

**SCHOOL OF BASIC SCIENCES AND RESEARCH**

**Department of Mathematics**

**LAB REPORT FILE**

**Course Title: Statistics lab-3**

**Course Code: BDA 251**

**Program-Bachelor of Science (Hons.)-Data Science**

**Semester: 3rd**

**Session 2022-2023**

**Student Name: - Manish kumar**

**System ID: - 2021509877**

**Roll No. - 2107158014**

**Submitted to**

**Dr. Renu mam (Assistant Professor)**

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. N.** | **Date** | **Title** | **Signature** |
|  |  | Problem-based on the least square method  Problem-based on the simple linear regression  Problem-based on the multiple linear regression.  Problem based on the Paired samples t-test  Problem based on 2 samples t-test assuming equal and unequal variance.  Problem based on Z-test.  Problem based on F-test  Problem based on one-way ANOVA.  Problem based on Two-way ANOVA with replication.  Problem based on Two-way ANOVA without replication. |  |

**Practical- 1**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem-based on the least square method.

* **Theory:**

Principle of Least Square- The least square method is the process of obtaining the best-fitting curve or line of best fit for the given data set by reducing the sum of the squares of the offsets (residual part) of the points from the curve. Least-square method is the curve that best fits a set of observations with a minimum sum of squared residuals or errors.

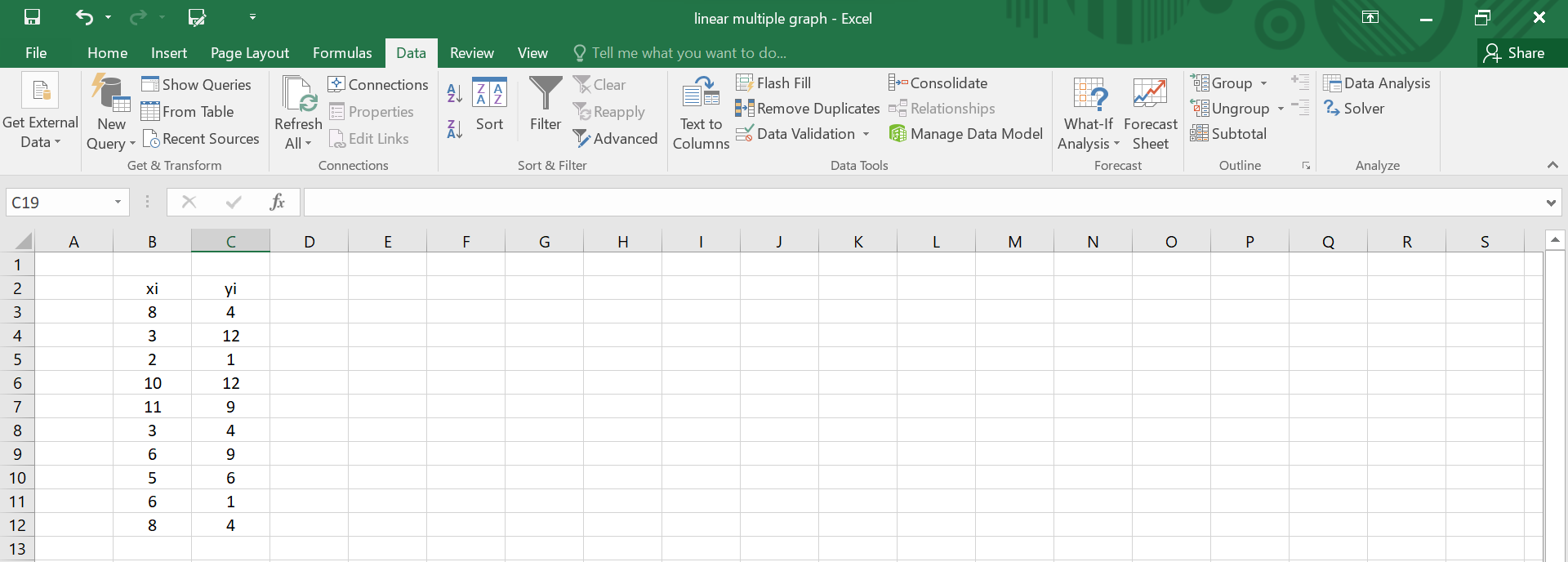
* **Problem:**

Consider the time series data given below:

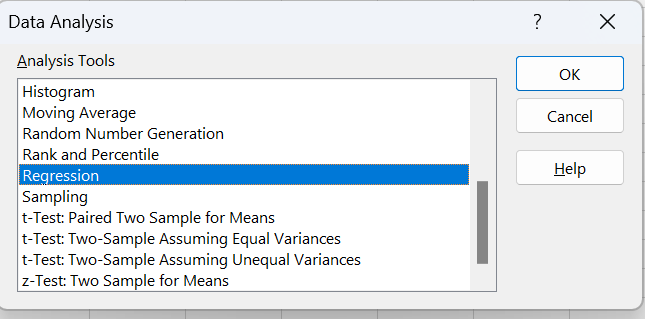
|  |  |
| --- | --- |
| xi | yi |
| 8 | 4 |
| 3 | 12 |
| 2 | 1 |
| 10 | 12 |
| 11 | 9 |
| 3 | 4 |
| 6 | 9 |
| 5 | 6 |
| 6 | 1 |
| 8 | 14 |

Determine the equation of line of best fit for data and then plot the line.

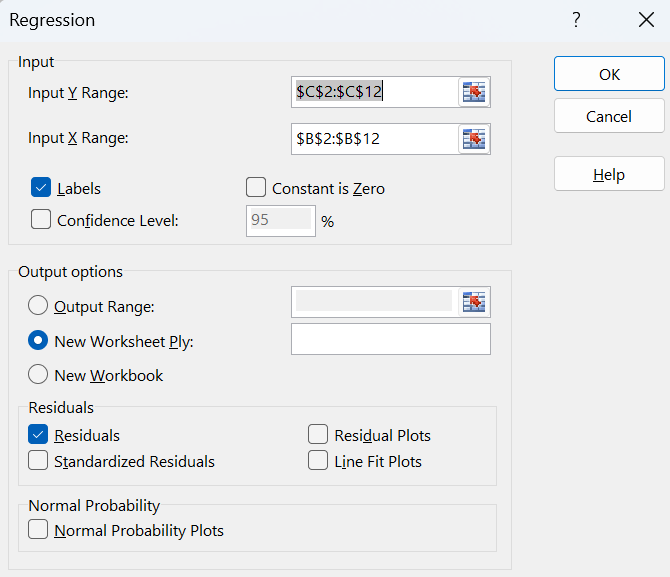
* **STEPS:**
* Click on Data Analysis on Data tab.



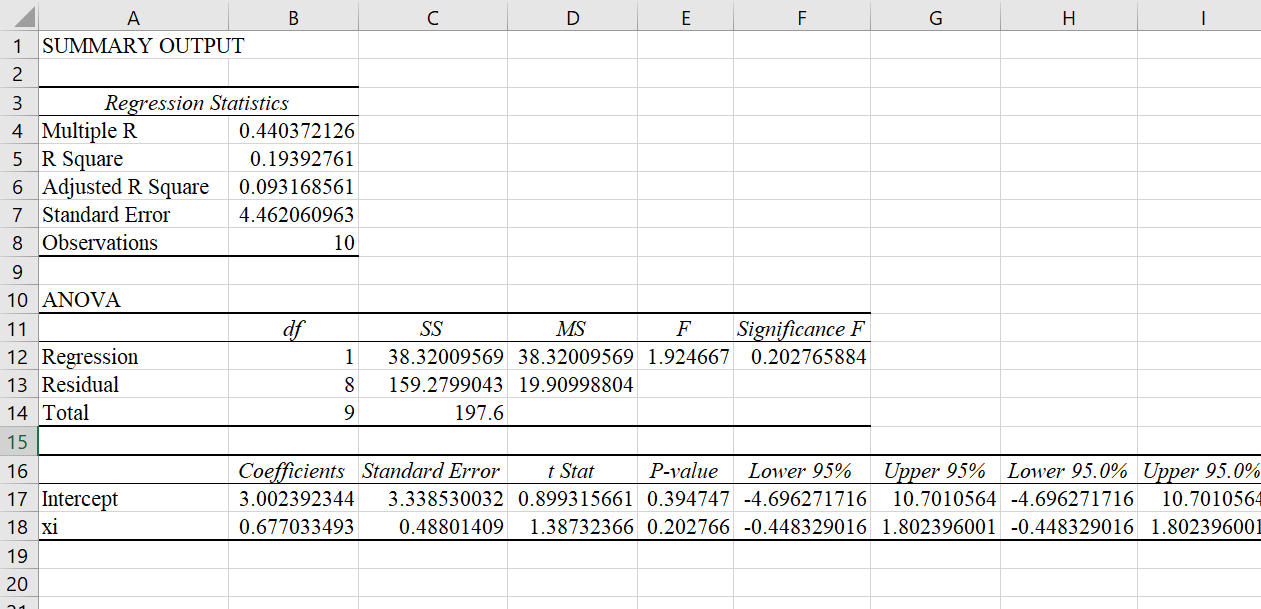
* Choose the regression:

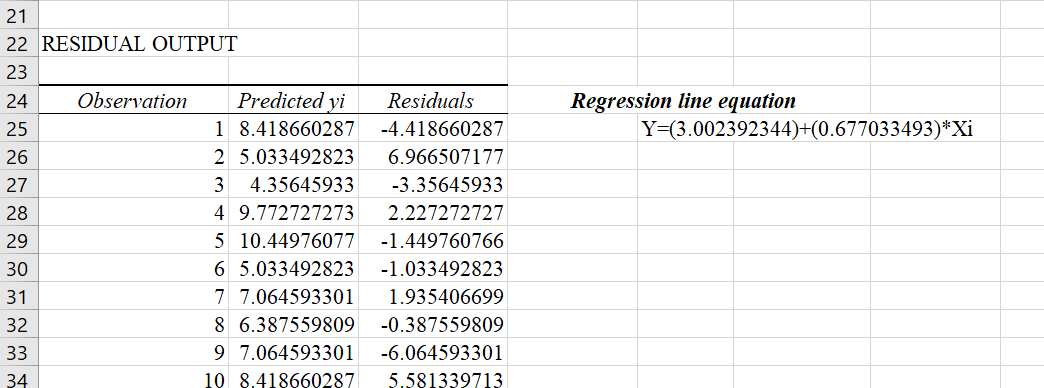


* Select the range of Dependent variable(Y) and Independent variable(X) and Click ok.



* GRAPH:
* OUTPUT:





* **OBSERVATION:**

Regression line equation is given by:

Y=a+b\*x

Y = (3.002392344) + (0.677033493) \* Xi

Where,

X=Independent variable

Y=Dependent variable

a =intercept = 3.002392344

b = slope = 0.677033492

**System id: 2021509877**

**Date:**

**Practical – 2**

* **Aim:**

Problem-based on the simple linear regression

* **Theory:**

**Simple Linear Regression**:

Simple linear regression is used to model the relationship between two continuous variables. Often, the objective is to predict the value of an output variable (or response) based on the value of an input (or predictor) variable.

We might also recognize the equation as the **slope formula.** The equation has the form

Y= a + bX, where Y is the dependent variable (that’s the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the [slope](https://tinyurl.com/y2ebznan) of the line and a is the [y-intercept](https://www.statisticshowto.com/y-intercept/).

* **Problem:**

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

What linear regression equation best predicts statistics performance, based on math aptitude scores?

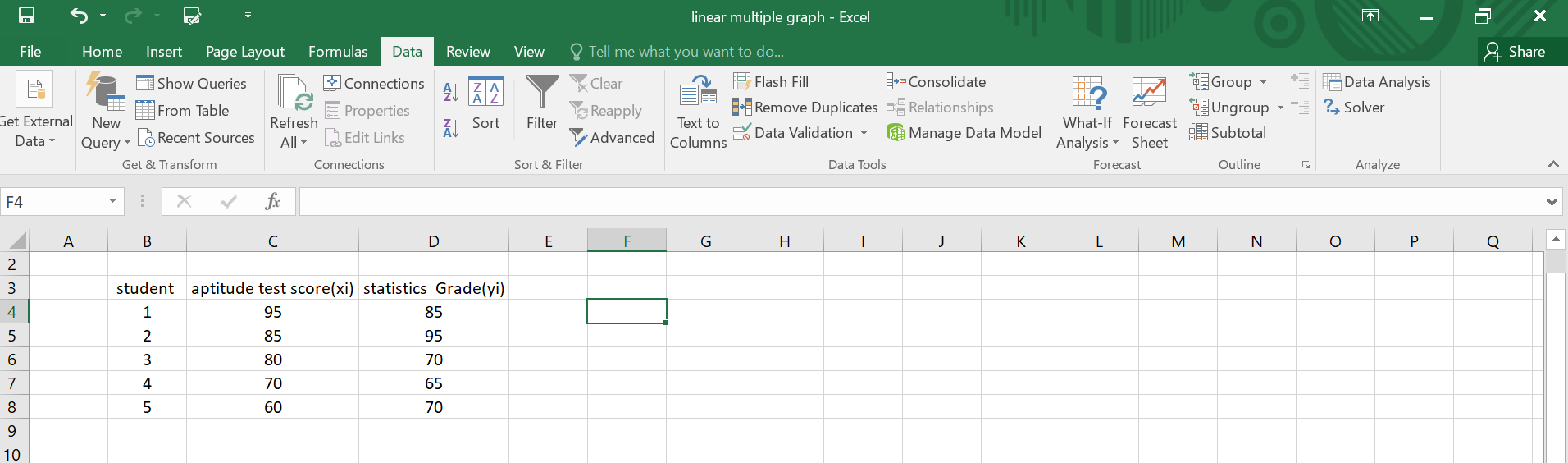
If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

How well does the regression equation fit the data?

