# 



**School of InfoComm Technology**



**Data Visualisation**

Diploma in Financial Informatics

April 2021 Semester

**ASSIGNMENT 2**

**(Individual Assignment)**

**Submission Deadline:**

**15th August 2021, 23:59PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **P04** |
| **Student Name** | **:** | Keane Dominic Travasso |
| **Student Number** | **:** | S10202630(G) |

**Penalty for late submission:**

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

**NO** submission will be accepted after 22th August 2021, 10AM.

**Introduction**

I am part of the Market Research Team for Fitness Trainer Pte Ltd, which is a retail business specializing in the sales of stationary bikes. I have been tasked to investigate whether there are differences across the models in respect to the customer characteristics based on the data collected across the past 3 months. I will be using Jupyter Notebook and Python to create my visualizations and dashboards.

**1. Project Objectives**

For this project, I will be coming up with 3 dashboards. The first being, the Customer Demographic Dashboard, followed by the Customer Fitness Dashboard and lastly, the Data Relationship Dashboard. These dashboards will incorporate multiple visualizations that will analyze the Customer Characteristics based on the bicycle products purchased.

**The questions I will be creating visualizations on are:**

**Customer Demographic Dashboard**

* Which products are the most popular amongst customers?
* What is the distribution of the Customers’ Marital Status for all the products?
* Which products do Customers with a higher income tend to get?
* What is the distribution of the Customer’s Gender for all the products?
* What is the distribution of the Customer’s Age for all the products?
* What is the distribution of the products sold at the different branches?
* Which branch do Customers with a higher income tend to shop at?

**Customer Fitness Dashboard**

* Which product gets used the most in a week by the Customers?
* How fit are the customers who bought each product?
* What is the mean miles that customers said they would cover for all the products?
* What is the distribution of the usage for each product?
* What is the distribution of the Fitness level of Customers for each product?
* What is the distribution of the miles covered by the Customers for each product?
* How does the miles covered by each product compare to each other?

**Data Relationship Dashboard**

* What is the correlation between all the Customer Characteristic Data in the dataset?
* What is the correlation between the Fitness Level of the Customers and the Miles covered?
* What is the correlation between the number of Usage days and the Miles covered?
* What is the correlation between the number of years of Education received by the customers and the income they make?

**Why should Fitness Trainer Pte Ltd. care about the Exploratory Questions I have identified?**

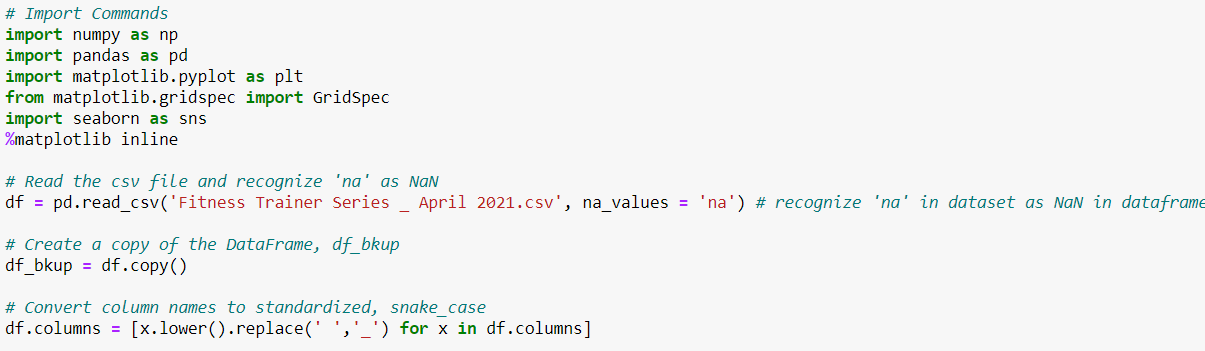
I believe that Fitness Trainer Pte Ltd should care about the Exploratory Questions that I have identified as the questions tell the company more about their customer’s details and characteristics as well as how the characteristics differ across the different bicycle models. Through this, Fitness Trainer Pte Ltd will know how to appeal to the customers based on the characteristics identified, as well as what changes they should make when marketing their bicycles in order to get more sales.

**2. Data Preparation**

**Introduction to Data Cleaning**

The dataset provided was a little bit messy due to certain values being different from the other values in the respective column. Alongside this, there were also quite a bit of ‘na’ values that had to be cleaned in order to use the dataset to create visualizations. I expect to do quite a bit of cleaning in order to ensure that the dataset can be used to come up with exploratory questions and to create visualizations based on them.

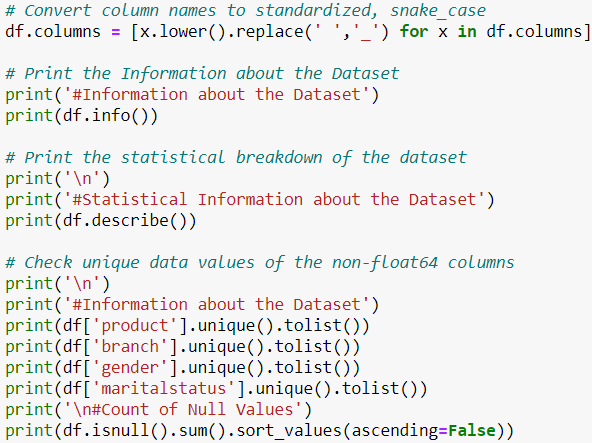
To Prepare the data to be used in order to create the visualizations, I first had to import the commands required to work on the data, as well as importing the ‘Fitness Trainer Series \_ April 2021.csv’ file in order to utilize the data in my visualization later on. I also created a backup of the dataset just in case something happens to it. From there, I converted the column headers of the dataset to snake-case so that it was more standardized.



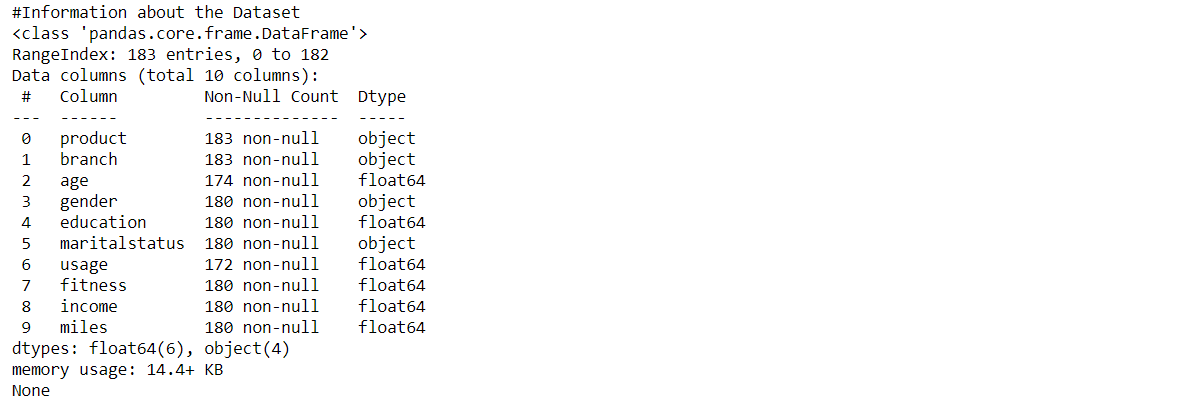
After doing that, I printed out the information about the dataset using, df.info() in order to learn what all the data columns were, how many null values were present in the dataset, as well as the data type of all the data in the columns. Following that, I printed out a description on the dataset using, df.describe() which provided me with the statistical data needed when I attempt to fill in the data that is null.

