

Análise Inteligente em Sistemas de 'Big Data' Influência salarial na criminalidade

Maya Gomes^{pg57891} and Rui Cerqueira^{pg57902}

Departamento de Engenharia informática, Universidade do Minho, Braga, Portugal.

Contributing authors: pg57891@alunos.uminho.pt; pg57902@alunos.uminho.pt;

Abstract

Este relatório apresenta um estudo sobre a relação entre crimes e salários nos Estados Unidos, com ênfase na comparação entre taxas de criminalidade estadual e os níveis salariais dos indivíduos, segmentados por nível educacional. O objetivo principal é investigar como as disparidades salariais e o nível de escolaridade influenciam a prevalência de crimes e índices de violência em diferentes estados ao longo dos anos. A análise utiliza dados estatísticos para identificar padrões e correlações entre esses dois fatores. Através de uma abordagem quantitativa, o estudo busca evidenciar como as variações no poder aquisitivo e no acesso à educação impactam os índices de criminalidade, oferecendo *insights* importantes para políticas públicas relacionadas à segurança e educação.

1 Contextualização

A relação entre a criminalidade e os salários é um tema amplamente debatido em estudos sociais e económicos. Diversas teorias sugerem sobre fatores como a desigualdade de renda, o acesso à educação e condições económicas que influenciam diretamente os índices de criminalidade numa sociedade. Nos Estados Unidos, esta questão adquire uma relevância particular devido às significativas disparidades regionais, tanto nos níveis de renda quanto na educação. As disparidades salariais e educacionais entre os estados refletem condições socioeconómicas locais que impactam diretamente as taxas de criminalidade.

Estados com menores níveis de renda e educação tendem a apresentar índices mais elevados de criminalidade, enquanto que aqueles com melhor acesso à educação e maior poder aquisitivo exibem menores taxas de violência. Além disso, a crise económica e as flutuações nos mercados de trabalho também desempenham um papel crucial, potencializando fatores como desemprego e desigualdade social. Este estudo tem como objetivo investigar como os salários, divididos por nível educacional, influenciam a criminalidade em diferentes estados ao longo do tempo, analisando as interações entre estas variáveis e as suas repercussões nas taxas de criminalidade e violência. [1]

2 Problema e Objetivos

A relação entre salários, educação e criminalidade é um fenómeno complexo e multifacetado que afeta diversas dimensões da sociedade. Nos Estados Unidos, as disparidades salariais e educacionais entre os estados geram contextos económicos distintos, os quais podem impactar diretamente os índices de criminalidade e violência. Em regiões com baixos níveis de renda e acesso restrito à educação, as taxas de criminalidade tendem a ser mais altas, enquanto estados com melhores condições económicas e educacionais apresentam índices de violência mais baixos. Apesar da relevância do tema, a interação exata entre essas variáveis, particularmente a correlação entre o nível salarial, o grau de escolaridade e as taxas de criminalidade, ainda carece de uma análise mais aprofundada e quantitativa, baseada em dados robustos e atualizados.

Este estudo tem como objetivo principal investigar como as variações nos salários, segmentadas por nível educacional, influenciam as taxas de criminalidade e violência em diferentes estados dos Estados Unidos ao longo dos anos. Para alcançar este objetivo, o estudo propõe-se a:

- Analisar a correlação entre salários por estado e taxas de criminalidade: Examinar como as flutuações nos salários impactam os índices de crimes violentos e não violentos em diferentes estados.
- Investigar o impacto do nível educacional na criminalidade: Analisar como o acesso à educação influencia as taxas de criminalidade, considerando a relação entre educação e oportunidades econômicas.
- Explorar a interação entre variáveis econômicas e demográficas: Estudar como o crescimento populacional, a densidade populacional e outras variáveis demográficas se associam aos dados de salários e criminalidade.

Com isso, espera-se oferecer uma análise detalhada e fundamentada da dinâmica entre os salários, a educação e os índices de criminalidade nos Estados Unidos, contribuindo para a compreensão de como as condições econômicas e educacionais podem ser um fator determinante na prevenção da violência e no fortalecimento da segurança pública. [2]

3 Estado da Arte

Palavras chaves: Crime, Salário, Educação, Fatores socioeconômicos, Ciência de dados, Análise de dados, Big Data

Aplicações da Análise Inteligente de Big Data nas Desigualdades socioeconômicas, Criminalidade e Educação

3.1 Análise do primeiro artigo: The effects of education on crime[3]

3.1.1 Resumo e conclusões obtidas

No artigo **The effects of education on crime** de W. Groot e H. M. van den Brink, os autores tentam encontrar uma relação entre a criminalidade e educação, analisando como o nível educacional influencia diferentes tipos de crime.

Os dados para este estudo tem origem num questionário feito em 1996 que entrevistou 2951 indivíduos com 15 anos ou mais, sendo que a parte das respostas foram preenchidas de forma anônima para obter respostas mais honestas e precisas. Através das respostas recolhidas deste questionário foi possível determinar que infrações e crimes menores como dirigir embriagado ou evasão da tarifa em transportes públicos são cometidos com mais frequência por pessoas com maior nível de educação, no entanto, a frequência de crimes mais graves como ameaçar e agredir indivíduos é maior entre pessoas com menor nível de educação.

Para analisar os dados, os autores aplicaram modelos probit, estimando a relação entre os anos de educação e a probabilidade de cometer diferentes tipos de crime em categorias como furtos, vandalismo, violência e fraude fiscal. Além disso foram utilizados modelos de regressão para avaliar a influência da educação sobre normas sociais e atitudes em relação à criminalidade. Para determinar o efeito da educação no comportamento criminal também foram utilizadas variáveis instrumentais (IV) para corrigir possíveis fatores de terceiros como país e gênero.

