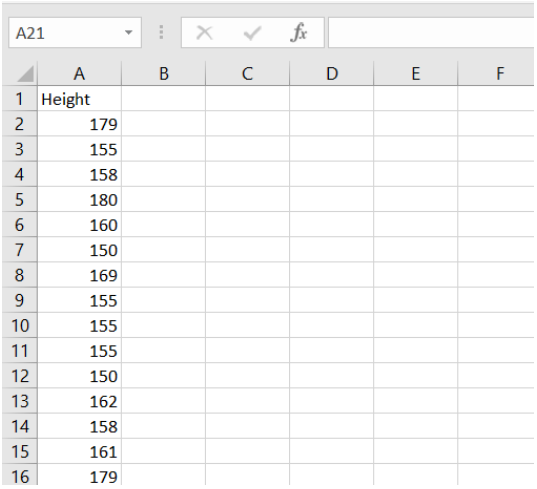
Practical 1

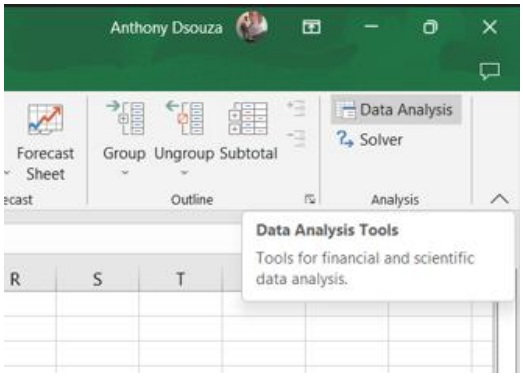
A.Write a program for obtaining descriptive statistics of data.

Step 1: Open your data in excel

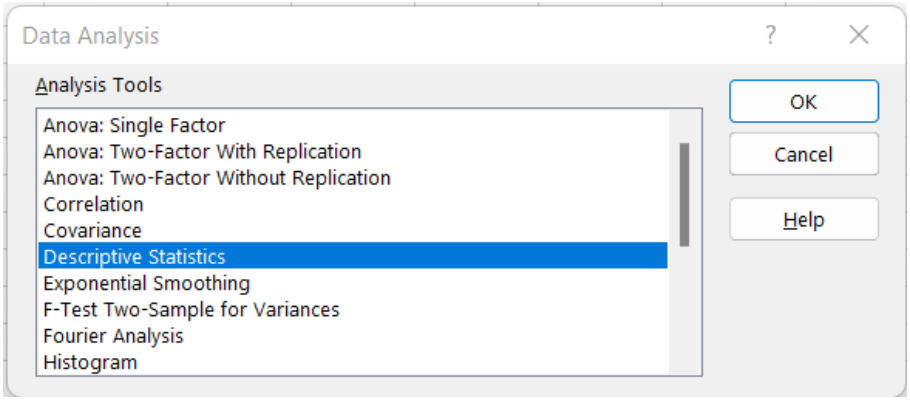


	A	B	C	D	E	F
1	Height					
2	179					
3	155					
4	158					
5	180					
6	160					
7	150					
8	169					
9	155					
10	155					
11	155					
12	150					
13	162					
14	158					
15	161					
16	179					

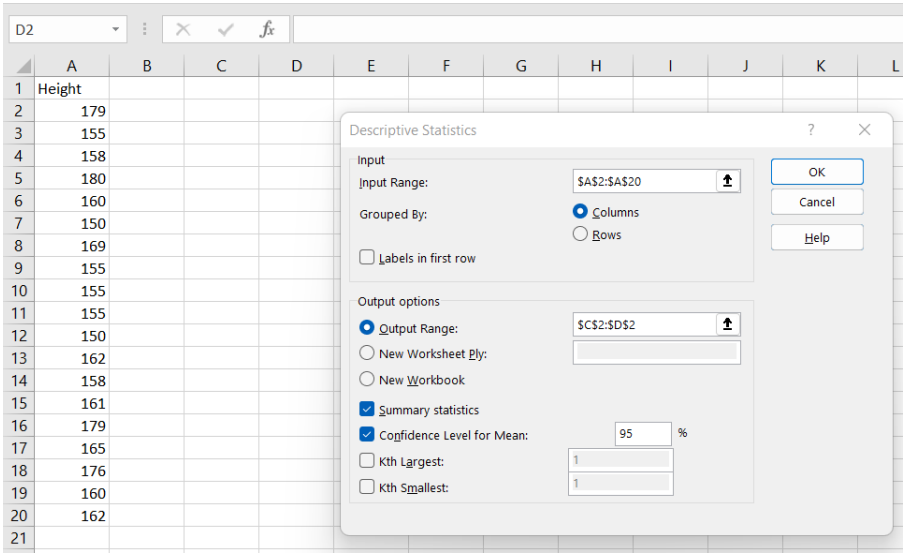
Step 2: From the Data Tool in the ribbon choose Data Analysis.



Step 3: Select the option of descriptive statistics



Step 4: Select an input range, output range, and check summary statistics and confidence level.



Step 5: The output may appear as follows.

1	Height		
2	179	*Column1*	
3	155		
4	158	Mean	162.5789474
5	180	Standard Error	2.212822533
6	160	Median	160
7	150	Mode	155
8	169	Standard Deviation	9.645469803
9	155	Sample Variance	93.03508772
10	155	Kurtosis	-0.561004927
11	155	Skewness	0.753286117
12	150	Range	30
13	162	Minimum	150
14	158	Maximum	180
15	161	Sum	3089
16	179	Count	19
17	165	Confidence Level(95.0%)	4.648967631
18	176		
19	160		
20	162		

Conclusion: we have successfully obtained the descriptive statistics in data.

B.Import data from different data sources (from Excel, csv, mysql, sql server, oracle to R /Python /Excel)

Code:

1 - SQLite3

```
import sqlite3 as sq

import pandas as pd

###############################################

Base='C:/VKHCG'

sDatabaseName=Base + '/01-Vermeulen/04-
Transform//SQLite/vermeulen.db'

conn = sq.connect(sDatabaseName)

###############################################

sFileName='C:/VKHCG/01-Vermeulen/01-Retrieve/01-EDS/02-
Python/Retrieve_IP_DATA.csv'

print('Loading :',sFileName)

IP_DATA_ALL_FIX=pd.read_csv(sFileName,header=0,low_memory=False)

IP_DATA_ALL_FIX.index.names = ['RowIDCSV']

sTable='IP_DATA_ALL'

print('Storing :',sDatabaseName,' Table:',sTable)

IP_DATA_ALL_FIX.to_sql(sTable, conn, if_exists="replace")

print('Loading :',sDatabaseName,' Table:',sTable)

TestData=pd.read_sql_query("select * from IP_DATA_ALL;", conn)

print('## Data Values')

print(TestData)

print('## Data Profile')
```

```
print('Rows :',TestData.shape[0])

print('Columns :',TestData.shape[1])

print('Successful')1
```

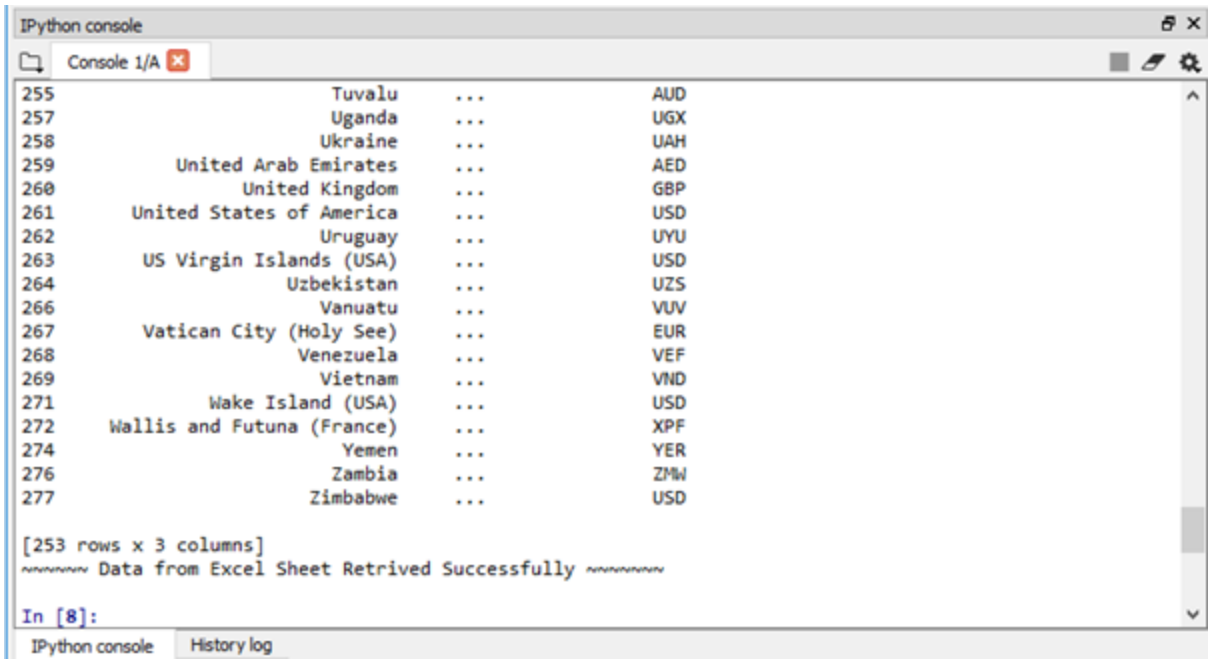
Output:

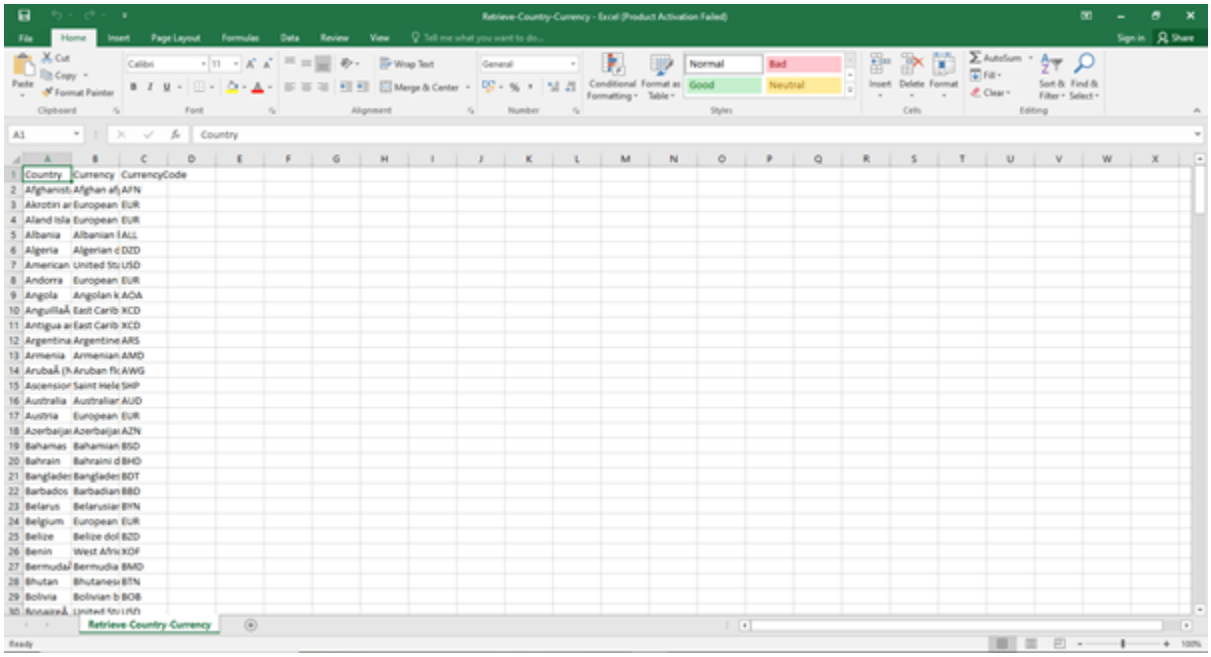
```
Loading : C:/VKHCG/01-Vermeulen/01-Retrieve/01-EDS/02-Python/
Retrieve_IP_DATA.csv
Storing : C:/VKHCG/01-Vermeulen/04-Transform//SQLite/
vermeulen.db  Table: IP_DATA_ALL
Loading : C:/VKHCG/01-Vermeulen/04-Transform//SQLite/
vermeulen.db  Table: IP_DATA_ALL
## Data Values
         RowIDCSV    RowID  ...  First.IP.Number
Last.IP.Number
0               0        0  ...        692781056
692781567
1               1        1  ...        692781824
692783103
2               2        2  ...        692909056
692909311
3               3        3  ...        692909568
692910079
4               4        4  ...        693051392
693052415
...           ...      ...  ...              ...
...
1247497   1247497  1247497  ...       1068157850
1068157850
1247498   1247498  1247498  ...       1334409600
1334409607
1247499   1247499  1247499  ...       1596886528
1596886783
1247500   1247500  1247500  ...       1742189568
1742190591
1247501   1247501  1247501  ...       1905782573
1905782573

[1247502 rows x 11 columns]
## Data Profile
Rows : 1247502
```

2 - Excel

```
import os

import pandas as pd

Base='C:/VKHCG'

sFileDir=Base + '/01-Vermeulen/01-Retrieve/01-EDS/02-Python'
```

```
CurrencyRawData = pd.read_excel('C:/VKHCG/01-Vermeulen/00-
RawData/Country_Currency.xlsx')

sColumns = ['Country or territory', 'Currency', 'ISO-4217']

CurrencyData = CurrencyRawData[sColumns]

CurrencyData.rename(columns={'Country or territory': 'Country',
'ISO-4217':

'CurrencyCode'}, inplace=True)

CurrencyData.dropna(subset=['Currency'],inplace=True)

CurrencyData['Country'] = CurrencyData['Country'].map(lambda x:
x.strip())

CurrencyData['Currency'] = CurrencyData['Currency'].map(lambda
x:

x.strip())

CurrencyData['CurrencyCode'] =
CurrencyData['CurrencyCode'].map(lambda x:

x.strip())

print(CurrencyData)

print('~~~~~~ Data from Excel Sheet Retrived Successfully
~~~~~~~')

sFileName=sFileDir + '/Retrieve-Country-Currency.csv'

