**PSIT1P1**

**RESEARCH IN COMPUTING PRACTICAL**

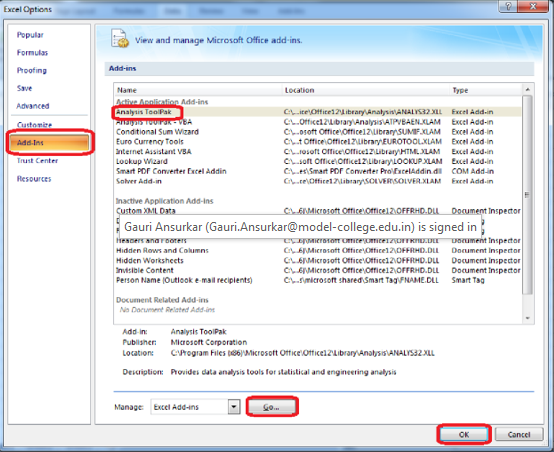
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sr No. | Practical No. | | Name of the Practical | Datasets | Date | Signature |
| 1. | 1. | A | Write a program for obtaining descriptive statistics  of data. |  |  |  |
| 2. | B | Import data from different data sources (from Excel,  csv, mysql, sql server, oracle to R/Python/Excel) |  |  |  |
| 3. | 2. | B | Perform suitable analysis of given secondary data. |  |  |  |
| 4. | 3. | A | Perform testing of hypothesis using one sample t- test. | ages.csv |  |  |
| 5. | B | Perform testing of hypothesis using two sample t-test. |  |  |  |
| 6. | C | Perform testing of hypothesis using paired t-test. | blood\_pressure.csv |  |  |
| 7. | 4. | A | Perform testing of hypothesis using chi-squared goodness-of-fit test. |  |  |  |
| 8. | B | Perform testing of hypothesis using chi-squared Test of Independence | Students\_Score.xlsx |  |  |
| 9. | 5. | A | Compute different types of correlation |  |  |  |
| 10. | 6. | A | Perform testing of hypothesis using one-way ANOVA. | scores.xlsx |  |  |
| 11. |  | B | Perform testing of hypothesis using two-way ANOVA. | ToothGrowth.csv |  |  |
| 12. | 7. | A | Perform linear regression for prediction |  |  |  |
| 13. | 8. | A | Perform Logistic regression | quality.csv |  |  |
| 14. | 9. | A | Perform testing of hypothesis using Z-test | blood\_pressure.csv |  |  |

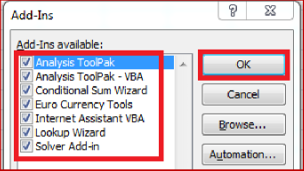
Practical 1:

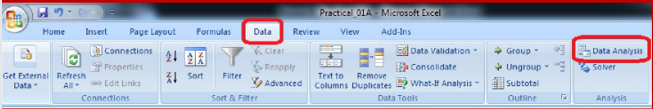
1. Write a program for obtaining descriptive statistics of data

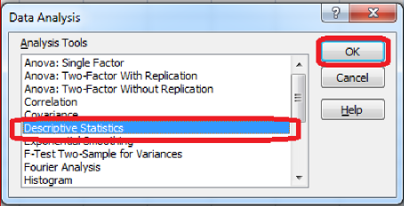
**Using Excel**

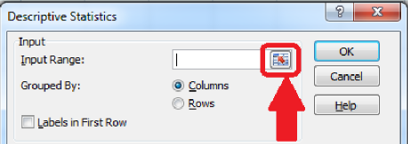
Go to File Menu **** Options **** Add-Ins**** Select Analysis ToolPak**** Press OK



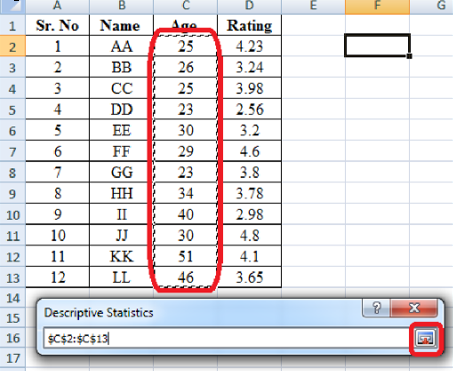


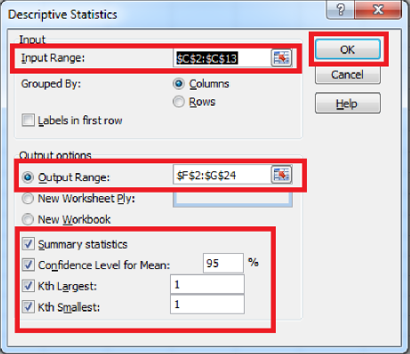




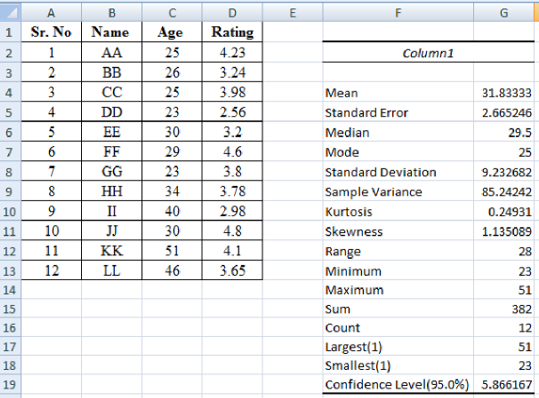


Select the data range from the excel worksheet.





OUTPUT:



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

NOTE: Create database in MySqL named as itvoyagers using command:

create database itvoyagers;

import mysql.connector

#creating connection object

db=mysql.connector.connect(user='root',passwd='',host='127.0.0.1',database='itvoyagers')

# prepare a cursor object using cursor() method

cur = db.cursor()

# execute SQL query using execute() method.

cur.execute("SELECT VERSION()")

# Fetch a single row using fetchone() method.

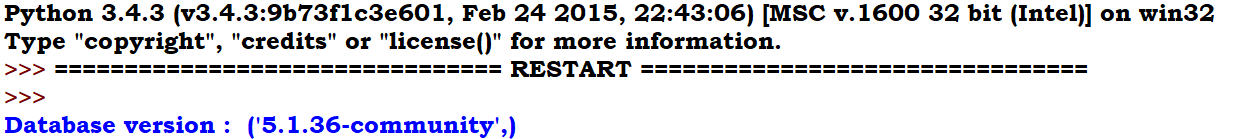
data = cur.fetchone()

print ("Database version : " , data)

# disconnect from server

db.close()

