MQAM - Grupo 15
Técnica de ANOVA

Mateus Caetano da Silva - 12543989
Victor Augusto Costa Monteiro - 8942937
Vinícius Henrique Crispim Rosa - 9395067

O Dataset

Título

Global YouTube Analytics 2023

Tamanho

995 linhas e 28 colunas

Origem

Kaggle.com

Descrição

Dados analíticos dos 995 canais com mais inscritos no YouTube em 2023. Coleção abrangente de informações relacionadas aos canais, como métricas de desempenho, renda mensal e renda anual, e informações relacionadas ao país de origem do canal, como demografia, coordenadas e estatísticas econômicas.

Variáveis do Dataset

28 variáveis no total

Qualitativa nominal	Qualitativa ordinal	Quantitativa discreta	Quantitativa contínua
8	5	7	8
Youtuber Categoria titulo país país_abrev tipo ano_criacao mes_criacao	rank rank_visualizacoes rank_por_pais rank_por_tipo	num_inscritos total_visualizacoes qtd_videos visualizacoes_30_dias inscricoes_30_dias populacao populacao_urbana	menor_ganho_mensal maior_ganho_mensal menor_ganho_anual maior_ganho_anual pop_curs_ens_sup taxa_desemprego latitude longitude

Sobre a Análise

Pergunta a ser respondida

Os principais canais do YouTube em diferentes países tiveram um crescimento, em número de inscritos, igual nos últimos 30 dias?

Varáveis utilizadas

inscritos_ultimos_30_dias
país

Estatística Descritiva das Variáveis

inscritos_ultimos_30_dias

- Variável quantitativa discreta
- Indica a quantidade de inscritos que o canal ganhou nos últimos 30 dias

pais

- Variável qualitativa nominal
- Indica o país de origem do canal
- 50 países únicos entre as 995 entradas na base

pais

Valores únicos

```
 [1] "India"          "United States"        "nan"
 [4] "Japan"          "Russia"               "South Korea"
 [7] "United Kingdom" "Canada"               "Brazil"
[10] "Argentina"      "Chile"                "Cuba"
[13] "El Salvador"    "Pakistan"             "Philippines"
[16] "Thailand"       "Colombia"             "Barbados"
[19] "Mexico"         "United Arab Emirates" "Spain"
[22] "Saudi Arabia"   "Indonesia"            "Turkey"
[25] "Venezuela"      "Kuwait"               "Jordan"
[28] "Netherlands"    "Singapore"            "Australia"
[31] "Italy"          "Germany"              "France"
[34] "Sweden"         "Afghanistan"          "Ukraine"
[37] "Latvia"         "Switzerland"          "Vietnam"
[40] "Malaysia"       "China"                "Iraq"
[43] "Egypt"          "Andorra"              "Ecuador"
[46] "Morocco"        "Peru"                 "Bangladesh"
[49] "Finland"        "Samoa"
```

inscritos _ultimos_30_dias

Estatística descritiva

Média	349079
Mínimo	1
1ºQuartil	100000
Mediana	200000
3ºQuartil	400000
Máximo	8000000
NA’s	337
Variância Amostral	3.77433e+11
Desvio Padrão	614355,4
Variância Populacional	3.76859e+11

Teste ANOVA

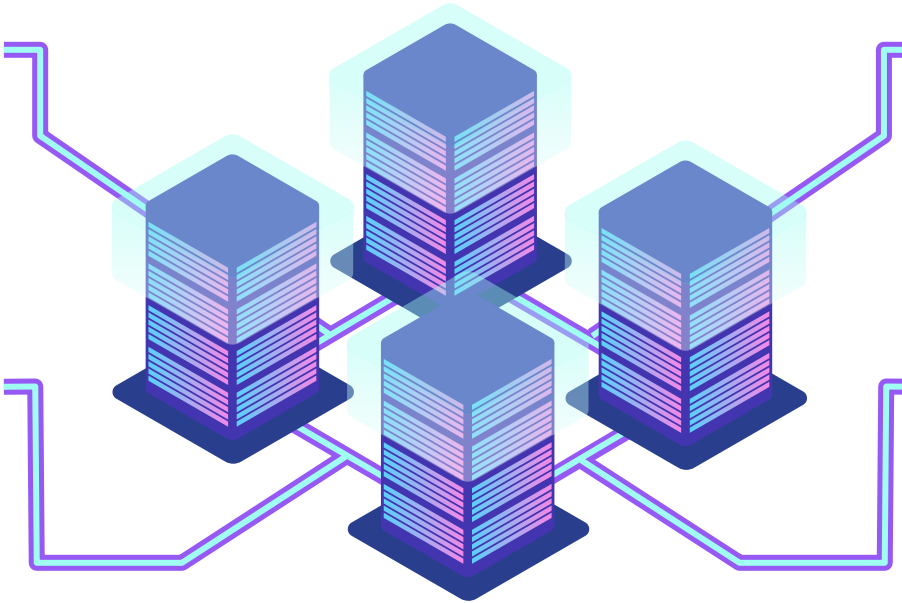
- Teste para comparação de médias de populações independentes
- Compara a variação devida ao tratamento (variação intra grupos) com a variação devida ao acaso (variação inter grupos)

