Statistical Analysis

Case Studies

1. Housing Price Analysis
2. Car Accidents in United States
3. Probability of being Myopic
4. Study Survey
5. Traffic Congestion
6. Real Estate Analysis
7. Process Improvements
8. Promotion in NYPD
9. Nutrition Education Program

EV

Housing Price Analysis

A sample of housing prices from a neighborhood in Ontario is listed:

1. Use Excel functions to give the answer of the values in blue area directly.

2. Organize the data into frequency distribution with seven class intervals appropriately (Outliers should not be included). Expand the table to have the following columns: relative frequency, cumulative frequency and mid point.

3. Draw the histogram and polygon to show the frequency distribution.

4. Draw the histogram and polygon to show cumulative frequency distribution.

5. Calculate the mean, median and sample standard deviation based on the frequency distribution developed in Point 2.
(Note: They are different from the values in blue area. Use the formula for grouped data and information in the frequency distribution table. Please use the formula for median and standard deviation below.)

List Price of houses

$ 204,900	$ 297,900	$ 635,000	$ 1,025,000
$ 210,000	$ 298,800	$ 635,000	$ 1,358,000
$ 214,900	$ 299,900	$ 640,000	$ 1,410,000
$ 219,877	$ 300,000	$ 649,900	$ 1,609,000
$ 228,500	$ 309,900	$ 702,500	$ 1,850,000
$ 229,900	$ 319,900	$ 725,800	$ 2,120,000
$ 229,900	$ 319,900	$ 742,000	$ 2,150,000
$ 234,900	$ 328,000	$ 756,000	$ 2,850,000
$ 235,800	$ 338,000	$ 757,000	
$ 239,900	$ 349,500	$ 778,200	
$ 244,900	$ 349,800	$ 792,100	
$ 249,000	$ 349,900	$ 807,000	
$ 249,900	$ 349,900	$ 821,000	
$ 249,900	$ 359,800	$ 825,000	
$ 250,000	$ 359,900	$ 827,000	
$ 253,000	$ 365,000	$ 832,100	
$ 254,800	$ 389,900	$ 835,000	
$ 259,900	$ 389,900	$ 836,000	
$ 264,900	$ 398,800	$ 839,000	
$ 265,000	$ 459,900	$ 839,500	
$ 269,000	$ 469,900	$ 841,000	
$ 270,000	$ 499,000	$ 850,000	
$ 274,000	$ 565,900	$ 855,000	
$ 274,900	$ 578,000	$ 858,000	
$ 275,900	$ 591,000	$ 859,000	
$ 278,877	$ 598,000	$ 886,000	
$ 279,888	$ 605,000	$ 890,000	
$ 279,900	$ 615,000	$ 899,100	
$ 279,900	$ 621,500	$ 921,000	
$ 279,900	$ 625,000	$ 925,000	
$ 289,900	$ 628,000	$ 985,000	
$ 292,500	$ 632,000	$ 992,300	
		$ 995,000	

Question Part I	
Minimum Value	2,04,900
Maximum Value	$2,850,000
Range	$2,645,100
First Quartile	$279,888
Second Quartile (Median)	$469,900
Third Quartile	$827,000
Inter Quartile Range	$547,112
30 Percentile	$293,580
Lower Limit of Inner Fence	-$540,780
Upper Limit of Inner Fence	$1,647,668
Lower Limit of Outer Fence	-$1,361,448
Upper Limit of Outer Fence	$2,468,336
Percentage of Prices less than $500,000	51%
Number of Mild Outliers	3
Number of Extreme Outliers	1
Sample Mean	$ 6,04,771
Sample Standard Deviation	447661.60

Frequency Distribution Table

Lower Limit	Upper Limit	Frequency(f)	Midpoint(x)	Cum Freq(cf)	Rel Freq(rf)	fx	f(x^2)	Class Interval
204900	405485	51	305192.5	51	0.5050	15564817.5	4750265564868.75	204900-405485
405486	606071	8	505778.5	59	0.0792	4046228.0	2046495128498.00	405486-606071
606072	806657	16	706364.5	75	0.1584	11301832.0	7983212909764.00	606072-806657
806658	1007243	22	906950.5	97	0.2178	19952911.0	18096302607905.50	806658-1007243
1007244	1207829	1	1107536.5	98	0.0099	1107536.5	1226637098832.25	1007244-1207829
1207830	1408415	1	1308122.5	99	0.0099	1308122.5	1711184475006.25	1207830-1408415
1408416	1609001	2	1508708.5	101	0.0198	3017417.0	4552402675944.50	1408416-1609001
		101				56298864.5	40366500460819.20	

Calculating the class width

(Max value-Min value)/No. of classes

$ 200,586

Round off to 2,00,585

GROUPED DATA VALUES

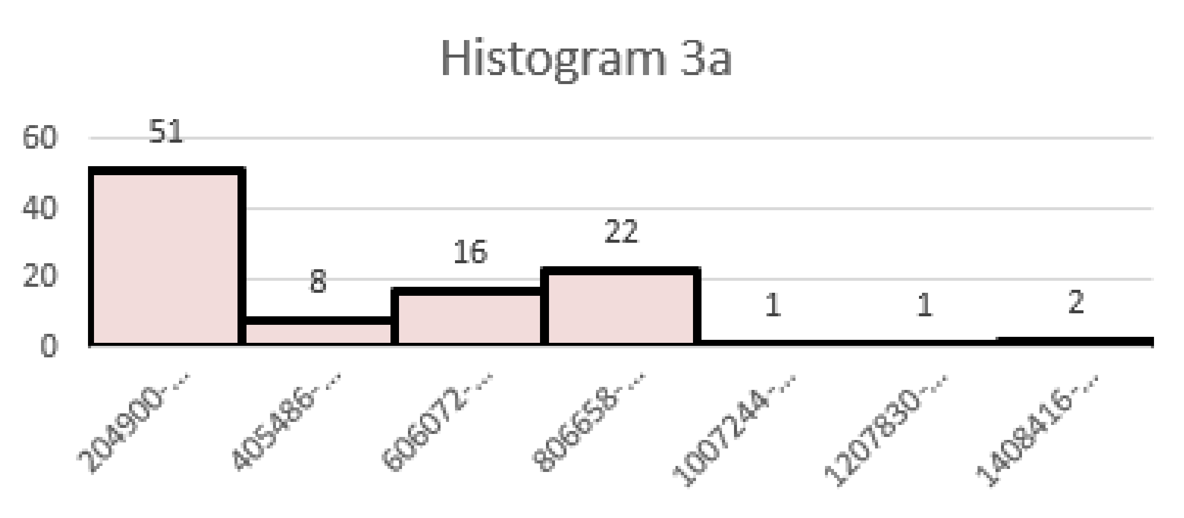
5a. Mean
557414.5

5b.Median
403519.47

5c. Standard Deviation
298257.27

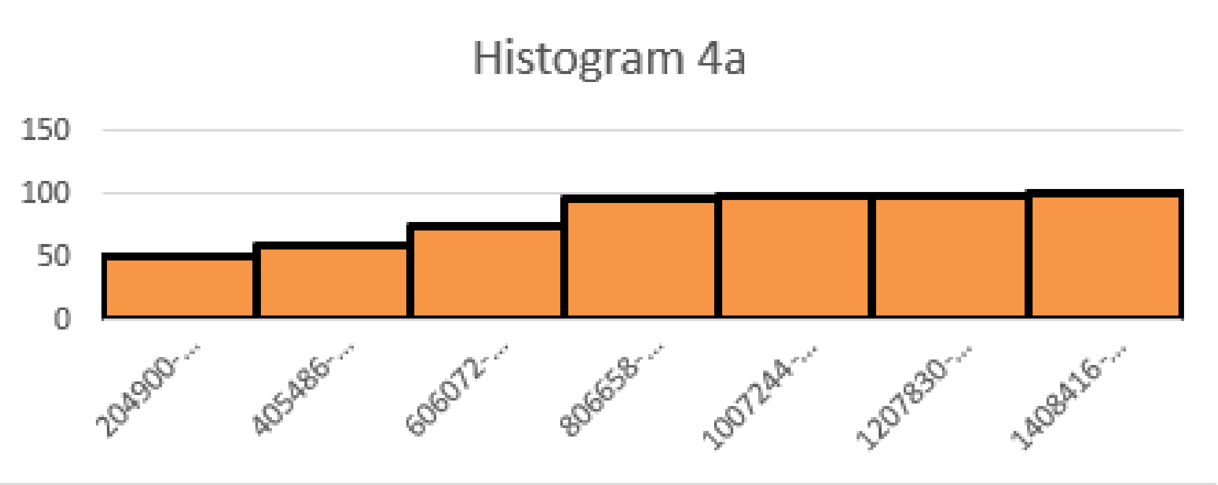
3a. Creating a Histogram

Class Interval	Frequency
204900-405485	51
405486-606071	8
606072-806657	16
806658-1007243	22
1007244-1207829	1
1207830-1408415	1
1408416-1609001	2



