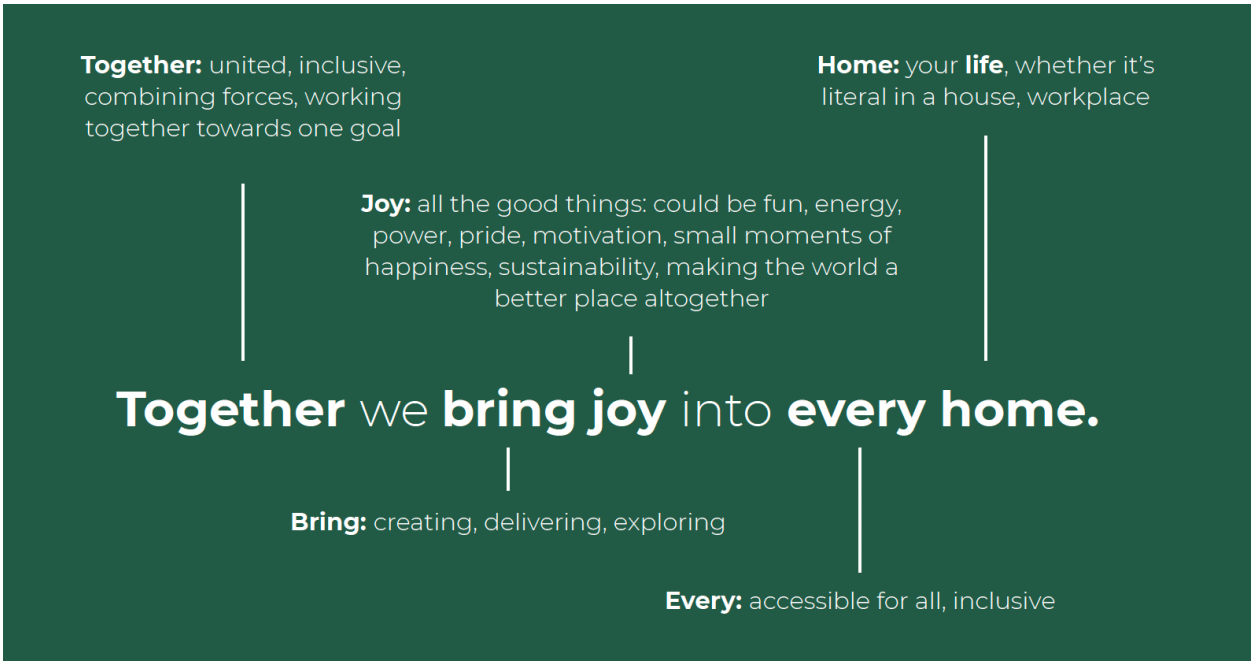
About Wehkamp.nl

Wehkamp.nl is one of the brands of RFS Holland Holding B.V. (The holding company of wehkamp and Tinka) with 2.6 million active customers, more than 600,000 visitors per day and over 11 million shipments per year, Wehkamp is a specialist in the field of online shopping. The product range consists of over 400,000 items of more then 2,500 well-known brands mainly in the categories fashion, living, beauty and baby/child. Wehkamp sells large international brands, local brands, small brands and exclusive brands. In addition, Wehkamp produces collections under private label.

Wehkamp is part of the British private equity company Apax Partners. More than 1.100 people work at Wehkamp, divided over four different locations: the head office, the photo studio and two distribution centers in Zwolle and Maurik. Zwolle is home to the world's largest automated distribution center for online retail. Wehkamp has a tech and data team of 118 fte.

Wehkamp believes in combining fashion and lifestyle in a smart way with online technology and a great shopping experience, in order to offer a relevant and convincing platform to its customers and partners. Meanwhile, 75% of the customers shop mobile and mainly via the app. Almost the entire assortment is ordered today, tomorrow in house with free shipping. Since 2023 customers have to pay a small amount for their returns. Wehkamp's mission is simple: Together we bring joy into every home.



Wehkamp.nl and data

Wehkamp collects data on all aspects of the customer journey of the visitors of their website, and the data science department analyzes these data to generate insights to improve business decisions.

For the course “Data Engineering for MADS” we have the opportunity to work with an anonymized sample of their data. This still means that there are a few restrictions how we are allowed to work with this data (the rules of the game). We will start working with these data during after Week 3 of the course. All personalized data from customers and visitors of Wehkamp has been removed or is replaced with a hash value.