CurrencyData.to_csv(sFileName, index = False)
```



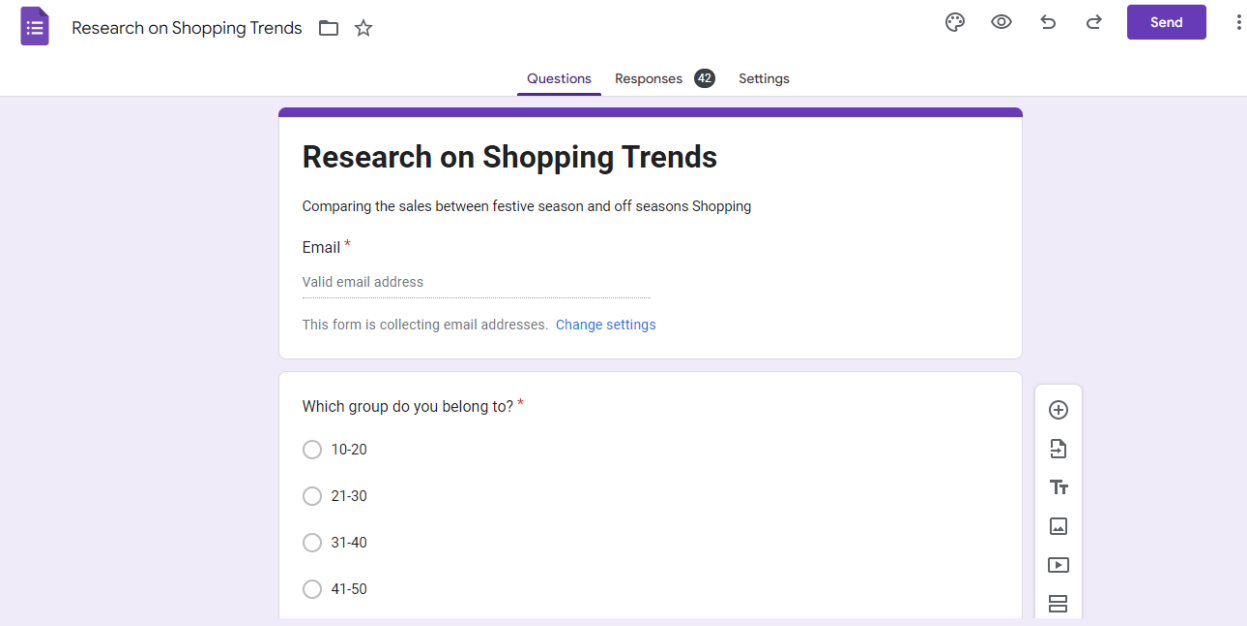


Conclusion: we have successfully imported data from SQLite3 and Excel in this program.

Practical 2

A.Design a survey form for a given case study, collect the primary data and analyse it.

Step 1: In order to make a survey we used Google forms. Our case study's aim was to find out which age group did Shopping in festive seasons.



Step 2: The questions and options entered were as follows.

Q1.Which group do you belong to?
(a)10-20 (b)21-30 (c)31-40 (d)41-50 (e)51 Above

Q2.What is your Gender?
(a)Male (b)Female

Q3.What is your Occupation?
(a)Business (b)Student (c)Employee (d)Retired (e)Self Employed (f)Household

Q4.What's your Monthly Income?
(a)10,000 - 20,000 (b)21,000 - 30,000 (c)31,000 - 40,000 (d)41,000 and above (e)N/A

Q5.How often do you shop?

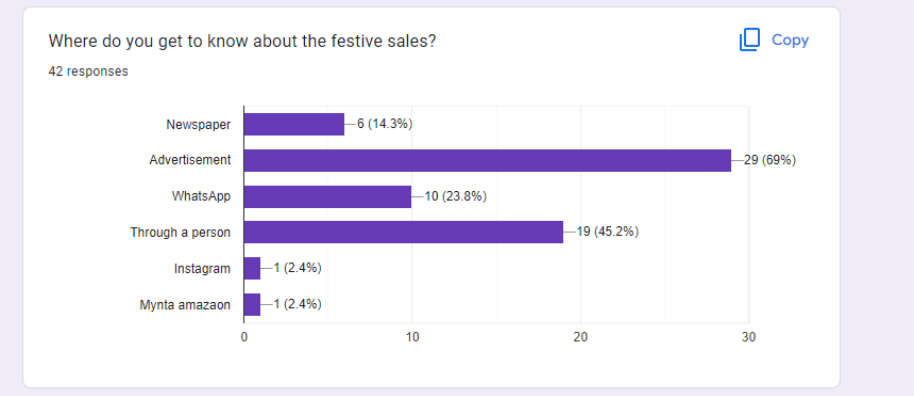
(a)Monthly (b)Occasionaly (c)Festive Seasons (d)Rarely

Q6.Purpose of Shopping?
(a)Personal Use (b)Gifting

Q7.Where do you get to know about the festive sales?
(a)Newspaper (b)Advertisement (c)WhatsApp (d)Through a person

Q8.How would you rate your overall shopping experience during festive season?
(a)Excellent (b)Good (c)Average (d)Poor

Step 3 : Google forms tool was chosen because analysis becomes easier on Google forms.



Conclusion: we have successfully made the survey form for the given case study.

B.Perform analysis of given secondary data.

Step 1: Open Data in excel

1	age	male	female	total
2	0 to 4	328759	307079	635838
3	4 to 9	315119	293664	608783
4	10 to 14	311456	290598	602054
5	15 to 19	312831	293313	606144
6	20 to 24	311077	295739	606816
7	25 to 29	284258	273379	557637
8	30 to 34	255596	247383	502979
9	35 to 39	248575	241938	490513
10	40 to 44	232217	226914	459131
11	45 to 49	202633	201142	403775
12	50 to 54	176241	176440	352681
13	55 to 59	153494	156283	309777
14	60 to 64	114194	121200	235394
15	65 to 69	83129	92071	175200
16	70 to 74	65266	77990	143256
17	75 to 79	43761	56895	100656
18	80 to 84	25060	37873	62933
19	85+	14164	28156	42320
20				
21		3477830	3418057	6895887

Step 2: Calculate the total sum of each column. Select the cell for Sum→ add formula SUM in formula bar→ select the range.

1	age	male	female	total
2	0 to 4	328759	307079	635838
3	4 to 9	315119	293664	608783
4	10 to 14	311456	290598	602054
5	15 to 19	312831	293313	606144
6	20 to 24	311077	295739	606816
7	25 to 29	284258	273379	557637
8	30 to 34	255596	247383	502979
9	35 to 39	248575	241938	490513
10	40 to 44	232217	226914	459131
11	45 to 49	202633	201142	403775
12	50 to 54	176241	176440	352681
13	55 to 59	153494	156283	309777
14	60 to 64	114194	121200	235394
15	65 to 69	83129	92071	175200
16	70 to 74	65266	77990	143256
17	75 to 79	43761	56895	100656
18	80 to 84	25060	37873	62933
19	85+	14164	28156	42320
20				
21		3477830	3418057	6895887

Step 3: Calculate the percentage of male in cell E. Use formula -1*100*B2/$D$21

SUM | =-1*100*B2/$D$21

	A	B	C	D	E	F
1	age	male	female	total	Male(%)	
2	0 to 4	328759	307079	635838	=-1*100*B2/$D$21	
3	4 to 9	315119	293664	608783		
4	10 to 14	311456	290598	602054		
5	15 to 19	312831	293313	606144		
6	20 to 24	311077	295739	606816		
7	25 to 29	284258	273379	557637		
8	30 to 34	255596	247383	502979		
9	35 to 39	248575	241938	490513		
10	40 to 44	232217	226914	459131		
11	45 to 49	202633	201142	403775		
12	50 to 54	176241	176440	352681		

Step 4: Calculate the percentage female in cell F. Use formula 100*C2/$D$21

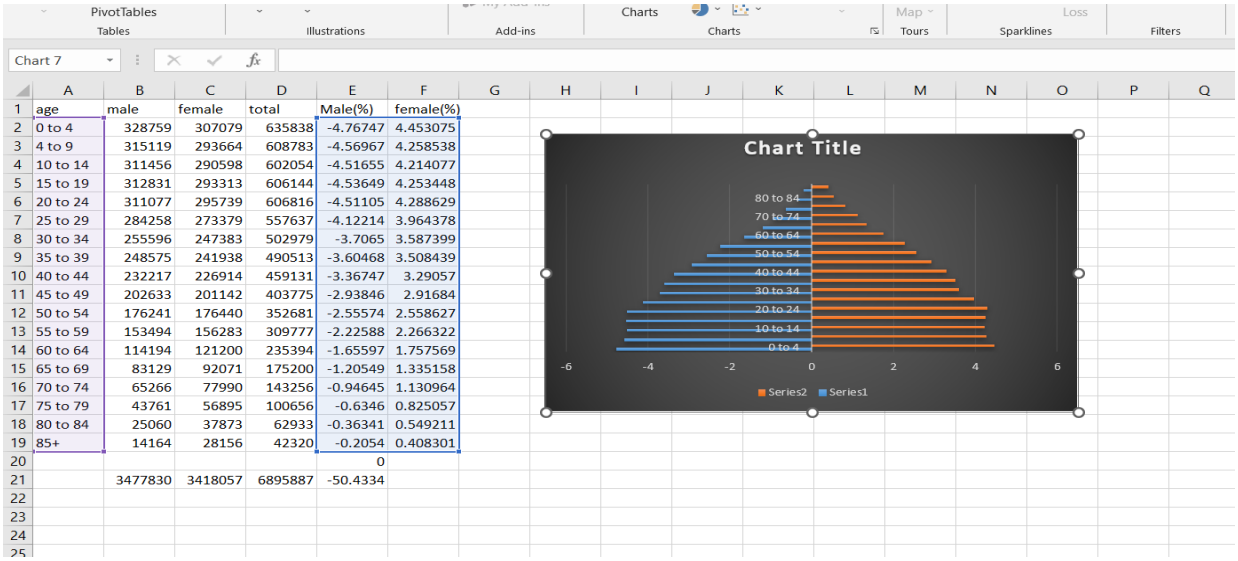
F2 | =100*C2/$D$21

	A	B	C	D	E	F
1	age	male	female	total	Male(%)	female(%)
2	0 to 4	328759	307079	635838	-4.76747	4.453075
3	4 to 9	315119	293664	608783	-4.56967	4.258538
4	10 to 14	311456	290598	602054	-4.51655	4.214077
5	15 to 19	312831	293313	606144	-4.53649	4.253448
6	20 to 24	311077	295739	606816	-4.51105	4.288629
7	25 to 29	284258	273379	557637	-4.12214	3.964378
8	30 to 34	255596	247383	502979	-3.7065	3.587399
9	35 to 39	248575	241938	490513	-3.60468	3.508439
10	40 to 44	232217	226914	459131	-3.36747	3.29057
11	45 to 49	202633	201142	403775	-2.93846	2.91684
12	50 to 54	176241	176440	352681	-2.55574	2.558627
13	55 to 59	153494	156283	309777	-2.22588	2.266322
14	60 to 64	114194	121200	235394	-1.65597	1.757569
15	65 to 69	83129	92071	175200	-1.20549	1.335158

Step 5: The modified data may seem as follows.

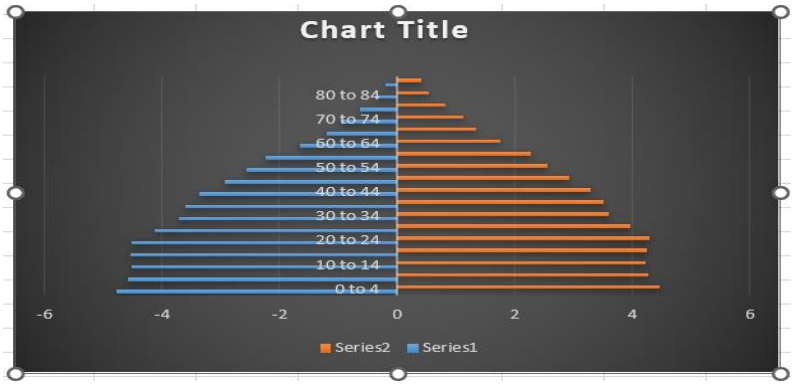
	A	B	C	D	E	F
1	age	male	female	total	Male(%)	female(%)
2	0 to 4	328759	307079	635838	-4.76747	4.453075
3	4 to 9	315119	293664	608783	-4.56967	4.258538
4	10 to 14	311456	290598	602054	-4.51655	4.214077
5	15 to 19	312831	293313	606144	-4.53649	4.253448
6	20 to 24	311077	295739	606816	-4.51105	4.288629
7	25 to 29	284258	273379	557637	-4.12214	3.964378
8	30 to 34	255596	247383	502979	-3.7065	3.587399
9	35 to 39	248575	241938	490513	-3.60468	3.508439
10	40 to 44	232217	226914	459131	-3.36747	3.29057
11	45 to 49	202633	201142	403775	-2.93846	2.91684
12	50 to 54	176241	176440	352681	-2.55574	2.558627
13	55 to 59	153494	156283	309777	-2.22588	2.266322

Step 6: For analysis go to Insert → Bar → 2D



	A	B	C	D	E	F
1	age	male	female	total	Male(%)	female(%)
2	0 to 4	328759	307079	635838	-4.76747	4.453075
3	4 to 9	315119	293664	608783	-4.56967	4.258538
4	10 to 14	311456	290598	602054	-4.51655	4.214077
5	15 to 19	312831	293313	606144	-4.53649	4.253448
6	20 to 24	311077	295739	606816	-4.51105	4.288629
7	25 to 29	284258	273379	557637	-4.12214	3.964378
8	30 to 34	255596	247383	502979	-3.7065	3.587399
9	35 to 39	248575	241938	490513	-3.60468	3.508439
10	40 to 44	232217	226914	459131	-3.36747	3.29057
11	45 to 49	202633	201142	403775	-2.93846	2.91684
12	50 to 54	176241	176440	352681	-2.55574	2.558627
13	55 to 59	153494	156283	309777	-2.22588	2.266322
14	60 to 64	114194	121200	235394	-1.65597	1.757569
15	65 to 69	83129	92071	175200	-1.20549	1.335158
16	70 to 74	65266	77990	143256	-0.94645	1.130964
17	75 to 79	43761	56895	100656	-0.6346	0.825057
18	80 to 84	25060	37873	62933	-0.36341	0.549211
19	85+	14164	28156	42320	-0.2054	0.408301
20					0	
21		3477830	3418057	6895887	-50.4334	

Step 7 :Drag the data and set the graph for analysis.



Conclusion: we have successfully performed the analysis for the given secondary data.

Practical 3

A.Perform testing of hypothesis using one sample t-test.

Code:

```
from scipy.stats import ttest_1samp

import numpy as np

ages=np.genfromtxt('H:/ages.csv')

print(ages)

ages_mean=np.mean(ages)

print(ages_mean)

tset,pval=ttest_1samp(ages,30)

print('p-values-',pval)

if pval<0.05:#alpha value is 0.05

     print("we are rejecting null hypothesis")

else:

     print("we are accepeting null hypothesis")
```

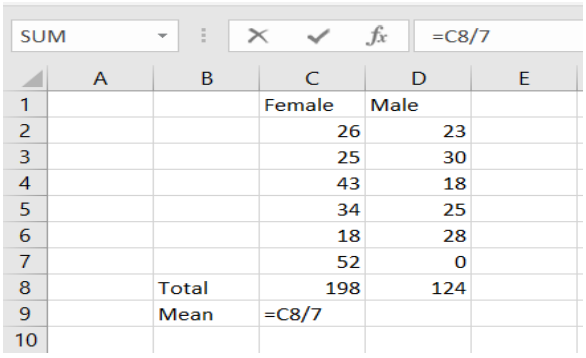
Output:

```
In [3]: runfile('H:/3a.py', wdir='H:')
[20. 30. 25. 13. 16. 17. 34. 35. 38. 42. 43. 45. 48. 49. 50. 51. 54. 55.
 56. 59. 61. 62. 18. 22. 29. 30. 31. 39. 52. 53. 67. 36. 47. 54. 40. 40.
 35. 22. 59. 58. 30. 43. 22. 45. 21. 59. 51. 47. 25. 58. 50. 23. 24. 45.
 37. 59. 28. 28. 48. 42. 54. 36. 36. 24. 26. 24. 50. 48. 34. 44. 56. 55.
 35. 33. 39. 53. 34. 28. 56. 24. 21. 29. 28. 58. 35. 57. 26. 25. 59. 56.
 22. 57. 48. 33. 23. 26. 57. 32. 53. 31. 35. 44. 54. 25. 31. 58. 26. 32.
 26. 50. 41. 49. 26. 33. 34. 24. 43. 42. 51. 36. 38. 38. 40. 38. 56. 39.
 23. 33. 53. 30. 38.]
39.47328244274809
p-values- 5.362905195437013e-14
we are rejecting null hypothesis
```

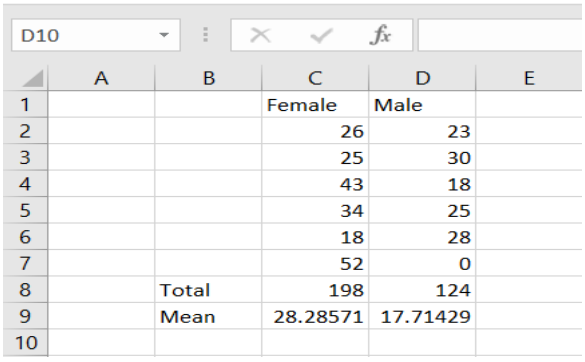
Conclusion: we have successfully performed testing of hypothesis using one sample t-test.

B.Write a program for t-test comparing two means for independent samples.

Step 1: Calculate the mean of the samples (Total /Count)

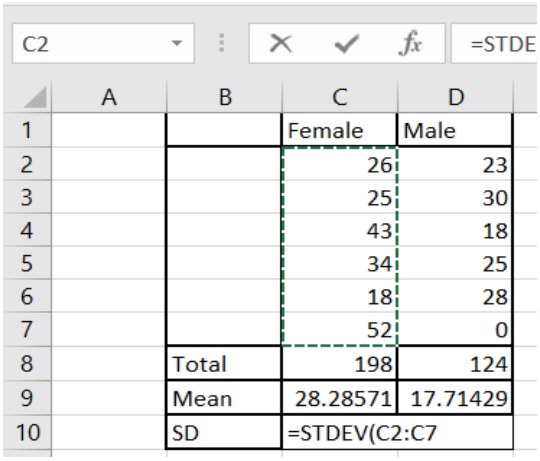


	A	B	C	D	E
1			Female	Male	
2			26	23	
3			25	30	
4			43	18	
5			34	25	
6			18	28	
7			52	0	
8		Total	198	124	
9		Mean	=C8/7		
10					



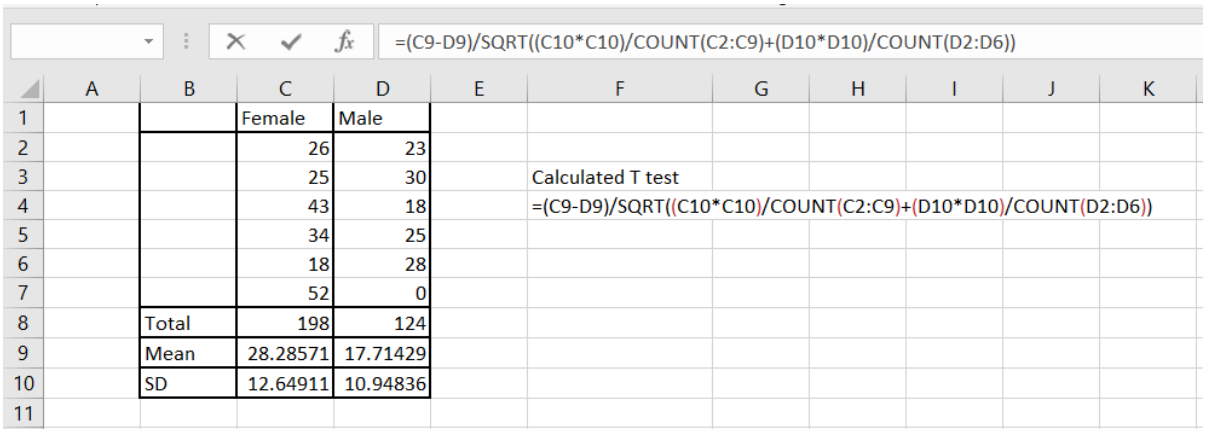
	A	B	C	D	E
1			Female	Male	
2			26	23	
3			25	30	
4			43	18	
5			34	25	
6			18	28	
7			52	0	
8		Total	198	124	
9		Mean	28.28571	17.71429	
10					

Step 2: Calculate the SD of samples STDEV(range)



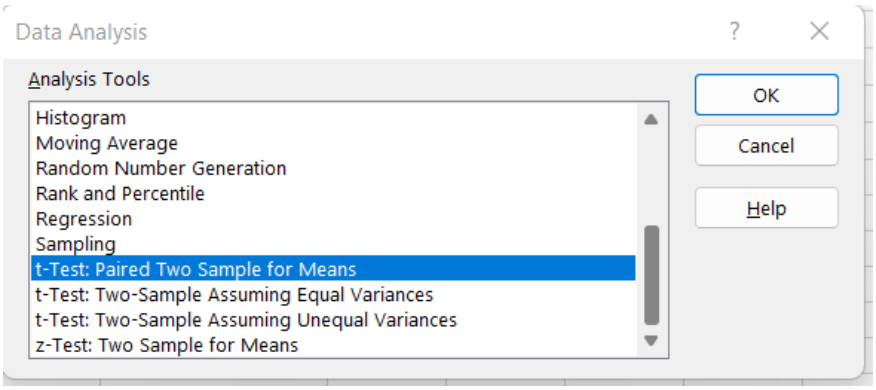
	A	B	C	D
1			Female	Male
2			26	23
3			25	30
4			43	18
5			34	25
6			18	28
7			52	0
8		Total	198	124
9		Mean	28.28571	17.71429
10		SD	=STDEV(C2:C7	

Step 3: Calculate the t-test

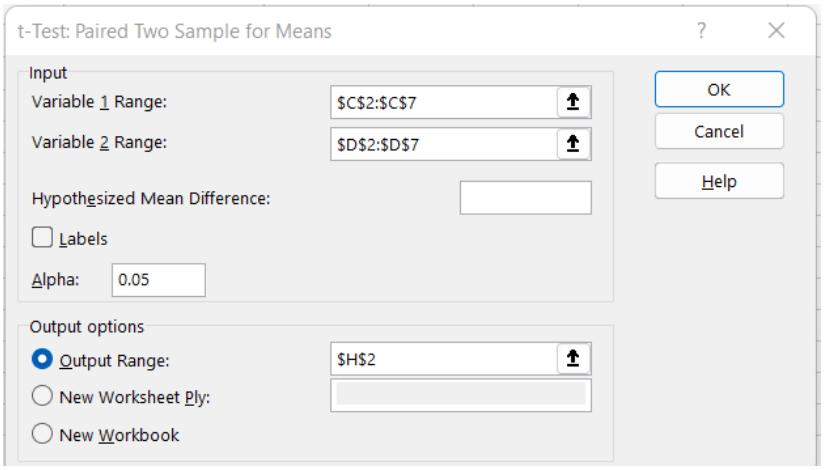


	A	B	C	D	E	F	G	H	I	J	K
1			Female	Male							
2			26	23							
3			25	30		Calculated T test					
4			43	18		=(C9-D9)/SQRT((C10*C10)/COUNT(C2:C9)+(D10*D10)/COUNT(D2:D6))					
5			34	25							
6			18	28							
7			52	0							
8		Total	198	124							
9		Mean	28.28571	17.71429							
10		SD	12.64911	10.94836							
11											

Step 4: Apply t-test paired two Samples for Means.



Step 5: Apply input Range and Output range.



Step 6: Hence we reject null hypothesis.

H15 | t Critical two-tail

	A	B	C	D	E	F	G	H	I	J	K
1			Female	Male							
2			26	23				t-Test: Paired Two Sample for Means			
3			25	30		Calculated T test					
4			43	18		1.594185981			*Variable 1*	*Variable 2*	
5			34	25				Mean	33	20.66666667	
6			18	28				Variance	160	119.8666667	
7			52	0				Observations	6	6	
8		Total	198	124				Pearson Correlation	-0.889613782		
9		Mean	28.28571	17.71429				Hypothesized Mean Difference	0		
10		SD	12.64911	10.94836				df	5		
11								t Stat	1.316901108		
12								P(T<=t) one-tail	0.122498157		
13								t Critical one-tail	2.015048373		
14								P(T<=t) two-tail	0.244996315		
15								t Critical two-tail	2.570581836		

Conclusion: we have successfully compared the two means of independent samples for a t-test.

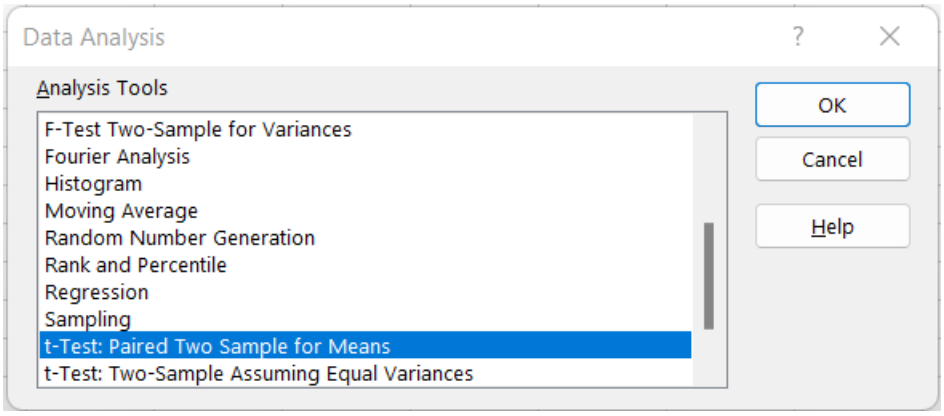
C.Perform testing of Hypothesis using paired t-test.

Step 1: Load the Data

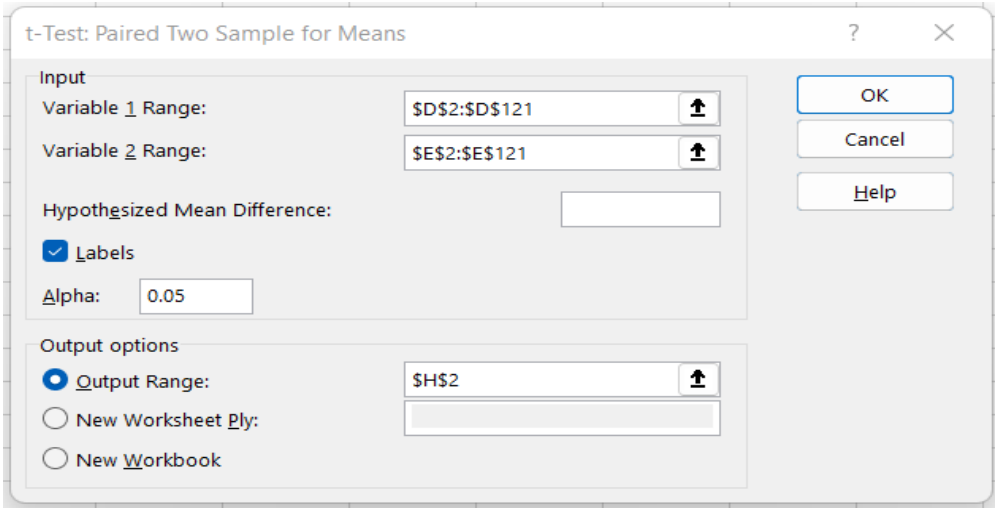
L13

	A	B	C	D	E	F
1	patient	gender	agegrp	bp_before	bp_after	
2	1	Male	30-45	143	153	
3	2	Male	30-45	163	170	
4	3	Male	30-45	153	168	
5	4	Male	30-45	153	142	
6	5	Male	30-45	146	141	
7	6	Male	30-45	150	147	
8	7	Male	30-45	148	133	
9	8	Male	30-45	153	141	
10	9	Male	30-45	153	131	
11	10	Male	30-45	158	125	
12	11	Male	30-45	149	164	
13	12	Male	30-45	173	159	

Step 2: Data -> Data Analysis -> t-test between samples for means.



Step 3: Choose the Input and Output range.



Step 4: Since the samples Means of the Data samples are not equal we reject the null hypothesis that they might be equal.

H21

	A	B	C	D	E	F	G	H	I	J	K
1	patient	gender	agegrp	bp_before	bp_after						
2	1	Male	30-45	143	153			t-Test: Paired Two Sample for Means			
3	2	Male	30-45	163	170						
4	3	Male	30-45	153	168				*143*	*153*	
5	4	Male	30-45	153	142			Mean	156.5630252	151.3445378	
6	5	Male	30-45	146	141			Variance	129.2820111	202.6853725	
7	6	Male	30-45	150	147			Observations	119	119	
8	7	Male	30-45	148	133			Pearson Correlation	0.161241417		
9	8	Male	30-45	153	141			Hypothesized Mean Difference	0		
10	9	Male	30-45	153	131			df	118		
11	10	Male	30-45	158	125			t Stat	3.403463555		
12	11	Male	30-45	149	164			P(T<=t) one-tail	0.000454744		
