# 



# School of InfoComm Technology

**Data Visualisation**

Diploma in Financial Informatics

April 2021 Semester

**ASSIGNMENT 2**

**(Individual Assignment)**

**Submission Deadline:**

**15th August 2021 (Thursday), 23:59PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **P04** |
| **Student Name** | **:** | Chen Han |
| **Student Number** | **:** | S10202961H |

**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.

Link to presentation: <https://web.microsoftstream.com/video/46b51e38-132a-4296-80ac-2b91b69d3482>

**1. Project Objectives**

**Goal: To investigate whether there are differences across the models with respect to customer characteristics.**

**Model Evaluation (Treadmill model sales understanding)**

1. Which model has least amount of sales for the past three months (or by geographic area)?

**Customer Characteristics vs Sales of the three models**

2. How is gender affects customer purchasing behavior on the three models?

3. How is the age affects customer purchasing behavior on the three models?

4. How is education level affects customer purchasing behaviors on the three models?

5. How is marital status affects customer purchasing behaviors on the three models?

6. How is the income affects customer purchasing behaviors on the three models?

**User Usage Goal and Expectations**

7. How is the customer self-rated fitness affects their expected miles walk/run?

8. How does the customer rate their fitness?

9. What are the customers expected average number of usage per week?

10. What are the customers expected number of miles walk/run each week?

**2. Data Preparation**

* **State of data**

The initial state of data is not cleaned. There are total of 183 rows of data, other than Product and Branch, all other columns contain more than missing values.

For column MaritalStatus, there are few irregular values like “S” and “P”, which supposed to be “Single” and “Partnered”. Thus, a substantial data cleanup is definitely needed to fill up all missing values and make sure data in MaritalStatus is regular and consistent.

* **Data cleansing**

1. First by defining a list of possible missing values (e.g. “na”). So that when reading the data, it will check and replace the missing values with NULL.

*missing\_values = ["n/a", "na", "N.A.", "--"]*

*df = pd.read\_csv("Fitness Trainer Series \_ April 2021.csv", na\_values = missing\_values)*

RangeIndex: 183 entries, 0 to 182

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Product 183 non-null object

1 Branch 183 non-null object

2 Age 174 non-null float64

3 Gender 180 non-null object

4 Education 180 non-null float64

5 MaritalStatus 180 non-null object

6 Usage 172 non-null float64

7 Fitness 180 non-null float64

8 Income 180 non-null float64

9 Miles 180 non-null float64

dtypes: float64(6), object(4)

2. Check which are the columns contains NULL values.

*df.isnull().any()*

Product False

Branch False

Age True

Gender True

Education True

MaritalStatus True

Usage True

Fitness True

Income True

Miles True

3. Next, fill in the NULL values in **Age, Education, Usage, Fitness, Income and Miles** with the median of all the values of that column.

And convert the data type to **int** instead of **float** by default.

*df['Age'].fillna(df['Age'].median(),inplace = True)*

*df["Age"] = df["Age"].astype(int)*

*df['Education'].fillna(df['Education'].median(),inplace = True)*

*df["Education"] = df["Education"].astype(int)*

*df['Usage'].fillna(df['Usage'].median(),inplace = True)*

*df["Usage"] = df["Usage"].astype(int)*

*df['Fitness'].fillna(df['Fitness'].median(),inplace = True)*

*df["Fitness"] = df["Fitness"].astype(int)*

*df['Income'].fillna(df['Income'].median(),inplace = True)*

*df["Income"] = df["Income"].astype(int)*

*df['Miles'].fillna(df['Miles'].median(),inplace = True)*

*df["Miles"] = df["Miles"].astype(int)*

4. Next, handling irregular values for MaritalStatus, convert all “S” and “P” into “Single” and “Partner”.

*df["MaritalStatus"] = df["MaritalStatus"].apply(lambda x: "Single" if x == "S" else x)*

*df["MaritalStatus"] = df["MaritalStatus"].apply(lambda x: "Partnered" if x == "P" else x)*

5. Lastly, I realized that there are 3 missing values in Gender and MaritalStatus, so I made a decision to drop these three rows, because in these three rows, only Product and Branch columns are filled with value, and the rest are NA values, which does not help in the data visualization later.

*df = df.dropna(axis='rows')*

Int64Index: 180 entries, 0 to 182

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Product 180 non-null object

1 Branch 180 non-null object

2 Age 180 non-null int32

3 Gender 180 non-null object

4 Education 180 non-null int32

5 MaritalStatus 180 non-null object

6 Usage 180 non-null int32

7 Fitness 180 non-null int32

8 Income 180 non-null int32

9 Miles 180 non-null int32

dtypes: int32(6), object(4)

**3. Visualizations**

* + Describe the process to perform univariate and multivariate analysis on your datasets.

I first by visualizing all unique values that appears in each column of the cleansed dataset.

*#Visualise unique values in the columns again*

*print("Product",df["Product"].unique())*

*print("Branch",df["Branch"].unique())*

*print("Age",df["Age"].unique())*

*print("Gender",df["Gender"].unique())*

*print("Education",df["Education"].unique())*

*print("MaritalStatus",df["MaritalStatus"].unique())*

*print("Usage",df["Usage"].unique())*

*print("Fitness",df["Fitness"].unique())*

*print("Income",df["Income"].unique())*

*print("Miles",df["Miles"].unique())*

**Output:**

Product ['FT100' 'FT400' 'FT700']

Branch ['Jurong' 'Tampines' 'Woodlands' 'Tiong Bahru']

Age [18 19 20 21 22 23 24 25 26 27 28 30 31 32 33 34 35 36 37 38 39 40 41 43 44 46 47 50 45 48 42]

Gender ['Male' 'Female']

Education [14 15 12 13 16 18 20 21]

MaritalStatus ['Single' 'Partnered']

Usage [3 2 4 5 6 7]

Fitness [4 3 2 1 5]

Income [ 29562 31836 30699 32973 35247 37521 36384 38658 40932 34110

39795 42069 44343 45480 46617 48891 53439 43206 52302 51165

50028 54576 68220 55713 60261 67083 56850 59124 61398 57987

64809 47754 65220 62535 48658 54781 48556 58516 53536 61006

57271 52291 49801 62251 64741 70966 75946 74701 69721 83416

88396 90886 92131 77191 52290 85906 103336 99601 89641 95866

104581 95508]

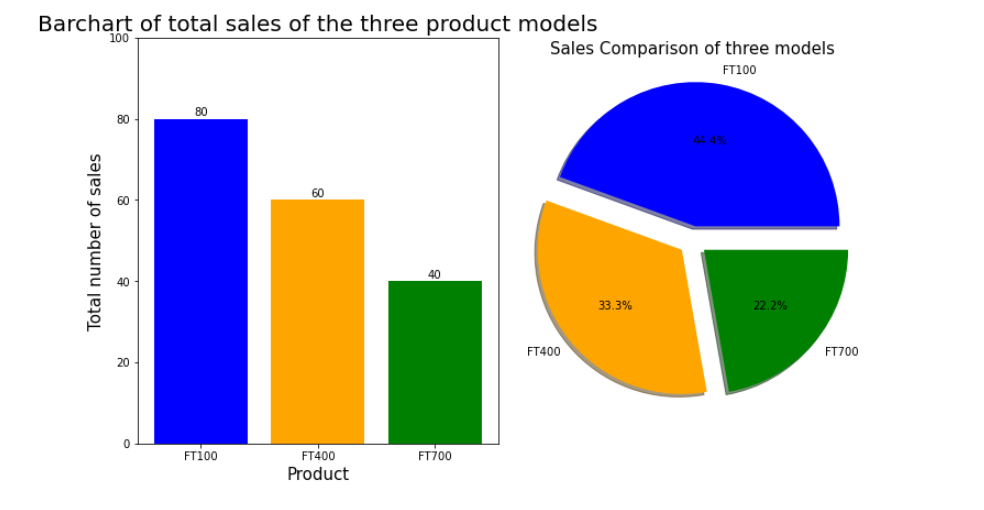
Miles [112 75 66 85 47 141 103 94 113 38 188 56 132 169 64 53 106 95 212 42 127 74 170 21 120 200 140 100 80 160 180 240 150 300 280 260 360]

From here, I came out with some ideas on how to visualize these data.