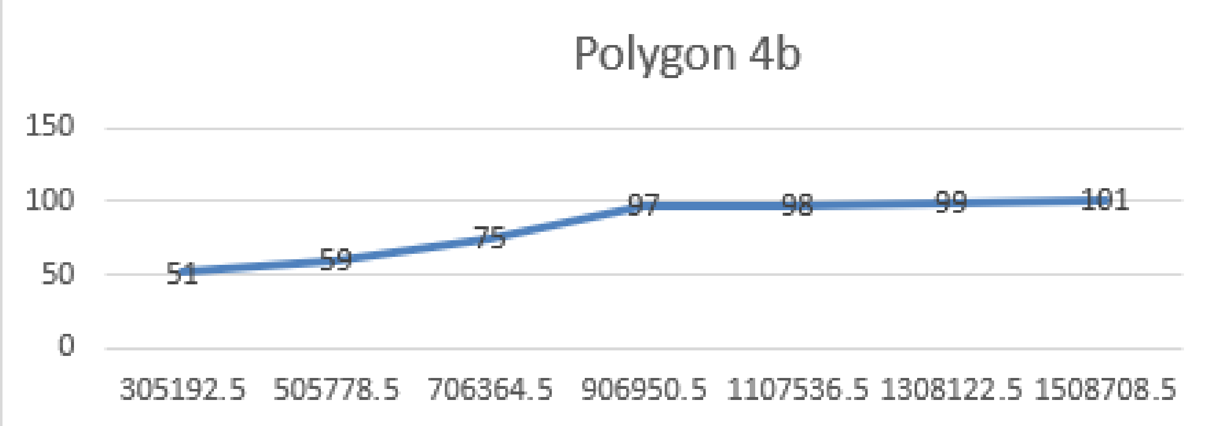
a. Creating a Histogram (Cumulative Frequency)

lass Interval	CF
04900-405485	51
05486-606071	59
06072-806657	75
06658-1007243	97
007244-1207829	98
207830-1408415	99
408416-1609001	101



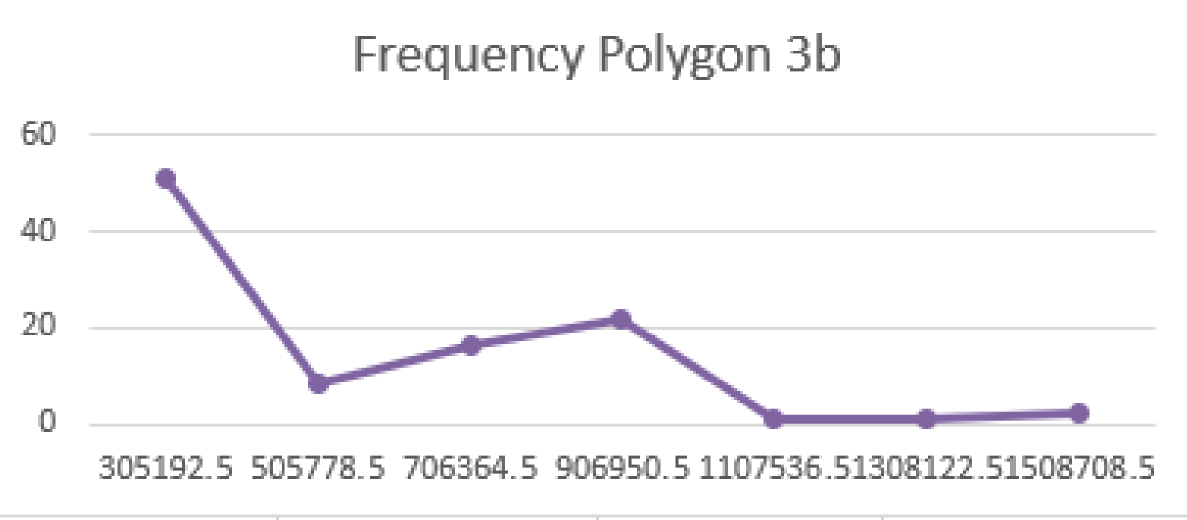
b. Creating a Polygon (Cumulative Frequency)

Midpoint	CF
305192.5	51
505778.5	59
706364.5	75
906950.5	97
1107536.5	98
1308122.5	99
1508708.5	101



3b.Creating a Polygon

Midpoint	Frequency
305192.5	51
505778.5	8
706364.5	16
906950.5	22
1107536.5	1
1308122.5	1
1508708.5	2



Conditional Probability- Car Accidents in USA

The U.S. National Highway Traffic Safety Administration gathers data concerning the causes of highway crashes where at least one fatality has occurred. The following probabilities were determined from the 1998 annual study (BAC is blood-alcohol content). *(Source: Statistical Abstract of the United States, 2000,* Table 1042.)

$P(BAC = 0 \mid \text{Crash with fatality}) = .616$
$P(BAC \text{ is between } .01 \text{ and } .09 \mid \text{Crash with fatality}) = .300$
$P(BAC \text{ is greater than } .09 \mid \text{Crash with fatality}) = .084$

Over a certain stretch of highway during a 1-year period, suppose the probability of being involved in a crash that results in at least one fatality is .01. It has been estimated that 12% of the drivers on this highway drive while their BAC is greater than .09. Determine the probability of a crash with at least one fatality if a driver drives while legally intoxicated (BAC greater than .09).

SOLUTION

Let Event A= BAC > 0.09 so
P(A)= 0.12

Let Event B = Crash with atleast one Fatality so
P(B)= 0.01

Given, **P(A/B)= 0.084** that means that Probabilty of BAC>0.09 given that crash with fatality happened.

As per the quetion, we need to find the probability of a crash with at least one fatality (B) given that BAC >0.09(A). So, find P(B/A)

P(B/A)= P(A&B)/P(A)....................................... (Equation 1)
We only know P(A)=0.12

But, P(A/B)= P(A&B)/P(B)
0.084= P(A&B)/0.01 so
P(A&B)=0.00084 (Putting this value in Equation 1)

P(B/A)= 0.00084/0.12
= 0.007

Thus, the probability of crash with atleast one fatality if the driver drives while legally intoxicated is 0.007

Probability of Being Myopic- Using Binomial Distribution

Researchers at the University of Pennsylvania School of Medicine theorized that children under 2 years old who sleep in rooms with the light on have a 40% probability of becoming myopic by age 16. Suppose that researchers found 25 children who slept with the light on before they were 2.

a. What is the probability that 10 of them will become myopic before age 16? - 0.16
b. What is the probability that fewer than 5 of them will become myopic before age 16? - 0.0095
c. What is the probability that more than 15 of them will become myopic before age 16? - 0.013
d. What is the probability that at least 3 of them will become myopic before age 16? - 0.99
e. What is the probability that at most 20 of them will become myopic before age 16? - 0.99
f. How many children will be expected to become myopic before age 16? - 10

PART A:	X=10, P= 0.4, N=25	Using Binomial Distribution formula	
	Probability that 10 of them will become myopic before age 16		**0.161157939**
PART B:	X= 1 to 4, p=0.4, n=25	Using Binomial Distribution Formula	
	P(X=1)		0.00005
	P(X=2)		0.00038
	P(X=3)		0.00194
	P(X=4)		0.00710
	P(X=1)+ P(X=2)+P(X=3)+P(X=4)		**0.00947**
PART C:	**0.013169073**		

PART D:	it means atmost 2	
Using BD	formula	**0.999571**
PART E:	METHOD 1	**0.999992**
	METHOD 2	**0.999992**
PART F:	Expected Value=Mean	
	Mean=	N*P
		10

Study Statistics- Using Normal Distribution

The amount of time devoted to studying statistics each week by students who achieve a grade of A in the course is a normally distributed random variable with a mean of 7.5 hours and a standard deviation of 2.1 hours.

a. What proportion of A students study for more than 10 hours per week? - *(3 out of 25)*
b. Find the probability that an A student spends between 7 and 9 hours studying. - 0.356
c. What proportion of A students spend fewer than 3 hours studying? - *(1 out of 50)*
d. What is the amount of time below which only 5% of all A students spend studying? - *4 hours*

a) z-score = (10 - 7.5) / 2.1 = 1.19
P(X < 10) = 0.8830 ~ 0.88
P (X > 10) = 1 - 0.88 = 0.12
Proportion of A students who study > 10 hours per week = 12/100 = 3/25

0.12 Using the formula

b) z-score = (9 - 7.5) / 2.1 = 0.7143
P (X < 9) = 0.7611
z-score = (7 - 7.5) / 2.1 = -0.2381
P (X < 7) = 0.4052
P (7 < X < 9) = 0.7611 - 0.4052 = 0.3559

0.36 Using the formula

c) z-score = (3 - 7.5) / 2.1 = −2.1429
P (X < 3) = 0.0162 ~ 0.02

Proportion of A students who study < 3 hours per week = 2/100 = 1/50

0.02 Using the formula

d) z-score associated with 5% is -1.65
-1.65 = (X - 7.5) / 2.1
X - 7.5 = -1.65 * 2.1 = −3.465
X = 7.5 -3.465 = 4.035 ~4
Only 5% of all A students spend studying less than 4

4.0 Using the formula

Traffic Congestion in USA - Using Hypothesis Testing

Traffic congestion seems to worsen each year. This raises the question, How much does roadway congestion cost the United States annually? The Federal Highway Administration's Highway Performance Monitoring System conducts an analysis to produce an estimate of the total cost. Drivers in the 73 most congested areas in the United States were sampled, and each driver's congestion cost in time and gasoline was recorded. The total number of drivers in these 73 areas was 128,000,000.

a. Estimate with 95% confidence the total cost of congestion in the 73 areas. (Adapted from the Statistical Abstract of the United States, 2006, Table 1082.)

b. If an organization claims that the total cost of congestion in the 73 areas is greater than $420, do you agree with it based on this sample result?

c. If an organization claims that the total cost of congestion in the 73 areas is less than $450, do you accept it based on this sample result?