No final deste artigo foi possível concluir que é possível obter uma redução significativa nos custos sociais do crime por meio do investimento na educação, que a probabilidade de cometer crimes mais graves como roubo, vandalismo e agressões diminui com os anos de educação, no entanto a probabilidade de cometer fraudes fiscais aumenta.

Este artigo utiliza um conjunto de dados significativo, no entanto a amostra é restrita a indivíduos da Holanda e foi coletada em 1996, logo é difícil de ser comparada com os dias atuais em que os sistemas educacionais e economia sofreram grandes alterações. Logo para o nosso trabalho utilizaremos dados mais recentes e possivelmente de diferentes regiões para aumentar a validade da investigação.

3.2 Análise do segundo artigo: Exploration of the Hidden Influential Factors on Crime Activities: A Big Data Approach[4]

3.2.1 Resumo e conclusões obtidas

Este artigo propõe uma estrutura metodológica de "Big Data" para analisar os fatores da taxa de criminalidade. Ao longo do artigo são utilizados diversos métodos desde árvores de decisão de aumento de gradiente (GBDT), eliminação recursiva de características (RFE) e sistemas de informação geográfica (GIS) com o objetivo de filtrar, classificar e analisar características importantes. Os resultados do artigo demonstram que certas características são encontrados como fatores importantes na taxa de criminalidade e de agressões graves: o estado civil, origens afro-americanas, a situação econômica, a educação e os locais de interesse. Desta forma, conseguimos uma semelhança para e com o nosso estudo devido aos temas da situação econômica e da educação.

Apesar de este estudo utilizar um grande conjunto de dados sobre Nova York e mencionar a possível utilização de GIS para fusão de outros dados de outras cidades, a sua aplicação a outras cidades ou países pode ser limitada, pois a criminalidade é influenciada por fatores culturais e institucionais específicos de cada região e isso é algo que queremos ter em conta na nossa investigação.

3.3 Análise do terceiro artigo: Review of Economic Dynamics: Crime and the minimum wage[5]

3.3.1 Resumo e conclusões obtidas

O objetivo deste artigo é estabelecer a relação entre o salário mínimo e a taxa de criminalidade, e quantificar os efeitos dos mesmos. Desta forma, o estudo demonstra a relação entre a taxa agregada de criminalidade e o salário mínimo. Esta possui a forma de um U devido a dois efeitos opostos: o efeito do salário e o efeito do desemprego. Para além disso, o estudo demonstra que a taxa de criminalidade é minimizada quando o salário mínimo é 0,91 do salário médio dos jovens de 16 a 19 anos. No entanto, o salário mínimo que minimiza o crime não é o salário mínimo que maximiza o bem-estar, pois o salário mínimo afeta não apenas o crime, mas todos os resultados do mercado de trabalho. Para a análise, foram utilizadas certas técnicas como a regressão não paramétrica para testar a previsão do modelo, a estimação de coeficientes para quartis e análise da igualdade dos mesmos e também foram utilizadas funções quadráticas de forma a analisar as tendências temporais lineares e quadráticas.

Embora o artigo tenha fornecido uma análise significativa dos dados, há oportunidades para uma exploração mais profunda. O documento poderia ser completada com uma análise mais detalhada dos dados, utilizando técnicas adicionais de exploração para identificar padrões ocultos ou *insights* valiosos. Ao aplicar mais ferramentas de *insights* e técnicas avançadas de visualização, seria possível aprimorar a interpretação dos dados e fornecer uma compreensão mais robusta e precisa dos fenômenos observados. Este aprofundamento permitiria não apenas refinar os resultados, mas também descobrir novas variáveis ou correlações que poderiam impactar positivamente as conclusões do estudo.

3.4 Análise do quarto artigo: Crime and the minimum wage [6]

3.4.1 Resumo e conclusões obtidas

Este artigo considera possíveis conexões entre o crime e o mercado de trabalho, utilizando a introdução do Salário Mínimo Nacional no Reino Unido como uma forma de investigar o que acontece com o crime quando os salários dos trabalhadores de baixa renda recebem um aumento significativo. Os resultados fornecem um ângulo interessante em contraposição a outros trabalhos nesta área, que têm tendido a concluir que as oportunidades salariais são relevantes para o crime. Além disso, parece que a introdução do salário mínimo no mercado de trabalho do Reino Unido esteve diretamente ligada à redução do crime em áreas com trabalhadores de baixa remuneração. Para tal, o artigo relacionou as mudanças nas várias taxas de criminalidade antes e depois da introdução do salário mínimo com a proporção inicial de trabalhadores de baixa remuneração.

A pesquisa deste artigo concentra-se nas mudanças imediatas após a introdução do salário mínimo. No entanto, não há uma análise do impacto a longo prazo sobre a criminalidade e isso é algo que pretendemos investigar. Tal como os outros artigos este também se foca exclusivamente numa área neste caso o Reino Unido, utilizando dados de áreas específicas.

3.5 Justificação do Estudo

Apesar de já existir literatura que explora a relação entre criminalidade e salários assim como entre salários e educação o nosso trabalho pretende não só ampliar o escopo desta análise ao considerar as diferenças regionais e temporais nos Estados Unidos mas também explorar de que forma a criminalidade, salários e educação se relacionam entre si, oferecendo uma visão mais completa deste problema.

A principal lacuna identificada reside na ausência de análises que integrem simultaneamente as dimensões regionais, temporais e educacionais na relação entre salários e criminalidade. Grande parte dos estudos existentes aborda essas variáveis de forma isolada ou em contextos limitados, sem considerar as interações dinâmicas entre elas ao longo do tempo e entre diferentes estados. Assim, este trabalho propõe-se a preencher esta lacuna ao adotar uma abordagem abrangente e longitudinal que combina dados de múltiplos estados dos EUA, segmentando os salários por nível educacional. Com isso, pretende-se oferecer uma compreensão mais profunda e refinada sobre como esses fatores interagem na explicação das variações nas taxas de criminalidade.

