Helin Varoglu 29237

PROJECT REPORT

Motivation

My main reason for working on the project is to keep track of my monthly expenses and determine the categories I spend the most on, and to determine an appropriate savings plan in the future. Thanks to the analysis, I was able to make a more detailed analysis of where and how much I spent. My hypothesis consisted of two basic parts: The first part was that my monthly expenses constantly changed from month to month due to the intensity in my personal life, and the second part was that the first 3 categories with the most spending did not change mostly (over 50%). At the same time, visualizing all these analyzes will provide the opportunity for a more detailed analysis.

Data Source

This data was received manually through the "Garanti BBVA" banking application. The reason why it is done manually is because it is a banking application and it is a highly protected application, therefore BeautifulSoup etc. are not valid.

Data Analysis

In the first stage, general EDA and visualization techniques were applied in the "EDA.ipynb" file. Here the data was brought into a suitable dataframe form. By using these techniques, data such as the total spendings of each month, the 3 categories with the most spendings in general, and how much was spent in which categories each month were obtained. These results were visualized using bar chart, stacked bar chart and scatter plot.

In the second step, the necessary machine learning and analysis techniques were determined. Linear regression analysis was used to prove whether total spendings, which is the first part of the hypothesis, changes monthly and does not remain constant. Here, months were determined as a categorical independent variable and total monthly expenses, which were the dependent variable, were calculated. Monthly expenses were created as a table and the slope, intercept and r-squared were calculated by finding the linear equation. At the end, a model for linear regression emerged.

Chi-square distribution was used for the second part of the hypothesis, that is, the frequency of the top 3 spending categories is more than 50%. The reason for this is to be able to draw the correlation between discrete and categorical variables. The top 3 most spent categories of each month were determined, and then the frequency table of the 3 most spent categories (of total) in each month was presented. Then, the expected data were calculated using the chi-square distribution formula based on the given data. This is the expected data formula used:

Expected Count = ((row total)(column total))/table total*

And this is the formula used for chi-square:

$$X^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

Where *O* is for the observed, *E* is for the expected and k is the total number of cells.

Findings

For the EDA part, these are main conclusions that can be drawn:

- The most spent months are (in the descending order): 2023 November, 2023 March, 2023 April, 2023 September, 2023 February, 2023 August.
- As can be deduced from the table, spendings did not remain constant over the months and varied.
- The most spent categories from all of the months are (in the descending order): Subscription Services, Dining, Online Shopping.
- Likewise, these 3 categories are seen as the 3 most frequently spent categories, respectively: Subscription Services (7 times), Dining (6 times), Online Shopping (6 times).

From this section, I made two general conclusions about myself. First of all, I usually spend the most around holidays or special occasions (New Year's Eve, my relatives' birthdays, etc.). This was something I expected, but what I didn't expect was that the category I spent the most on was subscription services. I expected the remaining two to fall into the top 3 most spent categories, but the third one was unexpected for me to receive subscription services category while I was waiting for groceries category.

For the testing part, I argued that my spendings did not remain constant and varied from month to month, as the first half of my hypothesis suggested. This hypothesis was confirmed by the test we conducted. It was observed that when the r-squared value was low, the determination was low, meaning there was no stability in the data. Likewise, the high standard deviation and the fact that the data were far from a linear equation in the model shown proved that the spendings were not constant over the months. In this way, the first part of the thesis was confirmed.

As a result of the frequency tables and probability calculations made for the second part of the hypothesis, it was proven that the frequencies of 3 categories that spent the most in total were more than 50%. However, due to the small number of data and the expected values being zero, the alpha value remained smaller than the p-value, causing the null hypothesis to cannot be rejected. Although it did not state that the second part of this hypothesis was wrong, it did state that it needed to be tested with more data.

Limitations and Future Work

Since the time period in which I spent the most was the last year (the end of the pandemic, the schools were fully opened, etc.), I wanted to analyze this year on the project. However, due to the scarcity of data, there was not enough evidence to reject the null hypothesis. When I do my next project, I will make sure to collect more and more comprehensive data, even if I only want to examine it for a limited time.