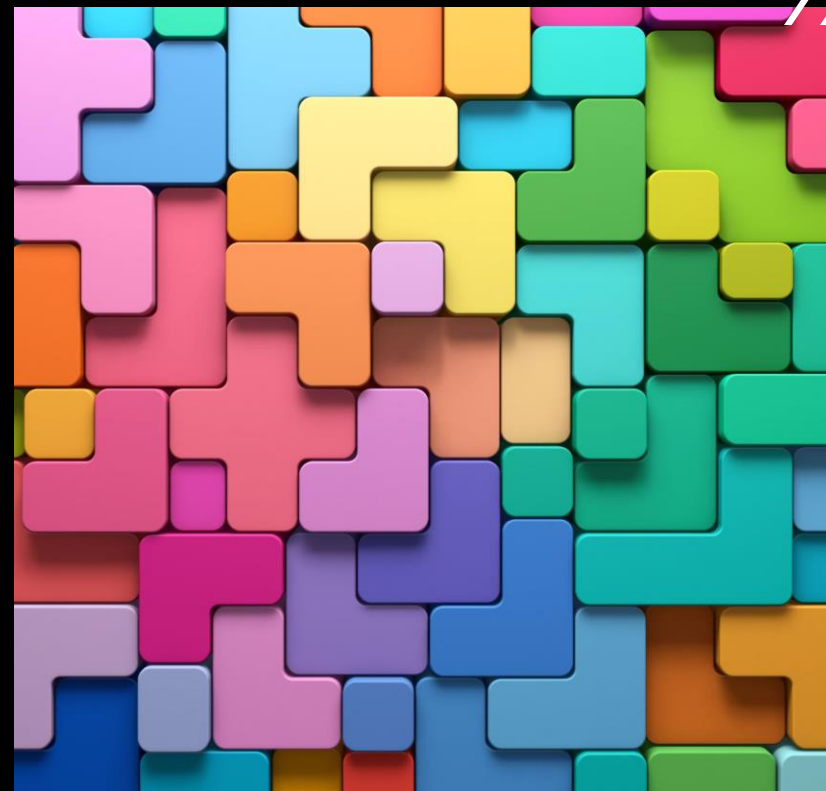


TOPICS IN DATA ANALYSIS

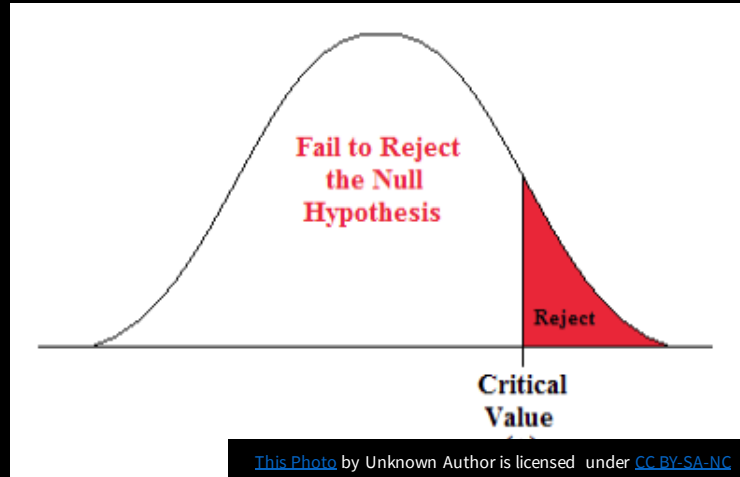
SUDOKU PUZZLES

BY: HARIMUKESH KALAIYARSON,
OBINNA NJOKU, SAMANYU SEN, SAHIL
SINGH

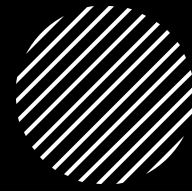




[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)



[This Photo](#) by Unknown Author is licensed under [CC BY-SA-NC](#)

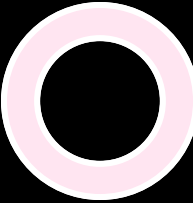


Introduction

- In this project we have come up with multiple hypotheses from the sudoku **data set**.
- We used different methods like t-test and chi-squared test to either accept or reject null hypothesis
- We used SAS and R as our primary coding language.