13	12	Male	30-45	173	159			t Critical one-tail	1.657869522		
14	13	Male	30-45	165	135			P(T<=t) two-tail	0.000909488		
15	14	Male	30-45	145	159			t Critical two-tail	1.980272249		
16	15	Male	30-45	143	153						
17	16	Male	30-45	152	126						
18	17	Male	30-45	141	162						
19	18	Male	30-45	176	134						
20	19	Male	30-45	143	136						
21	20	Male	30-45	162	150						

Conclusion: we have performed the testing of hypothesis using paired t-test.

Practical 4

A.Perform testing of hypothesis using chi-squared goodness-of-fit test.

Step 1: Load the data.

D2

	A	B	C	D
1	O	E		
2	29	21.33		
3	24	21.33		
4	22	21.33		
5	19	21.33		
6	21	21.33		
7	18	21.33		
8	19	21.33		
9	20	21.33		
10	21	21.33		
11	18	21.33		
12	20	21.33		
13	23	21.33		
14				

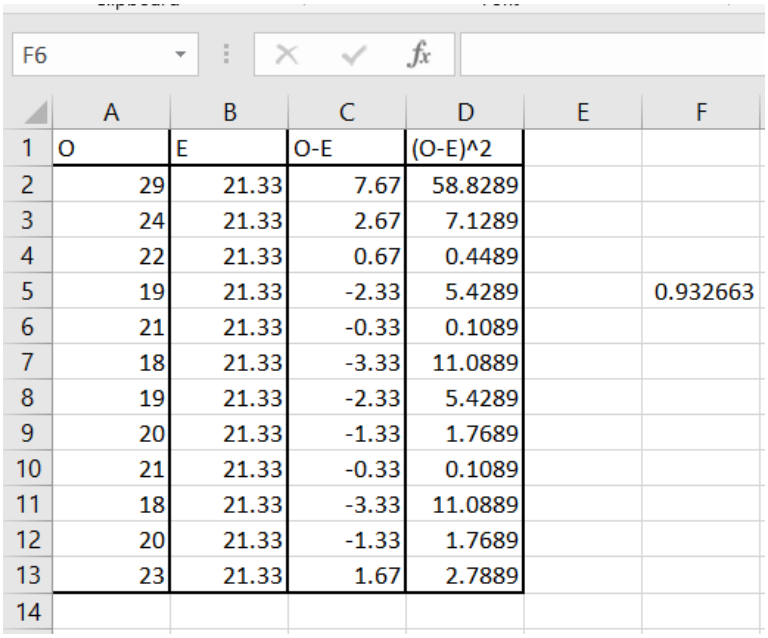
Step 2: Type CHITEST and select Actual(observed value range) and Expected value range.

F5 =CHITEST(A2:A13,B2:B13)

	A	B	C	D	E	F	G	H
1	O	E	O-E	(O-E)^2				
2	29	21.33	7.67	58.8289				
3	24	21.33	2.67	7.1289				
4	22	21.33	0.67	0.4489				
5	19	21.33	-2.33	5.4289		=CHITEST(A2:A13,B2:B13)		
6	21	21.33	-0.33	0.1089				
7	18	21.33	-3.33	11.0889				
8	19	21.33	-2.33	5.4289				
9	20	21.33	-1.33	1.7689				
10	21	21.33	-0.33	0.1089				
11	18	21.33	-3.33	11.0889				
12	20	21.33	-1.33	1.7689				
13	23	21.33	1.67	2.7889				
14								

Step 3: This calculated value is less than table value which is 19.68. Hence we accept null hypothesis.

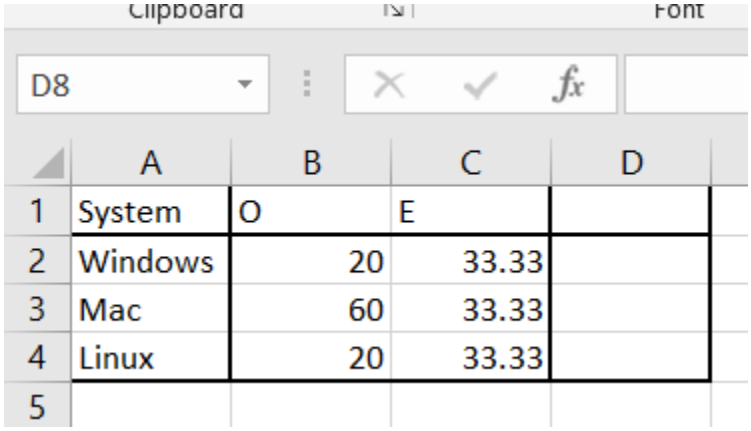
i.e 0.932663 < 19.68. H0 accepted.



	A	B	C	D	E	F
1	O	E	O-E	(O-E)^2		
2	29	21.33	7.67	58.8289		
3	24	21.33	2.67	7.1289		
4	22	21.33	0.67	0.4489		
5	19	21.33	-2.33	5.4289		0.932663
6	21	21.33	-0.33	0.1089		
7	18	21.33	-3.33	11.0889		
8	19	21.33	-2.33	5.4289		
9	20	21.33	-1.33	1.7689		
10	21	21.33	-0.33	0.1089		
11	18	21.33	-3.33	11.0889		
12	20	21.33	-1.33	1.7689		
13	23	21.33	1.67	2.7889		
14						

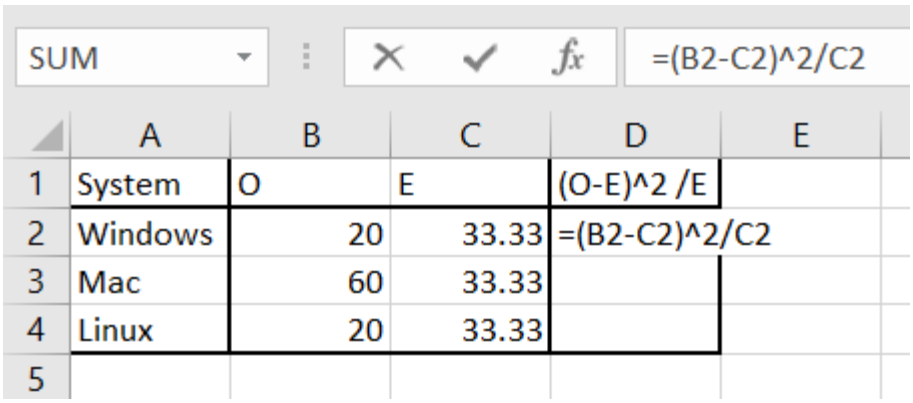
B.Perform testing of hypothesis using chi-squared test of independence.

Step 1: Load the data

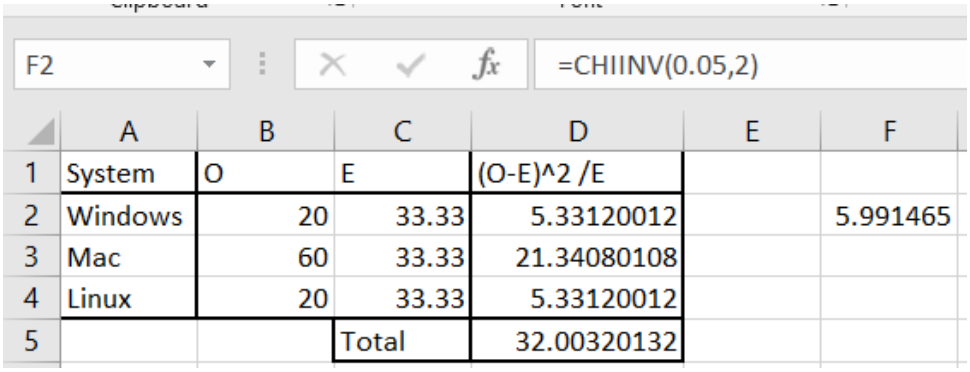


	A	B	C	D
1	System	O	E	
2	Windows	20	33.33	
3	Mac	60	33.33	
4	Linux	20	33.33	
5				

Step 2: Calculate the value of (O-E)^2 /E



Step 3: Calculate Chi square at 5% confidence and degree of freedom n-1=2(in our case)

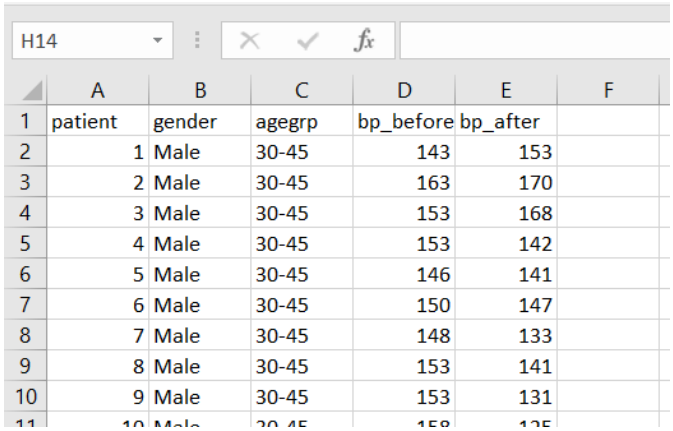


Conclusion: we have successfully performed the chi-squared test.

Practical 5

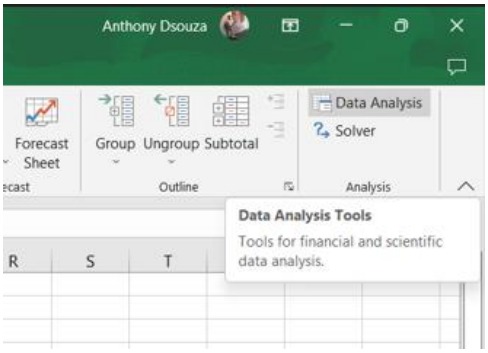
Perform testing of hypothesis using Z-test.

Step 1: Load the data

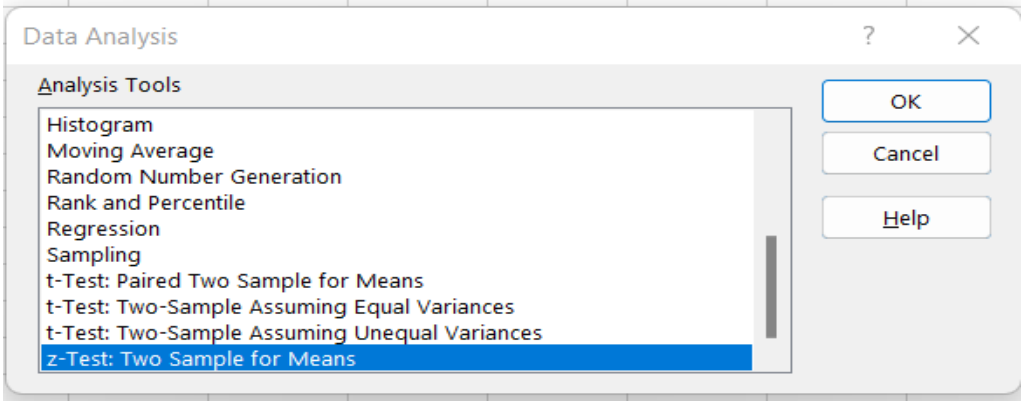


	A	B	C	D	E	F
1	patient	gender	agegrp	bp_before	bp_after	
2	1	Male	30-45	143	153	
3	2	Male	30-45	163	170	
4	3	Male	30-45	153	168	
5	4	Male	30-45	153	142	
6	5	Male	30-45	146	141	
7	6	Male	30-45	150	147	
8	7	Male	30-45	148	133	
9	8	Male	30-45	153	141	
10	9	Male	30-45	153	131	

Step 2: To apply Z test we need a sample size over 30. Here our sample size is 120 data points , so to apply Z-test go to Data -> Data analysis.



Step 3: Select Z-Test: Two Sample for means.

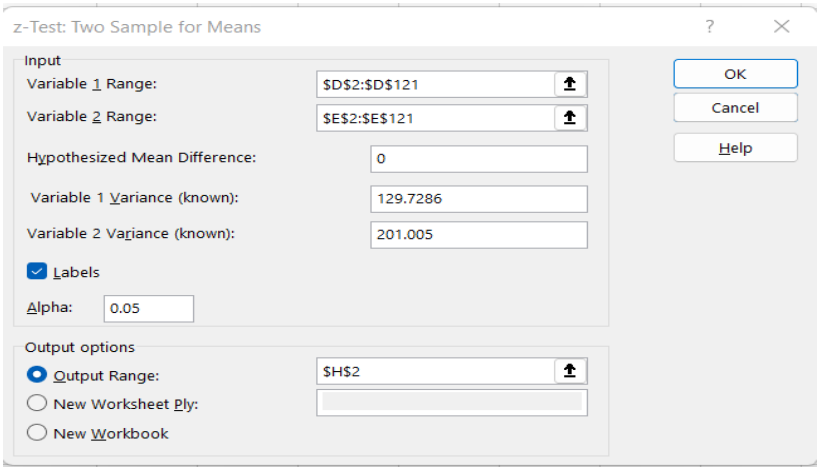


Step 4: Varience can be calculated using VAR.S (range)

G2 =VAR.S(D2:D121)

	A	B	C	D	E	F	G	H
1	patient	gender	agegrp	bp_before	bp_after			
2	1	Male	30-45	143	153		129.7286	
3	2	Male	30-45	163	170			
4	3	Male	30-45	153	168		201.005	
5	4	Male	30-45	153	142			
6	5	Male	30-45	146	141			
7	6	Male	30-45	150	147			
8	7	Male	30-45	148	133			
9	8	Male	30-45	153	141			

Step 5: Set the Variable 1 and 2 range.



Step 6: Output

G2 =VAR.S(D2:D121)

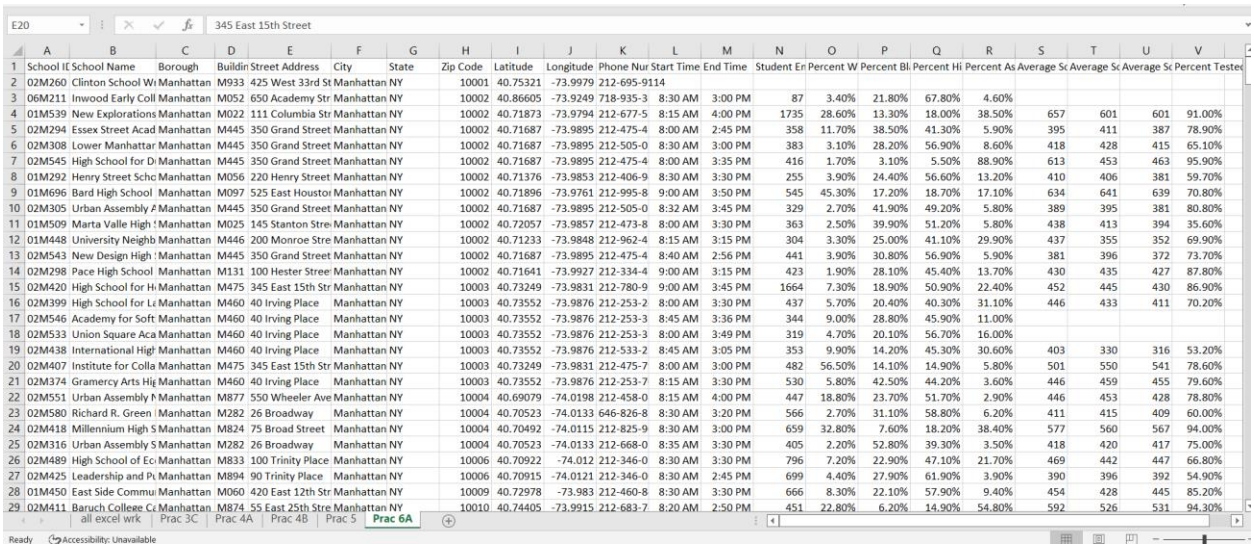
	A	B	C	D	E	F	G	H	I	J	K	L
1	patient	gender	agegrp	bp_before	bp_after							
2	1	Male	30-45	143	153		129.7286			z-Test: Two Sample for Means		
3	2	Male	30-45	163	170							
4	3	Male	30-45	153	168		201.005				*143*	*153*
5	4	Male	30-45	153	142					Mean	156.5630252	151.3445378
6	5	Male	30-45	146	141					Known Variance	129.7286	201.005
7	6	Male	30-45	150	147					Observations	119	119
8	7	Male	30-45	148	133					Hypothesized Mean Difference	0	
9	8	Male	30-45	153	141					z	3.13024954	
10	9	Male	30-45	153	131					P(Z<=z) one-tail	0.000873289	
11	10	Male	30-45	158	125					z Critical one-tail	1.644853627	
12	11	Male	30-45	149	164					P(Z<=z) two-tail	0.001746579	
13	12	Male	30-45	173	159					z Critical two-tail	1.959963985	
14	13	Male	30-45	165	135							

Conclusion: we have successfully performed the Z-Test hypothesis.

Practical 6

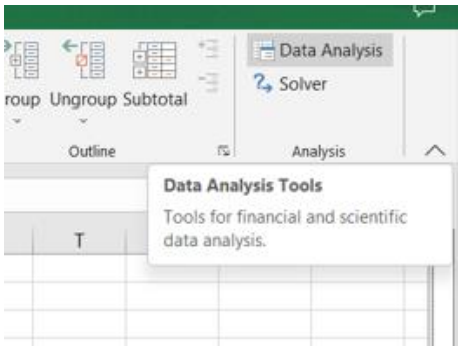
A.Perform testing of hypothesis using One-way ANOVA.

Step 1: Load the data

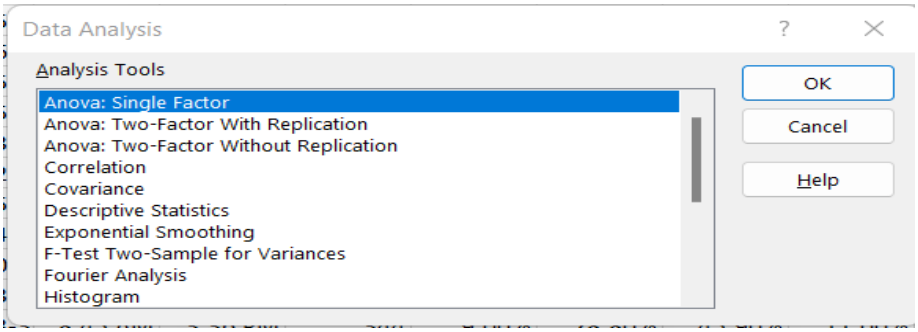