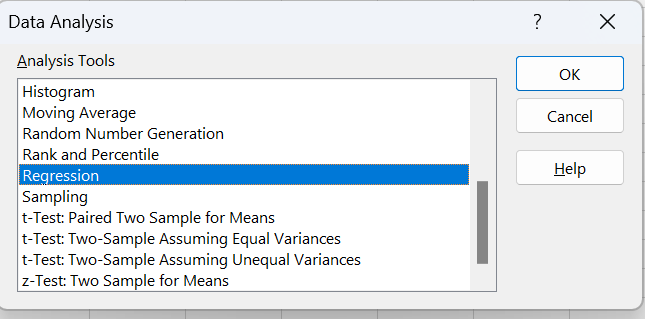
In the table below, the xi column shows scores on the aptitude test. Similarly, the yi column shows statistics grades

|  |  |  |
| --- | --- | --- |
|  |  |  |
| student | aptitude test score(xi) | statistics Grade(yi) |
| 1 | 95 | 85 |
| 2 | 85 | 95 |
| 3 | 80 | 70 |
| 4 | 70 | 65 |
| 5 | 60 | 70 |
|  |  |  |

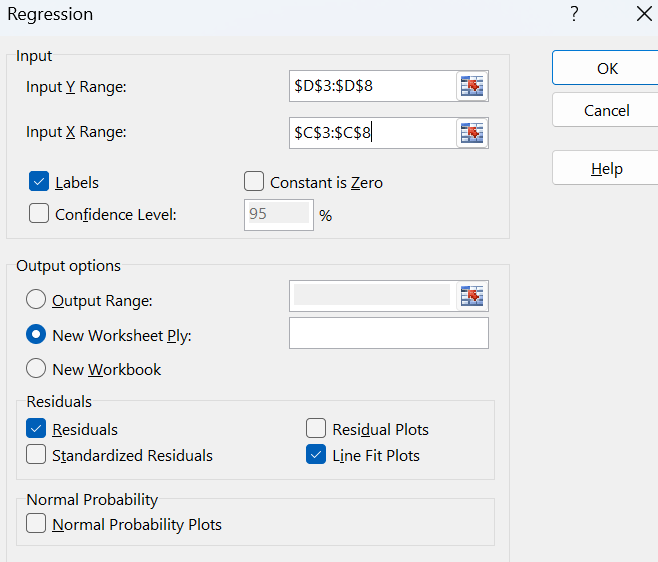
* **STEPS:**
* Click on Data Analysis on Data tab.



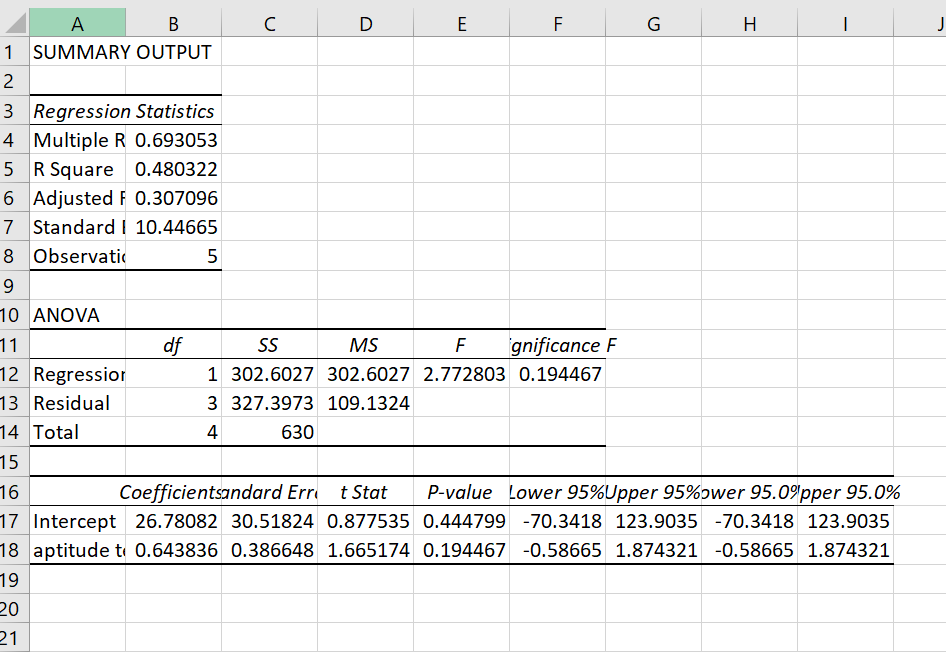
* Choose the regression:

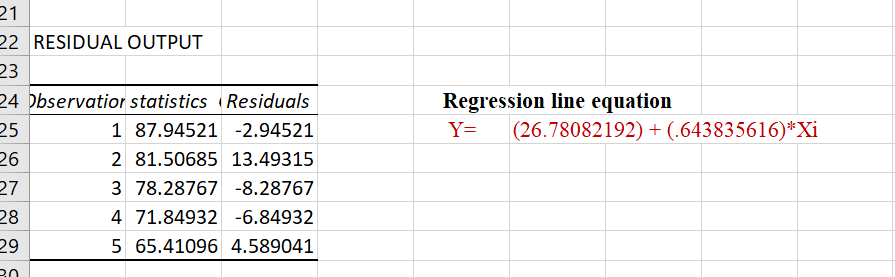


* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **Graph:**
* OUTPUT:





**OBSERVATION:**

* Regression line equation is given by:
* Y=a+b\*x
* Y = (26.78082192) + (0.643835616) \* Xi

Where,

* X=Independent variable
* Y=Dependent variable
* a =intercept = 26.78082192
* b = slope = 0.643835616

**Practical-3**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem-based on the multiple linear regression.

* **Theory:**

**Multiple Linear Regression-:**

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the [dependent variable](https://corporatefinanceinstitute.com/resources/knowledge/terms/dependent-variable/) are known as independent or explanatory variables.

* **Problem:**

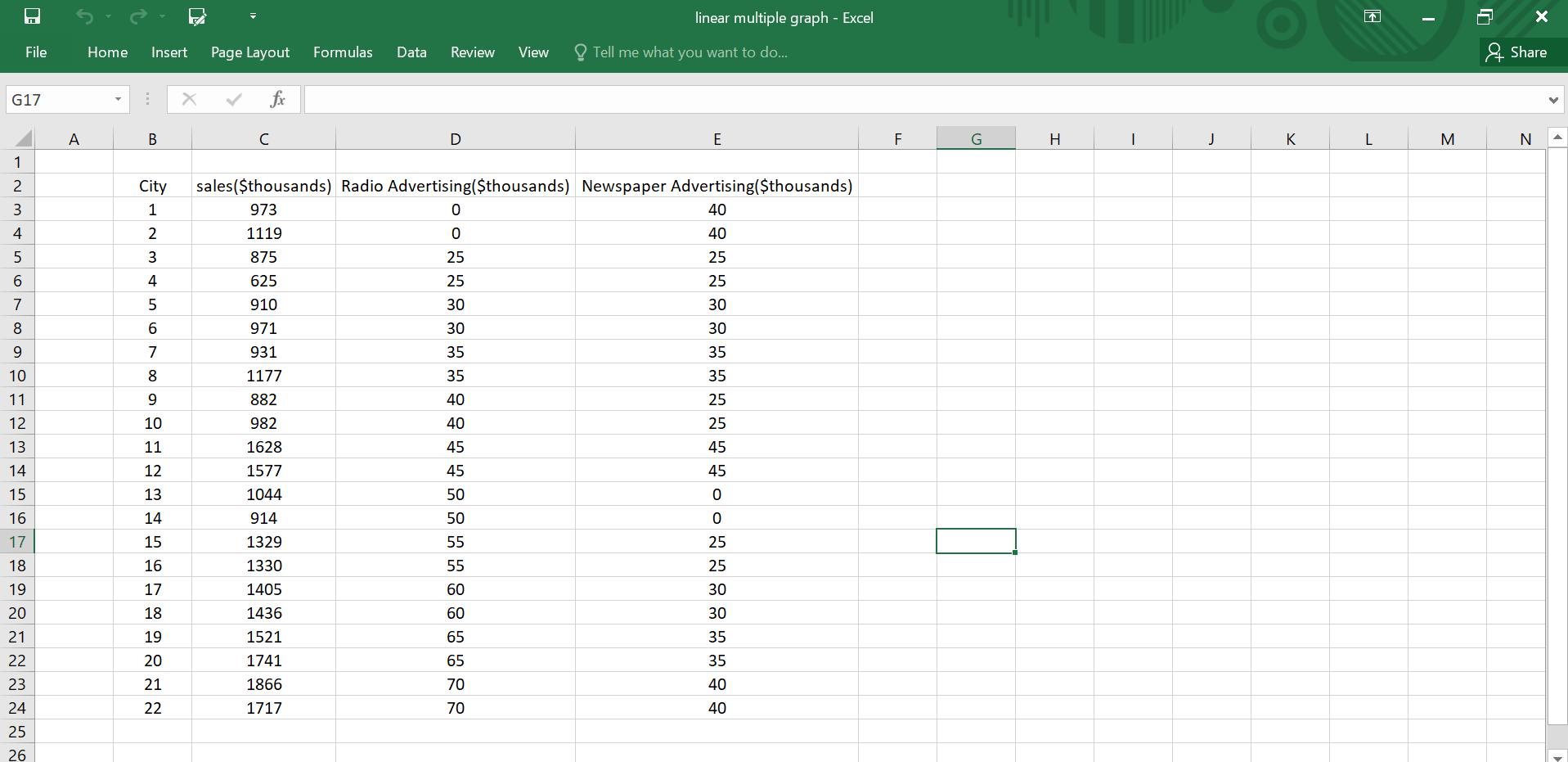
The business problem facing a consumer product company is to measure the effectiveness of different types of advertising media in the promotion of its products. Specifically, the company is interested in the effectiveness of radio advertising and newspaper advertising (including the cost of discount coupons).

During a one-month test period, data were collected from a sample of 22 cities with approximately equal population. Each city is allocated a specific expenditure level for radio advertising and for newspaper advertising. The sales of the product (in $thousand) and also the levels of media expenditure (in $thousand) during the test month are recorded, with the following result shown below and stored in advertise.

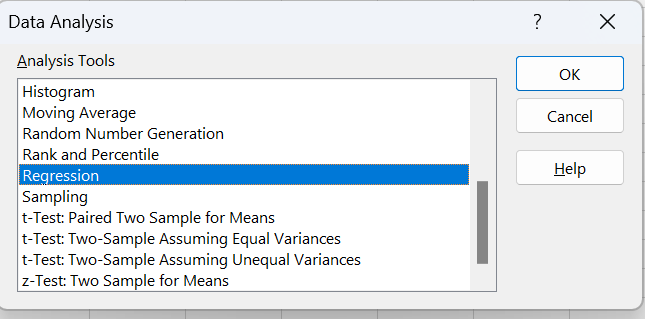
* **State the multiple regression equation**
* **Interpret the meaning of the slopes, b1 and b2, in this problem.**
* **Interpret the meaning of the regression coefficient, b0**

|  |  |  |  |
| --- | --- | --- | --- |
| City | sales($thousands) | Radio Advertising($thousands) | Newspaper Advertising($thousands) |
| 1 | 973 | 0 | 40 |
| 2 | 1119 | 0 | 40 |
| 3 | 875 | 25 | 25 |
| 4 | 625 | 25 | 25 |
| 5 | 910 | 30 | 30 |
| 6 | 971 | 30 | 30 |
| 7 | 931 | 35 | 35 |
| 8 | 1177 | 35 | 35 |
| 9 | 882 | 40 | 25 |
| 10 | 982 | 40 | 25 |
| 11 | 1628 | 45 | 45 |
| 12 | 1577 | 45 | 45 |
| 13 | 1044 | 50 | 0 |
| 14 | 914 | 50 | 0 |
| 15 | 1329 | 55 | 25 |
| 16 | 1330 | 55 | 25 |
| 17 | 1405 | 60 | 30 |
| 18 | 1436 | 60 | 30 |
| 19 | 1521 | 65 | 35 |
| 20 | 1741 | 65 | 35 |
| 21 | 1866 | 70 | 40 |
| 22 | 1717 | 70 | 40 |

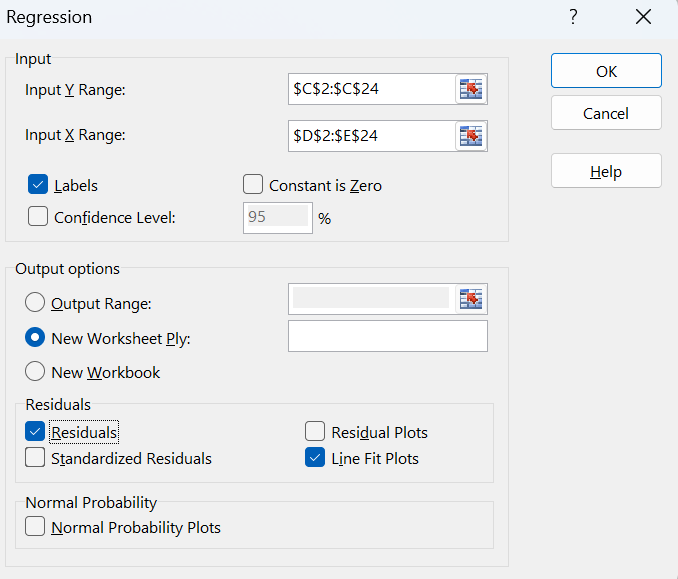
* **STEPS:**
* Click on Data Analysis on Data tab.



