Data Engineering 2 Term Project Report

Team No.1

Dajka Daniel
Ozan Kaya
Eszter Diamant

Introduction

Our initial idea was to scrutinize the E-Commerce traffic change during the global pandemic. Allegedly the covid pandemic has accelerated the E-Commerce globally forcing people to order online in times of lockdowns and restrictions. Typically, e-commerce datasets are proprietary, therefore we could not find appropriate data to conduct this specific research.

We have settled with an older dataset on E-Commerce Sales from 2010. The chosen E-Commerce Data contains 540 thousand sales records from a UK Retailer between 2010 and 2011. We were curious to find out, if the weather has any effect on the E-Commerce sales. It was suspected, that on rainy and cold days, people tend to spend more time with online shopping whereas in warm and sunny days, people are less inclined to spend ordering online.

First, we are presenting the E-Commerce Sales Dataset in more detail, than the process of receiving the weather data is introduced. After, the workflow in Knime is presented, detailing the data pipeline from loading the data until the analytics section. After the data cleaning has been completed, the analytics is briefly shown. In conclusion, there was no effect found of weather influencing E-Commerce Sales in the UK.

Dataset

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. It is avalilabe in the following link: https://www.kaggle.com/carrie1/ecommerce-data . The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. It has 540 thousand records that ranges for 1 year with daily records from 37 countries. It has a customer id, product info, quantity sold and price in it. We can check if people choose e-commerce sites in rainy/cold days perhaps? We can also control for regular customers and focus on new customers or vice versa.

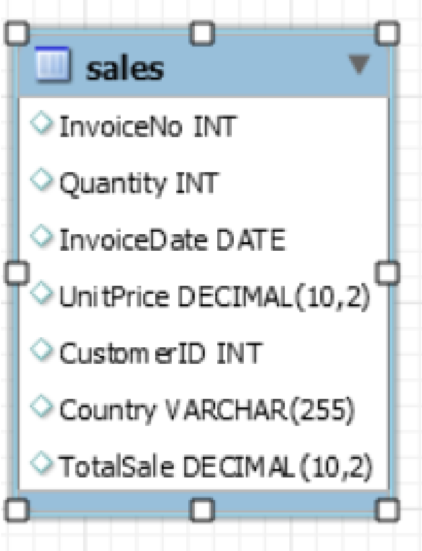
However, the dataset has data from many countries therefore we thought that the top five countries should be analysed. The top 5 countries in terms of order counts are

1. UK with 23.494 orders
2. Germany with 603 orders
3. France with 461 orders
4. Ireland with 360 orders
5. Belgium with 119 orders

We have realized that beside the UK orders, the number are insignificant which lead us to contrentrate on the first country.

The final version of dataset:

Considering the low number of frequencies outside of the United Kingdom, the analysis will use the data from this region. Also, as it was mentioned, only small sales volumes are included assuming that they are made by individuals than wholesalers. Furthermore, in order to be able work with total sales, a new column was added in which the quantity and price were multiplied with each other. After the process, the dataset was imported to MySQL and used as the database. During the data table load, the InvoiceDate was changed from date time to simple date datatype for further purposes.



sales
InvoiceNo INT
Quantity INT
InvoiceDate DATE
UnitPrice DECIMAL(10,2)
CustomerID INT
Country VARCHAR(255)
TotalSale DECIMAL(10,2)

Weather API

To get the Weather Data we have requested an API from NOAA. Since we had the Daily E-Commerce Data from the UK, we were interested in the temperature and rainfall in the UK between 1st of December 2010 and 9th of December 2011.

All in all, there are 144 weather station in the UK. Since London is the most densely populated area and we expect to receive most sales from London, we have chosen one weather station, the London Heathrow Station that is labelled with the stationID: UKM00003772. NOAA provides very detailed weather categories, for us, the most relevant were average daily temperature and precipitation.

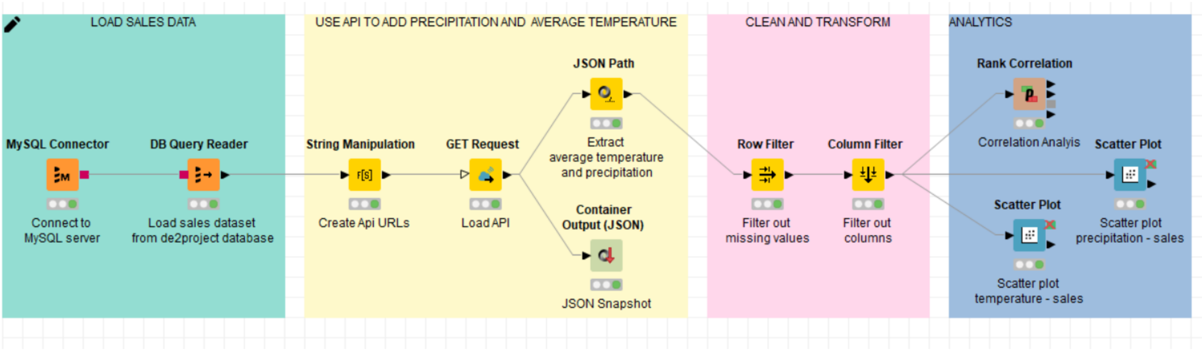
After defining the needed parameters with the help of postman, we could create the needed URL link:

https://www.ncdc.noaa.gov/cdo-web/api/v2/data?startdate=2010-12-01&enddate=2011-12-09&locationid=FIPS:UK&datacategoryid=TAVG&datacategoryid=PRCP&units=metric&datasetid=GHCND&stationid=GHCND:UKM00003772

KNIME Pipeline

The KNIME Pipeline is attached in a separate file called. The whole workflow is separated to four parts:

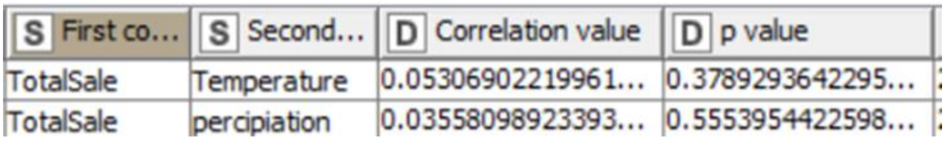
1. **load data** from MySQL database which contains the e-commerce dataset
2. **collect** the daily precipitation and average temperature and **connect to the data table** with the help **of API** defining above
3. **the dataset cleaning** which contains the exclude of missing value and unrelated columns
4. **the analytics** part including a basic correlation analysis and two scatter plots.



Analysis

Hypothesis testing:

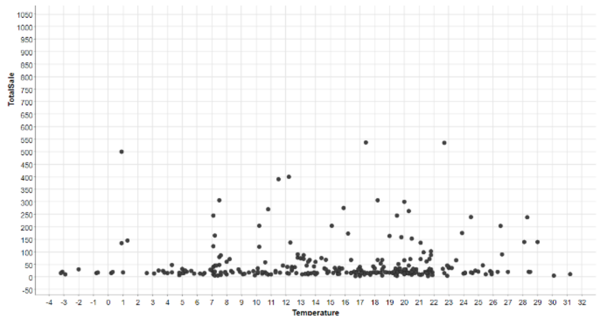
Our null hypothesis was that there is no causal connection between weather and E-Commerce Sales in the UK. Based on the analysis, we could not reject the alternative hypothesis and have to accept the null hypothesis. However, the p-values are significantly high therefore the thesis might risk false negative.

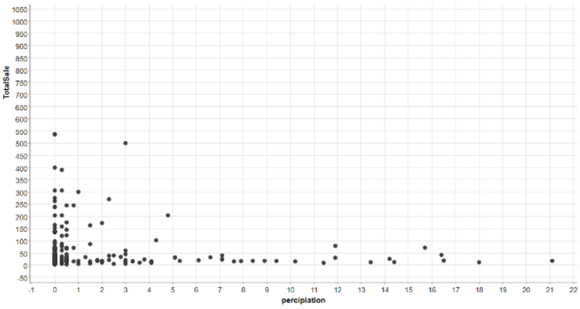


S First co...	S Second...	D Correlation value	D p value
TotalSale	Temperature	0.05306902219961...	0.3789293642295...
TotalSale	percipiation	0.03558098923393...	0.5553954422598...

Correlation:
Two basic correlation was analysed, more precisely, the correlation between the average temperature and sales and between precipitation and sales again. Even if the hypothesis was not rejected, a basic correlation analysis was implemented. It also proved the correct hypothesis considering that the correlation coefficient equalled to zero in both tested cases.

The scatterplots also visualize the same:





Conclusion

To sum up, our research question was whether there is correlation between the weather and e-commerce sales. As it was presented, it can be claimed that there is not. For the analysis, KNIME pipeline was implemented. However, it was also stated that the dataset mostly contains data on wholesale e-commerce, in other words, business to business, transactions and even excluding the greater part and focusing on individuals, it still might affect the analysis outcome.

Project workload allocated as:

Dani (API, Documentation)
Eszter (KNIME workflow)
Ozan (Data Cleaning, MySQL)