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	School ID	School Name	Borough	Building	Street Address	City	State	Zip Code	Latitude	Longitude	Phone Nur	Start Time	End Time	Student En	Percent W	Percent Bl	Percent Hi	Percent As	Average Sc	Average Sc	Average Sc	Percent Tested
2	02M260	Clinton School Wi	Manhattan	M933	425 West 33rd St	Manhattan	NY	10001	40.75321	-73.9979	212-695-9114											
3	06M211	Inwood Early Coll	Manhattan	M052	650 Academy Str	Manhattan	NY	10002	40.86605	-73.9249	718-935-3	8:30 AM	3:00 PM	87	3.40%	21.80%	67.80%	4.60%				
4	01M539	New Explorations	Manhattan	M022	111 Columbia Str	Manhattan	NY	10002	40.71873	-73.9794	212-677-5	8:15 AM	4:00 PM	1735	28.60%	13.30%	18.00%	38.50%	657	601	601	91.00%
5	02M294	Essex Street Acad	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.9895	212-475-4	8:00 AM	2:45 PM	358	11.70%	38.50%	41.30%	5.90%	395	411	387	78.90%
6	02M308	Lower Manhattar	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.9895	212-505-0	8:30 AM	3:00 PM	383	3.10%	28.20%	56.90%	8.60%	418	428	415	65.10%
7	02M545	High School for Di	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.9895	212-475-4	8:00 AM	3:35 PM	416	1.70%	3.10%	5.50%	88.90%	613	453	463	95.90%
8	01M292	Henry Street Schc	Manhattan	M056	220 Henry Street	Manhattan	NY	10002	40.71376	-73.9853	212-406-9	8:30 AM	3:30 PM	255	3.90%	24.40%	56.60%	13.20%	410	406	381	59.70%
9	01M696	Bard High School	Manhattan	M097	525 East Houstor	Manhattan	NY	10002	40.71896	-73.9761	212-995-8	9:00 AM	3:50 PM	545	45.30%	17.20%	18.70%	17.10%	634	641	639	70.80%
10	02M305	Urban Assembly A	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.9895	212-505-0	8:32 AM	3:45 PM	329	2.70%	41.90%	49.20%	5.80%	389	395	381	80.80%
11	01M509	Marta Valle High S	Manhattan	M025	145 Stanton Stre	Manhattan	NY	10002	40.72057	-73.9857	212-473-8	8:00 AM	3:30 PM	363	2.50%	39.90%	51.20%	5.80%	438	413	394	35.60%
12	01M448	University Neighb	Manhattan	M446	200 Monroe Stre	Manhattan	NY	10002	40.71233	-73.9848	212-962-4	8:15 AM	3:15 PM	304	3.30%	25.00%	41.10%	29.90%	437	355	352	69.90%
13	02M543	New Design High S	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.9895	212-475-4	8:40 AM	2:56 PM	441	3.90%	30.80%	56.90%	5.90%	381	396	372	73.70%
14	02M298	Pace High School	Manhattan	M131	100 Hester Stree	Manhattan	NY	10002	40.71641	-73.9927	212-334-4	9:00 AM	3:15 PM	423	1.90%	28.10%	45.40%	13.70%	430	435	427	87.80%
15	02M420	High School for Hi	Manhattan	M475	345 East 15th Str	Manhattan	NY	10003	40.73249	-73.9831	212-780-9	9:00 AM	3:45 PM	1664	7.30%	18.90%	50.90%	22.40%	452	445	430	86.90%
16	02M399	High School for Le	Manhattan	M460	40 Irving Place	Manhattan	NY	10003	40.73552	-73.9876	212-253-2	8:00 AM	3:30 PM	437	5.70%	20.40%	40.30%	31.10%	446	433	411	70.20%
17	02M546	Academy for Soft	Manhattan	M460	40 Irving Place	Manhattan	NY	10003	40.73552	-73.9876	212-253-3	8:45 AM	3:36 PM	344	9.00%	28.80%	45.90%	11.00%				
18	02M533	Union Square Aca	Manhattan	M460	40 Irving Place	Manhattan	NY	10003	40.73552	-73.9876	212-253-3	8:00 AM	3:49 PM	319	4.70%	20.10%	56.70%	16.00%				
19	02M438	International High	Manhattan	M460	40 Irving Place	Manhattan	NY	10003	40.73552	-73.9876	212-533-2	8:45 AM	3:05 PM	353	9.90%	14.20%	45.30%	30.60%	403	330	316	53.20%
20	02M407	Institute for Colla	Manhattan	M475	345 East 15th Str	Manhattan	NY	10003	40.73249	-73.9831	212-475-7	8:00 AM	3:00 PM	482	56.50%	14.10%	14.90%	5.80%	501	550	541	78.60%
21	02M374	Gramercy Arts Hig	Manhattan	M460	40 Irving Place	Manhattan	NY	10003	40.73552	-73.9876	212-253-7	8:15 AM	3:30 PM	530	5.80%	42.50%	44.20%	3.60%	446	459	455	79.60%
22	02M551	Urban Assembly M	Manhattan	M877	550 Wheeler Ave	Manhattan	NY	10004	40.69079	-74.0198	212-458-0	8:15 AM	4:00 PM	447	18.80%	23.70%	51.70%	2.90%	446	453	428	78.80%
23	02M580	Richard R. Green	Manhattan	M282	26 Broadway	Manhattan	NY	10004	40.70523	-74.0133	646-826-8	8:30 AM	3:20 PM	566	2.70%	31.10%	58.80%	6.20%	411	415	409	60.00%
24	02M418	Millennium High S	Manhattan	M824	75 Broad Street	Manhattan	NY	10004	40.70492	-74.0115	212-825-9	8:30 AM	3:00 PM	659	32.80%	7.60%	18.20%	38.40%	577	560	567	94.00%
25	02M316	Urban Assembly S	Manhattan	M282	26 Broadway	Manhattan	NY	10004	40.70523	-74.0133	212-668-0	8:35 AM	3:30 PM	405	2.20%	52.80%	39.30%	3.50%	418	420	417	75.00%
26	02M489	High School of Ec	Manhattan	M833	100 Trinity Place	Manhattan	NY	10006	40.70922	-74.012	212-346-0	8:30 AM	3:30 PM	796	7.20%	22.90%	47.10%	21.70%	469	442	447	66.80%
27	02M425	Leadership and Pu	Manhattan	M894	90 Trinity Place	Manhattan	NY	10006	40.70915	-74.0121	212-346-0	8:30 AM	2:45 PM	699	4.40%	27.90%	61.90%	3.90%	390	396	392	54.90%
28	01M450	East Side Commu	Manhattan	M060	420 East 12th Str	Manhattan	NY	10009	40.72978	-73.983	212-460-8	8:30 AM	3:30 PM	666	8.30%	22.10%	57.90%	9.40%	454	428	445	85.20%
