Análise de Dados

Universidade Fernando Pessoa

Integração e exploração de dados de um conjunto de dados de múltiplas fontes

Elaborado por:
Gonçalo Cunha, 2022110211
Vasco Martins, 2022121836

Índice

Resumo

Este projeto integrou três fontes de dados – World Happiness (felicidade nacional), survey de tempo em redes sociais (perfil individual) e prevalência de transtornos mentais (Global Burden of Disease) – por meio de inner join em “Country”. Após padronizar nomes de países e filtrar saúde mental para 2017, obteve-se um conjunto final de utilizadores cujos países têm cobertura completa em todas as tabelas.

1. Introdução

Em um cenário movido a dados, combinar diferentes bases aumenta a profundidade dos insights. Este estudo teve o objetivo de investigar como o comportamento individual em redes sociais (horas diárias, demografia, renda) se relaciona a indicadores nacionais de felicidade e saúde mental, todos referentes a 2019. Para isso, foram usados:

- **World Happiness 2017** (Score, GDP per capita, social support etc.),
- **Survey de uso de redes em 2017**(age, gender, time_spent, platform, income etc.)
- **Prevalência de transtornos mentais** (schizophrenia %, depression %, anxiety % etc.).

Após padronizar “Country” e aplicar inner join, formou-se um dataset coeso.

2. Fontes de Dados (Data Sources)

2.1 Dataset de World Happiness (Indicadores de Felicidade)

Este conjunto de dados provém do Kaggle, uma plataforma de compartilhamento de dados amplamente utilizada por cientistas de dados e pesquisadores. Ele é baseado no *World Happiness Report*, uma publicação anual que classifica países de acordo com o bem-estar percebido de sua população. O principal objetivo é medir o nível de felicidade ou satisfação com a vida em diferentes nações, considerando fatores econômicos, sociais e de saúde.

Contexto e Fonte:

- Elaborado pela Rede de Soluções para o Desenvolvimento Sustentável das Nações Unidas (UN Sustainable Development Solutions Network).
- Baseia-se em entrevistas aos habitantes de cada país, que avaliam sua satisfação geral de vida em uma escala de 0 a 10 (escala de Gallup World Poll).
- Complementado por indicadores auxiliares (PIB per capita, apoio social, expectativa de vida saudável etc.) que explicam as diferenças de pontuação entre países.

2.2 Dataset de Tempo Passado nas Redes Sociais

Este conjunto de dados coleta informações demográficas e comportamentais de utilizadores em relação ao uso de redes sociais, com foco em "time_spent" (tempo gasto). Serve para análises de hábitos de navegação, segmentação de público e correlação entre características pessoais e plataformas preferidas.

Contexto e Fonte:

- Geralmente construído a partir de surveys (pesquisas online), questionários ou logs de aplicativos que registam quanto tempo o utilizador dedica a cada plataforma social.
- Inclui atributos demográficos (idade, gênero, localização) e socioeconômicos (renda, profissão).

2.3 Dataset de Doenças Mentais (por País e Ano)

Este conjunto de dados agrupa estimativas de prevalência de diferentes transtornos mentais em percentagem da população, para vários países ao longo dos anos.

Contexto e Fonte:

- Mede a carga de doenças mentais em termos percentuais, possibilitando comparações entre países e análises de evolução temporal (anos de 1990 em diante).

3. Metodologia de Integração

Para unir os três datasets e garantir que apenas países **presentes em todas as três fontes** sejam considerados, optou-se pelo método de **inner join**. O fluxo de junção foi o seguinte:

Padronização da coluna de país:

- Em df_mental_health, a coluna original "Entity" foi renomeada para **"Country"**.

- Em df_time_on_social_media, a coluna original “location” foi renomeada para **“Country”**.

- Em df_world_happiness, a coluna original “Country or region” foi renomeada para **“Country”**.

Filtragem do dataset de saúde mental para o ano de 2017:

- Foi garantido que Year fosse tipo inteiro e, em seguida, aplicou-se df_mental_health[df_mental_health["Year"] == 2017].

- O resultado, chamado df_mental_2017, contém apenas registros de prevalências de transtornos mentais relativos a 2017.

Merge (inner) entre Saúde Mental 2017 e World Happiness 2017:

- Foi feito um merged_df(df_mental_2017, df_world_happiness, on="Country", how="inner").

- Como df_world_happiness já é, naturalmente, composto por registros de 2017, não houve necessidade de adicionar coluna “Year” nele.

