TRABALHO FINAL
BANCO DE DADOS II _ 2023.1
FELIPE FREITAS SILVA, LUIZA HELLER KROEFF PLÁ, PAOLA ANACLETO LOPES

1. Contextualização

A Inteligência de Negócio (BI) tem como principal função o fornecimento de informações para possibilitar que diferentes organizações cheguem a decisões da forma mais estratégica possível. É uma abordagem que utiliza a análise de dados, tirando proveito da crescente disponibilidade de dados abertos, provenientes de diferentes fontes, o que fornece às empresas “insights” relevantes sobre mercado, concorrência e clientela.

Neste projeto de BI, será explorada a utilização das fontes de dados abertos listadas abaixo, provenientes da plataforma Kaggle, para analisar e relacionar as informações disponíveis, de forma a obter uma visão completa e abrangente das informações. O objetivo deste projeto é entender os fatores que influenciam o sucesso acadêmico e fornecer informações que auxiliem na melhora do sistema educacional geral.

Fonte de Dados 1: “Students Performance in Exams”

Fonte de Dados 2: “Students Exam Scores: Extended Dataset”

2. Questões de Negócio

Neste projeto, fomos atrás de fatores que poderiam influenciar a educação de forma geral. Para isso formulamos questões, para as quais fomos atrás de resposta, via análise dos dados abertos, citados anteriormente.

Abaixo, segue as perguntas para as quais buscamos respostas com a análise dos dados encontrados:

Pergunta 1: Quais as características que mais afetam os resultados em exames?

Pergunta 2: Existem fatores que interagem e, como resultado, influenciam as pontuações nos testes?

Pergunta 3: Quais fatores socioeconômicos (como nível de educação dos pais, renda familiar, etc.) estão correlacionados com o desempenho dos alunos nos exames?

Pergunta 4: Como os diferentes grupos étnicos se comparam em termos de desempenho nos exames?

3. Fonte de Dados e Estrutura

As fontes de dados selecionadas para este projeto vêm do Kaggle, uma plataforma online, subsidiária da empresa Google, para competições focadas na ciência de dados, que “datasets” abertos para uso. Para o trabalho, selecionamos os conjuntos: “Students Performance in Exams” e “Students Exam Scores: Extended Dataset”, sendo que ambos se relacionam por conterem informações sobre o desempenho de estudantes, incluindo diversos atributos que permitem uma análise mais conclusiva dos dados apresentados.

A seguir, informações mais específicas acerca de cada Dataset:

“Students Performance in Exams”:

Link: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?resource=download

Formato: CSV

Estrutura: Essa fonte contém informações sobre a performance de estudantes em exames, incluindo suas notas nas áreas de matemática, leitura e escrita. São apresentados ainda, informações demográficas, tais como: gênero, etnia, nível de educação parental, alimentação no período de almoço e participação em cursos preparatórios.

“Students Exam Scores: Extended Dataset”:

Link: https://www.kaggle.com/datasets/desalegngeb/students-exam-scores

Formato: CSV

Estrutura: Esse Dataset fornece, além do que foi listado para a primeira fonte, algumas informações adicionais, que são: estado civil dos pais, participação em esportes, se é ou não primeiro filho(a), número de irmãos, meios de transportes e quantidade de horas usadas para estudo na semana.

4. Modelagem Multidimensional

A modelagem multidimensional é uma técnica fundamental no campo da Inteligência de Negócios (BI) que visa organizar os dados de maneira estruturada e intuitiva, permitindo análises eficientes e compreensão dos relacionamentos entre medidas e dimensões. Neste projeto, criamos um modelo multidimensional utilizando o esquema Star Schema, que inclui uma tabela-fato central, medidas e várias tabelas dimensionais.

A tabela-fato é o ponto central do modelo, contendo as métricas ou medidas a serem analisadas. No nosso caso, a tabela-fato é chamada de "ExamScores" e possui medidas relacionadas às notas dos alunos em diferentes disciplinas, como matemática, leitura e escrita. Essas medidas são quantitativas e representam os indicadores de desempenho dos alunos.

