Table of Contents

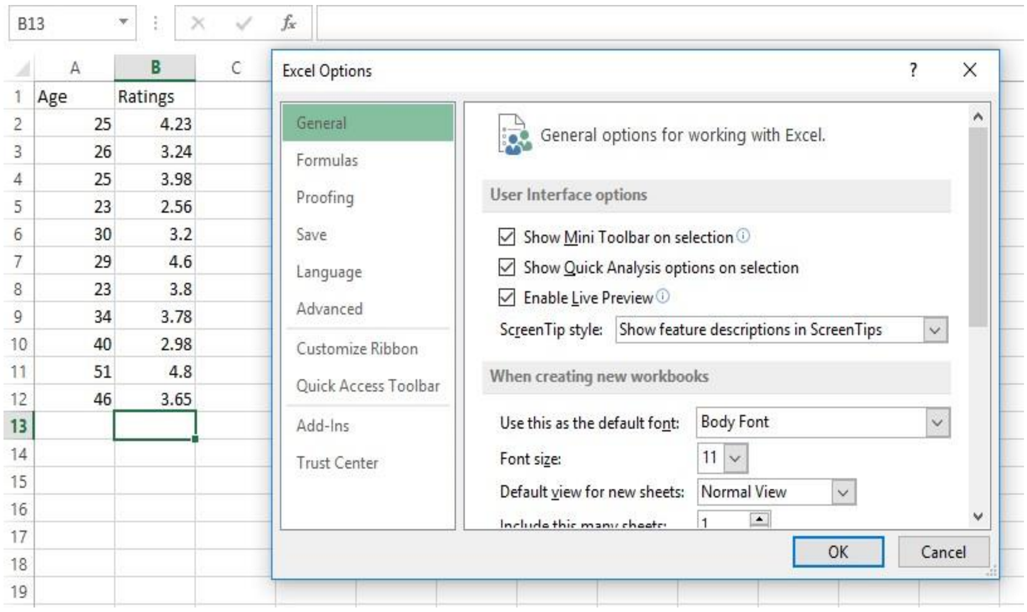
Sr. No	Practical No		Name of the Practical	Signature
1)	**1**	**A**	Write a program for obtaining descriptive statistics of data.	
2)		**B**	Import data from different data sources (from Excel, csv, mysql, sql server, oracle to R/Python/Excel)	
3)	**2**	**A**	Design a survey form for a given case study, collect the primary data and analyze it	
4)		**B**	Perform suitable analysis of given secondary data.	
5)	**3**	**A**	Perform testing of hypothesis using one sample t-test.	
6)		**B**	Perform testing of hypothesis using two sample t-test.	
7)		**C**	Perform testing of hypothesis using paired t-test.	
8)	**4**	**A**	Perform testing of hypothesis using chi-squared goodness-of-fit test.	
9)		**B**	Perform testing of hypothesis using chi-squared Test of Independence	
10)	**5**		Perform testing of hypothesis using Z-test.	
11)	**6**	**A**	Perform testing of hypothesis using one-way ANOVA.	
12)		**B**	Perform testing of hypothesis using two-way ANOVA.	
13)		**C**	Perform testing of hypothesis using multivariate ANOVA (MANOVA).	
14)	**7**	**A**	Perform the Random sampling for the given data and analyse it.	

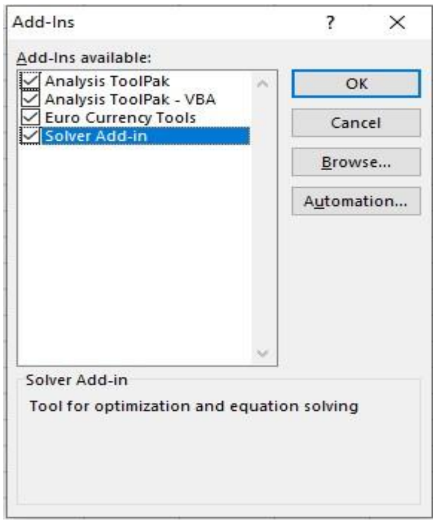
15)		**B**	Perform the Stratified sampling for the given data and analyse it.	
16)	**8**		Compute different types of correlation.	
17)	**9**	**A**	Perform linear regression for prediction.	
18)		**B**	Perform polynomial regression for prediction.	
19)	**10**	**A**	Perform multiple linear regression.	
20)		**B**	Perform Logistic regression.	

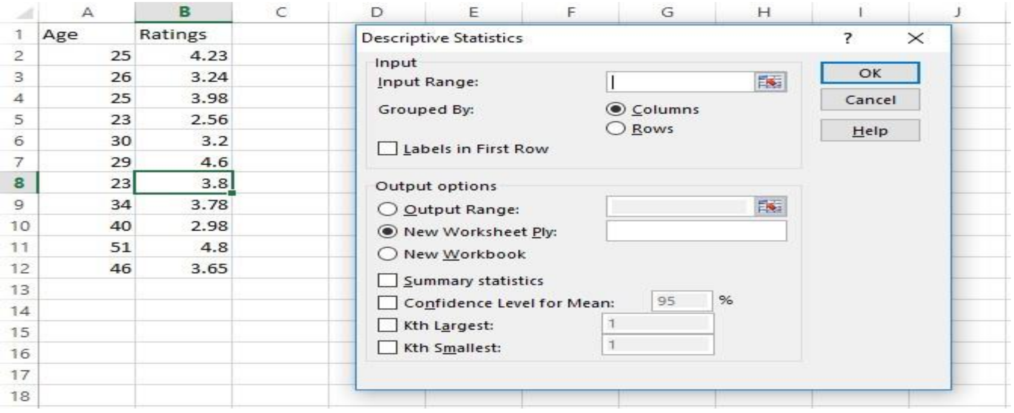
<u>**PRACTICAL NO: 1**</u>

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

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include the standard deviation, variance, the minimum and maximum variables, and the kurtosis and skewness.







	A	B	C	D	E	F
1	Age	Ratings				
2	25	4.23			*Column1*	
3	26	3.24				
4	25	3.98			Mean	32
5	23	2.56			Standard E	2.913916
6	30	3.2			Median	29
7	29	4.6			Mode	25
8	23	3.8			Standard D	9.664368
9	34	3.78			Sample Va	93.4
10	40	2.98			Kurtosis	-0.12813
11	51	4.8			Skewness	1.046398
12	46	3.65			Range	28
13					Minimum	23
14					Maximum	51
15					Sum	352
16					Count	11
17					Largest(1)	51
18					Smallest(1	23
19					Confidenc	6.49261
20						

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

<u>Code:</u>

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'] =

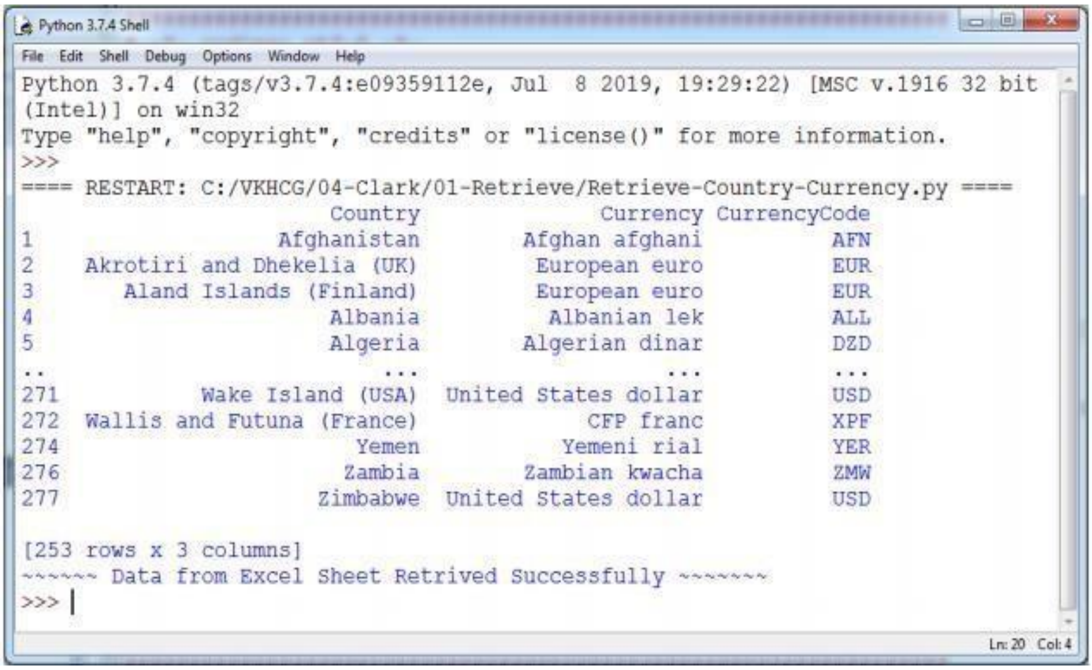
CurrencyData['Currency'].map(lambda x: x.strip()) CurrencyData['CurrencyCode'] =

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)

<u>Output:</u>



```
Python 3.7.4 Shell
File Edit Shell Debug Options Window Help
Python 3.7.4 (tags/v3.7.4:e09359112e, Jul  8 2019, 19:29:22) [MSC v.1916 32 bit
(Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
==== RESTART: C:/VKHCG/04-Clark/01-Retrieve/Retrieve-Country-Currency.py ====
                          Country              Currency CurrencyCode
1                     Afghanistan        Afghan afghani          AFN
2    Akrotiri and Dhekelia (UK)         European euro          EUR
3       Aland Islands (Finland)         European euro          EUR
4                         Albania          Albanian lek          ALL
5                         Algeria        Algerian dinar          DZD
..                            ...                   ...          ...
271           Wake Island (USA)  United States dollar          USD
272  Wallis and Futuna (France)              CFP franc          XPF
274                         Yemen           Yemeni rial          YER
276                        Zambia        Zambian kwacha          ZMW
277                      Zimbabwe  United States dollar          USD

[253 rows x 3 columns]
~~~~~~ Data from Excel Sheet Retrived Successfully ~~~~~~~
>>> |
Ln: 20  Col: 4
```

PRACTICAL NO. 2

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

A survey on “Social Media” using google form then send it to everyone to know about their views or perspectives regarding social media. After that the data have been collected and then analysis is done on it by using excel.

Social Media

Social media have its own advantages and disadvantages. This questionnaire section just shows us how often are people connected/addicted to the Social media.

* Required

1. Gender *

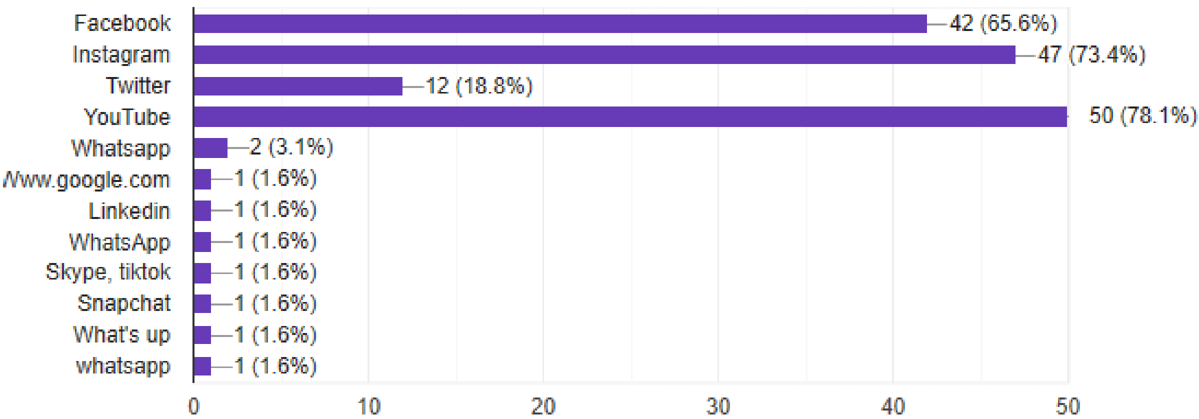
- Male
- Female

2. What age category do you belong? *

- 13 years or younnger
- 16-20
- 21-26
- 27 or above

	A	B	C	D	E	F	G	H
1	Timestamp	Score	1. Gender	2. What age category do	3. What social networking	4. How many hours a day	5. Why do you use these	6. Do you think Privacy P
2	12/19/2019 11:43:16		Female	21-26	Instagram	1-2 hours	To be updated with recen	Agree
3	12/19/2019 11:55:34		Male	21-26	Facebook, Instagram, Yo	4 or more	To be updated with recen	Strongly agree
4	12/19/2019 11:57:57		Male	21-26	Facebook, YouTube	4 or more	To be updated with recen	Neutral
5	12/19/2019 12:01:39		Male	21-26	Facebook, Instagram, Yo	4 or more	To be updated with recen	Neutral
6	12/19/2019 12:06:46		Male	21-26	Instagram	4 or more	Contact and connect fami	Agree
7	12/19/2019 12:14:18		Male	21-26	Facebook, Instagram, Yo	4 or more	To be updated with recen	Agree
8	12/19/2019 12:15:48		Male	21-26	Facebook, Instagram, Yo	3-4	To be updated with recen	Agree
9	12/19/2019 12:29:17		Male	21-26	YouTube	3-4	To be updated with recen	Neutral
10	12/19/2019 12:32:48		Male	21-26	Facebook, Instagram, Yo	1-2 hours	To be updated with recen	Agree
11	12/19/2019 12:36:24		Male	21-26	Facebook, Instagram, Yo	1-2 hours	To be updated with recen	Agree
12	12/19/2019 12:36:29		Male	16-20	Facebook	1-2 hours	To be updated with recen	Disagree
13	12/19/2019 12:39:58		Male	21-26	Instagram	3-4	To be updated with recen	Neutral
14	12/19/2019 12:39:59		Female	16-20	Facebook, Instagram, Yo	3-4	To be updated with recen	Agree
15	12/19/2019 12:40:25		Male	21-26	Instagram	3-4	To be updated with recen	Neutral
16	12/19/2019 12:48:29		Female	21-26	Facebook, Instagram, Yo	1-2 hours	To be updated with recen	Agree
17	12/19/2019 12:55:32		Female	21-26	Instagram	1-2 hours	Sharing / Liking Posts	Strongly agree
18	12/19/2019 12:57:33		Male	16-20	Facebook, Instagram	3-4	To be updated with recen	Agree
19	12/19/2019 13:00:28		Male	21-26	Instagram	1-2 hours	To be updated with recen	Neutral
20	12/19/2019 13:01:16		Female	27 or above	Facebook, YouTube	1-2 hours	To be updated with recen	Neutral
21	12/19/2019 13:04:59		Female	27 or above	Facebook, Instagram, Yo	1-2 hours	To be updated with recen	Neutral

Perform analysis of given secondary data.



PRACTICAL NO. 3

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

One sample t-test: The One Sample t Test determines whether the sample mean is statistically different from a known or hypothesized population mean. The One Sample t Test is a parametric test.

Code:

```
From scipy.stats import ttest_1samp

import numpy as np

ages = np.genfromtxt('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:

        print(" we are rejecting null hypothesis")

else:

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

Output:

```
In [4]: runfile('K:/Research In Computing/Practical Material/Programs/
Practical_05/Prac_3A.py', wdir='K:/Research In Computing/Practical Material/
Programs/Practical_05')
[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
```

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

	A	B
1	Men	Women
2	181	160
3	169	150
4	160	160
5	170	175
6	175	160
7	158	170
8	152	160
9	172	150
10	160	155
11	175	162
12	180	165
13	170	148
14	165	159
15	180	163
16	155	170
17	159	178
18	163	180
19	171	156

H0 – Height of men and women are same

H1 – Height of men and women are the different

(-tcritical two-tail<tsat<tcritical two tail) then accept

F-Test Two-Sample for Variances

	Variable 1	Variable 2
Mean	167.5	162.2777778
Variance	79.5588	87.03594771
Observations	18	18
df	17	17
F	0.91409	
P(F<=f) one-tail	0.42762	
F Critical one-tail	0.44016	
	reject equal variance hypothesis	

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	Variable 2
Mean	167.5	162.278
Variance	79.5588	87.0359
Observations	18	18
Hypothesized Mean Difference	0	
df	34	
t Stat	1.71657	
P(T<=t) one-tail	0.04758	
t Critical one-tail	1.69092	
P(T<=t) two-tail	0.09516	
t Critical two-tail	2.03224	

C) Perform testing of hypothesis using paired t-test

	A	B	C	D	E	F
1	patient	gender	agegrp	bp_before	bp_after	diffrerence
2	1	Male	30-45	143	153	-10
3	2	Male	30-45	163	170	-7
4	3	Male	30-45	153	168	-15
5	4	Male	30-45	153	142	11
6	5	Male	30-45	146	141	5
7	6	Male	30-45	150	147	3
8	7	Male	30-45	148	133	15
9	8	Male	30-45	153	141	12
10	9	Male	30-45	153	131	22
11	10	Male	30-45	158	125	33
12	11	Male	30-45	149	164	-15
13	12	Male	30-45	173	159	14
14	13	Male	30-45	165	135	30
15	14	Male	30-45	145	159	-14
16	15	Male	30-45	143	153	-10
17	16	Male	30-45	152	126	26
18	17	Male	30-45	141	162	-21
19	18	Male	30-45	176	134	42
20	19	Male	30-45	143	136	7
21	20	Male	30-45	162	150	12
22	21	Male	46-59	149	168	-19
23	22	Male	46-59	156	155	1
24	23	Male	46-59	151	136	15
25	24	Male	46-59	159	132	27
26	25	Male	46-59	164	160	4
27	26	Male	46-59	154	160	-6

H0 - The mean difference between sample 1 and sample 2 is equal to 0.

H1- The mean difference between sample 1 and sample 2 is not equal to 0

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	156.45	151.3583333
Variance	129.7285714	201.004972
Observations	120	120
Pearson Correlation	0.159118103	
Hypothesized Mean [	0	
df	119	
t Stat	3.337187051	
P(T<=t) one-tail	0.000564896	
t Critical one-tail	1.657759285	
P(T<=t) two-tail	0.001129791	
t Critical two-tail	1.980099876	

Column1	
Mean	156.45
Standard Error	1.0397
Median	154.5
Mode	162
Standard Deviation	11.39
Sample Variance	129.73
Kurtosis	-0.439
Skewness	0.5542
Range	47
Minimum	138
Maximum	185
Sum	18774
Count	120
Confidence Level(95.	2.0588

PRACTICAL NO. 4

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

System	O	Ei	$\sum \frac{(O_i - E_i)^2}{Ei}$
Windows	20	33.33%	
Mac	60	33.33%	
Linux	20	33.33%	

Output:

	A	B	C	D	E	F	G	H	I
1									
2		H0 : The population distribution of the variable is the same as the proposed distribution							
3		HA : The distributions are different							
4									
5									
6		Type	O	Ei	Calculated				
7		windows	20	33.33	5.33120012				
8		Mac	60	33.33	21.34080108				
9		Linux	20	33.33	5.33120012				
10		Total	100	100.00	32.00320132				
11									
12				Table value	5.991				
13									
14				H0 accepted					