Since the goal of this assignment is to investigate whether there are differences across the models with respect to customer characteristics. So grouped bar chart is a good tool to compare the difference in sales of the three models by customer characteristics. Other than grouped bar charts that shows only the count of the sales, violin chart can also be used to observe the distribution of numeric data by the product models such as Usage and Fitness. A scatter plot can also be used to see the relationship between two numeric variables such as the correlation between Fitness and Miles.

**Visualizations**

1. Overall sales comparison between the three models.



This visualization consists of one bar chart and one pie chart, showing the sales comparison of the three product models. It shows that model FT100 has the highest amount of sales of 80 (44.4%) as compared to FT400 and FT700, it has double of sales to FT700. This tells us that close to half of the customers chose FT100 over the other two models. Also FT700 has the least amount of sales of only 22.2% among others. Therefore, in the data exploratory process later, I will take a deeper look into this model and investigate the reason for it having the lowest sales. Also through data visualization, suggest various ways to improve sales on FT700 model.

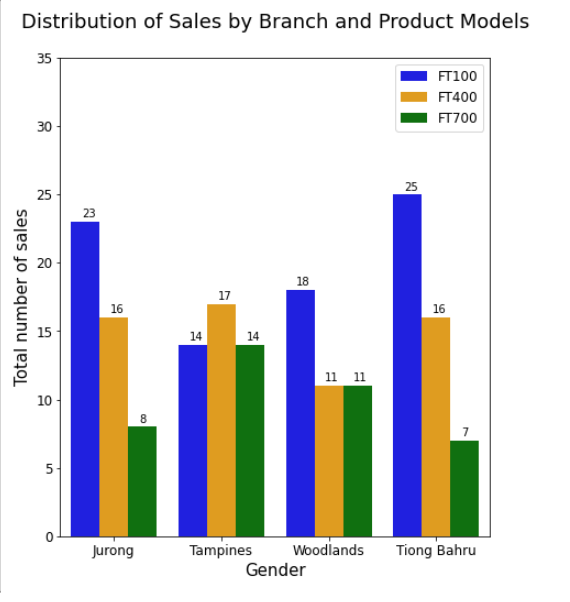
Business/research question answered:

Which model has least amount of sales for the past three?

- Model FT700 has the least amount of sales.

Visualization used: One bar chart and one pie chart. Using bar chart allows the user to visualize the difference in the sales of each model easily, the text label on the bar tells how many sales does one model has. Whereas using a pie chart on the side helps to create contrast between model FT100 and FT700 occupying the largest and smallest portion of the circle. It also gives the overall percentage of sales by each model.

2. Comparing sales of Product models by different branches



This visualization shows a grouped bar chart that compares sales over branch and product models. It shows that Jurong and Tiong Bahru branches has a significant difference in the sales across the models. Thus, these two branches should promote model FT700 as they are having the lowest sales.

Business/research question answered:

Which model has least amount of sales for the past three months by geographic area?

* Model FT700 has the least amount of sales in Jurong and Tiong Bahru area.

Visualization used: A grouped bar chart, as it is a great tool to compare multiple sets of data items. In this case, it is used to compare the difference in sales of three models by four different branches.

3. Comparing sales of Product models by Gender

Chart, pie chart

Description automatically generated

This visualization shows a grouped bar chart together with a pie chart that compare the sales over gender and product models. It shows that male customers has purchased more products than female customers in overall. For male customers, it seems that each model has a equal chance of being chosen by the customer, but for the female customers, the sales of FT700 is very low as compared to other models. This means that model FT700 is more popular in male customers comparing to female customers. Thus, to understand the customer more, the company should conduct a survey with female customers to find out why there is a low sales.

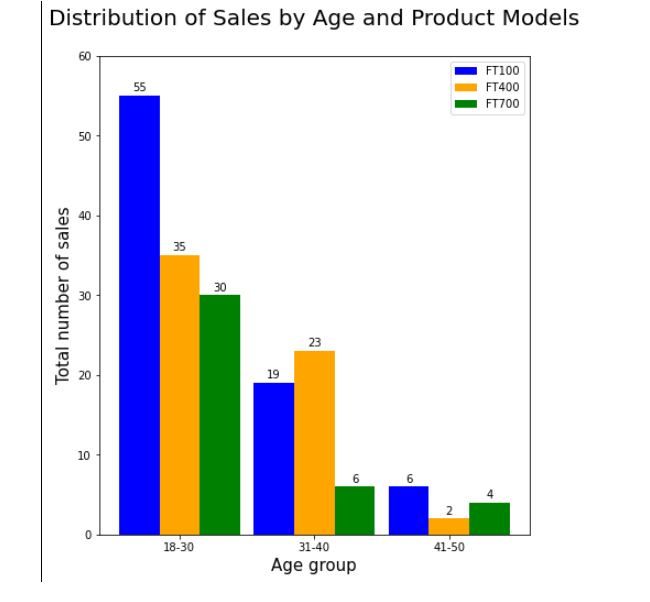
Business/research question answered:

How is gender affects customer purchasing behavior on the three models?

* Male customers are more likely to purchase model FT700 than female customers.

Visualization used: A grouped bar chart and a pie chart. Using grouped bar chart helps to compare the difference in sales of three models by the gender. While the pie chart on the right helps to give an overall visual of the difference in overall sales by gender.

4. Comparing sales of Product models by different age groups



This visualization shows a grouped bar chart that compare the sales over different age group and product models. I have split the age into 3 age groups, what are 18 to 30, 31 to 40 and 41 to 50 years old. This bar chart shows that most of the purchases are made from customers in age group 1, which are young people from age 18 to 30. This means that all the three models are popular in the young generations, and the company can use this as a guidline to sell products to target customers who are around 18 to 30 years old.

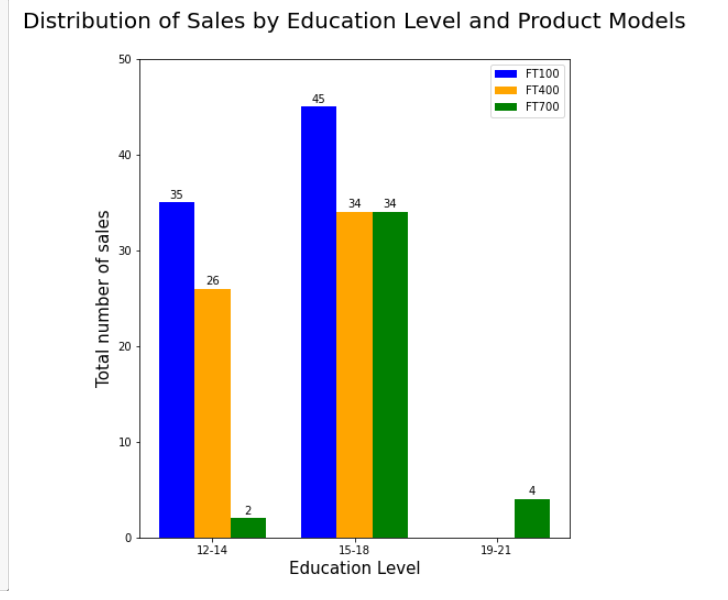
Business/research question answered:

How is the age affects customer purchasing behavior on the three models?

* Customer aged between 18 and 30 are more likely to purchase the product.

Visualization used: A grouped bar chart, as it is a great tool to compare multiple sets of data items. In this case, it is used to compare the difference in sales of three models by three different age groups.

5. Comparing sales of Product models by different education level



This visualization shows a grouped bar chart that compare the sales over different education level and product models. Similar to the previous visualization, I have split the education into 3 groups, what are 12 to 14, 15 to 18 and 19 to 21 years of education. This bar chart shows that most of the purchases are made from customers in the middle education level, who are probably undergraduate students. This suggest that this product is quite complex in terms of functionality and requires a level of education to understand and use.