29	02M411	Baruch College Ca	Manhattan	M874	55 East 25th Stre	Manhattan	NY	10010	40.74405	-73.9915	212-683-7	8:20 AM	2:50 PM	451	22.80%	6.20%	14.90%	54.80%	592	526	531	94.30%

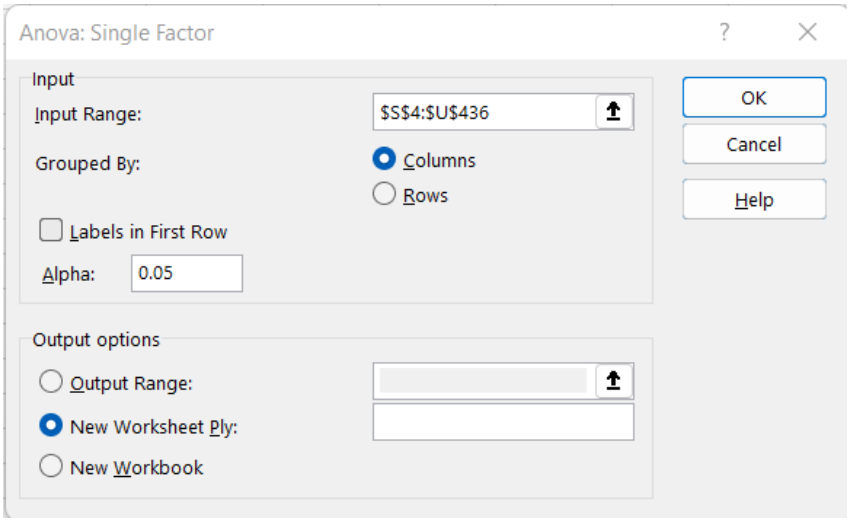
Step 2: To Calculate Single factor(one-way) anova we use cells S,T,U .
Data -> Data Analysis.



Step 3: Select Anova: Single Factor



Step 4: Select the cells of S-T-U for input range



Step 5: Output

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Column 1	375	162354	432.944	5177.144		
6	Column 2	375	159189	424.504	3829.267		
7	Column 3	375	156922	418.4587	4166.522		
8							
9							
10	ANOVA						
11	*ce of Varia*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between G	39700.57	2	19850.28	4.520698	0.01108	3.003745
13	Within Grc	4926677	1122	4390.977			
14							
15	Total	4966377	1124				
16							

Conclusion: we have successfully performed one-way ANOVA.

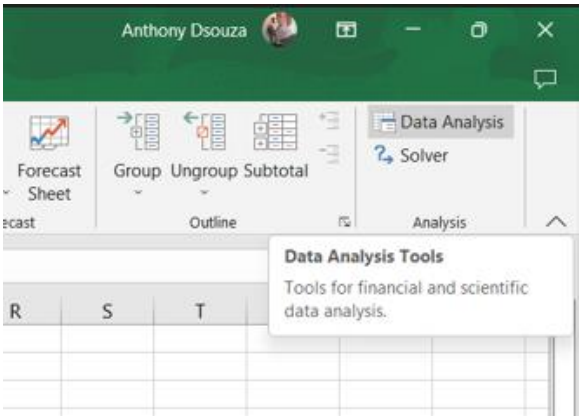
B.Perform testing of hypothesis using Two-way ANOVA.

Step 1: Load the data.

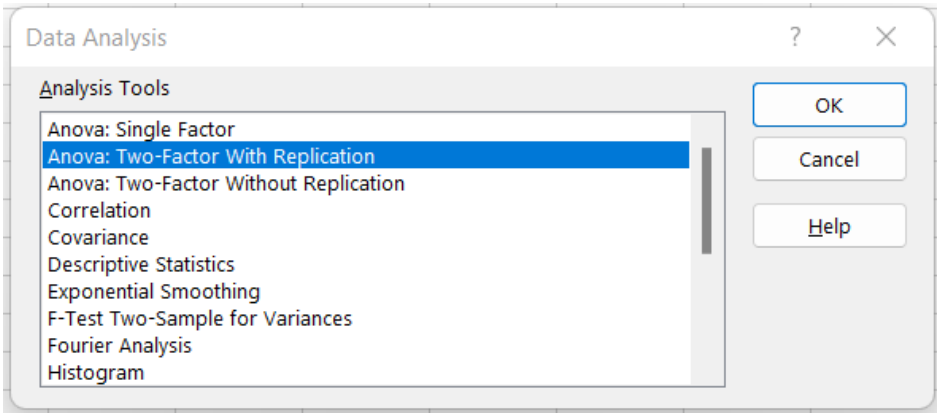
E18

	A	B	C	D	E
1		len	supp	dose	
2	1	4.2	VC	0.5	
3	2	11.5	VC	0.5	
4	3	7.3	VC	0.5	
5	4	5.8	VC	0.5	
6	5	6.4	VC	0.5	
7	6	10	VC	0.5	
8	7	11.2	VC	0.5	
9	8	11.2	VC	0.5	
10	9	5.2	VC	0.5	
11	10	7	VC	0.5	
12	11	16.5	VC	1	
13	12	16.5	VC	1	
14	13	15.2	VC	1	

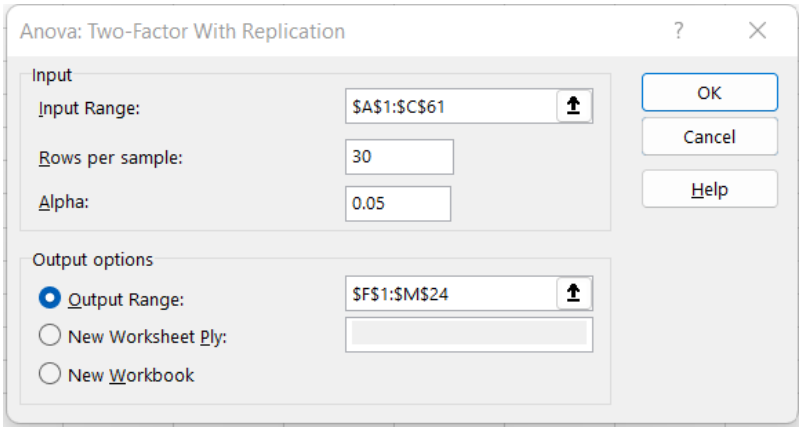
Step 2: Data -> Data Analysis



Step 3: Select Anova: Two Factor with Replication



Step 4: Select input and output range



Step 5: Output

Anova: Two-Factor With Replication							
SUMMAR	len	dose	Total				
1							
Count	30	30	60				
Sum	508.9	35	543.9				
Average	16.9633	1.16667	9.065				
Variance	68.3272	0.4023	97.2233				
31							
Count	30	30	60				
Sum	619.9	35	654.9				
Average	20.6633	1.16667	10.915				
Variance	43.6334	0.4023	118.285				
Total							
Count	60	60					
Sum	1128.8	70					
Average	18.8133	1.16667					
Variance	58.512	0.39548					
ANOVA							
Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
Sample	102.675	1	102.675	3.64208	0.05881	3.92288	
Columns	9342.15	1	9342.15	331.384	8.5E-36	3.92288	
Interactior	102.675	1	102.675	3.64208	0.05881	3.92288	
Within	3270.19	116	28.1913				
Total	12817.7	119					

Conclusion: we have successfully performed two way ANOVA.

C.Perform testing of hypothesis using MANOVA.

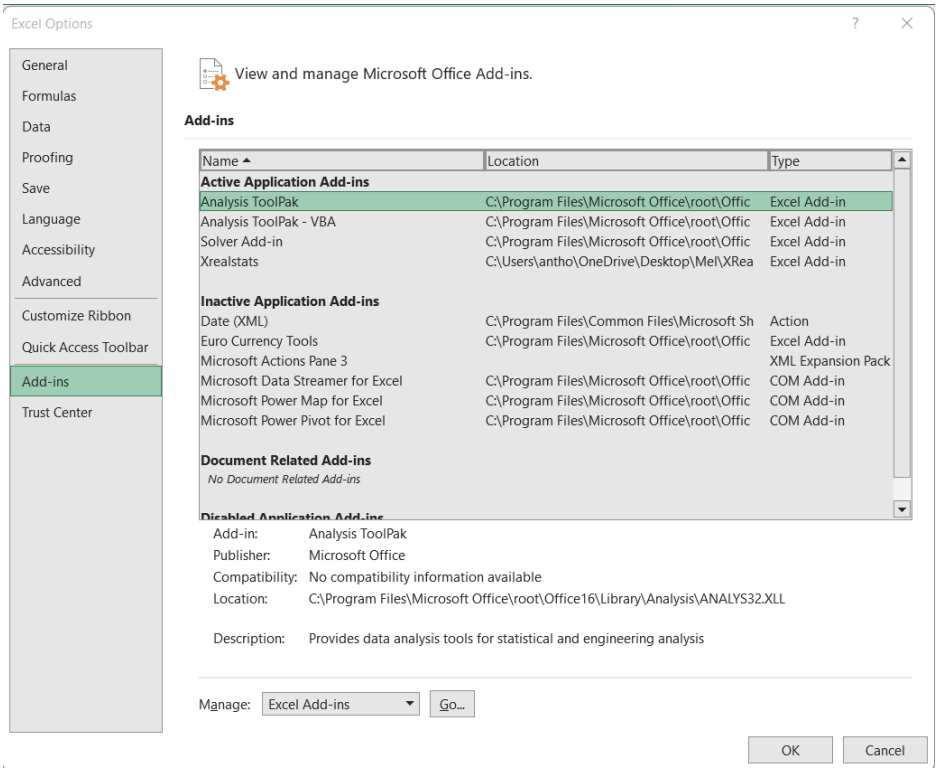
Step 1: Load the Data.

H13

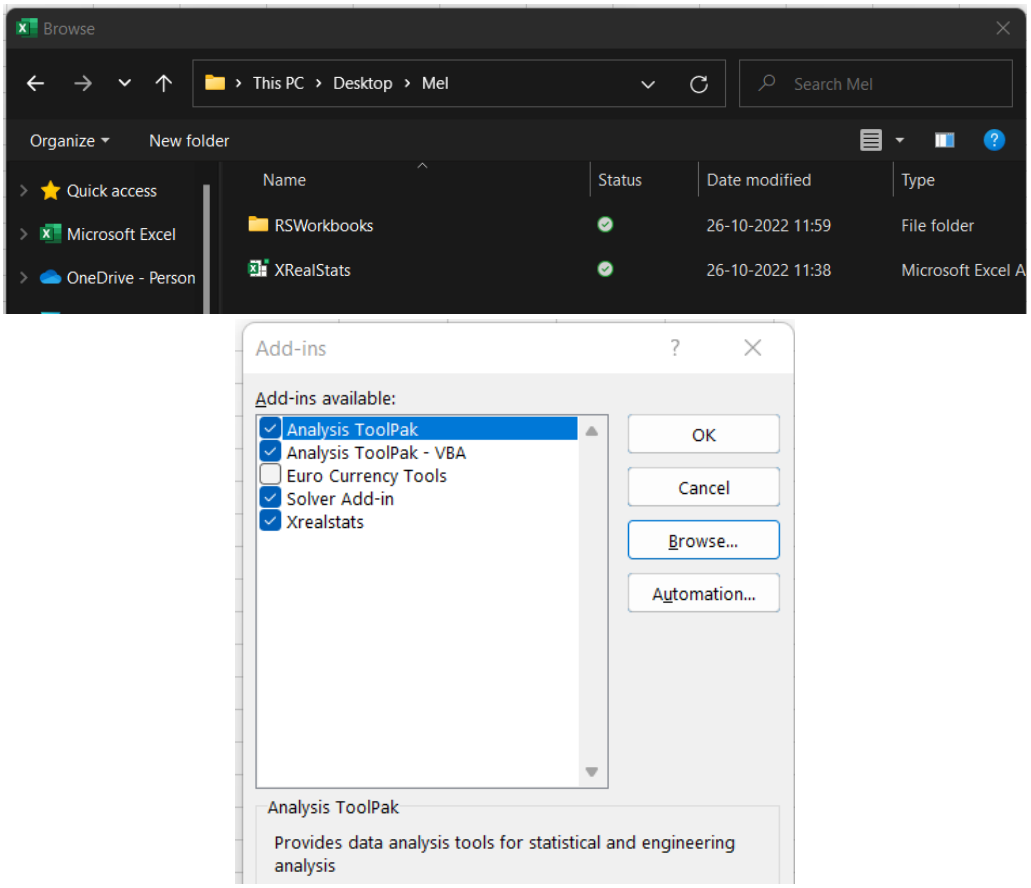
	A	B	C	D	E
1	Gender	Economics	Kindness	Optimism	
2	male	wealthy	5	3	
3	male	wealthy	4	6	
4	male	wealthy	3	4	
5	male	wealthy	2	4	
6	male	wealthy	4	6	
7	male	wealthy	3	6	
8	male	middle	5	4	
9	male	middle	5	5	
10	male	middle	7	5	

Step 2: Install Add-in in excel.

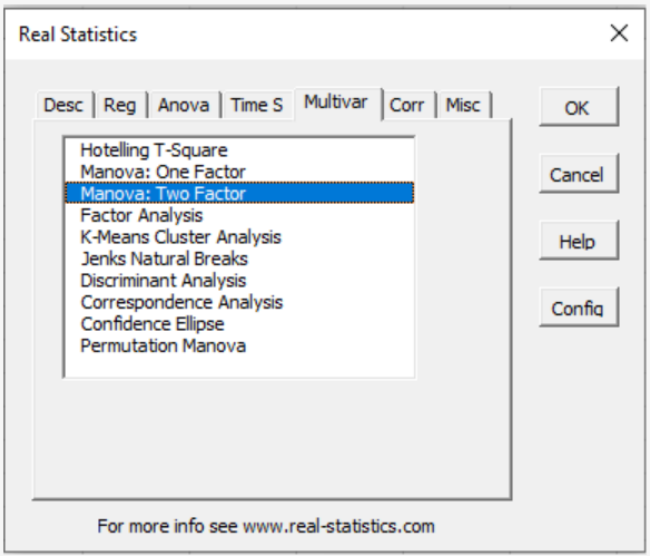
Select **File -> Help | Options -> Add-Ins** and click on the **Go** button at the bottom of the window.



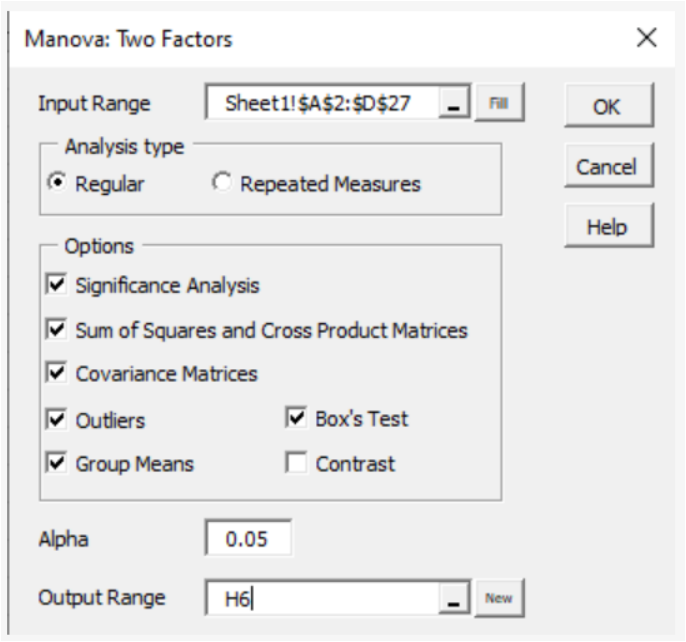
Step 3: Click on browse and select XrealStats file. Check the following check boxes.



Step 4: Once loaded , CTRL+M to perform MANOVA. Click Multivar -> Manova: two factor -> OK.



Step 5: Select the data excluding column names. Select a cell for output



Step 6: Output

Two-Way MANOVA							SSCP Matrices	
fact A	*stat*	*df1*	*df2*	*F*	*p-value*	*part eta-sq*	Tot	
Pillai Trac	0.214135	2	18	2.452347	0.114323	0.214135	130	64
Wilk's Lam	0.785865	2	18	2.452347	0.114323	0.214135	64	142
Hotelling	0.272483	2	18	2.452347	0.114323	0.214135		
Roy's Lg R	0.272483						Row (A)	
							27.08333	17.5
fact B	*stat*	*df1*	*df2*	*F*	*p-value*	*part eta-sq*	17.5	11.30769
Pillai Trac	0.339044	4	38	1.939195	0.123777	0.169522		
Wilk's Lam	0.820428	4	36	1.969888	0.120034	0.179572	Column (B)	
Hotelling	0.467615	4	34	1.987365	0.118648	0.189501	31.90278	28.18056
Roy's Lg R	0.425176						28.18056	30.36111
fact AB	*stat*	*df1*	*df2*	*F*	*p-value*	*part eta-sq*	Interaction (AB)	
Pillai Trac	0.249126	4	38	1.35172	0.268831	0.124563	5.963889	0.019444
Wilk's Lam	0.871554	4	36	1.326381	0.278864	0.128446	0.019444	21.0312
Hotelling	0.304976	4	34	1.296146	0.291053	0.132312		
Roy's Lg R	0.260917						Res	
							65.05	18.3
							18.3	79.3

Conclusion: we have successfully performed MANOVA.

Practical 7

A.Perform the Random sampling for the given data and analyse it.

Step 1: Load the data.

I16 | sfe

	A	B	C	D	E	F	G	H	I	J	K
1	Sr. No.	Roll No.	Name	Gender	Grade		Sr. No.	Roll No.	Name	Gender	Grade
2	1	1	xyz	Male	O		24	3	add	Female	A
3	2	2	mno	Male	A		25	7	sd	Female	A
4	3	5	pqr	Male	O		26	9	hg	Female	C
5	4	13	abc	Male	B		27	11	dg	Female	D
6	5	16	xyz	Male	O		28	14	pgu	Female	F
