

1 Identificação do dataset escolhido

O conjunto de dados contém uma coleção abrangente de informações relacionadas aos 995 canais com mais inscritos no YouTube, incluindo métricas de desempenho, informações demográficas, detalhes geográficos e estatísticas econômicas dos países de origem dos canais. Isso pode ser usado para análises sobre o sucesso dos canais e a influência de fatores geográficos e socioeconômicos em seus desempenhos. O dataset foi retirado do [Kaggle](#), um repositório online com diversos tipos diferentes de datasets que são disponibilizados publicamente por diversos estudantes e profissionais do mundo inteiro. A última atualização reportada no site do Kaggle para esse dataset é de agosto de 2023, a coleta das informações e dados é utilizando Python. o dataset tem 995 linhas e 28 colunas.

2 Variáveis

As variáveis presentes no dataset são as seguintes:

- rank: Posição do canal de acordo com o número de inscritos
Qualitativa ordinal
- Youtuber: Nome do canal do Youtube
Qualitativa nominal
- inscritos: Número de inscritos no canal
Quantitativa discreta
- total_visualizacoes: Total de visualização contando todos os vídeos do canal
Quantitativa discreta
- categoria: Categoria ou nicho do canal
Qualitativa nominal
- titulo: Título do canal do Youtube
Qualitativa nominal
- qtd_videos: Total de vídeos postados no canal
Quantitativa discreta
- pais: País de origem do canal
Qualitativa nominal
- pais_abrev: Abreviação do país
Qualitativa nominal
- tipo: Tipo do canal (individual ou marca)
Qualitativa nominal
- rank_visualizacoes: Posição do canal de acordo com o total de visualizações
Qualitativa ordinal
- rank_por_pais: Posição do canal de acordo com o número de inscritos dentro do seu país de origem
Qualitativa ordinal
- rank_por_tipo: Posição do canal baseado em seu tipo (individual ou marca)
Qualitativa ordinal

- `visualizacoes_ultimos_30_dias`: Total de visualizações nos últimos 30 dias
Quantitativa discreta
- `menor_ganho_mensal`: Menor ganho mensal estimado do canal
Quantitativa contínua
- `maior_ganho_mensal`: Maior ganho mensal estimado do canal
Quantitativa contínua
- `menor_ganho_anual`: Menor ganho anual estimado do canal
Quantitativa contínua
- `maior_ganho_anual`: Maior ganho anual estimado do canal
Quantitativa contínua
- `inscricoes_ultimos_30_dias`: Número de novos inscritos nos últimos 30 dias
Quantitativa discreta
- `ano_criacao`: Ano em que o canal foi criado
Qualitativa nominal
- `mes_criacao`: Mês em que o canal foi criado
Qualitativa nominal
- `dia_criacao`: Dia em que o canal foi criado
Qualitativa nominal
- `porc_populacao_cursando_ens_superior`: Porcentagem da população matriculada no ensino superior no país de origem
Quantitativa contínua
- `populacao`: População do país de origem
Quantitativa discreta
- `taxa_desemprego`: Taxa de desemprego do país de origem
Quantitativa contínua
- `populacao_urbana`: População urbana no país de origem
Quantitativa discreta
- `latitude`: Latitude do país de origem
Quantitativa contínua
- `longitude`: Longitude do país de origem
Quantitativa contínua

Qualitativa Nominal	Qualitativa Ordinal	Quantitativa Discreta	Quantitativa Contínua	Total
9	4	7	8	28

Tabela quantificando cada tipo de variável presente no dataset.

3 Observações, casos ou instâncias

As variáveis qualitativas nominais são bem diversas, se referem principalmente a nomes e datas. As variáveis mais interessantes são a categoria, que especifica o nicho para o qual o conteúdo do canal é voltado e o tipo do canal, que determina se a produção de conteúdo é gerida por uma pessoa ou é uma marca.

- `youtuber`
- `categoria`

- titulo
- pais
- pais_abrev
- tipo
- ano_criacao
- mes_criacao
- dia_criacao

As variáveis qualitativas ordinais são as que existem em menor quantidade nesse dataset, mas não são as menos importantes. São as variáveis que definem o rank atual dos 995 maiores canais e subranks que organizam os canais por tipo, país de origem (utilizando a quantidade de inscritos) e rankeando os 995 canais por total de visualizações. São as seguintes variáveis:

- rank
- rank_visualizacoes
- rank_por_pais
- rank_por_tipo

Neste dataset as variáveis quantitativas discretas se referem de pessoas reais interagindo com uma plataforma. As observações são uma contagem dessas interações contabilizam a quantidade de cliques que um determinado canal do YouTube recebeu ao longo da sua existência. Algumas das variáveis nesse caso são:

- inscritos
- total_visualizacoes
- qtd_videos
- visualizacoes_ultimos_30_dias
- inscricoes_ultimos_30_dia
- populacao
- populacao_urbana

As últimas duas variáveis são variáveis referentes ao país de origem do canal do YouTube.

As variáveis quantitativas contínuas se referem à renda do canal, em dólares, à taxas socio-econômicas em porcentagem, como a porcentagem da população cursando ensino superior e desemprego e também tem duas medidas em graus, que são a latitude e longitude atribuídas aos países de origem do canal. As variáveis são as seguintes:

- menor_ganho_mensal
- maior_ganho_mensal
- menor_ganho_anual
- maior_ganho_anual
- porc_populacao_cursando_ens_superior
- taxa_desemprego
- latitude
- longitude

4 Estatística descritiva

Foram gerados alguns gráficos para uma análise inicial dos dados. Os alvos foram as variáveis julgadas mais importantes e relevantes para esse dataset. Um deles é o da quantidade de canais por país 1, uma distribuição de quantos canais existem para cada país presente no banco de dados. Muitos canais são dos Estados Unidos, é o país mais presente no dataset, seguido por Índia e Brasil. A terceira maior ocorrência é de colunas não preenchidas, representadas pelo valor "nan". Esse gráfico foi gerado com o seguinte código em R:

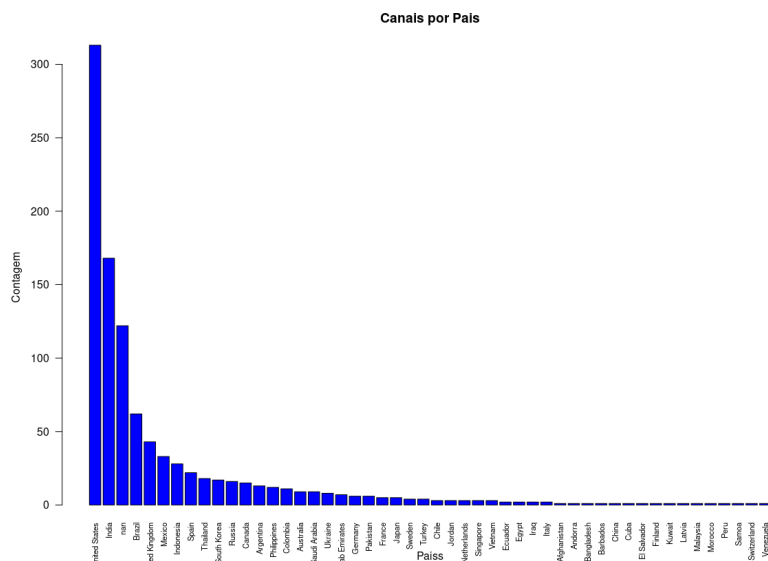


Figura 1: Gráfico da distribuição de canais por país.

```
1 countryTable <- table(database$Country) \\  
2 countryDF <- data.frame(Country = names(countryTable), Contagem = as.numeric(countryTable))  
3 countryDF <- countryDF[order(countryDF$Contagem, decreasing = TRUE), ]  
4 countryDF  
5 barplot(countryDF$Contagem, names.arg = countryDF$Country, col = "blue", main = "Canais por País",  
  , xlab = "País", ylab = "Contagem", las=2, cex.names = 0.7)
```

O ano de criação dos canais segue uma distribuição muito mais equilibrada 2. A maior parte dos canais foi criado em 2014 e em 2006, com um outlier muito interessante de 1970, que é o canal do próprio YouTube, criado junto com a plataforma e que é mantido até hoje como o canal da plataforma. O código que

