Exploratory Data Analysis – Project Proposal

Question/need:

- Which train stations had the highest total number of cuny student entries in 2021?
- Which train stations had the highest %-share of cuny student entries in 2021?
- How did these figures evolve over single months?
- Case: A New York-based fintech wants to buy efficient display advertising inventar at train stations to market its brand new college loan offerings. The company also thinks about creating a short-term investment portfolio for elderly people and using display ads at train stations for marketing.

Data Description:

- Data on number of entries at New York City Subway stations between 2019 (pre-covid) and 2021 (during covid): http://web.mta.info/developers/turnstile.html
- Data on fares measured by weekly number of MetroCard swipes of customers entering New York City Subway etween 2019 (pre-covid) and 2021 (during covid): http://web.mta.info/developers/fare.html
- Individual sample of analysis: number of entries at 1st Jan 2021 at Whitehall station and the corresponding share of student metro-card used on average during the week 1st Jan 2021 belongs to
- Features expected to work with from Turnstile-data:
  - C/A, UNIT, SCP, STATION, DATE, TIME, ENTRIES, DESC
- Features expected to work with from Fares-data:
  - REMOTE (UNIT), STATION, STUDENT, CUNY-60, CUNY-120, SEN/DIS, [total sum of all different MetroCard swipes in single weeks]
- Approach: In a first iteration, the analysis will focus on cuny students, then expand on high school-students (prospective college students) and eldery people
- Comment on limitation: Since the cuny students represent only a subset of all college students, there may be a risk of the sample of fare data not being representative to overall student population in New York

Tools:

- Ingesting raw data from both sources into SQL and conducting preliminary data analysis on both tables (e.g. drop unnecessary columns, rows, join of tables)
- Querying into Python via SQLAlchemy
- Further exploration in Python (creating percentages of student MetroCard-swipes in single weeks, applying percentages to corresponding entries, calculating entries by day by station)
- Visualization in Python using linecharts from matplotlib/seaborn interactive visualization via Bokeh, eventually create visual map to show city areas with high frequencies of student / elderly metro users

MVP Goal:

- MVP contains a line chart displaying the top 10 train stations development of total number of student entries as well as %-share development of student entries per month in 2021