7	6	17	mno	Male	O		29	25	yt	Female	O
8	7	34	pqr	Male	D		30	36	f	Female	C
9	8	35	abc	Male	O		31	40	gh	Female	C
10	9	38	xyz	Male	O		32	41	t	Female	B
11	10	42	mno	Male	C		33	46	es	Female	A
12	11	43	pqr	Male	O		34	47	gg	Female	B
13	12	45	abc	Male	O		35	10	gfh	Female	D
14	13	48	xyz	Male	B		36	20	ghy	Female	C
15	14	49	mno	Male	O		37	21	tgs	Female	B
16	15	50	pqr	Male	D		38	72	sfe	Female	O
17	16	51	abc	Male	O		39	73	sgg	Female	D
18	17	54	xyz	Male	A		40	75	dgt	Female	A
19	18	55	mno	Male	B		41	77	dgt	Female	B
20	19	56	pqr	Male	D		42	82	jgt	Female	D
21	20	59	abc	Male	O		43	84	ghh	Female	D
22	21	61	xyz	Male	F		44	91	fer	Female	D
23	22	62	mno	Male	A		45	95	gyd	Female	D
24	23	63	pqr	Male	C		46	4	gtt	Female	C

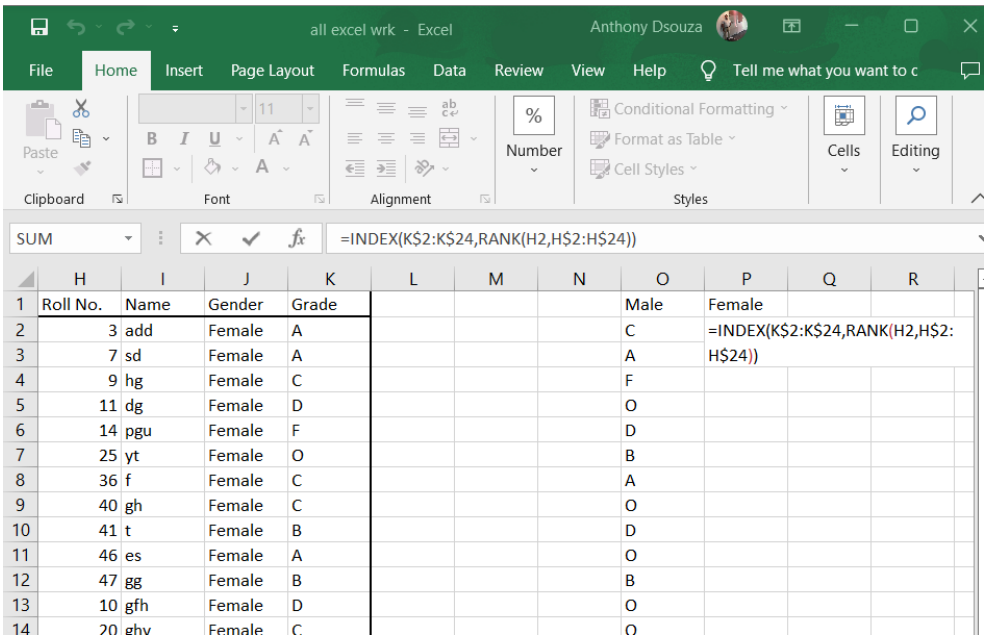
Step 2: Set Cell O1= Male and Cell P2= Female

To generate a random sample for male students from given population go to Cell O1 and type =INDEX(E$2:E$62,RANK(B2,B$2:B$62))

SUM | =INDEX(E$2:E$62,RANK(B2,B$2:B$62))

	H	I	J	K	L	M	N	O	P	Q	R
1	Roll No.	Name	Gender	Grade				=INDEX(E$2:E$62,RANK(B2,B$2:B$62))			
2	3	add	Female	A							
3	7	sd	Female	A							
4	9	hg	Female	C							
5	11	dg	Female	D							
6	14	pgu	Female	F							
7	25	yt	Female	O							
8	36	f	Female	C							
9	40	gh	Female	C							
10	41	t	Female	B							
11	46	es	Female	A							
12	47	gg	Female	B							
13	10	gfh	Female	D							

Step 3: Repeat the step to perform random sampling on Female data.



	H	I	J	K	L	M	N	O	P	Q	R
1	Roll No.	Name	Gender	Grade				Male	Female		
2	3	add	Female	A				C	=INDEX(K$2:K$24,RANK(H2,H$2:		
3	7	sd	Female	A				A	H$24))		
4	9	hg	Female	C				F			
5	11	dg	Female	D				O			
6	14	pgu	Female	F				D			
7	25	yt	Female	O				B			
8	36	f	Female	C				A			
9	40	gh	Female	C				O			
10	41	t	Female	B				D			
11	46	es	Female	A				O			
12	47	gg	Female	B				B			
13	10	gfh	Female	D				O			
14	20	ghy	Female	C				O			

Step 4: Output

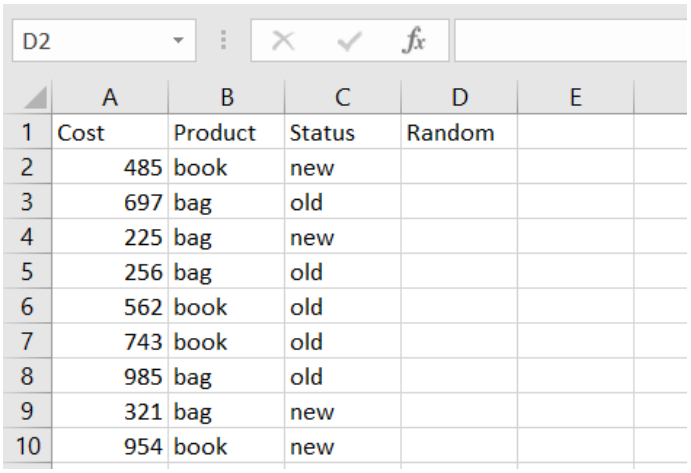
O1 Male

	F	G	H	I	J	K	L	M	N	O	P
1		Sr. No.	Roll No.	Name	Gender	Grade				Male	Female
2		24	3	add	Female	A				C	C
3		25	7	sd	Female	A				A	D
4		26	9	hg	Female	C				F	D
5		27	11	dg	Female	D				O	B
6		28	14	pgu	Female	F				D	A
7		29	25	yt	Female	O				B	B
8		30	36	f	Female	C				A	C
9		31	40	gh	Female	C				O	D
10		32	41	t	Female	B				D	B
11		33	46	es	Female	A				O	A
12		34	47	gg	Female	B				B	B
13		35	10	gfh	Female	D				O	D
14		36	20	ghy	Female	C				O	D
15		37	21	tgs	Female	B				C	O
16		38	72	sfe	Female	O				O	C
17		39	73	sgg	Female	D				O	C
18		40	75	dgt	Female	A				D	O
19		41	77	dgt	Female	B				O	F
20		42	82	jgt	Female	D				O	D
21		43	84	ghh	Female	D				B	C
22		44	91	fer	Female	D				O	A
23		45	95	gyd	Female	D				A	A
24		46	4	gtt	Female	C				O	D
25											

Conclusion: we have successfully performed Random sampling and analyze it.

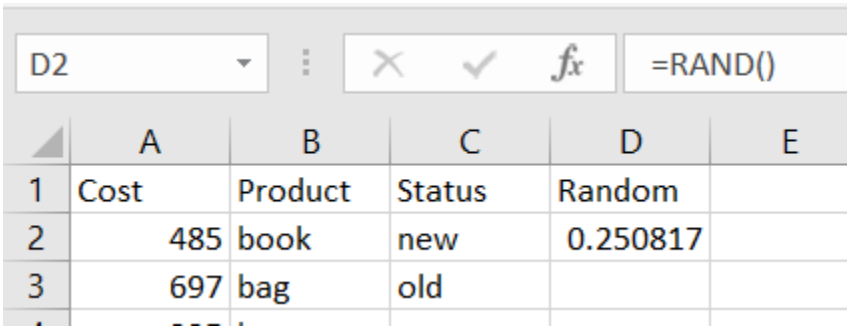
B.Perform the Stratified sampling for the given data and analyse it

Step 1: Load the data.



	A	B	C	D	E
1	Cost	Product	Status	Random	
2	485	book	new		
3	697	bag	old		
4	225	bag	new		
5	256	bag	old		
6	562	book	old		
7	743	book	old		
8	985	bag	old		
9	321	bag	new		
10	954	book	new		

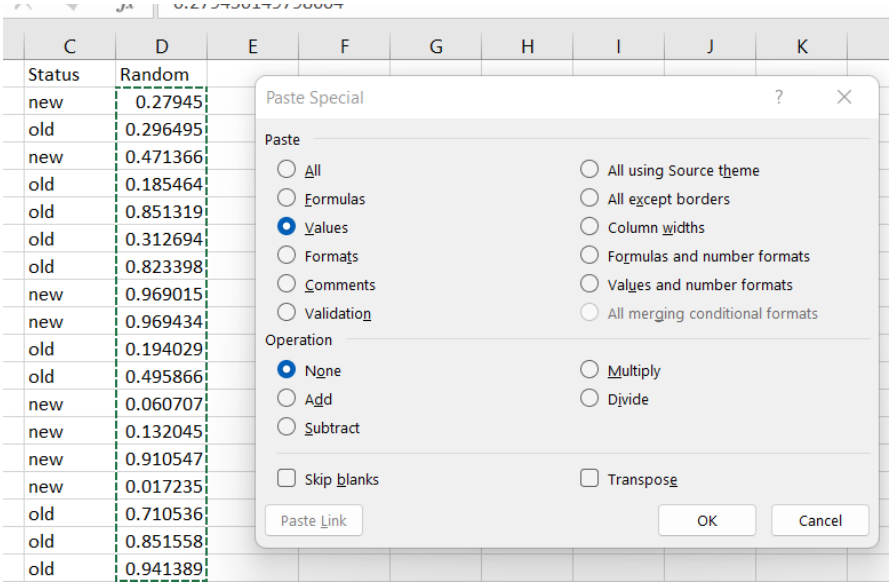
Step 2: Assign Random values using RAND() function.



D2 =RAND()

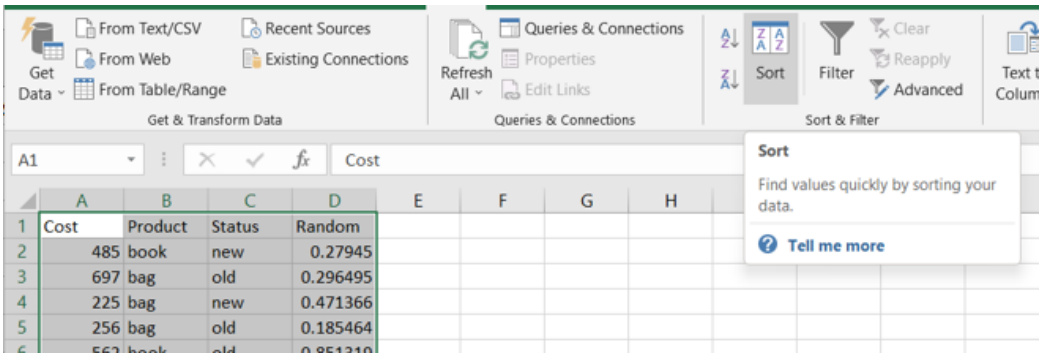
	A	B	C	D	E
1	Cost	Product	Status	Random	
2	485	book	new	0.250817	
3	697	bag	old		

Step 3: Copy the entire column D and paste only values

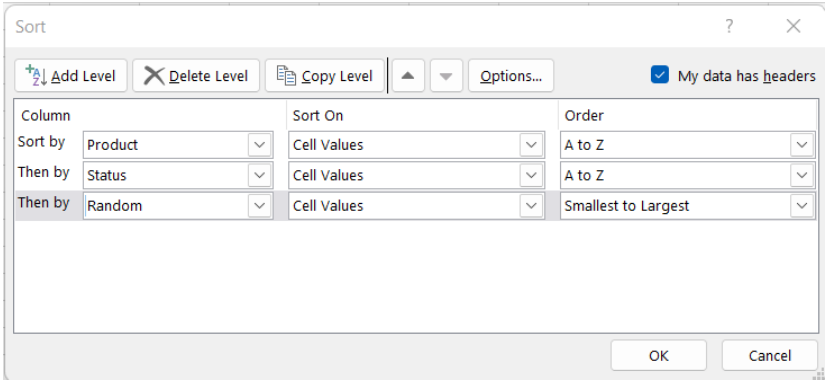


C	D
Status	Random
new	0.27945
old	0.296495
new	0.471366
old	0.185464
old	0.851319
old	0.312694
old	0.823398
new	0.969015
new	0.969434
old	0.194029
old	0.495866
new	0.060707
new	0.132045
new	0.910547
new	0.017235
old	0.710536
old	0.851558
old	0.941389

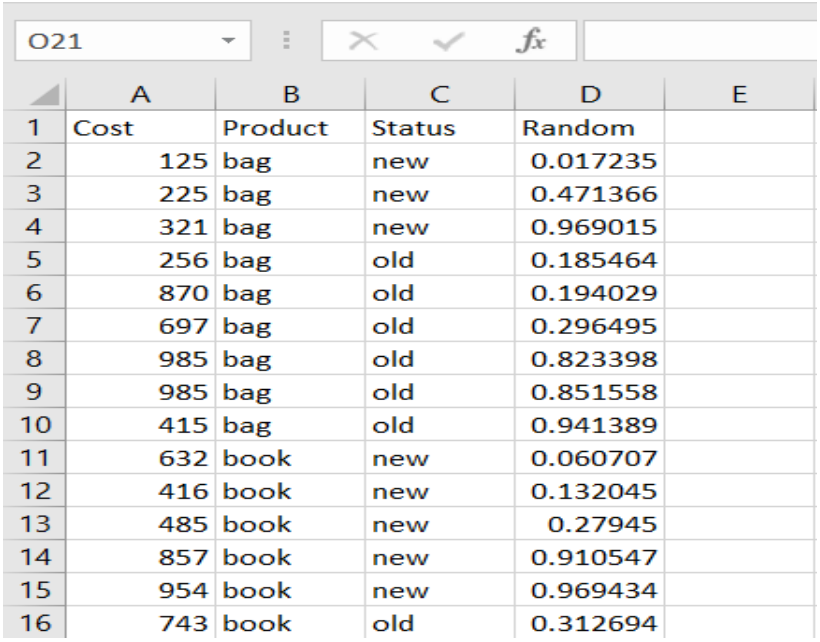
Step 4: To perform stratified sampling and obtain data where we get 2 new books cost, 2 old books cost, 2 new bags cost, 2 old bags cost. Select the data and click on sort.



Step 5: Sort the data by product , status and random.



Step 6: the output obtained will be as follows.



O21

	A	B	C	D	E
1	Cost	Product	Status	Random	
2	125	bag	new	0.017235	
3	225	bag	new	0.471366	
4	321	bag	new	0.969015	
5	256	bag	old	0.185464	
6	870	bag	old	0.194029	
7	697	bag	old	0.296495	
8	985	bag	old	0.823398	
9	985	bag	old	0.851558	
10	415	bag	old	0.941389	
11	632	book	new	0.060707	
12	416	book	new	0.132045	
13	485	book	new	0.27945	
14	857	book	new	0.910547	
15	954	book	new	0.969434	
16	743	book	old	0.312694	

Conclusion: we have successfully performed stratified sampling and analyse it.

Practical 8

Write a program for computing different correlation

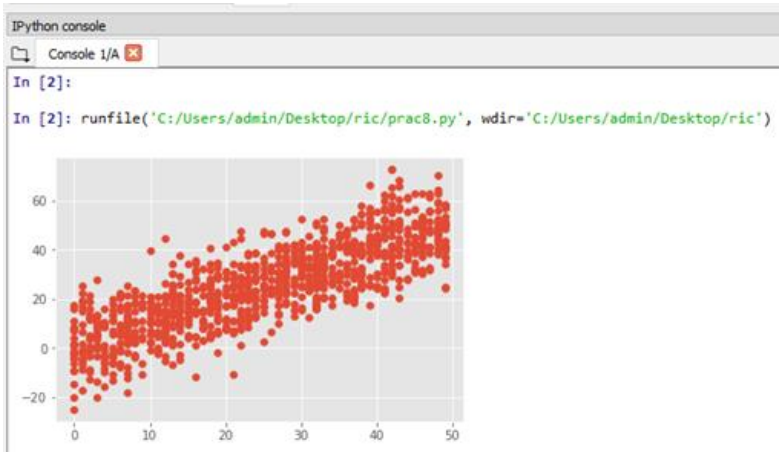
A.Positive Correlation.

Code:

```
import matplotlib

import numpy as np

import matplotlib.pyplot as plt

np.random.seed(1)

# 1000 random integers between 0 and 50

x = np.random.randint(0, 50, 1000)

# Positive Correlation with some noise
```

```
y = x + np.random.normal(0, 10, 1000)

np.corrcoef(x, y)

matplotlib.style.use('ggplot')

plt.scatter(x, y)

plt.show()
```

Output:



Conclusion: we have successfully computed positive correlation.

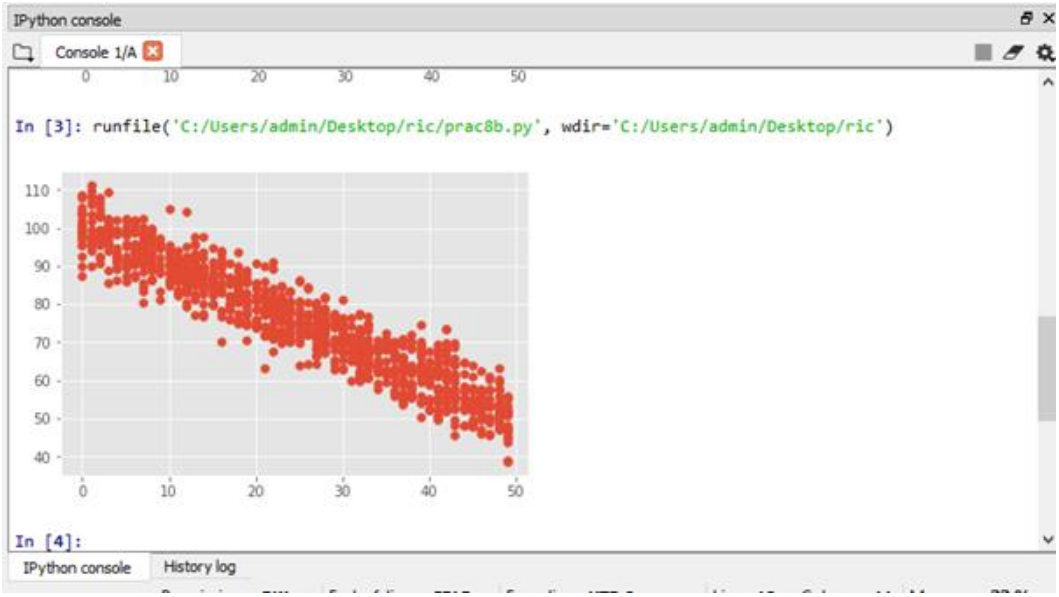
B.Negative Correlation.

Code:

```
import matplotlib

import numpy as np

import matplotlib.pyplot as plt

np.random.seed(1)

# 1000 random integers between 0 and 50

x = np.random.randint(0, 50, 1000)

# Negative Correlation with some noise

y = 100 - x + np.random.normal(0, 5, 1000)
```

```
np.corrcoef(x, y)

plt.scatter(x, y)

plt.show()
```

Output:



Conclusion: we have successfully computed negative correlation.

C.No/Weak Correlation.

Code:

```
import numpy as np

import matplotlib.pyplot as plt

np.random.seed(1)

x = np.random.randint(0, 50, 1000)

y = np.random.randint(0, 50, 1000)

np.corrcoef(x, y)

plt.scatter(x, y)

plt.show()
```

Output:

Conclusion: we have successfully computed no/weak correlation.

Practical 9

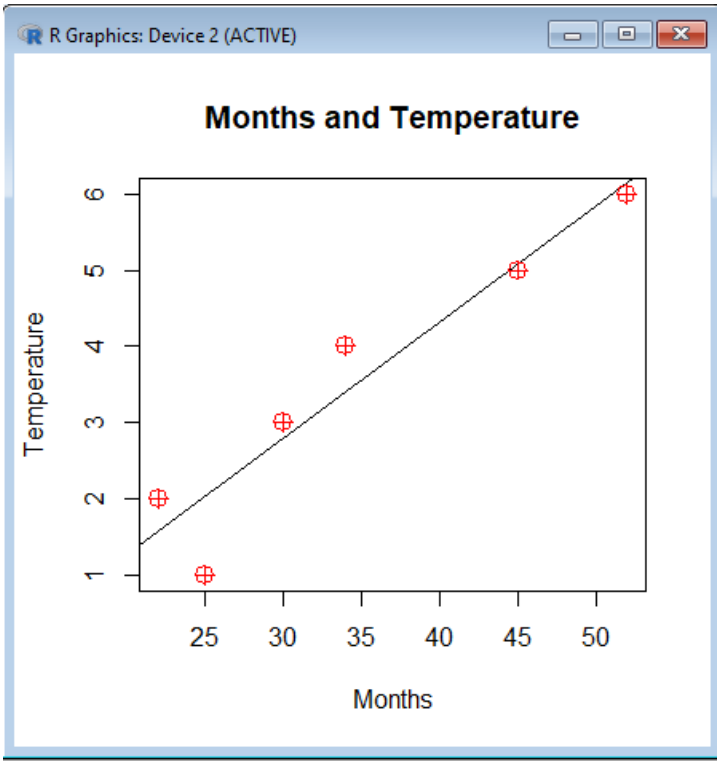
A.Write a program to Perform linear regression for prediction.

Code:

```
> #Perform linear regression

> m<-c(1,2,3,4,5,6)

> t<-c(25,22,30,34,45,52)

> #Label the chart

> png(file="Linear Regression")

> plot(t,m,col="red",main="Months and
Temperature",abline(lm(m~t)),cex=1.6,pch=10,xlab="Months",ylab="
Temperature")
```

```
> dev.off()

null device

          1

> plot(t,m,col="red",main="Months and
Temperature",abline(lm(m~t)),cex=1.6,pch=10,xlab="Months",ylab="
Temperature")
```

```
> #Perform linear regression
> m<-c(1,2,3,4,5,6)
> t<-c(25,22,30,34,45,52)
> #Label the chart
> png(file="Linear Regression")
> plot(t,m,col="red",main="Months and Temperature",abline(lm(m~t)),cex=1.6,pch=$
> dev.off()
null device
          1
> plot(t,m,col="red",main="Months and Temperature",abline(lm(m~t)),cex=1.6,pch=$
>
```



Conclusion: we have successfully performed linear regression for predicted.

B.Polynomial Regression.

Code:

```
> #Polynomial Regression

> set.seed(16)

> x<-0:50

> y<-2.3-15.1*x+1.2*x^2+rnorm(length(x),20,50)

> plot(x,y)

> fit <- lm(y ~ 1 + x + I(x^2))

> points(x, predict(fit), type="l")

> summary(fit)
```

Output:

```
Call:
lm(formula = y ~ 1 + x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-92.173 -28.968   3.673  24.953  97.269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.84216   18.36178   1.843   0.0715 .
x           -15.28705    1.69836  -9.001 7.07e-12 ***
I(x^2)        1.20126    0.03285  36.569  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.44 on 48 degrees of freedom
Multiple R-squared:  0.996,     Adjusted R-squared:  0.9959
F-statistic:  6034 on 2 and 48 DF,  p-value: < 2.2e-16

>
```

Conclusion: we have successfully performed polynomial regression in R.

Practical 10

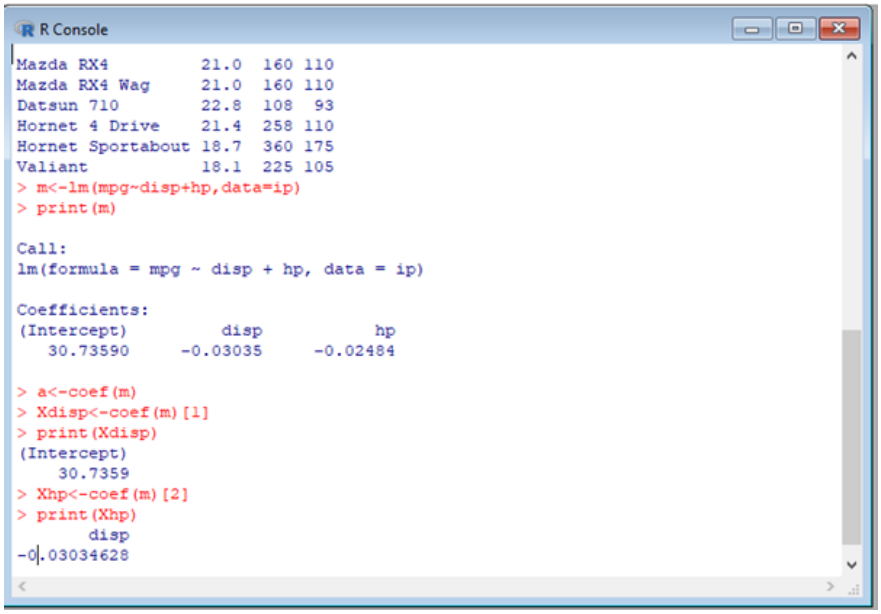
A.Multiple linear regression.

Code:

```
> #Multiple linear regression

>ip<-mtcars[,c("mpg","disp","hp")]

> print(head(ip))

> m<-lm(mpg~disp+hp,data=ip)

> print(m)

> a<-coef(m)

>Xdisp<-coef(m)[1]

> print(Xdisp)

>Xhp<-coef(m)[2]

> print(Xhp)
```

Output:

```
> print(head(ip))
                  am    wt  mpg
Mazda RX4          1 2.620 21.0
Mazda RX4 Wag      1 2.875 21.0
Datsun 710         1 2.320 22.8
Hornet 4 Drive     0 3.215 21.4
Hornet Sportabout  0 3.440 18.7
Valiant            0 3.460 18.1
```



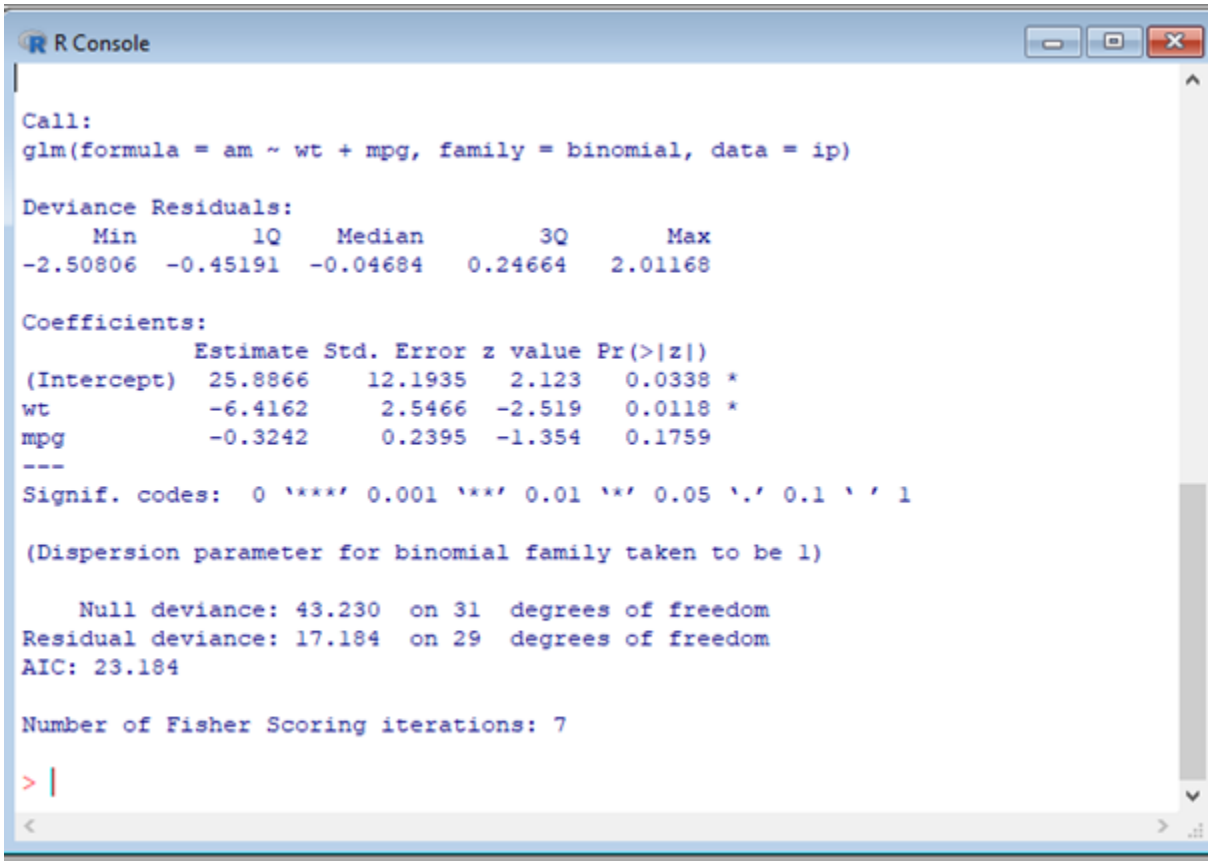
Conclusion: we have successfully performed multiple linear regression.

B.Logistic Regression.

Code:

```
> #Logistic regression

>ip<-mtcars[,c("am","wt","mpg")]

> print(head(ip))

>am.data<-glm(formula=am~wt+mpg,data=ip,family=binomial)

> summary(am.data)
```

Output:



Conclusion: we have successfully performed logistic regression.