

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

1.Introdução	2
2.Fonte dos dados (Data Sources)	3
2.1 Dataset de World Happiness (Indicadores de Felicidade).....	3
2.2 Dataset de Tempo Passado nas Redes Sociais.....	3
2.3 Dataset de Doenças Mentais (por País e Ano)	3
3. Metodologias de Integração e Importação.....	3
3.1 Vantagens do Inner Join em Relação ao Outer Join	5
4. Análise e Tratamento de Valores em Falta e de Valores Atípicos	6
4.1 Análise de Valores em Falta	6
4.2 Tratamento dos Valores em Falta	6
Preenchimento de Valores em Falta através da média:	6
Preenchimento de Valores em Falta através da mediana:	6
Preenchimento de Valores em Falta através da moda:	6
4.3 Análise de Valores Atípicos (outliers).....	7
4.4 Tratamento de Valores Atípicos (outliers)	7
5. Estatísticas.....	8
5.1 Frequência Absoluta	8
5.2 Frequência Relativa	8
5.3 Frequência Comutativa.....	9
6. Vizualização de Dados	9
6.1 Tempo médio gasto por plataforma e gênero	9
6.2 Índice de Rendimento Médio vs Felicidade por País	10
6.3 Evolução Anual: Tempo gasto em redes sociais vs Transtornos de Ansiedade (%).....	11
7. Conclusão	12

1.Introdução

Este projeto integrou três fontes de dados dataset World Happiness (felicidade nacional),dataset de tempo em redes sociais e dataset de prevalência de transtornos mentais (Global Burden of Disease), e explora relações entre tempo de uso de redes sociais doenças mentais e rank de felicidade global.

2. Fonte dos dados (Data Sources)

2.1 Dataset de World Happiness (Indicadores de Felicidade)

Este conjunto de dados 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.

Link: <https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2017.csv>

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.

Link: <https://www.kaggle.com/datasets/imyjoshua/average-time-spent-by-a-user-on-social-media>

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.

Link: <https://www.kaggle.com/datasets/imtkaggleteam/mental-health>

3. Metodologias de Integração e Importação

Quando tentámos carregar o ficheiro CSV com percentagens de “Schizophrenia (%)” e “Eating disorders (%)”, deparamos-nos com um erro de “mixed types”: algumas células continham valores como “n/d” ou percentagens com vírgulas, enquanto outras eram números puros. O Pandas tentou inferir o tipo automaticamente e falhou, tratando a coluna como object (equivalente a texto) e impedindo qualquer operação matemática.

Para resolver, obrigámos o Pandas a ler essas colunas como strings (usando `dtype=str`), garantindo que todo o conteúdo viria como texto bruto. De seguida, aplicámos `pd.to_numeric(..., errors='coerce')` a cada uma, convertendo as strings numéricas em floats e transformando valores inválidos em NaN. Assim passámos a ter colunas numéricas limpas, prontas para cálculos.

```
#Passar para numérico onde tem string
df_mental_health = pd.read_csv(file_path1, low_memory=False, dtype={
    'Schizophrenia (%)': str,
    'Eating disorders (%)': str
})

# Converter coluna 5 para float
df_mental_health['Schizophrenia (%)'] = pd.to_numeric(df_mental_health['Schizophrenia (%)'], errors='coerce')
# Converter coluna 6 para float
df_mental_health['Eating disorders (%)'] = pd.to_numeric(df_mental_health['Eating disorders (%)'], errors='coerce')
```

Para unir os três datasets e garantir que apenas países presentes em todas as três fontes sejam considerados, optamos 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”.

```
# Merge datasets
df_mental_health = df_mental_health.rename(columns={"Entity": "Country"})
df_time_on_social_media = df_time_on_social_media.rename(columns={"location": "Country"})
df_world_happiness = df_world_happiness.rename(columns={"Country or region": "Country"})

merged_df = pd.merge(df_mental_health, df_time_on_social_media, on="Country", how='inner')
merged_df2 = pd.merge(merged_df, df_world_happiness, on="Country", how='inner')
```

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 registros 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:

- Pelo menos um valor de prevalência de transtorno mental em 2017.
- Pelo menos um valor de felicidade (Score) para 2017.
- Pelo 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

```
# Handling Missing Data
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)
```

4.2 Tratamento dos 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)
```

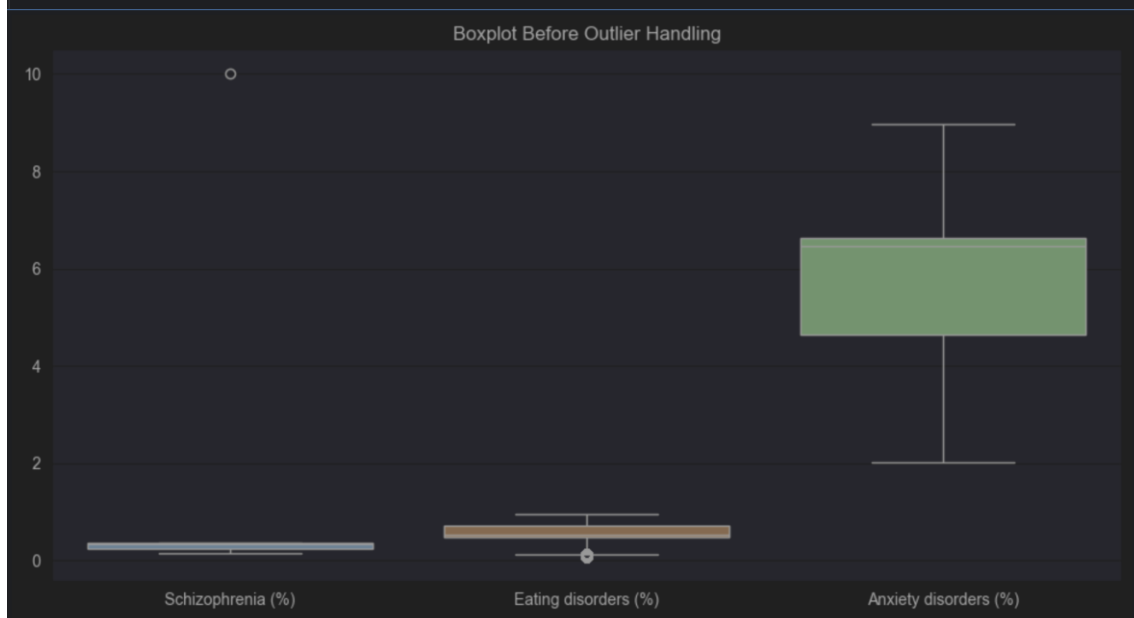
4.3 Análise de Valores Atípicos (outliers)

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:

```
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)]
print(outliers)
```

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

5. Estatísticas

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:")
print(gender_freq)
```

5.2 Frequência Relativa

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:")
print(gender_relative)
```


5.3 Frequência Comutativa

Frequência acumulada (ou frequência comutativa) 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:")
print(time_spent_cumfreq)
```

6. Vizualização de Dados

6.1 Tempo médio gasto por plataforma e gênero

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

Link1:

<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>

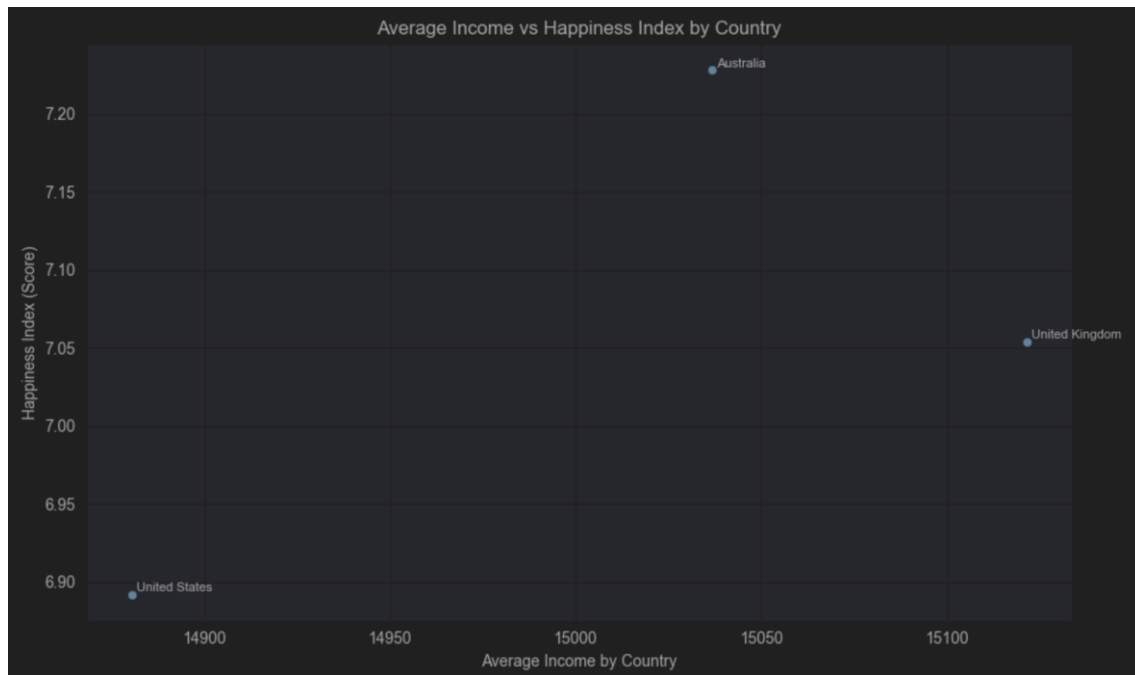
Link2:

<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 analisar se o rendimento afeta a felicidade das pessoas, analisamos que países como a Austrália e o Reino Unido onde o rendimento médio é maior o índice de felicidade é maior do que por exemplo nos Estados Unidos, o que nos permite concluir que em países em que o rendimento é superior o índice de felicidade é maior.

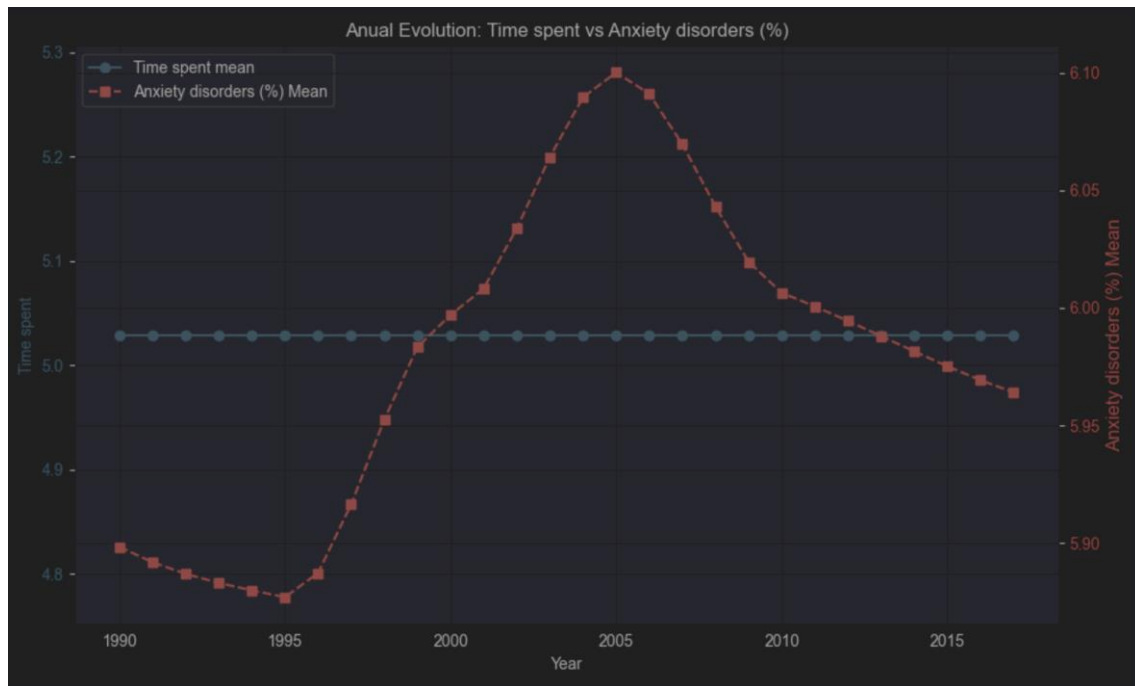


6.3 Evolução Anual: Tempo gasto em redes sociais vs Transtornos de Ansiedade (%)

Ao elaborar um gráfico com o objetivo de analisar se o tempo gasto em redes sociais contribui para transtorno de ansiedade analisamos 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”**.

Link:

<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 sociais 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.