MIS381N (Fall 2019) Homework5

In this homework, you will analyze the IRS 990 Index. Load the file **full index_2016.csv** into Databricks. You can complete the assignment either using SQL or Python support in Databricks.

Q1: Compare different Return Types

As with 1040 individual tax return forms, US 990 filers may file different versions of the form, including 990 and 990EZ. Write a program to output the number of returns filed for each return type. The return type is indicated in Column G of the Excel file and the column is titled Return Type.

Q2. Calculate popular dates to file

List the top ten dates when returns were submitted. Are there any patterns? Why do you think the dates that are near the top are there?

Hints

Look into orderBy or sort

Check the deadlines for 990 filings for more information

Q3. Calculate most popular months to file (10 pts)

Create a new table that shows how many returns were submitted in each month of 2016. Order it by the number of returns submitted.

Hints

Functions in pyspark.sql.functions will be useful. Look at to_date and month in particular.