# Week 7 Methods: Migration

ECON 125: The Science of Population

# Setup

Today, we analyze Mexican migration

We use a random sample drawn from the 2020 Mexican census

The main dataset includes:

- Household-level and individual-level variables
- One observation per adult aged 20-64

Later, we will merge this dataset with municipality-level data on area poverty

Start by setting up R and loading the first dataset

```
# Load tidyverse and clear the R environment
library(tidyverse)
rm(list=ls())

# Load dataset
load(url("https://github.com/tomvogl/econ125/raw/main/data/mex2020.rds"

# Ask R to not use scientific notation
options(scipen = 999)
```

# Glimpse

```
glimpse(mex2020)
```

```
## Rows: 684,508
## Columns: 14
## $ hhid      <dbl> 1000, 1000, 3000, 4000, 4000, 4000, 6000, 6000, 10
## $ mun       <int> 1001, 1001, 1001, 1001, 1001, 1001, 1001, 1001, 10
## $ locsize   <fct> "100,000 or more inhabitants", "100,000 or more in
## $ hhsize    <dbl> 3, 3, 1, 3, 3, 3, 4, 4, 5, 5, 6, 6, 5, 5, 3, 3, 5,
## $ migrants  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ remitt    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ head      <dbl> 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1,
## $ male      <dbl> 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1,
## $ age       <dbl> 55, 21, 60, 62, 54, 32, 40, 41, 32, 31, 34, 29, 46
## $ educ      <dbl> 16, 14, 17, 18, 18, 16, 16, 16, 9, 16, 8, 11, 16,
## $ country5  <fct> "Mexico", "Mexico", "Mexico", "Mexico", "Mexico",
## $ mun5      <dbl> 1001, 1001, 1001, 1001, 1001, 1001, 1005, 1005, 10
## $ migcause5 <fct> "NIU (not in universe)", "NIU (not in universe)",
## $ sample    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

# Understanding Emigration

Key variables for studying emigration:

- `migrants` = number of HH members who left Mexico in last 5 years
- `hhsize` = number of HH members currently
- `remitt` = 1 if HH received remittances in last 5 years, 0 otherwise

```
mex2020 |> summarise(avg_migrants = mean(migrants),
                     avg_hhsize = mean(hhsize),
                     share_remitt = mean(remitt))
```

```
## # A tibble: 1 x 3
##   avg_migrants avg_hhsize share_remitt
##          <dbl>      <dbl>        <dbl>
## 1       0.0241       4.32       0.0808
```

Could note some interesting facts, but the unit of observation is wrong

# Individual vs. Household Level

The variables on the previous slide are HH-level

Confusing to analyze them in individual-level data

Let's create a HH-level data frame by keeping only HH heads

```
mex2020_hh <- mex2020 |> filter(head==1)

mex2020_hh |> summarise(avg_migrants = mean(migrants),
                        avg_hhsize = mean(hhsize),
                        share_remitt = mean(remitt))
```

```
## # A tibble: 1 x 3
##   avg_migrants avg_hhsize share_remitt
##          <dbl>      <dbl>        <dbl>
## 1       0.0242       3.73       0.0880
```

Key findings

- Less than 1% of HH members emigrated in last 5 years
- 9% of HHs received remittances in last 5 years $\rightarrow$ many long-ago emigrants

# Distribution of Number of Migrants

What is the distribution of migrants per household?

Instead of `group_by()`, convenient to use `count()`

```
mex2020_hh |> count(migrants) |> mutate(pct = 100*n/sum(n))
```

```
## # A tibble: 9 x 3
##   migrants      n       pct
##      <dbl>  <int>     <dbl>
## 1        0 293892 98.0
## 2        1   5300  1.77
## 3        2    580  0.193
## 4        3    145  0.0483
## 5        4     58  0.0193
## 6        5     18  0.006
## 7        6      3  0.001
## 8        7      2  0.000667
## 9        8      2  0.000667
```

# Characteristics of Migrant-Sending vs. Non-Migrant-Sending HHs

Can we learn about determinants of emigration by looking at HH characteristics?

Let's compute average characteristics for HH with and without recent emigrants

```
mex2020_hh <- mex2020_hh |> mutate(anymigrants = if_else(migrants>0, 1,

mex2020_hh |>
  group_by(anymigrants) |>
  summarise(avg_size = mean(hhsize),
            avg_age = mean(age),
            avg_educ = mean(educ),
            share_male = mean(male))
```

```
## # A tibble: 2 x 5
##   anymigrants avg_size avg_age avg_educ share_male
##         <dbl>    <dbl>   <dbl>    <dbl>      <dbl>
## 1           0     3.73    44.0     9.75      0.708
## 2           1     3.78    44.9     9.52      0.593
```

HHs with emigrants have slightly less educated heads $\rightarrow$ negative selection?

Not so fast $\rightarrow$ more female heads, likely due to endogenous HH structure

# Area-Level Predictors of Emigration

To avoid bias from endogenous HHs, better to look at area predictors

Dataset already has locality size → less than 2500 considered rural

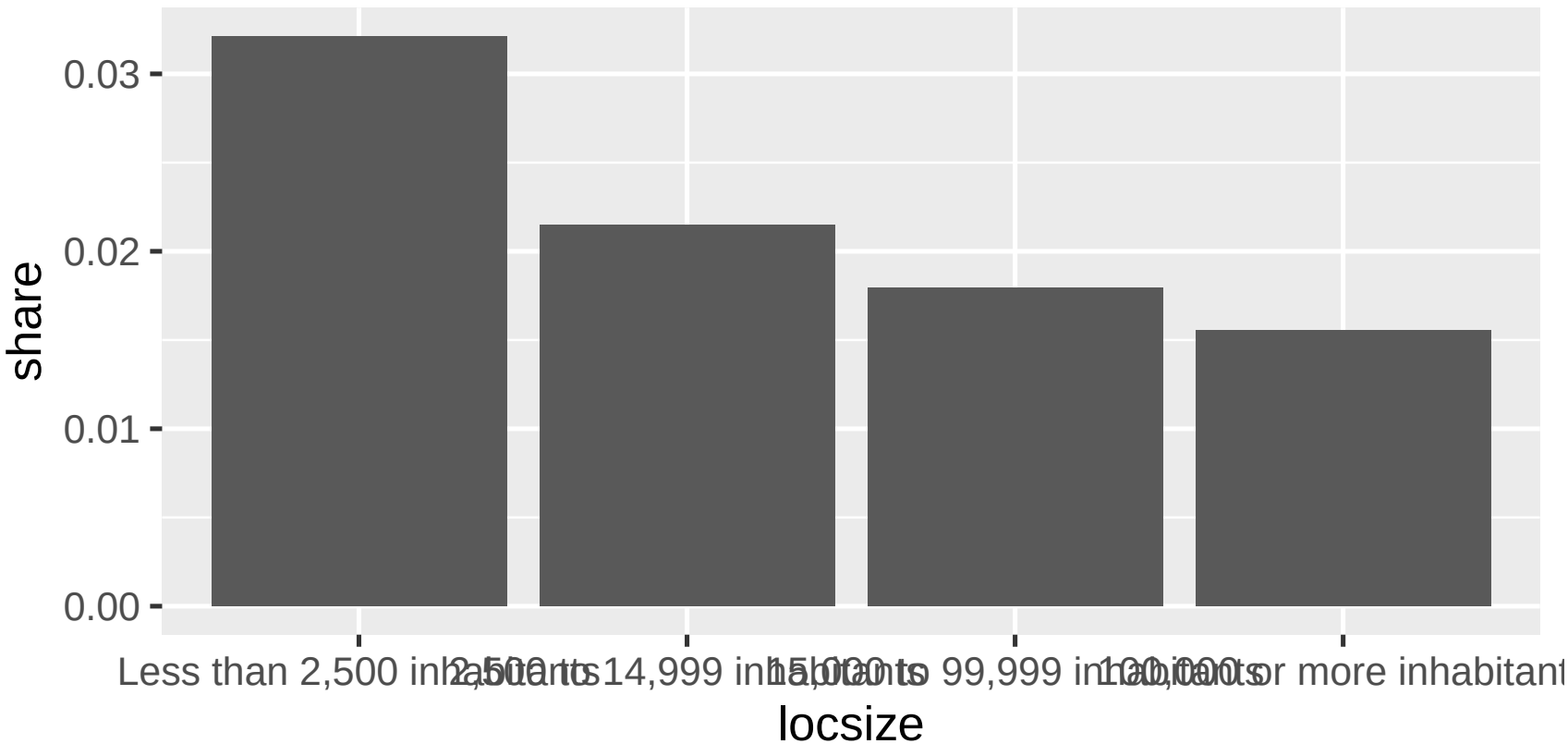
```
mex2020_hh |> count(locsize) |> mutate(pct = 100*n/sum(n))
```

```
## # A tibble: 4 x 3
##   locsize                          n   pct
##   <fct>                        <int> <dbl>
## 1 Less than 2,500 inhabitants   62044  20.7
## 2 2,500 to 14,999 inhabitants   47885  16.0
## 3 15,000 to 99,999 inhabitants  53499  17.8
## 4 100,000 or more inhabitants  136572  45.5
```

# Emigration Shares by Locality Size

HHs in smaller (generally more rural) localities more likely to send migrants

```
tbl <-
  mex2020_hh |>
  group_by(locsize) |>
  summarise(share = mean(anymigrants))

ggplot(tbl, aes(x=locsize,y=share)) +
  geom_col()
```



# Introducing Outside Data on Area Poverty

Locality size a bit hard to interpret

We'll use the Mexican government's measures of municipality "marginalization"

```
marg2020 <- read_csv(url("https://github.com/tomvogl/econ125/raw/main/d
glimpse(marg2020)
```

```
## Rows: 2,469
## Columns: 9
## $ state      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2
## $ name_state <chr> "Aguascalientes", "Aguascalientes", "Aguascalient
## $ mun        <dbl> 1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1
## $ name_mun   <chr> "Aguascalientes", "Asientos", "Calvillo", "Cosío"
## $ population <dbl> 948990, 51536, 58250, 17000, 129929, 47646, 57369
## $ index_raw  <dbl> 60.31880, 56.54607, 57.05825, 57.11403, 59.01176,
## $ grade      <chr> "1 Very low", "1 Very low", "1 Very low", "1 Very
## $ index_norm <dbl> 0.9445084, 0.8854328, 0.8934528, 0.8943262, 0.924
## $ index_rank <dbl> 2435, 1816, 1932, 1948, 2323, 2265, 2076, 2048, 1
```

# Introducing Outside Data on Area Poverty

The data include a coarse "grade" and a continuous "index" of marginalization

For simplicity, we'll use the 5-category "grade"

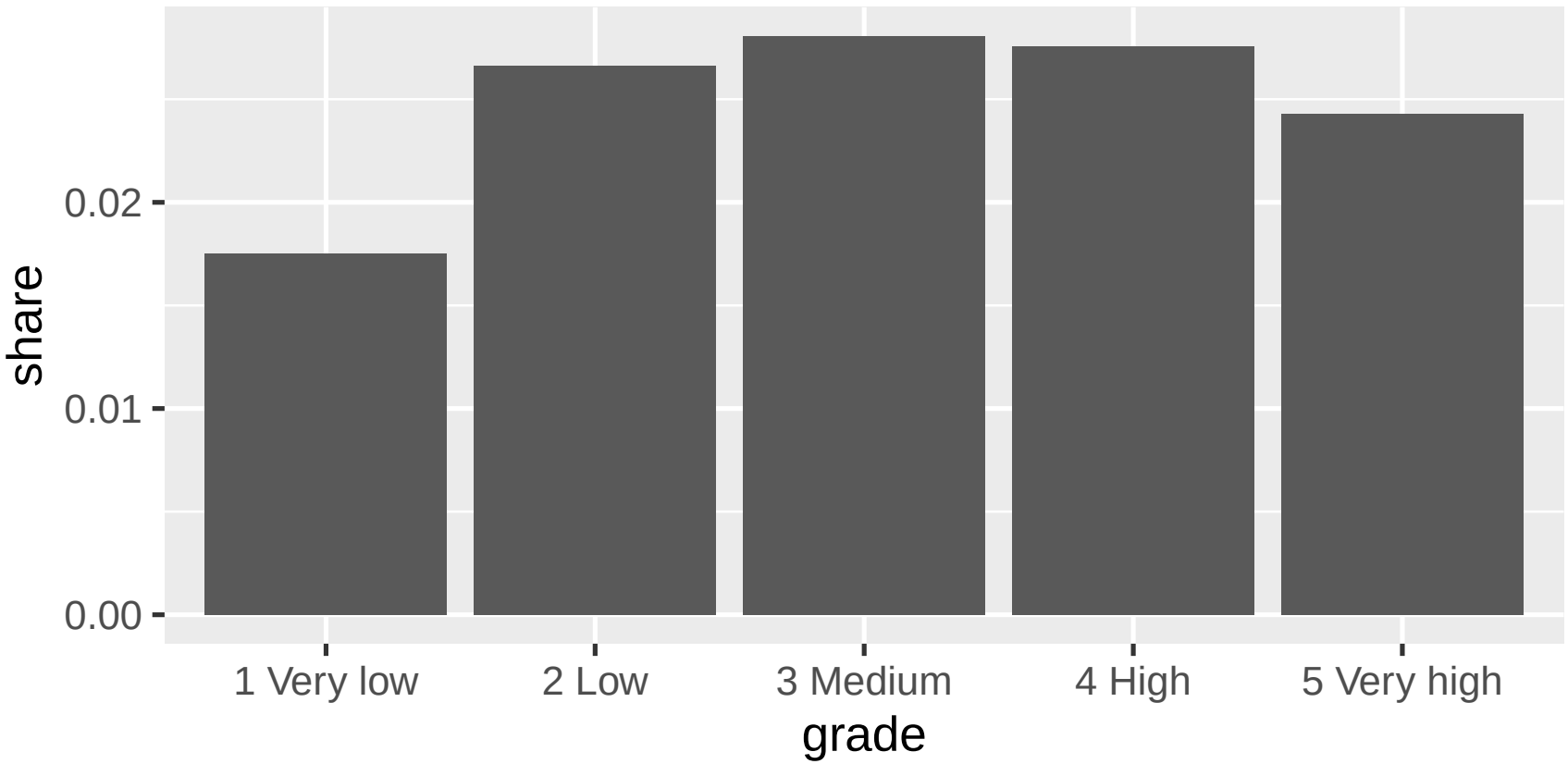
```
marg2020 |> count(grade) |> mutate(pct = n/sum(n))
```

```
## # A tibble: 5 x 3
##   grade           n    pct
##   <chr>       <int>  <dbl>
## 1 1 Very low    655 0.265
## 2 2 Low         530 0.215
## 3 3 Medium      494 0.200
## 4 4 High        586 0.237
## 5 5 Very high   204 0.0826
```

# Emigration Shares by Municipality Marginalization

HHs in poorer (but not poorest) areas more likely to send migrants

```
mex2020_hh <- mex2020_hh |> left_join(marg2020, by = "mun")

tbl <- mex2020_hh |> group_by(grade) |>
  summarise(share = mean(anymigrants))

ggplot(tbl, aes(x = grade, y = share)) +
  geom_col()
```



# Interpretation

Mexican migrants less likely to come from poorest/richest parts of the country

Consistent with the results Abramitzky and Boustan report in their article

Mexican migrants come from the middle of the distribution

No strong pattern of positive or negative selection

# Immigration

The results so far have been about **em**migration: leaving Mexico

The data also tell us about **im**migration: coming to Mexico

`country5` = **individual's** residence in 2015 → switch back to individual-level

```
mex2020 |> count(country5) |> mutate(pct = 100*n/sum(n)) |> arrange(-n)
```

```
## # A tibble: 40 x 3
##    country5           n      pct
##    <fct>          <int>    <dbl>
##  1 Mexico        680538 99.4
##  2 United States   2653  0.388
##  3 Venezuela        224  0.0327
##  4 Colombia         122  0.0178
##  5 Honduras         102  0.0149
##  6 Cuba              96  0.0140
##  7 Guatemala         86  0.0126
##  8 Argentina         59  0.00862
##  9 Spain             54  0.00789
## 10 Brazil            52  0.00760
## # i 30 more rows
```

# Immigration Shares by Municipality Marginalization: Table

Basically all immigrants to Mexico in 2015-2020 came from the US

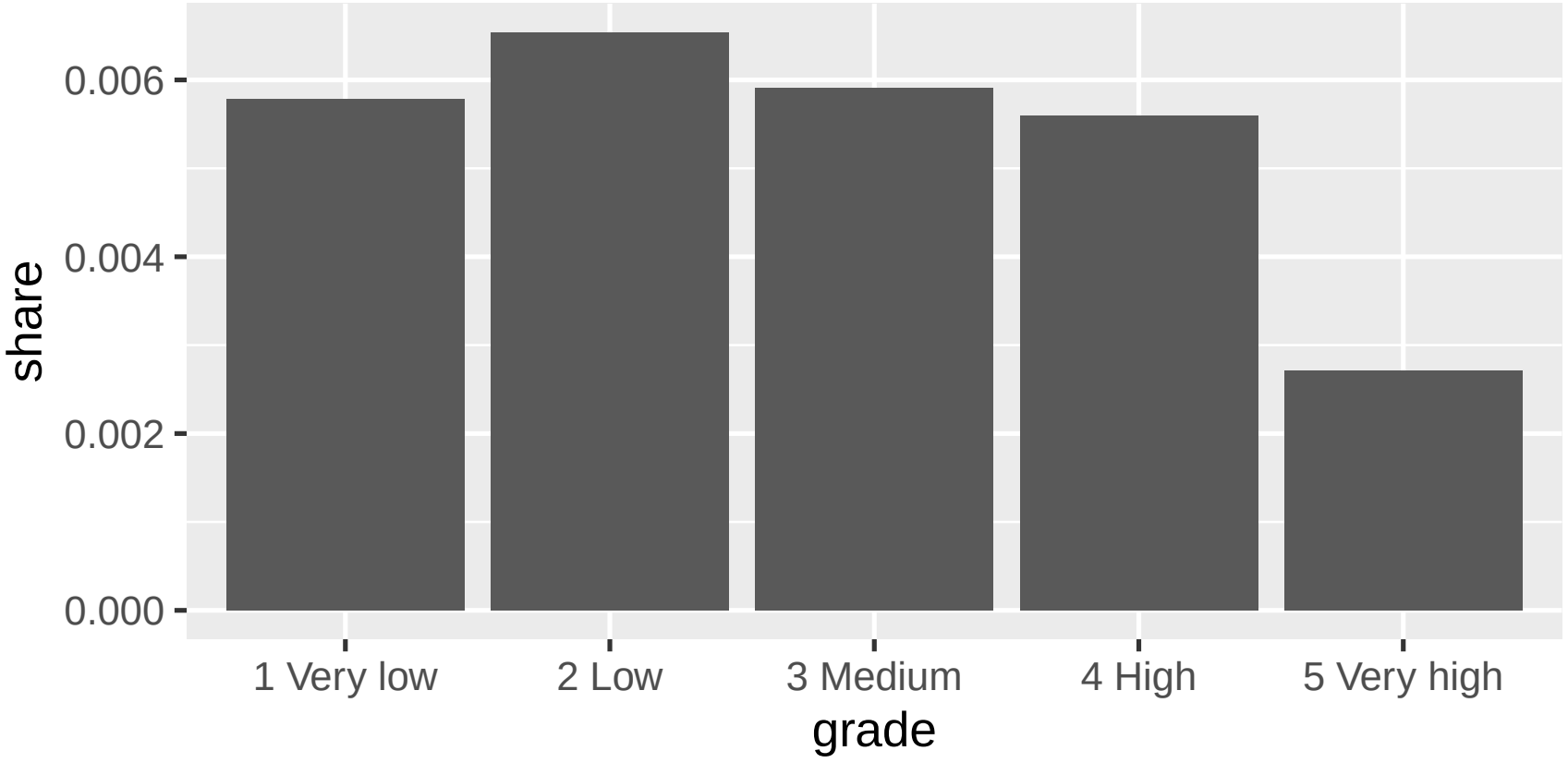
Avoided "very high" marginalization municipalities

```
mex2020 <- mex2020 |>
  mutate(immigrant = if_else(country5!="Mexico", 1, 0)) |>
  left_join(marg2020, by = "mun")

tbl <- mex2020 |>
  group_by(grade) |>
  summarise(share = mean(immigrant))

tbl
```

```
## # A tibble: 5 x 2
##   grade          share
##   <chr>          <dbl>
## 1 1 Very low  0.00579
## 2 2 Low       0.00654
## 3 3 Medium    0.00591
## 4 4 High      0.00560
## 5 5 Very high 0.00272
```

# Immigration Shares by Municipality Marginalization: Graph

```
ggplot(tbl, aes(x = grade, y = share)) +
  geom_col()
```




Some similarities to the emigration graph, but some differences

- More immigrants settling in "very low" than in "very high"
- Incentive to relocate to higher opportunity areas, even if returning to Mexico

# Internal Migration

Mexico also has a lot of internal migration

Here we will define internal migration as movement across municipalities

How common was internal migration in 2015-20? Check whether `mun == mun5`

Some individuals have `NA` for `mun5`, mostly because they lived outside Mexico

```
summary(mex2020$mun5)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1001   11014   15099   16734   22014   32058    7807
```

Drop `NA`s and generate internal migration dummy variable

```
mex2020 <-
  mex2020 |>
  drop_na(mun5) |>
  mutate(migrant = if_else(mun!=mun5, 1, 0))
```

# Internal Migrants: Population Share and Characteristics

6% of the Mexican adult population moved municipalities during 2015-2020

```
mex2020 |> summarise(share = mean(migrant))
```

```
## # A tibble: 1 x 1
##    share
##    <dbl>
## 1 0.0633
```

How were migrants different from non-migrants?

```
mex2020 |>
  group_by(migrant) |>
  summarise(avg_age = mean(age),
            avg_educ = mean(educ),
            share_male = mean(male))
```

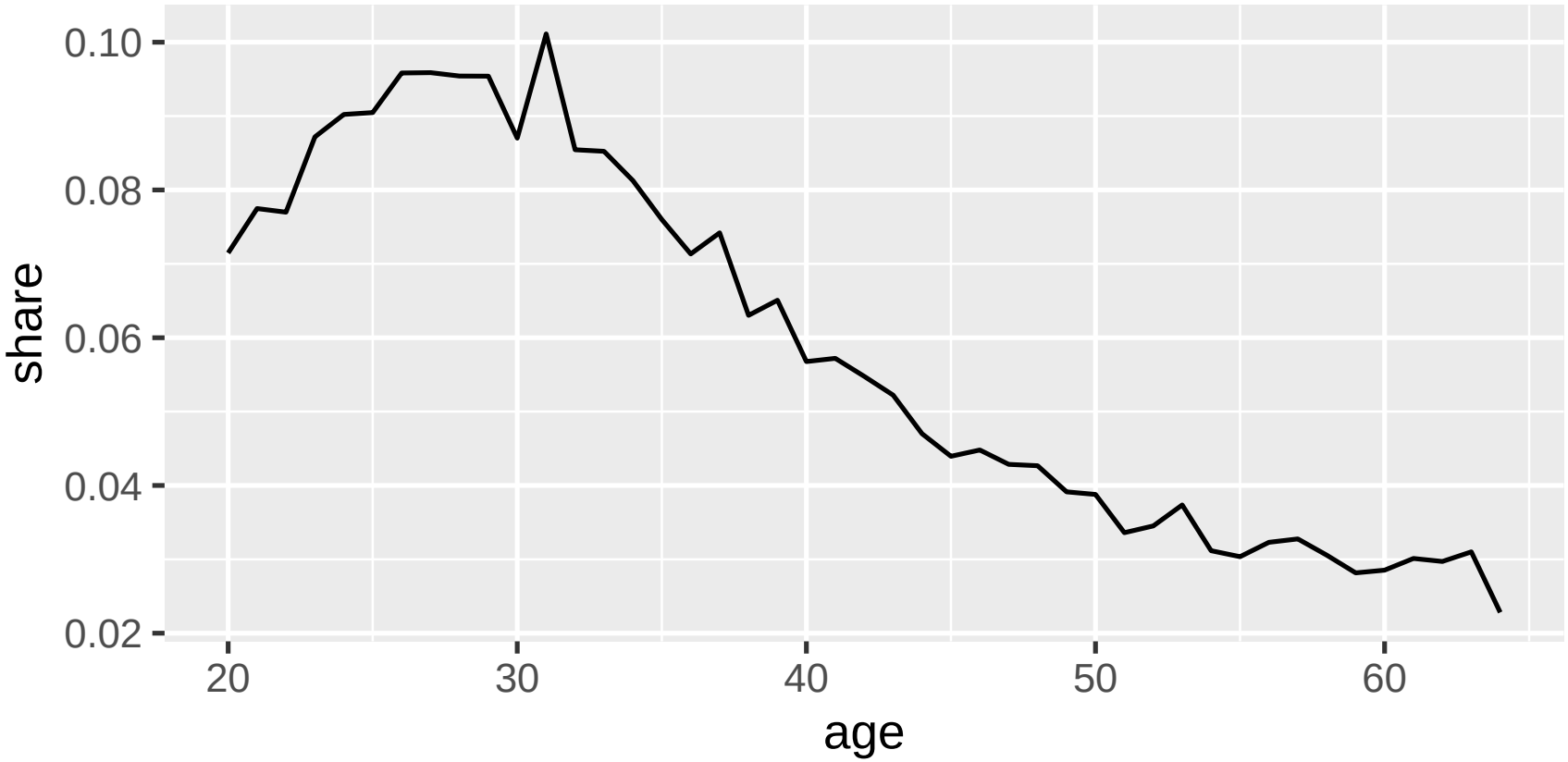
```
## # A tibble: 2 x 4
##   migrant avg_age avg_educ share_male
##     <dbl>   <dbl>    <dbl>      <dbl>
## 1       0    39.2     10.0      0.483
## 2       1    34.7     11.5      0.493
```

Selection! Migrants are younger and more educated than non-migrants

# Age and Internal Migration

Let's dig into the age-migration relationship a bit more

```
tbl <- mex2020 |> group_by(age) |> summarise(share = mean(migrant))

ggplot(tbl, aes(x = age, y = share)) +
  geom_line()
```



# Interpreting the Age Patterns

People in their 20s were most likely to move

- Common for young people to be more mobile
- Could reflect cohort effects $\rightarrow$ not possible to check in cross-section

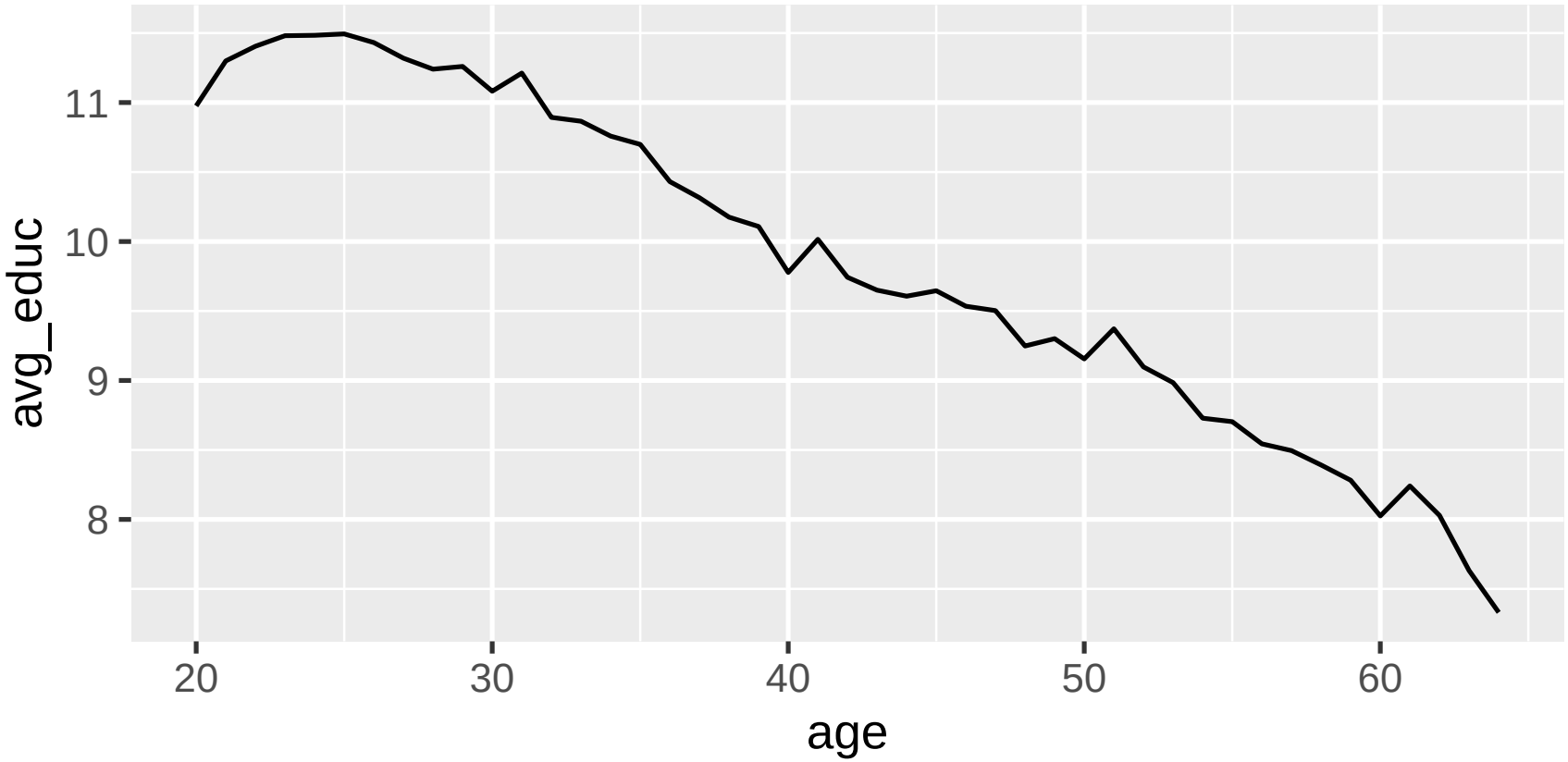
Cohort effects are likely to be important for confounding role of education

- Recent cohorts more educated, more likely to move

# Age and Education

Age is related to education, but this is really a cohort phenomenon

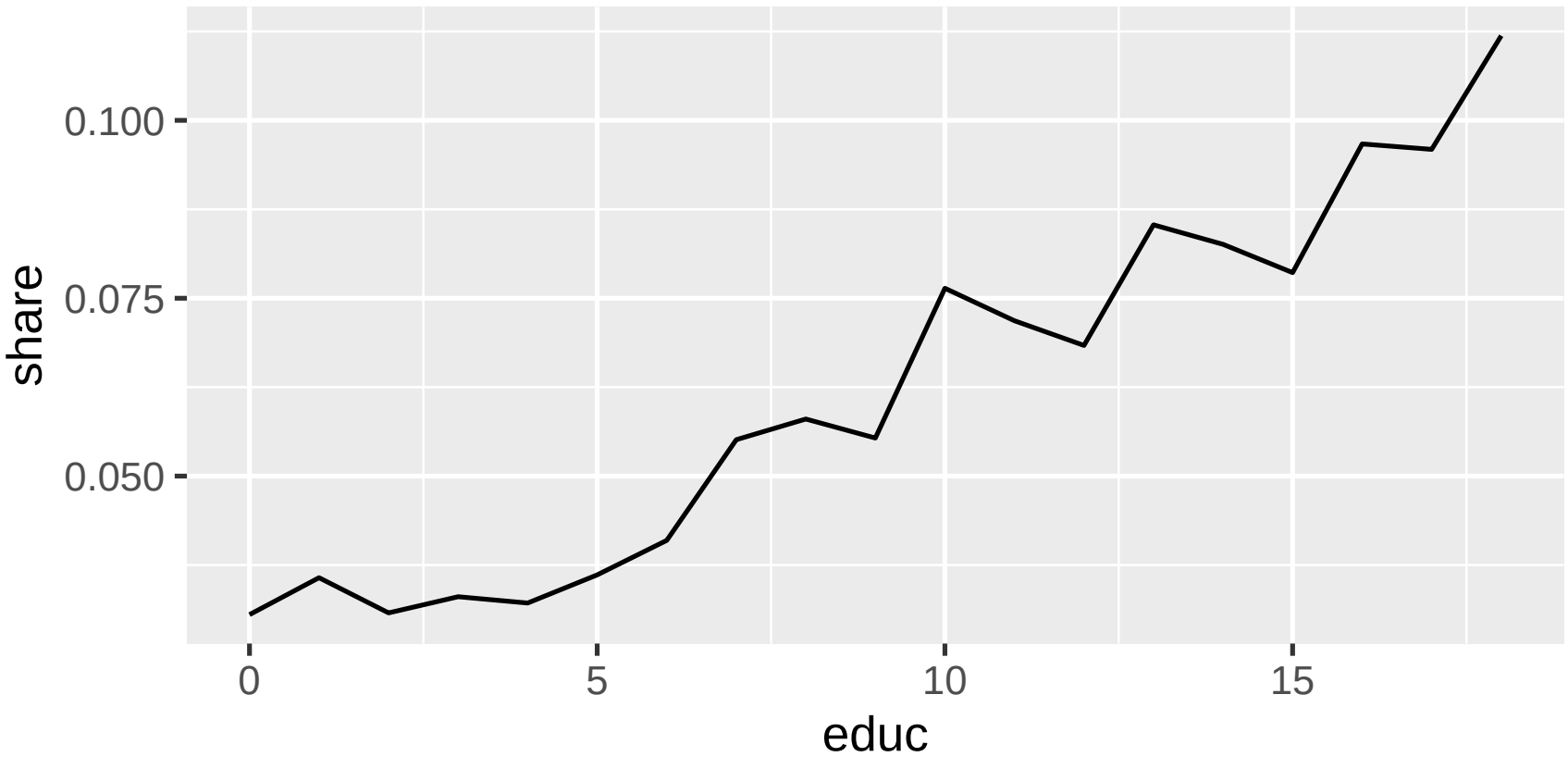
```
tbl <- mex2020 |> group_by(age) |> summarise(avg_educ = mean(educ))

ggplot(tbl, aes(x = age, y = avg_educ)) +
  geom_line()
```



# Education and Internal Migration

Let's dig into the education-migration relationship a bit more

```
tbl <- mex2020 |> group_by(educ) |> summarise(share = mean(migrant))

ggplot(tbl, aes(x = educ, y = share)) +
  geom_line()
```




Very clear positive selection

# Disentangling the Roles of Age and Education

How can we disentangle these two forces?

Standard approach: regression adjustment

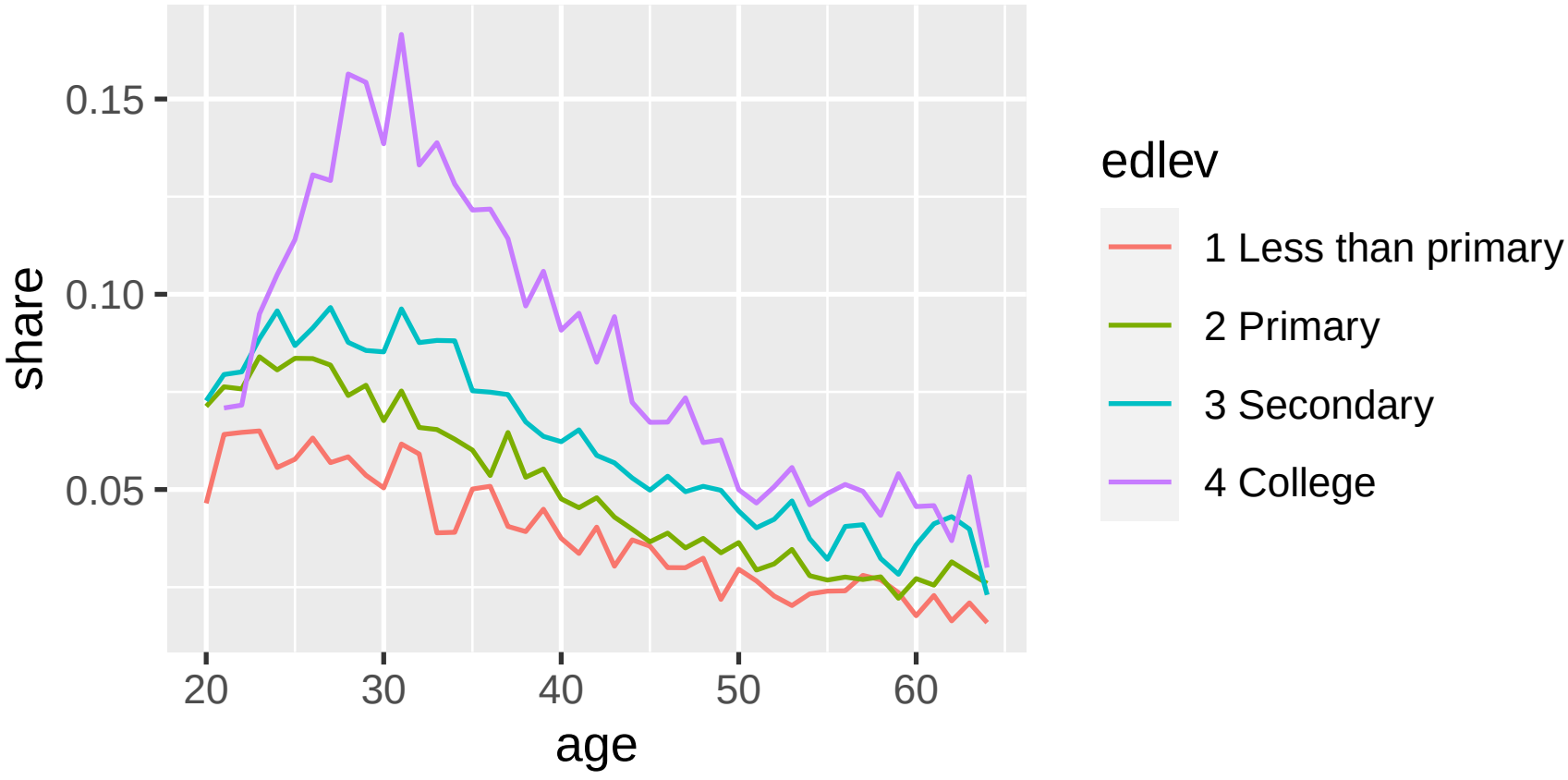
But we are not running regressions in this class!

As an alternative, we can draw separate age-migration graphs by education level

```
mex2020 <-
  mex2020 |>
  mutate(edlev = case_when(educ<6 ~ "1 Less than primary",
                           educ>=6&educ<12 ~ "2 Primary",
                           educ>=12&educ<16 ~ "3 Secondary",
                           educ>=16 ~ "4 College"))
```

# Age and Internal Migration by Education Level

```
tbl <- mex2020 |>
  group_by(edlev, age) |>
  summarise(share = mean(migrant))

ggplot(tbl, aes(x = age, y = share, color=edlev)) +
  geom_line()
```



# Interpreting the Age and Education Patterns

Age and education independently predict migration

More educated migrate more at almost every age

Young migrate more than old in every education group, but peak age varies

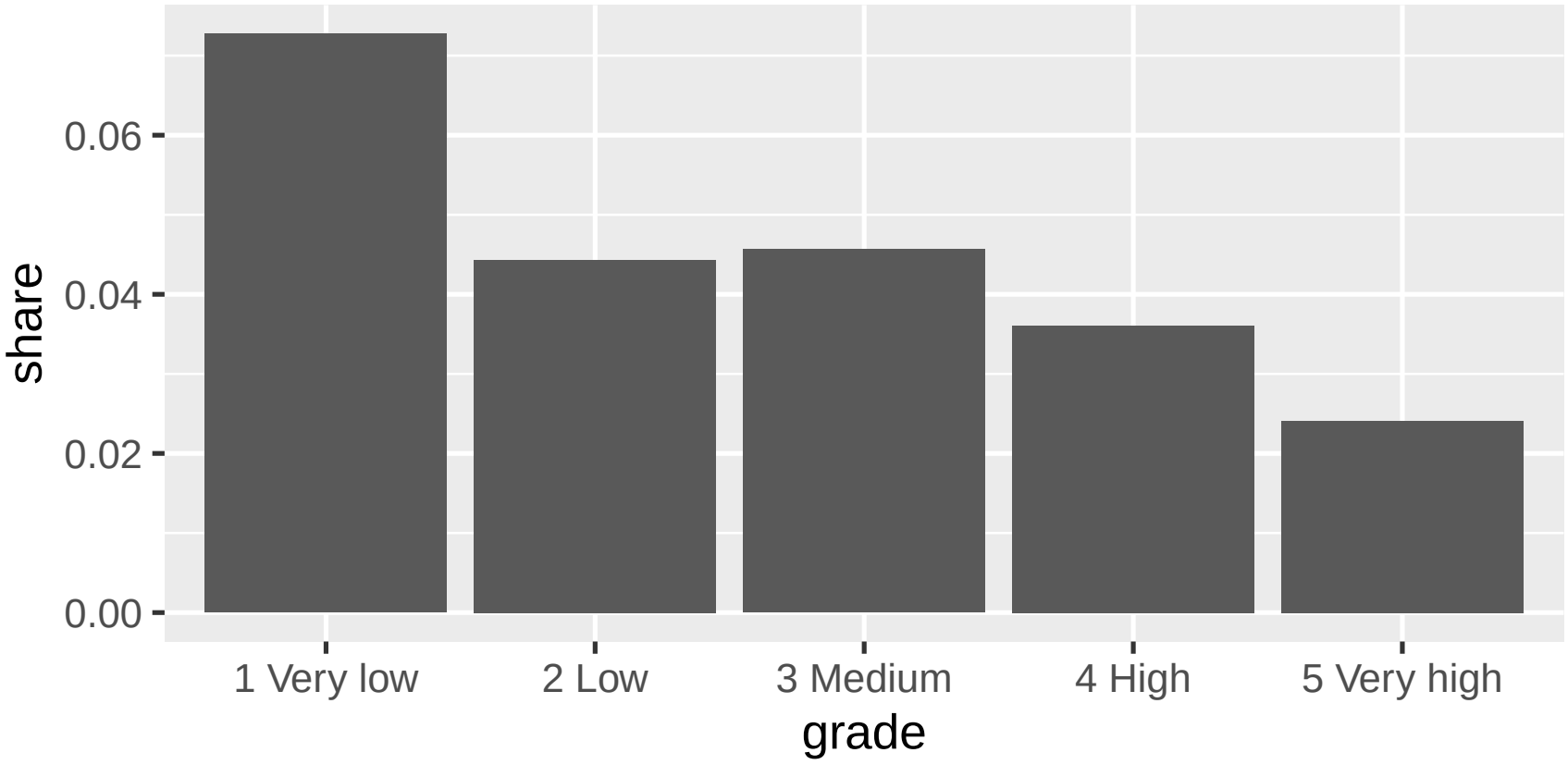
Lots of mobility for college-educated just after finishing college

But even for individuals in their 60s, migration rates highest for college, then secondary, then primary, then less

# Migrant Status by Destination Municipality Marginalization

Do high opportunity areas tend to receive more internal migrants? **Yes**

```
tbl <- mex2020 |> group_by(grade) |> summarise(share = mean(migrant))

ggplot(tbl, aes(x = grade, y = share)) +
  geom_col()
```



# Origin Municipality Marginalization

Also interesting to study the marginalization level of **origin** municipalities

We need to merge in marginalization data again, this time by lagged municipality

First rename variables in the `marg2020` data frame to avoid duplicate names

```
marg2020 <- marg2020 |>
  select(mun, grade, index_rank) |>
  rename(mun5 = mun, grade5 = grade, index_rank5 = index_rank)
```

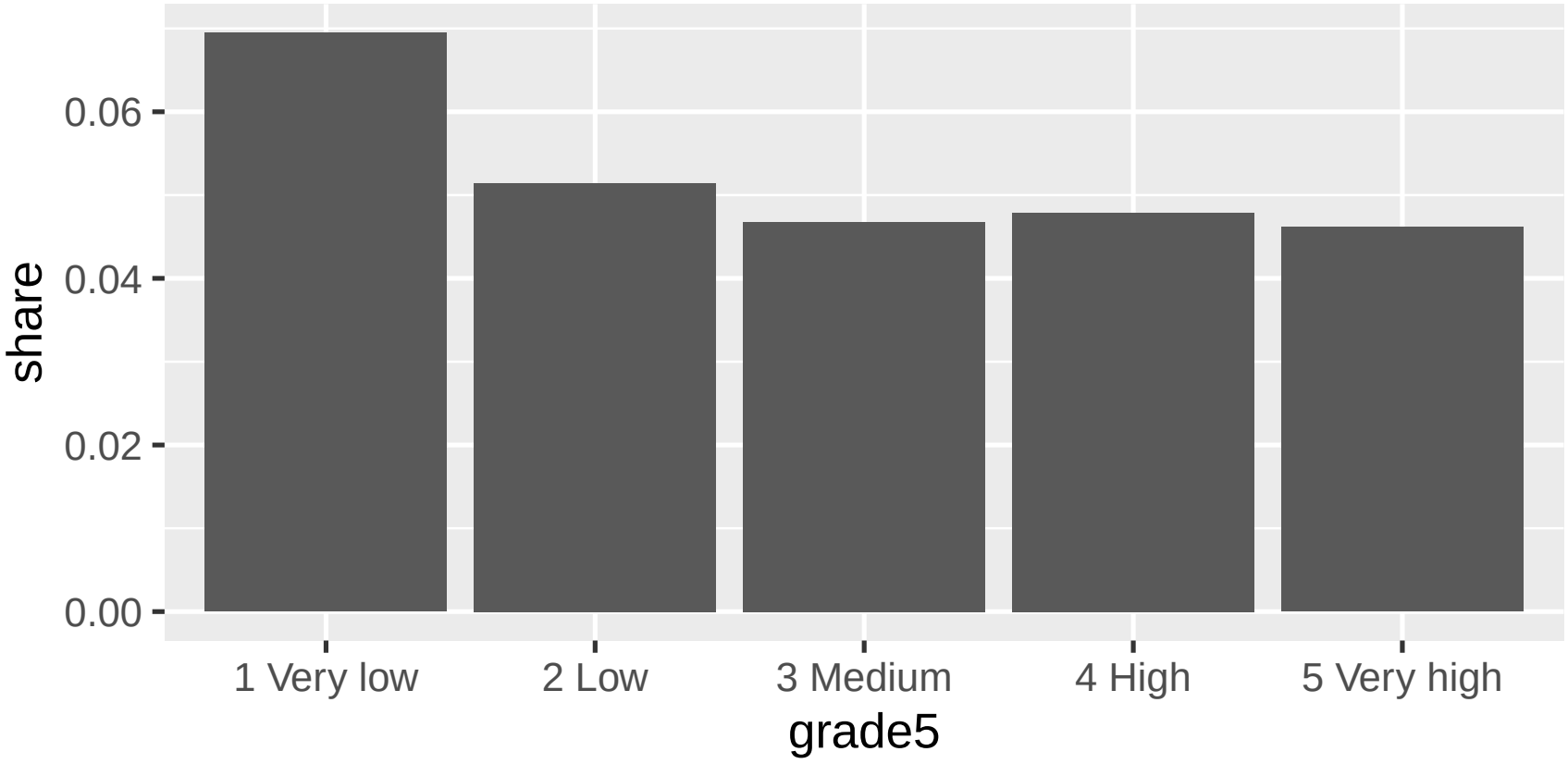
Now merge into census dataset by `mun5`

```
mex2020 <- mex2020 |>
  left_join(marg2020, by = "mun5")
```

# Migrant Status by Origin Municipality Marginalization

Do low opportunity areas tend to send more internal migrants? **No**

```
tbl <- mex2020 |> group_by(grade5) |> summarise(share = mean(migrant))

ggplot(tbl, aes(x = grade5, y = share)) +
  geom_col()
```



# Origin-Destination Matrix

Preceding results suggest many migrants move from *very low* to *very low*

We can check by tabulating `grade5` with `grade`

We'll deviate from `tidyverse` syntax because it's is much easier in base `R`

```
migrants <- mex2020 |> filter(migrant==1)

table(migrants$grade5, migrants$grade)
```

```
##
##                1 Very low 2 Low 3 Medium 4 High 5 Very high
##   1 Very low        27881  2469     1405    887          253
##   2 Low              3223   612      390    238           68
##   3 Medium           1656   345      253    188           24
##   4 High             1308   306      270    168           51
##   5 Very high         498   144       84     85           37
```

The rows correspond to origins, and the columns correspond to destinations

# Refining the Origin-Destination Matrix

Easier to interpret with relative frequencies ($\frac{n}{N}$) rather frequencies ($N$)

```
tbl <- table(migrants$grade5, migrants$grade)

round(100*prop.table(tbl), 1)
```

```
##
##                1 Very low 2 Low 3 Medium 4 High 5 Very high
##   1 Very low         65.1   5.8      3.3    2.1          0.6
##   2 Low               7.5   1.4      0.9    0.6          0.2
##   3 Medium            3.9   0.8      0.6    0.4          0.1
##   4 High              3.1   0.7      0.6    0.4          0.1
##   5 Very high         1.2   0.3      0.2    0.2          0.1
```

65% of internal migration is from "very low" to "very low"

Most internal migration is not "moving to opportunity," but you can find it

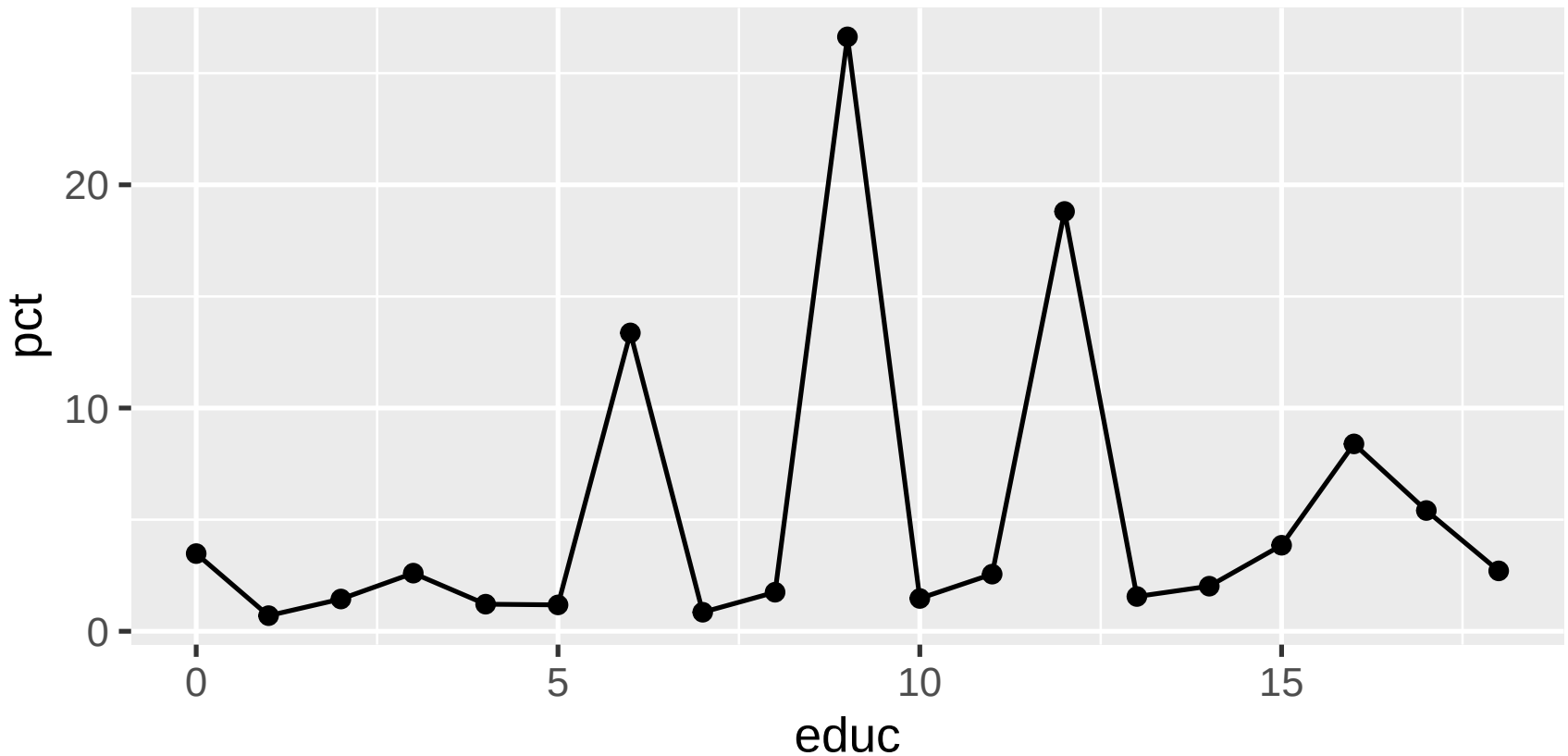
E.g., "very high" to "very low" twice as likely as "very low" to "very high"

# Distribution of Years of Education Among Mexican Adults

PS 4 asks you to compare the education distributions of Mexican immigrants to the US and Mexicans in Mexico

One way to represent the distribution is with a histogram of years of education

```
tbl <- mex2020 |> count(educ) |> mutate(pct = 100*n/sum(n))
ggplot(tbl, aes(x=educ, y=pct)) +
  geom_point() +
  geom_line()
```



# Distribution of Education Levels among Mexican Adults

Another nice way to represent it is with a tabulation of highest level completed

This table is relevant for your problem set!

```
mex2020 |> count(edlev) |> mutate(pct = 100*n/sum(n))
```

```
## # A tibble: 4 x 3
##   edlev                    n   pct
##   <chr>                <int> <dbl>
## 1 1 Less than primary  71894  10.6
## 2 2 Primary           315530  46.6
## 3 3 Secondary         177553  26.2
## 4 4 College           111724  16.5
```