Relatório de Visualização de Dados - Projecto 2 (Power BI)

Faculdade de Ciências da Universidade de Lisboa
Duarte Balata (46304), Inês Almeida (50921) e Miguel Nunes (40790)

O presente trabalho focou-se na utilização da ferramenta de software Power BI para a exploração de relações entre diferentes variáveis de um determinado conjunto de dados. O conjunto de dados a utilizar neste trabalho foi dado a escolher aos alunos, de entre duas temáticas distintas:

1) Dados sobre Ensino Público em Portugal;
2) Dados sobre sinais fisiológicos.

Para este estudo foi escolhido o conjunto de dados sobre o **Ensino Público em Portugal**, facultado pela Direção-Geral de Estatísticas da Educação e Ciência (DGEEC).

De forma a encontrar relações de interesse e com significado para o entendimento e melhoria do ensino público em Portugal, as variáveis deste conjunto de dados foram cruzadas com variáveis adicionais, pertencentes ao conjunto de dados disponíveis nos repositórios PORDATA e Instituto Nacional de Estatística (INE). De entre todo o conjunto de variáveis incluídas nestes repositórios, era pedido a priori que se utilizassem as variáveis “Poder de compra per capita” e “Taxa de natalidade por 1000 habitantes”, deixando à escolha dos alunos a seleção de pelo menos duas outras variáveis disponíveis neste conjunto.

Desta forma, após avaliar a existência ou ausência de correlações entre diversas variáveis destes repositórios e as variáveis disponíveis relativas ao ensino público, foram adicionalmente escolhidas para este trabalho as variáveis “Crimes por categoria por mil habitantes” e “Número médio de alunos por computador com ligação à Internet no ensino básico e secundário público: total e por nível de ensino”, ambas obtidas no portal **PORDATA**. A segunda foi posteriormente dividida em “Número médio de alunos por computador com ligação à Internet no 3º ciclo” e em “Número médio de alunos por computador com ligação à Internet no ensino secundário”.

Por outro lado, as variáveis relativas ao ensino público que se destacaram e revelaram correlações interessantes de incluir e interpretar neste estudo foram: “Escolas” – variável representativa de todas as escolas estudadas; “Estudantes (Total)” – variável representativa do número total de estudantes por escola; “TEIP” – variável binária indicativa se a escola faz ou não parte de Territórios Educativos de Intervenção Prioritária (TEIP); “Abandono escolar ou risco de abandono escolar (secundário) (%)” –

variável representativa da taxa de abandono e desistência escolar; “Nível de ocupação escolar (3º ciclo – secundário)” – variável que quantifica o nível de ocupação das escolas por alunos; “Média de anos de escolaridade das mães dos alunos no nível de ensino secundário” – variável que representa o número médio de anos que as mães dos alunos do ensino secundário de uma dada escola estudaram ao longo da vida.

Uma vez que os dados do Ensino Público que caracterizam os alunos remontam ao ano letivo de 2016/2017, os dados das variáveis retiradas do portal PORDATA e aqui utilizados correspondem também ao ano de 2017, para que as variáveis possam ser correctamente comparadas e a informação extrapolada seja fidedigna e representativa de uma tendência ou correlação com verdadeiro significado.

Segue-se a lista organizada do conjunto de variáveis utilizadas neste trabalho:

Variáveis retiradas do portal PORDATA

1. Poder de compra per capita
2. Taxa de natalidade por 1000 habitantes
3. Crimes por categoria por mil habitantes
4. Número médio de alunos por computador com ligação à Internet no ensino básico e secundário público: total e por nível de ensino

Variáveis do ensino público em Portugal

1. Escolas
2. Estudantes (Total)
3. TEIP
4. Abandono escolar ou risco de abandono escolar (secundário) (%)
5. Nível de ocupação escolar (3º ciclo – secundário)
6. Média de anos de escolaridade das mães dos alunos no nível de ensino secundário

Depois de selecionadas e recolhidas as variáveis para este estudo, as tabelas que continham as variáveis foram tratadas e editadas antes e após serem importadas para o software Power BI. Neste processo de edição os dados foram organizados, o nome das colunas e das tabelas foi modificado ou abreviado, o domínio dos dados de cada variável foi verificado e/ou rectificado, e as instâncias com entrada igual a “n” (correspondente à designação “não aplicável”) foram substituídas por “null” (valores nulos). Adicionalmente, os valores das variáveis “Crimes por categoria por mil

habitantes” e “Média de anos de escolaridade das mães dos alunos no nível de ensino secundário” foram transformados e agrupados em categorias discretas (bins) para uma melhor representação e interpretação dos dados.

Técnicas de visualização de dados oferecidas pelo Power BI

O software Power BI oferece um conjunto de técnicas de visualização bastante diversas e que podem ser utilizadas em contextos diferentes, conjugando diversos tipos variáveis. O objectivo da utilização de diferentes técnicas consiste em explorar a melhor representação e visualização possível, que facilite uma fácil interpretação por parte do utilizador e extrapolação da informação contida num determinado conjunto de dados.

No seu conjunto, as técnicas de visualização standard disponíveis no software Power BI incluem:

1. **Gráficos de barras e colunas agrupadas ou empilhadas**

   É um método de visualização muito comum e prático, que facilita a representação de uma ou mais variáveis com valores distintos para diferentes categorias, representadas por barras ou colunas de cores diferentes.

2. **Gráfico de linhas**

   É um dos métodos de visualização mais simples e comuns, que ilustra o valor que uma ou mais variáveis apresentam ao longo de um eixo. Este eixo pode representar datas (evolução das variáveis ao longo do tempo), categorias ou valores de uma segunda variável.

3. **Gráfico de áreas**

   É uma representação semelhante ao gráfico de linhas mas com a área abaixo da linha preenchida, demonstrando a evolução de variáveis em função de outras variáveis. Podem facilitar uma rápida interpretação dos resultados e de uma tendência ao longo do eixo entre variáveis que apresentem áreas diferentes. É uma representação útil para variáveis que representem quantidades.

4. Gráfico de áreas empilhadas

É uma técnica de visualização semelhante ao gráfico de áreas acima descrito, mas essencialmente direcionado para a comparação de áreas de categorias distintas. Esta comparação é facilitada pela apresentação das áreas das diferentes variáveis no topo umas das outras, cada uma começando no ponto onde a anterior terminou (com exceção da primeira variável, que começa no eixo horizontal do gráfico). Esta técnica permite ao utilizador distinguir facilmente qual das variáveis possui uma maior área em cada ponto do intervalo de valores analisado.

5. Gráfico de combinação (colunas + linhas)

Combina dois tipos de gráficos diferentes (de colunas e de linhas) permitindo correlacionar múltiplas variáveis de uma forma intuitiva. Ajuda também a economizar espaço numa dashboard por permitir conjugar diferentes visualizações numa única, sem perder clareza na interpretação e extrapolação dos resultados.

6. Gráfico de friso

É um tipo de visualização que permite facilmente perceber que variável apresenta os valores mais elevados num determinado momento, ou para uma determinada categoria ou valor de outra variável. A evolução de cada variável ao longo do eixo dos xx está representada por colunas, que estão ligadas entre si por um sombreado de cor única e próxima da cor das próprias colunas, para uma fácil distinção das diferentes variáveis e respectivos valores em cada ponto do eixo horizontal.

7. Gráfico de cascata

É uma representação útil sob a forma de colunas para entender o avanço e recuo (aumento e decréscimo) dos valores de uma certa variável em função do

tempo, ou de categorias ou valores de outra variável. As colunas que representam avanços têm uma cor diferente das colunas que representam recuos, pelo que a interpretação da sua evolução é fácil e intuitiva.

8. Gráfico de funil

Trata-se de uma representação de barras horizontais sequenciais, em que cada uma das barras corresponde a uma determinada categoria e o seu tamanho varia em função do valor que uma dada variável apresenta para essa mesma categoria, representada sob a forma de percentagem. Neste tipo de visualização a representação das categorias é feita de cima para baixo, das percentagens mais altas para as mais baixas, o que leva a que o gráfico apresente a forma afunilada que dá nome a esta representação visual.

9. Gráfico de dispersão

É um dos tipos de visualizações mais comuns para representar correlações entre duas ou mais variáveis. Cada variável é representada num dos eixos do gráfico, com o seu intervalo de valores, e a dispersão de pontos representados faz a correlação entre os valores das diferentes variáveis. Através deste tipo de visualização é possível extrapolar de forma intuitiva tendências de correlações lineares ou não entre as variáveis, ou encontrar maiores concentrações de pontos num determinado intervalo para cada uma, ou ainda encontrar valores outliers.

10. Gráficos circulares (“queijo”) e em anel

São visualizações que representam diferentes partes de um todo, ou a divisão de um valor total por diferentes categorias, apresentando uma configuração circular totalmente preenchida (semelhante a um queijo, quando estão representadas as diferentes categorias) ou aberta no meio (semelhante a um anel). As diferentes categorias destes gráficos são usualmente representados em forma de percentagens.

11. Gráficos Treemap

Trata-se de uma representação hierárquica de diferentes variáveis, divididas em retângulos de tamanhos e cores diferentes, sendo o tamanho do retângulo proporcional ao valor da variável em causa. As diferentes variáveis são ordenadas pelo tamanho do seu retângulo, ficando o maior retângulo no canto superior esquerdo e o menor retângulo no canto inferior direito.

12. Representação de mapas

O software Power BI possibilita também a visualização dos dados de diferentes variáveis sobre mapas, podendo estes ser de diversos tipos. Entre eles, destacam-se os mapas simples, correspondentes a uma determinada região geográfica definida no estudo em causa, e onde se podem inserir os dados com uma referência espacial precisa. Esta ligação dos dados ao mapa facilita a interpretação dos mesmos quando existe uma correlação entre a diferença de valores e a localização geográfica. Os dados podem também ser representados em mapas sob a forma de cores que o preenchem, e cuja luminosidade determina o *range* de valores distintos entre regiões distintas (mapas preenchidos). Por outro lado, os dados não têm de ter uma referência espacial, e as regiões de interesse podem ser simplesmente representadas através de um mapa de formas e cores, isolado de qualquer conexão ao mapa real (mapas de forma).

13. Cartões de valores (únicos ou múltiplos)

Correspondem a representações visuais simples de um único valor de uma variável (cartão), ou de um ou mais valores de diferentes variáveis por fila (cartão de linhas múltiplas). É útil para verificar de forma prática e rápida qual é o valor específico de uma dada variável quando se interage de forma dinâmica com o conjunto de visualizações de uma dashboard.

14. Gráficos KPI (indicador chave de performance) e Medidor

São dois gráficos diferentes, que representam um progresso percentual ou numérico alcançado para uma dada variável em função de um certo objetivo. O gráfico KPI mostra a evolução de uma variável ao longo do eixo horizontal e o ponto exacto onde o valor do progresso atual dessa variável se encontra. O gráfico medidor faz o mesmo, mas com uma representação semelhante ao velocímetro de um carro em que a agulha mostra o objectivo, enquanto que o progresso alcançado está representado com uma determinada cor até esse ponto.

15. Gráfico de principais influenciadores (ou influenciadores chave)

É um tipo de representação que ajuda a perceber quais os fatores (variáveis) que mais influenciam e contribuem para um certo valor de outra variável.

16. Representação de tabelas e matrizes

São visualizações simples dos dados, correspondentes a uma ou mais variáveis do nosso dataset, sob a forma de tabelas organizadas com diferentes linhas e colunas. Permitem ao utilizador selecionar determinadas linhas de uma variável ou uma variável em detrimento das restantes, e observar interactivamente e de forma destacada os dados selecionados nas restantes visualizações que compõem a dashboard.

17. Representação de segmentação de dados

É uma técnica de visualização que, de forma semelhante às tabelas e matrizes, permite selecionar quais as categorias ou valores de uma variável que queremos visualizar temporária e simultaneamente nas restantes técnicas de visualização que utilizam essa mesma variável, geradas e integradas de forma dinâmica na mesma dashboard.

18. Representação de árvore de decomposição

Esta é uma técnica de visualização relativamente recente do Power BI e de particular interesse nas áreas de Aprendizagem Automática e Inteligência Artificial. Permite representar os valores de uma variável sob a forma de percentagem em cada dimensão da árvore, de uma forma hierárquica e explicada segundo as categorias de outras variáveis. A variável escolhida em cada nível/dimensão deve estar relacionada com as variáveis dos níveis anteriores, de forma integrativa e sequencial para que a interpretação dos resultados seja intuitiva e a extrapolação dos mesmos tenha significado.

19. Representação de perguntas e respostas

Este tipo de visualização permite, de forma muito direta, usar a nossa linguagem natural (como o Inglês) para fazer questões específicas sobre o conjunto de dados que foi inicialmente importado para o software Power BI. As perguntas podem incidir sobre uma ou mais variáveis e esta técnica de visualização apresenta respostas diretas às perguntas feitas, sob a forma de valores das variáveis em causa.

Para além das visualizações standard do Power BI, é ainda possível importar outras visualizações *custom* de um alargado leque de opções disponíveis no repositório online do software, desenvolvidas por empresas ou particulares. Independentemente do alargado conjunto de visualizações disponíveis no software Power BI, a utilização de cada visualização depende sempre do objectivo que se pretende alcançar e das variáveis em causa.

Visualizações criadas, interpretação e discussão dos resultados

Numa fase inicial do trabalho foram importados para o PowerBI os conjuntos de dados relativos ao ensino em escolas públicas portuguesas e os dados recolhidos do portal PORDATA. Foi então criada uma ligação entre os conjuntos através da variável geográfica “Municípios” e seguido o processo de edição e tratamento dos dados anteriormente descrito. Após a realização do tratamento básico dos dados, procedeu-

se à exploração dos mesmos com recurso às técnicas de visualização oferecidas pelo software.

Primeiramente, numa das janelas do dashboard foi criado um cartão de valor único com a variável “Estudantes (Total)” presente no dataset relativo ao ensino em escolas públicas portuguesas. Através desta visualização, é possível tomar conhecimento da dimensão do universo amostral de alunos em estudo. Para além disto, devido às funcionalidades interativas dos dashboards deste software (discutidas posteriormente) e como explicado acima na descrição desta visualização, é também possível determinar a quantidade de alunos pertencentes a cada um dos valores ou categorias analisadas nas restantes visualizações desenvolvidas e incluídas no dashboard.

De seguida, foram realizados gráficos, na tentativa de evidenciar relações interessantes entre as variáveis escolares escolhidas para este estudo e os dados referentes às autarquias retirados do portal PORDATA. De modo a manter a coerência temática e assim tirar melhor partido da funcionalidade interativa dos dashboards providenciada pelo software PowerBI, todas as variáveis escolhidas e analisadas neste trabalho centram-se na temática das condições socioeconómicas das escolas e dos municípios e na forma como estas impactam diversos setores que podem ser determinantes para o sucesso académico dos estudantes.

Em primeiro lugar, analisou-se a quantidade de escolas em que é aplicado o programa **TEIP** (Programa Territórios Educativos de Intervenção Prioritária). Este programa consiste numa iniciativa governamental, implementada ao longo de diversas escolas por todo o país, e que tem como finalidade a redução do abandono escolar e do absentismo em regiões económica e socialmente desfavorecidas. Para tal, recorreu-

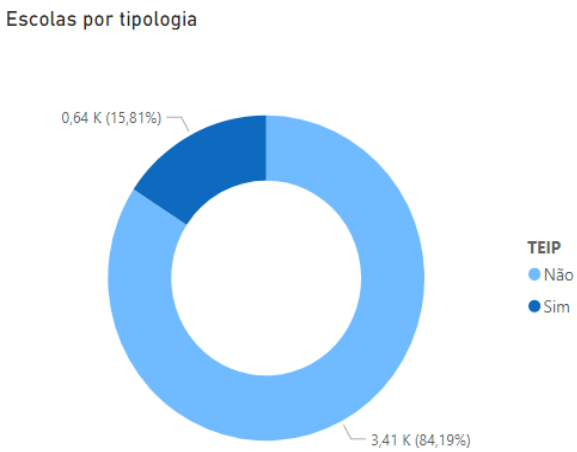

Figura 1 – Proporção de escolas públicas abrangidas pelo programa TEIP.

se a um gráfico de anel, no qual foi introduzido como valor da contagem a variável “Escolas” do dataset das escolas públicas e como legenda a variável “TEIP”, cujos valores na tabela foram alterados para ‘Sim’ e ‘Não’ de modo a aumentar a interpretabilidade das visualizações. Assim, através da análise desta representação torna-se possível verificar que 640 escolas (15.81% das escolas) são abrangidas pelo programa TEIP (Figura 1).

Como previamente descrito, sendo um dos objetivos centrais deste programa a prevenção e redução do abandono escolar, decidiu-se então comparar a relação entre os níveis de abandono escolar no ensino secundário dos alunos pertencentes a escolas TEIP e não-TEIP através da utilização de um gráfico de barras horizontal. Para a construção deste gráfico foram utilizadas as variáveis “Abandono escolar ou risco de abandono escolar (secundário) (%)“ e “TEIP”, tendo-se recorrido à média da primeira variável no campo dos valores (eixo dos xx) e a segunda variável no campo da categoria (eixo dos yy). Ao analisar esta representação torna-se então bastante fácil de compreender que apesar do esforço realizado pelo governo na criação deste programa,

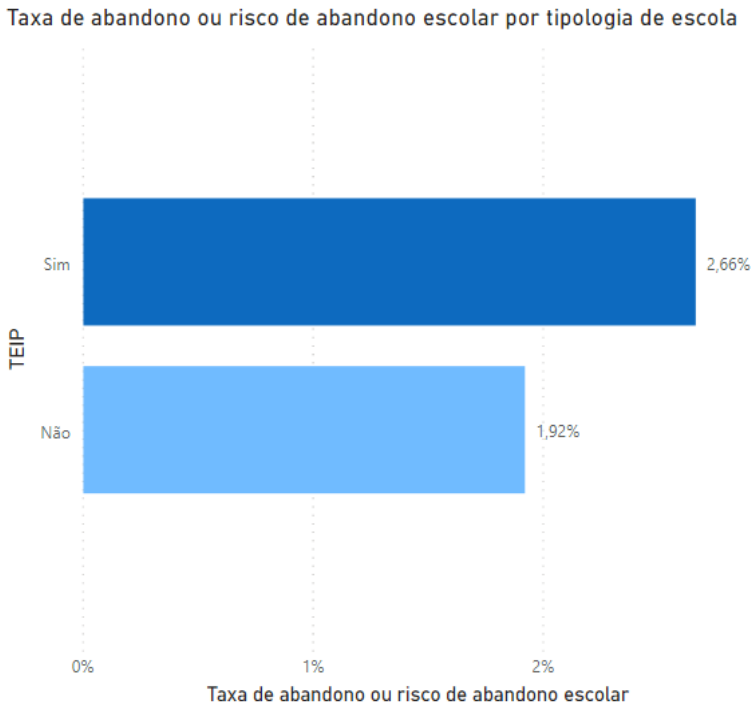

Figura 2 - Abandono escolar ou risco de abandono escolar (secundário) (%) por tipologia de escola

continua a existir uma maior percentagem de abandono ou risco de abandono escolar entre os alunos de escolas abrangidas por este programa (2,66%) do que em escolas nas quais o programa não foi implementado (1,92%) (Figura 2).

Com o intuito de tentar perceber a forma como a criminalidade dos municípios tem impacto sobre a necessidade de implementação deste programa nas suas escolas, decidiu-se utilizar um gráfico de colunas 100% empilhadas, em que se colocou como variável de eixo a “criminalidade por 1000 habitantes” proveniente do dataset recolhido do portal PORDATA. Esta variável foi agrupada e discretizada em quatro níveis, correspondendo o “Nível 1” aos valores de menor criminalidade e o “Nível 4” aos de maior criminalidade. Para que fosse possível a realização desta visualização foram ainda utilizadas as variáveis “Estudantes” e “TEIP” do dataset das escolas públicas, tendo sido a primeira utilizada para os valores do gráfico e a segunda para a construção da legenda. A partir desta visualização é possível confirmar que existe uma maior proporção de estudantes em escolas TEIP em municípios com maior criminalidade do que em municípios com menor criminalidade (Figura 3).

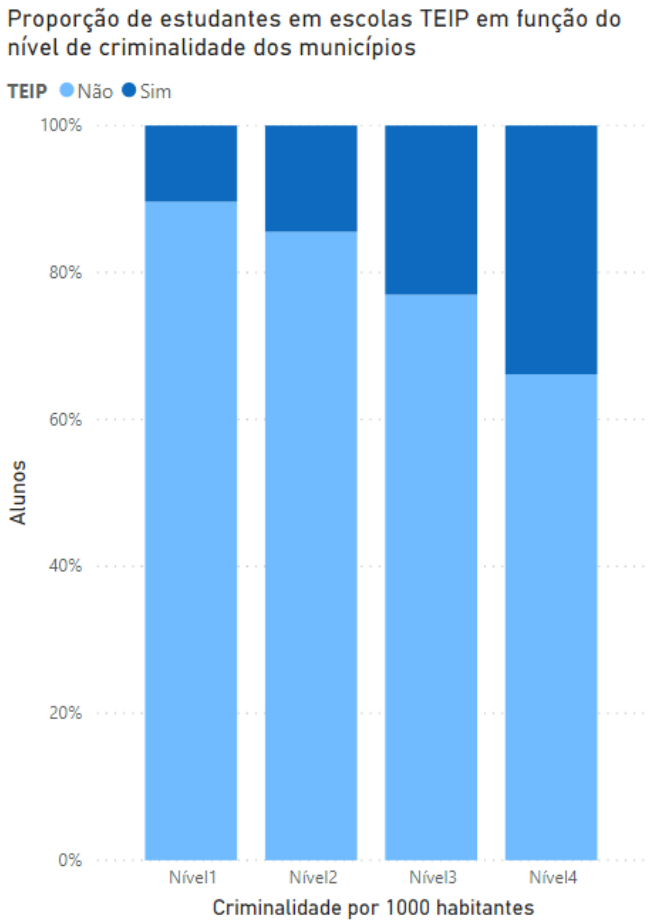

Figura 3 – Aumento da proporção de estudantes em escolas TEIP em função do nível de criminalidade dos municípios. A taxa de criminalidade por município varia entre 8,6 e 80,5 crimes por 1000 habitantes e encontra-se discretizada em 4 bins de igual largura, sendo que o nível1 corresponde aos índices de criminalidade mais baixa e o nível4 à criminalidade mais elevada.

De seguida, foi construído um gráfico de áreas empilhadas com valores da média de duas outras variáveis disponíveis no portal PORDATA, o “Número médio de alunos por computador com ligação à Internet no 3º ciclo” e o "Número médio de alunos por computador com ligação à Internet no ensino secundário”, tendo sido utilizada para o eixo dos xx a variável “Nível de ocupação escolar (3º ciclo – secundário)”, uma variável categórica com valores de 1 a 5, presente no conjunto de dados das escolas públicas portuguesas. Com esta visualização pretendeu-se estudar o acesso dos estudantes a computadores em função do nível de ocupação das escolas, tendo sido possível concluir que quanto maior o nível de ocupação das escolas maior é o número de alunos por cada computador, uma vez que se regista uma tendência positiva, tanto ao nível do ensino de 3º ciclo como secundário. Esta representação dá-nos uma indicação para a existência de um elevado número de estudantes sem acesso a computadores pessoais, que dependem dos computadores das escolas para a realização das suas atividades letivas. Assim sendo, com um aumento no nível de ocupação das escolas verifica-se uma maior necessidade de partilha de equipamentos informáticos entre estudantes, o que pode levar a uma potencial diminuição do rendimento escolar dos mesmos (Figura 4).

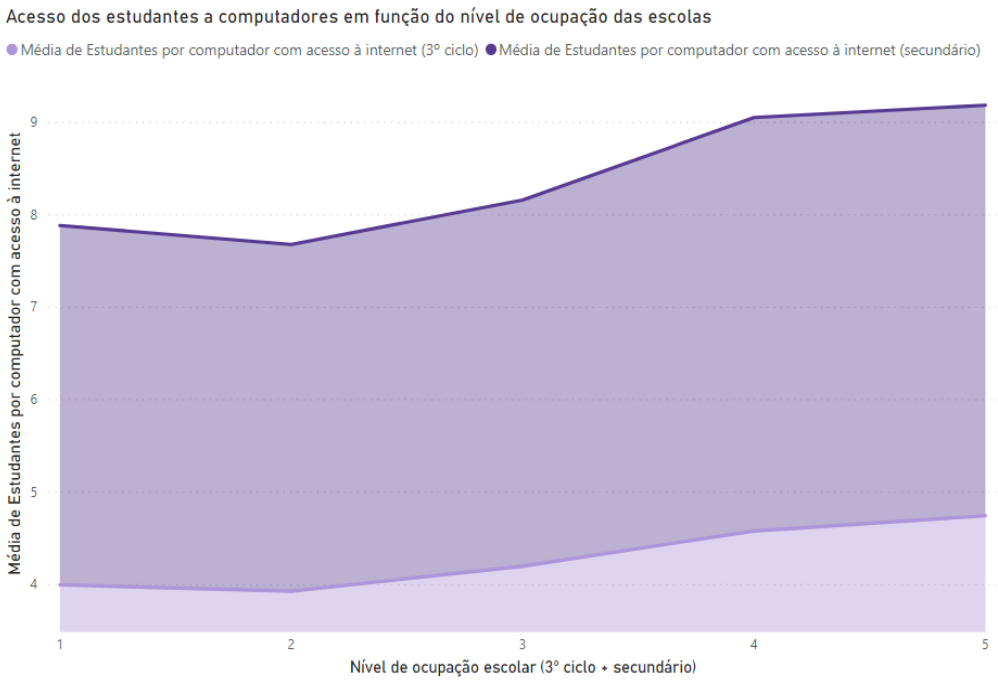

Figura 4 – Aumento do número de estudantes por computador com acesso à internet, ao nível do 3º ciclo e ensino secundário em função do nível de ocupação das escolas que frequentam.

Adicionalmente, considerou-se pertinente estudar a taxa de abandono do ensino secundário em função do nível de educação escolar das mães dos alunos e do poder

de compra per capita de cada município. Elaborou-se então um gráfico de combinação, mais precisamente um gráfico de linhas e colunas. Para os valores da linha deste gráfico foi utilizada a média da variável “poder de compra per capita” presente no dataset recolhido do PORDATA. Para os valores das colunas foi utilizada a média da variável “Taxa de abandono ou risco de abandono escolar” e para o eixo partilhado a variável “Nível de educação escolar das mães dos alunos do ensino secundário” discretizada sob a forma de categorias.

Graças a esta visualização é possível concluir que a probabilidade de abandono ou risco de abandono escolar diminui quanto maior for o número de anos de escolaridade das mães dos alunos. É ainda possível verificar que quanto mais elevado o nível de escolaridade das mães, maior se torna o poder de compra per capita da região. É então possível fazer várias interpretações dos dados expostos neste gráfico. Por um lado, é possível propor a explicação de que o maior nível de escolaridade da geração das mães dos alunos seja a causa do enriquecimento da região, que por sua vez confere às famílias a capacidade de proporcionar aos alunos melhores condições de vida e estudo, o que leva à consequente diminuição da taxa de abandono escolar. Por outro lado, é possível, ou até provável, que o poder de compra per capita nos municípios portugueses se tenha mantido relativamente estável ao longo do período de uma ou mais gerações. Deste modo, as mães de alunos que nasceram à partida em regiões com maiores rendimentos, tiveram a possibilidade de estudar durante mais tempo e consequentemente proporcionar melhores condições aos seus filhos.

De qualquer forma, independentemente da ordem de dependência das variáveis representadas no gráfico, torna-se sempre possível concluir pela observação do mesmo que os alunos filhos de mães mais educadas, têm geralmente acesso a melhores condições de vida (em termos económicos), o que leva a uma redução acentuada na taxa de abandono escolar (Figura 5).

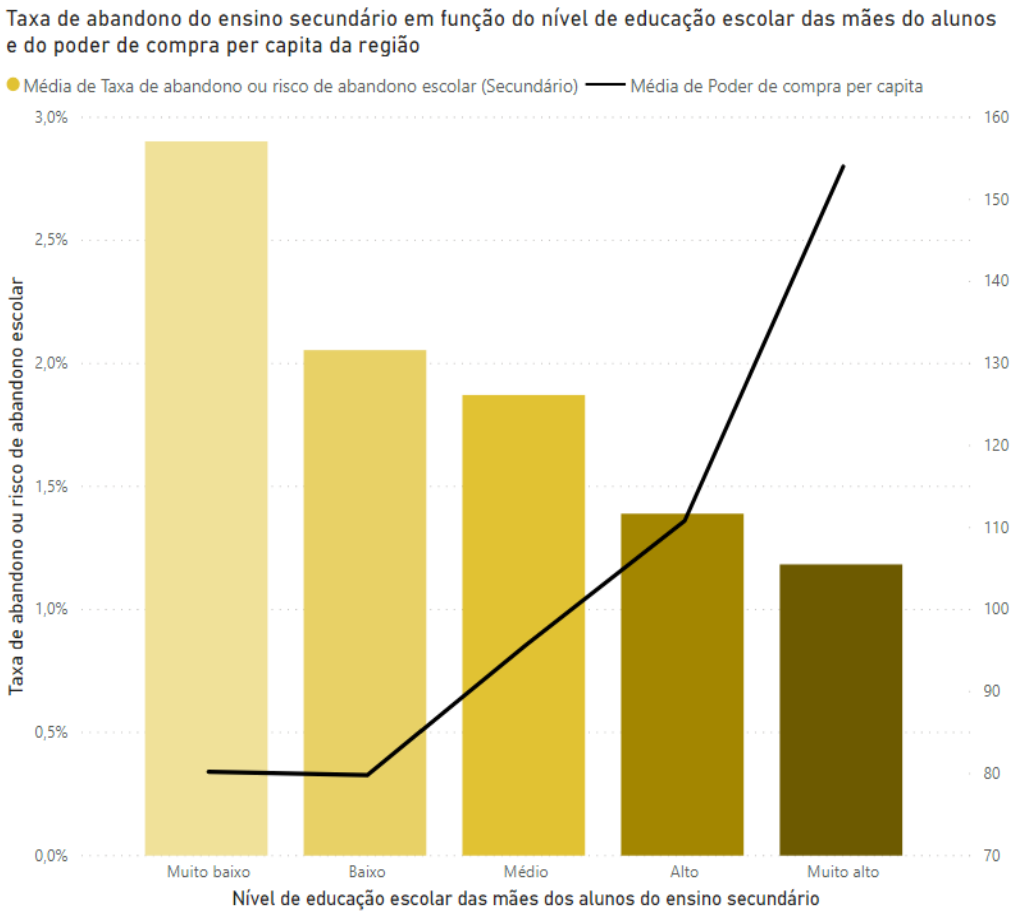

Figura 5 – Redução da abandono do ensino secundário em função do aumento nível de educação escolar das mães dos alunos e do poder de compra per capita da região. A variável 'nível de educação' foi gerada a partir da variável 'anos de educação das mães de alunos do ensino secundário' (com uma distribuição entre 6,2 e 14,7 anos) através da sua discretização em 5 bins de igual largura.

Por fim, foram realizadas mais duas representações em que se incluiu uma nova variável, "taxa de natalidade por 1000 habitantes" do dataset importado do portal PORDATA. Inicialmente, tentou-se encontrar uma correlação entre o nível de educação das mães dos alunos e a taxa de natalidade dos municípios. Teoricamente, poderia ser colocada a hipótese de que a população mais educada teria um menor número de filhos, uma vez que começam a ter filhos mais tarde devido aos anos despendidos nos estudos. Deste modo, com o auxílio de um gráfico de dispersão, utilizou-se para os valores do eixo dos xx a média da variável "Média de anos de escolaridade das mães dos alunos no nível de ensino secundário" e para os valores do eixo dos yy a média da "taxa de natalidade por 1000 habitantes" tendo-se verificado uma forte correlação entre estas variáveis, mas contrariando as expectativas iniciais. Através da análise desta visualização podemos então concluir que quanto maior o número de anos de escolaridade das mães de alunos do ensino secundário, maior é a taxa de natalidade desses municípios. Apesar de poder parecer contraintuitivo à primeira vista, esta

correlação pode ser fundamentada pela relação previamente observada entre o nível de educação das mães dos alunos e o poder de compra per capita da região. Indicando assim que apesar do tempo despendido nos estudos pela população, verifica-se que o fator preponderante para o aumento da taxa de natalidade ao nível das municipalidades portuguesas estará mais fortemente relacionado com a capacidade económica das famílias do que com a disponibilidade adicional das mulheres que optaram por abandonar a atividade letiva (Figura 6).

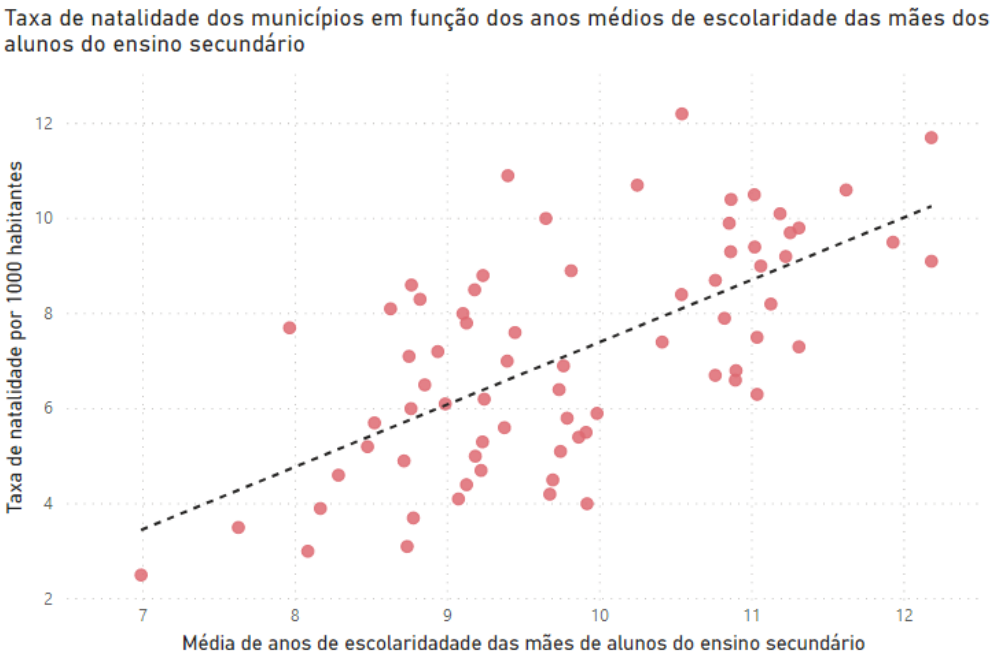

Figura 6 – Aumento da taxa de natalidade média dos municípios em função da média de anos de escolaridade das mães dos alunos do ensino secundário.

A última visualização criada neste trabalho foi um mapa (com recurso à ferramenta da ESRI para PowerBI), com o objetivo de se representar o poder de compra per capita e a taxa de natalidade por 1000 habitantes por cada município, de modo a corroborar a interpretação feita dos dados previamente analisados. Assim, para a localização espacial utilizou-se a variável geográfica "Municipalidade, País", que foi adicionada pela concatenação do nome do país ('Portugal') à variável 'Municipalidade' disponível no dataset das escolas, de modo a restringir os pontos representados ao território nacional. Para a cor dos pontos foi utilizada a variável "taxa de natalidade por 1000 habitantes" e para o tamanho dos pontos a variável "poder de compra per capita". Com esta representação é possível concluir, como previamente especulado, que os municípios com menor poder de compra (círculos mais pequenos) se encontram fortemente associados a taxas de natalidade mais reduzidas (círculos mais claros), como é particularmente visível na região interior-norte do país (Figura 7).

Figura 7 – Média do poder de compra per capita e taxa de natalidade por 1000 habitantes por município. Os pontos de cor mais escura representam taxas de natalidade mais elevadas, enquanto que os pontos de maiores dimensões representam os valores mais elevados de poder de compra per capita.

Finalmente, ao observar todas as representações abordadas num dashboard comum (Figura 8), torna-se possível tirar o máximo partido das capacidades interativas do software PowerBI.

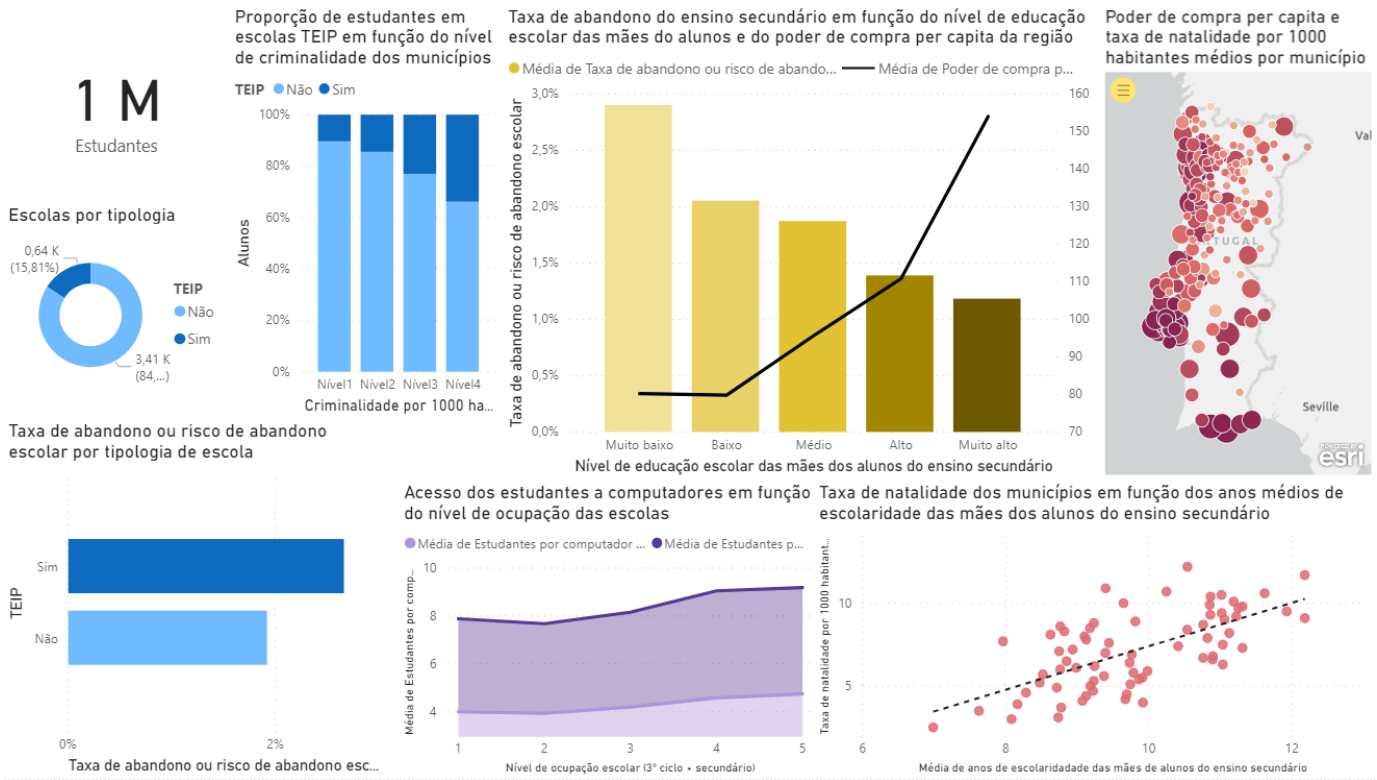

Figura 8 - Dashboard com representação conjunta das visualizações utilizadas para a exploração das variáveis em estudo.

Por exemplo, na visualização de dashboard abaixo apresentada, é possível observar como todos os parâmetros anteriormente descritos se relacionam, quando temos em conta apenas as escolas abrangidas pelo programa TEIP. Em primeiro lugar é possível observar que o número total de estudantes nessas escolas é de 150 mil. É também possível observar que as mães dos alunos pertencentes a estas escolas, apresentam níveis de escolaridade tendencialmente baixos, o que poderá contribuir para as elevadas taxas de abandono escolar. Estas escolas, consideradas de risco elevado, estão em geral distribuídas em redor da região da grande Lisboa e do grande Porto. Para além de nos ajudar a compreender melhor os resultados, esta interatividade pode ainda ajudar a encontrar novos padrões ocultos nos dados ou a expor os resultados de forma dinâmica aquando da apresentação dos mesmos (Figura 9).

