Education and family structure

Mehdi Khan

December 13, 2017

load the libraries

```
suppressMessages(suppressWarnings(library(dplyr)))
suppressMessages(suppressWarnings(library(stringr)))

suppressMessages(suppressWarnings(library(psych)))
suppressMessages(suppressWarnings(library(ggplot2)))

suppressMessages(suppressWarnings(library(devtools)))
suppressMessages(suppressWarnings(library(stats)))
suppressMessages(suppressWarnings(library(tidyr)))
suppressMessages(suppressWarnings(library(ggpubr)))
```

Introduction

Data about Households, family structures and their characterstics in 50 US states published by US Census Bureau was examined in this project to see if the data provides any insight on the impact of family structures on education of people in a society.

Data collection

The data represents selected social chracterstics in the United States in the period of 5 years from 2011 to 2015 and compiled by American Community Survey,US Census Bureau. This publicly available data was downloaded in CSV format for this project.

Data Source

The data was published by US Census Bureau and posted in American Fact Finder website: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_DP02&src=pt

data prepration:

First the data was imported in R:

```
originDS <- read.csv("https://raw.githubusercontent.com/kmehdi2017/projectProp/master/ProjectProposal/A
    sep = ",", stringsAsFactors = FALSE)
```

The original data had a lot of variables, which are not relevant to this study, therefore a subset of the data was extracted. The following are the variables with their descriptions that were selected for the project:

GEO.display.label: Geography HC01_VC04: Estimate: Total households - Family households (families)
HC01_VC06: Estimate: Total households - Family households (families) - Married-couple family
HC01_VC08: Estimate: Total households - Family households (families) - Male householder, no wife present
HC01_VC10: Estimate: Total households - Family households (families) - Female householder, no husband

present HC01_VC76: Estimate: SCHOOL ENROLLMENT - Population 3 years and over enrolled in school HC01_VC91: Estimate: EDUCATIONAL ATTAINMENT - Population 25 years and over - Bachelor's degree

```
vars <- c("GEO.display.label", "HC01_VC04", "HC01_VC06", "HC01_VC08",
    "HC01_VC10", "HC01_VC76", "HC01_VC91")

familyEduDS <- originDS[-1, vars]

head(familyEduDS)
```

```
##   GEO.display.label HC01_VC04 HC01_VC06 HC01_VC08 HC01_VC10 HC01_VC76
## 2           Alabama   1238967    880942     78073    279952   1206014
## 3            Alaska    167562    124649     14733     28180    195151
## 4           Arizona   1581380   1142828    131803    306749   1754549
## 5          Arkansas    759924    558920     50484    150520    750024
## 6        California   8732734   6245351    759047   1728336  10579176
## 7          Colorado   1300972   1003324     91627    206021   1395787
##   HC01_VC91
## 2    478812
## 3     83201
## 4    753425
## 5    267741
## 6   5002596
## 7    847977
```

providing meaningful names to columns:

```
columnNames <- c("states", "total_family", "married_couple_family",
    "husband_only_family", "wife_only_family", "school_enrollment",
    "bachelor_degree")

colnames(familyEduDS) <- columnNames

head(familyEduDS)
```

```
##       states total_family married_couple_family husband_only_family
## 2    Alabama      1238967                880942               78073
## 3     Alaska       167562                124649               14733
## 4    Arizona      1581380               1142828              131803
## 5   Arkansas       759924                558920               50484
## 6 California      8732734               6245351              759047
## 7   Colorado      1300972               1003324               91627
##   wife_only_family school_enrollment bachelor_degree
## 2           279952           1206014          478812
## 3            28180            195151           83201
## 4           306749           1754549          753425
## 5           150520            750024          267741
## 6          1728336          10579176         5002596
## 7           206021           1395787          847977
```

Research question

Does the family structure of single parents or two parents families have any impact on the number of educated people in a society ?

case

Each case represents a state in the United States, there are 51 of them.

Type of study

This is an observational study

Response

The response variables are the estimates of school enrollment, and number of bachelor degree holders. Both of them are numerical.

Explanatory

The explanatory variables are the estimates of family types (i.e. single-parents family and two-parents family) and are numerical.

further data prepration

The data types of the fields were converted to numeric for calculation:

```
familyEduDS$total_family <- as.numeric(familyEduDS$total_family)
familyEduDS$married_couple_family <- as.numeric(familyEduDS$married_couple_family)
familyEduDS$husband_only_family <- as.numeric(familyEduDS$husband_only_family)
familyEduDS$wife_only_family <- as.numeric(familyEduDS$wife_only_family)
familyEduDS$school_enrollment <- as.numeric(familyEduDS$school_enrollment)
familyEduDS$bachelor_degree <- as.numeric(familyEduDS$bachelor_degree)
```

Five derived fields were created that describe: 1. the number of single parent families in each state 2. average school enrollment per family in each state 3. average bachelor degree holders per family in each state 4. ratio of two-parents families in each state 5. ratio of single-parents families in each state

```
familyEduDS <- mutate(familyEduDS, single_parent_family = husband_only_family +
    wife_only_family)
familyEduDS <- mutate(familyEduDS, avg_enrollment = round(school_enrollment/total_family,
    2))
familyEduDS <- mutate(familyEduDS, avg_bachelor = round(bachelor_degree/total_family,
    2))
familyEduDS <- mutate(familyEduDS, ratio_both_parents = round(married_couple_family/total_family,
    2))
familyEduDS <- mutate(familyEduDS, ratio_single_parents = round(single_parent_family/total_family,
    2))
```

Analysis approach:

correlation analysis was used to find if there is any correlation between the type of family structures and the number of school enrollment and the number of bachelor degree holders.

Hypothesis:

Null Hypothesis, Ho: family structures does not affect education i.e.There is no correlation between family structures and education, correlation coefficients = 0

Alternative Hypothesis, Ha: family structures does affect education There is correlation between family structures and education, correlation coefficients != 0

descriptive Analysis:

```
describe(familyEduDS$married_couple_family)
```

```
##    vars  n    mean      sd median  trimmed      mad   min     max   range
## X1    1 51 1107424 1181185 752359 880225.4 708491.5 65383 6245351 6179968
##    skew kurtosis       se
## X1 2.28     6.15 165398.9
```

```
describe(familyEduDS$single_parent_family)
```

```
##    vars  n     mean       sd median  trimmed      mad   min     max
## X1    1 51 407488.5 471125.8 294017 311463.8 283268.5 29874 2487383
##      range skew kurtosis      se
## X1 2457509 2.42     6.67 65970.8
```

```
describe(familyEduDS$avg_enrollment)
```

```
##    vars  n mean  sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 51 1.05 0.1   1.03    1.04 0.06 0.86 1.39  0.53 1.33     3.31 0.01
```

```
describe(familyEduDS$avg_bachelor)
```

```
##    vars  n mean  sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 51  0.5 0.1   0.49     0.5 0.07 0.32 0.89  0.57 1.02     3.13 0.01
```

```
describe(familyEduDS$ratio_both_parents)
```

```
##    vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis
## X1    1 51 0.74 0.05   0.74    0.74 0.04 0.55 0.82  0.27 -1.38     4.23
##      se
## X1 0.01
```

```
describe(familyEduDS$ratio_single_parents)
```

```
##    vars  n mean   sd median trimmed  mad  min  max range skew kurtosis
## X1    1 51 0.26 0.05   0.26    0.26 0.04 0.18 0.45  0.27 1.38     4.23
##      se
## X1 0.01
```

```
summary(familyEduDS$avg_enrollment)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.860   0.995   1.030   1.046   1.080   1.390
```

```
summary(familyEduDS$avg_bachelor)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3200  0.4400  0.4900  0.5022  0.5650  0.8900
```

```r
summary(familyEduDS$ratio_both_parents)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5500  0.7200  0.7400  0.7408  0.7750  0.8200
```

```r
summary(familyEduDS$ratio_single_parents)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1800  0.2250  0.2600  0.2592  0.2800  0.4500
```

```r
IQR_enrollment <- 1.08 - 0.995
IQR_bachelor <- 0.565 - 0.44

IQR_single_parents <- 0.28 - 0.225
IQR_two_parents <- 0.775 - 0.72

IQR_enrollment
```

```
## [1] 0.085
```

```r
IQR_bachelor
```

```
## [1] 0.125
```

```r
IQR_single_parents
```

```
## [1] 0.055
```

```r
IQR_two_parents
```

```
## [1] 0.055
```

```r
ggplot(familyEduDS, aes(x = avg_enrollment, fill = "red", col = "blue",
    alpha = 0.2)) + geom_histogram(position = "identity", bins = 20,
    show.legend = FALSE, binwidth = 0.05) + theme(plot.title = element_text(size = 12,
    color = "blue", hjust = 0.5)) + ggtitle("distribution of average school enrollment per family ") +
    xlab("average school enrollment per family")
```

