Dr. Konrad Völkel
Heinrich-Heine-Universität Düsseldorf
Wintersemester 2024/25

Data Science 2
Blatt 9, Abgabe am 18.12.2024 um 12:00

Bitte geben Sie ein Notebook und die daraus erstellte PDF-Datei ab.

Aufgabe 1. *Führen Sie eine explorative Datenanalyse auf dem New York Yellow Taxi Cab Datensatz 2023 (wahlweise 2009–2024) durch.*

Sie werden später noch mit den Daten arbeiten (Projekt: Implementierung von BFR-Clustering), daher lohnt es sich, die Analyse für den späteren Projektbericht gleich sauber in einem Jupyter Notebook zu dokumentieren.

- *Datenquelle: www.nyc.gov/site/tlc/about/tlc-trip-record-data.page*
- *siehe auch data.cityofnewyork.us/Transportation/2023-Yellow-Taxi-Trip-Data/4b4i-vvec*
- *Tipp: das Einlesen von Parquet-Dateien ist mit Pandas/PyArrow einfach und hier beschrieben:*
  - *www.nyc.gov/assets/tlc/downloads/pdf/working_parquet_format.pdf*
  - *wesmckinney.com/blog/python-parquet-multithreading/*
- *Tipp 2: Wenn der DataFrame zu groß für's RAM wird, könnte Polars helfen: pola-rs.github.io/polars-book/user-guide/*