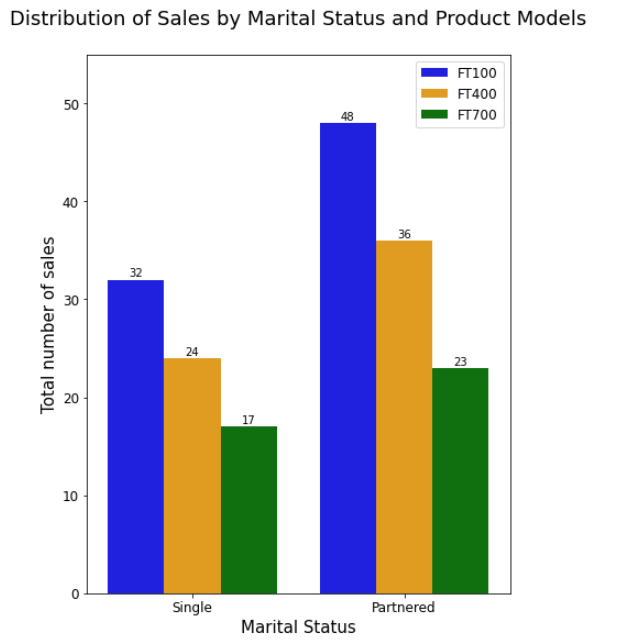
Business/research question answered:

How is education level affects customer purchasing behaviors on the three models?

* Customer who had between 15 to 18 years of education are more likely to purchase the product.

Visualization used: A grouped bar chart, as it is a great tool to compare multiple sets of data items. In this case, it is used to compare the difference in sales of three models by three different education levels.

6. Comparing sales of Product models by marital status



This visualization shows a grouped bar chart that compares the sales over different marital status and product models. It shows that most of the purchases are made from customers who are partnered.

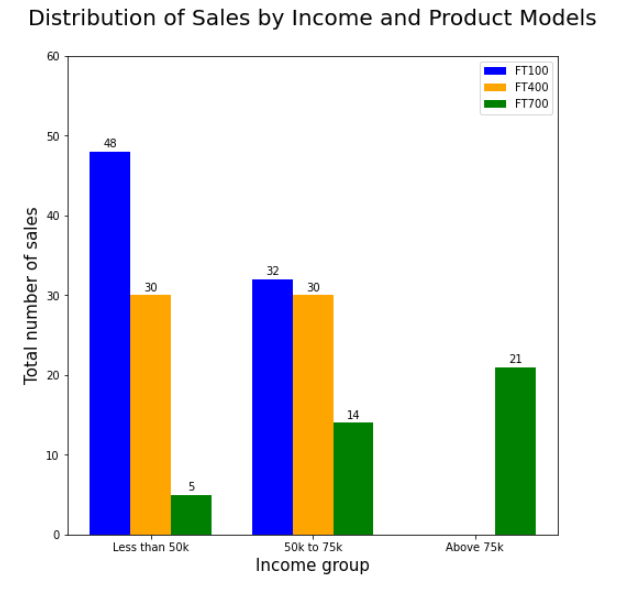
Business/research question answered:

How is marital status affects customer purchasing behaviors on the three models?

* Customer who are married are more likely to purchase the product.

Visualization used: A grouped bar chart, as it is a great tool to compare multiple sets of data items. In this case, it is used to compare the difference in sales of three models by two different marital status.

7. Comparing sales of Product models by different income groups



This visualization shows a grouped bar chart that compares the sales over different income groups and product models. It shows that model FT100 and FT400 are more cater to low and middle income group whereas FT700 is more cater to middle and high income group. Thus, I can infer that model FT700 costs much higher than FT100 and FT400 which cause it to having the lowest sales. If the company wants to increase the sales of FT100 alone, should make it more affordable for low and middle income group.

Business/research question answered:

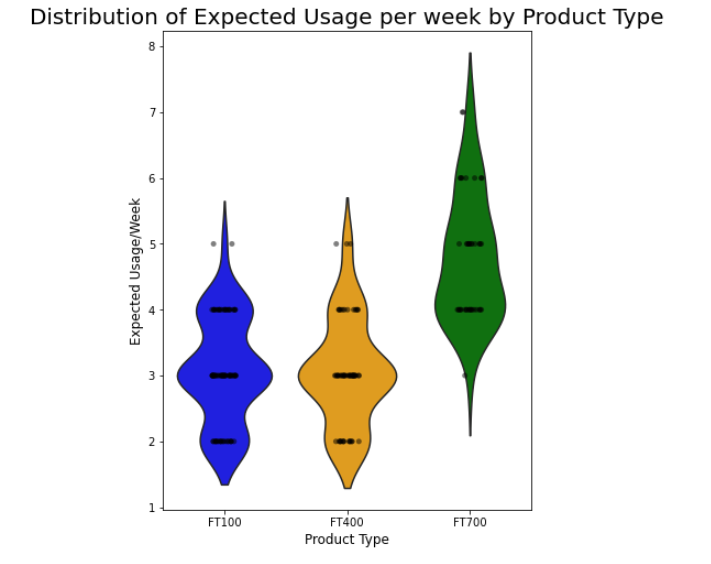
How is the income affects customer purchasing behaviors on the three models?

