**Data Science**

**Curso I – Primeiros Passos:**

Link da documentação das bibliotecas que utilizaremos nesse curso:

- Pandas: <https://pandas.pydata.org/pandas-docs/stable/>;

- Seaborn: <http://seaborn.pydata.org/introduction.html>.

Vou colocar todo arquivo escrito na pasta da aula, mas aqui está o link do drive, de qualquer forma: <https://drive.google.com/drive/folders/1YV0fPKe-_f5beekL2cBUVSOsZ_2G6uaS>.

1. **Aula 1 – Data Science – Dados e Visualizações:**
   1. Podemos sim alterar o nome das colunas passando uma lista com os novos nomes a partir do atributo colunas do nosso df:

notas.columns = ['usuarioId', 'filmeId', 'nota', 'momento']

* 1. O pandas possui visualização e construção de dados a partir do método plot().
     1. O plot recebe o parâmetro kind=’’, onde podemos escolher o tipo de gráfico, como um hist de histograma, por exemplo. Se colocar somente plot() ele cria um plot normal onde não vemos nada:

notas.nota.plot(kind='hist')



notas.nota.plot()



* + - 1. O hist é um tipo de gráfico que conta uma história, mostrando no eixo x as notas que um filme recebeu e no y a quantidade de notas. Basicamente o inverso do plot.
    1. S.median(): retorna a mediana de uma series, mostrando exatamente qual o valor que dividi um conjunto de dados no meio:

notas.nota.median()



* + 1. Importamos a seaborn e fizemos um boxplot para visualizar melhor os gráficos:

import seaborn as sns

sns.boxplot(notas.nota)



* + - 1. A mediana é o traço no meio do box. Quartil 25 é o final esquerdo do box e 75 o final direito. As linhas com traço na ponta dos lados esquerdo e direito são, respectivamente os limites inferior e superior, ou seja, da mediana para as laterais do box temos 25% cada lado, os li e ls são outros 25%, somando 50% no box e 50% fora.
      2. Com esse gráfico podemos concluir que metade das pessoas, ou seja, o box inteiro, deram notas entre 3 e 4.
  1. O que aprendemos:
     1. Importar pandas que é uma biblioteca de análise de dados;
     2. Ler dados CSV;
     3. Fazer upload de arquivo para o Google Colab;
     4. Renomear colunas;
     5. Contar dados;
     6. Melhorar a visualização das informações.

1. **Aula 2 – Análise Exploratória:**
   1. Podemos usar queries para procurar coisas específicas nos nossos dados e selecionar a coluna que queremos, bem como a análise estatística depois dela:

notas.query('filmeId == 1').nota.mean()



* + 1. Dissemos que queremos a média das notas de todo filme que tenha o id == 1 dentro do nosso df de notas.
    2. Se quisermos tirar a média de todas as notas para cada id de filme teria que fazer isso infinitamente, mas, graças ao pandas, podemos agrupar todos os filmes pelo seu id com o df.groupby(‘coluna’) e tirar a média:

notas.groupby('filmeId').mean()



* + - 1. Contudo queremos só a coluna de notas, então podemos selecionar somente ela. O retorno será uma series. Para transformar em df basta colocar um to\_frame(). Como já vimos em outros cursos, podemos colocar tudo numa mesma linha para facilitar, ficando assim:

notas.groupby('filmeId').nota.mean().to\_frame()



* + 1. Após atribuir a uma variável, podemos fazer um histograma para ver a distribuição da média das notas:

medias\_por\_filme = notas.groupby('filmeId').nota.mean().to\_frame()

medias\_por\_filme.plot(kind='hist')



* + 1. Ou então pedir um sns.boxplot(medias\_por\_filme), porém, aprendi que ele não aceita frames, apenas series, nesse caso removi o s.to\_frame() da linha acima, rodei novamente e ele aceitou fazer o gráfico:

medias\_por\_filme = notas.groupby('filmeId').nota.mean()

sns.boxplot(medias\_por\_filme)



* + - 1. Geralmente boxplots são exibidos na vertical e não horizontal. Para alterar isso basta passar nosso s como y:

sns.boxplot(y=medias\_por\_filme)



* + 1. Para fazer um histograma no seaborn a nomenclatura muda, se tornando sns.distplot(), ou seja, gráfico de distribuição. Sua visualização é diferente também:

sns.distplot(medias\_por\_filme)



* + - 1. Podemos notar que esse histograma é diferente do que o pandas fez, mas isso ocorre de acordo com a distribuição de barras que cada um determina fazer.
      2. O pandas fez 10 separações/caixinhas/barras para distribuir as notas, já o seaborn fez MUITO mais, por isso a exibição fica diferente, mas isso não altera o resultado do gráfico, uma vez que os valores são um só.
      3. Entretanto, se quisermos podemos escolher quantas caixinhas o seaborn vai exibir, isto é, quantos bins=n queremos que ele faça para exibir o gráfico, passando esse como segundo parâmetro/argumento na hora de fazer o plot:

sns.distplot(medias\_por\_filme, bins=10)



* + - 1. Note que ao colocar 10 a distribuição fica exatamente igual ao do pandas.
      2. Precisamos tomar cuidado pois dependendo do número que escolhermos podemos passar diferentes mensagens com nosso gráfico, desse modo podendo passar uma mensagem errônea para frente ou até mesmo uma mentira. Precisamos ter cuidado com isso!
      3. Na prática o próprio seaborn já faz uns cálculos para determinar qual seria a quantidade de bins necessárias para cada gráfico.
      4. Além do hist. Ele também faz uma curva para mostrar mais suavemente como ficam essas notas de uma outra forma.
    1. Tanto o pandas quanto o seaborn utilizam uma biblioteca bem baixo nível para produzir esses gráficos. A biblioteca que ambos usam é a matplotlib.pyplot, sendo importada como plt. Biblioteca essa que também possui seu próprio histograma:

import matplotlib.pyplot as plt

plt.hist(medias\_por\_filme)



* + - 1. Essa lib possui diversos métodos, parâmetros que podem ser passados para alterar e modificar nossos gráficos de diversos modos que não vamos ver nesse curso, mas sim no data visualization mais pra frente.
    1. Voltando um pouco nas anotações vimos que o boxplot do seaborn ficou achatado quando mudamos para a horizontal. Para alterar isso podemos utilizar a biblioteca plt, uma vez que ela é quem está sendo rodada por baixo dos panos pelas outras libs que estamos utilizando. Para configurar o tamanho da imagem do gráfico utilizamos o plt.figure(figsize=(x,y)), podendo ser totalmente personalizado:

plt.figure(figsize=(5,8))

sns.boxplot(y=medias\_por\_filme)



* 1. Histograma e boxplot:
     1. O histograma também é conhecido como "distribuição de frequências", e é representado em um gráfico de colunas ou de barras a partir de um conjunto de dados previamente tabulado e dividido em classes (uniformes ou não uniformes);
     2. Já o boxplot, também conhecido como "diagrama de caixa", é um método alternativo para o histograma. Uma de suas particularidades é mostrar a faixa em que está concentrada a média de valores.
     3. Referências:
        1. <https://pt.wikipedia.org/wiki/Diagrama_de_caixa>;
        2. <https://pt.wikipedia.org/wiki/Histograma>.
  2. O que aprendemos:
     1. Trabalhar com query;
     2. O que são bins;
     3. Filtrar apenas uma coluna.

