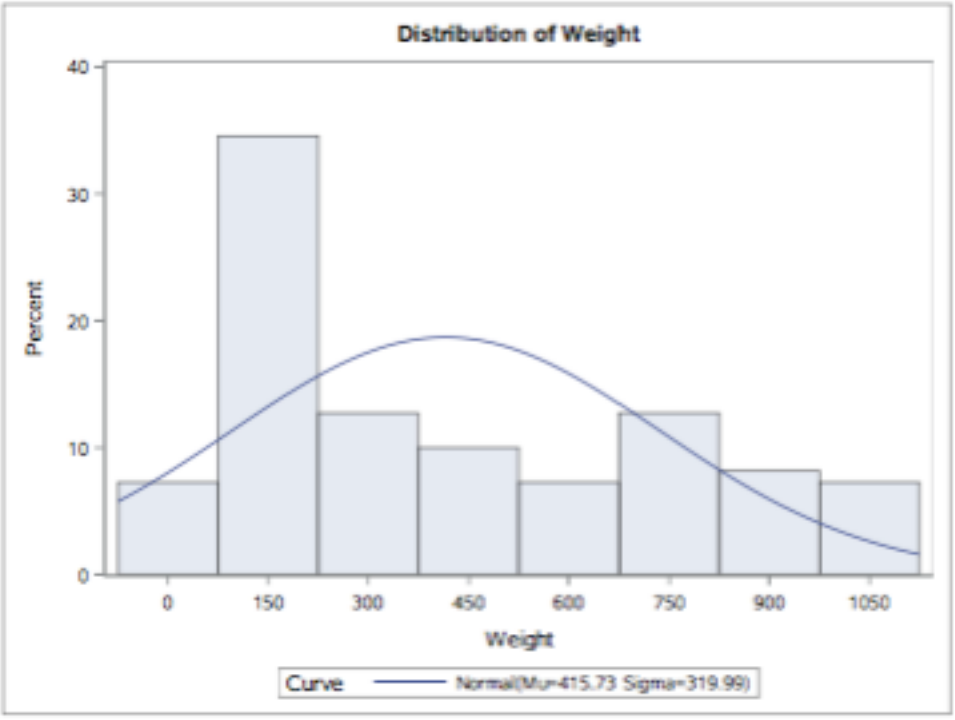
For the skewness, Bream, Perch and Roach have skewness measure as 0.19, 0.84 and 0.97. Generally, positive measure indicates right long tail, and right-skewness.

