STA10003 FOUNDATIONS OF STATISTICS WORK INTEGRATED LEARNING (WIL) ASSIGNMENT PART 1

This **Assignment Part 1** is worth **20%** of your final mark for STA10003.

Scenario

You have been employed as a new graduate researcher at Social Interactions Australia. Social Interactions Australia conducts research about relationships and other social issues. You have been given a dataset which contains the results of a survey given to Melbourne adults who visited a psychologist in 2022. You have been asked to analyse the data and answer several questions of interest that are presented on the following pages.

Data Preparation

For your assignment you should use the data set **STA10003 Sem 2 2024 Assignment Data.sav** provided in **Week 06: Assignment Part 1 Instructions & Data File** which can be accessed within the Week 06: Assignment Part 1 page. You must use SPSS to draw a random sample of 5000 cases from the 6000 cases in the data file. You will conduct your analysis on this sample of 5000 cases. Instructions on how to generate your random sample are on pages 5 to 8 of this document.

Submission Instructions

- Your submission must be a single Word file or PDF file.
- Although a cover page is not required, you should include your name and student number within the document [e.g., in footer].
- You must submit your file via the **SUBMIT ASSIGNMENT** button on the Week 06: Assignment Part 1 page in Canvas by Sunday September 8 by 11:59pm.
- Only the last document you submit will be marked.
- Once submitted, please review your submission to ensure the correct file has been submitted.
- This is an individual assignment. Do not share your work with other students. They will have a different sample of data, so any copying will be detected.

Plagiarism and Maintaining Academic Integrity

This is an individual assignment, so you are expected to complete it by yourself. It is important that you demonstrate good Academic Integrity by ensuring that your assignment work is entirely your own, because this shows that you have understood what you have learnt. You should not use the work of anyone else, including that of another students, for this assignment. Whilst you may seek help if there is anything about this assignment you do not understand, the assignment must be your own work.

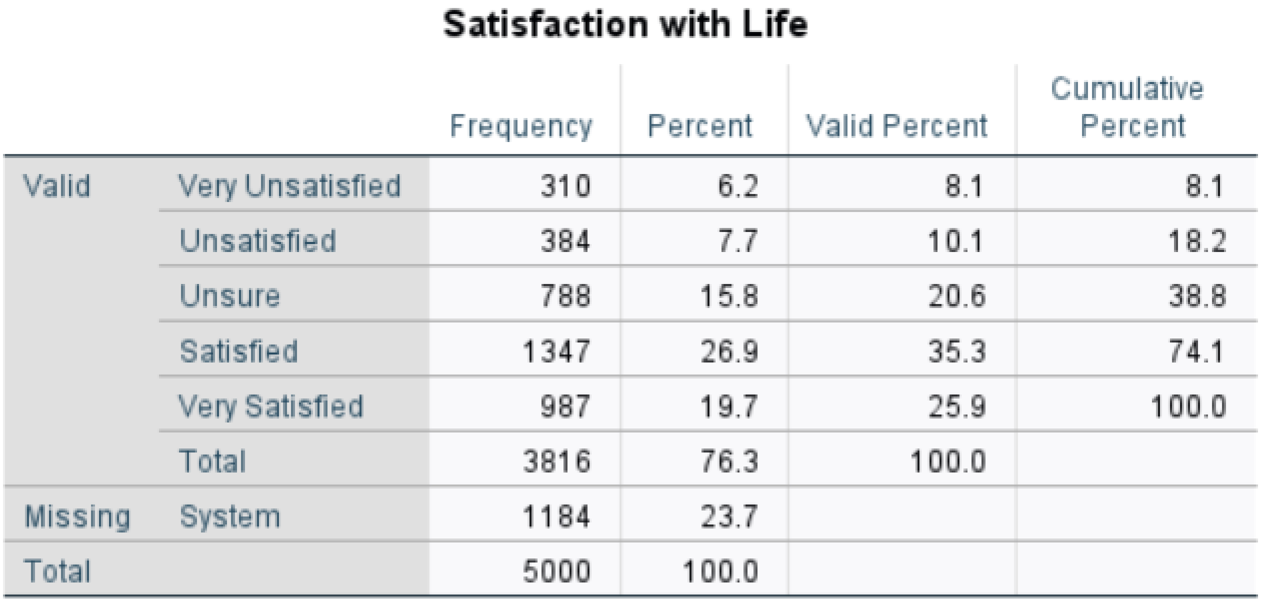
Use of generative AI such as ChatGPT is not permitted. Any form of copying or submitting work that is not entirely done by you is a breach of academic integrity and could attract academic penalties. Your work will be checked for breaches of academic integrity.

For your Assignment Part 1, you are required to complete the first three (3) questions by producing the appropriate analyses using SPSS and writing the relevant report for each question. You are also required to complete questions 4 and 5, which contain short answer questions.

For each of the first three questions requiring SPSS, you should include the relevant output immediately following your report. Graphs which are part of the report should be included within the report as shown in the report writing examples used in the course materials – as shown in the lectures and in tutorials. The document *Reading B: Reporting Information about Single Variables* also has more report writing examples.

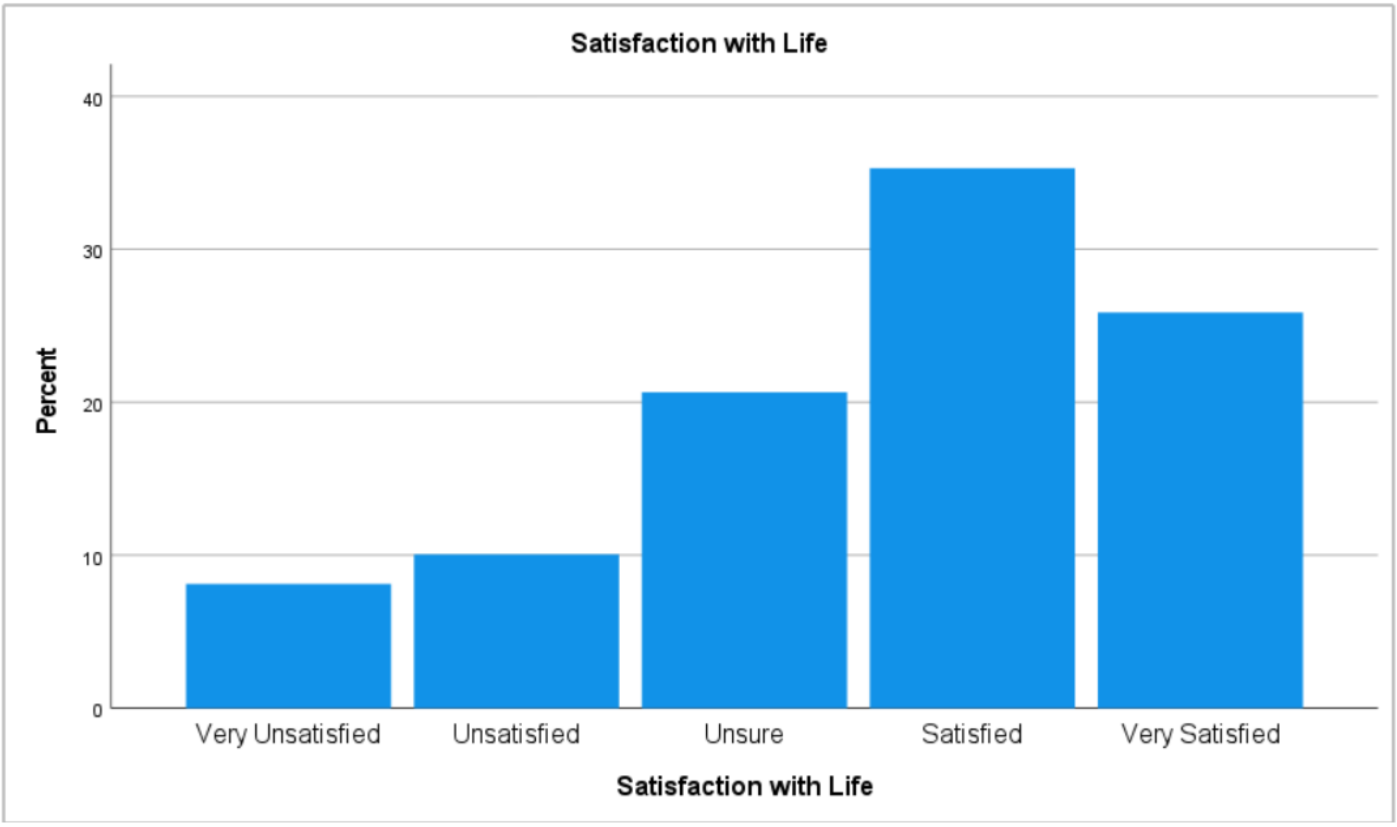
Question 1: Satisfaction with life

The variable Life measures the level of satisfaction with life. Using SPSS, produce the relevant graph and table to summarise the Life variable and write a paragraph explaining the key features of the data observed in the output in the style presented in the course materials.



