Student:

Huynh hoang tuan anh  001323726

comp 1810

[**Executive Summary** 2](#_Toc173603181)

[**Introduction** 2](#_Toc173603182)

[**Data Cleaning and Preparation** 2](#_Toc173603183)

[**Task A: Extracting Relevant Columns and Data Preparation** 2](#_Toc173603184)

[**Task B: Sales Analysis by Month** 6](#_Toc173603185)

[**Task D: Descriptive Analysis Using R** 10](#_Toc173603186)

[**Task E: Analysis Using the Starwars Dataset** 11](#_Toc173603187)

[**Conclusion:** 14](#_Toc173603188)

# **Executive Summary**

This paper offers recommendations for optimizing website traffic and income production through an analysis of clickstream data from an online retail store. To complete the assigned tasks, the dataset "e-shop clothing 20081.xlsx","shop clothing infor 2008.txt","input.xlsx" was examined using the proper data analytics tools. The report includes descriptive analysis of another dataset using R, sales and revenue trend visualization, and data cleaning. Moreover, insights obtained with R's dplyr on the "starwars" dataset are offered. The findings aim to guide decision-making for enhancing the online retail shop's performance.

# **Introduction**

This report's goal is to use data analytics methods to extract useful information from an online retailer's clickstream dataset. In order to support strategic decisions targeted at boosting website traffic and income generation, this report will look at sales patterns, revenue trends, and descriptive analysis.

# **Data Cleaning and Preparation**

Cleaning and prepping the dataset for analysis was the first stage. This includes verifying data integrity, fixing data formats, and addressing missing values. The identification of pertinent columns for the analysis was predicated on their importance in comprehending sales, revenue, and user activity on the website.

# **Task A: Extracting Relevant Columns and Data Preparation**

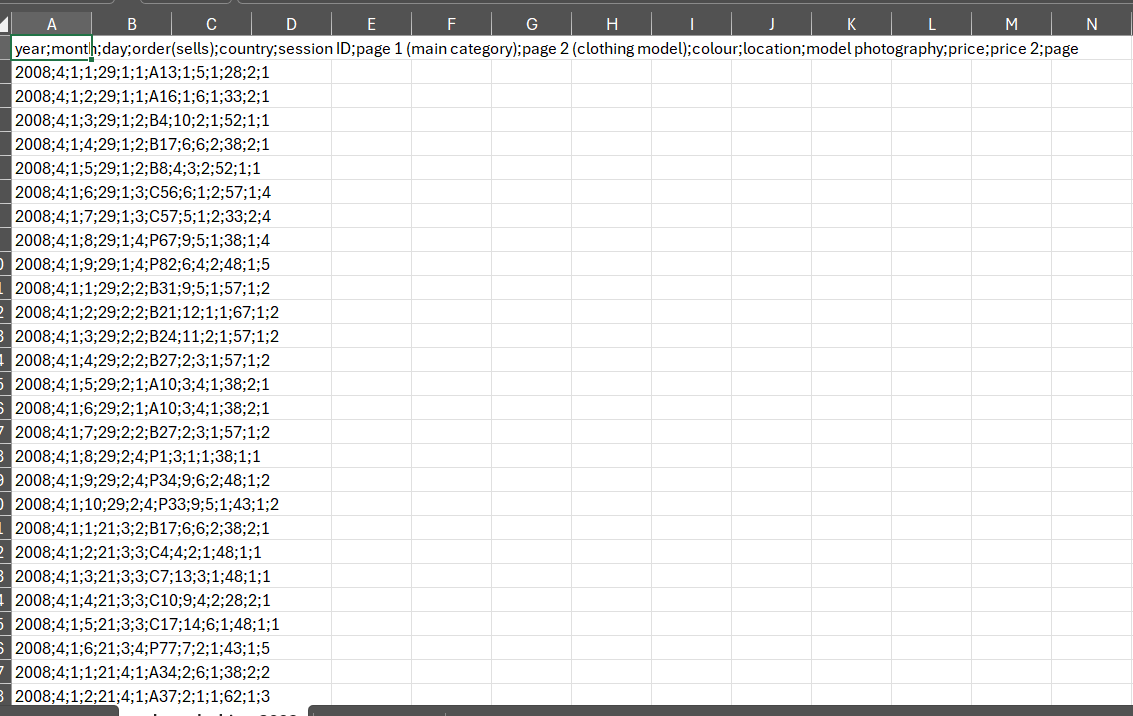
The columns extracted for analysis were chosen based on their relevance to understanding sales and revenue metrics. Key columns included product year, month, day, order, country, page 1 (main category), page 2 (clothing model), colour, location, model photography, price in us dollars, price 2 -> variable informing whether the price of a particular product is higher than

the average price for the entire product category, page -> page number within the e-store website (from 1 to 5) These columns provide insights into customer purchasing behavior and revenue generation trends.