THE SUDOKU DATA SET



IN MAYNOOTH UNIVERSITY BETWEEN 2009-2013, AN IN-CLASS EXPERIMENT WAS CARRIED OUT WITH EIGHT DIFFERENT CLASSES. FOR EACH EXPERIMENT, THE STUDENTS IN THE CLASS WERE GIVEN A SUDOKU PUZZLE TO COMPLETE. THERE WERE FOUR DIFFERENT TYPED OF PUZZLES WHICH WERE EFFECTIVELY THE SAME PUZZLE BUT WITH DIFFERENT SYMBOLS.

THIS DATA SET HAD 10 VARIABLES!!!

- CLASS(1-8)
- BEFORE1 (HAD PLAYED SUDOKU BEFORE)
- TYPE(NUMBERS, GREEK, LETTERS OR SYMBOLS)
- CORRECT(YES OR NO)
- TIME1(TIME TO COMPLETION)
MINS
SECONDS
- TIME2(TIME TO COMPLETION IN SECONDS)
- BEFORE2(HAVE PLAYED SUDOKU IN LAST 3 MONTHS)
- LOGIC(YES OR NO)

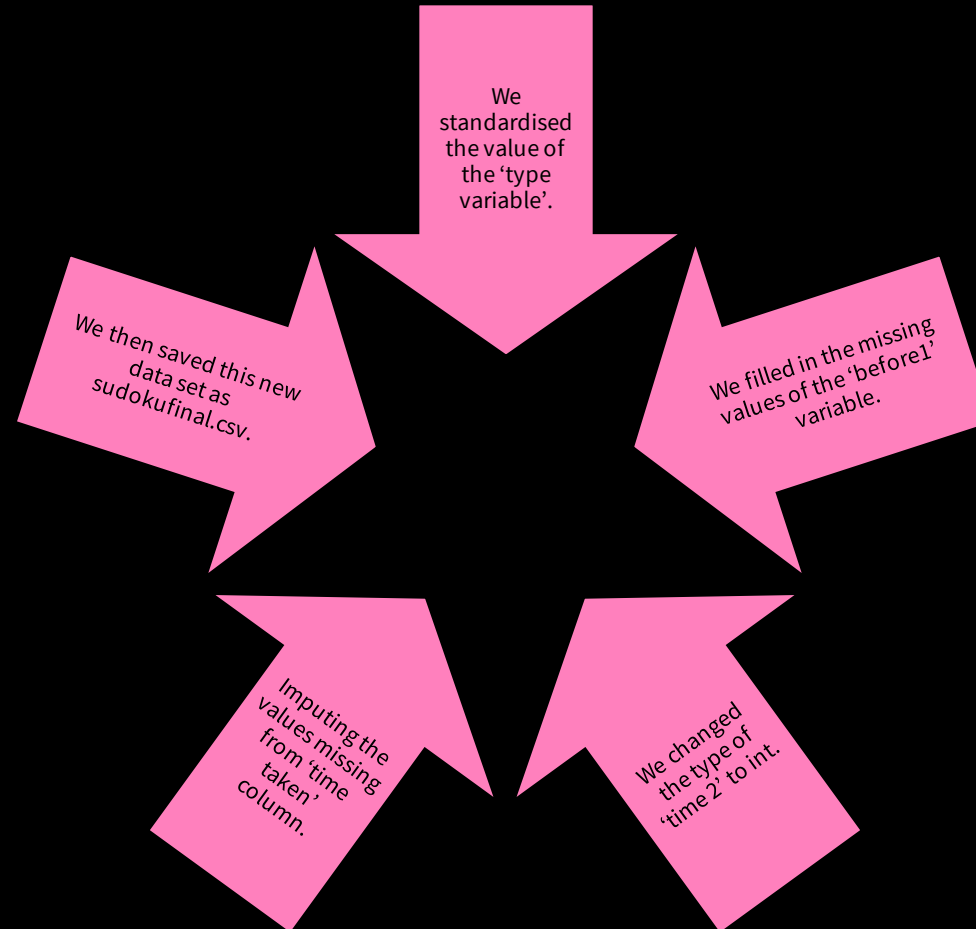
THE WORK BEGINS!!!

