DATA101 Assignment 1, Semester 2 2025

Due Date: 25 July 2025

Student ID Number: 83972706

Total Marks: 20

Instructions:

- Complete all questions in R Markdown format
- If using this .Rmd file for your solution, please enter your student ID number in the space provided above
- Submit your completed assignment as a PDF through LEARN by 11:59pm on the due date
- Show all your R code and explain your reasoning where required
- Code that produces errors will receive minimal marks

Question One (7 marks)

Below you are given the daily step counts recorded by a fitness tracker for 10 consecutive days:

8453 6721 9546 7832 5298 10234 8765 7123 9871 6540

Answer the following:

a) Write the R code to create a vector with these daily step counts that allows you to perform statistical analysis on them. (2 marks)

```
##  [1]  8453  6721  9546  7832  5298 10234  8765  7123  9871  6540
```

b) Calculate the mean daily steps and the standard deviation of these measurements using R. (1 mark)

```
## [1] 8038.3
```

```
## [1] 1616.68
```

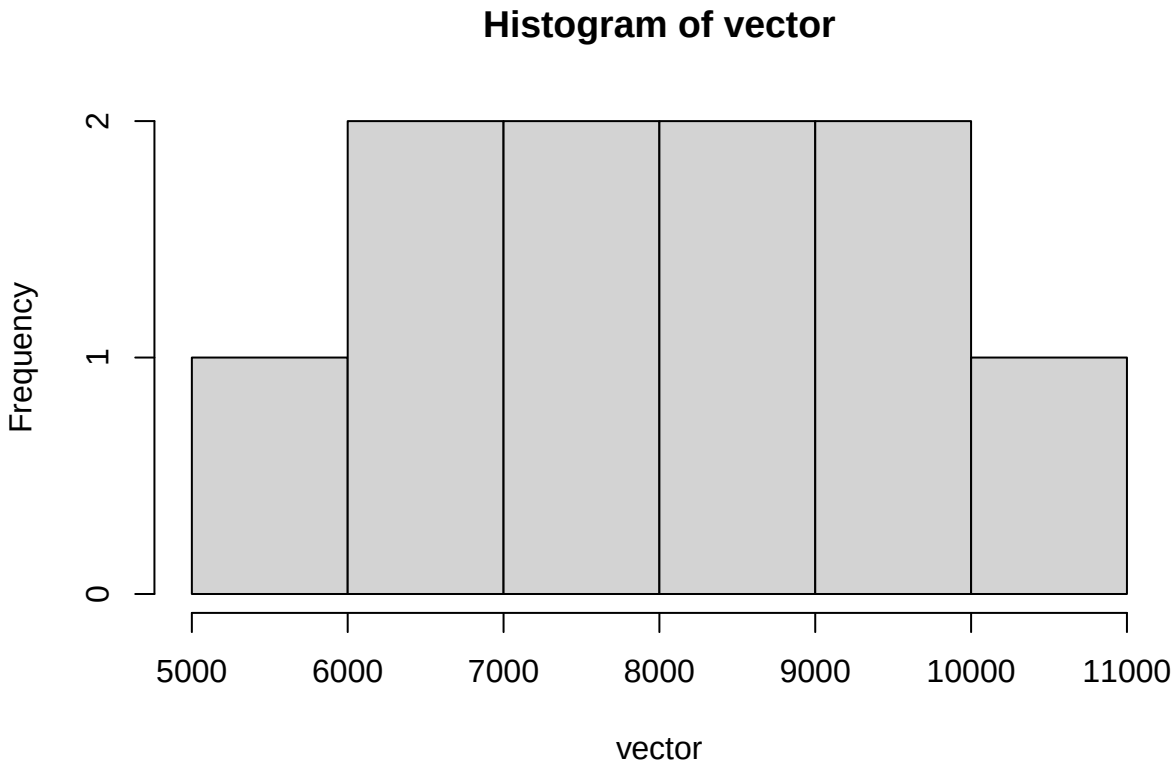
c) What is the median daily step count? Explain the steps you would follow to calculate this median manually without using R functions. (2 marks)

```
## [1] 8142.5
```