This is the raw of the data, we can see that is the challenging for us to understand clearly which information is provided in the excel file, so initial step is make it look manageable. By using text to column in Data tab, don’t forget select the column contain all the data in this case all data have written on column A.

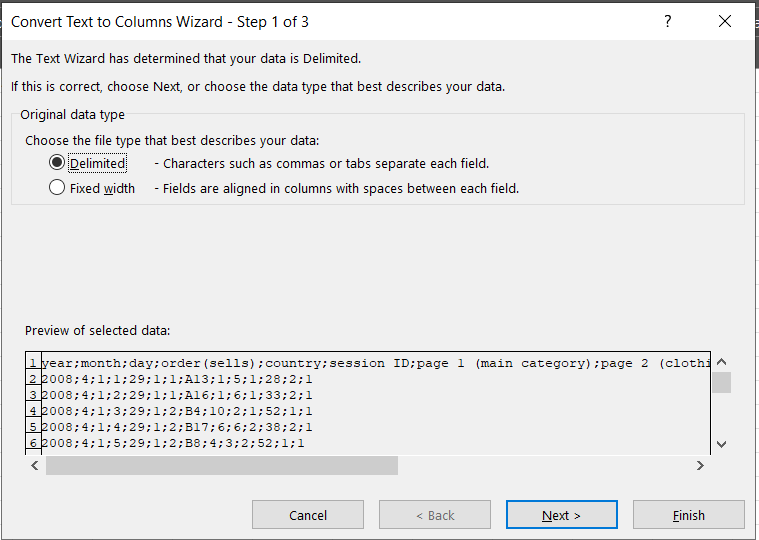
Do the same with the the ‘**input.xlsx’** excel file but the **input.xlsx**

