OS APPs MAIS RENTÁVEIS NA APP STORE E GOOGLE PLAY

PROJETO PARA PORTFOLIO EM CIÊNCIA DE DADOS

Supondo que trabalho em uma empresa que desenvolve apps gratuitos que lucram com propagandas, meu objetivo é analisar dados para ajudar os desenvolvedores a compreender qual o tipo de aplicativo atrai mais usuários.

BANCO DE DADOS

para efeitos de estudo seria muito custoso analisar mais de 4milhões de apps, sendo assim, optei por trabalhar com dois bancos de dados menores:

- 10.000 aplicações Android na Google Play (dados coletados em 08/2018); https://www.kaggle.com/lava18/google-play-store-apps (https://www.kaggle.com/lava18/google-play-store-apps)
- 7.000 aplicações iOS na App Store (dados coletados em 07/2017); https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps (https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps)

PASSO A PASSO

1. **abrir os dados** em uma lista de listas separando o cabeçalho das demais linhas:

In [1]:
```
from csv import reader

#google play
aberto = open('googleplaystore.csv')
ler = reader(aberto)
android = list(ler)
android_header = android[0]
android_corpo = android[1:]

#app store
aberto  = open('AppleStore.csv')
ler = reader(aberto)
ios = list(ler)
ios_header = ios[0]
ios_corpo = ios[1:]
```

2. crio uma função para **explorar os dados** possibilitando a impressão de linhas e colunas específicas e para descobrir o número de linhas e colunas de cada uma:

In [2]:
```
def explora_dados(dados, inicio, fim, linhas_colunas = False):
    dados_corte = dados[inicio:fim]
    for linha in dados_corte:
        print(linha)
        print('\n')

    if linhas_colunas:
        print('num de linhas:', len(dados))
        print('num de colunas', len(dados[0]))
```

In [3]:
```
teste = explora_dados(android_corpo, 2, 5, True)
print(teste)
print('\n')
print(android_header)
```

```
['U Launcher Lite – FREE Live Cool Themes, Hide Apps', 'ART_AND_DESIGN', '4.7',
'87510', '8.7M', '5,000,000+', 'Free', '0', 'Everyone', 'Art & Design', 'August
1, 2018', '1.2.4', '4.0.3 and up']


['Sketch - Draw & Paint', 'ART_AND_DESIGN', '4.5', '215644', '25M', '50,000,000
+', 'Free', '0', 'Teen', 'Art & Design', 'June 8, 2018', 'Varies with device',
'4.2 and up']


['Pixel Draw - Number Art Coloring Book', 'ART_AND_DESIGN', '4.3', '967', '2.8
M', '100,000+', 'Free', '0', 'Everyone', 'Art & Design;Creativity', 'June 20, 2
018', '1.1', '4.4 and up']


num de linhas: 10841
num de colunas 13
None


['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price',
'Content Rating', 'Genres', 'Last Updated', 'Current Ver', 'Android Ver']
```

3. **limpar os dados**: deletar o que não for relevante no meu escopo, como apps pagos ou em língua não inglesa.

3.1. *remover apps sem alguns parâmetros*:

In [4]:
```
#data cleaning
print(android_corpo[10472])
# linha não tem categoria
```

```
['Life Made WI-Fi Touchscreen Photo Frame', '1.9', '19', '3.0M', '1,000+', 'Fre
e', '0', 'Everyone', '', 'February 11, 2018', '1.0.19', '4.0 and up']
```

In [5]:
```
print(len(android_corpo))
del android_corpo[10472]
print(len(android_corpo))
```

```
10841
10840
```

3.2 *remover apps em duplicidade*:

obs: esse paço não é realizado com ios pois sua lista não tem duplicidade, para verificação basta chegar a coluna 'id' na lista ios_header.

In [6]:
```
unicos = []
duplicados = []

for app in android_corpo:
    nome = app[0]
    if nome in unicos:
        duplicados.append(nome)
    else:
        unicos.append(nome)

print(len(duplicados))
```

```
1181
```

para remover os apps em duplicidade verifiquei qual teria o maior valor de reviews, tendo em vista que essa seria a versão de dados mais atual.

In [7]:
```
reviews_max = {}

for app in android_corpo:
    nome = app[0]
    reviews = float(app[3])
    if nome in reviews_max and reviews_max[nome] < reviews:
        reviews_max[nome] = reviews
    elif nome not in reviews_max:
        reviews_max[nome] = reviews

# verificação da limpeza esperada x limpeza feita
print(len(android_corpo) - 1181)
print(len(reviews_max))
```

```
9659
9659
```

usando o dicionário acima para remover os app duplicados, criei uma nova lista limpa e outra com os nomes já adicionados:

In [8]:
```
android_limpa = []
jadd = []

for app in android_corpo:
    nome = app[0]
    reviews = float(app[3])
    if reviews_max[nome] == reviews and nome not in jadd:
        android_limpa.append(app)
        jadd.append(nome)

#teste de limpeza esperada = 9659
print(explora_dados(android_limpa, 0, 2, True))
```

```
['Photo Editor & Candy Camera & Grid & ScrapBook', 'ART_AND_DESIGN', '4.1', '15
9', '19M', '10,000+', 'Free', '0', 'Everyone', 'Art & Design', 'January 7, 201
8', '1.0.0', '4.0.3 and up']


['U Launcher Lite – FREE Live Cool Themes, Hide Apps', 'ART_AND_DESIGN', '4.7',
'87510', '8.7M', '5,000,000+', 'Free', '0', 'Everyone', 'Art & Design', 'August
1, 2018', '1.2.4', '4.0.3 and up']


num de linhas: 9659
num de colunas 13
None
```

3.3 *remover apps de língua não inglesa*:

- encontrar os apps cujo nome contenha caracteres que não são letras do alfabeto Inglês (a - z), números (0 - 9), pontuações (., !, ?, ;) e outros simbolos (+, *, /).
- para isso utilizei a função built-in ord() que disponibiliza um número de codificação correspondente a cada caractere.
- os caracteres mais comuns do Inglês estão no range de 0 - 127 de acordo com a ASCII
- como existem apps que contém caracteres acima de 127 e mesmo assim são de língua inglesa, como 'Instachat 😜' ou 'Docs To Go™ Free Office Suite', optei deletar os que teriam mais de 3 caracteres acima de 127.

In [9]:
```
def ingles(string):
    non_ascii = 0

    for caractere in string:
        if ord(caractere) > 127:
            non_ascii += 1

    if non_ascii > 3:
        return False
    else:
        return True

#teste
print(ingles('teste'))
print(ingles('ブロックパズ'))
print(ingles('Instachat 😜'))
print(ingles('Docs To Go™ Free Office Suite'))
```

```
True
False
True
True
```

In [10]:
```
#nova lista de apps apenas na língua inglesa
android_ingles = []
ios_ingles = []

for app in android_limpa:
    nome = app[0]
    if ingles(nome) is True:
        android_ingles.append(app)

for app in ios_corpo:
    nome = app[1]
    if ingles(nome) is True:
        ios_ingles.append(app)

#numero de apps em inglês:
print(len(android_ingles))
print(len(ios_ingles))
```

```
9614
6183
```

3.4 *remover apps pagos*

In [11]:
```
android_gratuito = []
ios_gratuito = []

for app in android_ingles:
    valor = app[7]
    if valor == '0':
        android_gratuito.append(app)

for app in ios_ingles:
    valor = app[4]
    if valor == '0.0':
        ios_gratuito.append(app)

#quantidade de apps gratuitos em casa plataforma
print(len(android_gratuito))
print(len(ios_gratuito))
```

```
8864
3222
```

4. **análise**:

o objetivo final é lançar a aplicação no Google Play e na App Store, portanto é preciso encontrar perfis de aplicativos que tenham sucesso em ambos os mercados.

4.1 *gêneros mais comuns em cada mercado*:

4.1.1 dentro da função "tabelafreq" preciso contar quantas vezes cada gênero aparece em casa lista de apps. para isso, crio um dicionário com os gênreos e sua respectiva frequência, a qual será posteriormente transformada em porcentagem.

In [13]:
```
def tabelafreq(dataset, index):
    tabela = {}
    total = 0

    for app in dataset:
        total += 1
        valor = app[index]
        if valor in tabela:
            tabela[valor] += 1
        else:
            tabela[valor] = 1

    tabela_porcentagem = {}
    for chave in tabela:
        porcentagem = (tabela[chave] / total) * 100
        tabela_porcentagem[chave] = porcentagem

    return tabela_porcentagem
```

4.1.2 colocar a tabela criada via dicionário em ordem decrescente para visualizar com maior clareza os gêneros mais comuns:

- transformo o dicionário em tupla;
- uso a função sorted() para colocar em ordem crescente;
- uso o reverse para colocar em ordem decrescente.

