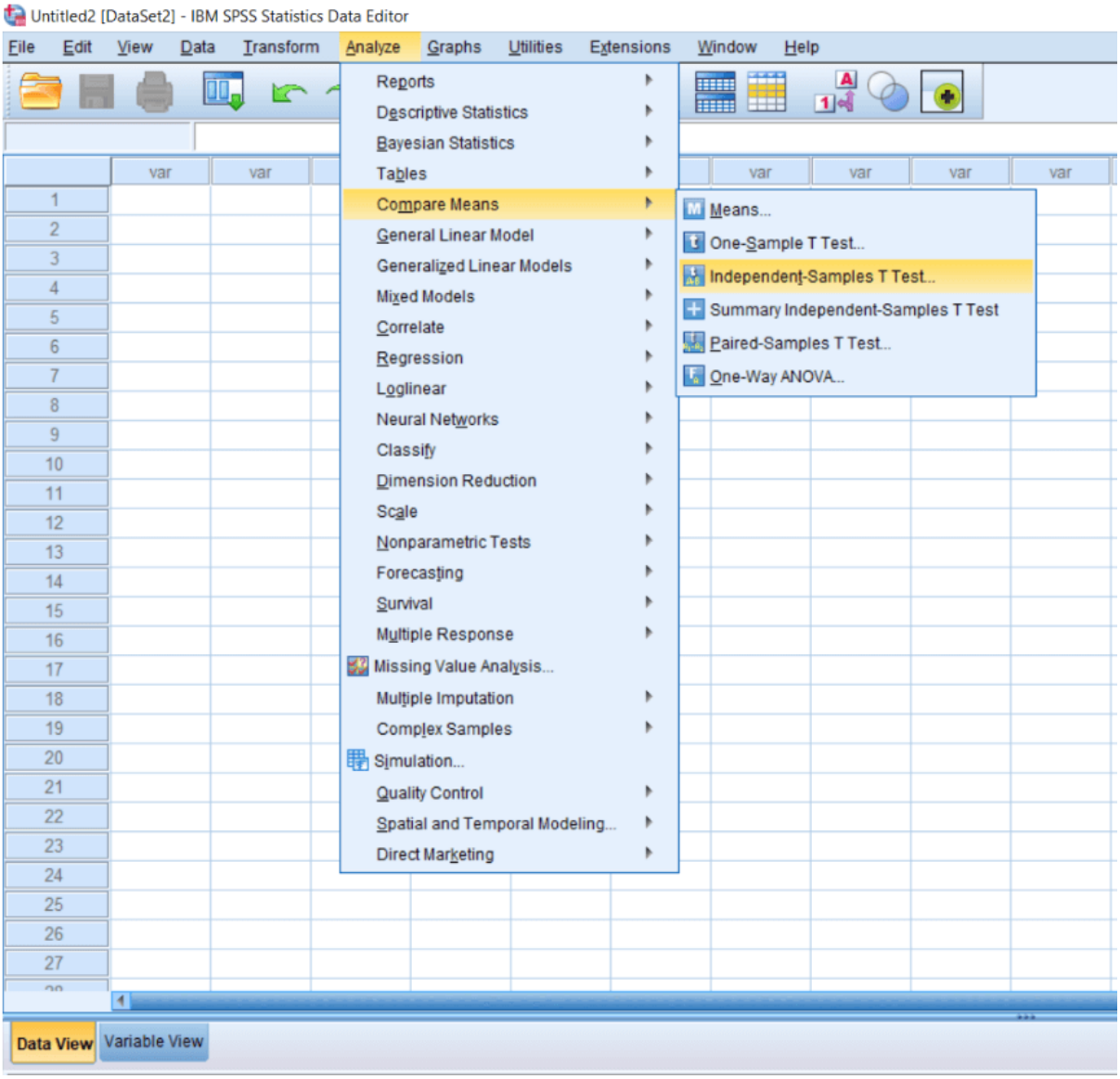
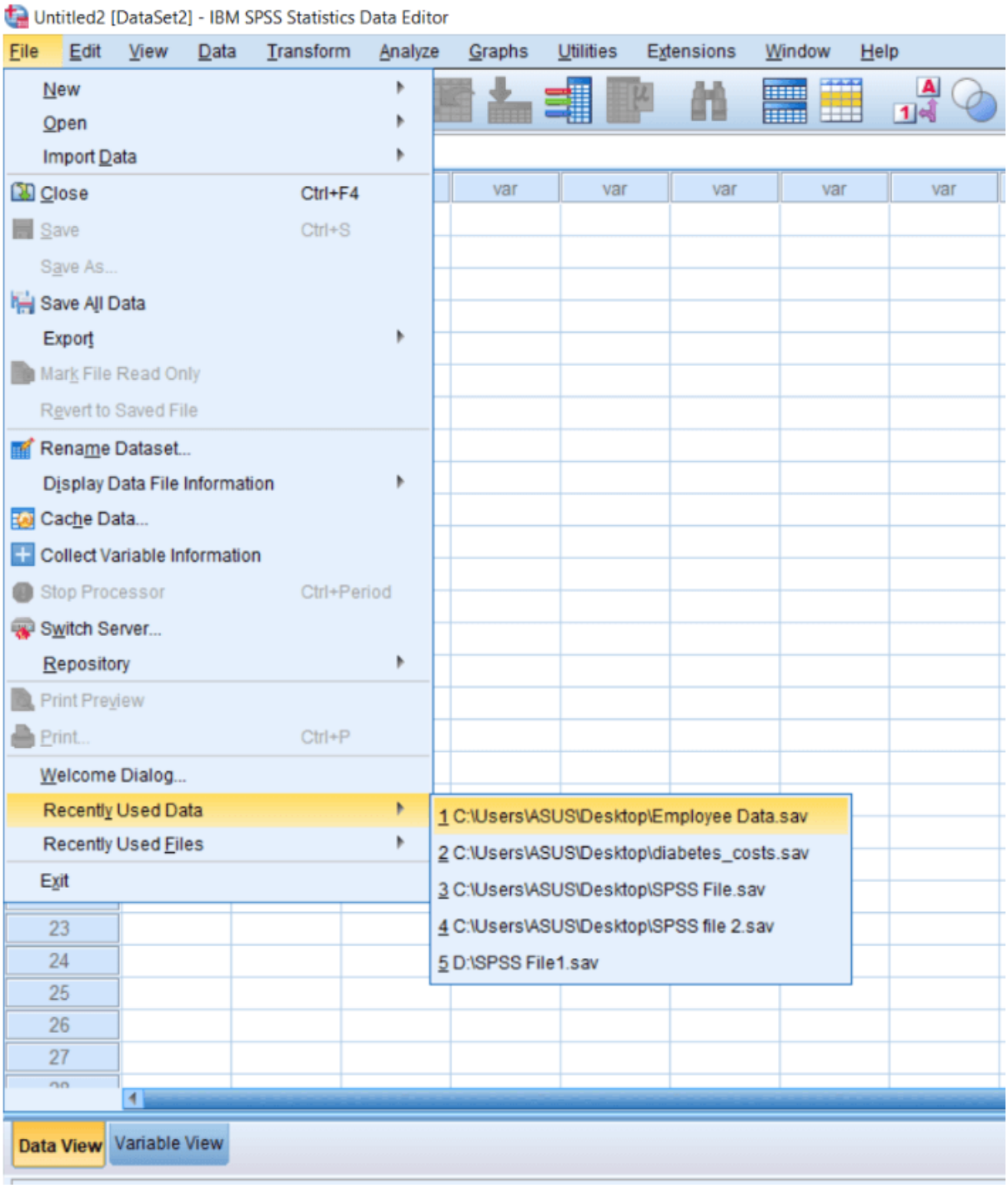
Define Input of Independent Sample T-Test

