Question 1

(a) Does mobile phone usage affect young adult brain's memory functions?

(b) The response variable is the score (out of 100) on a 30 minute memory test taken by the subject. It is a discrete numeric data. The primary explanatory variable is the average hours of non-essential mobile phone usages per day across a month, grouped between low (less than 1 hour), medium (1 hour to 2 hours), and high (more than 2 hours). It is an ordered categorical statistic.

(c)

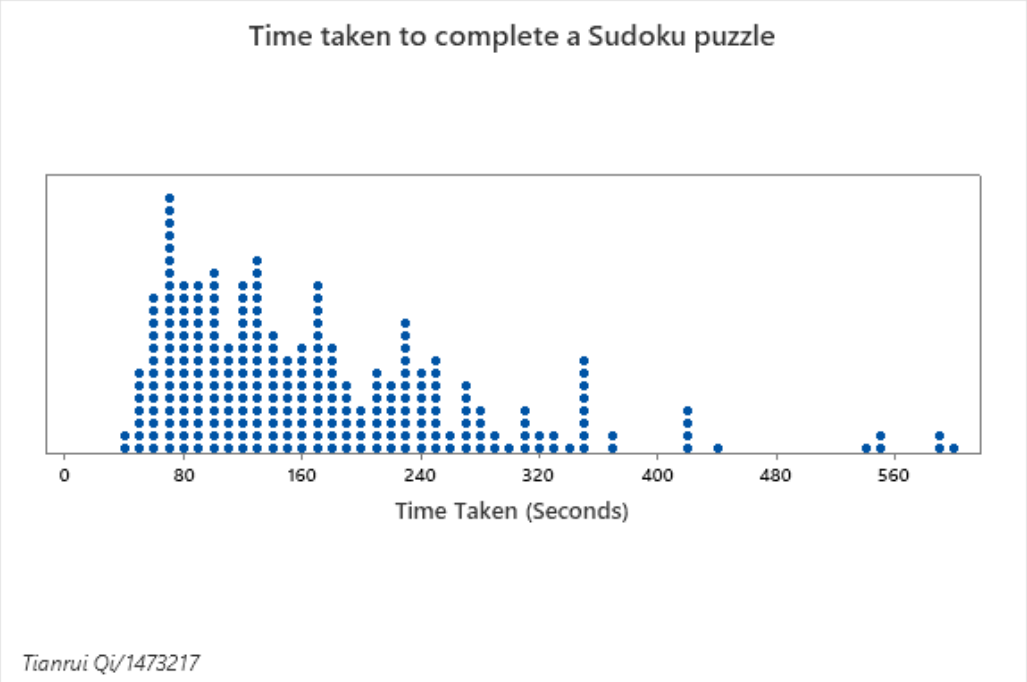
Randomization The study would randomly sample young adults through both online surveys and in public at concentrated areas. It will also randomize the assigning of subjects into the study groups (of low, medium and high mobile phone usages) through a computer randomizer. This will generally to reduce the impacts of confounding variables.

Control The study will enact control through restrictions based on age (say from 18 to 30), as to reduce the effects of confounding variables underlying age and brain memory functions. A sufficient protocol to reduce bias could be to conduct the memory tests in the morning starting at the same time for the same duration. Blinding is not possible for the tests are objective, and the subjects clearly knowing their experimental groups. With capacity, we may perform blocking on the subject genders. This helps to eliminate the confounding impacts of gender.

Replication To increase replication, the study aims to sample around 500 to 1000 subjects to ensure a sufficient numbers subjects as to drown out the unaccounted confounding variables. It is also possible to perform another memory test in conjunction with the existing morning test, but in the afternoon, to increase precision through repetition in measurements.

Question 2

(a) Figure



Statistics

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Time (Seconds)	269	171.90	106.25	40.00	90.50	142.00	228.37	600.00	137.87

(b)

- The distribution is positively skewed with a tail to the right. Its shape resembles a skewed bell curve.
- The median of the distribution is 142 seconds, with the mean larger at 171.9 seconds.
- The spread, using the IQR, is quite large at 137.87 seconds. When measuring in standard deviations, the data points are on average 106.25 seconds away from the mean.
- The outliers in this dataset would lie below $142 - 1.5 \times 137.87 = -64.805$ or above $142 + 1.5 \times 137.87 = 348.805$. There are 4 outliers around the 350 range, 2 around the 370 range,

