Trabalho Prático 2

Ari Gonçalves da Silva Filho, Caio Rodrigues Costa, Gustavo Henrique Silva Paiva, João Vitor Soares Santos, Túlio Henrique Rodrigues Costa

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais – Belo Horizonte – MG – Brasil

{arigsf, caiorc, gust4vo, jvss2023, tuliohrc}@ufmg.br

1. Objetivos

O objetivo desta análise é investigar a relação entre o salário e a carga horária dos profissionais de educação com o desempenho dos alunos nas escolas públicas. Para o primeiro, será considerada a média salarial e de horas trabalhadas Já para o segundo, será utilizado o Índice de Desenvolvimento da Educação Básica (IDEB). As hipóteses a se provar, ou não, estão descritas abaixo.

Hipótese 1: Existe uma correlação positiva entre o salário médio dos profissionais de educação nos municípios e o desempenho dos alunos no IDEB. A suposição é que salários mais altos podem atrair profissionais mais qualificados e motivados, o que resulta em melhor desempenho dos alunos.

Hipótese 2: Existe uma correlação negativa entre as horas médias trabalhadas pelos profissionais de educação e o desempenho dos alunos no IDEB. A suposição é que uma carga de trabalho excessiva pode levar ao esgotamento dos profissionais, o que impacta negativamente a qualidade do ensino e, consequentemente, o desempenho dos alunos.

Hipótese 3: O salário ou as horas trabalhadas dos profissionais de educação não afetam significativamente o desempenho dos alunos no IDEB.

2. Metodologia

A metodologia adotada incluiu as etapas a seguir.

1. Coleta de Dados:
    - Os dados foram obtidos a partir de arquivos públicos disponibilizados pelo INEP. As URLs dos arquivos utilizados estão listados abaixo.

      Remuneração Docentes Brasil 2019,

      Remuneração Docentes UF 2019,

      Remuneração Docentes Municípios 2019,

      Divulgação Brasil IDEB 2019,

      Divulgação Regiões UFs IDEB 2019,

      Divulgação Anos Iniciais Municípios 2019,

      Divulgação Anos Finais Municípios 2019,

      Divulgação Ensino Médio Municípios 2019,

      Divulgação Anos Iniciais Escolas 2019,

      Divulgação Anos Finais Escolas 2019,

      Divulgação Ensino Médio Escolas 2019.
2. Processamento de Dados:
    - Os dados foram extraídos dos arquivos Excel utilizando a biblioteca pandas, as colunas e linhas indesejadas foram removidas e foi realizado alguns ajustes como indexação por número ao invés de letras, que é o padrão no Excel.
    - Por não se tratar de um sistema muito complexo, foi utilizado o SQLite para processamento de manipulação do banco de dados.
3. Organização e Normalização dos Dados:
    - Um modelo ER (Entidade-Relacionamento) foi elaborado para melhor visualizar e entender a estrutura das tabelas e suas relações.

- As tabelas foram organizadas conforme o modelo, incluindo a definição de chaves primárias e tipos de dados apropriados para cada coluna e posteriormente salvas no SQLite

4. Análise e Visualização:
    - As consultas SQL foram elaboradas especificamente de forma a analisar regiões e estados separadamente e utilizando matplotlib, foram gerados gráficos de correlação para visualização

2.1. Ambiente de desenvolvimento

Foi escolhido desenvolver todo o código em um notebook. Uma das maiores motivações para tanto foi a sua execução celular, que permite executar apenas partes específicas do código. Isso é importante porque apenas o download e processamento dos dados totalizam cerca de dez minutos. Os arquivos e o código estão disponíveis em um repositório público do GitHub, que pode ser acessado por este link.

2.2. Ferramentas utilizadas

- SQLite: Utilizado para armazenar e manipular os dados.
- Pandas: Utilizado para a leitura dos arquivos Excel, processamento e filtragem dos dados.
- Matplotlib: Utilizado para a geração de gráficos e visualizações.
- Python (Jupyter Notebook): O ambiente de desenvolvimento utilizado para implementar o código e realizar as análises.