In [14]:
```
def tabela_decresente(dataset, index):
    tabela = tabelafreq(dataset, index)
    tabela_tupla = []
    for chave in tabela:
        chave_val_tupla = (tabela[chave], chave)
        tabela_tupla.append(chave_val_tupla)

    tabela_sorted = sorted(tabela_tupla, reverse = True)
    for e in tabela_sorted:
        print(e[1], ':', e[0])
```

In [15]:
```
#aplicando as funções nos mercados:

print(tabela_decresente(ios_gratuito, 11))
print('\n')
print(tabela_decresente(android_gratuito, 1))
print('\n')
print(tabela_decresente(android_gratuito, 9))
```

```
Games : 58.16263190564867
Entertainment : 7.883302296710118
Photo & Video : 4.9658597144630665
Education : 3.662321539416512
Social Networking : 3.2898820608317814
Shopping : 2.60707635009311
Utilities : 2.5139664804469275
Sports : 2.1415270018621975
Music : 2.0484171322160147
Health & Fitness : 2.0173805090006205
Productivity : 1.7380509000620732
Lifestyle : 1.5828677839851024
News : 1.3345747982619491
Travel : 1.2414649286157666
Finance : 1.1173184357541899
Weather : 0.8690254500310366
Food & Drink : 0.8069522036002483
Reference : 0.5586592178770949
Business : 0.5276225946617008
```

4.1.3 resultados parciais:

App Store

- o gênero mais comum é games com 58% e em segundo lugar entretenimento com aproximadamente 8%, seguido por aplicativos de foto e vídeo próximos a 5%. Apenas 3,7% dos aplicativos são para educação e 3,3% para redes sociais.
- a impressão geral é que a App Store, dentro do escopo de análise, é dominada por aplicativos que são projetados para se divertir (jogos, entretenimento, foto e vídeo, redes sociais, esportes, música, etc.), enquanto aplicativos com propósitos práticos (educação, compras, serviços públicos, produtividade, estilo de vida, etc.) são mais raros.

Google Play

- O cenário parece significativamente diferente: não existem muitos aplicativos projetados para se divertir e parece que um bom número de aplicativos são projetados para fins práticos (família, ferramentas, negócios, estilo de vida, produtividade, etc.). No entanto, ao investigar mais a fundo, descobri que a categoria família (quase 19% dos aplicativos) são principalmente jogos para crianças.

Conclusões parciais

- aplicativos práticos parecem ter uma melhor representação na Google Play em comparação com a App Store, que é mais voltada ao entretenimento.
- o fato de um gênero de apps ser mais numeroso não significa que eles tenham o maior número de usuários - a demanda pode não ser a mesma que a oferta.
- portanto, não é possível recomendar um perfil de aplicativo apenas com essas informações. Ainda se faz necessário analisar a quantidade de usuários e descobrir quais gêneros têm mais usuários.

4.2 *Aplicativos mais populares por gênero em cada mercado*

para essa análise trabalharei com a média de instalações por gênero na Google play e na App Store o número de avaliações de usuários por gênero.

4.2.1 App Store:

In [17]:
```
generos_ios = tabelafreq(ios_gratuito, 11)

for genero in generos_ios:
    total = 0 # total de avaliações, pois não há número de instalações
    len_genero = 0 # quantos apps esse gênero tem
    for app in ios_gratuito:
        genero_app = app[11]
        if genero_app == genero:
            n_avaliacao = float(app[5])
            total += n_avaliacao
            len_genero += 1

    media_ios = total / len_genero
    print(genero, ':', media_ios)
```

```
Social Networking : 71548.34905660378
Photo & Video : 28441.54375
Games : 22788.6696905016
Music : 57326.530303030304
Reference : 74942.11111111111
Health & Fitness : 23298.015384615384
Weather : 52279.892857142855
Utilities : 18684.456790123455
Travel : 28243.8
Shopping : 26919.690476190477
News : 21248.023255813954
Navigation : 86090.33333333333
Lifestyle : 16485.764705882353
Entertainment : 14029.830708661417
Food & Drink : 33333.92307692308
Sports : 23008.898550724636
Book : 39758.5
Finance : 31467.944444444445
Education : 7003.983050847458
Productivity : 21028.410714285714
Business : 7491.117647058823
Catalogs : 4004.0
Medical : 612.0
```

conclusões parciais:

- o gênero mais popular é o de navegação, seguido por referência, redes sociais e música.
- todavia não recomendaria o perfil de navegação (waze e google maps), tampouco o de redes sociais (Facebook, Pinterest, Skype) pois sua média é fortemente influenciada por alguns gigantes como os acima citados.
- sendo assim, minha recomendação seria um app no perfil de referência ou música, ou até quem sabe, algo que una música e referência.

4.2.2 Google Play:

o número de instalções aparece como 100+, 1,000+, 5,000+, etc. Sendo assim, para realizar os cálculos será preciso remover as vírgulas e os caracteres de adição para não gerar erro.

In [19]:
```
genero = tabelafreq(android_gratuito, 1)

for categoria in genero:
    total = 0
    len_categoria = 0
    for app in android_gratuito:
        app_categoria = app[1]
        if app_categoria == categoria:
            n_instal = app[5]
            n_instal = n_instal.replace('+', '')
            n_instal = n_instal.replace(',', '')
            total += float(n_instal)
            len_categoria += 1

    media_android = total / len_categoria
    print(categoria, ':', media_android)
```

```
ART_AND_DESIGN : 1986335.0877192982
AUTO_AND_VEHICLES : 647317.8170731707
BEAUTY : 513151.88679245283
BOOKS_AND_REFERENCE : 8767811.894736841
BUSINESS : 1712290.1474201474
COMICS : 817657.2727272727
COMMUNICATION : 38456119.167247385
DATING : 854028.8303030303
EDUCATION : 1833495.145631068
ENTERTAINMENT : 11640705.88235294
EVENTS : 253542.22222222222
FINANCE : 1387692.475609756
FOOD_AND_DRINK : 1924897.7363636363
HEALTH_AND_FITNESS : 4188821.9853479853
HOUSE_AND_HOME : 1331540.5616438356
LIBRARIES_AND_DEMO : 638503.734939759
LIFESTYLE : 1437816.2687861272
GAME : 15588015.603248259
FAMILY : 3695641.8198090694
MEDICAL : 120550.61980830671
SOCIAL : 23253652.127118643
SHOPPING : 7036877.311557789
PHOTOGRAPHY : 17840110.40229885
SPORTS : 3638640.1428571427
TRAVEL_AND_LOCAL : 13984077.710144928
TOOLS : 10801391.298666667
PERSONALIZATION : 5201482.6122448975
PRODUCTIVITY : 16787331.344927534
PARENTING : 542603.6206896552
WEATHER : 5074486.197183099
VIDEO_PLAYERS : 24727872.452830188
NEWS_AND_MAGAZINES : 9549178.467741935
MAPS_AND_NAVIGATION : 4056941.7741935486
```

conclusões parciais:

- em média, os aplicativos de comunicação têm o maior número de instalações. Este número é fortemente distorcido por alguns aplicativos que têm mais de um bilhão de instalações

(WhatsApp, Facebook Messenger, Skype, Google Chrome, Gmail e Hangouts).

- o mesmo raciocínio se aplica para o segundo colocado, que são os players de vídeo como Youtube, Google Play Filmes e TV etc e o terceiro colocado, os apps sociais, como Facebook, Instagram, Google+, etc.
- a principal preocupação é que esses gêneros de aplicativos possam parecer mais populares do que realmente são. Além disso, esses nichos parecem ser dominados por alguns gigantes contra os quais é difícil competir.
- os livros e o gênero de referência também parecem bastante populares, com um número médio de instalações de 8.767.811. É interessante explorar isso com mais profundidade tendo em vista que é um gênero que também tem potencial para funcionar na App Store.
- não foi encontrada uma categoria específica para música, esse perfil de aplicativo deve fazer parte dos apps de entretenimento, os quais aparecem com uma boa média.

Conclusão

como o objetivo era recomendar um gênero de app que mostra potencial para ser lucrativo tanto na App Store quanto no Google Play, recomenda-se aprofundar o estudo em torno das categorias de livros e referências, bem como de música.

In [ ]: