# 



# School of InfoComm Technology

**Data Visualisation**

Diploma in Financial Informatics

October 2020 Semester

**ASSIGNMENT 2**

**(Individual Assignment)**

**Submission Deadline:**

**11th February 2021 (Thursday), 12:00PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **P01** |
| **Student Name** | **:** | Tan Zi Xu |
| **Student Number** | **:** | S10185168 |

**Penalty for late submission:**

10% of the marks will be deducted every calendar day after the deadline.

**NO** submission will be accepted after 19th February (Friday), 10AM.

Contents

[School of InfoComm Technology 1](file:///D:\NP%20FI%20Year%202\Sem%202\DV\Assignment%202\DV%20Assignment%202%20Cover%20Page.docx#_Toc63891710)

[Project Objective 4](#_Toc63891711)

[Data Preparation 5](#_Toc63891712)

[Initial Overview 5](#_Toc63891713)

[Importing Data 6](#_Toc63891714)

[Data Cleansing 6](#_Toc63891715)

[Null Values 7](#_Toc63891716)

[Wrong Input 11](#_Toc63891717)

[Wrong Datatypes 11](#_Toc63891718)

[Adding Column 12](#_Toc63891719)

[Column Encoding for Heatmap Prep 13](#_Toc63891720)

[Visualisation 15](#_Toc63891721)

[Univariate 16](#_Toc63891722)

[Product 16](#_Toc63891723)

[Branches 17](#_Toc63891724)

[Age & Age Group 18](#_Toc63891725)

[Gender 20](#_Toc63891726)

[Education 21](#_Toc63891727)

[Marital Status 22](#_Toc63891728)

[Usage 23](#_Toc63891729)

[Fitness 24](#_Toc63891730)

[Income & Income Bin 25](#_Toc63891731)

[Miles & MilesPerUsage(MPS) 26](#_Toc63891732)

[Multivariate 27](#_Toc63891733)

[Heatmap 27](#_Toc63891734)

[Education & Income 28](#_Toc63891735)

[Age & Miles (product) 29](#_Toc63891736)

[Miles & Income (product) 30](#_Toc63891737)

[Viz for dashboard 32](#_Toc63891738)

[Is there any gender preference to a particular product? 32](#_Toc63891739)

[Is there any age preference to a particular product? 33](#_Toc63891740)

[Does income affect customer preference to a particular product? 34](#_Toc63891741)

[What is the product distribution amongst the branches? 35](#_Toc63891742)

[What is the age group distribution amongst the branches? 36](#_Toc63891743)

[What is the income class distribution amongst the branches? 37](#_Toc63891744)

[Is the customer mindful of their fitness in relation to any personal variables? 38](#_Toc63891745)

[Is their target goal reasonable? 39](#_Toc63891746)

[How do they plan on achieving the target? 40](#_Toc63891747)

[Is there any usage difference between genders; marital status a factor? 41](#_Toc63891748)

[Dashboard 42](#_Toc63891749)

# Project Objective

In this project, I am market research analyst who has just joined Cardio Good Fitness (CGF), a treadmill retail business. While I do have a set of data on CGF’s customer, I still lack understanding of CGF’s product and customer demographic. As a market research analyst, those are vital information that I will be working with. In my opinion, product understanding may be important but it is not as important as customer understanding. Hence, I have a set of research questions that I want to answer through visualisation. The questions are categorized by 3 different topics, but they are all designed to give the marketing team meaningful insight.

The 3 different topics are mainly focused on the customer aspect. Such that, I will mainly discuss on the customer’s demographic, preference, their goal setting or usage tendency and any region based analysis.

**Demographical / Preferential Product Based Questions:**

* Is there any gender preference to a particular product?
* Is there any age preference to a particular product?
* Does income affect customer preference to a particular product?

**Geographical Based Questions:**

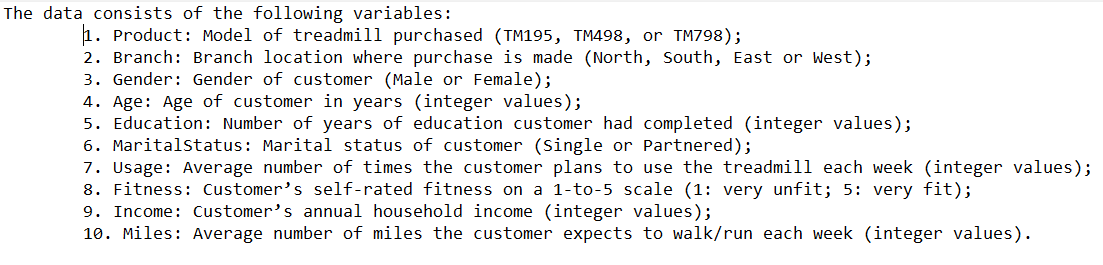
* What is the product distribution amongst the branches?
* What is the age group distribution amongst the branches?
* What is the income class distribution amongst the branches?

**Customer Goal / Usage Based Questions:**

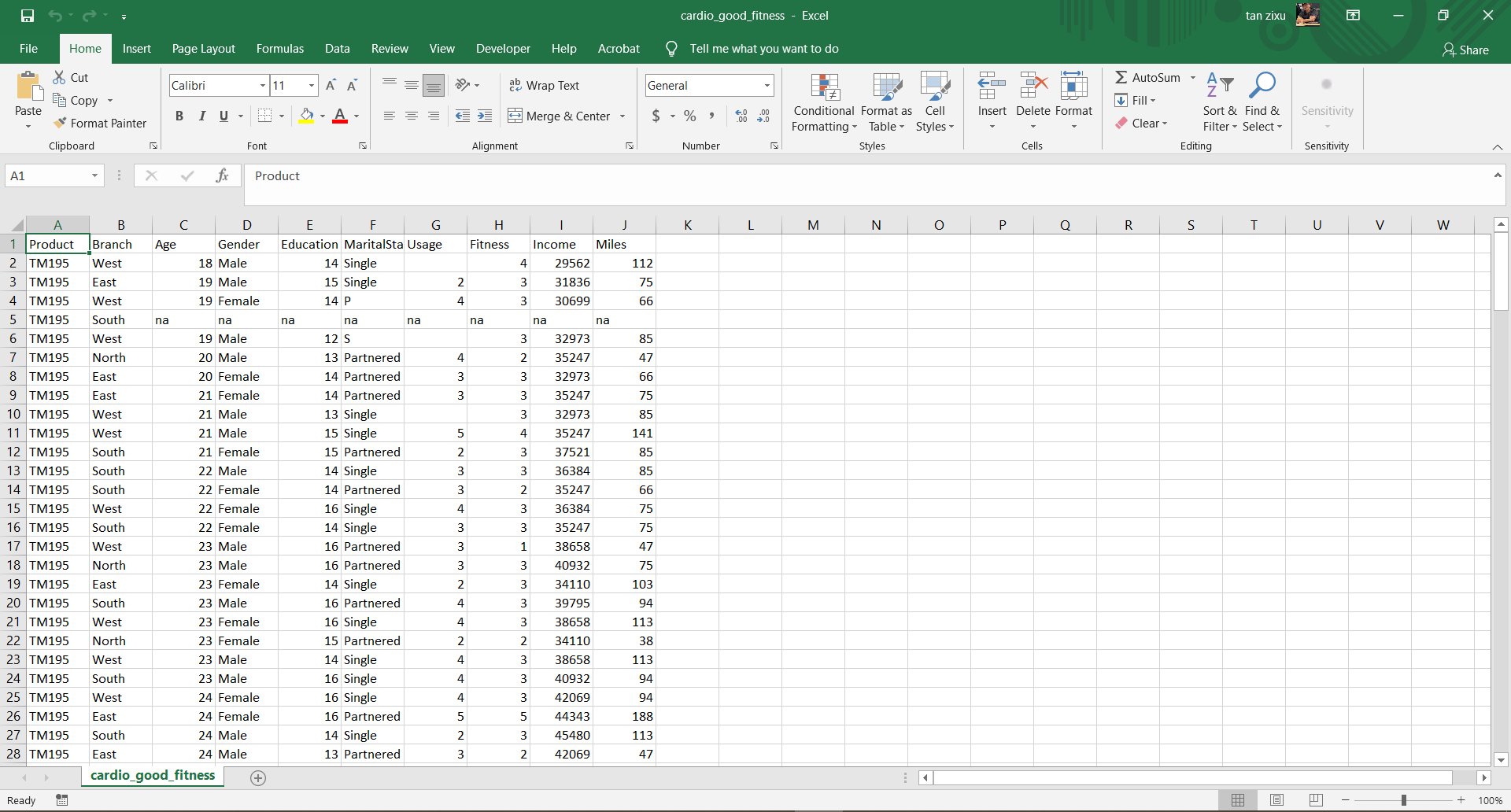
* Is the customer mindful of their fitness in relation to any personal variables?
* Is their target goal reasonable?
* How do they plan on achieving the target?
* Is there any usage difference between genders; marital status a factor?

# Data Preparation

## Initial Overview



Before I start any work related to the data preparation or think about visualizing it, I would always refer to the data dictionary first. As this first step will help me grasp the data I will be working with and it would also serve as a reference point at any stage of the project. Next, I will take a look at the data in its raw state, in the form that I received it in. This helps me get a feel of the data structure and through this brief scanning, I would try to identify any noteworthy finds.



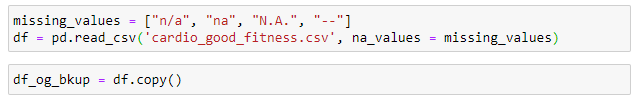
In this brief observation, I find the overall format looks to be clean and organised, the data file is also in a csv format. As the structure is similar to what the data dictionary stated, there doesn’t seem to have a lot of weird data inputs randomly.

Following up, I noted that there were some cells listed as “na”, I believed it is supposed to indicate the cell is to be a null value. This was an important find as the library, ‘Pandas’, that I will be using for Python programming on the Jupyter Notebook platform does not recognise “na” as null values.

Another find was that some rows had a lot of those supposed null values. A few recurring thoughts was whether to drop those rows or assign common values in the data set. However, the decision will be made later on.

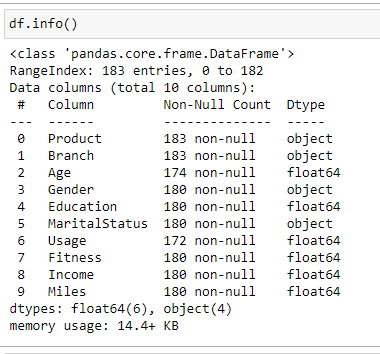
## Importing Data

Moving onto to actually working on the data, the csv file will first be imported into the python file. Utilising the pandas library, the data file will be read and stored as a data frame, the core component in this project for subsequent work with data visualizing. However, due to how pandas are unable to recognise certain values as null, a set of input was made for the reader will recognise the null values. Without this crucial step, the data frame will assign columns with “na” as object when they are supposed to be integers or decimal numbers. If this issue was not spotted and I was unaware of the “na” inputs, I would also be unable to switch the data type of affected columns. Lastly, I made a backup copy of the untouched data frame just in case I need it in the subsequent steps.

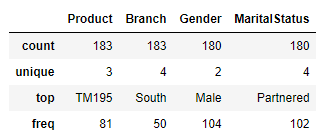


## Data Cleansing

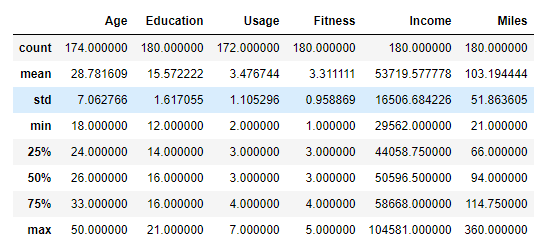
With a data frame, I am now able to use various functions for various goals in manipulating the data. First of such goals was to get a consolidated understanding of the data frame. In this case, I used the info function to consolidate various basic information of the data columns within the data frame.



From this output, I identified that there are 10 data columns and 183 rows. 4 of the data columns are also of the object data type and the rest of which are decimal numbers. From here, I noted there are definitely some data columns with null values out there. Regarding data type, the 6 columns with float64 are clearly in the wrong data type. As stated in the dictionary, those columns are to be in integer value. On the other hand, to get more in-depth information of each column, I need to use the describe function. Due to the differing data type amongst data columns, columns with object data type will run on a separate describe function.



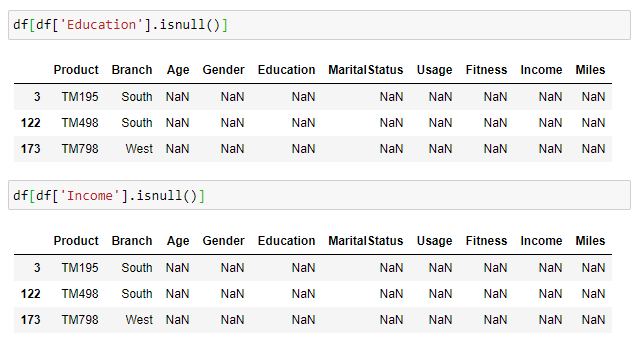
Running through the describe output of object data columns, only MaritalStatus seems to have a unexpected output. When referred to the data dictionary, the other 3 columns has the expected unique output matching the number of different input the columns could have. MaritalStatus on the other hand has 2 additional type of unique input in this data frame, adding on to the list of what needs to be prep.



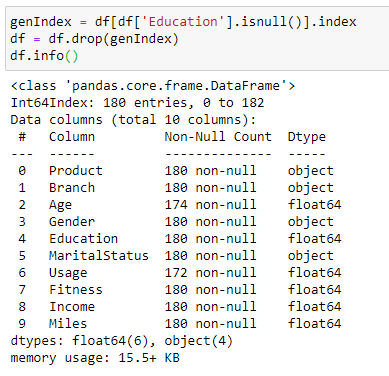
On the surface, the describe function output for float64 data columns seems to be normal. However, there are lingering hints that the columns hold values with decimal number and this has to be sorted out as the dictionary states the columns are to be integer data type. I also foresee that I will need group up or bin certain columns so that it is easier to perceive the column.

### Null Values

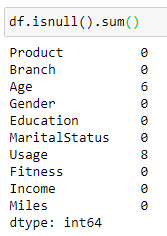
First off, I want to deal with the rows that has null value in multiple columns. When recalling about the brief observation of the raw data, those rows seem to be missing everything except for Product and Branch. At the same time, a few of the other columns seems to be have the same number of null counts. I believe this coincident might be what I am looking for and I check for those rows with two of the suspected columns, Education and Income.



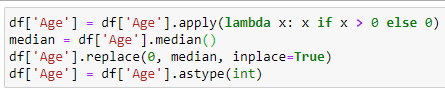
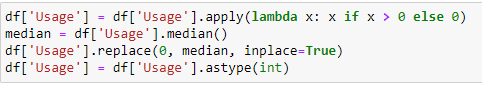
Here I found data entries with null value in multiple columns along with their index. The index from both searches also matched, which leads me to believe it will be similar result even if I search the other columns. From here, I decided to drop those 3 data entries. The primary reason being, inputs that I would have to assign are too much. Even if I did assign values and classify it as some type of generic data entries based on common inputs from the whole data set, there are additional layer of work added on. I would have to assign common values from the same Product and Branch it was from. Which is for consistency and integrity purposes, but there only 3 record that needs this and they are all different products from each other. The workload of it is not justified for its size, if there were more records of such cases, I would have done it. Speaking on the number of these null entries, 3 out of the 183 data entries only amounts to 1.63834% of the total dataset. Which would be fine if I dropped the rows, considering the context does not make it so I have to salvage each and every single entry. Hence, the decision was to drop those 3 data entries, leaving me with 180 entries left.



Moving on to fix other null values, I used the sum function on the isnull function. Which has help me consolidate the number of null cells within each column of the data frame.

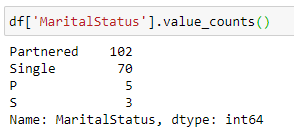


It is here that I identify the remaining null values are found in Age and Usage columns. For the Age column, I do not believe there is much to discuss on with what I can do except for assigning an appropriate value such as median. However, there is a couple things to consider when dealing with the Usage column while avoiding to drop the data. I could either assign 0 or the median value of the Usage column. On the surface, there is an argument that the usage could not be 0 as there is a Miles input, making it unlikely that a customer did not use it when it states miles were recorded. However, referring to the dictionary will show that both columns are independent from one another. As they are both what the customer intends to do or expectation, but at the same time it does not mean both columns run in parallel and are not correlated in the slightest bit. Hence, validating assigning 0 as a viable solution but, I opt for giving the median value instead. As the context was usage is the average number of times the customer plan to use the treadmill each week. Which raise the question, “Who in the right mind will buy a treadmill and not use it?”. To answer that, some possibilities could be resellers or people are buying the product as a gift thus, not knowing what should be the usage. In either case where who or why they bought the treadmill, it will still likely be used in a way, making the median a more sensible choice instead of 0.

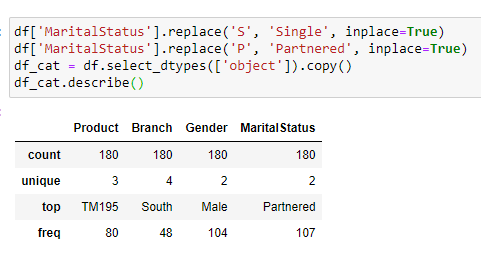
 

### Wrong Input

Next up, I will deal with wrong inputs, referring back to the use of the describe function earlier on, MaritalStatus seems to have 2 additional type of inputs. Using the value\_counts function, I can list out the unique inputs along with the frequency of them appearing.

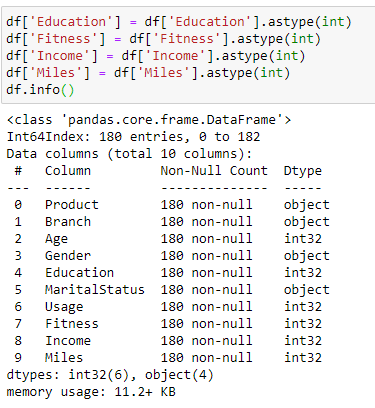


Based on the function output, I am assuming that the “S” input is representing the “Single” input and “P” to “Partnered” as another form. To resolve this, I make use of the replace function and standardize the inputs into the 2 specified inputs mentioned in the dictionary.



### Wrong Datatypes

Lastly, the final part of my cleansing to-do list, I have to switch the appropriate columns into the correct data type, integer.



## Adding Column

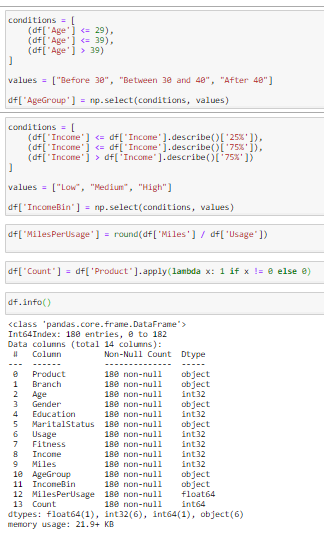
In this project, I have determined I need to add a few more columns to assist me in creating visualization, some of which could aid in understanding the customer better.

The first new column that I will introduce is, “AgeGroup”. This column will classify the age of the customer into ranges that is easier group up. The ranges will be set in a way I can classify customer who are in their 20’s, 30’s or beyond 40 years of age. I find this set of classification appropriate to represent part of the customer demographic. As it shows the changes from young adults and progressively entering the ‘middle age’ class.

Next, I have the “IncomeBin” column, I find the range of the income was rather huge, in a sense it is big in terms of quantity but not the gap between minimum and maximum. As I would not care too much about the gap in a continuous data column, I find it more important to classify the range into categories. Which will make it easier to understand how the income of our customer stands across the given dataset. The categories I set were “Low”, “Medium” and “High”. It will correspond by assigning low to below the 1st quartile, medium is the interquartile range and high would be anything beyond the 3rd quartile.

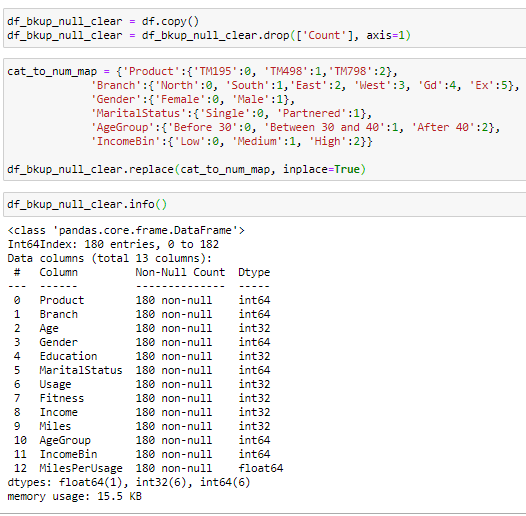
Thirdly, I want to introduced the “MilesPerUsage” column. While I believe the Miles and Usage column are likely to be related in one way or another, I think this column will assist in answering one of the research questions. The column is made by taking the Miles input divided by the Usage input.

Lastly, the final column to be introduced is “Count” column. It will serve a general-purpose column that indicates the validity of the data entry. It will be used for counting purposes when plotting charts and visualizations.



## Column Encoding for Heatmap Prep

As part of the multivariate analysis, I plan to use a heatmap and identify some correlation using heatmap. To do that, I will need to encode my object data columns into integers. Making use of a dictionary and the replace function, I can map the new encoded values to replace the old given values. Prior to the encoding, I first create a copy of the data frame for encoding, this way, I can still use the original values for plotting the other visualizations. I would also need to drop the Count column in the copy as it is a column introduced assist in plotting charts and it serves no purpose towards the correlation.



# Visualisation

In this section, I will be plotting out a series of visualization sets for univariate analysis, multivariate analysis and the set for answering the research question.

The series will start with a set of simple charts for the univariate analysis, these simple charts include the likes of vertical bar charts, boxplots, pie charts and histograms. There will 13 individual visualization corresponding to the 13 columns, including the newly added columns. In these set, I aim to get a basic understanding of CGF’s product and its customers, the visualization will give an individual breakdown of each data column. As I will be speeding through this section, I will first provide the general reason for choosing any of those charts for the visualization. The bar charts are used for data column that I deem to be categorical regardless if it is in numeric form as the input. The bar charts are vertical instead of horizontal as in each categorical data column, there really isn’t any column with a wide multitude of unique input. Which stylistically makes sense for me to read it vertical in my opinion instead of horizontal, which I would use if there were more unique inputs. The pie charts are used for data columns with binary inputs regardless of alphabetical or numerical form. As I believe in such cases with binary input or output, a pie chart would feel like a better fit compared to a bar chart. As for continuous data column, the choices are split between boxplots and histograms. Boxplot will be used for columns that I want to search for outliers and histograms will be used for columns with unusually large range as compared to other columns.

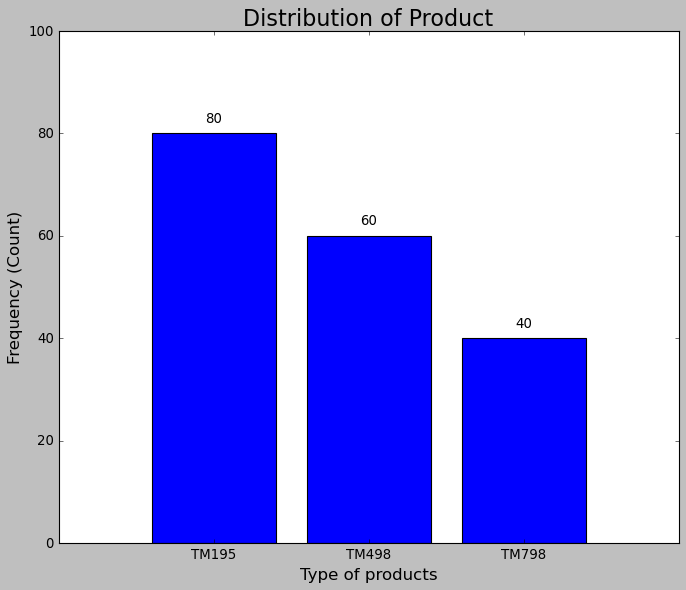
In the next set, the multivariate analysis will begin with the heatmap that was intended to be part of this analysis. The heatmap is an enabling visualization for reference on correlation between the columns. It will serve as a basis as to which relationship between the columns that I should explore into. Most of the visualization here will scatterplots to study the correlation at deeper context compared to the heatmap. The aim of this analysis is to help get a deep understanding after a shallow run through in the univariate analysis. The exploration behind the correlation of multiple variables can provide me insight as a new analyst to this data set.

Lastly, the final set of visualization will be the supporting viz for those research question set above. The charts used are primarily bar charts, violin & swarm plot combo and scatterplots. These supporting visualizations will be consolidated into dashboards, the consolidation will be organised what needs to be seen or discuss within the topic context. The choices for the type of visualization here was much more limited as compared to the other sets. This is because I want the visualization to be as simple as possible, I want the information to be visually straightforward. Hence, I am using charts that I have done before, that are simple and easy to understand. To elevate the mundane feels of such simple charts, colours were chosen to brighten both the visualizations and dashboard in its overall view.

## Univariate

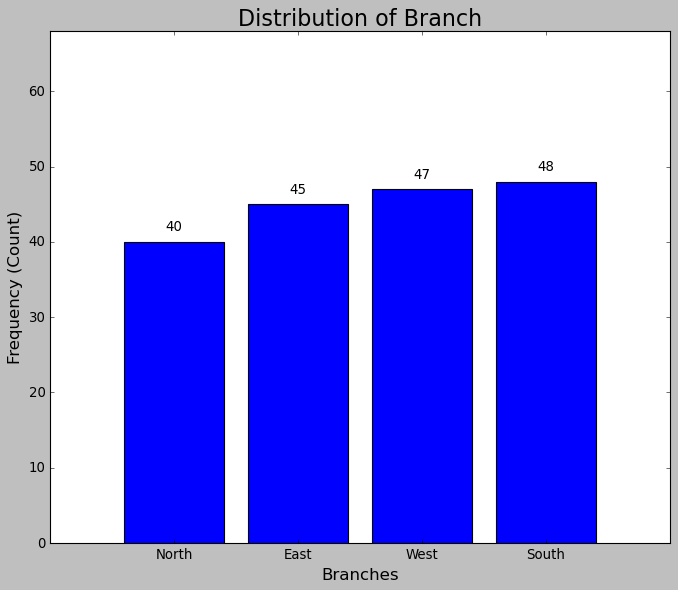
To start things off in analysis, this univariate analysis section will be a shallow layer of analysis to understand each individual data column. This section is not designed as in-depth exploration but it will go through some obvious observation that can be seen in the visualization of this set. There will be some speculation or thoughts that I have on the data column or even light recommendation on the topic.

### Product



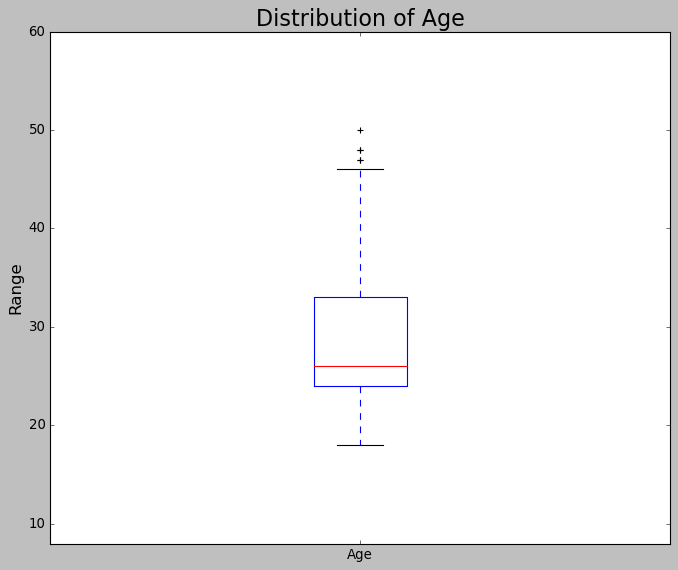
Looking at this bar chart, I can understand how well each of the product line is doing. At the same time, I have some question, doubt or exploration that I would like to delve into with the subsequent visualizations. From here, we can see the highest selling product is the “TM195” product, which almost comprised half of treadmill sold. As a new analyst, I want to know how the products were advertised or marketed. This is because looking at this chart or statistic, I want to know what variables or feature set the products apart from each other. It is important as it briefly taps into what our customer are looking for in the machine they want to buy. Developing on this point as part of exploration will help me grasp the trend and better understand both products and consumers. Other than that, this particular visualization does not offer much on its own.

### Branches

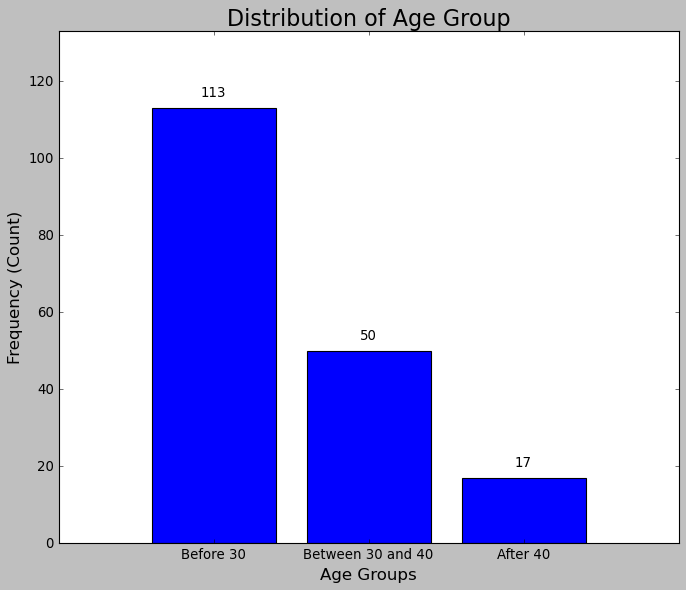


Looking at this chart, I can say that sales are just about evenly distributed. The chart on its own does not provides any strong starting point for speculative exploration. However, it does say that this particular dataset provides a fair point for comparison regionally. If there were lacking data entries from any branches, the branch with the short end would be as well represented as its counterparts.

### Age & Age Group

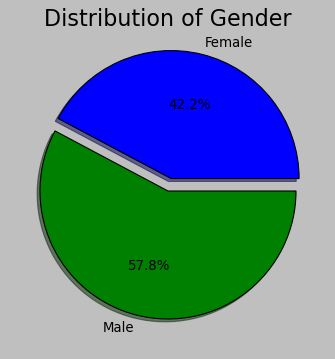


As I am observing the context of age, I find it would make sense to cover age and age group column together. Starting this section off, the boxplot of the age column has reflected the range of the age amongst the customer. We find that there are 3 instances of outlier listed in the age column, which seems to appear after the 45 mark. On its own it does not amount to much significance. Hence, I added the age group column earlier to assist visualizing the distributing within the age column.



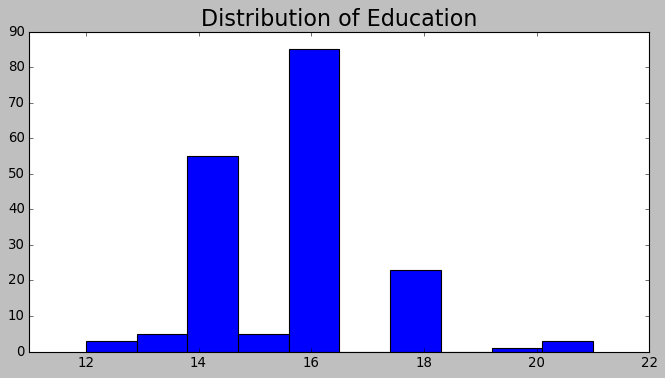
From the bar chart, we can see most of our customer are below their 40 years of age. The lack of customers who are above 40 years of age is expected and the popularity with younger people in contrast is remarkably high. After all, younger people would seem like the more sensible target audience to go after, they are younger and would most likely use equipment such as treadmills. On the other hand, the older folks might prefer the traditional route of going outdoor and hence, does not see the need to get a treadmill. However, if a pandemic were to take place, people are not able to go out and exercise, it will create a demand for indoor equipment such as treadmill. Which raise the possibility if CGF could develop a product and market it to older people, those who are heading into their “middle age”. Which I think is a potential market for CGF to explore and work towards.

### Gender



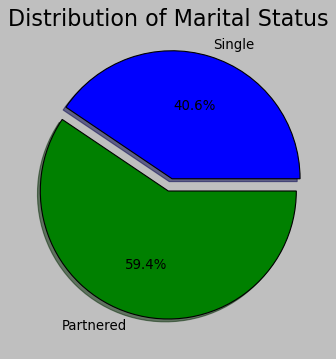
The gender distribution of our customer seems to be fairly reasonable, while it is not evenly distributed between the 2, it is also not dominated by a particular gender. This information does have its good and bad points. Starting with the bad implication, the products are catered to both genders and both sides are sizable contributors. Which sounds good but, it limits how CGF can market to their audience, CGF’s product will not be marketed as gender specialized product. It limits the potential to monopolize on a set of target audience and opportunities to explore other target audience. On the other hand, it is also good that they are working with generalize product for both genders. CGF has proven to be able to cater their product for both genders and focus other variables that are marketable.

### Education



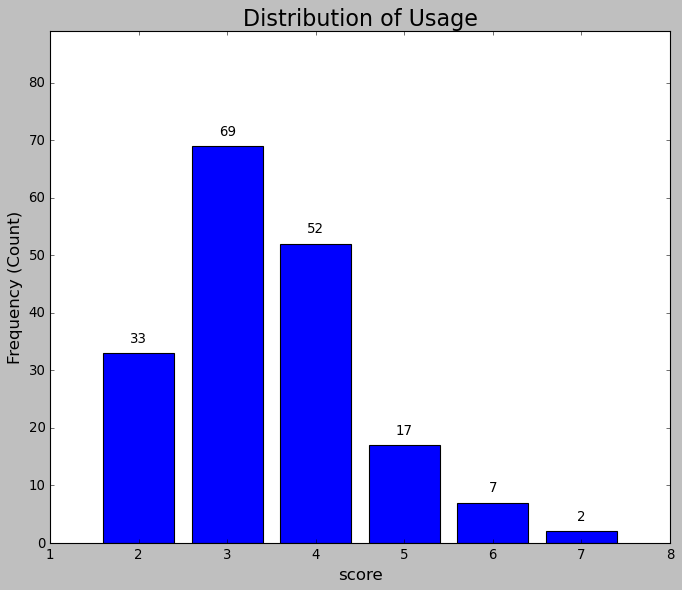
In this histogram, we can see the distribution of education by number of years amongst our customer. The majority of them have at least 14 years of education, in certain context this information might be useful. However, I do not think this column offers any meaningful insights within the context of this project. At the same time, there could be interesting correlation between education and other variables.

### Marital Status



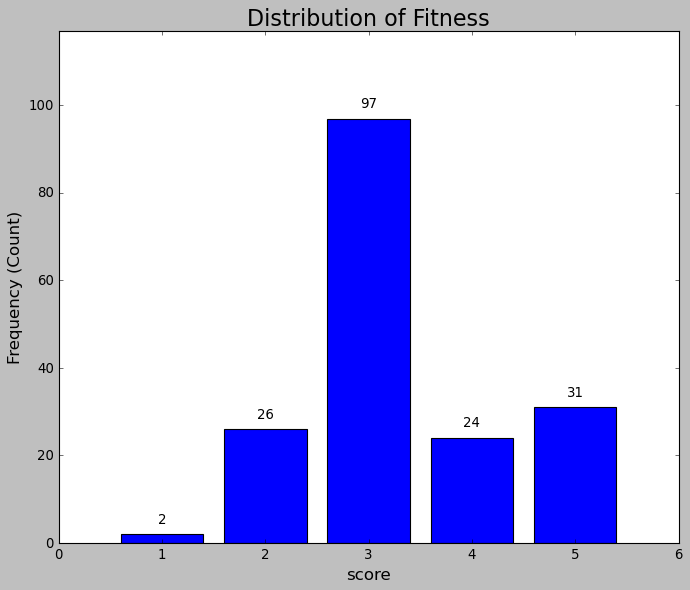
Marital status, it was not the first detail that comes to mind when I think about the customer demographic. However, I believe this data column could have some interaction with another data column. In one of the research questions listed above, I wanted to explore if marital status affects the rate of usage along with gender. Noting the importance of this data column, it shown that 3 in 5 customers are married.