* Choose the regression:



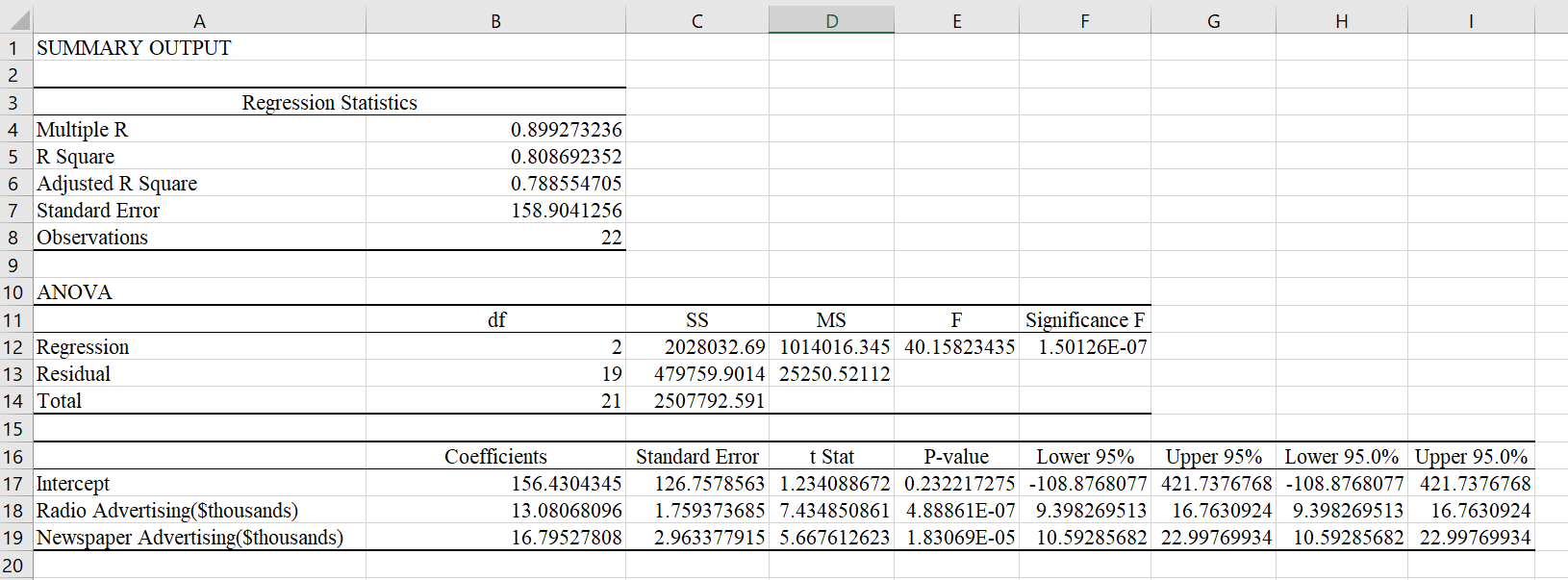
* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.

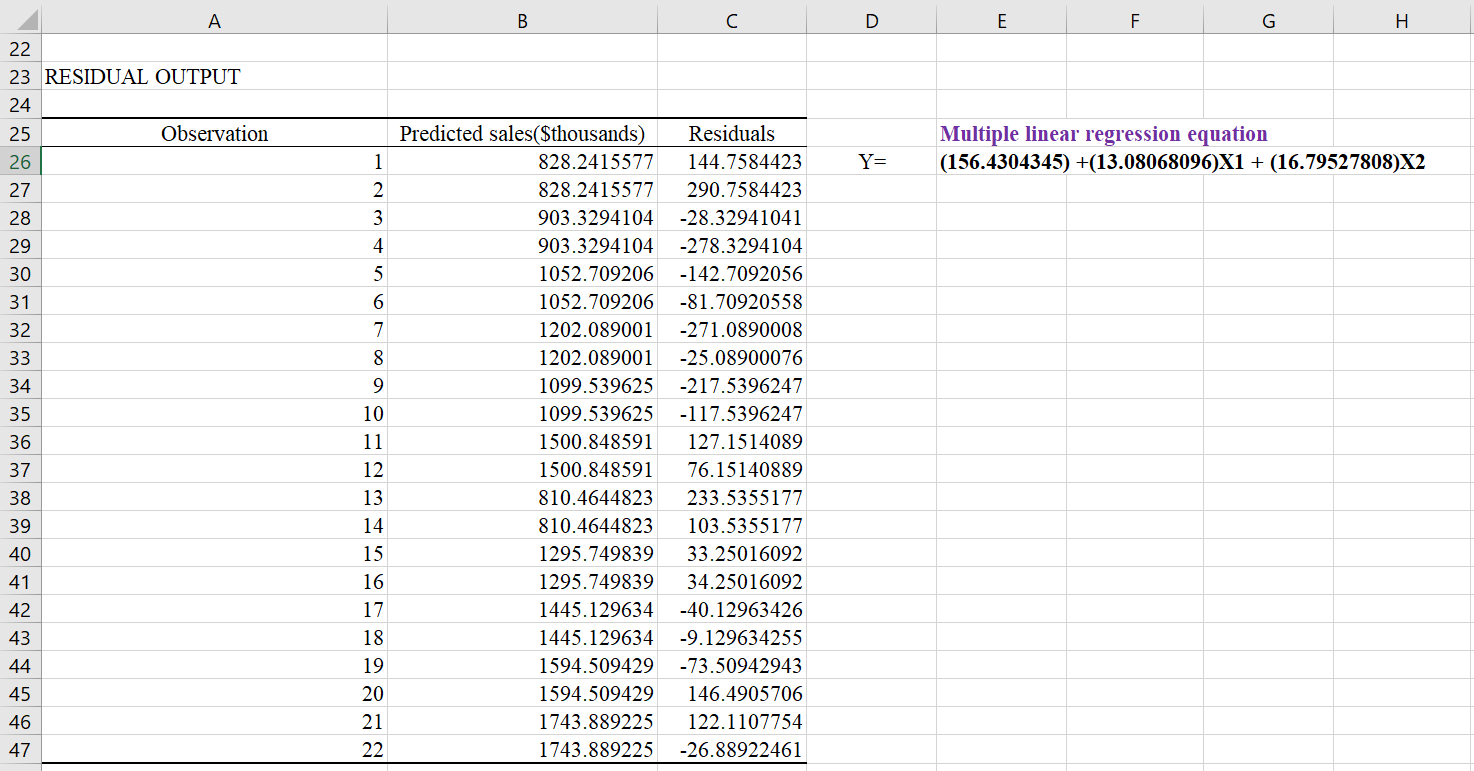


* **Graph:**

.

* Output:





* **OBSERVATION:**

Regression line equation is given by:

Y=b0 + b1\*x1 + b2\*x2 + …………. + bn\*xn

Y = (156.4304345) +(13.08068096) X1 + (16.79527808) X2

Where,

X1=Independent variable

X2= Independent variable

Y=Dependent variable

b0 =intercept = 156.4304345

b1 = slope = 13.08068096

b2 = slope = 16.79527808

* **State the multiple regression equation**

The multiple linear regression equation will be .

* **Interpret the meaning of the slopes, b1 and b2, in this problem.**

For a given amount of newspaper advertising, each increase of $1000 in radio advertising is estimated to result in a mean increase in sales of $13,081. For a given amount of radio advertising, each increase of $1000 in newspaper advertising is estimated to result in a mean increase in sales of $16,795.

* **Interpret the meaning of the regression coefficient, b0 .**

When there is no money spent on radio advertising and newspaper advertising, the estimated mean amount of sales is $156,430.44.

According to the above results, newspaper advertising is more effective as each increase of $1000 in newspaper advertising will result in a higher mean increase in sales than the same amount of increase in radio advertising.

**Practical-4**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem based on the Paired samples t-test.

* **Theory:**

**Paired t-test:**

Paired sample t-test, commonly known as dependent sample t-test is used to find out if the difference in the mean of two samples is 0. The test is done on dependent samples, usually focusing on a particular group of people or thing. In this, each entity is measured twice, resulting in a pair of observations.

**We can use this when:**

* Two similar (twin like) samples are given. [Eg, Scores obtained in English and Math (both subjects)]
* The dependent variable (data) is continuous.
* The observations are independent of one another.
* The dependent variable is approximately normally distributed.

**Null Hypothesis: H0: μd = 0**

**Alternative Hypothesis: H1: μd ≠ 0**

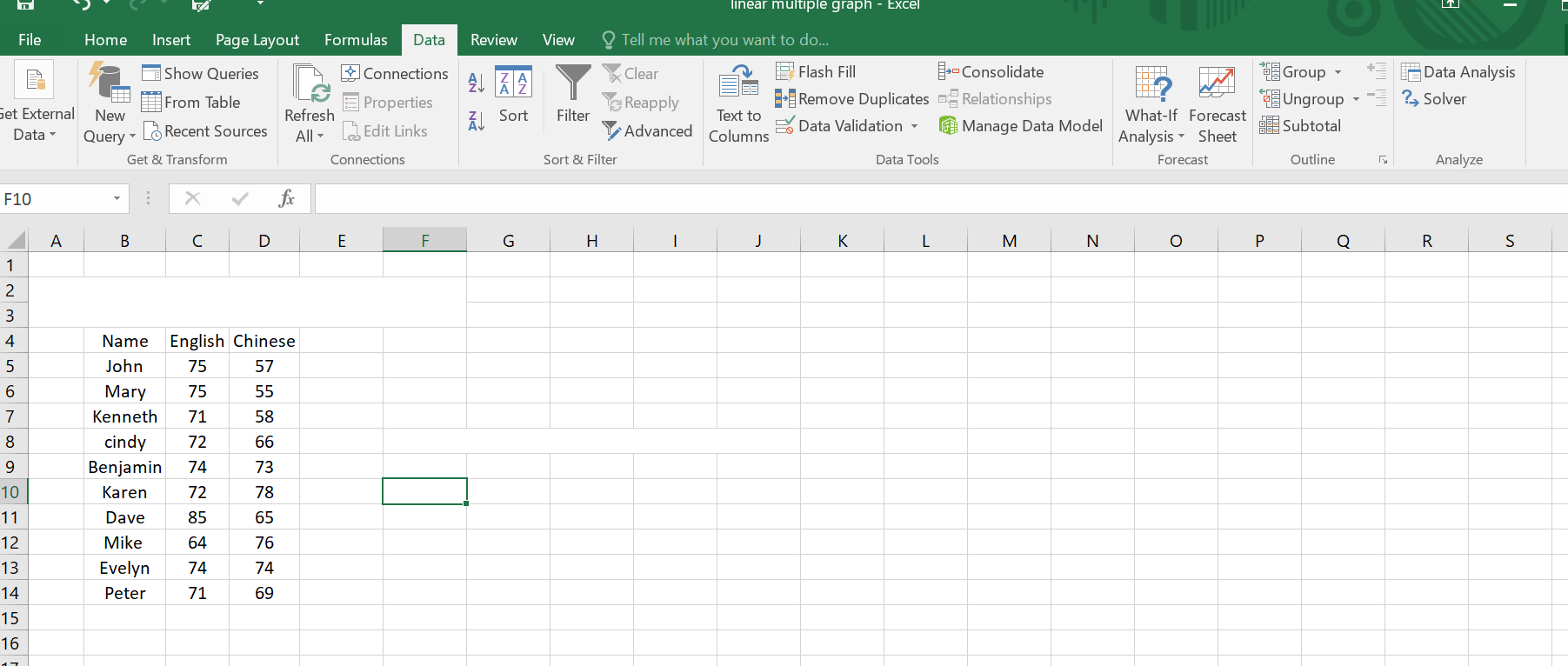
**Paired t-test formula: t =**

* **Problem:**

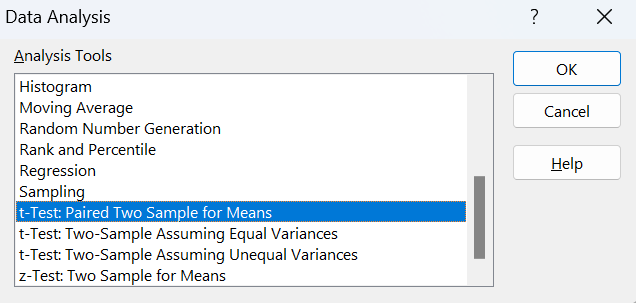
Problem based on the Paired samples t-test.

|  |  |  |
| --- | --- | --- |
| Name | English | Chinese |
| John | 75 | 57 |
| Mary | 75 | 55 |
| Kenneth | 71 | 58 |
| cindy | 72 | 66 |
| Benjamin | 74 | 73 |
| Karen | 72 | 78 |
| Dave | 85 | 65 |
| Mike | 64 | 76 |
| Evelyn | 74 | 74 |
| Peter | 71 | 69 |

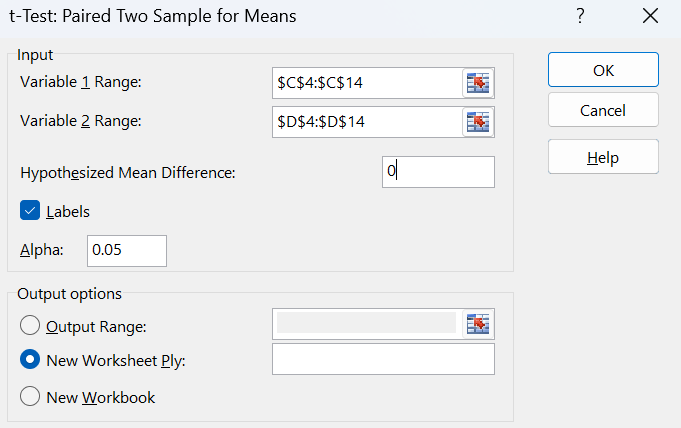
* **STEPS:**
* Click on Data Analysis on Data tab.