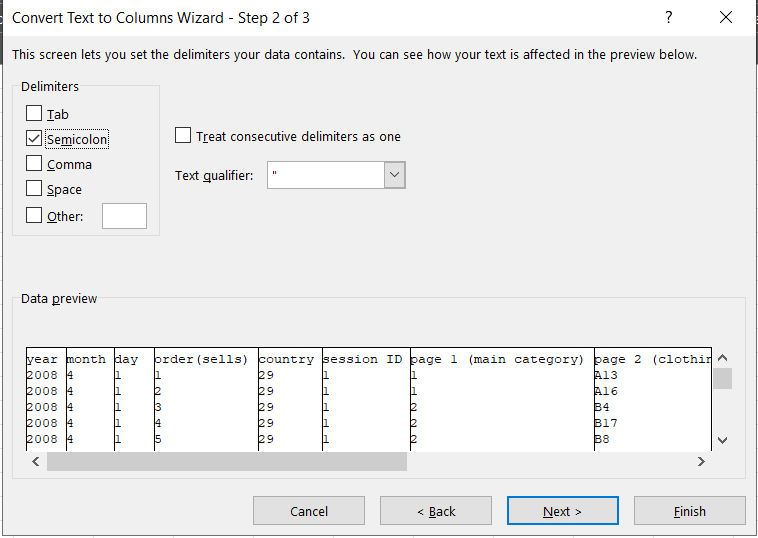
**Preparing For Tasks:**

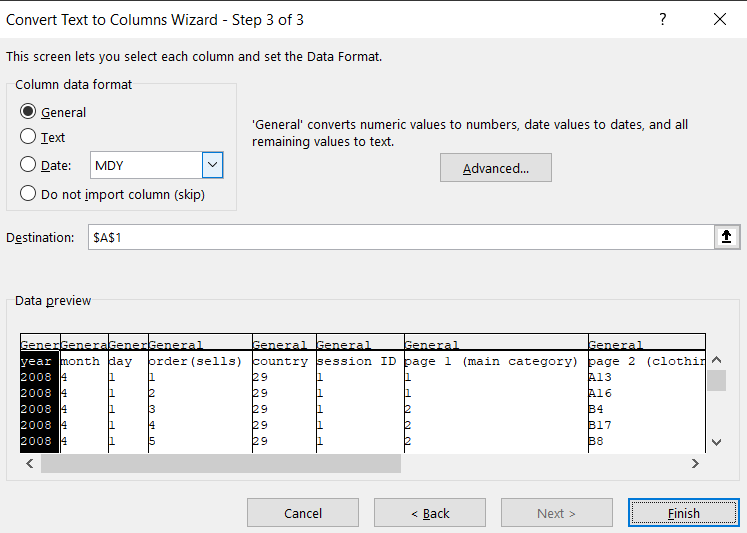


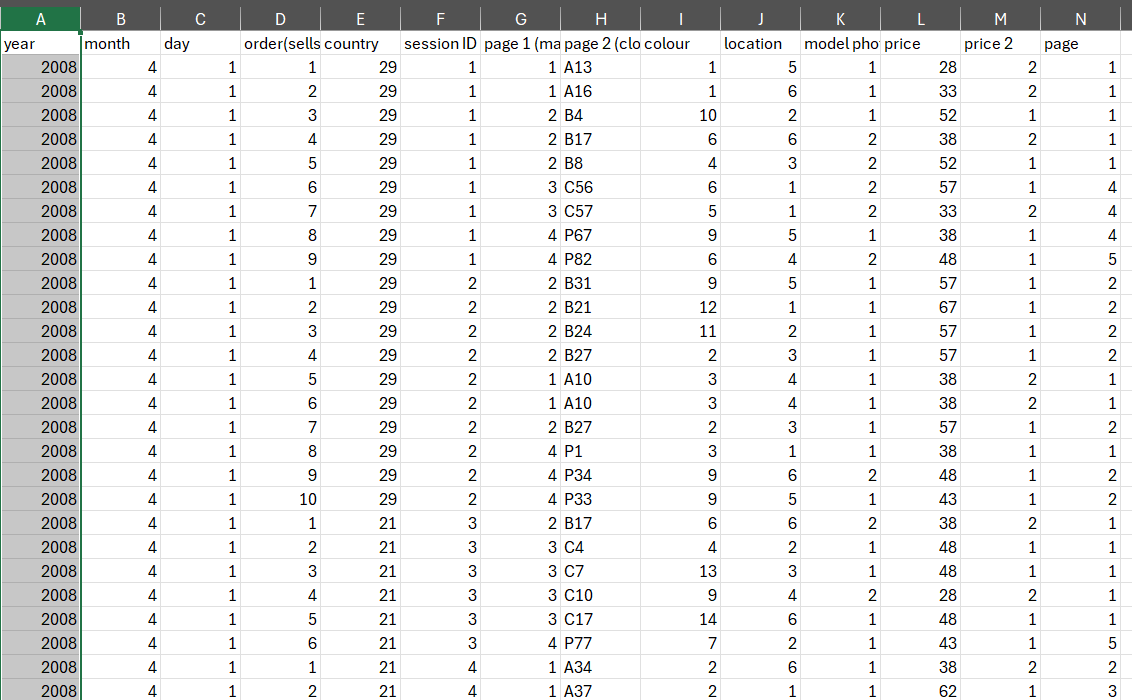
\

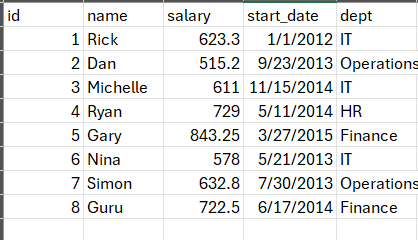
Choosing delimited



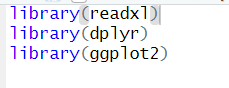
Select semicolon



Click Finish and let excel handle all for you:  


**Input.xlsx:**After doing the same like the steps before you will get result   


**Data Reading Prepare and Chart:**

****

Using **readxl** library help us in reading excel file, **ddlyr** help us doing the task E which provide the starwar.csv and **ggplot2** provided us creating chart function

**Prepare Data for Tasks B and C:**  

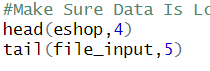

In **readxl** library provided us **read\_excel** function that help us to read excel file, in that function you have to integrate into the function the location of the excel file, in this case the R file and the excel file is the same location so that we just insert into the name of the file. Furthermore, reusing the excel file for another tasks I stored it to eshop variable.