### Usage



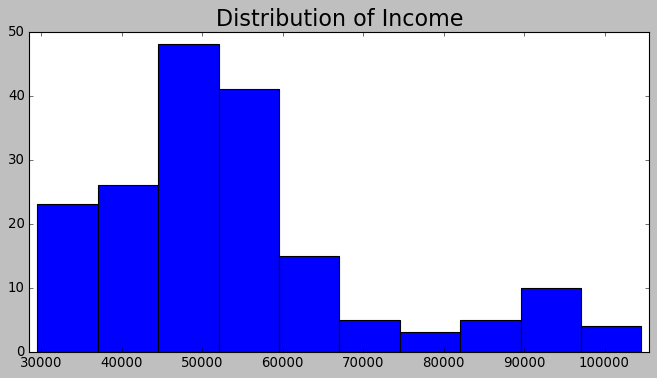
In this chart, we see the distribution of usage, it shows a representation of how often the customer will use the treadmill on average for each week. We can see that majority of the customer would use the treadmill at least 2 to 4 times weekly. Which does sound fairly reasonable for an average consumer, but there are also people who plans to use it daily. This data column seems to be very useful regarding about insights on user goals setting or usage habits.

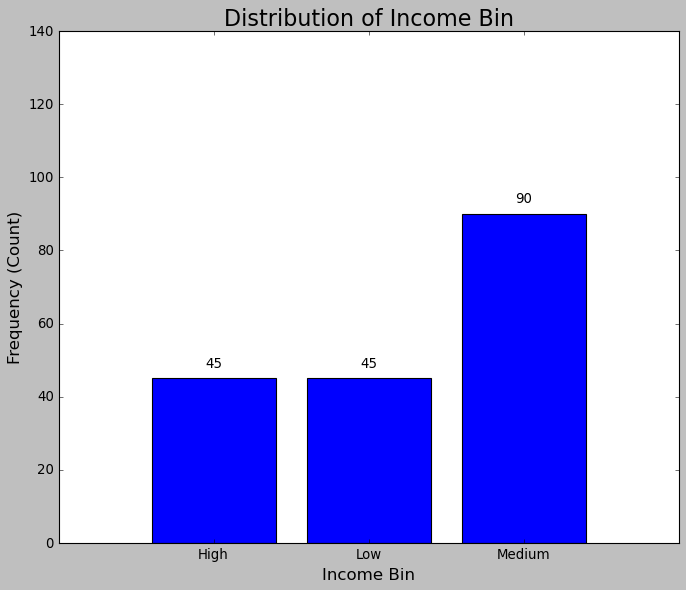
### Fitness



In this chart, we see that most of the customer find themselves to fairly average in terms of fitness. While only half of the customer rates themselves to be average, there are customer who themselves tilting toward either end of the spectrum. With this information, I can draw out some leads for exploration between the different variables.

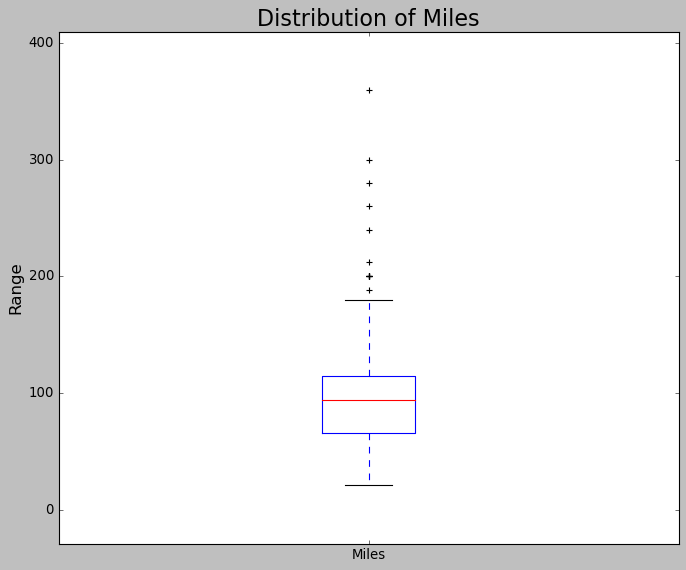
### Income & Income Bin

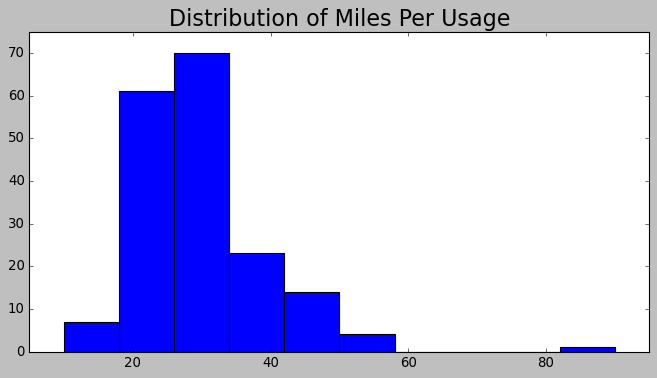




In the distribution of the income data column, the histogram seems to be skewed to the right. To look at the data column in a more normally distributed manner, Income Bin was introduced and we can see that income are categorise into class from low to medium to high. Both of which can be used for further exploration in multivariate analysis.

### Miles & MilesPerUsage(MPS)



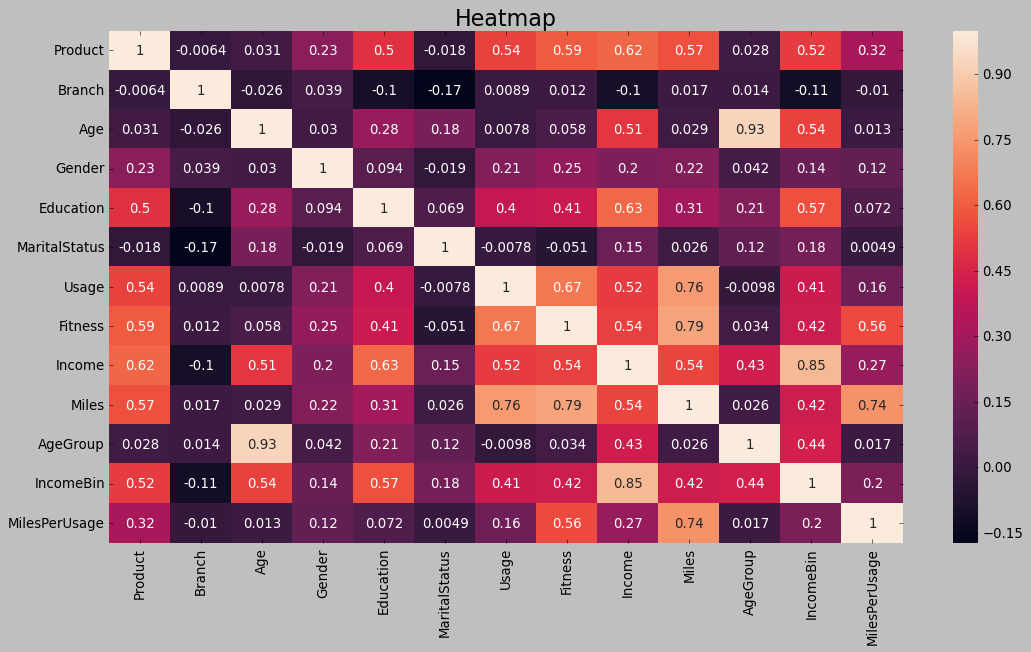


In this boxplot, there are some instances of outliers within the range of the Miles data column. From this plot alone, there are few things I have thought about with this data column. Such that, are people who set high miles would be someone with a goal that is really high, it makes me wonder how would they achieve it. I also thought if some those customers are overly ambitious with their goal setting. Hence, the Miles Per Usage column was added but, further analysis will answer those questions.

## Multivariate

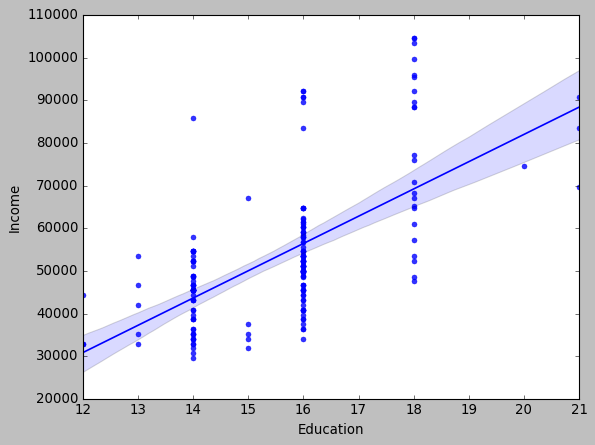
This multivariate analysis section is an attempt to further explore speculation I have on the data set. Its primary goal is to look into potential relationship between the data columns, at the same time providing a deeper understanding of the data columns that would be missed univariate analysis. While the univariate analysis only draws out insight based on individual variables, the multivariate analysis draw leads together and explore the hidden layers in the data set.

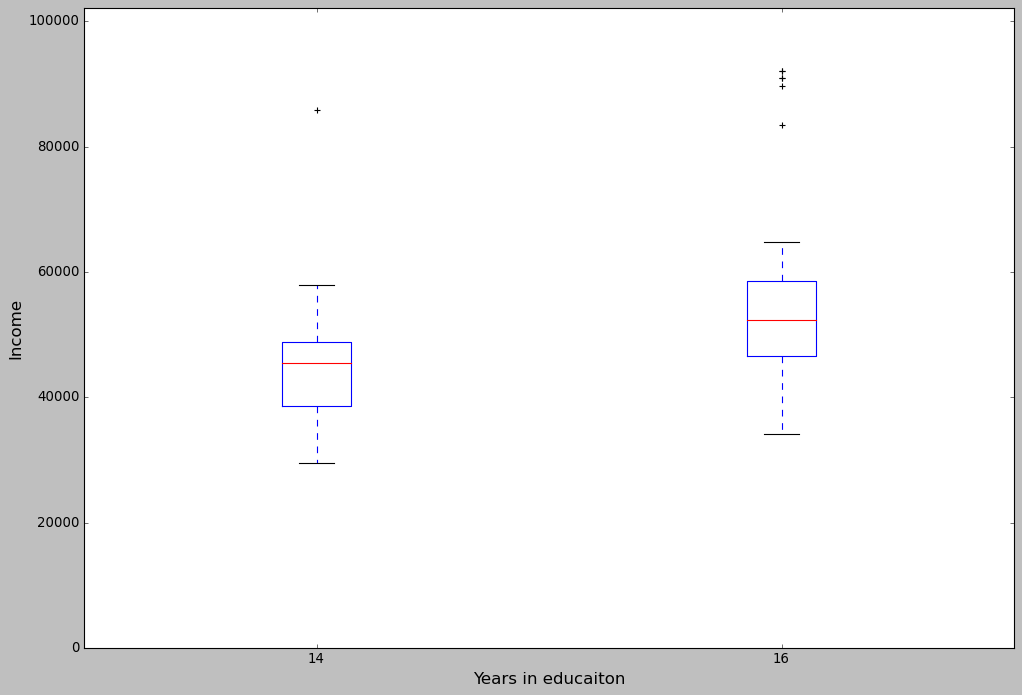
### Heatmap



Using the separate copy of the data frame, along with some encoding, this heatmap was made possible. The purpose of this heatmap is to be foundation for the multivariate analysis here. From this plot alone, we can identify and limit the correlation that I should look into. In this case, correlation with scores higher than 0.6 are what I am looking to explore in the remainder of this multivariate analysis. Adding on to the requirement for my chosen correlated targets, this section will only look into correlation that is not part of the research questions. As some of these correlations are what I will looking into on another set of visualization, which I would like to save for later. In the remainder of this section, I have chosen to look into the correlation of Education & Income, Age & Miles based on products and Miles & Income based on product. Education and Income is a relationship that I would like find out but not as part of my research question in this project. As for the other 2 combinations, I would like to use them to further understand the customers who are buying which product. On its own, I could classify it as a topic that is part of my research questions but, I am basing my research more on customers. While topics on customer can be vague, the research questions were designed to look into different areas about the customers.

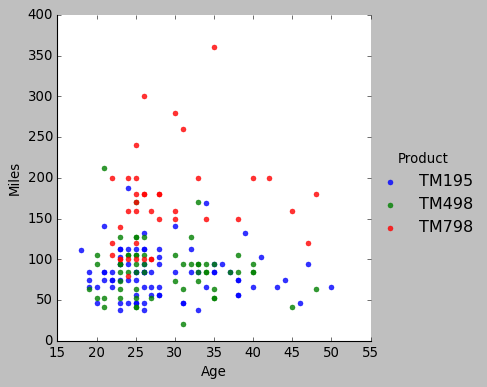
### Education & Income





Firstly, the education and income correlation. I initially thought that is already a connection between the 2 variables. However, I wanted to spread it out on a visualization to search for any other findings that I could not speculate. From this scatterplot, it was expected to see how education and income scales with each other. Such that, higher education leads to higher income. However, an interesting find was the lowest point in the range of income between people who have 14 or 16 years of education only had what seems to be slight difference. Whereas, the jump between 16 to 18 years seems more noticeable, so I created another boxplot to look closer between 14 and 16 year of education. In the new boxplot, the difference is much more noticeable as compared to what we can see in the scatterplot. It really goes to show the importance of context, where the gap looks different when I try to limit the dataset.

### Age & Miles (product)



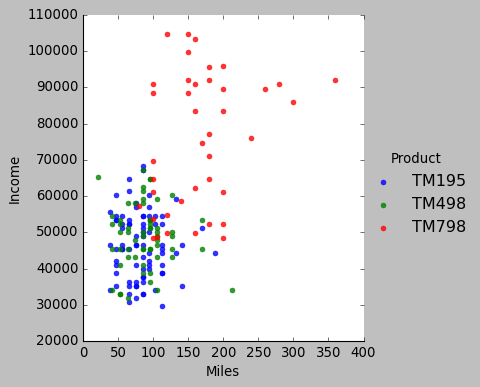
Next up, is the correlation between Age and Miles columns but, there is the additional variable, Products. This chart will enable some connection amongst the 3 columns to be drawn out.

Between Age and Miles itself, we can see that majority of the customer have the theirs goals of reaching certain miles limited to nothing beyond the 150 or 200 mark. This observation draws speculation on whether fitness is a determining factor in their goal setting. At the same time, it should be noted that the cluster of points between 20 and 30 years of age. That is because most of CGF’s customers are attributed as part of that age group.

Moving onto Age and Product, there actually does not seem to have much correlation between the 2 variables. We can see all 3 products are sold to various customers of different age. This evaluation may seem puzzling but the influence of having more younger customer affects the validity of fair evaluation. As the other age group, would not be as well represented, a fair evaluation relating to age and product is just not possible.

Lastly, there is the Miles and product correlation. In this case, the relationship is much more visible. Based on how the points of each product are spread out along the Miles, product “TM798” seems to be a product designed or advertised as a high-performance machine. The inference is supported by the fact only the demand of this product is clearer after the 150 miles mark. On the other hand, the other 2 products seem to be limited and hardly appear after the 150 miles threshold. Hence, suggesting “TM798” is a high performance or at least advertised to be while the other 2 seems to be commercialized products for the average consumer.

### Miles & Income (product)



In this scatterplot between Miles and Income along with type of product, there is a recurring relationship brought up again. Such that, I will focus on the newer correlation that has not been discussed about.

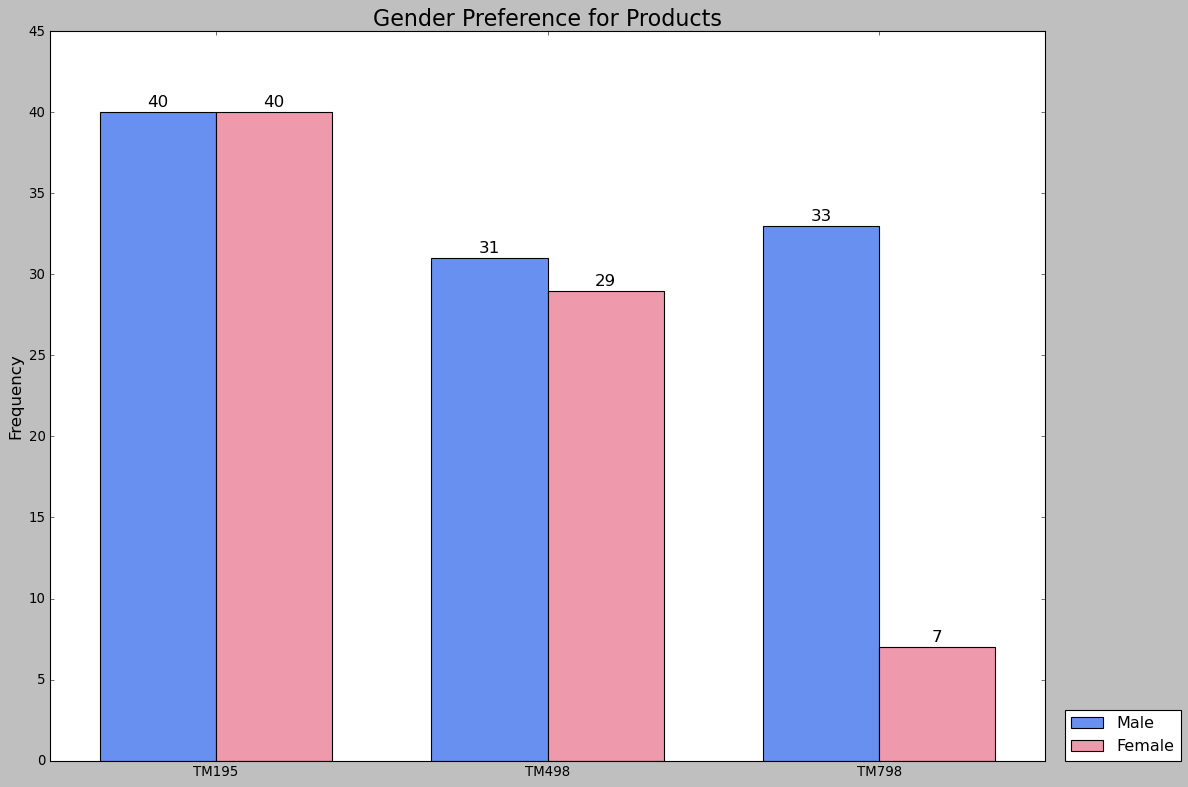
Between the primary variables of Miles and Income, I do not see a clear correlation or insight behind the relationship. Based on the position of each points, the scaling of both columns does not seem to be very clear. However, we can make speculation that products of higher end performance are sold at much high but reasonable price. Hence, the customer with higher income could afford it.

Which pushes the analysis into the relationship between income and products. There is clearly visible that people with a higher income input opted for product “TM798” over the other 2 products. There were some people with lower income but still pushed for “TM798” but, people around the same income group or class went for “TM195” or “TM498”. It is an indicator that the products are likely constructed and marketed to the average buyer. They are customers with fair budget, reason goal setting, ones who are part of the main consumer type. All in all, they are normal buyers but people who opted for “TM798” are ones that I speculated to one of a few types. They are likely to be exercise enthusiast or in general people with extra budget to spare. Even from this chart alone, there are still some information that is missing or lacking to find out why would customer buy a certain product based on budget alone. However, we establish the fact that “TM798” is CGF’s high performance product that is advertised to a specific target group such as enthusiast with money to spare.

## Viz for dashboard

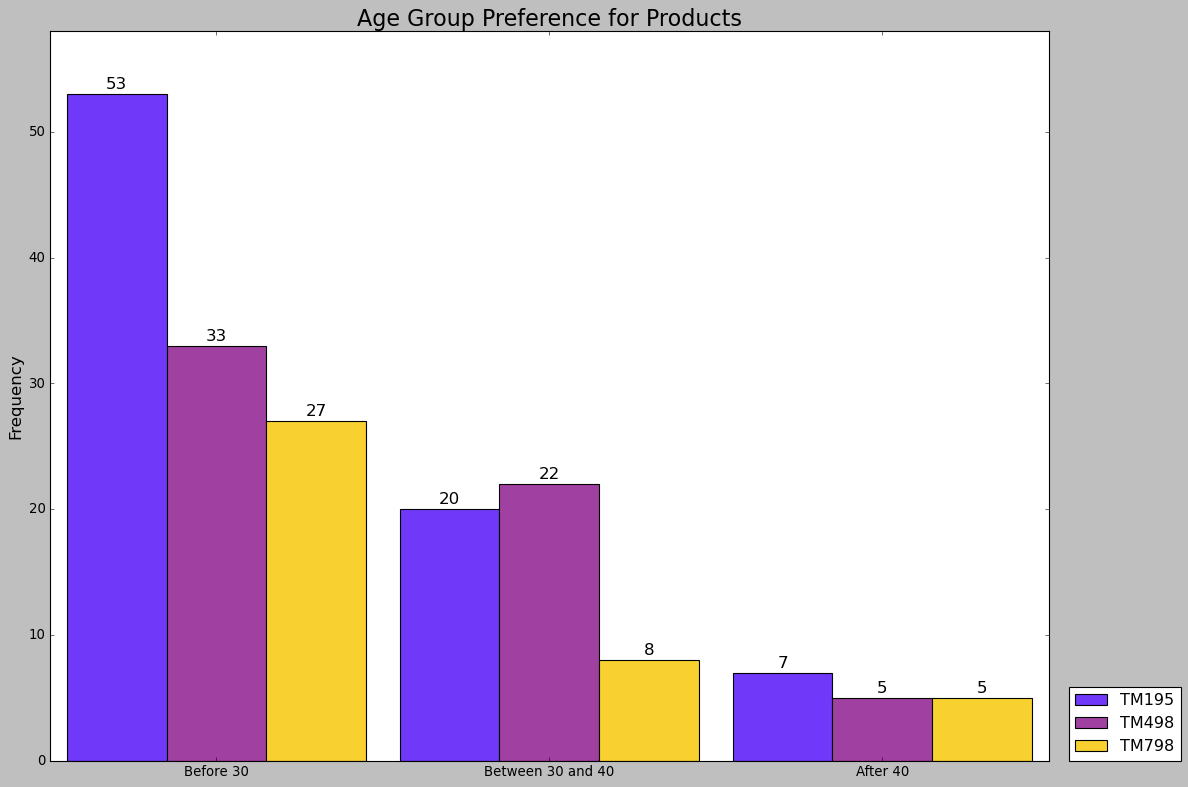
This particular set of visualizations will be my key visualizations for answering the research questions. All of which will be categorized into topics and placed into the corresponding dashboards. Each of the visualization are designed for a question that is made to assist the marketing team.

### Is there any gender preference to a particular product?



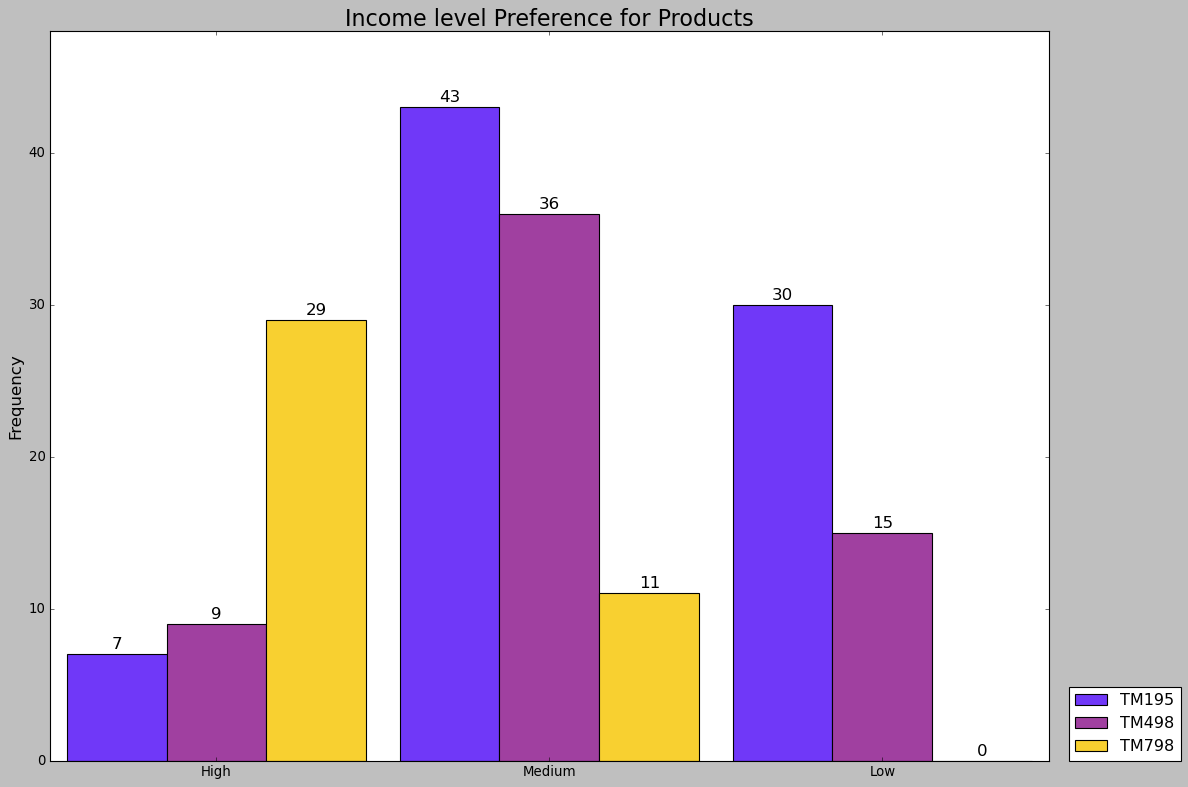
For our first visualization, the research question it stems from is “Is there any gender preference to a particular product?”. This is part of the demographical / preferential based question topic, which primarily focus on customer preference to a product based on the customer’s demographic. In this example, gender is used as the variable for research. This is because gender objectively speaking is a common classification point, it is commonly used to study difference between customer preferences. Despite the gender distribution being slightly disproportioned, this is an attempt to fairly evaluate the connection between gender and product. Look at a specific gender on its own, I will go with the male first. All 3 products seem to be selling evenly, there is no obvious signs of preferential influence. On other hand, there is a gap between female customers who bought product “TM798” and those we did not. An inference here is that female customers do not see the need to buy a higher performance machine. It is likely they are looking for products that are for the general consumers and getting “TM798” might be an overkill for them in general. Hence, we do not see as much female customers buying product “TM798”. Lastly, I will be comparing the difference between the 2 genders. Focusing on the stark difference between both genders for product “798”, it is likely that male customers felt the need to get a better product. Such that product “798” fits what they are looking for, which result in a demand by male customer as compared to female. All in all, this chart does not necessarily find out the gender having a direct influence on preferential needs among customer but it does explore their potential needs as a factor.

### Is there any age preference to a particular product?



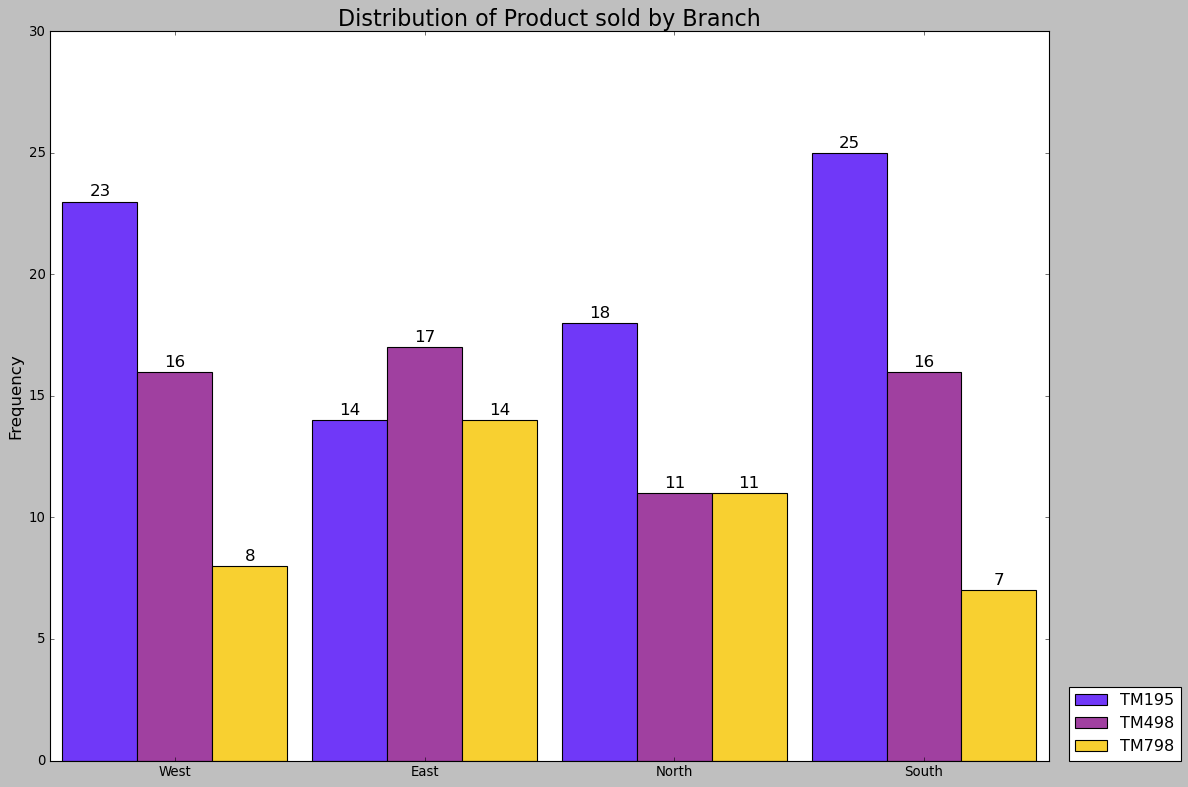
Moving on to the next visualization, this the second visualization out of the 3 on demographical / preferential based topic. In this chart, we are exploring if the age of the customer shows any form of preference but I will be using age group. This is to collectively show a better representation through classifying the age into groups. In this chart, we see the discrepancy of total product sold between the different age group. Which would be unfair to evaluate between the different age group directly. Hence, we will be looking at each age group individually and consolidate the findings. Starting with the group before 30 years of age, we see the contrast between each product and I believe it is due to the fact some products are more popular than each other. Other than that, people before 30 would prefer to buy product “TM195”. Moving onto the people who are in between 30 and 40, the most popular product sold was the “TM498”. A thing to note that in this age group, the sale of product “TM195” is trailing behind the leading product and product “TM798” just does not seem to be doing well. As for the people who are beyond 40, the sample is rather low, it is evenly distributed with marginal difference. However, product “TM195” is leading the sales in this age group. In a nutshell, our findings on product preference based on age group were product “TM195” leading in 2 age groups and product “TM498” is leading in the remaining age group. What I can infer from this finding is that product “TM798” target audience is a different set of customers that is bounded to the age as a variable. While age is not necessarily a strong variable when looking at preference in this case, we did also find that leading product is catered people of different age group. The marketing or development team could look to stimulate this feature in future or existing product to replicate a similar success in product “TM195”

### Does income affect customer preference to a particular product?



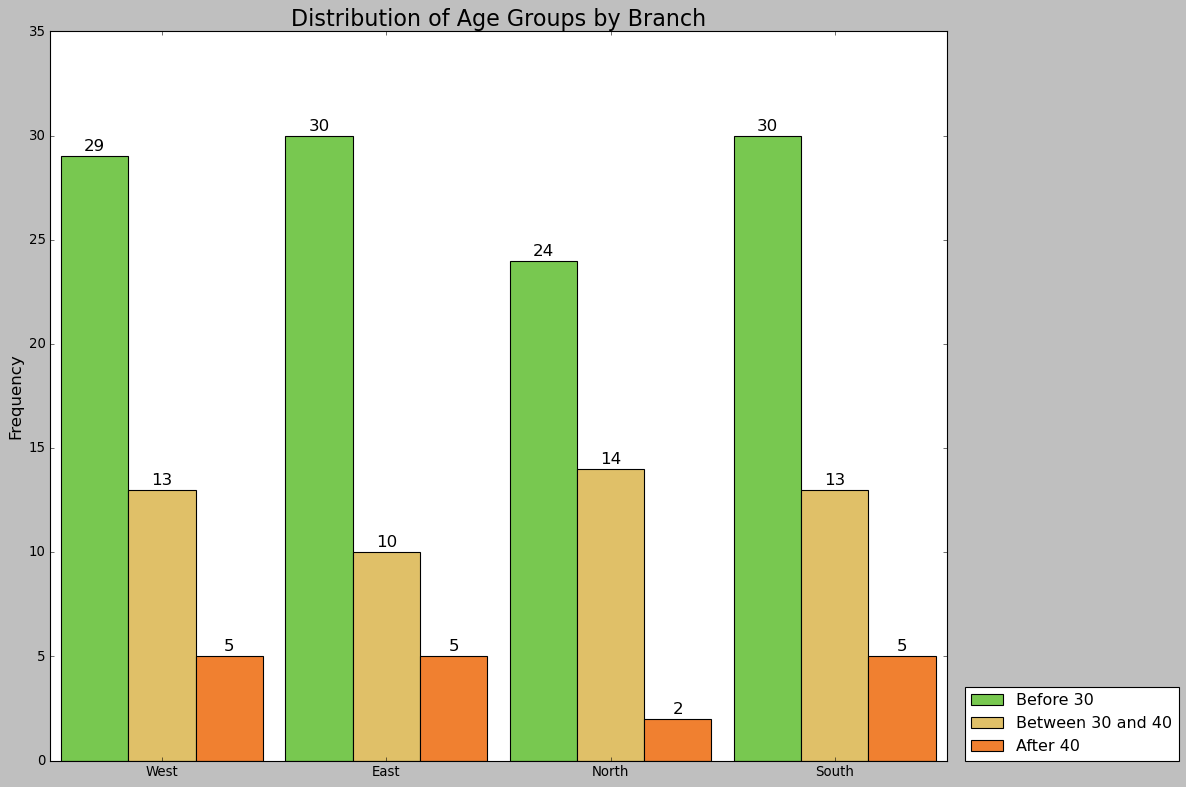
The last of the visualization on demographical and preferential topic, this chart will be looking into the income of the customer. Similar to the age preference, the income have been grouped into class, utilizing the Income Bin column, each of the data entries are assign to either be low, medium or high. In this chart, I will be look at each individual income class for my evaluation during the analysis. Firstly, in the high-income class, more than half of the customer opted to buy product “TM798”. Next, we observe a much higher amount of customer buying the other 2 products, whereas the people who buy “TM798” in the medium income class is lower. Lastly, in the low-income class, we observe lesser overall sales when compare to the medium income class. However, we also see no sales of product “TM798” in this income class. Drawing all these observations across the 3 income classes, I came up with a plausible reason for each observation. Starting from the gap of overall sales between the low and medium income class, it is due to how things are classified. The low-income class comprise of the customer below the first quartile. Whereas medium income class are made up of customer who are in the interquartile range. Hence, the disproportion in customer count. Next, the observation of seeing products being popular in the different income classes can be explained from this chart. In product “TM798”, we saw sales in high and medium classes and no sale in low class. To my current understanding, product “TM798” is CGF’s best product in terms of capabilities and the price of it would be the highest. We draw the connection that customer with a higher income has higher budget and can afford the product. As for the other 2 products, they saw more sales in the medium and lower classes, suggesting it is perhaps entry level or standard product. Between the 2 types of product, it is more likely for people who can afford the better product to get it. Hence, we see the high sales of product “TM798” in high income class. An interesting initiative that we can touch on from this chart alone is, “How CGF can get more customer in medium income class to buy product TM798?”. As most of the customer would be in the medium class, it is ideal to garner more sales in this market.

### What is the product distribution amongst the branches?



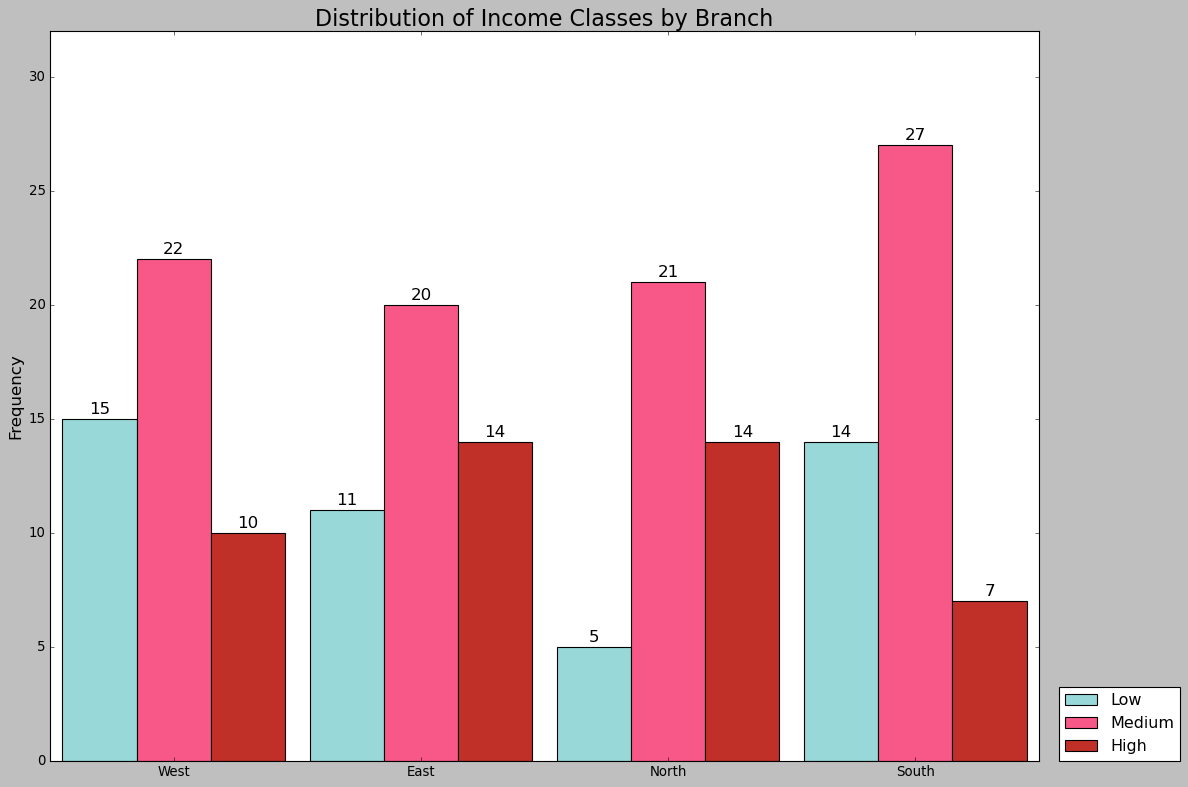
Moving onto the series of visualization for the geographical based question, the first question is exploring the distribution of each product regionally. Looking into the different regions, this chart shows how each product fare individually. Start with the west branch, the most popular product is “TM195” and figure of the distribution looks similar to the overall product distribution. Moving onto the east branch, Product “TM498” is the best-selling product while the other 2 had the same number of sales. In the north branch, we observe a similar pattern with the east branch but, the best-selling product is “TM195”. Lastly, the south branch is following the same distribution figure as the one in west branch. Through this preliminary observation, we know how each branch is faring in terms of sales for each product. However, this chart has another layer to it. I can look individually at each product while comparing their sales across the 4 branches. Starting with product “TM195”, it is the best-selling product across 3 different branches and its not far behind the leading product in the branch it failed attain the lead spot. As for product “TM498”, sales are about evenly distributed across the 4 branches but the sale in North is lagging behind a little. Lastly, for product “TM798”, the sales performance is split. The product selling in west and south is not as well as their counterparts in east and north. Before summarising all these observations, a thought to note is that demand for certain product will vary individually or overall, across the different region. Assuming that the demand of each product is negligible in this analysis. Product “TM195” is doing poorer in East and north branch when compared to the other 2 regions. Sales for product “TM498” in the north branch can be worked on. Lastly, product “798” sales in west and south branch is not doing too well.

### What is the age group distribution amongst the branches?



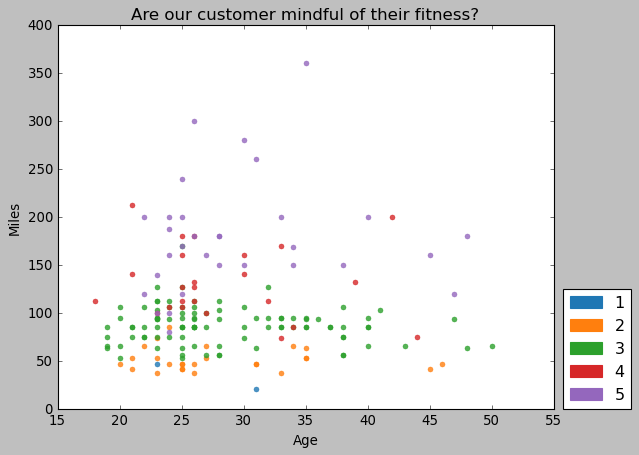
In this second question for the geographical based question, I am exploring the distribution of the customer age across the branches. Using the age group column, I will be looking at the number of customers in each branch who are in their 20’s. 30’s or beyond 40. Across the 4 branches, each of them has a similar distribution figure to one another. This is likely due to the overall distribution of customer age group but, I was expecting to see some difference between the branches in the first place. While the result of this finding is not exactly interesting or exciting, it is still meaningful in a sense that there is a clear visualization of the customer age group distribution.

### What is the income class distribution amongst the branches?



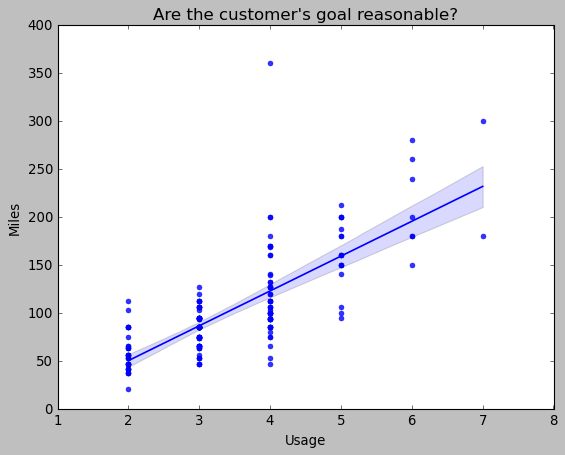
In the final installment for geographical based question, I am looking into the distribution of the customer’s income among the branches based on class I assigned. Predominantly, the medium income class has the most count across the 4 branches. Whereas, the difference between each branch comes down to some having more low income over high income count or vice versa. Such that west and south branch has more low-income count over the high-income class and the remaining 2 branches has it the other way round. From these observations, we can state that customer in north or east region are likely to have a higher income as compared to the other 2 branches. Which push for the possibility to have special advertising targeting people in different region based on the general income level of the branch.

### Is the customer mindful of their fitness in relation to any personal variables?



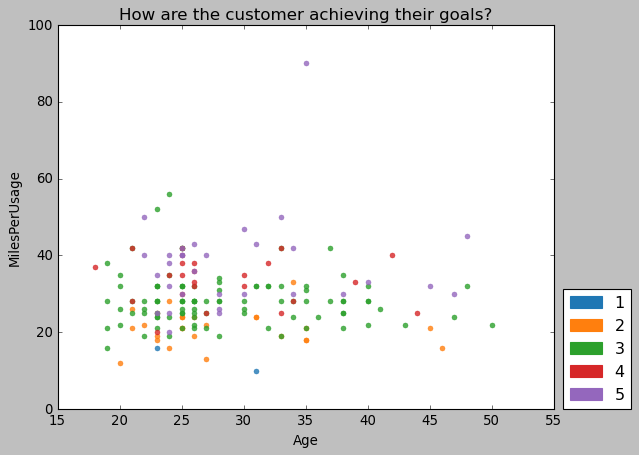
Moving on the final set of questions, the topic that the question was based around were on customer goals. In this set of question, it will feel slightly different from the other 2 topics as some questions were designed based on the preceding ones. Start with the first question, “Are the customers mindful of their fitness?”. This question was designed to draw the correlation between the Miles that the customer set and their age initially. However, I also saw an opportunity to find any correlation to the level of fitness they think they are at. Starting with the initial focus of the analysis between miles and age, there does not seem to have any correlation with each other. Both variables do not seem to scale with each other, people of different age each set their goals differently. Moving onto fitness with age, we see that between each tick of the age, the points that were plotted has at least 4 different levels of fitness set by the customer. Through this observation, we can say that age alone is not necessarily a direct variable that influence how customer see their fitness is at. To the point that I think the customer of various age is likely to view their fitness based on other variables. Hence, we move to look at the last connection with fitness in this chart, the miles. Immediately we can see a noticeable difference, only customers with level 4 or 5 fitness are setting mileage goals beyond 150 miles. Establishing the pattern where people with higher fitness would set higher goals. A notable find that I observe in these 3 interactions is that when customers are setting goals. Static variables such as age is unlikely to influence the goal setting whereas, dynamic variable that the customer can control will scale along.

### Is their target goal reasonable?



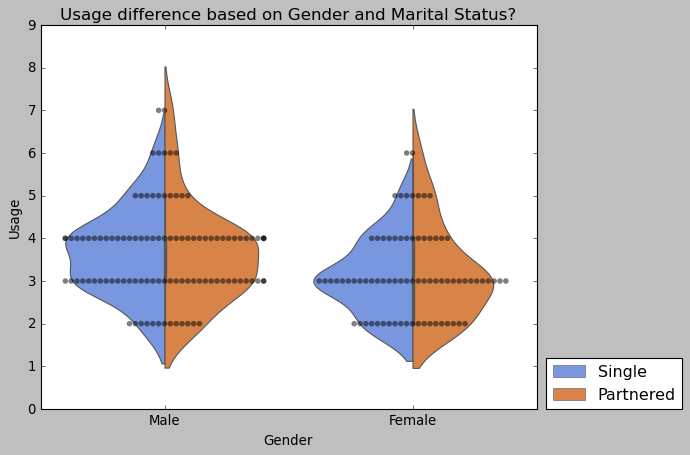
After noticing miles as form of goal setting and that dynamic variables which depend on the customer perspective are like to be correlated, I want to look into similar variables and analyse it. Which help me came up with the next question on whether the customer goals are reasonable such that it is not too ambitious. Noting how Usage is a similar variable to miles, I used both of these variables to make a regression plot. Looking at this visualization, we see that customer with higher goals set also plan to use the treadmill more often each week. Such customer who plans to achieve certain miles every week would also manage the number of usages per week. Which makes their goals manageable to achieve and from this chart, I can draw another speculation. It is possible that when customers are setting their miles, they could be calculating it based on how many miles they plan to run with each usage. It is highly possible that some customers already planned out their usage on average before they set their miles as well. However, this point will only remain as speculation as I lack information to corroborate this point. In a nutshell, we did get an answer for this question where the customer goals are manageable but I am left with more questions that may remain unanswered.

### How do they plan on achieving the target?



After exploring how manageable the customer goals are, the possibility of miles being set after usage was fresh on my mind. Which means there is likely another half to the correlation between age and miles. This is because if usage were to play a role in determining their goals or how they evaluate their fitness, the first chart of this customer goal topic would not be entirely valid. I will need to explore the other half and do a similar chart but with the influence from the planned usage. To do that, I set up another column where I take miles divided by usage to determine the miles planned per usage. Running though with the same variables of age and fitness, this visualization is made similar to its preceding version without usage in the initial analysis. While range is now lower, majority of the points plotted have been lowered in comparison to the old chart. There is the exception of that outlier of a customer who thinks he or she is very fit, set a high target for miles but plans to accomplish within a limited number of usages. Well for this case, it is negligible as all the points were lowered as expected in this version. In this chart, we can see the effect when we take into account that customers are managing their goals at a rate it is suitable or at least normal.

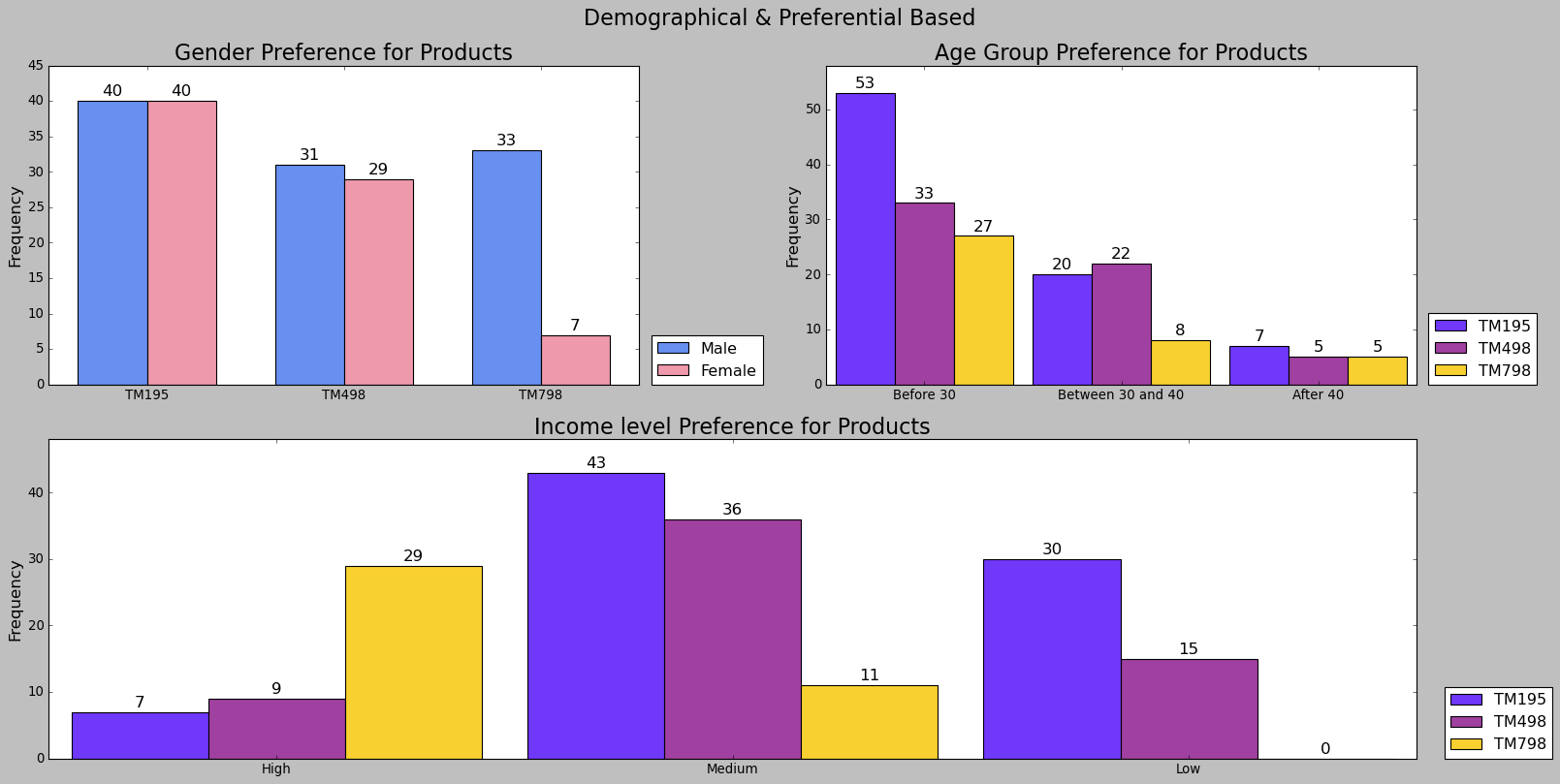
### Is there any usage difference between genders; marital status a factor?

****

Moving away from the fitness focus analysis, the usage seems to have an underlying find that can explored on. The initial question was to find a link between gender of the customer and their intended usage count. It was from there I felt that I could explore into marital status, such that “Does marital status affect the usage rate?”. To answer the question, I determine the use of a swarm plot could be useful here and I found out that violin also can add in hue. Which allowed me to add in marital status as an additional variable into this chart for this analysis. Start with gender, it is noticeable that more males are intending to use the treadmill at least 4 times a week when compared to female customers. Moving on, we see that figure of the violin plot where 1 half is about single males and the other half is partnered are similar in shapes. The only difference I can see is that the plot on partnered customer is slightly bigger and this observation is applied on the female customer as well. Which leads me to believe that I have the answer for both the question. Firstly, talking about the gender and usage thing, there is a usage difference between genders. However, I think the fact that there is more male customer over female might play a part here. As for the marital status observations, the size difference could be due to the discrepancy between count for both marital statuses. Objectively speaking, I do not see marital status affecting usage due to the similarity in shape of the violin plots. Although, there are minor difference at each of the curvature.

# Dashboard

***Demographical & Preferential based question:***

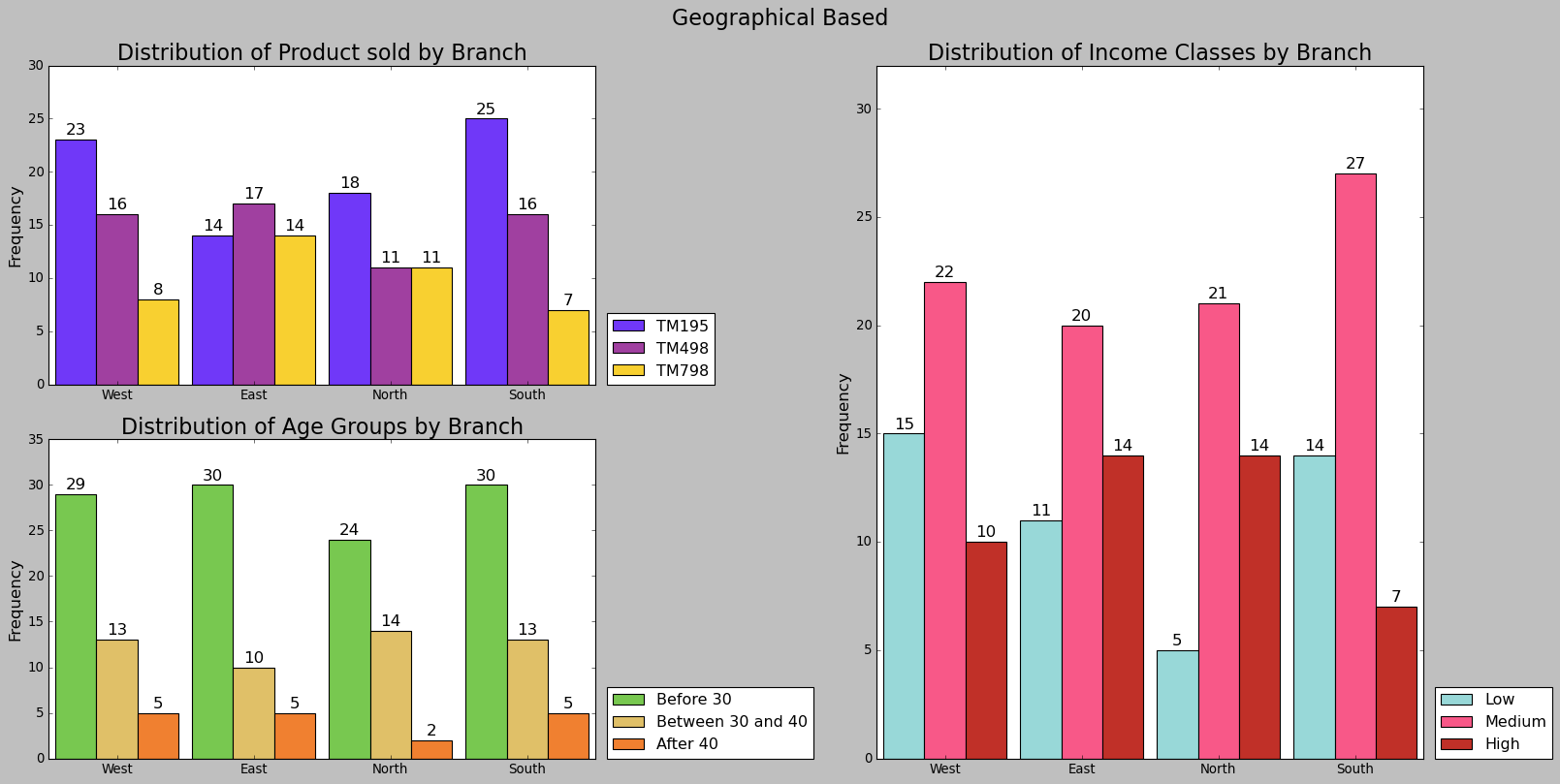
****

**Top left:** **Is there any gender preference to a particular product?**

**Top right:** **Is there any age preference to a particular product?**

**Bottom row:** **Does income affect customer preference to a particular product?**

***Geographical based question:***

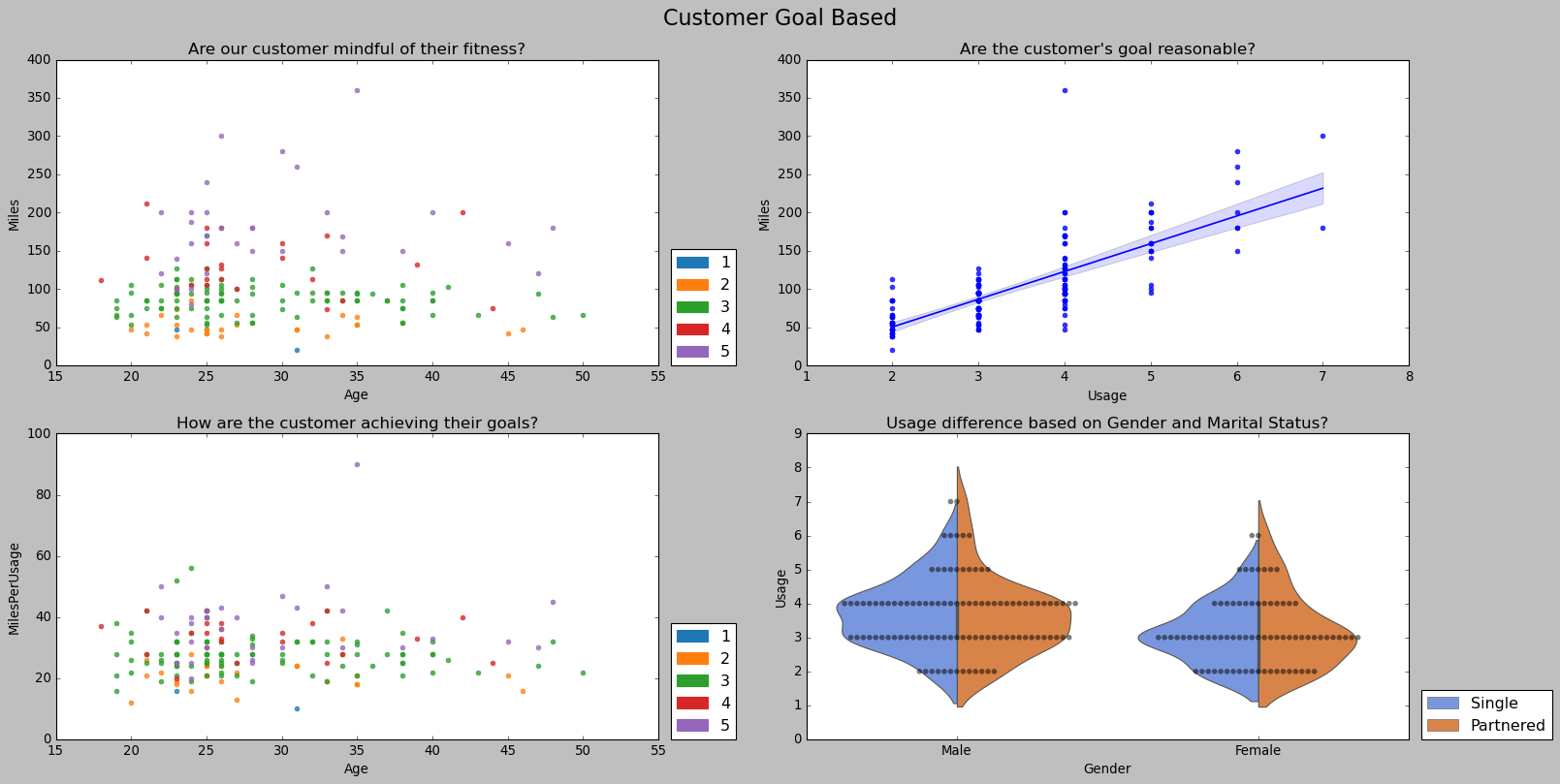
******

**Top left: What is the product distribution amongst the branches?**

**Bottom left:** **What is the age group distribution amongst the branches?**

**Right column:** **What is the income class distribution amongst the branches?**

***Customer based question:***

****

**Top left:** **Is the customer mindful of their fitness in relation to any personal variables?**

**Top right: Is their target goal reasonable?**

**Bottom left:** **How do they plan on achieving the target?**

**Bottom right: Is there any usage difference between genders; marital status a factor?**