

DATA SCIENCE

# Predição de Renda



Leonardo Fiatkoski



# Tópicos

---

- Definição e Objetivo
- Descrição das Variáveis
- Análise Exploratória
- Análise Multivariada
- Modelagem Preditiva
- Análise dos Resultados



## PREDIÇÃO DE RENDA



Definição e Objetivo

## DEFINIÇÃO DO PROBLEMA

O dataset foi extraído em 1994 de uma base do census dos Estados Unidos e contém diversas informações relativas a vida profissional de uma amostra, com a finalidade de verificar possíveis desigualdades salariais e de condições de trabalho entre as pessoas.

## OBJETIVO

Realizar uma análise exploratória para validar a hipótese de que existe desigualdade salarial na população, além da criação de um modelo preditivo para inferir quais pessoas receberão mais de 50 mil dólares por ano.



## PREDIÇÃO DE RENDA



Descrição das Variáveis

# DESCRIÇÃO DAS VARIÁVEIS

Variável	Descrição	Tipo	Pré-Processamento
age	Idade	int	-
workclass	Categoria da empresa em que trabalha;	object	Removidos valores marcados como '?'.
education	Grau de educação	object	-
maritalstatus	Estado civil	object	-
occupation	Área de atuação	object	Removidos valores marcados como '?'.
relationship	Posição na família	object	Remoção dos '_' das informações de Husband e Wife.
race	Raça	object	-
sex	Sexo	object	-
capitalgain	Ganho de capital durante o ano	int	-
capitalloss	Perda de capital durante o ano	int	-
hoursperweek	Horas trabalhadas por semana	int	-
nativecountry	País de origem	object	Separação dos 3 países com maior volume e agrupamento dos demais.
over50k	Salário acima ou abaixo de 50k ano (target)	object	Binarização no momento da modelagem preditiva.

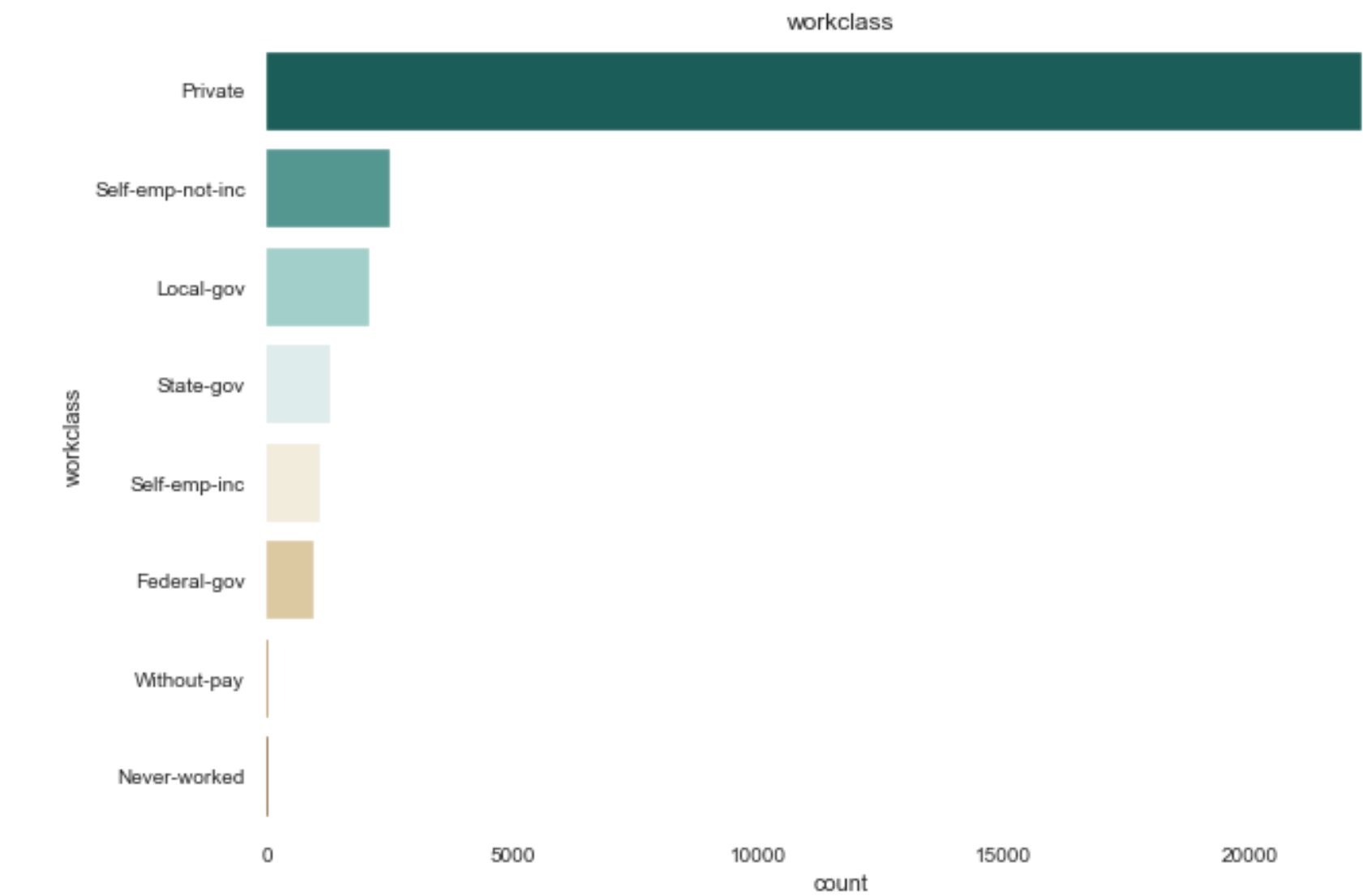
Poucas operações de pré-processamento necessárias. Não localizei nenhuma variável missing, apenas variáveis com valores deturbados (que foram removidos), ajustes no label e remoção de espaços em branco no início e fim da linha.



