Investigating Factors Influencing Academic Achievement in English Towns

Zakk Heile & Julia Healey-Parera

```r
library(MASS)
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.2
```

```
Warning: package 'readr' was built under R version 4.3.2
```

```
Warning: package 'lubridate' was built under R version 4.3.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x dplyr::select() masks MASS::select()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(broom)
library(glm2)
```

```
Attaching package: 'glm2'

The following object is masked from 'package:MASS':

    crabs
```

```r
library(dplyr)
library(Stat2Data)
library(pROC)
```

```
Warning: package 'pROC' was built under R version 4.3.3

Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'

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

    cov, smooth, var
```

```r
library(yardstick)
```

```
Attaching package: 'yardstick'

The following object is masked from 'package:readr':

    spec
```

```r
library(ggplot2)
library(janitor)
```

```
Warning: package 'janitor' was built under R version 4.3.3


Attaching package: 'janitor'
```

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

    chisq.test, fisher.test
```

```r
library(here)
```

```
here() starts at C:/Users/zakkh/STA 210/STA210Final
```

```r
library(fs)
library(withr)
library(lmtest)
```

```
Warning: package 'lmtest' was built under R version 4.3.3
```

```
Loading required package: zoo
```

```
Warning: package 'zoo' was built under R version 4.3.2
```

```

Attaching package: 'zoo'

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

    as.Date, as.Date.numeric
```

```r
#edu <- read.csv("data/english_education.csv")
edu <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mast
```

```
Rows: 1104 Columns: 31
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (13): town11cd, town11nm, size_flag, rgn11nm, coastal, coastal_detailed,...
dbl (18): population_2011, ks4_2012_2013_counts, key_stage_2_attainment_scho...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
working_dir <- here::here("data")

xls_path <- withr::local_tempfile(fileext = ".xlsx")
download.file(
  "https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/educationandchildcare/dat
  xls_path,
  mode = "wb"
)

english_education <- readxl::read_xlsx(xls_path, sheet = "Data", na = "*") |>
  janitor::clean_names()

readr::write_csv(
  english_education,
  fs::path(working_dir, "english_education.csv"))
```

```
# bc <- boxcox(lm(highest_level_qualification_achieved_b_age_22_average_score ~ population
# lambda <- bc$x[which.max(bc$y)]
# english_education$population_2011_bc <- (english_education$population_2011^lambda - 1) /

# level4qual_residents35_64_2011
# activity_at_age_19_full_time_higher_education
# activity_at_age_19_appprenticeships

english_education$rgn11nm <- as.character(english_education$rgn11nm)

english_education$rgn11nm_combined <- "Other"

#Combining regions where it makes sense geographically
english_education$rgn11nm_combined[english_education$rgn11nm %in% c("South East", "South W
english_education$rgn11nm_combined[english_education$rgn11nm == "North East"] <- "North Ea
english_education$rgn11nm_combined[english_education$rgn11nm == "North West"] <- "North We

english_education$rgn11nm_combined <- factor(english_education$rgn11nm_combined)
english_education$rgn11nm_combined <- factor(english_education$rgn11nm_combined)

#Baseline
english_education$rgn11nm_combined <- relevel(english_education$rgn11nm_combined, ref = "S

#Final model that includes an interaction term
m1 <- lm(
  highest_level_qualification_achieved_b_age_22_average_score   ~
    level4qual_residents35_64_2011*level_3_at_age_18 +
    activity_at_age_19_full_time_higher_education +
    level4qual_residents35_64_2011 +
    rgn11nm_combined +
    level_3_at_age_18   +
    key_stage_4_attainment_school_year_2012_to_2013,
  data = english_education
  )
summary(m1)
```

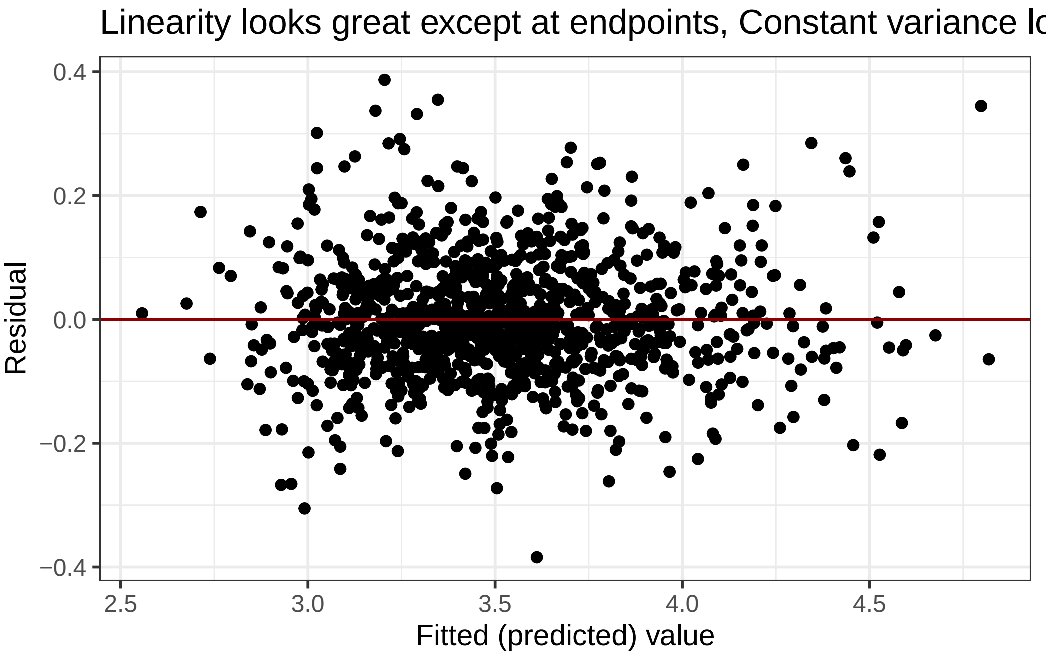
```
Call:
lm(formula = highest_level_qualification_achieved_b_age_22_average_score ~
    level4qual_residents35_64_2011 * level_3_at_age_18 + activity_at_age_19_full_time_higher_
        level4qual_residents35_64_2011 + rgn11nm_combined + level_3_at_age_18 +
        key_stage_4_attainment_school_year_2012_to_2013, data = english_education)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38428 -0.05926 -0.00773  0.05484  0.38699

Coefficients:
                                                         Estimate Std. Error
(Intercept)                                             1.8241957  0.0737405
level4qual_residents35_64_2011Low                       0.2471406  0.0764781
level4qual_residents35_64_2011Medium                    0.2875048  0.0783453
level_3_at_age_18                                       0.0144596  0.0012259
activity_at_age_19_full_time_higher_education           0.0200731  0.0006169
rgn11nm_combinedNorth East                              0.0918015  0.0129024
rgn11nm_combinedNorth West                              0.0571483  0.0093396
rgn11nm_combinedOther                                   0.0323596  0.0071201
key_stage_4_attainment_school_year_2012_to_2013         0.0031308  0.0005040
level4qual_residents35_64_2011Low:level_3_at_age_18    -0.0033080  0.0012279
level4qual_residents35_64_2011Medium:level_3_at_age_18 -0.0040273  0.0012151
                                                       t value Pr(>|t|)
(Intercept)                                             24.738  < 2e-16 ***
level4qual_residents35_64_2011Low                        3.232 0.001268 **
level4qual_residents35_64_2011Medium                     3.670 0.000255 ***
level_3_at_age_18                                       11.795  < 2e-16 ***
activity_at_age_19_full_time_higher_education           32.538  < 2e-16 ***
rgn11nm_combinedNorth East                               7.115 2.03e-12 ***
```

```
rgn11nm_combinedNorth West                                6.119 1.31e-09 ***
rgn11nm_combinedOther                                     4.545 6.11e-06 ***
key_stage_4_attainment_school_year_2012_to_2013           6.212 7.42e-10 ***
level4qual_residents35_64_2011Low:level_3_at_age_18     -2.694 0.007170 ** 
level4qual_residents35_64_2011Medium:level_3_at_age_18  -3.314 0.000949 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09655 on 1088 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.9282,	Adjusted R-squared:  0.9275 
F-statistic:  1406 on 10 and 1088 DF,  p-value: < 2.2e-16
```

```
#transforming to log odds
#english_education <- english_education %>%
#  mutate(log_odds_ks4asy = log(key_stage_4_attainment_school_year_2012_to_2013 / (1 #- ke

#m1 <- lm(education_score ~ population_2011, data = english_education)

#summary(m1)

#m1 <- lm(education_score ~  factor(size_flag), data = english_education)

#summary(m1)

#better option - population_2011
#m1 <- lm(education_score ~  factor(size_flag)+factor(university_flag)+factor###(job_densi
#summary(m1)

#english_education <- na.omit(english_education) #complete case analysis for all variables

english_education$rgn11nm <- as.character(english_education$rgn11nm)

english_education$rgn11nm_combined <- "Other"

english_education$rgn11nm_combined[english_education$rgn11nm %in% c("South East", "South W
english_education$rgn11nm_combined[english_education$rgn11nm == "North East"] <- "North Ea
english_education$rgn11nm_combined[english_education$rgn11nm == "North West"] <- "North We

english_education$rgn11nm_combined <- factor(english_education$rgn11nm_combined)
english_education$rgn11nm_combined <- factor(english_education$rgn11nm_combined)
```

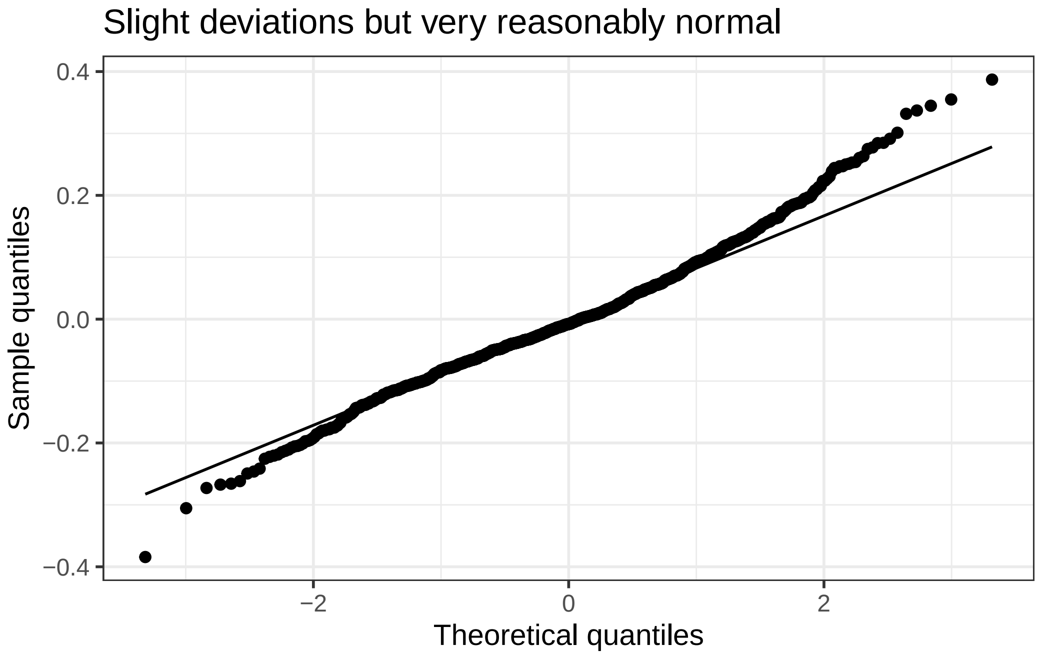
```
english_education$rgn11nm_combined <- relevel(english_education$rgn11nm_combined, ref = "S
```

```
#Residual Plot
ggplot(m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "darkred") +
  labs(x = "Fitted (predicted) value", y = "Residual", title = "Linearity looks great exce
  theme_bw()
```



```
m1_aug <- augment(m1)
```

```
#QQ Plot
ggplot(m1, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw() +
  labs(x = "Theoretical quantiles",
       y = "Sample quantiles", title = "Slight deviations but very reasonably normal")
```


Slight deviations but very reasonably normal

```
#Our custom method of assessing constant variance - splitting into even intervals, calcula
quantiles <- quantile(m1_aug$.fitted, probs = seq(0, 1, by = 0.2))

variance_intervals_df <- data.frame(interval = character(0), variance = numeric(0))

for (i in 1:(length(quantiles) - 1)) {
  subset_data <- m1_aug %>%
    filter(.fitted >= quantiles[i] & .fitted < quantiles[i + 1])

  interval_name <- paste(round(quantiles[i], 2), "-", round(quantiles[i + 1], 2), sep="")
  variance_value <- var(subset_data$.resid)

  variance_intervals_df <- rbind(variance_intervals_df,
                           data.frame(interval=interval_name,
                           variance=variance_value))
}

print(variance_intervals_df)
```

```
   interval    variance
1 2.56-3.22 0.010439236
2  3.22-3.4 0.008603479
```

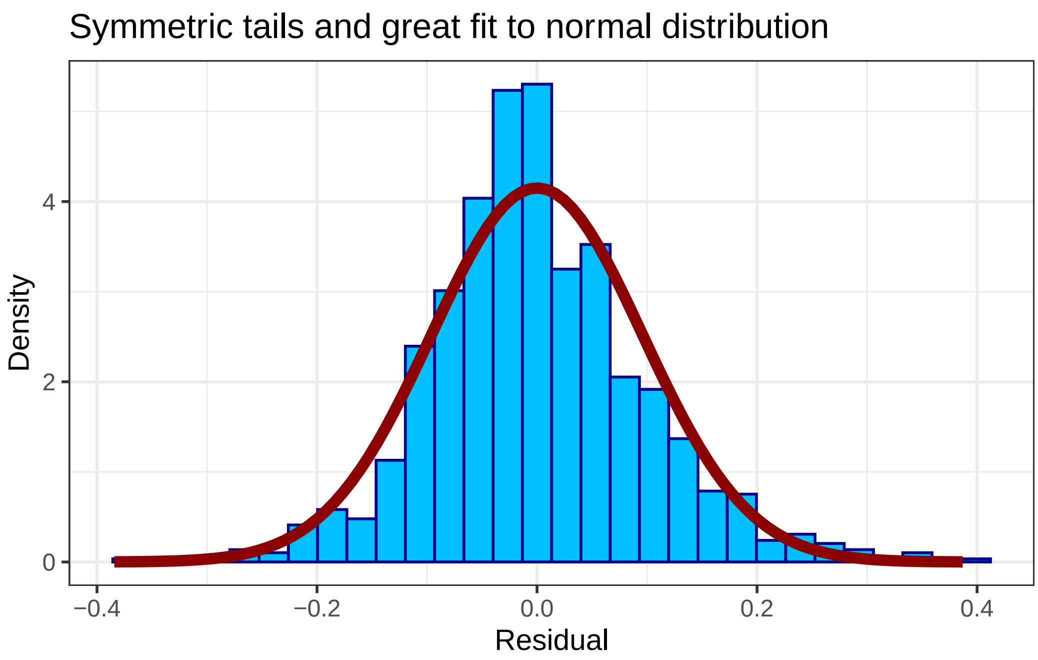
```
3  3.4-3.56 0.007706760
4 3.56-3.78 0.009201645
5 3.78-4.82 0.010143797
```

```r
#Formal test for constant variance
bptest_result <- bptest(m1)

print(bptest_result)
```

```
	studentized Breusch-Pagan test

data:  m1
BP = 27.512, df = 10, p-value = 0.00216
```

```r
resid_mean <- mean(m1_aug$.resid, na.rm = TRUE)
resid_sd <- sd(m1_aug$.resid, na.rm = TRUE)

#Histogram compared to normal distribution
ggplot(m1_aug, aes(x = .resid)) +
  geom_histogram(aes(y = ..density..),
                 fill = "deepskyblue", color = "darkblue", bins = 30) +
  stat_function(fun = dnorm,
                args = list(mean = resid_mean, sd = resid_sd),
                color = "darkred", lwd = 2) +
  labs(x = "Residual", y = "Density", title = "Symmetric tails and great fit to normal dis
  theme_bw()
```

```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
```



```
#transforming back
#predicted_log_odds <- predict(m1, type = "response")
#predicted_proportions <- exp(predicted_log_odds) / (1 + exp(predicted_log_odds))
```

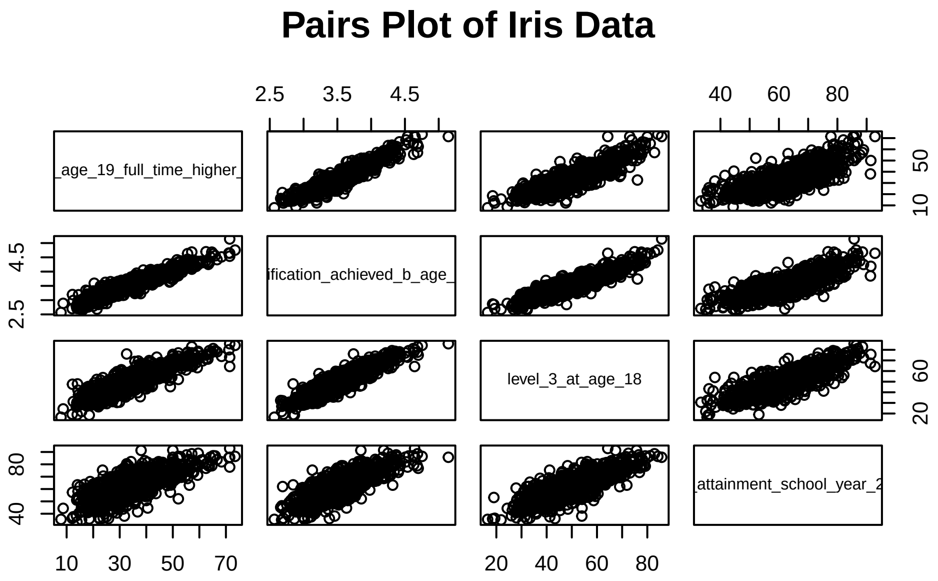
```
#Finding min and max values taken on by variables, columns of the datasaet.
min_max_education_score <- english_education %>%
  summarise(min_education_score = min(education_score, na.rm = TRUE),
            max_education_score = max(education_score, na.rm = TRUE))

min_max_population_2011 <- english_education %>%
  summarise(min_population_2011 = min(population_2011, na.rm = TRUE),
            max_population_2011 = max(population_2011, na.rm = TRUE))

min_max_highest_qualification <- english_education %>%
  summarise(min_highest_qualification = min(highest_level_qualification_achieved_b_age_22_
            max_highest_qualification = max(highest_level_qualification_achieved_b_age_22_

min_max_level_3_age_18 <- english_education %>%
  summarise(min_level_3_age_18 = min(level_3_at_age_18, na.rm = TRUE),
            max_level_3_age_18 = max(level_3_at_age_18, na.rm = TRUE))
```

```
min_max_activity_age_19 <- english_education %>%
  summarise(min_activity_age_19 = min(activity_at_age_19_employment_with_earnings_above_10
            max_activity_age_19 = max(activity_at_age_19_employment_with_earnings_above_10

min_max_key_stage_4 <- english_education %>%
  summarise(min_key_stage_4 = min(key_stage_4_attainment_school_year_2012_to_2013, na.rm =
            max_key_stage_4 = max(key_stage_4_attainment_school_year_2012_to_2013, na.rm =

min_max_df <- bind_rows(
  min_max_education_score,
  min_max_population_2011,
  min_max_highest_qualification,
  min_max_level_3_age_18,
  min_max_activity_age_19,
  min_max_key_stage_4
)

print(min_max_df)
```

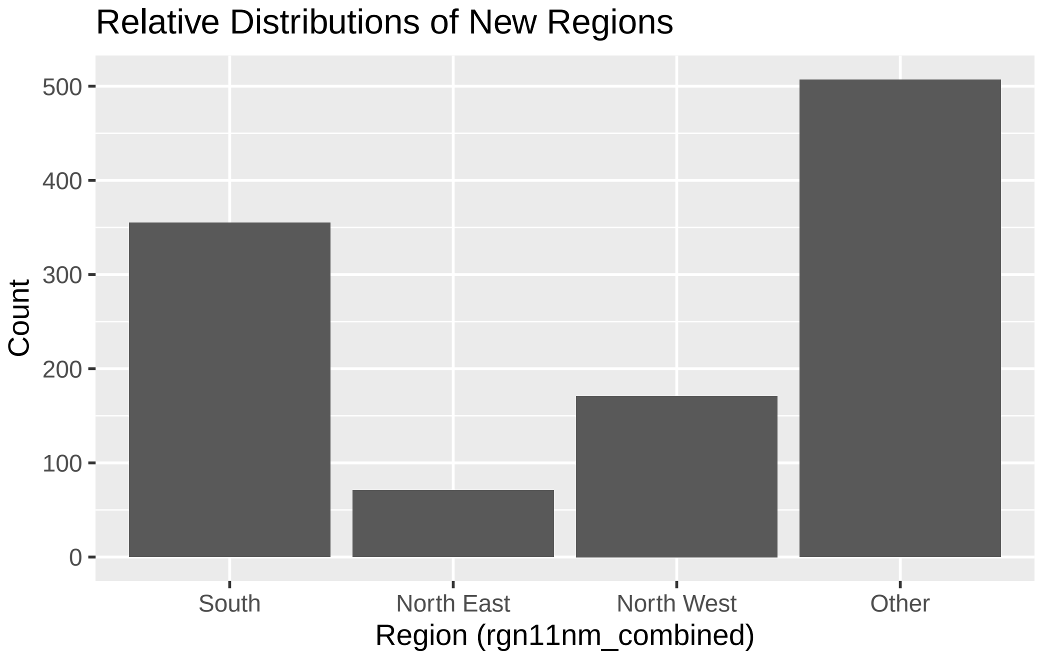
```
# A tibble: 6 x 12
  min_education_score max_education_score min_population_2011
                <dbl>               <dbl>               <dbl>
1               -10.0                11.9                  NA
2                NA                  NA                  5003
3                NA                  NA                    NA
4                NA                  NA                    NA
5                NA                  NA                    NA
6                NA                  NA                    NA
# i 9 more variables: max_population_2011 <dbl>,
#   min_highest_qualification <dbl>, max_highest_qualification <dbl>,
#   min_level_3_age_18 <dbl>, max_level_3_age_18 <dbl>,
#   min_activity_age_19 <dbl>, max_activity_age_19 <dbl>,
#   min_key_stage_4 <dbl>, max_key_stage_4 <dbl>
```

```
edu_pairs <- english_education |>
  select(
    activity_at_age_19_full_time_higher_education,
    highest_level_qualification_achieved_b_age_22_average_score,
    level_3_at_age_18,
    key_stage_4_attainment_school_year_2012_to_2013) |>
  mutate()
```

```
pairs(edu_pairs[, 1:4], main = "Pairs Plot of Iris Data")
```



```
ggplot(english_education, aes(x = factor(rgn11nm_combined))) +
  geom_bar() +
  labs(x = "Region (rgn11nm_combined)", y = "Count", title = "Relative Distributions of Ne
```


Relative Distributions of New Regions
Count
500
400
300
200
100
0
South
North East
North West
Other
Region (rgn11nm_combined)