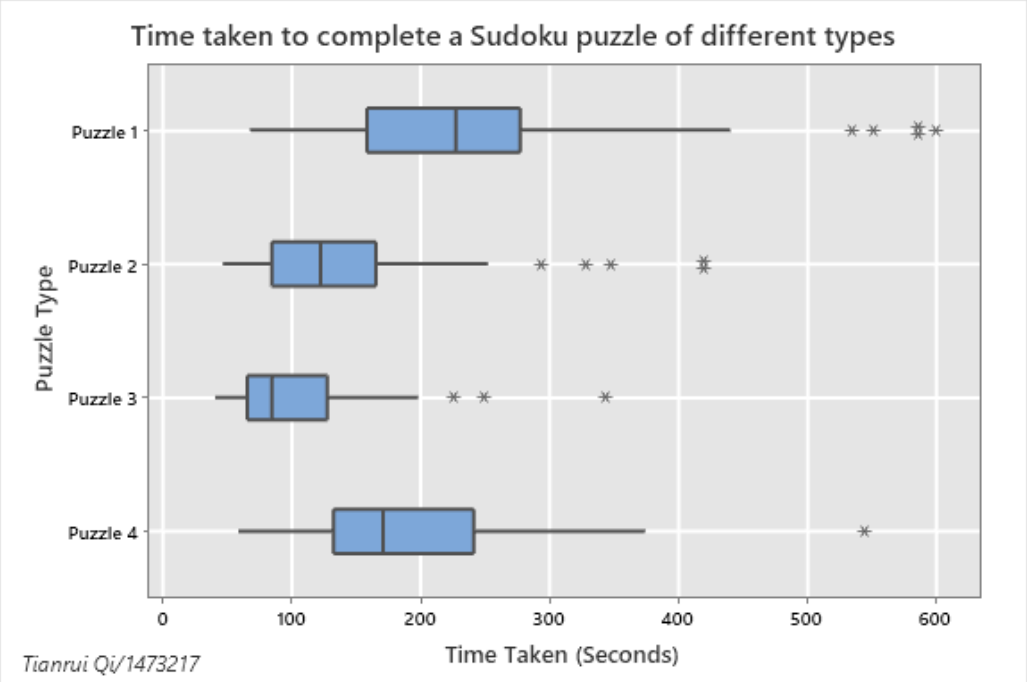
5 around the 420 range, 3 around the 540 range, and 3 around the 590 range. Overall, the dataset contains quite a lot of outliers.

(c) The lower scoring times are all completed on puzzle types of 2 and 3, namely the alphabetical and numeric Sudoku puzzles. Moreover, all the subjects with such times have had previous experience with Sudoku.

Question 3

(a) Barring some minuscule differences in the coloring of the papers, this was a fair comparison between Sudoku puzzles of different symbol types. The reasoning behind is the identical placements of the 6 symbols on the starting tiles between each puzzle types, and a one-to-one mapping of symbols between the puzzles: the beta on puzzle 1 is equivalent to the number 2 in puzzle 3. This ensures that the puzzles are only different by the symbols presented, not in the difficulty inherent of the puzzle.

(b) Figure



Statistics

Variable	Puzzle number	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Time (Seconds)	Puzzle 1	67	244.5	125.5	67.0	158.0	227.0	277.0	600.0	119.0
	Puzzle 2	75	139.30	80.14	46.00	84.00	122.00	165.00	419.00	81.00
	Puzzle 3	59	104.93	57.88	40.00	65.00	84.00	127.00	343.00	62.00
	Puzzle 4	68	194.5	93.4	58.0	132.0	170.5	241.0	545.0	109.0

(c) From the boxplots, we can see various outliers to the right as well as some skewness for each puzzle (puzzle 3 and 4 seems positively skewed). This indicates that we should use the medians for a less-biased comparison of the distributions.

Statistically, we can see that puzzle 3 had the least median time taken of (84 seconds), and puzzle 1 took the longest median time to complete (227 seconds), with puzzle 2 being second place (122 seconds) and puzzle 4 being third place (170.5 seconds). This order is also exhibited throughout the entire 5 number summaries.

Since the data originated from a designed study and the puzzles only differed in symbol types, the data supports the idea that the symbols in these Sudoku puzzle affects the time taken to successfully complete it for students in this class. Numbers lead to the fastest times and Greek letters to the slowest.

