Instituto Politécnico de Setúbal

Escola Superior de Tecnologia do Barreiro

Projeto de "Análise e Tratamento de Dados Multivariados"

Licenciatura em Bioinformática

Análise de Dados do COVID-19

janeiro de 2021

Grupo 1
Eduardo Palma 201900054
Nuno Melo 201700465
Ricardo Santos 201700524

Docente
Raquel Barreira

Índice

1 Introdução

O projeto consiste em analisar um conjunto de dados relativos à COVID-19, disponibilizado pela docente. Recorrendo ao R, realizamos uma análise exploratória aos dados, fazendo uma limpeza prévia e preparação dos dados para as etapas seguintes. Obtivemos estatísticas descritivas dos dados, análises de regressão linear multivariada, análise de clusters e análises de componentes principais.
Este subconjunto de dados que nos foram fornecidos encontram-se disponíveis no GitHub, dizem respeito a 31/07/2020 [1].

1.1 R e RStudio

R é uma linguagem com foco em análises estatísticas e gráficas. O R é um sistema de computação científica e estatística, programável e que permite o tratamento de vários tipos de dados.
O RStudio é um ambiente de interface gráfica para o R, ou melhor, um IDE (Integrated Development Environment). Atualmente o RStudio é considerado o melhor IDE para quem programa em R. Além de uma interface mais amigável, possui diversas funcionalidades que facilitam a aprendizagem e a produtividade.

1.2 O que é um dataframe?

Um dataframe em R é uma tabela em que cada coluna contém valores para uma variável e cada linha contém um conjunto de valores para cada coluna. Há duas funções importantes quando lidamos com dataframes: str(), que nos devolve a estrutura do dataframe, e summary(), que nos apresenta uma série de estatísticas sobre os dados.

2 Estatística Descritiva

É a parte da estatística que procura somente descrever e avaliar um certo grupo sem tirar conclusões sobre um grupo maior. Fornecido um certo conjunto de dados relativo a uma amostra de uma população, podemos sempre apresentá-los, ou organizá-los de duas formas distintas:

- Recorrendo a gráficos e/ou tabelas;
- Apresentando medidas de posição e/ou dispersão

2.1 Estatística Descritiva Univariada

A Estatística Univariada compreende todos os mecanismos de Estatística Descritiva que possibilitam a análise de cada variável separadamente e também inferências para determinada variável, podendo esta ser medida para uma ou mais amostras independentes. A palavra "univariada" subentende que há apenas uma variável dependente [2].

2.2 Análise Descritiva dos Dados

Em primeiro lugar, colocou-se o conjunto de dados num dataframe a que deu-se o nome *df*. Utilizou-se o comando *summary* para observar as variáveis, as númericas com os respetivos valores (Mínimo, Quartis, Média, Mediana, Máximo, NA's) e as não numéricas com o respetivo tamanho e classe. Na tabela 1, apresenta-se os valores dos NA's por coluna.

```
summary(df)
```

Variável	**Número de NA's**
total cases per million	1
total deaths per million	6
reproduction rate	5
total tests per thousand rate	25
positive rate	23
tests per case	24
stringency index	2
population density	2
median age	0
aged 70 older	0
gdp per capita	2
extreme poverty	21
cardiovasc death rate	1
diabetes prevalence	2
female smokers	9
male smokers	9
handwashing facilities	26
hospital beds per thousand	5
life expectancy	0
human development index	1
total cases	1
total deaths	6
total tests	25
population	0

Tabela 1: NA's existentes em cada variável

Utilizou-se o comando *str* para observar os tipos de todas as variáveis, que estão descritos na tabela 2.

```
str(df)
```

O comando usado para visualizar o valor do desvio-padrão é o *sd*, como mostra em seguida o desvio-padrão para o total de casos.

Numérica	Categórica
total cases per million	iso code
total deaths per million	continent
reproduction rate	
total tests per thousand rate	
positive rate	
tests per case	
stringency index	
population density	
median age	
aged 70 older	
gdp per capita	
extreme poverty	
cardiovasc death rate	
diabetes prevalence	
female smokers	
male smokers	
handwashing facilities	
hospital beds per thousand	
life expectancy	
human development index	
total cases	
total deaths	
total tests	
population	

Tabela 2: Tipos das variáveis

```
sd(total_casos)
```

O comando usado para visualizar a variação de dados observados de uma variável numérica por meio de quartis é o *boxplot*, como mostra em seguida para o total de casos do continente asiático.

```
boxplot(dfTotalAsia$total_cases)
```

Observando o *boxplot* do total de casos na Ásia, na figura 1 consegue-se perceber que existem alguns *outliers*, como é o caso da Índia, que é o *outlier* mais extremo. Em seguida removeu-se os *outliers* que estão a desvirtuar a visualização dos dados, na figura 2.

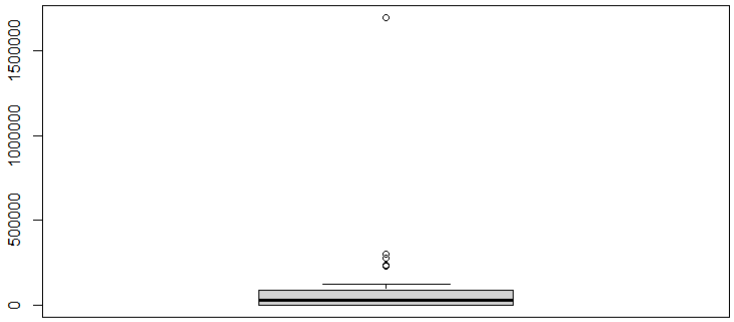


Figura 1: Boxplot para o total de casos na Ásia

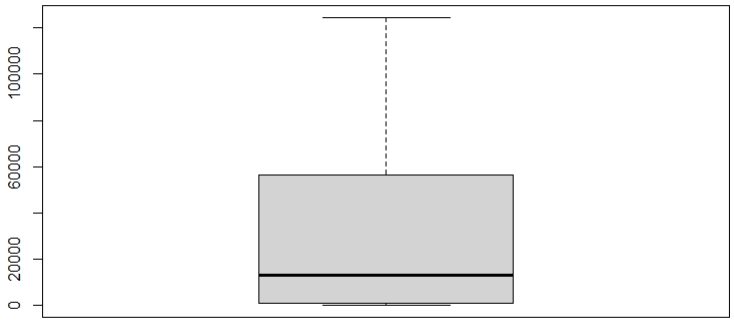


Figura 2: *Boxplot* para o total de casos na Ásia excluindo os *outliers*

Assim, com a remoção dos *outliers* visualizou-se que o 3ºQuartil é maior que o 1ºQuartil. Isto é, o número de casos está acima da mediana.
Na figura 3, observa-se o total de casos na Oceânia, onde verifica-se que o 3ºQuartil novamente é maior que o 1ºQuartil e por isso existem muitos casos acima da mediana.

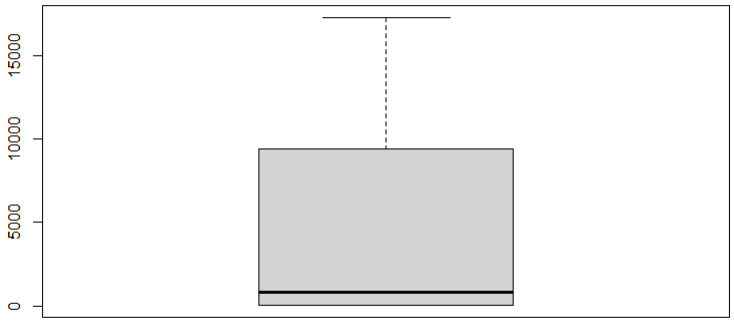


Figura 3: *Boxplot* para o total de casos na Oceânia

Na figura 4, verificou-se que existem *outliers* e procedeu-se à sua remoção, figura 5.

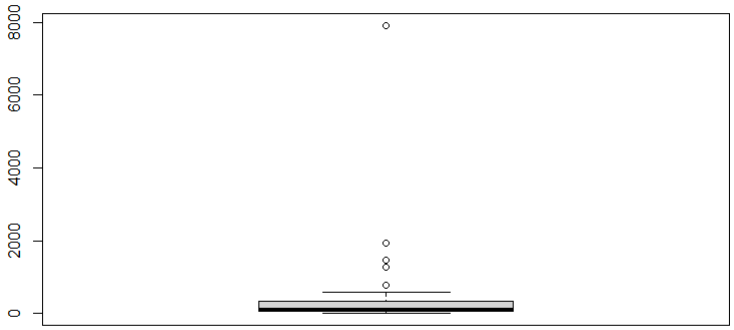


Figura 4: *Boxplot* para a densidade populacional na Ásia

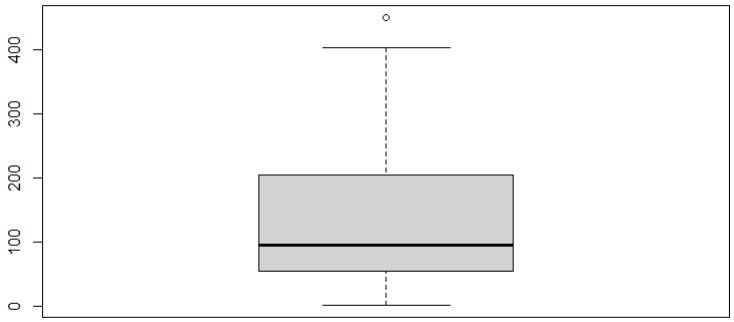


Figura 5: *Boxplot* para a densidade populacional na Ásia sem *outliers*

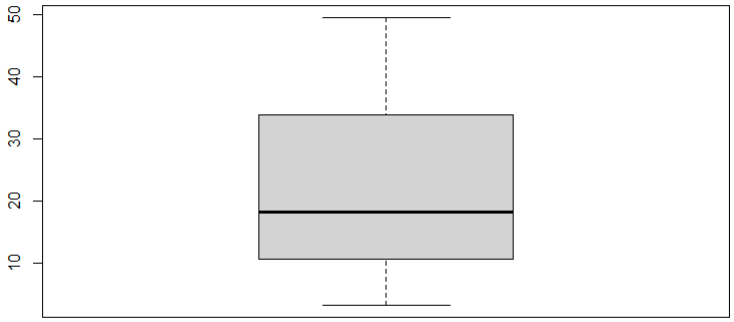


Figura 6: *Boxplot* para a densidade populacional na Oceânia

Na figura 5, após terem sido removidos vários *outliers*, percebeu-se que não faria sentido continuar a remover, pois se continuasse-mos a removê-los estaríamos a eliminar valores importantes para análise . Observou-se que a densidade populacional na Oceânia é muito inferior à densidade populacional na Ásia. No continente asiático a mediana encontra-se abaixo da mediana da Oceânia, mas continua a existir uma assimetria da distribuição, figura 6.

2.3 Análise de Regressão Linear Multivariada

2.3.1 Estatística Descritiva Bivariada

A Estatística Bivariada inclui métodos de análise de duas variáveis, podendo ser ou não estabelecida uma relação de causa/efeito entre elas. São exemplos típicos de métodos de análise bivariada o teste para a independência de duas variáveis (vulgarmente conhecido por teste do R-Quadrado) e o estudo da relação linear entre duas variáveis, quer através dos coeficientes de correlação linear de Pearson ou Spearman, quer do modelo clássico de regressão linear simples[4].

2.3.2 Regressão Linear

Num problema de regressão, à variável x chamamos variável independente ou controlada, e à variável Y chamamos variável dependente ou variável resposta. Neste caso, estamos interessados no estudo do comportamento de Y sendo conhecida x, i.e, no estudo da dependência de Y em relação a x, ou seja, na regressão de Y em x. Exemplos: Dependência da maturidade escolar em relação à idade.

2.3.3 Medidas de Correlação

As medidas de correlação quantificam a intensidade e a direção da associação entre duas variáveis. As medidas de correlação/associação usadas com mais frequência são:

- coeficiente de correlação de Pearson *(Rp)*;
- coeficiente V de Cramer *(V)*;
- coeficiente Phi(ϕ).

Utilizou-se o comando *ggplot*, figura 8, para visualizar o gráfico de total de casos por país, em seguida testou-se vários métodos para determinar a correlação associada às variáveis com o comando *cor*. Na regressão linear entre as variáveis população e total de casos utilizou-se o comando *lm* para obter o coeficiente de interseção e o comando *abline* para visualizar a reta de regressão. E ainda utilizou-se o comando *summary* para obter o R^2 de 0.46, logo existe alguma relação entre as variáveis.

```
cor(populacao,total_casos, method="pearson")
cor(populacao,total_casos, method="spearman")
cor(populacao,total_casos, method="kendall")
modelo_reg<-lm(populacao~total_casos)
modelo_reg
abline(modelo_reg)
```

Na figura 7, apresenta-se o cálculo dos coeficientes de correlação, pelo método de *Pearson* obteve-se um coeficiente de 0.68, o que indica que está correlacionado. Pelos métodos de *Spearman* e *Kendall*, 0.47 e 0.32 respetivamente, não apresentam grande correlação.

```
> cor(populacao,total_casos, method="pearson")
[1] 0.6831928
> cor(populacao,total_casos, method="spearman")
 #calculo do coeficiente de spearman
[1] 0.4772629
>
> cor(populacao,total_casos, method="kendall")
[1] 0.3289796
```

Figura 7: *Output* dos coeficientes das correlações

```
library(ggplot2)
dispersao4<-ggplot(data=df1,aes(x=population, y=total_cases)
    )+geom_point(aes(color=continent))
dispersao4
dispersao4+xlab("Populacao")+ylab("Total de Casos")+ggtitle(
    "Grafico de dispersao")
```

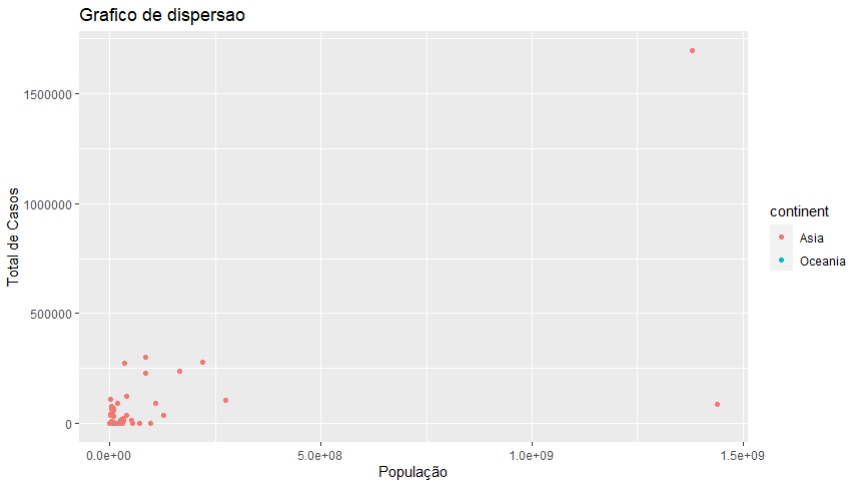


Figura 8: Gráfico de dispersão População vs. Total de Casos

Interpretando o gráfico, figura 8, conseguimos perceber que existe dois *outliers* na Ásia e que também existe mais infetados do que na Oceânia. Em seguida, figura 9 retirou-se os *outliers*.

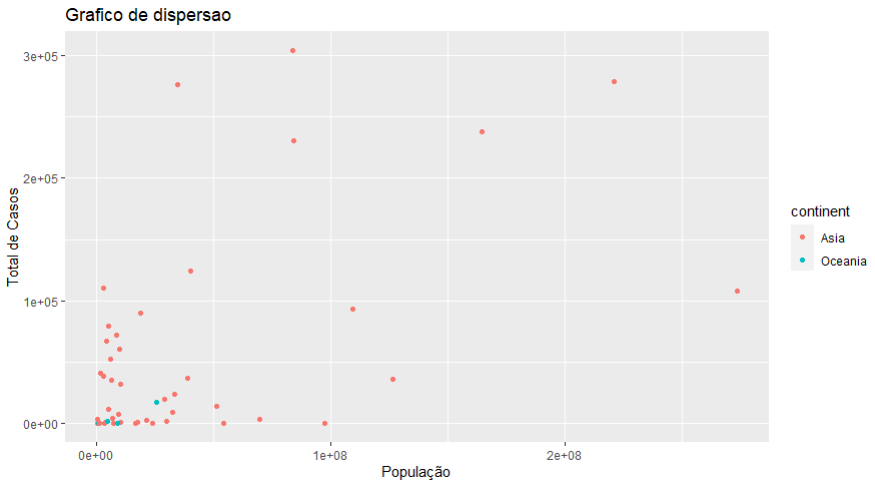


Figura 9: Gráfico de dispersão População vs. Total de Casos sem *outliers*

Com o comando *ggplot*, mostrado abaixo, analisou-se a relação entre o total de casos e as mulheres fumadoras.

```
1 dispersao2<-ggplot(data=df_semIndia,aes(x=female_smokers, y=
      total_cases))+geom_point(aes(color=continent))
2 dispersao2
3 dispersao2+xlab("Percentagem de Mulheres fumadoras")+ylab("
      Total de Casos")+ggtitle("Total de Casos vs. Mulheres
      Fumadoras")
```

Após a execução deste comando obteve-se o gráfico de dispersão, na figura 10.

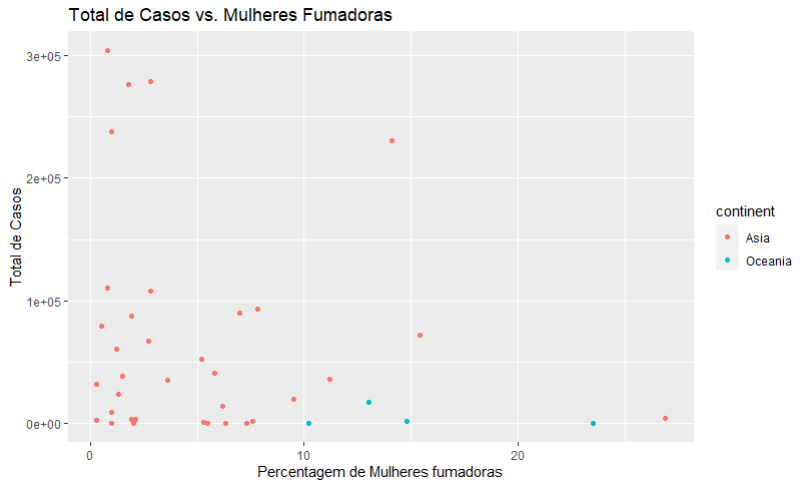


Figura 10: *ggplot* para Total de Casos vs. Percentagem de Mulheres Fumadoras

Observou-se na figura 10, que não existe uma relação entre o total de casos e a percentagem de mulheres fumadoras.
Utilizou-se o comando *ggplot* para criar um gráfico, figura 11, que mostra a relação entre a densidade populacional, total de casos e a mediana das idades, na Ásia.

```
1 dispersao_casospopulacaoidade<-ggplot(data=df_semIndiaChina,
    aes(population_density,total_cases))+geom_point(aes(color
    =median_age))+geom_text(aes(label=iso_code),hjust=1.2,
    vjust=1)+theme_bw()
2 dispersao_casospopulacaoidade+xlab("Densidade Populacional")
    +ylab("Total de Casos")+ggtitle("Grafico de dispersao")+
    labs(colour="Idade Media")
```

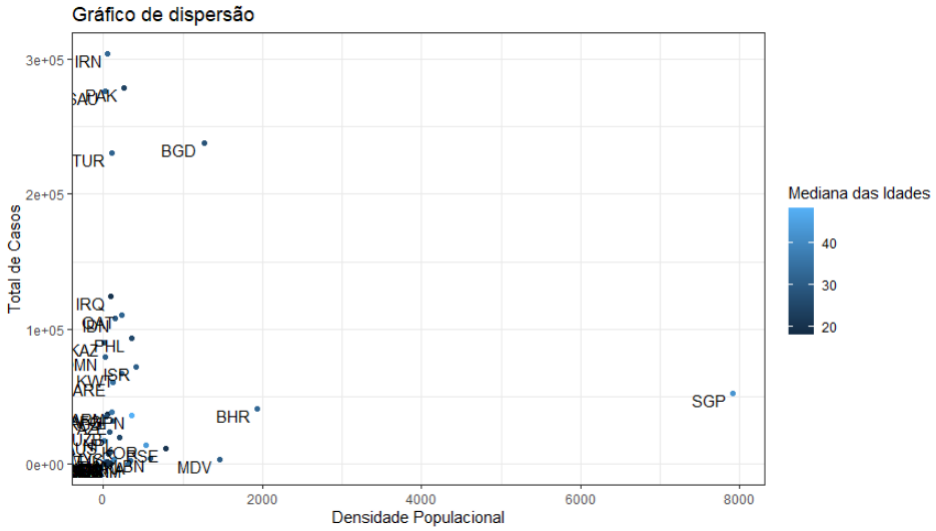
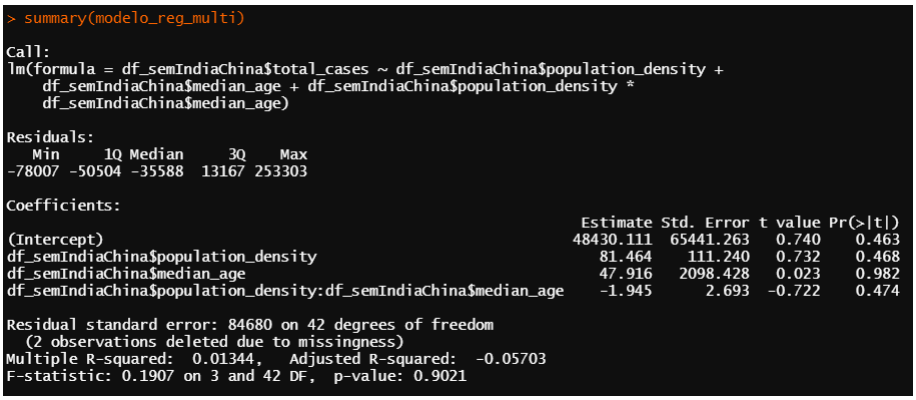


Figura 11: *ggplot* para Densidade Populacional vs. Total de Casos

Observou-se na figura 11 que não existe relação entre a densidade populacional e o total de casos, em grande parte dos países da Ásia, os valores são diretamente proporcionais.
Utilizou-se o comando *lm* para fazer a regressão linear entre o total de casos e a densidade populacional, anteriormente os *outliers* (Índia e China) foram removidos para uma melhor visualização.

```
modelo_reg_multi<-lm(df_semIndiaChina$total_cases~df_
    semIndiaChina$population_density+df_semIndiaChina$median_
    age+df_semIndiaChina$population_density*df_semIndiaChina$
    median_age)
```



```
> summary(modelo_reg_multi)

Call:
lm(formula = df_semIndiaChina$total_cases ~ df_semIndiaChina$population_density +
    df_semIndiaChina$median_age + df_semIndiaChina$population_density *
    df_semIndiaChina$median_age)

Residuals:
   Min     1Q Median     3Q    Max
-78007 -50504 -35588  13167 253303

Coefficients:
                                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                    48430.111  65441.263   0.740    0.463
df_semIndiaChina$population_density                               81.464    111.240   0.732    0.468
df_semIndiaChina$median_age                                       47.916   2098.428   0.023    0.982
df_semIndiaChina$population_density:df_semIndiaChina$median_age   -1.945      2.693  -0.722    0.474

Residual standard error: 84680 on 42 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.01344,   Adjusted R-squared:  -0.05703
F-statistic: 0.1907 on 3 and 42 DF,  p-value: 0.9021
```

Figura 12: *Output* da regressão linear

Como o R^2 é -0.05703 indica que não existe nenhuma relação entre as variáveis em estudo. De seguida irá-se procurar as variáveis que melhor se relacionam. Utilizou-se novamente o comando *lm* para fazer a regressão linear entre o desenvolvimento humano, o total de mortes e a percentagem de mortes por doenças cardiovasculares.

```
1 modelo_reg_multi<-lm(df_semIndiaChina$human_development_
      index~df_semIndiaChina$cardiovasc_death_rate+df_
      semIndiaChina$total_deaths+df_semIndiaChina$cardiovasc_
      death_rate*df_semIndiaChina$human_development_index)
2 modelo_reg_multi
3 summary(modelo_reg_multi)
```

Legenda:
CDR - cardiovasc death rate;
TD - total deaths;
HD - human development index

Logo, a equação do plano ajustada é dada por Y = 8.071e-01 -1.843e-03CDR + -1.966e-06TD + (2.320e-03HDCDR). Como mostra na figura 13 o R^2 é 0.8042 o que indica que existe relação entre as variáveis em estudo.

```
> summary(modelo_reg_multi)

Call:
lm(formula = df_semIndiaChina$human_development_index ~ df_semIndiaChina$cardio
vasc_death_rate +
    df_semIndiaChina$total_deaths + df_semIndiaChina$cardiovasc_death_rate *
    df_semIndiaChina$human_development_index)

Residuals:
      Min        1Q    Median        3Q       Max
-0.127639 -0.026649 -0.007202  0.023364  0.103929

Coefficients:

   Estimate Std. Error t value Pr(>|t|)
(Intercept)
  8.071e-01  1.995e-02  40.464  < 2e-16 ***
df_semIndiaChina$cardiovasc_death_rate
 -1.843e-03  1.506e-04 -12.242 9.30e-15 ***
df_semIndiaChina$total_deaths
 -1.966e-06  2.926e-06  -0.672    0.506
df_semIndiaChina$human_development_index:df_semIndiaChina$cardiovasc_death_rate
  2.320e-03  2.378e-04   9.756 6.75e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05496 on 38 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.8185,    Adjusted R-squared:  0.8042
F-statistic: 57.13 on 3 and 38 DF,  p-value: 3.773e-14
```

Figura 13: *Output* da regressão linear

Após diversas tentativas para escolher as variáveis que melhor se relacionavam, o melhor R-quadrado a que conseguiu-se chegar foi de 0.8042 utilizando 3 variáveis, Desenvolvimento Humano, Percentagem de mortes por doenças cardiovasculares e total de mortes. Para visualizar num gráfico de dispersão utilizou apenas 2 delas, Desenvolvimento Humano e Percentagem de mortes por doenças cardiovasculares, representado na figura 14. Desta forma, este foi o melhor gráfico que conseguiu-se obter. O gráfico mostra alguma relação entre elas, consegue-se observar que existe um grupo de pontos a meio do gráfico.

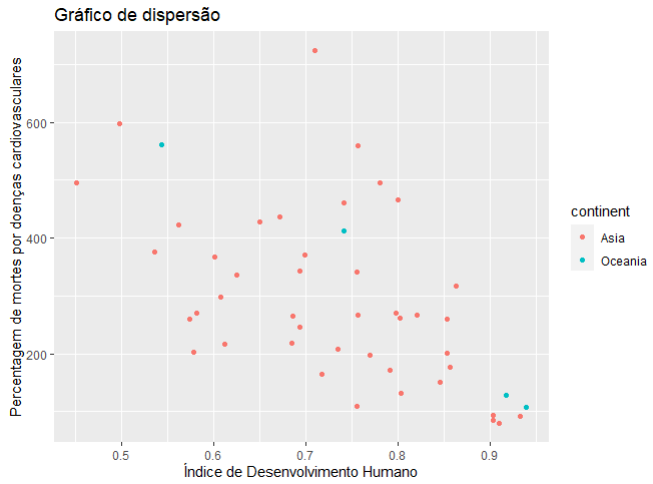


Figura 14: Diagrama de dispersão para as variáveis Índice Desenvolvimento Humano vs. Percentagem de mortes por doenças cardiovasculares

2.4 Análise de Clusters

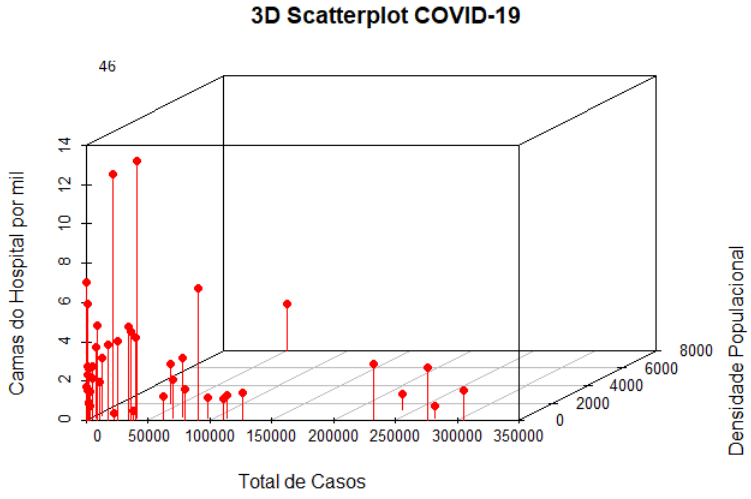


Figura 15: *Scatterplot* do Total de Casos vs. População vs. Camas do Hospital

A partir da figura 15, conclui-se que quanto menos casos existem menos camas por hospital são utilizadas e que o contrário também acontece, ainda se per-

cebe que em países com maior densidade populacional o total de casos é maior. Consegue-se fazer 2 *clusters*, um deles com a maior parte dos países que têm menos densidade populacional e menos total de casos, e no outro *cluster* com apenas 5 países que têm maior número de total de casos. Existe ainda um *outlier* a destacar que é a Singapura (39) que apresenta uma elevada densidade populacional comparada à de outros países.
Para analisarmos os *clusters* que existem entre as variáveis total de casos e a densidade populacional, começou-se porque colocar as duas variáveis num *dataframe*. Em seguida, utilizou-se o método *Elbow* e o método *Silhouette* para visualizar qual o número ótimo de *clusters*.

```
df_2variaveis<-data.frame(df_semIndiaChina$total_cases,df_
    semIndiaChina$population_density)
df_2variaveis

library ( factoextra )
fviz_nbclust(df_2variaveis , FUN = hcut , method = "wss", k.
    max = 10)
fviz_nbclust(df_2variaveis , FUN = hcut , method = "
    silhouette", k.max = 10)
```

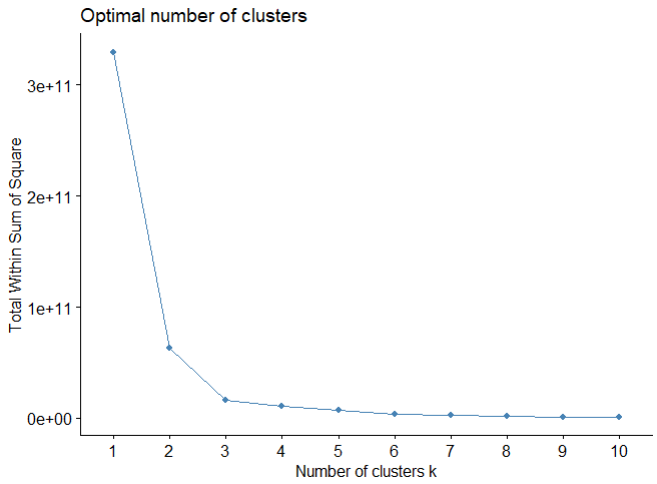


Figura 16: Método *Elbow*

Na figura 16, pelo método *Elbow* identifica-se 3 *clusters* como número ótimo.

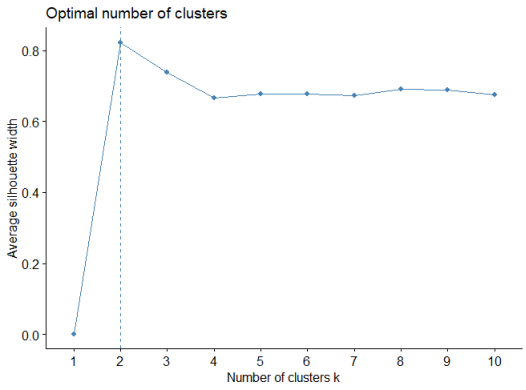


Figura 17: Método *Silhouette*

Na figura 17, pelo método *Silhouette* identica-se 2 *clusters* como número ótimo. Começou-se por usar o comando *scale* para colocar as variáveis na mesma escala e o comando *plot* para representar o dendograma.

```
scaled_df<-scale(df_2variaveis)
dist_df<-dist (scaled_df,method = "euclidean" )
dist_df
AC <- hclust (dist_df, method = "ward")
plot ( AC , hang = -1)
```

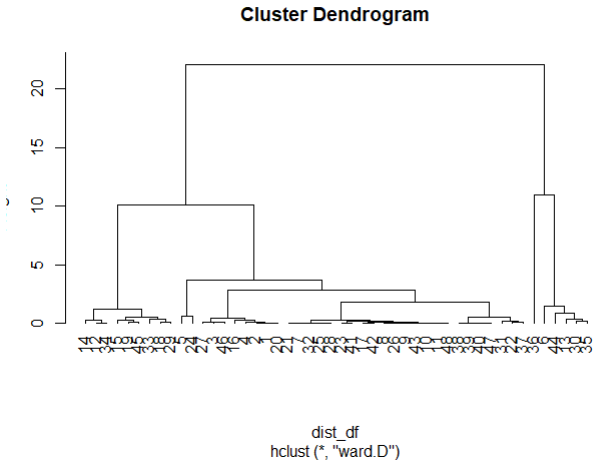


Figura 18: Dendograma de *Clusters*

Na figura 18 observa-se que existem 2 *clusters*.

2.5 Análise de Componentes Principais

Utilizou-se o comando *fviz pca* para representar o gráfico do PCA num *Screeplot*.

```
df_active<-df_semIndiaChina[,c(3,5,15,16,19,25)]
df_active
res.pca <- prcomp(df_active, scale = T)
res.pca
fviz_eig(res.pca)
grupos <- as.factor(df_semIndiaChina$continent)
fviz_pca_biplot(res.pca,
                label="var",
                habillage=grupos,
                legend.title = "Continente",
                title= "Grafico de PCA")
```

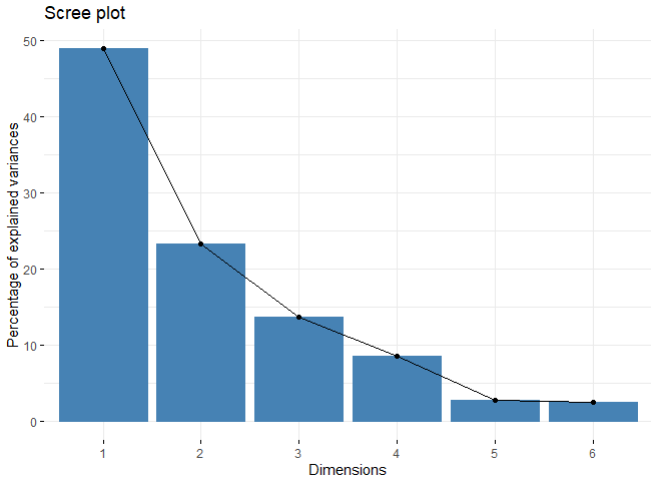


Figura 19: *Scree plot* das dimensões

Na figura 19, as duas primeiras componentes principais explicam 73% da variância total. Portanto retêm-se apenas as duas primeiras componentes principais.

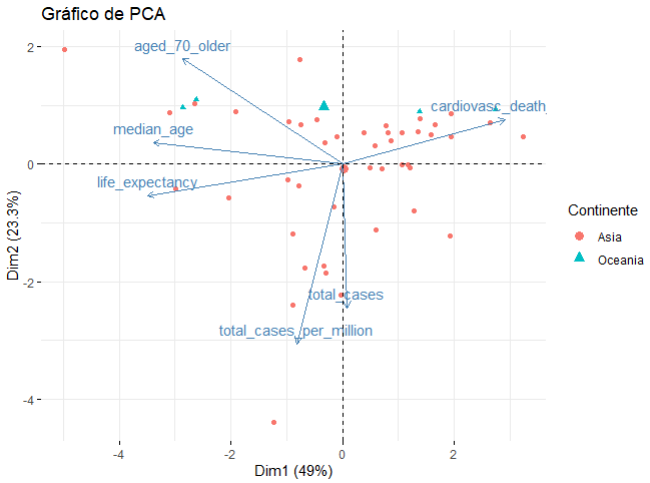


Figura 20: Gráfico de PCA

Na figura 20, representou-se graficamente considerando no eixo das abcissas a Dimensão 1 e para o eixo das ordenadas a Dimensão 2.

Figura 21: Gráfico de Correlação do PCA

Na figura 21, observa-se a correlação entre as dimensões.

3 Conclusões

Este trabalho teve como objetivo fazer uma análise de dados multivariados relacionados com os dados de um determinado dia da doença COVID-19.

Começou-se por analisar descritivamente os dados, de 26 variáveis escolheu-se as mais pertinentes e removeu-se os *outliers* para não desvirtuar a visualizacão dos gráficos. De seguida, realizou-se a regressão linear multivariada onde obteve-se valores de correlação para os diferentes métodos de correlação e em seguida fez-se vários gráficos de dispersão para visualizar a relação existente entre as variáveis.

Analisou-se a regressão linear entre o Desenvolvimento Humano, Percentagem de mortes por doenças cardiovasculares e total de mortes, onde obteve-se um valor de R^2=0.8042 que mostra que existe uma boa relação entre as variáveis.

Na análise de *clusters* utilizou-se várias métodos para determinar quantos *clusters* existem. A realçar os métodos de Elbow e Silhouette em que se obteve diferentes conclusões. Pelo método de Elbow identificou-se 3 clusters como número ótimo, enquanto que, pelo método de Silhouette apenas 2 clusters. Conclui-se então que existe 2 clusters, pois o método Silhouette é mais preciso, a média da medida Silhouette para todos os elementos para os diferentes números de clusters e o número ótimo é o que maximiza essa média.

Na análise de componentes principais, utilizou-se 6 dimensões onde as duas primeiras se destacam tendo 73% da variância total. Portanto retêm-se apenas as duas primeiras componentes principais.

Logo após, representou-se graficamente as relações entre as componentes principais, que se conclui que na Ásia existe mais mortes por doenças cardiovasculares e que na Oceânia a população é mais envelhecida. Com este trabalho adquirimos conhecimentos na linguagem de programação R que será muito importante no futuro.

Referências

[1] http://www.luisam.utad.pt/EstatDescritiva20probabil.pdf, Consultado a 04/01/21

[2] https://alexandreramos.blogs.sapo.pt/7901.html, Consultado a 04/01/21

[3] https://moodle.ips.pt/2021/course/view.php?id=2227 Consultado a 26/01/21