Index

Result of analysis with K-means (report in English).. 2

Data processing.. 2

K-means clustering...3

Correlation between variables and their matrix..6

Conclusion...7

References used.. 7

Result of analysis with K-means (report in English)

It is important to emphasize that the observations that will be made in this document always refer in the **first instance** to the data obtained from the Kaggle page *"Mall Customer Segmentation Data"* of the user Yanyan Wu. That said, any external source that helps understand the data will be referenced in the **last heading** so that the reader can corroborate its veracity.

The meaning of the columns in the table is:

- **"CustomerID"**: Customer number in our database.
- **"Gender"**: Biological sex of the client.
- **"Age"**: Customer age.
- **"Annual Income (k$)"**: Client's annual salary based on one thousand dollars ($1k = $1000).
- **"Spending Score (1-100)"**: It is a metric that helps companies segment customers based on their spending patterns and purchasing behavior where values close to 1 indicate very low spending behavior, and 100 indicates very high spending behavior..

Data processing

Before processing the data, we must **eliminate or adjust the columns** that may cause us problems when it comes to understanding the **trend of the group** that buys in this market. A column that is **not useful** for our analysis is *"CustomerID"*, although it is useful. To know how many clients are being taken into account in the database (which is 200 clients), within a deeper analysis, we can not take the column into account, we can see this as when the first one arrives on a plane: you must occupy Your corresponding position and arriving earlier will not influence the time you take off from the runway (be sure to bring your phone or tablet to entertain yourself while the rest board).

On the other hand, the *"Gender"* column specifies the **sex of the client** as Male and Female, on the page I created their classification is shown with these words, but this representation of the column is not useful in the analysis, since we will obtain an error when processing this type of data and not numbers like the rest. This is why we use a data science technique known as "One Hot" **coding**, which transforms the data into numerical characters that are simpler to understand using the method we **apply with our artificial intelligence**. The question we ask ourselves is: Is this row of our data table is of type "Female"? If it is true, we will replace the word with the number 1, otherwise it is zero 0.

Since we want to compare variables that are on **different scales** (such as years and thousands of dollars or $k\$$ of annual income), we will use a **mathematical standardization method** known as "Z-score normalization", its main objective is to center the data around the mean and scale them according to their standard deviation, so that all values have a mean of 0 and a standard deviation of 1, it could be said in a very simplified way that standardization takes the mean as a reference and adjusts the data so that they present a lower rate of relative variability around the mean, which makes it easier for the algorithm to not only process the data, but also more easily visualize outliers. Its mathematical formula is:

$$Z = \frac{x - \mu}{\sigma}$$

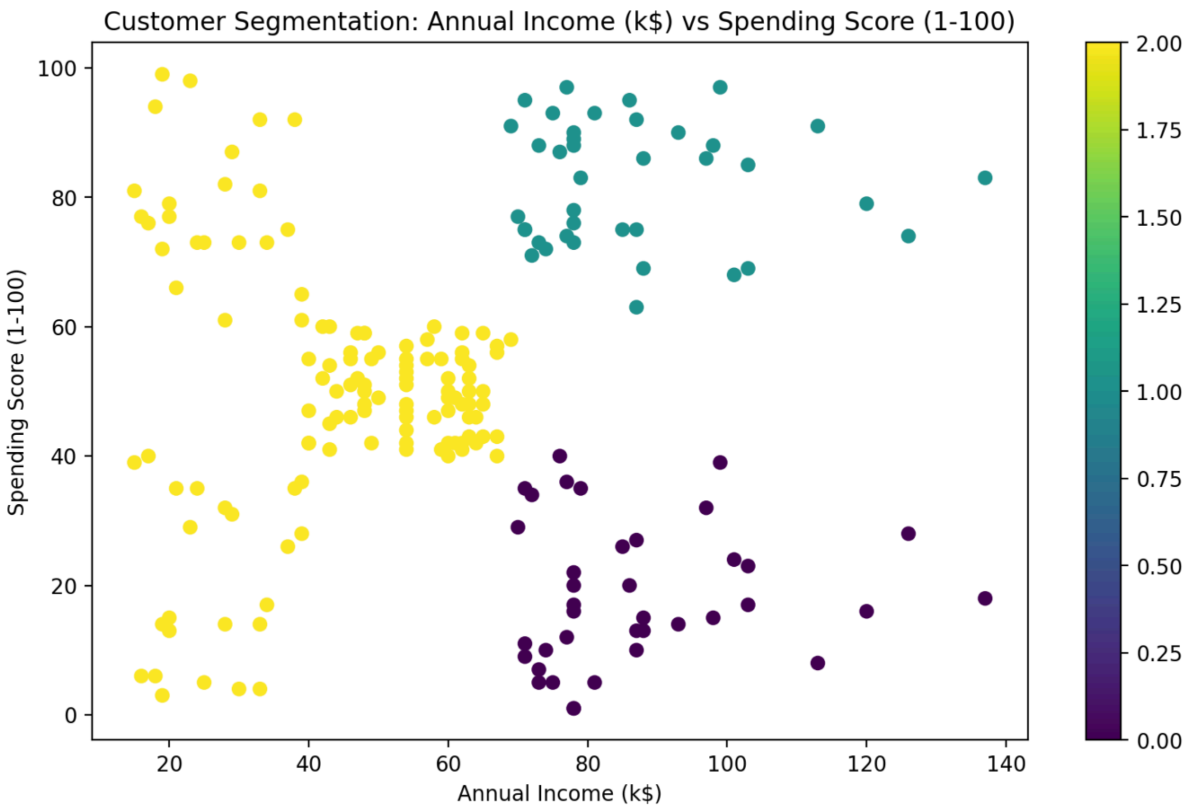
Where the variable (x) is the value present in the table to be standardized, (μ) is the mean of the variable or characteristic, (σ) is the standard deviation of the variable or characteristic, finally (Z) is the Z-score that tells us how many standard deviations a data is from our set (having a data with $Z = 2$ is equivalent to saying that it deviates from the mean by 2 times the standard deviation[1]).