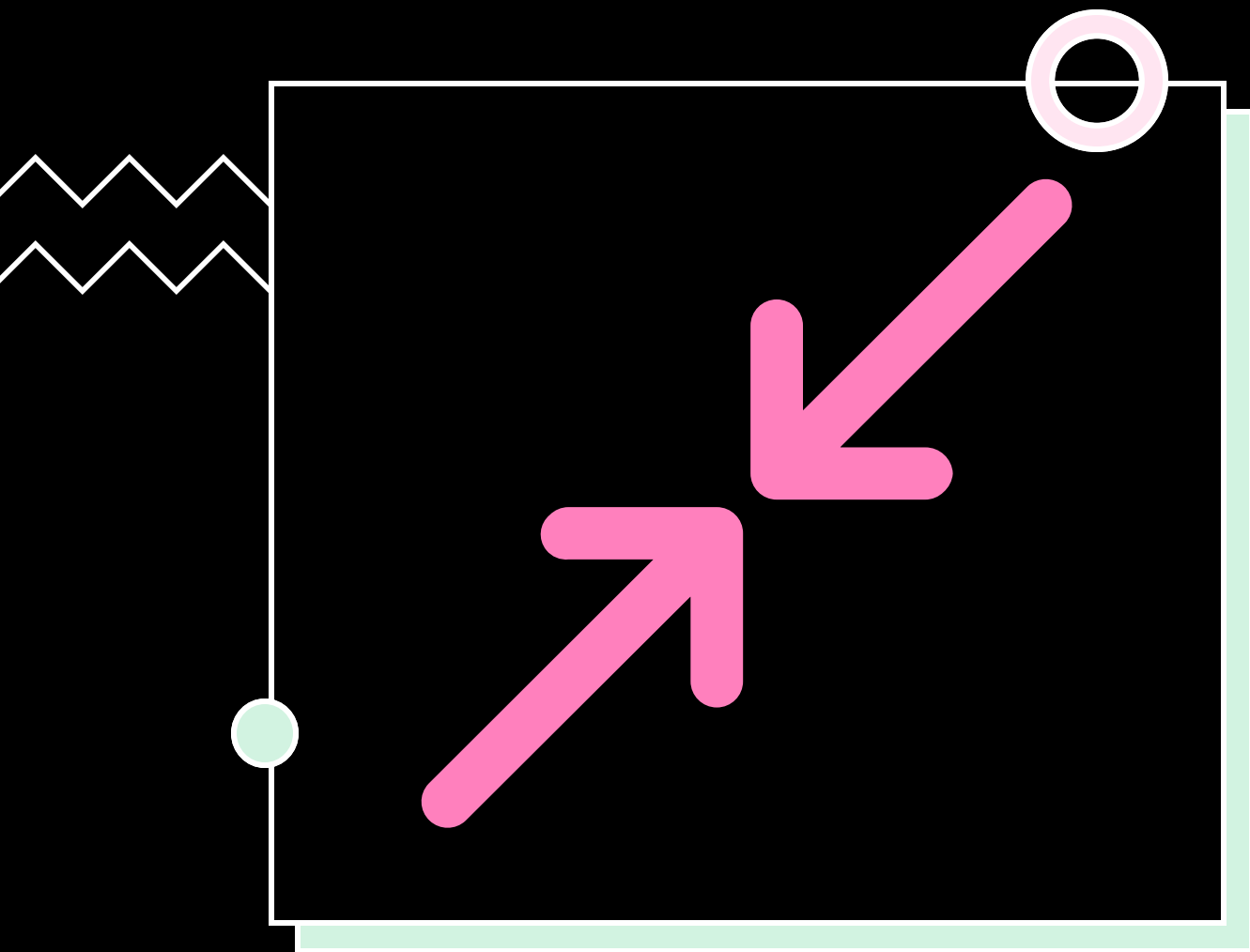
- Before we started hypothesis testing a few necessary steps had to be taken.
- Firstly, using the data from sudokucombined.csv, we cleaned the data in SAS.





HOW DID WE CLEAN THE DATA?

