Proposta de Projeto
if1015 - Data Science
Claudio Carvalho (cco2) & Guilherme Lima (ggfl) & Henrique Caúla (lhtc)

Coleta e Integração dos Dados:
Pretendemos usar o dataset do site MyAnimeList, o site de maior relevância no contexto de animes e mangás, contendo uma robusta quantia de informações e metadados sobre estes, assim como avaliações de usuários.
O conjunto de dados em questão está disponível no site Kaggle (Link), tendo sido coletado pelo usuário Azathoth, que vem mantendo-a atualizada desde o primeiro semestre de 2018. Segundo esse usuário, seu dataset não é o primeiro com essa finalidade, mas vem como uma alternativa melhor por ser o mais completo disponível atualmente, contendo a maior parte do conteúdo disponível por meio de scraping no site em questão e tendo sido atualizado constantemente desde seu lançamento. Pela nossa análise e pesquisa em conjuntos de dados semelhantes, esse realmente aparenta ser o caso.
Dentro do conjunto que analisaremos, existem três arquivos: AnimeList.csv, UserList.csv e UserAnimeList.csv. O primeiro contém informações sobre todos os animes indexados pelo site, desde o básico como seu nome e gênero até dados mais específicos como estúdio de produção e data de veiculação. O segundo por sua vez contém dados sobre os usuários cadastrados no site desde os mais básicos como username até mais específicos como gênero, localização e data de nascimento. No terceiro, encontram-se informações sobre as avaliações dos usuário sobre os animes, registrando as notas que atribuíram às obras assim como a data da avaliação e mais.
Para ler esses dados, pretendemos utilizar as tecnologias vistas em sala de aula, com ênfase na biblioteca de Pandas para Python. Integrar os dados será simples, baseando-nos principalmente nos ID's dos usuários e animes para cruzá-los. Caso sentirmos necessidade de utilizar alguma métrica sociodemográfica para realizar análises aprofundadas sobre o perfil dos usuários, podemos utilizar alguns outros conjuntos de dados que encontramos como os disponíveis no UN Data (Link), dentre outros.

Pré-Processamento
Considerando que o conjunto de dados a ser utilizado já foi coletado por outro usuários, que já o pré-processou até certo ponto, essa tarefa para nós deve ser relativamente fácil. Após algum estudo sobre a relevância de que dados ao certo serão de interesse a serem extraídos desse dataset, podemos dropar algumas linhas com conteúdo indesejado.
Além disso, esses datasets terão que ser quebrados em múltiplos dataframes para que possam ser tratados individualmente, cada um com seu propósito.
Como aprendido em sala também pretendemos ver os outliers e analisá-los separadamente para ver se são relevantes para as hipóteses que estamos achando
Caso optarmos por usarmos dados de fontes adicionais, claro, teremos que fazer a ligação entre estes, uniformizando aspectos como datas e linguagens que podem diferir no processo.

Análise Exploratória e Hipóteses
Nossa hipótese inicial é que, partindo de um cruzamento de dados seguido de análise sobre os conjuntos com os quais estaremos trabalhando, podemos ser capazes de predizer se um novo anime

ao lançar obterá sucesso (tanto de avaliações quanto de visualizações) ou não, assim como o perfil dos espectadores que o acompanharão.
Analisaremos todos os aspectos, partindo do mais óbvio como relacionar determinados perfis de usuários com o gênero da obra, produtora ou época de lançamento. Porém, no decorrer do projeto, temos certeza que vamos encontrar variáveis inesperadas agregando valor à análise.
Pretendemos ainda testar algumas hipóteses de menor escala, partindo de alguns pressupostos comuns que o público em geral tem sobre o perfil dos espectadores de anime (por exemplo, que o gênero Yaoi é mais assistido por mulheres) e testar sua validade. Certamente, com isso, conseguiremos quebrar alguns estereótipos e possivelmente gerar um meio mais inclusivo na comunidade como um todo.
Caso achemos que necessitemos de mais informações gerais sobre regiões e vermos que isso é relevante para algumas conclusões nós podemos usar, como dito antes, os dados da UN para gerar hipóteses mais globais e com dados menos de nicho do que nas propostas iniciais.

Detecção de Padrões
Nosso foco inicialmente será de tentar relacionar certas variáveis nos data frames usando regressão linear e logística, pois possivelmente possuímos dados que são categóricos(Se certo usuário viu ou não tal anime) e também quantitativos. Depois disso por interesse nosso e por achar que essa estratégia cabe para nossas hipóteses, usaremos árvores de decisão para predizer certos eventos como explicado no tópico anterior. Caso observemos que seria útil outra das estratégias, usaremos também para explanar e/ou explicar mais hipóteses.

Análise dos Resultados
A estória dos dados será contada por meio de gráficos interativos que mostrem um bom conjunto exemplo do tipo de correlações que conseguimos fazer para ligar os dados que tínhamos inicialmente e os resultados encontrados a partir destes. Esses exemplos, claro, devem ser bem escolhidos para ser realmente representativos do processo em geral pelo qual passamos, e não apenas do que fora encontrado ao final.
Além disso, claro, a atração principal será uma ferramenta para que possamos mostrar a utilidade de nossa análise, sendo capaz de fazer a predição de sucesso para animes futuros baseado nos padrões encontrados até então. Essa ferramenta poderá ser testada utilizando exemplos reais de animes que lançaram dentre o tempo da última atualização dos datasets e o momento atual, checando se a predição será feita corretamente ou não.