Visualização de Dados

2021/2022

2º Projeto

Visualização de Dados Científicos dashboards usando o Power BI Desktop

Grupo VD01

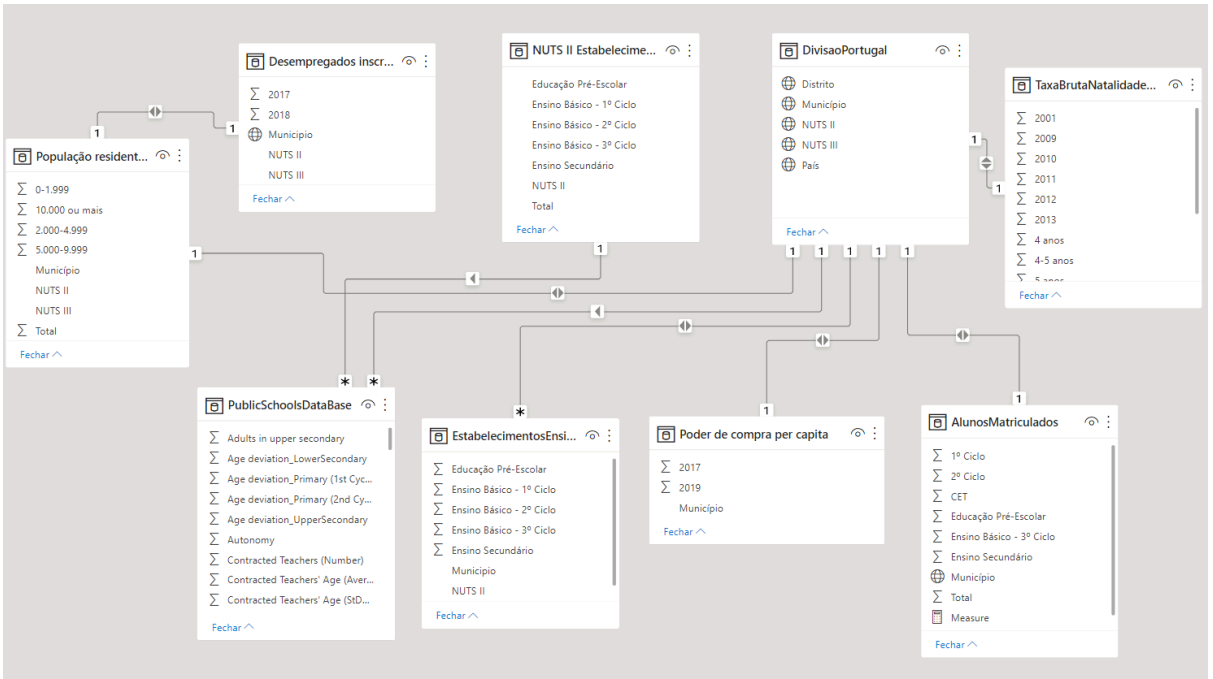
André Monteiro n.º 51718
Carolina Magro n.º 55817
Tiago Duarte n.º 55780

Introdução e tratamento das bases de dados utilizadas

Neste projeto tem-se como objetivo a exploração do PowerBI, uma ferramenta que permite a construção de dashboards para visualização de dados. Recorreram-se a dashboards para identificar correlações entre as variáveis existentes no conjunto de dados disponibilizado pelas docentes e variáveis adicionais que foram obtidas no portal PORDATA. Mais especificamente, os dados são os seguintes:

- Dados de escolas públicas pertencentes ao território de Portugal Continental, facultados pela Direção-Geral de Estatísticas da Educação e Ciência (DGEEC) e disponibilizados pelas docentes;
- População residente segundo os Censos (2011): total e por dimensão dos lugares; PORDATA - População residente segundo os Censos: total e por dimensão dos lugares
- Desempregados inscritos nos centros de emprego e de formação profissional; PORDATA - Desempregados inscritos nos centros de emprego e de formação profissional
- Matriculados por níveis de ensino; PORDATA - Alunos matriculados nos ensinos pré-escolar, básico e secundário público: total e por nível de ensino
- Taxa bruta de natalidade; PORDATA - Taxa bruta de natalidade
- Estabelecimentos pré-escolar, básico e secundário; PORDATA - Estabelecimentos nos ensinos pré-escolar, básico e secundário: por nível de ensino
- Poder de compra per capita; PORDATA - Poder de compra per capita

Na imagem seguinte estão representadas as várias tabelas utilizadas e as relações entre si.



Para que todos os dados de municípios fossem ligados entre as tabelas principais e as adicionais foi criada uma nova tabela, denominada “DivisãoPortugal". Esta tabela contém o NUTS III, NUTS II e o distrito de cada município. Foi necessária a sua criação dado que surgiam dificuldades quando os dados principais e adicionais se encontravam ligados diretamente. O mesmo acabou por não ser necessário na Tipologia de Áreas Urbanas

(TIPAU), dado que se fez a ligação apenas à base de dados das escolas públicas. Futuramente, caso fosse necessário adicionar mais alguma base de dados com áreas urbanas, seria adicionada outra tabela como foi feito com a tabela “DivisãoPortugal”.

Técnicas de visualização de dados disponíveis no PowerBI

Gráficos de barras empilhadas e Gráfico de colunas empilhadas: Representam barras/colunas respeitando os dois eixos de coordenadas X e Y. É possível apresentar várias variáveis empilhadas na vertical (colunas) ou na horizontal (barras).

Gráfico de barras agrupadas e Gráfico de colunas agrupadas: Semelhante ao anterior, diferem no método de apresentação, que em vez de ser sequencial (em cima) é feito um agrupamento (ao lado). Nestas visualizações é mais fácil a interpretação para a apresentação de valores com menor relação entre si.

Gráfico de barras 100% empilhadas e Gráfico de colunas 100% empilhadas: É possível representar valores em barras/colunas em que todas têm o mesmo tamanho, mas em que os valores são distribuídos por essa área de acordo com a sua percentagem de ocorrência.

Gráfico de linhas: são utilizados para valores quantitativos num intervalo contínuo e são representados através de pontos ligados por uma linha. É útil para análise de tendências e variação das variáveis escolhidas nos eixos.

Gráfico de área: Semelhante ao anterior, com a diferença de ser possível colorir a área formada entre a linha dos dados e os eixos do gráfico.

Gráfico de áreas empilhadas: Semelhante ao anterior, mas neste não há sobreposição de cores. Somam-se os valores para cada ponto (semelhante ao “gráfico de colunas/empilhadas”) havendo um empilhamento de dados.

Gráfico de linhas e de colunas empilhadas: Aglutina o “gráfico de linhas” e o “gráfico de colunas empilhadas”.

Gráfico de linhas e de colunas agrupadas: Semelhante ao anterior, com a diferença que gera novas colunas distintas para cada variável. As linhas representam várias variáveis que facilitam a comparação com as colunas.

Gráfico do friso: É útil para mostrar as alterações de classificação, uma vez que mostra a classificação mais elevada na zona superior do gráfico.

Gráfico de cascata: É utilizado para demonstrar o efeito de valores positivos ou negativos num valor inicial.

Funil: Os dados são divididos em conjuntos (normalmente dados progressivamente menores) e são representados em forma de funil.

Gráfico de dispersão: É possível representar 4 dimensões de informação neste gráfico. Como valores de coordenadas dos eixos x e y, em que o tamanho dos pontos

representados varia consoante uma variável e é possível atribuir cor para distinguir os pontos. A quarta dimensão costuma ser para mostrar a variação destes valores ao longo do tempo.

Gráfico circular: Divide sectores ilustrando uma proporção numérica. A soma de todas as divisões é igual a 1 unidade.

Gráfico em anel: Muito semelhante ao anterior, mas com centro vazio, sendo possível adicionar mais informação no mesmo.

Treemap: É possível visualizar a estrutura hierárquica de um diagrama de árvore, com a vantagem de ser possível, através da área de cada quadrado, evidenciar as quantidades para cada categoria.

Mapa: Permite sobrepor a um mapa variáveis e dados através de grafismos (círculos, por exemplo).

Mapa de manchas: Permite sobrepor a um mapa variáveis e dados através de manchas coloridas, definidas através de outra variável.

Medidor: É útil para representar, máximo, mínimo e médias.

Cartão: Apresenta uma visualização básica de uma variável de dados quantitativos.

Cartão de linhas múltiplas: Como o "Cartão", no entanto permite o preenchimento de várias linhas com múltiplas variáveis.

KPI: "Key Performance Indicator" é útil para medir o progresso dum objetivo.

Segmentação de dados: Apresenta-se como um conjunto de campos que podem ser selecionados e permitir a filtragem dos elementos interativos no mesmo *dashboard*.

Tabela: Uma tabela comum que pode ser formatada tal como no Excel.

Matriz: Representa uma relação entre variáveis e a sua dimensão varia conforme número e índices utilizados.

R script visual: Importa e desenvolve scripts em linguagem R.

Elemento visual em Python: Semelhante ao anterior, mas em linguagem Python.

Principais influenciadores: Permite a análise dos dados e apresenta os influenciadores principais.

Árvore de decomposição: É útil para visualização de dados ramificados, em que é possível selecionar um nó específico e mostrar apenas a sua ramificação.

Perguntas e Respostas: Permite a interrogação do utilizador acerca dos dados apresentados no dashboard e a resposta pelo PowerBI.

Elementos importados

O PowerBI oferece a possibilidade de importar outros tipos de gráficos que não são apresentados na área principal da ferramenta.

Histograma - O histograma irá fazer a distribuição dos elementos por intervalos contínuos, sendo que cada uma das barras irá representar a frequência dos dados em cada intervalo. Neste gráfico é mais perceptível onde a maior quantidade de valores está concentrada e quais os extremos desses valores.

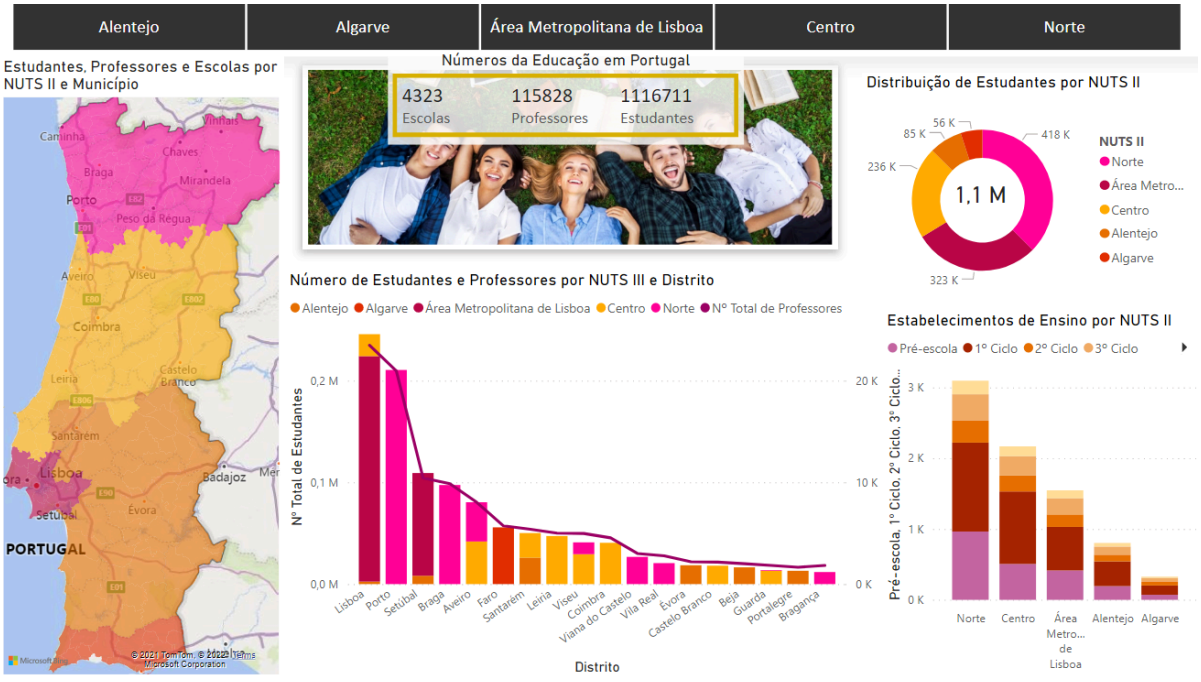
Heat Map - permitem mostrar a densidade, através de cores, de uma determinada variável no mapa. Graças a este mapa é possível detectar onde existe maior e menor densidade através de uma escala de cores.

Advanced Card - O cartão avançado permite mudar mais facilmente a sua aparência do que o "Cartão".

Visualizações criadas

De seguida serão apresentadas as visualizações criadas em cada dashboard, explicando e interpretando as diferentes representações gráficas. Para facilidade de referência e simplicidade, as visualizações são descritas da esquerda para a direita. Em todos os dashboards é apresentada uma "segmentação de dados" no topo, para se poderem filtrar os dados por NUTS II (Alentejo, Algarve, Área Metropolitana de Lisboa (AML), Centro, Norte) individualmente ou em conjunto, com exceção do último onde tal pode ser feito noutro elemento gráfico.

Dashboard 1 - Ensino Público



No primeiro *dashboard* apresentado, exploram-se os dados em relação à distribuição de alunos, escolas e professores em Portugal Continental, tendo sido geradas 6 visualizações.

1. Visualização “Cartão de linhas múltiplas” com informação acerca do número de estudantes, número de escolas e número de professores. Este número irá alterar consoante a filtragem feita no “Segmentação de dados” ou num elemento do *dashboard*. Esta visualização fornece ao leitor informações concretas e sintetizadas do que é observado no *dashboard*. O número total de professores foi calculado, criando uma nova coluna “Professores (Total)” através da soma das colunas “Contracted Teachers (Number)” e “Permanent Teachers (Number)”. Foi ainda adicionada uma imagem relativa a estudantes para colocar como fundo e aumentou-se a transparência do cartão.

