Fonte: https://www.census.gov/

O Departamento do Censo dos Estados Unidos (DCEU), oficialmente Bureau of the Census, é a principal agência governamental do sistema estatístico federal do país e responsável por produzir dados sobre a população e economia. O departamento faz parte do Departamento do Comércio e seu diretor é nomeado pelo presidente dos Estados Unidos.

O dataset fornecido pela Clicksign contém 48802 linhas após as transformações e 13 colunas com as seguintes variáveis

age: A idade de um indivíduo

workclass: Um termo geral para representar a situação de emprego de um indivíduo

fnlwgt: Peso final. Em outras palavras, esse é o número de pessoas que o censo acredita que a entrada representa

education: O nível mais alto de educação alcançado por um indivíduo

education-num: Anos de estudo

marital-status: Estado civil de um indivíduo. Married-civ-spouse corresponde a cônjuge civil, enquanto Cônjuge Married-AF-spouse é cônjuge das Forças Armadas

occupation: O tipo geral de ocupação de um indivíduo

relationship: Representa o que esse indivíduo é em relação aos outros. Por exemplo, um indivíduo pode ser Marido. Cada entrada tem apenas um atributo de relacionamento e é um tanto redundante com o estado civil. Podemos nem fazer uso desse atributo;

race: Descrição da raça de um indivíduo

sex: O sexo biológico do indivíduo

capital-gain: Ganhos de capital para um indivíduo

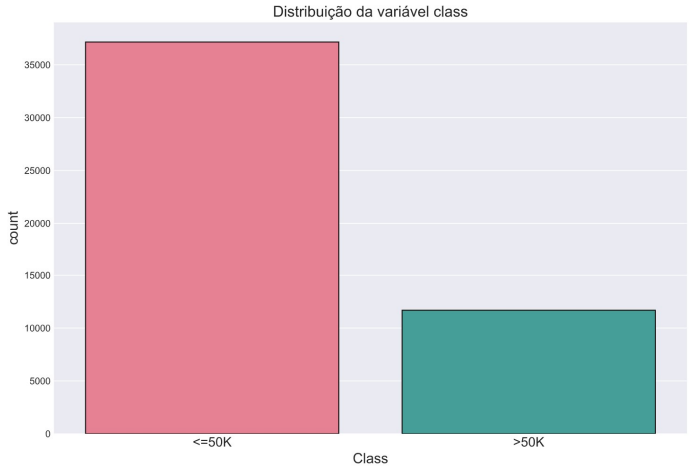
capital-loss: Perda de capital para um indivíduo

hours-per-week: As horas que um indivíduo relatou trabalhar por semana

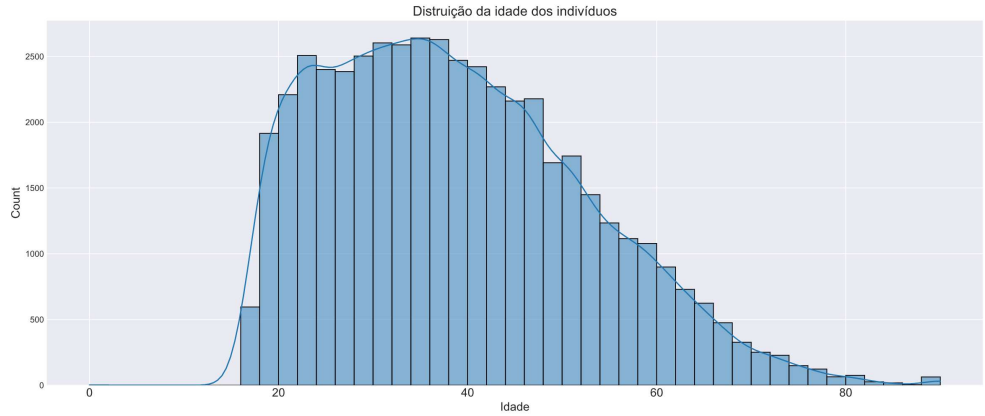
native-country: País de origem de um indivíduo

class: Se um indivíduo ganha ou não mais de $ 50.000 por ano.

Para iniciar vamos checar a distribuição dos indíviduos quanto alguns critérios que podem ser bem relevantes no que diz respeito aos ganhos anuais. A proporção de indivíduos com ganhos acima de $50000 é bem menor o que já é esperado devido a questões de desigualdade.

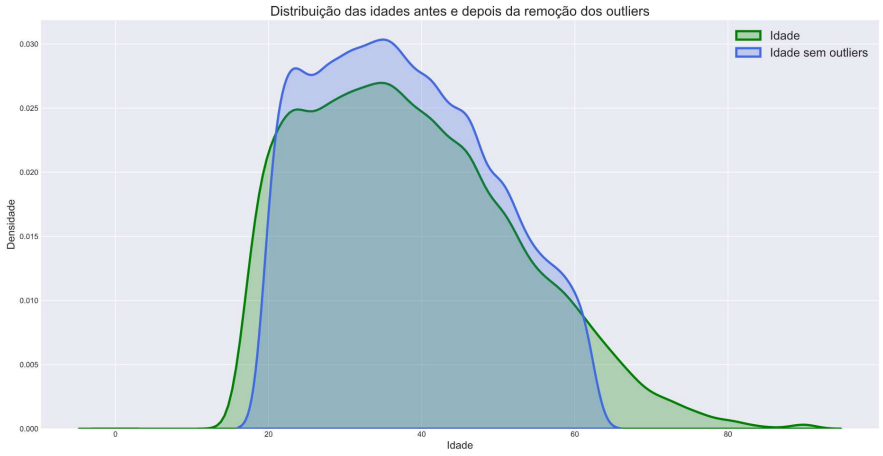


Como sabemos a maioria dos profissionais tende a aumentar os seus ganhos ao decorrer dos anos, mas baseando-se nisso qual seria a distribuição da idade destes indivíduos?

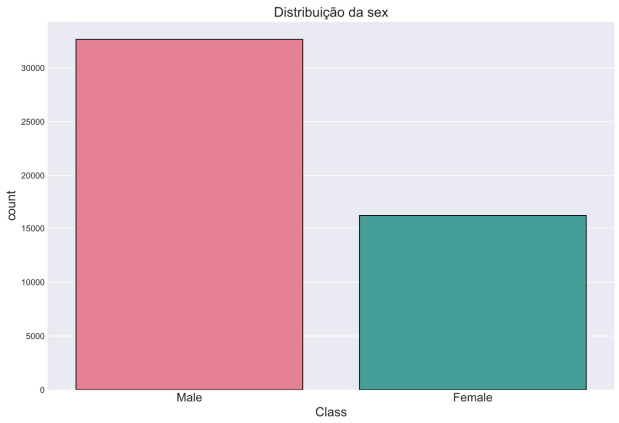


Temos uma distribuição assimetrica positiva onde a grande maioria dos indivíduos são jovens e a medida que a idade aumenta no gráfico a quantidade de indivíduos diminui. Abaixo vemos um outro gráfico da idade comparando o dataset com e sem outliers. No caso deste dataset os outliers são os indivíduos com idade abaixo de 19 e acima de 63 anos.

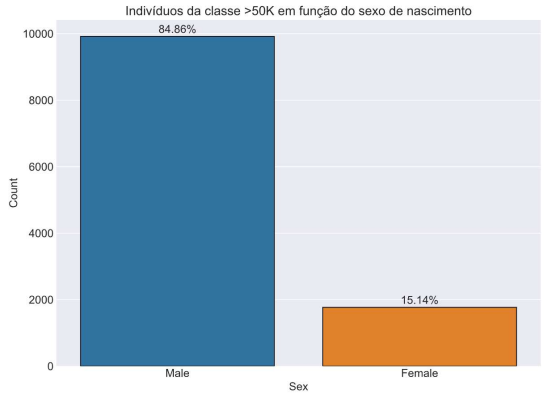
Detalhe, quando se exclui os outliers (área azual) a distribuição fica próxima de uma distribuição normal.

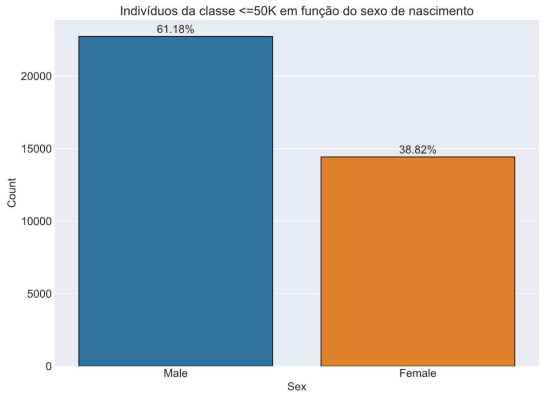


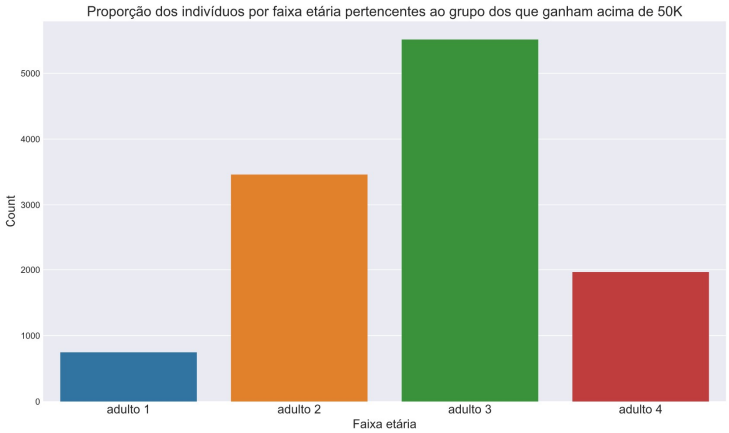
Ainda falando-se das idades dos indivíduos é importante informar que a mediana é 37 anos.



Passando para o sexo dos indivíduos, o dataset possuei uma quantidade bem maior de homens do que mulheres. Nos gráficos abaixo é possível enxergar a desigualdade quando se analisa os gráficos olhando primeiro o da esquerda e depois o da direita.

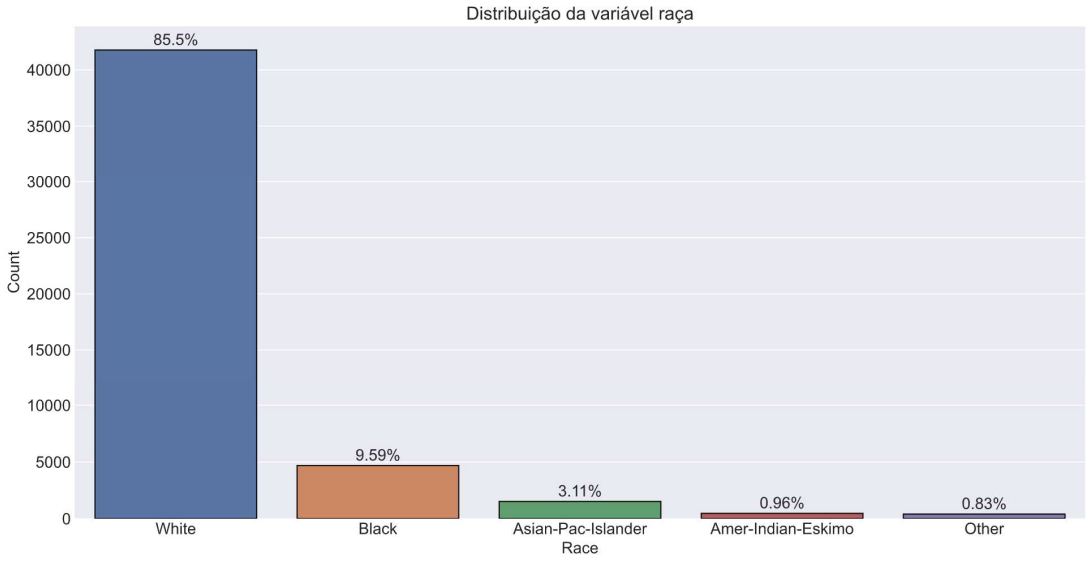






Neste gráfico acima conseguimos ver que o auge dos profissionais acontece na faixa etária adulto 3 que é dos 40 aos 55 anos. Logo, já temos algumas informações úteis sobre o dataset como mulheres realmente ganham menos, o auge profissional acontece com a experiência e está entre os 40 e 50 anos.

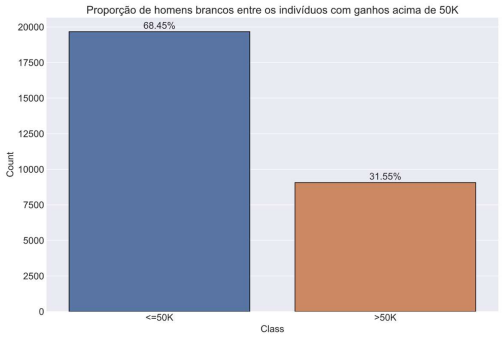
Vamos checar a questão das raças para ver se realmente existe diferenças de ganhos anuais. De início já podermos ver que a quantidade de indivíduos brancos é em torno de 85%, isso é um forte indicação de que as minorias não tem espaço no mercado de trabalho.

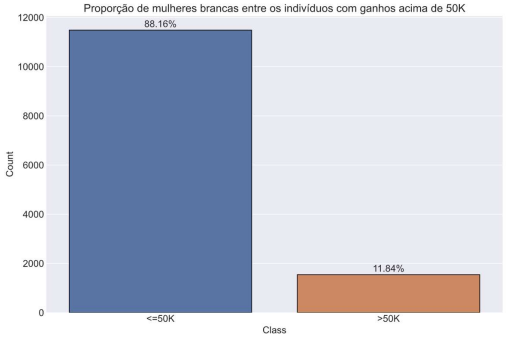


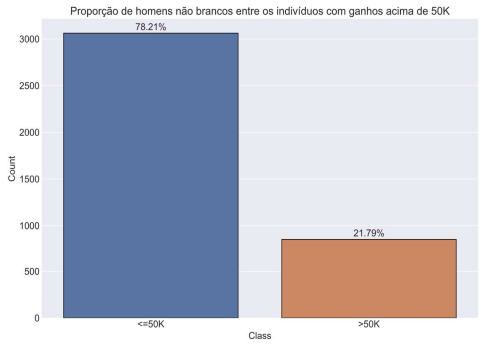
Hoje em dia muito se fala nas minorias e como o mercado de trabalho não dá espaço para estas pessoas, baseando-se nisso podemos tentar confimar se isso é verdade realmente.

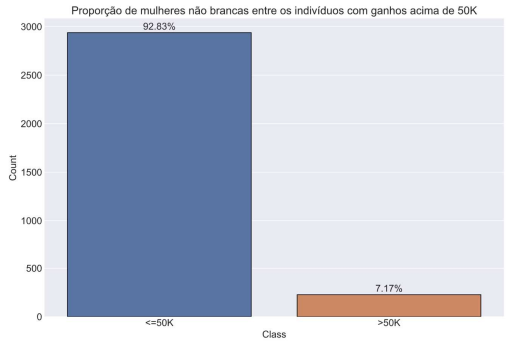
Em mais um gráfico plotado com os dados do dataset observa-se que entre homens e mulheres existe uma diferença grande de ganhos anuais.

O grupo das mulheres que integram o grupo com maiores ganhos é menor e quando se acrescenta a característica da raça cai mais ainda.









Todas estas informações e outras a seguir foram extraídas do dataset guiando-se atráves de algumas hipóteses que foram criadas sem ter olhado os dados antecipadamente para evitar ser tendêncioso. A seguir segue o resumo:

- H1: A maioria dos indivíduos trabalham na iniciativa privada
  VERDADEIRO

- H2: Mulheres possuem menos ganhos anuais
  VERDADEIRO

- H3: Negros e outras minorias possuem menos ganhos anuais
  VERDADEIRO

- H4: Os anos de escola não necessariamente influenciam nos ganhos anuais
  FALSO

- H5: Maior número de horas de trabalho por semana não necessariamente indicam maiores ganhos
  VERDADEIRO

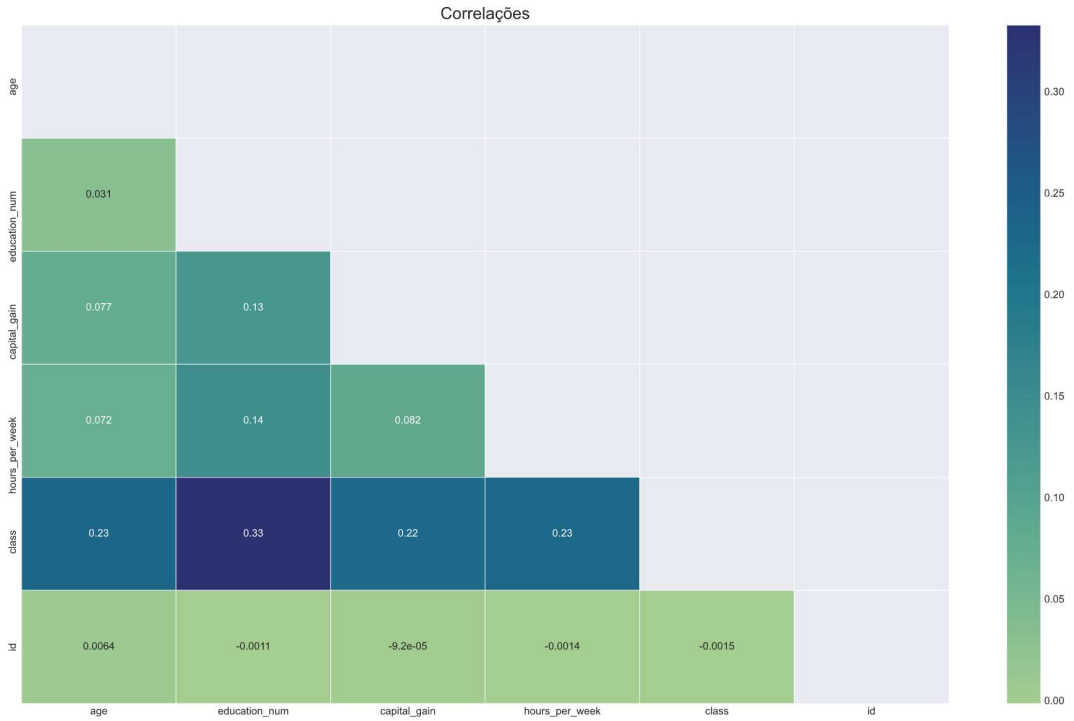
- H6: Profissões ligadas a área de tecnologia proporcionam maiores ganhos
  FALSO

- H7: Indivíduos com alguma relação com as forças armadas possuem maiores ganhos anuais
  FALSO

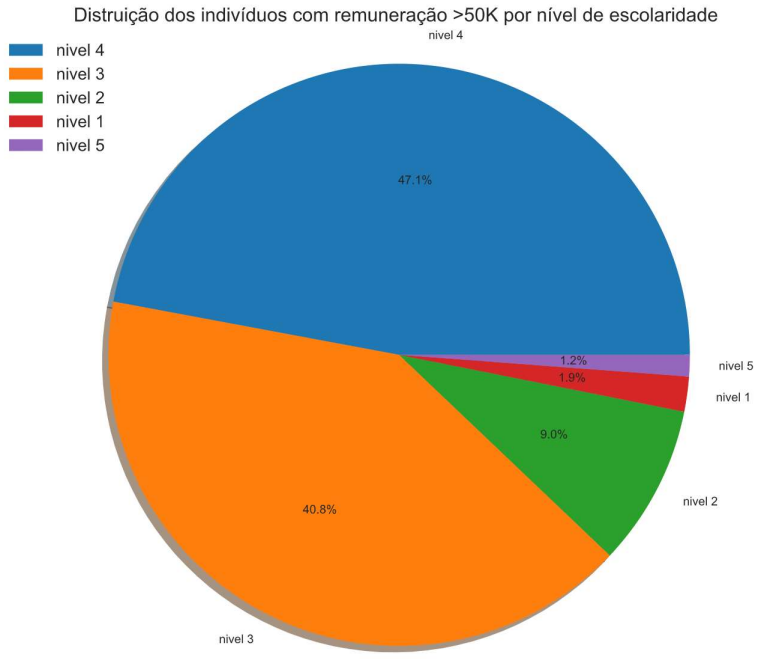
- H8: Indivíduos fora de um relacionamento possuem maiores ganhos
  FALSO

- H9: Indivíduos americanos possuem maiores ganhos comparados a outras nacionalidades
  VERDADEIRO

Uma análise de correlações foi feita para ver quais variáveis numéricas possuem maior correlação com a variável resposta. Os anos de estudos que são representados pela variável education_num foi a variável com maior correlação.



E isso se confirmou no gráfico de pizza que mostra que os indivíduos dos níveis 3 e 4 que seriam entre 8 e 14 anos de estudos compõem a maior parte do grádico.



Conclusão

O objetivo desta breve análise dos dados foi tentar confirmar ou reprovar algumas hipóteses que são baseadas no que se costuma ouvir sobre as características do mercado de trabalho americano. Idéias já conhecidas como americanos ganham mais, mulheres e integrantes de grupos classificados como minorias possuem menores ganhos, anos de estudos refletem em melhores ganhos e o auge profissional é entre os 30 ou 40 sempre são frases famosas que ouvimos por aí.

A grande maioria se confirmou atráves da breve análise dos dados do Censo americano fornecido pela Clicksign.

O notebook contém todos os gráficos e comentários da análise dos dados lembrando que houveram transformações nesses dados em uma etapa anterior e esses dados foram hospedados em um banco de dados RDS na Amazon.

Em uma situação em que esses dados fossem utilizados para uma predição, a idéia seria criar um modelo de classificação para classificar as pessoas que ganhariam mais ou menos de 50K. Para finalizar, nesta situação hipotética seria bom que o dataset fosse ampliado para um número maior do que as atuais 48000 linhas.