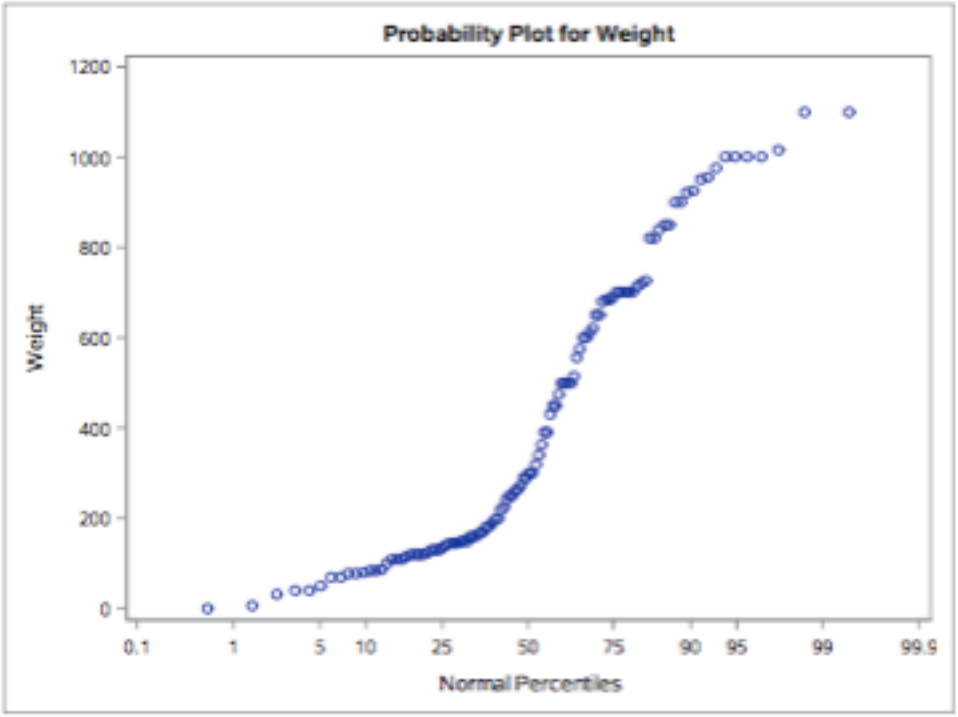
Exercise2

(a)

Variable: Weight

Tests for Normality				
Test		**Statistic**		**p Value**
Shapiro-Wilk	**W**	0.894113	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.159381	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.722376	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	4.282585	**Pr > A-Sq**	<0.0050