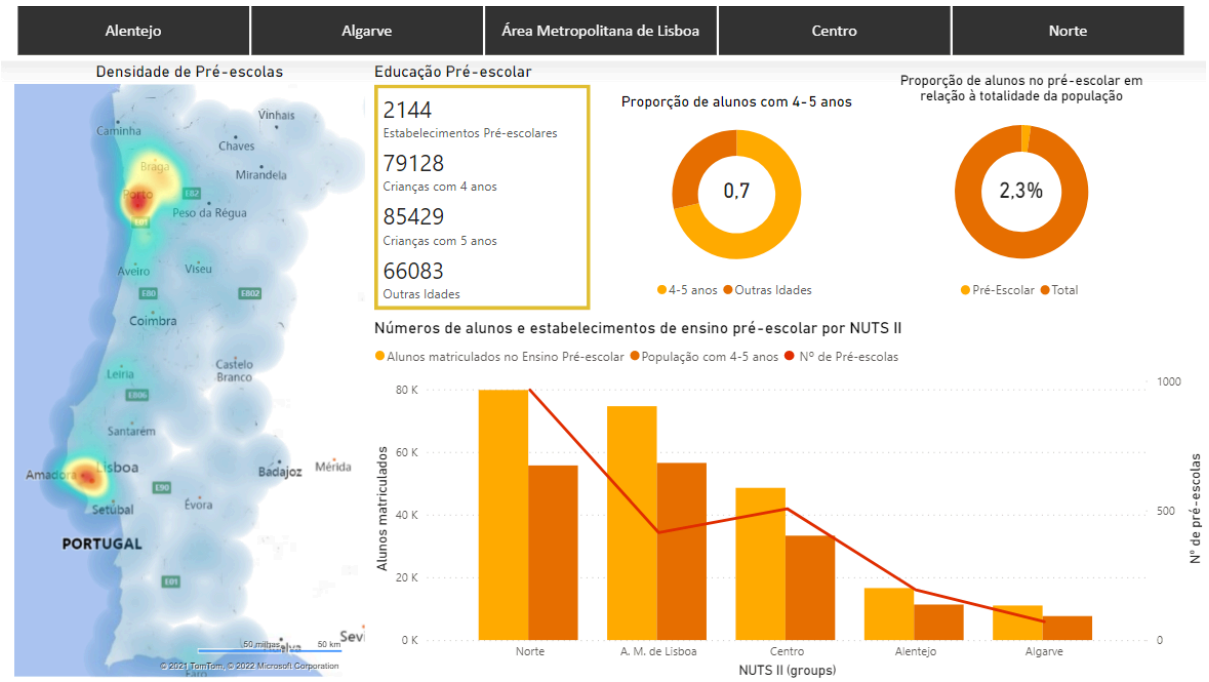
2. Visualização “Mapa de manchas” de Portugal Continental, preenchido com os diferentes municípios e coloridos de acordo com as NUTS II. Foram adicionados os parâmetros de número de alunos, professores e escolas ao campo "Descrições" para que possam ser representados estes valores cada vez que o leitor clicar ou passar o rato em cada município.

3. Visualização “Gráfico de linhas e colunas agrupadas”, onde mostra o número total de estudantes (eixo dos y) agrupados por NUTS II e de acordo com os Distritos que pertencem (eixo dos x).
Há ainda uma linha referente ao número de professores existentes por cada Distrito, o valor referente a cada ponto da linha pode ser lido à direita no gráfico ou colocando o cursor por cima do ponto de interesse.
Como demonstra o gráfico e o que seria expectável, existe um maior número de alunos no distrito de Lisboa e Porto, seguindo-se predominantemente Setúbal, Braga e Aveiro. Nos territórios referentes ao NUTS II (Centro e Alentejo) existem menos alunos.
O número de professores que é representado pela linha já referida anteriormente acaba por acompanhar a distribuição de alunos existente, o que seria expectável pois é necessário um número maior de professores onde existe uma maior agregação de alunos.

4. Visualização “Gráfico em anel” que mostra a proporção de alunos nas diferentes NUTS II.
Esta visualização além de apresentar os NUTS II que estão divididos por cores, também mostra o número de estudantes em cada NUTS II.
Foi adicionada uma visualização "Cartão" ao centro desta, contendo o número de alunos total selecionado, com uma casa decimal. Este número irá alterar da mesma forma que altera no “Cartão de linhas múltiplas”, caso alteremos a filtragem feita na “Segmentação de dados” ou num elemento do dashboard.
Observa-se que o número de alunos é maior no Norte, seguindo da Área Metropolitana de Lisboa, Centro, Alentejo e finalmente Algarve. Observa-se também que o número de alunos diminui de acordo com a seguinte sequência: AML, Norte, Centro, Alentejo e Algarve.

5. Visualização “Gráfico de colunas empilhadas”, representando a soma dos estabelecimentos de ensino, bem como a sua tipologia (pré-escolar, 1º ciclo, 2º ciclo e secundário) dentro de cada NUTS II. Através desta visualização é possível observar que o número total de estabelecimentos de ensino é mais elevado do Norte, seguindo-se do Centro, AML, Alentejo e Algarve. Segundo os outros gráficos referidos anteriormente seria expectável a AML conter mais escolas que o Centro. No entanto, o que se verifica é a falta de escolas na AML comparativamente ao número de alunos presentes nesta zona, o que resulta numa sobrecapacidade das mesmas.

Dashboard 2 - Pré-escola



No segundo *dashboard* é abordado o Ensino Pré-escolar. Para este *dashboard* recorreu-se à taxa de natalidade para calcular o número de crianças com idades entre 4 e 5 anos. Como os dados dos alunos matriculados são de 2018, foi possível utilizar os dados de natalidade de 2012 e 2013 e inferir o número de crianças com 4 e 5 anos. É ainda pressuposto que as crianças que nasceram naquele município se mantiveram no mesmo município.
Foram ainda utilizados os dados dos alunos matriculados.

1. Visualização "*Heat map*" que apresenta a densidade de pré-escolas em Portugal Continental. Esta foi importada utilizando o "*PowerBi App source*". Tal como já concluído nas outras visualizações, pode-se observar uma maior densidade de pré-escolas nas Áreas Metropolitanas de Lisboa e Porto, mantendo-se intermédia ao longo do litoral entre estas duas cidades. O resto do país acaba por apresentar densidades mais baixas, especialmente no caso do Alentejo que tem uma densidade bastante baixa.

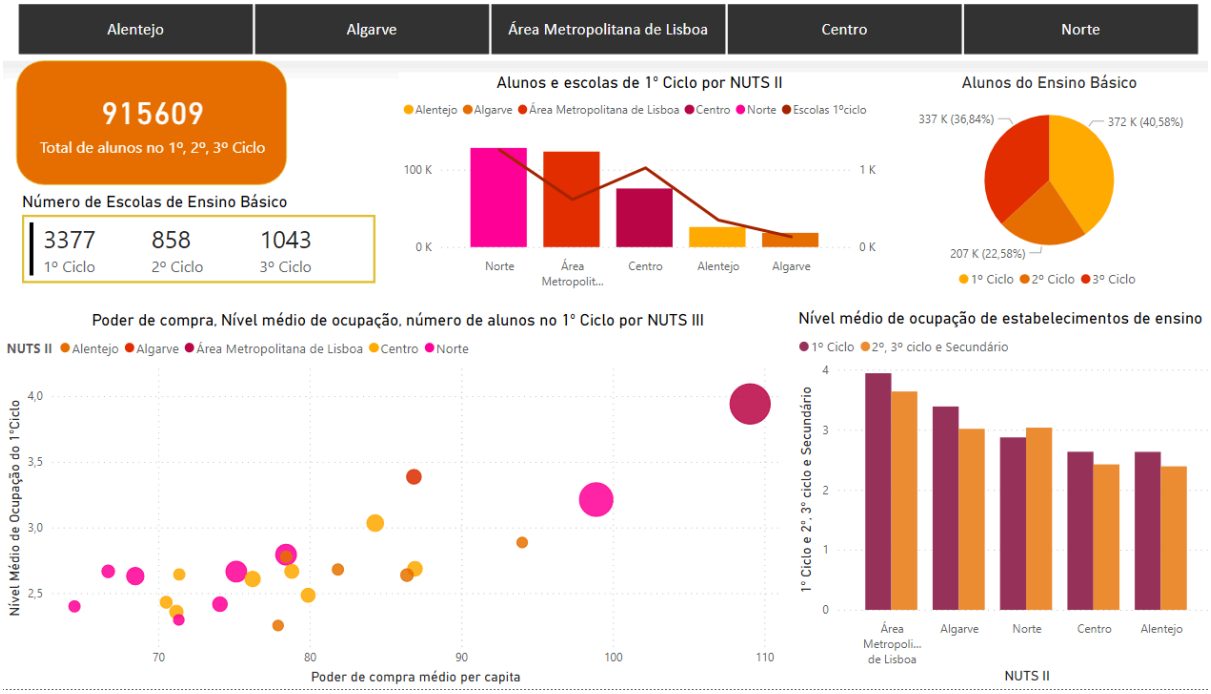
2. Visualização "Cartão de linhas múltiplas" com o número de estabelecimentos pré-escolares, o número de crianças matriculadas com 4 anos, 5 anos e outras idades. Este número irá alterar consoante a filtragem feita no "Segmentação de dados" ou num elemento do *dashboard*.

3. Visualização "Gráfico de linhas e colunas agrupadas", que permite visualizar a proporção de alunos de 4-5 anos de idade em relação à totalidade de alunos no ensino pré escolar.
Foi também adicionada uma linha de tendência da quantidade de pré-escolas existentes por cada NUTS II, o valor referente a cada ponto da linha pode ser observado no eixo direito no gráfico ou passando o cursor por cima do mesmo.
Novamente, tal como no dashboard anterior, pode-se observar que o rácio entre número de alunos e número de pré-escolas encontra-se constante nas NUTS II excepto na AML, onde é bastante superior. Conclui-se que há poucos estabelecimentos pré-escolares na AML comparativamente ao número de crianças existentes.

4. Visualização “Gráfico em anel” que mostra a proporção entre a soma dos alunos entre 4 e 5 anos de idade, no ensino pré escolar público, e o número total de alunos de outras idades na mesma situação. Tal como no *dashboard* 1, foi adicionada uma visualização “Cartão” dentro desta visualização, mas desta vez em percentagem. Ambas visualizações funcionam em conjunto e irão alterar-se de acordo com o filtro selecionado.

5. Outra visualização “Donut Chart” que irá funcionar da mesma forma que a visualização anterior. Nesta, é representada a proporção de alunos no ensino pré-escolar em relação ao número total de população residente no país.

Dashboard 3 - Ensino Básico



Neste terceiro *dashboard* são utilizados os dados do ensino básico (1º ciclo, 2º ciclo e 3º ciclo do ensino). Para tal, foram criadas visualizações que permitem perceber a relação entre variáveis como:

- O número e distribuição de alunos
- Número de escolas
- Influência do poder de compra per capita
- Nível de ocupação

1. Visualização “*Advance Card*” com o número total de estudantes de Portugal Continental matriculados no Ensino Básico. Esta visualização foi também importada utilizando o “*PowerBi App source*”, e trata-se de uma versão mais completa e avançada do “Cartão de linhas múltiplas”.

2. Visualização “Cartão de linhas múltiplas” com o número de escolas do Ensino Básico de Portugal Continental. Este número, à semelhança de situações anteriormente descritas, variará conforme o filtro aplicado.

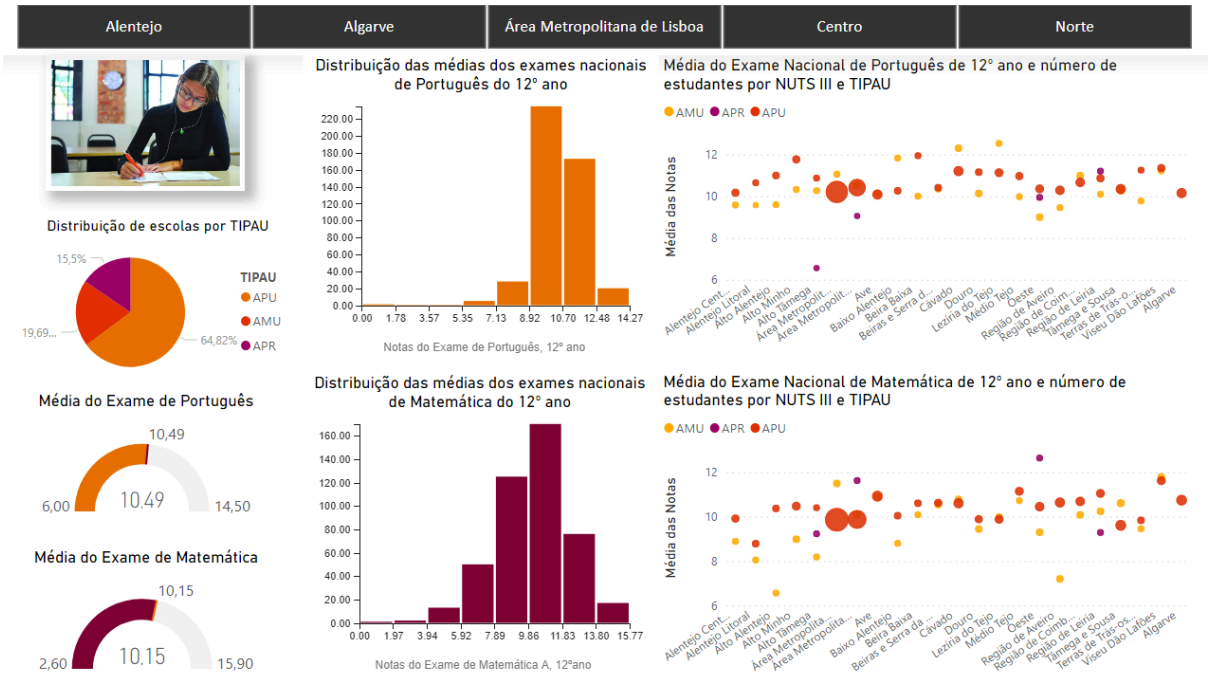
3. A visualização “Gráfico de dispersão” permite relacionar o poder de compra per capita com o nível médio de ocupação do 1º ciclo e a densidade populacional, nas diferentes NUTS III. Quanto maior for cada círculo, maior é a densidade populacional e este posiciona-se no eixo dos x de acordo com o poder de compra per capita, e no eixo dos y de acordo com o Nível médio de ocupação do 1º ciclo. A sua cor indica a que NUTS II corresponde.
É possível observar um maior poder de compra per capita nas Áreas Metropolitanas de Lisboa e Porto, o que está certamente relacionado com a densidade populacional elevada e, por consequência, o nível de ocupação elevado nas escolas de 1º ciclo.

4. A visualização “Gráfico de linhas e de colunas agrupadas” representa o número de alunos como colunas e como linhas temos o número de escolas, distribuídos por NUTS II.
Nesta visualização voltam-se a concluir que o Norte tem um maior número de alunos, seguindo-se a AML, Centro, Alentejo e Algarve. Novamente observa-se a possibilidade da sobrelotação das escolas de 1º ciclo na AML, dado o número elevado de alunos vs o número de escolas existentes.

5. A visualização “Gráfico circular” apresenta a proporção do total de alunos por nível de ensino. Dado que a quantidade de anos escolares em cada ciclo é diferente (quatro anos no 1º ciclo, 2 anos no 2º ciclo e 3 anos no 3º ciclo), há uma maior quantidade de alunos no 1º e 3º ciclos, em comparação com o 3º.

6. Visualização “Gráfico de colunas agrupadas” com o nível médio de ocupação dos estabelecimentos de ensino nas diferentes NUTS II, para o 1º Ciclo e para a média dos restantes níveis de ensino (por impossibilidade de separação destes últimos). Nota-se que a Área Metropolitana de Lisboa apresenta um maior nível médio de ocupação em relação aos restantes distritos. A esta seguem-se o Algarve, Norte, Centro e Alentejo.

Dashboard 4 - Exames Nacionais e TIPAU por NUTS III



Neste *dashboard* foi abordada a potencial relação entre a “tipologia da área urbana" e as notas dos exames nacionais das disciplinas de português e de matemática. Agruparam-se as variáveis pelos diferentes NUTS III e verificaram-se também as respectivas tipologias “APU” - área predominantemente urbana, “APR”- área predominantemente rural, “AMU” - área mediamente urbana”.
Foram utilizados 4 tipos de representações gráficas:

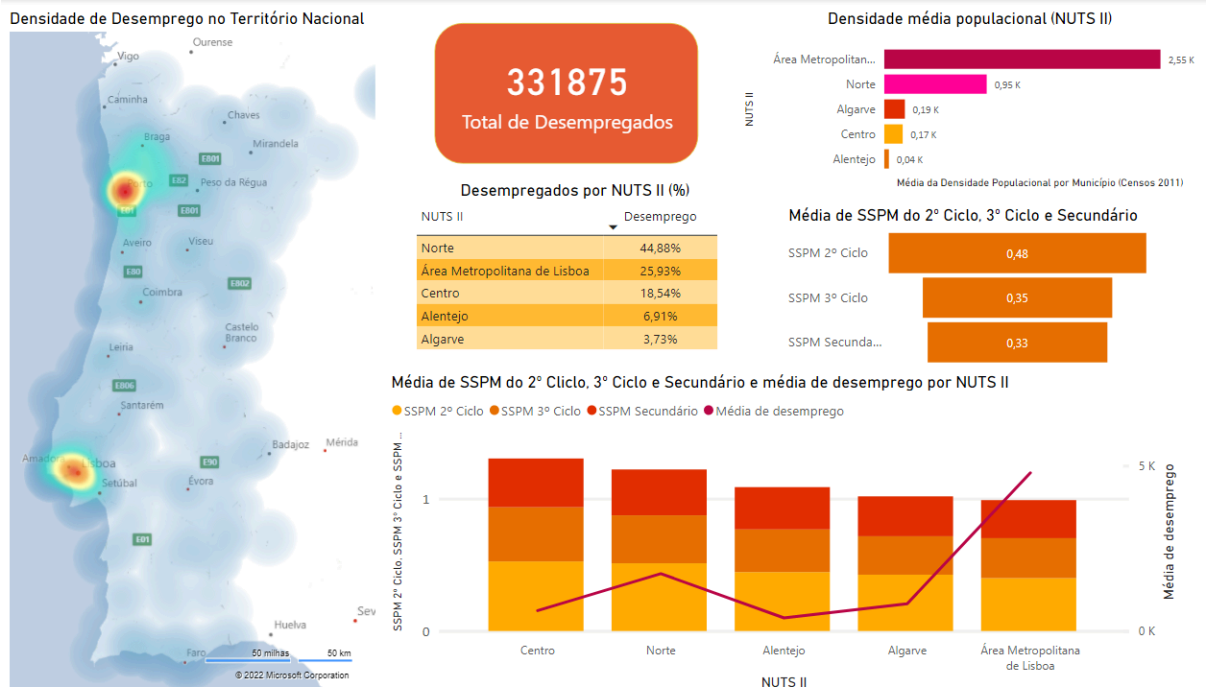
1. Visualização “gráfico circular”, com a distribuição das escolas pelos vários tipos de área urbana, dividido por AMU, APR e APU. Foi possível verificar que quanto menor a urbanização de uma área, menor é o número de escolas presentes.

2. Através de dois gráficos “medidores”, representam-se os valores mínimos, médios e máximos dos resultados dos dois exames nacionais. Os valores mínimo e máximo, são relativos aos maiores e menores valores de classificação média apresentados por cada escola, para cada tipo de exame. O valor evidenciado no medidor representa a média nacional de cada exame. A barra preenchida dá-nos uma percepção da "distância" a que uma determinada região está a alcançar o máximo ou mínimo do gráfico, bem como a distância à média nacional.
Ao analisar a variação das médias em relação às regiões e às TIPAU através da componente interativa, conclui-se que, de uma forma geral, as médias rondam os 10 valores, com a exceção do exame de Português nas APR em que a média é de 9.19 valores. É assim possível aferir através da interatividade quais as zonas em que obtêm os piores e os melhores resultados, e ainda verificar que as notas dos dois exames de português e de matemática diferem, mesmo que ligeiramente.

3. Para demonstrar a distribuição das várias médias dos exames, utilizam-se “histogramas” em que o eixo dos xx apresenta as médias e o eixo dos yy apresenta o número de escolas em que estas médias foram alcançadas. Foram também importados utilizando o “*PowerBi App source*”.
No exame de Português, a maioria dos resultados das médias das escolas encontram-se nos intervalos de notas entre 8.92 a 10.70 valores, em 235 escolas. No exame de matemática há uma maior distribuição dos resultados (e consequentemente das frequências) do que no exame de Português. Há mais resultados entre 9.86 a 11.83 valores, em 170 escolas.

4. Os “gráficos de dispersão” representam uma relação entre as zonas NUTS III, as médias das notas de cada exame e a TIPAU correspondente. A variação de cores representa as diferentes TIPAU, enquanto o tamanho dos pontos representa a densidade populacional de estudantes de cada zona.
Quanto ao exame de Português, as zonas com médias mais altas são locais medianamente urbanos (AMU) como a Lezíria do Tejo (12.53 valores) e Cávado (12.30 valores). Nota-se que apesar de uma diferença mínima, é possível verificar um maior conjunto de melhores notas em regiões predominantemente urbanas (APU). As piores notas estão no Alto Tâmega e na AML. No exame de Matemática, observa-se uma tendência de melhores notas nas zonas predominantemente rurais (APR). Como no Oeste, com 12.64 valores e a nas APR da AMP, com 11.63 valores.

Dashboard 5 - Desemprego e SSPM



Neste *dashboard* foi abordada a potencial relação entre os "Desempregados inscritos nos centros de emprego e de formação profissional" e o "Successful Student Paths Measure" (SSPM). Agruparam-se as variáveis pelos diferentes NUTS II.
Além destas variáveis foram utilizadas ainda outras presentes na base de dados fornecida (Dados do ensino público). Todos são referentes a 2018, para a visualização ser o mais coerente possível.

1. A visualização "*Heat map*" contém a densidade de desemprego em Portugal Continental. Foi utilizado este tipo de mapa pois é mais eficaz de entender onde se distribui a maior densidade de desempregados inscritos nos centros de emprego e de formação profissional. Pode-se observar uma maior densidade de desempregados nas Áreas Metropolitanas de Lisboa e Porto, mantendo-se intermédia ao longo do litoral entre estas duas cidades. Expectavelmente, o resto do país acaba por apresentar densidades mais baixas, especialmente no caso do Alentejo que tem uma densidade bastante baixa.

2. Visualização do "*Advance Card*" com o número total de estudantes desempregados inscritos nos centros de emprego e de formação profissional. Este cartão permite ter uma noção geral da quantidade de desemprego existente no país.

3. A visualização "Tabela" apresentada relaciona a percentagem de desempregados pelos diferentes tipos de NUTS II. Nesta tabela podem-se filtrar os dados presentes neste *dashboard* clicando em cada linha correspondente à zona a selecionar. É ainda possível ordenar os dados por ordem de grandeza de desemprego ou por ordem alfabética das zonas.
É possível, assim, a observação dos dados em pormenor que estão representados no *"Heat map",* em que se verifica que o desemprego é maior nas AML e AMP. Já no Algarve e Alentejo a percentagem de desemprego é bastante menor, porém a densidade populacional também.

4. A visualização “Gráfico de barras empilhadas”, demonstra a densidade média populacional de cada município, por NUTS II.
Pode-se observar que a AML concentra a maior densidade populacional, com cerca de 2550 pessoas/km^2 em cada município, enquanto que no Alentejo este número desce para cerca de 44 pessoas/km^2.

5. Foi usada a visualização “Funil” onde, com as variáveis SSPM, foi determinada a média para cada ciclo de estudos (2º ciclo, 3º ciclo e Secundário). Sempre que se filtra por uma região em específico, os dados do funil irão alterar para as médias dessa mesma região, isto graças a interação que existe no *dashboard*.
Observa-se que em todas as regiões há uma predominância do sucesso escolar no 2º ciclo do Ensino Básico e que esse sucesso vai decrescendo até ao Ensino Secundário.

6. A visualização “gráfico de linhas e colunas empilhadas” além de apresentar a soma das classificações médias SSPM das escolas, também contém uma linha que define o total do desemprego. O conjunto utilizado para ambos é o NUTS II.
A linha tem um valor máximo na AML. Este valor é bastante alto em relação aos outros que são apresentados. Curiosamente, a AML é o local onde se encontra o menor valor de sucesso dos estudantes, tal pode ser observado na na visualização “Funil”.
Pode-se inferir que um dos potenciais motivos do desemprego ser elevado na AML está relacionado com o menor valor de sucesso estudantil. No entanto, tal não é linear, uma vez que na região Norte os valores de SSPM são os segundos mais altos e mesmo assim esta é a segunda região com maior desemprego.

Avaliação do PowerBI

A utilização do software foi intuitiva principalmente para quem está habituado à interface de utilização do Word, Power Point, Excel e até Access. As funcionalidades que o Power BI suporta são variadas e abrangentes, incluindo a possibilidade de importar dados de fontes e formatos muito variados, assim como a edição no próprio programa dos dados importados, o que se revela útil quando se trabalha em grupo, sem a necessidade de ter as bases de dados presentes no sistema para as editar e visualizar. Para além das variadas visualizações disponíveis, há uma boa oferta de visualizações adicionais para descarregar gratuitamente.

No entanto, o software tem algumas limitações, como a atribuição errada de municípios a outros locais que não Portugal, mesmo especificando o País em questão. No caso deste trabalho, o local “Lagos” não apareceu no mapa do *Dashboard* 1, por ser também uma cidade na Nigéria. Outra questão é o facto das cores das visualizações se alterarem quando o ficheiro é aberto e reaberto, problema que só ocorre pontualmente. Ao utilizar o “*Heat map”* também se encontram entraves, uma vez que para a distribuição de dados aparecer diante do mapa é necessário mexer na posição do mesmo ou aumentar/diminuir zoom. Por fim, a interface de utilização poderia ser melhorada se fosse possível editar campos de informação das visualizações, não só o texto como a sua posição relativa, e dimensão.