M-IE_2.01 Data Mining, WS 2021/22

Clustering of retail data

Due: Sunday, 3 April 2022, 11:59 PM

Mark as done

Assume that you are a data analyst working for a retail company. You were given a sample of data that contains web sale transactions from 1st April to 1st May (without considering the transaction of return). (Attachment: Web_Baskets_2020.zip)

The following are the attributes of the dataset:

- **date:** date of the transaction in the format yyyy-mm-dd.
- **transaction_id:** unique identifier associated with each transaction.
- **article_id**: unique identifier of the article purchased.
- **article_cat:** category the article belonged to.
- **article_name:** short name of the article sold.
- **quantity:** number of that specific *article* sold in the given *transaction_id.*

As part of your job, you are creating transparency in the online business. To help you the management sent you the following questions:

- What kind of products are purchased the most?
- Which are the most successful categories in our online business?
- Which days of the week do we make the most/least transactions?
- How are the different categories and articles performing?

You were given the following tasks:

1. Find the five best-selling categories.

a. For this you should transform your data (group by categories), showing the sum of monthly sales per category.

b. Name the categories and visualize the respective best-selling article of each of the five categories per day in one picture.

2. You should prepare the data for clustering analysis. To cluster the articles and the categories, you will need to extract some information.

a. You should extract at least four new attributes. For example; find out the number of units sold, average units of article per day/basket/weekday, avg. additional units (of other articles) per basket, …

b. For each created variable, provide a description of what it measures.

c. Visualize at least two different distributions for at least two of the newly created variables. Try to explain what the pictures show.

d. Normalize the variables to prepare them as input for clustering and provide the R code on how to calculate the extracted variables. (Document the code such that it explains what you are doing/want to do.)

3. Use the extracted information from **2.** and find clusters of articles.

a. Try different cluster sizes and choose the best clustering.

b. Show whether (or how good) the clusters from your best clustering match the categories.

c. Explain your approach and results.

Hint: Even if the clustering doesn't work well, explain what you did and what you tested.

Create a report of 3-5 pages in which you explain your findings of the business. You will use tasks 1-3 to give the structure to your report. Additionally, you should deliver the R code with explanations (short comments). You can add a comment in R with '#' at the start of the line.

In the end, your report should have the following structure:

1. Introduction (give a brief motivation of the report and provide some summarized key figures about the data)
2. Online Best-Sellers. (visualization and description from **1.**)
3. Key performance indicators (Provide specific key figures based on quantities, explain how they are calculated, and deliver an interpretation)
4. Clustering results (Visualization and interpretation. Present your approach to clustering with the given sample.)
5. Summary / Conclusion

Web_Baskets_2020.zip 28 January 2022, 10:05 PM

Submission status

Submission status	No attempt
Grading status	Not marked
Time remaining	42 days 11 hours
Last modified	-
Submission comments	Comments (0)