**Prepare Data for Task D:**

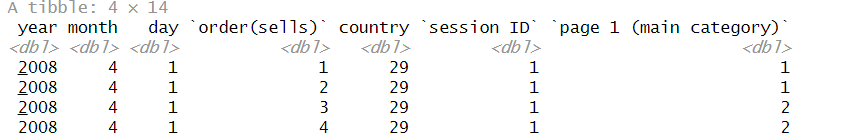
****

Do the same prepare data for tasks B and C steps

**Make sure the data is prepared**



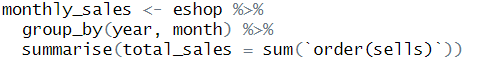
Using the head function and tail function in helping us knowing the data have been loaded, if the data is loaded, it led to this result



# **Task B: Sales Analysis by Month**

Using appropriate charts, the sales trends by month were visualized. This analysis helps in understanding seasonal variations in sales and identifying peak months for revenue generation. The insights gained contribute to planning marketing campaigns and optimizing inventory based on demand fluctuations.

After loading the data, I will use **ggplot2** library to create the chart the task B, before showing the chart we have to determine the right column, in this case calculating the sales by month, I use year, month and total sale of the month.



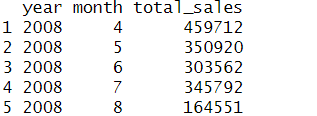
**eshop %>%:**   
To begin, open the eshop data table.   
To transport data through a sequence of operations, use the %>% operator.

**group\_by(month, year):**   
The dplyr function group\_by is used to group data according to one or more variables.   
Year and month are the two columns used to group the data in this instance.   
This implies that the year and month values will be used to categorize all of the rows in the eShop table.

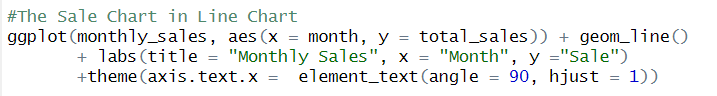
**summarise(order(sells) + total\_sales):**   
The dplyr function summarise is used to combine previously grouped data groups into a summary table.

**Order(sells) + sum(total\_sales):**   
Determine the total value of each group's order(sells) column, or for each pair of years and months.   
In the summary table, a new column named total\_sales is given this total amount.

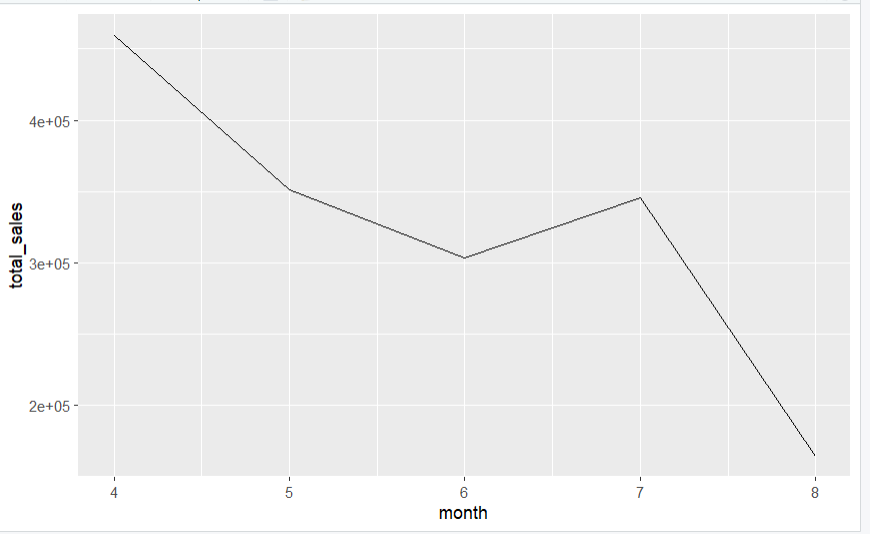
**Showing the mothly\_sales:**



**Showing the Chart:**

****

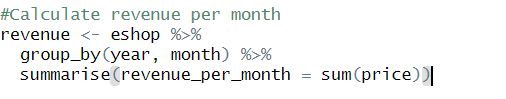
Using **ggplot** to provide the chart, in this case I want to show the trend of the sale by month so I use the Line Chart



**Task C: Revenue Analysis by Month**

Revenue trends were analyzed by aggregating sales data for each month and plotting total revenue over time. This visualization highlights revenue growth patterns and identifies opportunities for increasing revenue through targeted strategies such as promotional offers during peak sales months.