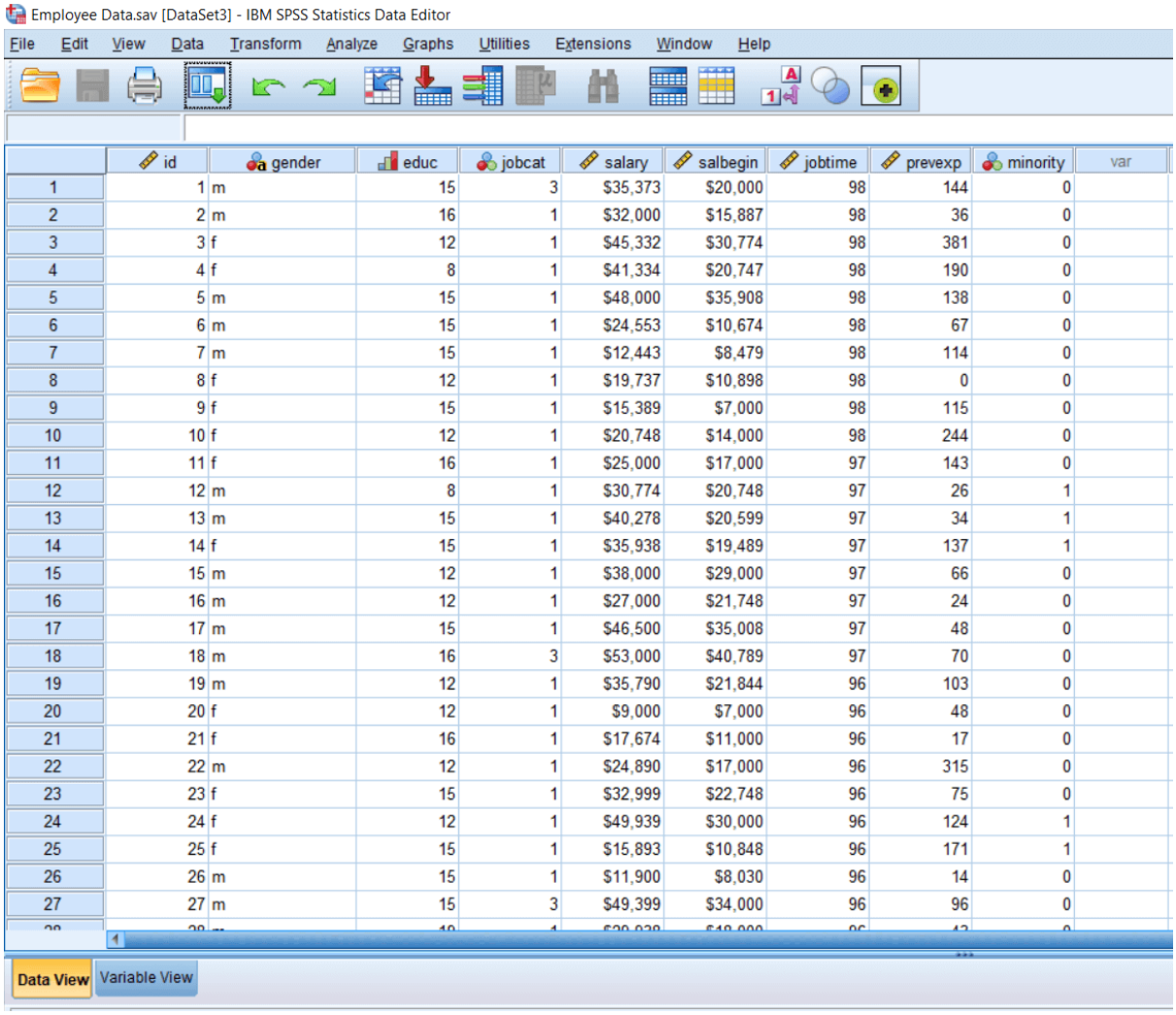
In this section, we will learn **Independent sample T-test**, and how to calculate the differences between two group **Means**. When we have two groups to compare and we want to find out whether there are significant differences between the two groups or not, we can go for a **Mean comparison** between two groups. **Independent sample t-test** is a powerful test for finding out the group differences between two group means. To calculate the **Independent sample T-test**, we will go to the **Analyze menu** and then go to **Compare Means**. Now we can see the **Independent sample T-test** like this:



The symbol of **Independent sample T-test** read as **t A-B**. It means the group we are comparing **A** and **B** are **independent** of each other. **For example**, suppose we want to compare the **salary** of **males** and **females** or the **population** of two **cities** like Delhi or Mumbai. In this case, the groups are not related to each other. So we can go for an **Independent sample t-test**. To calculate the independent sample t-test, we will open our **Data set**. We will go to the **File menu**, then go to **Recently used Data** as follows:

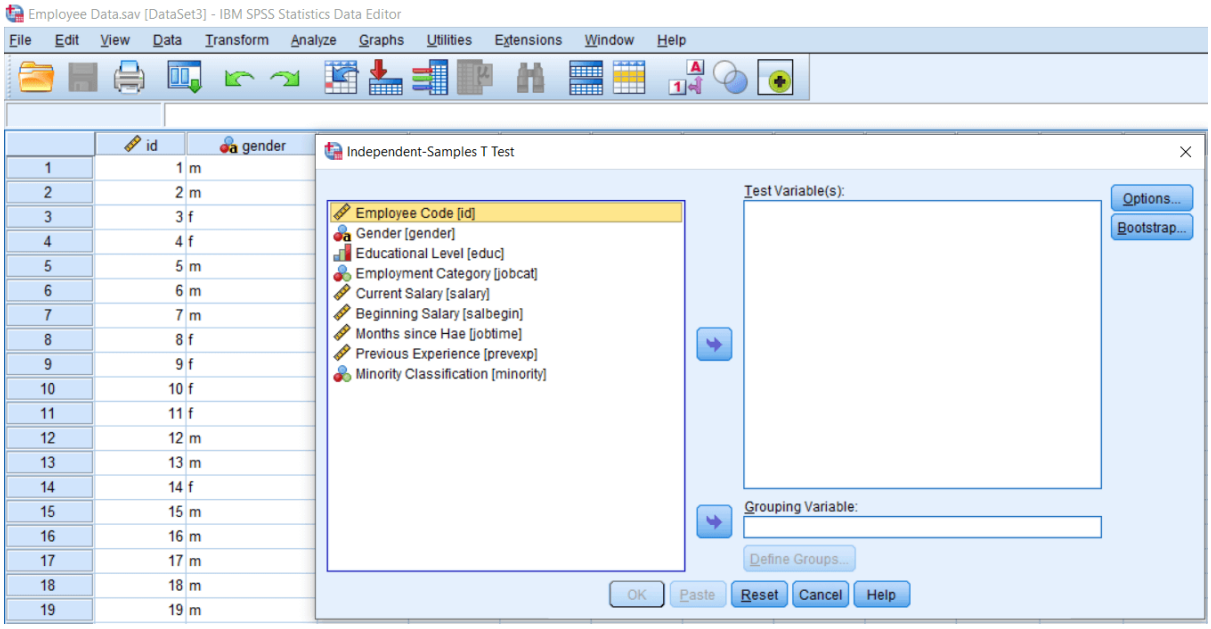


Now we will click on the above **Employee Data** option and see our **Employee Data** set as follows:

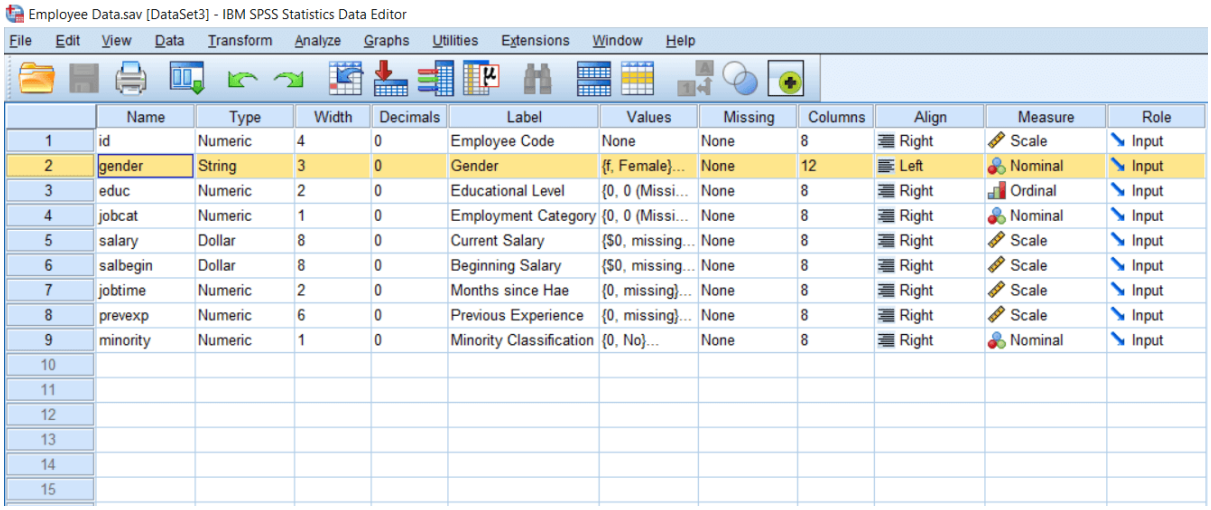


	id	gender	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var
1	1	m	15	3	$35,373	$20,000	98	144	0	
2	2	m	16	1	$32,000	$15,887	98	36	0	
3	3	f	12	1	$45,332	$30,774	98	381	0	
4	4	f	8	1	$41,334	$20,747	98	190	0	
5	5	m	15	1	$48,000	$35,908	98	138	0	
6	6	m	15	1	$24,553	$10,674	98	67	0	
7	7	m	15	1	$12,443	$8,479	98	114	0	
8	8	f	12	1	$19,737	$10,898	98	0	0	
9	9	f	15	1	$15,389	$7,000	98	115	0	
10	10	f	12	1	$20,748	$14,000	98	244	0	
11	11	f	16	1	$25,000	$17,000	97	143	0	
12	12	m	8	1	$30,774	$20,748	97	26	1	
13	13	m	15	1	$40,278	$20,599	97	34	1	
14	14	f	15	1	$35,938	$19,489	97	137	1	
15	15	m	12	1	$38,000	$29,000	97	66	0	
16	16	m	12	1	$27,000	$21,748	97	24	0	
17	17	m	15	1	$46,500	$35,008	97	48	0	
18	18	m	16	3	$53,000	$40,789	97	70	0	
19	19	m	12	1	$35,790	$21,844	96	103	0	
20	20	f	12	1	$9,000	$7,000	96	48	0	
21	21	f	16	1	$17,674	$11,000	96	17	0	
22	22	m	12	1	$24,890	$17,000	96	315	0	
23	23	f	15	1	$32,999	$22,748	96	75	0	
24	24	f	12	1	$49,939	$30,000	96	124	1	
25	25	f	15	1	$15,893	$10,848	96	171	1	
26	26	m	15	1	$11,900	$8,030	96	14	0	
27	27	m	15	3	$49,399	$34,000	96	96	0	

This is an **Employee data set** where we have the **id** of an employee, their **gender, education, job** category, **salary, beginning salary, job timing, previous experience**, and whether they belong to **minority** or **majority** group. In this case, suppose we want to test that there is a significant difference between the **Salary** of males and females. To test that, we can conduct an **Independent sample t-test**. Similarly, suppose we want to determine whether people from **minority** categories are taking a **lesser** amount of Salary than people from the **majority** community. In that case, we can again calculate the **Independent sample t-test**. To test the independent sample t-test, we will go to the **Analyze menu** and then go to **Compare Means** option. In the **Compare Means** option, we locate the **Independent sample t-test**. When we click on it, we will see a dialog box like this:

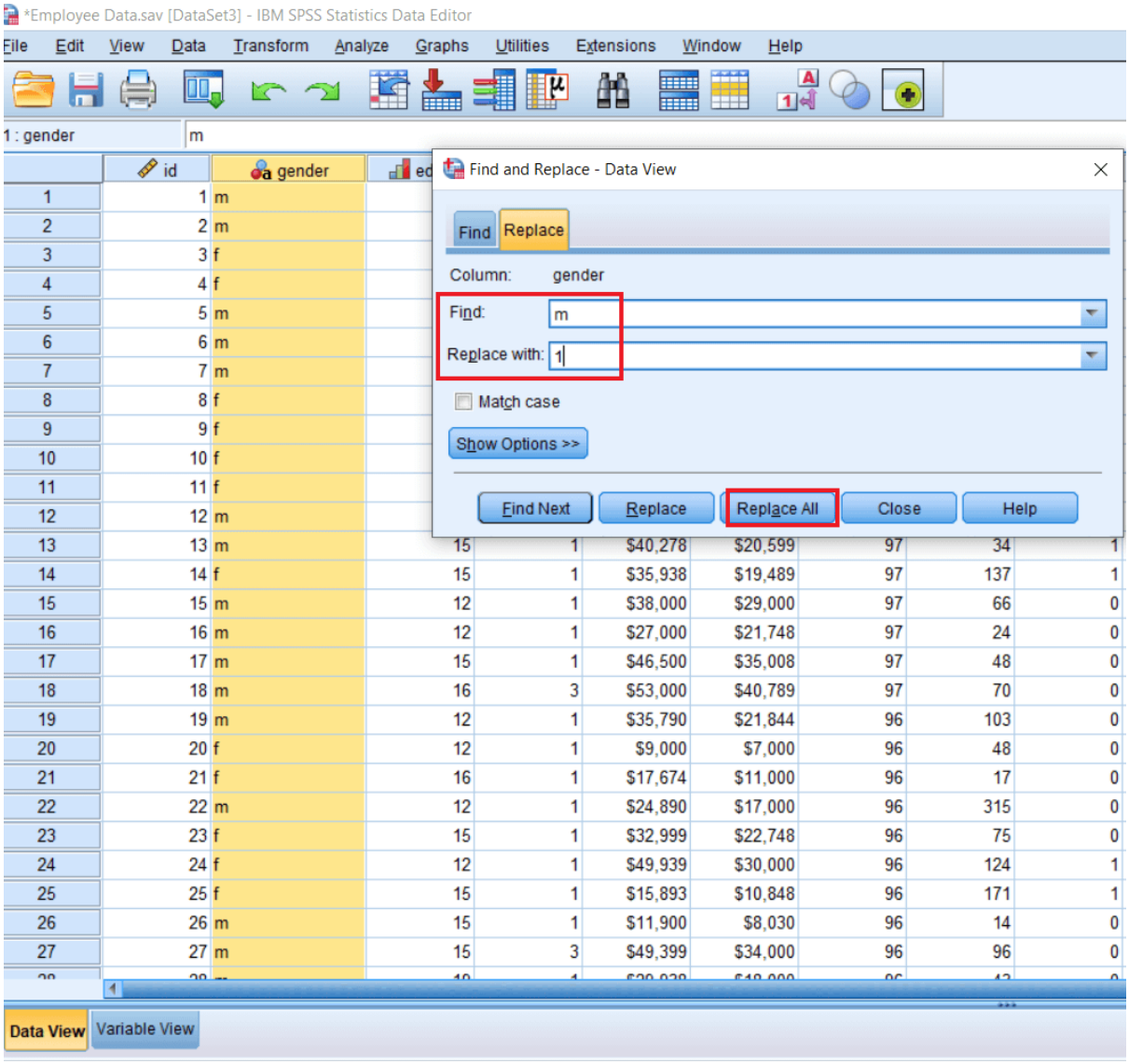


Now we want to compare people across **Gender**. In this case, Gender has been defined as a **String variable**. So to calculate any meaningful test, we need to define all variables as a **Numeric** variable. So we will change the definition of the gender variable. We will turn it into a **Numeric** variable from **String**. So we will go to our **Variable view** and look at the **Gender** as follows:

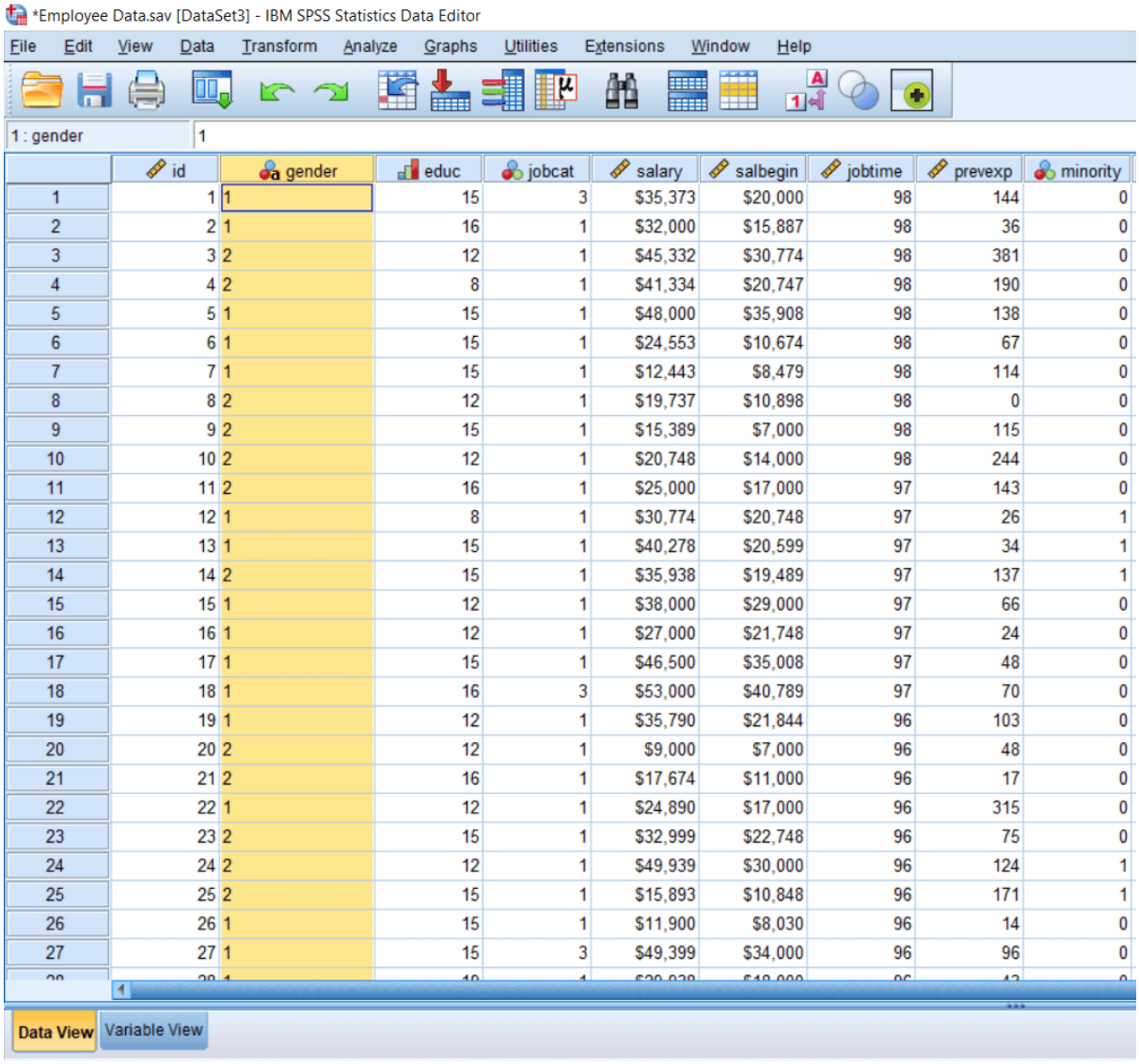


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	4	0	Employee Code	None	None	8	Right	Scale	Input
2	gender	String	3	0	Gender	{f, Female}...	None	12	Left	Nominal	Input
3	educ	Numeric	2	0	Educational Level	{0, 0 (Missi...	None	8	Right	Ordinal	Input
4	jobcat	Numeric	1	0	Employment Category	{0, 0 (Missi...	None	8	Right	Nominal	Input
5	salary	Dollar	8	0	Current Salary	{$0, missing...	None	8	Right	Scale	Input
6	salbegin	Dollar	8	0	Beginning Salary	{$0, missing...	None	8	Right	Scale	Input
7	jobtime	Numeric	2	0	Months since Hae	{0, missing}...	None	8	Right	Scale	Input
8	prevexp	Numeric	6	0	Previous Experience	{0, missing}...	None	8	Right	Scale	Input
9	minority	Numeric	1	0	Minority Classification	{0, No}...	None	8	Right	Nominal	Input
10											
11											
12											
13											
14											
15											

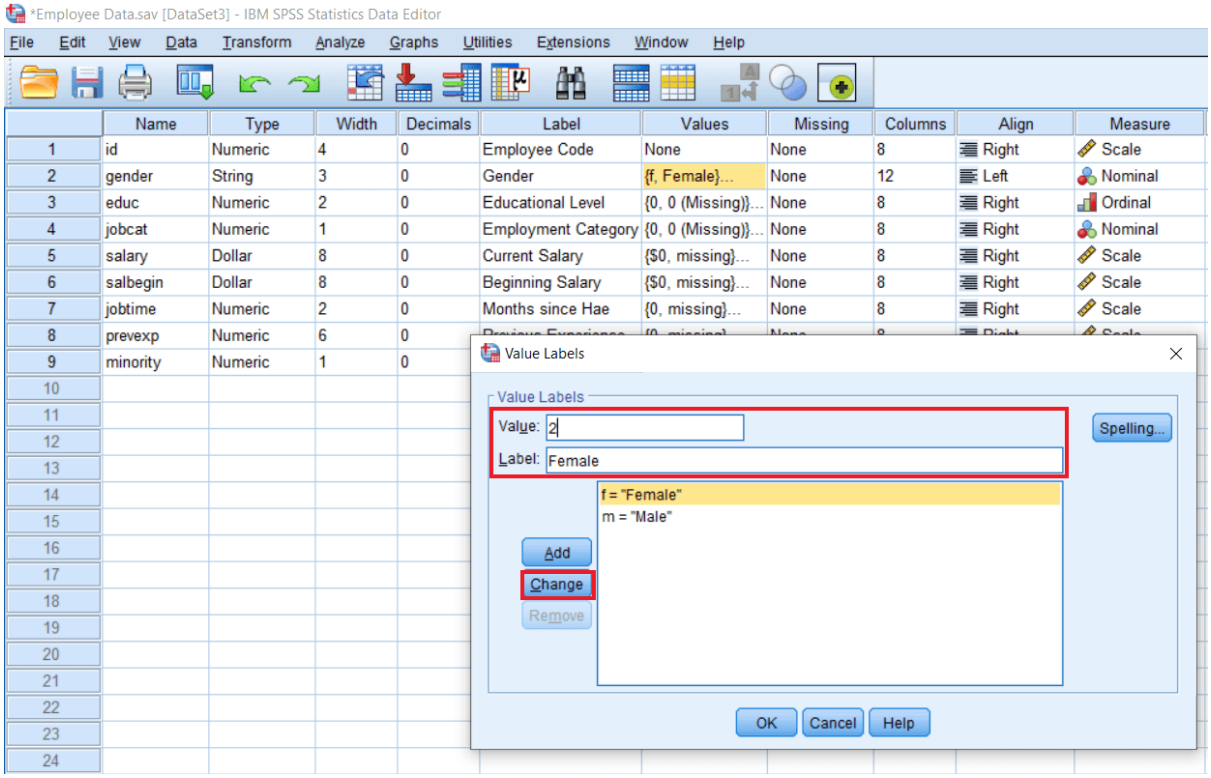
Since it's a String, so first, we need to convert the value. So we will select the row, press **Ctrl+F**, and then click on **Replace**. We will find **m** and replace it by **1** and then click on **Replace all**.



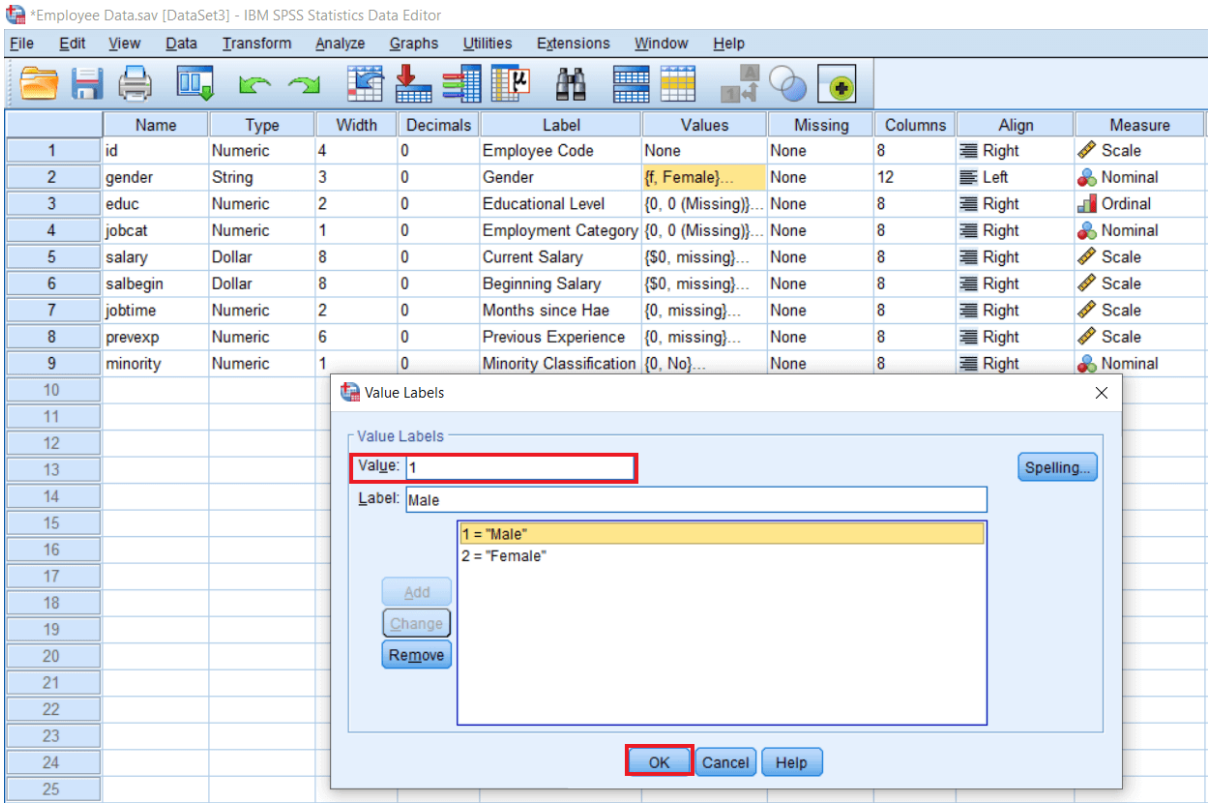
Similarly, we will write **f** and replace it by **2** and click on **Replace all**. After this, we will see the following changes in the Gender variable:



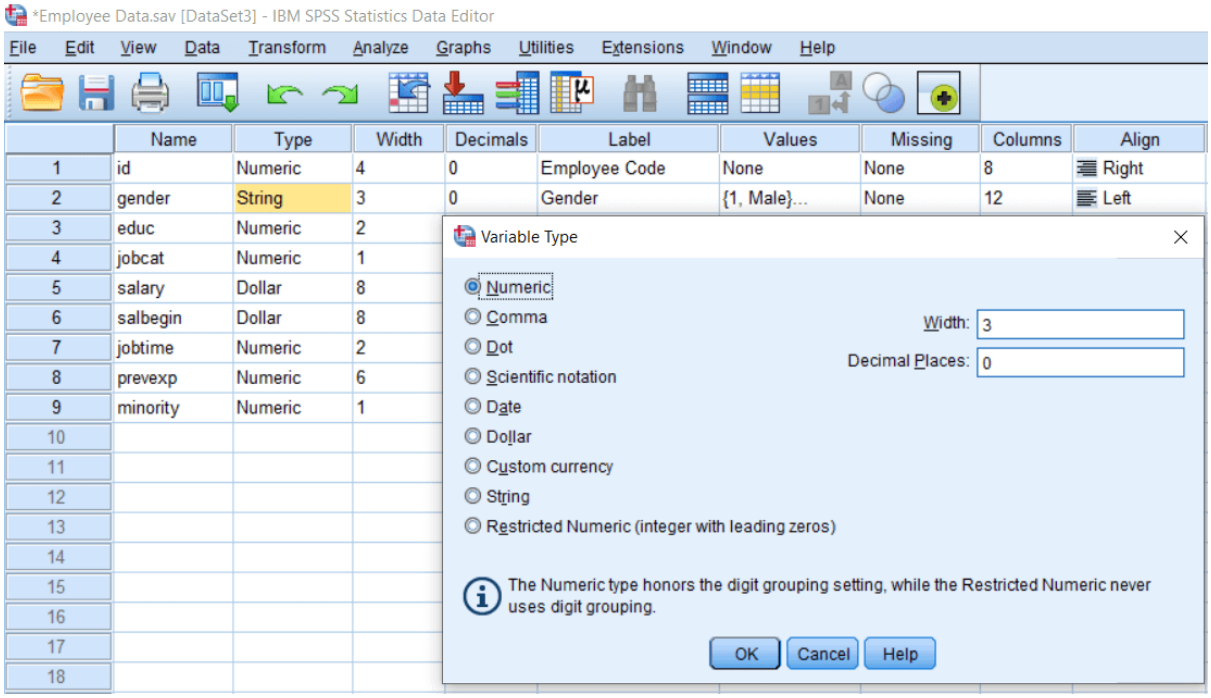
Now we have to redefine this Gender variable. So we will go to the **Variable view** option and click on the **Value step** of the Gender variable. Now we will select the **female** option and define **Value** as **2** and click on **Change** as follows:



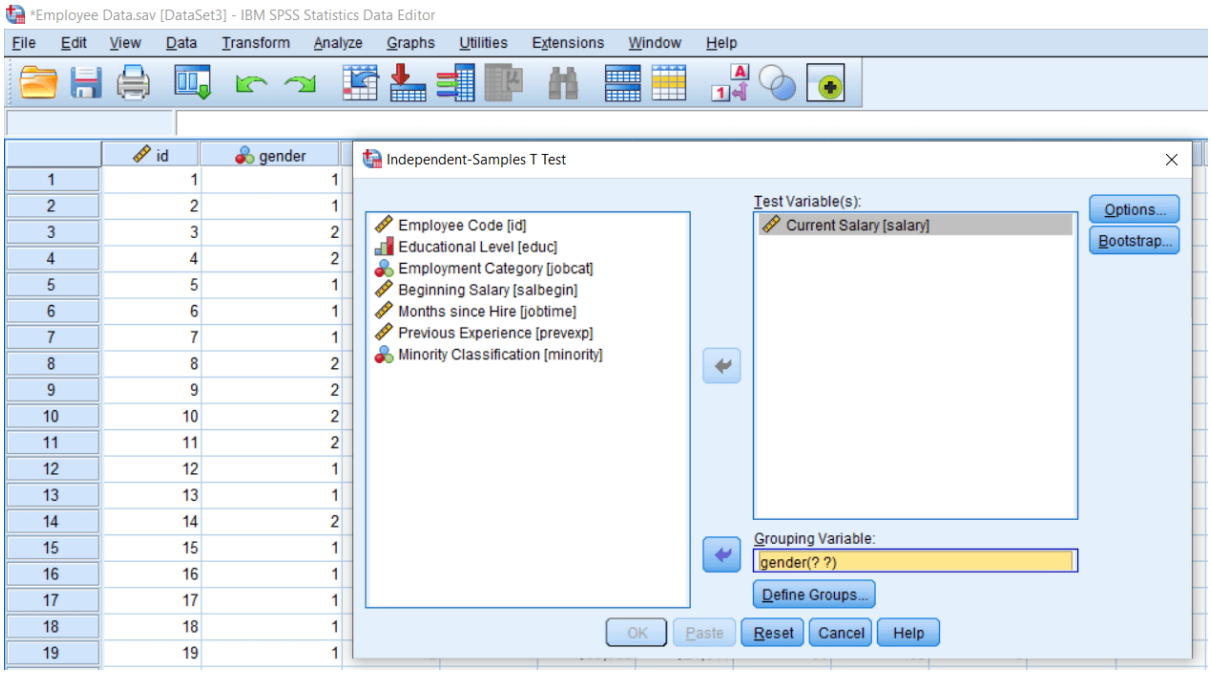
Again select the **male** option and define **Value** as **1** and click on **Change**. Now press **Ok**.



Now we can change the Gender from **String** to **Numeric** variable like this:

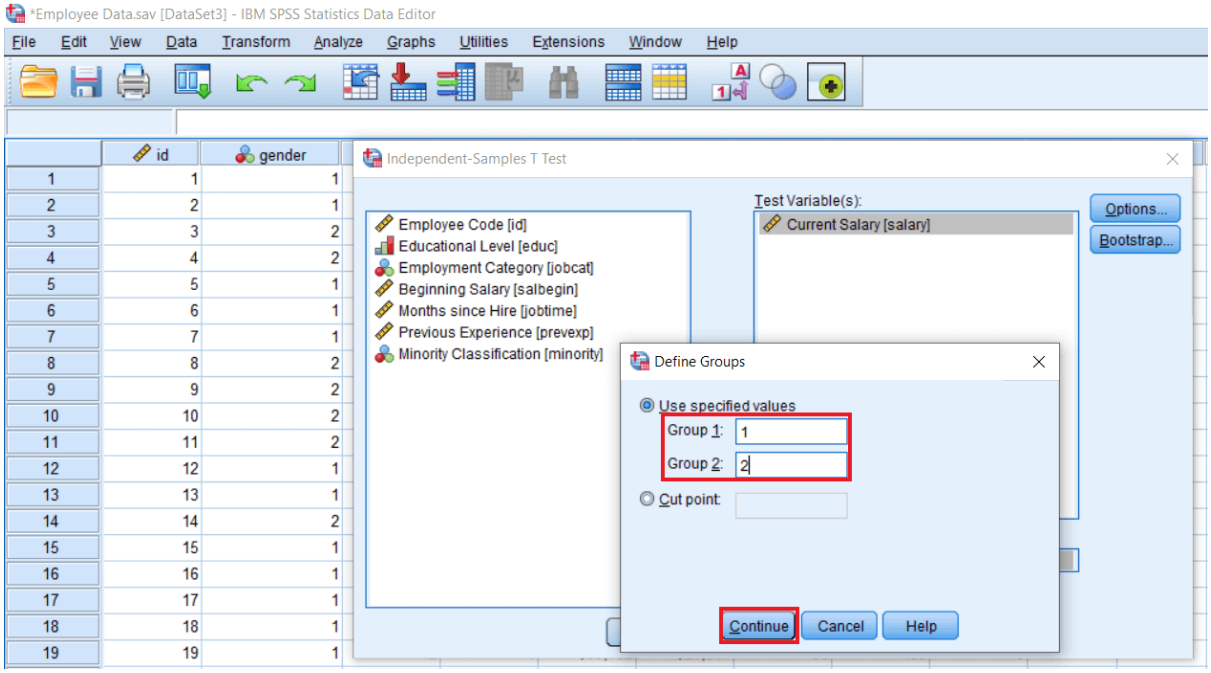


Now variables have been defined. We will go to **Compare Means** option and click on an **Independent sample t-test**. So we want to compare genders for their **salary**. We will take the **Salary** as our **Test variable** and **Gender** as a **Grouping variable**. So the **Salary** is our **dependent** variable, and **Gender** is our **independent** variable.

