Grupo 30 - Projeto 2 - Visualização de Dados

Alexandre Monforte (54491), Ariana Dias (53687), Maria Eduarda Pimentel (54525), Ricardo Pedro (55309)

Estudo e visualização de um conjunto de dados, obtidos apartir do site da FCUL, utilizando o *software PowerBI* da *Microsoft*. Análise do abandono escolar usando os dados dos anos letivos 2016/17 até 2019/20, por Grau, Curso e Departamento. Tentativa de encontrar possíveis relações entre os dados. Exploração de várias possibilidades de representação dos dados com o *PowerBI* e apontar suas possíveis limitações.

I. INTRODUÇÃO

O objetivo do presente trabalho foi analisar dados da Faculdade de Ciências da Universidade de Lisboa de modo a fazer uma apreciação sobre o sucesso escolar dos alunos ao longo dos seus percursos académicos e procurar possíveis explicações para o abandono escolar. Para isto, as três principais bases de dados consistiram em:

- Sucesso escolar por UCs e por departamentos
- Abandono escolar
- Colocações e opções de candidatura no Concurso Nacional de Acesso

Todos estes dados foram obtidos no *site* da Faculdade, representado na Figura 1, em formato Excel, abrangendo quatro anos letivos: 2016/17, 2017/18, 2018/19 e 2019/20. O tratamento dos mesmos será explicado na Secção II.

Figura 1. Portal do *website* da Faculdade de Ciências da Universidade de Lisboa onde foram obtidos os dados utilizados para o presente trabalho.

II. VARIÁVEIS ANALISADAS E RESPECTIVAS VISUALIZAÇÕES

PowerBI

O tratamento dos dados foi feito com recurso à ferramenta *PowerBI*, da *Microsoft*. *Power BI* é um conjunto de ferramentas de análise de dados e visualização que ajuda os usuários a transformar dados em *insights* e compartilhar esses *insights* com os outros. Ele inclui uma série de ferramentas de análise, incluindo *dashboards*, relatórios e painéis que permitem aos usuários criar visualizações de dados e compartilhar as informações de maneira fácil e intuitiva. *Power BI* é uma plataforma de análise de dados muito poderosa e flexível que é amplamente utilizada em empresas de todos os tamanhos em todo o mundo.

Este permite a visualização dos dados em vários estilos, alguns destes representados na Figura 2. Para este trabalho utilizou-se:

- clustered bar chart
- clustered column chart
- pie chart
- line chart
- 100% stacked column chart

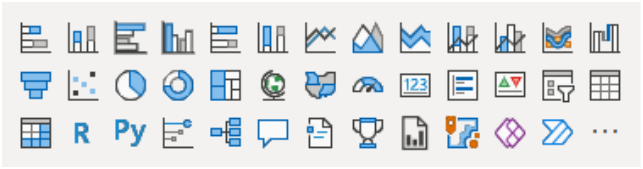


Figura 2. Tools predefinidas disponíveis pelo *PowerBI*.

Abandono Escolar

A realização da análise do abandono escolar fez-se por Grau (Licenciatura e Mestrado Integrado, Mestrado e

Doutoramento) e por curso (os 16 cursos que a FCUL disponibiliza).

Iniciou-se por analisar como evoluía o abandono escolar por Graus, ao longo dos anos que cada grau necessita para a sua conclusão: 3 anos licenciatura, 5 anos mestrado integrado, 2 anos mais dissertação para o mestrado e somente a dissertação para o grau doutoramento. Para a representação desses dados, optou-se por usar gráfico de barras. Este tipo de comparação é ideal para comparar de uma forma direta as diferentes variáveis. Para este estudo, decidiu-se organizar o abandono, em percentagem, agrupando para cada ano de estudo do grau, os anos letivos de 2016/17 até ao ano letivo 2019/2020 (ano do início da pandemia). Teria sido interessante comparar o abandono entre os anos anteriores à pandemia COVID-19 com os anos durante a pandemia, mas por não termos acesso a anos posteriores ao ano letivo 2019/2020, esse estudo tornou-se impossível. Contudo, decidimos, ao mesmo tempo da análise da evolução da percentagem de abandono por ano, verificar se o primeiro ano da pandemia afetou a percentagem de abandono. Da Fig.3 até à Fig.5, encontram-se os gráficos do estudo em questão, cujos dados foram obtidos dos excels *Abandono por curso - 20XX/YY* (onde *XX* e *YY* representam o ano letivo) retirados do site da FCUL.

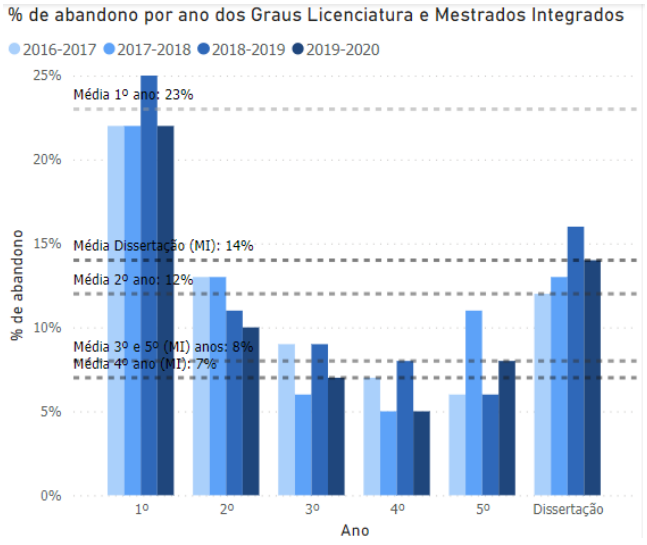


Figura 3. Percentagem de abandono por ano dos Graus Licenciatura (do 1ºano até ao 3ºano) e Mestrado Integrado (do 1ºano até ao 5ºano), agrupados por anos letivos de 2016/17 até 2019/20. A linha a tracejado representa a média por ano do Grau.

Ao longo desta análise, deparámo-nos com uma limitação do *PowerBI*. De facto, a ideia original era criar pontos que representassem a média por ano do Grau e ligá-los de modo a poder mais facilmente qual era a tendência da evolução da percentagem de abandono. Foi possível criar esse gráfico, contudo a posição dos pontos estava errada em relação aos eixos, mesmo com o

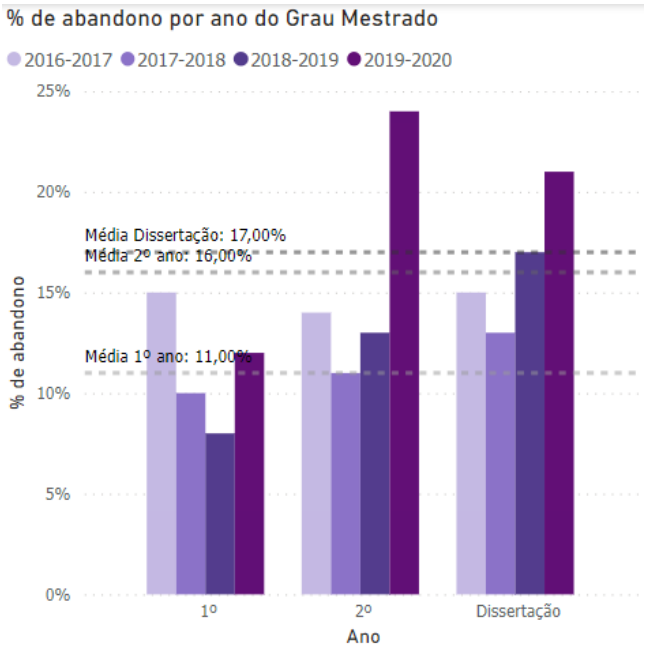


Figura 4. Percentagem de abandono por ano do Grau Mestrado, agrupados por anos letivos de 2016/17 até 2019/20. A linha a tracejado representa a média por ano do Grau.

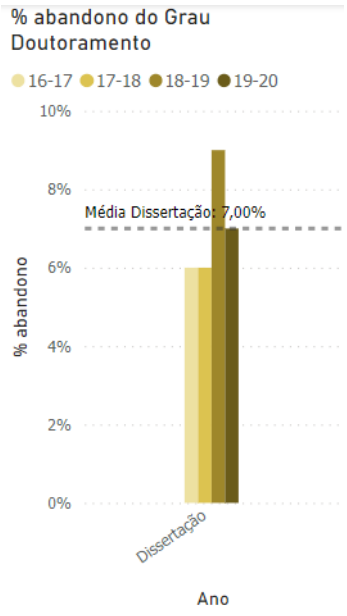


Figura 5. Percentagem de abandono por ano do Grau Doutoramento, agrupados por anos letivos de 2016/17 até 2019/20. A linha a tracejado representa a média por ano do Grau.

valor dos pontos correta. Esse problema pode ser visto na Fig.6. Tendo sido impossível corrigir esse problema, esse tipo de representação fora abandonado.

Analisando os resultados obtidos, pela Fig.3 é de notar que a percentagem de abandono vai sempre diminuindo do 1ºano até ao 3ºano. A análise para os anos posteriores é apenas feita para o Grau Mestrado Inte-

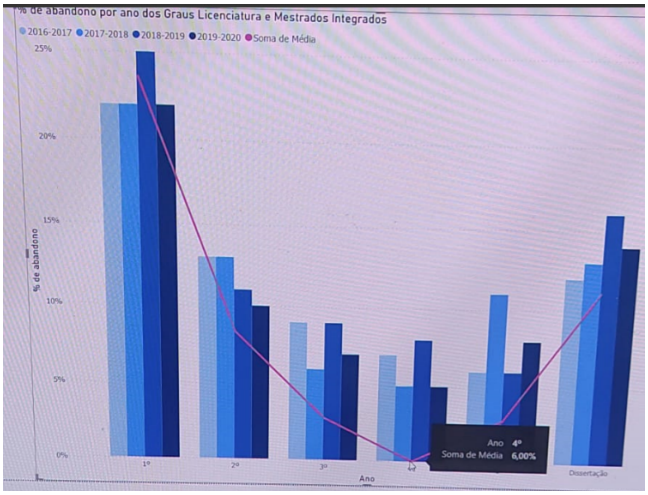


Figura 6. Percentagem de abandono por ano dos Graus Licenciatura (do 1ºano até ao 3ºano) e Mestrado Integrado (do 1ºano até ao 5ºano), agrupados por anos letivos de 2016/17 até 2019/20. A linha representa a evolução da média por ano do Grau. Insidência na posição errada do ponto assinalado com o eixo vertical.

grado. Do 3ºano para o 4ºano a percentagem de abandono diminui ligeiramente e do 4ºano para o 5ºano ela volta a aumentar para o mesmo valor que tinha para o 3ºano. É de realçar que a percentagem de abandono aquando da Dissertação tem um aumento quase para o dobro (comparando com o 5ºano) mas mesmo assim esse valçor não é tão elevado que no 1ºano. Comparando os grupos de anos letivos, não existe nenhuma correlação entre eles, por exemplo, não é pelo ano letivo 2018/19 ter tido a maior percentagem de abandono no 1ºano que foi o que teve a maior taxa de abandono para os anos posteriores comparando os anos letivos entre si. Comparando os anos sem pandemia com o ano letivo 2019/20, não parece que o aparecimento da COVID-19 tenha influenciado o abandono dos alunos nesses Graus.

Analisando os resultados obtidos para o Grau Mestrado (Fig.4), a tendência da taxa de abandono inverte-se, ou seja, onde nas Licenciaturas e Mestrados Integrados a tendência era ir diminuindo o abandono, neste grau o abandono vai aumentando ao longo dos anos. Nesta situação, podemos verificar que no ano letivo 2019/20 (aparecimento da pandemia) existe uma discrepância significativa com os anos anteriores, apenas para o 2ºano e dissertação. Nestes anos a COVID-19 pode ter sido a responsável para tal discrepância. Contudo, não temos explicação porque tal situação não aconteceu no caso da Dissertação de alunos do Grau Mestrado Integrado (Fig.3 e do Grau Doutoramento Fig.5).

Por último, na Fig.5 apenas podemos comparar a percentagem de abandono por ano pois o Doutoramento apenas é composto pela Dissertação. Apenas podemos observar que o ano com o maior abandono fora 2018/19.

Seguidamente, decidimos realizar uma comparação entre os Graus Licenciatura e Mestrado Integrado, cujo gráfico encontra-se presente na Fig.7.

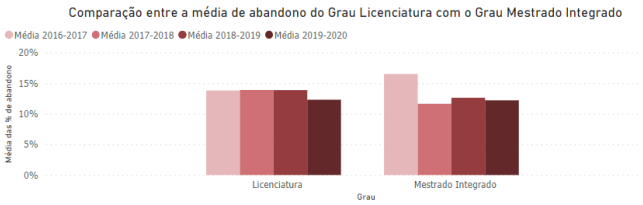


Figura 7. Percentagem de abandono dos Graus Licenciatura e Mestrado Integrado, agrupados por anos letivos de 2016/17 até 2019/20.

Realizou-se a média para o Grau Licenciatura (13,5%) e para o Grau Mestrado Integrado (13,2%) de modo a procer-se à comparação entre ambos. Com base nessas médias, conclui-se que não há um grau que se destaque relativamente ao outro. Apenas podemos referir que o ano letivo 2016/17 teve uma maior percentagem de abandono comparativamente aos outros anos somente para o Grau Mestrado Integrado.

Após analisar o abandono escolar por Grau, analisou-se o abandono escolar por Curso do ano letivo 2016/17 até o ano letivo 2019/20. Agora, decidiu-se comparar a percentagem de abandono por curso (Fig.8) com a número de inscritos por curso (Fig.9), de modo a verificar a existência ou não de uma correlação. Para a representação dos dados, escolheu-se um gráfico de barras na horizontal. Desta forma podemos comparar as percentagens de abandono e o número de inscritos entre os cursos. A comparação entre os gráficos apenas poderá ser feita em termos de *ranking* pois não estamos a trabalhar nas mesmas unidades, mas veremos que esse tipo de comparação é suficente para tirar conclusões. É de notar que esta análise apenas foi feita do 1º até ao 3º anos exluíndo assim o 4ºano e 5ºano reservados para os Mestrados Integrados. Tal escolha deveu-se ao facto de tornar mais simples a representação evitando uma análise apenas entre 3 cursos (os 3 únicos cursos de Mestrado Integrado). De modo a tornar simples a comparação entre os *rankings* dos cursos em cada gráfico, podemos usar uma *tool* do *PowerBI* em que clicando sobre um curso, destaca esse curso em toda a folha e é rápido mudar para outro curso, bastando clicar nesse outro curso.

Pela Fig.8 é de notal que em todas as situações a percentagem de abandono correspondente ao 1ºano é sempre igual ou superior aos restantes, corroborando os resultados obtidos anteriormente. Embora nem sempre aconteça, no geral a perecentagem de abandono do 2ºano é superior da do 3ºano.

Da Fig.9 retira-se que os cursos de Biologia e de Engenharia Informática são os cursos que têm o maior número de inscritos. É interessante notar que para os cur-

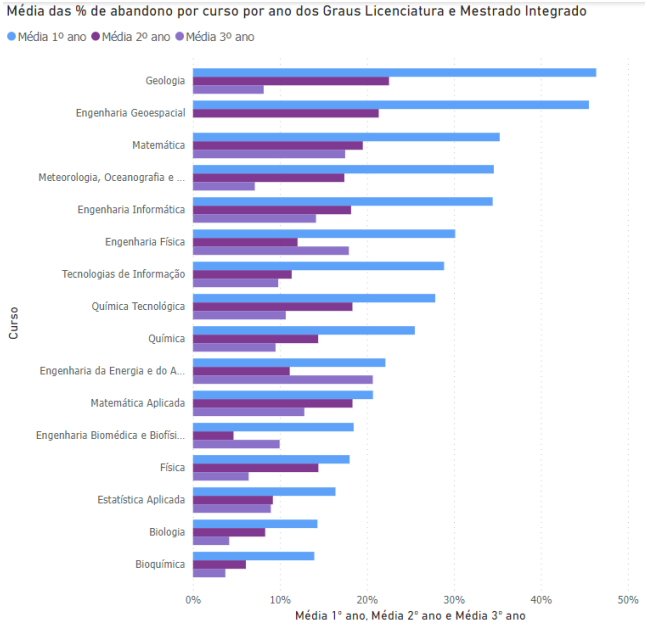


Figura 8. Percentagem de abandono por curso agrupados do 1º até ao 3º anos, cujos os valores provém da média dos anos letivos 2016/17 até 2019/20 para cada curso.

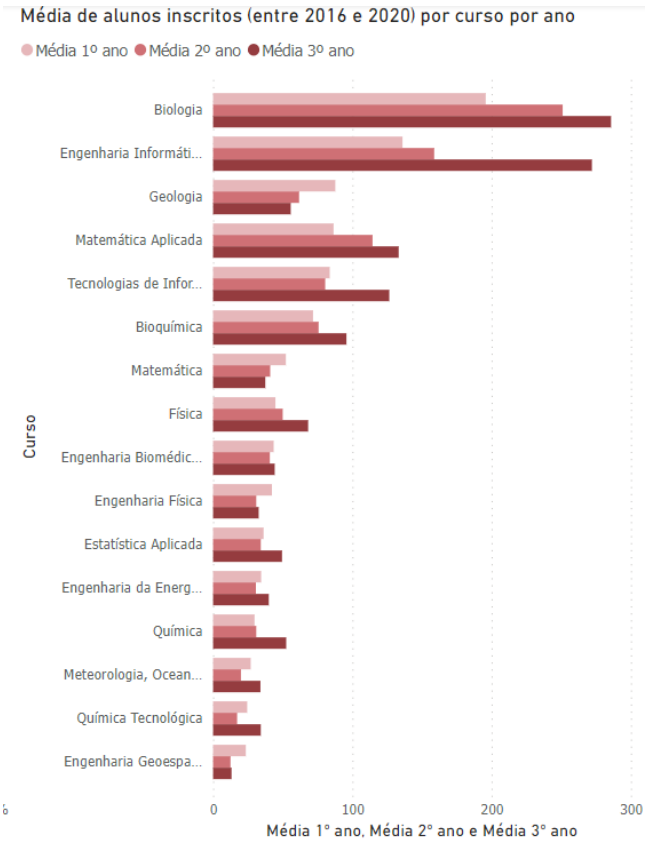


Figura 9. Valor absoluto do número de alunos inscritos por curso agrupados do 1º até ao 3º anos, cujos os valores provém da média dos anos letivos 2016/17 até 2019/20 para cada curso.

sos com maior número de inscritos (>100) a tendência é ter um maior número de alunos inscritos no ano seguinte que no ano anterior (havendo uma excessão para o curso de Geologia).

Infelizmente, não é possível fazer uso da *tool* do *PowerBI* mencionada anteriormente para selecionar os cursos neste relatório, que constituiu uma grande vantagem no uso do *PowerBI* para tratamento estatístico, mas realizando uma comparação atenta entre as Fig.8 e Fig.9, é possível notar que não existe qualquer relação entre haver um curso com um grande ou pequeno número de alunos inscritos e a sua percentagem de abandono e vice-versa.

Sucesso Escolar

Os dados de "Sucesso Escolar"referem-se às quantidades de alunos aprovados, reprovados, não avaliados e desistentes em cada UC. As UCs também estão organizadas por departamento. Os dados, vindos do site da FCUL, estão em tabelas Excel, porém, estas não estão padronizadas para os diversos anos analisados.

Nos anos de 2016/17 e 2017/18, os dados de cada semestre estavam em ficheiros diferentes, enquanto nos anos de 2018/19 e 2019/20 ambos os semestres estavam presentes num mesmo ficheiro. Os dados dos anos mais antigos também possuíam colunas ocultas que não possuíam dados, mas que apareciam quando a tabela era carregada para a *dashboard* do Power BI. Assim, de forma a criar uma base de dados robusta e uniforme, foi preciso padronizar os dados manualmente, mantendo apenas as colunas comuns a todos os anos. Foram excluídas as UCs cujo departamento estava classificado como "ULisboa", "Outro"ou "Externo", uma vez que o objetivo era comparar apenas os departamentos pertencentes à FCUL.

As diferentes UCs são identificadas de acordo com os seus códigos numéricos. Há casos de disciplinas que ocorrem em ambos os semestres, logo, os seus códigos aparecem repetidos. De forma a criar uma chave única para as UCs, nos casos em que estas ocorrem duas vezes no mesmo ano, foram adicionadas as letras "A"e "B"no final do código numérico, de forma a diferenciar um semestre do outro (respetivamente, primeiro e segundo). Este tratamento acabou por mostrar-se desnecessário, uma vez que a análise dos dados foi feita por departamento e não por UC por motivos de simplicidade (visualizar em simultâneo as mais de 800 UCs disponibilizadas pela Faculdade não seria proveitoso para os fins deste trabalho), mas consideramos que esta seja uma boa prática.

Por fim, foram criadas novas medidas com base naquelas já existentes, de forma a auxiliar a visualização

e a análise dos dados. Estas foram:

- Número de Desistentes (foram considerados como desistentes tanto os alunos que desistiram formalmente da UC quanto aqueles que não foram avaliados naquele ano letivo)
- Desistentes/Inscritos
- Reprovados/Inscritos
- Outliers - Aprovados/Inscritos
- Outliers - Desistentes/Inscritos
- Outliers - Reprovados/Inscritos

Nota: os *outliers* serão explicados mais à frente.

Com isto tudo feito nas tabelas de cada um dos anos, elas foram todas unidas numa única tabela num ficheiro Excel, que então foi carregada para o Power BI. Na *dashboard* do PowerBI, foi preciso editar o fomato dos dados, uma vez que todas as colunas foram assumidas como sendo do tipo "Texto".

Numa primeira abordagem, fez-se uma análise geral da evolução das diferentes taxas com os anos. Nesta página, utiliza-se a ferramenta "Segmentação de dados"para poder visualizar, através da interação do utilizador, um departamento ou um ano de cada vez.

O primeiro gráfico criado para obter uma visão geral dos dados foi um gráfico de linhas com a evolução temporal do valor médio de três das taxas estudadas (Aprovados/Inscritos, Desistentes/Inscritos e Reprovações/Inscritos). O resultado obtido está na Figura 10. O gráfico mostra uma melhoria geral no sucesso escolar ao fim dos quatro anos, com a taxa média de Aprovados/Inscritos aumentando, e as outras duas diminuindo. Analisando este mesmo gráfico para cada um dos departamentos, utilizando a ferramenta mencionada anteriormente, verifica-se que este comportamento também se repete individualmente para cada um dos departamentos.

O segundo gráfico utilizado para analisar as taxas de aprovação ao longo dos anos é um histograma, comparando as taxas de Aprovados/Inscritos e Aprovados/Avaliados (Figura 11). Vê-se que também a taxa de Aprovados/Avaliados cresce, passando de 90.55% em 2016/17 para 93.73% em 2019/20, totalizando um crescimento de 3.18%, enquanto a média de Aprovados/Inscritos teve um crescimento de 4.21% (de 77.55% para 81.76%)

Na segunda abordagem, mais extensa, os dados foram agrupados de acordo com os departamentos responsáveis. Foi utilizado um gráfico de barras empilhadas para visualizar as diferentes taxas para cada departamento (Figura 12). Deste gráfico, conclui-se que:

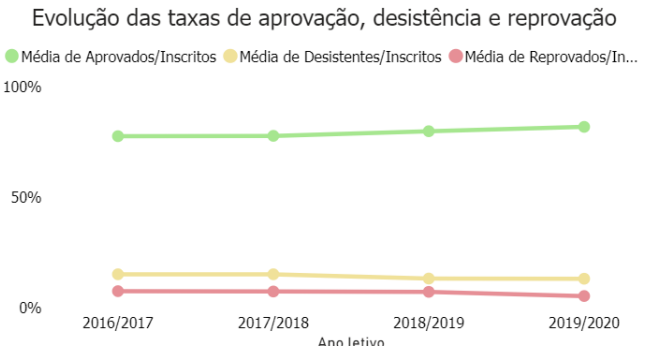


Figura 10. Gráfico de linhas com a média das taxas de Aprovados/Inscritos, Desistentes/Inscritos e Reprovados/Inscritos nos quatro anos analisados.

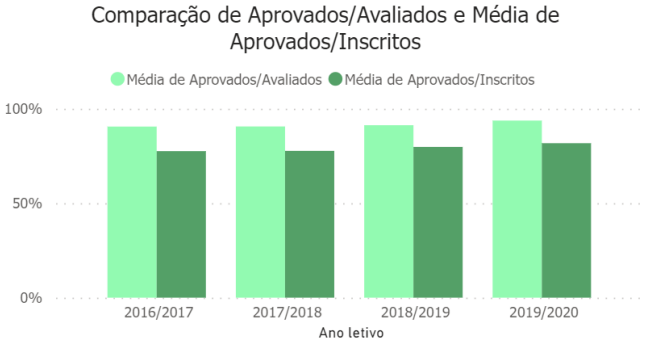


Figura 11. Histograma comparando as médias de Aprovados/Inscritos com as médias de Aprovados/Avaliados para os quatro anos em estudo.

- O DM é o departamento com maior taxa média de Reprovados/Inscritos (13.32%); maior taxa média de Desistentes/Inscritos (21.97%); e menor taxa média de Aprovados/Inscritos (64.71%).
- O DBA é o departamento com maior taxa média de Aprovações/Inscritos (90.78%) e menor taxa média de Desistentes/Inscritos (2.49%).
- O DHFC é o departamento com menor taxa média de Reprovados/Inscritos (2.49%)

Foi feita ainda uma análise das UCs consideradas *outliers* para cada uma das taxas medidas. Isto foi feito utilizando as estatísticas do *Diagrama de Caixa* (ou diagrama de extremos e quartis). Não foi encontrada no PowerBI uma ferramenta que permitisse calcular diretamente os quartis da amostra, nem representar graficamente o diagrama. Por isso, optou-se por fazer os cálculos à parte, cujos resultados estão na **Tabela I**.

Para permitir a identificação dos *outliers*, foram criadas colunas no ficheiro Excel dedicadas à classificação das UCs dentro deste quesito. Isto foi feito colocando "1"caso a UC se classificasse como *outlier*, e "0"caso contrário. Assim, foi possível utilizar o PowerBI para

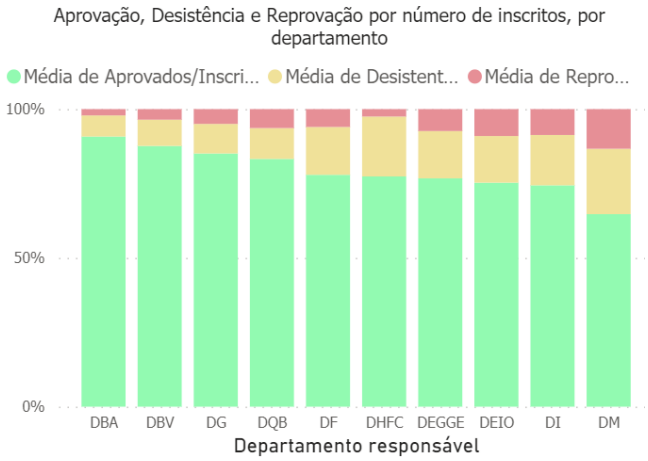
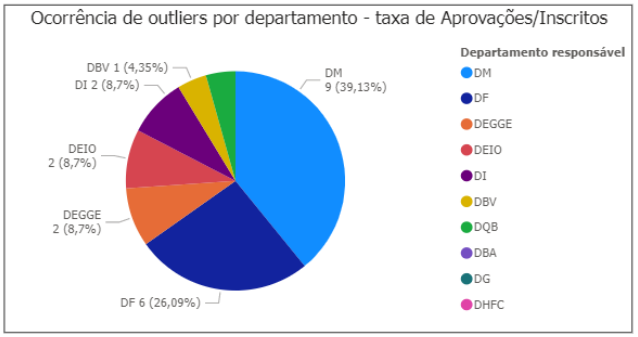
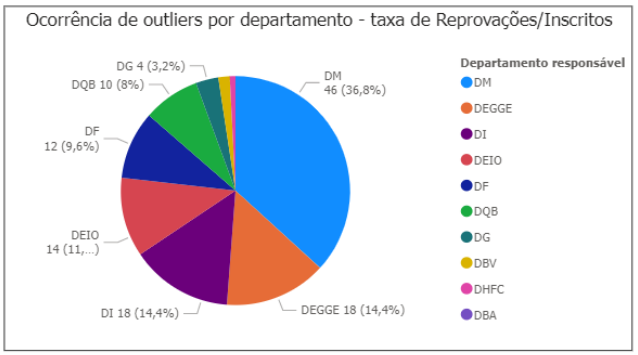


Figura 12. Gráfico de barras empilhadas representando a média dos quatro anos das taxas de Aprovados/Inscritos,Desistentes/Inscritos e Reprovados/Inscritos para cada um dos departamentos.

Taxa	Ano letivo	Critério de *outliers*
Aprovados/ Inscritos	2016/17	< 29.0 %
	2017/18	< 24.2 %
	2018/19	< 26.5 %
	2019/20	< 36.5 %
Desistentes/ Inscritos	2016/17	> 51.0 %
	2017/18	> 50.6 %
	2018/19	> 50 %
	2019/20	> 43.6 %
Reprovados/ Inscritos	2016/17	> 31.3 %
	2017/18	> 31.3 %
	2018/19	> 27.8 %
	2019/20	> 21.9 %

Tabela I. Lista do critério de classificação de *outliers* utilizando o diagrama de extremos e quartis para cada uma das taxas analisadas, para os quatro anos em estudo.





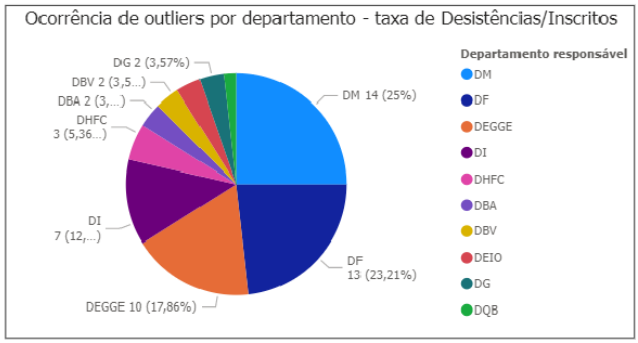


Figura 13. Gráficos de fatias com a contagem dos *outliers* em cada departamento, para cada uma das taxas.

contar quantas UCs *outliers* cada departamento possuía, utilizando o modo de resumo "Contagem".

A partir disto, são feitos três gráficos de fatias para analisar a quantidade de *outliers* que um departamento possui, ao longo dos quatro anos letivos estudados. O resultado desta análise, para cada uma das taxas, está representado na Figura 13. Nesta página também foi adicionada a ferramenta "Segmentação de Dados", para permitir a análise também ano a ano. Esta análise mostrou-se interessante, uma vez que as contagens de *outliers* por departamento variam muito ano a ano

Pode-se ver dos gráficos que, nos três casos, o DM é o que acumula mais *outliers*, seguido pelo DF na contagem dos *outliers* de Aprovações/Inscritos Reprovações/Inscritos e pelo DEGGE e pelo DI na contagem de Reprovações/Inscritos. É de se notar também que não há nenhum departamento sem *outliers*.

Abandono Escolar 1ºAno vs Alunos colocados em 1ª Opção

Na nossa última análise, quisemos ver se existia alguma correlação entre a percentagem de alunos que entrou na sua 1ª escolha na inscrição para o ensino superior com o abandono escolar desses alunos no seu 1ºano de faculdade cujos valores são as médias dos anos 2016/17 a 2019/2020.

Ambos os dados foram retirados do site da FCUL e tiveram de ser adicionados manualmente ao mesmo excel.