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

	A	B	C	D
1	User ID	Age	Gender	Grade
2	15624510	19	Male	O
3	15810944	35	Male	O
4	15668575	26	Female	O
5	15603246	27	Female	O
6	15804002	19	Male	O
7	15728773	27	Male	O
8	15598044	27	Female	O
9	15694829	32	Female	O
10	15600575	25	Male	O
11	15727311	35	Female	O
12	15570769	26	Female	O
13	15606274	26	Female	O
14	15746139	20	Male	O
15	15704987	32	Male	O
16	15628972	18	Male	O
17	15697686	29	Male	O
18	15733883	47	Male	O
19	15617482	45	Male	O
20	15704583	46	Male	O
21	15621083	48	Female	O
22	15649487	45	Male	O
23	15736760	47	Female	o
24	15714658	48	Male	D

H0 : The performance of girls students is same as boys students.

H1 : The performance of boys and girls students are different.

F	G	H	I	J	K	L
	O	A	B	C	D	Total
Girls	16	8	4	9	12	49
Boys	18	9	5	3	15	50
Total	34	17	9	12	27	99
	O	A	B	C	D	Total
Girls	16.82828	8.414141	4.454545	5.939394	13.36364	49
Boys	17.17172	8.585859	4.545455	6.060606	13.63636	50
Total	34	17	9	12	27	99
	O	A	B	C	D	
Girls	0.040768	0.020384	0.046382	1.577149	0.139147	
Boys	0.039952	0.019976	0.045455	1.545606	0.136364	
X2	3.611182					
DF=4	Table Value		9.488			H0 Rejected

PRACTICAL NO. 5

Performing testing of hypothesis using Z-Test.

Use a Z-Test if:

Your sample size is greater than 30. Otherwise, use a t test.

Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point.

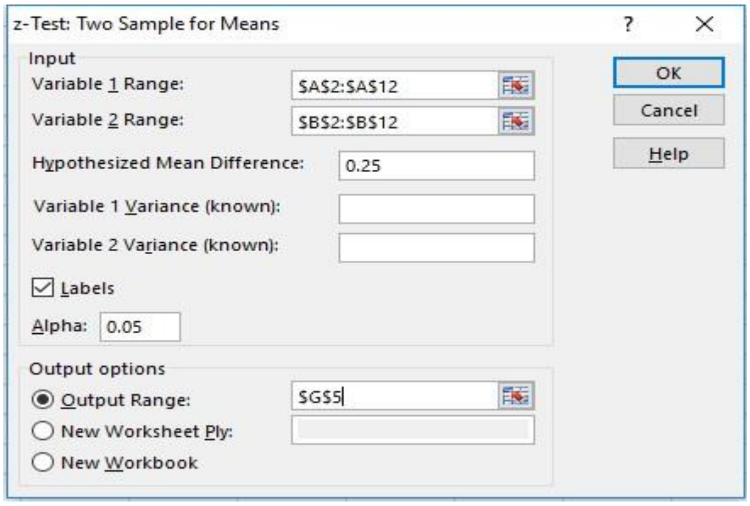
Your data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter.

Your data should be randomly selected from a population, where each item has an equal chance of being selected.

Sample sizes should be equal if at all possible.

H0: There is no difference between blood pressure before and after.

	A	B	C	D	E
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
14	13	Male	30-45	165	135
15	14	Male	30-45	145	159
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



G	H	I	J	K	L	M	N	O
Column1			*Column1*			z-Test: Two Sample for Means		
Mean	156.45		Mean	151.3583			*Variable 1*	*Variable 2*
Standard E	1.039746		Standard E	1.294234		Mean	156.45	151.3583
Median	154.5		Median	149.5		Known Va	129.7286	201.005
Mode	162		Mode	147		Observati	120	120
Standard I	11.38985		Standard I	14.17762		Hypothesi	0	
Sample Va	129.7286		Sample Va	201.005		z	3.066983	
Kurtosis	-0.43859		Kurtosis	-0.50515		P(Z<=z) or	0.001081	
Skewness	0.554244		Skewness	0.393365		z Critical c	1.644854	
Range	47		Range	60		P(Z<=z) tw	0.002162	
Minimum	138		Minimum	125		z Critical t	1.959964	
Maximum	185		Maximum	185				
Sum	18774		Sum	18163				
Count	120		Count	120				
H0- There is no difference between the blood pressure before and after								
H0 rejected								

PRACTICAL NO. 6

A) Perform testing of hypothesis using One-Way ANOVA.

ANOVA Assumptions:

The dependent variable (SAT scores in our example) should be continuous.

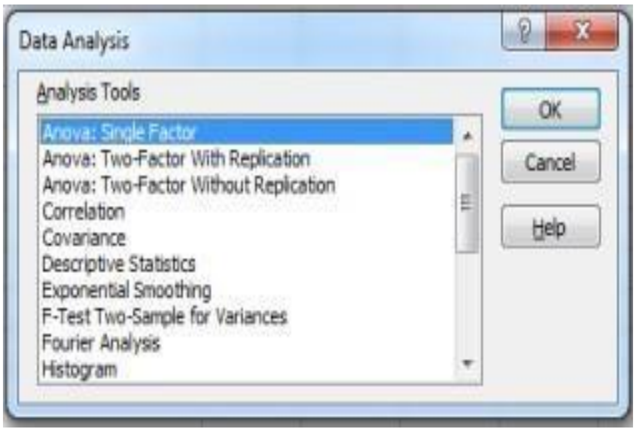
The independent variables (districts in our example) should be two or more categorical groups.

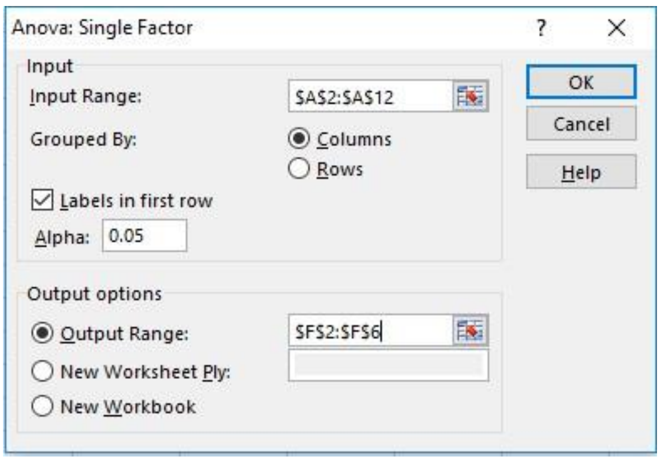
There must be different participants in each group with no participant being in more than one group. In our case, each school cannot be in more than one district.