Cost ($)	483	354	269	293
749	508	430	451	760
381	483	615	331	362
461	331	331	531	372
247	402	384	227	497
252	253	510	379	356
501	371	491	382	454
653	587	545	411	422
507	526	527	439	443
293	297	490	676	445
534	455	473	330	310
308	260	447	504	230
669	470	229	332	586
257	749	280	493	401
375	443	538	498	354
327	746	577	349	381
377	314	351	415	322
301	186	266	343	
604	418	511	489	
372	280	495	331	
237	648	532	459	
558	411	756	436	
242	509	394	482	
382	434	364	443	
392	326	332	462	
562	420	444	480	
557	439	302	485	
356	489	614	284	
250	465	354	380	
314	556	314	412	
575	224	327	355	
509	336	517	441	
456	364	252	473	
261	485	204	599	
455	409	556	374	
328	314	410	537	
	384	517	356	
	262	541		
	321	439		
	474	246		
	393	79		

PART-A	
Average(μ)	422.3636364
Standard Deviation(σ)	122.7739555
n	176
df	175
t-Stat	1.973612462
Confidence Interval	μ± t-stat*(σ/√n)
Calculating Lower Limit	404.0989679
Calculating Upper Limit Limit	440.6283048

PART-B	
We will have to formulate the hypothesis for this part: $H_o = 420$ $H_A > 420$	
We know that α = 0.05, so inorder to ascertain if we can accept our null hypothesis or not we should find the t-value and then the corresponding p-	
t-value	0.255405796
p-value	0.399354814
Since, p > α, so we fail to reject our null hypothesis. There is no sufficient evidence to support the claim that the total cost of congestion in the 73 areas is greater than $420	

PART-C	
We will have to formulate the hypothesis for this part: $H_o = 450$ $H_A < 450$	
We will find p-value and compare it with α to make the decision about accepting or rejecting the null hypothesis	
t-value	-2.986283153
p-value	0.001614454
Since, p < α, so we can reject the null hypothesis. There is sufficient evidence to claim that the total cost of congestion in the 73 areas is less than $450	

Real Estate Analysis - Using T:Test for two samples

Residents of neighbouring towns have an ongoing disagreement over who lays claim to the higher average price of of a single family home. Since you live in one of these towns, you decide to obtain a random sample of homes listed for sale with a majoy local realtor to investigate if there is actually any difference in the average home price.

a. Using the data provided, check the conditions (Independence, Randomization, Normal Condition and Variance Condition) for this test.

b. Write the null and alternative hypotheses for this test.

c. Test the hypotheses and find the p-value.

d. Make a conclusion about this test.

TOWN 1	
ID	Price
70615991	399900
70669695	429900
70650547	499000
70616722	669000
70667851	690000
70656875	699000
70610315	815900
70644981	929000
70626337	1365000
70658257	1395000
70576192	1650000
70642564	1695000
70547973	1750000
70654435	1995000
70642052	2100000
70624345	2750000
70657912	2950000

Mean: γ_1	1340100
SD1: σ_1	795584
n1:	17

TOWN 2	
ID	Price
70546158	489000
70660264	759900
70569799	799000
70649392	999000
70656361	1099000
70538352	1395000
70636923	1450000
70651181	1475000
70521906	1650000
70576614	1990000
70641078	1999800
70650711	2395000
70597605	2999000

Mean: γ_2	1499977
SD2 : σ_2	709909
n2:	13

Before performing a two-sample t-test, three conditions must be checked. First, the data in each group must be drawn independently and at random from its own homogenous population or generated by a randomized comparative experiment. The given data satisfy this assumption because it is given in the problem statement that the samples are random, and there is no reason to believe the data influence each other.

Next, the data in both groups must be approximately normal. The data appear normally distributed because the data are approximately symmetric instead of being skewed, and there are no outliers.

Finally, the two groups must be independent of each other. The given data satisfy this independent groups assumption because the two samples are not related. A home can only be in one town.

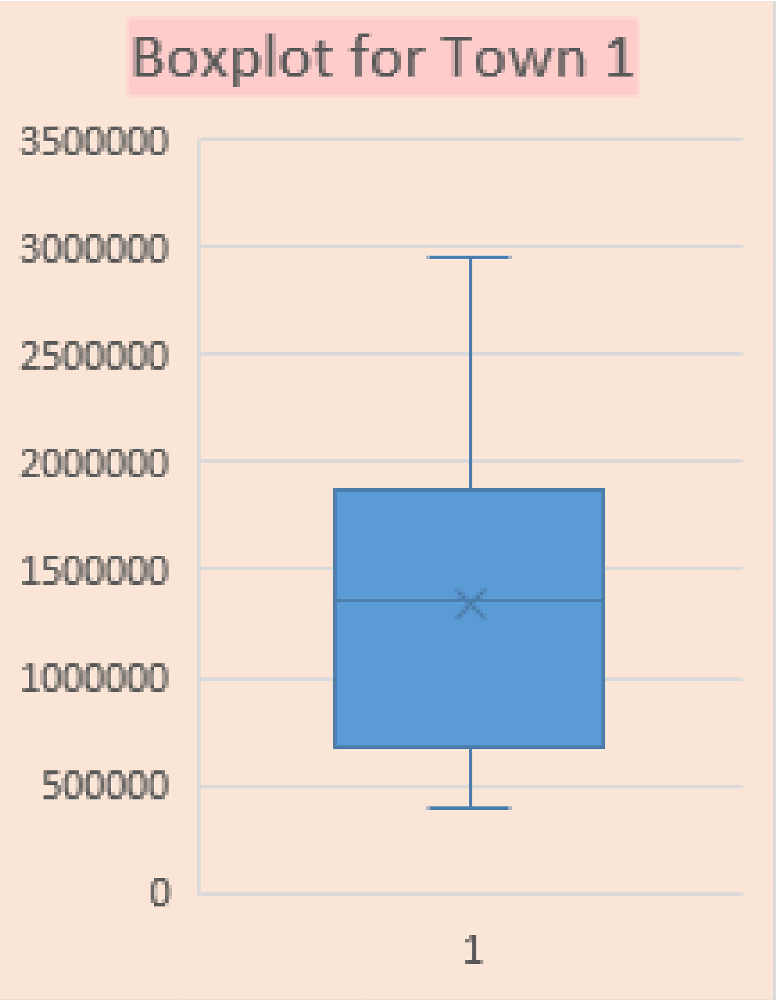
Part-B

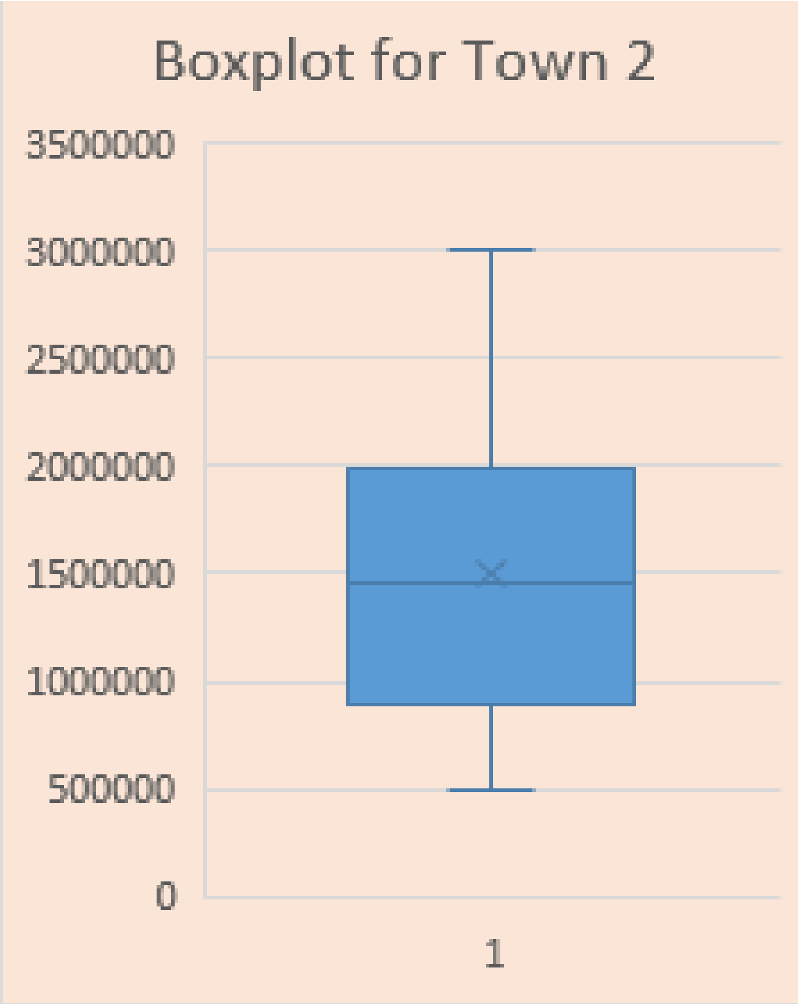
H_o:	$\mu_1 - \mu_2 = 0$
H_A:	$\mu_1 - \mu_2 \neq 0$

