Assignment 5

*This assignment can be **solved in groups** with up to 3 people. Please upload only one solution of a group to Moodle but check the crosses in the campus system individually. Make sure to name all group members in your submission by including a file **group.txt** containing the names of all group members.*

Exercise 1 – Define a DWH Schema

Download the dataset dwh.zip from Moodle. After unzipping the download consists of the following csv files containing transaction data from a Kaggle competition on H&M e-commerce data from 2019.
transactions.csv - purchase transactions with date, price, info on the sales channel and references to the customer and article
(**t_dat**,*customer_id*,*article_id*,price,sales_channel_id)

articles.csv – comprehensive master data on articles with taxonomic data on product type and group
(**article_id**,product_code,prod_name,product_type_no,product_type_name,product_group_name,graphical_appearance_no,graphical_appearance_name,colour_group_code,colour_group_name,perceived_colour_value_id,perceived_colour_value_name,perceived_colour_master_id,perceived_colour_master_name,department_no,department_name,index_code,index_name,index_group_no,index_group_name,section_no,section_name,garment_group_no,garment_group_name,detail_desc)

customers.csv - customer data with age, flags on activity status and a hex postal code
(**customer_id**,FN,Active,club_member_status,fashion_news_frequency,age,postal_code)
Since postal_code is a hex string, compute the modulo 10, such that the remainder locates the customer to one of the following Swedish regions:
- 1 → Stockholm
- 2 → Södermanland / Östergötland
- 3 → Jönköping
- 4 → Skåne
- 5 → Kronoberg / Kalmar
- 6 → Värmland / Dalarna
- 7 → Gävleborg / Västernorrland
- 8 → Västerbotten / Norrbotten
- 9 → Blekinge
- 0 → Gotland

Enrich the transaction data also with the daily historic weather data in the file **open-meteo.csv** from 2019.
(**day**, weather_code)

The weather_codes (wmo code) can be looked up from the website of the National Oceanic and Atmospheric Administration Agency:
https://www.nodc.noaa.gov/archive/arc0021/0002199/1.1/data/0-data/HTML/WMO-CODE/WMO4677.HTM

Reason about the files contained in dwh.zip and create a target DWH schema using relational tables where purchase transactions with given prices are facts. Include all dimensions necessary to be able to reason about time (day, weekday, weeks, months and seasons), regions, product types, product groups, graphical appearances, colors, customer age range, active and club member status as well as weather on the date of purchase. Make your own modeling decisions (i.e. on using outrigger tables or junk dimensions) and document the reason, why you did it one way or another. Also document your own assumptions about selecting a reasonable level of granularity (e.g. aggregating weather codes) or defining value ranges (e.g. for customer age).

Exercise 2 – Create A DWH

Build an ETL pipeline to load the csv files into your schema in PostgreSQL. It is up to you on how to write the ETL scripts for loading the data. Using Jupyter Notebook and Python is one of the potential options.

- In case of data quality problems, you may need to fix the issues, document the changes, and assumptions.

- For the sake of simplicity, you do not need to create surrogate keys to replace existing keys.

Upload: Please upload a complete script for the data import. If manual steps are required, document them / provide the changed files.

Exercise 3 – DWH Querying

The data should now be shown as a pivot table with the dimensions of Swedish region, time and product type showing the aggregate sales transactions. Write one query for each of the following tasks:

a) How many customers did at least one purchase.
b) How many articles have been sold in 2019.
c) Aggregate sales by graphical appearance name.
d) Aggregate sales for each Swedish region by product type and season.
e) Aggregate sales for the region Stockholm by product category and month.
f) Color groups that led to highest aggregate sales per season.

Exercise 4 – DWH Querying Customer Segmentation and Analytics

---

Analyze the customers based on their generated revenue. Define at least 3 queries on your own to segment customers according to their characteristics (such as age range or location) or article preferences and their contribution to aggregate sales.

Exercise 5 – DWH Querying Article Portfolio Analytics

Analyze article types and groups based on their generated revenue. Define at least 3 queries on your own to analyze articles according to their characteristics and their contribution to aggregate sales.

Submit the SQL queries with documentation.