The dependent variable should be approximately normally distributed for each category. Variances of each group are approximately equal.

H0: There is no variation in number of coffees.

	A	B	C
1	Shift	No. of Coffees	
2	Day	1	
3	Day	3	
4	Day	4	
5	Day	0	
6	Day	2	
7	Second	7	
8	Second	2	
9	Second	1	
10	Second	6	
11	Night	6	
12	Night	8	
13	Night	3	
14	Night	7	
15	Night	6	





H0-There is no variation in number of coffees

Day	Second	Night
1	7	6
3	2	8
4	1	3
0	6	7
2		6

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	5	10	2	2.5
Column 2	4	16	4	8.666667
Column 3	5	30	6	3.5

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	40	2	20	4.4	0.039446	3.982298
Within Groups	50	11	4.545455			
Total	90	13				

H0 Rejected

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

	A	B	C	D
1		supp	len	dose
2	1	VC	4.2	0.5
3	2	VC	11.5	0.5
4	3	VC	7.3	0.5
5	4	VC	5.8	0.5
6	5	VC	6.4	0.5
7	6	VC	10	0.5
8	7	VC	11.2	0.5
9	8	VC	11.2	0.5
10	9	VC	5.2	0.5
11	10	VC	7	0.5
12	11	VC	16.5	1
13	12	VC	16.5	1
14	13	VC	15.2	1
15	14	VC	17.3	1
16	15	VC	22.5	1
17	16	VC	17.3	1
18	17	VC	13.6	1
19	18	VC	14.5	1
20	19	VC	18.8	1
21	20	VC	15.5	1
22	21	VC	23.6	2
23	22	VC	18.5	2

Hypothysis for rows: There is no significant difference between the suppliment

Hypothysis for column: There is no significant difference between the len and dose

Anova: Two-Factor With Replication

SUMMARY	len	dose	Total
VC			
Count	30	30	60
Sum	508.9	35	543.9
Average	16.96333	1.166667	9.065
Variance	68.32723	0.402299	97.22333
OJ			
Count	30	30	60
Sum	619.9	35	654.9
Average	20.66333	1.166667	10.915
Variance	43.63344	0.402299	118.2854
Total			
Count	60	60	
Sum	1128.8	70	
Average	18.81333	1.166667	
Variance	58.51202	0.39548	

ANOVA

Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	102.675	1	102.675	3.642079	0.058808	3.922879
Columns	9342.145	1	9342.145	331.3838	8.55E-36	3.922879
Interactior	102.675	1	102.675	3.642079	0.058808	3.922879
Within	3270.193	116	28.19132			
Total	12817.69	119				

Hypothysis accepted that there is no significant difference between the suppliments

Hypothysis rejected that there is a significant difference between the len and dose

C) Perform testing of hypothesis using MANOVA.

	A	B	C	D
1	Gender	Economic	Kindness	Optimism
2	male	wealthy	5	3
3	male	wealthy	4	6
4	male	wealthy	3	4
5	male	wealthy	2	4
6	male	middle	4	6
7	male	middle	3	6
8	male	middle	5	4
9	male	middle	5	5
10	male	poor	7	5
11	male	poor	4	3
12	male	poor	3	1
13	male	poor	7	2
14	female	wealthy	2	3
15	female	wealthy	3	5
16	female	wealthy	5	3
17	female	wealthy	4	2
18	female	middle	9	8
19	female	middle	6	5
20	female	middle	7	6
21	female	middle	8	9
22	female	poor	8	9

F	G	H	I	J	K	L
Two-Way MANOVA						
fact A	*stat*	*df1*	*df2*	*F*	*p-value*	*part eta-sq*
Pillai Trac	0.190764	2	16	1.885866	0.183909	0.190764
Wilk's Lan	0.809236	2	16	1.885866	0.183909	0.190764
Hotelling	0.235733	2	16	1.885866	0.183909	0.190764
Roy's Lg R	0.235733					
fact B	*stat*	*df1*	*df2*	*F*	*p-value*	*part eta-sq*
Pillai Trac	0.340249	4	34	1.742501	0.163458	0.170125
Wilk's Lan	0.8181	4	32	1.778757	0.157443	0.1819
Hotelling	0.479878	4	30	1.799541	0.155008	0.193509
Roy's Lg R	0.448078					
fact AB	*stat*	*df1*	*df2*	*F*	*p-value*	*part eta-sq*
Pillai Trac	0.612127	4	34	3.748958	0.012446	0.306063
Wilk's Lan	0.66397	4	32	4.048738	0.009098	0.33603
Hotelling	1.148132	4	30	4.305494	0.007171	0.364703
Roy's Lg R	1.031635					

N	O
SSCP Matrices	
Tot	
104.9565	59.86957
59.86957	110.6087
Row (A)	
12.5247	15.41502
15.41502	18.97233
Column (B)	
31.15295	22.95885
22.95885	19.37655
Interaction (AB)	
11.02887	4.745695
4.745695	40.59314
Res	
50.25	16.75
16.75	31.66667

Q	R
Group Covariance Matrices	
female	middle
1.666667	2
2	3.333333
female	poor
7.583333	2.083333
2.083333	0.916667
female	wealthy
1.666667	-0.5
-0.5	1.583333
male	middle
0.916667	-0.75
-0.75	0.916667
male	poor
4.25	2.083333
2.083333	2.916667

PRACTICAL NO. 7

A) Perform the Random Sampling for the given data and analyze it.