3. Organização dos dados

3.1. Modelagem entidade-relacionamento

O modelo ER não fica com boa visibilidade neste documento. Para acessá-lo, clique neste link.

4. Análise crítica das fontes dos dados

Ao analisar as fontes de dados utilizadas, é possível identificar diversas dificuldades e limitações que impactam a qualidade dos dados disponibilizados. Em primeiro lugar, é notável um problema significativo de atualização. Para que se possa fazer um cruzamento certeiro de informações, dado o contexto de determinada época, é preciso que os anos de análise sejam iguais. Porém, enquanto o Indicador da Educação Básica (IDEB) foi realizado pela última vez em 2021 e acontece a cada 2 anos, a remuneração média dos docentes foi realizada pela última vez em 2020. Dessa forma, teve que se retroceder o ano de análise, a fim de que se possa analisar o mesmo período, no caso 2019.

Outra limitação é a própria forma com que os dados são disponibilizados. Nesse caso, ambos os indicadores foram revelados por meio de planilhas. Entretanto, é interessante notar que elas não são tão amigáveis para processos de automatização para extração e análise de dados. Dessa forma, torna-se necessário fazer uma intensa preparação manual, para que esses dados tenham um primeiro tratamento para serem usados.

A estrutura da planilha também representa um desafio. No caso do IDEB, por exemplo, há a junção de diversos anos em um só arquivo. Todos os dados, sendo eles taxas de aprovação, indicadores de rendimento, notas do SAEB e resultado do IDEB, dos anos de 2005 a 2019 se encontram numa única planilha. Essa agregação resulta num arquivo complexo e grande, que dificulta a navegação e a extração de dados específicos. Dessa forma, torna o processo de análise mais lento e sujeito a falhas, em virtude dessa estrutura desorganizada.

Além disso, a qualidade dos dados é um aspecto importante a ser julgado. Um dos pontos mais nítidos é o excesso de valores esparsos, faltantes e nulos. Exemplo disso é na planilha da remuneração dos docentes. Nela é possível verificar diversos valores estranhos, como “a”, “d”, que não remetem a nenhum significado, o que torna o dado obsoleto. Tais valores inválidos exigem uma atenção especial durante a limpeza, pois podem distorcer os resultados das análises se não forem tratados adequadamente.

Outra prova disso é na planilha do Ideb. Nela é possível observar diversos valores preenchidos com “-”, ou seja, não há dados disponíveis sobre aquele assunto, o que novamente torna o dado obsoleto. Dessa forma, há cidades sem dados salariais de professores, como em Colorado do Oeste (RO). Isso pode ocorrer devido a falhas na coleta dos dados ou a própria inexistência de informações para aqueles contextos. A ausência de dados completos compromete a integridade das análises, haja vista que dificulta uma visualização abrangente e concisa do cenário educacional.

Essa visualização concisa também é dificultada pela falta de padronização na organização dos dados internamente. Na planilha do Ideb, os valores de desempenho da rede estadual de ensino são separados no âmbito de cada município. Na planilha dos rendimentos salariais, por sua vez, os salários dos docentes da rede estadual de ensino não estão separados na esfera de cada município, mas no estado como um todo. Dessa maneira, foi considerado o estado inteiro como média para as escolas estaduais dos municípios.

Portanto, a análise das fontes de dados utilizadas revela uma série de problemas que passam desde a desatualização até a inconsistência. Por isso, é extremamente importante investir nos processos de coleta, padronização e divulgação dos dados. Isso é extremamente essencial tanto no contexto da Lei de Acesso à Informação (LAI), quanto na análise exploratória dos dados.

5. Análise de correlação

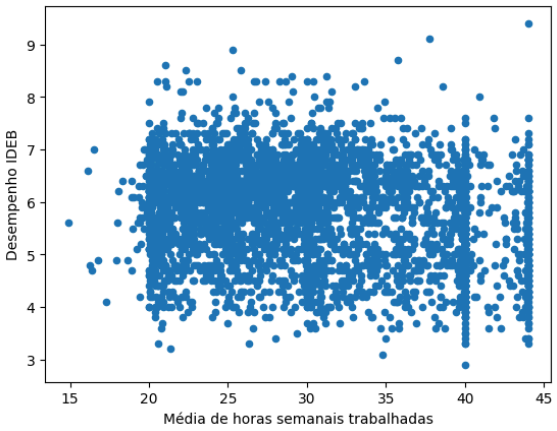
Para a análise de correlação, foi utilizada uma função *corr()* do Pandas, referente à correlação de Pearson. Ela é uma medida estatística amplamente utilizada para avaliar a força e a direção da relação linear entre duas variáveis. O coeficiente varia entre -1 e +1, sendo que

- +1 indica uma correlação positiva perfeita, em que as variáveis aumentam juntas na mesma proporção.
- -1 indica uma correlação negativa perfeita, onde uma variável aumenta enquanto a outra diminui na mesma proporção.
- Quanto mais próximo o coeficiente de +1 ou -1 mais forte é a correlação linear, e quanto mais próximo 0 menos forte é a mesma.

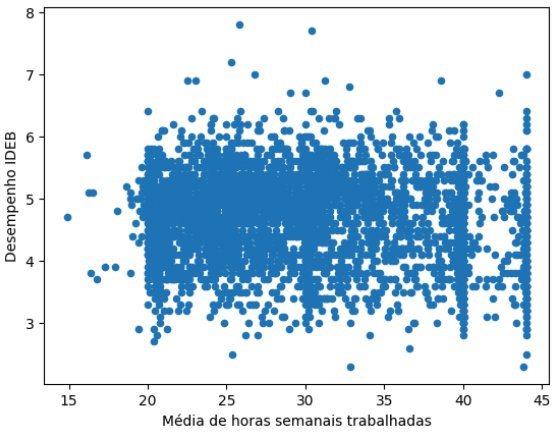
A seguir serão mostrados alguns gráficos resultantes. Inicialmente foi pensado, para fins de simplificação, em levar em conta apenas o desempenho do IDEB do Ensino Fundamental I. Isso porque é uma das amostras mais ideais, dado a enorme diferença de aplicantes da prova Brasil entre essa categoria e a do ensino médio. Porém, é relevante observar a diferença do impacto que o salário e as horas trabalhadas pelos professores têm nas diferentes etapas da educação básica.

5.1. Resultados nacionais: Horas trabalhadas x Desempenho IDEB

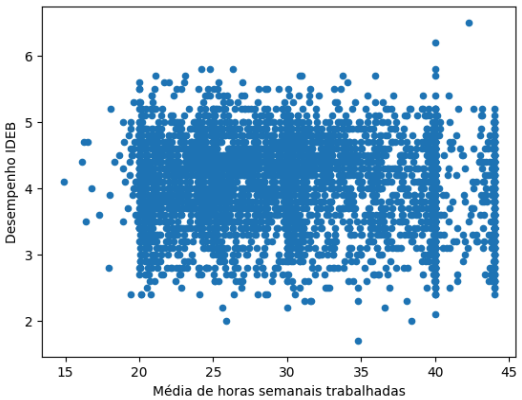
No que se refere à amostra de todos os municípios brasileiros, sem distinção de estados, a correlação entre salário e horas trabalhadas é negativa, como se espera pela hipótese 2. Porém, a intensidade dessa correlação é baixa. Observa-se a sutil perda de correlação conforme mais avançada é a etapa de educação básica.



Ensino Fundamental 1 (c = -0.229)



Ensino Fundamental 2 (c = -0.139)



Ensino Médio (c = -0.097)

5.2. Resultados Estaduais: Horas trabalhadas x Desempenho IDEB

Para não haver uma grande extensão de gráficos para cada estado, foi levantado tudo em uma tabela para melhor análise. O Distrito Federal ficou de fora da tabela por falta de amostras.

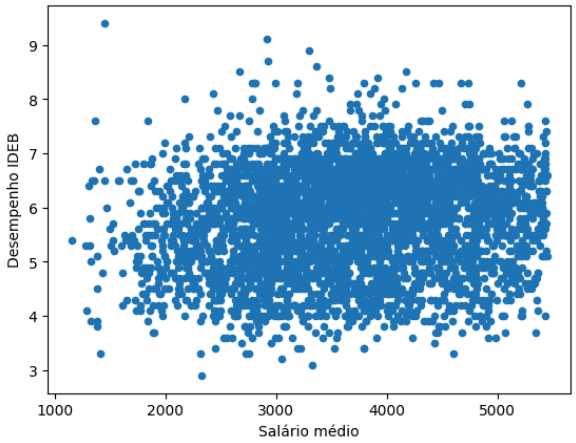
Conforme se observa na maioria dos estados, há uma correlação negativa, mesmo que pequena, entre desempenho no IDEB e horas trabalhadas pelos professores, o que sustenta a hipótese 2. Tambem é possível identificar em boa parte deles o distanciamento de uma correlação negativa conforme se passa de EF1 -> EF2 -> EM (Ensinos Fundamentais 1 e 2, e Ensino Médio).

Estados	Correlação - EF1	Correlação - EF2	Correlação - EM
Acre	-0.429	-0.116	0.109
Alagoas	-0.124	-0.053	-0.156
Amapá	-0.153	-0.629	0.395
Amazonas	0.139	0.129	-0.014
Bahia	-0.177	-0.114	-0.065
Ceará	-0.013	-0.012	-0.036
Espírito Santo	-0.160	-0.084	-0.296
Goiás	-0.041	0.000	-0.054
Maranhão	0.132	0.075	-0.199
Mato Grosso	-0.115	-0.099	0.105
Mato Grosso do Sul	-0.09	-0.027	-0.128
Minas Gerais	-0.142	-0.133	-0.039
Pará	0.000	-0.116	-0.074
Paraíba	-0.206	-0.117	-0.261
Paraná	0.141	-0.019	-0.009
Pernambuco	0.031	0.018	0.064
Piauí	-0.124	-0.168	-0.095
Rio de Janeiro	-0.236	-0.282	-0.145
Rio Grande do Norte	-0.069	0.116	-0.062
Rio Grande do Sul	0.028	-0.015	-0.094
Rondônia	0.237	0.019	0.021
Roraima	-0.837	-0.576	-0.380
Santa Catarina	-0.062	0.007	-0.180

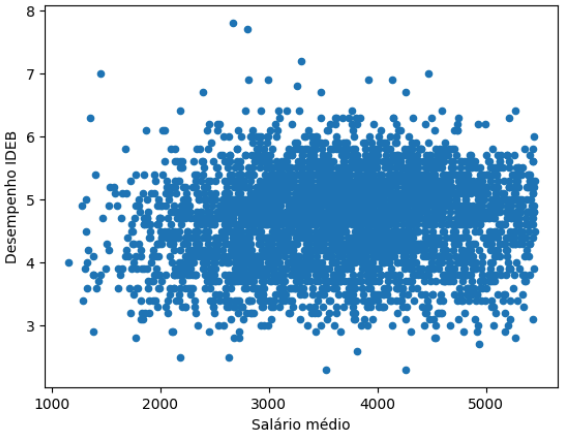
São Paulo	-0.023	0.011	-0.013
Sergipe	-0.153	0.129	-0.051
Tocantins	0.004	0.081	0.051

5.3. Resultados nacionais: Salário médio x Desempenho IDEB

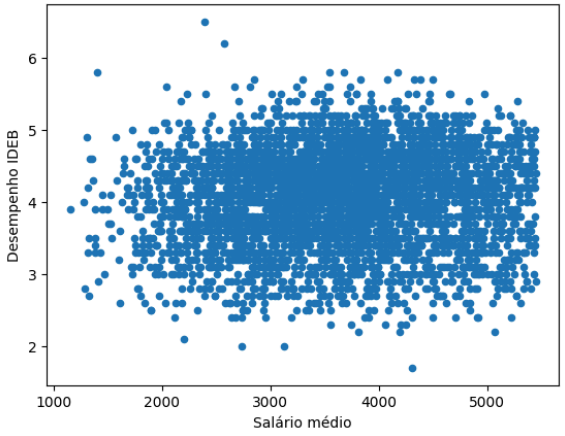
No que se refere à amostra de todos os municípios brasileiros, sem distinção de estados, a correlação entre salário e horas trabalhadas é positiva, como se espera pela hipótese 1. Porém, a intensidade dessa correlação também é baixa, como se observa pelo gráfico de dispersão apresentado, sendo retirado os outliers de salário. Nota-se aqui também uma sutil perda de correlação quanto mais avançada é a etapa da educação básica.



Ensino Fundamental 1 - c = 0,150



Ensino Fundamental 2 (c = 0.123)



Ensino Médio - c = 0.078

5.4. Resultados estaduais: Salário médio x Desempenho IDEB

Conforme se observa em boa parte dos estados, há uma correlação positiva, mesmo que pequena, entre desempenho no IDEB e salário dos professores, o que sustenta a hipótese 1. Tambeḿ é possível vislumbrar que boa parte deles o distanciamento de uma correlação positiva conforme se passa de EF1 -> EF2 -> EM (Ensinos Fundamentais 1 e 2, e Ensino Médio).

Estados	Correlação - EF1	Correlação - EF2	Correlação - EM
Acre	0.518	-0.025	-0.161
Alagoas	-0.031	-0.060	0.012
Amapá	0.271	-0.101	0.394
Amazonas	0.066	0.104	0.177
Bahia	0.021	-0.055	-0.075
Ceará	-0.154	-0.119	0.112
Espírito Santo	-0.090	-0.027	0.217
Goiás	0.070	0.008	0.043
Maranhão	-0.066	-0.103	0.106
Mato Grosso	0.239	0.104	-0.135
Mato Grosso do Sul	0.125	0.128	0.231
Minas Gerais	0.206	0.184	0.103
Pará	0.038	0.147	0.101
Paraíba	0.126	0.115	0.175
Paraná	0.038	0.046	0.057
Pernambuco	-0.113	-0.064	-0.020
Piauí	0.010	0.092	-0.103
Rio de Janeiro	0.174	0.294	0.108
Rio Grande do Norte	0.051	-0.087	-0.028
Rio Grande do Sul	0.022	0.035	0.030
Rondônia	-0.104	-0.019	0.045
Roraima	0.896	0.744	0.684
Santa Catarina	0.278	0.163	0.157
São Paulo	0.089	0.045	0.118
Sergipe	-0.083	-0.265	0.025
Tocantins	0.189	0.154	0.051

6. Conclusão

A partir dos dados e gráficos analisados, conclui-se que a carga horária e o salário dos professores impactam sim na qualidade de ensino dos alunos, e a maioria dos estados segue, mesmo que sutilmente, as hipóteses 1 e 2. Contudo, esse impacto não é tão significativo quanto esperado, sugerindo que o salário dos docentes esteja atrás, em uma escala de importância, de outros fatores envolvendo por exemplo a infraestrutura das escolas e municípios, que são limitantes físicos na qualidade de ensino. E quando estes limitantes físicos não estão presentes, o papel do professor passa a ser mais impactante, e consequentemente o salário que ele recebe vai influenciar mais significativamente no desempenho dos alunos.

Ressalta-se também uma “falha” na hipótese 1, uma vez que ele não leva em conta que, embora o salário maior de fato atraia professores mais qualificados, ele também pode atrair pessoas sem vocação para docência mas que se prestam ao cargo puramente devido ao dinheiro.

Além disso, é perceptível por meio da análise dos dados que os alunos da primeira etapa do Ensino Básico, o Ensino Fundamental 1, são os mais impactados com relação ao salário e carga horária dos professores e deduz-se que isso seja pelo fato de que é nesta etapa que os alunos são introduzidos à alfabetização e aos pilares do aprendizado no geral, sendo portanto mais sensíveis à qualidade de ensino de seus docentes.

7. Referências

Educação. Disponível em: <http://portal.mec.gov.br/component/tags/tag/31991 >. Acesso em: 24 jun. 2024.

Numeracy, maths and statistics - academic skills kit. Disponível em: <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html >. Acesso em: 24 jun. 2024.

SHARMA, T. Advantages of jupyter Notebook. Disponível em: <https://python.plainenglish.io/advantages-of-jupyter-notebook-5e41497bb4b2 >. Acesso em: 24 jun. 2024