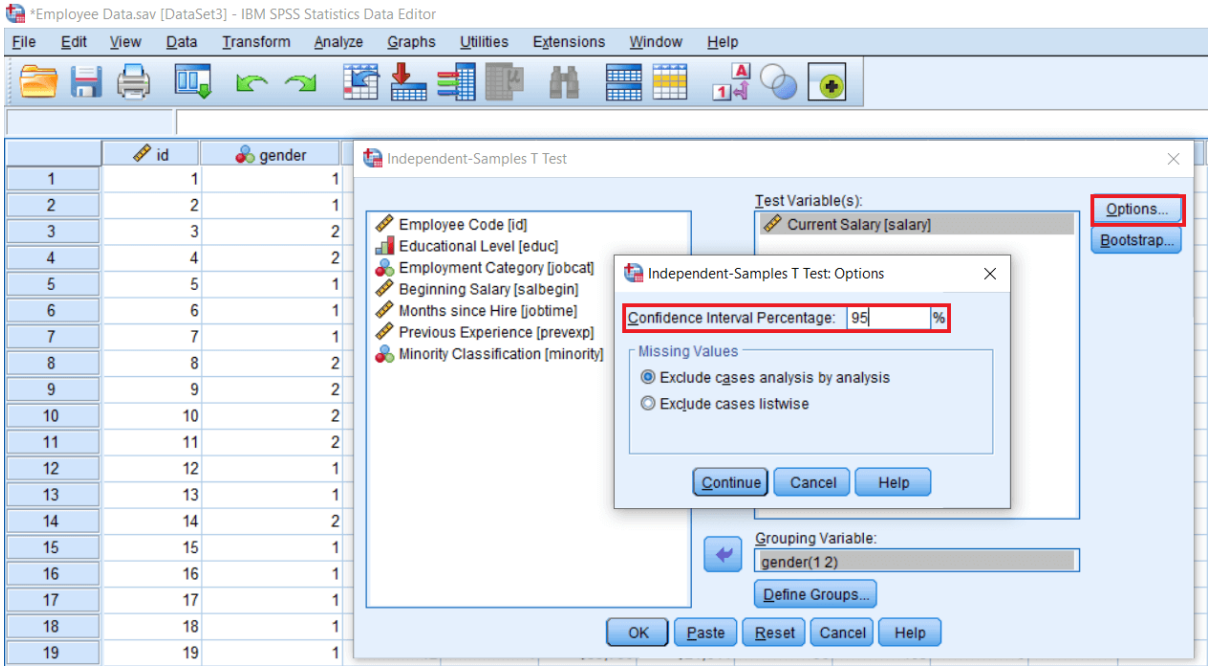


In the above image, under the **Grouping variable**, we can see two **question marks**. It means we need to define our **groups**. So click on **Define Groups** option and write **1** for **Group 1** and **2** for **Group 2**. So, 1 is for males, and 2 is for females. We can also define the **Cut point** instead of defining the groups. **For example**, suppose we have an **exact salary** and want to take a **cutoff** salary, which could be a median salary

or any salary, suppose **10000**. In that case, SPSS will compare two groups **less** than 10000 and **more** than 10000, and we will do significant testing between these two groups. Currently, we are using our group definition, so click on **Continue** like this:

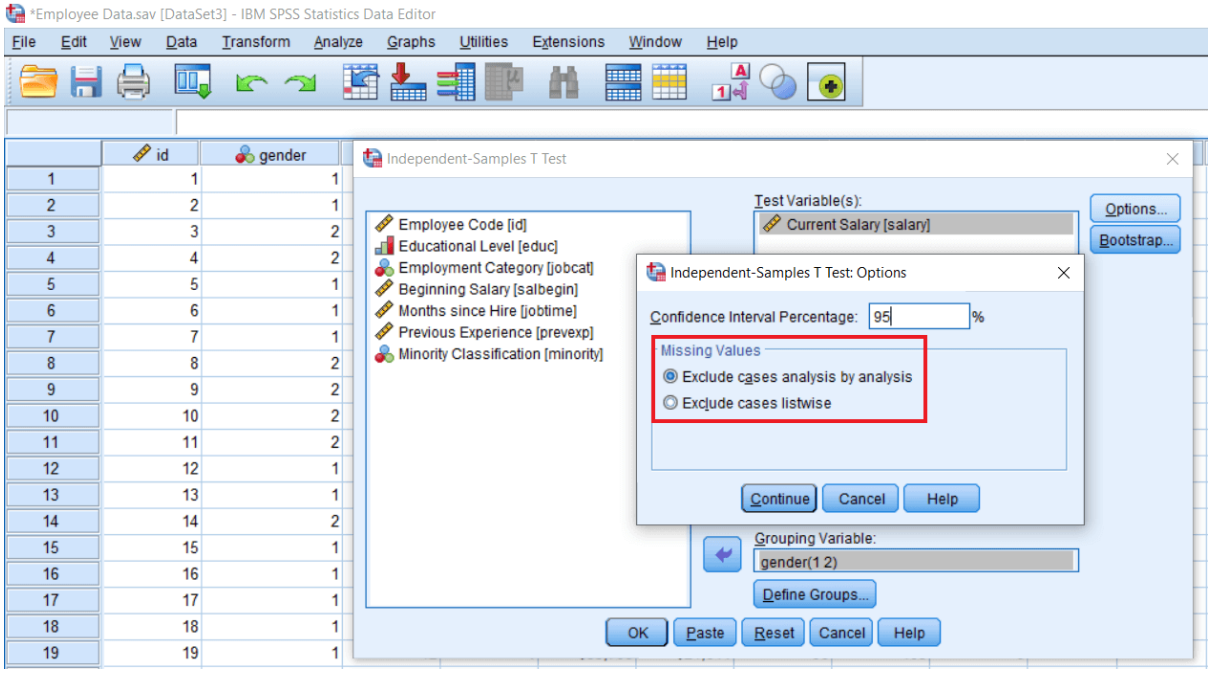


Now we will click on **Options** where we can select a **95% Confidence Interval** by default. If we want to change, we change it and make it as **99%**, but let's begin with the default value **95%**.