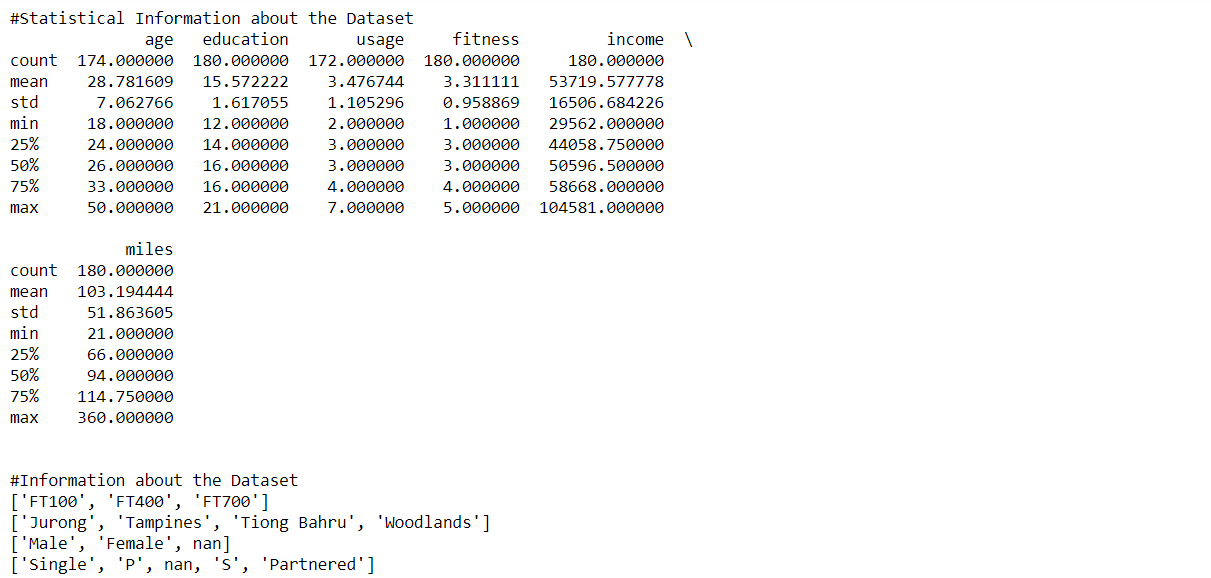
I then used df[x'].unique().tolist() to print out the unique data for the columns, Product, Branch, Gender and Marital Status by replacing x with the column name. This allows me to check all the unique values for each column to see if there are any unwanted values that need to be replaced.

Lastly, I printed out the number of null values per column using, df.isnull().sum().sort\_values(ascending=False) so that I would have a cleared count of the number of null values.



The result of these commands was,





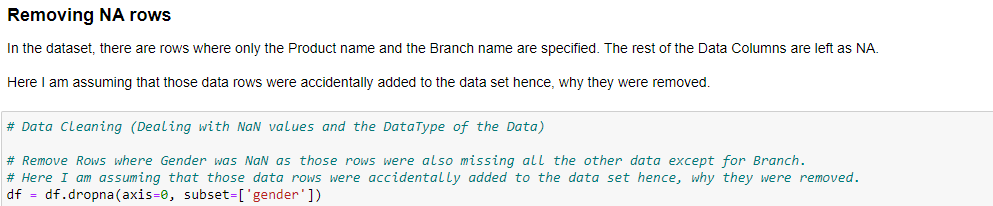


**Removing NA Row**

In order to ensure that the data can be used effectively to create the visualizations, I proceeded to use python commands to clean the data. The first thing I did was to remove the rows of data from the dataset where the Gender was ‘na’. The reason I did this was because those rows of data were missing all values except for the Product and Branch.

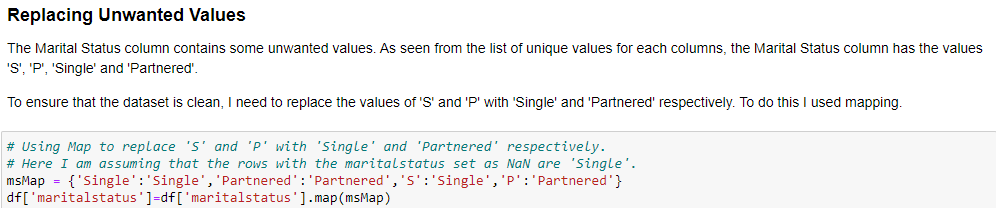


As such, I made an assumption that those rows were added into the dataset accidentally and hence should be removed.



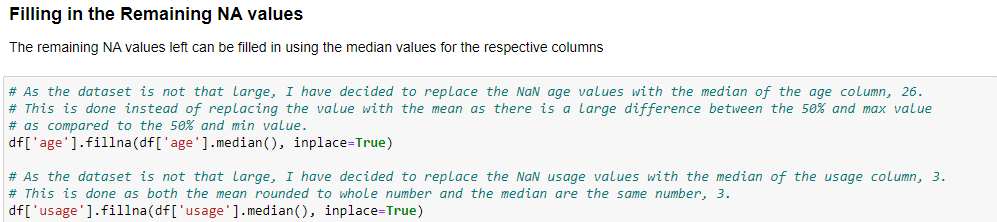
**Replacing Unwanted Values**

After removing those rows, I decided to clean the values for Marital Status. Under this column, there were 5 different values, ‘Single’, ‘P’, nan, ‘S’ and ‘Partnered’. The rows that had the Marital Status set to nan were removed earlier. Here, I assumed that ‘P’ was the same as ‘Partnered’ while ‘S’ was the same as ‘Single’. I then created a map called msMap and had it initialized as {'Single':'Single','Partnered':'Partnered','S':'Single','P':'Partnered'}. From there, I mapped the Marital Status column values using the map and the command df['maritalstatus']=df['maritalstatus'].map(msMap). This got rid of the extra values by changing them to be more consistent with the dataset.



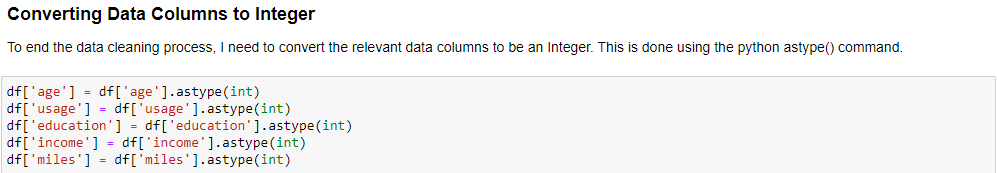
### Filling in the Remaining NA values

I then proceeded to clean the rest of the nan values. As the dataset provided was small, I decided to use the fillna() command in order to keep the rows rather than removing them completely. Using the commands, df['age'].fillna(df['age'].median(), inplace=True) and df['usage'].fillna(df['usage'].median(), inplace=True), I replaced the nan values in the age column and usage column with the respective median values for each column. I did this as after looking at the description of the dataset earlier, I saw that the median values were all whole numbers as compared to the mean values. After doing some research, I also found out that replacing nan values with the median value is usually recommended.



### Converting Data Columns to Integer

After the nan values had been dealt with, I decided to convert the respective columns to the integer type as the specifications mentioned that the values have to be in the integer format. To do this, I used the command df['x'] = df['x'].astype(int) and replaced the x with the column names for age, usage, education, income and miles.



### Final Check

I then printed the command df.isnull().sum().sort\_values(ascending=False) to check for any nan values in the columns, which showed 0. In this cell, I also defined 2 color palettes which will be used later on in the visualizations, productColors and fitnessColors.



**3. Visualizations**

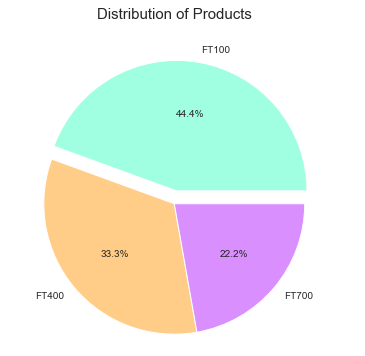
**Analyzing the Dataset**

In order to perform univariate & multivariate analysis on the dataset, I first had to take a look at the different data columns available for me to use when creating the visualizations. After looking at the data columns and thinking about the dashboards I wanted to create from the data, I split the data columns into 2 different categories.

The first being the Customer Demographic Data which consisted of the columns, age, gender, education, marital status and income. The second category I identified was Customer Fitness Data which consisted of the columns, usage, fitness and miles. From there, I focused on certain data columns when coming up with potential visualizations to answer the exploratory questions identified earlier.

**1. Which products are the most popular amongst customers?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

When analyzing the dataset, I wanted there to be a question that explores the spread of the products among the customers. Hence, I assigned the command, df.groupby('product')['branch'].count() to a variable called productDist. Through this command, productDist holds the count for each product assigned to the respective product names. I then declared another variable called products and assigned it the command, productDist.index. This makes the products variable hold the names for all of the products. I then plotted the pie chart using productDist and products, after sorting the values of productDist in ascending order. In the Pie Chart, I made it explode for FT100 so that I could bring the users attention to the largest slice in the pie chart.

**Reason for Picking Visualization:**

The reason why I chose a pie chart with percentage labels to answer this question was because I felt that pie charts best showed the proportion of the count of the different products. The pie chart is an easy chart to understand as the different ‘pies’ make up 1 whole. I also felt that including the percentages in the center of each pie, would allow the user to understand the proportion better as compared to a pie chart without the percentage labels. Another reason why I chose the pie chart was because there were only 3 products to visualize and thus, the pie chart produced would not look too messy and complicated.

The user can look at the Pie Chart and immediately know which product is the most popular by looking at the different slices and the percentage of each slice.

**Core Findings & Insights Found:**

From the analysis and visualization, some core findings identified are,

* The FT100 bicycle model is the most popular model with 44.44% followed by the FT400 at 33.33% and the FT700 at 22.22%.
* The FT100 is 2 times more popular than the FT700.

**Alternatives Considered:**

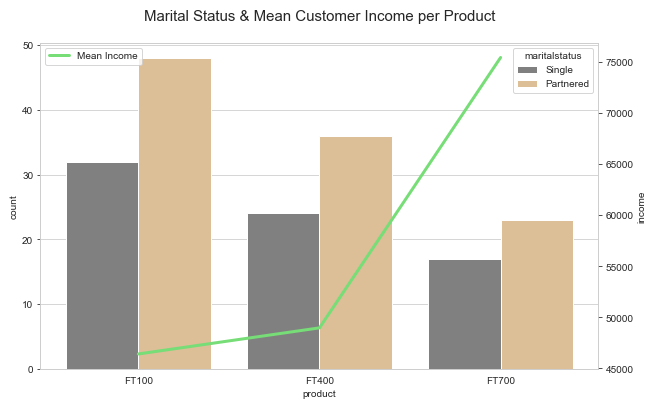
Bar Chart:

An alternative I considered while thinking of charts to answer this exploratory question was a bar chart. In the end, I decided against it as I felt like bar charts did not convey the proportion as a whole as well as a Pie Chart does.

**2. What is the distribution of the Customers’ Marital Status for all the products?**

**3. Which products do Customers with a higher income tend to get?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

I then created a dual axis chart, which contains a bar graph and line chart which overlap due to the x-axis being the product name. I did this as I wanted to explore how the customer’s marital status differed across the 3 products. To make the bar chart, I used a count plot with the x-axis as ‘product’. Due to the nature of the count plot, the y-axis would be the count of the number of customers. In order to show the viewer the difference in marital status, I set the hue of the chart to be the ‘marital status’. Next, in order to display how the mean income of the customer’s differs across the 3 products, I used a line chart. The line chart has its x-axis as ‘product’ while its y-axis is the ‘income’. I did not have to manually find the mean of the income across the products as python does it for me.

**Reason for Picking Visualization:**

I chose to use a dual axis chart as my visualization as it can effectively answer 2 exploratory questions at once, which enables me to display more data to the viewer. The bar chart portion, relays the proportion of Single customers to Partnered customers across the 3 products well while the line chart relays how the mean income of the customer changes across the 3 products.

The viewer can tell the proportion of the Single customers to Partnered customers, based on the proportion difference between the gold bar and the gray bar. The user can also tell which product customers with higher incomes tend to buy by looking at which point in the green line chart is the highest.

**Core Findings & Insights Found:**

From the analysis and visualization, some core findings identified are,

* More Customers are Partnered as compared to Single across all products.
* Customers with a higher income tend to purchase the FT700 with a mean income of around 75000, followed by the FT400 at around 49000 and lastly the FT100 at around 46000.

**Alternatives Considered:**

Pie Chart:

I decided against using a pie chart for this question as I realized that it would be hard to show how the proportion differs across the products this way.

Stacked Bar Chart:

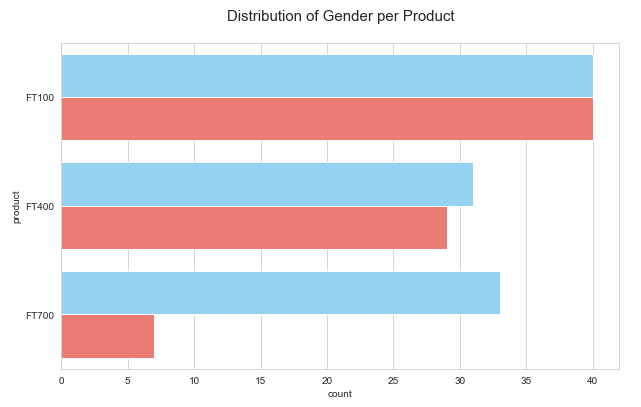
I did not end up using the stacked bar chart to display the marital status of customers across the products as I felt that using the stacked bar chart would make comparing the proportions much more difficult.

Scatter Plot & Box Plot:

I considered using a scatter plot and a box plot in order to display the distribution. In the end, I decided against using these plots as I wanted to create a dual axis chart and hence needed the visualization to be a line chart.

**4. What is the distribution of the Customer’s Gender for all the products?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

To visualize the distribution of the gender of the customer for each product, I decided to use a horizontal grouped bar chart. To do this, I used a count plot with the y-axis as ‘product’. Due to the nature of count plots, the x-axis will be the count. I then set the hue of the bar chart to be ‘gender’.

**Reason for Picking Visualization:**

I decided to use a grouped bar chart in order to display the proportion of the Male Customers to the Female Customers for all of the products as I felt like this was the best graph for this situation. I say this because the chart effectively plots the count of customers for each product, split into the gender, which is exactly what I want.

The viewer can easily tell the proportion between the male and female customers for each product from this graph by comparing the blue colored bar to the red colored one.

**Core Findings & Insights Found:**

From the analysis and visualization, some core findings identified are,

* Most of the customers are male across all products.
* The FT100 product has a balanced proportion of males to females at around 40 per gender.
* FT700 & FT400 have more male users compared to female users, with FT700 having a larger difference of around 27 customers.

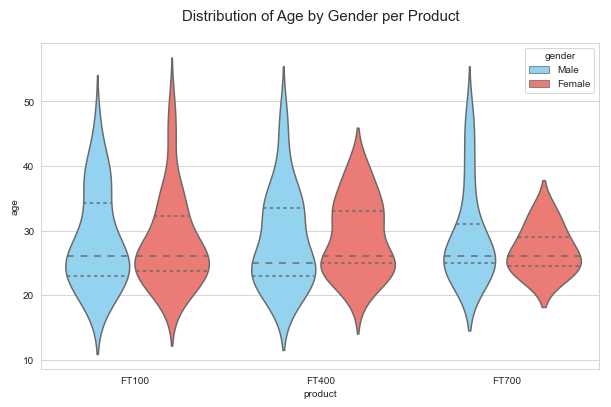
**Alternatives Considered:**

None:

I feel like the grouped bar graph created using the count plot and hue set to gender is the perfect graph for this situation.

**5. What is the distribution of the Customer’s Age for all the products?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

In order to visualize the distribution in the customer’s age for each product, I decided to use a violin chart, with the x-axis set to ‘product’ and the y-axis set to ‘age’. I have also set the hue to ‘gender’. This allows more analysis to be done as there is an extra variable included in the visualization.

**Reason for Picking Visualization:**

I decided to use a violin chart as the visualization as I feel that it shows the spread of data the best all while looking really nice. The violin chart also varies in width based on the number of data points, which makes it very useful when analyzing the data to find some insights. The violin chart also has the inner lines set to show the median and interquartile range. This makes the chart more useful as it allows viewers to tell which part of the chart the data points are concentrated at.

**Core Findings & Insights Found:**

From the analysis and visualization, some core findings identified are,

* The spread of the Customer’s age for males remains consistent throughout all 3 products.
* The FT100 has a larger spread of age for females while the FT700 has the lowest spread of age for females.
* Most customers tend to be between the ages of 20 and 30.

**Alternatives Considered:**

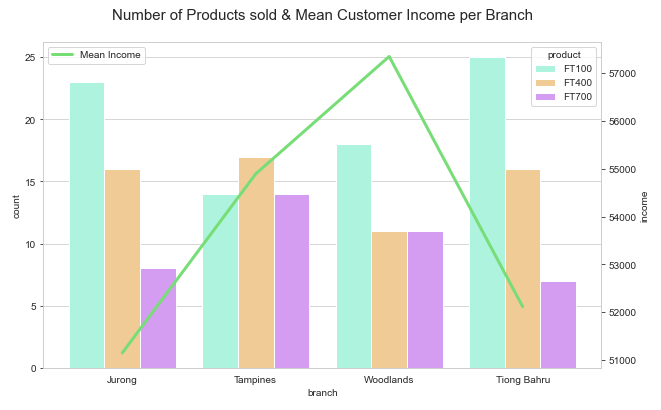
Box Plot:

In the end, I decided against using a box plot as I felt like a box plot did not look as appealing as the violin chart. The box plot would also not represent the concentration of data points well as a box plot’s width remains consistent throughout. Hence, the violin plot has an advantage over the box plot in the sense that it’s width can vary based on the number of data points at a certain position.

**6. What is the distribution of the products sold at the different branches?**

**7. Which branch do Customers with a higher income tend to shop at?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

In order to show the distribution of products by the branch, I have chosen to use a dual axis chart. The bar chart portion of the dual axis chart has its x-axis set to ‘branch’ while it’s y-axis is set to be the count. This represents the number of each product sold per branch. The line chart portion of the dual axis chart has its x-axis set to ‘branch’ while it’s y-axis is set to ‘income’. Thanks to python, the income is automatically made into the mean for the branch. This shows the difference in the mean income of the customers depending on the branch.

**Reason for Picking Visualization:**

The reason why I decided to use a dual axis chart was so that it could answer 2 exploratory questions while only taking up 1 chart space. I decided that since I already had a chart that showed the distribution in the products bought at each branch, I should have one for the mean income of the customers based on the branch at which they bought their products.

The reason why I chose the bar graph to display the distribution of the products bought from the different branches was because I felt that bar charts allowed the data to be easily comparable while looking clean and simple to understand.

**Core Findings & Insights Found:**

From the analysis and visualization, some core findings identified are,

* All branches except for Tampines have the FT100 as their best-selling product.
* The Woodlands branch has the highest mean customer income of around 57000 while the Jurong branch has the lowest mean customer income of around 51000.
* Despite Woodlands having the highest mean customer income, the sales of the FT700 are balanced with the sales of the FT400, despite an earlier graph showing that the FT700 is usually bought by customers with a high mean income.

**Alternatives Considered:**

Pie Chart:

I decided against using a pie chart for this question as I realized that it would be hard to show how the proportion differs across the products this way.

Stacked Bar Chart:

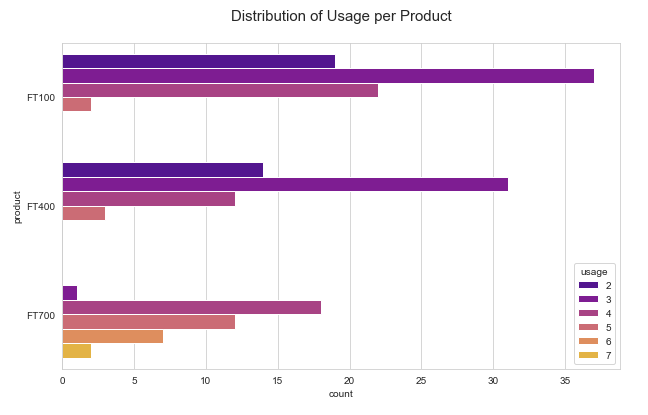
I did not end up using the stacked bar chart to display the product count across the branches as I felt that using the stacked bar chart would make comparing the proportions much more difficult.

Scatter Plot & Box Plot:

I considered using a scatter plot and a box plot in order to display the distribution. In the end, I decided against using these plots as I wanted to create a dual axis chart and hence needed the visualizations to be a line chart and bar graph.

**8. Which product gets used the most in a week by the Customers?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

To visualize the usage days of each product, I decided to use a horizontal bar chart with the y-axis set to ‘product’ while the x-axis is the count. I then set the hue to be ‘usage’ and set the palette to be a color scheme called ‘plasma’ so that the color of the usage days becomes brighter the more there are.

**Reason for Picking Visualization:**

I went with a horizontal bar chart in order to visualize the distribution of usage days as bar charts allow for easier comparison between both the usage days within the same product and usage days against other products. The color scheme used also helps differentiate the different usage day values to make comparing the values easier.

**Core Findings & Insights Found:**

From the analysis and visualization, some core findings identified are,

* The most usage days for FT100 and FT400 is 3 days.
* The most usage days for FT700 is 4 days.
* FT700 is the only product with 6 and 7 usage days.

**Alternatives Considered:**

Pie Chart:

I decided against it as pie charts are bad at visualizing data with 2 values, in this case usage and product.

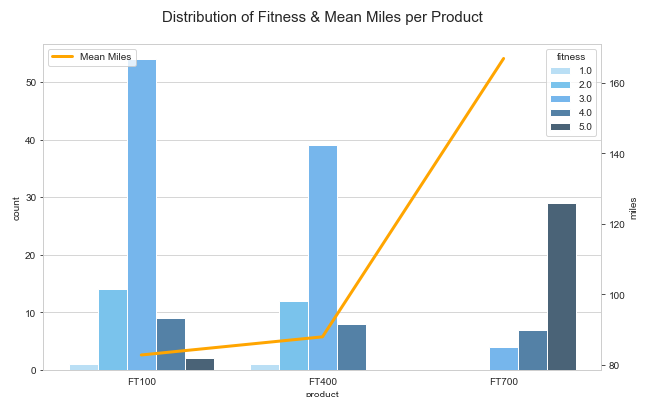
Box Plot & Violin Plot:

I was also considering using a plot like a box plot or a violin plot to show the spread of the data. In the end, I did not use them as I was going to add a few violin charts to the dashboard already and did not want there to be too many of the same graphs.

**9. How fit are the customers who bought each product?**

**10. What are the mean miles that customers said they would cover for all the products?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

In order to showcase the distribution of the customer fitness level per product, I decided to use a dual axis chart. The bar chart portion of the dual axis chart has the x-axis set to be ‘product’ while the y-axis is set to count. The bar chart has a hue of ‘fitness’, which will display the counts of the different fitness levels for each product. The dual axis chart also has a line chart which represents the mean miles covered per product. The line chart has the x-axis set to ‘product’, while the y-axis is set to ‘miles’.

**Reason for Picking Visualization:**

The reason why I decided to use a dual axis chart was so that it could answer 2 exploratory questions while only taking up 1 chart space. I decided that since I already had a chart that showed the different fitness level distribution for each product, I should have one for the mean miles covered based on the product. These 2 charts will complement each other and will save space at the same time.

The reason why I chose the bar graph to display the distribution of the fitness level per product was because I felt that bar charts allowed the data to be easily comparable while looking clean and simple to understand.

**Core Findings & Insights Found:**

From the analysis and visualization, some core findings identified are,

* FT100 has a higher spread of fitness levels as it includes 1-5.
* FT700 has a higher concentration of customers that are fit, only having fitness levels of 3, 4 and 5.
* FT700 also has the highest number of customers that rated themselves a 5.
* The FT100 has the lowest mean miles at around 82 followed by FT400 at around 86 and FT700 at around 170.

**Alternatives Considered:**

Pie Chart & Others:

In the end, I decided against using a pie chart as pie charts are bad at visualizing data that spans across 2 variables, in this case branch and products. I also considered using other charts to showcase the spread in the fitness level across the different products, but decided against it as I wanted to fit it into a dual axis chart.

**11. What is the distribution of the usage for each product?**

**12. What is the distribution of the Fitness level of Customers for each product?**

**13. What is the distribution of the miles covered by the Customers for each product?**

**Main Visualization:**

****

****

****

**Performing Univariate & Multivariate Analysis:**

To effectively answer questions 11-13 which deal with the distribution of data based on the products, I decided to use violin charts to answer them. The charts vary in terms of their y-axis, with the first being usage, the second being fitness and the third being miles. These violin charts all have the same x-axis which is ‘product’. All the violin plots also have the hue set to ‘gender’, so that they can display a more detailed data spread. Alongside this, the violin plots all have the inner set to ‘quartile’ so the lines shown in each violin plot display the median and the interquartile range.

**Reason for Picking Visualization:**

The reason why I picked violin plots to answer questions 11-13 is because I believe that violin charts best display the spread of data and the concentration of data points. This is shown when the width of the violin plot changes according to how many data values there are at a specific point in the graph. I have also chosen to include gender into the violin plots to make the visualizations more detailed. Another reason why I picked the violin plot to answer the questions was because violin plots look appealing and can still look detailed.

**Core Findings & Insights Found:**

Distribution of Usage by Gender per Product:

* The spread of usage for both genders in FT100 is relatively the same, with values from 2-5 usage days.
* The spread of data for females in FT400 is larger than that of the males, with the spread ranging from 2-5 usage days.
* The data points for FT700 are higher than that of FT100 and FT400 ranging from 3-8 usage days.
* The spread of data for males is larger compared to females in FT700.

Distribution of Fitness by Gender per Product:

* All violin plots have more data points concentrated, with a few outliers, causing the violin chart to stretch.
* FT100 and FT400 have the same median of a fitness level of 3.
* FT700 has its data values higher than FT100 and FT400 with a median of a fitness level of 5 for both males and females.

Distribution of Miles by Gender per Product:

* All violin plots have more data points concentrated, with a few outliers, causing the violin chart to stretch.
* The spread of data increases from FT100 to FT400 to FT700.
* FT700 has a significantly larger spread for both genders compared to the other products. The miles range from around 10 to around 410 for FT700.

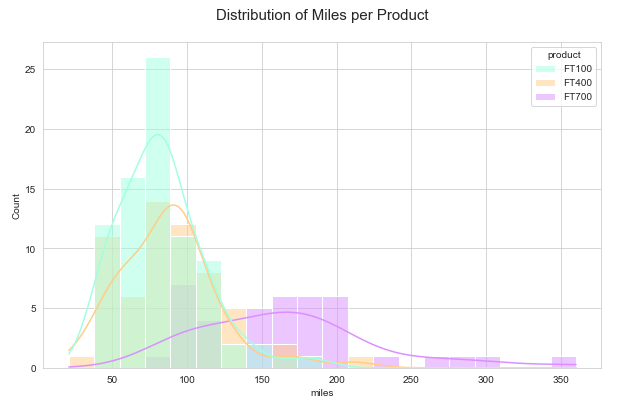
**Alternatives Considered:**

Box Plot:

I was also considering using a box plot to show the spread of the data. In the end, I did not use it as the violin plot portrays the data the same way as a box plot but with additional information like at which point the data points are most concentrated. The violin plot also looked much more appealing to the eyes compared to the box plot.

**14. How do the miles covered by each product compare to each other?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

To compare how the miles covered differ from product to product, I decided to use a histogram with kde lines. The histogram has the x-axis set to ‘miles’ while the hue is set to ‘product’. This plots the distribution of miles by the product against each other.

**Reason for Picking Visualization:**

I chose this visualization to display the distribution in miles as I felt that it best showed off the spread of the miles and the count of the miles. This plot also allows the comparison of the spread and count of miles across products. Alongside these reasons, the histogram plot with kde looks really pleasant.

**Core Findings & Insights Found:**

* FT100 has a lower spread of miles, but a higher count of data from around 50 to 175 miles.
* FT400 has a little higher spread as compared to FT100 but has a lower count, ranging from 0 to 225 miles.
* FT700 has a significantly larger spread of miles and a significantly lower count. The spread ranges from 75 to 350 Miles

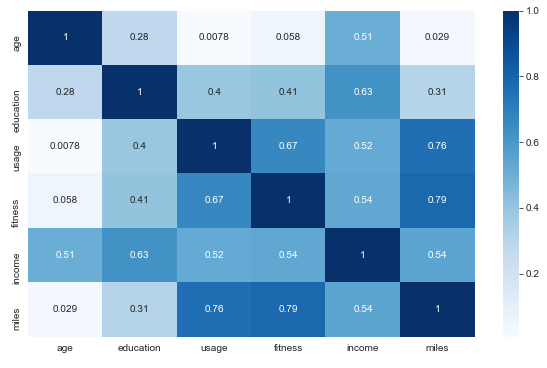
**Alternatives Considered:**

Box Plots & Violin Plots:

I was also considering using a plot like a box plot or a violin plot to show the spread of the data. In the end, I did not use them as I was going to add a few violin charts to the dashboard already and did not want there to be too many of the same graphs.

**15. What is the correlation between all the Customer Characteristic Data in the dataset?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

The heatmap compares the correlation between all the data columns in the excel sheet provided. It does this by plotting a value created by using the corr() python command. The closer the number on the heatmap to 1, the more the relationship between the 2 variables.

**Reason for Picking Visualization:**

I chose to use a heatmap in my visualizations as I felt like the heatmap is the best graph to show the total correlation between all of the data columns. I decided to add the color scheme of ‘Blues’ as it looked really pleasing to the eyes.

At first glance any viewer will be able to tell if 2 variables have a strong or weak correlation based on the shade of blue the cell is. The darker blue the cell, the more the correlation.

**Core Findings & Insights Found:**

* The variable fitness and miles have the strongest correlation amongst all the variables.
* The variable usage and miles have the second strongest correlation amongst all the variables.
* There is a pretty strong correlation between the number of education years and the income.

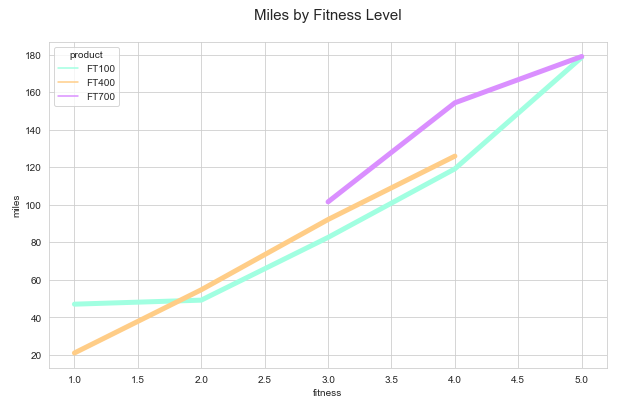
**Alternatives Considered:**

None:

I cannot think of any better charts to show the relationship between the different data variables.

**16. What is the correlation between the Fitness Level of the Customers and the Miles covered?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

In order to display the relationship between the fitness level and the miles covered per product, I used a line plot with the x-axis set to ‘fitness’, the y-axis set to ‘miles’ and the hue set to ‘product’. The resulting plot displayed shows how the relationship is for each product.

**Reason for Picking Visualization:**

I chose to use a line plot in order to plot the relationship of fitness and miles. This is because line plots are simple graphs that display the linear relationship between 2 variables. By using a line plot and setting the hue to be ‘product’ the plot is able to display the relationship for all products, which would be useful for the marketing team when collecting data.

**Core Findings & Insights Found:**

* The relationship is linear and the more the fitness level of the customer, the more the miles covered. This is true for all products.

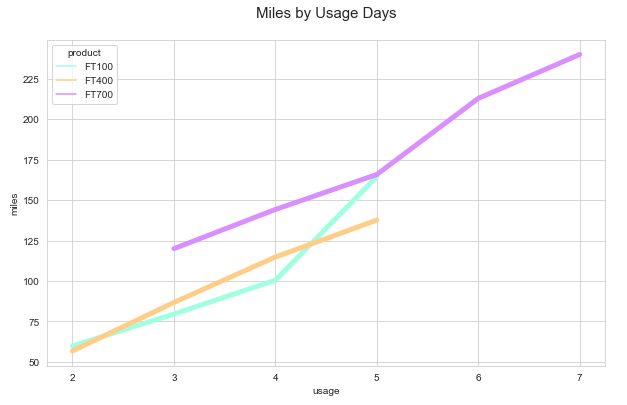
**Alternatives Considered:**

Scatterplot:

I decided against using a scatterplot as I felt like it would make the dashboard I put the visualization in look too complicated. I also liked how simple and easy to read the line plot looks.

**17. What is the correlation between the number of Usage days and the Miles covered?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

In order to display the relationship between the usage days and the miles covered per product, I used a line plot with the x-axis set to ‘usage’, the y-axis set to ‘miles’ and the hue set to ‘product’. The resulting plot displayed shows how the relationship is for each product.

**Reason for Picking Visualization:**

I chose to use a line plot in order to plot the relationship of usage and miles. This is because line plots are simple graphs that display the linear relationship between 2 variables. By using a line plot and setting the hue to be ‘product’ the plot is able to display the relationship for all products, which would be useful for the marketing team when collecting data.

**Core Findings & Insights Found:**

* The relationship is linear and the more the usage days of the customer, the more the miles covered. This is true for all products.

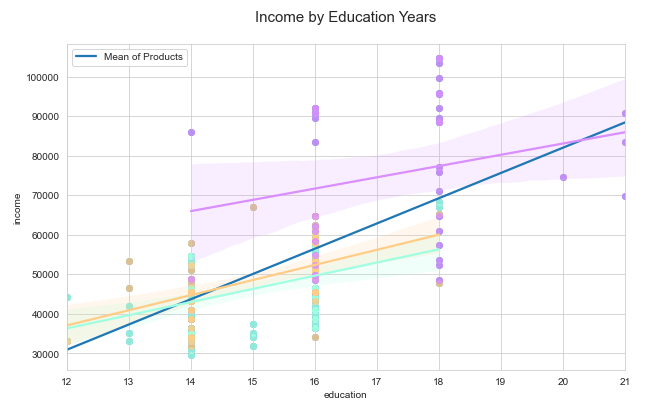
**Alternatives Considered:**

Scatterplot:

I decided against using a scatterplot as I felt like it would make the dashboard I put the visualization in look too complicated. I also liked how simple and easy to read the line plot looks.

**18. What is the correlation between the number of years of Education received by the customers and the income they make?**

**Main Visualization:**

****

**Performing Univariate & Multivariate Analysis:**

To display the correlation between the number of years of education and the income of the customer, I decided to use a linear regression model fit plot. This regplot has it’s x-axis set to ‘education’ while the y-axis is set to ‘income’. Due to the nature of reg plots, there is no hue command in order to show the relationship for each product. Hence, in order to overcome this problem, I plotted more reg plots over the initial one with the data set to be the product names. This enabled me to show the relationship for all products along with the specific data points, all in the relevant product color.

**Reason for Picking Visualization:**

I chose to use the Regression Model Fit plot to visualize the relationship between the education years and the income as the regression model shows the relationship line along with the different data points. This allows the plot to be more detailed and lets the viewer of the plot view more data.

**Core Findings & Insights Found:**

* There is a positive linear relationship between the number of education years and the income made by the customer. This relationship applies to all products.

**Alternatives Considered:**

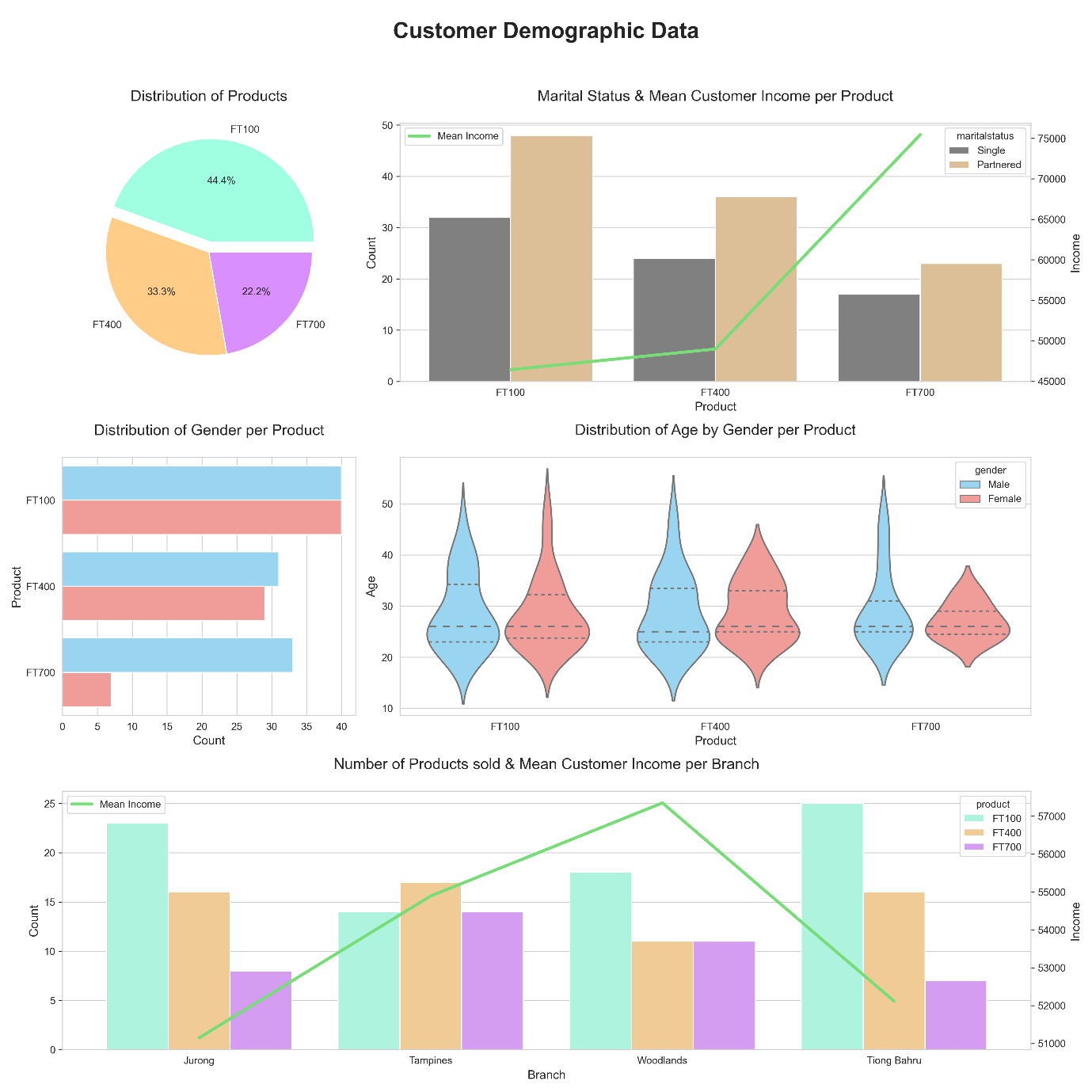
Line Plot:

I decided against using a line plot as I felt that the regression plot works better in this scenario. This is because it can both show the different data points to the user and even plots a line which showcases the relationship.