Rules of the game are:

- Do not share or distribute the data or results that you are allowed to work with during the course.
- Do not share the username and password that was provided to you to access the database.
- If you want to work with the data and results on your personal device, this should be locked with a username and strong password.
- Delete all data and tables after finalizing this course from your network drive at the UWP or your personal device.
- Do not copy the data on a USB stick, or any other small portable device. You are only allowed to work with the data in the virtual environment of the University of Groningen via the university workspace (UWP) or your own laptop.

You will receive full details on how to access the database in week 3.

Questions to resolve in the assignments.

The year-on-year growth of sales in the category beachwear is +18%, which is good. But Wehkamp has discovered also quite some volatility in the customer behavior like sales, visits and conversion in this category. Of course this is partially due to seasonal influences. One interesting relationship is still uninvestigated: the influence of the temperature. Of course, the behavior of customers is also affected by other factors like customer background variables (age, gender), their search behavior, their browsing behavior, and so on.

Your task is to translate the management dilemma into a research question and data requirements (Group assignment 1) and to create a data set which is suited for analysis that can generate insights in how weather conditions in addition to other factors drive customer behavior in the category beachwear (Group assignment 2).

Group Assignment 1.

During the tutorials of the first two weeks of the course you will working on translating Wehkamp’s management dilemma into a research question, and to delineate corresponding data requirements.

During the tutorial of week 1, you are supposed to go through the steps that were discussed in the lecture with the goal to define the research question, and its potential refinement into sub questions and factors. You have a 10-minute interview with one of the tutors, as they can be viewed as (external) stakeholders and industry experts. You can use this interview as an

opportunity to gain information about the management dilemma, management question and research question. The schedule for these interviews will be shared during the tutorial. Prepare this interview: think about which information you need and which assumptions or hypotheses you want to check.

In addition, we highly recommend that you also resort to other sources of information to guide your explorations. Specifically, we recommend that you look at how secondary data and academic and professional literature can be helpful in zooming in from the management dilemma to the research question, and how the research question can be divided up into sub questions and factors and specific hypotheses. Please refrain from emailing the guest lecturer or other employees of Wehkamp.nl, as they will not have time to respond to a bombardment of emails.

Subsequently, in the tutorial of week 2, the focus is on translating the research question into data requirements. This will include investigating how relevant constructs that you need for answering the research question(s) can be extracted or derived from the Wehkamp database, but also what external data sources you need to include in your data collection plans. An important part of your thinking needs to address the question how you will combine the different sources, and you need to be able to present a clear idea of how the final data set will look like (e.g. what columns will your data set include, and how many rows will your data set have). The teachers that are present during the tutorial have ample experience with the data and with constructing data sets from multiple sources, and you can ask for their assistance if your group does not know how to proceed.

The deliverable for Assignment 1 is a 10 page (max, excluding title slide) PowerPoint in which you summarize your plans in a concise research proposal. Minimal font size is 12. In the appendix you should add a slide where you explain if and how GenAI is used, i.e. by sharing the prompts, and shortly reflect on the output it has generated.

Topics that should be included and will be graded are (weight of that topic is between brackets):

Deriving the research question (45%)

1. Introduction that includes:

    - Definition of management dilemma plus motivation
    - Translation into management question plus motivation
    - Translation into research question plus motivation

2. Hypothesis tree that shows:

    - Refinement into sub- questions and factors and hypotheses

3. Motivation for the hypotheses:

- Which literature is used
- Motivation for the hypotheses

Data requirements and data model (45%)

1. Operationalization overview (can be combined with the variable description):

    - Operationalization of constructs of research question. I.e. which data, variables and calculations are needed to test your hypothesis.

2. Variable description including:

    - Variable name
    - Variable description
    - Type of variable:
        - Measurement level: text vs numeric (binary, ordinal, nominal, continuous)
        - Declared/appended/overlaid/implied information
        - Internal / external sources: add also which internal source/table is used
        - Structured / unstructured sources
    - Description of dimensions of data set (i.e. expected number of rows and columns) including example table of the first 5 rows.

    Grading will be based on:

    - Selection of included variables
    - Creativity of implied variable definitions
    - Creativity of inclusion of external variables
    - Quality of description content of variables in the data set

3. Data model:

    - Quality of data model: i.e. complete overview of used sources, their relation and key identifiers

Overall quality (10%)

This contains:

- the readability of the presentation,
- the presence and quality of the reflection of using GenAI. For a sufficient grade you should describe the prompts that have been used, the adjustments you've made to use it and/or how you used the output.

The due date for this assignment is September 17, 2024 at 1pm. You are requested to hand in one assignment per group, via Brightspace. Your submission will be automatically checked by URKUND, our plagiarism checker.

GOOD LUCK!