* Customer from lower income groups are more likely to purchase model FT100 and FT400, whereas customers from higher income group are more likely to purchase model FT700,

Visualization used:

A grouped bar chart, as it is a great tool to compare multiple sets of data items. In this case, it is used to compare the difference in sales of three models by three different income groups.

8. Comparing sales of Product models by customer expected usage per week.



This visualization shows a violin chart that compares the sales of the three models over the expected usage per week by the customer. It shows that customers who purchased model FT700 has a higher expected usage per week than model FT100 and FT400, this means that model FT700 user has higher fitness goal than other model users.

Business/research question answered:

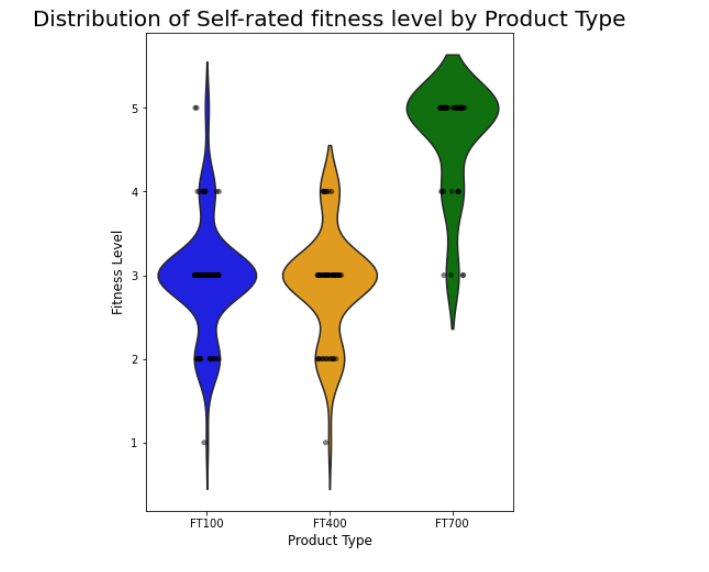
What are the customers expected average number of usage per week?

* Most of the FT100 and FT400 users have expected usage per week at round 2 to 5 times. While most of the FT700 users have expected usage per week at round 4 to 7 times.

Visualization used:

A violin chart, which is a great tool to use when one wants to observe the distribution of numeric data, and are especially useful when one wants to make a comparison of distributions between multiple groups.

9. Comparing sales of Product models by customer self-rated fitness level



This visualization shows a violin chart that compares the sales of the three models over the customer self-rated fitness level (1 to 5) . It shows that customers who purchased model FT700 has a higher overall fitness rating than other models. This also means that model FT700 user has higher fitness goal than other model users.

Business/research question answered:

How does the customer rate their fitness?

* Model FT700 users are have higher self-rated fitness of around 3 to 5, where other models are more concentrated in 2 to 3.

Visualization used:

A violin chart, which is a great tool to use when one wants to observe the distribution of numeric data, and are especially useful when one wants to make a comparison of distributions between multiple groups.

10. Comparing sales of Product models by different branches



This visualization shows a violin chart that compares the sales of the three models over the expected miles walk/run per week by the customer. It shows that customers who purchased model FT700 has a higher miles walk/run than other models. This also means that model FT700 user has higher fitness goal than other model users.

