# SG_Project1

Sidney Gehring

2025-02-05

```r
# Import dependencies
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.95 loaded
```

```r
library(tidyr)
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.3.3
```

```r
theme_set(theme_bw())

# Import data
data <- read.csv("mobility-all.csv", header = TRUE)
```

```r
glimpse(data)
```

```
## Rows: 741
## Columns: 43
## $ ID                           <int> 100, 200, 301, 302, 401, 402, 500, 601, 602,~
## $ Name                         <chr> "Johnson City", "Morristown", "Middlesboroug~
## $ Mobility                     <dbl> 0.06219881, 0.05365194, 0.07263514, 0.056281~
## $ State                     <chr> "TN", "TN", "TN", "TN", "NC", "VA", "NC", "N~
## $ Population                <int> 576081, 227816, 66708, 727600, 493180, 92753~
## $ Urban                     <int> 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1,~
## $ Black                     <dbl> 0.021, 0.020, 0.015, 0.056, 0.174, 0.224, 0.~
## $ Seg_racial                <dbl> 0.090, 0.093, 0.064, 0.210, 0.262, 0.137, 0.~
## $ Seg_income                <dbl> 0.035, 0.026, 0.024, 0.092, 0.072, 0.024, 0.~
## $ Seg_poverty               <dbl> 0.030, 0.028, 0.015, 0.084, 0.061, 0.015, 0.~
## $ Seg_affluence             <dbl> 0.038, 0.025, 0.026, 0.102, 0.081, 0.028, 0.~
## $ Commute                   <dbl> 0.325, 0.276, 0.359, 0.269, 0.292, 0.313, 0.~
## $ Income                    <int> 31560, 29959, 22328, 35884, 38892, 31265, 36~
## $ Gini                      <dbl> 0.468, 0.435, 0.441, 0.508, 0.466, 0.444, 0.~
## $ Share01                   <dbl> 13.459, 10.631, 10.691, 15.080, 11.917, 10.6~
## $ Gini_99                   <dbl> 0.333, 0.328, 0.334, 0.358, 0.346, 0.338, 0.~
## $ Middle_class              <dbl> 0.548, 0.538, 0.467, 0.504, 0.500, 0.538, 0.~
## $ Local_tax_rate            <dbl> 0.020, 0.023, 0.015, 0.019, 0.018, 0.015, 0.~
## $ Local_gov_spending        <int> 1886, 2004, 1190, 2357, 1891, 1558, 1932, 16~
## $ Progressivity             <dbl> 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1,~
## $ EITC                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ School_spending           <dbl> 5.185, 4.506, 5.614, 4.900, 5.463, 5.740, 5.~
## $ Student_teacher_ratio     <dbl> NA, NA, 15.1, NA, 15.4, NA, 16.7, 16.2, 12.3~
## $ Test_scores               <dbl> 2.728, -3.400, -9.315, -6.032, -2.297, 4.227~
## $ HS_dropout                <dbl> -0.015, -0.024, -0.005, -0.011, 0.023, NA, 0~
## $ Colleges                  <dbl> 0.014, 0.009, 0.045, 0.011, 0.014, 0.011, 0.~
## $ Tuition                   <int> 4817, 4762, 11840, 3480, 9715, 1113, 4528, 8~
## $ Graduation                <dbl> -0.002, -0.101, 0.111, -0.024, 0.052, -0.116~
## $ Labor_force_participation <dbl> 0.587, 0.625, 0.479, 0.615, 0.656, 0.599, 0.~
## $ Manufacturing             <dbl> 0.237, 0.238, 0.234, 0.146, 0.215, 0.395, 0.~
## $ Chinese_imports           <dbl> 5.294, 3.030, 2.063, 1.078, 1.016, 3.277, 2.~
## $ Teenage_labor             <dbl> 0.004, 0.005, 0.003, 0.004, 0.004, 0.003, 0.~
## $ Migration_in              <dbl> 0.006, 0.016, 0.008, 0.016, 0.022, 0.007, 0.~
## $ Migration_out             <dbl> 0.005, 0.014, 0.012, 0.014, 0.019, 0.010, 0.~
## $ Foreign_born              <dbl> 0.012, 0.023, 0.007, 0.020, 0.053, 0.025, 0.~
## $ Social_capital            <dbl> -0.298, -0.767, -1.270, -0.222, -0.018, -0.9~
## $ Religious                 <dbl> 0.514, 0.544, 0.668, 0.602, 0.488, 0.454, 0.~
## $ Violent_crime             <dbl> 0.001, 0.002, 0.001, 0.001, 0.003, 0.002, 0.~
## $ Single_mothers            <dbl> 0.190, 0.185, 0.211, 0.206, 0.220, 0.241, 0.~
## $ Divorced                  <dbl> 0.110, 0.116, 0.113, 0.114, 0.092, 0.096, 0.~
## $ Married                   <dbl> 0.601, 0.613, 0.590, 0.575, 0.586, 0.580, 0.~
## $ Longitude                 <dbl> -82.43639, -83.40725, -83.53533, -84.24279, ~
## $ Latitude                  <dbl> 36.47037, 36.09654, 36.55154, 35.95226, 36.0~
```

```r
str(data)
```

```
## 'data.frame':    741 obs. of  43 variables:
##  $ ID                       : int  100 200 301 302 401 402 500 601 602 700 ...
##  $ Name                     : chr  "Johnson City" "Morristown" "Middlesborough" "Knoxville" ...
##  $ Mobility                 : num  0.0622 0.0537 0.0726 0.0563 0.0448 ...
##  $ State                    : chr  "TN" "TN" "TN" "TN" ...
##  $ Population               : int  576081 227816 66708 727600 493180 92753 1055133 90016 64676 354533
##  $ Urban                    : int  1 1 0 1 1 0 1 0 0 1 ...
##  $ Black                    : num  0.021 0.02 0.015 0.056 0.174 0.224 0.218 0.032 0.029 0.207 ...
##  $ Seg_racial               : num  0.09 0.093 0.064 0.21 0.262 0.137 0.22 0.114 0.131 0.139 ...
##  $ Seg_income               : num  0.035 0.026 0.024 0.092 0.072 0.024 0.068 0.012 0.005 0.045 ...
##  $ Seg_poverty              : num  0.03 0.028 0.015 0.084 0.061 0.015 0.058 0.009 0.004 0.044 ...
##  $ Seg_affluence           : num  0.038 0.025 0.026 0.102 0.081 0.028 0.077 0.012 0.006 0.045 ...
##  $ Commute                 : num  0.325 0.276 0.359 0.269 0.292 0.313 0.305 0.289 0.325 0.299 ...
##  $ Income                  : int  31560 29959 22328 35884 38892 31265 36582 31544 30683 33417 ...
##  $ Gini                    : num  0.468 0.435 0.441 0.508 0.466 0.444 0.524 0.446 0.356 0.471 ...
##  $ Share01                 : num  13.5 10.6 10.7 15.1 11.9 ...
##  $ Gini_99                 : num  0.333 0.328 0.334 0.358 0.346 0.338 0.341 0.32 0.269 0.341 ...
##  $ Middle_class            : num  0.548 0.538 0.467 0.504 0.5 0.538 0.51 0.56 0.608 0.529 ...
##  $ Local_tax_rate          : num  0.02 0.023 0.015 0.019 0.018 0.015 0.017 0.014 0.014 0.018 ...
##  $ Local_gov_spending      : int  1886 2004 1190 2357 1891 1558 1932 1661 1208 2499 ...
##  $ Progressivity           : num  0 0 0 0 1 0 1 1 0 0 ...
##  $ EITC                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ School_spending         : num  5.18 4.51 5.61 4.9 5.46 ...
##  $ Student_teacher_ratio   : num  NA NA 15.1 NA 15.4 NA 16.7 16.2 12.3 15.9 ...
##  $ Test_scores             : num  2.73 -3.4 -9.31 -6.03 -2.3 ...
##  $ HS_dropout              : num  -0.015 -0.024 -0.005 -0.011 0.023 NA 0.016 0.021 NA NA ...
##  $ Colleges                : num  0.014 0.009 0.045 0.011 0.014 0.011 0.014 0.011 NA 0.02 ...
##  $ Tuition                 : int  4817 4762 11840 3480 9715 1113 4528 880 NA 7264 ...
##  $ Graduation              : num  -0.002 -0.101 0.111 -0.024 0.052 -0.116 -0.017 -0.123 NA 0.007 ..
##  $ Labor_force_participation: num  0.587 0.625 0.479 0.615 0.656 0.599 0.666 0.617 0.594 0.63 ...
##  $ Manufacturing           : num  0.237 0.238 0.234 0.146 0.215 0.395 0.261 0.275 0.321 0.295 ...
##  $ Chinese_imports         : num  5.29 3.03 2.06 1.08 1.02 ...
##  $ Teenage_labor           : num  0.004 0.005 0.003 0.004 0.004 0.003 0.004 0.003 0.004 0.004 ...
##  $ Migration_in            : num  0.006 0.016 0.008 0.016 0.022 0.007 0.017 0.012 0.006 0.017 ...
##  $ Migration_out           : num  0.005 0.014 0.012 0.014 0.019 0.01 0.015 0.012 0.006 0.016 ...
##  $ Foreign_born            : num  0.012 0.023 0.007 0.02 0.053 0.025 0.05 0.027 0.023 0.029 ...
##  $ Social_capital          : num  -0.298 -0.767 -1.27 -0.222 -0.018 ...
##  $ Religious               : num  0.514 0.544 0.668 0.602 0.488 0.454 0.434 0.561 0.43 0.596 ...
##  $ Violent_crime           : num  0.001 0.002 0.001 0.001 0.003 0.002 0.003 0.003 0.001 0.003 ...
##  $ Single_mothers          : num  0.19 0.185 0.211 0.206 0.22 0.241 0.237 0.165 0.167 0.246 ...
##  $ Divorced                : num  0.11 0.116 0.113 0.114 0.092 0.096 0.096 0.087 0.089 0.099 ...
##  $ Married                 : num  0.601 0.613 0.59 0.575 0.586 0.58 0.56 0.632 0.622 0.561 ...
##  $ Longitude               : num  -82.4 -83.4 -83.5 -84.2 -80.5 ...
##  $ Latitude                : num  36.5 36.1 36.6 36 36.1 ...
```

```
new_data <- data %>% select(Mobility, School_spending, Student_teacher_ratio, Test_scores, HS_dropout,

missing_values <- colSums(is.na(new_data))
print(missing_values)
```

```
##              Mobility       School_spending Student_teacher_ratio
##                    12                    10                    30
##           Test_scores            HS_dropout              Colleges
##                    36                   148                   157
##               Tuition            Graduation
##                   161                   160
```

```
mobility_data <- new_data %>% drop_na()
colSums(is.na(mobility_data))
```

```
##              Mobility       School_spending Student_teacher_ratio
##                     0                     0                     0
##           Test_scores            HS_dropout              Colleges
##                     0                     0                     0
##                Tuition            Graduation
##                      0                     0
```

```
glimpse(mobility_data)
```

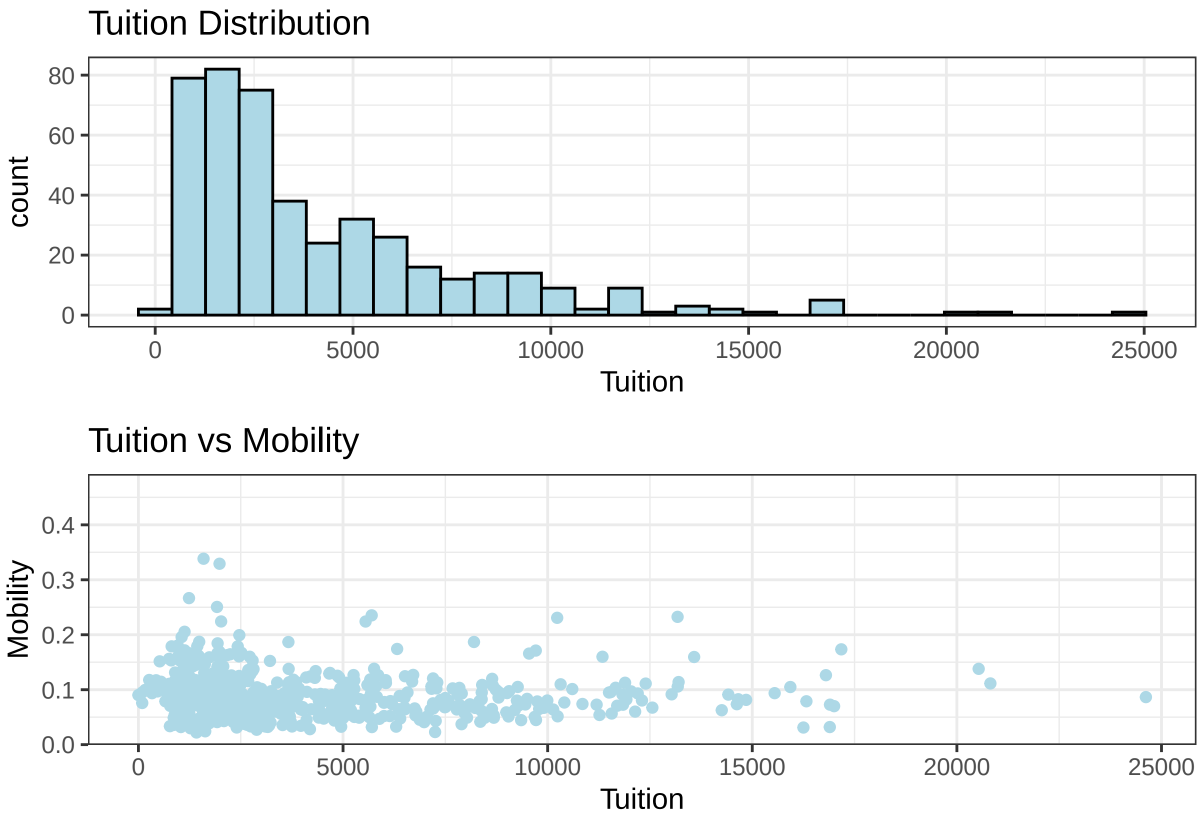
```
## Rows: 449
## Columns: 8
## $ Mobility              <dbl> 0.07263514, 0.04480079, 0.04739694, 0.05166313, ~
## $ School_spending       <dbl> 5.614, 5.463, 5.008, 5.296, 5.251, 5.674, 5.264,~
## $ Student_teacher_ratio <dbl> 15.1, 15.4, 16.7, 16.2, 16.4, 17.0, 13.9, 15.7, ~
## $ Test_scores           <dbl> -9.315, -2.297, -3.716, 6.698, 2.412, -2.291, 9.~
## $ HS_dropout            <dbl> -0.005, 0.023, 0.016, 0.021, 0.030, 0.023, 0.013~
## $ Colleges              <dbl> 0.045, 0.014, 0.014, 0.011, 0.015, 0.013, 0.026,~
## $ Tuition               <int> 11840, 9715, 4528, 880, 10244, 4790, 1674, 888, ~
## $ Graduation            <dbl> 0.111, 0.052, -0.017, -0.123, -0.102, -0.021, 0.~
```

```
p1 <- ggplot(mobility_data, aes(x= Tuition)) + geom_histogram(bins = 30, fill = 'lightblue', color = 'bl

p2 <- ggplot(data, aes(x = Tuition, y= Mobility)) + geom_point(color = "lightblue") + ggtitle("Tuition v

p1 / p2
```

```
## Warning: Removed 161 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

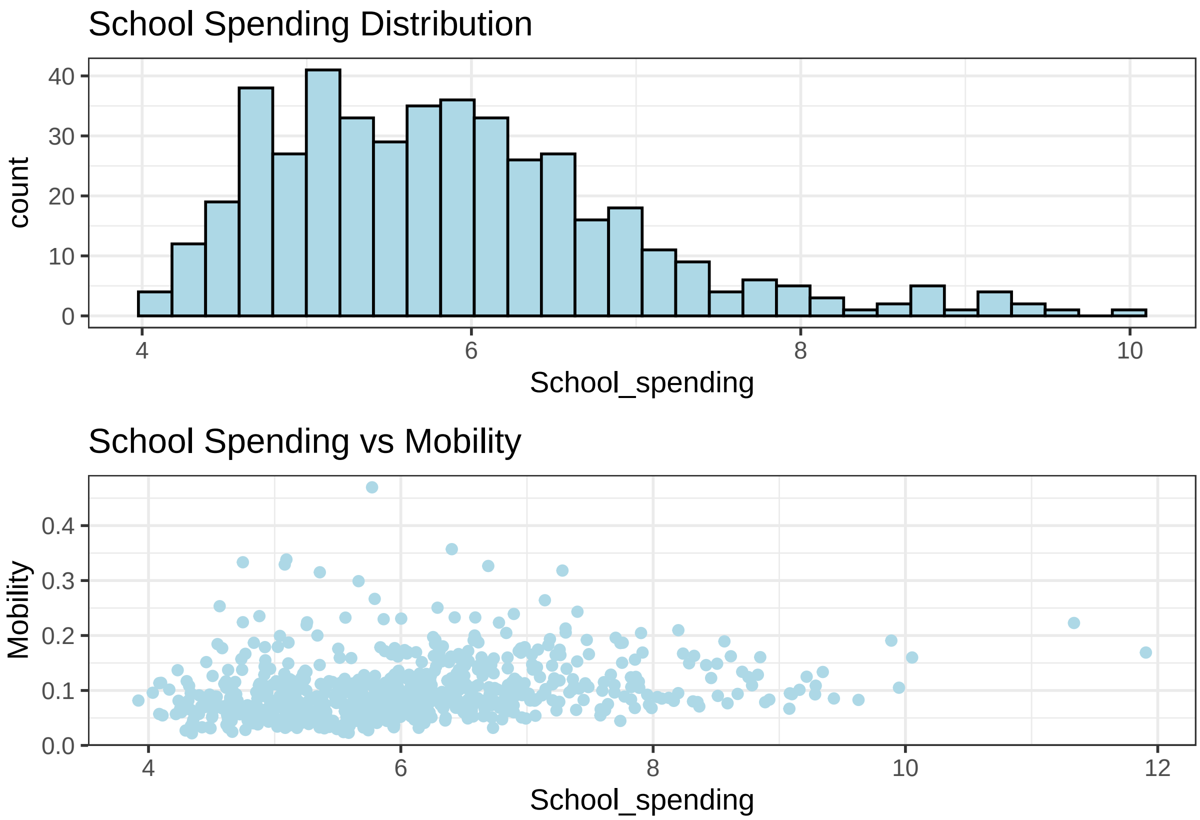


```
p1 <- ggplot(mobility_data, aes(x= School_spending)) + geom_histogram(bins = 30, fill = 'lightblue', co

p2 <- ggplot(data, aes(x = School_spending, y= Mobility)) + geom_point(color = "lightblue")  + ggtitle(

p1 / p2
```

```
## Warning: Removed 21 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


School Spending Distribution

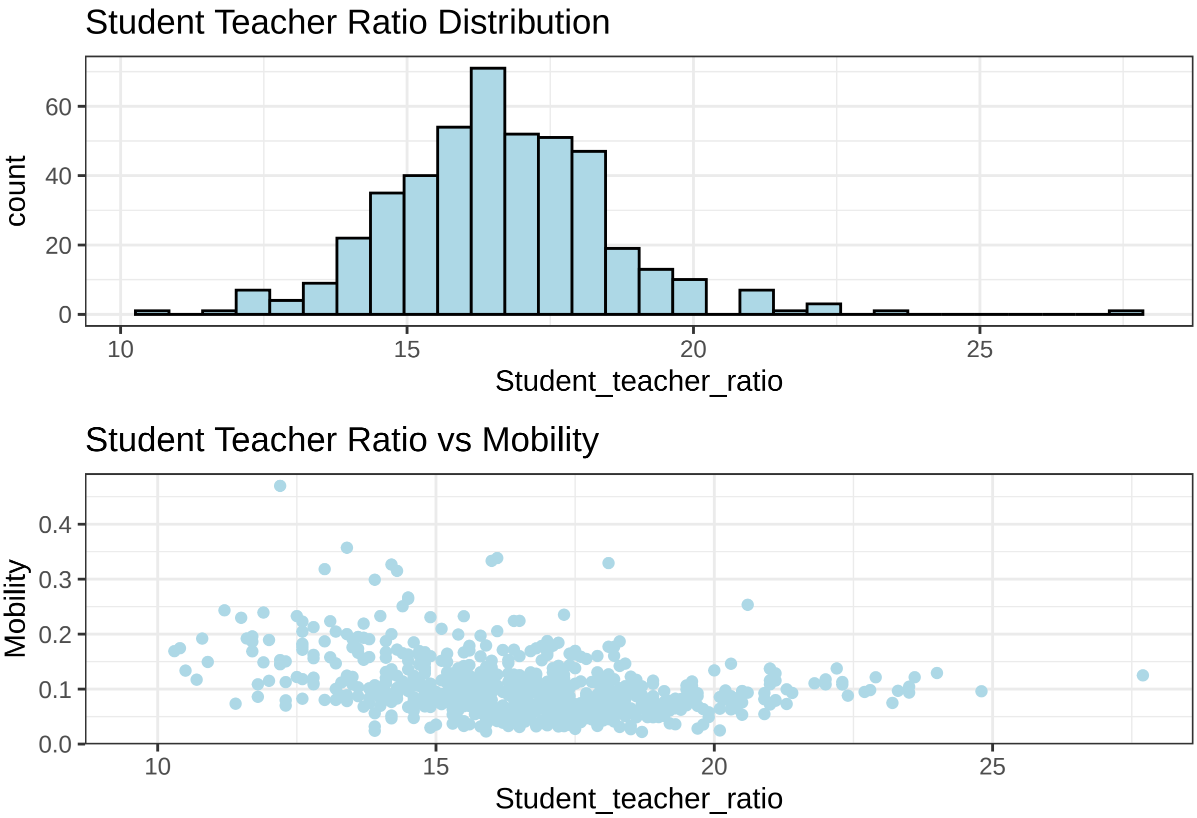
School Spending vs Mobility

```
p1 <- ggplot(mobility_data, aes(x= Student_teacher_ratio)) + geom_histogram(bins = 30, fill = 'lightblu

p2 <- ggplot(data, aes(x = Student_teacher_ratio, y= Mobility)) + geom_point(color = "lightblue")  + gg

p1 / p2
```

```
## Warning: Removed 42 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

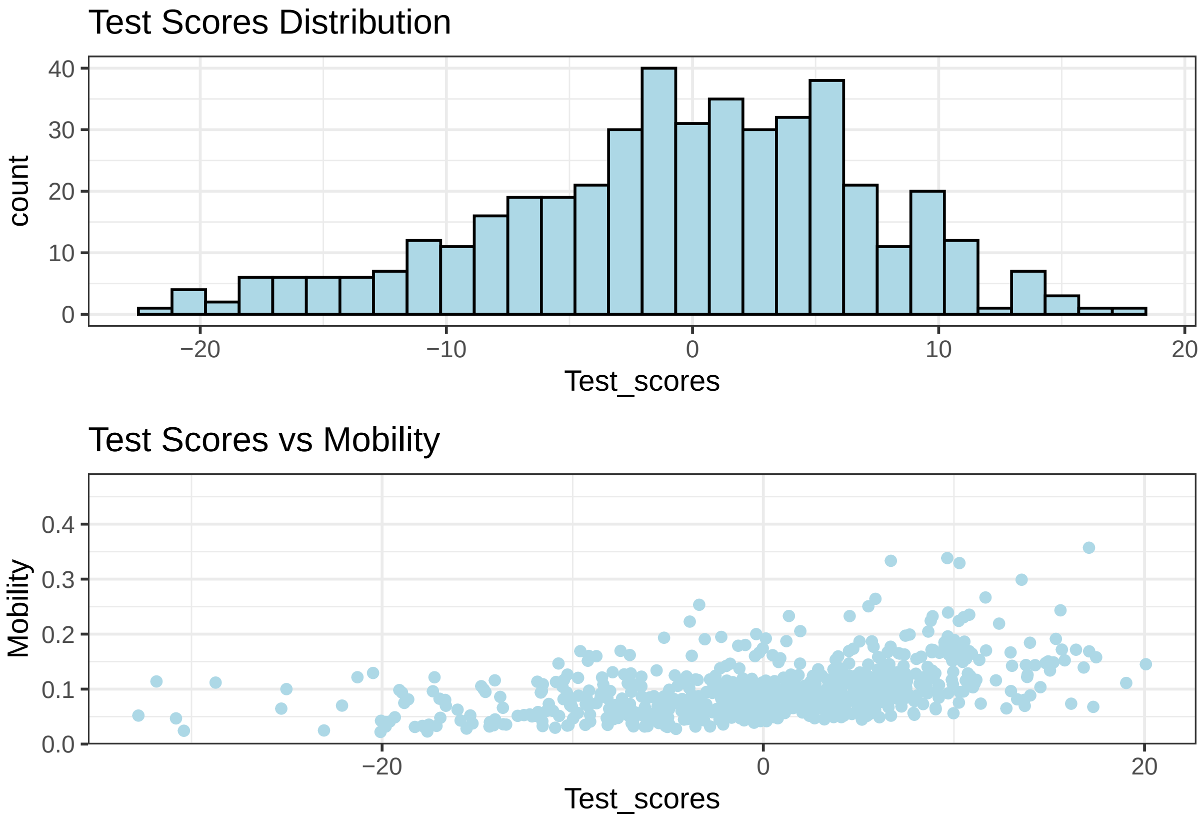


```
p1 <- ggplot(mobility_data, aes(x= Test_scores)) + geom_histogram(bins = 30, fill = 'lightblue', color =

p2 <- ggplot(data, aes(x = Test_scores, y= Mobility)) + geom_point(color = "lightblue")  + ggtitle("Tes

p1 / p2
```

```
## Warning: Removed 36 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

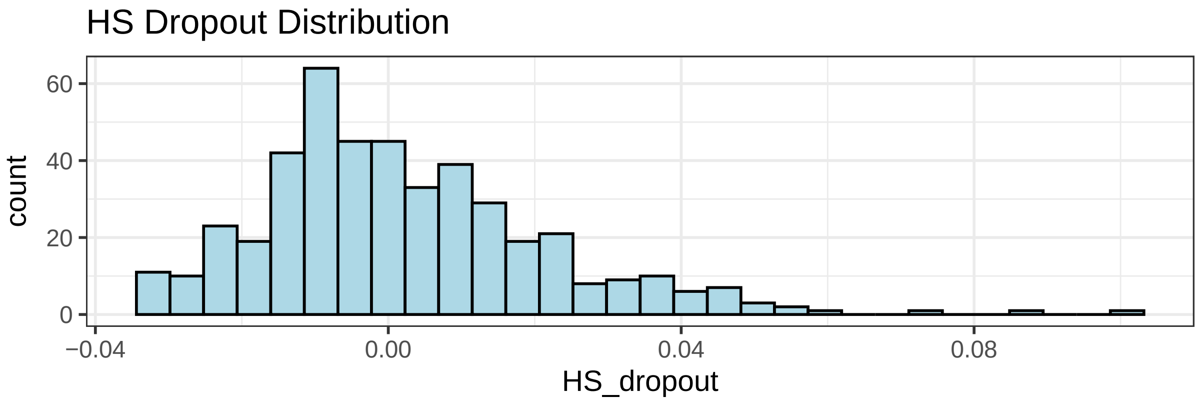


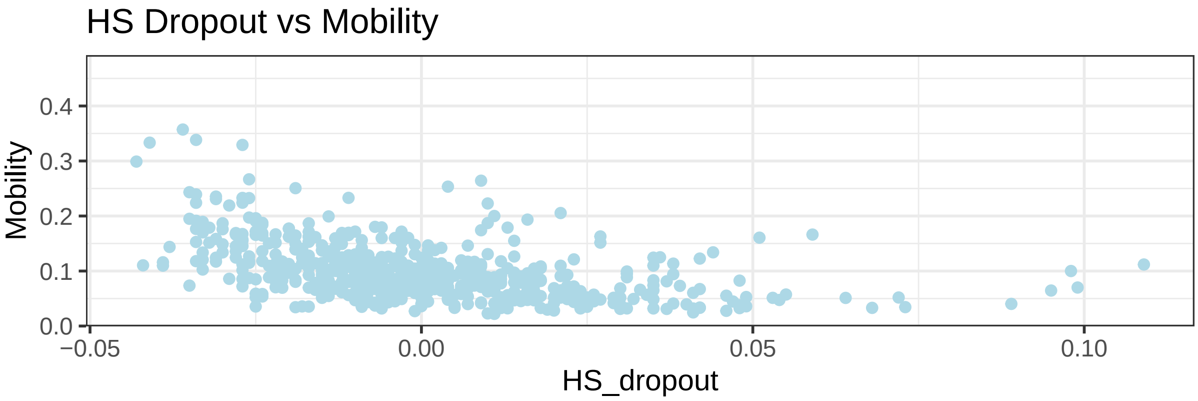
```
p1 <- ggplot(mobility_data, aes(x= HS_dropout)) + geom_histogram(bins = 30, fill = 'lightblue', color =

p2 <- ggplot(data, aes(x = HS_dropout, y= Mobility)) + geom_point(color = "lightblue")  + ggtitle("HS Dı

p1 / p2
```

```
## Warning: Removed 148 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```





```
cat("The tuition mean is", mean(mobility_data$Tuition, na.rm = TRUE), "\n")
```

```
## The tuition mean is 4158.717
```

```
cat("The tuition sd is", sd(mobility_data$Tuition, na.rm = TRUE), "\n")
```

```
## The tuition sd is 3645.11
```

```
cat("The school spending mean is", mean(mobility_data$School_spending, na.rm = TRUE), "\n")
```

```
## The school spending mean is 5.857416
```

```
cat("The school spending sd is", sd(mobility_data$School_spending, na.rm = TRUE), "\n")
```

```
## The school spending sd is 1.078322
```

```
cat("The teacher student ratio mean is", mean(mobility_data$Student_teacher_ratio, na.rm = TRUE), "\n")
```

```
## The teacher student ratio mean is 16.62316
```

```r
cat("The teacher student sd is", sd(mobility_data$Student_teacher_ratio, na.rm = TRUE), "\n")
```

```
## The teacher student sd is 1.868416
```

```r
cat("The test score mean is", mean(mobility_data$Test_scores, na.rm = TRUE), "\n")
```

```
## The test score mean is -0.4693363
```

```r
cat("The test score sd is", sd(mobility_data$Test_scores, na.rm = TRUE), "\n")
```

```
## The test score sd is 7.43072
```

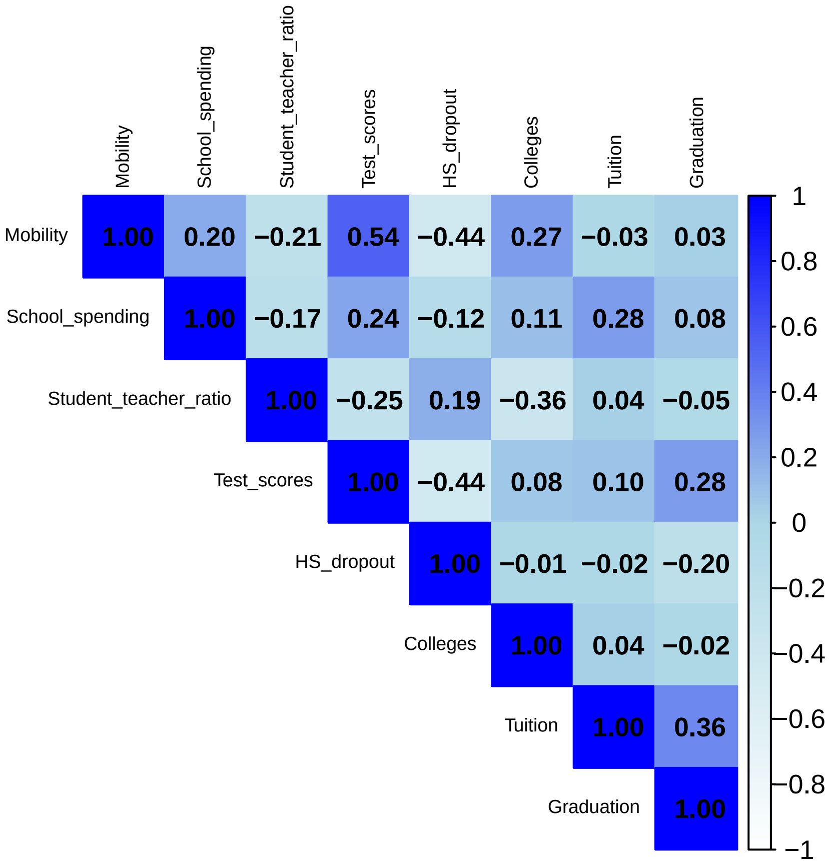
```r
cat("The hs dropout mean is", mean(mobility_data$HS_dropout, na.rm = TRUE), "\n")
```

```
## The hs dropout mean is 0.00198441
```

```r
cat("The hs dropout sd is", sd(mobility_data$HS_dropout, na.rm = TRUE), "\n")
```

```
## The hs dropout sd is 0.01940331
```

```r
cor_matrix <- cor(mobility_data, use = "pairwise.complete.obs")
corrplot(cor_matrix,
         method = "color",
         type = "upper",
         col = colorRampPalette(c("white", "lightblue", "blue"))(200),
         tl.cex = 0.6,
         number.cex = 0.8,
         addCoef.col = "black",
         tl.col = "black")
```



| | Mobility | School_spending | Student_teacher_ratio | Test_scores | HS_dropout | Colleges | Tuition | Graduation |
|---|---|---|---|---|---|---|---|---|
| Mobility | 1.00 | 0.20 | −0.21 | 0.54 | −0.44 | 0.27 | −0.03 | 0.03 |
| School_spending | | 1.00 | −0.17 | 0.24 | −0.12 | 0.11 | 0.28 | 0.08 |
| Student_teacher_ratio | | | 1.00 | −0.25 | 0.19 | −0.36 | 0.04 | −0.05 |
| Test_scores | | | | 1.00 | −0.44 | 0.08 | 0.10 | 0.28 |
| HS_dropout | | | | | 1.00 | −0.01 | −0.02 | −0.20 |
| Colleges | | | | | | 1.00 | 0.04 | −0.02 |
| Tuition | | | | | | | 1.00 | 0.36 |
| Graduation | | | | | | | | 1.00 |