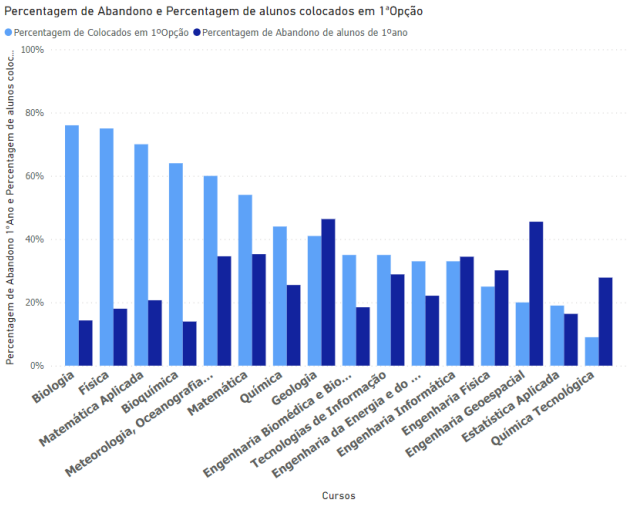


Figura 14. Percentagem de Abandono 1ºano e Percentagem de alunos colocados em 1ª opção

Depois de analisar os dados da figura 14 apesar de estarmos à espera de ver um aumento do abandono escolar com a diminuição da percentagem de alunos colocados em 1ª opção, pois eram alunos que não tinham entrado na sua primeira escolha, chegámos à conclusão que não há uma correlação forte entres as duas variáveis.

Uma possível explicação é que, mesmo não tendo escolhido o curso em que foram colocados como primeira opção, os alunos acabaram por ter sempre alguma afinidade ou satisfação com o mesmo, e por isso não se observa esta correlação.

III. CONCLUSÕES E APRECIAÇÃO CRÍTICA

Conclusões acerca dos dados estudados

Depois de analisar os três conjuntos de dados (Abandono escolar, Sucesso escolar e Colocações), podemos retirar que:

- Não há correlação entre o número de inscritos num curso e a percentagem de abandono nos mesmos (ver as Figs: 8 e 9).
- Não há correlação forte entre a percentagem de alunos inscritos que tinham o curso como primeiro lugar e a percentagem de desistências. Seria de se esperar que quanto mais pessoas que passaram em primeiro lugar, menor as taxas de desistências (isto verifica-se para alguns cursos, mas não todos - ver Fig: 14).
- As taxas de desistências dos cursos também não estão correlacionadas diretamente com o sucesso nas UCs dos respectivos departamentos. O Departamento de Matemática é o que possui pior desempenho (querendo-se dizer que é o departamento que tem as maiores taxas de desistência e reprovação, e menores taxas de aprovações), mas os seus dois cursos de licenciatura (Matemática e Matemática Aplicada) não são os que possuem as maiores taxas de desistência (ver Figs: 3, 12, 13). Assim, concluímos que as taxas de desistência das UCs de um departamento não são um bom indicativo da apreciação dos alunos dos cursos destes departamentos pelos seus cursos. Isto pode ser devido ao facto de que as UCs, apesar de pertencerem a um departamento, são atendidas por alunos de vários departamentos, que podem ter mais ou menos afinidade com o assunto estudado.
- Não temos dados suficientes para analisar os efeitos da pandemia de COVID-19 no sucesso escolar e no abandono. Porém, os dados de 2019/20 englobam o primeiro semestre de aulas em período pandémico, e não mostram uma piora nestas duas taxas (ver Figs: 3, 4, 5, 10).

Assim, para encontrar causas que levam os alunos a desistirem dos seus cursos seriam necessários mais dados sobre o perfil socioeconómico dos alunos da FCUL, bem como a sua apreciação pessoal do curso e da faculdade, ao invés de contarmos apenas com estatísticas acerca da faculdade e dos cursos em si.

Apreciação Crítica

Quanto à utilização do *PowerBI* foram encontradas algumas dificuldades e desagrados, nomeadamente:

- a falta de ferramentas para uma análise estatística mais aprofundada, por exemplo o cálculo de quartis (medida utilizada neste trabalho), tendo sido necessário recorrer ao *Excel* para o fazer;
- é preciso fazer um pré-processamento dos dados mais minucioso, para evitar possíveis confusões aquando a visualização destes.
- a facilidade/dificuldade de utilização desta ferramenta depende da padronização prévia dos dados em estudo.

No entanto, o conjunto de visualizações já disponíveis no *PowerBI* permitiu a construção rápida de visualizações intuitivas dos dados, que foram essenciais para se retirar conclusões dos dados em estudo. A possibilidade da criação de relatórios interativos no *PowerBI* verificou ser uma grande vantagem uma vez que permitiu comparar os dados consoante as várias categorias, neste caso, os cursos, os departamentos e os anos.