Updated statistics of New York taxis

We have to compute a set of statistic from the publicly available dataset of New York taxis trips: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Initially, we are only interested in Yellow taxi trips (you can ignore the others). The meaning of each field in the CSV files can be found in this file: https://data.cityofnewyork.us/api/views/biws-g3hs/files/eb3ccc47-317f-4b2a-8f49-5a684b0b1ecc?download=true&filename=data_dictionary_trip_records_yellow.pdf

For these trips, we would like to know:

- The average price per mile traveled by the customers of taxis.
- The distribution of payment types (how many trips are paid with each type of payment)
- The following custom indicator: (amount of tip + extra payment) / trip distance

All these metrics should be stored as objects in a JSON file on a folder in the local file system with the following name convention <year><month><day>_yellow_taxi_kpis.json. For example: "20210101_yellow_taxi_kpis.json".

In addition, we would like to update these metrics every day in a Linux server, given that we expect to append more data as new CSV chunks every day (you can simulate that by chunking one of the monthly available files or downloading more than one month of data).

Submission

You have two options:

1. Bundle your code in a zip file and send it to us by email
2. Create a private Git repository in GitLab / GitHub and share a link to the repo with us

Assessment

What are we looking for?:

- **Clarity**: readability of the code is specially valued
- **Efficiency**: metric computation should be efficient and scale to bigger datasets
- **Documentation**: functions and unclear statements should be properly documented
- **Reproducibility**: dependencies are documented and easy to install with a package manager/docker