Satisfaction with Life

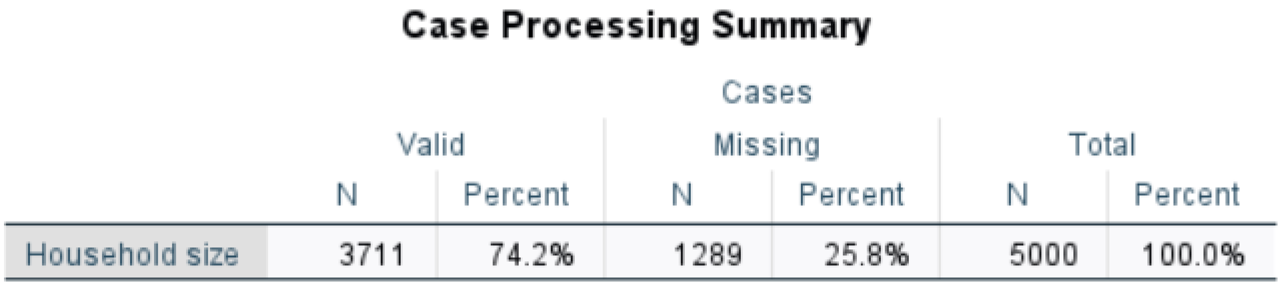
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very Unsatisfied	310	6.2	8.1	8.1
	Unsatisfied	384	7.7	10.1	18.2
	Unsure	788	15.8	20.6	38.8
	Satisfied	1347	26.9	35.3	74.1
	Very Satisfied	987	19.7	25.9	100.0
	Total	3816	76.3	100.0	
Missing	System	1184	23.7		
Total		5000	100.0		



The provided bar graph demonstrates the distribution of life satisfaction among the 3,816 respondents from a total sample of 5,000. It is clear that "Satisfied" emerged as the highest response during the survey, approximately 35.3% while "Very Satisfied" was substantially lower, precisely 25.9%. Additionally, those who expressed "Unsure" about their life followed by 20.6%. Moreover, the percentage of people reported "Unsatisfied" and "Very Unsatisfied" with their life displayed a lower ratio, with 10.1 and 8.1, respectively.

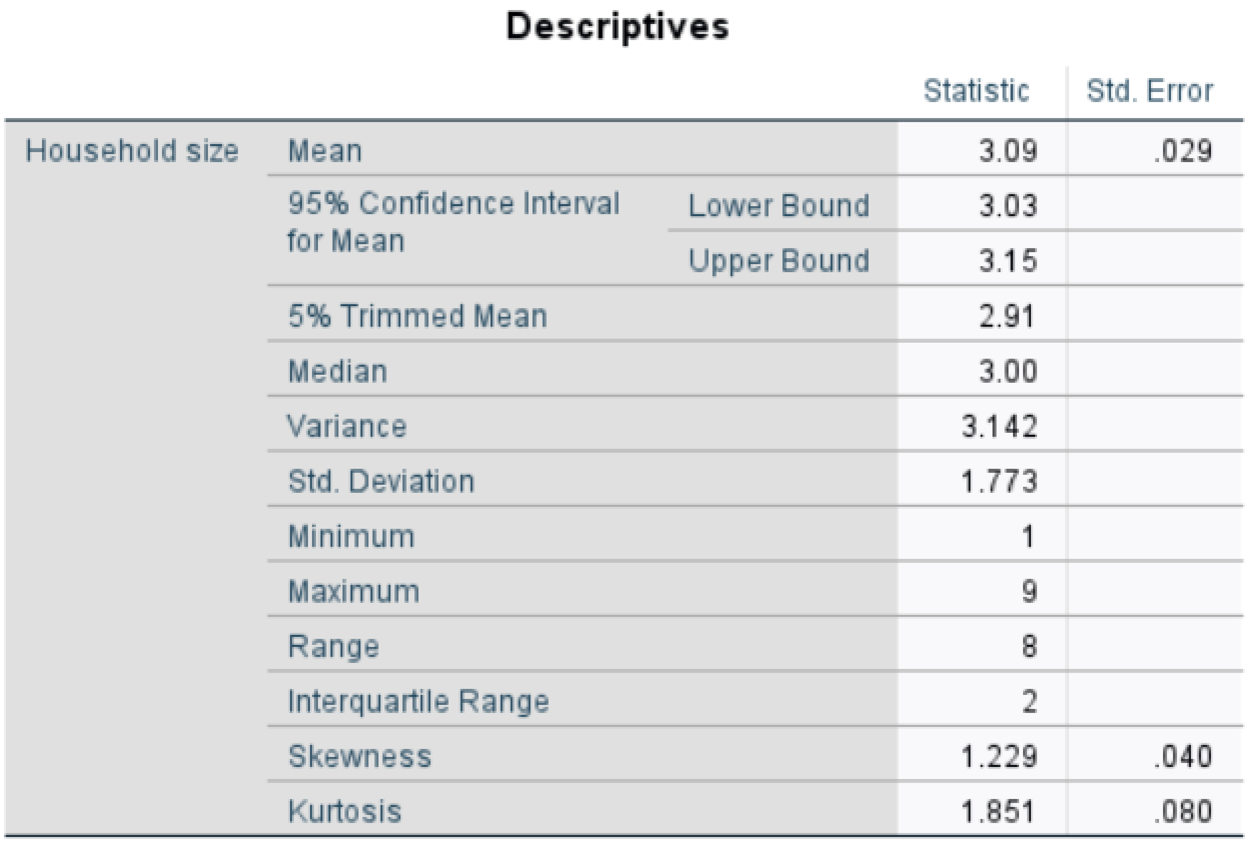
Question 2: Household size

The variable Household measures the household size. Using SPSS, produce the relevant graph and tables to summarise the Household variable and write a paragraph explaining the key features of the data observed in the output in the style presented in the course materials.



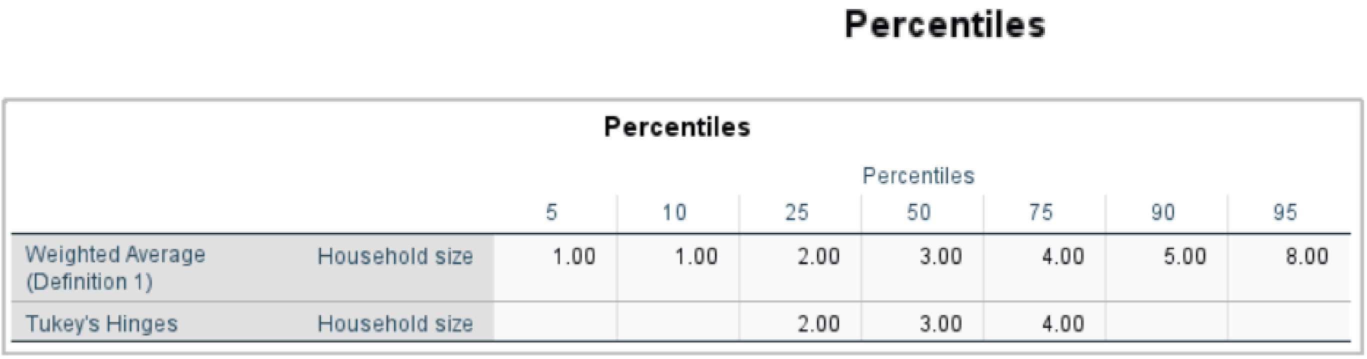
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Household size	3711	74.2%	1289	25.8%	5000	100.0%



Descriptives

			Statistic	Std. Error
Household size	Mean		3.09	.029
	95% Confidence Interval for Mean	Lower Bound	3.03	
		Upper Bound	3.15	
	5% Trimmed Mean		2.91	
	Median		3.00	
	Variance		3.142	
	Std. Deviation		1.773	
	Minimum		1	
	Maximum		9	
	Range		8	
	Interquartile Range		2	
	Skewness		1.229	.040
	Kurtosis		1.851	.080

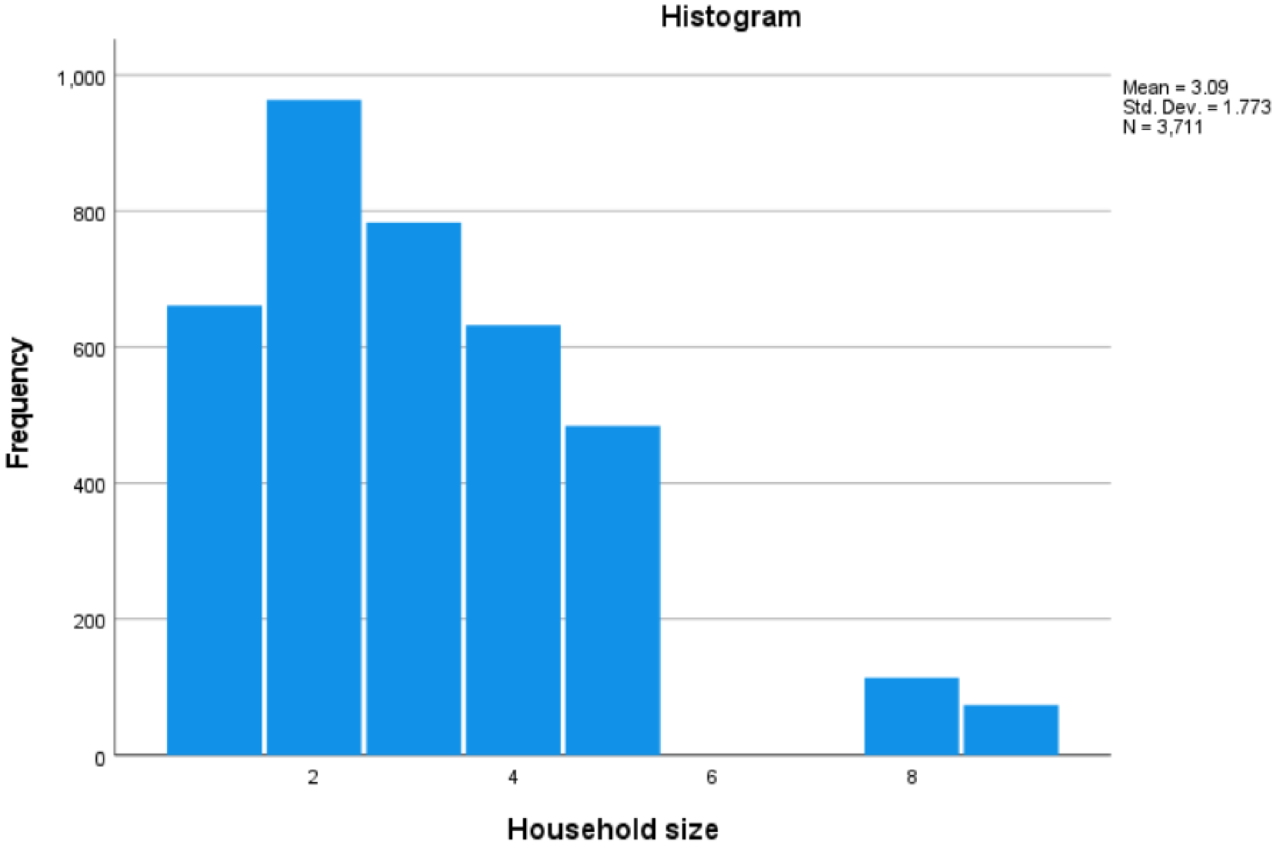


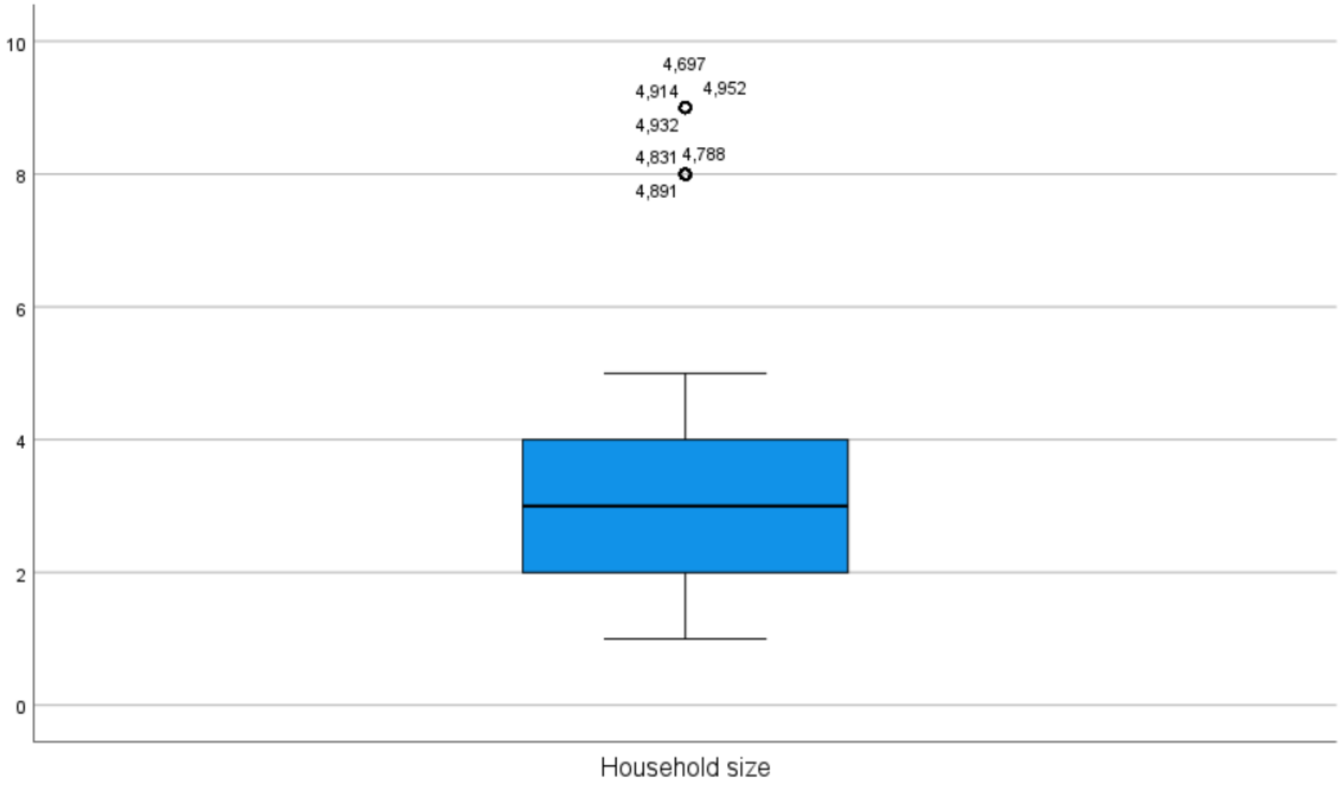
Percentiles

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Household size	1.00	1.00	2.00	3.00	4.00	5.00	8.00
Tukey's Hinges	Household size			2.00	3.00	4.00		

Household size

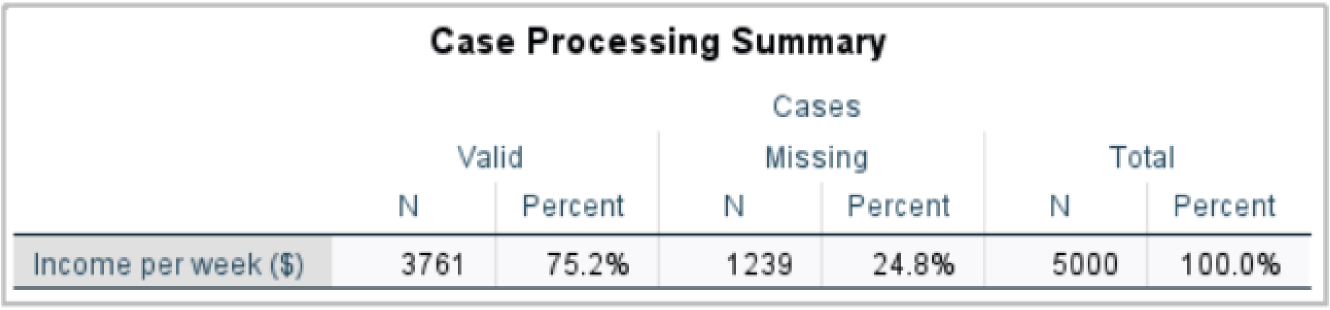




The distribution of household size within a sample of 5000 cases is visualized in the accompanying figures. It can be seen that the given histogram exhibited a positive skewness, with the percentage of household size being 74.2% and n=3711. Additionally, due to the median size being 3, half of the households had 3 members or less, while the 50% interquartile range comprised households from sizes 2 to 4, highlighting the central tendency of household size within this sample. Furthermore, the presence of a rare household with 9 members underscores the extent of the positive skew and, consequently, the presence of outliers in the distribution.

