UNIVERSIDADE
LUSÓFONA

Relatório Engenharia de Dados

Nayara Y. Rodrigues

a22311306

O presente projeto busca realizar o trabalho de transferência de dados, conjuntamente com a armazenagem dos dados em uma Base de Dados, limpeza, validação e estudo analítico dos dados, na busca de proporcionar a elaboração completa de um pipeline básico que um Engenheiro de Dados desenvolve, conjuntamente com o estudo dos dados, trabalho este de um Cientista de Dados.

Dificuldades:

No presente projeto não encontrei grandes desafios, tanto na criação da base de dados, a sua conexão ao script, limpeza dos dados e EDA.

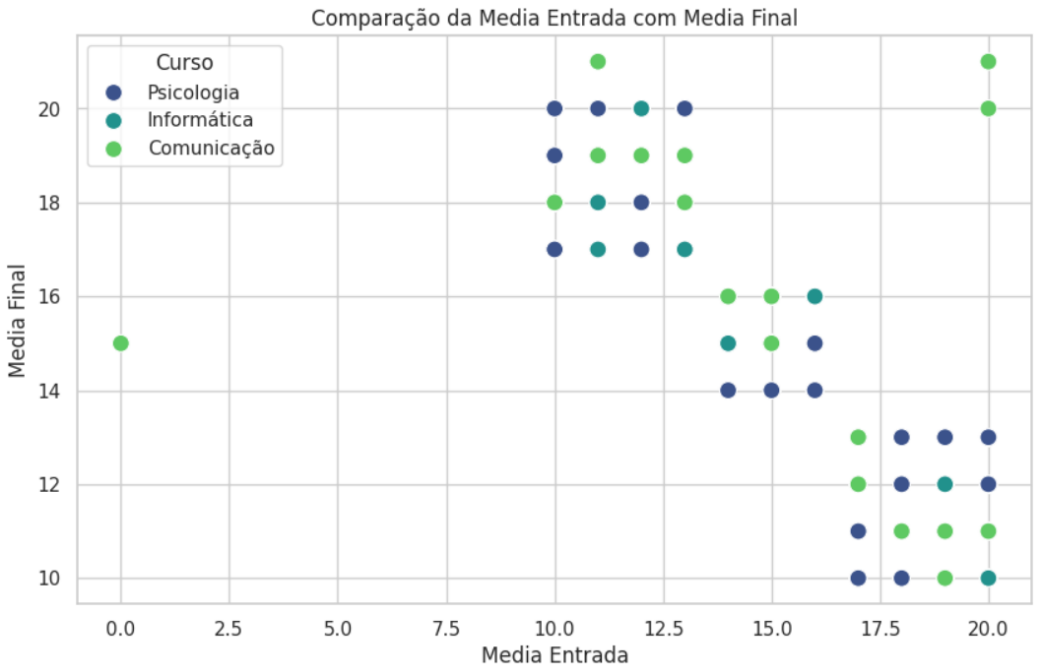
Problemas Encontrados:

Os dados possuem erros que afetam o estudo onde havia valores a NULL ou NaN, Apelidos escritos de forma errada, nome de cursos que não se encontravam na lista geral pontuação fora do âmbito normal utilizado pelo país para aceitação no curso e conclusão do mesmo. Durante a validação dos dados era possível notar a presença de NULL ou de dados não condizentes com o pedido, sendo necessário não considerar como validos.

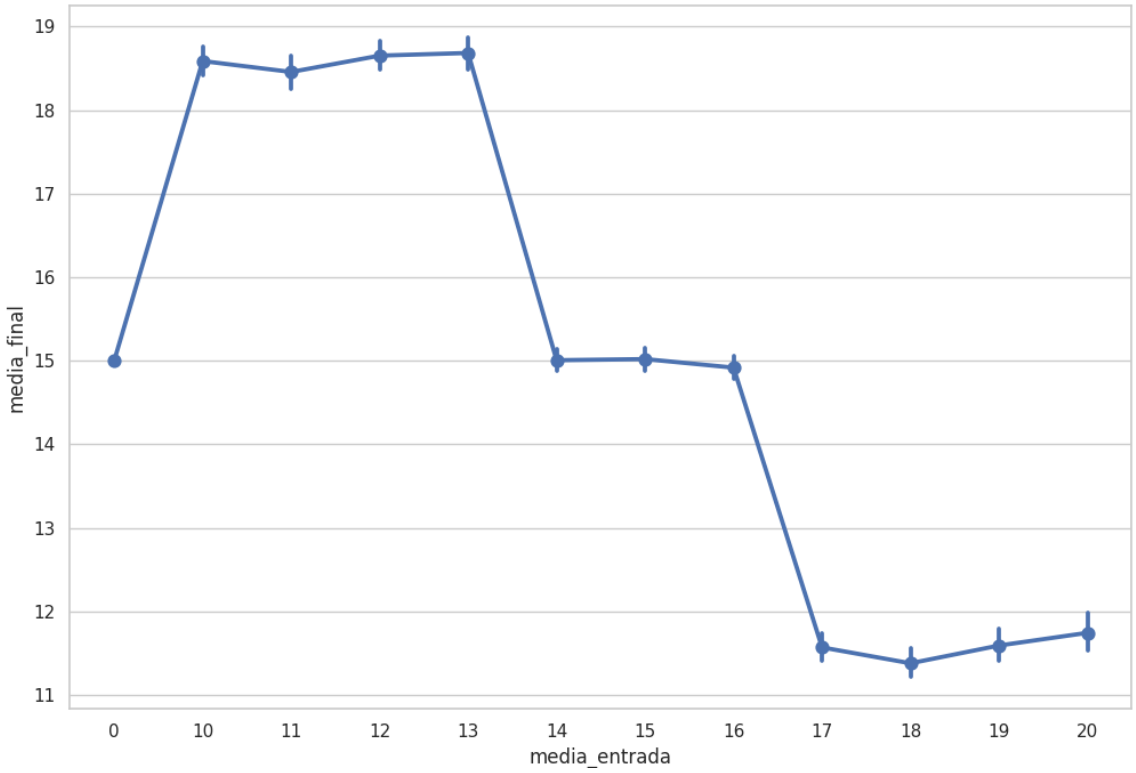
Resultados EDA:

Os resultados buscados são baseados nos seguintes questionamentos:

1. Existe alguma relação entre a média de entrada na licenciatura e a média de conclusão da licenciatura?

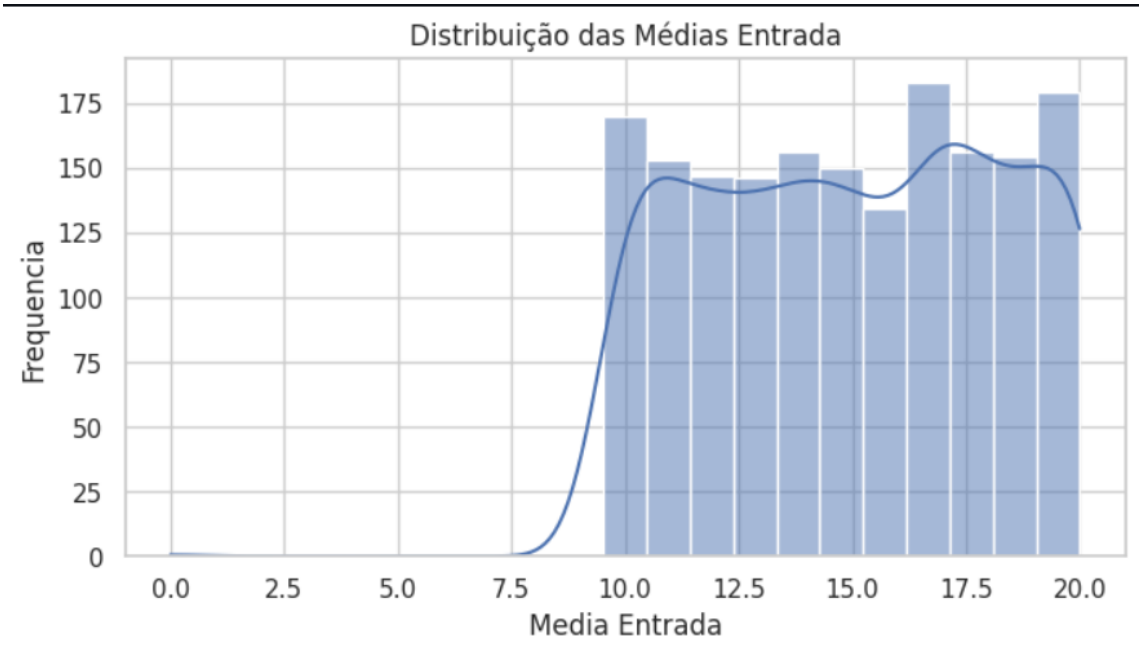


A pergunta da fase 1 questiona sobre a relação entre estudantes e suas médias de entrada e suas médias finais. Ao realizar a correlação entre ambas as variáveis é possível notar que há uma tendência geral onde alunos com médias de entrada alta não necessariamente mantem valores elevados durante o curso. Há uma tendência em manter as mesmas notas em alguns casos onde há situações que há uma variação entre 2 à 3 pontos para mais ou para menos entre o valor de ingresso e o valor final. Há a presença de outliers neste dataset, principalmente do curso de comunicação.



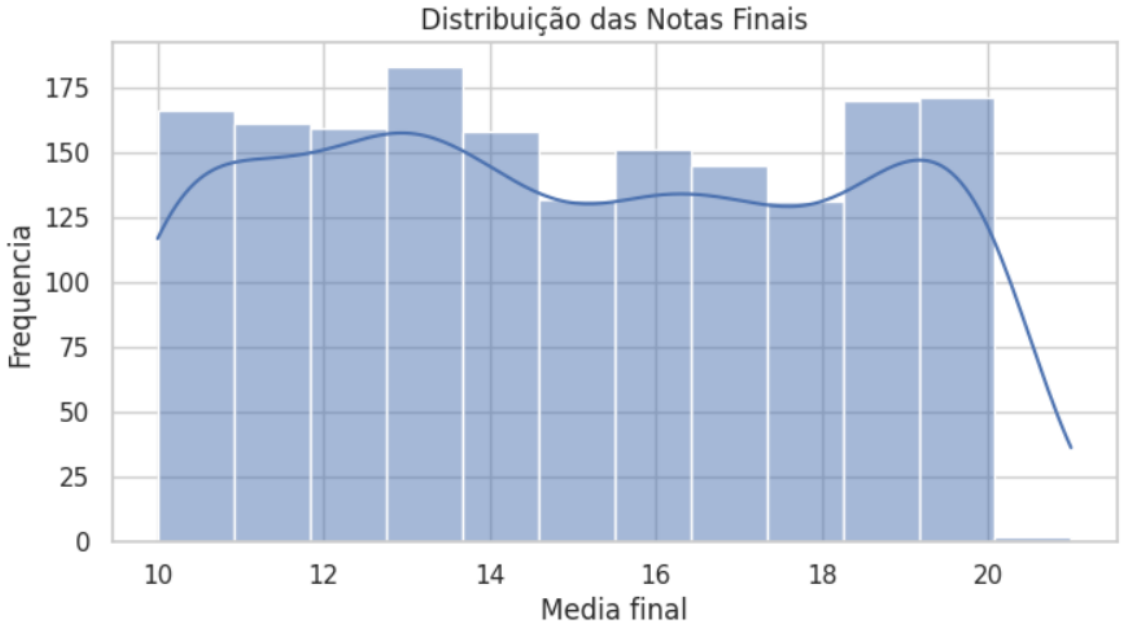
Outro gráfico para corroborar a observação anterior, há uma queda significativa dos valores das notas das médias dos alunos no final da licenciatura.

2. Distribuição das médias de entrada, final e conclusão do curso.



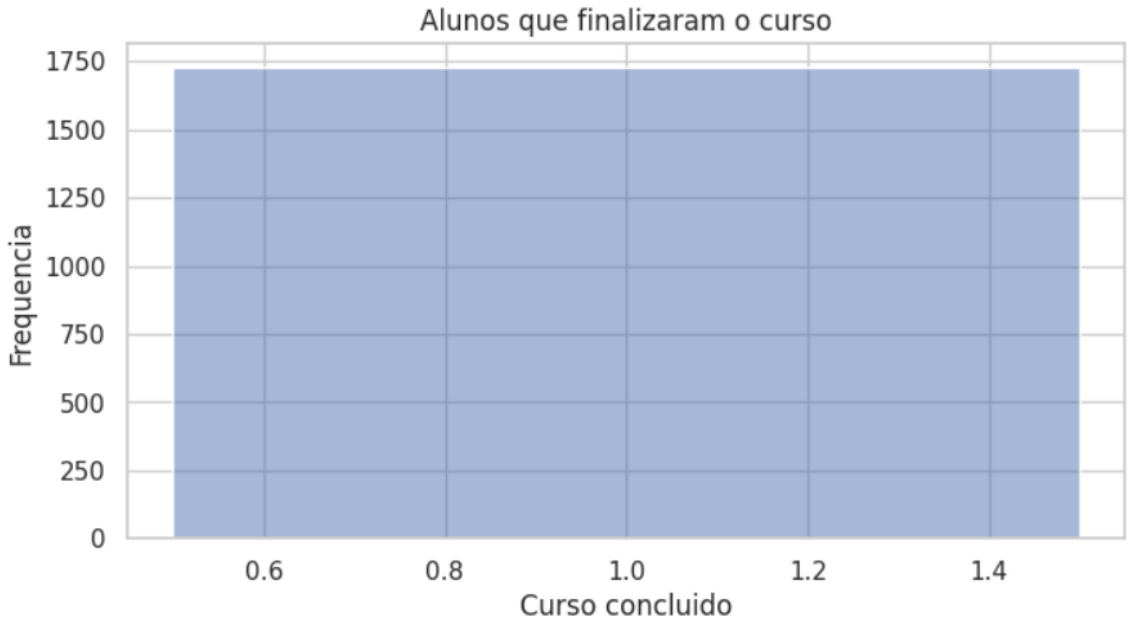
Alunos ingressão no curso com médias entre 10 e 20, tendo variações onde valores 17 são relativamente mais frequentes.

3. Distribuição médias finais



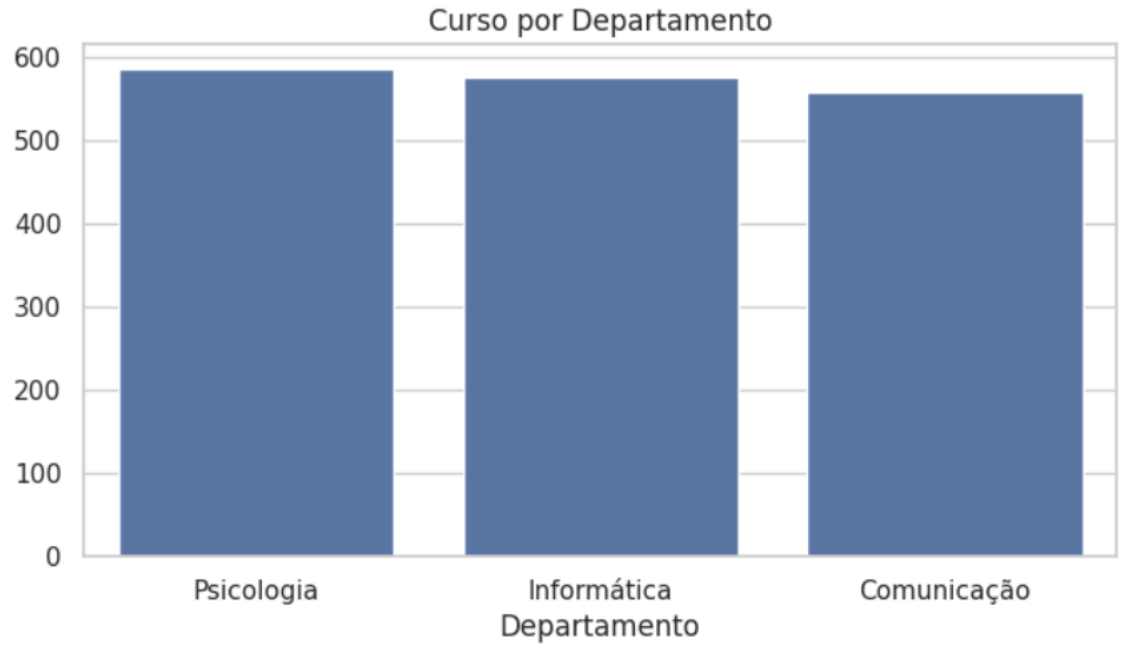
Alunos concluem o curso com médias entre 10 e 20, tendo variações onde valores 13 são relativamente mais frequentes e 19 e 20 também possuem um destaque.

4. Frequência de alunos que terminaram o curso.



Todos neste Dataset após realizar a limpeza concluíram o curso.

5. Curso por Departamento



As quantidades de dados advindos por departamento são relativamente iguais, sendo assim um dataset balanceado.

Conclusão

O presente projeto foi interessante de ser realizado, os dados eram simples, mas possuíam pontos curiosos para serem validados. Os problemas que adivinham dos dados adquiridos da base de dados invalidavam os resultados adquiridos na EDA nesse sentido era importante os excluírem e desenvolver um dataset sem falhas. Os dados oferecidos eram balanceados não existindo uma grande diferença na quantidade dos alunos por curso. Os resultados finais da EDA foram interessantes, pois não esperava uma queda qualitativa na pontuação finais dos alunos com média elevada nos cursos.