

Data Screening and Cleaning.

The following written guide is relating to screening and cleaning the data that you have imputed or transferred onto SPSS. This should be conducted **BEFORE** you conduct any statistical analysis on your data set.

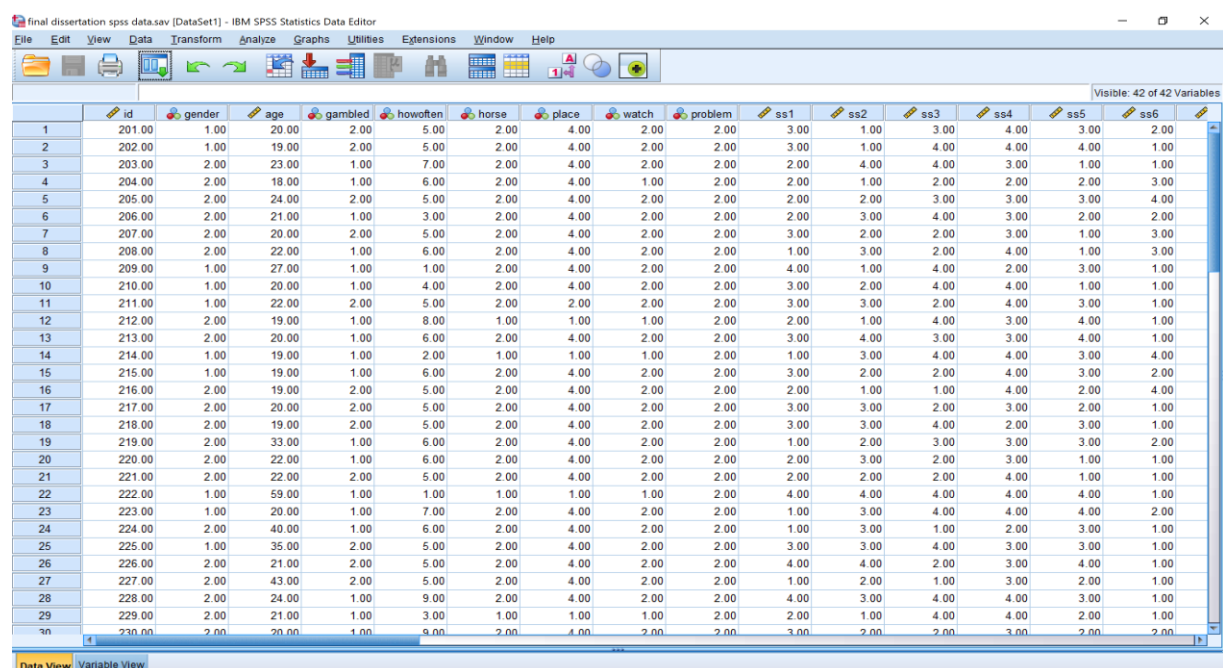
What is meant by screening and cleaning is checking the data sets for any errors and then correcting or dealing with the errors / outliers effectively. It is easy, particularly when you dealing with larger data sets, to make errors. For example a slip of the finger which means you have typed in 23 instead of 2 could have a major impact on the data set and subsequent analysis. Some analyses are particularly sensitive to outliers, spending the time to screen and then clean the data prior to conducting the analysis is far more effective than doing it later on. It can take time, particularly if you have a data set with a lot of variables, but it is really important.

Screening the data involves 2 processes:

- 1) Checking for any errors.
- 2) Finding and correcting the errors.

When checking for errors what you are likely to do is run some form of descriptive analysis to identify any values in your data set that are markedly different, wrong or completely inaccurate.

To start this session firstly the crib sheet will show how we generate descriptive statistics, then look at how the descriptive statistics can help clean the data. Using the data set below as a templated data set.

