# Group Project I - College Scorecard Dataset

*Kevin Hunt, Noah Johnson, Lydia Laseur, Michael McCormack*

*November 4, 2017*

### College Scorecard Data Aggregation and Cleaning

## Replace PrivacySuppressed

To view the R code for cleaning the College Scorecard Dataset, assessing the quality of the data, and adding latitude, longitude and rurality see the clean_all.R file.

We have chosen 65 columns to work with:

| | |
|---|---|
| year | program_percentage.military |
| id | program_percentage.multidiscipline |
| ope8_id | program_percentage.parks_recreation_fitness |
| ope6_id | program_percentage.philosophy_religious |
| name | program_percentage.theology_religious_vocation |
| city | program_percentage.physical_science |
| state | program_percentage.science_technology |
| zip | program_percentage.psychology |
| location.lat | program_percentage.security_law_enforcement |
| location.lon | program_percentage.public_administration_social_service |
| degrees_awarded.predominant_recoded | program_percentage.social_science |
| degrees_awarded.predominant | program_percentage.construction |
| degrees_awarded.highest | program_percentage.mechanic_repair_technology |
| locale | program_percentage.precision_production |
| program_percentage.agriculture | program_percentage.transportation |
| program_percentage.resources | program_percentage.visual_performing |
| program_percentage.architecture | program_percentage.health |
| program_percentage.ethnic_cultural_gender | program_percentage.business_marketing |
| program_percentage.communication | program_percentage.history |
| program_percentage.communications_technology | attendance.academic_year |
| program_percentage.computer | tuition.in_state |
| program_percentage.personal_culinary | tuition.out_of_state |
| program_percentage.education | faculty_salary |
| program_percentage.engineering | share_25_older |
| program_percentage.engineering_technology | share_firstgeneration |
| program_percentage.language | share_firstgeneration_parents.middleschool |
| program_percentage.family_consumer_science | share_firstgeneration_parents.highschool |
| program_percentage.legal | share_firstgeneration_parents.somecollege |
| program_percentage.english | demographics.age_entry |
| program_percentage.humanities | demographics.over_23_at_entry |
| program_percentage.library | demographics.first_generation |
| program_percentage.biological | rurality |
| program_percentage.mathematics | |

When analyzing the data, please keep in mind that not every year has good completeness of the data. Use completeness.by.year to explore the data before you begin visualization. Example:

```
completeness.by.year[, c("year", "locale")]
```

| year | locale |
|---|---|
| 1996_97 | 0.0000000 |
| 1997_98 | 0.0000000 |
| 1998_99 | 0.0000000 |
| 1999_00 | 0.0000000 |
| 2000_01 | 0.0000000 |
| 2001_02 | 0.0000000 |
| 2002_03 | 0.0000000 |
| 2003_04 | 0.0000000 |
| 2004_05 | 0.0000000 |
| 2005_06 | 0.0000000 |
| 2006_07 | 0.0000000 |
| 2007_08 | 0.0000000 |
| 2008_09 | 0.0000000 |
| 2009_10 | 0.0000000 |
| 2010_11 | 0.0000000 |
| 2011_12 | 0.0000000 |
| 2012_13 | 0.0000000 |
| 2013_14 | 0.0000000 |
| 2014_15 | 0.0000000 |
| 2015_16 | 0.9412617 |

## Visualizing the College Scorecard Dataset

Please select a question (or come up with your own) and add some visualizations to display any relationships of interest. Most questions we came up with involved the rurality of the schools. The orginial dataset includes this information for the 2015-2016 school year in the "locale" column. For all the years, I've added a column of rurality to the dataset (original data can be found in ziprural.txt or downloaded here). The dictionary for explanations of these codes is as follows:

| **A1** | **Rural Urban Continuum Code** | | |
|---|---|---|---|
| **Code** | **Description** | **Counties** | **2000 Population** |
| **Metro Counties:** | | | |
| 1 | Counties in metro areas of 1 million population or more | 413 | 149,224,067 |
| 2 | Counties in metro areas of 250,000 to 1 million population | 325 | 55,514,159 |
| 3 | Counties in metro areas of fewer than 250,000 population | 351 | 27,841,714 |
| **Nonmetro Counties:** | | | |
| 4 | Urban population of 20,000 or more, adjacent to a metro area | 218 | 14,442,161 |
| 5 | Urban population of 20,000 or more, not adjacent to a metro area | 105 | 5,573,273 |
| 6 | Urban population of 2,500 to 19,999, adjacent to a metro area | 609 | 15,134,357 |
| 7 | Urban population of 2,500 to 19,999, not adjacent to a metro area | 450 | 8,463,700 |
| 8 | Completely rural or < 2,500 urban population, adjacent to a metro area | 235 | 2,425,743 |
| 9 | Completely rural or < 2,500 urban population, not adjacent to a metro area | 435 | 2,802,732 |

**Questions we discussed in class:**

**Rurality story: Where did they come from? Where did they go? Where did they come from, Cotton Eye Joe?**
Do more rural schools show a higher percentage of returning adults than metro schools?
Do more rural schools have a higher percentage of first generation students?
How do the degrees awarded differ between rurality classes (do rural areas award more technical degrees)?
Are more rural schools cheaper?

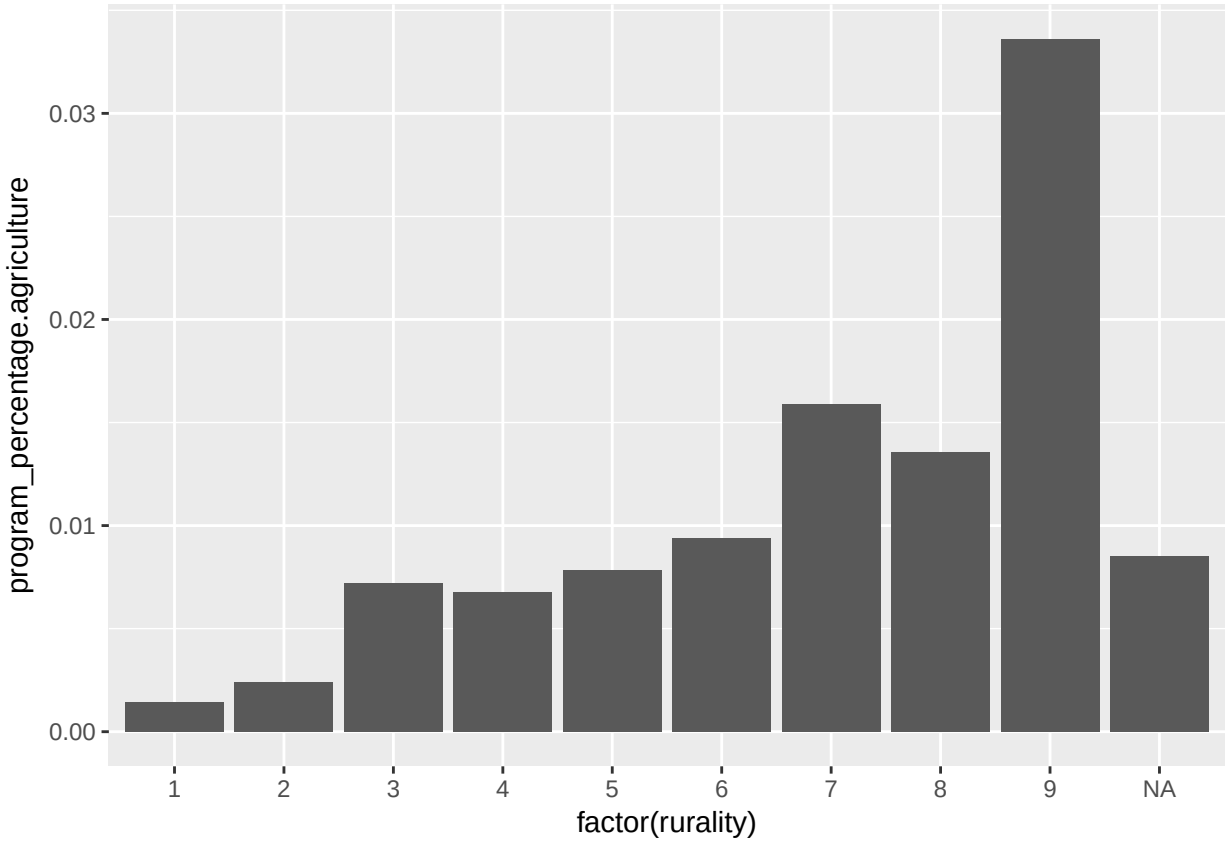
**Other Questions** How does the average faculty salary of a school scale with the cost of living in that area? (requires finding cost of living)
We can also ask some summary statistics of schools grouping by year: what has changed for schools in 20 years?

After digging through the data below, I'm not sure rurality is the best measure to look for differences in degrees awarded (9 the most rural has the highest percentage of history degrees?). Perhaps poverty levels would be more interesting.

I just threw together some simple ggplots, feel free to continue exploring anything that jumps out at you. I am by no means claiming all these for myself.

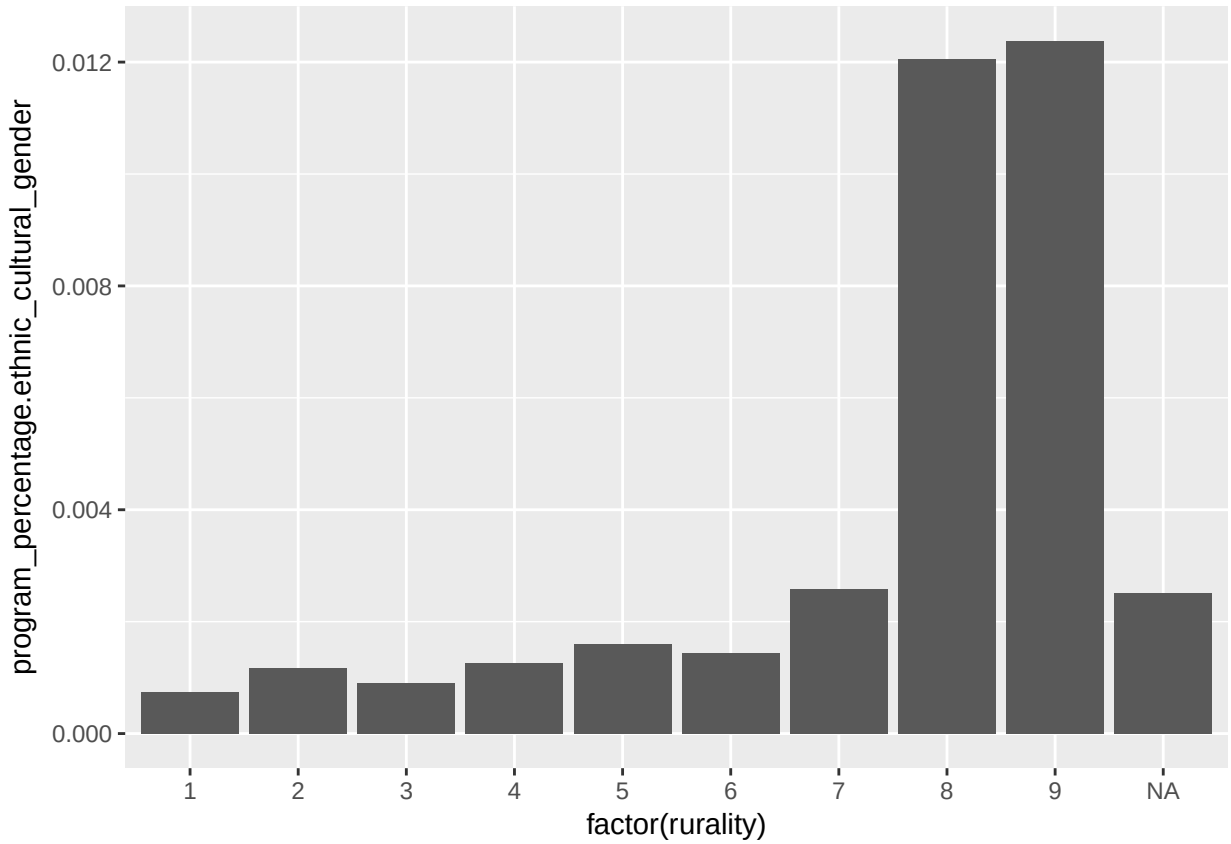
```
library(ggplot2)
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=program_percentage.agricul
```

```
## Warning: Removed 739 rows containing non-finite values (stat_summary).
```



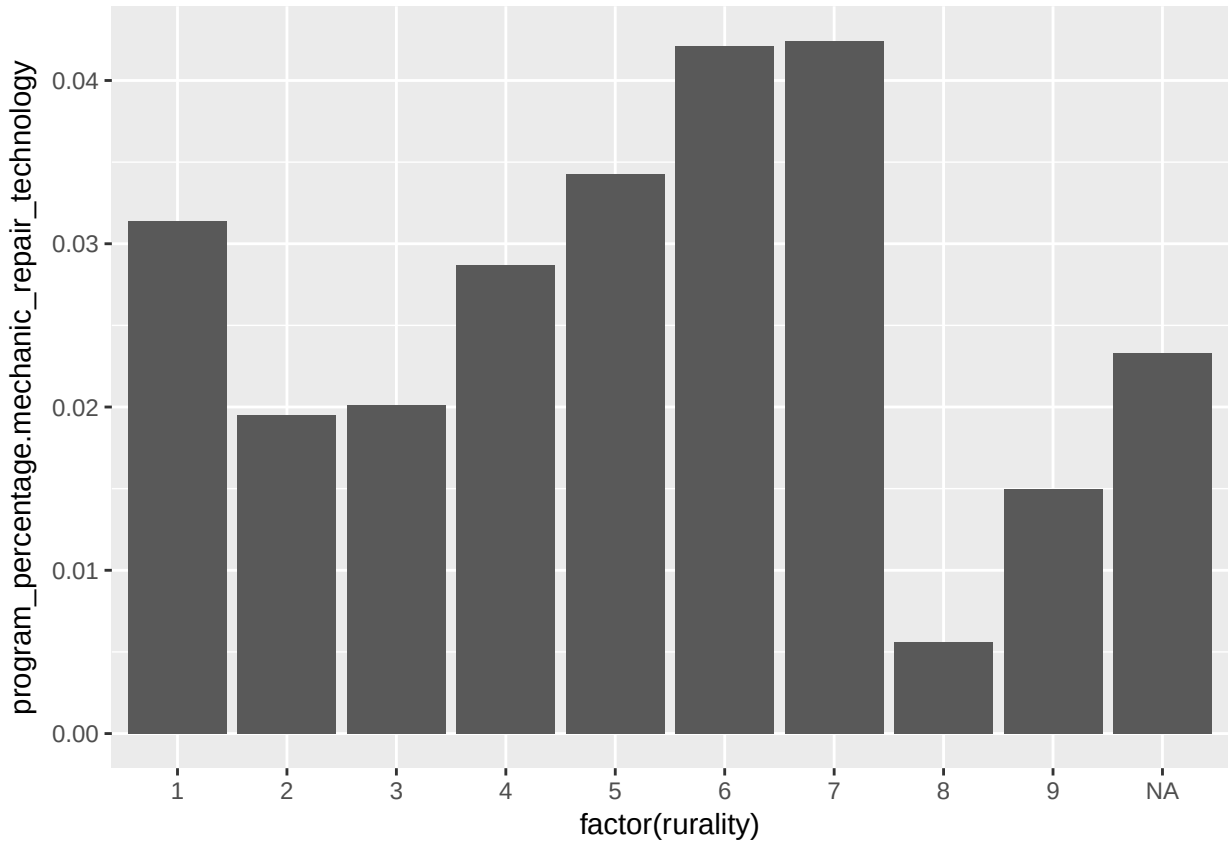

```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=program_percentage.ethnic
```

```
## Warning: Removed 739 rows containing non-finite values (stat_summary).
```



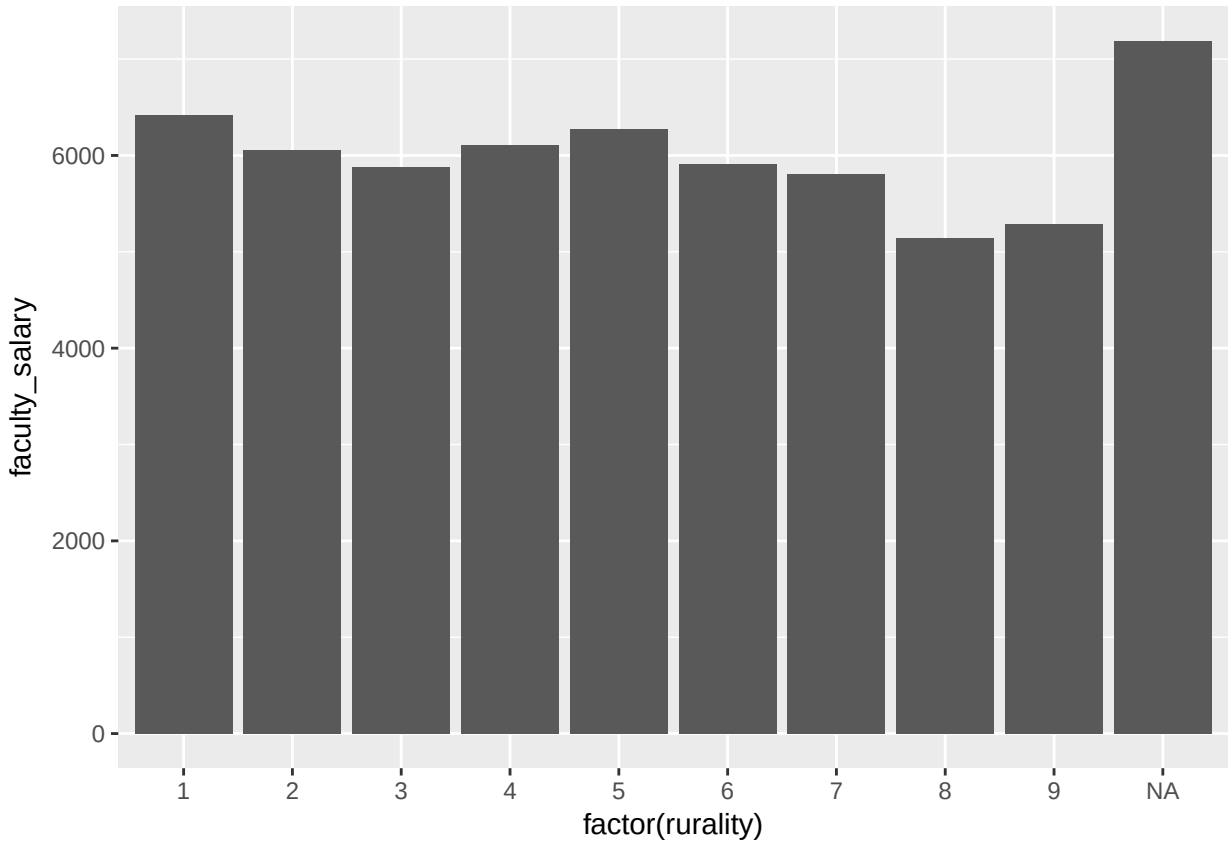

```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=program_percentage.mechani
```

```
## Warning: Removed 739 rows containing non-finite values (stat_summary).
```



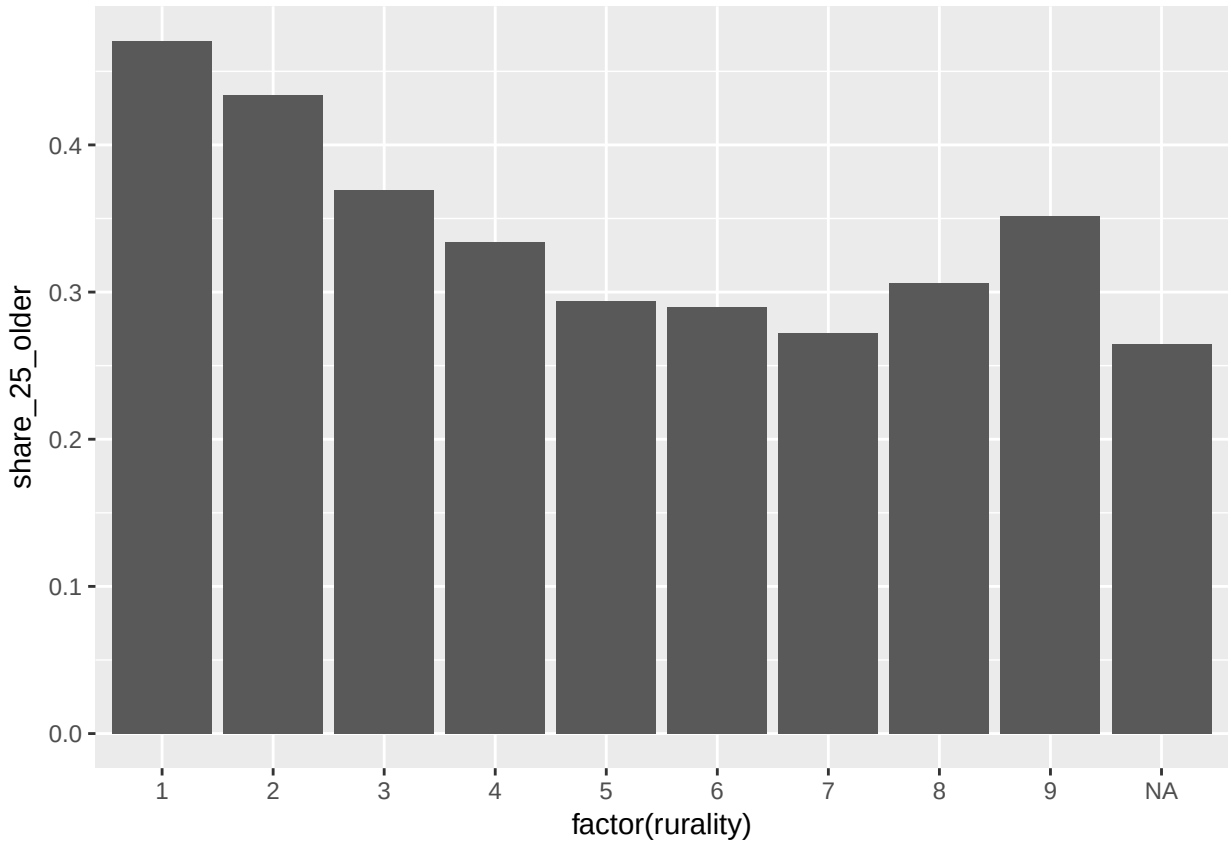

```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=faculty_salary)) + stat_su
```

```
## Warning: Removed 3139 rows containing non-finite values (stat_summary).
```



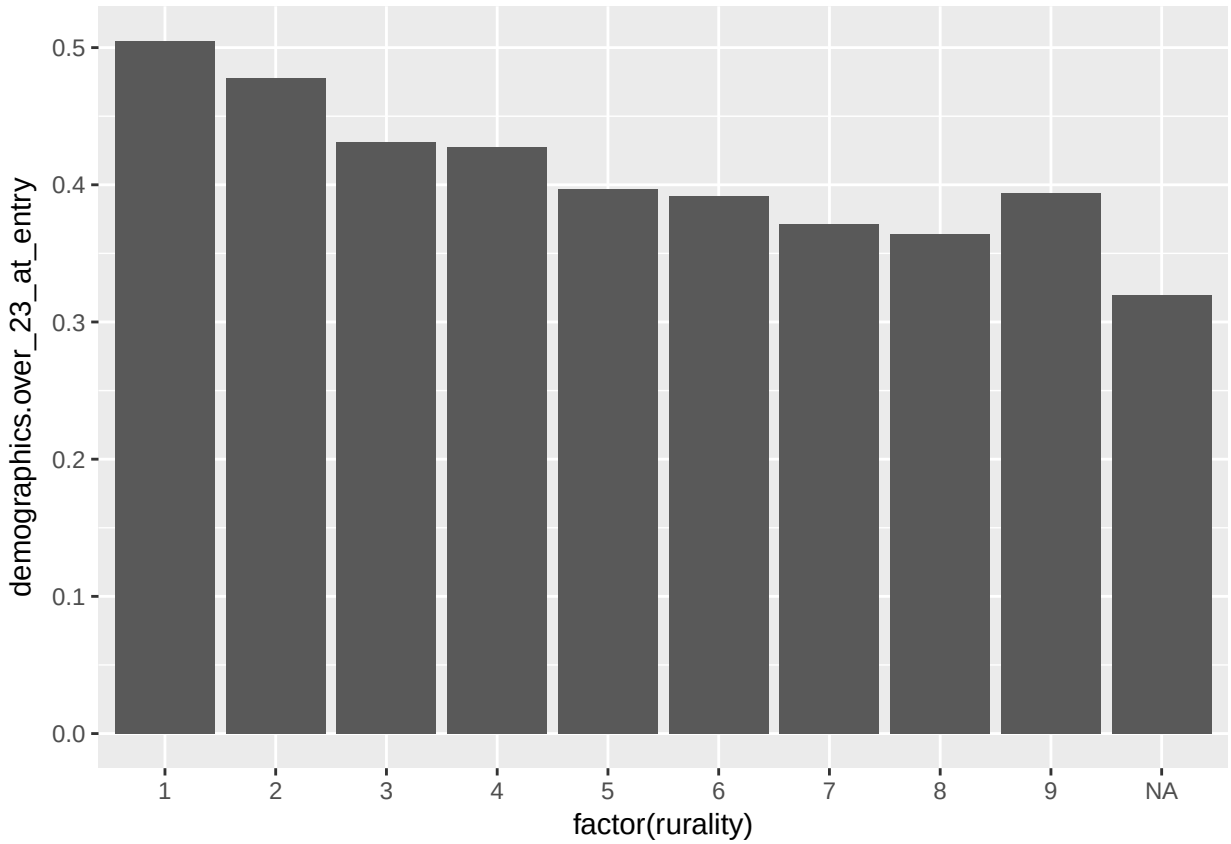

```
# Looks like metro areas have more return adults
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=share_25_older)) + stat_su
```

```
## Warning: Removed 805 rows containing non-finite values (stat_summary).
```



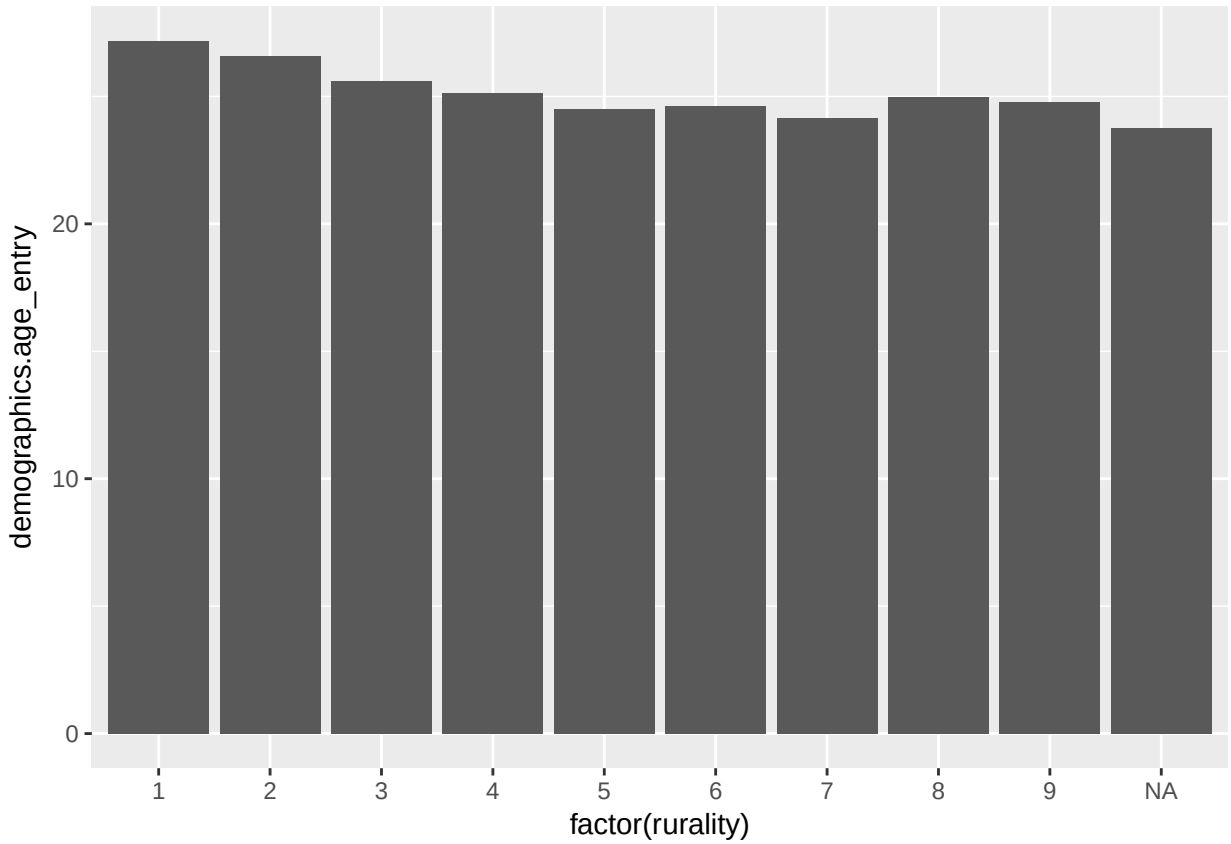

```
ggplot(collegeData[collegeData$year == "2005_06",], aes(x=factor(rurality), y=demographics.over_23_at_e
```

```
## Warning: Removed 285 rows containing non-finite values (stat_summary).
```



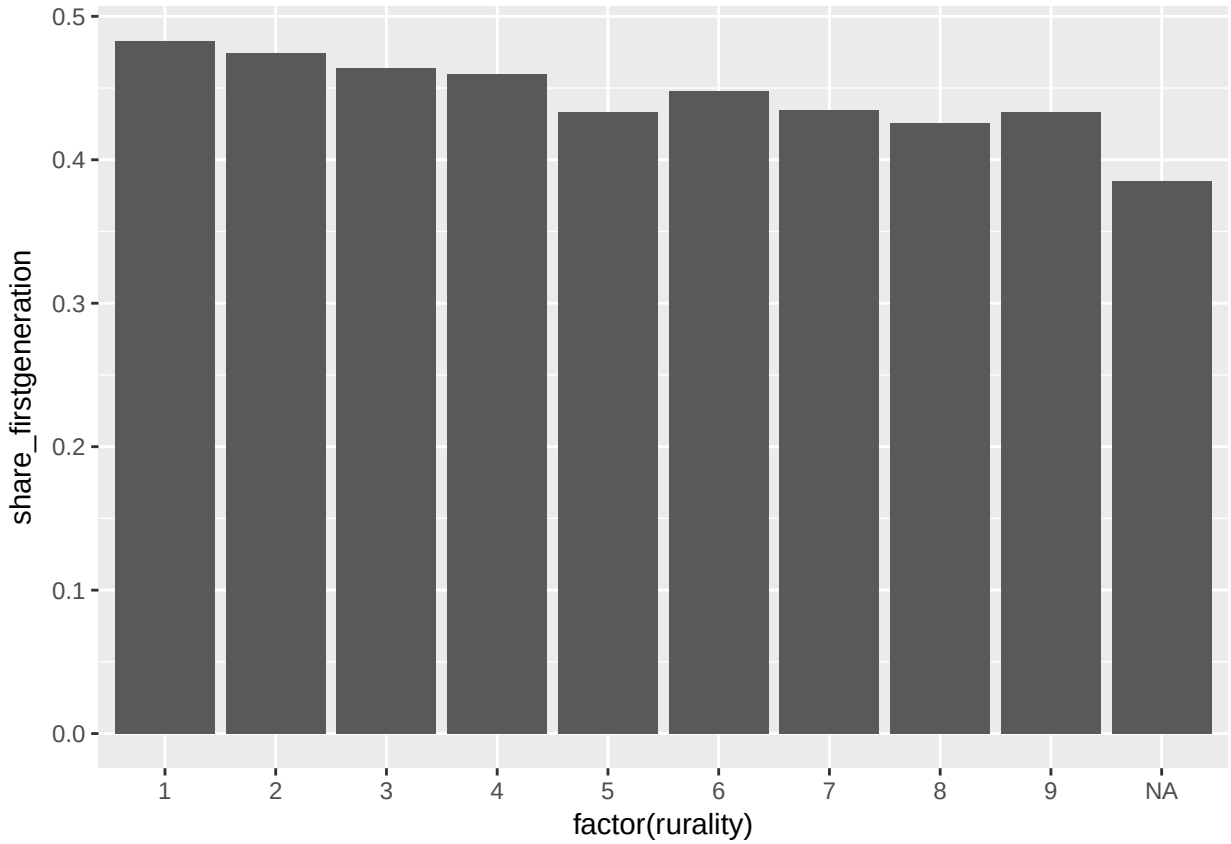

```
# But average age remains in the 20s across ruralities
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=demographics.age_entry)) +
```

```
## Warning: Removed 333 rows containing non-finite values (stat_summary).
```



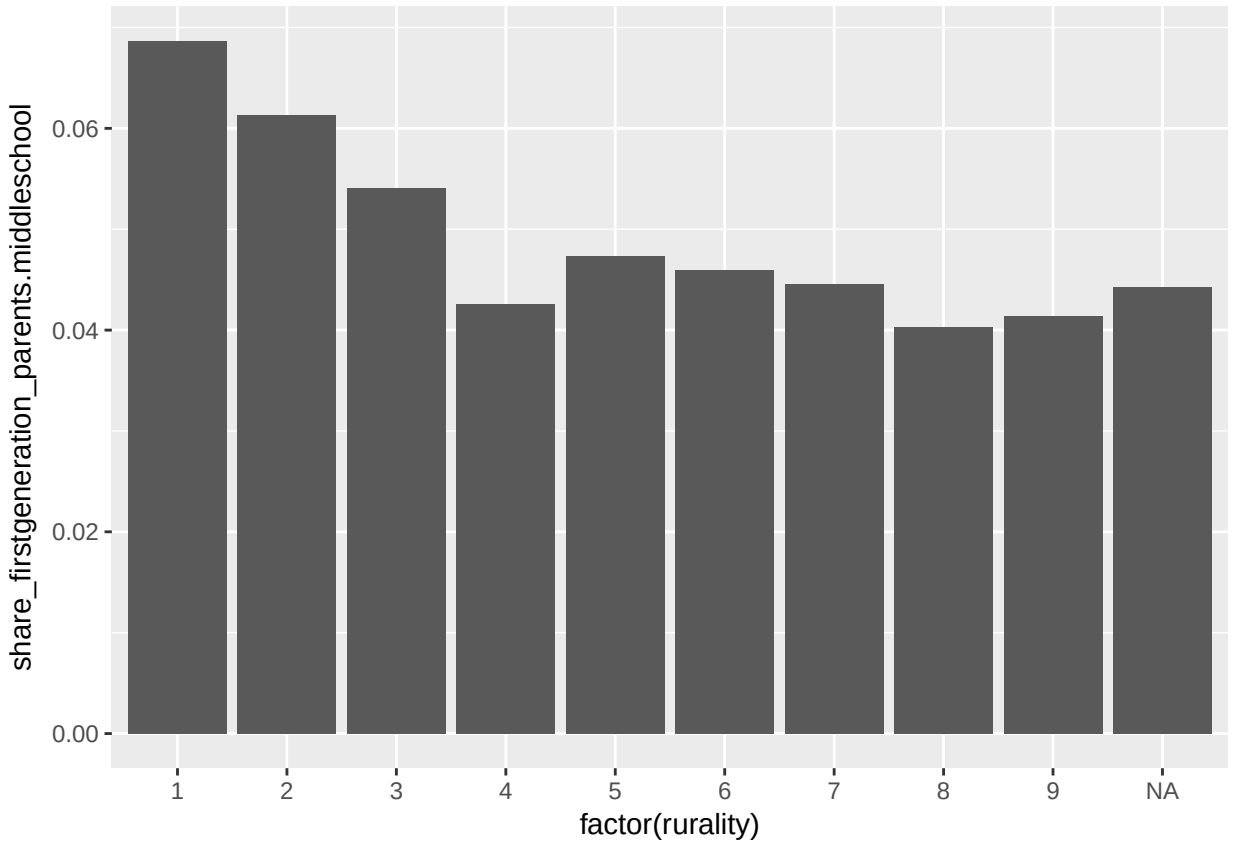

```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=share_firstgeneration)) +
```

```
## Warning: Removed 1118 rows containing non-finite values (stat_summary).
```



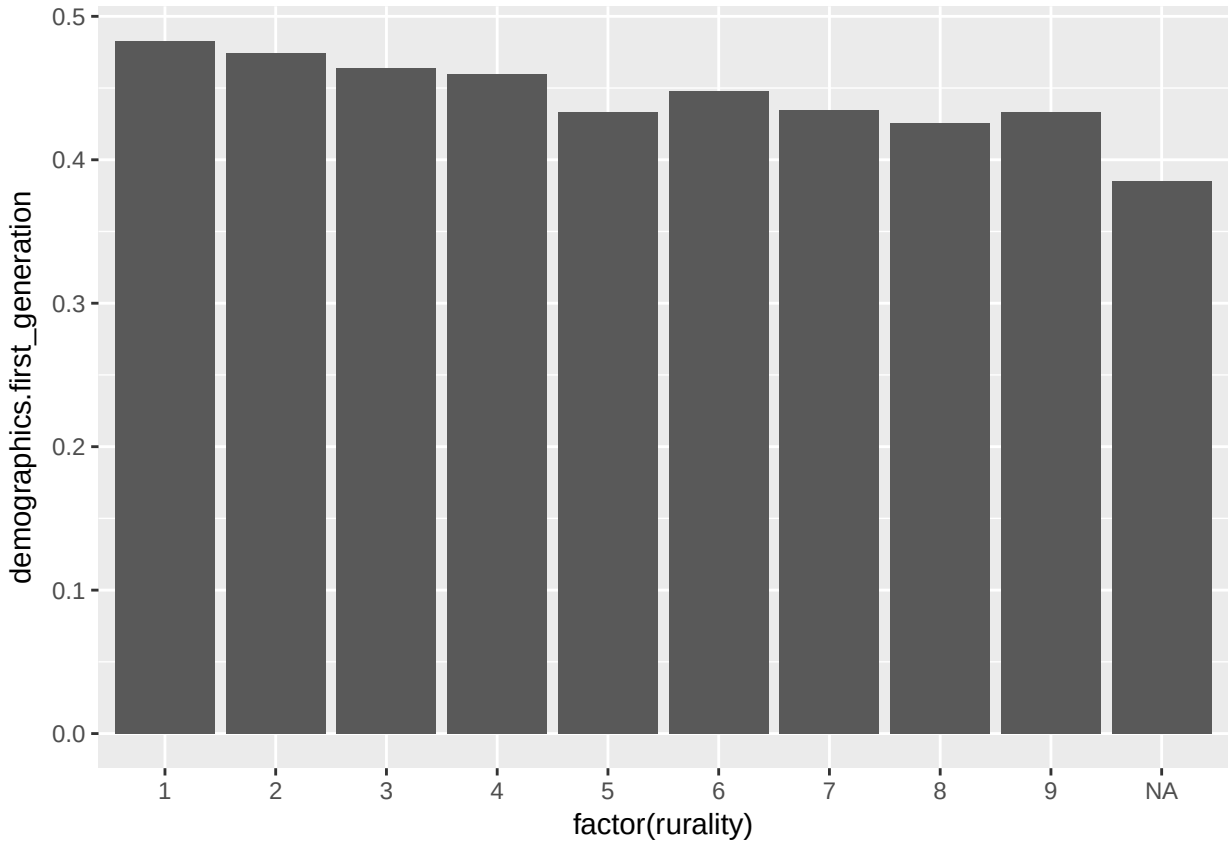

```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=share_firstgeneration_par
```

```
## Warning: Removed 2595 rows containing non-finite values (stat_summary).
```

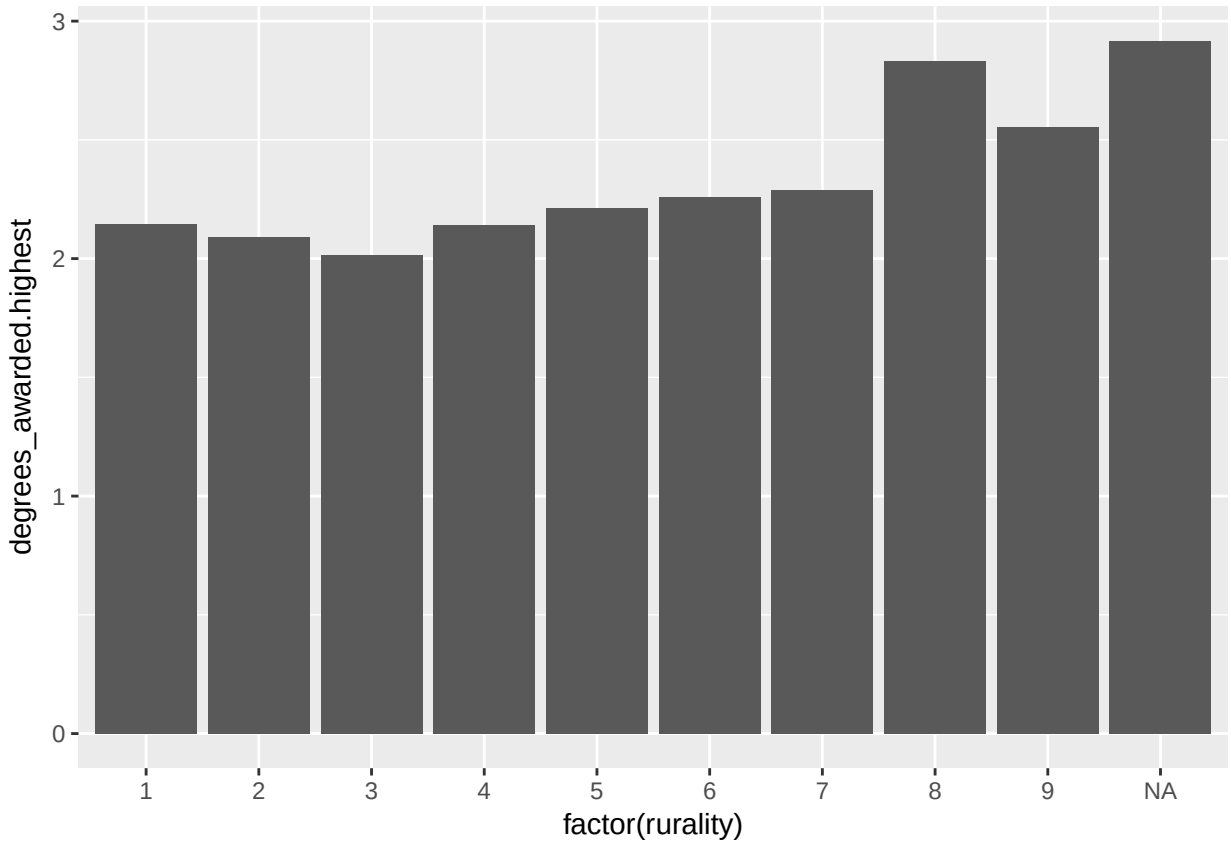



```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=demographics.first_genera
```

```
## Warning: Removed 1118 rows containing non-finite values (stat_summary).
```

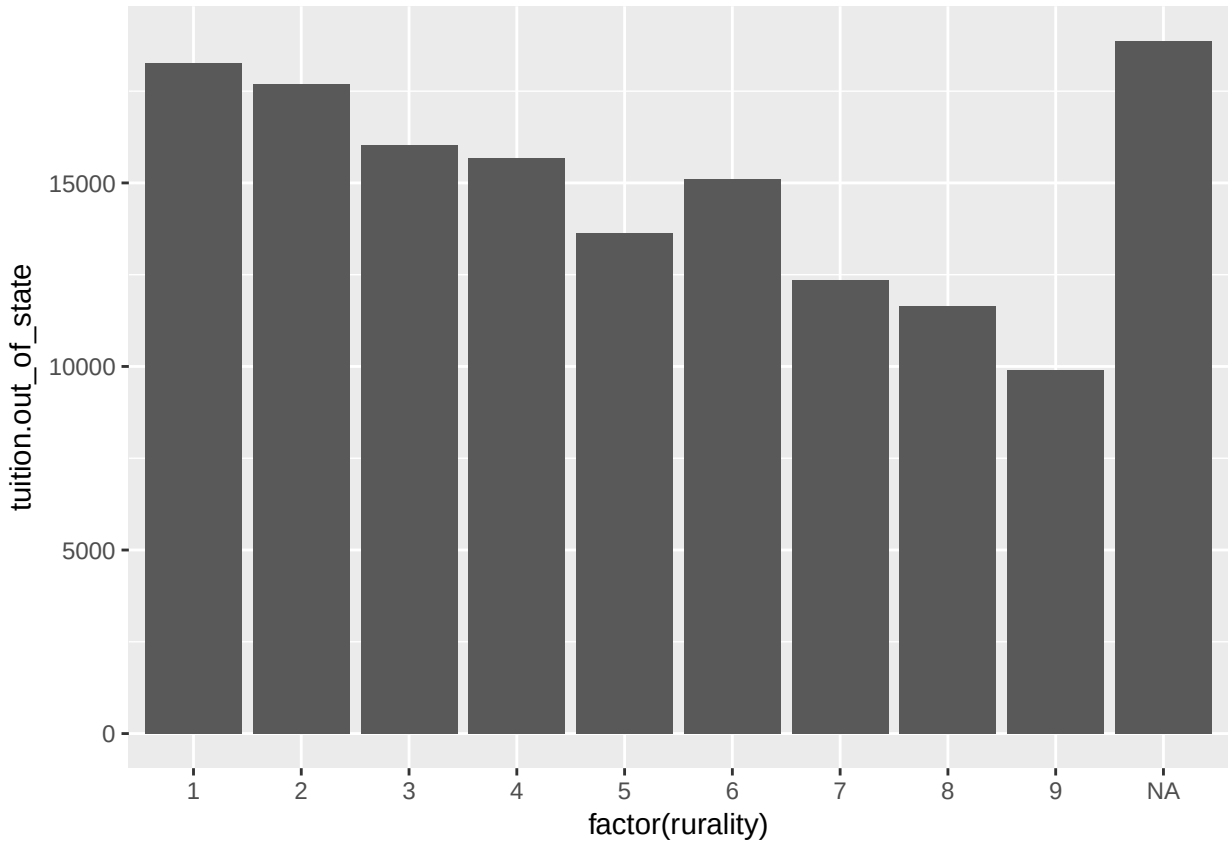



```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=degrees_awarded.highest))
```



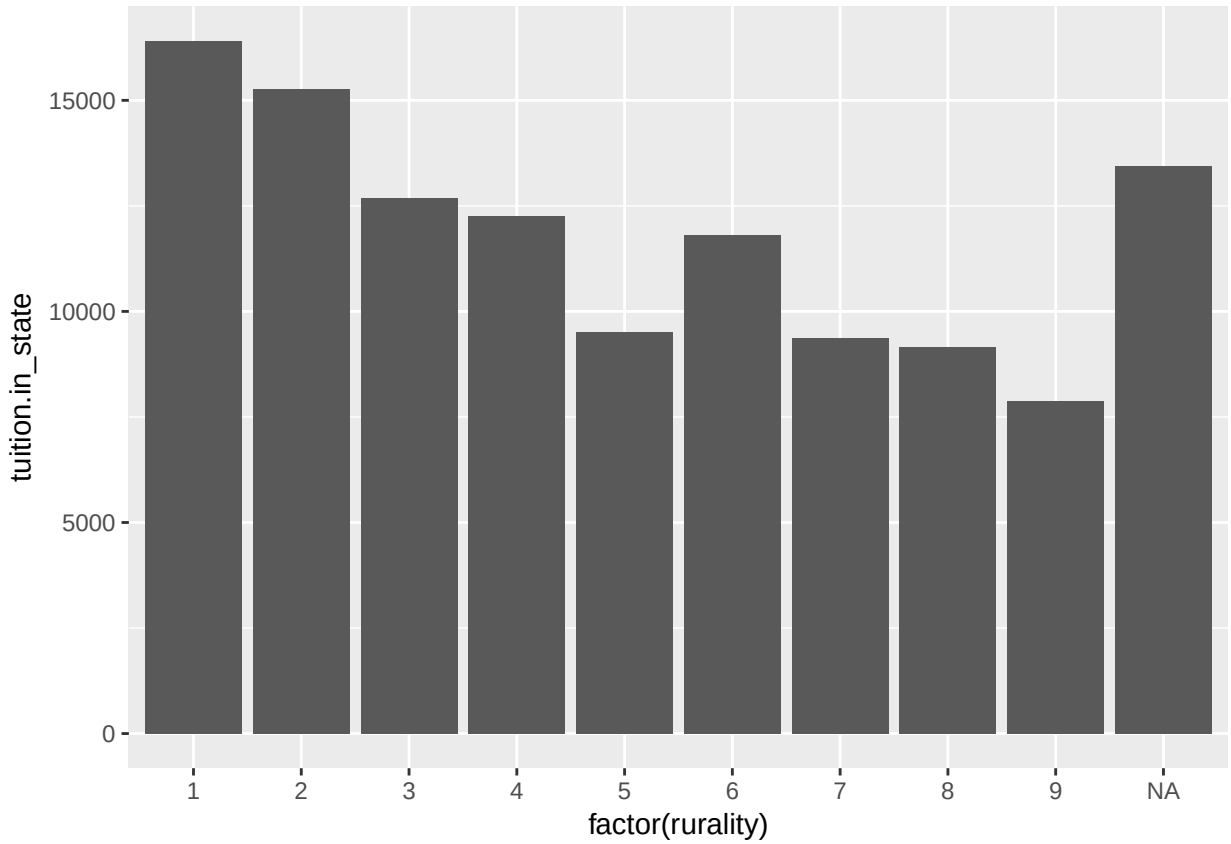

```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=tuition.out_of_state)) + s
```

```
## Warning: Removed 3504 rows containing non-finite values (stat_summary).
```



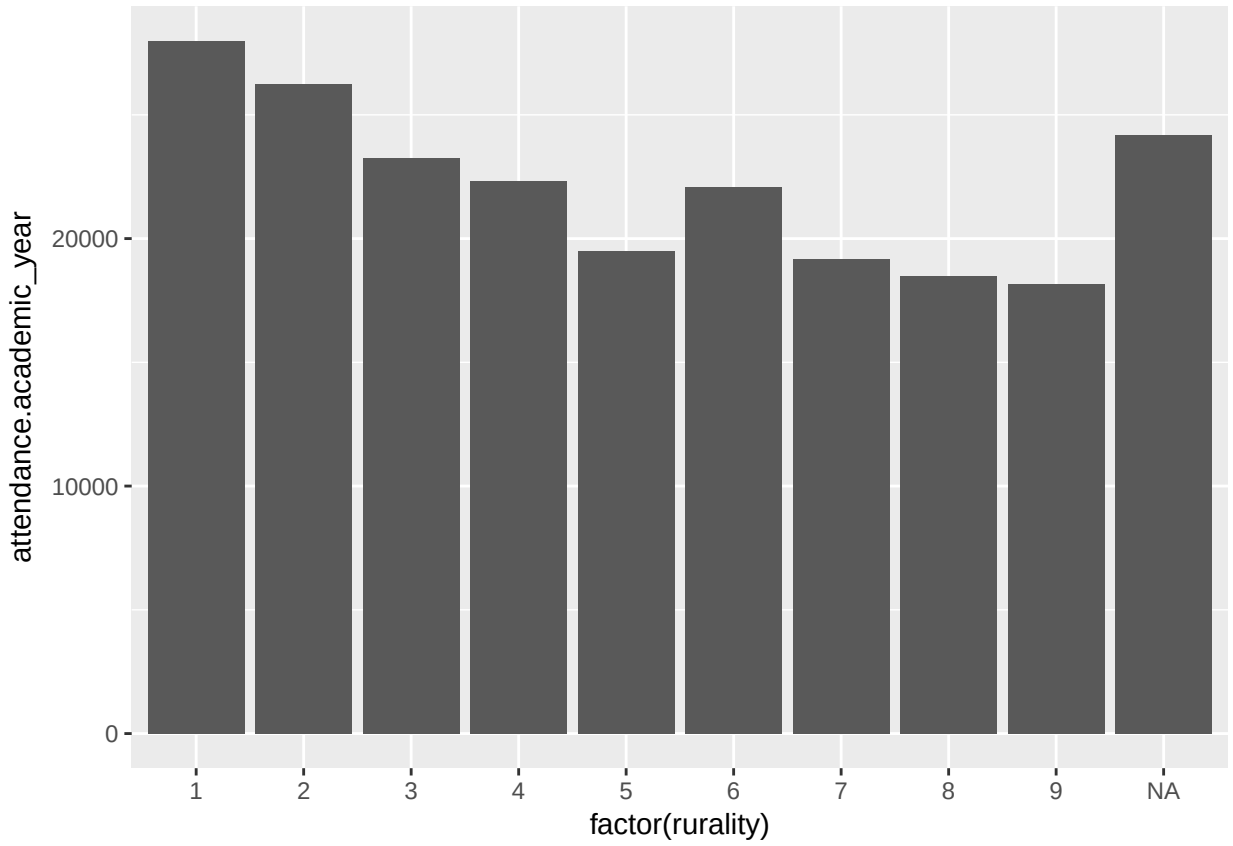

```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=tuition.in_state)) + stat_
```

```
## Warning: Removed 3290 rows containing non-finite values (stat_summary).
```




```
ggplot(collegeData[collegeData$year == "2015_16",], aes(x=factor(rurality), y=attendance.academic_year)
```

```
## Warning: Removed 3673 rows containing non-finite values (stat_summary).
```


attendance.academic_year
20000
10000
0
1
2
3
4
5
6
7
8
9
NA
factor(rurality)