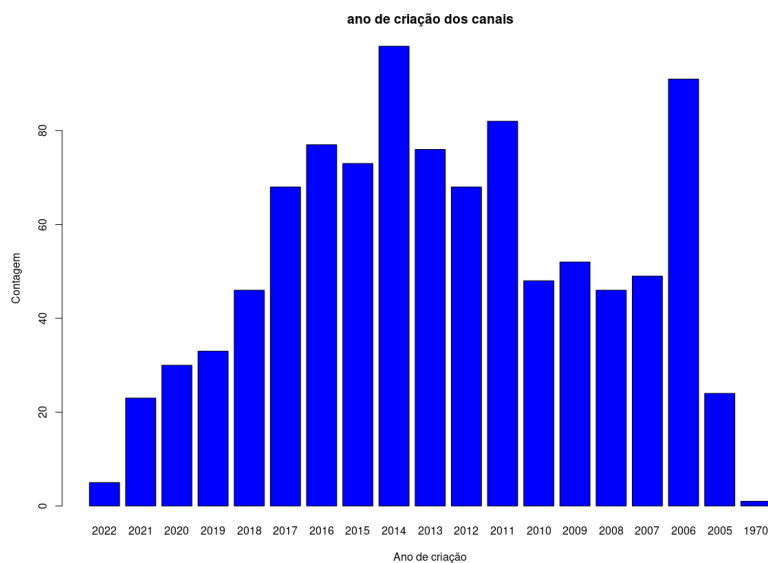


Figura 2: Canais agrupados por ano de criação.

gerou esse gráfico é o seguinte:

```

1 createdYearTable <- table(database\$created\_year)
2 createdYearDF <- data.frame(Year = names(createdYearTable), Contagem = as.numeric(createdYearTable)
3 )
4 createdYearDF <- createdYearDF[order(createdYearDF$Year, decreasing = TRUE),]
5 barplot(createdYearDF$Contagem, names.arg = createdYearDF$Year, col = "blue", main = "ano de
6 cria o dos canais", xlab = "Ano de cria o", ylab = "Contagem")

```

Os canais também foram agrupados pela categoria, que é uma representação simples do tipo de conteúdo que o canal produz 3. As categorias de Entretenimento, Musica e Pessoas&Blogs são as três mais com mais canais. Existem alguns canais sem classificação, mas nessa análise eles tem uma quantidade muito mais inexpressiva. O código em R que cria essa tabela é o seguinte:

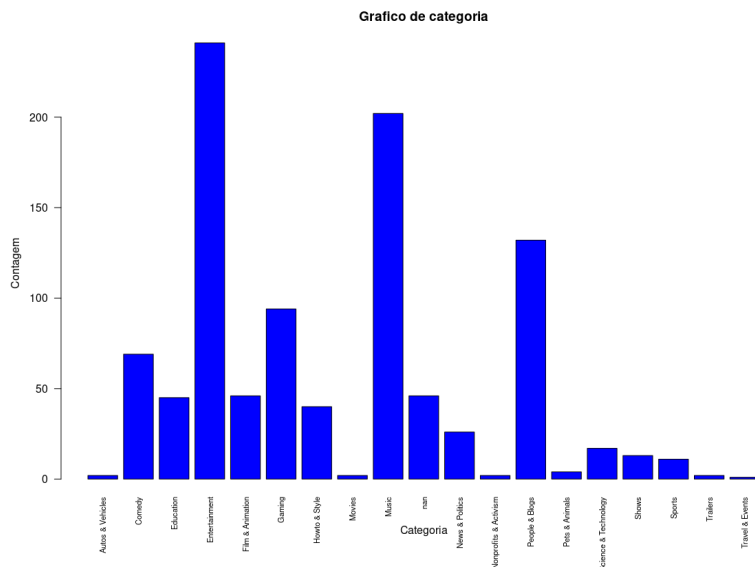


Figura 3: Canais agrupados por categorias.

```

1 categoryTable <- table(database$category)
2 summary(categoryTable)
3 createBarPlot("Canais por Categoria", "Categoria", "Contagem", categoryTable)

```

Análise dos inscritos por boxplot 4

Alguns dados para o box-plot dos inscritos estão na imagem 5.

Os dados e o grafico do boxplot foram gerados com o seguinte código:

```

1 database$subscribers <- as.numeric(database$subscribers)
2 df <- na.omit(database$subscribers)
3 quartis <- quantile(df)
4 medidas\_dispersao(df)
5 plot(df, ylab = "Inscritos", main="Inscritos por canal")
6 boxplot(df, main="Inscritos por canal")

```

Análise ganhos mensais máximo demonstrado na figura 7 e mínimo na figura 6 por box plot.

Os dados referentes aos box plots para o valor máximo e mínimo mensal são, estão respectivamente nas figuras, 9 e 8.

5 Que tipo de pesquisa/pergunta você pretende fazer com este dataset?

O dataset contém dados analíticos dos 995 canais com mais inscritos no YouTube em 2023. O conjunto de dados contém uma coleção abrangente de informações relacionadas a canais do YouTube, incluindo métricas de desempenho, renda, nicho de conteúdo, total de vídeos e também métricas do país de origem do canal, como informações demográficas, detalhes geográficos e estatísticas econômicas. A ideia do trabalho é de correlacionar essas informações e encontrar consequências para estes serem os maiores canais do YouTube. Alguns usos possíveis são:

- tentar entender o que esses canais fazem, quais audiências eles atingem, para serem os maiores do mundo na maior e mais popular plataforma de vídeos longos do mundo

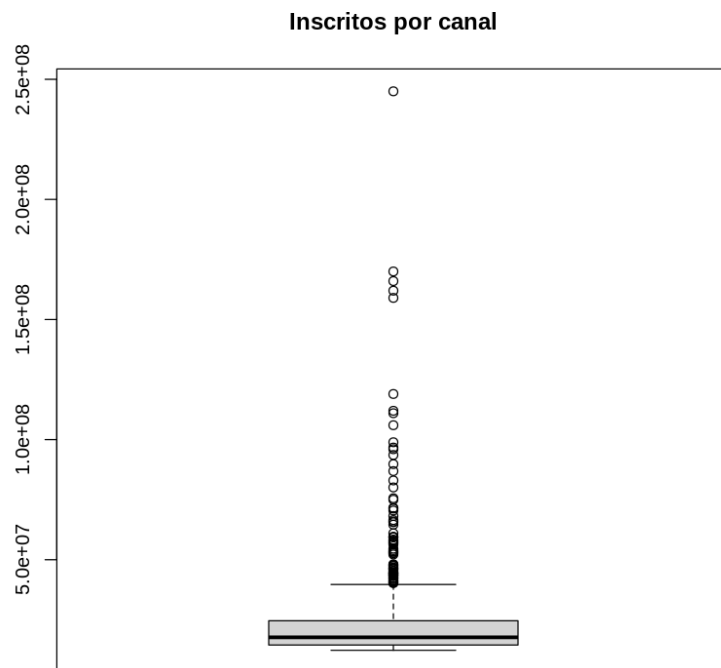


Figura 4: Box plot dos valores para inscritos nos canais.



Figura 5: Dados referentes à construção do box plot para os inscritos no canal.

- países que mais consomem o conteúdo do YouTube, o que tem de similar entre esses países e abstrair uma razão socio-econômica para o sucesso de um canal do YouTube, verificando se o país de origem do canal influencia no sucesso.
- análise de rendimento dos produtores de conteúdo, avaliando qual variável pode ser a maior responsável por essa renda, se o total de inscritos, o nicho de conteúdo do canal ou até o país de origem.
- descobrir conteúdo mais populares e frequência de postagem para determinadas audiências
- investigar como alguns conteúdos ganharam popularidade e correlacionar com período de maior audiência

Um título adequado ao trabalho seria "Análises e correlações estatísticas dos 995 canais com mais inscritos do YouTube em 2023".

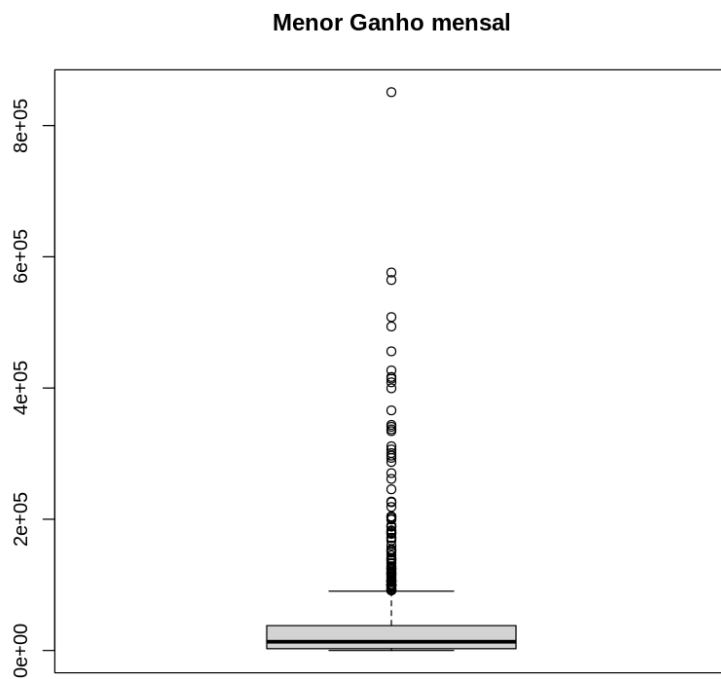


Figura 6: Box plot dos valores para o menor ganho mensal reportado pelo canal.

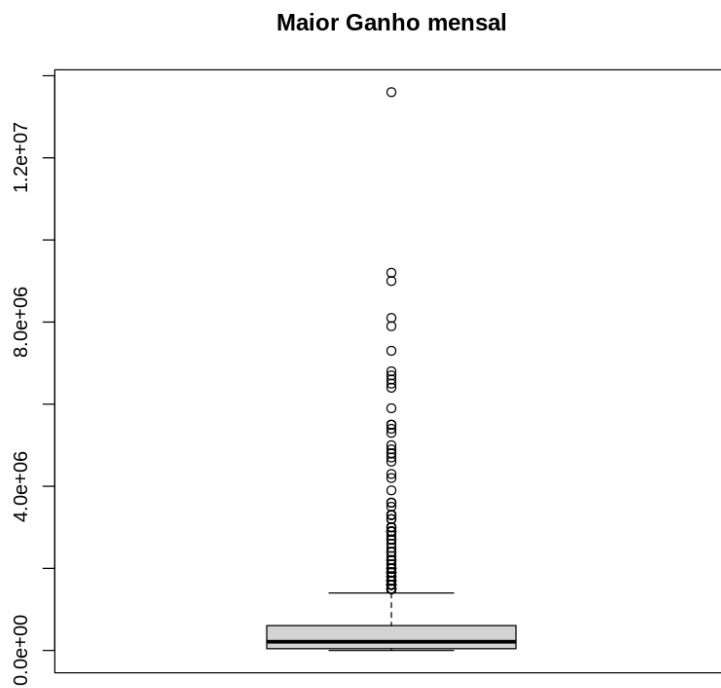


Figura 7: Box plot dos valores para o maior ganho mensal reportado pelo canal.

\$Variância	5163676228.16782
\$DesvioPadrao	71858.7240922619
\$Amplitude	850900
\$IQR	35200
\$MaxValue	850900
\$MinValue	0
\$moda	
\$media	36886.148281407
\$quartis	
0%:	0
25%:	2700
50%:	13300
75%:	37900
100%:	850900

Figura 8: Dados referentes à construção do box plot para o menor ganho mensal reportado pelo canal.

\$Variância	1319333598065.87
\$DesvioPadrao	1148622.4784784
\$Amplitude	13600000
\$IQR	563300
\$MaxValue	13600000
\$MinValue	0
\$moda	
\$media	589807.84758794
\$quartis	
0%:	0
25%:	43500
50%:	212700
75%:	606800
100%:	13600000

Figura 9: Dados referentes à construção do box plot para o maior ganho mensal reportado pelo canal.