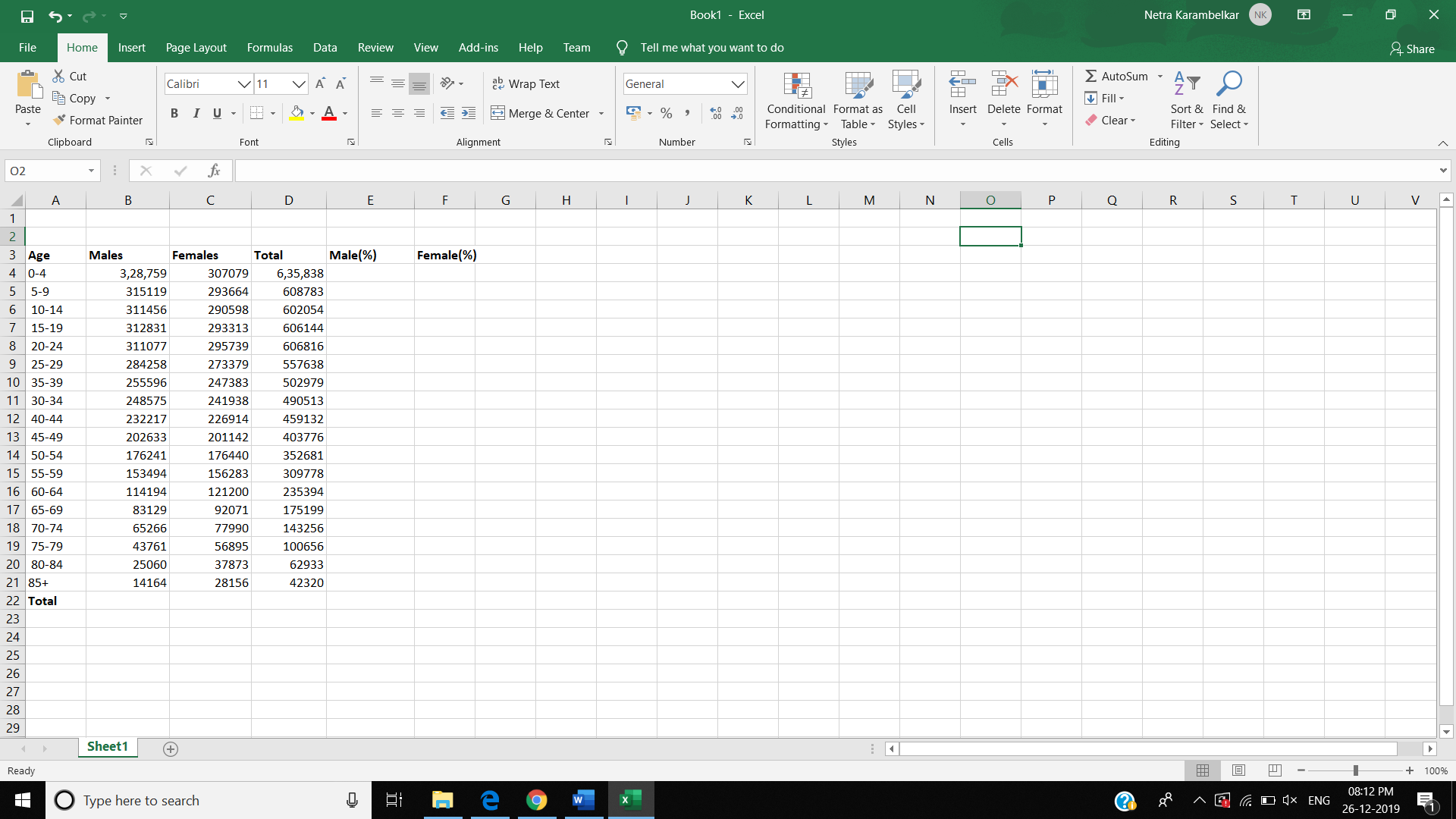
OUTPUT



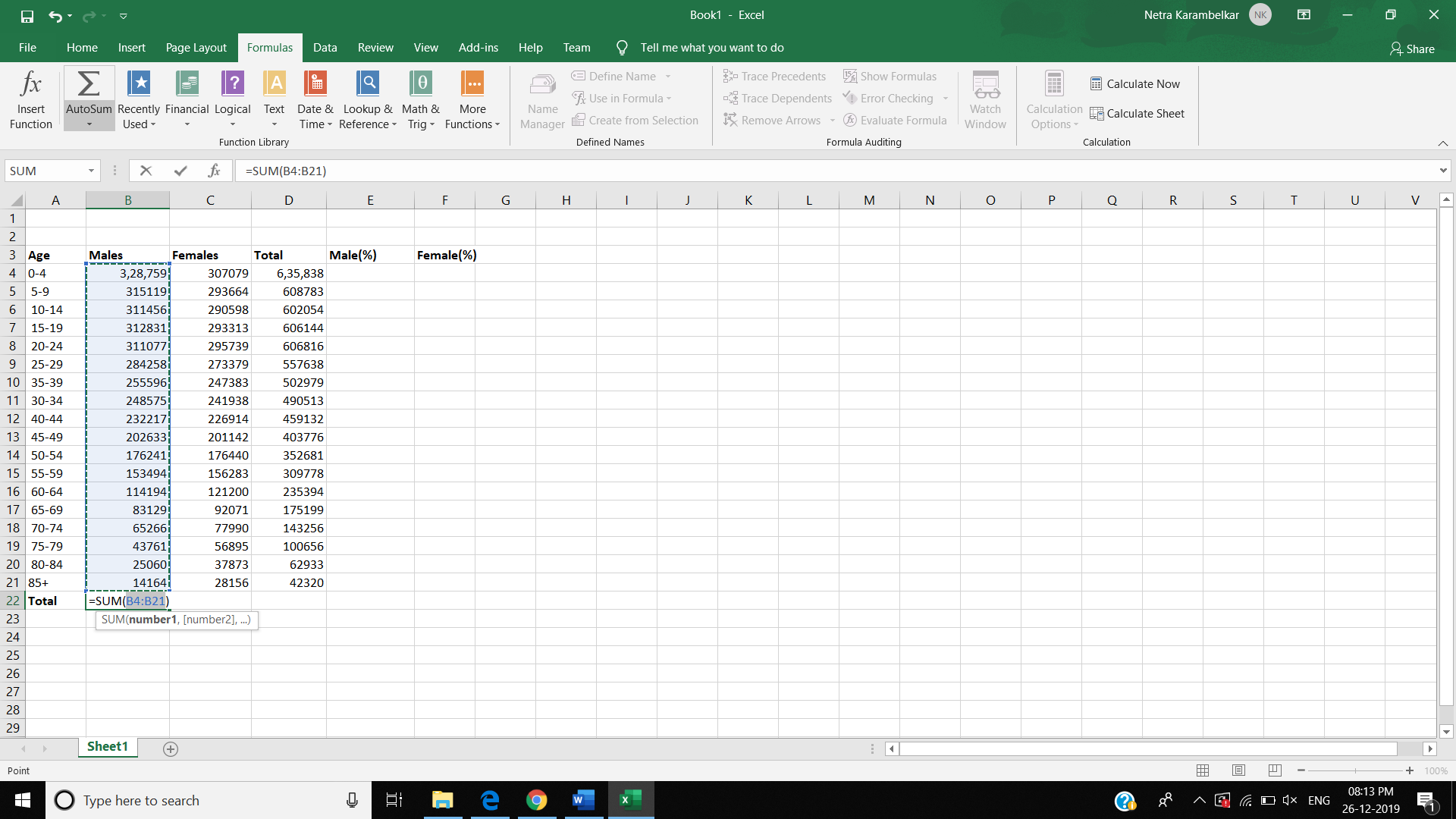
Practical 2:

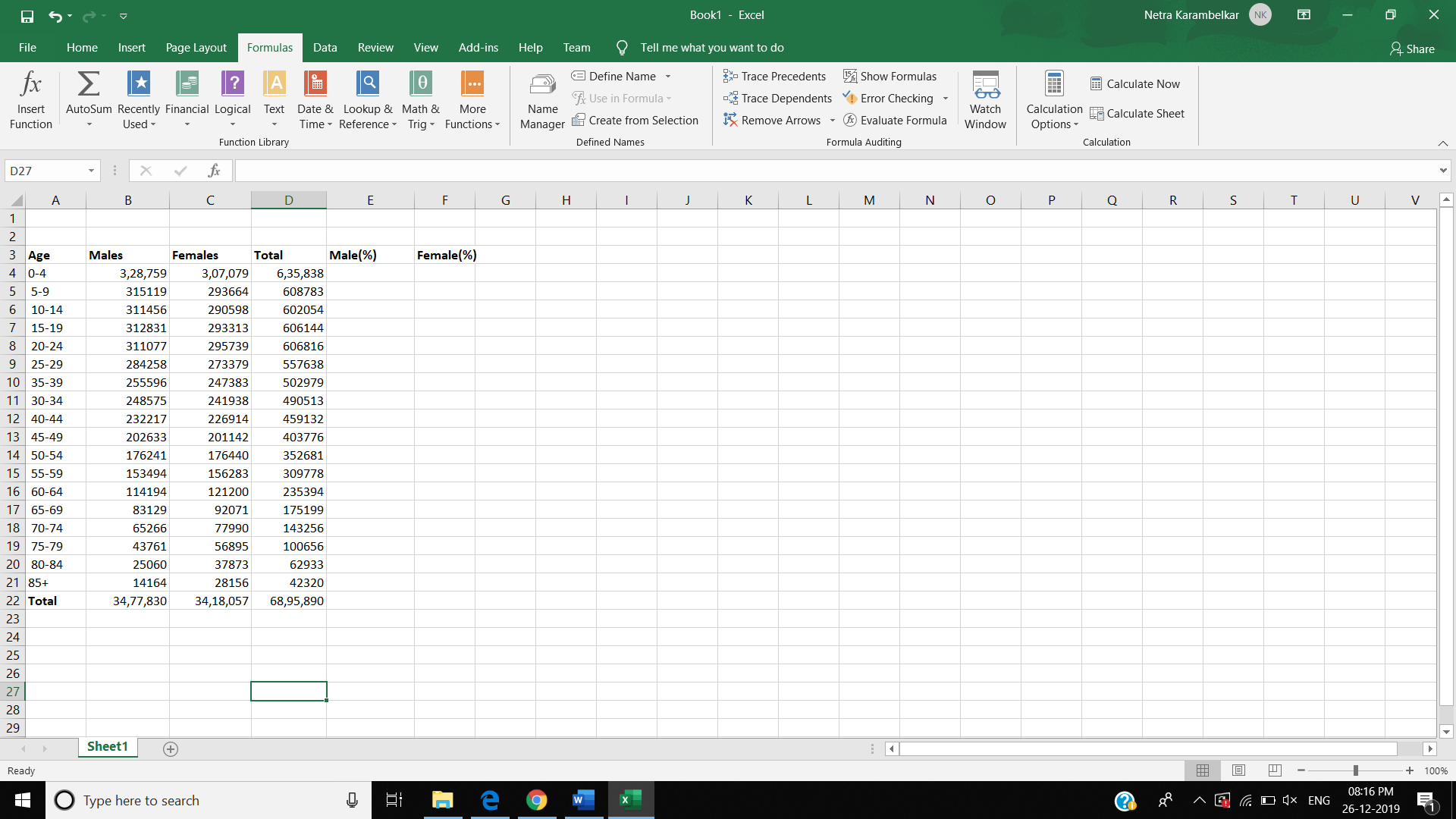
B. Write a program for obtaining descriptive statistics of data

**Step 1 -** Analyse the given Population Census Data for Planning and Decision Making by using the size and composition of populations.

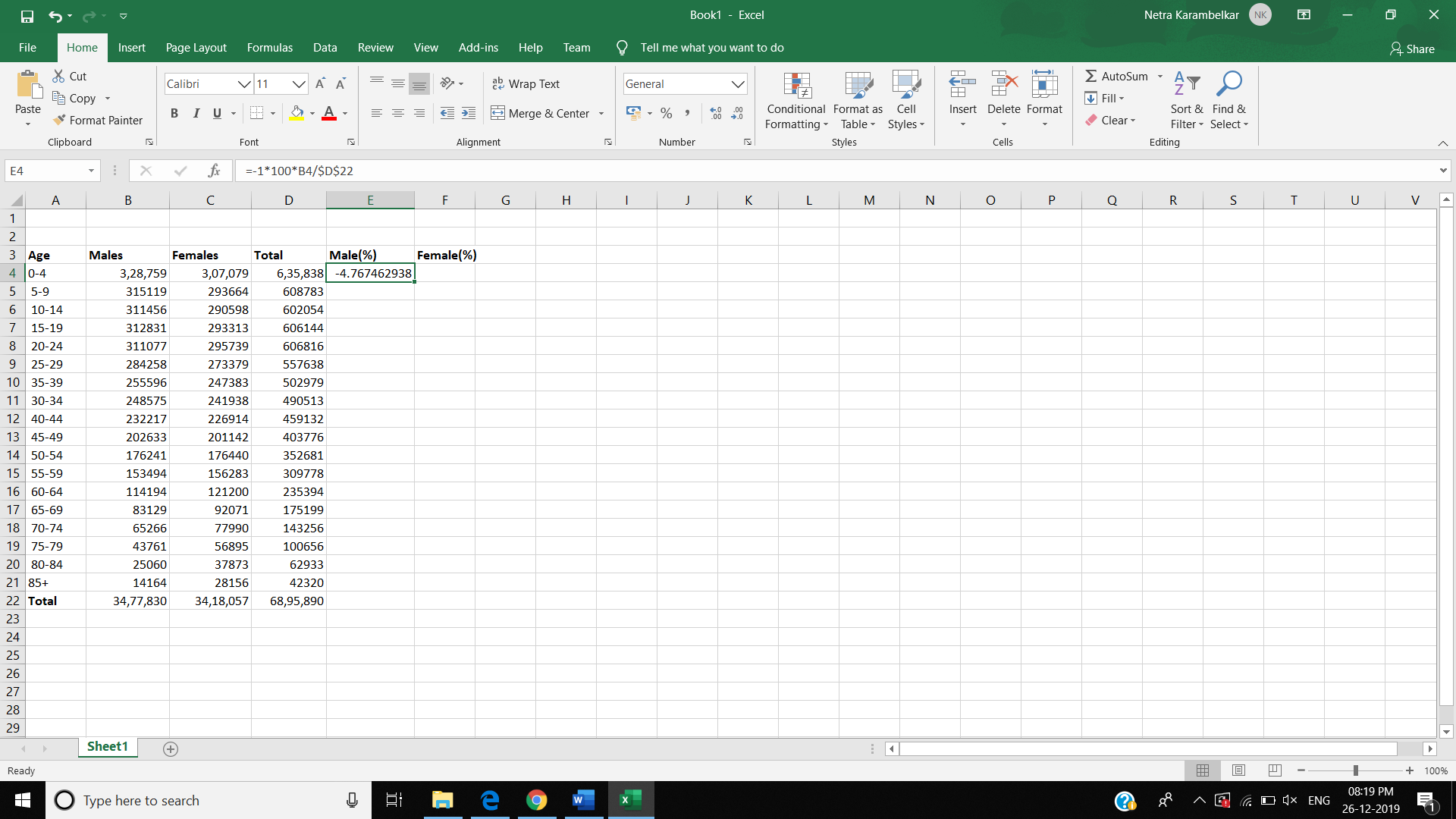


**Step 2 -** Put the cursor in cell B22 and click on the AutoSum and then click Enter. This will calculate the total population. Then copy the formula in cell D22 across the row 22.

