GROUP 2: FINAL PROJECT

Dania Usman, Jomaris Banua, Shruti Gajre

2024-12-13

1) Data Cleaning

Variable	Reason for Selection
`town11nm`	Identifies the geographical area, essential for studying regional patterns.
`size_flag`	Captures town size, a key variable for analyzing differences in educational outcomes.
`income_flag`	Represents income levels, crucial for examining socioeconomic influences on education.
`uni_flag`	Indicates if a town has a university, potentially impacting higher education access.
`qual_residents`	Reflects the proportion of educated adults, a possible community influence on students.
`GCSEs`	Serves as a proxy for high school graduation rates.
`college_grad`	Represents college graduation rates.

Exclusion Criteria
- Dropped variables not directly related to the research question, such as minor demographic details or redundant information. - Variables with incomplete or irrelevant data for analyzing educational attainment (e.g., administrative or non-educational metrics).

```r
# Selecting relevant columns and renaming variables for clarity
eng_ed <- eng_ed |>
  select(town11nm, size_flag, income_flag, university_flag,
         level4qual_residents35_64_2011,
         key_stage_4_attainment_school_year_2012_to_2013,
         highest_level_qualification_achieved_by_age_22_level_6_or_above,
         ) |>
    mutate(
    size_flag = factor(size_flag),
    income_flag = factor(income_flag),
    uni_flag= factor(university_flag),
    qual_residents=factor(level4qual_residents35_64_2011),
    GCSEs=key_stage_4_attainment_school_year_2012_to_2013,
    college_grad =
      highest_level_qualification_achieved_by_age_22_level_6_or_above
   )
```

```r
#removing repeat columns
eng_ed <- eng_ed |>
  select(town11nm, size_flag, income_flag, uni_flag,qual_residents,
                 GCSEs, college_grad   )
```

Reason for Collapsing Town Size Categories

Collapsed Level	Original Categories Included	Reason
`Large`	"City", "Medium Towns", "Large Towns", "Outer London BUA", "Inner London BUA", "Not BUA"	Reflects similarities in urban characteristics, such as population density and access to resources.
`Small`	"Small Towns", "Other Small BUAs"	Focuses on smaller, less urbanized communities to highlight contrasts with larger urban areas.

Additional Notes

- **Simplification Benefits:** Collapsing categories reduces noise and ensures clearer distinctions between urban and rural-like areas.
- **Comparison Focus:** The new levels, `Large` and `Small`, facilitate easier interpretation of differences in educational attainment based on town size.

```r
#factor collapse to combine mid and large size towns and cities into one level
levels(eng_ed$size_flag)
```

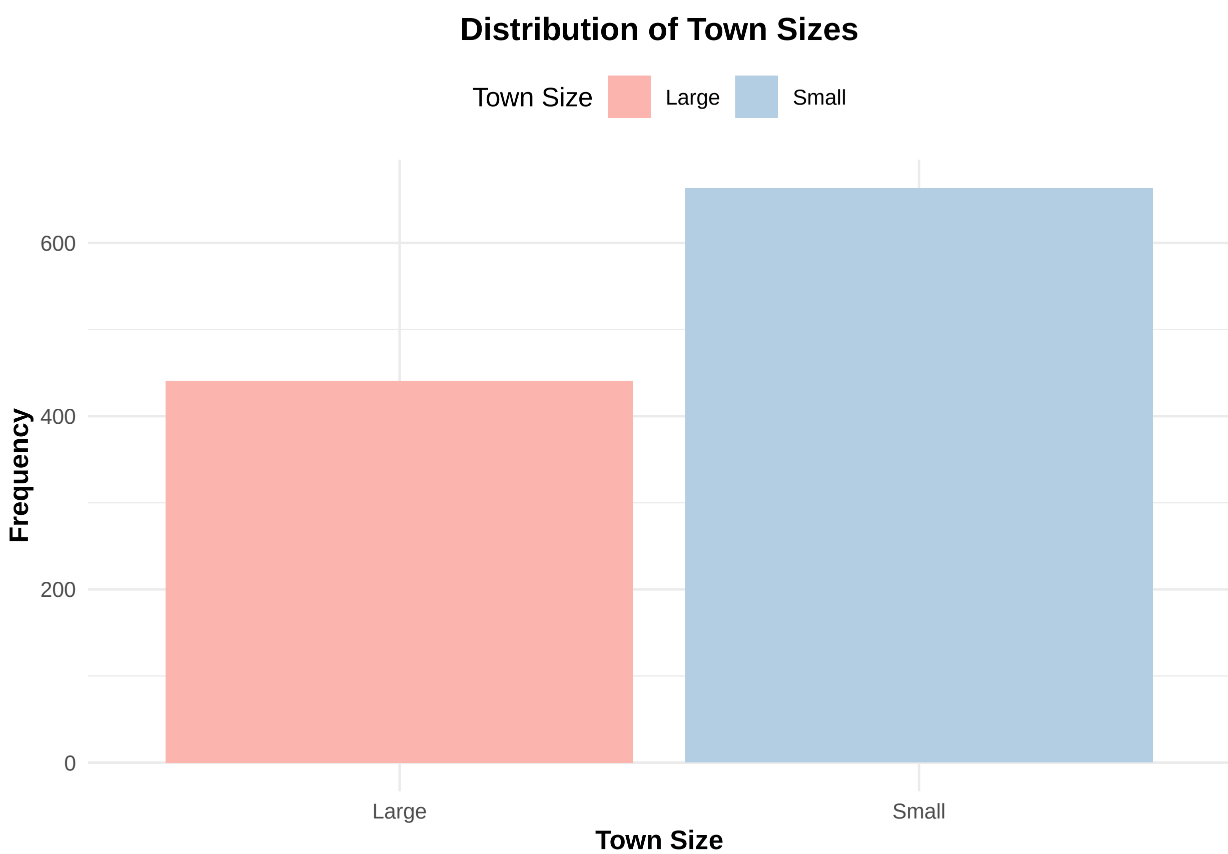
```
## [1] "City"              "Inner London BUA" "Large Towns"      "Medium Towns"
## [5] "Not BUA"           "Other Small BUAs" "Outer london BUA" "Small Towns"
```

```r
eng_ed<- eng_ed |>
mutate(size_flag = fct_collapse(size_flag,
Large = c("City", "Medium Towns", "Large Towns", "Outer london BUA",
          "Inner London BUA", "Not BUA"),
Small = c("Small Towns",  "Other Small BUAs")
))
```

2) EDA

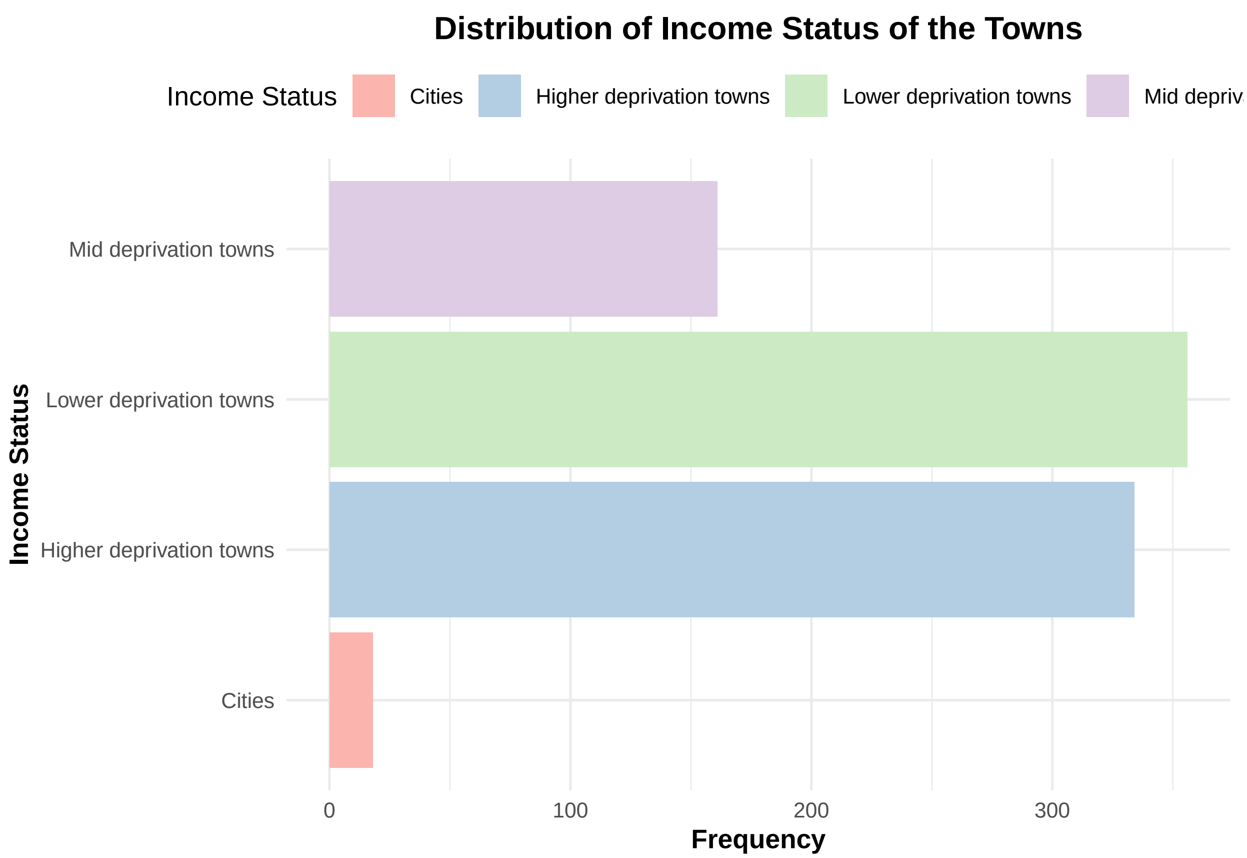
```r
 #distribution of the size_flag
ggplot(data = eng_ed, aes(x = size_flag, fill = size_flag)) +
  geom_bar() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(
    title = "Distribution of Town Sizes",
    x = "Town Size",
    y = "Frequency",
    fill = "Town Size"
```

```
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    legend.position = "top"
  )
```

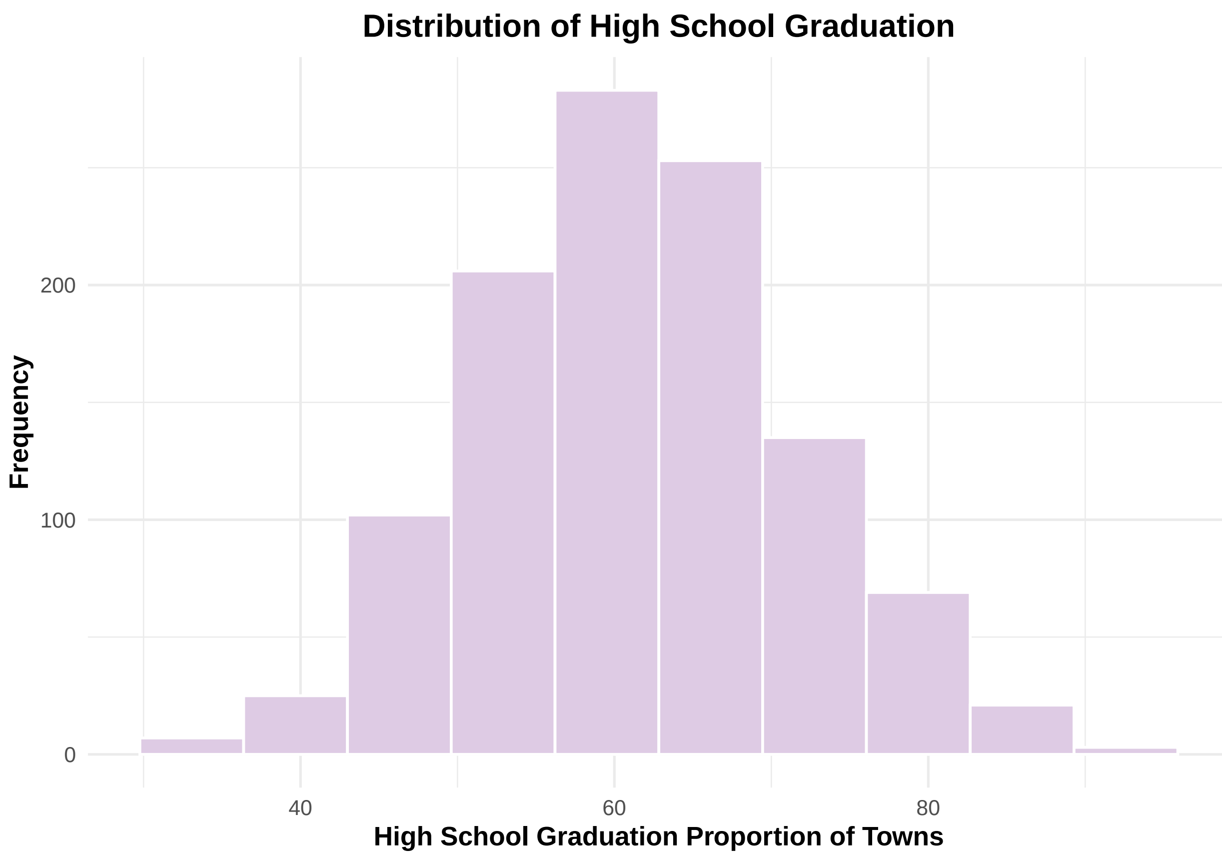



```
#distribution of the income_flag
eng_ed |>
  drop_na() |>
  ggplot(aes(y = income_flag, fill = income_flag)) +
  geom_bar() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(
    title = "Distribution of Income Status of the Towns",
    y = "Income Status",
    x = "Frequency",
    fill = "Income Status"
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
```

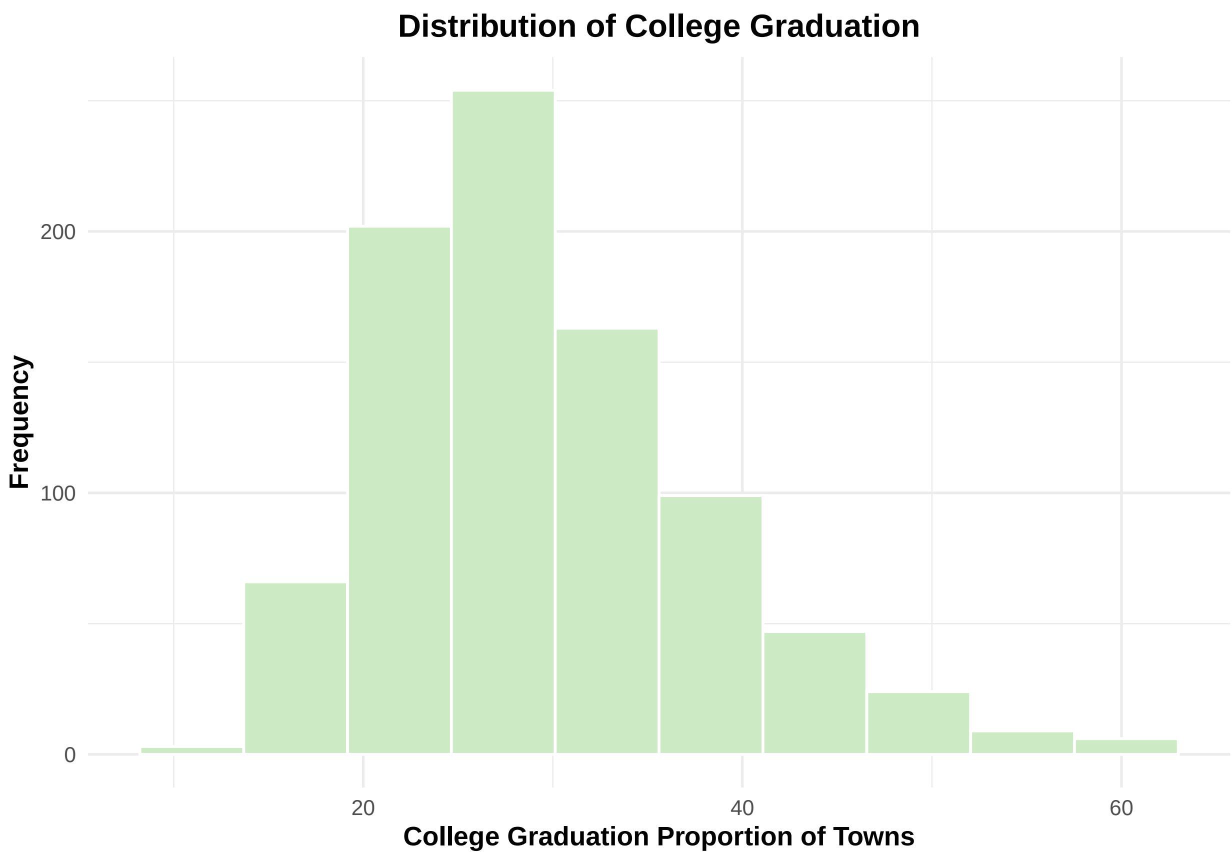
```
    axis.title.y = element_text(face = "bold"),
    legend.position = "top"
  )
```




```
# Distribution of High School Graduation
ggplot(eng_ed, aes(x = GCSEs)) +
  geom_histogram(
    bins = 10,
    col = "white",
    fill = "#DECBE4"
  ) +
  labs(
    title = "Distribution of High School Graduation",
    x = "High School Graduation Proportion of Towns",
    y = "Frequency"
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold")
  )
```




```
# Distribution of College Graduation
ggplot(eng_ed, aes(x = college_grad)) +
  geom_histogram(
    bins = 10,
    col = "white",
    fill = "#CCEBC5"
  ) +
  labs(
    title = "Distribution of College Graduation",
    x = "College Graduation Proportion of Towns",
    y = "Frequency"
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold")
  )
```




```
# Summary Statistics Grouped by Town Size
eng_ed |>
  select(size_flag, income_flag, GCSEs, college_grad) |>
  tbl_summary(
    by = size_flag,
     # Display mean (SD) for continuous variables
    statistic = list(all_continuous() ~ "{mean} ({sd})"),
    # Treat dichotomous variables as categorical
    type = list(all_dichotomous() ~ "categorical"),
    # Round to 2 decimal places
    digits = list(all_continuous() ~ c(2, 2)),
    label = list(
      income_flag = "Income Flag",
      GCSEs = "High School Graduation",
      college_grad = "College Graduation"
    )
  ) |>
  as_kable()
```

Characteristic	**Large** N = 441	**Small** N = 663
Income Flag		
Cities	18 (4.1%)	0 (0%)
Higher deprivation towns	221 (50%)	212 (32%)
Lower deprivation towns	117 (27%)	327 (49%)

Characteristic	**Large** N = 441	**Small** N = 663
Mid deprivation towns	82 (19%)	123 (19%)
Unknown	3	1
High School Graduation	60.20 (8.60)	62.02 (10.94)
College Graduation	27.53 (7.39)	31.59 (9.02)
Unknown	1	230

3) Data Analysis

TEST 1: High School Graduation Rates: Small v Large Towns

Is there a difference in high school graduation rates between small v large towns/cities?

- H0: There is no mean difference between the high school graduation rates between small v large towns/cities
- H1: There is a difference between the high school graduation rates between small v large towns/cities
- Test: Difference between two means

Conditions

- *Independence*: We can assume the observations of one town to another are independent from each other.
- *Normality*: Populations are approximately normal
- *Sample Size*: Both groups have n larger than 30

```
small_town <- eng_ed |>
  filter(size_flag == "Small") #filtered data by small town

large_town <- eng_ed |>
  filter(size_flag == "Large") #filtered data by large town


nrow(small_town)
```

```
## [1] 663
```

```
nrow(large_town)
```

```
## [1] 441
```

Power Analysis

```
mean_small_gcse <- mean(small_town$GCSEs, na.rm = TRUE)
mean_large_gcse <- mean(large_town$GCSEs, na.rm = TRUE)

sd_pooled_gcse <- sqrt((
  (sd(small_town$GCSEs, na.rm = TRUE)^2 +
     sd(large_town$GCSEs, na.rm = TRUE)^2) / 2
))
```

```r
# Cohen's d (effect size)
cohens_d_gcse <- abs(mean_small_gcse - mean_large_gcse) / sd_pooled_gcse

# Sample size
n_small <- nrow(small_town)
n_large <- nrow(large_town)

# Power analysis
power_gcse <- pwr.t.test(
  d = cohens_d_gcse,
  n = min(n_small, n_large),
  sig.level = 0.05,
  type = "two.sample"
)$power

cat(
  "Results Summary for GCSE Difference Test:\n",
  "------------------------------------------\n",
  "Observed Cohen's d (Effect Size):", round(cohens_d_gcse, 3), "\n",
  "Sample Size (Small Towns):", n_small, "\n",
  "Sample Size (Large Towns):", n_large, "\n",
  "Computed Power for GCSE Difference Test Power Analysis", round(power_gcse, 3), "\n"
)
```

```
## Results Summary for GCSE Difference Test:
##  ------------------------------------------
##  Observed Cohen's d (Effect Size): 0.185
##  Sample Size (Small Towns): 663
##  Sample Size (Large Towns): 441
##  Computed Power for GCSE Difference Test Power Analysis 0.782
```

Test for Difference in Means

```r
t_test_result <- t.test(small_town$GCSEs, large_town$GCSEs)
cat("T-test Result: \n")
```

```
## T-test Result:
```

```r
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  small_town$GCSEs and large_town$GCSEs
## t = 3.0771, df = 1071.4, p-value = 0.002144
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6579486 2.9738975
## sample estimates:
## mean of x mean of y
##  62.02080  60.20487
```

- Decision: The p-value is less than 0.05, we reject the null hypothesis
- Conclusion: We have enough evidence that there is a significant difference in high school graduation rates between small towns and large towns/cities.

TEST 2: Difference in College Graduation Rates: Small v Large Towns

Is there a difference in college graduation rates between small v large towns/cities?

- H0: There is no difference in college graduation rates between small vs large towns/cities.
- H1: There is a difference in college graduation rate between small vs large towns/cities.
- Test: Difference between two means

Conditions

- *Independence*: We can assume the observations of one town to another are independent from each other.
- *Normality*: Populations are approximately normal
- *Sample Size*: Both groups have n larger than 30

Power Analysis

```
mean_small_college <- mean(small_town$college_grad, na.rm = TRUE)
mean_large_college <- mean(large_town$college_grad, na.rm = TRUE)

sd_pooled_college <- sqrt((
  (sd(small_town$college_grad, na.rm = TRUE)^2 +
     sd(large_town$college_grad, na.rm = TRUE)^2) / 2
))

# Cohen's d (effect size)
cohens_d_college <- abs(mean_small_college - mean_large_college) /
  sd_pooled_college

# Power analysis for two-sample t-test
power_college <- pwr.t.test(
  d = cohens_d_college,
  n = min(n_small, n_large),
  sig.level = 0.05,
  type = "two.sample"
)$power


cat(
  "Results Summary for College Graduation Difference Test:\n",
  "--------------------------------------------------------\n",
  "Observed Cohen's d (Effect Size):", round(cohens_d_college, 3), "\n",
  "Sample Size (Small Towns):", n_small, "\n",
  "Sample Size (Large Towns):", n_large, "\n",
  "Computed Power for College Graduation Difference Test:", round(power_college, 3), "\n"
)
```

```
## Results Summary for College Graduation Difference Test:
##  ------------------------------------------------------
##  Observed Cohen's d (Effect Size): 0.492
##  Sample Size (Small Towns): 663
##  Sample Size (Large Towns): 441
##  Computed Power for College Graduation Difference Test: 1
```

Test for Difference in Means

```r
t_test_college_grad <- t.test(small_town$college_grad, large_town$college_grad)
cat("T-test Results for College Graduation Rates: \n")
```

```
## T-test Results for College Graduation Rates:
```

```r
print(t_test_college_grad)
```

```
##
##  Welch Two Sample t-test
##
## data:  small_town$college_grad and large_town$college_grad
## t = 7.2603, df = 833.25, p-value = 8.864e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.959436 5.152498
## sample estimates:
## mean of x mean of y
##  31.58938  27.53341
```

- Decision: The p-value is less than 0.05, we reject the null hypothesis
- Conclusion: We enough evidence that there is a difference in college graduation rate between small vs large towns/cities.

TEST 3: Association between Income and Town Size

Are income and town sizes associated?

- H0: Income levels are independent of town size.
- H1: Income levels are associated with town size
- Test: Chi-squared test of independence

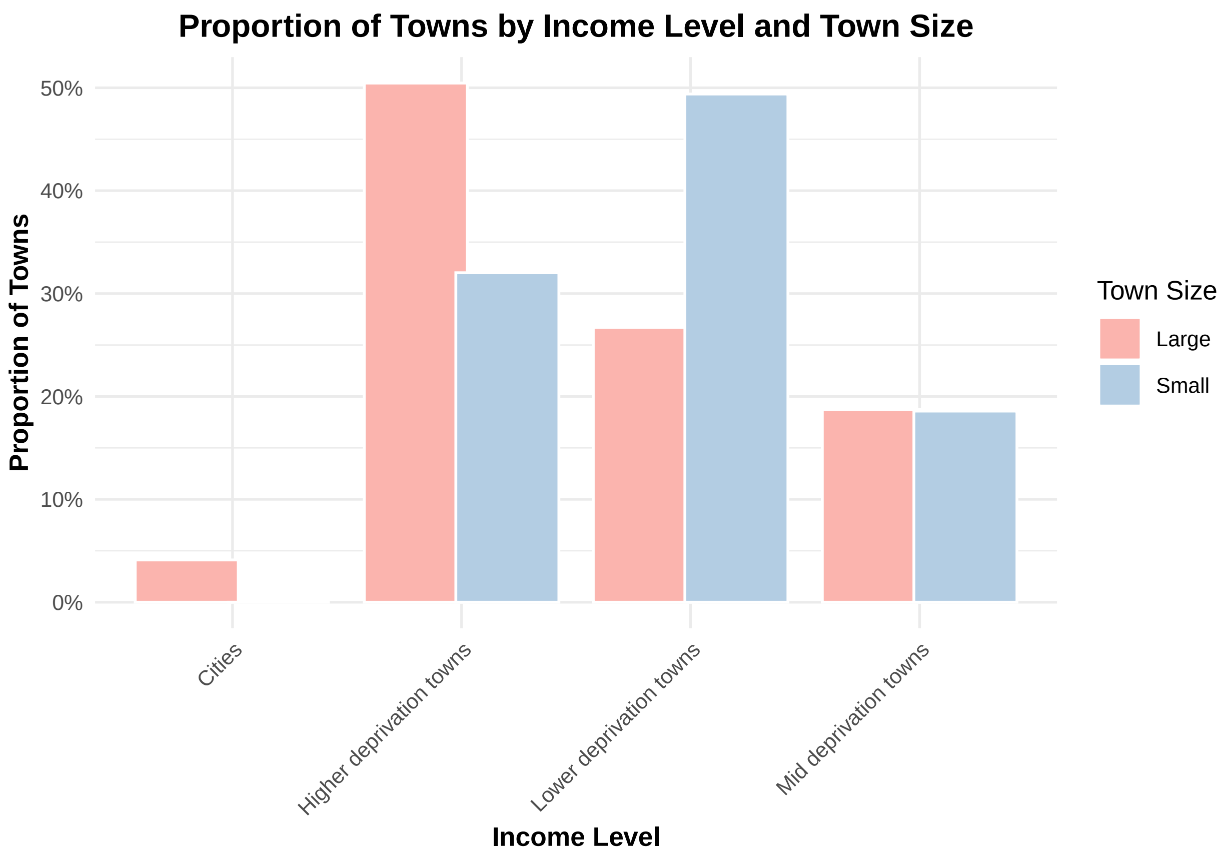
Conditions

- *Independence*: We can assume the observations of one town to another are independent from each other.
- *Expected Counts*: All counts are greater than 5.

```r
income_size_table <- table(eng_ed$size_flag, eng_ed$income_flag)
proportions <- prop.table(income_size_table, margin = 1)

proportions_df <- as.data.frame(as.table(proportions))
colnames(proportions_df) <- c("Town_Size", "Income_Level", "Proportion")

# Bar chart
ggplot(proportions_df, aes(x = Income_Level, y = Proportion,
                           fill = Town_Size)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), col = "white") +
  labs(
    title = "Proportion of Towns by Income Level and Town Size",
    x = "Income Level",
    y = "Proportion of Towns",
    fill = "Town Size"
  ) +
  scale_y_continuous(labels = scales::percent_format()) +
  scale_fill_manual(
    values = c("Small" = "#B3CDE3", "Large" = "#FBB4AE")
  ) +
  theme_minimal(base_size = 10) +  # Consistent font size
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



Power Analysis

```
chisq_test <- chisq.test(eng_ed$size_flag, eng_ed$income_flag)

income_size_table <- table(eng_ed$size_flag, eng_ed$income_flag)

# Calculate Cramér's V
cramers_v <- sqrt(chisq_test$statistic / (sum(income_size_table) *
                                           (min(dim(income_size_table)) - 1)))


# Df
df <- chisq_test$parameter

# Total sample size
n_total <- sum(income_size_table)

# Compute power for chi-squared test
power_chisq <- pwr.chisq.test(
  w = cramers_v,
  N = n_total,
  df = df,
  sig.level = 0.05
)$power
```

```r
cat(
  "Results Summary for Chi-Squared Test:\n",
  "------------------------------------\n",
  "Observed Cramér's V (Effect Size):", round(cramers_v, 3), "\n",
  "Degrees of Freedom:", df, "\n",
  "Total Sample Size:", n_total, "\n",
  "Computed Power for Chi-Squared Test:", round(power_chisq, 3), "\n"
)
```

```
## Results Summary for Chi-Squared Test:
##  ------------------------------------
##  Observed Cramér's V (Effect Size): 0.276
##  Degrees of Freedom: 3
##  Total Sample Size: 1100
##  Computed Power for Chi-Squared Test: 1
```

```r
test <- chisq.test(eng_ed$income_flag, eng_ed$size_flag)
cat("Expected counts:", test$expected)
```

```
## Expected counts: 7.167273 172.4127 176.7927 81.62727 10.83273 260.5873 267.2073 123.3727
```

```r
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  eng_ed$income_flag and eng_ed$size_flag
## X-squared = 83.562, df = 3, p-value < 2.2e-16
```

- Decision: We reject the null hypothesis.
- Conclusion: There is strong statistical evidence to conclude that income levels and town size are not independent.

TEST 4: Relationship between College Graduation Rate & Highschool Graduation Rate

Is high school completion a good predictor of college degree completion?

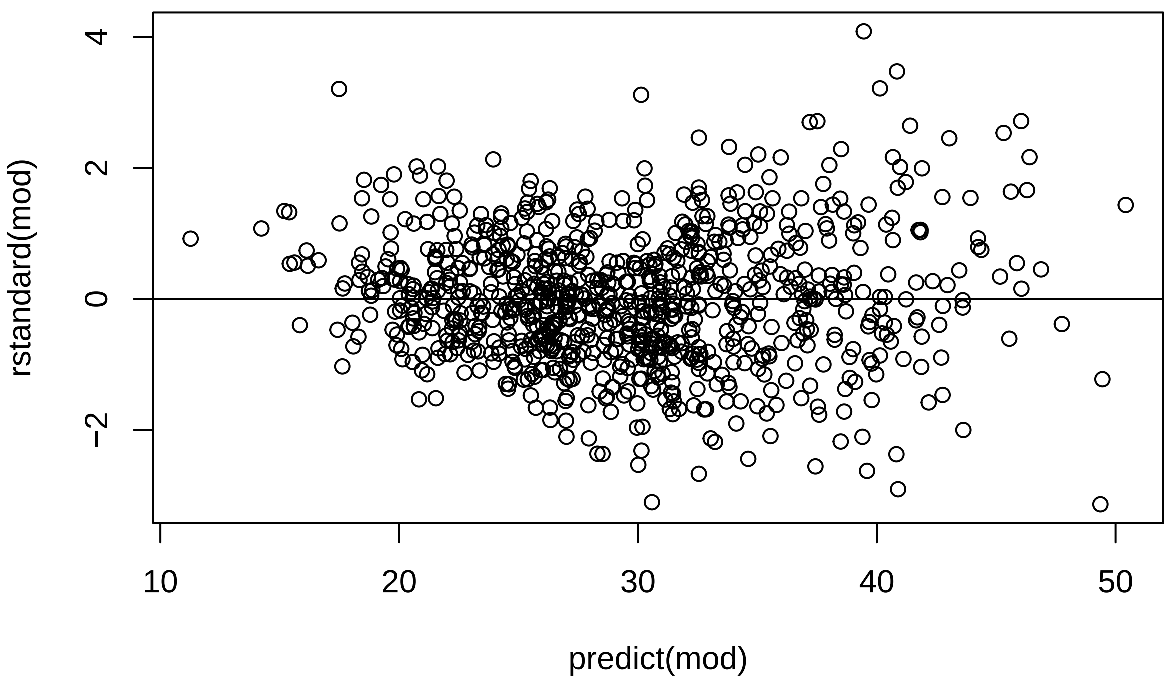
- H0: There is no linear relationship between high school and college completion.
- H1: There is a linear relationship between high school and college completion.
- Test: Linear Regression

Conditions

- *Linearity*: Data appears to be linear
- *Independence:* Errors seem to have no pattern
- *Constant Variance*: Seems Constant
- *Normality*: The qq plots appear normal

```r
#Independence

mod<- lm(college_grad ~ GCSEs, data =eng_ed)
plot(predict(mod), rstandard(mod))
abline(h=0)
```

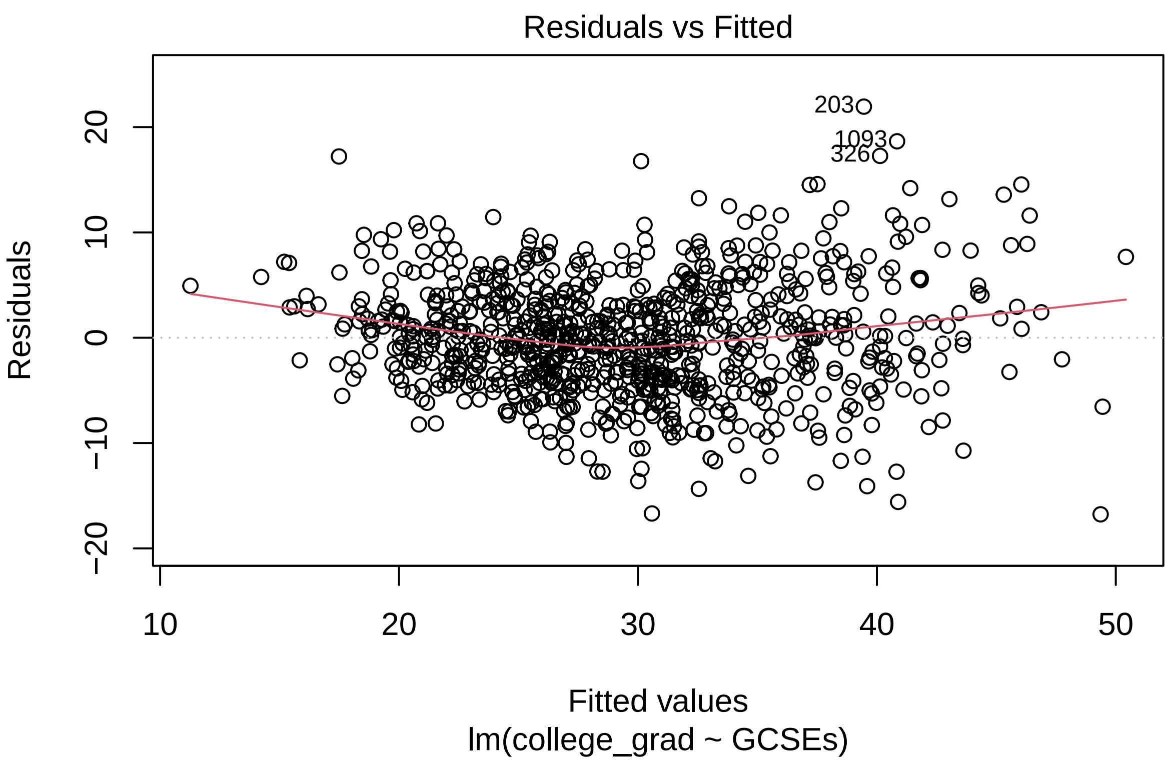


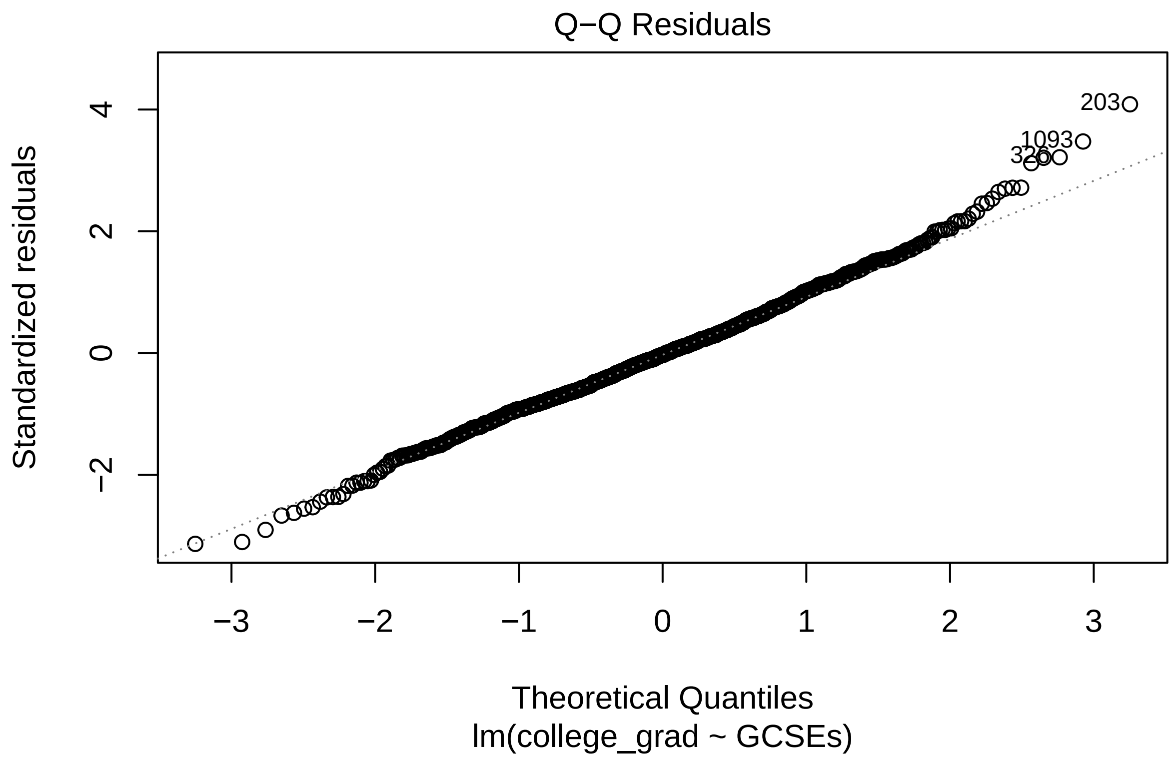

```r
summary(mod)
```

```
##
## Call:
## lm(formula = college_grad ~ GCSEs, data = eng_ed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.765  -3.612  -0.180   3.296  21.944
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.94830    1.19441  -10.84   <2e-16 ***
## GCSEs         0.68247    0.01896   36.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.38 on 871 degrees of freedom
##   (231 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.598,  Adjusted R-squared:  0.5976
## F-statistic:  1296 on 1 and 871 DF,  p-value: < 2.2e-16
```

```r
#constant variance
#normality
plot(mod, 1:2)
```





Power Analysis

```
r_squared <- summary(mod)$r.squared

# Size (Cohen's f2)
f2 <- r_squared / (1 - r_squared)

n <- nrow(eng_ed)
num_predictors <- 1

# Power analysis
power_regression <- pwr.f2.test(
  u = num_predictors,
  v = n - num_predictors - 1,
  f2 = f2,
  sig.level = 0.05
)$power

cat(
  "Results Summary for Regression Test:\n",
  "-------------------------------------\n",
  "R-squared (Goodness-of-Fit):", round(r_squared, 3), "\n",
  "Effect Size (Cohen's f²):", round(f2, 3), "\n",
  "Sample Size:", n, "\n",
```

```
  "Computed Power for Regression Test:", round(power_regression, 3), "\n"
)
```

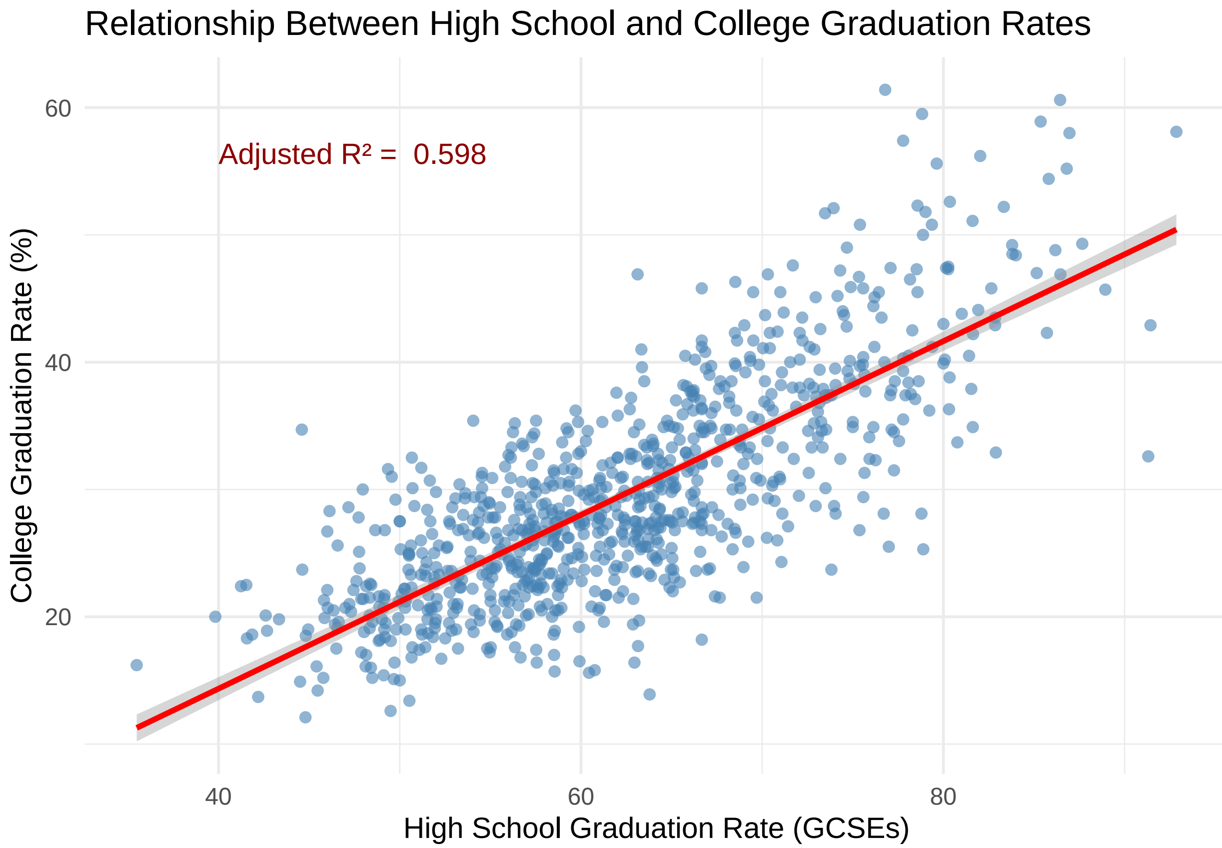
```
## Results Summary for Regression Test:
##  -------------------------------------
##  R-squared (Goodness-of-Fit): 0.598
##  Effect Size (Cohen's f²): 1.488
##  Sample Size: 1104
##  Computed Power for Regression Test: 1
```

```
clean_data <- eng_ed[!is.na(eng_ed$GCSEs) & !is.na(eng_ed$college_grad), ]

# Fit the linear regression model
model <- lm(college_grad ~ GCSEs, data = clean_data)

# Calculate Adjusted R-squared
adj_r2 <- summary(model)$adj.r.squared

# Create scatter plot with regression line
ggplot(clean_data, aes(x = GCSEs, y = college_grad)) +
  geom_point(color = "steelblue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  annotate("text", x = 40, y = max(clean_data$college_grad) - 5,
           label = paste("Adjusted R² = ", round(adj_r2, 3)),
           color = "darkred", size = 4, hjust = 0) +
  labs(
    title = "Relationship Between High School and College Graduation Rates",
    x = "High School Graduation Rate (GCSEs)",
    y = "College Graduation Rate (%)"
  ) +
  theme_minimal()
```



- Decision: We reject the null hypothesis.
- Conclusion: The results provide strong evidence of a statistically significant positive linear relationship between high school graduation rates (`GCSEs`) and college graduation rates (`college_grad`). This indicates that high school completion is a good predictor of college degree completion. Specifically, the model estimates that for every 1% increase in high school graduation rates, college graduation rates increase by approximately 0.682%, on average. The model explains 59.8% of the variability in college graduation rates ($R^2 = 0.598$).