Análise de Dados I - Fundamentos de estatística

Lab 1 - Estatística descritiva

A base de dados a ser utilizada, adults.csv, é um extrato do censo estadunidense de 1994. A base é composta por 15 atributos:

age (idade): continuo.
workclass (classe de trabalho): Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous. (meio complicado de explicar o significado. Mais em http://web.cs.wpi.edu/~cs4341/C00/Projects/fnlwgt)
education (nível de educação): Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num (código para o nível educacional): continuous.
marital-status (estado civil): Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation (profissão): Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race (etnia): White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex (gênero): Female, Male.
capital-gain (ganho de capital): continuous.
capital-loss (perda de capital): continuous.
hours-per-week (horas de trabalho por semana): continuous.
native-country (país de origem): United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
Class (Classe): - se ganhou mais ou menos que US$ 50.000,00
Wages (salário): o salário percebido naquele ano.

De posse desses dados, espera-se que você sumarize (utilizando todos os métodos estudados) os valores e suas dispersões (claro, aqueles que possam ser sumariados). Identifique qual métrica você considera ser, com base no assunto estudado, a mais adequada para a sumariação do valor e dispersão de cada atributo.

Em seguida, identifique qual profissão, gênero, cor, país de origem e nível educacional possuem maior **média** salarial. Já conhece o plot? Trace gráficos para visualizar melhor o resultado! Muda algo se considerarmos apenas quem ganha menos de 20.000 ou mais do que 80.000?

Gere, em uma mesma imagem, os boxplots de todos os atributos numéricos. Por fim, calcule a proporção entre a média salarial e quantidade média de horas de trabalho.

Esteja atento a todas as observações de entrega definidas no piazza!