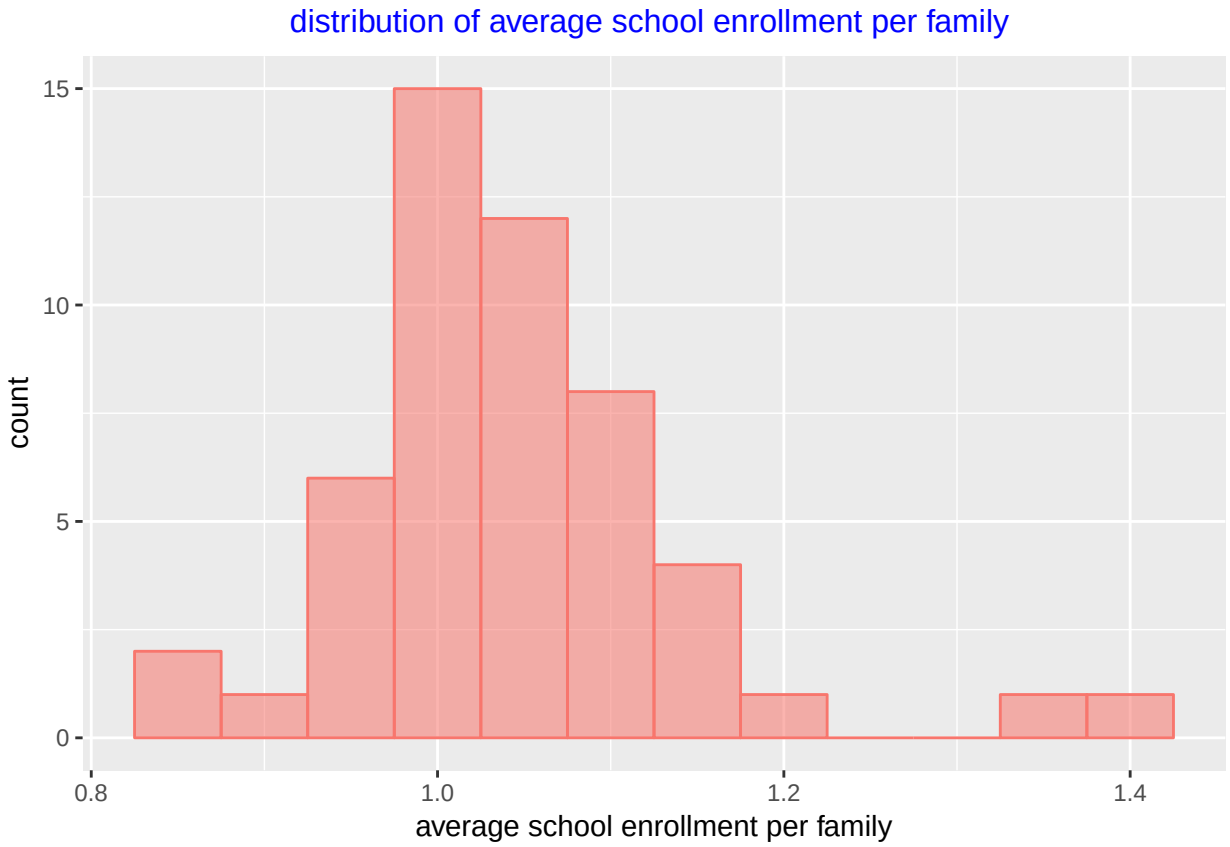

Figure 1.

```
ggplot(familyEduDS, aes(x = avg_bachelor, fill = "blue", col = "red",
    alpha = 0.2)) + geom_histogram(position = "identity", bins = 20,
    show.legend = FALSE) + theme(plot.title = element_text(size = 12,
    color = "blue", hjust = 0.5)) + ggtitle("distribution of average bachelor degree holder per family
    xlab("average bachelor degree holder per family")
```

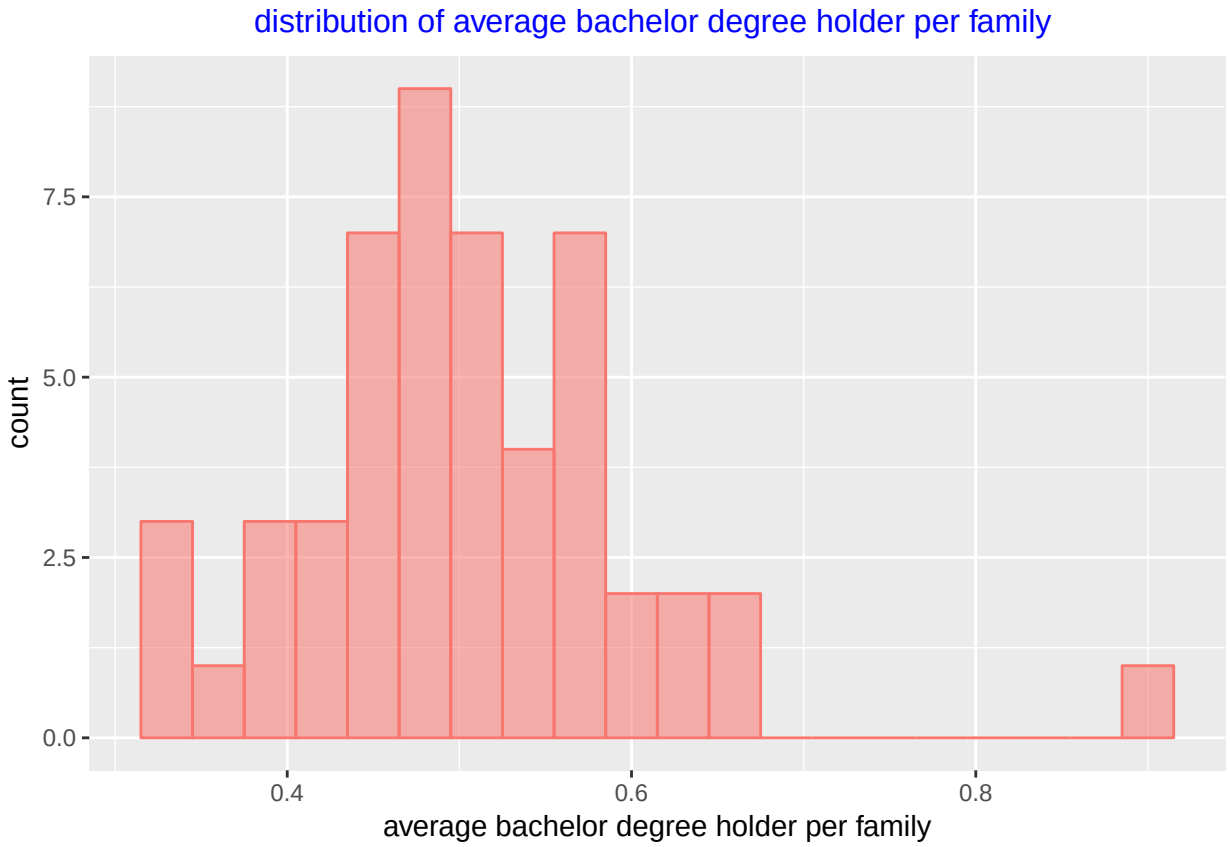

Figure 2.

```
ggplot(familyEduDS, aes(x = ratio_both_parents, fill = "blue", col = "red",
    alpha = 0.2)) + geom_histogram(position = "identity", bins = 20,
    show.legend = FALSE, binwidth = 0.01) + theme(plot.title = element_text(size = 12,
    color = "blue", hjust = 0.5)) + ggtitle("distribution of the ratios of families of both parents") +
    xlab("ratios of families of both parents")
```

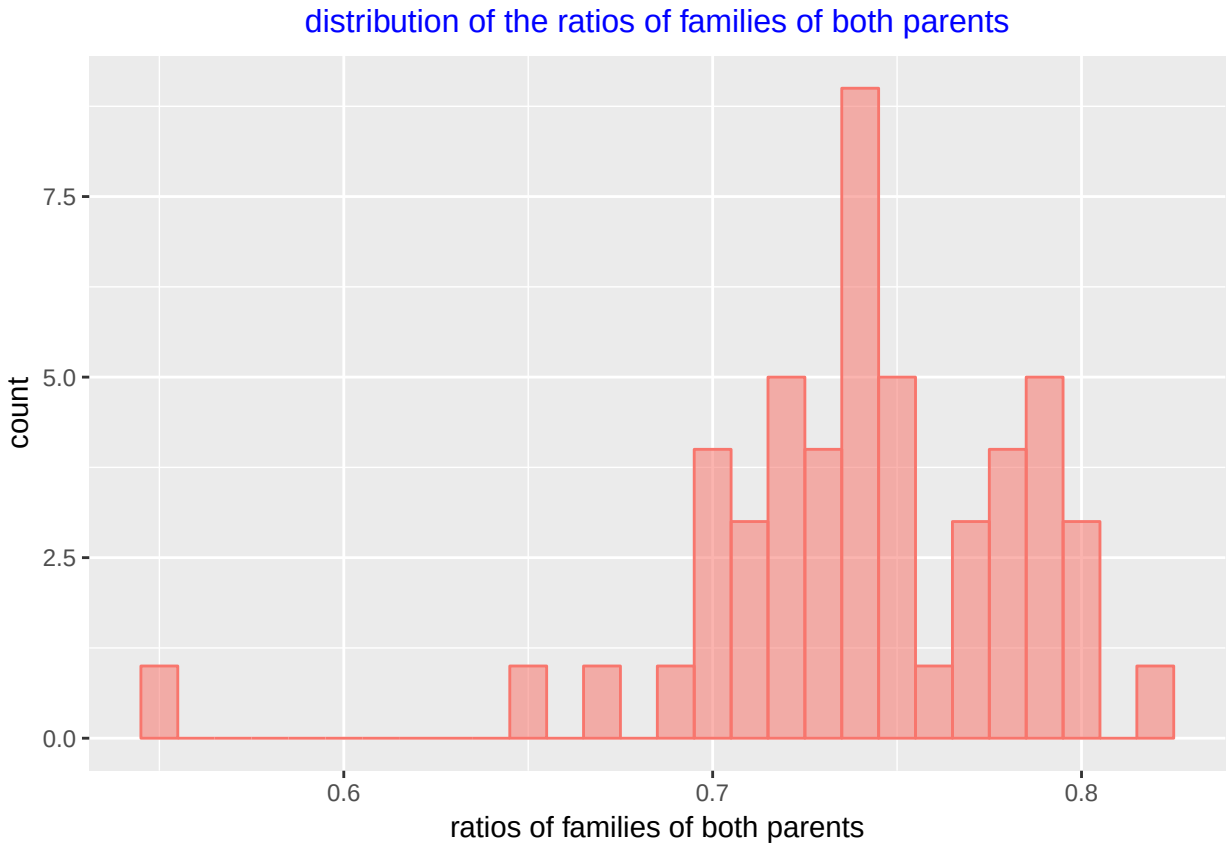

Figure 3.

```
ggplot(familyEduDS, aes(x = ratio_single_parents, alpha = 0.2)) +
    geom_histogram(position = "identity", bins = 20, show.legend = FALSE,
        binwidth = 0.01, col = "blue", fill = "blue") + theme(plot.title = element_text(size = 12,
    color = "blue", hjust = 0.5)) + ggtitle("distribution of the ratios of families of single parents")
    xlab("ratios of families of single parents")
```

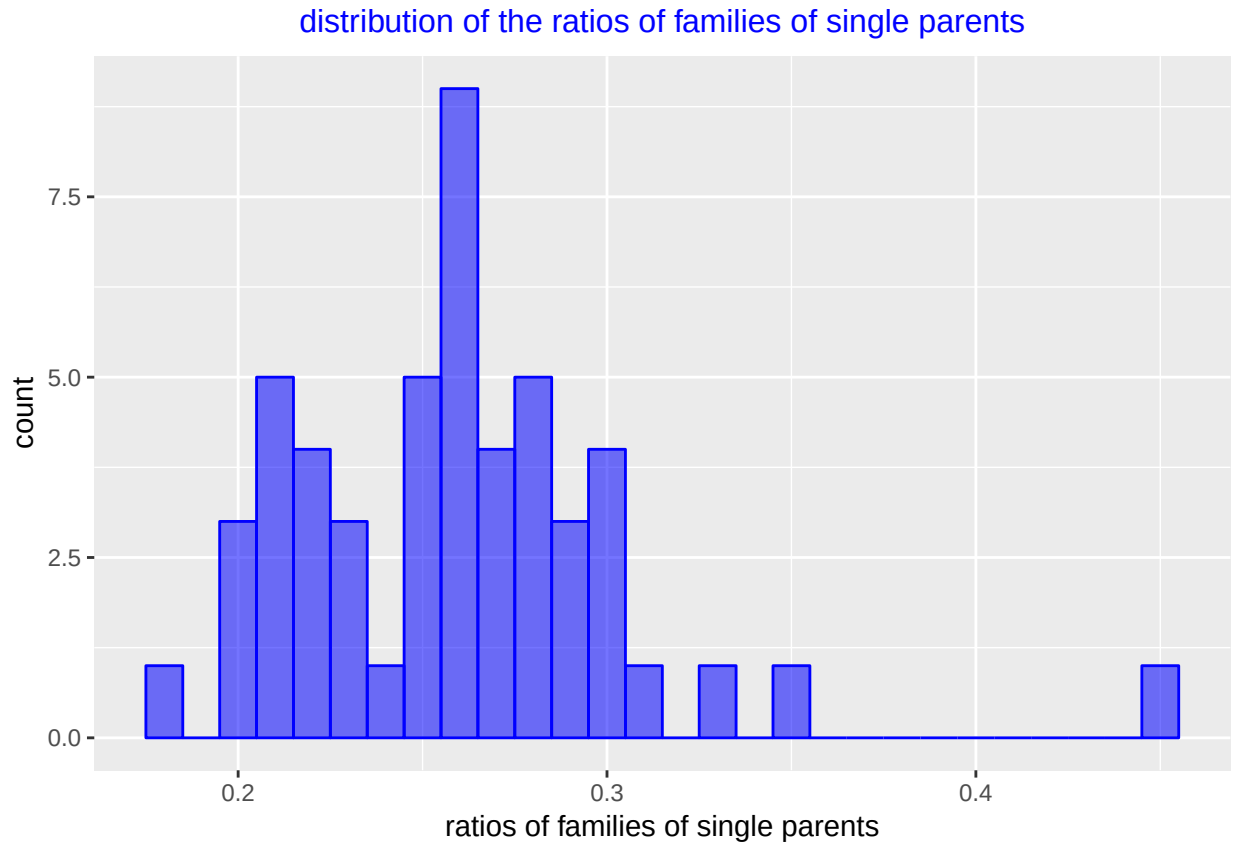

Figure 4.

```
ggplot(familyEduDS, aes(x = ratio_both_parents, y = avg_enrollment)) +
    geom_point(color = "red") + ggtitle("two-parents families vs school enrollment") +
    xlab("ratio of families with both parents") + ylab("average school enrollment per family") +
    geom_smooth(method = "auto", col = "red")
```

```
## `geom_smooth()` using method = 'loess'
```

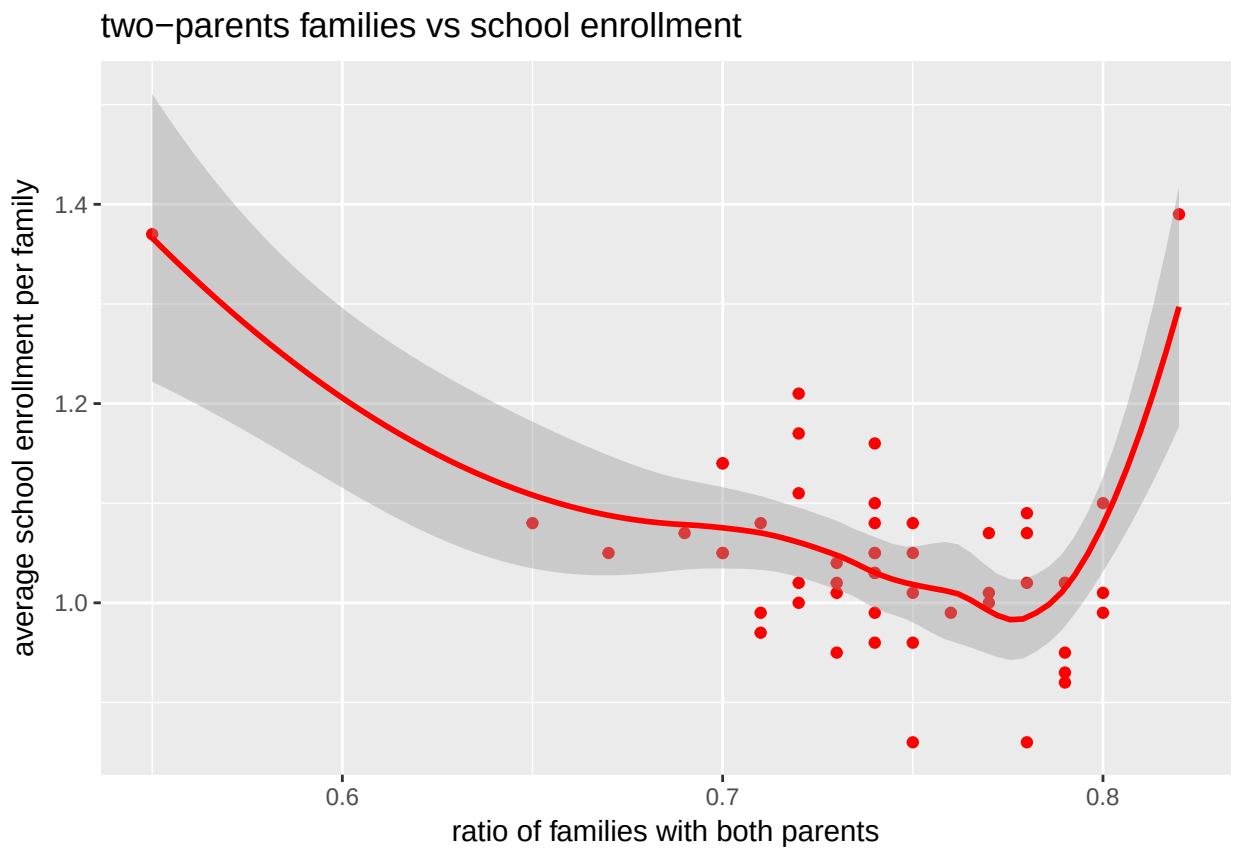

Figure 5.

The figure 5 shows a mostly linearity between school enrollment and two-parents families but two outliers on both ends heavily impact the relationship.

```
ggplot(familyEduDS, aes(x = ratio_both_parents, y = avg_bachelor)) +
    geom_point(color = "red") + ggtitle("two-parents families vs bachelor degree holders") +
    xlab("ratio of families with both parents") + ylab("average bachelor degree holders per family") +
    geom_smooth(method = "auto", col = "red")
```

```
## `geom_smooth()` using method = 'loess'
```

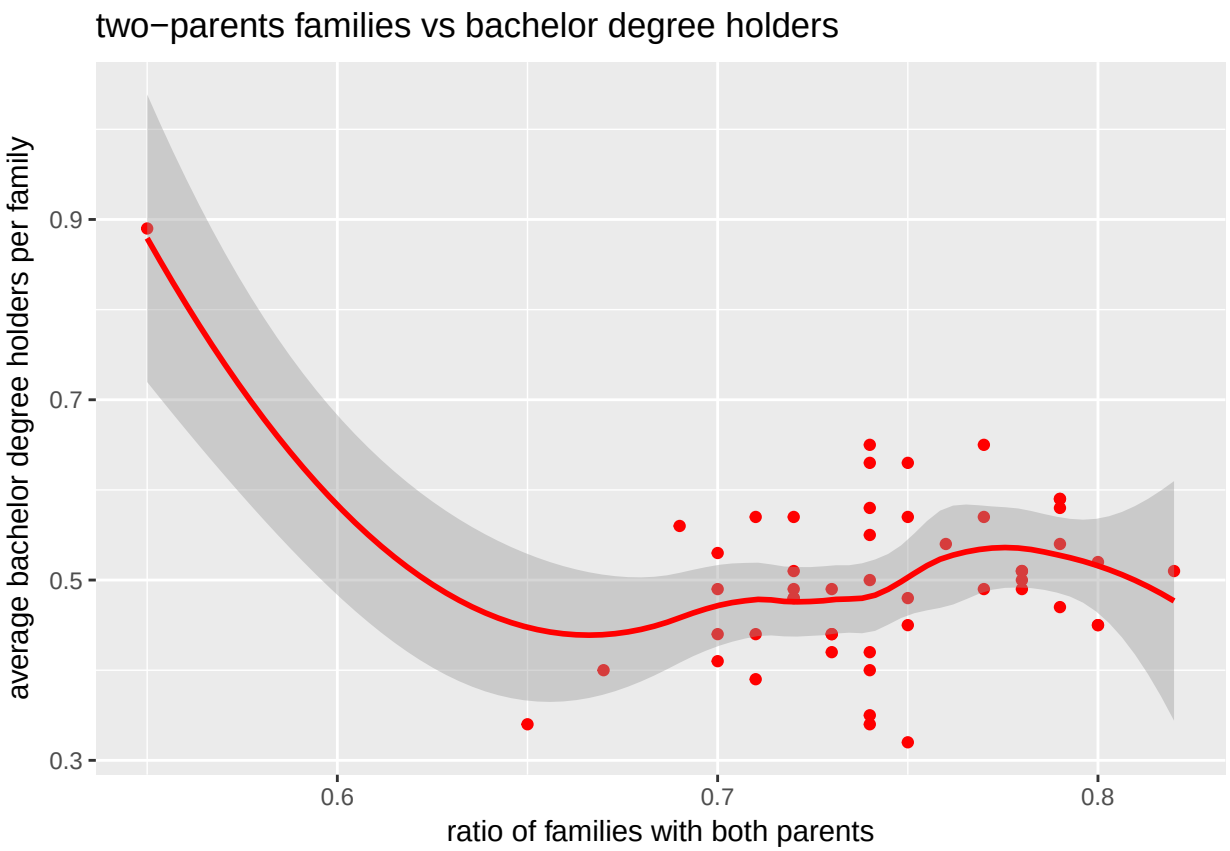

Figure 6.

The figure 6 also shows a mostly linearity between bachelor degree holders and two-parents families but one extreme outliers on one end heavily impact the relationship.

```
ggplot(familyEduDS, aes(x = ratio_single_parents, y = avg_enrollment)) +
    geom_point(color = "blue") + ggtitle("single parents families vs school enrollment") +
    xlab("ratio of families of single parents") + ylab("average school enrollment per family") +
    geom_smooth(method = "auto")
```

```
## `geom_smooth()` using method = 'loess'
```

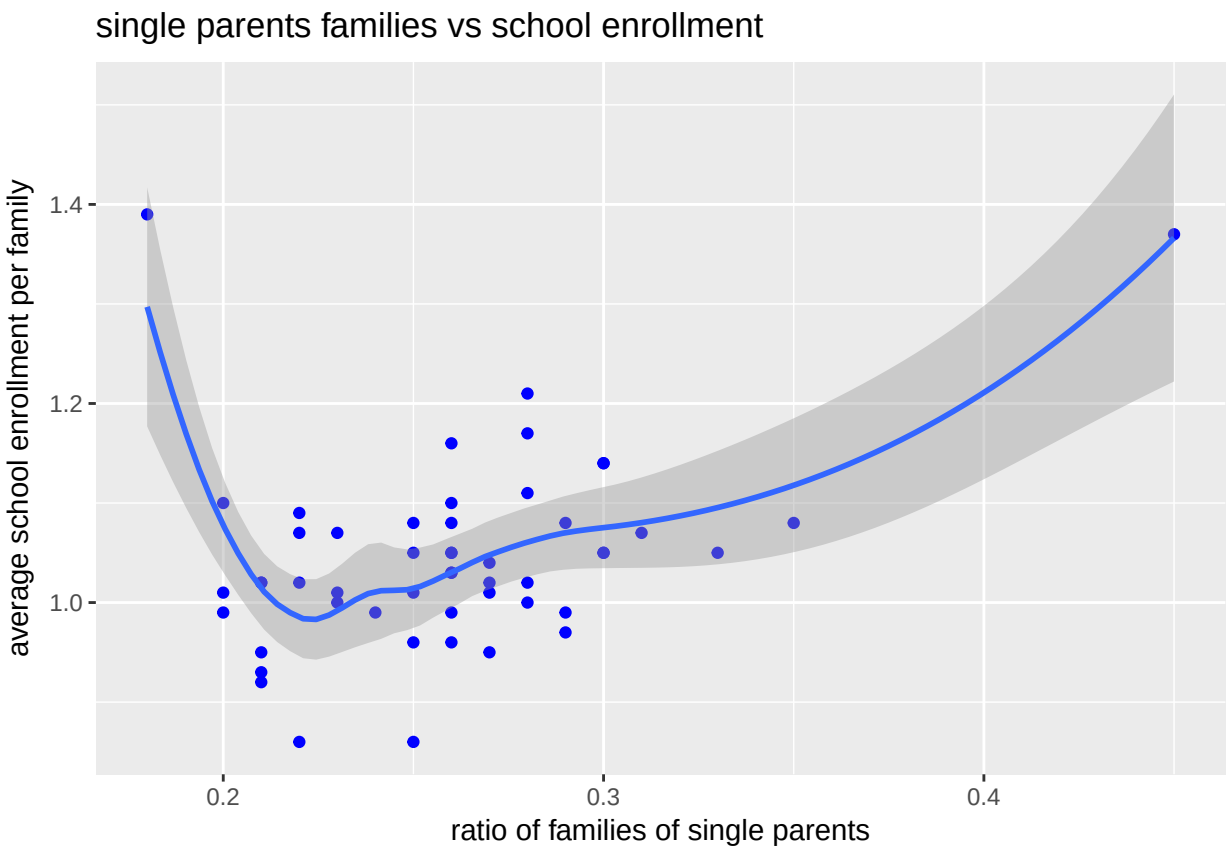

Figure 7.

The figure 7 also shows a mostly linearity between school enrollment and single parents families but two outliers on both ends heavily impact the relationship.

```
ggplot(familyEduDS, aes(x = ratio_single_parents, y = avg_bachelor)) +
    geom_point(color = "blue") + ggtitle("single parents families vs bachelor degree holders") +
    xlab("ratio of families of single parents") + ylab("average bachelor degree holders per family") +
    geom_smooth(method = "auto")
```

```
## `geom_smooth()` using method = 'loess'
```

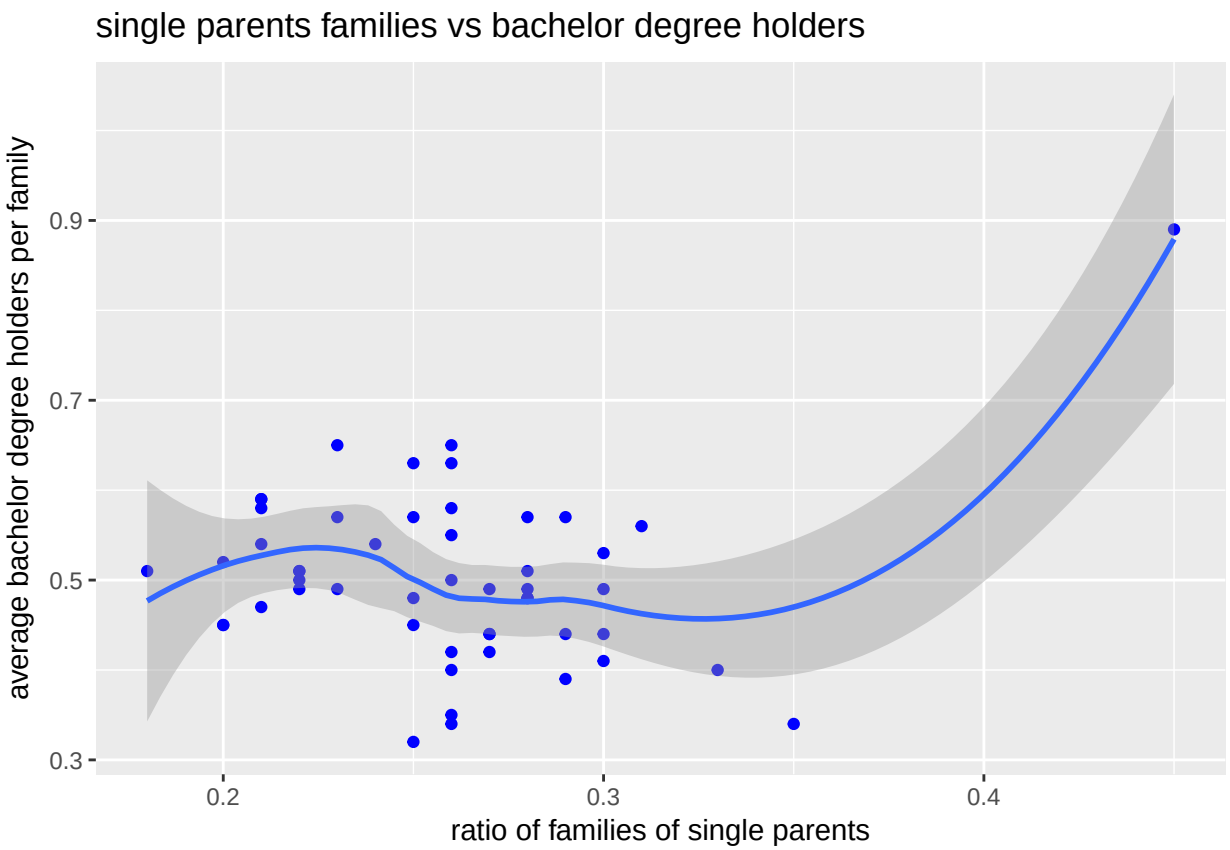

Figure 8.

The figure 8 shows a mostly linearity between bachelor degree holders and single parents family but one extreme outliers on one end heavily impact the relationship.

correlation analysis

Conditions check: All the histograms (from figure 1,2,3 and 4) shows the distributions of all the variables of interst are near normal with some skews which are the effect of outliers. All the variables are numerical.All the scatterplots (figure 5,6,7,8) show that the linearity condition is met for all the variables. Since the data represnts the whole population (50 states) the independence condition is not relevant.

So conditions are met except the presence of outliers.

cocorrelation tests without removing outliers:

Average school enrollment and two-parents families:

```
cor.test(familyEduDS$ratio_both_parents, familyEduDS$avg_enrollment,
    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  familyEduDS$ratio_both_parents and familyEduDS$avg_enrollment
## t = -2.7312, df = 49, p-value = 0.008747
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.58088046 -0.09768515
## sample estimates:
##        cor
## -0.3634837
```

```
ggscatter(familyEduDS, x = "ratio_both_parents", y = "avg_enrollment",
    add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson",
    xlab = "ratio of two-parents families", ylab = "avg. enrollment in schools per family",
    color = "red")
```

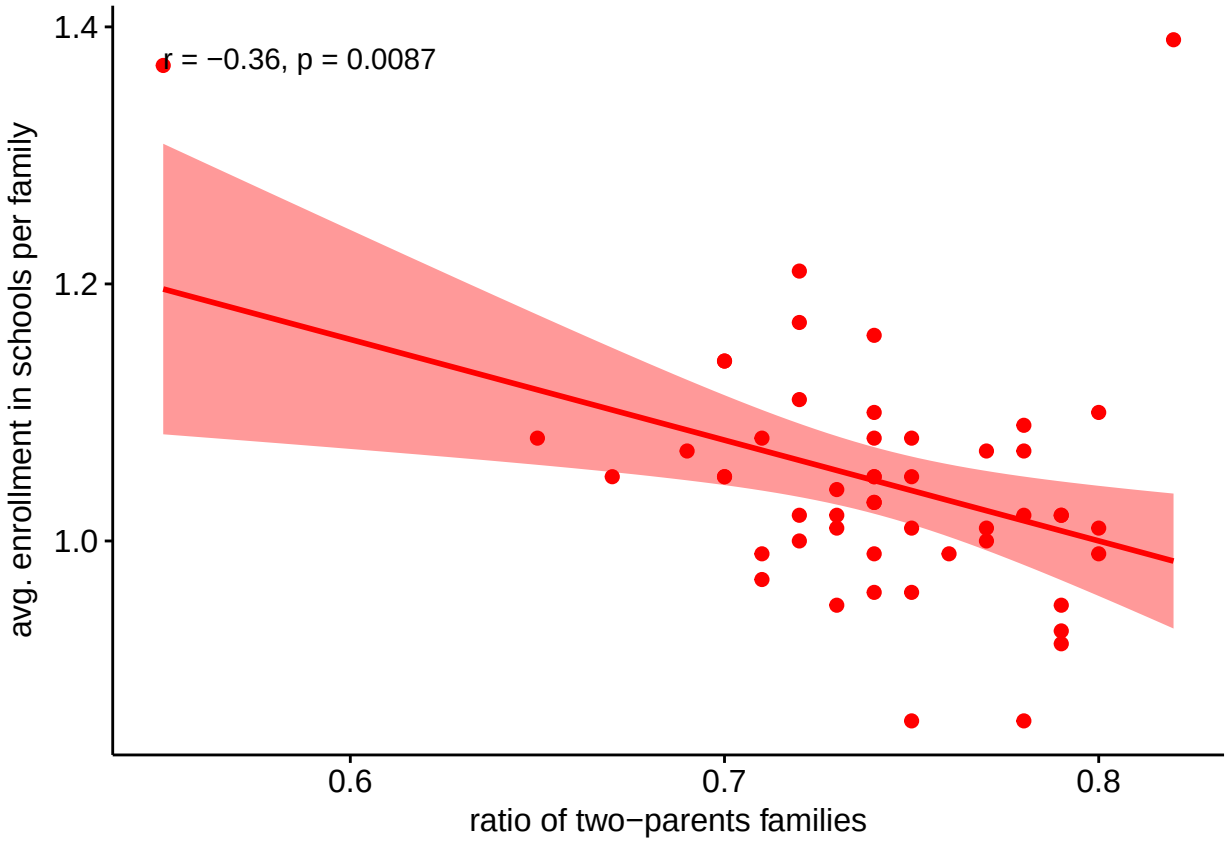

Figure 9.

The test and the figure 9 shows a negative correlation between school enrollment and both-parents families with correlation coefficients (r) of -0.36

Average school enrollment and single parents family:

```
cor.test(familyEduDS$ratio_single_parents, familyEduDS$avg_enrollment,
    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  familyEduDS$ratio_single_parents and familyEduDS$avg_enrollment
## t = 2.7312, df = 49, p-value = 0.008747
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09768515 0.58088046
## sample estimates:
##       cor
## 0.3634837
```

```
ggscatter(familyEduDS, x = "ratio_single_parents", y = "avg_enrollment",
    add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson",
    xlab = "ratio of single-parents families", ylab = "avg. enrollment in schools per family",
    color = "blue")
```

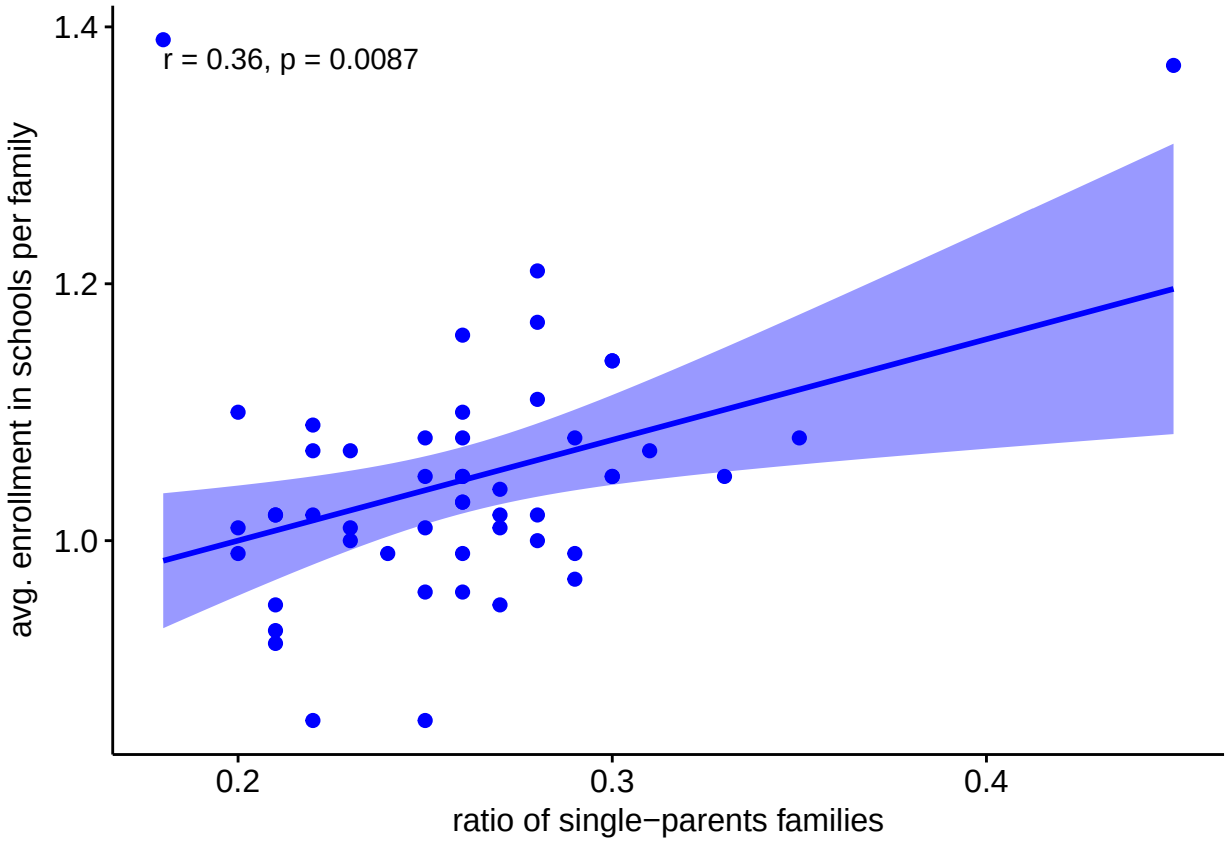

Figure 10.

The test and the figure 10 shows a positive correlation between school enrollment and single-parents families with correlation coefficients (r) of 0.36

Average bachelor degree holders and two-parents families

```
cor.test(familyEduDS$ratio_both_parents, familyEduDS$avg_bachelor,
    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  familyEduDS$ratio_both_parents and familyEduDS$avg_bachelor
## t = -0.94704, df = 49, p-value = 0.3483
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3950579  0.1469422
## sample estimates:
##        cor
## -0.1340707
```

```
ggscatter(familyEduDS, x = "ratio_both_parents", y = "avg_bachelor",
    add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson",
    xlab = "ratio of two-parents families", ylab = "avg. bachelor degree holders per family",
    color = "red")
```

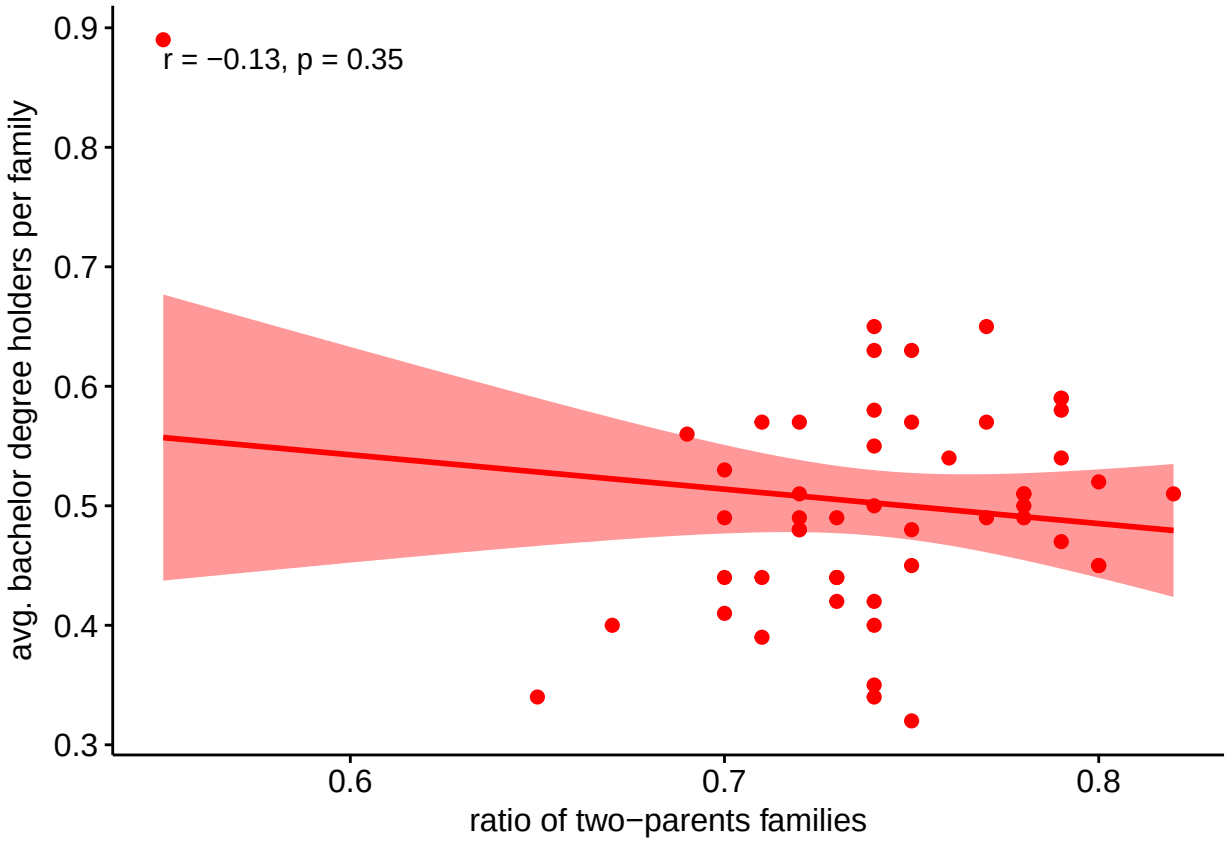

Figure 11.

The test and the figure 11 shows a negative correlation between bachelor degree holders and two-parents families with a correlation coefficients (r) of -0.13

Average bachelor degree holders and single parents families

```
cor.test(familyEduDS$ratio_single_parents, familyEduDS$avg_bachelor,
    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  familyEduDS$ratio_single_parents and familyEduDS$avg_bachelor
## t = 0.94704, df = 49, p-value = 0.3483
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1469422  0.3950579
## sample estimates:
##       cor
## 0.1340707
```

```
ggscatter(familyEduDS, x = "ratio_single_parents", y = "avg_bachelor",
    add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson",
    xlab = "ratio of single-parents families", ylab = "average bachelor degree holders per family",
    color = "blue")
```

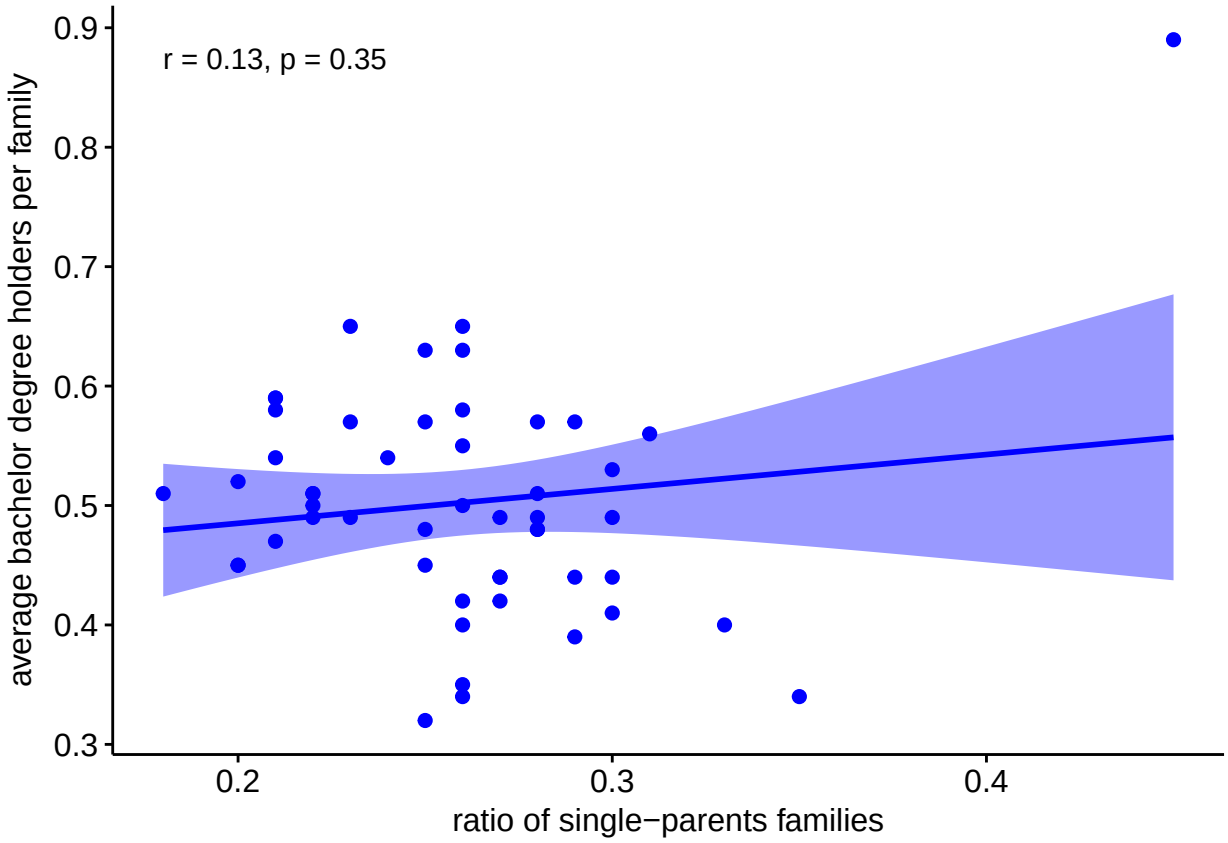

Figure 12.

The test and the figure 12 shows a positive correlation between average bachelor degree holders per family and single-parents families with correlation coefficients (r) of 0.13

So all the above tests show that there are a positive correaltions between single parent families with both school enrollment and average bachelor degree holders i.e. a community with single parent families would have more educated population while two parents families have negative correlations with both measures of education (i.e. school enrollment and number of bachelor degree holders)

cocorrelation tests without outliers:

Since not having outliers is a condition for correlation analysis, all the outliers were removed and the similar tests were done again on the revised data.

Finding outliers:

```
familyEduDS[(familyEduDS$avg_enrollment < 0.995 - 1.5 * IQR_enrollment) |
    (familyEduDS$avg_enrollment > 1.08 + 1.5 * IQR_enrollment), ]
```

```
##                 states total_family married_couple_family
## 5           California      8732734               6245351
## 9  District of Columbia       118737                 65383
## 20               Maine       347579                270147
## 45                Utah       680007                554555
## 49       West Virginia       479803                361652
##    husband_only_family wife_only_family school_enrollment bachelor_degree
## 5               759047          1728336          10579176         5002596
## 9                10502            42852            162835          105880
## 20               24446            52986            299595          178375
## 45               38394            87058            942989          347460
## 49               33962            84189            410745          152377
##    single_parent_family avg_enrollment avg_bachelor ratio_both_parents
## 5               2487383           1.21         0.57               0.72
## 9                 53354           1.37         0.89               0.55
## 20                77432           0.86         0.51               0.78
## 45               125452           1.39         0.51               0.82
## 49               118151           0.86         0.32               0.75
##    ratio_single_parents
## 5                  0.28
## 9                  0.45
## 20                 0.22
## 45                 0.18
## 49                 0.25
```

```
familyEduDS[(familyEduDS$avg_bachelor < 0.44 - 1.5 * IQR_bachelor) |
    (familyEduDS$avg_bachelor > 0.565 + 1.5 * IQR_bachelor), ]
```

```
##                 states total_family married_couple_family
## 9 District of Columbia       118737                 65383
##   husband_only_family wife_only_family school_enrollment bachelor_degree
## 9               10502            42852            162835          105880
##   single_parent_family avg_enrollment avg_bachelor ratio_both_parents
## 9                53354           1.37         0.89               0.55
##   ratio_single_parents
## 9                 0.45
```

```
familyEduDS[(familyEduDS$ratio_both_parents < 0.72 - 1.5 * IQR_two_parents) |
    (familyEduDS$ratio_both_parents > 0.775 + 1.5 * IQR_two_parents),
    ]
```

```
##                 states total_family married_couple_family
## 9 District of Columbia       118737                 65383
##   husband_only_family wife_only_family school_enrollment bachelor_degree
## 9               10502            42852            162835          105880
##   single_parent_family avg_enrollment avg_bachelor ratio_both_parents
## 9                53354           1.37         0.89               0.55
##   ratio_single_parents
## 9                 0.45
```

```
familyEduDS[(familyEduDS$ratio_single_parents < 0.225 - 1.5 * IQR_single_parents) |
    (familyEduDS$ratio_single_parents > 0.28 + 1.5 * IQR_single_parents),
    ]
```

```
##                 states total_family married_couple_family
## 9 District of Columbia       118737                 65383
##   husband_only_family wife_only_family school_enrollment bachelor_degree
## 9               10502            42852            162835          105880
##   single_parent_family avg_enrollment avg_bachelor ratio_both_parents
## 9                53354           1.37         0.89               0.55
##   ratio_single_parents
## 9                 0.45
```

District of Columbia is the outlier in all the variables while California, Utah, Maine and West Virginia are outliers only in scholl enrollment variables.

Outliers were removed and two seperate datasets were created:

```
EduDS <- familyEduDS[-c(9), ]
EduDS_enroll <- familyEduDS[-c(5, 9, 20, 45, 49), ]
```

correlation test:

Average school enrollment and two-parents family

```
cor.test(EduDS_enroll$ratio_both_parents, EduDS_enroll$avg_enrollment,
    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  EduDS_enroll$ratio_both_parents and EduDS_enroll$avg_enrollment
## t = -2.4519, df = 44, p-value = 0.01825
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5787578 -0.0627270
## sample estimates:
##       cor
## -0.3467116
```

```
ggscatter(EduDS_enroll, x = "ratio_both_parents", y = "avg_enrollment",
    add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson",
    xlab = "ratio of two-parents families", ylab = "avg. enrollment in schools per family",
    color = "red")
```

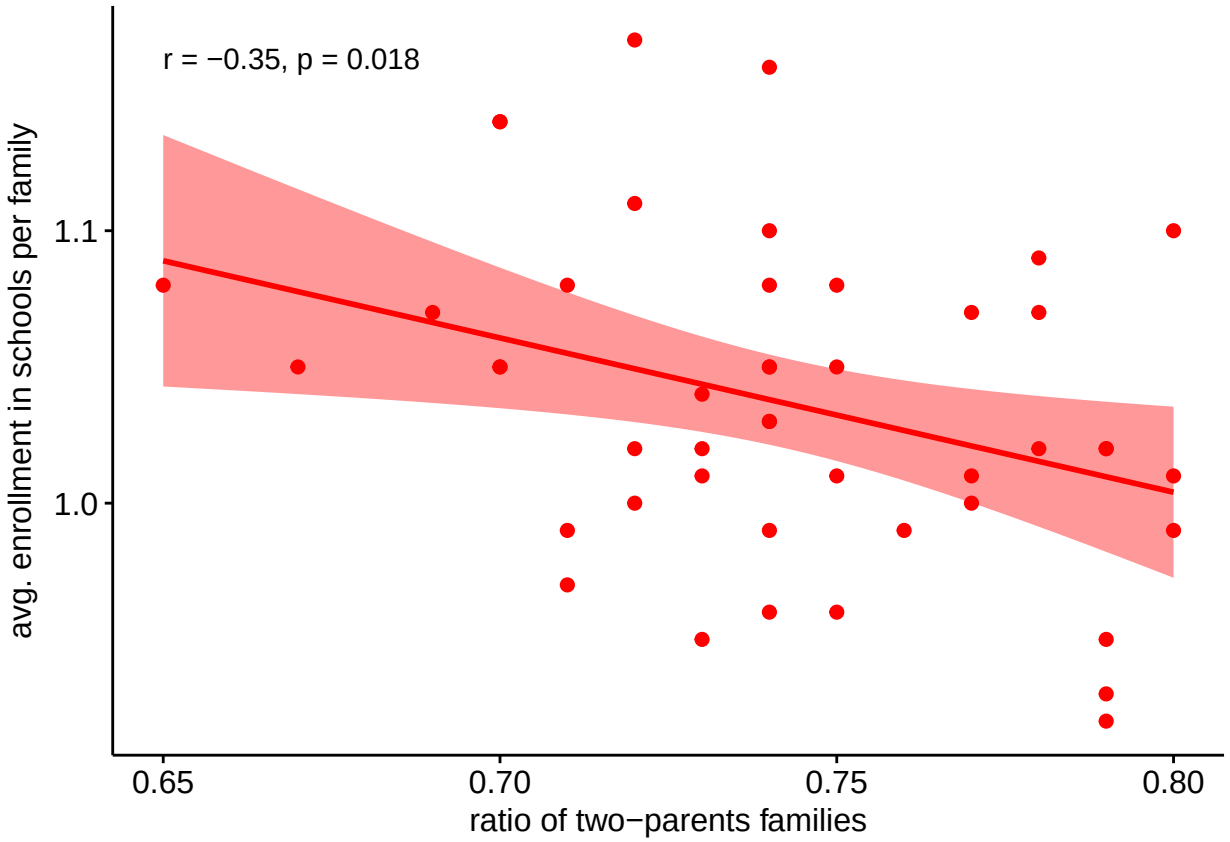

Figure 13.

The test and the figure 13 shows a negative correlation between school enrollment and two-parents families with correlation coefficients (r) of -0.35

Average school enrollment and single parents family

```
cor.test(EduDS_enroll$ratio_single_parents, EduDS_enroll$avg_enrollment,
    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  EduDS_enroll$ratio_single_parents and EduDS_enroll$avg_enrollment
## t = 2.4519, df = 44, p-value = 0.01825
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0627270 0.5787578
## sample estimates:
##       cor
## 0.3467116
```

```
ggscatter(EduDS_enroll, x = "ratio_single_parents", y = "avg_enrollment",
    add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson",
    xlab = "single-parents families", ylab = "avg. enrollment in schools per family",
    color = "blue")
```

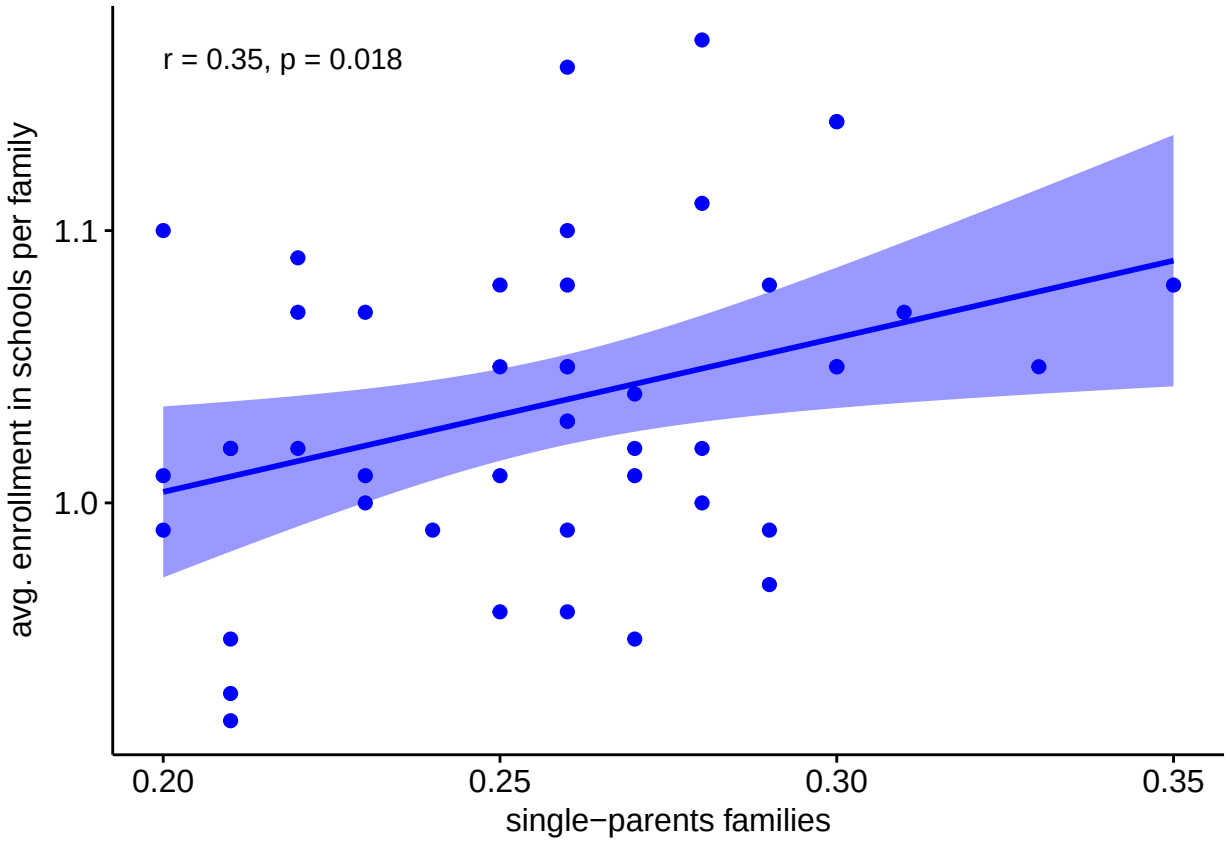

Figure 14.

The test and the figure 14 shows a positive correlation between school enrollment and single-parents families with correlation coefficients (r) of 0.35

Average bachelor degree holders and two-parents family

```
cor.test(EduDS$ratio_both_parents, EduDS$avg_bachelor, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  EduDS$ratio_both_parents and EduDS$avg_bachelor
## t = 2.3108, df = 48, p-value = 0.02518
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04173024 0.54661051
## sample estimates:
##       cor
## 0.3164028
```

```
ggscatter(EduDS, x = "ratio_both_parents", y = "avg_bachelor", add = "reg.line",
    conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "both-parents families",
    ylab = "avg. bachelor degree holders per family", color = "red")
```

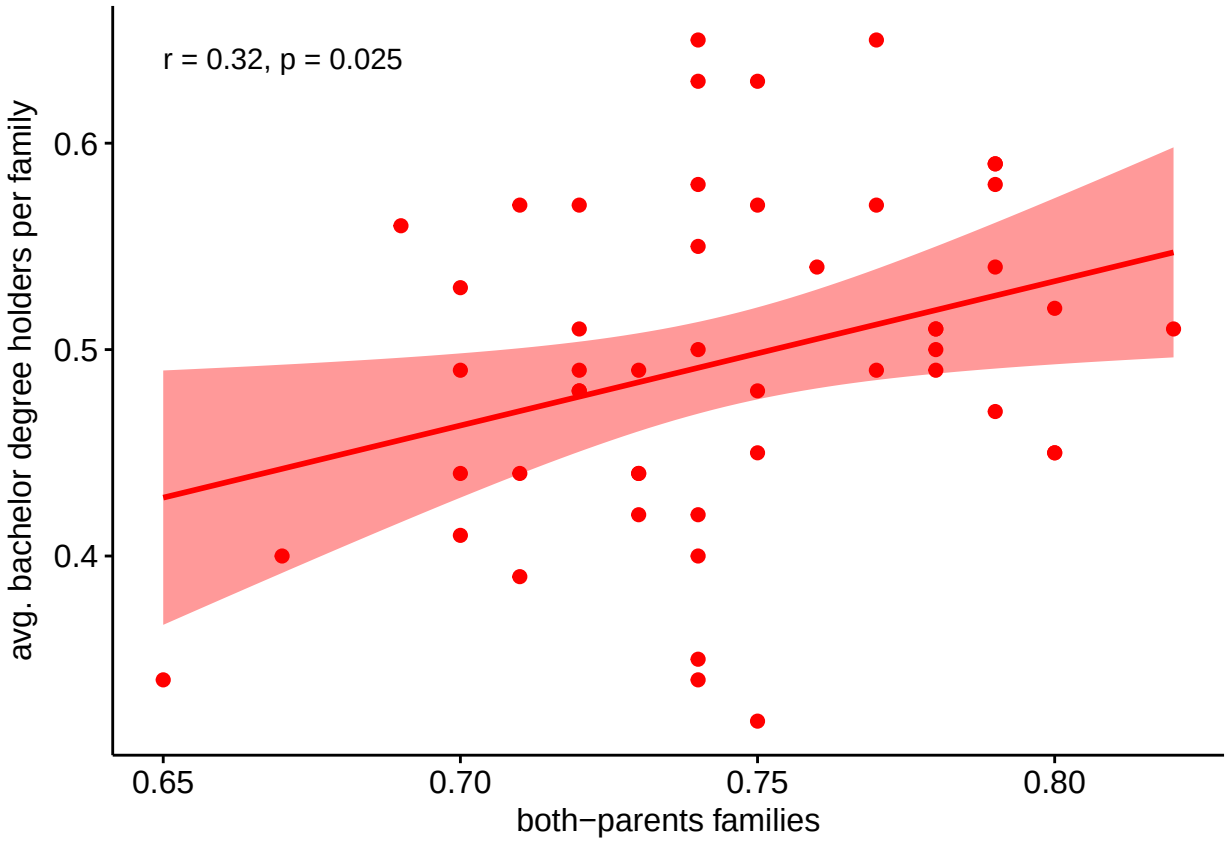

Figure 15.

The test and the figure 15 shows a positive correlation between bachelor degree holders and two-parents families with correlation coefficients (r) of 0.32

Average bachelor degree holders and single parents family

```
cor.test(EduDS$ratio_single_parents, EduDS$avg_bachelor, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  EduDS$ratio_single_parents and EduDS$avg_bachelor
## t = -2.3108, df = 48, p-value = 0.02518
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.54661051 -0.04173024
## sample estimates:
##        cor
## -0.3164028
```

```
ggscatter(EduDS, x = "ratio_single_parents", y = "avg_bachelor", add = "reg.line",
    conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "single-parents families",
    ylab = "avg. bachelor degree holders per family", color = "blue")
```

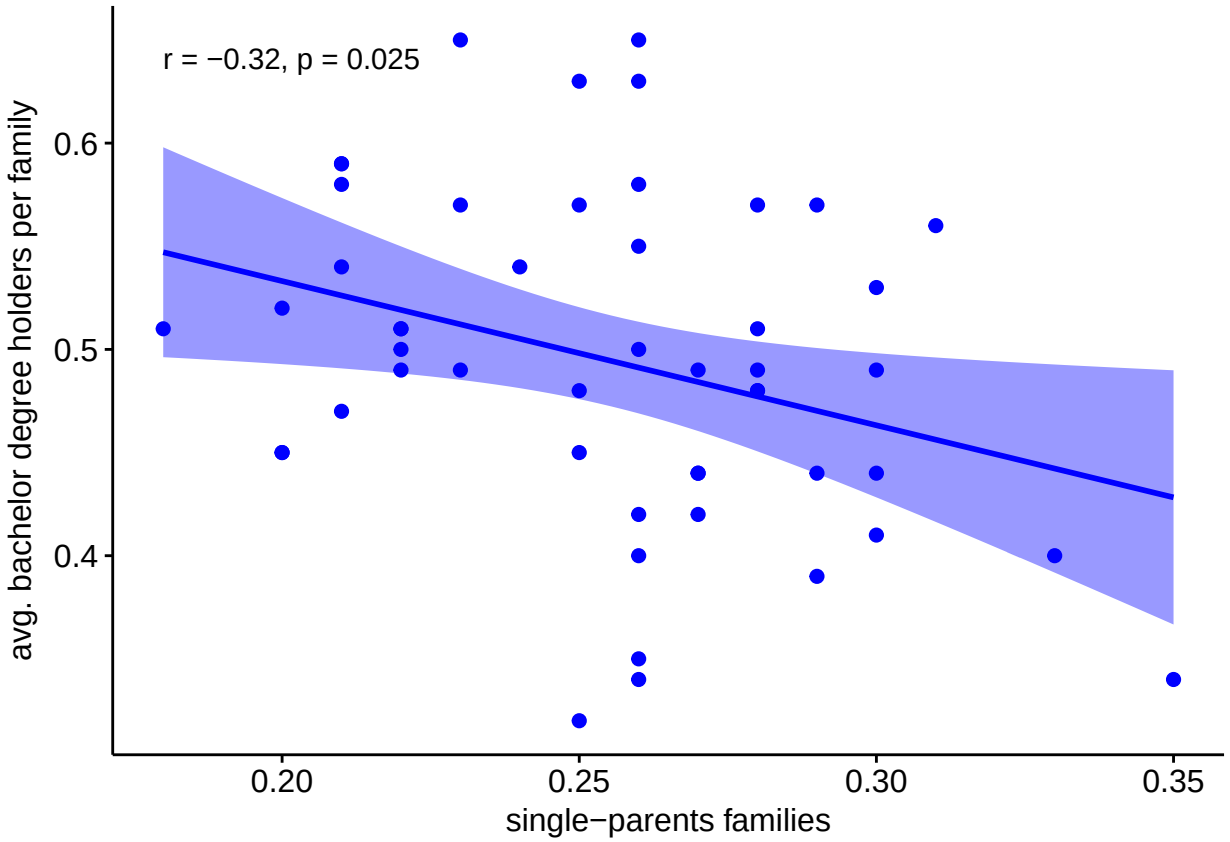

Figure 16.

The test and the figure 16 shows a negative correlation between average bachelor degree holders and single-parents families with correlation coefficients (r) of -0.32

CONCLUSION

Since the data represents the whole population (50 States) the statistical significance is meaningless here. There is no standard error and the p-value is irrelevant. Therefore the correlation coefficients found here represent the poulation correlation coefficients. So after removing the ouliers the result show that:

two-parents families have a positive correlation with population with graduate (bachelor) degree holders but have a negative coorelation with school enrollment. Single-parents familes have the same correlation but in the opposite directions.

Only 12.25% (0.35ˆ2) variablity in average scool enrollment and only 10.24% (0.32ˆ2) variablity in average bachelor degree holders can be explained by the family structure variables.

So there is an impact of family structures on the education of people and we reject the Null hypothesis.

Further analysis:

If the purpose of the analysis is to predict the affect of family structures on education in future cases , then the dataset may be considered as the smaple dataset from a population of an infinite cases of the future. Asuming the above, a regression analysis was done between average bachelor degree holders per family and the ratio of two-parents families

```
fit_bachelor <- lm(avg_bachelor ~ ratio_both_parents, data = EduDS)
```

Histogram of residuals

```
hist(fit_bachelor$residuals, col = "green")
```

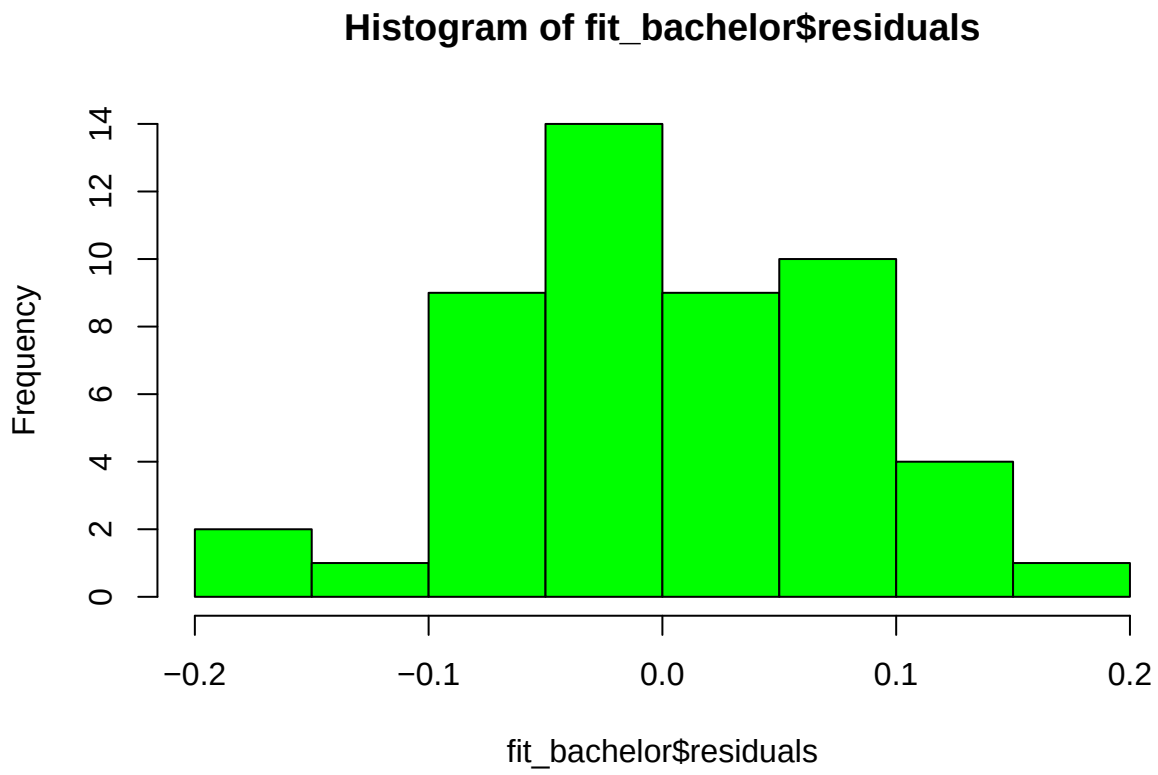

Figure 17.

```
qqnorm(fit_bachelor$residuals)
qqline(fit_bachelor$residuals, col = "blue")
```

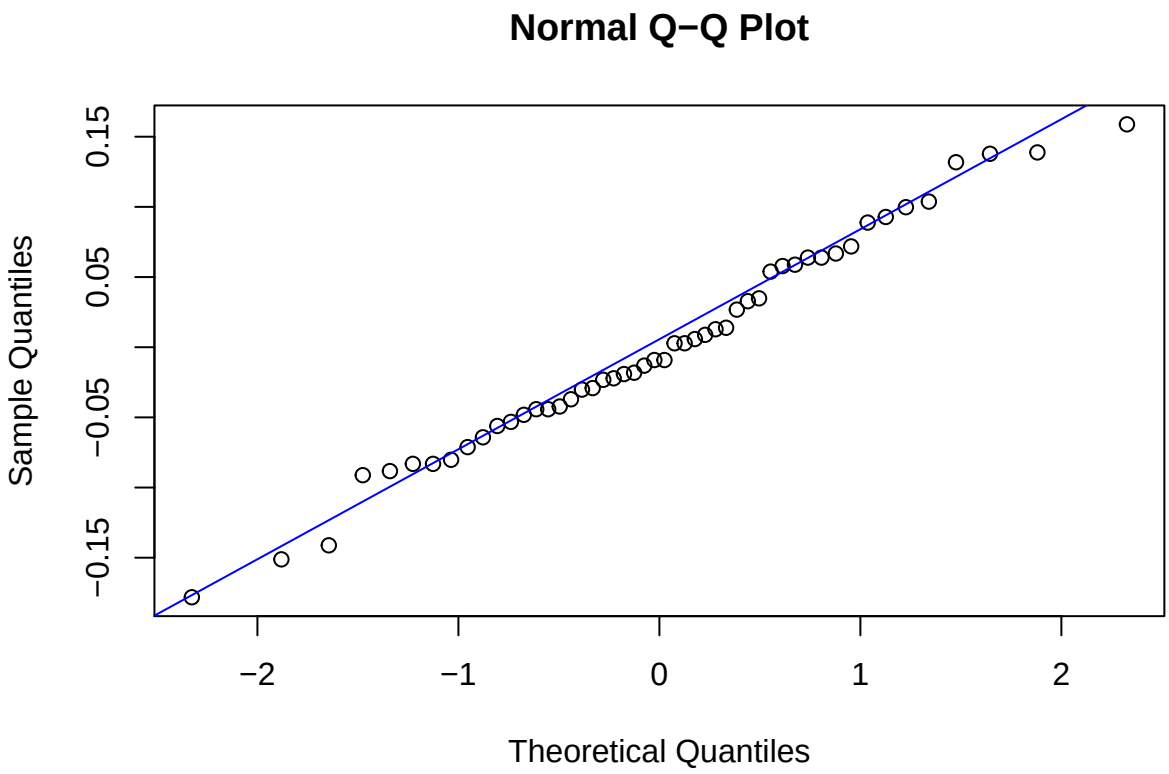

Figure 18.

```
plot(fit_bachelor$residuals ~ EduDS$ratio_both_parents)
abline(h = 0, lt = 2, col = "red")
```

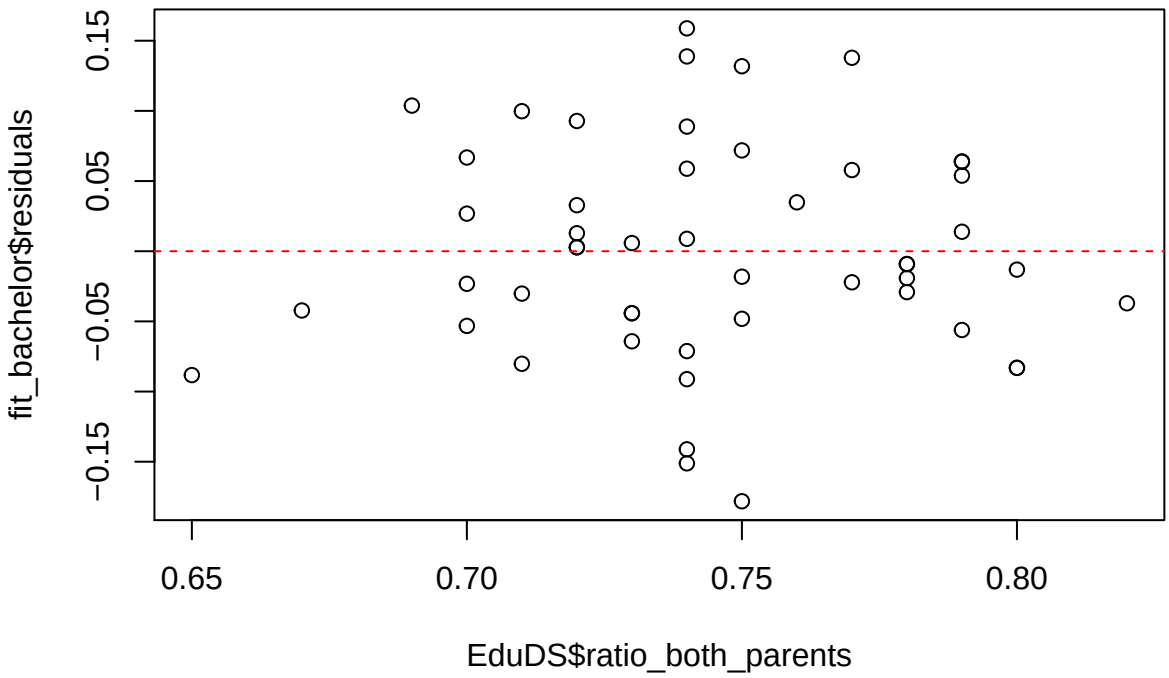

Figure 19.

Conditions check for simple regression analysis:

a. Linearity check: Figure 19 (scatterplot) shows a linear trend of the data. so linearity condtion is satisfied.

b. Nearly normal residuals: Both the histogram (Figure 17) and qqplot and qqline plots (Figure 18) show that the residuals are nearly normally distributed. So the condition is also met.

c. constant variability: The figure 19 also shows the residuals are scattered around the horizontal line almost at a constant variability, so this condition is also satisfied.

d. Independent observations: If we consider the datset as the sample from an infinite population of future cases we can assume that the sample size is less than 10% of the population, so independence is reasonable.

Therefor all the conditions of simple regression analysis are met.

```
summary(fit_bachelor)
```

```
##
## Call:
## lm(formula = avg_bachelor ~ ratio_both_parents, data = EduDS)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.178175 -0.047180 -0.009147  0.058573  0.158816
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.02612    0.22552  -0.116   0.9083
## ratio_both_parents  0.69906    0.30252   2.311   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07751 on 48 degrees of freedom
## Multiple R-squared:  0.1001, Adjusted R-squared:  0.08136
## F-statistic:  5.34 on 1 and 48 DF,  p-value: 0.02518
```

The P Value is smaller than .05 and the Coefficients of both parents is greater than zero. So two-parent family structure does have an affect on the average number of bachelor degree holders per family. Since the R-squared value is 0.10, only 10% variablity in the average bachelor degree holders can be explained by the ratio of two-parents families.

So family structures have impacts on education, therefore null hypothesis is rejected.