	A	B	C	D	E	F	G	H	I
1	sr.no	roll no	name	gender	grade	random sample			
2	2	1002	tushar	male	o	0.712009077			
3	11	1011	shilpa	female	o	0.154686678		1018	
4	10	1010	sanjay	male	o	0.128870233		1001	
5	16	1016	jeevanya	female	o	0.088487238		1005	
6	9	1009	mayuresh	male	o	0.226328459		1011	
7	4	1004	umesh	male	o	0.780785169		1016	
8	18	1018	pallavi	female	o	0.918268098		1013	
9	20	1020	ashwini	female	o	0.889405784			
10	17	1017	tanvi	female	o	0.786585271			
11	3	1003	avaneesh	male	o	0.354570913			
12	12	1012	mangla	female	o	0.717085544			
13	1	1001	sonu	male	o	0.804576547			
14	19	1019	chaitali	female	o	0.080311045			
15	6	1006	chaitanya	male	d	0.613922527			
16	7	1007	rudransh	male	d	0.486692263			
17	5	1005	tanish	male	d	0.081390688			
18	14	1014	shalini	female	d	0.610977591			
19	8	1008	medhansh	male	d	0.616955848			
20	13	1013	neeta	female	d	0.251357692			
21	15	1015	shravani	female	d	0.22923629			

	A	B	C	D	E	F	G	H	I
1	sr.no	roll no	name	gender	grade	random sample			
2	2	1002	tushar	male	o	0.712009077			
3	11	1011	shilpa	female	o	0.154686678		1010	
4	10	1010	sanjay	male	o	0.128870233		1004	
5	16	1016	jeevanya	female	o	0.088487238		1017	
6	9	1009	mayuresh	male	o	0.226328459		1001	
7	4	1004	umesh	male	o	0.780785169		1007	
8	18	1018	pallavi	female	o	0.918268098		1008	
9	20	1020	ashwini	female	o	0.889405784			
10	17	1017	tanvi	female	o	0.786585271			
11	3	1003	avaneesh	male	o	0.354570913			
12	12	1012	mangla	female	o	0.717085544			
13	1	1001	sonu	male	o	0.804576547			
14	19	1019	chaitali	female	o	0.080311045			
15	6	1006	chaitanya	male	d	0.613922527			
16	7	1007	rudransh	male	d	0.486692263			
17	5	1005	tanish	male	d	0.081390688			
18	14	1014	shalini	female	d	0.610977591			
19	8	1008	medhansh	male	d	0.616955848			
20	13	1013	neeta	female	d	0.251357692			
21	15	1015	shravani	female	d	0.22923629			

B) Perform the Stratified Sampling for the given data and analyze it.

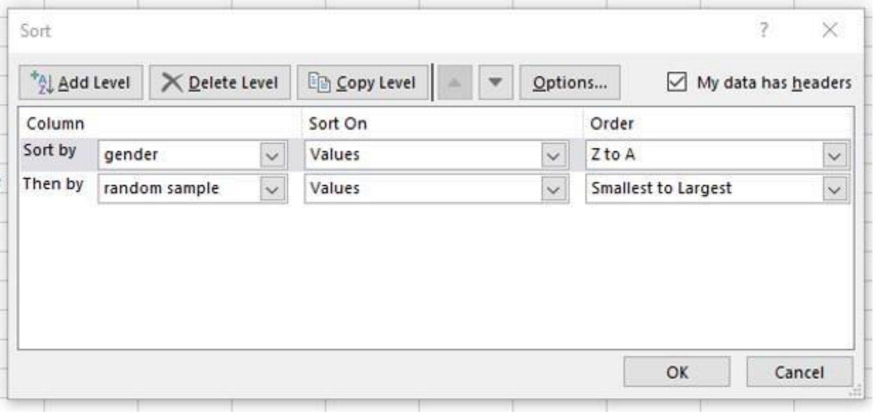
	A	B	C	D	E	F
1	sr.no	roll no	name	gender	grade	random sample
2	5	1005	tanish	male	d	0.081390688
3	10	1010	sanjay	male	o	0.128870233
4	9	1009	mayuresh	male	o	0.226328459
5	3	1003	avaneesh	male	o	0.354570913
6	7	1007	rudransh	male	d	0.486692263
7	6	1006	chaitanya	male	d	0.613922527
8	8	1008	medhansh	male	d	0.616955848
9	2	1002	tushar	male	o	0.712009077
10	4	1004	umesh	male	o	0.780785169
11	1	1001	sonu	male	o	0.804576547
12	19	1019	chaitali	female	o	0.080311045
13	16	1016	jeevanya	female	o	0.088487238
14	11	1011	shilpa	female	o	0.154686678
15	15	1015	shravani	female	d	0.22923629
16	13	1013	neeta	female	d	0.251357692
17	14	1014	shalini	female	d	0.610977591
18	12	1012	mangla	female	o	0.717085544
19	17	1017	tanvi	female	o	0.786585271
20	20	1020	ashwini	female	o	0.889405784
21	18	1018	pallavi	female	o	0.918268098

Stratified random sample for male

1003
1006
1002
1009
1009

stratified random sample for female

1015
1019
1016
1017
1014



PRACTICAL NO. 8

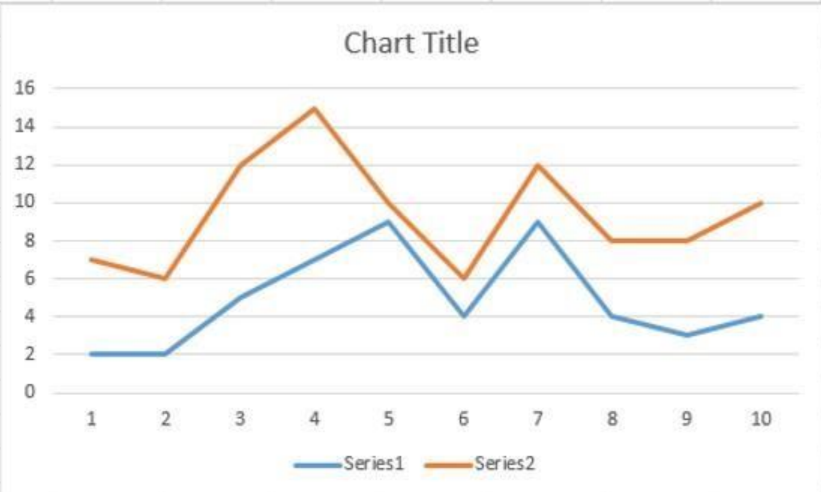
A) Write a program for computing different correlation.

Boys	Girls
2	5
2	4
5	7
7	8
9	1
4	2
9	3
4	4
3	5
4	6

-0.2457

	Column 1	*Column 2*
Column 1	1	
Column 2	-0.2457	1

There is no correlation between boys and girls marks



PRACTICAL NO. 9

A) Write a program to perform Linear Regression for prediction.

Linear regression is a basic and commonly used type of predictive analysis.

