**Bigquery Activity**

**Goal**

Practice using BigQuery, along with data upload, data warehouse architecture, and querying.

**Description**

You are given three csv files:

student\_exam\_results.csv

* Shows math, reading, and writing scores for each student

student\_exam\_survey.csv

* Shows results of a short survey each student was asked to complete. Includes parental level of education, lunch plan, and test preparation course. The values in these three columns include numerically encoded values that equate to these:

|  |  |  |
| --- | --- | --- |
| **Parental Level of Education** | **Lunch Plan** | **Test Preparation Course** |
| 1 = Some High School | 1 = Free/Reduced | 0 = None |
| 2 = High School | 2 = Standard | 1 = Completed |
| 3 = Some College | - |  |
| 4 = Associate’s Degree |  |  |
| 5 = Bachelor’s Degree |  |  |
| 1. = Master’s Degree |  |  |

student\_list.csv

* Shows student information including gender and year group

Your first course of action is to prepare the data. You should design a data warehouse with this data using a star schema. (It might be easiest for you to first generate a single large data with all the data denormalized). You may need to create your own data to include the decoded values for student\_exam\_survey.csv.

Once you have your data warehouse completed, find this information:

* Average scores across subjects of students grouped by year group.
* Average scores across subjects of students grouped by gender
* How does parental level of education affect student test scores?
* How well did the test preparation course help students?

Steps:

1. explain each variable, its type, brief description etc.
2. pre-process data explaining in detail any changes, null values removed, type casting if necessary, or any new fields created to hold new data created from existing data
3. perform exploratory data analysis with plots of all variables, note any potential outliers or influential points, include scatter plots or histograms or pie charts etc, whichever appropriate, any also perform simple two-way comparisons with variables
4. create a star schema with a corresponding ERD in MySQL or any tool you prefer, and be sure to present this ERD in presentation
5. upload data to BigQuery data warehouse with your star schema
6. perform queries in BigQuery, save these queries, and take screen shots in big query performing them, then also present queries live in BigQuery during the presentation. The queries should answer the foure questions given. Also come up with at least 3 more questions of your own that you answer with queries in BigQuery. Try to make them very analytical in nature.

Example Analysis answers and questions: variable A increases as variable B increases and we might postulate further increase in A leads to increases in B with a scatter plot, regression line, and regression equation with corresponding R-squared (coefficient of determination) value in Excel. This could answer the question is variable B correlated with variable A? Or can we use any variables increase to predict the increase of another variable? Or are any variables correlated (then look at other variables relationships too).

Or another example could be Students in category C who are Male and have parental education level X perform better on math Tests as compared to Students in category C who are female and have parental education level X as exhibited by Y percentage of these male students in C scoring above w... and only Z percentage of female students in C with parental education level X scoring above w..so the question could be Is there a good predictor of performance on math tests for students in group C or make it broader and show more analysis.

I want you to be creative and analytical, really look at the data, explore the data, and come up with your own analytical questions and answers.

**Presentation**

You will give a 5 minute presentation of your findings on February 9, 2024 Be prepared to show an ERD of your data warehouse, as well as the results of your queries.

**Strech Goals:**

* Write a few more questions to answer
* Create a Star Schema and answer a few questions for queries for your own dataset

Data Variables Explained:

1. Student ID: ID number used to uniquely identify each row inside the student\_list. INT
2. Parental\_level\_of\_education: Shows the highest level of education that a parent received. INT
3. Lunch\_plan: Indicates the level of lunch a student has enrolled into. INT
4. Test\_preparation\_course: Indicates whether the student has successfully completed a test preparation course or not. TEXT
5. Test\_ID: ID Number used to uniquely identify each row pertaining to the test. INT