Project Name

Project Tagline

Najma Osman

02 February 2021

1 Project Overview (600 words)

1.1 Early Exploration Results

1.1.1 What did you find?

1.1.2 Next Steps

1.2 Key Data Elements/Fields

1.3 Key Questions

1.3.1 Q1

1.3.2 Q2

1.3.3 Q3

1.3.4 Q4

1.3.5 Q5

2 Data

2.1 Source + Type

I've chosen a public dataset from Ontario's Data Catalogue. This dataset is on school information and student demographics in the province of Ontario, not including private schools, Education and Community Partnership Programs, or summer and night schools. It includes data on the following:

- board information
- school information
- grade 3 and 6 EQAO achievements for reading, writing and mathematics
- grade 9 EQAO academic and applied student achievements
- grade 10 OSSLT student achievement
- student demographic percentages on student parents, special education, first language spoken, and new students to Canada

The data is reported by:

- Board School Identification Database (BSID) and Ontario School Information System (OnSIS),
- OnSIS Preliminary 2018-2019 (Student Population)
- the Education Quality and Accountability Office (EQAO), specifically the 2018-2019 data,
- and the 2016 census.

2.2 Cleaning Requirements (500 words)

This dataset does not need a significant amount of cleaning. I've only used the package janitor by Firke (2021) to modify the column names into snake case (variable_name) format, to access column names more easily. All other changes focused on data reduction to remove any unwanted fields, which I'll go into detail on in section 2.2.6.

2.2.1 Data Quality

2.2.2 Anonymization

This dataset pulls from OnSIS and Statistics Canada, which both suppress results for variables based on scool population size to protect student privacy. The following methods were used to ensure anonymity:

- randomly rounding percentages up or down depending on school enrollment,
  - 0 - NA
  - 1-49 - SP
  - 50-99 - round up or down to a multiple of 10
  - 100-499 - round up or down to a multiple of 5
  - 500-4,999 - round up or down to the ones digit
  - 5,000 + - round up or down to one decimal place
- not publicly reporting data where enrollment is less than 10

2.2.3 Inconsistencies

2.2.4 Missing Data

There 4 instances of missing data in this set, which is already handled by the cataloguer:

- where student population information isn't available due to the school board not providing the data to the ministry (denoted by NA),
- where schools the school does not have EQAO results (denoted by ND),
- where the number of students participating is fewer than 10 and anonymity isn't ensured (denoted by NR),
- and where the results are repressed due to school enrollment of fewer than 50 students (denoted by SP)

2.2.5 Outliers

2.2.6 Unwanted Data

Since my research questions/interests specifically focus on _, I removed the following columns from the dataset:

- school number, board type, building suite, P.O. Box, Street, postal code
- phone number, fax number, school website, board website
- grade 3 and 6 results (EQAO, achievement of provincial standard (percentage of student and change over 3 years))
- extract date

I also removed rows containing:

- elementary schools

2.2.7 Grouping/Aggregate Data

3 Expected Outcomes (400 words)

What do you expect to learn? We are not asking to predict the results but asking what you will be able to tell a story on? Housing prices? Pandemic control? Marketing results?

4 Expected Challenges (400 words)

Getting access to the latest data Finding an expert on the topic Finding detailed data Data consistency ...

5 Next Steps (300 words)

Your skills and ideal team member skills

External resources & support: What other resources and support can you use to shape your story and provide context

6 Supporting Documents

References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.