Além da tabela-fato, temos várias tabelas dimensionais que representam as diferentes características ou dimensões pelas quais os dados serão analisados. Cada dimensão possui uma tabela associada contendo atributos relevantes para essa dimensão. No nosso modelo, as dimensões incluem:

1. "Gender" (Gênero): Essa dimensão representa os diferentes gêneros dos alunos. A tabela dimensional correspondente possui atributos como sexo e cod_gender, que identificam e descrevem os gêneros.
2. "Race" (Raça/Etnia): Essa dimensão representa as diferentes raças ou etnias dos alunos. A tabela dimensional correspondente possui atributos como cod_race e group, que identificam e descrevem as raças ou etnias.
3. "Parental Level of Education" (Nível de Educação dos Pais): Essa dimensão representa o nível de educação dos pais dos alunos. A tabela dimensional correspondente possui atributos como cod_parentEducation e education_level, que identificam e descrevem os diferentes níveis de educação dos pais.
4. "Lunch" (Refeição): Essa dimensão representa o tipo de refeição dos alunos, indicando se eles têm uma refeição paga ou gratuita. A tabela dimensional correspondente possui atributos como cod_lunchType e type, que identificam e descrevem os diferentes tipos de refeição.
5. "Test Preparation Course" (Curso Preparatório para o Teste): Essa dimensão representa se o aluno realizou ou não um curso preparatório para o teste. A tabela dimensional correspondente possui atributos como cod_course e courseName, que identificam e descrevem os diferentes cursos preparatórios.

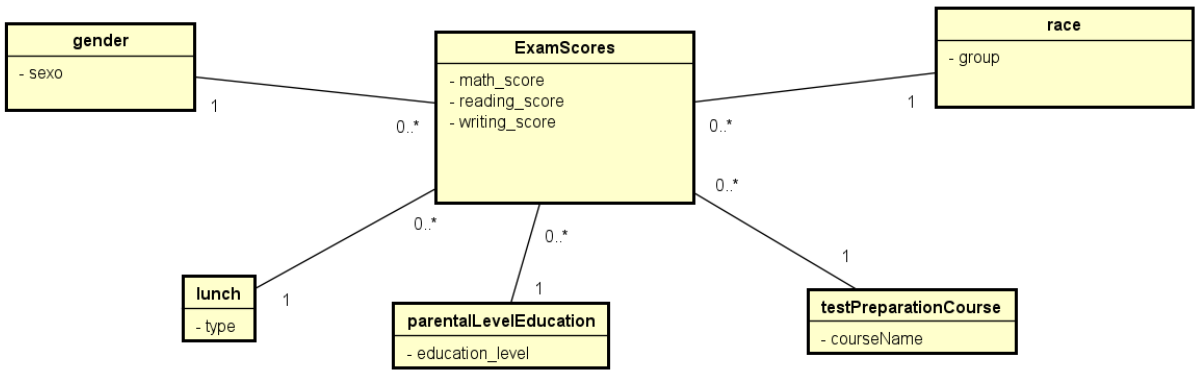
A relação entre a tabela-fato e as tabelas dimensionais é estabelecida por meio de chaves estrangeiras, permitindo a análise integrada das medidas em relação às diferentes características dos alunos. Essa estrutura em estrela (Star Schema) facilita a

navegação e a compreensão dos dados, possibilitando a execução de consultas eficientes e a obtenção de insights significativos por meio das análises multidimensionais.