Business/research question answered:

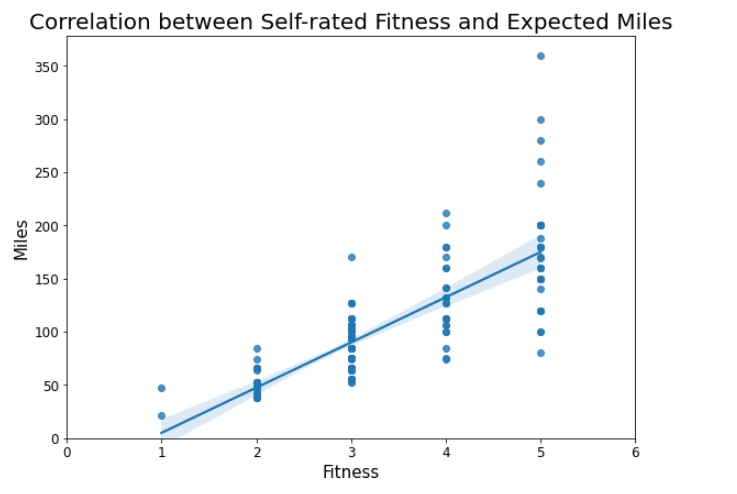
What are the customers expected number of miles walk/run each week?

* The customer who have purchased model FT100 and FT400 are expected to walk/run around 100 miles per week. While the customer who have purchased model FT700 have miles expected walk/run of 100 miles and above.

Visualization used:

A violin chart, which is a great tool to use when one wants to observe the distribution of numeric data, and are especially useful when one wants to make a comparison of distributions between multiple groups.

11. Comparing sales of Product models by different branches

****

This visualization shows a scatter plot with a best-fit line that helps to visualize the correlation between the customer self-rated fitness level and the expected miles walk/run per week. The linear best-fit line shows an increasing trend, which means the higher the fitness level of a customer, the longer distance that he planning to walk/run.

From here, we can claim that the dataset is accurate and reliable.

Business/research question answered:

How is the customer self-rated fitness affects their expected miles walk/run?

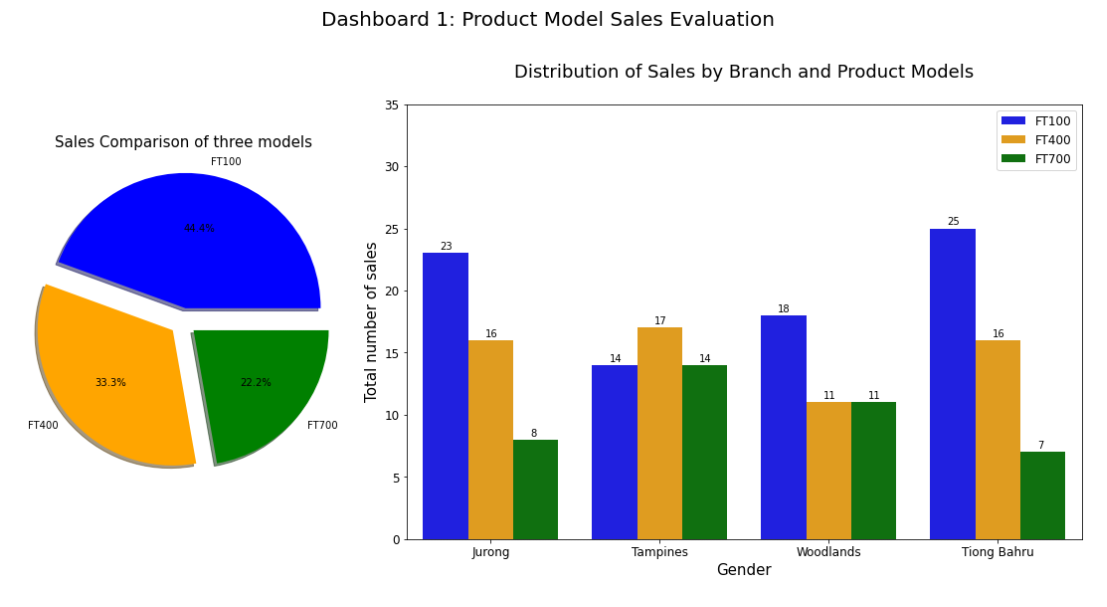
* The higher the self-rated fitness level of a customer, the longer the distance that he expected to walk/run.