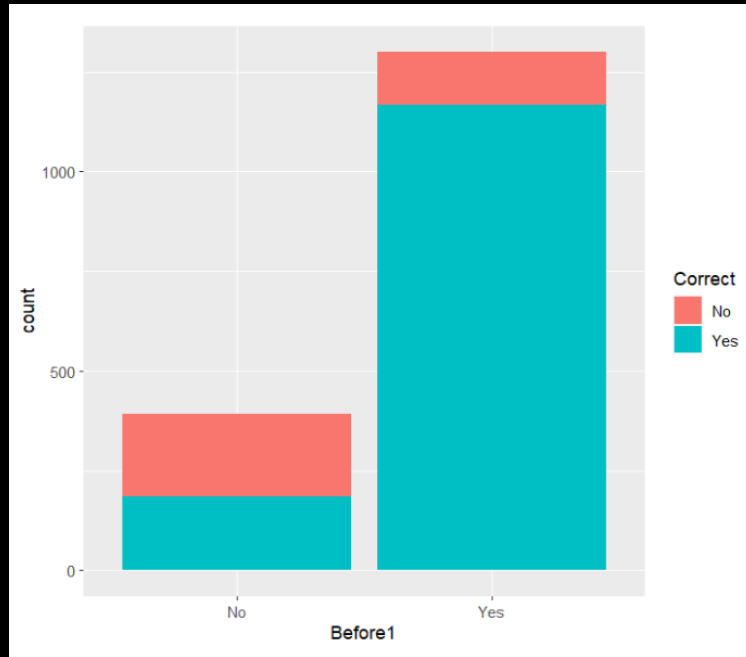




T H E W O R K
C O N T I N U E S ! ! !

WE THEN USED R TO SEE
THE CORRELATION
BETWEEN THE DIFFERENT
VARIABLES.

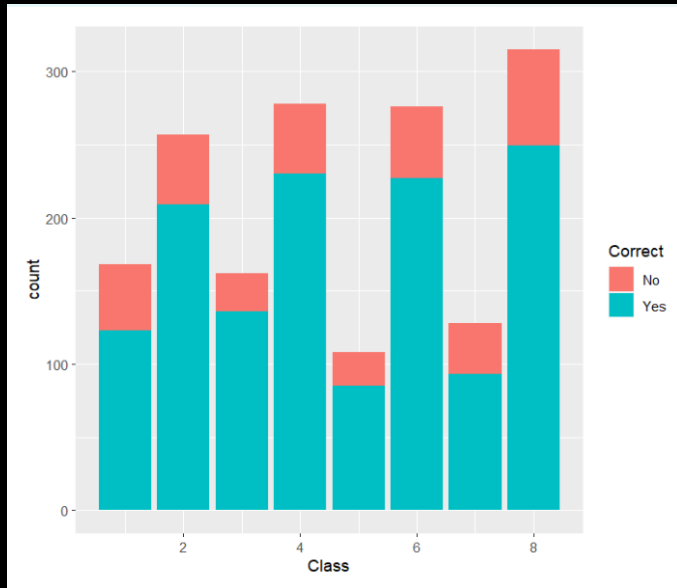




R SESSION

- We looked at the ratio of correct and incorrect solutions, for the Before1 variable (i.e. people that had played sudoku before: yes or no).
- We observed that people that had not played sudoku before were more likely to get an incorrect solution compared to those who had played sudoku

