Missing Analysis

Dominik Glandorf

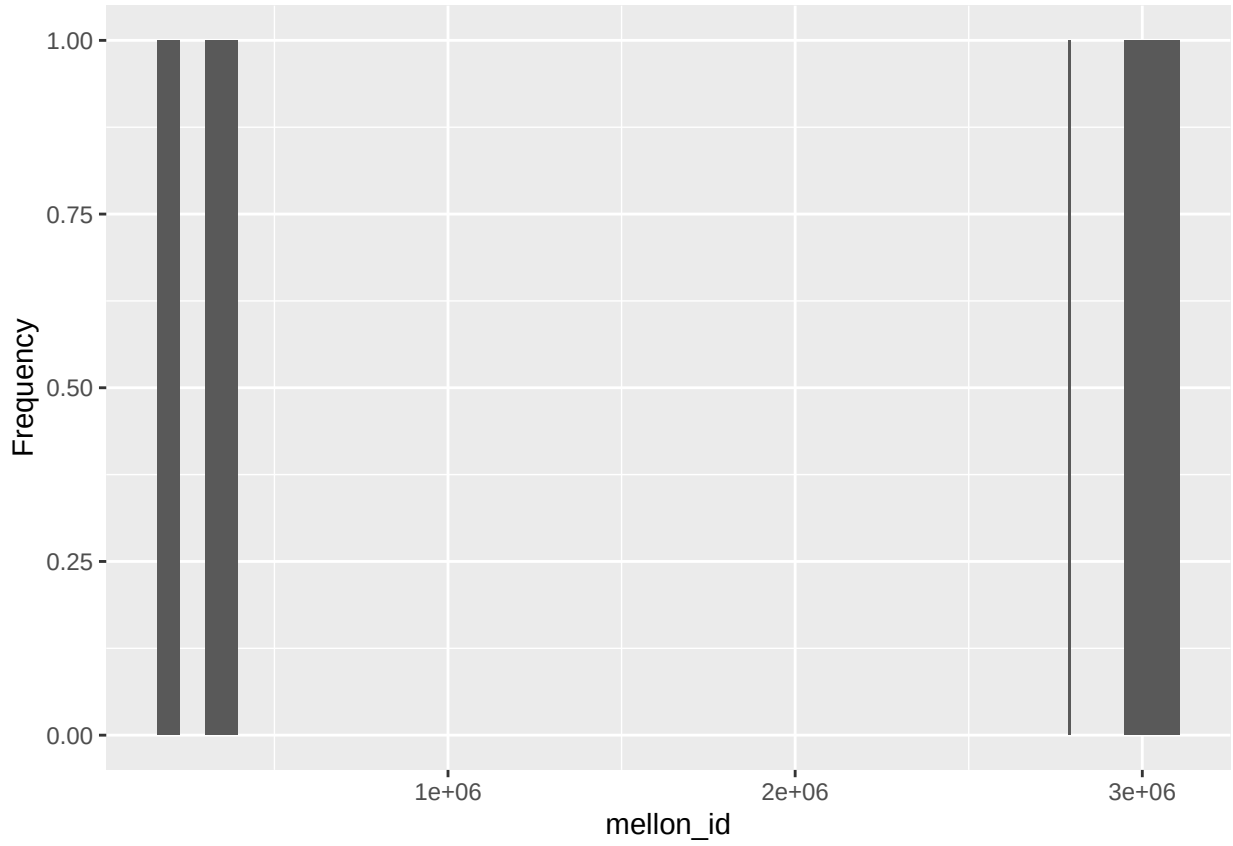
2022-11-22

Background variables

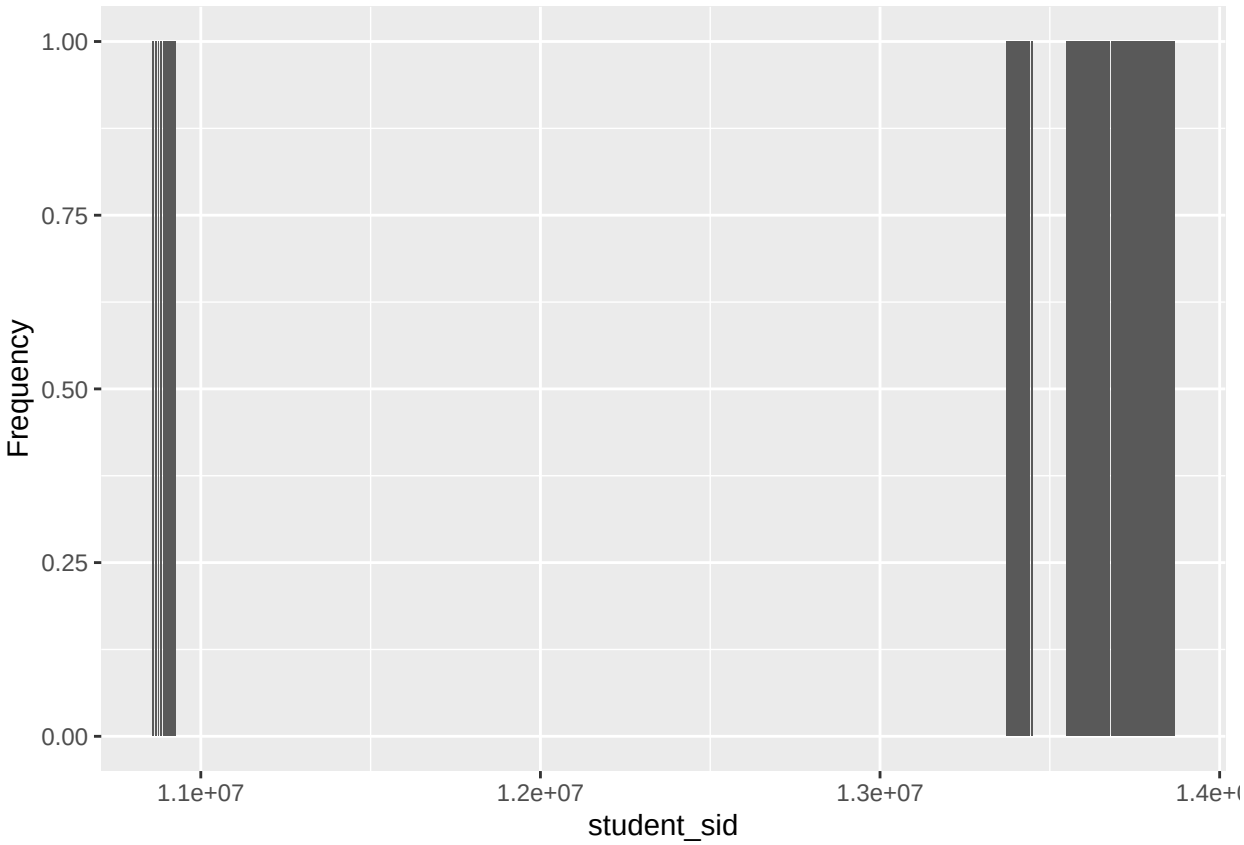
The spreadsheet knows 161 variables. The corresponding file contains 176 columns. The expected variables freshman,transfer are not in the dataset. Therefore exist variables pascd,hscd,must_hsid,ncessch,hs_address,hs_city,hs_state

Requested and available variables

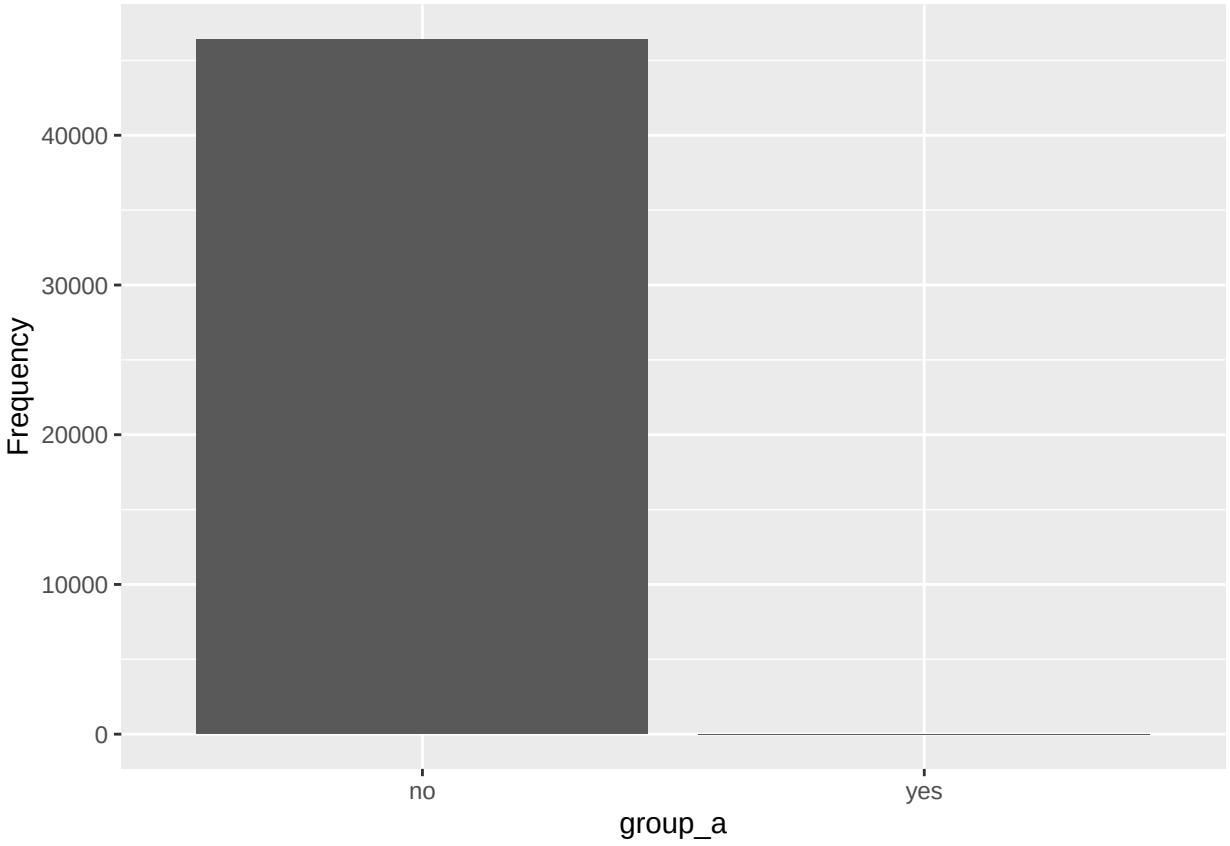
```
[1] Variable: mellon_id, type: numeric
[1] Values (46408 unique): 300890, 189970, 314750, 173027, 195012, ...
[1] Missing: 0%
```



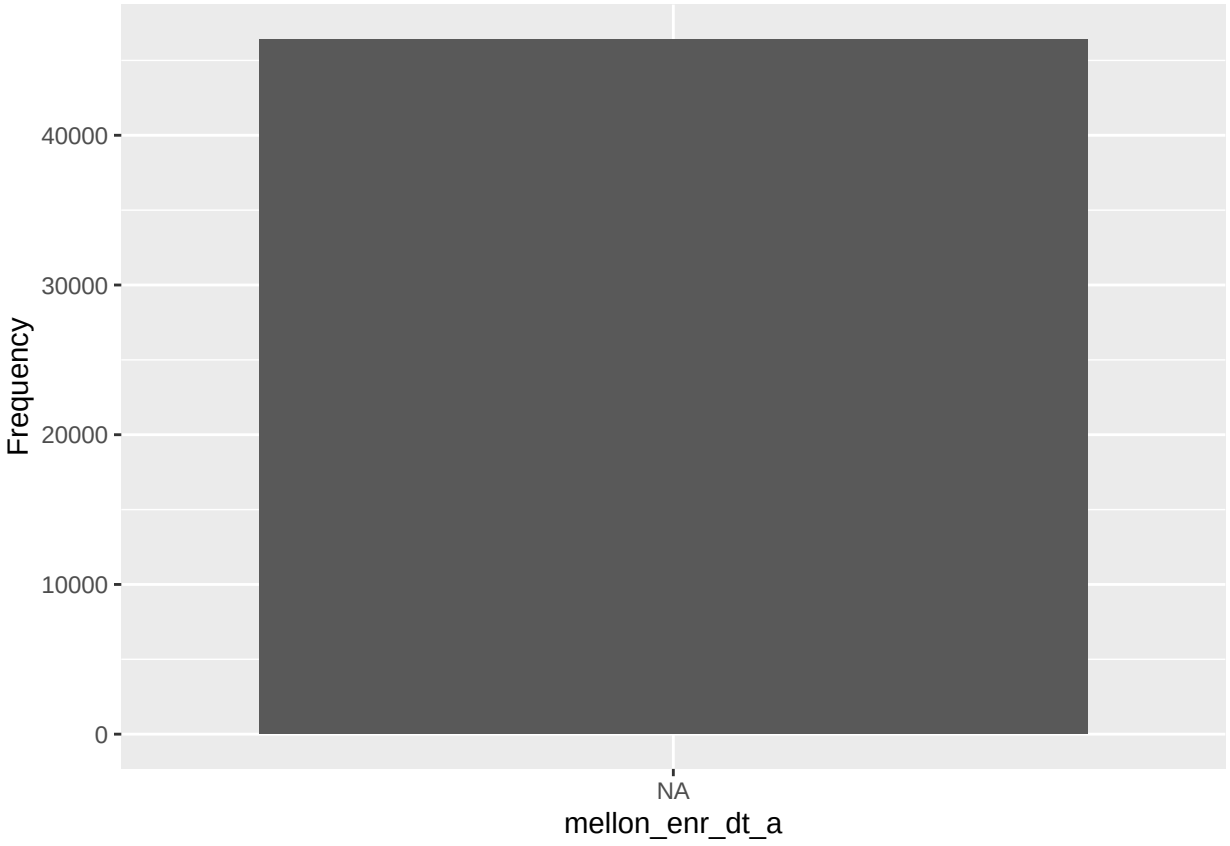
```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: student_sid, type: numeric
[1] Values (46408 unique): 10857406, 10859561, 10861911, 10861952, 10865378, ...
[1] Missing: 0%
```



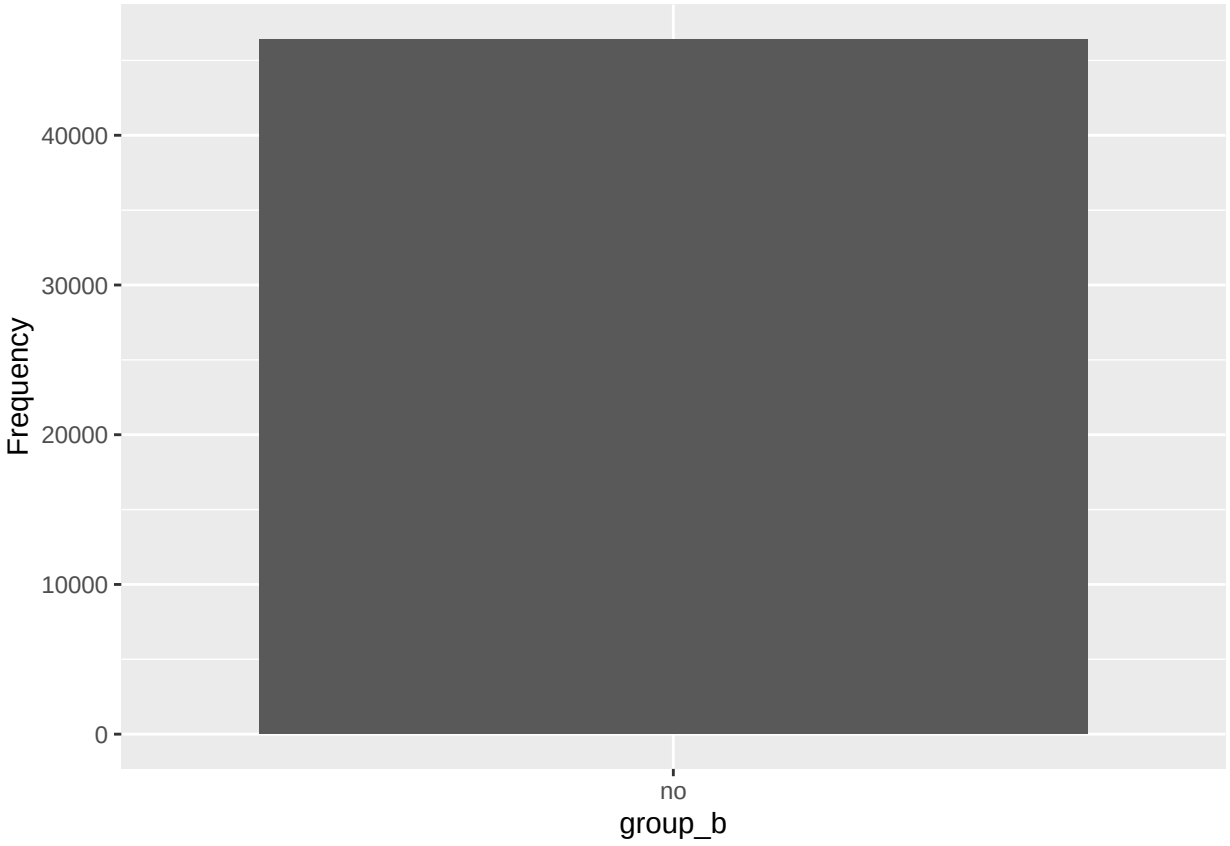

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: group_a, type: character
[1] Values (2 unique): no, yes
[1] Missing: 0%
```



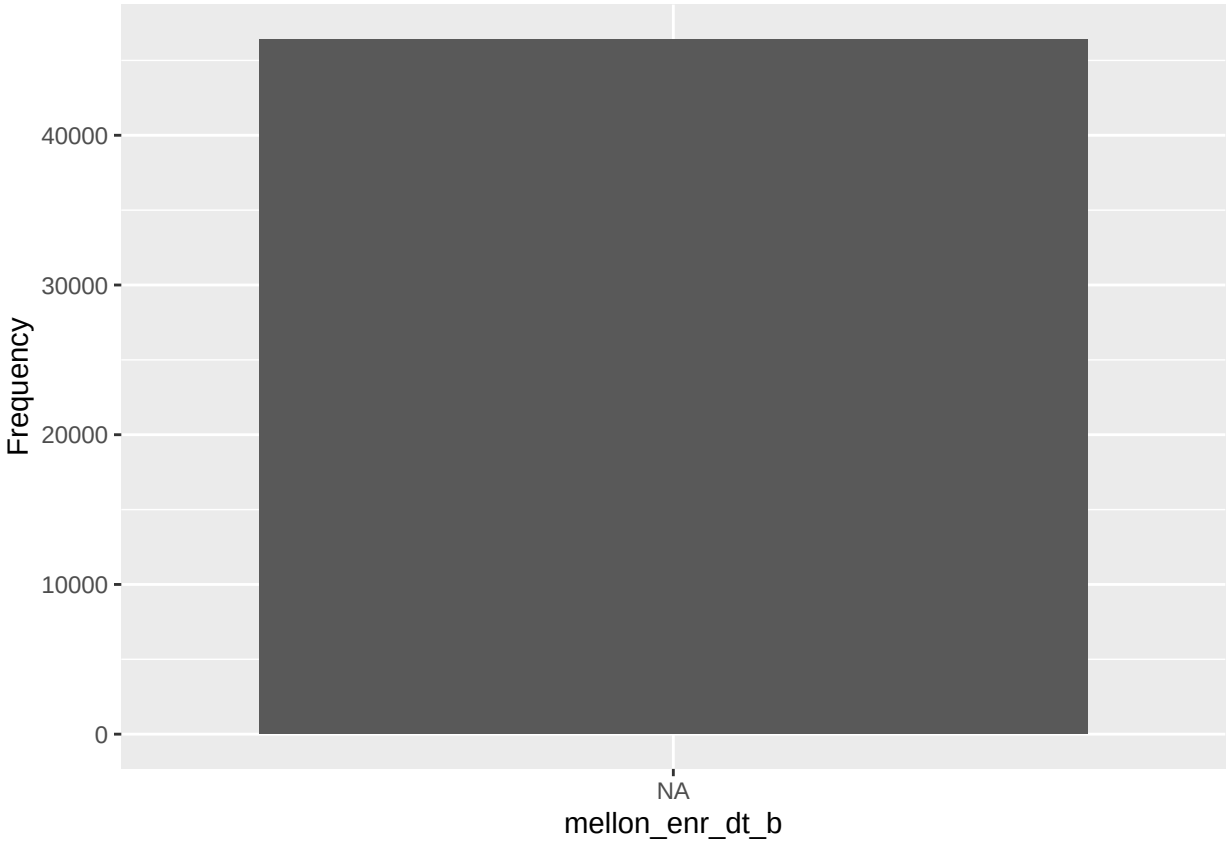

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: mellon_enr_dt_a, type: character
[1] Values (1 unique): NA
[1] Missing: 100%
```



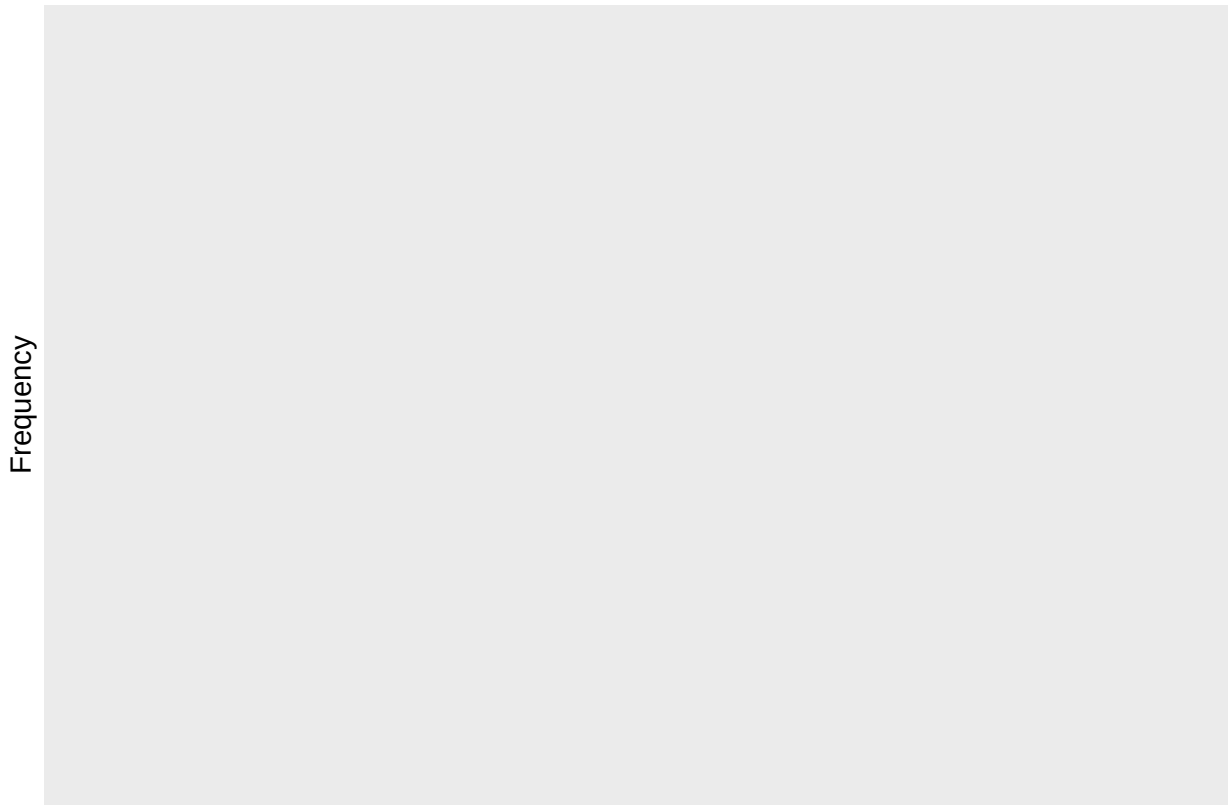

```
[1] should not be used as a predictor
[1] ___________________________
[1] Variable: group_b, type: character
[1] Values (1 unique): no
[1] Missing: 0%
```



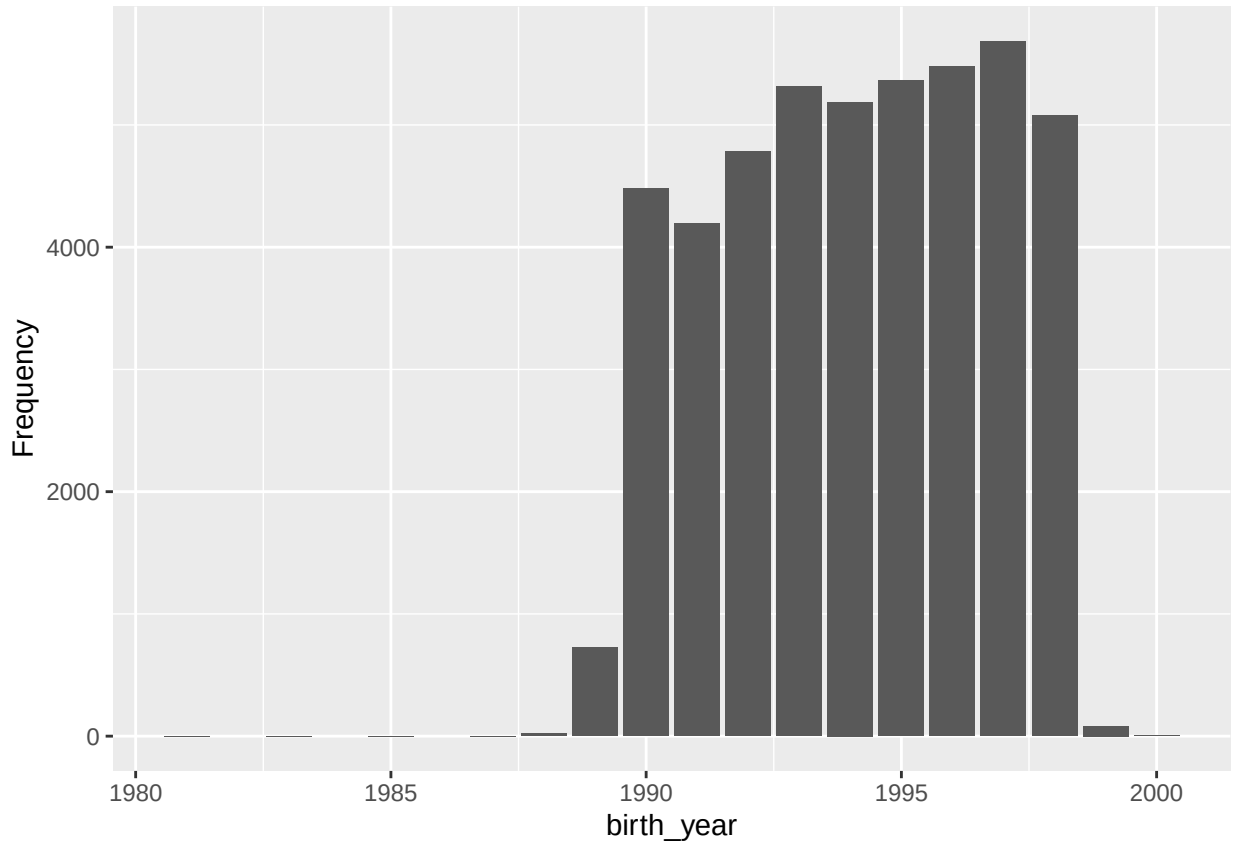

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: mellon_enr_dt_b, type: character
[1] Values (1 unique): NA
[1] Missing: 100%
```



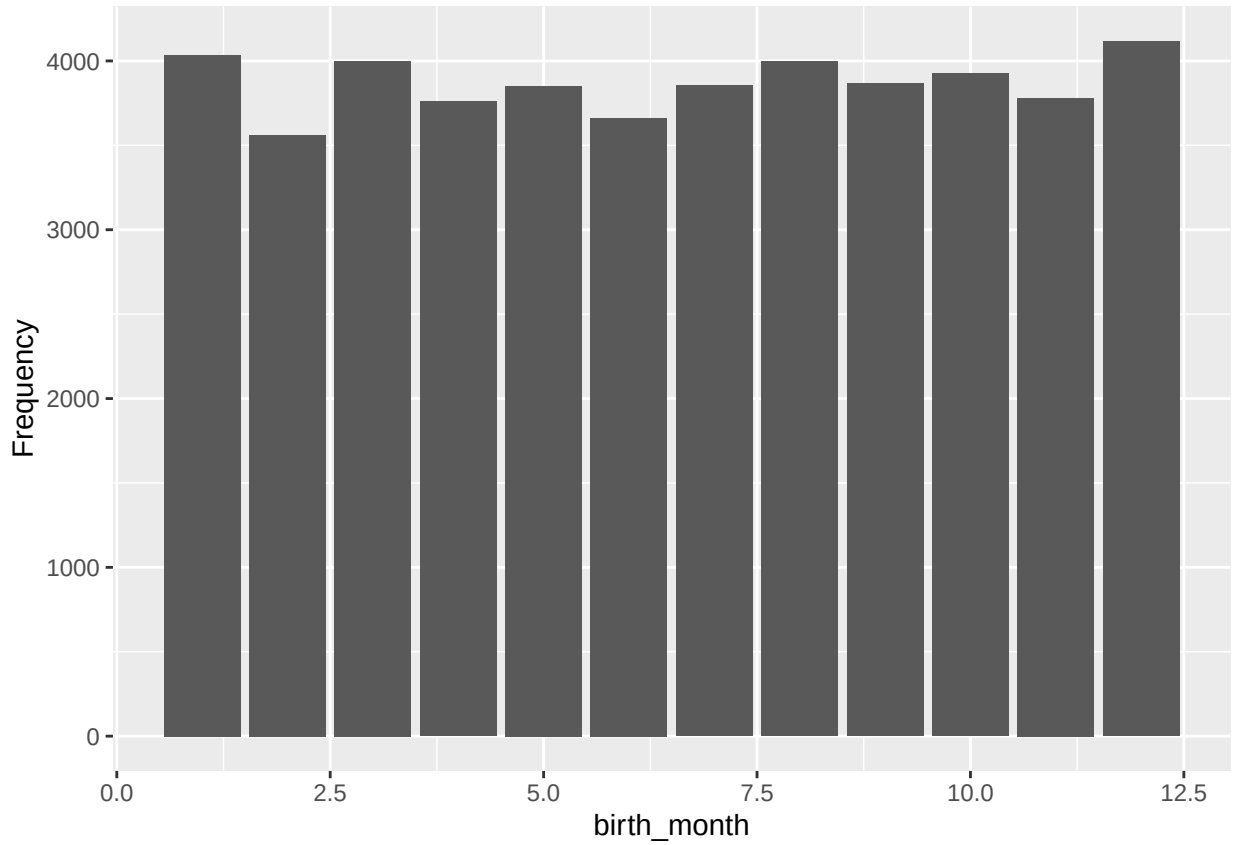

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: mellon_yr, type: numeric
[1] Values (1 unique): NA
[1] Missing: 100%

Warning in min(x): kein nicht-fehlendes Argument für min; gebe Inf zurück

Warning in max(x): kein nicht-fehlendes Argument für max; gebe -Inf zurück

Warning in min(diff(sort(x))): kein nicht-fehlendes Argument für min; gebe Inf
zurück

Warning: Removed 46408 rows containing non-finite values (‘stat_count()‘).
```



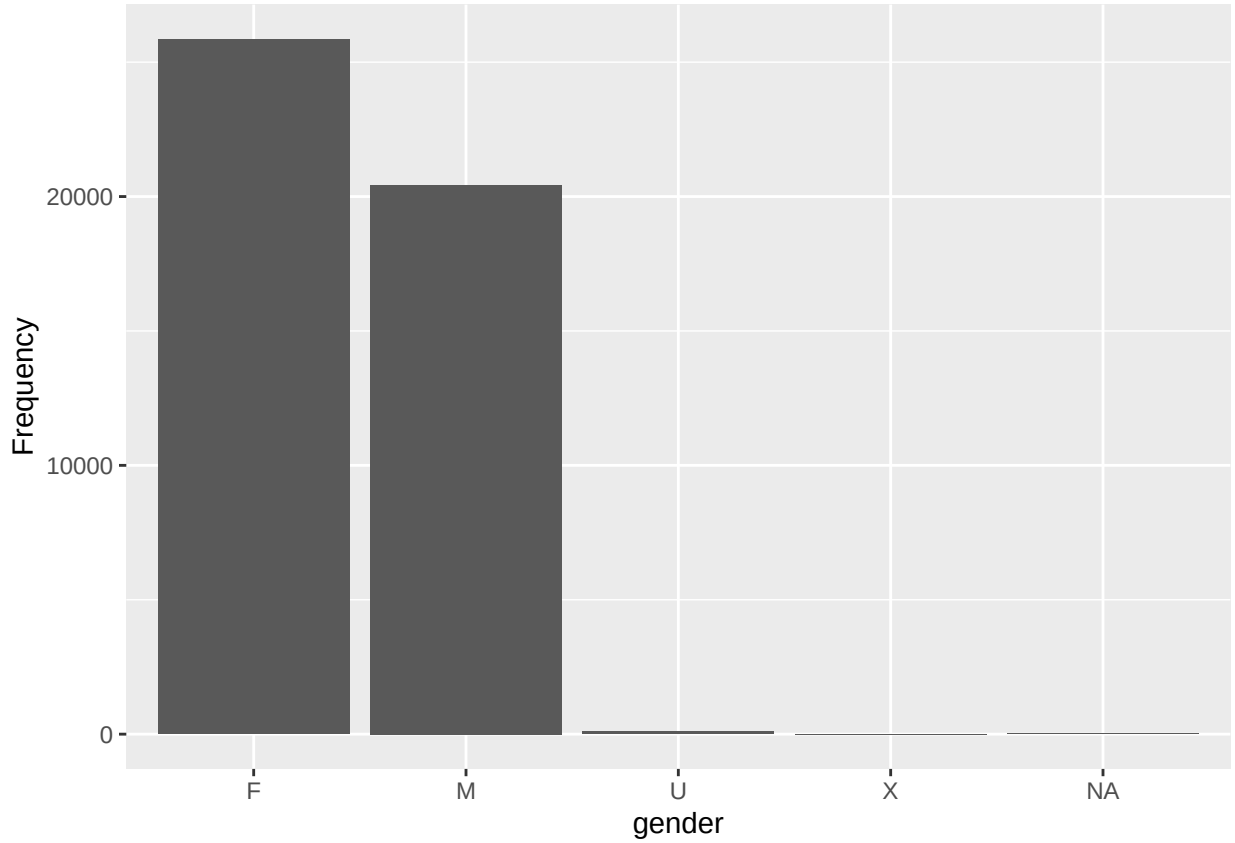

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: birth_year, type: numeric
[1] Values (17 unique): 1990, 1991, 1989, 1992, 1997, ...
[1] Missing: 0%
```



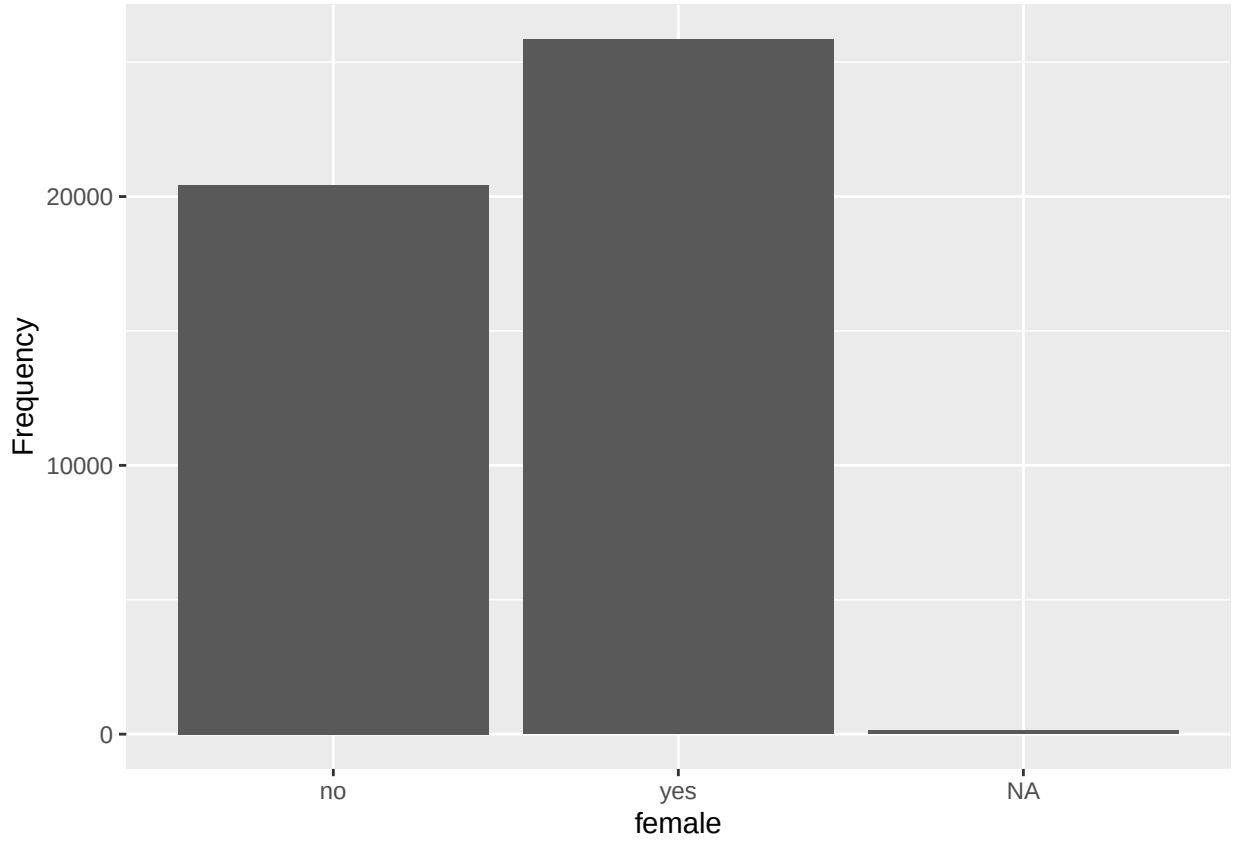

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: birth_month, type: numeric
[1] Values (12 unique): 5, 4, 7, 2, 10, ...
[1] Missing: 0%
```




```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: gender, type: character
[1] Values (5 unique): F, M, NA, X, U
[1] Missing: 0.1%
```

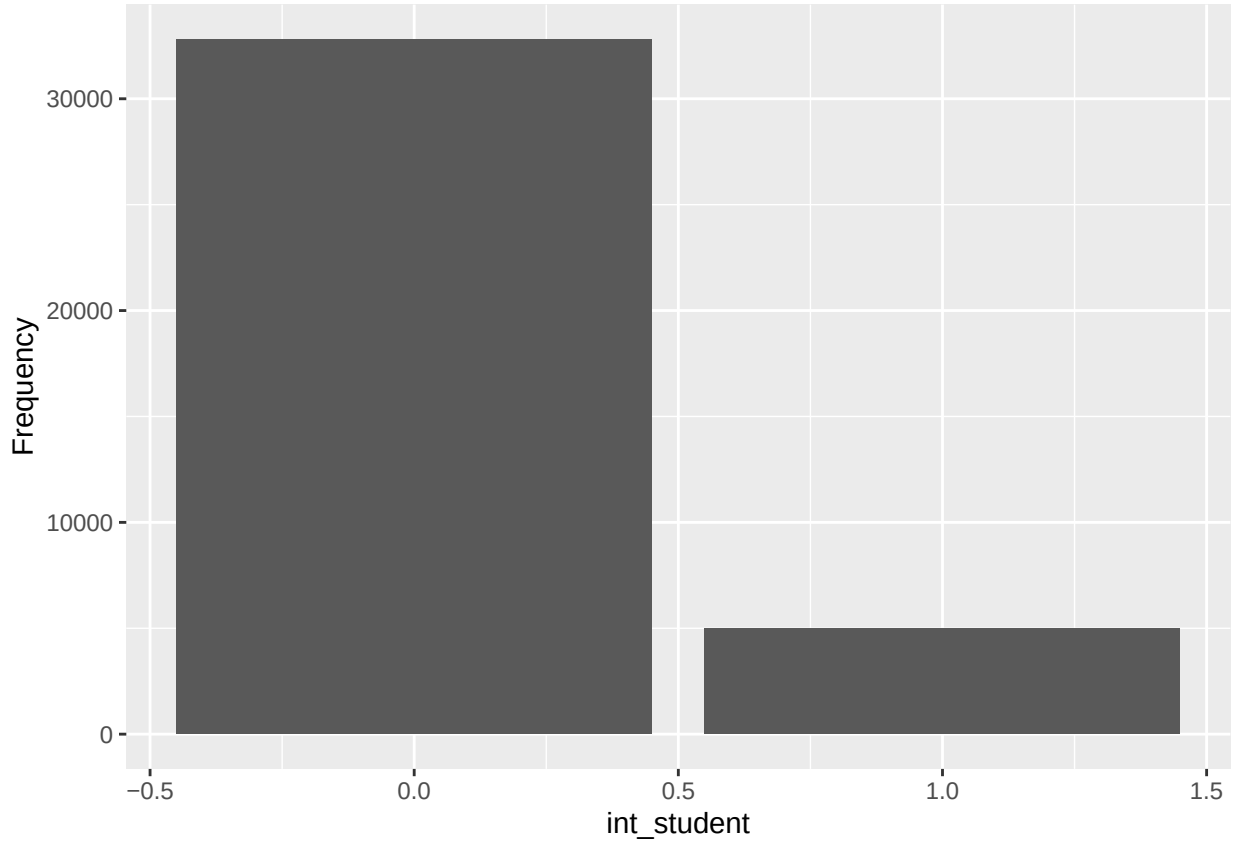



```
[1] ____________________________
[1] Variable: female, type: character
[1] Values (3 unique): yes, no, NA
[1] Missing: 0.3%
```

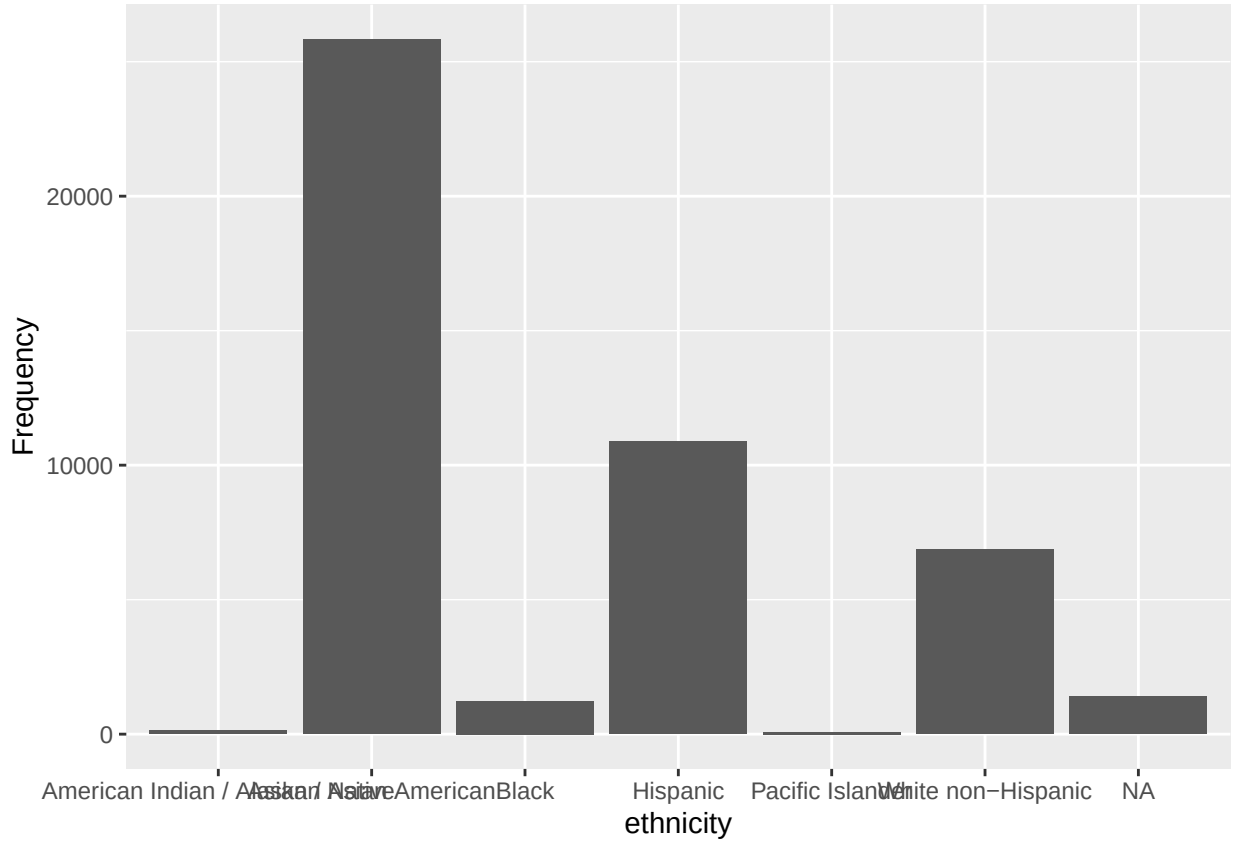



```
[1] ____________________________
[1] Variable: int_student, type: numeric
[1] Values (3 unique): 0, 1, NA
[1] Missing: 18.6%
  Group.1  int_student
1     F08 0.9960886571
2     F09 0.9955467590
3     F10 0.0002267574
4     F11 0.0001953507
5     F12 0.0001968892
6     F13 0.0000000000
7     F14 0.0003699593
8     F15 0.0001740644
9     F16 0.0000000000
```

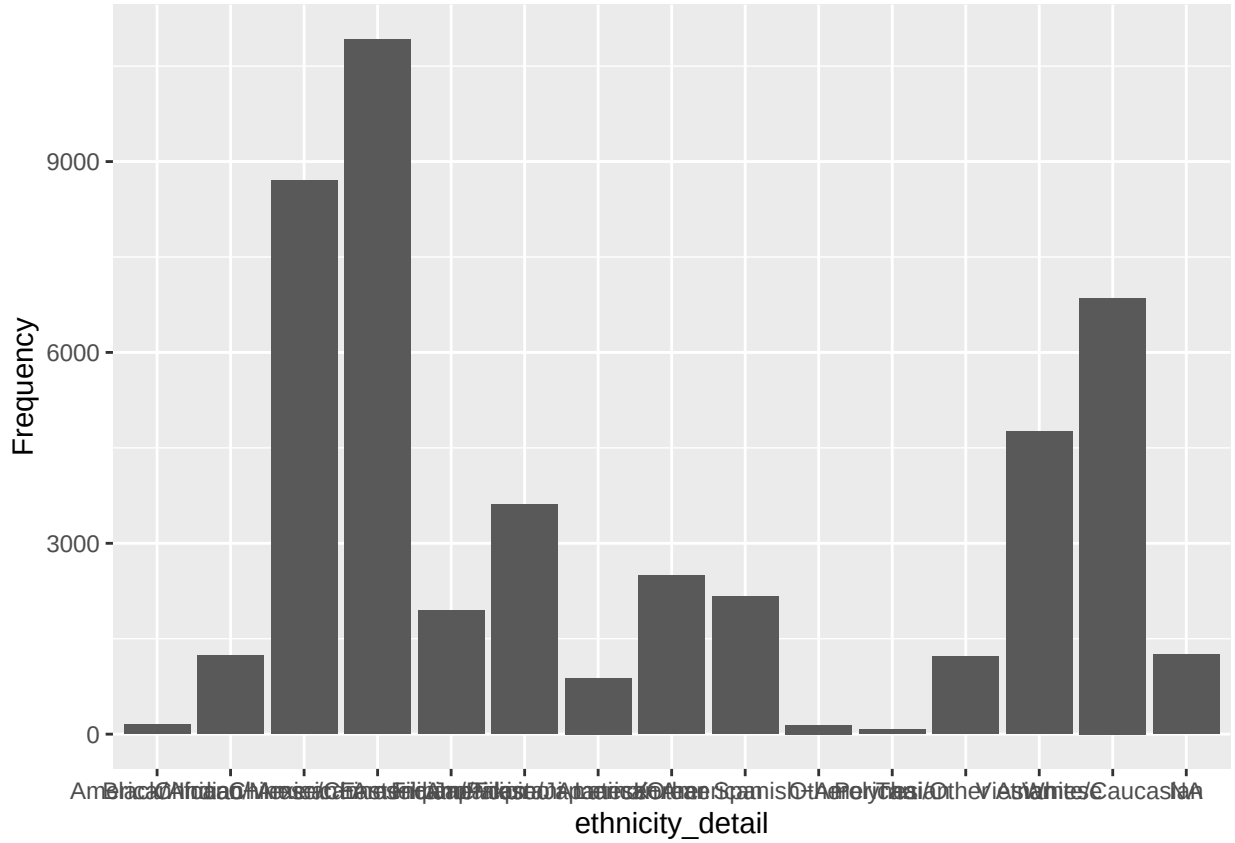
```
Warning: Removed 8614 rows containing non-finite values (‘stat_count()‘).
```



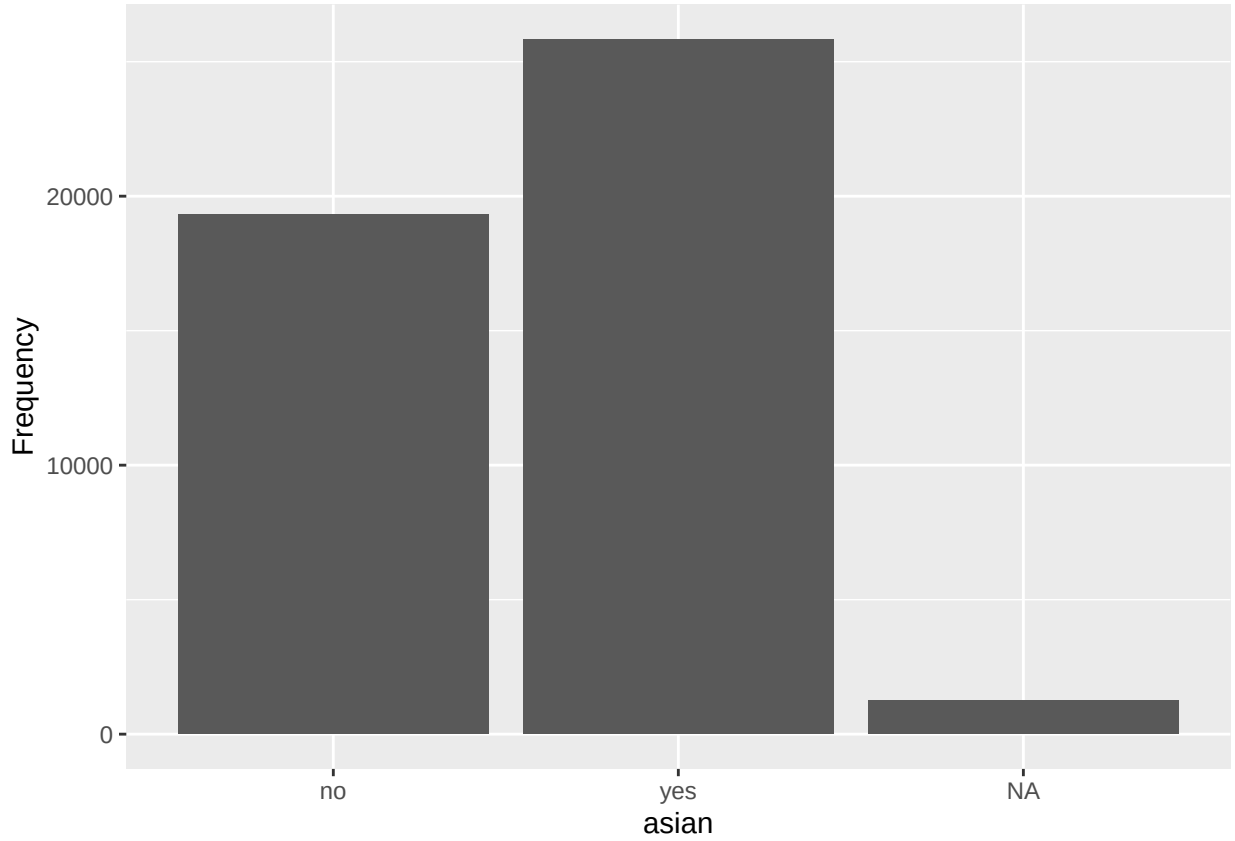

```
[1] ____________________________
[1] Variable: ethnicity, type: character
[1] Values (7 unique): Asian / Asian American, White non-Hispanic, Hispanic, NA, Black, ...
[1] Missing: 3%
[1] Most missing: F08 5.6%, Least missing: F11 1.5%
```



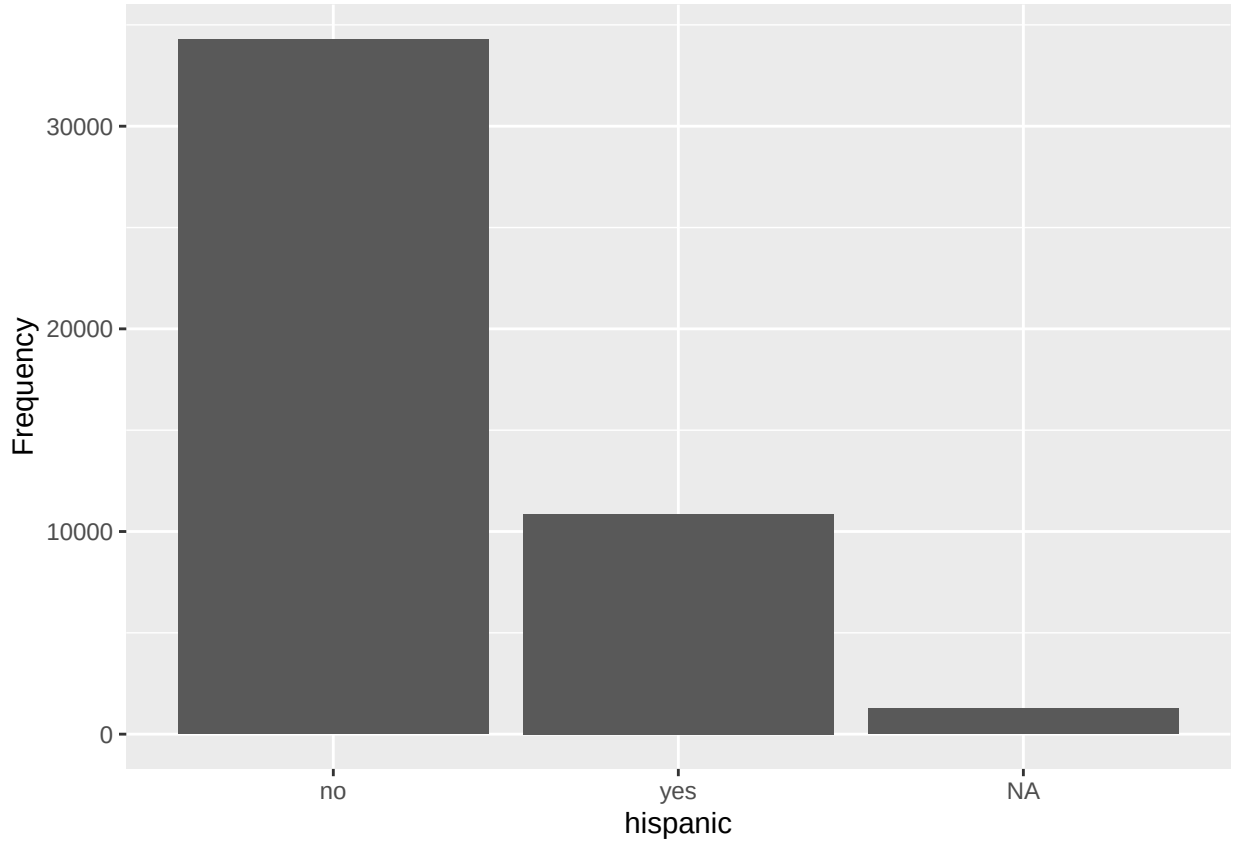

```
[1] ____________________________
[1] Variable: ethnicity_detail, type: character
[1] Values (15 unique): Filipino/Filipino American, White/Caucasian, Thai/Other Asian, Chicano/Mexican-A
[1] Missing: 2.7%
[1] Most missing: F10 4.4%, Least missing: F11 1.5%
```



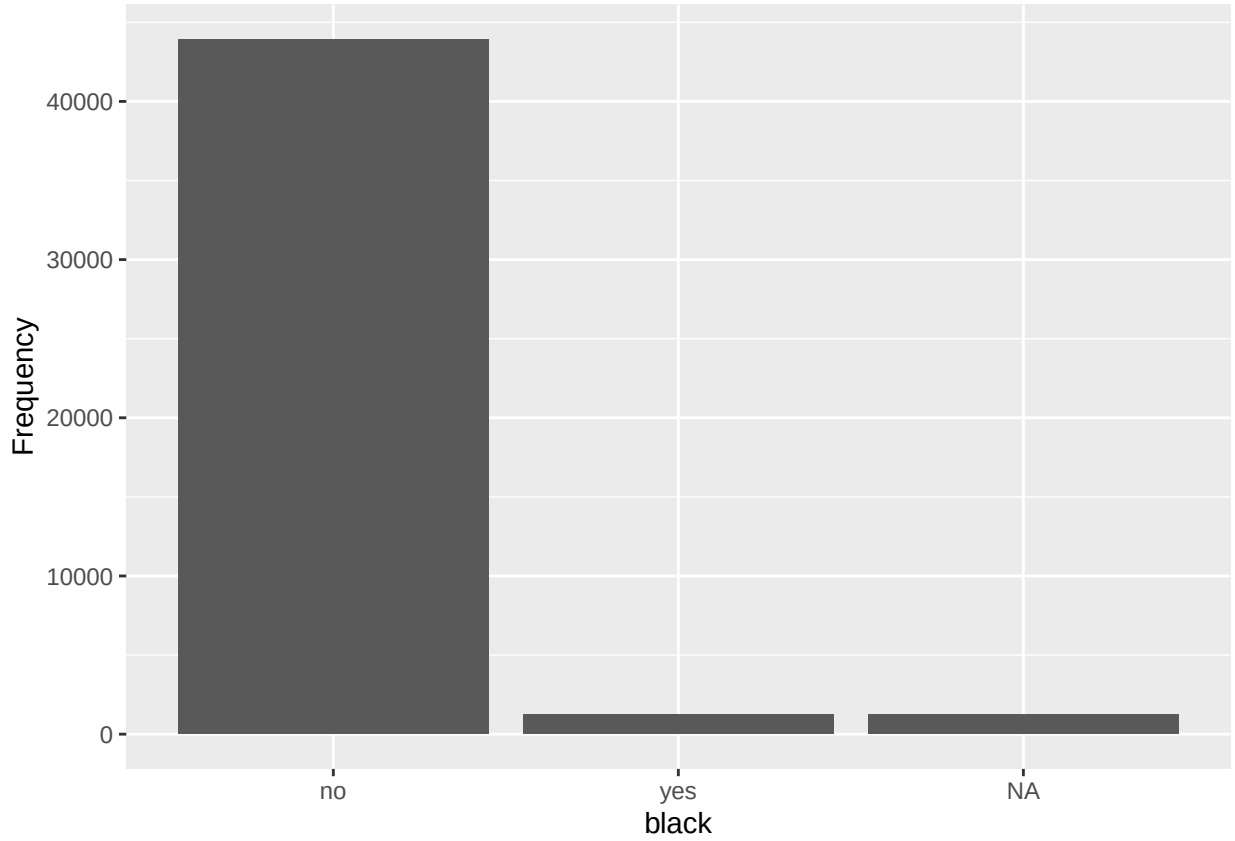

```
[1] ____________________________
[1] Variable: asian, type: character
[1] Values (3 unique): yes, no, NA
[1] Missing: 2.7%
[1] Most missing: F10 4.4%, Least missing: F11 1.5%
```



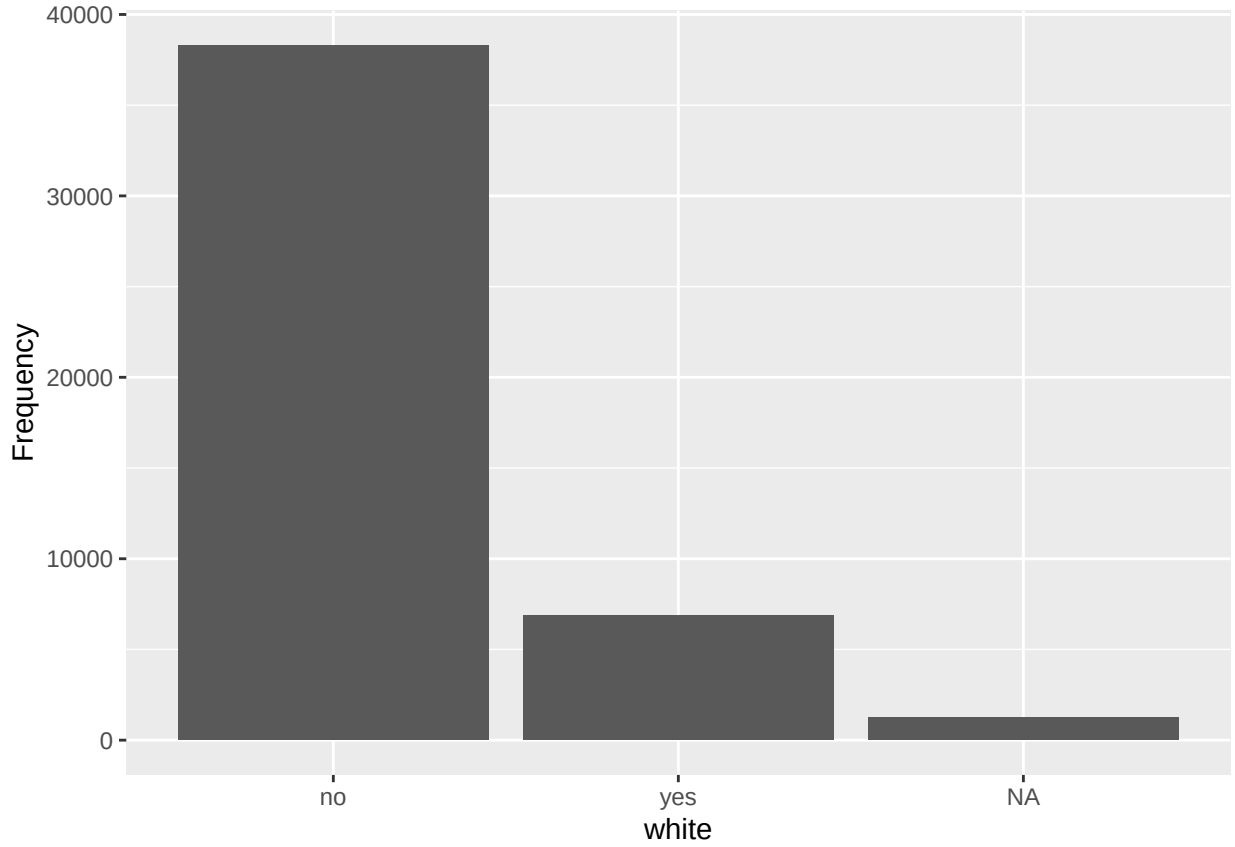

```
[1] ____________________________
[1] Variable: hispanic, type: character
[1] Values (3 unique): no, yes, NA
[1] Missing: 2.7%
[1] Most missing: F10 4.4%, Least missing: F11 1.5%
```



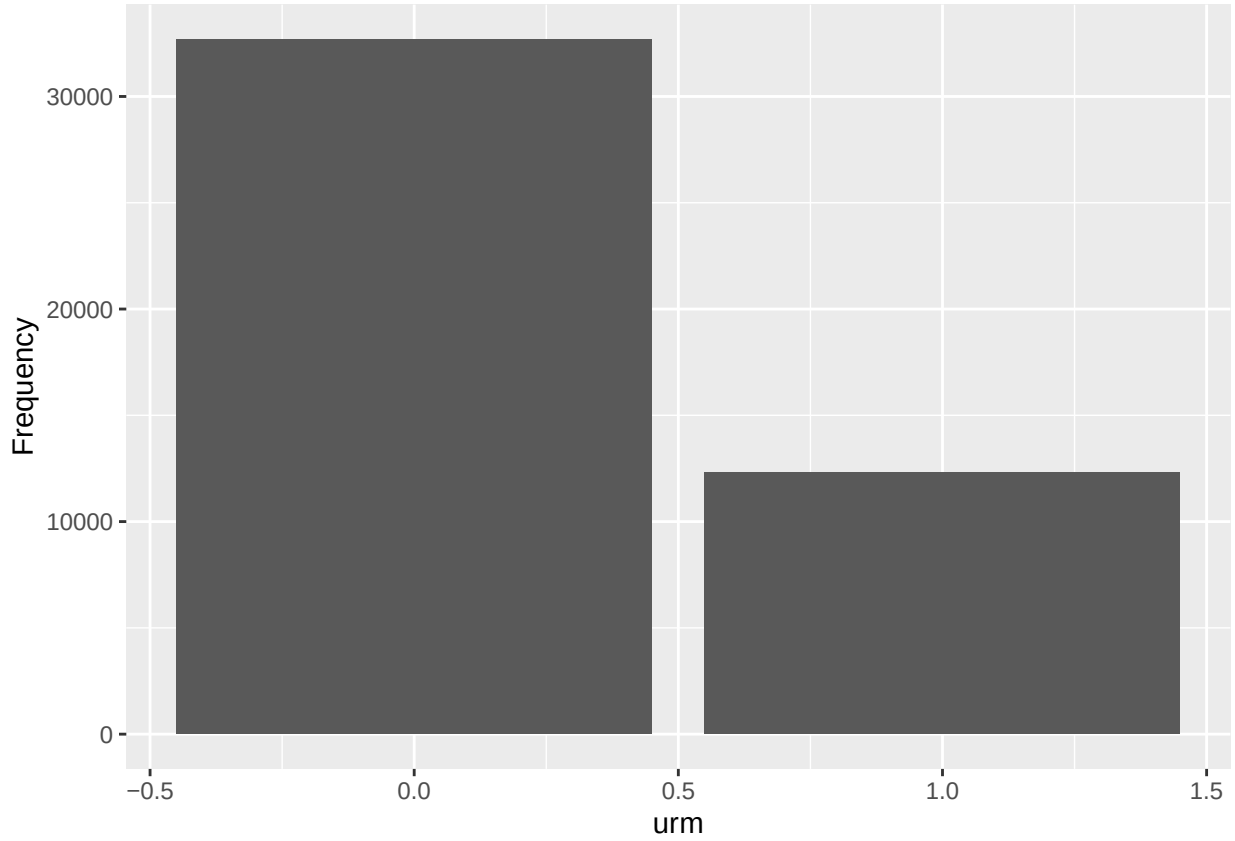

```
[1] ____________________________
[1] Variable: black, type: character
[1] Values (3 unique): no, NA, yes
[1] Missing: 2.7%
[1] Most missing: F10 4.4%, Least missing: F11 1.5%
```



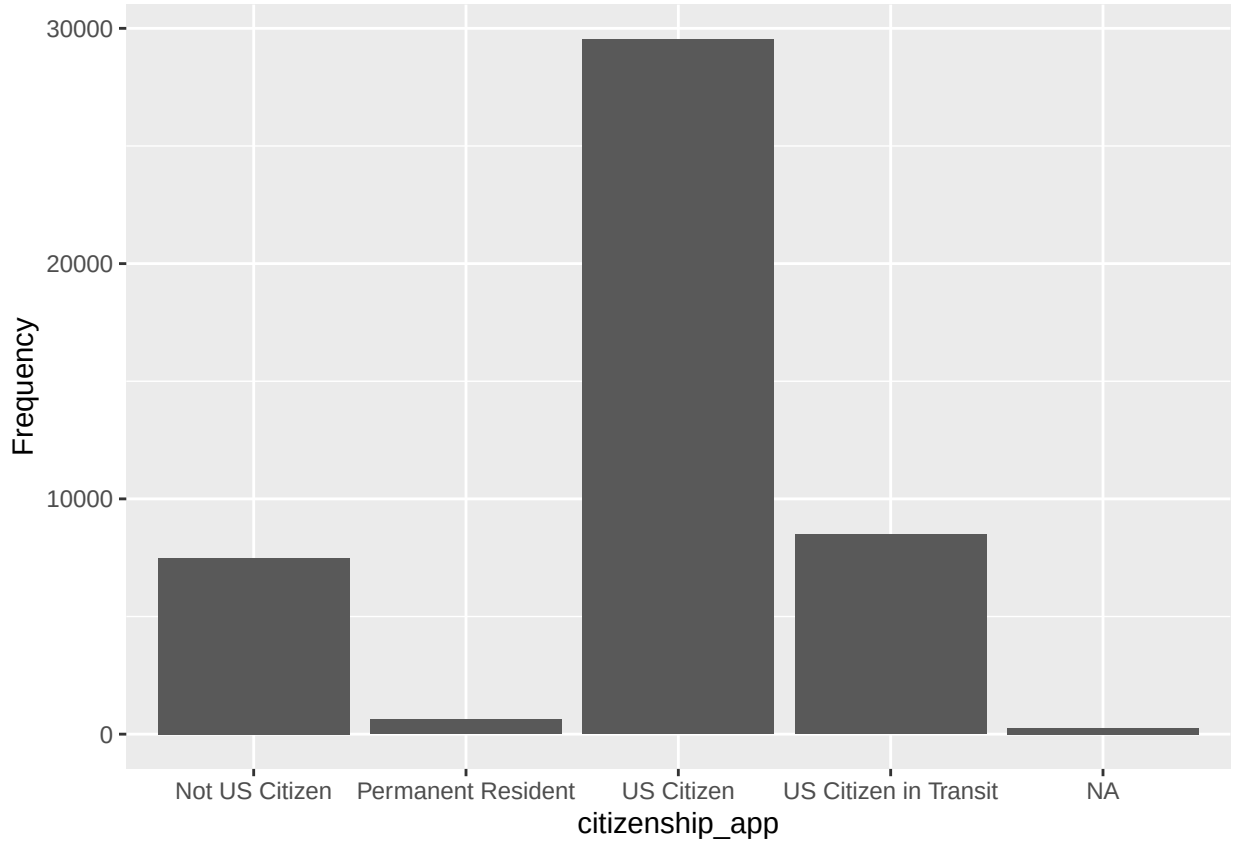

```
[1] ____________________________
[1] Variable: white, type: character
[1] Values (3 unique): no, yes, NA
[1] Missing: 2.7%
[1] Most missing: F10 4.4%, Least missing: F11 1.5%
```



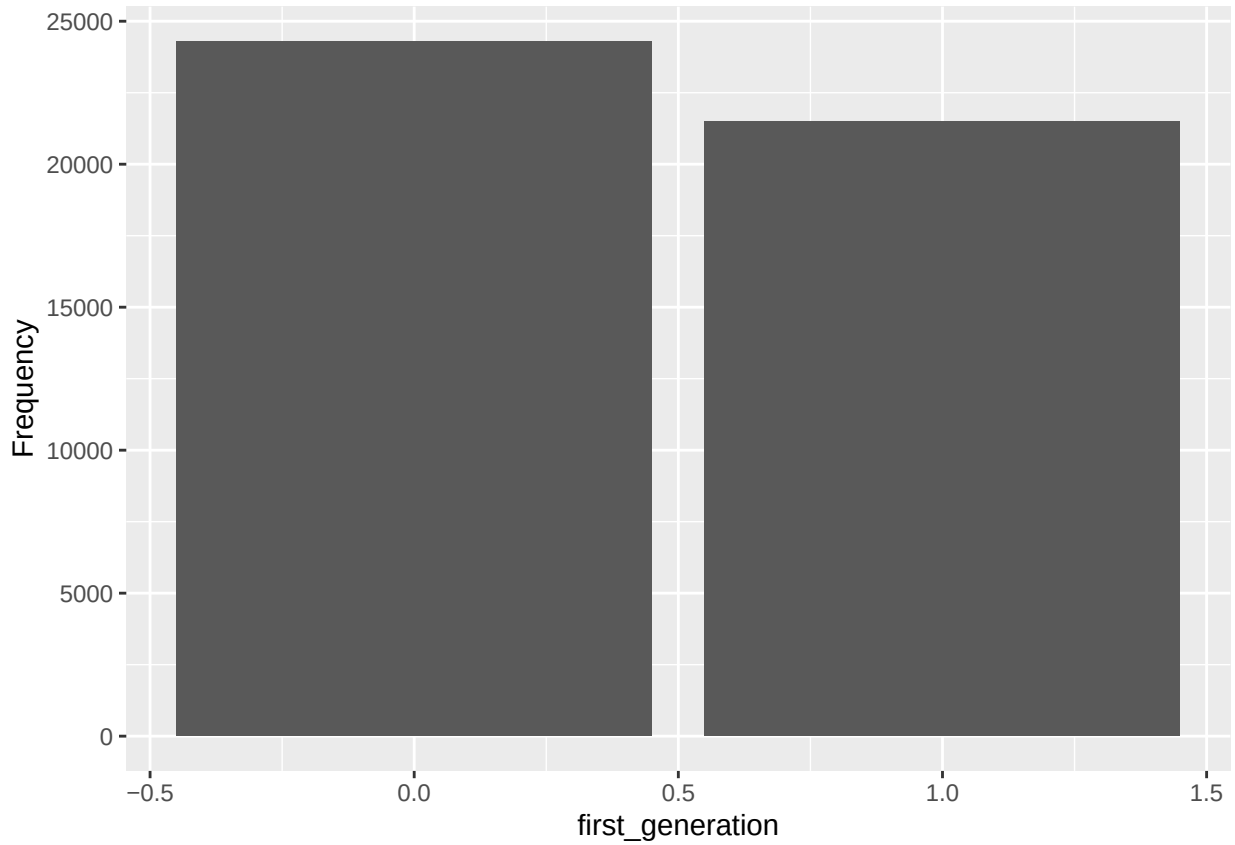

```
[1] ____________________________
[1] Variable: urm, type: numeric
[1] Values (3 unique): 0, 1, NA
[1] Missing: 3%
[1] Most missing: F08 5.6%, Least missing: F11 1.5%

Warning: Removed 1399 rows containing non-finite values (‘stat_count()‘).
```



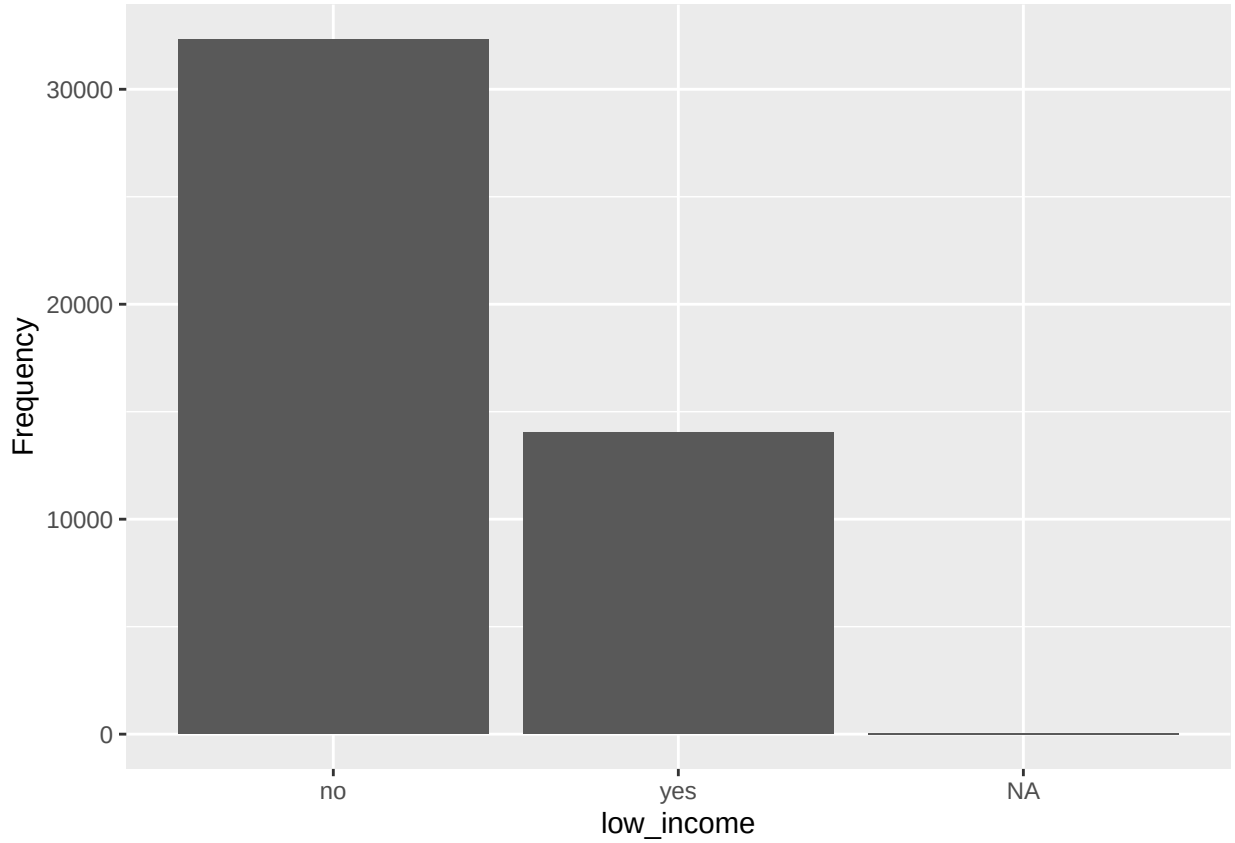

```
[1] ____________________________
[1] Variable: citizenship_app, type: character
[1] Values (5 unique): US Citizen in Transit, Not US Citizen, Permanent Resident, NA, US Citizen
[1] Missing: 0.6%
```



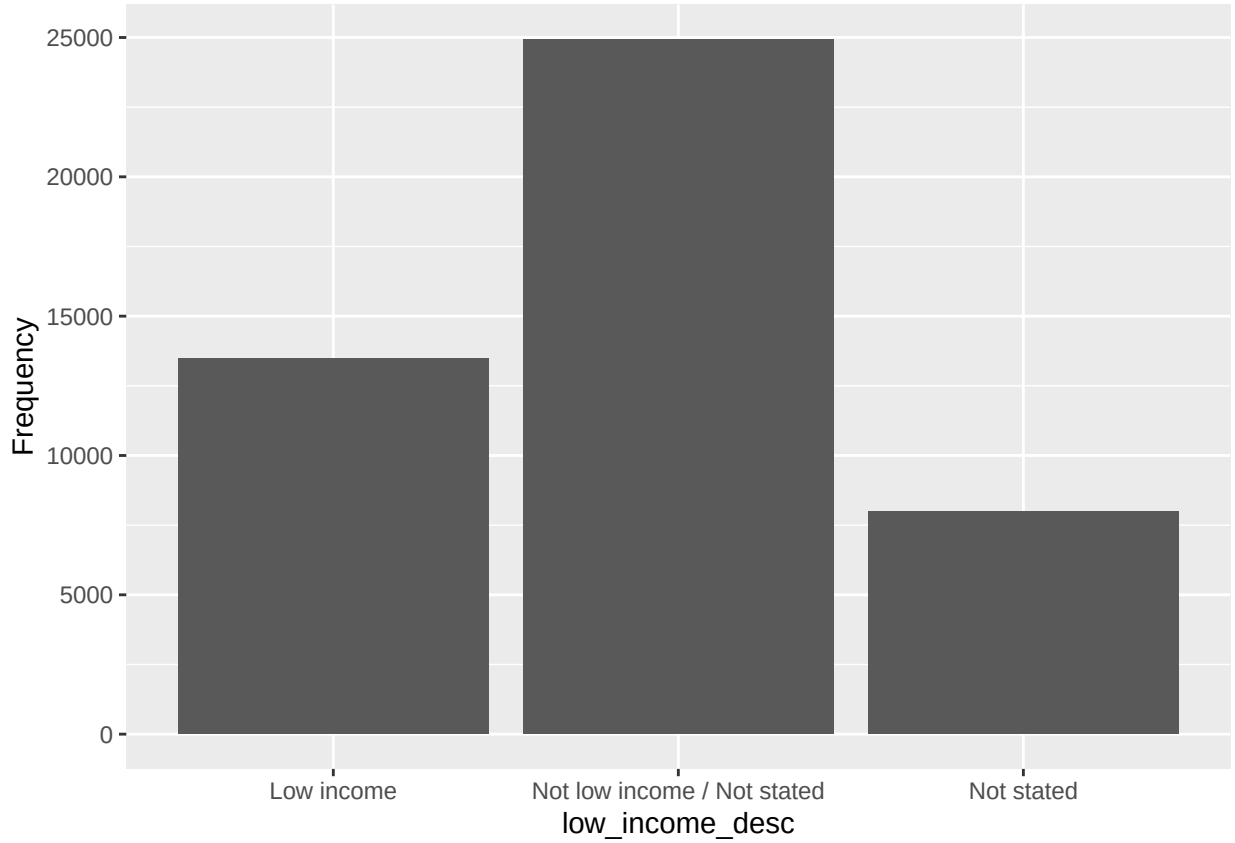

```
[1] ____________________________
[1] Variable: first_generation, type: numeric
[1] Values (3 unique): 0, 1, NA
[1] Missing: 1.3%
[1] Most missing: F08 3.2%, Least missing: F16 0.4%

Warning: Removed 622 rows containing non-finite values (‘stat_count()‘).
```



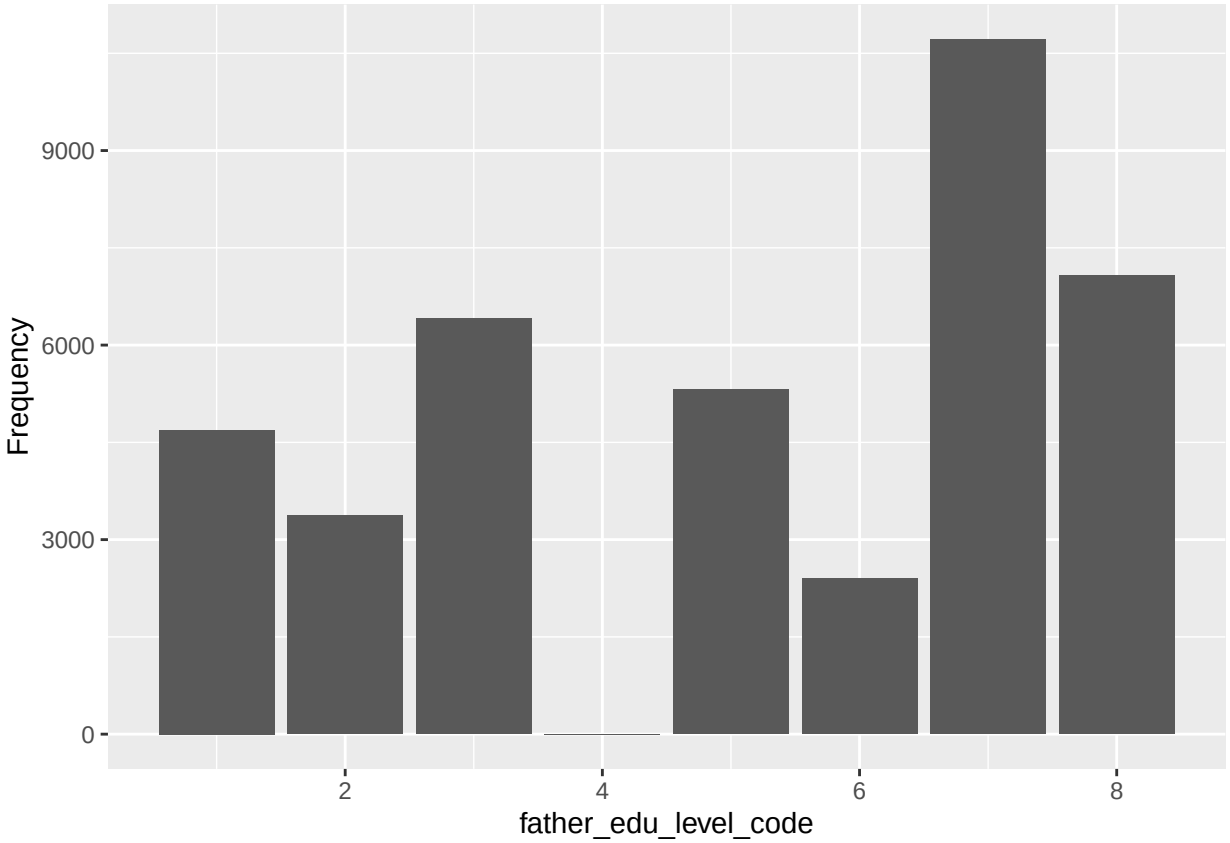

```
[1] ___________________________
[1] Variable: low_income, type: character
[1] Values (3 unique): no, yes, NA
[1] Missing: 0.1%
```



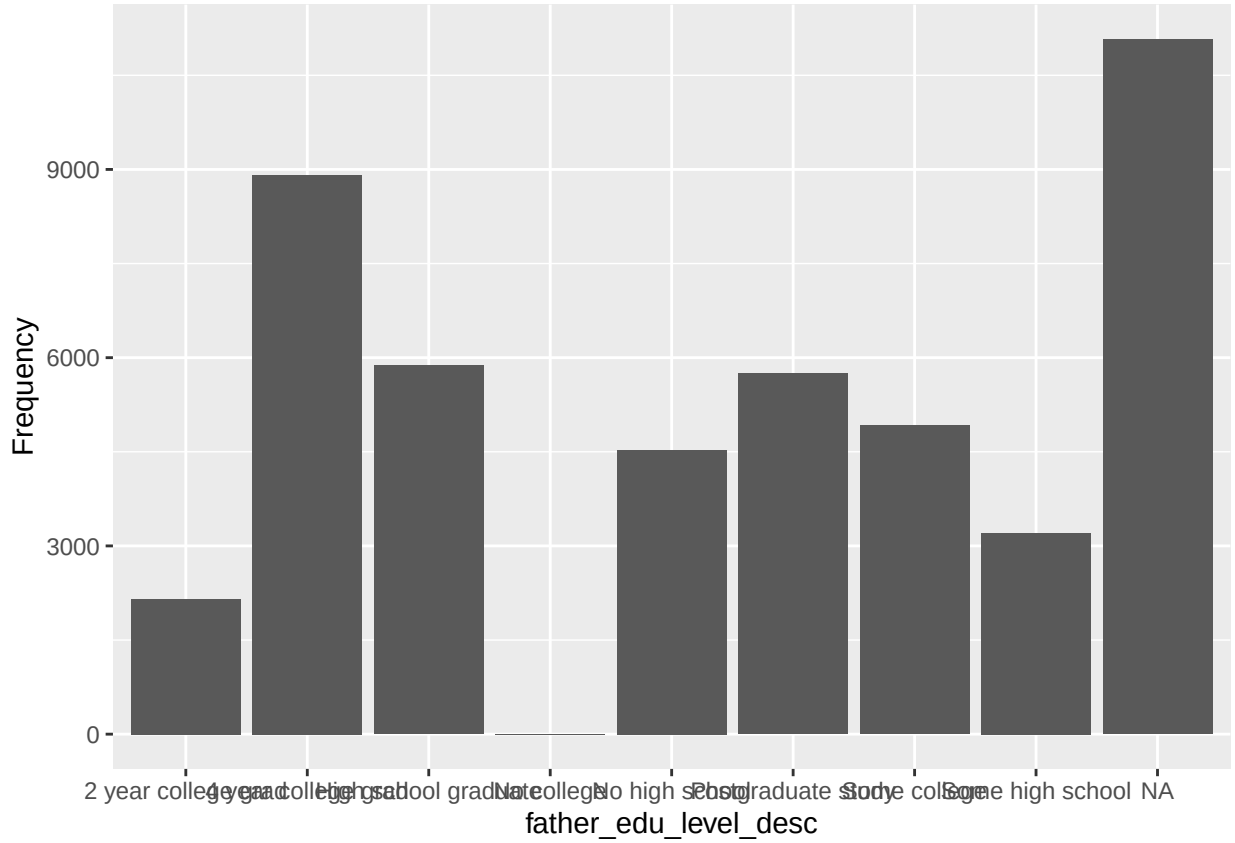

```
[1] ____________________________
[1] Variable: low_income_desc, type: character
[1] Values (3 unique): Not low income / Not stated, Low income, Not stated
[1] Missing: 0%
```



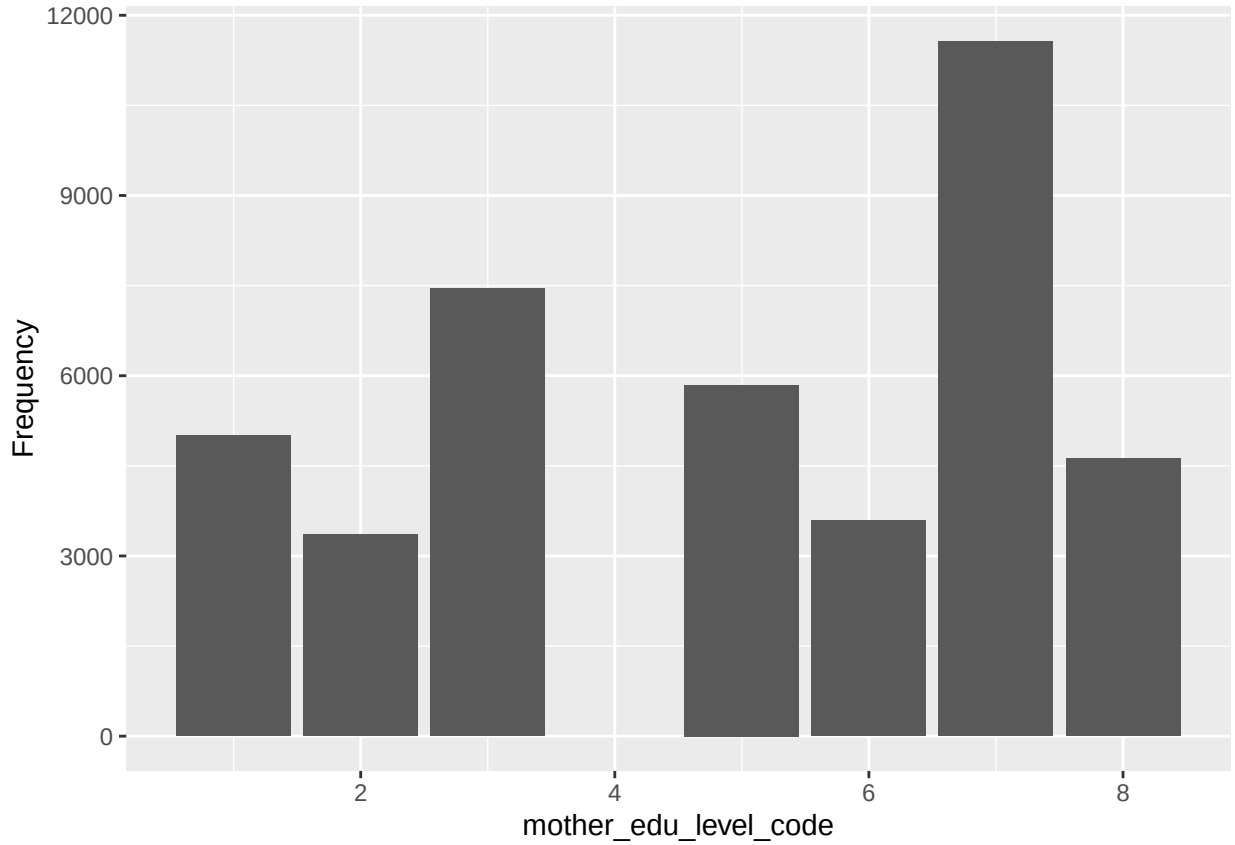

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: father_edu_level_code, type: numeric
[1] Values (9 unique): 7, 8, 5, NA, 3, ...
[1] Missing: 13.9%
[1] Most missing: F08 43.7%, Least missing: F11 6.8%

Warning: Removed 6441 rows containing non-finite values (‘stat_count()‘).
```



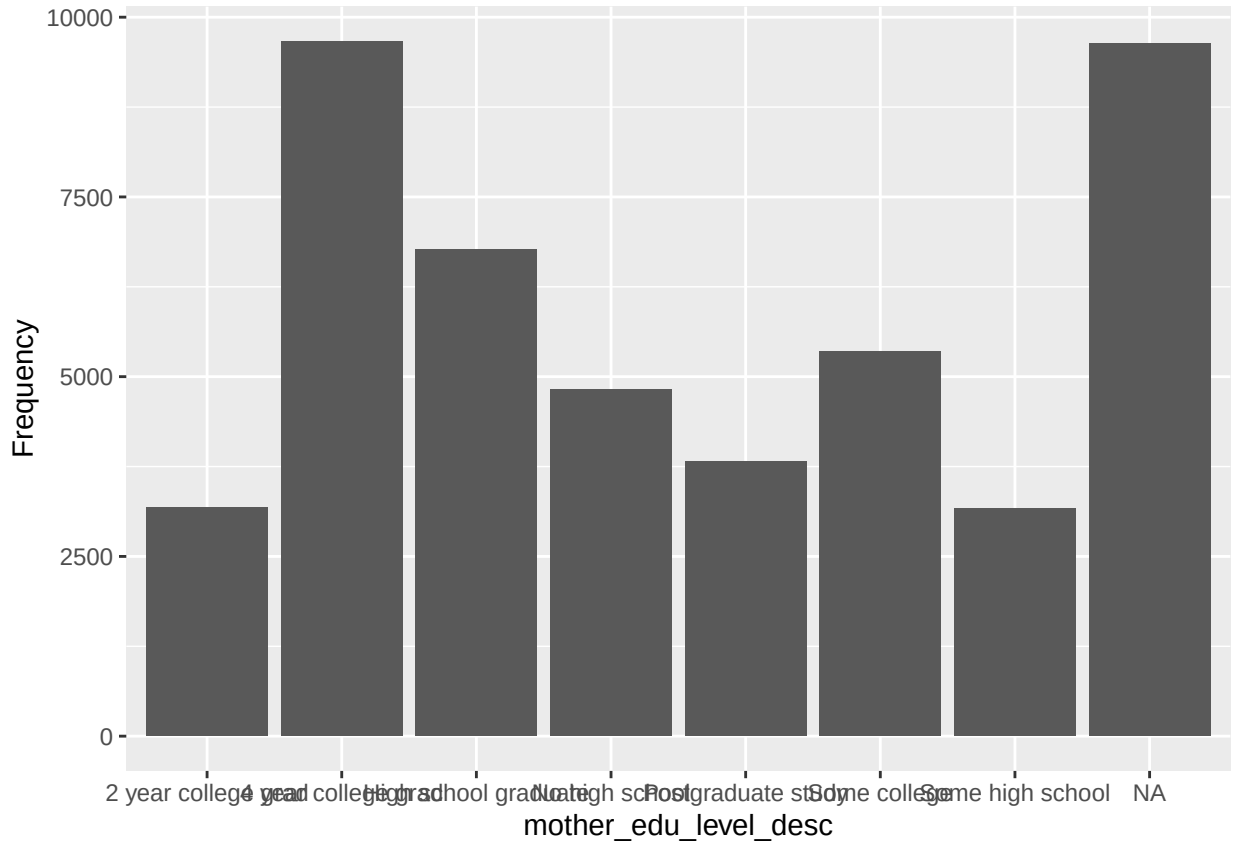

```
[1] ____________________________
[1] Variable: father_edu_level_desc, type: character
[1] Values (9 unique): 4 year college grad, Postgraduate study, Some college, NA, High school graduate,
[1] Missing: 23.9%
[1] Most missing: F08 43.7%, Least missing: F15 15.4%
```



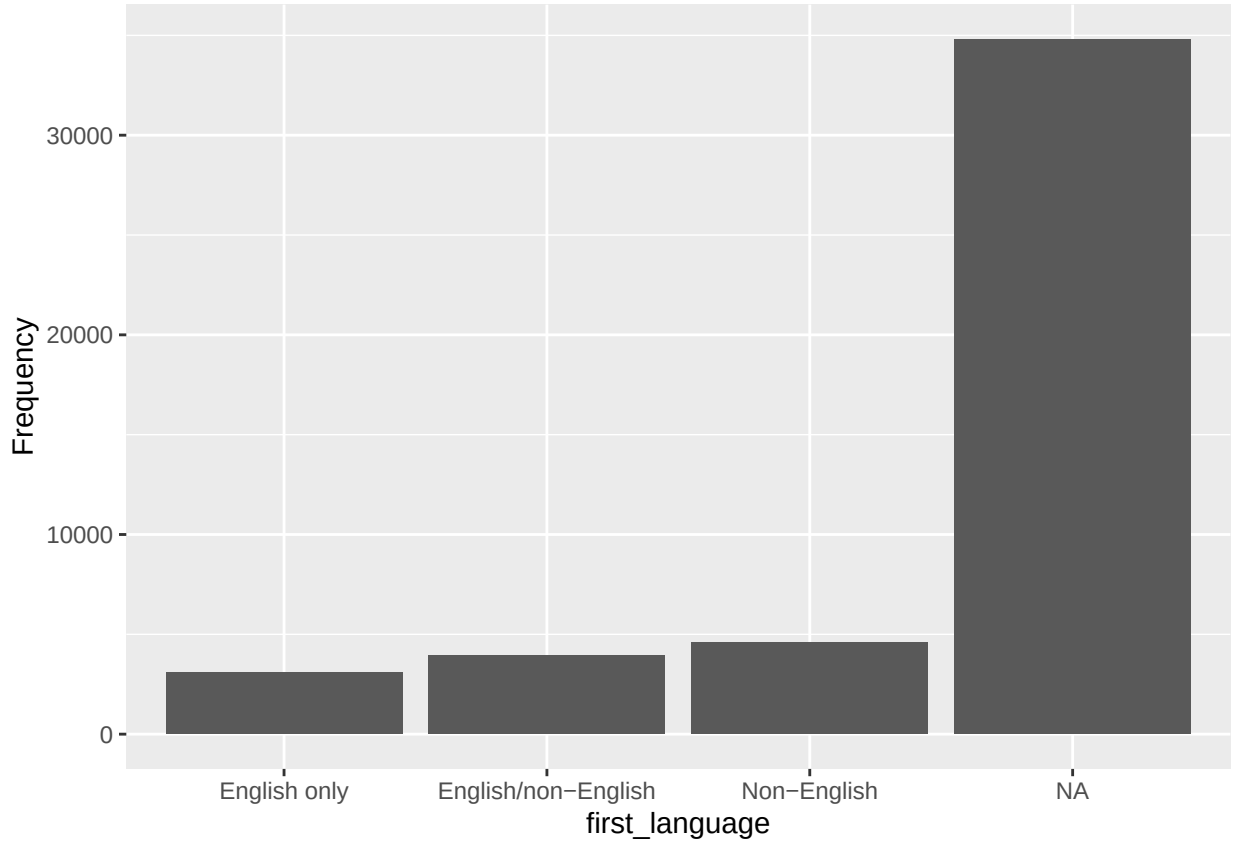

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: mother_edu_level_code, type: numeric
[1] Values (8 unique): 7, 8, 3, 1, 2, ...
[1] Missing: 10.7%
[1] Most missing: F08 41.7%, Least missing: F11 3.4%

Warning: Removed 4969 rows containing non-finite values (‘stat_count()‘).
```



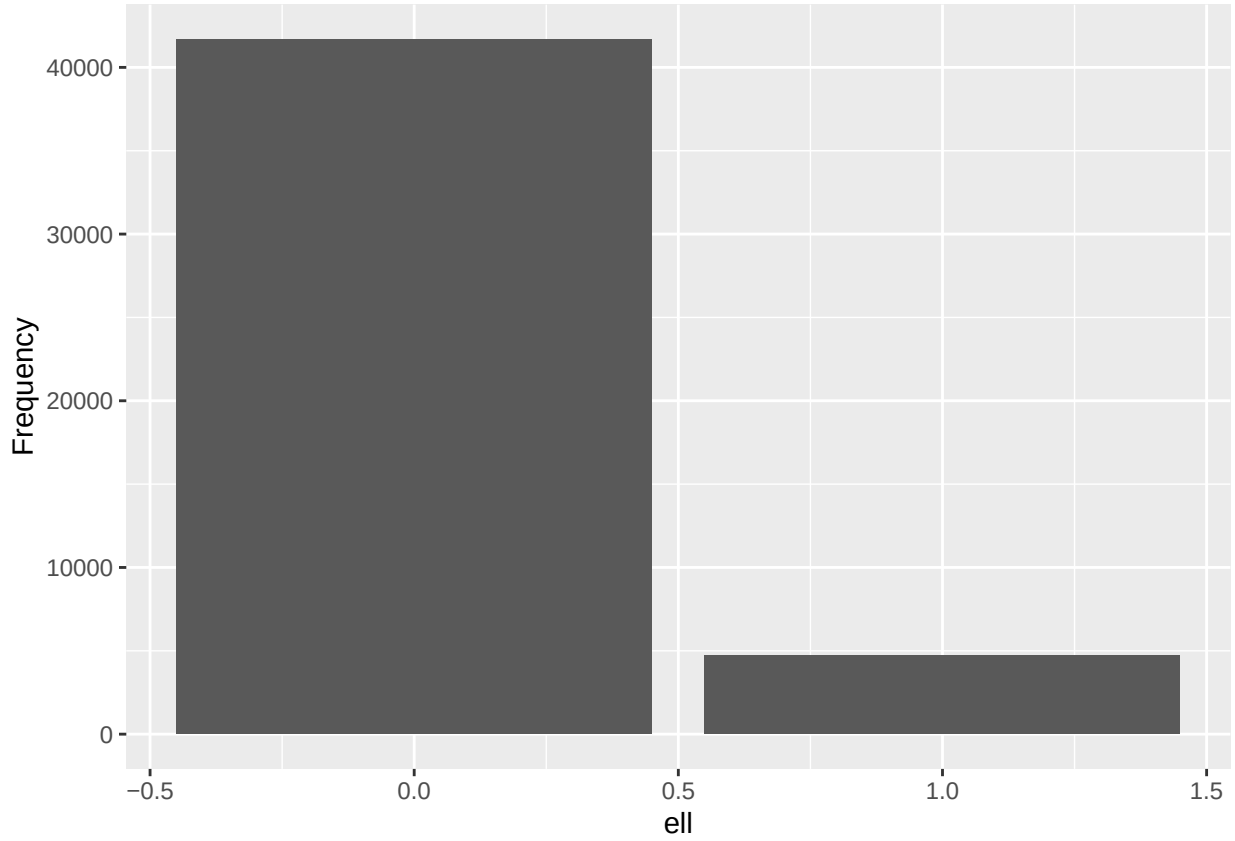

```
[1] ____________________________
[1] Variable: mother_edu_level_desc, type: character
[1] Values (8 unique): 4 year college grad, Postgraduate study, High school graduate, No high school, S
[1] Missing: 20.8%
[1] Most missing: F08 41.7%, Least missing: F15 11.4%
```



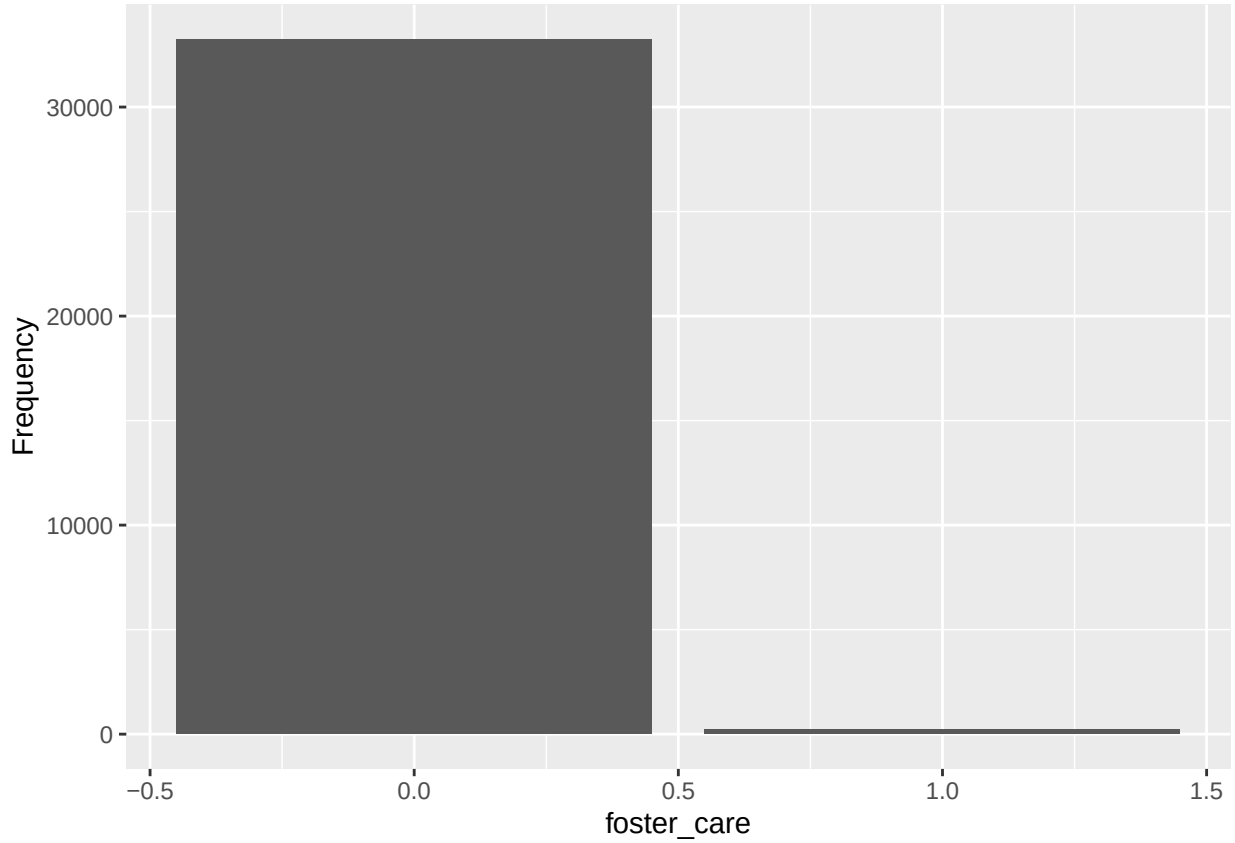

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: first_language, type: character
[1] Values (4 unique): English only, English/non-English, Non-English, NA
[1] Missing: 75%
  Group.1 first_language
1     F08     0.99608866
2     F09     0.99554676
3     F10     0.99591837
4     F11     0.98886501
5     F12     0.98582398
6     F13     0.97056659
7     F14     0.84961154
8     F15     0.21566580
9     F16     0.09499163
```



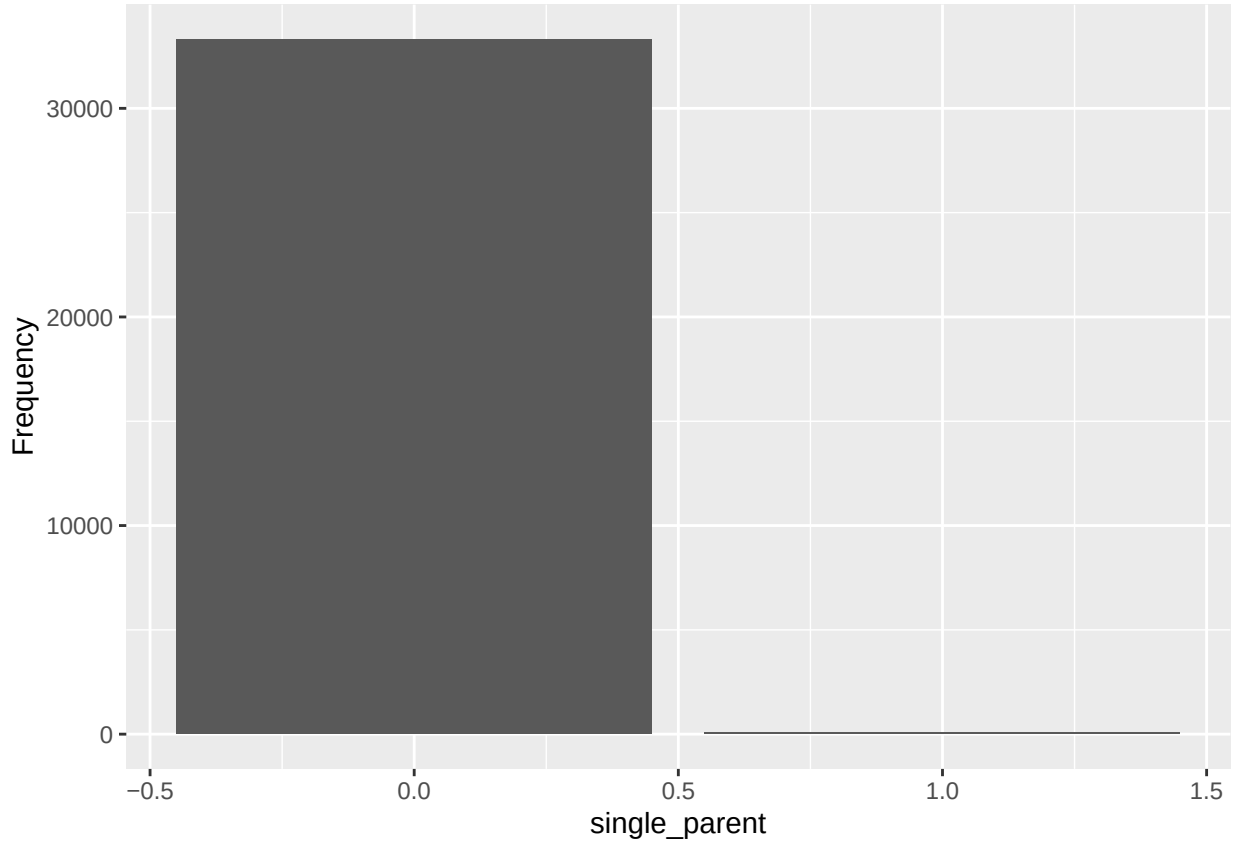

```
[1] ____________________________
[1] Variable: ell, type: numeric
[1] Values (2 unique): 0, 1
[1] Missing: 0%
```



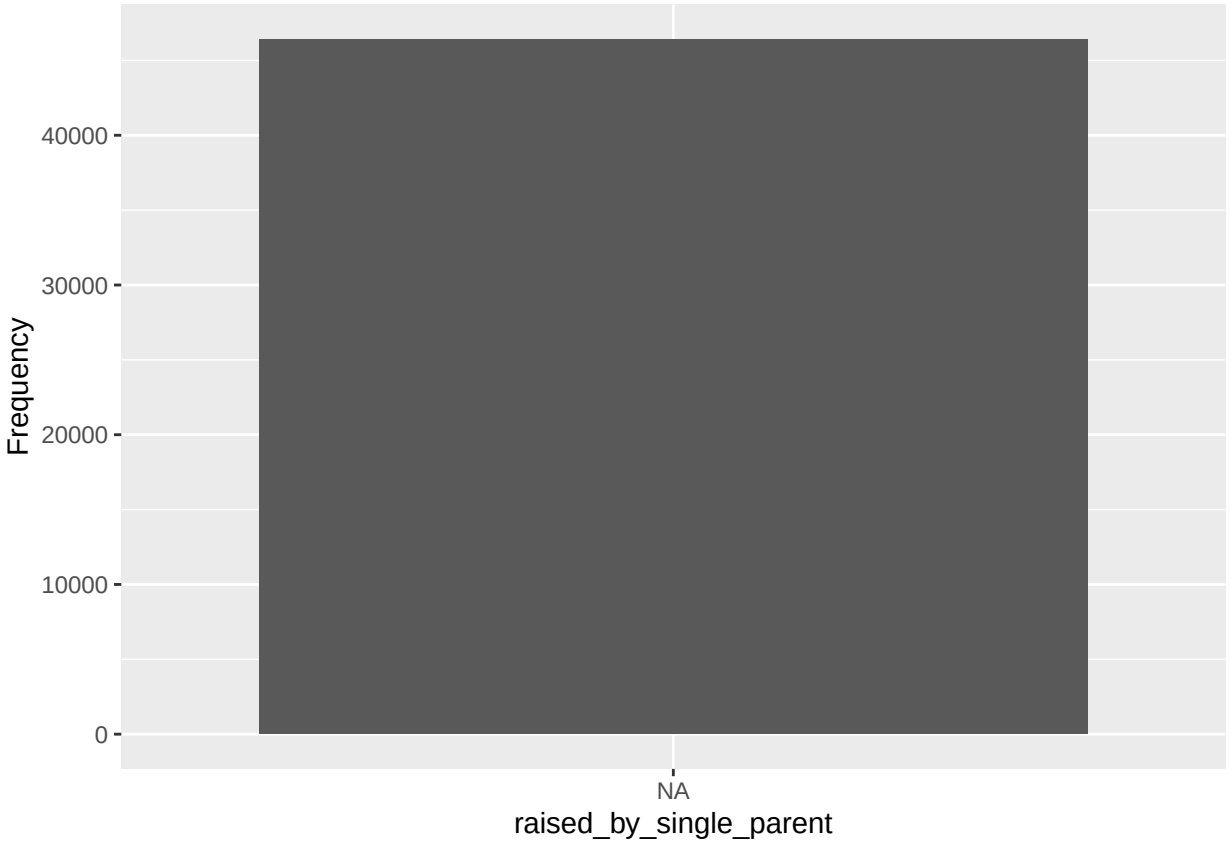

```
[1] ____________________________
[1] Variable: foster_care, type: integer
[1] Values (3 unique): 0, 1, NA
[1] Missing: 27.9%
  Group.1 foster_care
1     F08   0.9913081
2     F09   0.9925779
3     F10   0.9927438
4     F11   0.0000000
5     F12   0.0000000
6     F13   0.0000000
7     F14   0.0000000
8     F15   0.0000000
9     F16   0.0000000

Warning: Removed 12952 rows containing non-finite values (‘stat_count()‘).
```



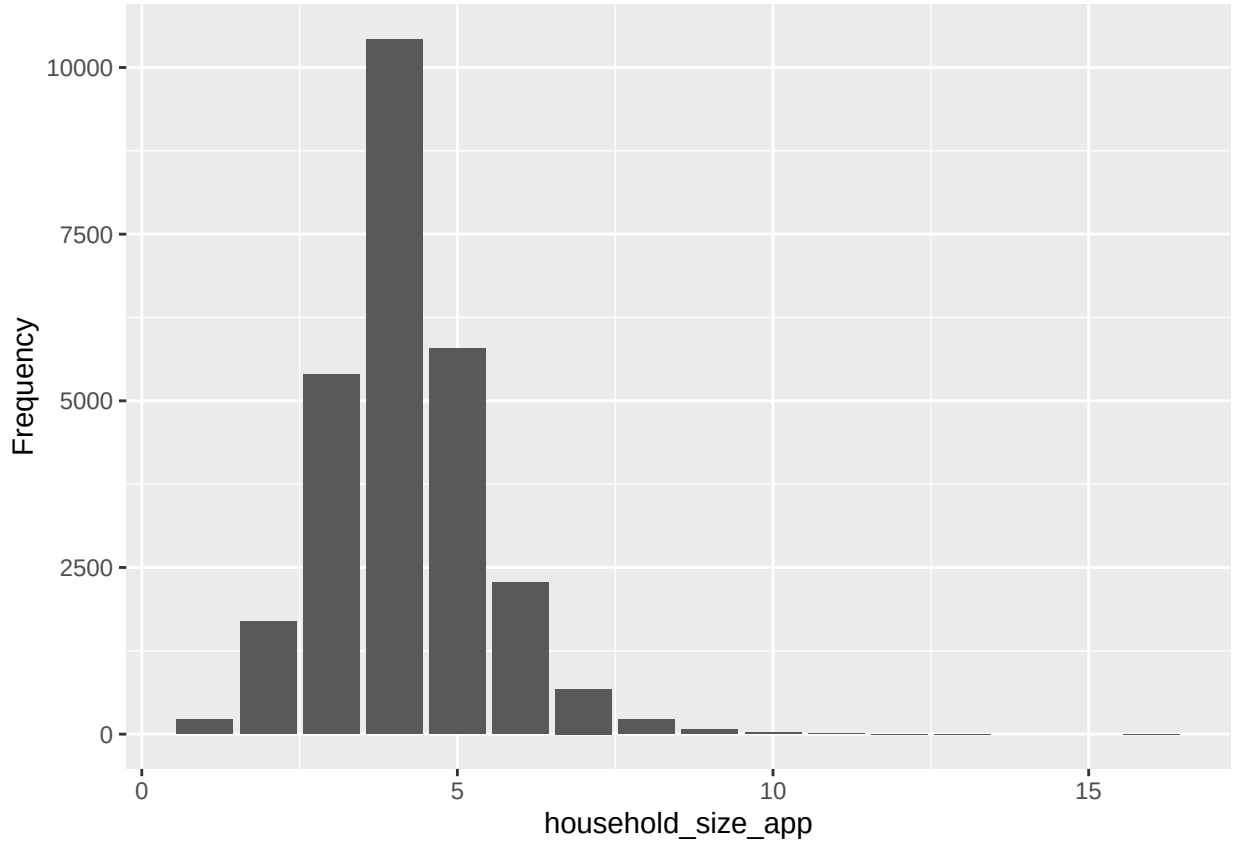

```
[1] ____________________________
[1] Variable: single_parent, type: numeric
[1] Values (3 unique): NA, 0, 1
[1] Missing: 28.1%
  Group.1 single_parent
1     F08  0.9952194698
2     F09  0.9970311727
3     F10  0.9968253968
4     F11  0.0001953507
5     F12  0.0003937783
6     F13  0.0003679176
7     F14  0.0001849797
8     F15  0.0006962576
9     F16  0.0012178414

Warning: Removed 13024 rows containing non-finite values (‘stat_count()‘).
```




```
[1] ____________________________
[1] Variable: raised_by_single_parent, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
```

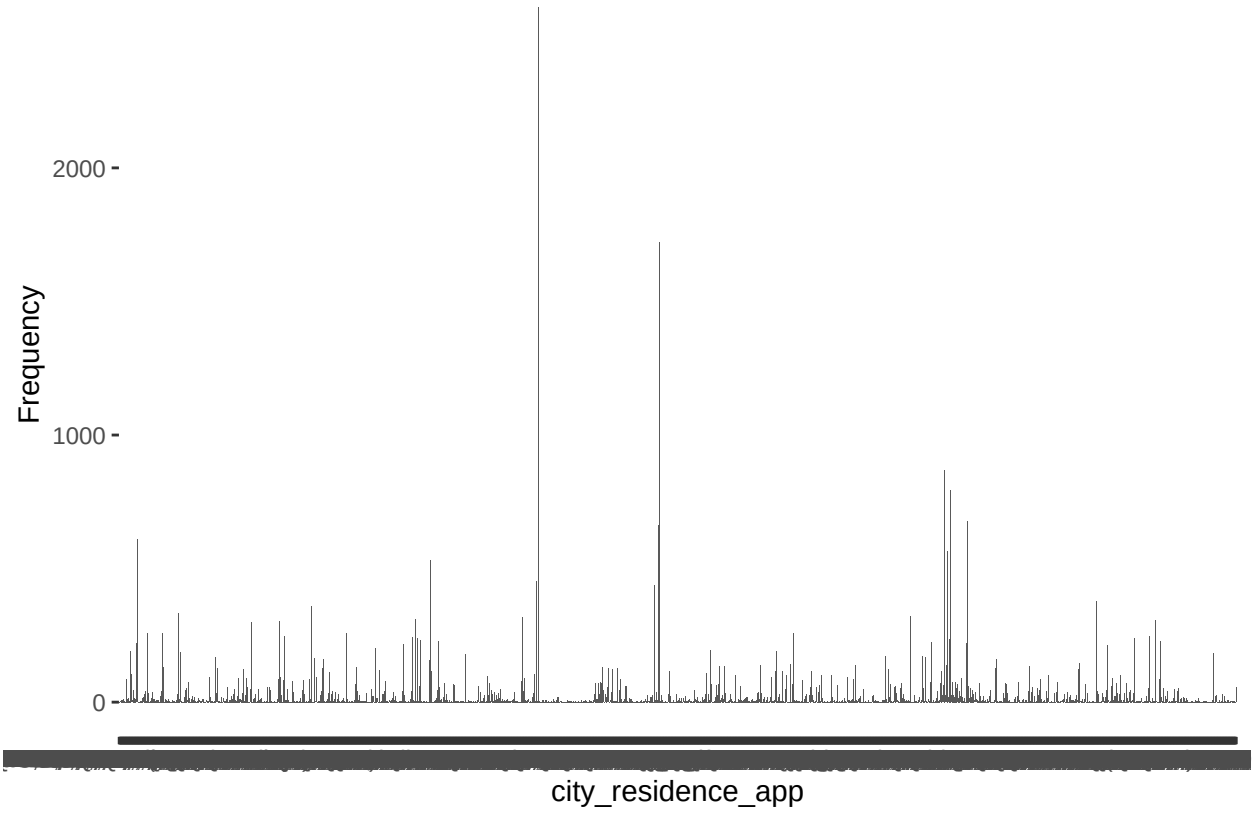



```
[1] ____________________________
[1] Variable: household_size_app, type: numeric
[1] Values (15 unique): NA, 4, 6, 7, 5, ...
[1] Missing: 42.3%
  Group.1 household_size_app
1     F08          0.3828770
2     F09          0.2936665
3     F10          0.2061224
4     F11          0.1103731
5     F12          0.1614491
6     F13          0.2049301
7     F14          0.3200148
8     F15          0.8628372
9     F16          1.0000000

Warning: Removed 19613 rows containing non-finite values (‘stat_count()‘).
```

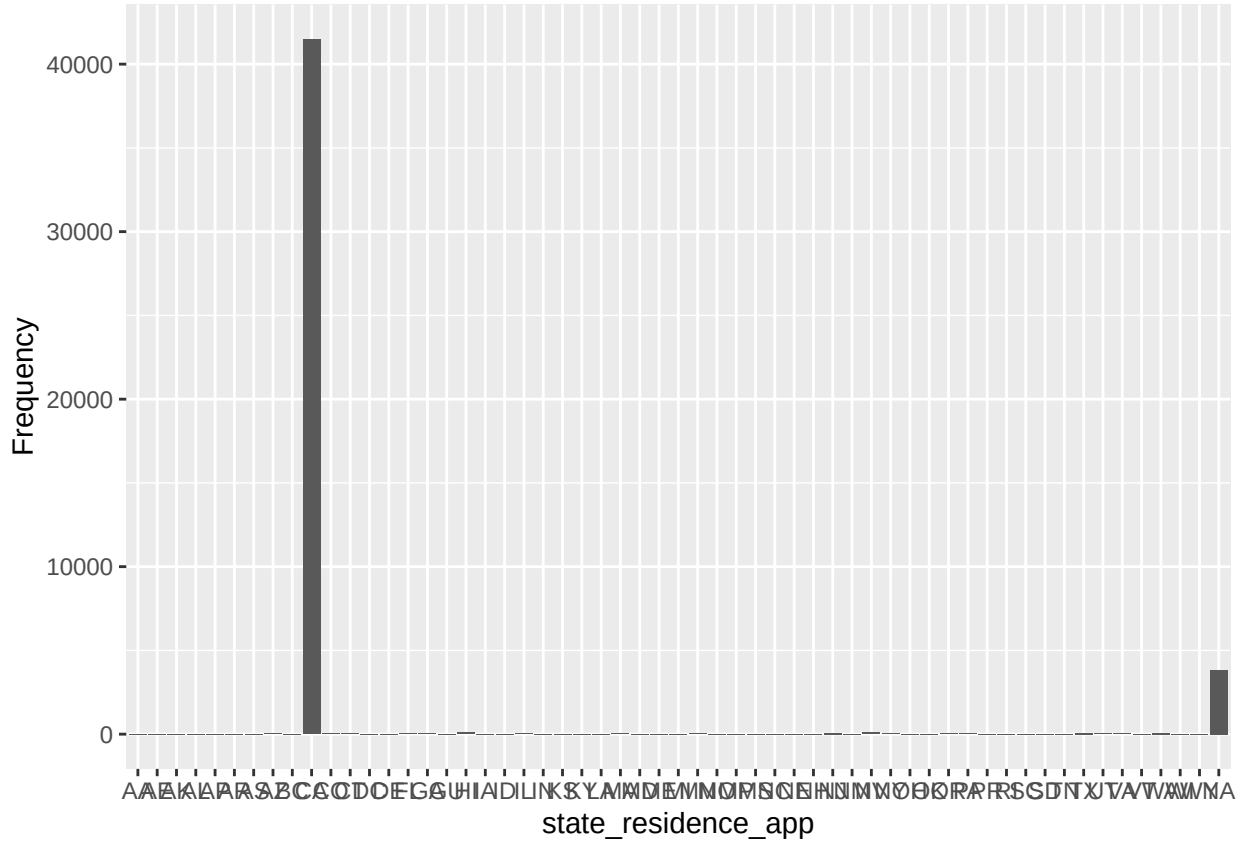



```
[1] ____________________________
[1] Variable: city_residence_app, type: character
[1] Values (2969 unique): NA, POMONA, ELK GROVE, WESTMINSTER, LOS ANGELES, ...
[1] Missing: 0.1%

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>
```

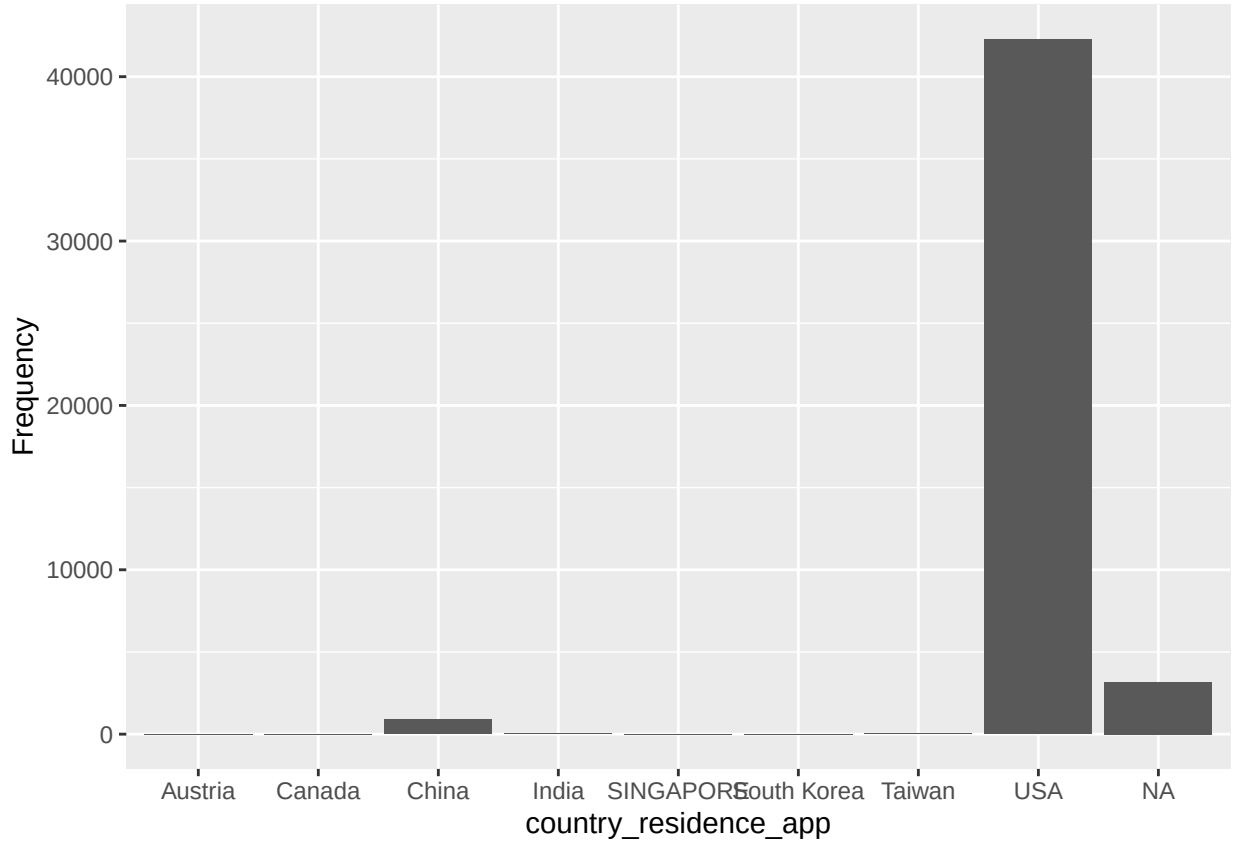
```
Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>

Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
Konvertierungsfehler für 'LA HABRÁ HEIGHTS, CA' in 'mbcsToSbcs': Punkt ersetzt
<c1>
```



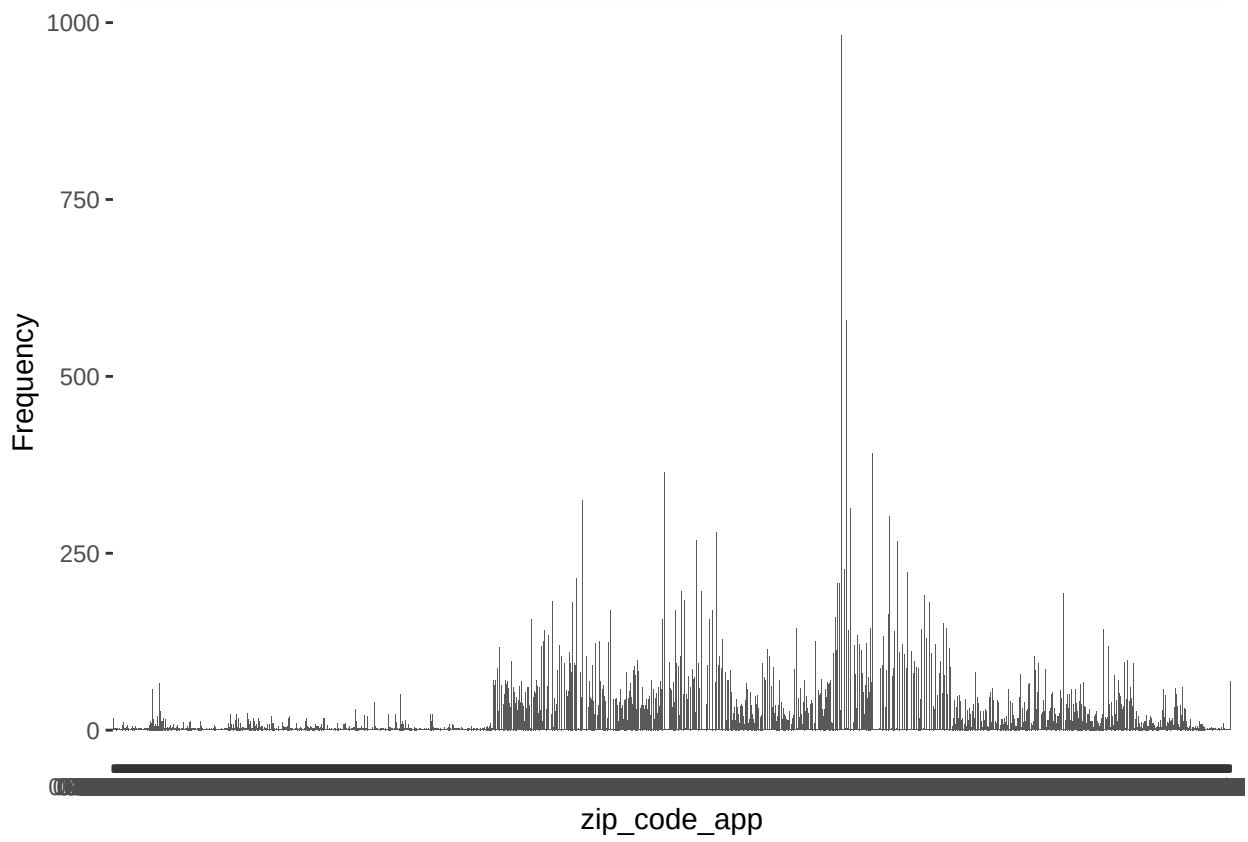

```
[1] ____________________________
[1] Variable: state_residence_app, type: character
[1] Values (57 unique): NA, CA, CO, AZ, DE, ...
[1] Missing: 8.3%
[1] Most missing: F16 17.2%, Least missing: F09 1%
```



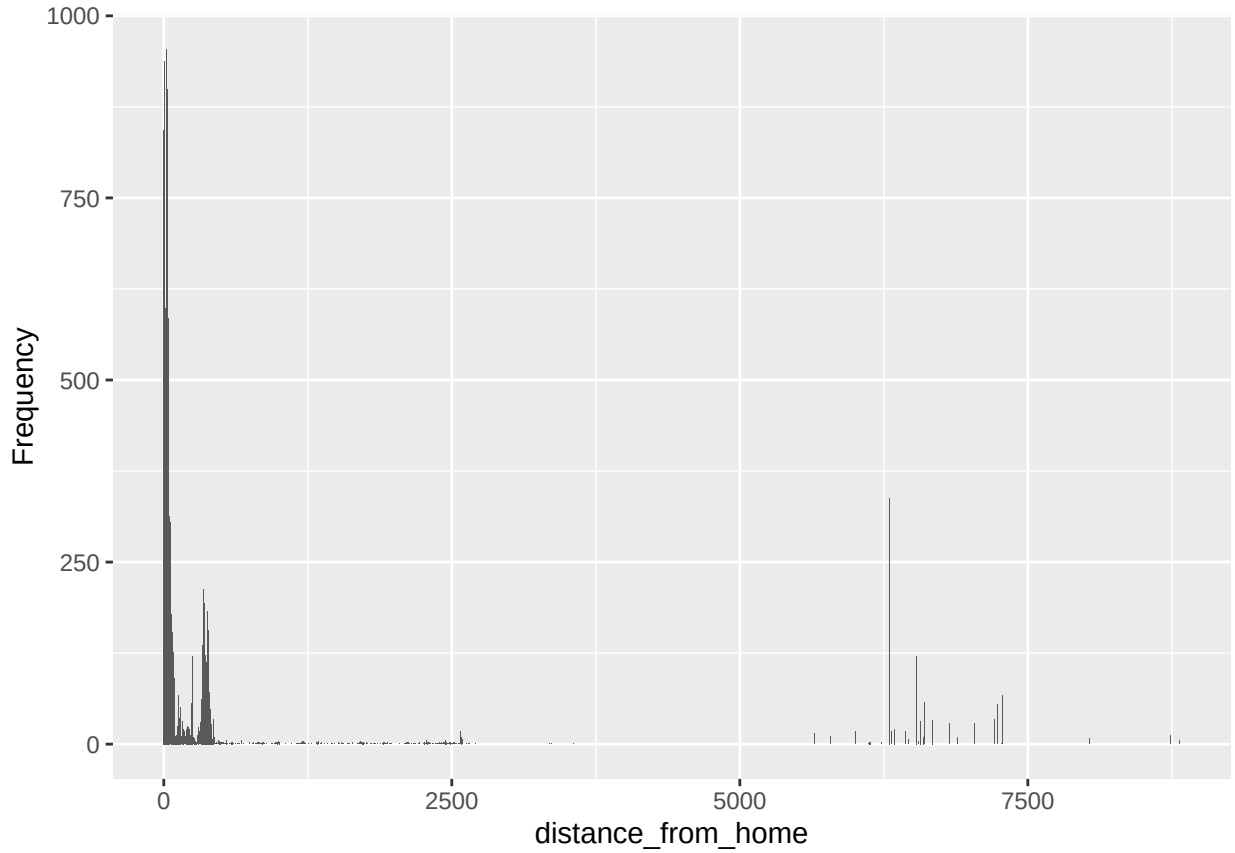

```
[1] ____________________________
[1] Variable: country_residence_app, type: character
[1] Values (9 unique): NA, USA, China, South Korea, Taiwan, ...
[1] Missing: 6.8%
[1] Most missing: F15 10.8%, Least missing: F09 1.5%
```



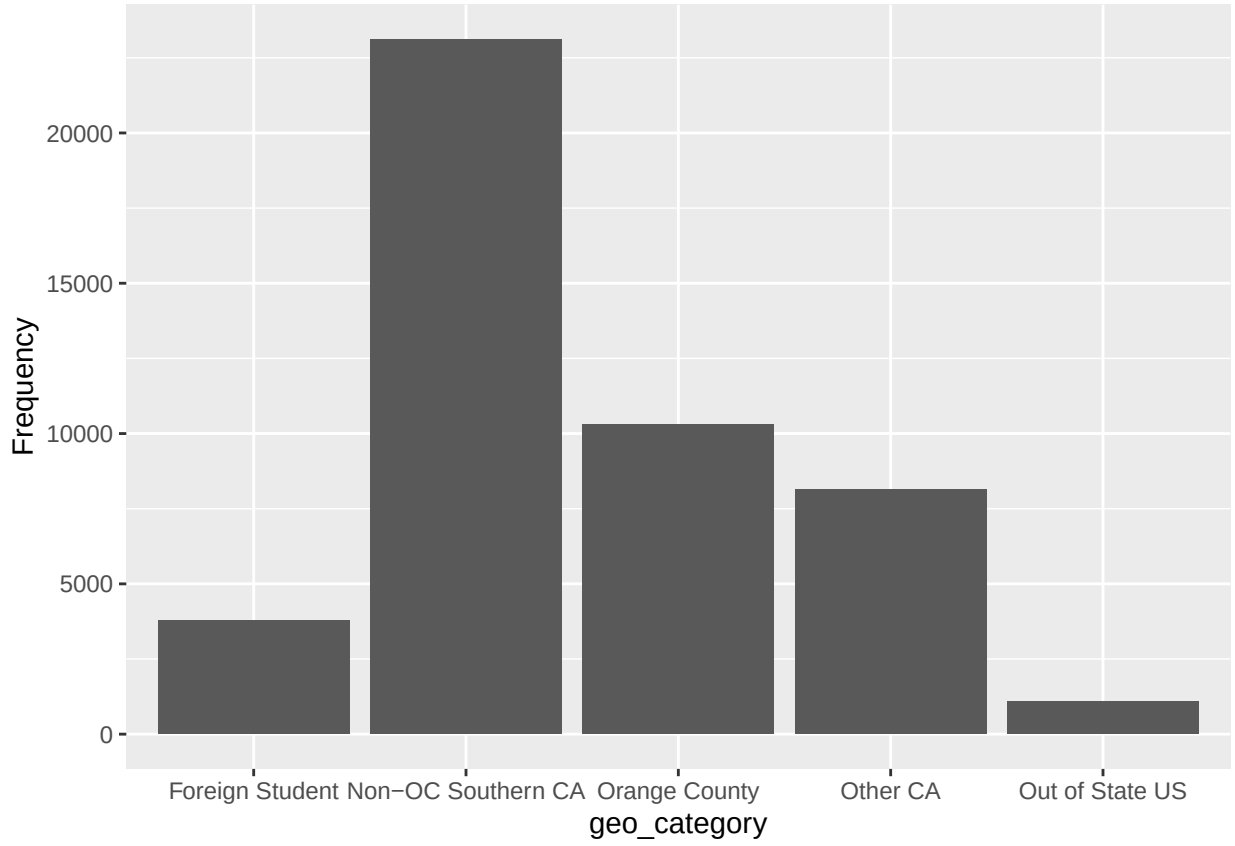

```
[1] ____________________________
[1] Variable: zip_code_app, type: character
[1] Values (6525 unique): NA, 91767, 95757, 92683, 90022, ...
[1] Missing: 0.1%
```



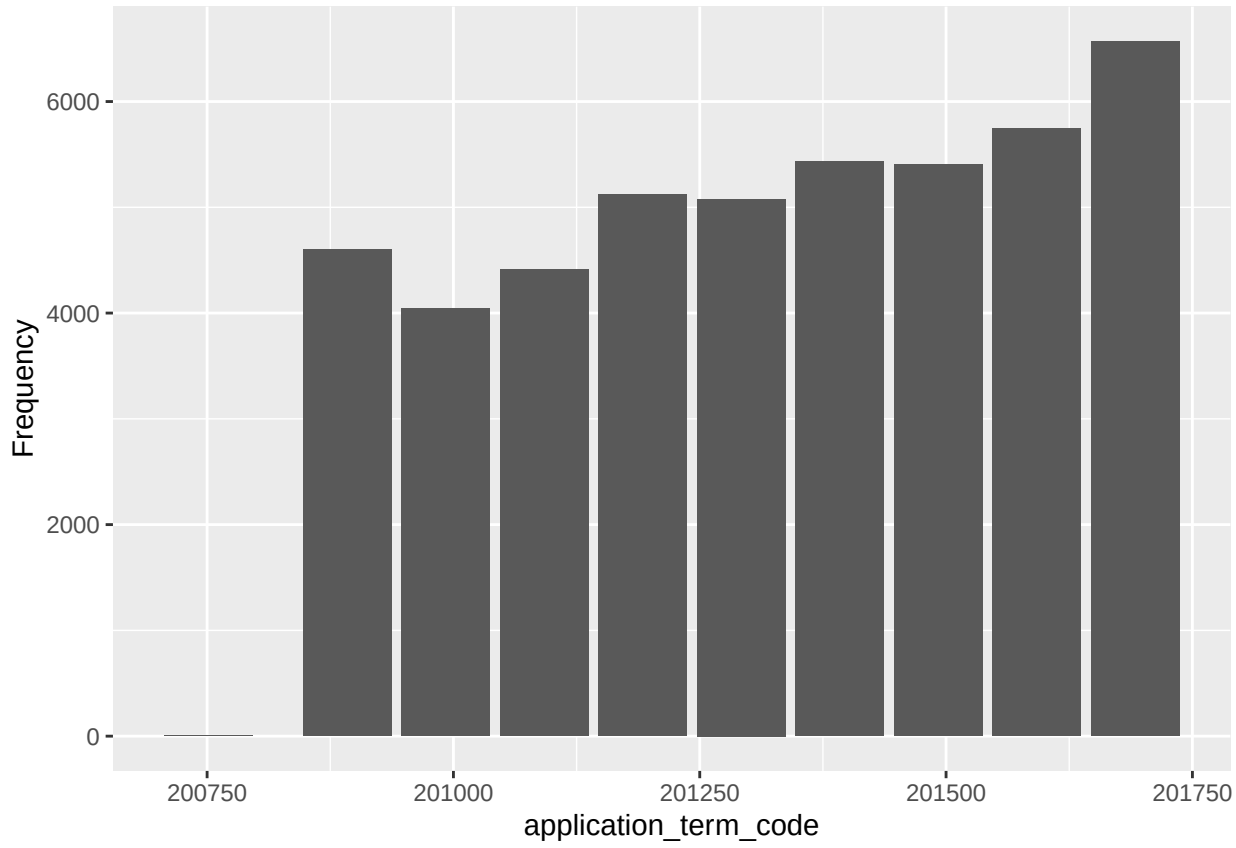

```
[1] ____________________________
[1] Variable: distance_from_home, type: numeric
[1] Values (721 unique): NA, 6302, 6670, 6567, 6604, ...
[1] Missing: 27.9%
  Group.1 distance_from_home
1     F08         0.01607997
2     F09         0.01484414
3     F10         0.02380952
4     F11         0.04766556
5     F12         0.09155346
6     F13         0.13134658
7     F14         0.23862375
8     F15         0.77963446
9     F16         0.83711372

Warning: Removed 12930 rows containing non-finite values (‘stat_count()‘).
```



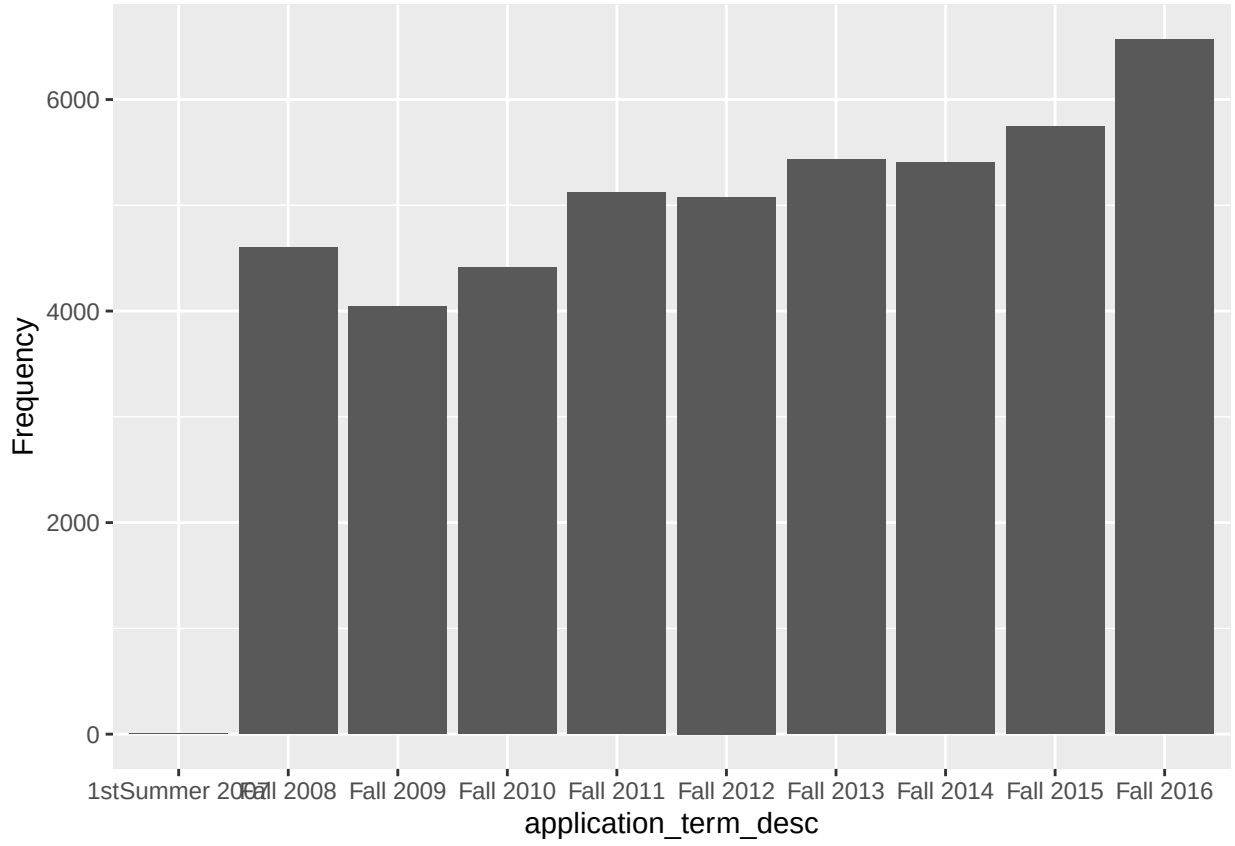

```
[1] ____________________________
[1] Variable: geo_category, type: character
[1] Values (5 unique): Other CA, Non-OC Southern CA, Orange County, Out of State US, Foreign Student
[1] Missing: 0%
```



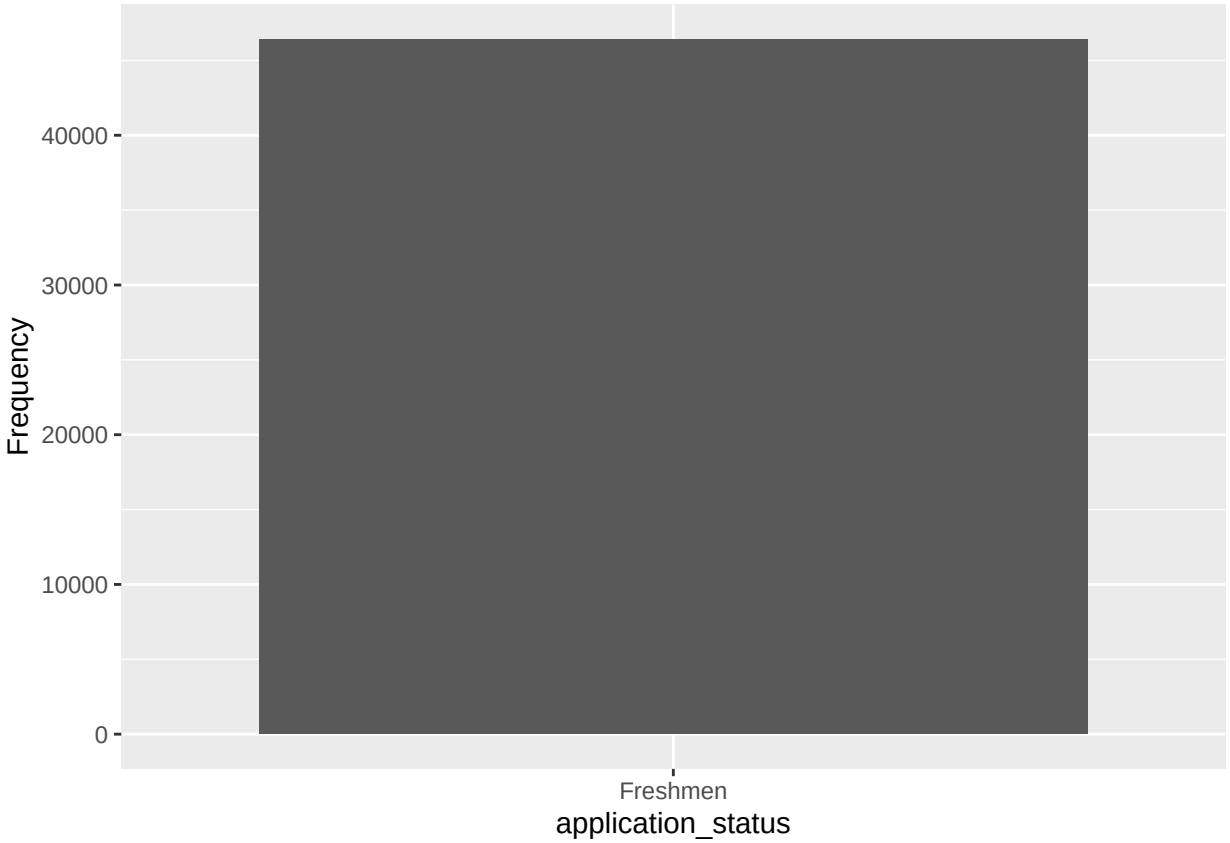

```
[1] ____________________________
[1] Variable: application_term_code, type: numeric
[1] Values (10 unique): 200892, 200992, 201092, 201592, 201692, ...
[1] Missing: 0%
```



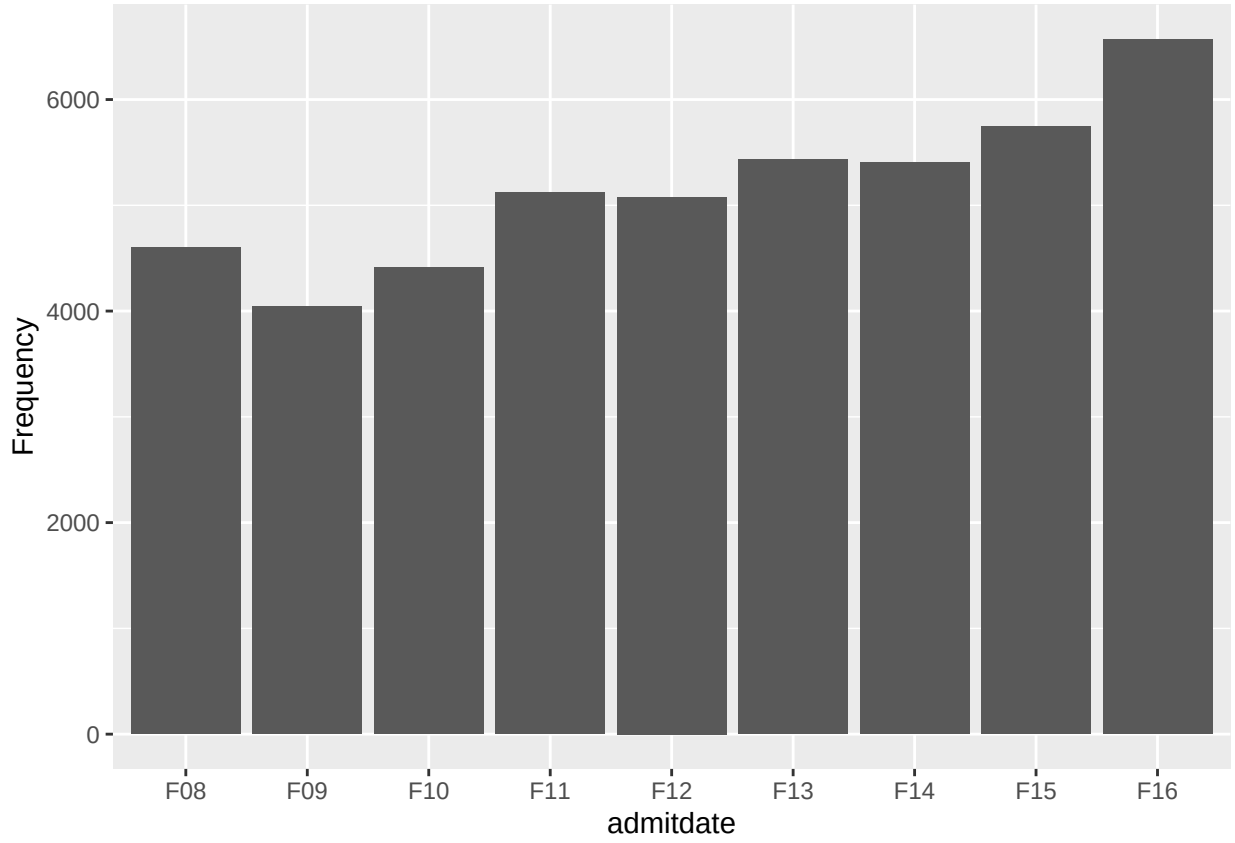

```
[1] ____________________________
[1] Variable: application_term_desc, type: character
[1] Values (10 unique): Fall 2008, Fall 2009, Fall 2010, Fall 2015, Fall 2016, ...
[1] Missing: 0%
```



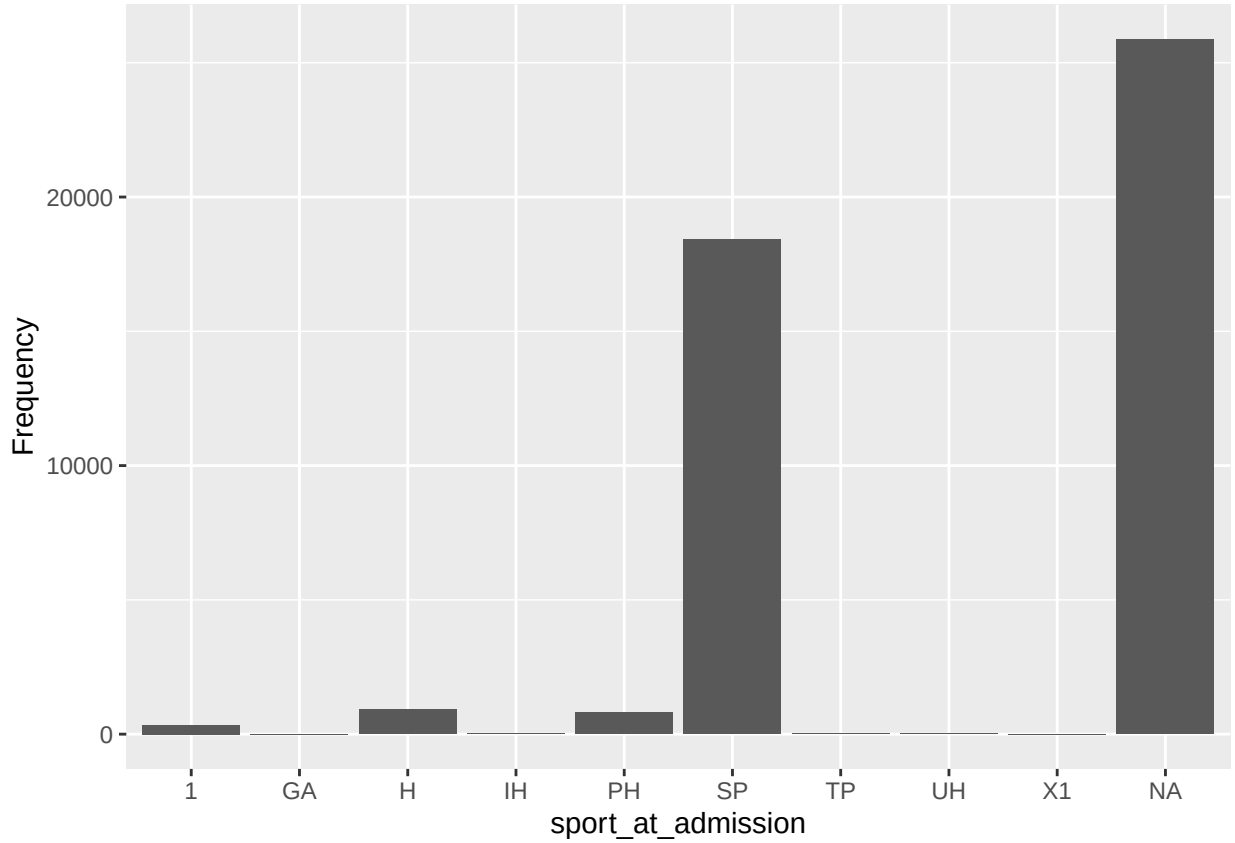

```
[1] should not be used as a predictor
[1] ____________________________
[1] Variable: application_status, type: character
[1] Values (1 unique): Freshmen
[1] Missing: 0%
```



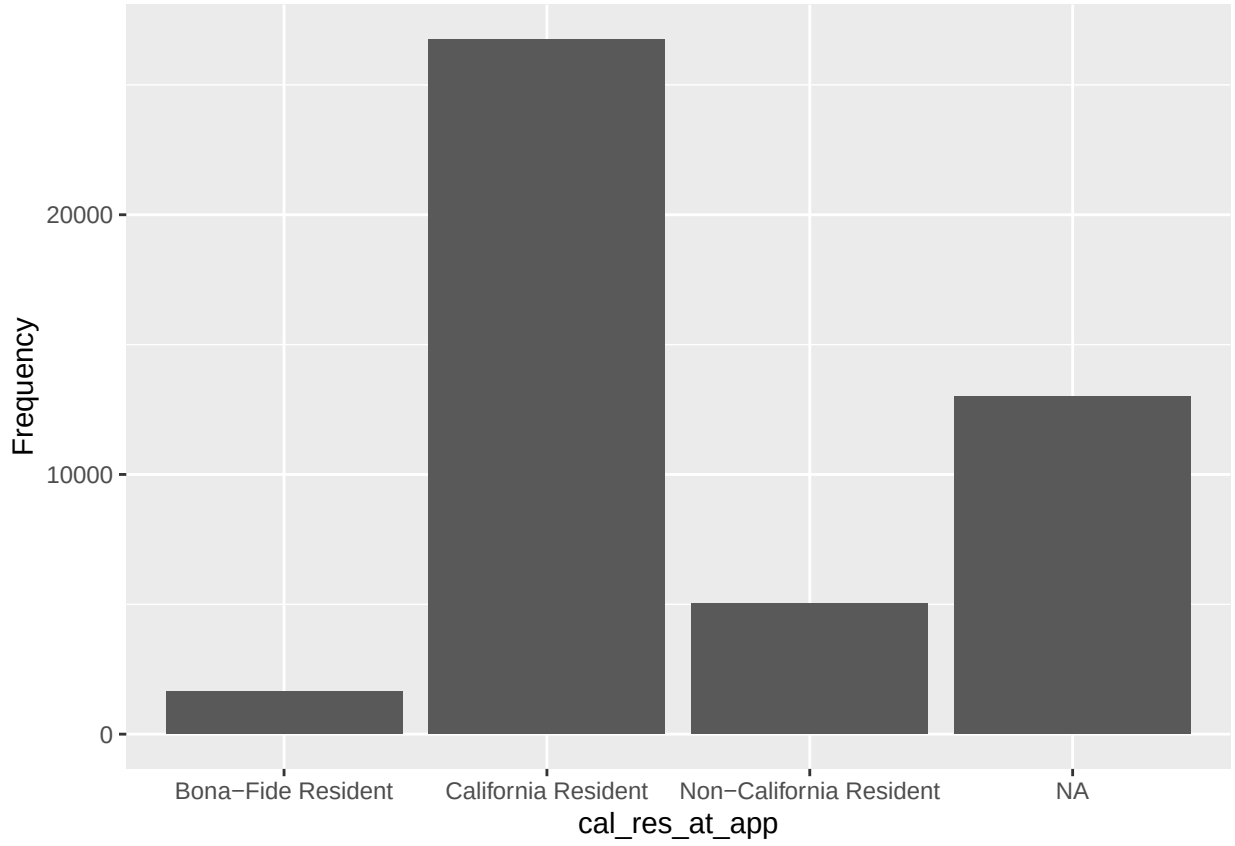

```
[1] ____________________________
[1] Variable: admitdate, type: character
[1] Values (9 unique): F08, F09, F10, F15, F16, ...
[1] Missing: 0%
```



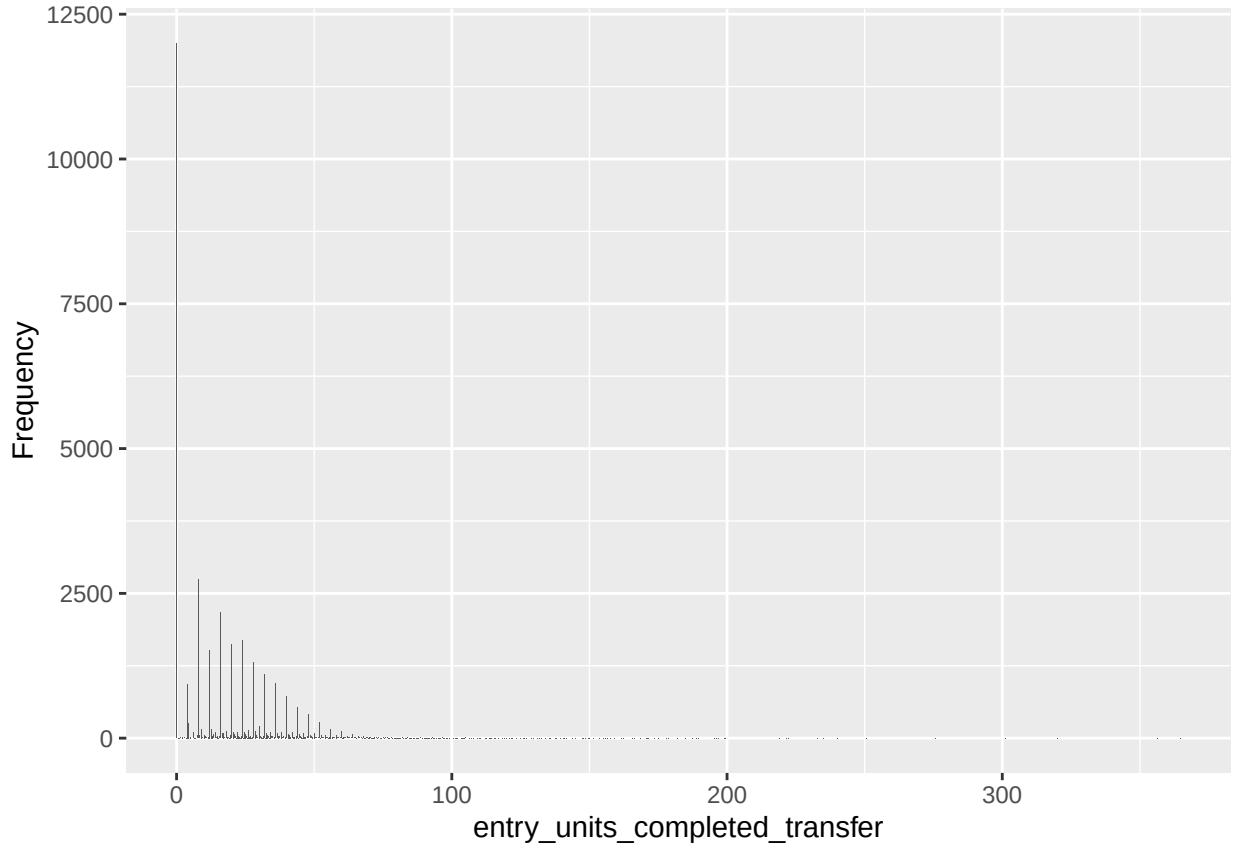

```
[1] ____________________________
[1] Variable: sport_at_admission, type: character
[1] Values (10 unique): NA, SP, PH, H, UH, ...
[1] Missing: 55.7%
  Group.1 sport_at_admission
1     F08         0.99913081
2     F09         1.00000000
3     F10         1.00000000
4     F11         0.06602852
5     F12         0.05906675
6     F13         0.08480500
7     F14         0.21254162
8     F15         0.80087032
9     F16         0.90911859
```



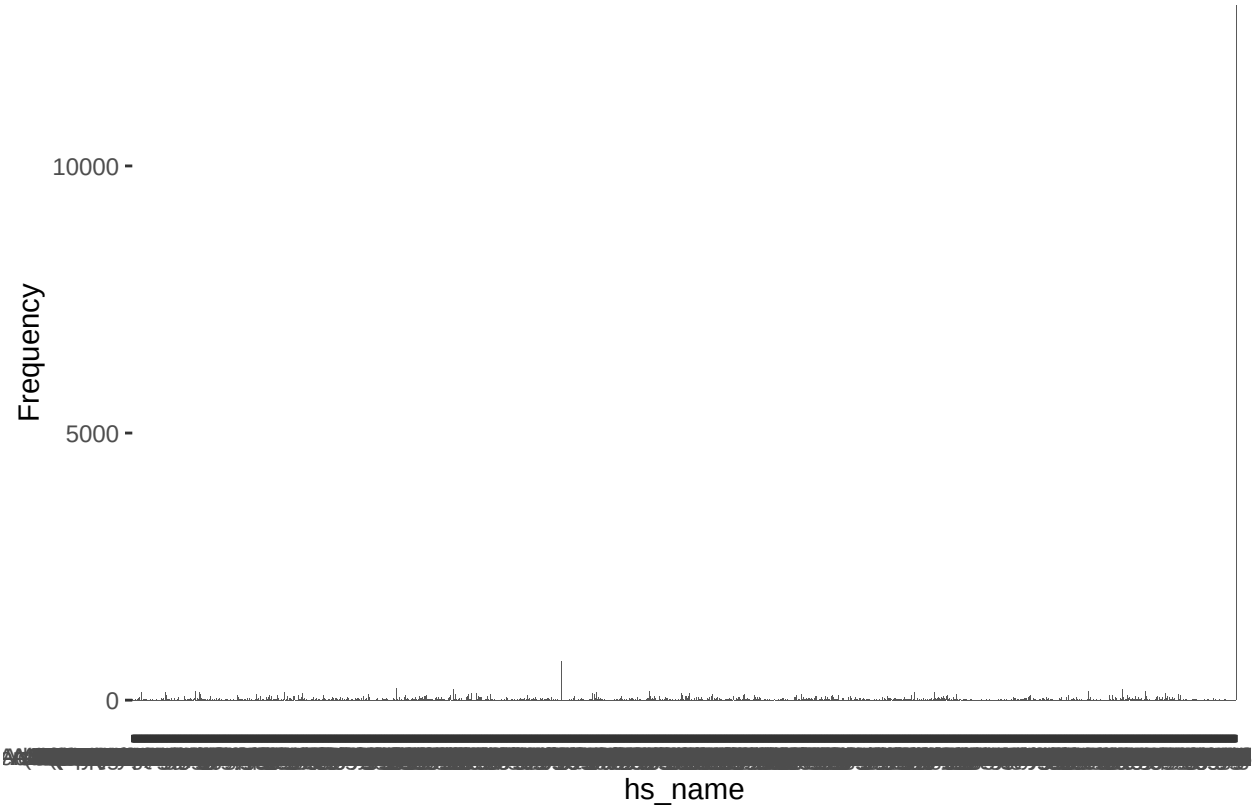

```
[1] ____________________________
[1] Variable: cal_res_at_app, type: character
[1] Values (4 unique): NA, Bona-Fide Resident, California Resident, Non-California Resident
[1] Missing: 28%
  Group.1 cal_res_at_app
1     F08      0.9952195
2     F09      0.9970312
3     F10      0.9968254
4     F11      0.0000000
5     F12      0.0000000
6     F13      0.0000000
7     F14      0.0000000
8     F15      0.0000000
9     F16      0.0000000
```



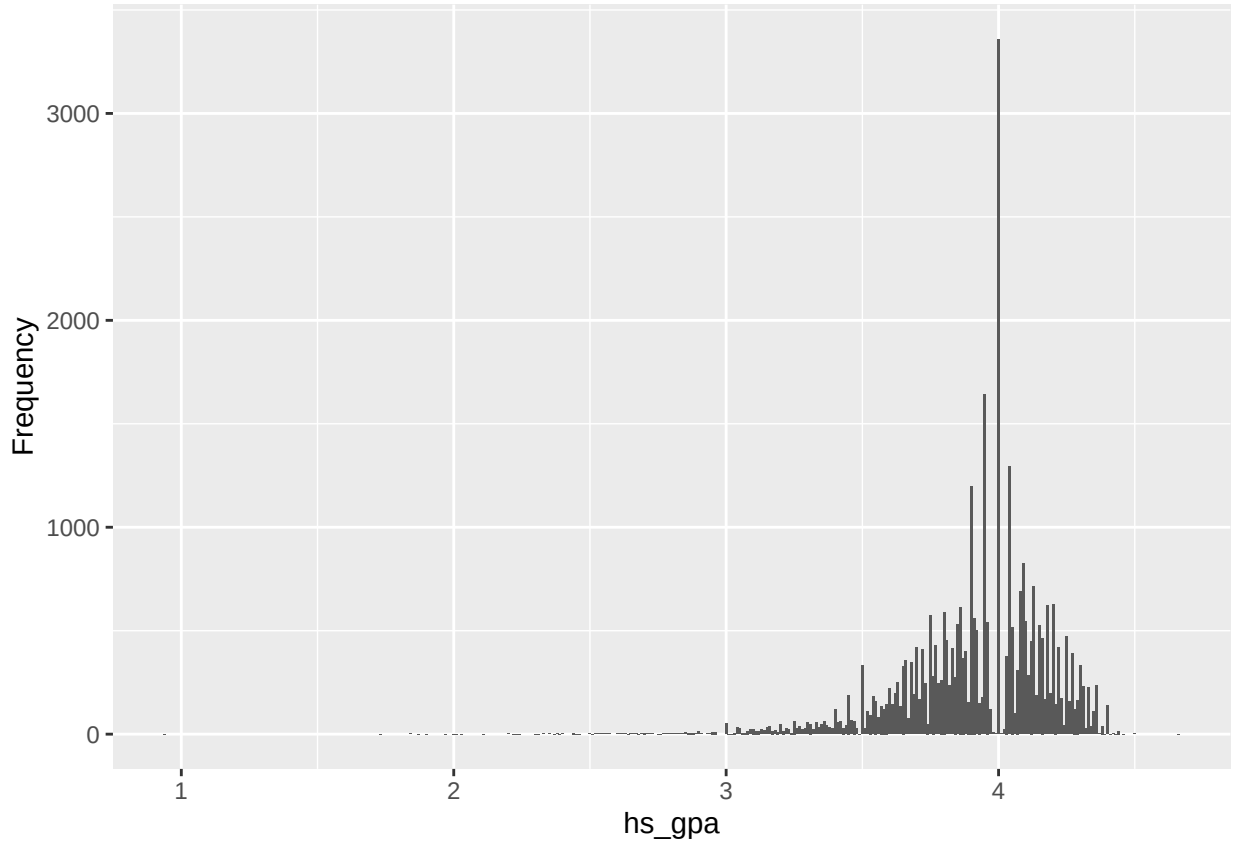

```
[1] ____________________________
[1] Variable: entry_units_completed_transfer, type: numeric
[1] Values (370 unique): NA, 0, 24, 32, 12, ...
[1] Missing: 28%
  Group.1 entry_units_completed_transfer
1     F08                      0.9952195
2     F09                      0.9970312
3     F10                      0.9968254
4     F11                      0.0000000
5     F12                      0.0000000
6     F13                      0.0000000
7     F14                      0.0000000
8     F15                      0.0000000
9     F16                      0.0000000

Warning: Removed 13006 rows containing non-finite values (‘stat_count()‘).
```



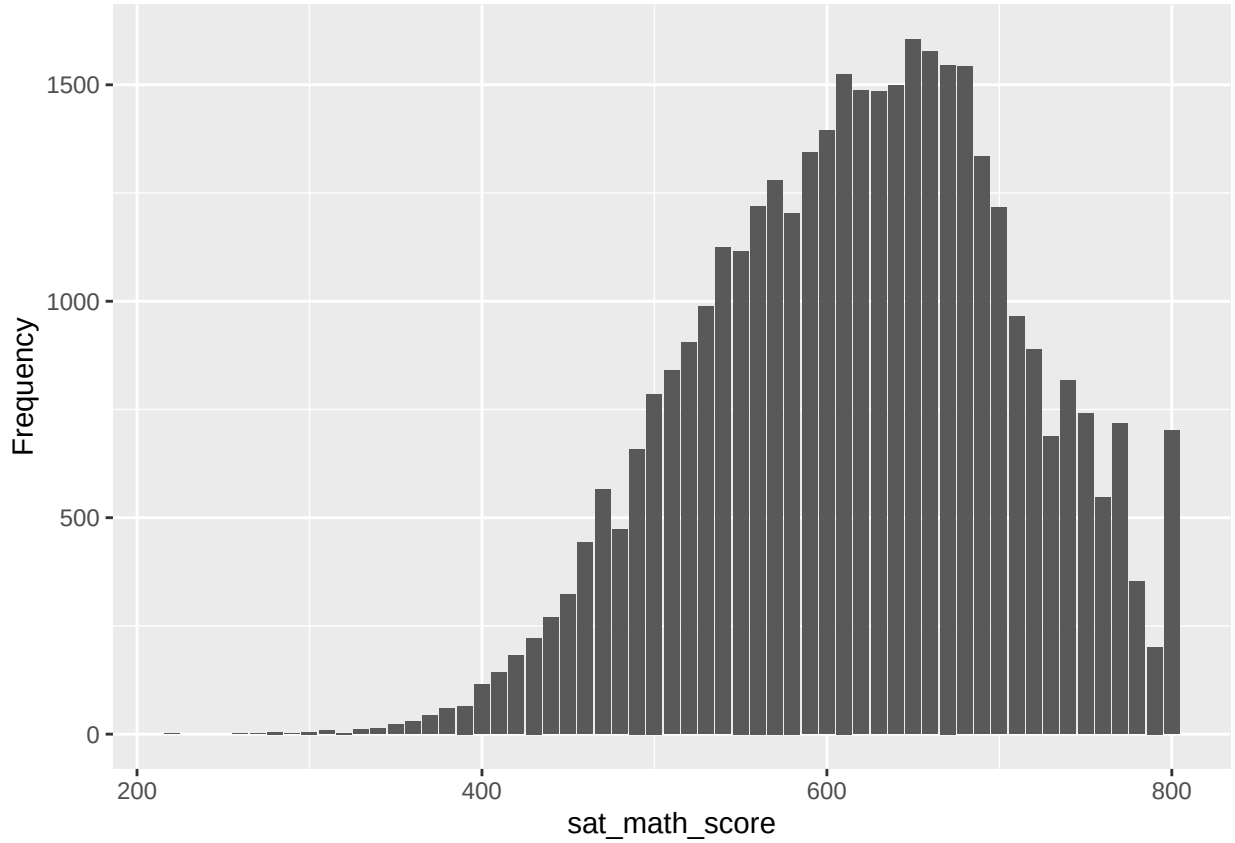

```
[1] ____________________________
[1] Variable: hs_name, type: character
[1] Values (4218 unique): NA, POMONA HS, FRANKLIN, LA QUINTA HS, J A GARFIELD HS, ...
[1] Missing: 28%
  Group.1   hs_name
1     F08 0.9952195
2     F09 0.9970312
3     F10 0.9968254
4     F11 0.0000000
5     F12 0.0000000
6     F13 0.0000000
7     F14 0.0000000
8     F15 0.0000000
9     F16 0.0000000
```



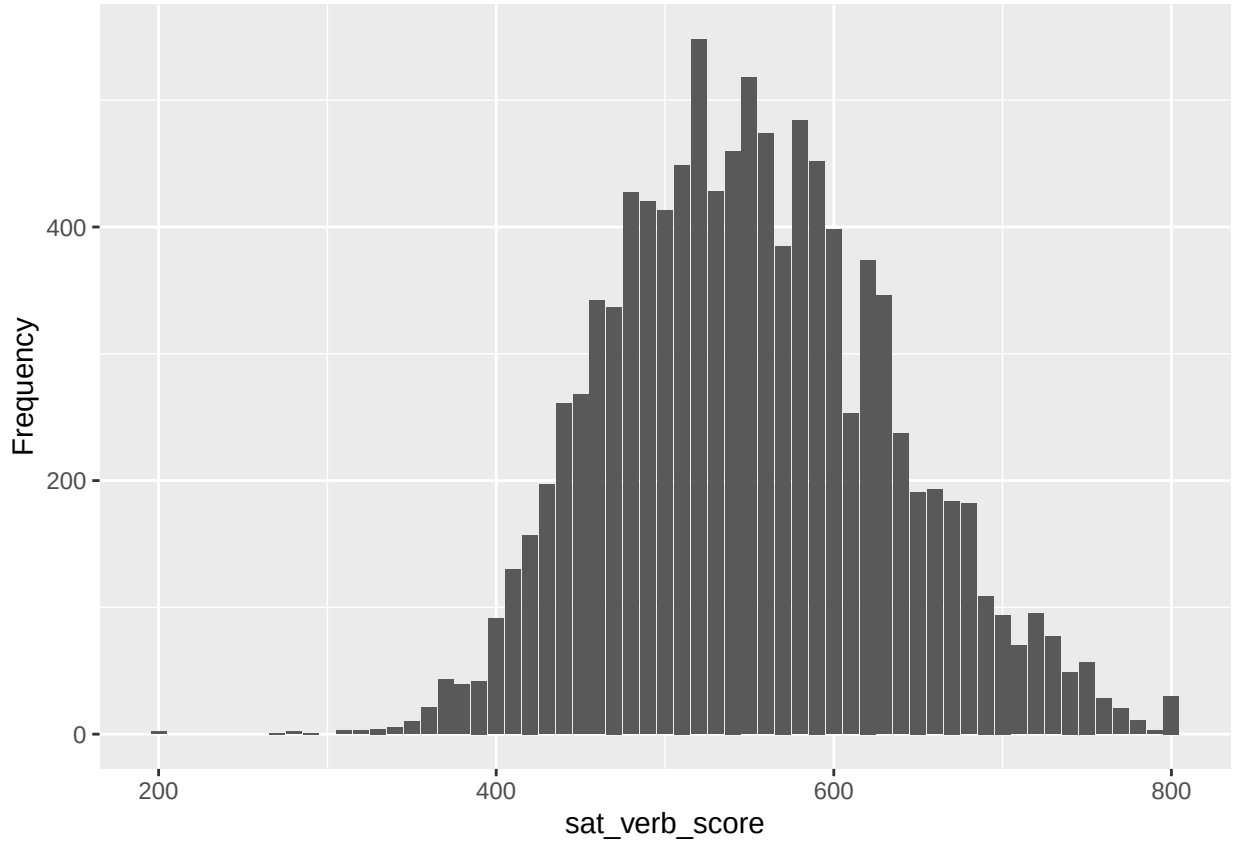

```
[1] ____________________________
[1] Variable: hs_gpa, type: numeric
[1] Values (217 unique): NA, 3.45, 4.0300002, 3.9000001, 4, ...
[1] Missing: 28.1%
  Group.1       hs_gpa
1     F08 0.9952194698
2     F09 0.9972785750
3     F10 0.9970521542
4     F11 0.0011721039
5     F12 0.0003937783
6     F13 0.0005518764
7     F14 0.0001849797
8     F15 0.0006962576
9     F16 0.0012178414

Warning: Removed 13032 rows containing non-finite values (‘stat_count()‘).
```



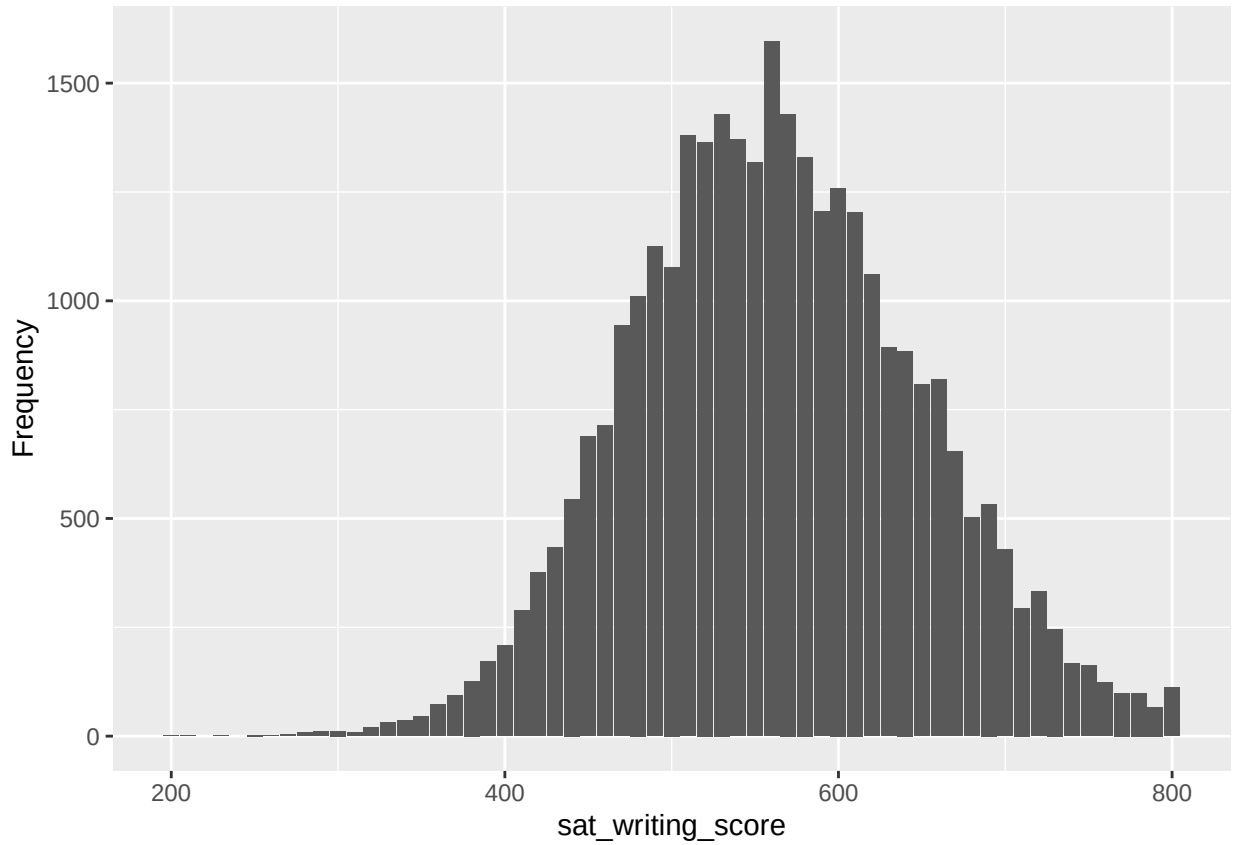

```
[1] ____________________________
[1] Variable: sat_math_score, type: numeric
[1] Values (57 unique): NA, 530, 540, 590, 610, ...
[1] Missing: 19.6%
[1] Most missing: F10 24.8%, Least missing: F15 11.8%

Warning: Removed 9112 rows containing non-finite values (‘stat_count()‘).
```



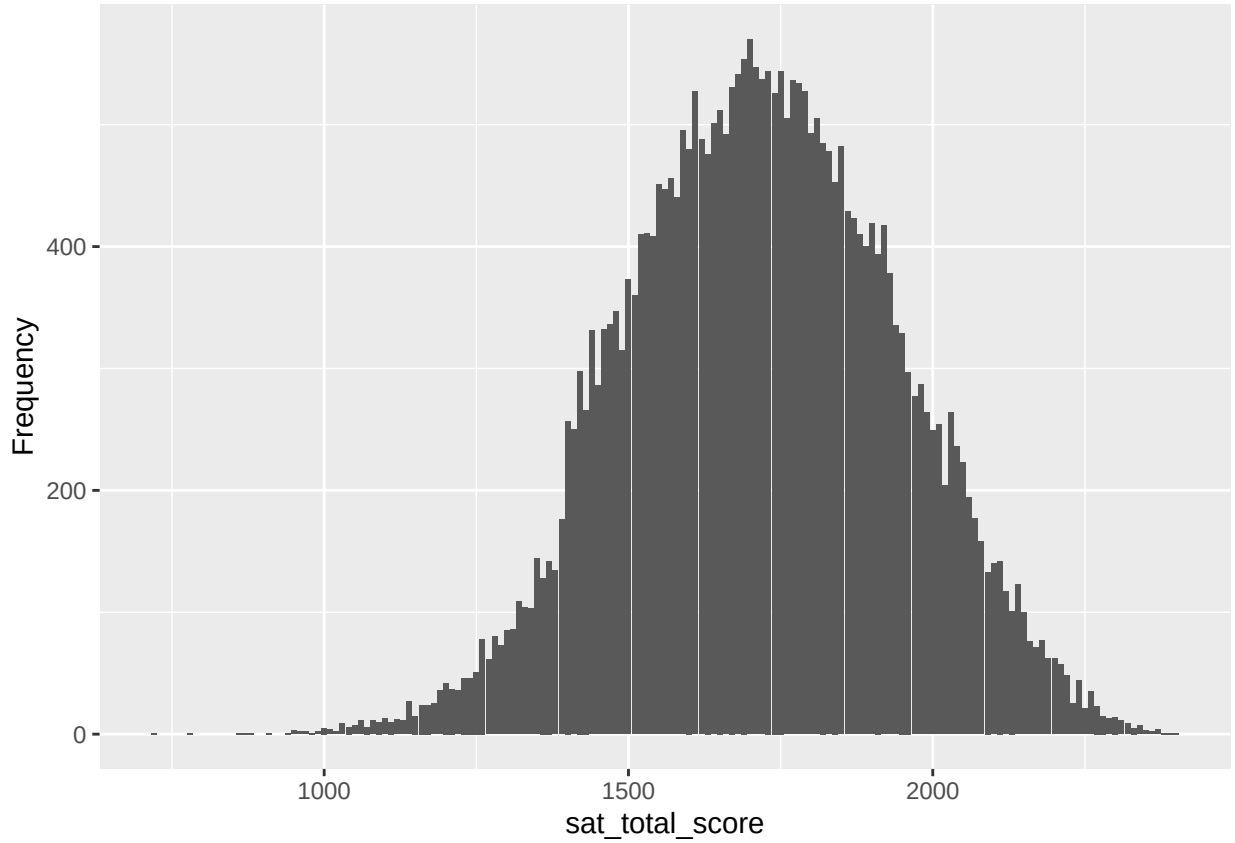

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: sat_verb_score, type: numeric
[1] Values (55 unique): NA, 550, 590, 540, 430, ...
[1] Missing: 77.6%
  Group.1 sat_verb_score
1     F08      1.0000000
2     F09      1.0000000
3     F10      1.0000000
4     F11      0.9888650
5     F12      0.9868084
6     F13      0.9713024
7     F14      0.8571957
8     F15      0.2790252
9     F16      0.2047496

Warning: Removed 35990 rows containing non-finite values (‘stat_count()‘).
```



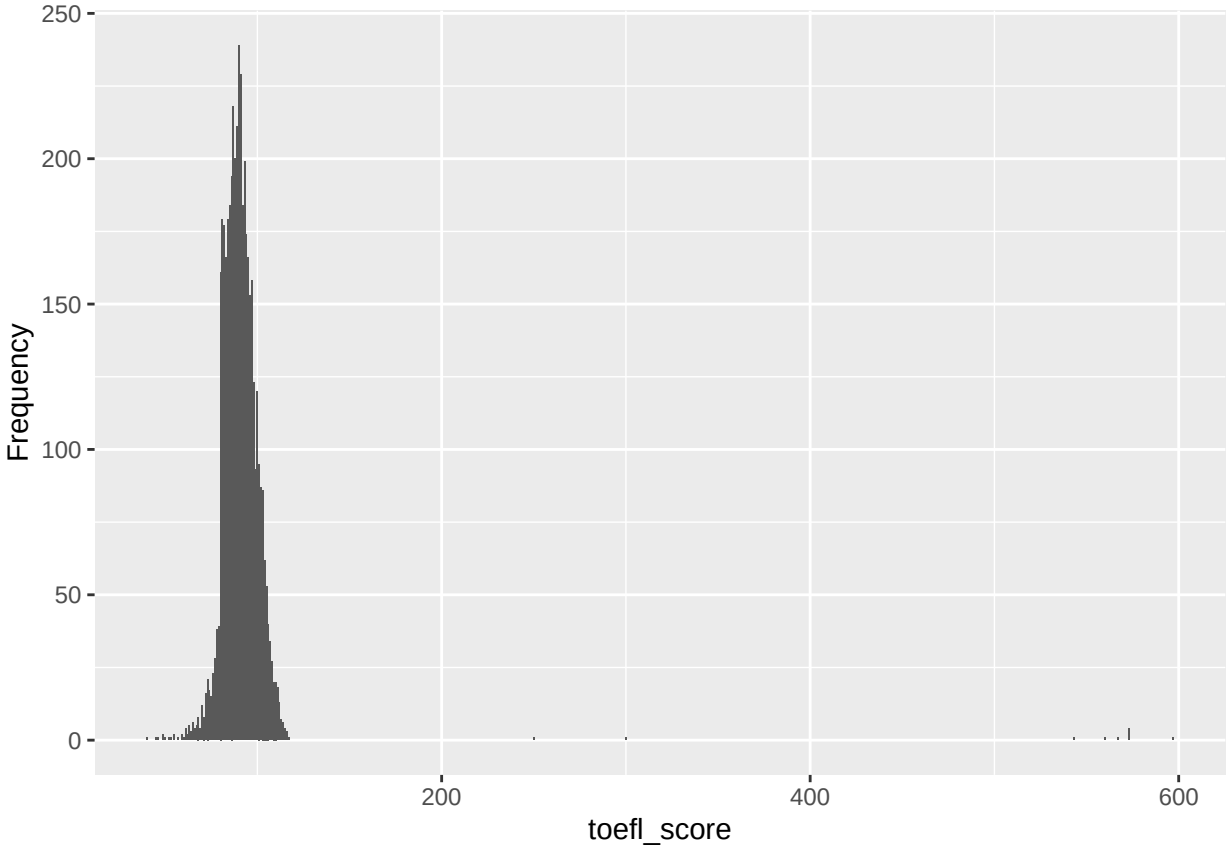

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: sat_writing_score, type: numeric
[1] Values (60 unique): NA, 580, 560, 540, 520, ...
[1] Missing: 32.7%
  Group.1 sat_writing_score
1     F08        0.99630595
2     F09        0.99777338
3     F10        0.99841270
4     F11        0.03106075
5     F12        0.03740894
6     F13        0.04194260
7     F14        0.05512394
8     F15        0.08442124
9     F16        0.12147968

Warning: Removed 15179 rows containing non-finite values (‘stat_count()‘).
```




```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: sat_total_score, type: numeric
[1] Values (154 unique): NA, 1660, 1690, 1670, 1680, ...
[1] Missing: 32.7%
  Group.1 sat_total_score
1     F08      0.99630595
2     F09      0.99777338
3     F10      0.99841270
4     F11      0.02949795
5     F12      0.03681827
6     F13      0.04120677
7     F14      0.05512394
8     F15      0.08442124
9     F16      0.12147968

Warning: Removed 15164 rows containing non-finite values (‘stat_count()‘).
```

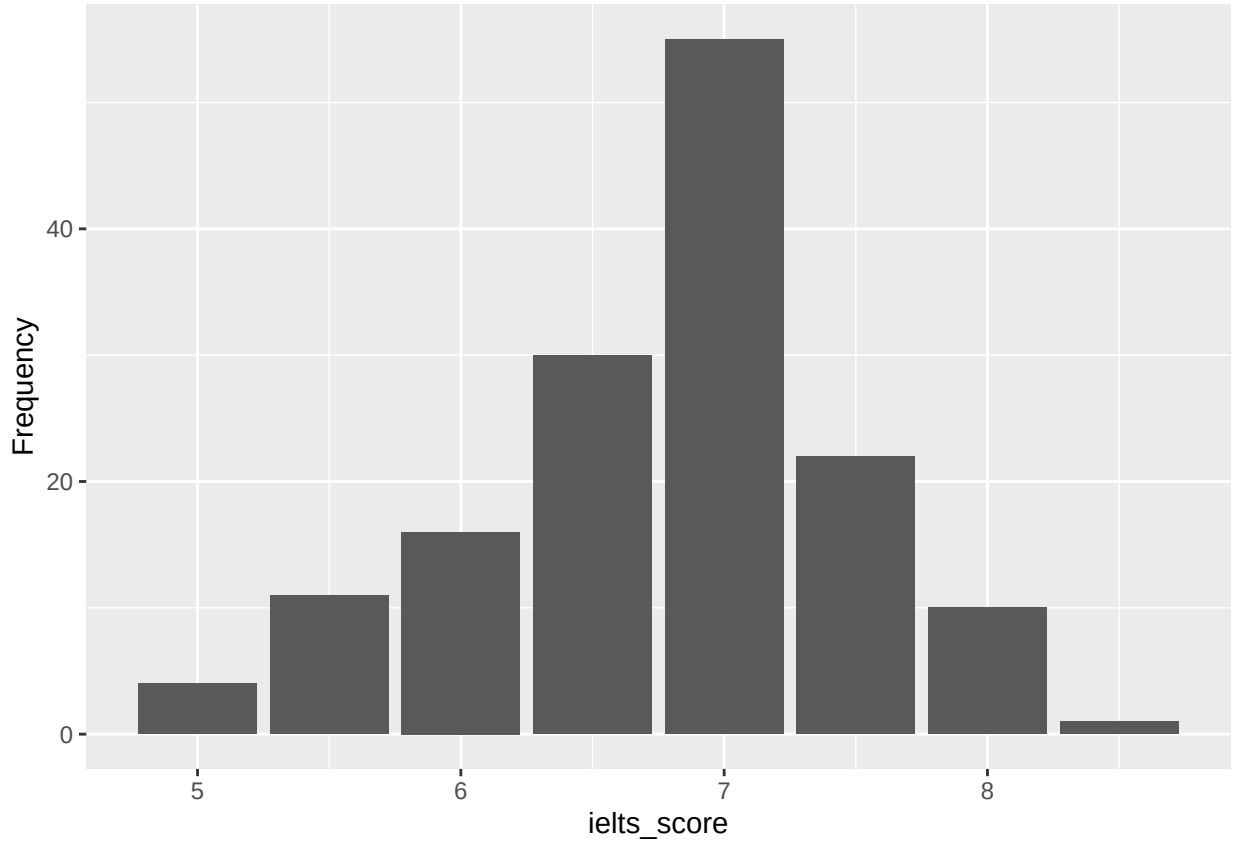



```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: toefl_score, type: numeric
[1] Values (76 unique): NA, 83, 100, 94, 93, ...
[1] Missing: 90.2%
[1] Most missing: F08 100%, Least missing: F16 79.1%

Warning: Removed 41843 rows containing non-finite values (‘stat_count()‘).
```

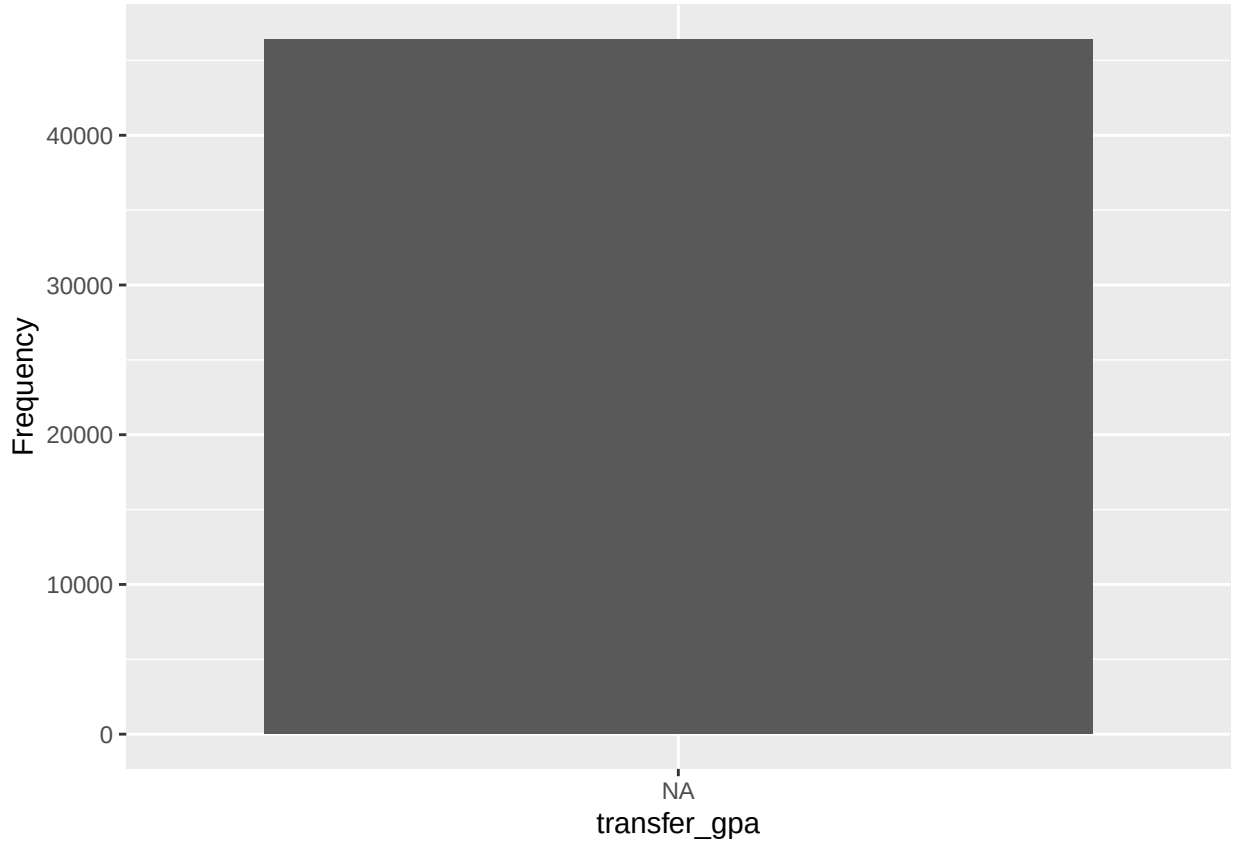



```
[1] ____________________________
[1] Variable: ielts_score, type: numeric
[1] Values (9 unique): NA, 7.5, 6.5, 6, 7, ...
[1] Missing: 99.7%
```

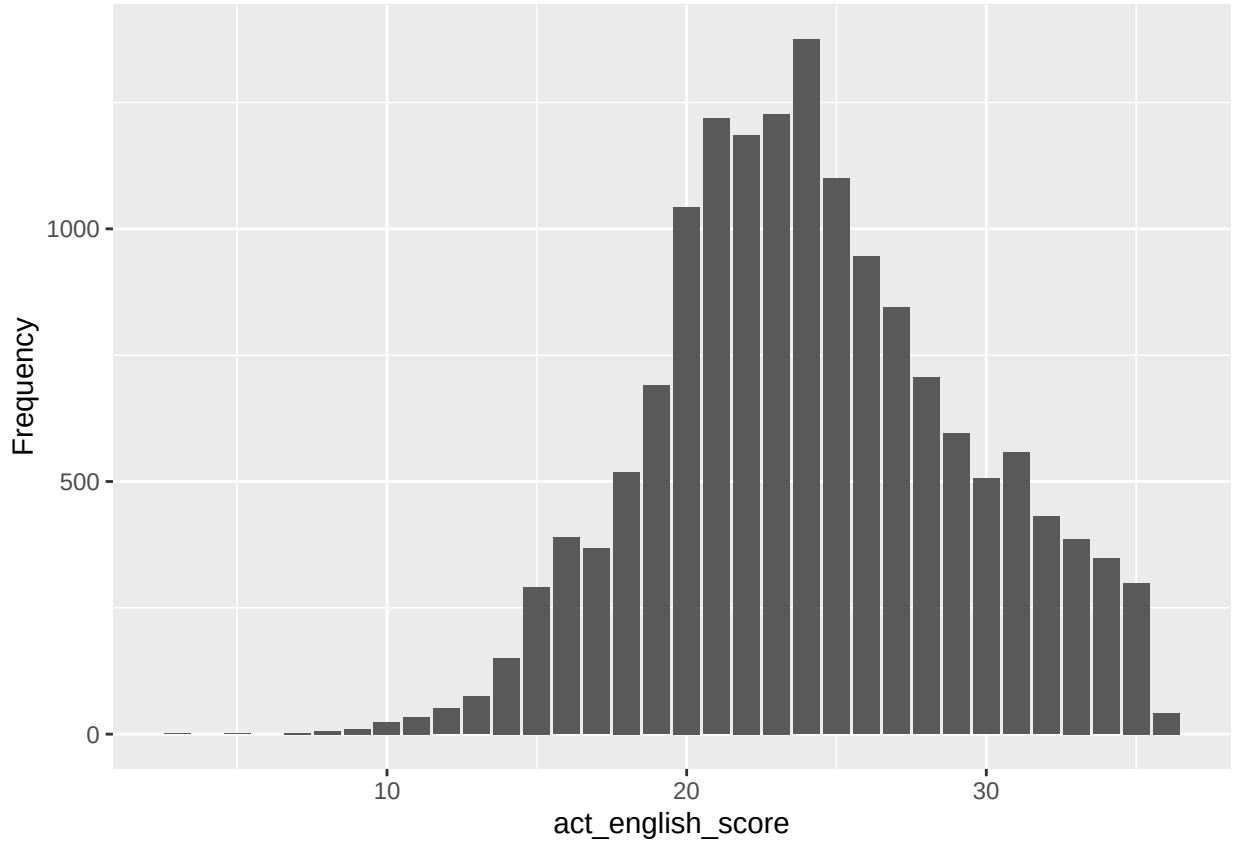
```
Warning: Removed 46259 rows containing non-finite values (‘stat_count()‘).
```



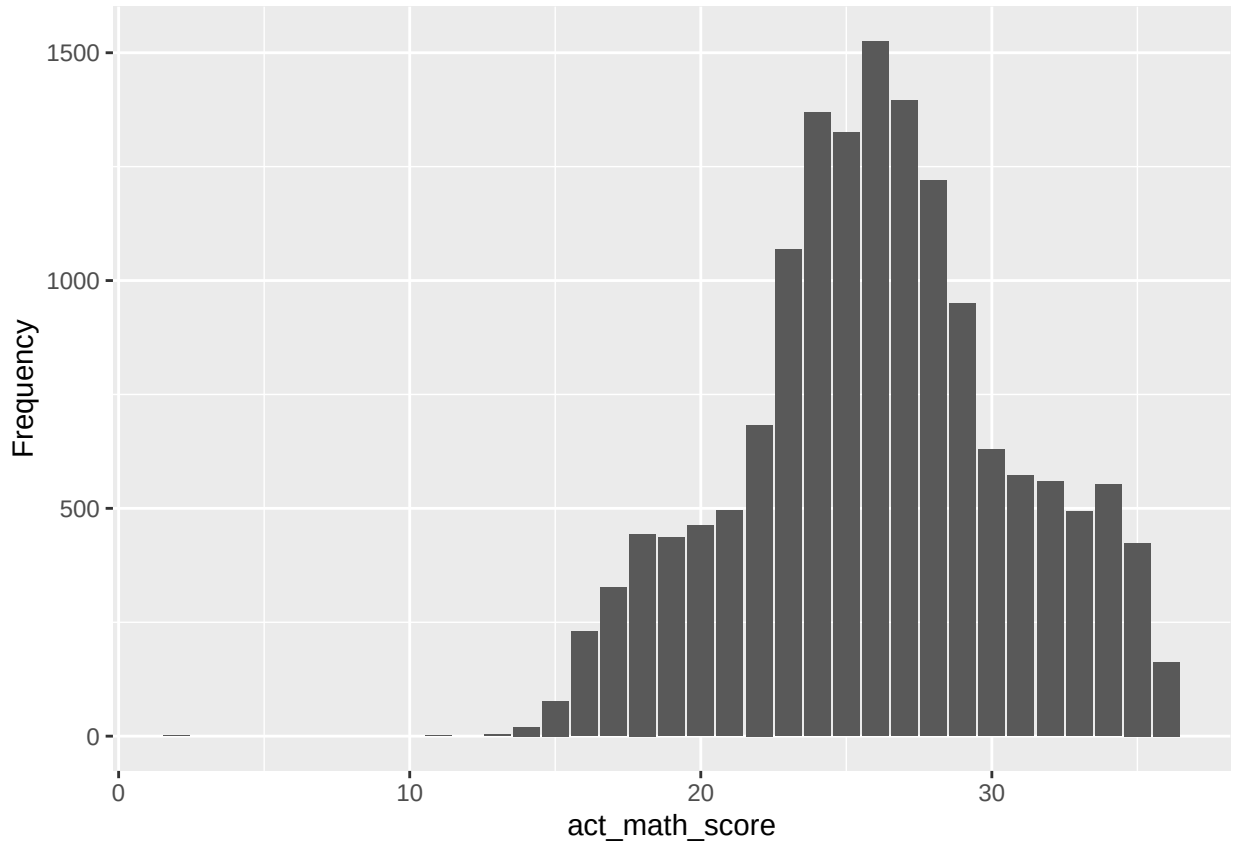

```
[1] ____________________________
[1] Variable: transfer_gpa, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
```



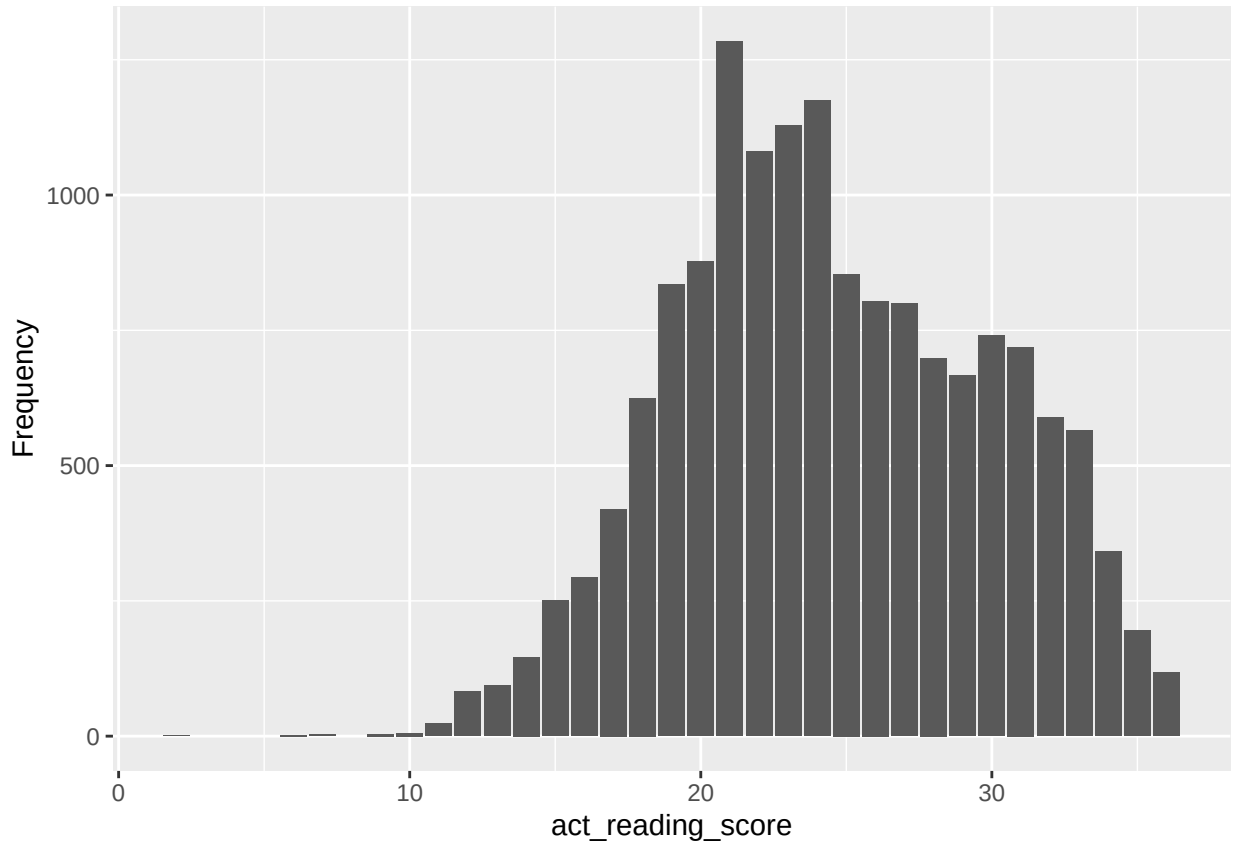

```
[1] ____________________________
[1] Variable: act_english_score, type: numeric
[1] Values (33 unique): NA, 25, 24, 16, 21, ...
[1] Missing: 66.8%
  Group.1 act_english_score
1     F08         0.9989135
2     F09         0.9987630
3     F10         0.9995465
4     F11         0.5540145
5     F12         0.5678283
6     F13         0.5603385
7     F14         0.5421754
8     F15         0.5166232
9     F16         0.4994672

Warning: Removed 30988 rows containing non-finite values (‘stat_count()‘).
```



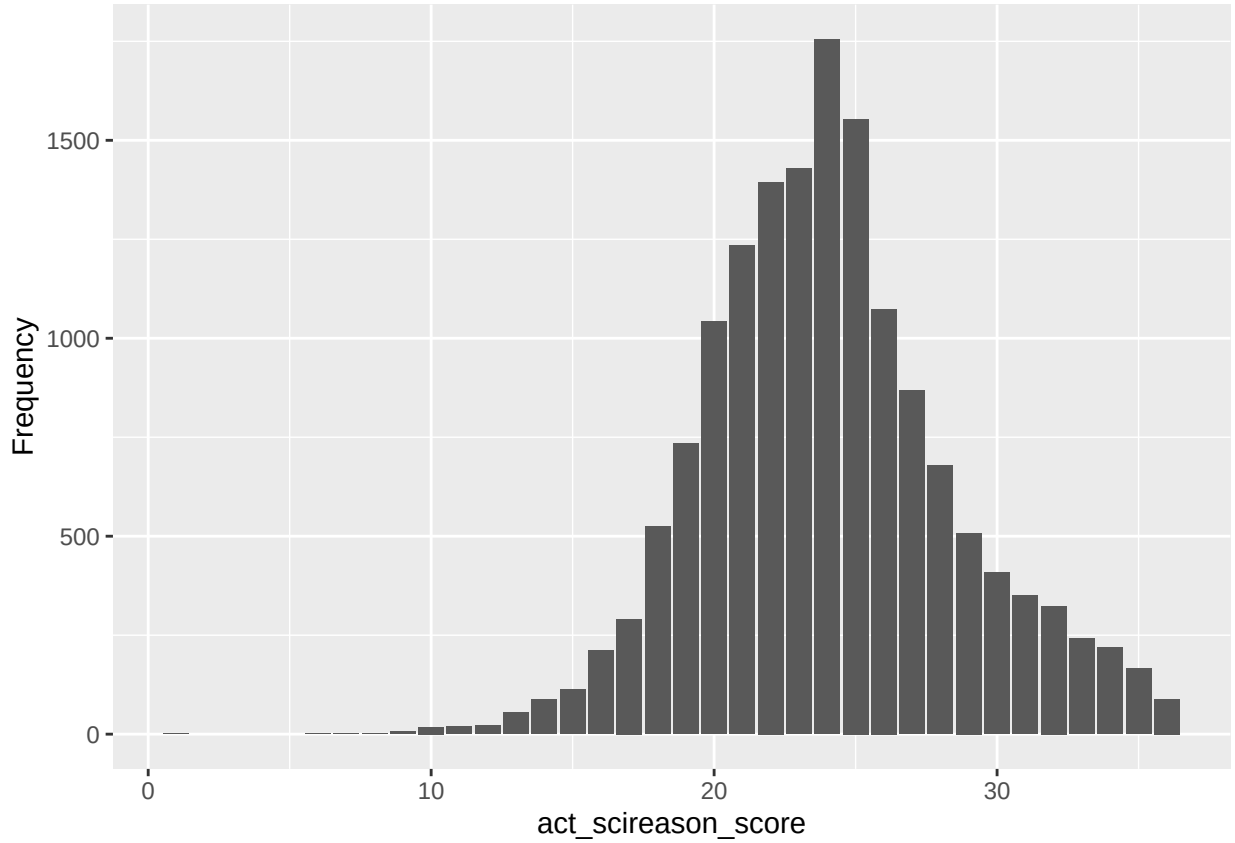

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: act_math_score, type: numeric
[1] Values (27 unique): NA, 27, 26, 19, 30, ...
[1] Missing: 66.8%
  Group.1 act_math_score
1     F08      0.9989135
2     F09      0.9987630
3     F10      0.9995465
4     F11      0.5540145
5     F12      0.5678283
6     F13      0.5603385
7     F14      0.5421754
8     F15      0.5166232
9     F16      0.4994672

Warning: Removed 30988 rows containing non-finite values (‘stat_count()‘).
```



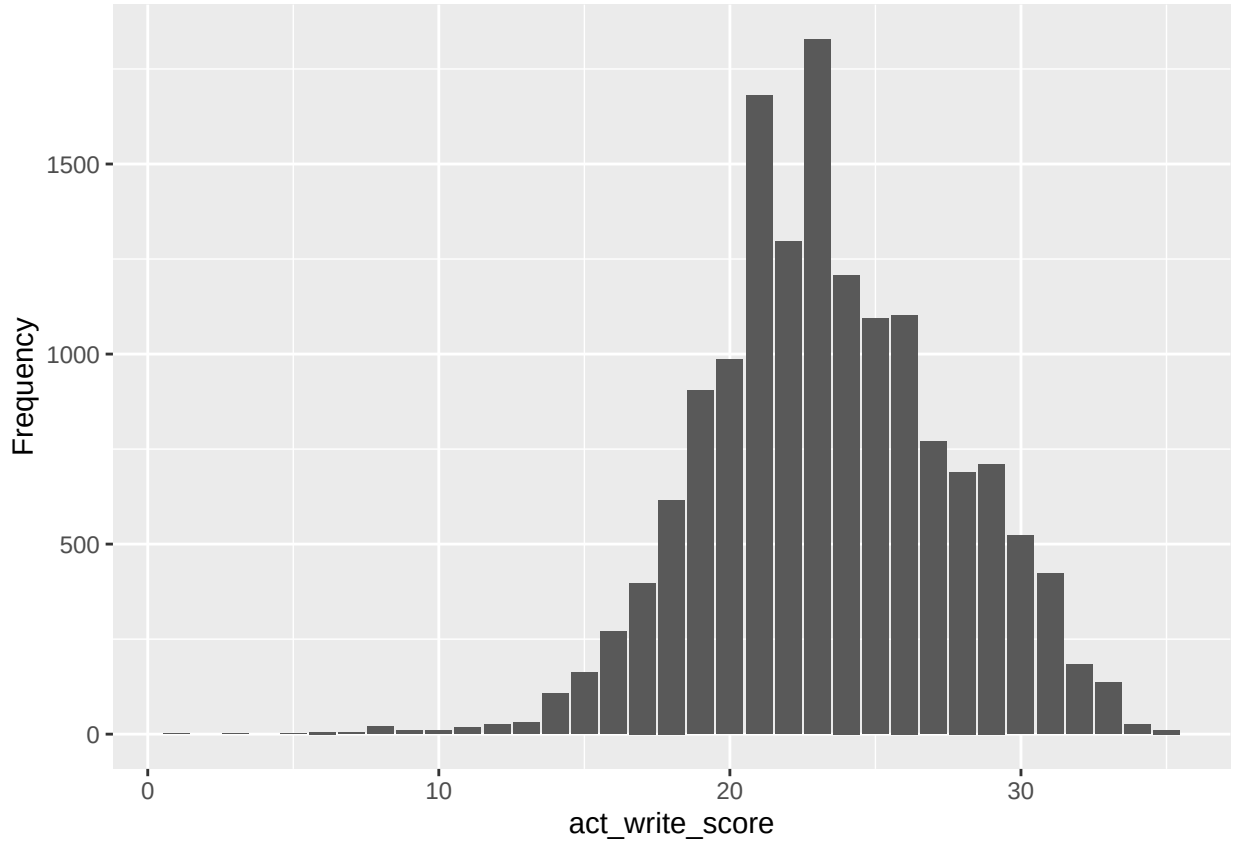

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: act_reading_score, type: numeric
[1] Values (32 unique): NA, 25, 24, 19, 22, ...
[1] Missing: 66.8%
  Group.1 act_reading_score
1     F08         0.9989135
2     F09         0.9987630
3     F10         0.9995465
4     F11         0.5540145
5     F12         0.5678283
6     F13         0.5603385
7     F14         0.5421754
8     F15         0.5166232
9     F16         0.4994672

Warning: Removed 30988 rows containing non-finite values (‘stat_count()‘).
```



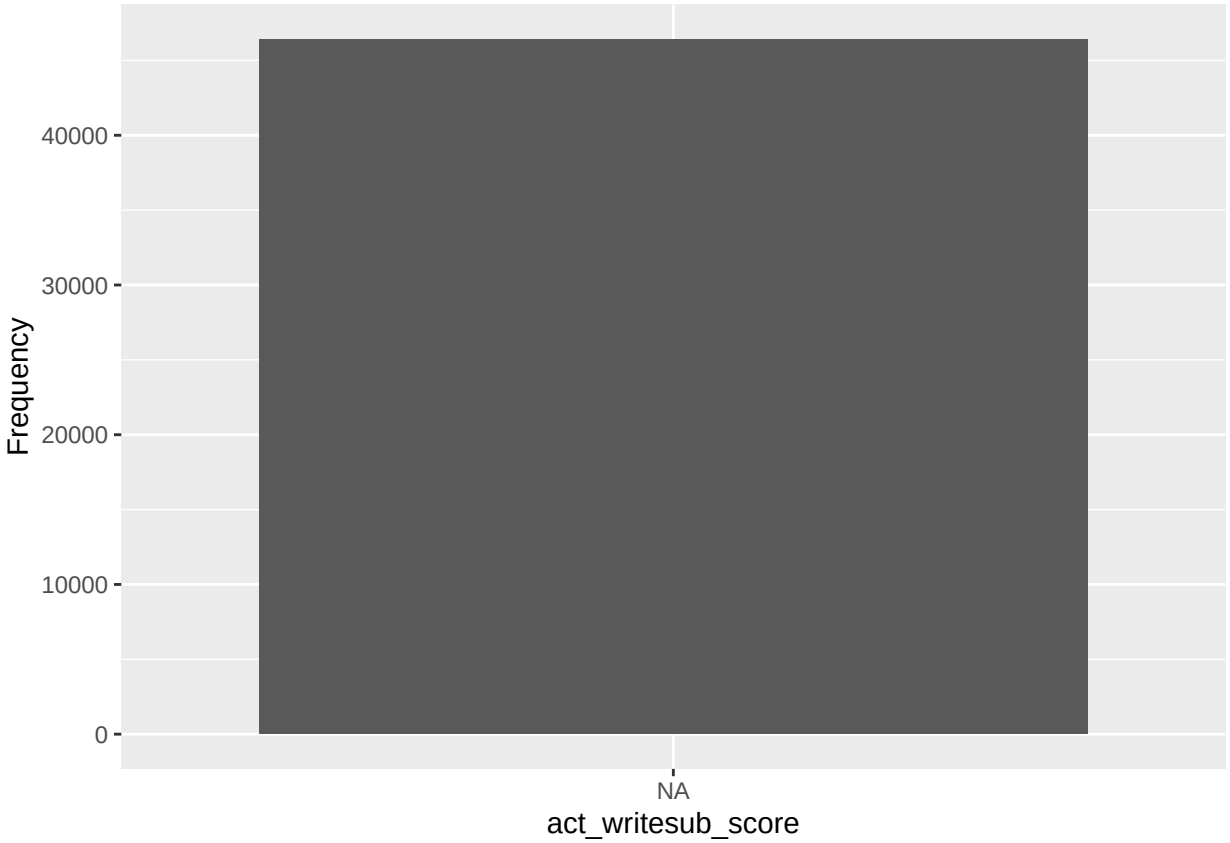

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: act_scireason_score, type: numeric
[1] Values (33 unique): NA, 23, 26, 25, 19, ...
[1] Missing: 66.8%
  Group.1 act_scireason_score
1     F08           0.9989135
2     F09           0.9987630
3     F10           0.9995465
4     F11           0.5540145
5     F12           0.5678283
6     F13           0.5603385
7     F14           0.5421754
8     F15           0.5166232
9     F16           0.4994672

Warning: Removed 30988 rows containing non-finite values (‘stat_count()‘).
```



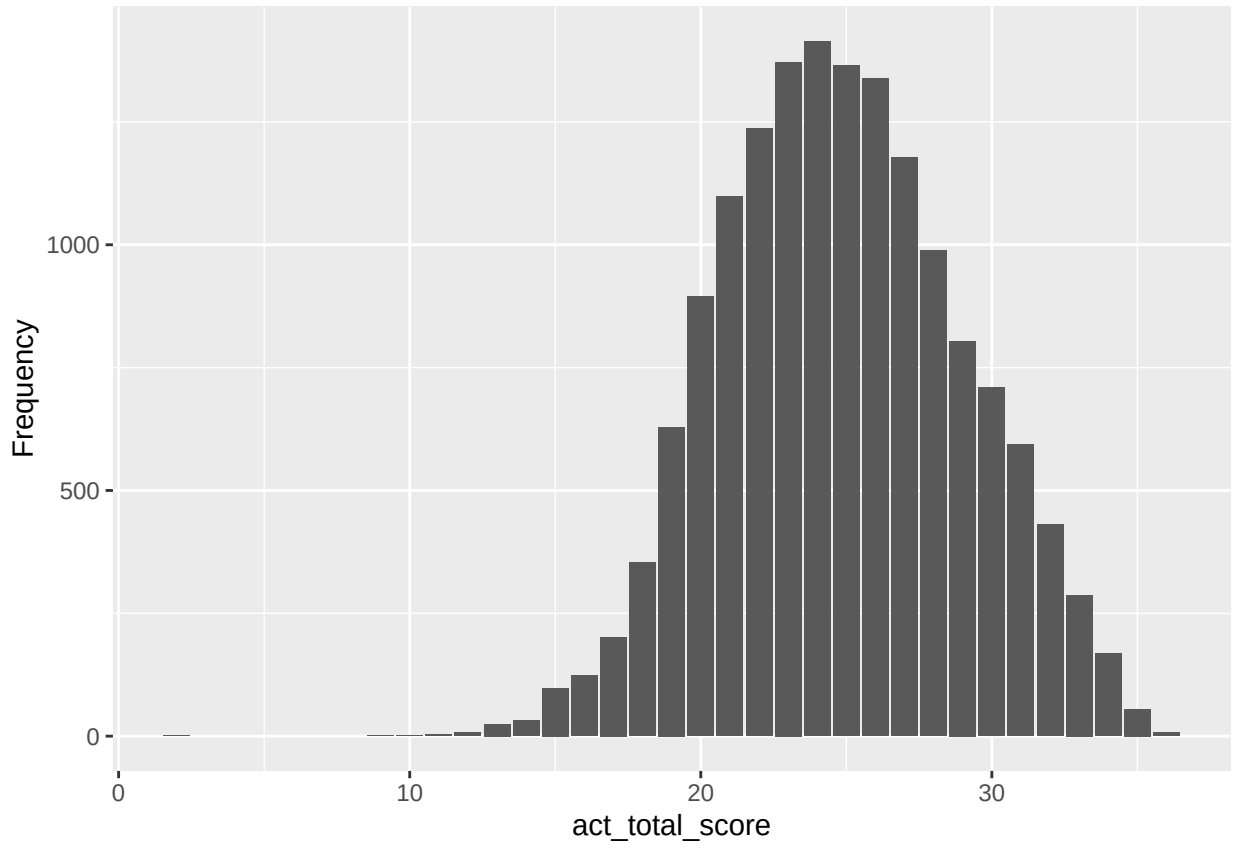

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: act_write_score, type: numeric
[1] Values (34 unique): NA, 25, 23, 19, 30, ...
[1] Missing: 67.1%
[1] Most missing: F10 100%, Least missing: F16 50.6%

Warning: Removed 31156 rows containing non-finite values (‘stat_count()‘).
```



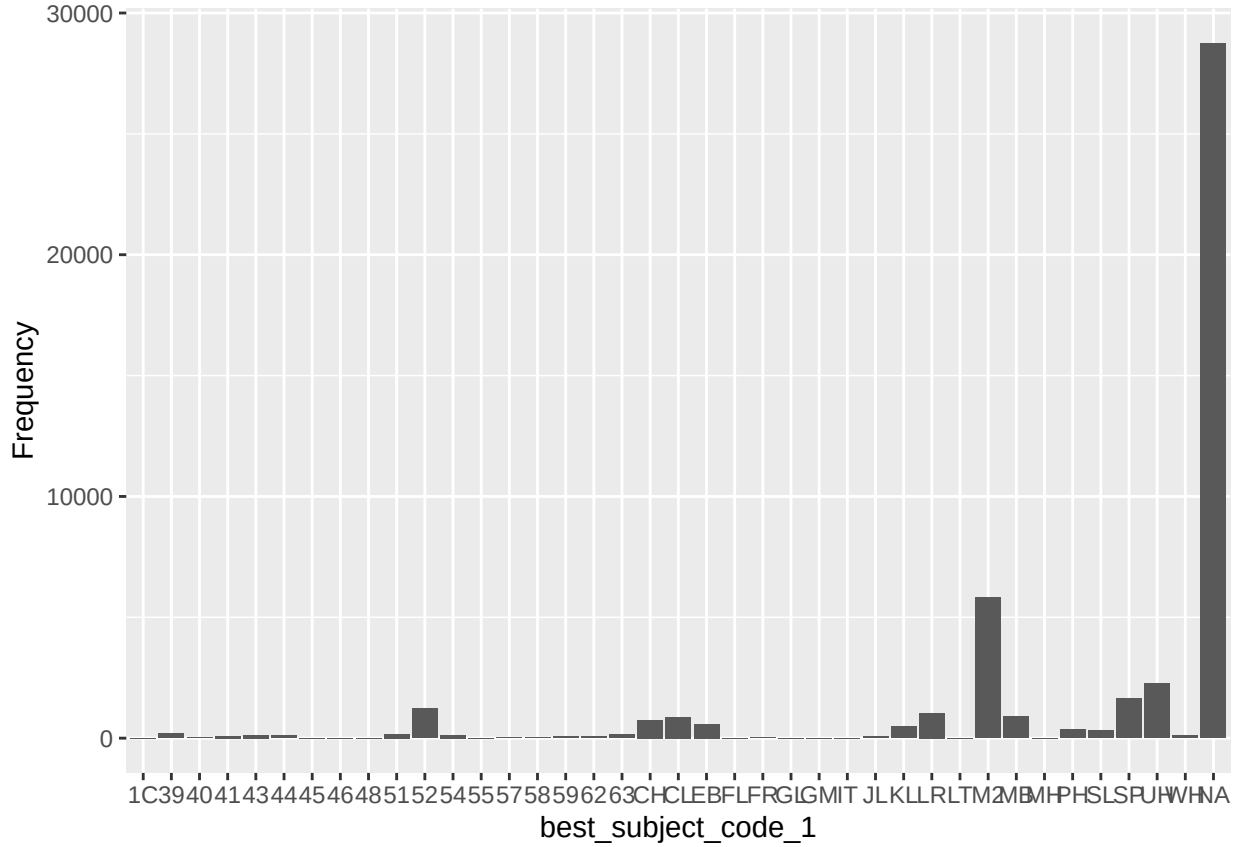

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: act_writesub_score, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
```



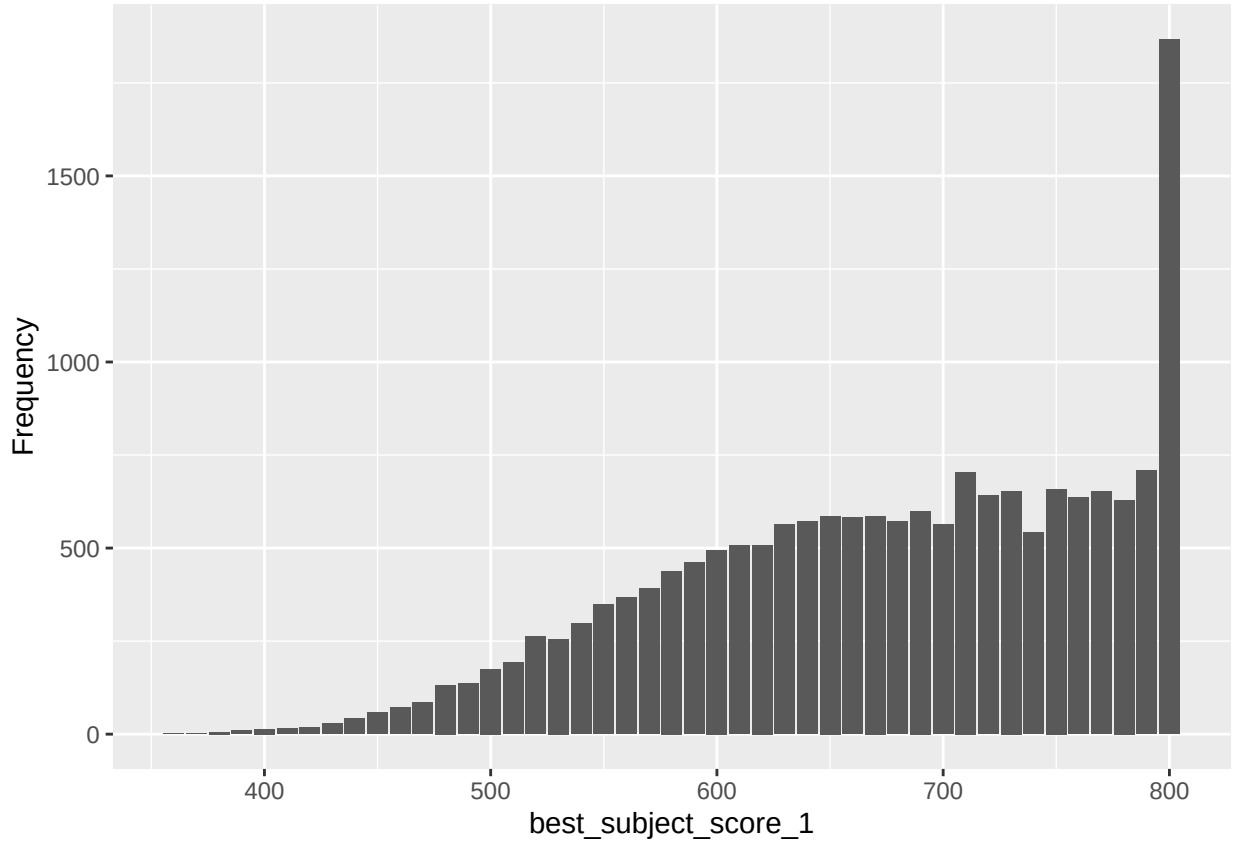

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: act_total_score, type: numeric
[1] Values (30 unique): NA, 25, 22, 20, 30, ...
[1] Missing: 66.8%
  Group.1 act_total_score
1     F08       0.9989135
2     F09       0.9987630
3     F10       0.9995465
4     F11       0.5540145
5     F12       0.5678283
6     F13       0.5603385
7     F14       0.5421754
8     F15       0.5166232
9     F16       0.4994672

Warning: Removed 30988 rows containing non-finite values (‘stat_count()‘).
```



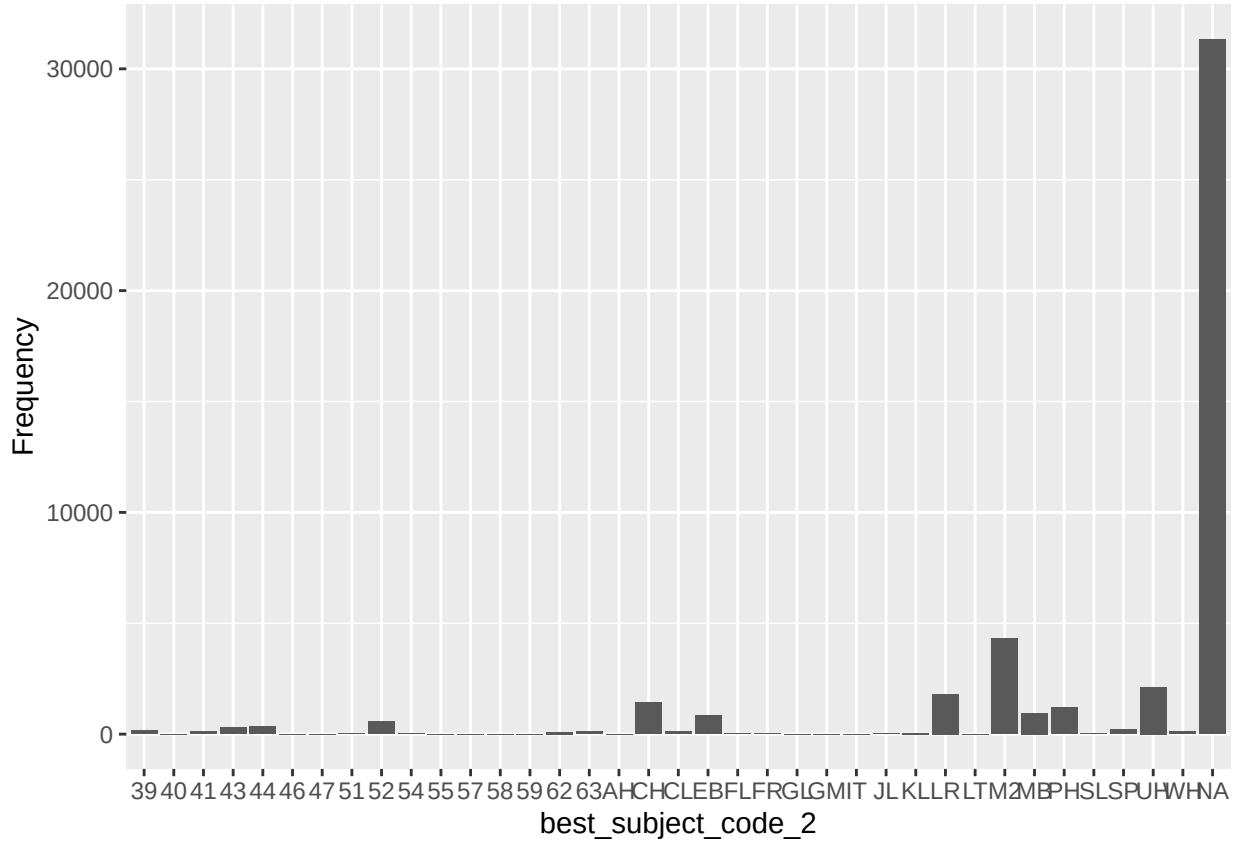

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: best_subject_code_1, type: character
[1] Values (39 unique): NA, UH, 39, M2, 52, ...
[1] Missing: 61.9%
  Group.1 best_subject_code_1
1     F08         0.996305954
2     F09         0.997773380
3     F10         0.998412698
4     F11         0.002148857
5     F12         0.454223272
6     F13         0.519499632
7     F14         0.558638550
8     F15         0.614273281
9     F16         0.613944284
```



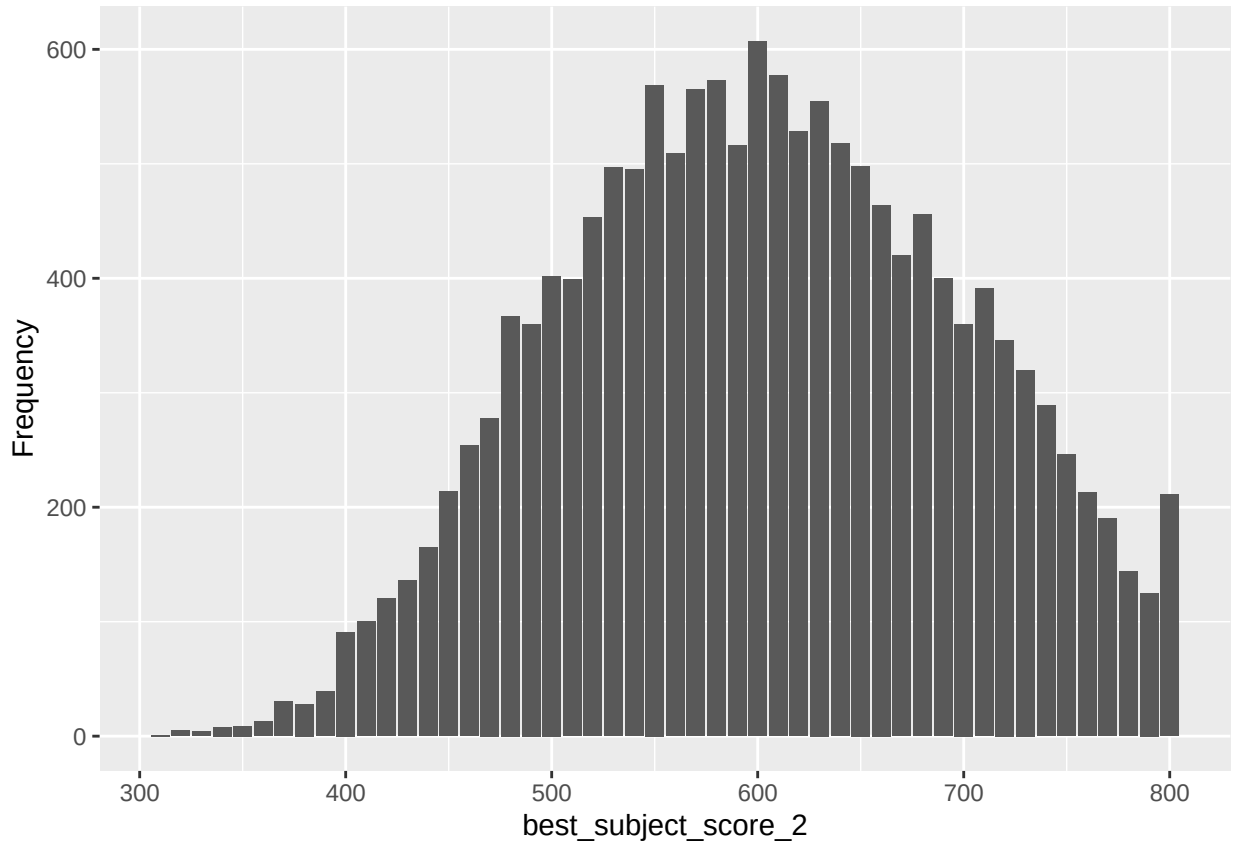

```
[1] ____________________________
[1] Variable: best_subject_score_1, type: integer
[1] Values (46 unique): NA, 540, 610, 650, 720, ...
[1] Missing: 62%
  Group.1 best_subject_score_1
1     F08          0.996305954
2     F09          0.997773380
3     F10          0.998412698
4     F11          0.002148857
5     F12          0.454223272
6     F13          0.519499632
7     F14          0.558638550
8     F15          0.618276762
9     F16          0.613944284

Warning: Removed 28768 rows containing non-finite values (‘stat_count()‘).
```



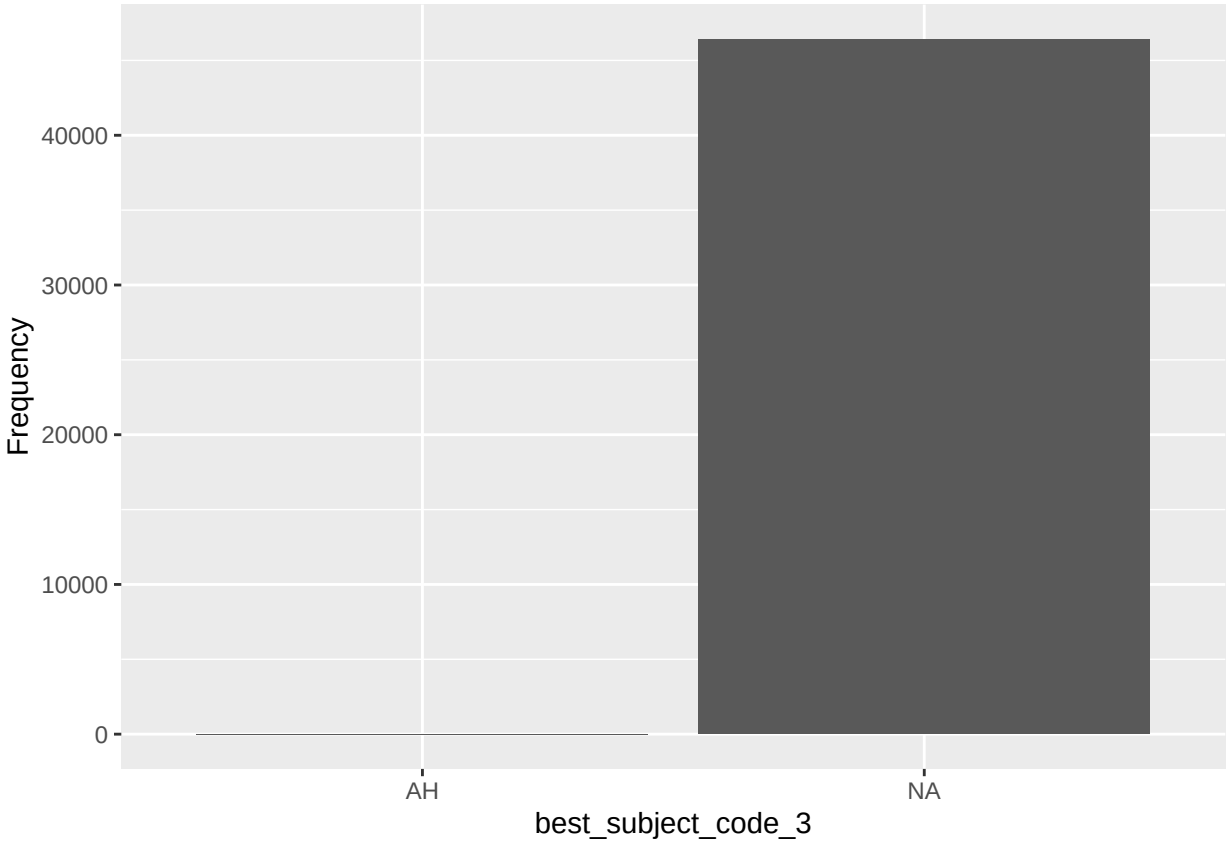

```
[1] ____________________________
[1] Variable: best_subject_code_2, type: character
[1] Values (37 unique): NA, M2, 52, EB, PH, ...
[1] Missing: 67.5%
  Group.1 best_subject_code_2
1     F08         0.996305954
2     F09         0.997773380
3     F10         0.998412698
4     F11         0.002539559
5     F12         0.546367395
6     F13         0.609639441
7     F14         0.651128376
8     F15         0.697302002
9     F16         0.712589435
```



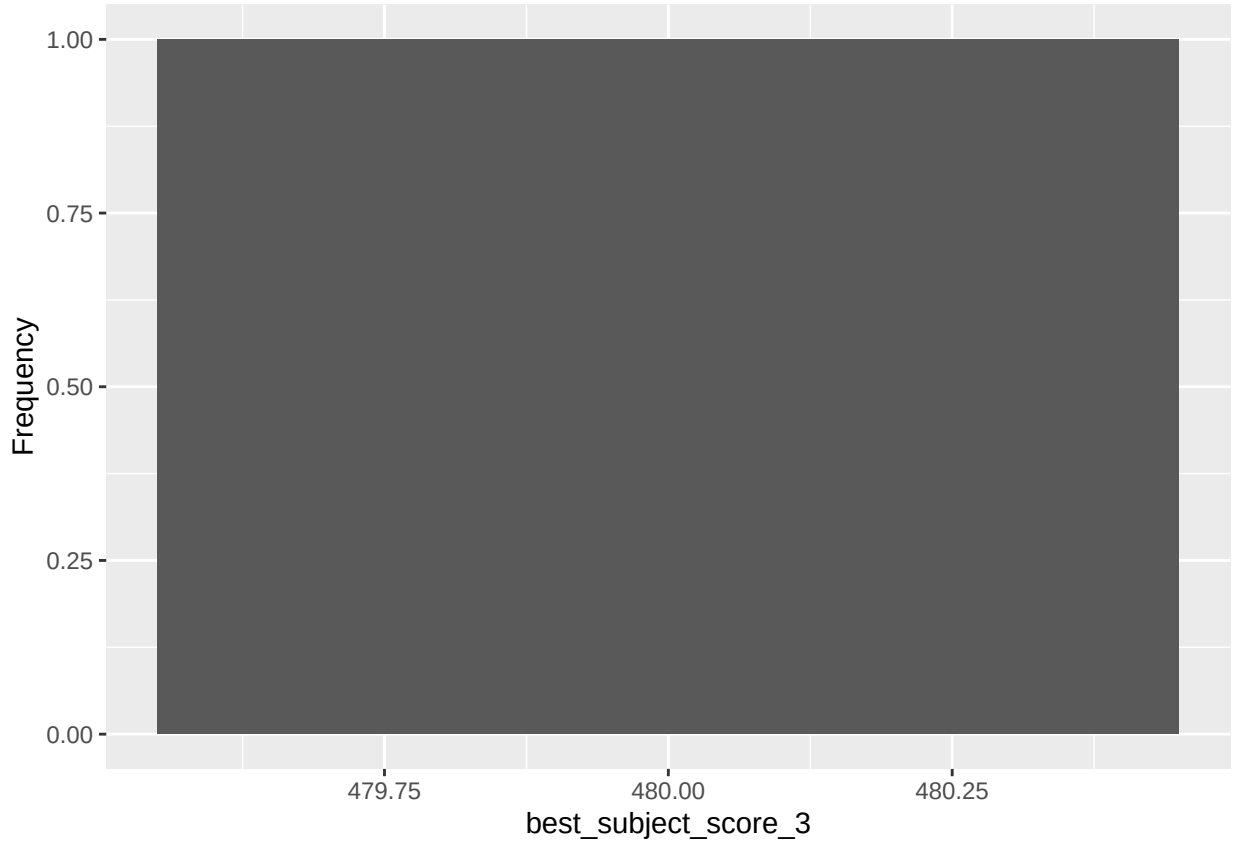

```
[1] ____________________________
[1] Variable: best_subject_score_2, type: integer
[1] Values (51 unique): NA, 590, 600, 730, 540, ...
[1] Missing: 67.6%
  Group.1 best_subject_score_2
1     F08          0.996305954
2     F09          0.997773380
3     F10          0.998412698
4     F11          0.002539559
5     F12          0.546367395
6     F13          0.609639441
7     F14          0.651128376
8     F15          0.700609225
9     F16          0.712741665

Warning: Removed 31350 rows containing non-finite values (‘stat_count()‘).
```



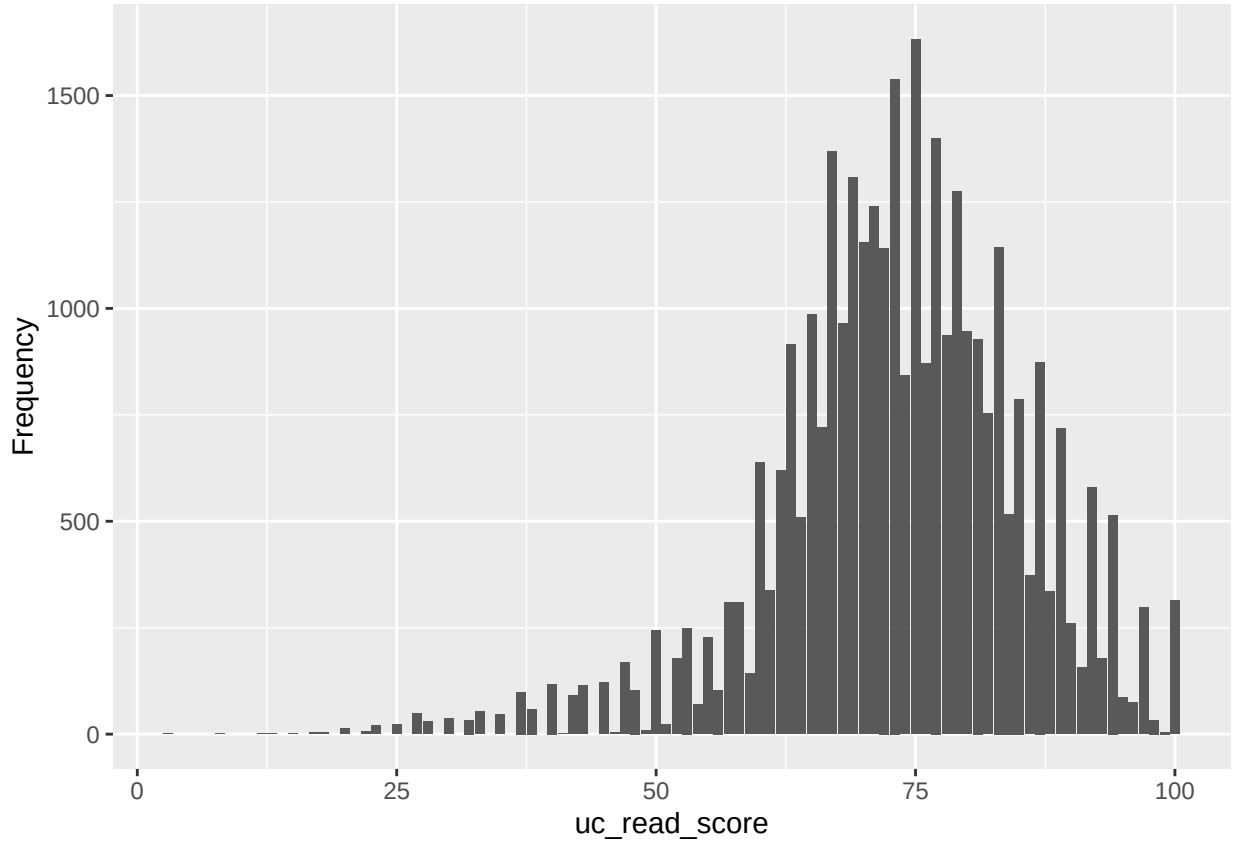

```
[1] ____________________________
[1] Variable: best_subject_code_3, type: character
[1] Values (2 unique): NA, AH
[1] Missing: 100%
```



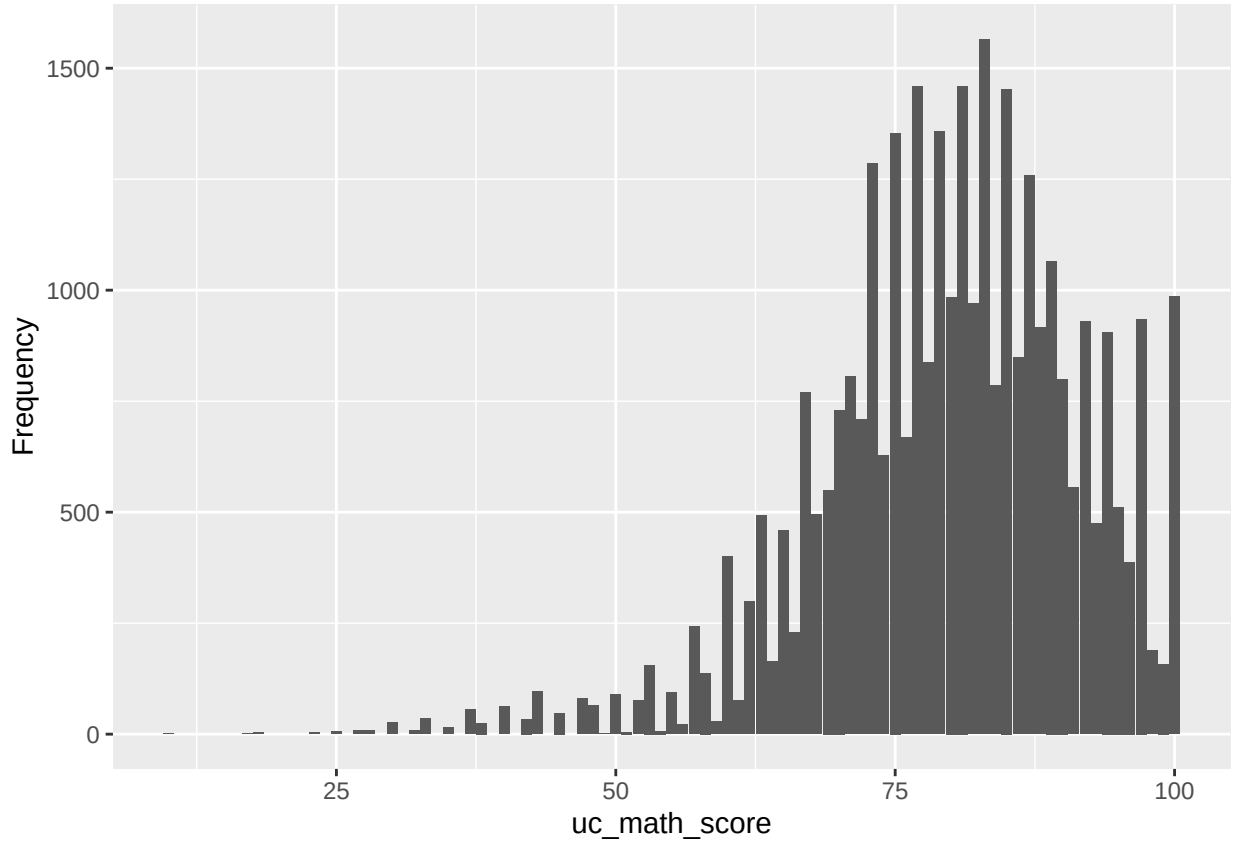

```
[1] ____________________________
[1] Variable: best_subject_score_3, type: integer
[1] Values (2 unique): NA, 480
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```




```
[1] ____________________________
[1] Variable: uc_read_score, type: numeric
[1] Values (80 unique): NA, 77, 79, 75, 71, ...
[1] Missing: 28.2%
  Group.1 uc_read_score
1     F08  0.9963059539
2     F09  0.9977733795
3     F10  0.9981859410
4     F11  0.0023442079
5     F12  0.0013782241
6     F13  0.0011037528
7     F14  0.0009248983
8     F15  0.0010443864
9     F16  0.0015223017

Warning: Removed 13066 rows containing non-finite values (‘stat_count()‘).
```

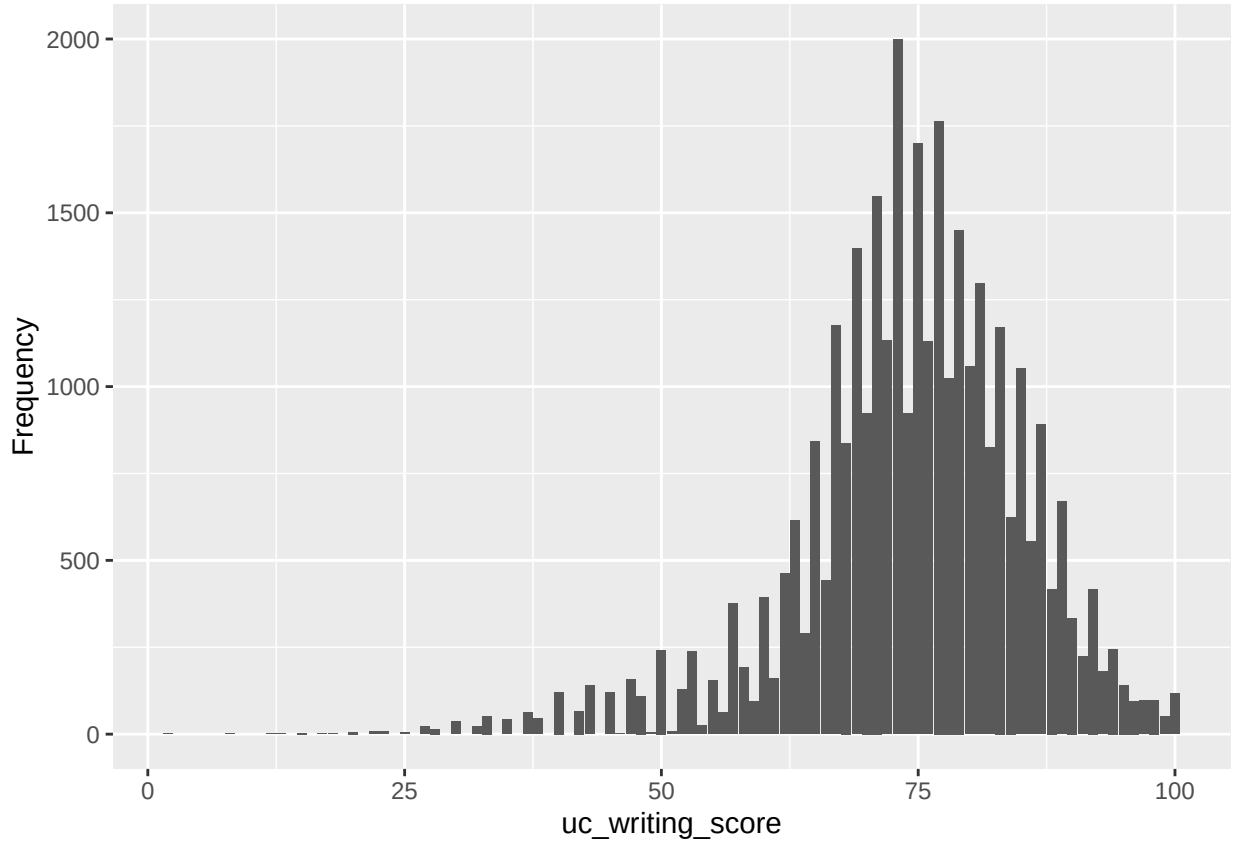



```
[1] ____________________________
[1] Variable: uc_math_score, type: numeric
[1] Values (72 unique): NA, 81, 74, 65, 72, ...
[1] Missing: 28.1%
  Group.1 uc_math_score
1     F08  0.9963059539
2     F09  0.9977733795
3     F10  0.9981859410
4     F11  0.0015628052
5     F12  0.0013782241
6     F13  0.0011037528
7     F14  0.0009248983
8     F15  0.0010443864
9     F16  0.0015223017

Warning: Removed 13062 rows containing non-finite values (‘stat_count()‘).
```



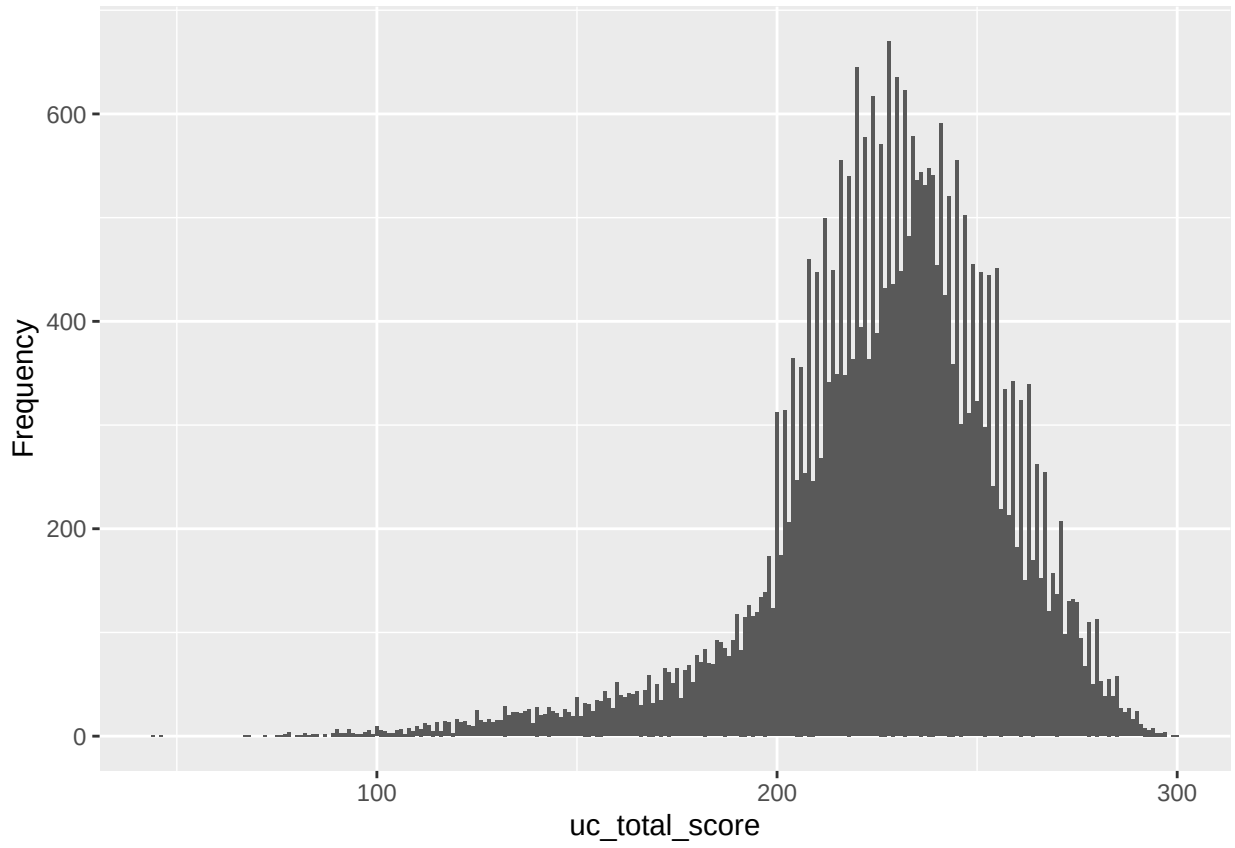

```
[1] ____________________________
[1] Variable: uc_writing_score, type: numeric
[1] Values (79 unique): NA, 77, 76, 73, 72, ...
[1] Missing: 28.2%
  Group.1 uc_writing_score
1     F08     0.9963059539
2     F09     0.9977733795
3     F10     0.9981859410
4     F11     0.0021488572
5     F12     0.0015751132
6     F13     0.0014716703
7     F14     0.0009248983
8     F15     0.0010443864
9     F16     0.0015223017
```

```
Warning: Removed 13068 rows containing non-finite values (‘stat_count()‘).
```

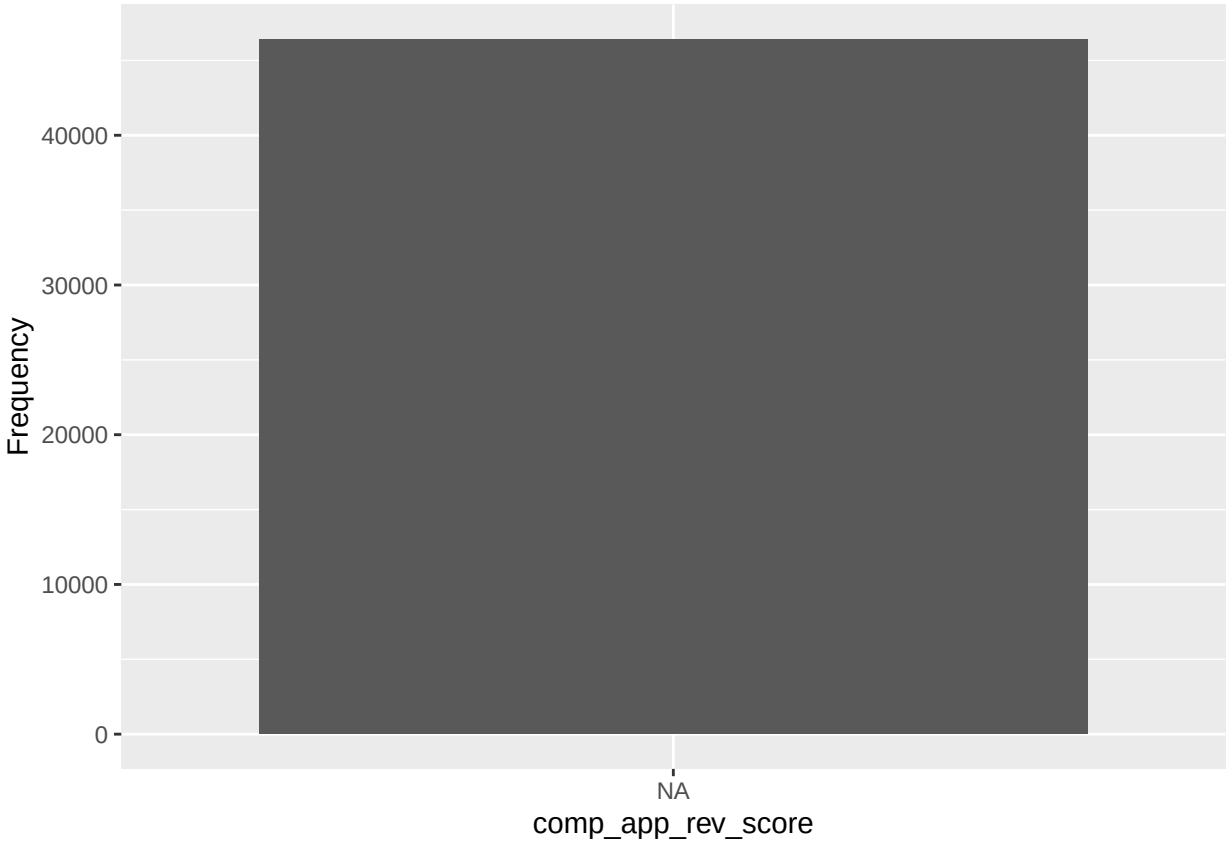



```
[1] ___________________________
[1] Variable: uc_total_score, type: numeric
[1] Values (228 unique): NA, 232, 229, 228, 208, ...
[1] Missing: 28.1%
  Group.1 uc_total_score
1     F08   0.9963059539
2     F09   0.9977733795
3     F10   0.9981859410
4     F11   0.0015628052
5     F12   0.0013782241
6     F13   0.0011037528
7     F14   0.0009248983
8     F15   0.0010443864
9     F16   0.0015223017
```

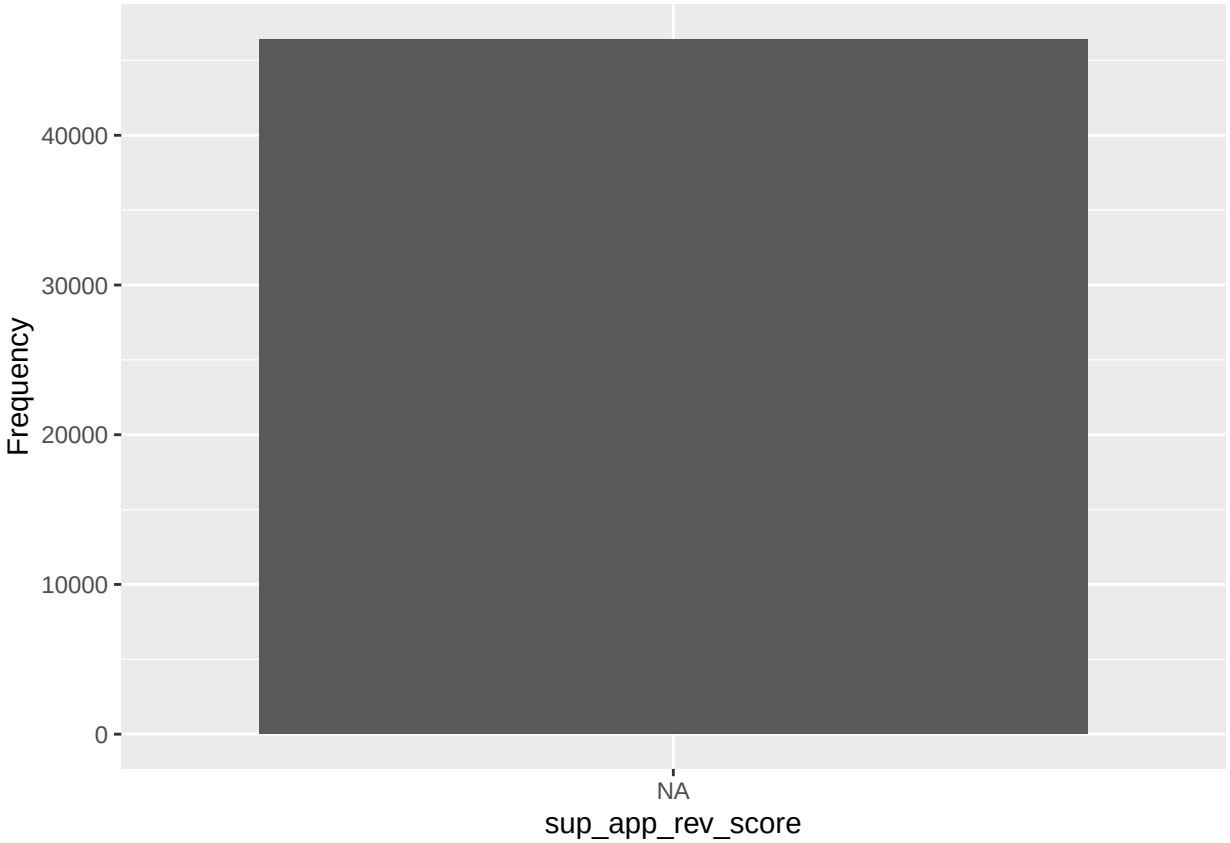
```
Warning: Removed 13062 rows containing non-finite values (‘stat_count()‘).
```



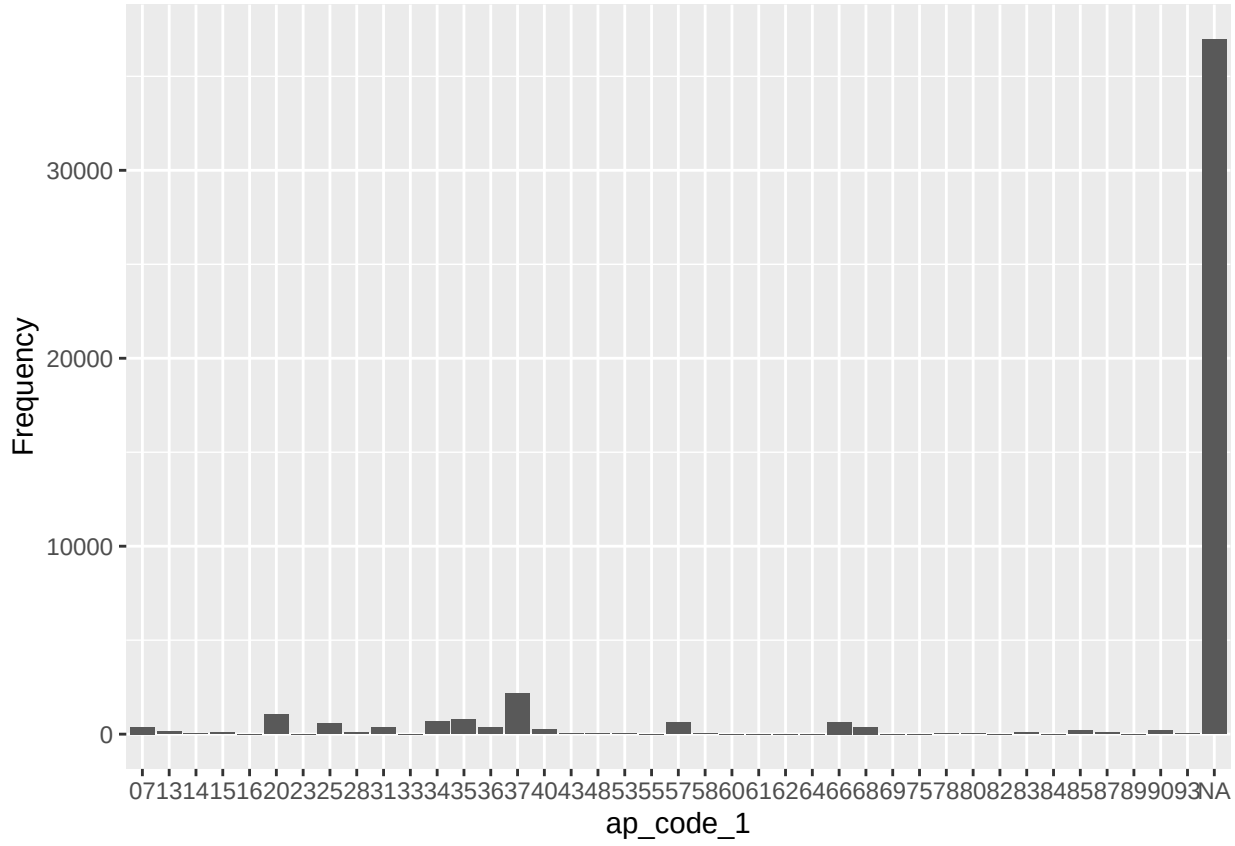

```
[1] ____________________________
[1] Variable: comp_app_rev_score, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
```



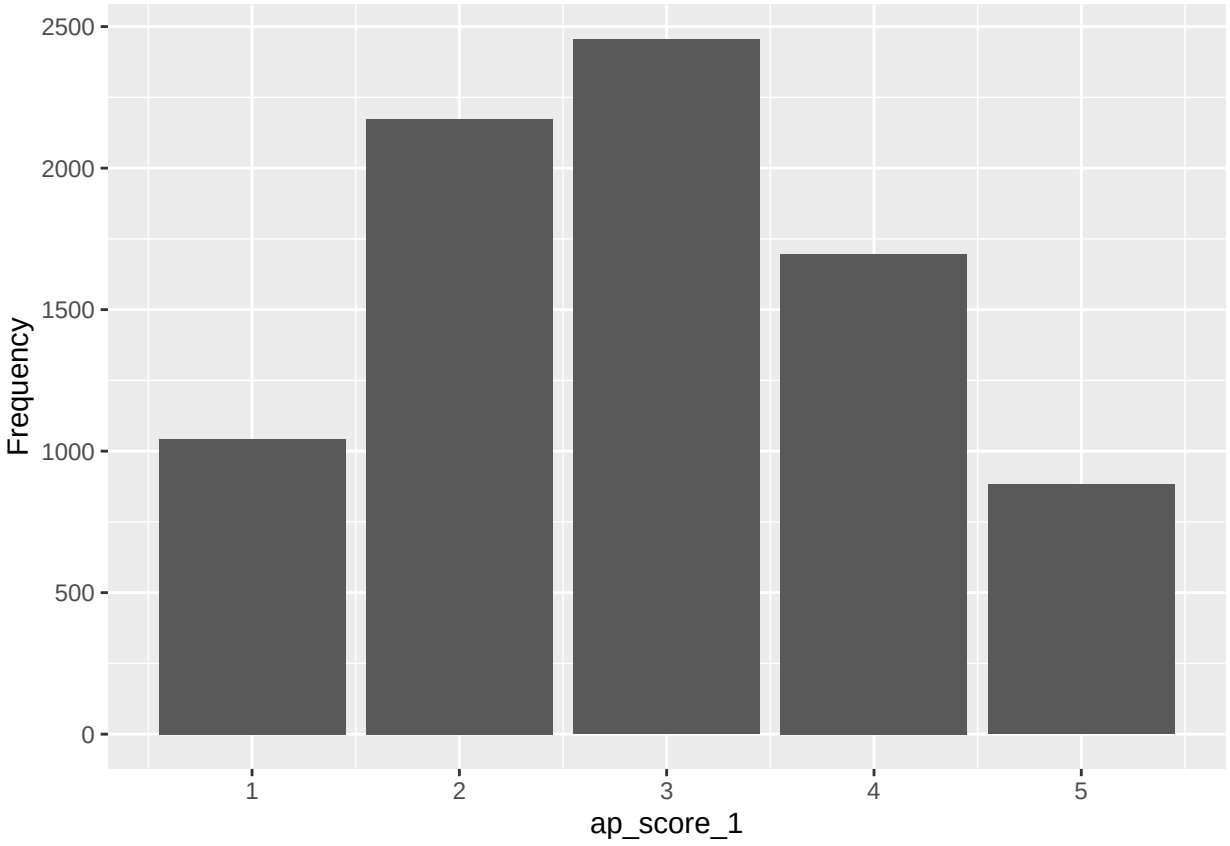

```
[1] ____________________________
[1] Variable: sup_app_rev_score, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
```



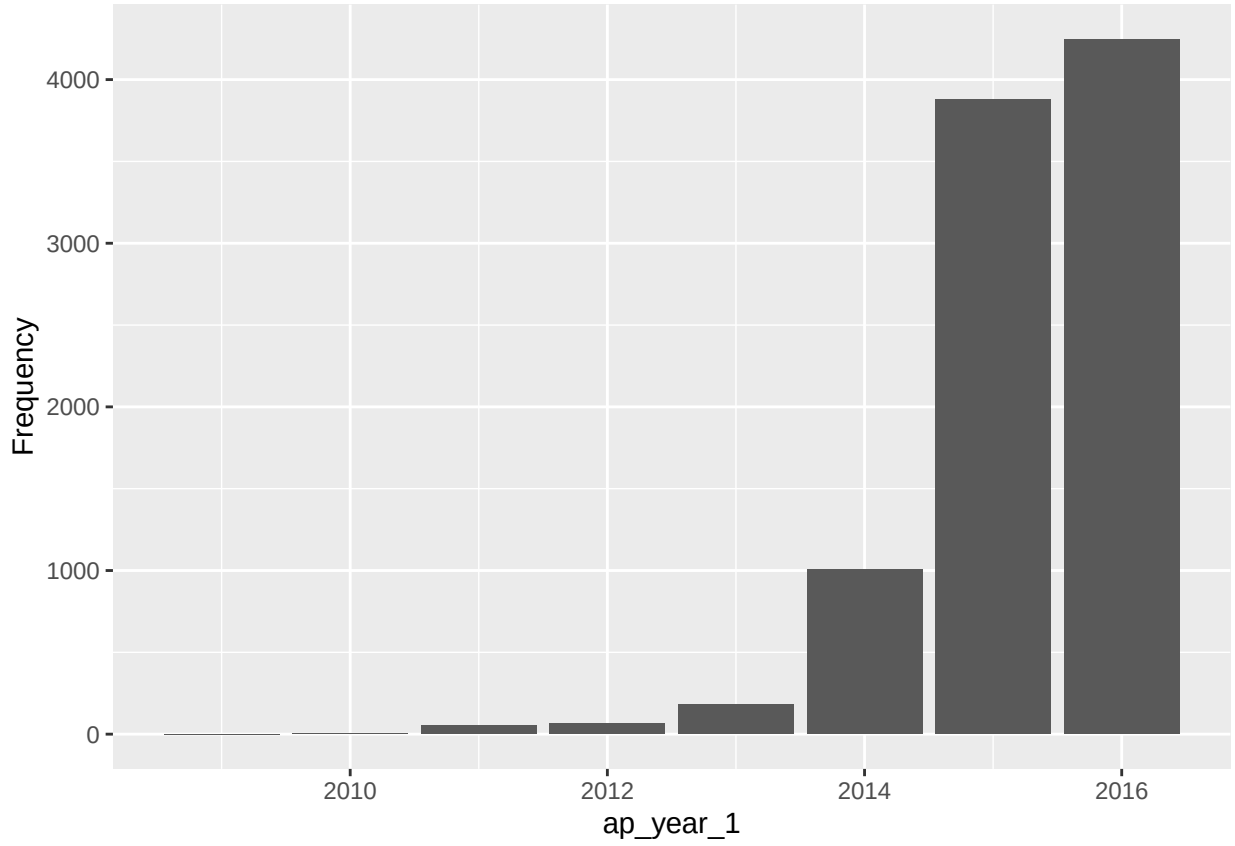

```
[1] ____________________________
[1] Variable: ap_code_1, type: character
[1] Values (41 unique): NA, 25, 14, 20, 35, ...
[1] Missing: 79.7%
  Group.1 ap_code_1
1     F08 1.0000000
2     F09 1.0000000
3     F10 1.0000000
4     F11 0.9902325
5     F12 0.9893680
6     F13 0.9760854
7     F14 0.8742138
8     F15 0.3536989
9     F16 0.2673162
```



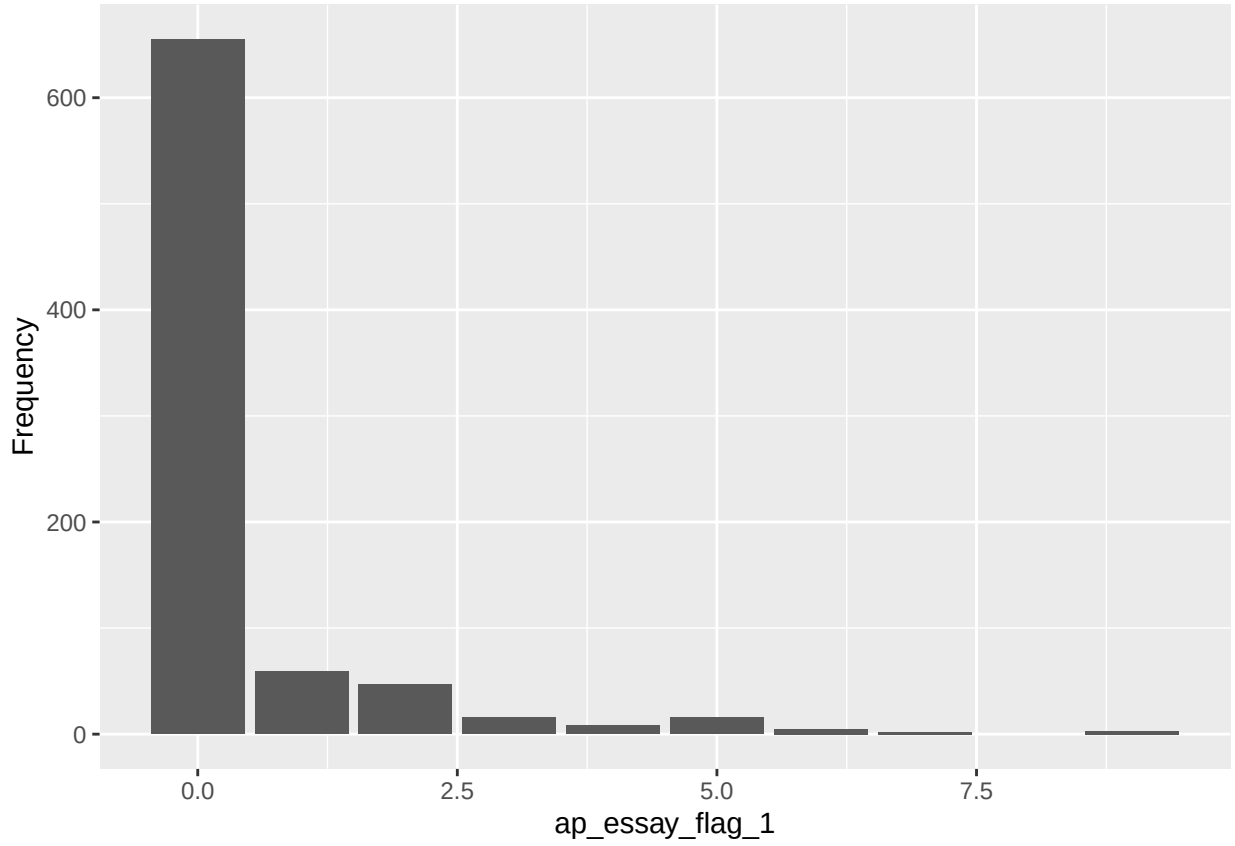

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_1, type: numeric
[1] Values (6 unique): NA, 2, 5, 3, 4, ...
[1] Missing: 82.2%
  Group.1 ap_score_1
1     F08  1.0000000
2     F09  1.0000000
3     F10  1.0000000
4     F11  0.9908185
5     F12  0.9907462
6     F13  0.9792127
7     F14  0.8843877
8     F15  0.4259356
9     F16  0.3728117

Warning: Removed 38158 rows containing non-finite values (‘stat_count()‘).
```



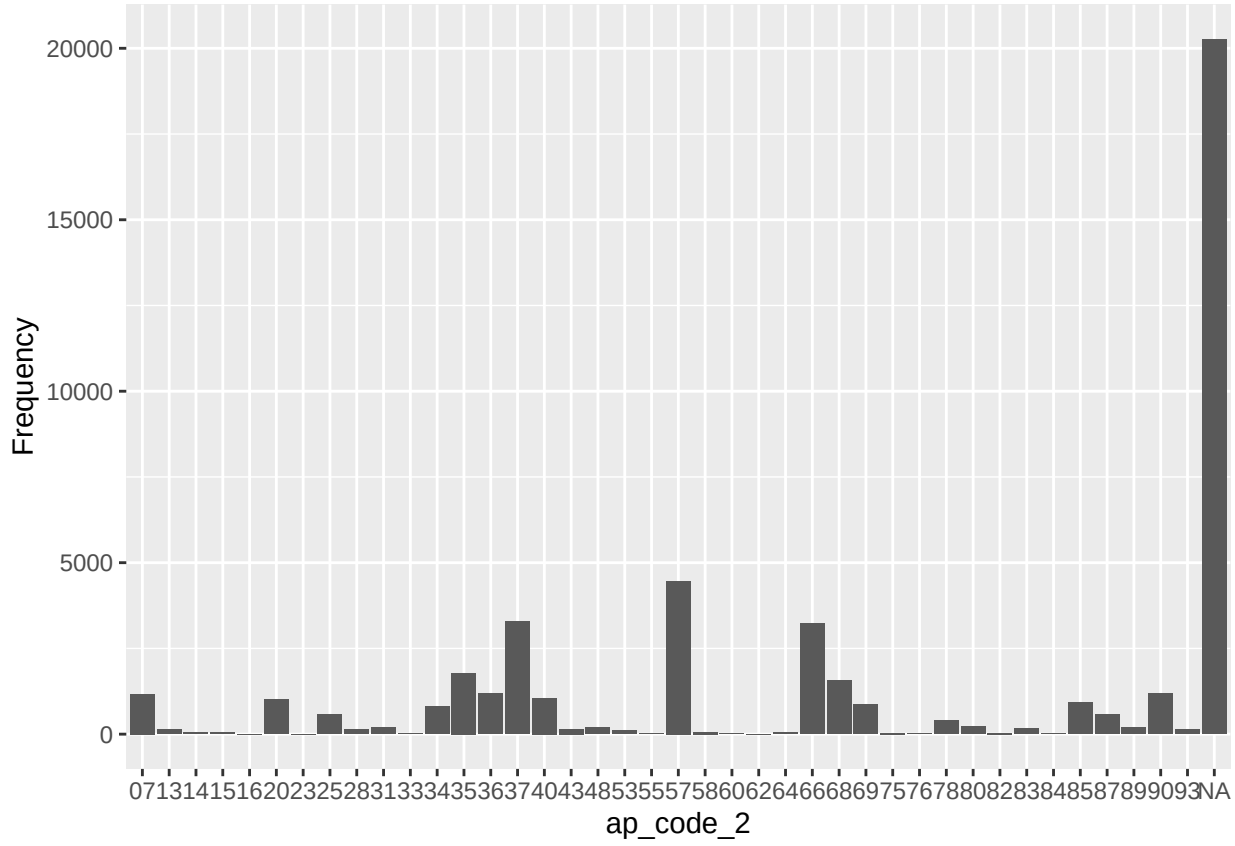

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_1, type: integer
[1] Values (9 unique): NA, 2015, 2016, 2014, 2013, ...
[1] Missing: 79.7%
  Group.1 ap_year_1
1     F08 1.0000000
2     F09 1.0000000
3     F10 1.0000000
4     F11 0.9902325
5     F12 0.9893680
6     F13 0.9760854
7     F14 0.8742138
8     F15 0.3536989
9     F16 0.2673162

Warning: Removed 36968 rows containing non-finite values (‘stat_count()‘).
```



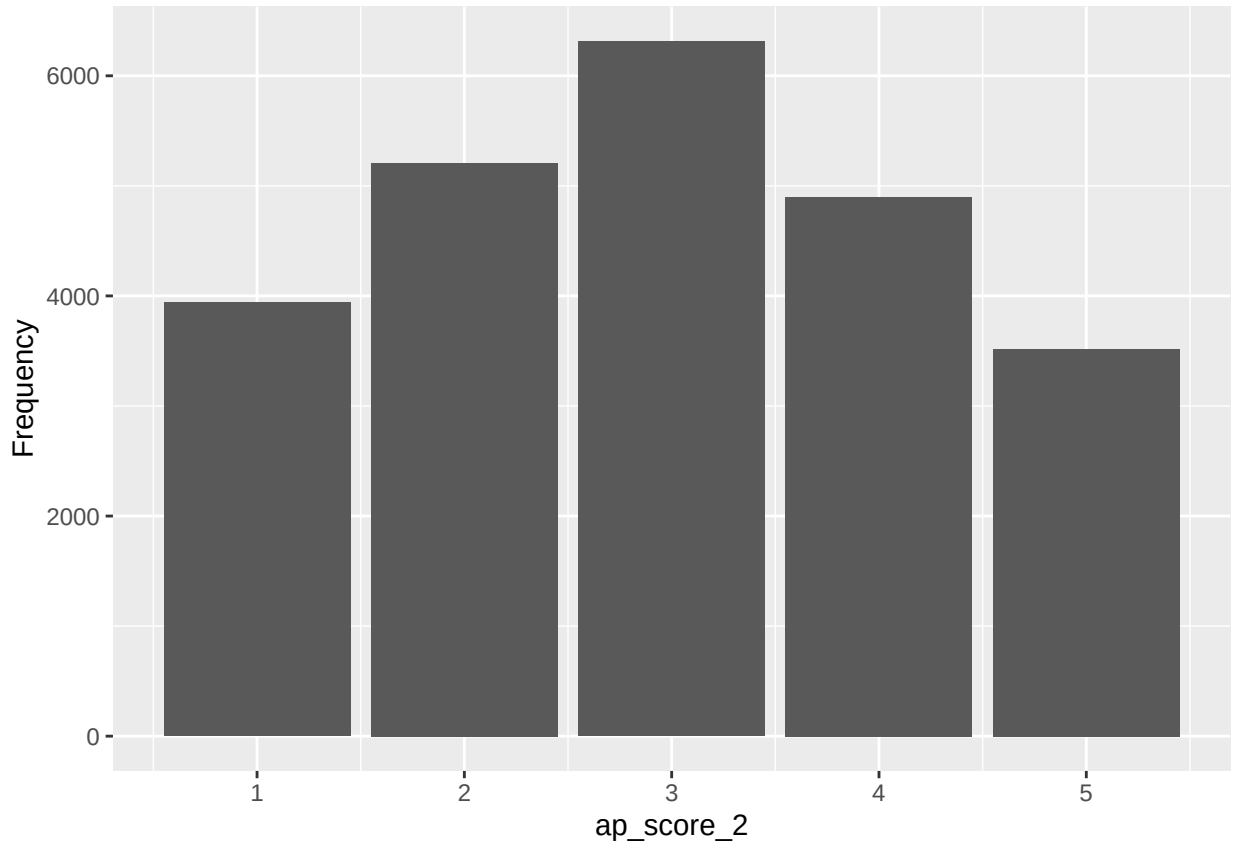

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_1, type: numeric
[1] Values (10 unique): NA, 0, 2, 1, 4, ...
[1] Missing: 98.3%
[1] Most missing: F08 100%, Least missing: F14 88.8%

Warning: Removed 45597 rows containing non-finite values (‘stat_count()‘).
```



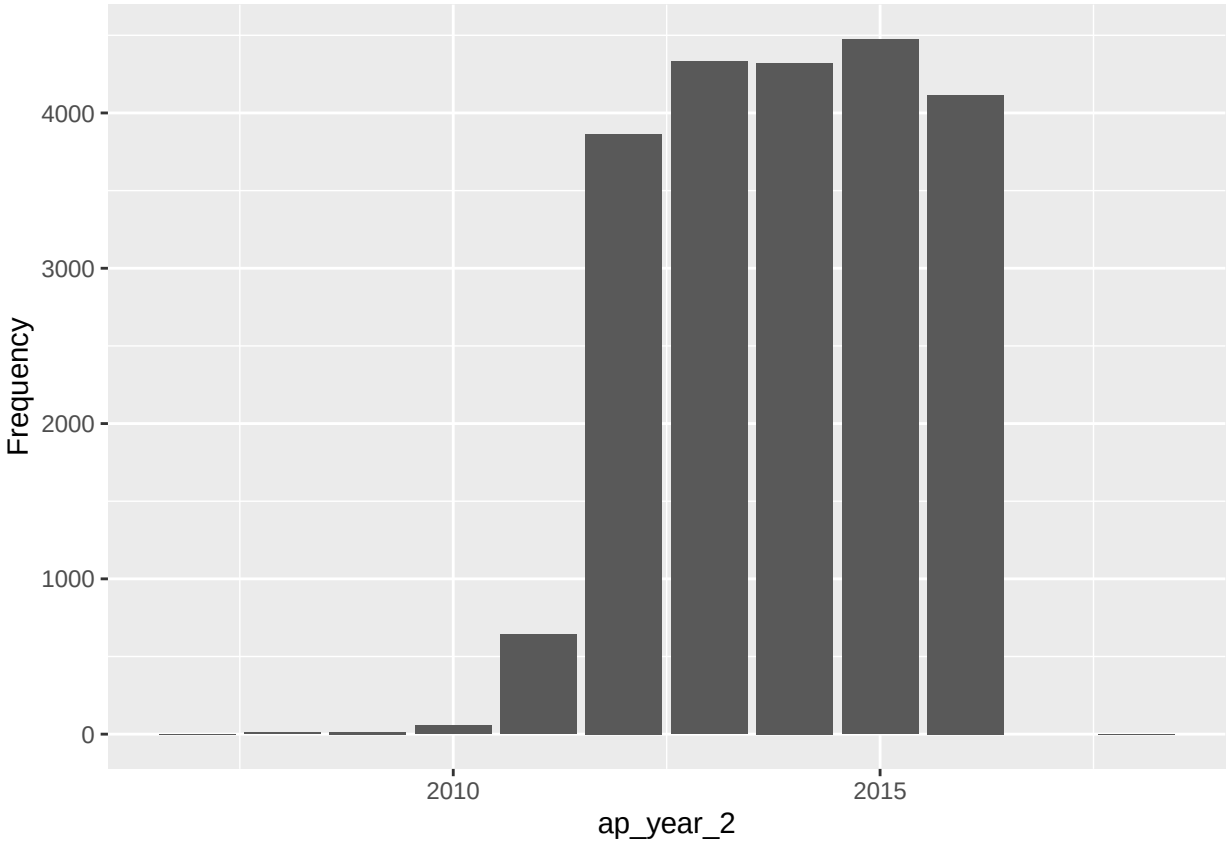

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_2, type: character
[1] Values (41 unique): NA, 35, 07, 66, 90, ...
[1] Missing: 43.7%
  Group.1 ap_code_2
1     F08 0.9967405
2     F09 0.9975260
3     F10 0.9981859
4     F11 0.1574526
5     F12 0.2327230
6     F13 0.2058499
7     F14 0.2230855
8     F15 0.2382942
9     F16 0.2371746
```



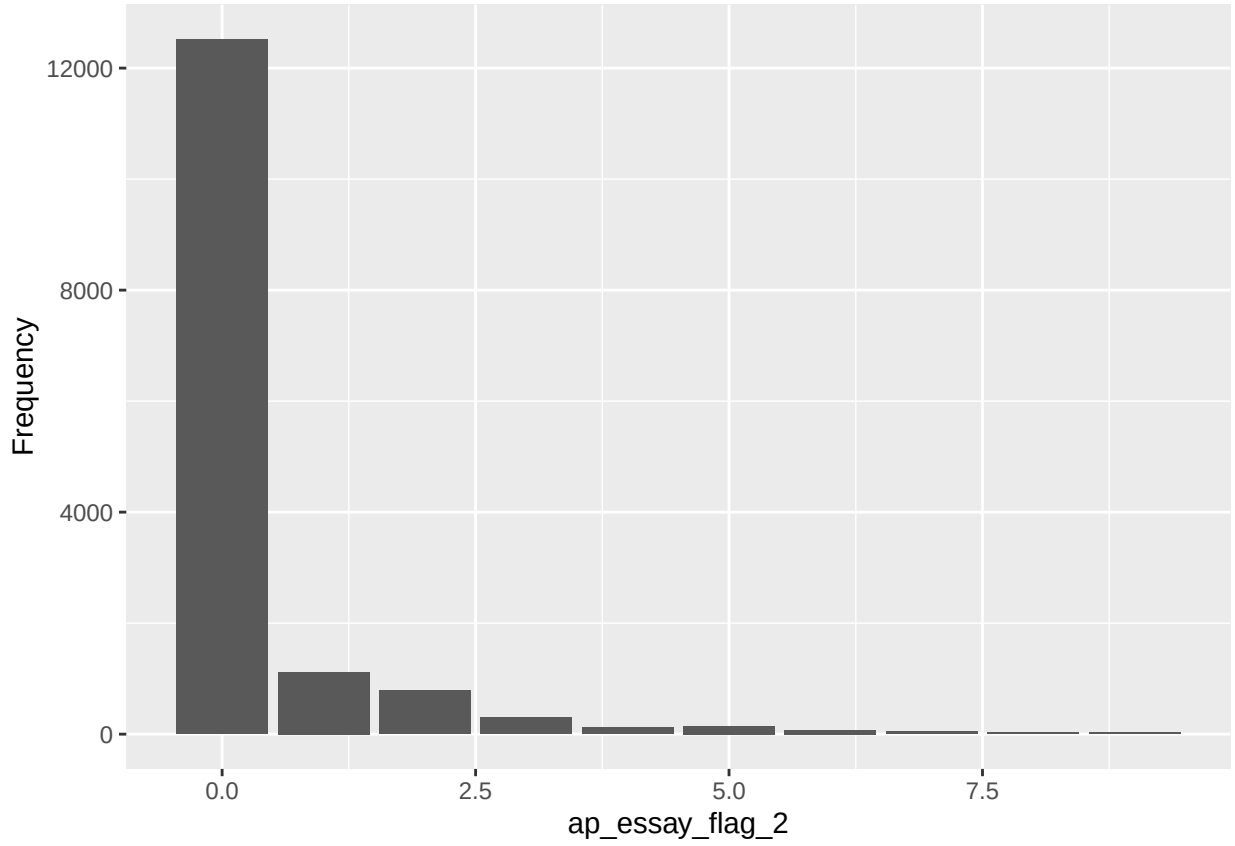

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_2, type: numeric
[1] Values (6 unique): NA, 1, 3, 4, 5, ...
[1] Missing: 48.5%
  Group.1 ap_score_2
1     F08  0.9969578
2     F09  0.9975260
3     F10  0.9984127
4     F11  0.2236765
5     F12  0.3041937
6     F13  0.2433775
7     F14  0.2719201
8     F15  0.3117493
9     F16  0.3397777

Warning: Removed 22529 rows containing non-finite values (‘stat_count()‘).
```



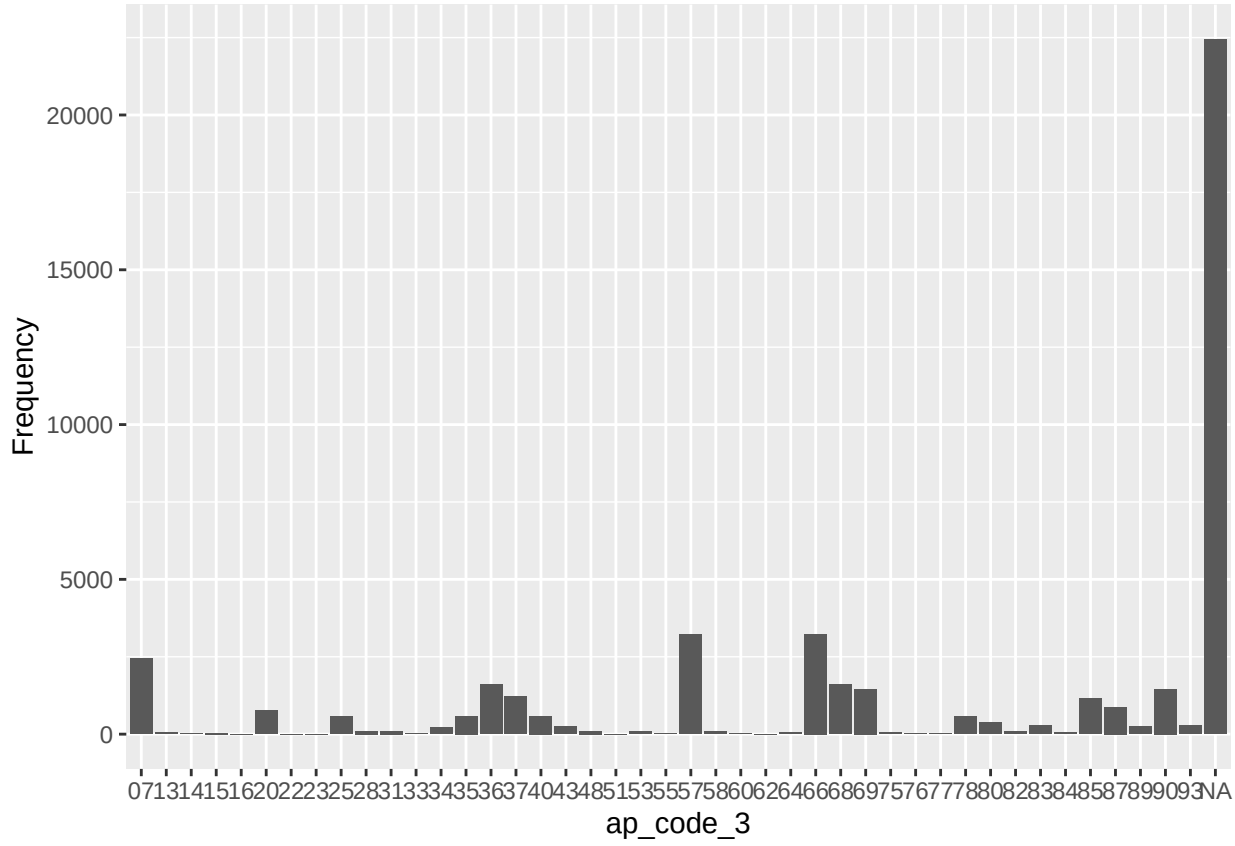

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_2, type: integer
[1] Values (12 unique): NA, 2015, 2016, 2014, 2012, ...
[1] Missing: 53%
  Group.1 ap_year_2
1     F08 0.9976097
2     F09 0.9980208
3     F10 0.9981859
4     F11 0.9988279
5     F12 0.2331168
6     F13 0.2060338
7     F14 0.2230855
8     F15 0.2382942
9     F16 0.2371746

Warning: Removed 24577 rows containing non-finite values (‘stat_count()‘).
```



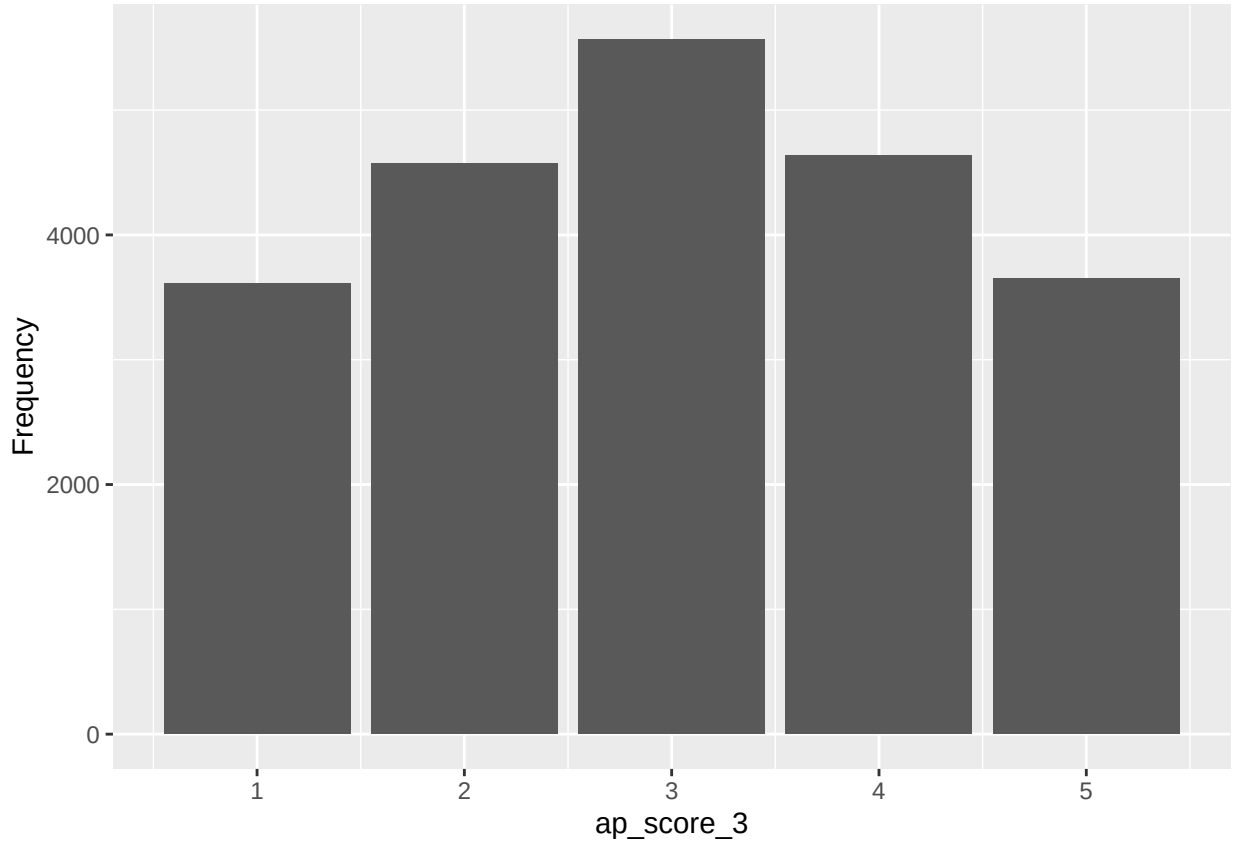

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_2, type: numeric
[1] Values (11 unique): NA, 0, 1, 2, 3, ...
[1] Missing: 67.3%
  Group.1 ap_essay_flag_2
1     F08       0.9978270
2     F09       0.9982682
3     F10       0.9988662
4     F11       0.2510256
5     F12       0.3297893
6     F13       0.2545990
7     F14       0.2887532
8     F15       0.9998259
9     F16       0.9995433

Warning: Removed 31247 rows containing non-finite values (‘stat_count()‘).
```



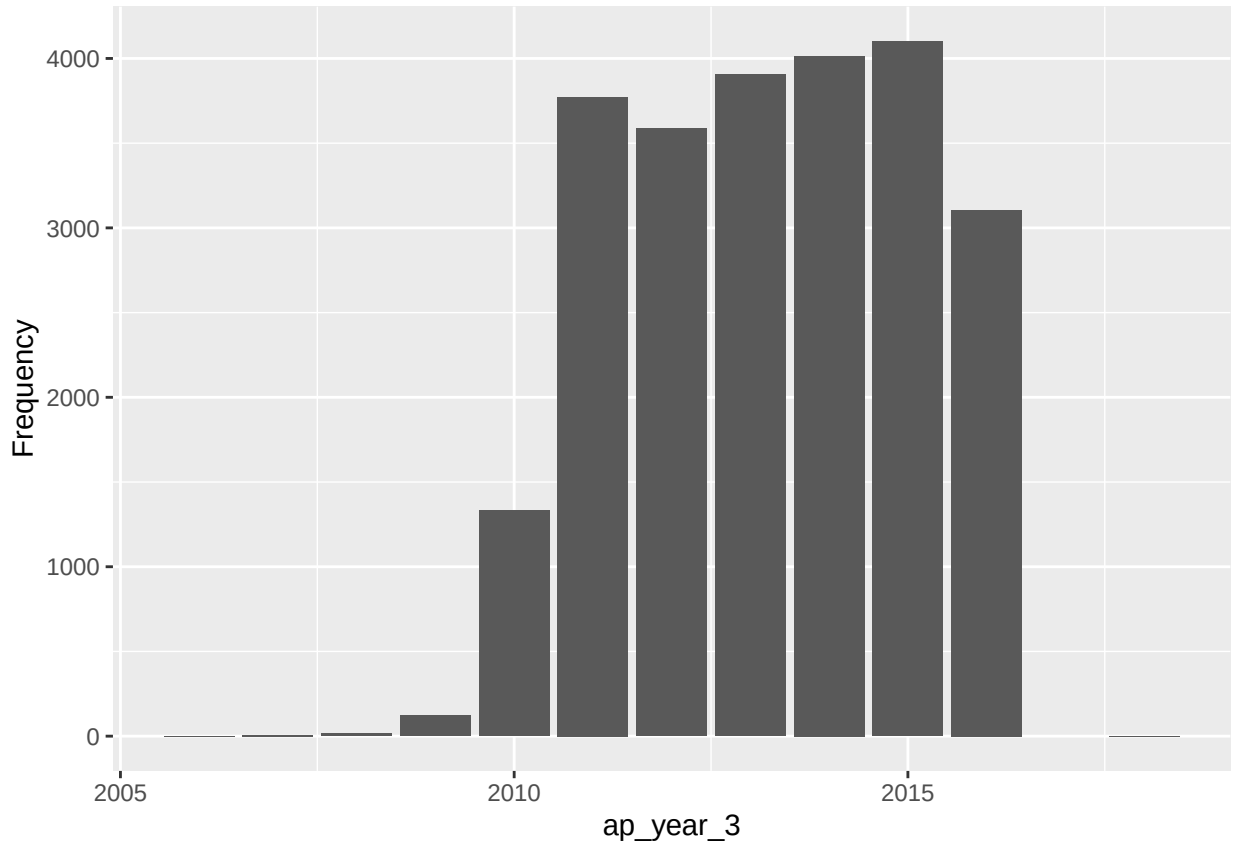

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_3, type: character
[1] Values (44 unique): NA, 37, 25, 36, 93, ...
[1] Missing: 48.4%
  Group.1 ap_code_3
1     F08 0.9969578
2     F09 0.9975260
3     F10 0.9986395
4     F11 0.2410627
5     F12 0.3089191
6     F13 0.2709713
7     F14 0.2772845
8     F15 0.2988686
9     F16 0.2933475
```



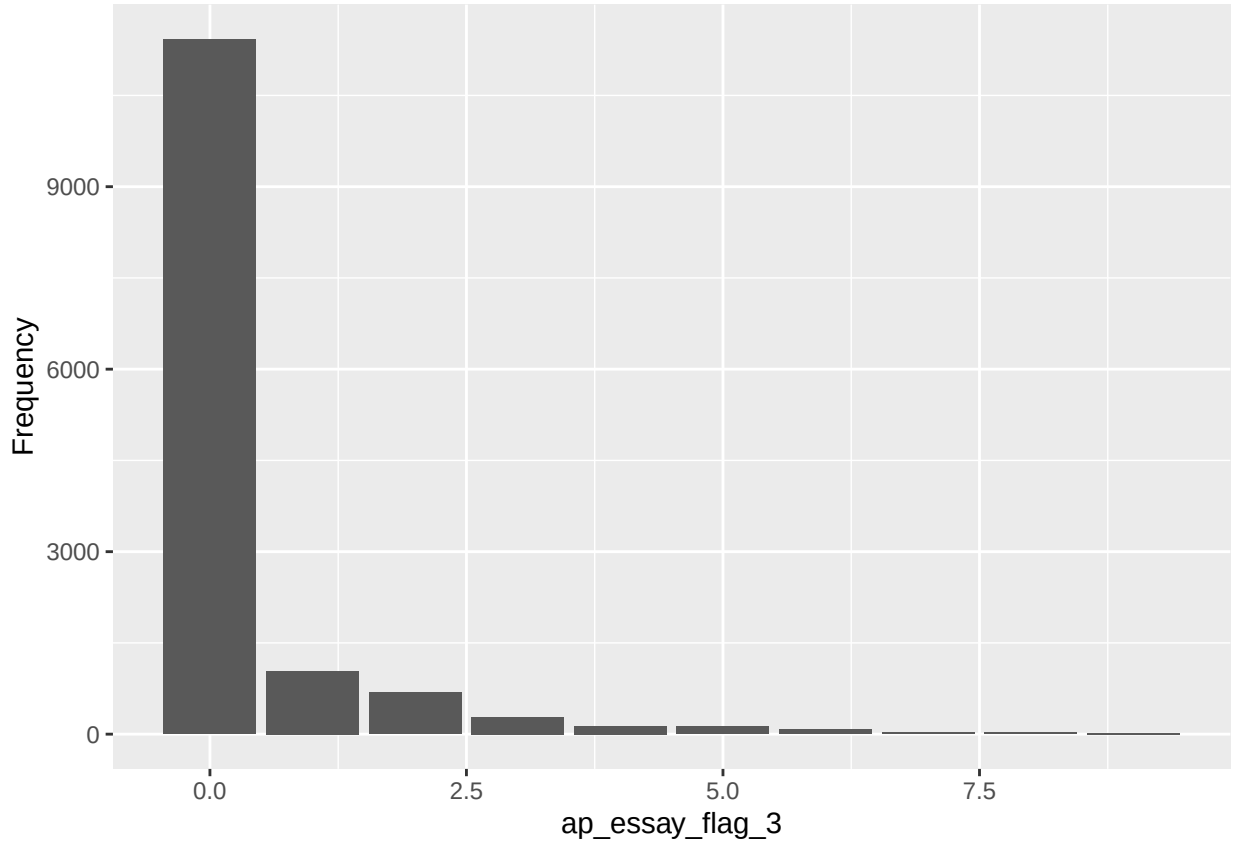

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_3, type: numeric
[1] Values (6 unique): NA, 3, 2, 4, 5, ...
[1] Missing: 52.5%
  Group.1 ap_score_3
1     F08  0.9969578
2     F09  0.9977734
3     F10  0.9988662
4     F11  0.3024028
5     F12  0.3634574
6     F13  0.3101545
7     F14  0.3214946
8     F15  0.3566580
9     F16  0.3775308

Warning: Removed 24373 rows containing non-finite values (‘stat_count()‘).
```



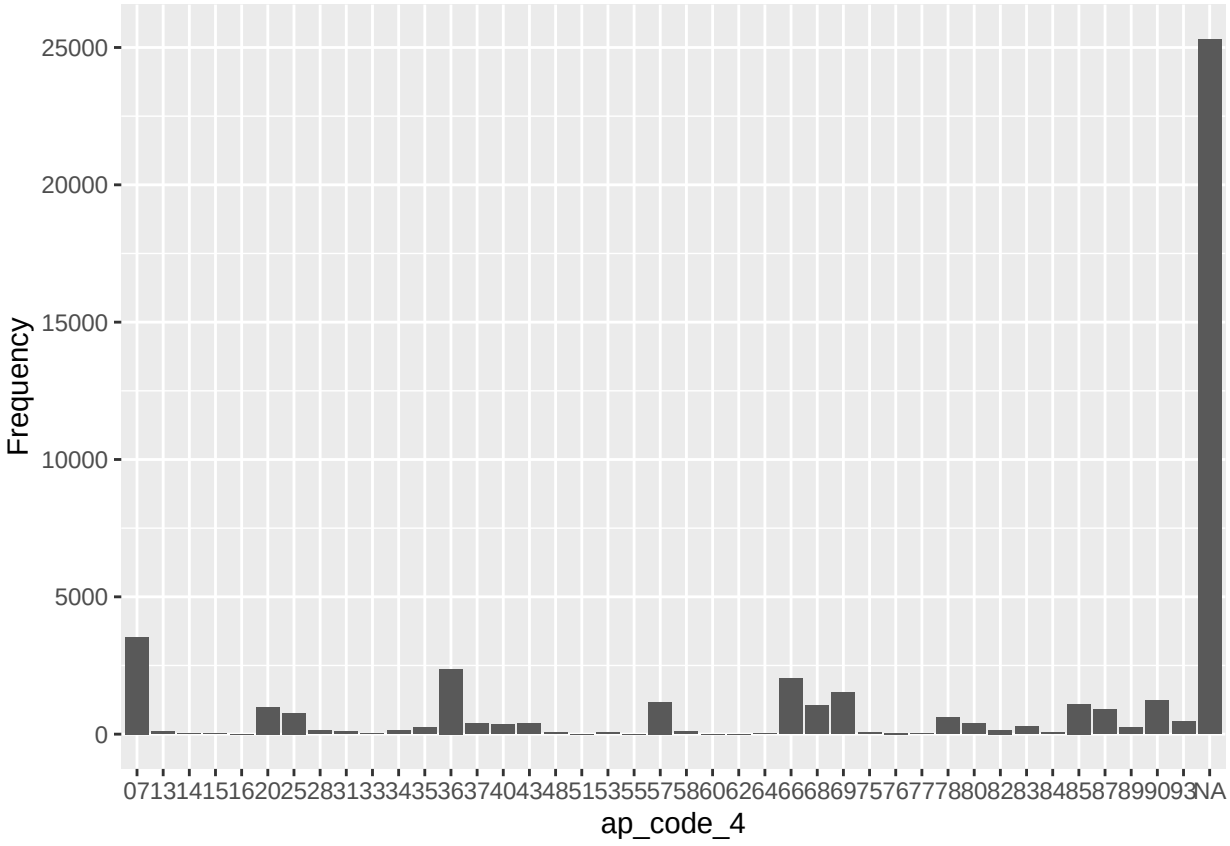

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_3, type: integer
[1] Values (13 unique): NA, 2015, 2014, 2013, 2016, ...
[1] Missing: 48.4%
  Group.1 ap_year_3
1     F08 0.9969578
2     F09 0.9975260
3     F10 0.9986395
4     F11 0.2410627
5     F12 0.3089191
6     F13 0.2709713
7     F14 0.2772845
8     F15 0.2988686
9     F16 0.2933475

Warning: Removed 22443 rows containing non-finite values (‘stat_count()‘).
```



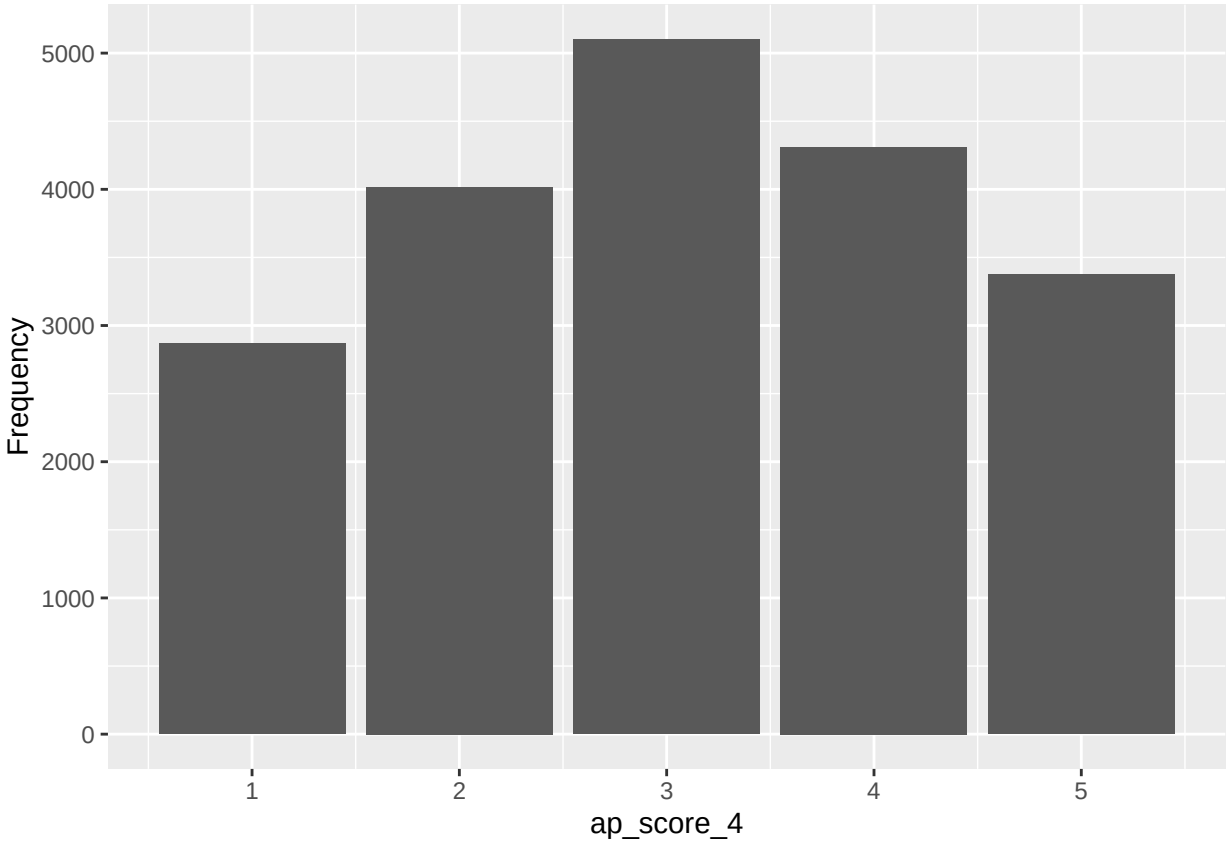

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_3, type: numeric
[1] Values (11 unique): NA, 0, 4, 1, 2, ...
[1] Missing: 70.2%
  Group.1 ap_essay_flag_3
1     F08       0.9980443
2     F09       0.9982682
3     F10       0.9990930
4     F11       0.3272123
5     F12       0.3884623
6     F13       0.3228477
7     F14       0.3377728
8     F15       0.9998259
9     F16       0.9995433

Warning: Removed 32573 rows containing non-finite values (‘stat_count()‘).
```



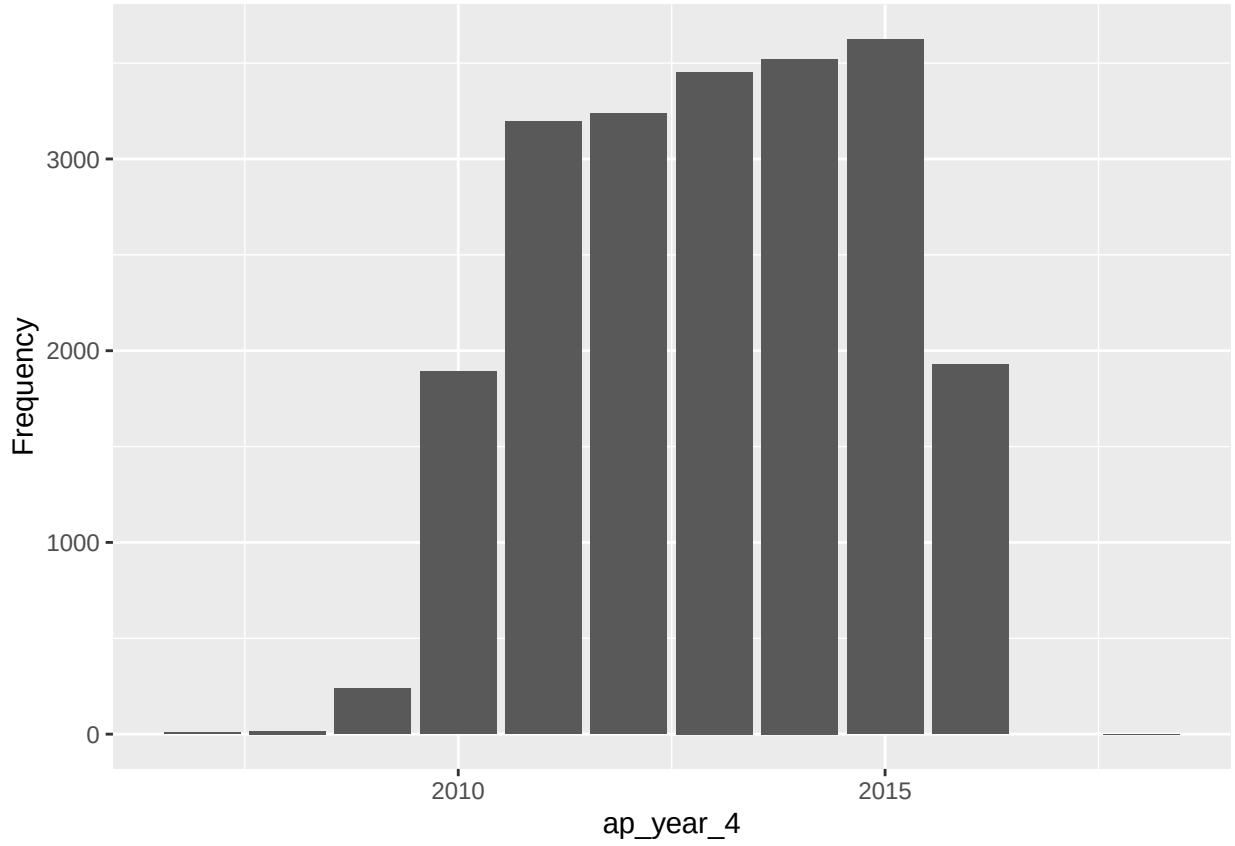

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_4, type: character
[1] Values (42 unique): NA, 57, 93, 07, 40, ...
[1] Missing: 54.5%
  Group.1 ap_code_4
1     F08 0.9978270
2     F09 0.9977734
3     F10 0.9988662
4     F11 0.3385427
5     F12 0.4016539
6     F13 0.3600074
7     F14 0.3540511
8     F15 0.3808529
9     F16 0.3705282
```



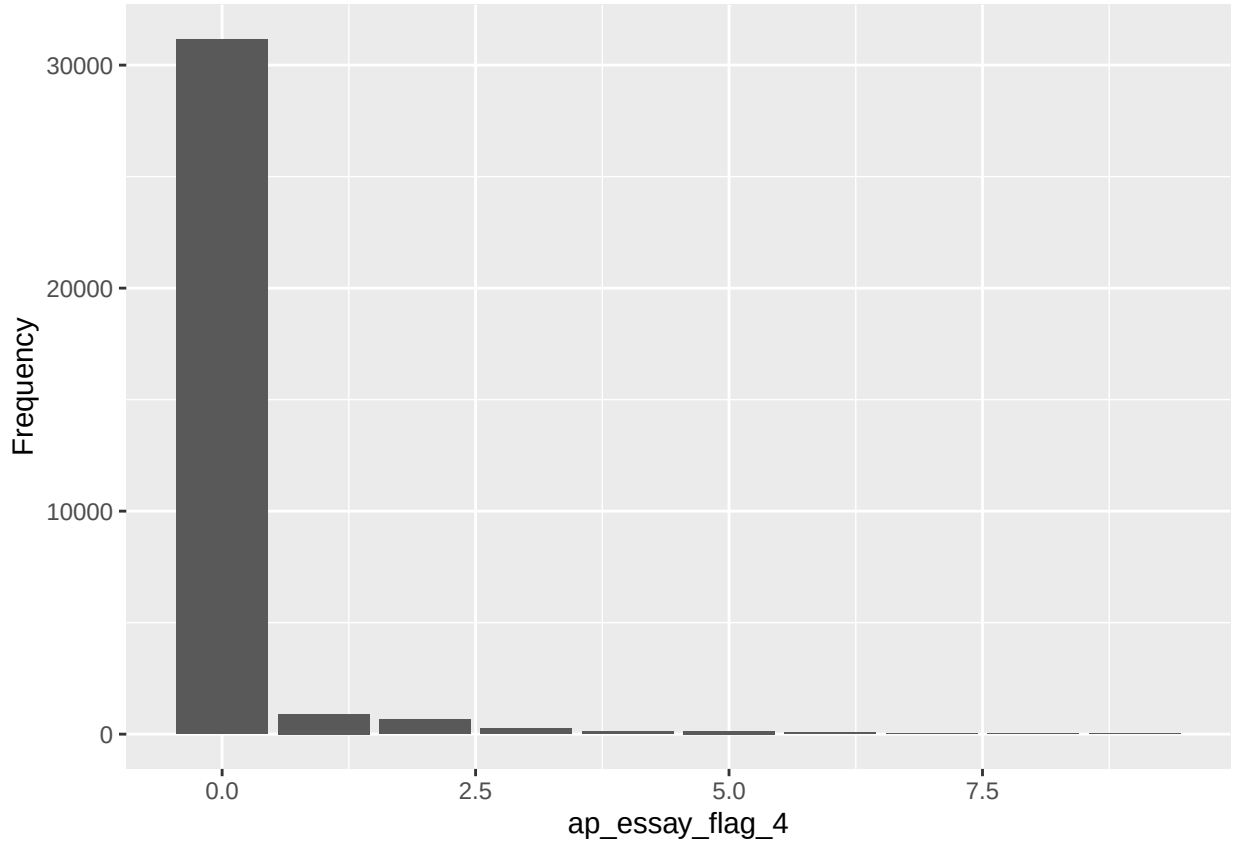

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_4, type: numeric
[1] Values (6 unique): NA, 1, 3, 4, 2, ...
[1] Missing: 57.6%
  Group.1 ap_score_4
1     F08  0.9978270
2     F09  0.9977734
3     F10  0.9990930
4     F11  0.3854268
5     F12  0.4426068
6     F13  0.3938558
7     F14  0.3879023
8     F15  0.4283725
9     F16  0.4239610

Warning: Removed 26736 rows containing non-finite values (‘stat_count()‘).
```



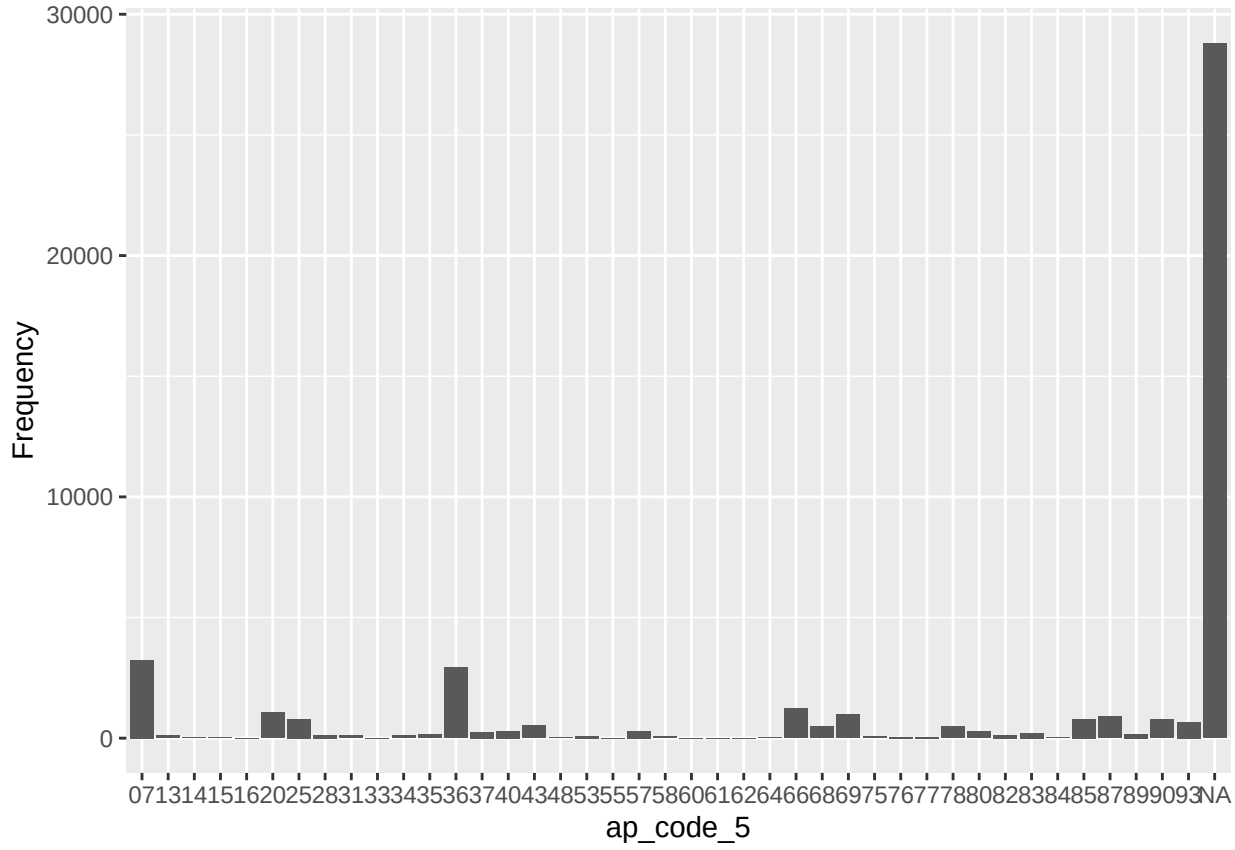

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_4, type: integer
[1] Values (12 unique): NA, 2015, 2014, 2013, 2016, ...
[1] Missing: 54.5%
  Group.1 ap_year_4
1     F08 0.9978270
2     F09 0.9977734
3     F10 0.9988662
4     F11 0.3385427
5     F12 0.4016539
6     F13 0.3600074
7     F14 0.3540511
8     F15 0.3808529
9     F16 0.3705282

Warning: Removed 25296 rows containing non-finite values (‘stat_count()‘).
```



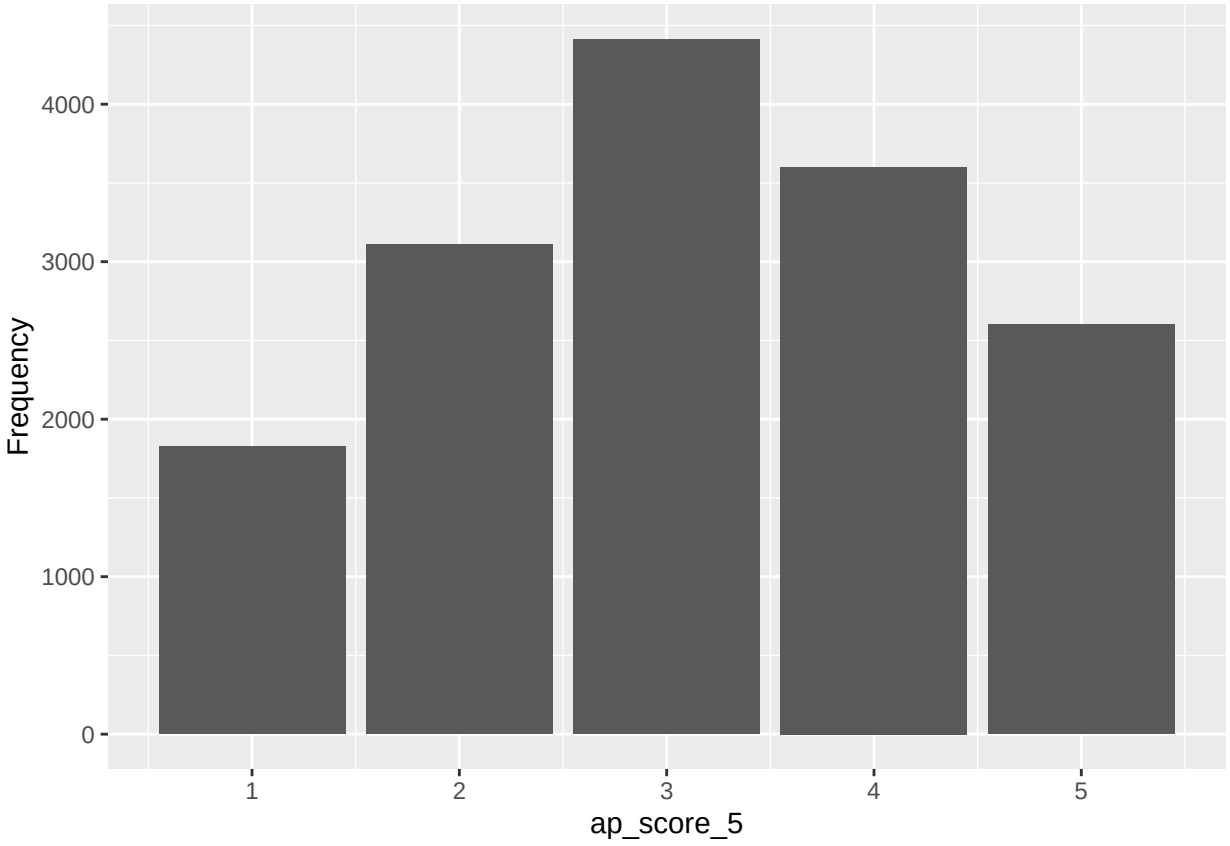

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_4, type: numeric
[1] Values (11 unique): NA, 0, 1, 5, 9, ...
[1] Missing: 28.1%
  Group.1 ap_essay_flag_4
1     F08    0.9952194698
2     F09    0.9970311727
3     F10    0.9968253968
4     F11    0.0001953507
5     F12    0.0003937783
6     F13    0.0003679176
7     F14    0.0001849797
8     F15    0.0006962576
9     F16    0.0012178414

Warning: Removed 13024 rows containing non-finite values (‘stat_count()‘).
```



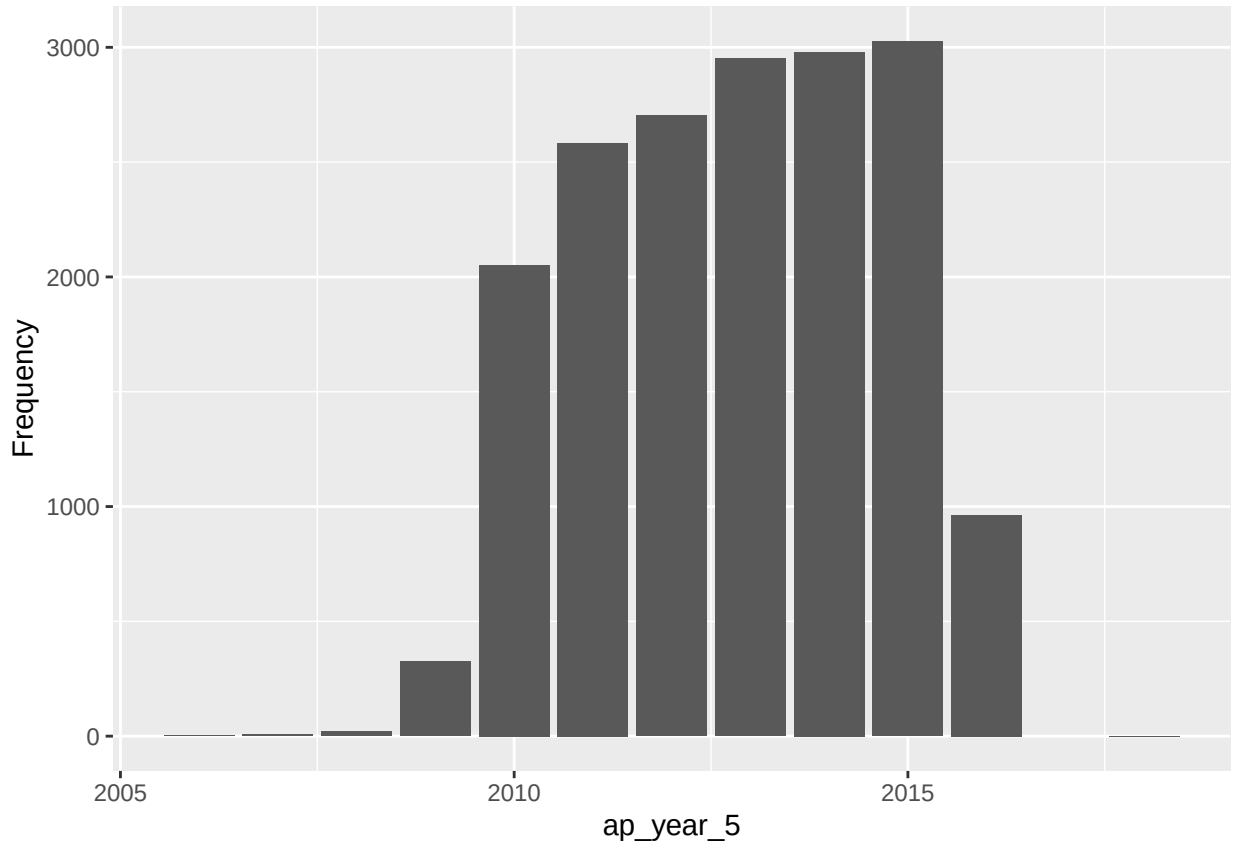

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_5, type: character
[1] Values (42 unique): NA, 66, 87, 83, 07, ...
[1] Missing: 62%
  Group.1 ap_code_5
1     F08 0.9980443
2     F09 0.9982682
3     F10 0.9988662
4     F11 0.4551670
5     F12 0.5146682
6     F13 0.4793966
7     F14 0.4557899
8     F15 0.4694517
9     F16 0.4644543
```



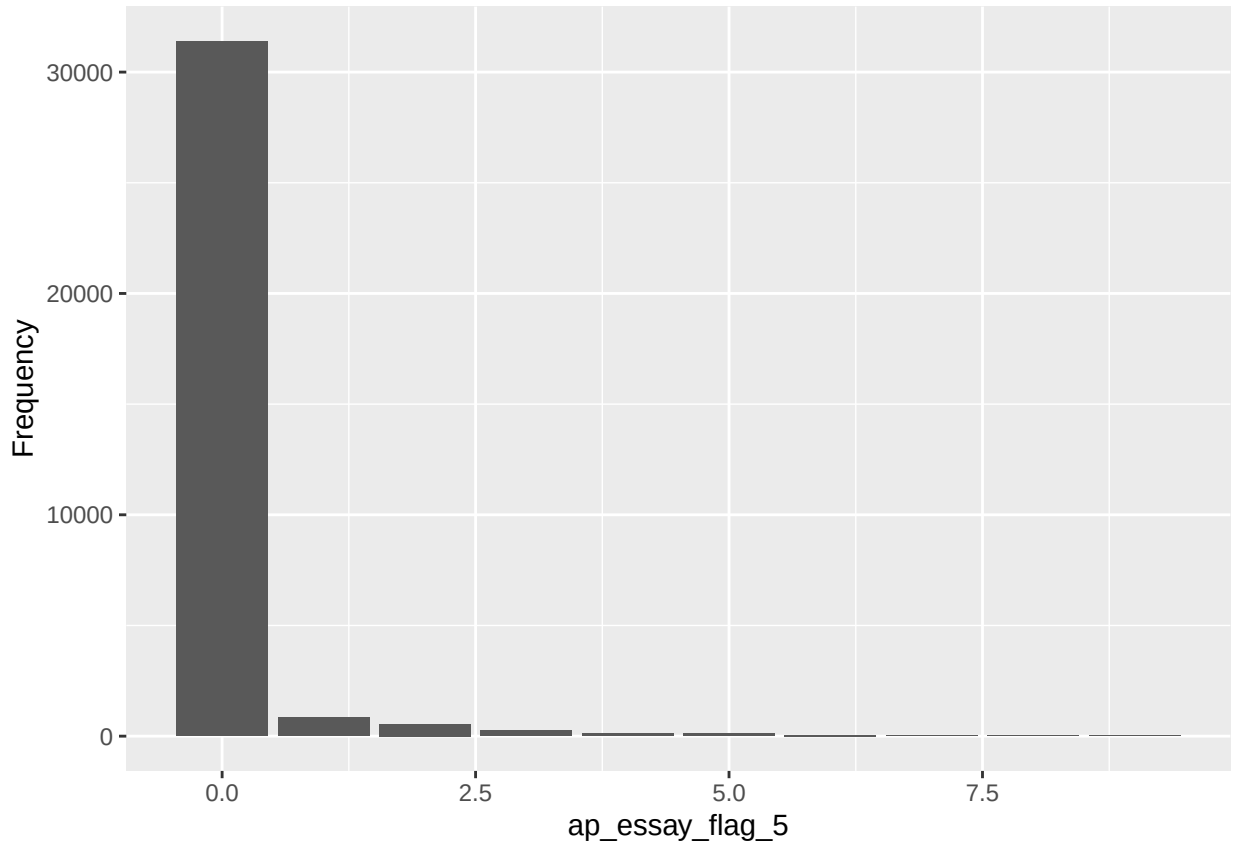

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_5, type: numeric
[1] Values (6 unique): NA, 2, 5, 3, 4, ...
[1] Missing: 66.5%
  Group.1 ap_score_5
1     F08  0.9980443
2     F09  0.9982682
3     F10  0.9990930
4     F11  0.4813440
5     F12  0.5412483
6     F13  0.5011038
7     F14  0.4796522
8     F15  0.7009574
9     F16  0.4971837

Warning: Removed 30857 rows containing non-finite values (‘stat_count()‘).
```



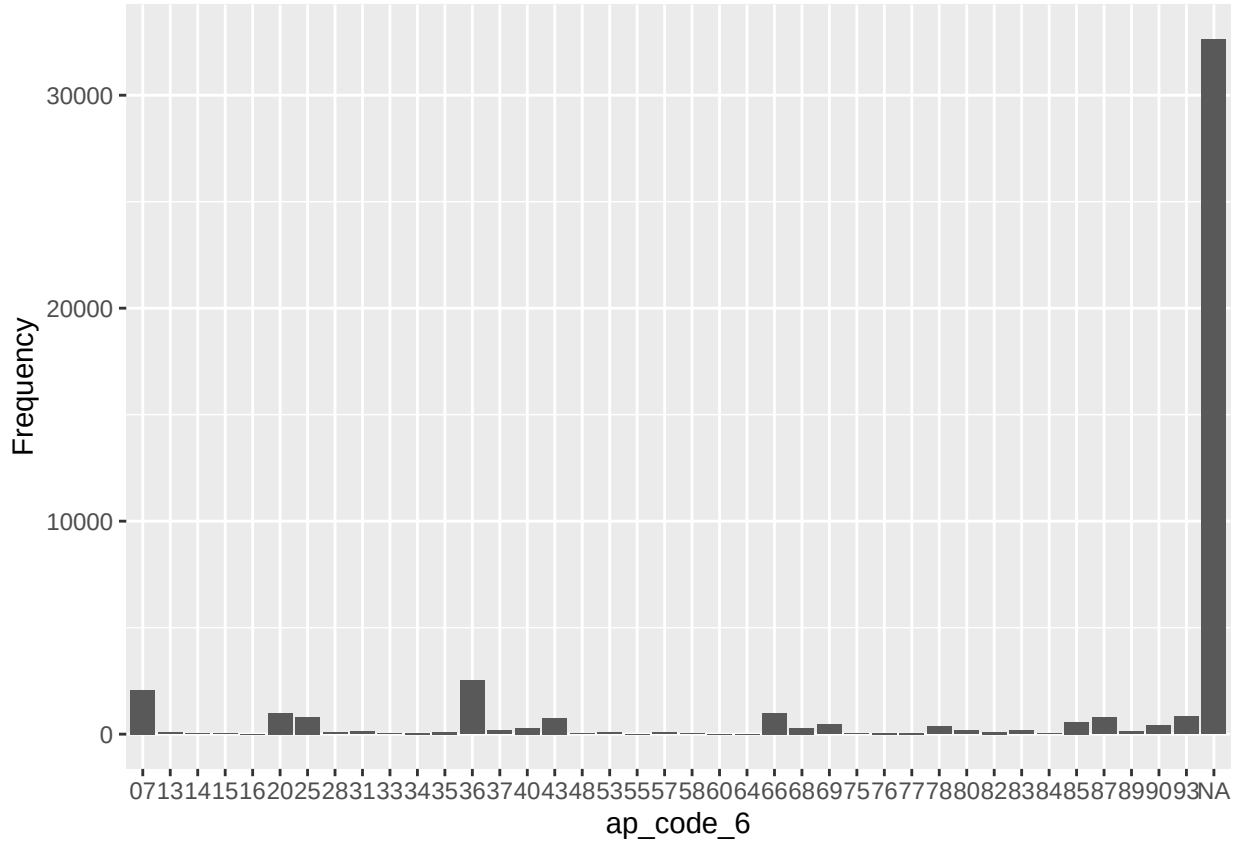

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_5, type: integer
[1] Values (13 unique): NA, 2015, 2014, 2016, 2013, ...
[1] Missing: 62%
  Group.1 ap_year_5
1     F08 0.9980443
2     F09 0.9982682
3     F10 0.9988662
4     F11 0.4551670
5     F12 0.5146682
6     F13 0.4793966
7     F14 0.4557899
8     F15 0.4694517
9     F16 0.4644543

Warning: Removed 28795 rows containing non-finite values (‘stat_count()‘).
```



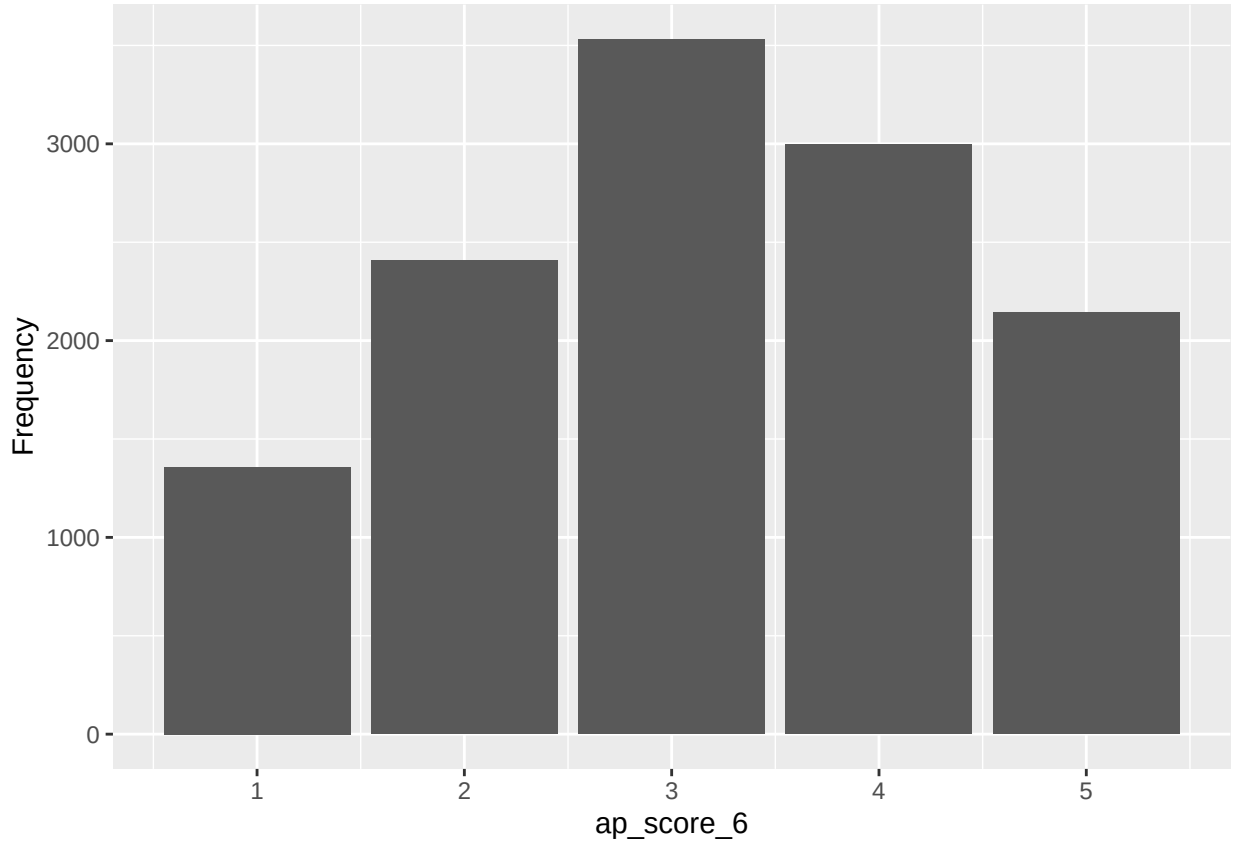

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_5, type: numeric
[1] Values (11 unique): NA, 0, 5, 1, 3, ...
[1] Missing: 28.1%
  Group.1 ap_essay_flag_5
1     F08    0.9952194698
2     F09    0.9970311727
3     F10    0.9968253968
4     F11    0.0001953507
5     F12    0.0003937783
6     F13    0.0003679176
7     F14    0.0001849797
8     F15    0.0006962576
9     F16    0.0012178414

Warning: Removed 13024 rows containing non-finite values (‘stat_count()‘).
```



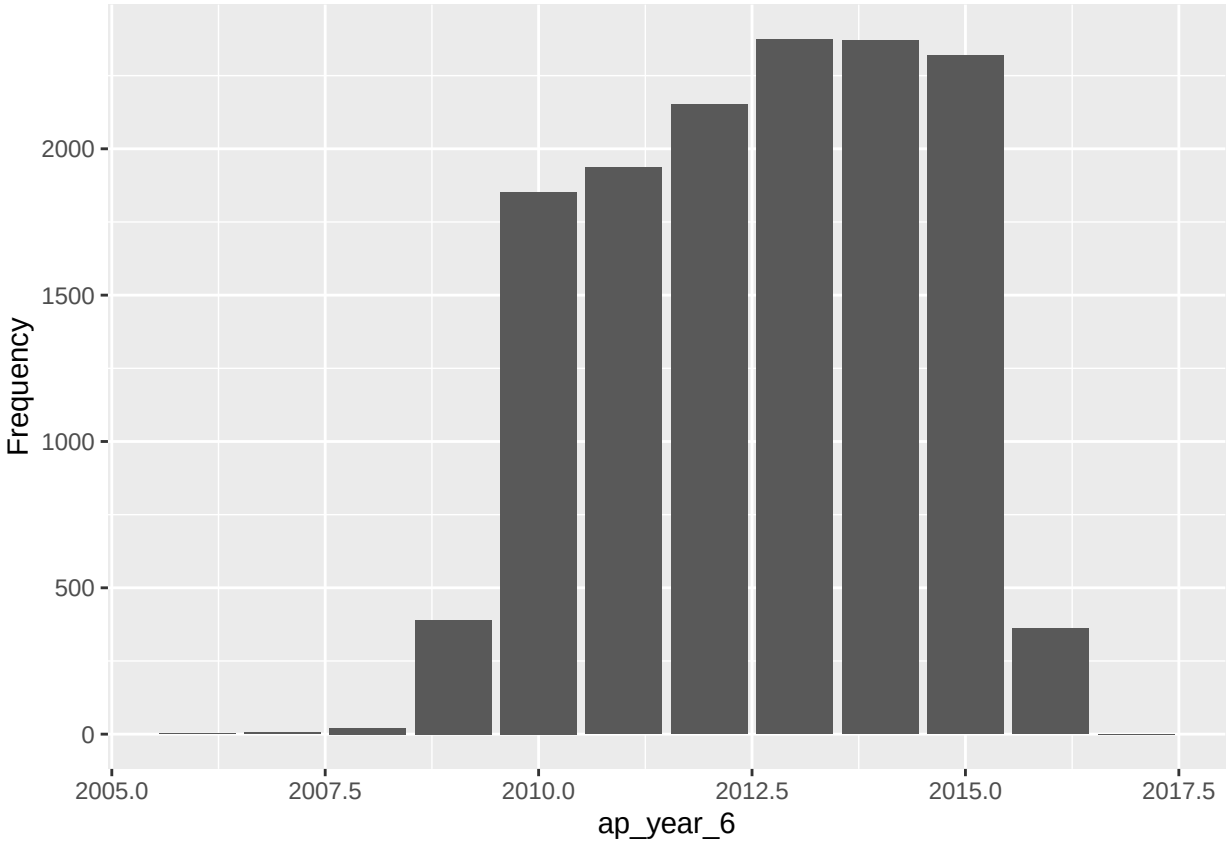

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_6, type: character
[1] Values (40 unique): NA, 07, 93, 90, 20, ...
[1] Missing: 70.3%
[1] Most missing: F10 99.9%, Least missing: F14 56.9%
```



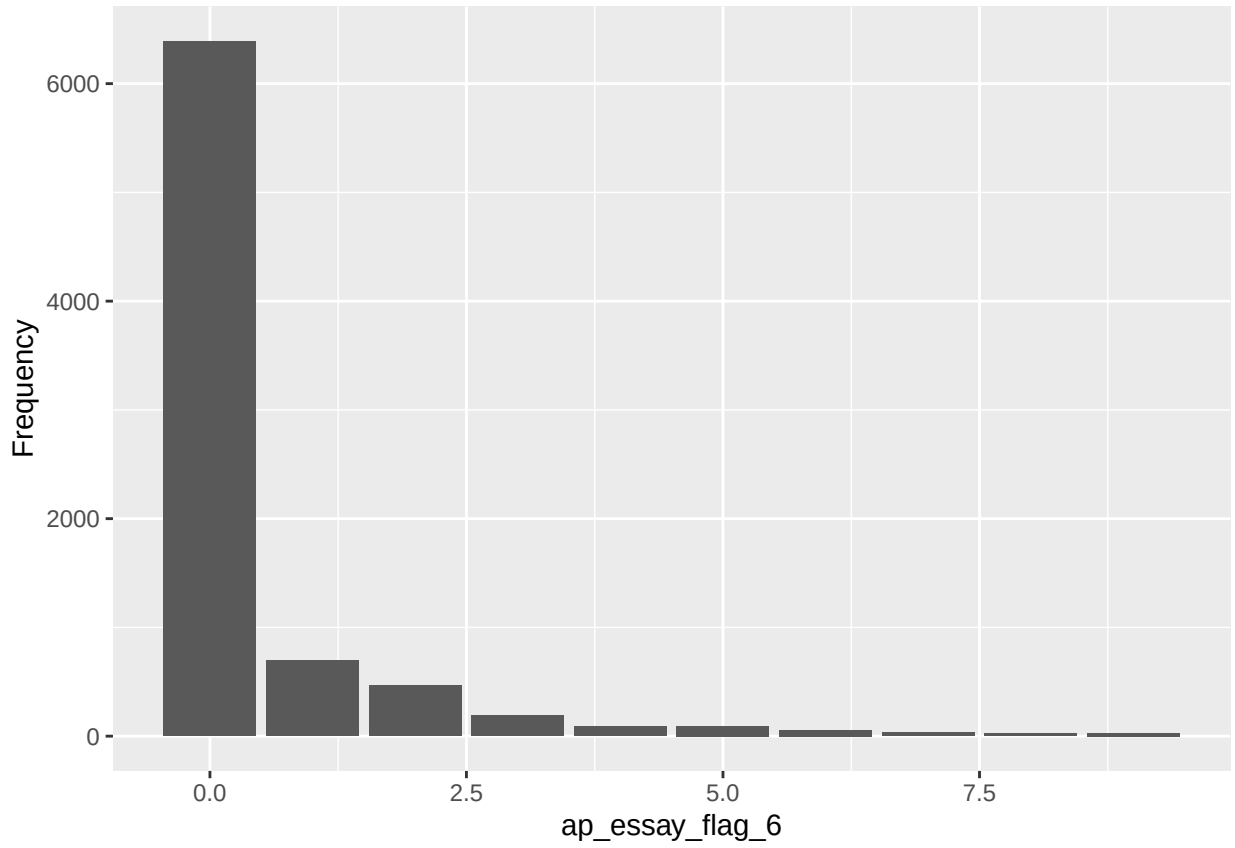

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_6, type: numeric
[1] Values (6 unique): NA, 1, 4, 3, 2, ...
[1] Missing: 73.2%
[1] Most missing: F10 99.9%, Least missing: F14 58.2%

Warning: Removed 33973 rows containing non-finite values (‘stat_count()‘).
```



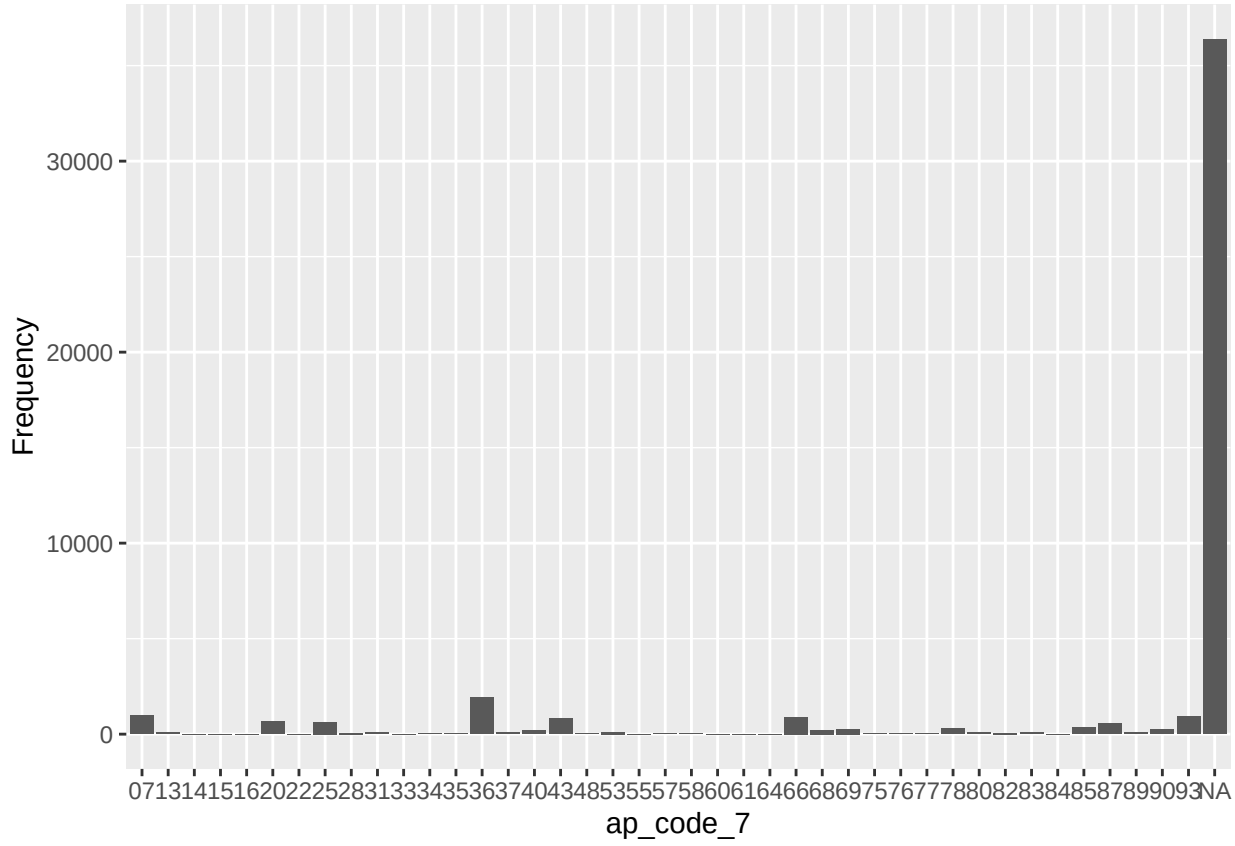

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_6, type: integer
[1] Values (13 unique): NA, 2014, 2015, 2013, 2016, ...
[1] Missing: 70.3%
[1] Most missing: F10 99.9%, Least missing: F14 56.9%

Warning: Removed 32623 rows containing non-finite values (‘stat_count()‘).
```



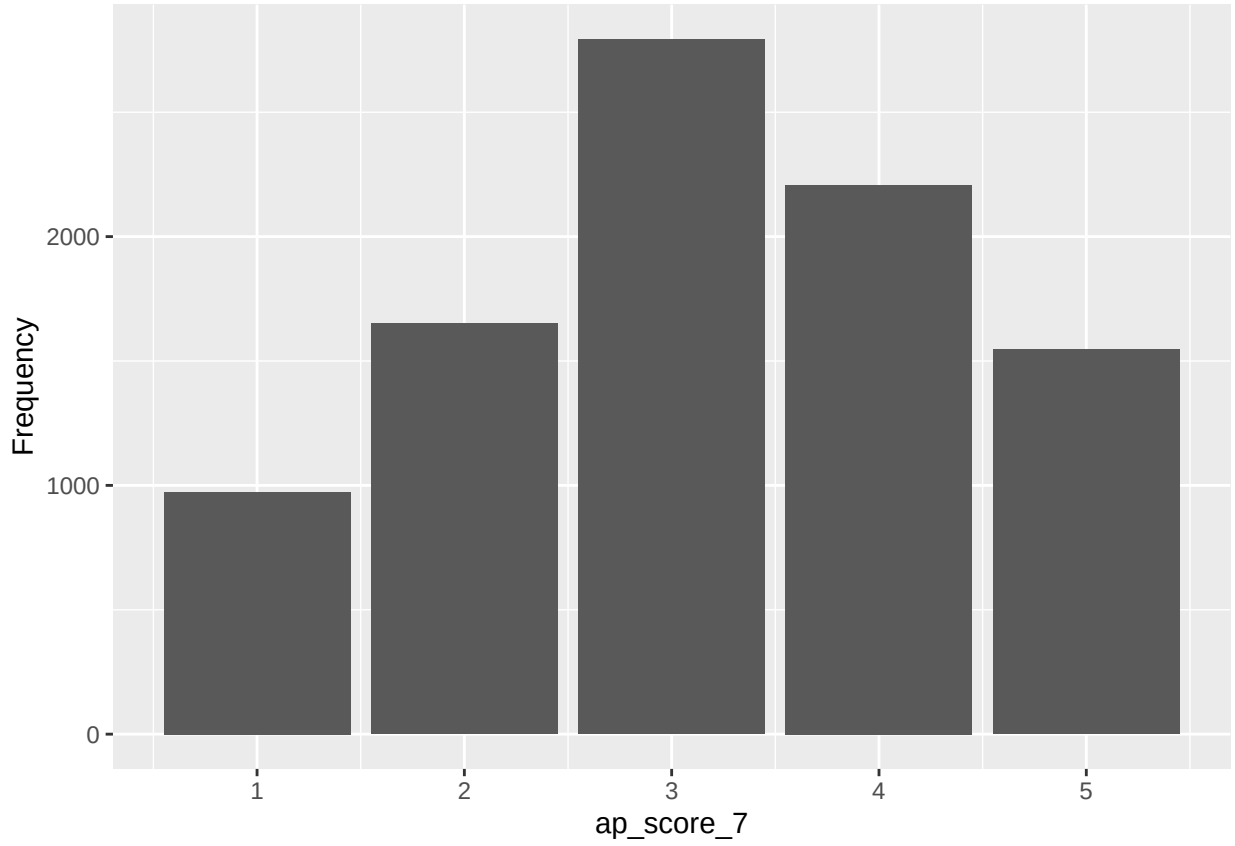

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_6, type: numeric
[1] Values (11 unique): NA, 2, 0, 3, 4, ...
[1] Missing: 82.6%
[1] Most missing: F15 100%, Least missing: F14 58.8%

Warning: Removed 38345 rows containing non-finite values (‘stat_count()‘).
```



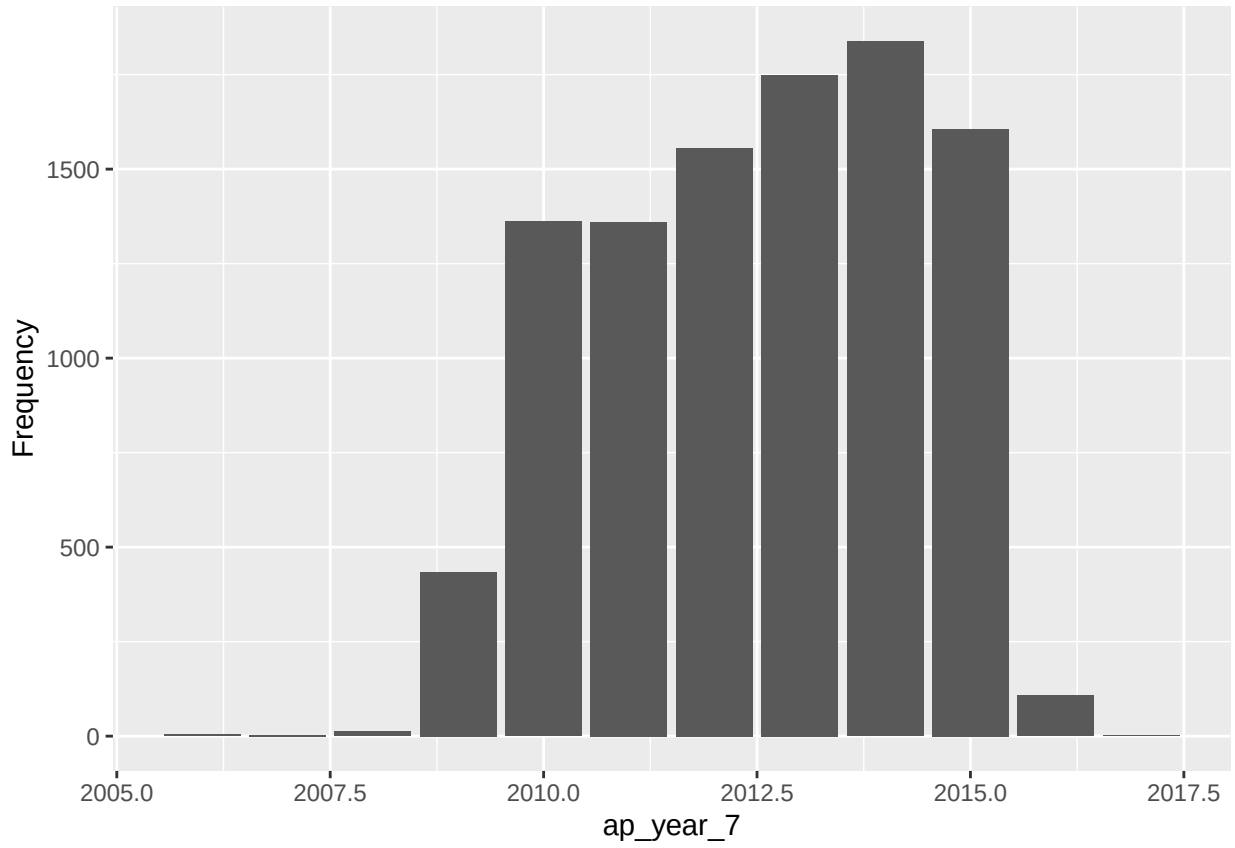

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_7, type: character
[1] Values (42 unique): NA, 36, 07, 66, 90, ...
[1] Missing: 78.4%
[1] Most missing: F10 99.9%, Least missing: F16 67.2%
```



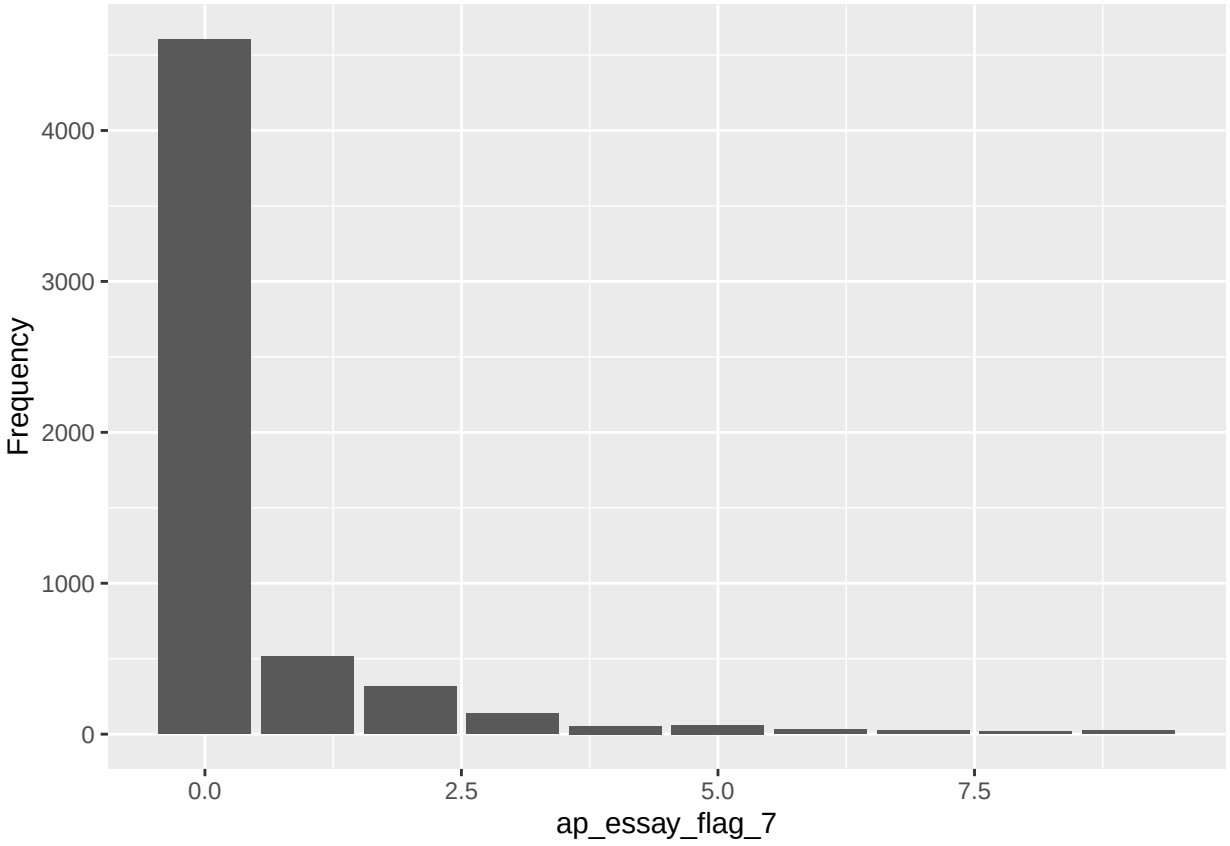

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_7, type: numeric
[1] Values (6 unique): NA, 3, 2, 1, 5, ...
[1] Missing: 80.2%
[1] Most missing: F10 99.9%, Least missing: F16 67.6%

Warning: Removed 37239 rows containing non-finite values (‘stat_count()‘).
```



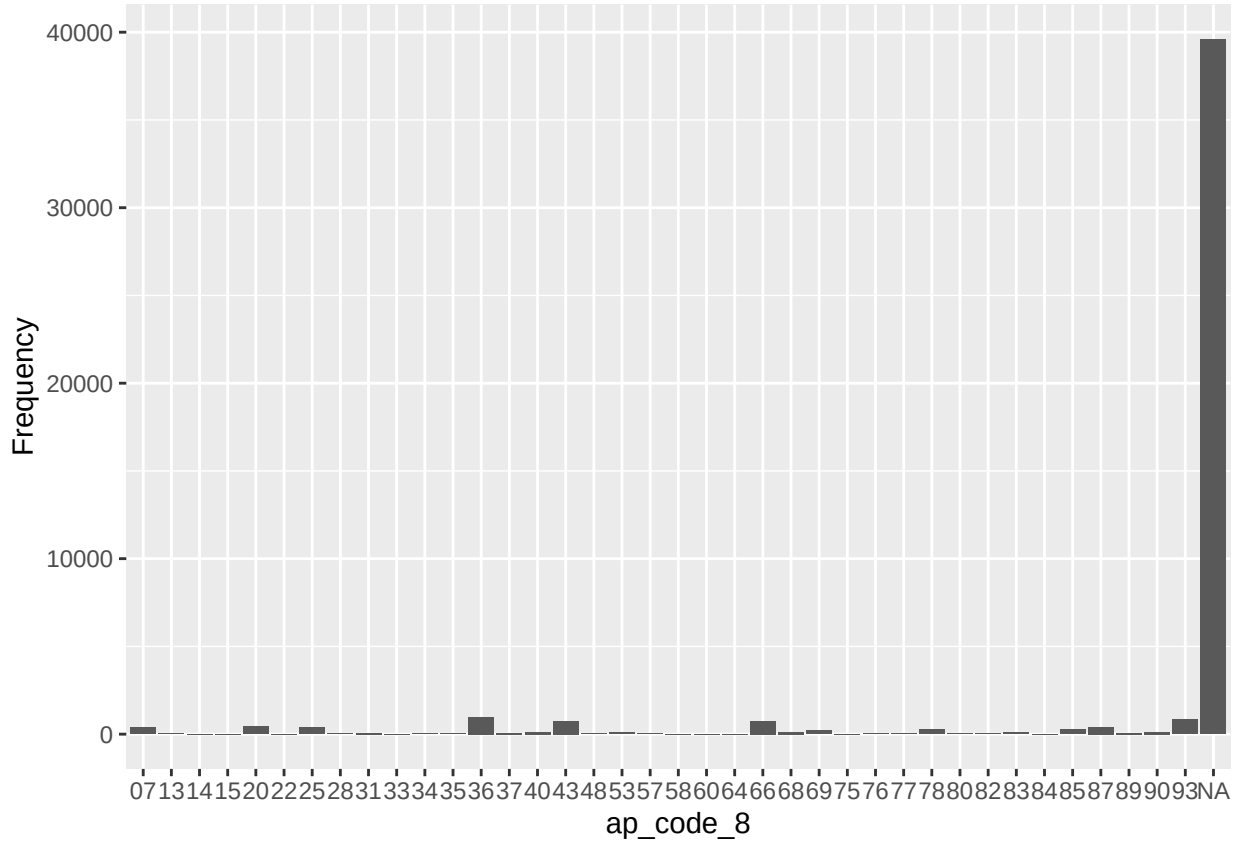

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_7, type: integer
[1] Values (13 unique): NA, 2014, 2015, 2013, 2016, ...
[1] Missing: 78.4%
[1] Most missing: F10 99.9%, Least missing: F16 67.2%

Warning: Removed 36376 rows containing non-finite values (‘stat_count()‘).
```



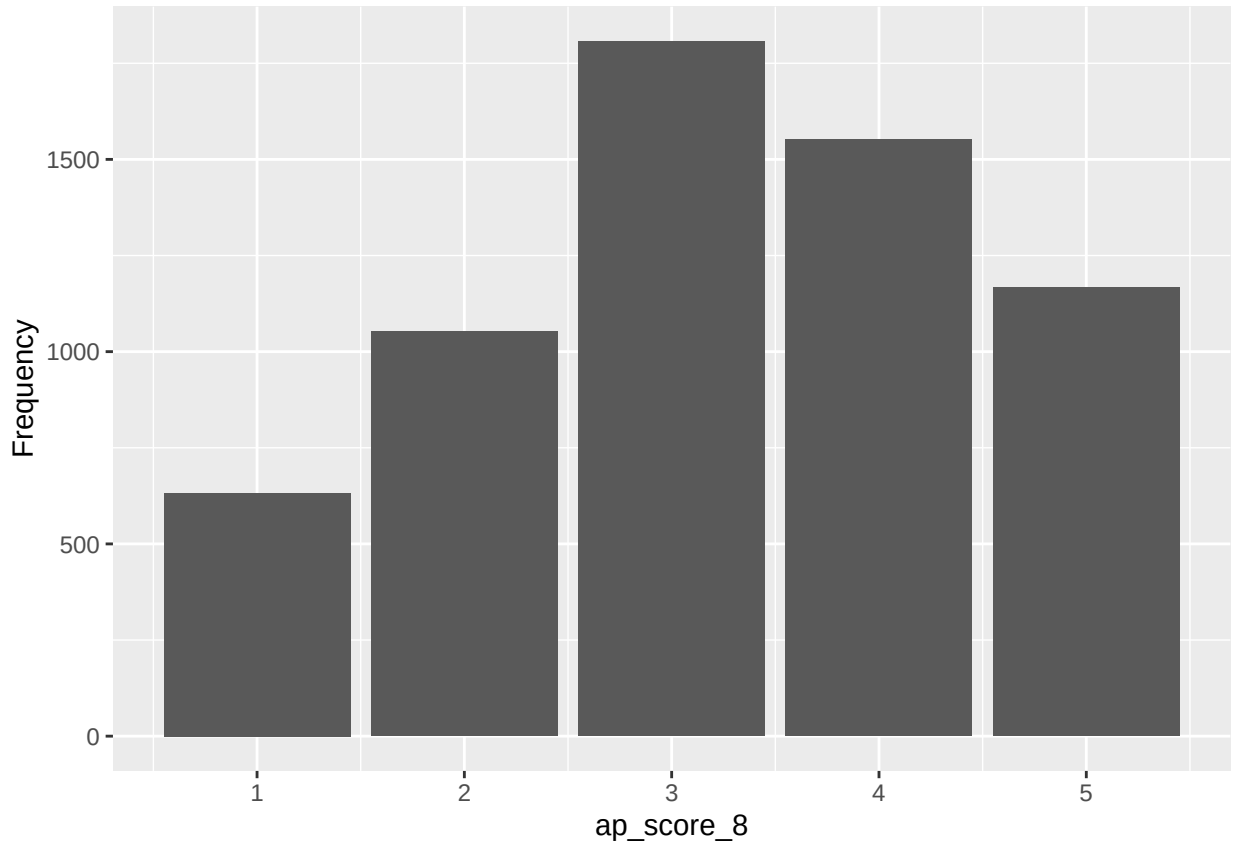

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_7, type: numeric
[1] Values (11 unique): NA, 3, 0, 1, 5, ...
[1] Missing: 87.5%
[1] Most missing: F15 100%, Least missing: F14 69.4%

Warning: Removed 40624 rows containing non-finite values (‘stat_count()‘).
```



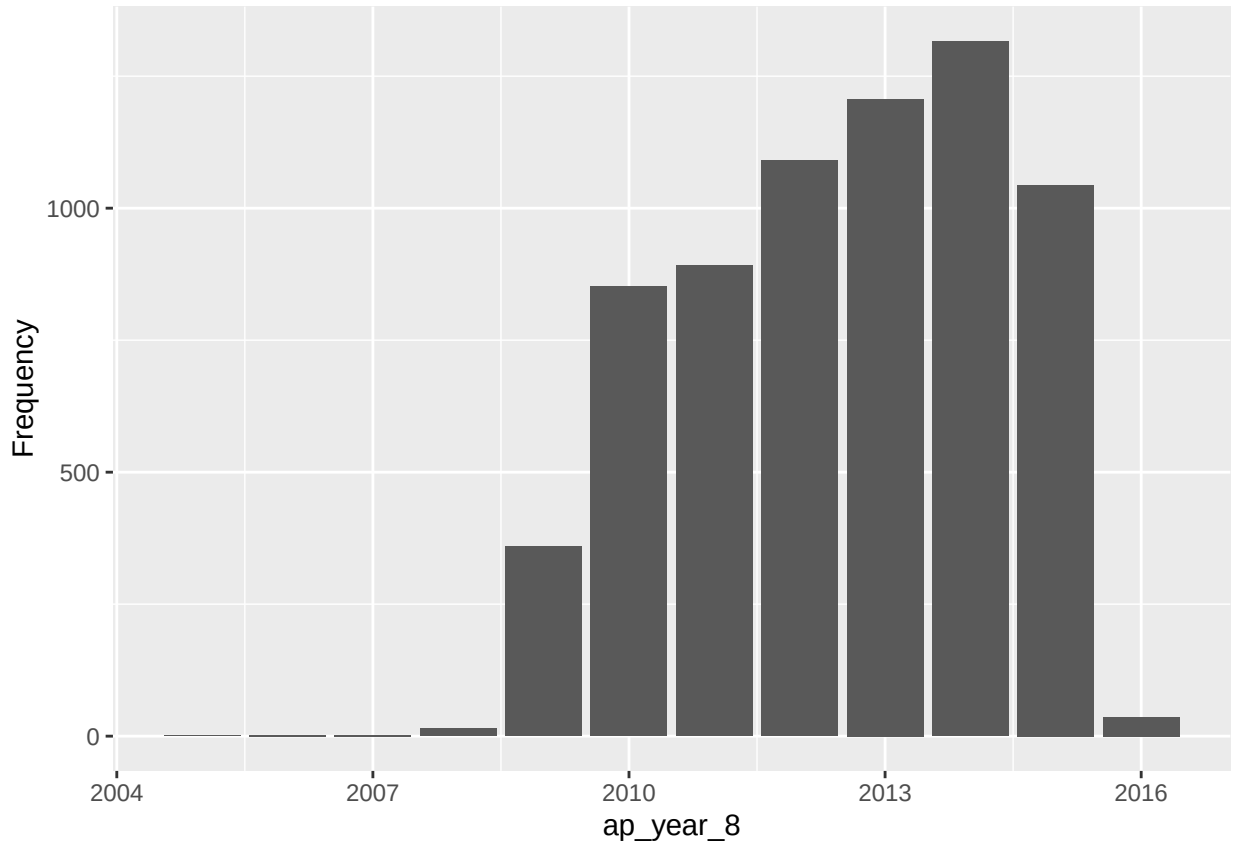

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_8, type: character
[1] Values (39 unique): NA, 78, 20, 36, 93, ...
[1] Missing: 85.3%
[1] Most missing: F10 99.9%, Least missing: F16 75.8%
```



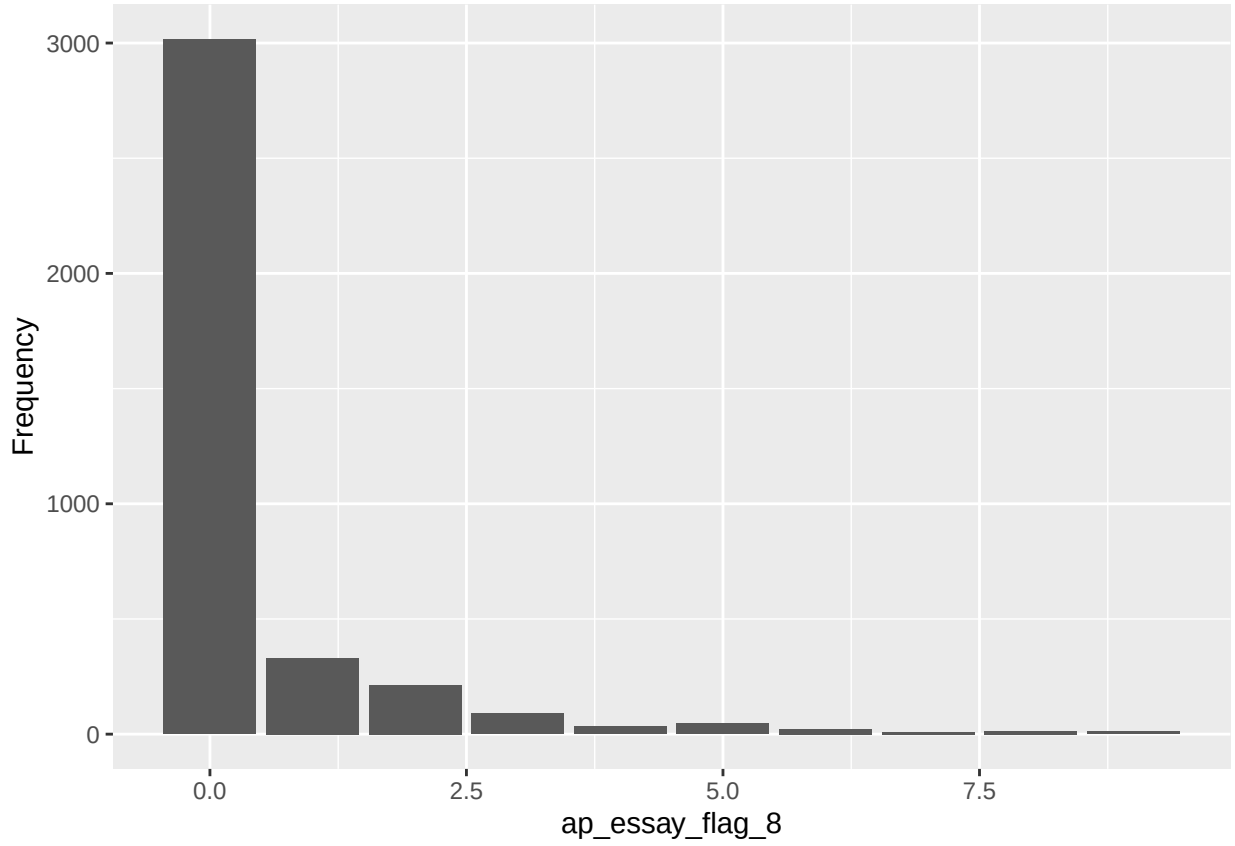

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_8, type: numeric
[1] Values (6 unique): NA, 2, 3, 5, 4, ...
[1] Missing: 86.6%
[1] Most missing: F08 99.9%, Least missing: F16 76%

Warning: Removed 40198 rows containing non-finite values (‘stat_count()‘).
```



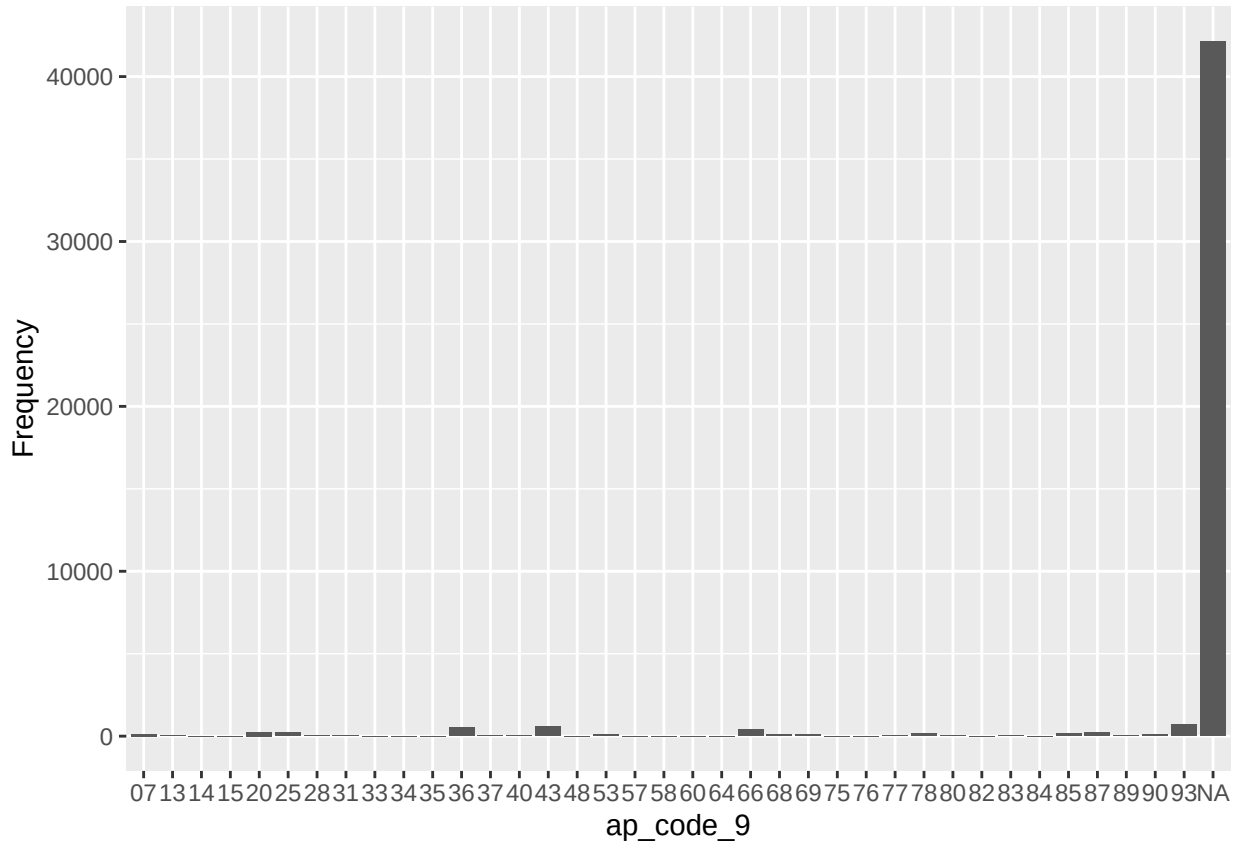

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_8, type: integer
[1] Values (13 unique): NA, 2014, 2015, 2013, 2012, ...
[1] Missing: 85.3%
[1] Most missing: F10 99.9%, Least missing: F16 75.8%

Warning: Removed 39593 rows containing non-finite values (‘stat_count()‘).
```



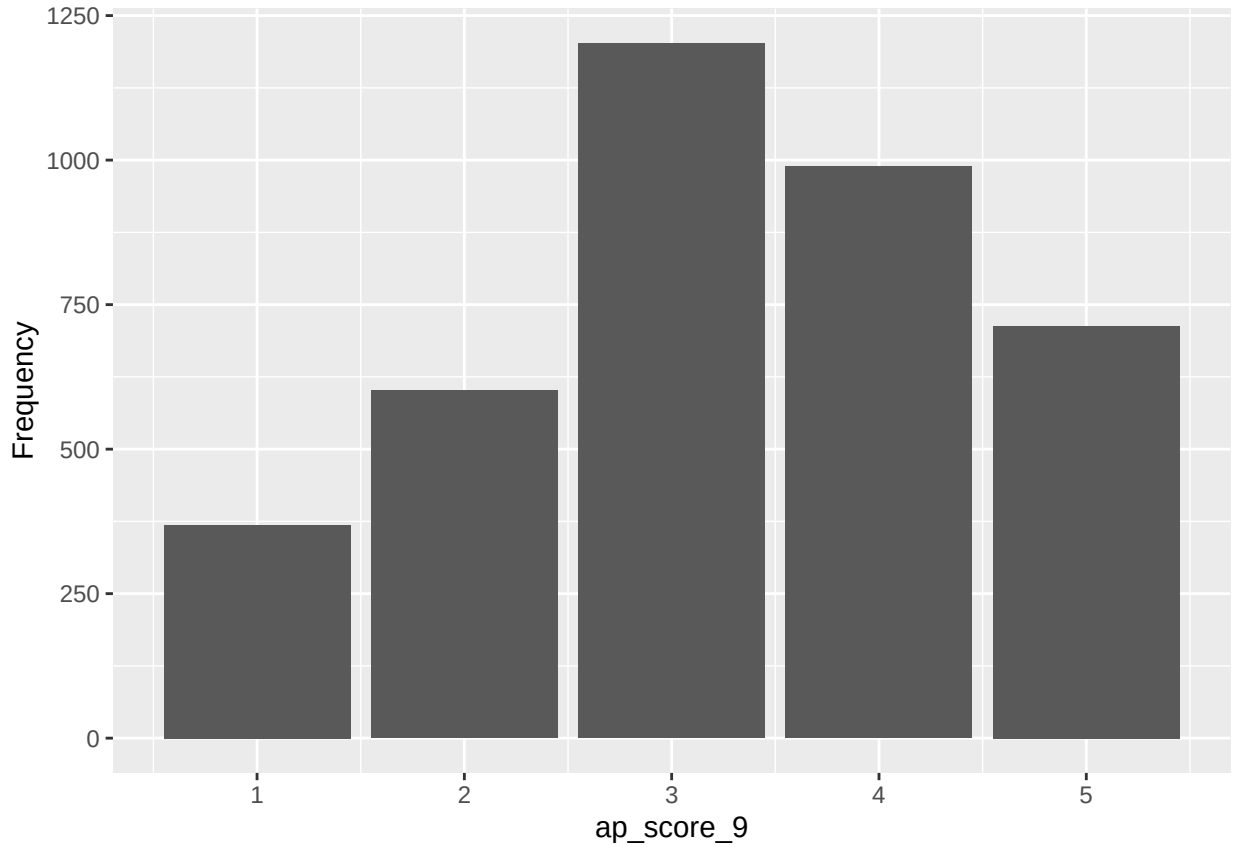

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_8, type: numeric
[1] Values (11 unique): NA, 0, 3, 1, 5, ...
[1] Missing: 91.8%
[1] Most missing: F15 100%, Least missing: F14 79.2%

Warning: Removed 42624 rows containing non-finite values (‘stat_count()‘).
```



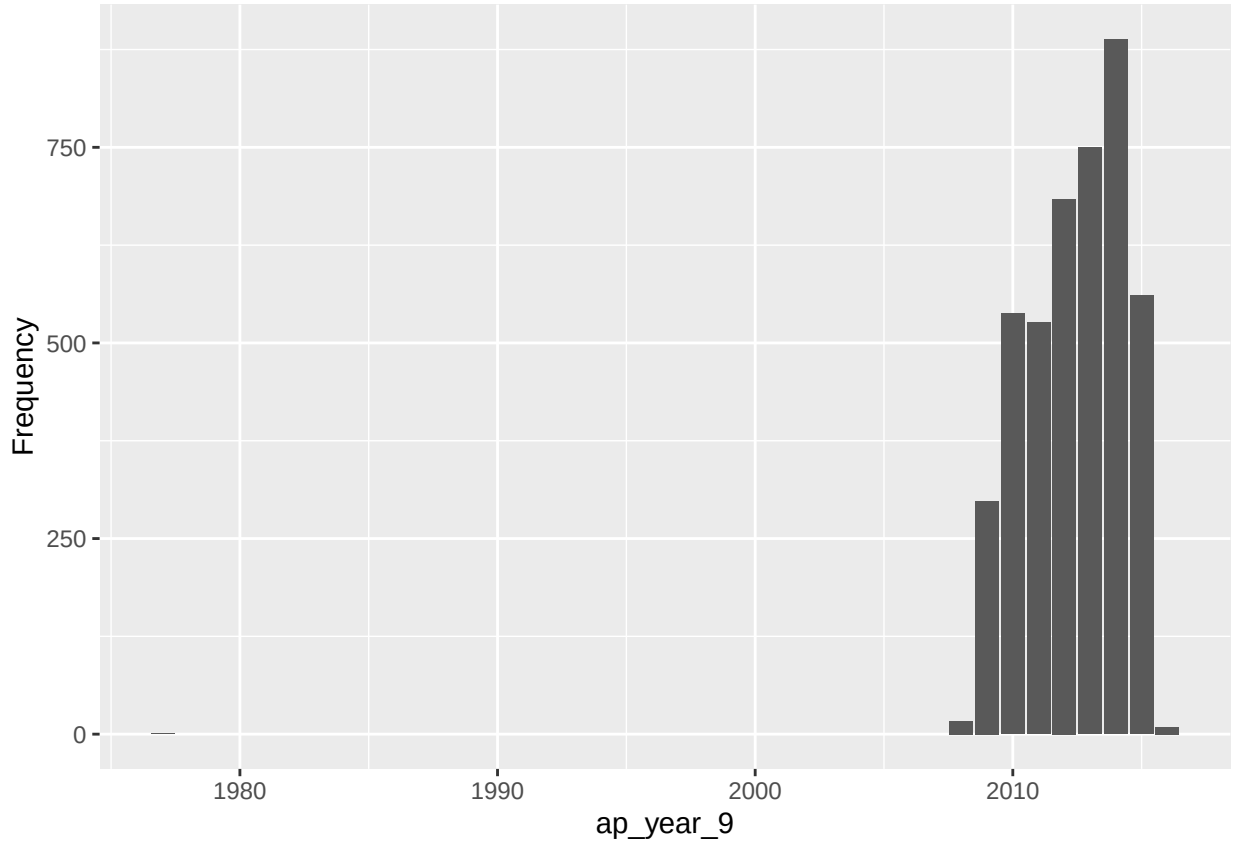

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_9, type: character
[1] Values (38 unique): NA, 36, 93, 87, 53, ...
[1] Missing: 90.8%
[1] Most missing: F08 100%, Least missing: F16 83.9%
```



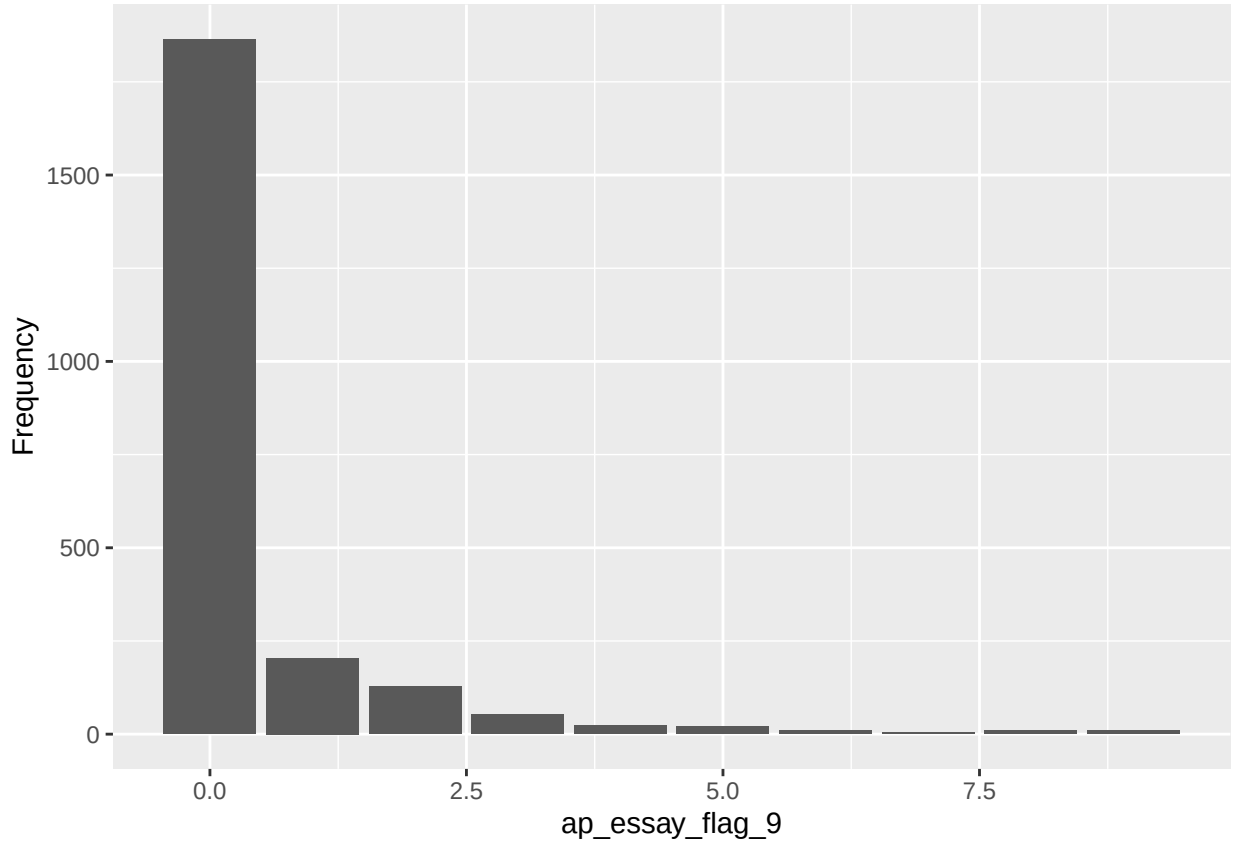

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_9, type: numeric
[1] Values (6 unique): NA, 3, 5, 2, 4, ...
[1] Missing: 91.7%
[1] Most missing: F08 100%, Least missing: F16 84%

Warning: Removed 42533 rows containing non-finite values (‘stat_count()‘).
```



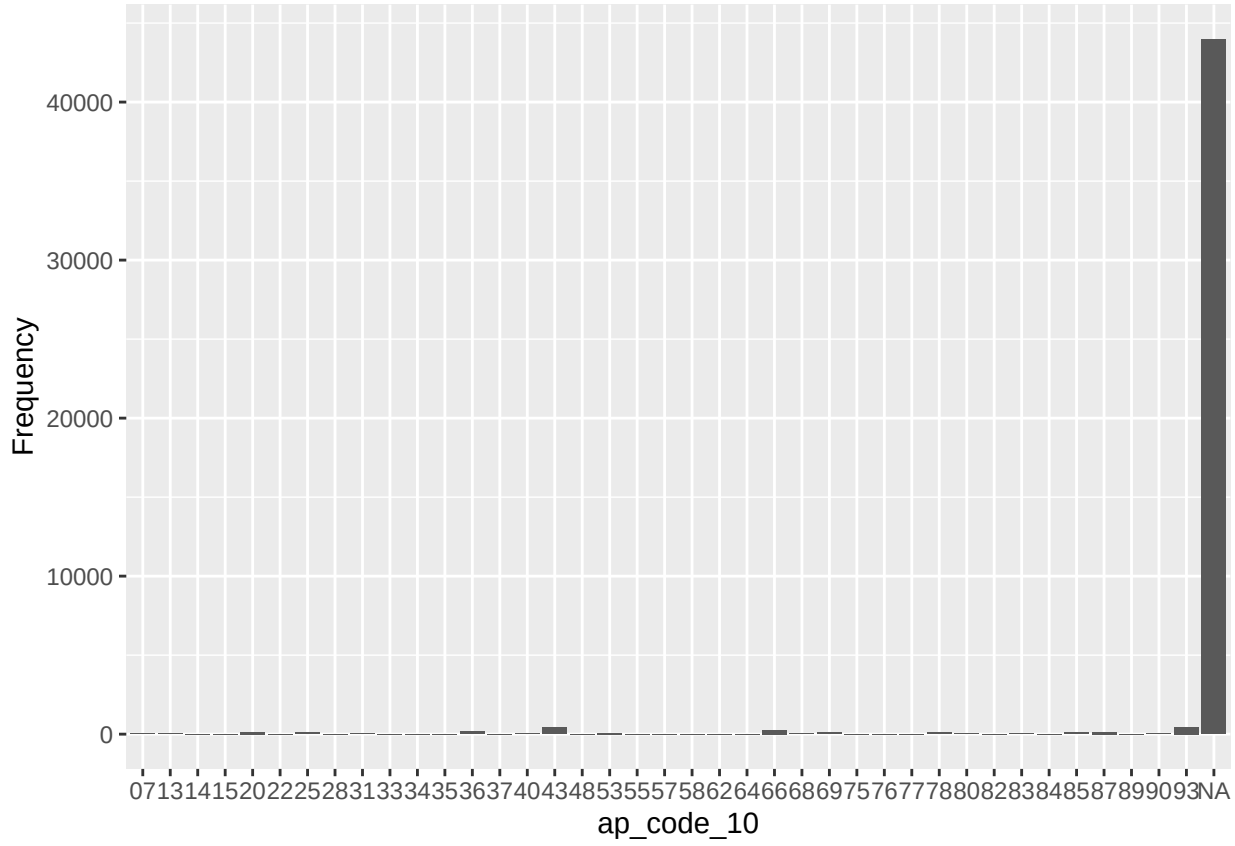

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_9, type: integer
[1] Values (11 unique): NA, 2014, 2015, 2013, 2012, ...
[1] Missing: 90.8%
[1] Most missing: F08 100%, Least missing: F16 83.9%

Warning: Removed 42136 rows containing non-finite values (‘stat_count()‘).
```



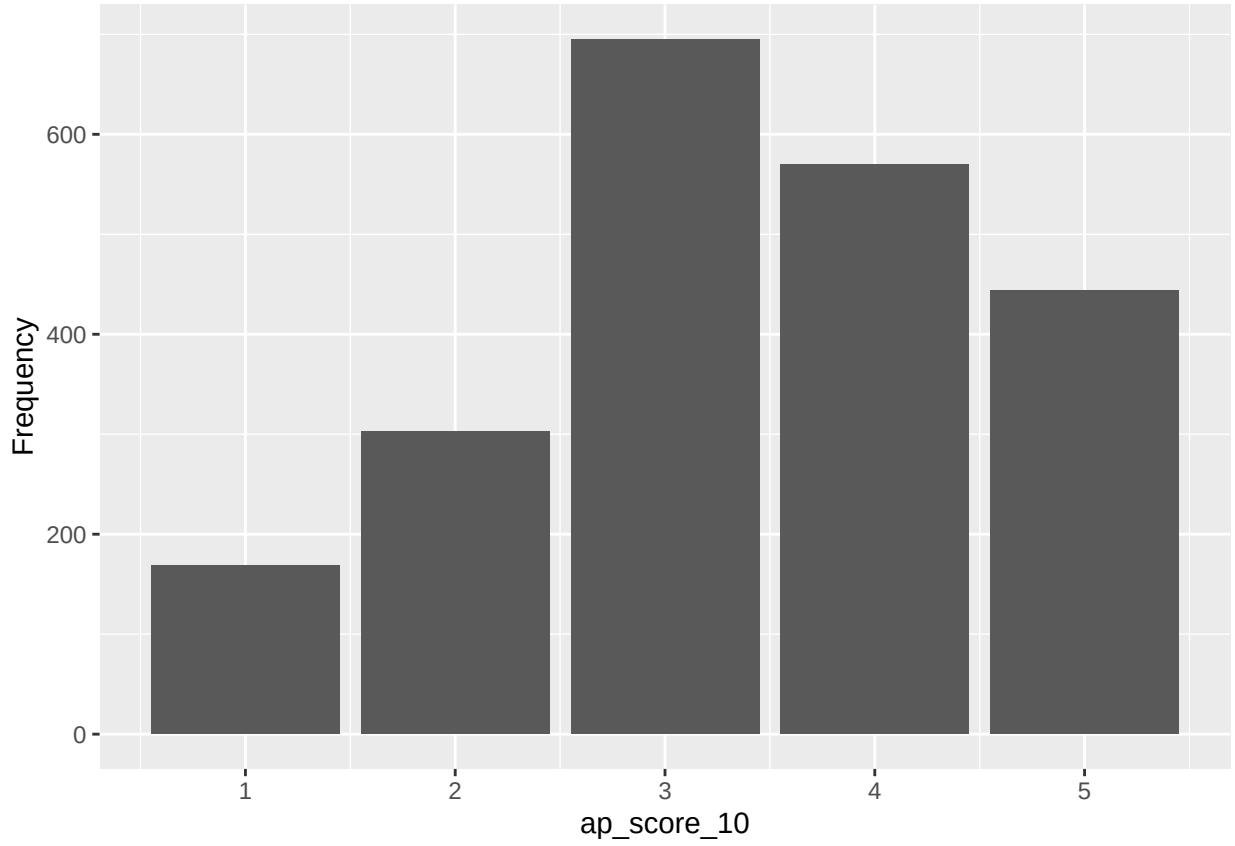

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_9, type: numeric
[1] Values (11 unique): NA, 0, 1, 2, 5, ...
[1] Missing: 95%
[1] Most missing: F08 100%, Least missing: F14 86.8%

Warning: Removed 44083 rows containing non-finite values (‘stat_count()‘).
```



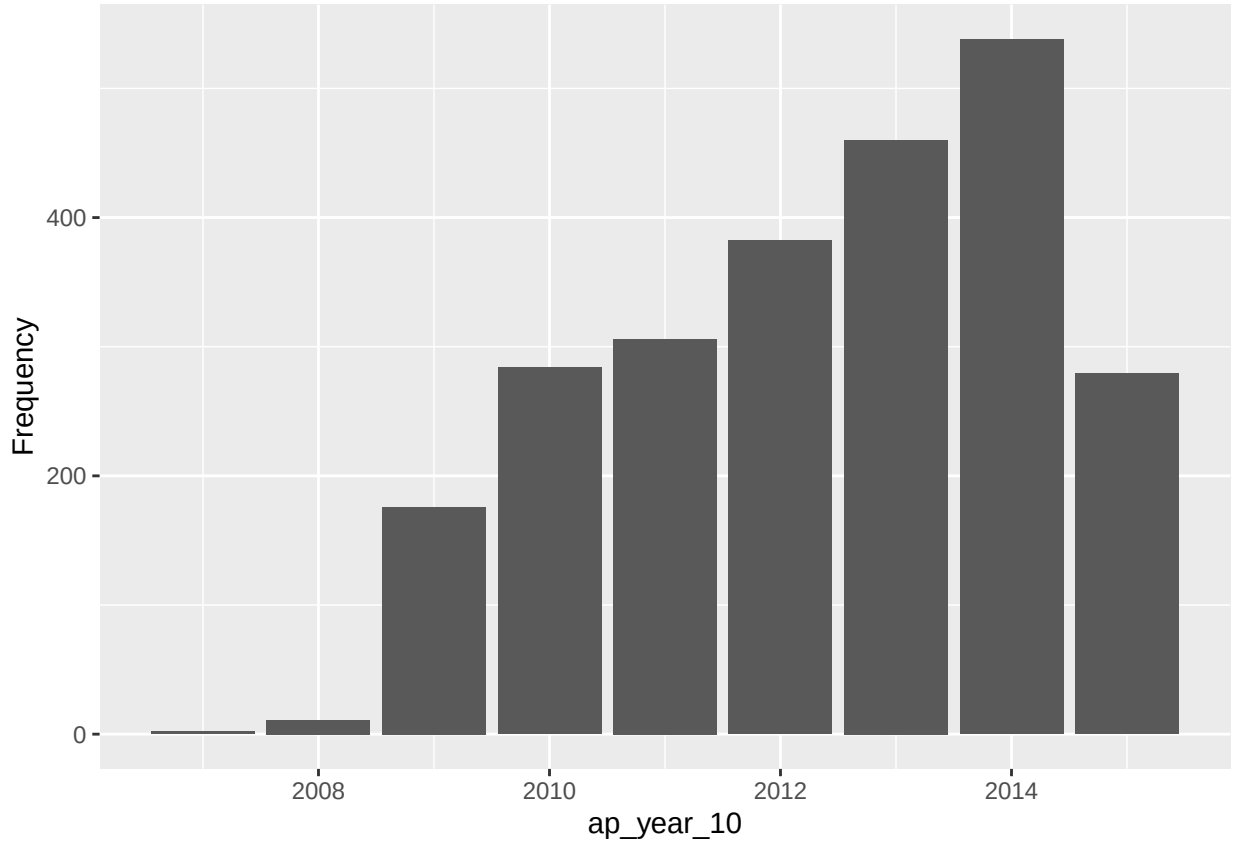

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_10, type: character
[1] Values (40 unique): NA, 66, 53, 93, 85, ...
[1] Missing: 94.7%
[1] Most missing: F08 100%, Least missing: F16 90.1%
```



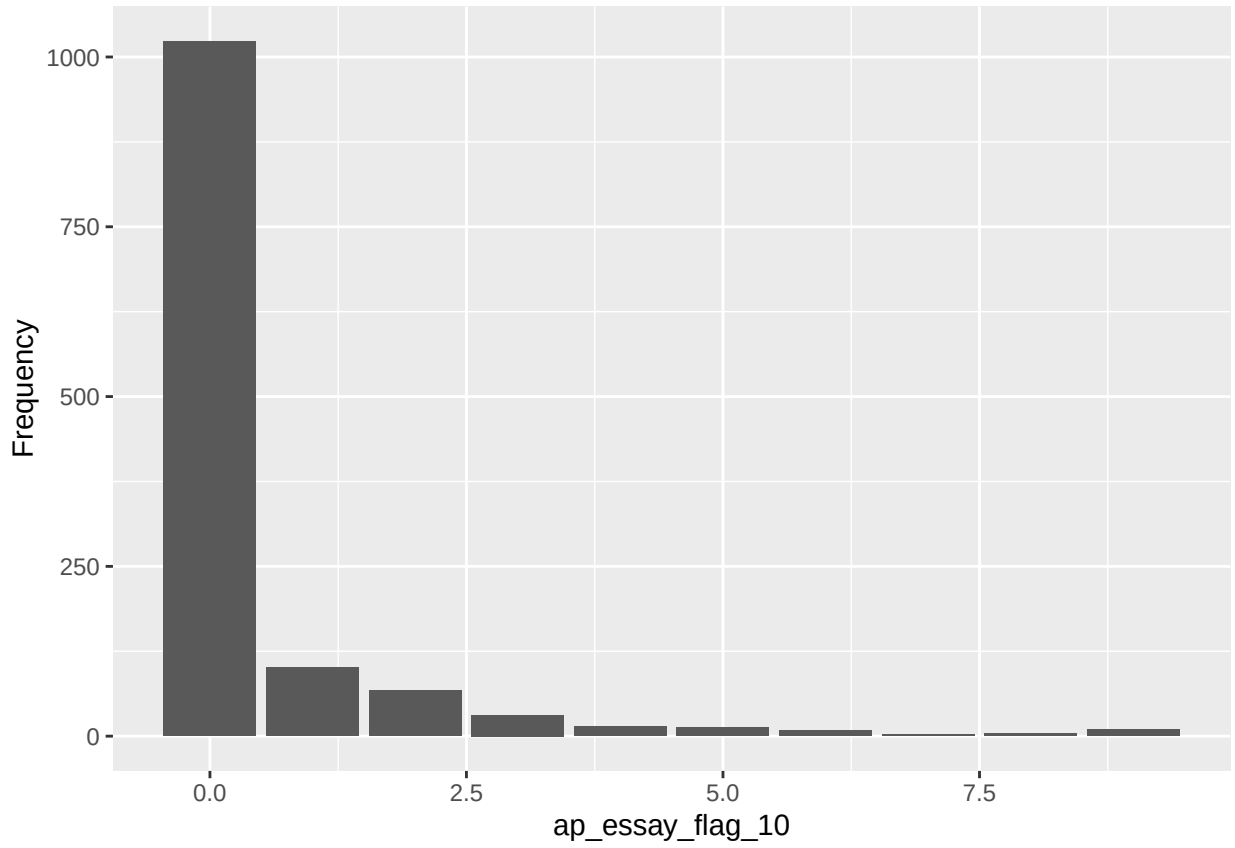

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_10, type: numeric
[1] Values (6 unique): NA, 3, 4, 2, 5, ...
[1] Missing: 95.3%
[1] Most missing: F08 100%, Least missing: F16 90.2%

Warning: Removed 44227 rows containing non-finite values (‘stat_count()‘).
```



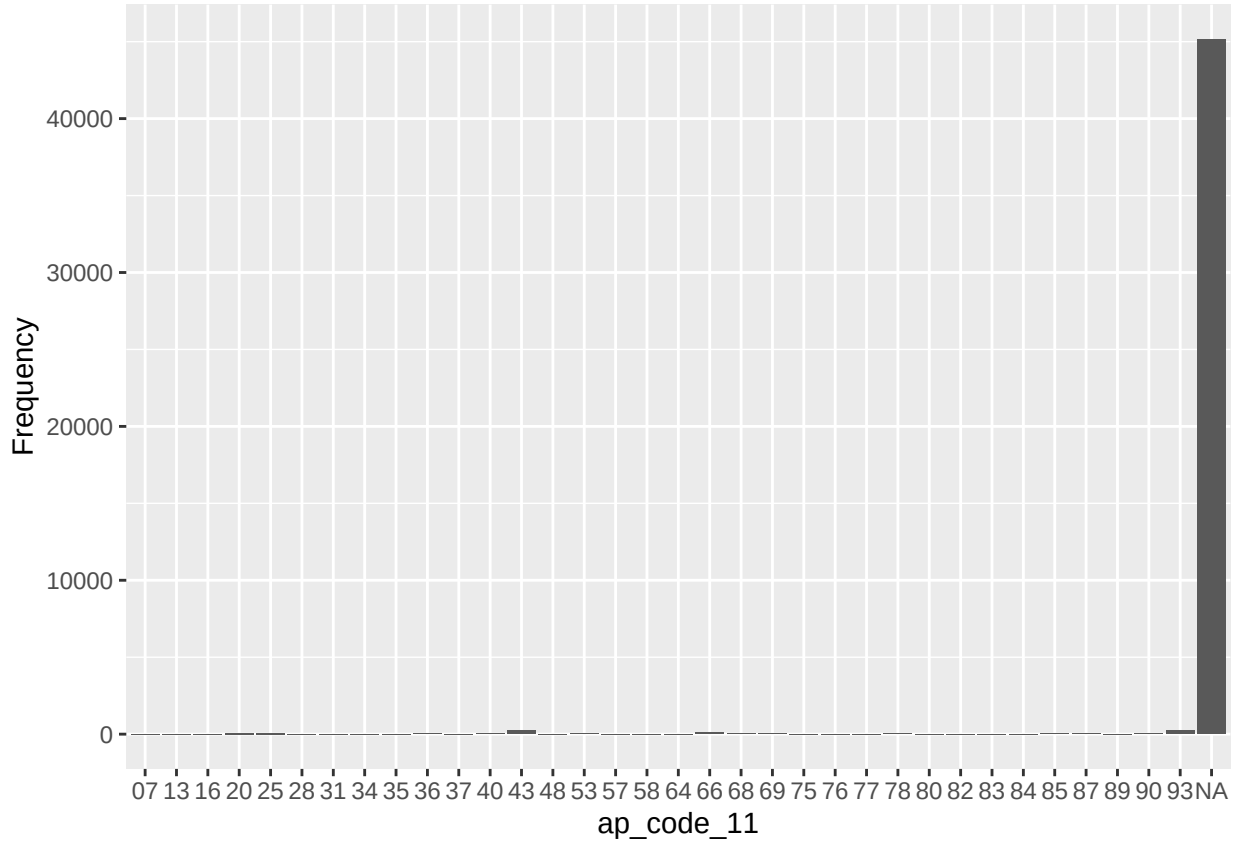

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_10, type: integer
[1] Values (10 unique): NA, 2014, 2013, 2015, 2012, ...
[1] Missing: 94.7%
[1] Most missing: F08 100%, Least missing: F16 90.1%

Warning: Removed 43970 rows containing non-finite values (‘stat_count()‘).
```



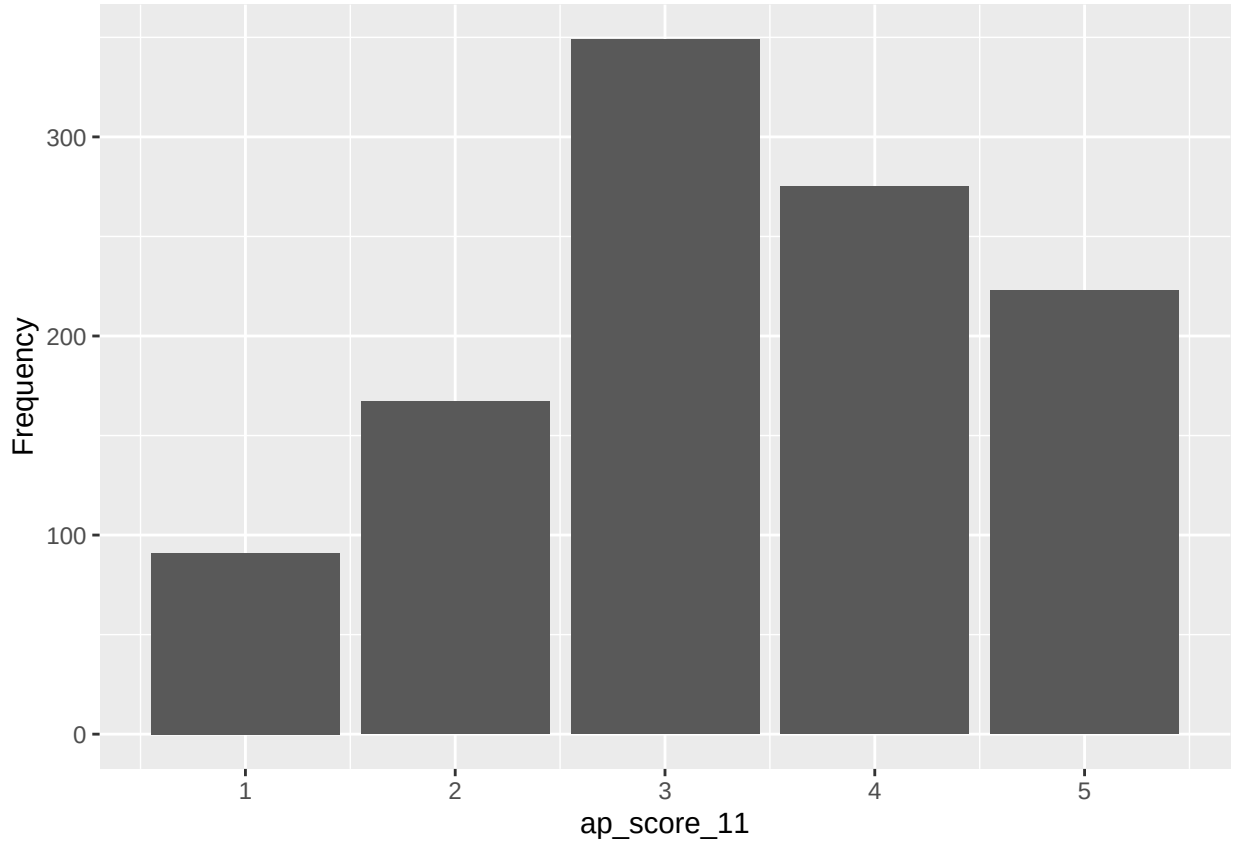

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_10, type: numeric
[1] Values (11 unique): NA, 0, 1, 2, 3, ...
[1] Missing: 97.3%
[1] Most missing: F15 100%, Least missing: F14 92.6%

Warning: Removed 45135 rows containing non-finite values (‘stat_count()‘).
```



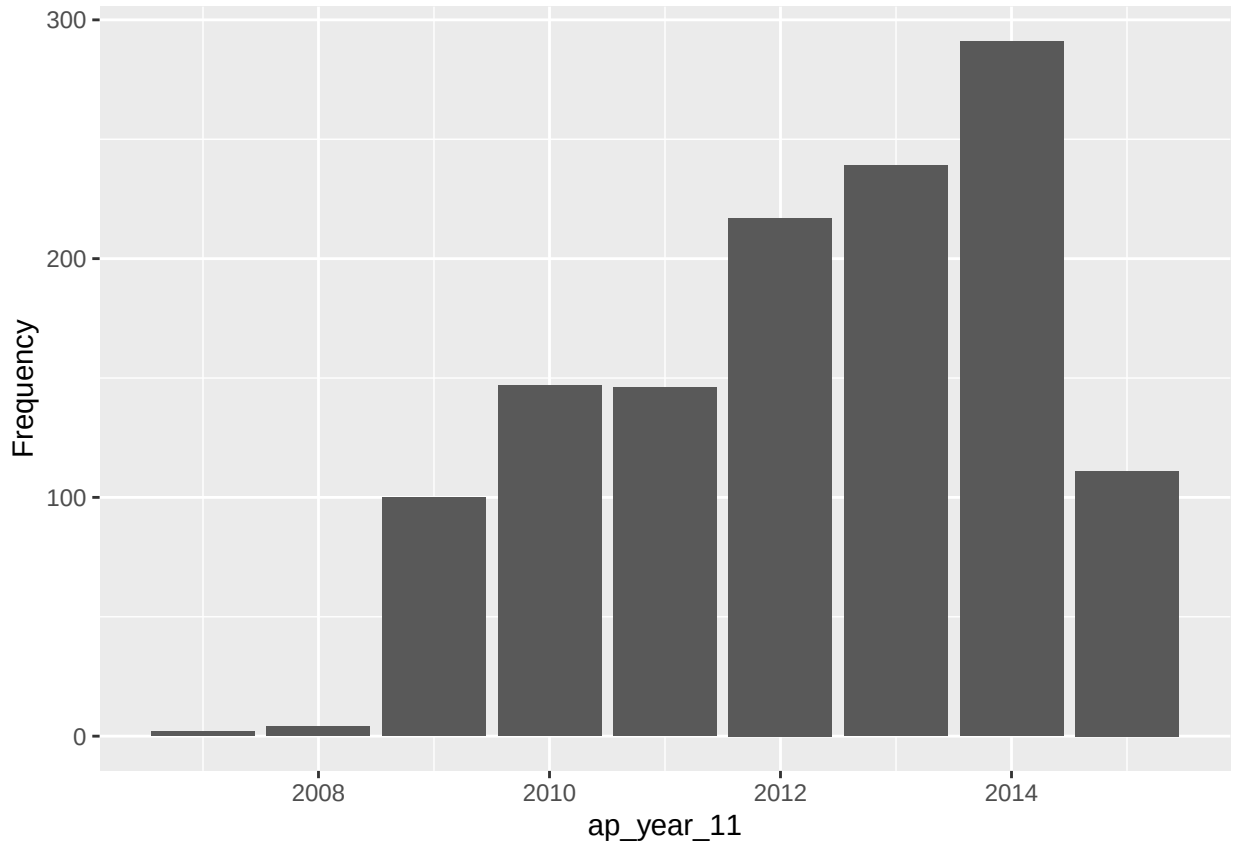

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_11, type: character
[1] Values (35 unique): NA, 43, 66, 90, 69, ...
[1] Missing: 97.3%
[1] Most missing: F08 100%, Least missing: F16 94.7%
```



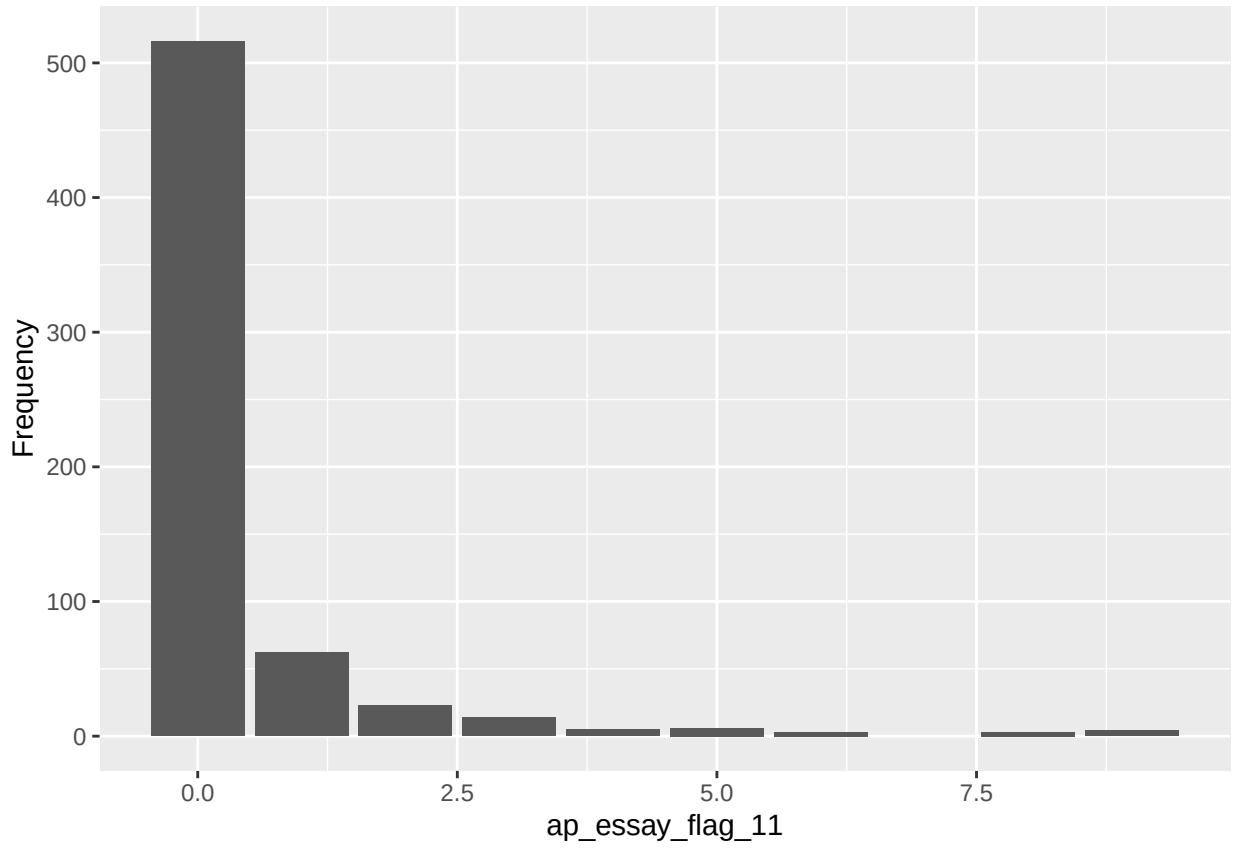

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_11, type: numeric
[1] Values (6 unique): NA, 3, 4, 5, 1, ...
[1] Missing: 97.6%
[1] Most missing: F08 100%, Least missing: F16 94.7%

Warning: Removed 45303 rows containing non-finite values (‘stat_count()‘).
```



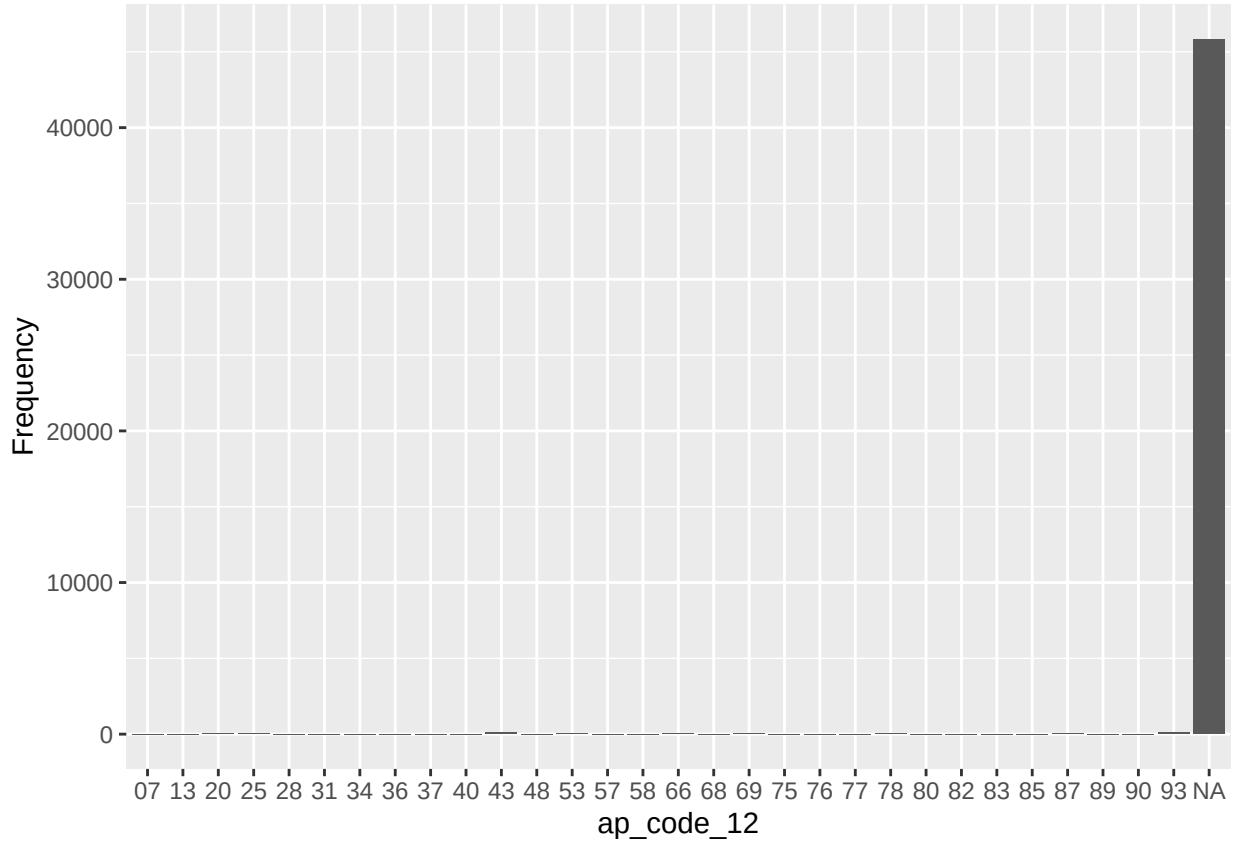

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_11, type: integer
[1] Values (10 unique): NA, 2013, 2014, 2015, 2011, ...
[1] Missing: 97.3%
[1] Most missing: F08 100%, Least missing: F16 94.7%

Warning: Removed 45151 rows containing non-finite values (‘stat_count()‘).
```



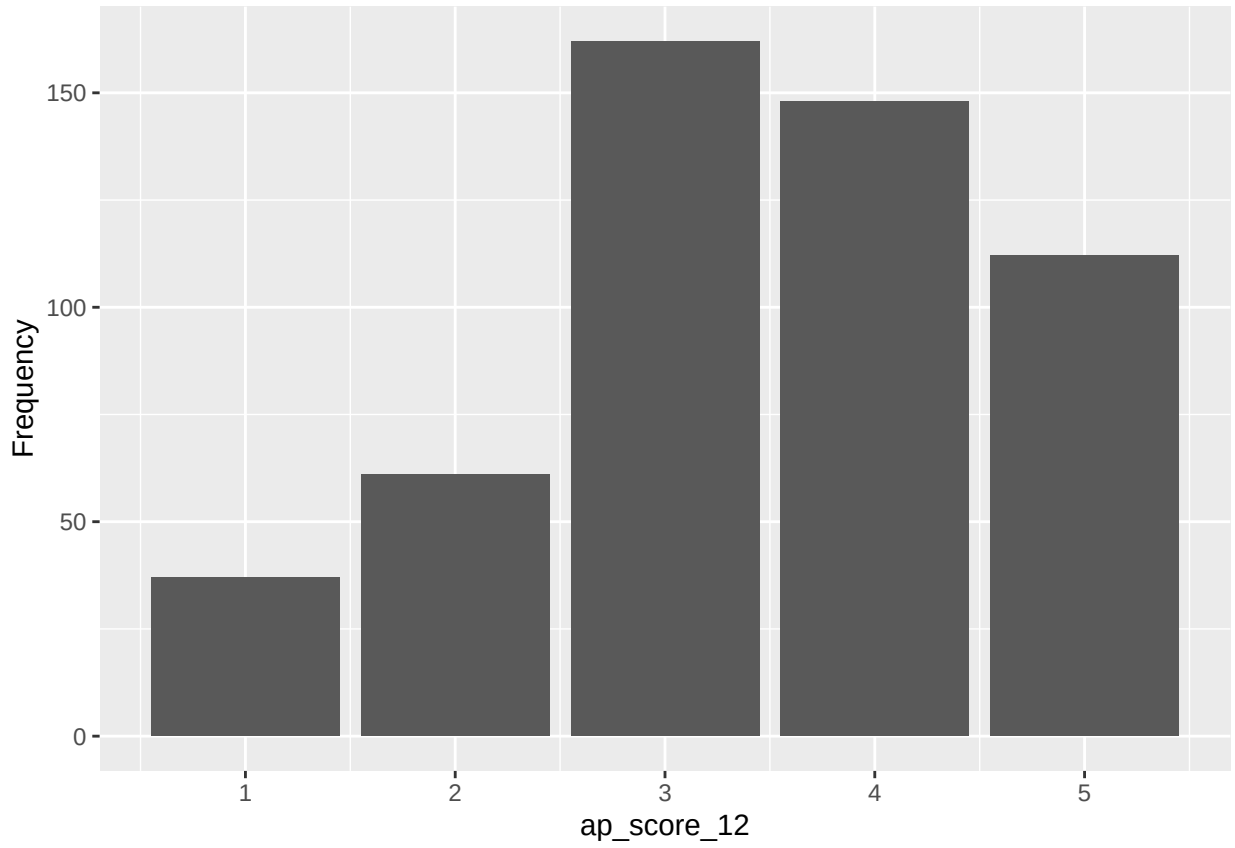

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_11, type: numeric
[1] Values (10 unique): NA, 0, 1, 2, 5, ...
[1] Missing: 98.6%
[1] Most missing: F15 100%, Least missing: F14 96.3%

Warning: Removed 45772 rows containing non-finite values (‘stat_count()‘).
```



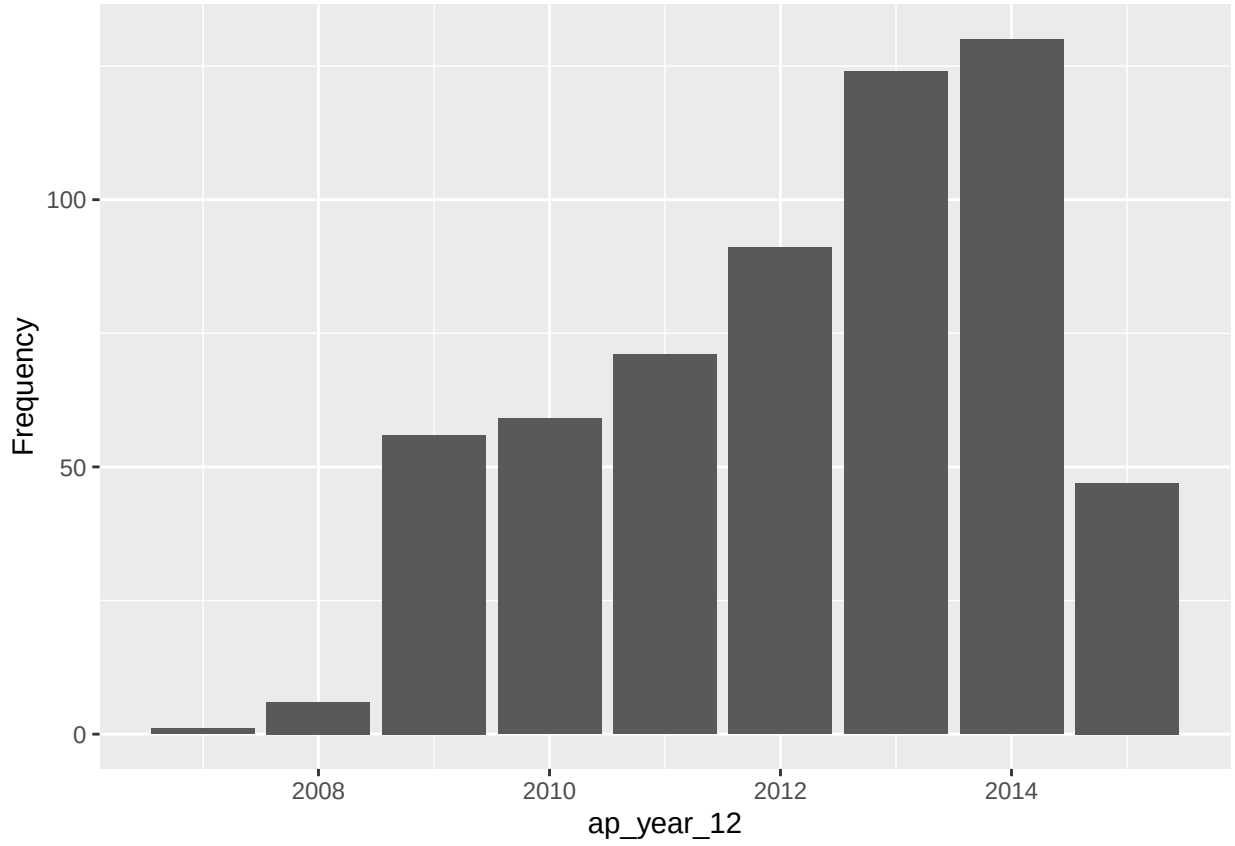

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_12, type: character
[1] Values (31 unique): NA, 43, 66, 78, 90, ...
[1] Missing: 98.7%
[1] Most missing: F09 100%, Least missing: F16 97.4%
```



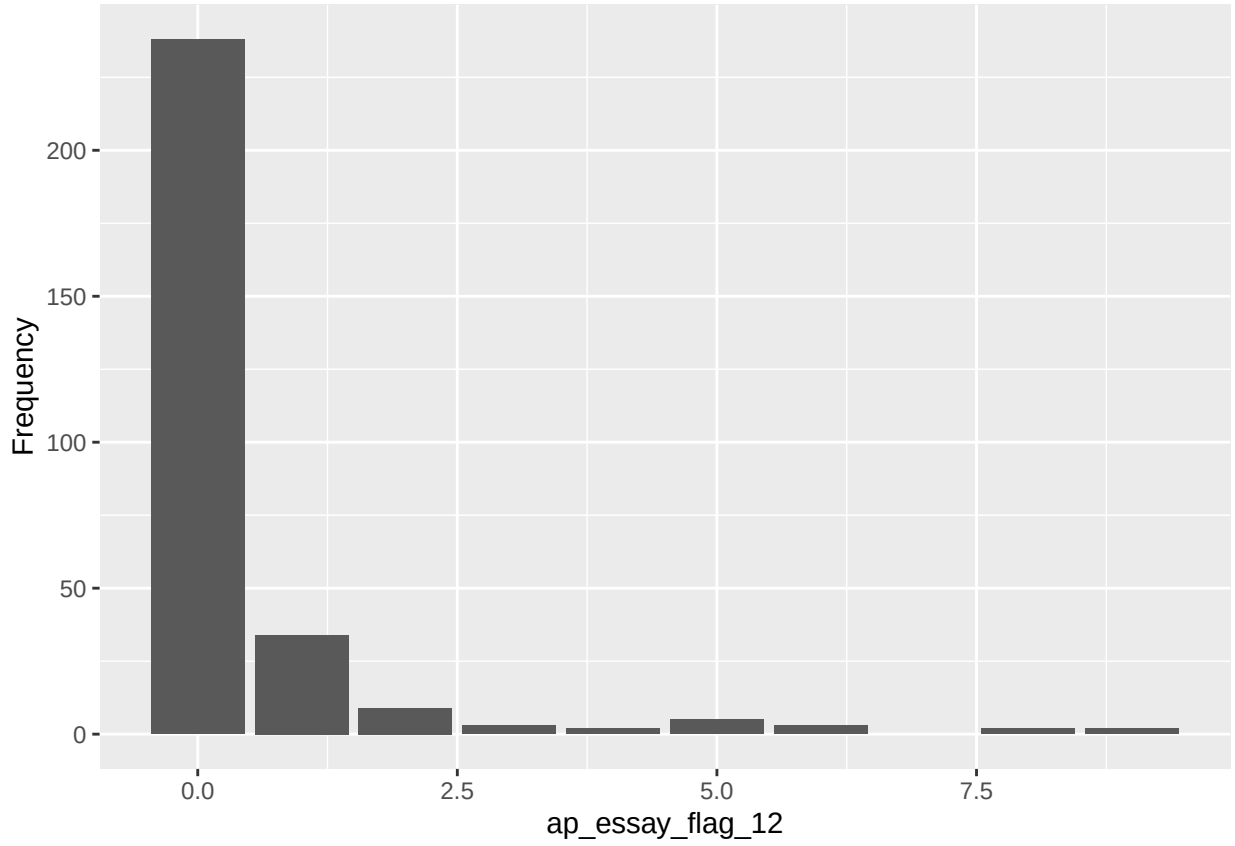

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_12, type: numeric
[1] Values (6 unique): NA, 3, 1, 5, 4, ...
[1] Missing: 98.9%
[1] Most missing: F09 100%, Least missing: F16 97.5%

Warning: Removed 45888 rows containing non-finite values (‘stat_count()‘).
```



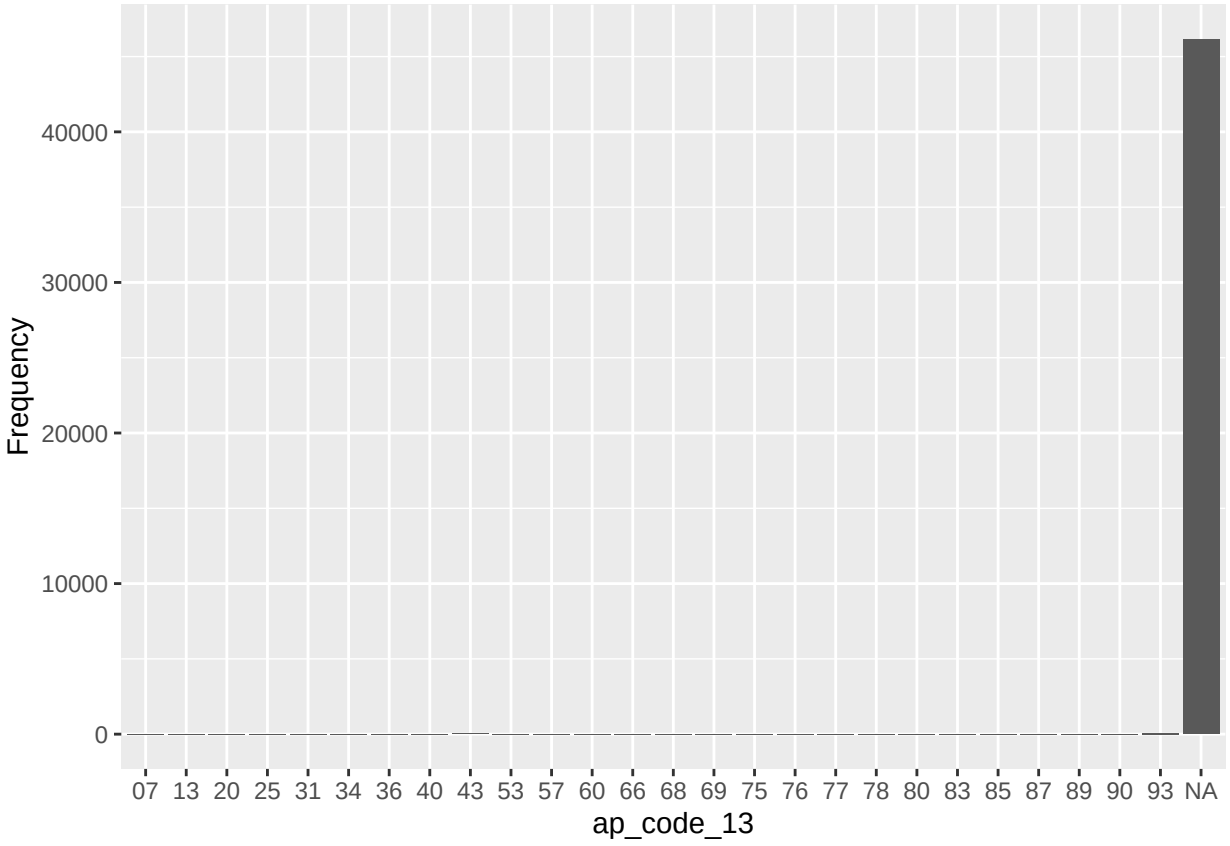

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_12, type: integer
[1] Values (10 unique): NA, 2013, 2014, 2015, 2011, ...
[1] Missing: 98.7%
[1] Most missing: F09 100%, Least missing: F16 97.4%

Warning: Removed 45823 rows containing non-finite values (‘stat_count()‘).
```



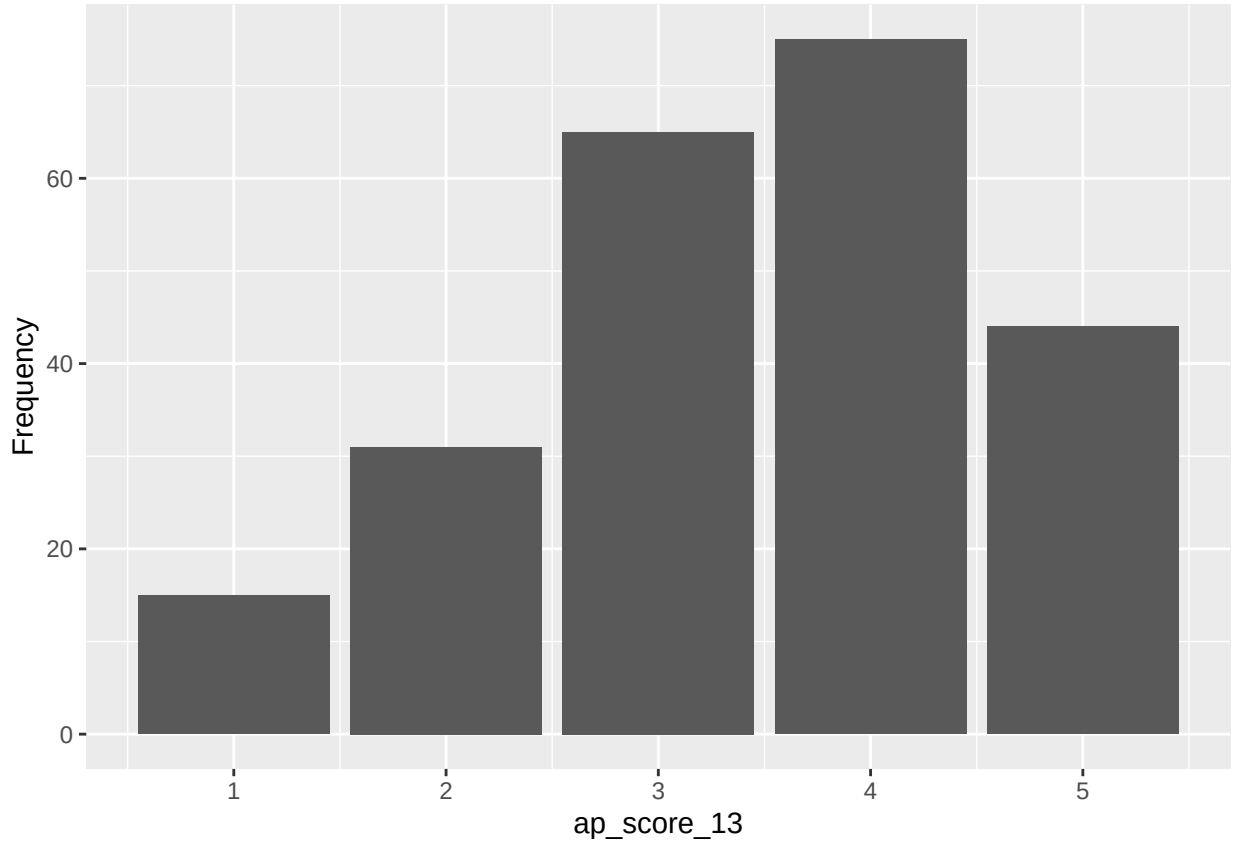

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_12, type: numeric
[1] Values (10 unique): NA, 0, 1, 5, 2, ...
[1] Missing: 99.4%

Warning: Removed 46110 rows containing non-finite values (‘stat_count()‘).
```



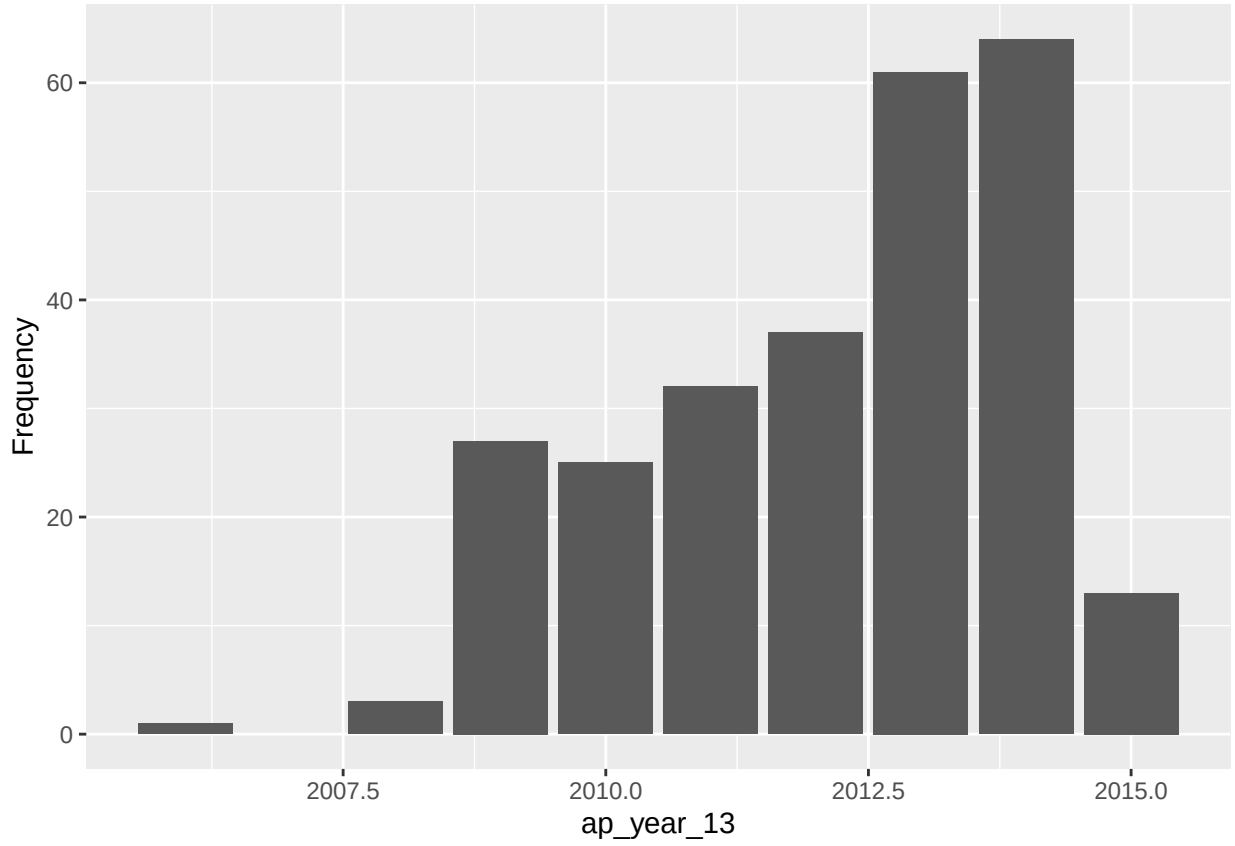

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_13, type: character
[1] Values (27 unique): NA, 83, 57, 93, 43, ...
[1] Missing: 99.4%
```



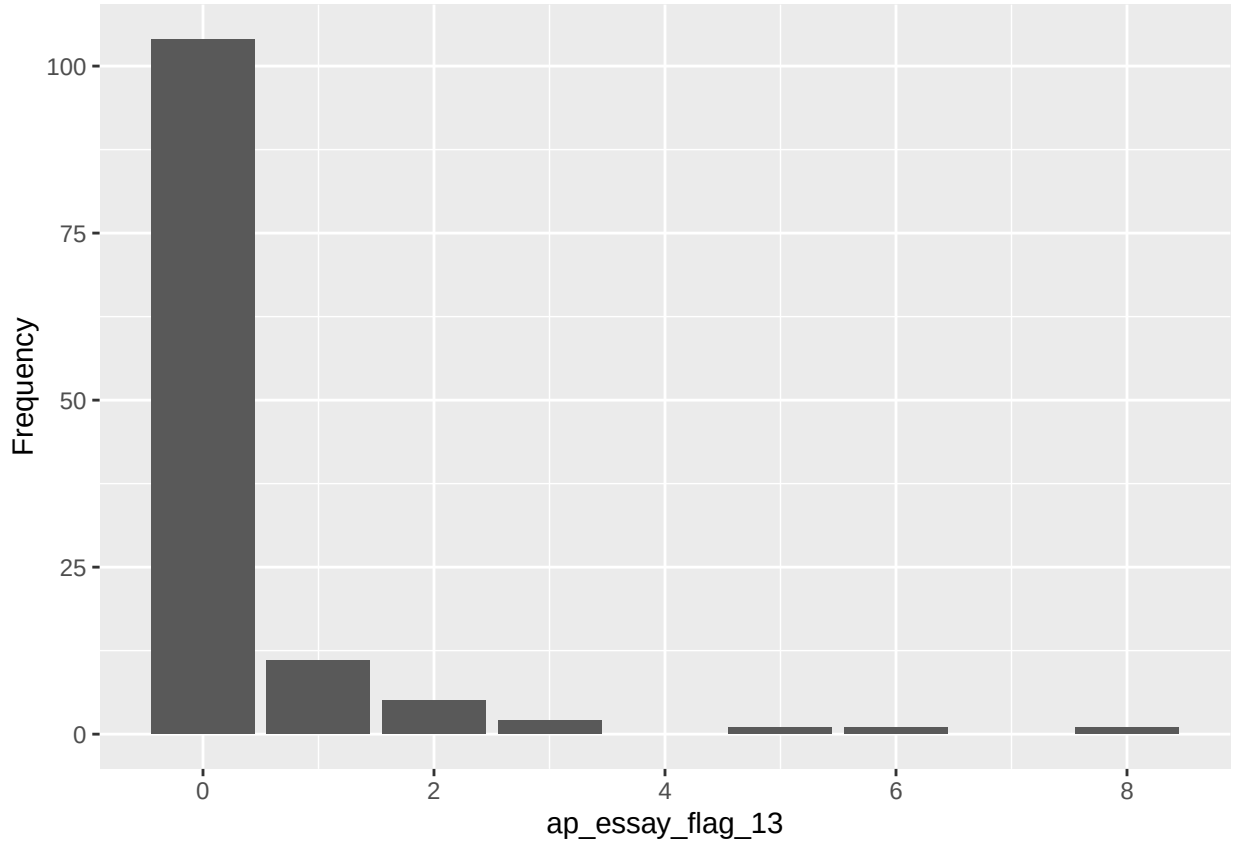

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_13, type: numeric
[1] Values (6 unique): NA, 4, 3, 5, 1, ...
[1] Missing: 99.5%

Warning: Removed 46178 rows containing non-finite values (‘stat_count()‘).
```



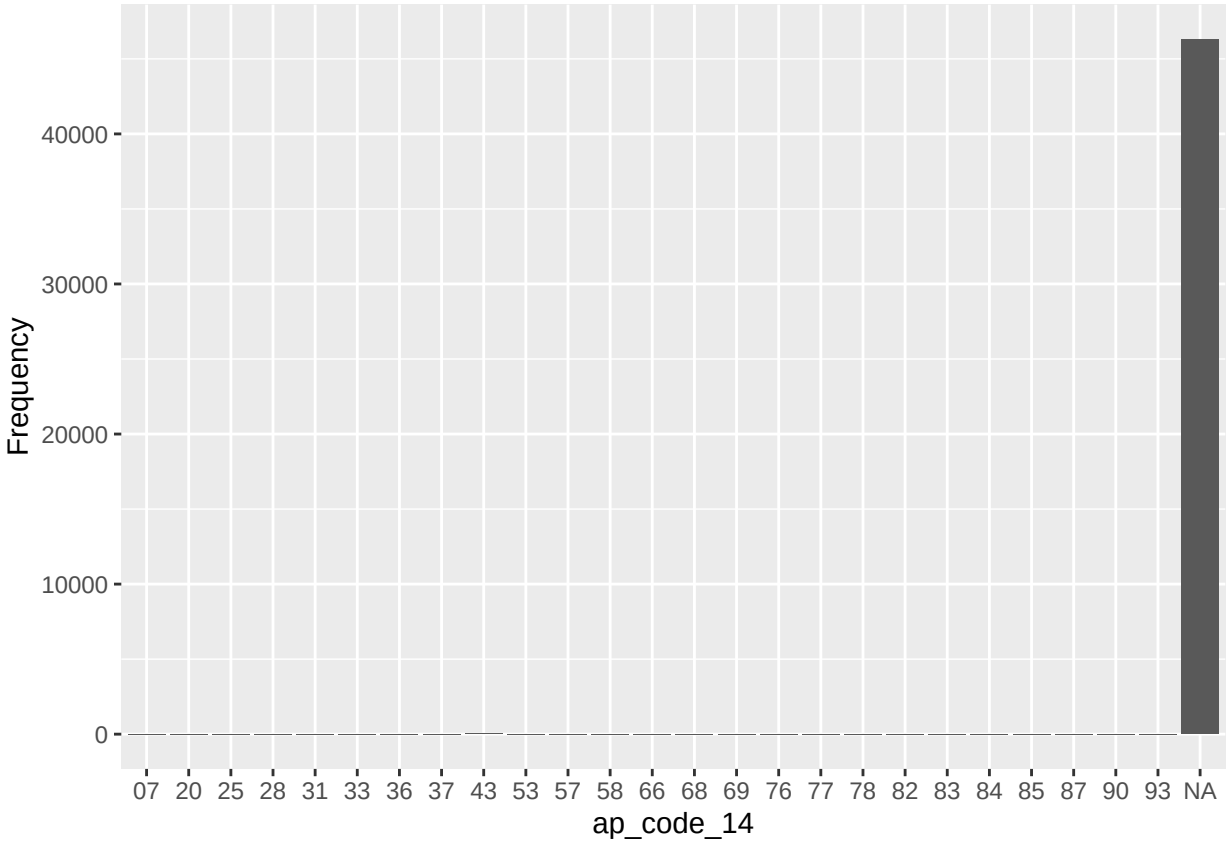

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_13, type: integer
[1] Values (10 unique): NA, 2015, 2014, 2013, 2012, ...
[1] Missing: 99.4%

Warning: Removed 46145 rows containing non-finite values (‘stat_count()‘).
```



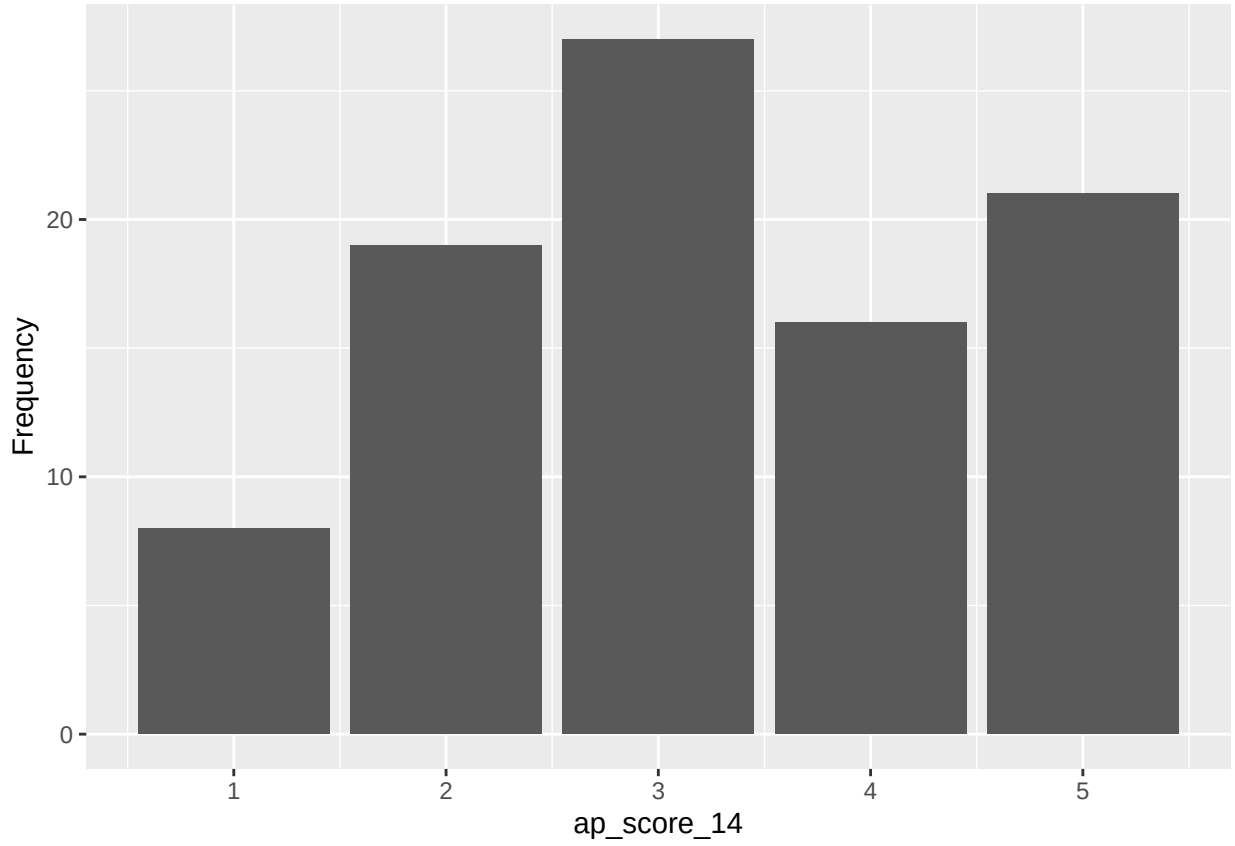

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_13, type: numeric
[1] Values (8 unique): NA, 0, 2, 1, 5, ...
[1] Missing: 99.7%

Warning: Removed 46283 rows containing non-finite values (‘stat_count()‘).
```



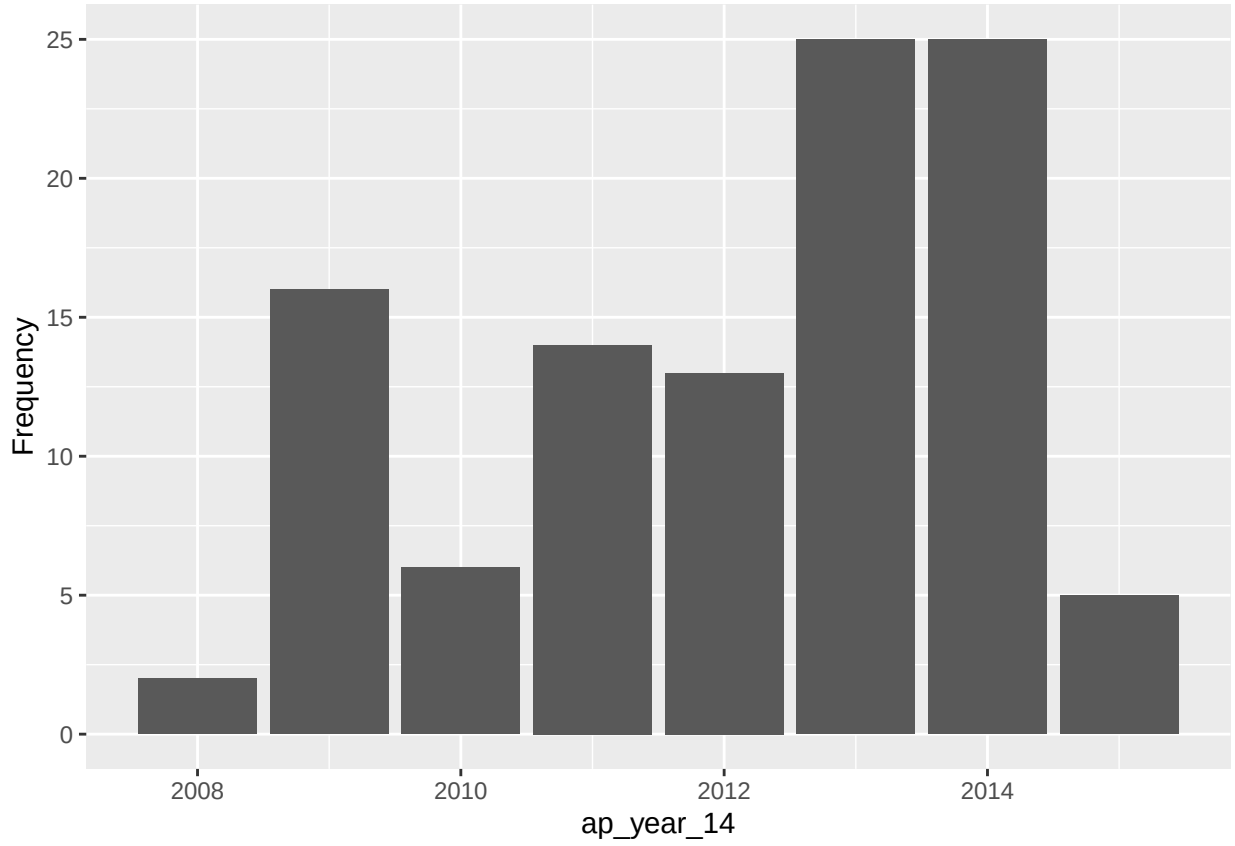

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_14, type: character
[1] Values (26 unique): NA, 43, 87, 28, 78, ...
[1] Missing: 99.8%
```



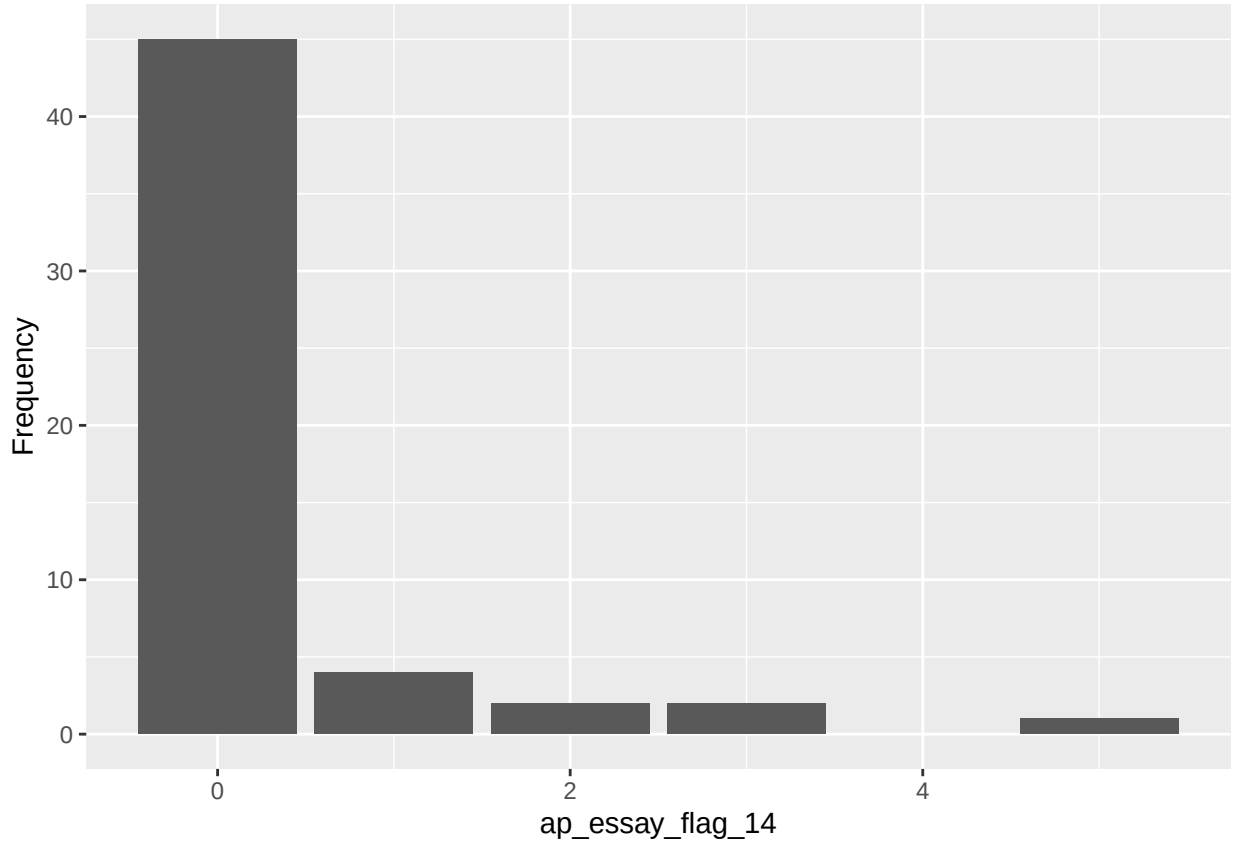

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_14, type: numeric
[1] Values (6 unique): NA, 4, 5, 1, 3, ...
[1] Missing: 99.8%

Warning: Removed 46317 rows containing non-finite values (‘stat_count()‘).
```



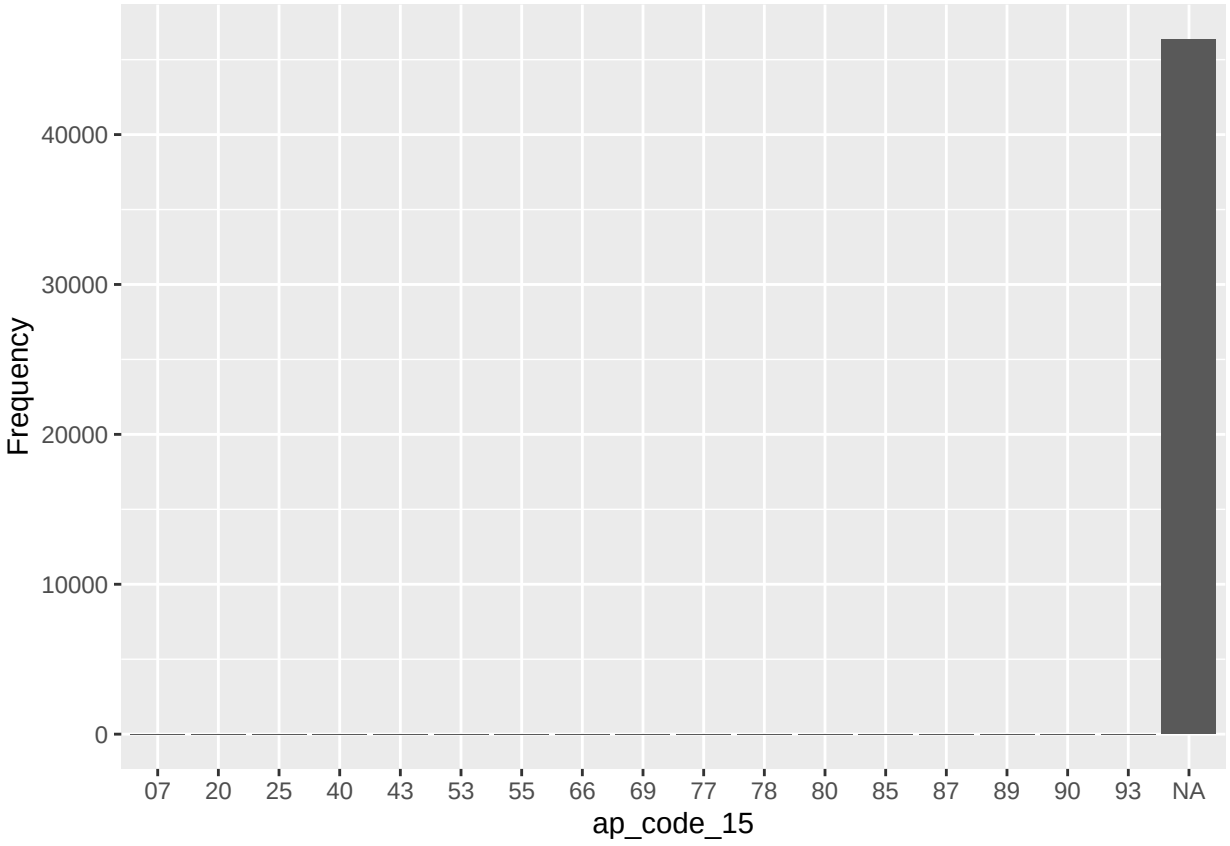

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_14, type: integer
[1] Values (9 unique): NA, 2014, 2013, 2012, 2015, ...
[1] Missing: 99.8%

Warning: Removed 46302 rows containing non-finite values (‘stat_count()‘).
```



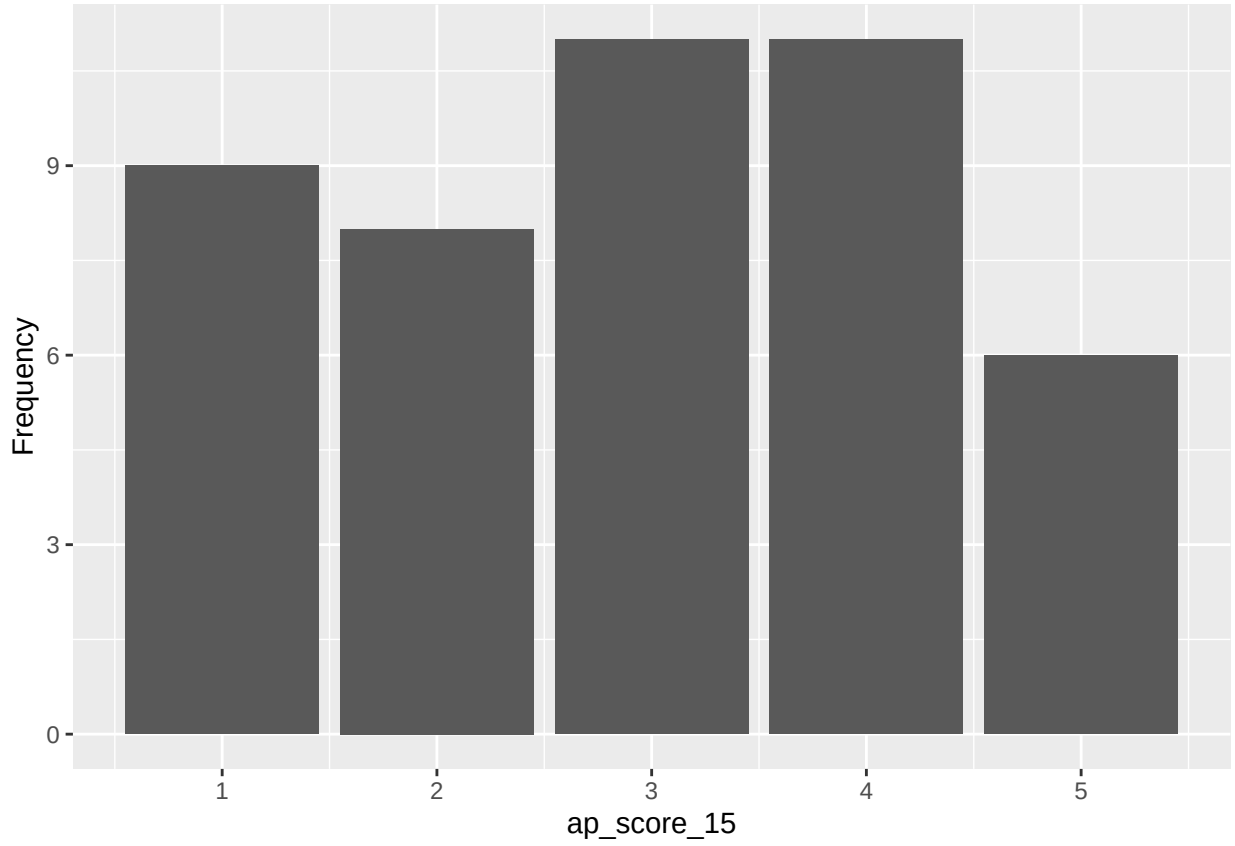

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_14, type: numeric
[1] Values (6 unique): NA, 0, 1, 2, 5, ...
[1] Missing: 99.9%

Warning: Removed 46354 rows containing non-finite values (‘stat_count()‘).
```



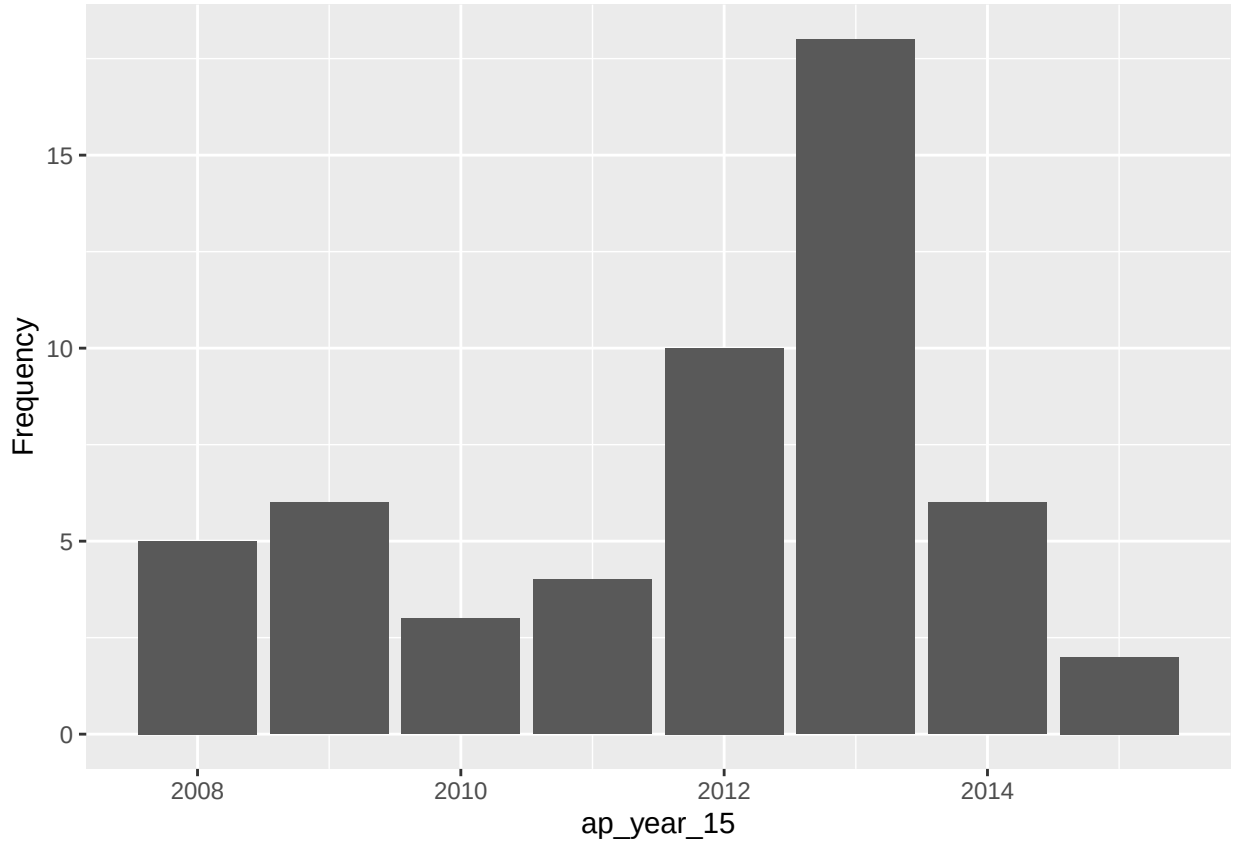

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_15, type: character
[1] Values (18 unique): NA, 66, 53, 90, 07, ...
[1] Missing: 99.9%
```



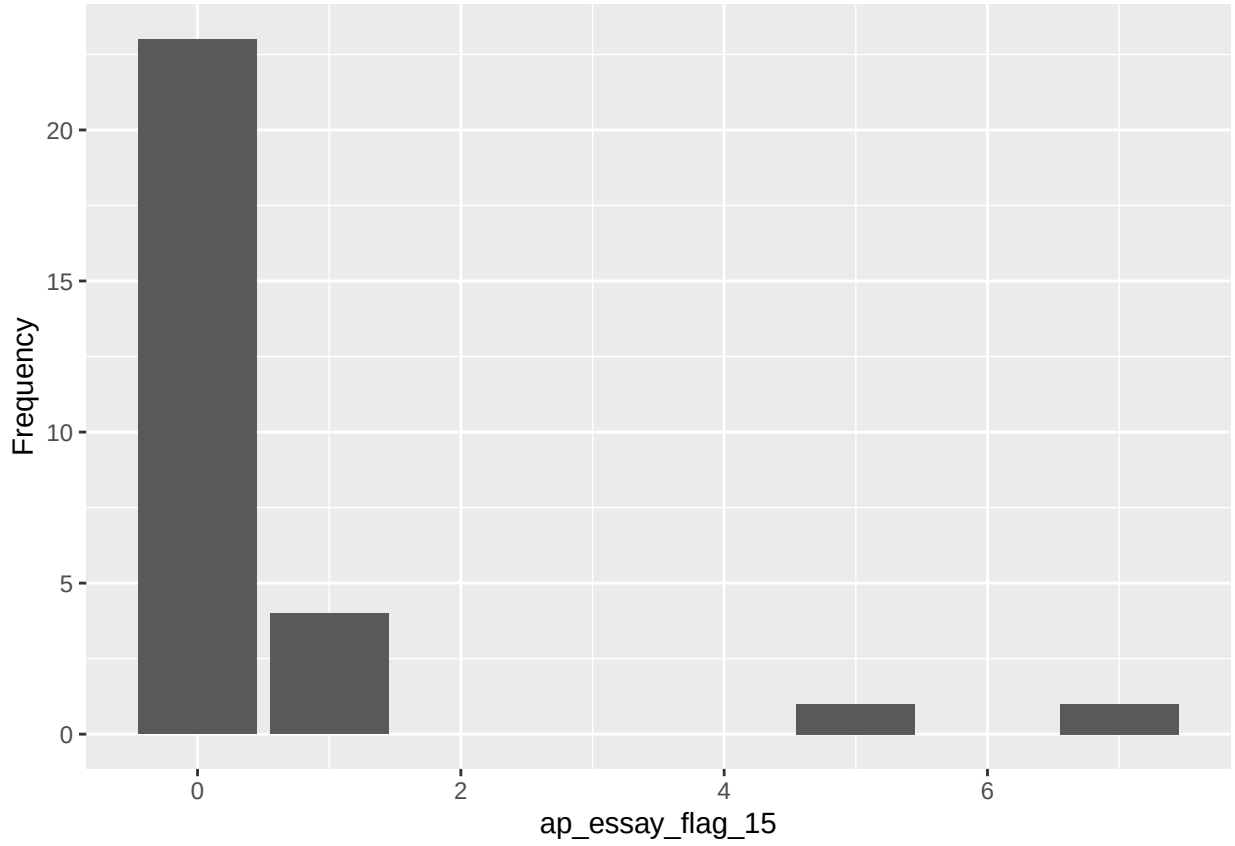

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_15, type: numeric
[1] Values (6 unique): NA, 2, 4, 3, 1, ...
[1] Missing: 99.9%

Warning: Removed 46363 rows containing non-finite values (‘stat_count()‘).
```



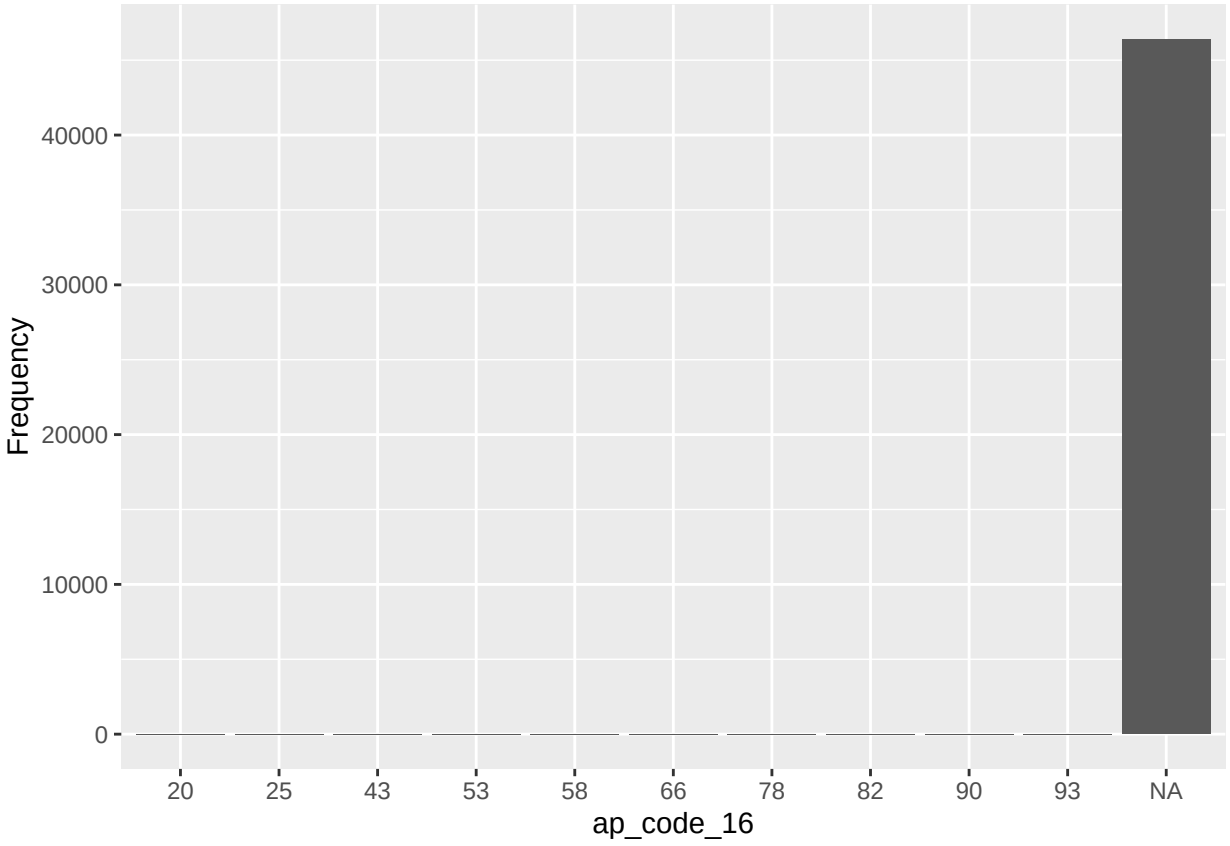

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_15, type: integer
[1] Values (9 unique): NA, 2013, 2014, 2012, 2015, ...
[1] Missing: 99.9%

Warning: Removed 46354 rows containing non-finite values (‘stat_count()‘).
```



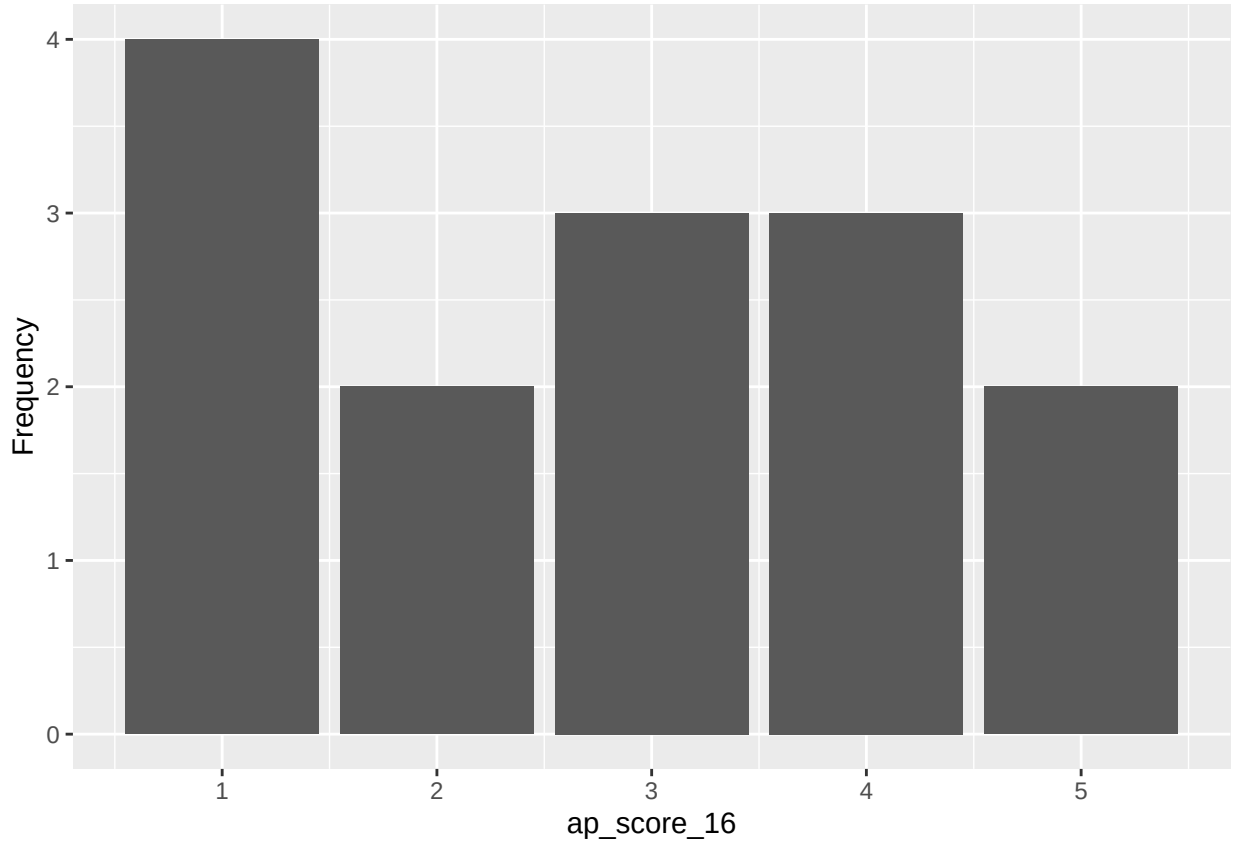

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_15, type: numeric
[1] Values (5 unique): NA, 0, 7, 5, 1
[1] Missing: 99.9%

Warning: Removed 46379 rows containing non-finite values (‘stat_count()‘).
```



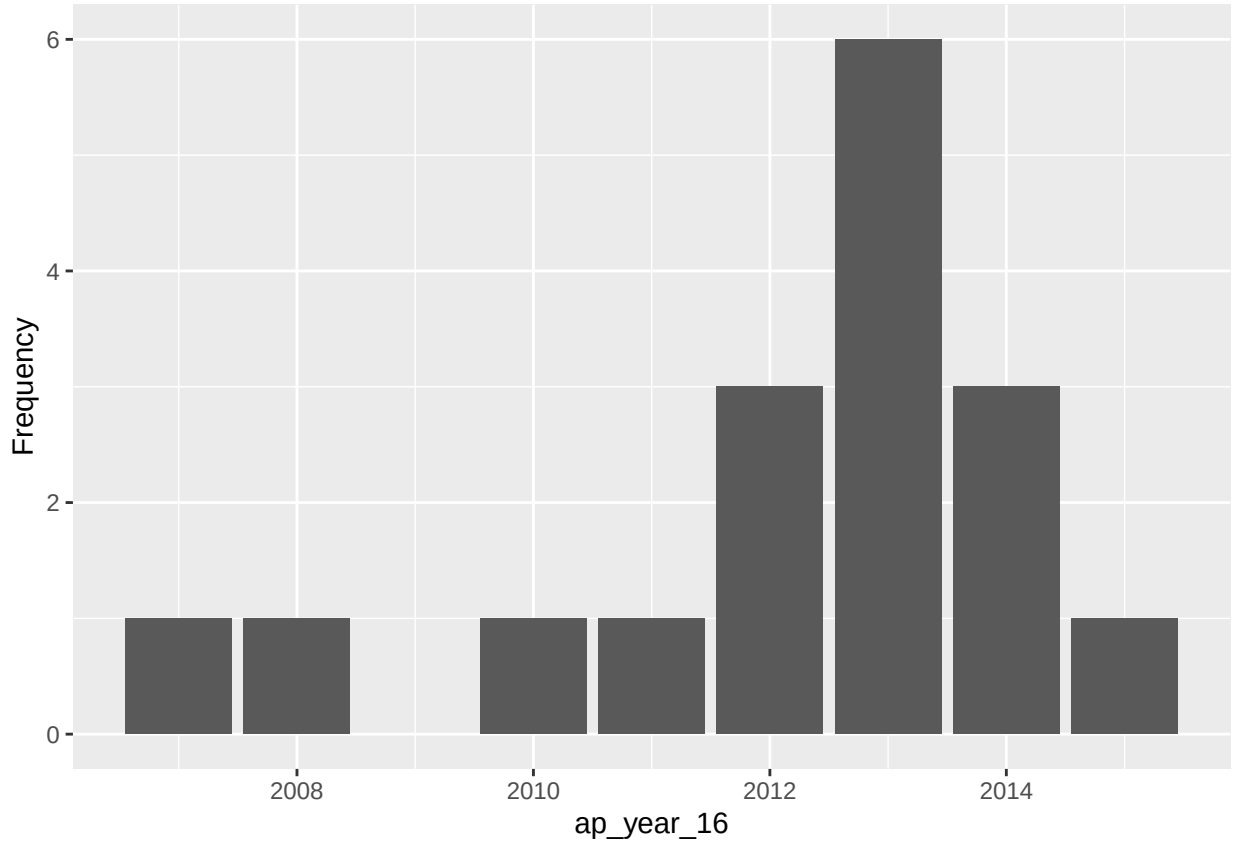

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_16, type: character
[1] Values (11 unique): NA, 93, 25, 43, 90, ...
[1] Missing: 100%
```



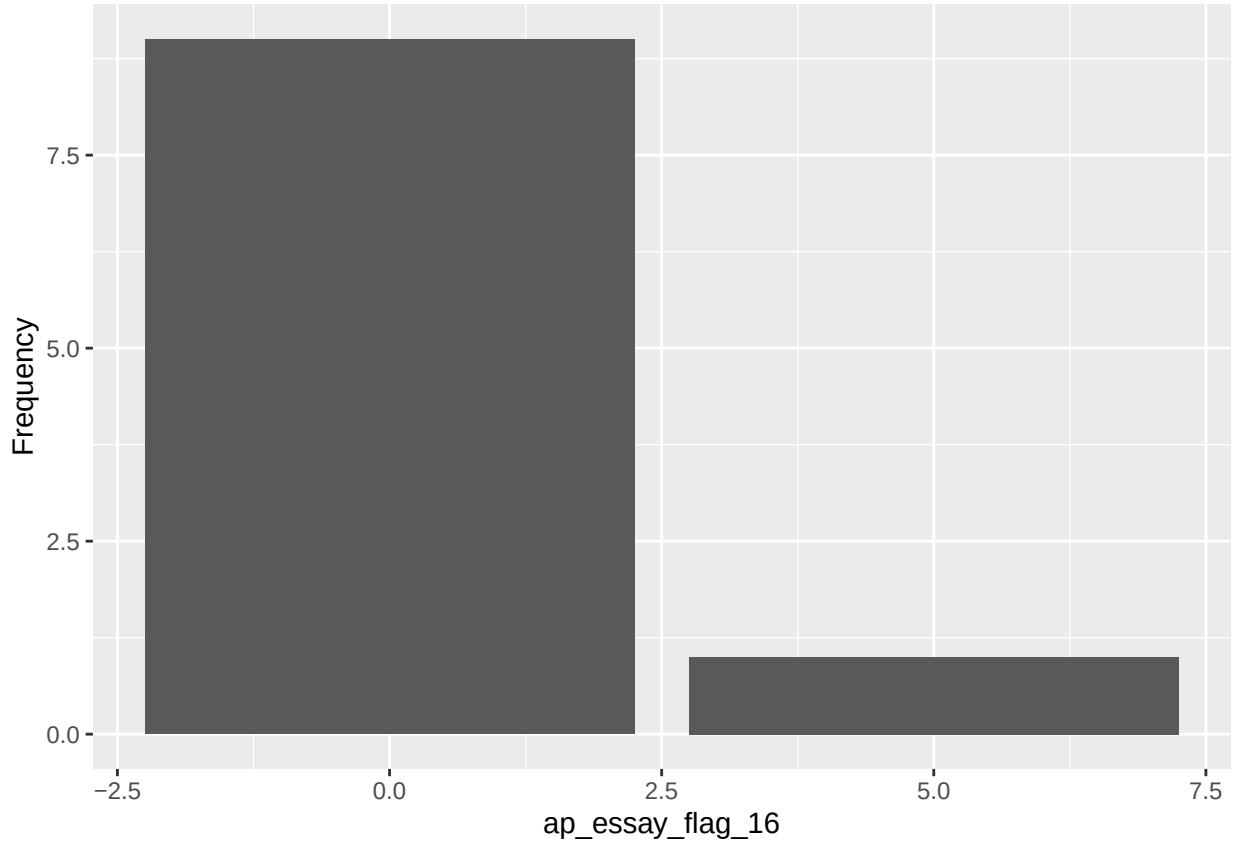

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_16, type: numeric
[1] Values (6 unique): NA, 3, 1, 4, 2, ...
[1] Missing: 100%

Warning: Removed 46394 rows containing non-finite values (‘stat_count()‘).
```



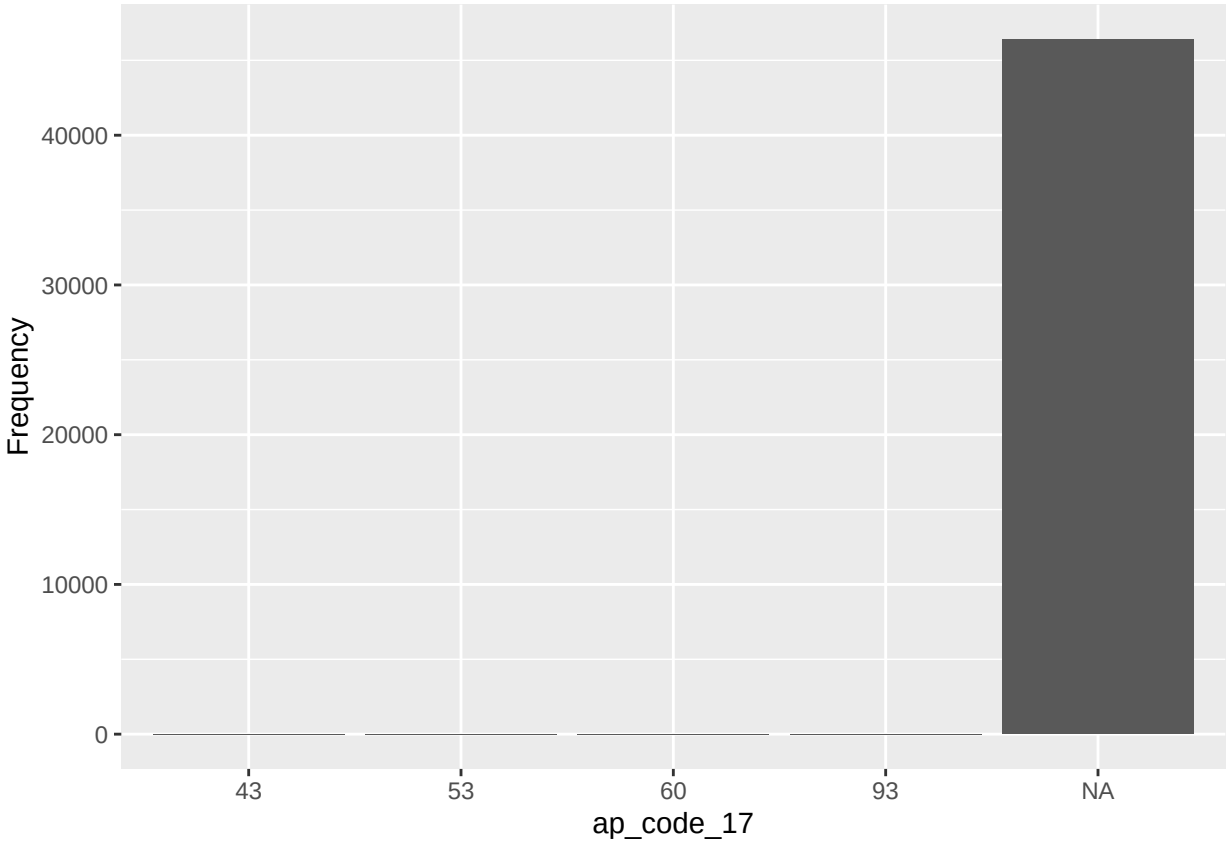

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_16, type: integer
[1] Values (9 unique): NA, 2013, 2014, 2015, 2012, ...
[1] Missing: 100%

Warning: Removed 46391 rows containing non-finite values (‘stat_count()‘).
```



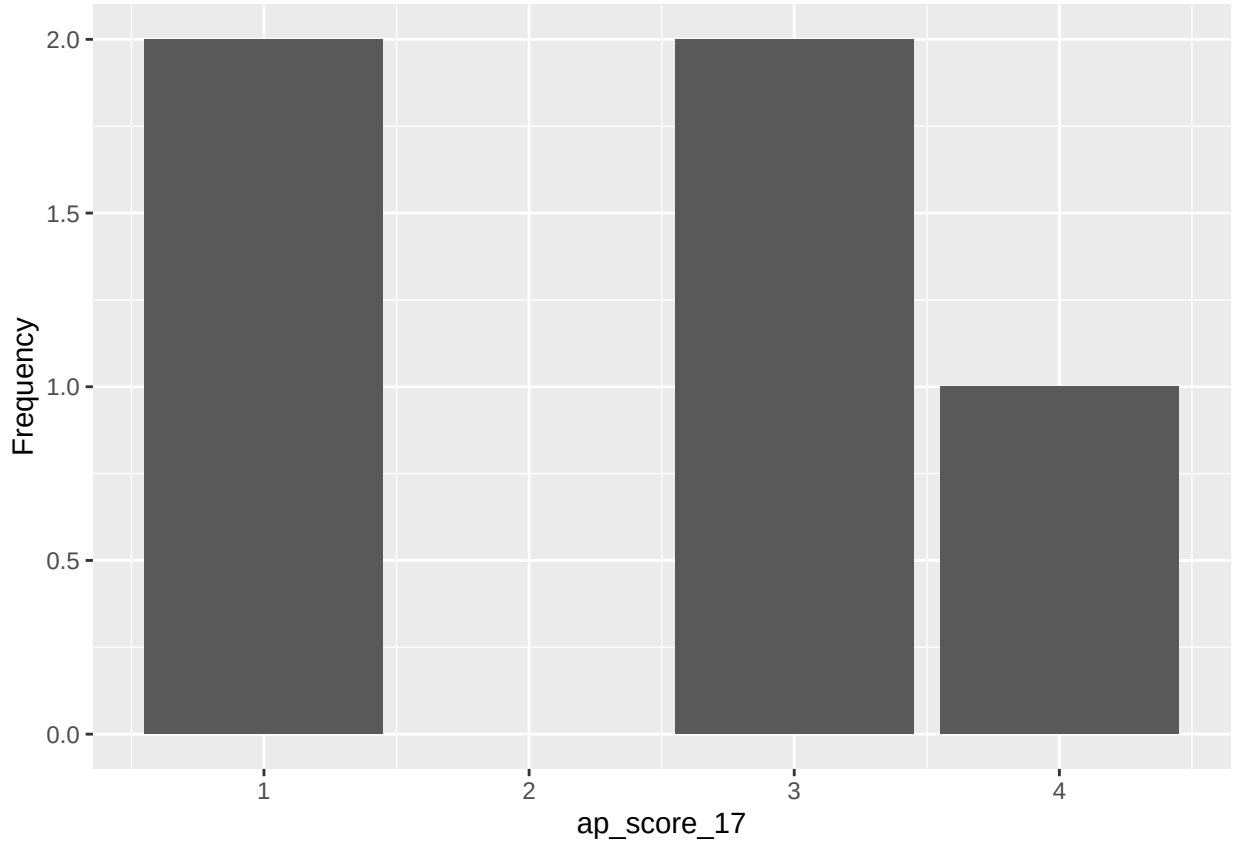

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_16, type: numeric
[1] Values (3 unique): NA, 0, 5
[1] Missing: 100%

Warning: Removed 46398 rows containing non-finite values (‘stat_count()‘).
```



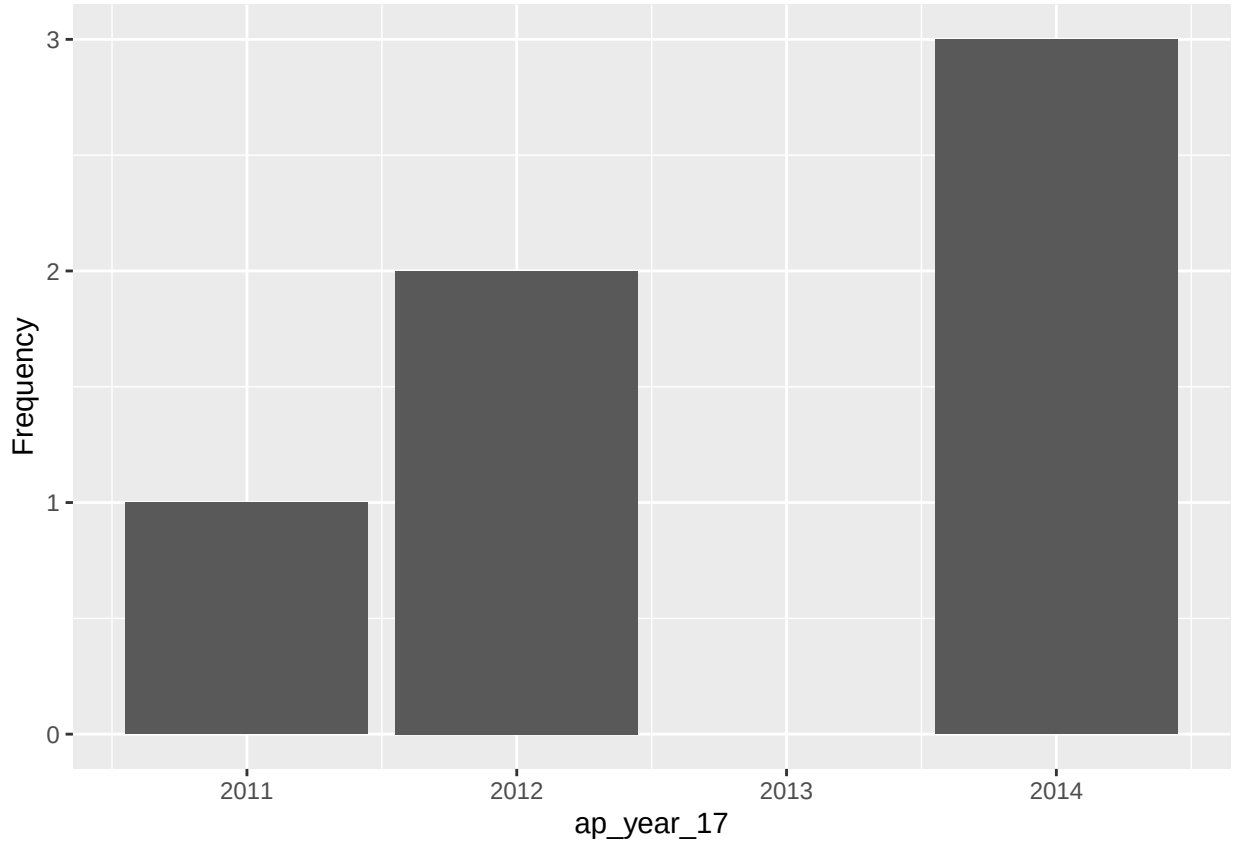

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_17, type: character
[1] Values (5 unique): NA, 93, 53, 43, 60
[1] Missing: 100%
```



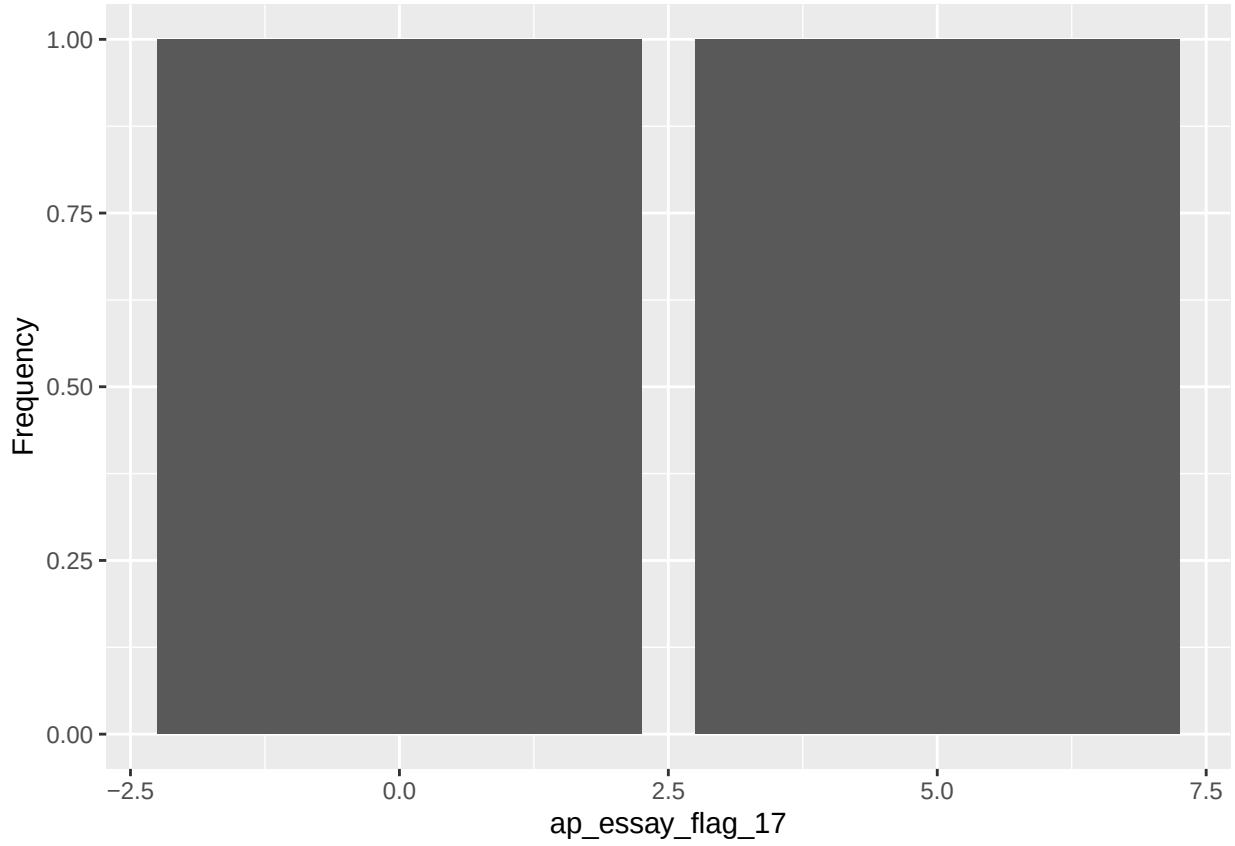

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_17, type: numeric
[1] Values (4 unique): NA, 1, 3, 4
[1] Missing: 100%

Warning: Removed 46403 rows containing non-finite values (‘stat_count()‘).
```



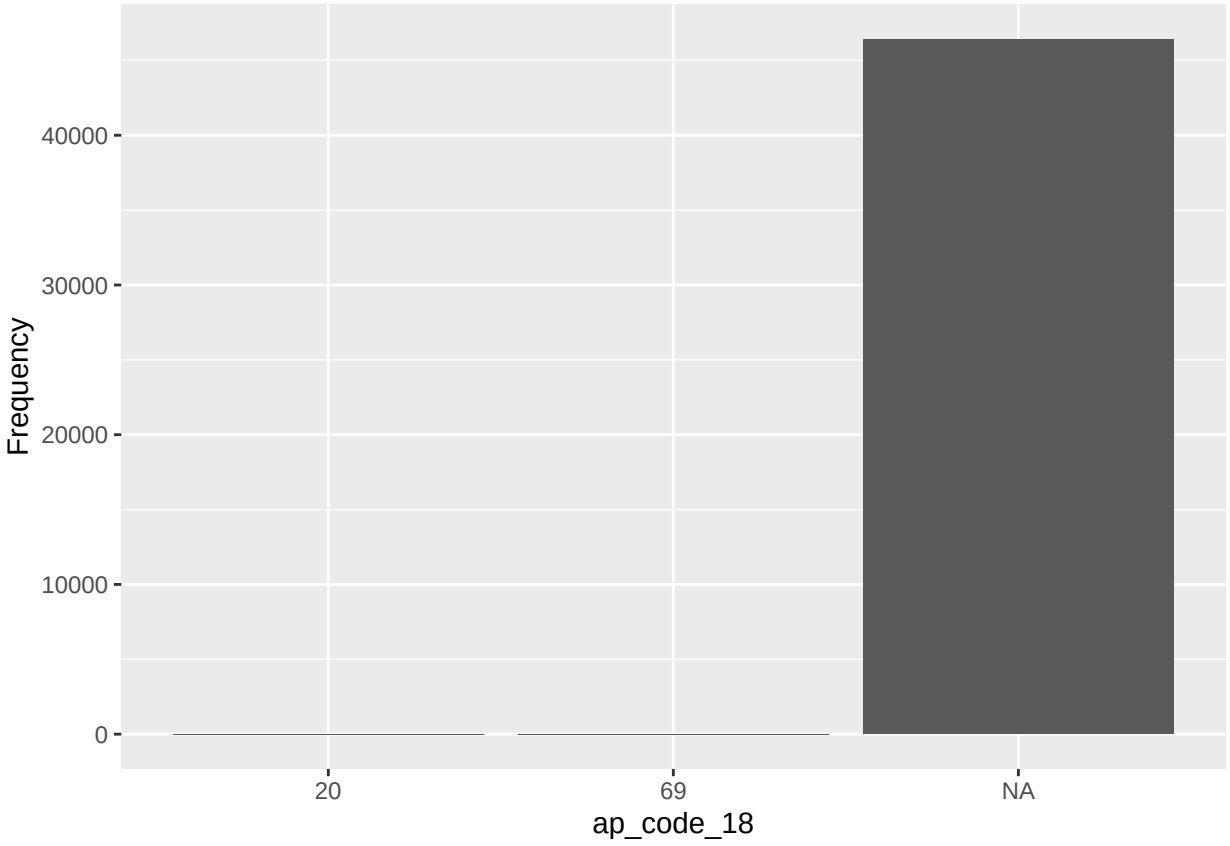

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_17, type: integer
[1] Values (4 unique): NA, 2012, 2014, 2011
[1] Missing: 100%

Warning: Removed 46402 rows containing non-finite values (‘stat_count()‘).
```



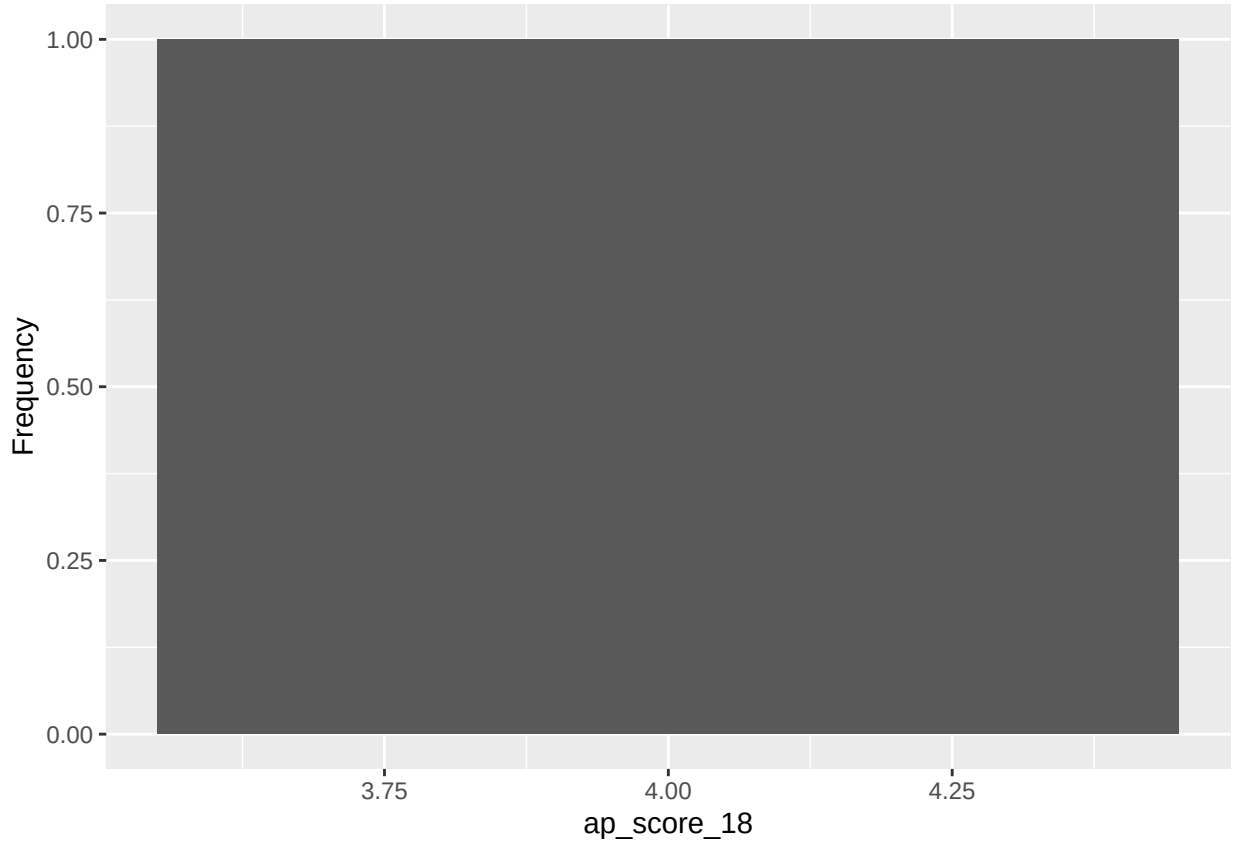

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_17, type: numeric
[1] Values (3 unique): NA, 5, 0
[1] Missing: 100%

Warning: Removed 46406 rows containing non-finite values (‘stat_count()‘).
```



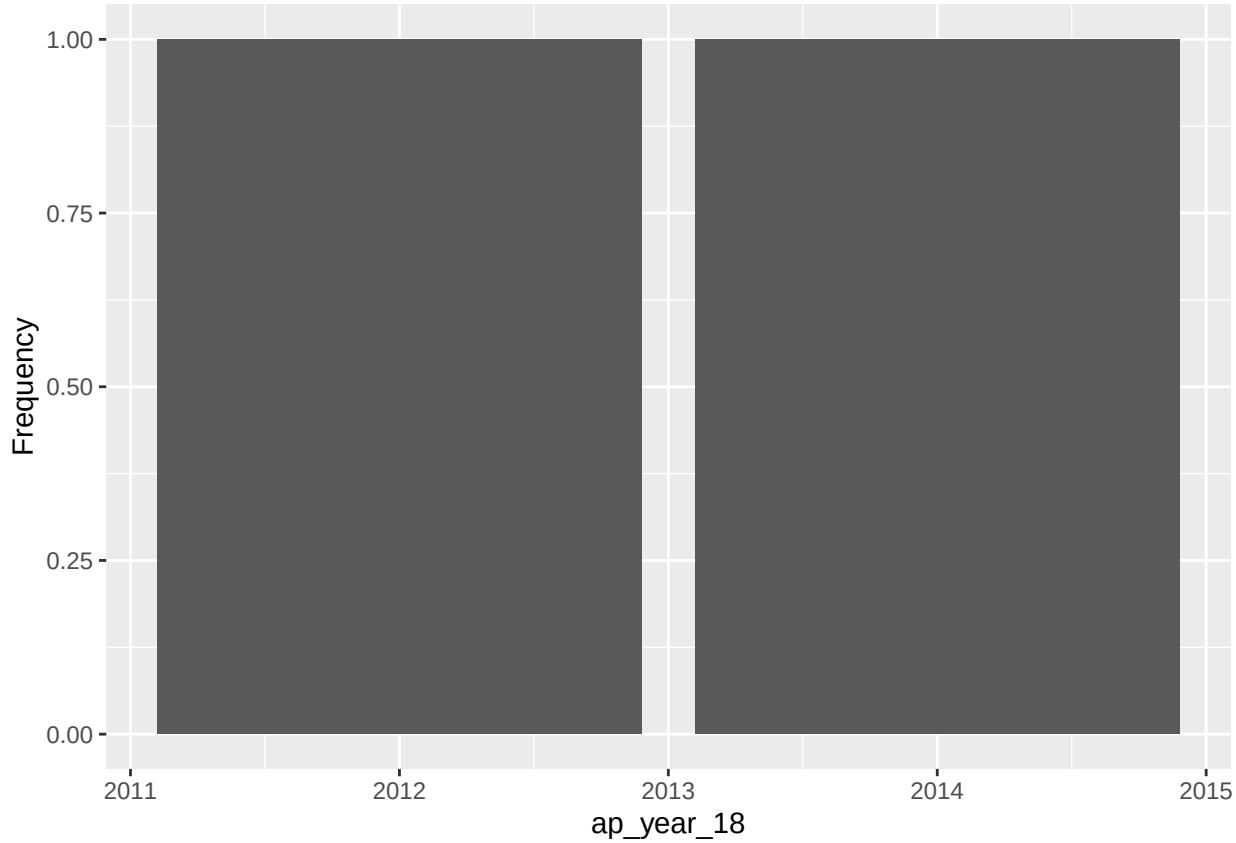

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_18, type: character
[1] Values (3 unique): NA, 20, 69
[1] Missing: 100%
```



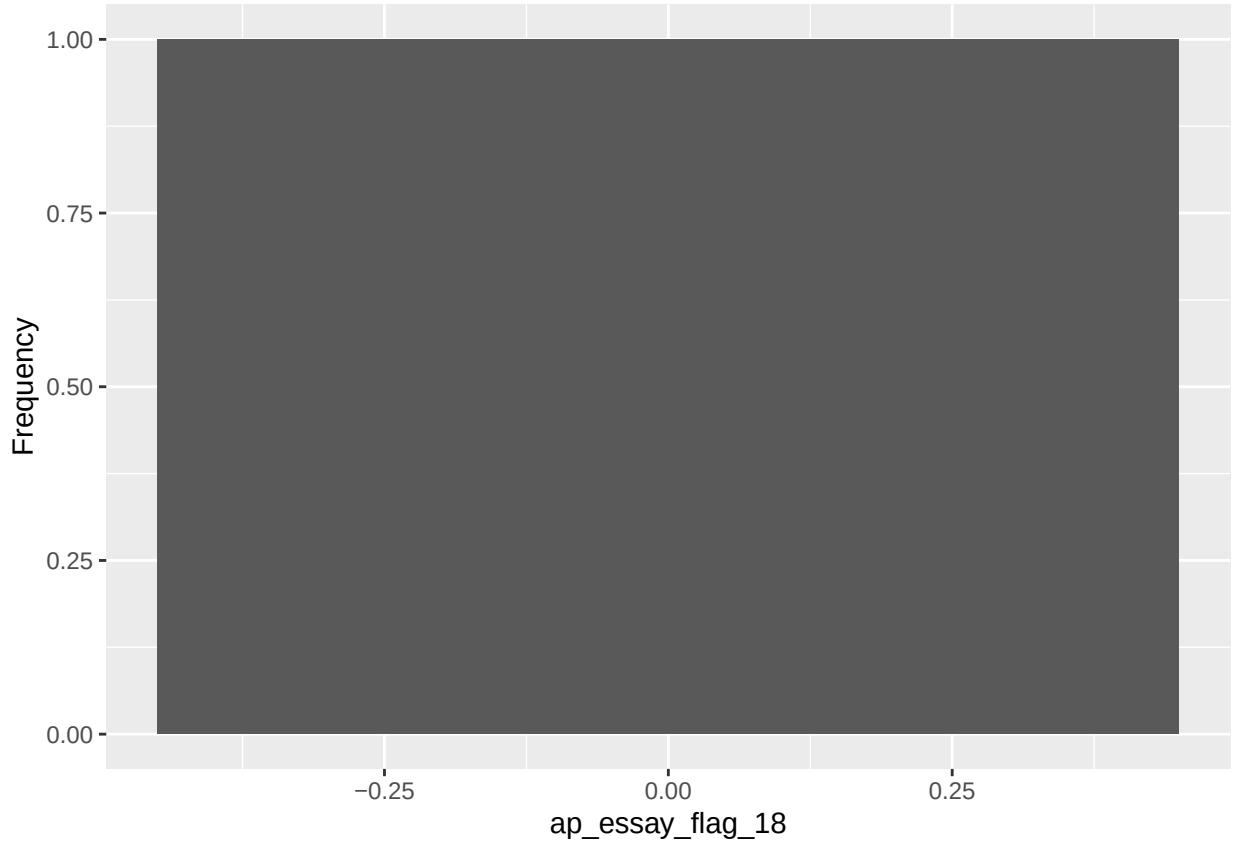

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_18, type: numeric
[1] Values (2 unique): NA, 4
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



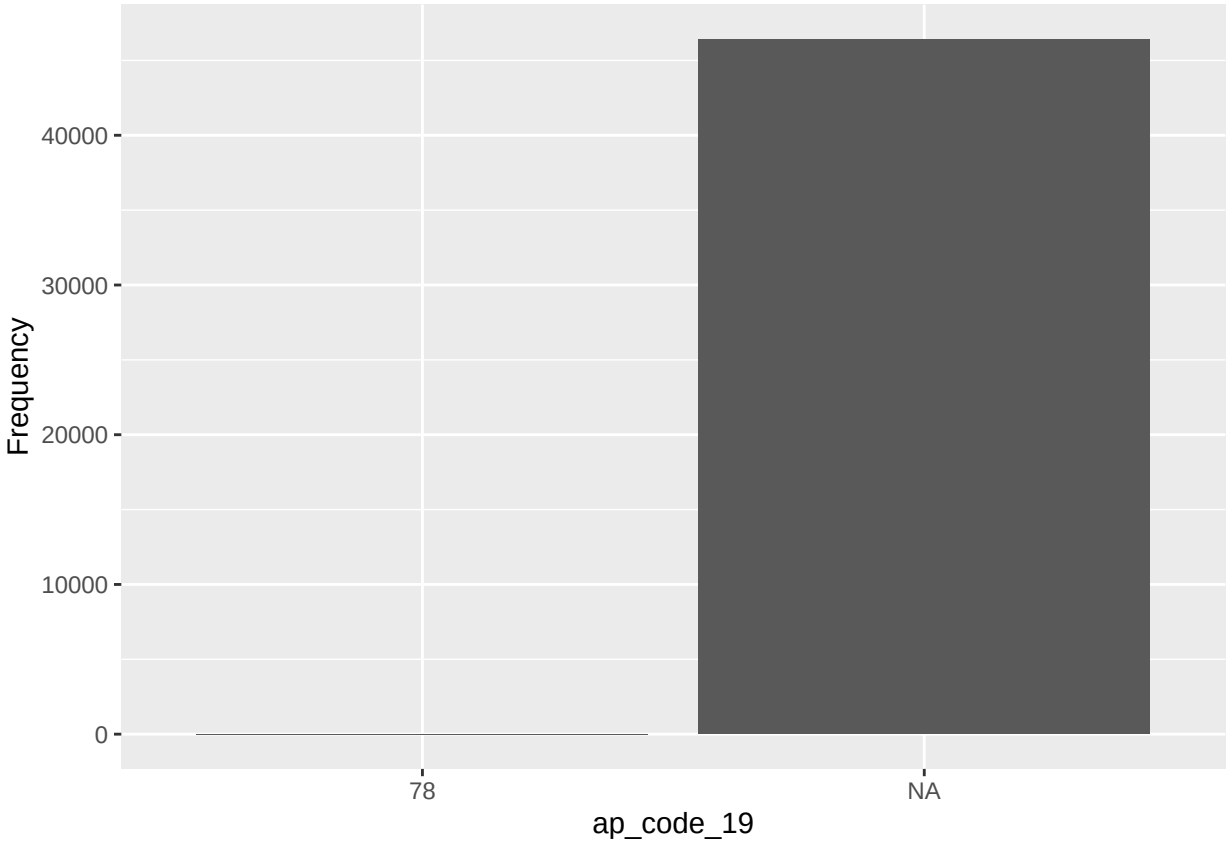

```
[1] is used in feature engineering and hence not included
[1] ----------------------------
[1] Variable: ap_year_18, type: integer
[1] Values (3 unique): NA, 2012, 2014
[1] Missing: 100%

Warning: Removed 46406 rows containing non-finite values (‘stat_count()‘).
```



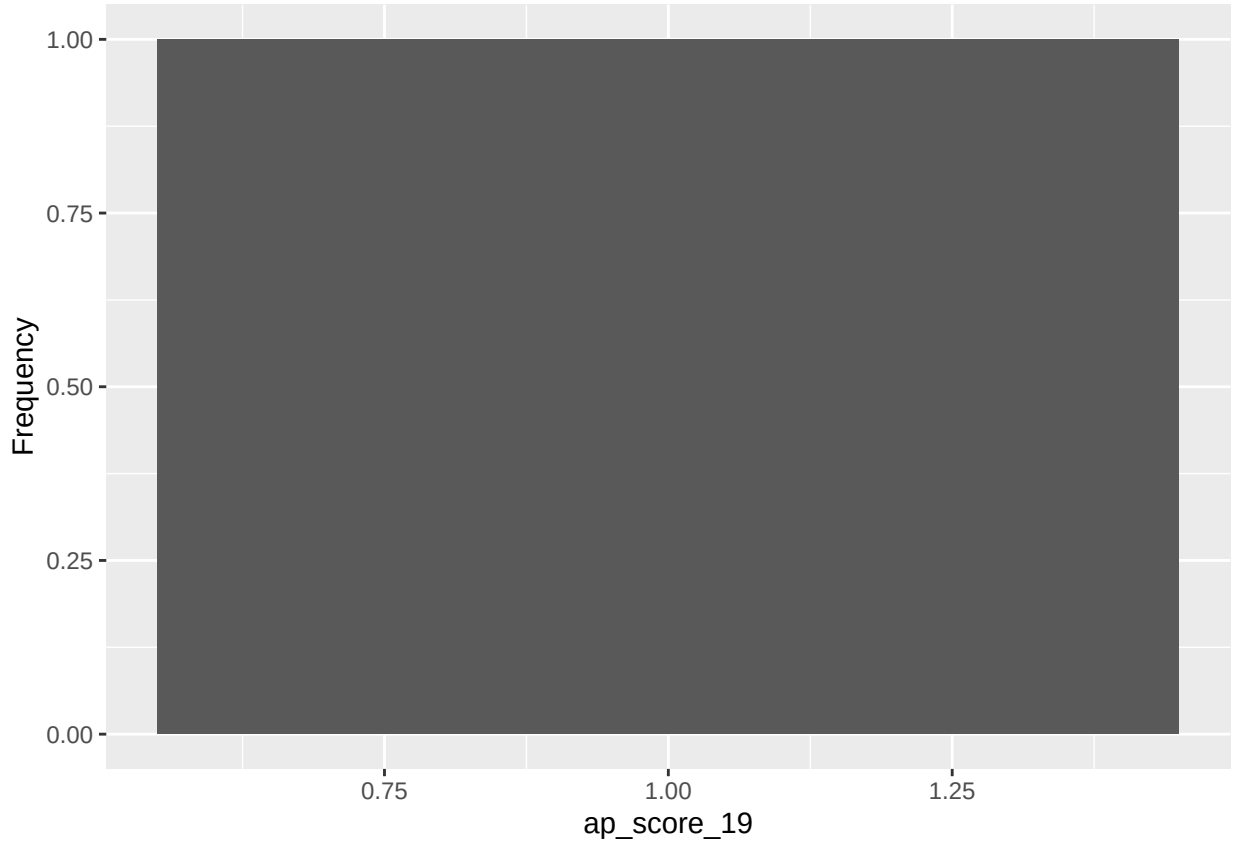

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_18, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



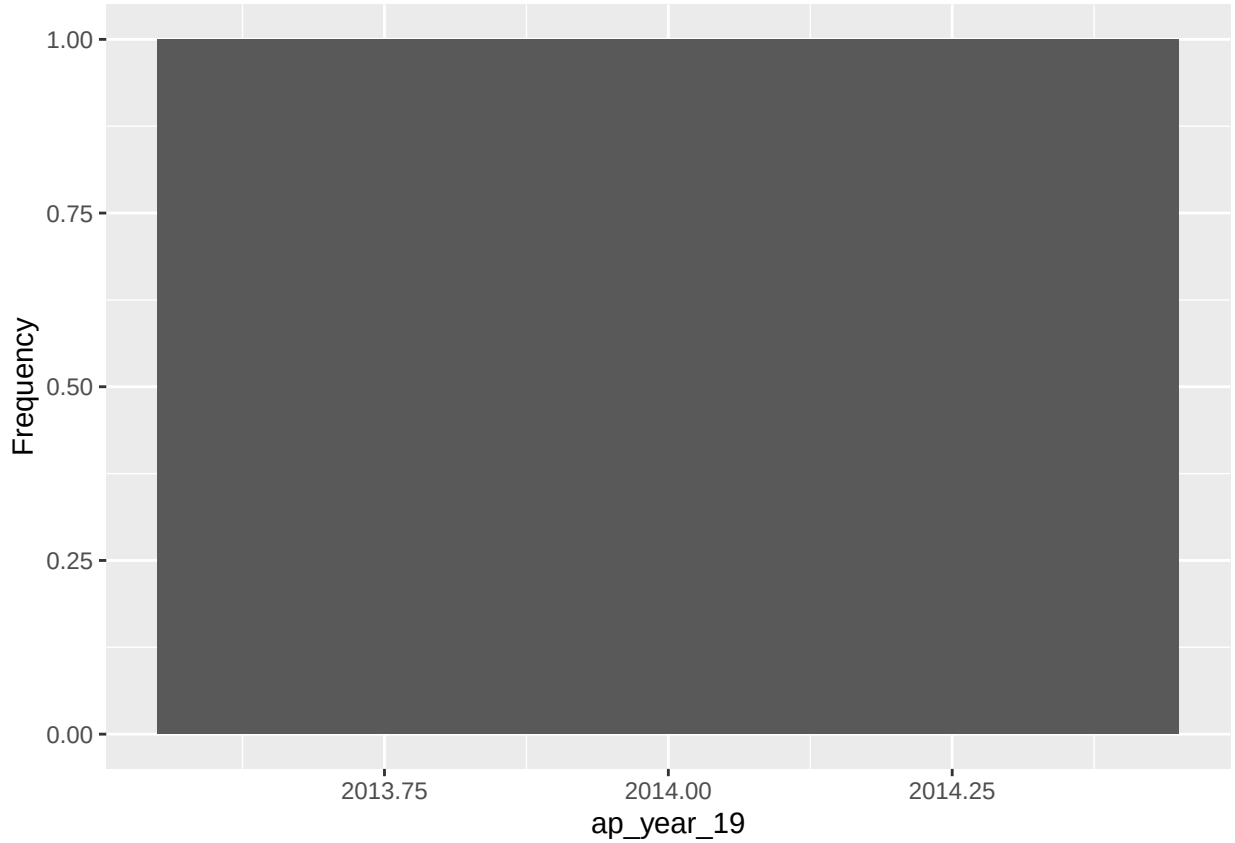

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_19, type: character
[1] Values (2 unique): NA, 78
[1] Missing: 100%
```



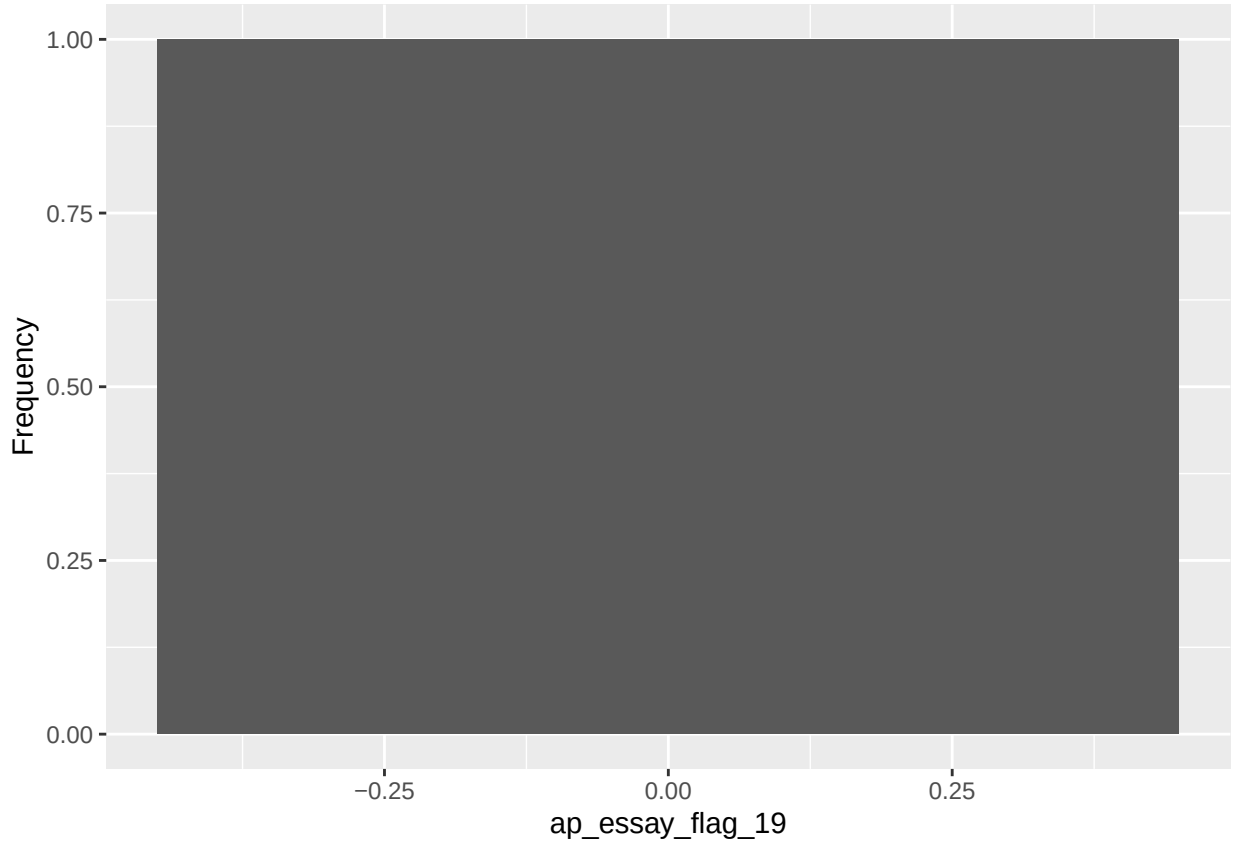

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_19, type: numeric
[1] Values (2 unique): NA, 1
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



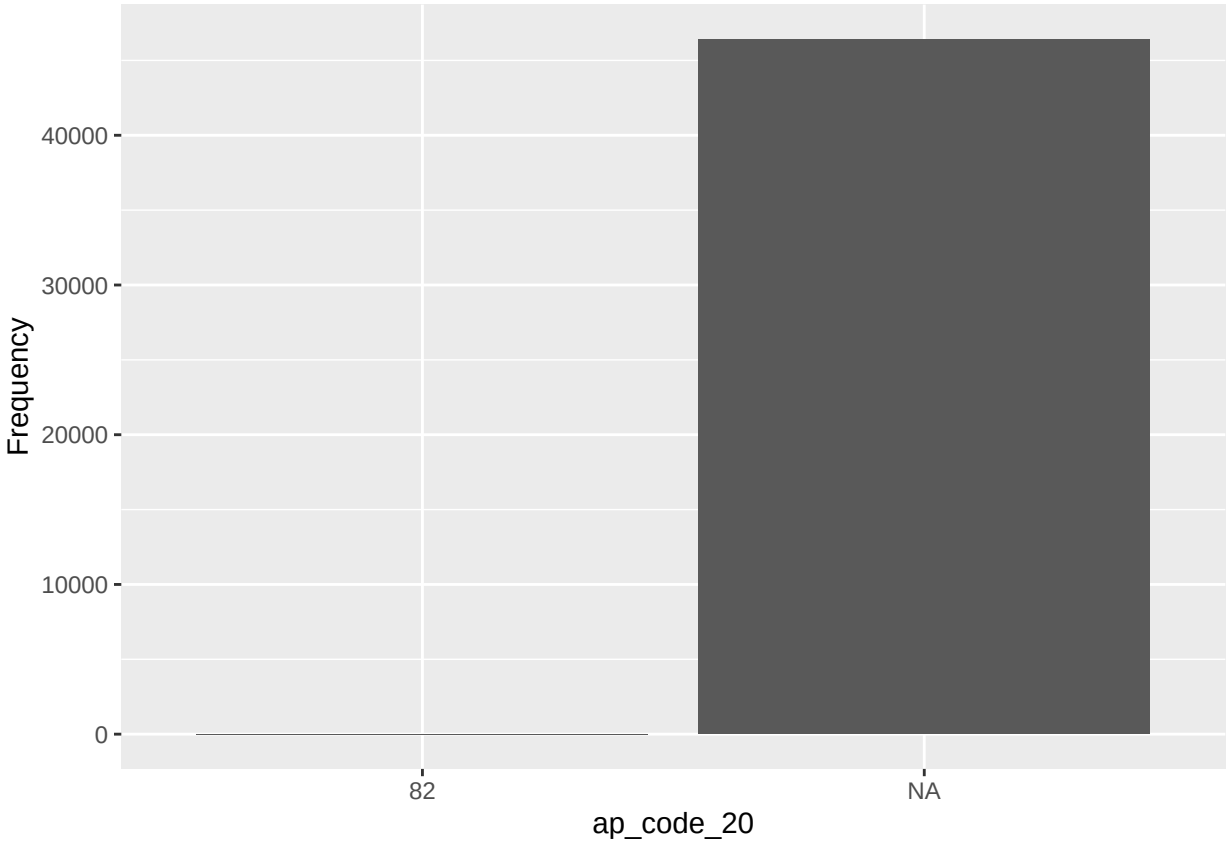

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_19, type: integer
[1] Values (2 unique): NA, 2014
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



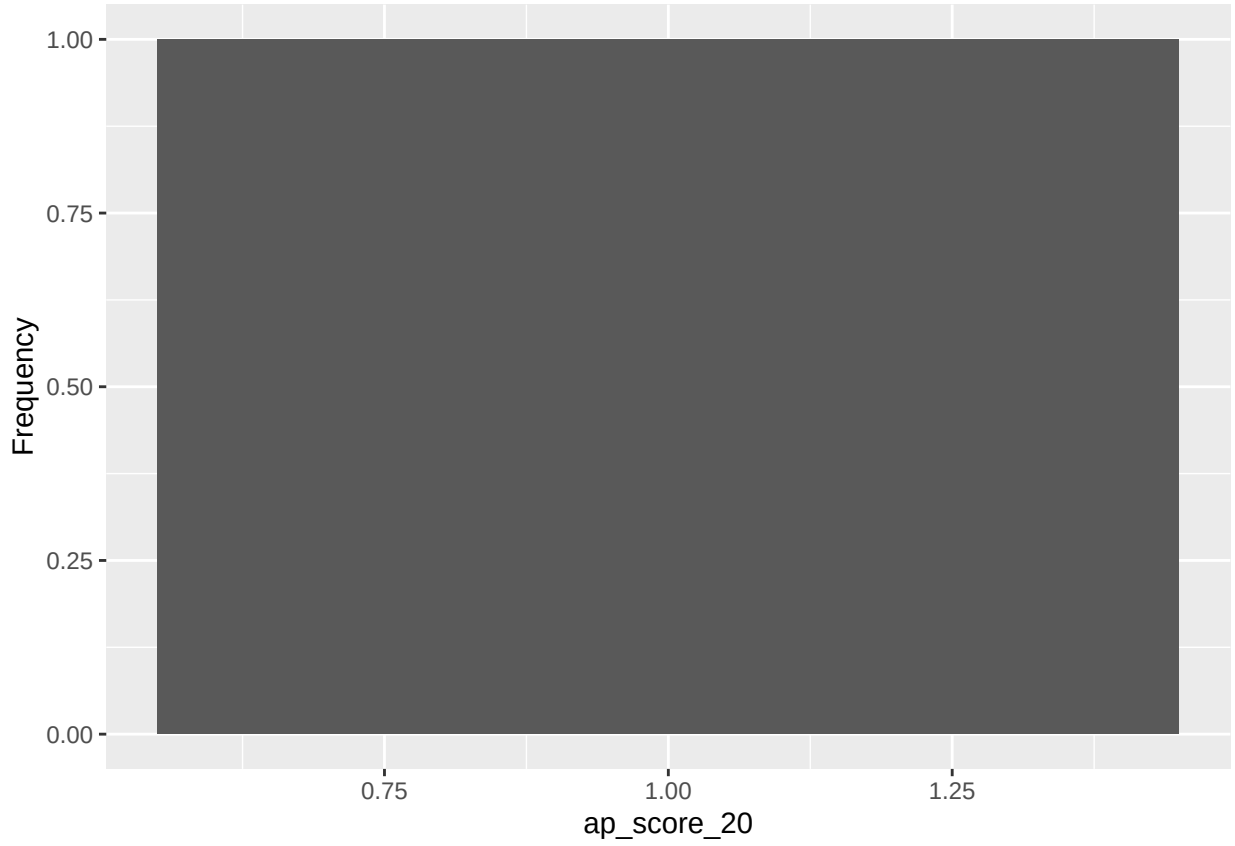

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_19, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



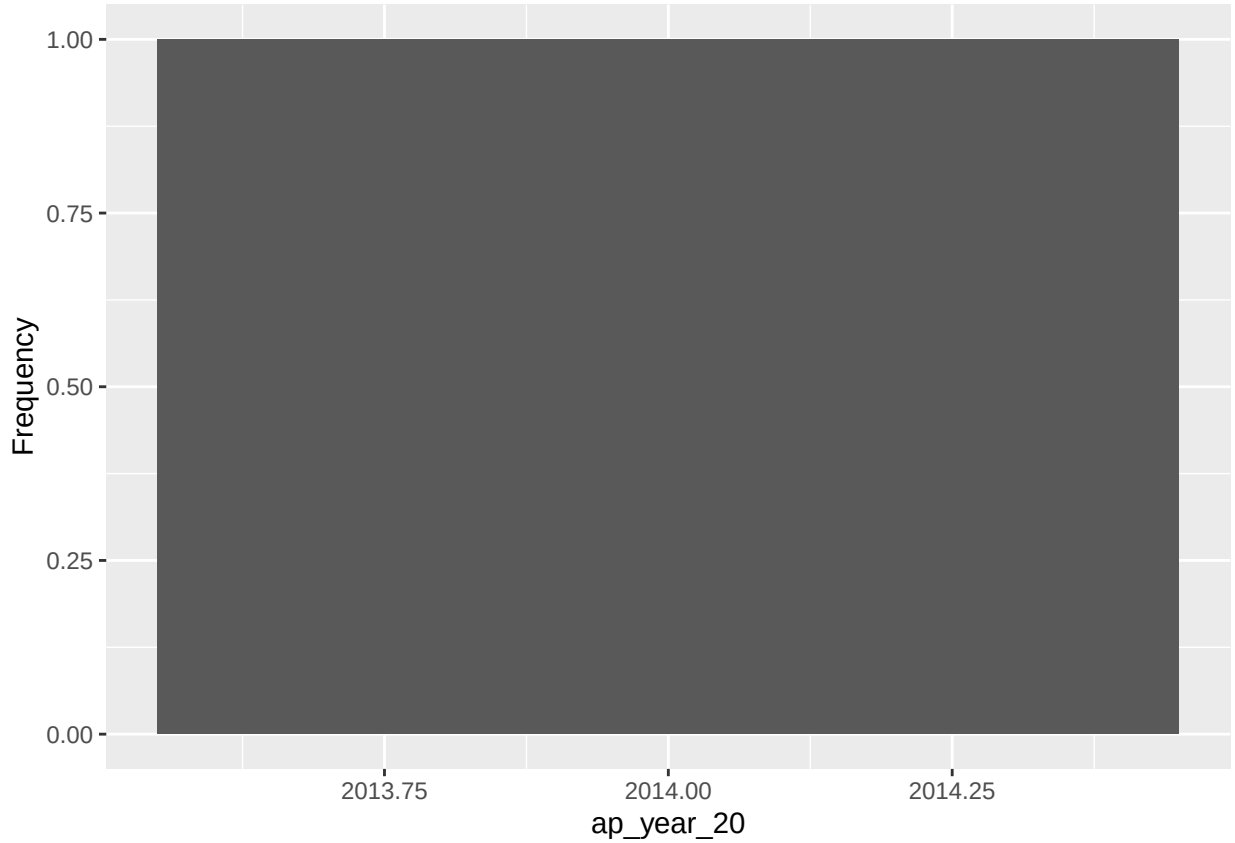

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_code_20, type: character
[1] Values (2 unique): NA, 82
[1] Missing: 100%
```



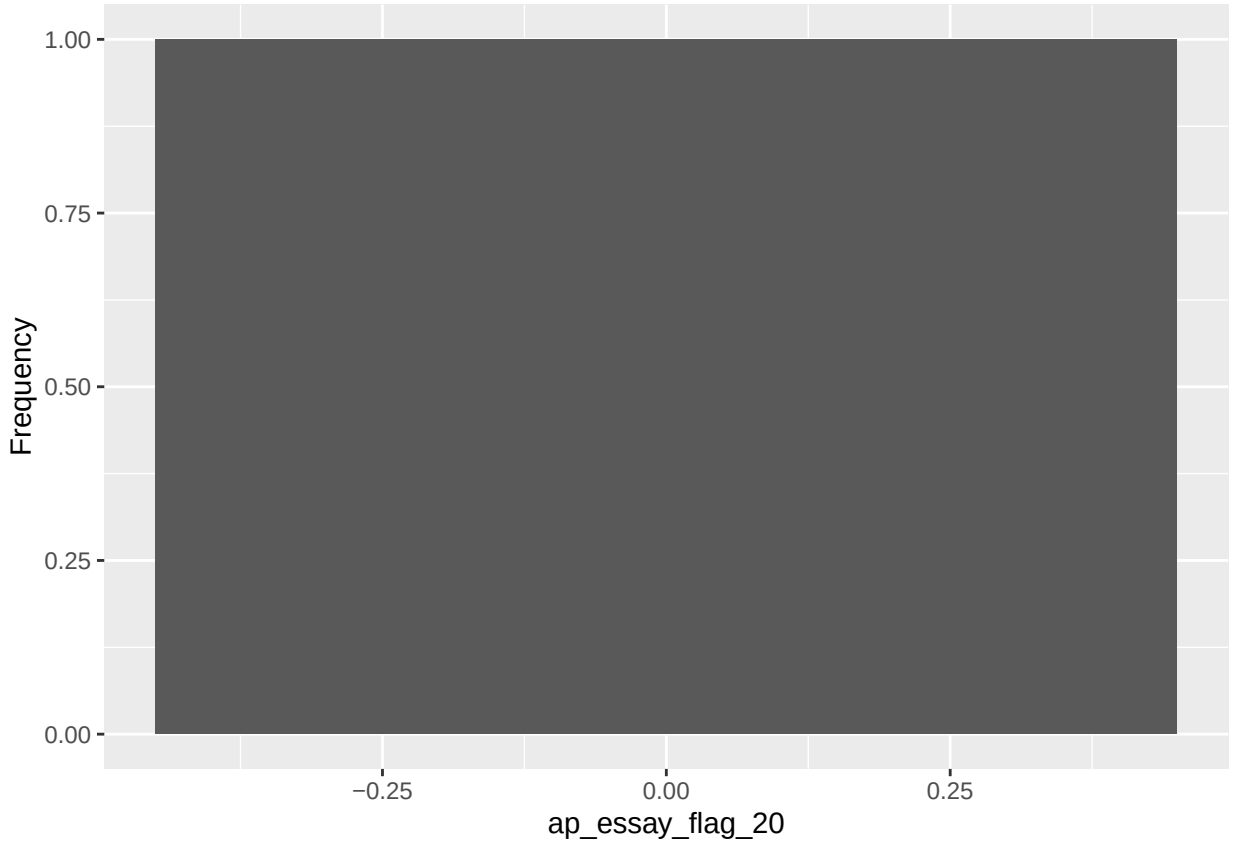

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_score_20, type: numeric
[1] Values (2 unique): NA, 1
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



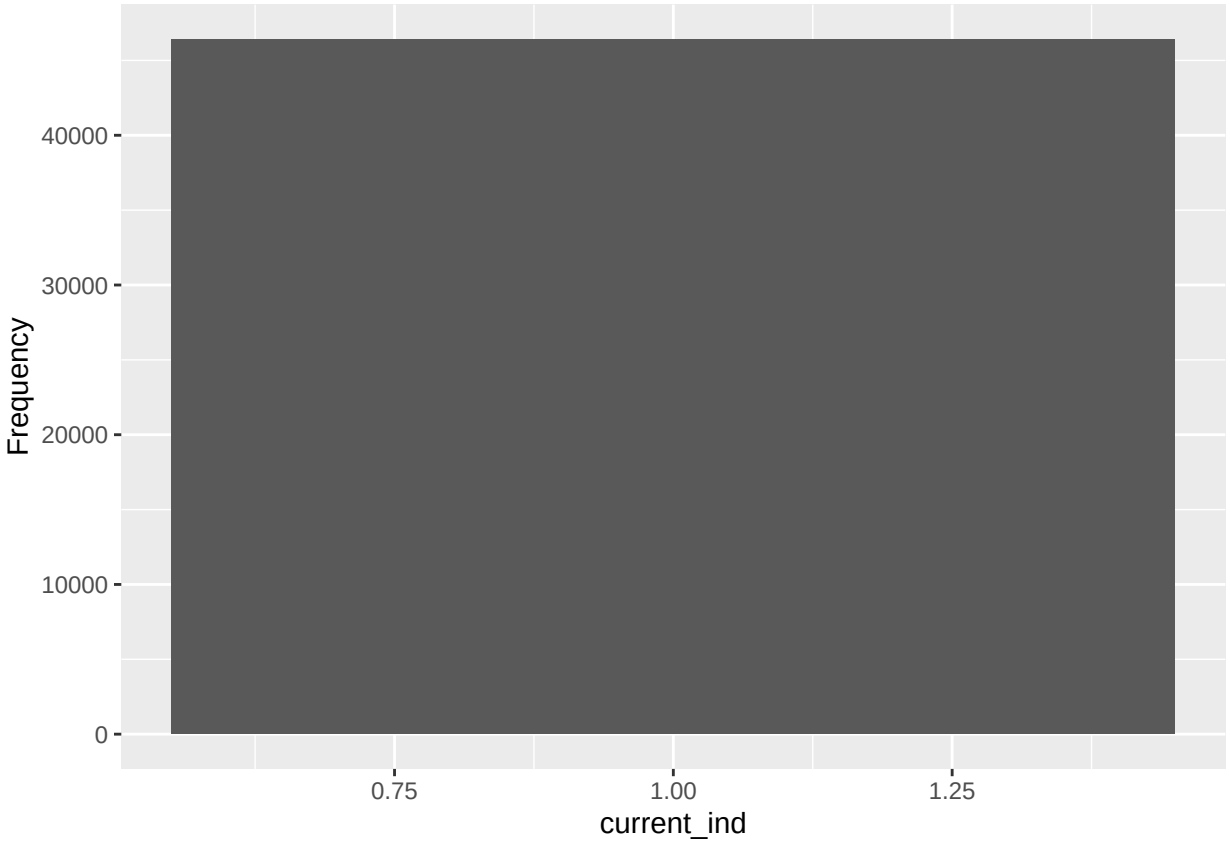

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_year_20, type: integer
[1] Values (2 unique): NA, 2014
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



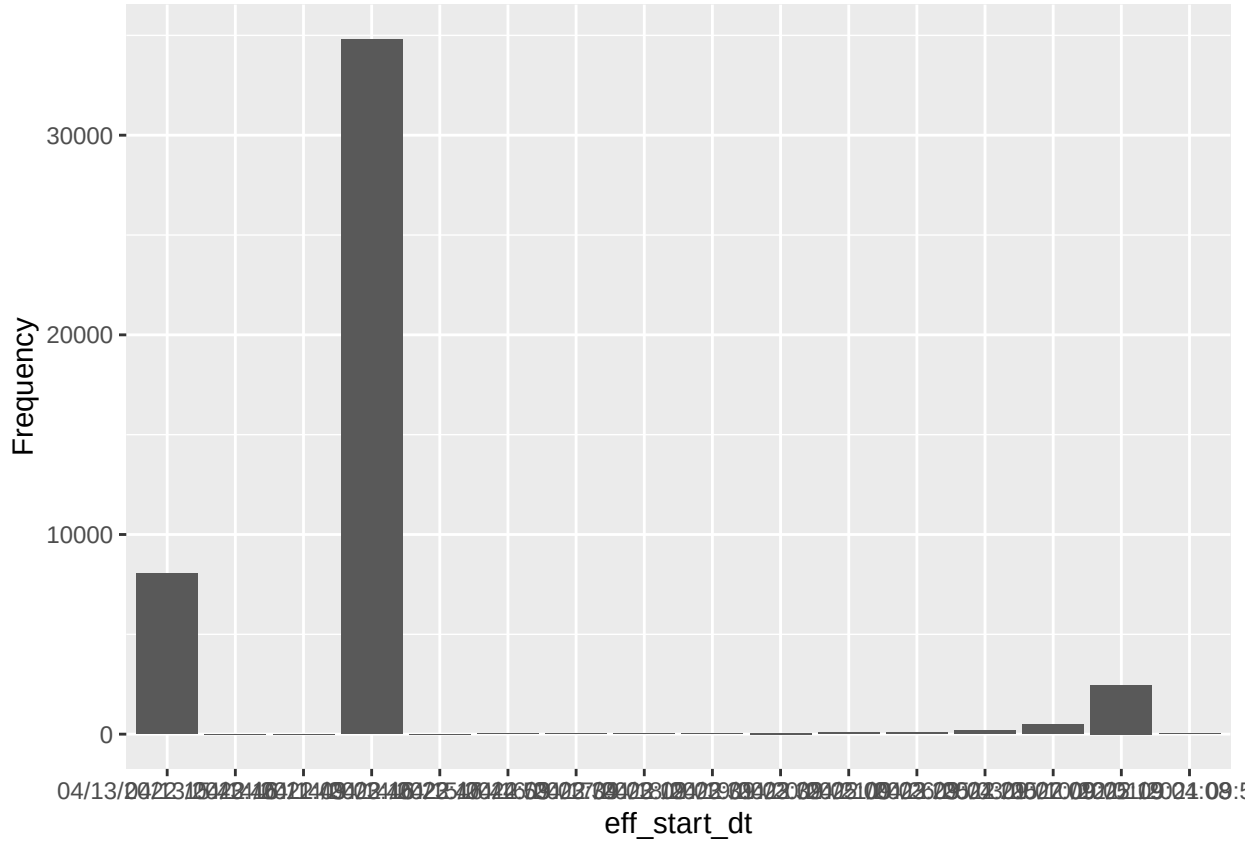

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: ap_essay_flag_20, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 100%

Warning: Removed 46407 rows containing non-finite values (‘stat_count()‘).
```



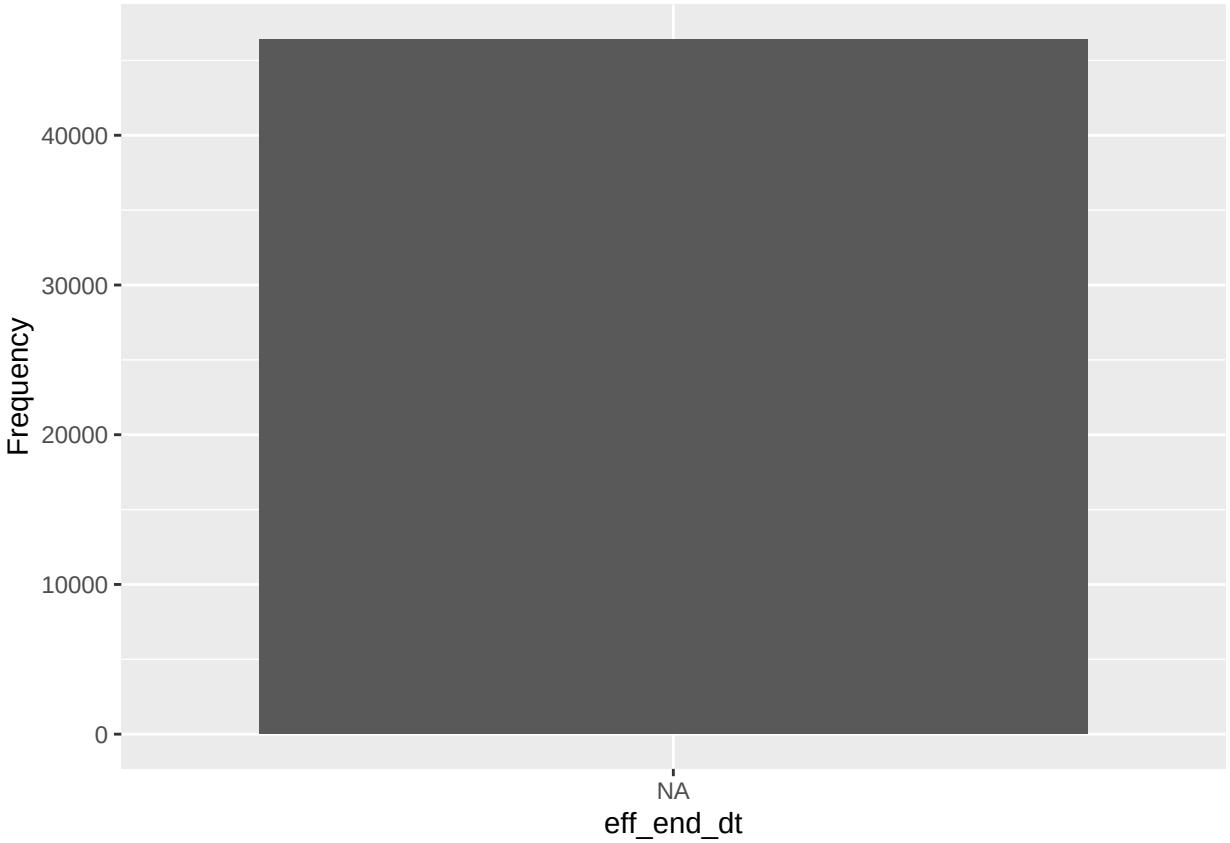

```
[1] is used in feature engineering and hence not included
[1] ____________________________
[1] Variable: current_ind, type: numeric
[1] Values (1 unique): 1
[1] Missing: 0%
```



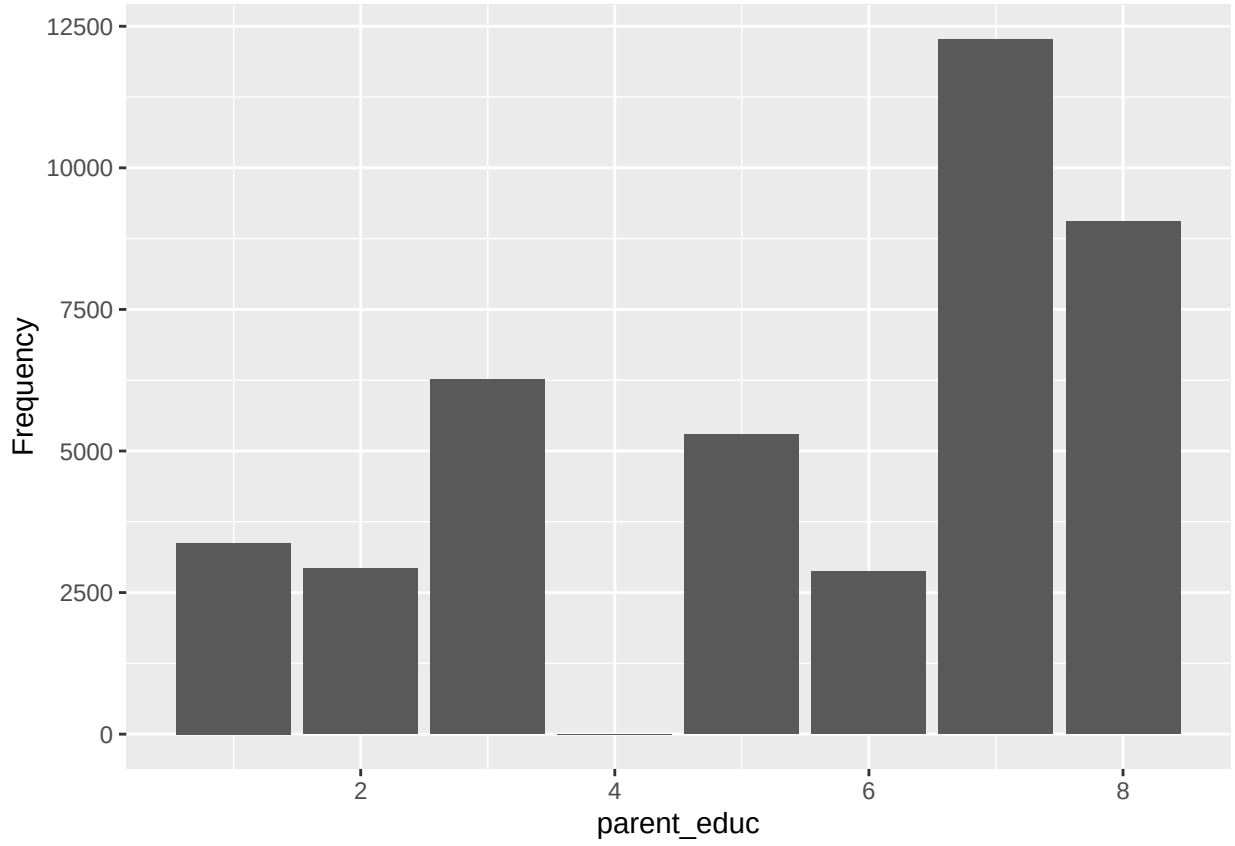

```
[1] ____________________________
[1] Variable: eff_start_dt, type: character
[1] Values (16 unique): 12/01/2021 09:59:41, 04/13/2022 15:43:46, 04/13/2022 16:11:43, 04/14/2022 09:03
[1] Missing: 0%
```



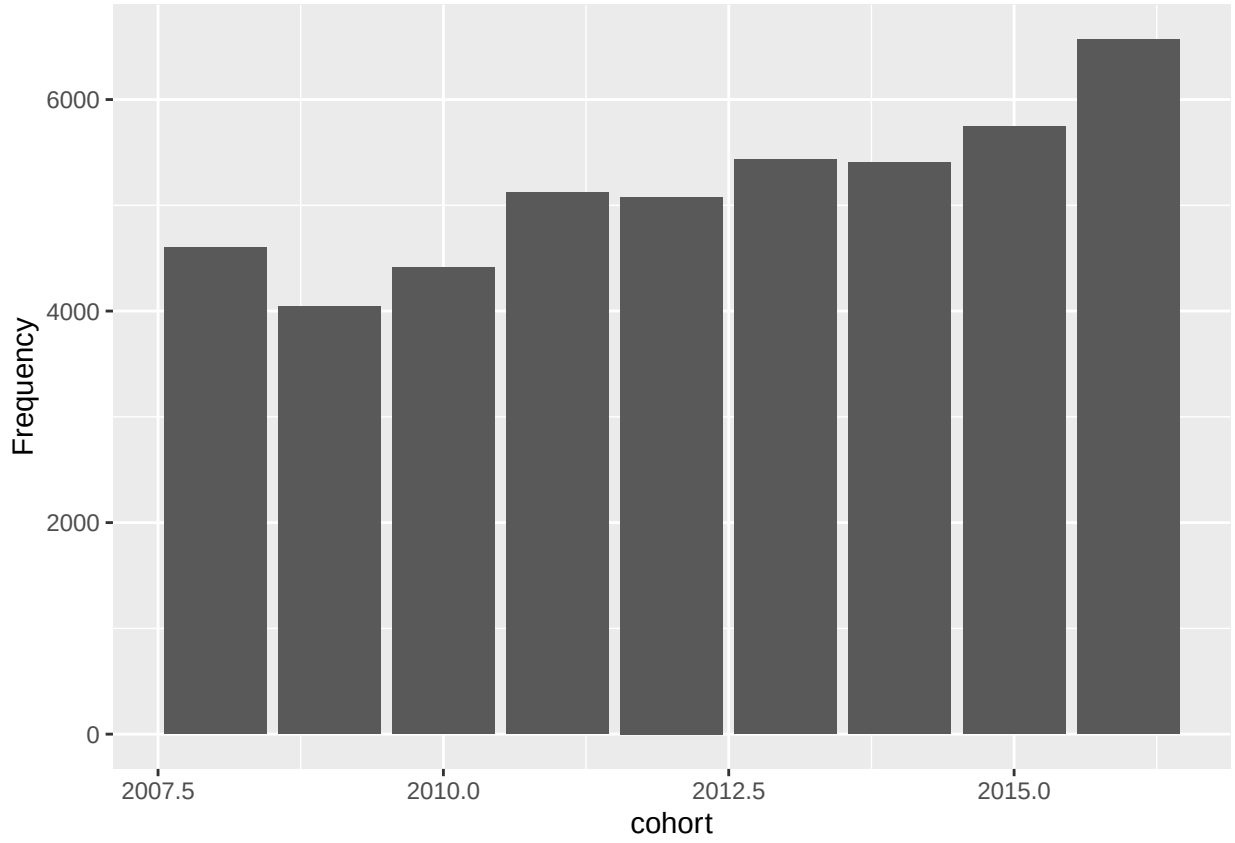

```
[1] ____________________________
[1] Variable: eff_end_dt, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
```




```
[1] ____________________________
[1] Variable: parent_educ, type: numeric
[1] Values (9 unique): 7, 8, 5, 3, NA, ...
[1] Missing: 9.4%
[1] Most missing: F08 41.5%, Least missing: F11 1.9%

Warning: Removed 4344 rows containing non-finite values (‘stat_count()‘).
```




```
[1] ___________________________
[1] Variable: cohort, type: numeric
[1] Values (9 unique): 2008, 2009, 2010, 2015, 2016, ...
[1] Missing: 0%
```




```
[1] should not be used as a predictor
[1] ____________________________
```

These are the predictors that remain when filtering for less than 50% missing, our blacklist and variables that need to be somehow transformed first:

```
 [1] "gender"                "female"                "int_student"
 [4] "ethnicity"             "asian"                 "hispanic"
 [7] "black"                 "white"                 "urm"
[10] "citizenship_app"       "first_generation"      "low_income"
[13] "father_edu_level_code" "mother_edu_level_code" "ell"
[16] "foster_care"           "single_parent"         "country_residence_app"
[19] "geo_category"          "application_term_code" "application_status"
[22] "admitdate"             "cal_res_at_app"        "current_ind"
[25] "parent_educ"
```

Variables that are unexpectedly in the dataset

```
## [1] Variable: pascd, type: numeric
## [1] Values (4 unique): NA, 2, 3, 6
## [1] Missing: 0.1%
## [1]
## [1] _____________________
## [1] Variable: hscd, type: numeric
```

```
## [1] Values (4209 unique): NA, 1937020, 3439963, 3033570, 1933380, ...
## [1] Missing: 28.1%
## [1]
## [1] ____________________
## [1] Variable: must_hsid, type: numeric
## [1] Values (5493 unique): NA, 110813, 101584, 100704, 102236, ...
## [1] Missing: 28%
## [1]
## [1] ____________________
## [1] Variable: ncessch, type: character
## [1] Values (530 unique): NA, 061440001678, 062271003139, 040625000589, 062569007784, ...
## [1] Missing: 94.6%
## [1]
## [1] ____________________
## [1] Variable: hs_address, type: character
## [1] Values (533 unique): NA, 41717 PALM AVENUE, 3501 NORTH BROADWAY, 21200 NORTH 83 AVENUE, 4500 N T
## [1] Missing: 94.6%
## [1]
## [1] ____________________
## [1] Variable: hs_city, type: character
## [1] Values (451 unique): NA, FREMONT, LOS ANGELES, PEORIA, MOORPARK, ...
## [1] Missing: 94.6%
## [1]
## [1] ____________________
## [1] Variable: hs_state, type: character
## [1] Values (51 unique): NA, CA, AZ, DE, TX, ...
## [1] Missing: 94.6%
## [1]
## [1] ____________________
## [1] Variable: hs_zip, type: numeric
## [1] Values (514 unique): NA, 94539, 90031, 85382, 93021, ...
## [1] Missing: 94.6%
## [1]
## [1] ____________________
## [1] Variable: hs_ceeb, type: numeric
## [1] Values (1732 unique): NA, 4400, 4670, 4839, 4098, ...
## [1] Missing: 33.2%
## [1]
## [1] ____________________
## [1] Variable: uncapped_gpa, type: numeric
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: unweighted_gpa, type: numeric
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: grad_major_1, type: character
## [1] Values (96 unique): NA, 97220, 55120, 65867, 967A0, ...
## [1] Missing: 30.8%
## [1]
## [1] ____________________
```

```
## [1] Variable: grad_major_2, type: character
## [1] Values (73 unique): NA, 65836, 65246, 967A1, 7908Q, ...
## [1] Missing: 90.2%
## [1]
## [1] ____________________
## [1] Variable: grad_major_3, type: character
## [1] Values (26 unique): NA, 97220, 97632, 65246, 65836, ...
## [1] Missing: 99.7%
## [1]
## [1] ____________________
## [1] Variable: grad_major_1_desc, type: character
## [1] Values (94 unique): NA, Criminology, Law And Society, Biological Sciences, Sociology, Public Heal
## [1] Missing: 30.8%
## [1]
## [1] ____________________
## [1] Variable: grad_major_2_desc, type: character
## [1] Values (72 unique): NA, International Studies, Economics, Public Health Policy, Education Scienc
## [1] Missing: 90.2%
## [1]
## [1] ____________________
## [1] Variable: grad_major_3_desc, type: character
## [1] Values (26 unique): NA, Criminology, Law And Society, Psychology And Social Behavior, Economics,
## [1] Missing: 99.7%
## [1]
## [1] ____________________
```

We could think about using pascd.

Student by term

The spreadsheet knows 159 variables. The corresponding file contains 183 columns. The expected variables student_term_sid (pk),class_standing,cumulative_units_attempted_graduate,cumulative_units_attempted_online,cumula are not in the dataset. Therefore exist variables created_dttm,current_units_attempted_transfer,current_units_attempted_

Requested and available variables

```
[1] Variable: mellon_id, type: numeric
[1] Values (46408 unique): 214379, 221837, 328990, 212451, 326308, ...
[1] Missing: 0%
[1] should not be used as a predictor
[1] ____________________
[1] Variable: group_a, type: numeric
[1] Values (2 unique): 0, 1
[1] Missing: 0%
[1] should not be used as a predictor
[1] ____________________
[1] Variable: mellon_enr_dt_a, type: character
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: group_b, type: numeric
```

```
[1] Values (1 unique): 0
[1] Missing: 0%
[1] should not be used as a predictor
[1] ____________________
[1] Variable: mellon_enr_dt_b, type: character
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: term_code, type: numeric
[1] Values (43 unique): 201814, 201592, 201514, 201692, 201403, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: term_desc, type: character
[1] Values (43 unique): Spring 2018, Fall 2015, Spring 2015, Fall 2016, Winter 2014, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: term_part_code, type: character
[1] Values (4 unique): 14, 92, 03, NA
[1] Missing: 0%
[1] ____________________
[1] Variable: term_part_desc, type: character
[1] Values (7 unique): Spring, Fall, Winter, NA, Winter Quarter, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: enrollment_open_date, type: character
[1] Values (3 unique): NA, 11/23/2020 00:00:00.000, 02/24/2020 00:00:00.000
[1] Missing: 100%
[1] ____________________
[1] Variable: enrollment_close_date, type: character
[1] Values (3 unique): NA, 12/14/2020 00:00:00.000, 03/11/2020 00:00:00.000
[1] Missing: 100%
[1] ____________________
[1] Variable: instruction_start_date, type: character
[1] Values (3 unique): NA, 01/06/2020 00:00:00.000, 03/30/2020 00:00:00.000
[1] Missing: 100%
[1] ____________________
[1] Variable: instruction_end_date, type: character
[1] Values (3 unique): NA, 03/13/2020 00:00:00.000, 06/05/2020 00:00:00.000
[1] Missing: 100%
[1] ____________________
[1] Variable: citizenship, type: character
[1] Values (3 unique): US Citizen, Not US Citizen, NA
[1] Missing: 0%
[1] ____________________
[1] Variable: household_size, type: numeric
[1] Values (14 unique): NA, 5, 4, 6, 2, ...
[1] Missing: 85.6%
[1] ____________________
[1] Variable: city_residence, type: character
[1] Values (2282 unique): SAN GABRIEL, RIALTO, GARDEN GROVE, IRVINE, EL MONTE, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: state_residence, type: character
[1] Values (59 unique): CA, NV, NA, HI, NY, ...
```

```
[1] Missing: 4.9%
[1] ____________________
[1] Variable: country_residence, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: zip_code, type: character
[1] Values (6829 unique): 91775, 92377, 92841, 92612, 91731, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: housing_status, type: character
[1] Values (3 unique): NA, On-campus, Off-campus
[1] Missing: 83%
[1] ____________________
[1] Variable: housing_status_desc, type: character
[1] Values (4 unique): NA, On-campus: UCI, Off-campus: Commuter, On-campus: ACC
[1] Missing: 83%
[1] ____________________
[1] Variable: housing_complex_name, type: character
[1] Values (15 unique): NA, Mesa Court, Off-campus commuter, Middle Earth, Vista del Campo, ...
[1] Missing: 83%
[1] ____________________
[1] Variable: currently_enrolled, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: cumulative_term_enroll, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: levelcd, type: character
[1] Values (6 unique): 4, 2, 3, 1, NA, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: pascd, type: numeric
[1] Values (2 unique): 2, NA
[1] Missing: 0%
[1] ____________________
[1] Variable: active_student, type: numeric
[1] Values (2 unique): 0, 1
[1] Missing: 0%
[1] ____________________
[1] Variable: new_student, type: character
[1] Values (4 unique): 0, C, R, P
[1] Missing: 0%
[1] ____________________
[1] Variable: full_time, type: numeric
[1] Values (2 unique): 0, 1
[1] Missing: 0%
[1] ____________________
[1] Variable: ferpa_blocked, type: character
[1] Values (3 unique): NA, X, Z
[1] Missing: 100%
[1] ____________________
```

```
[1] Variable: year_study, type: character
[1] Values (5 unique): Senior, Sophomore, Junior, Freshman, Limited
[1] Missing: 0%
[1] ____________________
[1] Variable: second_baccalaureate, type: numeric
[1] Values (3 unique): NA, 1, 0
[1] Missing: 100%
[1] ____________________
[1] Variable: dual_degree, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 91.9%
[1] ____________________
[1] Variable: dual_degree_active, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 91.9%
[1] ____________________
[1] Variable: honors, type: numeric
[1] Values (3 unique): 0, 1, NA
[1] Missing: 0%
[1] ____________________
[1] Variable: honors_program, type: character
[1] Values (6 unique): NA, H, 0, C, S, ...
[1] Missing: 96.9%
[1] ____________________
[1] Variable: deans_list, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: concurrent_law, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 91.9%
[1] ____________________
[1] Variable: ucdc, type: numeric
[1] Values (3 unique): NA, 0, 1
[1] Missing: 91.9%
[1] ____________________
[1] Variable: absentia, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 91.9%
[1] ____________________
[1] Variable: rotc, type: numeric
[1] Values (7 unique): NA, 0, 3, 1, 9, ...
[1] Missing: 91.7%
[1] ____________________
[1] Variable: reduced_ed_fee, type: numeric
[1] Values (3 unique): NA, 0, 1
[1] Missing: 91.9%
[1] ____________________
[1] Variable: education_abroad, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: education_opportunity, type: logical
[1] Values (1 unique): NA
```

```
[1] Missing: 100%
[1] ____________________
[1] Variable: esl_program, type: numeric
[1] Values (2 unique): 0, 1
[1] Missing: 0%
[1] ____________________
[1] Variable: sport, type: character
[1] Values (8 unique): 1, NA, A, Z, 0, ...
[1] Missing: 51.7%
[1] ____________________
[1] Variable: sport_gender, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: entry_level_writing, type: character
[1] Values (5 unique): NA, Requirement satisfied, Course passed, Test passed, Requirement not satisfied
[1] Missing: 91.9%
[1] ____________________
[1] Variable: american_history_desc, type: character
[1] Values (4 unique): NA, Requirement satisfied, Course passed, Requirement not satisfied
[1] Missing: 91.9%
[1] ____________________
[1] Variable: american_intitutions_desc, type: character
[1] Values (4 unique): NA, Requirement satisfied, Course passed, Requirement not satisfied
[1] Missing: 91.9%
[1] ____________________
[1] Variable: probation_status, type: character
[1] Values (3 unique): N, Y, NA
[1] Missing: 0%
[1] ____________________
[1] Variable: registration_status, type: character
[1] Values (7 unique): Continuing, New, Returning, Withdraw, NA, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: veteran_status, type: numeric
[1] Values (3 unique): NA, 0, 1
[1] Missing: 91.9%
[1] ____________________
[1] Variable: pell_eligible_flag, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: major_count, type: numeric
[1] Values (4 unique): 1, 2, 3, 4
[1] Missing: 0%
[1] ____________________
[1] Variable: netpayer_fee_assessed, type: numeric
[1] Values (41 unique): 0, 5149.5298, 5150.54, 5223.6401, 5207.04, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: current_units_attempted_grade, type: numeric
[1] Values (65 unique): NA, 16, 12, 17, 17.5, ...
[1] Missing: 91.9%
[1] ____________________
```

```
[1] Variable: current_units_attempted_pnp, type: numeric
[1] Values (41 unique): NA, 0, 4, 1.3, 2, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_attempted_lowerdiv, type: numeric
[1] Values (62 unique): NA, 4, 0, 13.5, 9.3000002, ...
[1] Missing: 94.8%
[1] ____________________
[1] Variable: current_units_attempted_upperdiv, type: numeric
[1] Values (46 unique): NA, 12, 16, 18, 4, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_attempted_graduate, type: numeric
[1] Values (10 unique): NA, 0, 4, 2, 5, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_attempted_online, type: numeric
[1] Values (18 unique): NA, 0, 8, 4, 2, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_attempted_oncampus, type: numeric
[1] Values (92 unique): NA, 16, 20, 17, 17.5, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_completed_graded, type: numeric
[1] Values (62 unique): NA, 16, 12, 17, 8, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_completed_pnp, type: numeric
[1] Values (34 unique): NA, 0, 4, 1, 2, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_completed_lowerdiv, type: numeric
[1] Values (43 unique): NA, 0, 8, 4, 9, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_completed_upperdiv, type: numeric
[1] Values (36 unique): NA, 0, 20, 4, 8, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_completed_graduate, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: current_units_completed_online, type: numeric
[1] Values (16 unique): NA, 0, 4, 8, 2, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_completed_oncampus, type: numeric
[1] Values (76 unique): NA, 16, 20, 17, 8, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: current_units_completed_transfer, type: numeric
[1] Values (1 unique): 0
```

```
[1] Missing: 0%
[1] ____________________
[1] Variable: current_units_completed_total, type: numeric
[1] Values (88 unique): 14, 12, 8, 16, 15.5, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: cumulative_units_attempted_grade, type: numeric
[1] Values (535 unique): NA, 173, 172, 155, 218, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: cumulative_units_attempted_pnp, type: numeric
[1] Values (356 unique): NA, 0, 1, 2, 7.9000001, ...
[1] Missing: 71%
[1] ____________________
[1] Variable: cumulative_units_attempted_lower, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: cumulative_units_attempted_upper, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: cumulative_units_attempted_total, type: numeric
[1] Values (1446 unique): NA, 173, 179.89999, 179, 218, ...
[1] Missing: 91.9%
[1] ____________________
[1] Variable: cumulative_units_completed_pnp, type: numeric
[1] Values (317 unique): NA, 0, 1, 2, 7.9000001, ...
[1] Missing: 71%
[1] ____________________
[1] Variable: cumulative_units_completed_total, type: numeric
[1] Values (3031 unique): 182, 45, 225, 146, 216, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: gpa_term, type: numeric
[1] Values (396 unique): 3.8, 3.53, 2.4200001, 3.5999999, 3.1700001, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: gpa_cumulative, type: numeric
[1] Values (390 unique): 3.25, 3.53, 2.8299999, 3.6900001, 3.01, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: major_code_1, type: character
[1] Values (106 unique): 836, 345, 284, 090, 75A, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: major_name_1, type: character
[1] Values (106 unique): INTERNATIONAL STUDIES, ENGLISH, CIVIL ENGINEERING, ART, PHARMACEUTICAL SCIENCES
[1] Missing: 0%
[1] ____________________
[1] Variable: major_name_abbrev_1, type: character
[1] Values (158 unique): Bio-Ua, English, P H Sci, Art, PhrmSci, ...
[1] Missing: 0.3%
[1] ____________________
```

```
[1] Variable: major_school_code_1, type: character
[1] Values (16 unique): 55, 60, 96, 57, 93, ...
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_school_name_1, type: character
[1] Values (16 unique): School of Biological Sciences, School of Humanities, Program in Public Health, S
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_school_name_abbrev_1, type: character
[1] Values (18 unique): Bio Sci, Humanities, PubHlth, Arts, Pharmacy, ...
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_subcampus_1, type: character
[1] Values (3 unique): General campus, Health science, NA
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_funding_1, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: major_cip_code_1, type: character
[1] Values (90 unique): 30.9999, 23.0101, 51.2299, 50.0101, 51.2010, ...
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_cip_code_historical_1, type: character
[1] Values (98 unique): 30.9999, 23.0101, 51.2299, 50.0101, 51.2010, ...
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_cip_name_1, type: character
[1] Values (90 unique): Multi-/Interdisciplinary Studies, Other, English Language and Literature, Genera
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_cip_series_code_1, type: character
[1] Values (21 unique): 30, 23, 51, 50, 26, ...
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_cip_series_name_1, type: character
[1] Values (22 unique): Multi/Interdisciplinary Studies, English Language and Literature/Letters, Healt
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_cip_category_code_1, type: character
[1] Values (69 unique): 30.99, 23.01, 51.22, 50.01, 51.20, ...
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_cip_category_name_1, type: character
[1] Values (69 unique): Multi/Interdisciplinary Studies, Other, English Language and Literature, Genera
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_stem_1, type: numeric
[1] Values (3 unique): 0, 1, NA
[1] Missing: 0.3%
[1] ____________________
[1] Variable: major_degree_granting_1, type: numeric
[1] Values (3 unique): NA, 1, 0
```

```
[1] Missing: 91.9%
[1] ____________________
[1] Variable: major_graduated_1, type: character
[1] Values (147 unique): NA, Civil Engineering, Mathematics, URBAN STUDIES, PSYCH & SOC BEHAVIOR, ...
[1] Missing: 90.3%
[1] ____________________
[1] Variable: major_code_2, type: character
[1] Values (94 unique): NA, 836, 632, 780, 851, ...
[1] Missing: 92.1%
[1] ____________________
[1] Variable: major_name_2, type: character
[1] Values (94 unique): NA, INTERNATIONAL STUDIES, PSYCHOLOGY AND SOCIAL BEHAVIOR, PSYCHOLOGY_BA, SOCIA
[1] Missing: 92.1%
[1] ____________________
[1] Variable: major_name_abbrev_2, type: character
[1] Values (93 unique): NA, Intl St, Psy Beh, Psych, SocEcol, ...
[1] Missing: 92.2%
[1] ____________________
[1] Variable: major_school_code_2, type: numeric
[1] Values (13 unique): NA, 65, 97, 77, 60, ...
[1] Missing: 92.2%
[1] ____________________
[1] Variable: major_school_name_2, type: character
[1] Values (13 unique): NA, School of Social Sciences, School of Social Ecology, School of Engineering,
[1] Missing: 92.2%
[1] ____________________
[1] Variable: major_school_name_abbrev_2, type: character
[1] Values (14 unique): NA, Soc Sci, Soc Ecol, Engr, Humanities, ...
[1] Missing: 92.2%
[1] ____________________
[1] Variable: major_subcampus_2, type: character
[1] Values (3 unique): NA, General campus, Health science
[1] Missing: 92.2%
[1] ____________________
[1] Variable: major_funding_2, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: major_cip_code_2, type: numeric
[1] Values (68 unique): NA, 42.2707, 45.1001, 45.1101, 42.2799, ...
[1] Missing: 98.8%
[1] ____________________
[1] Variable: major_cip_code_historical_2, type: numeric
[1] Values (69 unique): NA, 42.2707, 45.1001, 45.1101, 42.2799, ...
[1] Missing: 98.8%
[1] ____________________
[1] Variable: major_cip_name_2, type: character
[1] Values (68 unique): NA, Social Psychology, Political Science and Government, General, Sociology, Res
[1] Missing: 98.8%
[1] ____________________
[1] Variable: major_cip_series_code_2, type: numeric
[1] Values (20 unique): NA, 42, 45, 23, 50, ...
[1] Missing: 98.8%
[1] ____________________
```

```
[1] Variable: major_cip_series_name_2, type: character
[1] Values (21 unique): NA, Psychology, Social Sciences, English Language and Literature/Letters, Visual
[1] Missing: 98.8%
[1] ____________________
[1] Variable: major_cip_category_code_2, type: numeric
[1] Values (55 unique): NA, 42.27, 45.1, 45.11, 23.01, ...
[1] Missing: 98.8%
[1] ____________________
[1] Variable: major_cip_category_name_2, type: character
[1] Values (55 unique): NA, Research and Experimental Psychology, Political Science and Government, Soci
[1] Missing: 98.8%
[1] ____________________
[1] Variable: major_stem_2, type: numeric
[1] Values (3 unique): NA, 0, 1
[1] Missing: 92.2%
[1] ____________________
[1] Variable: major_degree_granting_2, type: numeric
[1] Values (2 unique): NA, 1
[1] Missing: 98.8%
[1] ____________________
[1] Variable: major_graduated_2, type: character
[1] Values (102 unique): INTERNATIONAL STUDIES, ENGLISH, CIVIL ENGINEERING, ART, PHARMACEUTICAL SCI, ..
[1] Missing: 5.8%
[1] ____________________
[1] Variable: major_code_3, type: character
[1] Values (49 unique): NA, 666, 540, 16A, 277, ...
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_name_3, type: character
[1] Values (49 unique): NA, PHYSICS, MATHEMATICS, CHICANO/LATINO STUDIES, MECHANICAL ENGINEERING, ...
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_name_abbrev_3, type: character
[1] Values (46 unique): NA, Physics, Math, Chc/Lat, Engr ME, ...
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_school_code_3, type: numeric
[1] Values (11 unique): NA, 62, 65, 77, 60, ...
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_school_name_3, type: character
[1] Values (11 unique): NA, School of Physical Sciences, School of Social Sciences, School of Engineeri
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_school_name_abbrev_3, type: character
[1] Values (11 unique): NA, PhySci, Soc Sci, Engr, Humanities, ...
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_subcampus_3, type: character
[1] Values (3 unique): NA, General campus, Health science
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_funding_3, type: logical
[1] Values (1 unique): NA
```

```
[1] Missing: 100%
[1] ____________________
[1] Variable: major_stem_3, type: numeric
[1] Values (3 unique): NA, 1, 0
[1] Missing: 99.8%
[1] ____________________
[1] Variable: major_degree_granting_3, type: numeric
[1] Values (2 unique): NA, 1
[1] Missing: 100%
[1] ____________________
[1] Variable: major_graduated_3, type: character
[1] Values (35 unique): NA, URBAN STUDIES, BIOMEDICAL ENGR, MATERIALS SCI ENGR, SOCIAL ECOLOGY, ...
[1] Missing: 99.9%
[1] ____________________
[1] Variable: major_code_4, type: character
[1] Values (5 unique): NA, 0EH, 632, 867, 912
[1] Missing: 100%
[1] ____________________
[1] Variable: major_name_4, type: character
[1] Values (5 unique): NA, PSYCHOLOGICAL SCIENCE, PSYCHOLOGY AND SOCIAL BEHAVIOR, SOCIOLOGY, URBAN STUDI
[1] Missing: 100%
[1] ____________________
[1] Variable: major_name_abbrev_4, type: character
[1] Values (4 unique): NA, PsychSc, Sociol, UrbanSt
[1] Missing: 100%
[1] ____________________
[1] Variable: major_school_code_4, type: numeric
[1] Values (3 unique): NA, 97, 65
[1] Missing: 100%
[1] ____________________
[1] Variable: major_school_name_4, type: character
[1] Values (3 unique): NA, School of Social Ecology, School of Social Sciences
[1] Missing: 100%
[1] ____________________
[1] Variable: major_school_name_abbrev_4, type: character
[1] Values (3 unique): NA, Soc Ecol, Soc Sci
[1] Missing: 100%
[1] ____________________
[1] Variable: major_subcampus_4, type: character
[1] Values (2 unique): NA, General campus
[1] Missing: 100%
[1] ____________________
[1] Variable: major_funding_4, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: major_stem_4, type: numeric
[1] Values (2 unique): NA, 0
[1] Missing: 100%
[1] ____________________
[1] Variable: major_degree_granting_4, type: numeric
[1] Values (2 unique): NA, 1
[1] Missing: 100%
[1] ____________________
```

```
[1] Variable: major_graduated_4, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: major_minor, type: character
[1] Values (85 unique): NA, GENDER & SEXUALITY STUDIES, PSYCHOLOGY, KOREAN LIT & CULTURE, ART HISTORY,
[1] Missing: 77.8%
[1] ____________________
[1] Variable: current_ind, type: logical
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
```

These are the predictors that remain when filtering for less than 50% missing, our blacklist and variables that need to be somehow transformed first:

```
 [1] "term_part_code"                   "term_part_desc"
 [3] "citizenship"                      "levelcd"
 [5] "pascd"                            "active_student"
 [7] "new_student"                      "full_time"
 [9] "year_study"                       "honors"
[11] "esl_program"                      "probation_status"
[13] "registration_status"              "major_count"
[15] "current_units_completed_transfer" "major_subcampus_1"
[17] "major_stem_1"
```

Variables that are unexpectedly in the dataset

```
## [1] Variable: created_dttm, type: character
## [1] Values (68 unique): 04/08/2022 13:10:30.000, 04/09/2022 09:08:26.000, 04/08/2022 12:20:39.000, 1
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: current_units_attempted_transfer, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: current_units_attempted_total, type: numeric
## [1] Values (92 unique): NA, 16, 20, 17, 17.5, ...
## [1] Missing: 91.9%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_attempted_gradu, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_attempted_onlin, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
```

```
## [1] Variable: cumulative_units_attempted_oncam, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_attempted_trans, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_completed_grade, type: numeric
## [1] Values (493 unique): NA, 173, 168, 155, 218, ...
## [1] Missing: 91.9%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_completed_lower, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_completed_upper, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_completed_gradu, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_completed_onlin, type: numeric
## [1] Values (73 unique): NA, 4, 0, 8, 28, ...
## [1] Missing: 91.9%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_completed_oncam, type: numeric
## [1] Values (1470 unique): NA, 173, 175.89999, 175, 218, ...
## [1] Missing: 91.9%
## [1]
## [1] ____________________
## [1] Variable: cumulative_units_completed_trans, type: numeric
## [1] Values (281 unique): 0, 12, 16, 44, 66.5, ...
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: total_terms_enrolled_excluding_s, type: numeric
## [1] Values (23 unique): 0, 12, 9, 14, 6, ...
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: total_terms_enrolled_including_s, type: numeric
## [1] Values (28 unique): NA, 14, 12, 9, 11, ...
## [1] Missing: 91.9%
## [1]
```

```
## [1] ____________________
## [1] Variable: major_cip_code_3, type: numeric
## [1] Values (21 unique): NA, 40.0801, 45.1101, 42.0101, 27.0101, ...
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_code_historical_3, type: numeric
## [1] Values (21 unique): NA, 40.0801, 45.1101, 42.0101, 27.0101, ...
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_name_3, type: character
## [1] Values (21 unique): NA, Physics, General, Sociology, Psychology, General, Mathematics, General,
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_series_code_3, type: numeric
## [1] Values (12 unique): NA, 40, 45, 42, 27, ...
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_series_name_3, type: character
## [1] Values (12 unique): NA, Physical Sciences, Social Sciences, Psychology, Mathematics and Statisti
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_category_code_3, type: numeric
## [1] Values (20 unique): NA, 40.08, 45.11, 42.01, 27.01, ...
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_category_name_3, type: character
## [1] Values (20 unique): NA, Physics, Sociology, Psychology, General, Mathematics, ...
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_code_4, type: numeric
## [1] Values (4 unique): NA, 42.2799, 45.1101, 45.1201
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_code_historical_4, type: numeric
## [1] Values (4 unique): NA, 42.2799, 45.1101, 45.1201
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_name_4, type: character
## [1] Values (4 unique): NA, Research and Experimental Psychology, Other., Sociology, Urban Studies/Af
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_series_code_4, type: numeric
## [1] Values (3 unique): NA, 42, 45
## [1] Missing: 100%
```

```
## [1]
## [1] ____________________
## [1] Variable: major_cip_series_name_4, type: character
## [1] Values (3 unique): NA, Psychology, Social Sciences
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_category_code_4, type: numeric
## [1] Values (4 unique): NA, 42.27, 45.11, 45.12
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: major_cip_category_name_4, type: character
## [1] Values (4 unique): NA, Research and Experimental Psychology, Sociology, Urban Studies/Affairs
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: pell_grant_eligibility, type: character
## [1] Values (4 unique): NA, N, P, F
## [1] Missing: 88.5%
## [1]
## [1] ____________________
## [1] Variable: pell_grant_received, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: md5, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: effective_date, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: end_effective_date, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: mapping_nm, type: logical
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
## [1] Variable: maj_tag1, type: numeric
## [1] Values (4 unique): 2, 1, 15, 8
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: maj_tag2, type: numeric
## [1] Values (6 unique): 17, 9, 1, 15, 8, ...
```

```
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: maj_tag3, type: numeric
## [1] Values (5 unique): 17, 1, 9, 15, 16
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: maj_tag4, type: numeric
## [1] Values (3 unique): 17, 1, 15
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: acadyr, type: character
## [1] Values (14 unique): 2017-2018, 2015-2016, 2014-2015, 2016-2017, 2013-2014, ...
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: mellon_yr, type: numeric
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
## [1] ____________________
```

We could think about using maj_tag1,maj_tag2,maj_tag3,maj_tag4.

Student by term by course

The spreadsheet knows 52 variables. The corresponding file contains 55 columns. The expected variables are not in the dataset. Therefore exist variables year,acadyr,mellon_yr

Requested and available variables

```
[1] Variable: mellon_id, type: numeric
[1] Values (46281 unique): 3022066, 167308, 2959537, 2971243, 3026122, ...
[1] Missing: 0%
[1] should not be used as a predictor
[1] ____________________
[1] Variable: group_a, type: numeric
[1] Values (2 unique): 0, 1
[1] Missing: 0%
[1] should not be used as a predictor
[1] ____________________
[1] Variable: mellon_enr_dt_a, type: character
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: group_b, type: numeric
[1] Values (1 unique): 0
[1] Missing: 0%
[1] should not be used as a predictor
```

```
[1] ____________________
[1] Variable: mellon_enr_dt_b, type: character
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: term_code, type: numeric
[1] Values (42 unique): 200892, 200903, 200914, 200992, 201003, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: term_desc, type: character
[1] Values (42 unique): Fall 2008, Winter 2009, Spring 2009, Fall 2009, Winter 2010, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: course_code, type: NA
[1] Values (24810 unique): 20075, 20000, 20001, 20005, 20006, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: course_code_concat, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: course_title, type: NA
[1] Values (6026 unique): NA, FEAR-REBEL SLAVES, RACE AND GENDER, ART&COLONIAL AFRICA, RACE & ETHNICITY
[1] Missing: 52.4%
[1] ____________________
[1] Variable: course_dept_code_and_num, type: NA
[1] Values (4449 unique): AFAM 158, AFAM 40A, ANTHRO 121D, ANTHRO 139, ANTHRO 169, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: course_section_num, type: NA
[1] Values (1142 unique): A, 1, 5, 6, B, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: course_section_title, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: school_code, type: NA
[1] Values (22 unique): 65, 08, 77, 62, 97, ...
[1] Missing: 71%
[1] ____________________
[1] Variable: school_name_abbrev, type: NA
[1] Values (21 unique): Soc Sci, Unaffiliated, Engr, Phys Sci, Soc Ecol, ...
[1] Missing: 71%
[1] ____________________
[1] Variable: dept_code, type: NA
[1] Values (97 unique): AFAM, NA, ANTHRO, ART STU, ASIANAM, ...
[1] Missing: 72.9%
[1] ____________________
[1] Variable: dept_name_abbrev, type: NA
[1] Values (111 unique): AFAM, ANTHRO, ARTHIS, ARTS, ARTSTU, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: course_level, type: NA
```

```
[1] Values (3 unique): Undergraduate, NA, Graduate
[1] Missing: 72.9%
[1] ____________________
[1] Variable: course_type, type: NA
[1] Values (12 unique): NA, LEC, DIS, LAB, TUT, ...
[1] Missing: 52.4%
[1] ____________________
[1] Variable: meeting_schedule, type: NA
[1] Values (2402 unique): NA, T T 09:30-10:50, T T 12:30-01:50P, M 02:00-04:50P, T T 11:00-12:20, ...
[1] Missing: 8.1%
[1] ____________________
[1] Variable: meeting_location, type: NA
[1] Values (947 unique): NA, DBH 1427, HOB2 131, PCB 1200, CRCC 102, ...
[1] Missing: 8.8%
[1] ____________________
[1] Variable: enroll_restrictions, type: NA
[1] Values (20 unique): NA, A, B, X, T, ...
[1] Missing: 72.2%
[1] ____________________
[1] Variable: enroll_restrictions_desc, type: NA
[1] Values (15 unique): NA, Prerequisite required, Pass/Not Pass option only, Authorization code require
[1] Missing: 95.5%
[1] ____________________
[1] Variable: class_weeks, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: instructor_id, type: NA
[1] Values (9837 unique): NA, 1719045, 1712940, 1716270, 1711402, ...
[1] Missing: 53%
[1] ____________________
[1] Variable: instructor_title, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: honors_course, type: NA
[1] Values (3 unique): 0, NA, 1
[1] Missing: 72.9%
[1] ____________________
[1] Variable: transferring_inst, type: NA
[1] Values (2 unique): &, NA
[1] Missing: 72.9%
[1] ____________________
[1] Variable: online_course, type: NA
[1] Values (3 unique): 0, NA, 1
[1] Missing: 72.9%
[1] ____________________
[1] Variable: min_units, type: NA
[1] Values (12 unique): 4, 0, 2, 1, 3, ...
[1] Missing: 0%
[1] ____________________
[1] Variable: max_units, type: NA
[1] Values (13 unique): 4, 0, 2, 1, 3, ...
[1] Missing: 7.1%
```

```
[1] ____________________
[1] Variable: max_seats, type: NA
[1] Values (399 unique): NA, 18, 16, 316, 501, ...
[1] Missing: 92.6%
[1] ____________________
[1] Variable: students_waitlisted, type: NA
[1] Values (24 unique): NA, 0, 10, 4, 2, ...
[1] Missing: 92.6%
[1] ____________________
[1] Variable: seats_requested, type: NA
[1] Values (752 unique): NA, 14, 28, 29, 39, ...
[1] Missing: 52.4%
[1] ____________________
[1] Variable: in_progress, type: NA
[1] Values (2 unique): NA, 1
[1] Missing: 92.6%
[1] ____________________
[1] Variable: enrollment_status, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: course_syllabus, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: section_start_at, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: section_end_at, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: credit_code, type: NA
[1] Values (4 unique): NA, NAT, PN, WC
[1] Missing: 99.3%
[1] ____________________
[1] Variable: repeat_code, type: NA
[1] Values (8 unique): NA, NAT, G0, G1, PG, ...
[1] Missing: 99.3%
[1] ____________________
[1] Variable: workload_credit_only, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: include_gpa, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: include_units_attempted, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: include_units_completed, type: NA
```

```
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: units_attempted, type: NA
[1] Values (17 unique): 4, 0, 2, 1, 3, ...
[1] Missing: 6.6%
[1] ____________________
[1] Variable: units_completed, type: NA
[1] Values (33 unique): 4, NA, 3, 2, 5, ...
[1] Missing: 45.7%
[1] ____________________
[1] Variable: midterm_grade, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: final_score_canvas, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
[1] Variable: final_grade, type: NA
[1] Values (26 unique): A-, B-, B+, B, CR, ...
[1] Missing: 2.5%
[1] ____________________
[1] Variable: final_grade_date, type: NA
[1] Values (42 unique): NA, 2019-12-19, 2019-12-23, 2020-01-28, 2020-01-16, ...
[1] Missing: 99.3%
[1] ____________________
[1] Variable: final_grade_official, type: NA
[1] Values (1 unique): NA
[1] Missing: 100%
[1] ____________________
```

These are the predictors that remain when filtering for less than 50% missing, our blacklist and variables that need to be somehow transformed first:

```
NULL
```

Variables that are unexpectedly in the dataset

```
## [1] Variable: year, type: NA
## [1] Values (15 unique): 2008, 2009, 2010, 2011, 2012, ...
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: acadyr, type: character
## [1] Values (14 unique): 2008-2009, 2009-2010, 2010-2011, 2011-2012, 2012-2013, ...
## [1] Missing: 0%
## [1]
## [1] ____________________
## [1] Variable: mellon_yr, type: numeric
## [1] Values (1 unique): NA
## [1] Missing: 100%
## [1]
```

```
## [1] ____________________
```

We could think about using .