Visualization used:

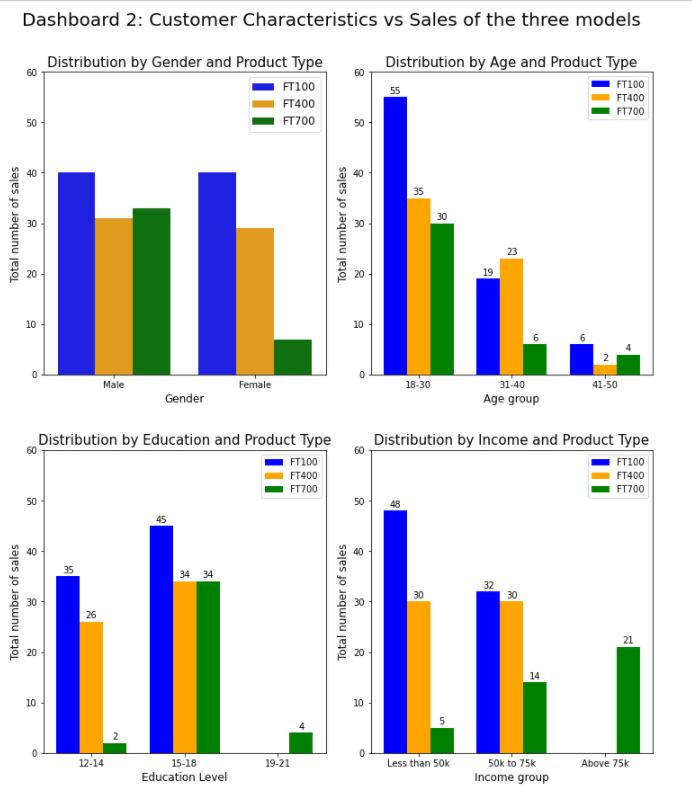
A scatter graph, which use dots to represent values for Fitness and Miles and shows the relationship by these two numeric variables, with a best-fit line showing the trend.

**4. Dashboards**

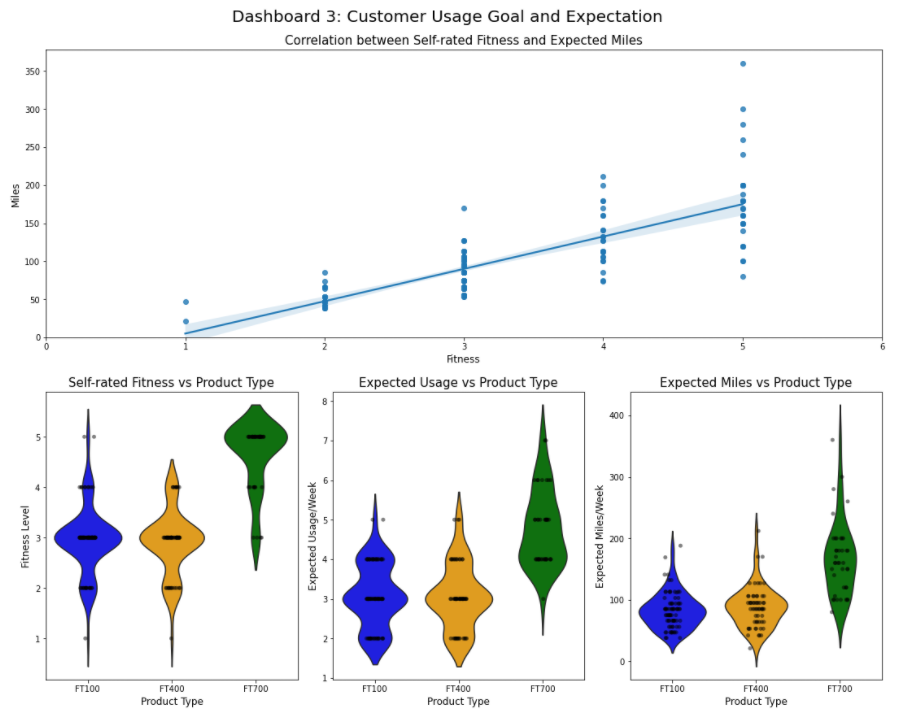
**Dashboard 1**

****

**Dashboard 2**

****

**Dashboard 3**

****

**References**

<https://chartio.com/learn/charts/violin-plot-complete-guide/>