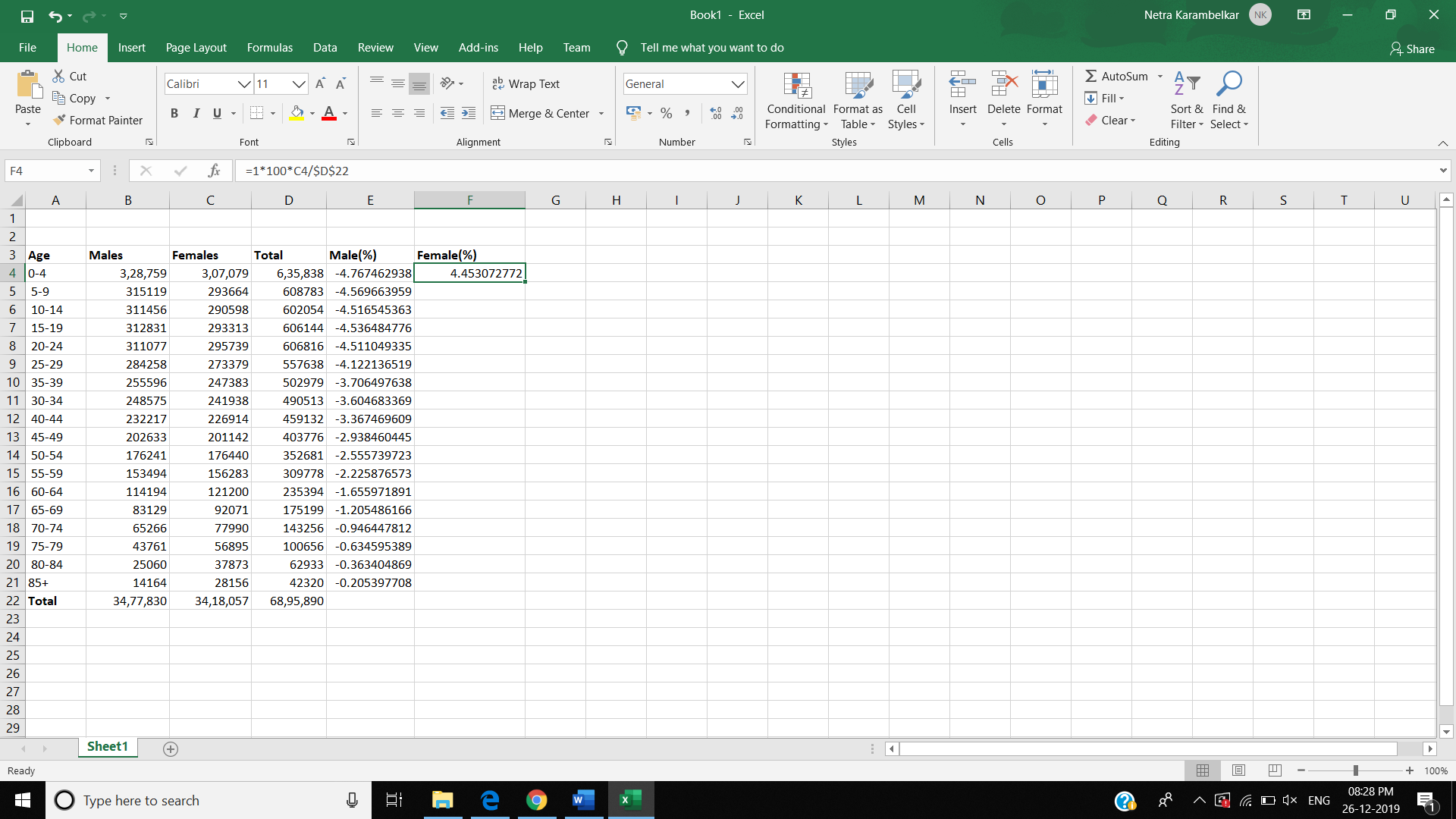




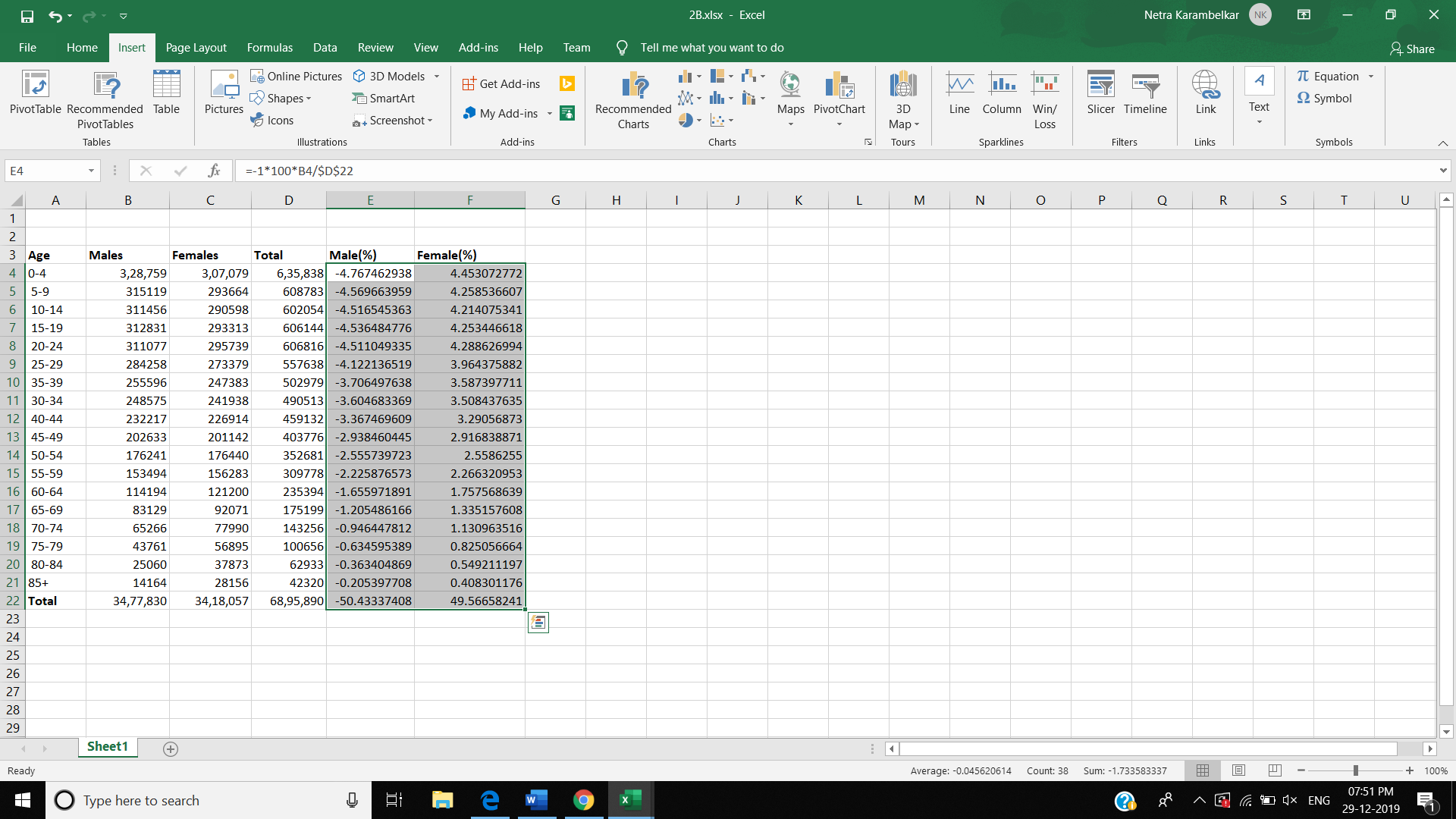
**Step 3 -** To calculate the percent of males in cell E4, enter the formula = -1\*100\*B4/$D$22. And copy the formula in cell E4 down to cell E21.

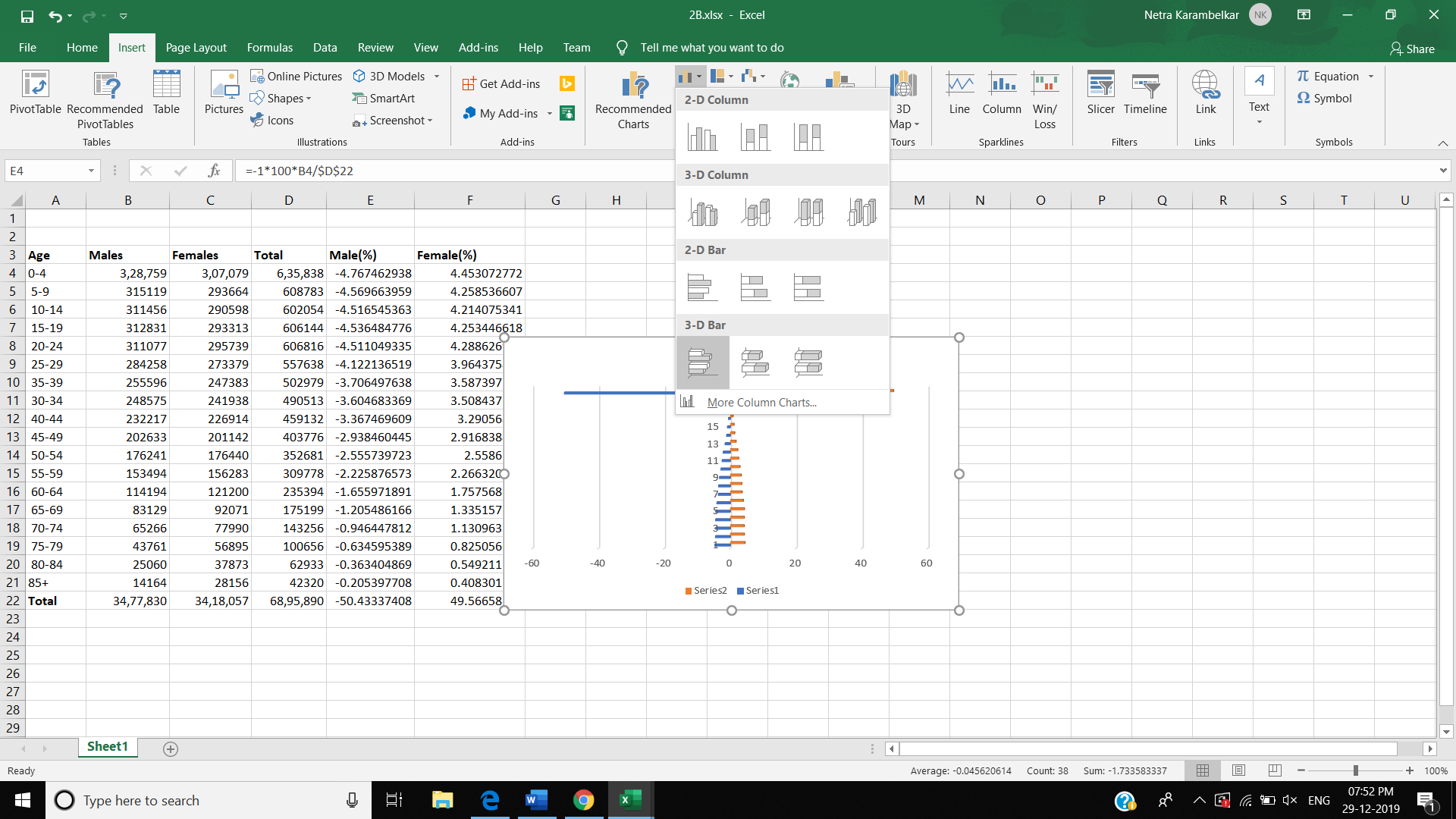


**Step 4 -** To calculate the percent of females in cell F4, enter the formula =1\*100\*C4/$D$22. Copy the formula in cell F4 down to cell F21.

