Homework 2

The examination of the module *31-M29 Data Science* consists of a portfolio of four programming tasks and a final exam. This is the second programming task. Please hand in your solution via "LernraumPlus" by January 29, 2020.

Bicycle Traffic in Essen (22.5 Points)

The website `https://open.nrw/dataset/radverkehrszaehlungen-es` provides a dataset collected with an automated system that counts the number of bicyclists at 4 different streets in the city of Essen. Your task is it to perform an EDA of the collected data since September 2018. In LernraumPlus you will find a subset of the dataset that you should use for your analysis.
Hand in **two non-graphical and five graphical representations** of the provided dataset that describe your findings and provide interesting insights (e.g., discovered patterns or anomalies). At least one of you visualizations should explore the influence of the weather on the number of bicyclists. To this end you should use the daily weather data from `https://www.ecad.eu/dailydata/index.php` for the weather station *Essen-Bredeney* (STAID: 4074). You are also allowed to combine the given data with data from other sources, but please point these sources out in your submission. 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:

- 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. **Note however** that we will give a bonus point to the **most original/interesting representation** proposed. As originality is a main criteria for our subjective judgment, if the same representation is handed in by multiple students then it cannot be awarded the bonus point.

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? Are the representations not misleading? 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?