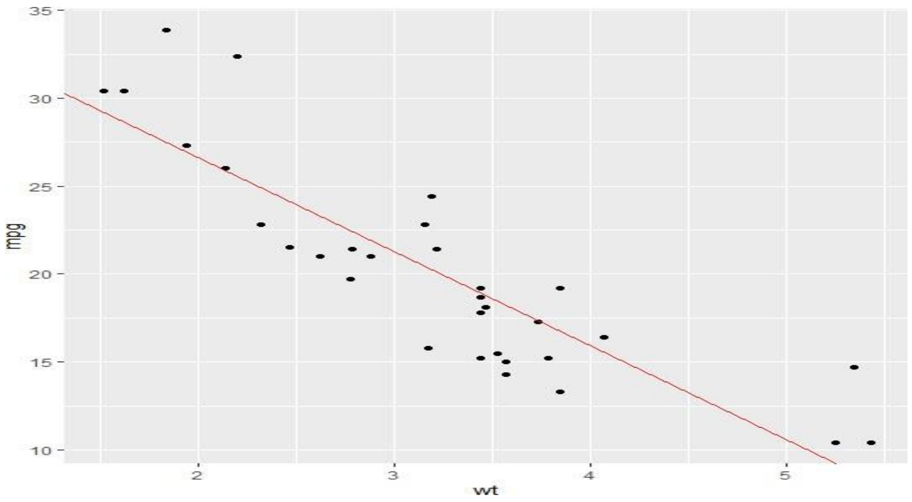
The overall idea of regression is to examine two things:

Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

```
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 3.6.2
> ggplot(data=mtcars, aes(x=wt, y=mpg)) +geom_point()
> mpg_model<-lm(mpg~wt,data=mtcars)
> summary(mpg_model)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```




```
> ggplot(data=mtcars,aes(x=wt,y=mpg))+geom_point()+geom_abline(intercept=37.2851,slope=-5.3445,color="red")
> preds<-predict(mpg_model,newdata=mtcars)
> preds
          Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive
          23.282611           21.919770           24.885952           20.102650
  Hornet Sportabout             Valiant          Duster 360           Merc 240D
          18.900144           18.793255           18.205363           20.236262
           Merc 230            Merc 280           Merc 280C          Merc 450SE
          20.450041           18.900144           18.900144           15.533127
         Merc 450SL         Merc 450SLC  Cadillac Fleetwood Lincoln Continental
          17.350247           17.083024            9.226650            8.296712
  Chrysler Imperial            Fiat 128         Honda Civic      Toyota Corolla
           8.718926           25.527289           28.653805           27.478021
      Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28
          24.111004           18.472586           18.926866           16.762355
   Pontiac Firebird           Fiat X1-9       Porsche 914-2        Lotus Europa
          16.735633           26.943574           25.847957           29.198941
     Ford Pantera L        Ferrari Dino       Maserati Bora          Volvo 142E
          20.343151           22.480940           18.205363           22.427495
```

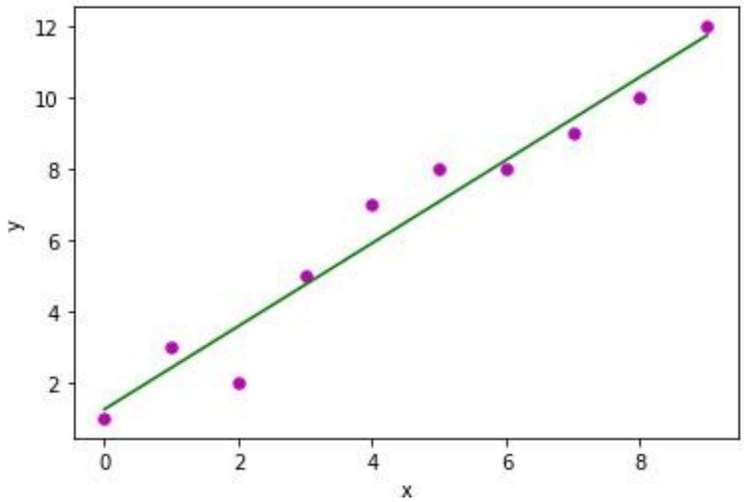
B) **Perform Polynomial Regression for prediction.**

Code:

```
import numpy as np

import matplotlib.pyplot as plt

def estimate_coef(x, y):

n = np.size(x)

m_x, m_y = np.mean(x), np.mean(y)

SS_xy = np.sum(y*x) - n*m_y*m_x

SS_xx = np.sum(x*x) - n*m_x*m_x

b_1 = SS_xy / SS_xx

b_0 = m_y - b_1*m_x

return(b_0, b_1)

def plot_regression_line(x, y, b):

plt.scatter(x, y, color = "m",marker = "o", s = 30)

y_pred = b[0] + b[1]*x

plt.plot(x, y_pred, color = "g")

plt.xlabel('x')

plt.ylabel('y')

plt.show()

def main():

x = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

y = np.array([1, 3, 2, 5, 7, 8, 8, 9, 10, 12])

b = estimate_coef(x, y)

print("Estimated coefficients:\nb_0 = {} b_1 = {}".format(b[0], b[1]))

plot_regression_line(x, y, b)

if __name__ == " main ": main()
```

Output:

```
In [1]: runfile('C:/Users/SIAC/9b.py', wdir='C:/Users/SIAC')
Estimated coefficients:
b_0 = 1.2363636363636363 b_1 = 1.1696969696969697
```



PRACTICAL NO. 10

A) Write a Program for Multiple Linear Regression analysis.

```
import numpy as np
import matplotlib as mpl
from mpl_toolkits.mplot3d
import Axes3D
import matplotlib.pyplot as plt
def generate_dataset(n):
x = []
y = []
random_x1 = np.random.rand()
random_x2 = np.random.rand()
for i in range(n):
x1 = i
x2 = i/2 + np.random.rand()*n
x.append([1, x1, x2])
y.append(random_x1 * x1 + random_x2 * x2 + 1)
return np.array(x), np.array(y)
x, y = generate_dataset(200)
mpl.rcParams['legend.fontsize'] = 12
fig = plt.figure()
ax = fig.gca(projection ='3d')
ax.scatter(x[:, 1], x[:, 2], y, label ='y', s = 5)
ax.legend()
ax.view_init(45, 0)
plt.show() d
ef mse(coef, x, y):
```

