# Executive Summary

## Understand The Data

### Stage Overview

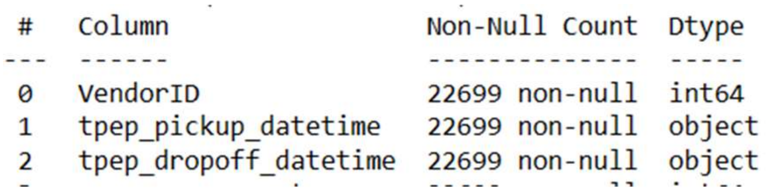
This stage aims to investigate and understand the data provided, use a data frame constructed within Python, perform a cursory inspection of the provided dataset, and inform team members of my findings.

## Key Findings

- The data types of tpep_pickup_datetime and tpep_dropoff_datetime columns are not appropriate.
- PULocationID and DOLocation columns are not useful if we don't have the name or Latitude and longitude lines of each zone.
- There are validity issues in the data:
  - Negative amount of money in columns related to total_amount
  - Observation in Ratecode & extra columns with values are not in their  intervals

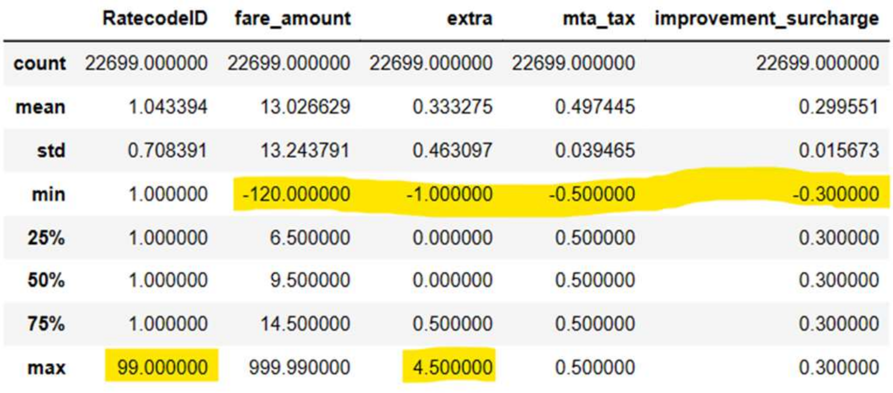
## Screenshots

*Structure  Issues*



```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   VendorID               22699 non-null  int64
 1   tpep_pickup_datetime   22699 non-null  object
 2   tpep_dropoff_datetime  22699 non-null  object
```

*Validity Issues*



|  | RatecodeID | fare_amount | extra | mta_tax | improvement_surcharge |
|---|---|---|---|---|---|
| count | 22699.000000 | 22699.000000 | 22699.000000 | 22699.000000 | 22699.000000 |
| mean | 1.043394 | 13.026629 | 0.333275 | 0.497445 | 0.299551 |
| std | 0.708391 | 13.243791 | 0.463097 | 0.039465 | 0.015673 |
| min | 1.000000 | -120.000000 | -1.000000 | -0.500000 | -0.300000 |
| 25% | 1.000000 | 6.500000 | 0.000000 | 0.500000 | 0.300000 |
| 50% | 1.000000 | 9.500000 | 0.000000 | 0.500000 | 0.300000 |
| 75% | 1.000000 | 14.500000 | 0.500000 | 0.500000 | 0.300000 |
| max | 99.000000 | 999.990000 | 4.500000 | 0.500000 | 0.300000 |

## Next Steps

- Modify tpep_pickup/dropoff_datatime data type to data time and split them into month, day, hour  columns.
- Try to get the ID values data of PULocation and DOLocation columns If you can’t remove them.
- Clean the data from validity issues.
- Analyze the data.