Variável Quantitativa

Exige que a variável dependente seja quantitativa, contínua ou discreta

Homoscedasticidade

Exige que a variância inter grupos seja semelhante



Distribuição Normal

Exige que a distribuição da variável dependente seja normal dentro de cada grupo

Variável Qualitativa

As variáveis independentes devem se qualitativas, ordinal ou nominal

Avaliação de Suposições por Grau de Certeza

95%

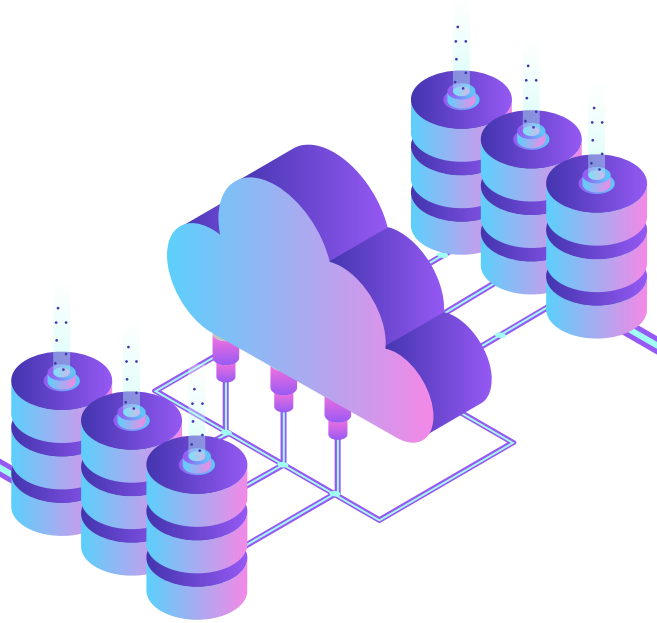
Adota-se como base para avaliar os resultados o grau de certeza de 95%

0.05

Portanto, o valor referência para comparação é 0.05, ou 5%

P > 0.05

Para valores maiores que 0.05, aceita-se a hipótese nula



Hipótese Nula

Também chamada de H0, geralmente é a hipótese que diz que não há diferença entre as amostras

Hipótese Alternativa

Hipótese que será considerada caso a hipótese nula seja rejeitada

P < 0.05

Para valores menores que 0.05, rejeita-se a hipótese nula

Hipóteses Estatísticas Definidas

H_0

Hipótese de igualdade

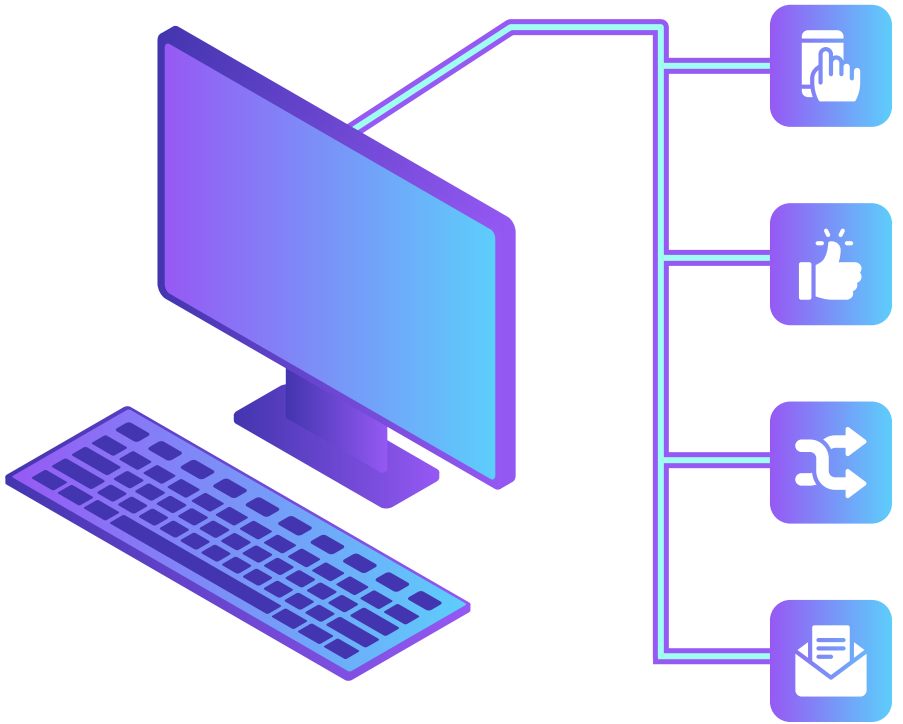
Os grupos apresentaram crescimento igual entre si dos inscritos nos últimos 30 dias.

H_1

Hipótese alternativa

Os grupos não apresentaram crescimento igual entre si dos inscritos nos últimos 30 dias.

