INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APU2F2211CS(DA) TP065116 NEAW AIK KA

HAND OUT DATE: 20 FEBRUARY 2023
HAND IN DATE: 8 MARCH 2023
WEIGHTAGE: 50%

INSTRUCTIONS TO CANDIDATES:

1. **Submit your assignment at the administrative counter**
2. **Student are advised to underpin their answer with the use of references (cited using the Harvard Name System of Referencing)**
3. **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld**
4. **Cases of plagiarism will be penalized**
5. **The assignment should be bound in an appropriate style (comb bound or stapled).**
6. **Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.**
7. **You must obtain 50% overall to pass this module.**

Table of Contents

Table of Figures

Introduction

This dataset is about the replacement of students in an Australian school. This dataset provides student information that may be used by an organization to identify any potential problems with student placement on campuses. The marketing department needs to analyze this dataset in order to find any hidden issues among the students and offer useful information for making decisions. The attributes in the dataset have included gender, age, address living, education of parents, family support, academic test result, and other personal information. The target variable of the analysis is the salary of individuals and there are students that are placed and not placed in the dataset for analyzing. We will proceed with pre-processing and analysis of this dataset in the following section.

Data import and Library using

```
#====Package Installed====
install.packages("dplyr")
install.packages("ggplot2")
install.packages("janitor") #Data Cleaning
install.packages("RColorBrewer") #Data Visualization (Colour Platte)
install.packages("hrbrthemes")
install.packages("tidyverse")
install.packages("ggpubr")
install.packages("reshape2")
install.packages("ggExtra")
install.packages("hrbrthemes")
install.packages("viridis")
install.packages("fmsb")
install.packages("GGally")
install.packages("remotes")
remotes::install_github("davidsjoberg/ggsankey")
install.packages("circlize")
```

```
#====Load Library====
library(dplyr)
library(ggplot2)
library(janitor)
library(RColorBrewer)
library(hrbrthemes)#Extra features
library(tidyverse)#Extra features
library(ggpubr)
library(reshape2)#Extra features
library(ggExtra)#extra features
library(viridis)
library(fmsb)#Extra features
library(GGally)#Extra features
library(ggsankey)#Extra features
library(circlize)#Extra features
```

Figure 1.0.1 Library using

The library we have used in this analysis includes “dplyr”, “ggplot2”, “janitor”, “RcolorBrewer”, “hrbrthemes”, “tidyverse”, “ggpubr”, “reshape2”, “ggExtra”, “viridis”, “fmsb”, “GGally”, “ggsankey”, “circlize”.

```
#====import data======
df <- read.csv("~/Placement_Data_Full_Class.csv", header=T)
```

Figure 1.0.2 import data

Then, we import the dataset by using read.csv() function.

Data Pre-processing

```
#====Exploratory data analysis EDA====
dim(df)
View(df)
arrange(df, sl_no)
str(df)
summary(df)
dfnumeric <- select_if(df, is.numeric)
dfnumeric
dfchar <- df %>% select_if(is.character)
dfchar
summary(dfnumeric)

factordf <- lapply(dfchar, factor)
lapply(factordf, levels)
```

Figure 2.0.1 Data Preprocessing

Data Preprocessing is one of the steps of EDA. In this section, we will proceed with the preprocessing according to the R code in the figure above.

```
> dim(df)
[1] 17007    25
```

Firstly, we check the dimension of the dataset. This dataset has 17007 rows of data and 25 attributes within it.

```
> str(df)
'data.frame':   17007 obs. of  25 variables:
 $ sl_no         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ gender        : chr  "M" "M" "M" "M" ...
 $ age           : int  23 19 19 21 22 19 19 18 19 21 ...
 $ address       : chr  "U" "U" "U" "U" ...
 $ Medu          : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu          : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob          : chr  "at_home" "at_home" "at_home" "health" ...
 $ Fjob          : chr  "teacher" "other" "other" "services" ...
 $ famsup        : chr  "no" "yes" "no" "yes" ...
 $ paid          : chr  "no" "no" "yes" "yes" ...
 $ activities    : chr  "no" "no" "no" "yes" ...
 $ internet      : chr  "no" "yes" "yes" "yes" ...
 $ ssc_p         : num  67 79.3 65 56 85.8 ...
 $ ssc_b         : chr  "State" "State" "Private" "Central" ...
 $ hsc_p         : num  91 78.3 68 52 73.6 ...
 $ hsc_b         : chr  "State" "Central" "Private" "State" ...
 $ hsc_s         : chr  "Commerce" "Science" "Arts" "Science" ...
 $ degree_p      : num  58 77.5 64 52 73.3 ...
 $ degree_t      : chr  "Sci&Tech" "Sci&Tech" "Comm&Mgmt" "Sci&Tech" ...
 $ workex        : chr  "No" "Yes" "No" "No" ...
 $ etest_p       : num  55 86.5 75 66 96.8 ...
 $ specialisation: chr  "Mkt&HR" "Mkt&Fin" "Mkt&Fin" "Mkt&HR" ...
 $ mba_p         : int  78 80 77 50 86 63 59 83 51 67 ...
 $ status        : chr  "Placed" "Placed" "Placed" "Not Placed" ...
 $ salary        : int  350000 200000 350000 NA 250000 NA NA 300000 350000 NA ...
```

Figure 2.0.2 str() function result

Then, we check the data types and data within each of the columns in the dataset so that we can have a better understanding of the data types of each column.

Missing data

```
> summary(df)
     sl_no          gender              age          address              Medu            Fedu           Mjob              Fjob       
 Min.   :    1   Length:17007       Min.   :18.00   Length:17007       Min.   :0.000   Min.   :0.000   Length:17007       Length:17007      
 1st Qu.: 4252   Class :character   1st Qu.:19.00   Class :character   1st Qu.:2.000   1st Qu.:1.000   Class :character   Class :character  
 Median : 8504   Mode  :character   Median :20.00   Mode  :character   Median :3.000   Median :2.000   Mode  :character   Mode  :character  
 Mean   : 8504                      Mean   :20.49                      Mean   :2.513   Mean   :2.489                                        
 3rd Qu.:12756                      3rd Qu.:22.00                      3rd Qu.:4.000   3rd Qu.:3.000                                        
 Max.   :17007                      Max.   :23.00                      Max.   :4.000   Max.   :4.000                                        

    famsup              paid           activities          internet            ssc_p          ssc_b              hsc_p       
 Length:17007       Length:17007       Length:17007       Length:17007       Min.   :40.89   Length:17007       Min.   :37.00  
 Class :character   Class :character   Class :character   Class :character   1st Qu.:61.00   Class :character   1st Qu.:61.00  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :72.00   Mode  :character   Median :72.00  
                                                                             Mean   :72.44                      Mean   :72.45  
                                                                             3rd Qu.:84.00                      3rd Qu.:84.00  
                                                                             Max.   :95.00                      Max.   :97.70  

    hsc_b              hsc_s             degree_p      degree_t           workex            etest_p      specialisation    
 Length:17007       Length:17007       Min.   :50.00   Length:17007       Length:17007       Min.   :50.00   Length:17007      
 Class :character   Class :character   1st Qu.:61.00   Class :character   Class :character   1st Qu.:61.00   Class :character  
 Mode  :character   Mode  :character   Median :72.00   Mode  :character   Mode  :character   Median :72.00   Mode  :character  
                                       Mean   :72.39                                         Mean   :72.32                     
                                       3rd Qu.:84.00                                         3rd Qu.:84.00                     
                                       Max.   :95.00                                         Max.   :98.00                     

     mba_p           status              salary       
 Min.   :50.00   Length:17007       Min.   :200000  
 1st Qu.:61.00   Class :character   1st Qu.:250000  
 Median :72.00   Mode  :character   Median :300000  
 Mean   :72.54                      Mean   :308532  
 3rd Qu.:84.00                      3rd Qu.:350000  
 Max.   :95.00                      Max.   :500000  
                                    NA's   :8265    
 '
```

Figure 2.0.1 missing data detection

Then, by using the summary() function, we can have a comprehensive view to the data of each column, including their missing data. However, there is no missing data for us to remove in this dataset. Hence, we will proceed to check the inconsistent data in the dataset.

Inconsistent data

```
> factordf <- lapply(dfchar, factor)
> lapply(factordf, levels)
$gender
[1] "F" "M"

$address
[1] "R" "U"

$Mjob
[1] "at_home"  "health"   "other"    "services" "teacher"

$Fjob
[1] "at_home"  "health"   "other"    "services" "teacher"

$famsup
[1] "no"  "yes"

$paid
[1] "no"  "yes"

$activities
[1] "no"  "yes"

$internet
[1] "no"  "yes"

$ssc_b
[1] "Central" "Private" "State"

$hsc_b
[1] "Central" "Private" "State"

$hsc_s
[1] "Arts"     "Commerce" "Science"

$degree_t
[1] "Comm&Mgmt" "Others"    "Sci&Tech"

$workex
[1] "No"  "Yes"

$specialisation
[1] "Mkt&Fin" "Mkt&HR"

$status
[1] "Not Placed" "Placed"
```

Figure 2.0.1 Inconsistent data detection

By using “lapply” function, we can check the levels of all columns at once, the result shows that all the format of categorical data in this dataset has no abnormal or wrong input data.

Outliers’ detection

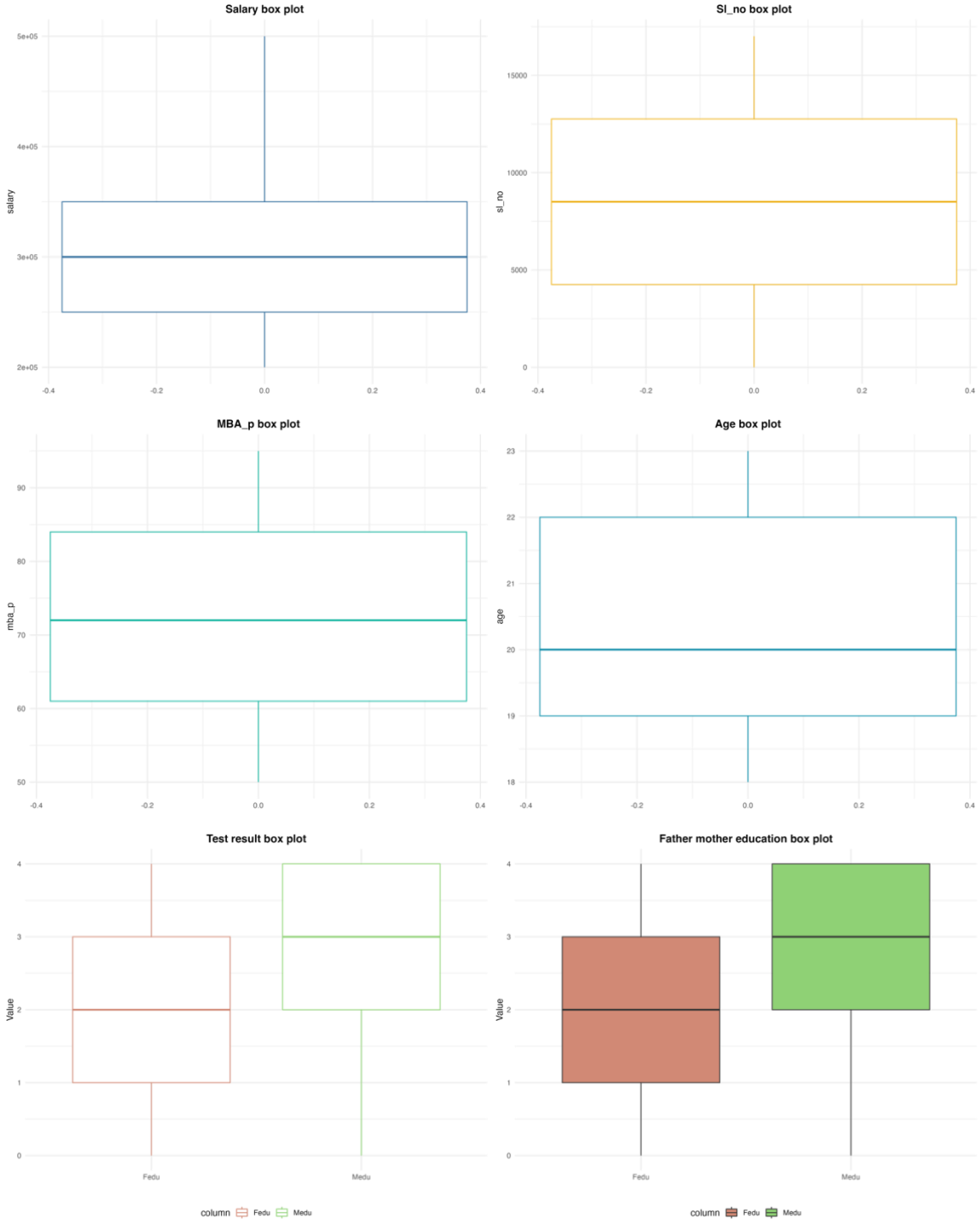
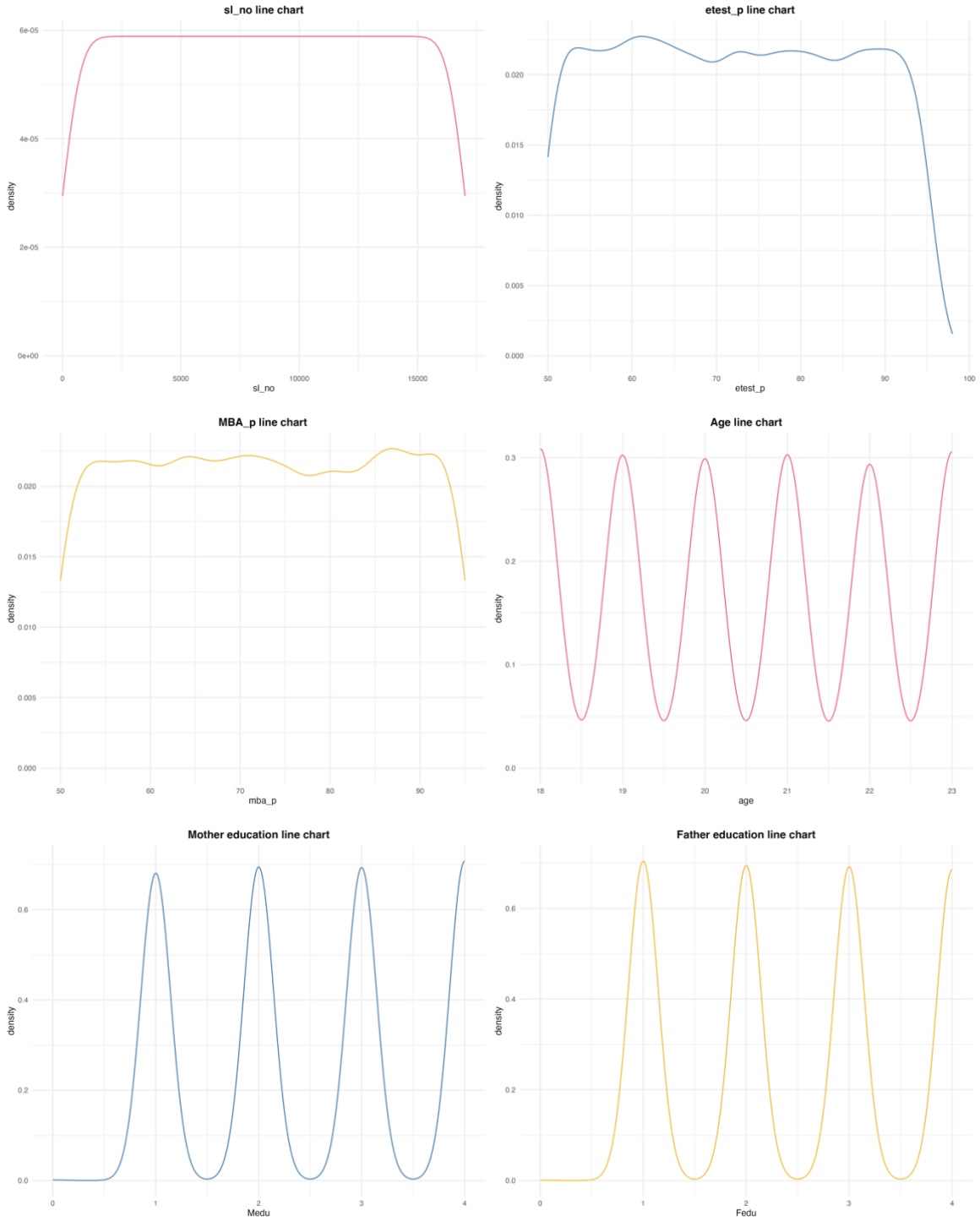


Figure 2.0.1 Outliers detection with boxplot

After that, we utilized the boxplot drawing using the geom_boxplot() function to detect the outliers of numeric data. After drawing the boxplot of numeric data, outliers have not been found in those columns.



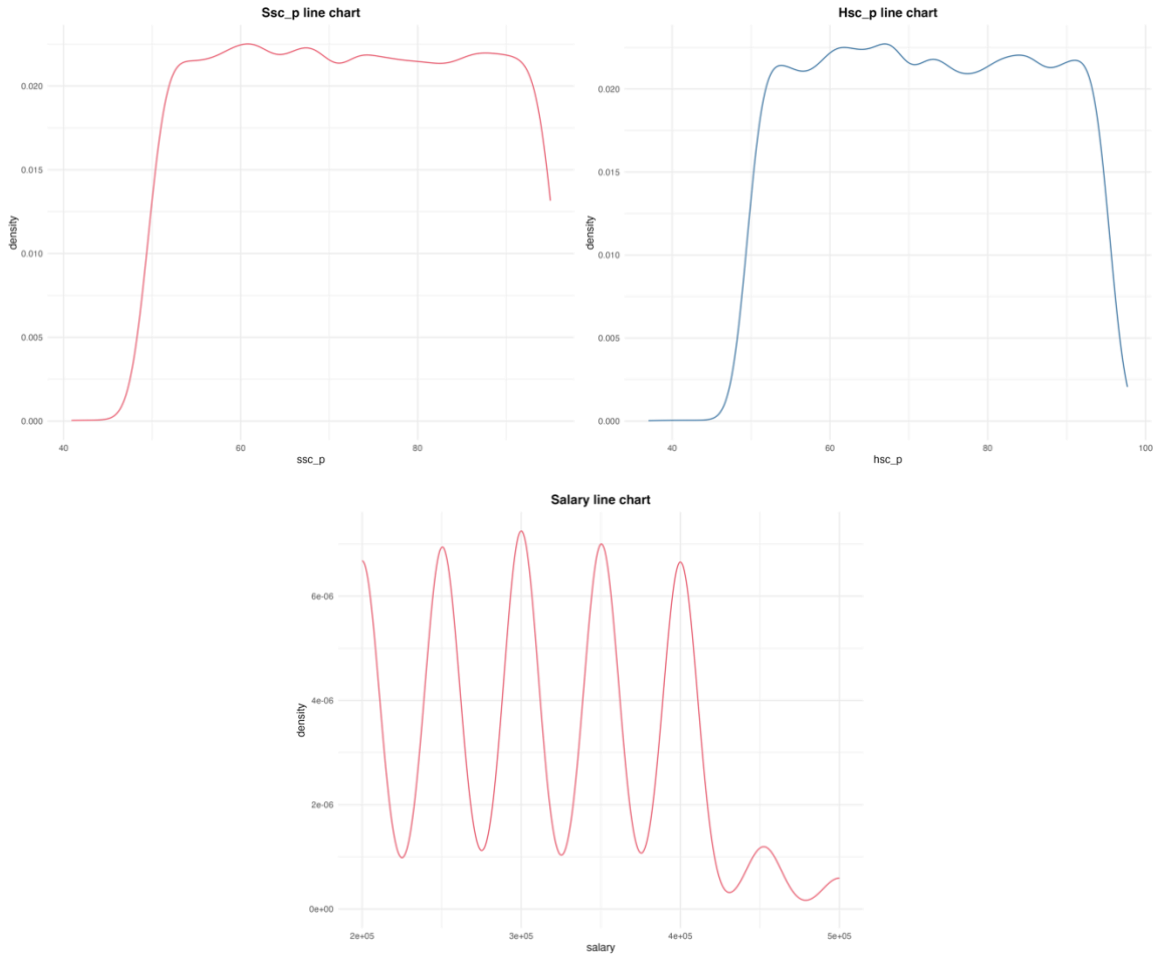


Figure 2.0.2 Distribution analysis with line chart

For the distribution of the numeric data, the distribution of its value is average and it shows no abnormal data like negative data in the columns. Hence, we are not required to eliminate any data that occurs in the dataset.

Data transformation

```
#transformation====
df <- mutate(df, age_group = ifelse(age <= 20, "18-20", ifelse(age > 20, "20-23", "23-")))
df <- df %>% mutate(Fedu_level = ifelse(Fedu==0, "none",
            ifelse(Fedu==1,"primary",ifelse(Fedu==2, "9th Grade",
                                            ifelse(Fedu==3,"secondary education", "higher education")))) )
df <- df %>% mutate(Medu_level = ifelse(Medu==0, "none", ifelse(Medu==1,"primary",
              ifelse(Medu==2, "9th Grade", ifelse(Medu==3,"secondary education", "higher education")))) )
df <- df %>% mutate(age_group = ifelse(age <= 20, "18-20", "21-23"))
df <- df %>% mutate(ssc_level = ifelse(ssc_p >= 85, "HD",
                    ifelse(ssc_p>=75, "DN",
                    ifelse(ssc_p >= 65,"CR", ifelse(ssc_p>=50, "PS",ifelse(ssc_p>=47,"MF","FL"))))))
df <- df %>% mutate(hsc_level = ifelse(hsc_p >= 85, "HD",
              ifelse(hsc_p>=75, "DN", ifelse(hsc_p >= 65,"CR", ifelse(hsc_p>=50, "PS",
              ifelse(hsc_p>=47,"MF","FL"))))))
df <- df %>% mutate(degree_level = ifelse(degree_p >= 85, "HD", ifelse(degree_p>=75, "DN",
              ifelse(degree_p >= 65,"CR", ifelse(degree_p>=50, "PS",ifelse(degree_p>=47,"MF","FL"))))))
df <- df %>% mutate(etest_level = ifelse(etest_p >= 85, "HD",
                ifelse(etest_p>=75, "DN", ifelse(etest_p >= 65,"CR",
                ifelse(etest_p>=50, "PS",ifelse(etest_p>=47,"MF","FL"))))))
df <- df %>% mutate(mba_level = ifelse(mba_p >= 85, "HD", ifelse(mba_p>=75, "DN",
                                    ifelse(mba_p >= 65,"CR", ifelse(mba_p>=50, "PS",ifelse(mba_p>=47,"MF","FL"))))))
quantile(df$salary, probs = c(0.4,0.8,1), na.rm = T)
df <- df %>% mutate(salary_class = ifelse(is.na(salary), NA, ifelse(salary>=4e+05, "T20",
                                    ifelse(salary>=3e+05,"M40", "B40"))))
df$address <- ifelse(df$address == "U", "Urban", "Rural")
dfrural <- df[df$address=="Rural",]
dfurban <- df[df$address=="Urban",]
```

Figure 3.0.1 Data Transformation

In the data transformation, we categorize the age group into two groups 18-20 and 21-23, it will be easier for us when we need to do the analysis by dividing the samples into different age groups.

Then add a new column that implies the father and mother's education in a more straightforward way that can tell us that the education is 9th Grade instead of a number like “1”. By doing so, we can easily make the graph in analyzing more visualizable in the report. For the address column, urban and rural are shown in “U” and “R”, and we change to urban and rural as well.

Next, we divide the student’s result in all tests according to the level of High distinction, distinction, credit, pass, marginal fail, and fail. The category name and range are based on the resource we obtain from the Australia USC student portal. (USC, n.d.)

Grade	Parameterspan>
High Distinction (HD)	85%-100%
Distinction (DN)	75%-84%
Credit (CR)	65%-74%
Pass (PS)	50%-64 %
Marginal Fail (MF)	47%-49%
Fail (FL)	0%-47%
Fail Absent (FA)	0%-47%

Figure 3.0.2 Result level division

Following that, we categorize the pupils' income class like how the Malaysian government categorized the income group, which is Tertiary 20% (T20), Middle 40%(M40), and Below 40% (B40). To do this we check the percentage of quantiles at 40% and 80% to determine the range. For B40, the income range is below 30,000, M40 is 30,000 – 40,000, and T20 is 40,000 and above.

```
> quantile(df$salary, probs = c(0.4,0.8,1), na.rm = T)
  40%   80%  100%
3e+05 4e+05 5e+05
```

Figure 3.0.3 quantiles of salary

Lastly, to make our analysis easier in analyzing the difference between rural and urban areas, we filter the original data separately into dfrural and dfurban.

Heatmap correlation analysis

```
#===heatmap====
corrMap <- round(cor(dfnumeric), 3)
corrMap[lower.tri(corrMap)] <- NA
meltedMap <- melt(corrMap)
ggplot(meltedMap, aes(x = Var1, y =Var2, fill=value)) + geom_tile() +
  geom_text(aes(Var1, Var2, label=value), color="white", size=4) +
  scale_fill_distiller(palette="Reds")
ggsave("~/Graph/heatmap.png")
```

Figure 4.0.1 heatmap correlation analysis

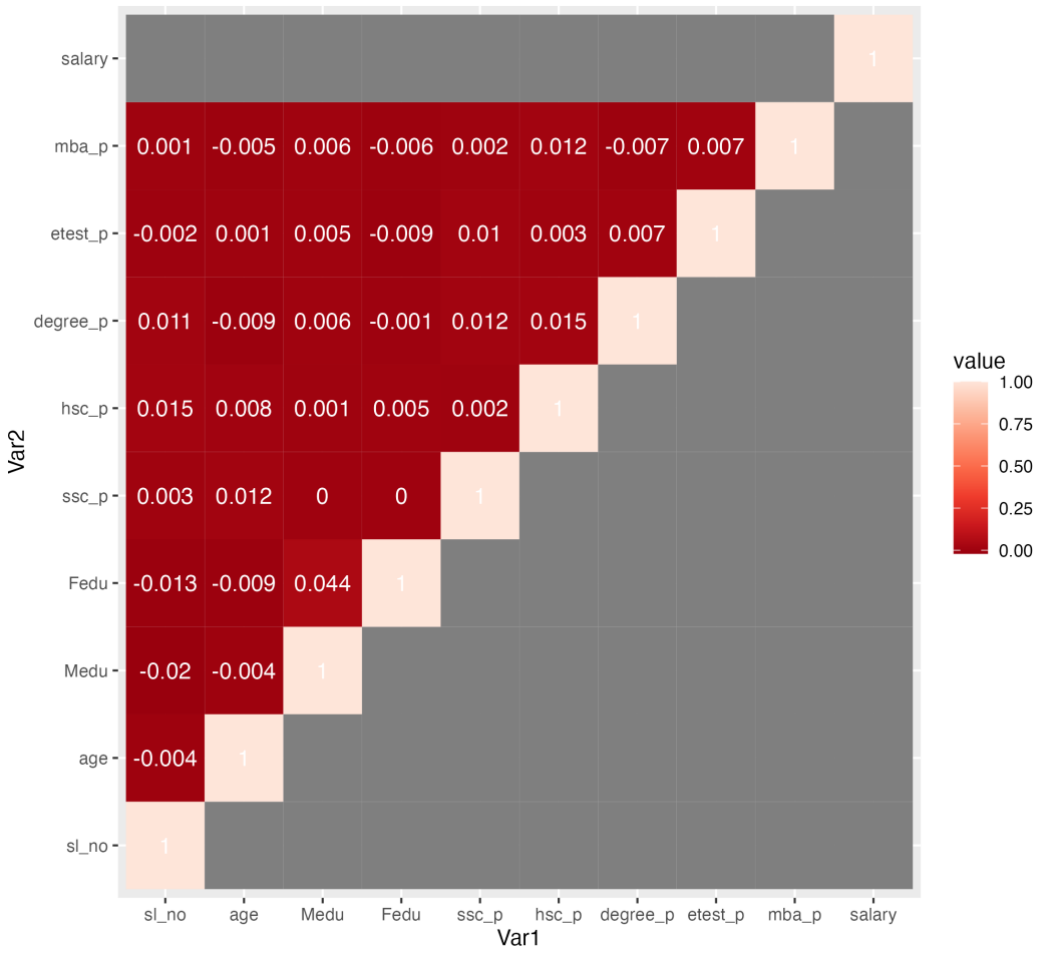


Figure 4.0.2 heatmap

For the data preprocessing before analysis, we examine the correlation index among each attribute by using R code. However, all the correlation indexes in this dataset between the two attributes are quite low.

Question 1: Are educational resources different in urban and rural areas?

Analysis 1.1: Are Father's education in different areas between individuals diverse?

```
#====Analysis 1.1:  Are Father's education in different areas between individuals diverse?====
dfedu <- df %>% select(address, Fedu_level) %>%
  group_by(address) %>% count(Fedu_level) %>%
  mutate(percent=round(n/sum(n), 2)) %>% mutate(cumulative=cumsum(percent)) %>%
  mutate(ymin = cumulative - percent)
dfedu
ggplot(dfedu, aes(ymax = cumulative, ymin = ymin, xmax = 4, xmin=3, fill=Fedu_level)) +
  geom_rect(color = "white") +
  geom_label(x = 3.5, aes(y = (cumulative + ymin)/2), label=paste0(dfedu$percent*100, "%"),
             size = 3) +
  coord_polar(theta = "y")+ facet_grid(~address) +#Extra features
  theme_ipsum_ps(grid="XY", axis="xy") +scale_color_ipsum() + scale_fill_ipsum() +
  theme_void() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color = "white"))+
  labs(fill = "Education Level", title = "Father Education in different areas") + xlim(2,4)
```

Figure 1.1.1 Relationship of Father education and address code

This code first selects two columns (address and Fedu_level) from the data frame df using the select function from the dplyr package. Then, it groups the resulting data frame by address and Fedu_level and counts the number of occurrences for each combination using group_by and count. It calculates the percentage of each combination by dividing the count by the total number of records for that address using mutate. It also computes the cumulative percentage for each group and creates a ymin column for the lower boundary of each rectangle in the plot using cumsum and mutate.

The resulting dfedu data frame is used to create a stacked bar chart showing the percentage of father education levels for each address using ggplot2 and geom_rect. The geom_label function is used to display the percentage values in the middle of each rectangle. The coord_polar function is used to transform the rectangular plot into a polar plot. The facet_grid function is used to create a separate plot for each address.

Other functions such as theme_ipsum_ps, scale_color_ipsum, scale_fill_ipsum, theme_void, theme, labs, and xlim are used to customize the appearance of the plot. Finally, the resulting plot is saved to a file using ggsave.

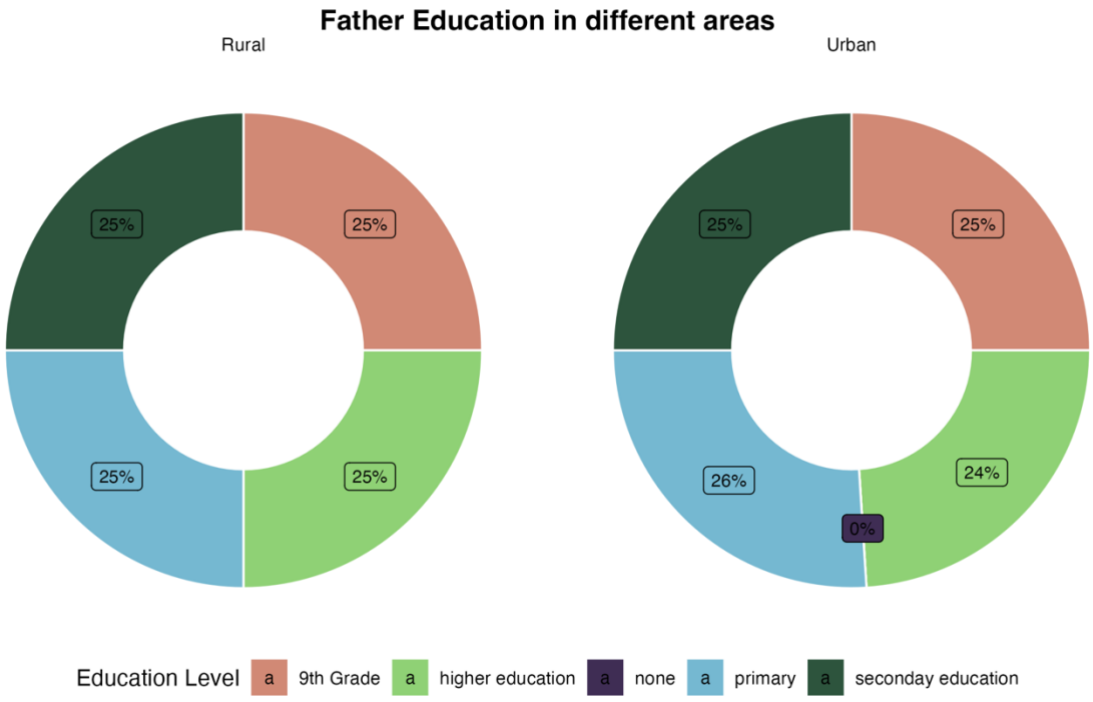


Figure 1.1.2 Relationship of Father education and address graph

The doughnut chart illustrates the differences in the level of education among fathers in rural and urban areas. The data is categorized into five levels of education: 9th grade, higher education, no education, primary education, and secondary education.

According to the chart, the percentage of fathers with the same level of education is relatively consistent across rural areas, with slight differences in percentage. However, the figure suggests that there are no significant differences in the percentage of fathers with different levels of education between rural and urban areas.

The possible reason could be that fathers in urban and rural areas have the same access to education due to the availability of schools, universities, and other educational resources. Further analysis may be needed to understand the reasons behind these differences and their implications for families and communities.

Analysis 1.2: Are Mother's education diverse between different areas?

```
#====Analysis 1.2:  Are Mother's education diverse between different areas?====
dfedu <- df %>% select(address, Medu_level) %>%
  group_by(address) %>% count(Medu_level) %>%
  mutate(percent=round(n/sum(n), 2)) %>% mutate(cumulative=cumsum(percent)) %>%
  mutate(ymin = cumulative - percent)
dfedu
ggplot(dfedu, aes(ymax = cumulative, ymin = ymin, xmax = 4, xmin=3, fill=Medu_level)) +
  geom_rect(color = "white") +
  geom_label(x = 3.5, aes(y = (cumulative + ymin)/2), label=paste0(dfedu$percent*100, "%"),
             size = 3) +
  coord_polar(theta = "y")+ facet_grid(~address)+
  labs(fill = "Education Level", title = "Mother Education in different areas") + xlim(2,4) +
  theme_ipsum_ps(grid="XY", axis="xy") +scale_color_ipsum() + scale_fill_ipsum() +
  theme_void() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
```

Figure 1.2.1 Relationship of address and Mother education code

Firstly, we load the data frame 'df', selects the columns 'address' and 'Medu_level', groups the data by 'address' and 'Medu_level', counts the frequency of each 'Medu_level' in each 'address', and calculates the percentage and cumulative percentage of each 'Medu_level' for each 'address'.

The resulting data frame 'dfedu' is then used to create a polar area chart using 'ggplot'. The 'ggplot' code creates a polar area chart with 'ymax' representing the cumulative percentage of each 'Medu_level', 'ymin' representing the cumulative percentage minus the percentage of each 'Medu_level', 'xmax' representing 4, and 'xmin' representing 3, with each 'Medu_level' represented by a fill color. The code adds a label showing the percentage of each 'Medu_level' and uses 'facet_grid' to create a separate chart for each 'address'. The code also sets the chart title, x-axis limits, and various theme settings.

Finally, the chart is saved to a file named '1p2.png' with a width of 8 inches. This code is similar to the previous code for the father's education but uses 'Medu_level' instead of 'Fedu_level' and creates a chart for the mother's education.

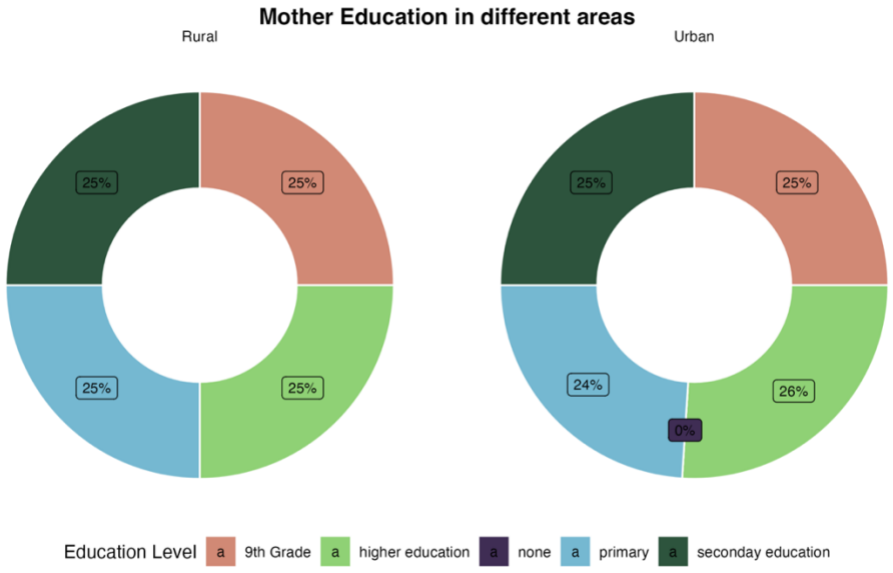


Figure 1.2.2 Relationship of address and Mother education code graph

The figure shows the distribution of mother education levels in rural and urban areas. The mother's education levels are categorized into 9th grade, higher education, none, primary, and secondary education.

In rural areas, the percentage of mothers with each education level is almost the same, with around 25% of mothers having 9th-grade education, higher education, primary education, or secondary education.

In urban areas, the percentage of mothers with 9th-grade education is 25%, which is the same as in rural areas. However, the percentage of mothers with higher education is slightly higher at 26%. Additionally, a very small percentage of mothers in urban areas have no education. The cumulative percentage of mothers with education levels from 9th grade to secondary education in both areas is 100%, showing that mother from both rural and urban areas has the same education level.

This could be due to the fact that the education system in both rural and urban areas is relatively similar, and mothers have access to similar educational resources regardless of their location. Additionally, there may be initiatives and policies in place to ensure that mothers in rural areas have access to education and can achieve similar levels of education as those in urban areas. However, more detailed analysis and data are needed to fully understand the factors contributing to the similarity in mother education levels between rural and urban areas.

Analysis 1.3: Do people in different areas prefer secondary school boards?

```
#====Analysis 1.3:  Do people in different areas prefer secondary school boards?====
dfboard <- df %>% select(address, ssc_b) %>% count(address, ssc_b) %>% group_by(address)
dfboard
ggplot(dfboard, aes(x = address, y = n, fill = ssc_b)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Secondary school board chosen in different areas") +scale_color_ipsum() +
  scale_fill_ipsum() + theme_minimal()+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p3p1.png")

ggplot(dfboard, aes(x = ssc_b, y = n, group = address, color = address, fill = address)) +
  geom_polygon(alpha = 0.2) + geom_point(shape = 24, size = 4) +
  coord_polar(start = -pi/6) + ylab("Population count") + xlab("Secondary school board") +
  ggtitle("Secondary school board chosen in different areas") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p3p2.png")
```

Figure 1.3.1 Relationship of address and Secondary School boards code

First, the code selects the "address" and "ssc_b" columns from the "df" data frame, counts the frequency of each unique combination of "address" and "ssc_b", and groups the results by "address". The resulting data frame "dfboard" contains the count of each combination of "address" and "ssc_b".

The first plot is created using the "ggplot2" package, where the x-axis represents the different areas ("address"), the y-axis represents the count of SSC boards chosen in each area, and the fill color represents the different SSC boards. The "geom_bar" function is used to create a bar chart, where each bar represents the count of each SSC board chosen in each area. The "position = 'dodge'" argument is used to display the side of the bar by side for each area.

The second plot is a polar coordinate plot created using the "ggplot2" package. The x-axis represents the different SSC boards, the y-axis represents the count of SSC boards chosen in each area, and the color represents the different areas. The "geom_polygon" function is used to create a polygonal shape for each area, where the vertices of the polygon corresponding to the count of SSC boards chosen in each area for each SSC board. The "geom_point" function is used to add a point for each area in the center of the polar coordinate plot. The "coord_polar" function is used to transform the Cartesian coordinates into polar coordinates.

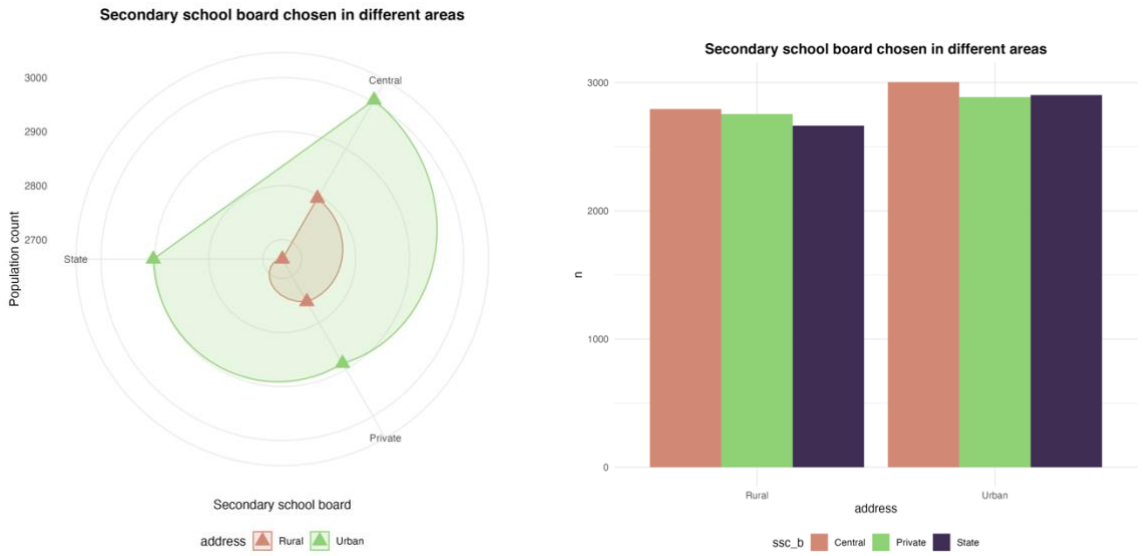


Figure 1.3.2 Relationship of address and Secondary school boards graphs

The bar chart and polygon diagram show that a slight difference causes the number of students in rural and urban areas to be the highest for Central board education, followed by State board education and Private board education. Although it seems to be a big difference in the polygon diagram, it is because the gap set in the polygon diagram is considerably small with 100 people per gap. In rural areas, the number of students is almost the same for Central, Private, and State board education, with around 2,700 to 2,800 students in each category. In urban areas, the number of students is slightly higher for Central board education, with around 3,000 students, followed by State board education and Private board education with almost the same number of students. It is obvious that the students in this dataset have no preference in choosing secondary school boards.

Usually, the number of students in urban areas like Central board should be the highest one as its reputation for providing quality education and preparing students for competitive exams, is highly valued by parents and students alike. Private board education, on the other hand, maybe less popular due to the high cost of tuition and fees, which may be unaffordable for many families. The State board of education may also have a lower number of students due to the perception of lower-quality education compared to the Central board. Hence, further observations and data collection should be conducted to prove this analysis result.

Analysis 1.4: Do people in different areas have different choices in choosing their high school board?

```
#====Analysis 1.4:  Do people in different areas have different choices in choosing their high school board?====
dfboard <- df %>% select(address, hsc_b) %>% count(address, hsc_b) %>% group_by(address)
dfboard
ggplot(dfboard, aes(x = address, y = n, fill = hsc_b)) +
  geom_bar(stat = "identity", position = "dodge")+
  scale_color_ipsum() + scale_fill_ipsum() +  theme_minimal()+
  ggtitle("High School Board chosen in different areas") + ylab("count") +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p4p1.png")


ggplot(dfboard, aes(x = hsc_b, y = n, group = address, color = address, fill = address)) +
  geom_polygon(alpha = 0.2) + geom_point(shape = 24, size = 4) +#Extra features
  coord_polar(start = -pi/6) + ylab("Population count") + xlab("Secondary school board") +
  ggtitle("High School Board chosen in different areas") +
  theme(legend.position = "bottom", plot.title = element_text(hjust = 0.5)) +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal()+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p4p2.png")
```

Figure 1.4.1 Relationship of address and High school boards code

We start by selecting the columns "address" and "hsc_b" from the data frame "df" using the dplyr package's select function. Then, we count the number of occurrences of each combination of "address" and "hsc_b" using the count function and group the result by "address" using the group_by function. The resulting data frame is stored in "dfboard".

Next, we create a bar plot using the ggplot2 package, with "address" on the x-axis and the count of each combination of "address" and "hsc_b" on the y-axis. We use the fill aesthetic to color the bars based on the "hsc_b" variable. We add a title to the plot, adjust the legend position, and save it as a PNG file using the ggsave function.

Then, we create a polar coordinate plot using ggplot2 with "hsc_b" on the x-axis and the count of each combination of "address" and "hsc_b" on the y-axis. We use the color aesthetic to distinguish between each "address" and the fill aesthetic to color the polygon based on the "address" variable. We add a title to the plot, adjust the legend position, and save it as a PNG file using the ggsave function.

Overall, these code snippets aim to visualize the distribution of the "hsc_b" variable across different "address" categories in the dataset "df".

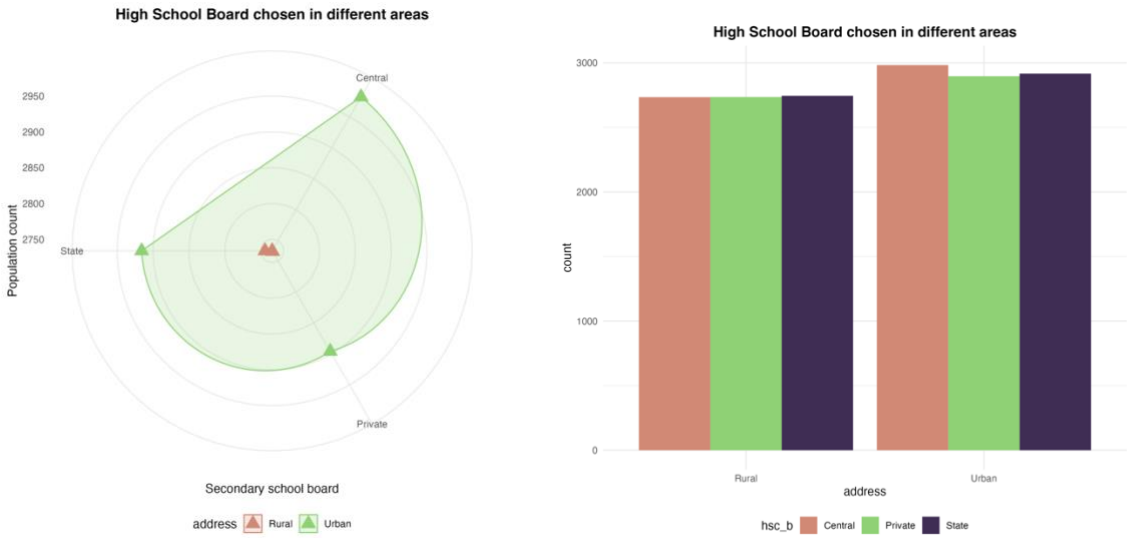


Figure 1.4.2 Relationship of address and High school boards graphs

The bar chart and polygon diagram drawn represent the distribution of the type of board (HSC_B) in rural and urban areas. For the bar chart, the 3 different types of boards, which are Central, Private, and State, reside on the perimeter line of the circle. The triangle marks the frequency of the students joining the board. The closer the triangle is to the perimeter line, the more students join the particular school board. The gap in the polygon diagram is comparatively bigger it shows a big difference between the choices of people from urban and rural areas. However, when we keep examining the data using the bar chart, it shows there is only a tiny difference between the number of joining students. In both urban and rural areas, the highest frequency is for the State Board, while the lowest frequency is for the Central board. The gap between the student number for each type of high school board also remains at 100 to 200.

Overall, there is not a significant difference between the distribution of the types of boards in rural and urban areas.

Analysis 1.5: Will the school boards affect the number of students joining in the extra-curricular activities?

```
#====Analysis 1.5:  Will the school boards affect the number of students joining in the extra-curricular activities?====
dfboard <- df %>% select(ssc_b, hsc_b, activities) %>%
  mutate(ssc_b = ifelse(ssc_b == "Central", "Secondary Central",
                        ifelse(ssc_b == "Private", "Secondary Private", "Secondary State"))) %>%
  mutate(hsc_b = ifelse(hsc_b == "Central", "High Central",
                        ifelse(hsc_b == "Private", "High Private", "High State"))) %>%
  pivot_longer(cols = c(ssc_b, hsc_b), values_to = "board") %>% count(board,activities)
dfboard
ggplot(dfboard, aes(board, n, fill=activities)) +
  geom_bar(stat = "identity", position = "stack")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Extra-curricular activities in different school board")+ ylab("count") +xlab("board of school") +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p5.png")
```

Figure 1.5.1 Relationship of school boards and extra-curricular activities code

In this code, we are first selecting the columns 'ssc_b', 'hsc_b', and 'activities' from a data frame 'df'. We then use the 'mutate' function to replace the values in the 'ssc_b' and 'hsc_b' columns with new values based on specific conditions. Specifically, we replace "Central", "Private", and "State" values with "Secondary Central", "Secondary Private", and "Secondary State" for the 'ssc_b' column, and "High Central", "High Private", and "High State" for the 'hsc_b' column.

Next, we use the 'pivot_longer' function to reshape the data so that the 'ssc_b' and 'hsc_b' columns are combined into a single column called 'board', and the corresponding values are placed in a new column called 'value'. We then count the number of occurrences of each 'board' and 'activities' combination using the 'count' function.

Finally, we create a stacked bar chart using 'ggplot', where the 'board' values are displayed on the x-axis, the count of each 'board' and 'activities' combination is displayed on the y-axis, and the 'activities' values are represented by different fill colors. We also add a title, x-axis, and y-axis labels, and adjust the theme and legend settings. The resulting plot shows the distribution of extracurricular activities in different types of school boards.

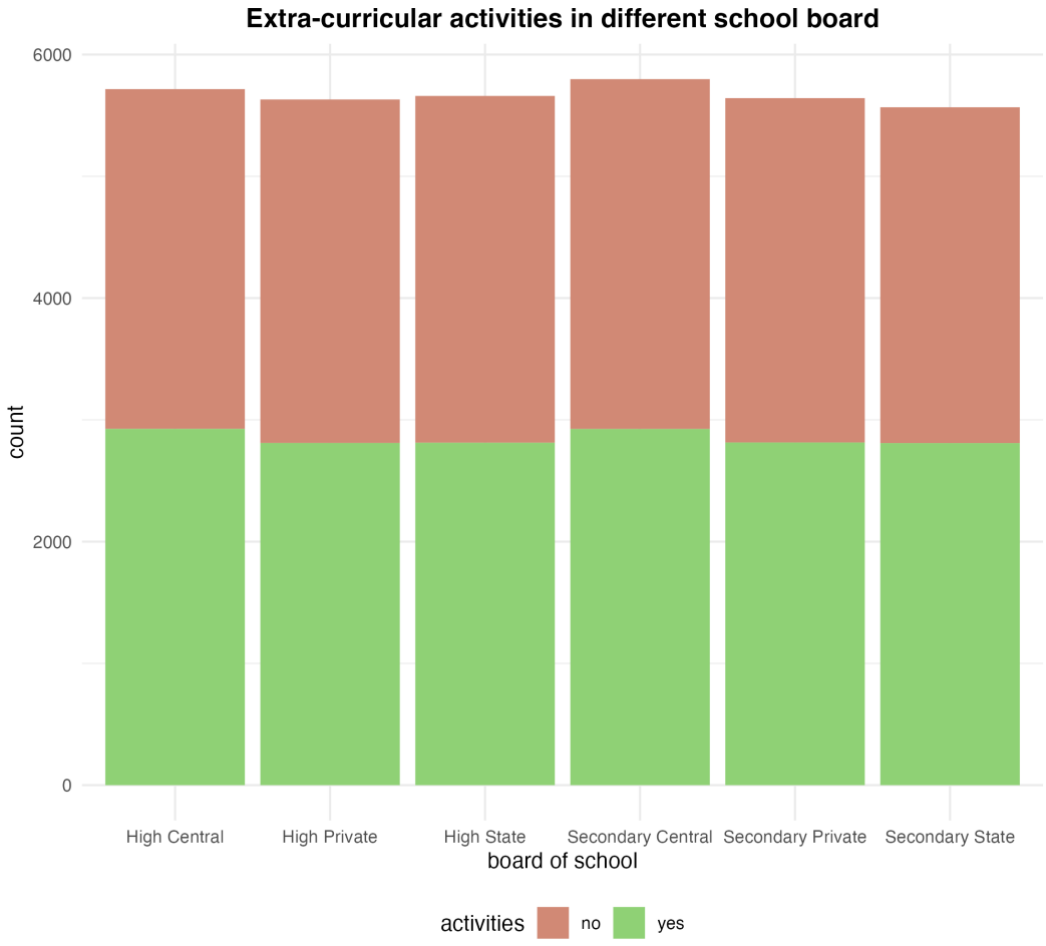


Figure 1.5.2 Relationship of school boards and extra-curricular activities graph

The bar chart depicts the number of students who participate in extracurricular activities or not, based on their board type (High Central, High Private, High State, Secondary Central, Secondary Private, and Secondary State). The chart has 6 bars for each board type.

One possible reason for the differences in participation in extracurricular activities among students from different board types could be the availability of resources and opportunities. For example, students from High Private boards may have access to more resources and opportunities for extracurricular activities due to higher fees and better infrastructure. On the other hand, students from Secondary State boards may have limited access to resources and opportunities for extracurricular activities due to lower funding and less infrastructure. This could contribute to the differences in participation in extracurricular activities among students from different board types.

However, the stacked bar chart drawn based on this dataset shows no difference between the number of students from different school board joining in extra-curricular activities.

Analysis 1.6: Will the school boards affect the number of students joining extra-paid classes?

```
#====Analysis 1.6:  Will the school boards affect the number of students joining extra-paid classes?====
dfboard <- df %>% select(ssc_b, hsc_b, paid) %>%
  mutate(ssc_b = ifelse(ssc_b == "Central", "Secondary Central",
                        ifelse(ssc_b == "Private", "Secondary Private", "Secondary State"))) %>%
  mutate(hsc_b = ifelse(hsc_b == "Central", "High Central",
                        ifelse(hsc_b == "Private", "High Private", "High State"))) %>%
  pivot_longer(cols = c(ssc_b, hsc_b), values_to = "board") %>% count(board,paid)
dfboard
ggplot(dfboard, aes(board, n, fill=paid)) + geom_bar(stat = "identity", position = "stack")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Extra paid classes in different school board")+ ylab("count") +xlab("board of school") +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p6.png")
```

Figure 1.6.1 Relationship of school boards and extra-paid classes code

In Analysis 1.6, we are analyzing the relationship between school boards and the number of students joining extra-paid classes. To do this, we start by selecting the columns 'ssc_b', 'hsc_b', and 'paid' from the original dataset 'df'. We then use the 'mutate' function to replace the values of 'ssc_b' and 'hsc_b' with more readable names and use 'pivot_longer' to transform the data from wide to long format, with 'board' as the new column containing the values of 'ssc_b' and 'hsc_b'. We then count the number of students for each combination of 'board' and 'paid' using the 'count' function and store the results in the 'dfboard' data frame.

We then create a stacked bar chart using 'ggplot' to visualize the relationship between 'board' and 'paid'. We use the 'geom_bar' function to create the stacked bars, with 'fill' set to 'paid'. We also use the 'scale_color_ipsum' and 'scale_fill_ipsum' functions to set the colors of the chart and use 'theme_minimal' to apply a minimalistic theme. We add a title to the chart using 'ggtitle' and set the labels of the x-axis and y-axis using 'xlab' and 'ylab', respectively. Finally, we use 'theme' to adjust the position and appearance of the legend, title, and background of the chart and save the chart as a PNG file using 'ggsave'.

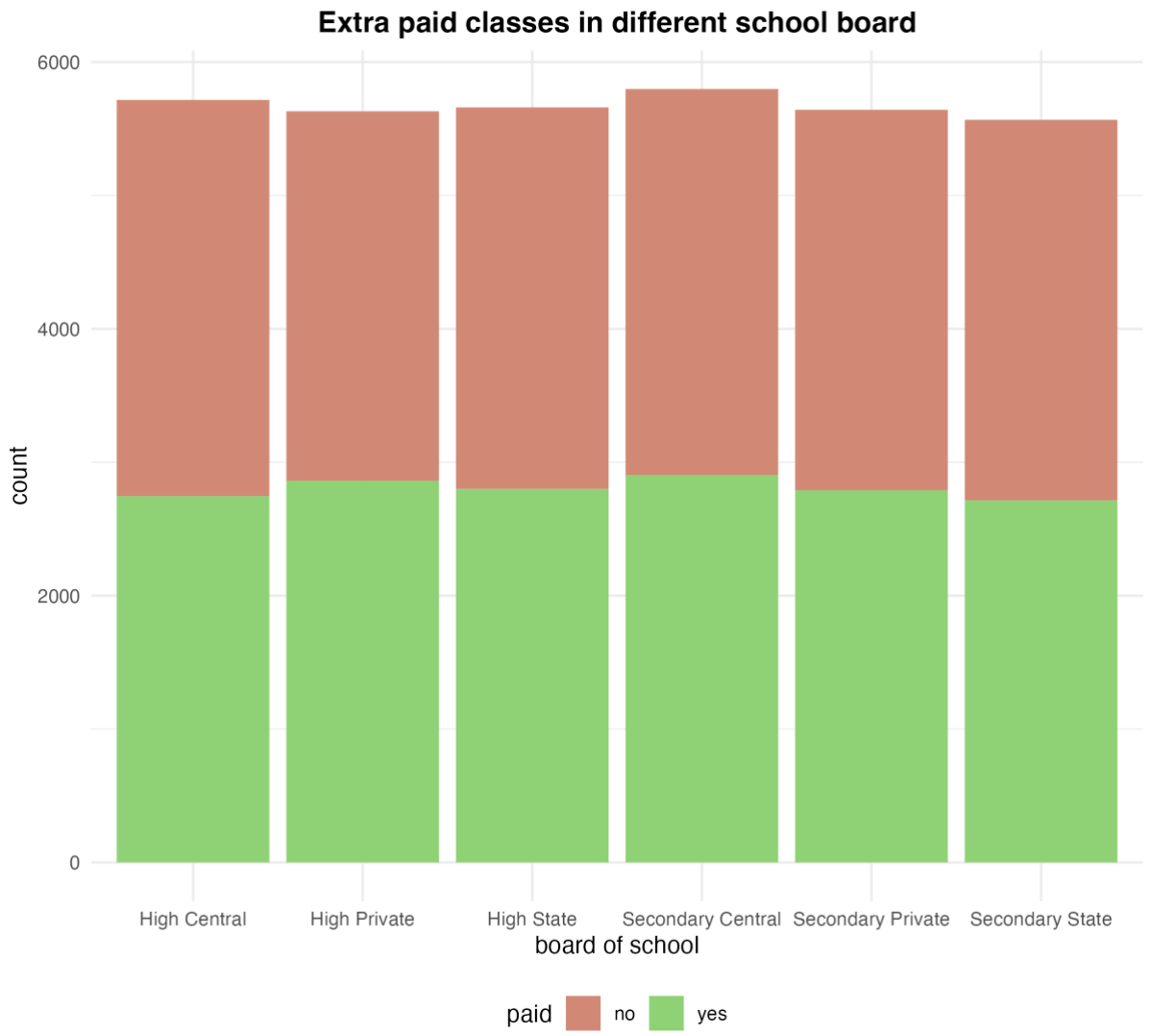


Figure 1.6.2 Relationship of school boards and extra-curricular classes graph

The bar chart shows the frequency of students who pay for extra classes or tuition separated by their education board (high school or secondary school) and location (Central, Private, or State).

One possible reason for the differences in the frequency of paid coaching classes or tuition could be related to the quality of education provided by the different types of schools. Students attending government-run schools (State) may be less satisfied with the quality of education and feel the need for additional coaching classes or tuition to supplement their learning. Conversely, students attending private schools may receive better quality education and may not feel the need for additional coaching classes or tuition.

Analysis 1.7: Is there any difference for students from different areas in terms of willingness in joining extra-paid classes?

```
#====Analysis 1.7:  Is there any difference for students from different areas in terms of willingness in joining extra-paid classes?====
dfpaid <- df %>% select(address, paid) %>% count(address, paid)
dfpaid
ggplot(dfpaid, aes(x=paid, y=n, fill=address)) +
  geom_bar(position = "stack", stat = "identity") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Extra paid classes in different areas")+ ylab("count") +xlab("board of school") +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p7.png")
```

Figure 1.7.1 Relationship of extra-paid classes and address code

In this analysis, we are examining whether there is any difference in willingness among students from different areas to join extra-paid classes. To do this, we first select the columns "address" and "paid" from the data frame and then count the number of students from each area who are willing to join extra-paid classes. We then create a bar plot using ggplot, where the x-axis represents whether the student is willing to join extra-paid classes or not, and the y-axis represents the count of students. The bars are stacked by the "address" column, which represents the different areas. Finally, we add some aesthetic elements to the plot, such as title, x-label, y-label, color, and fill, and save it as a png file.

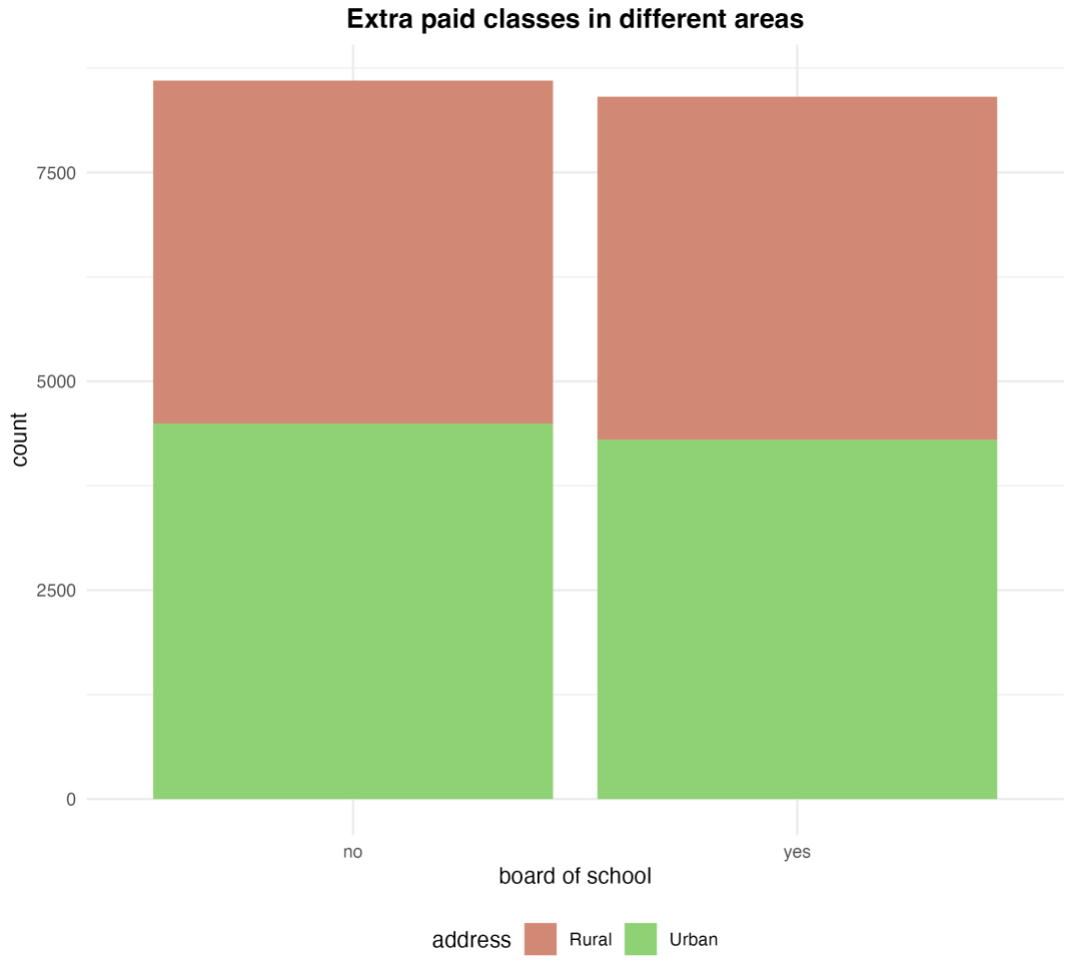


Figure 1.7.2 Relationship of extra-paid classes and address graph

The bar chart represents the number of students from rural and urban areas who have paid and have not paid for additional classes. The x-axis represents the address (rural/urban) and the y-axis represents the number of students. There are two bars for each address, one representing the number of students who have paid and the other representing the number of students who have not paid.

The possible reason for the difference in the number of students who have paid for additional coaching classes between rural and urban areas occurs is the difference in the availability of coaching centers. Urban areas may have more coaching centers, providing easier access for students to attend such classes, whereas rural areas may have limited coaching centers with fewer facilities, making it difficult for students to the availability of such classes. However, this phenomenon has not occurred in the observations.

In a summary, students from different areas in this dataset have no difference in the mindset of joining extra classes.

Analysis 1.8: Is there any difference for students from different areas in joining extra-curricular activities?

```
#====Analysis 1.8:  Is there any difference for students from different areas in joining extra-curricular activities?====
dfactivity <- df %>% select(address, activities) %>% count(address, activities)
dfactivity
ggplot(dfactivity, aes(x=activities, y=n, fill=address)) +
  geom_bar(position = "stack", stat = "identity") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Extra-curricular Activities in different areas")+ ylab("count") +xlab("board of school") +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p8.png")
```

Figure 1.8.1 Relationship of extra-curricular activities and address code

In this analysis, we are looking at whether there is any difference in the willingness of students from different areas in joining extra-curricular activities.

The code starts by selecting the address and activities columns from the df data frame and counting the number of occurrences for each combination of values using the count() function. The resulting data frame is stored in the variable dfactivity.

Then, a bar plot is created using the ggplot() function, with dfactivity as the data source. The x-axis is set to activities, the y-axis is set to n (the count of each combination of address and activities), and the fill color is set to address. The geom_bar() function is used to create a stacked bar chart with bars for each combination of activities and address.

Finally, various formatting options are applied to the plot using the theme () function, such as setting the legend position to the bottom and the plot background color to white. The resulting plot is saved as a PNG file using the ggsave() function.

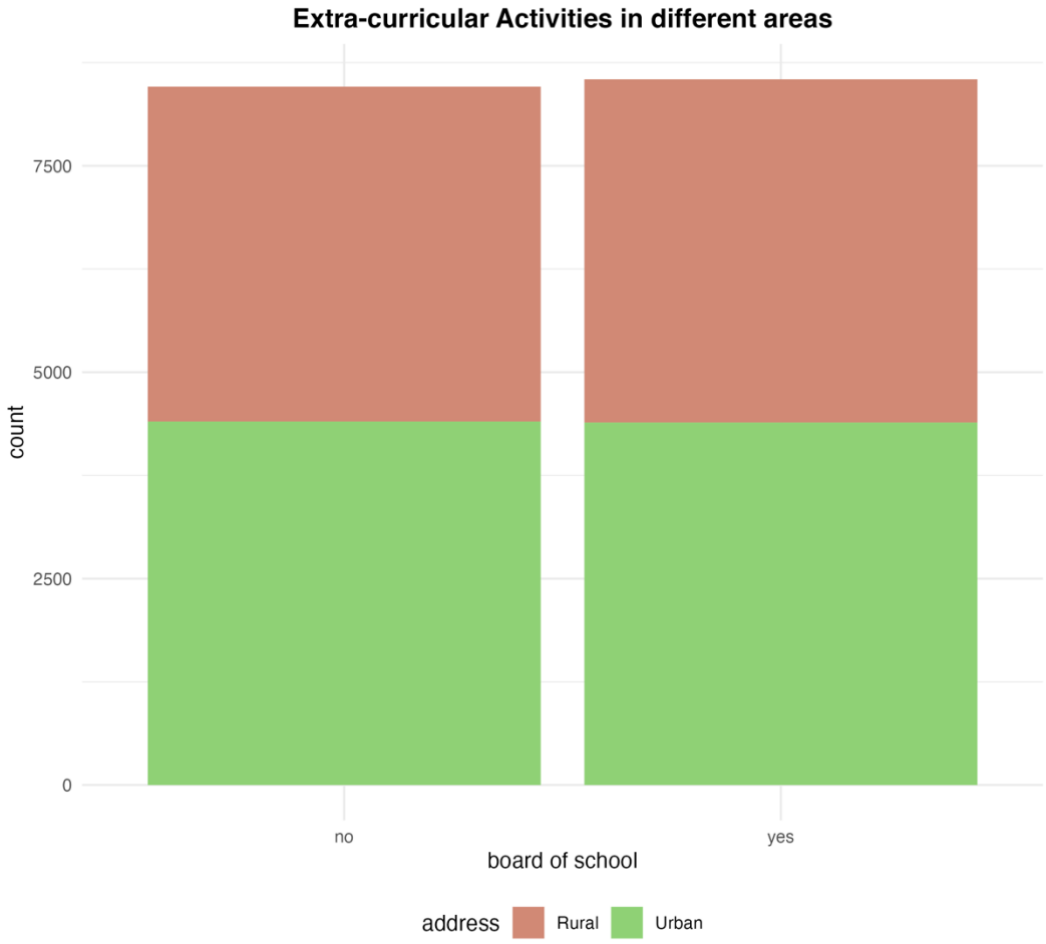


Figure 1.8.2 Relationship of extra-curricular activities and address graph

The bar chart represents the distribution of students' participation in extra-curricular activities based on their address. It shows that there are more students in urban areas participating in extra-curricular activities compared to those in rural areas. However, the difference is not significant, as the number of students in both areas who participate in extra-curricular activities is relatively close.

The differences in participation rates in extracurricular activities between rural and urban students could be attributed to the varying availability of resources and opportunities in their respective areas. Urban areas typically have more resources and facilities, such as sports clubs, community centers, and cultural centers, which offer a wider range of extracurricular options for students. Additionally, urban areas may receive more financial support from local governments or private organizations to fund these activities. However, the observations and data from the diagram do not support this conclusion, suggesting that further analysis is needed to better understand the underlying factors contributing to the differences in participation rates.

Analysis 1.9: Does the internet access for students different in urban and rural areas?

```
#====Analysis 1.9:  Does the internet access for students different in urban and rural areas?=====
dfinternet <- df %>% select(address, internet) %>% count(address, internet)
dfinternet
ggplot(dfinternet, aes(x=internet, y=n, fill=address)) +
  geom_bar(position = "stack", stat = "identity") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Internet access in different areas")+ ylab("count") +xlab("board of school") +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p9p1.png")
```

Figure 1.9.1 Relationship of internet access and address code

In this code, the data frame df is used to create a new data frame dfinternet that selects the columns address and internet and then counts the number of students in each category.

Then, ggplot function is used to create a stacked bar plot showing the number of students with and without internet access in urban and rural areas. The x-axis represents the internet access, and the y-axis represents the count of students. The fill color represents the address (urban or rural) of the students.

The scale_color_ipsum() and scale_fill_ipsum() functions are used to set the color scheme of the plot. theme_minimal() function is used to set the theme of the plot, and ggtitle(), ylab(), xlab(), and theme() functions are used to set the title, x-axis label, y-axis label, and various plot parameters respectively.

Finally, ggsave() function is used to save the plot in a PNG file format with the given file path.

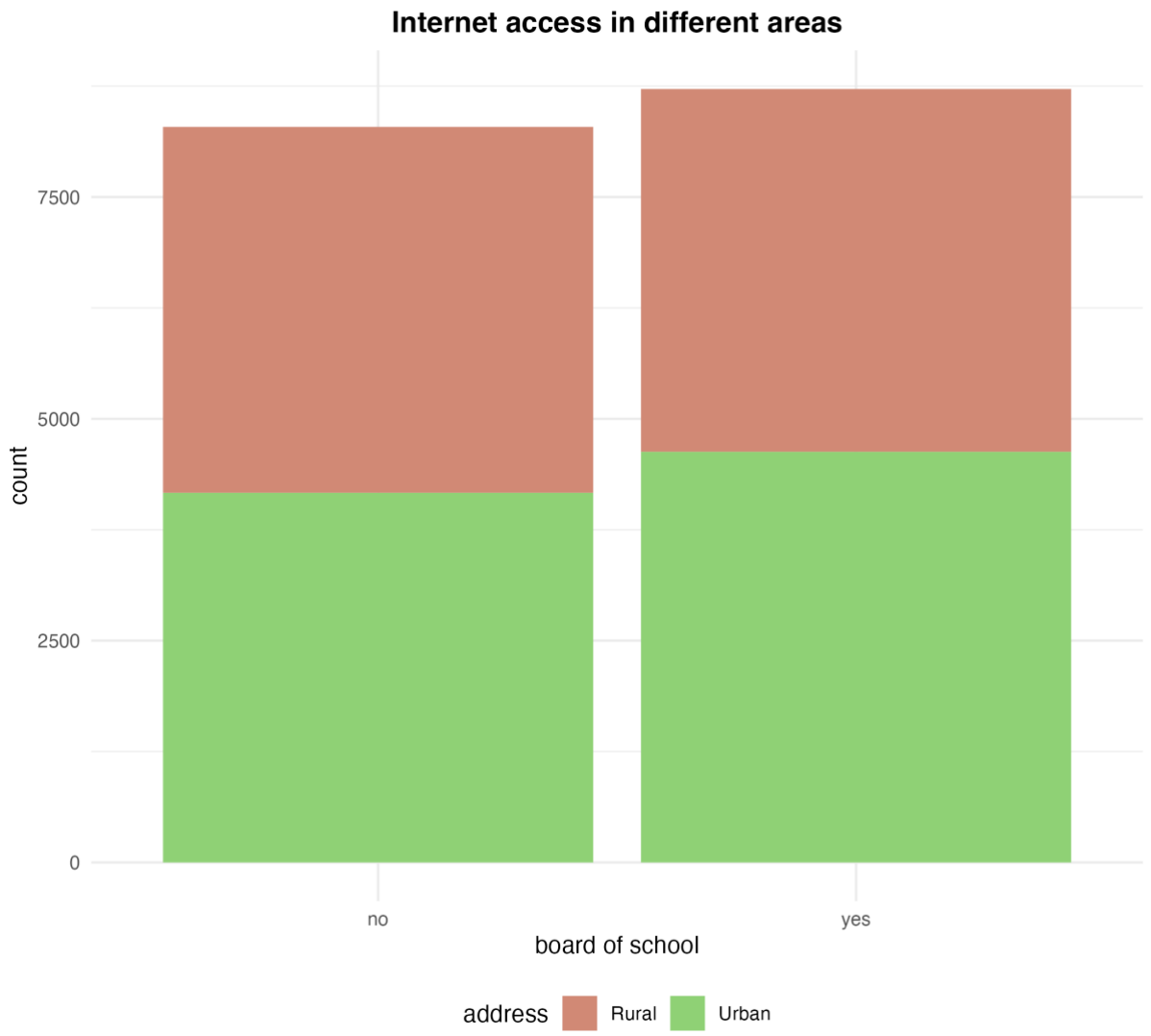


Figure 1.9.2 Relationship of internet access and address graph

The stacked bar chart above is the stacked count of the students in rural and urban areas whether they have internet access or not. For these two bars, the rural and urban areas' students both occupied half of the students with internet access and no internet access.

The heights of the bars represent the frequency of each category. the bar for "yes" in the "internet" variable is slightly taller than the bar for "no", indicating that the frequency of "yes" is slightly higher than "no" for the "internet" variable.

The diagram shows that there are similar proportions of rural and urban students with internet access. One possible reason for this could be the widespread availability of internet connectivity across both urban and rural areas due to the increasing usage and importance of the internet in daily life. However, it is important to note that the dataset may not capture differences in the quality or reliability of internet access in rural versus urban areas, which

could impact students' ability to access online resources and participate in remote learning activities. Therefore, further research on the quality and accessibility of internet connectivity in different areas may be necessary to fully understand the potential impact on student's academic performance and opportunities.

Summary for Analysis 1.9 the difference between extra-curricular activities, internet access, and extra-paid classes for students from urban and rural areas

```
#Summary 1.9====
dfsummary <- df %>% select(address, paid, activities, internet) %>%
  count(address, paid, activities, internet)
dfsummary
ggplot(dfsummary, aes(x = activities, y = n, fill = internet)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(paid~address) + scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Extra-curricular Activities, internet access and
Extra-paid classes in different areas")+
  ylab("count") +xlab("Extra-curricular activities") +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p9Sum1.png")

ggplot(dfsummary, aes(x = interaction(address, paid, internet), y = n, fill = activities)) +
  geom_col() +
  scale_fill_manual(values = c("#4E79A7", "#F28E2B"), labels = c("no", "yes")) +
  labs(x = "Address, Paid, and Internet", y = "Count") +
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Extra-curricular Activities, internet access and
Extra-paid classes in different areas")+ ylab("count")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p9Sum2.png")
```

Figure summary 1.0.1 Summary for analysis 1.9 code

First, the dfsummary table is created using the select() and count() functions from the dplyr package. It counts the number of students by their address, participation in extra-paid classes, extra-curricular activities, and access to the internet.

Next, the ggplot() function from the ggplot2 package is used to create a grouped bar chart with different facets for each combination of extra-paid classes and addresses. The x-axis represents the extra-curricular activities, while the y-axis shows the count of students. The bars are colored by their access to the internet. The facet_grid() function creates separate panels for each combination of extra-paid classes and address.

Then, we create a stacked bar chart showing the count of students in different areas who participate in extra-curricular activities and have internet access, with the fill color representing whether they attend extra-paid classes or not. The x-axis is an interaction of the address, paid, and internet variables and the y-axis represent the count of students. The legend at the bottom indicates the fill color for the "no" and "yes" values of the "paid" variable. The title of the graph is "Extra-curricular Activities, internet access and Extra-paid classes in different areas" and the axis labels are "Address, Paid, and Internet" and "Count". The graph is saved as a PNG file with the file name "1p9Sum2.png" in the Graph folder.

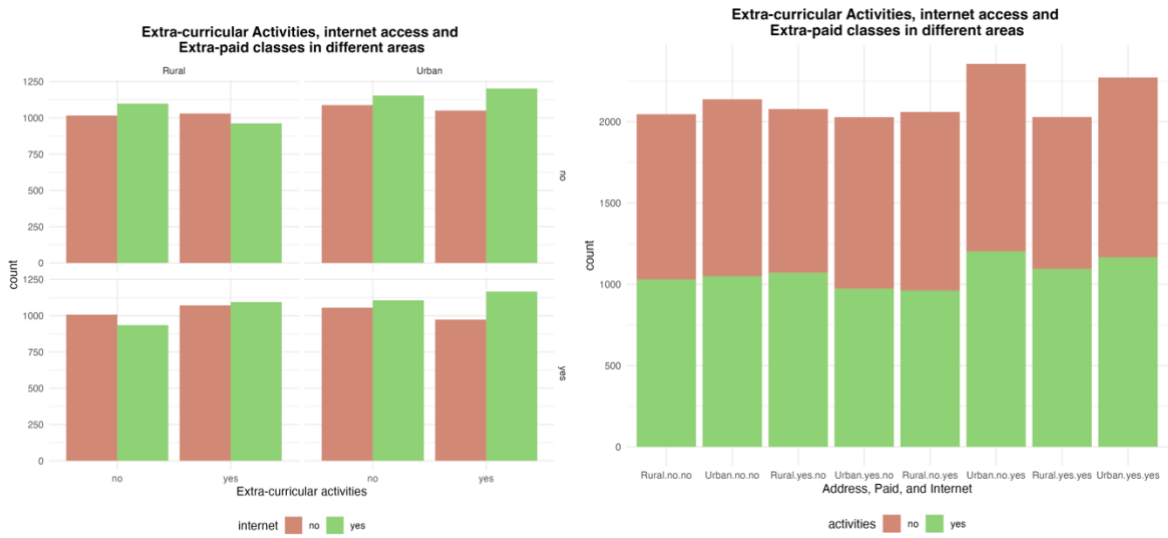


Figure summary 1.0.2 Summary 1.9 graphs

By joining all the attributes in the two bar charts above, we examine the proportion for all attributes in various situations and given conditions. The proportion for all situations is all occupying half of the bars, meaning that the conclusions above show no relationships among all attributes including activities, internet, paid, and areas correct. Notably, the analysis from this dataset is not the truth and needs further examination.

Analysis 1.10: Is there any tendency of students from different areas and school boards in choosing specialization?

```
#====Analysis 1.10: Is there any tendency of students from different areas and school boards in choosing specialization?====
dfspec <- df %>% select(address, hsc_b, hsc_s) %>% count(address, hsc_b, hsc_s) %>%
  make_long(address, hsc_b, hsc_s)
dfspec
ggplot(dfspec, aes(x = x, next_x = next_x, node = node,next_node = next_node,
                   fill = factor(node))) + geom_sankey() + theme_sankey(base_size = 16) +
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Specialisation chosen in different areas and school board")+ xlab("Choice")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p10.png")
```

Figure 1.10.1 Relationship of address, school boards and specialization code

In this code, we use the %>% pipe operator to perform a series of operations on the dataset df to investigate the tendency of students from different areas and school boards in choosing specializations.

First, we select the columns "address", "hsc_b", and "hsc_s" using select(), and then count the number of occurrences of each combination of these variables using count(). We then transform the dataset from wide to long format using the custom function make_long(), which creates a new dataset dfspec.

Finally, we create a Sankey diagram using ggplot() and geom_sankey() functions to visualize the flow of students in choosing specializations in different areas and school boards. The fill aesthetic is used to differentiate between the different nodes, which correspond to the different choices of specializations. The ggtitle(), xlab(), and theme() functions are used to customize the plot's title, x-axis label, and theme, respectively.

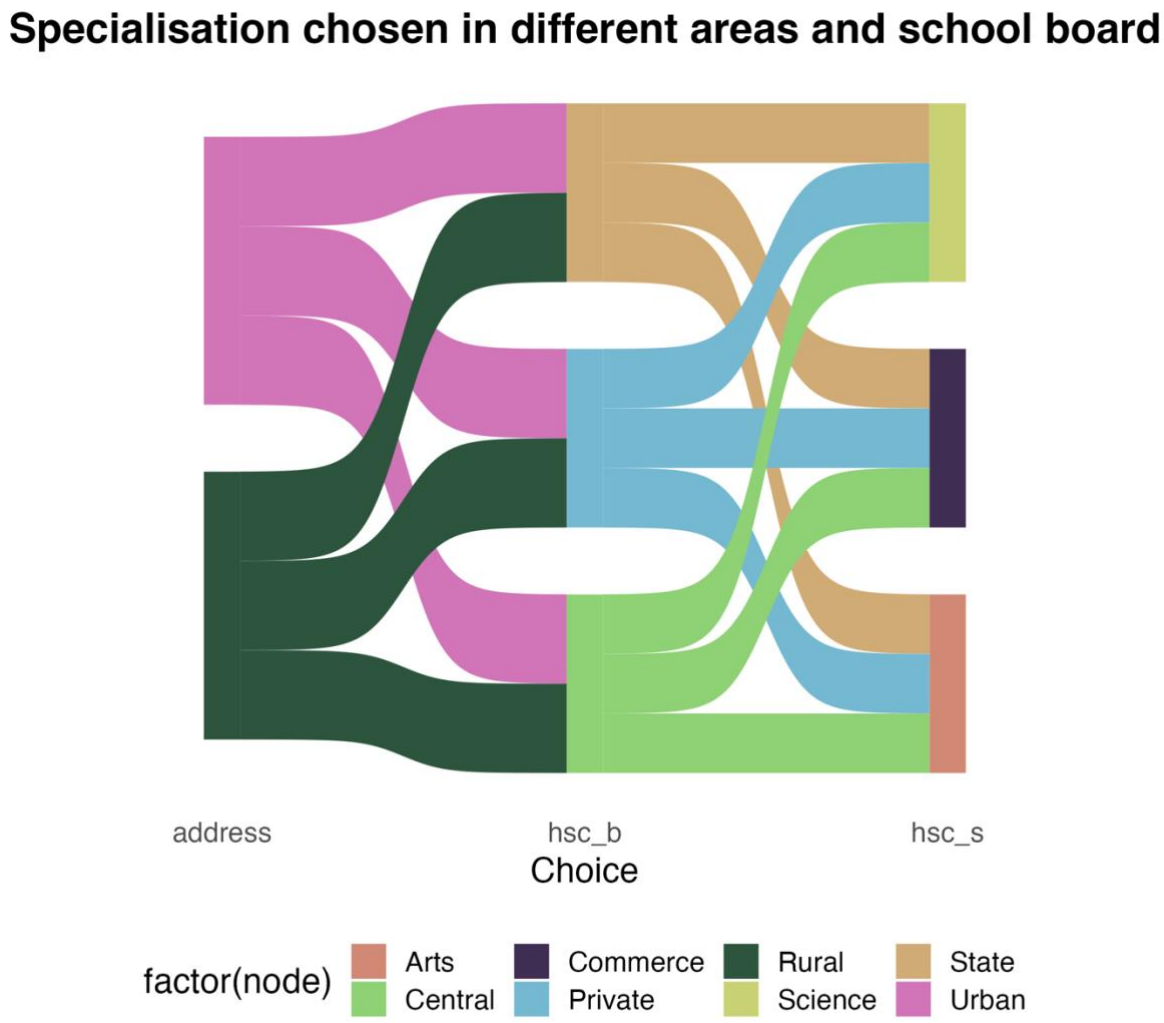


Figure 1.10.2 Relationship of address, school boards and specialization graph

The Sankey diagram above represents the tendency of students from different areas in choosing high school board (HSC_B) and high school specialization (HSC_S).

The HSC_B column consists of three categories: Central, Private, and State. The HSC_S column consists of three categories: Arts, Commerce, and Science. The address column consists of two categories: Rural and Urban.

Looking at the diagram, we can see that the number of students who choose private, central, and state school boards is average. After that, when they choose the specialization for their high school education, there is no specialization that is widely chosen by the students.

There could be several reasons why the number of students in each branch is average. For instance, it could be that students in these areas have different backgrounds and they tend to choose different fields based on their understanding and societal norms.

Analysis 1.11: Do the students from different areas choose different degree types?

```
#====Analysis 1.11: Do the students from different areas choose different degree types?====
dfdeg <- df %>% select(address, degree_t) %>% count(address, degree_t) %>% arrange(address, n)
dfdeg
ggplot(dfdeg, aes(address, n, fill = degree_t)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Degree types chosen in different areas")+ ylab("count")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/1p11.png")
```

Figure 1.11.1 Relationship of address and degree types code

I used the dplyr package in R to create a new data frame dfdeg that selects the address and degree_t columns from the original data frame and counts the number of occurrences for each combination of address and degree_t. The resulting data frame is sorted by address and n columns in ascending order.

Next, I used the ggplot2 package to create a stacked bar plot using ggplot() function. The x-axis shows the different addresses, and the y-axis shows the count of each degree type. The bars are filled with colors based on the different degree types.

I also added a title to the plot using ggtitle() function, labeled the y axis using ylab() function, and positioned the legend at the bottom of the plot using theme() function.

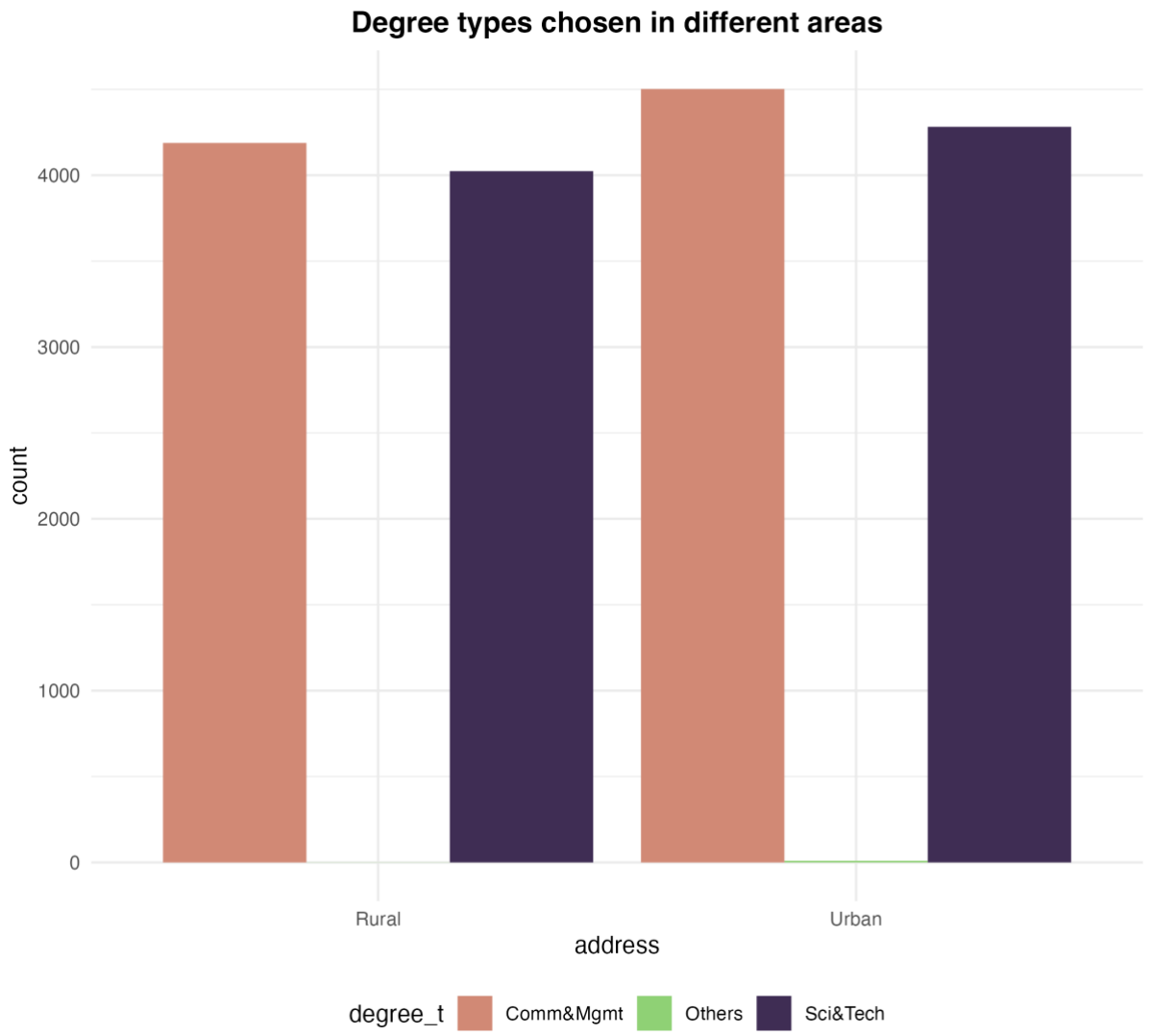


Figure 1.11.2 Relationship of address and degree types graph

This table shows the number of students in different degree categories, categorized by their address (rural or urban). The degree types are "Others", "Sci&Tech", and "Comm&Mgmt".

From the figure, it can be seen that the students in the urban's "Comm&Mgmt" category, with a total of 4500 students, followed by urban's "Sci&Tech" with 4300 students. The "Others" category has very few students in both rural and urban areas.

A possible reason for this tiny difference within hundreds of students could be the demand for jobs in the "Comm&Mgmt" and "Sci&Tech" categories may have the same popularity in both rural and urban areas.

Analysis 1.12: Do the students from different areas choose different specializations?

```
#====Analysis 1.12: Do the students from different areas choose different specializations?====
dfspec <- df %>% select(address, specialisation) %>%
  group_by(address, specialisation) %>% summarize(count = n())
dfspec
chordDiagram(dfspec)+
  title("Specialisation chosen in different areas", cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/1p12.png")
dev.off()
```

Figure 1.12.1 Relationship of address and specialization code

I analyzed the data on the specializations chosen by students from different areas using the 'dfspec' data frame. The 'select' function was used to extract the 'address' and 'specialisation' columns from the original data frame, and the 'group_by' function was used to group the data by address and specialization. The 'summarize' function was then used to count the number of occurrences of each address-specialization pair. The resulting data was then plotted using the 'chordDiagram' function to show the relationships between the different specializations and their respective addresses. Finally, the plot was saved in a low-resolution png format using the 'dev.copy' and 'dev.off' functions.

Figure 1.12.2 Relationship of address and specialization graph

The given data represents the count of students who have chosen different specializations in two areas, Rural and Urban. The specializations provided include Mkt&Fin (Marketing and Financial) and Mkt&HR (Marketing and Human Resources). It appears that the number of students who have chosen these specializations is quite similar across both rural and urban areas.

One possible reason for this could be that in recent times, there has been significant growth in the demand for both marketing and finance-related job roles across various sectors and industries, which might have influenced the choice of students in choosing these specializations. Additionally, there might not be significant differences in the level of exposure and awareness of these specializations across urban and rural areas, which could also explain the similar count of students across both areas.

Summary for Question 1 Analysis

Since we have drawn the graphs and examined the data that is filtered from the dataset, counting the percentage of students that whether they have equal educational resources or admission to their desired schools, there are no attributes that show they are highly related and lead to the inequality of educational resources distribution. From this dataset, whether the students are from rich or normal families, parents are highly-educated or not, and they have the same opportunity in accepting high-quality education.

Question 2: Does family support affect the educational resources that the students get and their educational background?

Analysis 2.1: Does family support directly affect the rural area student's secondary school results?

```
#====Analysis 2.1:  Does family support directly affect the rural area student's secondary school results?====
dfsup <- dfrural %>% select(famsup, ssc_p) %>%
  pivot_longer(cols = c(ssc_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
dfsup
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to secondary school result in rural area")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal()+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p1p1.png")
```

Figure 2.1.1Relationship of family support and secondary school result in rural area code

In this analysis, we first create a new data frame dfsup by selecting the famsup and ssc_p columns from the dfrural data frame. We then use pivot_longer to convert the ssc_p column into the long format, so that it can be plotted on the y-axis. The resulting dfsup data frame is then grouped by famsup.

We then create a scatter plot using ggplot with famsup on the x-axis, percentage on the y-axis, and famsup as the color of the points. We add a title to the plot and adjust the theme settings, including changing the background color to white and placing the legend at the bottom.

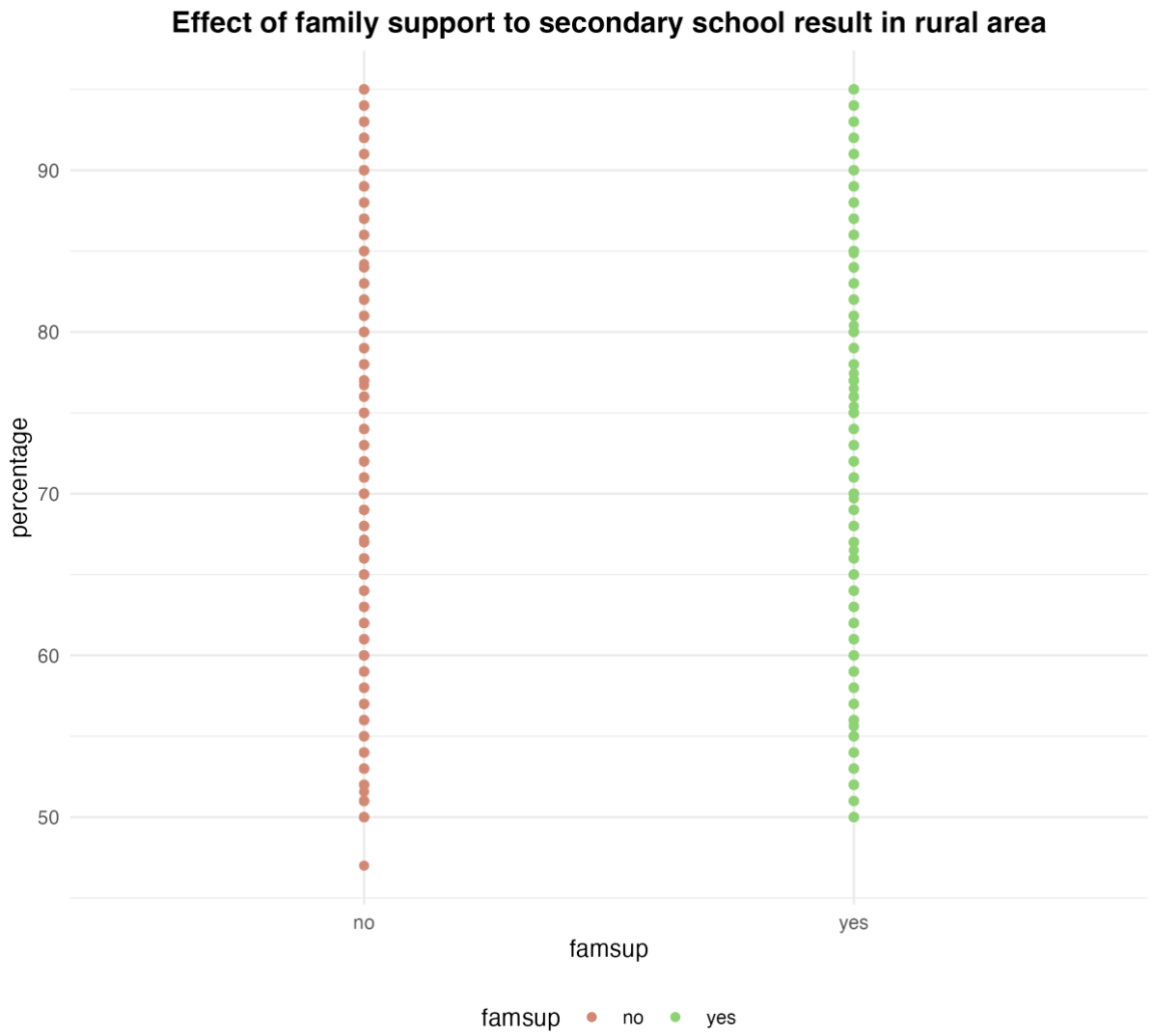


Figure 2.1.2 Relationship of family support and secondary school result in rural area graph

The data in this graph describes the family support (famsup) for students, with a secondary school result (percentage). The table suggests that there is no significant difference in secondary school results whether they get their family support, as the distribution of results for both "yes" and "no" are very similar. One possible reason for this is that the students in this dataset could get any results no matter having family support or not.

Analysis 2.2: Does family support directly affect the urban area student's secondary school results?

```
#====Analysis 2.2:  Does family support directly affect the urban area student's secondary school results?====
dfsup <- dfurban %>% select(famsup, ssc_p) %>%
  pivot_longer(cols = c(ssc_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
dfsup
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to secondary school result in urban area")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p1p2.png")
```

Figure 2.2.1 Relationship of family support and secondary school result in urban area code

In this analysis, we first create a new dataframe dfsup by selecting the famsup and ssc_p columns from the dfrural dataframe. We then use pivot_longer to convert the ssc_p column into long format, so that it can be plotted on the y-axis. The resulting dfsup dataframe is then grouped by famsup.

We then create a scatter plot using ggplot with famsup on the x-axis, percentage on the y-axis, and famsup as the color of the points. We add a title to the plot and adjust the theme settings, including changing the background color to white and placing the legend at the bottom.

The plot shows the relationship between family support and secondary school results for rural area students, where the x-axis represents whether the student received family support (yes or no) and the y-axis represents the percentage of marks obtained in the secondary school examination.

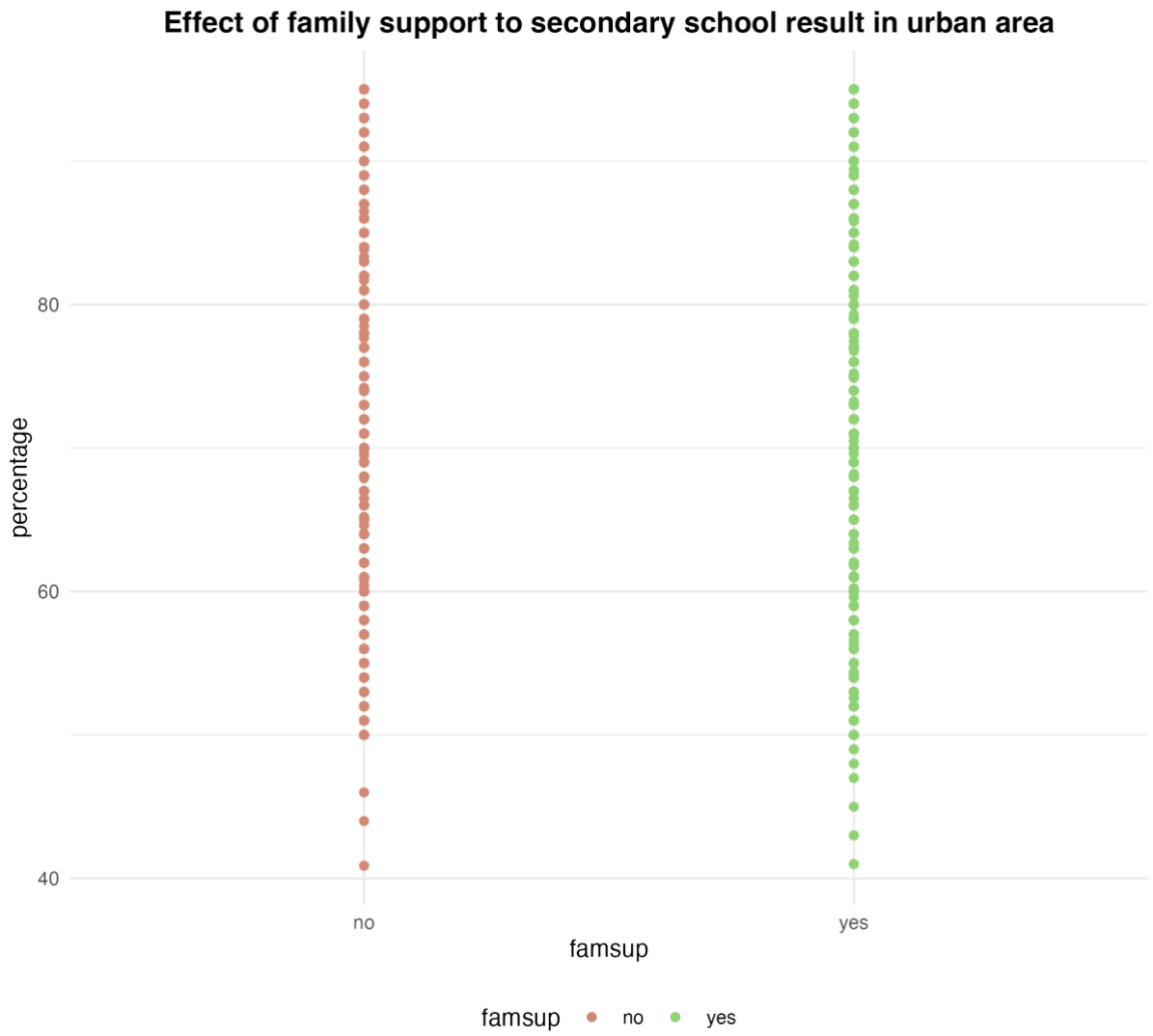


Figure 2.2.2 Relationship of family support and secondary school result in urban area graph

The data for urban area students suggests that there is no significant difference in secondary school results based on whether they receive family support or not, as the distribution of results for both "yes" and "no" are very similar. One possible reason for this could be that family support may not be the only factor affecting a student's academic performance. Other factors such as the student's own motivation, natural ability, and access to resources and support from schools could also play a significant role. Therefore, it is important to consider these factors in addition to family support when trying to understand and improve student performance. Overall, the data suggest that family support may not be as strong of a predictor of academic success as previously thought.

Analysis 2.3: Does family support directly affect the rural area student's high school results?

```
#====Analysis 2.3:  Does family support directly affect the rural area student's high school results?====
dfsup <- dfrural %>% select(famsup, hsc_p) %>%
  pivot_longer(cols = c(hsc_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
dfsup
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to high school result in rural area") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p2p1.png")
```

Figure 2.3.1 Relationship of family support and high school result in rural area code

We start by selecting the 'famsup' and 'hsc_p' columns from the 'dfrural' dataset. We then use the pivot_longer function to transform the 'hsc_p' column from wide format to long format. This is done to create a 'level' column and a 'percentage' column. The 'level' column indicates the level of education (in this case, 'hsc_p' indicates high school percentage), and the 'percentage' column contains the actual percentage values.

We then group the data by 'famsup' column. Finally, we create a scatter plot using ggplot() function where 'famsup' is represented on the x-axis, 'percentage' is represented on the y-axis, and 'famsup' is used as the color. The title of the plot is "Effect of family support to high school results in rural areas". We use the scale_color_ipsum() and scale_fill_ipsum() functions to set the color scheme for the plot. Finally, we use the theme_minimal() function to set the theme of the plot, and add a few more elements to the theme using the theme()

function.

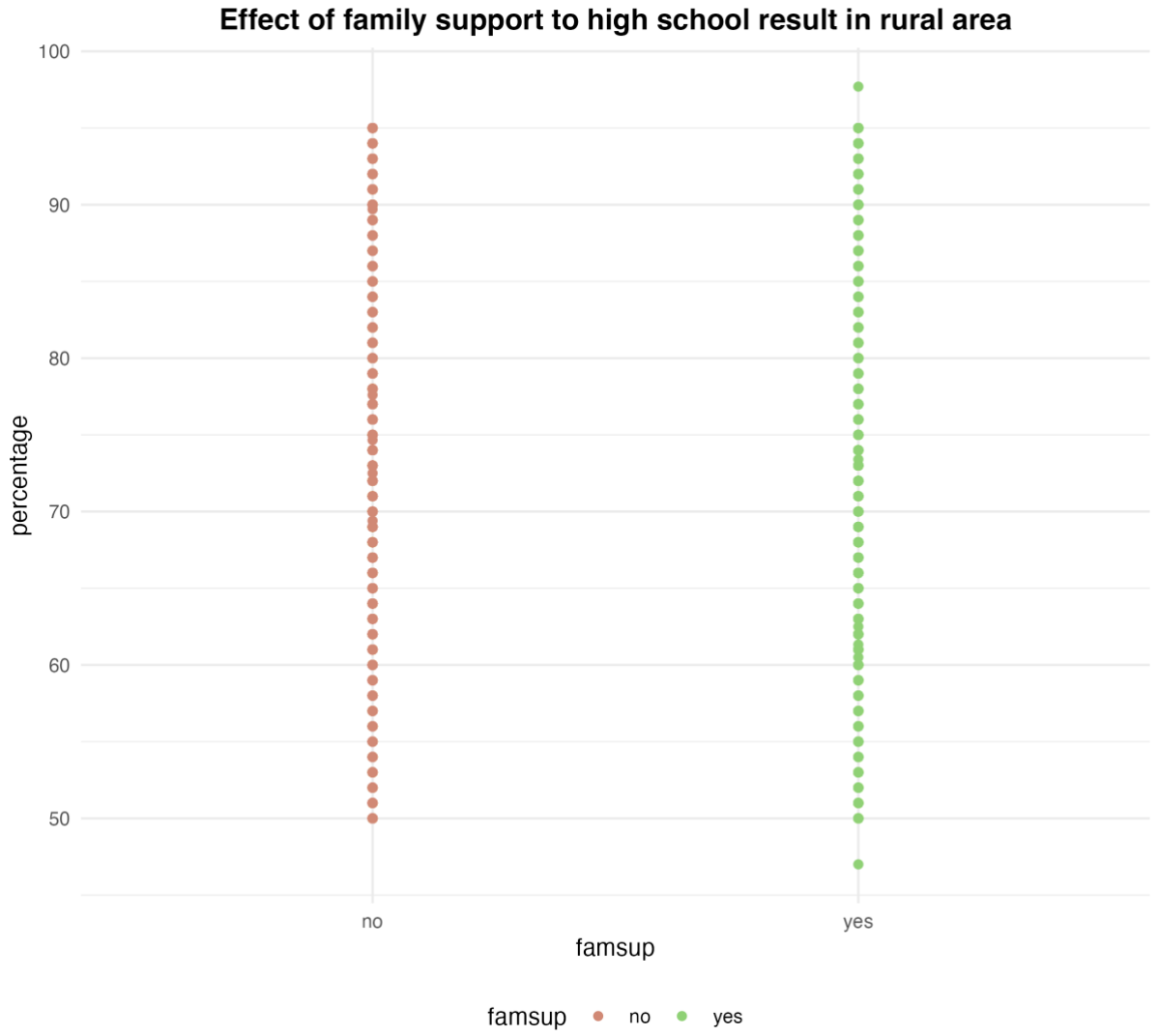


Figure 2.3.2 Relationship of family support and high school result in rural area graph

The data table shows that there is no significant difference in high school results for students who receive family support (famsup) compared to those who do not. This suggests that family support may not be a determining factor for academic success in high school. One possible reason for this could be that factors such as student motivation, study habits, and natural ability play a greater role in determining academic success than family support alone. Further research may be needed to explore the complex relationship between family support and academic achievement in high school. In summary, the data suggest that family support may not be a significant factor in high school academic success.

Analysis 2.4: Does family support directly affect the urban area student's high school results?

```
#====Analysis 2.4:  Does family support directly affect the urban area student's high school results?====
dfsup <- dfurban %>% select(famsup, hsc_p) %>%
  pivot_longer(cols = c(hsc_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to high school result in urban area")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p2p2.png")
```

Figure 2.4.1 Relationship of family support and high school result in urban area code

This code uses the dfurban data frame to analyze whether family support directly affects urban area students' high school results. The code selects the famsup and hsc_p columns from the dfurban data frame using select(). Then, it uses pivot_longer() to reshape the data from wide to a long format, making it easier to plot. The function group_by() groups the data by the famsup variable.

The resulting data frame is plotted using ggplot(). The x variable is set to famsup, the y variable is set to percentage, and the color variable is set to famsup. The plot shows the relationship between family support and high school results for urban area students. Finally, ggtitle(), scale_color_ipsum(), scale_fill_ipsum(), theme_minimal(), legend.position, plot.title, and plot.background is used to customize the plot's title, color scheme, and background.

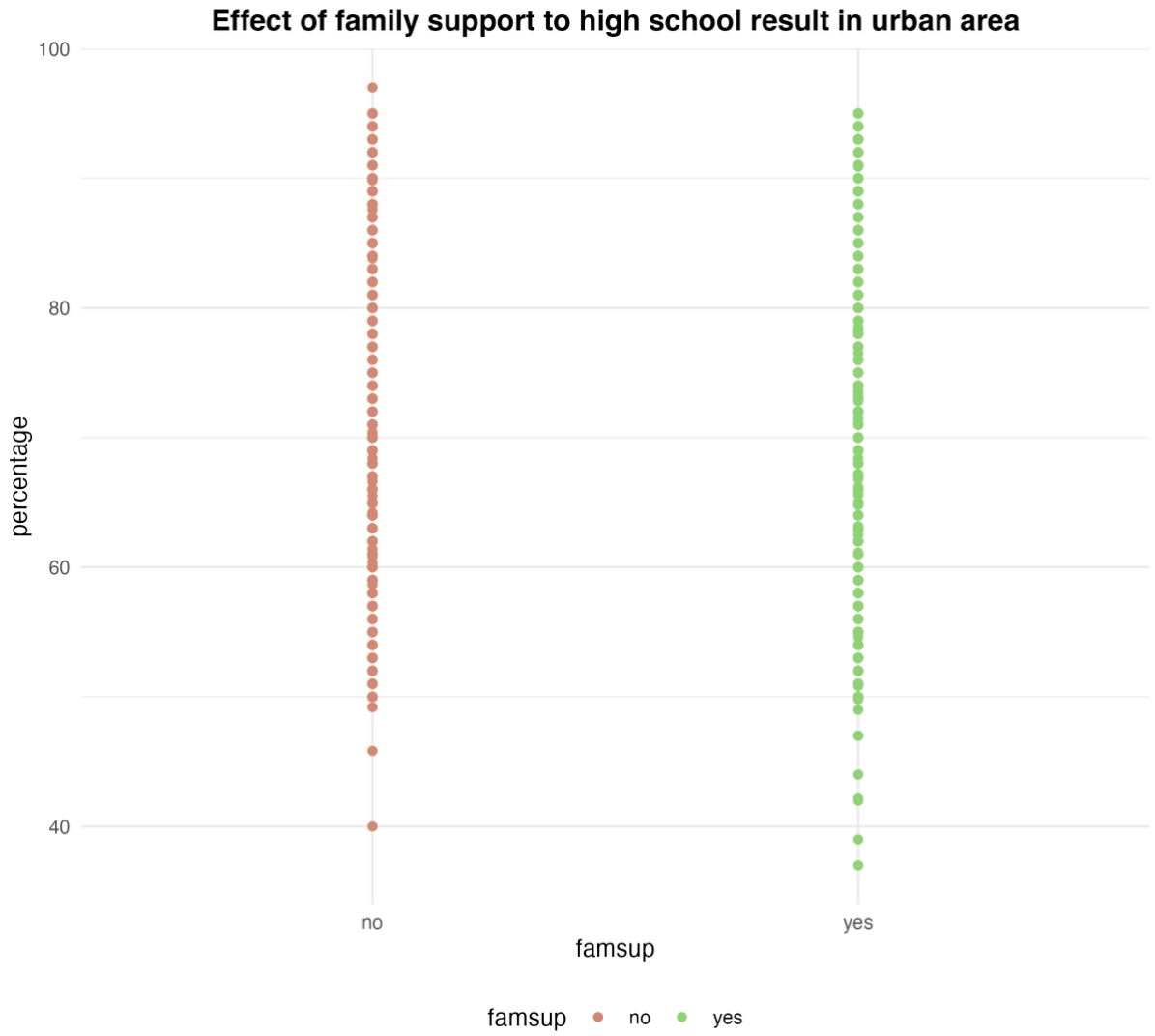


Figure 2.4.2 Relationship of family support and high school result in urban area graph

If we consider the same data for high school results from urban area students, we can see that there is no significant difference in the distribution results of students who receive family support and those who do not. This suggests that family support is not a significant factor in determining high school results for urban area students. One possible reason for this could be that other factors such as individual effort, quality of education, and access to resources may have a greater impact on high school results for urban area students. In summary, the data suggest that family support may not be a significant factor in determining high school results for both rural and urban area students.

Analysis 2.5: Does family support directly affect the rural area student's degree results?

```
#====Analysis 2.5:  Does family support directly affect the rural area student's degree results?====
dfsup <- dfrural %>% select(famsup, degree_p) %>%
  pivot_longer(cols = c(degree_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to degree result in rural area")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p3p1.png")
```

Figure 2.5.1 Relationship of family support and degree result in rural area code

In this analysis, we are investigating whether family support has a direct impact on rural students' degree results. To do this, we first select the 'famsup' and 'degree_p' columns from the 'dfrural' data frame. Then, we use the 'pivot_longer' function to transform the 'degree_p' column into a long format so that we can plot it later. After that, we group the data by 'famsup' using the 'group_by' function.

Finally, we create a scatter plot using 'ggplot' function and plot the 'percentage' column against 'famsup'. We color-code the plot using 'famsup' column and add a title to the plot. We also use the 'scale_color_ipsum' and 'scale_fill_ipsum' functions to set the color palette and use the 'theme_minimal' function to adjust the plot's appearance.

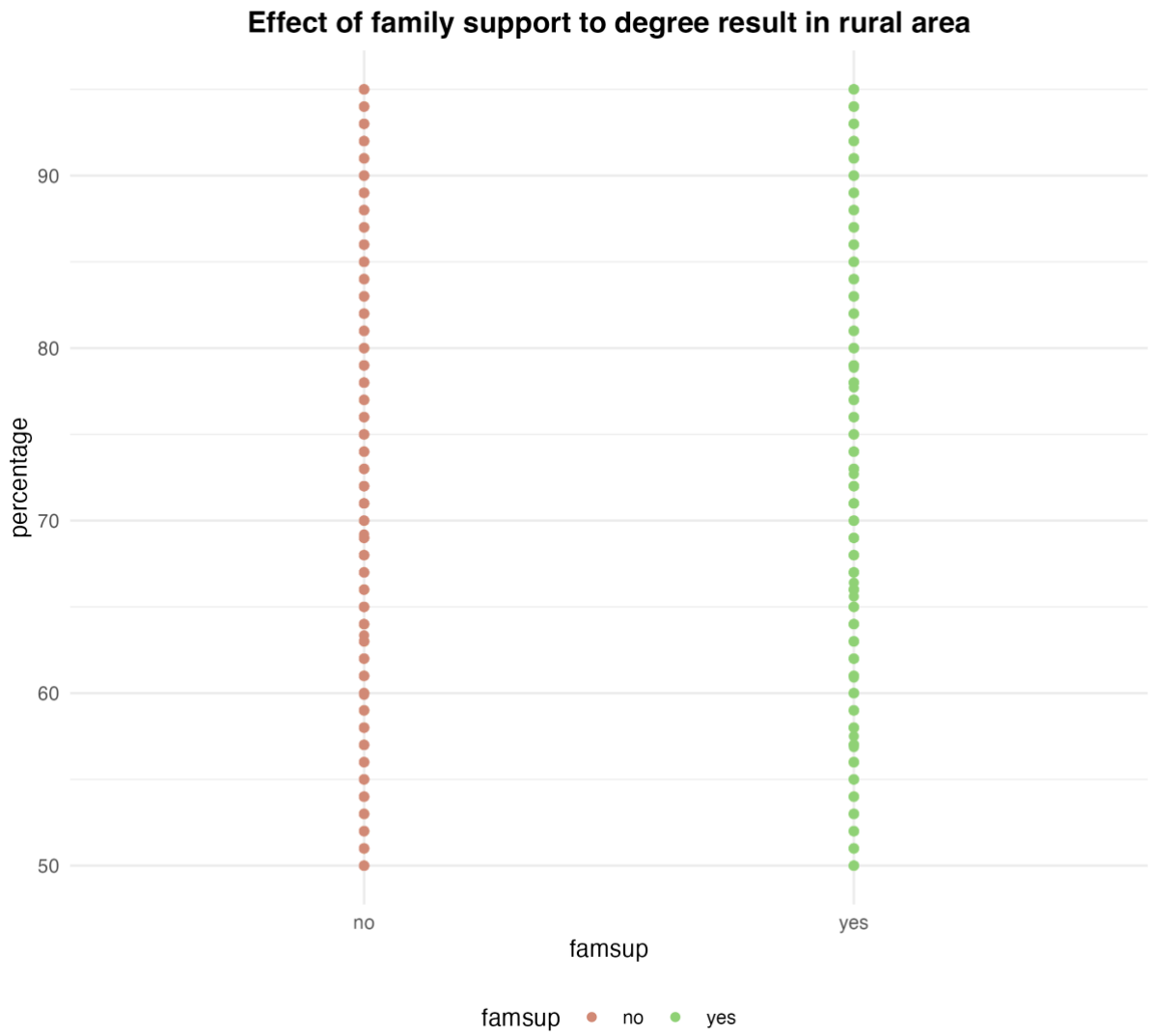


Figure 2.5.2 Relationship of family support and degree result in rural area graph

The distribution diagram shows that the distribution of students from rural areas with family support (yes) or without (no) is almost the same for degree results. This suggests that family support may not have a significant impact on the degree result for students in rural areas. One possible reason for this is that in rural areas, students may rely less on family support and more on their own efforts and resources to achieve their academic goals.

In summary, the data suggest that family support may not have a significant impact on academic performance for students in rural areas, at least in terms of their degree results. Other factors such as personal motivation, access to resources, and quality of education may play a more critical role.

Analysis 2.6: Does family support directly affect the urban area student's degree school results?

```
#====Analysis 2.6:  Does family support directly affect the urban area student's degree school results?====
dfsup <- dfurban %>% select(famsup, degree_p) %>%
  pivot_longer(cols = c(degree_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to degree result in urban area")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p3p2.png")
```

Figure 2.6.1 Relationship of family support and degree result in urban area code

We are selecting the columns "famsup" and "degree_p" from the "dfurban" data frame using the select function. Then, we are using the pivot_longer function to convert the "degree_p" column from wide to long format, creating a new column called "level" to store the original column name and a new column called "percentage" to store the corresponding values. We then group the resulting data frame by the "famsup" column.

We then create a scatter plot using the ggplot2 library, where the x-axis represents the "famsup" variable, and the y-axis represents the "percentage" variable. We use the geom_point function to plot the data points, and color the points based on the "famsup" variable. We add a title to the plot using the ggtitle function. We also adjust the color and fill scale using scale_color_ipsum() and scale_fill_ipsum() functions, respectively. Finally, we adjust the theme of the plot using theme_minimal() and customize the legend, title, and background using the theme() function.

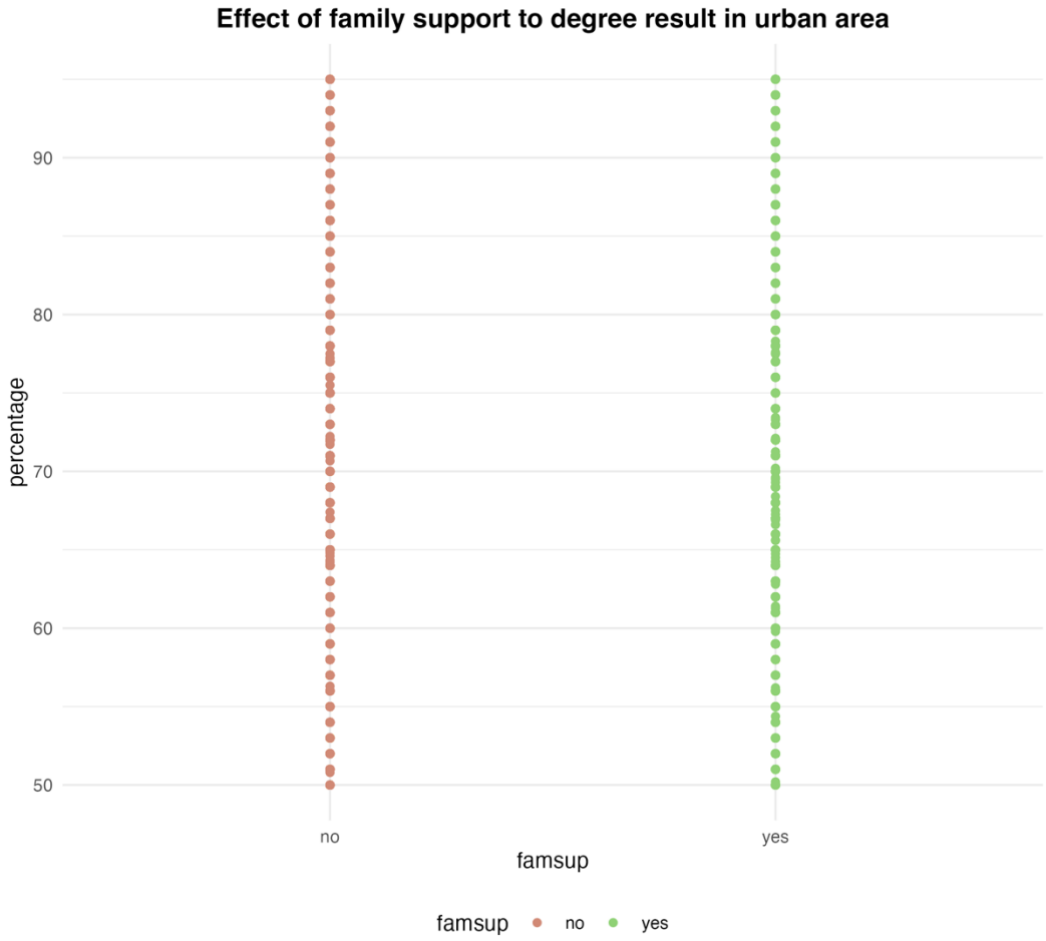


Figure 2.6.2 Relationship of family support and degree result in urban area graph

The data for urban area students also shows a similar pattern in the distribution of degree results between those who have family support and those who do not. There is no significant difference in the distribution of degree results between the two groups, indicating that family support does not play a significant role in determining the degree result for urban area students.

A possible reason for this similarity in the distribution of degree results could be that other factors, such as academic ability and motivation, have a greater influence on a student's performance in university. Additionally, urban areas may have more opportunities and resources available to students, such as access to better schools and libraries, which could contribute to a more even distribution of degree results between those with and without family support.

In summary, the data suggest that for both rural and urban area students, family support does not appear to significantly impact their degree results, indicating that other factors may be more influential in determining academic performance.

Analysis 2.7: Does family support directly affect the rural area student's MBA results?

```
#====Analysis 2.7:  Does family support directly affect the rural area student's MBA results?====
dfsup <- dfrural %>% select(famsup, mba_p) %>%
  pivot_longer(cols = c(mba_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) +
  geom_point() + theme(legend.position = "bottom", plot.title = element_text(hjust = 0.5)) +
  ggtitle("Effect of family support to MBA result in rural areas")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p4p1.png")
```

Figure 2.7.1 Relationship of family support and MBA result in rural area code

In this analysis, we are examining whether family support directly affects the MBA results of students living in rural areas. We start by selecting the "famsup" and "mba_p" columns from the "dfrural" dataset. Then, we use the "pivot_longer()" function to convert the "mba_p" column from wide format to long format. We group the resulting data by "famsup" using the "group_by()" function.

Next, we create a scatter plot using the "ggplot()" function, where we plot the "percentage" on the y-axis and "famsup" on the x-axis, with each family support category represented by a different color. We set the title of the plot using the "ggtitle()" function and adjust the legend and plot appearance using various "theme" functions. Finally, we use the "scale_color_ipsum()" and "scale_fill_ipsum()" functions to use a color palette consistent with the "ggplot2" package.

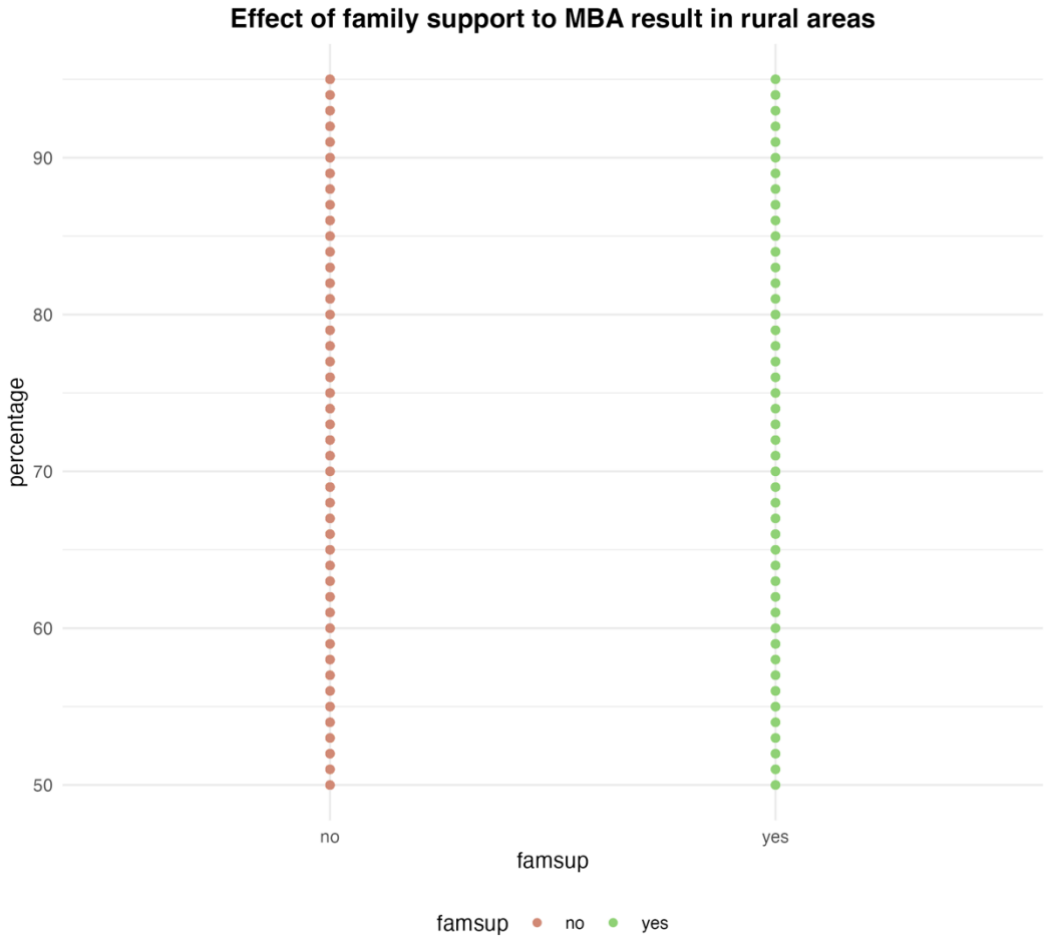


Figure 2.7.2 Relationship of family support and MBA result in rural area graph

The distribution of students from rural areas with and without family support (famsup) for their MBA result is similar, with no concentration in any particular range. This suggests that there is no significant difference in MBA results for students from rural areas who receive family support and those who do not. One possible reason for this could be that the students in this dataset are self-motivated and independent, and their family support has little impact on their academic performance.

It is important to note that these conclusions are based solely on the data presented in the graph, and further research would be necessary to fully understand the relationship between family support and academic performance for students in rural areas. Nonetheless, the data suggest that family support may not have a significant impact on the MBA results of rural area students.

Analysis 2.8: Does family support directly affect the urban area student's MBA results?

```
#====Analysis 2.8:  Does family support directly affect the urban area student's MBA results?====
dfsup <- dfurban %>% select(famsup, mba_p) %>%
  pivot_longer(cols = c(mba_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to MBA result in urban area")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p4p2.png")
```

Figure 2.8.1 Relationship of family support and MBA result in urban area code

We start by selecting the famsup and mba_p columns from the dfurban data frame. Then, we use the pivot_longer function to convert the mba_p column into a long format with two columns - level and percentage. We group the data by famsup using the group_by function.

Next, we create a scatter plot using ggplot to visualize the relationship between famsup and percentage. We set the x axis to famsup, y-axis to percentage, and the color to famsup. We also set the plot title to "Effect of family support to MBA result in urban area". Finally, we customize the plot using scale_color_ipsum, scale_fill_ipsum, theme_minimal,

legend.position, plot.title, and plot.background functions.

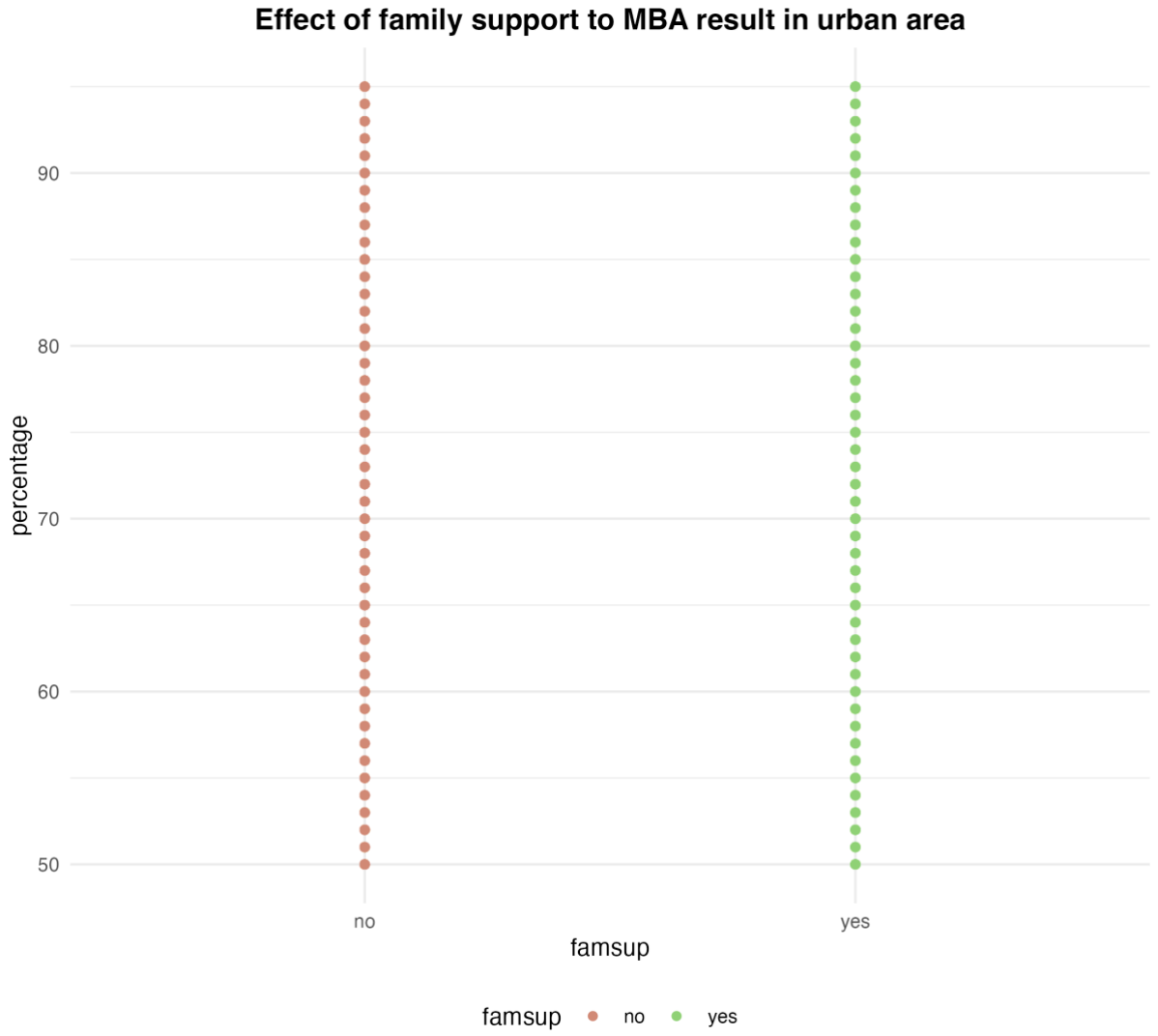


Figure 2.8.2 Relationship of family support and MBA result in urban area graph

The MBA result for urban students also shows a similar situation where the distribution of degree results is the same for both "yes" and "no" family support. This suggests that family support does not have a significant impact on MBA results for urban students. A possible reason for this could be that factors other than family support, such as personal motivation and access to resources, may play a more significant role in determining MBA outcomes for urban students.

Analysis 2.9: Does family support have gained effect on employability test results for students in rural areas?

```
#====Analysis 2.9:  Does family support have gained effect on employability test results for students in rural areas?====
dfsup <- dfrural %>% select(famsup, etest_p) %>%
  pivot_longer(cols = c(etest_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to employability test result in rural area")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p5p1.png")
```

Figure 2.9.1 Relationship of family support and employability test result in rural area code

In this analysis, we are investigating whether family support has a significant effect on employability test results for students in rural areas. We first select the columns "famsup" and "etest_p" from the "dfrural" data frame using the select() function. We then use the pivot_longer() function to convert the wide format of the "etest_p" variable into a long format with the column "level" indicating the variable name and "percentage" indicating the variable value. We group the resulting data frame by "famsup".

We then create a scatter plot using ggplot() and aes() functions to plot the "percentage" variable on the y-axis against the "famsup" variable on the x-axis. We use the geom_point() function to add points to the plot and the ggtitle() function to add a title to the plot. We also adjust the legend, plot title, and plot background using the scale_color_ipsum(), scale_fill_ipsum(), and theme() functions.

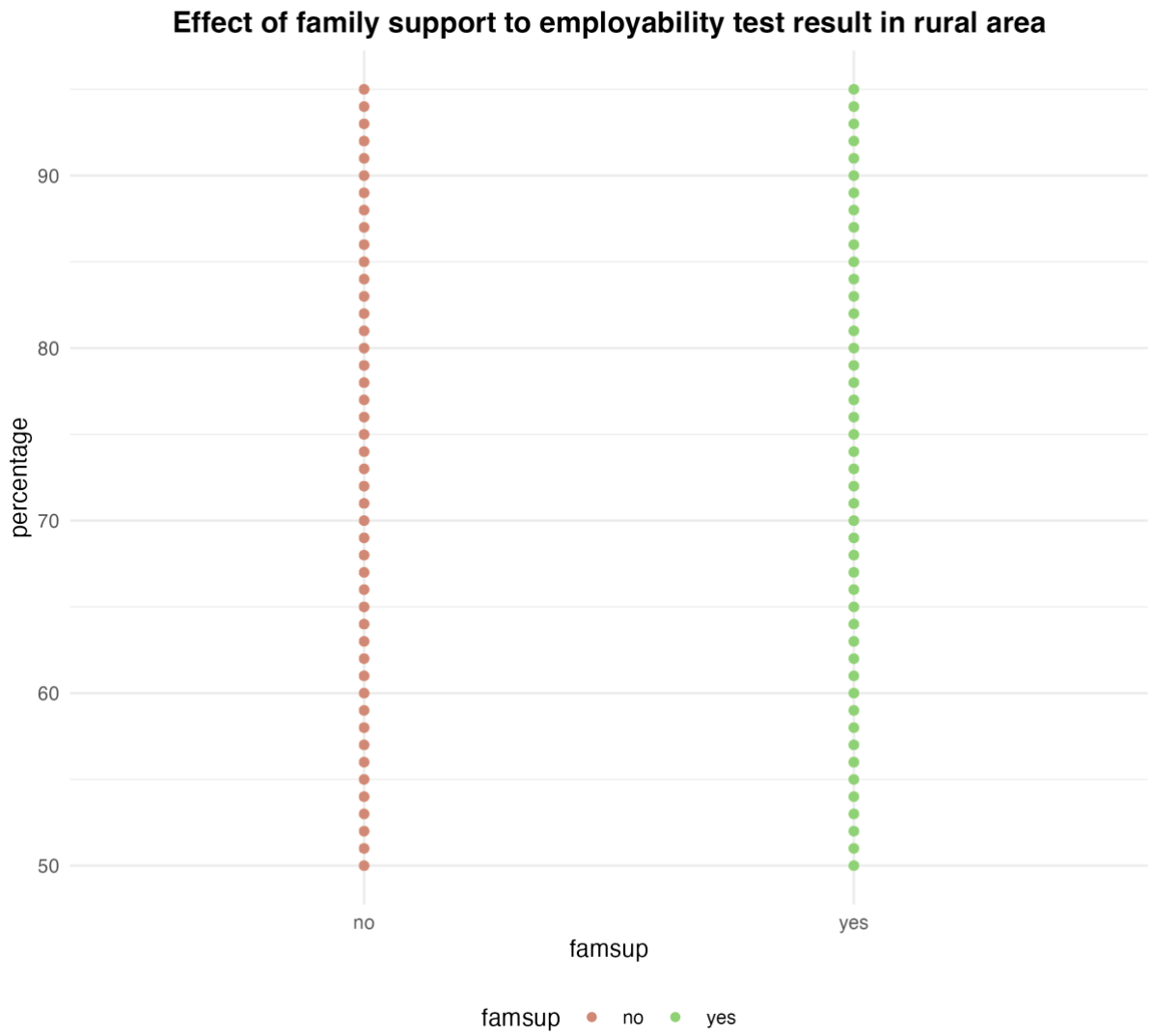


Figure 2.9.2 Relationship of family support and employability test result in rural area graph

The employability test results for rural area students indicate that receiving family support (famsup "yes") or not (famsup "no") does not have a significant impact on the distribution of results. This finding suggests that other factors may play a more important role in determining a student's employability, such as the quality of education and opportunities available to them. It's possible that students in rural areas have limited access to employment opportunities, which can impact their chances of securing a job regardless of family support. Overall, the data implies that the relationship between family support and employability for rural area students may not be as strong as previously thought.

Analysis 2.10: Does family support have gained effect on employability test results for students in urban areas?

```
#====Analysis 2.10: Does family support have gained effect on employability test results for students in urban areas?====
dfsup <- dfurban %>% select(famsup, etest_p) %>%
  pivot_longer(cols = c(etest_p), names_to = "level", values_to = "percentage") %>%
  group_by(famsup)
ggplot(dfsup, aes(x = famsup, y=percentage, colour=famsup)) + geom_point() +
  ggtitle("Effect of family support to employability test result in urban area")+
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p5p2.png")
```

Figure 2.10.1 Relationship of family support and employability test result in urban area code

The dfsup data frame is created by selecting the famsup and etest_p columns from the dfurban data frame, which contains the urban area student data. The pivot_longer() function is used to reshape the data so that the etest_p values are in a single column named percentage. The group_by() function is used to group the data by famsup.

The ggplot() function is used to create a scatter plot of percentage values against famsup values. The geom_point() function adds points to the plot. The ggtitle() function sets the plot title to "Effect of family support to employability test result in urban area". The scale_color_ipsum() and scale_fill_ipsum() functions are used to set the color and fill schemes for the plot. The theme_minimal() function sets the plot theme to a minimal style. Finally, the theme() function is used to customize the legend, plot title, and background.

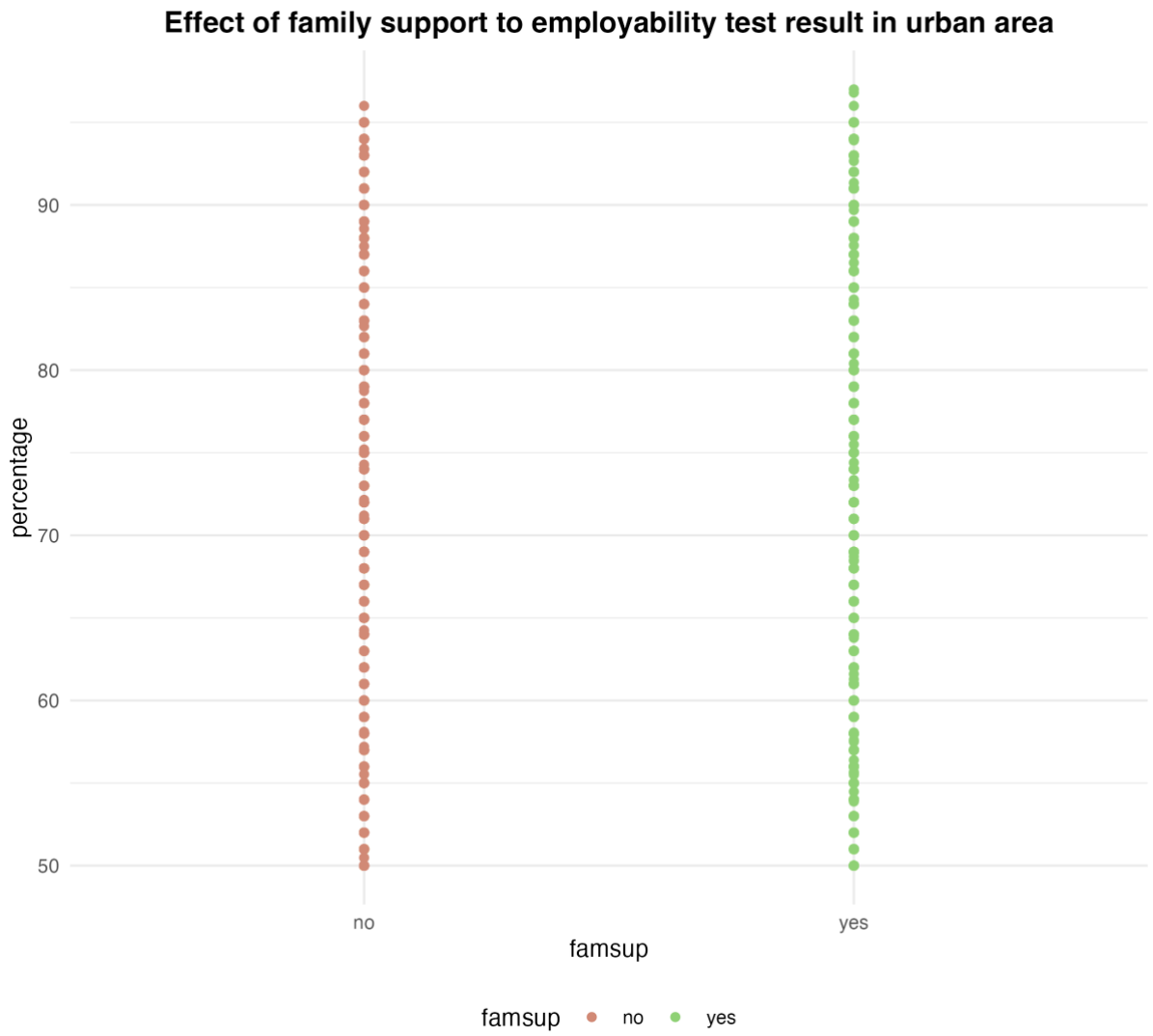


Figure 2.10.2 Relationship of family support and employability test result in urban area graph

From the diagram, the distribution for students who get family support has the same probability of getting any result. This means that students who receive more family support do not perform better on employability tests than those who receive less family support.

One possible reason for this could be that employability tests are more heavily influenced by external factors such as education, experience, and job-specific skills, which may not necessarily be related to family support. Therefore, while family support may play a role in overall success and well-being, it may not directly impact employability test results.

In summary, a lack of effect of family support on employability test results in urban areas may result in a similar distribution of scores across all levels of family support. This could be due to the fact that employability tests are more influenced by external factors such as education and experience, rather than family support alone.

Summary 2.10 The effect of family support to academic and employability test result in different areas

```
#====Summary 2.10 The effect of family support to academic and employability test result in different areas====
dfsup <- df %>% select(address, famsup, ssc_p, hsc_p, degree_p, mba_p, etest_p) %>%
  pivot_longer(cols = c(ssc_p, hsc_p, degree_p, mba_p), names_to = "level",
               values_to = "percentage") %>% group_by(address, famsup)
dfsup
ggplot(dfsup, aes(x = etest_p, y=percentage, colour=level)) + geom_point() +
  facet_grid(address~famsup) +
  ggtitle("Effect of family support to academic and
 employability test result in different areas")+
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p5Sum1.png")
```

Figure Summary 0.1 Summary 2.10 code

We start by creating a new data frame dfsup which includes the columns address, famsup, and the percentages of ssc_p, hsc_p, degree_p, mba_p, and etest_p. We then pivot the data frame to make the percentages into a single column named percentage and the columns ssc_p, hsc_p, degree_p, and mba_p into a new column named level. We group the data frame by address and famsup.

Next, we create a scatter plot using ggplot, where the x-axis represents etest_p, the y-axis represents percentage, and the points are colored by the different levels. We use facet_grid to create a grid of plots based on the address and famsup variables. Finally, we add a title and some styling to the plot using theme_minimal (), scale_color_ipsum(), scale_fill_ipsum(), and theme(). We save the plot using ggsave().

Overall, the code allows us to visualize the effect of family support on academic and employability test results for students in different areas.

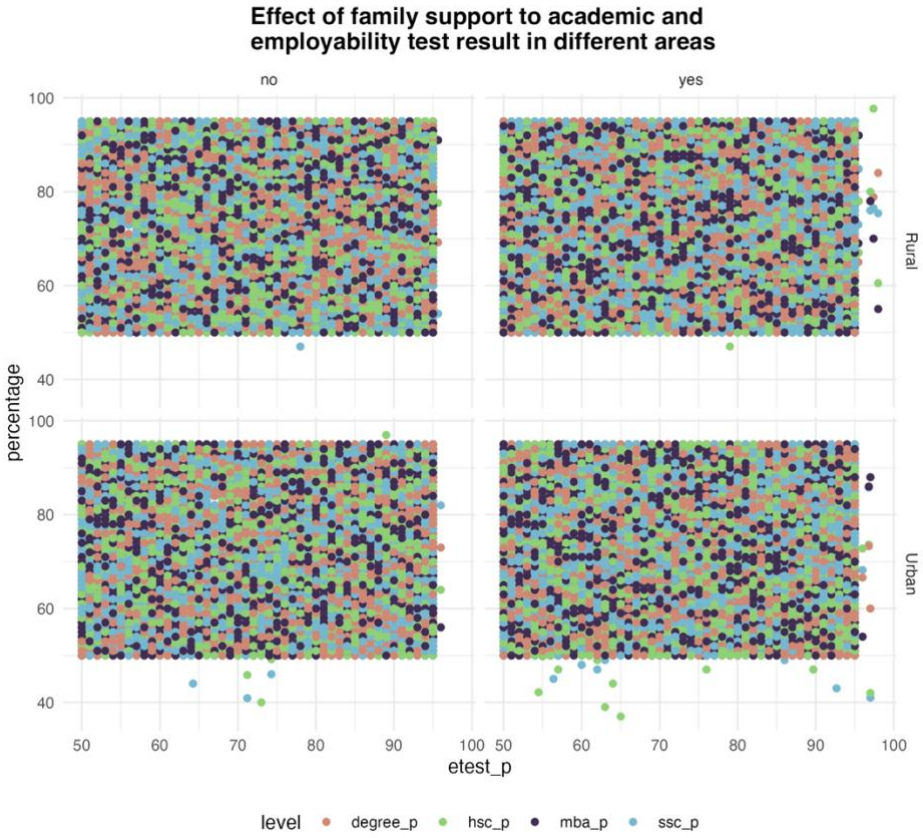


Figure Summary 0.2 Summary 2.10 graph

The data presented in the graph suggests that there is no significant difference in academic performance (measured by secondary school results, degree results, MBA results, and employability test results) for students who receive family support (famsup "yes") compared to those who do not (famsup "no"). This similarity in the distribution of results suggests that family support may not be the determining factor for academic success or employability for students. Other factors, such as individual effort, natural ability, motivation, access to resources, and quality of education, may play a more critical role. It is important to consider these factors in addition to family support when trying to understand and improve student performance and employability. Additionally, the data suggest that family support may not significantly impact academic performance or employability for both rural and urban area students. However, further research may be necessary to fully understand the complex relationship between family support and academic performance/employability.

So, we come up to take the top students who have gotten High Distinction (HD) results in all tests from the dataset and draw the line chart for their results to check whether any trend among the attributes when it is applied to the top students only.

Analysis 2.11: Do all the results affect by family support when only top students are observed?

```
#====Analysis 2.11: Do all the results affect by family support when only top students are observed?====
dfsummary <- df %>% filter(ssc_level == "HD" & hsc_level == "HD" &
                           degree_level == "HD" & mba_level == "HD" & etest_level == "HD") %>%
  select(famsup, ssc_p, hsc_p, degree_p, mba_p, etest_p) %>% arrange(.)
ggparcoord(dfsummary, columns = c(2,3,4,5,6), groupColumn = 1, showPoints = T)+
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of family support to academic and
 employability test result in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p5Sum2.png")
```

Figure 2.11.1 Relationship between family support and top students code

In this analysis, we filtered out the top students from the dataset based on their academic performance levels in secondary school certificate (ssc), higher secondary certificate (hsc), undergraduate degree (degree), Master of Business Administration (mba), and employability test (etest). We then selected the columns that represent the students' family support and academic and employability test results.

Next, we created a parallel coordinate plot using ggparcoord function from GGally package to visualize the relationship between family support and academic and employability test results for top-performing students. We used the "groupColumn" parameter to group the plot by family support and used "showPoints" parameter to display data points in the plot.

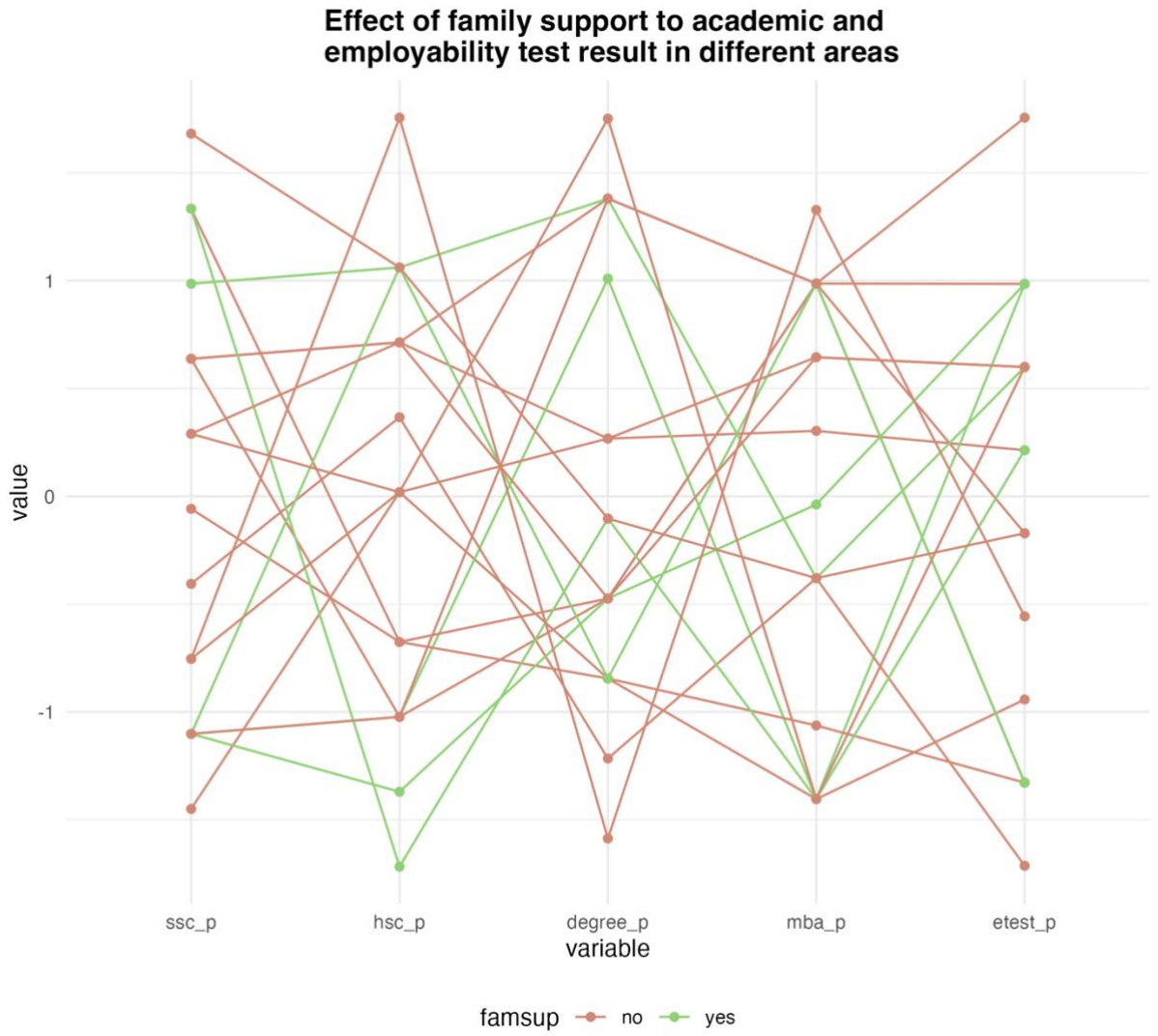


Figure 2.11.2 Relationship between family support and top students' graph

As the line chart above, the results of top students in the 5 tests all fluctuated. This means that even if they are top students with all High Distinction results on all tests. No matter they are supported by their family in education or not, they are surely having ups and down in all the exams.

Analysis 2.12: Do family support determines the difference for number of students in accessing to internet, attending extra-paid classes and extra-curricular activities?

```
#====Analysis 2.12: Do family support determines the difference for number of students in accessing to internet,
#attending extra-paid classes and extra-curricular activities?====
dfsup <- df %>% select(famsup, internet, activities, paid) %>%
  pivot_longer(cols = c(internet, activities, paid), names_to = "condition",
               values_to = "binary") %>% count(famsup, condition, binary)
dfsup
ggplot(dfsup, aes(binary, n,group=condition, color=condition)) + geom_point() +
  geom_line()+ facet_wrap(~famsup)+
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of family support to internet access
 extra-curricular activities and extra-paid classes")+ylab("count")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/2p6p1.png")
```

Figure 2.12.1 Relationship between family support and the number of students accessing to the internet, attending extra-paid classes, and extra-curricular activities code

In this analysis, we are exploring the relationship between family support and whether or not students have access to the internet, attend extra-curricular activities, and attend extra-paid classes. We start by selecting the relevant columns from the data frame and pivoting them into a long format. We then count the number of occurrences of each combination of family support, condition (i.e., internet access, extra-curricular activities, and extra-paid classes), and binary value (i.e., yes or no).

After that, we use ggplot to create a line chart that displays the count of each combination for each level of family support. The x-axis displays the binary value (yes or no), the y-axis displays the count, and the lines are colored by the condition. Finally, we save the resulting graph.

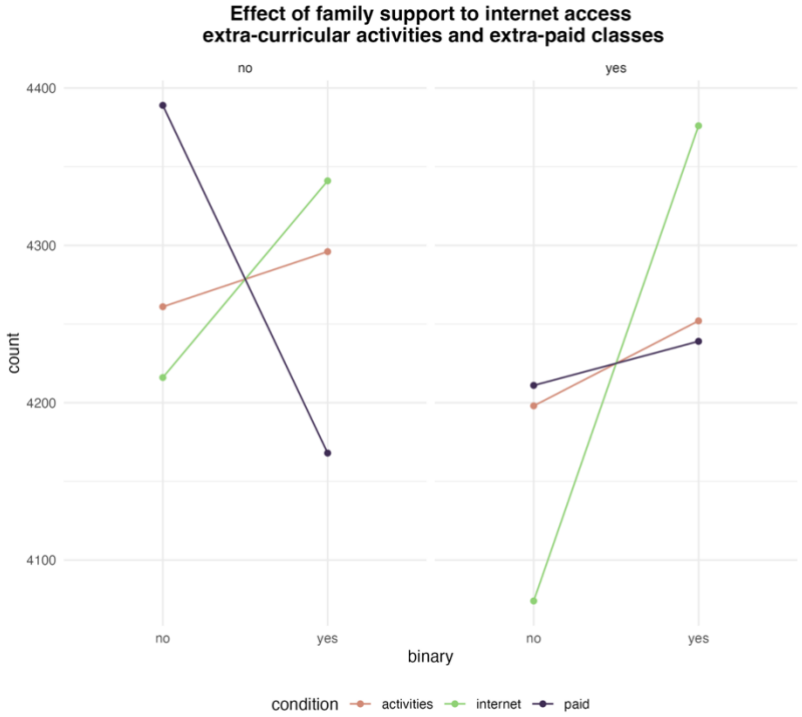


Figure 2.12.2 Relationship between family support and number of students in accessing to internet, attending extra-paid classes, and extra-curricular activities graph

The diagram provides information about students' family support and their involvement in extracurricular activities, extra-paid classes, and internet access. It shows the difference of students who have family support and those who do not have family support, and for each group, it shows the number of students who participate in each activity and those who do not. Then connect the ‘yes’ and ‘no’ numbers with a line to check their difference in the number

From the graph, we can observe that the number of students who participate in extracurricular activities, extra-paid classes, and internet access is similar for both groups. Although it seems a big difference in the diagram, the difference in number is only 500 students maximum in the dataset.

There could be several reasons for these observations. One possible reason is that the school might provide equal opportunities and resources for all students, regardless of their family background, to ensure a level playing field. Another reason could be that the students themselves, with or without family support, have similar interests and motivations to participate in these activities.

Overall, this table provides valuable information about the student's participation in extracurricular activities, extra-paid classes, and internet access, which can help identify any disparities and address them accordingly.

Analysis 2.13: Does the father's occupation play a critical role in the children's secondary school results in rural areas?

```
#====Analysis 2.13: Does the father's occupation play a critical role in the children's secondary school results in rural areas?====
dfjob <- dfrural %>% select(address, Fjob, ssc_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=ssc_p, color=Fjob)) + geom_point() +
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to secondary school result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p7.png", plot=g,width=8,height=8,dpi=300)
```

Figure 2.13.1 Relationship between father's job and secondary school result in rural areas code

In this analysis, we are exploring the relationship between a father's occupation and secondary school results for students living in rural areas. To begin, we select the relevant variables from the original dataframe and group the data by address and father's occupation using the group_by function. We then create a new data frame called dfjob.

Next, we create a scatter plot using the ggplot function and pass in dfjob as the data argument. We set Fjob as the x-axis variable, ssc_p as the y-axis variable, and Fjob as the color variable to differentiate between different occupations. We add a minimal theme and a title to the plot using the theme_minimal and ggtitle functions, respectively. We also set the legend to be displayed at the bottom and center the plot title using the theme function.

To visualize the distribution of the ssc_p variable, we add marginal density plots to the sides of the main scatter plot using the ggMarginal function. We set the type argument to "density" to display the density plots and margins to 'y' to display the density plots on the y-axis. Finally, we set the color of the density plots to purple and the size of the lines to 4 using the color and size arguments, respectively.

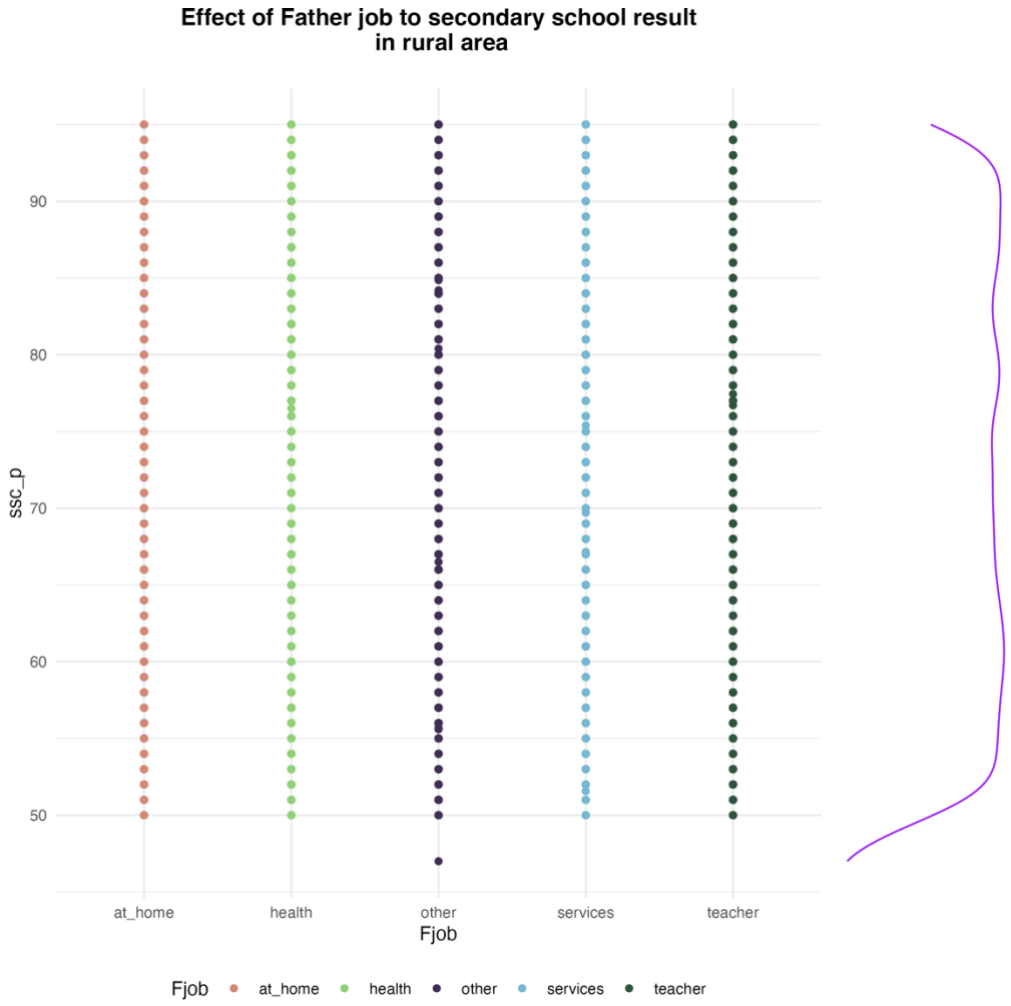


Figure 2.13.2 Relationship between father's job and secondary school result in rural areas graph

The resulting diagram with the distribution of students' secondary school results shows no relationship with their fathers in rural areas, which means that there is no significant correlation between these two variables. This suggests that the father's occupation may not be a critical factor in determining children's academic performance in rural areas.

One possible reason for this could be those other factors such as parental involvement, access to educational resources, and quality of teaching may have a greater impact on children's academic performance in rural areas than on fathers' occupation alone. Additionally, it is possible that the sample size or the study design did not allow for detecting a significant relationship between the variables.

It is important to note that further research may be needed to fully understand the relationship between fathers' occupations and children's academic performance in rural areas.

Analysis 2.14: Does the father's occupation play a critical role in the children's secondary school results in urban areas?

```
#====Analysis 2.14: Does the father's occupation play a critical role in the children's secondary school results in urban areas?====
dfjob <- dfurban %>% select(address, Fjob, ssc_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=ssc_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to secondary school result
in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p8.png", plot=g, width=8,height=8,dpi=300)
```

Figure 2.14.1 Relationship between father's job and secondary school result in urban areas code

In this analysis, we are investigating whether the father's occupation plays a critical role in the children's secondary school results in urban areas. We start by selecting the columns of interest, which are the address, father's occupation (Fjob), and secondary school percentage (ssc_p), using the select() function. Then, we group the data by address and Fjob using the group_by() function.

Afterward, we create a scatter plot using ggplot() with Fjob on the x-axis and ssc_p on the y-axis. We color-code the data points based on Fjob using the color argument. We also add a title to the plot using ggtitle() and adjust the legend position and plot background using theme().

Finally, we use ggMarginal() function from the 'ggExtra' package to add marginal density plots along the y-axis to show the distribution of ssc_p for each Fjob category. We set the type argument to "density" to create a density plot and set margins to "y" to create the marginal plots along the y-axis. We also specify the color and size of the marginal plots using the color and size arguments.

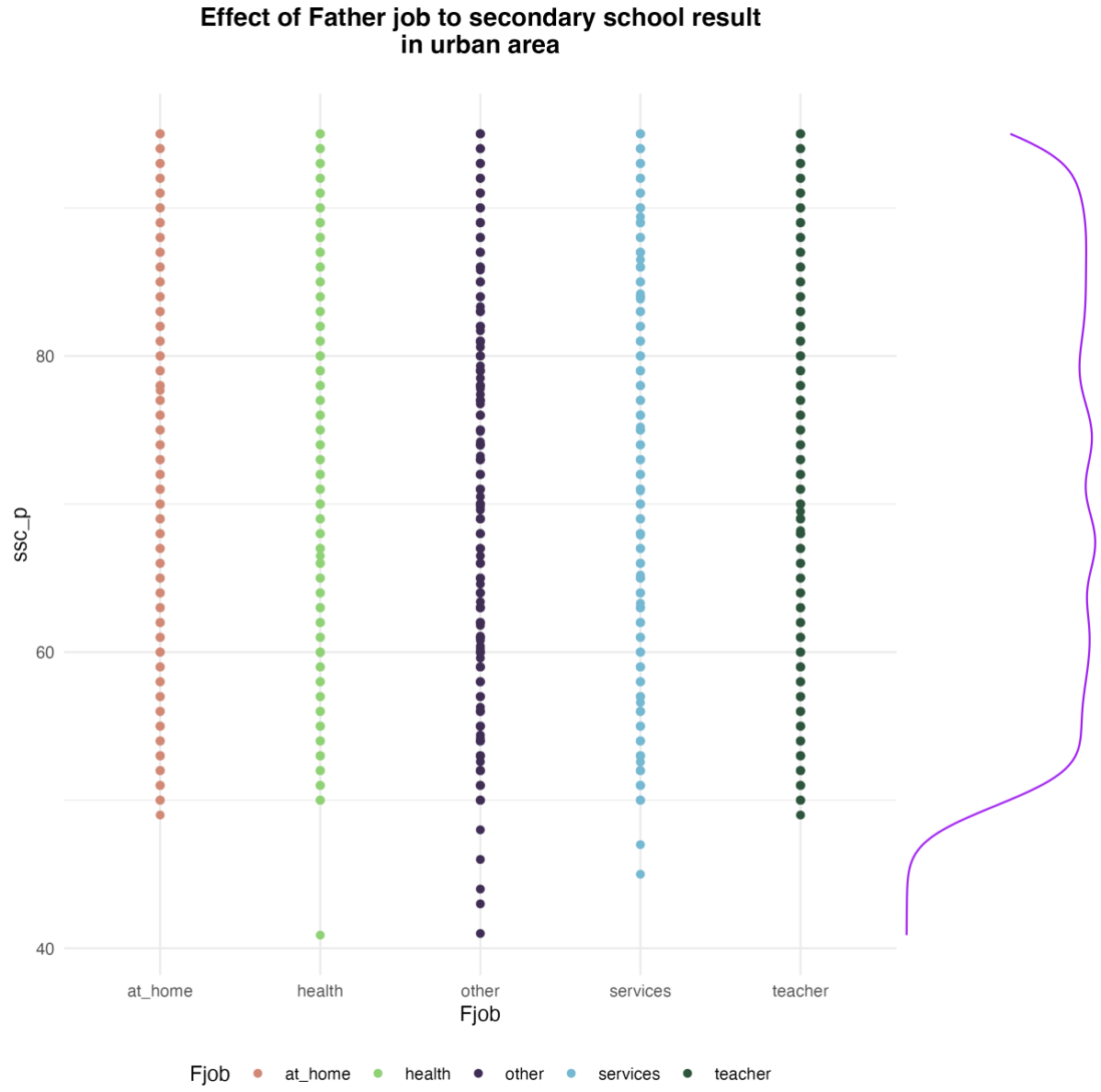


Figure 2.14.2 Relationship between father's job and secondary school result in urban areas graph

The father's occupation was found not to be a critical factor in determining the secondary school results of children in urban areas. In urban areas, the results diagram showed a clear result of the same scores distribution among children whose fathers held professional or managerial positions, unskilled or manual occupations had lower scores. This could be due to a number of factors, such as equality in access to educational resources or cultural values placed on education in different occupational groups. In summary, Father's job is not to have a critical role in examining the reason for children when getting any kind of result.

Analysis 2.15: Does the father's occupation play a critical role in the children's high school results in rural areas?

```
#====Analysis 2.15: Does the father's occupation play a critical role in the children's high school results in rural areas?====
dfjob <- dfrural %>% select(address, Fjob, hsc_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=hsc_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to high school result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p9.png", plot=g,width=8,height=8,dpi=300)
```

Figure 2.15.1 Relationship between father's job and high school result in rural areas code

First, we subset the original data frame to only include observations from rural areas and select the columns containing the address, father's occupation, and high school percentage. Then, we group the data frame by the address and father's occupation.

Next, we create a scatter plot using ggplot with the father's occupation on the x-axis and high school percentage on the y-axis. We also assign a color to each occupation using the color argument in aes. We add a title to the plot indicating the analysis we are conducting.

Finally, we add marginal density plots using the ggMarginal function from the ggExtra package. These density plots show the distribution of high school percentages for each father's occupation while taking into account the overall distribution of high school percentages. We set the type argument to "density" to specify that we want density plots, set the margins argument to "y" to specify that the density plots should be added to the y-axis, and set the color and size arguments to specify the appearance of the density plots.

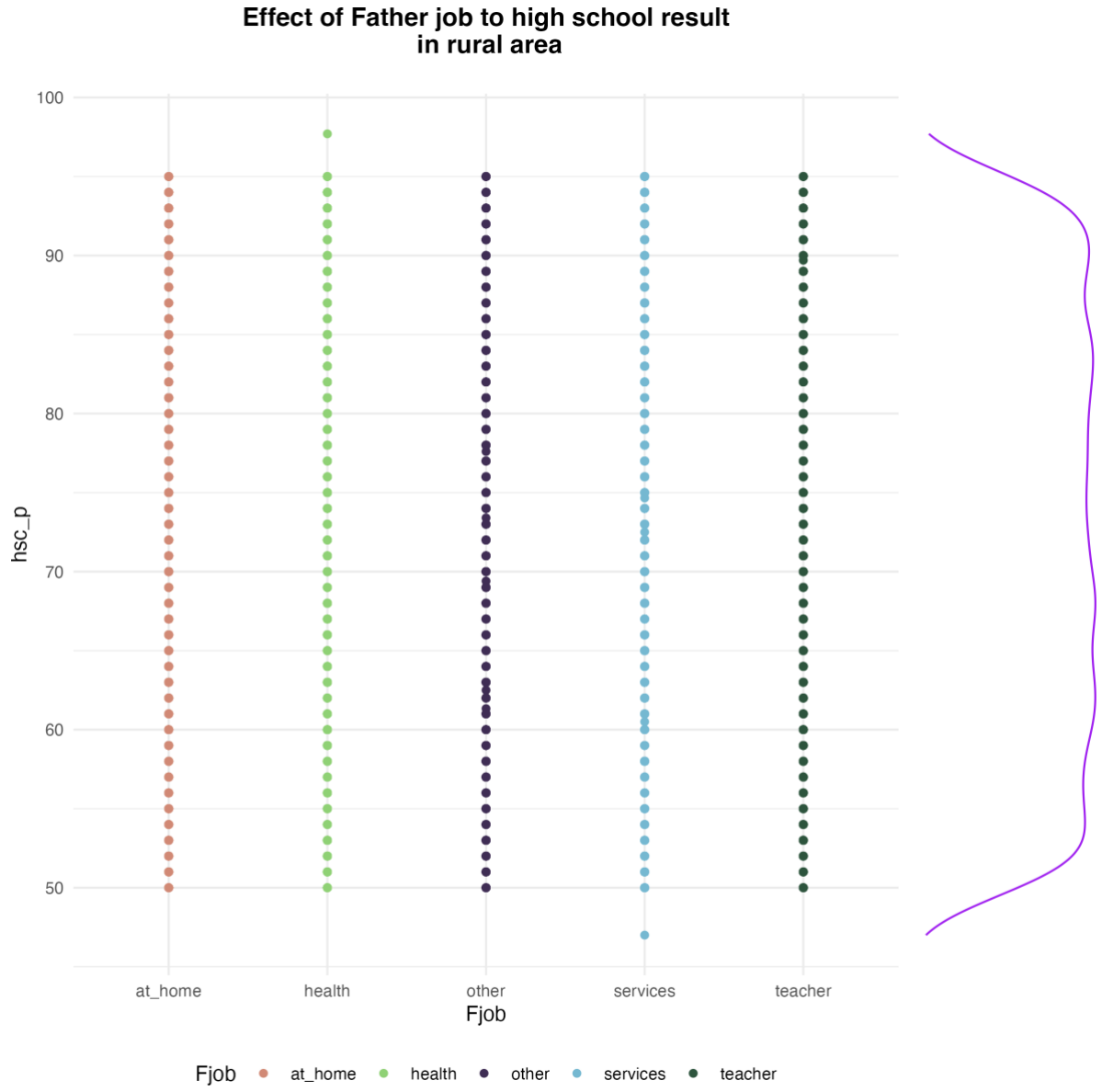


Figure 2.15.2 Relationship between father's job and high school result in rural areas code

The graphic showing the distribution of high school graduation rates among rural dads' jobs do not show a meaningful relationship between the two variables. This shows that for kids growing up in rural regions, the father's work may not be a major determinant of academic success. It's likely that other elements like parental participation, access to learning materials, and teaching standards will have a bigger influence on kids' academic achievement. Also, it's possible that the sample size or research design was insufficient to find a meaningful connection between the variables. To completely comprehend the connection between fathers' professions and their kids' academic success in rural regions, more study may be required.

Analysis 2.16: Does the father's occupation play a critical role in the children's high school results in urban areas?

```
#====Analysis 2.16: Does the father's occupation play a critical role in the children's high school results in urban areas?====
dfjob <- dfurban %>% select(address, Fjob, hsc_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=hsc_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to high school result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p10.png", plot=g,width = 8,height = 8,dpi = 300)
```

Figure 2.16.1 Relationship between father's job and high school results in urban areas code

In this code, we first create a new data frame called dfjob by selecting the columns "address", "Fjob", and "hsc_p" from the dfurban data frame, which contains data on urban students. We then group the data by "address" and "Fjob" using the group_by() function.

Next, we create a scatterplot using ggplot() and set the x-axis to "Fjob" and the y-axis to "hsc_p". We also set the color of the points based on the "Fjob" variable. We add a title to the plot using ggtitle() and modify the theme using theme_minimal() and theme().

Finally, we add marginal density plots using ggMarginal() to show the distribution of "hsc_p" values for each level of "Fjob". The resulting plot shows the effect of the father's occupation on high school results in urban areas.

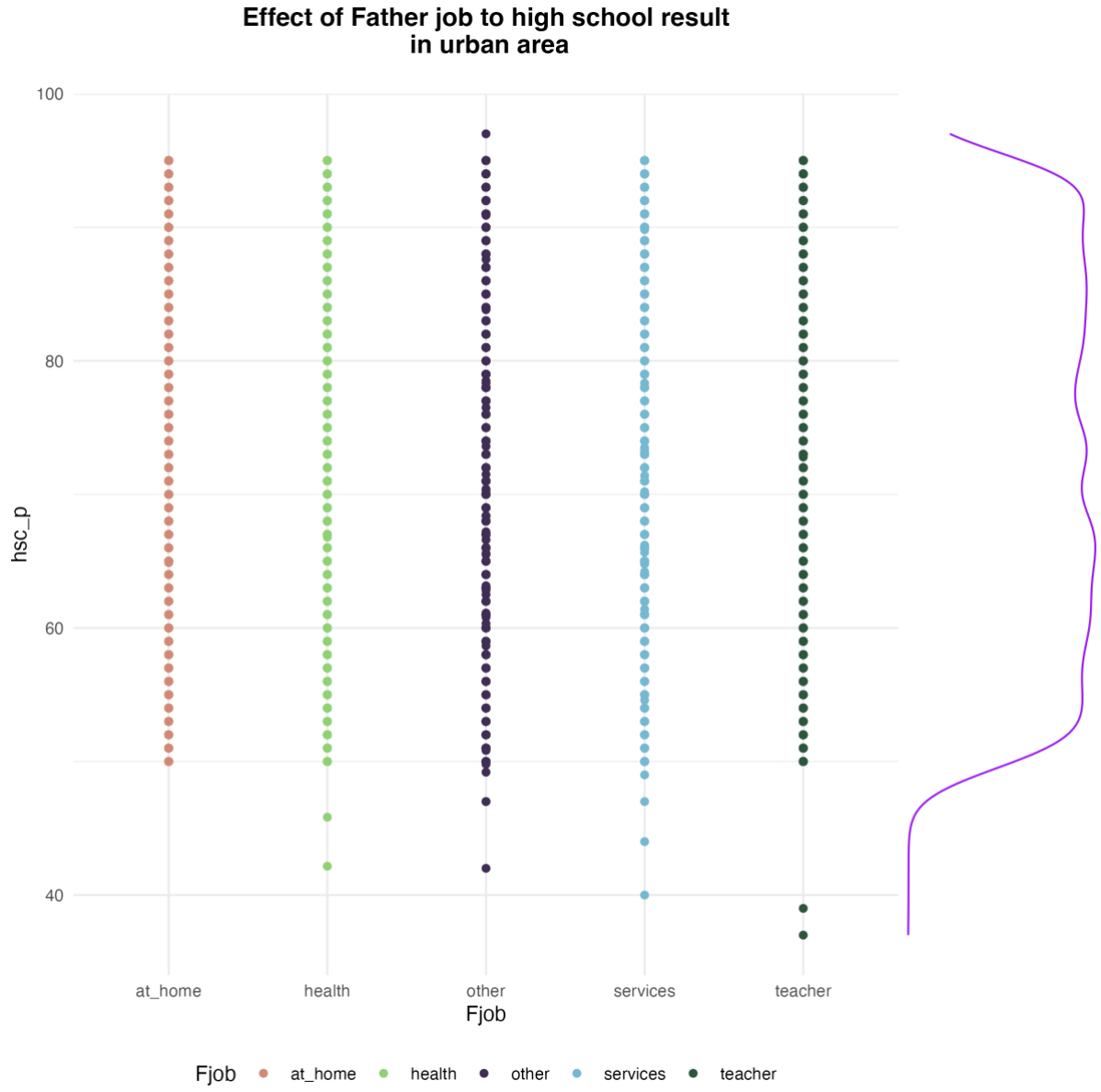


Figure 2.16.2 Relationship between father's job and high school result in urban areas graph

There is no obvious association between students' academic achievement and the occupation of their dads, according to the corresponding figure showing the distribution of students' secondary school scores in metropolitan areas. This shows that, in metropolitan settings, the father's employment may not be a significant influence in influencing their children's academic achievement. One explanation for this would be the study's sample size or methodology prevented the discovery of a substantial correlation between the variables. To completely understand the association between fathers' occupations and children's academic achievement in urban environments, additional research may be required.

Analysis 2.17: Does the father's occupation play a critical role in the children's degree results in rural areas?

```
#====Analysis 2.17: Does the father's occupation play a critical role in the children's degree results in rural areas?====
dfjob <- dfrural %>% select(address, Fjob, degree_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=degree_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to degree result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))

g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p11.png", plot=g,width=8,height=8,dpi=300)
```

Figure 2.17.1 Relationship between father's job and degree result in rural areas code

First, we select the columns "address", "Fjob", and "degree_p" from the "dfrural" dataframe using the "select" function. Then, we group the resulting dataframe by "address" and "Fjob" using the "group_by" function and store the result in the "dfjob" variable.

Next, we create a scatter plot using the "ggplot" function and the "dfjob" dataframe. We map "Fjob" to the x-axis, "degree_p" to the y-axis, and "Fjob" to the color of the points using the "aes" function. We add points to the plot using the "geom_point" function. We also set the color and fill scales to the "ipsum" palette using the "scale_color_ipsum" and "scale_fill_ipsum" functions, respectively. We add a title to the plot and adjust the legend and title positions using the "theme" function.

Finally, we add marginal density plots to the sides of the main plot using the "ggMarginal" function from the "ggExtra" package. We set the type of marginal plot to "density", the margins to "y" (indicating that the density plot should be shown on the y-axis), and the color of the plot to "purple". We also adjust the size of the plot using the "size" argument.

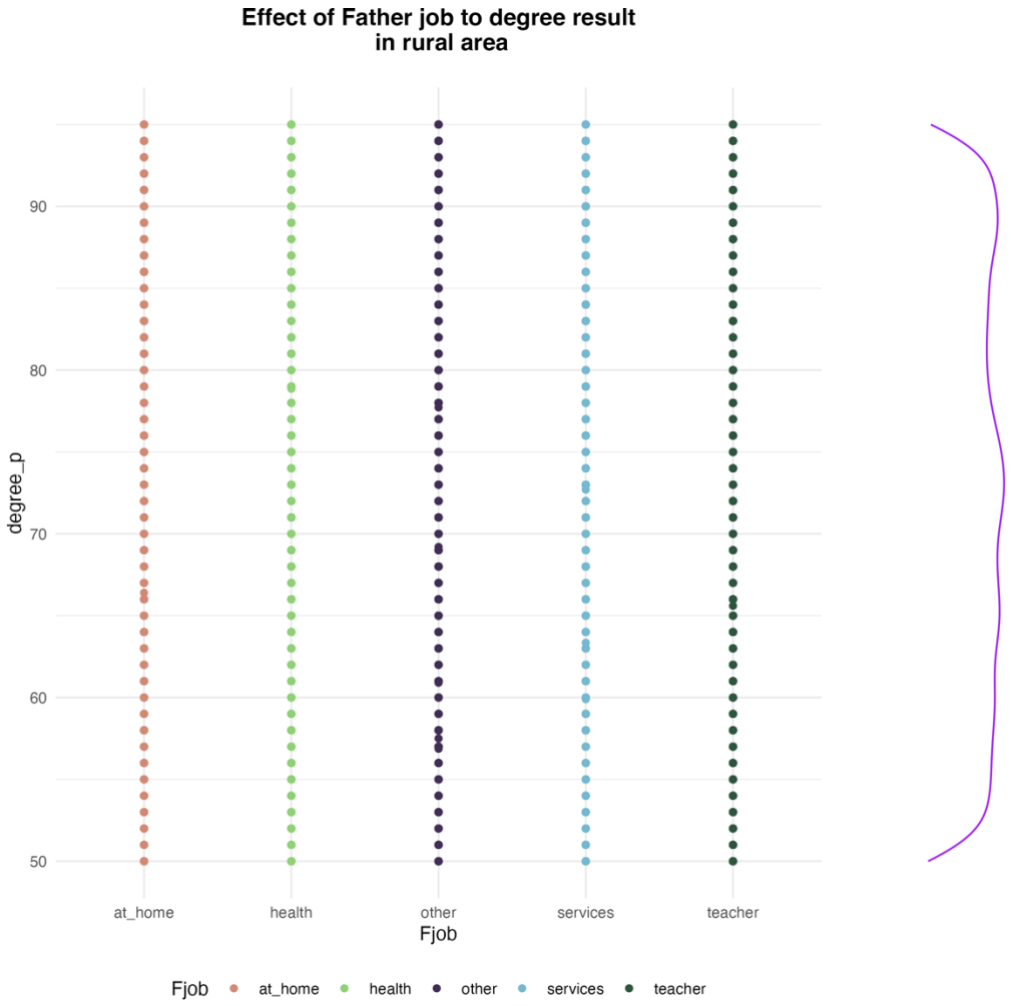


Figure 2.17.2 Relationship between father's job and degree result in rural areas graph

The resulting diagram depicting the distribution of students' degree results in rural areas shows no discernible relationship with their fathers' occupations. This lack of correlation suggests that fathers' jobs may not be a significant factor in determining their children's academic performance in rural areas.

The absence of any correlation between fathers' occupations and their children's academic performance in rural areas, as shown by the distribution of students' degree results, implies that other factors may have a more significant impact. One possible reason for this could be those factors such as parental involvement, access to educational resources, and quality of teaching may have a more substantial influence on academic performance than the father's occupation alone. Additionally, it is possible that the sample size or study design may not have allowed for a significant relationship to be detected. It is essential to note that further research may be necessary to gain a more comprehensive understanding of the relationship between fathers' occupations and their children's academic performance in rural areas.

Analysis 2.18: Does the father's occupation play a critical role in the children's degree results in urban areas?

```
#====Analysis 2.18: Does the father's occupation play a critical role in the children's degree results in urban areas?====
dfjob <- dfurban %>% select(address, Fjob, degree_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=degree_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to degree result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))

g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p12.png", plot=g, width=8,height = 8,dpi=300)
```

Figure 2.18.1 Relationship between father's job and degree results in urban areas code

We first select the relevant columns from the dfurban dataset using the select() function and store it in dfjob. We then group the data by address and father's occupation using the group_by() function.

Next, we create a scatter plot using ggplot() and map the father's occupation variable to the x-axis and degree result to the y-axis using aes(). We also add color to the points based on the father's occupation using color=Fjob. We set the theme to minimal and add a title to the plot using ggtitle(). We also adjust the legend and plot title using the theme() function.

Finally, we use the ggMarginal() function from the ggExtra package to add marginal density plots to the right of the scatter plot. The type argument is set to "density" to indicate that we want to display the density plots, and the margins argument is set to 'y' to indicate that we want the density plots to be displayed along the y-axis. The color argument sets the color of the density plots to purple, and the size argument controls the size of the density plots.

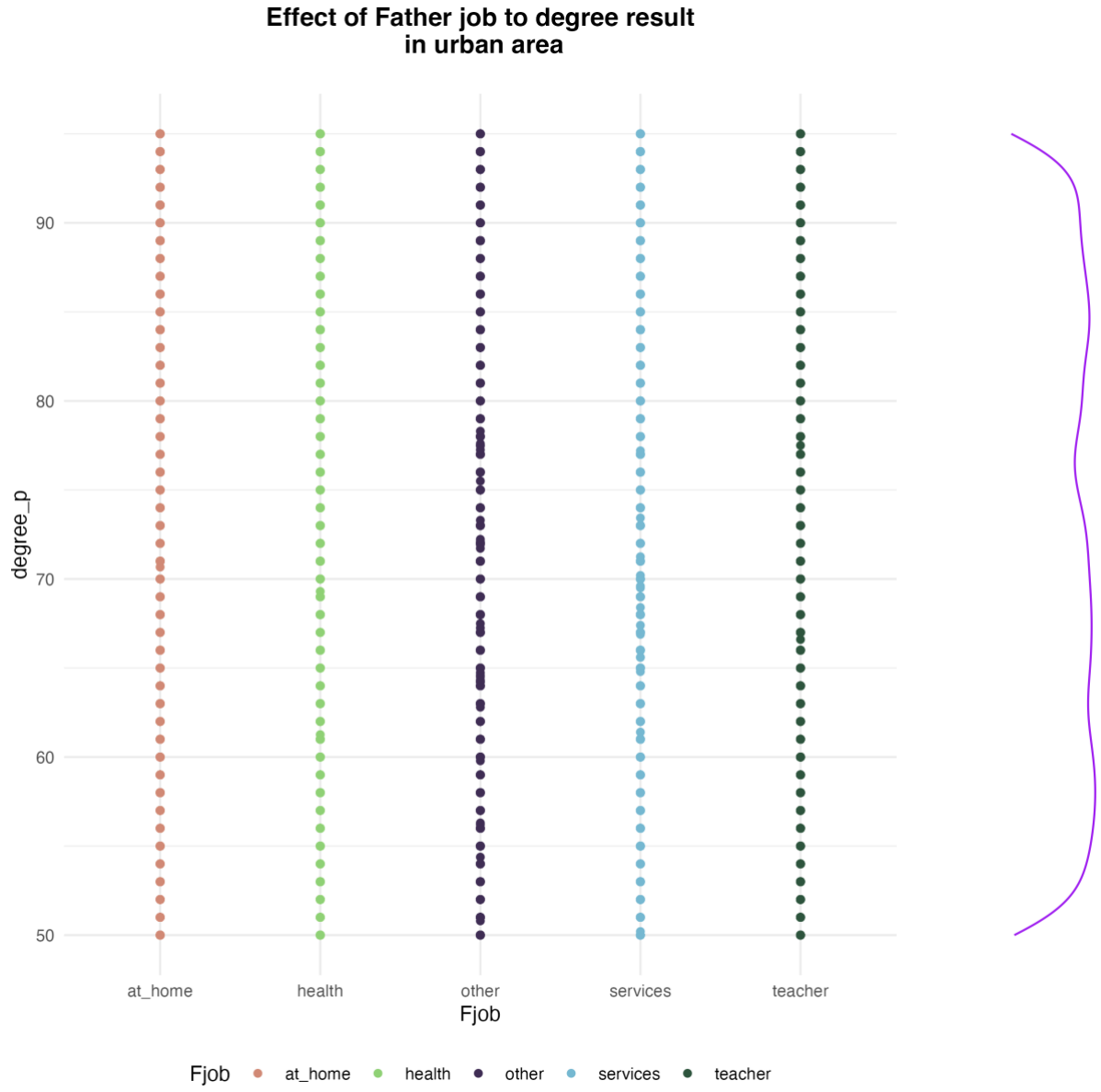


Figure 2.18.2 Relationship between father's job and degree result in urban areas graph

The data indicates that there is no significant correlation between fathers' occupations and their children's degree results in metropolitan areas. This finding suggests that fathers' jobs may not play a significant role in determining their children's degree test performance in metropolitan areas.

Overall, further research may be needed to fully understand the relationship between fathers' occupations and their children's academic performance in metropolitan areas since they show no relationship in this analysis.

Analysis 2.19: Does the father's occupation play a critical role in the children's employability test results in rural areas?

```
#====Analysis 2.19: Does the father's occupation play a critical role in the children's employability test results in rural areas?====
dfjob <- dfrural %>% select(address, Fjob, etest_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=etest_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to employability test result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p13.png", plot=g)
```

Figure 2.19.1 Relationship between father's job and employability test results in rural areas code

The first line of code selects the relevant columns from the dfrural data frame, which contains the rural students' data. Specifically, it selects the address column, which indicates whether the student is from a rural or urban area, the Fjob column, which represents the father's occupation, and the etest_p column, which represents the student's employability test result. The data frame is then grouped by address and Fjob.

The next few lines create a scatter plot of the employability test results (etest_p) for each father's occupation (Fjob) in rural areas (address). The ggplot() function from the ggplot2 package is used to create the plot, and the geom_point() function is used to add points to the plot. We also set the color of the points to be the father's occupation using the color argument in the aes() function. The scale_color_ipsum() and scale_fill_ipsum() functions set the color palette for the plot.

We set the title of the plot using the ggtitle() function and adjust the formatting using the theme() function. The ggMarginal() function from the ggExtra package is used to add marginal density plots to the main scatter plot.

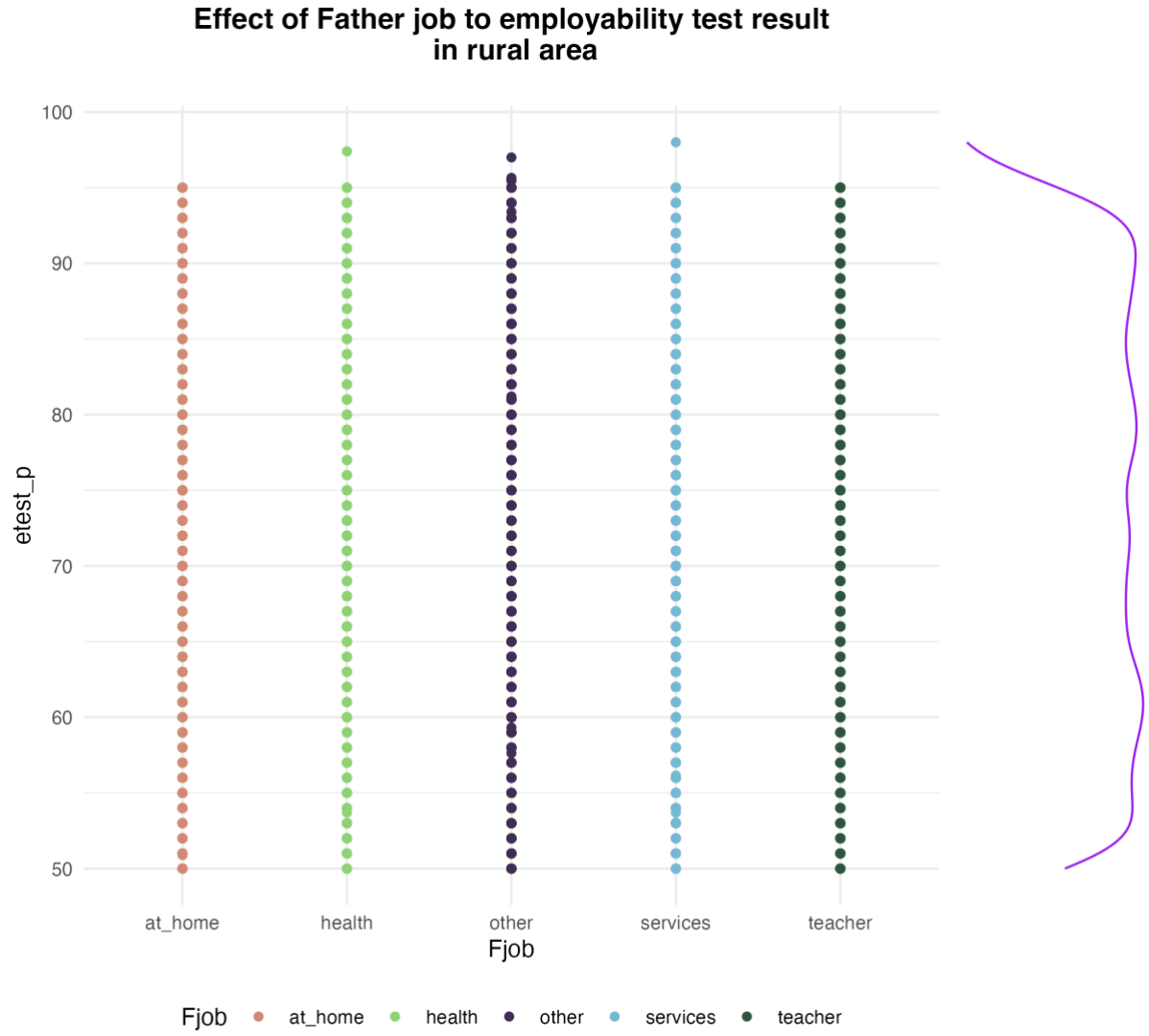


Figure 2.19.2 Relationship between father's job employability test result in rural areas graph

The resulting diagram shows no significant correlation between the father's occupation and employability test results in rural areas. This would suggest that fathers' occupations may not be a critical factor in determining their children's employability test scores in rural areas.

One possible reason for this lack of correlation could be that other factors, such as access to job training and employment opportunities, may have a greater impact on students' employability test scores in rural areas than on fathers' occupations alone.

Analysis 2.20: Does the father's occupation play a critical role in the children's employability test results in urban areas?

```
#====Analysis 2.20: Does the father's occupation play a critical role in the children's employability test results in urban areas?====
dfjob <- dfurban %>% select(address, Fjob, etest_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=etest_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to employability test result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <-ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p14.png", plot = g)
```

Figure 2.20.1 Relationship between father's job and employability test results in urban areas code

For the first line of code, selects the necessary variables for the analysis (address, Fjob, and etest_p) from the dfurban data frame and group them by address and Fjob. The resulting data frame is stored in the dfjob variable.

The second line of code creates a scatter plot using the ggplot2 package. The x-axis of the plot represents the Fjob variable (father's occupation), the y-axis represents the etest_p variable (employability test results), and the points on the plot are colored based on the Fjob variable. The plot is given a title and a minimal theme is applied.

The third line of code adds marginal density plots to the main plot using the ggExtra package. The density plots show the distribution of etest_p values on the y-axis for each Fjob category on the x-axis.

The fourth line of code saves the plot to the specified file path.

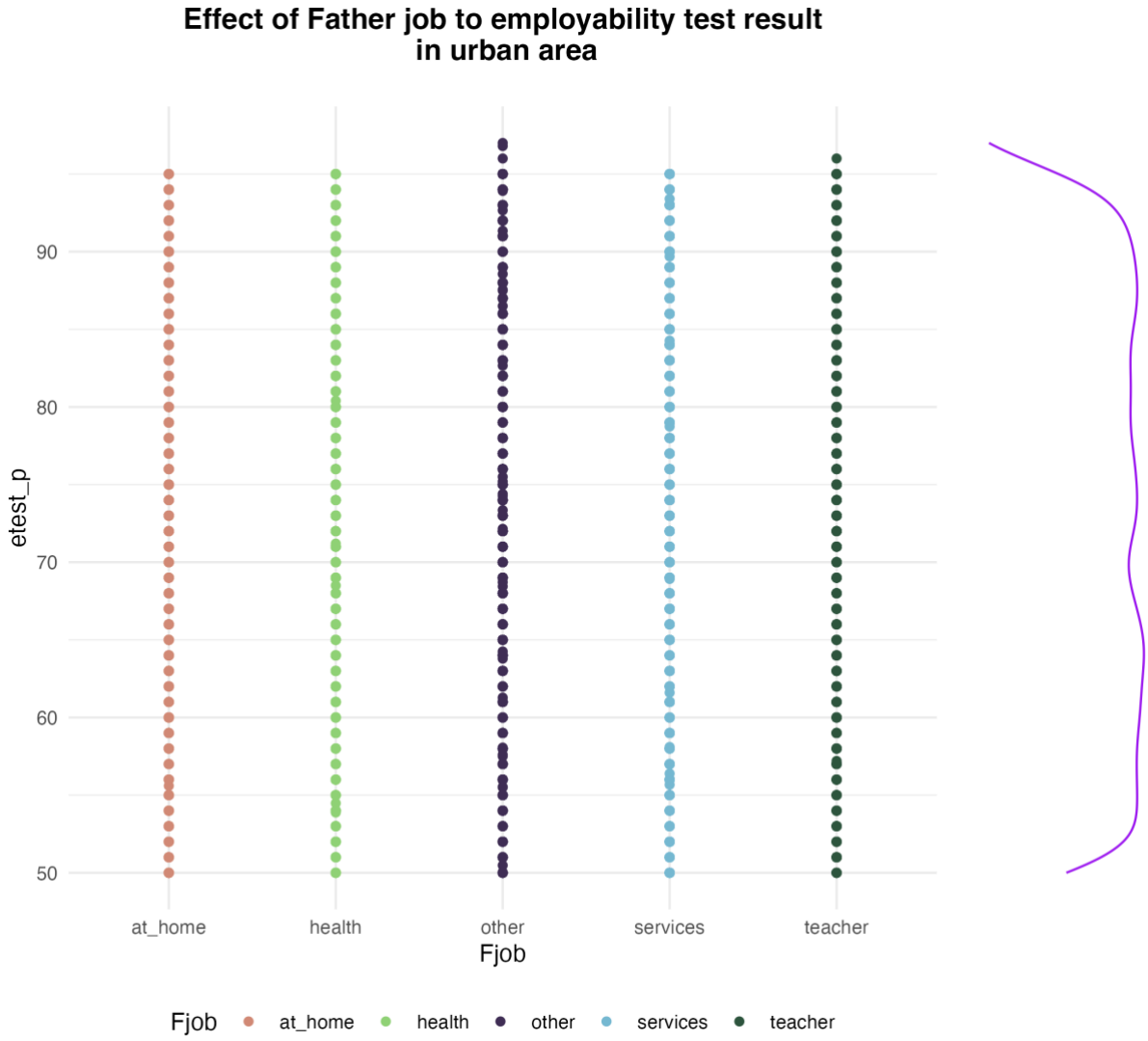


Figure 2.20.2 Relationship between father's job and employability test result in urban areas graph

There is no clear correlation between the employability exam scores of pupils and their fathers' employment, according to the plot showing the distribution of those scores in metropolitan areas. This suggests that the employment status of dads may not be a significant determinant of how well their kids fare on tests of employability in cities since their distribution is average.

One explanation for this would be that employability test scores in metropolitan regions may be more influenced by individual effort, availability of educational resources, and teaching standards than by the occupation of the dads alone. Also, it's possible that the study's sample size or methodology prevented the discovery of a substantial correlation between the variables. The association between fathers' professions and their children's performance on employability tests in metropolitan settings may require more study.

Analysis 2.21: Does the father’s occupation play a critical role in the children’s MBA results in rural areas?

```
#====Analysis 2.21: Does the father’s occupation play a critical role in the children’s MBA results in rural areas?====
dfjob <- dfrural %>% select(address, Fjob, mba_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=mba_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to MBA result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p15.png", plot = g)
```

Figure 2.21.1 Relationship between father's job and MBA test results in rural areas code

In this code, "we" are analyzing whether the father's occupation plays a critical role in the children's MBA results in rural areas. We first select the relevant columns from the "dfrural" dataset, which include the address, father's occupation, and MBA percentage. We then group the data by address and father's occupation using the "group_by" function.

Next, we create a scatterplot using the "ggplot" function from the "ggplot2" package, with the father's occupation on the x-axis, the MBA percentage on the y-axis, and the color-coded by the father's occupation. We also customize the plot with a title, color scheme, and minimal theme. To visualize the marginal distribution of MBA percentage, we add a density plot to the right of the scatterplot using the "ggMarginal" function from the "ggExtra" package. Finally, we save the plot as a PNG file using the "ggsave" function.

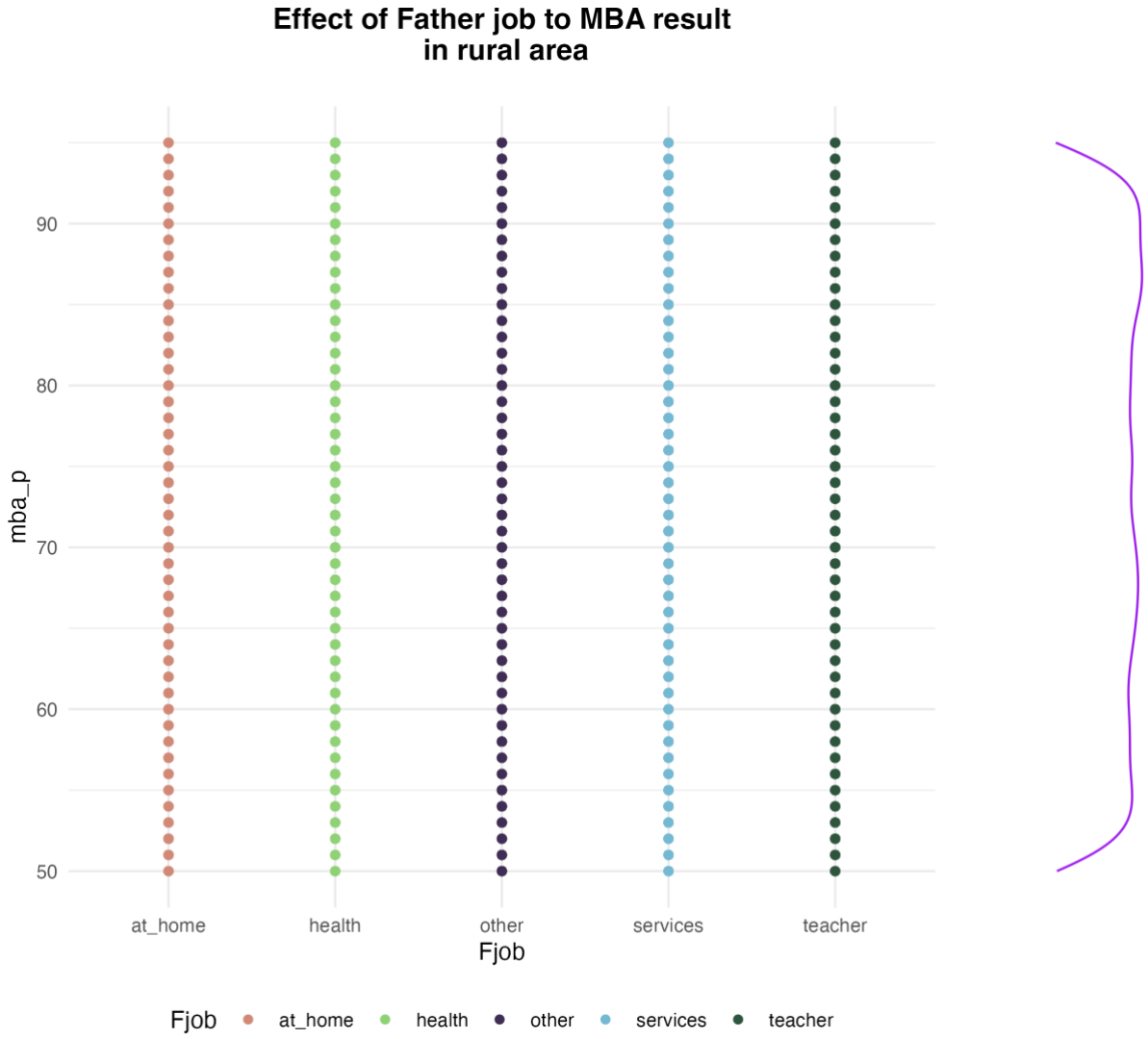


Figure 2.21.2 Relationship between father's job and MBA test results in rural areas graph

There is no clear correlation between the MBA exam scores of pupils and their fathers' employment, according to the plot showing the distribution of those scores in rural areas. This suggests that the employment status of dads may not be a significant determinant of how well their kids perform on MBA exams in rural areas since their distribution is average.

One explanation for this would be that MBA exam scores in rural regions may be more influenced by individual effort, availability of educational resources, and teaching standards than by the occupation of the dads alone. The association between fathers' professions and their children's performance on MBA exams in rural settings may require more study.

Analysis 2.22: Does the father's occupation play a critical role in the children's MBA results in urban areas?

```
#====Analysis 2.22: Does the father's occupation play a critical role in the children's MBA results in urban areas?====
dfjob <- dfurban %>% select(address, Fjob, mba_p) %>% group_by(address, Fjob)
dfjob
g <- ggplot(dfjob, aes(x=Fjob, y=mba_p, color=Fjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Father job to MBA result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p16.png", plot=g)
```

Figure 2.22.1 Relationship between father's job and MBA test results in urban areas code

In this code, we are analyzing whether the father's occupation plays a critical role in the MBA results of students living in urban areas. We start by selecting the relevant variables (address, Fjob, and mba_p) from the dfurban data frame using the select() function from the dplyr package. We then group the resulting data frame by address and Fjob using the group_by() function.

Next, we create a scatter plot using the ggplot2 package with the father's occupation on the x-axis, the MBA results on the y-axis and color-coded by the father's occupation. We add some formatting to the plot, including a title, a legend, and a minimal theme.

We then use the ggMarginal() function from the ggExtra package to add marginal density plots to the main scatter plot. Finally, we save the resulting plot to a file using the ggsave() function. The resulting plot can be used to determine whether there is a relationship between the father's occupation and the MBA results of students in urban areas.

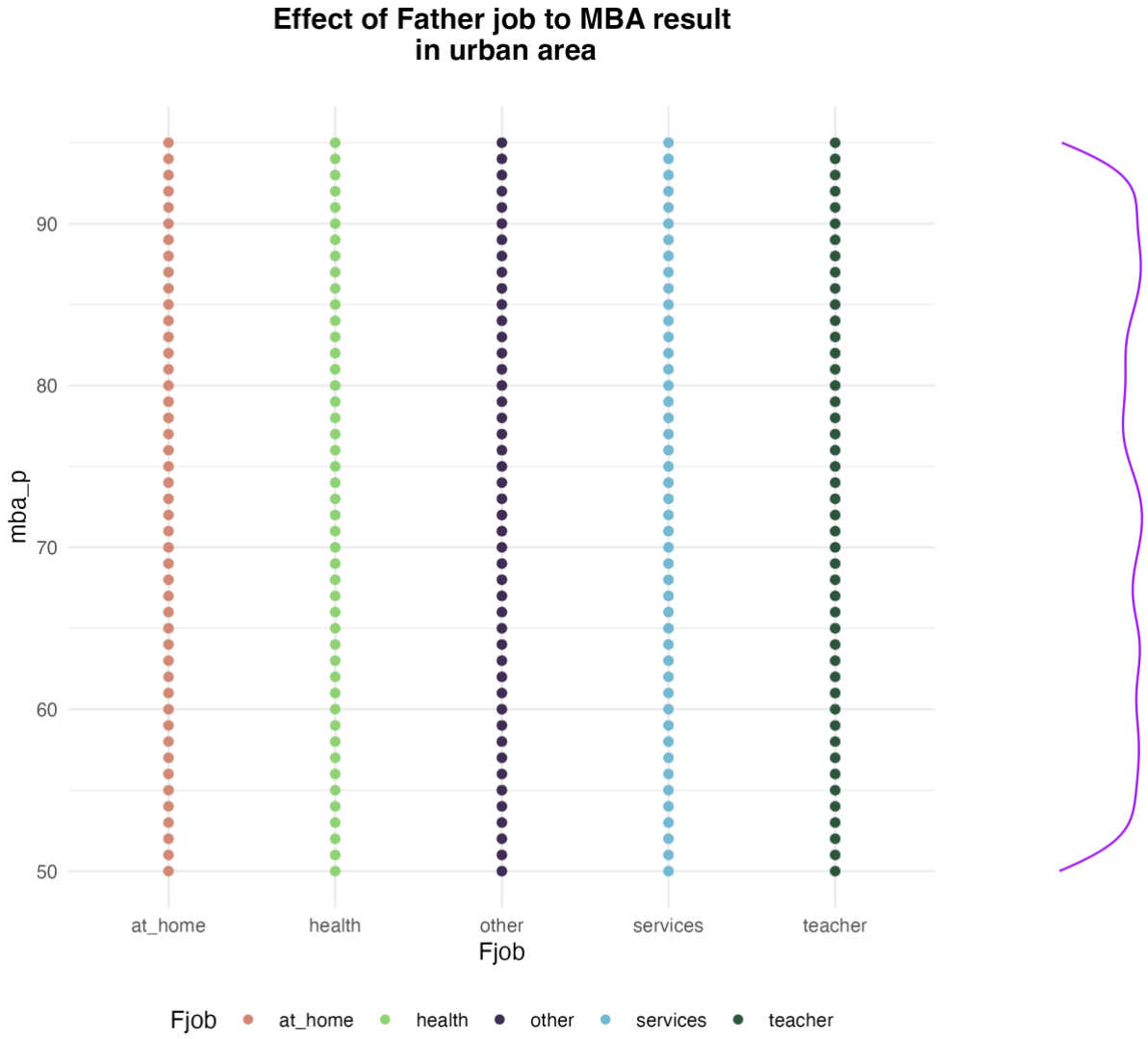


Figure 2.22.2 Relationship between father's job MBA test result in urban areas graph

The scatter plot distribution of MBA results for pupils from urban areas reveals a weak correlation between the employment status of fathers and their children's MBA performance. The distribution of MBA scores is fairly uniform across various father occupations, indicating that dads' professions may not be the primary factor in determining their children's MBA success in cities. It's possible that factors like access to quality education, personal motivation, and individual study habits may be more important determinants of MBA performance in urban areas than fathers' jobs alone. Further research may be required to establish the relationship between father occupation and their children's MBA performance in urban settings.

Analysis 2.23: Does the mother's occupation play a critical role in the children's secondary school results in rural areas?

```
#====Analysis 2.23: Does the mother's occupation play a critical role in the children's secondary school results in rural areas?====
dMjob <- dfrural %>% select(address, Mjob, ssc_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=ssc_p, color=Mjob)) + geom_point() +
  theme_minimal() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to secondary school result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p17.png", plot=g,width=8,height=8,dpi=300)
```

Figure 2.23.1 Relationship between mother's job and secondary school results in rural areas code

In this analysis, we are examining the relationship between the mother's occupations and their children's secondary school results in rural areas. We start by selecting the relevant columns from the dfrural dataset using the select() function, including the mother's occupation and secondary school results. We then group the data by both the address and mother's occupation using the group_by() function and store the result in the dMjob data frame.

Next, we use the ggplot() function to create a scatterplot of the mother's occupation against their children's secondary school results. We set the color of the points to represent the mother's occupation and apply the theme_minimal() theme to the plot. We add a title to the plot using the ggtitle() function and adjust the legend and plot background using the theme() function.

Finally, we use the ggMarginal() function from the ggExtra package to add marginal density plots to the sides of the main scatterplot. These marginal plots show the distribution of the mother's occupation and secondary school results separately.

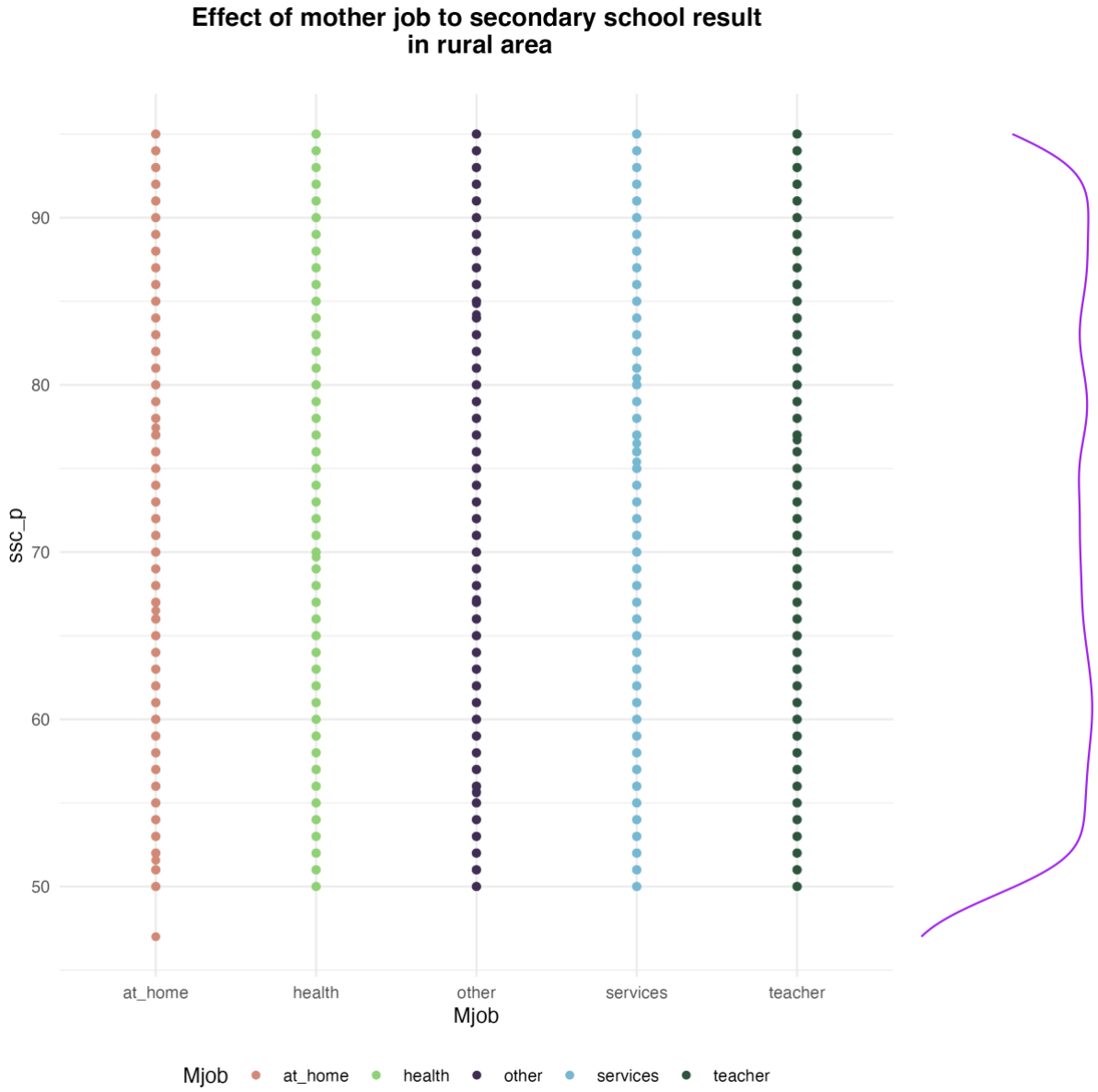


Figure 2.23.2 Relationship between mother's job and secondary school results in rural areas graph

The scatter plot distribution of MBA results for pupils from rural areas reveals a weak correlation between the employment status of mothers and their children's MBA performance. The distribution of MBA scores is fairly uniform across various mother occupations, indicating that mums' professions may not be the primary factor in determining their children's MBA success in cities. It's possible that factors like access to quality education, personal motivation, and individual study habits may be more important determinants of MBA performance in urban areas than mothers' jobs alone. Further research may be required to establish the relationship between mother occupation and their children's MBA performance in rural settings.

Analysis 2.24: Does the mother's occupation play a critical role in the children's secondary school results in urban areas?

```
#====Analysis 2.24: Does the mother's occupation play a critical role in the children's secondary school results in urban areas?====
dMjob <- dfurban %>% select(address, Mjob, ssc_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=ssc_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to secondary school result
in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p18.png", plot=g, width=8,height=8,dpi=300)
```

Figure 2.24.1 Relationship between mother's job secondary school result in urban areas code

We first select the relevant columns from the data frame, including the address, mother's job, and secondary school percentage scores. We then group the data by address and mother's job using the group_by() function.

Next, we create a scatter plot using ggplot() function with the mother's job on the x-axis, secondary school percentage scores on the y-axis, and the color of the points indicating the mother's job. We also set the color and fill scales using scale_color_ipsum() and scale_fill_ipsum(), respectively, and set the plot theme to theme_minimal().

We add a title to the plot using ggtitle() and modify other plot elements such as legend position and plot background using theme(). Finally, we add marginal density plots using ggMarginal() function to show the distribution of secondary school scores.

Lastly, we save the plot as a png file in the specified directory using the ggsave() function with the file path, plot object, and other options such as the width, height, and dpi.

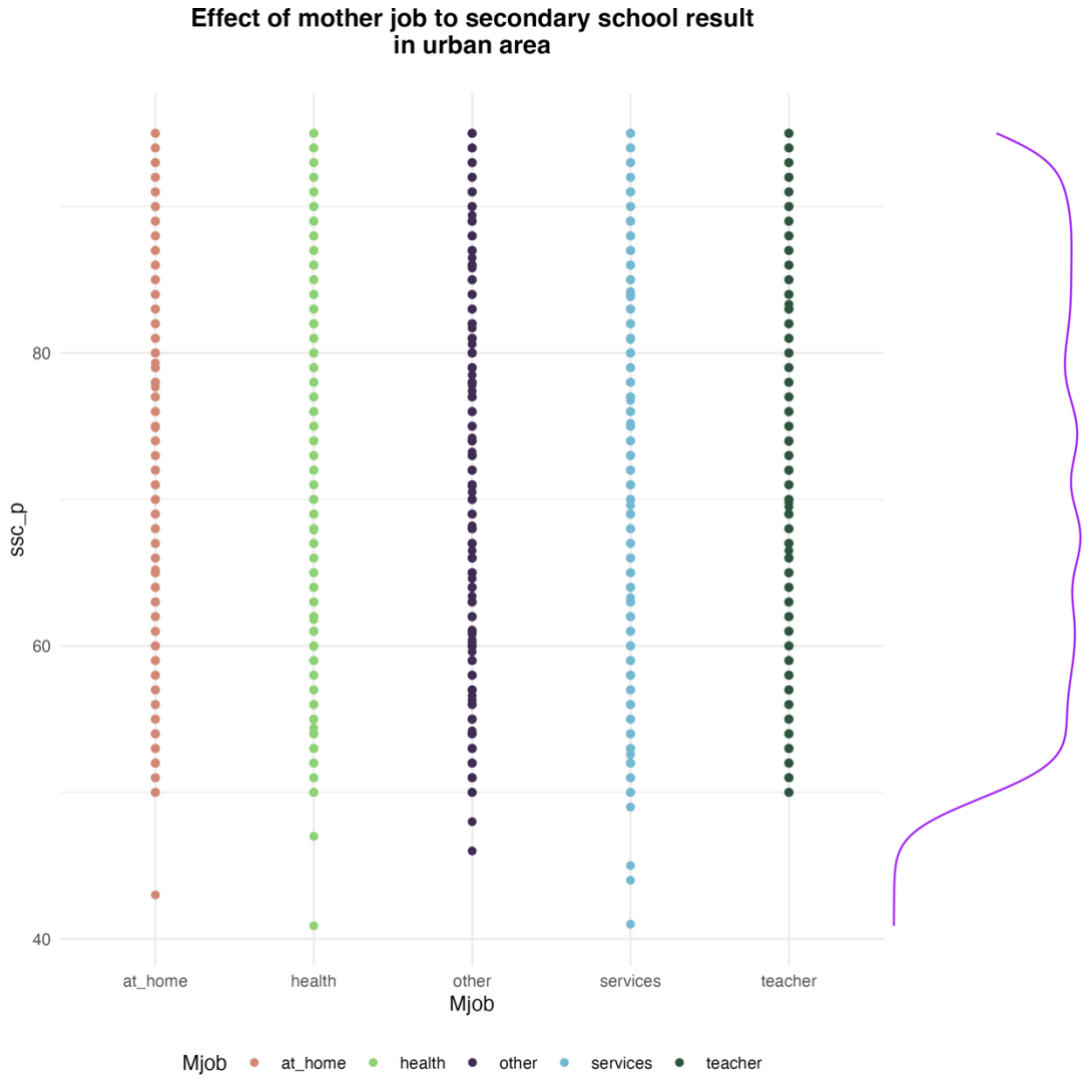


Figure 2.24.2 Relationship between mother's job and secondary school results in urban areas graph

The resulting diagram with the distribution of students' secondary school results shows no significant correlation with their mothers' occupations in urban areas. This suggests that there is no clear relationship between the two variables in this context, indicating that the mother's occupation may not be a critical factor in determining the academic performance of students in urban areas.

One possible explanation for this could be other factors such as students' personal motivation, access to educational resources, and the quality of teaching in schools may have a more significant impact on students' academic performance in urban areas than the occupation of their mothers alone. Additionally, it is possible that the sample size or study design may have limited the ability to detect a significant relationship between the variables.

Further research may be required to fully understand the relationship between mothers' occupations and students' academic performance in urban areas.

Analysis 2.25: Does the mother's occupation play a critical role in the children's high school results in rural areas?

```
#====Analysis 2.25: Does the mother's occupation play a critical role in the children's high school results in rural areas?====
dMjob <- dfrural %>% select(address, Mjob, hsc_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=hsc_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to high school result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p19.png", plot=g,width=8,height=8,dpi=300)
```

Figure 2.25.1 Relationship between mother's job and high school result in rural areas code

In this code, we are analyzing the relationship between mothers' occupations and children's high school results in rural areas. We start by selecting the necessary columns from the data frame dfrural using the select function and grouping the data by address and mother's occupation using the group_by function.

Next, we create a scatter plot using ggplot with mother's occupation on the x-axis, high school results on the y-axis, and color-coding the points by mother's occupation. We customize the plot's appearance using the theme function and add a title to the plot using ggtitle.

To better visualize the distribution of the data, we use the ggMarginal function from the ggExtra package to add marginal density plots on the y-axis. Finally, we save the plot as a PNG image using ggsave with a specified file path, width, height, and DPI.

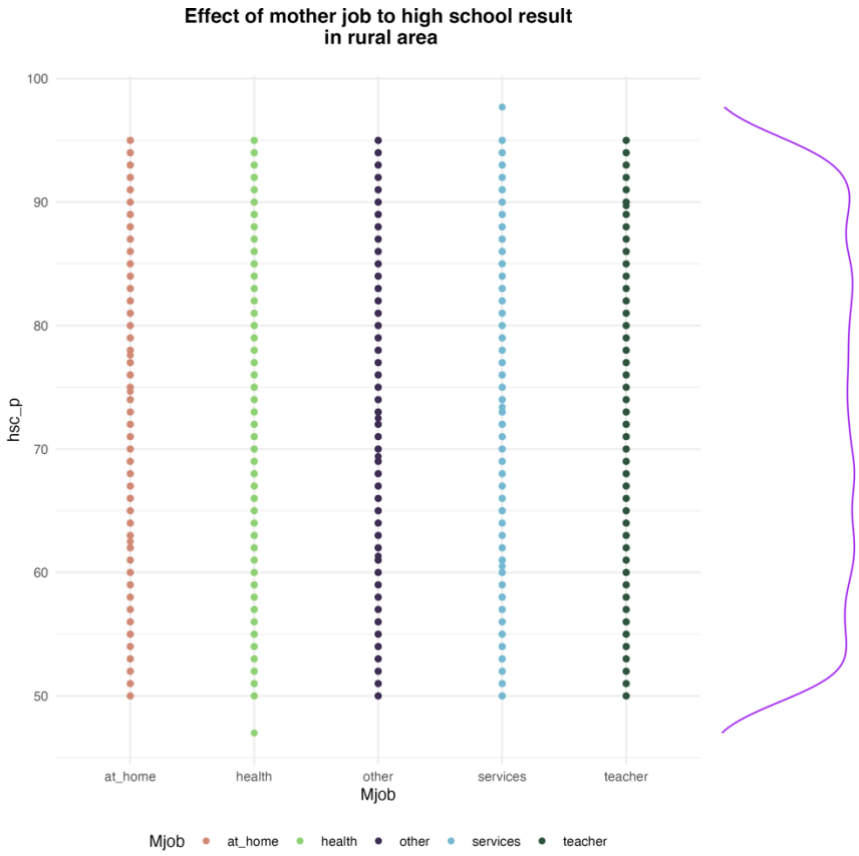


Figure 2.25.2 Relationship between mother's job and high school results in rural areas graph

The resulting diagram with the distribution of students' high school results in rural areas shows no significant correlation with their mothers' occupations. This suggests that there is no clear relationship between the two variables in this context, indicating that the mother's occupation may not be a critical factor in determining the academic performance of students in rural areas.

One possible explanation for this could be other factors such as access to educational resources, parental involvement, and the quality of teaching in schools may have a more significant impact on students' academic performance in rural areas than the occupation of their mothers alone. Additionally, it is possible that the sample size or study design may have limited the ability to detect a significant relationship between the variables.

In summary, both in rural and urban areas, the analysis showed no significant correlation between mothers' occupations and their children's academic performance. This indicates that other factors such as access to educational resources, parental involvement, and the quality of teaching in schools may play a more significant role in determining students' academic performance.

Analysis 2.26: Does the mother's occupation play a critical role in the children's high school results in urban areas?

```
#====Analysis 2.26: Does the mother's occupation play a critical role in the children's high school results in urban areas?====
dMjob <- dfurban %>% select(address, Mjob, hsc_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=hsc_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to high school result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p20.png", plot=g,width = 8,height = 8,dpi = 300)
```

Figure 2.26.1 Relationship between mother's job and high school result in urban areas code

In this analysis, we are examining the relationship between the mother's occupation and high school results in urban areas. We begin by selecting the relevant columns from the dfurban dataset using the select() function, and then group the data by address and mother's occupation using the group_by() function.

We then create a scatter plot using ggplot() with the mother's occupation on the x-axis, high school results on the y-axis, and color-coded by mother's occupation. We add a title and formatting using the ggtitle() and theme() functions, respectively.

To better visualize the relationship between mother's occupation and high school results, we add marginal density plots using the ggMarginal() function. Finally, we save the plot as a png file using the ggsave() function with the desired file path, width, height, and resolution.

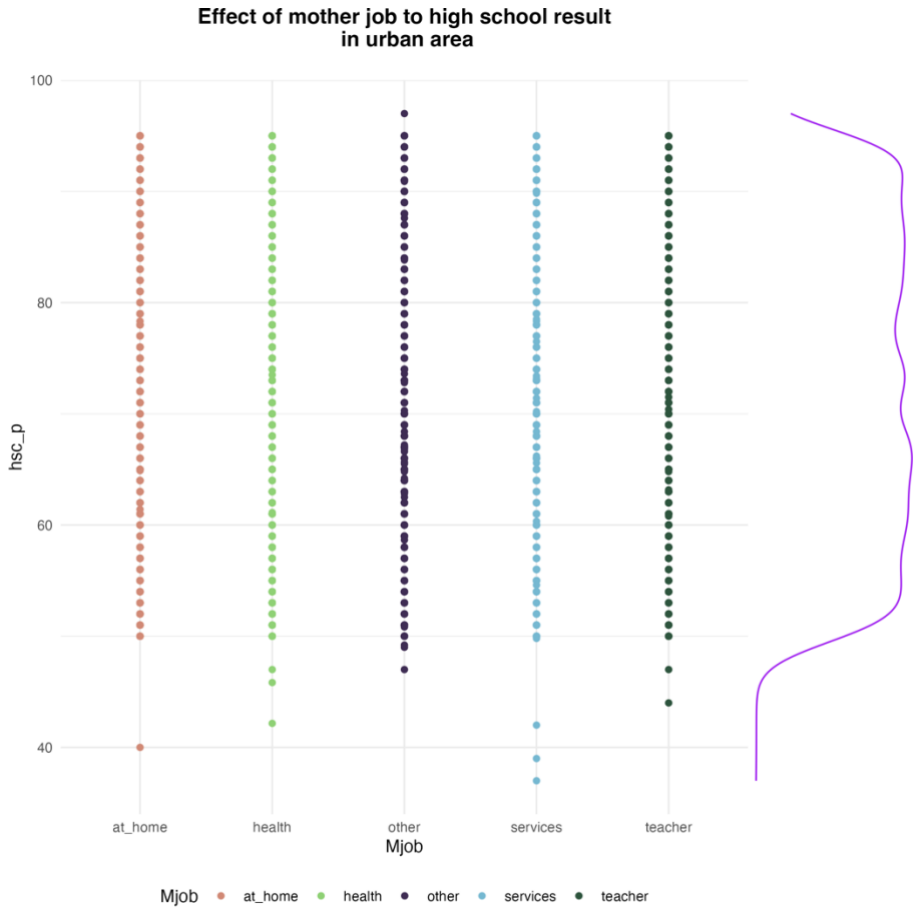


Figure 2.26.2 Relationship between mother's job and high school results in urban areas graph

The resulting diagram with the distribution of students' secondary school results shows no significant correlation with their mothers' occupations in urban areas. This suggests that there is no clear relationship between the two variables in this context, indicating that the mother's occupation may not be a critical factor in determining the academic performance of students in urban areas.

One possible explanation for this could be other factors such as student's personal motivation, access to educational resources, and the quality of teaching in schools may have a more significant impact on students' academic performance in urban areas than the occupation of their mothers alone. Additionally, it is possible that the sample size or study design may have limited the ability to detect a significant relationship between the variables.

In summary, while the analysis shows no significant relationship between mothers' occupations and their children's high school results in urban areas, this does not discount the potential influence of other factors on academic performance.

Analysis 2.27: Does the mother's occupation play a critical role in the children's MBA results in rural areas?

```
#====Analysis 2.27: Does the mother's occupation play a critical role in the children's MBA results in rural areas?====
dMjob <- dfrural %>% select(address, Mjob, degree_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=degree_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of Mother job to degree result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))

g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p21.png", plot=g,width=8,height=8,dpi=300)
```

Figure 2.27.1 Relationship between mother's job MBA test result in rural areas code

For the first line of the code, we create a new data frame called "dMjob" by selecting three columns from the original data frame "dfrural": "address" (which indicates whether the student is from a rural or urban area), "Mjob" (which indicates the mother's occupation), and "degree_p" (which indicates the student's degree result). Then, the "group_by" function groups the data by address and Mjob, creating subsets of the data for each combination of these two variables.

The next lines of the code create a scatter plot called "g" using the "ggplot" function from the "ggplot2" package in R. The x-axis represents the mother's occupation, the y-axis represents the student's degree result, and the color of the points represents the mother's occupation. The "scale_color_ipsum" and "scale_fill_ipsum" functions are used to set the color palette of the plot, while the "theme_minimal" function sets the overall style of the plot. The "ggtitle" function adds a title to the plot, and the "theme" function sets the position of the legend and the appearance of the plot title and background.

The "ggMarginal" function from the "ggExtra" package is then used to add marginal density plots to the top and bottom of the main plot. Finally, the "ggsave" function saves the plot to a file named "2p21.png" in the specified directory.

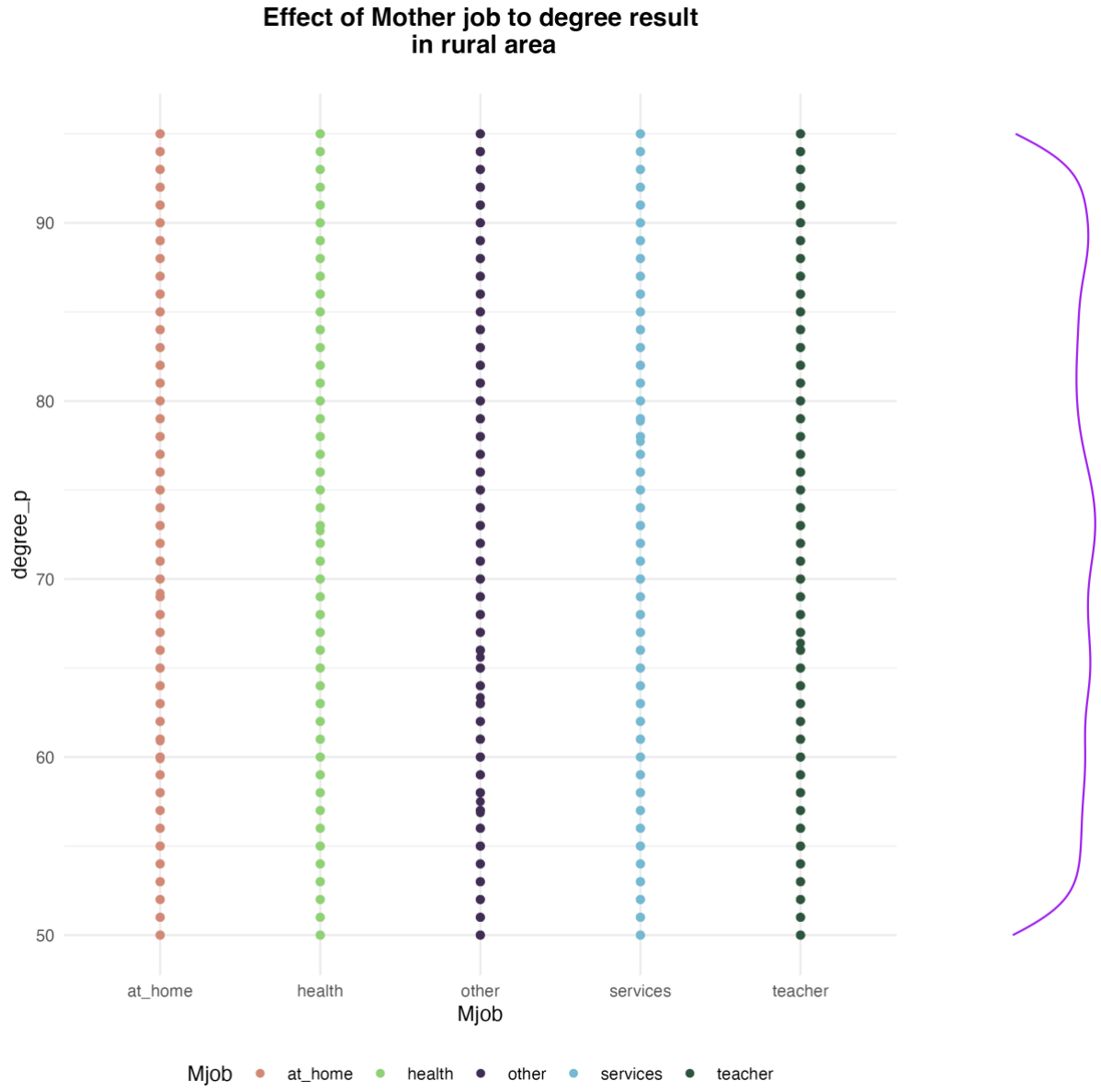


Figure 2.27.2 Relationship between mother's job MBA test result in rural areas graph

The resulting graph showing the distribution of MBA subject results in urban areas indicates no clear correlation between students' academic performance and their mothers' occupations. This suggests that the mother's profession may not be a significant factor in determining their children's performance in MBA subjects in urban areas.

One explanation for this could be those other factors, such as students' individual effort, access to educational resources, and quality of teaching, may have a more significant impact on MBA subject performance in urban areas than the mother's occupation alone. Additionally, it is possible that the sample size or study design may have limited the detection of a significant relationship between the variables.

Analysis 2.28: Does the mother’s occupation play a critical role in the children’s degree results in urban areas?

```
#====Analysis 2.28: Does the mother’s occupation play a critical role in the children’s degree results in urban areas?====
dMjob <- dfurban %>% select(address, Mjob, degree_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=degree_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to degree result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))

g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p22.png", plot=g, width=8,height = 8,dpi=300)
```

Figure 2.28.1 Relationship between mother's job MBA test result in urban areas code

Firstly, selects the "address", "Mjob", and "degree_p" columns from the dataset and groups them by the address and mother's occupation.

Next, it creates a scatter plot using ggplot with "Mjob" on the x-axis, "degree_p" on the y-axis, and the color of the points representing different types of jobs. It also sets a title and adjusts the plot's theme.

Then adds marginal density plots using the ggMarginal function to show the distribution of "degree_p" values on the y-axis. Finally, the plot is saved as a png file in the specified directory.

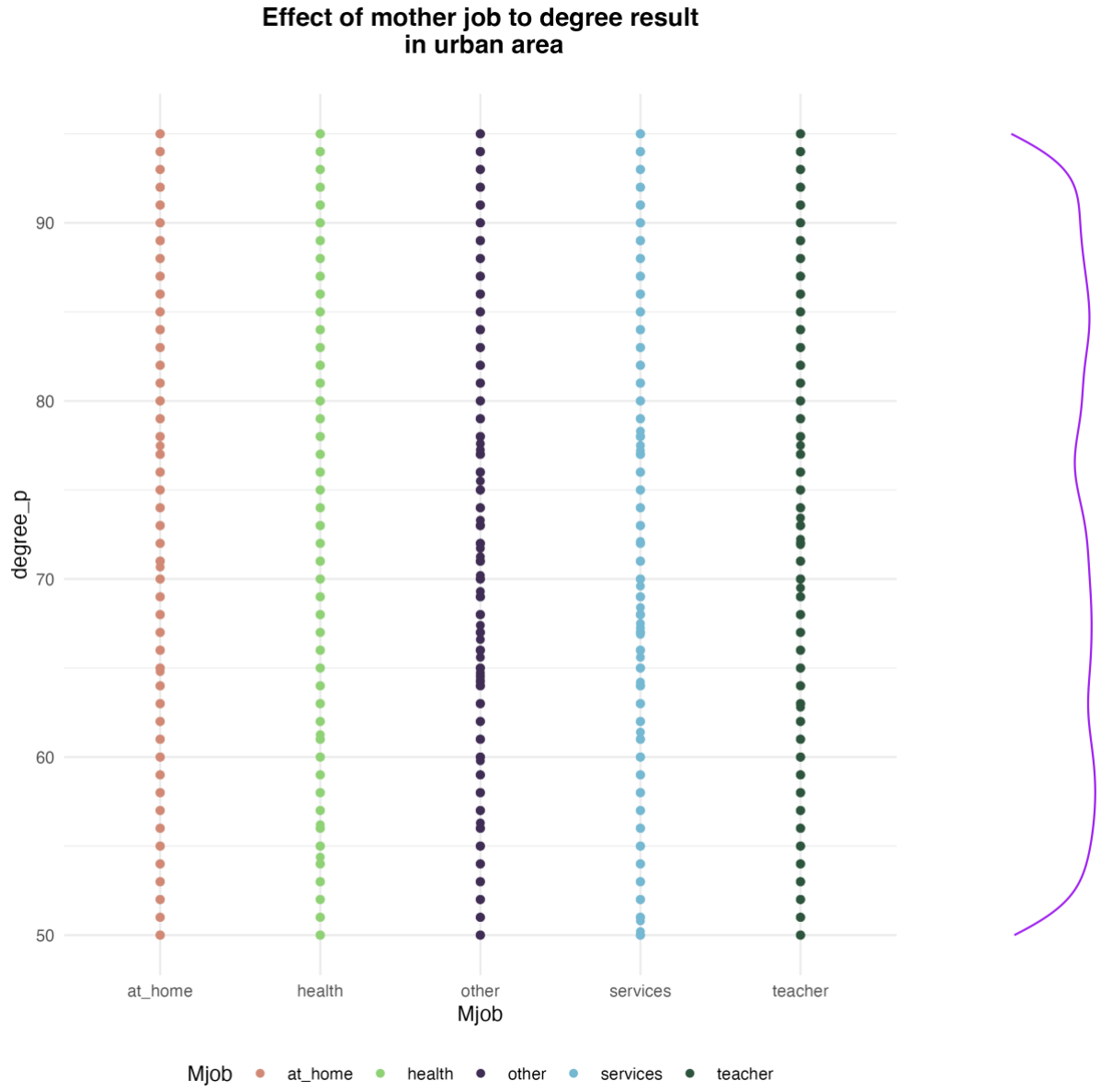


Figure 2.28.2 Relationship between mother's job MBA test result in urban areas graph

The resulting data analysis shows no significant correlation between mothers' occupations and their children's degree results in urban areas. This suggests that the mother's occupation may not play a critical role in determining the academic success of their children in this context. One possible explanation could be that other factors such as the quality of higher education institutions, the field of study chosen by the students, and their personal drive and ambition may have a greater impact on their academic performance in achieving a degree.

In summary, while the analysis shows no significant relationship between mothers' occupations and their children's degree results in urban areas, this does not rule out the potential impact of other factors.

Analysis 2.29: Does the mother's occupation play a critical role in the children's employability test results in rural areas?

```
#====Analysis 2.29: Does the mother's occupation play a critical role in the children's employability test results in rural areas?====
dMjob <- dfrural %>% select(address, Mjob, etest_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=etest_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to employability test result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p23.png", plot=g)
```

Figure 2.29.1 Relationship between mother's job employability test results in rural areas code

In this analysis, we are investigating whether there is a relationship between the mother's occupation and children's employability test results in rural areas. To do so, we first select the relevant columns from our dataset using the select() function, which selects the address, mother's occupation (Mjob), and employability test results (etest_p) columns from the rural dataset (dfrural). We then group the data by address and mother's occupation using the group_by() function and store the result in the dMjob object.

Next, we create a scatter plot using the ggplot() function, where the mother's occupation is on the x-axis, employability test results are on the y-axis, and the color of the points represents the mother's occupation. We use geom_point() to add the data points to the plot, and scale_color_ipsum() and scale_fill_ipsum() to add color to the plot. We also add a title and adjust the theme of the plot using various theme functions.

To further explore the relationship between the mother's occupation and employability test results, we add marginal density plots to both sides of the scatter plot using the ggMarginal() function. Finally, we save the plot as a PNG file using the ggsave() function.

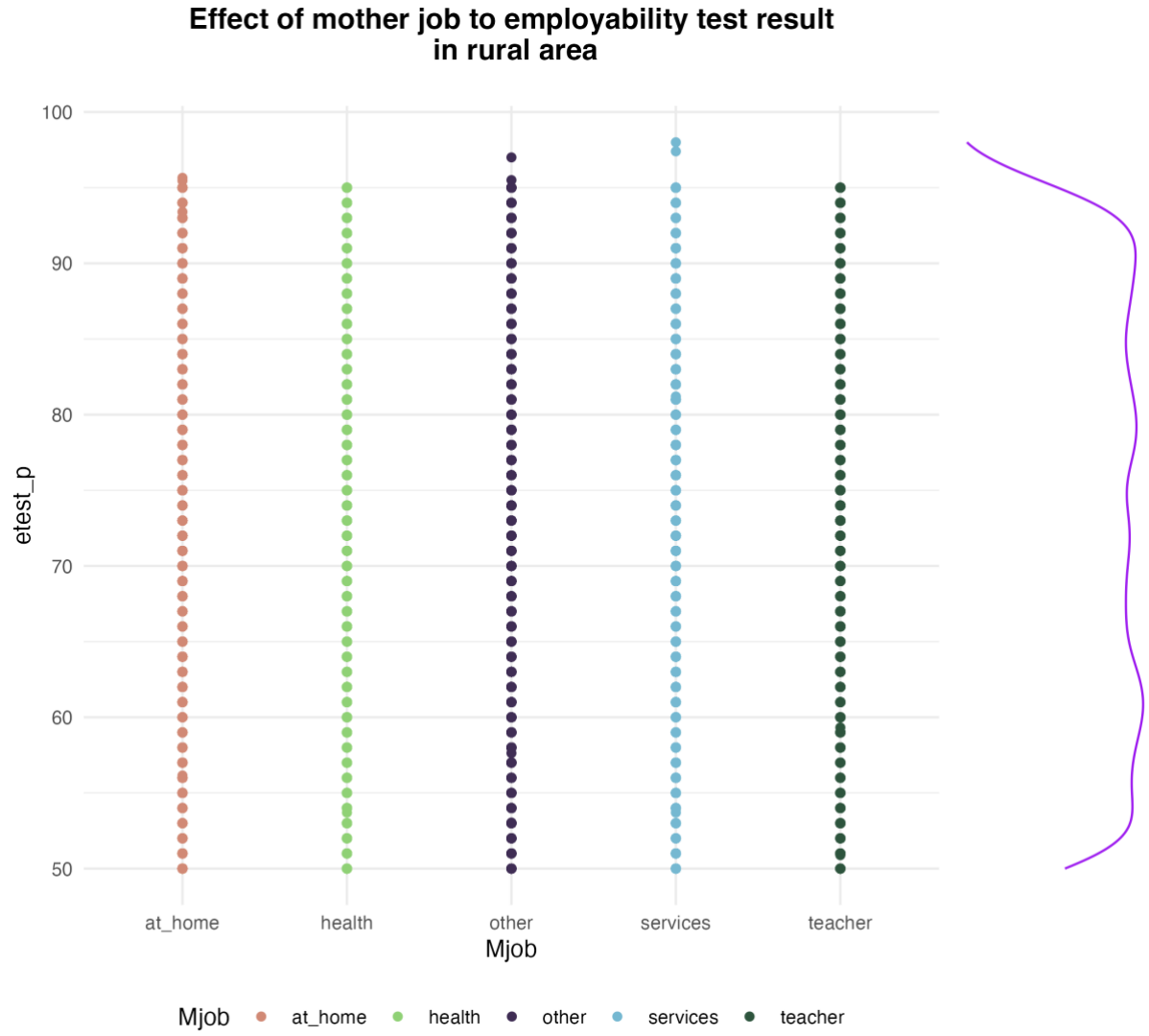


Figure 2.29.2 Relationship between mother's job employability test result in rural areas graph

When examining the relationship between mothers' occupations and children's employability test results in rural areas, it is important to consider the potential impact of contextual factors. In rural areas, job opportunities may be limited, and access to education and training may be restricted, which could have significant implications for the employability of students. However, the analysis here has not displayed any trend between the mother's job and their children's employability results in rural areas.

However, the unique circumstances of rural communities underscore the importance of considering the broader social, economic, and cultural context when examining the role of various factors in determining the relationship between a mother's occupation and career outcomes. Henceforth, further analysis should be considered to examine this relationship in a proper way.

Analysis 2.30: Does the mother's occupation play a critical role in the children's employability test results in urban areas?

```
#====Analysis 2.30: Does the mother's occupation play a critical role in the children's employability test results in urban areas?====
dMjob <- dfurban %>% select(address, Mjob, etest_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=etest_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to employability test result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <-ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p24.png", plot = g)
```

Figure 2.30.1 Relationship between mother's job employability test result in urban areas code

In this analysis, we are exploring whether the mother's occupation has a critical role in the children's employability test results in urban areas. We start by selecting the relevant columns from the "dfurban" dataset using the "select" function and grouping the data by the mother's occupation and address using the "group_by" function.

Then, we create a scatterplot using the "ggplot" function, where we set the x-axis to the mother's occupation, y-axis to the employability test results, and color of the points to represent different mother occupations. We also add a title and adjust the theme of the plot using the "ggtitle" and "theme" functions, respectively.

Finally, we add marginal density plots on the y-axis using the "ggMarginal" function and save the plot as a png file using the "ggsave" function.

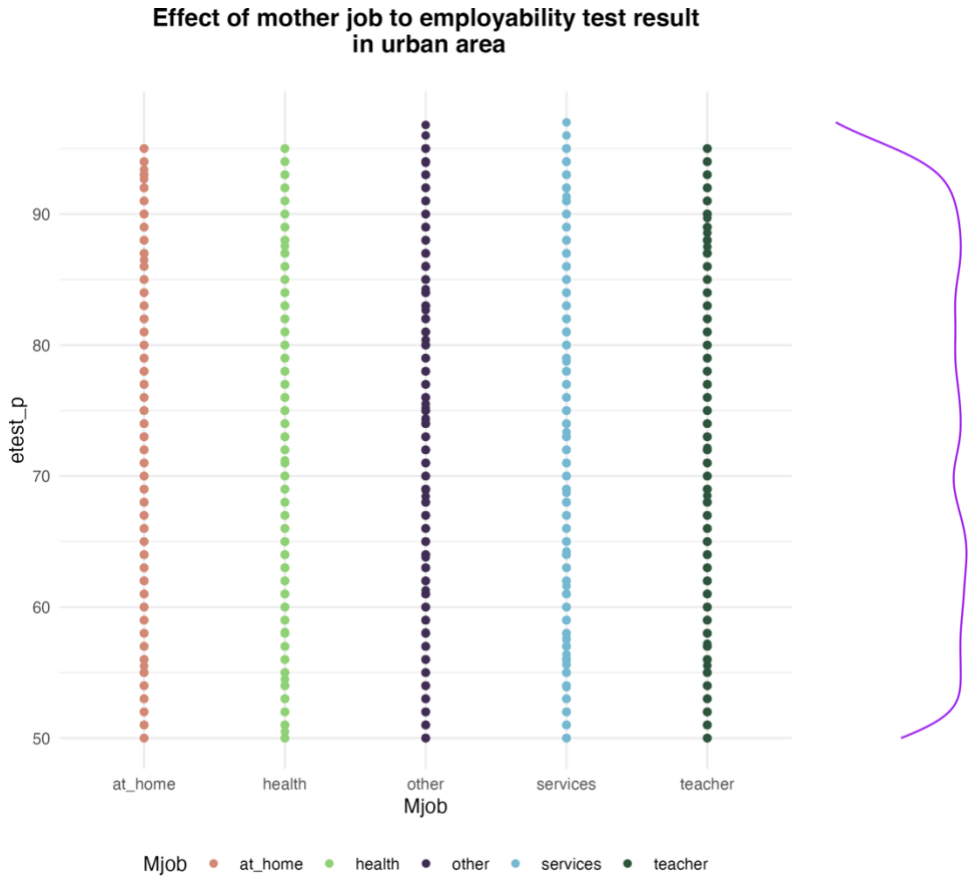


Figure 2.30.2 Relationship between mother's job employability test result in urban areas graph

When it comes to the relationship between mothers' occupations and children's employability test results in urban areas, it is possible that the same factors identified in the rural areas analysis may also be relevant. Access to educational resources, quality of teaching, and personal motivation may all play important roles in determining employability outcomes, regardless of the urban or rural context.

In summary, the relationship between a mother's occupation and children's employability test results in urban areas is likely influenced by a range of factors, including access to resources and job opportunities. Further research would be necessary to fully understand the nature of this relationship and its unique characteristics in urban areas.

Analysis 2.31: Does the mother's occupation play a critical role in the children's MBA results in rural areas?

```
#====Analysis 2.31: Does the mother's occupation play a critical role in the children's MBA results in rural areas?====
dMjob <- dfrural %>% select(address, Mjob, mba_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=mba_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to MBA result
 in rural area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)
g
ggsave("~/Graph/2p25.png", plot = g)
```

Figure 2.31.1 Relationship between mother's job MBA result in rural areas code

To analyze the relationship between a mother's occupation and her children's MBA results in rural areas, I used the dplyr package to select the relevant variables from my rural dataset (dfrural), including the mother's occupation (Mjob) and her child's MBA score (mba_p).

Then, I grouped the data by both address and mother's occupation using the group_by function, which allowed me to create a plot that shows the distribution of MBA scores by mother's occupation.

I used ggplot2 to create the plot, which includes points colored by the mother's occupation and a title that explains the purpose of the plot. I also added theme elements to make the plot more readable, such as a minimal theme and a legend at the bottom.

Finally, I used the ggMarginal function from the ggExtra package to add marginal density plots that show the distribution of MBA scores and the mother's occupation separately. I saved the plot as a PNG file using the ggsave function.

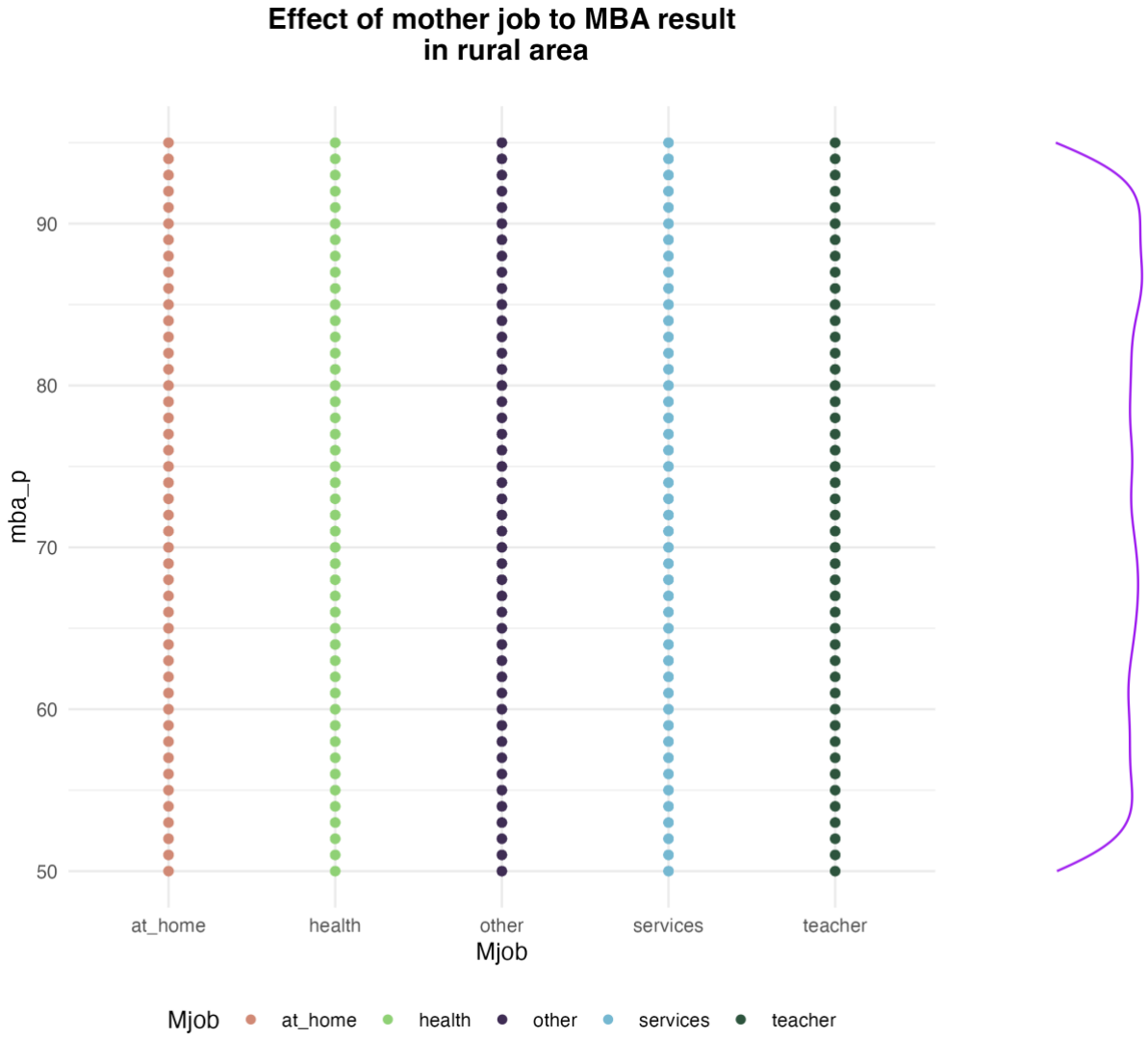


Figure 2.31.2 Relationship between mother's job MBA result in rural areas graph

When it comes to the relationship between a mother's occupation and children's MBA results in rural areas, the nature of this relationship has no difference with urban areas due to a range of contextual factors.

