Task 2 - Loading and Saving Data on AWS S3

Q1: Updates based on the instructor's comments and more brainstorming.

a. Comments:

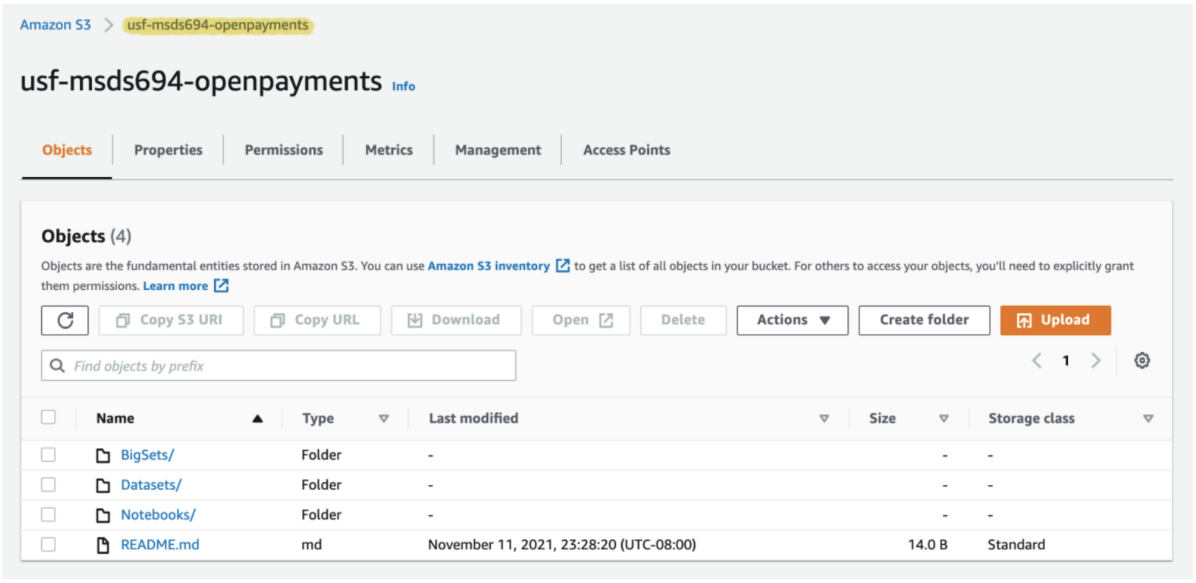
Integrate the data with something else - ex. insurance data, hospitalization cases, the number of doctors for each specialty in each state, opioid addiction cases, etc.

b. Updates:

We are investigating more datasets to augment the open payments information. Currently, we add the data of prescription drugs prescribed by individual physicians and other health care providers. We are thinking that there might be a relationship between two datasets. We select the year 2019 as our analyzing subject. So the datasets uploaded are as follows,

- BigSets/MUP_DPR_RY21_P04_V10_DY19_NPIBN_1.csv: Medicare Provider Utilization and Payment Data in 2019 [link]
- BigSets/OP_DTL_GNRL_PGYR2019_P06302021.csv: Open Payment data in 2019 [link]
- Datasets/OP_PH_PRFL_SPLMTL_P06302021.csv: Open Payment Physician Information data in 2019
- Datasets/VSRR_Provisional_Drug_Overdose_Death_Counts.csv: Provisional Drug Overdose Death Counts [link]

Q2: Load data to S3



Q3: Check the size of the bucket with the loaded data

```
$ aws s3api list-objects --bucket "usf-msds694-openpayments"  --output json --query "[sum(Contents[].Size), length(Contents[])]"
```

```
! aws s3api list-objects --bucket "usf-msds694-openpayments" \
  --output json --query "[sum(Contents[].Size), length(Contents[])]"
```

```
[
    9732296517,
    16
]
```

Q4: Connection between EMR and S3 (by EMR)