1. **Aula 3 – Variáveis:**
   1. Site para baixar arquivos com 5000 databases de filmes chamado the movie database (TMDB 5000): <https://www.kaggle.com/tmdb/tmdb-movie-metadata>.
      1. Virão 2 arquivos: um com os créditos, ou seja, as pessoas que participaram do filme e as informações sobre o filme em si.
   2. Categórica Nominal: variável de nome responsável por categorizar dados, como o caso de língua original de um filme. É um tipo de variável que não possui grandeza ou ordem, como as numéricas.
   3. Categórica Ordinal: ou seja, com ordem. Temos como exemplo a escolaridade, onde temos primeiro, segundo e terceiro grau por exemplo.
      1. Não possuí soma, ou seja, o primeiro mais o segundo não é igual ao 3°, por exemplo. No caso da escolaridade não é a mesma distância que o 1° grau tem do 2° que tem do 3°. É diferente das numéricas onde o 1 + 2 = 3.
      2. Não tem uma facilidade de realizar operações aritméticas.
   4. Quantitativo: variáveis de quantidade, como o nome diz. Temos como exemplo o orçamento/budget de filmes.
      1. Pode ser inclusive contínuo, ou seja, pode 0, pode ser 500 e pode ter infinitos valores no meio de ambos, como por exemplo, ter os números 15; 15,2; 15,37.... e assim por diante infinitamente.
   5. Quantitativo Intervalar: não possuí infinitos números decimais, por exemplo. É sempre determinado que vai de 1 em 1 ou de .5 em .5, como no caso de votos e notas de filmes, respectivamente.
   6. Para saber mais ordinais e nominais:
      1. As variáveis categóricas são aquelas que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.
      2. Quando estamos trabalhando com categorias ordinais, existe uma ordenação entre as categorias. Por exemplo: escolaridade (primeiro, segundo e terceiro graus), estágios de uma doença (inicial, intermediário, terminal), mês de observação (janeiro, fevereiro, ..., dezembro), etc.
      3. Referência: <http://leg.ufpr.br/~silvia/CE055/node8.html>.
   7. O que aprendemos:
      1. Importar um CSV;
      2. Identificar o tipo de uma variável observando seu conteúdo;
      3. O que é categórica ordinal;
      4. O que é quantitativa.
2. **Aula 4 – Data Visualization:**
   1. Para visualizar categorias ordinais podemos utilizar o value\_counts() para ver quantas vezes uma língua aparece em um df de filmes, por exemplo, uma vez que o value\_counts() conta tudo, inclusive str’s:

tmdb['original\_language'].value\_counts()



* + 1. Para remover o index de uma series após converter ela para frame, usamos o método reset\_index():

tmdb['original\_language'].value\_counts().to\_frame().reset\_index()



* + - 1. Desse modo possuímos 2 colunas ao invés de 1 índice e 1 colunas, transformando assim uma série em um df.
      2. Note que os nomes das colunas estão bagunçados, podemos concertar facilmente como já sabemos:

contagem\_de\_lingua = tmdb['original\_language'].value\_counts().to\_frame().reset\_index()

contagem\_de\_lingua.columns = ['original\_language', 'total']

contagem\_de\_lingua.head()



* + 1. O seaborn possuí diversas formas de plotar categoricals. Podemos simplesmente procurar na documentação.
       1. Vamos usar o barplot(x, y, df):

sns.barplot(x = 'original\_language', y = 'total', data = contagem\_de\_lingua)



* + 1. Tivemos muito trabalho para chegar nesse gráfico, mas caso queiramos podemos utilizar o sns.catplot(x, kind=’’, data=df), que ao invés de fazer todo esse tratamento, podemos só pegar no df com todas as informações e pedir para que o seaborn faça um plot por categoria, passando quais são as categorias no x e dizendo qual o tipo de plot que queremos, nesse caso passando o ‘count’ como parâmetro para que ele conte quantas vezes cada categoria apareceu e depois fazendo as barras. No data passamos o df com todas as informações que importamos lá no começo.
       1. Podemos receber um erro se tentar rodar logo de cara dizendo que esse método não existe, porque, apesar dele existir, está apenas nas versões 0.9+ do seaborn e, ao verificar qual estamos rodando no colab com um print(), veremos que estamos rodando a 0.7. No meu caso ele está em uma versão recente:

print(sns.\_\_version\_\_)