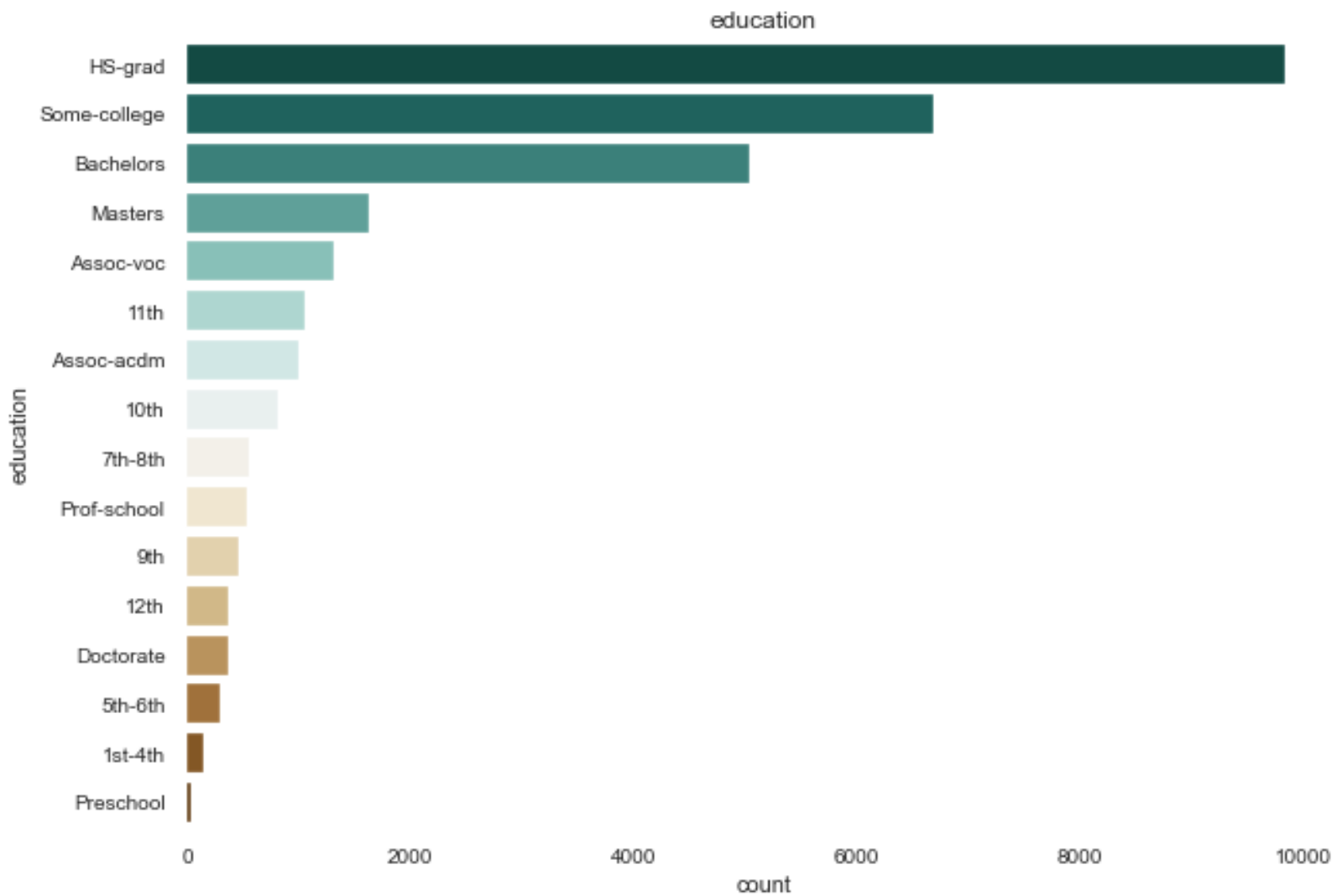
## PREDIÇÃO DE RENDA



Análise Exploratória

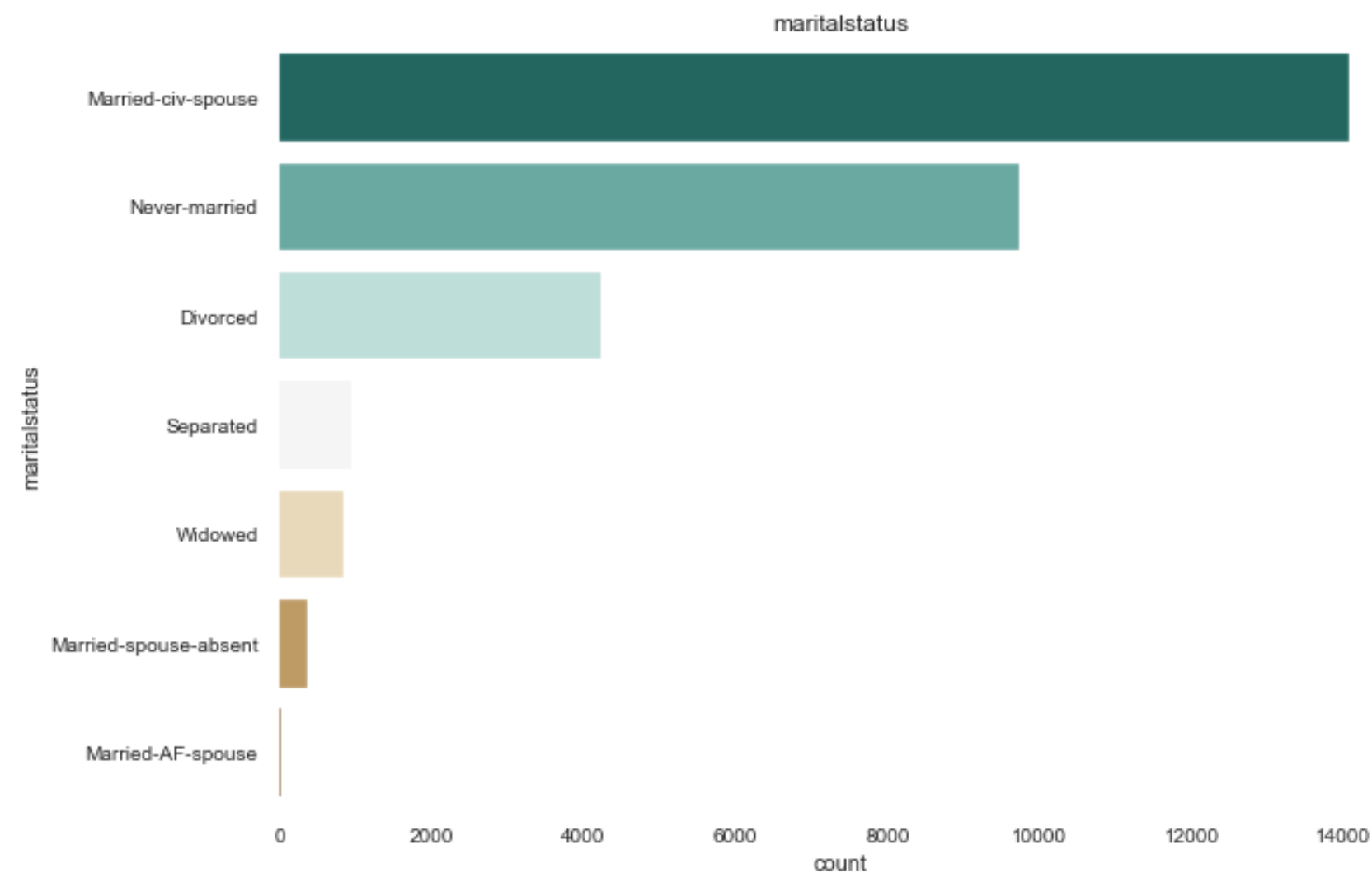


**74%** da base trabalha em **instituições privadas**, seguido por empreendedores e trabalhadores de estatais. Os declarantes que **não recebem salário ou que nunca trabalharam** somam **menos de 1%** de representatividade.

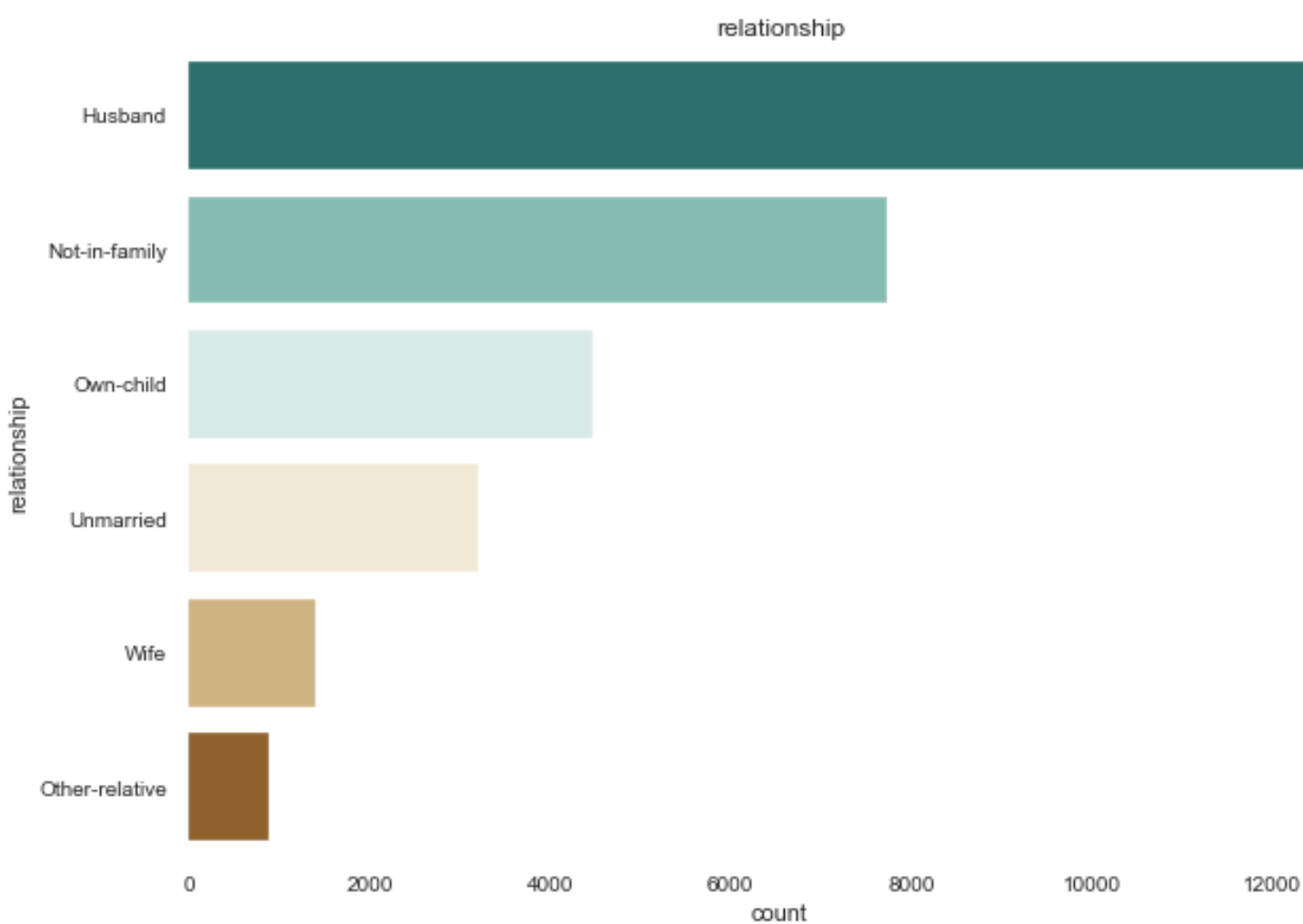


**33%** concluíram apenas a '**high school**', seguido com 22% dos registros como 'some college' que significa estar cursando a faculdade ou qualquer outro curso acima da high school e **17% terminaram a faculdade**.

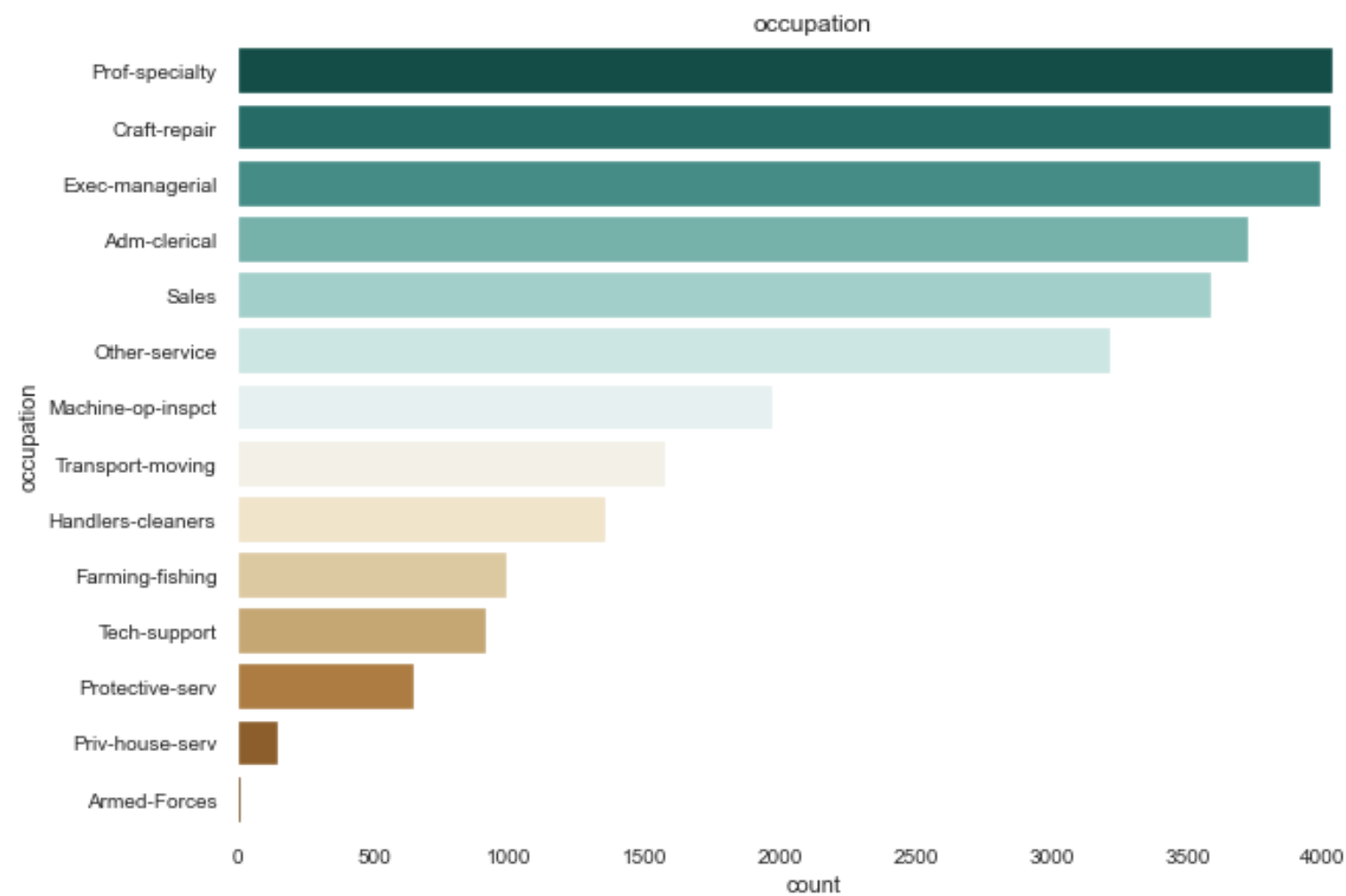




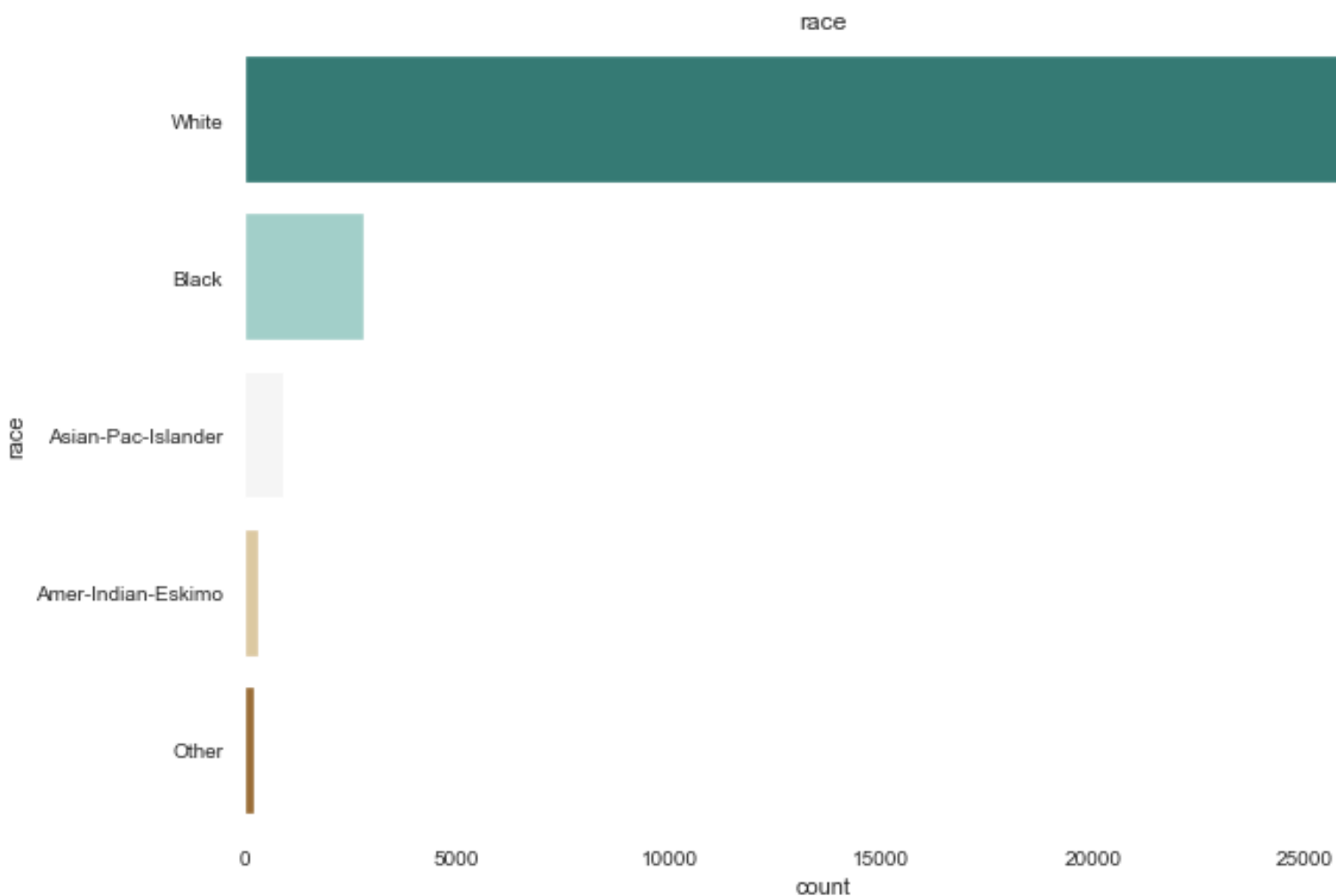
**47%** declararam que estão **casados**, 32% nunca casaram e 14% estão divorciados. Os 'divorciados' e 'separados' foram mantidos em variáveis diferentes por entender que se tratam de coisas diferentes. Os separados podem não ter assinado o divórcio no papel, com isso são legalmente considerados casados.



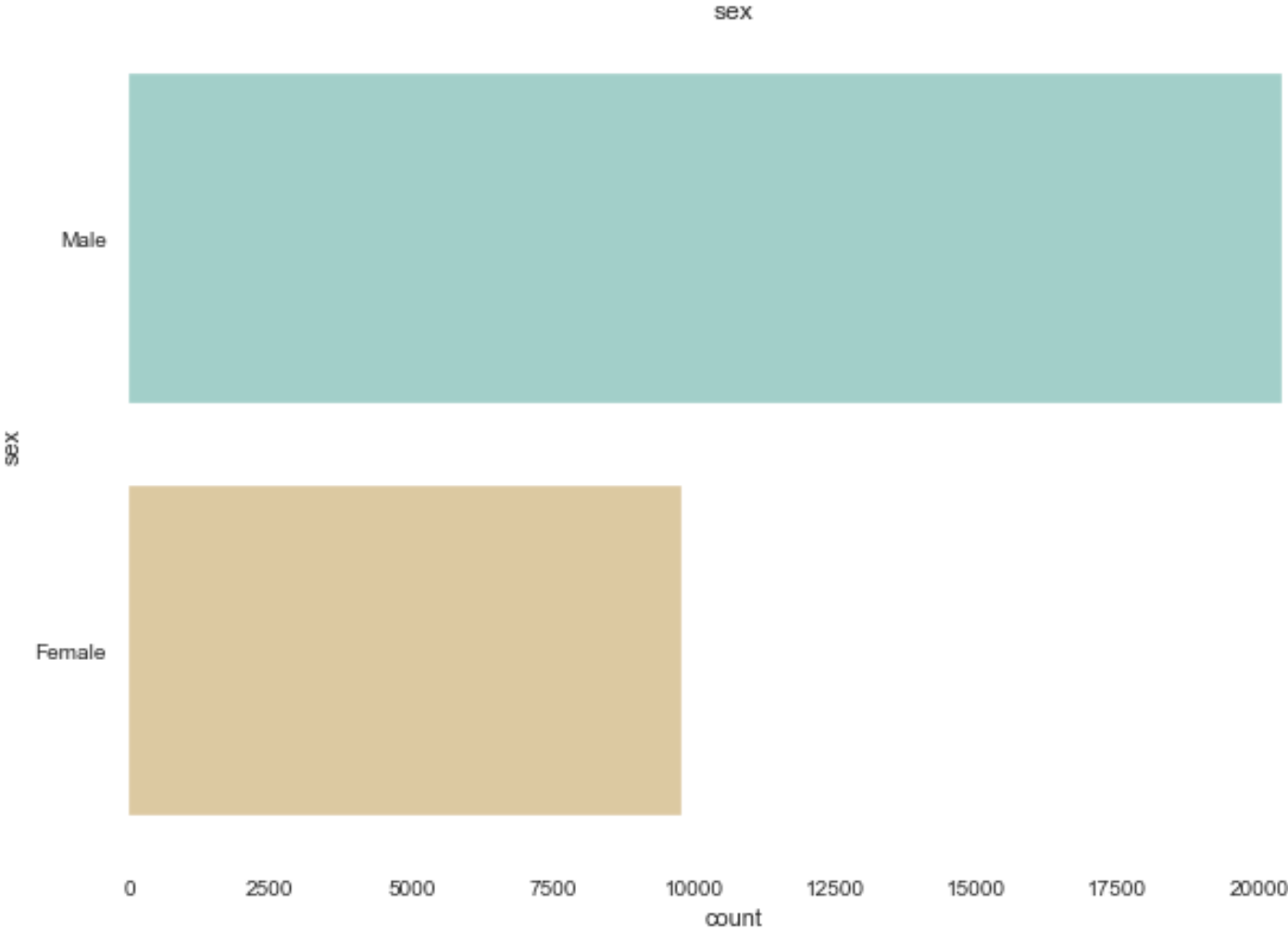
**41%** foram registrados como sendo o marido no status da relação (as esposas foram registradas com 6%, totalizando os 47% da variável 'maritalstatus'), seguido por 26% que declararam não pertencerem a uma família. Por se tratar de uma variável bem semelhante com a 'maritalstatus', iremos removê-la para evitar multicolinearidade.



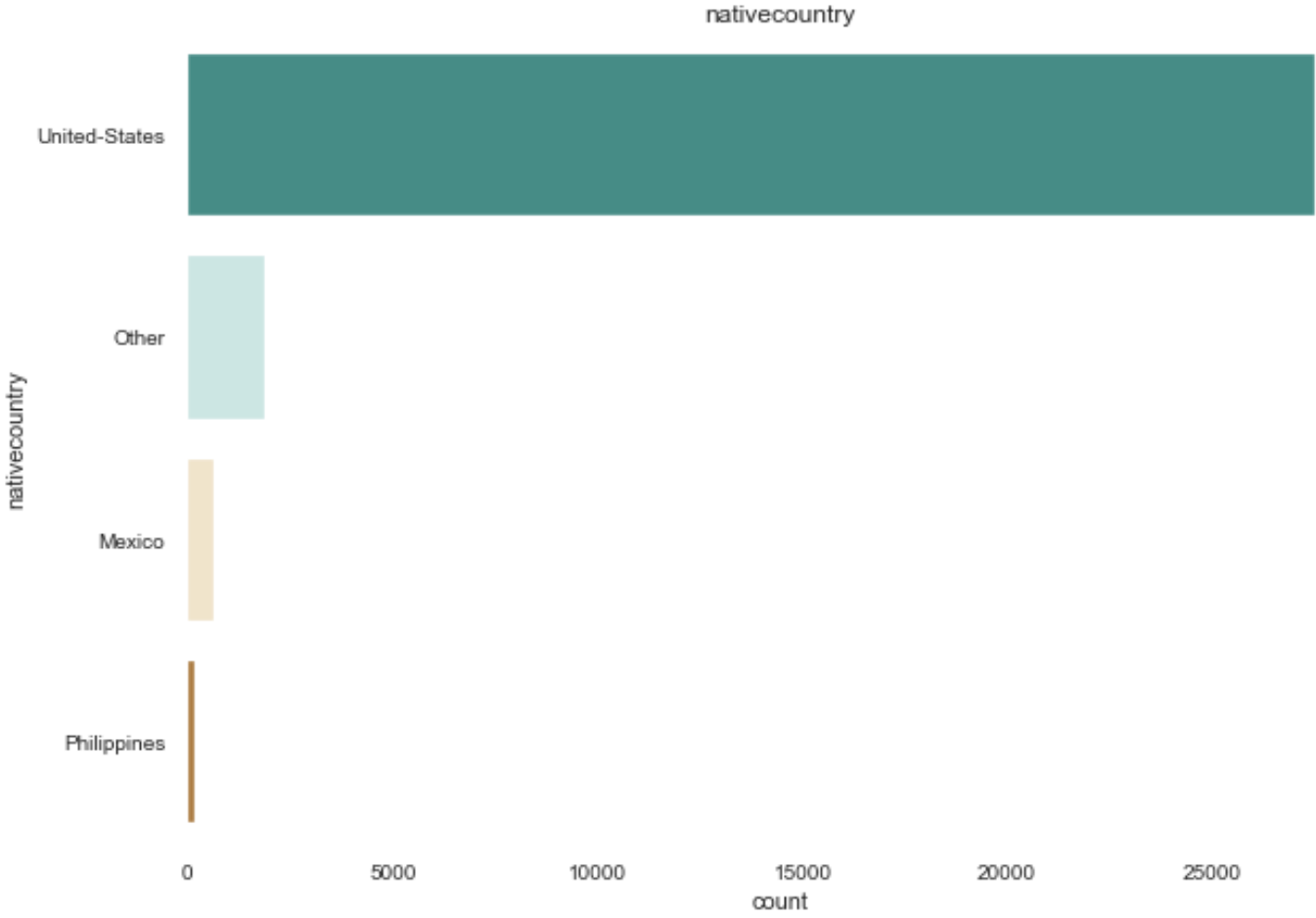
Professores Especialistas, Mecânicos e Executivos representam **13%** da base **cada**, seguidos pelos profissionais do setor administrativo e vendas com 12% casa e 11% declararam 'outros serviços', os **demais cargos** possuem **menos de 10%** de representatividade cada um.



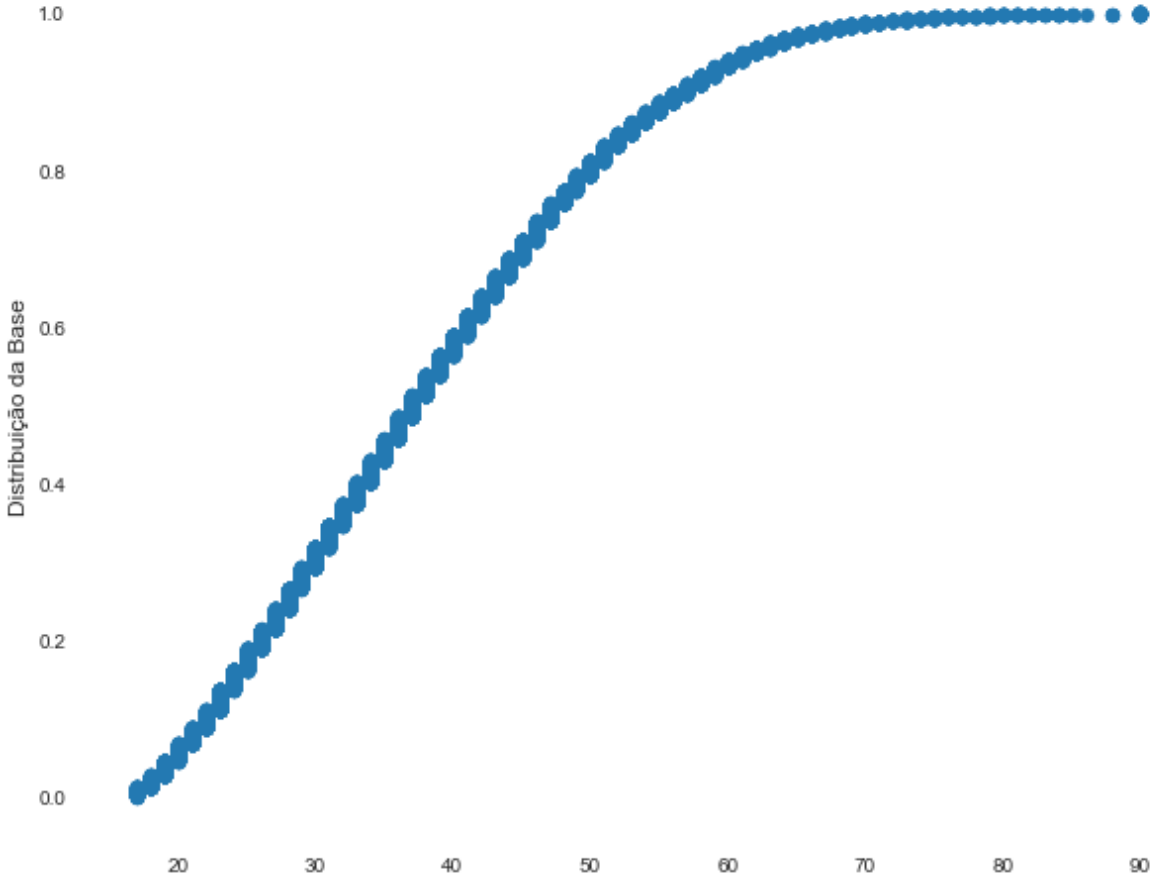
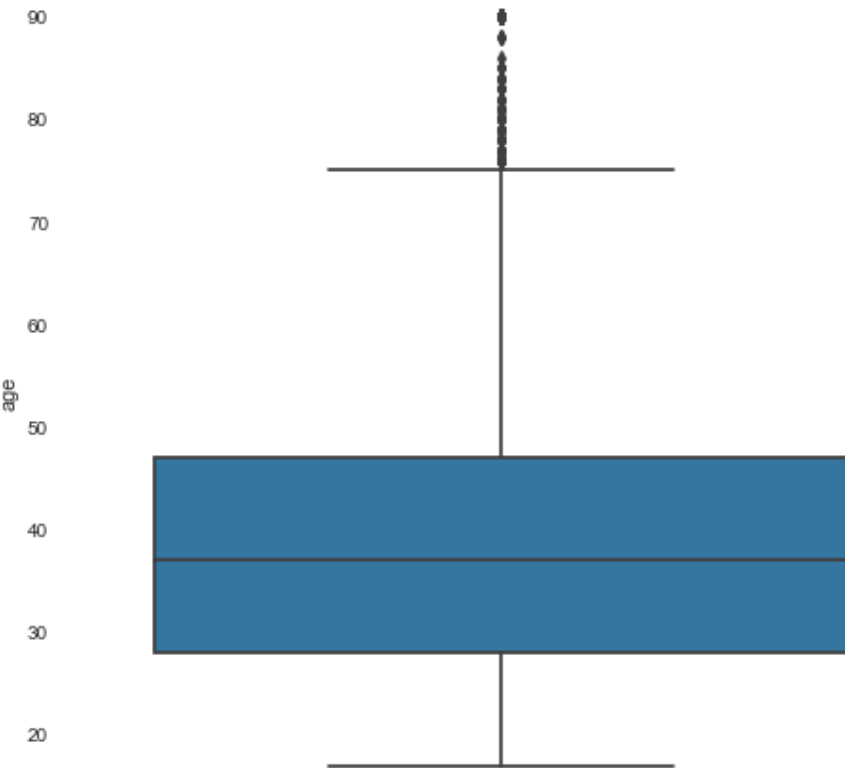
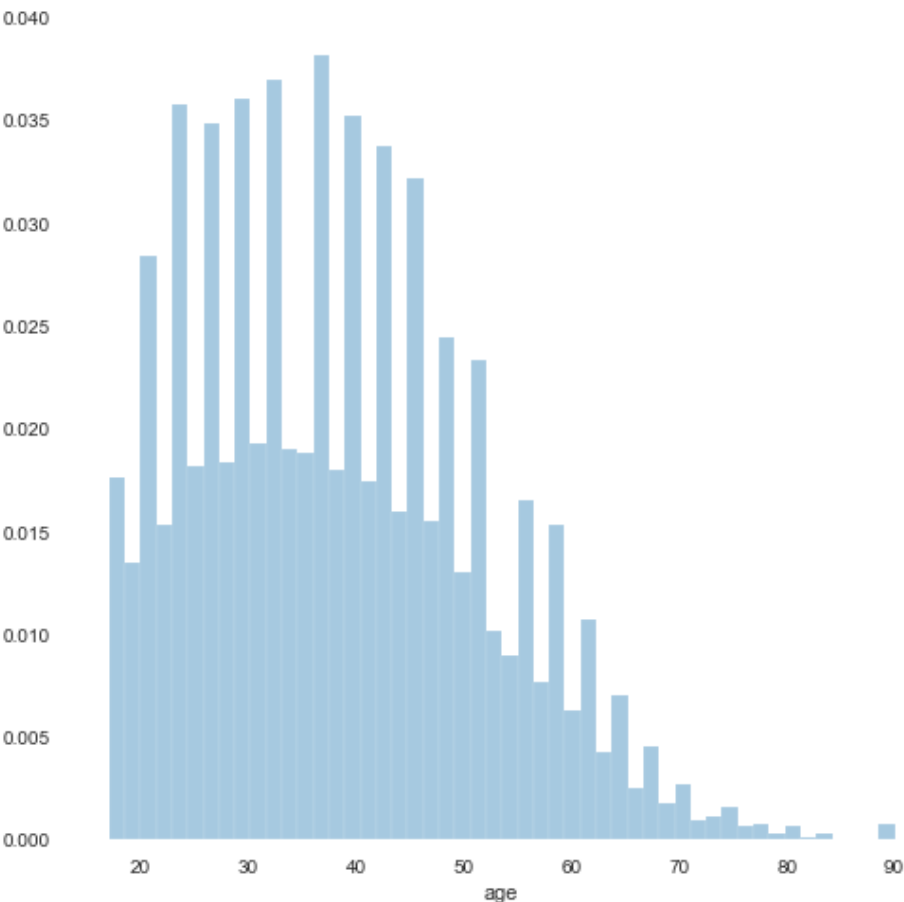
**86%** se consideraram de etnia branca (atualmente a representatividade no país está em 75%), contra **9% de negros**, 3% asiáticos, 1% indianos e 1% de outras etnias.



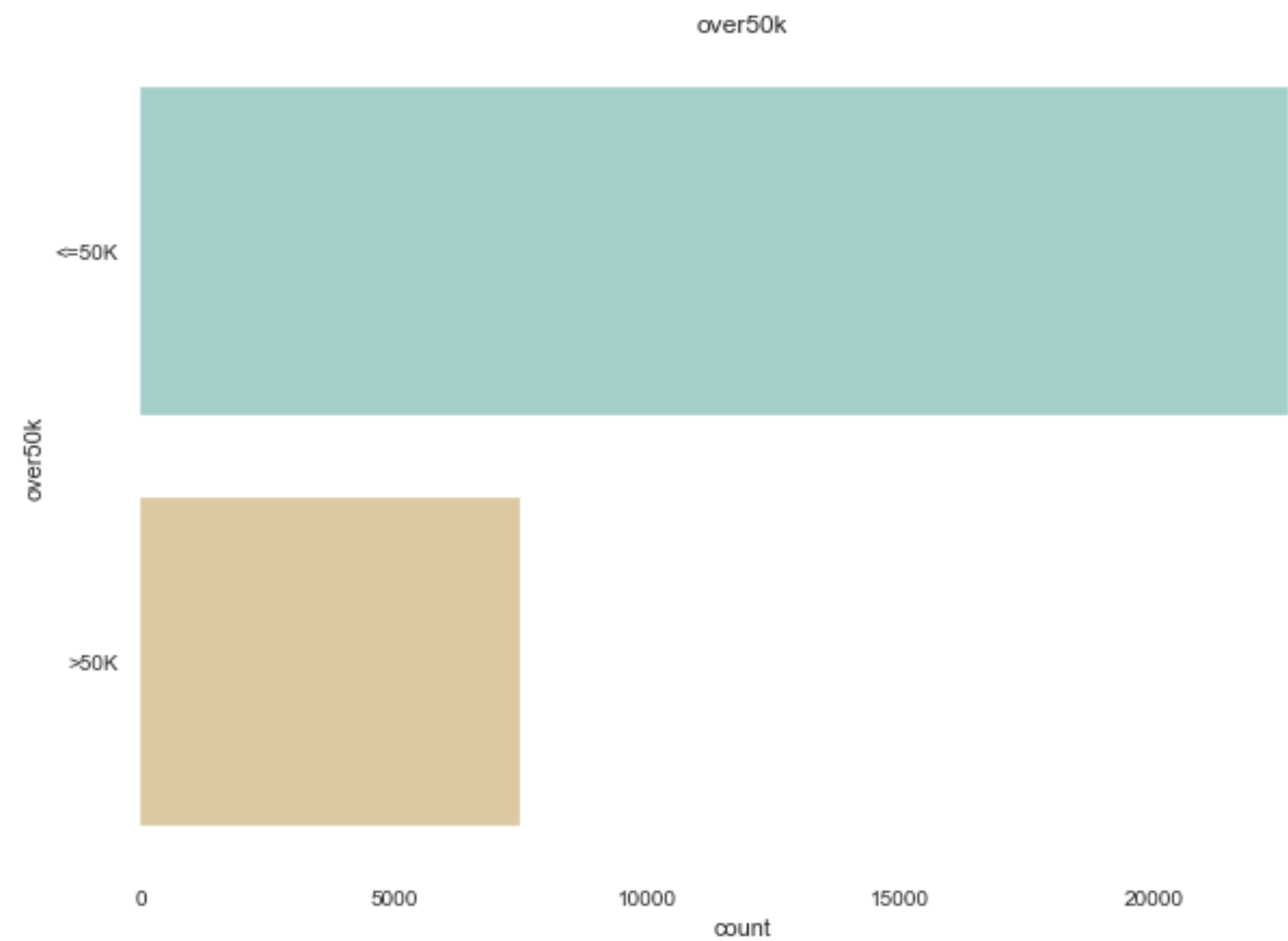
**68%** da base é constituída por **homens**, contra 32% de mulheres.



**91%** da base é constituída por cidadãos **americanos**, o agrupamento das diversas nacionalidades representa 6% da base, seguido de 2% de mexicanos e 1% por filipínos.



Nossa base é constituída por pessoas entre 17 e 90 anos, com **média de 38 anos** e **mediana de 37 anos**. Podemos notar um aumento no número de trabalhadores entre 20 e 30 anos, seguido por uma queda mais acentuada a partir dos 35~40 anos. Alguns outliers foram detectados acima dos 75 anos, mas serão mantidos por fazerem parte da análise em questão. Além disso, o ECDF nos mostra que cerca de **50% da distribuição** total da base possui cerca de **40 anos** de idade.



Agora vamos para nossa variável target, se o respondente **recebe mais de 50 mil dólares por ano** ou não. Os que responderam **sim** constituem **75%** da base, contra 25% que responderam não.

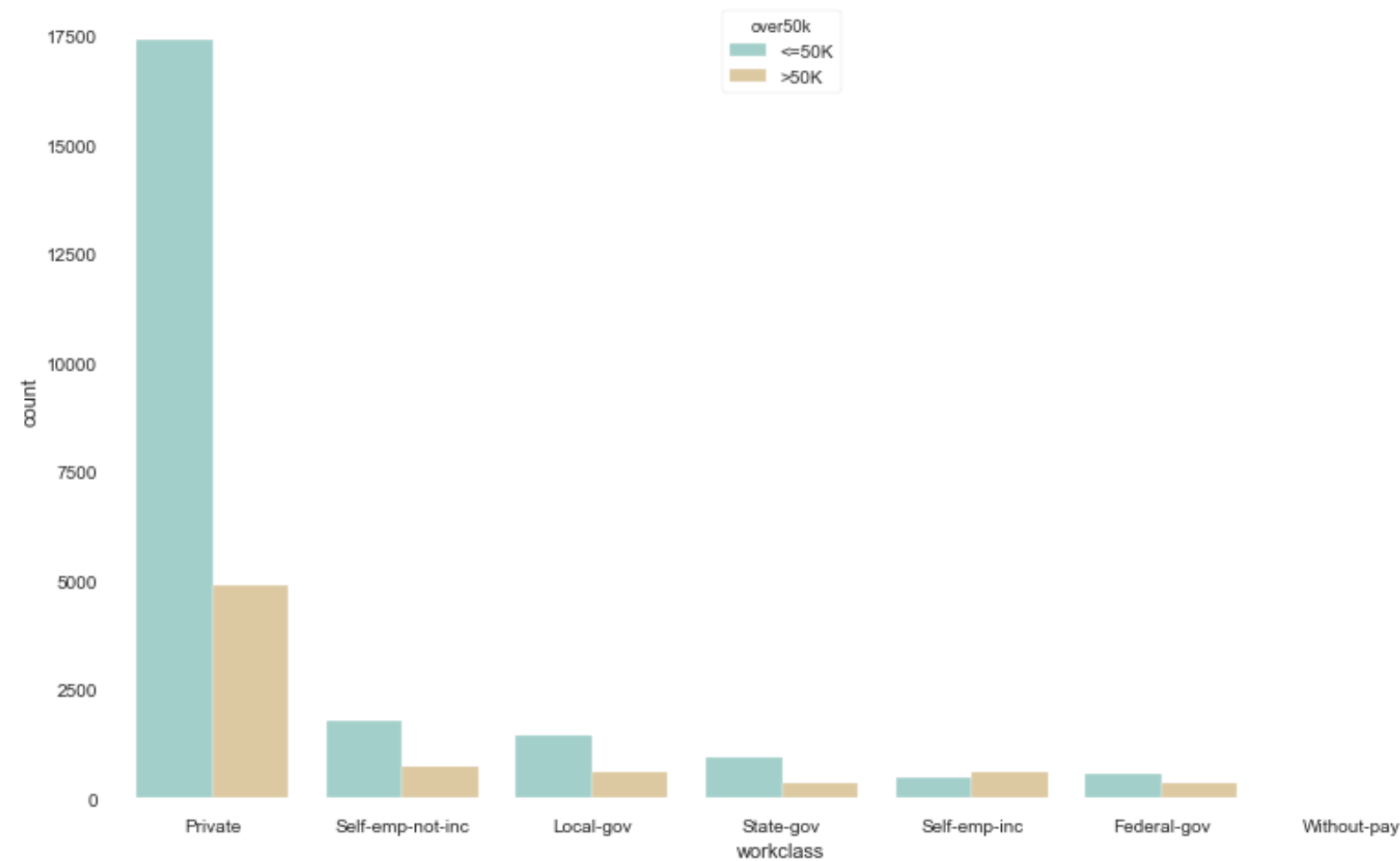
obs: as classes não serão balanceadas no algoritmo, não acredito que esta proporção seja suficientemente desbalanceada para aplicação de over ou undersample.



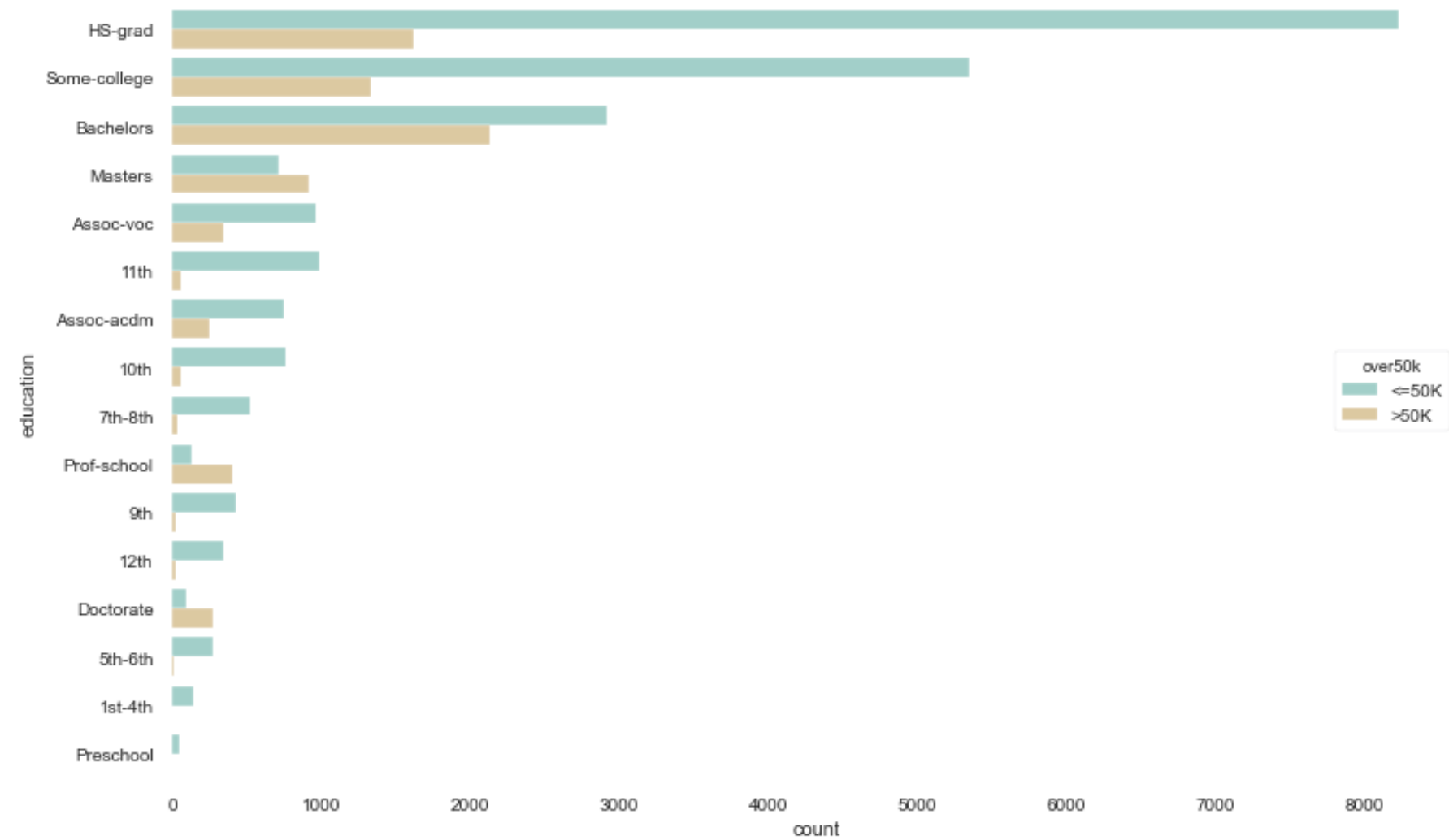
## PREDIÇÃO DE RENDA



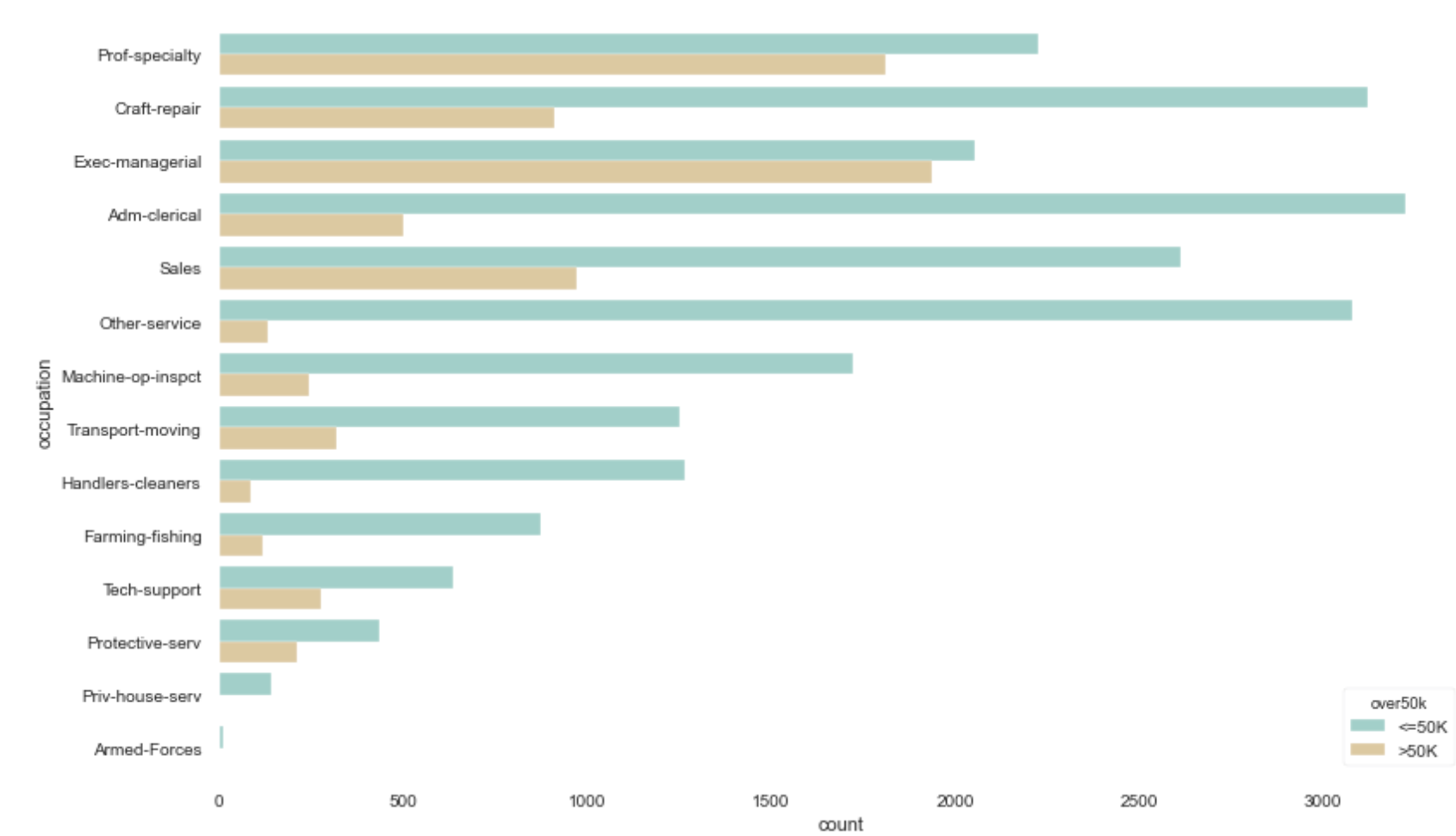
Análise Multivariada



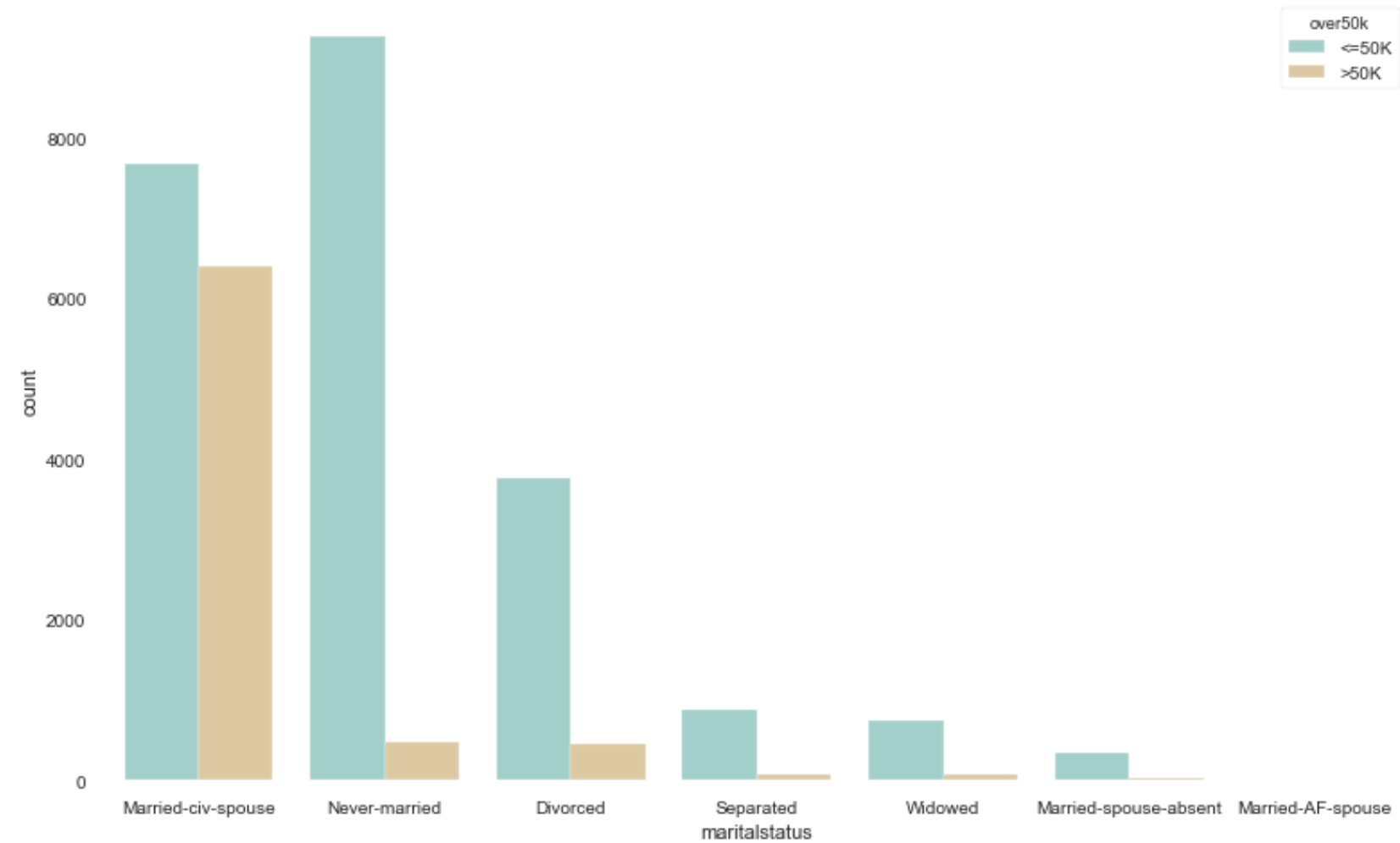
Os **empreendedores lideram** os ganhos acima de 50k com **55%** de representatividade, seguido dos **profissionais federais** com **38%**. Os funcionários de **empresas privadas** possuem o **menor índice** de ganhos com apenas **21%** ganhando acima de 50k por ano (curioso percebem que mesmo com este baixo índice, as instituições privadas ainda abrigam 74% dos profissionais)



O grau de educação com maior índice de ganhos é o de '**professional school**' com **74% ganhando acima de 50k**, que é uma modalidade de graduação para formar profissionais especialistas em determinado tema (conversa com a variável de ocupação profissional), seguido por Doutores, Mestres e Bacharelados. As pessoas que concluíram **até o 'ensino médio'** possuem um índice de ganhos acima de 50k de **menos de 10% cada**.

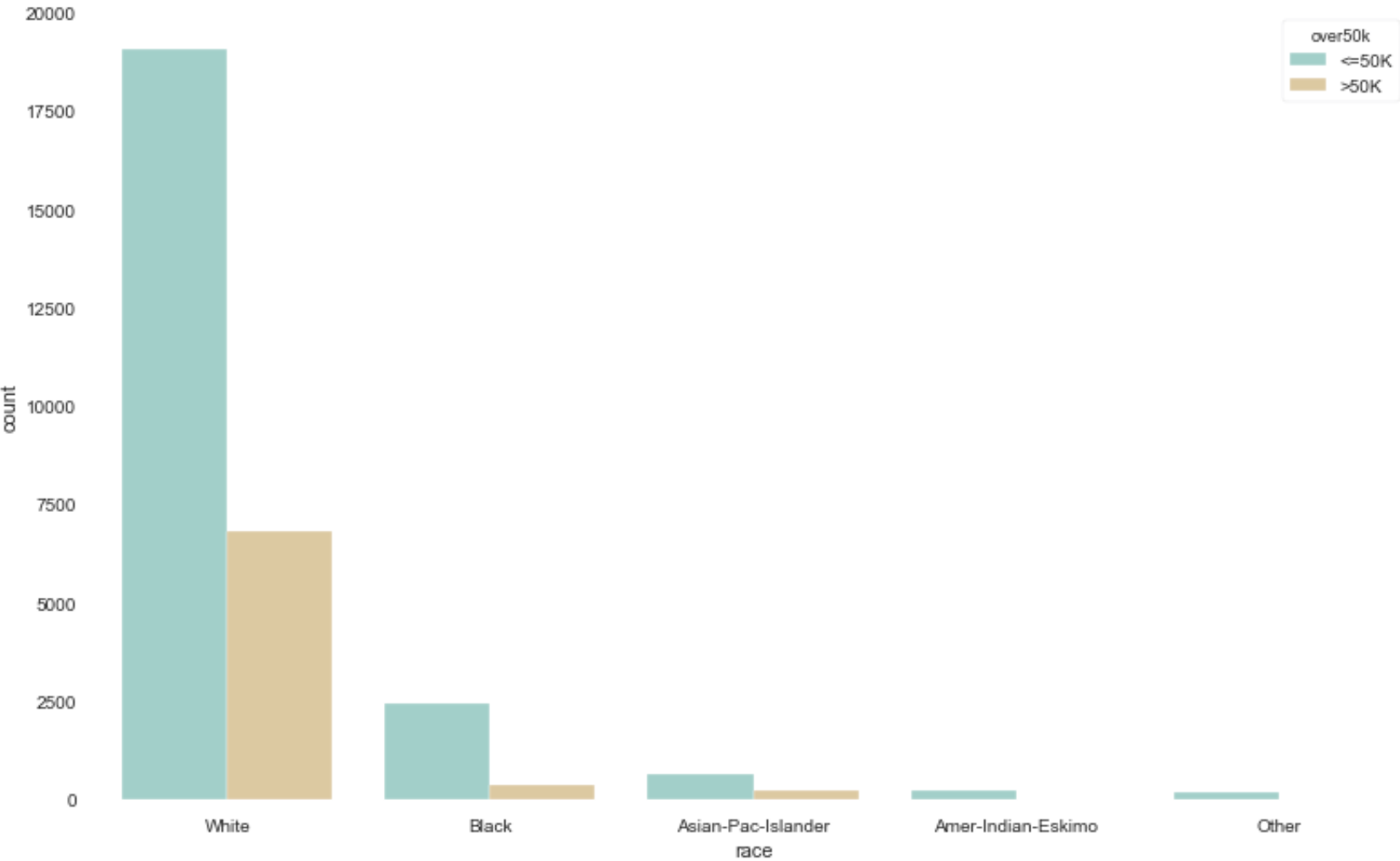


As profissões com **maior percentual de ganhos** acima de 50k é a de **Executivo** com **48%**, seguido por **Professores Especialistas** com **44%**, porém, apesar da representatividade ser maior, ainda são **classes impuras** por **não serem homogêneas**. Empregada doméstica, Outros Serviços, Limpadoras e Forças Armadas são as profissões com o **menor índice de ganhos** acima de 50k, todas representam **1% ou menos** e possuem uma grande homogeneidade, negativa.

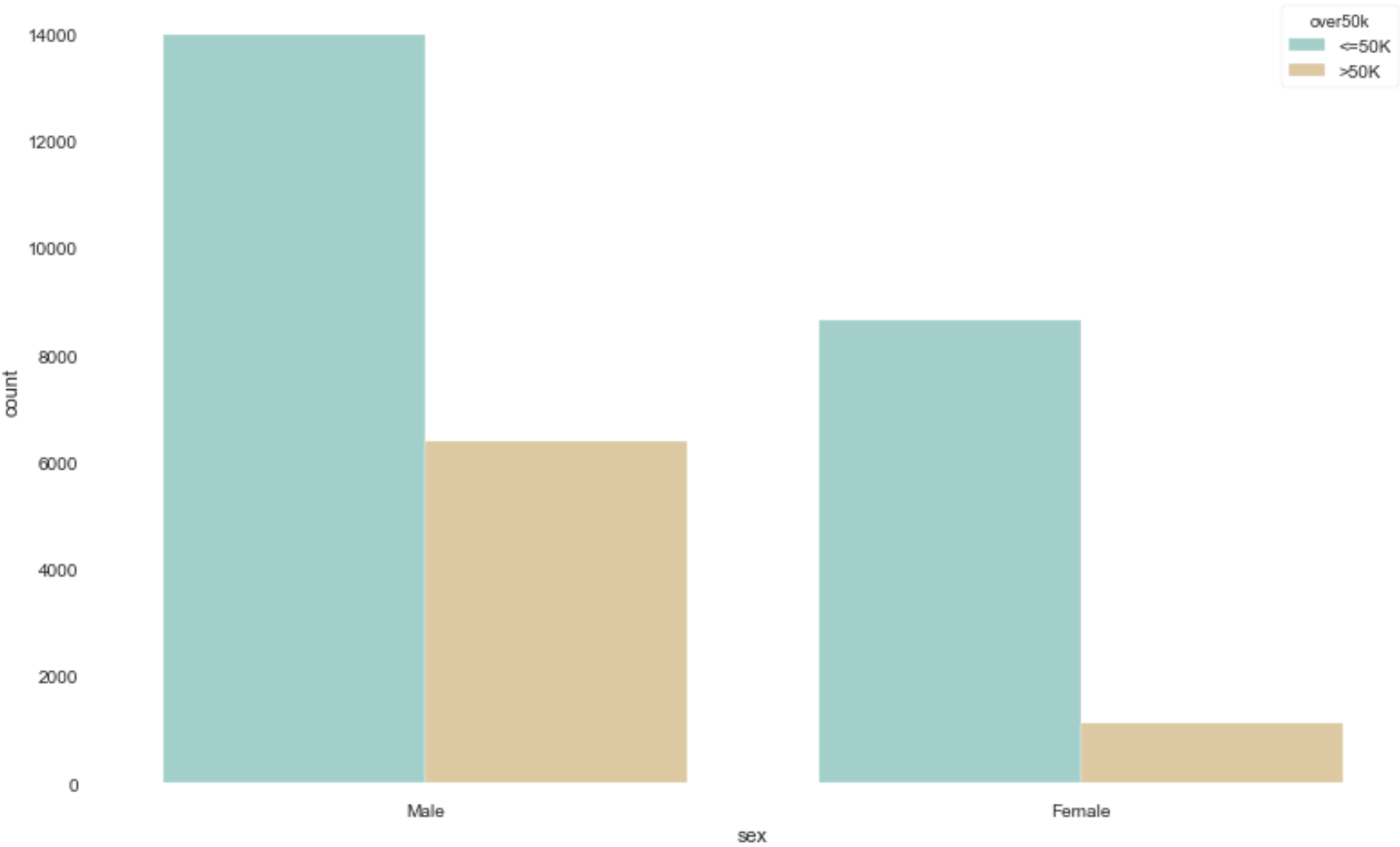


Os **casados** lideram o ranking de ganhos acima de 50k com mais de **45% de representatividade** (casais inteligentes enriquecem juntos), todos os **outros estados civis** possuem **menos de 10%** ganhando acima de 50k.

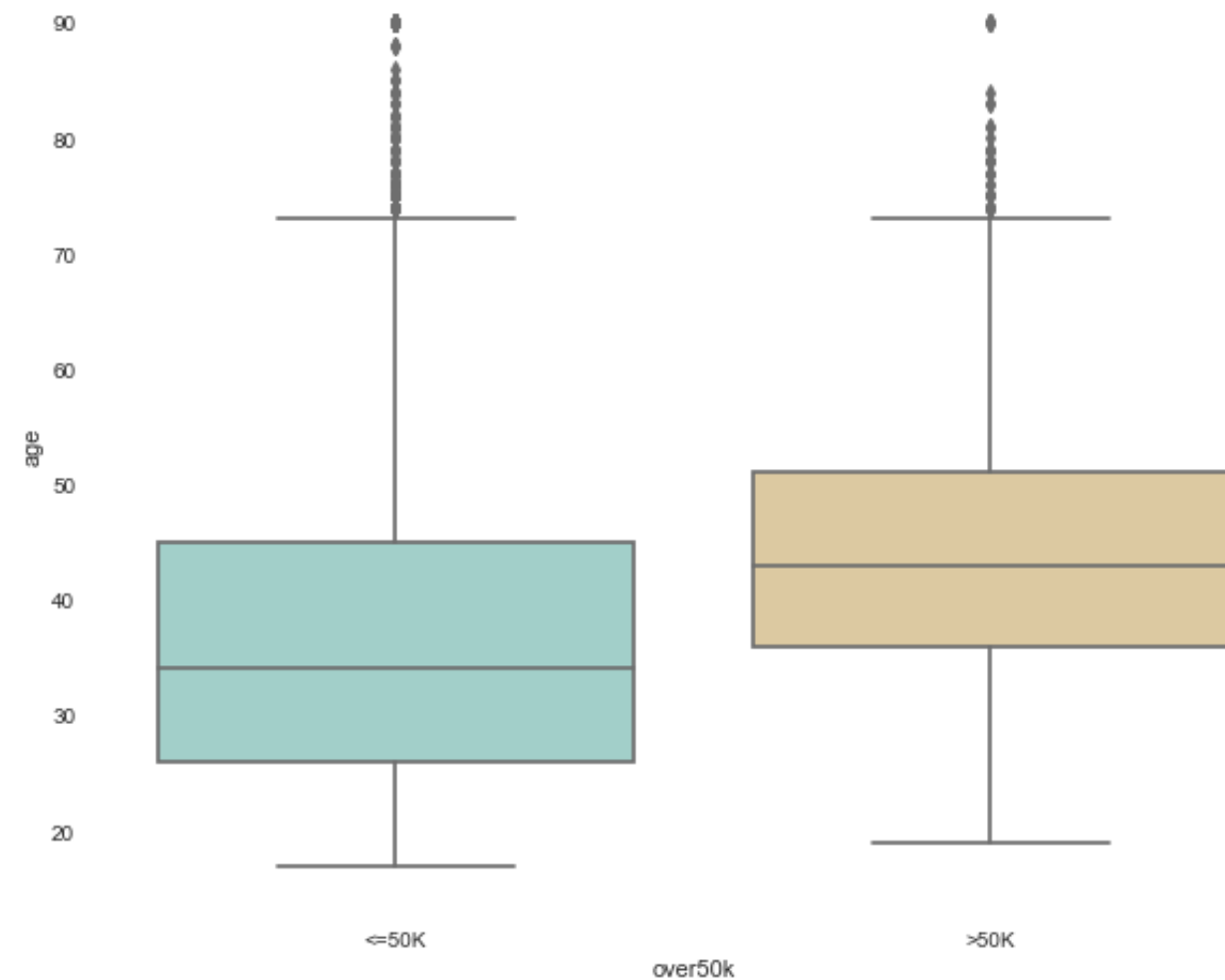




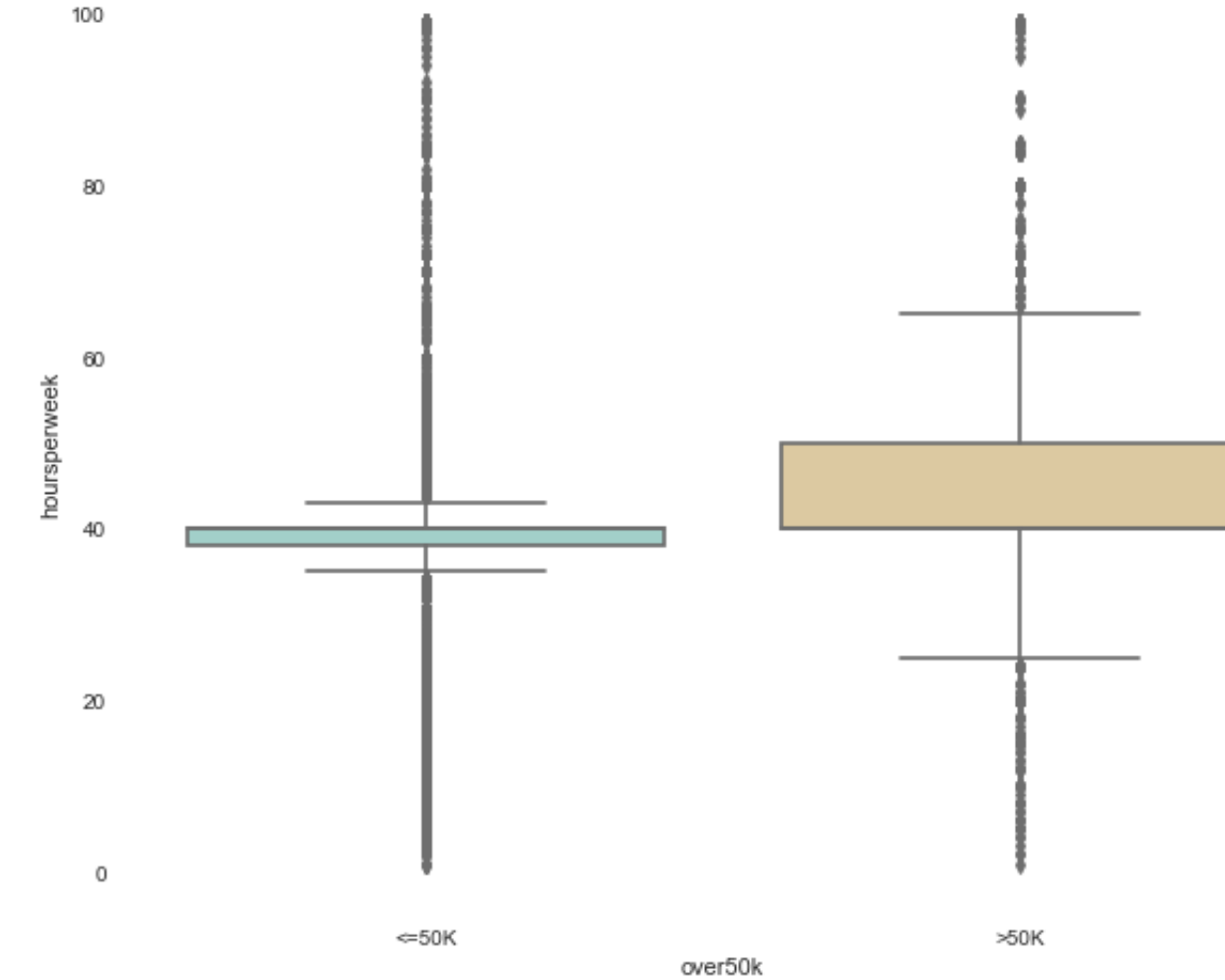
Falando em etnia, os **asiáticos** possuem a **maior representatividade nos altos ganhos** com cerca de **27%**, seguido por brancos com 26%. Negros e Indianos possuem 12% e 11% respectivamente, seguidos por outros com 9%.



Na variável sexo, os **homens** com ganhos acima de 50k são **31%**, contra **apenas 12% das mulheres** (pois é...)



Em relação a idade, percebemos uma **forte relação** da idade **com os ganhos**, a **média de idade** de quem ganha **menos** de 50k por ano é de **33 anos**, já a **média** de idade de quem ganha **mais de 50k** é de **44 anos**.



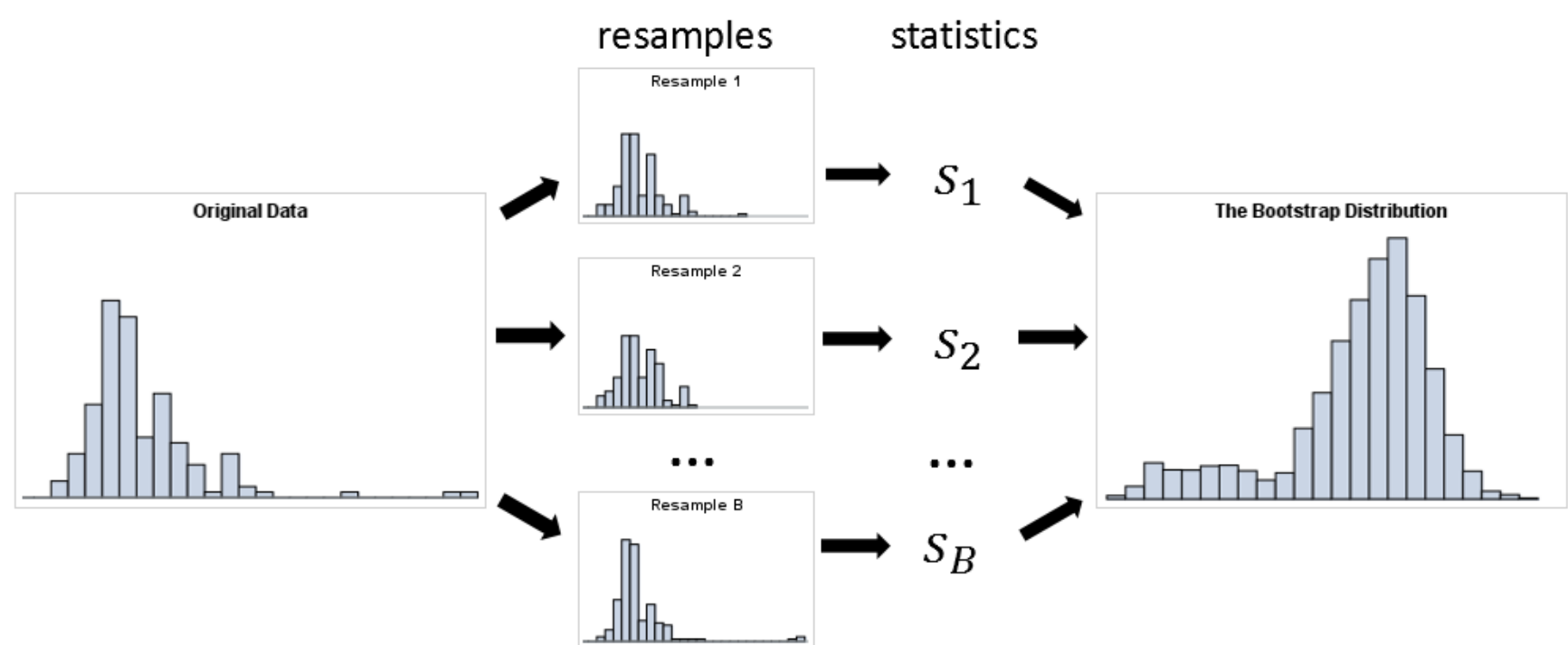
O gráfico nos indica que aqueles com maior média de horas trabalhadas possuem uma maior relação com os ganhos acima de 50k.



## PREDIÇÃO DE RENDA



Modelagem Preditiva



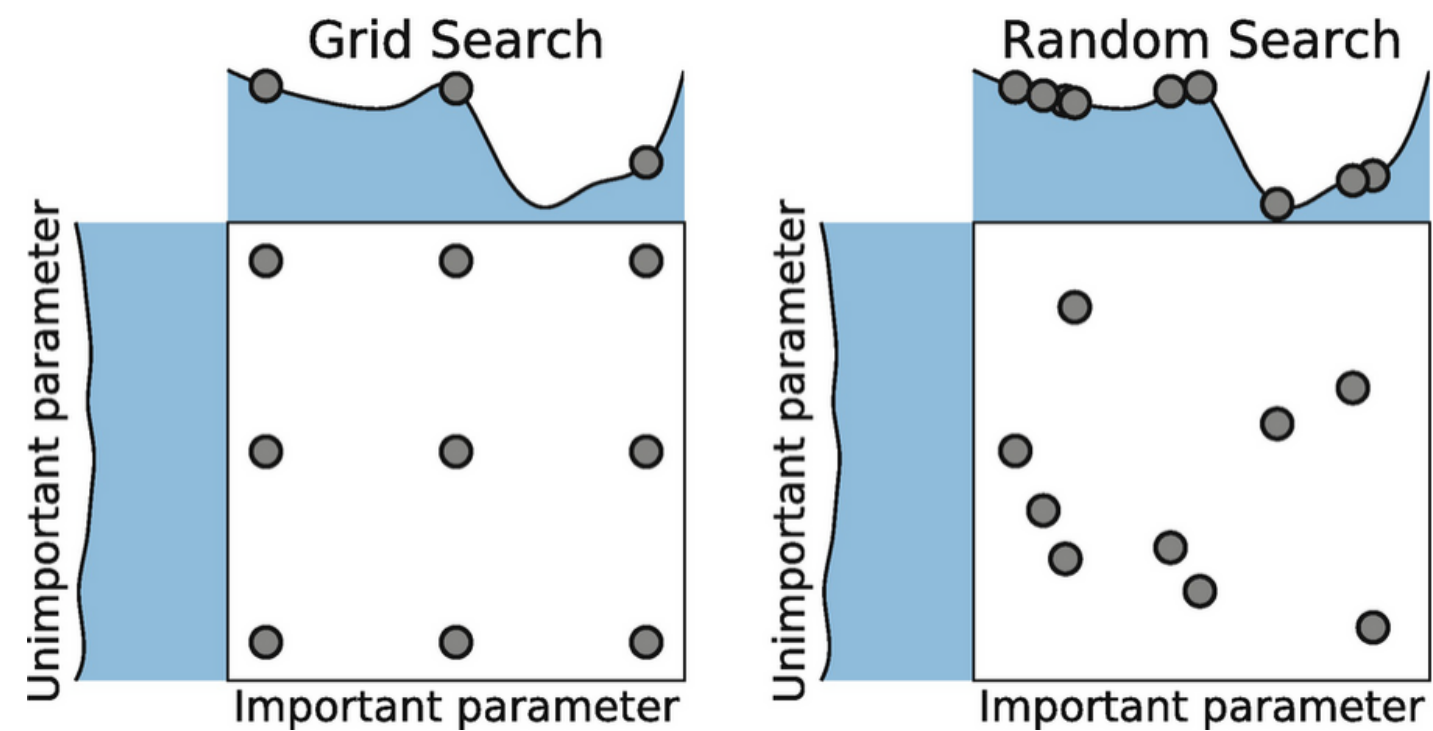
Para criação do modelo base, utilizei uma técnica de 'bootstrapping' para separar os dados de treino e teste. Esta tecnica consiste em criar N amostragens diferentes para treino e teste do modelo aliado com Cross-Validation para melhor aproximar o treinamento do modelo da amostra e reduzir a chance de overfitting. Além disso, removi as variáveis de 'capitalgain' e 'capitalloss' por possuirem mais de 90% dos dados vazios, e a variável 'relationship' por representar a mesma informação da variável 'maritalstatus'.

As métricas utilizadas para o modelo base foram **F1 Score** e **ROC\_AUC**, e, em ambas, o **melhor desempenho** foi do **Gradient Boosting** com **62,7%** e **88,9%** respectivamente.

```
LogisticRegression: 0.618170 (0.014298)
KNeighborsClassifier: 0.581155 (0.017727)
DecisionTreeClassifier: 0.553303 (0.014512)
RandomForestClassifier: 0.603067 (0.013398)
GradientBoostingClassifier: 0.627141 (0.011314)

LogisticRegression: 0.879128 (0.007201)
KNeighborsClassifier: 0.817435 (0.009417)
DecisionTreeClassifier: 0.731140 (0.007733)
RandomForestClassifier: 0.857024 (0.006693)
GradientBoostingClassifier: 0.889096 (0.006281)
```

# HYPERPARAMETER TUNNING



O tuning nos hiperparâmetros foi realizado utilizando RandomizedSearchCV, que consiste em capturar uma amostra do range de parâmetros registrados no teste para determinar o melhor conjunto para treinamento do modelo.

Após o treinamento com o tuning, os melhores hiperparâmetros encontrados foram:

**max\_depth=15**  
**max\_leaf\_nodes=10**  
**min\_samples\_leaf=30**  
**n\_estimators=500**

```
GradientBoostingClassifier(max_depth=15, max_leaf_nodes=10, min_samples_leaf=30,  
                           n_estimators=500)
```

	precision	recall	f1-score	support
<= 50k	0.87	0.92	0.90	5642
> 50k	0.73	0.60	0.66	1899
accuracy			0.84	7541
macro avg	0.80	0.76	0.78	7541
weighted avg	0.84	0.84	0.84	7541

Lista CV - ROC\_AUC: [0.89 0.87 0.9 0.88 0.9 0.87 0.9 0.87 0.89 0.89]  
Média - ROC\_AUC: 0.88

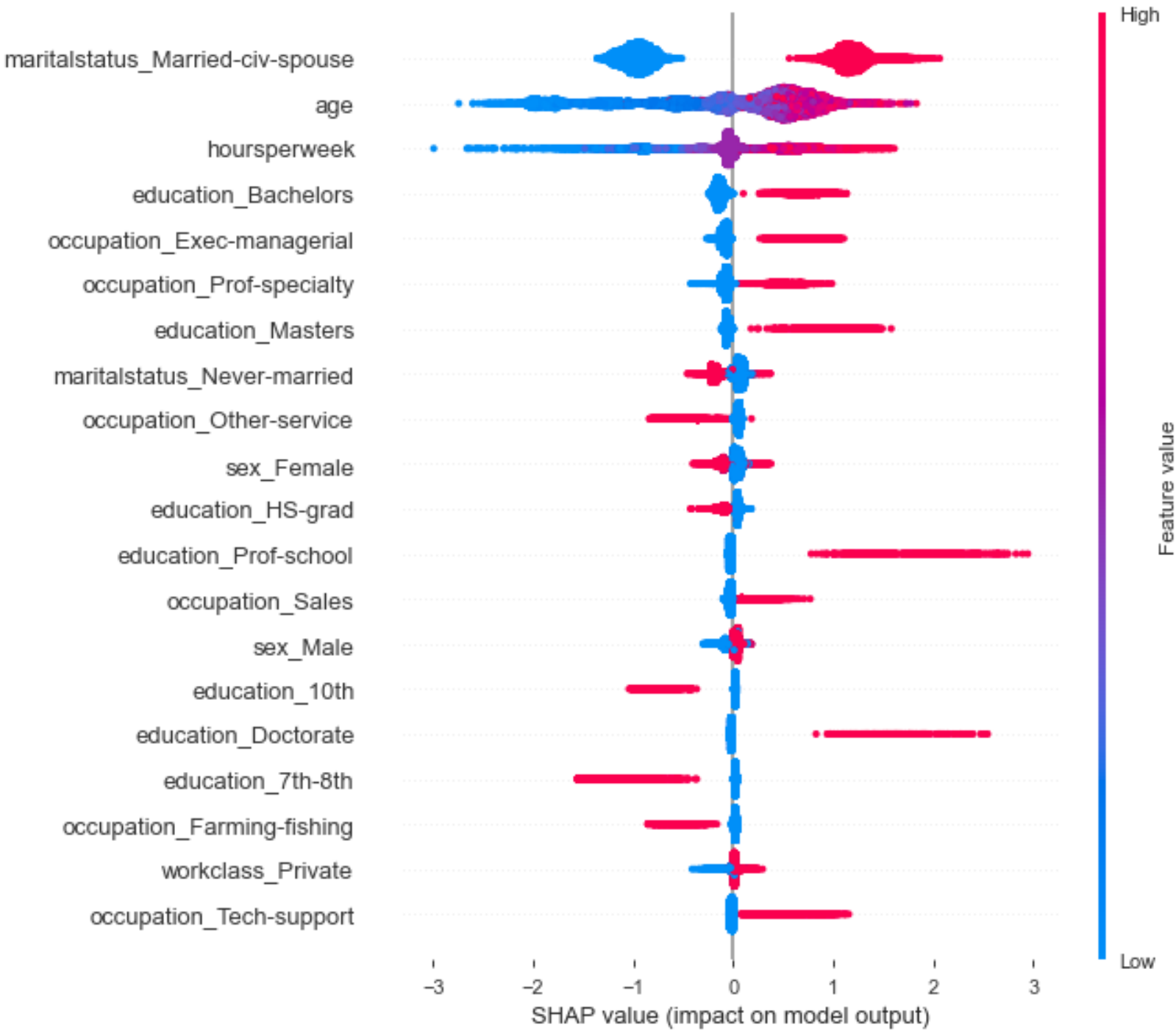
Lista CV - F1: [0.66 0.6 0.65 0.64 0.68 0.63 0.68 0.63 0.67 0.64]  
Média - F1: 0.65

Percebe-se que o modelo previu de maneira mais precisa a classe abaixo de 50k, isso se deu por conta da separação das classes mais homogêneas serem justamente as que se tratavam de ganhos mais baixos

Com cross-validation 10 percebemos que os resultados são bem próximos, com isso podemos descartar o under e overfitting.

Resultados finais:  
**F1: 65%**  
**ROC\_AUC: 88%**

Podemos notar que a ROC\_AUC se manteve nos 88% sem tuning, porém, **ganhamos 3% de F1 Score**, o que pode-se considerar um bom ganho por **reduzir o índice de FP e FN**.



Com o SHAP podemos ver quanto cada variável contribuiu para o resultado final da predição:

- Quanto maior o valor da variável, maior a intensidade da cor vermelha;
- Quanto mais pra direita o dado se encontra, maior a força da predição para a classe acima de 50k.

obs: em caso de variável categórica, 1 é vermelho e 0 é azul.

E como se era esperado por conta da análise multivariada, as variáveis com maior contribuição no modelo foram o status de casado, idades altas, maior média de horas trabalhadas, educação a partir de bacharelado e cargos de executivos e professores especialistas.



DATA SCIENCE

# Obrigado.

---

Leonardo Fiatkoski