Segundo Exercício para a G1

Taxas de Desemprego (OECD)

Faça o download dos dados de desemprego mantidos pela OECD (donwload full indicator data (.csv) em https://data.oecd.org/unemp/unemployment-rate.htm) e copie o arquivo .csv para um subdiretório `data/` abaixo do diretório do script.

Leitura Inicial dos Dados

Crie um script que leia os dados do arquivo CSV, se certificando que os valores alfanuméricos sejam tratados como `string` e não `factor`.

Examinando os Dados

Examine os dados lidos e responda:

a) Quantos registros e atributos há?

b) Quantos países/regiões estão incluídos nos dados?

Extraindo Subconjuntos de Dados

Extraia o subconjunto dos dados relacionados apenas aos indicadores anuais (excluindo os indicadores trimestrais e mensais) e responda as seguintes perguntas. (Obs: Certifique-se de que a coluna relacionada às datas seja da class `numeric`.)

a) Quantos registros permanecem?

b) De quantos anos há dados?

c) Qual é o período de tempo (primeiro ano e último ano)?

d) Qual é o primeiro e o último ano de que se tem dados de cada país? Elabore uma tabela com as colunas `location`, `first.year`, `last.year`, ordenada por `first.year` e depois por `location`. Para exibir a tabela, utilize a função `kable` do pacote `knitr`.

e) Desde o início da coleta dos dados em cada país, há algum ano de que não se tenha dados?

f) Desde qual ano a OECD tem dados de desemprego do Brasil?

g) De quantos e quais anos se tem dados de desemprego do Brasil?

Estatísticas Descritivas Básicas

a) Qual é o país/ano com a menor taxa de desemprego? E o com a maior taxa de desemprego?

b) Quais as médias de taxa de desemprego de cada país?

c) Qual é o país com a menor taxa média? E o país com a maior taxa média?

Distribuição das Taxas de Desemprego por País

a) Elabore um gráfico de boxplots com os valores de desemprego por país/região.

b) Qual é a amplitude (máx-mín) de variação da taxa de desemprego de cada país/região? Ordene por amplitude, da maior para a menor.

c) Qual é o país com maior variação (máx-min)?

Gráficos

Evolução das Taxas de Desemprego

Faça um gráfico de linhas representando a evolução das taxas de desemprego. As linhas correspondentes a um país/região devem ser representadas em cor cinza (`gray75`), e a média da OECD em vermelho (`darkred`).

Países com maior amplitude de variação

Elabore um gráfico de linhas com os 5 países com maior amplitude, cada qual com uma cor e com o nome do país à direita de cada linha.

Países com Menor Variação

Elabore um gráfico de linhas com os 5 países com menor variação (excetuando-se a média da OECD), cada qual com uma cor e com o nome do país à direita de cada linha.

Relação entre Taxas de Desemprego e Produto Interno Bruto

Busque no site da OECD as taxas de Produto Interno Bruto (GDP: Gross Domestic Product) de diversos países em: https://data.oecd.org/gdp/gross-domestic-product-gdp.htm.

Gráfico de Dispersão

a) Elabore um gráfico de dispersão relacionando as taxas de desemprego e o produto interno bruto. Represente os dados sobre o Brasil em uma cor diferente dos demais países.

b) Calcule a correlação linear entre os dois conjuntos de dados. Avalie se o grau de correlação é forte, moderado, fraco ou nenhum.

c) Elabore um novo gráfico de dispersão, somente com os países do G7: https://en.wikipedia.org/wiki/Group_of_Seven_(G7).

d) Calcule e exiba no gráfico a linha de regressão linear para os dados do G7. Qual é o root-mean-square error? E o R-squared? O que esses resultados significam?

e) Calcule a correlação linear entre os dois conjuntos de dados do G7. Avalie se o grau de correlação é forte, moderado, fraco ou nenhum.

Clusterização das Taxas de Desemprego

Hierarchical Clustering

Faça uma análise de clusters considerando os dados de 2010 a 2014, buscando identificar quais países/regiões se assemelham com relação às taxas de desemprego. Primeiro, utilizem o *hierarchical clustering* para ajudar a decidir (aproximadamente) o número de clusters para aplicar o *k-means*.

Dica: utilize a função `reshape` para construir a matriz de valores que será utilizada para calcular a matriz de distâncias.

Clusterização Utilizando o K-Means

Utilizem o k-means para clusterizar os dados de 2010 a 2014 Avaliem a qualidade da clusterização utilizando a silhouette. Façam a análise com três números diferentes de clusters, para obter uma melhor avaliação com silhouette.

Qual foi a melhor clusterização que você obteve, e por quê?