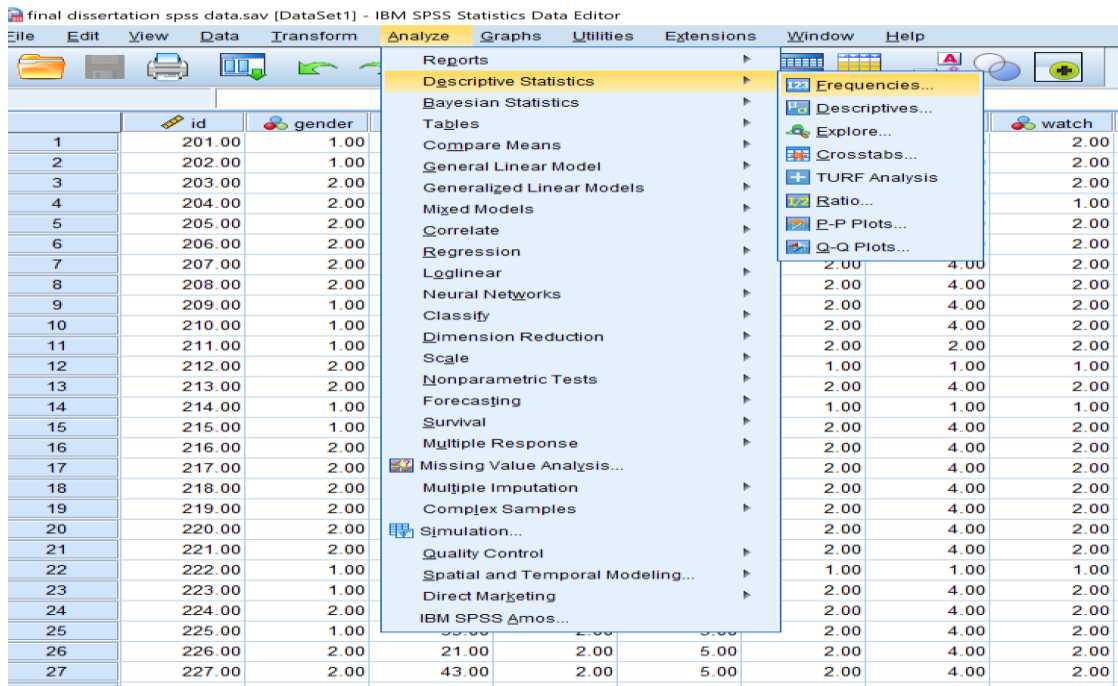


The screenshot displays the IBM SPSS Statistics Data Editor window. The title bar reads "final dissertation spss data.sav [DataSet1] - IBM SPSS Statistics Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window, and Help. The toolbar contains icons for file operations, editing, and analysis. The main data grid shows 30 rows (cases) and 42 columns (variables). The variables are: id, gender, age, gambled, howoften, horse, place, watch, problem, ss1, ss2, ss3, ss4, ss5, and ss6. The data values are numerical, with some cases showing missing values (e.g., 201.00, 202.00, 203.00, etc.). The status bar at the bottom indicates "Variable View" and "Visible: 42 of 42 Variables".

