Universidade do Minho
Escola de Engenharia
Mestrado em Engenharia Informática

Unidade Curricular de Análise Inteligente em Sistemas de "Big Data"

Ano Letivo de 2022/2023

Impacto da Inflação na Educação

António Luís de Macedo Fernandes PG50229 João Silva Torres PG50499
Mário Jorge Amaral Pinto Correia PG51246

4 de junho de 2023

Data de Receção	
Responsável	
Avalição	
Observações	

Resumo

Este relatório foi feito de forma faseada e pretende esclarecer todos os pontos que foram propostos pela equipa docente.

No âmbito do desenvolvimento do projeto da unidade curricular de Análise Inteligente em Sistemas de "Big Data", foi-nos fornecido, pela equipa docente, um dataset sobre inflação e foi-nos proposto selecionar dois ou três datasets complementares.

Dito isto, o primeiro inclui seis medidas de inflação: *Headline consumer price index (CPI) inflation*, *Food CPI inflation*, *Energy CPI inflation*, *Core CPI inflation*, *Producer price index inflation* e *Gross domestic product deflator*. Optamos por dar mais ênfase ao *Headline consumer price index (CPI) inflation*, devido à natureza do nosso tema. Quanto aos datasets complementares, selecionamos três. O primeiro é o *PIB-GDP Global since 1960-2021* que ilustra o valor do produto interno bruto para vários países ao longo dos anos; O segundo é o *Government expenditure on education, total (% of GDP)* que consiste num indicador que diz qual a percentagem do PIB de um determinado país utilizado em educação, pelo governo; O terceiro é o *Literacy rate, adult total* que indica, para cada país, qual a percentagem de alfabetização de indivíduos (masculino e feminino).

Assim, o tema no nosso trabalho é calcular métricas que relacionem o aparecimento da inflação com o produto interno bruto de diversos países e como isto afetou o setor financeiro de diferentes áreas da educação. Através do dataset fornecido da inflação (CPI) vamos relacioná-lo com o PIB a nível mundial, com a percentagem que o governo investe na educação (também a nível mundial) e com o nível de alfabetização.

Ao longo deste relatório vamos também explicar pormenorizadamente todos os datasets e o seu respetivo processamento e tratamento. De seguida, iremos abordar como foi feito o armazenamento dos dados e a visualização dos mesmos. Por fim, temos uma secção onde discutimos os resultados obtidos e outra com conclusões e trabalho futuro.

Em conclusão, a equipa pensa que o nosso projeto respeita as especificações solicitadas e que, de uma maneira geral, foi realizado um bom trabalho.

Área de Aplicação: Calcular métricas que relacionem o aparecimento da inflação com o produto interno bruto de diversos países e como afetou o setor financeiro de diferentes áreas da educação.

Palavras-Chave: Bases de Dados, Big Data, Inflação, PIB, Educação, $Python$, $MongoDB$, $PowerBI$.

Índice

Lista de Figuras

Lista de Tabelas

1 Introdução

Nesta primeira secção será feito um enquadramento teórico do nosso trabalho, ou seja, a relação do mesmo com a atualidade. Deste modo, justificaremos o que nos motivou a desenvolvê-lo.

Será também abordado todo o planeamento e a gestão de tempo na realização do projeto, bem como todos os recursos utilizados para o bom funcionamento do mesmo.

1.1 Contextualização

A inflação é um fenómeno económico que existe desde a antiguidade. Porém, atualmente é caracterizada pelo aumento contínuo e generalizado dos preços dos bens e serviços na economia.

A inflação em larga escala, como a conhecemos hoje, teve início após a Segunda Guerra Mundial. Durante as décadas de 1950 e 1960, muitos países adotaram políticas que incentivaram o crescimento económico e à criação de empregos. Estas políticas eram caracterizadas por ter altos níveis de gastos públicos e baixas taxas de juro. Isto levou a um aumento de dinheiro em circulação na economia.

Esse aumento na oferta de dinheiro levou a um aumento da procura por bens e serviços, o que, por sua vez, levou a um aumento nos preços. Em muitos países, essa situação foi agravada por choques externos, como a crise do petróleo de 1973 e a recessão global que se seguiu. Recentemente, acontecimentos como o aparecimento do covid19, que levou a uma crise pandémica, e a invasão russa na Ucrânia agravaram a inflação.

Assim, este fenómeno monetário pode ter um impacto significativo no setor educacional. As altas taxas de inflação podem afetar negativamente a economia como um todo, reduzindo a capacidade do governo de investir em educação e outros serviços públicos. O aumento dos preços pode levar a um aumento na procura por salários mais altos, o que pode levar a uma redução no orçamento disponível para outras áreas, incluindo a educação. Por outro lado, se o governo pode manter a inflação sob controlo, pode haver mais recursos disponíveis para investir em programas educacionais e em infraestrutura escolar.

Este trabalho vem então ajudar a tirar conclusões quanto a este problema. De facto, este visa calcular métricas que relacionem o aparecimento da inflação com o produto interno bruto de diversos países e assim verificar como afetou o setor financeiro de diferentes áreas da educação.

1.2 Apresentação do Caso de Estudo

O caso de estudo deste artigo científico é o cálculo de métricas que relacionem o aparecimento da inflação com o produto interno bruto de diversos países e como afetou o setor da educação. Com isto, pretendemos pegar no tema principal, a inflação, e ver como pode afetar a educação. Para tal, utilizamos métricas como literacia e inflação de bens essenciais e como estes podem afetar diretamente a educação.

1.3 Motivação e Objetivos

A motivação por trás deste projeto surge da importância que a educação tem na sociedade. De facto, esta pode ser afetada por diferentes fatores económicos, tal como a inflação, que pode ter um grande impacto neste setor. Por isso, este estudo concentra-se em entender como este desiquilíbrio económico afeta diretamente a literacia e outros indicadores educacionais.

Os objetivos deste estudo são múltiplos. Em primeiro lugar, pretendemos calcular métricas que relacionem o aparecimento da inflação com o produto interno bruto de diversos países. Em segundo lugar, queremos entender como a inflação afeta o setor da educação e a literacia em particular. Para alcançar estes objetivos, serão utilizadas diferentes medidas, incluindo a inflação de bens essenciais.

Assim, esperamos que este estudo ajude a aumentar a consciencialização sobre a importância da estabilidade económica para a área da educação.

1.4 Estrutura do Relatório

Na primeira secção, "Introdução", é feita uma contextualização do tema e é apresentado o caso de estudo, sendo também definidos os objetivos a serem alcançados.

A segunda secção, "Estado da Arte", apresenta uma revisão bibliográfica sobre o tema em questão, com o objetivo de situar o leitor sobre as pesquisas e descobertas já realizadas sobre o assunto.

Na terceira secção, "Arquitetura", os datasets escolhidos são explicados, bem como o tratamento feito aos mesmos. São também explicados os processos de processamento e armazenamento dos dados e a visualização dos resultados obtidos.

Na quarta secção, "Resultados e Discussão", são apresentados os resultados obtidos a partir da análise dos dados, seguidos de uma discussão crítica acerca dos mesmos.

Por fim, na secção "Conclusões e Trabalho Futuro", são apresentadas as principais conclusões

a partir dos resultados obtidos e são sugeridos possíveis caminhos para futuras pesquisas sobre o tema.

O relatório inclui também uma lista de siglas e acrónimos utilizados, bem como anexos que possam ajudar na compreensão dos resultados apresentados.

2 Estado da Arte

Um projeto recente é o estudo "Big Data for Monitoring Education Inflation", realizado em 2019 por Julia Jessen e Florian Ramsauer. Este estudo utilizou técnicas de Big Data para monitorizar a inflação na educação em países subdesenvolvidos e/ou de baixa renda. Os resultados mostraram que a inflação na educação é um problema significativo em muitos desses países, afetando a acessibilidade e a qualidade da mesma. Além disso, o estudo destacou a importância do uso de dados em tempo real para lidar com o problema.

Outro projeto relevante é o "Inflation and Its Impact on Education in Latin America", realizado por Ana Corbacho e outros autores em 2015. Este estudo utilizou dados de 17 países da América Latina para analisar a relação entre inflação e investimento em educação. Os resultados mostraram que este fenómeno económica tem um impacto negativo no investimento nesta área, afetando a qualidade da mesma e a igualdade de oportunidades. Além disso, o estudo também destacou a importância de políticas públicas para combater a inflação e melhorar a qualidade da educação.

Outro estudo interessante é o "Inflation and Education: Evidence from Indonesia", realizado por Harry Patrinos e outros autores em 2016. Este estudo utilizou dados da Indonésia para analisar o impacto da inflação no desempenho dos alunos em testes padronizados. Os resultados mostraram que a inflação tem um impacto negativo no desempenho dos alunos, especialmente aqueles que pertencem a famílias com rendimentos baixos. Além disso, o estudo também destacou a importância da transparência e da gestão eficiente dos recursos financeiros para garantir a qualidade da educação.

Por fim, o estudo "Exploring the Effects of Inflation on Education using Big Data Analytics", realizado por M. Raquibul Islam e outros autores em 2020, utiliza técnicas de Big Data para analisar o impacto da inflação na educação em Bangladesh. O estudo utiliza dados de diversas fontes, incluindo pesquisas nacionais e internacionais, para avaliar o impacto da inflação no desempenho dos alunos e na igualdade de oportunidades na educação. Os resultados mostraram que esta alteração económica tem um impacto significativo no desempenho dos alunos e na qualidade da educação em Bangladesh.

Em resumo, os projetos de Big Data nesta área mostram a importância de compreender e monitorizar a inflação para garantir a qualidade e a acessibilidade da educação. Os estudos destacam que este fenómeno tem um impacto significativo no investimento em educação, no desempenho dos alunos e na igualdade de oportunidades na educação. Além disso, também mostram a importância do uso de técnicas de Big Data para monitorizar e avaliar a inflação em tempo real e para informar políticas públicas para melhorar a qualidade da educação.

Em conclusão, os projetos apresentados nesta secção mostram a relevância dos estudos nesta área. Os resultados destacam a necessidade de políticas públicas para combater a inflação e melhorar a qualidade da educação em todo o mundo. A utilização de técnicas de Big Data pode fornecer informações relevantes para os responsáveis pelas políticas. Desta forma ajuda a garantir um futuro melhor para as gerações futuras.

3 Arquitetura

Nesta secção serão apresentados os conjuntos de dados utilizados nesta análise, bem como o seu processamento, armazenamento e visualização. Serão utilizados quatro conjuntos de dados: *CPI inflation*, *PIB-GDP global since 1960-2021*, *Government expenditure* e *Literacy rate, adult total*. Estes conjuntos de dados serão então processados e armazenados para possibilitar a sua análise e visualização. Estes processos serão importantes para entender as tendências dos dados ao longo do tempo e fornecer *insights* sobre as mudanças nos indicadores económicos e a sua relação com a educação.

3.1 Datasets

Nesta subsecção, serão apresentados os quatro conjuntos de dados utilizados nesta análise. (*Data Acquisition*) O primeiro foi dado pelos docentes e os restantes foram selecionados em diferentes fontes. O formato de todos os datasets utilizados é CSV.

3.1.1 CPI inflation

(Ver fonte)

Como primeiro dataset selecionamos o *CPI inflation* que corresponde ao *Headline consumer price index* de vários países em diferentes anos. A fonte deste dataset é o The World Bank. Este abrange 209 países no período de 1970 a 2022. Por este motivo, contém então 209 entradas e 52(+2) colunas.

- Country code - código de 3 letras que corresponde a um país.
- IMF Country code - código de 3 números que corresponde a um país.
- Country - nome do país por extenso.
- Indicator Type - que tipo de indicador se trata, neste caso, é Inflation.
- Series Name - nome do indicador, Headline Consumer Price Inflation.
- Ano (1970-2022) - Anos entre 1970 e 2022.

Em relação ao tamanho, ocupa 6,46 MBytes de memória.

O *Headline consumer price index* consiste numa medida estatística que mede as variações médias de preços de um determinado conjunto de bens e serviços comprados por famílias e indivíduos. Isto é, para cada país é atribuído um valor que representa a inflação destes determinados produtos num determinado ano. Estes podem ser positivos ou negativos e indicam se os valores destes bens, para um determinado valor base, foram muito inflacionados ou não.

Como dito anteriormente, através da variação destes produtos e bens essenciais, conseguimos assim relacionar com um dos tema principais do nosso trabalho, a inflação. Para além disso, através deste indicador podemos concluir sobre o custo de vida da população nos diferentes anos e assim relacionar também diretamente com a educação. Pois se o custo destes bens está inflacionado, então o custo de vida aumenta e diretamente os gastos associados ao setor educativo também aumentam.

3.1.2 PIB-GDP Global since 1960-2021

(Ver fonte)

O segundo dataset é o *PIB-GDP Global by countries since 1960 to 2021* e irá ilustrar o valor do produto interno bruto para vários países ao longo dos anos. A fonte é o Kaggle. Este estudo contém dois datasets, mas o que iremos utilizar é o *countries_gdp_hist*, que contém 13025 entradas e as seguintes features:

- country_code - código de 3 letras que corresponde a um país.
- region_name - nome da região (Americas, ..).
- sub_region_name - nome da subregião.
- intermediate_region - nome da região intermédia.
- country_name - nome do país por extenso.
- income_group - grupo que indica os níveis de renda por capita.
- year - ano em questão.
- total_gdp - valor do pib.
- total_gdp_million - valor do pib por milhão.

Em relação ao tamanho, ocupa 378 kBytes de memória.

As entradas corresponderão aos valores do PIB nos diferentes anos por country codes.

Com este dataset pretendemos analisar o *income* total dos vários países, com o objetivo de relacionar com a taxa de inflação do dataset anterior e, assim, retirar conclusões tendo em conta a temática geral do estudo: a educação.

3.1.3 Government Expenditure on Education

(Ver fonte)

O terceiro dataset selecionado é *Government expenditure on education, total (% of GDP)*. Este consiste num indicador selecionado no The World Bank. Dito isto, indica, para cada país, qual a percentagem do PIB utilizada em educação. Por este motivo, cada linha corresponderá aos diferentes países (ver o números).

- Country Name - nome do país por extenso.
- Country Code - código de 3 letras que corresponde a um país.
- Indicator Name - nome do indicador a tratar
- Indicator Code - código do indicador a tratar
- Ano (1960-2022) - Anos entre 1960 e 2022.

Utilizamos este dataset de forma a analisar qual o investimento dos diferentes países no que toca a educação e assim poder retirar conclusões sobre o aumento/diminuição do custo escolar.

Em relação ao tamanho, ocupa 168 kBytes de memória, quando extraído.

3.1.4 Literacy rate, adult total

(Ver fonte)

O último dataset é o *Literacy rate, adult total (% of people ages 15 and above)*. Este dataset consiste noutro indicador selecionado no The World Bank. Este indicador irá, para cada país, indicar qual a percentagem de indivíduos (masculino e feminino) literados. O termo literados consiste na taxa de alfabetização, nomeadamente ter a capacidade de ler e escrever.

- Country Name - nome do país por extenso.
- Country Code - código de 3 letras que corresponde a um país.
- Indicator Name - nome do indicador a tratar
- Indicator Code - código do indicador a tratar

- Ano (1960-2022) - Anos entre 1960 e 2022.

Este dataset poderá fornecer consequências do possível aumento da inflação na educação que se reflete na educação dos indiviuos dos diferentes países.

Em relação ao tamanho, ocupa 120 kBytes de memória, quando extraído.

3.2 Processamento

3.2.1 Pré-processamento de dados

Nesta subsecção referimos qual o tratamento de dados realizado sobre os dados anteriormente referidos.

Missing Values

De forma a tratar dos valores em falta no nosso conjunto de dataset, inicialmente começamos por identificar quais as colunas que continham estes valores. Verificámos que os mesmos incidam principalmente nos indicadores do *the World Bank* escolhidos, em que existiam anos que apresentavam *missing values*. Dito isto, lidamos com estes valores, alterando-os para a média da váriavel correspondente ao país, de forma a preencher os valores em falta.

Feature Selection

De forma a simplificar a interpretação dos nossos datasets fizemos a seleção dos atributos mais pertinentes de cada um. As colunas que selecionamos foram:

- *CPI inflation*: country code, year e o valor de CPI;
- *PIB global*: country code, year e total_gdp;
- *Government Expenditure on Education*: country code, year e o valor de gasto na educação do governo;
- *Literacy rate*: country code, year e literacy_rate.

3.2.2 Merge de Datasets

Após o tratamento de dados, realizamos o merge dos datasets escolhidos através do country code, uma vez que esta coluna está presente em todos os escolhidos. Sendo assim a key para

o processo de junção dos datasets.

Assim, ao realizar este processo, as entradas do dataset final irão consistir nos diferentes country codes para cada ano, isto é, cada país da coluna country code.

No final do processo de Merge, obtemos como resultado um dataset com diferentes dados, sendo estes, neste caso, o CPI, o PIB, o gasto na educação de um governo e a taxa de literacia. A estes dados estarão associados o ano e o país a que se referem.

3.2.3 Parquet

Após ter sido realizado o merge dos nossos datasets para um dataset final, realizamos a conversão do mesmo para o formato parquet. O tempo de execução desta operação foi de 0,5 segundos. As principais vantagens deste formato são:

- Armazenamento eficiente - Economiza espaço de armazenamento utilizando compactação em coluna altamente eficiente e esquemas de codificação flexíveis para colunas com diferentes tipos de dados;
- Maior Desempenho e Transferência de dados - Faz uso de técnicas como salto de dados, em que as consultas que buscam valores de coluna específicos, não precisam de ler toda a linha de dados;
- Maior compatibilidade com ferramentas de Big Data;

3.2.4 Ferramentas - Tratamento de Dados

Em termos de processamento utilizamos a linguaguem *Python* e as bibliotecas *Pandas* e *Numpy*.

- Pandas - é uma biblioteca de análise de dados de alto desempenho que fornece estruturas de dados fáceis de utilizar e ferramentas de manipulação de dados. É construído sobre a biblioteca NumPy e é utilizado em tarefas de limpeza, preparação e análise de dados.
- Numpy - (Numerical Python) é uma biblioteca fundamental para computação numérica em Python. Fornece uma estrutura de matriz multidimensional eficiente e várias funções para realizar operações numéricas avançadas nomeadamente utilizadas em modelos de aprendizagem.

Estas bibliotecas foram utilizadas para o processo de aquisição de dados e para o tratamento dos diferentes datasets assim como a junção dos mesmos.

3.3 Armazenamento

Nesta secção iremos abordar a próxima fase do nosso projeto, que é referente ao *Data Storage*. Inicialmente iremos fazer uma análise comparativa das diversas ferramentas sugeridas para o armazenamento de dados e qual a utilizada para o nosso trabalho. Para além disso, também iremos indicar como foi feito todo o processo e qual a estratégia de extração de conhecimento .

3.3.1 Ferramentas - Armazenamento de Dados

As ferramentas que foram consideradas para esta fase foram:

- Mongo DB - É uma base de dados documental NoSQL e uma ferramenta open source utilizada para o armazenamento de elevados volumes de dados. Utiliza uma linguagem não estruturada para queries e consultas sobre os dados.

  As principais vantagens são:

  - Elevada performance;
  - Alta disponibilidade dos dados;
  - Simples de utilizar;
  - Flexível para vários tipos de dados;
  - Sharding - Replicação horizontal;

  Apresenta também algumas desvantagens como:

  - Joins - o processo de junção de documentos pode ser bastante complexa;
  - Elevado uso de memória;

- Apache Cassandra - É um sistema de gestão de bases de dados distribuído, altamente escalável e altamente disponível, projetado para lidar com grandes volumes de dados num ambiente distribuído.

- Apache HBase - HDFS - É um banco de dados NoSQL de código aberto, distribuído e escalável, projetado para executar operações de leitura/gravação em grande escala em dados estruturados. É construído em cima do HDFS (Hadoop Distributed File System) e fornece acesso rápido e aleatório aos dados.

A ferramenta escolhida pelo grupo foi o MongoDB que para além de apresentar todas as vantagens enumeradas acima, também se trata de um ferramenta mais familiarizada pelo

grupo. Para além disso apresenta uma boa documentação e por isso tornou-se mais fácil de ser utilizada.

3.3.2 Processo de Armazenamento

Para o armazenamento de dados, o nosso grupo definiu a utilização do MongoDB.

Para podermos armazenar os dados e tornar este processo o mais eficiente possível, criamos um *script* em *python* que automatizava este processo.

Depois de todo o tratamento de dados, e também da conclusão do processo de merge, seguimos para a parte de armazenamento. Utilizando o **PyMongo** (libraria python), criamos uma ligação à nossa base de dados MongoDB denominada "AIS", e a partir daqui enviamos todos os dados tratados para uma coleção.

A coleção tem o nome de "dados2"e contém os seguintes campos:

- _id
- country code
- gdp
- govexp
- inflation
- literacy_rate
- year

Para esta fase do projeto, beneficiámos do Docker, onde criamos uma instância do MongoDB no localhost, e também o MongoDBCompass, onde verificamos a criação das coleções.

Estes dois recursos tornaram-se indispensáveis neste processo devido às suas funcionalidades.

O MongoDB Compass é uma interface gráfica de utilizador (GUI) para o MongoDB e permite explorar e visualizar facilmente os dados armazenados no MongoDB, para além de oferecer recursos para consultas, visualização de esquemas e criação de consultas ad-hoc.

3.3.3 ETL

Seguindo a lógica do nosso projeto descrita até agora, o processo de extração de conhecimento seguiu as fases **E**xtract, **T**ransform e **L**oad (ETL).

- Extract - Consiste no processo de extração de dados referido na secção dos datasets. Isto é, quais foram as fontes de dados utilizados e como foram extraídas.
- Transform - É referente ao pré-processamento de dados indicados em que tivemos de transformar os dados, de forma a serem utilizados de forma mais intuitiva e fácil.
- Load - Refere-se ao processo de armazenamento dos dados já tratados na nossa Base de Dados MongoDB.

3.4 Visualização

A última fase deste projeto é referente a *Data Visualization*. Inicialmente iremos referir quais as ferramentas consideradas pelo grupo e qual a utilizada. Para além disso, iremos referir quais as dashboards utilizadas e criadas para retirar conlusões e conhecimento útil sobre o nosso caso de estudo e que será exposto na secção final dos resultados.

3.4.1 Ferramentas - Visualização de Dados

Para a fase de visualização de dados, as ferramentas consideradas pelo grupo foram:

- Power BI - trata-se de uma ferramenta de visualização de dados desenvolvida pela Microsoft. Oferece recursos avançados para a criação de painéis, relatórios e gráficos a partir de diversas fontes de dados. As principais vantagens são:
  - Interface intuitiva;
  - Fácil de utilizar;
  - Conexão a diferentes fontes de dados;
  - Recursos avançados de visualização;

  A principal desvantagem é a limitação de recursos da sua versão gratuita.
- Tableau - é uma ferramenta que permite amplas opções de visualização de dados e que lida bem com um grande volume da dados e de diversas fontes.
- Google Data Studio - é uma ferramenta online para a criação de relatórios e painéis a partir de banco de dados, introduzido pelo Google.

Apesar de se tratarem todas de ferramentas de visualização ótimas, utilizamos o software Power BI uma vez que já nos encontramos familiarizados com o funcionamento da mesma e é mais popular na comunidade.

3.4.2 Processo de Visualização de Dados

Após a fase anterior correspondente ao armazenamento de dados, o processo para a visualização dos nossos dados consistiu em fazer a ligação da nossa base de dados criada em Mongo BD com o software do Power BI Desktop. Apesar das dificuldades iniciais a realizar esta conexão, o grupo conseguiu ultrapassar a dificuldade e o processo de envio de dados foi bastante simples e rápido.

3.4.3 Dashboards

As dashboards construídas para a visualização dos nossos dados foram:

- Gráfico de colunas empilhadas
- Gráfico de dispersão
- Segmentação de Dados
- Gráfico de áreas empilhadas

Na criação das dashboards pretendemos analisar e relacionar as métricas escolhidas no nosso caso de estudo.

(As dashboards encontram-se representadas na secção seguinte juntamente com a análise e conclusões retiradas.)

4 Resultados e Discussão

Nesta secção iremos fazer uma ilustração dos diferentes dashboards construídos na ferramenta de visualização. Também iremos fazer uma análise critica sobre cada um e extrair conclusões sobre o nosso caso de estudo. Iremos inicialmente ilustrar diferentes gráficos a nível global e independentes do ano selecionado, a fim de extrair conclusões sobre as métricas utilizadas e como se relacionam. Numa seguinte fase, iremos, através de segmentação de dados, selecionar Portugal e um intervalo de anos (1980-2020) a fim de analisar os diferentes dados e que tipo de conclusões podemos extrair.

4.1 Dashboards - Nível Mundial

Os seguintes gráficos de dispersão irão ilustrar o comportamento das métricas utilizadas (PIB, Percentagem do PIB em Educação, Taxa de literacia e Taxa de inflação de bens essenciais) para todos os países e sem segmentação de anos.

Percentagem do PIB em Educação e Taxa de Inflação

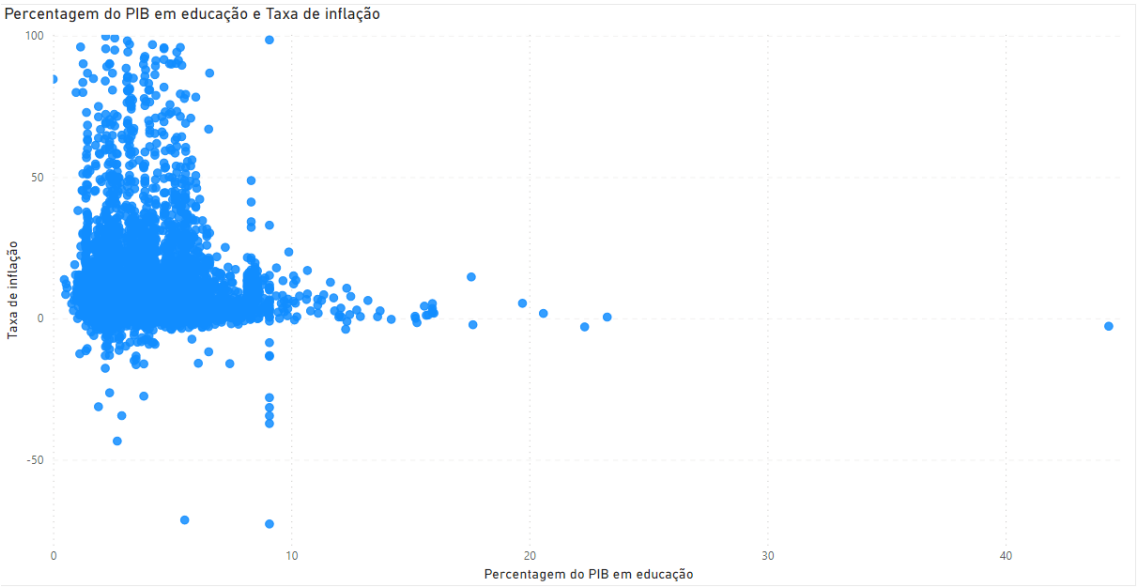

Figura 4.1: Percentagem do PIB em Educação e Taxa de Inflação

Através desta dashboard conseguimos comparar a taxa de inflação de bens essenciais com a percentagem do PIB gasto em educação. Observa-se que para uma percentagem inferior a 10 encontram-se os valores mais elevados de inflação registados nos vários países. Conseguimos concluir que a inflação é mais notória nestes bens quando o governo não investe o seu produto interno bruto na educação. Também se observa um outlier em que a percentagem é igual a 40 e a taxa de inflação é negativa(-2,76). Podemos assim afirmar que quando o governo investe e fornece mais ajudas neste setor, a taxa de inflação diminui e assim o custo de educação também.

Percentagem do PIB em Educação e PIB

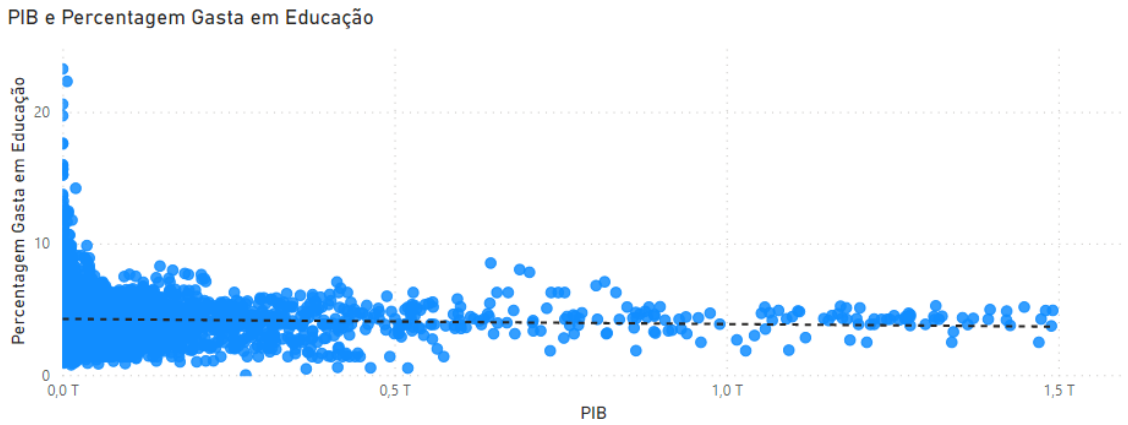

Figura 4.2: Percentagem do PIB em Educação e PIB

Esta dashboard compara para os diversos valores do PIB como varia a percentagem gasta em educação. Conseguimos analisar que a percentagem não depende diretamente do PIB do país, pois conseguimos verificar que para um baixo valor do produto interno bruto, encontram-se as percentagem mais elevadas capturadas. Por outro lado, conseguimos verificar que a média de percentagem do PIB a nível global tende a ser inferior a 5% independentemente do income dos diferentes países. (linha preta).

Taxa de Literacia e Taxa de Inflação

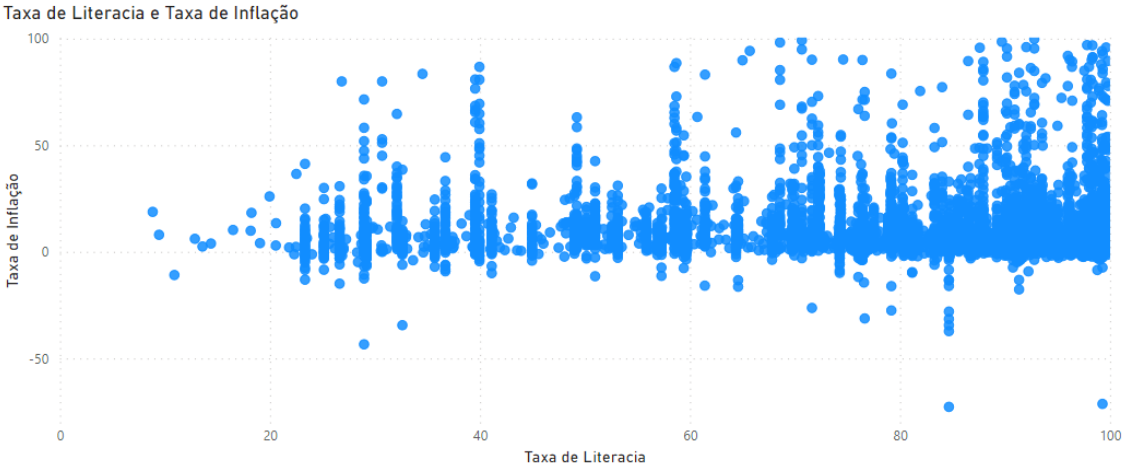

Figura 4.3: Taxa de Literacia e Taxa de Inflação

Nesta dashboard tentamos procurar alguma relação entre a taxa de literacia e a taxa de inflação. No entanto, como se pode observar estas métricas não se relacionam pois para um valor elevado da taxa de inflação encontrámos tanto valores pequenos como elevados de literacia. Para melhor concluir sobre estas métricas devemos analisar um determinado país e um intervalo de anos.

Taxa de Literacia e Percentagem do PIB em Educação

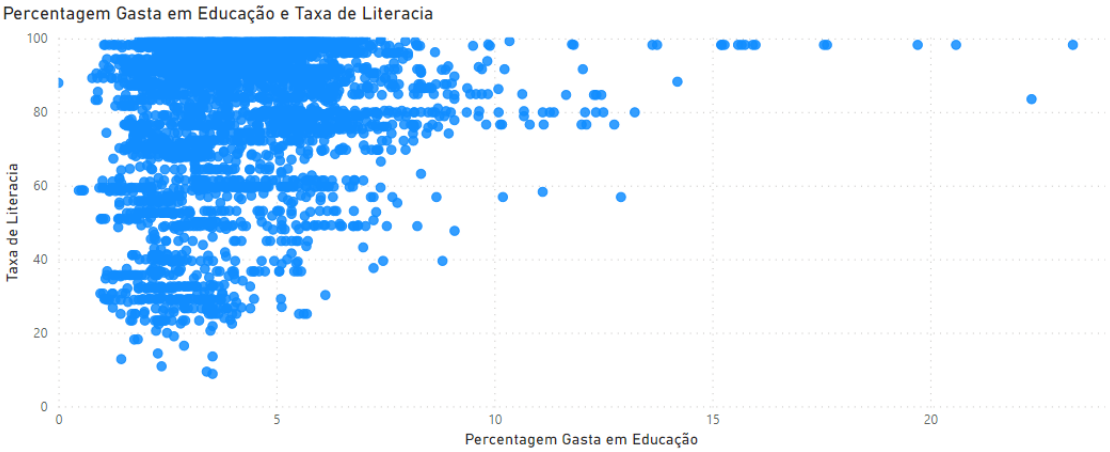

Figura 4.4: Taxa de Literacia e Percentagem do PIB em Educação

Ao analisar estas duas métricas conseguimos concluir que para uma menor percentagem de literacia (menos de 40%) a percentagem gasta em educação não saiu dos 5%-10%. Isto indica que nos anos em que se registara baixos valores de literacia, o governo também não investia neste setor. No entanto, como a falta de literacia é menos comum com o passar dos anos e com o elevado acesso a informação, para valores de 90 para cima, o valor da percentagem

do PIB já oscila mais. Ou seja, já dependerá do país em questão para analisar melhor estas métricas.

PIB e Taxa de Inflação

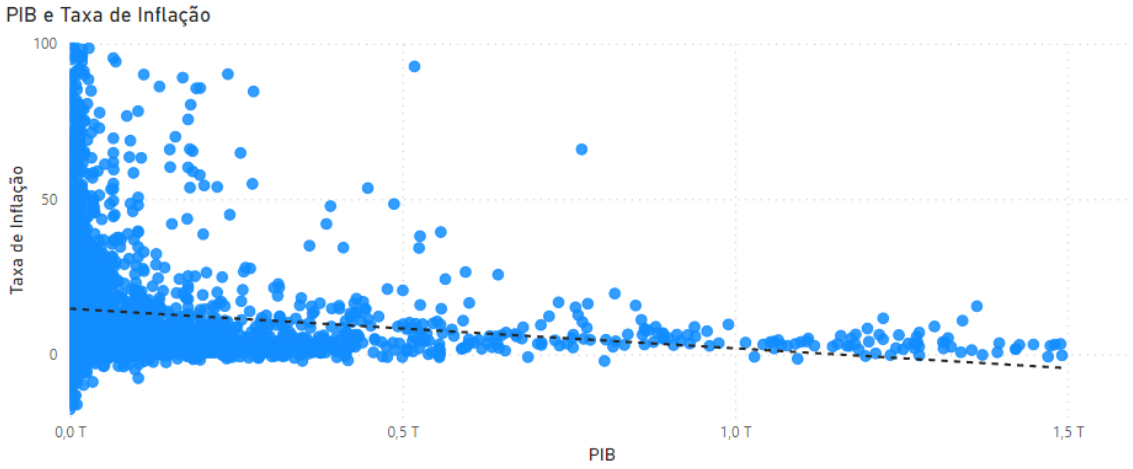

Figura 4.5: PIB e Taxa de Inflação

Através desta dashboard conseguimos comparar os diferentes valores do produto interno bruto com a taxa de inflação de bens essenciais. Conseguimos analisar que para maiores valores de inflação associam-se produtos internos mais baixos. Para além disso através da linha de tendência (linha preta) conseguimos concluir que com o aumento do produto interno bruto de um país, a taxa de inflação tende a diminuir.

PIB e Taxa de Literacia

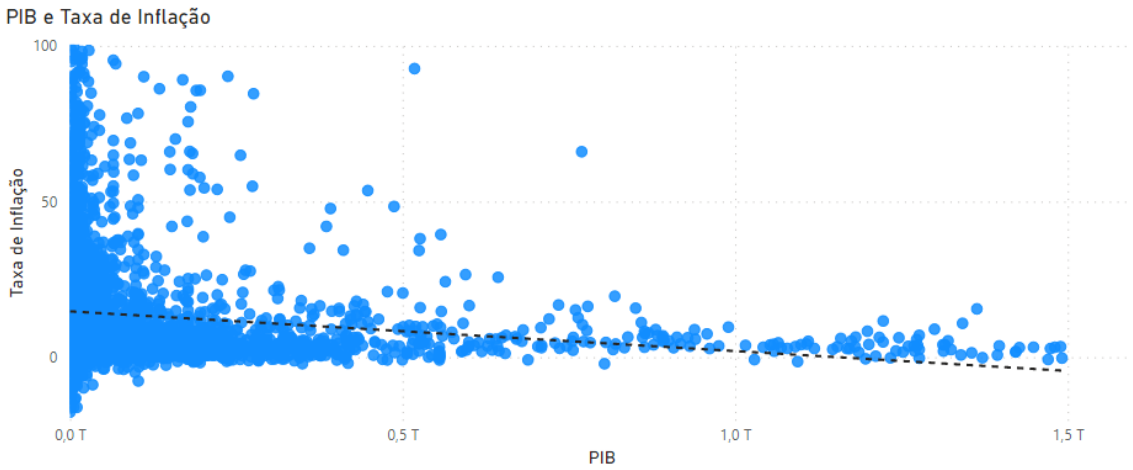

Figura 4.6: PIB e Taxa de Literacia

Neste gráfico de dispersão conseguimos relacionar o PIB de um país com a taxa de inflação. Através da linha de tendência (linha preta) conseguimos concluir que a taxa de inflação tende a diminuir com o aumento do PIB de um determinado país. Como se pode observar os maiores valores de inflação encontram-se quando o PIB é menor.

4.2 Dashboards - Caso Específico (Portugal - 1980-2021)

De forma a visualizar e analisar melhor os nossos parâmetros de estudo, isto é, Taxa de Inflação, PIB, Percentagem do PIB em Educação e Taxa de Literacia, criamos duas dashboards de Segmentação de Dados. Na primeira iremos filtrar os resultados escolhendo um país e na segunda um intervalo de anos (entre 1970-2021). Para os exemplos a seguir, escolhemos Portugal e o intervalo de anos entre 1980 e 2021. Optamos por este intervalo uma vez que os missing values dos nossos datasets estavam mais presentes entre 1960 e 1980.

Segmentação de Dados - Código do País e Ano

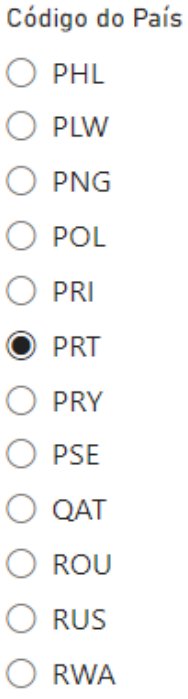

Figura 4.7: Segmentação por País

Figura 4.8: Segmentação por intervalo de anos

Através destas segmentação de dados conseguimos selecionar qual o país e intervalo de anos que queremos analisar.

Taxa de Inflação por Ano

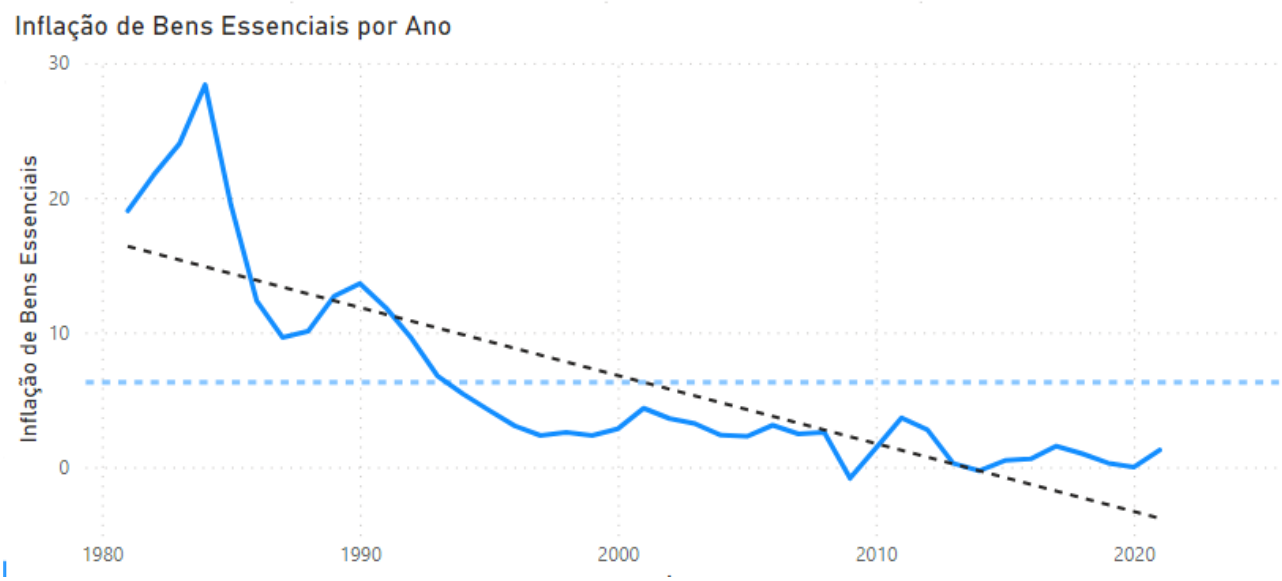

Figura 4.9: Taxa de Inflação por Ano

Neste gráfico de linhas conseguimos observar para Portugal, entre 1980 e 2021 como oscilou a taxa de inflação. A linha preta corresponde a uma linha de tendência e a linha azul corresponde à média. Conseguimos observar que a taxa de inflação sofreu um pico em 1984, no entanto tem vindo a diminuir ao longo dos anos. Por outro lado, a média dos valores estão acima de 0, ou seja, indica que os produtos essenciais encontram-se geralmente inflacionados.

Taxa de Literacia por Ano

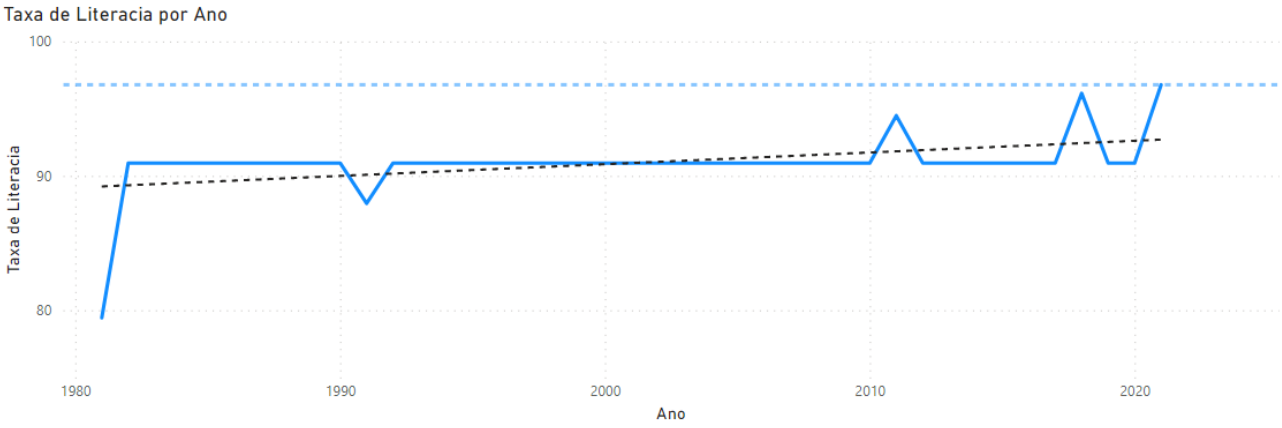

Figura 4.10: Taxa de Literacia por Ano

Através novamente deste gráfico de linhas observamos a variação da taxa de literacia dos indivíduos em Portugal. Conseguimos ver que a taxa de literacia ronda os 90 e que a tendência é aumentar ao longo do anos (linha preta). Também se observa que o maior valor registado da taxa de literacia foi no ano de 2021 (linha azul). Com isto podemos concluir que com o passar dos anos e com o maior acesso de informação, a taxa de literacia tende a aumentar.

PIB por Ano

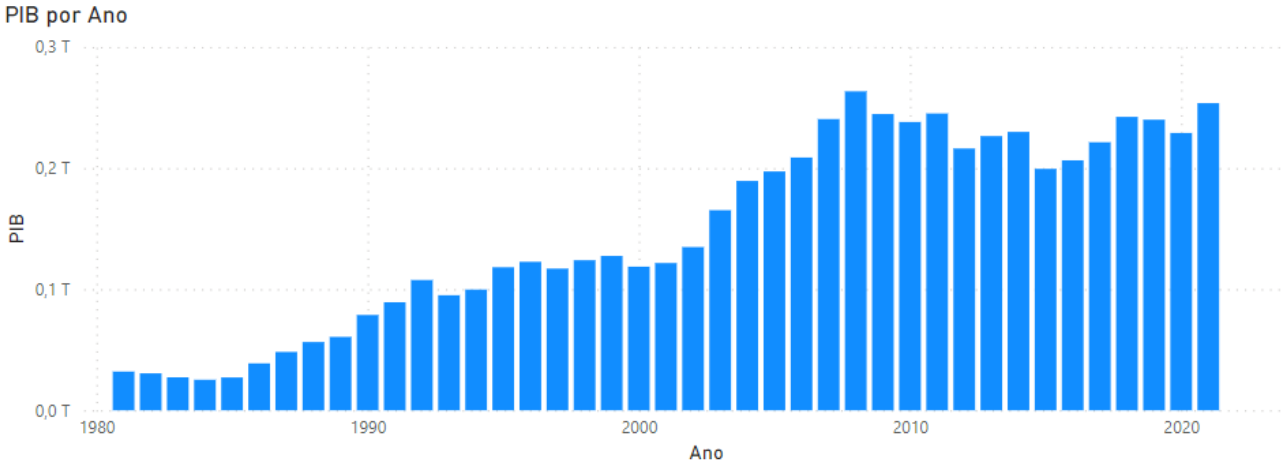

Figura 4.11: PIB por Ano

Através deste gráfico de colunas empilhadas conseguimos observar como variou o PIB em Portugal. Conclui-se que tem vindo a aumentar, apesar de sofrer algumas oscilações a partir de 2010

Percentagem Gasta em Educação por Ano

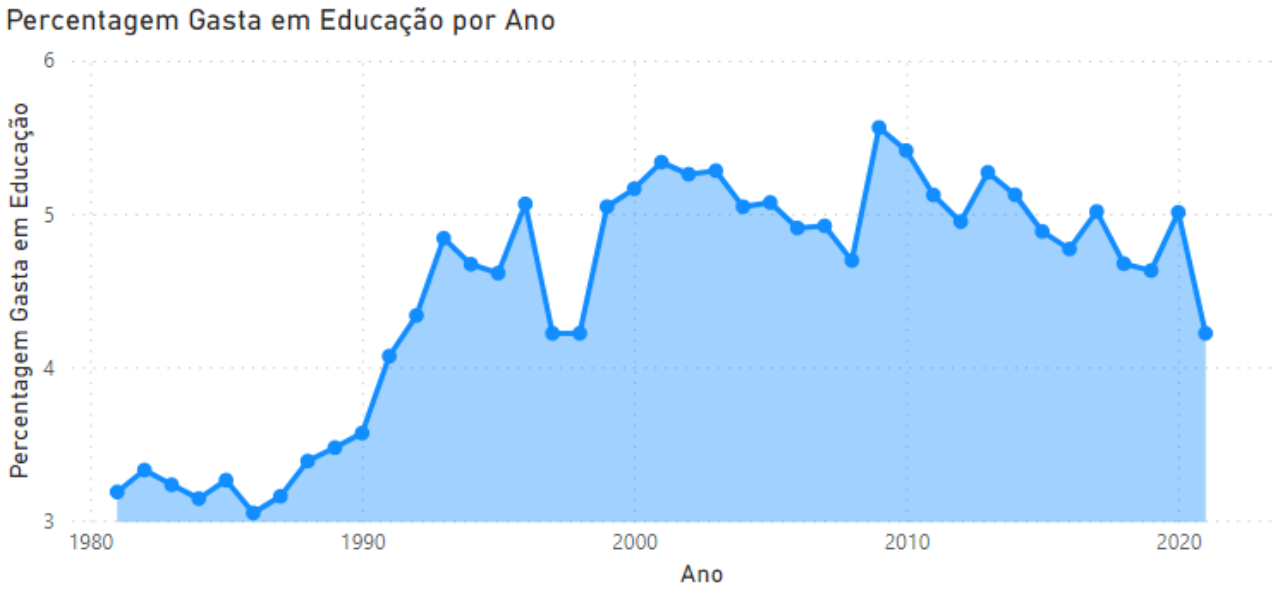

Figura 4.12: Percentagem Gasta em Educação por Ano

Neste gráfico de áreas empilhadas, analisamos como foi a variação da percentagem gasta em educação. Podemos verificar que de 1980 a 1990 o valor era muito baixo mas que sofreu um aumento significativo até 2000. Desse ano até 2020 podemos verificar que o valor da percentagem vai oscilando entre 4,5 e 5,5.

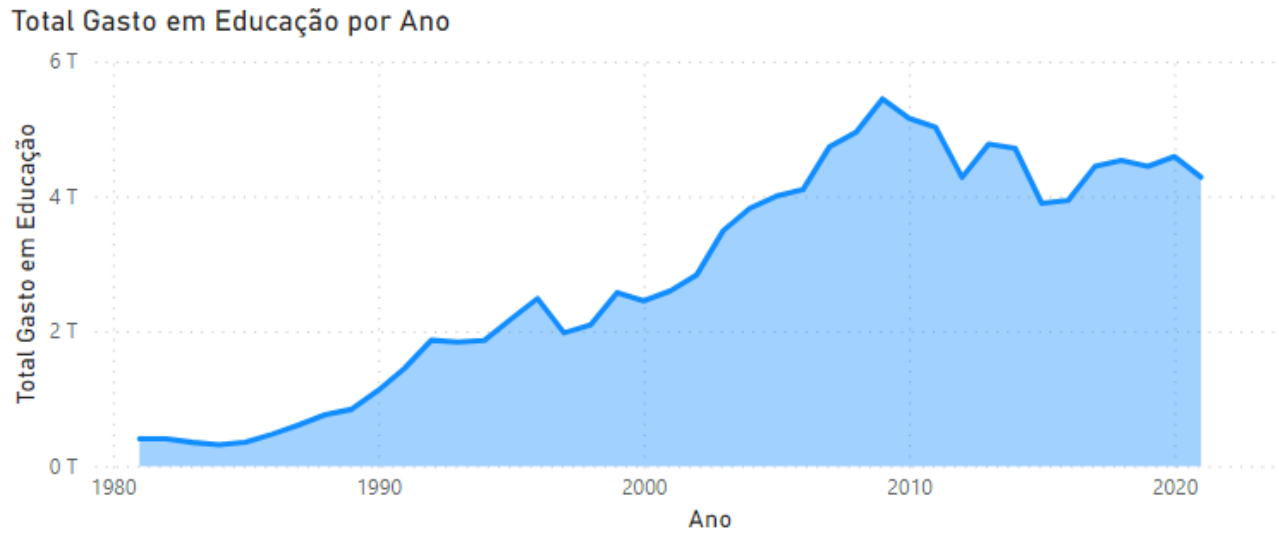

Figura 4.13: Total Gasto em Educação por Ano

De forma a complementar esta análise, criámos uma nova medida em que calculamos o valor exato gasto no setor educativo por ano. Com este novo gráfico conseguimos confirmar o que

foi dito anteriormente e reforçar que ouve um aumento de investimento até 2008 e a partir daí até 2021 tem oscilado pouco.

Nesta secção apenas ilustramos a análise e conclusões retiradas para o caso especifico de Portugal e de 1980 a 2021. Este painel, no entanto, pode ser utilizado para qualquer outro país e com um intervalo de anos diferente. Logo, as conclusões e análises para outros casos podem ser muito distintas das que foram feitas aqui.

5 Conclusões e Trabalho Futuro

Em conclusão, o trabalho prático da unidade curricular de Análise Inteligente em Big Data permitiu-nos não só aprofundar os conhecimentos adquiridos durante as aulas, como também adquirir novos na área e como aplicá-los.

Observámos que o processo de análise envolve vários processos, sendo os principais, o Tratamento de Dados, Armazenamento e Visualização. Para além disto, concluímos também que incentivou a nossa procura por ferramentas novas de acordo com as nossas necessidades e o papel fundamental que estas têm em Big Data.

O maior desafio que surgiu durante o desenvolvimento do projeto, foi o Tratamento de Dados devido a ser um processo trabalhoso, e também devido à quantidade de Datasets e a sua grande quantidade de informação. Foi necessário bastante tempo para a realização desta parte sendo consequentemente a parte mais trabalhosa do trabalho prático.

Para além disto também existiram desafios na conexão da nossa ferramenta de visualização com a ligação à Base de Dados, que conseguimos ultrapassar.

Relativamente ao nosso tema conseguimos estabelecer relações entre a inflação e os seus efeitos na educação a nível global, através da visualização de diversos gráficos, de onde, após análise e compreensão, conseguimos retirar conclusões que permitiram os resultados e conclusão deste trabalho.

Através da análise, planeamento e implementação dos processos envolvidos pudemos explorar melhor quais as vantagens e desvantagens de cada abordagem, bem como a sua aplicabilidade, e, assim, melhorar a nossa capacidade de decisão e *problem solving*.

Como trabalho futuro, gostaríamos de adicionar novas e diferentes fontes de dados que permitissem novas conclusões e conhecimento útil sobre o nosso estudo. Para além disso, a utilização de fontes que sejam dinâmicas e que apresentem valores ainda mais recentes do que os apresentados seria uma vertente a melhorar neste projeto. Outro aspeto seria a utilização de diferentes ferramentas de forma a expandir o nosso conhecimento na área de Big Data, quer para o armazenamento quer para a visualização de dados.

Por fim, ao realizar este trabalho prático, adquirimos uma compreensão mais aprofundada dos principais desafios a enfrentar ao implementar os diversos processos relacionados com Big Data. Esta experiência permitiu-nos melhorar significativamente o nosso conhecimento geral nesta área.

Referências

(1) Dataset CPI Inflation:

<https://www.worldbank.org/en/research/brief/inflation-database >

(2) Dataset PIB Global:

<https://www.kaggle.com/datasets/fredericksalazar/
pib-gdp-global-by-countries-since-1960-to-2021

(3) Dataset Government expenditure on education:

<https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS?view=chart

(4) Dataset Literacy Rate:

<https://data.worldbank.org/indicator/SE.ADT.LITR.ZS?view=chart

(5) Parquet:

<https://www.upsolver.com/blog/apache-parquet-why-use>

(6) Mongo DB:

<https://www.mongodb.com/atlas/database>

Lista de Siglas e Acrónimos

BD Base de Dados

PIB Produto Interno Bruto

GDP Gross Domestic Product

CPI Consumer Price Index