Data cleaning and transformation

Contents

Introduction

Data analysis is a process that involves collecting, organizing, and interpreting data to answer questions or solve problems. One of the tools that can be used for data analysis is Excel, a spreadsheet software that allows users to manipulate and visualize data in various ways. In this project, we will demonstrate how to use Excel to perform some common data analysis tasks, such as cleaning data, transforming data, and creating dashboards.

Cleaning data is the first step of data analysis, and it involves removing or correcting any errors, inconsistencies, or missing values in the data set. This can improve the quality and accuracy of the data and make it easier to work with.

Transforming data is the next step of data analysis, and it involves changing the structure or format of the data to make it more suitable for analysis. This can involve creating new variables, aggregating, or summarizing data, splitting, or combining data, or reshaping data from wide to long format or vice versa.

Creating dashboards is the final step of data analysis, and it involves presenting and communicating the results of the analysis in a clear and concise way. Dashboards are interactive reports that display key metrics and trends using charts, tables, slicers, timelines, or other visual elements and can help users to monitor performance, identify patterns or anomalies, compare scenarios, or make decisions based on data.

Duplicate Removal

The first step is to remove any duplicate records from the data set. Duplicate records are rows that have the same values for all or some of the variables. They can happen due to data entry errors, merging of different sources, or other reasons. Such records can affect the quality of the data and lead to wrong calculations and interpretations. Therefore, we need to identify and delete them before proceeding with the analysis. This will ensure that our data is accurate and consistent.

Columns

We want to make some changes to the dataset to make it more readable and useful. The first column, Id, will remain unchanged because it is important to have a unique identifier for each row.

The second column, Marital Status, will be modified by replacing M and S with Married and Single, respectively. This can be done by using the find and replace function (ctrl + H). This will make it easier for the user to understand and use this column.

The third column, Gender, will also be modified in a similar way, by replacing M and F with Male and Female, respectively.

The fourth column, Income, will be formatted as currency instead of general, to show the exact amount of income for each row.

The fifth column, Age, will be grouped into categories based on age ranges. We will use an IF formula to assign each row a label of Adolescent, Middle Age or Senior, depending on whether the age is less than 31, between 31 and 55, or greater than 55. This will help us create better visualizations without having too many age values.

Pivot Table Creation

Table 1:

To create a summary of our data, we used a pivot table to display the average income of different groups of customers. We grouped them by gender (male or female) and by whether they bought a bike or not (yes or no). This way, we could see how these factors affected the income level of our customers.

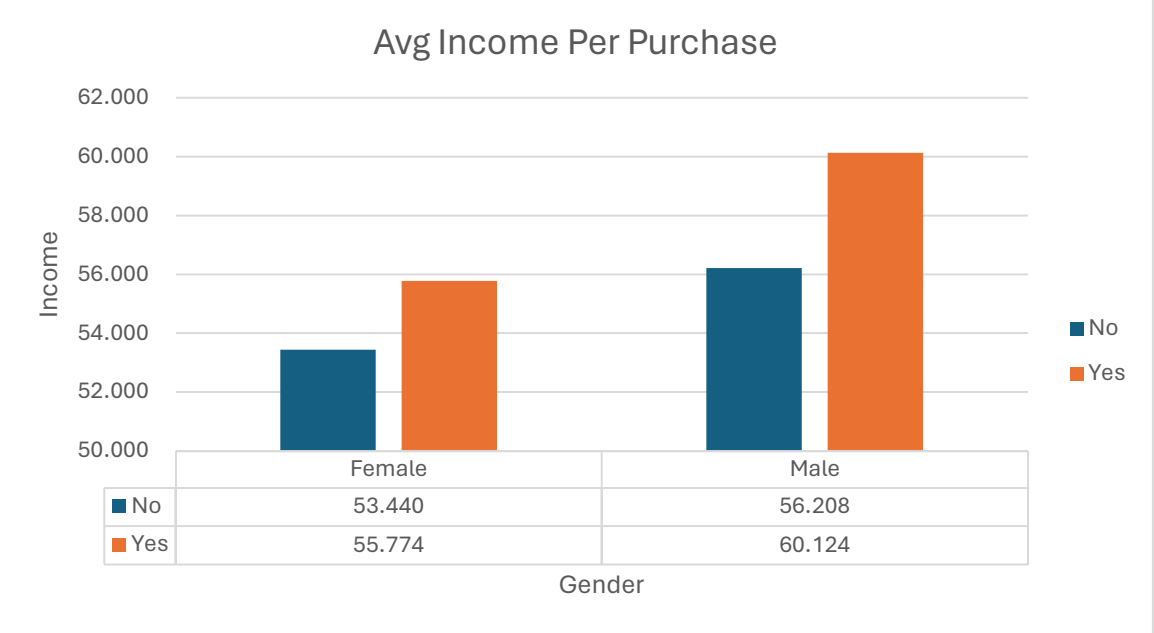
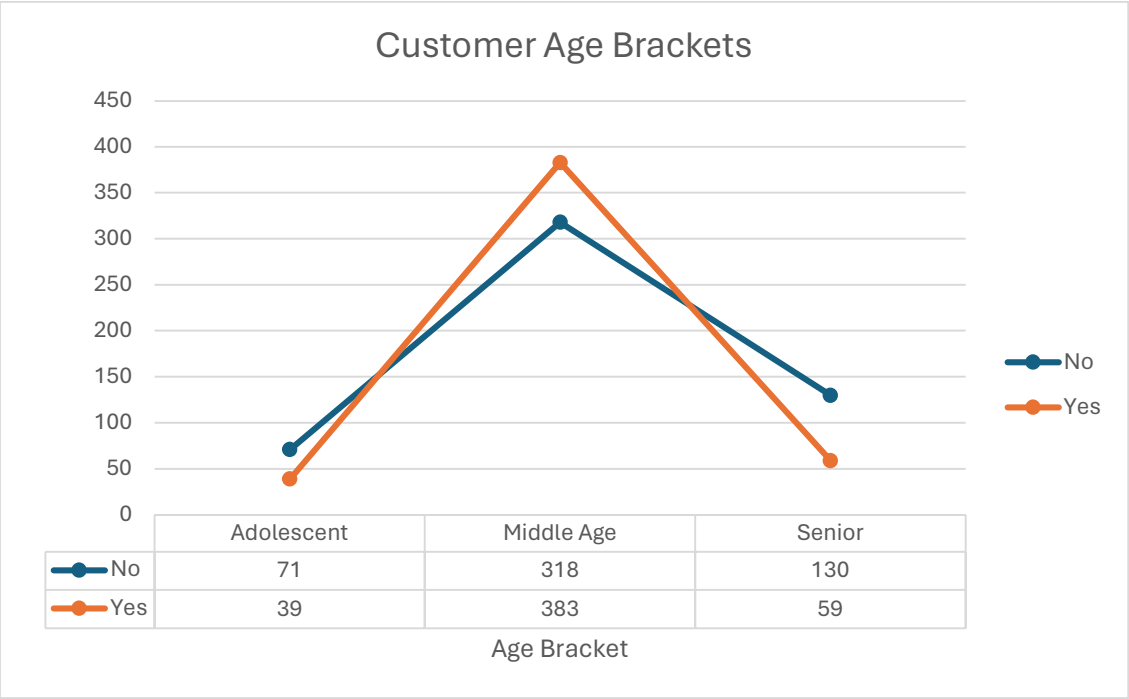


Table 2:

One of the factors that can influence the choice of a bike is the commuting distance. People who need to travel long distances may prefer bikes that are comfortable, fast and fuel-efficient. On the other hand, people who use bikes for short trips may opt for bikes that are easy to manoeuvre, affordable and eco-friendly. Therefore, it is important to know the average commuting distance of bike buyers and how it affects their preferences and satisfaction.

Table 3:

To analyse the relationship between age and bike purchase, we can use a pivot table that summarizes the data by the age groups we created before. The pivot table will show the count of customers who bought a bike and those who did not, for each age group. This way, we can see how the bike purchase behaviour varies across different age ranges.



Dashboard

The dashboard was improved by arranging the pivot tables next to each other and removing the gridlines for a cleaner look. To allow for better visualization and analysis, slicers were inserted based on region, marital status, and education. The data reveals some interesting insights.

Observations

According to our data analysis, most of our customers belong to the middle age group, while the younger and older age groups are less represented. This suggests that our products and services appeal more to people in their 30s and 40s than to those in their teens or 60s and above.

Single women buy more than single men in the young and middle age groups, but this changes for the older group, where men buy much more. Married men also buy more than married women for the young and middle age groups, but not for the older group, where they buy the same amount. The main takeaway from this is that single middle-aged women are the best customers for our business.

The sales data shows that the north American region has the highest purchases of our products, followed by the European and pacific markets. This is surprising because we expected the European market to be more receptive to our ecofriendly and sustainable features, as they are known for their environmental awareness and policies. However, it seems that the north American consumers value our products more, despite their lower environmental standards and regulations. This could indicate a gap between perception and reality, or a difference in consumer preferences and behaviour.

According to our data, the most frequent users of our bikes are those who have a bachelor's degree. This is a surprising finding because they also rank second in terms of income level. One might expect that the lowest-income group would rely on bikes as a cheap mode of transportation, or that the highest-income group would enjoy bikes as a luxury or a hobby, but our numbers suggest otherwise.

Conclusions

Based on the data analysis, it is clear that our products and services appeal more to people in their 30s and 40s. Single middle-aged women are the best customers for our business. Therefore, we should focus on creating marketing campaigns that target this demographic. According to, social media is the best way to reach female millennials, with Facebook and Instagram being the most popular platforms. It is important to note that we should avoid stereotyping or generalizing our marketing messages. Instead, we should narrow our marketing demographic to be as specific as possible and develop content and messages that speak to this more refined audience.

In terms of geography, the north American region has the highest purchases of our products, followed by the European and pacific markets. We should continue to focus on these regions

and tailor our marketing campaigns to the specific needs and preferences of each region. For example, we could highlight the eco-friendly and sustainable features of our products in the European market, where environmental awareness and policies are more prevalent.

Finally, we should consider targeting those who have a bachelor's degree, as they are the most frequent users of our bikes. This group ranks second in terms of income level, which suggests that they value the convenience and health benefits of biking over other modes of transportation. We could create marketing campaigns that highlight the health benefits of biking and how it can improve one's quality of life.

Overall, our sales strategies should focus on creating targeted marketing campaigns that appeal to specific demographics and regions, while highlighting the unique features and benefits of our products.