Agrupamento dos Países para a Análise



Top 5 mais frequentes

Estados Unidos, Índia, Brasil , Reino Unido e México

Top 4 anglófonos

Estados Unidos, Índia, Reino Unido e Canadá

Top 5 asiáticos

Índia, Indonésia, Tailândia, Coreia do Sul e Filipinas

Top 5 europeus

Reino Unido, Espanha, Russia, Ucrânia e Alemanha

Estatística Descritiva Top 5 Países

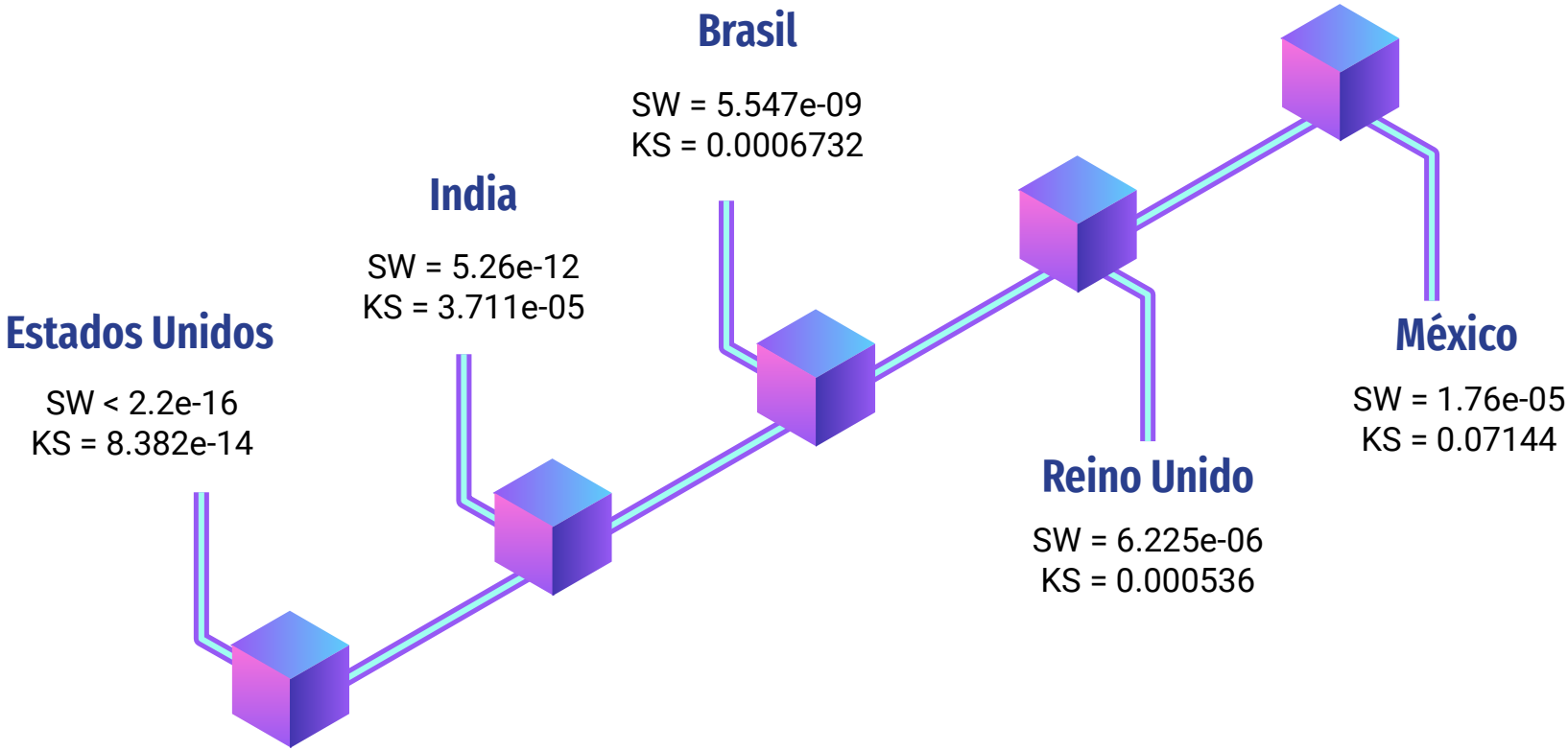
- Alguns dados importantes para a ANOVA, separados por grupos e pertinentes à variável inscritos_últimos_30_dias

	Estados Unidos	Índia	Brasil	Reino Unido	México
Média	**349822.8**	**392638.9**	**218181.8**	**177189.6**	**176470.6**
Mediana	**2e+05**	**3e+05**	**1e+05**	**1e+05**	**1e+05**
Variância	**6.053e+11**	**1.191e+11**	**6.278e+10**	**2.563e+10**	**1.816e+10**

Código Teste de Normalidade

```r
for(pais in paises_interesse){
  dados_filtrados <- subset(database, Country == pais)
  dados_filtrados <- na.omit(dados_filtrados)
  print(paste("teste de normalização do ano de criação dos canais do pais", pais))
  z_scores <- scale(dados_filtrados$subscribers_for_last_30_days)
  limite_z_score <- 4
  dados_sem_outliers <- dados_filtrados$subscribers_for_last_30_days[abs(z_scores) <= limite_z_score]


  media <- mean(dados_sem_outliers)
  desvioPadrao <- sd(dados_sem_outliers)
  cat("Media da distribuição -", media, "\n")
  cat("Desvio Padrão da distribuição -", desvioPadrao, "\n")
  shapiroTest <- shapiro.test(dados_sem_outliers)
  ksTest <- ks.test(dados_sem_outliers, pnorm, media, desvioPadrao)
  print(shapiroTest)
  print(ksTest)
}
```

