ANÁLISE EXPLORATÓRIA NO CONJUNTO DE DADOS *ADULT*

Adair da S. Oliveira Junior, Anderson A. dos Santos, Ederson R. da Costa

[1]Faculdade de Computação – Universidade Federal de Mato Grosso do Sul (UFMS)
Av. Costa e Silva, s/n – 79.070-900 – Campo Grande – MS – Brazil

{adairsojr, anderson.asantos3}@gmail.com, edersondacosta@hotmail.com

Abstract. *This article presents the explored results of the data combinations from the adult census of the year 1994 of the United States of America. The visualization techniques employed allowed a better analysis and comparison of several data, such as: wages and hours worked among the most varied profiles of individuals; another analysis made it possible to detail the relationship between gender and marital status, as well as several other relationships. The exploration of the data also enabled the elaboration of a correlation matrix, with the objective of displaying the highest relatable values, and thus obtaining important information in the comparison of data.*

Resumo. *Este artigo apresenta os resultados explorados das combinações de dados do censo de indivíduos adultos do ano de 1994 dos Estados Unidos da América. As técnicas de visualização empregadas possibilitaram uma melhor análise e comparação de diversos dados, tais como: salários e horas trabalhadas entre os mais variados perfis de indivíduos; outra análise possibilitou detalhar a relação entre gênero e estado civil, bem como diversas outras relações. A explorarão dos dados também possibilitou a elaboração de uma matriz de correlação, com o objetivo de exibir os maiores valores relacionáveis, e assim obter informações importantes na comparação de dados.*

1. Introdução

Uma análise exploratória, basicamente, consiste em detectar padrões, tendências e relações em dados, o que também está associado com a visualização destes dados [Andrienko and Andrienko 2006]. Baseado nestes conceitos, este artigo tem por objetivo responder a alguns questionamentos da Atividade I, da disciplina de Mineração de Dados, do Programa de Pós-Graduação da Faculdade de Computação (Facom) da Universidade Federal de Mato Grosso do Sul (UFMS).

O conjunto de dados (*dataset*) explorado é do censo de 1994, dos Estados Unidos da América [Becker 1994], e possui informações como, por exemplo, renda, idade, educação, raça e gênero dos entrevistados. O *dataset* é composto por 15 atributos, os quais representam informações dos indivíduos, e um total de 32561 amostras, correspondentes a quantidade de pessoas entrevistadas.

A análise dos dados foi realizada utilizando a linguagem de programação Python[1], a biblioteca de análise de dados Pandas[2] e a biblioteca de plotagem Matplotlib[3]. Uma

[1]Disponível em: `https://www.python.org/`.
[2]Disponível em: `https://pandas.pydata.org/`.
[3]Disponível em: `https://matplotlib.org/`.

visão geral do conjunto de dados é apresentada na Tabela 1, em que é possível observar informações como o nome dos atributos, o tipo dos dados e as informações contidas em cada atributo.

Table 1. Informações gerais do conjunto de dados.

Atributo	**Tipo**	**Informações**
age	Racional	*Continuous*
workclass	Nominal	*Private*, *Self-emp-not-inc*, *Self-emp-inc*, *Federal-gov*, *Local-gov*, *State-gov*, *Without-pay*, *Never-worked*
fnlwgt	Racional	*Continuous*
education	Ordinal	*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	Ordinal	*Continuous*
marital-status	Nominal	*Married-civ-spouse*, *Divorced*, *Never-married*, *Separated*, *Widowed*, *Married-spouse-absent*, *Married-AF-spouse*
occupation	Nominal	*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	Nominal	*Wife*, *Own-child*, *Husband*, *Not-in-family*, *Other-relative*, *Unmarried*
race	Nominal	*White*, *Asian-Pac-Islander*, *Amer-Indian-Eskimo*, *Other*, *Black*
sex	Nominal	*Female*, *Male*
capital-gain	Racional	*Continuous*
capital-loss	Racional	*Continuous*
hours-per-week	Racional	*Continuous*
native-country	Nominal	*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	Racional	$> 50K$, $<= 50K$

Com base nas informações descritas, cada uma das questões da atividade é abordada em uma seção. Assim sendo, a Seção 2 descreve a relação entre gênero e estado civil,

a Seção 3 mostra a relação entre gênero e estado civil, a Seção 4 apresenta a relação entre ganho capital e ganho anual, a Seção 5 descreve a relação entre ganho anual por profissão e, por fim, a Seção 6 faz outras análises, tendo como base uma matriz de correlação.

2. Relação entre Gênero e Estado Civil

Questão 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 relação entre gênero e estado civil é apresentada na Figura 1, com um gráfico de barras agrupadas, em que cada grupo representa a quantidade de homens e mulheres para um dado estado civil.

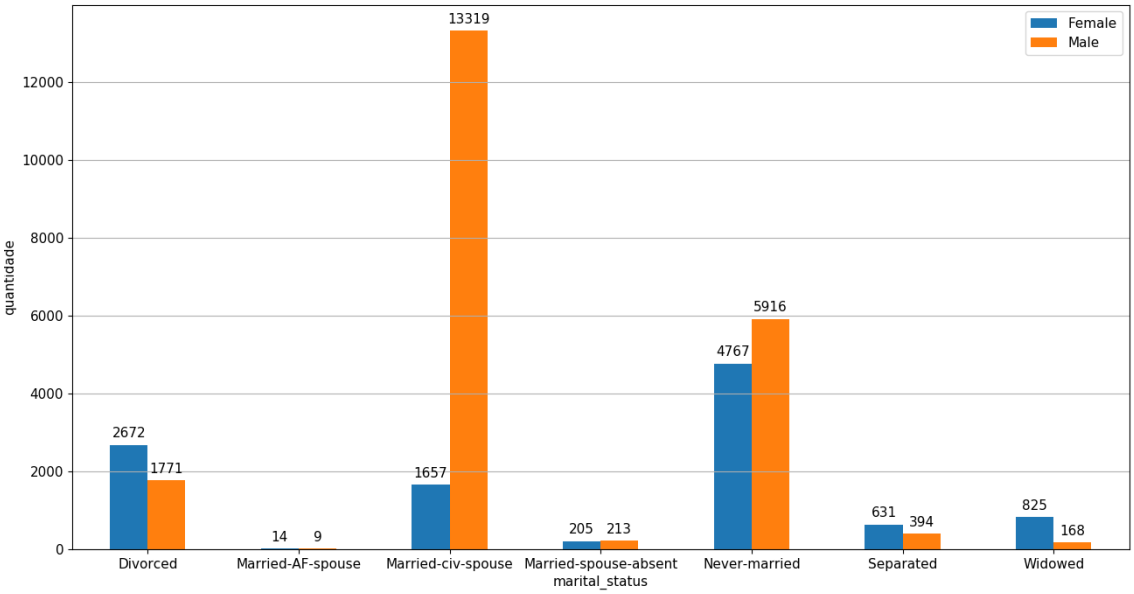

Figure 1. Relação entre gênero e estado civil.

É possível observar na Figura 1 que a quantidade de mulheres divorciadas foi superior aos homens divorciados. Já os dados de indivíduos casados com parceiros das Forças Armadas possuem a menor quantidade de amostras, 23 no total, tendo mais mulheres do que homens casados com conjugues das Forças Armadas. Nos dados sobre indivíduos casados civilmente, os homens estão superiores às mulheres. A relação de conjugues casados ausentes possuem dados parecidos entre homens e mulheres, contando com 205 mulheres e 213 homens. Os homens que nunca casaram superam as mulheres nunca casadas. Já nos dados sobre indivíduos separados ou viúvos, as mulheres superaram os homens.

3. Relação entre Ganho de Capital e Ganho Anual

Questão 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?

Duas classes são atribuídas aos entrevistados, de acordo com seu ganho anual, em que na primeira classe os entrevistados possuem renda anual menor ou igual a

US$50.000,00 ($<= 50K$), e na segunda classe os entrevistados possuem uma renda anual superior a US$50.000,00 ($> 50K$). Alguns entrevistados possuem ganho capital igual a zero e, considerando que a renda destes indivíduos pode prejudicar a visualização dos dados, na Tabela 2 são apresentados os valores do ganho médio anual para cada classe, incluindo os indivíduos com renda capital igual a zero e também excluindo tais amostras.

Table 2. Ganho capital médio de acordo com o ganho anual.

Ganho anual	Ganho médio US$ (com 0)	Ganho médio US$ (sem 0)
$<= 50K$	148,75	3.552,81
$> 50K$	4.006,14	18.731,16

Com os dados apresentados na Tabela 2, pode-se concluir que a média de renda anual, considerando os indivíduos com ganho capital igual a zero, é baixa para os indivíduos da classe $<= 50K$. Quando estes dados são removidos, a média de ganho capital aumenta consideravelmente, para as duas classes. Outra relação a ser analisada é a frequência dos valores de ganho capital de acordo com a classe. Para esta análise, foi utilizado um histograma, que mostra a frequência de uma determinada faixa de valores. Para cada classe ($<= 50K$ e $> 50K$), foi calculado um histograma, de 50 barras, apresentado na Figura 2.

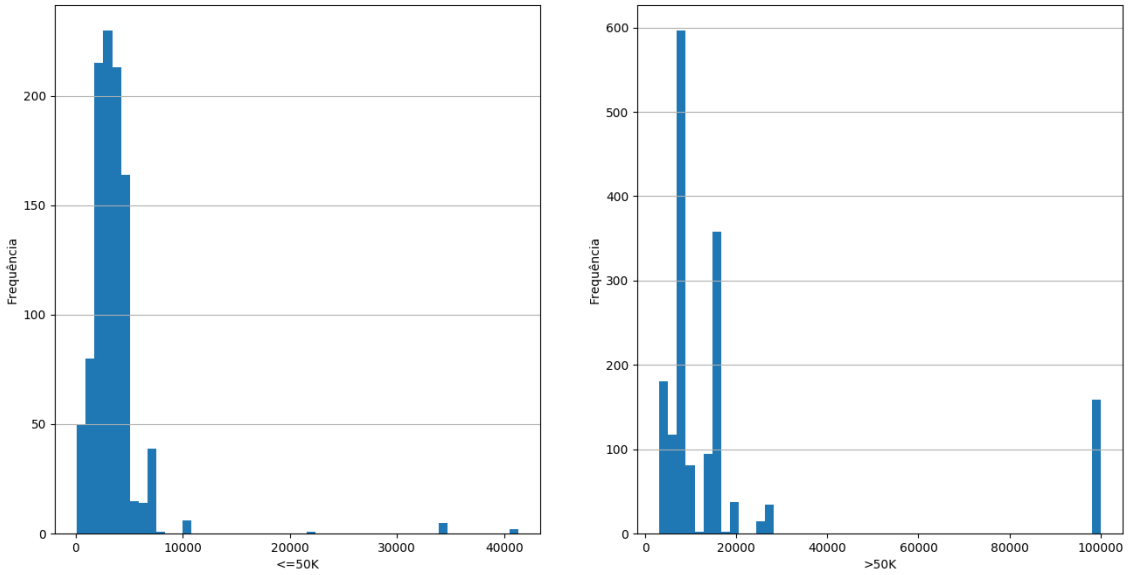

Figure 2. Relação entre Ganho de Capital e Ganho Anual.

Na Figura 2 são mostrados dois histogramas, referentes aos salários de cada uma das classes $<= 50K$ e $> 50K$. O resultado apresentado não representam os indivíduos que apresentam ganho capital igual a zero. Observa-se que a faixa de salários da classe $<= 50K$ está entre zero e aproximadamente US$10.000,00, apresentando mais frequência em torno de US$3.000,00, o que está próximo do valor apresentado na Tabela 2. Na classe $> 50K$, os salários possuem mais variações, e a maioria das pessoas com salários acima de US$50.000,00 obtiveram lucros, na maior parte dos casos, entre zero e US$20.000,00, tendo alguns indivíduos com ganho capital de até US$100.000,00.

4. Relação entre Ganho Anual e Profissão

Questão 3: Quais são as profissões em que mais pessoas tem ganho anual $50K$? Quais profissões tem mais de $50K$?

Para responder a esta questão, foi utilizado um gráfico de barras agrupadas, em que cada grupo representa a quantidade de indivíduos de cada uma das classes $<= 50K$ e $> 50K$ para uma dada profissão. O gráfico da relação entre ganho anual e profissão é apresentado na Figura 3. No primeiro grupo, representado por "?", a profissão dos entrevistados não foi informada.

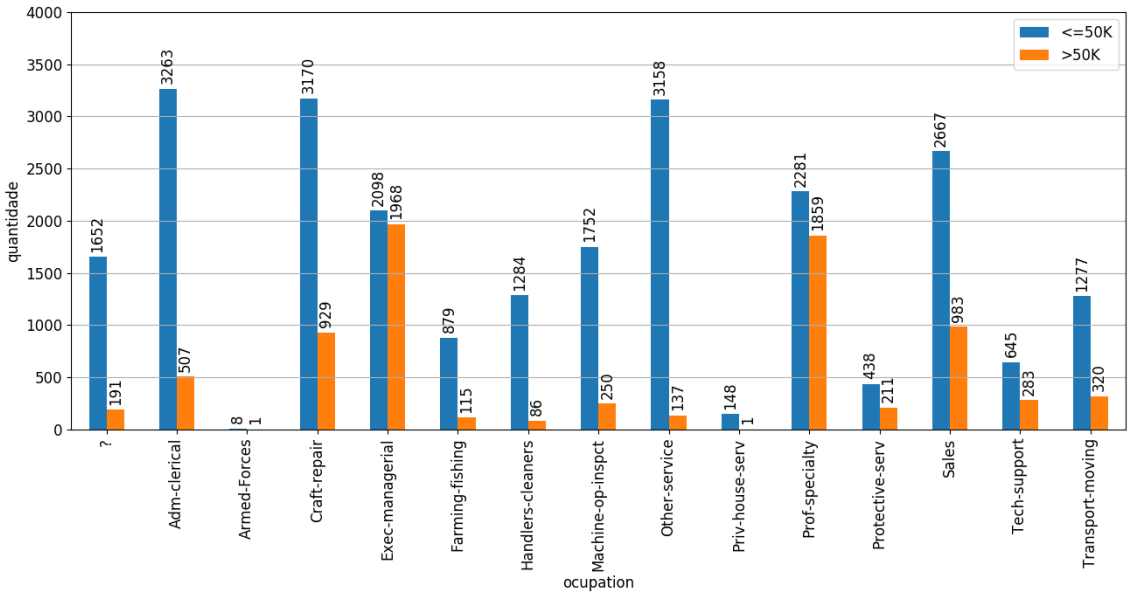

Figure 3. Relação entre Ganho anual e Profissão.

É possível observar na Figura 3 que em todas as profissões, a maioria dos entrevistados possuem ganho anual abaixo de US$50.000,00. De forma similar, porém, em menor número, em todas as profissões analisadas existem profissionais que ganham acima de US$50.000,00. Dentre as profissões analisadas o *Exec-managerial* e *Prof-speciality* são os profissionais que possuem a maior parte dos indivíduos com salários acima de U$ 50.000,00.

5. Relação entre Relacionamento, Horas Trabalhadas e Idade

Questão 4: Quem trabalha mais horas em média: o marido, a esposa, pessoas com filhos etc? Este comportamento varia dependendo da idade?

A primeira parte desta questão, em relação a quem trabalha mais horas em média, pode ser respondida a partir da Tabela 3, que mostra a média de horas trabalhadas por semana para cada um dos relacionamentos.

Table 3. Relação entre relacionamento e média de horas trabalhadas por semana

Relacionamento	Média de horas trabalhadas
Not-in-Family	40.60
Own-Child	33.27
Wife	36.86
Other-Relative	37.01
Husband	44.12
Unmarried	39.10

Na Tabela 3, em média, *Husband* e *Not-in-Family* trabalham mais que 40 horas semanais, enquanto que indivíduos em outras situações de relacionamento trabalham menos.

Para responder a segunda parte da questão, referente a condição de que a quantidade de horas trabalhadas por semana pode variar com a idade, um gráfico para cada situação de relacionamento, com a relação entre idade (representada pelo eixo x) e horas por semana (representada pelo eixo y), é mostrado na Figura 4.

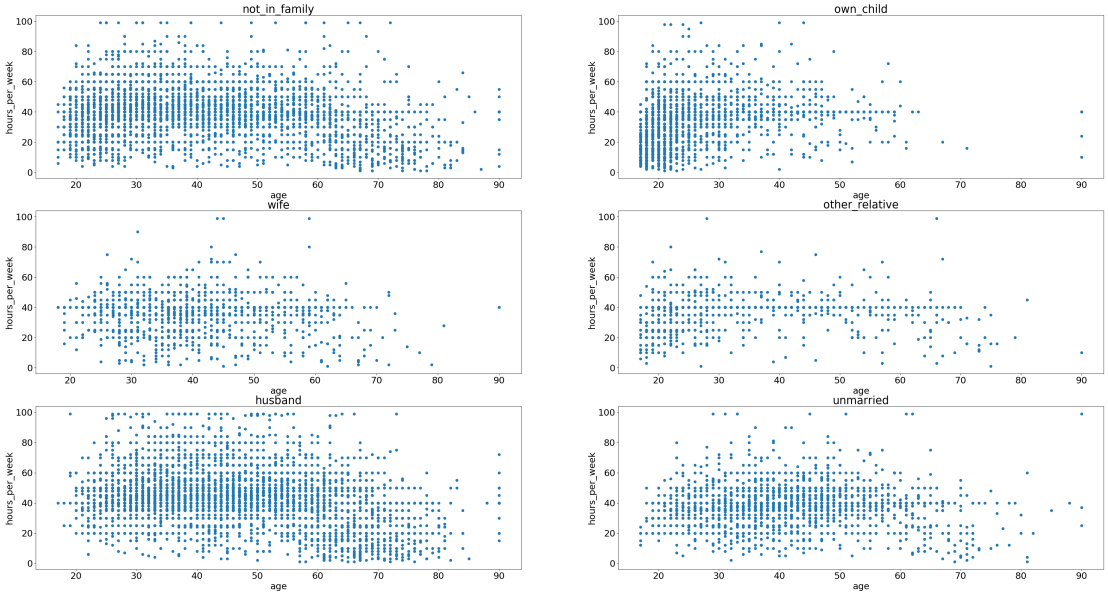

Figure 4. Relação entre Idade, Horas Trabalhadas e Relacionamento.

Na Figura 4, nota-se que para cada situação de relacionamento, os dados se agrupam em determinada faixa de horas trabalhadas e idade. Por exemplo, indivíduos na situação *Not-in-Family* e *Husband* mostram maior concentração de pontos entre as faixas de 20 a 60 horas e 20 a 70 anos. No caso dos indivíduos em situações *Own-Child*, *Wife*, *Other-Relative* e *Unmarried*, com o aumento da idade, a quantidade de indivíduos que trabalham mais horas semanais diminui. Isto indica que para algumas situações de relacionamento, a carga horária de trabalho diminui com o aumento da idade.

6. Demais Relações

Questão 5: O que mais você consegue explorar nesta base? Apresente pelo menos mais uma relação que conseguir encontrar utilizando uma técnica de visualização.

Diversas outras relações entre os atributos podem ser feitas. Como auxílio para encontrar as correlações de atributos que mais se destacam, foi elaborada uma matriz de correlação. Para o cálculo da matriz de correlação, foi utilizado o coeficiente de correlação Pearson, que basicamente, quanto mais próximo o valor de correlação entre dois atributos é próximo de um, mais correlação existe entre os dados. A força de correlação pode ser medida da seguinte forma: entre 0.1 e 0.29 existe uma correlação baixa, entre os valores de 0.3 e 0.49 a correlação é considerada média, já acima de 0.5 a correlação é considerada alta [Benesty et al. 2009].

Para a elaboração da matriz de correlação, os atributos não numéricos, como por exemplo, gênero e raça, foram codificados para números, iniciando do zero até a quantidade de classes subtraída de um. Por exemplo, para o atributo *sex*, as classes são *Female* e *Male*. Assim sendo, por ordem alfabética, *Female* é codificado para zero e *Male* é codificado para um. A técnica de codificação utilizada foi a *LabelEncoder*[4], implementada na biblioteca de aprendizado de máquina *scikit-learn*[5]. A matriz de correlação dos atributos é exibida na Figura 5.

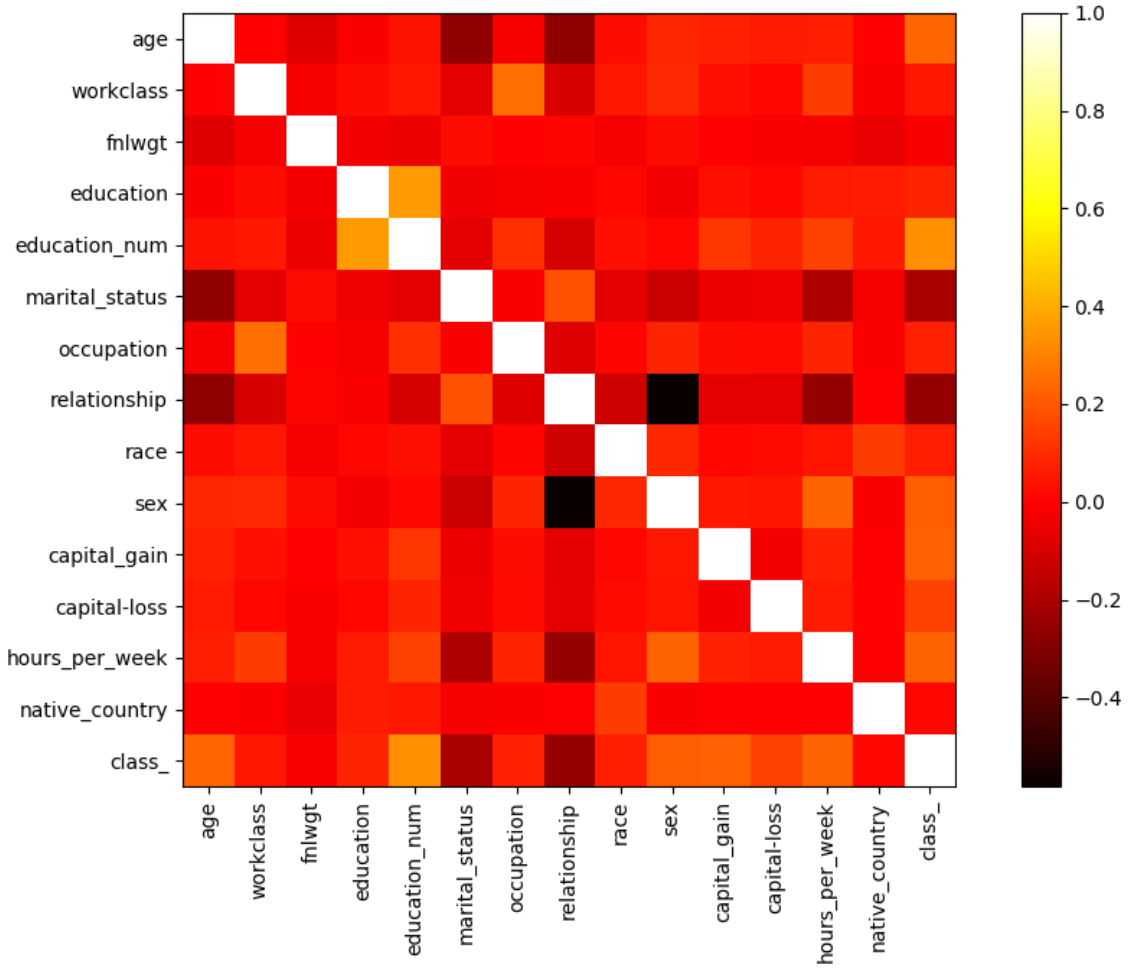

Figure 5. Matriz de Correlação.

[4]Disponível em: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html`.

[5]Disponível em: `https://scikit-learn.org/stable/index.html`.

Na Figura 5, entre as correlações existentes, as que mais se destacam estão associadas ao ganho anual (*class_*). Tendo como base estas relações, foram efetuadas análises correlacionando a educação, idade, gênero e horas trabalhadas com o ganho anual.

6.1. Relação entre Ganho Anual e Educação

A análise de correlação entre ganho anual por educação visa exibir o salário dos indivíduos de acordo com seu nível de escolaridade. Assim sendo, foi elaborado um gráfico de barras, em que cada barra representa um grau de escolaridade e a quantidade de indivíduos por classe é exibida em percentual. O resultado de tal relação é exibido na Figura 6.

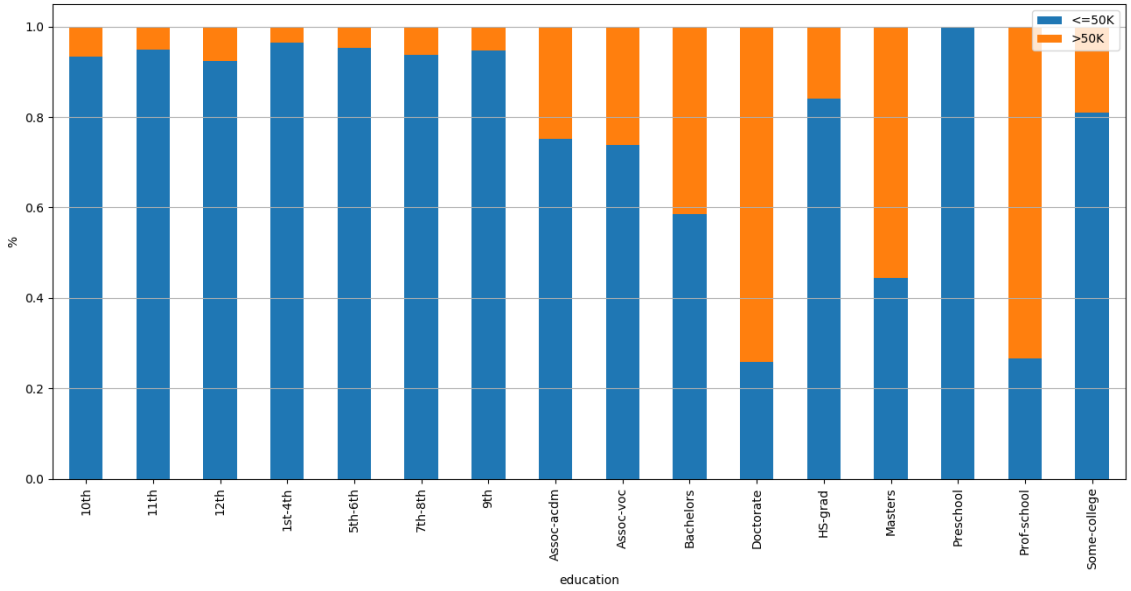

Figure 6. Relação entre Educação e Ganho Anual.

Percebe-se na Figura 6 que a classe de indivíduos formados em escolas profissionais, indivíduos com doutorado e mestrado, tendem a ter salários acima de US$50.000,00. Em contrapartida, a maioria dos indivíduos com outra escolaridade ganha abaixo de US$50.000,00.

6.2. Relação entre Ganho Anual e Idade

Outra análise de correlação efetuada foi entre ganho anual por idade, que visa exibir o salário dos indivíduos de acordo com sua idade. Desta maneira, foi elaborado um gráfico de barras, em que cada barra representa as idades dos indivíduos, agrupadas em intervalos de 10 em 10 anos, bem como o percentual de indivíduos de cada classe. A relação entre os dados pode ser vista na Figura 7.

A partir da visualização da relação entre idade e ganho anual, foi possível perceber as pessoas entre 41 e 60 anos, possuem a maior parcela de indivíduos que tendem a ganhar salários acima de US$50.000,00. Estima-se que indivíduos mais jovens tendem a ganhar menos devido ao fato de estarem na fase de início em algum emprego.

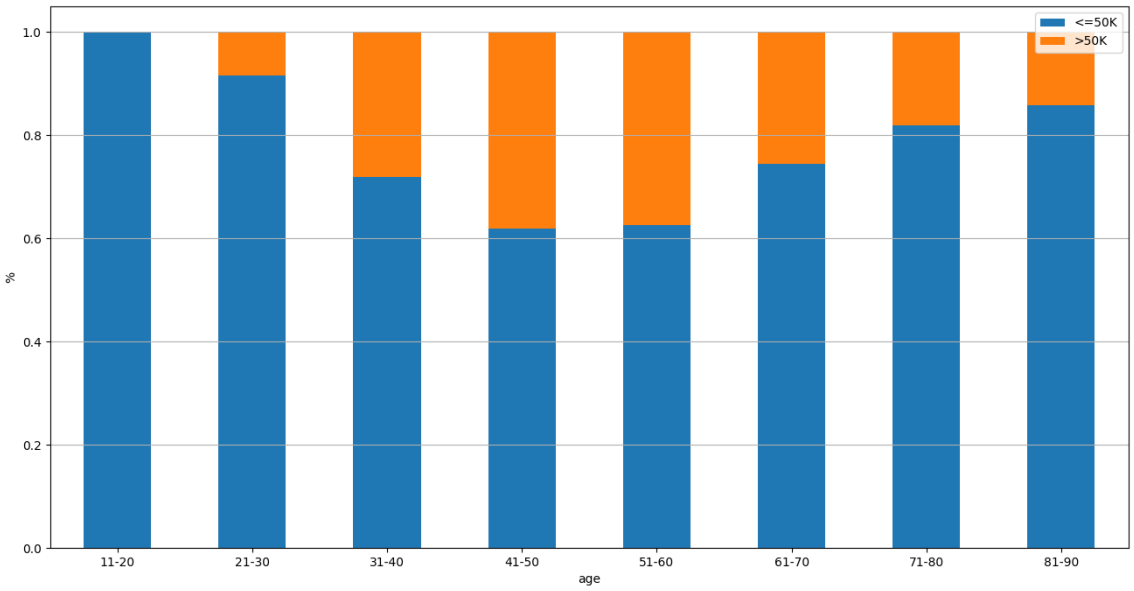

Figure 7. Relação entre Idade e Ganho Anual.

6.3. Relação entre Ganho Anual e Gênero

Foi efetuada uma análise de correlação entre ganho anual e gênero, que buscou identificar o salário dos indivíduos baseado no gênero. Para esta análise, um gráfico de barras mostra o percentual de mulheres e homens para cada uma das classes $<= 50K$ e $> 50K$. A relação entre os dados é exibida na Figura 8.

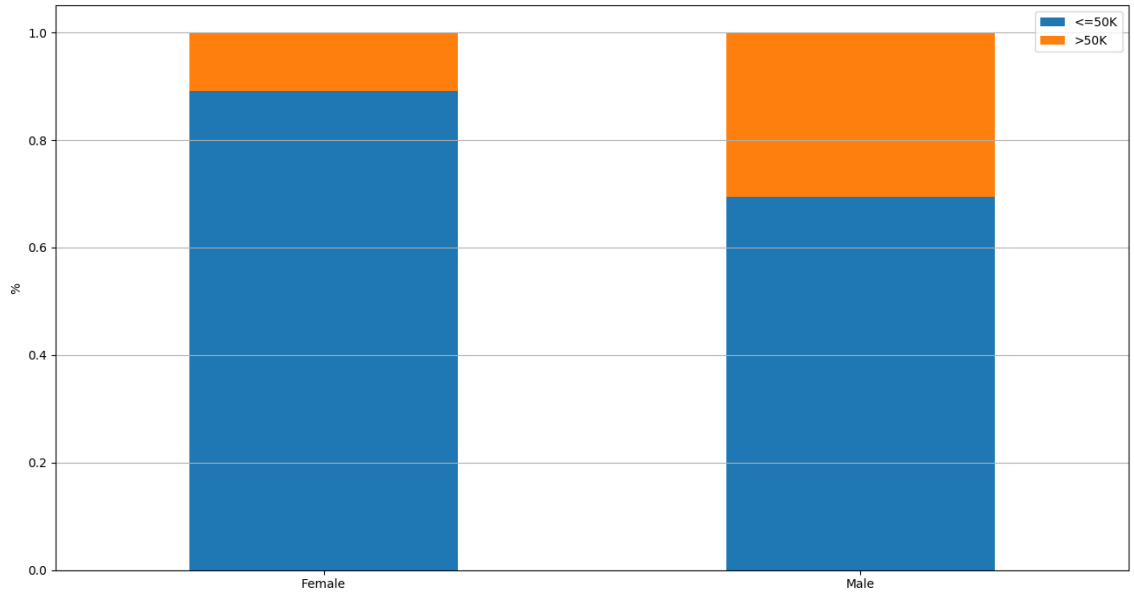

Figure 8. Relação entre Gênero e Ganho Anual.

Na Figura 8, nota-se que a grande parte dos indivíduos do gênero feminino possui ganho anual abaixo de US$50.000,00. As pessoas do gênero masculino também possuem, em sua grande maioria, ganho anual menor do que US$50.000,00, porém, apresentam um

maior percentual de indivíduos que ganham mais que US$50.000,00 em relação ao gênero feminino.

6.4. Relação entre Ganho Anual por Horas Trabalhadas

Por último foi efetuada uma análise de correlação entre ganho anual e horas trabalhadas, que teve como objetivo identificar o salário dos indivíduos de acordo com as horas trabalhadas. Para visualizar os dados, foi elaborado um gráfico de barras, em que cada barra apresenta as horas semanais trabalhadas, agrupadas de 10 em 10, com o percentual de cada uma das classes $<= 50K$ e $> 50K$. A relação entre as informações é detalhada na Figura 9.

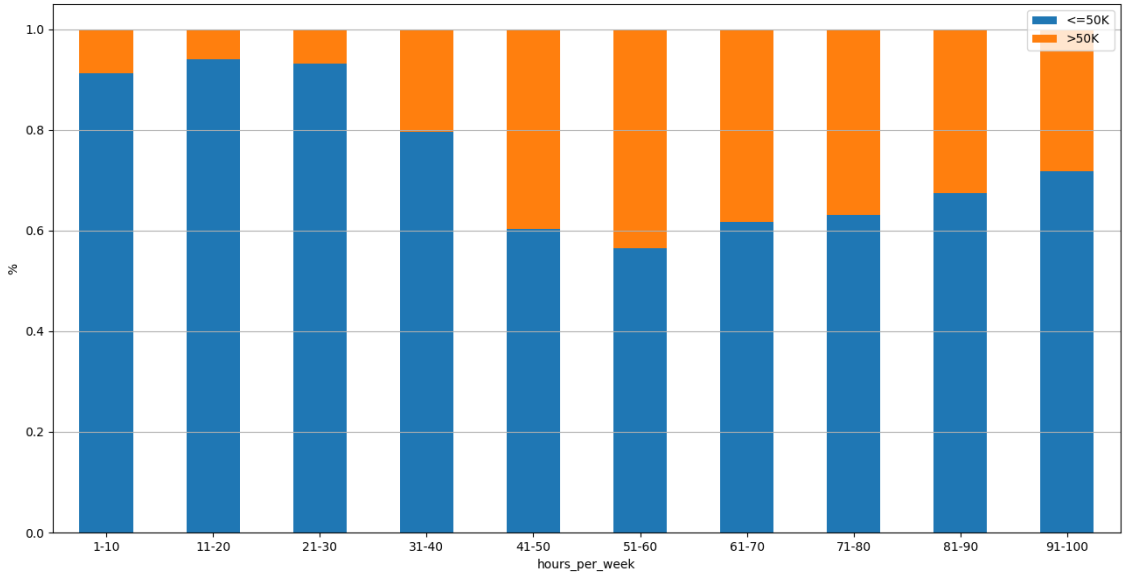

Figure 9. Relação entre Horas Trabalhadas e Ganho Anual.

Percebe-se na Figura 9 que pessoas que trabalham mais horas semanais tendem a um maior ganho anual, principalmente para indivíduos que trabalham mais que 40 horas semanais. A grande maioria dos indivíduos que trabalham abaixo de 40 horas semanais recebem salários abaixo de US$50.000,00.

7. Conclusão

Pôde-se perceber que as técnicas de visualização utilizadas neste artigo foram fundamentais para auxiliar na análise e compreensão dos dados. Foi possível identificar as mais variadas relações entre os dados, o que possibilitou visualizar situações claras de comparações de faixa etária, profissão, gênero, estado civil entre outros. Também foi possível elaborar uma matriz de correlação a fim de analisar outras relações, tal matriz exibiu o maior laço de relação entre os dados, o que possibilitou relacionar educação, idade, gênero e horas trabalhadas com o ganho anual.

References

Andrienko, N. and Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media.

Becker, B. (1994). Adult data set. `http://archive.ics.uci.edu/ml/datasets/Adult`. Acessado em 05/04/2019.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.