(PART-C) t-Test: Two-Sample Assuming Unequal Variances

(Used the Data Analysis in DATA)

	Town 1	*Town 2*
Mean	1340100	1499976.923
Variance	632954666250.00	503970670256.41
Observations	17	13
Mean Difference	0	
df	27	
t Stat	-0.579936071	
P(T<=t) one-tail	0.283383125	
t Critical one-tail	1.703288446	
P(T<=t) two-tail	**0.567**	
t Critical two-tail	2.051830516	





Part-D

Since p > 0.05, so w fail to reject the null hypothesis and conclude there is insuffcient evidence to prove that average price of houses is different for houses in both towns.

Process Improvement - Using Two Factor ANOVA Test with Replication

Design	System	Time
1	1	4.5
2	1	3.3
3	1	3.4
1	1	4
2	1	3
3	1	2.9
1	1	4.2
2	1	3
3	1	3.2
1	1	4.5
2	1	3.5
3	1	3.2
1	1	3.8
2	1	2.8
3	1	3
1	2	3
2	2	3.8
3	2	3.6
1	2	2.8
2	2	4
3	2	3.5
1	2	3
2	2	3.5
3	2	3.8
1	2	4
2	2	4.2
3	2	4.2
1	2	3
2	2	3.6
3	2	3.8

Sorting data by Workspace Design- 1, 2 and 3

Organized Data

Item#	Group	Storage 1	Storage 2
1	WD1	4.5	3
4		4	2.8
7		4.2	3
10		4.5	4
13		3.8	3
2	WD2	3.3	3.8
5		3	4
8		3	3.5
11		3.5	4.2
14		2.8	3.6
3	WD3	3.4	3.6
6		2.9	3.5
9		3.2	3.8
12		3.2	4.2
15		3	3.8

Results - ANOVA two factor test with replication

Anova: Two-Factor With Replication

SUMMARY	Storage 1	Storage 2	Total
WD1			
Count	5.00	5.00	10.00
Sum	21.00	15.80	36.80
Average	4.20	3.16	3.68
Variance	0.10	0.23	0.44
WD2			
Count	5.00	5.00	10.00
Sum	15.60	19.10	34.70
Average	3.12	3.82	3.47
Variance	0.08	0.08	0.21
WD3			
Count	5.00	5.00	10.00
Sum	15.70	18.90	34.60
Average	3.14	3.78	3.46
Variance	0.04	0.07	0.16
Total			
Count	15.00	15.00	
Sum	52.30	53.80	
Average	3.49	3.59	
Variance	0.33	0.21	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Workdesign	0.31	2.00	0.15	1.56	0.23	3.40
Storage	0.07	1.00	0.07	0.76	0.39	4.26
Interaction	4.88	2.00	2.44	24.72	0.00	3.40
Within	2.37	24.00	0.10			

One way to improve a process is to eliminate non-value-added activities (e.g. extra movements) and wasted effort (e.g. looking for materials). A consultant was hired to improve the efficiency in a large shop floor operation. She tested three different workspace designs and two different storage/retrieval systems. She measured process flow time for three randomly selected operations through each of of the combinations of workspace design and storage/retrieval systems.

a. Is this an experiment or observational study? Explain
It is an experiment because a consultant was hired to improve the efficiency in a large shop floor operation and she measures the flow time through a combinaton of work design and storage systems.

b. Use the data provided to run two-factor ANOVA.
The two factor ANOVA with replication was run after arranging the data as shown on the left. The results of ANOVA are on the next sheet.

c. What is the response variable?
Process Flow Time

d. How many treatments are involved?
There are 3 levels in the workspace design-1,2,3 and there are 2 levels in storage sytem-1,2. So total number of treatments will be: 3x2= 6. So, Treatments=6

e. Based on your ANOVA results, does workspace design impact process flow time?
Based on the P-value of workspace design which is 0.23 we can say that p>(alpha=0.05) so we fail to reject our null hypothesis and conclude that workspace design does not impact the the process flow time.

f. Based on your ANOVA results, does retrieval system impact process flow time?
Based on the P-value of storage systems which is 0.39 we can say that p>(alpha=0.05) so we fail to reject our null hypothesis and conclude that storage systems does not impact the the process flow time.

g. How does the interaction perform based on your ANOVA result?
The p-value for interaction is 0 which is less than alpha=0.05.This means that we can conclude there is significant interaction impact and hence, can agree with the fact that that workspace design & storage systems have an impact on the process time flow.

Promotion in NYPD - Using Chi Square Test

THE DATA COLLECTED SHOWS THE RANK ATTAINED BY MALE AND FEMALE OFFICERS IN THE NEW YORK CITY POLICE DEPARTMENT (NYPD). DO THESE DATA INDICATE THAT MEN AND WOMEN ARE EQUITABLY REPRESENTED AT ALL LEVELS OF THE DEPARTMENT?

A. WHAT'S THE PROBABILITY THAT A PERSON SELECTED AT RANDOM FROM NYPD IS A FEMALE?

B. WHAT'S THE PROBABILITY THAT A PERSON SELECTED AT RANDOM FROM NYPD IS A DETECTIVE?

C. ASSUMING NO BIAS IN PROMOTIONS, HOW MANY FEMALE DETECTIVES WOULD YOU EXPECT THE NYPD TO HAVE?

D. TO SEE IF THERE IS EVIDENCE OF THE DIFFERENCES IN RANKS ATTAINED BY MALES AND FEMALES, WOULD YOU TEST GOODNESS-OF-FIT OR HOMOGENEITY (INDEPENDENCE)?

E. STATE THE HYPOTHESES.

F. TEST THE CONDITIONS.

G. HOW MANY DEGREES OF FREEDOM ARE THERE?

H. FIND THE CHI-SQUARE VALUE AND THE ASSOCIATED P-VALUE.

I. STATE YOUR CONCLUSION.

J. IF YOU CONCLUDED THAT THE DISTRIBUTIONS ARE NOT THE SAME, ANALYZE THE DIFFERENCES USING THE STANDARDIZED RESIDUALS OF YOUR CONCLUSIONS.

Rank	Number of Females	Number of Males	*Total number for each rank*
Officer	4281	21900	***26181***
Detective	806	4058	***4864***
Sergeant	415	3898	***4313***
Lieutenant	89	1333	***1422***
Captain	12	359	***371***
Higher Ranks	10	218	***228***
	Sum of Females	*Sum of Males*	***Overall Total***
Total	***5613***	***31766***	***37379***

Part-A

Probability that a person selected at random from NYPD is a female is

P(Female)	*0.150164531*

Part -B

Probability that a person selected at random from NYPD is a detective is 0.13

P(Detective)	*0.130126542*

Part-C

Female detectives the NYPD is expected to have is 730

No. of Female Detectives	730.4002782

PART -D

Since we want to know if men and women are equitably represented at all levels of the department, therefore we should test for homogeneity

PART-E

Since we have 2 varibles gender and rank. Therefore, our null hypothesis and alternative hypothesis will be:

H_0: $(P\ Male\ Officer) = P(Female\ Officer),\ P(Male\ Detective) = P(Female\ Detective), P(Male\ Sergeant) = P(Female\ Sergeant)$, *P(Male Lieutenant) = P(Female Lieutenant), P(Male Captain)= P(Female Captain), P(Male Higher Ranks)=P(Female Higher Ranks)*

H_A: *There is different proportion for atleast one rank*

$H0$: Rank is independent of gender, HA: Rank is dependent on gender

PART-F

There are 3 conditions to use for Chi-Square test for Homogeneity. They are as follows:

1) The variables should be categorical. Here, rank and gender are both categorical.
2) We should assume that the sample is randomly selected
3) The Expected count for each cell should be minimum 5. We can calculate it:

Rank	Number of Females	Number of Males	*Total number for each rank*
Officer	4281	21900	***26181***
Detective	806	4058	***4864***
Sergeant	415	3898	***4313***
Lieutenant	89	1333	***1422***
Captain	12	359	***371***
Higher Ranks	10	218	***228***
	Sum of Females	*Sum of Males*	***Overall Total***
Total	***5613***	***31766***	***37379***

PART-F

Testing for expected count condition shows that all the values are greater than 5.