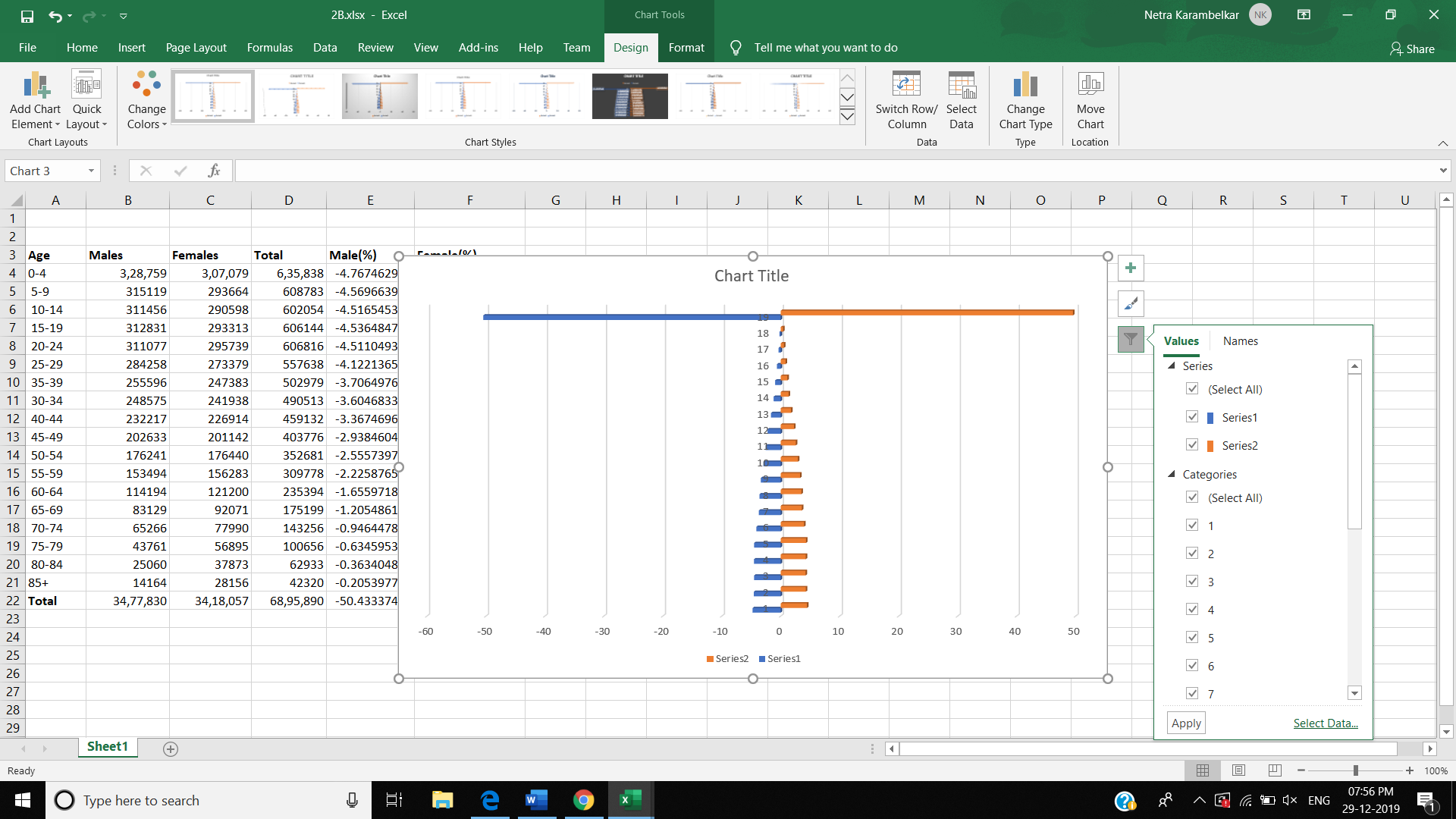


**Step 5 -** To build the population pyramid, we need to choose a horizontal bar chart with two series of data (% male and % female) and the age labels in column A as the Category X-axis labels. Highlight the range E3:F21 and under inset tab, under horizontal bar charts select clustered bar chart.

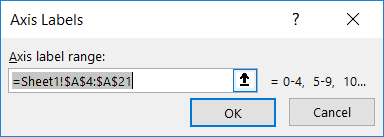




**Step 6 –** Go to Charts Filter option and click on Select Data.

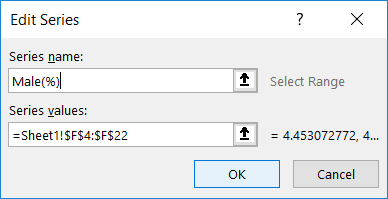
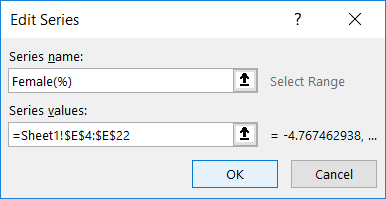


**Step 7 –** In Select Data Source Tab 🡪 Horizontal (Category) axis label and select the age range from A4:A21



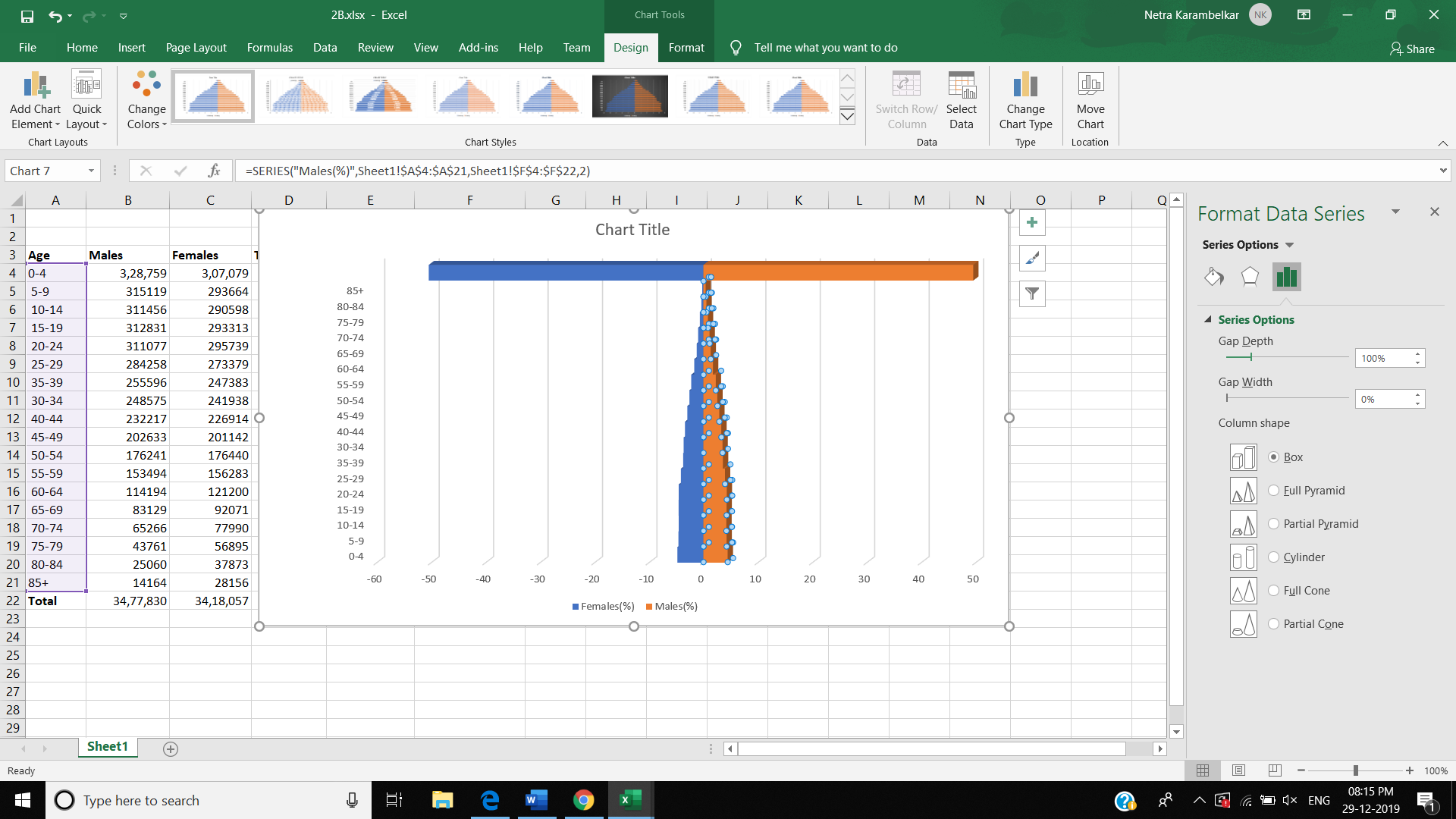
Also, in Legend entries (series) 🡪 Series1 🡪 Edit 🡪 Series Name = Female(%).

Series2 🡪 Edit 🡪 Series Name = Male(%). And finally click on Ok.

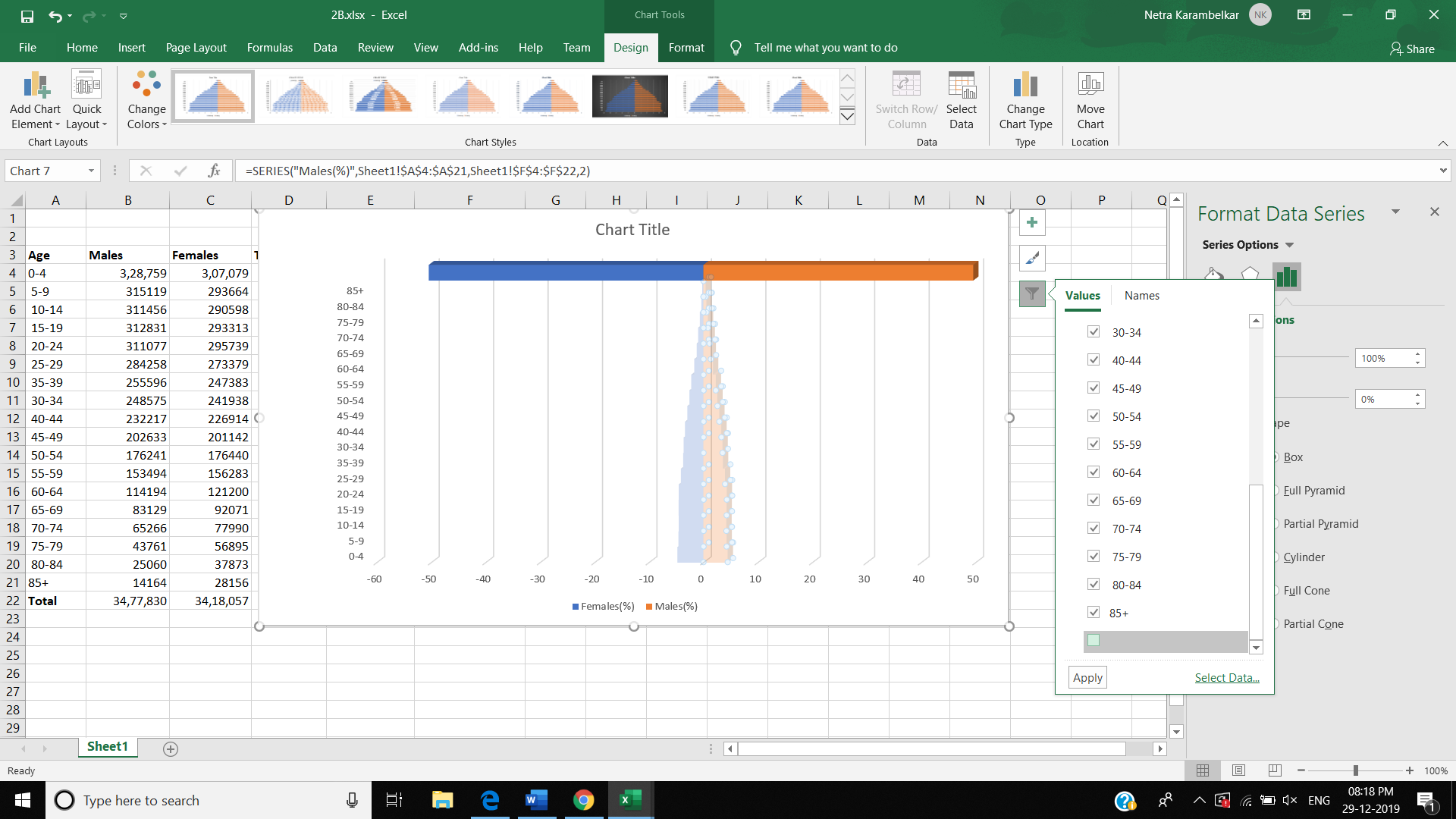


**Step 8** – Now in the chart 🡪 Right click on Age range 🡪 Format Axis 🡪 Axis Options 🡪 In Tick Marks 🡪 Major Type, Minor Type = None 🡪 In Labels 🡪 Labels Position = Low.