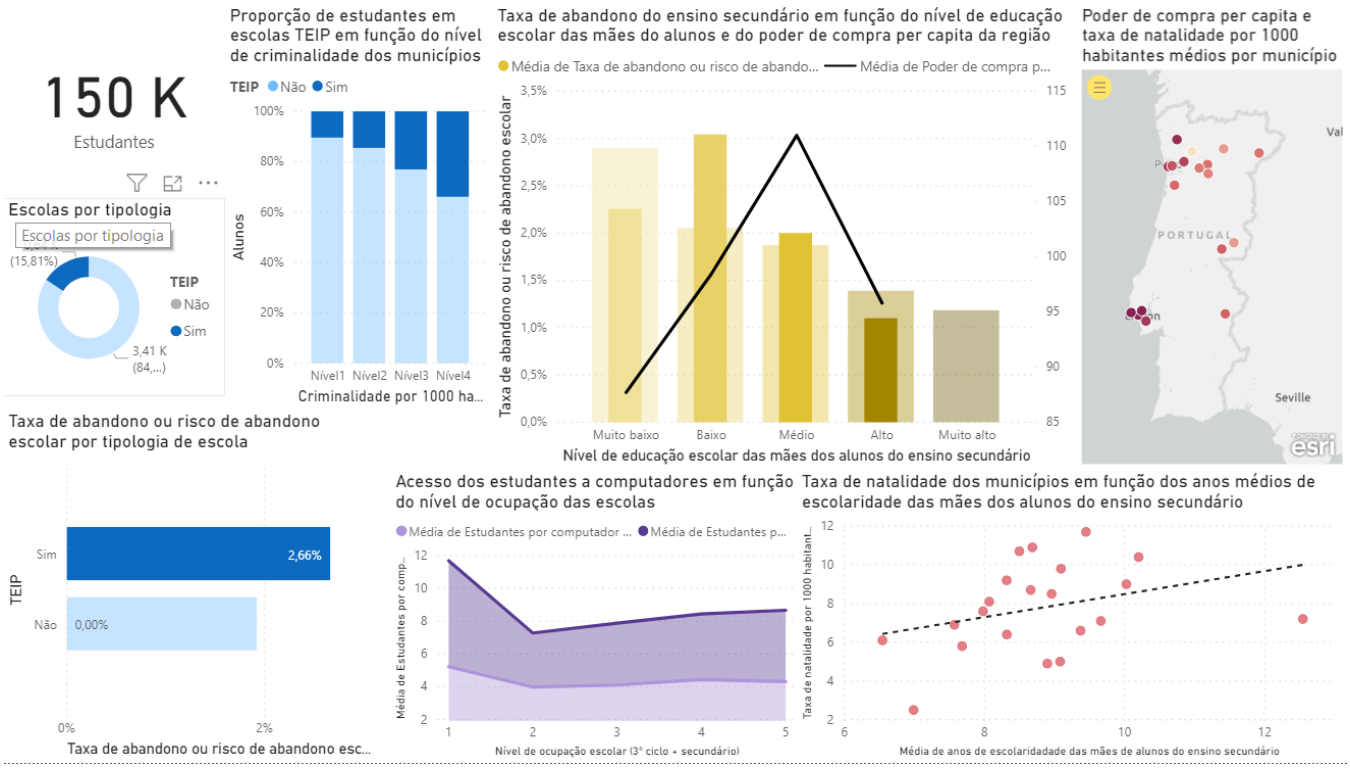

Figura 9 – Dashboard com a variável 'escolas TEIP' selecionada e representação conjunta das visualizações utilizadas para a exploração das variáveis em estudo.

Apreciação geral do software

Através da realização deste trabalho, concluiu-se que o software PowerBI é uma ferramenta versátil que permite ao utilizador criar uma multiplicidade de visualizações de dados simples ou complexas, com recurso a uma interface gráfica, geralmente intuitiva e familiar para os utilizadores frequentes das ferramentas Office.

A possibilidade de importar tabelas de dados das mais diversas fontes, tais como, folhas excel, ficheiros de texto, bases de dados SQL, diversas plataformas de armazenamento de dados em nuvem, etc, conferem a esta ferramenta uma versatilidade singular entre as ferramentas de visualização de dados com interface gráfica.

O PowerBI oferece também uma incrível facilidade na ligação de dados entre tabelas, através da sua ferramenta de criação de modelos. Esta ferramenta permite gerir de forma simples a cardinalidade das ligações entre tabelas, assim como a sua direção, facilitando a exploração dos dados.

Outra vantagem do software é a possibilidade de importação de métodos de visualização alternativos, desenvolvidos por utilizadores ou empresas, expandindo ainda mais o leque de alternativas de visualização disponíveis para as várias dezenas.

Apesar de tudo isto, a funcionalidade mais interessante do software, consiste na possibilidade de elaboração de dashboards interativos, nos quais os dados se encontram conectados entre visualizações, permitindo uma exploração prática e dinâmica do dataset de forma a encontrar padrões ocultos nos dados e explorar facilmente correlações entre variáveis.

Contudo, o software não deixa de apresentar alguns pontos negativos, alguns dos quais foram notados no decorrer da realização deste trabalho. De facto, apesar de apresentar uma interface gráfica intuitiva que permite desempenhar diversas tarefas complexas com simplicidade, o software perde consideravelmente ao tornar morosas ou menos intuitivas tarefas que seriam fáceis de realizar com recurso a outras ferramentas.

Por exemplo, no caso das visualizações do tipo ‘gráfico de colunas empilhadas', houve numa fase inicial a necessidade de trocar a ordem das colunas no gráfico. Uma vez que todas as colunas apresentavam o mesmo número de valores e a sua divisão era feita apenas pela variável inserida no campo ‘legenda’, tornou-se impossível ordenar as colunas pela ordem desejada (caso esta não seja a alfabética). Depois de algum tempo de pesquisa, constatou-se que a solução para este problema passaria pela criação de uma tabela intermédia na qual os dados estariam ordenados da forma especificada, apenas para proceder à ordenação das colunas. Esta solução não só não é intuitiva, como é demorada e torna o ficheiro mais pesado, uma vez que requer o armazenamento de uma maior quantidade de dados. Para além disto, a falta de possibilidade de alteração direta da legenda das figuras constitui também um grande entrave na utilização da ferramenta, que poderia ser facilmente mitigado.

Por fim, apesar de poderoso, o editor de tabelas apresenta uma curva de aprendizagem acentuada e torna por vezes complexas algumas tarefas (como a criação de colunas de agregação de dados) que seriam relativamente simples de efetuar num software programático, ou até no Microsoft Excel, desenvolvido pela mesma companhia.

No geral, este software apresenta um potencial elevado, particularmente ao nível da utilização empresarial, permitindo que qualquer utilizador, mesmo sem experiência no ramo da informática, possa ter acesso a ferramentas de visualização avançadas e técnicas de gestão de dados que até à data se encontravam disponíveis apenas em ferramentas programáticas.

No que diz respeito à utilização em meio científico, esta ferramenta encontra uma forte competição nas bibliotecas de visualização disponíveis para a linguagem Python, tais como o Matplotlib ou Seaborn, que devido à sua modularidade e integração na linguagem Python permitem aos utilizadores mais avançados um maior controlo sobre o tratamento de dados e sobre os diversos detalhes das visualizações geradas,

sem comprometer a capacidade de integração com diversos sistemas de armazenamento de dados.

A possibilidade de interação direta com os gráficos da dashboard para edição de títulos, legendas ou ordenamento das variáveis em exibição, tornaria a ferramenta potencialmente mais apelativa para os diversos grupos de utilizadores. A implementação de uma interface interativa (semelhante à do microsoft excel) para edição das tabelas, poderia também tornar o software mais apelativo para novos utilizadores e menos demorado para os utilizadores mais experientes.

Conclusão

Através das diferentes técnicas de visualização disponíveis no software Power BI, como a construção de mapas, gráficos em anel, gráficos de dispersão, gráficos de linhas e colunas, cards, gráficos de áreas empilhadas, entre outras, foi possível explorar o conjunto de dados fornecido sobre o ensino público português e relacioná-lo com variáveis importadas do portal PORDATA que se achou serem de interesse para a análise. Assim sendo, foi possível concluir através da construção de um dashboard que o nível de escolaridade das mães é um parâmetro bastante importante por ter influencia na taxa de abandono ou risco de abandono escolar e uma forte correlação com as variáveis poder de compra e taxa de natalidade da região. Foi também possível visualizar que as escolas abrangidas pelo programa TEIP são as escolas localizadas em regiões de maior criminalidade e onde se verifica um maior abandono escolar. Para concluir, o software PowerBI demonstrou-se bastante útil, na medida em que facilitou a exploração dos dados em estudo e permitiu a criação de gráficos iterativos de elevada qualidade visual que facilitam a divulgação explicita dos resultados obtidos.