Teste de Normalidade Top 5 Países Mais Frequentes



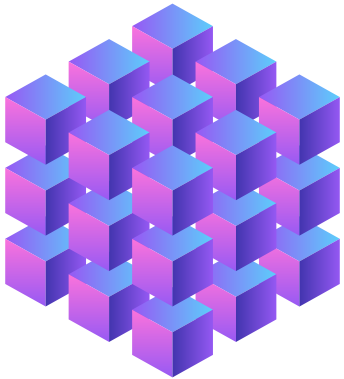
Código Teste de Homocedasticidade e ANOVA

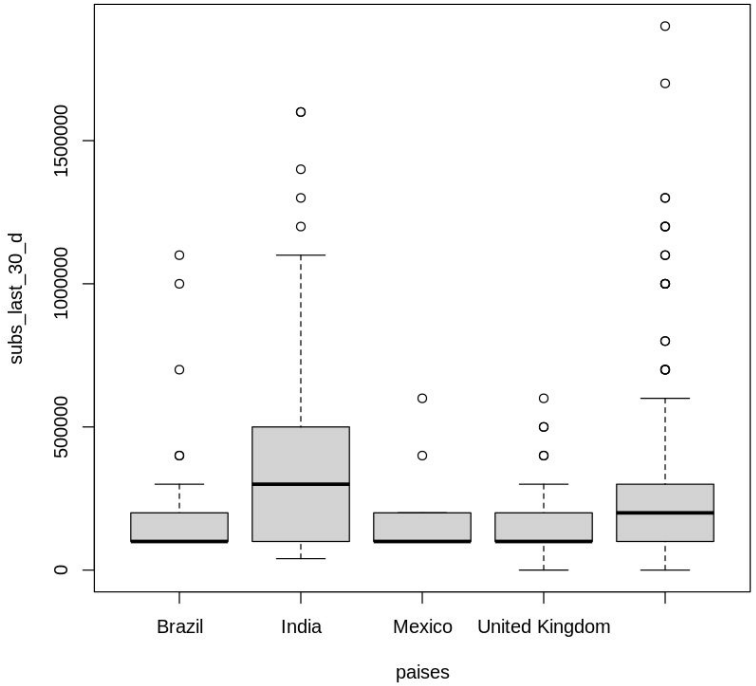
```
subs_last_30_d<-c()
paises<-c()
for(pais in paises_interesse){
  dados_filtrados <- subset(database, Country == pais)
  dados_filtrados <- na.omit(dados_filtrados)
  z_scores <- scale(dados_filtrados$subscribers_for_last_30_days)
  limite_z_score <- 4
  dados_sem_outliers <- dados_filtrados$subscribers_for_last_30_days[abs(z_scores) <= limite_z_score]

  subs_last_30_d <- c(subs_last_30_d, dados_sem_outliers)
  paises <- c(paises, rep(pais, length(dados_sem_outliers)))
}

boxplot(subs_last_30_d ~ paises)
leveneTest(y = subs_last_30_d, group = paises)
```

```
modelo_anova <- aov(subs_last_30_d ~ paises)
resultado_anova <- summary(modelo_anova)

print(resultado_anova)
```

Teste de Homoscedasticidade Top 5 países Mais Frequentes

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	4	3.355856	0.01015875
	410	NA	NA





Teste ANOVA

F Value

Teste F da análise ANOVA tem o valor de 5.449

P-Value

P-Value associado ao fator F com 4 graus de liberdade tem o valor de 0.000277

Rejeita Hipótese nula

O P-Value é menor que o valor fixado de 0.05

Top 5 Países

```
            Df    Sum Sq   Mean Sq F value   Pr(>F)
paises       4 1.924e+12 4.809e+11   5.449 0.000277 ***
Residuals  410 3.618e+13 8.825e+10
```

H_0: Existe uma relação no crescimento dos inscritos dos últimos 30 dias (igualdade)