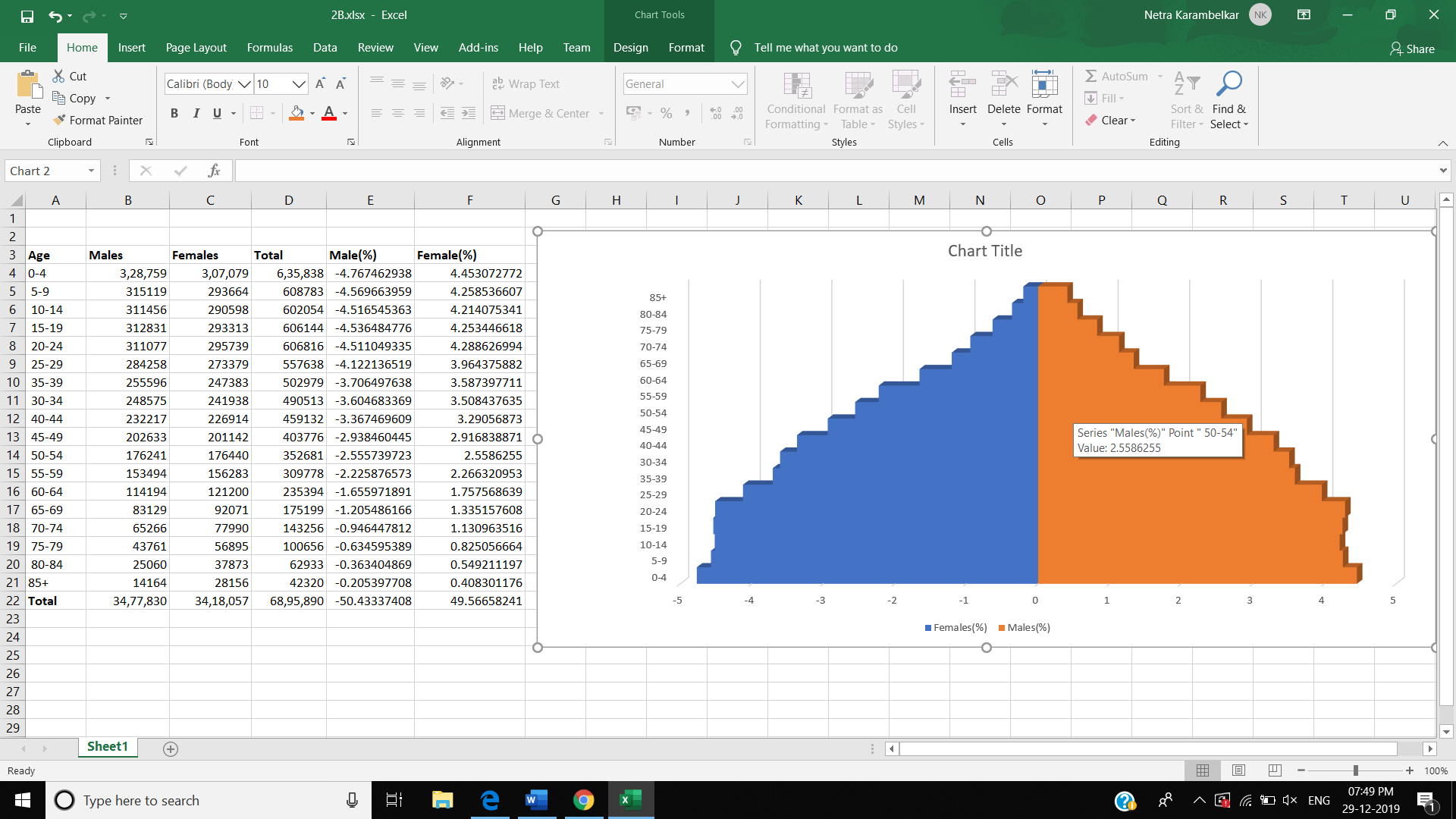
**Step 9** – Now put the tip of your mouse arrow on anywhere on the bars of chart 🡪 Right Click 🡪 Format Data Series 🡪 Set the Overlap to 100 and Gap Width to 0 🡪 Click OK.



**Step 10 -** Go to Charts Filter option 🡪 Scroll down Categories 🡪 Untick the option after 85+ and Click on Apply.



**OUTPUT:**



Practical 3:

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

USING PYTHON

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:

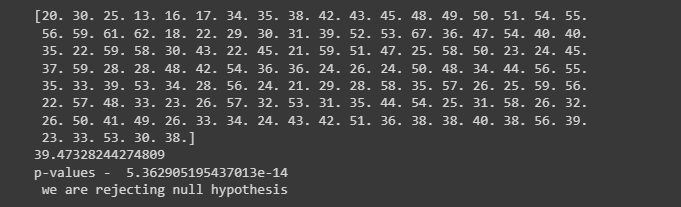
    # alpha value is 0.05

    print(" we are rejecting null hypothesis")

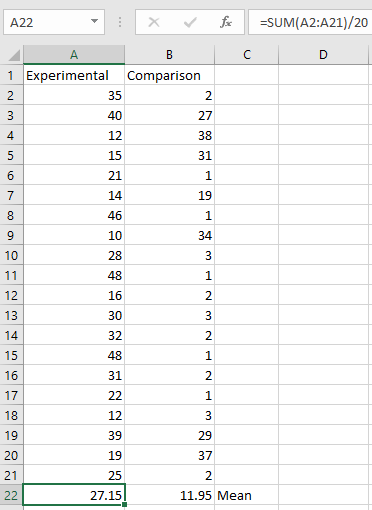
else:

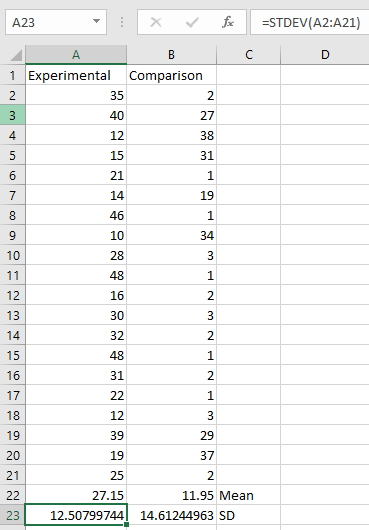
    print("we are accepting null hypothesis")

OUTPUT:



USING EXCEL





Experimental Data

To calculate Standard Mean go to cell A22 and type =SUM(A2:A21)/20 To calculate Standard Deviation go to cell A23 and type =STDEV(A2:A21)

Comparison Data

To calculate Standard Mean go to cell B22 and type =SUM(B2:B21)/20

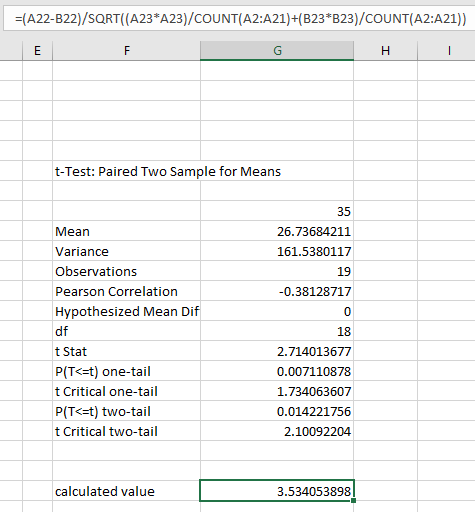
To calculate Standard Deviation go to cell B23 and type =STDEV(B2:B21) To find T-Test Statistics go to data **** Data Analysis

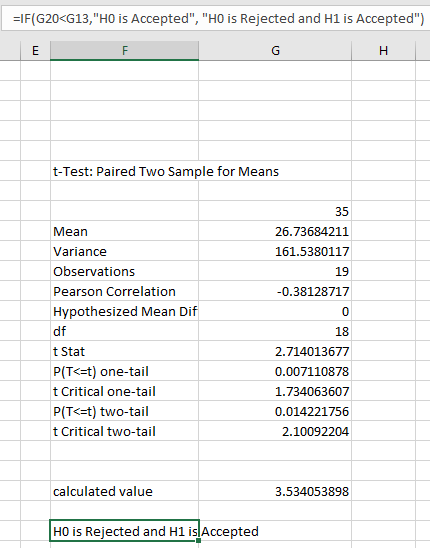
To caluculate the T-Test square value go to cell E20 and type =(A22-B22)/SQRT((A23\*A23)/COUNT(A2:A21)+(B23\*B23)/COUNT(A2:A21))

Now go to cell E20 and type

=IF(E20<E12,"H0 is Accepted", "H0 is Rejected and H1 is Accepted")

Our calculated value is larger than the tabled value at alpha = .01, so we reject the null hypothesisand accept the alternative hypothesis, namely, that the difference in gain scores is likely the resultof the experimental treatment and not the result of chance variation.





1. Perform testing of hypothesis using two sample t-test.

import numpy as np

from scipy import stats

from numpy.random import randn

N = 20

#a = [35,40,12,15,21,14,46,10,28,48,16,30, 32,48,31,22,12,39,19,25]

#b = [2,27,31,38,1,19,1,34,3,1,2,1,3,1,2,1,3,29,37,2]

a = 5 \* randn(100) + 50

b = 5 \* randn(100) + 51

var\_a = a.var(ddof=1)

var\_b = b.var(ddof=1)

s = np.sqrt((var\_a + var\_b)/2)

t = (a.mean() - b.mean())/(s\*np.sqrt(2/N))

df = 2\*N - 2

#p-value after comparison with the t

p = 1 - stats.t.cdf(t,df=df)

print("t = " + str(t))

print("p = " + str(2\*p))

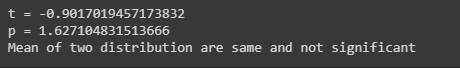
if t> p :

print('Mean of two distribution are differnt and significant')

else:

print('Mean of two distribution are same and not significant')

OUTPUT:



1. Perform testing of hypothesis using paired t-test.

from scipy import stats

import matplotlib.pyplot as plt

import pandas as pd

df = pd.read\_csv("blood\_pressure.csv")

print(df[['bp\_before','bp\_after']].describe())

#First let’s check for any significant outliers in

#each of the variables.

df[['bp\_before', 'bp\_after']].plot(kind='box')

# This saves the plot as a png file

plt.savefig('boxplot\_outliers.png')

# make a histogram to differences between the two scores.

df['bp\_difference'] = df['bp\_before'] - df['bp\_after']

df['bp\_difference'].plot(kind='hist', title= 'Blood Pressure Difference Histogram')

#Again, this saves the plot as a png file

plt.savefig('blood pressure difference histogram.png')

stats.probplot(df['bp\_difference'], plot= plt)

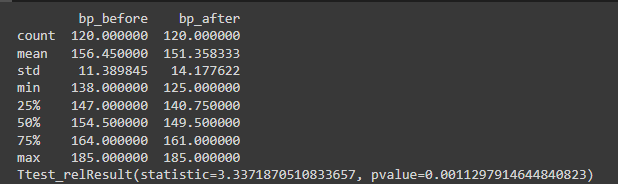
plt.title('Blood pressure Difference Q-Q Plot')

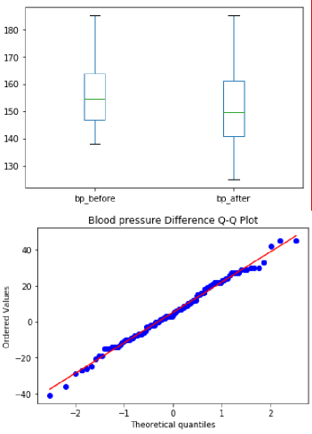
plt.savefig('blood pressure difference qq plot.png')

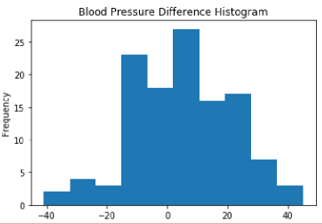
stats.shapiro(df['bp\_difference'])

stats.ttest\_rel(df['bp\_before'], df['bp\_after'])

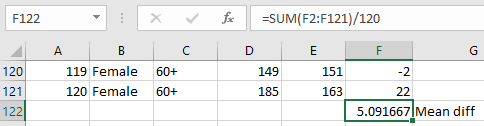
OUTPUT:

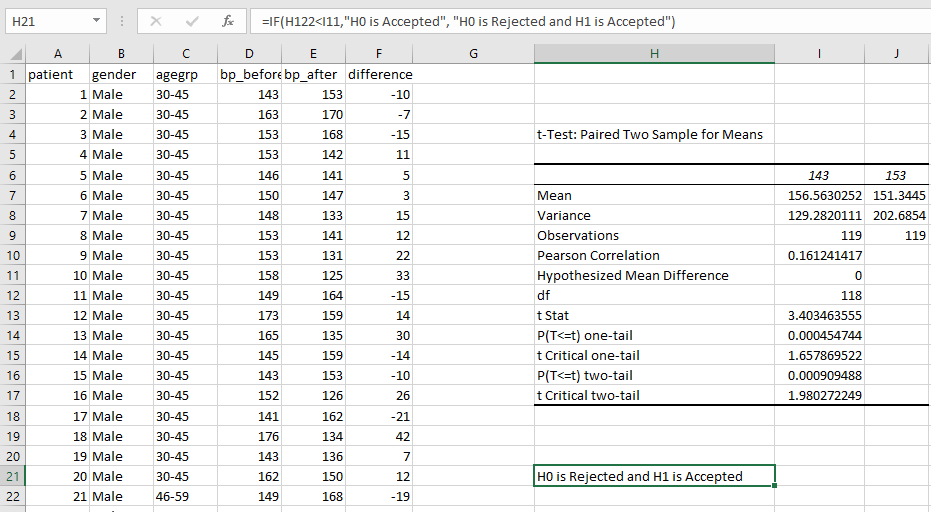




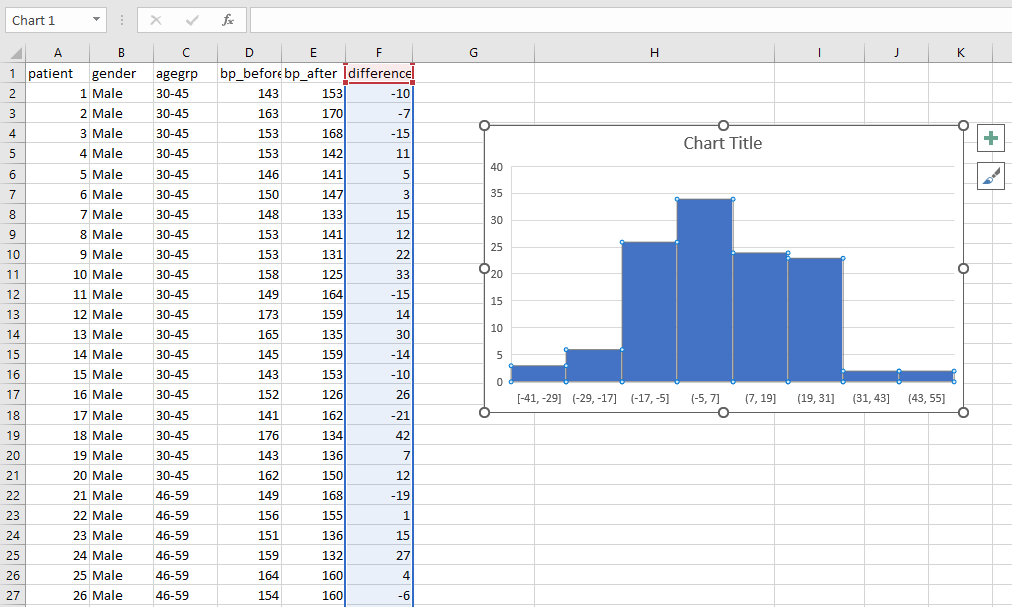


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Patient | gender | agegrp | bp\_before | bp\_after | Difference |
| 1 | Male | 30-45 | 143 | 153 | -10 |
| 2 | Male | 30-45 | 163 | 170 | -7 |
| 3 | Male | 30-45 | 153 | 168 | -15 |
| 4 | Male | 30-45 | 153 | 142 | 11 |
| 5 | Male | 30-45 | 146 | 141 | 5 |
| 6 | Male | 30-45 | 150 | 147 | 3 |
| 7 | Male | 30-45 | 148 | 133 | 15 |
| 8 | Male | 30-45 | 153 | 141 | 12 |
| 9 | Male | 30-45 | 153 | 131 | 22 |
| 10 | Male | 30-45 | 158 | 125 | 33 |
| 11 | Male | 30-45 | 149 | 164 | -15 |
| 12 | Male | 30-45 | 173 | 159 | 14 |
| 13 | Male | 30-45 | 165 | 135 | 30 |
| 14 | Male | 30-45 | 145 | 159 | -14 |
| 15 | Male | 30-45 | 143 | 153 | -10 |
| 16 | Male | 30-45 | 152 | 126 | 26 |
| 17 | Male | 30-45 | 141 | 162 | -21 |
| 18 | Male | 30-45 | 176 | 134 | 42 |
| 19 | Male | 30-45 | 143 | 136 | 7 |
| 20 | Male | 30-45 | 162 | 150 | 12 |
| 21 | Male | 46-59 | 149 | 168 | -19 |
| 22 | Male | 46-59 | 156 | 155 | 1 |
| 23 | Male | 46-59 | 151 | 136 | 15 |
| 24 | Male | 46-59 | 159 | 132 | 27 |
| 25 | Male | 46-59 | 164 | 160 | 4 |
| 26 | Male | 46-59 | 154 | 160 | -6 |
| 27 | Male | 46-59 | 152 | 136 | 16 |
| 28 | Male | 46-59 | 142 | 183 | -41 |
| 29 | Male | 46-59 | 162 | 152 | 10 |
| 30 | Male | 46-59 | 155 | 162 | -7 |
| 31 | Male | 46-59 | 175 | 151 | 24 |
| 32 | Male | 46-59 | 184 | 139 | 45 |
| 33 | Male | 46-59 | 167 | 175 | -8 |
| 34 | Male | 46-59 | 148 | 184 | -36 |
| 35 | Male | 46-59 | 170 | 151 | 19 |
| 36 | Male | 46-59 | 159 | 171 | -12 |
| 37 | Male | 46-59 | 149 | 157 | -8 |
| 38 | Male | 46-59 | 140 | 159 | -19 |
| 39 | Male | 46-59 | 185 | 140 | 45 |
| 40 | Male | 46-59 | 160 | 174 | -14 |
| 41 | Male | 60+ | 157 | 167 | -10 |
| 42 | Male | 60+ | 158 | 158 | 0 |
| 43 | Male | 60+ | 162 | 168 | -6 |
| 44 | Male | 60+ | 160 | 159 | 1 |
| 45 | Male | 60+ | 180 | 153 | 27 |
| 46 | Male | 60+ | 155 | 164 | -9 |
| 47 | Male | 60+ | 172 | 169 | 3 |
| 48 | Male | 60+ | 157 | 148 | 9 |
| 49 | Male | 60+ | 171 | 185 | -14 |
| 50 | Male | 60+ | 170 | 163 | 7 |
| 51 | Male | 60+ | 175 | 146 | 29 |
| 52 | Male | 60+ | 175 | 160 | 15 |
| 53 | Male | 60+ | 172 | 175 | -3 |
| 54 | Male | 60+ | 173 | 163 | 10 |
| 55 | Male | 60+ | 170 | 185 | -15 |
| 56 | Male | 60+ | 164 | 146 | 18 |
| 57 | Male | 60+ | 147 | 176 | -29 |
| 58 | Male | 60+ | 154 | 147 | 7 |
| 59 | Male | 60+ | 172 | 161 | 11 |
| 60 | Male | 60+ | 162 | 164 | -2 |
| 61 | Female | 30-45 | 152 | 149 | 3 |
| 62 | Female | 30-45 | 147 | 142 | 5 |
| 63 | Female | 30-45 | 144 | 146 | -2 |
| 64 | Female | 30-45 | 144 | 138 | 6 |
| 65 | Female | 30-45 | 158 | 131 | 27 |
| 66 | Female | 30-45 | 147 | 145 | 2 |
| 67 | Female | 30-45 | 154 | 134 | 20 |
| 68 | Female | 30-45 | 151 | 135 | 16 |
| 69 | Female | 30-45 | 149 | 131 | 18 |
| 70 | Female | 30-45 | 138 | 135 | 3 |
| 71 | Female | 30-45 | 162 | 133 | 29 |
| 72 | Female | 30-45 | 157 | 135 | 22 |
| 73 | Female | 30-45 | 141 | 168 | -27 |
| 74 | Female | 30-45 | 167 | 144 | 23 |
| 75 | Female | 30-45 | 147 | 147 | 0 |
| 76 | Female | 30-45 | 143 | 151 | -8 |
| 77 | Female | 30-45 | 142 | 149 | -7 |
| 78 | Female | 30-45 | 166 | 147 | 19 |
| 79 | Female | 30-45 | 147 | 149 | -2 |
| 80 | Female | 30-45 | 142 | 135 | 7 |
| 81 | Female | 46-59 | 157 | 127 | 30 |
| 82 | Female | 46-59 | 170 | 150 | 20 |
| 83 | Female | 46-59 | 150 | 138 | 12 |
| 84 | Female | 46-59 | 150 | 147 | 3 |
| 85 | Female | 46-59 | 167 | 157 | 10 |
| 86 | Female | 46-59 | 154 | 146 | 8 |
| 87 | Female | 46-59 | 143 | 148 | -5 |
| 88 | Female | 46-59 | 157 | 136 | 21 |
| 89 | Female | 46-59 | 149 | 146 | 3 |
| 90 | Female | 46-59 | 161 | 132 | 29 |
| 91 | Female | 46-59 | 142 | 145 | -3 |
| 92 | Female | 46-59 | 162 | 132 | 30 |
| 93 | Female | 46-59 | 144 | 157 | -13 |
| 94 | Female | 46-59 | 142 | 140 | 2 |
| 95 | Female | 46-59 | 159 | 137 | 22 |
| 96 | Female | 46-59 | 140 | 154 | -14 |
| 97 | Female | 46-59 | 144 | 169 | -25 |
| 98 | Female | 46-59 | 142 | 145 | -3 |
| 99 | Female | 46-59 | 145 | 137 | 8 |
| 100 | Female | 46-59 | 145 | 143 | 2 |
| 101 | Female | 60+ | 168 | 178 | -10 |
| 102 | Female | 60+ | 142 | 141 | 1 |
| 103 | Female | 60+ | 147 | 149 | -2 |
| 104 | Female | 60+ | 148 | 148 | 0 |
| 105 | Female | 60+ | 162 | 138 | 24 |
| 106 | Female | 60+ | 170 | 143 | 27 |
| 107 | Female | 60+ | 173 | 167 | 6 |
| 108 | Female | 60+ | 151 | 158 | -7 |
| 109 | Female | 60+ | 155 | 152 | 3 |
| 110 | Female | 60+ | 163 | 154 | 9 |
| 111 | Female | 60+ | 183 | 161 | 22 |
| 112 | Female | 60+ | 159 | 143 | 16 |
| 113 | Female | 60+ | 148 | 159 | -11 |
| 114 | Female | 60+ | 151 | 177 | -26 |
| 115 | Female | 60+ | 165 | 142 | 23 |
| 116 | Female | 60+ | 152 | 152 | 0 |
| 117 | Female | 60+ | 161 | 152 | 9 |
| 118 | Female | 60+ | 165 | 174 | -9 |
| 119 | Female | 60+ | 149 | 151 | -2 |
| 120 | Female | 60+ | 185 | 163 | 22 |





A paired sample t -test was used to analyze the blood pressure before and after the intervention to test if the intervention had a significant affect on the blood pressure. The blood pressure before the intervention was higher (156.45 ± 11.39 units) compared to the blood pressure post intervention (151.36 ± 14.18 units); there was a statistically significant decrease in blood pressure (t(119)=3.34, p= 0.0011) of 5.09 units.