4 Recolha de dados

De forma a proceder à realização do trabalho foram escolhidos três datasets. O primeiro dataset é referente aos crimes por estados, o segundo aos salários mínimos por estado e o último à educação. Os dois primeiros datasets serão interligados pelas colunas referentes ao ano e ao estado. O terceiro, por não possuir uma coluna referente ao estado mas somente uma referente ao ano, será utilizado para estudo adicional. Os conjuntos de dados encontram-se no formato CSV.

4.1 Dataset 1: Statewise crime in USA - K means Clustering [7]

Este dataset agrupa os diversos estados com base nos vários crimes reportados. Desta forma, procura indicar quais estados são mais ou menos propensos a crimes e quais tipos de crimes ocorrem nos mesmos.

Os dados abrangem os 50 estados dos Estados Unidos durante o período de 1960-2019. O dataset é composto por possui 21 colunas e 3116 linhas. Todas as colunas, com exceção da coluna referente ao estado, possuem dados numéricos.

Relativamente a valores ausentes, o dataset não possui nenhum.

É também importante referir que colunas referentes a taxas referem-se ao número de crimes reportados por 100.000 habitantes. Assim, algumas das colunas são as seguintes:

- **Data.Population:** O número de pessoas que viviam neste estado no momento do relatório.
- **Data.Rates.Violent.Assault:** O número de crimes onde alguém tentou iniciar um contacto prejudicial ou ofensivo com outra pessoa, ou realizou uma ameaça.
- **Data.Rates.Violent.Murder:** O número de crimes onde alguém cometeu um homicídio.
- **Data.Rates.Violent.Rape:** O número de crimes onde alguém cometeu algum tipo de agressão sexual.
- **Data.Rates.Violent.Robbery:** O número de crimes onde alguém roubou, ameaçando de força ou colocando a sua vítima com medo.

4.2 Dataset 2: US Minimum Wage by State from 1968 to 2020[8]

Este dataset contém informação acerca do salário mínimo federal e estatal por hora nos Estados Unidos que os trabalhadores podem receber para garantir que os cidadãos tenham uma qualidade de vida mínima.

Os dados deste dataset abrangem os 50 estados dos Estados Unidos entre 1968 e 2020 e possui 15 colunas e 2863 linhas, nas quais apenas a coluna do estado representa um valor categórico e as restantes valores numéricos.

Relativamente a valores ausentes, o dataset possui *missing values* em duas colunas: *Department.Of.Labor.Cleaned.Low.Value.2020.Dollars.* e *Department.Of.Labor.Cleaned.Low.Value* que possuem 415 e 430 *missing values* respetivamente. Para além disso este dataset foi previamente tratado para valores nulos nessas colunas, onde os valores foram substituídos por zeros. No entanto, iremos analisar este tratamento e procurar outras alternativas a esta medida.

O dataset utiliza o valor da moeda de 2020 como referência para os salários no período de 1968 a 2020 para comparação.

- **State.Minimum.Wage:** Salário mínimo do estado no dia 1 de Janeiro do Ano.
- **State.Minimum.Wage.2020.Dollars:** State.Minimum.Wage em 2020 dollars.
- **Federal.Minimum.Wage:** Salário mínimo federal no dia 1 de Janeiro do Ano.
- **Federal.Minimum.Wage.2020.Dollars:** Federal.Minimum.Wage em 2020 dollars.
- **Effective.Minimum.Wage:** O salário mínimo que é aplicado no estado no dia de Janeiro. Porque o salário mínimo federal tem efeito se o salário mínimo do estado for menor que o federal. Este é o maior dos dois.
- **Effective.Minimum.Wage.2020.Dollars:** Effective.Minimum.Wage em 2020 dollars.
- **CPI.Average:** O valor médio do *Consumer Price Index* do Ano.

4.3 Dataset 3: Wages by Education in the USA [9]

Este dataset fornece uma visão sobre os salários médios por hora dos trabalhadores nos EUA, desagregados pelo nível de escolaridade mais elevado alcançado. Este conjunto de dados abrange o período de 1973 a 2022. O dataset possui 61 colunas e 51 linhas. Relativamente a valores ausentes, o dataset não possui nenhum. Todas as colunas possuem dados numéricos. Algumas das colunas são as seguintes:

- **high_school:** Salário médio por hora para indivíduos com educação de ensino médio.
- **bachelor's degree:** Salário médio por hora para indivíduos com diploma de licenciatura.
- **advanced degree:** Salário médio por hora para indivíduos com um grau avançado (como mestrado ou doutorado).
- **men_high_school:** Salário médio por hora para homens com educação de ensino médio.
- **women_high_school:** Salário médio por hora para mulheres com educação de ensino médio.

5 Tratamento dos Datasets

5.1 Remoção de colunas

No dataset "minimum_wage" [8] decidimos remover duas colunas, a coluna *Footnote* pois não possui dados relevantes e a coluna *Department.Of.Labor.Uncleaned.Data* pois está dividida nas colunas *Department.Of.Labor.Cleaned.Low.Value* e *Department.Of.Labor.Cleaned.High.Value*.

5.2 Valores Nulos

Na análise dos diferentes datasets, observamos que apenas o dataset sobre o salário mínimo [8] possuía valores nulos, no entanto o autor deste dataset já tinha feito um tratamento do mesmo, tendo criado diferentes colunas e substituindo os valores nulos por 0. Para tratarmos estes valores decidimos substituí-los por NAN, e utilizamos interpolação linear para tratamento final. Escolhemos esta abordagem porque os dados incluem valores monetários de períodos antigos e atuais, os quais estão sujeitos a fatores como inflação, logo métodos como a média poderiam distorcer os dados, enquanto que a interpolação linear permite uma estimativa mais coerente desses valores.