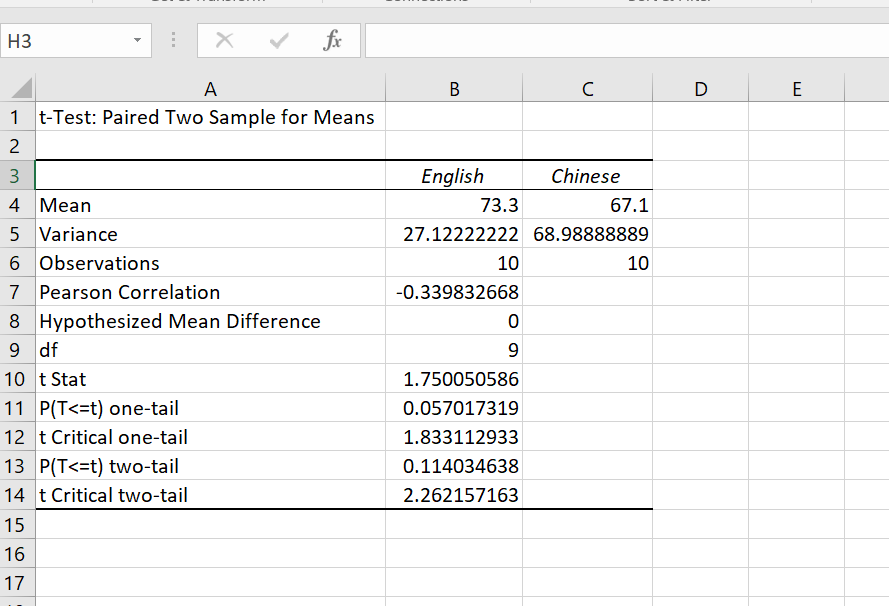
* Choose the regression:



* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **Output:**
* **Null hypothesis:** Average marks for english= Average marks for chinese

****

* **Observation:**

**Null hypothesis:** Average marks for english= Average marks for chinese

The means of test English is 73.3 and test Chinese mean is 67.1. The variance values are 27.12222222 and 68.98888889 respectively. The number of observations of both groups was 10. A p-value is a probability that the results from the sample dataset are occurred by chance. Low p-values are considered good. Excel provided us with both one-tail and two-tail T-Tests.

**Practical-5**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem based on 2 samples t-test assuming equal and unequal variance.

* **Theory:**

**T-TEST: TWO-SAMPLE ASSUMING EQUAL VARIANCES**

The t-Test Paired Two-Sample for Means tool performs a paired two-sample Student's t-Test to ascertain if the null hypothesis (means of two populations are equal) can be accepted or rejected. This test does not assume that the variances of both populations are equal.

**T-TEST: TWO-SAMPLE ASSUMING UNEQUAL VARIANCES**

This tool executes a two-sample student's t-Test on data sets from two independent populations with unequal variances.  This test can be either two-tailed or one-tailed contingent upon if we are testing that the two population means are different or if one is greater than the other.

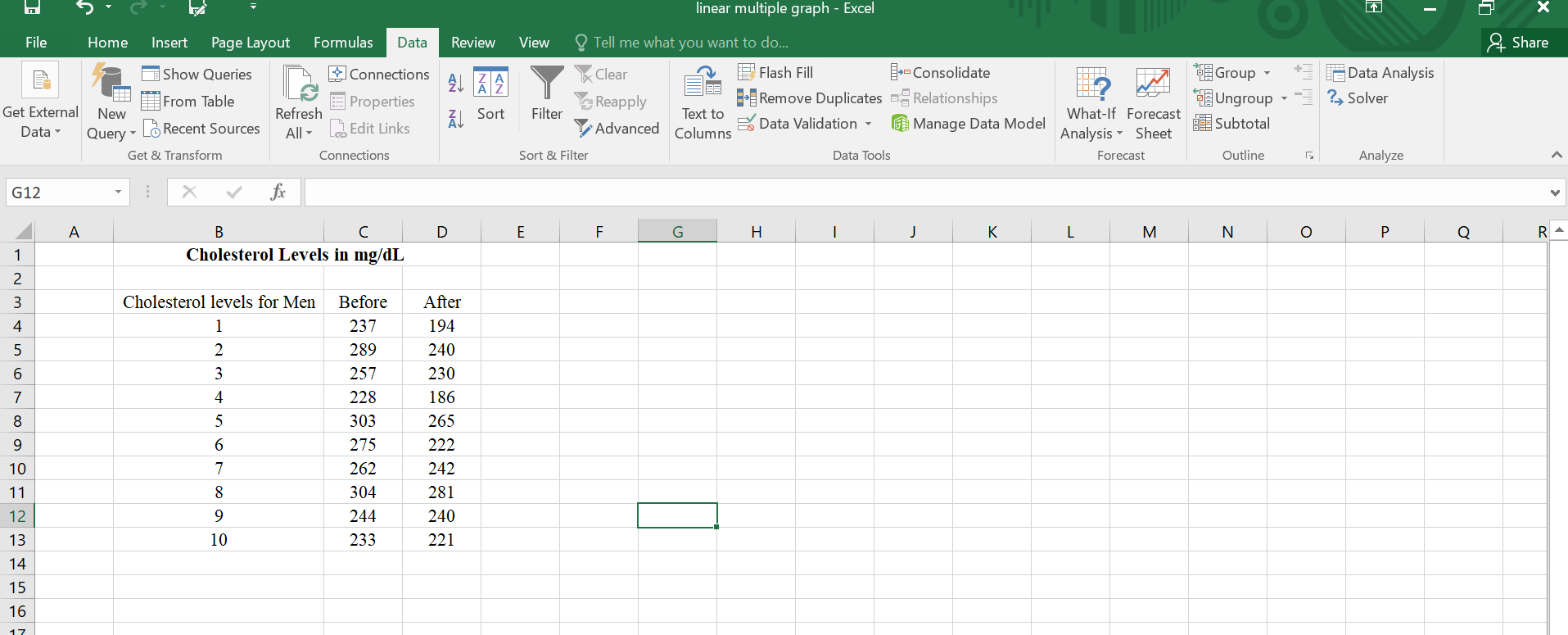
* **Problem:**

The sample data here is the cholesterol levels for 10 men diagnosed with high cholesterol. The first row represents levels from 10 men who did not use the drug and the second row from 10 different men who did use the drug.

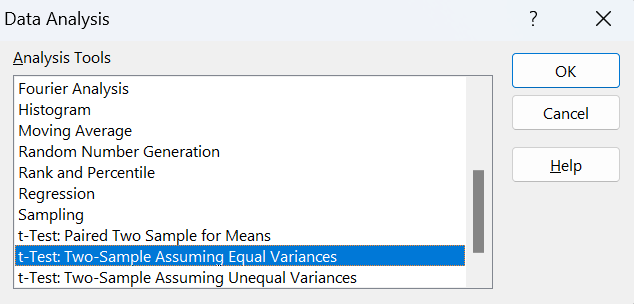
Test the claim that the mean level for all men who use the drug is less than the mean for all men who do not use it. Use a 0.01 significance level. You should enter the data into Excel as columns with headings in the first row.

|  |  |  |
| --- | --- | --- |
| Cholesterol Levels in mg/dL | | |
|  |  |  |
| Cholesterol levels for Men | Before | After |
| 1 | 237 | 194 |
| 2 | 289 | 240 |
| 3 | 257 | 230 |
| 4 | 228 | 186 |
| 5 | 303 | 265 |
| 6 | 275 | 222 |
| 7 | 262 | 242 |
| 8 | 304 | 281 |
| 9 | 244 | 240 |
| 10 | 233 | 221 |

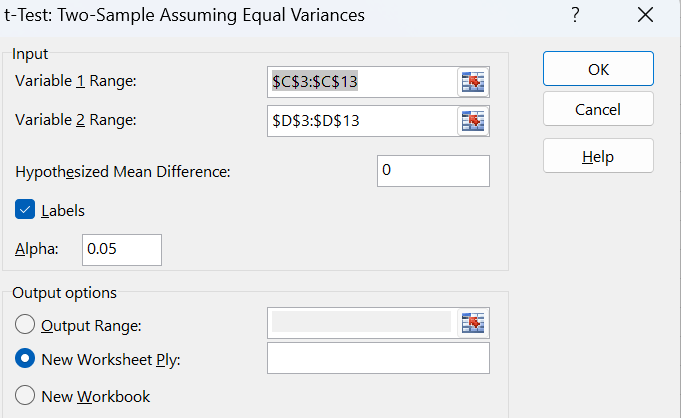
* **T-TEST: TWO-SAMPLE ASSUMING EQUAL VARIANCES:**
* **STEPS:**
* Click on Data Analysis on Data tab.



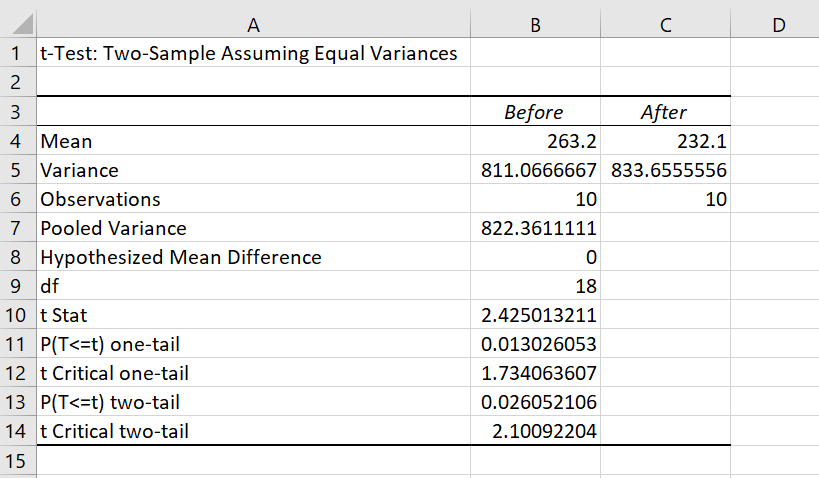
* Choose the regression:



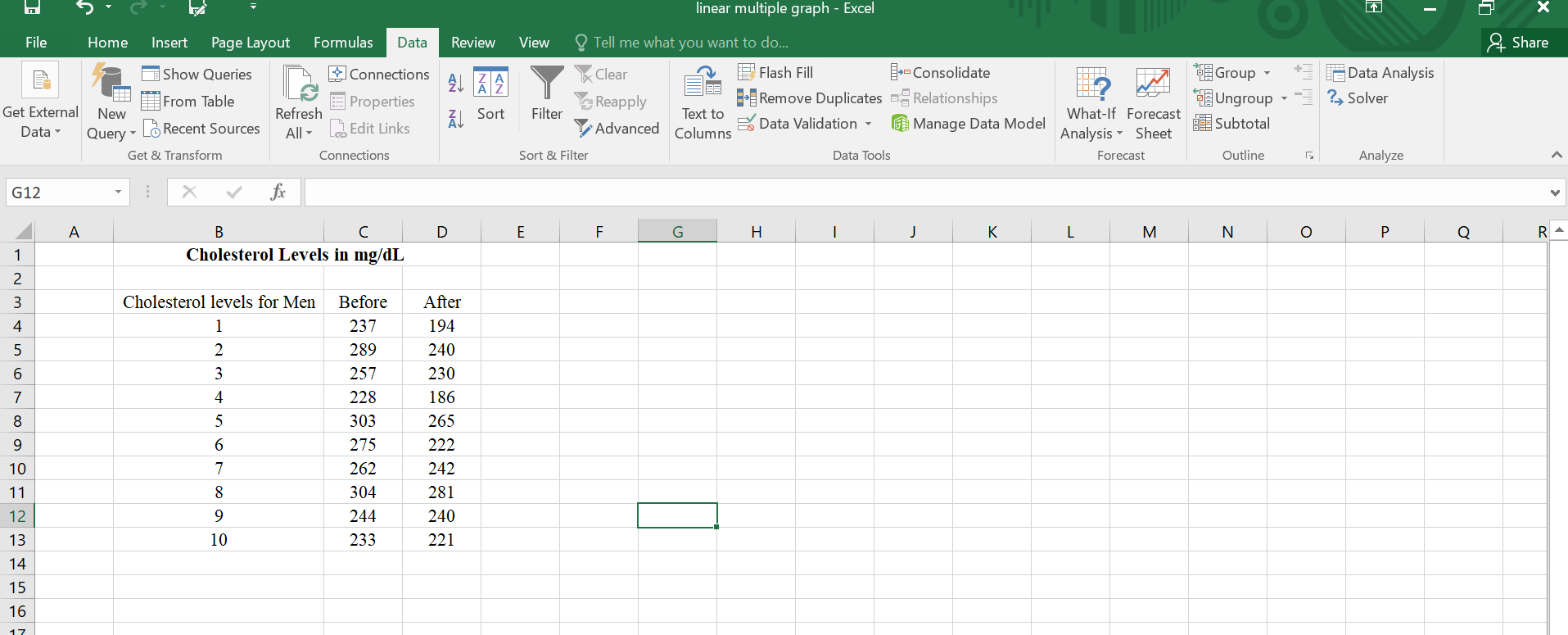
* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



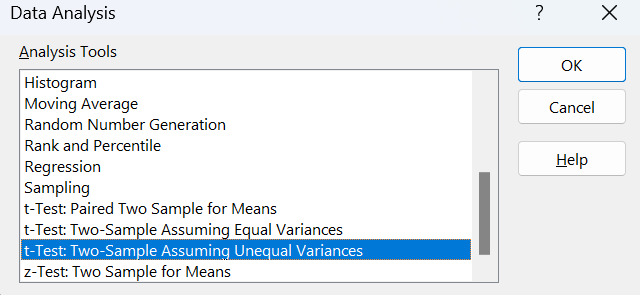
* **Output:**

****

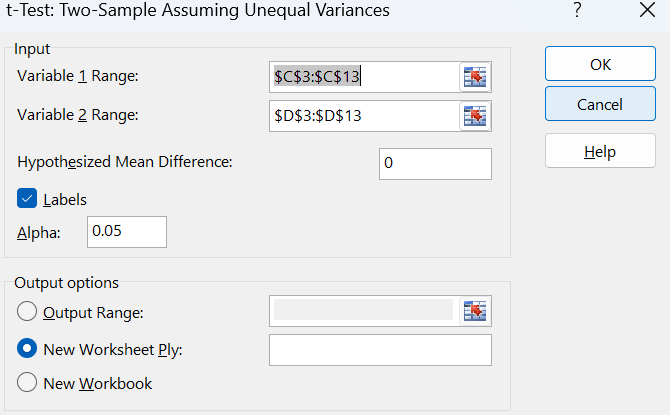
* **T-TEST: TWO-SAMPLE ASSUMING UNEQUAL VARIANCES:**
* **STEPS:**
* Click on Data Analysis on Data tab.