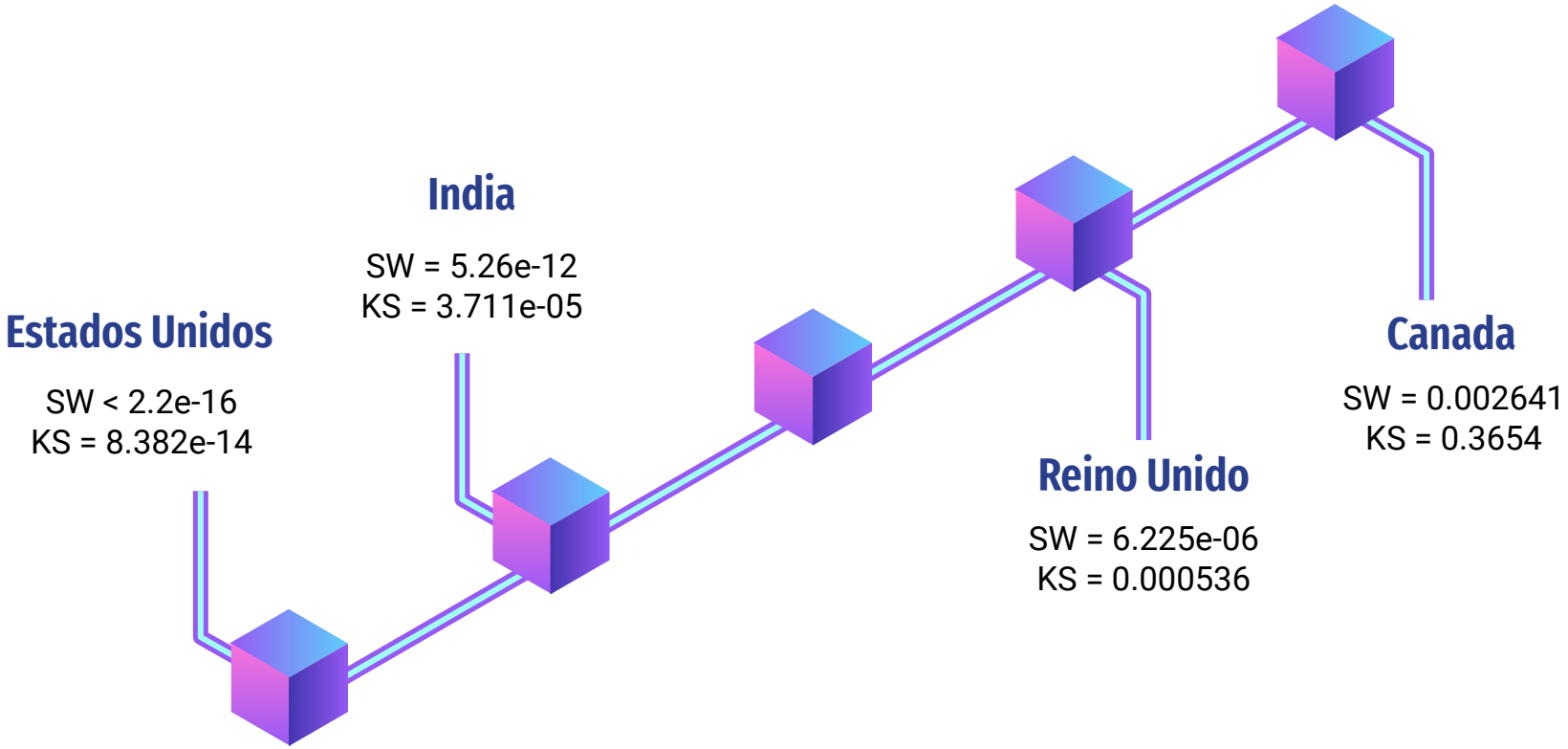
H_1: Não tem relação entre o crescimento de inscritos dos canais por país

Estatística Descritiva Top 4 Anglófonos

- Alguns dados importantes para a ANOVA, separados por grupos e pertinentes à variável inscritos_últimos_30_dias

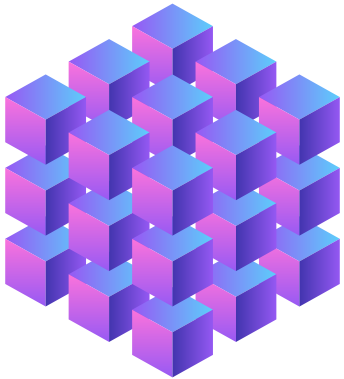
	Estados Unidos	Índia	Reino Unido	Canadá
Média	**349822.8**	**392638.9**	**177189.6**	**166666.7**
Mediana	**2e+05**	**3e+05**	**1e+05**	**1e+05**
Variância	**6.053e+11**	**1.191e+11**	**2.563e+10**	**1.467e+10**

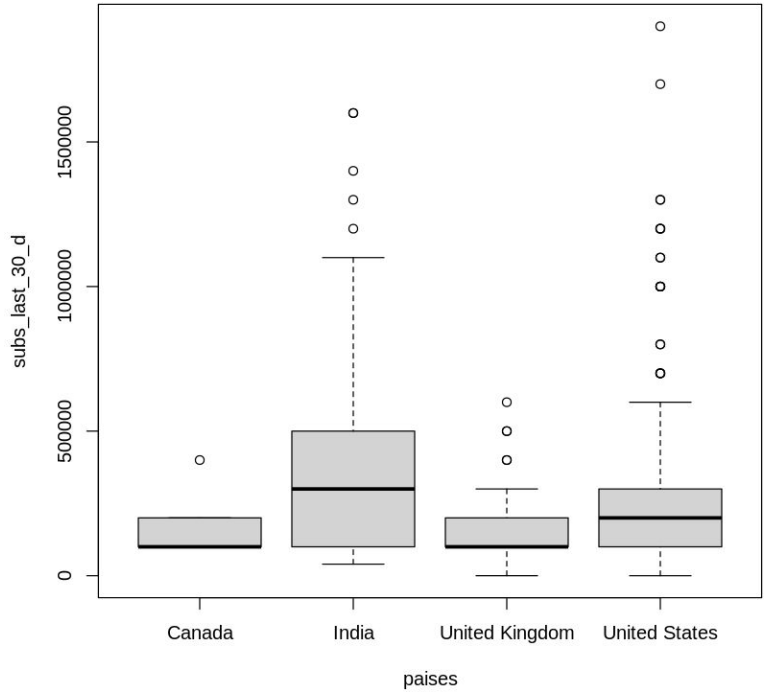
Teste de Normalidade Top 4 Anglofonos



Teste de Homoscedasticidade Top 4 países Anglofonos

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	3	2.922877	0.03390397
	367	NA	NA





