REPORT ASSIGNMENT 1

CHEMINFORMATICS WINTER SEMESTER 2022-23

Stephan Liu
Universität Tübingen
stephan.liu@student.uni-tuebingen.de

October 29, 2022

ABSTRACT

KNIME, also called Konstanz Information Miner, is an open-source data analytics program with the goal to advance the impact and understanding of data science through visual programming. [1]Thus, this program makes abstract tasks easier to understand as the problem is visualized. The graphical user interface is simple and clear and makes operations fairly user friendly.

1 Introduction

The importance of data and data sets cannot be neglected in science. The goal is to process these data in such a way that noise is minimized and still representative enough to be useable. The sheer volume, variance, and dimensionality of the data available make it difficult for the individual to identify, extract and interpret them correctly, which increases the importance of data analysis and statistical learning significantly. [2] Data analytics and machine learning simplify and advance our understanding in our endeavor to work with data.

Natural to say as the amount of data increases it gets gradually harder to keep clarity over the data and harder to reprocess the data. Accordingly, there are data processing programs that are used to maintain better uniformity. Besides well-known software like Rapidminer, IBM SPSS statistics and SAP analytics cloud, there are free open-source data analytics like KNIME that provide easy access for university purposes.

2 Material and Methods

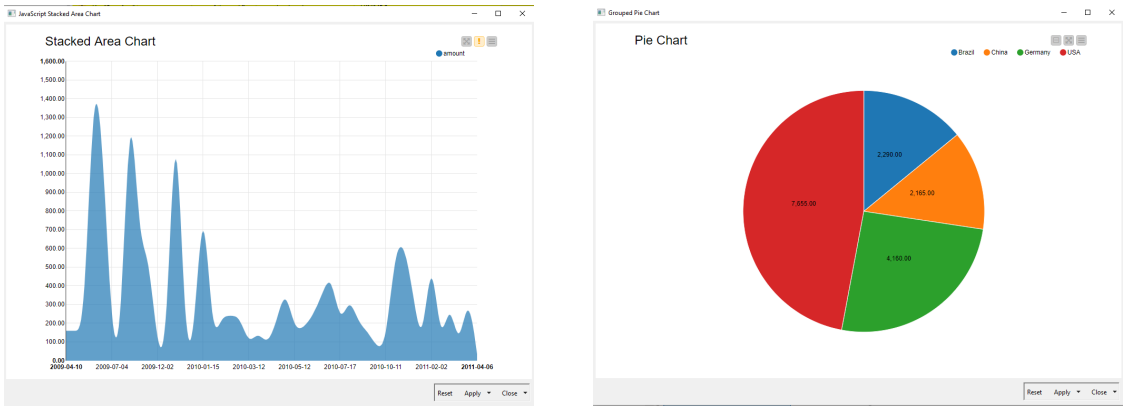
For the creation of the workflow, the steps of the Quick Start Guide on the official KNIME website were followed. [3] For this purpose a workflow consisting of the elements file reader, column filter, row filter was used and a visualization by a stacked area chart and by a Pie/donut chart is provided.

First, the file reader reads all data from the CSV file and displays them in tabular form. With the help of the column filter only the columns amount, country and date were considered. The filtered data is now passed to the row filter, which filters out all rows with "unknown" as the as country. The data is then visualized by graphs in form of a stacked area chart and in form of a pie chart.

[1] https://en.wikipedia.org/wiki/KNIME
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8511823/
[3] https://docs.knime.com/2021-06/analytics_platform_quickstart_guide/index.html#introduction

3 Results

Figure 1 shows the graphs generated by KNIME in form of a stacked area chart (a) and a pie chart (b). In the stacked area chart, the quantity sold is plotted on the y-axis against the respective date on the x-axis. The pie chart shows the total amount sold by country. Clearly is to be seen here that the USA takes the half of all sales followed by germany, brazil and finally china.



(a) Stacked area chart.

(b) Pie chart.

Figure 1: Results of the KNIME workflow

References

[1] Berthold, Michael R. and Cebron, Nicolas and Dill, Fabian and Gabriel, Thomas R. and Kötter, Tobias and Meinl, Thorsten and Ohl, Peter and Thiel, Kilian and Wiswedel, Bernd. KNIME - the Konstanz Information Miner: Version 2.0 and Beyond *SIGKDD Explor. Newsl.*, 2009.