	id	gender	age	gambled	howoften	horse	place	watch	problem	ss1	ss2	ss3	ss4	ss5	ss6
1	201.00	1.00	20.00	2.00	5.00	2.00	4.00	2.00	2.00	3.00	1.00	3.00	4.00	3.00	2.00
2	202.00	1.00	19.00	2.00	5.00	2.00	4.00	2.00	2.00	3.00	1.00	4.00	4.00	4.00	1.00
3	203.00	2.00	23.00	1.00	7.00	2.00	4.00	2.00	2.00	2.00	4.00	4.00	3.00	1.00	1.00
4	204.00	2.00	18.00	1.00	6.00	2.00	4.00	1.00	2.00	2.00	1.00	2.00	2.00	2.00	3.00
5	205.00	2.00	24.00	2.00	5.00	2.00	4.00	2.00	2.00	2.00	2.00	3.00	3.00	3.00	4.00
6	206.00	2.00	21.00	1.00	3.00	2.00	4.00	2.00	2.00	2.00	3.00	4.00	3.00	2.00	2.00
7	207.00	2.00	20.00	2.00	5.00	2.00	4.00	2.00	2.00	3.00	2.00	2.00	3.00	1.00	3.00
8	208.00	2.00	22.00	1.00	6.00	2.00	4.00	2.00	2.00	1.00	3.00	2.00	4.00	1.00	3.00
9	209.00	1.00	27.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00	4.00	2.00	3.00	1.00
10	210.00	1.00	20.00	1.00	4.00	2.00	4.00	2.00	2.00	3.00	2.00	4.00	4.00	1.00	1.00
11	211.00	1.00	22.00	2.00	5.00	2.00	2.00	2.00	2.00	3.00	3.00	2.00	4.00	3.00	1.00
12	212.00	2.00	19.00	1.00	8.00	1.00	1.00	1.00	2.00	2.00	1.00	4.00	3.00	4.00	1.00
13	213.00	2.00	20.00	1.00	6.00	2.00	4.00	2.00	2.00	3.00	4.00	3.00	3.00	4.00	1.00
14	214.00	1.00	19.00	1.00	2.00	1.00	1.00	1.00	2.00	1.00	3.00	4.00	4.00	3.00	4.00
15	215.00	1.00	19.00	1.00	6.00	2.00	4.00	2.00	2.00	3.00	2.00	2.00	4.00	3.00	2.00
16	216.00	2.00	19.00	2.00	5.00	2.00	4.00	2.00	2.00	2.00	1.00	1.00	4.00	2.00	4.00
17	217.00	2.00	20.00	2.00	5.00	2.00	4.00	2.00	2.00	3.00	3.00	2.00	3.00	2.00	1.00
18	218.00	2.00	19.00	2.00	5.00	2.00	4.00	2.00	2.00	3.00	3.00	4.00	2.00	3.00	1.00
19	219.00	2.00	33.00	1.00	6.00	2.00	4.00	2.00	2.00	1.00	2.00	3.00	3.00	3.00	2.00
20	220.00	2.00	22.00	1.00	6.00	2.00	4.00	2.00	2.00	2.00	3.00	2.00	3.00	1.00	1.00
21	221.00	2.00	22.00	2.00	5.00	2.00	4.00	2.00	2.00	2.00	2.00	2.00	4.00	1.00	1.00
22	222.00	1.00	59.00	1.00	1.00	1.00	1.00	1.00	2.00	4.00	4.00	4.00	4.00	4.00	1.00
23	223.00	1.00	20.00	1.00	7.00	2.00	4.00	2.00	2.00	1.00	3.00	4.00	4.00	4.00	2.00
24	224.00	2.00	40.00	1.00	6.00	2.00	4.00	2.00	2.00	1.00	3.00	1.00	2.00	3.00	1.00
25	225.00	1.00	35.00	2.00	5.00	2.00	4.00	2.00	2.00	3.00	3.00	4.00	3.00	3.00	1.00
26	226.00	2.00	21.00	2.00	5.00	2.00	4.00	2.00	2.00	4.00	4.00	2.00	3.00	4.00	1.00
27	227.00	2.00	43.00	2.00	5.00	2.00	4.00	2.00	2.00	1.00	2.00	1.00	3.00	2.00	1.00
28	228.00	2.00	24.00	1.00	9.00	2.00	4.00	2.00	2.00	4.00	3.00	4.00	4.00	3.00	1.00
29	229.00	2.00	21.00	1.00	3.00	1.00	1.00	1.00	2.00	2.00	1.00	4.00	4.00	2.00	1.00
30	230.00	2.00	20.00	1.00	8.00	2.00	4.00	2.00	2.00	3.00	2.00	2.00	3.00	2.00	2.00

To start the screening process through looking at the frequency analysis.

Analyse >>>> Descriptive Statistics >>>> Frequencies.



The frequencies analysis will tell you the “how many” for the variables you have selected. For example, how many male participants and female participants if you select the gender variables. This is particularly useful as part of the screening process for **categorical data**. Lets check the frequencies for gender, in this case you have imputed the data as 1 = Males and 2 = Females.

So then in the dialogue box place gender in the variables box and then click OK.



What comes up next is the SPSS output for the gender frequencies.

➔ Frequencies

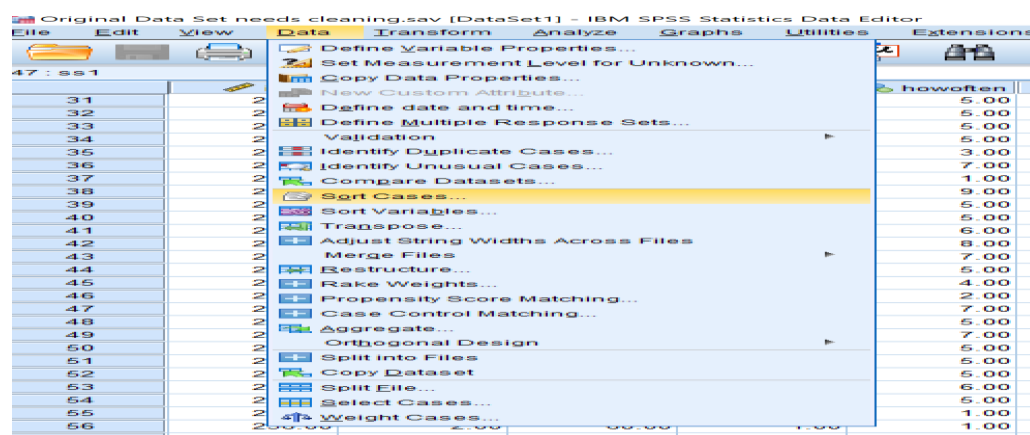
Statistics					
Gender					
N	Valid	66			
	Missing	0			

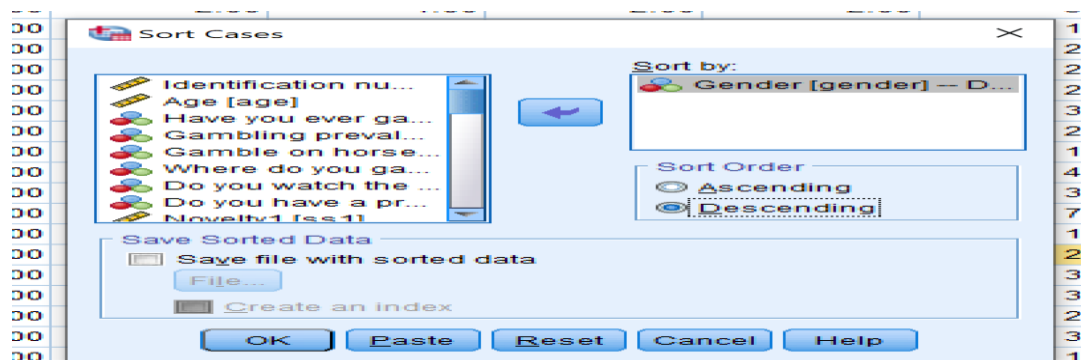
Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	19	28.8	28.8	28.8
	female	46	69.7	69.7	98.5
	5.00	1	1.5	1.5	100.0
	Total	66	100.0	100.0	

The first box tell us the total number of participants that had a value imputed. In this case all 66 participants had a value imputed so there was nothing blank. The second box is also important as this tells you how many males and how many female participants (according to the gender variable in the data set). Here you can see 19 Males and 46 Females but there is 1 participant labelled "5" this was not one of the options and is clearly an error in the data set, so this needs to be resolved. We need to firstly find this anomaly and replace it with the correct value.

The quickest way to do this is close the output and go back to the data set..... and then....

Data >>>>>> Sort Cases





We have spotted the error in the gender category so in the dialog box place gender in the “sort by” location and as we know the number 5 is higher than 1 or 2 check “Descending” sort order and then the participant with the number 5 under gender should be placed first. Next click ok.

You will then get an output box saying that this task has been done. Close this box.

Original Data Set needs cleaning.sav [Dataset] - IBM SPSS Statistics Data Editor

	id	gender	age	gambled	howoften	horse	place	watch
1	242.00	5.00	45.00	1.00	8.00	2.00	4.00	2.00
2	203.00	2.00	23.00	1.00	7.00	2.00	4.00	2.00
3	204.00	2.00	18.00	1.00	6.00	2.00	4.00	1.00
4	205.00	2.00	24.00	2.00	5.00	2.00	4.00	2.00
5	206.00	2.00	21.00	1.00	3.00	2.00	4.00	2.00

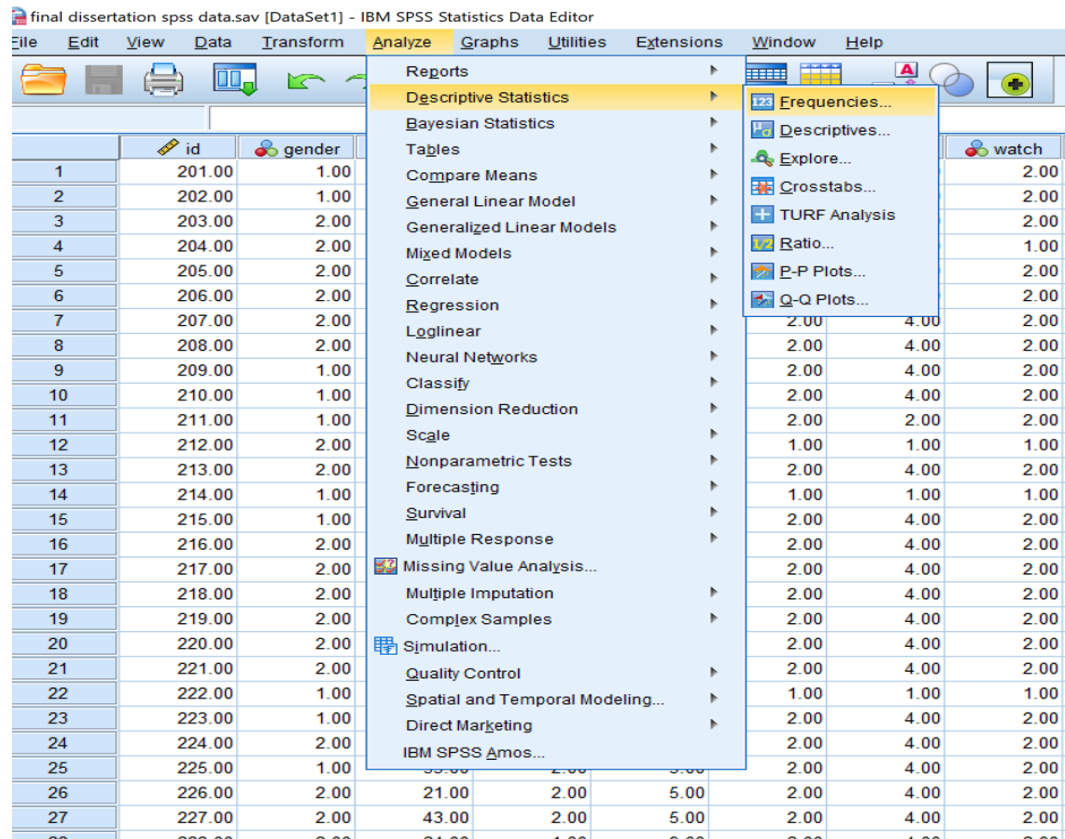
As you can see the data has now been reorganised so that those with the highest value in the gender category are placed on top. There you can see participant ID 242 has currently has a gender value of 5. This we know is incorrect – we check back with the original raw data for Participant 242 and we then note she has checked female – so we simply change that 5 to a 2 (which represents females). **THEN SAVE THE FILE**

Original Data Set needs cleaning.sav [Dataset] - IBM SPSS Statistics Data Editor

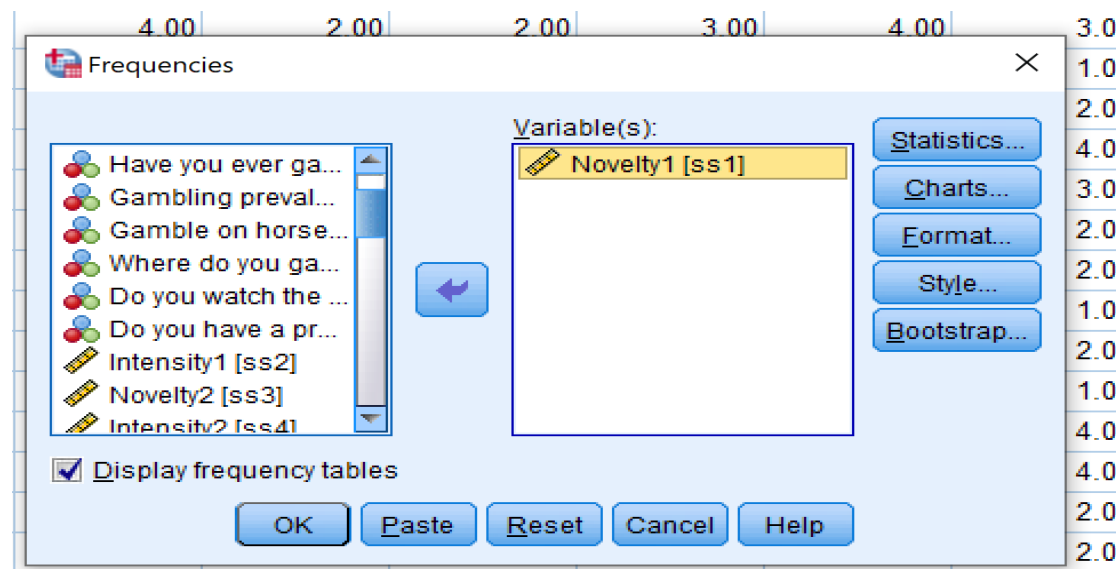
	id	gender	age	gambled	howoften	horse	place	watch
1	242.00	2.00	45.00	1.00	8.00	2.00	4.00	2.00
2	203.00	2.00	23.00	1.00	7.00	2.00	4.00	2.00
3	204.00	2.00	18.00	1.00	6.00	2.00	4.00	1.00

We can now be confident that there are no errors in imputing the gender variable in the dataset. For **all categorical variables** in the data set you would need to **repeat this process**. Let's do another example from this data set – namely the first question asked **SS1 variable**. This asked questions across a 4 point Likert scale so we know the values that should be in this variable are 1, 2, 3 or 4.

So again it is **Analyze >>>> Descriptive Statistics >>>> Frequencies**.



This time we are going to look at the SS1 data so in the dialogue box we place that as the variable of choice.



Then click Ok. You will then get the SPSS Output.

➔ Frequencies

Statistics

Novelty1

N	Valid	66
	Missing	0

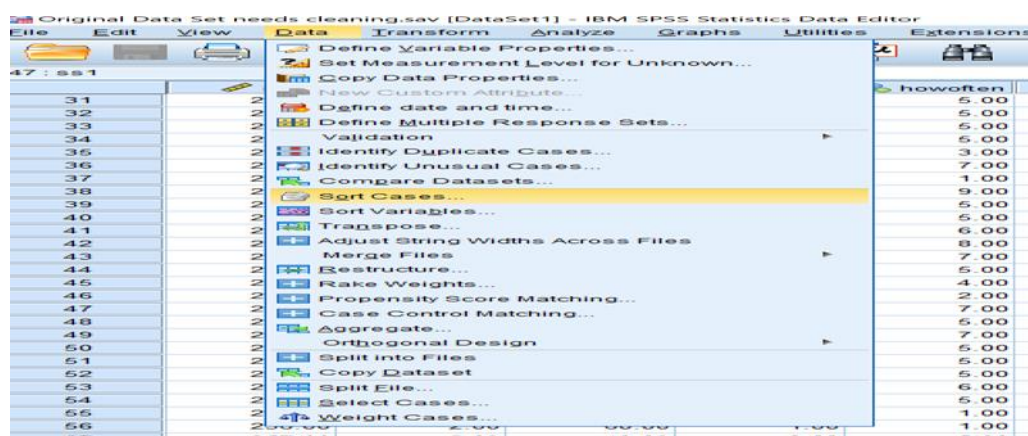
Novelty1

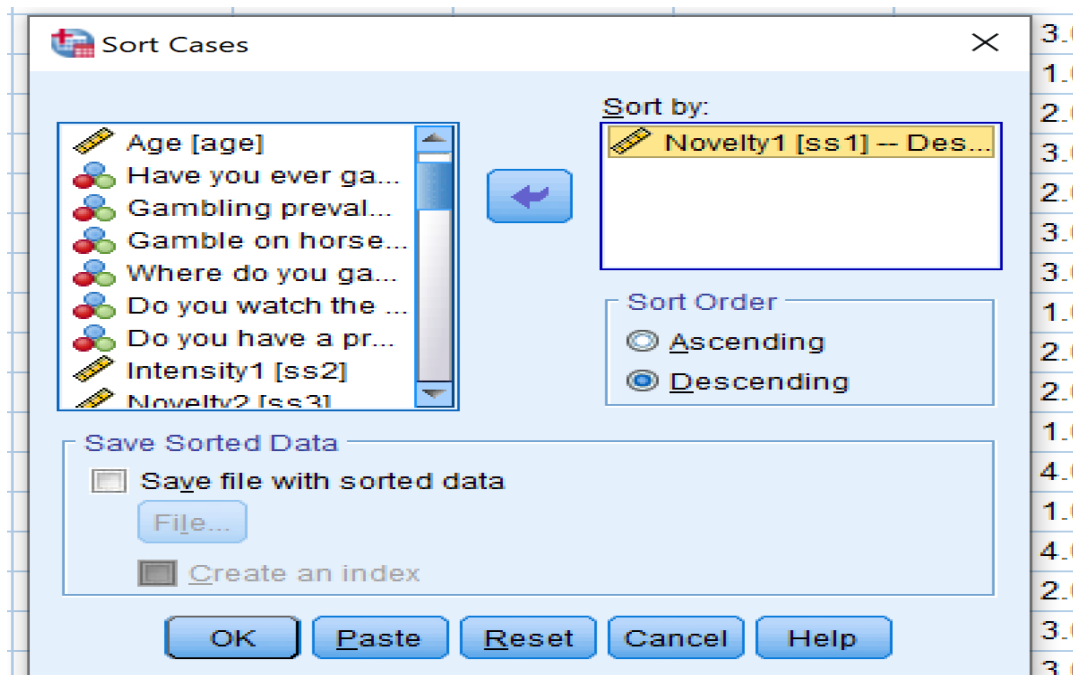
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	does not describe me at all	16	24.2	24.2	24.2
	does not describe me very well	21	31.8	31.8	56.1
	describes me somewhat	21	31.8	31.8	87.9
	describes me very well	7	10.6	10.6	98.5
	7.00	1	1.5	1.5	100.0
	Total	66	100.0	100.0	

As previously, the first box tell us the total number of participants that had a value imputed. So, in this case all 66 participants had a value imputed so there was nothing blank. The second box is also important as this tells you how many answered 1 for the question, 2 for the question, 3 for the question and 4 for the question. Here you can see 16 answered 1, 21 answered 2, 21 answered 3 and 7 answered 4 but there also 1 score of 7 and this was not one of the options and is clearly an error in the data set, so this needs to be resolved. We need to firstly find this anomaly and replace it with the correct value.

The quickest way to do this is close the output and go back to the data set..... and then....

Data >>>>>> Sort Cases





We have spotted the error in the SS1 category so in the dialog box place SS1 in the “sort by” location and as we know the number 7 is higher than 1, 2, 3, and 4 so check “Descending” sort order and then the participant with the number 7 under SS1 should be placed first. Next click ok.

You will then get an output box saying that this task has been done. Close this box.

*Original Data Set needs cleaning.sav [DataSet1] - IBM SPSS Statistics Data Editor

	id	gender	age	gambled	howoften	horse	place	watch	problem	ss1	ss2
1	245.00	1.00	21.00	1.00	4.00	1.00	2.00	1.00	2.00	7.00	1.00
2	226.00	2.00	21.00	2.00	5.00	2.00	4.00	2.00	2.00	4.00	4.00
3	228.00	2.00	24.00	1.00	9.00	2.00	4.00	2.00	2.00	4.00	3.00
4	209.00	1.00	27.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
5	222.00	1.00	59.00	1.00	1.00	1.00	1.00	1.00	2.00	4.00	4.00
6	243.00	1.00	21.00	1.00	7.00	2.00	4.00	2.00	2.00	4.00	4.00

As you can see the data has now been reorganised so that those with the highest value in the SS1 category are placed on top. There you can see participant ID 245 has currently has an SS1 value of 7. This we know is incorrect – we check back with the original raw data for Participant 245 and we then note the participant placed this score really as a 4 – so we simply change that 7 to a 4 – **THEN SAVE THE FILE.**

*Original Data Set needs cleaning.sav [DataSet1] - IBM SPSS Statistics Data Editor

	id	gender	age	gambled	howoften	horse	place	watch	problem	ss1	ss2
1	245.00	1.00	21.00	1.00	4.00	1.00	2.00	1.00	2.00	4.00	1.00
2	226.00	2.00	21.00	2.00	5.00	2.00	4.00	2.00	2.00	4.00	4.00
3	228.00	2.00	24.00	1.00	9.00	2.00	4.00	2.00	2.00	4.00	3.00
4	209.00	1.00	27.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00

We can now be confident that there are no errors in imputing the gender variable in the dataset. For **all categorical variables** in the data set you would need to **repeat this process**.

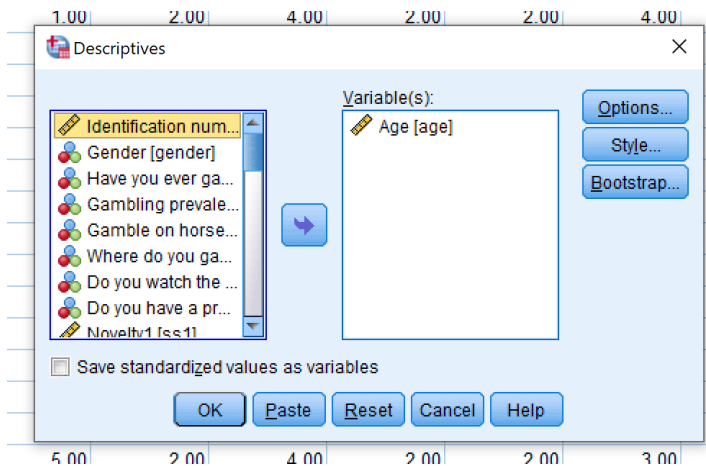
For continuous data – you can also screen and clean the data, again you will be needing to look at descriptive data but for continuous data the frequencies tab would **NOT** be appropriate for this data you would need to look at the **descriptives**. **Through looking at the descriptive statistics of continuous data we can see if the data range (maximum and minimum) falls within the acceptable levels.** Lets look at age (continuous data) as an example.