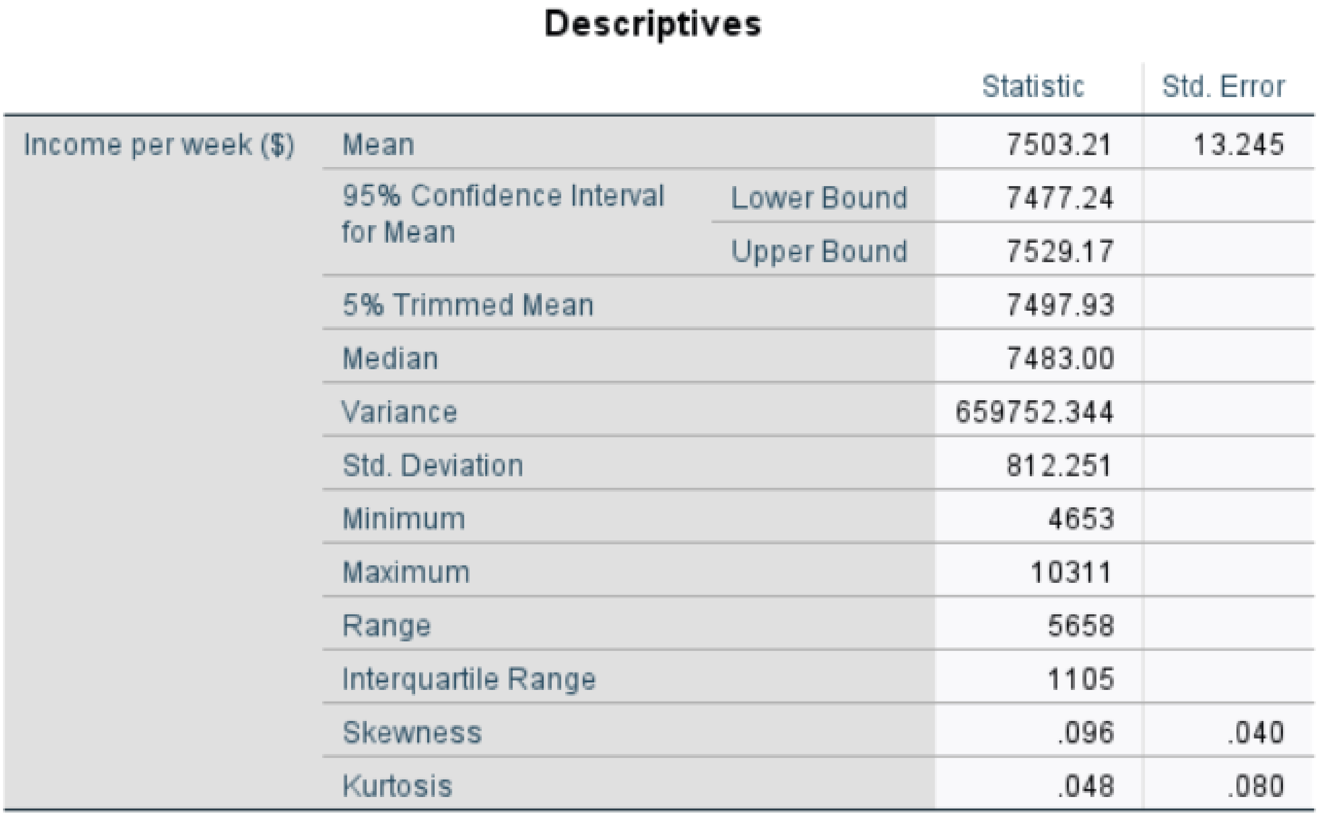
Question 3: Income per week ($)

The variable Income measures income per week in $. Using SPSS, produce the relevant graph and tables to summarise the Income variable and write a paragraph explaining the key features of the data observed in the output in the style presented in the course materials.



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Income per week ($)	3761	75.2%	1239	24.8%	5000	100.0%



Descriptives

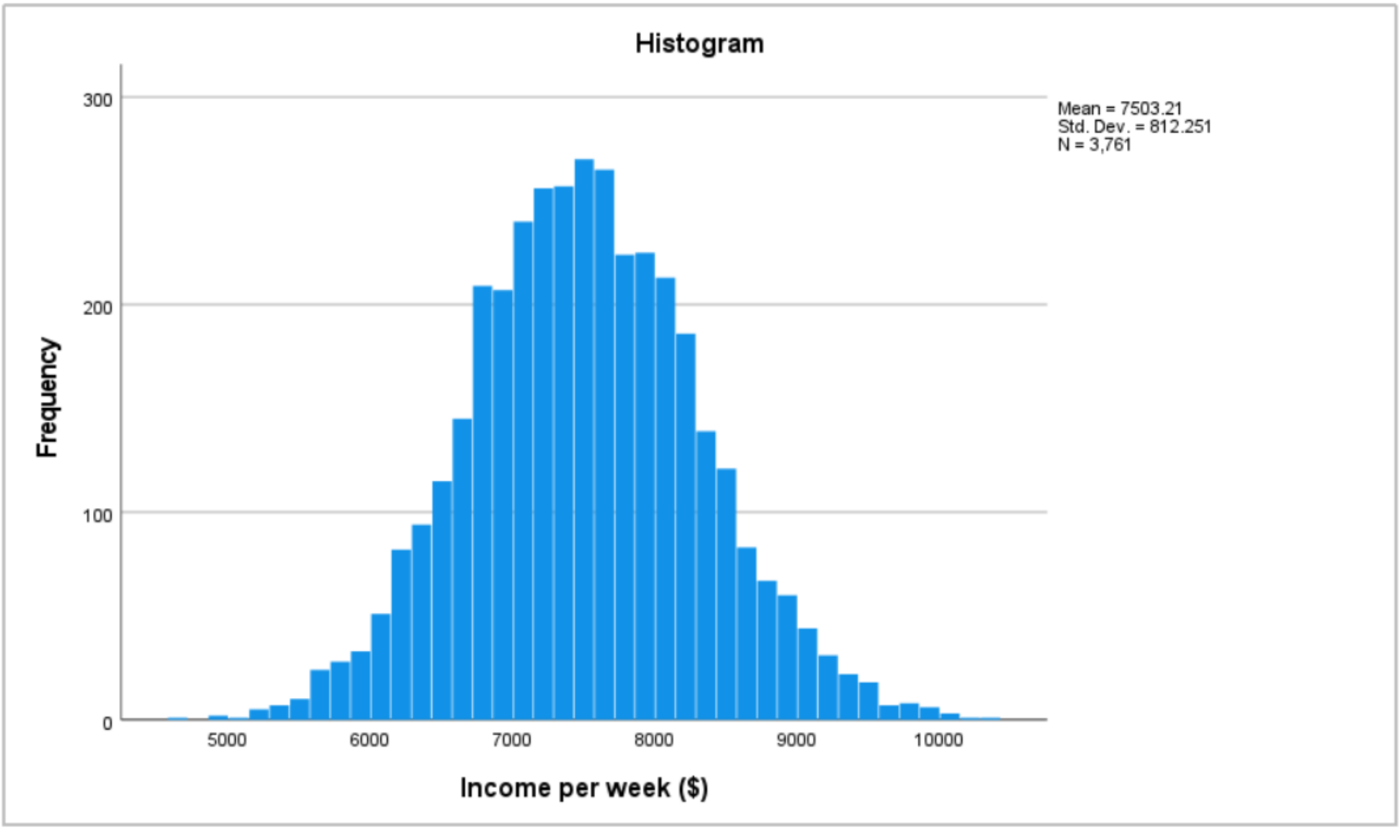
			Statistic	Std. Error
Income per week ($)	Mean		7503.21	13.245
	95% Confidence Interval for Mean	Lower Bound	7477.24	
		Upper Bound	7529.17	
	5% Trimmed Mean		7497.93	
	Median		7483.00	
	Variance		659752.344	
	Std. Deviation		812.251	
	Minimum		4653	
	Maximum		10311	
	Range		5658	
	Interquartile Range		1105	
	Skewness		.096	.040
	Kurtosis		.048	.080

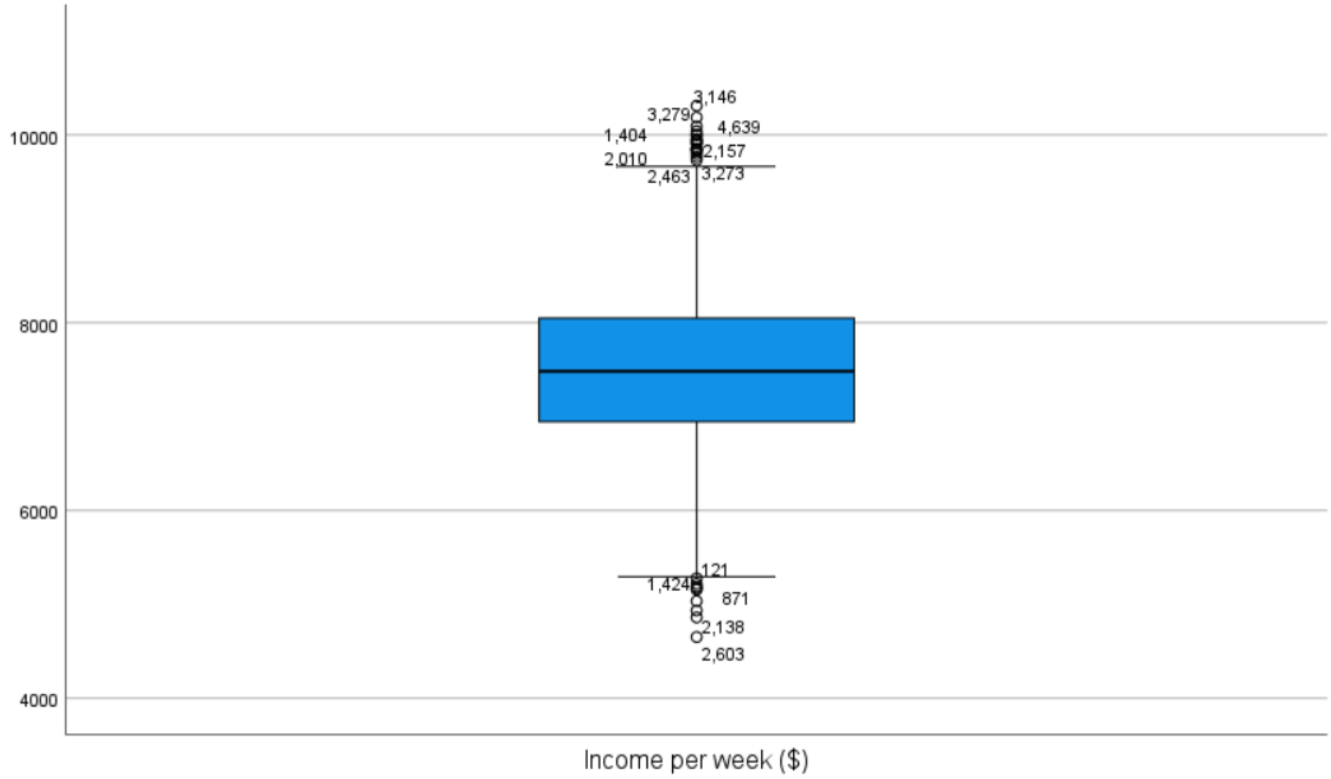
Percentiles

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Income per week ($)	6187.00	6482.00	6946.00	7483.00	8050.50	8538.60	8878.40
Tukey's Hinges	Income per week ($)			6947.00	7483.00	8050.00		

Income per week ($)





The distribution of weekly income ($), based on 3,761 valid responses from a sample of 5,000, is shown in the figures. A slight positive skew (0.096) is apparent, with a median income of $7,483 and an interquartile range of $6,946 to $8,050.50. Although most incomes fall within this range, several high-end outliers, up to $10,311, contribute to the positive skew and a mean income of $7,503.21. Specifically, the boxplot identifies outliers such as $1,404, $2,010, $2,157, $3,279, and $4,639, indicating a small subset with considerably higher earnings.

Question 4: [does not require SPSS]

Jasmine likes to take part in competitions where participants assemble mechanical puzzles. Participants use an assortment of metal and plastic parts to make a specific object. Participants are timed to see how long it takes them to assemble a mechanical puzzle. Jasmine participated in four competitions last year. Jasmine completed the car puzzle in 420 seconds, she completed the bus puzzle in 340 seconds, she completed the train puzzle in 670 seconds, and she completed the plane puzzle in 550 seconds. Completion times for people who did the car puzzle competition are normally distributed with a mean of μ = 469 seconds and a standard deviation σ = 54 seconds. Completion times for people who did the bus puzzle competition are normally distributed with a mean of μ = 300 seconds and a standard deviation σ = 29 seconds. Completion times for people who did the train puzzle competition are normally distributed with a mean of μ = 724 seconds and a standard deviation σ = 42 seconds. Completion times for people who did the plane puzzle competition are normally distributed with a mean of μ = 505 seconds and a standard deviation σ = 36 seconds. In which puzzle competition was Jasmine's performance best, relative to others who did the puzzle competitions? Justify your answer, quoting relevant statistics as part of your explanation.

Answer: For determining in which puzzle competition Jasmine performed best, relative to others who did the puzzle competition, we need to use the z-score formula below:

$$z = \frac{X-\mu}{\sigma}$$

* Car puzzle: $z = \frac{X-\mu}{\sigma} = \frac{420-469}{54} \approx -0.91$

* Bus puzzle: $z = \frac{X-\mu}{\sigma} = \frac{340-300}{29} \approx 1.38$

* Train puzzle: $z = \frac{X-\mu}{\sigma} = \frac{670-724}{42} \approx -1.29$

* Plane puzzle: $z = \frac{X-\mu}{\sigma} = \frac{550-505}{36} = 1.25$

Comparing the z-score between the puzzle competitions, Jasmine's z-score in the bus puzzle competition was the highest which means that her completion time for the bus puzzle was 1.38 standard deviation and below the mean completion time for that puzzle, highlighting a faster time than other competitors in this competition. Although Jasmine was better and shorter than the other participants in the plane puzzle competition, Jasmine's performance in the bus competition was the best. In conclusion, Jasmine's best performance, relative to others, was in the bus puzzle competition.

Question 5: [does not require SPSS]

Umami Papi is a crispy aromatic chili oil that is made in Melbourne. The company makes two versions of chili oil, Original and Extra Spicy. Umami Papi comes in two sizes, a 225g jar and a 750g jar. The production manager at the production plant wants to make sure that the machine that fills the jars is working correctly. The production manager takes a random sample of 1200 225g jars of Umami Papi Original chili oil produced in March 2024. This sample is then used to check if bottles are filled with the correct amount of chilli oil, or if perhaps the machine needs to be adjusted.

a. What is the population we can draw conclusions about in this study?

The population we can conclude about in this study is all 225g jars of Unami Papi Original chili oil that were produced in March 2024 at the Melbourne production plant..

To answer questions (b) to (d), only consider the sampling distribution shown in *Figure 1*.
No calculations are needed.

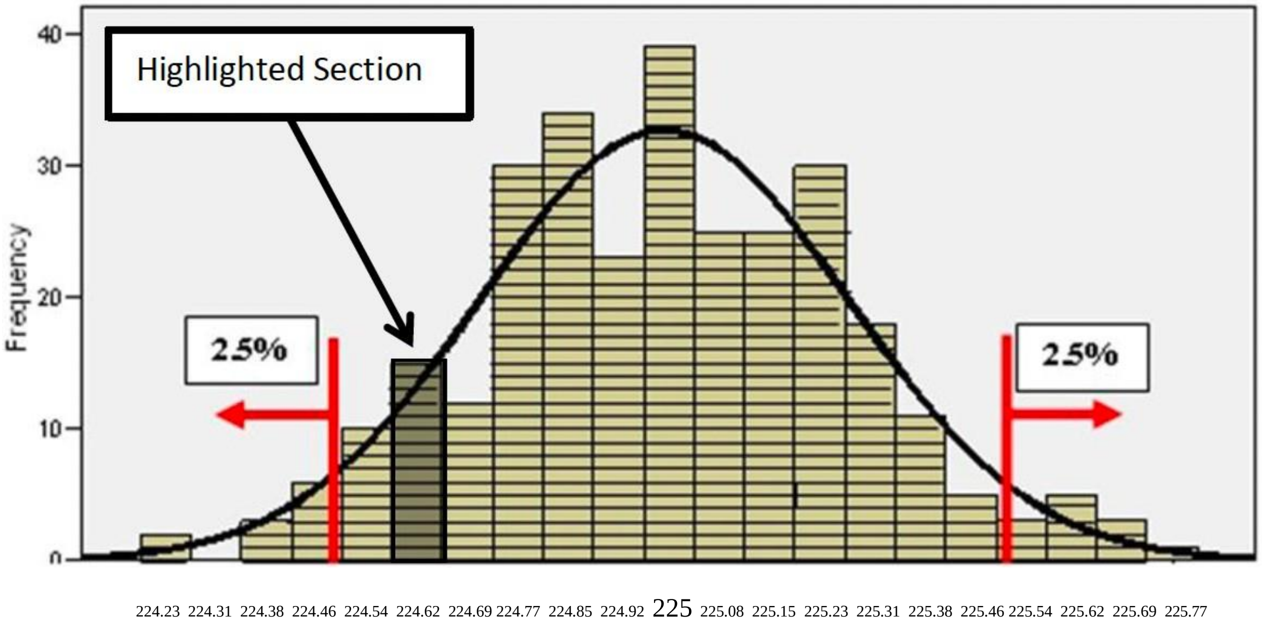


Figure 1: Distribution of sample means in 300 samples of size 1200, taken from a population where the mean is 225 and the standard deviation is 8

b. What does the highlighted section of the sampling distribution in *Figure 1* represent?
The highlight section of the sampling distribution in Figure 1 represents the extreme tails, capturing 5% (2.5% in each tail) of sample means that deviate the most from the population mean of 225. These represent the least probable sample means to observe if the true population means is 225.

c. The random sample of 1200 jars taken by the production manager had a mean weight of 225.62g. Does this sample look like it belongs to the sampling distribution displayed in Figure 1? Justify your answer.

