Project 0 - Solution

Jon Moreno-Medina

Question 1

R code that calculate how much is the square root of 12.

Solution

To estimate the square root of 12, we just use the code below, with the function `sqrt()`:

```
sqrt(12)
```

```
[1] 3.464102
```

Question 2

R code that loads the data for project 1 into a dataframe called 'atlas'.

Solution

As showed in class, we just need to go to the site of the course, and copy the line to load the data at the beginning of Project 1:

```
atlas <- readRDS(gzcon(url("https://raw.githubusercontent.com/jrm87/ECO3253_fall2023/maste
```

Here I will also show how the data looks in general, but using the function `head()`:

```
head(atlas)
```

```
  tract county state    cz     czname hhinc_mean2000 mean_commutetime2000
1 20100      1     1 11101 Montgomery       68638.73             26.17191
2 20200      1     1 11101 Montgomery       57242.51             24.80671
3 20300      1     1 11101 Montgomery       75647.73             25.32253
4 20400      1     1 11101 Montgomery       74852.05             22.96535
5 20500      1     1 11101 Montgomery       96174.77             26.22235
6 20600      1     1 11101 Montgomery       68095.77             21.63042
  frac_coll_plus2010 frac_coll_plus2000 foreign_share2010 med_hhinc2016
1          0.2544283          0.1564792       0.009950249         66000
2          0.2671937          0.1469317       0.016336633         41107
3          0.1641593          0.2244131       0.027095681         51250
4          0.2527439          0.2304688       0.015082644         52704
5          0.3750627          0.3211544       0.046488225         52463
6          0.2394235          0.1607055       0.024985302         63750
  med_hhinc1990 popdensity2000 poor_share2010 poor_share2000 poor_share1990
1         27375       195.7238     0.10503040     0.12681565     0.09887157
2         19000       566.3814     0.14759035     0.22705820     0.19833852
3         29419       624.1968     0.08038494     0.07664010     0.11398072
4         37891       713.8040     0.06322314     0.04548451     0.06789701
5         41516       529.9303     0.05956933     0.03679151     0.05473420
6         29000       408.3740     0.10523222     0.15216105     0.17814240
  share_black2010 share_hisp2010 share_asian2010 share_black2000
1      0.11924686     0.02301255     0.004707113      0.07548152
2      0.56497693     0.03456221     0.002304147      0.62209302
3      0.19804329     0.02579306     0.004743552      0.14914645
4      0.04673963     0.01937984     0.003647971      0.02589991
5      0.13969906     0.03297418     0.026032491      0.06009934
6      0.21155943     0.04798255     0.001635769      0.16903494
  share_white2000 share_hisp2000 share_asian2000 gsmn_math_g3_2013
1       0.8969287    0.006246747     0.003643936          2.759864
2       0.3546512    0.008456660     0.003171247          2.759864
3       0.8200060    0.016471997     0.003893381          2.759864
4       0.9378841    0.022168569     0.007288219          2.759864
5       0.8970199    0.015728477     0.010596027          2.759864
6       0.7992895    0.019538188     0.001480166          2.759864
  rent_twobed2015 singleparent_share2010 singleparent_share1990
1              NA              0.1139240             0.18118466
2             907              0.4884615             0.35245901
3             583              0.2280702             0.12590799
4             713              0.2275335             0.12676056
5             923              0.2596976             0.07436399
6             765              0.3163717             0.23800738
  singleparent_share2000 traveltime15_2010   emp2000 mail_return_rate2010
```

1	0.2509804	0.2730337	0.5673077	83.5
2	0.3925234	0.1520396	0.4931694	81.3
3	0.2448560	0.2055336	0.5785598	79.5
4	0.1907216	0.3506735	0.5965011	83.5
5	0.1680000	0.2504962	0.6612682	77.3
6	0.2889344	0.3416459	0.6426789	82.8

	ln_wage_growth_hs_grad	jobs_total_5mi_2015	jobs_highpay_5mi_2015
1	0.03823291	10109	3396
2	0.08930562	9948	3328
3	-0.17774254	10387	3230
4	-0.07231081	12933	3635
5	-0.09613968	12933	3635
6	-0.04856208	9193	3052

	nonwhite_share2010	popdensity2010	ann_avg_job_growth_2004_2013
1	0.16265690	504.7518	-0.006769223
2	0.61105990	1682.1705	-0.004253248
3	0.24755412	1633.4139	0.014217778
4	0.08116734	1780.0325	-0.019840827
5	0.21623629	2446.2622	0.018626856
6	0.27153760	1184.3721	-0.051587597

	job_density_2013	kfr_natam_p25	kfr_natam_p75	kfr_natam_p100	kfr_asian_p25
1	92.13305	NA	NA	NA	NA
2	971.31787	NA	NA	NA	NA
3	340.92007	NA	NA	NA	NA
4	207.38637	NA	NA	NA	NA
5	800.27264	NA	NA	NA	NA
6	336.77753	NA	NA	NA	NA

	kfr_asian_p75	kfr_asian_p100	kfr_black_p25	kfr_black_p75	kfr_black_p100
1	NA	NA	26819.20	45925.62	84689.84
2	NA	NA	18138.11	33841.53	60512.21
3	NA	NA	20514.96	34133.12	56515.76
4	NA	NA	12882.58	40333.60	105250.12
5	NA	NA	26594.34	42574.89	72564.73
6	NA	NA	19108.02	26062.19	35736.69

	kfr_hisp_p25	kfr_hisp_p75	kfr_hisp_p100	kfr_pooled_p25	kfr_pooled_p75
1	NA	NA	NA	27620.96	51530.51
2	NA	NA	NA	22303.06	46649.74
3	NA	NA	NA	28215.48	50753.54
4	26363.10	67532.27	NA	33330.90	52337.20
5	17233.77	44642.39	93976.28	34632.66	57007.41
6	NA	NA	NA	23583.01	47734.75

	kfr_pooled_p100	kfr_white_p25	kfr_white_p75	kfr_white_p100	count_pooled
1	78921.50	30327.95	50820.14	75126.03	519

```
2        74225.37      42188.81      54239.12      66645.70           530
3        76055.36      33670.45      51579.38      71990.97           960
4        72586.48      34181.05      52847.86      74330.25          1123
5        81792.41      39540.15      58699.04      80415.09          1867
6        75188.00      27834.53      51198.23      80143.85           994
  count_white count_black count_asian count_hisp count_natam
1         457          42           3          4           6
2         173         336           1          5           1
3         774         151           1         21           2
4        1033          40           6         37           0
5        1626         137          13         39           8
6         756         198           2         19           2
```

As can be seen, the data has 73,278 observations, one for each neighborhood in the US. It also includes 62 variables for each observation.

Question 3

R code that estimates the mean and standard deviation of the average income of children of parents in the percentile 25 and 75.

Solution

Recall that in this dataset the average income of children with parents in percentile 25 and 75 are `kfr_pooled_p25`and `kfr_pooled_p25`, respectively. See the Data Description of Project 1.

As showed in class (and in the Cheat Sheet Section of Project 1), we just need to use the function `mean()` and `sd()`, along with the option for `na.rm=TRUE` so we do not include any missing data (or NAs) in the calculation. Perhaps the last thing to remember here is that to select all the whole vector of data in any one variable in a dataset, we can use the operator `$`, as follows. I will save each number, and then print it, but you could have just printed it directly.

- The average income of children of parents in percentile 25 across the US is:

```r
avg_us_p25<-mean(atlas$kfr_pooled_p25, na.rm=TRUE)
avg_us_p25
```

```
[1] 34443.48
```

- The standard deviation of income of children of parents in percentile 25 across the US is:

```
sd_us_p25<-sd(atlas$kfr_pooled_p25, na.rm=TRUE)
sd_us_p25
```

```
[1] 8169.155
```

- The average income of children of parents in percentile 25 across the US is:

```
avg_us_p75<-mean(atlas$kfr_pooled_p75, na.rm=TRUE)
avg_us_p75
```

```
[1] 51500.78
```

- The standard deviation of income of children of parents in percentile 75 across the US is:

```
sd_us_p75<-sd(atlas$kfr_pooled_p75, na.rm=TRUE)
sd_us_p75
```

```
[1] 9491.954
```

Question 4

A ggplot graphic showing the distribution of the variables above across the US, Texas, Utah and South Carolina.

Solution

As showed in the Section on Data Visualization, we need to use the package `ggplot2`, and use the syntax for a histogram as follows. Also, I left an example in the Project 'test_project' in Posit Cloud, in the file `data_analysis.R` for you to look at to serve as the basis.

I will first load the package (this will only work if the package has already been installed in this project):

```
library(ggplot2)
```

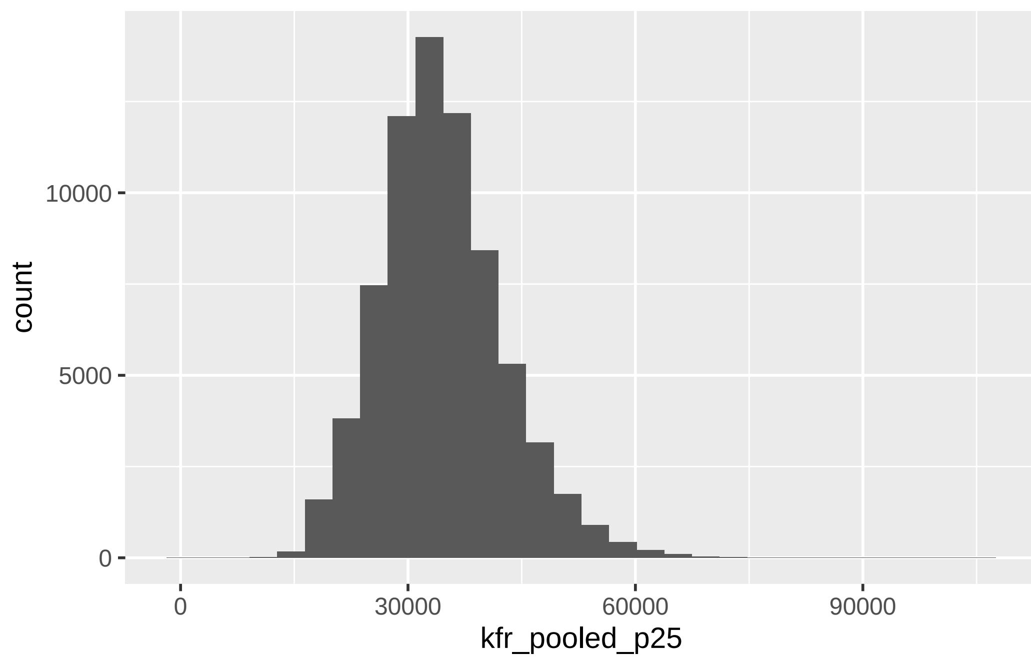
US Data - P25

To evaluate the data across the US, we just need to keep using our `atlas` dataset, as it includes all the neighborhoods in the country. Now, let's show the histogram of mobility for children with low income parents, `kfr_pooled_p25`:

```
# if you want the histogram, you can do this:
ggplot(data = atlas, aes(x=kfr_pooled_p25))+
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

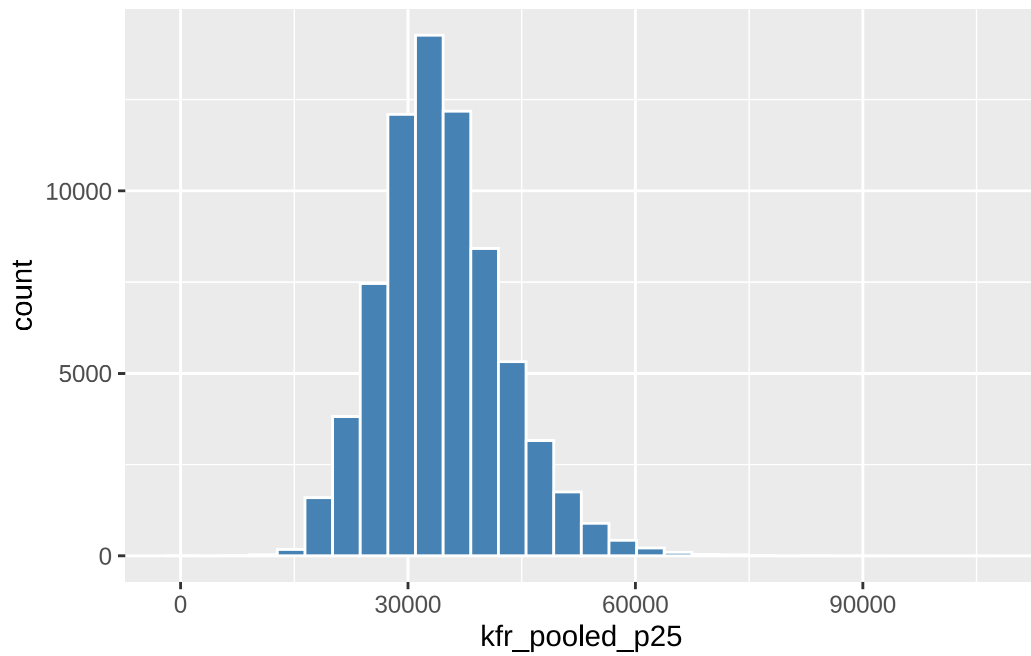
```
Warning: Removed 1267 rows containing non-finite values (`stat_bin()`).
```




```
# this is a nicer version of the one above:
ggplot(data = atlas, aes(x=kfr_pooled_p25))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 1267 rows containing non-finite values (`stat_bin()`).
```



Note: The above two plots represent the same info, but the second looks prettier to me than the first. I will use similar code as in the second going forward, but you could have showed the simple version too. It's up to you.

Regarding the interpretation, we can see that most neighborhoods provide an average income similar to the average neighborhood in the US ($34,443), but a few have very low income mobility (just over $15,000), while some other rare neighborhoods have a large measurement of mobility (over $50,000). Recall that this is the average income for children with parents with the same income level. This is a remarkable range in opportunity across the US geography that we see here in the data.

Texas Data - P25

Let's define a dataset with just the observations from TX. Recall that these are defined by the variable `state==48`. (You can find the whole list by state here - I just googled it, by the way.)

To filter the data like that, I need to use pipes (`%>%`), which require us to load the package `dplyr`. So let's do that first.

```
library(dplyr)
```

```
Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Ok, now let's select the observations for Texas in a new dataset that I will call `texas_atlas`:

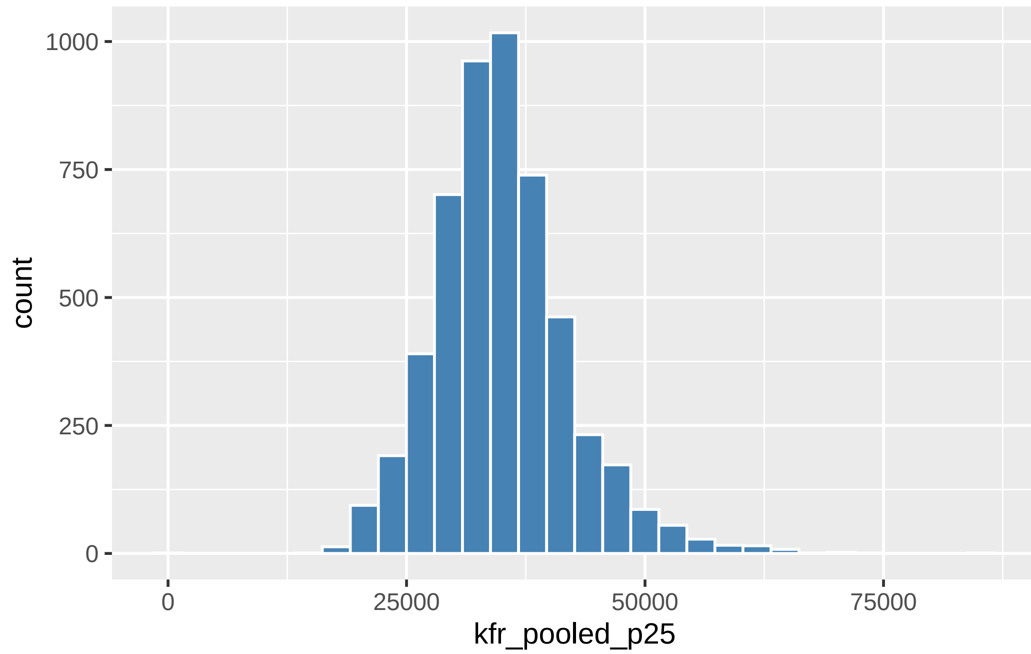
```r
texas_atlas<-atlas%>%
  filter(state==48)
```

Now, let's look at the distribution of mobility in Texas:

```r
ggplot(data = texas_atlas, aes(x=kfr_pooled_p25))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.


Warning: Removed 50 rows containing non-finite values (`stat_bin()`).
```



Utah Data - P25

For Utah, we do pretty much the same thing, now choosing the appropriate filter:

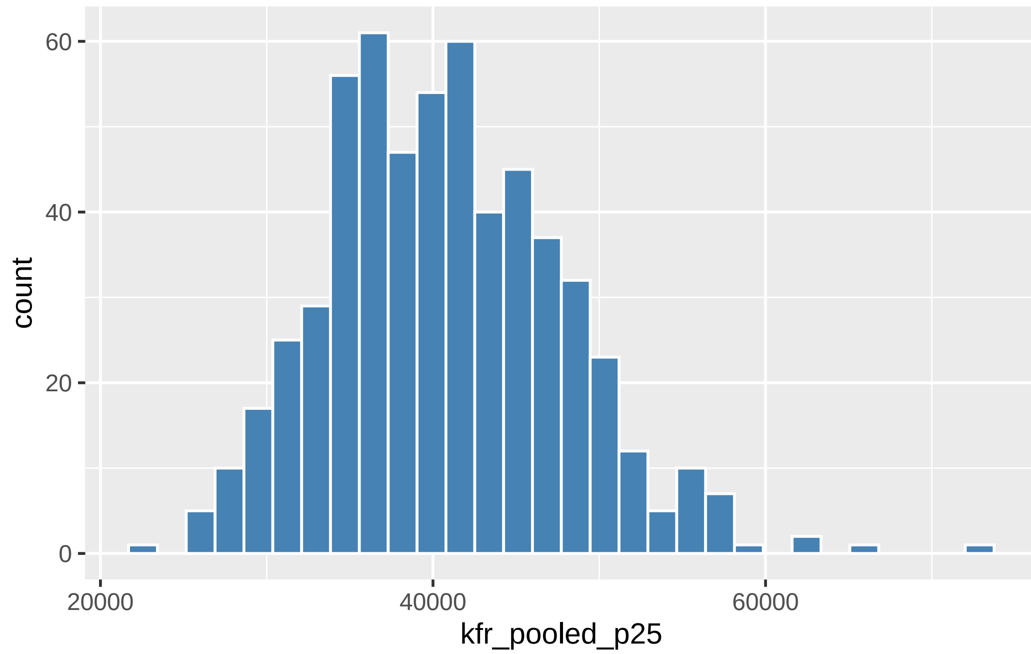
```
utah_atlas<-atlas%>%
  filter(state==49)
```

Now, let's look at the distribution of mobility in Texas:

```
ggplot(data = utah_atlas, aes(x=kfr_pooled_p25))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 6 rows containing non-finite values (`stat_bin()`).
```



South Carolina Data -P25

Same for South Carolina:

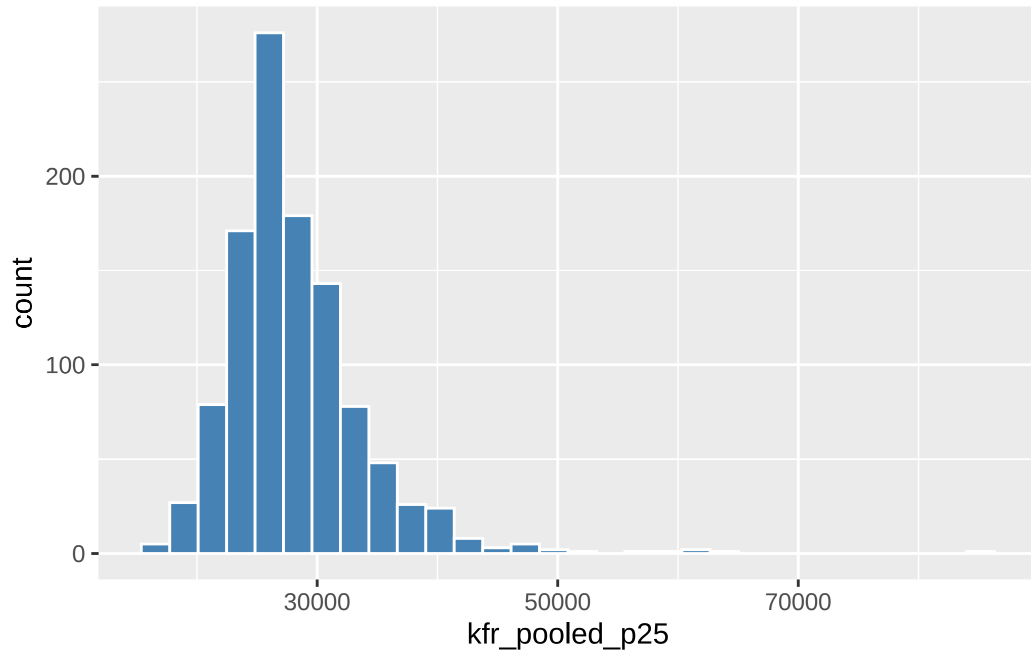
```
sc_atlas<-atlas%>%
  filter(state==45)
```

Now, let's look at the distribution of mobility in Texas:

```
ggplot(data = sc_atlas, aes(x=kfr_pooled_p25))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 11 rows containing non-finite values (`stat_bin()`).
```



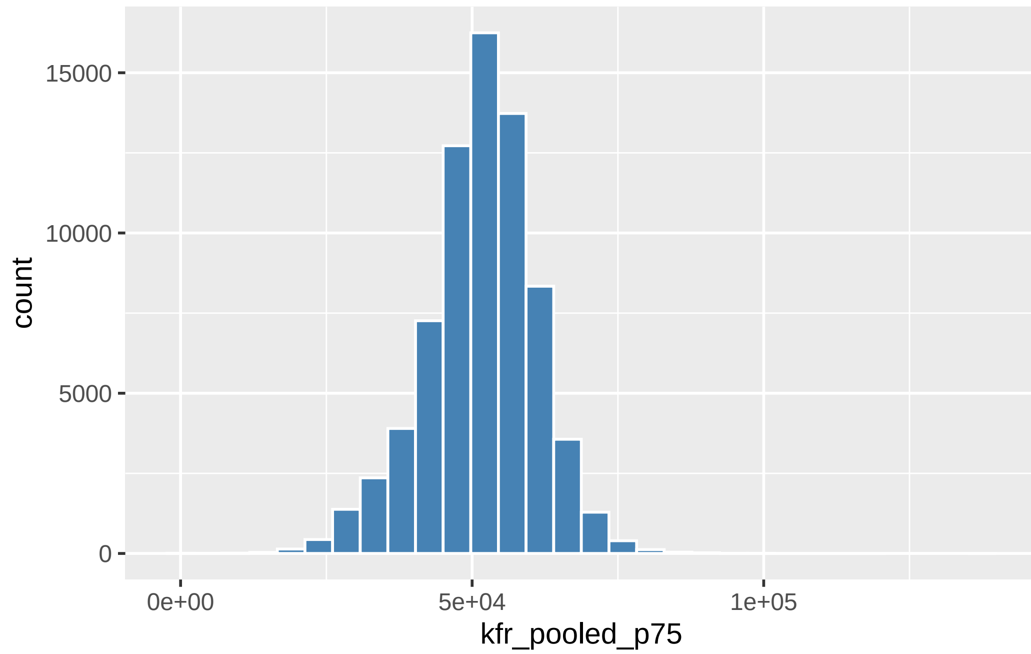
US Data - P75

Now, for the mobility of those children of high income parents (those in percentile 75), we do not need to define any of our data again. Just plot the correct variable and database.

```
ggplot(data = atlas, aes(x=kfr_pooled_p75))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 1266 rows containing non-finite values (`stat_bin()`).
```

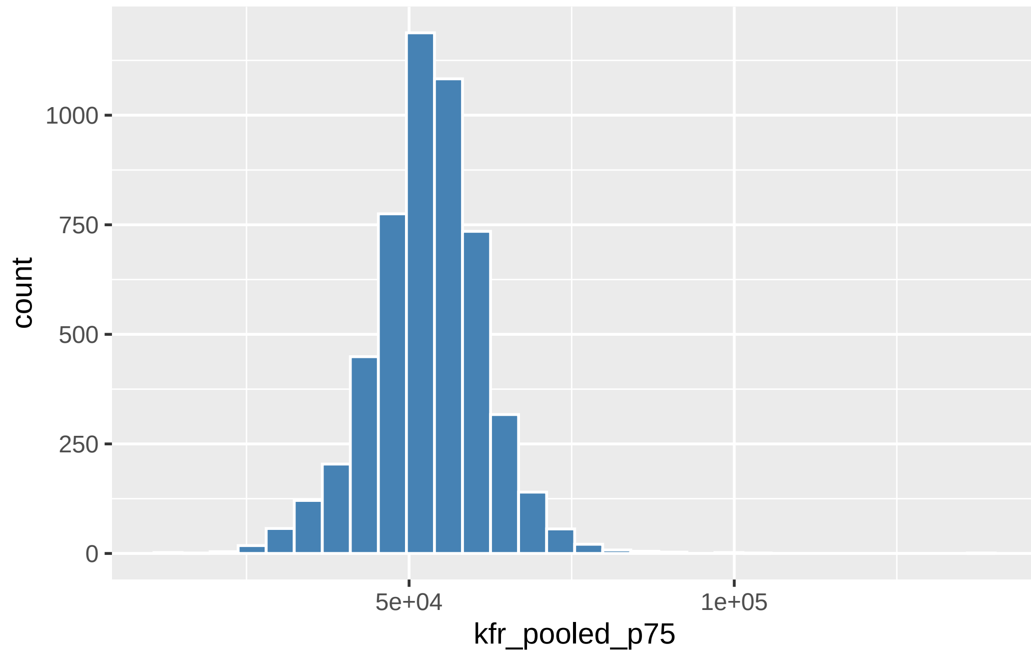


Texas Data - P75

```
ggplot(data = texas_atlas, aes(x=kfr_pooled_p75))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 48 rows containing non-finite values (`stat_bin()`).
```

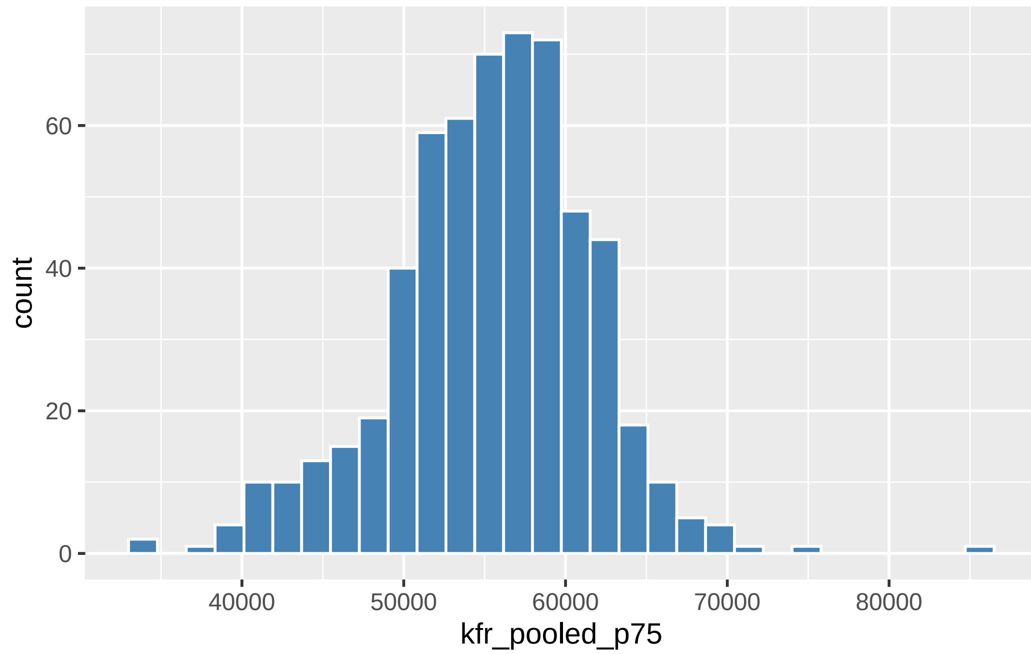


Utah Data - P75

```r
ggplot(data = utah_atlas, aes(x=kfr_pooled_p75))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 6 rows containing non-finite values (`stat_bin()`).
```

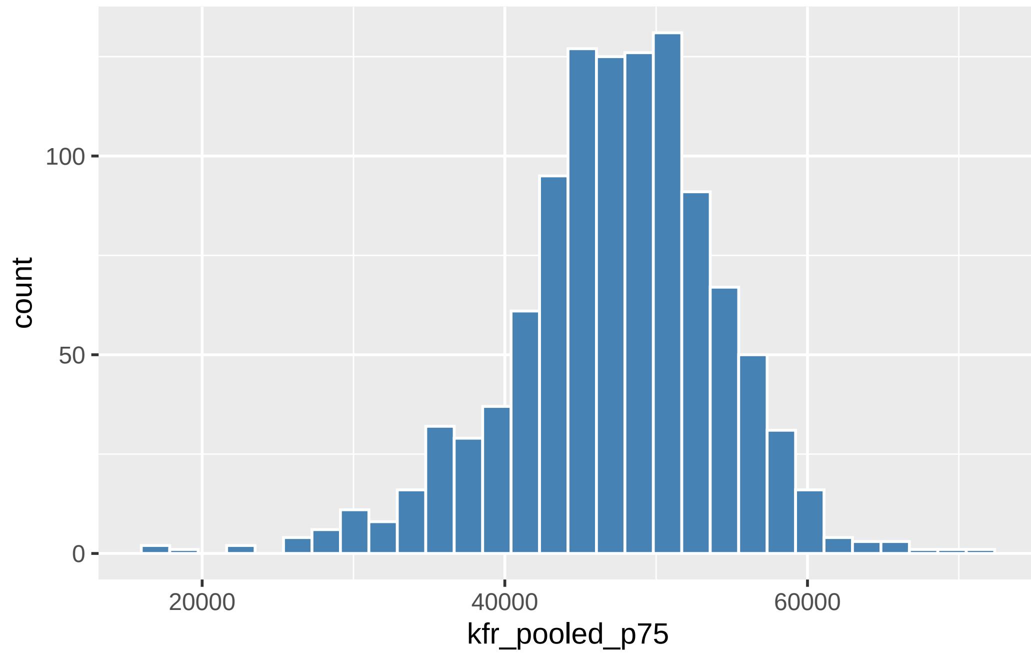


South Carolina Data - P75

```
ggplot(data = sc_atlas, aes(x=kfr_pooled_p75))+
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 11 rows containing non-finite values (`stat_bin()`).
```



Question 5

A simple description of what you see in those numbers and in those plots.

Solution

Overall, the distribution of income of children of low income parents in TX looks to be centered around 36,000 or so, similar to the overall US. In contrast, the distribution in Utah seems centered around 40,000, while that in South Carolina around 30,000 or even lower. These plots show that economic opportunity in Utah is on average likely better than in the overall US, while in South Carolina the reverse seems to happen.

The plots for the distribution for income of children of high income parents tell a similar story, although a bit more nuanced. Overall, it looks like Utah provides higher economic mobility both for children of low and high income parents, while South Carolina has lower mobility for both children of low and high income parents as well.

We will explore these issues further in class, and in Project 1.