This time..... Analyze >>>>> Descriptive Statistics >>>>> Descriptives

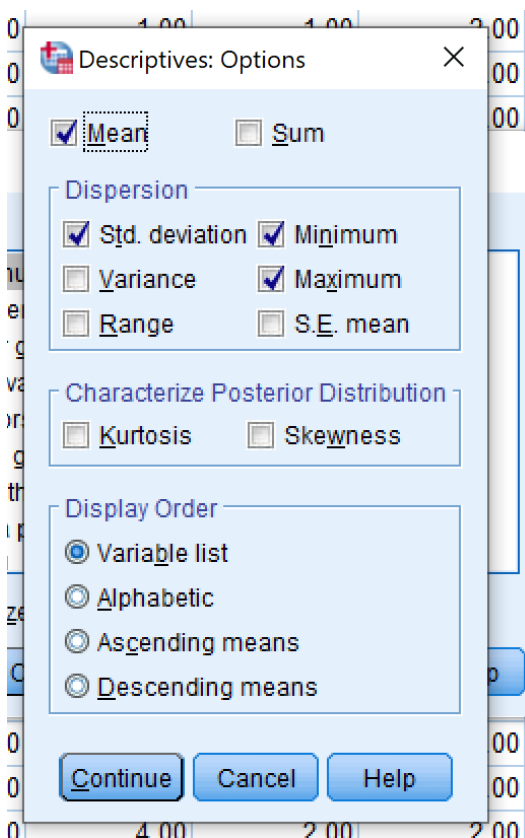
*Original Data Set needs cleaning.sav [DataSet1] - IBM SPSS Statistics Data Editor

	id	gender	age	gambled	howoften	horse	place	watch	problem	ss1	ss2
1	245.00	1.00	21.00	1.00	4.00	1.00	2.00	1.00	2.00	4.00	1.00
2	226.00	2.00	21.00	2.00	5.00	2.00	4.00	2.00	2.00	4.00	4.00
3	228.00	2.00	24.00	1.00	9.00	2.00	4.00	2.00	2.00	4.00	3.00
4	209.00	1.00	27.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
5	222.00	1.00	22.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
6	243.00	1.00	24.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
7	258.00	1.00	25.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
8	262.00	1.00	26.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
9	207.00	2.00	20.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
10	213.00	2.00	21.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
11	217.00	2.00	21.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
12	218.00	2.00	21.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
13	230.00	2.00	23.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
14	231.00	2.00	23.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
15	235.00	2.00	23.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
16	240.00	2.00	24.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
17	244.00	2.00	24.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
18	248.00	2.00	24.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
19	249.00	2.00	24.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
20	251.00	2.00	25.00	2.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
21	201.00	1.00	20.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
22	202.00	1.00	20.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
23	210.00	1.00	21.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
24	211.00	1.00	21.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00
25	215.00	1.00	21.00	1.00	1.00	2.00	4.00	2.00	2.00	4.00	1.00

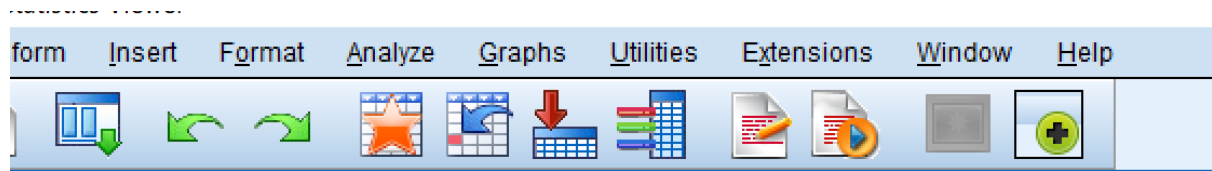
In the descriptives dialog box, place age in the variables section – then click options.



Under options, make sure the mean, std. deviation, minimum and maximum boxes are checked. Then click continue, then click ok.



This output box appears next



```
DESCRIPTIVES VARIABLES=age
  /STATISTICS=MEAN STDDEV MIN MAX.
```

➔ Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	66	18.00	60.00	25.3485	9.45840
Valid N (listwise)	66				

Here you can see that the minimum participant age (youngest participant) was 18 and the oldest was 60. You can also see that all 66 participants had a score for age that fell between 18 and 60. If however, you had a minimum score of less than 18 and you knew all participants were 18 or older then an error has been committed and you would need to clean the data in the manner described above using data >>>> sort cases. Here though you would be asking SPSS to place the order in ascending order as then the smallest number of age would come first.