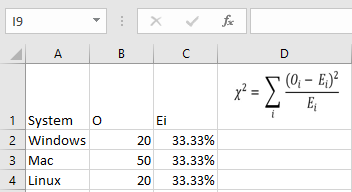


Practical 4:

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

Problem

An system administrator needs to upgrade the computers for his division. He wants to know what sort of computer system his workers prefer. He gives three choices: Windows, Mac, or Linux. Test the hypothesis or theory that an equal percentage of the population prefers each type of computer system



H0 : The population distribution of the variable is the same as the proposed distribution

HA : The distributions are different

To calculate the Chi –Squred value for Windows go to cell D2 and type =((B2-C2)\*(B2-C2))/C2

To calculate the Chi –Squred value for Mac go to cell D3 and type =((B3-C3)\*(B3-C3))/C3

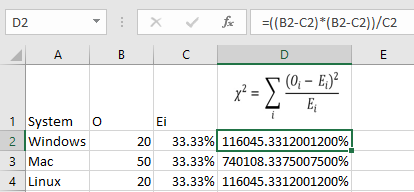
To calculate the Chi –Squred value for Mac go to cell D3 and type =((B4-C4)\*(B4-C4))/C4

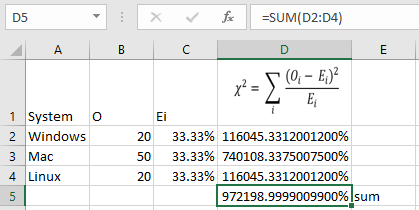
Go to Cell D5 for and type=SUM(D2:D4)

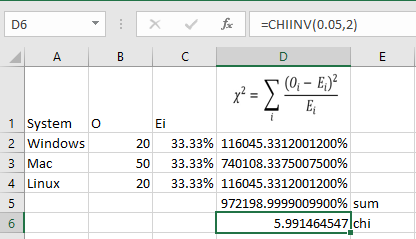
To get the table value for Chi-Square for α = 0.05 and dof = 2, go to cell D7 and type

=CHIINV(0.05,2)

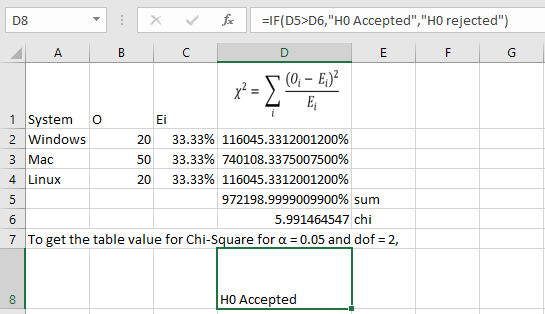
At cell D8 type =IF(D5>D7, "H0 Accepted","H0 Rejected")







OUTPUT:



1. Perform testing of hypothesis using chi-squared Test of Independence

In a study to understand the performance of M. Sc. IT Part -1 class, a college selects a random sample of 100 students. Each student was asked his grade obtained in B. Sc. IT. The sample is as given below

Sr. Roll No Student's Name Gen Grade

No

1 1 Gaborone m O

2 2 Francistown m O

3 5 Niamey m O

4 13 Maxixe m O

5 16 Tema m O

6 17 Kumasi m O

7 34 Blida m O

8 35 Oran m O

9 38 Saefda m O

10 42 Constantine m O

11 43 Annaba m O

12 45 Bejaefa m O

13 48 Medea m O

14 49 Djelfa m O

15 50 Tipaza m O

16 51 Bechar m O

17 54 Mostaganem m O

18 55 Tiaret m O

19 56 Bouira m O

20 59 Tebessa m O

21 61 El Harrach m O

22 62 Mila m O

23 65 Fouka m O

24 66 El Eulma m O

25 68 SidiBel Abbes m O

26 69 Jijel m O

27 70 Guelma m O

28 85 Khemis El Khechna m O

29 87 Bordj El Kiffan m O

30 88 Lakhdaria m O

31 6 Maputo m D

32 12 Lichinga m D

33 15 Ressano Garcia m D

34 19 Accra m D

35 27 Wa m D

36 28 Navrongo m D

37 37 Mascara m D

38 44 Batna m D

39 57 El Biar m D

40 60 Boufarik m D

41 63 OuedRhiou m D

42 64 Souk Ahras m D

43 71 Dar El Befda m D

44 86 Birtouta m D

45 18 Takoradi m C

46 22 Cape Coast m C

47 29 Kwabeng m C

48 30 Algiers m C

49 31 Laghouat m C

50 39 Relizane m C

51 52 Setif m C

52 53 Biskra m C

53 67 Kolea m C

54 100 AefnFakroun m C

55 26 Nima m B

56 32 TiziOuzou m B

57 33 Chlef m B

Sr. No Roll No Student's Name Gen Grade

62 3 Maun f O

63 7 Tete f O

64 9 Chimoio f O

65 11 Pemba f O

66 14 Chibuto f O

67 25 Mampong f O

68 36 Tlemcen f O

69 40 Adrar f O

70 41 Tindouf f O

71 46 Skikda f O

72 47 Ouargla f O

73 10 Matola f D

74 20 Legon f D

75 21 Sunyani f D

76 72 Teenas f D

77 73 Kouba f D

78 75 HussenDey f D

79 77 Khenchela f D

80 82 HassiBahbah f D

81 84 Baraki f D

82 91 Boudouaou f D

83 95 Tadjenanet f D

84 4 Molepolole f C

85 8 Quelimane f C

86 23 Bolgatanga f C

87 58 Mohammadia f C

88 83 Merouana f C

89 24 Ashaiman f B

90 76 N'gaous f B

91 90 Bab El Oued f B

92 92 BordjMenael f B

93 93 Ksar El Boukhari f B

94 74 Reghaa f A

95 78 Cheria f A

96 79 Mouzaa f A

97 80 Meskiana f A

98 81 Miliana f A

99 94 Sig f A

100 99 Kadiria f A

**Null Hypothesis - H0 :** The performance of girls students is same as boys students.

**Alternate Hypothesis - H1 :** The performance of boys and girls students are different.Open Excel Workbook

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **O** | **A** | **B** | **C** | **D** | **Total** |  |
| **Girls** | 11 | 7 | 5 | 5 | 11 | **39** | 6.075 |
| **Boys** | 30 | 4 | 3 | 10 | 14 | **61** | 6.075 |
| **Total** | 41 | 11 | 8 | 15 | 25 | **100** | **12.150** |
| **Ei** | **20.5** | **5.5** | **4** | **7.5** | **12.5** | **50** |  |

Prepare a contingency table as shown above.

To calculate Girls Students with ‘O’ Grade

Go to Cell N6 and type =COUNTIF($J$2:$K$40,"O")

To calculate Girls Students with ‘A’ Grade

Go to Cell O6 and type =COUNTIF($J$2:$K$40,"A")

To calculate Girls Students with ‘B’ Grade

Go to Cell P6 and type =COUNTIF($J$2:$K$40,"B")

To calculate Girls Students with ‘C’ Grade

Go to Cell Q6 and type =COUNTIF($J$2:$K$40,"C")

To calculate Girls Students with ‘D’ Grade

Go to Cell R6 and type =COUNTIF($J$2:$K$40,"D")

To calculate Boys Students with ‘O’ Grade

Go to Cell N7 and type =COUNTIF($D$2:$E$62,"O")

To calculate Boys Students with ‘A’ Grade

Go to Cell O7 and type =COUNTIF($D$2:$E$62,"A")

To calculate Boys Students with ‘B’ Grade

Go to Cell P7 and type =COUNTIF($D$2:$E$62,"B")

To calculate Boys Students with ‘C’ Grade

Go to Cell Q7 and type =COUNTIF($D$2:$E$62,"C")

To calculate Boys Students with ‘D’ Grade

Go to Cell R7 and type =COUNTIF($D$2:$E$62,"D")

**To calculated the expected value Ei**

Go to Cell N9 and type =N8/2

Go to Cell O9 and type =O8/2

Go to Cell P9 and type =P8/2

Go to Cell Q9 and type =Q8/2

Go to Cell R9 and type =R8/2

Go to Cell S6 and calculate total girl students = SUM(N6:R6)

Go to Cell S7 and calculate total girl students = SUM(N7:R7)

**Now Calculate**

Go to cell **T6** and type

=SUM((N6-$N$9)^2/$N$9,(O6-$O$9)^2/$O$9,(P6-$P$9)^2/$P$9,(Q6-Q$9)^2/$Q$9,

(R6-$R$9)^2/$R$9)

Go to cell **T7** and type

=SUM((N7-$N$9)^2/$N$9,(O7-$O$9)^2/$O$9,(P7-$P$9)^2/$P$9,(Q7-Q$9)^2/$Q$9,

(R7-$R$9)^2/$R$9)

To get the table value go to cell T11 and type **=CHIINV(0.05,4)**

Go to cell O13 and type =IF(T8>=T11," H0 is Accepted", "H0 is Rejected")

Practical 5:

1. Compute different types of correlation

**Positive Correlation**

import numpy as np

import matplotlib

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

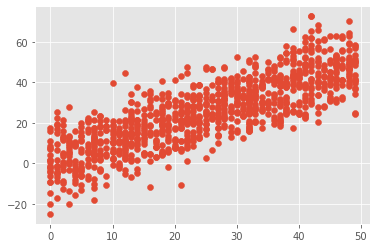
y = x + np.random.normal(0, 10, 1000)

np.corrcoef(x, y)

matplotlib.style.use('ggplot')

plt.scatter(x, y)

plt.show()



**Negative Correlation**

import numpy as np

import matplotlib

import matplotlib.pyplot as plt

np.random.seed(1)

# 1000 random integers between 0 and 50

x = np.random.randint(0, 50, 1000)

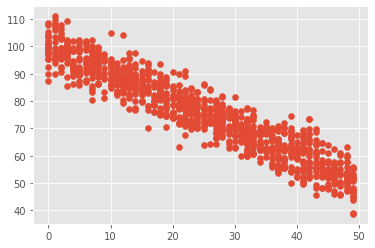
# Negative Correlation with some noise

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

np.corrcoef(x, y)

plt.scatter(x, y)

plt.show()



**No/Weak Correlation**

import numpy as np

import matplotlib.pyplot as plt

np.random.seed(1)

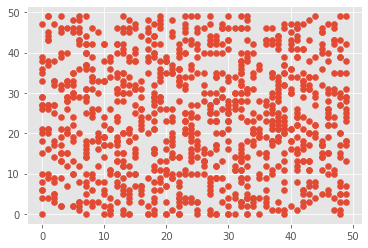
x = np.random.randint(0, 50, 1000)

y = np.random.randint(0, 50, 1000)

np.corrcoef(x, y)

plt.scatter(x, y)

plt.show()

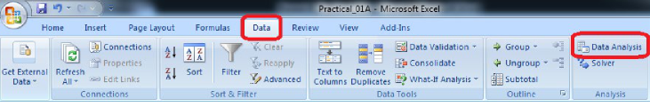


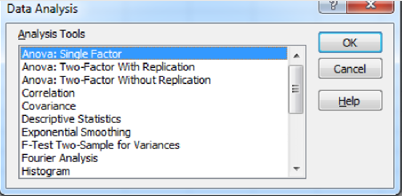
Practical 6:

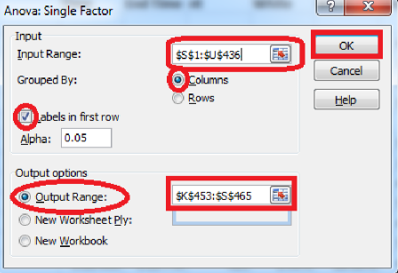
1. Perform testing of hypothesis using one-way ANOVA.

H0 - There are no significant differences between the Subject’s mean SAT scores. µ1 = µ2 = µ3 = µ4 = µ5

H1 - There is a significant difference between the Subject's mean SAT scores. To perform ANOVA go to data 🡪Data Analysis







**Input Range** : $S$1:$U$436*( Select columns to be analyzed in group)*

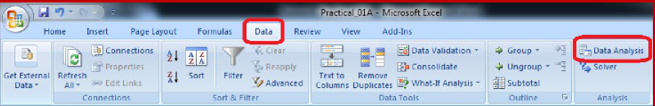
**Output Range** :$K$453:$S$465*( Can be any Range)*

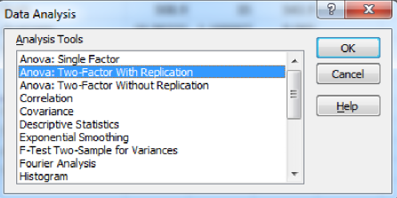
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| Average Score (SAT Math) | 375 | 162354 | 432.944 | 5177.144 |  |  |
| Average Score (SAT Reading) | 375 | 159189 | 424.504 | 3829.267 |  |  |
| Average Score (SAT Writing) | 375 | 156922 | 418.4587 | 4166.522 |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 39700.57 | 2 | 19850.28 | 4.520698 | 0.01108 | 3.003745 |
| Within Groups | 4926677 | 1122 | 4390.977 |  |  |  |
| Total | 4966377 | 1124 |  |  |  |  |

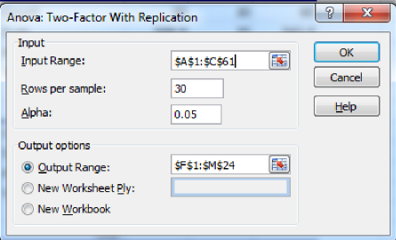
Since the resulting pvalue is less than 0.05. The null hypothesis (H0) is rejected and conclude that there is a significant difference between the SAT scores for each subject.