5.2.1 Interpolação

De seguida, como referido anteriormente, decidimos tratar os *missing values* com a interpolação. A interpolação é um método utilizado para estimar valores ausentes num conjunto de dados com base nos valores existentes. Desta forma, ela pode revelar-se mais vantajosa do que simplesmente substituí-los por um determinado valor (moda, média, etc) ou ainda removê-los e perder dados. Decidimos utilizar a interpolação linear, onde os valores ausentes são calculados através de uma reta que conecta os pontos adjacentes. A mesma foi aplicada às seguintes colunas: *'State.Minimum.Wage'*, *'State.Minimum.Wage.2020.Dollars'*, *'Department.Of.Labor.Cleaned.Low.Value'*, *'Department.Of.Labor.Cleaned.Low.Value.2020.Dollars'*, *'Department.Of.Labor.Cleaned.High.Value'*, *'Department.Of.Labor.Cleaned.High.Value.2020.Dollars'*.

5.3 Outliers

No nosso caso de estudo não existe motivo para realizar o tratamento de outliers porque o mesmo resultaria numa perda de informação e queremos realizar uma análise dos possíveis extremos existentes tanto no caso da taxa de criminalidade como no caso da análise salarial. De facto, procuramos

precisamente analisar as diferenças entre os salários, ou seja, se houver um salário muito mais alto ou muito mais baixo que outros, num determinado estado, esta informação é relevante para o nosso estudo, pelo que, não pode ser tratado. É importante referir que o mesmo aplica-se nos *datasets* da criminalidade e da educação.

5.4 Normalização - Padronização de variáveis

A normalização e a padronização são técnicas de pré-processamento de dados muito utilizadas para transformar variáveis e colocá-las em escalas adequadas para análise ou modelagem. Neste trabalho decidimos usar a Padronização de variáveis. Esta técnica permite transformar os dados de forma a que a distribuição da variável tenha média 0 e desvio padrão 1. Achemos a Padronização mais adequada, pois queremos que cada variável tenha a mesma importância, independentemente da sua escala original. Esta técnica é também mais robusta que a normalização, nomeadamente na presença de *outliers* [10].

5.5 Merge

Para efetuar o *merge*, utilizamos as colunas "Year" e "State" dos *datasets* sobre *minimum_wage* [8] e *state_crime* [7] com o método *Outer Join*. Por fim, criamos outro *dataset* ao realizar o *merge* do *dataset* obtido com o *dataset* de *wages_by_education* [9] na coluna *year*, também utilizando o método *Outer Join*, ou seja, como este *dataset* não possuía dados relativos aos estados, o valor de cada ano foi duplicado para todos os estados no *dataset* final.

Para ser possível o *merge* entre o *dataset* resultante da primeira junção e o *dataset* de *wages_by_education* [9] necessitamos alterar o nome da coluna "year" para "Year" de forma a não haver conflitos.

5.5.1 Remoção de incompatibilidades

Ao realizar o *merge* dos nossos *dataset* reparamos na presença de determinados *missing values*. Esta ausência de valores foi devida ao facto de um *dataset* ter mais entradas na coluna "State" do que outro. De facto, reparamos que o *dataset* "wages_by_education" [9] possui algumas ilhas americanas na coluna "State", sendo que na verdade estas ilhas não são estados. Estas linhas não possuíam correspondência no outro *dataset*. Desta forma, decidimos remover as mesmas.

Para além disso, a coluna "Year" do *dataset* sobre o salário mínimo [8] possui valores de 1968 até 2020, no *dataset* dos crimes possuíamos valores de 1960 até 2019 e o *dataset* da educação possui valores de 1973 até 2022. Desta forma, também notamos alguma presença de valores ausentes devido a esta incompatibilidade. Assim, escolhemos manter somente os dados da faixa 1960-2019 no caso do primeiro *dataset* de *merge* e 1973-2019 no caso do segundo. Para tal, removemos novamente as linhas que possuíam valores ausentes.

5.6 PySpark

Para o tratamento também testamos a ferramenta de *Big Data* PySpark, para avaliar a viabilidade do uso da ferramenta, realizamos um teste para comparar a execução tradicional em python e a execução utilizando PySpark e obtivemos os seguintes resultados:

Sem PySpark:

- Tempo de execução: 0.27 segundos
- Memória utilizada: 10.91 MB
- Uso final de CPU: 7.00%

Com PySpark:

- Tempo de execução: 14.01 segundos
- Memória total utilizada: 20.00 MB
- Uso final de CPU: 15.90%

Com os resultados conseguimos observar que para os nossos *datasets*, o uso de PySpark não é aconselhado. Apenas o processo de inicialização da sessão do PySpark já acrescenta tempo de execução maior do que a execução tradicional completa, além disso a utilização de memória e CPU também é maior com o Spark. Este resultado é devido ao PySpark ser feito para lidar com grandes volumes de dados e não para *datasets* mais pequenos como o nosso.