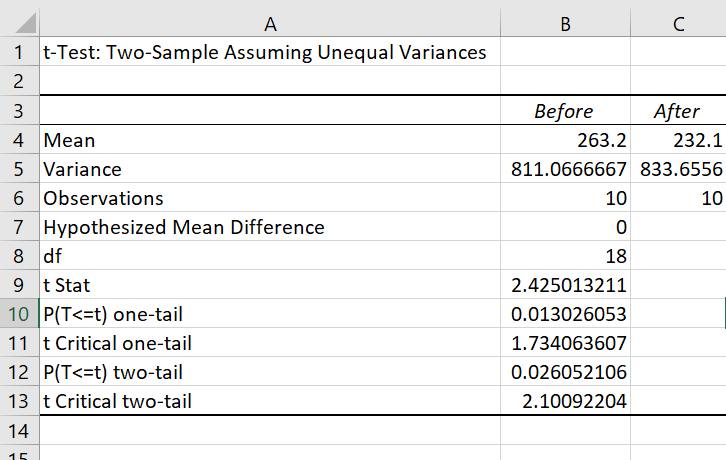
* Choose the regression:



* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **Output:**

****

* **Observation:**

**T-TEST: TWO-SAMPLE ASSUMING EQUAL VARIANCES:**

The means of test Before is 263.2 and test After mean is 232.1. The variance values are 811.0666667 and 833.6556 respectively. The number of observations of both groups was 10. A p-value is a probability that the results from the sample dataset are occurred by chance. Low p-values are considered good. Excel provided us with both one-tail and two-tail T-Tests.

**Practical-6**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem based on Z-test.

* **Theory:**

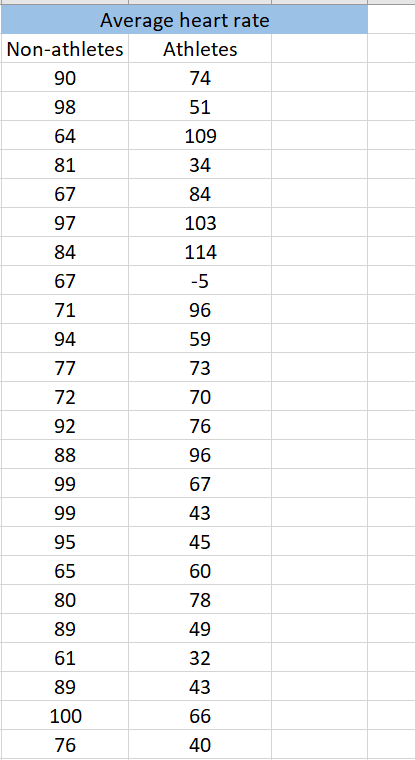
A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large. A z-test is a hypothesis test in which the z-statistic follows a normal distribution. A z-statistic, or z-score, is a number representing the result from the z-test.

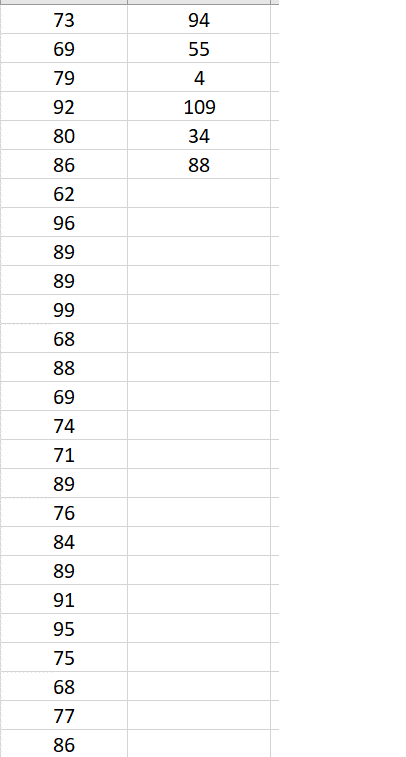
A z-test is used in hypothesis testing to evaluate whether a finding or association is statistically significant or not. In particular, it tests whether two means are the same (the null hypothesis). A z-test can only be used if the population standard deviation is known and the sample size is 30 data points or larger.

A z-test assumes that σ is known; a t-test does not. As a result, a t-test must compute an estimate s of the standard deviation from the sample. Under the null hypothesis that the population is distributed with mean μ, the z-statistic has a standard normal distribution, N (0,1).

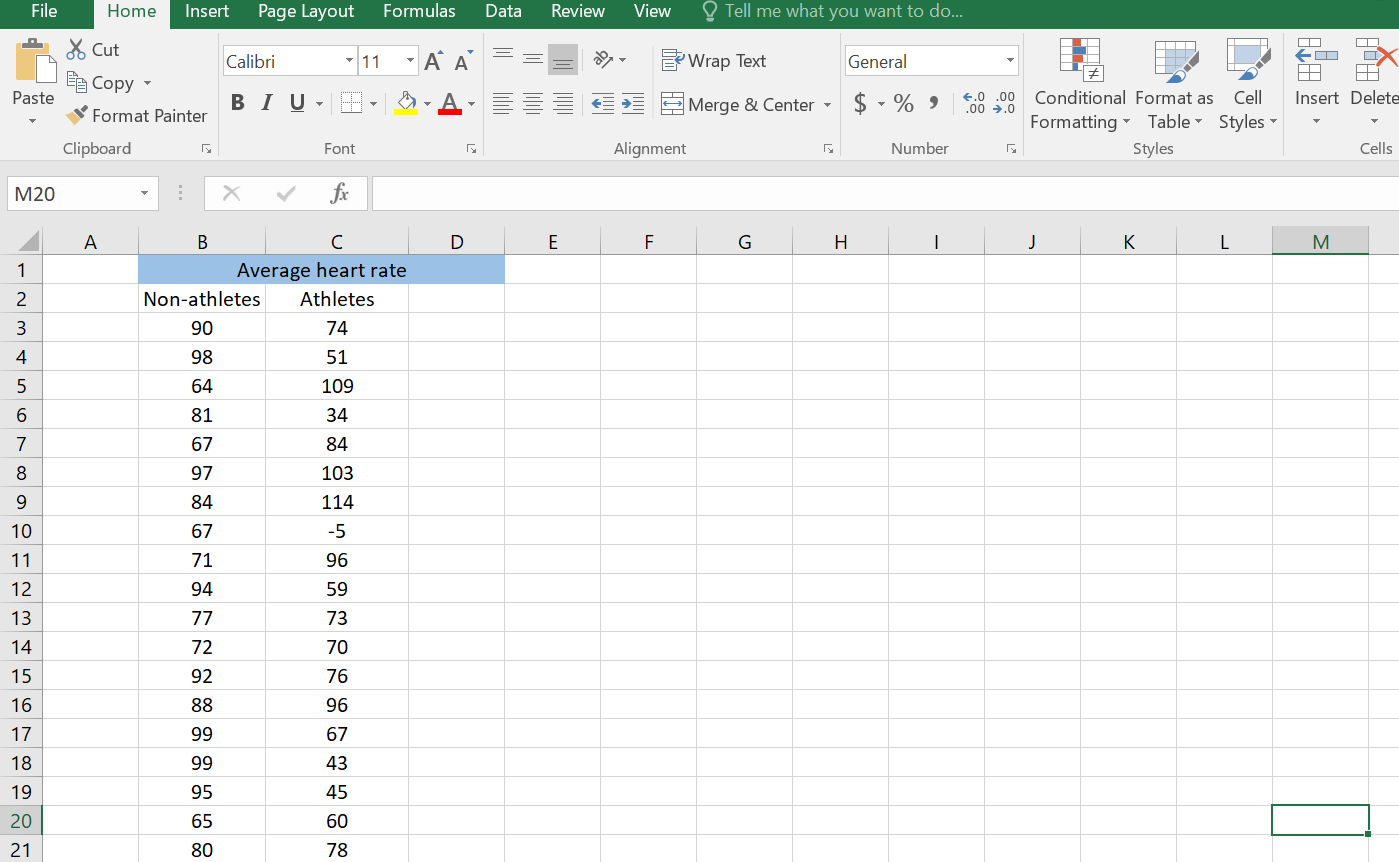
* **Problem:**

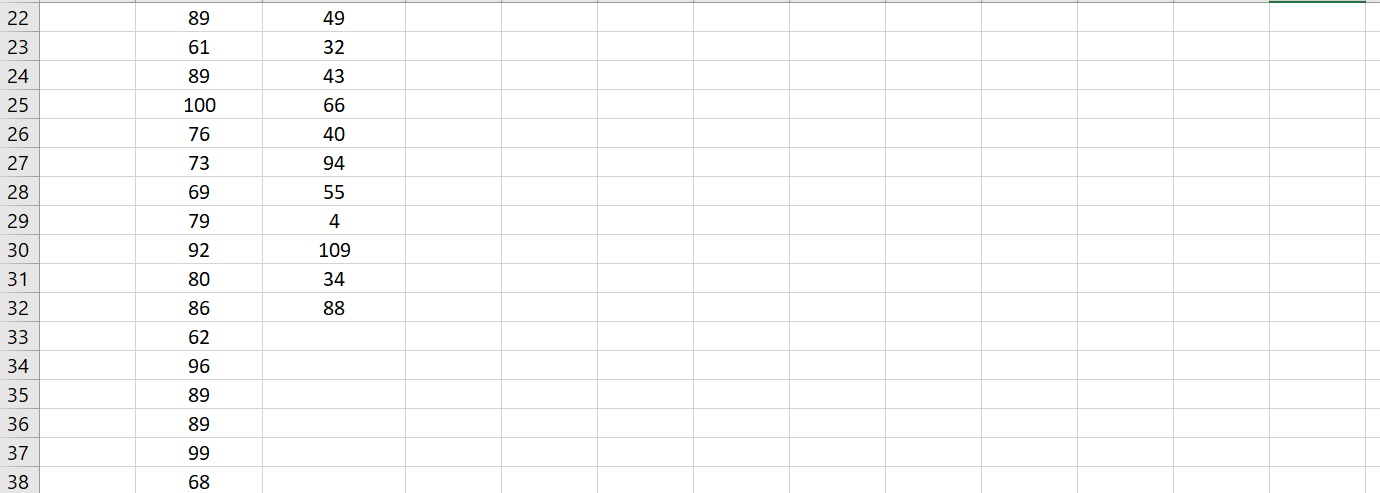
It is based on z-test.

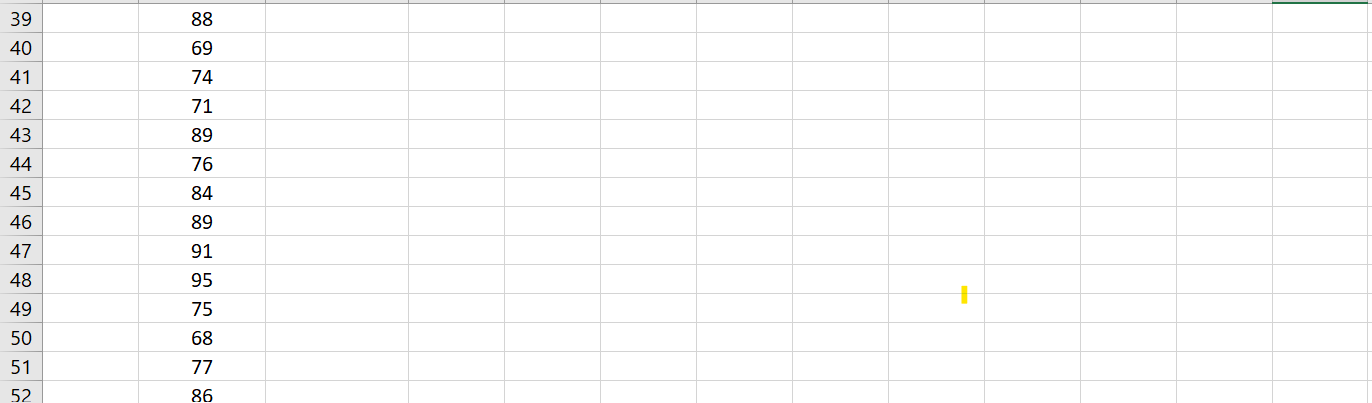
****



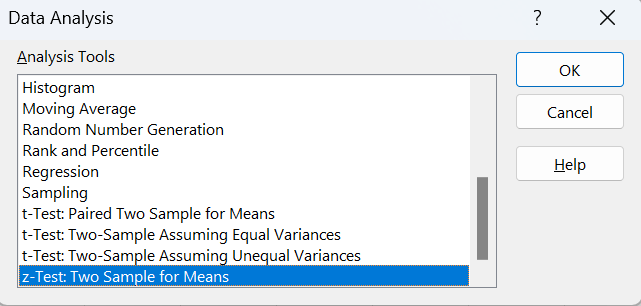
* **STEPS:**
* Click on Data Analysis on Data tab.