1. Perform testing of hypothesis using two-way ANOVA.

Go to Data tab **🡪** Data Analysis







Input Range - $A$1:$C$61

Rows Per Sample – 30 (Beacause 30 Patients are given each dose)

Alpha – 0.05

Output Range - $F$1:$M$24

**Output:**

**Anova: Two-Factor With Replication**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY | len | dose | Total |  |  |  |
|  | *1* |  |  |  |  |  |
| Count | 30 | 30 | 60 |  |  |  |
| Sum | 508.9 | 35 | 543.9 |  |  |  |
| Average | 16.96333 | 1.166667 | 9.065 |  |  |  |
| Variance | 68.32723 | 0.402299 | 97.22333 |  |  |  |
|  | *31* |  |  |  |  |  |
| 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* |  |  |  |  |  |  |
| *Variation* | *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 |
| Interaction | 102.675 | 1 | 102.675 | 3.642079 | 0.058808 | 3.922879 |
| Within | 3270.193 | 116 | 28.19132 |  |  |  |
| Total | 12817.69 | 119 |  |  |  |  |

P-value = 0.0588079 column in the ANOVA Source of Variation table at the bottom of the output. Because the p -values for both medicine dose and interaction are less than our significance level, these factors are statistically significant. On the other hand, the interaction effect is not significant because its p-value (0.0588) is greater than our significance level. Because the interaction effect is not significant, we can focus on only the main effects and not consider the interaction effect of the dose.

Practical 7:

1. Perform linear regression for prediction

Using R tools

> x <-c(151,174,138,186,128,136,179,163,152,131)

> y <-c(63,81,56,91,47,57,76,72,62,48)

> relation <-lm(y~x)

> print(relation)

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

-38.4551 0.6746

> print(summary(relation))

Call:

lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-6.3002 -1.6629 0.0412 1.8944 3.9775

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -38.45509 8.04901 -4.778 0.00139 \*\*

x 0.67461 0.05191 12.997 1.16e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.253 on 8 degrees of freedom

Multiple R-squared: 0.9548, Adjusted R-squared: 0.9491

F-statistic: 168.9 on 1 and 8 DF, p-value: 1.164e-06

> a <-data.frame(x=170)

> result <- predict(relation,a)

> print(result)

1

76.22869

> png(file= "linearregression.png")

> plot(y,x,col="blue",main= "Height & Weight Regression",abline(lm(x~y)0,cex=1.3,pch=16,xlab="WEIGHT in Kg",ylab="HEIGHT in cm")

Error: unexpected numeric constant in "plot(y,x,col="blue",main= "Height & Weight Regression",abline(lm(x~y)0"

> plot(y,x,col="blue",main= "Height & Weight Regression",abline(lm(x~y)0,cex=1.3,pch=16,xlab="WEIGHT in Kg",ylab="HEIGHT in cm")

Error: unexpected numeric constant in "plot(y,x,col="blue",main= "Height & Weight Regression",abline(lm(x~y)0"

> plot(y,x,col="blue",main= "Height & Weight Regression",abline(lm(x~y)),cex=1.3,pch=16,xlab="WEIGHT in Kg",ylab="HEIGHT in cm")

> dev.off()

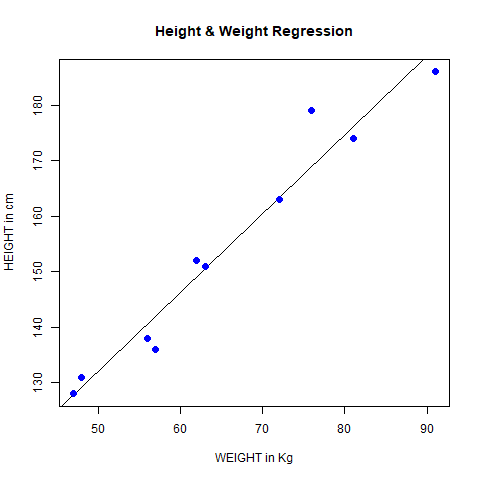
null device

1

> plot(y,x,col="blue",main= "Height & Weight Regression",abline(lm(x~y)),cex=1.3,pch=16,xlab="WEIGHT in Kg",ylab="HEIGHT in cm")

> plot(y,x,col="blue",main= "Height & Weight Regression",abline(lm(x~y)),cex=1.3,pch=16,xlab="WEIGHT in Kg",ylab="HEIGHT in cm")

>



Practical 8:

1. Perform Logistic regression

Using RTools

> quality <- read.csv('C:/Users/Gauri/Downloads/quality.csv')

> str(quality)

'data.frame': 131 obs. of 14 variables:

$ MemberID : int 1 2 3 4 5 6 7 8 9 10 ...

$ InpatientDays : int 0 1 0 0 8 2 16 2 2 4 ...

$ ERVisits : int 0 1 0 1 2 0 1 0 1 2 ...

$ OfficeVisits : int 18 6 5 19 19 9 8 8 4 0 ...

$ Narcotics : int 1 1 3 0 3 2 1 0 3 2 ...

$ DaysSinceLastERVisit: num 731 411 731 158 449 ...

$ Pain : int 10 0 10 34 10 6 4 5 5 2 ...

$ TotalVisits : int 18 8 5 20 29 11 25 10 7 6 ...

$ ProviderCount : int 21 27 16 14 24 40 19 11 28 21 ...

$ MedicalClaims : int 93 19 27 59 51 53 40 28 20 17 ...

$ ClaimLines : int 222 115 148 242 204 156 261 87 98 66 ...

$ StartedOnCombination: logi FALSE FALSE FALSE FALSE FALSE FALSE ...

$ AcuteDrugGapSmall : int 0 1 5 0 0 4 0 0 0 0 ...

$ PoorCare : int 0 0 0 0 0 1 0 0 1 0 ...

> table(quality$PoorCare)

0 1

98 33

> 98/131

[1] 0.7480916

> install.packages("caTools")

Installing package into ‘C:/Users/Gauri/Documents/R/win-library/4.1’

(as ‘lib’ is unspecified)

--- Please select a CRAN mirror for use in this session --- **(Select Canada NS / Canada US)**

also installing the dependency ‘bitops’

trying URL 'https://cran.utstat.utoronto.ca/bin/windows/contrib/4.1/bitops\_1.0-7.zip'

Content type 'application/zip' length 42557 bytes (41 KB)

downloaded 41 KB

trying URL 'https://cran.utstat.utoronto.ca/bin/windows/contrib/4.1/caTools\_1.18.2.zip'

Content type 'application/zip' length 316382 bytes (308 KB)

downloaded 308 KB

package ‘bitops’ successfully unpacked and MD5 sums checked

package ‘caTools’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Gauri\AppData\Local\Temp\RtmpQjbOOR\downloaded\_packages

> library(caTools)

Warning message:

package ‘caTools’ was built under R version 4.1.3

> set.seed(88)

> split = sample.split(quality$PoorCare, SplitRatio = 0.75)

> split

[1] TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE TRUE FALSE

[13] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

[25] FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE

[37] FALSE TRUE TRUE TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE TRUE

[49] FALSE TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

[61] TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE FALSE TRUE

[73] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE

[85] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE

[97] TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE

[109] TRUE FALSE FALSE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE

[121] FALSE TRUE TRUE FALSE TRUE TRUE FALSE TRUE TRUE TRUE FALSE

> qualityTrain = subset(quality, split == TRUE)

> qualityTest = subset(quality, split == FALSE)

> nrow(qualityTrain)

[1] 99

> nrow(qualityTest)

[1] 32

> QualityLog = glm(PoorCare ~ OfficeVisits + Narcotics,data=qualityTrain, family=binomial)

> summary(QualityLog)

Call:

glm(formula = PoorCare ~ OfficeVisits + Narcotics, family = binomial,

data = qualityTrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.06303 -0.63155 -0.50503 -0.09689 2.16686

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.64613 0.52357 -5.054 4.33e-07 \*\*\*

OfficeVisits 0.08212 0.03055 2.688 0.00718 \*\*

Narcotics 0.07630 0.03205 2.381 0.01728 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 111.888 on 98 degrees of freedom

Residual deviance: 89.127 on 96 degrees of freedom

AIC: 95.127

Number of Fisher Scoring iterations: 4

> predictTrain = predict(QualityLog, type="response")

> summary(predictTrain)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.06623 0.11912 0.15967 0.25253 0.26765 0.98456

> tapply(predictTrain, qualityTrain$PoorCare, mean)

0 1

0.1894512 0.4392246

> table(qualityTrain$PoorCare, predictTrain > 0.5)

FALSE TRUE

0 70 4

1 15 10

> 10/25

[1] 0.4

> 70/74

[1] 0.9459459

> table(qualityTrain$PoorCare, predictTrain > 0.2)

FALSE TRUE

0 54 20

1 9 16

> 16/25

[1] 0.64

> 54/74

[1] 0.7297297

> install.packages("ROCR")

Installing package into ‘C:/Users/Gauri/Documents/R/win-library/4.1’

(as ‘lib’ is unspecified)

also installing the dependencies ‘gtools’, ‘gplots’

trying URL 'https://cran.utstat.utoronto.ca/bin/windows/contrib/4.1/gtools\_3.9.2.zip'

Content type 'application/zip' length 366977 bytes (358 KB)

downloaded 358 KB

trying URL 'https://cran.utstat.utoronto.ca/bin/windows/contrib/4.1/gplots\_3.1.1.zip'

Content type 'application/zip' length 603166 bytes (589 KB)

downloaded 589 KB

trying URL 'https://cran.utstat.utoronto.ca/bin/windows/contrib/4.1/ROCR\_1.0-11.zip'

Content type 'application/zip' length 454019 bytes (443 KB)

downloaded 443 KB

package ‘gtools’ successfully unpacked and MD5 sums checked

package ‘gplots’ successfully unpacked and MD5 sums checked

package ‘ROCR’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Gauri\AppData\Local\Temp\RtmpQjbOOR\downloaded\_packages

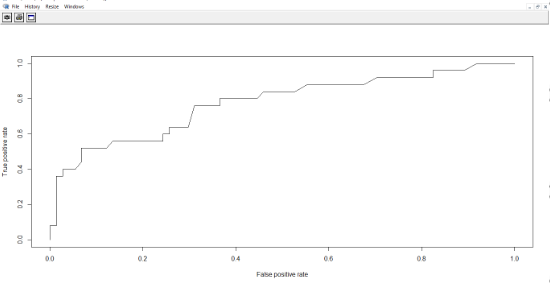
> library(ROCR)

Warning message:

package ‘ROCR’ was built under R version 4.1.3

> ROCRpred = prediction(predictTrain, qualityTrain$PoorCare)

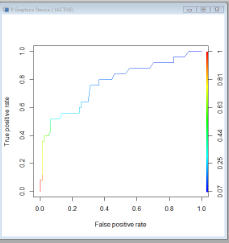
> ROCRperf = performance(ROCRpred, "tpr", "fpr")



> plot(ROCRperf)

> plot(ROCRperf, colorize=TRUE)

> plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))



Practical 9:

1. Perform testing of hypothesis using Z-test

from statsmodels.stats import weightstats as stests

import pandas as pd

from scipy import stats

df = pd.read\_csv("blood\_pressure.csv")

df[['bp\_before','bp\_after']].describe()

print(df)

ztest ,pval = stests.ztest(df['bp\_before'], x2=None, value=156)

print(float(pval))

if pval<0.05:

print("reject null hypothesis")

else:

print("accept null hypothesis")

OUTPUT:

