Visualização de dados - Tarefa 2

Aluno: Daniel de Miranda Almeida, Matrícula: 241708065

Conjunto de dados

Conjunto de dados escolhido: County-Level Trends in Outcomes (1978-1992 Cohorts) by Parental Income, Race, and Gender

Ele contém informações sobre a mobilidade social nos EUA, com diversas variáveis e agrupamentos:

- de gênero: masculino e feminino
- de raça: brancos, negros, hispânicos, asiáticos e AIAN's (American Indian Alaska Native)
- de renda parental: 1º, 25º, 50º, 75º e 100º percentis da distribuição nacional de renda

Além disso, o autor da base define diferentes *outcomes* nas distribuições:

- kfr: classificação percentil média (relativa a nascimento no mesmo ano) na distribuição nacional de renda familiar (i.e., renda pessoal acrescida da renda do cônjuge) medida aos 27 anos.
- kir: classificação percentil média (relativa a nascimento no mesmo ano) na distribuição nacional de renda pessoal (renda própria) medida aos 27 anos.
- emp: fração de pessoas empregadas aos 27 anos.
- kfi: renda familiar média aos 27 anos (em $ de 2023) obtida convertendo classificações percentil (kfr) para dólares usando a conversão percentil-dólar.
- kii: renda individual média aos 27 anos (em $ de 2023) obtida convertendo classificações percentil (kir) para dólares usando a conversão percentil-dólar.

Nas minhas análises vou me ater ao uso das medidas de kfi, porque acredito que só ela seja mais intuitiva no entendimento da mobilidade social ocorrendo.

Objetivo

Explorar como a mobilidade social se deu em diferentes grupos, e se essa mobilidade social ao longo dos anos trouxe alguma diminuição na desigualdade de riqueza.

Perguntas a responder

1. Quais são as diferenças existentes entre os grupos?
    a. Homens ascendem mais que mulheres?
    b. Qual raça tem maior tendência de ascensão?
2. Houve alguma redução na desigualdade social/distâncias entre classes?

Olhando para os dados

```
==============================
Number of columns: 540
Number of columns with null counts above average
340
Percentage of columsn with count of null values above average: 62.96 %
==============================
Number of rows: 3191
average null count: 1588.7740740740742
```

Dando uma primeira olhada nas colunas, é possível perceber que são muitas: 540. Mais do que isso, o conjunto de dados tem uma quantidade enorme de valores nulos, que se devem provavelmente à dados faltantes sobre um determinado grupo. São 340 (muito mais que a metade) colunas com o número de valores nulos acima da média. Isso significa que a maioria das linhas nessas colunas são nulas.

```
==============================
Removing columns with aian
Number of rows: 3191
average null count: 1340.691111111111
standard deviation of null count: 1009.1005764627582
==============================
Removing columns with asian
```

```
Number of rows: 3191
average null count: 1374.108888888889
standard deviation of null count: 1054.4950233470977
==============================
Removing columns with hisp
Number of rows: 3191
average null count: 1538.9577777777777
standard deviation of null count: 1169.947612896216
==============================
Removing columns with black
Number of rows: 3191
average null count: 1533.3666666666666
standard deviation of null count: 1170.4504248812407
==============================
Removing columns with white
Number of rows: 3191
average null count: 1865.0333333333333
standard deviation of null count: 965.2266619342247
```

Checando os valores nulos quando removemos diferentes raças do conjunto de dados, vemos um certo padrão: quando tiramos raças menos representativas, como AIAN e asiáticos, temos menos valores nulos e uma tendência de valores mais divergentes entre si. Essa divergência é intensificada quando tiramos os hispânicos e negros - um "meio termo", em termos de representatividade, entre brancos e AIAN's e asiáticos, creio eu - e tem seu menor valor (junto do maior valor médio de linhas nulas) quando retiramos os brancos. Isso mostra, entre outras coisas, que os dados da população branca são sensivelmente mais completos que os das demais raças.

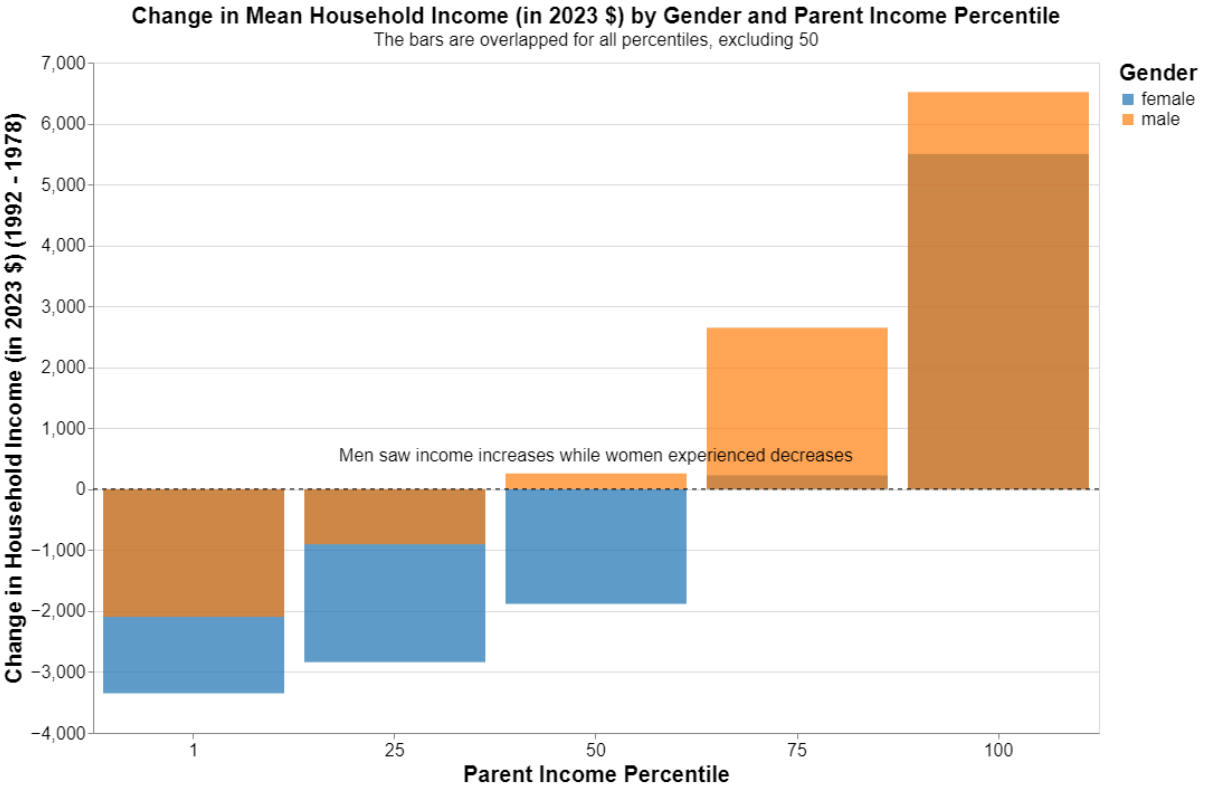
Nas linhas acima, `average null count` representa a quantidade média de linhas com valores nulos quando retiramos uma determinada raça da análise e `standard deviation of null count` é o desvio padrão de mesma distribuição.

Sanity check: os valores de kfi precisam ser todos maiores que 0, uma vez que reprentam renda média

Nenhum valor encontrado, então a variável está se comportando como esperado.

Investigando perguntas

Pergunta 1.1: Pessoas de diferentes sexos tiveram mais ou menos sucesso na mobilidade social?

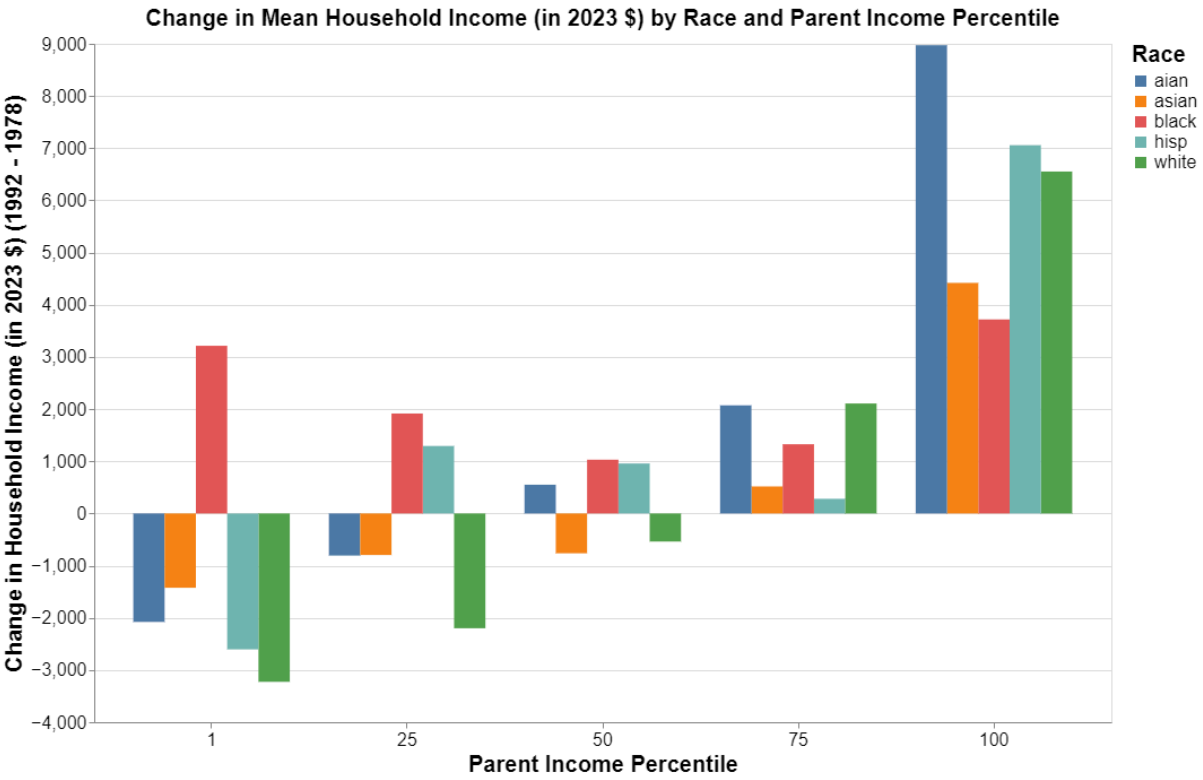


Com o gráfico, podemos ver que claramente os homens tiveram mais sorte que as mulheres para todos os percentis de renda parental: quando os valores são positivos, as barras do gênero masculino estão mais altas; quando os valores são negativos, as barras para o gênero feminino são mais baixas.

Isso indica que de fato, independentemente da renda parental, a mudança é sempre menor ou pior para pessoas do sexo feminino.

Além disso, outra coisa que é possível perceber no gráfico é uma ideia de como os percentis de renda parental influenciam na mobilidade social: quanto maior a renda parental, maior a tendência de uma ascensão. Como a divisão é feita por percentis, é possível perceber que para 50% da população na verdade houve uma tendência muito maior na diminuição da renda, o que indica que aqueles que tiveram aumentos expressivos na renda são os que já estavam nas classes mais abastadas.

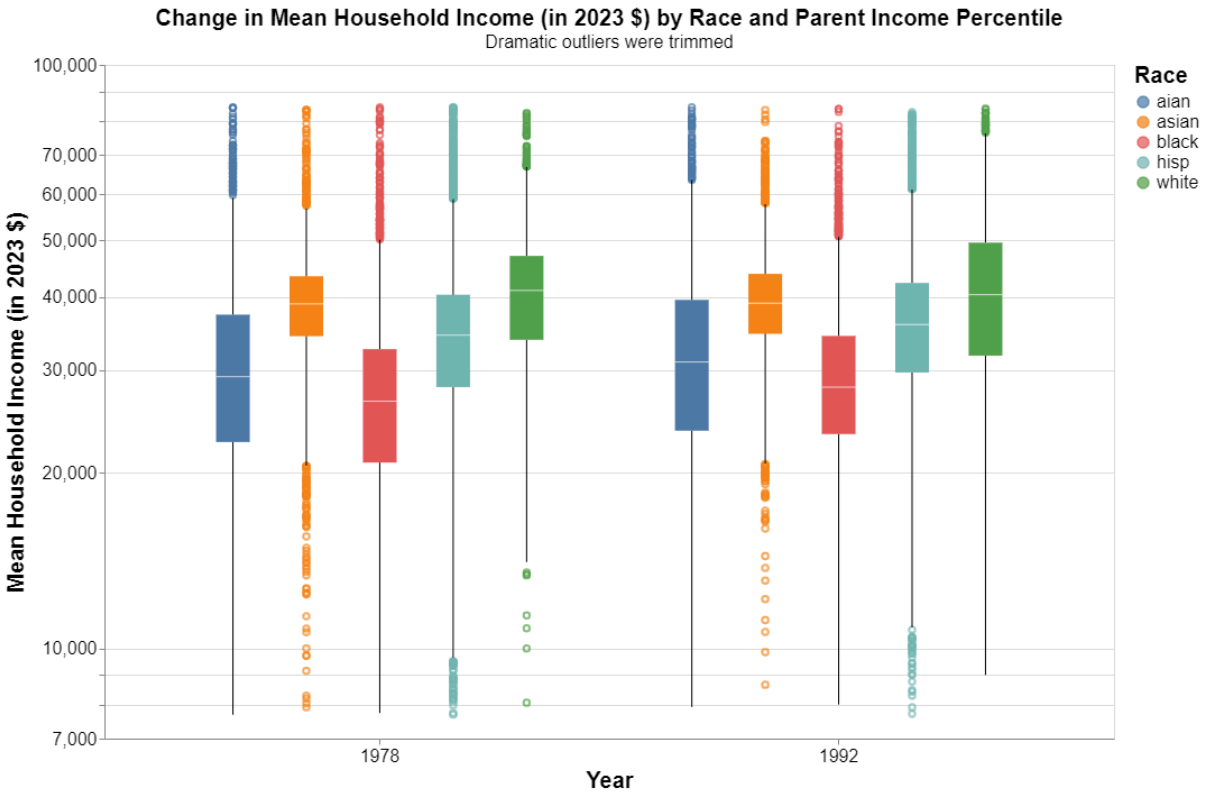
Pergunta 1.2: Pessoas de diferentes raças tiveram mais ou menos sucesso na mobilidade social?



Esse gráfico já traz mais informações interessantes. Primeiramente, assim como no gráfico com uma divisão somente por gênero, quanto maior a renda parental, maior a tendência de mobilidade.

Mas outro fato interessante surge: contrariando os dados para as outras raças, os negros são os únicos que tem um aumento médio na renda familiar independentemente da renda parental, um aumento que é expressivamente grande nos 1º e 25º percentis de renda parental!

Pergunta 2: Houve alguma redução na desigualdade social/distâncias entre classes?



Definir se houve uma mudança na desigualdade social é algo difícil de se fazer. Como existem abismos de diferença entre os mais ricos e os mais pobres, as distância são muito grandes. Mesmo com a retirada dos outliers mais dramáticos e o uso de uma escala logaritmica, ainda é possível perceber muitos outliers em todas as distribuiçõs de renda familiar média.

Como seria de se imaginar, as distribuições de renda para pessoas brancas tem quartis mais altos, enquanto para AIAN's e negros principalmente, os valores de quartis são todos mais baixos. Ao ponto em que os 75% mais pobres da população negra em 1978 tinha menos renda que os 25% mais pobres da população branca.

De maneira geral, as distribuições mudaram pouquíssimo de 1978 para 1992. É possível destacar que os outliers na porção inferior do boxplot para pessoas brancas somem ao longo do tempo, o que pode indicar uma ascensão de pessoas mais probres mas ou talvez um agravamento nas distâncias sociais.

Conclusão

As conclusões em que podemos chegar é de que:

- A maior parte da população teve uma diminuição na renda de 1978 para 1992, com apenas os que já eram mais ricos ficando mais ricos.

- Curiosamente hispânicos e principalmente negros tiveram um aumento na renda familiar média entre as pessoas com menores rendas parentais, o que mostra uma ascensão social bastante expressiva pra esses dois grupos
- O cenário de desigualdade praticamente se manteve o mesmo ao longo dos 14 anos de dados.