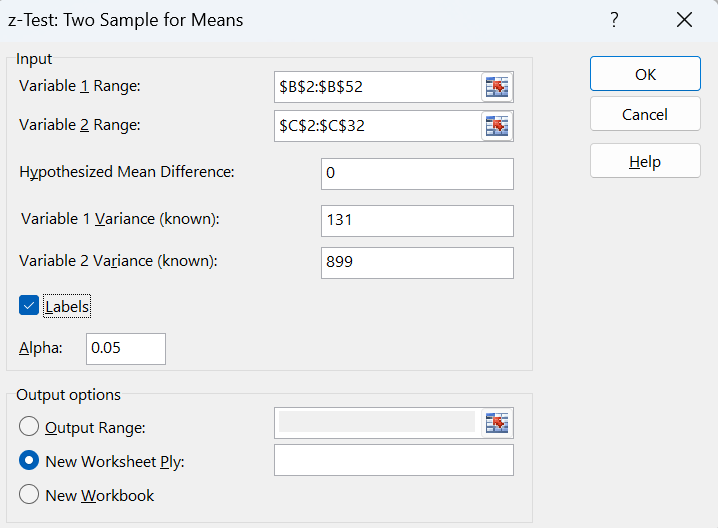




* Choose the regression:



* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **Output:**

**Hypothesis**

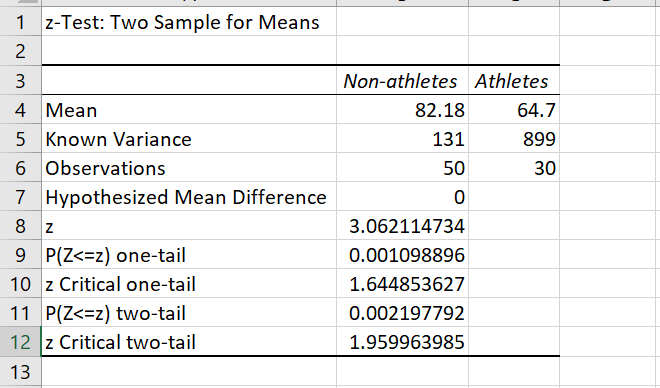
**H0** athletes = non-athletes

**H1**  athletes  non-athletes

**Variance**

**Non-athletes** 131

**Athletes** 899

****

* **Observation:**

The null hypothesis H0 is that the mean difference = 0. or in other words the means are the same.

The alternative hypothesis Ha is that the mean difference is > 0. or in other words that the mean of the trained population is larger

Non-athlets, athlete scompares the p-value (0.002) to the significance level (0.05) and interprets the result for you.

Gives *p*=0.002

**Since the p-value is < 0.05, we can Reject the Null Hypothesis (Means are Different/Means not the Same).**

**System id: 2021509877**

**Date:**

**Practical-7**

* **Aim:**

Problem based on F-test

* **Theory:**

An F-test returns the two-tailed probability that the variances in array1 and array2 are not significantly different. Use this function to determine whether two samples have different variances. For example, given test scores from public and private schools, you can test whether these schools have different levels of test score diversity.

**Syntax**

FTEST (array1, array2)

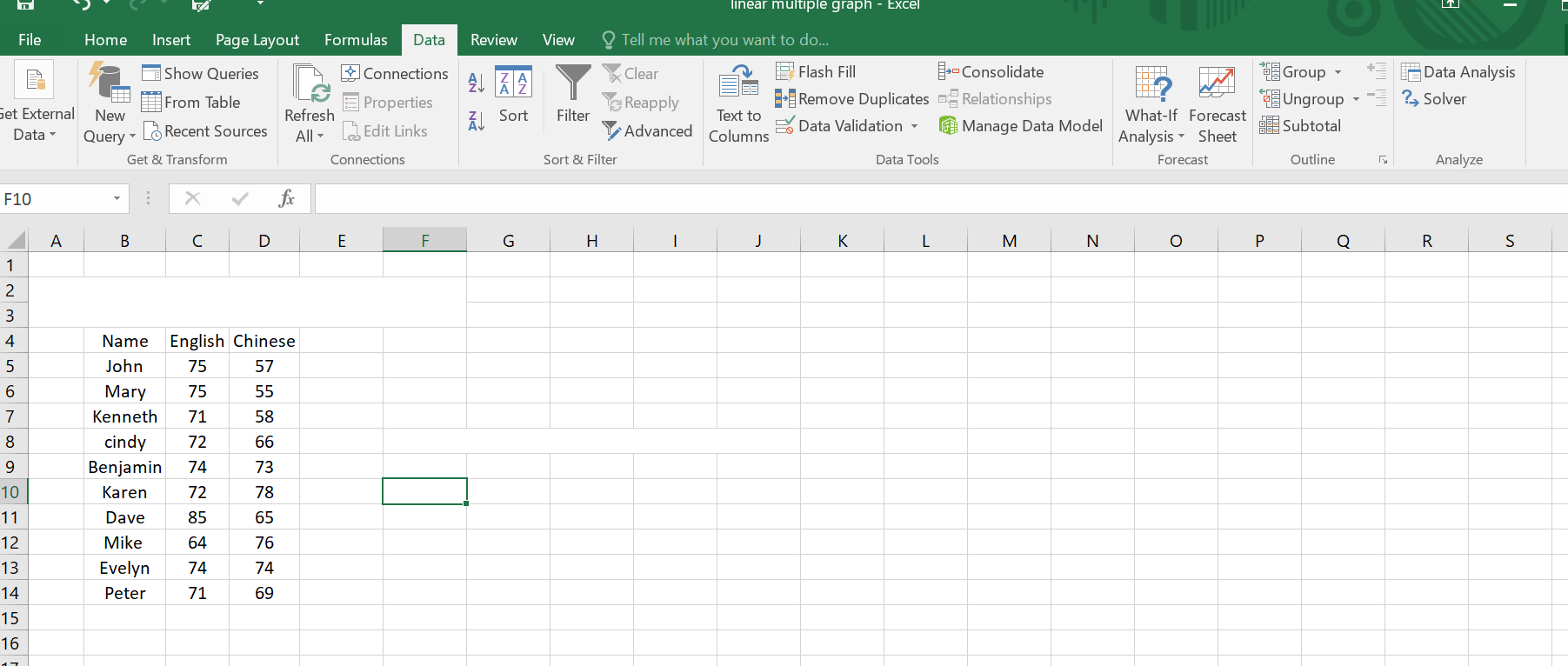
The FTEST function syntax has the following arguments:

* **Array1**     Required. The first array or range of data.
* **Array2**     Required. The second array or range of data
* **Problem:**

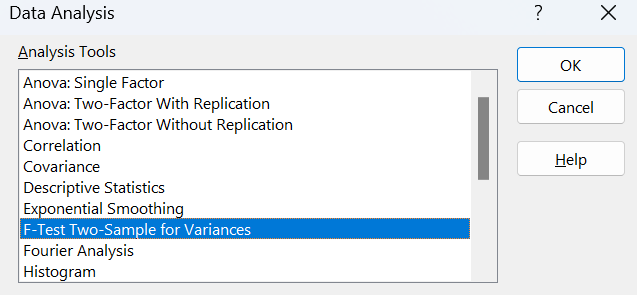
|  |  |  |
| --- | --- | --- |
| Name | English | Chinese |
| John | 75 | 57 |
| Mary | 75 | 55 |
| Kenneth | 71 | 58 |
| cindy | 72 | 66 |
| Benjamin | 74 | 73 |
| Karen | 72 | 78 |
| Dave | 85 | 65 |
| Mike | 64 | 76 |
| Evelyn | 74 | 74 |
| Peter | 71 | 69 |

Problem based on F –test.

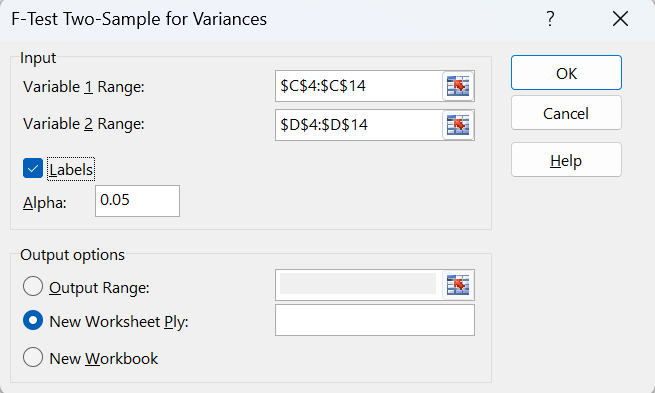
* **STEPS:**
* Click on Data Analysis on Data tab.



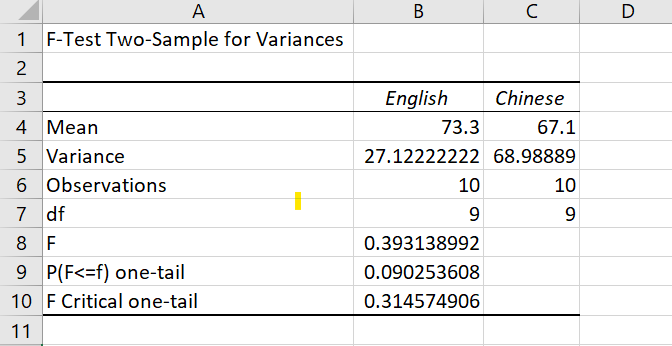
* Choose the regression:



* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **OUTPUT:**



* **Observation:**

**Null hypothesis:**

**H0=** =

**Alternate hypothesis:**

**H1=**

Decision Criteria: If the f statistic < f critical value then rejectthe null hypothesis

**Given:**

F=0.393138

P=0.09

Since the f-test satistic is in the rejection region (p-value<0.05), accept Null hypothesis (H0)

**Practical- 8**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem based on one-way ANOVA.

* **Theory:**

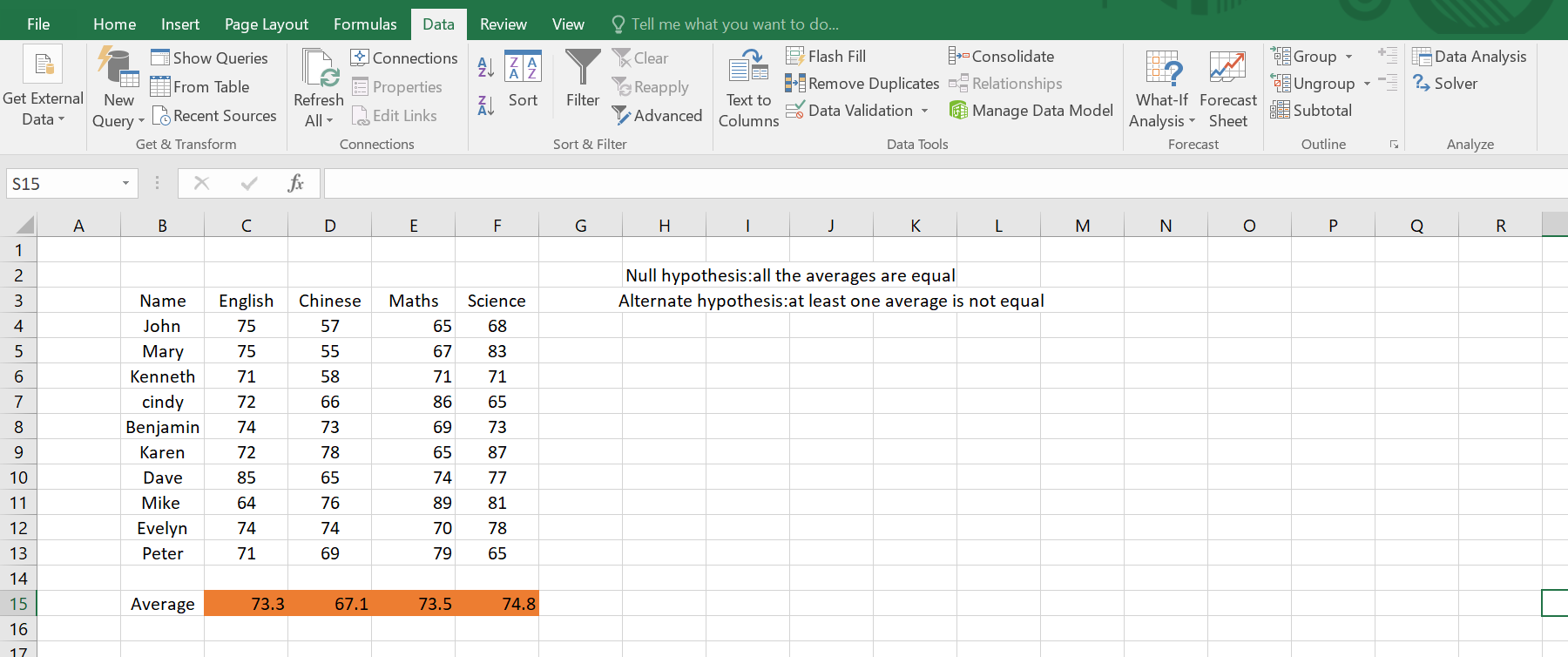
Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

* **Problem:**

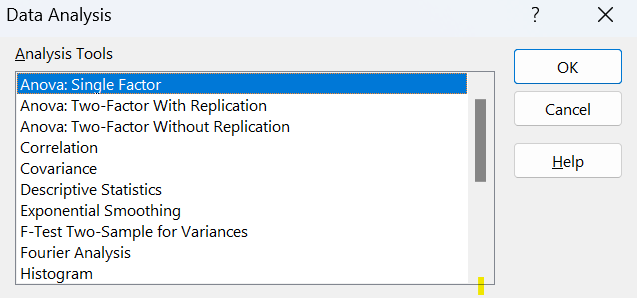
Problem based on one-way ANOVA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | English | Chinese | Maths | Science |
| John | 75 | 57 | 65 | 68 |
| Mary | 75 | 55 | 67 | 83 |
| Kenneth | 71 | 58 | 71 | 71 |
| cindy | 72 | 66 | 86 | 65 |
| Benjamin | 74 | 73 | 69 | 73 |
| Karen | 72 | 78 | 65 | 87 |
| Dave | 85 | 65 | 74 | 77 |
| Mike | 64 | 76 | 89 | 81 |
| Evelyn | 74 | 74 | 70 | 78 |
| Peter | 71 | 69 | 79 | 65 |

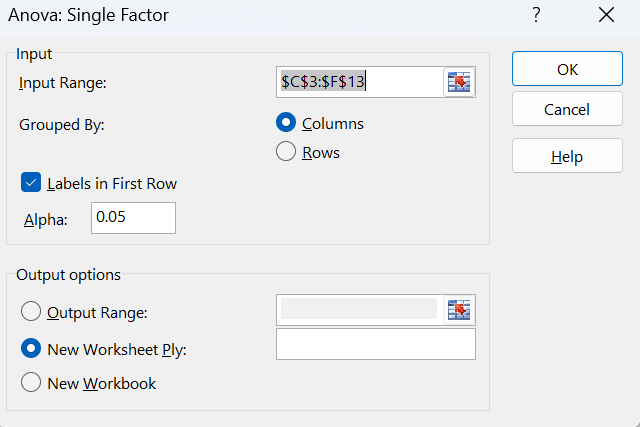
* **STEPS:**
* Click on Data Analysis on Data tab.