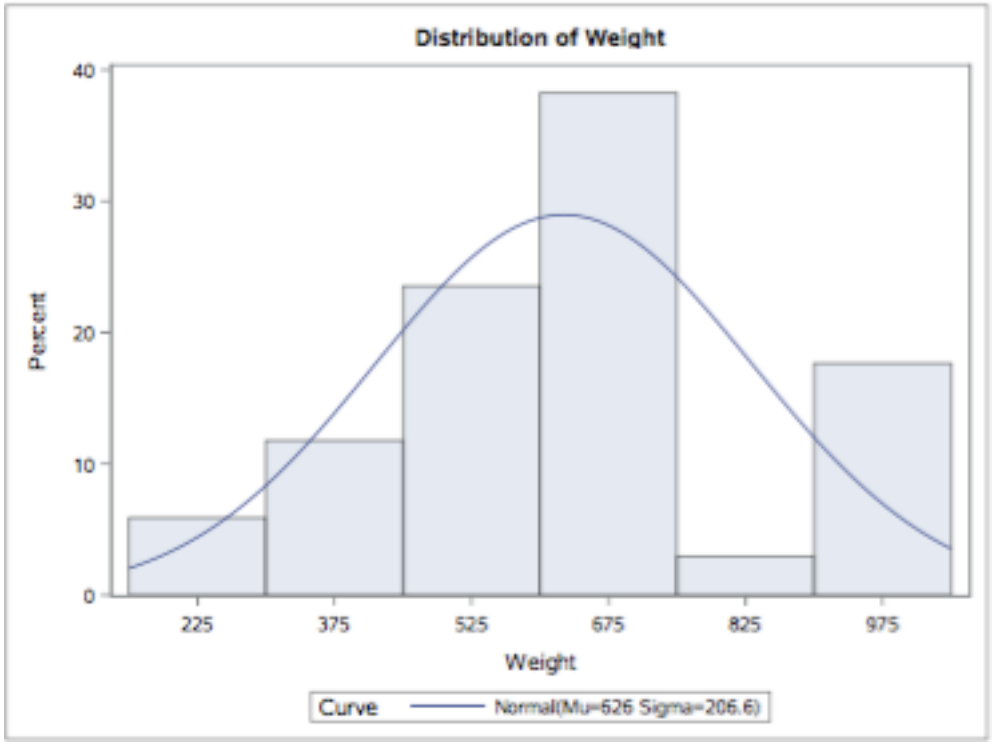


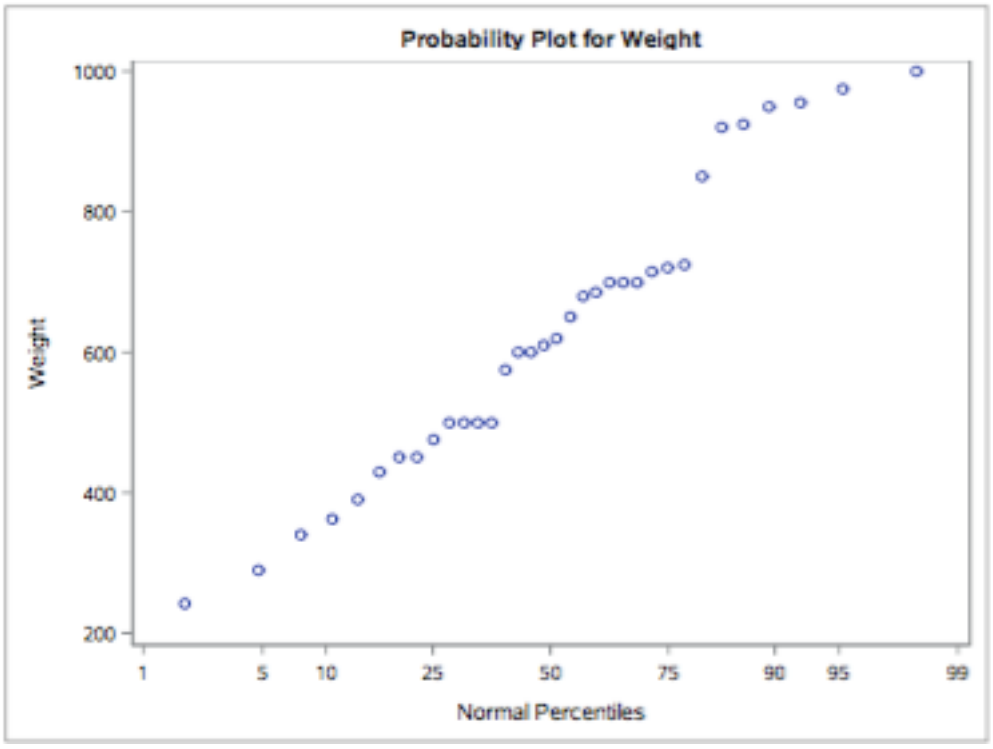


Weight does not seem to follow normal distribution. We can see the histogram is not symmetric and the line in the QQ-plot is not straight. The quantitative test results agree. Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling test (H0: normal distribution fits data) show p-value less than 0.05, thus null hypothesis is rejected. We should not assume normality for the weight.

(b)
1. Species=Bream

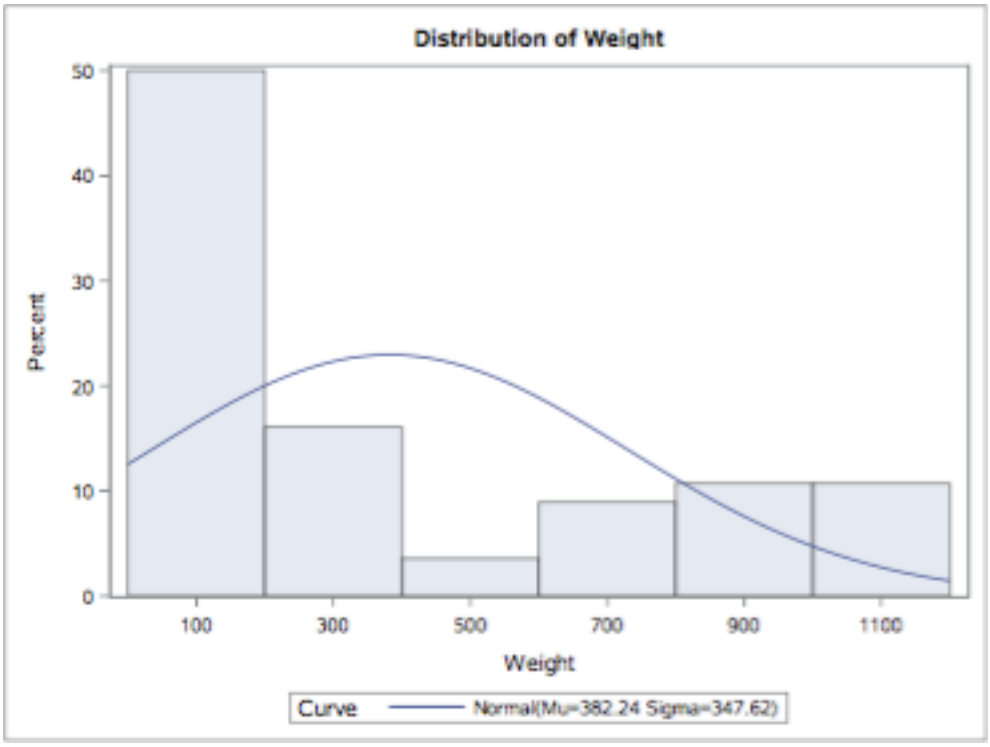
Tests for Normality				
Test		**Statistic**		**p Value**
Shapiro-Wilk	**W**	0.961518	**Pr < W**	0.2691
Kolmogorov-Smirnov	**D**	0.111376	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.058693	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.423242	**Pr > A-Sq**	>0.2500

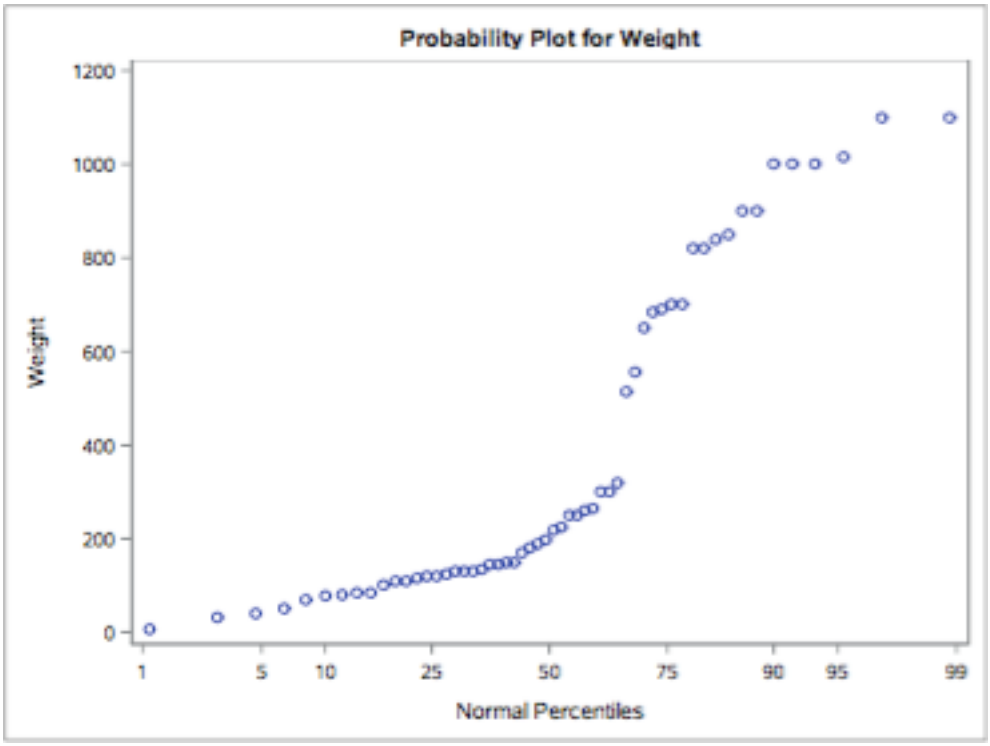




2. Species=Perch

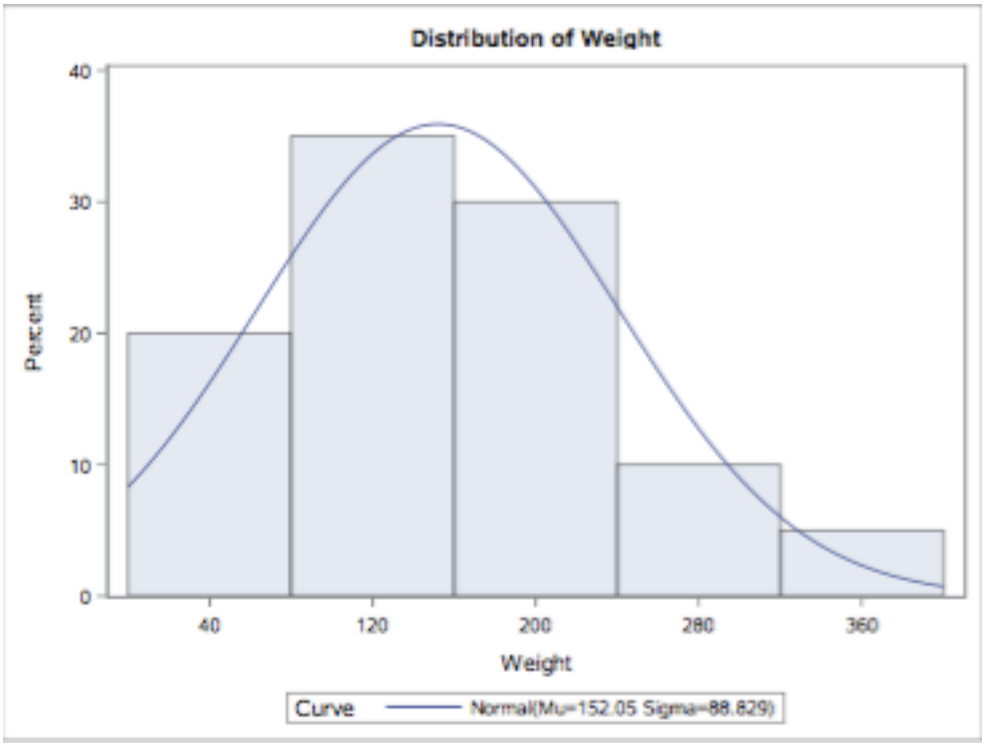
Tests for Normality				
Test		**Statistic**		**p Value**
Shapiro-Wilk	**W**	0.816849	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.239184	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.810687	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	4.362138	**Pr > A-Sq**	<0.0050

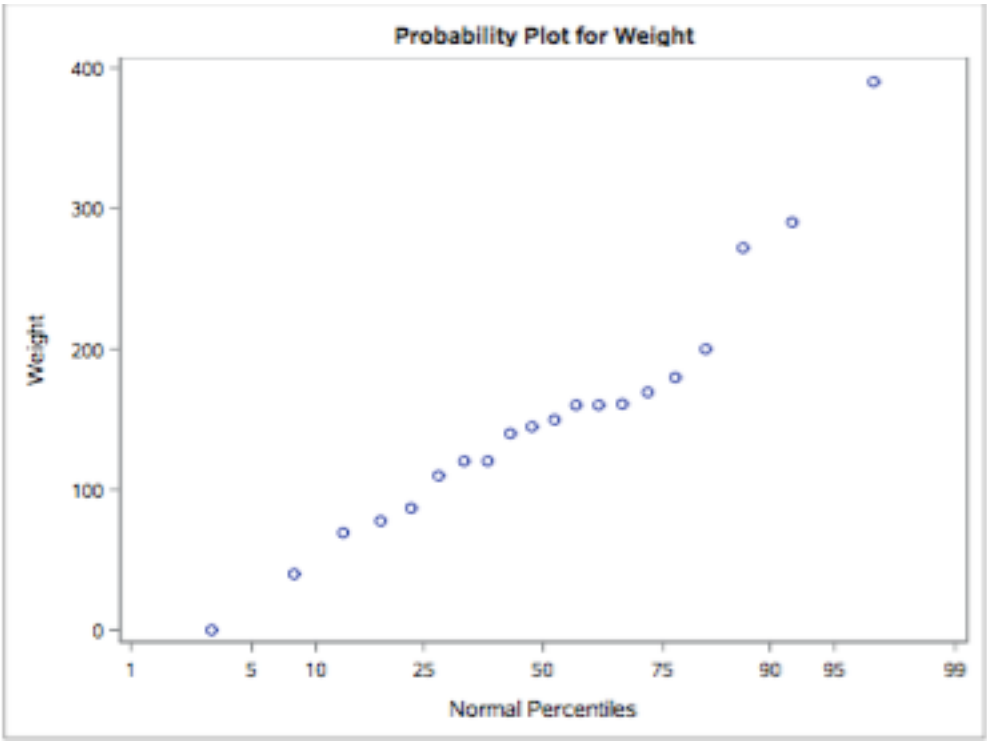




3.Species=Roach

Tests for Normality				
Test		**Statistic**		**p Value**
Shapiro-Wilk	**W**	0.932277	**Pr < W**	0.1708
Kolmogorov-Smirnov	**D**	0.176514	**Pr > D**	0.0986
Cramer-von Mises	**W-Sq**	0.102833	**Pr > W-Sq**	0.0968
Anderson-Darling	**A-Sq**	0.572359	**Pr > A-Sq**	0.1243





For Bream and Roach, All test p-values are greater than 0.05, so conclude that we do not have evidence to reject normality assumption. Although histograms do not have perfect symmetric shapes, they are not very bad. The observations in the QQ-plots are almost in a line.

For Perch, distribution of weight shows statistically significant difference from normal distribution. All test results give p- values less than 0.05, so we reject normality for Perch.

In conclusion, the distribution of weight for Perch is significantly different from normal distribution, but the distributions of weight for two other species, Bream and Roach, are normal.

Exercise 3

(a)

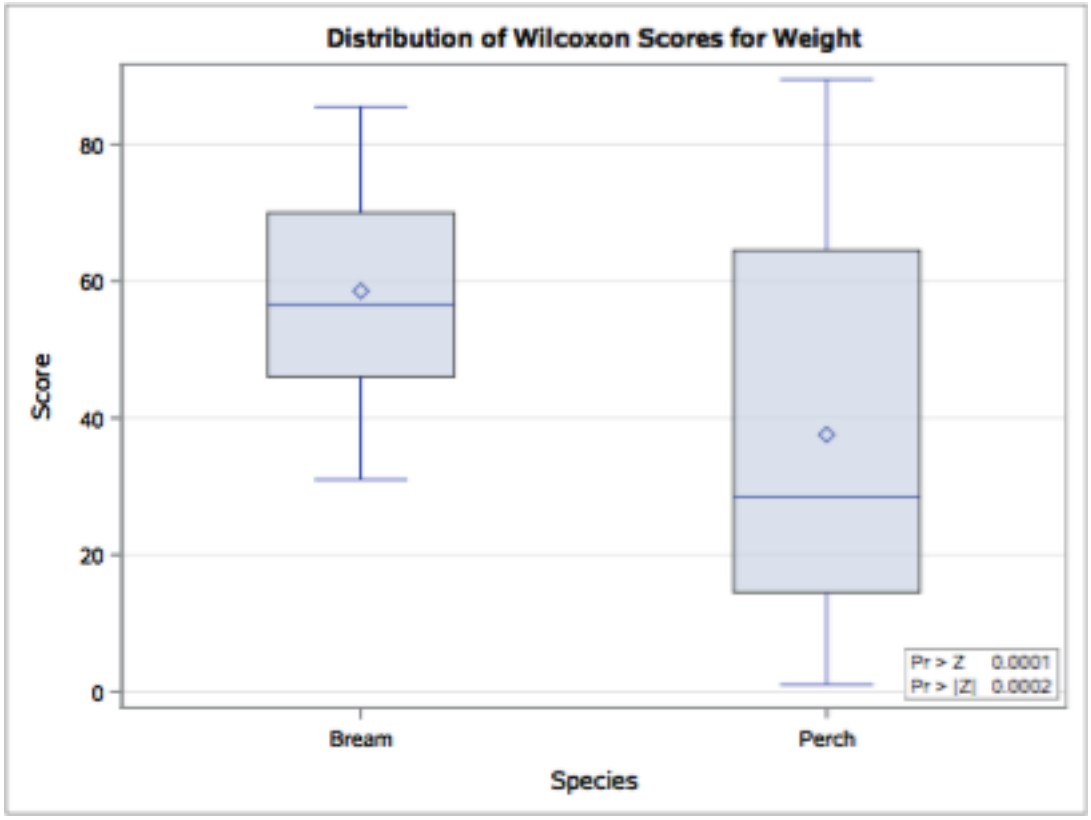
Wilcoxon Scores (Rank Sums) for Variable Weight Classified by Variable Species					
Species	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
Bream	34	1991.0	1547.0	120.131821	58.558824
Perch	56	2104.0	2548.0	120.131821	37.571429
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	1991.0000
Normal Approximation	
Z	3.6918
One-Sided Pr > Z	0.0001
Two-Sided Pr > |Z|	0.0002
t Approximation	
One-Sided Pr > Z	0.0002
Two-Sided Pr > |Z|	0.0004
Z includes a continuity correction of 0.5.	

In Exercise 2(b) we found that weight for Bream can be assumed as normal, but weight for Perch was significantly different from normal. Since we do not have normality for one group, we perform non-parametric tests to compare weight of two groups. Wilcoxon one-sided two-sample test gives p-value less than 0.05, and we can conclude that Bream fish have significantly heavier weight than Perch fish.

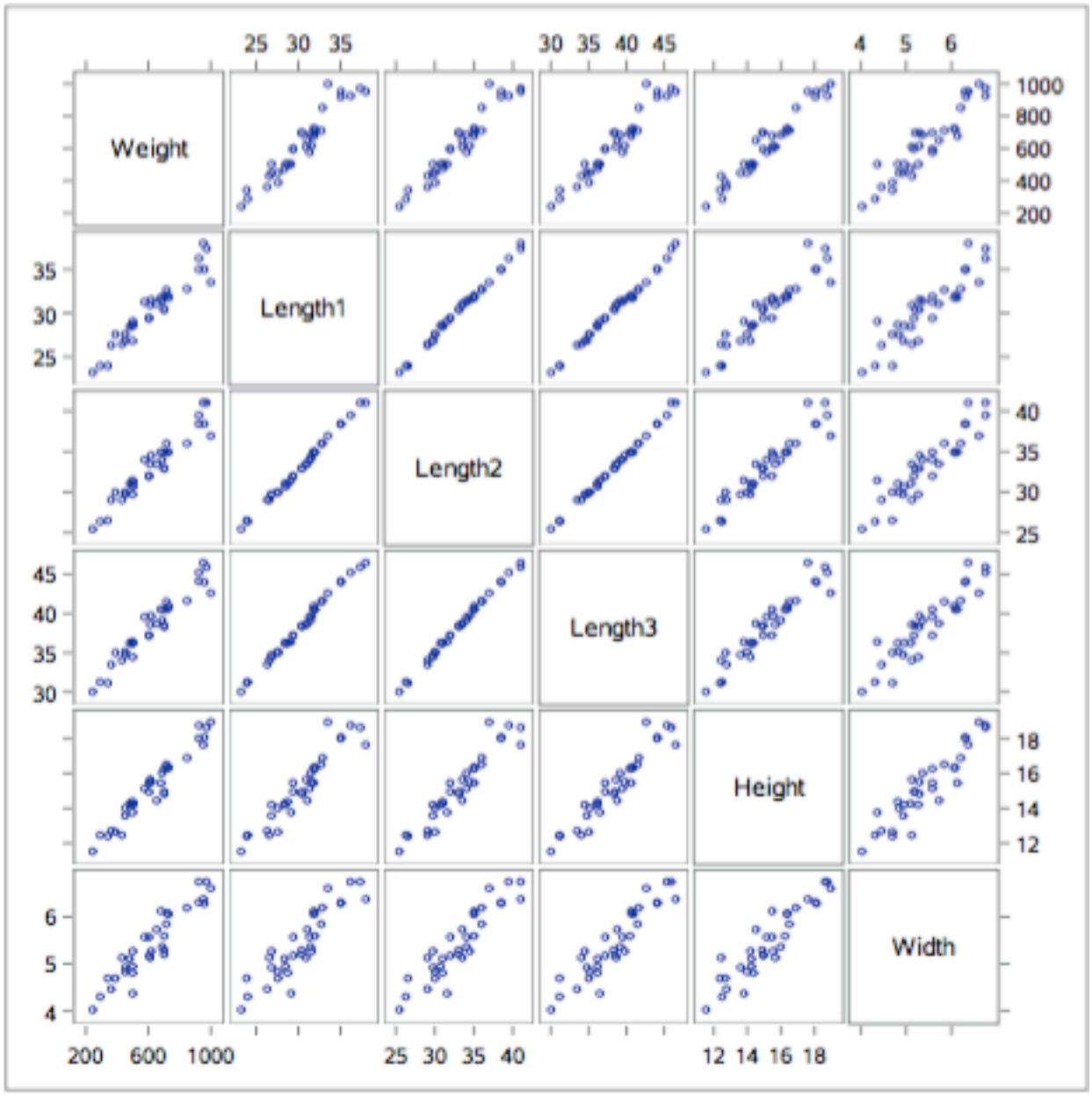
We can see this in the box plot as well, which shows generally higher ranks for Bream fish. Since the question asked about whether Bream's weight is greater than Perch's, we use the one-sided normal approximation (since the sample size is greater than 30). The p-value from the one-sided t-approximation is <0.0001, therefore we reject the null

hypothesis that the two species' weight come from the same distribution, and conclude that Bream's weight is stochastically heavier than Perch's.



Exercise 4

From the scatter plot, we can find linear relationships between each variables. all pairs of variables show strong linear associations. In addition, (length1, length2), (length2, length3) and (length1, length3) shows nearly perfect linear relationships.


Weight
Length1
Length2
Length3
Height
Width
25 30 35
30 35 40 45
4 5 6
1000
800
600
400
200
35
30
25
40
35
30
25
45
40
35
30
18
16
14
12
6
5
4
200 600 1000
25 30 35 40
12 14 16 18

Exercise 5

Pearson Correlation Coefficients, N = 34 Prob > |r| under H0: Rho=0						
	Weight	**Length1**	**Length2**	**Length3**	**Height**	**Width**
Weight	1.00000	0.95458 <.0001	0.96204 <.0001	0.96394 <.0001	0.97080 <.0001	0.93415 <.0001
Length1	0.95458 <.0001	1.00000	0.99776 <.0001	0.99636 <.0001	0.94169 <.0001	0.89994 <.0001
Length2	0.96204 <.0001	0.99776 <.0001	1.00000	0.99825 <.0001	0.95204 <.0001	0.91597 <.0001
Length3	0.96394 <.0001	0.99636 <.0001	0.99825 <.0001	1.00000	0.95492 <.0001	0.92171 <.0001
Height	0.97080 <.0001	0.94169 <.0001	0.95204 <.0001	0.95492 <.0001	1.00000	0.92632 <.0001
Width	0.93415 <.0001	0.89994 <.0001	0.91597 <.0001	0.92171 <.0001	0.92632 <.0001	1.00000

Spearman Correlation Coefficients, N = 34 Prob > |r| under H0: Rho=0						
	Weight	**Length1**	**Length2**	**Length3**	**Height**	**Width**
Weight	1.00000	0.96013 <.0001	0.95748 <.0001	0.95730 <.0001	0.95747 <.0001	0.92841 <.0001
Length1	0.96013 <.0001	1.00000	0.99809 <.0001	0.99702 <.0001	0.96775 <.0001	0.92327 <.0001
Length2	0.95748 <.0001	0.99809 <.0001	1.00000	0.99809 <.0001	0.96779 <.0001	0.92387 <.0001
Length3	0.95730 <.0001	0.99702 <.0001	0.99809 <.0001	1.00000	0.96469 <.0001	0.92586 <.0001
Height	0.95747 <.0001	0.96775 <.0001	0.96779 <.0001	0.96469 <.0001	1.00000	0.91688 <.0001
Width	0.92841 <.0001	0.92327 <.0001	0.92387 <.0001	0.92586 <.0001	0.91688 <.0001	1.00000

Firstly, both Pearson and Spearman correlations show that all pairs of variables have statistically significant correlations. And all have positive relationships with positive Pearson and Spearman coefficients.

Secondly, when it comes to the magnitude of the correlation, for Pearson coefficients, the correlation between (length1, length2), (length2, length3) and (length1, length3) are among the top three strongest with the measures more than 0.99. For the rest of pairs, correlation measures are between 0.90 to 0.97. We find similar result in the Spearman coefficients. Again, the pairs (length1, length2), (length2, length3) have the top three strongest positive relationships, and their estimated magnitudes are all more than 0.99. The other pairs have correlation coefficients around 0.92-0.97, which is almost similar to Pearson result.