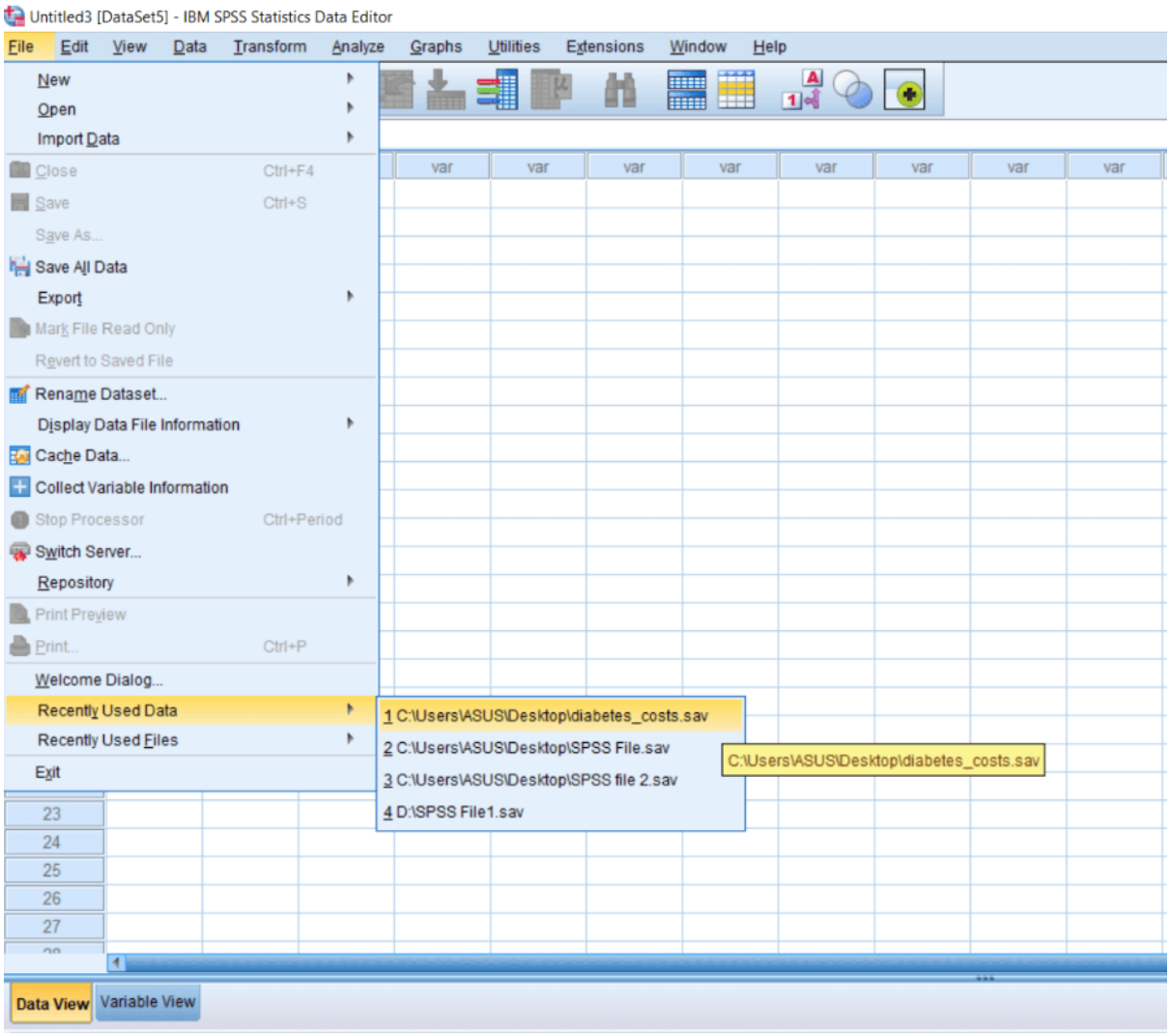
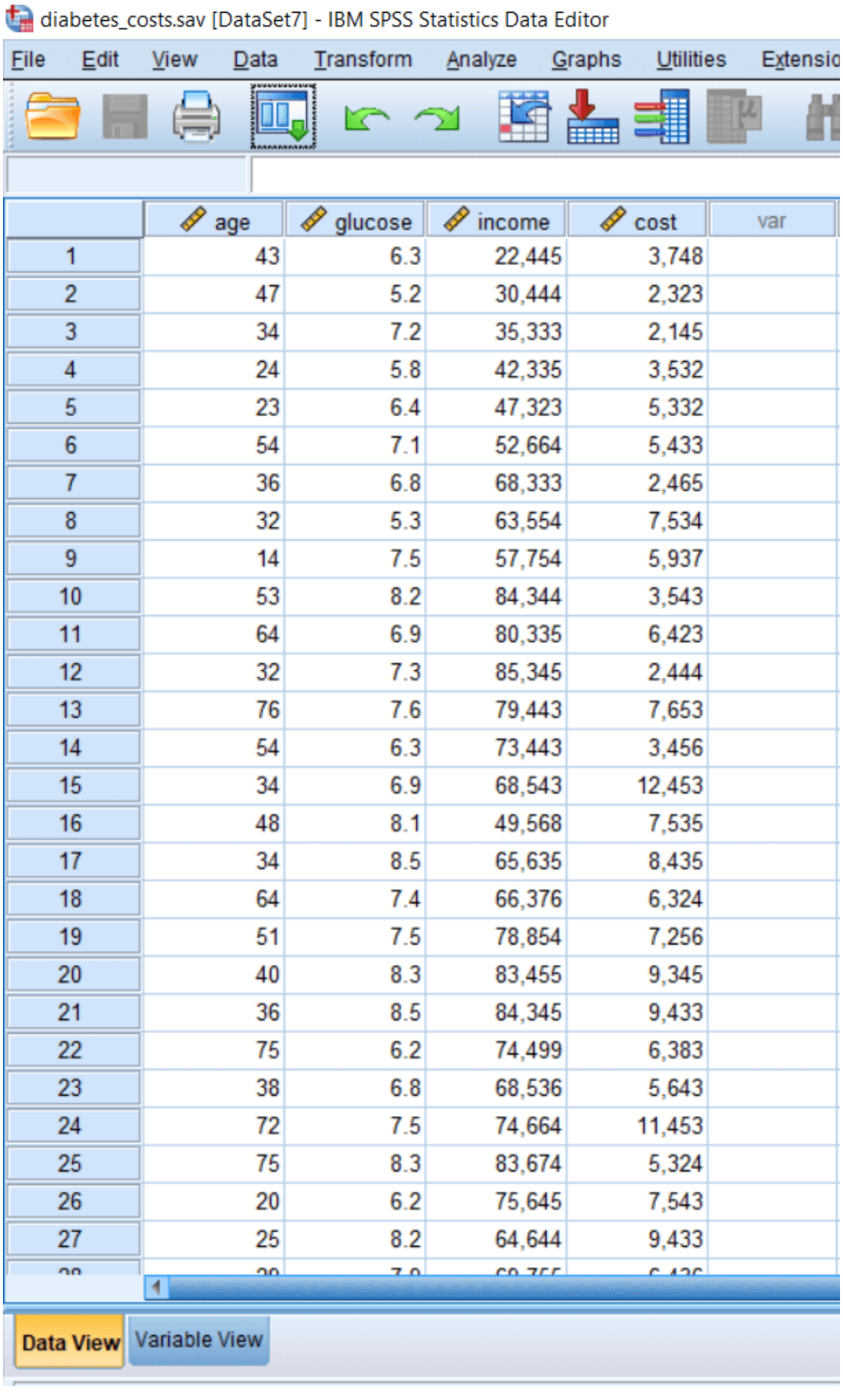
Descriptive Statistics

Setting Data for Descriptive Analysis

In this section, we will learn how we can calculate various shots of **Descriptive stats** using SPSS. **SPSS** is a wonderful software when we want to use **descriptive analysis** apart from the inferences analysis as well in our research. There are many ways through which we can calculate the **Descriptive analysis** in SPSS. Currently, we don't have any data. So first, we will **import** any data set in SPSS so that we can demonstrate how to use descriptive analysis. To import a new data set, we will go to the **File** menu, and then we can see many **data set** open here earlier. One of the data set is the **diabetes_costs** data set, as shown below:



It is a comparative **new data set** and less use. So we will use this data set. Click on this data set to open it. The following details will be open after click on this:



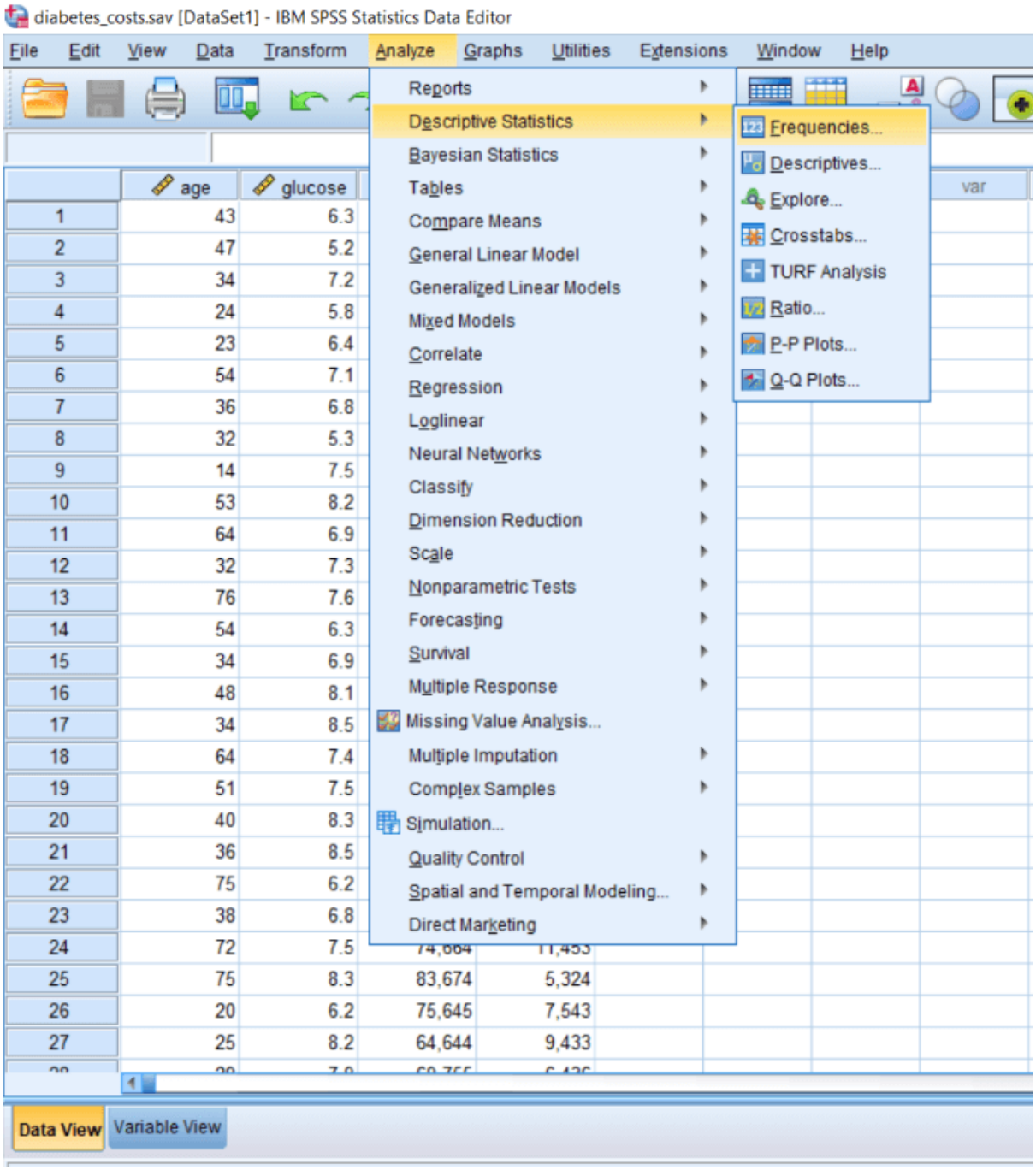
	age	glucose	income	cost	var
1	43	6.3	22,445	3,748	
2	47	5.2	30,444	2,323	
3	34	7.2	35,333	2,145	
4	24	5.8	42,335	3,532	
5	23	6.4	47,323	5,332	
6	54	7.1	52,664	5,433	
7	36	6.8	68,333	2,465	
8	32	5.3	63,554	7,534	
9	14	7.5	57,754	5,937	
10	53	8.2	84,344	3,543	
11	64	6.9	80,335	6,423	
12	32	7.3	85,345	2,444	
13	76	7.6	79,443	7,653	
14	54	6.3	73,443	3,456	
15	34	6.9	68,543	12,453	
16	48	8.1	49,568	7,535	
17	34	8.5	65,635	8,435	
18	64	7.4	66,376	6,324	
19	51	7.5	78,854	7,256	
20	40	8.3	83,455	9,345	
21	36	8.5	84,345	9,433	
22	75	6.2	74,499	6,383	
23	38	6.8	68,536	5,643	
24	72	7.5	74,664	11,453	
25	75	8.3	83,674	5,324	
26	20	6.2	75,645	7,543	
27	25	8.2	64,644	9,433	

This is our data set that currently we are going to use for the demonstration purpose for calculating various sorts of **descriptive stats**. This data set contains the **age** of the subject, **glucose** level, **income,** and the treatment **cost**. So it is about calculating the **treatment cost** of the subject based on the **age, glucose** level, and the **income** of the subject. So we can use this for various sorts of descriptive stats.

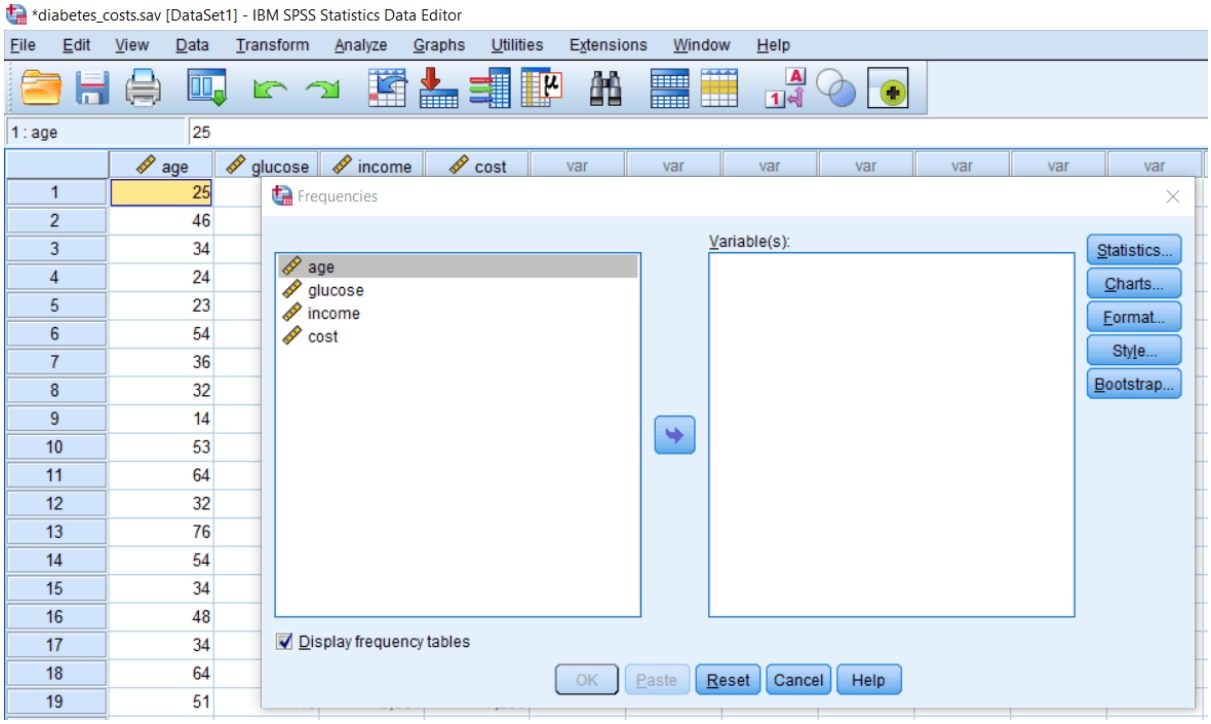
Descriptive stats can be useful in communicating the overall picture of our data set. **For example**, we might wish to communicate with someone and want to find the **average age** of the subject analysed in this data set and the **percentage** of the subjects below a particular age. For **example**, suppose we want to communicate that what **percentage** of subjects in this data set below **52** years or **60** years or any number of years. For this, we need to take the help of **percentiles**. If we need to communicate the **average age** of the subject, in that case, we wish to talk about the **mean** score, **average** scores. Then there are **median, modes, etc**.

We have a **glucose** level, **income**. So again, we might be interested in what is the **average in-come of the subjects which are there in the data set and what is the treatment cost.**

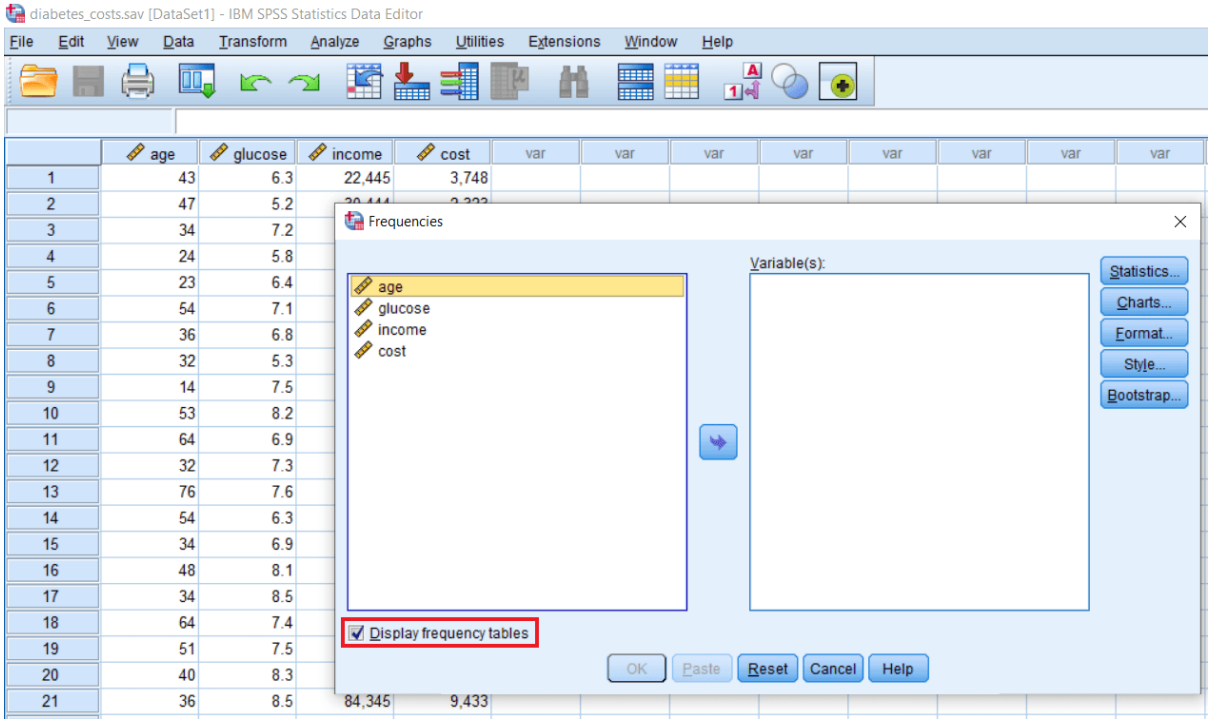
In this section, we will learn about the types of **Descriptive Statistics**. For this, we will go to the **Analyze** menu and then **Descriptive statistics**. Now we can see all the Descriptive statistics in the following image, and we can choose them according to our purpose.



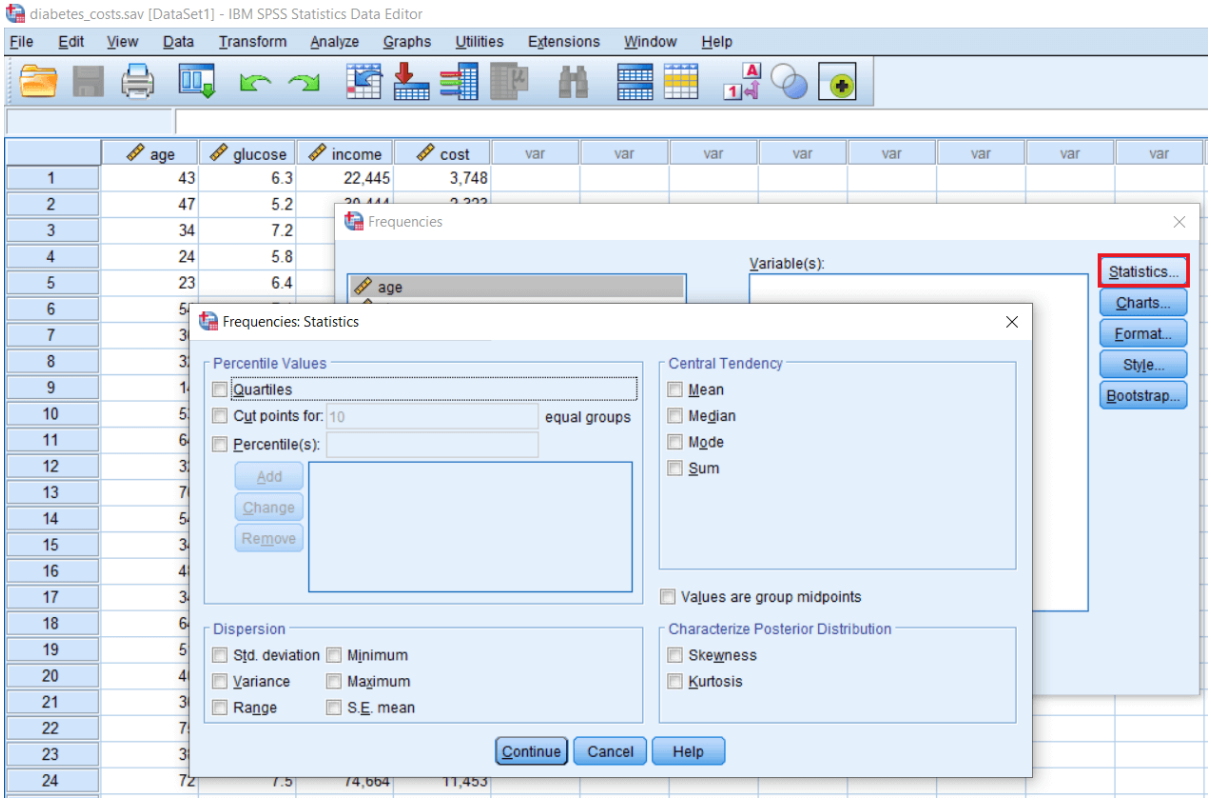
We can see that the first option is **Frequencies**, which is the most basic type of **Descriptive stats**. They simply tell the number of times a particular observation occurs in data. **For example**, we might be interested in knowing how many individuals in this data set are **25** years age, how many individuals are **46** years age or there is any repetition. There might be two-three or more people who are 25 years or 46 years. So for that, we may wish to calculate the **frequencies**. If we want to do that, click on **Frequencies**. Now we can see a **dialog box** like this:



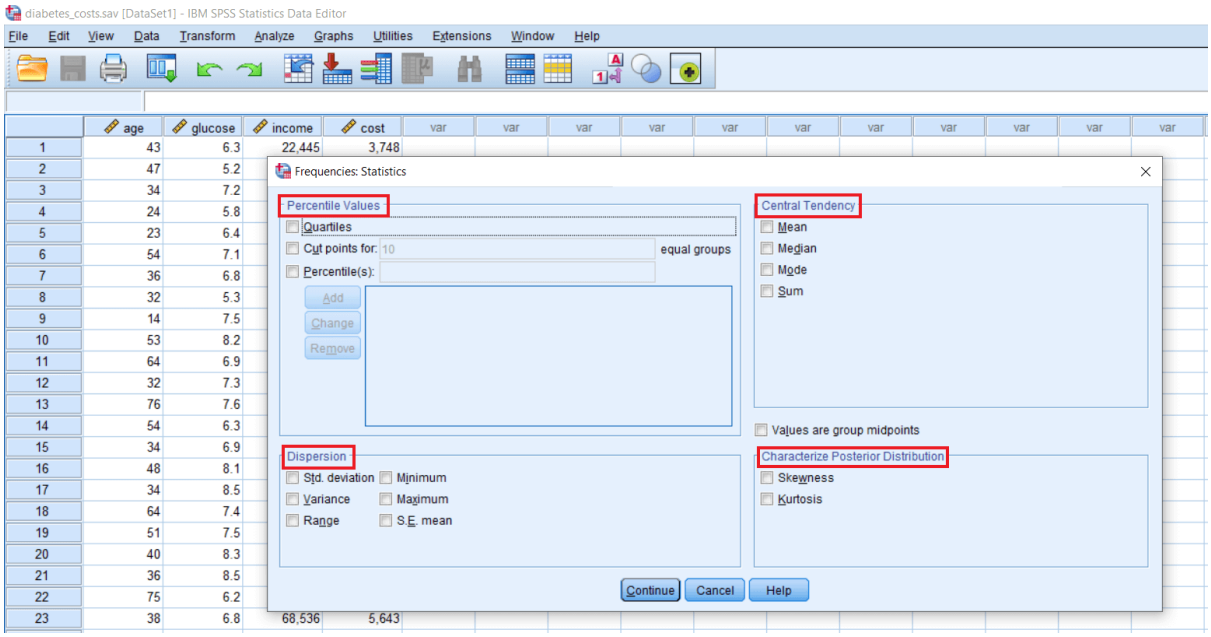
In the above dialog box, we can see an automatically checked box and that read as a **Display frequency table**. If we click on this, we are going to see the **frequency tables**.



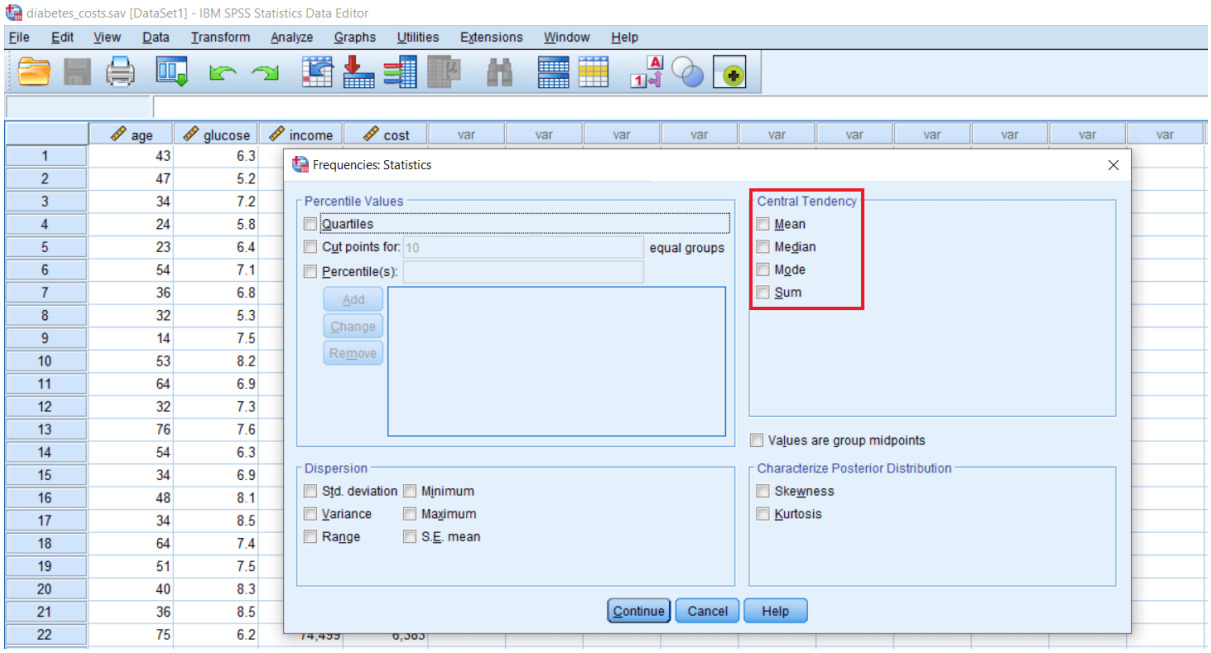
Apart from this, we have options like **Statistics**. When we click on this, we will see a dialog box like this:



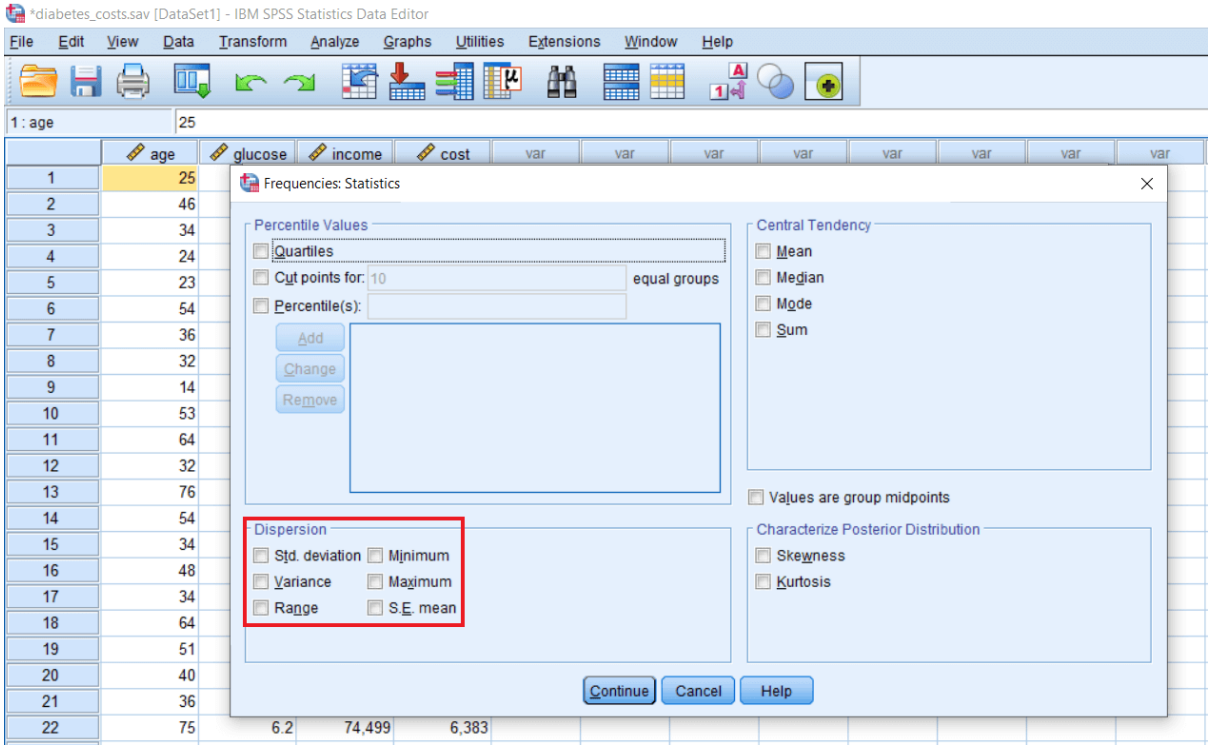
If we have no background in stats and just want to know what shots of descriptive stats are there, we can take the help of SPSS to teach ourselves the types of **descriptive stats**. In the above image, we can see how SPSS has categorized various types of descriptive stats for us. We have measures of **Central Tendency**. If we look at the dialog box, we can see there are four types of descriptive stats. The first one is a measure of **Central tendency**. The second one is a measure of **Dispersion**. The third one is a measure of **Distribution**. The fourth one is a measure of **Percentile scores**.



