**ATOM CAMP**

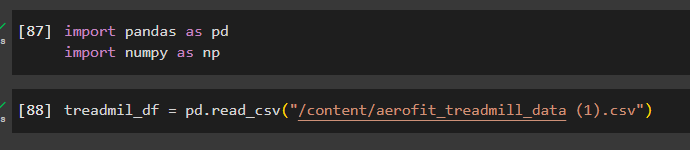
**AI AND DS BOOTCAMP**

EDA PORTFOLIO PROJECT REPORT

|  |  |
| --- | --- |
| **NAME** | Sarah Qasim |
| **MODULE NAME** | EDA |
| **DATE** | 14-12-2024 |

**1. Data Exploration and Processing:**

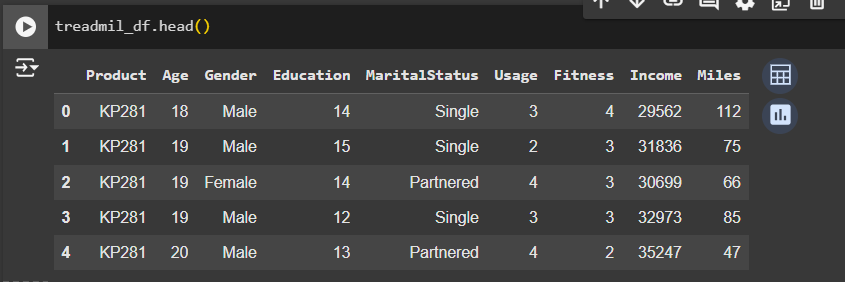
* **Importing data**



Importing pandas and numpy (numerical functions)

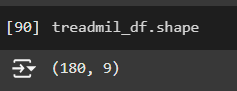
Importing csv file using pandas function: read\_csv

* **Reading dataframe**

****

Initial reading of the dataset using head function which shows the first five records in the dataset by default

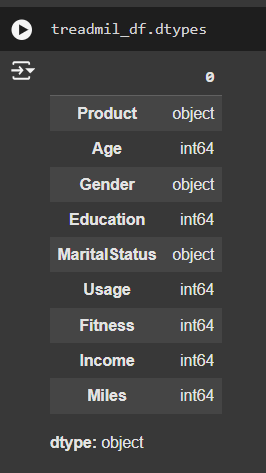
* **Check the shape of the dataframe**

****

Using the shape method to see the number of rows and columns

Here, we have 180 rows and 9 columns

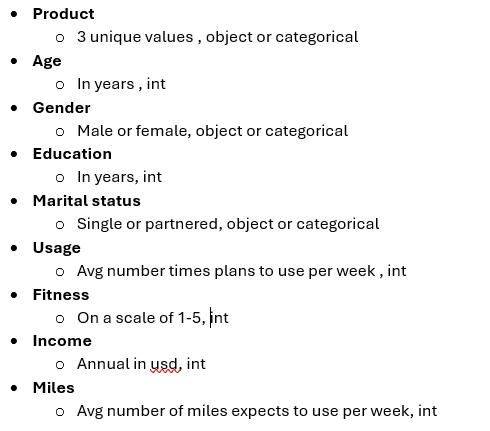
* **Datatype of each column**

****

Using dtypes method to get the data types for each of the 9 columns in the dataset.

Comparing these results to what I expected looking at the data in the columns.

What I expected:



So, no need to change any data type

* **Missing value detection**

A screenshot of a computer

Description automatically generated

Using the isna function to get Boolean output of True (where values are missing) and False (where values are not missing)

Each true corresponds to 1 so applying sum function counts all true values for each column.

Here, none of the values are missing

* **Checking duplicate values in the dataset**

A screen shot of a computer

Description automatically generated

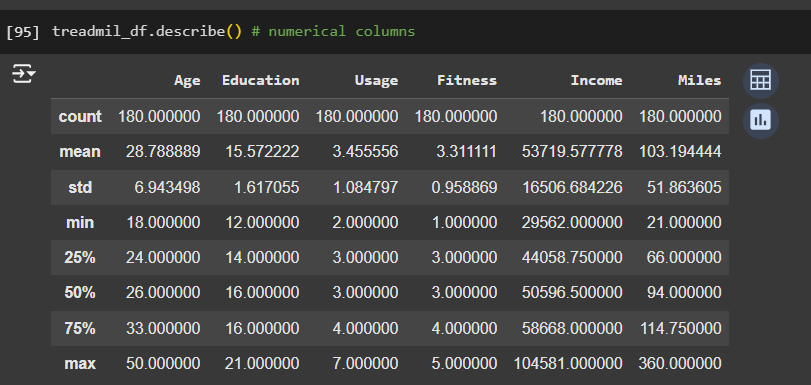
In the same way, the Boolean outputs for the duplicated function are counted using sum function.

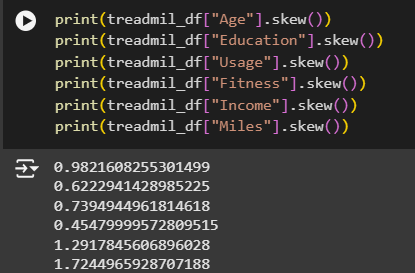
No duplicates here.

**2. Statistical Summary:**

Provide an analysis of the statistical summary in few lines for both categorical and numerical features.

* **Numerical columns**

****

****

**Age:** Skewed towards younger people, most people under 30

**Education:** Generally, evenly distributed with the mean and median being close to each other and all customers being at least educated till the 12th grade.

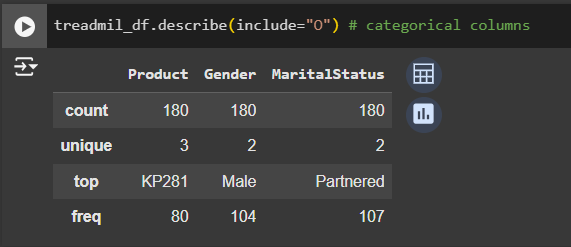
**Usage:** Slightly skewed towards less usage times per week with majority being less than or equal to 4 times a week.

**Fitness:** Mostly symmetrical with most customers rated themselves “3”

**Income:** Data is spread out with a wide range from 29.5k to 104k with most values less than the median i.e. 50.5k .

**Miles:** Data is spread out with a wide range from 21 to 360 miles with most values concentrated between the 25th and 50th percentile.

* **Categorical columns**



**Product:** The KPI281 is the most bought treadmill type.

**Gender:** Most customers are male.

**Marital status:** Most customers have partners.

**3. Non-Graphical Analysis:**

* **Value Counts for all categorical features**

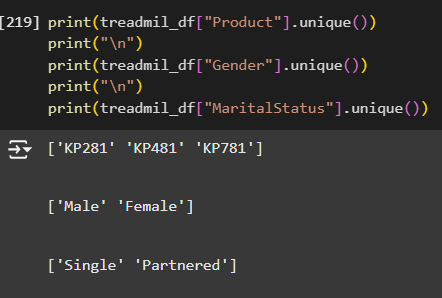
For all 3 categorical columns, using the function value\_counts which by default sorts the values in descending order

**A screenshot of a computer

Description automatically generated**

We see a more detailed look at the distribution of all categories in these columns.

* **Unique Attributes for all categorical features**

****

Here, we see the unique values or categories for these columns using unique function.

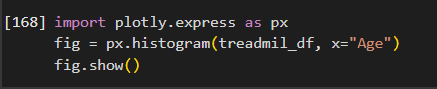
**4. Graphical Analysis**

**• Univariate Analysis - Numerical features**

Univariate refers to the analysis of data in one particular feature or column of the data

**o Distribution Plot**

For distribution of numerical columns, we will look at histograms using plotly package and its subpackage express to plot.



We import the express subpackage as px

This can be used to call the histogram function where the first argument is the dataset, and the x variable argument is the numerical column to be plotted.

**Age**

A graph of age and age

Description automatically generated

As seen earlier, age is skewed towards the younger ages

**Education**

**A graph of a graph

Description automatically generated**

Most people are educated till the 16th level while all people are at least educated till the 12th grade.

**Usage**

**A graph with a blue line

Description automatically generated with medium confidence**

Most people think they will use the treadmill they buy 4 or less times a week.

**Fitness**

**A graph with a bar and text

Description automatically generated with medium confidence**

Most people rate their fitness level right in the middle of the range at 3.

**Income**

**A graph of income

Description automatically generated**

Skewed highly towards the lower income values less than 60k per year

**Miles**

**A graph of a number of miles

Description automatically generated**

Highly skewed towards the lower miles with most people predicting they will cover less than 120 miles per week.

**o Count Plot**

This plot is similar to histogram but is more suited towards numerical values with some unique values.

In this case, better suited for Education with 8 unique grades, usage with a range of 1-7 and fitness with a range of 1-5 in which people have to choose from as we will see below with the result.

I am using seaborn and matplotlib’s pyplot subpackage here.

A screenshot of a computer program

Description automatically generated

Seaborn and matplotlib.pyplot are imported with their short form names.

A palette is created for all following graphs using a variable called custom which contains colours chosen from the matplot website

Plotting using the catplot function with argument data containing the dataset, argument x containing the column to be plotted and the kind (of graph) argument being count.

**Age**

**A graph of age and age

Description automatically generated**

**Education**

**A graph of a graph

Description automatically generated**

We can see the distinct grades of education much better in this plot

**Usage**

**A graph of a number of objects

Description automatically generated with medium confidence**

**Fitness**

**A graph of a bar graph

Description automatically generated**

**Income**

**A graph with purple lines

Description automatically generated**

**Miles**

**A graph of a number of miles

Description automatically generated**

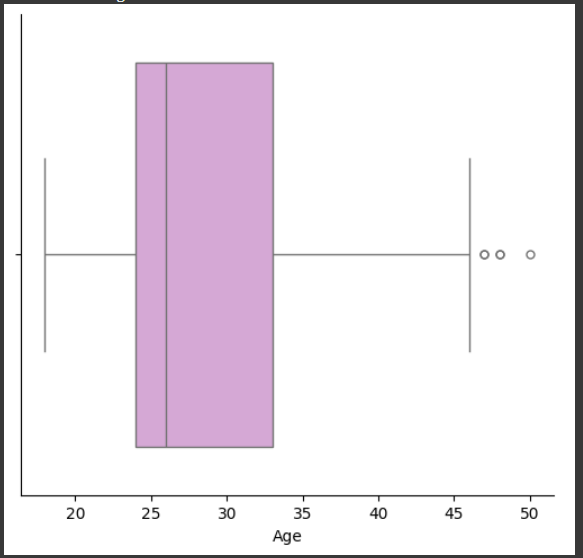
**o Box Plot**

A box plot looks at numerical data and divides it into

* A filled box that represents the interquartile range starting from the 25th quantile and ending at the 75th quantile with a median line inside the box
* There are whiskers coming out of the box on either side that represents the 25th quantile + 1.5\* interquartile and 75th quantile + 1.5\* interquartile
* Finally, any data point outside these whiskers are shown as dots. These represent the outliers in the data.

Made using seaborn’s catplot function where the kind argument is now “box”.

**Age**

****

We see a clear picture of the data points being concentrated between 24 and 33.

**Education**

**A graph with a purple rectangle

Description automatically generated**

We observe the median and 75th percentile values are the same and values are concentrated between 14 and 16.

**Usage**

**A graph with a purple rectangle

Description automatically generated**

**Fitness**

**A graph with a purple rectangle

Description automatically generated**

In both usage and fitness, the majority of values are between 3 and 4 with very few outliers.

**Income**

**A graph of a bar graph

Description automatically generated**

This column has a lot of outliers for customers with very high income as most of the values are between 45k and 60k.

**Miles**

A graph with a purple rectangular object

Description automatically generated

The miles column also has some outliers for very high miles expected by the customers to be covered as most customers expected to cover miles between 60 to 120 miles per week.