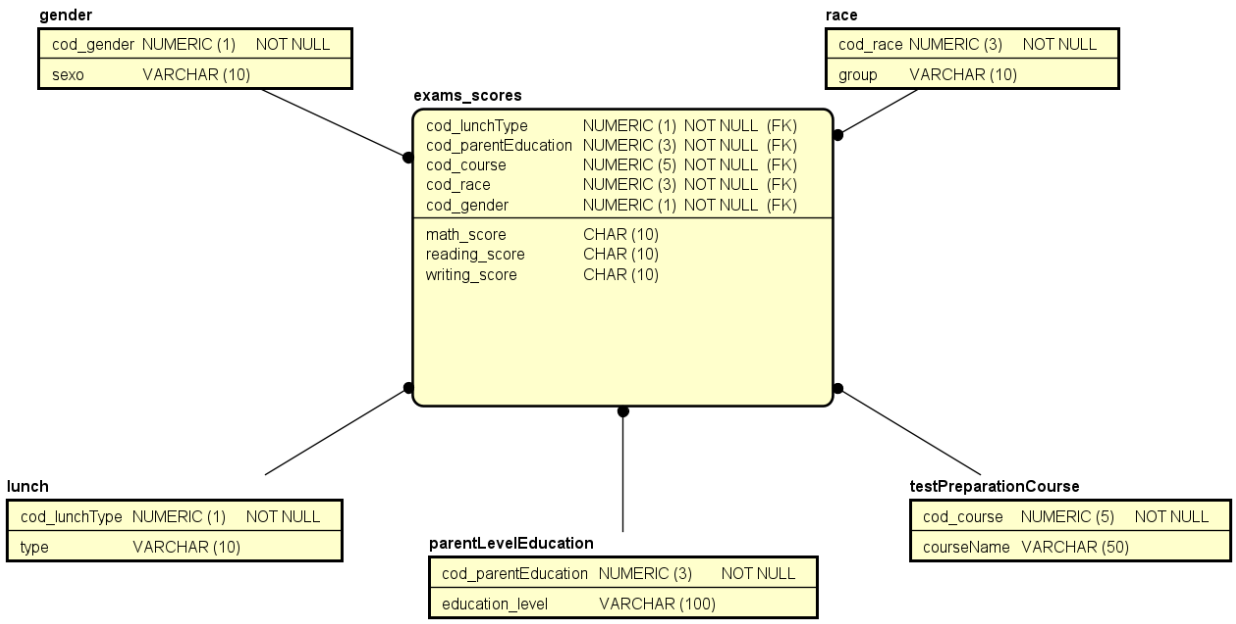
Em resumo, a modelagem multidimensional realizada neste projeto utiliza o esquema Star Schema para organizar os dados relacionados ao desempenho dos alunos. A tabela-fato "ExamScores" contém as medidas de notas em diferentes disciplinas, enquanto as tabelas dimensionais representam as características dos alunos, como gênero, raça/etnia, nível de educação dos pais, tipo de refeição e participação em curso preparatório. Essa modelagem permite uma análise abrangente e aprofundada dos dados, contribuindo para a tomada de decisões informadas e a obtenção de insights valiosos no contexto educacional.

5. No Astah: Realizar a modelagem multidimensional do cubo, com tabela-fato, medidas e dimensões, conforme exemplos.

Modelo Estrela: Cubos Com um Nível Hierárquico Nas Dimensões

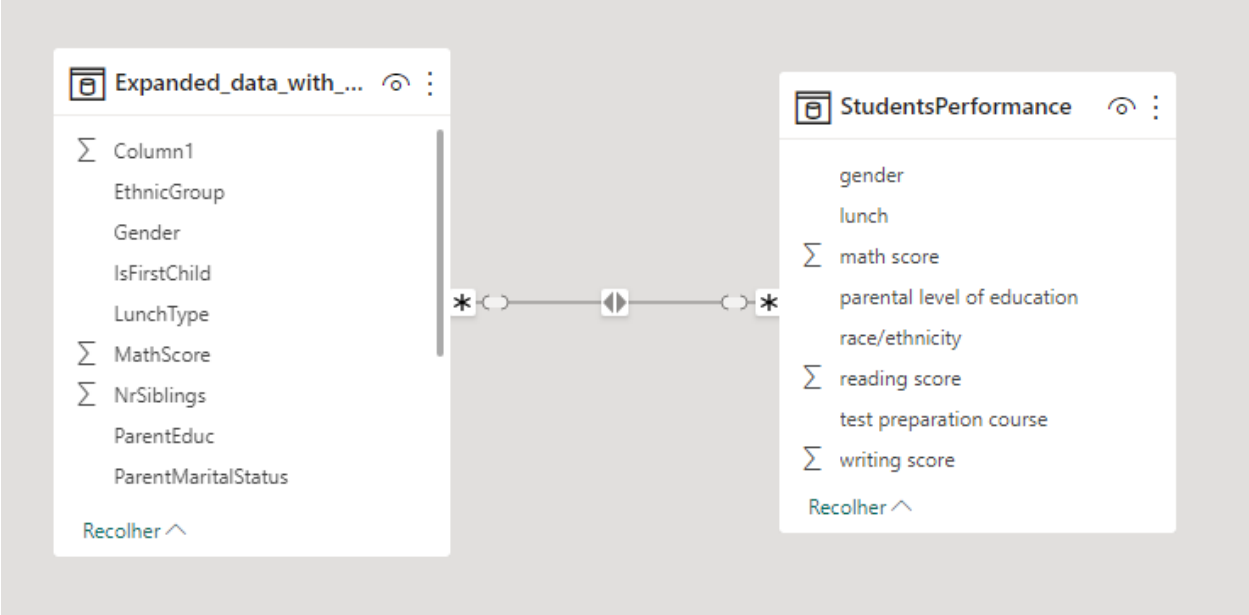


Modelo Estrela: Esquema Lógico

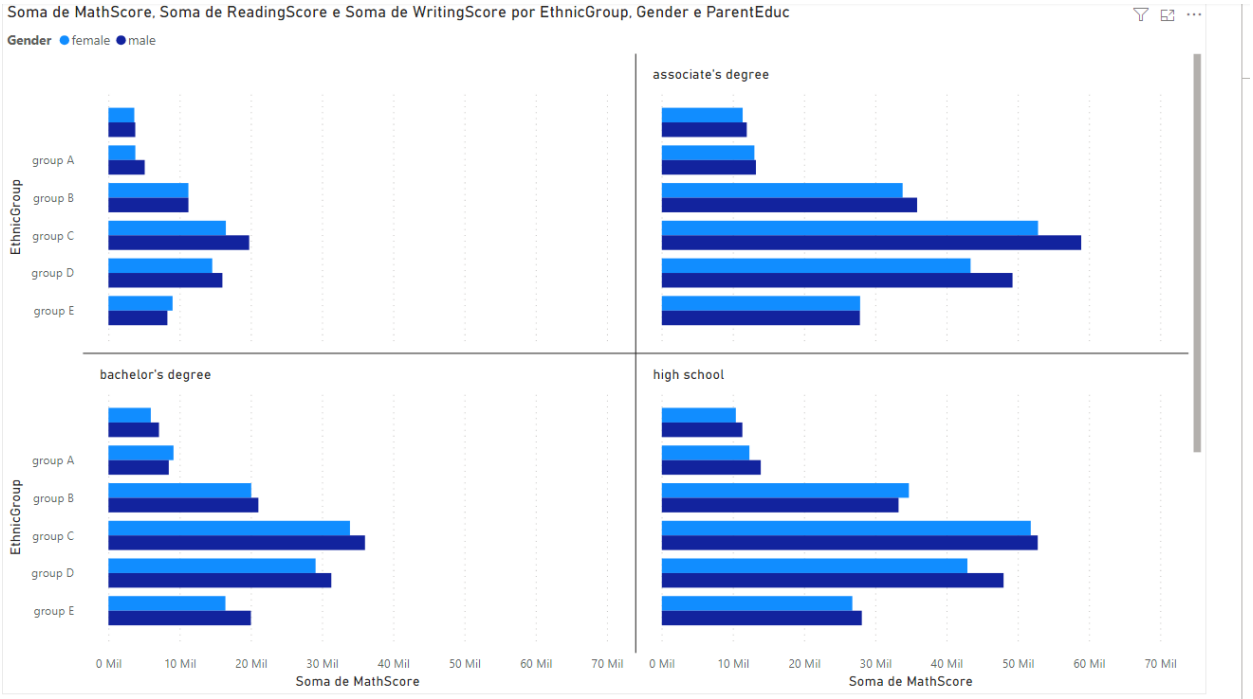


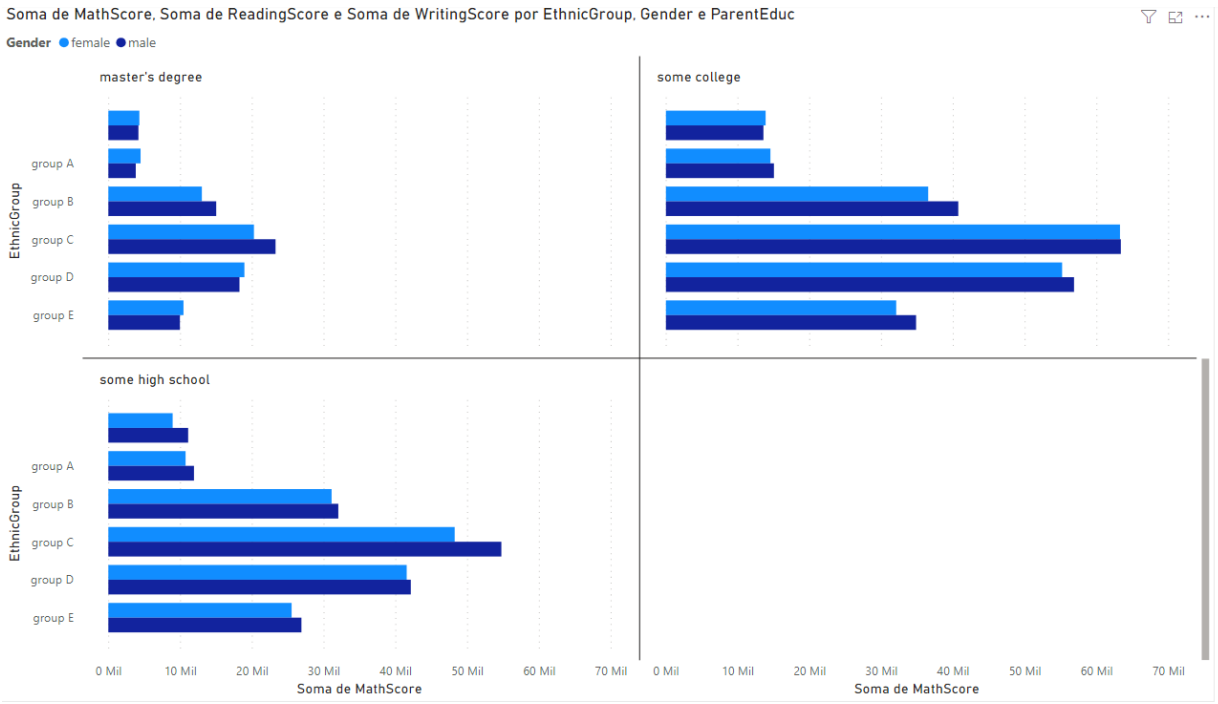
PARTE II

Modelagem lógica do cubo



Visualização


Soma de MathScore, Soma de ReadingScore e Soma de WritingScore por EthnicGroup, Gender e ParentEduc
Gender female male
associate's degree
bachelor's degree
high school
EthnicGroup
group A
group B
group C
group D
group E
0 Mil
10 Mil
20 Mil
30 Mil
40 Mil
50 Mil
60 Mil
70 Mil
Soma de MathScore


Soma de MathScore, Soma de ReadingScore e Soma de WritingScore por EthnicGroup, Gender e ParentEduc
Gender female male
master's degree
some college
some high school
EthnicGroup
group A
group B
group C
group D
group E
0 Mil
10 Mil
20 Mil
30 Mil
40 Mil
50 Mil
60 Mil
70 Mil
Soma de MathScore