- O resultado, chamado df_mental_2017, inclui apenas os países que estão presentes em **ambos** os datasets de saúde mental (2017) e de felicidade (2017).

- Caso um país não exista em uma das duas fontes, ele é automaticamente excluído da tabela final.

Merge (inner) entre merged_df e Redes Sociais:

- Em seguida, executou-se merged_df2(merged_df, df_time_on_social_media, on="Country", how="inner").

- Isso preserva apenas os registos em que o país do utilizador(no dataset de redes sociais) também tenha indicadores de saúde mental e felicidade em 2017.

- O dataset final, que chamamos de merged_df2, contém apenas:

    - Utilizadores cujos países aparecem **simultaneamente** em df_time_on_social_media, df_mental_2017 e df_world_happiness (todas em 2017).

    - As colunas combinadas dos três conjuntos de dados — ou seja, atributos individuais de cada usuário (idade, gênero, tempo gasto, etc.), prevalências de transtornos mentais de seu país e indicadores de felicidade nacional.

3.1 Vantagens do Inner Join em Relação ao Outer Join

- **Foco apenas nos países com cobertura completa**: Ao usar inner join, o dataset final não contém valores nulos resultantes de países que estariam ausentes em uma das fontes. Dessa forma, toda linha possui:
    - Ao menos um valor de prevalência de transtorno mental em 2017.
    - Ao menos um valor de felicidade (Score) para 2017.
    - Ao menos um registo individual associado ao país.
- **Análises sem necessidade de tratar NaNs de merge**: Todas as linhas de merged_df2 têm colunas preenchidas em cada bloco de informação, reduzindo a complexidade de EDA quando se lida com valores em falta resultantes de junção.
- **Consistência Geográfica e Temporal**: Garante que cada utilizador analisado pertença a um país que, em 2017, tenha tanto métricas de saúde mental quanto de felicidade disponíveis.
- **Menor volume de dados para análise**: Embora o outer join normalmente gere um número maior de linhas (incluindo "países fantasmas" e usuários isolados), o inner join produz um subconjunto menor, mais coeso, adequado para fins comparativos exatos em 2017.

4.Análise e Tratamento de Valores em Falta e de Valores Atípicos

Após o inner join não houve valores em falta, decidimos então para demonstração usar o outer join.

4.1 Análise de Valores em Falta

Executou-se:

```
print("\nMissing Values per Column:")
print(merged_df2.isnull().sum())

print("\nMissing Values per Column(%):")
missing_percentage = (merged_df2.isnull().sum() / len(merged_df2)) * 100
print(missing_percentage)
```

Para analisar os valores em falta por coluna e a sua percentagem.

4.2 Tratamento de Valores em Falta

Preenchimento de Valores em Falta através da média:

A **média** é o valor obtido somando todos os números de um conjunto e dividindo pelo total de elementos. Ela representa uma noção de equilíbrio dos dados.

```
# Fill with mean
for col in ["Schizophrenia (%)","Bipolar disorder (%)","Eating disorders (%)","Anxiety disorders (%)"]: merged_df2[col].fillna
 (merged_df2[col].mean(), inplace=True)
```

Preenchimento de Valores em Falta através da mediana:

A **mediana** é o número que ocupa a posição central em um conjunto de dados ordenados. Se houver um número par de elementos, é a média dos dois centrais.

```
# Fill with median
for col in ["Depression (%)","Alcohol use disorders (%)"]: merged_df2[col].fillna(merged_df2[col].median(), inplace=True)
```

Preenchimento de Valores em Falta através da moda:

A **moda** é o valor que mais se repete no conjunto, ou seja, o mais frequente entre os dados.

```
# Fill with mode
for col in ["platform","interests"]: merged_df2[col].fillna(merged_df2[col].mode()[0], inplace=True)
```

4.3 Análise de Valores Atípicos (outliers)

Valores Atípicos, outliers, são valores que **fogem do padrão** de um conjunto de dados — ou seja, são **muito maiores ou muito menores** do que a maioria dos outros valores.

Antes de aplicar técnicas de modelagem ou correlação, identificou-se a presença de valores extremos em variáveis contínuas de saúde mental:

- Boxplot de Valores Atípicos:

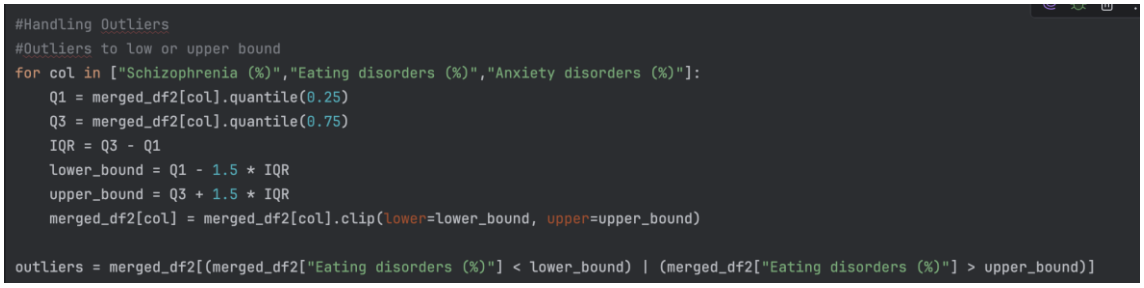
```
#Boxplot Before Handling Outliers
plt.figure(figsize=(12, 6))
sns.boxplot(data=merged_df2[["Schizophrenia (%)","Eating disorders (%)","Anxiety disorders (%)"]])
plt.title("Boxplot Before Outlier Handling")
plt.show()
```

4.4 Tratamento de Valores Atípicos (outliers)

Implementamos o seguinte código para lidar com Valores Atípicos:

1º Calculamos o primeiro quartil (Q1), ou seja, o valor abaixo do qual estão 25% dos dados.
2º Calculamos o terceiro quartil (Q3), ou seja, o valor abaixo do qual estão 75% dos dados.
3º Calculamos o **IQR** (Interquartile Range) que consiste na diferença entre o terceiro e o primeiro quartil. Representa a amplitude da "caixa" no boxplot: quanto maior for esse valor, mais espalhados estão 50% dos dados centrais.

4º Definimos o limite inferior e o limite superior.
5º Utilizamos o método “clip" para para “limitar” valores fora dos limites.



```
#Handling Outliers
#Outliers to low or upper bound
for col in ["Schizophrenia (%)","Eating disorders (%)","Anxiety disorders (%)"]:
    Q1 = merged_df2[col].quantile(0.25)
    Q3 = merged_df2[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    merged_df2[col] = merged_df2[col].clip(lower=lower_bound, upper=upper_bound)

outliers = merged_df2[(merged_df2["Eating disorders (%)"] < lower_bound) | (merged_df2["Eating disorders (%)"] > upper_bound)]
```

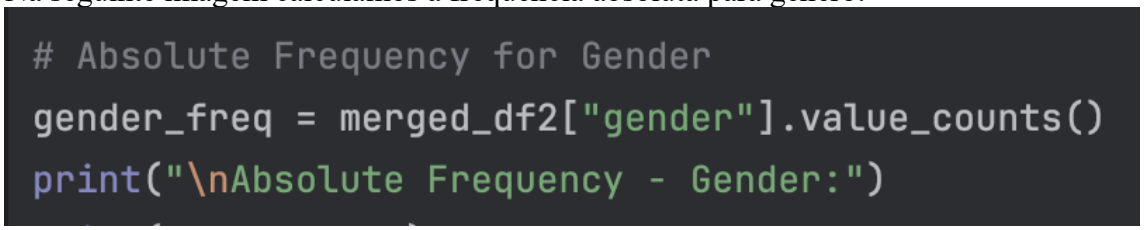
Dessa forma, ao fim do for, as três colunas estarão “limitadas” para que nenhum valor extremo ultrapasse o intervalo [Q1 − 1.5·IQR, Q3 + 1.5·IQR]. Isso ajuda em análises estatísticas posteriores, pois reduz a influência de pontos que seriam considerados atípicos.

5. Estatíticas

5.1 Frequência Absoluta

Frequência absoluta consiste no **número de vezes** que um determinado valor aparece num determinado conjunto de dados.

Na seguinte imagem calculamos a frequência absoluta para género:

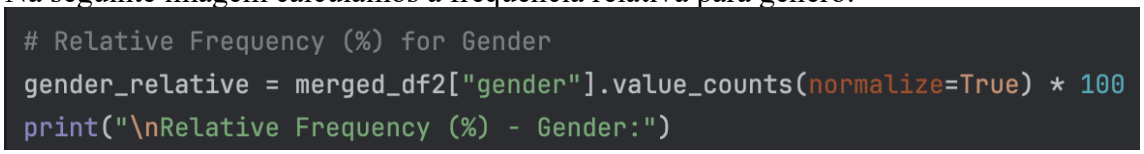


```
# Absolute Frequency for Gender
gender_freq = merged_df2["gender"].value_counts()
print("\nAbsolute Frequency - Gender:")
```

5.2 Frequência Relativa

A **frequência relativa** representa **a proporção (ou percentagem)** de vezes que um determinado valor aparece **em relação ao total de dados**.

Na seguinte imagem calculamos a frequência relativa para género:



```
# Relative Frequency (%) for Gender
gender_relative = merged_df2["gender"].value_counts(normalize=True) * 100
print("\nRelative Frequency (%) - Gender:")
```

5.3 Frequência Comutativa

Frequência acumulada (ou frequência comutativa, às vezes também chamada de frequência “cumulativa”) é a soma das frequências absolutas **sucessivas** até um certo ponto da lista de valores ou classes ordenadas. Em outras palavras, indica quantos elementos estão “até” aquele valor.

Na seguinte imagem calculamos a frequência comutativa para o tempo passado em redes sociais:

```
# Cumulative Frequency
time_spent_cumfreq = merged_df2["time_spent"].value_counts().sort_index().cumsum()
print("\nCumulative Frequency - Time spent on Social Media:")
```

6. Vizualização de Dados

6.1 Tempo médio gasto por plataforma e gênero

Ao elaborar um heatmap com o objetivo de analizar o tempo médio gasto por plataforma e gênero concluimos que mulheres são mais ativas em redes sociais que os homens, o que bate certo com o que podemos conferir nestas duas notícias:

https://gauchazh.clicrbs.com.br/comportamento/noticia/2022/03/mulheres-sao-mais-conectadas-do-que-os-homens-mas-acessam-menos-servicos-na-internet-cl0i2uxw50018017chlm51fy6.html

https://inovag.com.br/2019/05/27/homem-ou-mulher-quem-e-mais-ativo-nas-redes-sociais/

6.2 Índice de Rendimento Médio vs Felicidade por País

Ao elaborar um gráfico de pontos com o objetivo de analizar se o rendimento afeta a felicidade da pessoas e concluimos que em países em que o rendimento é superior o indice de felicidade é maior.

6.3 Evolução Anual: Tempo gasto em redes socias vs Transtornos de Ansiedade (%)

Ao elaborar um gráfico com o objetivo de analizar se o tempo gasto em redes sociais contribui para transtorno de ansiedade concluimos que não existe ligação que nos leve a dizer que o tempo gasto em redes sociais influencia o aparecimento de transtornos de ansiedade, o que contrasta com vários estudos sobre o assunto como a seguinte notícia que revela que “Excesso de redes sociais está associado a 45% dos casos de ansiedade em jovens”.

https://veja.abril.com.br/saude/excesso-de-redes-sociais-esta-associado-a-45-dos-casos-de-ansiedade-em-jovens/

Os resultados que obtivemos diferentes da realidade pode se dever ao facto de os datasets utilizados serem diferentes para o tempo gasto em redes socias e para as percentagens de transtornos de ansiedade e também pelo facto de serem pequenas amostras de pessoas comparativamente ao numero total de utilizadores de redes sociais.

7. Conclusão

De forma geral, as análises realizadas indicam três pontos principais:

1. **Tempo médio por plataforma e gênero**: observou-se que, no nosso conjunto de dados, as mulheres passam em média mais tempo em redes sociais do que os homens, resultado que confirma tendências apontadas por estudos externos.

2. **Rendimento médio vs. felicidade por país**: constatou-se uma relação positiva entre renda e satisfação de vida em nível nacional, isto é, países com maior índice de rendimento tendem a apresentar também um Score de felicidade mais elevado.

3. **Evolução anual de tempo em redes vs. transtornos de ansiedade**: não encontramos, nos dados analisados, evidência de que um maior uso de redes sociais esteja associado a elevações proporcionais na prevalência de transtornos de ansiedade. Embora exista literatura apontando correlações nesse sentido, a discrepância pode ser resultado de amostras restritas e de fontes distintas para tempo de uso e dados de saúde mental.

Em síntese, embora tenhamos identificado padrões coerentes em relação a gênero e renda, a ausência de relação clara entre uso de redes e ansiedade evidencia limitações dos conjuntos de dados utilizados – tais como o tamanho amostral e a falta de total sincronia temporal entre variáveis. Futuros estudos com amostras maiores e dados mais alinhados poderiam fornecer maior robustez às conclusões sobre o impacto das redes sociais na saúde mental.