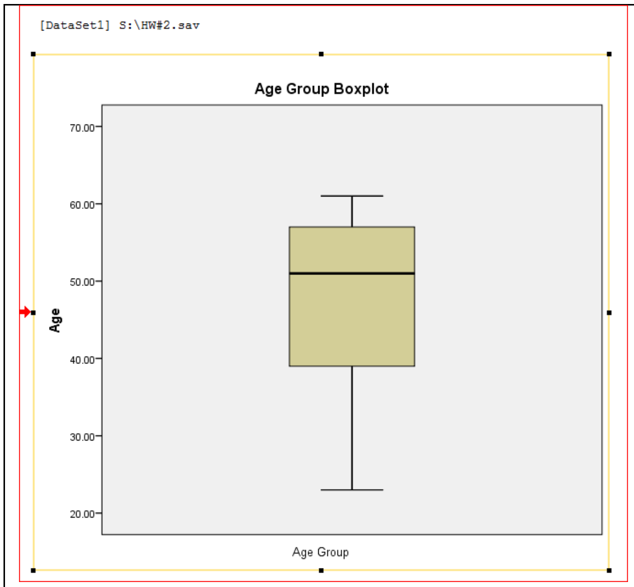
Problem 1:

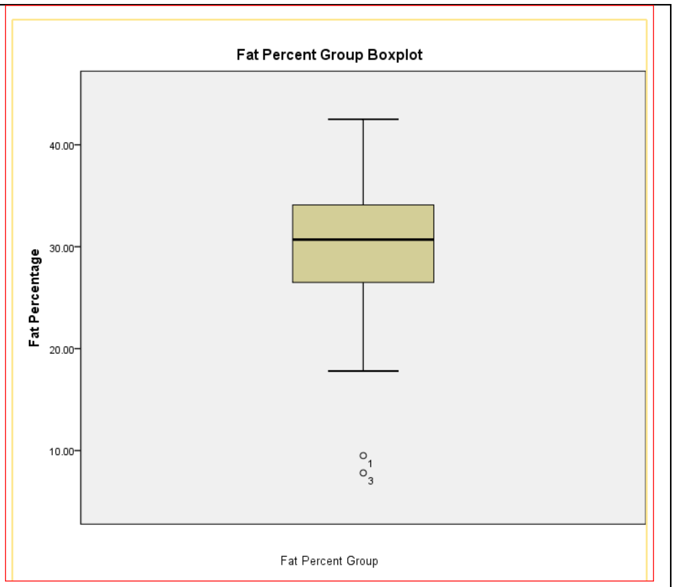
Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Use the data view to enter this into an SPSS table. Then do the following:
a. (3 points) Draw the box-plots for age and % fat. Explain what you can tell from this visualization of the distribution of the data.

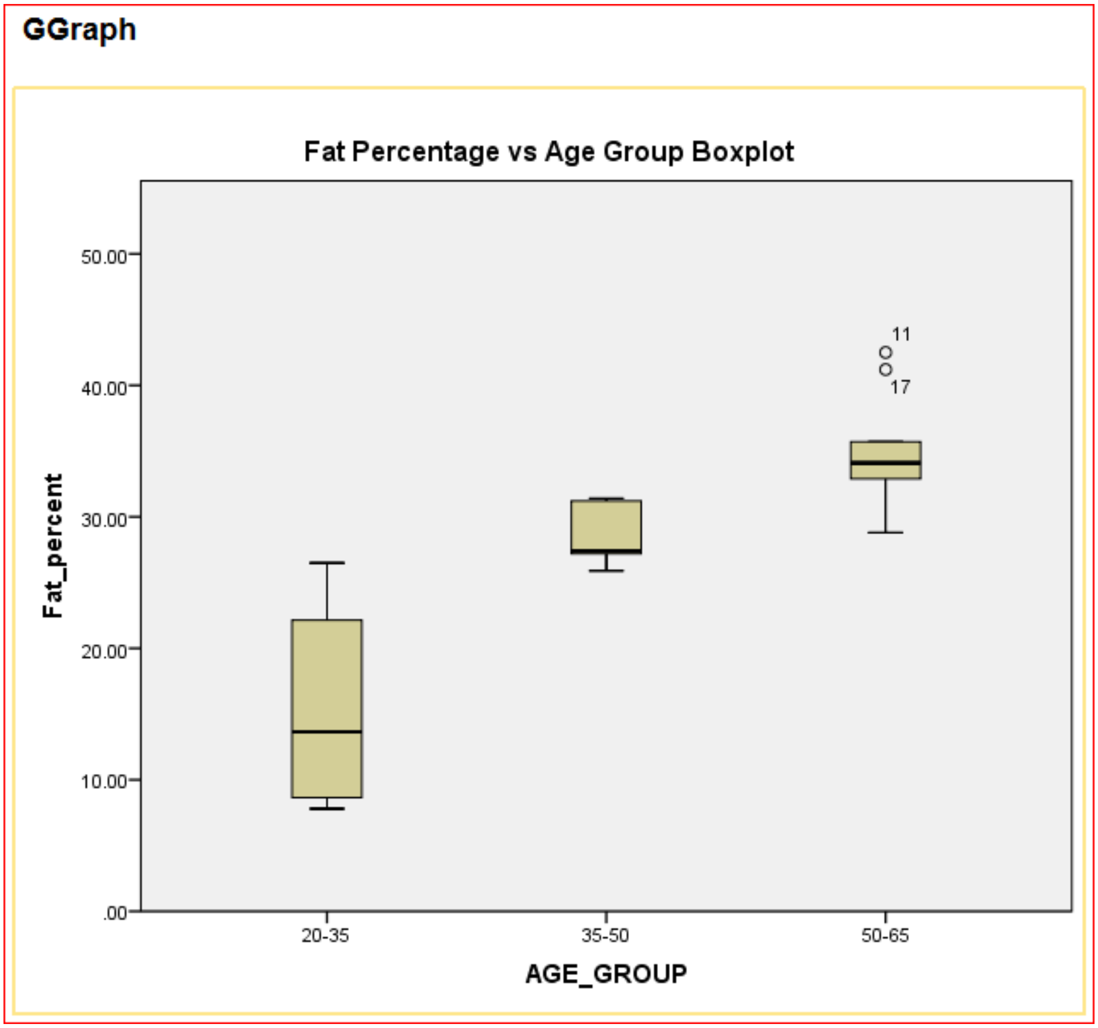
Ans:





From the above plots, we could observe that most of the data falls in the distribution in the Age group, there is No outliers within the group, the Statistics of this group is as the following: Median=51, Mean=46.4, S.D=13.2, Min=23, Max=61.

However, there are 2 outliners which appear at the lower part within the Fat% group are 9.5% and 7.8% of Fat. the Statistics of this group is as the following: Median=30.7, Mean=28.8, S.D=9.25, Min=7.8, Max=42.5.



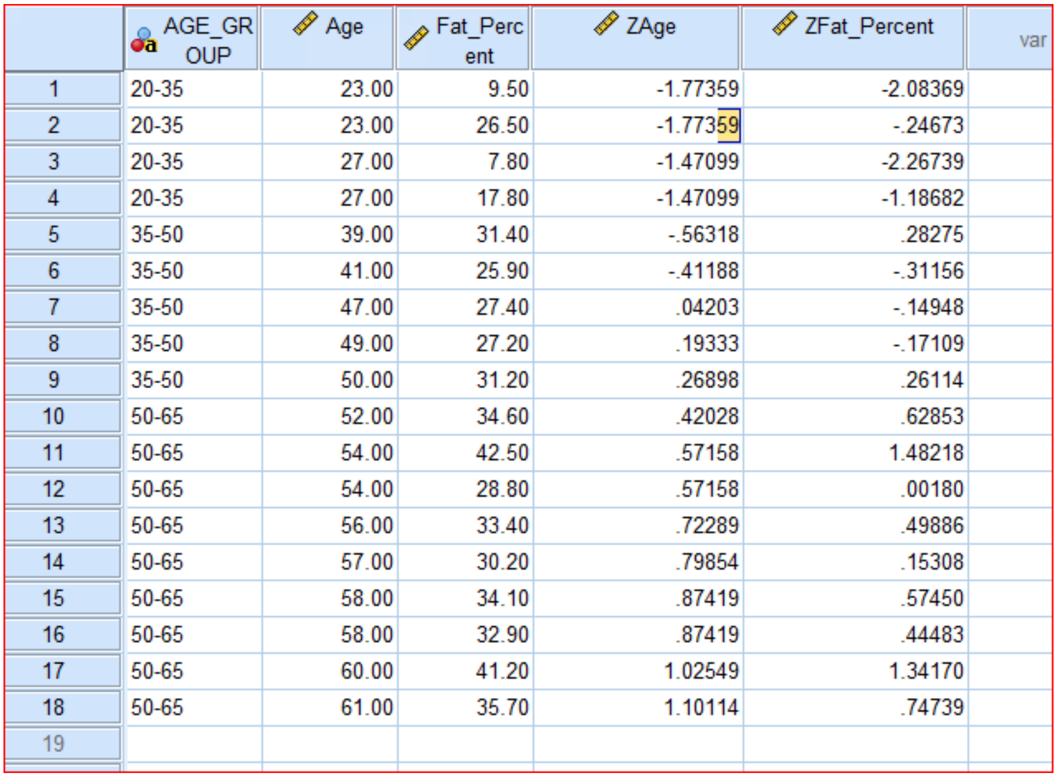
GGraph

Ans:

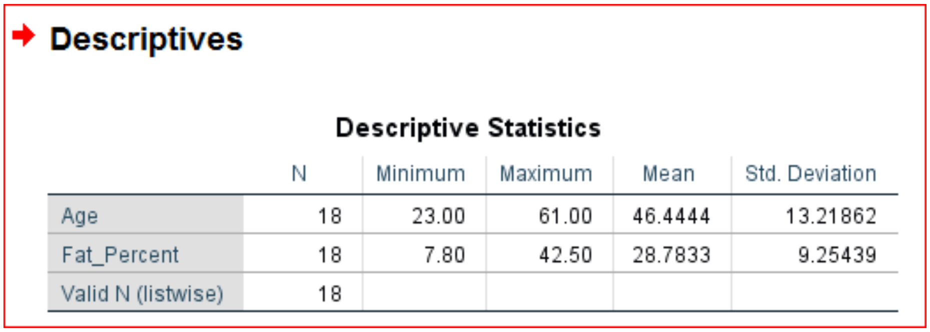
By looking at the above plot, there are 3 age group. The data of first 2 groups (Age: 20-35, 35-50) are very much falls in the range. On the other hand, there are 2 outliers which are 42.5% and 41.2% of fat. This result might be a closer view for Analyst to observe that there is some people having abnormal situation based on the report. Then, they might be able to do a further analysis on those people.

b. (3 points) Normalize the two attributes based on z-score normalization. Include an image showing the data table with this done.

Ans:



	AGE_GROUP	Age	Fat_Percent	ZAge	ZFat_Percent	var
1	20-35	23.00	9.50	-1.77359	-2.08369	
2	20-35	23.00	26.50	-1.77359	-.24673	
3	20-35	27.00	7.80	-1.47099	-2.26739	
4	20-35	27.00	17.80	-1.47099	-1.18682	
5	35-50	39.00	31.40	-.56318	.28275	
6	35-50	41.00	25.90	-.41188	-.31156	
7	35-50	47.00	27.40	.04203	-.14948	
8	35-50	49.00	27.20	.19333	-.17109	
9	35-50	50.00	31.20	.26898	.26114	
10	50-65	52.00	34.60	.42028	.62853	
11	50-65	54.00	42.50	.57158	1.48218	
12	50-65	54.00	28.80	.57158	.00180	
13	50-65	56.00	33.40	.72289	.49886	
14	50-65	57.00	30.20	.79854	.15308	
15	50-65	58.00	34.10	.87419	.57450	
16	50-65	58.00	32.90	.87419	.44483	
17	50-65	60.00	41.20	1.02549	1.34170	
18	50-65	61.00	35.70	1.10114	.74739	
19						



Descriptives

Descriptive Statistics

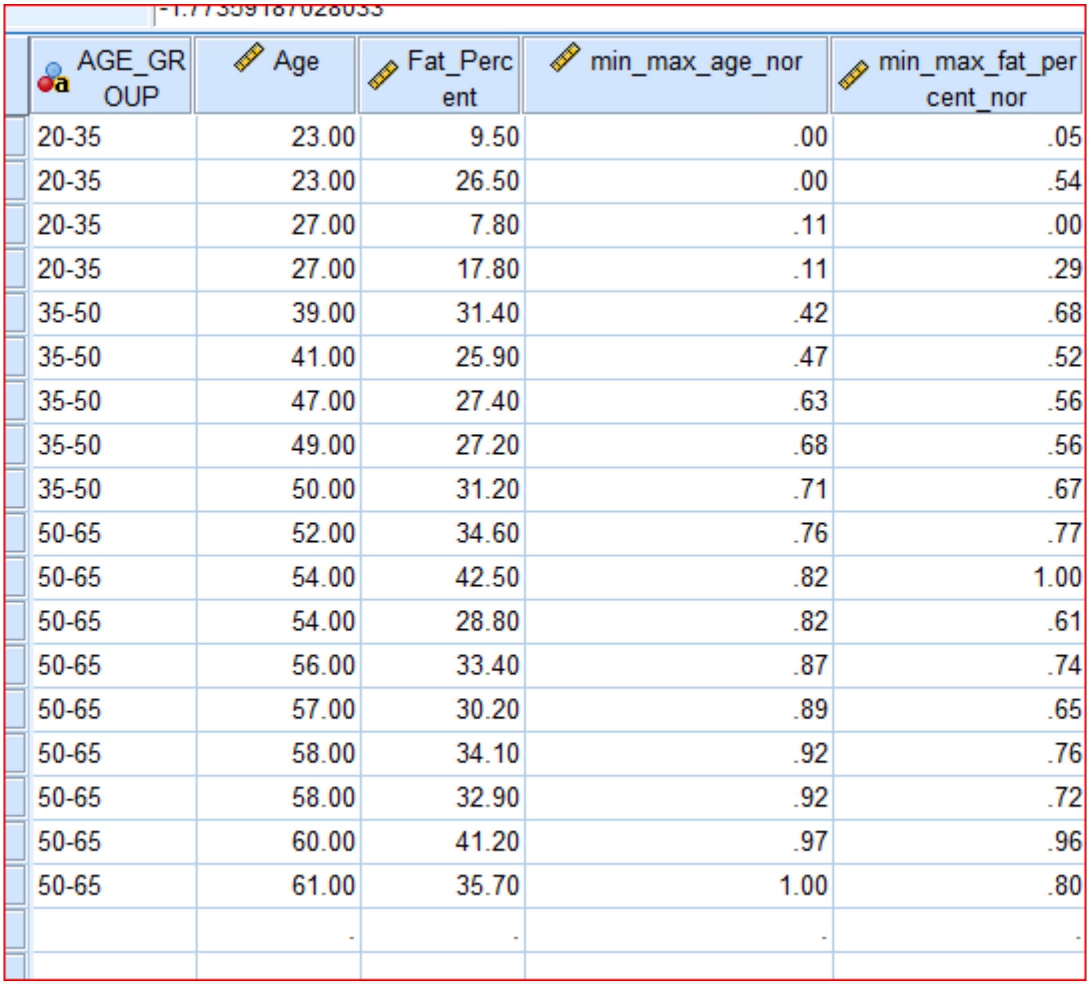
	N	Minimum	Maximum	Mean	Std. Deviation
Age	18	23.00	61.00	46.4444	13.21862
Fat_Percent	18	7.80	42.50	28.7833	9.25439
Valid N (listwise)	18				

The 4th and 5th columns show the z-score normalization of Age and Fat%. The Mean of z-score should be equals to ZERO, so that based on the answer in Part-a, the mean of Age group is 46.4, so that the row-7 z-score value is 0.042 which is the closest to ZERO because it is the value closest to Mean value. That happens the same on the Fat% variable, the Mean of Fat% is 28.8, the row-8 with Fat%27.2 which is the closest to the Mean, so that its z-score value is the closest to ZERO.

Also, we could observe that the Z-Score value is ranged according to the standard derivation, the value lesser than the mean (z-score =0), then their values will be showed Negative. On the contrary, Values above the Mean shows Positive values. If the z-score value is +1 which means the z-score value is 1 standard deviation above the mean; If the z-score value is -1 which means the z-score value is 1 standard deviation below the mean.

c. (3 points) Regardless of the original ranges of the variables, normalization techniques transform the data into new ranges that allow to compare and use variables on the same scales. What are the value ranges of the following normalization methods applied to this data? Explain your answer by explaining how the methods work on data in general.
i. Min-max normalization (use default target interval 0 to 1)

Ans:

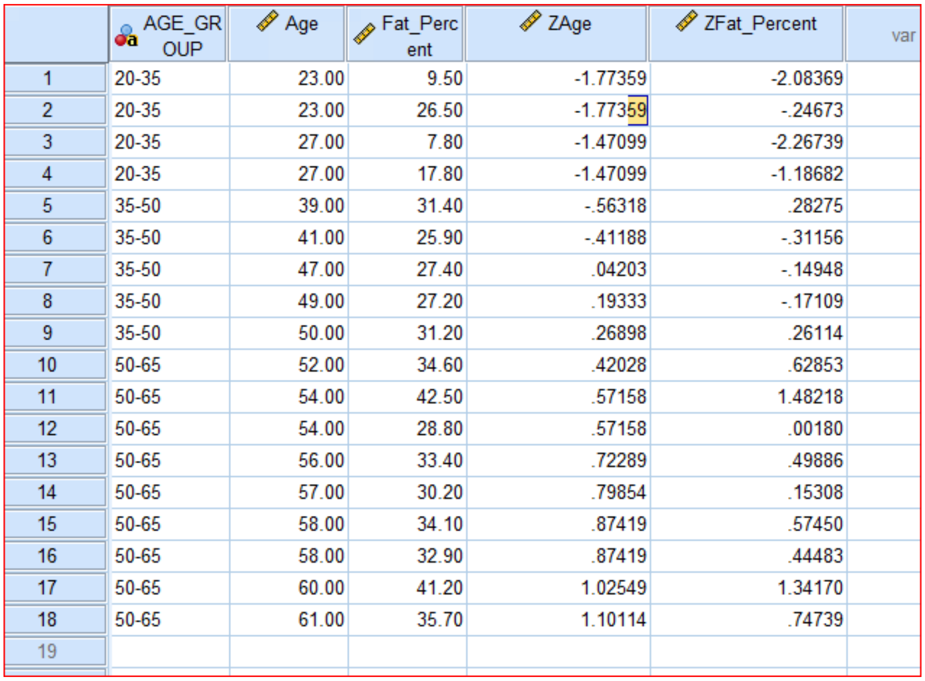


-1.77359187028033

AGE_GROUP	Age	Fat_Percent	min_max_age_nor	min_max_fat_percent_nor
20-35	23.00	9.50	.00	.05
20-35	23.00	26.50	.00	.54
20-35	27.00	7.80	.11	.00
20-35	27.00	17.80	.11	.29
35-50	39.00	31.40	.42	.68
35-50	41.00	25.90	.47	.52
35-50	47.00	27.40	.63	.56
35-50	49.00	27.20	.68	.56
35-50	50.00	31.20	.71	.67
50-65	52.00	34.60	.76	.77
50-65	54.00	42.50	.82	1.00
50-65	54.00	28.80	.82	.61
50-65	56.00	33.40	.87	.74
50-65	57.00	30.20	.89	.65
50-65	58.00	34.10	.92	.76
50-65	58.00	32.90	.92	.72
50-65	60.00	41.20	.97	.96
50-65	61.00	35.70	1.00	.80
	.	.	.	.

Based on the above, the values normalized ranged between 0 and 1 by using the following formula: $\frac{\text{value}-\min(\text{group})}{\min(\text{group})-\max(\text{group})}$ +(new_max - new_min) + new_min which new_max=1 and new_min=0. Basically, this method is used to consolidated the data into a smaller scale which might be easier for us to perform any analysis further. In addition, we should be aware that the entire set of data does NOT changed at all in terms of volume or scale. It just turned the dataset to a smaller scale.

ii. Z-score normalization



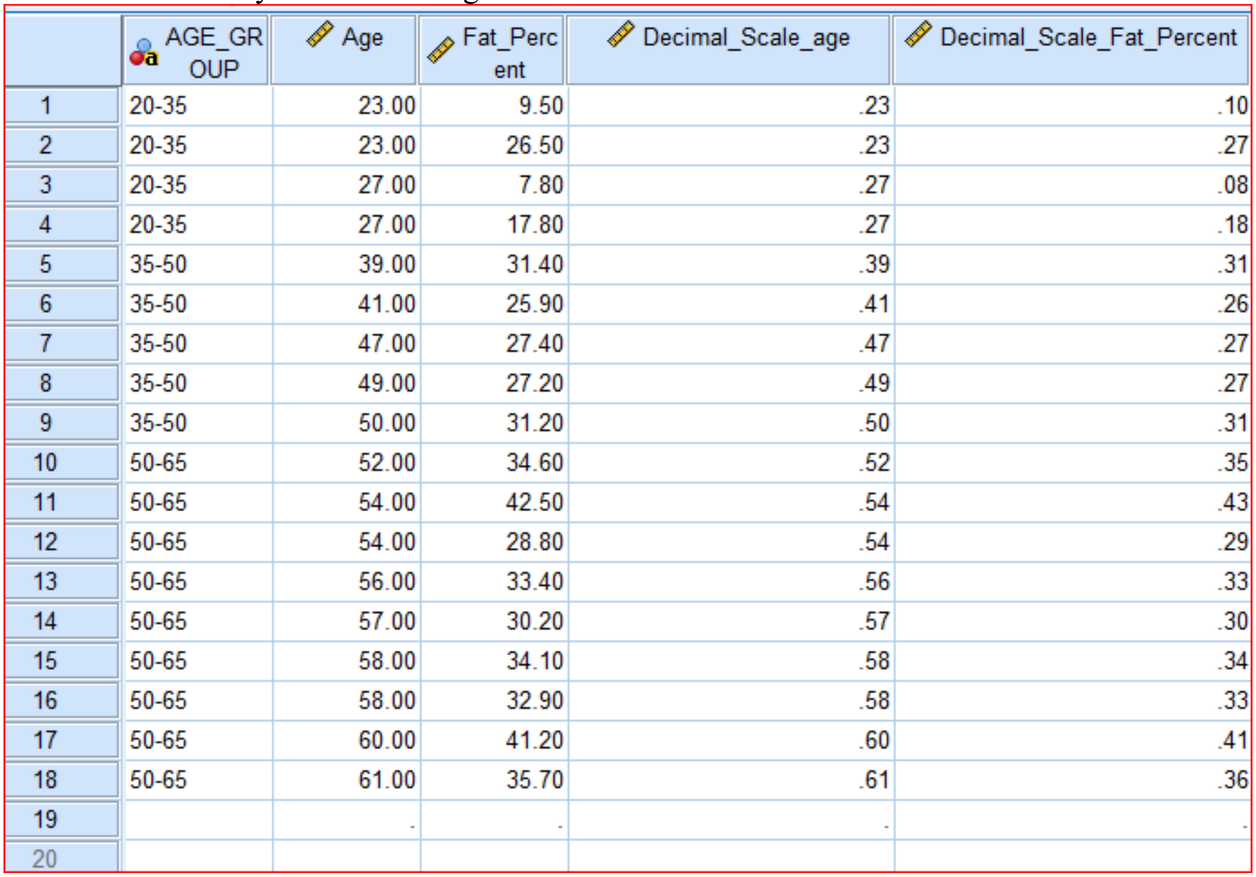
	AGE_GROUP	Age	Fat_Percent	ZAge	ZFat_Percent	var
1	20-35	23.00	9.50	-1.77359	-2.08369	
2	20-35	23.00	26.50	-1.77359	-.24673	
3	20-35	27.00	7.80	-1.47099	-2.26739	
4	20-35	27.00	17.80	-1.47099	-1.18682	
5	35-50	39.00	31.40	-.56318	.28275	
6	35-50	41.00	25.90	-.41188	-.31156	
7	35-50	47.00	27.40	.04203	-.14948	
8	35-50	49.00	27.20	.19333	-.17109	
9	35-50	50.00	31.20	.26898	.26114	
10	50-65	52.00	34.60	.42028	.62853	
11	50-65	54.00	42.50	.57158	1.48218	
12	50-65	54.00	28.80	.57158	.00180	
13	50-65	56.00	33.40	.72289	.49886	
14	50-65	57.00	30.20	.79854	.15308	
15	50-65	58.00	34.10	.87419	.57450	
16	50-65	58.00	32.90	.87419	.44483	
17	50-65	60.00	41.20	1.02549	1.34170	
18	50-65	61.00	35.70	1.10114	.74739	
19						

The Z-Score normalization actually does the similar job as the Min-max normalization method. But this method used another parameters to do the scaling. The formula we used to norimalize data is as the following:

$$\frac{Value - Mean(group)}{Standare\_Derivation(group)}$$

The Z-Score value is ranged according to the standard derivation, the value lesser than the mean (z-score =0), then their values will be showed Negative. On the contrary, Values above the Mean shows Positive values. In other words, If the z-score value is +1 which means the z-score value is 1 standard deviation above the mean; If the z-score value is -1 which means the z-score value is 1 standard deviation below the mean.

iii. Normalization by decimal scaling.



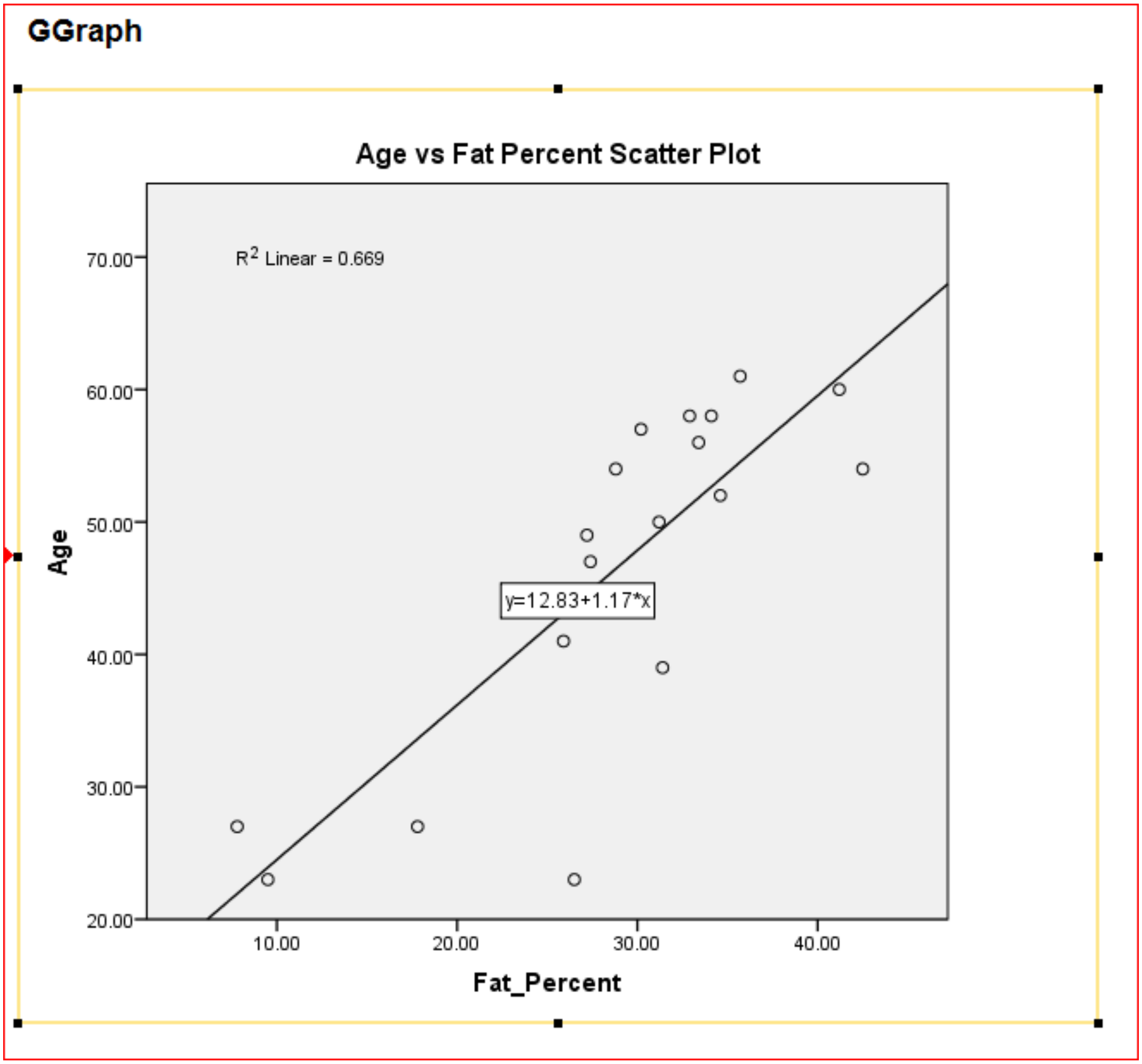
	AGE_GROUP	Age	Fat_Percent	Decimal_Scale_age	Decimal_Scale_Fat_Percent
1	20-35	23.00	9.50	.23	.10
2	20-35	23.00	26.50	.23	.27
3	20-35	27.00	7.80	.27	.08
4	20-35	27.00	17.80	.27	.18
5	35-50	39.00	31.40	.39	.31
6	35-50	41.00	25.90	.41	.26
7	35-50	47.00	27.40	.47	.27
8	35-50	49.00	27.20	.49	.27
9	35-50	50.00	31.20	.50	.31
10	50-65	52.00	34.60	.52	.35
11	50-65	54.00	42.50	.54	.43
12	50-65	54.00	28.80	.54	.29
13	50-65	56.00	33.40	.56	.33
14	50-65	57.00	30.20	.57	.30
15	50-65	58.00	34.10	.58	.34
16	50-65	58.00	32.90	.58	.33
17	50-65	60.00	41.20	.60	.41
18	50-65	61.00	35.70	.61	.36
19		.	.	.	.
20					

The Decimal Scaling method is also used to scaling down the original dataset without destroy the original pattern of the dataset. The formula we used for this method is as the following:

$$V' = \frac{Value}{10^n} \quad where\ n\ is\ the\ smallest\ integer\ such\ that\ Max(|V'| < 1)$$

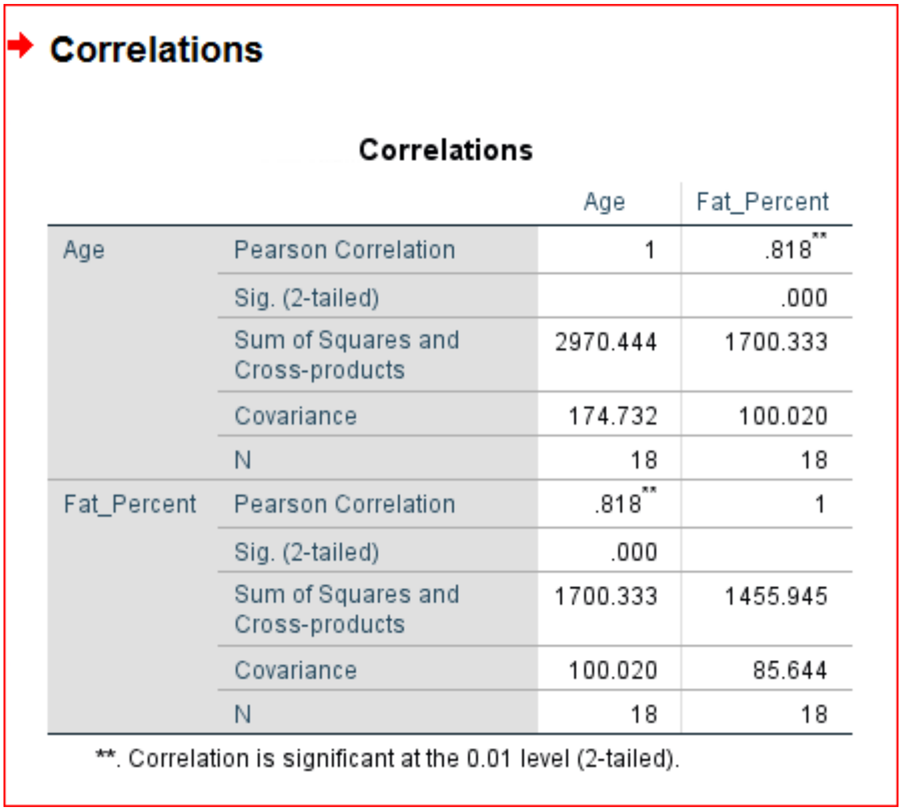
Decimal Scaling actually moves forth the decimal points of all the values. The number of decimal points that is required to move depends on the largest number within the group. For example, 61 is the largest number among the Age group, so the $V' = \frac{61}{10^2}\ where\ n = 2$ because 61 has 2 digits, and then the rest of group will be normalized the same way.

d. (3 points) Draw a scatter-plot based on the two variables and visually interpret the relationship between the two variables.
Ans:



Based on the above plot, we could observe that Age and Fat% are Positively correlated with $R^2 = 0.669$ although the R^2 value does not equal to 1. In other words, the value of Age increases, the value of Fat% increases as well. The plot also shows that there might be some outliers within the group, the young age 23 has 26.5% of fat, but that might not be true because we do not have a very good size of dataset for this analysis.

e. (3 points) Correlation is useful when integrating or cleaning data to see if two variables are so strongly correlated that they should be checked to see if they duplicate information. Get the full covariance and correlation matrix giving the relationships between all pairs of variables, even though there are only two. Are these two variables positively or negatively correlated?
Ans:



Correlations

Correlations

		Age	Fat_Percent
Age	Pearson Correlation	1	.818**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	2970.444	1700.333
	Covariance	174.732	100.020
	N	18	18
Fat_Percent	Pearson Correlation	.818**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	1700.333	1455.945
	Covariance	100.020	85.644
	N	18	18

**. Correlation is significant at the 0.01 level (2-tailed).

Based on the answer on part-D, the age and fat% are positively correlated with R $R^2 = 0.669$. In other words, the value of Age increases, the value of Fat% increases as well. With another way to observe this, we could check on the Covariance between Age and Fat%, the Cov value is 100.020 which is positive. So that the Cov value also supports our result (positively correlated).

Problem 2:
Suppose a group of 12 sales price records has been sorted as follows:
5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215
Partition them into bins by each of the following methods. Show which values are in which bins. Then smooth the data using the bins, and show the new set of smoothed values. Explain how each type of smoothing affect the data and the ways they are different.

a. (5 points) equal-depth partitioning with 3 values per bin

Number of value in each Bin= $\frac{12}{3} = 4$

Bin1: 5, 10, 11, 13,

Bin 2: 15, 35, 50, 55

Bing 3: 72, 92, 204, 215

After smoothing with bin mean:

Bin1: 9.75, 9.75, 9.75, 9.75

Bin 2: 38.75, 38.75, 38.75, 38.75

Bing 3: 145.75, 145.75, 145.75, 145.75

After smoothing with bin boundaries

Bin1: 5, 13, 13, 13,

Bin 2: 15, 55, 55, 55

Bing 3: 72, 72, 215, 215

Based on the above result, the smoothing are very similar result. The bin mean method generate a more FLAT value outcome because it could only show 3 values (9.75, 38.75,145.75). So that it might not be very representable. On the other hand, the bin boundaries method shows more values comparing the mean method. For these 2 methods on this Non-Uniform partition, the data might not affect a lot, especially on the bin mean method.

b. (5 points) equal-width partitioning with 3 bin

Ans:

Bin Width= $\frac{215-5}{3} = \frac{210}{3} = 70$

Bin1: [5-75] : 5, 10, 11, 13, 15, 35, 50, 55, 72

Bin 2: [75-145] : 92

Bing 3: [145-215] : 204, 215

After smoothing with bin mean:

Bin1: [5-75] : 30, 30, 30, 30, 30, 30, 30, 30, 30

Bin 2: [75-145] : 92

Bin 3: [145-215] : 311.5, 311.5

After smoothing with bin boundaries:

Bin1: [5-75] : 5, 5, 5, 5, 5, 5, 72, 72, 72

Bin 2: [75-145] : 92

Bin 3: [145-215] : 204, 215

According to the above result, we could observe that a very Right skewed distribution, especially after smoothing with bin Mean because most of the data fall in the Bin 1 category. So that using the bin mean method, we might defect the randomization of this particular data set because most data turns into value 30. The result might not be representable.

On the other hand, the replacement with bin boundaries method might generally generate a more reasonable result because it would diversify the values on each bin. In this case. The bin 1 area is still way larger than other bins.

By looking at the bin width and bin depth method, I would think that the bin boundaries method is more representable on this dataset because it would NOT differ much on the original dataset after smoothing, especially by using the bin boundaries smoothing method. Also, this non-uniform method could show more values than uniform method in this case.

Problem 3:

a. In real-world data, there are often rows that have missing values for some variables. Describe two methods for dealing with this problem.
Ans:
Method 1: We should ignore/ delete the tuple or record if there is missing values exists. In this case, we assume that there is NO Class labels on the table.

Method 2: We could generally fill in the missing value with the mean of attribute. Or we could fill out the missing value by using the mean value of that attribute belonging the that Class (if we have).

Method 3: We could also build a regression model by using known independent variable values to predict the dependent variable (i.e. missing value).

b. If we have class labels for our data, how can we use them to help get better estimates when filling in missing values?
Ans:
If we have class labels for our data, we could use the Mean value of the particular attribute belonging to that class to fill in the missing value.

c. Describe two issues that may come up during data integration.
Ans:
Issue 1:
Data value Conflict: While we integrate data into our schema. Data attributes values might have a variations on data format even they have the same meanings (e.g. kg , lb, time zone, language format etc).
Issue 2:
Redundancy data: When we do data integration from different tables, the integrated data table might then contain redundancy data which occupies countless space of database. This actually directly increase the business cost and process time. This happens due to data contains different format data (e.g. names , id and dateformat) because different tables are designed by different organization with different background and purpose.

Bonus Problem:
We discussed how a clustering of data can be used to smooth data, so let’s consider if it could be used for repairing missing data. We discussed how class labels can be used to improve the process of filling in missing values (and you wrote about it in 3b), and we discussed how a clustering result can be used similarly to class labels. Can we cluster data and use the clustering to fill in missing values? If so, how? If not, what problem would we encounter?

Ans:
For this question, I think it is very hard to say we should fill in the missing values with clustering analysis. It really depends on the real situation and data set format. Let’s discuss this matter in the following situation.

If the data comes in with a lot missing data (Let say Over 20% missing), I believe that we might probably need to check the consistence and accuracy of the data itself, and it probably not a good idea to replace missing values with clustering analysis. Because the missing value replacement would probably defect the similarity and randomization on the dataset itself and impact on the final model which might seriously affect the outcome of the users’ real life experience.

In some case, missing values comes in a very random (No particular pattern) from a dataset. Also, multiple predictor variables are available to predict the target value (i.e. response variable). In this kind of circumstance, we might probably could try to use the clustering analysis to replace the missing value by cluster represented values. With this process, we could build various model to compare the results to double check the missing values replacement would bring in positive or negative result to the response variable or class labels. However, we need to be positive about the data randomization before performing the replacement because sometimes missing values comes in to form other pattern (cluster), then we should pay more attention at this situation.