Correlation 2

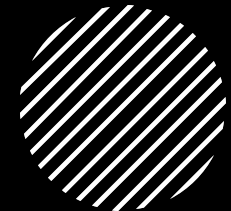
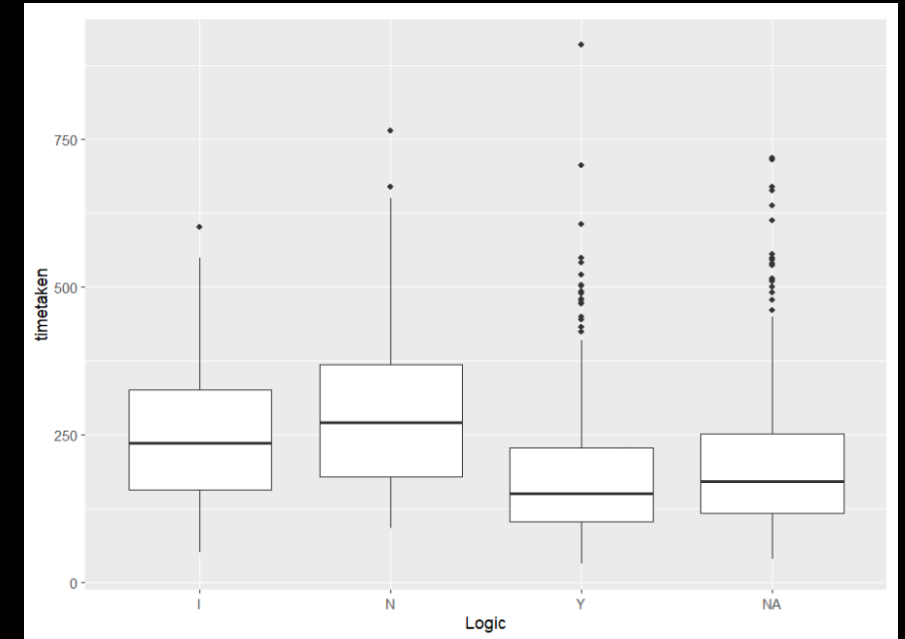


- The same ratio was also taken for the class variable.
- We found that in all the 8 classes the probability of getting the correct solution is always higher than getting an incorrect one.

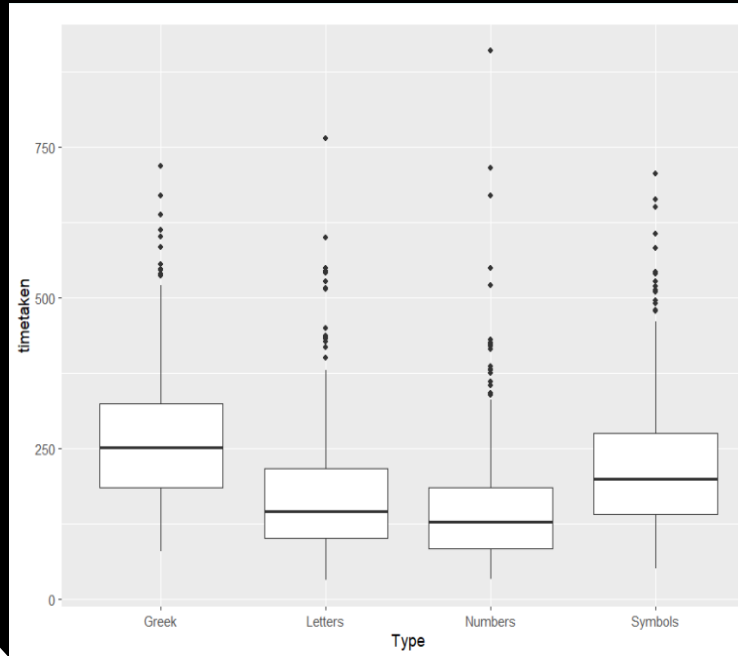


Correlation 3...

- We also took a look at the correlation between the logic and the time taken (i.e. if they liked playing sudoku does it affects the time taken to complete puzzle?)
- We noticed that on average, people that liked playing sudoku, took less time to finish the puzzle, compared to those that didn't like playing sudoku. People that were indifferent took less time to complete puzzle than those that didn't like sudoku but more time than those who liked sudoku.



CORRELATION 4...



- We finally looked at the correlation between the time taken to complete puzzle and the type of puzzle.
- We seen that the puzzles with numbers on average, would take the least time to complete, followed by letter puzzle, symbol puzzle came in third place and the Greek puzzle took the most time to complete.