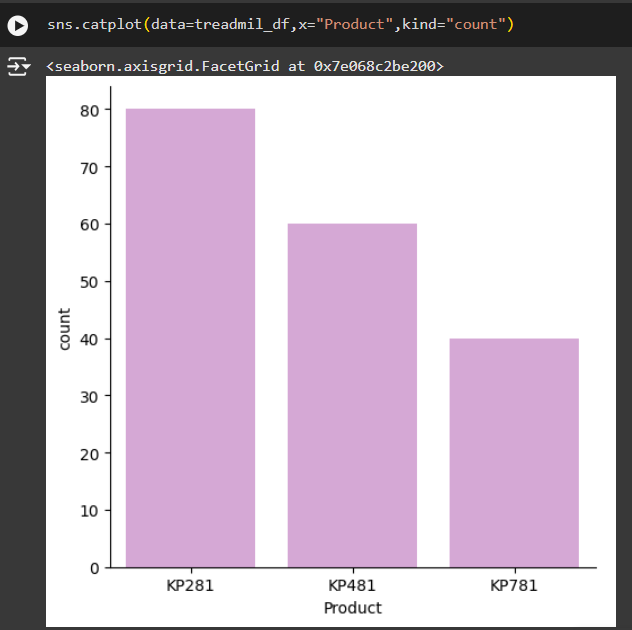
**• Univariate Analysis - Categorical features:**

**o Count Plot**

We now do the same count plot method for the three categorical columns which counts number of values in each category for that column similar to the non-graphical value\_counts function.

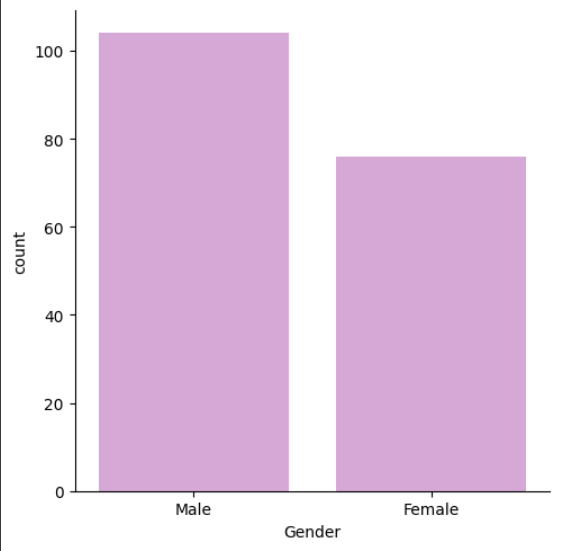
This is done as before using the catplot function inside seaborn with kind = “count”.

**Product**

****

We see a trend in the sale of products with the more expensive and advanced feature treadmills having the least number of customers.

**Gender**

****

Most customers are male, but females are not far behind making up about 43% of the total customers.

**Marital Status**

**A graph of a number of people

Description automatically generated with medium confidence**

Most people who bought the treadmills are partnered but single people are not far behind making up about 40% of the total customers.

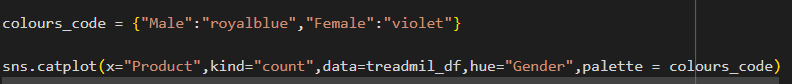
**Bivariate Analysis**

**Check features effect on the product purchased**

**PRODUCT VS GENDER**

These are two categorical columns so we can compare the two using multiple columns in a count graph.

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is count
* x is Product
* hue is Gender
* palette is a dictionary made earlier that stores “royalblue” for Male and “violet” for Female to show the gender contrast better.



A graph of different colored bars

Description automatically generated

We see that KP281 has an equal number of male and female customers, KP481 has fewer female customers while KP781 has the least number of female customers compared to male customers.

The total number of product types sold decreasing from KP281 to KP781

**PRODUCT VS AGE**

One column is categorical, and one is numerical so we can plot the two on a box graph which the categorical column being on the x axis and the numerical column on the y axis**.**

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is box
* x is Product
* y is Age
* palette is a pre-existing palette called “Set2”

A chart with different colored squares

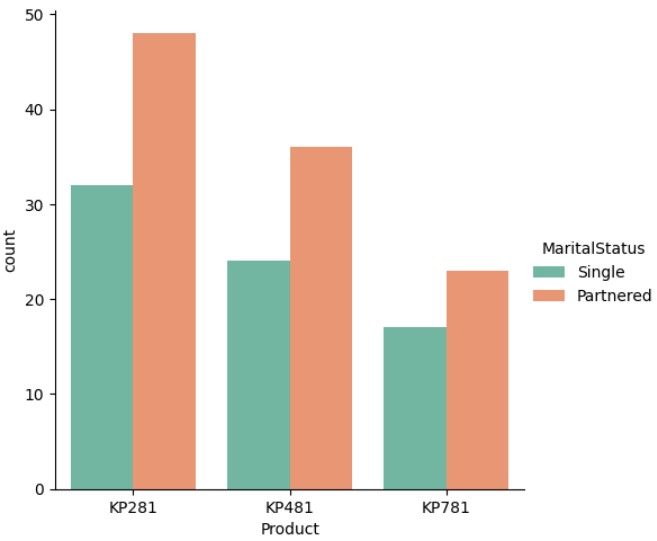
Description automatically generated

Most customers who buy the KP781 are much younger than those who buy the other treadmills. The interquartile range for the other two treadmills are almost the same between 23 and 33.

**PRODUCT VS MARITAL STATUS**

These are two categorical columns so we can compare the two using multiple columns in a count graph.

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is count
* x is Product
* hue is MaritalStatus
* palette is the pre-existing “Set2”



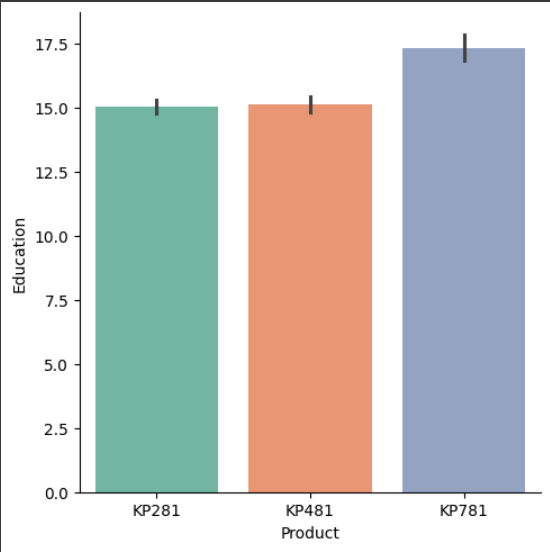
The number of partnered customers is higher in each product type.