Teste ANOVA

F Value

Teste F da análise ANOVA tem o valor de 5.486

P-Value

P-Value associado ao fator F com 4 graus de liberdade tem o valor de 0.00107

Rejeita Hipótese nula

O P-Value é menor que o valor fixado de 0.05

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
paises	3	1.523e+12	5.076e+11	5.486	0.00107	**
Residuals	367	3.396e+13	9.252e+10			

Top 4 Países Anglofonos

H_0: Existe uma relação no crescimento dos inscritos dos últimos 30 dias (igualdade)

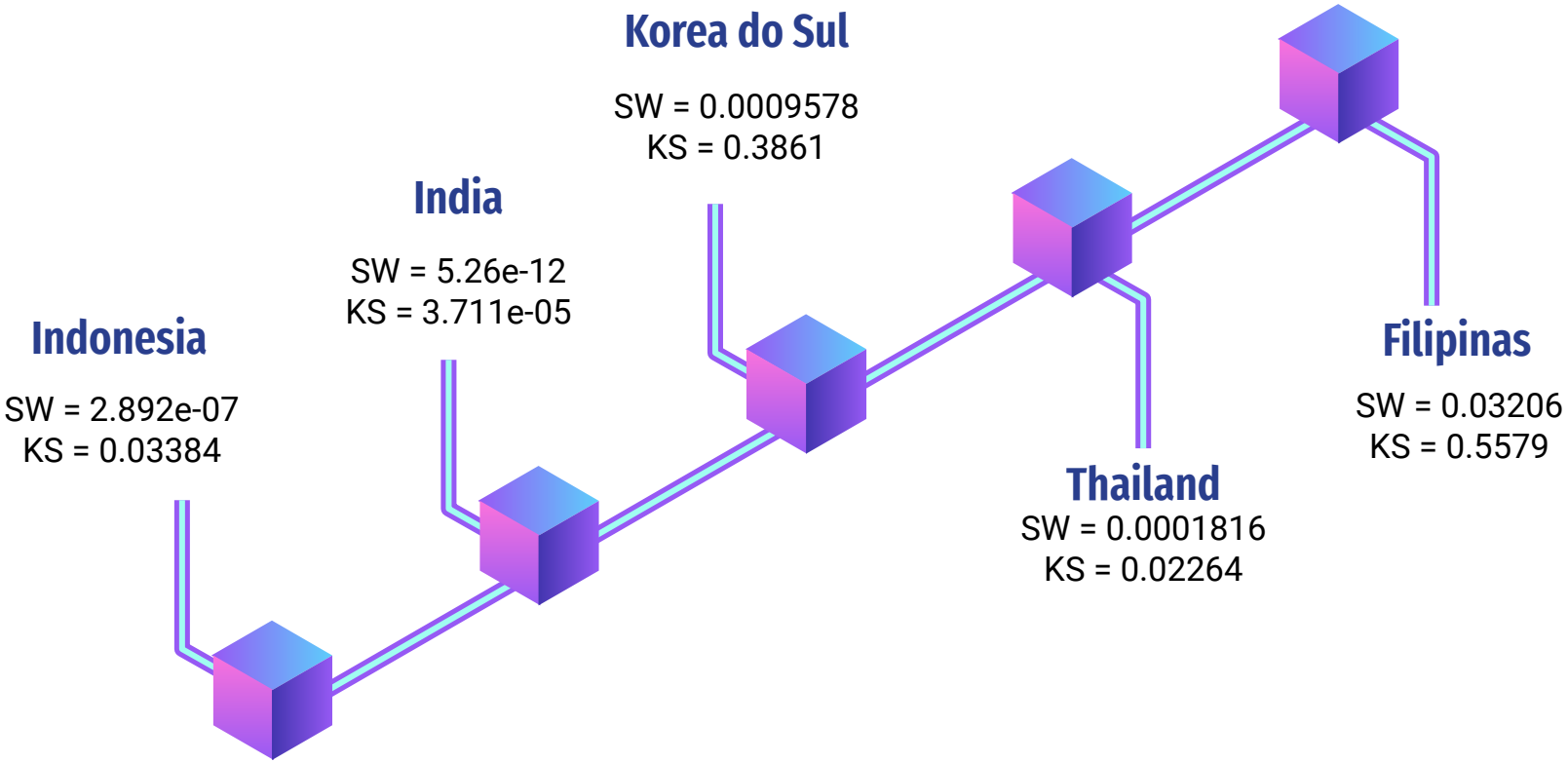
H_1: Não tem relação entre o crescimento de inscritos dos canais por país

Estatística Descritiva Top 5 Países Asiáticos

- Alguns dados importantes para a ANOVA, separados por grupos e pertinentes à variável inscritos_últimos_30_dias

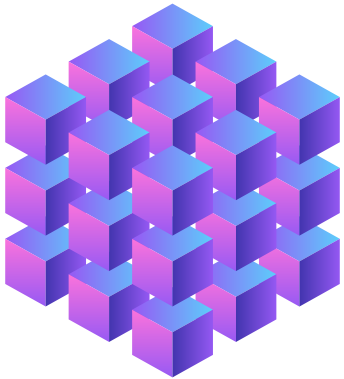
	Índia	Indonésia	Tailândia	Coreia do Sul	Filipinas
Média	**392638.9**	**620000.1**	**138461.5**	**788888.9**	**212500**
Mediana	**3e+05**	**1e+05**	**1e+05**	**4e+05**	**1.5e+05**
Variância	**1.191e+11**	**1.586e+15**	**4.231e+09**	**9.161e+11**	**2.125e+10**

