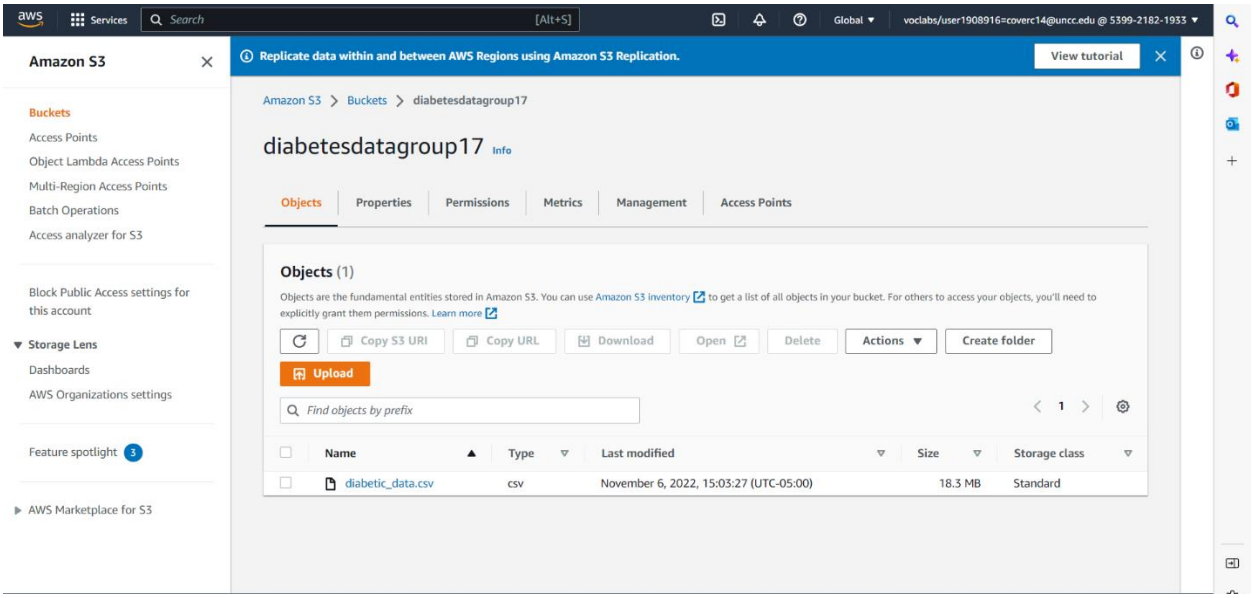
Data Preparation

- To begin our data preparation, we created an S3 bucket and uploaded our dataset. A screenshot of the bucket and data in it is below.



- We then used amazon Sage Maker to import all of the libraries and load the dataset for our initial data preparation, as shown below.

```
[6]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     from sklearn.cluster import KMeans
```

```
#Loading and displaying data

df = pd.read_csv("sample_data/diabetic_data.csv")
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
df.head(50)
```

- We then Checked the shape of our data (101,766 rows and 50 columns) and checked what data type each value/column represented.

- The next part was figuring out which data was incomplete or irrelevant. We used a simple code to count the number of instances in each column that had missing data, which was represented in the dataset as a "?"
- We also used the describe function to get a better understanding of each column as a whole.
- Next we dropped rows that were missing data in the "race", and all of the diagnosis columns.

```
#Checking for missing values in dataset the dataset represents missing
value as '?'
for col in df.columns:
   if df[col].dtype == object:
       print(col,df[col][df[col] == '?'].count())

df.describe().T

#dropping columns with higher missing values
df = df.drop(['weight','payer_code','medical_specialty'], axis = 1)

#dropping column because all records have the same value
df = df.drop(['citoglipton', 'examide'], axis = 1)

df= df.replace('?',np.nan)

# Drop rows that has NaN values on selected columns
df=df.dropna(subset=['race','diag_1','diag_2','diag_3','gender'])
```

This is how our data was prepared to be ready for visualization. More data manipulation might be required if issues arise in the future.