Group 4 Repeatable ETL Guide

Azure Eller, Widnie Dorilas, Juliann Groglio, Seungmin Kim

August 2, 2022

Introduction

Everyone in our group will work for a client with ties to the financial industry. We were tasked with finding data, gathering questions, and getting answers all within our industry. We decided to focus on school finances. We wanted to see the relationship between school finances in relation to data such as graduation rates and juvenile arrests. However, we were not specific enough and the direction was just too broad, covering all 50 states. Instead, we decided to focus on New York state school districts given that all of us at some point had lived there and were curious. We then started exploring the demographics and graduation rate data of New York state school districts and started wondering whether we could predict things such as dropout and graduation rates between districts using the information gathered.

Data Sources & Extraction

The first data source being used is the Census Bureau's 2019 Public Elementary-Secondary Finance Data where we used the 'All Data Items' that was provided (United States Census Bureau, 2022). We were able to download the 'All Data Items' as a CSV file by clicking on the title with an Excel icon next to it. The document contained two sheets, one with information about the file and one with the actual data. The data was organized by school districts and the state they belonged to along with columns that had broken down information on the different expenditures and revenues. Our second and third data sources came from the New York State Education Department's data website (New York State Education Department, 2022). Along their header, they have a button that directs to the downloads which are organized by year. Since we are working with 2019 data, we went down to the 2019-20 school year and downloaded our data which were the Enrollment and Graduation Rate databases. The databases were downloaded as CSVs. The enrollment database downloaded as a ZIP file and once unzipped contained two Microsoft Access Databases and a readme document that explained what the columns contained and the meanings of what was in the databases. The graduation rate database was similarly downloaded as a ZIP file with a folder inside. The folder contained a Microsoft Access database, an Excel CSV, and a readme explaining what the columns contained and their meanings.

Transformation & Loading

1. Create container in storage account (gen10datafund2205) and upload all CSVs to it
2. Create SQL Database
3. Create login
4. Connect to database using login credentials
5. Create a notebook in Databricks
6. Create a mount point and view all files in that mount point
7. Read in Education Finance Census CSV
8. Only keep columns 'STATE','NAME', 'TSTREV', 'TLOCREV', 'C05', 'E17', 'T06', 'Z32', 'TOTALREV', 'TFEDREV', 'C16', 'C19', 'C04', 'C08', 'TOTALEXP', 'TCURELSC', 'TCURINST', 'TCURSSVC', 'E07', 'E08', 'E09'
9. Rename "TSTREV" to "Total Revenue State"
10. Rename "TLOCREV" to "Total Revenue from Local Sources"
11. Rename "C05" to "SPED Rev State"
12. Rename "E17" to "COE Pupil Support"

13. Rename "T06" to "Property Tax Revenue"

14. Rename "Z32" to "Total Salaries and Wages"

15. Rename "TOTALREV" to "Total Revenue"

16. Rename "TFEDREV" to "Total Federal Revenue"

17. Rename "C16" to "FEDREV Math, Science, Teach Quality"

18. Rename "C19" to "FEDREV Vocational and Technical Education"

19. Rename "C04" to "STATEREV Staff Improvement Programs"

20. Rename "C08" to "STATEREV Gifted and Talented Programs"

21. Rename "TCURELSC" to "Total Current Spending on Programs"

22. Rename "TOTALEXP" to "Total Expenditure"

23. Rename "TCURINST" to "Current Spending on Instruction"

24. Rename "E17" to "Expenditure Pupil Support"

25. Rename "E07" to "Expenditure Instructional Staff Support"

26. Rename "TCURSSVC" to "Total Expenditure Support Services"

27. Rename "E08" to "Expenditure Gen. Administration"

28. Rename "E09" to "Expenditure School Administration"

29. Rename "STATE" to "StateID"

30. Rename "NAME" to "School District"

31. Only keep StateID number 33 (New York)

32. Add a space at the end of each value in column 'School_district'

33. Replace instances of ' SCH ' with ' SCHOOL '

34. Replace instances of ' DIST ' with ' DISTRICT '

35. Remove spaces at end that we created before

36. Run join in SQL database to see what did and did not join

37. Create list in databrick for school districts that came back successfully joined

38. Create another list of all school districts in the census dataframe

39. Compare lists to find differences

40. Create Excel sheet and go through all similarities and compare name differences

41. Read in that created CSV

42. Create list of the first column containing the naming conventions of Grad CSV

1. Create list of the second column containing the naming conventions of Census CSV

2. Change census df to a pandas df

3. Replace values in 'School_district' column with made list so all naming matches across datasets

4. Convert back to a pyspark df

5. Remove New York City #1-32 row

6. Drop duplicate school districts

7. Drop column 'StateID'

8. Read in Graduation Rate CSV

9. Drop columns: 'report_school_year','aggregation_index','aggregation_type','aggregation_code', 'entity_inactive_date','lea_beds','lea_name','nrc_code','county_code','nyc_id','boces_code', 'boces_code','boces_name','membership_code','membership_key','membership_desc', 'subgroup_code','subgroup_name','grad_pct','local_pct','reg_pct','reg_adv_pct',

'non_diploma_credential_pct','still_enr_pct','ged_pct','dropout_pct'

50. Rename "aggregation_name" to "School_district"

51. Rename "nrc_desc" to "Needs"

52. Rename "enroll_cnt" to "Enrollment_count"

53. Rename "grad_cnt" to "Graduation_count"

54. Rename "reg_cnt" to "Regular_degree_count"

55. Rename "reg_adv_cnt" to "Advanced_degree_count"

56. Rename "non_diploma_credential_cnt" to "Non_diploma_count"

57. Rename "county_name" to "County_name"

58. Drop "nyc_ind" column

59. Drop NAs and drop all values of "-"

60. Drop all New York City District # rows

61. Read in Demographic Factors CSV

62. Drop columns: 'YEAR','NUM_ELL', 'PER_ELL', 'NUM_Multi', 'PER_Multi', 'NUM_MIGRANT', 'PER_MIGRANT', 'NUM_FOSTER', 'PER_FOSTER', 'NUM_ARMED', 'PER_ARMED', 'ENTITY_CD'

63. Divide 'PER_HISP', 'PER_BLACK', 'PER_AM_IND', 'PER_ASIAN', 'PER_WHITE', 'PER_SWD', 'PER_FEMALE', 'PER_MALE', 'PER_ECDIS', 'PER_HOMELESS' by 100 to get decimals for percentages

64. Drop all New York City District # rows

65. Export clean files to container

66. Created producer and read in cleaned files

67. Create topic and run producer

68. Created consumer to consume producer messages

69. Create pipeline in data factory

70. Run pipeline to populate SQL database tables and load in data