Do the same Task B, in this case I determine that I have to use year, month and price column.



**eshop %>%:**

Start with the eshop data table.

Use the %>% operator to move data through a series of operations.

**group\_by(year, month):**

group\_by is a function in dplyr used to group data by one or more variables.

In this case, the data is grouped by two columns year and month.

This means that all rows in the eshop table will be grouped by year and month values.

**summarise(revenue\_per\_month = sum(price)):**

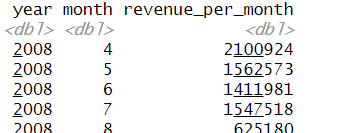
summarise is a function in dplyr used to create a summary table from previously grouped groups of data.

**revenue\_per\_month = sum(price):**

Calculate the total value of the price column for each group (that is, for each pair of year and month).

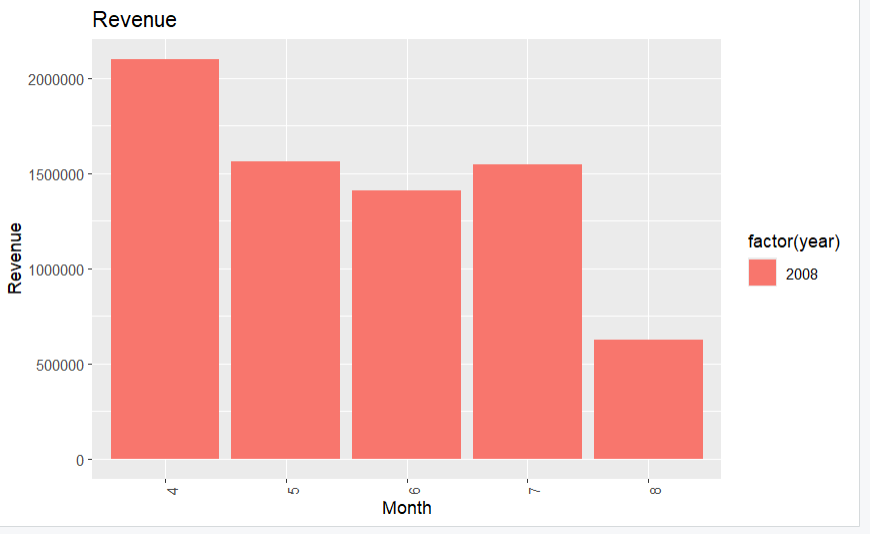
This total value is assigned to a new column called revenue\_per\_month in the summary table.

**Showing the revenue\_per\_month:**

****

**Showing The Chart:**

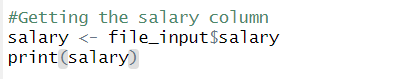
In this case I would like to use bar chart to show the comparation of the revenue monthly.



# **Task D: Descriptive Analysis Using R**

A separate dataset ("input.csv") was analyzed using R to compute descriptive statistics such as mean, median, mode, standard deviation, and variance of the salary column. These statistics provide a comprehensive overview of the salary distribution within the dataset, aiding in workforce planning and salary benchmarking.

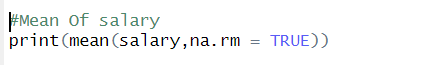
**Get the salary column:**

****

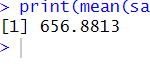
file\_input have been loaded before and **$salary** is the column of the file\_input make sure it contains the column or you will get error



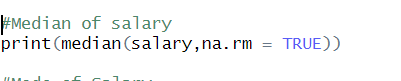
**Mean of the salary:**

****

**Result of Mean:**

****

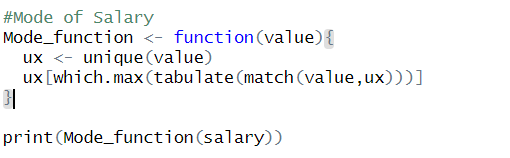
**Median of salary:**

****

**Result of Median:**

****

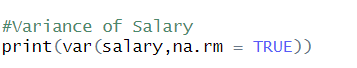
**Mode of Salary:**

****

**Result of Mode:**

****

**Variance of Salary:**

****

**Result of Variance:  
**

**Standard Deviation:**

****

**Result of Standard Deviation:  
**

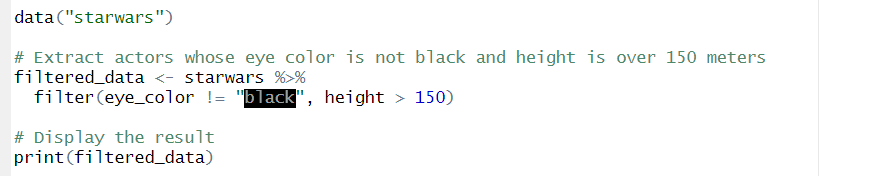
# **Task E: Analysis Using the Starwars Dataset**

The "starwars" dataset from the dplyr package in R was utilized to perform specific analyses:

**Part i:** Actors whose eye color is not black and height is over 150 cm were extracted. This segmentation helps in identifying distinct groups within the dataset based on physical attributes.

**Part ii:** A new column for Body Mass Index (BMI) was added to the dataset using the formula BMI=mass(height100)2BMI = \frac{mass}{(\frac{height}{100})^2}BMI=(100height​)2mass​. A scatter plot of height against BMI was created to visualize the relationship between these variables. This visualization aids in understanding the distribution of BMI across different heights among the actors.

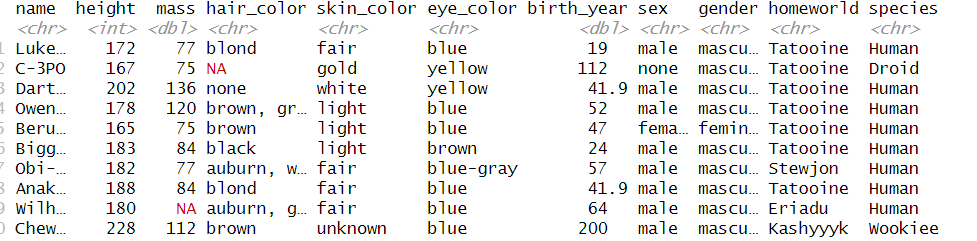
**i:**

****

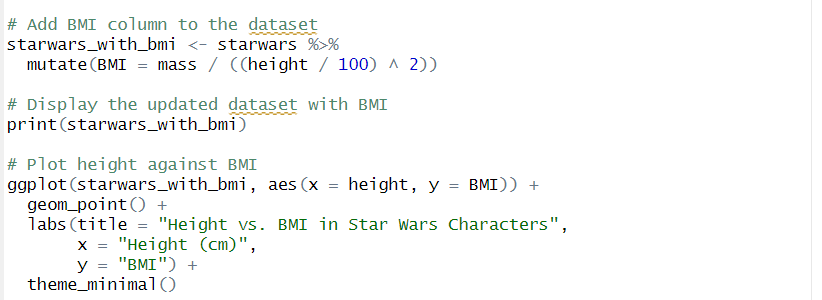
In this part, I used the **data(“starwars”)** to get all data inside starwars which mean starwars is already existed inside Rstudio.

Using **%>%** I can get all data in starwars then I use filter function to get the eye\_color which is a column in R and height is a column, then I stored it to **filtered\_data.**

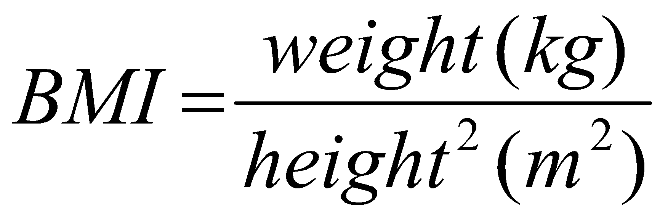
Finally, I used the **print()** function to display the result. After that, you will get this result.

****

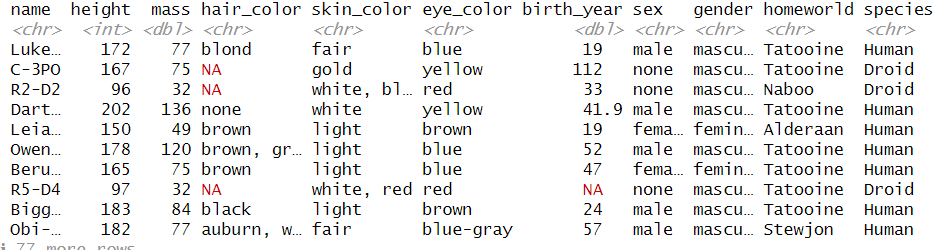
**ii:**

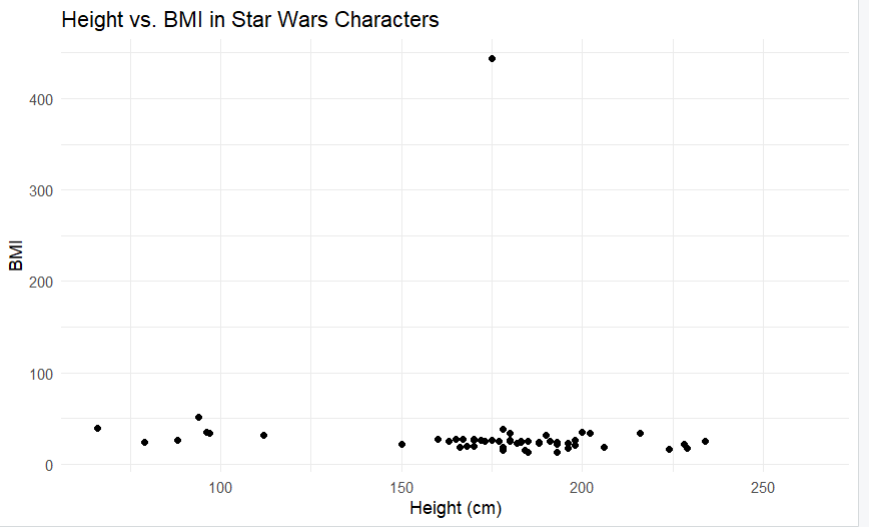


To get the part **ii** done, we have to add BMI column to the dataset, applying the same methods in the part **i** but we have to calculate the BMI with this formula:



In this case, mass is the weight then we display the result, you will get this result, if you do it right.



After that, I use **ggplot** to create visualization the BMI that we have gotten before:  


In this graph you can see the consistence base on the BMI that we have visualized

# **Conclusion:**

The clickstream information from an online retail company was thoroughly analyzed in this research, which offers insightful advice on how to maximize website traffic and income production. Following thorough data preparation, cleaning, and visualization along with descriptive statistical analysis in R, the following main conclusions and suggestions have been established:

**Data Preparation and Cleaning:**  
  
A crucial first step was guaranteeing the usability and integrity of the data. This required dealing with data types, dealing with null values, and extracting pertinent columns that were essential to comprehending revenue and sales indicators.

**Trends in Sales and Revenue:**  
**Task B**: Seasonal fluctuations and peak sales months were revealed by utilizing line charts to visualize the monthly sales trends. Planning successful marketing efforts and controlling inventory to meet demand require these knowledge.  
**Task C:** Bar charts were used to show monthly income trends and emphasize patterns of revenue growth. This study helps find ways to increase revenue by using focused tactics, including special deals during months with high sales.

**Analyzing Descriptively:**

**Task D:** Using R, descriptive statistics such as mean, median, mode, variance, and standard deviation were calculated for the wage column from a different dataset. These figures help with personnel planning and salary benchmarking by giving a thorough picture of the salary distribution.  
**Analyzing the Starwars dataset:**  
**Task E(i):** We extracted actors with non-black eye color and heights greater than 150 cm from the "starwars" dataset. This division facilitates the identification of discrete groups according to physical characteristics.  
**Task E(ii):** A scatter plot of height against BMI was made and a new BMI column was added to the "starwars" dataset. This graphic offers important insights into the relationship between height and BMI by illustrating how BMI is distributed throughout a range of heights.