Scatterplot:

I decided against using a scatterplot as I feel like the regression plot works better in this scenario. This is because it can both show the different data points to the user and even plots a line which showcases the relationship.

**4. Dashboards**

****

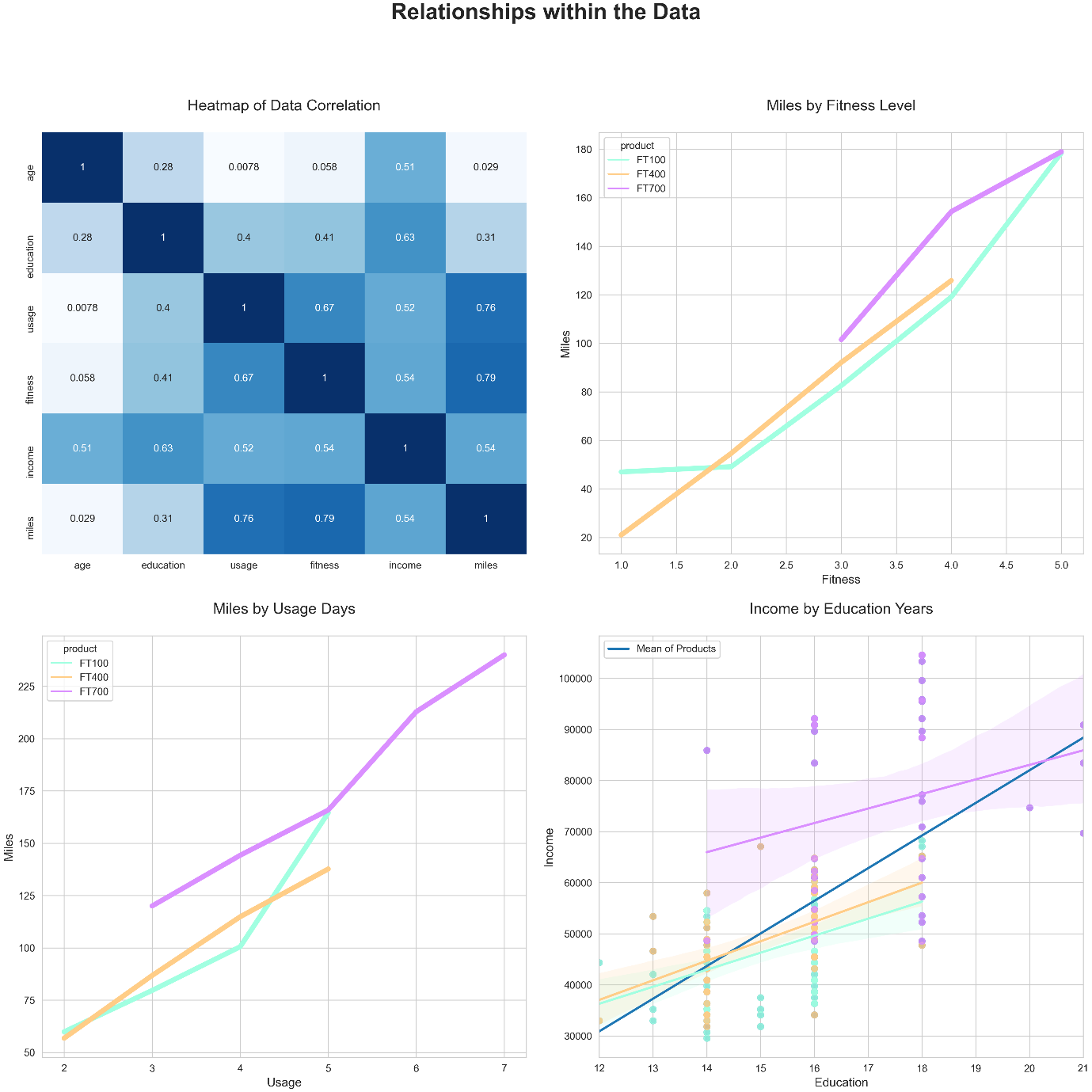
The first Dashboard I created using the Customer Characteristics data was the Customer Demographic Dashboard. This dashboard mainly covers how the demographic of the customer changes based on the product. This can be in the form of age, gender, marital status, income, etc.

Through the use of this dashboard, the marketing team will understand who the customers the bicycles are being bought by. This will allow for better marketing tactics as the team will know who to target and also will know which areas they need to promote the products more in are.

****

The second dashboard I created using the data provided is the Customer Fitness Dashboard. This dashboard covers the 3 fitness data columns in the dataset, fitness, usage and miles. Through the use of visualizations, the dashboard will convey the fitness level of the customers and how it changes across the products.

Through this dashboard, the marketing team will understand the general fitness level that each product is targeted and used for. This will allow them to increase their marketing tactics so that they can advertise the appropriate products based on the customer’s fitness preferences.

****

The third dashboard covers the relationships within the data. The first heatmap plot informs the user on how each data column is correlated to each other. In this case, the darker the cell, the more the correlation between the data columns. The other 3 plots in the dashboard show the relationship between some of the data columns across the products. One compares how the level of fitness affects the miles covered, another compares how the usage days affects the miles covered and lastly, one compares how the number of education years affects the income earned by the customer.

Through these graphs, the marketing team will understand how each data column is related to each other and what some of the interesting relationships are, so that they will be able to come up with interesting marketing tactics to appeal to the trends.

### Conclusion

### The FT100 is a beginner-friendly bicycle model that is also budget-friendly. As such the FT100 is the best-selling bicycle model sold by the company. This bicycle appeals to all ages and genders as seen by the spread in the data.

#### The FT400 is a mid-tier bicycle that appeals to customers who have a little more money to spend as compared to the FT100. The FT400 appeals to most age groups, with the less of the older female gender buying this model. This bicycle appeals towards customers who will use the bicycle to cover more miles and will be used more frequently than the FT100.

#### The FT700 is a high-tier expensive bicycle model that appeals to customers who have money to spend and want to buy a bicycle that will be used to travel more miles, will be used more frequently and will be used by fitter people as compared to the other models. This bicycle is bought by more males across a wide spread of age than women.