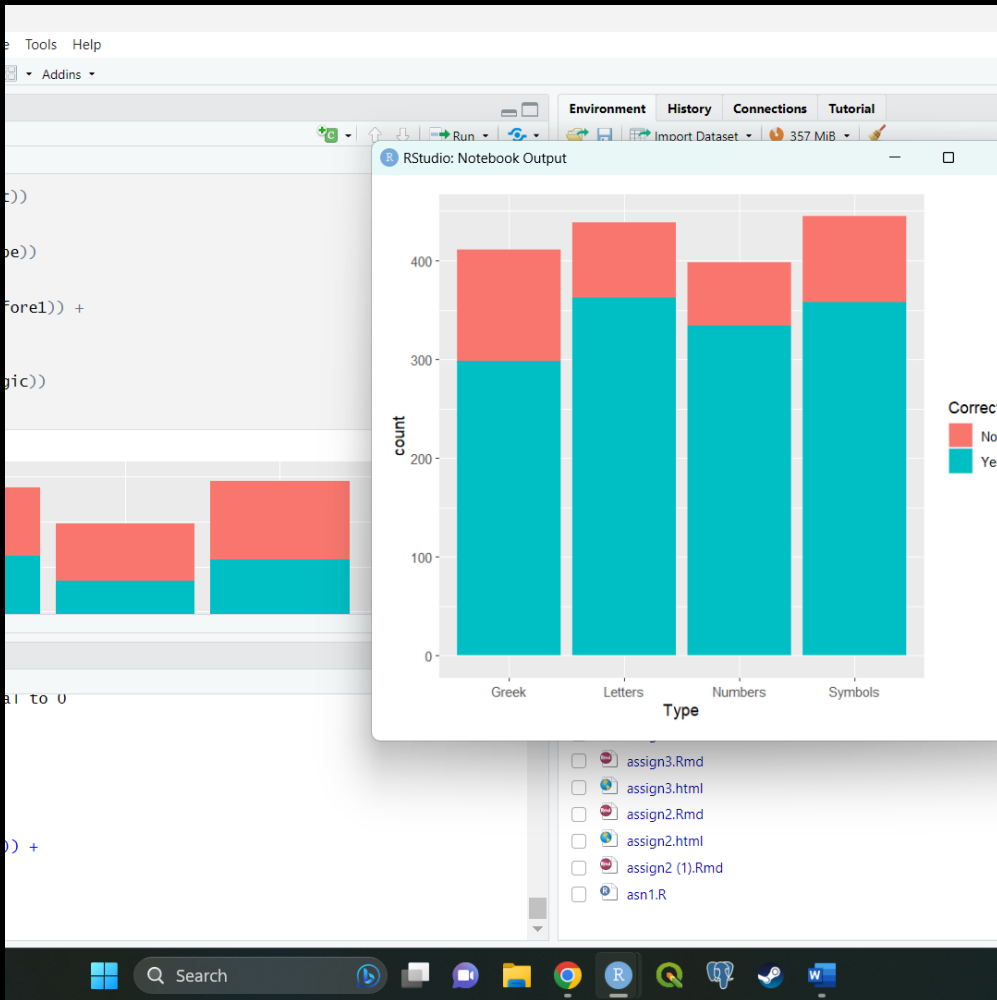


HYPOTHESES TESTING

WE ENDED UP TESTING 3 DIFFERENT
HYPOTHESES.

Hypotheses 1

- Is the probability of successfully completing the sudoku related to the type of sudoku played?
- H_0 : The probability of successfully completing the sudoku is similar for all types of sudoku played.
- H_a : The probability of successfully completing the sudoku is not similar for all types of sudoku played.
- A Pearson chi-squared test was used to test hypotheses and we estimated a p-value = 0.0001586.
- We went ahead and rejected the null hypotheses(H_0).
- Based on this test we can conclude that the probability of successfully completing the sudoku puzzle is related to the type of sudoku played.



Hypotheses test 2



The time taken to successfully complete puzzle is similar for all type of puzzle?



H0: The time taken to successfully complete the sudoku is similar for all types of sudoku played.



