Mineração dos Dados do Censo 1994 EUA

Higor H. P. Nucci - Univesidade Federal de Mato Grosso do Sul

Abril 2019

1 Introdução

O objetivo desta atividade prática é exercitar a análise exploratória de um conjunto de dados. Para isso, espera-se que a(o) estudante consiga, utilizando observações estatísticas e apresentações gráficas, responder questionamentos sobre as relações em um conjunto de dados.

2 Atividade

2.1 Observe a relação existente entre gênero e estado civil. Descreva suas observações. Lembre-se que a base de dados Adult é uma amostra com apenas maiores de idade ($age > 16$) e que trabalham ($hours-per-week > 0$)

A Figura 1 mostra a quantidade de homens e mulheres para cada um dos estados matrimoniais. O conjunto de dados apresentado mostra que existe uma quantidade discrepante de homens militares que são casados com civis. Em contrapartida, não existem muitos casos de militares casados com outros militares. Os outros estados matrimoniais mostram uma relação relativamente parecida.

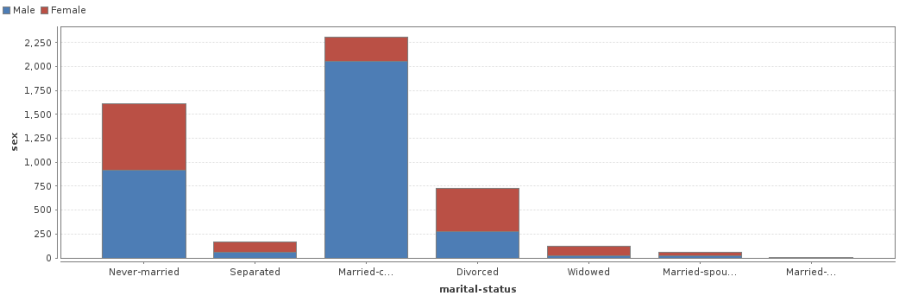

Figure 1: Comparação do estado civil de acordo com o sexo.

2.2 Investigue a relação entre ganho de capital e ganho anual: qual é o ganho médio de capital de acordo com o ganho anual? Observe também o histograma do ganho de capital. O que podemos inferir sobre essas observações?

Na Figura 2, que mostra o histograma do ganho de capital, pode-se observar que a grande maioria não declarou ou não recebe ganho algum. Dessa forma, fica difícil observar qual o motivo dessa ocorrência ou fazer relações com outros atibutos do conjunto de dados.

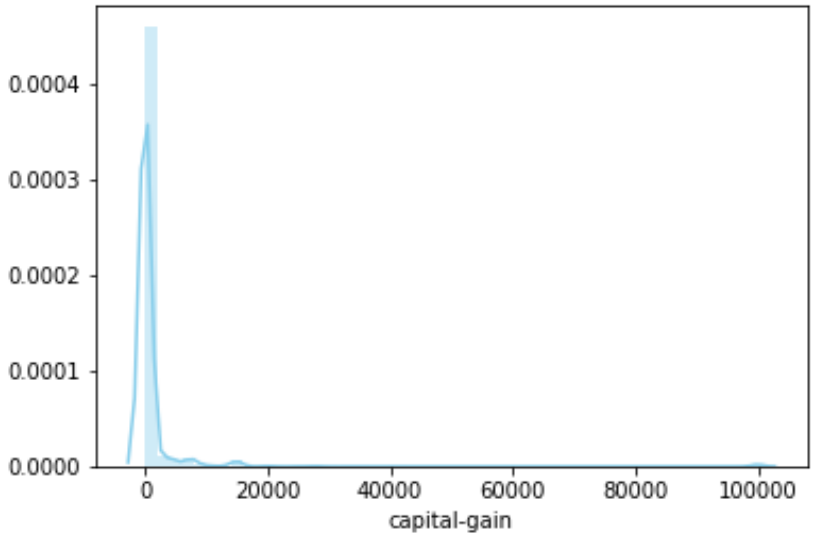

Figure 2: Histograma do ganho de capital.

Portanto, a retirada das amostras que não possuem ganho foi feita. Logo, a relação entre o ganho anual e o ganho de capital declarado pode ser observado na Figura 3. Pessoas que tem um ganho capital acima de $ 5000 tendem a receber mais que 50 mil anual. Isto posto, pode-se chegar a conclusão que o ganho anual está ligado ao ganho de capital declarado.

Os gráficos de violino permitem visualizar a distribuição de uma variável numérica para um ou vários grupos. É muito próximo de um boxplot, mas permite uma compreensão mais profunda da distribuição. Os violinos são particularmente adaptados quando a quantidade de dados é enorme e as observações individuais são impossíveis.

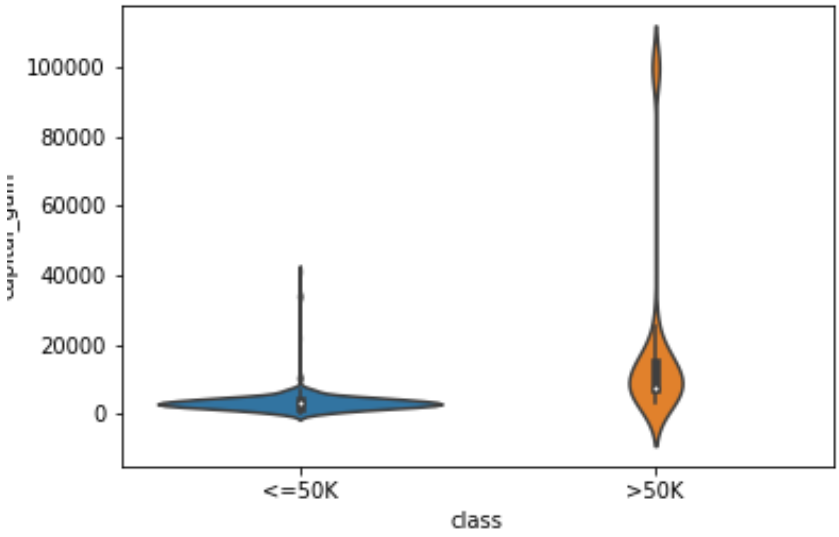

Figure 3: Ganho de capital por ganho anual.

2.3 Quais são as profissões em que mais pessoas tem ganho anual $ 50K? Quais profissões tem mais de $ 50K?

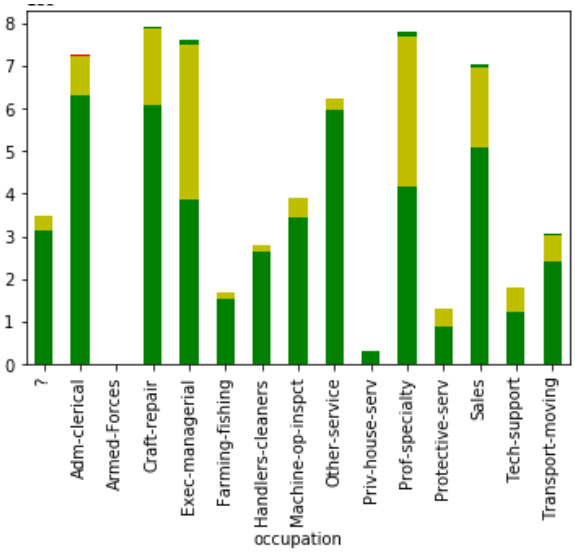

Figure 4: Ganho anual por profissões.

A Figura 4 mostra o ganho de capital anual por ocupação/profissão. A cor verde representa o ganho menor que 50 mil dólares anuais e a cor amarela representa o ganho maior que 50 mil dólares anuais.

O barplot foi utilizado pois pode exibir valores para vários níveis de agrupamento. Em vez de colocar as barras uma ao lado da outra, é possível empilhá-las, resultando em um barplot empilhado. De acordo com a Figura 4, as profissões mais bem remuneradas são de as de Executivo Gerencial e Prof Especialista e as que menos remuneram são Forças Armadas e Segurança Residencial Privado.

No entanto, É possível observar que em todas as profissões, a maioria dos entrevistados possuem ganhos anuais abaixo de U$ 50.000,00. Em contrapartida todas as profissões analisadas existem profissionais que ganham acima de U$ 50.000,00. Dentre as profissões analisadas o Executivo Gerencial é o profissional que possui a maior parte dos indivíduos com salários acima de U$ 50.000,00.

2.4 Quem trabalha mais horas em média: o marido, a esposa, pessoas com filhos, etc? Este comportamento varia dependendo da idade?

A Figura 5 mostra a média das horas trabalhas por tipo de relacionamento familiar. Ao observar a figura, nota-se que Maridos (*Husbands*) trabalham mais horas semanais. Por outro lado, pessoas que são filho único (*Own-child*) são as pessoas que trabalham menos horas.

Figure 5: Comparação das horas médias trabalhadas por tipo de relacionamento familiar.

No entanto, quando os mesmos dados são analisados levando em consideração a idade, pode-se notar pessoas que são filhos únicos são os que tem menor idade (¡=25 anos). Indivíduas classificados como maridos tendem a ser mais velhos e trabalharem mais.

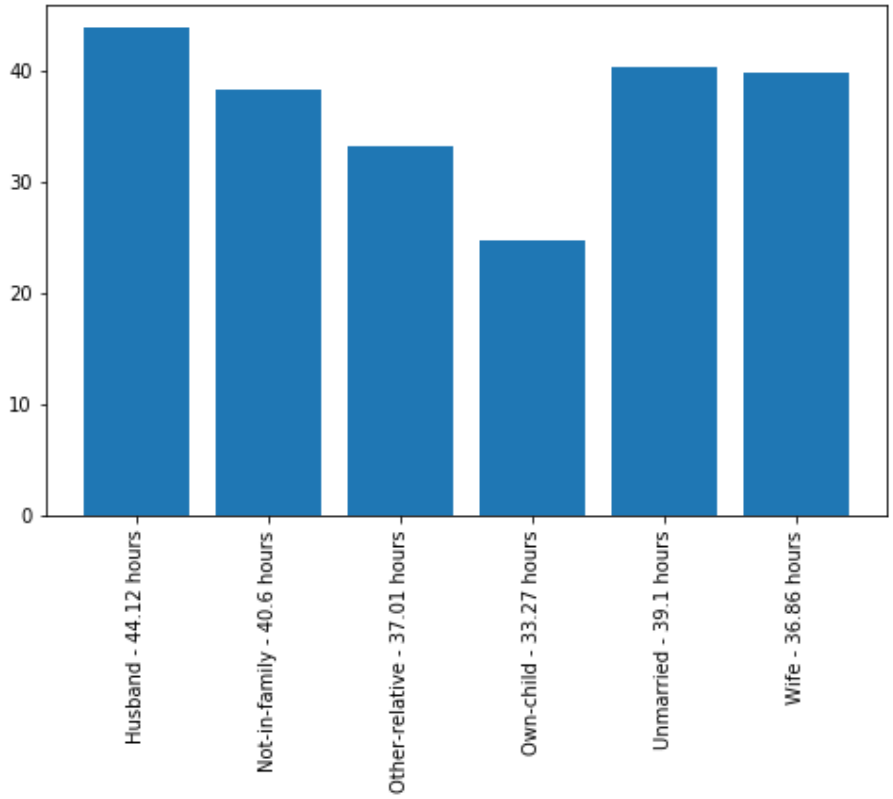

Figure 6: Comparação das horas médias trabalhadas por tipo de relacionamento familiar levando em consideração a idade.

2.5 O que mais você consegue explorar nesta base? Apresente pelo menos mais uma relação que conseguir encontrar utilizando uma técnica de visualização

Outra relação que pode-se perceber ao fazer a comparação da quantidade de pessoas pelo estado civil de acordo com o sexo e idade. A Figura 7 apresenta a quantidade de pessoas (eixo z) pelo seu estado civil (eixo x) de acordo com os sexos masculino (azul) e feminino (vermelho) e suas respectivas idades (eixo y). Pode-se notar que quanto mais velhos, menos mulheres casadas com civis aparecem no gráfico. A maioria das pessoas solteiras tem idade menor ou igual

a 35 anos.

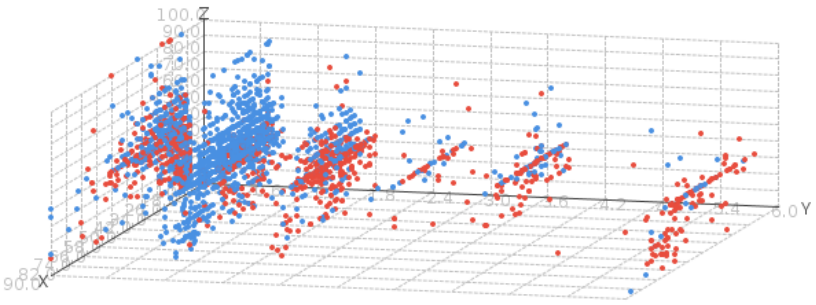

Figure 7: Comparação da quantidade de pessoas pelo estado civil de acordo com o sexo e idade.

3 Conclusão

A primeira atividade prática da disciplina de mineração de dados, do curso de pós-graduação em Ciência Computação da UFMS, pede uma análise detalhada dos dados contidos no relatório do Censo de 1994, feito nos EUA. Foram abordadas todas as questões relacionadas a base de dados de militares do censo de 1994 feito nos estados unidos. Pode-se notar que militares tendem a casar com civis ou serem solteiros.

A base de dados pode ser encontrada em `http://archive.ics.uci.edu/ml/datasets/Adult`. Cada atividade será relatada como uma subseção da Seção 2.