5.7 Base de dados

5.7.1 Análise teórica

Característica	MongoDB	SQLite	PostgreSQL
Modelo de Dados	NoSQL (Documentos JSON/BSON)	Relacional (SQL)	Relacional (SQL)
Escalabilidade	Alta (suporta sharding e replicação)	Limitada (single-file, embutido)	Alta (com replicação, mas não sharding nativo)
Concorrência	Boa (alta simultaneidade)	Limitada (bloqueio de banco)	Excelente (controle de concorrência)
Desempenho de Leitura	Rápido para grandes volumes de dados	Rápido para bancos pequenos/médios	Rápido, especialmente em consultas complexas
Desempenho de Escrita	Rápido para dados não estruturados	Bom, mas limitado a pequena escala	Bom, mas mais lento devido à ACID
Suporte a Transações	Suporte transacional, mas limitado	Suporte básico a transações	Completo suporte a transações ACID
Flexibilidade	Muito flexível (esquema dinâmico)	Muito simples (esquema fixo)	Estrutura rígida (esquema fixo)
Caso de Uso Ideal	Dados semi-estruturados e escaláveis	Aplicações simples e locais	Aplicações corporativas e transacionais complexas

5.7.2 Benchmarking

Para cada tipo de bases de dados decidimos criar a base de dados e importar o nosso dataset. De seguida, de forma a avaliar o desempenho criamos as seguintes queries:

- Query 1: Consultar a média dos salários mínimos por estado
- Query 2: Consultar as taxas de criminalidade por estado
- Query 3: Consultar o total de crimes violentos por estado
- Query 4: Consultar o índice de CPI por estado
- Query 5: Consultar o nível educacional médio de homens e mulheres por estado
- Query 6: Selecionar todos os dados / todas as linhas

Os resultados foram os seguintes:

Query	Tempo (em segundos)
MongoDB - Query 1	0.0254
MongoDB - Query 2	0.0340
MongoDB - Query 3	0.0223
MongoDB - Query 4	0.0179
MongoDB - Query 5	0.0207
MongoDB - Query 6	0.1167
SQLite - Query 1	0.0025
SQLite - Query 2	0.0012
SQLite - Query 3	0.0013
SQLite - Query 4	0.0011
SQLite - Query 5	0.0014
SQLite - Query 6	0.0012
Postgres - Query 1	0.0026
Postgres - Query 2	0.0024
Postgres - Query 3	0.0016
Postgres - Query 4	0.0021
Postgres - Query 5	0.0026
Postgres - Query 6	0.0991

Table 1 Tempos das Queries para MongoDB, SQLite e Postgres

5.8 Escolha final

A escolha do MongoDB como solução para um cenário de Big Data, mesmo quando se trata de um dataset estruturado, pode ser justificada por várias razões teóricas, especialmente quando comparamos com outros bancos de dados como SQLite e PostgreSQL. Embora as bases de dados relacionais, como o PostgreSQL, sejam mais adequadas para dados altamente estruturados e relações complexas, o MongoDB oferece vantagens cruciais para cenários de Big Data que tornam sua escolha vantajosa.

Embora o MongoDB seja uma base de dados NoSQL e, portanto, normalmente seja mais associado a dados não estruturados, ele também é altamente eficiente no armazenamento e consulta de dados semiestruturados. Este pode armazenar grandes volumes de dados de forma eficiente, sem a necessidade de esquemas rígidos, o que pode ser um grande benefício para Big Data, onde a flexibilidade no modelo de dados é muitas vezes necessária. No nosso caso, revela-se mais simples importar os dados num cenário não relacional do que construir uma arquitetura estrutural para uma base de dados relacional.

O MongoDB usa um modelo de dados baseado em documentos (JSON), que é muito mais flexível do que o modelo relacional do PostgreSQL ou do SQLite. Esse modelo permite que os dados sejam armazenados de forma mais natural, com a capacidade de incluir diversos tipos de dados (incluindo dados binários, arrays, e subdocumentos) num único registro.

O MongoDB é também amplamente utilizado em conjunto com ferramentas de Big Data e análise de dados, como Apache Spark e Hadoop. No nosso caso, não iremos utilizar o Spark como foi referido anteriormente, mas pode surgir a necessidade de utilizar ferramentas de big data, sendo por isso, mais adequado a utilização do MongoDB.

Concluindo, a sua arquitetura orientada a documentos e a sua escalabilidade tornam-no uma escolha popular para aplicativos de Big Data que exigem processamento distribuído.

6 PowerBI

6.1 Dashboard

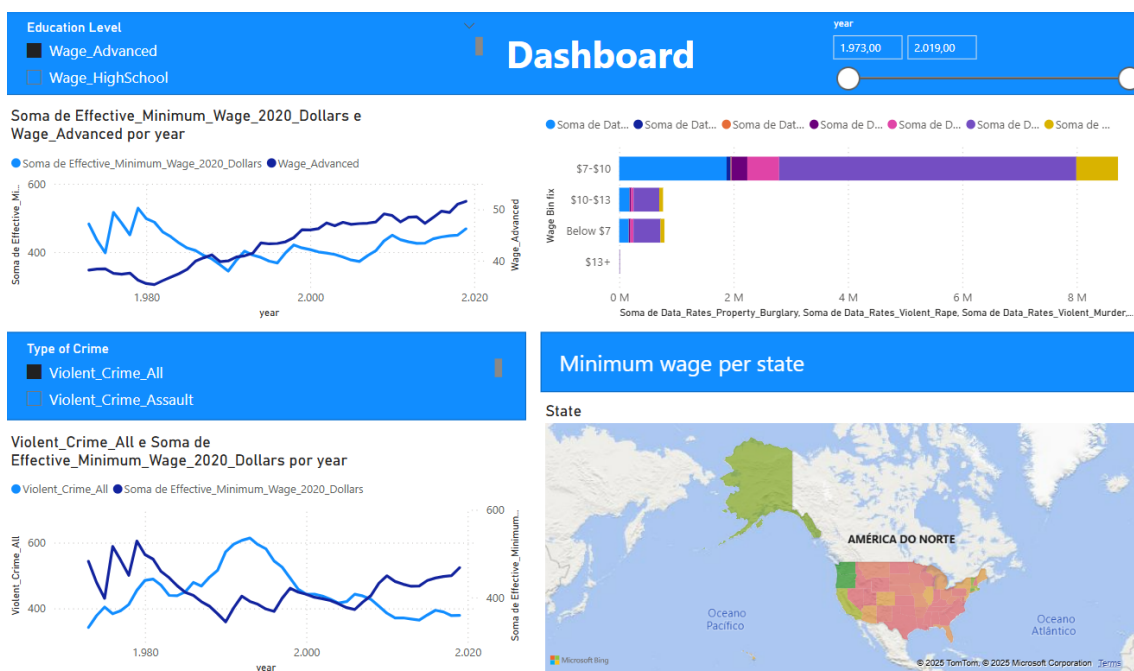


Fig. 1 Dashboard PowerBI

Na parte superior do *dashboard* foi criado um filtro que permite definir o ano para conseguirmos verificar a evolução deste problema e podermos observar os padrões ao longo dos anos. Para além disso, foi criado outro filtro baseado na coluna "State". Estes serão automaticamente aplicados a todos os gráficos do *dashboard* ao selecionarmos um estado a partir do mapa geográfico apresentado.

6.2 Gráfico 1: Salário Mínimo Efetivo e Salário por Nível Educacional Avançado

Para este gráfico, escolhemos criar um filtro que permitisse escolher uma coluna referente a níveis de educação avançados (como vemos na parte superior da Figura 2). A coluna escolhida será relacionada com a coluna "Effective.Minimum.Wage.2020.Dollars" que representa o salário mínimo em \$ do ano de 2020. Temos também em consideração a coluna "Year" no gráfico de forma a visualizarmos a evolução ao longo dos anos.

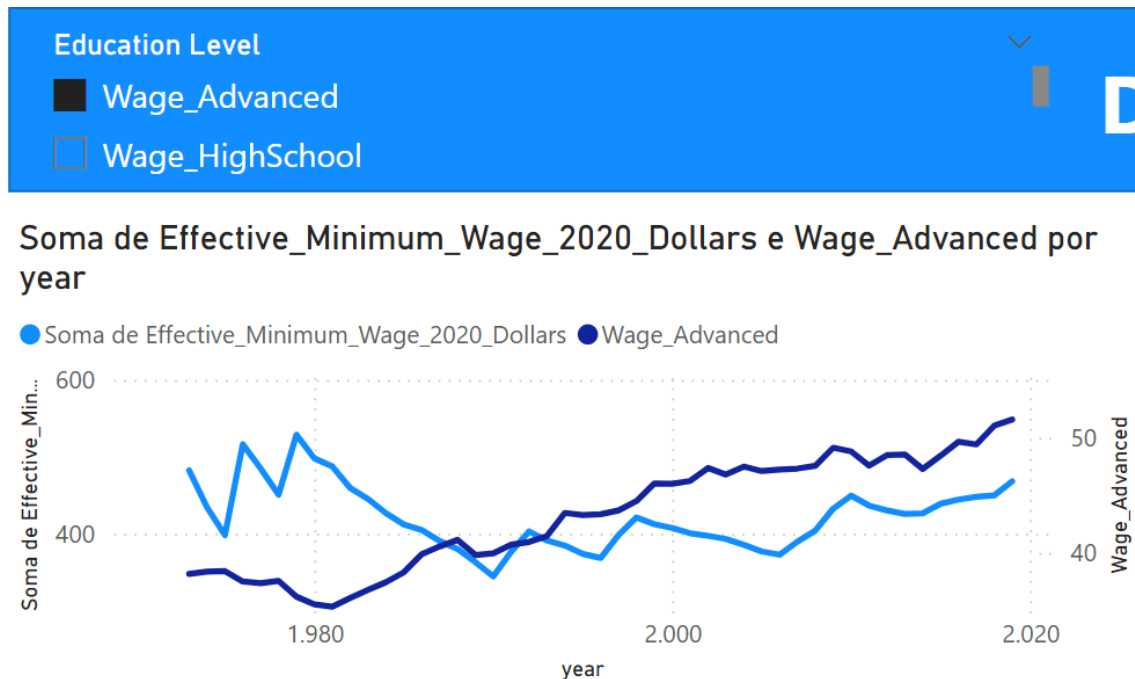


Fig. 2 Gráfico 1: Salário Mínimo Efetivo e Salário por Nível Educacional Avançado

Este gráfico de linha mostra a evolução temporal (de 1980 a 2020) do salário mínimo efetivo em dólares de 2020 (linha azul claro) comparado com os salários de pessoas com educação avançada (linha azul escuro). Observa-se que:

- O salário mínimo efetivo teve oscilações significativas, com pico em torno de 1980, seguido de declínio até aproximadamente 2000
- Os salários para pessoas com educação avançada mostram tendência de crescimento constante desde 1990
- Existe uma disparidade crescente entre salários de trabalhadores com educação avançada e o salário mínimo efetivo

6.3 Gráfico 2: Distribuição de Crimes por Faixa Salarial

Neste gráfico, decidimos relacionar a coluna do "Effective.Minimum.Wage.2020.Dollars", criando um bin da mesma, com diversas colunas de crime com o objetivo de verificar a relação entre os diversos salários e os crimes.