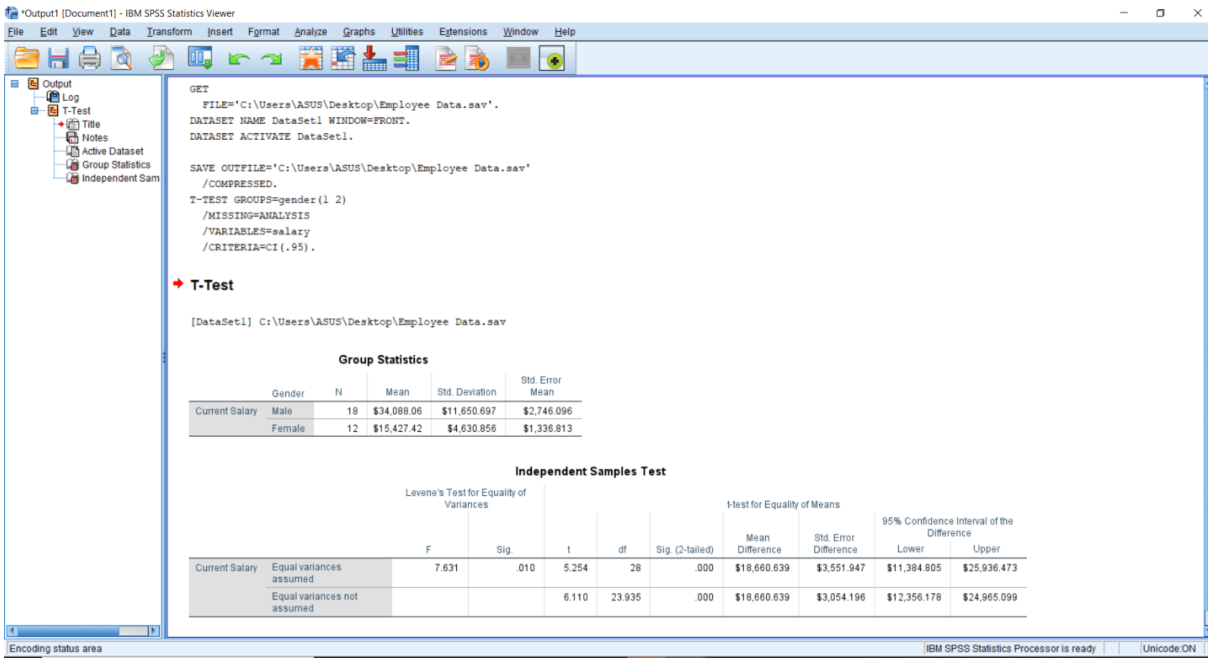


In the missing value, we are going with the default value **Exclude case analysis by analysis**. It basically leads to a lesser amount of data loss as compare to **Exclude cases**

listwise. So we will take it as an analysis by analysis method. Now we will click on **Continue**.

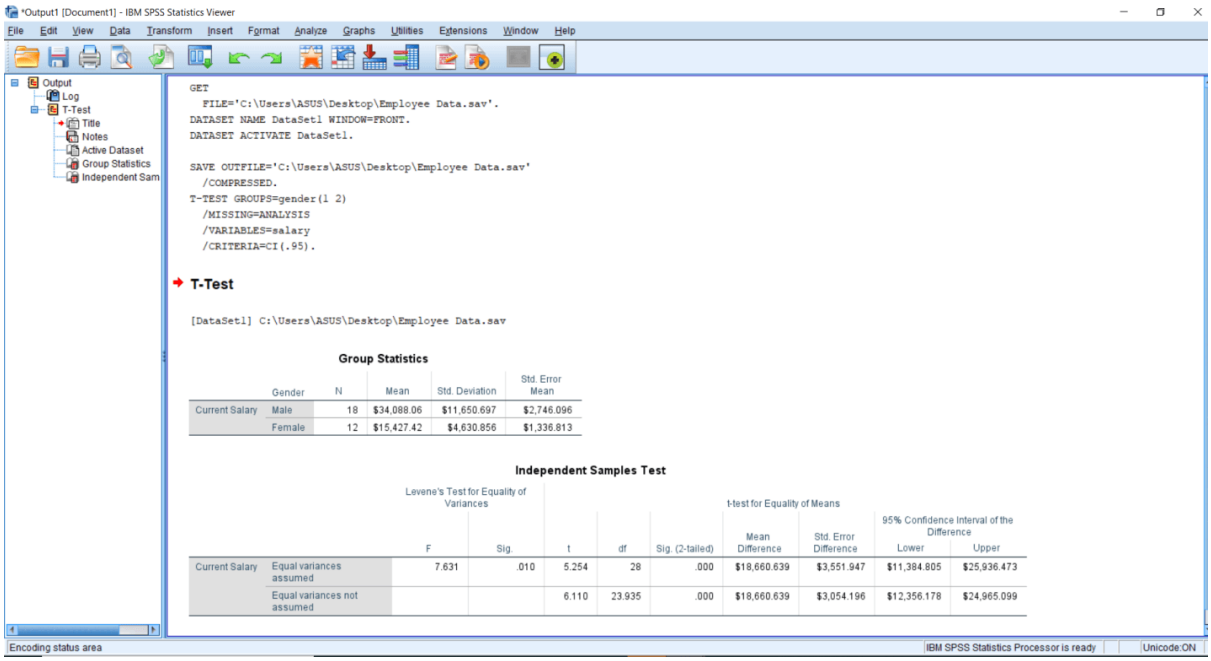


Now we will click on **Ok**, and after that, we will see the following Output:

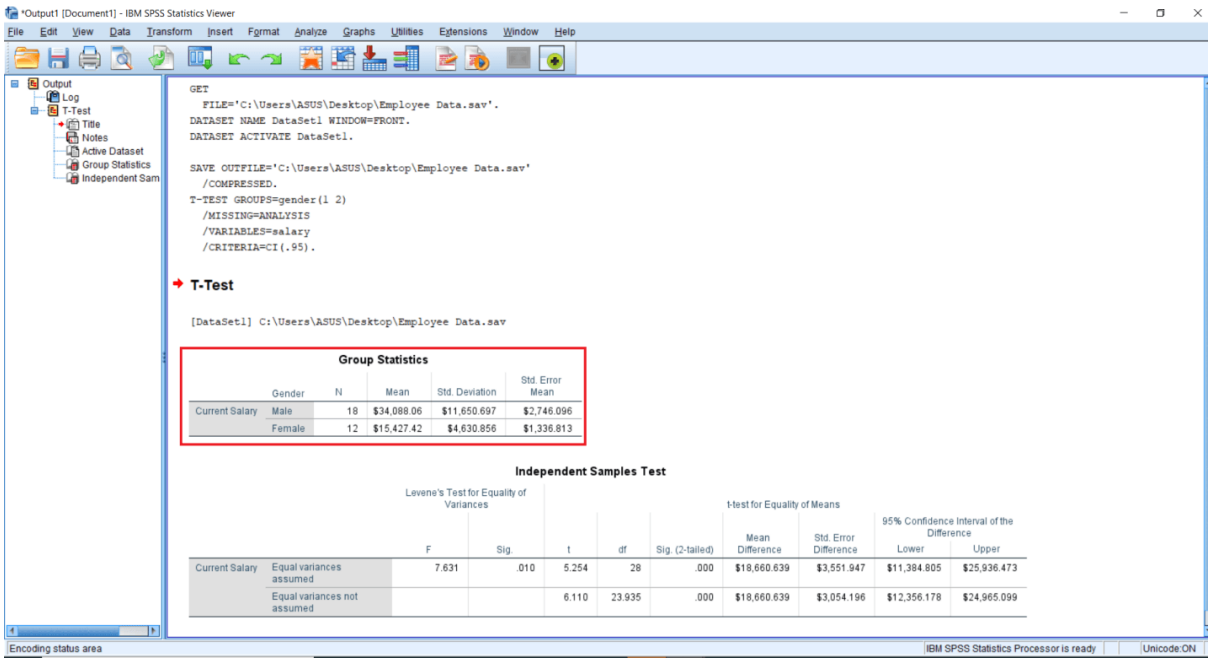


Output of Independent Sample T-test

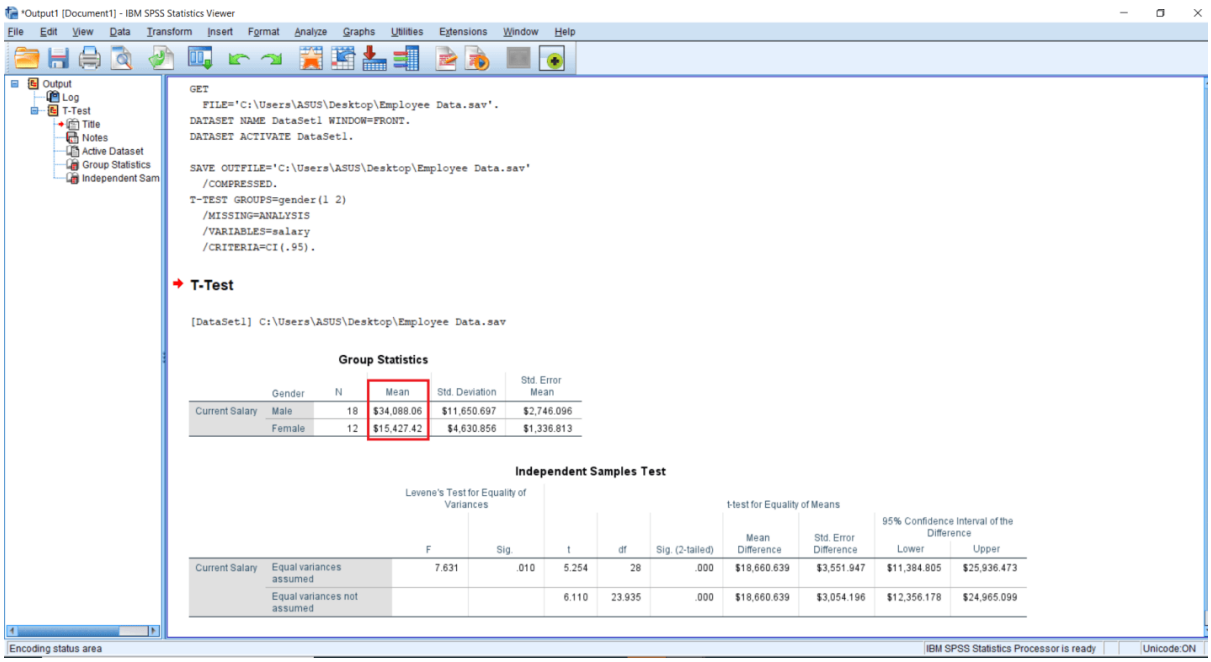
In this section, we will discuss the **Output** of the **Independent sample t-test**. The output of Independent sample t-test is given below, which is the output of the previous **Define Input option** file:



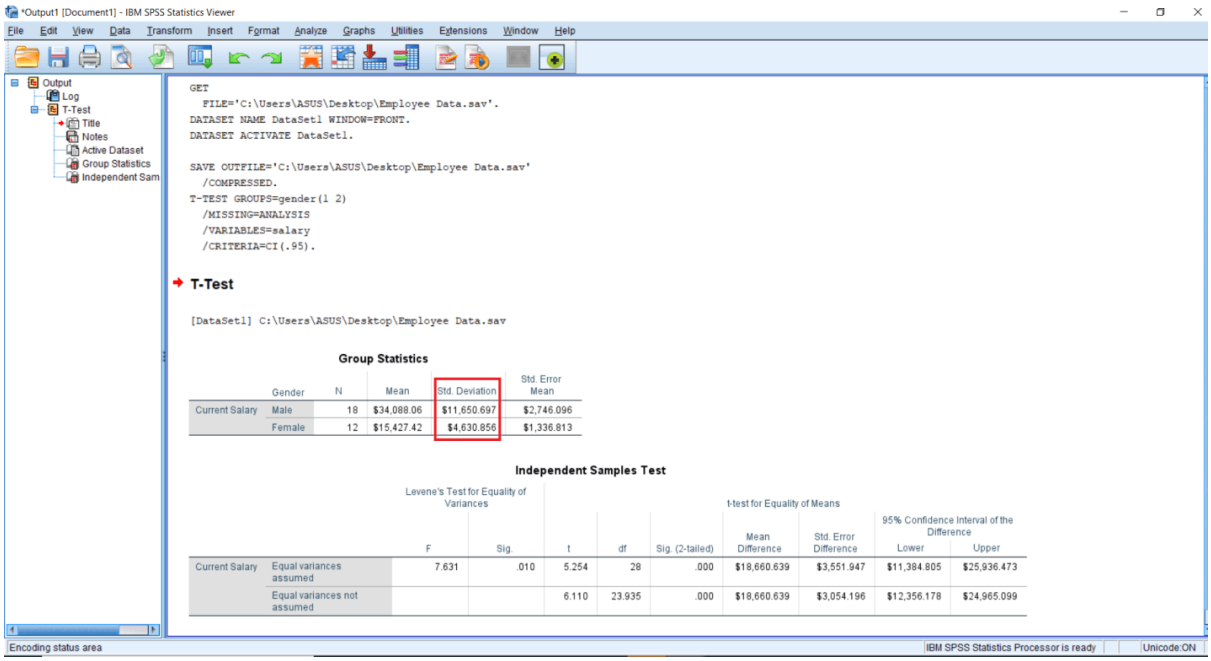
The first table is a **descriptive table**. So, it gives the **average salary**, no of **people** along with **means, Standard deviation, Standard Error mean**. We have **two groups** 1 and 2. **Current salary** is our **dependent** variable, and **Gender** is our **Independent** variable. We have **18 males** and **12 females**.



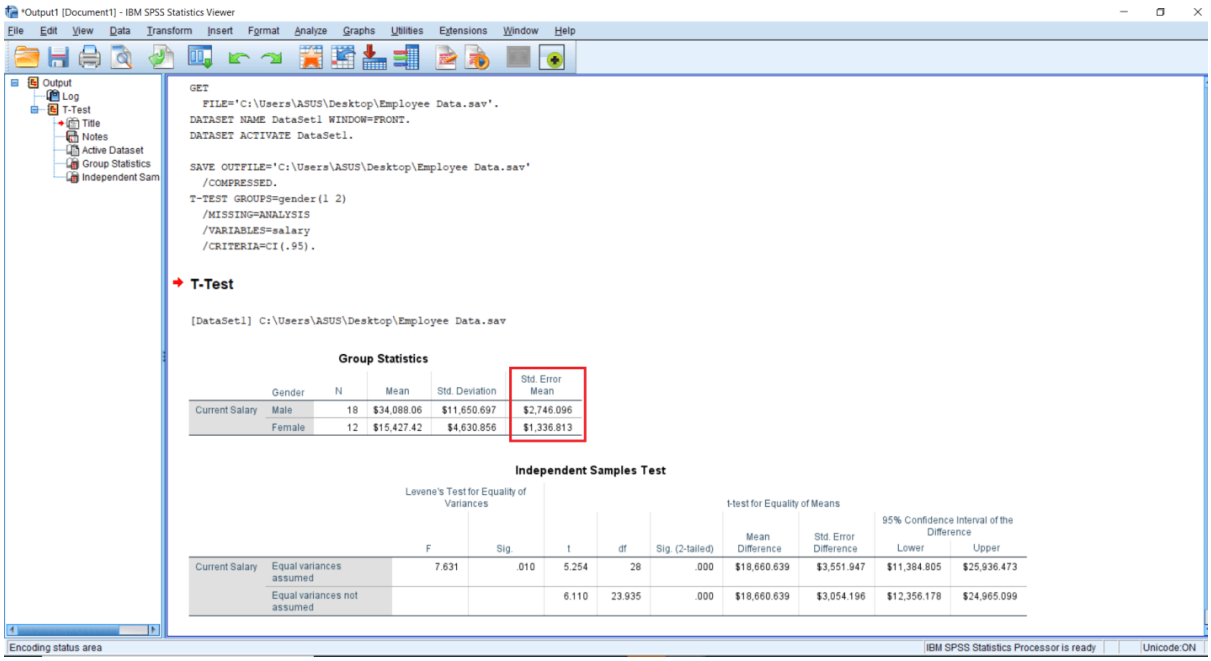
The average salary drawn by males is **34 thousand 008.06 dollars** huge compared to the females who are drawing on an average of **15 thousand 427.42 dollars**. So even if we don't do a significance testing and just look at these **Mean** figures, we are getting the impression that there is a huge difference between the **salary** of **males** and **females**. So we are expecting a significant effect of Gender on salary.



Then we have a **Standard Deviation** for these two groups. The males are drawing the higher salary again, but the variance of salary is quite high in the male group compared to the female group, in fact, more than twice.



The **Independent sample t-test** is an **inferential** test, and it is based on the sample characteristics. We are trying to draw **inferences** for the population. So **Standard Error of Mean** basically the first two standard deviations of various sample means that we draw from the **population**. Now imagine a situation where we are doing the study, and we are drawing a lot of **subsamples** from the population. We draw one sample like suppose we are having **10000 people** in the employee category, and out of those 10000, we randomly draw a **100 sample**. Now, this **100** will be having a **mean salary** and some **standard deviation**. Again we draw another **100** samples that will be having another **mean** and **standard deviation.** Similarly, we keep on drawing the samples from the population, and we are going to get **different means** for different samples. If there is a more variation in these samples that we are drawing, it means we are committing more amount of **Error** in the estimation of a population mean. So that's why we have a figure called **Standard error of Mean**. So that basically tells us the **average** amount of **error** when we draw n number of samples from the population. So our **Standard error of mean** is quite less than the **average salary**, but it's a bit **high** for the **male** group compared to the **female** group. Similarly, **Standard Deviation** is again **high** for **males** groups as compared to the **female** groups. It might be due to **outliers** since we did not eliminate the outliers, but it will come down if we eliminate the outliers.



All in all, we will keep in mind that our **Standard deviation** should not be more than our **Mean**. Similarly, our **Standard error** should not be more either **Standard Deviation** or **Mean**. So this is the interpretation for the **Descriptive statistics**. Generally, we are not supposed to report the entire thing, but only for the sake of understanding, we can know it.