With this data preprocessing, we ensure that we meet the requirements for an analysis that is easier to interpret and process.

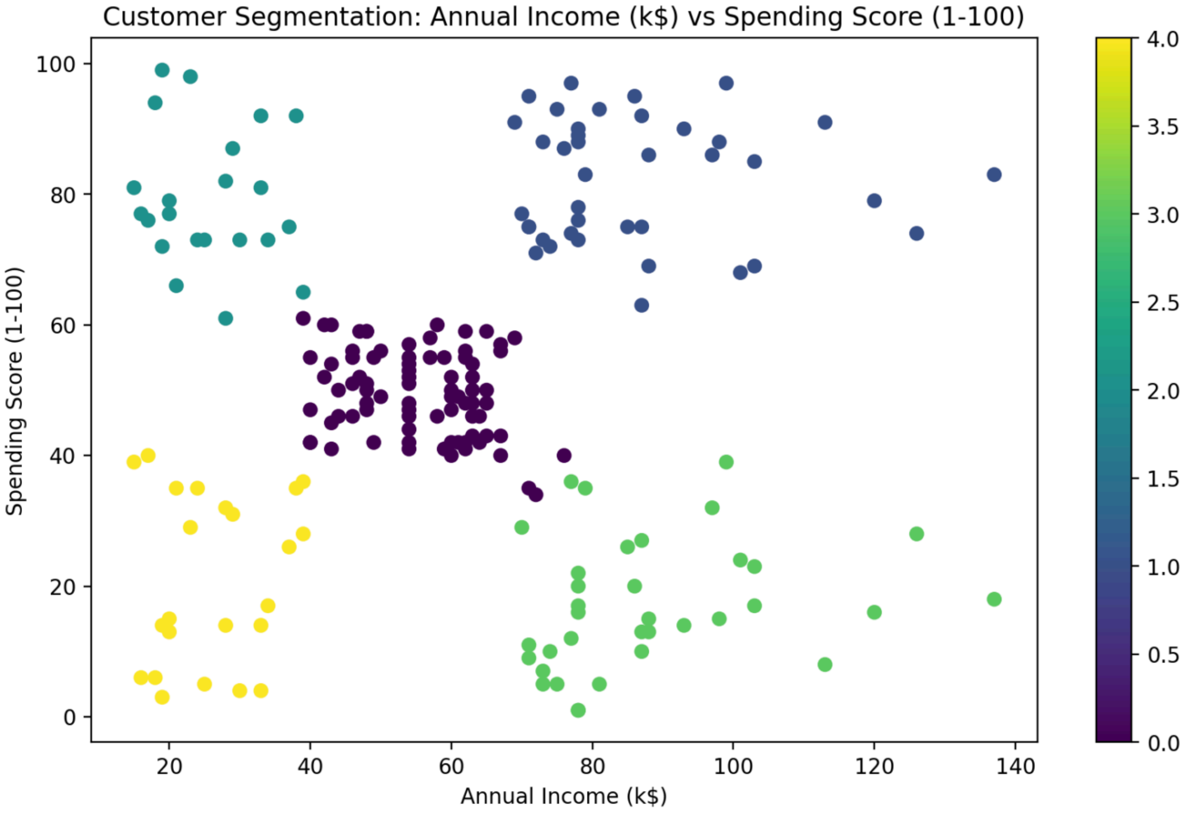
K-means clustering

I chose this **unsupervised machine learning technique of data science** because it **helps us visualize** how to group data sets by their common relationships that are less visible to a human, therefore, it allows us to better see the data that **share characteristics** with other **neighboring points**. (k-nearest neighbors or K-means). When applying this method with the number of groups at 3 ("Number of Clusters"), which is the default configuration, and we only look at the annual income ("Annual Income") and spending points ("Spending Score"), we find this graph:

[1] I recommend researching these terms so as not to lengthen the report.



As we can see in the graph, three groups are represented that are grouped based on colors to indicate the common behavior between them, however, the yellow group is very generalized, so we will increase the number of groups to 5 (“Number of Clusters"), with this we can obtain more categorized groups:



Excellent, now it is much clearer the categories that we have in this analysis based on these variables, a view of the graph shows how we have an average group in purple that seems to concentrate the majority of our clients, which encompasses an annual income between 40 and almost 80 thousand dollars annually with spending points between 40 and 60, these are very important