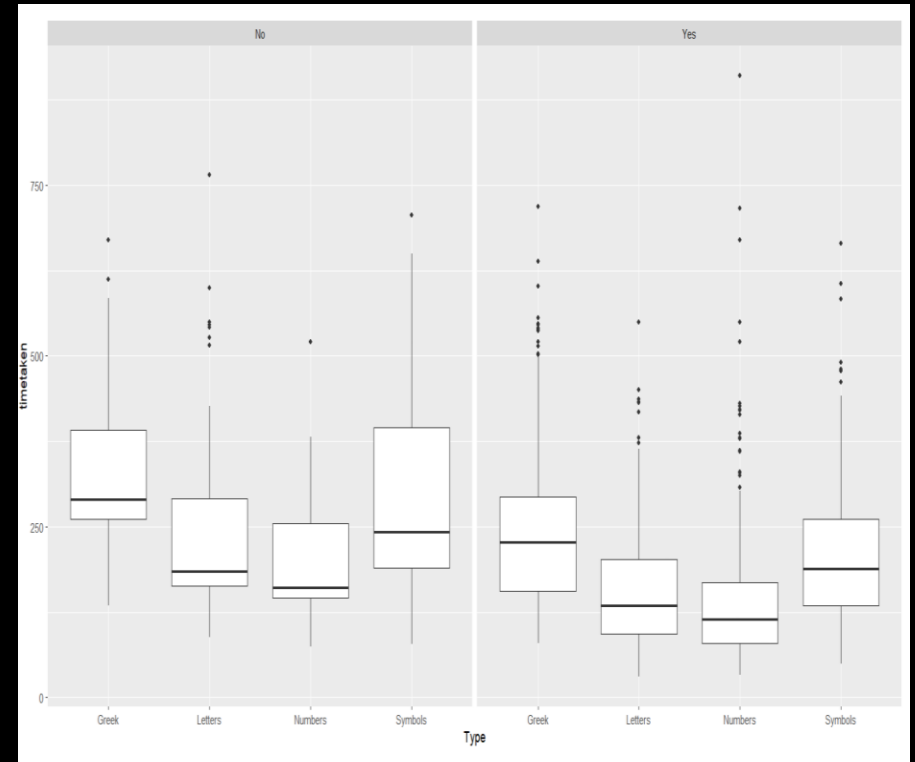
Ha: The time taken to successfully complete the sudoku is not similar for all types of sudoku played.



We applied a one sample t-test and we got a p-value=0.003867 so we reject the null hypotheses.



Based on this test we concluded that The time taken to successfully complete the puzzle is not similar for all types of sudoku played.



Hypotheses test 3



Do people who have solved sudoku before take less time to solve the puzzle correctly?



H0: The time taken to successfully complete the sudoku is similar for people who have played sudoku than those that have not.



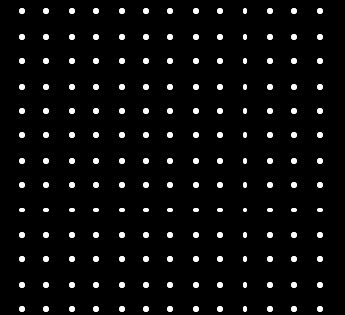
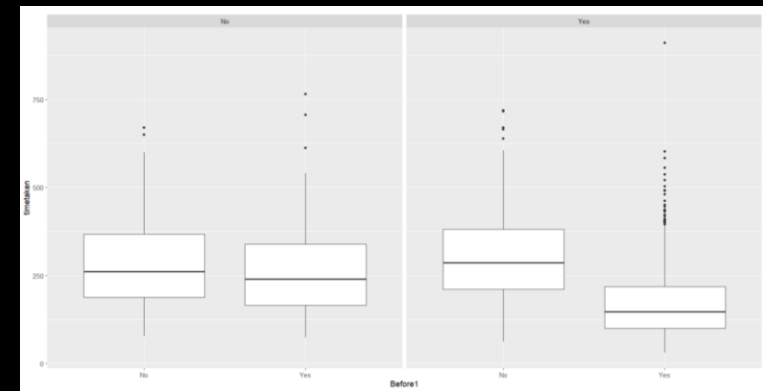
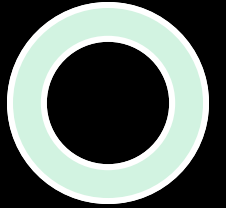
Ha: The time taken to successfully complete the sudoku is not similar for people who have done sudoku and those that have not.



We calculated the mean time taken for all types of sudoku, then applied a one sample t-test to get a p-value = 0.1795. Therefore, we fail to reject the null hypotheses.



Based on this test we can conclude The time taken to successfully complete the sudoku is similar for people who have played sudoku than those that have not.





CONCLUSION

- All the information above gives us a good insight into the factors that contribute to successful completion of the puzzles and the time taken to complete them.