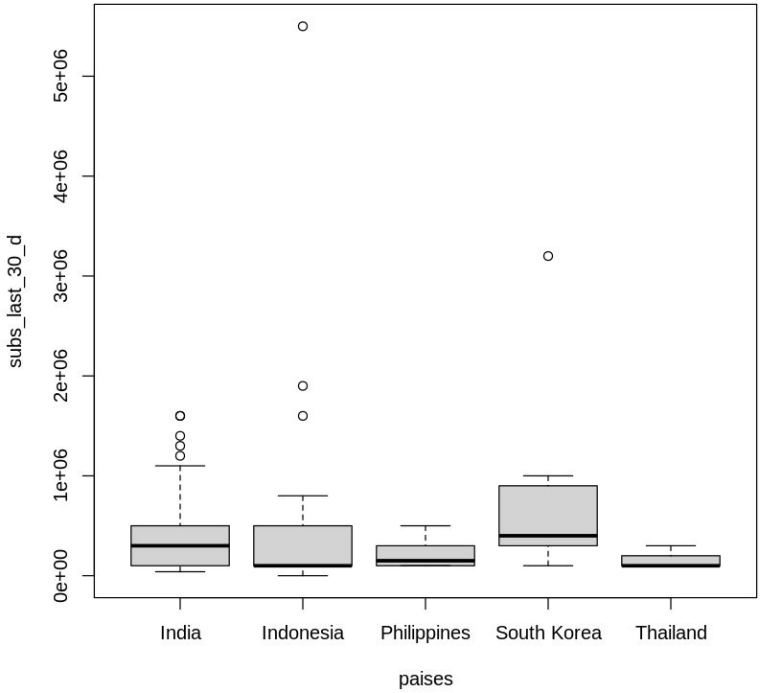
Teste de Normalidade Top 5 Países Asiáticos



Teste de Homoscedasticidade Top 5 países Asiáticos

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	4	3.367246	0.01090569
	188	NA	NA





Teste ANOVA

F Value

Teste F da análise ANOVA tem o valor de 3.202

P-Value

P-Value associado ao fator F com 4 graus de liberdade tem o valor de 0.0143

Rejeita Hipótese nula

O P-Value é menor que o valor fixado de 0.05

```
            Df    Sum Sq   Mean Sq F value Pr(>F)
paises       4 3.548e+12 8.871e+11   3.202 0.0143 *
Residuals  188 5.209e+13 2.771e+11
```

Top 5 Países Asiáticos

H_0: Existe uma relação no crescimento dos inscritos dos últimos 30 dias (igualdade)

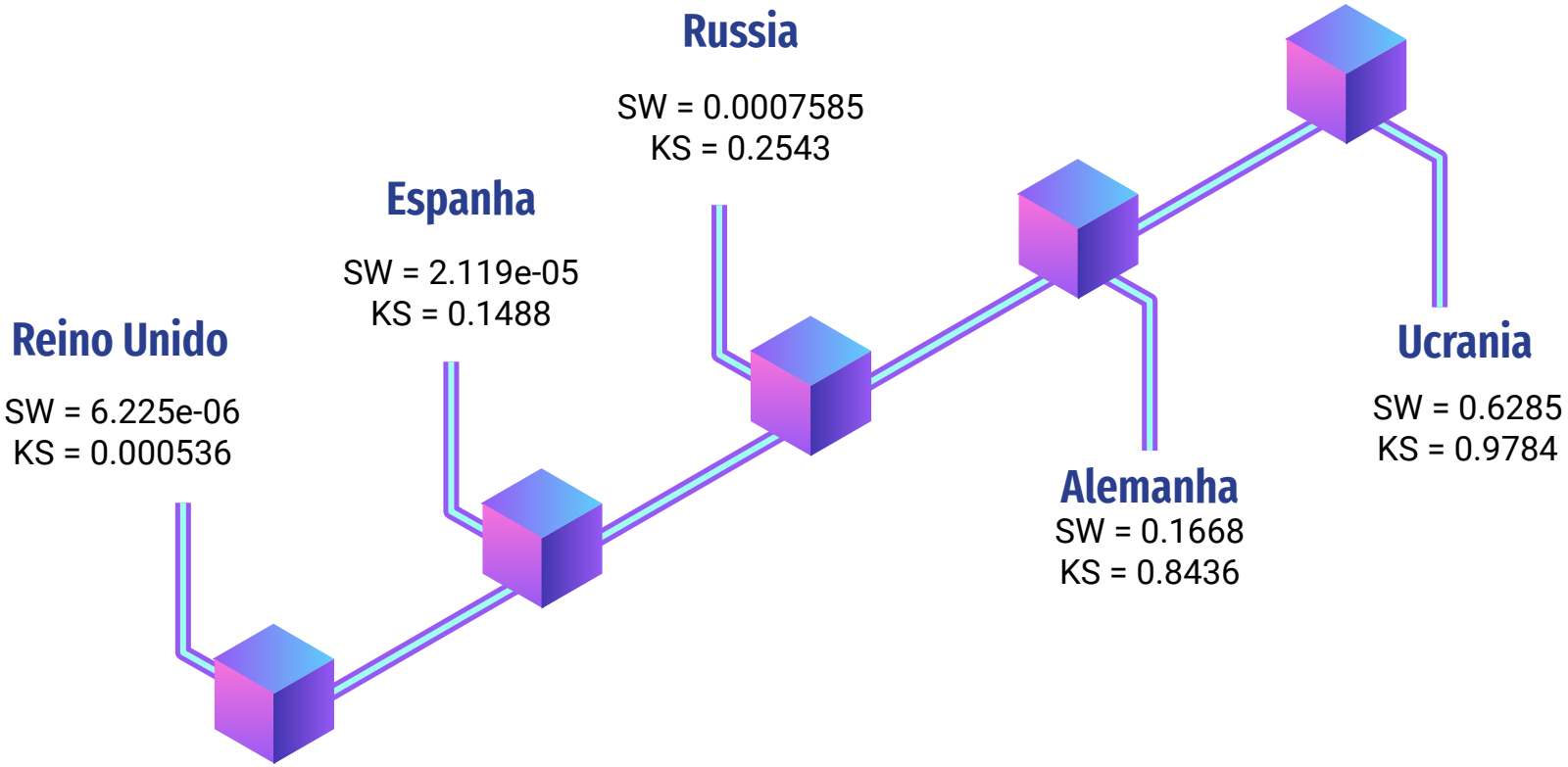
H_1: Não tem relação entre o crescimento de inscritos dos canais por país