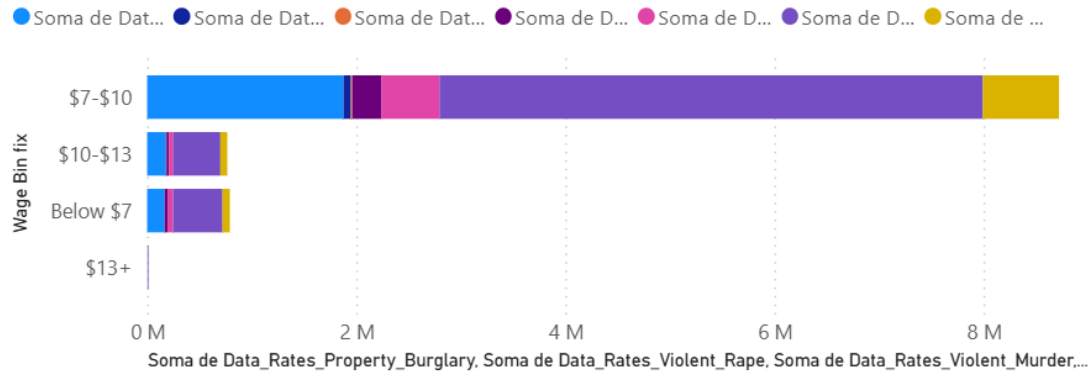


Fig. 3 Gráfico 2: Distribuição de Crimes por Faixa Salarial

Este gráfico de barras horizontais mostra a relação entre faixas salariais e diversas categorias de crimes:

- A faixa salarial de 7–10 apresenta o maior número de ocorrências criminais
- As faixas salariais mais altas (\$13+) têm significativamente menos registros de crimes
- Observa-se uma correlação inversa entre nível salarial e incidência criminal

6.4 Gráfico 3: Mapa de Salário Mínimo por Estado

Neste gráfico quisemos elaborar um gráfico sobre o mapa dos EUA para termos uma visualização generalizada dos salários mínimos do país.

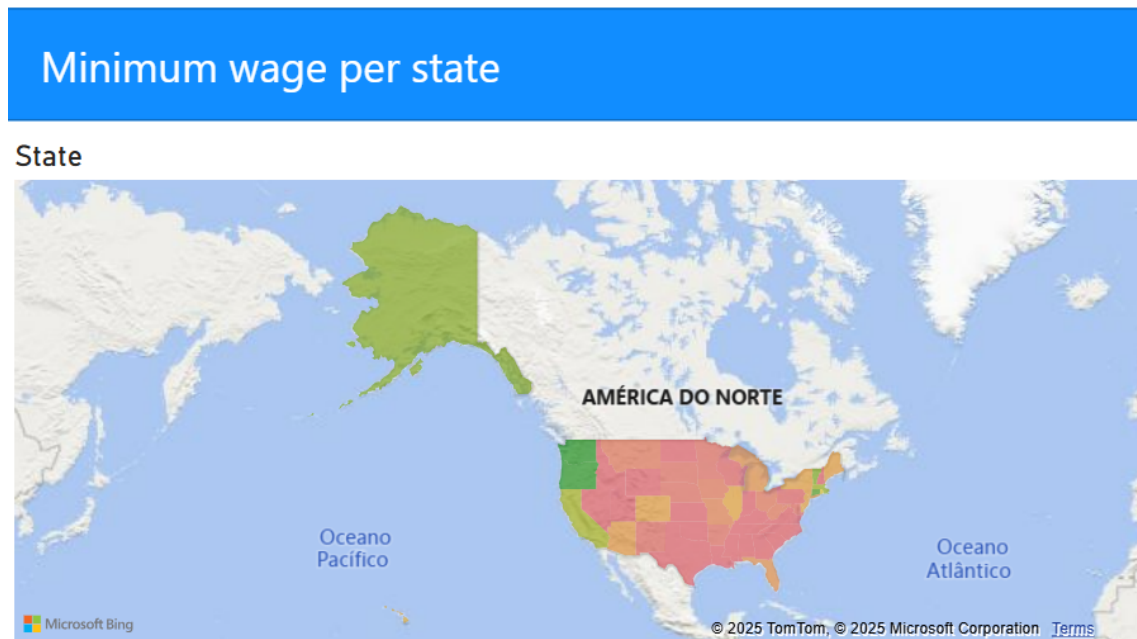


Fig. 4 Gráfico 3: Mapa de Salário Mínimo por Estado

O mapa mostra a distribuição geográfica do salário mínimo nos EUA:

- Estados a verde (como Alasca e Washington) parecem ter salários mínimos mais altos
- Estados a vermelho/rosa têm salários mínimos mais baixos
- Existe uma variação regional clara, com estados do oeste e nordeste geralmente tendo salários mínimos mais elevados

6.5 Gráfico 4: Criminalidade Violenta e Salário Mínimo

Neste gráfico, decidimos criar um filtro sobre os diversos tipos de crimes que existem (nomeadamente violentos e contra a propriedade) e comparar com o salário mínimo. A nossa ideia é analisar se Estados mais pobres têm crimes mais violentos ou contra a propriedade que estados mais ricos. No fundo, procuramos separar crimes violentos como homicídios, agressões físicas e assaltos à mão armada dos crimes patrimoniais como furtos e roubos sem violência, investigando se a pobreza está mais correlacionada com "crimes de desespero", ou crimes de património, ou seja, onde pessoas em situação de vulnerabilidade económica recorrem à violência ou furto como manifestação extrema de sua condição socio-económica.

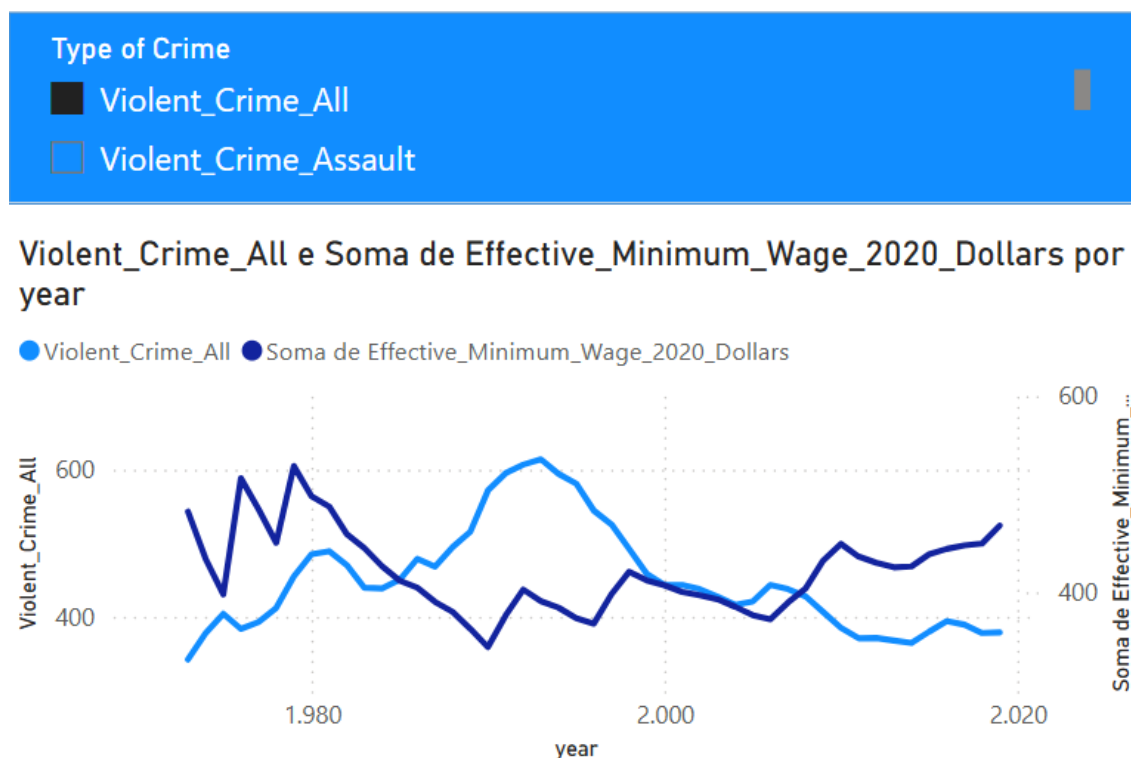


Fig. 5 Gráfico 4: Criminalidade Violenta e Salário Mínimo

Este gráfico analisa a relação entre criminalidade violenta (linha azul clara) e salário mínimo efetivo (linha azul escura) ao longo do tempo:

- Entre 1980-2000, há períodos em que o aumento da criminalidade coincide com a queda do salário mínimo
- Após 2000, enquanto o salário mínimo efetivo se manteve relativamente estável com leve aumento, a criminalidade violenta apresentou tendência de queda
- Isso sugere uma possível correlação negativa entre salário mínimo adequado e taxas de criminalidade

7 Conclusões gerais sobre o Dashboard

Com a nossa análise no PowerBi, conseguimos concluir que existe uma correlação visível entre salários baixos e maior incidência de crimes violentos. Estados com salários mínimos mais altos tendem a apresentar menores taxas de criminalidade, ou seja, as políticas de salário mínimo parecem ter impacto significativo nos indicadores sociais, incluindo taxas de criminalidade. Conseguimos também analisar que o crescimento da disparidade salarial entre trabalhadores com educação avançada e aqueles com salário mínimo verifica-se.

O *dashboard* apresenta uma análise complexa das relações entre fatores económicos (salários), educacionais e sociais (criminalidade) nos Estados Unidos, demonstrando como essas variáveis se influenciam mutuamente ao longo do tempo e nos diversos estados do país.

8 Conclusão

Com a realização deste trabalho fomos capazes de explorar a relação entre o valor do salário mínimo, os diferentes salários por nível de educação e o crime. Consideramos ter alcançado os objetivos pretendidos com a nossa investigação, com o dashboard construído conseguimos confirmar que a taxa de crimes tinha tendência a diminuir com o aumento do salário mínimo e os diferentes tipos de crime nos quais o aumento do salário mínimo tinha o maior efeito. Estes resultados variavam um pouco em certos estados mas isso pode ser devido a outros fatores fora da nossa área de estudo.

Concluindo este estudo permitiu-nos perceber a correlação entre o aumento do salário mínimo e redução da taxa de criminalidade. A análise revelou que determinados tipos de crime, especialmente os natureza económica, parecem mais sensíveis às variações no salário mínimo. Para investigações futuras, seria ideal alargar o escopo da análise para incluir variáveis adicionais para obter uma visão maior do problema.

References

- [1] Medicine, N.L.: Educational pathways and change in crime between adolescence and early adulthood. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5365088/> (2017)
- [2] Lance Lochner, U.o.W.O.L.O.C. Department of Economics: Education and crime. https://economics.uwo.ca/people/lochner/docs/edu_crime_lochner.pdf (2020)
- [3] W. Groota, . b, Brink b, H.M.: The effects of education on crime. <https://www.tandfonline.com/doi/epdf/10.1080/00036840701604412?needAccess=true>
- [4] JIANMING ZHOU¹, J.J.M.. ZHENG LI ², JIANG³, F.: Exploration of the hidden influential factors on crime activities: A big data approach. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=9143124>
- [5] Christine Brauna, b.: Review of economic dynamics: Crime and the minimum wage. <https://doi.org/10.1016/j.red.2019.02.002>
- [6] Hansen*, K., Machin**, S.: Crime and the minimum wage¹. <https://personal.lse.ac.uk/machin/pdf/Crime>
- [7] Kaggle: Statewise crime in usa - k means clustering. <https://www.kaggle.com/datasets/vikramamin/statewise-crime-in-usa-k-means-clustering>
- [8] Kaggle: Us minimum wage by state from 1968 to 2020. <https://www.kaggle.com/datasets/lislejoem/us-minimum-wage-by-state-from-1968-to-2017>
- [9] Kaggle: Wages by education in the usa (1973-2022). <https://www.kaggle.com/datasets/asaniczka/wages-by-education-in-the-usa-1973-2022>
- [10] Shaibu, S.: Normalização vs. padronização: Como saber a diferença. <https://www.datacamp.com/pt/tutorial/normalization-vs-standardization>