Rank	Number of Females	Number of Males
Officer	3931.457583	22249.54242
Detective	730.4002782	4133.599722
Sergeant	647.6596217	3665.340378
Lieutenant	213.5339629	1208.466037
Captain	55.71104096	315.288959
Higher Ranks	34.23751304	193.762487
	Sum of Females	*Sum of Males*
Total	***5613***	***31766***

PART-G

Degree of Freedom:(No. of rows-1)(No. of columns -1)

Degree of Freedom: **5**

PART-H			
Rank	**Female**	**Male**	**Total**
Officer	76.27%	68.94%	70.04%
Detective	14.36%	12.77%	13.01%
Sergeant	7.39%	12.27%	11.54%
Lieutenant	1.59%	4.20%	3.80%
Captain	0.21%	1.13%	0.99%
Higher Ranks	0.18%	0.69%	0.61%
TotAL	5613	31766	37379

PART-H(Expected Value calculation			
Rank	**Female**	**Male**	**Total**
Officer	3931.457583	22249.54242	26181
Detective	730.4002782	4133.599722	4864
Sergeant	647.6596217	3665.340378	4313
Lieutenant	213.5339629	1208.466037	1422
Captain	55.71104096	315.288959	371
Higher Ranks	34.23751304	193.762487	228

PART-H(Finding the residual)		
Rank	**Female**	**Male**
Officer	31.07750716	5.491344446
Detective	7.824912041	1.382649099
Sergeant	83.57862334	14.7682054
Lieutenant	72.62876457	12.83338335
Captain	34.29580688	6.060012718
Higher Ranks	17.15828593	3.031840929

ParT-H(Calculating Chi Square and P-Value)	
Chi Square χ^2	**290.1313359**
P-Value	**0.00**

Part-I

Since chi-square is very large and p-value is zero so we will reject null hypothesis. We can conclude that there is different proportion of male & female in atleast one of the ranks.

PART-J		
Rank	Female	Male
Officer	5.57	-2.34
Detective	2.80	***-1.18***
Sergeant	-9.14	3.84
Lieutenant	-8.52	3.58
Captain	-5.86	2.46
Higher Ranks	-4.14	***1.74***

In Part-J, If the standardized residual is beyond the range of ± 2, then that cell can be considered to be a major contributor or statistically significant.

So. all the cells except Male Detective and Male Higher Ranks are major contributors.

Nutrition Education Program- Using T -Test & Scatter Plots

Nutrition Education Programs

Nutrition education programs, which teach clients how to lose weight or reduce cholesterol levels through better eating patterns, have been growing in popularity. The nurse in charge of one such program at a local hospital wanted to know whether the programs actually work. A random sample was drawn of 33 clients who attended a nutrition education program for those with elevated cholesterol levels. The study recorded the weight, cholesterol levels, total dietary fat intake per average day, total dietary cholesterol intake per average day, and percent of daily calories from fat. These data were gathered both before and 3 months after the program.
The researchers also determined the clients' genders, ages, and heights. The data are stored in the following way:

Column 1: Gender (1 = female, 2 = male)
Column 2: Age
Column 3: Height (in meters)
Columns 4 and 5: Weight, before and after (in kilograms)
Columns 6 and 7: Cholesterol level, before and after
Columns 8 and 9: Total dietary fat intake per average day, before and after (in grams)
Columns 10 and 11: Dietary cholesterol intake per average day, before and after (in milligrams)
Columns 12 and 13: Percent daily calories from fat, before and after

The nurse would like the following information:

a. In terms of each of weight, cholesterol level, fat intake, cholesterol intake, and calories from fat, is the program a success?

b. Did the program affect the amount of reduction in each of weight, cholesterol level, fat intake, cholesterol intake, and calories from fat in females?

c. Does age affect the amount of reduction in weight, cholesterol level, fat intake, cholesterol intake, and calories from fat cholesterol?

Gender	Age	Height	Weight 1	Weight 2	Choles 1	Choles 2	TotFat 1	TotFat 2	DietC 1	DietC 2	PDCF 1	PDCF 2
1	22	1.60	74.20	71.70	6.82	7.50	19.3	21.2	88.1	154.9	25.2	23.6
1	30	1.62	99.60	96.60	5.73	5.30	39.3	27.7	239.0	149.5	45.7	30.5
1	34	1.73	71.80	71.30	6.26	6.64	71.6	43.3	168.6	156.4	31.2	24.9
1	40	1.50	56.00	53.80	6.82	7.68	38.1	29.8	102.3	75.7	47.3	37.5
2	40	1.75	86.30	87.40	7.22	6.67	94.1	70.2	368.9	256.4	39.9	29.8
2	40	1.75	97.00	96.00	5.42	4.86	88.5	64.8	233.3	190.3	33.1	32.3
1	41	1.65	73.90	71.10	6.78	5.57	17.0	22.1	39.6	59.7	27.3	22.6
2	43	1.78	104.80	99.00	9.02	7.61	114.8	33.6	532.1	178.4	37.7	27.5
2	43	1.73	96.30	96.60	7.52	6.68	117.3	72.0	939.8	261.0	38.4	34.2
2	45	1.75	91.50	86.30	7.03	5.25	94.5	55.7	299.8	282.6	45.8	29.8
1	46	1.63	48.80	47.90	5.50	4.56	110.7	61.4	368.6	110.9	41.7	31.2
2	46	1.74	68.80	75.90	6.67	5.76	114.4	93.4	381.9	298.6	47.8	39.5
2	46	1.74	100.00	100.70	6.45	7.02	34.1	14.1	90.0	26.2	23.1	20.0
1	49	1.59	59.00	58.20	6.49	5.47	45.9	30.1	82.7	75.0	36.0	28.8
1	50	1.68	85.20	83.10	7.80	7.12	49.2	37.8	192.6	230.5	35.8	28.4
2	51	1.75	79.80	78.40	7.20	6.18	63.4	68.6	246.4	588.4	40.6	34.9
2	51	1.70	80.50	73.50	6.41	5.42	56.2	20.7	144.7	71.7	34.5	24.3
1	52	1.61	95.20	95.20	6.86	6.25	47.6	16.6	131.2	78.3	34.7	20.9
1	53	1.63	65.20	63.00	7.20	7.39	48.2	36.5	95.2	101.4	32.6	32.5
2	53	1.77	97.80	94.00	6.00	6.32	38.5	32.0	126.6	111.0	39.2	34.9
2	53	1.73	82.60	81.80	7.04	7.53	131.5	31.1	489.6	102.1	50.4	40.4
1	54	1.58	65.00	63.70	5.83	5.71	63.7	49.5	252.2	193.2	38.6	33.8
1	55	1.74	69.90	66.80	6.59	5.97	48.4	31.8	226.6	160.3	33.7	32.1
1	56	1.65	68.40	68.90	7.93	6.49	57.7	36.5	70.8	87.3	32.3	25.2
1	56	1.53	95.80	95.30	6.96	6.32	80.1	50.4	263.8	152.9	35.6	31.2
2	57	1.74	98.90	101.40	6.29	6.07	108.3	90.0	329.0	363.6	37.8	41.8
1	58	1.57	52.30	52.80	7.49	6.55	55.0	62.4	275.5	215.6	49.3	38.2
2	58	1.70	103.50	103.50	6.70	6.05	52.3	72.2	172.3	151.5	35.1	34.4
2	63	1.79	83.20	81.40	7.73	5.96	92.1	101.1	386.9	373.2	38.8	45.0
2	63	1.78	80.30	76.80	7.01	6.54	44.9	36.8	125.0	81.5	22.5	19.8
1	64	1.65	60.70	60.50	7.17	6.00	21.5	24.2	124.2	104.1	19.7	15.3
2	65	1.69	79.00	77.90	6.70	6.16	62.4	41.5	192.8	149.1	38.0	40.8
2	69	1.66	65.00	64.00	8.07	6.21	75.9	62.5	219.8	253.8	36.3	31.0

PART - A

To compare the suceess of the education program, we will compare the results of the variables like weight, cholesterol level, fat intake, cholesterol intake, and calories from fat before & after the program. We will consider these as paired samples & perform 't-Test: Paired Two Sample for Means' Test. Then we will look at the P-value of these test reults. If it is less than the sigificance level which we assume to be 0.05, then we can say that the program is a success.

Comparison of weight before & after the program

	Weight 1	*Weight 2*
Mean	79.888	78.621
Variance	255.537	251.089
Observations	33.000	33.000
Pearson Correlation	0.988	
Hypothesized Mean Difference	0.000	
df	32.000	
t Stat	2.897	
P(T<=t) one-tail	0.003	
t Critical one-tail	1.694	
P(T<=t) two-tail	***0.007***	
t Critical two-tail	2.037	

Comparison of cholesterol level before & after the program

	Choles 1	*Choles 2*
Mean	6.870	6.267
Variance	0.583	0.618
Observations	33.000	33.000
Pearson Correlation	0.571	
Hypothesized Mean Difference	0.000	
df	32.000	
t Stat	4.825	
P(T<=t) one-tail	0.000	
t Critical one-tail	1.694	
P(T<=t) two-tail	***0.000***	
t Critical two-tail	2.037	

Comparison of total dietary fat intake per average day, before & after the program

	Total Fat 1	*Total Fat 2*
Mean	66.561	46.715
Variance	967.594	533.914
Observations	33.000	33.000
Pearson Correlation	0.635	
Hypothesized Mean Difference	0.000	
df	32.000	
t Stat	4.698	
P(T<=t) one-tail	0.000	
t Critical one-tail	1.694	
P(T<=t) two-tail	***0.000***	
t Critical two-tail	2.037	

Comparison of dietary cholesterol intake per average day, before & after the program

	DietC 1	*DietC 2*
Mean	242.421	177.124
Variance	30617.755	13032.424
Observations	33.000	33.000
Pearson Correlation	0.420	
Hypothesized Mean Difference	0.000	
df	32.000	
t Stat	2.289	
P(T<=t) one-tail	0.014	
t Critical one-tail	1.694	
P(T<=t) two-tail	***0.029***	
t Critical two-tail	2.037	

Hypothesis:

H0- The mean difference in the values for weight, cholesterol level, fat intake, cholesterol intake, and calories from fat before & after the program is zero

Ha-The mean difference in the values for weight, cholesterol level, fat intake, cholesterol intake, and calories from fat before & after the program is not zero

Conclusion: Since the p-values(P(T<=t) two-tail) is less than alpha(0.05) so we can reject the null hypothesis and conclude that values for variables are different.

The Nutrition Education Program is a success as it has changed the variables.

PART-B: NUTRITION

In this sheet, Gender 1= FEMALE

We will conduct a 't-Test: Two-Sample Assuming Unequal Variances' test and look at one-tailed test for reduction in the values of each variable with respect to gender. If the p-value for one-tailed test of less than 0.05 then we can conclude that gender impacts the reduction of certain variables

Gender	Age	Height	Weight 1	Weight 2	Choles 1	Choles 2	TotFat 1	TotFat 2	DietC 1	DietC 2	PDCF 1	PDCF 2
1	56	1.65	68.40	68.90	7.93	6.49	57.7	36.5	70.8	87.3	32.3	25.2
1	53	1.63	65.20	63.00	7.20	7.39	48.2	36.5	95.2	101.4	32.6	32.5
1	54	1.58	65.00	63.70	5.83	5.71	63.7	49.5	252.2	193.2	38.6	33.8
1	56	1.53	95.80	95.30	6.96	6.32	80.1	50.4	263.8	152.9	35.6	31.2
1	22	1.60	74.20	71.70	6.82	7.50	19.3	21.2	88.1	154.9	25.2	23.6
1	52	1.61	95.20	95.20	6.86	6.25	47.6	16.6	131.2	78.3	34.7	20.9
1	34	1.73	71.80	71.30	6.26	6.64	71.6	43.3	168.6	156.4	31.2	24.9
1	55	1.74	69.90	66.80	6.59	5.97	48.4	31.8	226.6	160.3	33.7	32.1
1	58	1.57	52.30	52.80	7.49	6.55	55.0	62.4	275.5	215.6	49.3	38.2
1	41	1.65	73.90	71.10	6.78	5.57	17.0	22.1	39.6	59.7	27.3	22.6
1	30	1.62	99.60	96.60	5.73	5.30	39.3	27.7	239.0	149.5	45.7	30.5
1	46	1.63	48.80	47.90	5.50	4.56	110.7	61.4	368.6	110.9	41.7	31.2
1	64	1.65	60.70	60.50	7.17	6.00	21.5	24.2	124.2	104.1	19.7	15.3
1	50	1.68	85.20	83.10	7.80	7.12	49.2	37.8	192.6	230.5	35.8	28.4
1	49	1.59	59.00	58.20	6.49	5.47	45.9	30.1	82.7	75.0	36.0	28.8

CONCLUSION:

We can see that the p-value for one tailed test for all the variables except weight & dietary cholesterol before and after the program is less than the significance level (0.05)

We can conclude that there is no reduction in weight & cholesterol intake from the program but cholesterol level, fat intake, and calories from fat for females from the adoption of nutrition education program has reduced.

So, the nutrition education program has proved beneficial for women in aspects where it caused reduction.

Comparison of dietary cholesterol intake per average day, before & after the program

	DietC 1	*DietC 2*
Mean	170.063	131.606
Variance	8637.321	2785.709
Observations	16.000	16.000
Hypothesized Mean Difference	0.000	
df	24.000	
t Stat	1.439	
P(T<=t) one-tail	*0.0814*	
t Critical one-tail	1.710	
P(T<=t) two-tail	0.163	
t Critical two-tail	2.063	

Comparison of Percent daily calories from fat, before & after the program

	PDCF 1	*PDCF 2*
Mean	35.419	28.544
Variance	63.067	37.697
Observations	16.000	16.000
Hypothesized Mean Difference	0.000	
df	28.000	
t Stat	2.739	
P(T<=t) one-tail	*0.005*	
t Critical one-tail	1.701	
P(T<=t) two-tail	0.011	
t Critical two-tail	2.048	

Comparison of weight before & after the program

	Weight1	*Weight 2*
Mean	71.313	69.994
Variance	242.216	235.761
Observations	16.000	16.000
Hypothesized Mean Difference	0.000	
df	30.000	
t Stat	0.241	
P(T<=t) one-tail	*0.4055*	
t Critical one-tail	1.697	
P(T<=t) two-tail	0.811	
t Critical two-tail	2.042	

Comparison of cholesterol level before & after the program

	Choles 1	*Choles 2*
Mean	6.764	6.283
Variance	0.485	0.748
Observations	16.000	16.000
Hypothesized Mean Difference	0.000	
df	29.000	
t Stat	1.737	
P(T<=t) one-tail	*0.0465*	
t Critical one-tail	1.699	
P(T<=t) two-tail	0.093	
t Critical two-tail	2.045	

Comparison of total dietary fat intake per average day, before & after the program

	Tot Fat 1	*Tot Fat 2*
Mean	50.831	36.331
Variance	563.122	191.642
Observations	16.000	16.000
Hypothesized Mean Difference	0.000	
df	24.000	
t Stat	2.111	
P(T<=t) one-tail	*0.0226*	
t Critical one-tail	1.710	
P(T<=t) two-tail	0.045	
t Critical two-tail	2.064	

PART-C(Scatter plots)

Age	Height	Weight 1	Weight 2	Choles 1	Choles 2	TotFat 1	TotFat 2	DietC 1	DietC 2	PDCF 1	PDCF 2
22	1.60	74.20	71.70	6.82	7.50	19.3	21.2	88.1	154.9	25.2	23.6
30	1.62	99.60	96.60	5.73	5.30	39.3	27.7	239.0	149.5	45.7	30.5
34	1.73	71.80	71.30	6.26	6.64	71.6	43.3	168.6	156.4	31.2	24.9
40	1.50	56.00	53.80	6.82	7.68	38.1	29.8	102.3	75.7	47.3	37.5
40	1.75	86.30	87.40	7.22	6.67	94.1	70.2	368.9	256.4	39.9	29.8
40	1.75	97.00	96.00	5.42	4.86	88.5	64.8	233.3	190.3	33.1	32.3
41	1.65	73.90	71.10	6.78	5.57	17.0	22.1	39.6	59.7	27.3	22.6
43	1.78	104.80	99.00	9.02	7.61	114.8	33.6	532.1	178.4	37.7	27.5
43	1.73	96.30	96.60	7.52	6.68	117.3	72.0	939.8	261.0	38.4	34.2
45	1.75	91.50	86.30	7.03	5.25	94.5	55.7	299.8	282.6	45.8	29.8
46	1.63	48.80	47.90	5.50	4.56	110.7	61.4	368.6	110.9	41.7	31.2
46	1.74	68.80	75.90	6.67	5.76	114.4	93.4	381.9	298.6	47.8	39.5
46	1.74	100.00	100.70	6.45	7.02	34.1	14.1	90.0	26.2	23.1	20.0
49	1.59	59.00	58.20	6.49	5.47	45.9	30.1	82.7	75.0	36.0	28.8
50	1.68	85.20	83.10	7.80	7.12	49.2	37.8	192.6	230.5	35.8	28.4
51	1.75	79.80	78.40	7.20	6.18	63.4	68.6	246.4	588.4	40.6	34.9
51	1.70	80.50	73.50	6.41	5.42	56.2	20.7	144.7	71.7	34.5	24.3
52	1.61	95.20	95.20	6.86	6.25	47.6	16.6	131.2	78.3	34.7	20.9
53	1.63	65.20	63.00	7.20	7.39	48.2	36.5	95.2	101.4	32.6	32.5
53	1.77	97.80	94.00	6.00	6.32	38.5	32.0	126.6	111.0	39.2	34.9
53	1.73	82.60	81.80	7.04	7.53	131.5	31.1	489.6	102.1	50.4	40.4
54	1.58	65.00	63.70	5.83	5.71	63.7	49.5	252.2	193.2	38.6	33.8
55	1.74	69.90	66.80	6.59	5.97	48.4	31.8	226.6	160.3	33.7	32.1
56	1.65	68.40	68.90	7.93	6.49	57.7	36.5	70.8	87.3	32.3	25.2
56	1.53	95.80	95.30	6.96	6.32	80.1	50.4	263.8	152.9	35.6	31.2
57	1.74	98.90	101.40	6.29	6.07	108.3	90.0	329.0	363.6	37.8	41.8
58	1.57	52.30	52.80	7.49	6.55	55.0	62.4	275.5	215.6	49.3	38.2
58	1.70	103.50	103.50	6.70	6.05	52.3	72.2	172.3	151.5	35.1	34.4
63	1.79	83.20	81.40	7.73	5.96	92.1	101.1	386.9	373.2	38.8	45.0
63	1.78	80.30	76.80	7.01	6.54	44.9	36.8	125.0	81.5	22.5	19.8
64	1.65	60.70	60.50	7.17	6.00	21.5	24.2	124.2	104.1	19.7	15.3
65	1.69	79.00	77.90	6.70	6.16	62.4	41.5	192.8	149.1	38.0	40.8
69	1.66	65.00	64.00	8.07	6.21	75.9	62.5	219.8	253.8	36.3	31.0