* + - 1. Do mesmo modo que quando utilizando uma IDE local precisaríamos instalar a biblioteca, precisaremos instalar no notebook do colab, uma vez que ele está rodando em uma máquina virtual cloud e provavelmente não tem as versões mais recentes de todas as bibliotecas.
      2. Para instalar a versão desejada da lib, fazemos do mesmo modo que localmente, usando o !pip install lib. Se quiser ainda pode passar qual a versão da lib que quer instalar, fazendo a instalação ao rodar a célula:

!pip install seaborn=0.9.0

* + - 1. Para saber qual versão aquele método/função está disponível é bem fácil: CONSULTE A DOCUMENTAÇÃO!!!
      2. Resultado com o sns.catplot():

sns.catplot(x = 'original\_language', kind='count', data = tmdb)



* + - 1. Equivalente ao mesmo que fazendo toda a seleção de dados.
      2. A diferença é que ele não fica ordenado de acordo com a aparição como fizemos na seleção. Apenas demos sorte de en ter ficado em primeiro, mas ao ver os outros menores, podemos notar não está do maior para o menor.
      3. Para oficializar a versão, pós instalação, precisamos resetar o runtime.
  1. Gráfico de pizza quase nunca é recomendado em nenhuma situação.
     1. Vimos que no gráfico acima quase não conseguimos ver a diferença de cada língua, então decidimos fazer um gráfico de pizza/pie(en) para ver se melhora a visualização.
     2. O seaborn não possui esse tipo de gráfico, mas o matplotlib sim, então utilizamos ele passando 2 parâmetros, os valores e as labels no plt.pie(x, labels), e ele nos retorna esse resultado horrendo:

plt.pie(contagem\_de\_lingua['total'], labels = contagem\_de\_lingua['original\_language'])

plt.show()



* + 1. Por isso que quase nunca é recomendada sua utilização, pois fica visualmente horrível e muito mais confuso e pior do que estaria em qualquer outro.
    2. Considerando isso é bem importante sabermos qual é a história que queremos contar com nosso gráfico. Nesse caso nós queremos mostrar o quanto en aparece em relação aos outros idiomas, nesse caso, podemos separar os dados em 2 variáveis, uma com o total de inglês e outra com o total do resto. Para fazer isso podemos fazer uma contagem de valor da coluna de original\_language no nosso df com todos os dados lá do início. Após fazer essa atribuição, podemos criar outras 2 variáveis, a total geral, que mostra a soma de todos os idiomas, e a total de inglês, onde usamos o s.loc[‘en’] para pegar somente a contagem de inglês do nosso total por língua. Para saber qual o total do resto de línguas basta subtrair o total de inglês com o total geral, obtendo assim o total do resto:

total\_por\_lingua = tmdb['original\_language'].value\_counts()

total\_geral = sum(total\_por\_lingua)

total\_de\_ingles = total\_por\_lingua.loc['en']

total\_do\_resto = total\_geral - total\_de\_ingles

print(total\_de\_ingles, total\_do\_resto)



* + 1. A partir disso podemos criar um dict chamado dados que aloca duas chaves com 2 valores cada: língua, onde recebe o valor de ‘inglês’ e ‘outros’, e total, que recebe as variáveis de total de inglês e total do resto, convertendo esse dict em um df no final para trabalharmos com ele para fazer o gráfico:

dados = {

    'lingua': ['ingles', 'outros'],

    'total': [total\_de\_ingles, total\_do\_resto]

}

dados = pd.DataFrame(dados)

dados



* + 1. Decidimos fazer um gráfico de barras com sns para melhor visualização, passando como parâmetros sns.barplot(x=’língua’, y=’total’, data = dados), onde o x é a label, y são os valores e data é nosso df recém criado:

sns.barplot(x = 'lingua', y = 'total', data = dados)



* 1. Criamos uma variável que seleciona todas as línguas menos o inglês e criamos um sns.catplot() com ele para ver a importância relativa das outras línguas, uma vez que quando estava com o inglês incluso não conseguíamos ver praticamente nada além dele:

filmes\_sem\_lingua\_original\_em\_ingles = tmdb.query("original\_language != 'en'")

sns.catplot(x = 'original\_language', kind = 'count', data = filmes\_sem\_lingua\_original\_em\_ingles)



* 1. O que aprendemos:
     1. Comparar categorias;
     2. O que é uma Series;
     3. Instalar o seaborn;
     4. Transformar em dataframe com a função to\_frame();
     5. Remover o índice para gerar duas colunas usando a função reset\_index();
     6. Utilizar o Categorical do seaborn.

1. **Aula 5 – Continuando Com Visualização:**
   1. Notamos que na figura acima ficou tudo amontoado no eixo x e, diferentemente do sns.barplot() que nos permite personalizar como quisermos usando a plt por ser baixo nível, o catplot(), como é alto nível, o próprio sns que configura pra gente.
      1. Por conta disso, não conseguimos alterar o tamanho da figura como antes, precisando assim ir consultar a documentação do catplot e descobrir que existe um parâmetro chamado height e outro aspect, estando configurados como 5 e 1 respectivamente por default, ou seja, quadrado.
      2. Para alterar isso e expandir o eixo x do nosso gráfico, precisamos alterar o aspec. Colocando 2, ou seja, o dobro, já deixa nosso gráfico com a visualização bem melhor:

sns.catplot(x = 'original\_language', kind = 'count', data = filmes\_sem\_lingua\_original\_em\_ingles, aspect=2)



* + 1. Contudo notamos que ainda está fora de ordem, ao contrário do que estamos acostumados a ver, sendo ela do mais popular/visto/que apareceu para o que menos, uma vez que são variáveis categóricas nominais e não existe ordem de sequência.
    2. Entretanto, ao consultar a documentação, vimos que existe o parâmetro order, que define a ordem que nosso gráfico será exibido. Com isso podemos passar para ela o index da contagem de quantas vezes cada língua apareceu no nosso df e ele faz na ordem de frequência:

total\_por\_lingua\_de\_outros\_filmes = tmdb.query("original\_language != 'en'").original\_language.value\_counts()

sns.catplot(x = 'original\_language', kind = 'count', data = filmes\_sem\_lingua\_original\_em\_ingles, aspect=2, order=total\_por\_lingua\_de\_outros\_filmes.index)



* + 1. A cor ainda não está no ideal, então precisamos consultar a documentação de palette de cores do seaborn. Descobrimos que possuem palettes para cada tipo de gráfico, seja quantitativo, sequencial ou qual for.
    2. No nosso caso queremos sequencial, que ele vá do mais escuro para os mais frequentes até os mais claros para os menos.
    3. “Mako”, essa é a string da palette de cores que precisamos atribuir ao parâmetro palette=str do nosso gráfico:

sns.catplot(x = 'original\_language', kind = 'count', data = filmes\_sem\_lingua\_original\_em\_ingles, aspect=2, palette='mako', order=total\_por\_lingua\_de\_outros\_filmes.index)



* 1. Para saber mais: seaborn:
     1. Com o Seaborn, somos capazes de criar vários tipos de gráficos ou visualizações. Ao longo deste curso, aprendemos a fazer os gráficos de barras (histogramas) e diagrama de caixa (boxplot).
     2. Outro exemplo de gráfico que pode ser criado com essa biblioteca é o gráfico linear:

import seaborn as sns

sns.set(style="ticks")

# Load the example dataset for Anscombe's quartet

df = sns.load\_dataset("anscombe")

# Show the results of a linear regression within each dataset

sns.lmplot(x="x", y="y", col="dataset", hue="dataset", data=df,

col\_wrap=2, ci=None, palette="muted", height=4,

scatter\_kws={"s": 50, "alpha": 1})



* + 1. O dataframe utilizado para essa plotagem já está dentro da biblioteca do Seaborn!
    2. Outros tipos de gráfico com o seaborn: <http://seaborn.pydata.org/examples/>.
  1. O que aprendemos:
     1. Reescalar o gráfico;
     2. Ordenar o gráfico;
     3. Alterar os tons do gráfico.