```
return np.mean((np.dot(x, coef) - y)**2)/2
def gradients(coef, x, y):
return np.mean(x.transpose()*(np.dot(x, coef) - y), axis = 1)
def multilinear_regression(coef, x, y, lr, b1 = 0.9, b2 = 0.999, epsilon = 1e-8):
prev_error = 0
m_coef = np.zeros(coef.shape)
v_coef = np.zeros(coef.shape) moment_m_coef = np.zeros(coef.shape)
moment_v_coef = np.zeros(coef.shape)
t = 0
while True:
error = mse(coef, x, y)
if abs(error - prev_error) <= epsilon:
break
prev_error = error
grad = gradients(coef, x, y)
t += 1
m_coef = b1 * m_coef + (1-b1)*grad
v_coef = b2 * v_coef + (1-b2)*grad**2
moment_m_coef = m_coef / (1-b1**t)
moment_v_coef = v_coef / (1-b2**t)
delta = ((lr / moment_v_coef**0.5 + 1e-8) *(b1 * moment_m_coef + (1-b1)*grad/(1-
b1**t)))
coef = np.subtract(coef, delta)
return coef
coef = np.array([0, 0, 0])
c = multilinear_regression(coef, x, y, 1e-1)
fig = plt.figure()
```

ax = fig.gca(projection ='3d')

ax.scatter(x[:, 1], x[:, 2], y, label ='y',s = 5, color ="dodgerblue")

ax.scatter(x[:, 1], x[:, 2], c[0] + c[1]*x[:, 1] + c[2]*x[:, 2],label ='regression', s = 5, color ="orange")

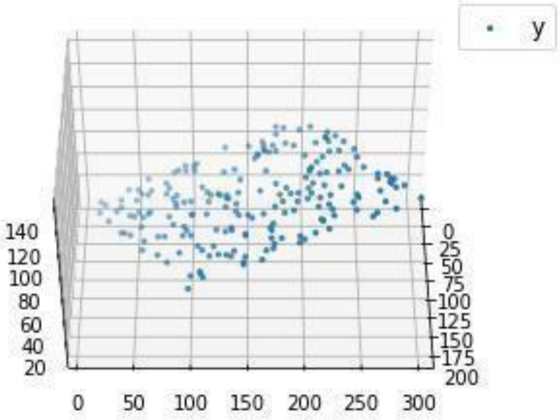
ax.view_init(45, 0)

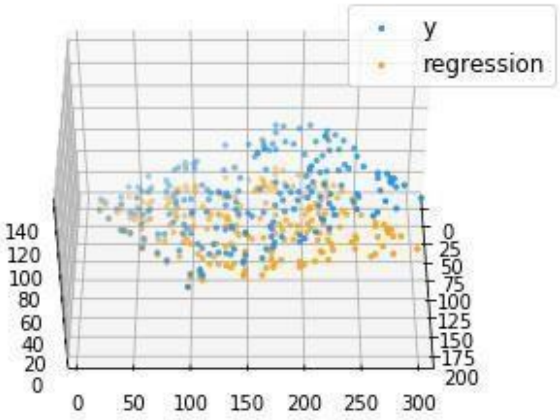
ax.legend()

plt.show()

Output:

```
In [1]: runfile('C:/Users/SIAC/10a.py', wdir='C:/Users/SIAC')
```





B) Perform Logistic Regression analysis.

Code:

```
import os
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import scipy.stats as stats
from sklearn import linear_model
from sklearn import preprocessing
from sklearn import metrics
matplotlib.style.use('ggplot')
plt.figure(figsize=(9,9))
def sigmoid(t):
return (1/(1 + np.e**(-t)))
plot_range = np.arange(-6, 6, 0.1)
y_values = sigmoid(plot_range)
 plt.plot(plot_range,y_values,color="red")
titanic_train = pd.read_csv("titanic_train.csv")
char_cabin = titanic_train["Cabin"].astype(str)
new_Cabin = np.array([cabin[0] for cabin in char_cabin])
titanic_train["Cabin"] = pd.Categorical(new_Cabin)
new_age_var = np.where(titanic_train["Age"].isnull(),
        28,
        titanic_train["Age"])
titanic_train["Age"] = new_age_var
label_encoder = preprocessing.LabelEncoder()
```

```
encoded_sex = label_encoder.fit_transform(titanic_train["Sex"])
log_model = linear_model.LogisticRegression()
log_model.fit(X = pd.DataFrame(encoded_sex),
        y = titanic_train["Survived"])
print(log_model.intercept_)
print(log_model.coef_)
preds = log_model.predict_proba(X= pd.DataFrame(encoded_sex))
preds = pd.DataFrame(preds)
preds.columns = ["Death_prob", "Survival_prob"]
pd.crosstab(titanic_train["Sex"], preds.ix[:, "Survival_prob"])
encoded_class = label_encoder.fit_transform(titanic_train["Pclass"])
encoded_cabin = label_encoder.fit_transform(titanic_train["Cabin"])
train_features = pd.DataFrame([encoded_class,
        encoded_cabin,
        encoded_sex,
        titanic_train["Age"]]).T
log_model = linear_model.LogisticRegression()
log_model.fit(X = train_features ,
        y = titanic_train["Survived"])
print(log_model.intercept_)
print(log_model.coef_)
preds = log_model.predict(X= train_features)
pd.crosstab(preds,titanic_train["Survived"])
log_model.score(X = train_features , y = titanic_train["Survived"])
metrics.confusion_matrix(y_true=titanic_train["Survived"], y_pred=preds)
print(metrics.classification_report(y_true=titanic_train["Survived"], y_pred=preds) )
titanic_test = pd.read_csv("titanic_test.csv")
```

```
char_cabin = titanic_test["Cabin"].astype(str)

new_Cabin = np.array([cabin[0] for cabin in char_cabin])

titanic_test["Cabin"] = pd.Categorical(new_Cabin)

new_age_var = np.where(titanic_test["Age"].isnull(),

        28,

        titanic_test["Age"])

titanic_test["Age"] = new_age_var

encoded_sex = label_encoder.fit_transform(titanic_test["Sex"])

encoded_class = label_encoder.fit_transform(titanic_test["Pclass"])

encoded_cabin = label_encoder.fit_transform(titanic_test["Cabin"])

test_features = pd.DataFrame([encoded_class,
encoded_cabin,encoded_sex,titanic_test["Age"]]).T

test_preds = log_model.predict(X=test_features)

submission = pd.DataFrame({"PassengerId":titanic_test["PassengerId"],

"Survived":test_preds})

submission.to_csv("tutorial_logreg_submission.csv", index=False)

print(pd)
```

Output: