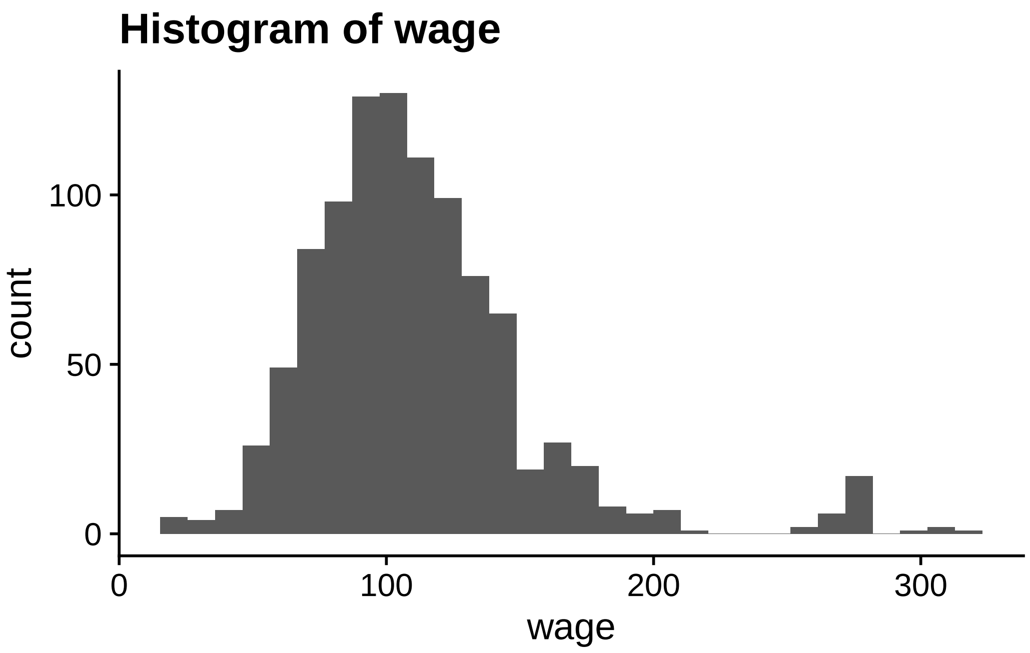
Initial data exploration

Sep 22, 2022

```
Rows: 1,000
Columns: 11
$ year       <dbl> 2003, 2006, 2003, 2005, 2007, 2006, 2005, 2004, 2005, 2008,~
$ age        <dbl> 45, 50, 38, 40, 40, 38, 49, 34, 51, 33, 34, 25, 27, 50, 35,~
$ maritl     <chr> "2. Married", "2. Married", "2. Married", "4. Divorced", "2~
$ race       <chr> "1. White", "1. White", "3. Asian", "1. White", "1. White",~
$ education  <chr> "3. Some College", "5. Advanced Degree", "4. College Grad",~
$ region     <chr> "2. Middle Atlantic", "2. Middle Atlantic", "2. Middle Atla~
$ jobclass   <chr> "1. Industrial", "2. Information", "2. Information", "2. In~
$ health     <chr> "1. <=Good", "2. >=Very Good", "2. >=Very Good", "2. >=Very~
$ health_ins <chr> "1. Yes", "2. No", "2. No", "2. No", "2. No", "2. No", "1. ~
$ logwage    <dbl> 4.875061, 5.360552, 5.301030, 3.920123, 5.079181, 4.544068,~
$ wage       <dbl> 130.98218, 212.84235, 200.54326, 50.40666, 160.64248, 94.07~
```

Table 1: Features with zero variance

feature	variance
region	0



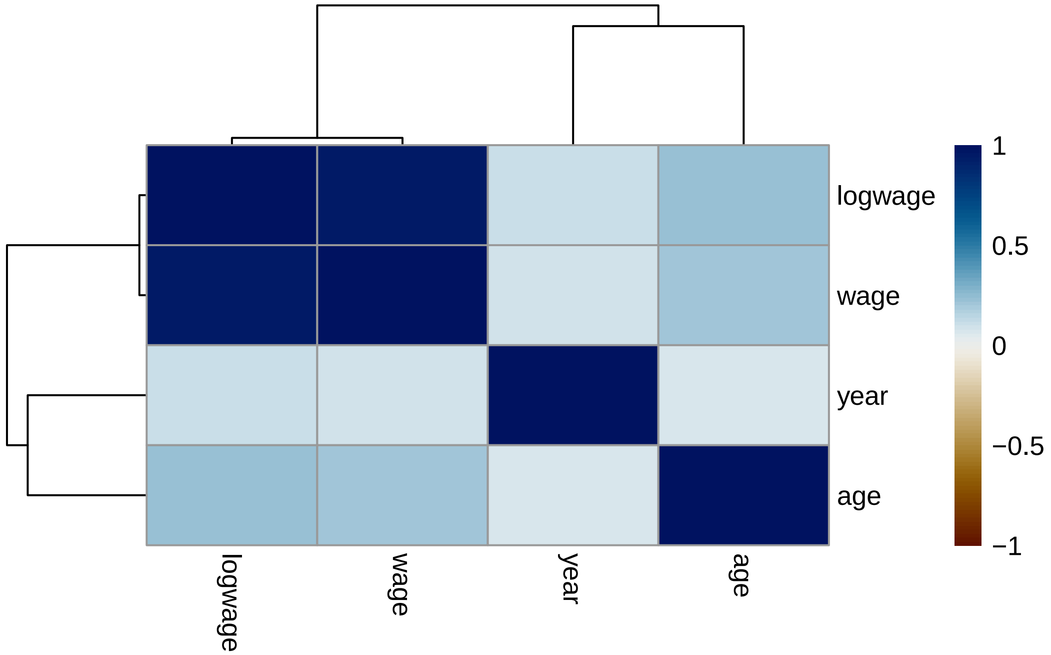


Figure 1: Heatmap of numerical variables

Table 2: Top skewness values (in absolute value)

feature	skewness	kurtosis
wage	1.7064384	4.7947921
year	0.1908600	-1.2004272
logwage	-0.1745861	1.8725763
age	0.0855479	-0.5871229

Table 3: Top kurtosis values (in absolute value)

feature	kurtosis	skewness
wage	4.7947921	1.7064384
logwage	1.8725763	-0.1745861
year	-1.2004272	0.1908600
age	-0.5871229	0.0855479