clients, since they not only buy at a stable average, they also, on average, have income that allows this activity to be sustainable over time (it is recommended to have take into account the minimum wage before drawing conclusions), in this document we will discuss the minimum wage by law at $7.25 per hour or $63,510, which we will represent approximately as 63k$ per year[2], although in other states and on federal land it may be different , many states have this minimum wage.

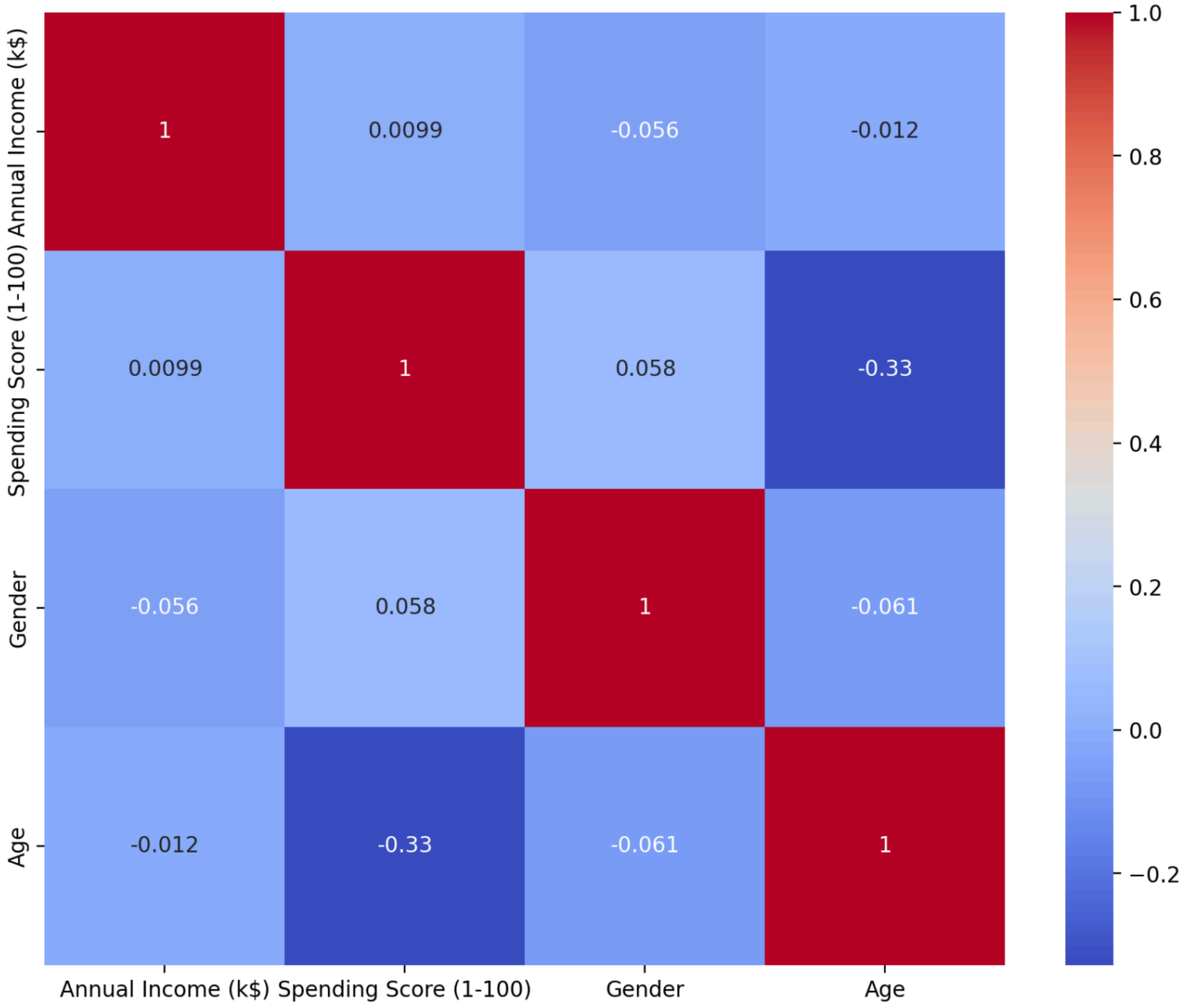
Analyzing the rest of the groups, the yellow group (bottom left) has a sustainable trend that adjusts in terms of how much they buy and earn annually, unlike the turquoise group (top left) who buy too much without taking into account their annual salary, which results in an unstable trend. On the other hand, it has managed to attract the attention of buyers who earn above 70k$, although we have to do a deeper investigation in the green group (below right) to observe the products they choose and their corresponding price in a weekly trend or monthly (which will allow us to earn more money by offering them the best products at the time they need it most), this is easy to understand if we think about a more obvious case, if we have a product that is purchased in one month of the year, we prepare not just a larger volume of these products, we also think about creating new variations of the product to see if we can have a new product that competes with the standard (preferably

[2] The references for this assumptions are 2, 3 and 4 in the last header “References used”.

better for customers and therefore they buy it more), in case where there is lower demand, the supply of the product and its variations will decrease.

Correlation between variables and their matrix

It is important to note that knowing the correlation or influence of one variable on another is crucial for deducing the behavior of the population we are studying. For this, we will use what we call a correlation matrix, which is a mathematical tool where each value in The matrix represents the correlation coefficient between two specific variables, this value is the linear relationship between variables. If we take the correlation matrix for our variables, we will obtain the following graph:



As we can see, it is divided into rows (horizontal) and columns (vertical), if we see a row marked by one of the variables, we can find the relationship with another variable seen in a column, for example: if we see the relationship between them variables (same row and column number), we will see that the relationship is 1, it makes sense because it makes sense with itself.

On the other hand, if it approaches zero or is far from 1 and -1, it means that it has no relationship with the variable. If we look at the matrix, almost all of them are zero, so the group we are studying does not influence too much between them.

Here there is a negative correlation between age (“Age”) and spending points (“Spending Score (1-100)”), the linear correlation is very weak but notable, which shows us that older people tend to spend less when buying.[3]

Conclusion

In conclusion, the analysis presented provides a detailed view of customer behavior based on their age, gender, annual income, and spending patterns. By preprocessing the data appropriately and applying clustering and correlation analysis techniques, it has been possible to identify specific groups of customers and their purchasing tendencies. These findings not only facilitate customer segmentation for marketing purposes, but also allow for a better understanding of consumption dynamics within the market.

Based on these results, companies can design more effective marketing strategies, optimize product offerings and improve customer satisfaction, better aligning their services with the needs and behaviors observed in each customer group. Therefore, this analysis is a valuable tool for making informed and strategic decisions in the commercial field.

References used

This document uses resources from:

1. “Mall Customer Segmentation Data” | Yanyan Wu | Kaggle.
2. “What are the annual earnings for a full-time minimum wage worker?” | Center for Poverty and Inequality Research.
3. “State Minimum Wage Laws” | U.S. Department of Labor | State Minimum Wage Laws.
4. “Minimum Wage” | U.S. Department of Labor | Minimum Wage.

[3] Remember: a weak correlation.