Homework 4

The examination of the module *31-M29 Data Science* consists this year of a portfolio of four programming tasks and a fifth additional programming task (that replaces the exam). This is the fourth programming task. Please hand in your solution via "LernraumPlus" by March 17, 2021.

EDA (22.5 Points)

The number of online shoppers has been growing steadily over the last months. To benefit from this trend online shops have to establish and maintain a fruitful relationship to its customers, which entails understanding the buying behaviour of their customers. Of special interest to online shops is it to understand how promotions influence the buying behavior and in what cases items are returned. The website `https://www.data-mining-cup.com/reviews/dmc-2016/` provides a dataset containing the shopping history of an online store along with a description. Your task is to analyze a subset of that dataset (available in the LernraumPlus). **Note** that the data has been **slightly modified by us** and that the description of the **value range in the data mining cup document is no longer correct!** Furthermore the column *returnQuantity* has been replaced by the column *returned* that indicates if any items have been returned. **Your task** is to perform an EDA of the dataset and create **six graphical representations** that describe your findings and provide interesting insights into the dataset (e.g., discovered patterns or anomalies) with a focus on the influence of promotions and the returning behavior. Upload your solution as an IPython notebook containing the representations, the code, and the descriptions. Make sure that all cells have been executed successfully so that the visualizations are visible.
Please also note the following additional instructions:

- Only use the dataset made available by us in the LernraumPlus. Do not use the data from the data mining cup website!
- Create graphical visualizations that follow the principals of graphical excellence. You should create fair representation that provide the viewer with meaningful insights into the data.
- Describe each data representation with 2 – 5 sentences. Emphasize which insights can be derived from the representation and why these insights are relevant. Argue why your representation is a fair representation of the data. The descriptions must be in English.
- Make sure that all representations highlight different aspects of the dataset. (The representations may partially overlap in the data used.)
- Make sure that all axes of your visualizations are labeled correctly.

- Use text cells to structure your notebook and to delimit different representations.
- To generate the graphical representation you are only allowed to use *matplotlib* and *seaborn*. You are **not** allowed to use the the .plot() function of pandas (`https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.DataFrame.plot.html`).
- **Working in groups is not permitted.** Do not share your code with the other students. You may discuss *what* kind of representations you are making with other students, but you may not discuss *how* you make them.

Grading

Grading will be performed based on the following criteria:

- **Correctness (4.5 Points)**: Are the representations showing what viewers think they are showing (e.g., based on the description or the labels)?
- **Design (3 Points)**: Are basic requirements requirements fulfilled (e.g., no overlapping axis labels, appropriate font size)?
- **Insights (12 Points)**: How difficult is it for the viewer to gain meaningful insights from the representations? Do the visualizations avoid misleading the viewer? Would a different visualization make it easier for the viewer to derive the same insights? How many different "insights" can be gained from the different representations?
- **Description/Notebook readability (3 Points):** Have all insights been explained in the description? Is the notebook well structured?

Late Submission Policy

Late submissions are penalized 5% per started hour after the deadline (e.g., a solution that is handed in 122 minutes after the deadline is penalized 15%). Note that these should still be handed in through LernraumPlus.