- The sample mean of 225.62g appears inconsistent with the sampling distribution in Figure 1. It is visually located far to the right, beyond the highlighted region representing the extreme 5% of likely values if the true mean were 225g. Calculating the z-score provides further evidence:

$$\text{z-score} = \frac{225.62-225}{\frac{8}{\sqrt{1200}}} \approx 2.685$$

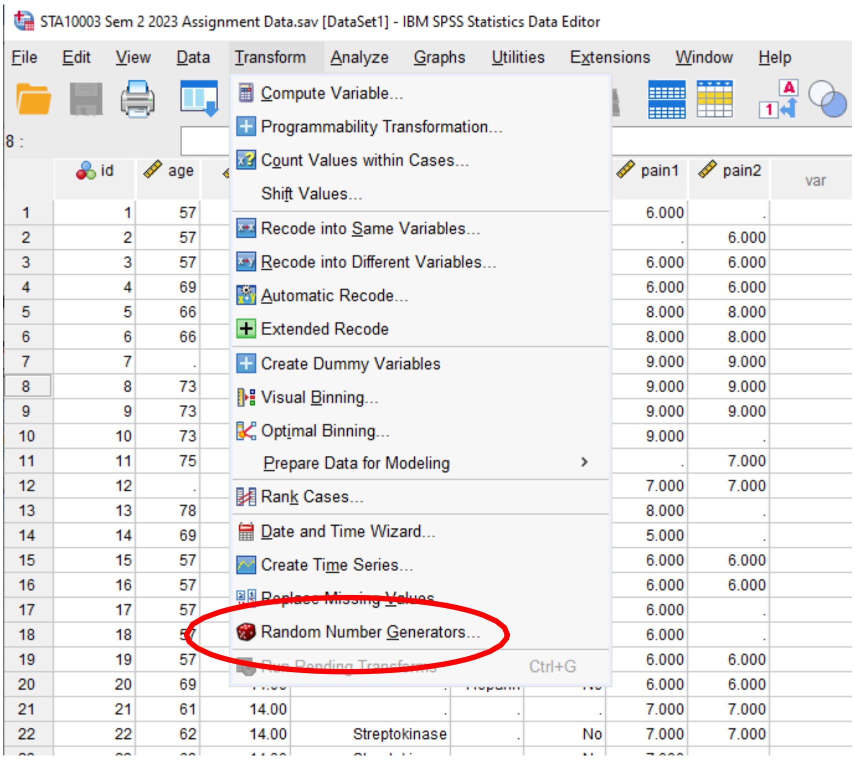
- This z-score (2.685) indicates the sample meaning is over two and a half standard errors above the hypothesized population mean. This large deviation suggests the sample likely does *not* belong to the distribution shown in Figure 1.

d. The production manager wants to know if bottles are filled with the correct amount of chilli oil. What specific conclusion can be made based on part (c)?

The elevated sample mean (225.62g) strongly suggests the filling process is overfilling the jars. This difference from the expected mean (225g) is unlikely due to random chance, as evidenced by the high z-score and its position outside the plausible range in Figure 1. The production manager should investigate and adjust the filling machine to address this overfilling issue.

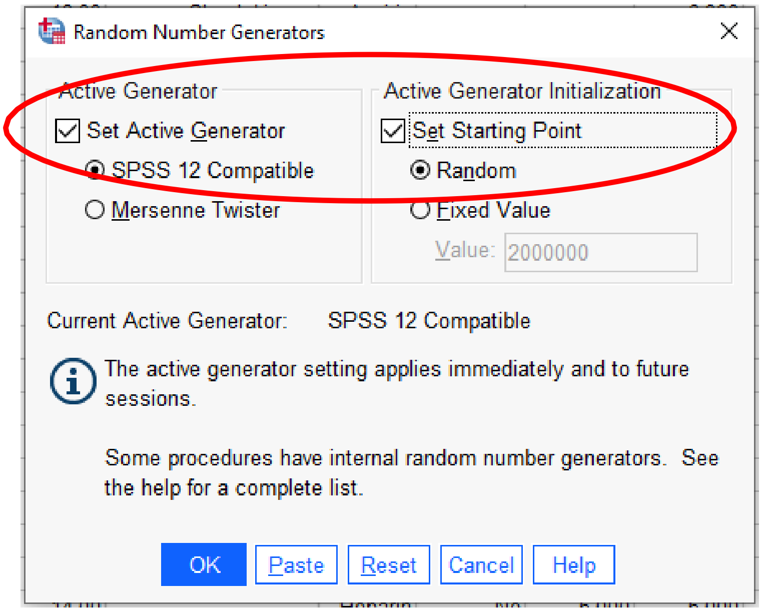
How to generate your random sample of 5000 observations.

1. Open the **STA10003 Sem 2 2024 Assignment Data.sav** data file. From the **Transform** drop-down menu, select **Random Number Generators** from the menu


STA10003 Sem 2 2023 Assignment Data.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Extensions
Window
Help
Compute Variable...
Programmability Transformation...
Count Values within Cases...
Shift Values...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Extended Recode
Create Dummy Variables
Visual Binning...
Optimal Binning...
Prepare Data for Modeling
Rank Cases...
Date and Time Wizard...
Create Time Series...
Replace Missing Values...
Random Number Generators...
Run Pending Transforms
Ctrl+G
8 :
var

	id	age	pain1	pain2
1	1	57	6.000	.
2	2	57	.	6.000
3	3	57	6.000	6.000
4	4	69	6.000	6.000
5	5	66	8.000	8.000
6	6	66	8.000	8.000
7	7	.	9.000	9.000
8	8	73	9.000	9.000
9	9	73	9.000	9.000
10	10	73	9.000	.
11	11	75	.	7.000
12	12	.	7.000	7.000
13	13	78	8.000	.
14	14	69	5.000	.
15	15	57	6.000	6.000
16	16	57	6.000	6.000
17	17	57	6.000	.
18	18	57	6.000	.
19	19	57	6.000	6.000
20	20	69	6.000	6.000
21	21	61	7.000	7.000
22	22	62	7.000	7.000

2. From the **Random Number Generators** dialogue box, click the boxes beside **Set Active Generator** and **Set Starting Point** as shown below. Then click **OK.**



Active Generator

☑ Set Active Generator

● SPSS 12 Compatible

○ Mersenne Twister

Active Generator Initialization

☑ Set Starting Point

● Random

○ Fixed Value

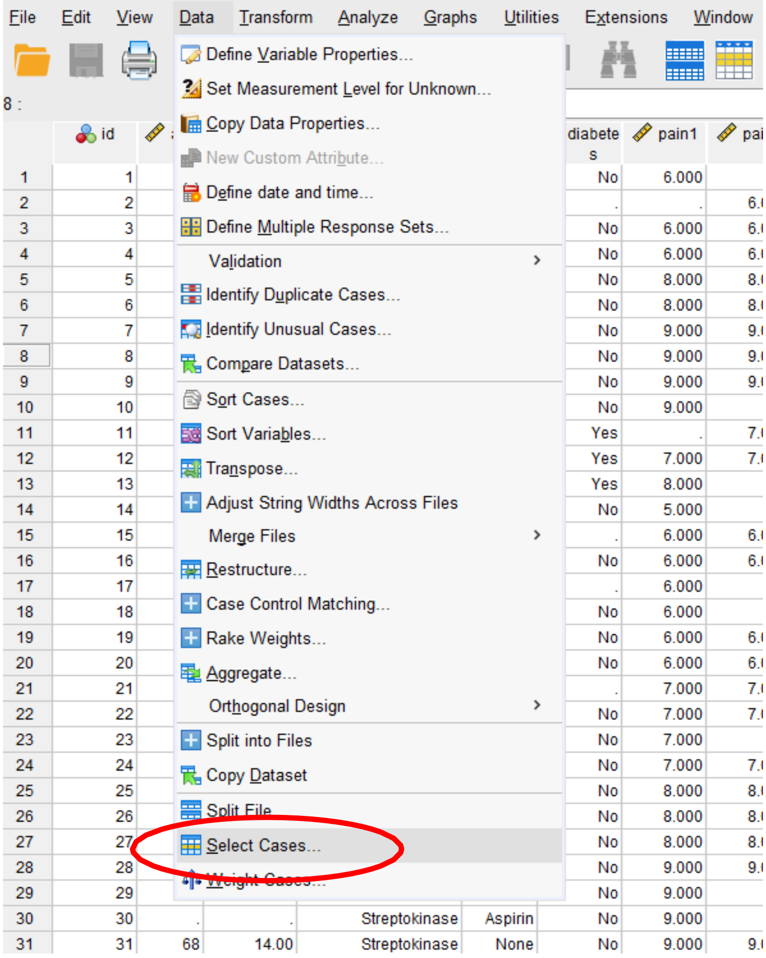
Value: 2000000

Current Active Generator: SPSS 12 Compatible

The active generator setting applies immediately and to future sessions.

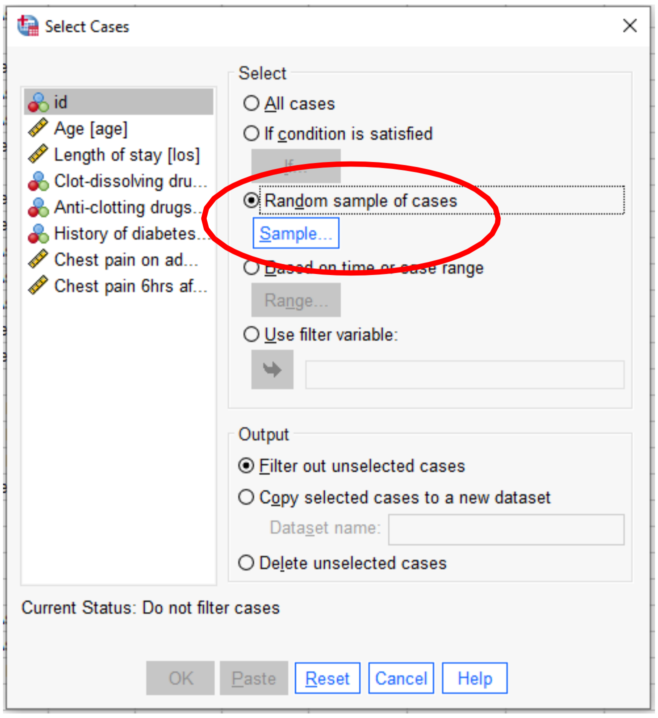
Some procedures have internal random number generators. See the help for a complete list.

3. From the **Data** drop-down menu, choose **Select Cases**



	id					diabetes	pain1	pai
1	1					No	6.000	
2	2					.	.	6.0
3	3					No	6.000	6.0
4	4					No	6.000	6.0
5	5					No	8.000	8.0
6	6					No	8.000	8.0
7	7					No	9.000	9.0
8	8					No	9.000	9.0
9	9					No	9.000	9.0
10	10					No	9.000	
11	11					Yes	.	7.0
12	12					Yes	7.000	7.0
13	13					Yes	8.000	
14	14					No	5.000	
15	15					.	6.000	6.0
16	16					No	6.000	6.0
17	17					.	6.000	
18	18					No	6.000	
19	19					No	6.000	6.0
20	20					No	6.000	6.0
21	21					.	7.000	7.0
22	22					No	7.000	7.0
23	23					No	7.000	
24	24					No	7.000	7.0
25	25					No	8.000	8.0
26	26					No	8.000	8.0
27	27					No	8.000	8.0
28	28					No	9.000	9.0
29	29					No	9.000	
30	30	.	.	Streptokinase	Aspirin	No	9.000	
31	31	68	14.00	Streptokinase	None	No	9.000	9.0

4. From the **Select Cases** dialogue box, choose **Random Sample of Cases** and then click the **Sample** button.



Select

○ All cases

○ If condition is satisfied

● Random sample of cases

○ Based on time or case range

○ Use filter variable:

Output

● Filter out unselected cases

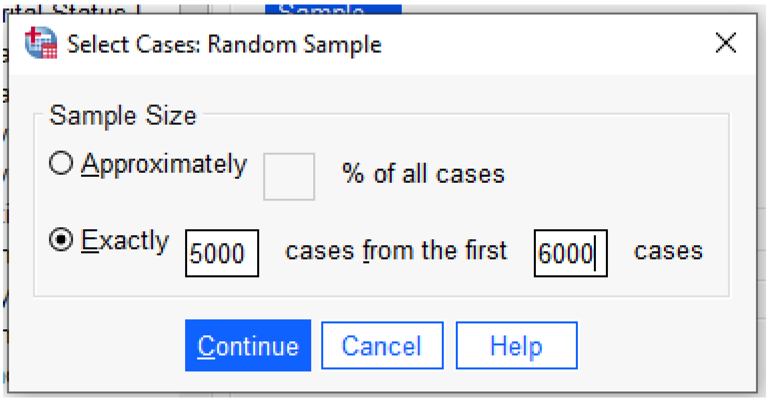
○ Copy selected cases to a new dataset

Dataset name:

○ Delete unselected cases

Current Status: Do not filter cases

5. From the **Select Cases: Random sample** dialogue box, click **Exactly** and type **5000** cases from the first **6000**.



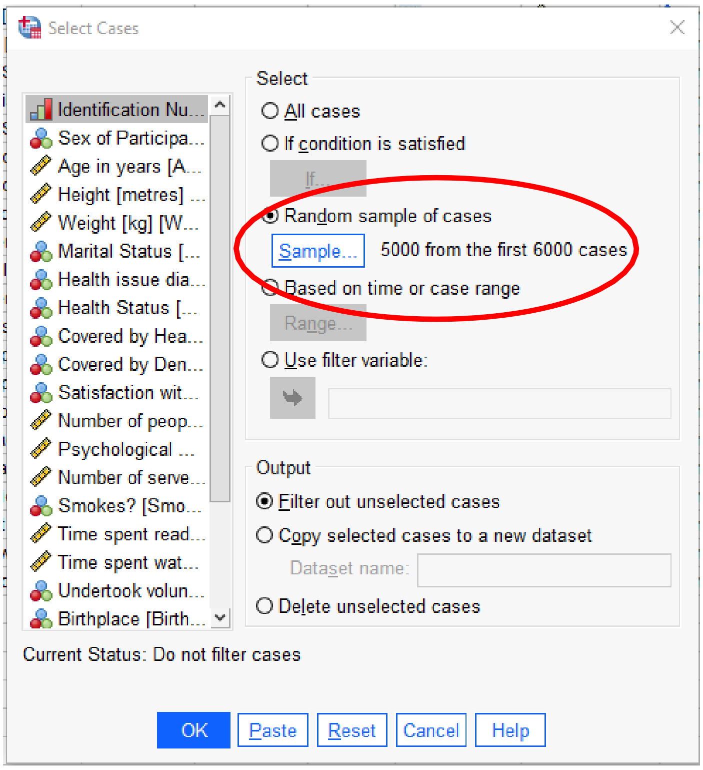
Select Cases: Random Sample

Sample Size

○ Approximately ____ % of all cases

● Exactly 5000 cases from the first 6000 cases

6. After entering the information above, click **Continue** [this returns you to the **Select Cases** Dialogue Box where you will see, next to the **Sample** button, confirmation of the 5000 cases selected].



Select Cases

Select

○ All cases

○ If condition is satisfied

● Random sample of cases

Sample... 5000 from the first 6000 cases

○ Based on time or case range

○ Use filter variable:

Output

● Filter out unselected cases

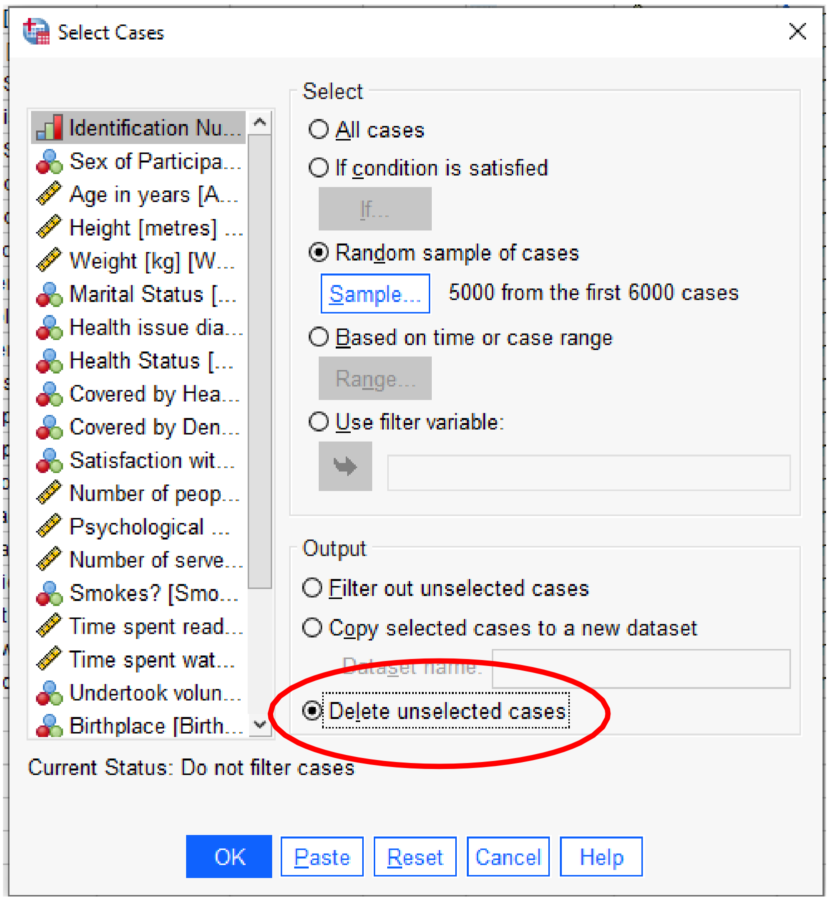
○ Copy selected cases to a new dataset

Dataset name:

○ Delete unselected cases

Current Status: Do not filter cases

7. We can remove the unselected cases by clicking the **Delete unselected cases** button under the **Output** heading.



Select Cases

Identification Nu...
Sex of Participa...
Age in years [A...
Height [metres] ...
Weight [kg] [W...
Marital Status [...
Health issue dia...
Health Status [...
Covered by Hea...
Covered by Den...
Satisfaction wit...
Number of peop...
Psychological ...
Number of serve...
Smokes? [Smo...
Time spent read...
Time spent wat...
Undertook volun...
Birthplace [Birth...

Select

○ All cases

○ If condition is satisfied

● Random sample of cases

Sample... 5000 from the first 6000 cases

○ Based on time or case range

○ Use filter variable:

Output

○ Filter out unselected cases

○ Copy selected cases to a new dataset

Dataset name:

● Delete unselected cases

Current Status: Do not filter cases

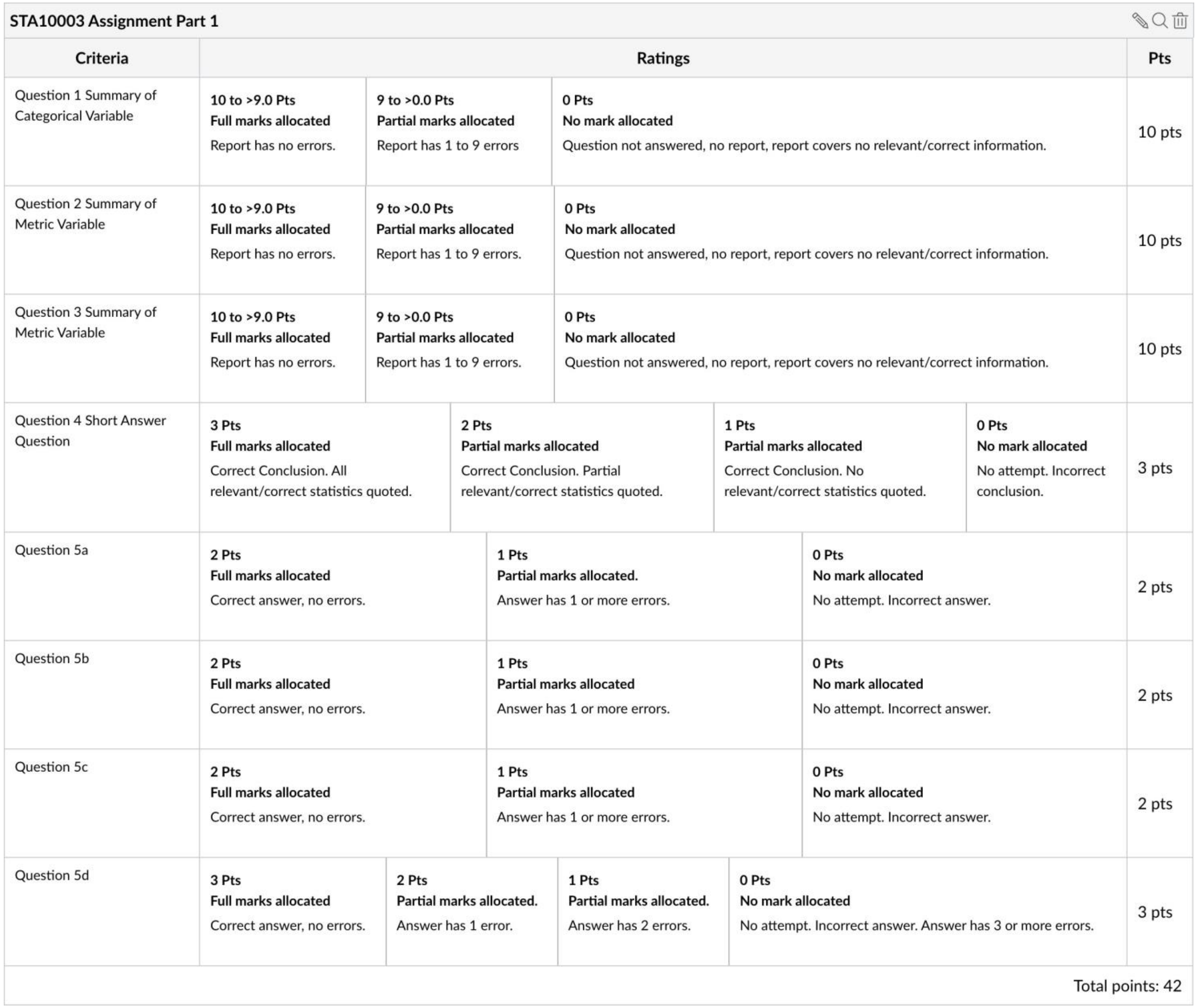
8. After clicking **OK**, your data set will now only show the 5000 cases selected.

9. You should now **save the data file with a new name**.

The data file is ready to use for your Assignment!

Note: This data file will also be used in the Assignment Part 2.

STA10003 Assignment Part 1 Marking Rubric [out of 42]



STA10003 Assignment Part 1

Criteria	Ratings				Pts
Question 1 Summary of Categorical Variable	**10 to >9.0 Pts** **Full marks allocated** Report has no errors.	**9 to >0.0 Pts** **Partial marks allocated** Report has 1 to 9 errors	**0 Pts** **No mark allocated** Question not answered, no report, report covers no relevant/correct information.		10 pts
Question 2 Summary of Metric Variable	**10 to >9.0 Pts** **Full marks allocated** Report has no errors.	**9 to >0.0 Pts** **Partial marks allocated** Report has 1 to 9 errors.	**0 Pts** **No mark allocated** Question not answered, no report, report covers no relevant/correct information.		10 pts
Question 3 Summary of Metric Variable	**10 to >9.0 Pts** **Full marks allocated** Report has no errors.	**9 to >0.0 Pts** **Partial marks allocated** Report has 1 to 9 errors.	**0 Pts** **No mark allocated** Question not answered, no report, report covers no relevant/correct information.		10 pts
Question 4 Short Answer Question	**3 Pts** **Full marks allocated** Correct Conclusion. All relevant/correct statistics quoted.	**2 Pts** **Partial marks allocated** Correct Conclusion. Partial relevant/correct statistics quoted.	**1 Pts** **Partial marks allocated** Correct Conclusion. No relevant/correct statistics quoted.	**0 Pts** **No mark allocated** No attempt. Incorrect conclusion.	3 pts
Question 5a	**2 Pts** **Full marks allocated** Correct answer, no errors.	**1 Pts** **Partial marks allocated.** Answer has 1 or more errors.	**0 Pts** **No mark allocated** No attempt. Incorrect answer.		2 pts
Question 5b	**2 Pts** **Full marks allocated** Correct answer, no errors.	**1 Pts** **Partial marks allocated** Answer has 1 or more errors.	**0 Pts** **No mark allocated** No attempt. Incorrect answer.		2 pts
Question 5c	**2 Pts** **Full marks allocated** Correct answer, no errors.	**1 Pts** **Partial marks allocated** Answer has 1 or more errors.	**0 Pts** **No mark allocated** No attempt. Incorrect answer.		2 pts
Question 5d	**3 Pts** **Full marks allocated** Correct answer, no errors.	**2 Pts** **Partial marks allocated.** Answer has 1 error.	**1 Pts** **Partial marks allocated.** Answer has 2 errors.	**0 Pts** **No mark allocated** No attempt. Incorrect answer. Answer has 3 or more errors.	3 pts

Total points: 42

Marking Details

Prior to submitting your Assignment, use the following checklist as a guide to ensure that you have provided all of the relevant information.

Q1 – Should include [*as appropriate*]:

A graph and Frequency output appropriate for a categorical variable.

A paragraph that includes mention of largest group, and all the other groups with correct percentages, and any other relevant patterns based on *Reading B: Reporting information about single variables*.

Q2 – Should include [*as appropriate*]:

A graph and Explore output appropriate for a metric variable.

A paragraph that includes a description of the shape of distribution, centre, spread, outliers if present based on *Reading B: Reporting information about single variables*.

Q3 – Should include [*as appropriate*]:

A graph and Explore output appropriate for a metric variable.

A paragraph that includes a description of the shape of distribution, centre, spread, outliers if present based on *Reading B: Reporting information about single variables*.

Q4 – The answer should be presented in a short paragraph, quoting relevant statistics in support of the response.

Q5 – The answers should be presented with sections (a) to (d) clearly identified.

Checklist:

- Correct variable used to produce output [note that many of the variables have similar names so it is important to double-check that the correct variable has been used]
- Correct procedure performed
- Graphs appropriately edited and labelled [eg edited variable names; "*Figure 1*. The distribution of …"]
- All figures quoted in report correct according to your own output
- Correctly referring to the sample or population when appropriate
- Proof reading of reports for errors