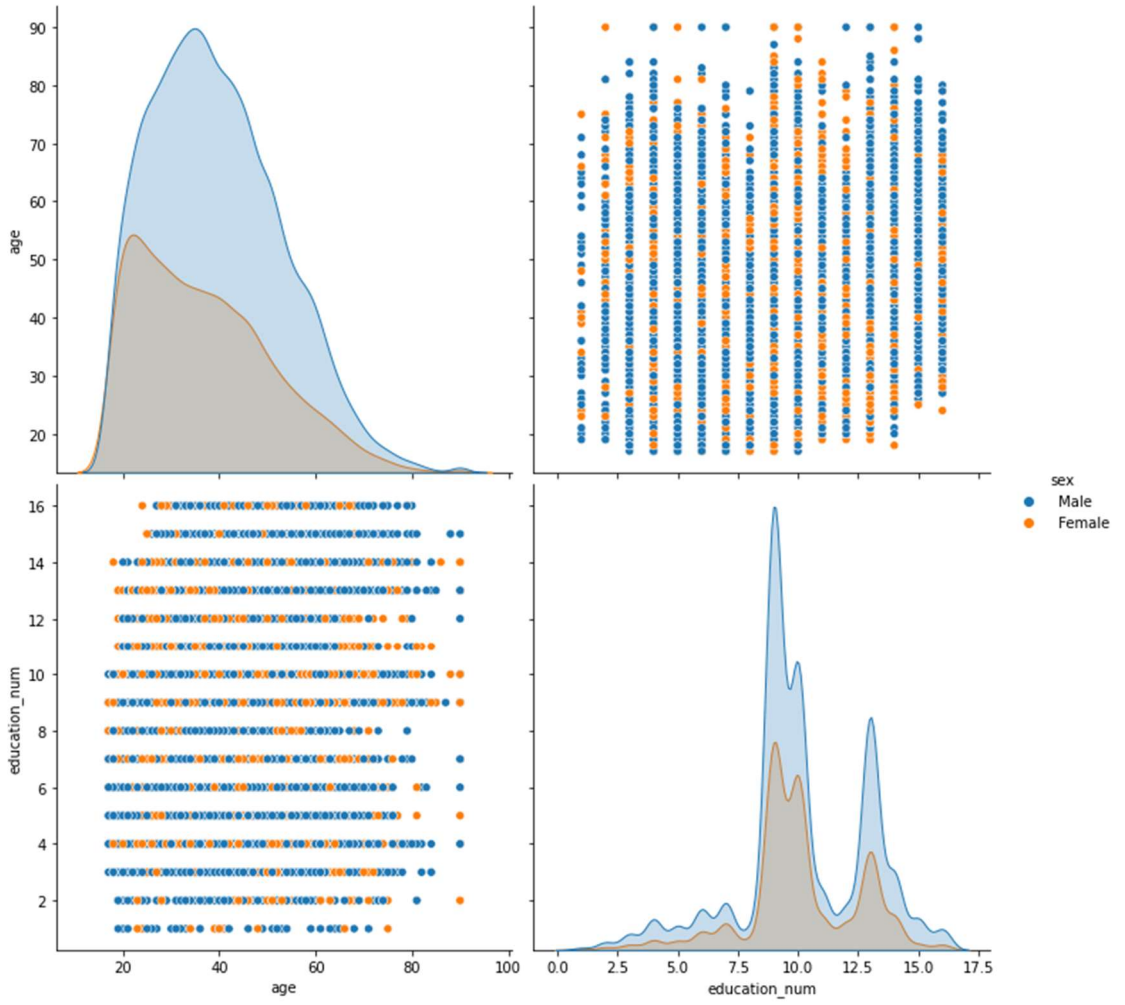
Descreva graficamente os dados disponíveis, apresentando as principais estatísticas descritivas. Comente o porquê da escolha dessas estatísticas.

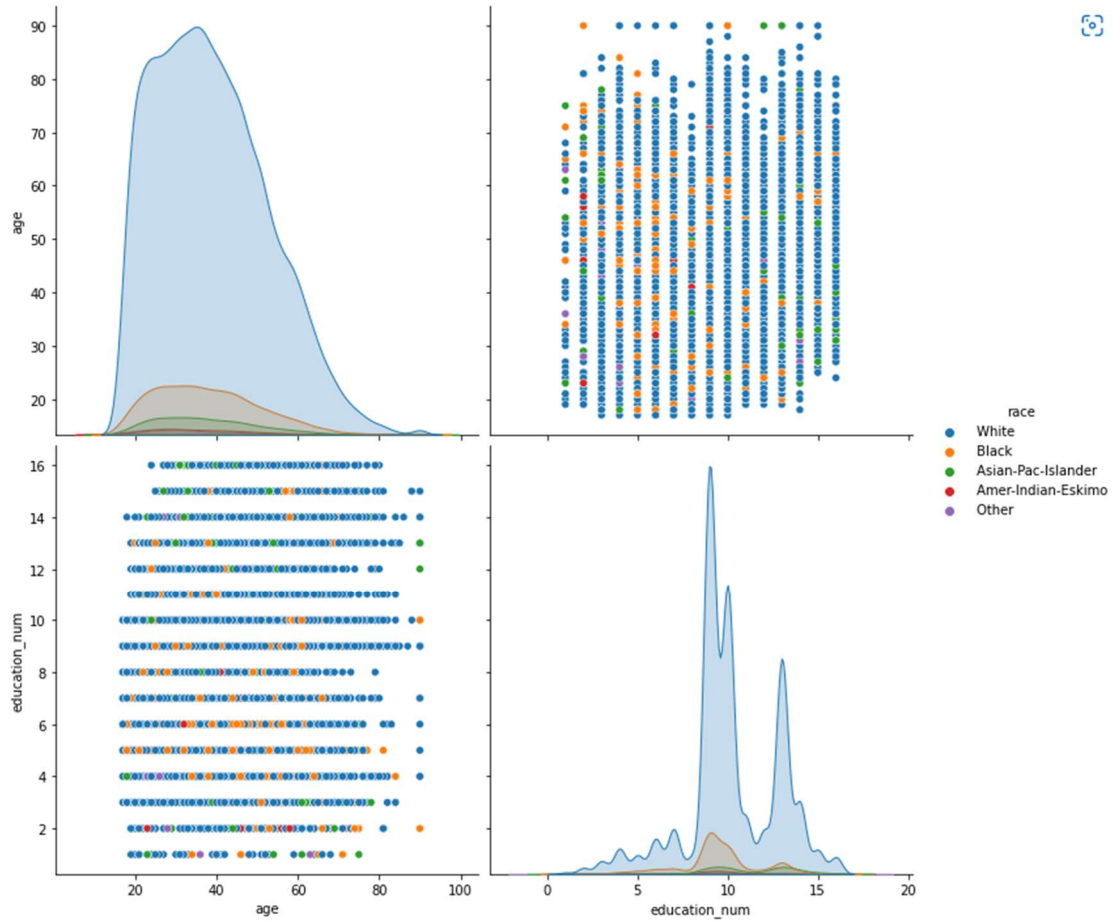
Neste caso foi feita uma análise entre os dados comparando os sexos disponíveis e os relacionando com a idade e educação afim de obser se há desifguadade entre gêneros na amostra. Observou-se um maior número de homens na amostra e um nível de educação que acompanha o gráfico populacional.

```
g = sns.pairplot(raw_data, hue = 'sex', height = 5, vars = ['age', 'education_num'])
```



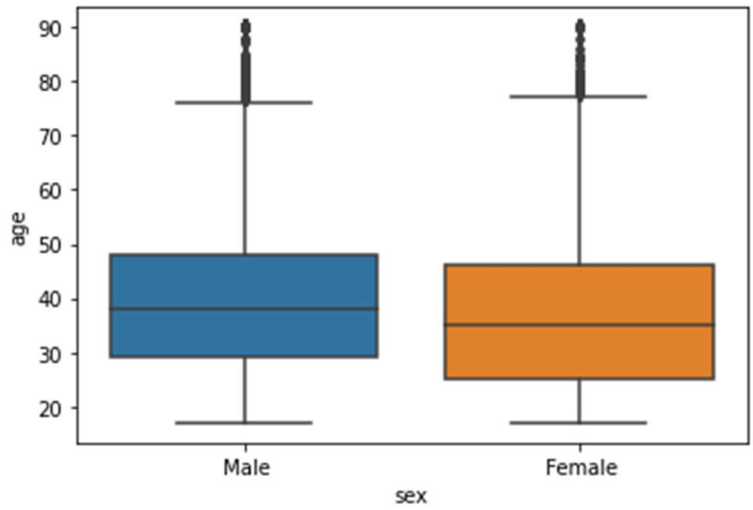
Semelhante a ideia anterior apresentada foi verificado graficamente o número de indivíduos de cada raça em comparação com a idade e edução. Foi observado um número muito superior de pessoas brancas, assim como observa-se que pessoas com alguma faculdade ou bacharéis dominam a população.

```
g = sns.pairplot(raw_data, hue = 'race', height = 5, vars = ['age', 'education_num'])
```

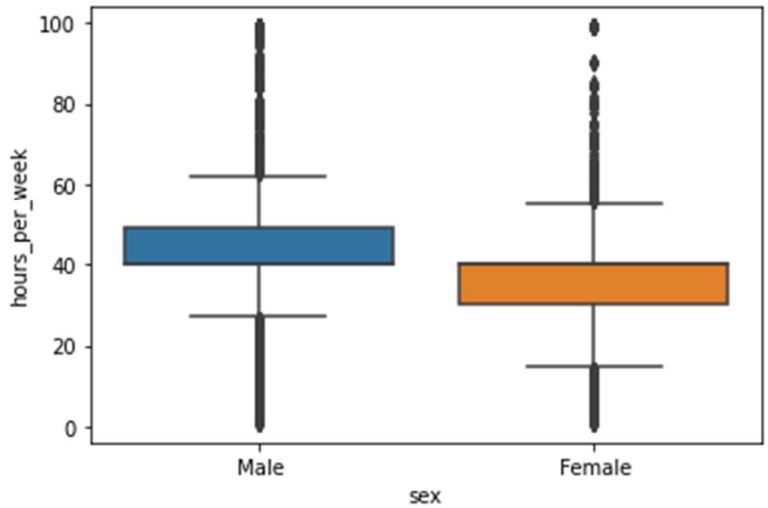


Após estas análises foi realizada a verificação de algumas populações com o propósito de de buscara alguma similaridade ou descrepância entre os dados. Foi observado que em todas as análises, independente de resultado, temos muitos outliers que devriam ser tratados para encontrar-mos um resultado mais apurado.

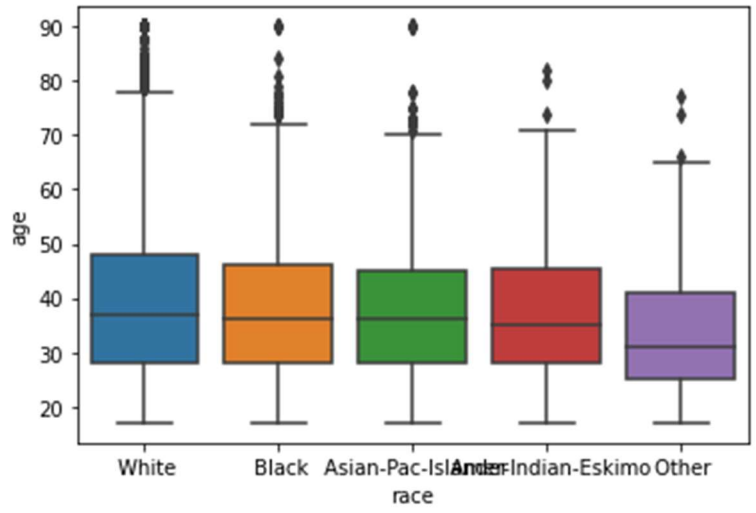
```
g =  sns.boxplot(x = 'sex', y = 'age', data = raw_data)
```



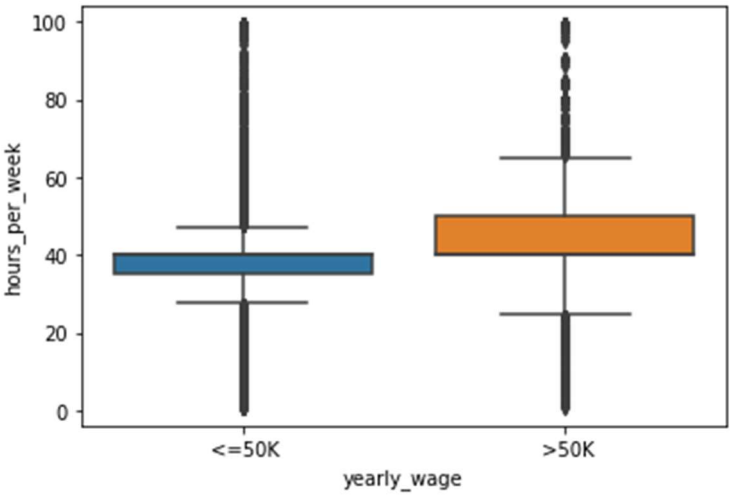

```
p =  g =  sns.boxplot(x = 'sex', y = 'hours_per_week', data = raw_data)
```



```
g =  sns.boxplot(x = 'race', y = 'age', data = raw_data)
```




```
g =  sns.boxplot(x = 'yearly_wage', y = 'hours_per_week', data = raw_data)
```



Explique como você faria a previsão do **salário** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

Para previsão do salário reduzimos a amostra para as seguintes colunas: 'age', 'education_num', 'occupation', 'sex', 'hours_per_week', 'native_country'.
O propópisto era ter confiança nos dados, mas que ao mesmo tempo reduzíssemos o peso da aplicação mantendo somente informações relacionadas a ganhos por trabalho.
‘age’ sabe-se que idade é um fator determinante na vida profissional e que por muitas vezes está relacionada a experiência e tão logo salário. ‘
‘education_num’ foi utilizado a coluna education_num pois já estava preparada com numerais, o que facilitaria a análise e sabe-se que a educação pode ser fator determinante de salário.
‘occupation’ mais um fator determiante de salário.
‘sex’ utilizado por observcarmos um variância na amostra entre sexos.
‘hours_per_week’ necessário por horas trabalhadas podem refletir em ganhos.
‘native_country’ necessário pois há desparidade entre moedas e salários entre países.

Utilizamos a regressão linear múltipla, a qual tem o propósito de analisar diversos dados e o relacionamento entre eles. Como nosso objetivo era exatemente este, correlacionar os dados de variáveis contínuas e categóricas e retornar um salário, foi o método ideal.