1. **Aula 6 – Dados e Estatística:**
   1. Podemos juntar 2 arrays np com o np.append():

np.append(np.array([2.5] \* 10), np.array([3.5] \* 10))

* 1. Precisamos saber como utilizer e analisar os dados fazendo testes estatísticos, não podemos simplesmente aceitar usar a média ou mediana de um dado sem saber sua distribuição, como por exemplo 20 notas para 2 filmes diferentes, enquanto 1 tem 10 notas 2.5 e 10 notas 3.5 tendo média/ mediana 3, o 2 tem 10 notas 1 e 10 notas 5, também tendo média/mediana 3, ou seja, o mesmo resultado final para distribuições totalmente diferentes.
     1. Para melhorar isso utilizamos os gráficos, mas mesmo assim precisamos saber qual gráfico utilizar, uma vez que dependendo dele os dados podem ficar sobrepostos ou com uma visualização muito ruim. Por isso é importante conhecer várias bibliotecas e tipos de visualização de dados diferentes, dessa forma sempre sabendo qual a melhor forma de expressar um dado:

import numpy as np

filme1 = np.append(np.array([2.5] \* 10), np.array([3.5] \* 10))

filme2 = np.append(np.array([5] \* 10), np.array([1] \* 10))

print(np.mean(filme1), np.mean(filme2))

print(np.median(filme1), np.median(filme2))



sns.distplot(filme1)

sns.distplot(filme2)



plt.hist(filme1)

plt.hist(filme2)



sns.boxplot(filme1)

sns.boxplot(filme2)



plt.boxplot([filme1, filme2])



* + 1. O seaborn é muito bom para fazer gráficos quando ainda não temos os dados separados, servindo para a parte de exploração de dados.
    2. No caso do nosso df das notas, podemos pedir um sns.boxplot(x =’filmeId’, y = ‘nota’, data = notas), para que ele faça um boxplot de todas as notas para cada filme, mas tem o contra de ficar SUPER poluído:

sns.boxplot(x = 'filmeId', y = 'nota', data = notas)



* + - 1. Podemos filtrar esse gráfico com um query:

sns.boxplot(x = 'filmeId', y = 'nota', data = notas.query('filmeId == [1,2,3,4,5]'))



* + 1. Se quisermos saber o quanto os quartis desviam do padrão, utilizamos o: desvio padrão, ou std():

print(np.mean(filme1), np.mean(filme2))

print(np.std(filme1), np.std(filme2))

print(np.median(filme1), np.median(filme2))



* + 1. Agora temos 2 informações, a média de cada filme e a dispersão das notas, ou seja, o quanto elas variam.
  1. Para saber mais: Desvio padrão:
     1. Em estatística, dispersão de dados pode ser chamada de variabilidade ou espalhamento e são dados estatísticos para mostrar uma determinada variação de dados, e com esses dados é possível obter a variância, o desvio padrão e a amplitude interquartil.
     2. Mas vamos demonstrar um exemplo de desvio padrão em uma sala de aula pensando na quantidade de aulas assistidas e de alunos que vão a aula. Teremos o seguinte bloco:



* + 1. Temos o desvio padrão quando temos 2 alunos com a quantidade 4 de aulas assistidas! E nas outras quantidades de aulas assistidas temos apenas 1 aluno participando, sendo assim temos um desvio quando temos 2 alunos com a quantidade de 4 aulas assistidas, porque ele quebra o padrão de ter um aluno por aula.
    2. Referência: <https://pt.wikipedia.org/wiki/Dispers%C3%A3o_estat%C3%ADstica>.