Problems

# Problem

I do not know how to force Python to automatically set working directory like in R. So, I have a separate script for it.

01\_Set the working directory.py

# Problem

Python does not provide table function.

In R

> table(df$age)

13 14 15 16 17 18 19 20 21

516 1318 2484 3357 3846 4476 3416 3610 3485

22 23 24 25 26 27 28 29 30

2999 2884 2550 2466 2220 2101 1986 1731 1873

Solution: I had to write it myself.

# Problem

How to save data so it will keep the info like correct categories and types specified in previous script.

So, I have chosen Parquet Format.

# Save the DataFrame to Parquet file

df.to\_parquet('data/data.parquet')

# Loading a DataFrame from a Parquet file

df = pd.read\_parquet('data/data.parquet')

The Apache Parquet project provides a standardized open-source columnar storage format for use in data analysis systems.

# Problem

How to make this expression readable:

# Group by the selected quasi-identifiers and calculate the size of each group

grouped = df.groupby(quasi\_identifiers).size().reset\_index(name='count').sort\_values(by='count', ascending=True)

Solution:

grouped = df.groupby(quasi\_identifiers

).size(

).reset\_index(name='count'

).sort\_values(by='count', ascending=True)

# Problem

# Problem