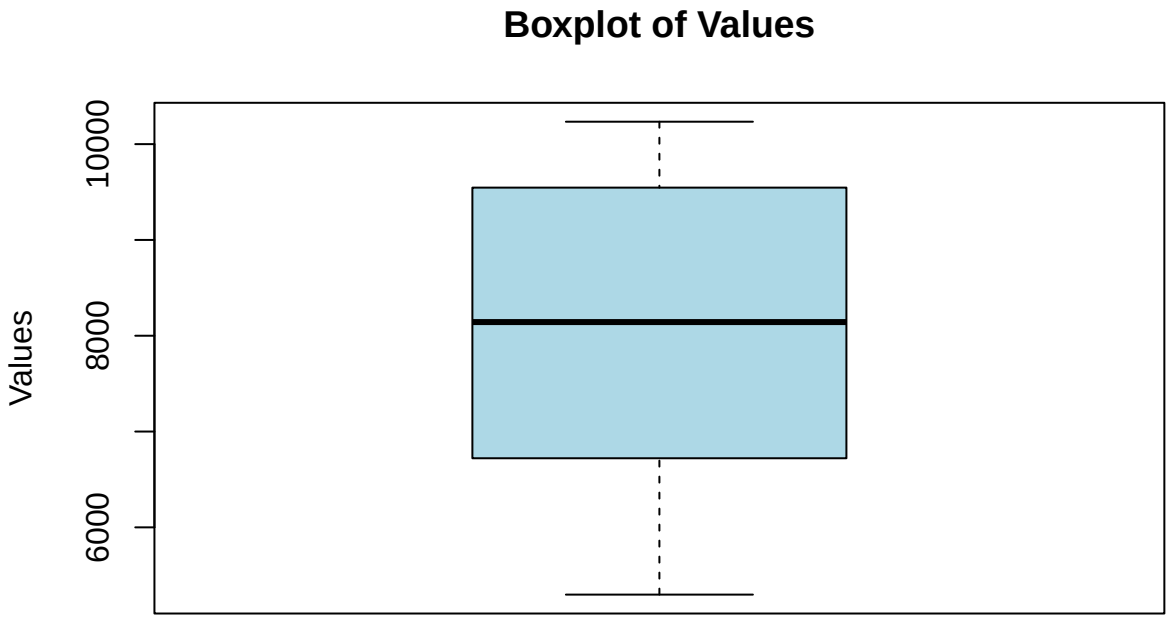
To do without R functions I would arrange the values in ascending order, if the number of values is odd then the middle value is the median, if it is even then it would be the average of the two middle values

d) Create a histogram and a boxplot of these step counts. Based on these visualizations, comment on the distribution of the data and identify any potential outliers. (2 marks)

Create a histogram of the vector



Create a box plot of the vector



The shape of the data appears to be quite symmetric.

The data ranges from just under 5,300 to over 10,200, indicating a wide spread.

There are no obvious outliers

Question Two: (6 marks)

The dataset `customer_data_2025.csv` contains information about customers of an online retailer, including their `time_on_site` (in minutes), `purchase_amount` (in dollars), `age` (in years), `clicks_before_purchase` (number of clicks), and `return_customer` (yes/no). The dataset can be downloaded from LEARN.

a) In this scenario, identify which variables could be considered explanatory variables and which would be the response variable if we wanted to predict purchase amount. Justify your answer. (1 mark)

Explanatory:

- time_on_site
- clicks_before_purchase
- age

time_on_site and clicks_before_purchase are both included as explanatory variables because they represent meaningful aspects of customer engagement. More time or more clicks may reflect greater interest or deliberation, which could reasonably influence how much a customer ends up spending (purchase_amount).

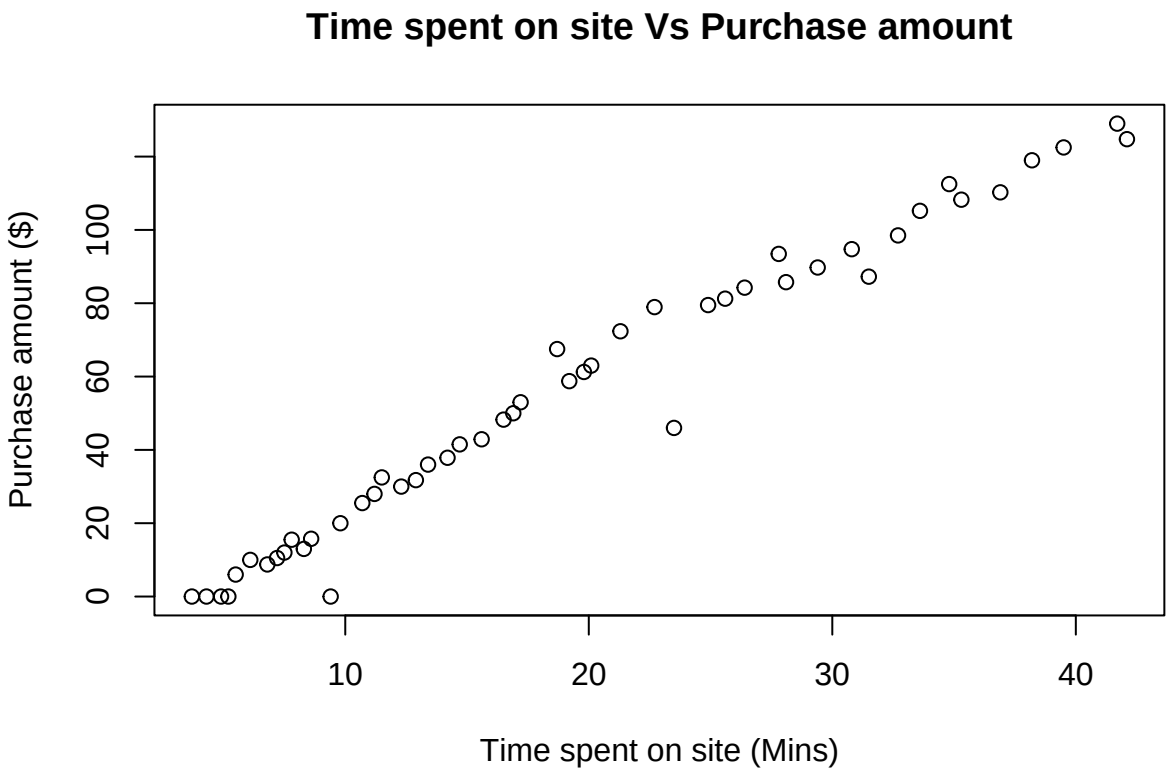
Response

- purchase_amount

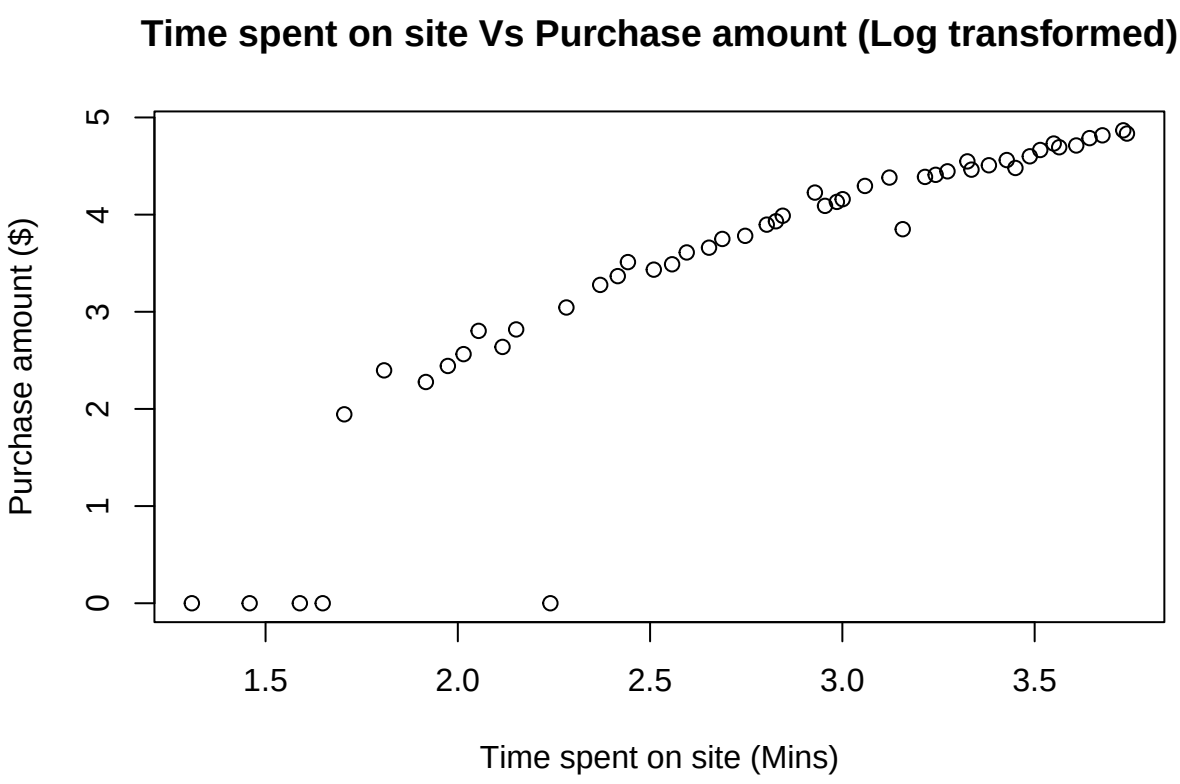
This variable will reflect how much the customer spent on the site so if this is lower we can see the customer spent less money on the site

b) Create a scatterplot to visualize the relationship between time spent on the website and purchase amount. Add appropriate axis labels and a title to your plot. Include a color dimension to represent whether the customer is a return customer or not. (1 mark)

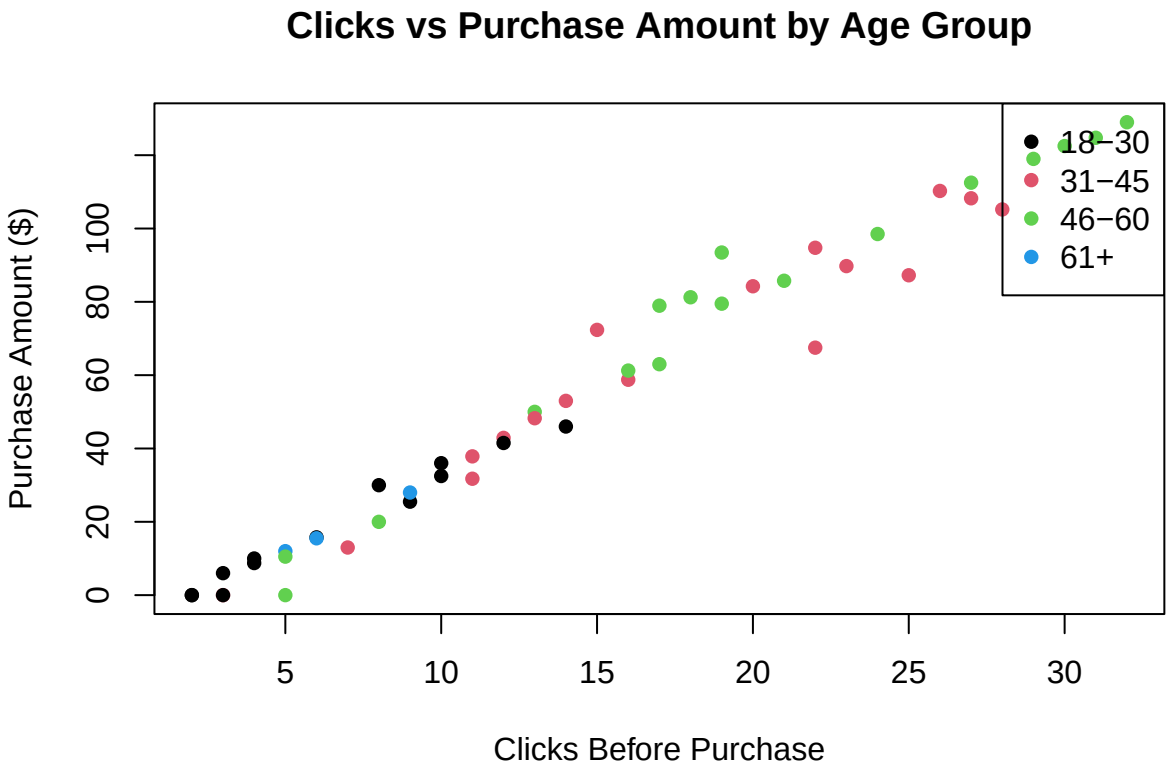
Create a scatterplot on the data



c) Transform the data by creating new variables: `log_time` = `log(time_on_site)` and `log_purchase` = `log(purchase_amount+1)`. Generate a scatterplot of these transformed variables. Additionally, create another visualization showing the relationship between `clicks_before_purchase` and `purchase_amount`, grouped by age categories (e.g., `18 - 30, 31 - 45, 46 - 60, 61+`). Explain how these transformations and groupings affect your understanding of the relationships in the data. (2 marks)


Time spent on site Vs Purchase amount (Log transformed)
Purchase amount ($)
0
1
2
3
4
5
1.5
2.0
2.5
3.0
3.5
Time spent on site (Mins)

Create a color for each age group



This allows me to see that the amount of clicks a user does on the site is reflected in their purchase amount but it appears that this is only the case when the user is $>$ 61 years of age.

d) After fitting a linear model to the transformed data, you find that:

```
purchase_amount = 2.75 * (time_on_site)^0.65 * (age)^0.12 * (0.95)^(clicks_before_purchase - 1)
```

Use this model to predict the purchase amount for a 35 year old return customer who spends 25 minutes on the website and makes 12 clicks before purchasing. (1 mark)

e) What additional information should a data scientist include when communicating the above result to the company? Consider aspects like prediction intervals, model assumptions, and limitations of the analysis. (1 mark)

They should include a prediction interval to show uncertainty, explain model assumptions (e.g. variable relationships, independence), and note limitations like potential data bias or changes in customer behavior that may affect accuracy.

Question Three: (7 marks)

You have been asked to determine the proportion of Auckland City residents who use public transportation as their primary method of commuting to work. Answer the following:

a) What is the population of interest? (1 mark)

Auckland city residents that commute to work

b) As it is not feasible to interview all residents, you are asked to draw a sample of 1500 residents from the city's database. To minimise bias, what property should the sampling procedure have? (1 mark)

The sample should be randomly selected, so that it is representative of the population and reduces selection bias

c) Are you trying to describe cause and effect, that is conduct an analytic study, or just describing, that is conduct a purely descriptive study? (1 mark)

Purely descriptive study

d) What is the population parameter in this example? How would you obtain an estimate of the population parameter in this example? (2 marks)

1550 (The size of the sample I was asked to collect) (Number of people that use transport as their primary commute)/1500

e) If you are undertaking this study now, i.e. in the year 2025, describe two potential biases that could arise in your study. (2 marks)

1. People who use public transport might be more willing to participate in surveys about commuting, skewing the results.
2. The proportion of people commuting at all may have changed significantly after the COVID-19 lockdowns, which could distort the apparent use of public transport among "commuters".