The number of partnered and single customers both decrease from KP281 to KP781

**PRODUCT VS EDUCATION**

One column is categorical, and one is numerical so we can plot the two on a bar graph which the categorical column being on the x axis and the numerical column on the y axis**.**

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is bar
* x is Product
* y is Education
* palette is a pre-existing palette called “Set2”

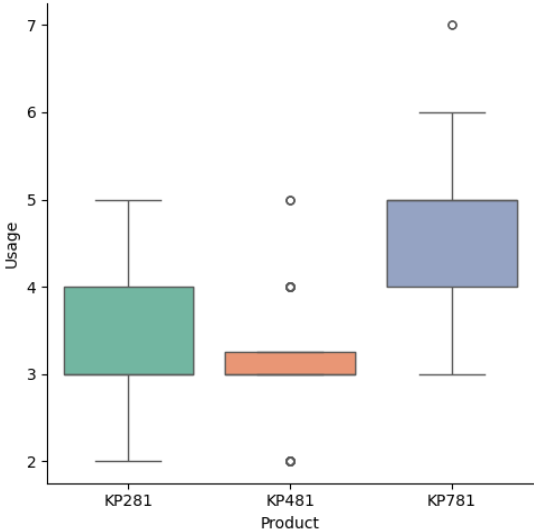
****

The mean education (default measure in bar plot) is 15 for both KP281 and KP481 while it is around 17 for KP781. So, on average customers buying KP781 are more educated.

**PRODUCT VS USAGE**

One column is categorical, and one is numerical with distinct range so we can plot the two on a box graph which the categorical column being on the x axis and the numerical column on the y axis**.**

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is box
* x is Product
* y is Usage
* palette is a pre-existing palette called “Set2”

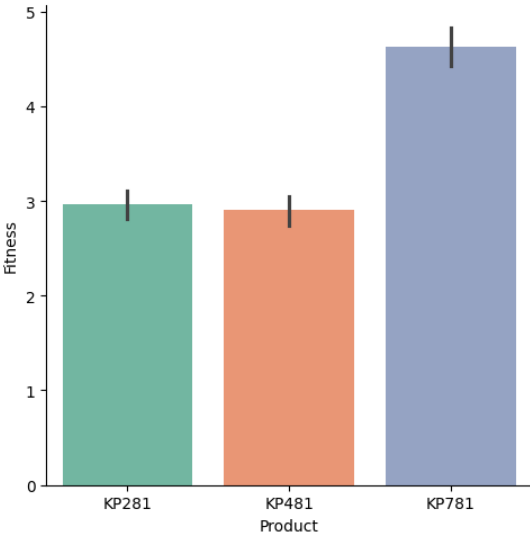
****

Most customers who buy the KP781 predict they will use the treadmill 4-5 times a week higher than customers who buy KP281 who mostly predict fitness level 3-4 times a week.

**PRODUCT VS FITNESS**

One column is categorical, and one is numerical so we can plot the two on a bar graph which the categorical column being on the x axis and the numerical column on the y axis**.**

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is bar
* x is Product
* y is Fitness
* palette is a pre-existing palette called “Set2”

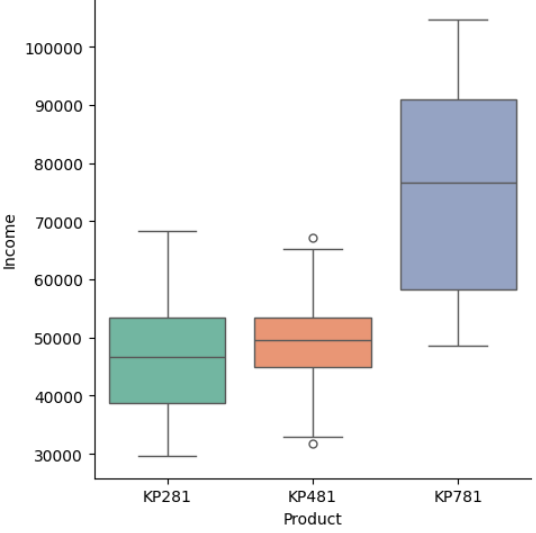
****

Customers who buy the KP781 rate their fitness level much higher than those who buy the other treadmills.

**PRODUCT VS INCOME**

One column is categorical, and one is continuous numerical so we can plot the two on a box graph which the categorical column being on the x axis and the numerical column on the y axis**.**

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is box
* x is Product
* y is Income
* palette is a pre-existing palette called “Set2”

****

The median income of people buying the KP781 is nearly double those buying the other two treadmills.

Nearly all of the highest income customers buy the KP781

**PRODUCT VS MILES**

One column is categorical, and one is continuous numerical so we can plot the two on a box graph which the categorical column being on the x axis and the numerical column on the y axis**.**

* We use seaborn’s catplot function
* Data is treadmil\_df
* Kind is box
* x is Product
* y is Miles
* palette is a pre-existing palette called “Set2”

**A graph of a bar chart

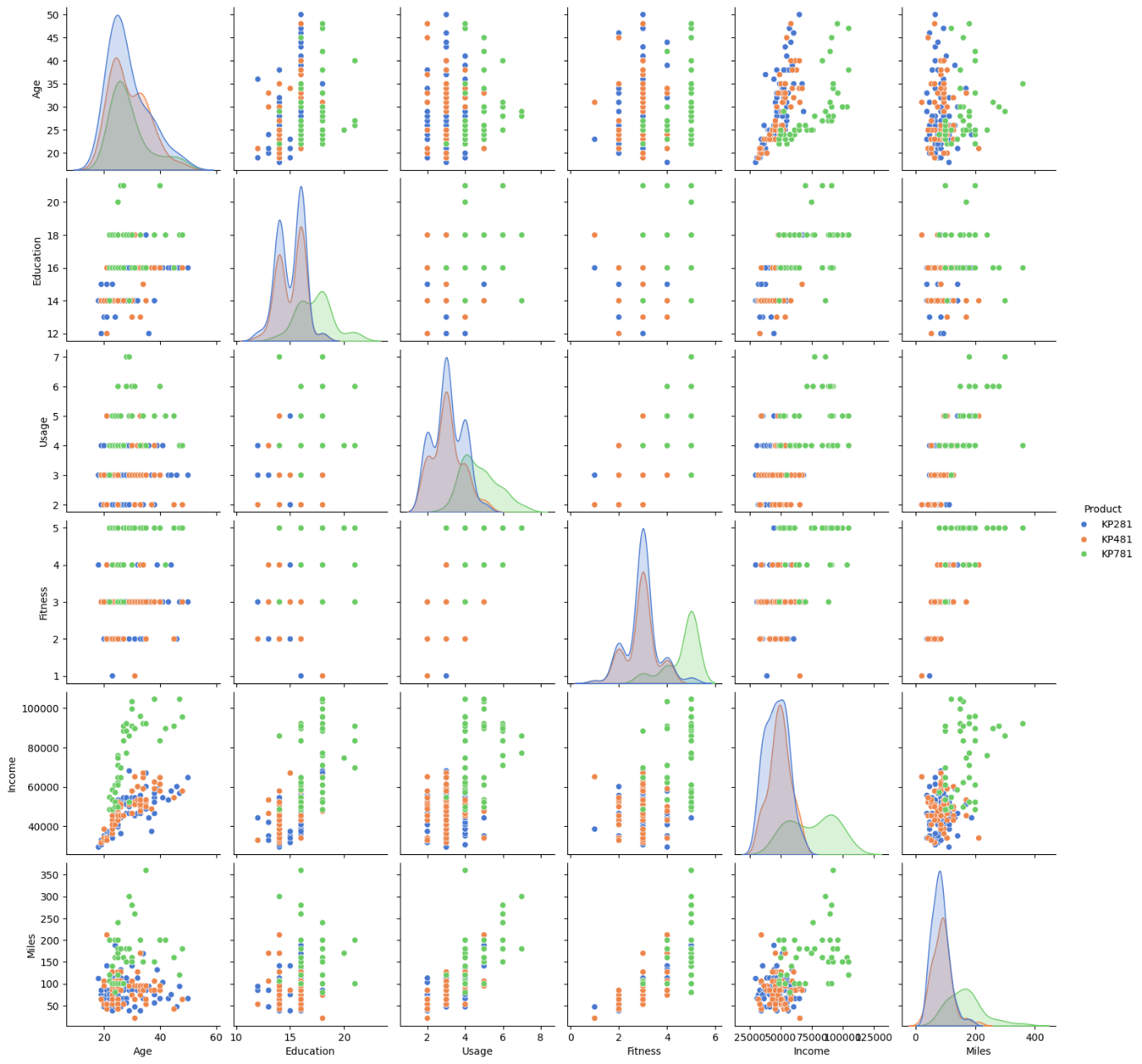
Description automatically generated with medium confidence**

The interquartile range of miles predicted to be covered in a week are higher for customers purchasing the KP781.

The interquartile ranges for customers purchasing the other two treadmills are between 65 to 100 miles per week with some higher values being categorized as outliers.

**• Multivariate Analysis**

* **Create Pairplot to show relationship of features**



Creates Pairplots of all numerical columns with hue set for the product type to see the effect on each variable.



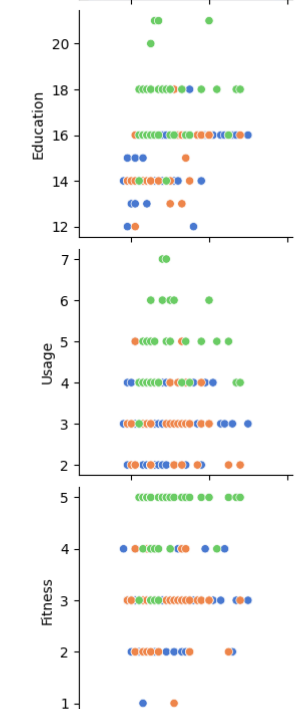
**Diagonal reading**

The diagonal shows histograms for each numerical column.

Most follow a gaussian distribution with one or two peaks

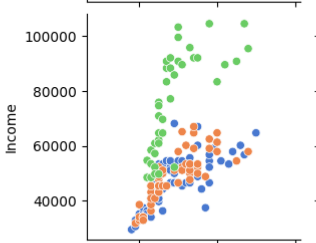
The customers buying KP781 find their peak in the higher values of each distribution.

**Age: Education, Usage, Fitness**



Age has no correlation with the Education, Usage or Fitness as seen here as people in the same age range are spread in all fitness, usage and education categories.

**Age: Income**

A group of text with numbers

Description automatically generated with medium confidence

Income and Age have a somewhat positive correlation as seen in the scatter plot. As age increases, the income generally increases.

Customers who buy the KP781 are younger and have higher income.

**Age: Miles**

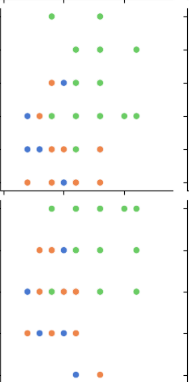
A graph with numbers and dots

Description automatically generatedA group of text with numbers

Description automatically generated with medium confidence

No clear correlation between the age and number of miles per week as most of it is concentrated in the lower part of the graph.

**Education: Usage, Fitness**



Education has no correlation with Usage or Fitness as seen here as people in the same range are spread in all fitness and usage categories.

**Education: Income**

**A screen shot of a chart

Description automatically generated**

A somewhat positive correlation between education (x axis) and income (y axis). As education level increases, the income generally increases with most people with higher income and education buying the KP781.

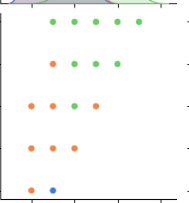
**Education: Miles**

A graph with dots and numbers

Description automatically generated

No clear correlation between the education and number of miles per week.

**Usage: Fitness**

****

There is a somewhat positive correlation between usage (x) and fitness (y) as seen here.

Customers who predict they will use the product more times per week generally rate their fitness higher as well.

**Usage: Income**

**A graph of dots and lines

Description automatically generated with medium confidence**

There is no clear correlation between the two as people who predict to use the product between 2-5 times a week are spread out on all income levels.

The only correlation is for customers who predict they will use the product 6-7 times a week which are exclusively higher income and purchase the KP781.

**Usage: Miles**

**A graph with dots and numbers

Description automatically generated**

There is a strong positive correlation between usage (x) and miles (y).

This is an expected result as time (usage) and distance (miles) are scientifically correlated variables.

**Fitness: Income**

**A graph of different colored dots

Description automatically generated**

There is a somewhat positive correlation between fitness(x) and income(y).

Higher fitness levels generally have higher income and are exclusively buying KP781.

For lower and middle fitness levels, incomes are also generally decreasing but the type of product they buy are mixed.

**Fitness: Miles**

**A graph of a graph with dots

Description automatically generated with medium confidence**

There is a strong positive correlation between fitness (x) and miles (y).

This is also expected as people who rate their fitness to higher tend to also rate the miles they will cover to be higher and vice versa.

**Income: Miles**

**A graph with many dots

Description automatically generated with medium confidence**

No clear correlation between the income and number of miles per week as most of it is concentrated in the lower part of the graph.

*Notice we only see correlation for a pair of variables once as correlation is same either way.*

**5. Correlation Analysis**

**Show the correlation matrix on heatmap and write your observation of findings in few lines.**



Using seaborn’s heatmap function with the first argument being the correlation function applied on only the numerical columns in the dataframe.

The second argument specifies a cmap using a pre-existing colour range from matplotlib called “GnBu” that transitions from green (lowest) and blue (highest) in terms of correlation.

The third argument sets annot to True to show the correlation values as well.



The heatmap reiterates all the correlation analysis we found earlier in pairplot confirming our results.

**6. Outlier Detection**

* **Check for the outliers by using the IQR method**.

To find IQR, subtract the 75th percentile of each column from its 25th percentile.

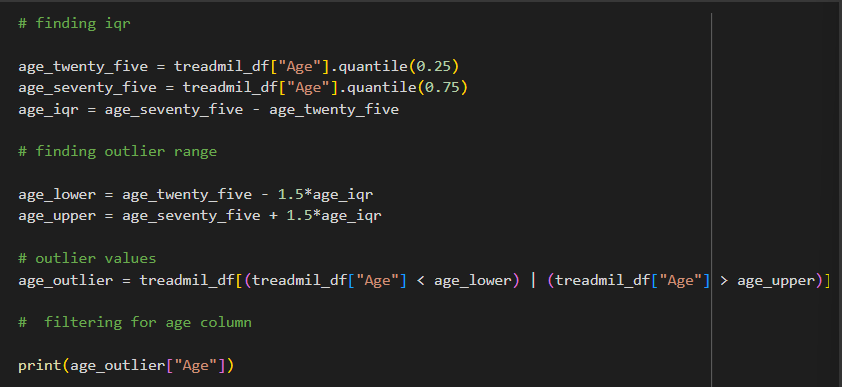
Lower outlier value = 25th percentile – 1.5\* IQR

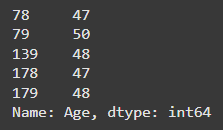
Upper outlier value = 75th percentile + 1.5\* IQR

To find outliers in dataset , we subset for values below the lower value or above the upper value.

This is repeated for all columns below.

**AGE**

****

****

Older customers (>= 47) are outliers. 5 such entries.

This shows that most customers are young or middle aged.

**EDUCATION**

**A computer screen shot of a program code

Description automatically generated**

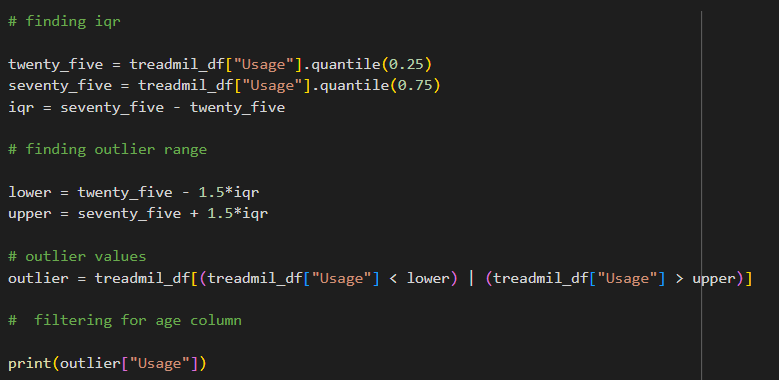
**A screen shot of a computer

Description automatically generated**

Customers educated (>=20) years are outliers. 4 such entries.

This shows that most customers are educated between 12 and 18 years.

**USAGE**

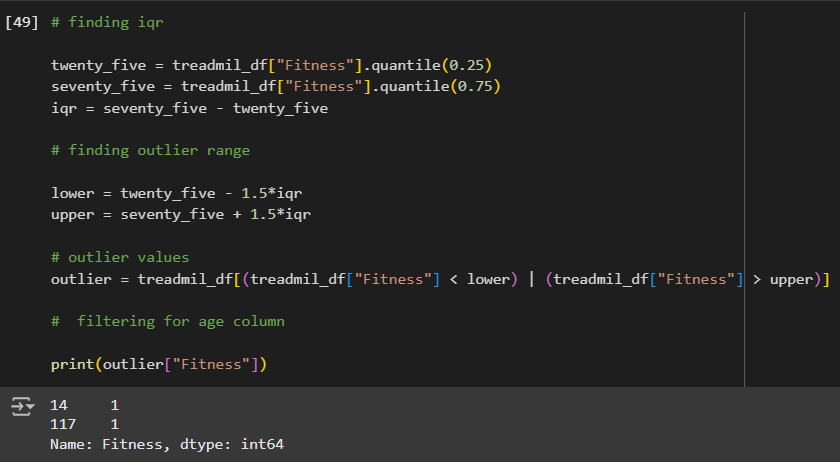
****

**A screenshot of a computer

Description automatically generated**

People who predicted they would use the treadmill nearly every day a week (6-7) are outliers. 9 such entries.

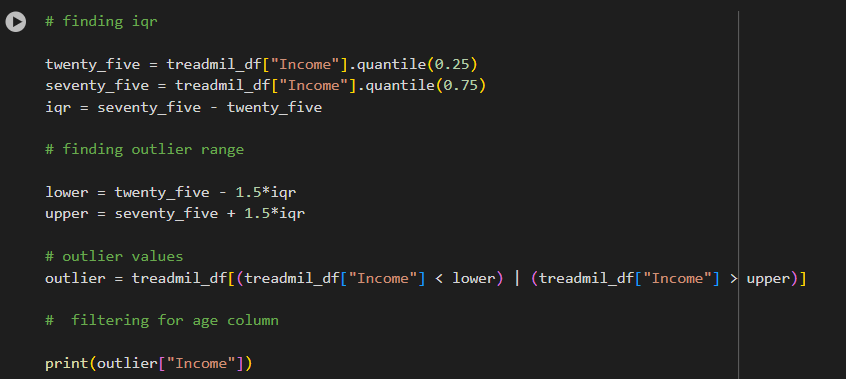
**FITNESS**

****

Interestingly, while people who said they would use the treadmills very often (6-7) were outliers, here customers rated their fitness the lowest were outliers (1).

Very few outliers here, 2 entries.

**INCOME**

****

**A screenshot of a computer screen

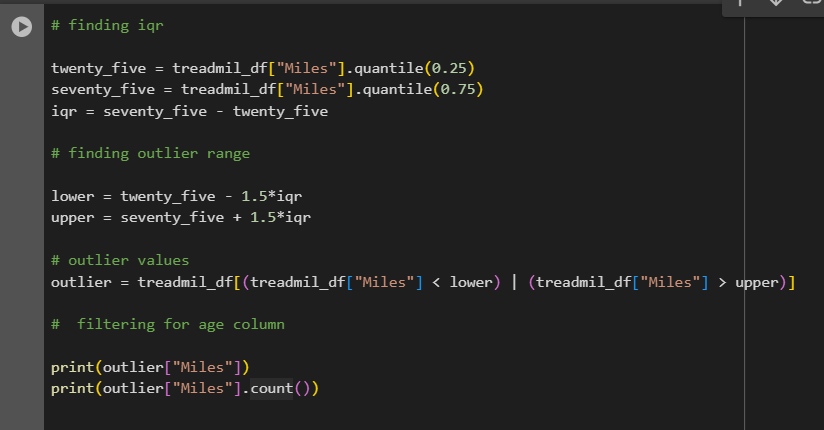
Description automatically generated**

Very high-income customers are outliers ( > 80581.875 {upper} )

Most number of outliers here: 19.

Most customers are lower or middle income.

**MILES**

****

**A screenshot of a computer

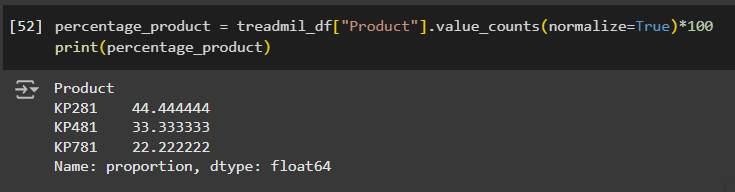
Description automatically generated**

Similar to usage, people who predicted they would cover very high number of miles per week are outliers.

High number of outliers: 13

**7. Conditional Probabilities**

**What percent of customers have purchased KP281, KP481, or KP781?**



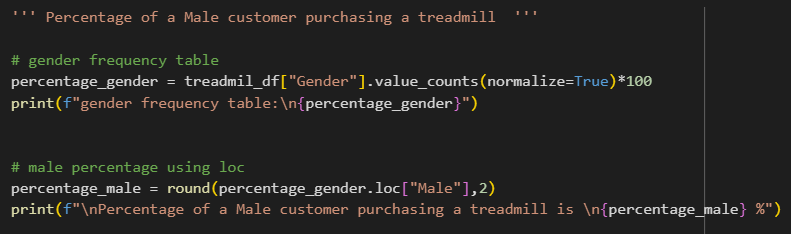
* We use value count function to create a frequency table using the column “Product”
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.

Here, number of purchases decreases from KP281 to KP781.

**Create frequency tables and calculate the percentage as follows:**

**Product – Gender**

* **Percentage of a Male customer purchasing a treadmill**



* We use value count function to create a frequency table using the column “Gender”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for male customers using loc and adding the string name “Male”.

**Result:**

A screen shot of a computer

Description automatically generated

* **Percentage of a Female customer purchasing KP781 treadmill**

A screen shot of a computer code

Description automatically generated

* We use value count function to create a frequency table using the columns “Gender” and “Product”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for female customers buying KP781 using loc and adding the string name “Male” and “KP781”.

**Result:**

A screenshot of a computer

Description automatically generated

* **Probability of a customer being a Female given that Product is KP281**

A screen shot of a computer program

Description automatically generated

* Here, we subset the dataset to only get rows where “Product” column contains KP281.
* We use value count function to create a frequency table using the column “Gender”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for male customers using loc and adding the string name “Female”.
* So, here we see the gender division for people purchasing the particular product.

A screenshot of a computer

Description automatically generated

**Product – Age**

* **Percentage of customers with age between 20s and 30s in customers.**

A computer screen shot of text

Description automatically generated

* Here, we created a new column (Age\_categories) using the lambda function to create 3 categories for customers (<20, between 20-39 , > 39 )
* Using the apply function to connect to the “Age” column in the dataset.

A computer screen shot of a black background

Description automatically generated

* We use value count function to create a frequency table using the column “Age\_categories”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for people in their 20s and 30s

using loc and adding the string name “20-39”.

**Result**

A screenshot of a computer

Description automatically generated

**Product – Income**

A computer screen with colorful text

Description automatically generated

* Here, we created a new column (Income\_categories) using the lambda function to create 3 categories for customers with (low income, medium income, high income) using the quantile function to divide the Income values in thirds.
* Using the apply function to connect to the “Income” column in the dataset.
* **Percentage of a low-income customer purchasing a treadmill**

A screen shot of a computer code

Description automatically generated

* We use value count function to create a frequency table using the column “Income\_categories”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for low income customers using loc and adding the string name “low income” to get percentage of people classified as low income purchasing a treadmill.

**Result**

A screen shot of a computer

Description automatically generated

* **Percentage of a high-income customer purchasing KP781 treadmill**

A screen shot of a computer program

Description automatically generated

* We use value count function to create a frequency table using the columns “Income\_categories” and “Product”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for high income customers buying KP781 using loc and adding the string name “high income” and “KP781”.

A screenshot of a computer

Description automatically generated

* **Percentage of customer with high-income salary buying treadmill given that Product is KP781**

A screen shot of a computer code

Description automatically generated

* Here, we subset the dataset to only get rows where “Product” column contains KP781.
* We use value count function to create a frequency table using the column “Income\_categoires”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for high income customers using loc and adding the string name “high income”.
* So, here we see the income division for people purchasing the particular product.

**Result**

A black background with white text

Description automatically generated

**Product – Fitness**

* **Percentage of customers that have fitness level 5**

A computer screen shot of a program

Description automatically generated

* We use value count function to create a frequency table using the column “Fitness”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for level 5 customers using loc and adding the int 5.

A screenshot of a computer

Description automatically generated

* **Percentage of a customer with Fitness Level 5 purchasing KP781 treadmill**

A screen shot of a computer program

Description automatically generated

* Here, we subset the dataset to only get rows where “Fitness” column contains 5.
* We use value count function to create a frequency table using the column “Product”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for KP781 using loc and adding the string name “KP781”.
* So, here we see the product division for people at a particular fitness level.

**Result**

A screen shot of a computer

Description automatically generated

**Product – Marital Status**

* **Percentage of a customers who are partnered using treadmills**

A computer screen shot of text

Description automatically generated

* We use value count function to create a frequency table using the column “MaritalStatus”.
* using the argument “normalize = True” to get each value as a fraction of total values
* then multiplying by 100 to convert fraction to percentage.
* Then, we filter out the percentage for partnered customers using loc and adding the string name “Partnered” .

A screenshot of a computer

Description automatically generated

**8. Actionable Insights & Recommendations**

**Insight 1:**

Our customer demographic is skewed towards young adults in their 20s and 30s.

A screenshot of a computer

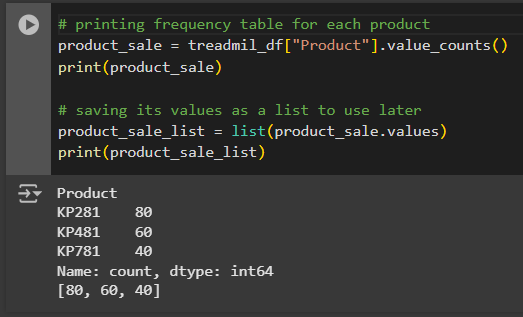
Description automatically generated

**Recommendation 1:**

* Focus marketing through social media platforms that reach this demographic ( TikTok, Instagram).
* Partner with fitness influencers on these platforms to promote the products.

**Insight 2:**

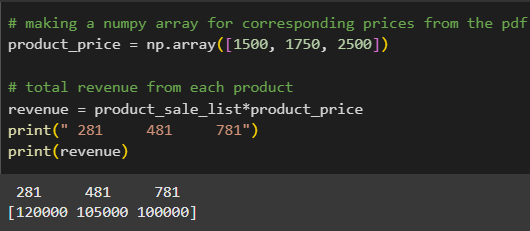
The KP281 is the best-selling treadmill type.



**Recommendation 2:**

* Ensure availability of KP281 in the inventory to meet high demand of customers.
* To keep these customers, provide points and discounts for them to upgrade to the other more expensive models as their usage and fitness level increases. This helps retain customers for a one-time sale product like a treadmill.

Finding revenue = sale\*price for each product:



**Insight 3:**

The KP781 generates high revenue, despite fewer sales, indicating its higher price point appeals to a smaller but important segment of the market.

Defining this segment:

* Gender

A grey background with white text

Description automatically generated

Predominantly male customers purchase the KP781

* Income

A black background with white text

Description automatically generated

Nearly all of the highest income customers buy the KP781

* Fitness

A screen shot of a computer

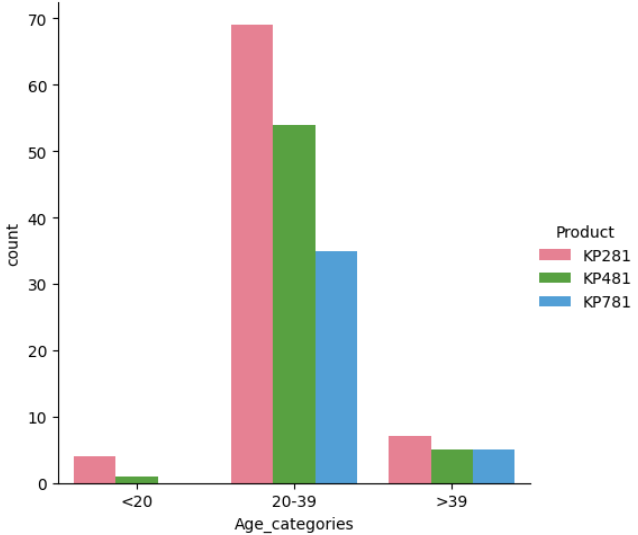
Description automatically generated

Nearly all of the fittest customers buy the KP781.

**Recommendation 3:**

* Increase sales by targeting male customers with messages around performance, endurance, and speed.
* Partnering with luxury gyms and training studios to present the treadmill as a choice for professionals.

**Insight 4:**

****

Older customers prefer the KP281 to the other treadmills.

Defining this segment further:

* Fitness

A graph with a number

Description automatically generated with medium confidence A row of purple rectangles

Description automatically generated

Most common fitness level is 3

**Recommendation 4:**

To expand the customer demographic to older customers, focus on marketing the KP281 as a user-friendly, low maintenance and cost-friendly product for beginners looking to undertake mild exercise in later years.