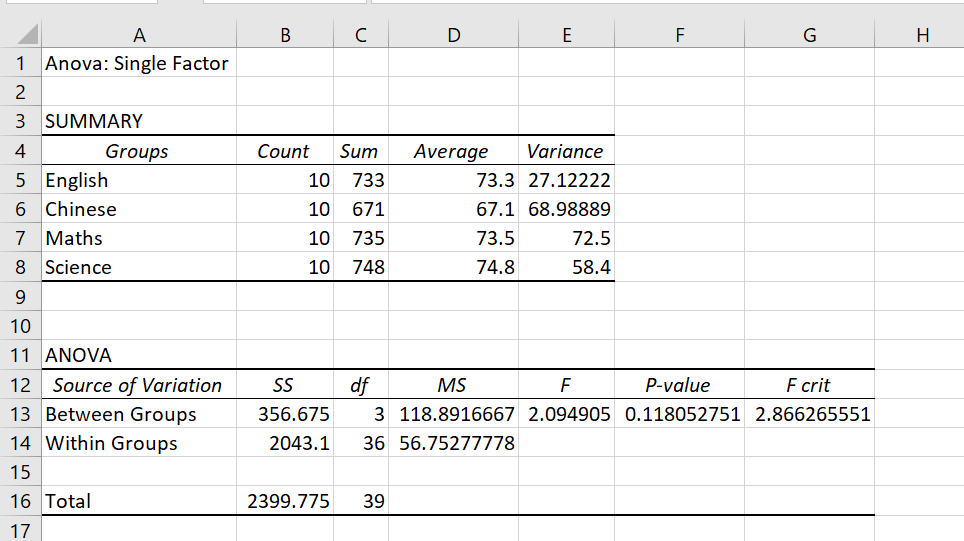


* Choose the regression:



* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **OUTPUT:** 

* **Observation:**

We can see that the value of P is 0.11 which is greater than 0.05(α), hence we will accept the NULL hypothesis.

**Practical-9**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem based on Two-way ANOVA with replication.

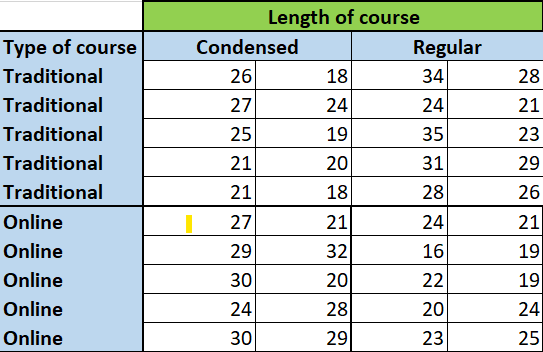
* **Theory:**

A two- way ANOVA with replication is performed when you have two modalities with several levels of the independent variable. For example, you might have group counseling and individual counseling, with symptoms of stress, depression and anxiety as levels.

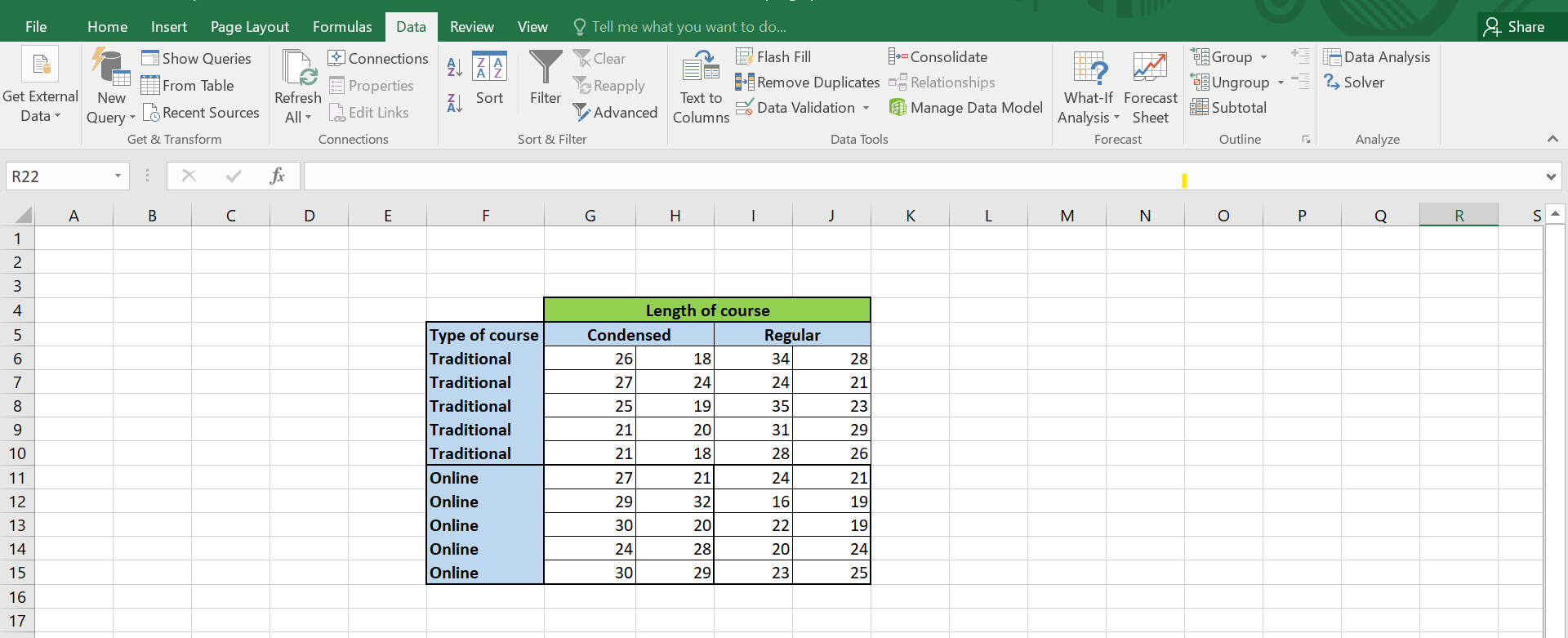
* **Problem:**

Two-way ANOVA with replication.

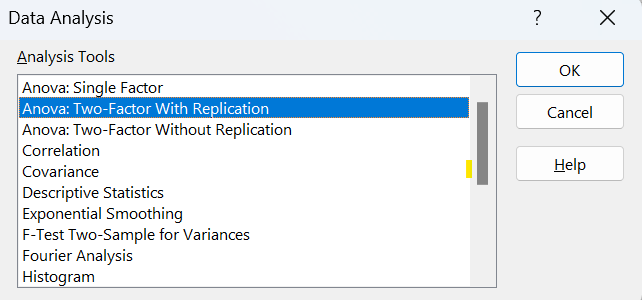
A nationwide company specializing in preparing students for college and graduate school entrance exams, such as the SAT, ACT, and LSAT, had the business objective of improving its ACT preparatory course. Two factors of interest to the company are the length of the course (a condensed 10-day period or a regular 30-day period) and the type of course (traditional classroom or online distance learning). The company collected data by randomly assigning 10 clients to each of the four cells that represent a combination of the length of the course and the type of course. The results are organized in the file ACT and presented in Table.



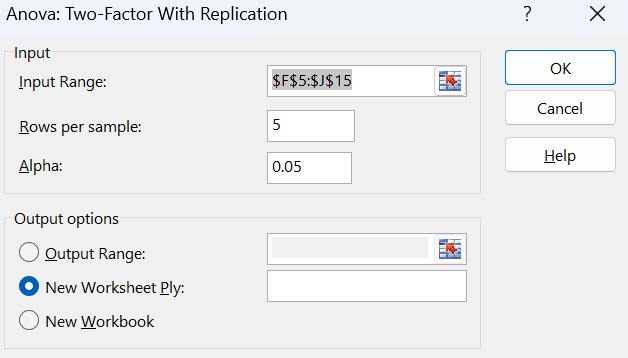
* **STEPS:**
* Click on Data Analysis on Data tab.



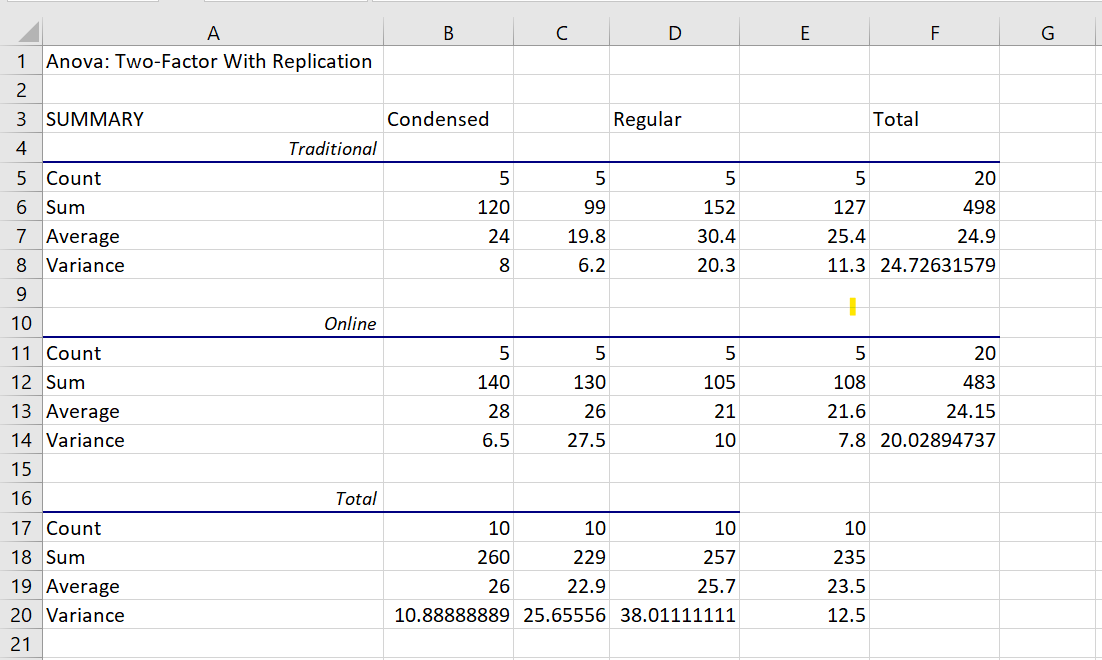
* Choose the regression:

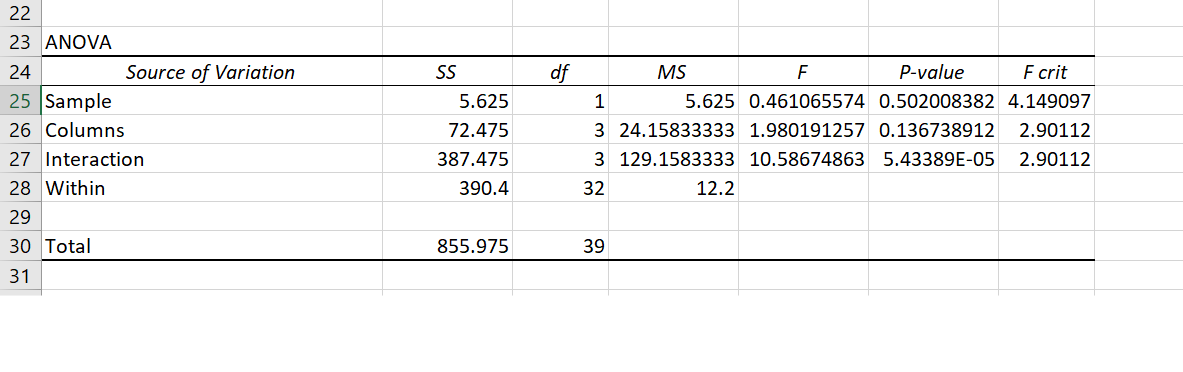


* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **OUTPUT:**





* **Problem2:**

Two-way ANOVA with replication.

The effects of developer strength (factor A) and development time (factor B) on the destiny of photographic plate film were being studied. Two strengths and two development times were used, and four replicates in each of the four cells were evaluated. The results (with larger being best) are stored in Photo and shown in the table:



At the 0.05 level of significance,

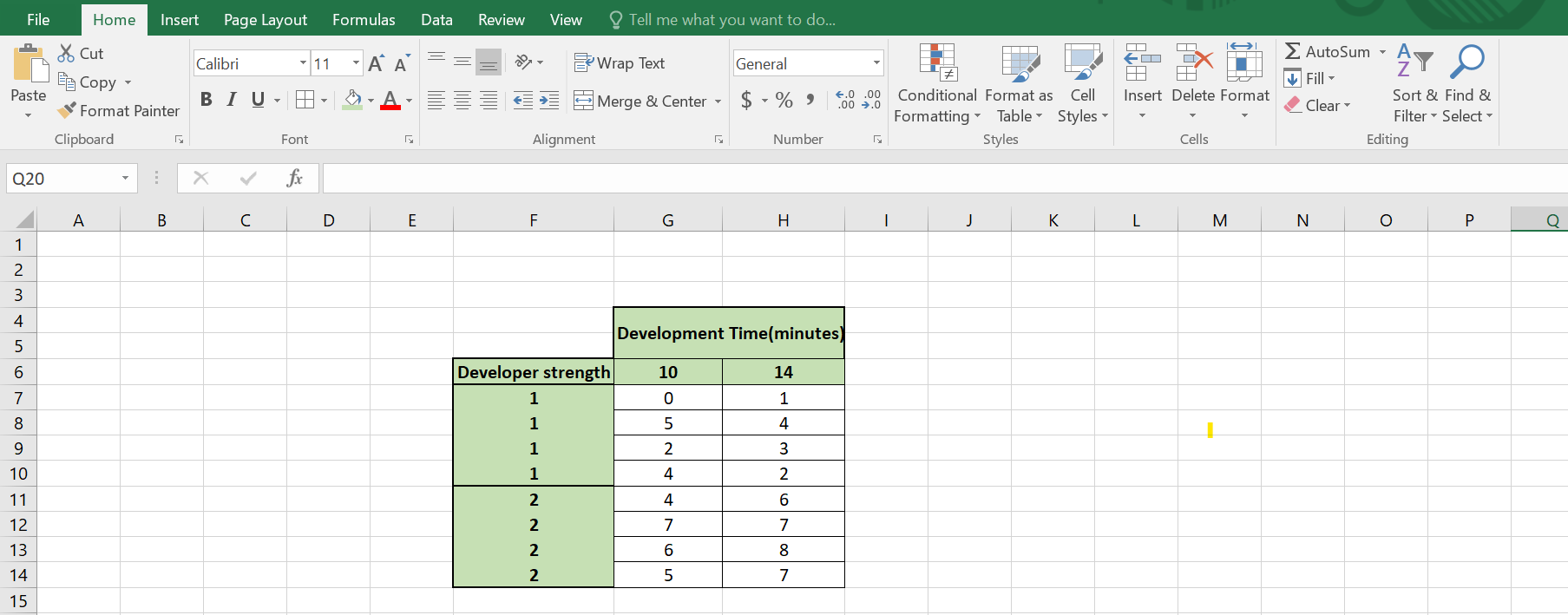
a.) Is there an interaction between developer strength and development time?

b.) Is there an effect due to developer strength?

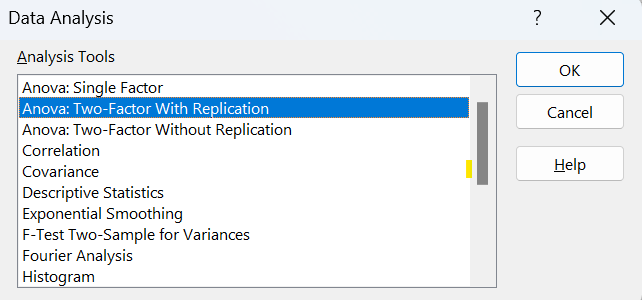
c.) Is there an effect due to development time?

d.) What can you conclude about the effect of developer strength and development time on density?

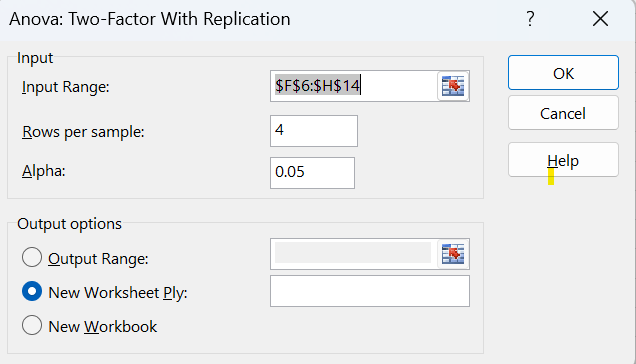
* **STEPS:**
* Click on Data Analysis on Data tab.



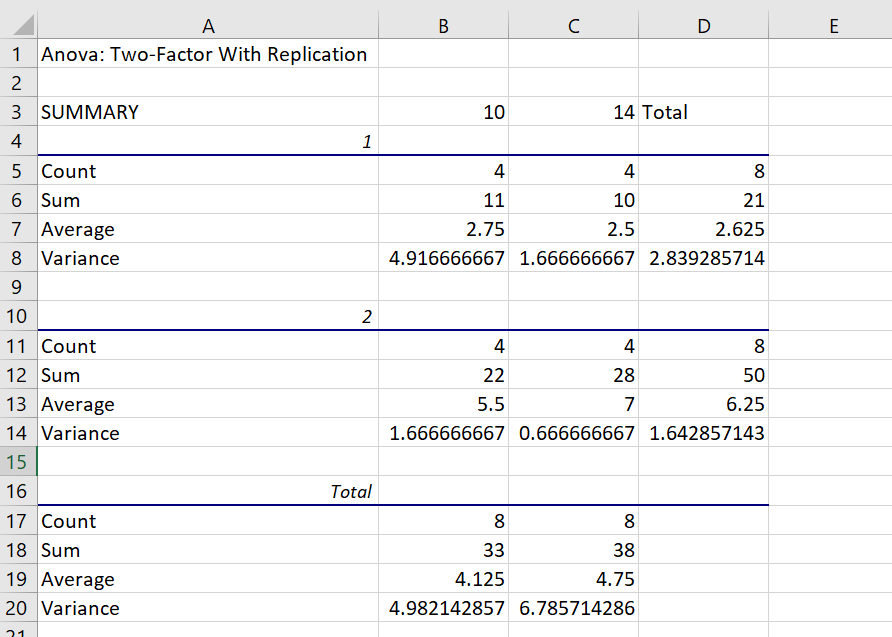
* Choose the regression:

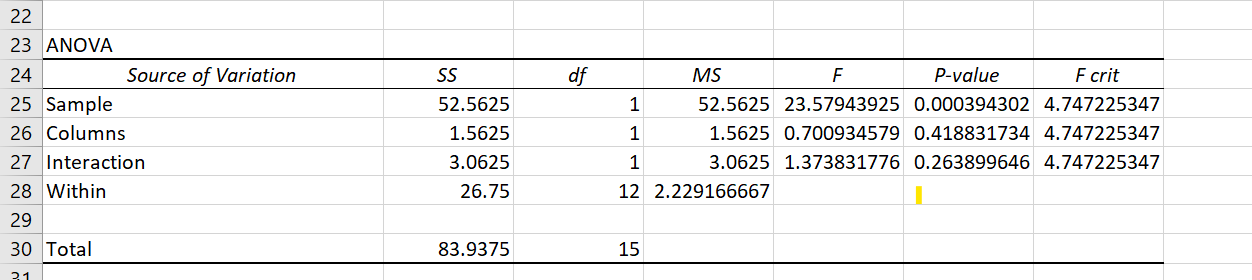


* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **OUTPUT:**





* **Observation:**

1. Because 𝐹𝑠𝑡𝑎𝑡 = 1.37 < 4.75, there is an interaction between developer strength and development time.

b.) Because 𝐹𝑠𝑡𝑎𝑡 = 23.58 > 4.75, there is no effect due to developer strength.

c.) Because 𝐹𝑠𝑡𝑎𝑡 = 0.70 < 4.75, there is an effect due to development time.

d.) Developer strength has a significant effect on density, but development time does not

**Practical:10**

**System id: 2021509877**

**Date:**

* **Aim:**

Problem based on Two-way ANOVA without replication.

* **Theory:**

Two-Factor ANOVA, also known as factorial analysis, is an extension to the one-way analysis of variance. In a two-factor analysis, there are two variables, rather than one as in a single factor analysis.

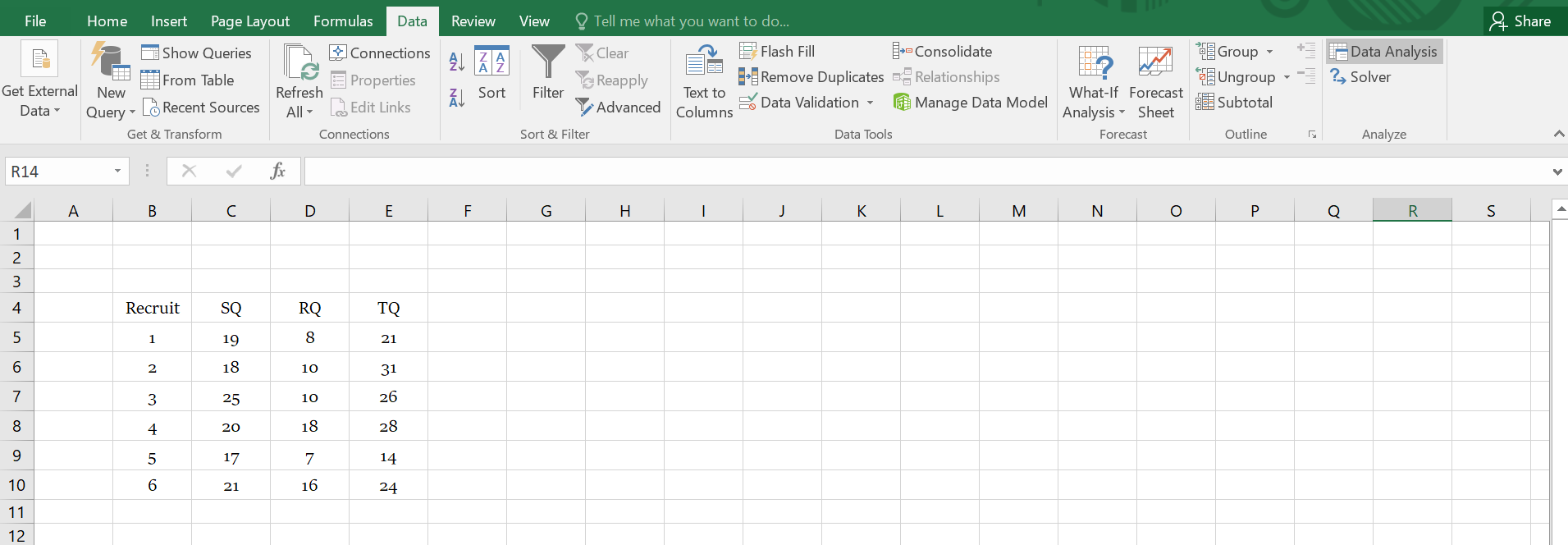
* **Problem:**

Problem based on Two-way ANOVA without replication.

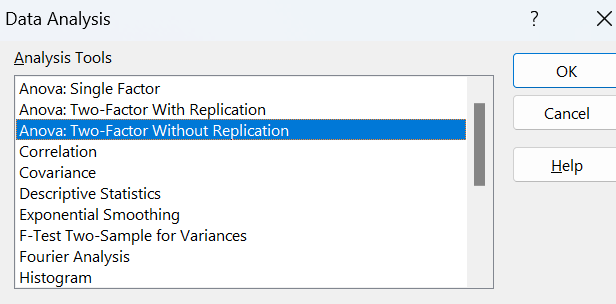
You are a Drill Sergeant for Phase two of basic training for the Marines on the East Coast. You’ve trained the new recruits and they are now being tested in three events. Swim Qualification(SQ), Rifle Qualification(RQ), and Team Qualification(TQ). Each recruit is tested in all three events, and all events were scored out of 35. What can you conclude from the results?

|  |  |  |  |
| --- | --- | --- | --- |
| Recruit | SQ | RQ | TQ |
| 1 | 19 | 8 | 21 |
| 2 | 18 | 10 | 31 |
| 3 | 25 | 10 | 26 |
| 4 | 20 | 18 | 28 |
| 5 | 17 | 7 | 14 |
| 6 | 21 | 16 | 24 |

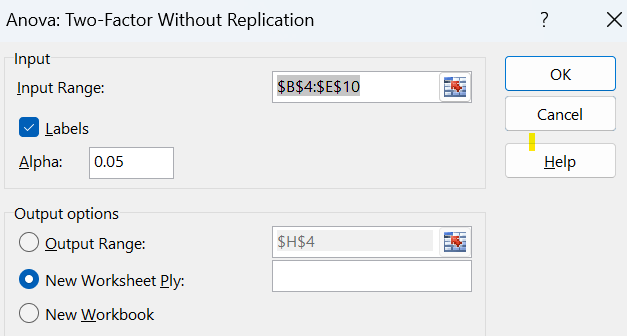
* STEPS:
* Click on Data Analysis on Data tab.



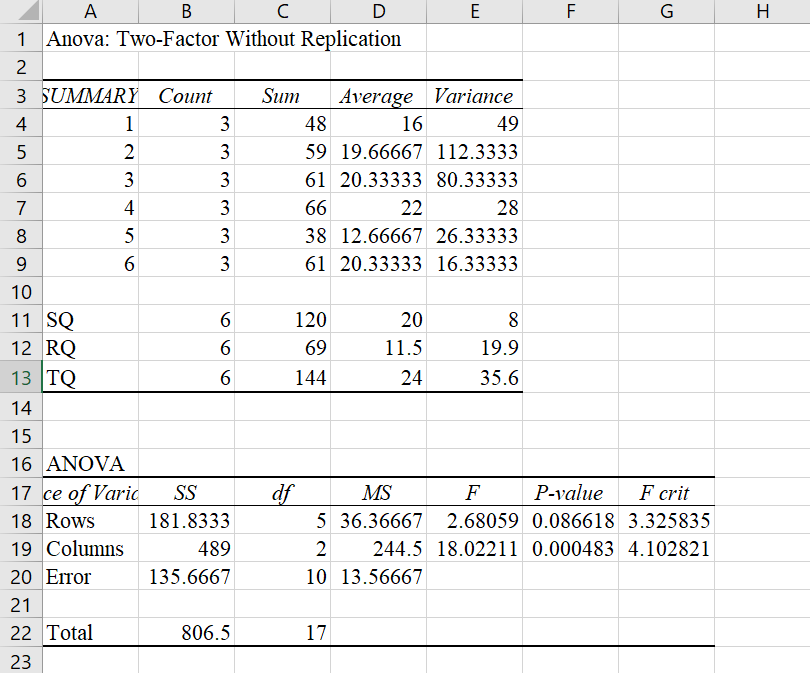
* Choose the regression:



* Select the range of Dependent variable(Y) and Independent variable(X) and Clic ok.



* **OUTPUT:**



* Obervation:

Recruit 4 was the best performing recruit within the group. However, there is no statistically significant difference between the recruits (as indicated by the data for Rows) There is however, a statistically significant difference between the scores of the three tests. Looking at the test means, we can see that the Rifle Qualification’s scores is much lower than the other two. Thus, either the Rifle Qualification is too hard, or the recruits are better suited to be Life Guards rather than Marines