Calculating correlation between age and each variable

	Weight 1	Weight 2
Age	-0.139	-0.124

	TotFat 1	TotFat 2
Age	0.051	0.225

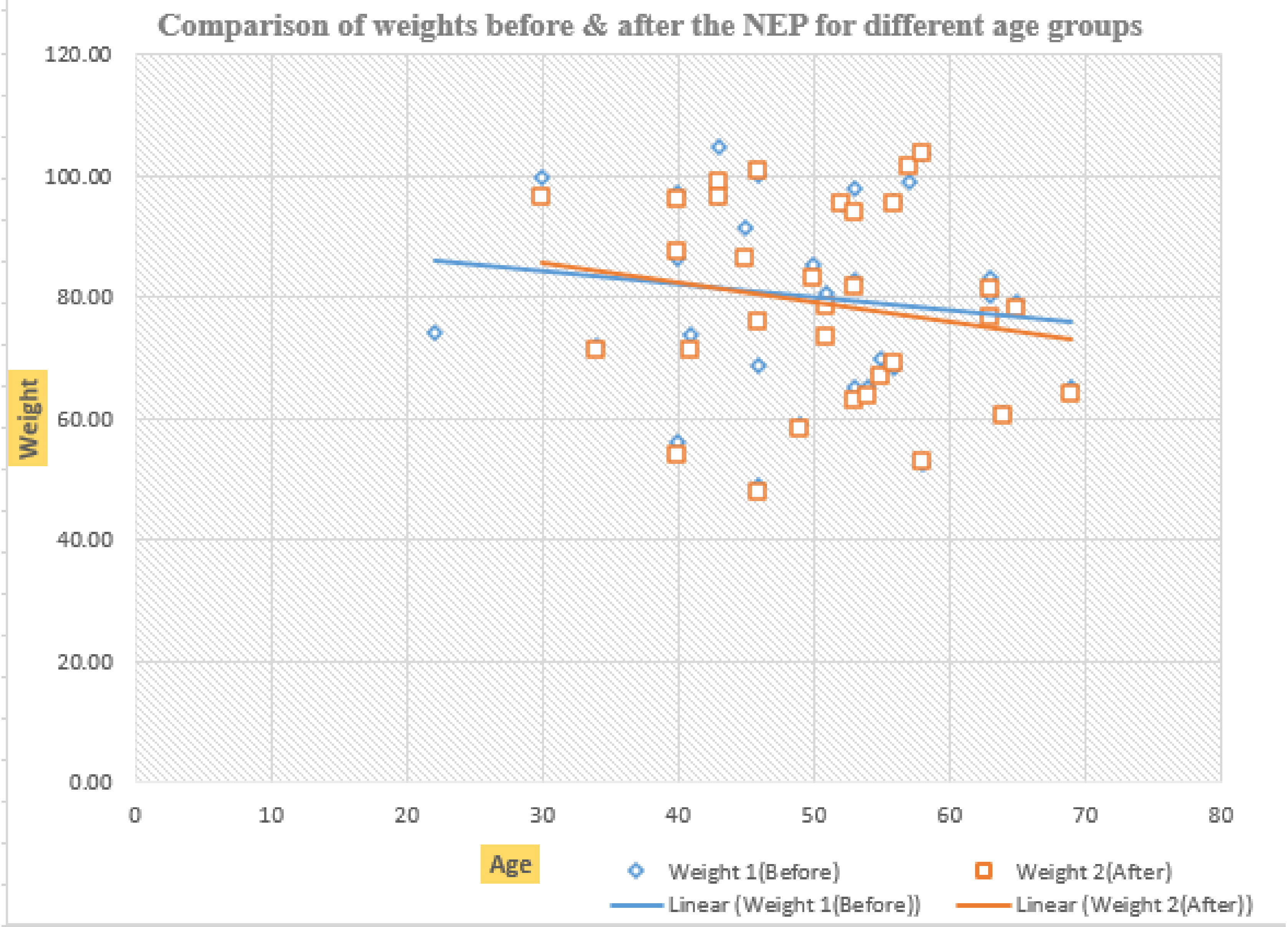
	Choles 1	Choles 2
Age	0.282	-0.088

	DietC 1	DietC 2
Age	-0.058	0.092

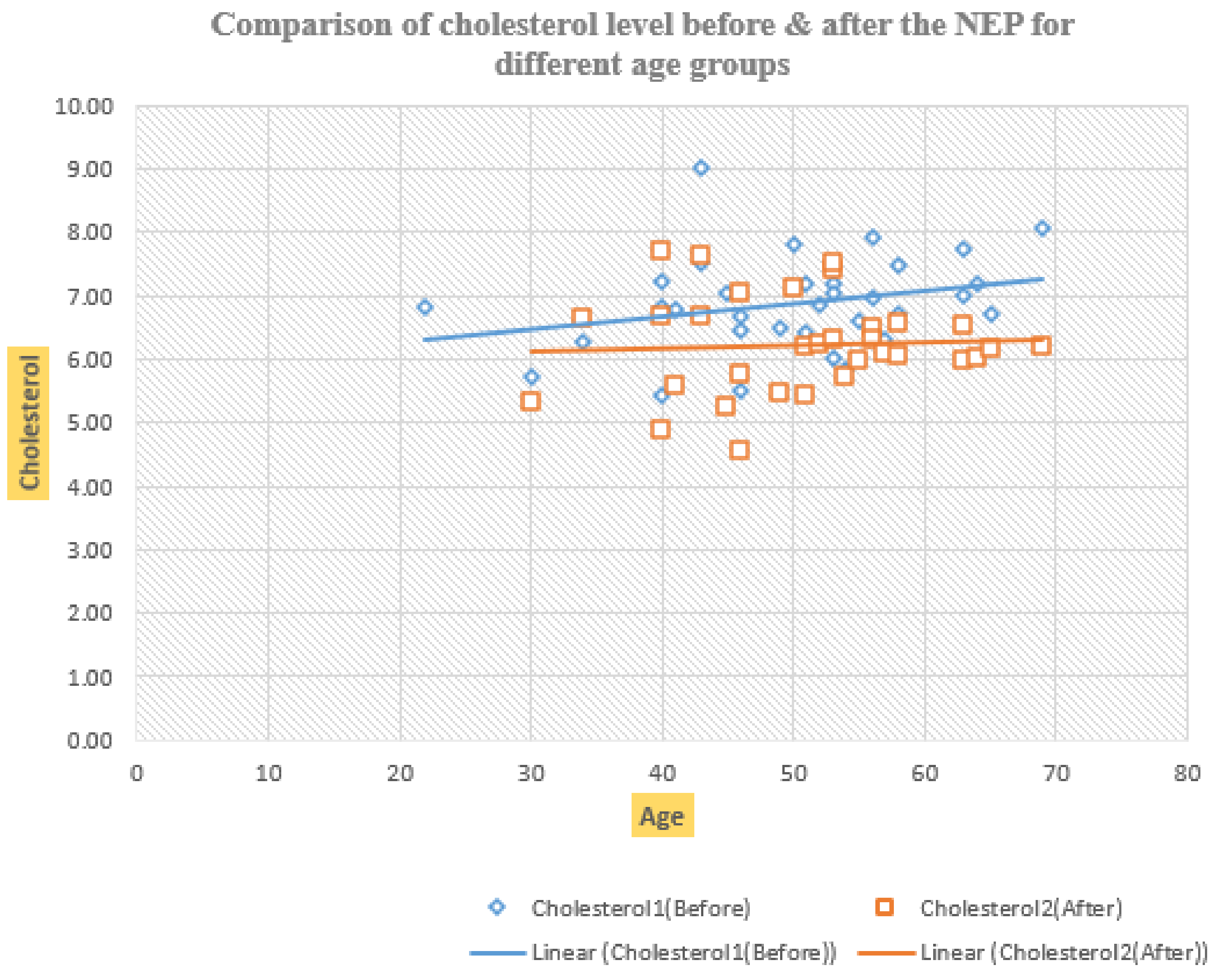
	PDCF 1	PDCF 2
Age	-0.078	0.150

None of the variables are highly corelated to age as they are way less than 0.7

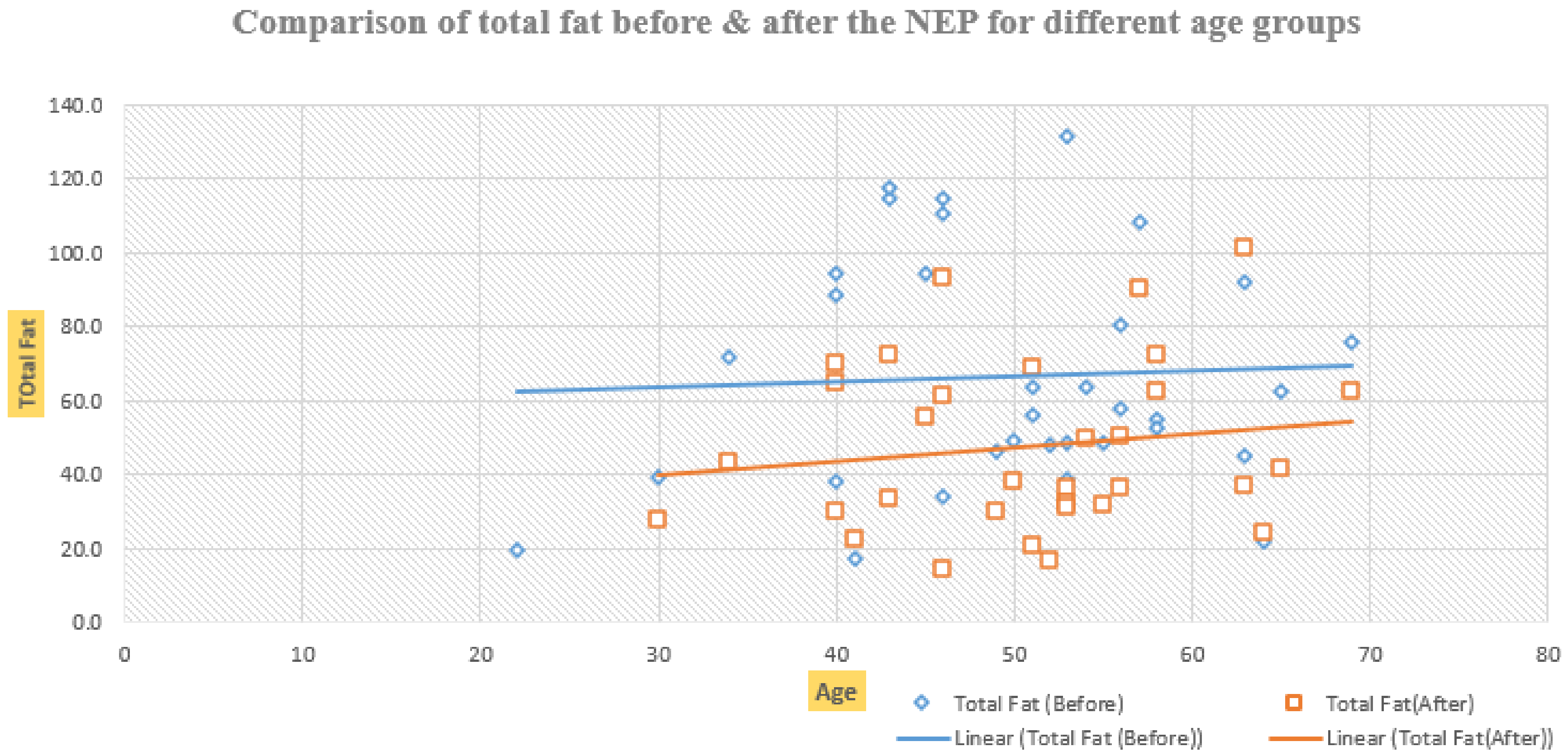
Weight1, Weight 2, Choles 2, DietC 1 and PDCF 1 are negatively correlated with age meaning an inverse relationship with age



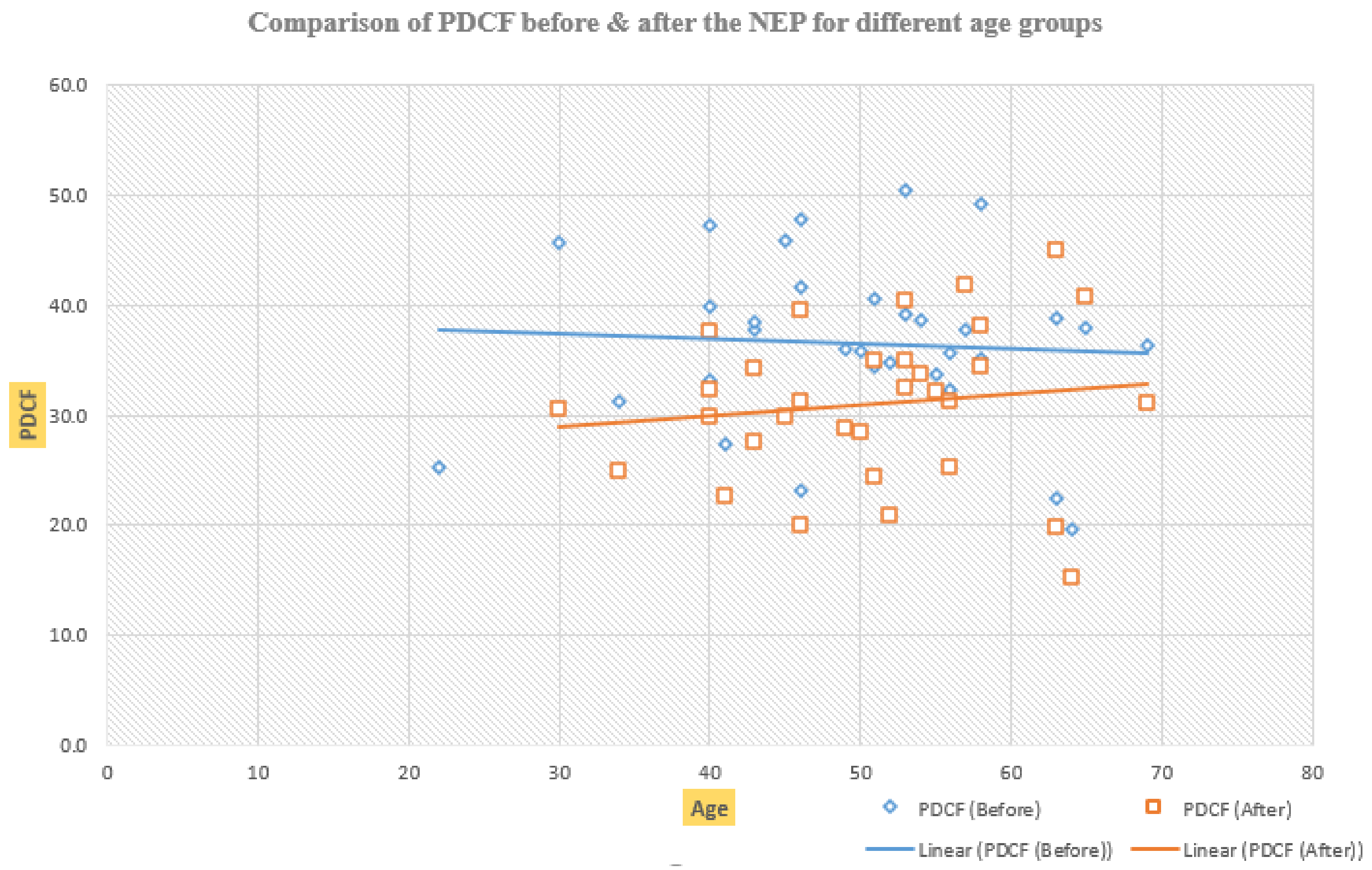
Comparison of weights before & after the NEP for different age groups
Weight
Age
Weight 1(Before)
Weight 2(After)
Linear (Weight 1(Before))
Linear (Weight 2(After))



Comparison of cholesterol level before & after the NEP for different age groups
Cholesterol
Age
Cholesterol1(Before)
Cholesterol2(After)
Linear (Cholesterol1(Before))
Linear (Cholesterol2(After))



Comparison of total fat before & after the NEP for different age groups
TOtal Fat
Age
Total Fat (Before)
Total Fat(After)
Linear (Total Fat (Before))
Linear (Total Fat(After))



Comparison of PDCF before & after the NEP for different age groups
PDCF
Age
PDCF (Before)
PDCF (After)
Linear (PDCF (Before))
Linear (PDCF (After))