Rasmus Made
Kristian Tamm
Harmo Ron Kallaste

Project B10: Grocery store spending analysis

Repo link: https://github.com/kristiantamm/SJ_AT_Bilance_projekt

Assessing our situation

Our inventory of resources is pretty minimal for this project. In terms of people as a resource, there is our team - Kristian, Rasmus and Harmo and our lab instructor Markus Haug, who is also connected with the business that is behind our project. Data that we are using to develop our workflow consists of 100,000 rows and 6 attributes. The data was synthesised by our lab instructor. As far as hardware resources, we are using our own personal laptops. Main software for developing our workflow is Jupyter Notebook, where we use Python and some publicly available libraries.

The finished project needs to be presented in the form of a poster on the 11th December. After the submission, we agree that the poster will be publicly available. The synthesised data was given to us by Bilance on condition that we sign an intellectual property document, which in conclusion states that we cannot demand anything from the company.

Terminology

1. Attribute a.k.a property - Measurable or observable characteristic of a data object.
   1.1 Examples: name, age, height, weight etc.
2. Synthesised data - artificially generated information designed to mimic real-world data.
3. Workflow - systematic sequence of tasks and processes involved in collecting, cleaning, analysing, and interpreting data to derive meaningful insights.

Identifying business goals

Background

Our client, Bilance Technology OÜ provides personal finance management service for private users. They provide the client with an overview of their expenses by categorising and tracking their payments. The client connects their bank with their app and based on the payment information that is provided by the bank, Bilance provides their service. As they handle a large amount of payment data every day, they would want to use that information in a way that would help them make statistics about the different enterprises where the clients go shopping.

Business goals

We intend to provide Bilance OÜ with the statistics of the patterns in which people visit different grocery stores and with the cost of the shopping basket in different stores.

Business success criteria

We deem our venture successful if we have achieved the business and data mining goals to the greatest extent possible. Should we be unable to fulfil them based on the provided data, that would be considered an exception.

Our data-mining goals

The main goal of our project is to find out the cost of a food basket by enterprise. The cost needs to be analysed in such a way that we can report how it changes over different periods of time, to for example see how inflation changes it. We also need to report on the information of how different clients visit different stores, find out if they have defined preferences and how their preferences change over time. We also need to determine if there are some enterprises where clients go to buy larger or smaller baskets.

Data understanding

The dataset provided to us is a collection of transaction records that are described by the following columns: id, formattedDate, merchantName, remittanceInfo, value. We got access to it from Bilance Technology OÜ. Since the original dataset contains real-world information that needs to be kept safe, we got a synthesised version of it. We have information about the date of purchase, where was the purchase made, which card was used and how much was spent. However the whole dataset is synthesised so the information is obscure in some instances.

The dataset contains above-mentioned data but there are many missing values, especially in merchantName and remittanceInfo so the data is incomplete.

Our main interest is in columns formattedDate, merchantName and value. The id column doesn't really give useful information and the currency column just lets us know that we need to convert the value in that row.

The columns:

id - each transaction is represented by an unique identifier

formattedDate - shows when the transaction was made

merchantName - shows which merchant was involved with the transaction

remittanceInfo - additional info about the purchase (card number, area of a country etc)

value - shows the amount of money that was spent during that transaction

currency - shows what currency was used in that transaction

The data structure is quite consistent. The dates are in the same format, the credit card numbers have a couple different forms but there are not incomprehensive amounts of them. There are many empty cells and null values. The missing entries indicate that the provided dataset needs a lot of cleaning. In cases like a null value in the "merchantName" column we have to check if the merchant can be taken from the "remittanceInfo" entry from the same row, if not then we might get very little information from that transaction. At this moment we are not sure if the strange information in "remittanceInfo" (random nonsense words, dates etc.) is due to the fact that the dataset was synthesised but we might synthesise it another time to fill in the blanks. Some entries in the "value" column also raise the question: what is the

difference between the numbers that have the "-" sign in front of them and those that don't. We might have a dataset that includes transactions where money is used to buy something and transactions where money is given back. But it is also possible that this is just a difference between different merchants or banks that the transactions have gone through.

704	06-03-2023	null	292384******9276 either JARVE SELVERI ISETEENINDUPARNU guy game	-10.74	USD

This is a row from the synthesised dataset. It's a great example of a row that we need to work more on to get good information from it. Firstly, we would need to convert the value by making it equal in euros. Secondly the merchant info is null so we would need to extract "SELVER" from the remittance info cell and put it instead of the null value.

1493	07-10-2023	KONSUM use East Nataliehaven	14/07/1985 will kaart...396009 RAADIMOISA KONSUM/TARTU VALD/EST include	-12.4	EUR

This is also a row from the synthesised dataset. It's a great example of a row that needs very little work to find it useful. We can almost clearly see the merchant in the desired cell and the remittance info even shows us the almost exact location of the store. We have the last numbers of the credit card and the currency is already euros.

1460	15-05-2023	Nichols, Davis and Young	race	-5	EUR

This is another row from the synthesised dataset. It's a great example of a row that gives us nearly no useful information. We can't see information about the merchant and there is nothing meaningful in the remittanceInfo cell either. The only thing we could do with this row is to take the date and the -5€ into consideration if we should ever compare the total money spent in a specific window of time or a certain month.

In conclusion we have a synthesised dataset of real-world transactions that are described by a date, information about the merchant and remittance, value and the currency used. There are many consistencies along the rows but there are lots of missing and null values. This indicates that we have to spend a great amount of time preparing the dataset. By cleaning,

processing the data and possibly imputing some values we can have a great dataset to analyse - we can possibly find interesting patterns among the data and provide useful visualisations.

Project plan

Task	Description	Time (h)	Contributors
1	Converting the currencies to EUR	3h	Everybody
2	Initial categorisation by merchantName	5h	Everybody
3	Calculating average shopping carts by store	8h	Everybody
4	Analysing the change of the shopping cart price through time	8h	Everybody
5	Initial visualisation of the data	4h	Everybody
6	Payment info analysis	16h	Everybody
7	Data visualisation based on location gathered from remittanceInfo	8h	Everybody

4. This task might be one of the hardest ones as it is rather abstract. We plan to use a machine learning model to understand the trends. The model would be based on time series forecasting.

6. This is the only task about which we are not sure if it can be done as per our initial inspections the payment info might not be complete enough to determine different users.