However, further research would be necessary to understand the specific factors that contribute to the relationship between a mother's occupation and children's MBA results in rural areas. The relationship only highlights the importance of considering the unique circumstances of each community when analyzing the role of various factors in determining academic performance.

Analysis 2.32: Does the mother's occupation play a critical role in the children's MBA results in urban areas?

```
#====Analysis 2.32: Does the mother's occupation play a critical role in the children's MBA results in urban areas?====
dMjob <- dfurban %>% select(address, Mjob, mba_p) %>% group_by(address, Mjob)
dMjob
g <- ggplot(dMjob, aes(x=Mjob, y=mba_p, color=Mjob)) + geom_point() +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Effect of mother job to MBA result
 in urban area")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
g <- ggMarginal(g, type="density", margins = 'y', color="purple", size=4)#Extra features
g
ggsave("~/Graph/2p26.png", plot=g)
```

Figure 2.32.1 Relationship between mother's job MBA result in urban areas code

I wrote this code to explore the relationship between the mother's occupations and their children's MBA results in urban areas. To do this, I selected the relevant columns from the urban area dataset and grouped the data by the address and mother's job. Then, I created a scatter plot with the mother's occupation on the x-axis and MBA results on the y-axis, using different colors for each job category. I added some extra features like color scheme, theme, and title to make the plot more informative and visually appealing. Finally, I used the ggMarginal function to add density plots on the y-axis margins to show the distribution of MBA results. I saved the plot as a PNG file for future reference.

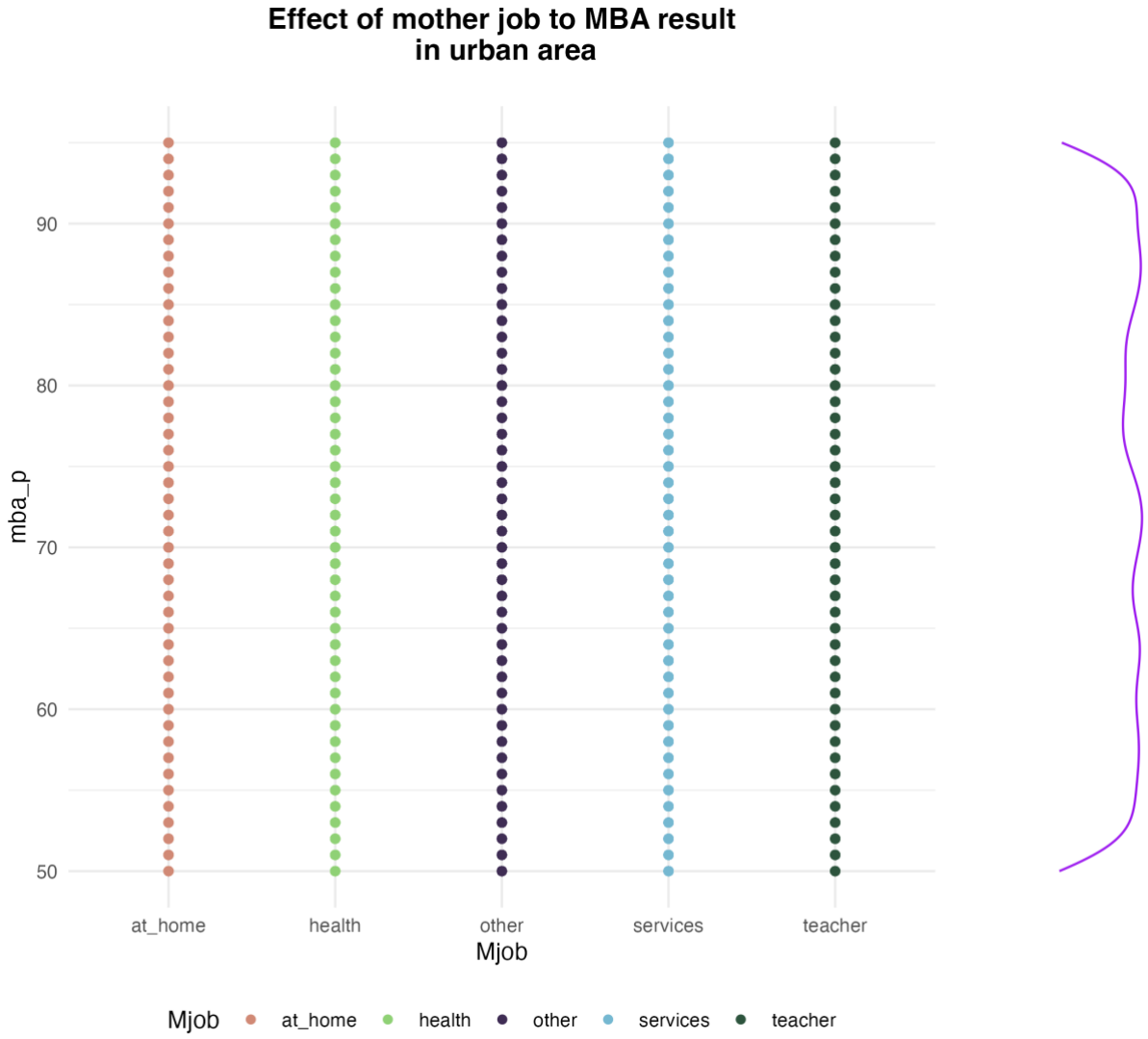


Figure 2.32.2 Relationship between mother's job MBA result in urban areas graph

When we looked at the distribution of MBA scores among students from metropolitan areas, we studied the association between mothers' occupations and results and found no evidence of a significant connection. This shows that a student's academic success in MBA programs in urban regions may not be significantly influenced by the mother's profession.

One reason for this finding is that other variables, besides the mother's work alone, maybe more important in influencing MBA performance in metropolitan regions than the quality of schooling, access to resources, and personal study habits. Also, it's possible that the study's sample size or methodology prevented the discovery of a substantial correlation between the variables.

Summary for Question 2

We have analyzed all the attributes that could have a relationship with family support in this dataset. However, family support is not actually an indicator that will be affecting one's test result, or income in both rural and urban areas. Whether the students are employed or not, they are not affected by their family support. This comes to the conclusion that students who have no family support can also be having success.

Question 3: Does gender discrimination occur in the different areas in the dataset?

Analysis 3.1: Do these students' parents in rural areas get equal chances of having an education during their education?

```
#====Analysis 3.1:  Do these students' parents in rural areas get equal chances of having an education during their education?====
dfedu <- dfrural %>% select(Fedu_level, Medu_level) %>%
  pivot_longer(cols = c(Medu_level, Fedu_level), names_to = "education", values_to = "education_level") %>%
  group_by(education) %>% count(education_level) %>%
  mutate(percent=round(n/sum(n), 2)) %>% mutate(cumulative=cumsum(percent)) %>%
  mutate(ymin = cumulative - percent)
dfedu
ggplot(dfedu, aes(ymax = cumulative, ymin = ymin, xmax = 4, xmin=3, fill=education_level)) +
  geom_rect(color = "white") +
  geom_label(x = 3.5, aes(y = (cumulative + ymin)/2), label=paste0(dfedu$percent*100, "%"),
             size = 3) +
  coord_polar(theta = "y")+
  facet_grid(~education) +
  labs(fill = "Education Level", title = "Education of Parents in rural areas") + xlim(2,4) +
  scale_color_ipsum() + scale_fill_ipsum()+ theme_void()+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p1.png")
```

Figure 3.1.1 Relationship between Father education and Mother education in rural areas code

I started by selecting the columns 'Fedu_level' and 'Medu_level' from the rural dataframe 'dfrural'. Then, I used the 'pivot_longer' function to gather these two columns into one column named 'education' and another column named 'education_level' to store the education level of the parents.

Next, I grouped the data by 'education' and 'education_level', counted the number of parents with each education level and calculated the percentage of parents in each category using 'mutate'. I also calculated the cumulative percentage for each category using 'cumsum' and the 'ymin' for each category using the 'mutate' function.

Then, I created a polar plot using 'ggplot' with the 'geom_rect' function to create rectangles for each category based on their percentage. I used 'geom_label' to add the percentage labels on each rectangle. I used 'facet_grid' to display separate polar plots for 'Fedu_level' and 'Medu_level'. I also used 'scale_fill_ipsum' and 'scale_color_ipsum' for aesthetic purposes.

Finally, I saved the plot as a png file in my local directory using the 'ggsave' function.

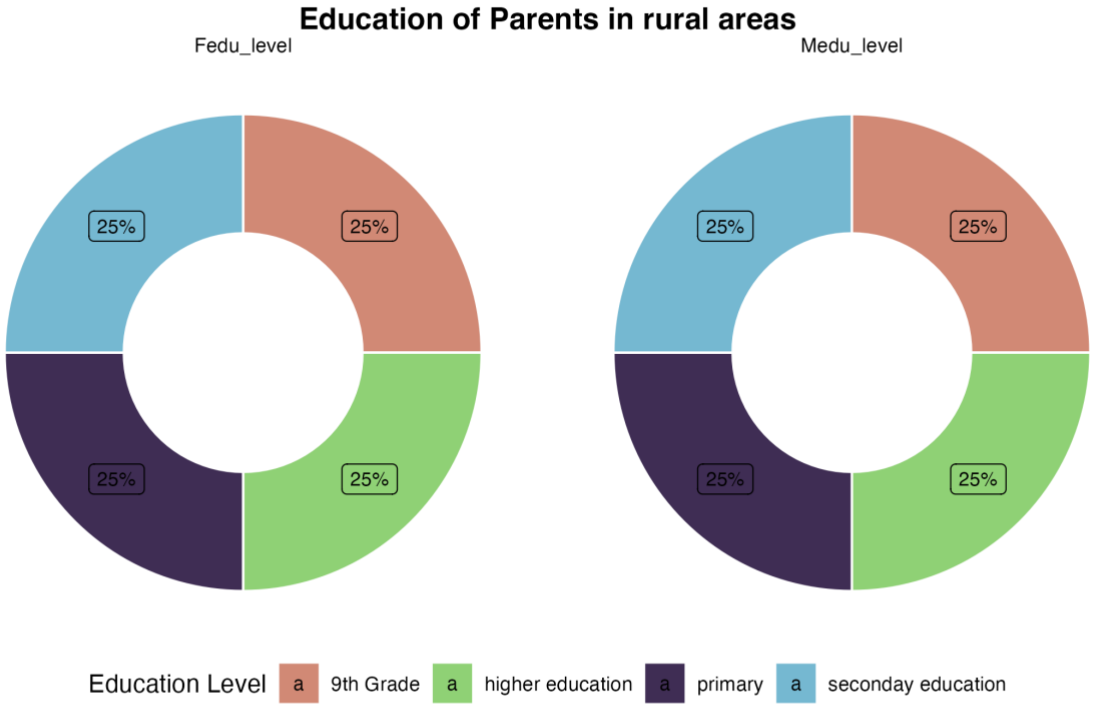


Figure 3.1.2 Relationship between Father education and Mother education in rural areas graph

The doughnut chart shows the education level of rural areas students' parents categorized into two types: "Fedu_level"(Father education) and "Medu_level"(Mother education). For "Fedu_level", the number of parents who have primary education, 9th grade, secondary education, and higher education is 25% equally. For "Medu_level", the number of parents who have secondary education, primary education, higher education, and 9th grade are also the same. In summary, most parents of the students in this dataset have either primary, secondary education, 9th Grade, or higher education, with a relatively equal distribution between fathers and mothers in the past.

Analysis 3.2: Do these students' parents in urban areas get equal opportunities in getting an education during their studying?

```
#====Analysis 3.2:  Do these students' parents in urban areas get equal opportunities in getting an education during their studying?====
dfedu <- dfurban %>% select(Fedu_level, Medu_level) %>%
  pivot_longer(cols = c(Medu_level, Fedu_level), names_to = "education",
               values_to = "education_level") %>%
  group_by(education) %>% count(education_level) %>%
  mutate(percent=round(n/sum(n), 2)) %>% mutate(cumulative=cumsum(percent)) %>%
  mutate(ymin = cumulative - percent)
dfedu
ggplot(dfedu, aes(ymax = cumulative, ymin = ymin, xmax = 4, xmin=3, fill=education_level)) +
  geom_rect(color = "white") +
  geom_label(x = 3.5, aes(y = (cumulative + ymin)/2), label=paste0(dfedu$percent*100, "%"),
             size = 3) +
  coord_polar(theta = "y")+
  facet_grid(~education) + theme_void() +
  labs(fill = "Education Level", title = "Education of Parents in urban areas") + xlim(2,4) +
  scale_color_ipsum() + scale_fill_ipsum() + theme_void()+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p2p1.png")
```

Figure 3.2.1 Relationship between Father education and Mother education in urban areas code

In this analysis, we look at whether urban parents have an equal opportunity to pursue an education throughout their children's schooling. We pivot the data in order to have one column for the type of education (father or mother) and one column for the education level in order to accomplish this. To start, we choose the columns that contain the educational backgrounds of both parents. The following steps are to organise the data by education type and level, count the instances, determine the percentage of each education level, and compute the cumulative percentage.

Next, for each education level, a rectangle is made on a polar plot, with the height of the rectangle being proportionate to the overall percentage of that level. Also, we include a label with the proportion of that education level for each rectangle. The final narrative is segmented according to the educational background (father's or mother's) and is given the proper titles and legends. In general, this methodology enables us to assess whether urban parents have equal opportunity to pursue an education throughout their children's schooling.

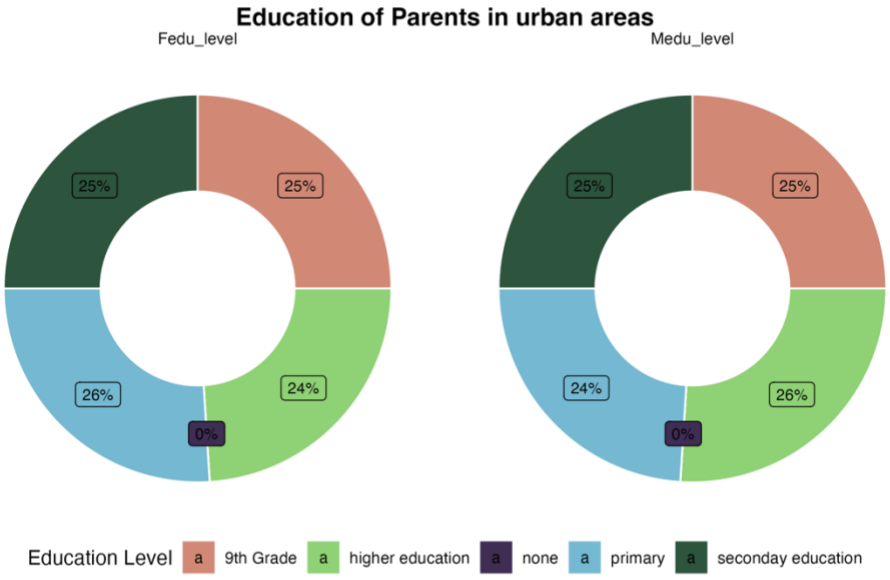


Figure 3.2.2 Relationship between Father education and Mother education in urban areas graph

The diagram above provides information on the education level of urban areas parents of students. The table is divided into two sections based on the mother's education level (Medu_level) and the father's education level (Fedu_level). The education levels range from 9th grade to higher education. The table also includes a category for "none" which represents parents who did not receive any formal education. The percentage of parents falling into each category is also provided.

For both Medu_level and Fedu_level, the largest percentage of parents received a primary education (24% and 26%, respectively). The second largest group for both is parents with secondary education (25% and 25%, respectively). The smallest group for both is parents with no formal education, which makes up less than 1% of the total.

In summary, the doughnut chart indicates that the majority of parents of students have received an education. The percentage of parents who have not received an education is relatively low. However, no matter whether it is the father or mother, they have equal opportunity in getting an education at their age.

Summary 3.1 and 3.2: Education equality for students' parents in the past time

```
#====Summary 3.1 and 3.2: Education equality for students' parents in the past time====
dfedu <- df %>% select(address, Medu, Fedu) %>% pivot_longer(cols = c(Medu, Fedu)
                                                             , names_to="parents", values_to = "education")  %>%
  group_by(address, parents, education) %>% summarize(count = n())
ggplot(dfedu, aes(x=education, y=count, fill=parents)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_viridis(discrete = T, name="") + facet_grid(~address) +
  coord_flip() + ggtitle("Parents education in different areas")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p2p2.png")

dfedu <- df %>% select(address, Fedu) %>% pivot_longer(cols = -address,
                                                       names_to="Father_education", values_to = "value") %>%
  group_by(address, Father_education, value) %>% count(value)
ggplot(dfedu, aes(x=value, y=n, fill=address)) + geom_bar(stat = "identity", position = "dodge") +
  xlab("Father education") + ylab("count") +
  ggtitle("Father education in different area") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p2p3.png")

dfedu <- df %>% select(address, Medu) %>%
  pivot_longer(cols = -address, names_to="Mother_education", values_to = "value") %>%
  group_by(address, Mother_education, value) %>% count(value)
ggplot(dfedu, aes(x=value, y=n, fill=address)) + geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Mother education in different area") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p2p4.png")
```

Figure Summary 0.1 Summary for 3.1 and 3.2 code

The summary code is using to analyze the education equality for students' parents in the past time. We first create a new dataframe (dfedu) by selecting the address, mother's education level (Medu), and father's education level (Fedu) columns from the original dataset. Then we use the pivot_longer function to gather the education levels of both parents into a single column. Next, we group the data by address, parents (i.e., father or mother), and education level to count the number of occurrences for each combination.

We then use ggplot to create a stacked bar chart that shows the count of parents' education levels in different areas (urban and rural). We use fill to differentiate between mother's and father's education levels, and we use facet_grid to separate the plots for urban and rural areas.

We also create two additional graphs to show the count of father's education level and mother's education level separately in different areas. For each graph, we use the pivot_longer function to gather the education level columns and count the number of occurrences for each value. We then use ggplot to create a bar chart that shows the count of education levels in different areas, with different colors representing different areas. We use xlab and ylab to label the x-axis and y-axis, respectively, and ggtitle to add a title to each

plot. We use scale_color_ipsum and scale_fill_ipsum to set the color palette, and theme_minimal to adjust the overall look of the plot. Finally, we use ggsave to save the graphs as png files.

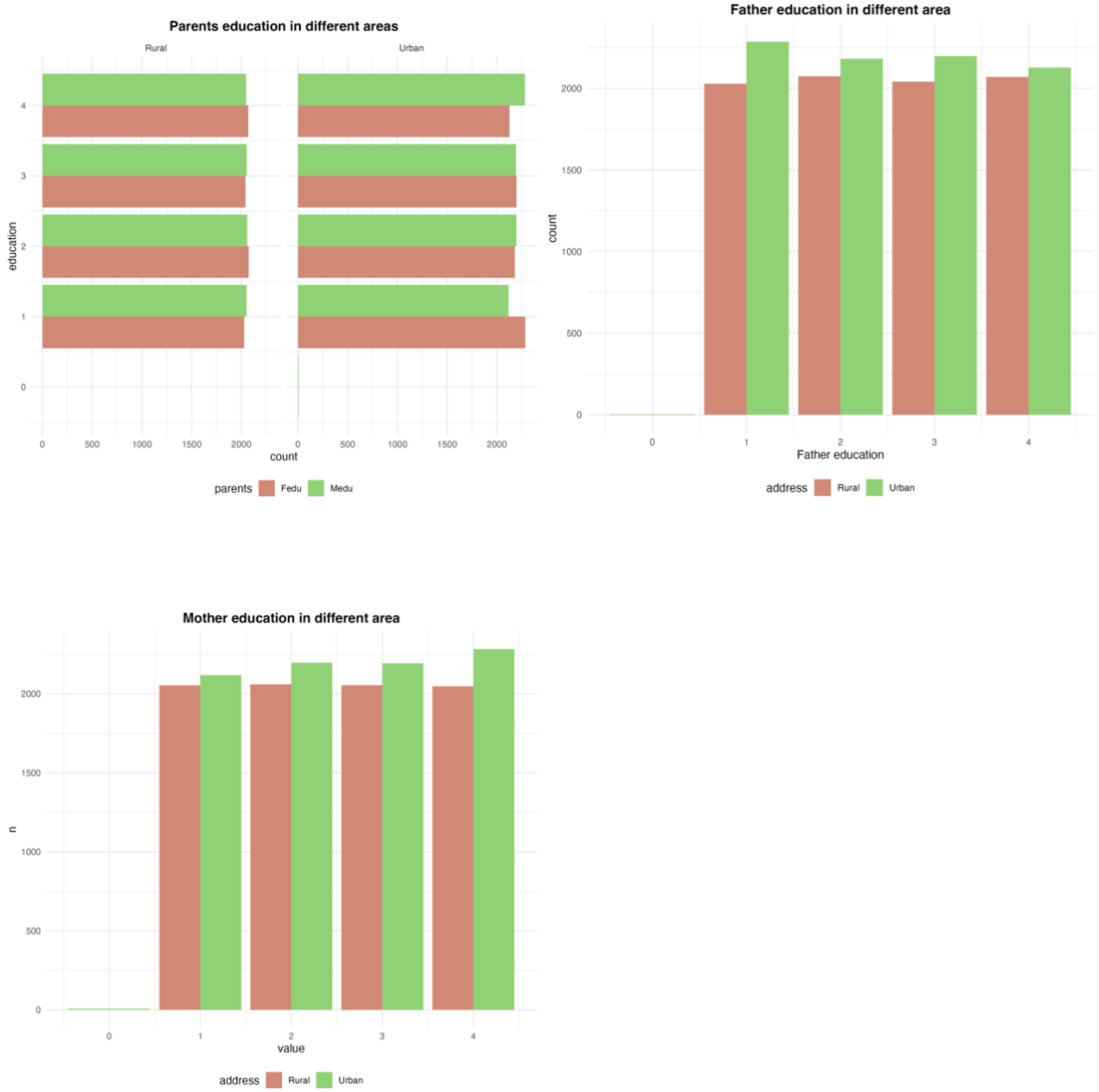


Figure Summary 0.2 Summary for 3.1 and 3.2 graph

Examining the parent's education equality is helpful in helping researchers to know the situation of education for children in the past time through indirect methods. After making a conclusion about the education equality of parents in the past in sections 3.1 and 3.2, we draw the graph using the bar chart to have a better understanding of the equal distribution and similar education percentage that we found by using the doughnut chart in the previous section. All of the difference between the education of father and mother is controlled in

around 500, which symbolizes that the education between gender in the past time has shown no discrimination whether the children are male or female.

Analysis 3.3: Is there any stereotype for parents in choosing occupations in rural areas?

```
#====Analysis 3.3:  Is there any stereotype for parents in choosing occupations in rural areas?====
df_gender <- dfrural %>% select(Fjob, Mjob) %>%
  pivot_longer(cols = c(Fjob,Mjob), names_to = "parents", values_to = "job")  %>% group_by(parents) %>%
  count(job) %>% mutate(percent=round(n/sum(n), 5)) %>% mutate(cumulative=cumsum(percent)) %>%
  mutate(ymin = cumulative - percent)
df_gender
ggplot(df_gender, aes(ymax = cumulative, ymin = ymin, xmax = 4, xmin=3, fill=job)) +
  geom_rect(color = "white") +
  geom_label(x = 3.5, aes(y = (cumulative + ymin)/2), label=paste0(df_gender$percent*100, "%"),
             size = 3) +
  coord_polar(theta = "y")+
  facet_grid(~parents) + theme_void() +
  labs(fill = "Job", title = "Job of Parents in rural area") + xlim(2,4) +
  scale_color_ipsum() + scale_fill_ipsum() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p3.png")
```

Figure 3.3.1 Relationship between Father's job and Mother's job in rural areas code

First, we extract the job information of fathers and mothers from the rural area dataset, and then transform the data to a long format using pivot_longer() function. Next, we group the data by parents and job, and count the number of occurrences for each combination. Then we calculate the percentage of each job category for each parent using mutate() function. We also calculate the cumulative percentage for each job category for each parent.

After that, we add a ymin column to the data frame to indicate the lower bound of each job category in the plot. We then create a polar bar chart for each parent using coord_polar() function. Finally, we add a label showing the percentage of each job category and a legend indicating the different job categories using geom_label() and labs() functions.

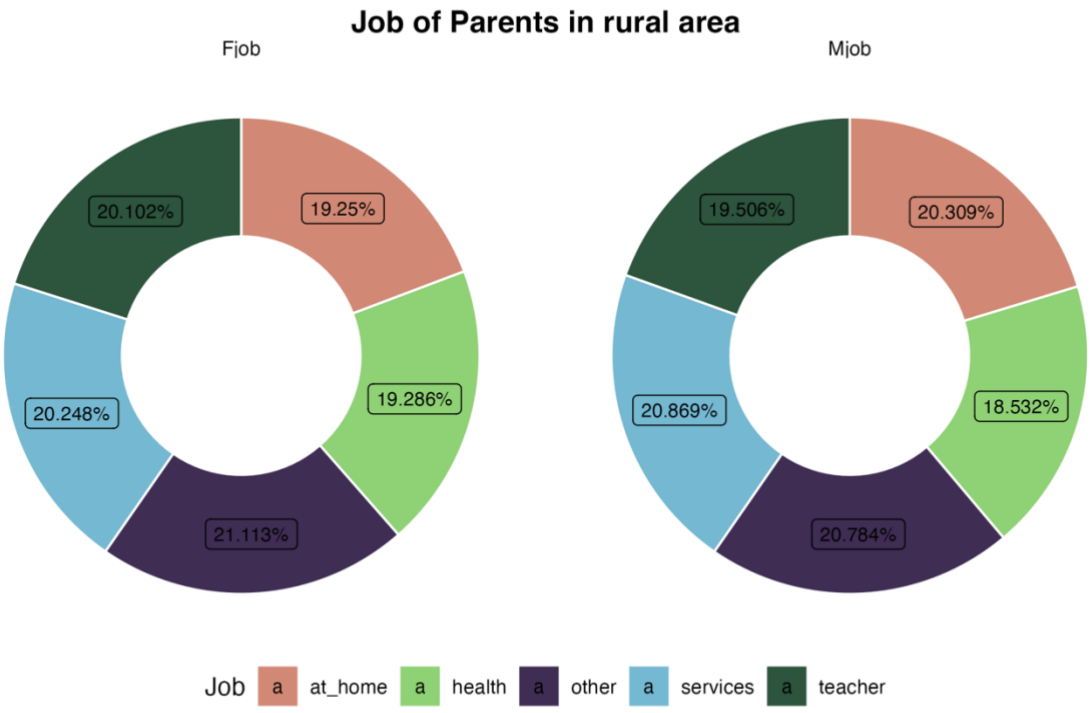


Figure 3.3.2 Relationship between Father's job and Mother's job in rural areas graph

The diagram shows the distribution of parents' jobs among students in the rural area. The percentages indicate that there is only a slight difference between 1 to 2% between the number of fathers and mothers who work in each category, suggesting that gender-based discrimination in career choices did not occur in the past among parents in this community. This could be due to various factors, such as cultural norms that encourage equal opportunities for men and women or limited job opportunities that require individuals to take whatever work is available regardless of gender. In summary, the data suggests that there is no significant gender-based discrimination in the occupational choices of parents in this rural area, which could have positive implications for promoting gender equality and diversity in the community.

Analysis 3.4: Is there any stereotype for parents in choosing occupations in urban areas?

```
#====Analysis 3.4:  Is there any stereotype for parents in choosing occupations in urban areas?====
df_gender <- dfurban %>% select(Fjob, Mjob) %>%
  pivot_longer(cols = c(Fjob,Mjob), names_to = "parents", values_to = "job")  %>% group_by(parents) %>%
  count(job) %>% mutate(percent=round(n/sum(n), 5)) %>% mutate(cumulative=cumsum(percent)) %>%
  mutate(ymin = cumulative - percent)
df_gender
ggplot(df_gender, aes(ymax = cumulative, ymin = ymin, xmax = 4, xmin=3, fill=job)) +
  geom_rect(color = "white") +
  geom_label(x = 3.5, aes(y = (cumulative + ymin)/2), label=paste0(df_gender$percent*100, "%"),
             size = 3) +
  coord_polar(theta = "y")+
  facet_grid(~parents) + theme_void() +
  labs(fill = "Job", title = "Job of Parents in urban areas") + xlim(2,4) +
  scale_color_ipsum() + scale_fill_ipsum() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p4.png")
```

Figure 3.4.1 Relationship between Father's job and Mother's job in urban areas code

In this analysis, we begin by selecting the 'Fjob' and 'Mjob' columns from the 'dfurban' dataframe, which contains data for urban areas. We then use the 'pivot_longer' function to reshape the data so that each row contains a parent (either 'Fjob' or 'Mjob') and the corresponding job. We group the data by parents and jobs, and count the number of occurrences of each job. We then calculate the percentage of each job, the cumulative percentage, and the minimum and maximum y-values for each job.

Next, we create a polar coordinate plot using the 'ggplot' function, where each job is represented by a filled segment of a circle. The x-axis represents the job categories, while the y-axis represents the cumulative percentage of parents with a given job. We use 'geom_rect' to draw the segments, 'geom_label' to add labels to each segment with the corresponding percentage, and 'coord_polar' to ensure the plot is circular. We then facet the plot by parents, add a title and legend, and adjust the appearance of the plot using various 'theme' functions.

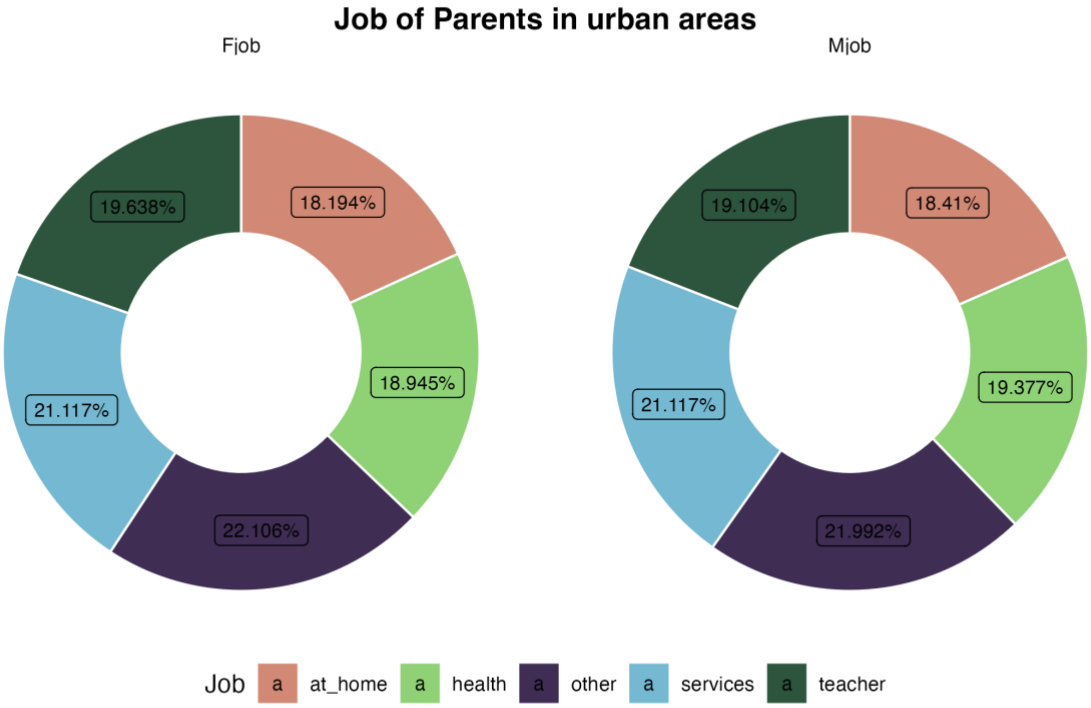


Figure 3.4.2 Relationship between Father's job and Mother's job in urban areas graph

In the urban areas, the distribution of parents' occupations is quite balanced between mothers and fathers, with no significant difference between the number of parents in each occupation category. This suggests that gender discrimination and stereotypes in choosing occupations may be less prevalent in urban areas. A possible reason for this could be greater exposure to a wider range of professions and job opportunities in urban areas. Overall, the data indicates that in both rural and urban areas, there is a relatively equal distribution of parents in different occupations.

Analysis 3.5: Is the student's family have a different attitude in supporting different gender children?

```
#====Analysis 3.5:  Is the student's family have a different attitude in supporting different gender children?=====
df_gender <- df %>% select(gender, address, famsup) %>% group_by(gender, address, famsup) %>% count(famsup)
df_gender
ggplot(df_gender, aes(gender, n, fill=gender)) +
  geom_segment( aes(x=gender, xend=gender, y=0, yend=n), color="grey") +
  geom_point( color="orange", size=4, position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + facet_grid(famsup~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Family support for different gender in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p5.png")
```

Figure 3.5.1 Relationship between Family support and gender code

In this analysis, we are investigating whether the student's family has a different attitude in supporting different gender children in different areas (urban or rural). First, we select the columns 'gender', 'address', and 'famsup' from the original dataframe and group them by 'gender', 'address', and 'famsup', and count the occurrence of 'famsup' for each group. The resulting dataframe is stored in 'df_gender'.

We then create a plot using ggplot that shows the count of 'famsup' for each gender in each area. We use geom_segment to draw vertical lines from the x-axis to the count for each group and geom_point to show the count for each group as a point on the plot. We use position='dodge' to separate the points for each group. We facet the plot by 'famsup' and 'address'. We also add a title and axis labels to the plot and adjust the theme to make it look better. Finally, we use the scale_color_ipsum and scale_fill_ipsum functions to change the colors of the plot.

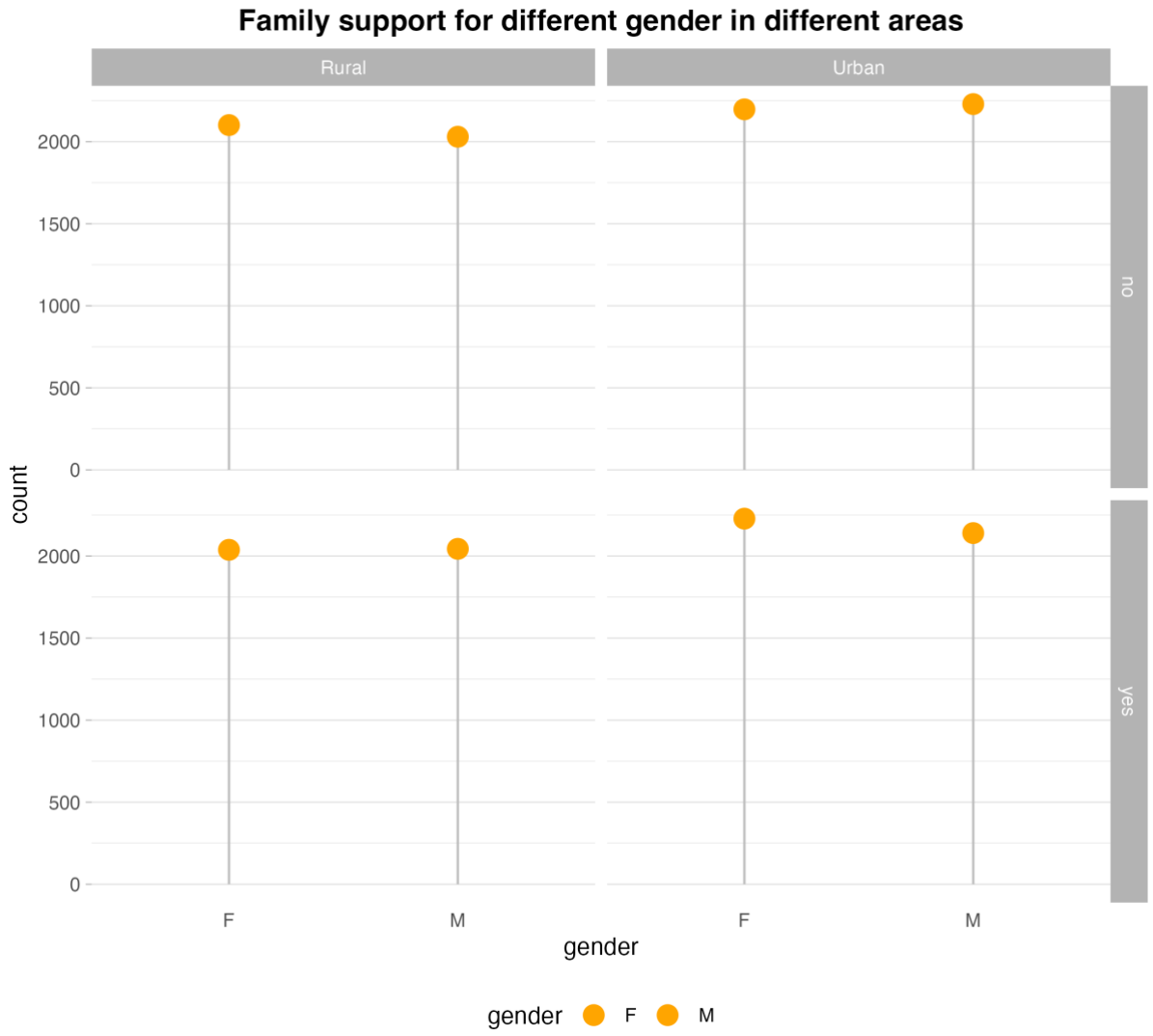


Figure 3.5.2 Relationship between Family support and gender graph

The data shows that there is no significant difference in family support for education between male and female students in rural and urban areas. In both rural and urban areas, there are slightly more females who receive family support compared to males, but the difference is not substantial. One possible reason for this could be that education is becoming increasingly important for both males and females in rural and urban areas, and families recognize the equal value of education in improving their children's future prospects regardless of gender. Another possible reason could be that government policies and programs aimed at promoting gender equality and education have helped to raise awareness and encourage families to support their children's education regardless of gender. Overall, this data suggests that family support for education in rural and urban areas is similar for both male and female students, which is a positive development for promoting education and gender equality.

Analysis 3.6: Are different gender students in different areas having a difference in joining the extra-paid classes?

```
#====Analysis 3.6:  Are different gender students in different areas having a difference in joining the extra-paid classes?====
df_gender <- df %>% select(gender, address, paid) %>% group_by(gender, address, paid) %>% count(paid)
df_gender
ggplot(df_gender, aes(gender, n, fill=gender)) +
  geom_segment( aes(x=gender, xend=gender, y=0, yend=n), color="grey") +
  geom_point(color="orange", size=4, position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + facet_grid(paid~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Extra-paid classes for different gender in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p6.png")
```

Figure 3.6.1 Relationship between address and joining extra-paid classes code

We start by selecting the gender, address, and paid columns from the original dataframe 'df' using the select function. Then we group the resulting dataframe by gender, address, and paid using the group_by function and count the occurrences of the paid column for each group using the count function. We then create a scatterplot using ggplot and map gender to the x-axis, the count of paid to the y-axis, and gender to the fill color. We use geom_segment to draw a vertical line for each gender-group combination, and geom_point to place a point on the line to represent the count of paid. We use facet_grid to create separate plots for each combination of paid and address, and add a y-axis label and title to the plot. Finally, we use the scale_color_ipsum and scale_fill_ipsum functions to adjust the color scheme, and the theme function to modify the appearance of the plot.

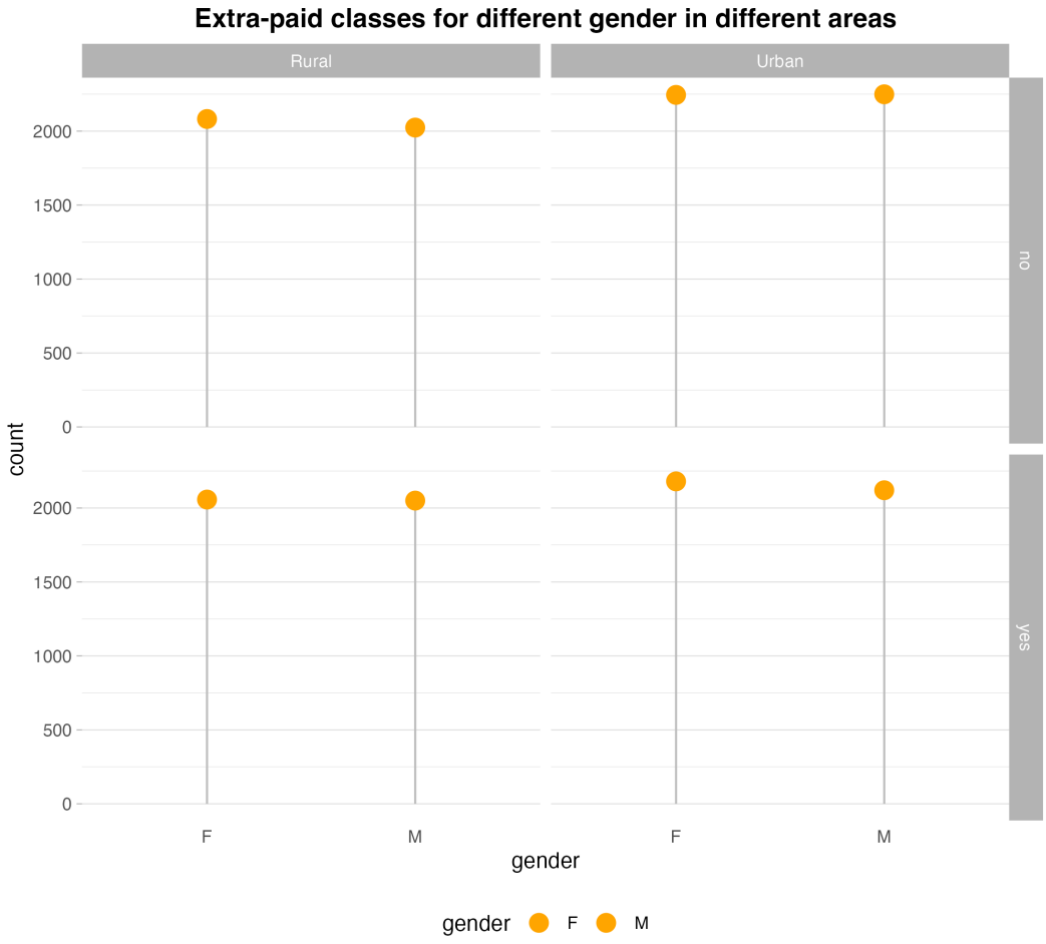


Figure 3.6.2 Relationship between address and joining extra-paid classes graph

Based on the data provided, there is no notable difference in the number of male and female students in rural and urban areas who participate in extra-paid classes. This suggests that both genders are equally encouraged to participate in these classes and there is no discrimination or stereotype present in the dataset.

There could be several possible reasons for this. Perhaps the schools in these areas have policies in place to encourage all students to participate in extra classes regardless of their gender or location. Additionally, parents may also be supportive of their children's education and willing to invest in extra classes regardless of their gender.

In conclusion, the data suggest that gender discrimination and stereotype do not seem to be present in terms of support for education and participation in extra classes in these rural and urban areas. Both male and female students are equally encouraged to participate in extra classes.

Analysis 3.7: Are different gender students in different areas having a difference in accessing the Internet?

```
#====Analysis 3.7:  Are different gender students in different areas having a difference in accessing the Internet?====
df_gender <- df %>% select(gender, address, internet) %>% group_by(gender, address, internet) %>% count(internet)
df_gender
ggplot(df_gender, aes(gender, n, fill=gender)) +
  geom_segment( aes(x=gender, xend=gender, y=0, yend=n), color="grey") +
  geom_point(color="orange", size=4, position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + facet_grid(internet~address) + ylab("count") +
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Internet access for different gender in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p7.png")
```

Figure 3.7.1 Relationship between address and internet accessing code

In this analysis, we are exploring whether different gender students in different areas have a difference in accessing the Internet. To do this, we first create a new dataframe df_gender that includes only the gender, address, and internet columns from the original dataframe df. We then group the data by gender, address, and internet, and count the number of observations for each combination of these variables using the count function.

We then create a scatterplot using ggplot, where the x-axis represents gender, the y-axis represents the count of observations, and the fill color represents gender. We use geom_segment to create vertical lines from the x-axis to the y-axis for each observation, and geom_point to plot the count of observations for each combination of gender, address, and internet. We use facet_grid to create separate plots for each combination of internet and address. We also add labels and formatting to the plot using various theme and ggtitle functions.

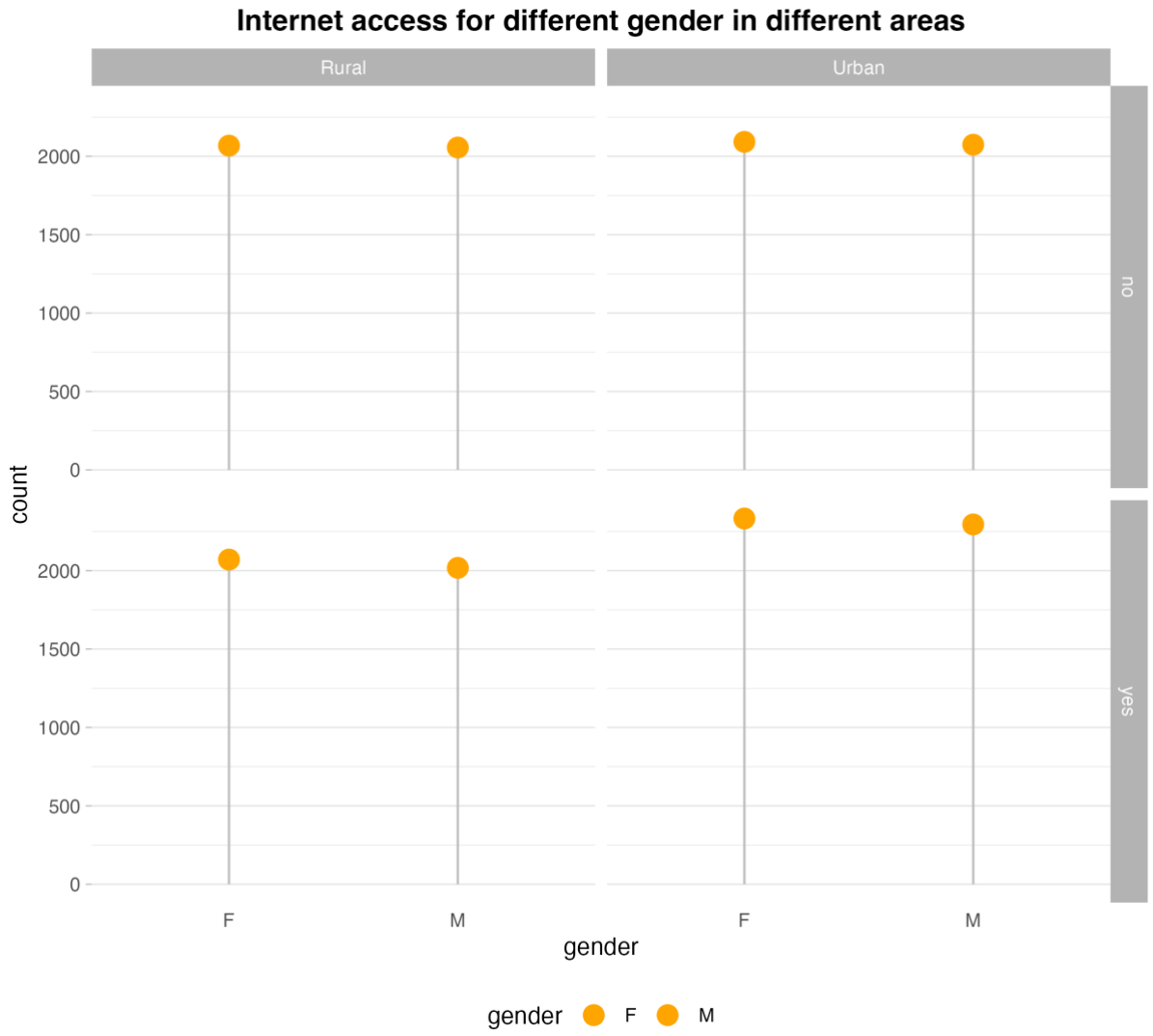


Figure 3.7.2 Relationship between address and internet accessing graph

The data shows that there is no significant difference between male and female students in urban and rural areas regarding internet access. Both genders are equally likely to have access to the internet regardless of where they live. This suggests that there is no gender discrimination or stereotyping in terms of encouraging internet access in this dataset.

One possible reason for this could be that the importance of internet access for education and personal development is widely recognized, and there are efforts to ensure that all students have equal opportunities to access it regardless of their gender or location.

In summary, the data suggest that there is no gender discrimination or stereotype in terms of encouraging internet access in this dataset, and both male and female students in urban and rural areas have equal opportunities to access the internet.

Analysis 3.8: Are different gender students in different areas having a difference in joining the extra-curricular activities?

```
#====Analysis 3.8:  Are different gender students in different areas having a difference in joining the extra-curricular activities?====
df_gender <- df %>% select(gender, address, activities) %>% group_by(gender, address, activities) %>% count(activities)
df_gender
ggplot(df_gender, aes(activities, n, fill=gender)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + ylab("count") + facet_grid(~address)+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Extra-curricular Activities for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p8.png")
```

Figure 3.8.1 Relationship between address and extra-curricular activities code

In this analysis, we want to explore whether different gender students in different areas have a difference in joining extra-curricular activities. To do so, we first create a new data frame called "df_gender" that selects three variables - gender, address, and activities - from our original data frame. We then group the data frame by gender, address, and activities and count the number of observations for each combination of variables.

Next, we use ggplot to create a bar chart that shows the count of students in each category of extra-curricular activities for each gender, with the bars for different genders positioned side-by-side using the "position = 'dodge'" argument. We also use the "facet_grid" function to create separate panels for each area, and the "scale_color_ipsum" and "scale_fill_ipsum" functions to set the color scheme. Finally, we add some cosmetic adjustments to the plot such as the plot title, axis labels, and background color.

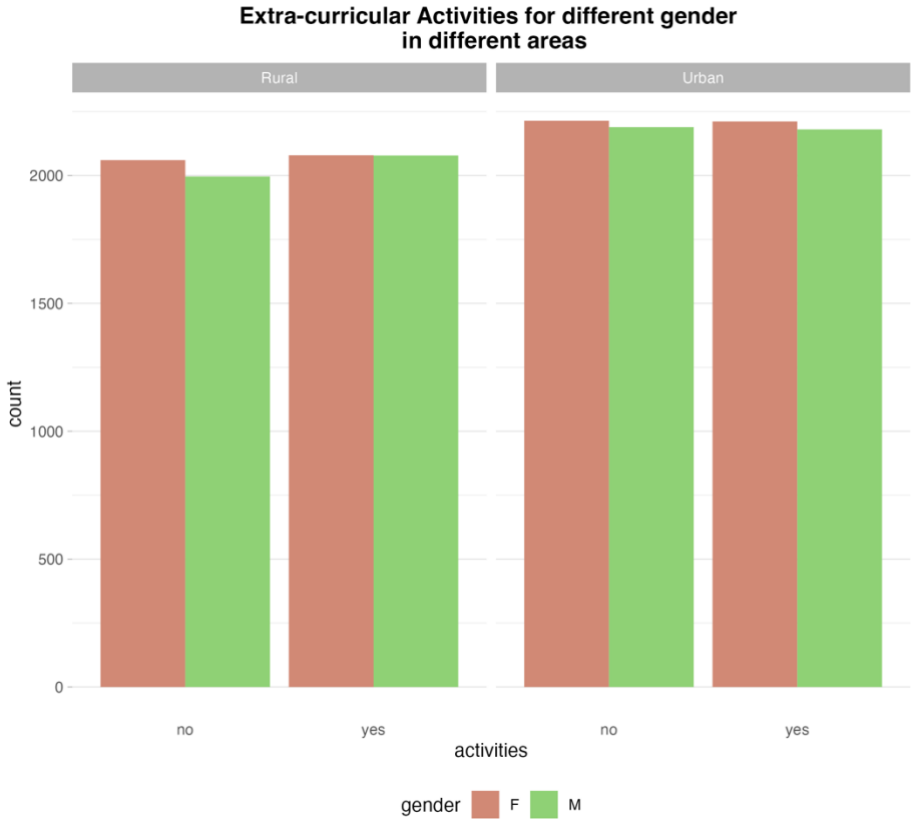


Figure 3.8.2 Relationship between address and extra-curricular activities graph

As we can see from the chart, the proportion of males and females who participate in extra-curricular activities is similar across both urban and rural areas. Whether in rural or urban areas, male and female has joined the extra-curricular activities and quit equally. This suggests that there is no significant difference in the support or encouragement that males and females receive with respect to participating in such activities.

One possible reason for this trend could be that schools and communities have taken steps to promote equal opportunities for males and females to participate in extra-curricular activities. This could include initiatives such as ensuring that activities are open to all genders, actively encouraging females to participate, and providing support for females who may face additional barriers to participation, such as social or cultural expectations.

Overall, this dataset suggests that there is no gender-based discrimination or stereotyping when it comes to participation in extra-curricular activities in both urban and rural areas. However, further analysis should be conducted as the gender in activities is hard to be controlled and this data shows an abnormal distribution for the gender joining the extra-curricular activities in both areas.

Analysis 3.9: Are different gender students in different areas having discrimination or gap in secondary school tests?

```
#====Analysis 3.9:  Are different gender students in different areas having discrimination or gap in secondary school tests?====
df_gender <- df %>% select(address, gender, ssc_level) %>% count(address, gender, ssc_level)
df_gender
ggplot(df_gender, aes(ssc_level, y=n)) +
  geom_segment( aes(x=ssc_level, xend=ssc_level, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address)+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Secondary school result for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p9.png")
```

Figure 3.9.1 Relationship between gender and secondary school test code

We first select the columns 'address', 'gender', and 'ssc_level' from the dataset and count the number of occurrences for each combination of these three variables using the count() function. We store this in a new dataframe called df_gender.

Next, we create a plot using ggplot2. We map the ssc_level column to the x-axis and the count of each combination to the y-axis. We use geom_segment() to create a bar chart with each bar colored in sky blue. We then add points on top of the bars to represent the count for each combination. The points are colored blue with an alpha value of 0.6. We use coord_flip() to flip the plot horizontally so that the bars are displayed vertically. We add facet_grid to plot separately for different gender students in different areas.

Finally, we add some theming to the plot using theme() to remove some of the grid lines and axis ticks. We use ggtitle() to add a title to the plot and center it with element_text(). We also set the legend position to the bottom and set the plot background to white.

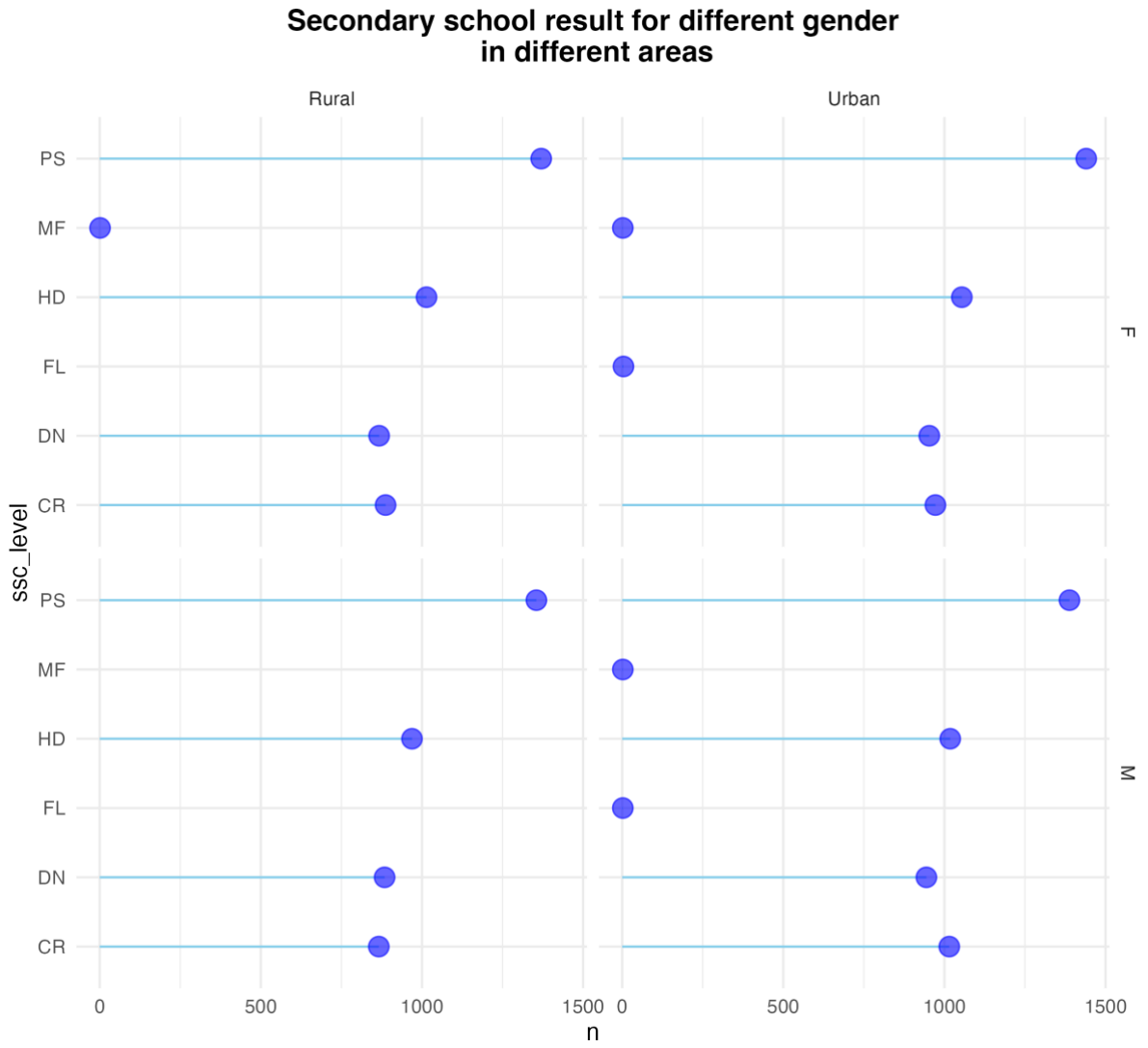


Figure 3.9.2 Relationship between gender and secondary school test graph

Based on the data, it appears that there is no significant difference in the distribution of secondary school test results between genders in different areas. The bar chart would show the number of students in each category of secondary school test result level (CR, DN, HD, MF, PS) for each gender in the urban and rural areas. One possible reason for this even distribution could be that the grading system is standardized and objective and is not influenced by gender or location. This suggests that the education system is fair and does not discriminate based on gender or area of residence. Overall, the data suggest that students, regardless of gender, have equal opportunities to achieve success in their secondary school tests.

Analysis 3.10: Are different gender students in different areas having a stereotype in joining the secondary school board?

```
#====Analysis 3.10: Are different gender students in different areas having a stereotype in joining the secondary school board?====
df_gender <- df %>% select(address, gender, ssc_b) %>% count(address, gender, ssc_b)
df_gender
ggplot(df_gender, aes(ssc_b, y=n)) +
  geom_segment( aes(x=ssc_b, xend=ssc_b, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Secondary school board for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p10.png")
```

Figure 3.10.1 Relationship between gender, address, joining secondary school boards code

In this analysis, we are exploring whether different gender students in different areas have stereotypes in joining the secondary school board. We first select the required columns (address, gender, and ssc_b) from the data frame using the select() function. We then group the data by address, gender, and ssc_b columns and count the number of occurrences of each combination using the count() function.

Next, we create a horizontal bar chart using ggplot(), with ssc_b on the x-axis and the count of each combination on the y-axis. We use geom_segment() to create a segment for each count, and geom_point() to place a point at the end of each segment. We flip the chart using coord_flip() to make the x-axis horizontal. We use facet_grid() to create a separate chart for each gender and address combination, and we set the y-axis label using ylab(). Finally, we set the chart's title, background, and legend position using ggtitle(), theme(), and scale_color_ipsum(), respectively.

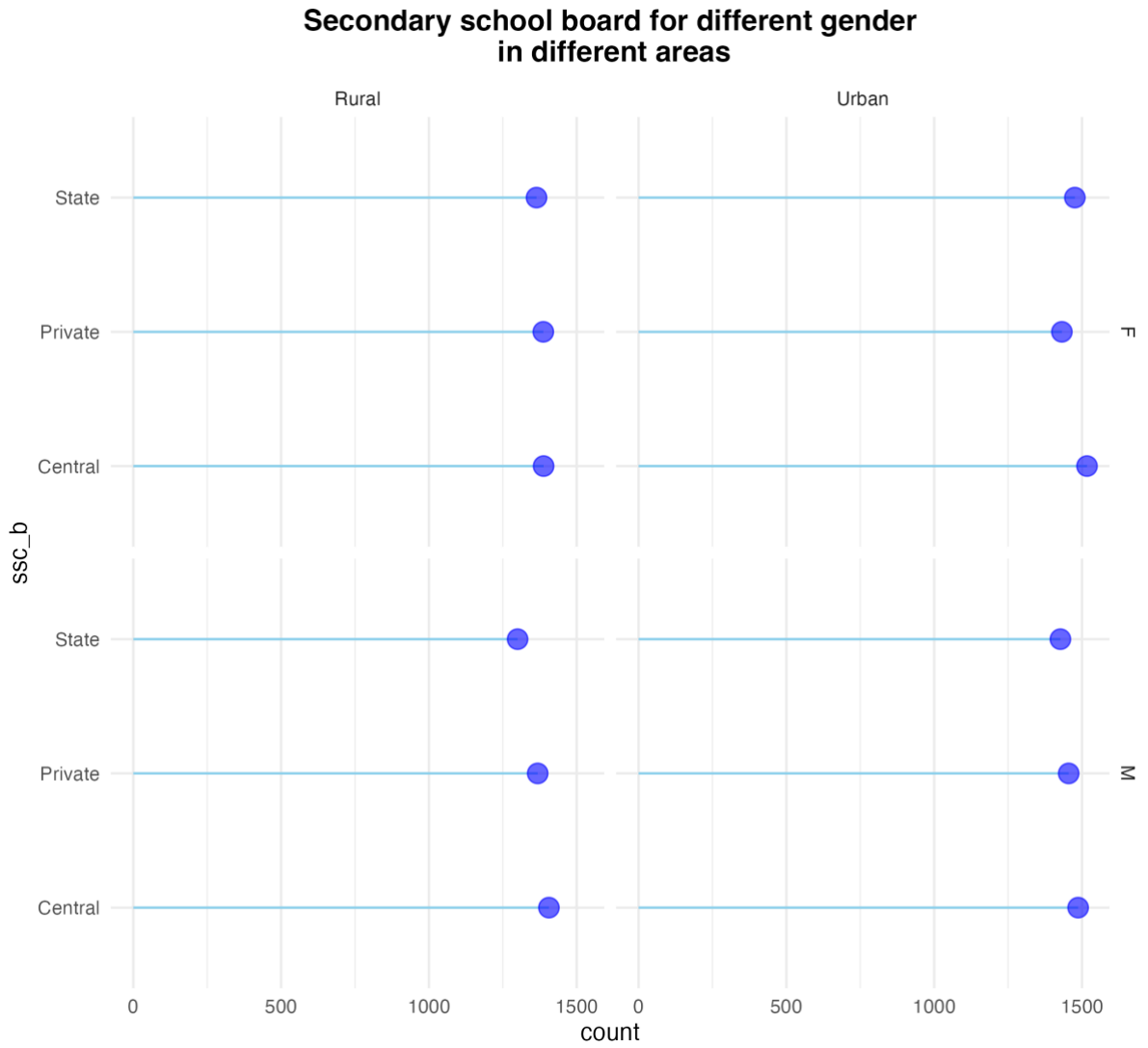


Figure 3.10.2 Relationship between gender, address, joining secondary school boards graph

The data provided shows the distribution of students based on their gender, location, and the board of their secondary school education. The data is divided into two categories, rural and urban, with each category having subcategories based on gender and board of education. The count column represents the number of students falling into each category.

The data indicate that the distribution of students across different school boards is generally uniform, with no significant variation observed in the number of students across different school boards. This suggests that there is no discrimination based on the board of the school in the admission process of students into secondary schools.

In summary, the data shows that the admission process of secondary schools in different school boards is not biased towards any particular gender.

Analysis 3.11: Are different gender students in different areas having discrimination or gap in high school tests?

```
#====Analysis 3.11: Are different gender students in different areas having discrimination or gap in high school tests?====
df_gender <- df %>% select(address, gender, hsc_level) %>% count(address, gender, hsc_level)
df_gender
ggplot(df_gender, aes(hsc_level, y=n)) +
  geom_segment( aes(x=hsc_level, xend=hsc_level, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("High school result for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p11.png")
```

Figure 3.11.1 Relationship between gender, address and high school test code

In this analysis, we are exploring whether there is any gender-based discrimination or gap in high school tests among students from different areas. We start by selecting the columns 'address', 'gender', and 'hsc_level' from the original data frame 'df' using the select() function. Then we count the number of observations for each combination of 'address', 'gender', and 'hsc_level' using the count() function and store the results in a new data frame called 'df_gender'.