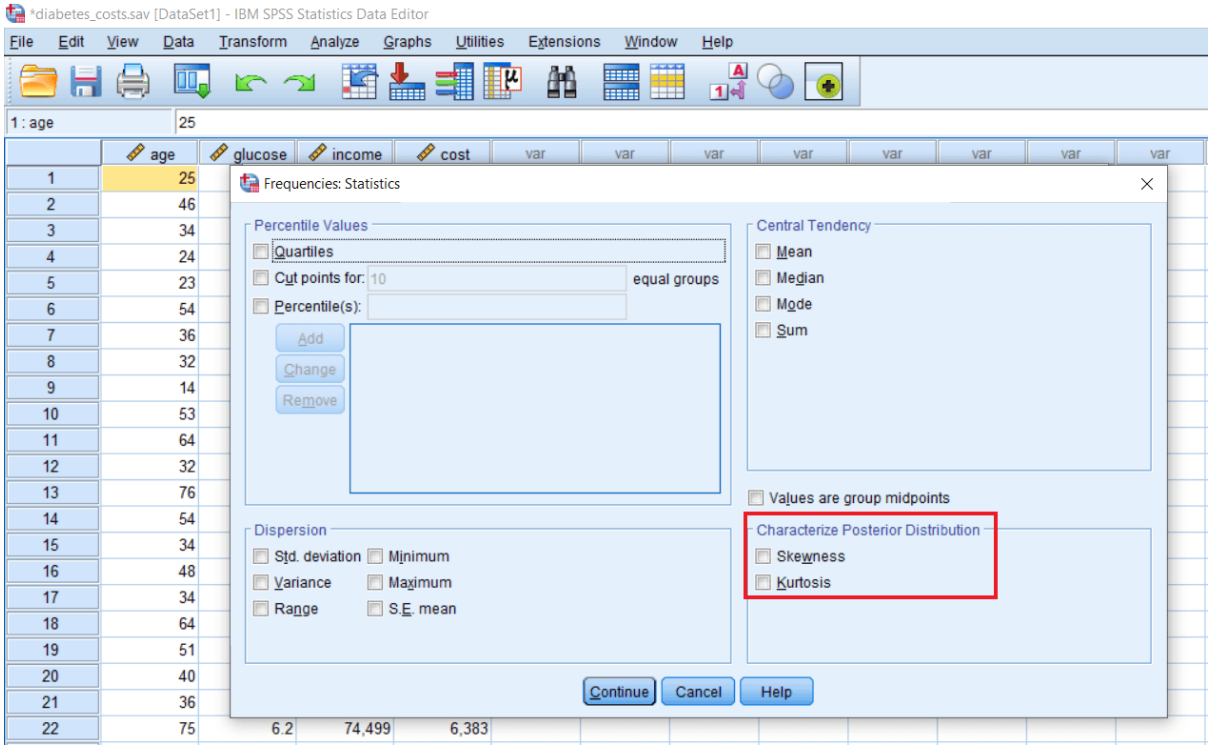
Central tendency shows us where our data is concentrated or where most of the observations lie. Central tendency contains **Mean, Median, Mode,** and **Sum** of all the scores.



Dispersion contains **Std. Deviation, Minimum, Maximum, Variance, Range, S.E. Mean** (standard error of the mean). If we are interested in communicating the spread of score or what extent scores are spread away from the mean or the central score, we can use the **measure of standard deviation. For example**, if we are in our **office**, we want to communicate that the **salary** of person A is **highly distinct** as compared to the salary of all other persons working in this office. If we say **distinct**, it might be **very less** or **very high**. It could be both the situation. So for that, we can look at the measure of **dispersion**. So, what is the standard deviation of salary scores in that particular office. So if the standard **deviation** is **high**, it means many individuals are earning very **high** in this office, and many individuals are earning very **less**. That's why we are having a huge spread of scores. There are **highs** and **lows,** so **minimum** and **maximum** would be huge. There will be a huge difference between the minimum and maximum scores. So that is going to spread our entire score in a **wide range**. So **Range** would again be **high. Range** basically tells us to what extent scores are dispersed or spread from each other.



Distribution contains **Skewness** and **Kurtosis**. If we want to communicate the **shape** of our distribution, we can use measures of **Distribution**. The shape of Distribution means how the **age** variable, **glucose** variable, or any variable that we are studying are distributed. If we plot them how they look and how we can visualize them, in that case, we have measures of **Distribution**. To measure, we have **Skewness** and **Kurtosis**.



Percentile score is one of the most fascinating commonly used measures of **Descriptive stats.** **Percentiles** are nothing but a type of score that tells us what percentage of people in a given set of

observations lie below a particular score. **For example**, if in this office, around **90%** of workers earn less than **10000 dollars**, so **90th** percentile is **10000** in this case. **Percentile** can be very **instructive** and can be used in many **competitive** examinations like **CAT** (common admission test for MBA). In that case, we don't get the percentages or averages. We will get the **percentile score**. It basically tells us the percentage of the candidates scoring below a particular candidate or particular score. To find the percentage, we will take the help of **percentiles** and **quartiles.** **Quartiles** are nothing but a type of percentiles. If we divide all the percentiles that can vary from 0 to 100 into 4 categories, 4 groups or 4 quarters, we call them quartiles.