Question 4

(a) If there were completely no association between the assigned puzzle types and Sudoku experiences of the subjects, we should observe the same percentage of less experienced subjects (23.05%) across all puzzles. While puzzle 2 and 4 track close to that proportion, puzzle 1 had an excess proportion of less experienced subjects (of 28.36%) while puzzle 3 had a deficit of such subjects (of 15.25%). This may suggest a weak association between puzzle types and subject experiences, hinting at an inadequate randomization process.

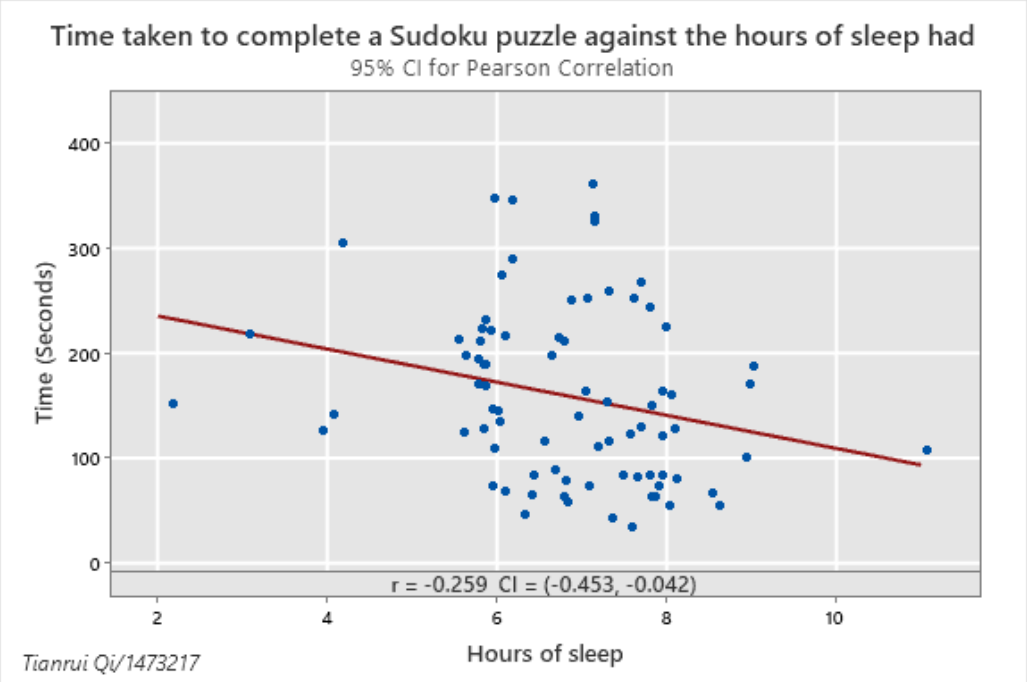
The randomization may also be inadequate for the number of subjects across each puzzle types is not even. We observe that only 21.93% of subjects were allocated to puzzle 3. Ideally, the percentages across each puzzle type should receive 25% to ensure balance. A more complete analysis on the significances of such deviations should leverage the Chi-squared test.

Rows: Have you played Sudoku before? Columns: Puzzle number

	Puzzle 1	Puzzle 2	Puzzle 3	Puzzle 4	All
No	19	17	9	17	62
	30.65	27.42	14.52	27.42	100.00
	28.36	22.67	15.25	25.00	23.05
Yes	48	58	50	51	207
	23.19	28.02	24.15	24.64	100.00
	71.64	77.33	84.75	75.00	76.95
All	67	75	59	68	269
	24.91	27.88	21.93	25.28	100.00
	100.00	100.00	100.00	100.00	100.00

Cell Contents
Count
% of Row
% of Column

(b) From the figure, there is a weak relationship between the hours of sleep the subjects had last night and their time taken to complete a Sudoku puzzle. The respective r value is negative but small at -0.259, indicating that higher hours of sleep is weakly associated with faster puzzle completions.



Question 5

(a)

Pro The graph has a clear title that includes both the primary explanatory variable (surgery types) and the response variable (regret rates).

Con The pie chart is wrongly used to display unrelated regret percentages between the studies of different populations; the sum of the percentages in the slices adds to 103%, a pie chart should have its slices add to 100%.

Con There are no units under the labels of each slice, allowing the reader to misinterpret the units as potentially cases. For instance, it should read "Bowel 32%" instead.

(b) Figure