Next, we create a plot using the ggplot2 library. We use the ggplot() function and specify the data to be used as 'df_gender' and aesthetics mapping for 'hsc_level' on x-axis and 'n' on the y-axis. We then add a segment using geom_segment() function to show the count of each category of 'hsc_level' for each 'gender' in a specific 'address'. We add a point using geom_point() function to show the total count of each category of 'hsc_level' for each 'gender' in a specific 'address'. We flip the coordinates using coord_flip() function to make it more readable.

Lastly, we add some customization to our plot using the theme() function. We remove the vertical lines using the panel.grid.major.y argument, remove the ticks on the y-axis using the axis. ticks.y argument, and remove the border of the plot using the panel. border argument. We add facet_grid() function to split the plot by 'gender' and 'address'. We add ylab() function to label the y-axis with 'count'. We add a title to our plot using ggtitle() function and specify its position and style using the theme() function. Finally, we add color schemes using scale_color_ipsum() and scale_fill_ipsum() functions and apply a minimal theme using theme_minimal() function.

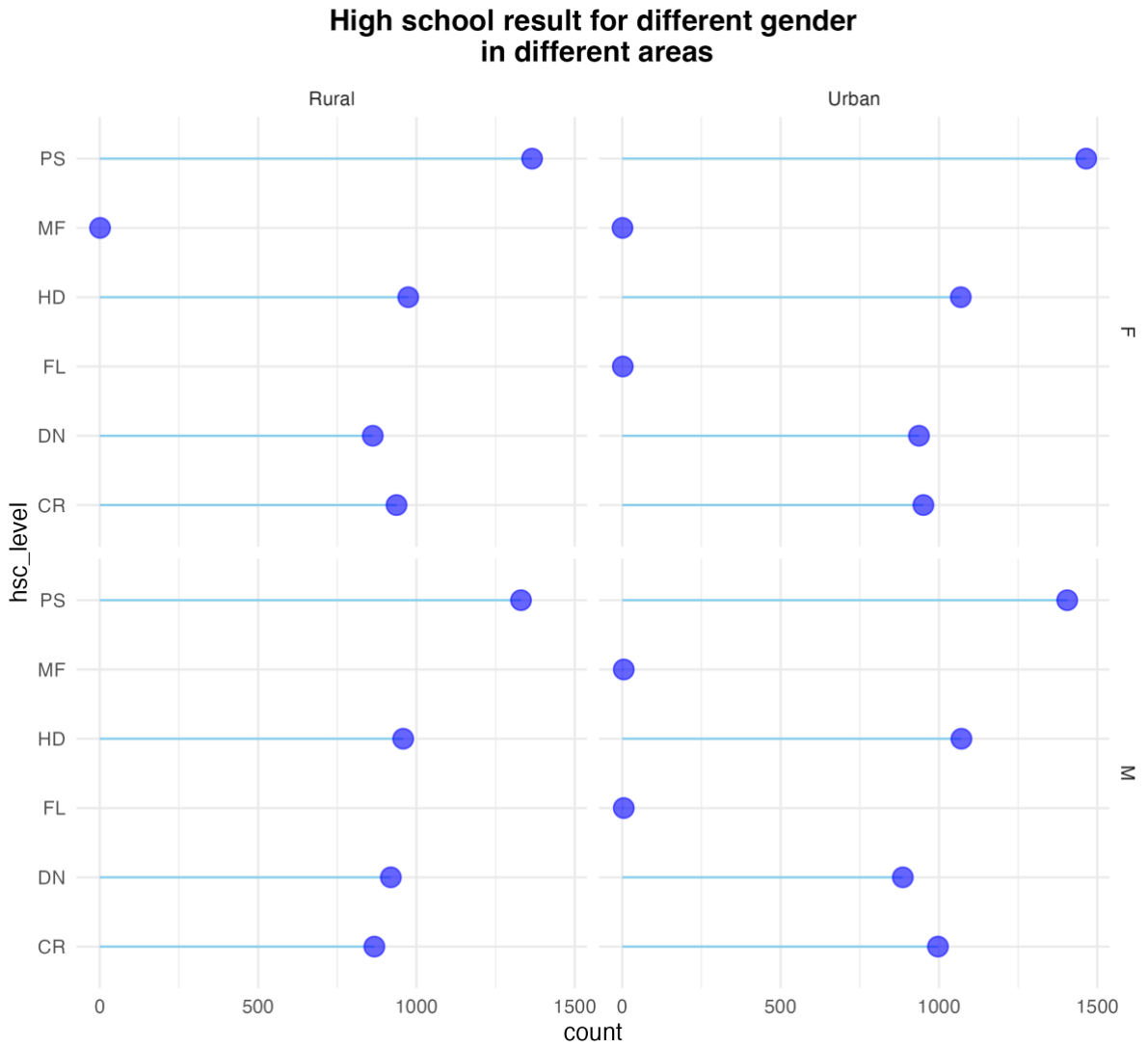


Figure 3.11.2 Relationship between gender, address and high school test graph

The lollipop chart represents the distribution of the HSC level results for students of different gender and areas. It shows that students from both areas have obtained similar high school results, with no significant difference in the distribution, whether they are boys or girls.

In summary, the data suggest that there is no significant discrimination in the distribution of HSC results based on gender from different locations. However, further analysis is required to determine the underlying factors that contribute to the observed differences in the proportion of students achieving different levels of results.

Analysis 3.12: Are different gender students in different areas having a stereotype in joining the high school board?

```
#====Analysis 3.12: Are different gender students in different areas having a stereotype in joining the high school board?====
df_gender <- df %>% select(address, gender, hsc_b) %>% count(address, gender, hsc_b)
df_gender
ggplot(df_gender, aes(hsc_b, y=n)) +
  geom_segment( aes(x=hsc_b, xend=hsc_b, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("High school board for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p12.png")
```

Figure 3.12.1 Relationship between gender, address and high school board code

In this analysis, we are examining if there is any gender-based stereotype among students from different areas when it comes to joining the high school board. We start by selecting the necessary columns from our data frame using the select() function. We then count the number of occurrences of each combination of address, gender, and high school board using the count() function.

Next, we create a bar chart using ggplot() and pass the df_gender data frame as an argument. We specify the hsc_b column as the x-axis and n as the y-axis. We add a vertical bar segment for each combination of hsc_b and n using geom_segment(). We also add a blue point for each combination using geom_point(). We flip the coordinates using coord_flip() so that the bars are horizontal.

We modify the theme of the plot using theme() to remove the vertical grid lines and the y-axis ticks. We also set the facet for the plot using facet_grid() to display separate graphs for each gender and address combination. Finally, we add a title to the plot using ggtitle(), set the theme to theme_minimal(), and adjust the position of the legend and plot title.

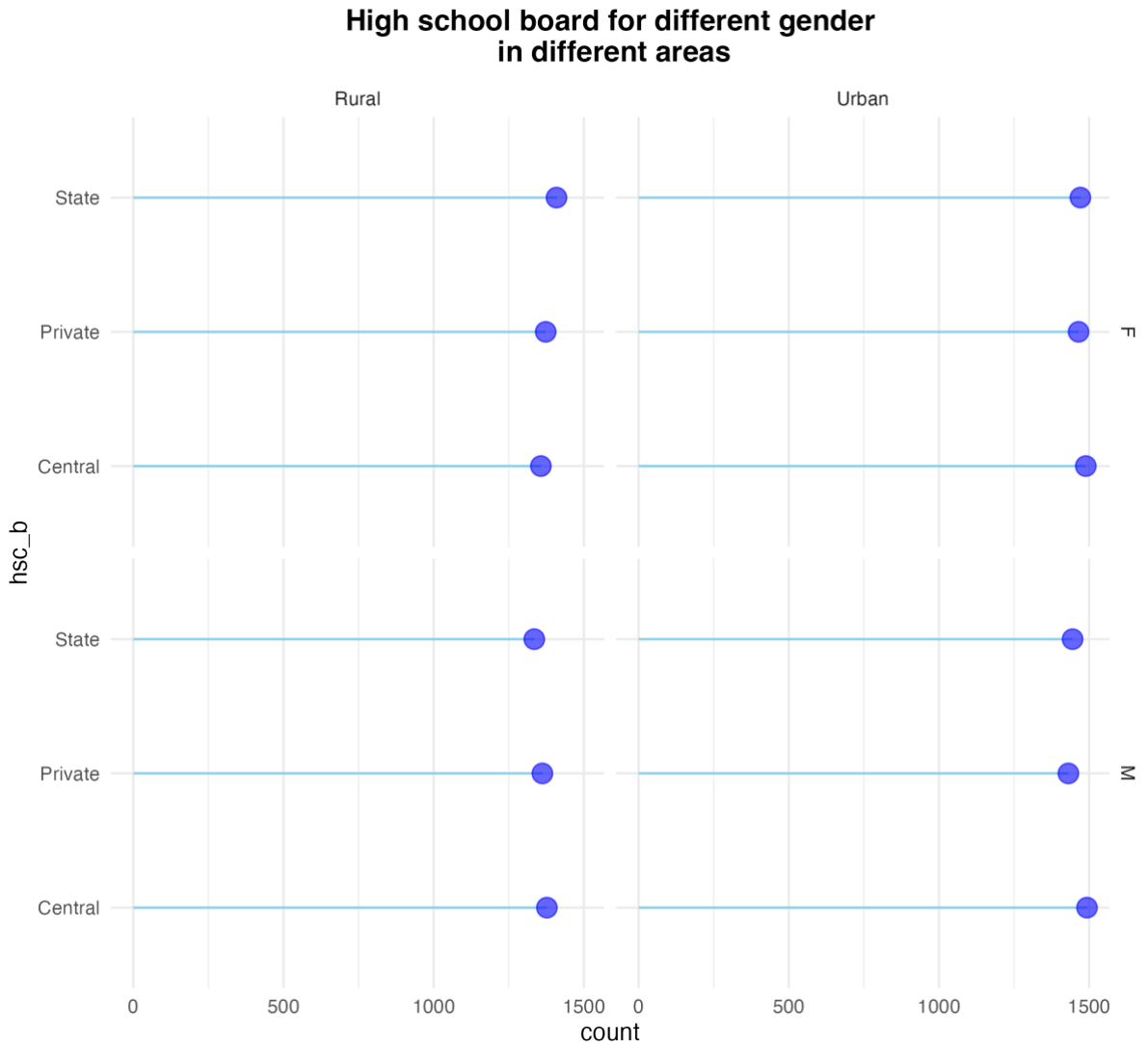


Figure 3.12.2 Relationship between gender, address and high school board graph

The chart represents the distribution of the high school board choice for students of different gender and areas. It shows that students from all areas and genders have obtained a similar preference in choosing high school boards, with no significant difference in the distribution.

One possible reason for this could be the similar quality of schools or educational facilities in different areas. It is also possible that students from both rural and urban areas have had the same access to resources or received the same support from their families or previous schools.

In summary, the data suggest that there is no significant discrimination in the distribution of high school choices based on gender or location. However, further analysis is required to determine the underlying factors that contribute to the observed differences in the proportion of students choosing different levels of schools.

Analysis 3.13: Are different gender students in different areas having a stereotype in choosing high school specialization?

```
#====Analysis 3.13: Are different gender students in different areas having a stereotype in choosing high school specialization?====
df_gender <- df %>% select(address, gender, hsc_s) %>% count(address, gender, hsc_s)
df_gender
ggplot(df_gender, aes(hsc_s, y=n)) +
  geom_segment( aes(x=hsc_s, xend=hsc_s, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("High school specialization for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p13.png")
```

Figure 3.13.1 Relationship between gender, address and high school specialization code

In this analysis, we are examining whether different gender students in different areas have a stereotype in choosing their high school specialization. We first create a new data frame, df_gender, by selecting the address, gender, and high school specialization columns from our original data frame, and then counting the number of occurrences of each combination of these variables.

Next, we create a bar plot using ggplot2 to visualize the counts of each high school specialization for each gender in each address category. We use the hsc_s variable for the x-axis and the count for the y-axis. We use geom_segment to create a horizontal line for each count and geom_point to add a blue point at the end of each line. We use coord_flip to flip the x- and y-axes to make the plot more readable, and we use facet_grid to create separate panels for each gender and address category.

Finally, we add axis labels, a title, and some formatting options to improve the appearance of the plot.

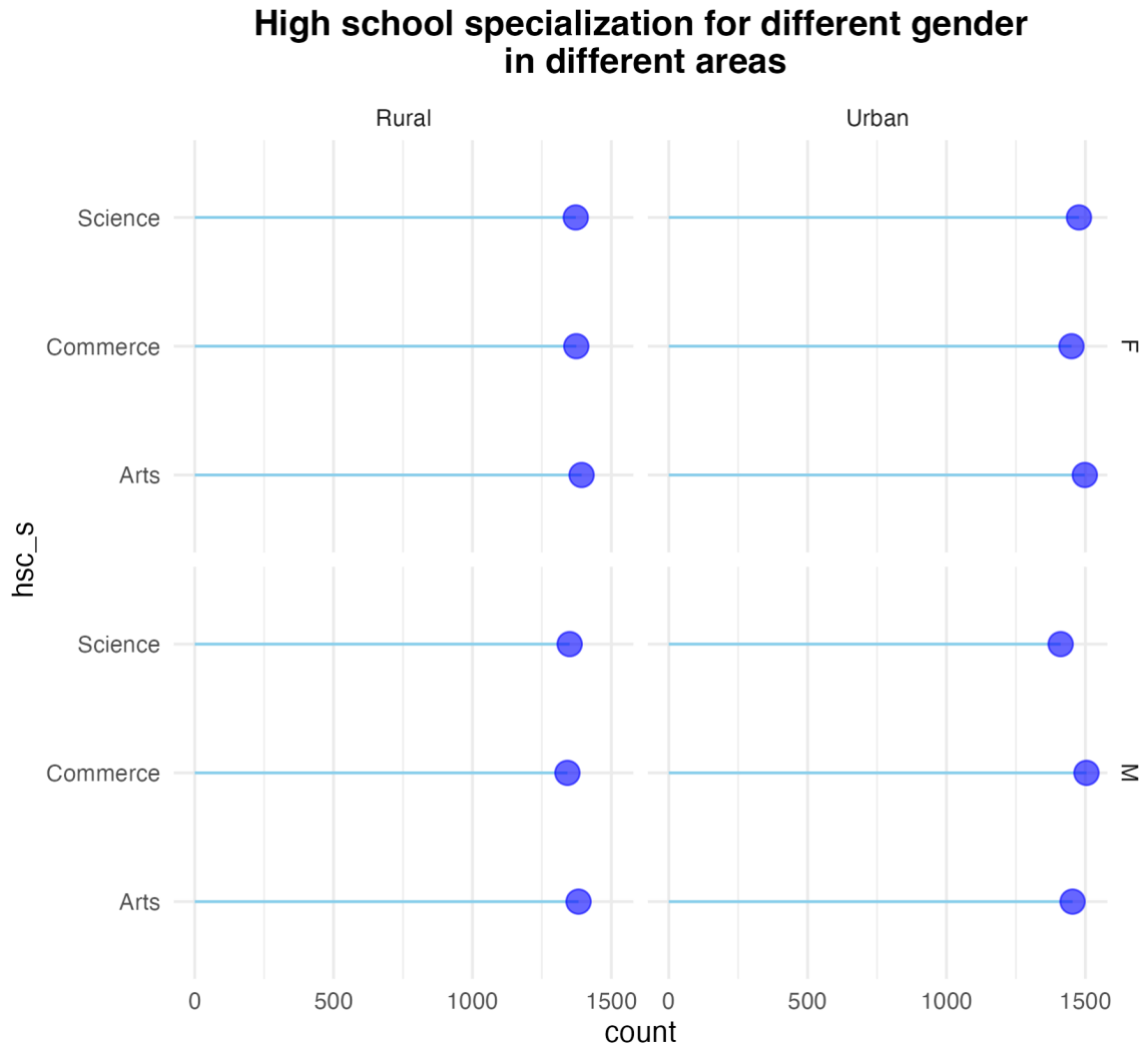


Figure 3.13.2 Relationship between gender, address and high school specialization graph

The data provided shows the distribution of students based on their gender, location, and the specialization of their high school education. The data is divided into two categories, rural and urban, with each category having subcategories based on gender and specialization. The count column represents the number of students falling into each specialization.

The data indicate that the distribution of students choosing different specializations is generally uniform, with no significant variation observed in the number of students selecting different specializations. This suggests that there is no discrimination based on the specialization in the admission process of students into high specializations. The data also shows that the admission process of high school specializations in different specializations is not biased towards any particular gender. In summary, the data shows that the admission process of high school specializations in different specializations is not biased towards any particular gender from any area.

Analysis 3.14: Are different gender students in different areas having discrimination or gap in high school tests?

```
#====Analysis 3.14: Are different gender students in different areas having discrimination or gap in high school tests?====
df_gender <- df %>% select(address, gender, degree_level) %>% count(address, gender, degree_level)
df_gender
ggplot(df_gender, aes(degree_level, y=n)) +
  geom_segment( aes(x=degree_level, xend=degree_level, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Degree result for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p14.png")
```

Figure 3.14.1 Relationship between gender, address and high school test code

In this analysis, we are investigating whether there is any gender-based discrimination or gap in the degree results of students in different areas. To do so, we first select the relevant columns from the data frame using the select() function, which includes the address, gender, and degree_level columns. Then, we count the number of students for each combination of address, gender, and degree_level using the count() function.

Next, we create a visualization of the data using the ggplot() function. We map the degree_level to the x-axis and the count of students to the y-axis. We use facet_grid() to create separate panels for each combination of gender and address. Finally, we customize the appearance of the plot using various theme functions to make it more readable and visually appealing.

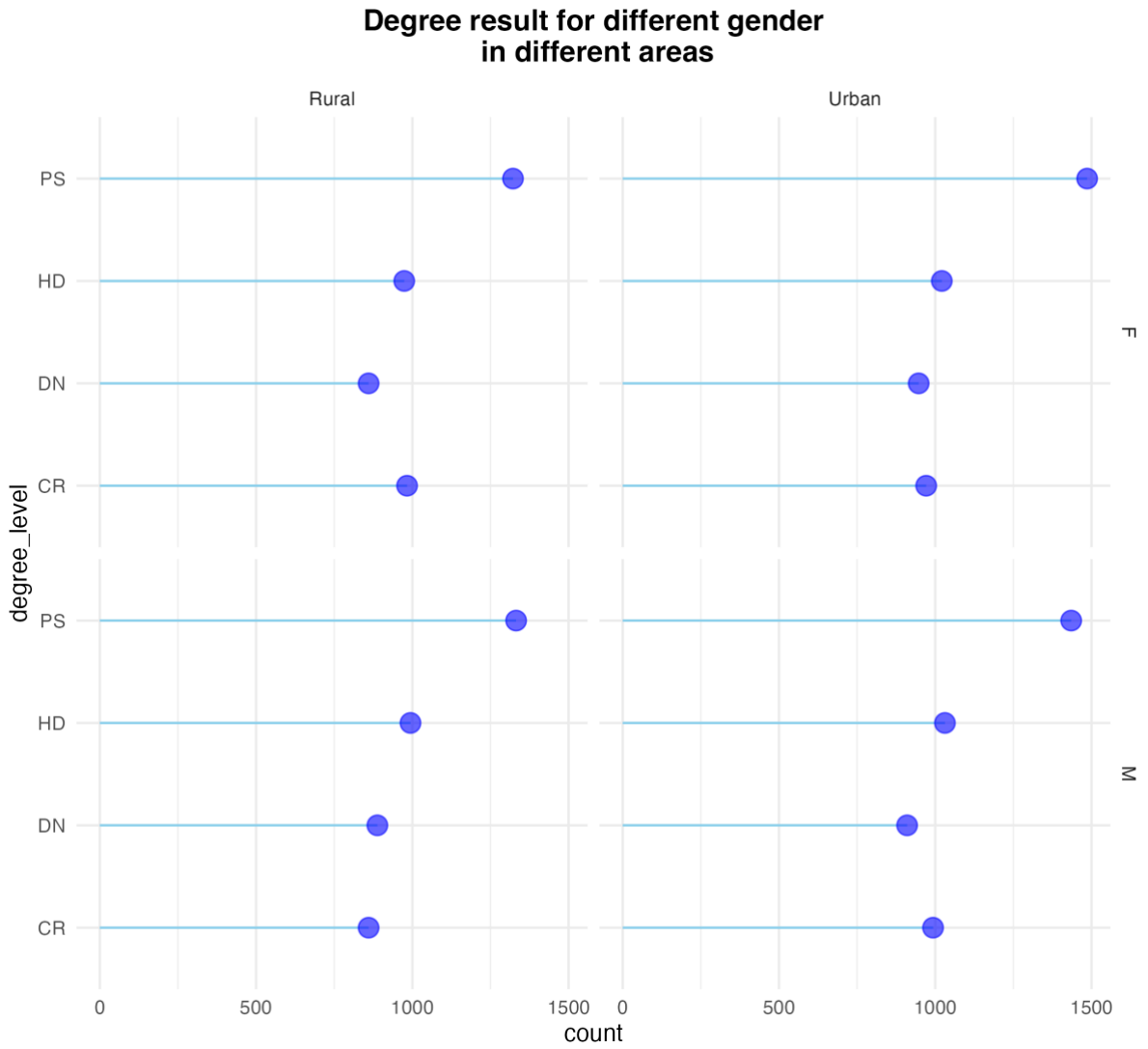


Figure 3.14.2 Relationship between gender, address and high school test graph

The data provided shows the distribution of students' high school test results based on their gender and location.

The data suggests that there is no significant variation in the number of students across different levels of high school tests based on gender or location. This indicates that the marking scheme of high schools in tests is not biased toward any particular gender or location. In summary, the data shows that there is no discrimination or gap in the marking of high school students' tests even some markers, in fact, will tend to give higher marks for the female that are writing in a more tangible way or have stereotypes to different gender at the moment they mark the papers.

Analysis 3.15: Are different gender students in different areas having a stereotype in choosing degree types?

```
#====Analysis 3.15: Are different gender students in different areas having a stereotype in choosing degree types?====
df_gender <- df %>% select(address, gender, degree_t) %>% count(address, gender, degree_t)
df_gender
ggplot(df_gender, aes(degree_t, y=n)) +
  geom_segment( aes(x=degree_t, xend=degree_t, y=0, yend=n), color="skyblue") +
  geom_point( size=5, color="red", fill=alpha("orange", 0.3), alpha=0.7, shape=21, stroke=2)+
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Degree types for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p15.png")
```

Figure 3.15.1 Relationship between gender, address and degree types code

In this analysis, we are investigating whether different gender of students in different areas have a stereotype in choosing degree types. To do so, we first select the necessary columns (address, gender, and degree_t) from the dataset using the select() function in the dplyr package. We then count the occurrences of each combination of address, gender, and degree_t using the count() function.

We then use the ggplot() function to create a bar plot of the counts for each combination of degree type and gender, grouped by address. We use the geom_segment() function to create vertical lines for each count, and the geom_point() function to plot the counts as points. We also use the facet_grid() function to create separate plots for each combination of gender and address.

In this plot, the x-axis represents the degree types, and the y-axis represents the count. We use different colors and shapes to distinguish the points for each combination of gender and address. Finally, we add appropriate labels and titles using the theme() function.

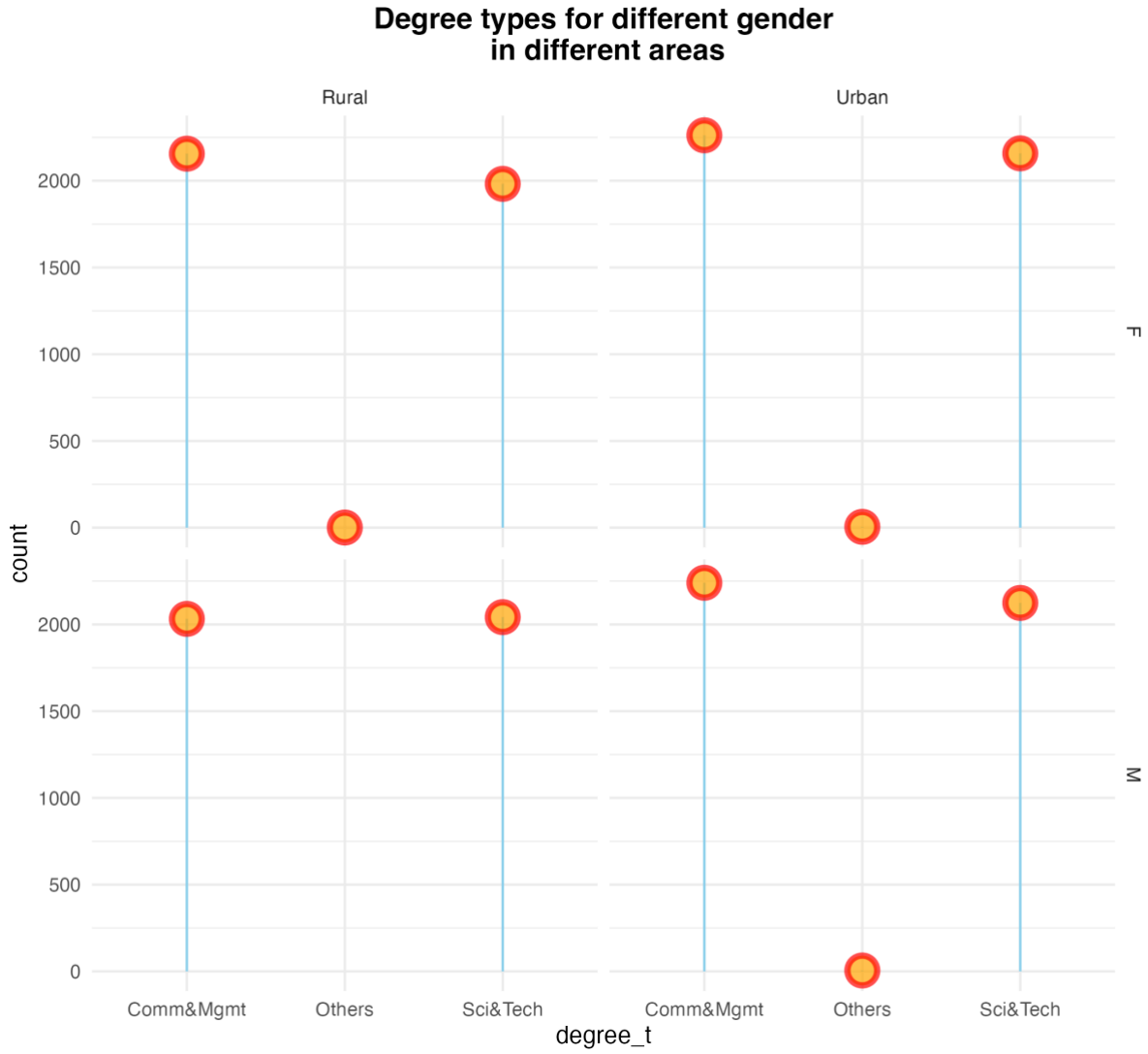


Figure 3.15.2 Relationship between gender, address and degree types graph

The data provided shows the distribution of students based on their gender, location, and the degree type they have chosen. The count column represents the number of students falling into each category. The data is divided into two categories, rural and urban, with each category having subcategories based on gender and degree type.

The data indicate that there is no tendency for students from different areas and genders to choose a particular degree type. Contrary to popular belief, the data suggests that there is no gender-based or location-based bias in the choice of degree type among the students.

In summary, the data shows that students from different areas and genders have an equal opportunity to choose the degree type they want to pursue, and there is no location-based or gender-based discrimination in the admission process of students into various degree programs.

Analysis 3.16: Are different gender students in different areas having discrimination or gap in employability tests?

```
#====Analysis 3.16: Are different gender students in different areas having discrimination or gap in employability tests?====
df_gender <- df %>% select(address, gender, etest_level) %>% count(address, gender, etest_level)
df_gender
ggplot(df_gender, aes(etest_level, y=n)) +
  geom_segment( aes(x=etest_level, xend=etest_level, y=0, yend=n), color="skyblue") +
  geom_point( size=5, color="red", fill=alpha("orange", 0.3), alpha=0.7, shape=21, stroke=2)+
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Employability test result for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p16.png")
```

Figure 3.16.1 Relationship between gender, address and employability test result code

In this analysis, we are examining if there is any discrimination or gap in employability tests among students of different genders in different areas.

The code first selects the columns related to the address, gender, and employability test result using the select() function from the dplyr package. It then counts the number of observations for each combination of address, gender, and employability test result using the count() function.

The resulting data frame df_gender is then used to create a ggplot visualization using the ggplot() function. The x-axis represents the employability test result, and the y-axis represents the count of observations for each category. The geom_segment() function is used to create vertical lines for each category, and the geom_point() function is used to create colored points for each observation. The facet_grid() function is used to create separate plots for each combination of gender and address. Finally, various theme() and ggtitle() functions are used to format and label the plot.

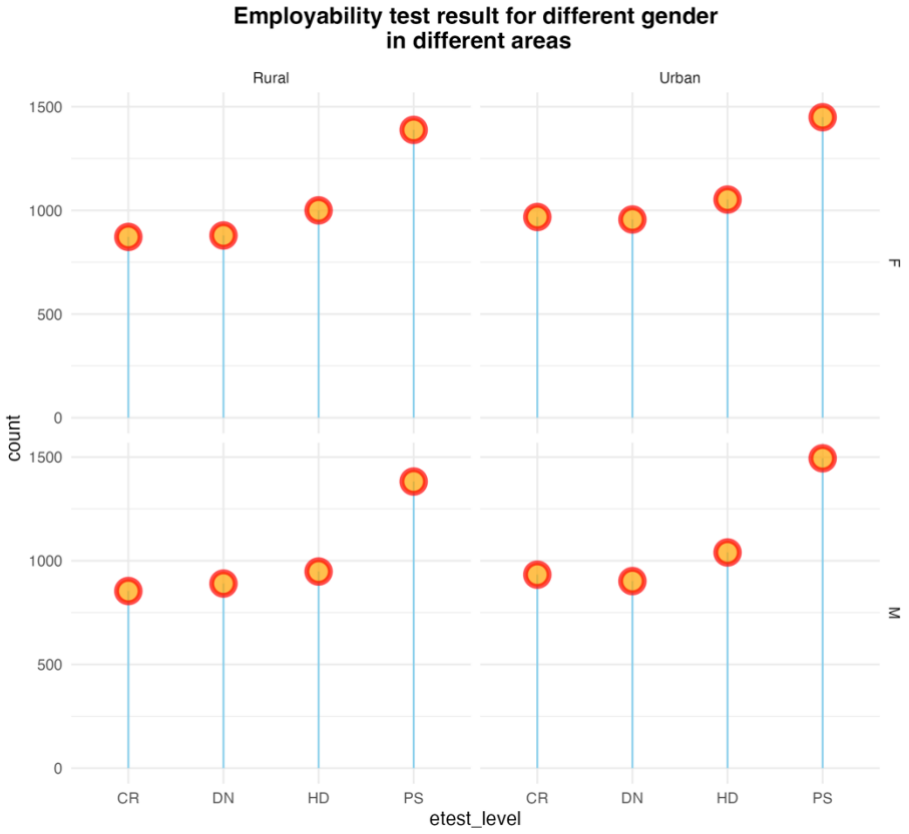


Figure 3.16.2 Relationship between gender, address and employability test result graph

The lollipop chart provided shows the results of employability tests for students from different areas and genders. There is no significant gap in employability test results between different genders of students from different areas. The data is evenly distributed across all subcategories, indicating that gender and area are not significant factors in employability.

It is important to note that employability tests include some uncontrollable elements that may affect the results, such as employer preferences and job requirements. For example, some employers may prefer to hire boys for physically demanding work, while some positions may require girls who are more detail-oriented, such as secretaries or administrative assistants. However, these uncontrollable elements do not seem to have a significant impact on the overall employability test results for students from different areas and genders, based on the dataset provided.

In summary, the dataset does not show any discrimination in employability test results between students from different areas and genders. While there may be some uncontrollable factors that affect employability, the dataset suggests that they do not have a significant impact on the results of the employability test.

Analysis 3.17: Are different gender students in different areas having discrimination or gap in MBA tests?

```
#====Analysis 3.17: Are different gender students in different areas having discrimination or gap in MBA tests?====
df_gender <- df %>% select(address, gender, mba_level) %>% count(address, gender, mba_level)
df_gender
ggplot(df_gender, aes(mba_level, y=n)) +
  geom_segment( aes(x=mba_level, xend=mba_level, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  theme_light() +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("MBA result for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p17.png")
```

Figure 3.17.1 Relationship between gender, address and MBA test result code

In this analysis, we are investigating whether there is a gender-based gap or discrimination in MBA test results among students from different areas.

First, we use the select function from the dplyr package to select the columns containing information about the students' addresses, gender, and MBA test results. Then we use the count function to count the number of students with each combination of address, gender, and MBA test result.

Next, we use the ggplot function to create a bar chart with MBA test results on the x-axis and the count of students on the y-axis. We use geom_segment and geom_point to create bars and data points for each category, respectively. We also use facet_grid to separate the chart into panels based on gender and address.

Finally, we add labels and formatting to the chart using various theme and scale functions. The title of this analysis is "MBA result for different gender in different areas".

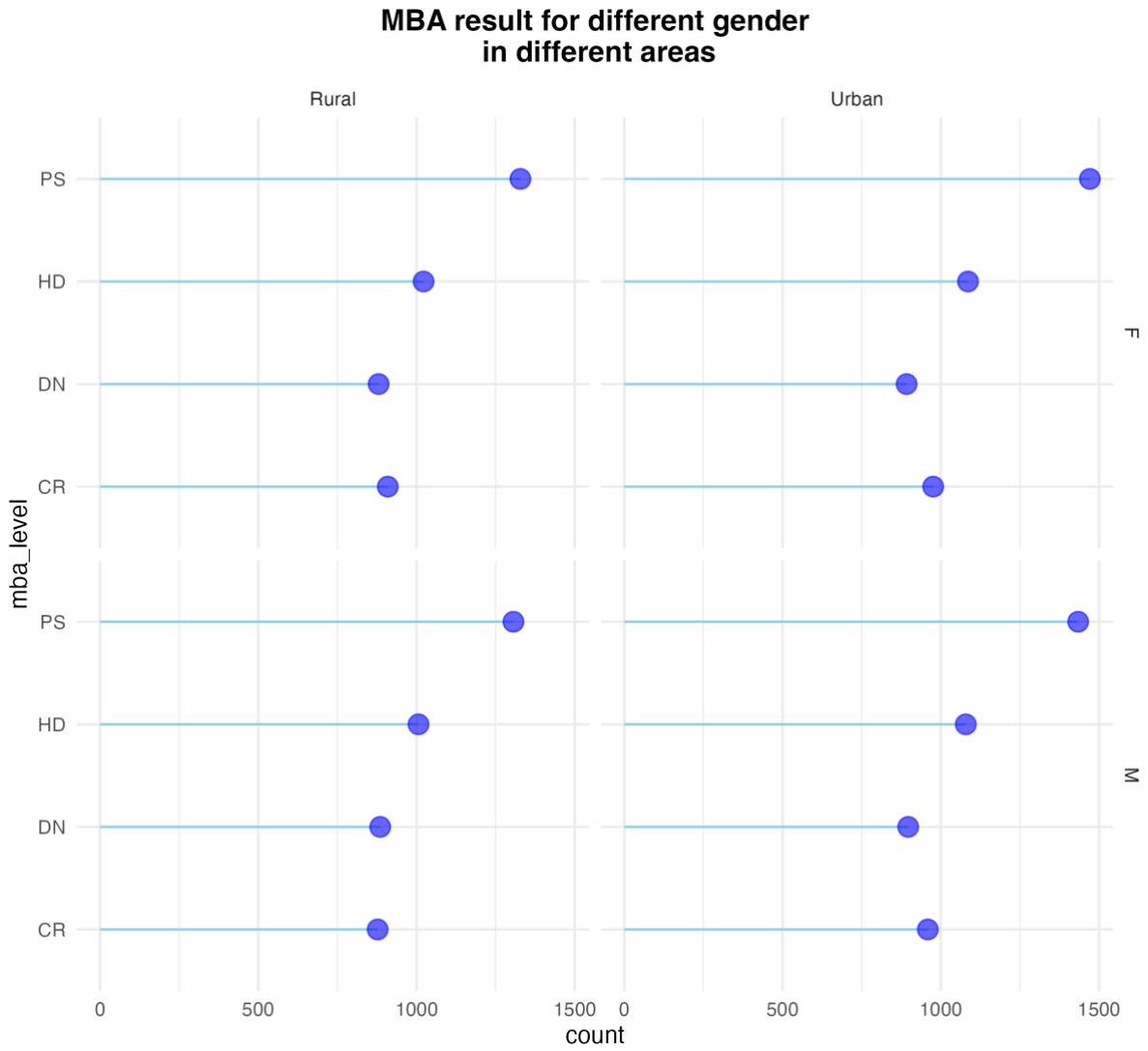


Figure 3.17.2 Relationship between gender, address and MBA test result graph

Based on the given diagram, there is no significant difference in the MBA test results of students from different genders and areas. The distribution and proportion of test results are almost similar for every subcategory. This indicates that the ability of students in MBA subjects is independent of their gender and the area they belong to.

In conclusion, the diagram does not provide any evidence of discrimination or gap in MBA test results among students from different areas and genders.

Analysis 3.18: Is there any discrimination that will affect the status of the student's placement in the dataset?

```
#====Analysis 3.18: Is there any discrimination that will affect the status of the student's placement in the dataset?====
df_gender <- df %>% select(address, gender, status) %>% count(address, gender, status)
df_gender
ggplot(df_gender, aes(status, y=n)) +
  geom_segment( aes(x=status, xend=status, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  theme_light() +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(gender~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Status for different gender
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p18.png")
```

Figure 3.18.1 Relationship between gender, address and status code

In this analysis, we are examining whether there is any discrimination affecting the status of students' placements in the dataset. We first select the columns of interest (address, gender, and status) and count the number of occurrences of each combination using the count() function. Then, we create a grouped bar chart using ggplot() to visualize the distribution of status for each combination of gender and address. The geom_segment() function creates a bar for each category and the geom_point() function adds a point to represent the count of each category. The facet_grid() function splits the chart into panels for each combination of gender and address, and coord_flip() rotates the chart to display the categories horizontally. Finally, we add labels, titles, and themes to improve the readability and aesthetics of the chart.

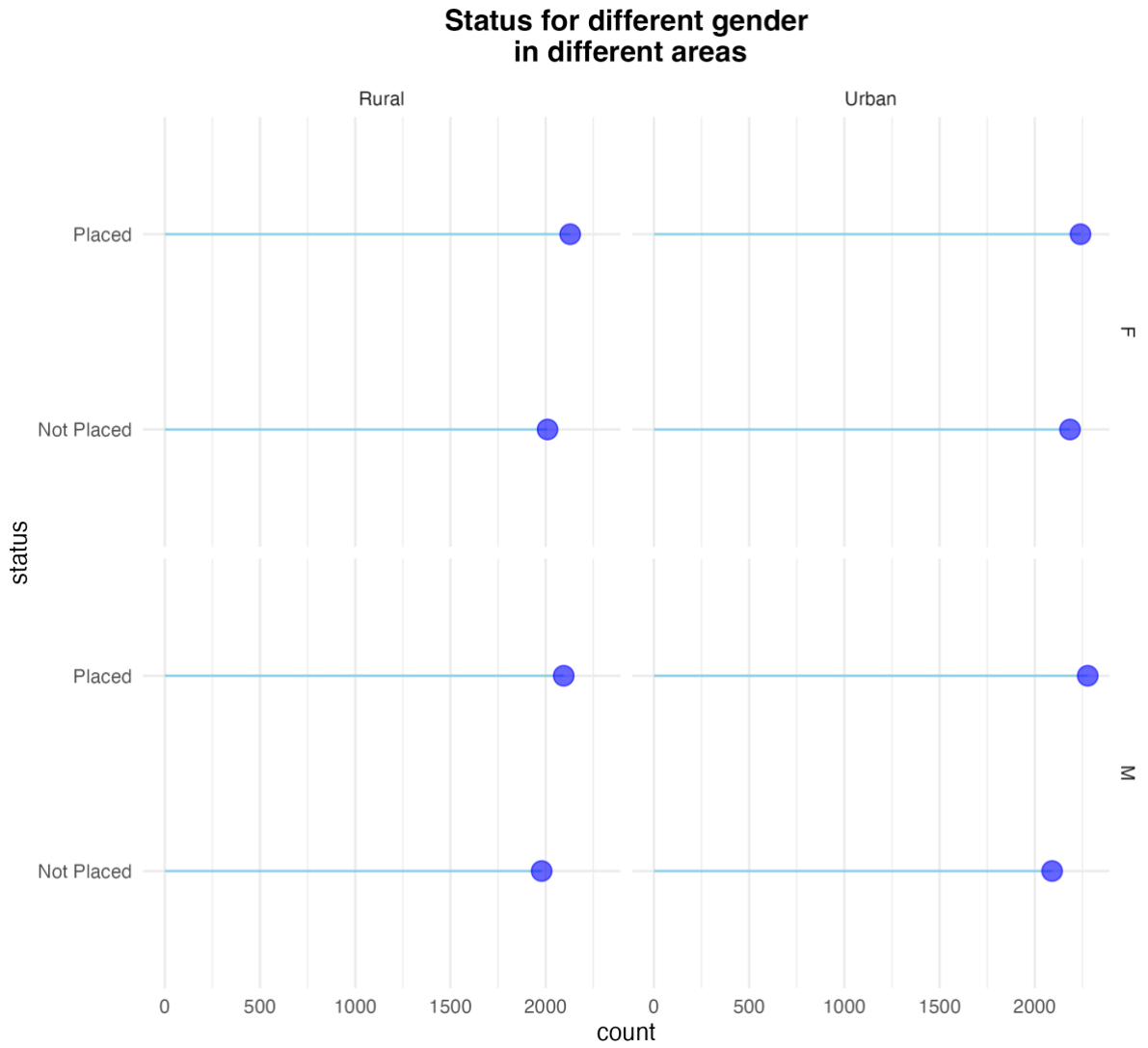


Figure 3.18.2 Relationship between gender, address and status graph

The data indicates that there is no significant difference in the employment status between males and females from rural and urban areas. The number of students who were placed and not placed in jobs is relatively similar across all groups. This suggests that gender and location do not play a significant role in determining employability in this particular dataset. Overall, this data shows that there is no discrimination in employment for students from different areas and genders.

Analysis 3.19: Is there any difference between males and females from different areas in choosing their specialization?

```
#====Analysis 3.19: Is there any difference between males and females from different areas in choosing their specialization?====
df_gender <- df %>% select(address, gender, specialisation) %>%
  count(address, gender, specialisation)
df_gender
to_add <- matrix(NA, empty_bar, ncol(df_gender))
colnames(to_add) <- colnames(df_gender)
df_gender <- rbind(df_gender, to_add)
df_gender$id <- seq(1, nrow(df_gender))
df_gender
df_gender <- df_gender %>% mutate(angle = 90 - 360 * (id-0.5) /nrow(df_gender)) %>%
  mutate(hjust = ifelse(angle < -90, 1, 0)) %>% mutate(angle = ifelse(angle == 90, angle+180, angle))
df_gender

ggplot(df_gender, aes(x=as.factor(id), y=n, fill=address, colour=gender)) +
  geom_bar(stat="identity") +
  ylim(-1900,2500) +
  theme_minimal() +
  theme(
    axis.text = element_blank(),
    axis.title = element_blank(),
    panel.grid = element_blank(),
    plot.margin = unit(rep(-1,4), "cm")
  ) +
  coord_polar(start = 0) +  theme_void()+
  geom_text(data=df_gender, aes(x=id, y=n+10, label=paste(address, gender, specialisation, sep = " "),
  hjust=hjust), color="black", fontface="bold",alpha=0.6, size=2.5, angle= df_gender$angle, inherit.aes = FALSE )+
  theme(legend.position = "right", plot.title = element_text(hjust = 0.5, face = "bold"),
  plot.background = element_rect(fill = "white", color = "white")) + ggtitle("Specialisation of different gender and areas")
ggsave("~/Graph/3p19.png")
```

Figure 3.19.1 Relationship between gender, address and specialization code

In this analysis, we are examining whether there are any differences in the specialization choices of male and female students from different areas in the dataset. We first create a new data frame df_gender by selecting the address, gender, and specialization columns from the original data frame and then counting the number of occurrences of each combination of these variables. We then add an empty row to the df_gender data frame and assign an ID to each row.

Next, we create a new column called angle in the df_gender data frame which will be used to create the polar coordinate plot. We calculate the angle using the ID and the number of rows in the df_gender data frame. We also create a new column called hjust which will be used to align the text labels later on. We set hjust to 1 if the angle is less than -90, and 0 otherwise. We also adjust the angle to be between 0 and 360 degrees.

We then create a polar coordinate plot using the ggplot() function. We use as.factor(id) and n as the x and y variables, respectively. We fill the bars based on the address variable and color them based on the gender variable. We also set the y-axis limits to be between -1900 and 2500 and remove the axis labels, grid lines, and plot margins. We then use coord_polar() to create a polar coordinate system. Finally, we add text labels to the plot using geom_text(), with the labels consisting of a concatenation of the address, gender, and specialization

variables. We also adjust the hjust and angle parameters based on the values in the df_gender dataframe. We set the legend position to the right and add a title to the plot.

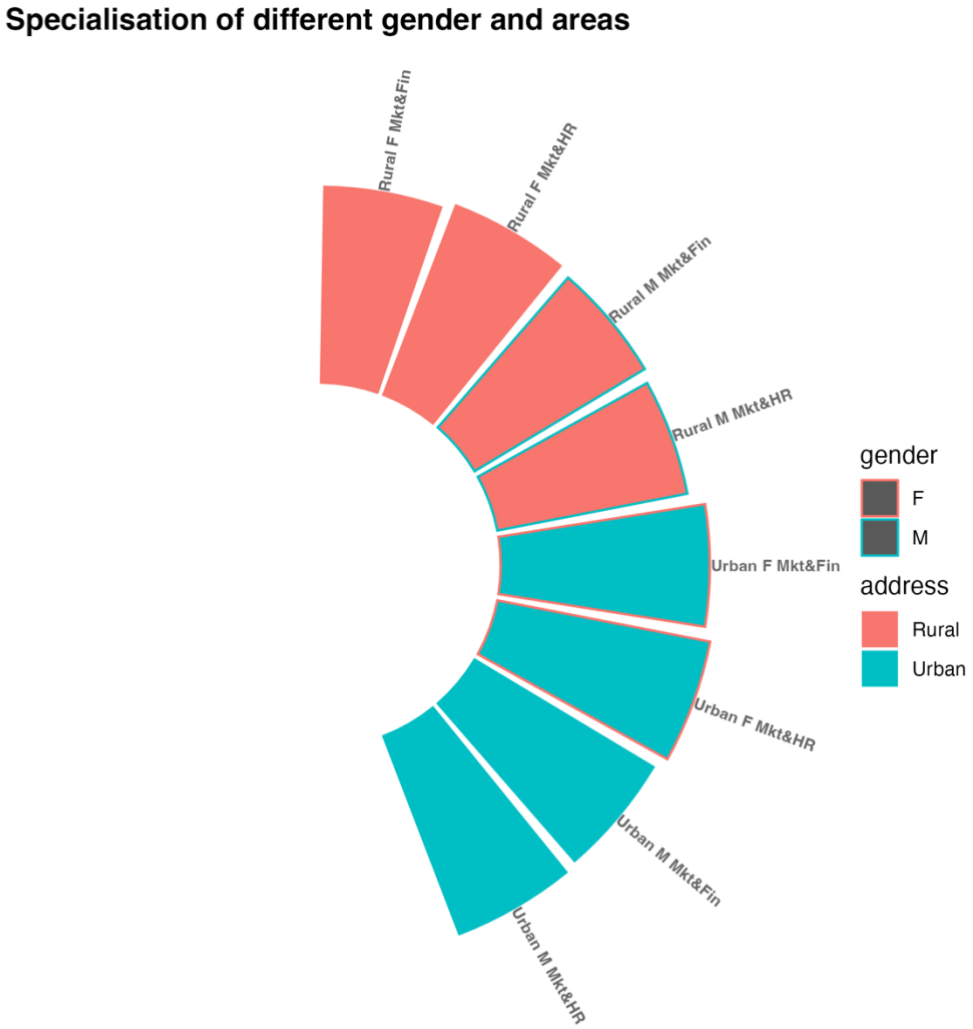


Figure 3.19.2 Relationship between gender, address and specialization graph

The circular bar plot suggests that there is no significant difference in the specialization choices of male and female students from rural and urban areas. This implies that there is no stereotype or societal pressure that forces males and females to pursue specific fields of study. The absence of such a bias in specialization choices indicates a level playing field for male and female students in terms of academic and career opportunities. Factors such as individual interests, personal aspirations, and future job prospects may have driven the specialization choices of the students in the dataset. Overall, the data presents a positive picture of gender equality in the field of education and employment.

Analysis 3.20: Do the males and females in different areas have equal chances to gain their work experience?

```
#====Analysis 3.20: Do the males and females in different areas have equal chances to gain their work experience? ====
df_exp <- df %>% select(workex, gender, address) %>%
  pivot_longer(cols = workex, names_to = "workex", values_to = "value") %>%
  group_by(gender, address) %>% count(value)
df_exp
ggplot(df_exp, aes(x = gender, y = n, fill = address)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~value, ncol = 1) + ylab("Count") +
  ggtitle("Gender work experience in different areas") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p20.png")
```

Figure 3.20.1 Relationship between gender, address and work experience code

In this analysis, we are investigating if males and females from different areas have equal chances of gaining work experience.

We first select the columns related to work experience, gender, and address from the dataset. Then, we pivot the data to make the work experience column more manageable. We group the data by gender, address, and work experience and count the number of occurrences for each combination.

Next, we plot a grouped bar chart with the gender on the x-axis, count on the y-axis, and address as the fill color. We use the facet_wrap function to split the plot by the different work experience values. We also add a title to the plot, adjust the theme, and set the legend position.

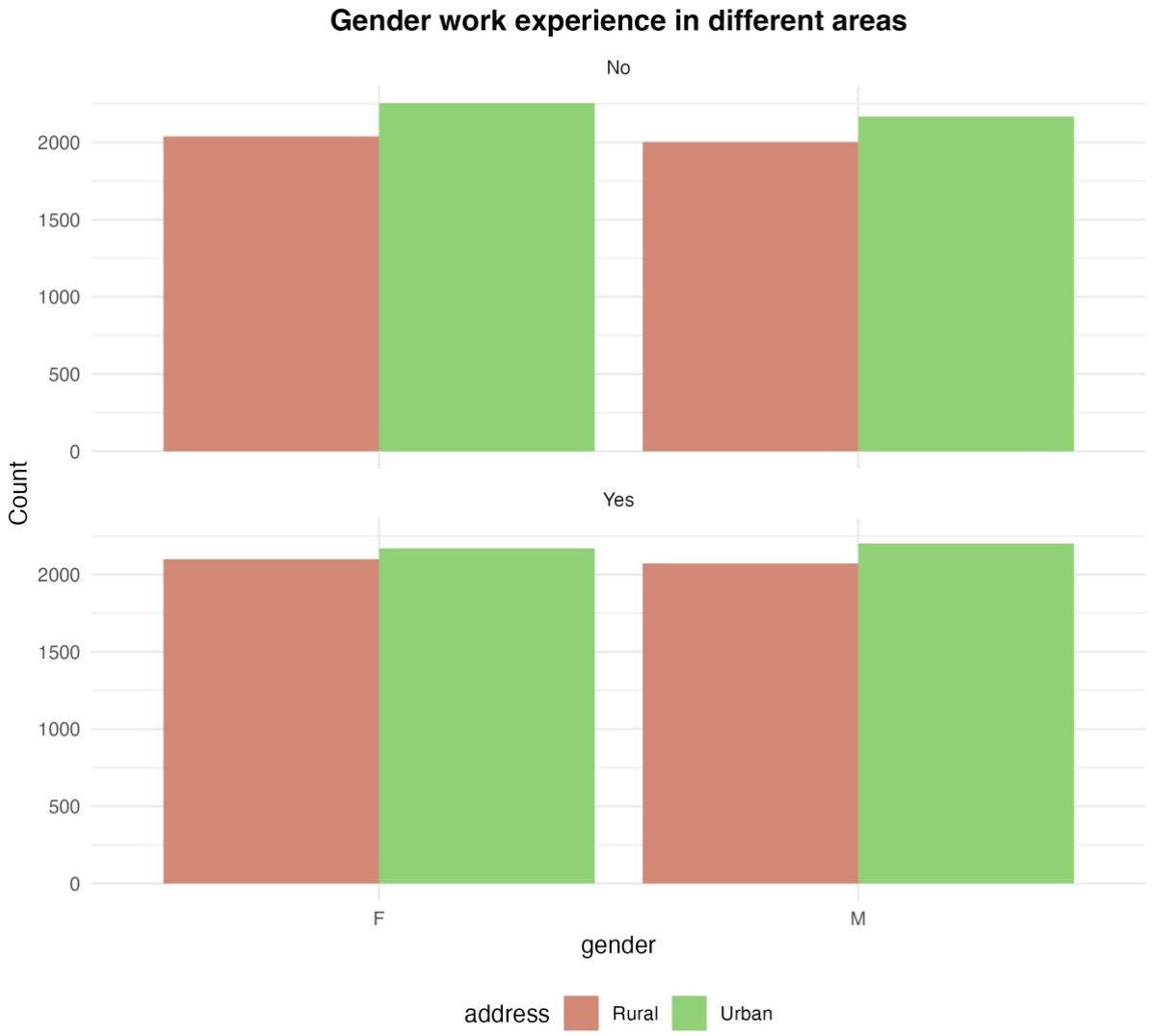


Figure 3.20.2 Relationship between gender, address and work experience graph

The data suggests that there is no significant difference between males and females in gaining work experience, regardless of their location. Both genders in rural and urban areas have similar numbers of people who gained and did not gain work experience. This implies that gender does not play a role in the likelihood of obtaining work experience. One possible reason for this could be that the job market is relatively equal for males and females, and job opportunities are based on qualifications and skills rather than gender. Another explanation could be that the sample size of this dataset is not large enough to detect any potential gender differences in gaining work experience. Overall, the data suggests that gender is not a significant factor in gaining work experience, and both males and females have equal opportunities in the job market.

Analysis 3.21: Is the job market providing a different salary for gender?

```
#====Analysis 3.21: Is the job market providing a different salary for gender?====
df_sal <- df %>%
  select(address, gender, salary) %>%
  filter(!is.na(salary)) %>%
  group_by(address, gender, salary) %>%
  count(salary)

ggplot(df_sal, aes(x=address, y = n, fill=address)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(salary ~ gender) +
  coord_flip() + ylim(0,500) + xlab("Area") +
  ylab("count") + ggtitle("Salary of gender in different areas")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/3p21.png")
```

Figure 3.21.1 Relationship between salary and gender code

In this analysis, we are selecting the columns "address", "gender" and "salary" from the data frame and filtering out the missing values of the salary column. We are then grouping the data by the "address", "gender", and "salary" columns and calculating the count of salaries in each group.

We are then creating a grouped bar plot using ggplot where the x-axis represents the different areas, and the y-axis represents the count of salaries. The plot is faceted by salary and gender, and the bars are filled with different colors representing different areas. The plot is flipped horizontally, and the y-axis is limited to a range of 0 to 500. The plot is also given a title and axis labels and the color scheme and theme are customized using the scale_color_ipsum(), scale_fill_ipsum(), and theme_minimal() functions.

Figure 3.21.2 Relationship between salary and gender graph

It's important to note that the data does show that the distribution of salary is similar between genders in both rural and urban areas, meaning that there is no discrimination in the job market. Other factors, such as education level, work experience, and negotiation skills, could also contribute to differences in salary between genders. It's also possible that discrimination exists but is not reflected in this particular dataset.

That being said, it is a positive sign that the data shows no significant difference in salary distribution between genders. This could be due to various reasons such as an increase in awareness and policies around gender equality, a more diverse pool of candidates applying for jobs, or a shift in societal attitudes towards gender roles. Overall, this data suggests that the job market is becoming more inclusive and merit-based, which is a positive development.

Summary for Question 3

As the analysis result above, we conclude that gender discrimination has not occurred in the areas included in this dataset, which is a good symbol that the people and employers, schools probably all highly emphasize one's skills, techniques, and other indicators that can bring revenue to the organization instead of having discrimination for any of the genders.

Question 4: Does the age of individuals affect their current circumstances?

Analysis 4.1: Does the presence of family support differ according to individuals' age?

```
#====Analysis 4.2:  Does age affect students' whether to involve themselves in extra-paid classes?====
df_age_group <- df %>% select(age_group, address, paid) %>%
  group_by(age_group, address, paid) %>% count(paid)
df_age_group
ggplot(df_age_group, aes(age_group, n, fill=age_group)) +
  geom_segment( aes(x=age_group, xend=age_group, y=0, yend=n), color="grey") +
  geom_point(color="orange", size=4, position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + facet_grid(paid~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Extra-paid classes for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p2.png")
```

Figure 4.1.1 Relationship between age and family support code

In this analysis, we are investigating whether the presence of family support differs based on individuals' age.

First, we create a new data frame df_age_group by selecting the age_group, address, and famsup columns from the original data frame df. We then group the new data frame by age_group, address, and famsup and count the occurrences of famsup.

Next, we use ggplot to create a segmented bar chart with age_group on the x-axis and the count of famsup on the y-axis. We use geom_segment to create the bar segments and geom_point to create the points on the chart. We also use facet_grid to separate the chart by address and famsup. Finally, we add labels and titles to the chart using ylab, ggtitle, and theme.

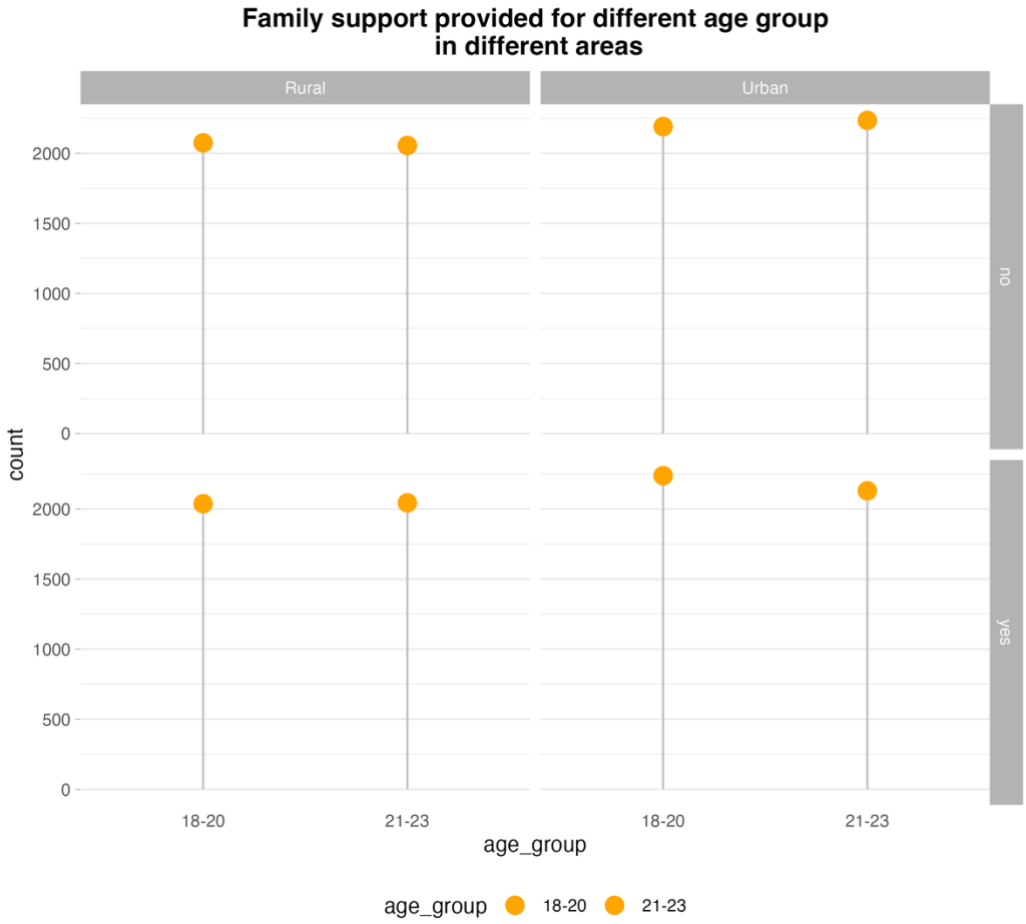


Figure 4.1.2 Relationship between age and family support graph

This dataset shows the distribution of students based on their age group, location (rural or urban), and whether they receive family support or not. The number of students who receive family support and those who do not is almost the same in both age groups (18-20 and 21-23) and in both rural and urban areas. This suggests that age does not have a significant impact on the presence of a student's family support.

The data indicate that male and female students in both rural and urban areas receive similar levels of family support when it comes to education. This could be due to the fact that the importance of education is widely recognized by families regardless of their gender. Additionally, it is possible that financial support for education is provided by the government or other organizations, which reduces the burden on families, so their families do not need to consider whether to sponsor their sons or daughters.

In summary, the data shows that gender does not appear to be a significant factor in getting family support in both rural and urban areas.

Analysis 4.2: Does age affect students' whether to involve themselves in extra-paid classes?

```
#Question 4: area age affect current circumstances====
#====Analysis 4.1:  Does the presence of family support differ according to individuals' age?=====
df_age_group <- df %>% select(age_group, address, famsup) %>%
  group_by(age_group, address, famsup) %>% count(famsup)
df_age_group
ggplot(df_age_group, aes(age_group, n, fill=age_group)) +
  geom_segment( aes(x=age_group, xend=age_group, y=0, yend=n), color="grey") +
  geom_point( color="orange", size=4, position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + facet_grid(famsup~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Family support provided for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p1.png")
```

Figure 4.2.1 Relationship between age and extra-paid classes code

In this analysis, we are exploring whether age affects students' involvement in extra-paid classes. We start by selecting the columns related to age group, address, and paid classes from the original dataframe. We then group the data by age group, address, and whether the students take extra-paid classes, and count the number of students falling into each group.

Next, we create a bar chart using ggplot2 library to visualize the data. The x-axis shows the age group, while the y-axis shows the count of students in each group. The bars are segmented by whether the students take extra-paid classes or not, with each segment colored according to the age group.

We also add some additional formatting to improve the visualization, such as removing the x-axis ticks and gridlines and adjusting the color scheme. Finally, we add a title to the chart to summarize the analysis.