Estatística Descritiva Top 5 Países Europeus

- Alguns dados importantes para a ANOVA, separados por grupos e pertinentes à variável inscritos_últimos_30_dias

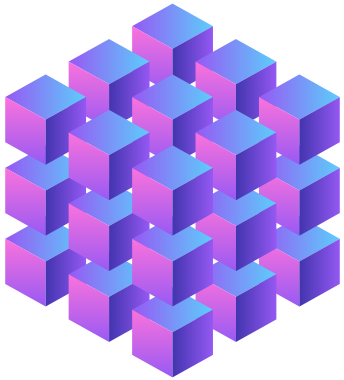
	Reino Unido	Espanha	Rússia	Ucrânia	Alemanha
Média	**177189.6**	**163691**	**392307.7**	**126002**	**450895**
Mediana	**1e+05**	**1e+05**	**2e+05**	**1e+05**	**4.5e+05**
Variância	**2.563e+10**	**4.853e+10**	**2.008e+11**	**8.379e+09**	**2.156e+11**

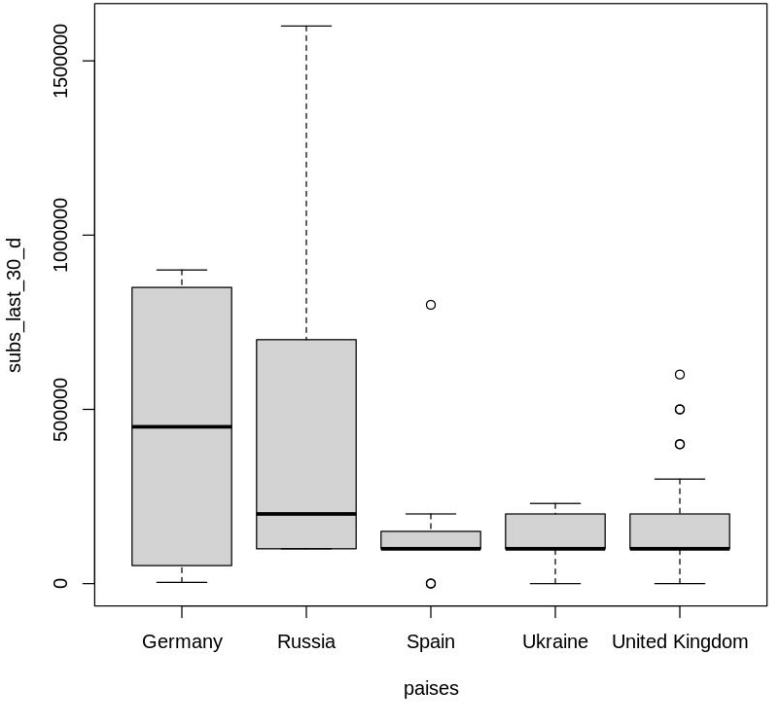
Teste de Normalidade Top 5 Países Europeus



Teste de Homoscedasticidade Top 5 países Europeus

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	4	3.031008	0.02492242
	55	NA	NA





Teste ANOVA

F Value

Teste F da análise ANOVA tem o valor de 2.308

P-Value

P-Value associado ao fator F com 4 graus de liberdade tem o valor de 0.0695

Não Rejeita Hipótese nula

O P-Value é maior que o valor fixado de 0.05

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
paises	4	7.117e+11	1.779e+11	2.308	0.0695
Residuals	55	4.241e+12	7.711e+10		

Top 5 Países Europeus

H_0: Existe uma relação no crescimento dos inscritos dos últimos 30 dias (igualdade)

H_1: Não tem relação entre o crescimento de inscritos dos canais por país

Obrigado!