The file’week6.py’ is my main file, I put my code of pd and dask together, but test separately.

File name: Cab\_data

Number of columns: 7

Number of rows: 359392

Size:20663KB

pd:

columns of files are: Index(['transaction\_id', 'date\_of\_travel', 'company', 'city', 'km\_travelled',

'price\_charged', 'cost\_of\_trip'],

dtype='object')

columns of YAML are: ['city', 'company', 'cost\_of\_trip', 'transaction\_id']

column name and column length validation failed

Following File columns are not in the YAML file ['km\_travelled', 'date\_of\_travel', 'price\_charged']

Following YAML columns are not in the file uploaded []

pd:validation failed

Dask:

columns of files are: Index(['Transaction ID', 'Date of Travel', 'Company', 'City', 'KM Travelled',

'Price Charged', 'Cost of Trip'],

dtype='object')

columns of YAML are: ['city', 'company', 'cost\_of\_trip', 'transaction\_id']

column name and column length validation failed

Following File columns are not in the YAML file ['km\_travelled', 'date\_of\_travel', 'price\_charged']

Following YAML columns are not in the file uploaded []

dask:validation failed

using dask and modin will be quicker than using pd