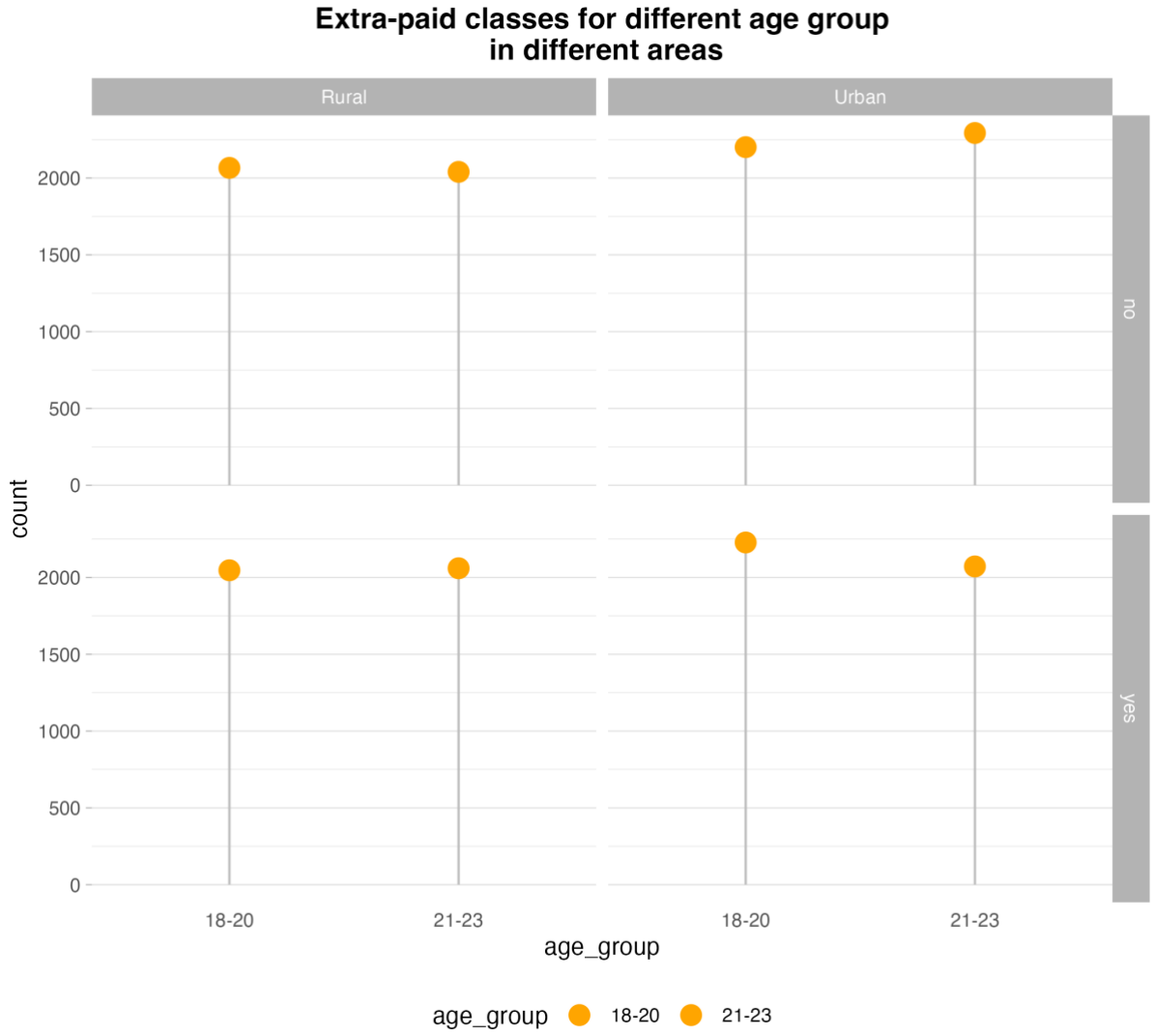


Figure 4.2.2 Relationship between age and extra-paid classes graph

In fact, when the students grow up, they will start to understand their position and what they want. This causes they might not be interested in joining many extra-paid classes like when they are still young.

However, the data shows the distribution of whether students aged 18-23 in rural and urban areas joining extra-paid classes average across all of the age groups in general.

In the 18-20 age group, slightly more students in urban areas joining extra-paid classes than those in rural areas, but the difference is not large. In the 21-23 age group, there is a slightly larger difference between urban and rural areas, with more students in urban areas joining in extra-paid classes than those in rural areas.

Overall, the data suggests that there may not be significant differences in joining extra-paid classes for students of different age groups in rural and urban areas.

Analysis 4.3: Is there any difference for students in different age groups in having internet access?

```
#====Analysis 4.3:  Is there any difference for students in different age groups in having internet access?====
df_age_group <- df %>% select(age_group, address, internet) %>%
  group_by(age_group, address, internet) %>% count(internet)
df_age_group
ggplot(df_age_group, aes(age_group, n, fill=age_group)) +
  geom_segment( aes(x=age_group, xend=age_group, y=0, yend=n), color="grey") +
  geom_point(color="orange", size=4, position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + facet_grid(internet~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Internet access for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p3.png")
```

Figure 4.3.1 Relationship between age and internet access code

In this analysis, we investigate whether there is any difference for students in different age groups in having internet access.

We start by selecting three variables from the dataset: age_group, address, and internet. It then groups the data by age_group, address, and internet, and counts the number of students falling into each combination of these variables.Next, creates a segmented bar chart using ggplot. The x-axis shows age groups, and the y-axis shows the count of students. The bars are segmented by the variable 'internet'. The bars are created using a combination of geom_segment and geom_point functions.The facet_grid function is used to display the segmented bars separately for each value of the 'address' variable, and ylab sets the label for the y-axis.Finally, various theme functions are used to customize the appearance of the chart, including the title, legend position, and plot background.

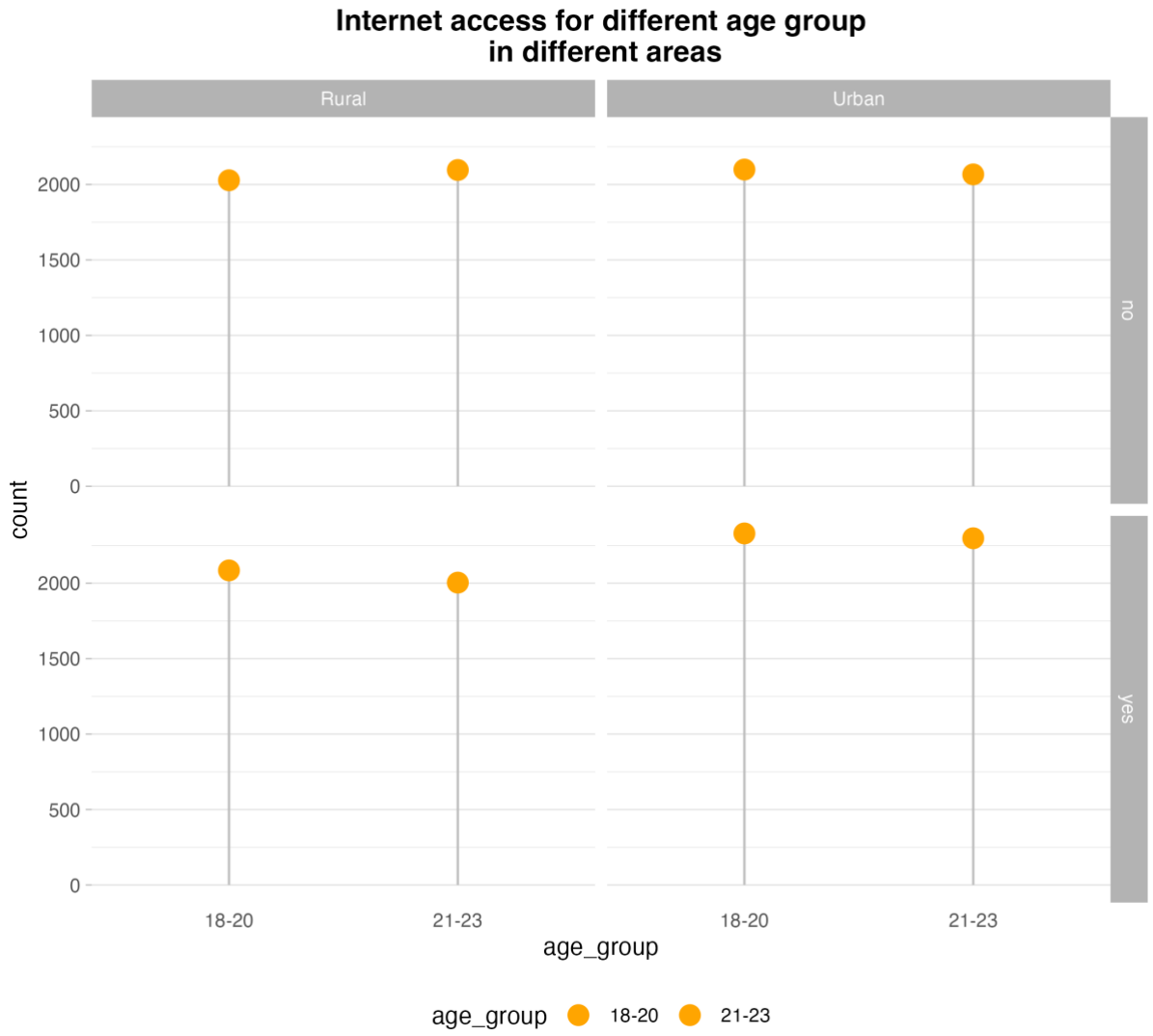


Figure 4.3.2 Relationship between age and internet access graph

The data shows the distribution of whether students aged 18-23 in rural and urban areas have internet access. In general, it seems that a similar proportion of students in rural and urban areas have internet access, and there is not a significant difference between age groups.

Overall, the data suggests that there may not be significant differences in access to the internet for students of different age groups in rural and urban areas.

Analysis 4.4: Do students of different age groups from different areas have different opinions on joining extra-curricular activities?

```
#====Analysis 4.4:  Do students of different age groups from different areas have different opinions on joining extra-curricular activities?====
df_age_group <- df %>% select(age_group, address, activities) %>%
  group_by(age_group, address, activities) %>% count(activities)
df_age_group
ggplot(df_age_group, aes(activities, n, fill=age_group)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_light() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank()
  ) + ylab("count") + facet_grid(~address)+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Extra-curricular activities for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p4.png")
```

Figure 4.4.1 Relationship between age and joining extra-curricular activities code

In this code, we are analyzing whether there is a difference in the opinions of students belonging to different age groups from different areas in joining extra-curricular activities.

We first select the columns of interest (age_group, address, and activities) from the original data frame and group the resulting data frame by age_group, address, and activities, then count the number of occurrences of each unique combination of the three variables. We store the resulting data frame in "df_age_group." Next, we create a bar plot using ggplot with activities on the x-axis and the count of each unique combination of age_group, address, and activities on the y-axis. We use the fill argument to color the bars by age_group, and we use facet_grid to separate the plots by address. Finally, we add appropriate themes and labels to the plot and save it as an image.

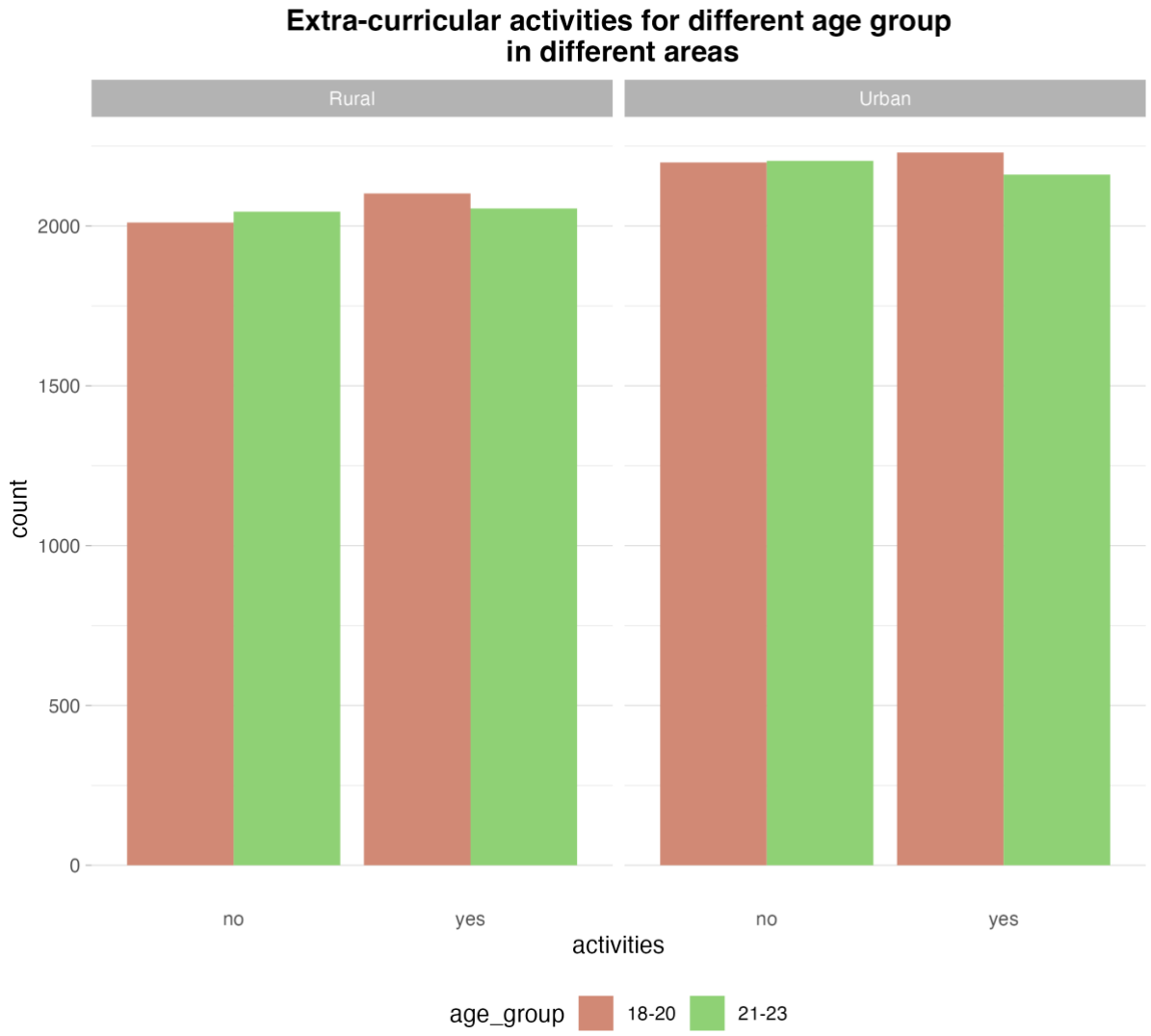


Figure 4.4.2 Relationship between age and joining extra-curricular activities graph

The data shows the distribution of whether students aged 18-23 in rural and urban areas join extra-curricular activities. In general, it seems that a similar proportion of students in rural and urban areas join extra-curricular activities, and there is not a significant difference between age groups.

In the 18-20 age group, slightly more students in urban areas joining extra-curricular activities than those in rural areas, but the difference is not large. In the 21-23 age group, there is a slight difference between urban and rural areas, with more students in urban areas joining extra-curricular activities than those in rural areas.

Overall, the data suggests that there may not be significant differences in joining extra-curricular activities for students of different ages in rural and urban areas.

Analysis 4.5: Do different age groups of students choose different secondary school boards in the past?

```
#====Analysis 4.5:  Do different age groups of students choose different secondary school boards in the past?====
df_age_group <- df %>% select(address, age_group, ssc_b) %>% count(address, age_group, ssc_b)
df_age_group
ggplot(df_age_group, aes(ssc_b, y=n)) +
  geom_segment( aes(x=ssc_b, xend=ssc_b, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(age_group~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() +
  ggtitle("Secondary school board for different age group
 in different areas")+theme_minimal()+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p5.png")
```

Figure 4.5.1 Relationship between age and joining secondary school board code

In this code, we are analyzing whether there is a difference in the opinions of students belonging to different age groups from different areas in joining extra-curricular activities.

We first select the columns of interest (age_group, address, and activities) from the original data frame and group the resulting data frame by age_group, address, and activities, then count the number of occurrences of each unique combination of the three variables. We store the resulting data frame in "df_age_group."

Next, we create a bar plot using ggplot with activities on the x-axis and the count of each unique combination of age_group, address, and activities on the y-axis. We use the fill argument to color the bars by age_group, and we use facet_grid to separate the plots by address. Finally, we add appropriate themes and labels to the plot and save it as an image.

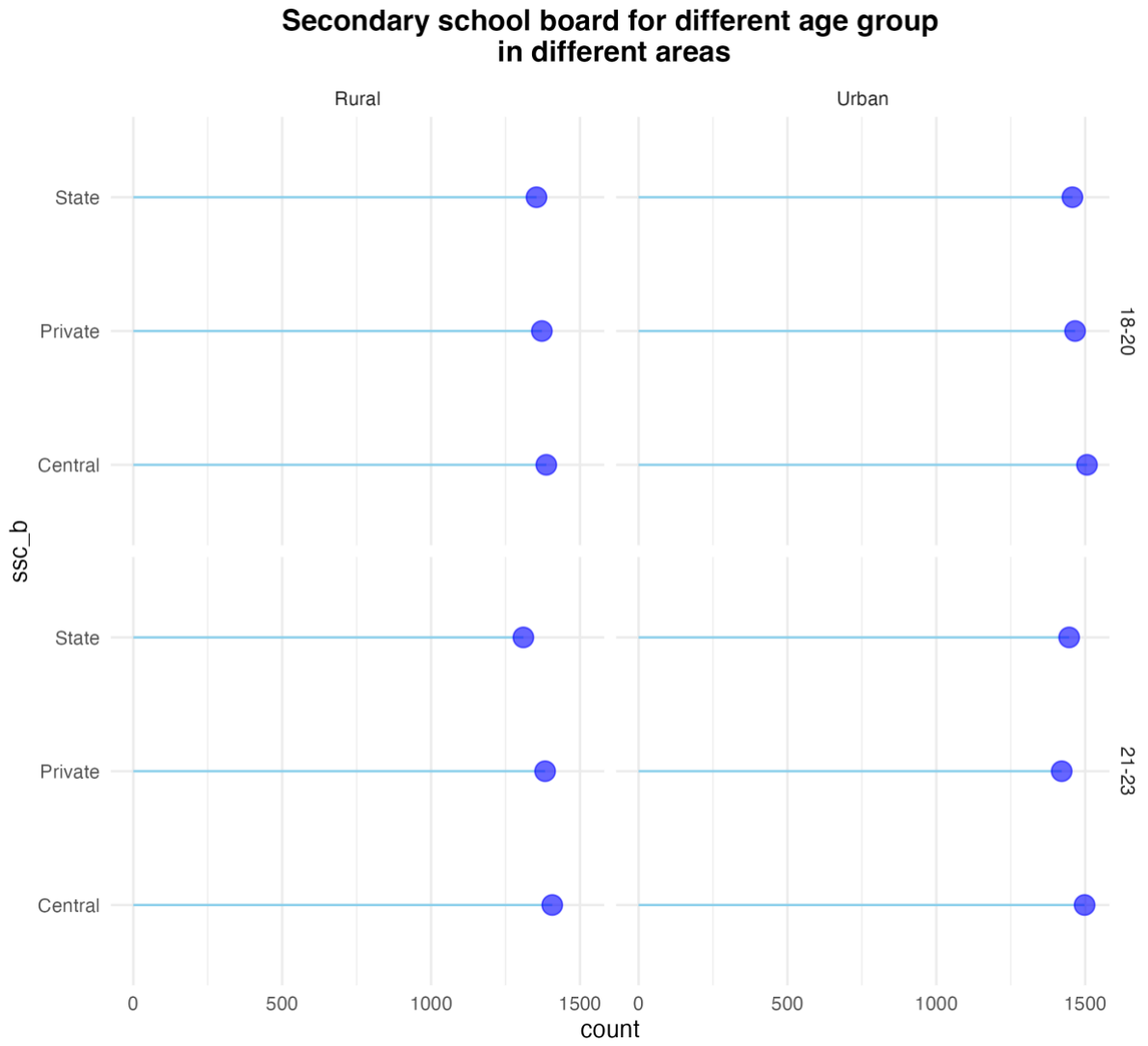


Figure 4.5.2 Relationship between age and joining secondary school board graph

The table shows the number of students enrolled in different secondary schools in rural and urban areas, divided into two age groups (18-20 and 21-23). This analysis is to examine the preference for choosing secondary school boards during their time. In fact, in different years, students or their parents might have different opinions in choosing secondary schools. Especially at that time, students have no absolute freedom in choosing their schools, while parents hold the bigger power in choosing secondary schools.

In the lollipop chart, the schools are categorized into three types based on their board affiliation: Central, Private, and State. In general, the number of students enrolled in Central schools, Private and State schools are similar across age groups. The number of students in the 18-20 age group is higher than the 21-23 age group in most schools, indicating that more students enroll in secondary. The conclusion of this analysis is that students of different age groups are not having any preference in choosing their secondary school board in the past.

Analysis 4.6: Do different age groups of students choose different high school boards in the past?

```
#====Analysis 4.6:  Do different age groups of students choose different high school boards in the past?====
df_age_group <- df %>% select(address, age_group, hsc_b) %>% count(address, age_group, hsc_b)
df_age_group
ggplot(df_age_group, aes(hsc_b, y=n)) +
  geom_segment( aes(x=hsc_b, xend=hsc_b, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(age_group~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("High school board for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p6.png")
```

Figure 4.6.1 Relationship between age and joining high school board code

In this analysis, we are exploring whether students from different age groups in different areas choose different high school boards in the past. We first use the select() function to choose the columns "address", "age_group", and "hsc_b" from the dataframe "df". We then use the count() function to count the number of occurrences of each combination of "address", "age_group", and "hsc_b". The resulting dataframe is stored in "df_age_group".

We then use ggplot() to create a bar chart with "hsc_b" on the x-axis and the count on the y-axis. We use geom_segment() to add vertical lines for each category, and geom_point() to add a point for each count. We use coord_flip() to flip the x and y axes and display the bars horizontally. We use facet_grid() to separate the plots by age group and address. We also use ylab() to add a label to the y-axis and ggtitle() to add a title to the plot.

Finally, we use theme() to customize the appearance of the plot, including removing the y-axis ticks and changing the color palette.

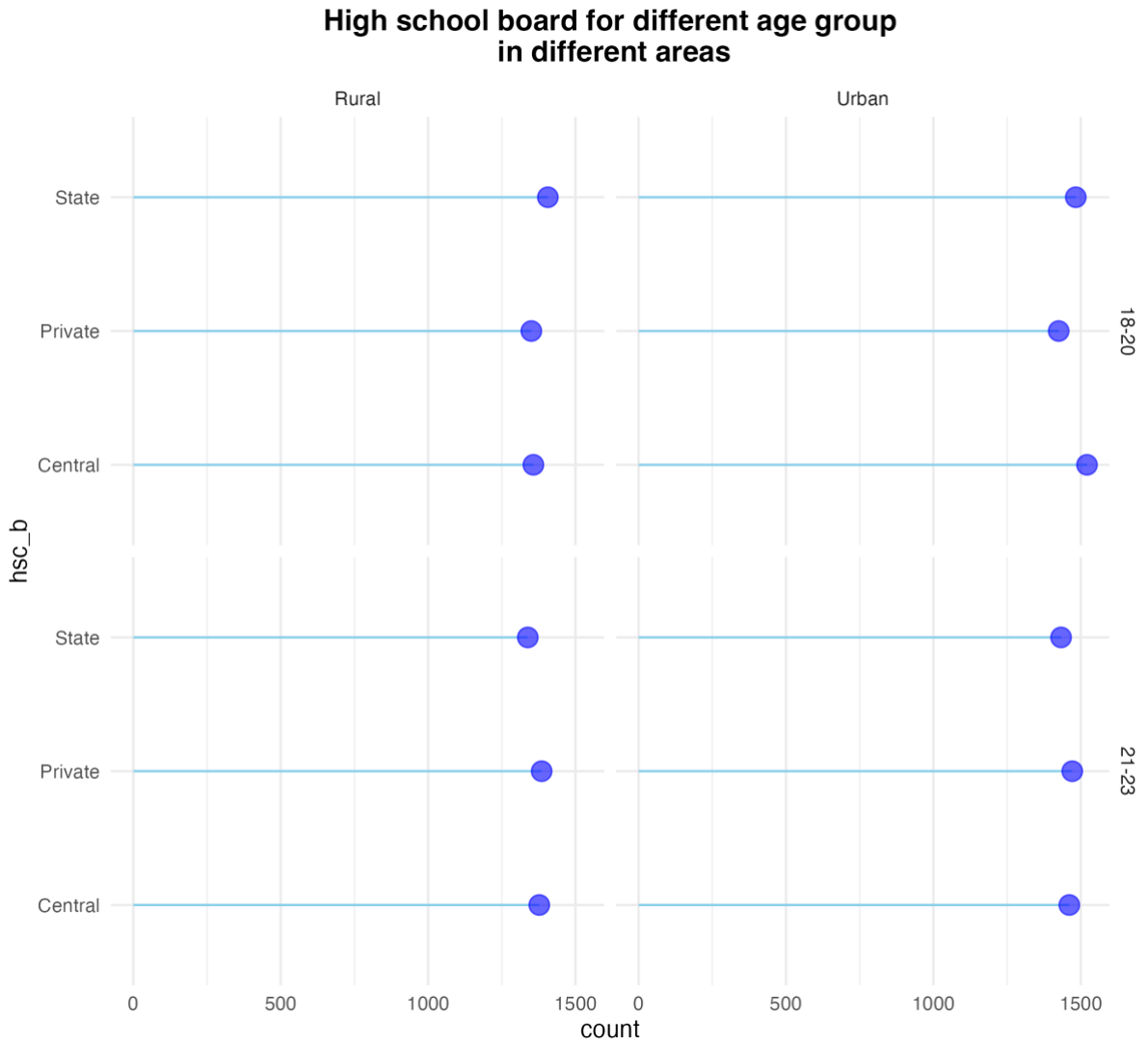


Figure 4.6.2 Relationship between age and joining high school board graph

The data shows the distribution of the number of students in high school based on their age group, address, and board of education. From the data, we can see that there is no significant difference in the number of students attending different types of high schools, whether they are from rural or urban areas. In both age groups (18-20 and 21-23), the number of students attending central, private, and state schools is distributed similarly. This indicates that students do not have a particular preference for a specific board of education, and the government may have set up education policies to make sure all types of schools provide similar quality of education.

In summary, the data suggest that the choice of the high school board is not influenced by the student's age or address, and the government may have established policies to ensure that all types of schools provide similar quality of education.

Analysis 4.7: Do different age groups of students choose different high school specializations in the past?

```
#====Analysis 4.7:  Do different age groups of students choose different high school specializations in the past?====
df_age_group <- df %>% select(address, age_group, hsc_s) %>% count(address, age_group, hsc_s)
df_age_group
ggplot(df_age_group, aes(hsc_s, y=n)) +
  geom_segment( aes(x=hsc_s, xend=hsc_s, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(age_group~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("High school specialisation for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p7.png")
```

Figure 4.7.1 Relationship between age and high school specialization code

First, we select the columns 'address', 'age_group', and 'hsc_s' from the dataset and count the number of occurrences of each combination of these variables using the count() function. We store the resulting dataframe in the df_age_group variable.

Then, we create a bar plot using ggplot(). We set the x-axis to be the high school specializations (hsc_s) and the y-axis to be the count (n). We use geom_segment() to draw a line segment for each specialization, with the length of the segment representing the count of that specialization. We use geom_point() to add a blue dot at the end of each segment to make the plot more readable.

We flip the x- and y-axes using coord_flip() to make the plot horizontal. We use facet_grid() to create separate plots for each combination of age_group and address. We label the y-axis as "count" using ylab(). We use the scale_color_ipsum() and scale_fill_ipsum() functions to set the colors to a pleasing palette. We set the theme to minimal using theme_minimal(), and add a title to the plot using ggtitle(). We set the legend position to the bottom using theme(legend.position = "bottom"), and set the plot title to be centered using plot.title = element_text(hjust=0.5, face = "bold"). Finally, we set the plot background to white using plot.background = element_rect(fill = "white", color="white").

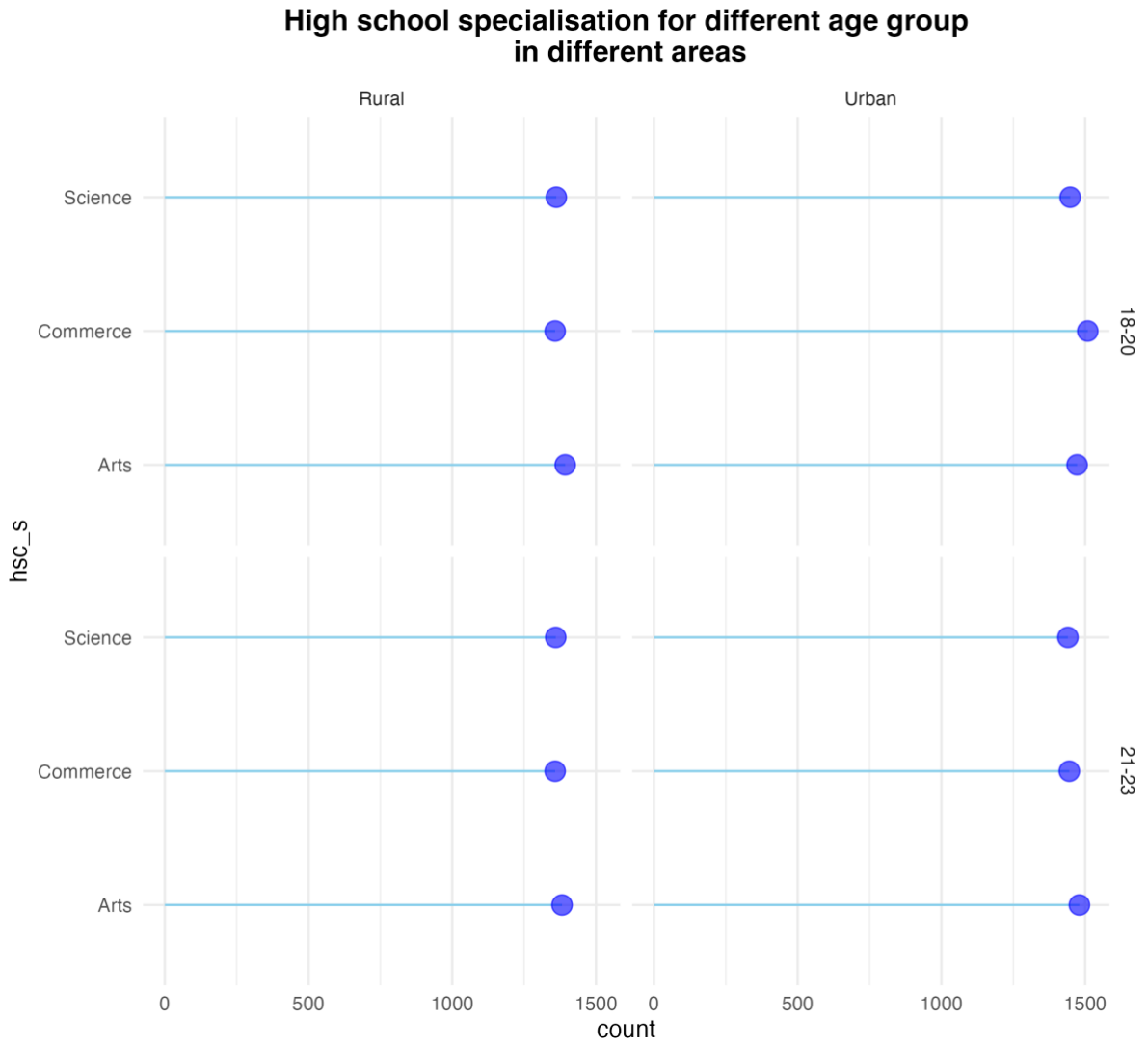


Figure 4.7.2 Relationship between age and high school specialization graph

The lollipop chart shows the distribution of students across different specializations in high school, grouped by address and age group. The number of students in each category is relatively similar, with no particular trend in any direction. This suggests that students are not significantly influenced by their age, location, or any other factor in choosing their specialization.

In summary, the data suggest that students have no clear preference for any particular specialization in high school, regardless of their age or location. Further research would be needed to understand why this is the case and what factors might influence students in their decision-making process.

Analysis 4.8: Do different age groups of students choose different degree types in the past?

```
#====Analysis 4.8:  Do different age groups of students choose different degree types in the past?====
df_age_group <- df %>% select(address, age_group, degree_t) %>% count(address, age_group, degree_t)
df_age_group
ggplot(df_age_group, aes(degree_t, y=n)) +
  geom_segment( aes(x=degree_t, xend=degree_t, y=0, yend=n), color="skyblue") +
  geom_point( size=5, color="red", fill=alpha("orange", 0.3), alpha=0.7, shape=21, stroke=2)+
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(age_group~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  ggtitle("Degree types for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p8.png")
```

Figure 4.8.1 Relationship between age and degree types code

In this analysis, we are looking at whether different age groups of students choose different degree types in the past. To do this, we first select the relevant columns from the data frame using the select function and then count the number of occurrences of each combination of address, age group, and degree type using the count function.

Next, we create a bar chart using ggplot2, where the x-axis represents the different degree types, and the y-axis represents the count of each combination of address, age group, and degree type. We use geom_segment to draw vertical lines for each bar and geom_point to add a point for each combination of address, age group, and degree type. We also use facet_grid to create separate plots for each combination of age group and address. Finally, we add appropriate labels and titles to the plot using various theme elements.

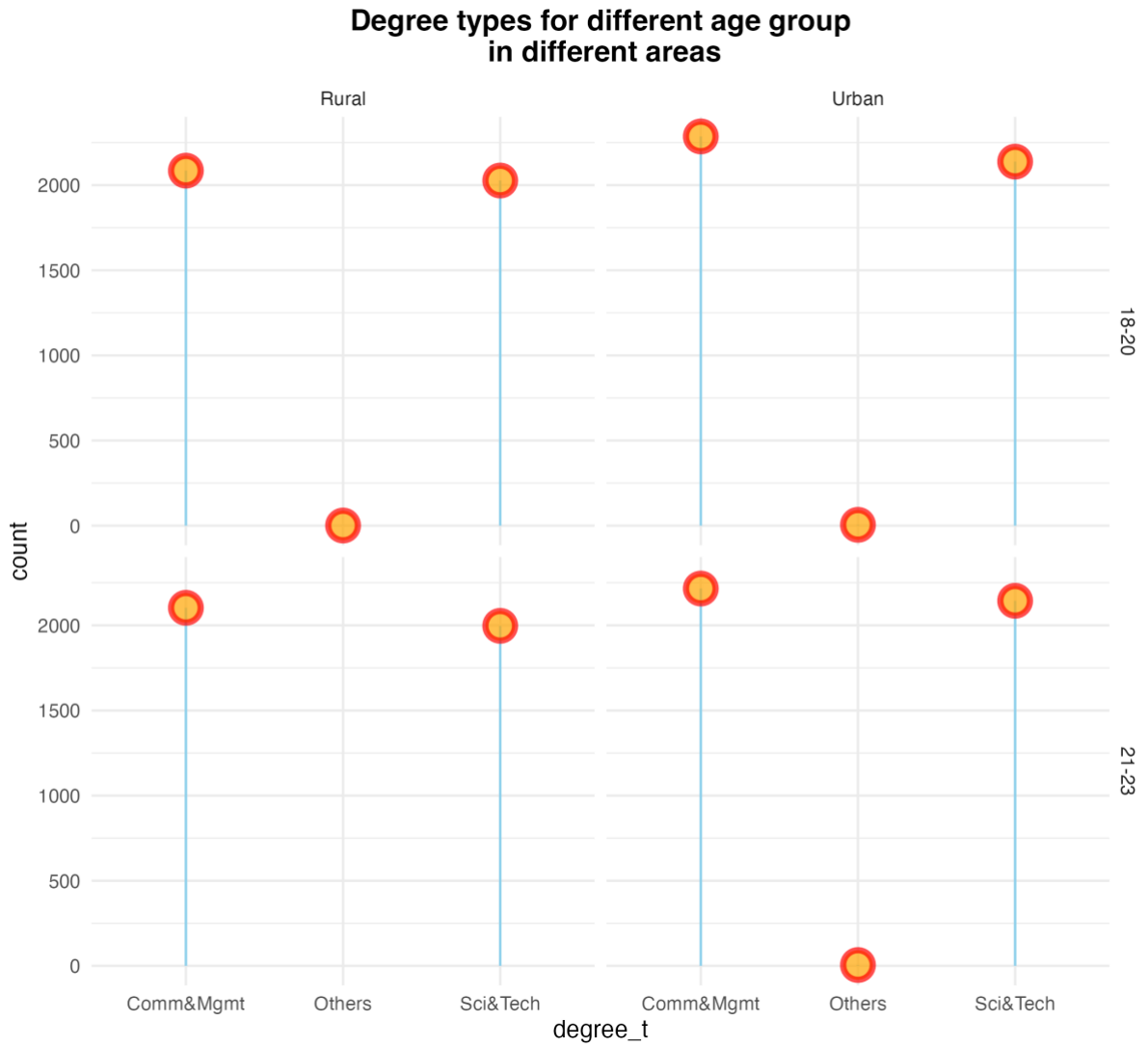


Figure 4.8.2 Relationship between age and degree types graph

The graph above shows the distribution of students based on their degree type and their address (urban or rural) for two different age groups (18-20 and 21-23). The majority of students in both age groups and areas are pursuing a degree in Commerce and Management (Comm&Mgmt). However, there is a small number of students pursuing a degree in Other fields.

We can see that the majority of students in both age groups are pursuing degrees in Comm&Mgmt. Meanwhile, there is a number of students pursuing degrees in Sci&Tech for both age groups. On the other hand, the number of students pursuing degrees in other fields is very small and does not show any clear trend as well.

Overall, it seems that there is not a significant difference in the preference for degree types between the two age groups. It is unclear why this difference exists, and further analysis would be needed to explore the reasons behind it.

Analysis 4.9: Do different age groups of students have their own preferences in choosing different specializations?

```
#====Analysis 4.9:  Do different age groups of students have their own preferences in choosing different specializations?====
df_age_group <- df %>% select(address, age_group, status) %>% count(address, age_group, status)
df_age_group
ggplot(df_age_group, aes(status, y=n)) +
  geom_segment( aes(x=status, xend=status, y=0, yend=n), color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  theme_light() +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) + facet_grid(age_group~address) + ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum()+
  ggtitle("Status for different age group
 in different areas")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p9.png")
```

Figure 4.9.1 Relationship between age and specialization code

In this analysis, we are exploring whether different age groups of students have their own preferences in choosing different specializations. To do this, we first select the relevant columns from the original dataset, which include the student's address, age group, and their specialization status. We then count the number of occurrences for each unique combination of address, age group, and specialization status using the count function.

Next, we create a bar plot using ggplot2 to visualize the results. We map the specialization status to the x-axis and the count to the y-axis. We use geom_segment to create bars for each category, and geom_point to show the count for each category. We also use facet_grid to create a separate plot for each age group and address combination.

Finally, we add appropriate titles and labels to the plot using ggtitle, ylab, and theme functions. We use a light theme for the plot and adjust the appearance of the gridlines and axis ticks to create a cleaner look.

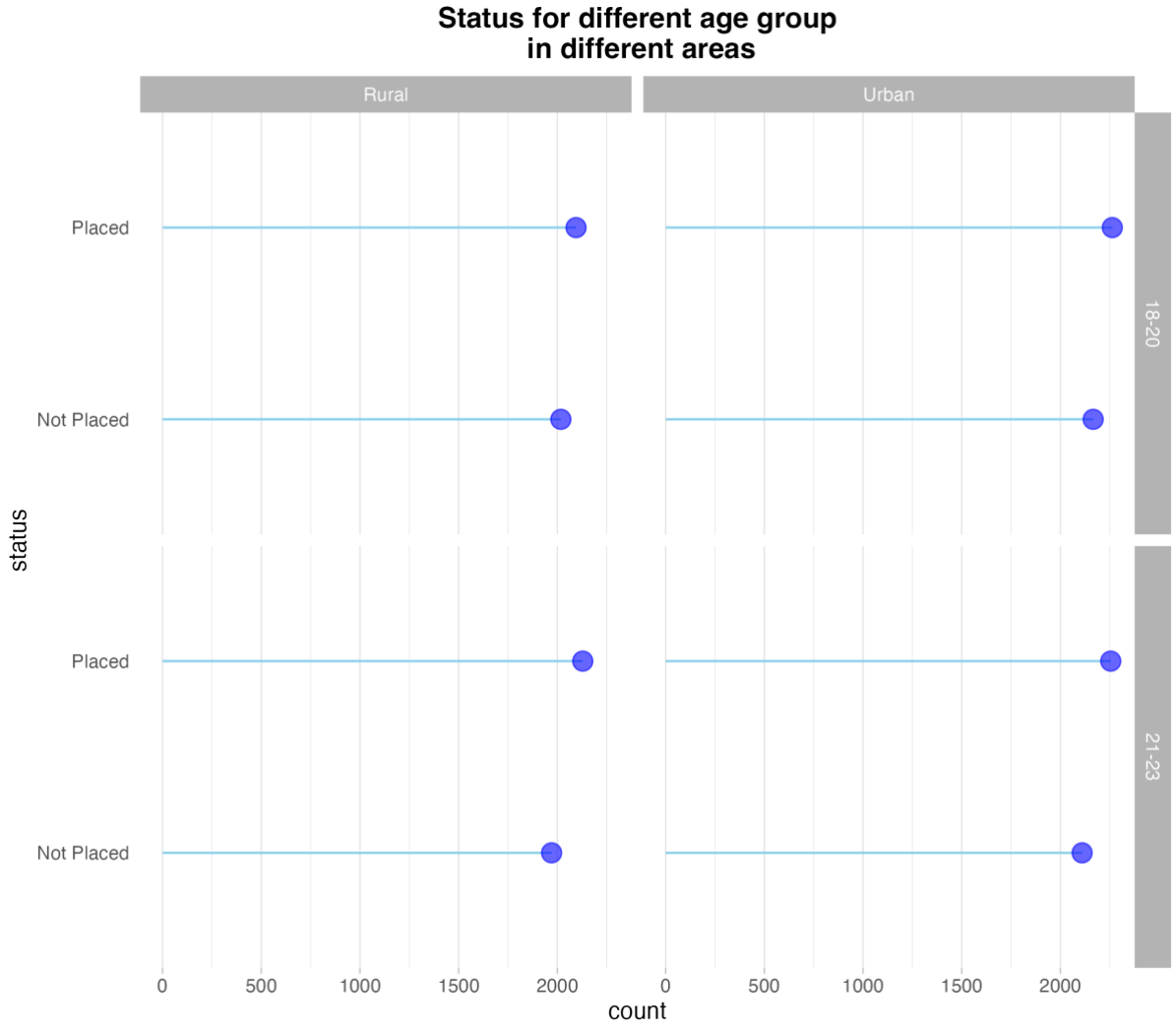


Figure 4.9.2 Relationship between age and specialization graph

Looking at the lollipop chart, it appears that age does not have a significant effect on whether a student is placed or not. None of the points in the graph is clearly higher than the other bars. Both the 18-20 and 21-23 age groups have similar numbers of students who are placed and not placed, regardless of their location (rural or urban).

One possible reason for this could be that the job market in this particular area is not affected by age, and employers are primarily concerned with the qualifications and skills of the candidates rather than their age.

In summary, the data shows that age does not seem to be a significant factor in whether a student is placed or not placed. The numbers of placed and not placed students are similar for both age groups and for both rural and urban students. This suggests that employers in the area may be primarily concerned with the qualifications and skills of candidates rather than their age.

Analysis 4.10: Do different age groups of students from different areas has their own preferences in choosing the specialization?

```
#====Analysis 4.10: Do different age groups of students from different areas has their own preferences in choosing the specialization?====
df_age_group <- df %>% select(address, age_group, specialisation) %>%
  count(address, age_group, specialisation)
df_age_group
to_add <- matrix(NA, empty_bar, ncol(df_age_group))
colnames(to_add) <- colnames(df_age_group)
df_age_group <- rbind(df_age_group, to_add)
df_age_group$id <- seq(1, nrow(df_age_group))
df_age_group
df_age_group <- df_age_group %>% mutate(angle = 90 - 360 * (id-0.5) /nrow(df_age_group)) %>%
  mutate(hjust = ifelse(angle < -90, 1, 0)) %>%
  mutate(angle = ifelse(angle == 90, angle+180, angle))
df_age_group

ggplot(df_age_group, aes(x=as.factor(id), y=n, fill=address, colour=age_group)) +
  geom_bar(stat="identity") +
  ylim(-1900,2500) +
  theme_minimal() +
  theme(
    axis.text = element_blank(),
    axis.title = element_blank(),
    panel.grid = element_blank(),
    plot.margin = unit(rep(-1,4), "cm")
  ) +
  coord_polar(start = 0) +  theme_void()+
  geom_text(data=df_age_group, aes(x=id, y=n+10,
                                   label=paste(address, age_group, specialisation, sep = " "), hjust=hjust),
            color="black", fontface="bold",alpha=0.6, size=2.5, angle= df_age_group$angle,
            inherit.aes = FALSE )+
  theme(legend.position = "right", plot.title = element_text(hjust = 0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white")) +
  ggtitle("Specialisation of different age group in different areas")
ggsave("~/Graph/4p10.png")
```

Figure 4.10.1 Relationship between age, address and specialization code

In this analysis, we first select the columns "address", "age_group", and "specialisation" from the data frame using the dplyr package's select function. We then group the data by "address", "age_group", and "specialisation" and count the number of occurrences using the count function.

Next, we add a new row of missing values to the data frame using the matrix function and rbind to ensure that we have enough space to create a circular bar plot. We then create an "id" column to give each row a unique identifier.

We use the mutate function to create a new column called "angle" to calculate the angle for each slice of the pie chart. We also create a new column called "hjust" to align the text for each slice of the pie chart.

Finally, we use the ggplot2 package to create a circular bar plot. We use the geom_bar function to create the bar chart, set the y-axis limits to ensure the chart is centered, and remove unnecessary elements from the chart using theme functions. We use coord_polar to create a circular chart, and we add labels to the chart using geom_text. We set the legend position to the right of the chart and add a title using ggtitle.

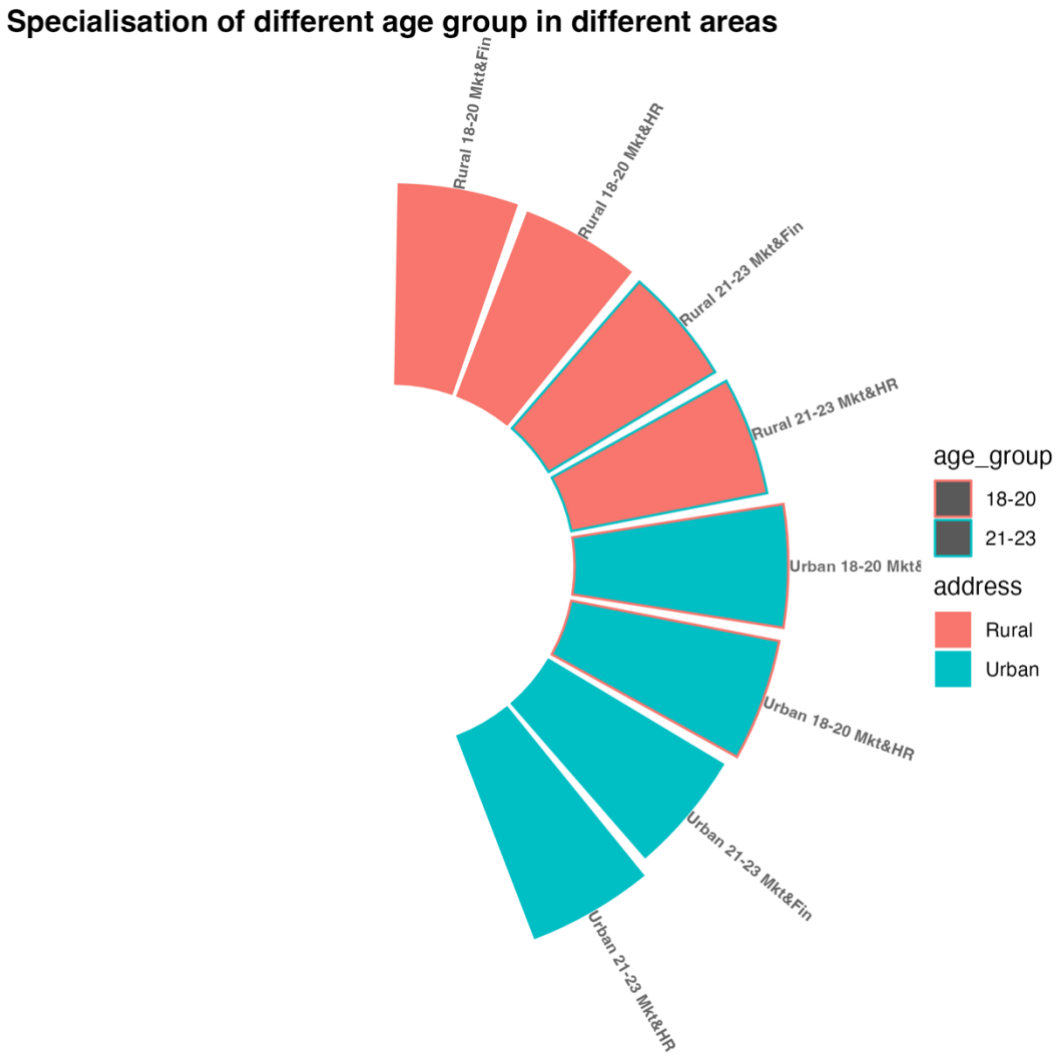


Figure 4.10.2 Relationship between age, address and specialization graph

From the given data, we can see that both age groups 18-20 and 21-23 have a similar preference for specializations in terms of Marketing and Finance (Mkt&Fin) and Marketing and Human Resources (Mkt&HR). However, the number of students in Mkt&Fin is a little bit higher in the 18-20 age group, while Mkt&HR is slightly higher in the 21-23 age group. In summary, since no bar is higher than any other bar in the circular bar plot, we conclude that the students with different age groups have not shown any tendency, interests or aspirations in choosing their specialization.

Analysis 4.11: Do different age groups of students have equality in gaining work experience?

```
#====Analysis 4.11: Do different age groups of students have equality in gaining work experience?====
df_exp <- df %>% select(workex, age_group, address) %>%
  pivot_longer(cols = workex, names_to = "workex", values_to = "value") %>%
  group_by(age_group, address) %>% count(value)
df_exp
ggplot(df_exp, aes(x = age_group, y = n, fill = address)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~value, ncol = 1) + ylab("count") +
  ggtitle("Work experience for different age group in different areas") +
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p11.png")
```

Figure 4.11.1 Relationship between age and work experience code

In this analysis, we are exploring whether different age groups of students have equality in gaining work experience. We first extract the columns related to work experience, age group, and address from the original dataset using the select() function. Then we use the pivot_longer() function to reshape the data from wide to long format so that we can count the number of occurrences of each work experience category for each age group and address combination. We group the data by age group, address, and work experience using the group_by() function and count the number of occurrences using the count() function.

We then create a bar plot using ggplot() and geom_bar() to visualize the number of students in each age group and address combination who have work experience or not. We use facet_wrap() to create separate plots for each work experience category. We set the y-axis label and the plot title using ylab() and ggtitle(), respectively. We also use the scale_color_ipsum() and scale_fill_ipsum() functions to set the color palette and the theme_minimal() function to set the theme. Finally, we adjust the position of the legend and the plot title using the theme() function.

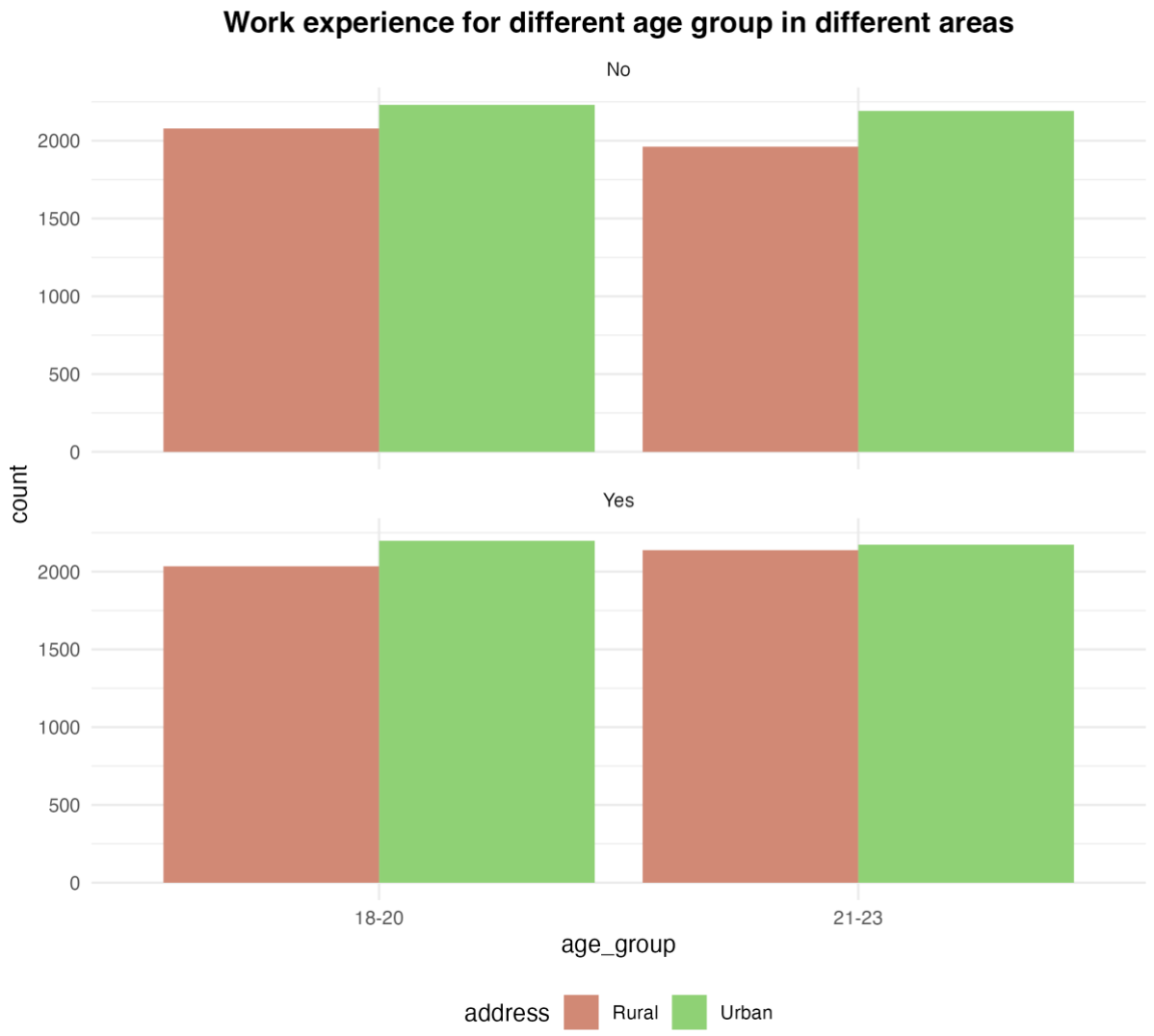


Figure 4.11.2 Relationship between age and work experience graph

The data shows the number of students based on their age group and work experience availability. Looking at the data, it appears that there is no clear relationship between age group and work experience availability. Across all age groups, there are similar numbers of students with and without work experience.

However, it is more logical to conclude that students in the older age group (21-23) may have more work experience compared to students in the younger age group (18-20) as they may have had more time to gain work experience. Additionally, students in urban areas may have more access to job opportunities, leading to a higher number of students with work experience compared to those in rural areas. However, this phenomenon has not been shown in this dataset. Overall, it seems that work experience is not tied to age group.

Analysis 4.12: Does age affect the salary of individuals in different areas?

```
#====Analysis 4.12: Does age affect the salary of individuals in different areas?====
df_sal <- df %>%
  select(address, age_group, salary) %>%
  filter(!is.na(salary)) %>%
  group_by(address, age_group, salary) %>%
  count(salary)
view(df_sal)
ggplot(df_sal, aes(x=address, y = n, fill=address)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(salary ~ age_group) +
  coord_flip() + ylim(0,500) + xlab("Area") + #Extra features
  ylab("count") + ggtitle("Salary of different age group in different areas")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/4p12.png")
```

Figure 4.12.1 Relationship between age, address and salary code

In this analysis, we are interested in studying the relationship between age, salary, and location. To do this, we first select the relevant columns (address, age_group, and salary) from our dataframe and remove any rows that have missing salary values. We then group the data by address, age_group, and salary, and count the number of occurrences for each salary value.

Next, we use ggplot to create a bar plot where the x-axis represents the different areas, the y-axis represents the count of individuals in each salary group, and the fill color represents the different areas. We use facet_grid to separate the different age groups by the different salary ranges. We flip the coordinate axis using coord_flip() and set the y-axis limit to 0-500. Finally, we add labels and a title to the plot using xlab(), ylab(), and ggtitle() functions.

Figure 4.12.2 Relationship between age, address and salary graph

Based on the graph given, we can conclude that the distribution of salary levels for students in the age groups of 18-20 and 21-23 is similar regardless of their area. One possible reason for this could be that the job market is not emphasizing hiring employees and providing salaries based on their age but on knowledge and other factors. In summary, the data suggest that the age of students is not a significant factor in determining their salary level and that the distribution of income is consistent across these two age groups.

Summary for Question 4

Does the age of individuals affect students' current circumstances? The above analysis has all points out that age is not a point that should be blamed when someone is not getting a job in this dataset. Age has never become a constraint for one to pursue a job, desired schools, specialization, and occupation, but probably their self-attitude will be the culprit that holds them down from being successful.

Question 5: What is the element that will affect individuals' salary?

Analysis 5.1: Does the individuals' gender affect the salary provided by their employers?

```
#Question 5: salary_class effect====
#====Analysis 5.1:  Does the individuals' gender affect the salary provided by their employers?====
df_sal <- df %>% filter(salary != 0)%>% select(gender, salary_class) %>%
  group_by(gender, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)
title("Effect of gender to salary",
      cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p1p1.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=gender)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Effect of gender to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p1p2.png")
```

Figure 5.1.1 Relationship between gender and salary code

First, we select the necessary columns "gender" and "salary_class" from the dataframe and group the data by "gender" and "salary_class". We then count the number of occurrences of each salary class for each gender group.

Next, we create a chord diagram of the counts using the chordDiagram function. We also add a title to the chord diagram using the title function and save the plot as a png file using dev.copy and dev.off functions.

Finally, we create a bar plot using ggplot to show the effect of gender on salary. We use the geom_bar function to create the plot and use the fill argument to differentiate the bars based on gender. We also add a title, label the axes and adjust the theme to make the plot visually appealing.

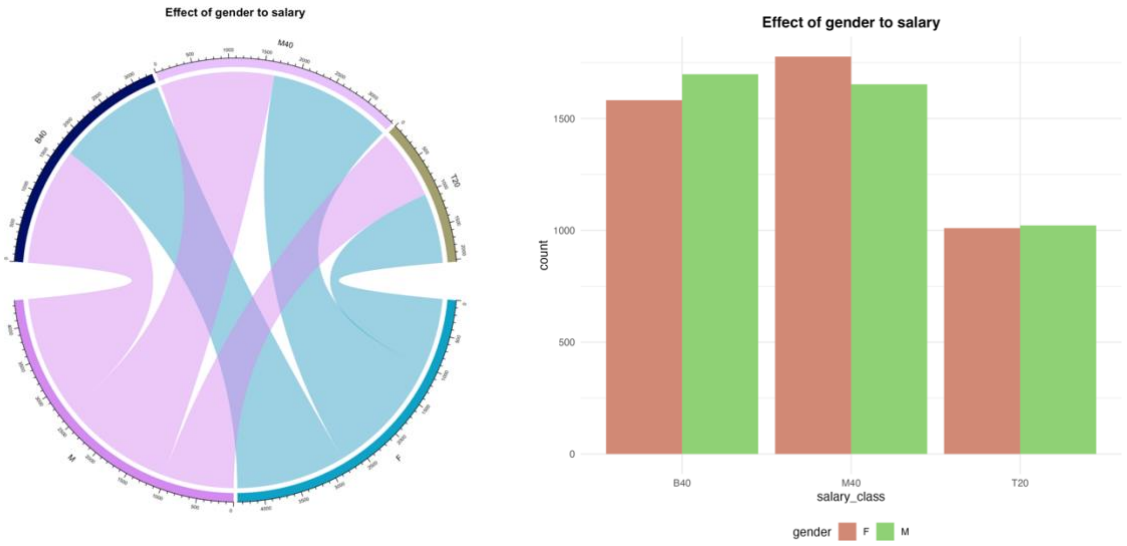


Figure 5.1.2 Relationship between gender and salary graphs

Based on the given diagram, we can observe that the number of males and females in each salary class is roughly similar in the chord diagram while in the attached bar chart, we found that there is still a slight difference that more females in the M40 class and slightly more males in the B40 class. However, without additional information, we cannot draw any definitive conclusions about the relationship between gender and salary. Possible reasons for the observed patterns could include differences in education level, work experience, job type or industry, negotiating skills, and gender bias in hiring or promotion. Further analysis and investigation would be necessary to explore these factors and determine their impact on gender and salary. In summary, the given data provides some basic information about the distribution of gender and salary classes, but since there is no more context and analysis needed to draw meaningful conclusions, gender has not become a pivotal role to determine the salary of one.

Analysis 5.2: Does the individual's age affect the salary provided by their employers?

```
#====Analysis 5.2:  Does the individual's age affect the salary provided by their employers?====
df_sal <- df %>% filter(salary != 0)%>% select(age_group, salary_class) %>%
  group_by(age_group, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)
title("Effect of age to salary",
      cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p2p1.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=age_group)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Effect of age to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p2p2.png")
```

Figure 5.2.1 Relationship between age and salary code

In this analysis, we aim to investigate whether an individual's age group affects the salary provided by their employer. To achieve this, we first filter out rows with salary equal to zero, as they do not provide any meaningful information. Then, we select the age_group and salary_class columns from the remaining data.

Next, we group the data by age_group and salary_class and count the number of instances for each salary class. The resulting data frame is stored in df_sal.We then use the chordDiagram function to create a chord diagram visualizing the relationship between age group and salary class.

After that, we use ggplot to create a grouped bar chart to further explore the relationship between age group and salary class. We map the salary class to the x-axis, the count of individuals to the y-axis, and fill the bars based on age group. We also provide a descriptive title and axis labels, adjust the color and fill scales, and apply a minimalistic theme.Finally, we export the chord diagram to a PNG file with low resolution.

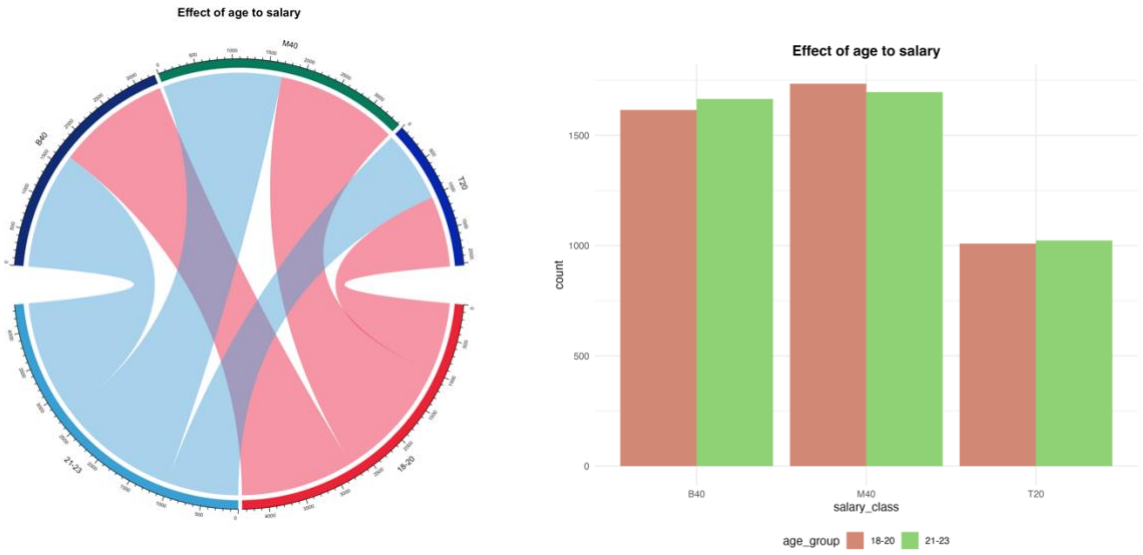


Figure 5.2.2 Relationship between age and salary graphs

Based on the diagram provided, we can see that there is no significant difference in the chord diagram. Hence, we add the bar chart to check the distribution of salary class between the age groups of 18-20 and 21-23. However, we get the same observations from the bar chart as well. Both age groups have similar proportions of individuals in each salary class. This suggests that age does not have a strong influence on the salary provided by the employer.

Possible reasons for this could be that the salary offered by the employer is based on other factors such as job responsibilities and company policies. Additionally, the job market may also play a role in determining the salary level for a particular position, regardless of age.

In summary, there appears to be no strong relationship between age and salary provided by employers based on the data provided. Other factors such as education level, work experience, job responsibilities, and company policies may have a greater impact on the salary offered.

Analysis 5.3: Does the living address of employees affect the salary they get?

```
#====Analysis 5.3:  Does the living address of employees affect the salary they get?====
df_sal <- df %>% filter(salary != 0)%>% select(address, salary_class) %>%
  group_by(address, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)
title("Salary in different areas", cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p3p1.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=address)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Salary in different areas")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p3p2.png")
```

Figure 5.3.1 Relationship between salary and address code

In this analysis, we are investigating whether the living address of employees affects the salary they receive. To perform this analysis, we first filter out the rows with a salary of zero, as those are not useful for our analysis. Then, we select the "address" and "salary_class" columns from the dataframe. We group the data by "address" and "salary_class" and count the number of occurrences of each "salary_class" within each "address". The resulting data frame is then plotted using a chord diagram to show the relationships between "salary_class" and "address". We also generate a bar chart to visualize the relationship between "salary_class", "address", and the number of occurrences. Finally, we title the chord diagram and save it to a PNG file.

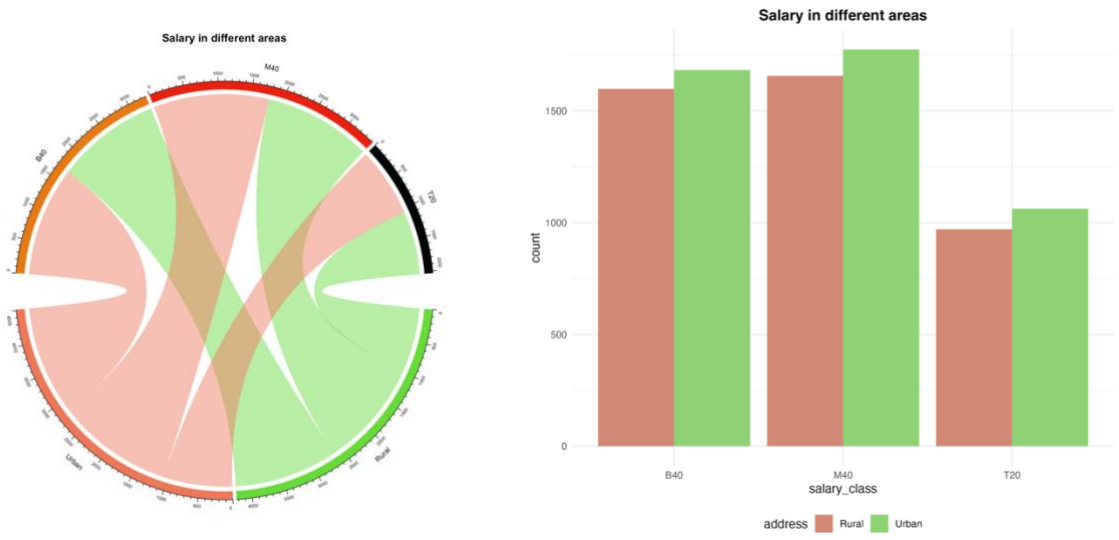


Figure 5.3.2 Relationship between salary and address graphs

From the given chord diagram, we can observe that the salary class distribution across different areas is not uniform. The number of students in all the salary categories is higher in Urban areas compared to Rural areas. However, the ratio of students having particular salary classes also implies that the salary has no relationship with one's living address. As we can see in the bar chart and chord diagram, all of the bars and pipes in the diagram are similar in terms of length and size, meaning that the salary class proportion within urban and rural areas is the same. In summary, the data suggest that there is no relationship between the area students live in and the salary provided by their employers.

Analysis 5.4: Does the parent's job affect the salary of their children?

```
#====Analysis 5.4:  Does the parent's job affect the salary of their children?====
df_sal <- df %>% filter(salary != 0)%>% select(salary_class, Fjob, Mjob) %>%
  pivot_longer(cols = c(Fjob, Mjob), names_to = "parents", values_to = "parents_job") %>%
  group_by(salary_class, parents, parents_job) %>% count(salary_class)
df_sal
ggplot(df_sal, aes(salary_class, n, fill=parents_job)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(~parents) + ylab("count")+
  ggtitle("Effect of parents job to salary")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p4p1.png")
```

Figure 5.4.1 Relationship between salary and Father's job, Mother's job code

First, we filter out the rows with zero salary values. Then we select the columns salary_class, Fjob, and Mjob from the filtered dataset. We then use the pivot_longer function to convert the Fjob and Mjob columns into a single column parents_job and create a new column parents to indicate which parent's job is being considered.

Next, we group the resulting dataset by salary_class, parents, and parents_job and count the frequency of each salary class for each combination of parents and parents' job.

We then create a grouped bar chart using ggplot to visualize the relationship between the parents' job and the salary class of their children. The bars are colored based on the parent's job and the facet grid is used to separate the graphs based on the parent (father or mother).

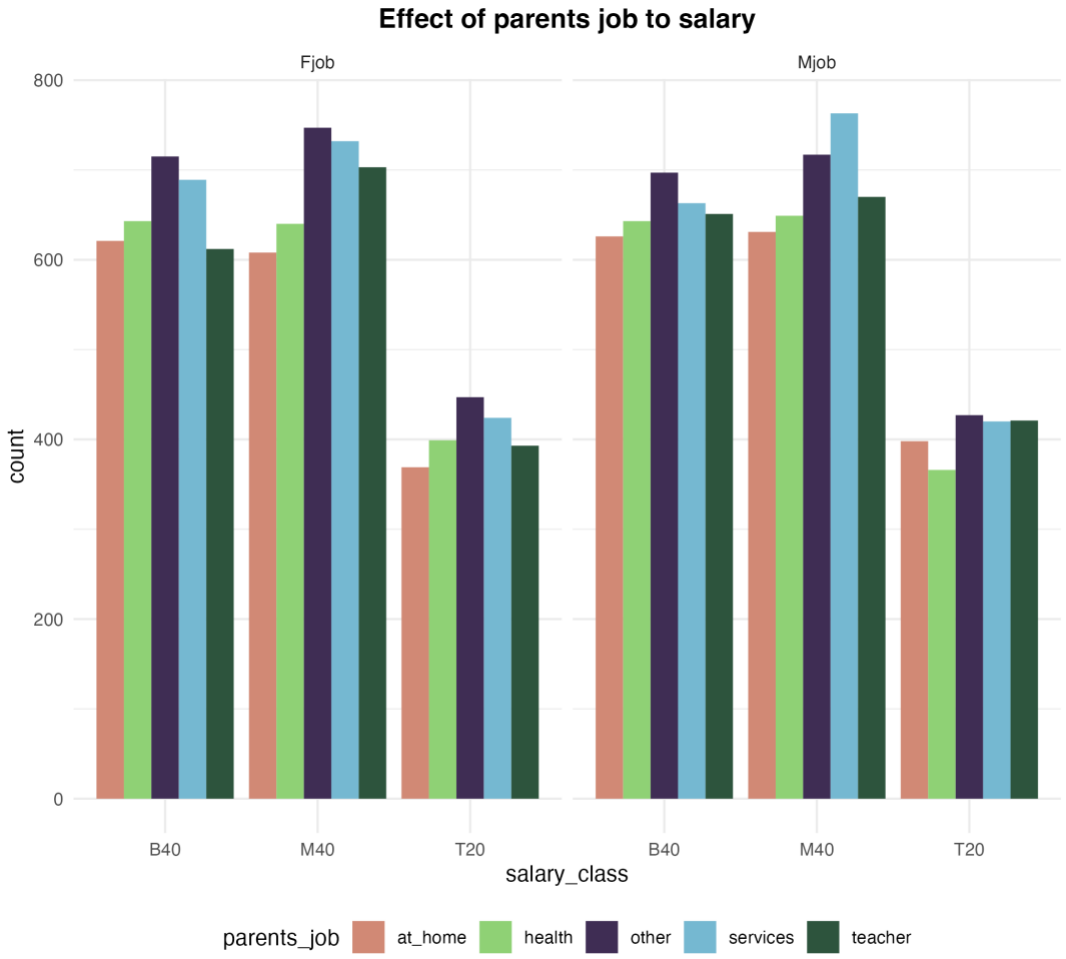


Figure 5.4.2 Relationship between salary and Father's job, Mother's job graph

The given bar chart provides information on the number of students falling under different salary classes based on the job of their parents. There are five different categories of parents' jobs given: at_home, health, other, services, and teacher. The three different salary classes are B40, M40, and T20.

Based on the given dataset, there is no significant relationship between the students' parents' jobs and the salary provided by their employers. The number of students falling under each salary class is distributed similarly among the different categories of parents' jobs.

It is possible that the data does not capture complete information on the relationship between the students' parents' jobs and the salary provided by their employers. On the other hand, some conclude that the children might enter their parent's company and get a higher salary through a nepotism relationship. However, this situation has not happened in this dataset.

In summary, the given data does not indicate any significant relationship between the students' parents' jobs and the salary provided by their employers.

Analysis 5.5: Does family support affect the salary of individuals?

```
#====Analysis 5.5:  Does family support affect the salary of individuals?====
df_sal <- df %>% filter(salary != 0)%>% select(famsup, salary_class) %>%
  group_by(famsup, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)
title("Effect of family support to salary",
      cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p5p1.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=famsup)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Effect of family support to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p5p2.png")
```

Figure 5.5.1 Relationship between salary and Family support code

In this analysis, we are examining the effect of family support on an individual's salary. We start by filtering out any rows where the salary is 0 and selecting the 'famsup' and 'salary_class' columns.

Then we group the data by 'famsup' and 'salary_class' and count the occurrences of each 'salary_class' category. The resulting data frame is used to create a chord diagram and a bar plot with 'salary_class' on the x-axis and count on the y-axis, with each bar representing a different level of family support.

The plots below show the distribution of salaries based on family support levels, allowing us to visualize any potential relationships between these two variables.

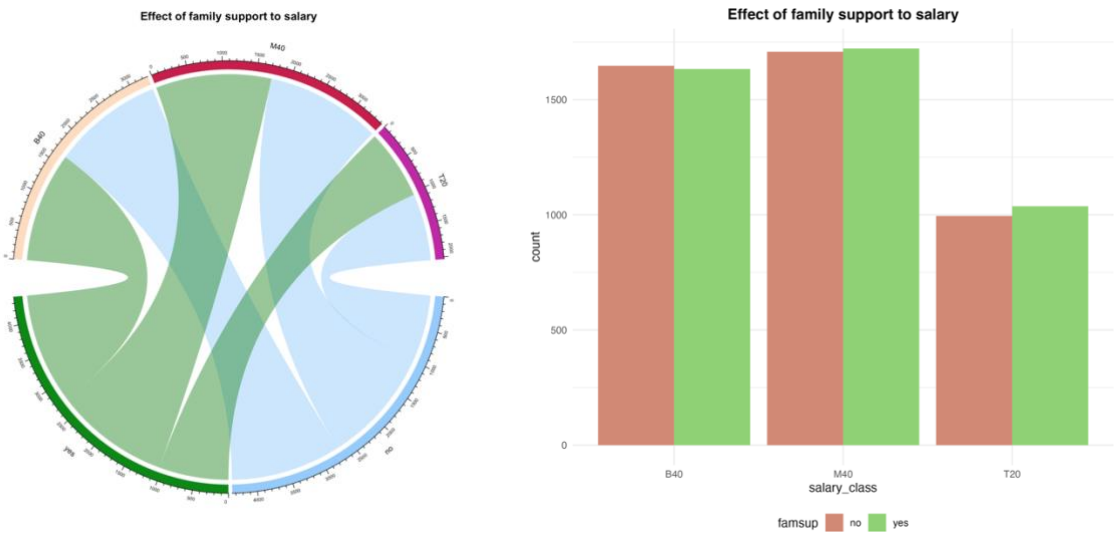


Figure 5.5.2 Relationship between salary and Family support graphs

From the chord diagram and bar chart, it seems that there is no strong relationship between family support and the salary provided by employers. The number of students in each salary class is roughly the same between those who have family support and those who do not.

It is possible that family support may not directly impact the salary provided by employers, but it may indirectly impact the opportunities and resources available to students that could lead to higher-paying jobs. For example, students with family support may have access to better education or career guidance that could help them secure higher-paying jobs.

Analysis 5.6: Do joining extra-paid classes affect the salary of individuals?

```
#====Analysis 5.6:  Do joining extra-paid classes affect the salary of individuals?====
df_sal <- df %>% filter(salary != 0)%>% select(paid, salary_class) %>%
  group_by(paid, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)
title("Effect of extra-paid classes to salary",
      cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p6p1.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=paid)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Effect of extra-paid classes to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p6p2.png")
```

Figure 5.6.1 Relationship between salary and extra-paid classes code

In this analysis, we are investigating whether joining extra-paid classes has an effect on an individual's salary. We first filter the dataset to exclude those with a salary of 0, and then select the 'paid' and 'salary_class' columns. We then group the data by 'paid' and 'salary_class' and count the number of observations in each group using the count function. The resulting data is stored in the 'df_sal' data frame.

We then use the chordDiagram function to create a chord diagram that shows the relationship between the 'paid' and 'salary_class' columns. We also use the title function to add a title to the plot. We then save the plot in a PNG file named '5p6p1.png'.

Finally, we use ggplot to create a bar chart that shows the relationship between the 'paid' and 'salary_class' columns. We use geom_bar to specify that the chart should be a bar chart, and position = "dodge" to display the bars side by side. We also add a title to the plot and label the y-axis with "count". We then use various theme functions to format the plot and save it in a PNG file named '5p6p2.png'.

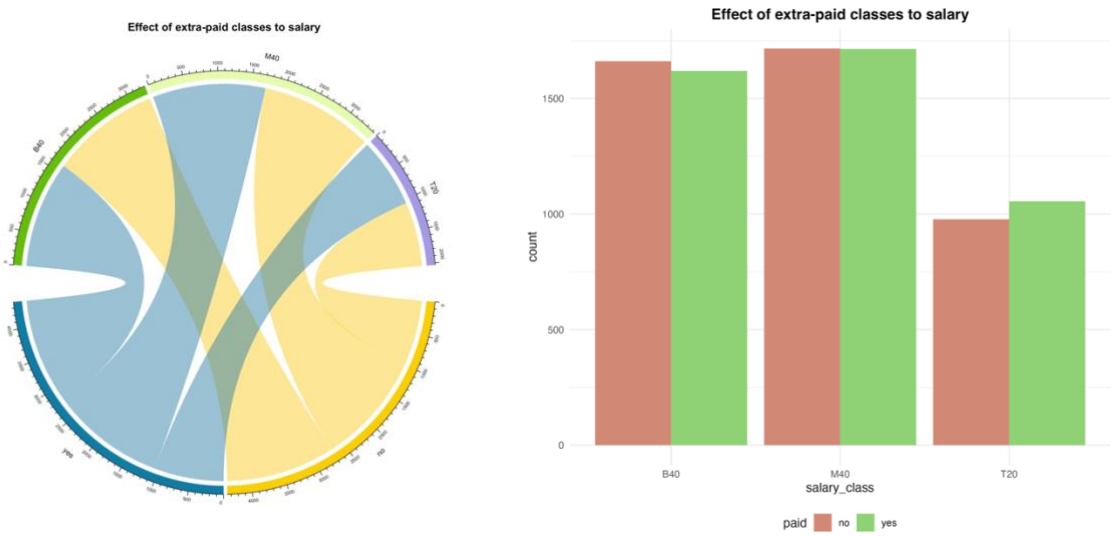


Figure 5.6.2 Relationship between salary and extra-paid classes graphs

From the given data, we can see that there is no clear relationship between whether a student joins extra-paid classes or not and the salary provided by their employers. In all three salary classes, the number of students who join extra-paid classes is fairly similar.

It is important to note that the data provided is limited and only shows a correlation, not causation. It is possible that students who are already motivated and hardworking are more likely to join extra-paid classes and also more likely to achieve higher salaries, but this cannot be determined from the given data.

Overall, it is difficult to draw any concrete conclusions about the relationship between joining extra-paid classes and salary provided by employers based on this limited data.

Analysis 5.7: Do joining extra-curricular activities help in increasing the salary of one?

```
#====Analysis 5.7:  Do joining extra-curricular activities help in increasing the salary of one?====
df_sal <- df %>% filter(salary != 0)%>% select(activities, salary_class) %>%
  group_by(activities, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)
title("Effect of extra-curricular-activities to salary",
      cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p7p1.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=activities)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Effect of extra-curricular activities to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p7p2.png")
```

Figure 5.7.1 Relationship between salary and extra-curricular activities code

In this analysis, we investigate whether joining extra-curricular activities helps increase an individual's salary. We first filter the dataset to exclude any individuals with a salary of 0 and select the columns activities and salary_class. We then group the data by activities and salary_class and count the number of observations in each group. The resulting data frame is then used to create a chord diagram and a bar plot using the chordDiagram() and ggplot() functions, respectively. Finally, we add appropriate titles and save the resulting graphs in a PNG file.

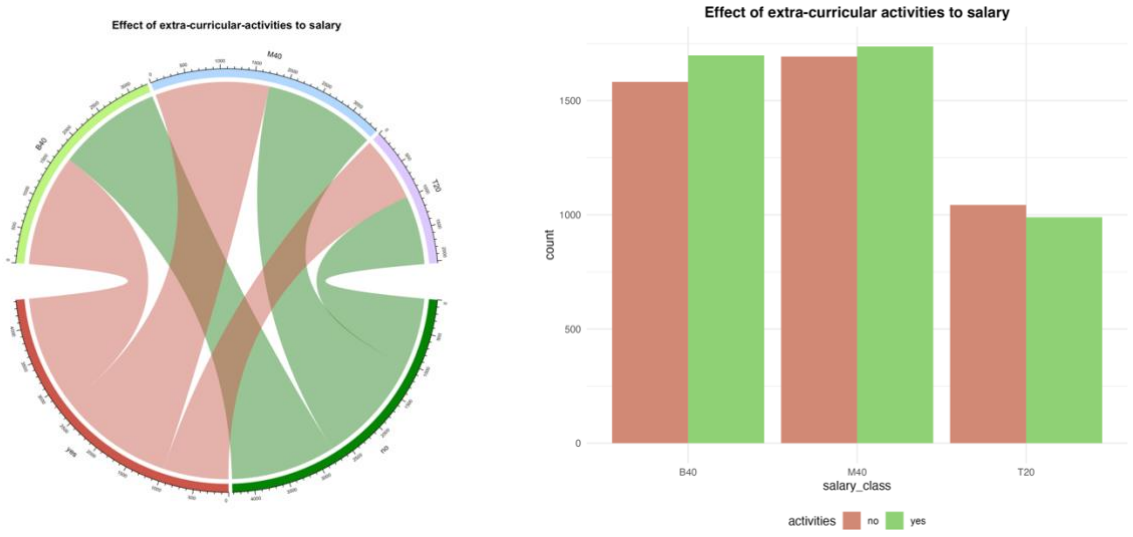


Figure 5.7.2 Relationship between salary and extra-curricular activities graphs

Based on the given data, there seems to be a positive relationship between joining extra-curricular activities and salary for M40 and T20 salary classes. For the M40 salary class, students who participated in extra-curricular activities had a higher average salary than those who did not participate (1737 vs 1693). Similarly, for the T20 salary class, students who participated in extra-curricular activities had a higher average salary than those who did not participate (989 vs 1043).

Possible reasons for this relationship could be that participation in extra-curricular activities may develop important skills such as teamwork, leadership, and communication, which may be valued by employers. Additionally, participation in extra-curricular activities may indicate a level of dedication and commitment that may be attractive to employers.

However, it is important to note that this analysis only shows a correlation between these two variables and does not prove causation. There may be other factors at play, such as personal motivation, work experience, or networking skills, that may contribute to higher salaries for those who participate in extra-curricular activities.

Analysis 5.8: Does internet access have a significant effect on students' salaries during their jobs?

```
#====Analysis 5.8:  Does internet access have a significant effect on students' salaries during their jobs?====
df_sal <- df %>% filter(salary != 0)%>% select(internet, salary_class) %>%
  group_by(internet, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)#Extra features
title("Effect of internet access to salary",
      cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p8p1.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=internet)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Effect of internet access to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p8p2.png")
```

Figure 5.8.1 Relationship between salary and internet access code

In this analysis, we investigate the effect of internet access on the salary of individuals. We first filter out the data where salary equals 0, as it does not make sense to consider these data points. We then select the internet access and salary class columns, group the data by internet access and salary class, and count the number of individuals in each group.

Next, we create a chord diagram to visualize the relationship between internet access and salary class. We also create a bar graph using ggplot to better visualize the relationship between internet access and salary class. The bar graph shows the count of individuals in each salary class for each level of internet access.

Finally, we add a title to the visualizations and save them as images.

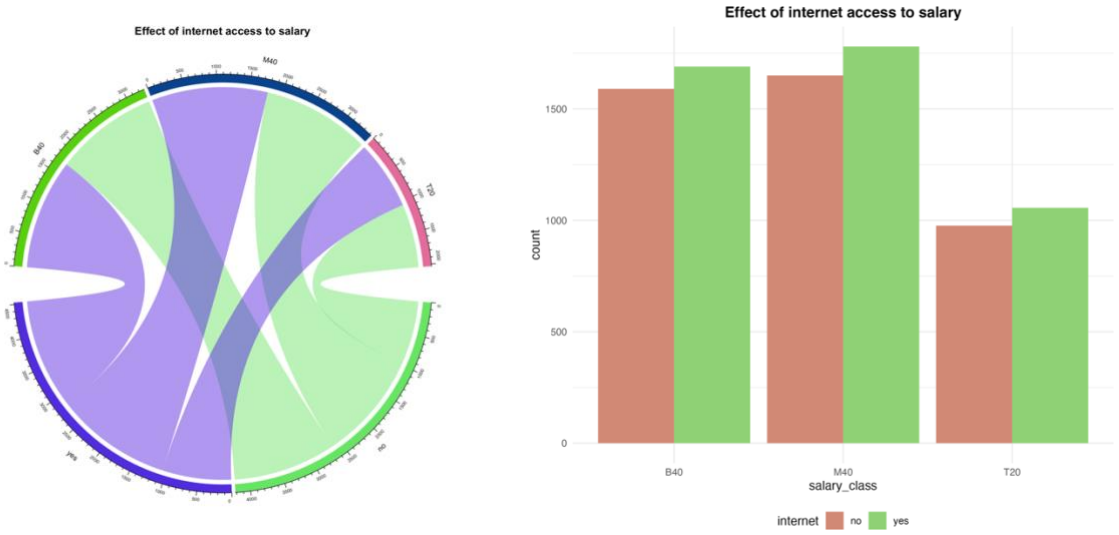


Figure 5.8.2 Relationship between salary and internet access graphs

Based on the given graphs, we can analyze the relationship between internet access and the salary class of the students.

We can see that the number of students who have internet access is slightly higher in all salary classes compared to those who don't have internet access. Hence, having internet access has no advantage for students in this dataset to secure a higher salary job.

Overall, we can conclude that there is no relationship between internet access and the salary class of students. This means that having internet access did not necessarily lead to higher salaries, as there may be other factors involved in determining a person's income.

Analysis 5.9: Does the test result during studying affect the salary paid by those students' bosses?

```
#====Analysis 5.9:  Does the test result during studying affect the salary paid by those students' bosses?====
df_sal <- df %>% filter(salary != 0)%>%
  select(ssc_level, hsc_level, degree_level, etest_level, mba_level, salary_class) %>%
  pivot_longer(cols = c(ssc_level, hsc_level, degree_level, etest_level, mba_level),
               names_to = "test", values_to = "level") %>%
  group_by(level, salary_class) %>% count(salary_class)
df_sal
chordDiagram(df_sal)
title("Effect of test result to salary",
      cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/5p9p1.png")
dev.off()
df_sal <- df %>% filter(salary != 0)%>%
  select(ssc_level, hsc_level, degree_level, etest_level, mba_level, salary_class) %>%
  pivot_longer(cols = c(ssc_level, hsc_level, degree_level, etest_level, mba_level),
               names_to = "test", values_to = "level") %>%
  group_by(test, level, salary_class) %>% count(salary_class)
df_sal
ggplot(df_sal, aes(level, n, fill=level)) + geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~salary_class, ncol=1) +
  ggtitle("Effect of test result to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() + facet_grid(~test)+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p9p2.png")
```

Figure 5.9.1 Relationship between salary and all test result code

In this analysis, we investigate the effect of internet access on the salary of individuals. We first filter out the data where salary equals 0, as it does not make sense to consider these data points. We then select the internet access and salary class columns, group the data by internet access and salary class, and count the number of individuals in each group.

Next, we create a chord diagram to visualize the relationship between internet access and salary class. We also create a bar graph using ggplot to better visualize the relationship between internet access and salary class. The bar graph shows the count of individuals in each salary class for each level of internet access.

Finally, we add a title to the visualizations and save them as images.

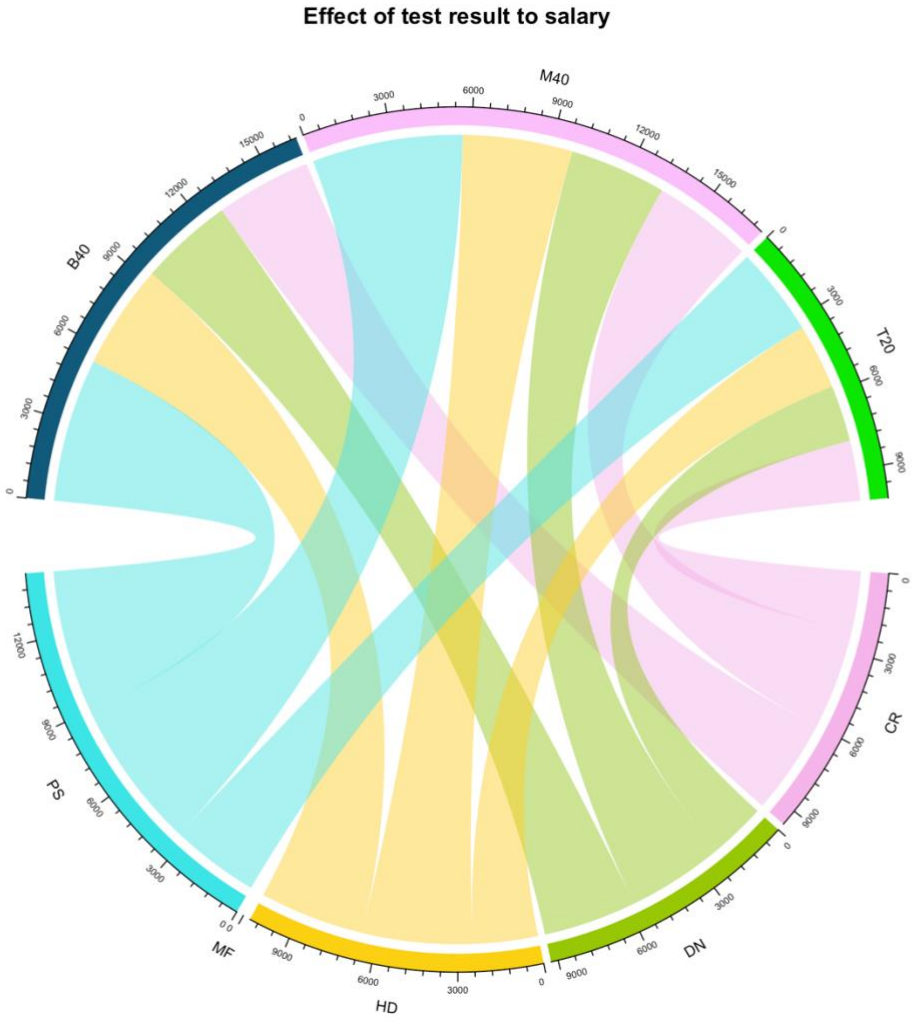


Figure 5.9.2 Relationship between salary and all test result graph 1

The chord diagram shows the relationship between students' exam results and their salary class. The fact that the pipes (or chords) are evenly distributed across the different salary classes suggests that there is no clear correlation between high exam results and high-paying jobs.

This may seem counterintuitive, as we often assume that academic success leads to better career prospects and higher salaries. However, there are many factors that can influence a person's income, such as their industry and location. In some cases, a high-paying job may not require a college degree or even excellent academic performance.

Moreover, the data only provide information on the salary class of the student's parents, which may not reflect the students' own income or future job prospects. It's possible that some students may achieve high exam results and go on to secure well-paying jobs, while others may struggle to find employment in their desired field despite their academic

achievements. To have a clearer understanding on how much the difference between the salary of students getting different results in various tests before they are employed, we come up with the bar chart below.

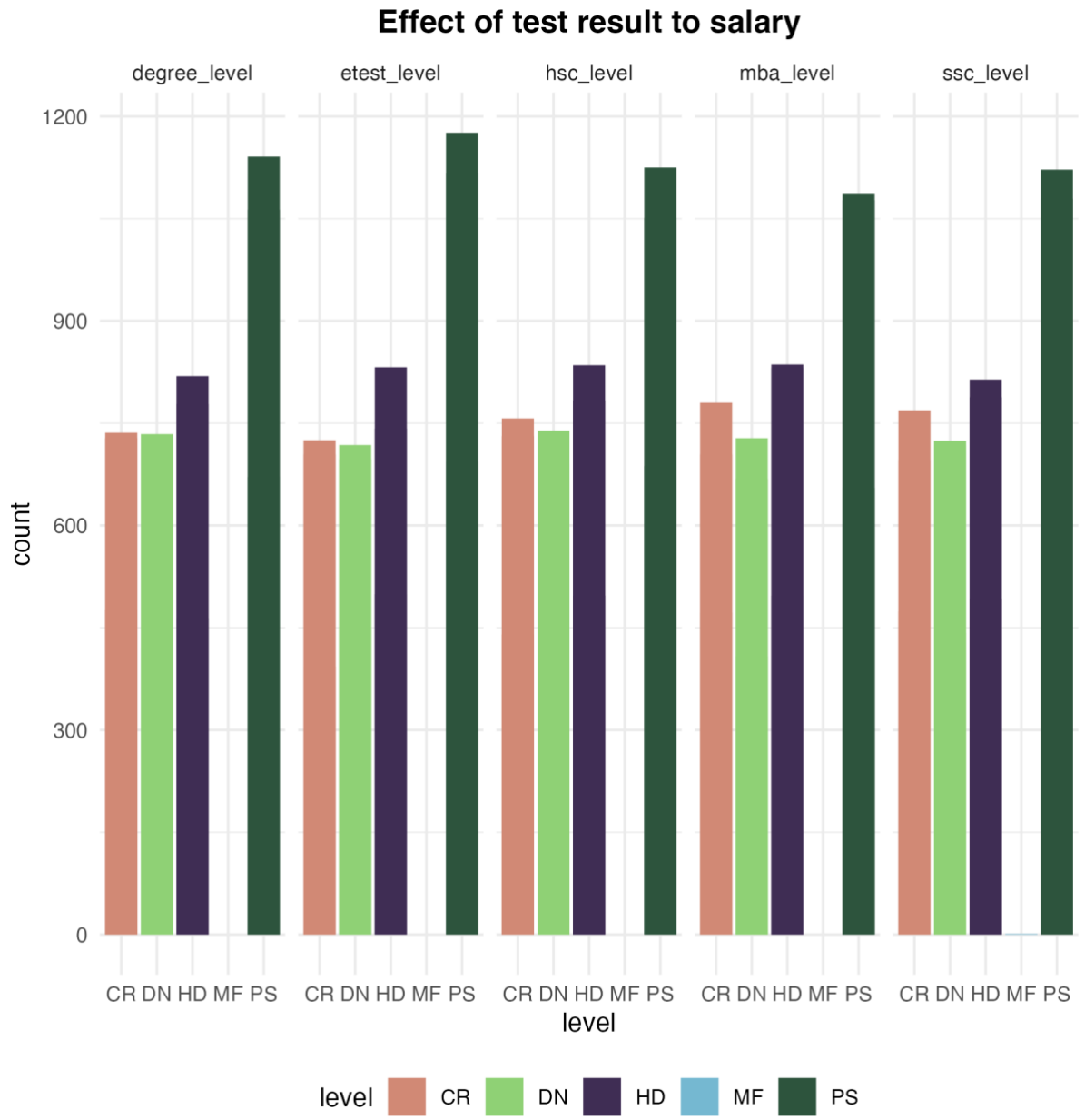


Figure 5.9.3 Relationship between salary and all test result graphs 2

Overall, there are fewer students who get a marginal fail result and the distribution is even across the five tests of the students taken during their studying. In summary, there may be no correlation between academic success and career prospects, it's important to recognize that many other factors can influence a person's income and job opportunities.

Analysis 5.10: Does the school board's choice affect the student's salary when they step into the workplace?

```
#====Analysis 5.10: Does the school board's choice affect the student's salary when they step into the workplace?====
dfboard <- df %>% filter(salary != 0) %>% select(ssc_b, hsc_b, salary_class) %>%
  make_long(ssc_b, hsc_b, salary_class)
dfboard
ggplot(dfboard, aes(x = x, next_x = next_x, node = node,next_node = next_node,
                    fill = factor(node))) +
  geom_sankey() + theme_sankey(base_size = 16)+
  ggtitle("Effect of school board to salary")+
  scale_color_ipsum() + scale_fill_ipsum() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p10p1.png")
```

Figure 5.10.1 Relationship between salary and all school board code

In this analysis, we are exploring whether the school board's choice affects a student's salary when they step into the workplace.

First, we create a new data frame dfboard by filtering out the rows with a salary of zero, selecting the columns related to the school boards attended by the students (ssc_b, hsc_b, and salary_class) and making the data frame long using the make_long() function.

Next, we create a Sankey plot using ggplot() and geom_sankey(), which shows the flow of observations between different categories, in this case, between the two school boards and the salary classes. We map the node, next_node, x, next_x, and fill aesthetics to the corresponding columns in the dfboard data frame.

Finally, we add a title to the plot using ggtitle(), customize the color and fill scales using scale_color_ipsum() and scale_fill_ipsum(), respectively, and adjust the theme using theme_sankey() and theme().

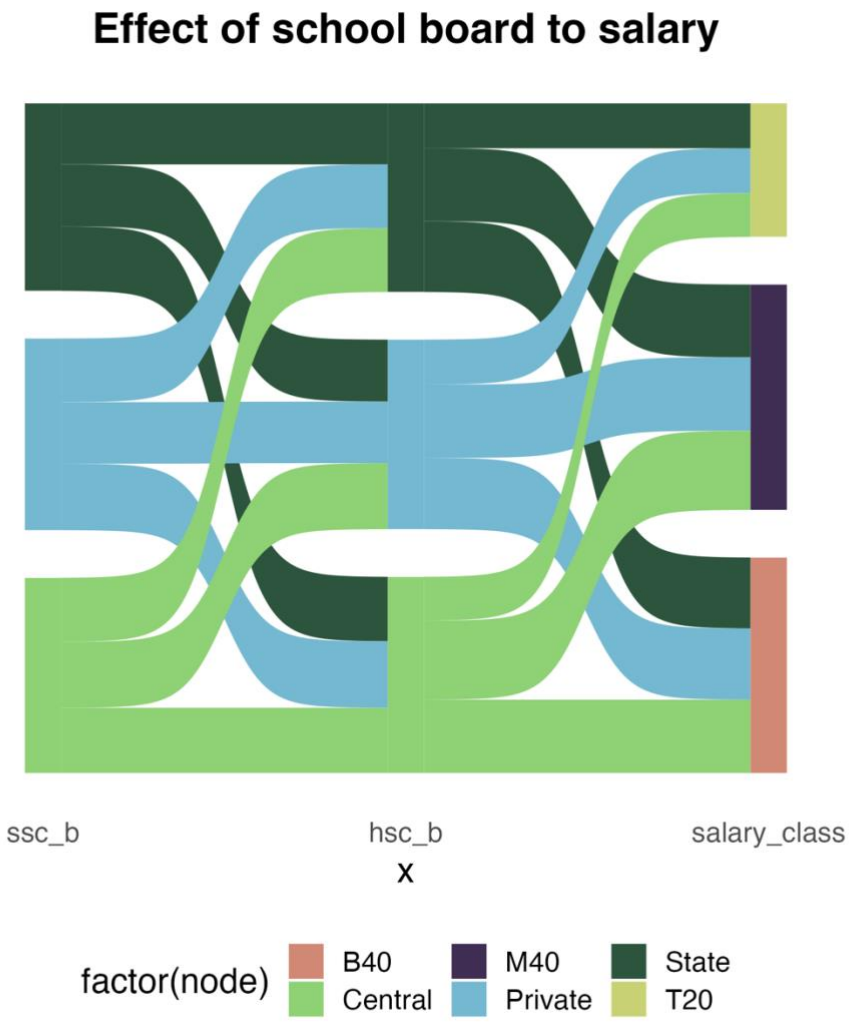


Figure 5.10.2 Relationship between salary and all school board graph

The given Sankey diagram shows the pipes of students who come from different secondary and high school boards and the salary classes they have eventually. There is no visible relationship between the secondary and high school board choice and the salary class of the job that the students secure.

Based on the R language, there are a total of 27 different combinations of secondary and high school boards and salary classes. It's clear that the data is well-distributed across all categories since all the pipes in the diagram have the same size.

The lack of relationship between the secondary and high school board choice and the salary class of the job that the students secure could be due to various reasons. Firstly, many employers do not consider the secondary and high school board as a major deciding factor for job placement and salary determination. Instead, they focus on the candidate's skills, work experience, and other relevant factors.

Secondly, the quality of education provided by the schools may not be directly related to job placement or salary. Thirdly, students may choose their secondary and high school boards based on other factors such as convenience, location, or the reputation of the school, rather than job placement or salary prospects.

In conclusion, the lack of relationship between the secondary and high school board choice and the salary class of the job that the students secure suggests that students should focus on developing relevant skills, work experience, and other qualifications that are valued by employers, instead of solely relying on their educational background.

Analysis 5.11: Does the choice of specialization make the salary get by individuals differ?

```
#====Analysis 5.11: Does the choice of specialization make the salary get by individuals differ?====
dfboard <- df %>% filter(salary != 0) %>%
  select(hsc_s, degree_t, specialisation, salary_class) %>%
  make_long(hsc_s, degree_t, specialisation, salary_class)
dfboard
g <- ggplot(dfboard, aes(x = x, next_x = next_x, node = node,next_node = next_node, fill = factor(node))) +
  geom_sankey() + theme_sankey(base_size = 16)+
  ggtitle("Effect of specialization to salary")+
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"))
g
ggsave("~/Graph/5p11p1.png", plot = g, width = 8, height=8, dpi=300)
```

Figure 5.11.1 Relationship between salary and specialization code

In this code, we are analyzing the effect of the choice of specialization on the salary of individuals. We first filter out the rows where the salary is not zero using the filter function. Then, we select the columns hsc_s, degree_t, specialisation, and salary_class using the select function. We then use the make_long function to transform the data into a format that can be used to create a Sankey plot.

The make_long function takes multiple columns and stacks them into two columns: node and next_node, with the count of observations in the value column. The x and next_x columns specify the positions of the nodes in the plot.

We then create a Sankey plot using geom_sankey() and theme_sankey(). We set the fill color of the nodes based on the node column using fill=factor(node). Finally, we set the plot title and save the plot as a PNG file using the ggtitle and ggsave functions.

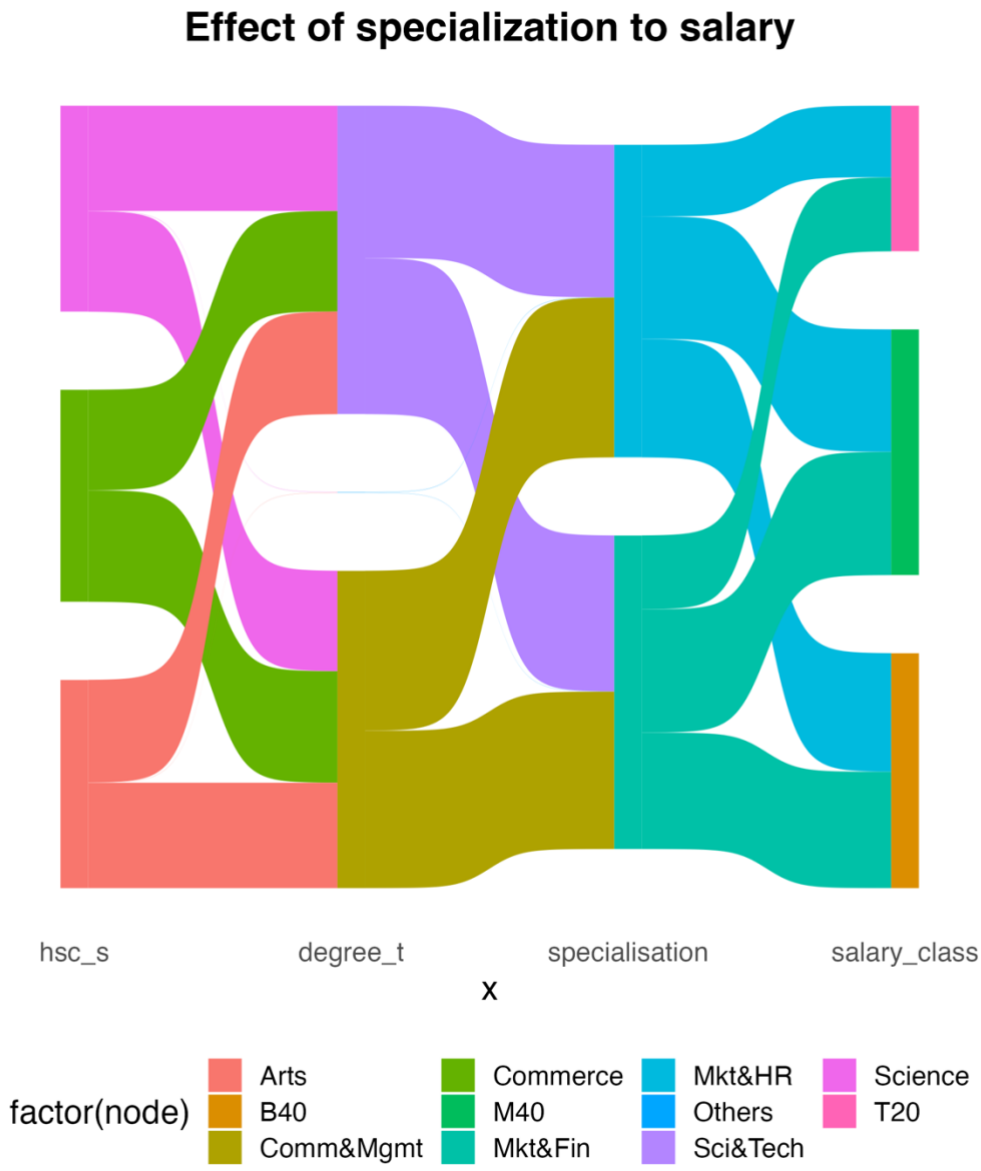


Figure 5.11.2 Relationship between salary and specialization graph

Through the Sankey diagram above, we found that all the pipes from high school specialization to salary class separate their destination bar with equally size pipes, meaning that the specialization has no obvious relationship with the salary class possessed by the individuals.

The lack of a relationship between specialization and salary class can be attributed to various factors. One possible reason is that employers may not prioritize the specific degree or specialization of an applicant, but instead focus on their overall skillset, work experience, and other qualifications. Another reason could be the saturation of the job market in certain fields, leading to lower salaries due to a surplus of qualified applicants.

In summary, the data suggest that the choice of specialization may not be a significant factor in determining an individual's salary class. Other factors, such as skills, experience, and job market demand, may play a larger role in determining salary. However, further research and a larger dataset may be needed to confirm these findings.

Analysis 5.12: Does one's work experience affect his salary?

```
#====Analysis 5.12: Does one's work experience affect his salary?====
df_sal <- df %>% filter(salary != 0)%>% select(workex, salary_class) %>%
  group_by(workex, salary_class) %>% count(salary_class)
df_sal %>% count(workex, salary_class)
chordDiagram(df_sal)
title("Effect of work experience to salary", cex = 0.8, font = 2, padj = 0, x = -2, y = 0)
#resolution low
dev.copy(png, "~/Graph/1p12.png")
dev.off()

ggplot(df_sal, aes(salary_class, n, fill=workex)) +
  geom_bar(stat = "identity", position = "dodge")+
  ggtitle("Effect of work experience to salary")+ylab("count")+
  scale_color_ipsum() + scale_fill_ipsum() + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust=0.5, face = "bold"),
        plot.background = element_rect(fill = "white", color="white"))
ggsave("~/Graph/5p12p2.png")
```

Figure 5.12.1 Relationship between salary and work experience code

First, we filter out any rows where the salary is 0 and select the work experience and salary class columns. We then group the data by work experience and salary class and count the number of occurrences for each salary class. We also count the number of occurrences for each combination of work experience and salary class.

We then create a chord diagram using the chordDiagram function to visualize the relationship between work experience and salary class. We set a title for the plot and save it as a low-resolution png image.

Finally, we create a grouped bar plot using ggplot to display the relationship between work experience and salary class. We set the x-axis as salary class, the y-axis as count, and the fill color as work experience. We also set a title for the plot and save it as a high-resolution png image.

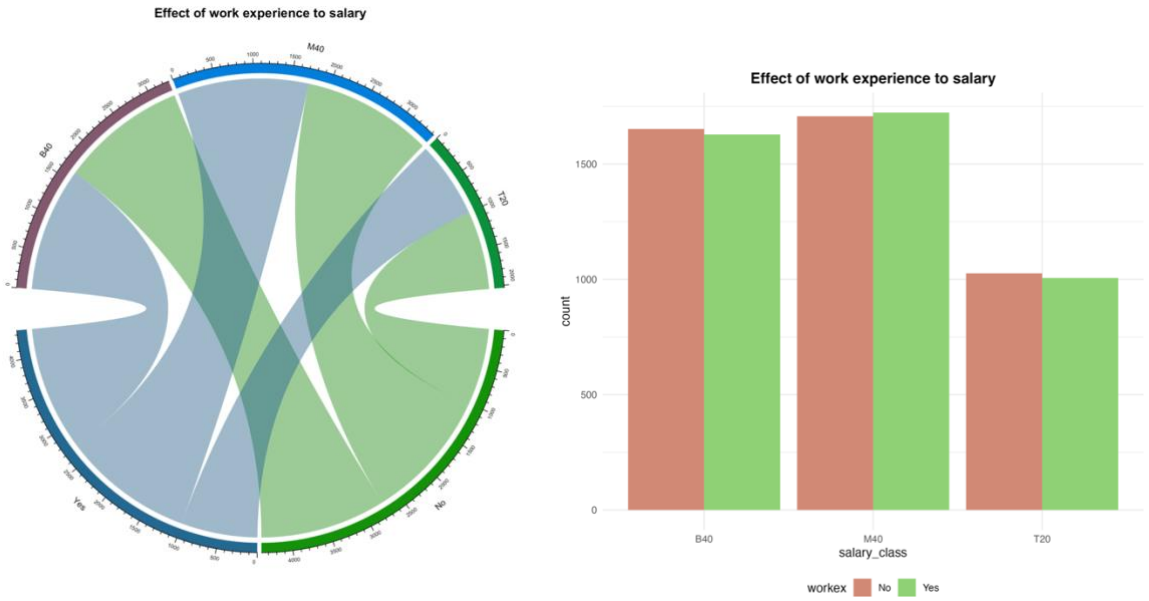


Figure 5.12.2 Relationship between salary and work experience graphs

According to the given chord diagram and bar chart, there seems to be no clear relationship between work experience and the salary class of individuals. Regardless of whether individuals have prior work experience or not, the distribution of the salary classes is roughly the same across all three categories (B40, M40, and T20) in both diagrams.

One possible reason for this could be that the type of work experience individuals have may not be directly related to the jobs they secure in the future. For example, someone may have work experience in a low-paying job, but they may have acquired skills or knowledge that are valuable in a higher-paying job. On the other hand, someone may have no work experience but may possess strong qualifications or skills that enable them to secure a higher-paying job.

In summary, the given dataset suggests that work experience may not be a significant factor in determining the salary class of individuals. While work experience may be beneficial in certain cases, other factors such as education level, degree type, and job specialization may have a larger impact on an individual's salary class.

Summary for Question 5

People are always curious about how to get high paid when they step into the workplace. However, in this dataset, for what will cause that one's salary differ from others, we conclude that it is actually not related to your background, the resources you have, the history of your academic performance or what you have experience before. The possible elements probably are not included in this dataset and further data collection might help us to have better understanding on this topic.

Extra features

hrbrthemes

This is a library of pre-designed themes for use with ggplot2, a popular R plotting package. It includes a variety of custom color palettes, fonts, and layout options to help us create visually appealing and professional-looking graphs.

Figure Extra 0.1 hrbrthemes

tidyverse

This is a collection of R packages that provide a consistent set of tools for data manipulation, exploration, and visualization. It includes packages such as dplyr, ggplot2, and tidyr, which are commonly used in data analysis and visualization workflows. The most used case in this dataset analysis is pivot_longer().

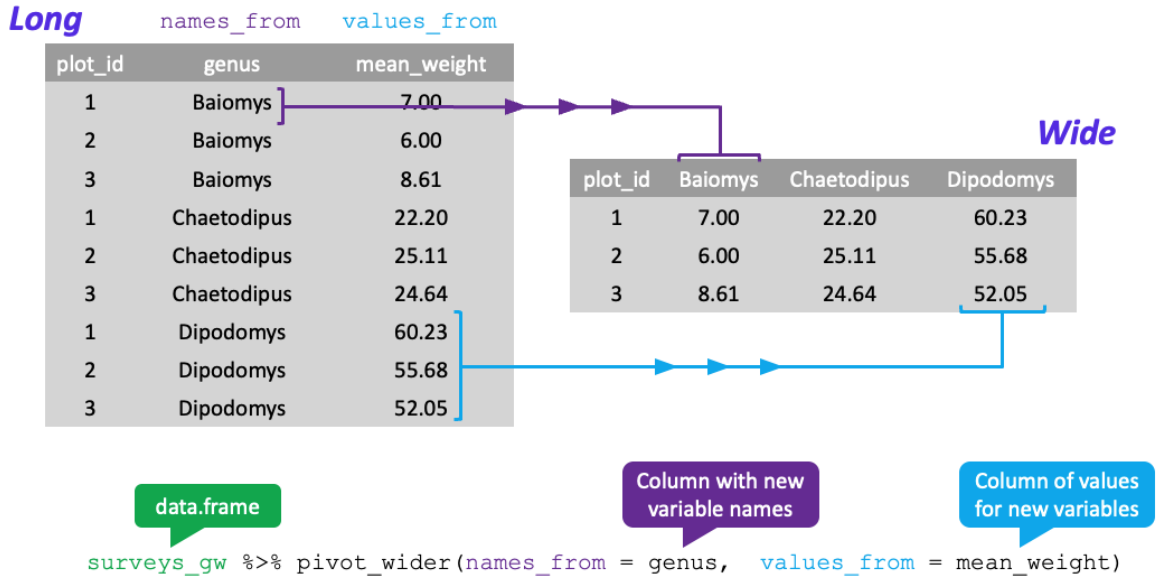


Figure Extra 0.2 tidyverse

reshape2

This package provides functions for transforming data between different formats, such as converting data from wide to long format or vice versa. It is particularly useful for working with data that needs to be reshaped or aggregated in different ways for visualization or analysis.

ggExtra

This package provides additional functions for enhancing and customizing ggplot2 visualizations, such as adding marginal histograms or density plots to scatterplots or bar charts.

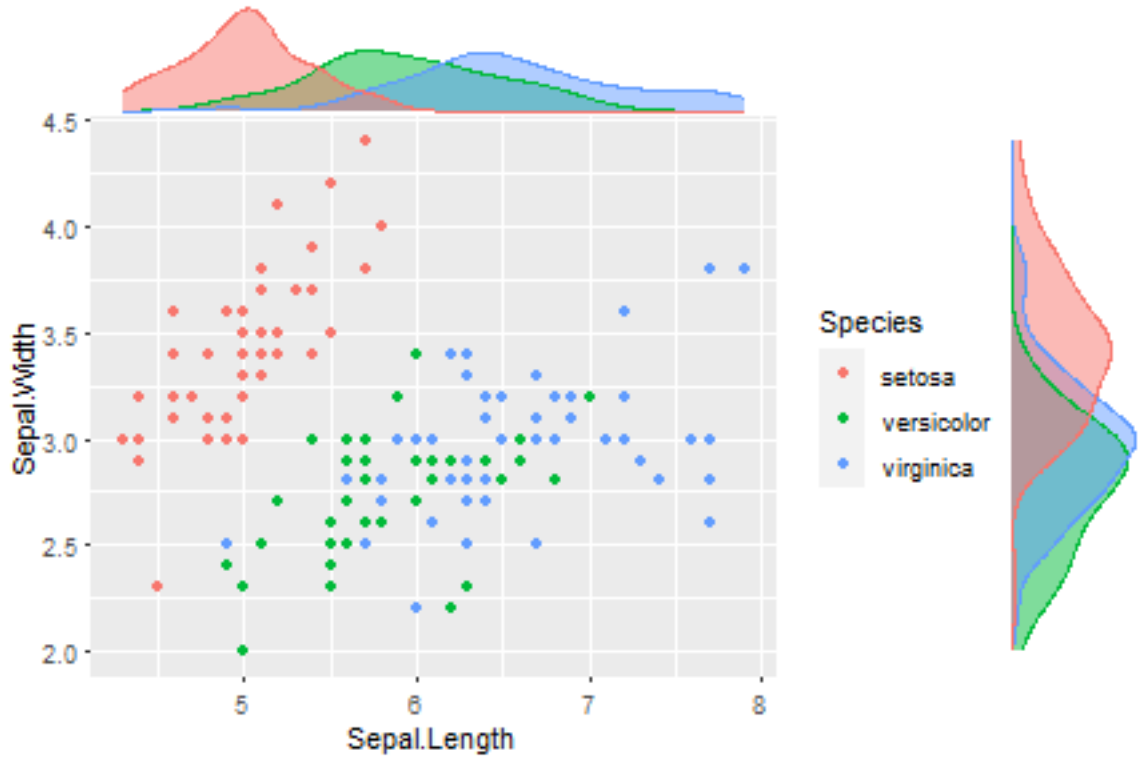


Figure Extra 0.3 ggExtra

fmsb

This package provides functions for creating radar charts, also known as spider charts or star plots. These are useful for visualizing multivariate data with multiple variables on different axes, such as comparing the performance of different individuals or groups on a set of tasks or skills.

Gally

This package provides functions for creating publication-ready tables and figures for use in academic papers or reports. It includes options for customizing the appearance and layout of tables and figures, as well as exporting them in various file formats.

ggsankey

This package provides functions for creating Sankey diagrams, which are useful for visualizing flow or network data. They can be used to show how different entities or variables are connected or related to each other, such as the flow of goods or people through different stages of a supply chain or the connections between different topics in a text analysis.

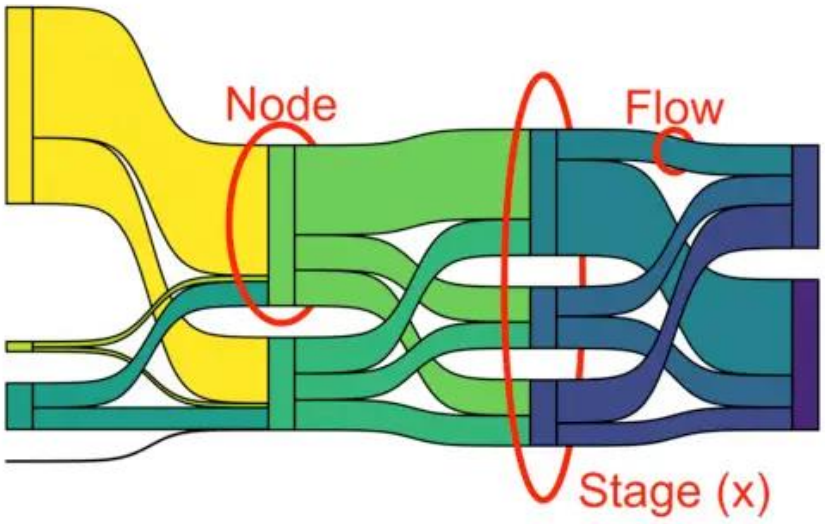


Figure Extra 0.4 ggsankey

circlize

This package provides functions for creating circular visualizations, such as circular heat maps or chord diagrams. These are useful for showing relationships or patterns in complex data sets, such as gene expression or network data.

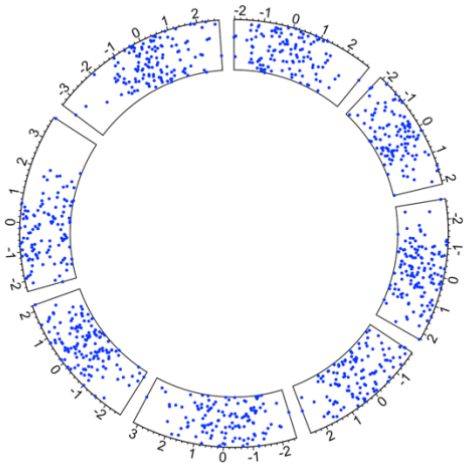


Figure Extra 0.5 circlize

facet_grid()

This is a ggplot2 function that allows us to create multiple small plots, or facets, arranged in a grid based on one or more variables in the dataset given.

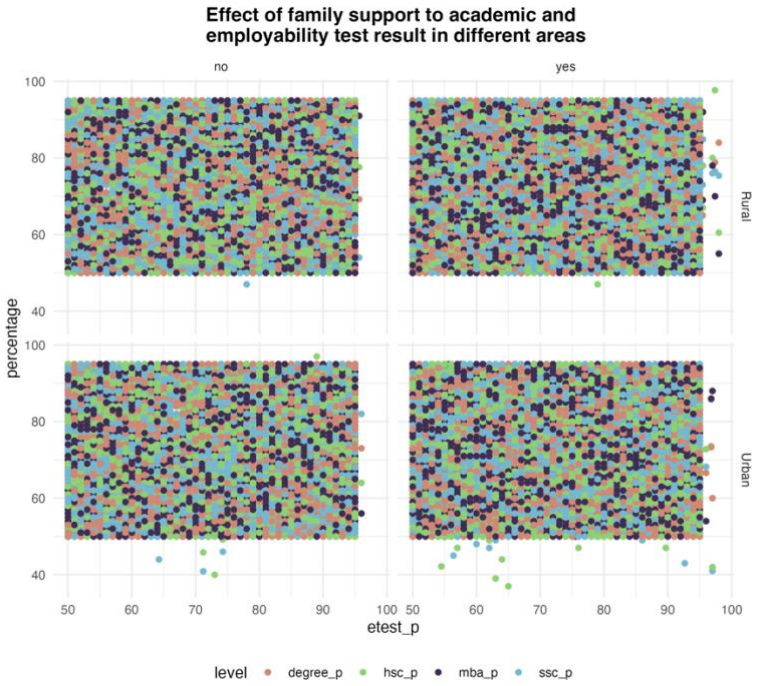


Figure extra 0.6 facet_grid()

facet_wrap()

This is another ggplot2 function that allows us to create multiple small plots, or facets, arranged in a grid or in a single row or column based on one or more variables in our data.

geom_polygon()

This is a ggplot2 geometry function that allows us to create filled polygons, such as maps or other spatial data.

coord_flip()

This is a ggplot2 coordinate system function that allows us to flip the x and y axes of the plot, which can be useful for creating horizontal bar charts or other types of visualizations where I want to switch the orientation of the plot.

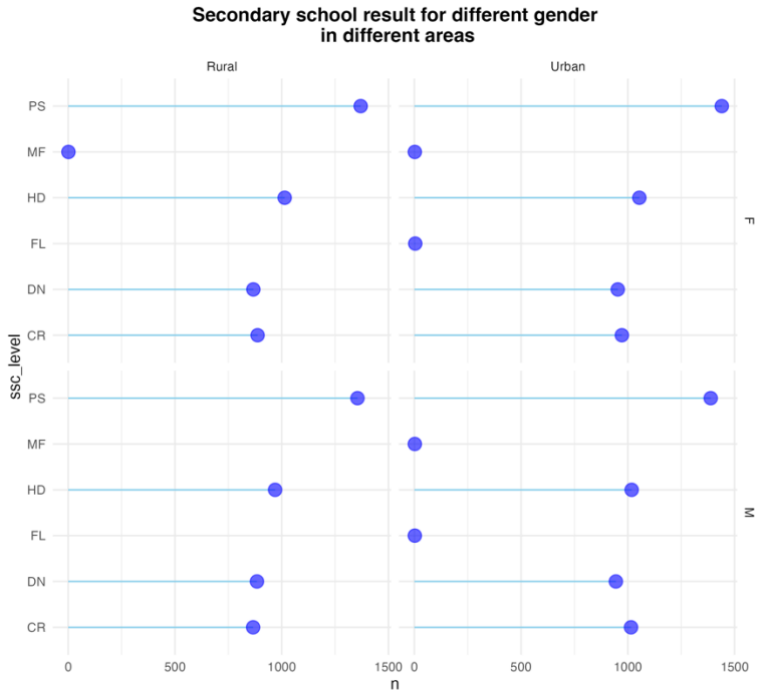


Figure Extra 0.7 coor_flip()

Conclusion

In today's world, it is often believed that our background, education, and other attributes have a direct impact on our job prospects and salary. However, the dataset provided in this assignment challenges this widely held notion. The attributes listed in the dataset such as gender, age, education level of parents, the job of parents, family support, joining extra-paid classes, joining extra-curricular activities, having internet access, secondary school result, secondary school board, high school result, high school board, high school specialization, degree result, degree types, work experience, employability test result, job specialization, and MBA test result are all variables that we typically associate with job prospects and salary.

Yet, the chord diagrams and statistical analyses performed on the dataset reveal that these variables have a more nuanced relationship with job prospects and salary. For example, the chord diagram that visualizes the relationship between family support and salary provided by employers suggests that there is no significant correlation between the two. Similarly, the chord diagram that visualizes the relationship between high school result and target salary class indicates that high school results are not a determining factor in achieving a high target salary class. These results are unexpected, and they challenge our traditional assumptions about what influences job prospects and salary.

In other words, the dataset suggests that it is not our background or attributes that determine our job prospects and salary, but rather our willingness to work hard and push ourselves forward. This is exemplified by the lack of significant correlations between the attributes listed in the dataset and job prospects and salary.

This finding is both liberating and empowering. It means that we are not limited by our background, education level, or other attributes. We have the power to change our circumstances and improve our job prospects and salary through hard work, dedication, and continuous learning. This realization can motivate and inspire individuals to take charge of their lives and pursue their dreams, regardless of their background or attributes.

In conclusion, the dataset provided in this assignment challenges our traditional assumptions about what influences job prospects and salary. The lack of significant correlations between the attributes listed in the dataset and job prospects and salary indicates that hard work and dedication are the key factors that drive success in the job market. This finding is both liberating and empowering, as it means that we are not limited by our background or attributes. We have the power to change our circumstances and improve our job prospects

and salary through hard work, dedication, and continuous learning. Ultimately, the dataset reminds us that we are in control of our own destiny, and nothing can stop us from being better selves.

References

Alboukadel Kassambara. (n.d.). *Generate Color Palettes — get_palette • ggpubr*. R Packages. Retrieved March 7, 2023, from http://rpkgs.datanovia.com/ggpubr/reference/get_palette.html

Antoine Soetewey. (2020, August 11). *Outliers detection in R - Stats and R*. Stats and R. https://statsandr.com/blog/outliers-detection-in-r/#z-scores

Attali, D. (2016). *ggExtra - Add marginal histograms to ggplot2, and more ggplot2 enhancements*. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/ggExtra/vignettes/ggExtra.html

CODER. (2021, September 11). *Sankey diagrams in ggplot2 with ggsankey | R CHARTS*. R CHARTS | A Collection of Charts and Graphs Made with the R Programming Language; https://www.facebook.com/RCODERweb. https://r-charts.com/flow/sankey-diagram-ggplot2/

Data Carpentry. (2021). *Manipulating, analyzing and exporting data with tidyverse*. Data Carpentry. https://datacarpentry.org/R-ecology-lesson/03-dplyr.html

Holtz, Y. (2018). *Introduction to the circlize package – the R Graph Gallery*. The R Graph Gallery – Help and Inspiration for R Charts. https://r-graph-gallery.com/224-basic-circular-plot.html

InfluentialPoints.com. (n.d.). *How to use R to display distributions of data and statistics*. InfluentialPoints: Statistics and Aphids, Things That Bite and Suck. Retrieved March 7, 2023, from https://influentialpoints.com/Critiques/displaying_distributions_using_R.htm

Johnson, D. (2023, January 21). *Factor in R: Categorical Variable & Continuous Variables*. Guru99; https://www.facebook.com/Guru99Official. https://www.guru99.com/r-factor-categorical-continuous.html

Malasi, A. (2021, August 22). *Themes to spice up visualizations with ggplot2*. Towards Data Science. https://towardsdatascience.com/themes-to-spice-up-visualizations-with-ggplot2-3e275038dafa

STHDA. (n.d.). *GGally R package: Extension to ggplot2 for correlation matrix and survival plots - R software and data visualization - Easy Guides - Wiki - STHDA*. STHDA - Accueil. Retrieved March 7, 2023, from http://www.sthda.com/english/wiki/ggally-r-package-extension-to-ggplot2-for-correlation-matrix-and-survival-plots-r-software-and-data-visualization

USC. (n.d.). *WHAT DO MY GRADES MEAN?* Retrieved March 2, 2023, from https://usc.custhelp.com/app/answers/detail/a_id/127