EDA_part2

2024-04-28

Team: Final Project Group 5

Members: Jinyu Wang, Letian Yu, Xiaoyan Liu, Yijia Xue, Ziao Zhang

Author: Letian

Last Update Date: April 28 2024

Introduction

This document contains parts of the EDA code for the DATA2020 final project.

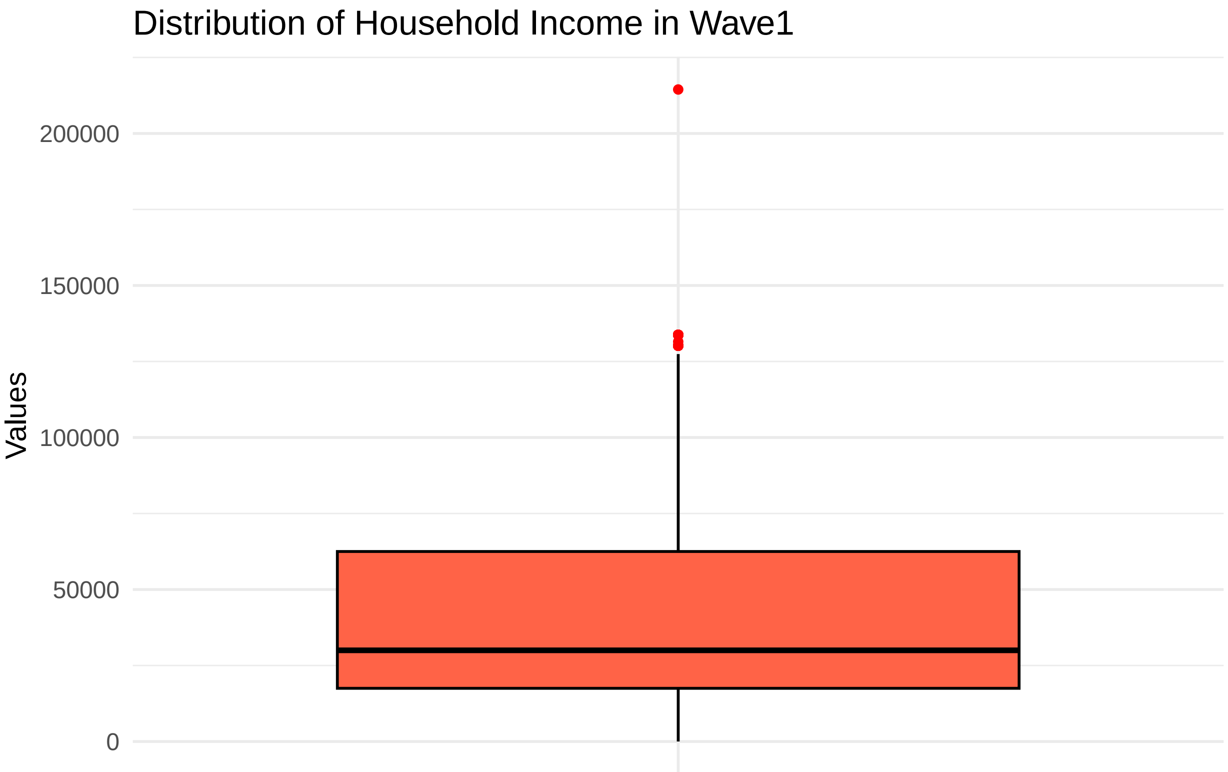
```
library(haven)
library(dplyr)
```

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

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

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

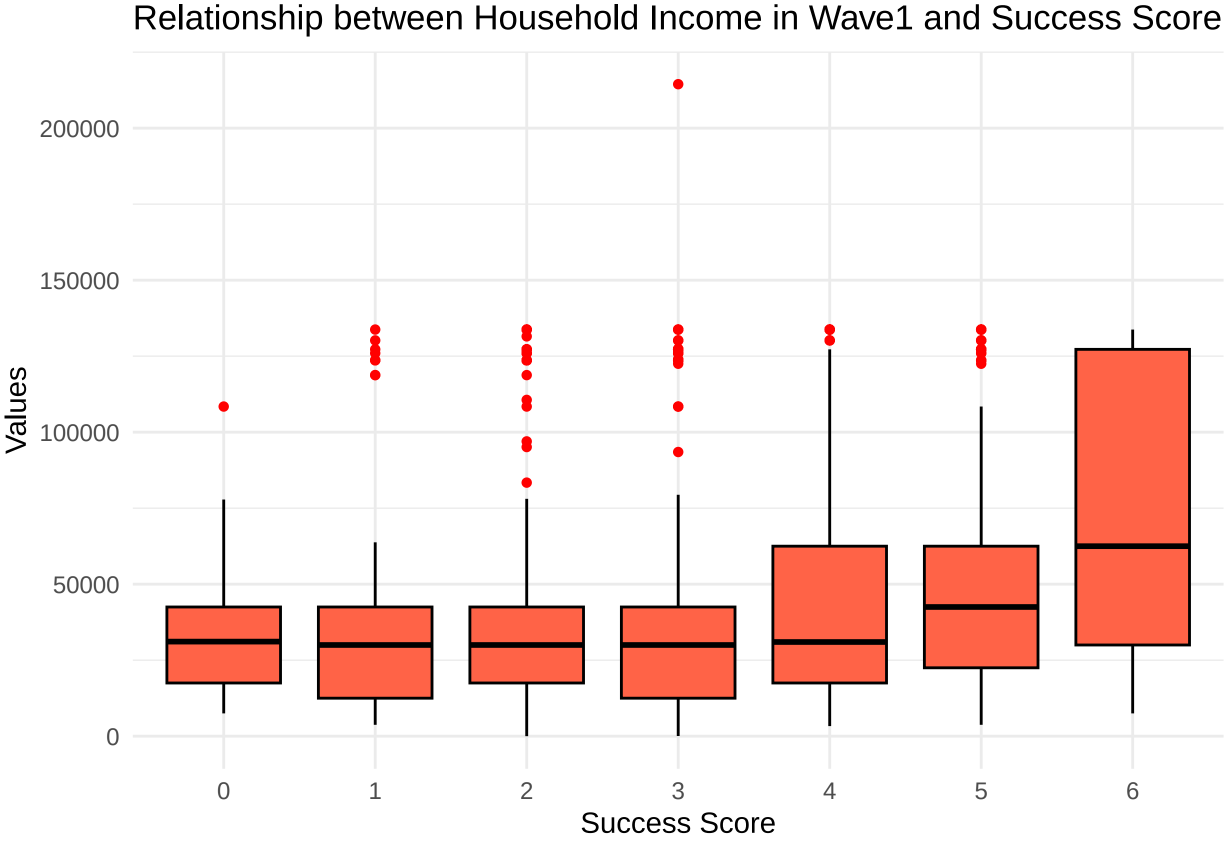
```
library(ggplot2)
df <- read_dta('../data/ff_data_x_preprocessed.dta')
df_explained <- read_dta('../data/ff_data_preprocessed.dta')
```

Continous Variable

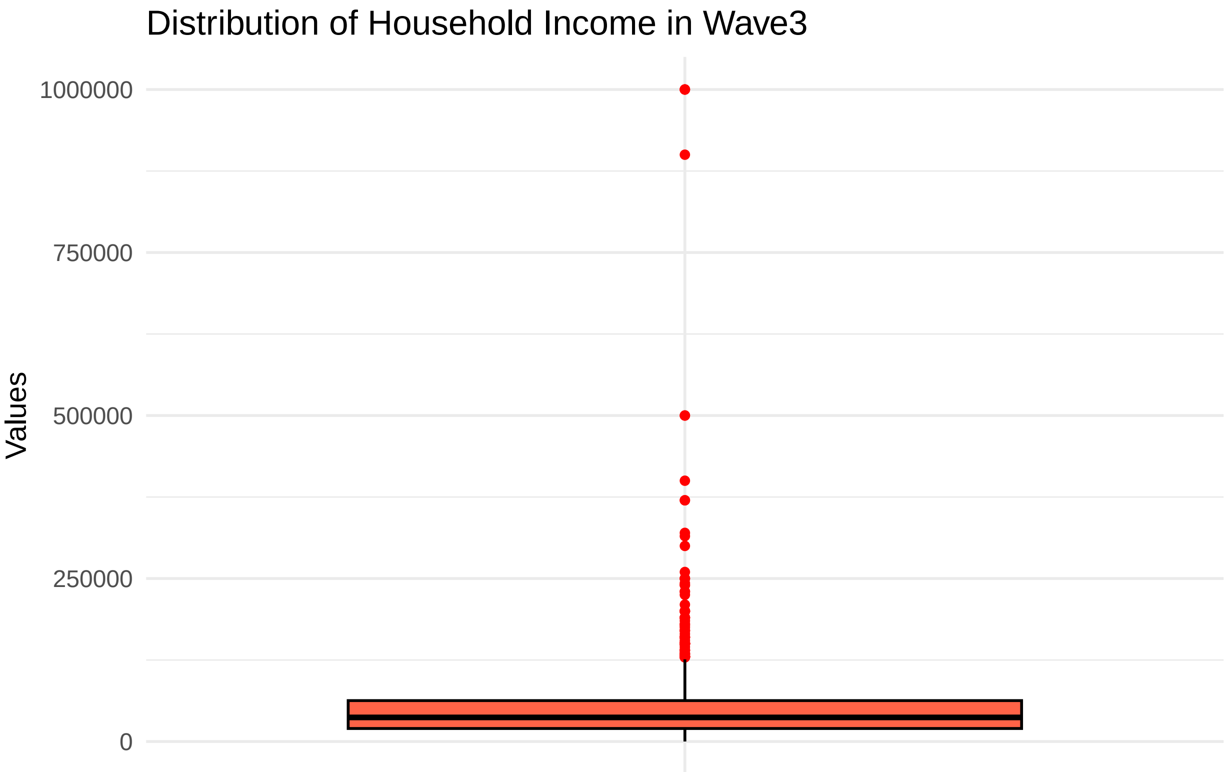
[‘cf1hhinc’, ‘f1j8’, ‘m1i2b’, ‘cf3hhinc’] cf1hhinc: Household income(wave1) f1j8: How much did you earn(father) in wave1? m1i2b: How much did you earn(mother) in wave1? cf3hhinc: Household income(wave3) Boxplot for continous variable

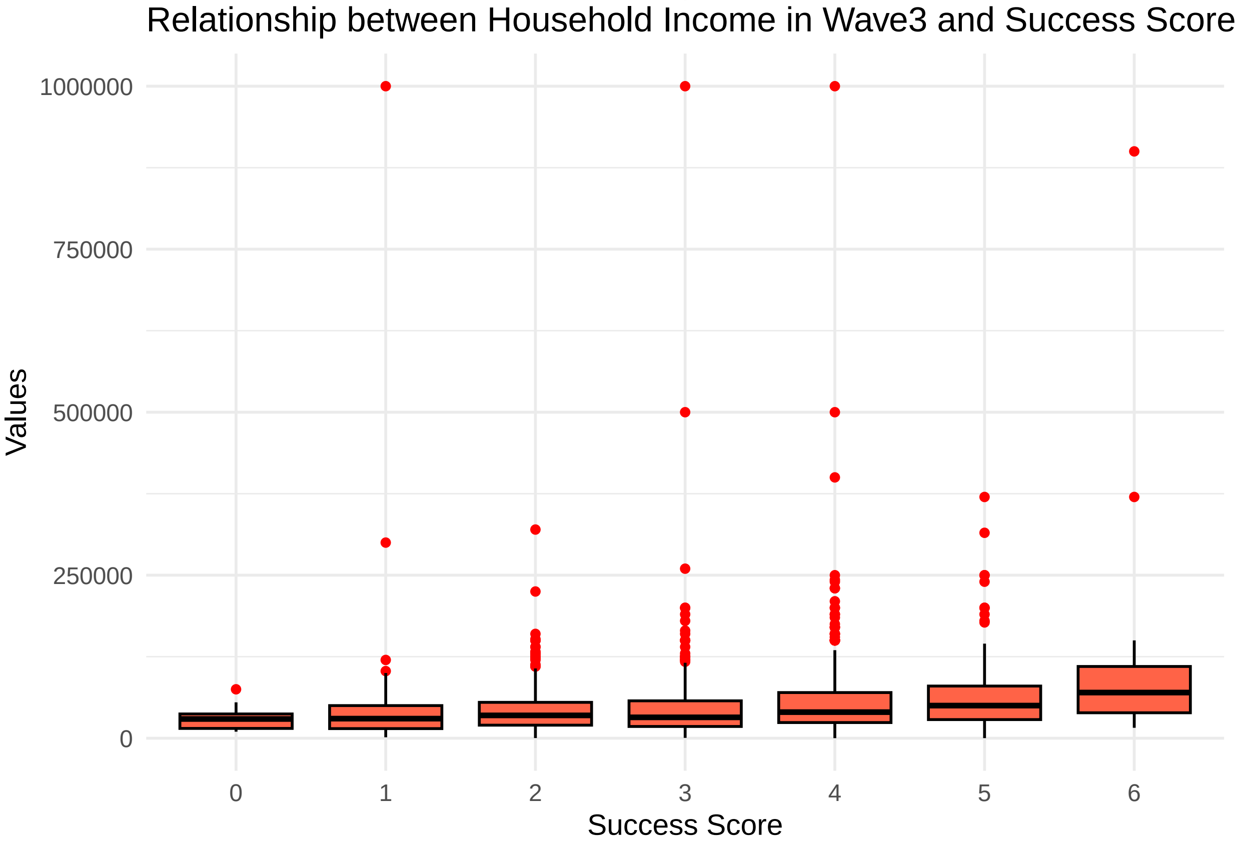


Distribution of Household Income in Wave1

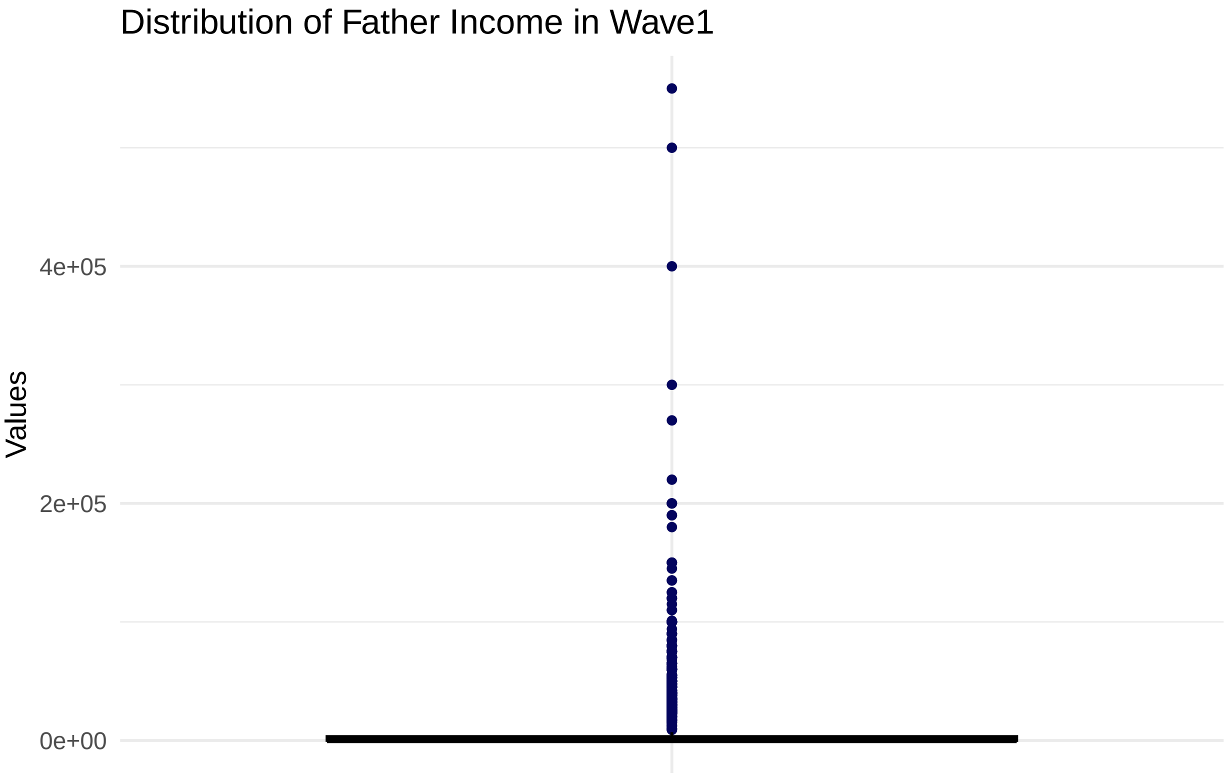


Relationship between Household Income in Wave1 and Success Score

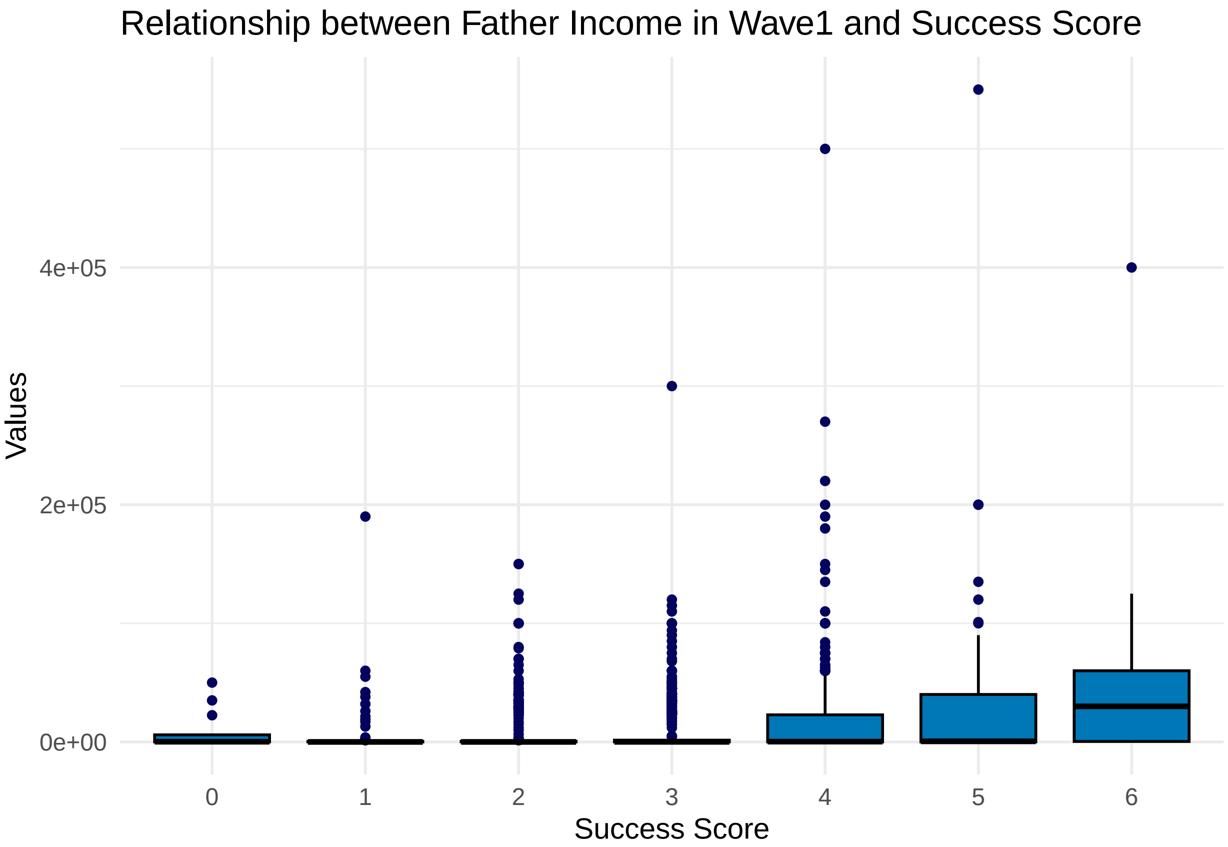




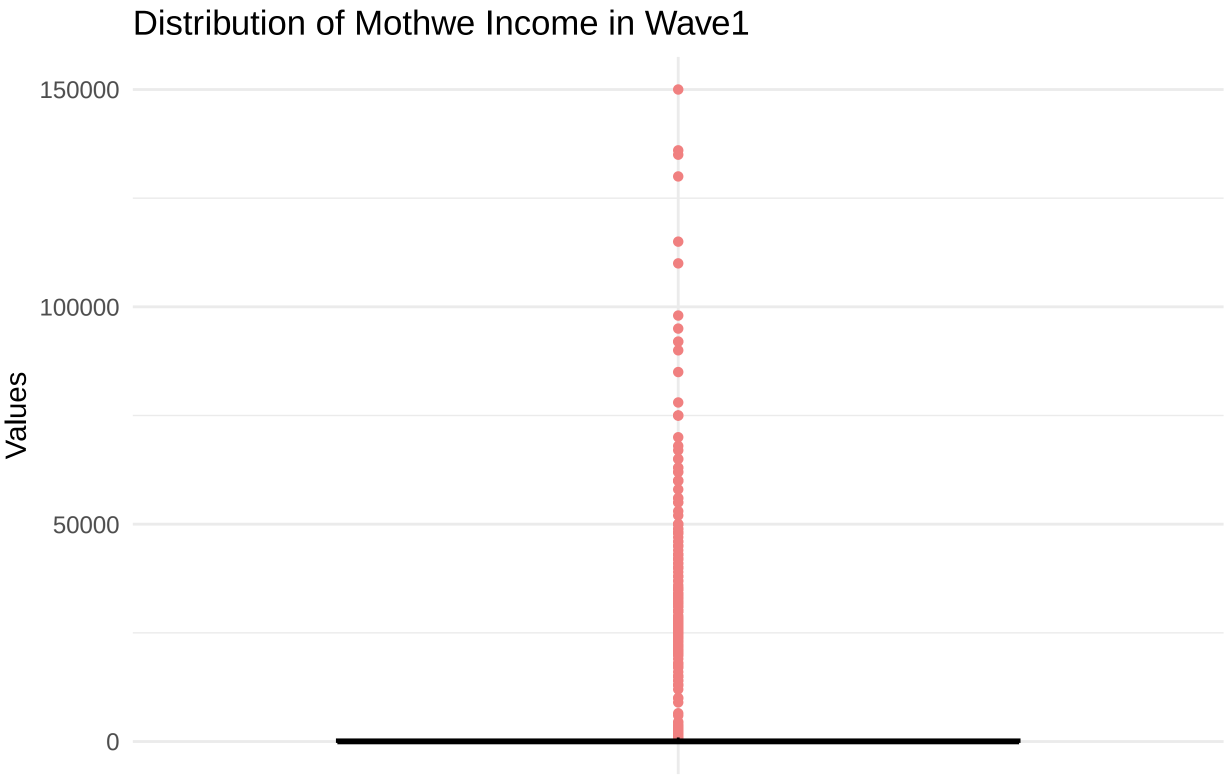
```r
#Wave 1 Father Income
ggplot(positive_data, aes(x = "",y = f1j8)) +
  geom_boxplot(fill = '#0077b6', color = "black", outlier.colour = "#03045e", outlier.shape = 16) +
  ggtitle("Distribution of Father Income in Wave1") +
  theme_minimal() +
  ylab("Values") +
  xlab("")
```



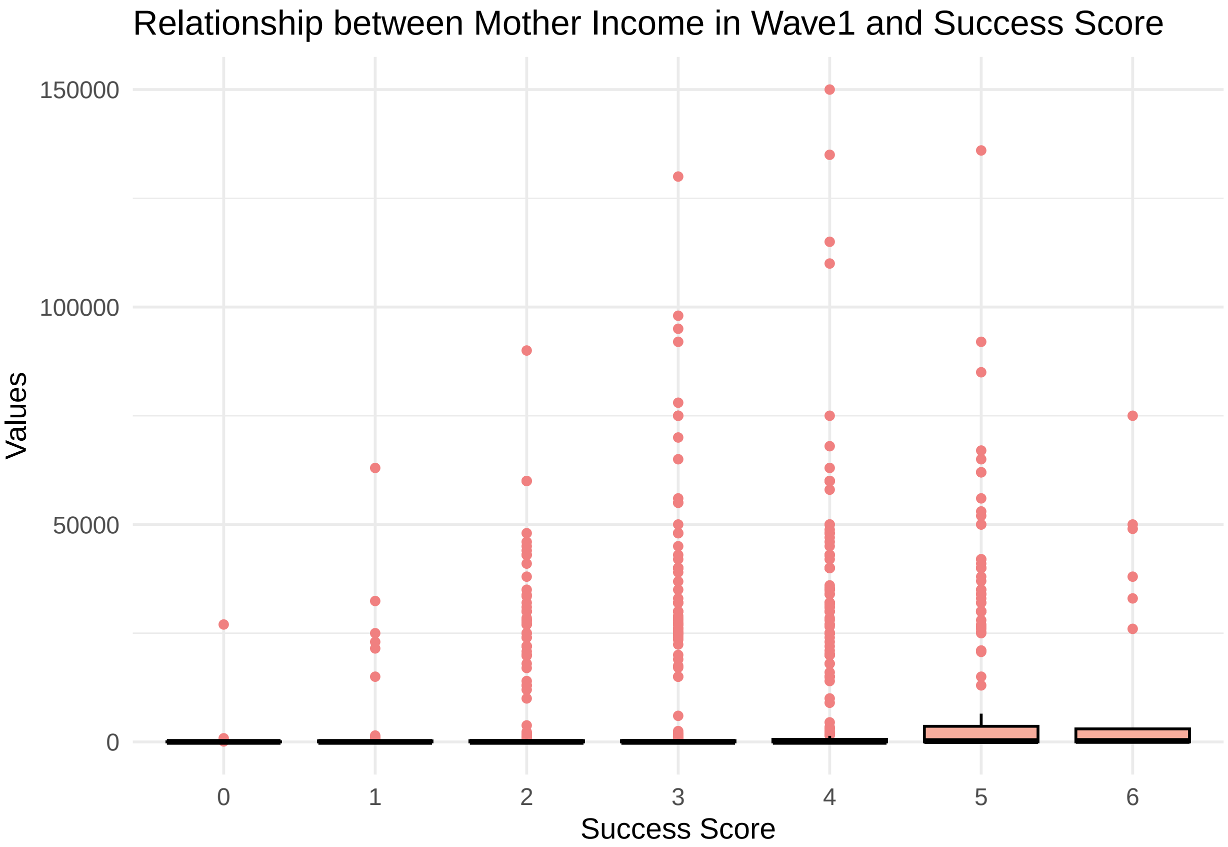
```
ggplot(positive_data, aes(x = as.factor(y_score),y = f1j8)) +
  geom_boxplot(fill = "#0077b6", color = "black", outlier.colour = "#03045e", outlier.shape = 16) +
  ggtitle("Relationship between Father Income in Wave1 and Success Score") +
  theme_minimal() +
  ylab("Values") +
  xlab("Success Score")
```



```
#Wave 1 Mother Income
ggplot(positive_data, aes(x = "",y = m1i2b)) +
  geom_boxplot(fill = '#f8ad9d', color = "black", outlier.colour = "#f08080", outlier.shape = 16) +
  ggtitle("Distribution of Mothwe Income in Wave1") +
  theme_minimal() +
  ylab("Values") +
  xlab("")
```

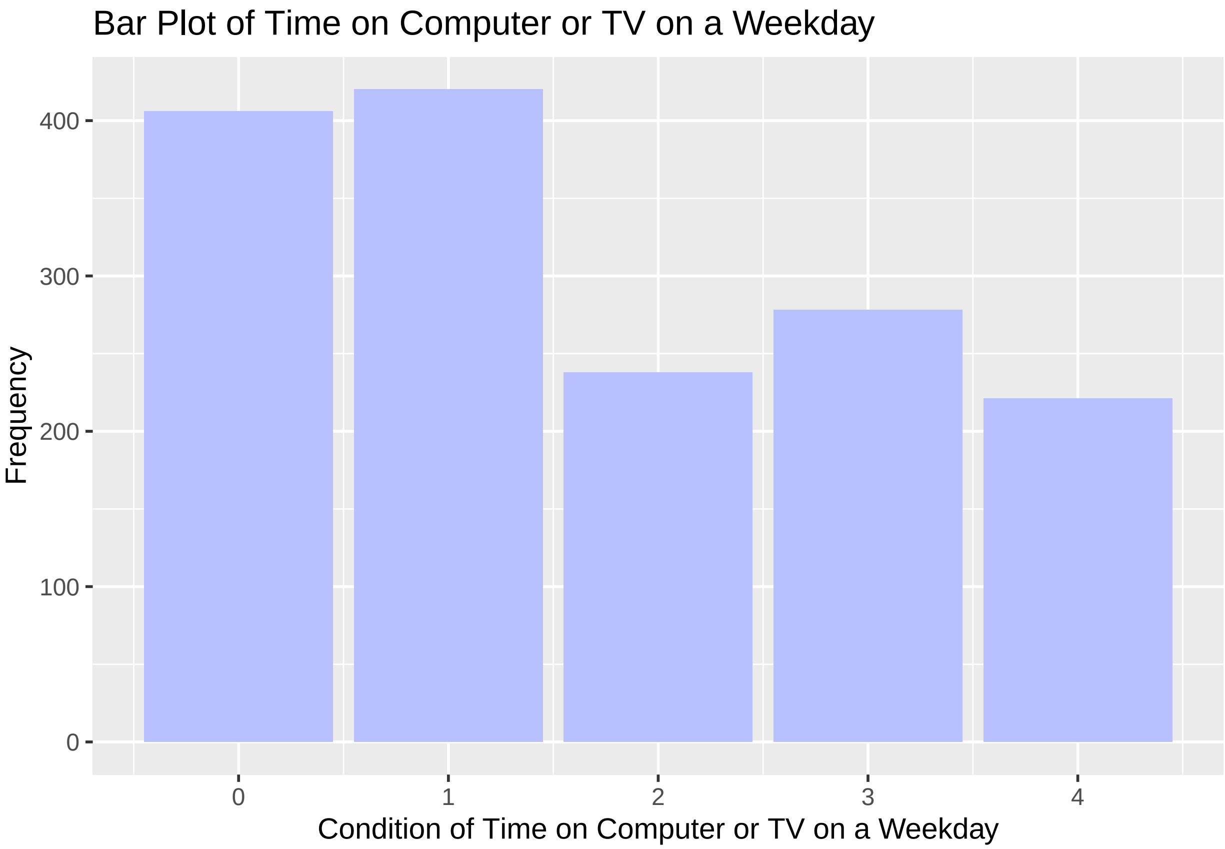


```r
ggplot(positive_data, aes(x = as.factor(y_score),y = m1i2b)) +
  geom_boxplot(fill = "#f8ad9d", color = "black", outlier.colour = "#f08080", outlier.shape = 16) +
  ggtitle("Relationship between Mother Income in Wave1 and Success Score") +
  theme_minimal() +
  ylab("Values") +
  xlab("Success Score")
```

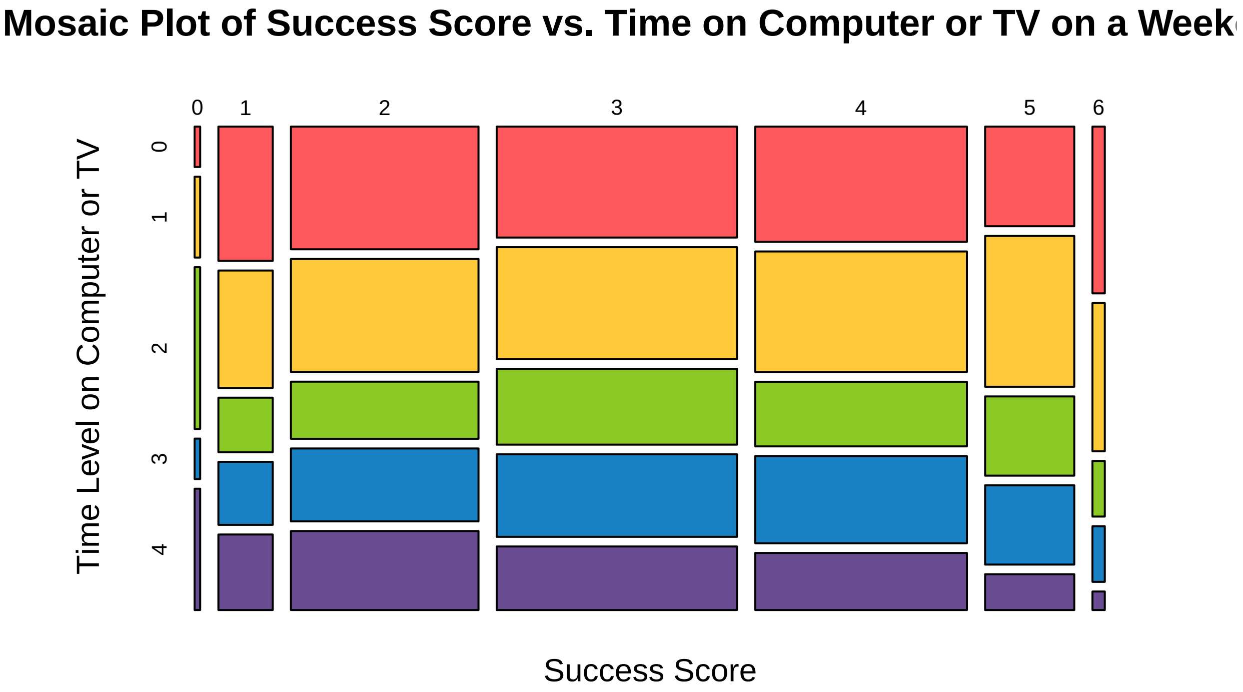


Categorical Variable: 'k5d1f', 'k5d1g', 'k5h1', 'k5d1e', 'k5f1f' k5d1f:Amount of time on a weekday you play computer games on the computer or TV k5d1g: Amount of time on a weekday you watch TV and movies k5h1: Condition of health in general k5d1e: Amount of time on a weekday you chat with friends on the computer k5f1f: Hurt an animal on purpose

```
positive_data <- positive_data %>%
  filter(k5d1f >= 0, k5d1g >= 0,k5h1>=0,k5d1e>=0,k5f1f>=0)

ggplot(positive_data, aes(x = k5d1f)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of Time on Computer or TV on a Weekday") +
  xlab("Condition of Time on Computer or TV on a Weekday") +
  ylab("Frequency")
```

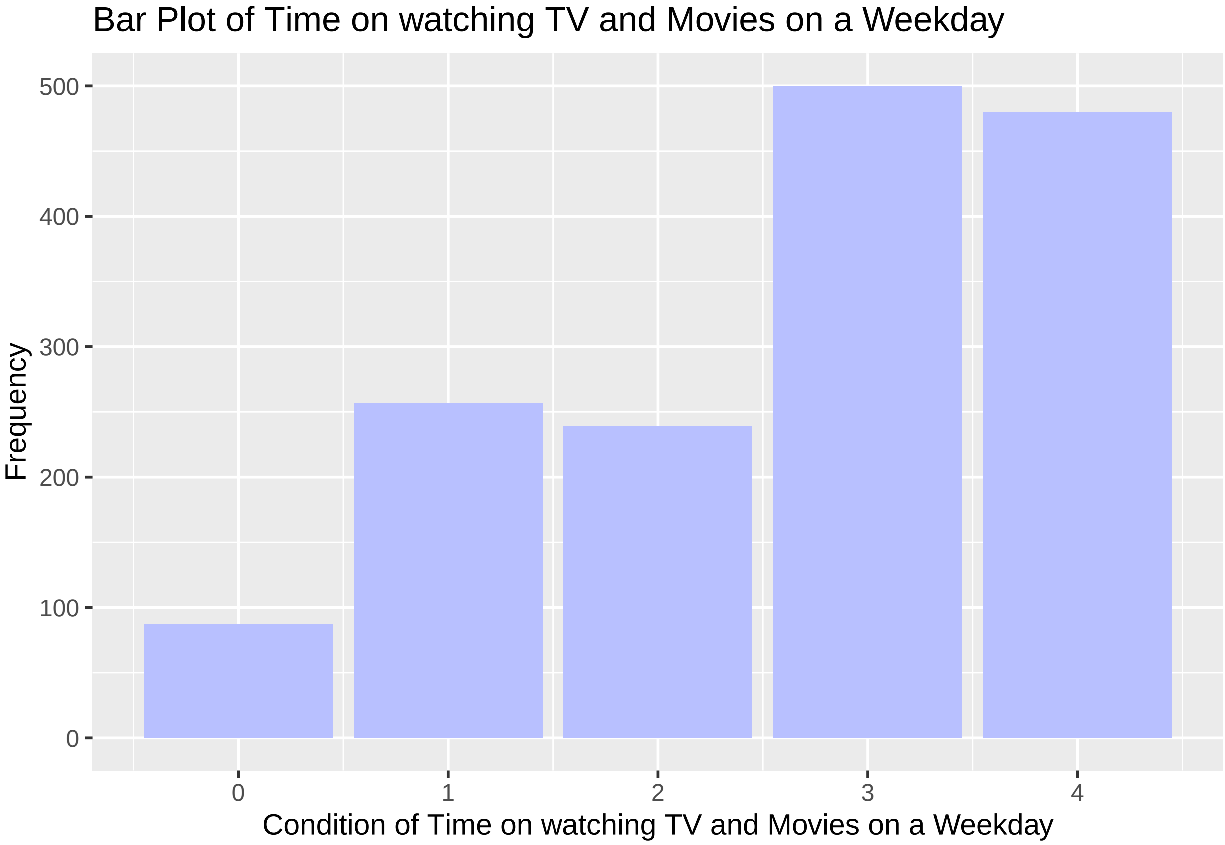


```
description_k5d1f <- c("0-none","1-half an hour or less","2-more than half an hour but less than an hou
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$k5d1f), main = "Mosaic Plot of Success Score vs. 
```

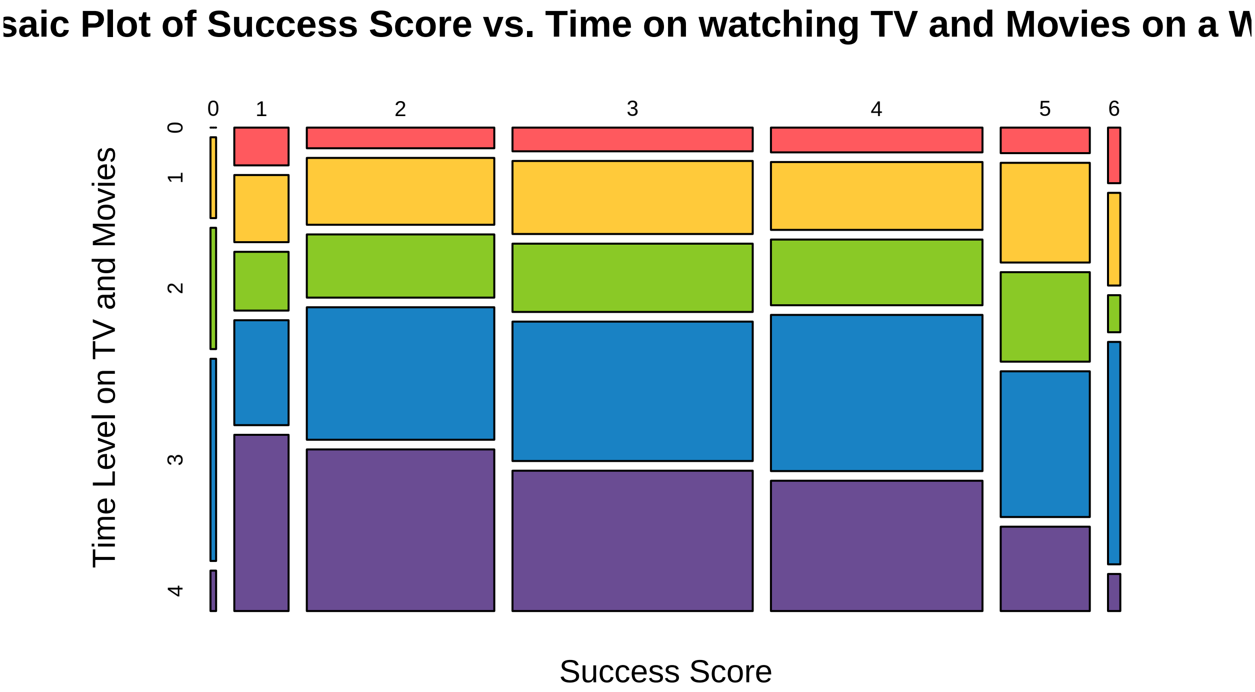


```
#par(xpd = TRUE, mar = par()$mar + c(0, 0, 0, 50))
#legend("topright", inset = c(-0.5, 0),legend = description_k5d1f, fill = colors, title = "Numbers", ce
```

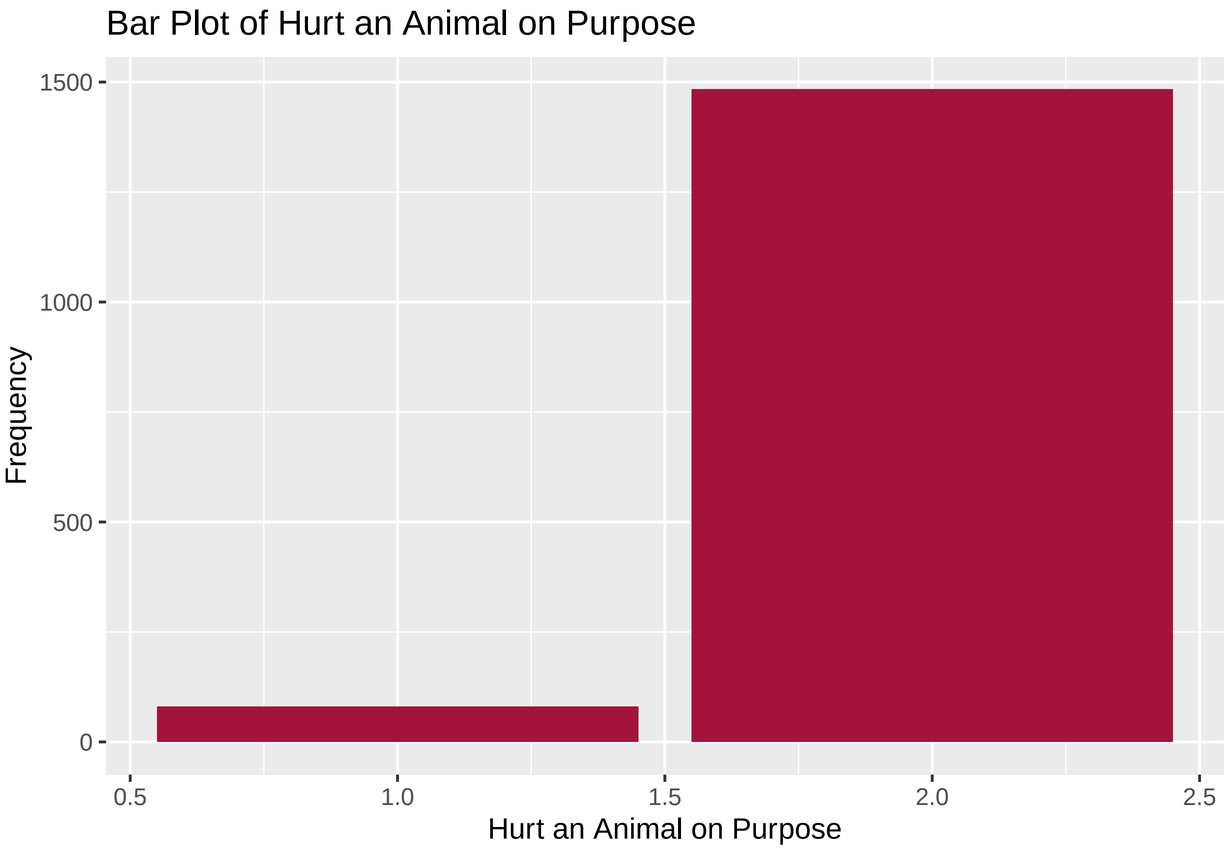
```
ggplot(positive_data, aes(x = k5d1g)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of Time on watching TV and Movies on a Weekday") +
  xlab("Condition of Time on watching TV and Movies on a Weekday") +
  ylab("Frequency")
```



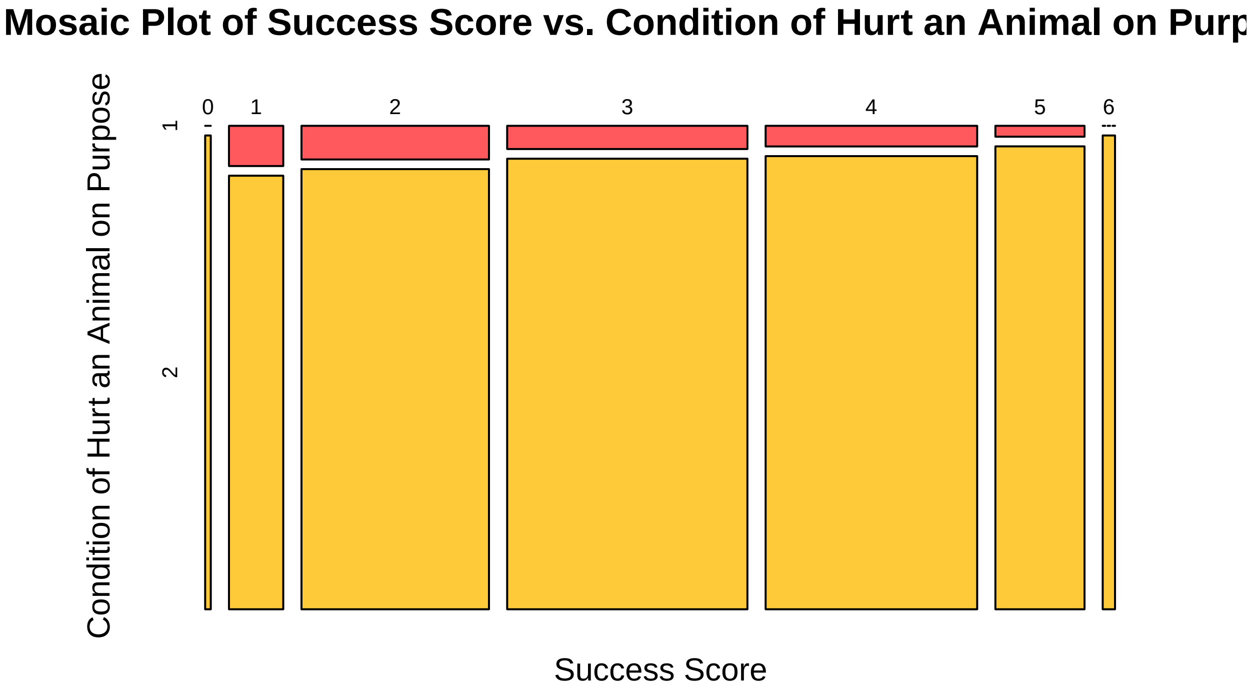
```
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$k5d1g), main = "Mosaic Plot of Success Score vs. '
```



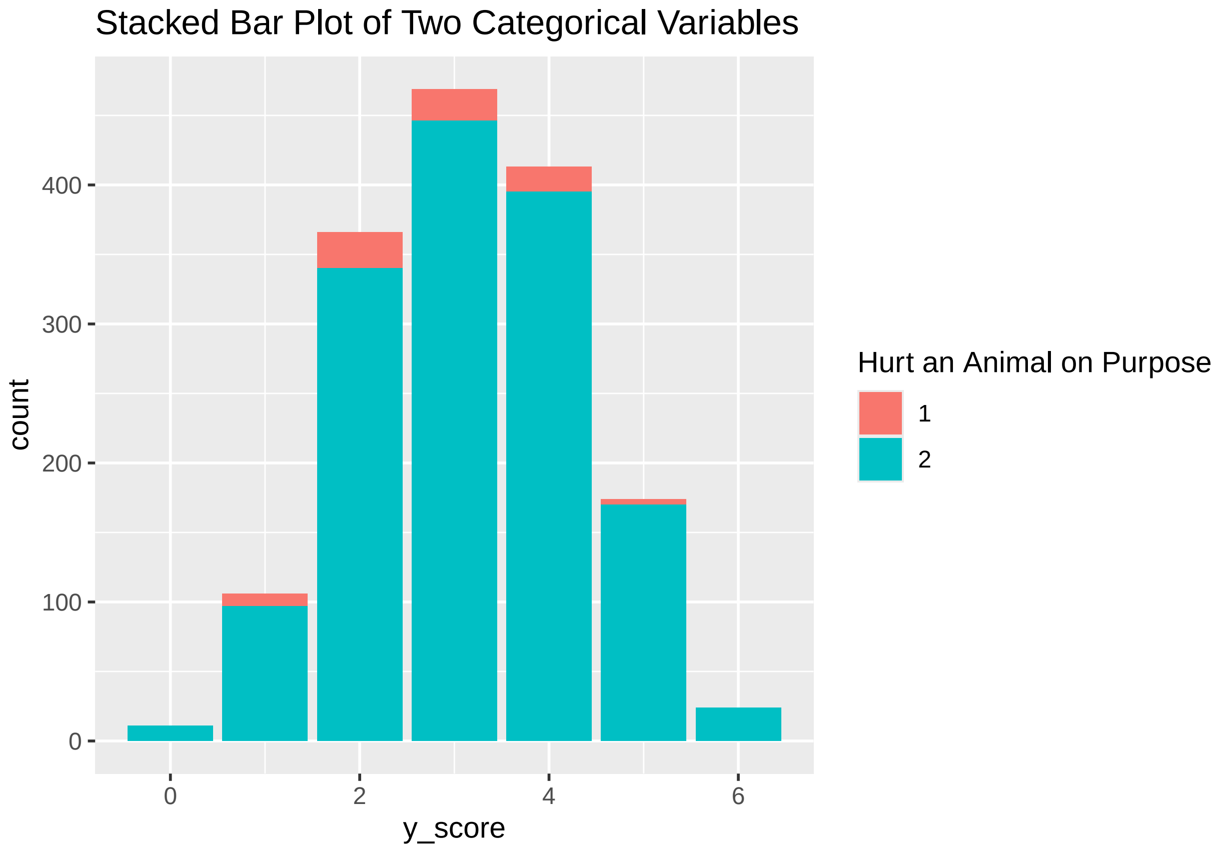
```
library(ggplot2)
#1 Yes - 2 No
ggplot(positive_data, aes(x = k5f1f)) +
  geom_bar(fill = "#a4133c") +
  ggtitle("Bar Plot of Hurt an Animal on Purpose") +
  xlab("Hurt an Animal on Purpose") +
  ylab("Frequency")
```



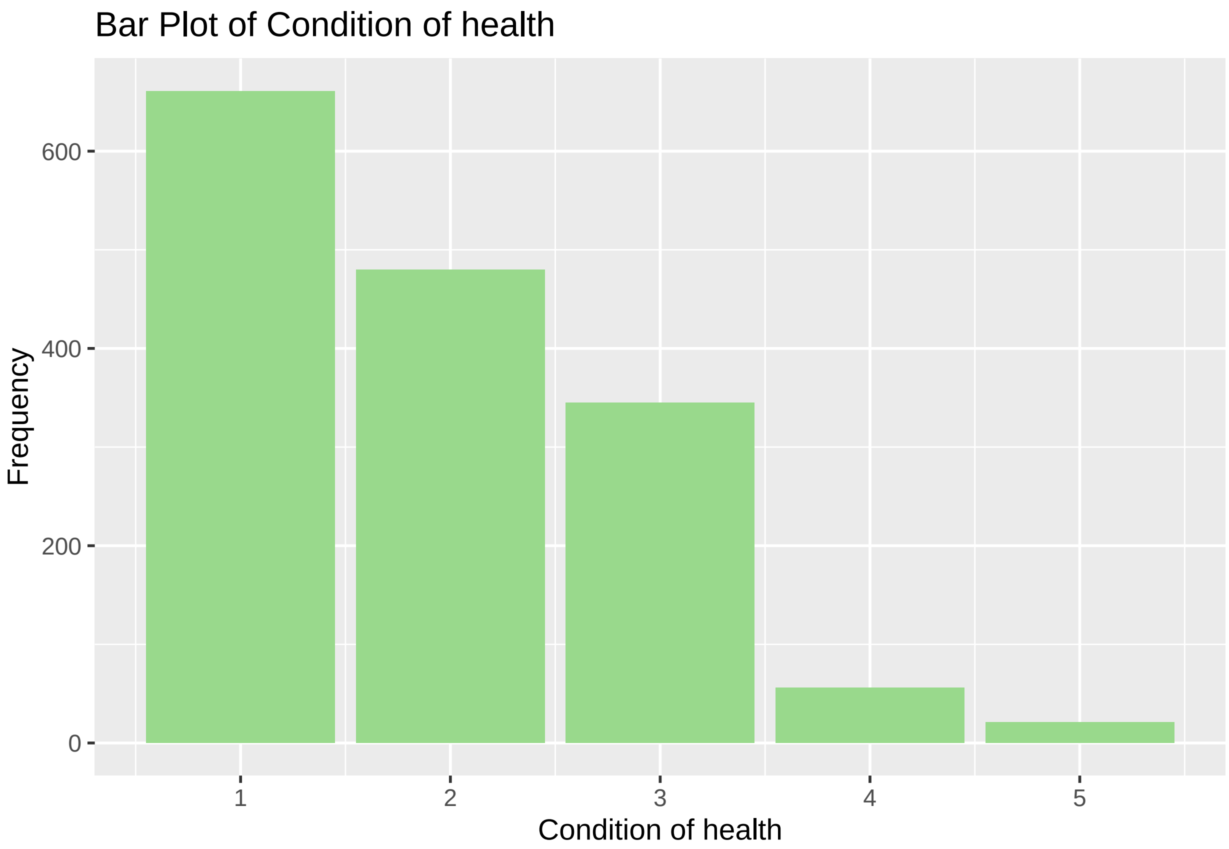
```
mosaicplot(table( positive_data$y_score,positive_data$k5f1f), main = "Mosaic Plot of Success Score vs.
```



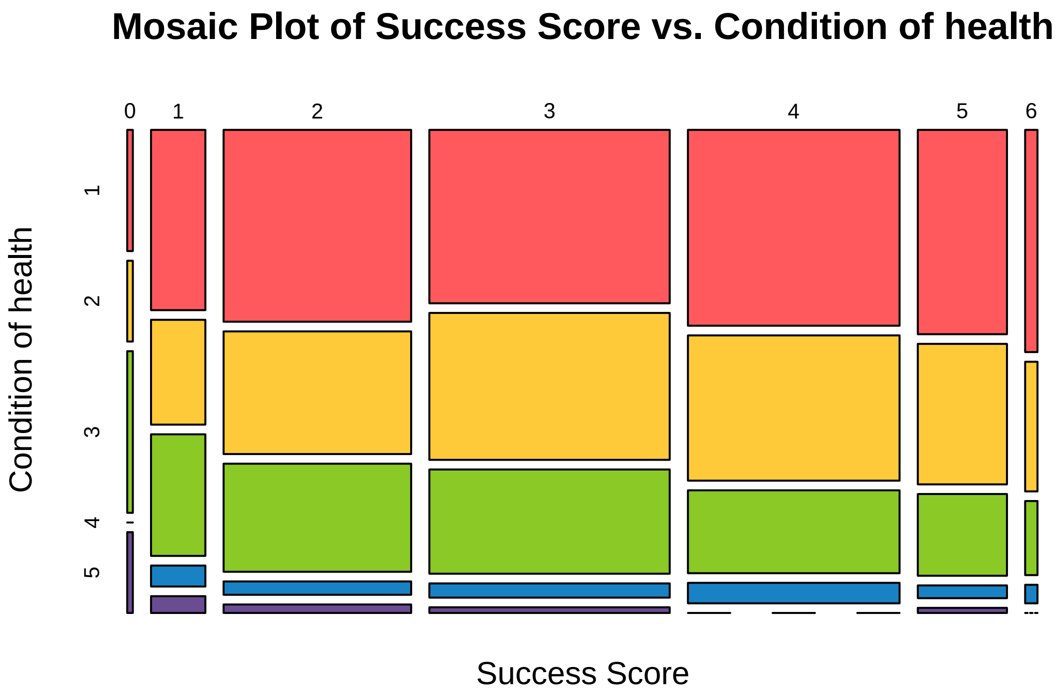
```
ggplot(positive_data, aes(x = y_score, fill = as.factor(k5f1f))) +
  geom_bar(position = "stack") +
  ggtitle("Stacked Bar Plot of Two Categorical Variables")+
  labs(fill = "Hurt an Animal on Purpose")
```



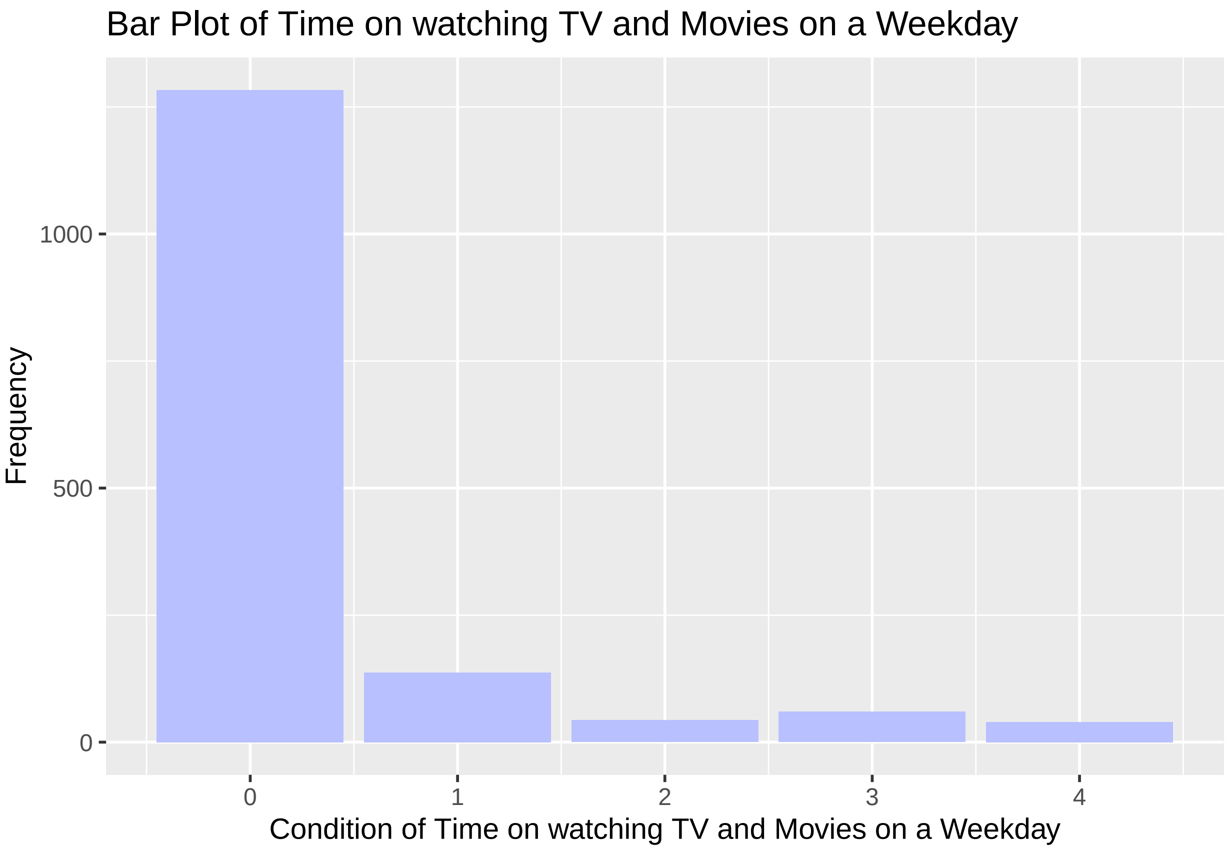
```
#k5h1:Condition of health in general
ggplot(positive_data, aes(x = k5h1)) +
  geom_bar(fill = "#99d98c") +
  ggtitle("Bar Plot of Condition of health") +
  xlab("Condition of health") +
  ylab("Frequency")
```



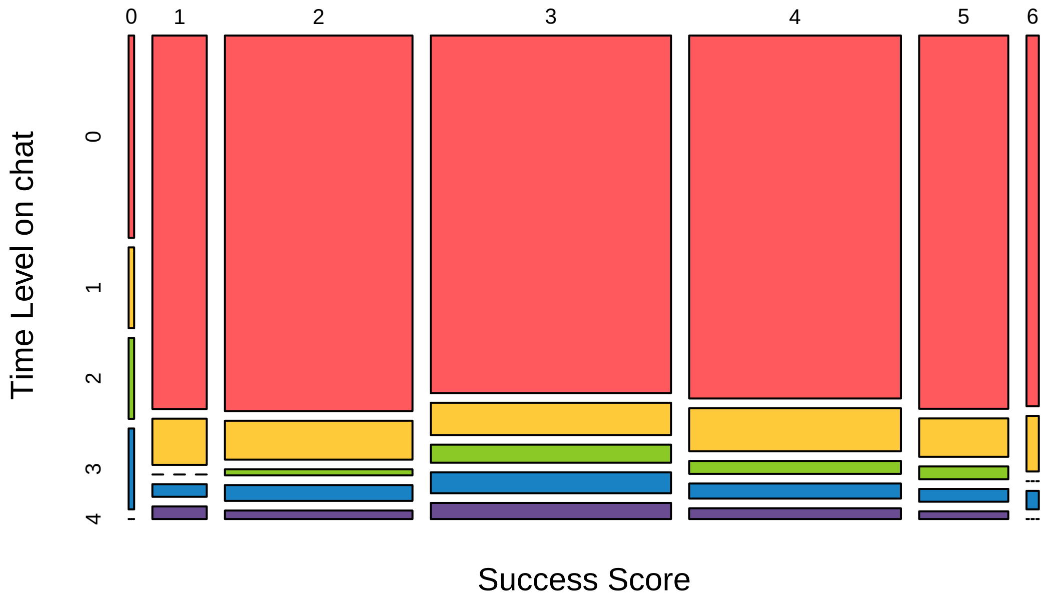
```
mosaicplot(table( positive_data$y_score,positive_data$k5h1), main = "Mosaic Plot of Success Score vs. Co
```



```
#k5d1e:Time on chat with friends on the computer(weekday)
#0  0 none; 1 half an hour or less;2 0.5 hour-1hour; 3 1-2 hours;4 more than 2 hours

ggplot(positive_data, aes(x = k5d1e)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of Time on watching TV and Movies on a Weekday") +
  xlab("Condition of Time on watching TV and Movies on a Weekday") +
  ylab("Frequency")
```

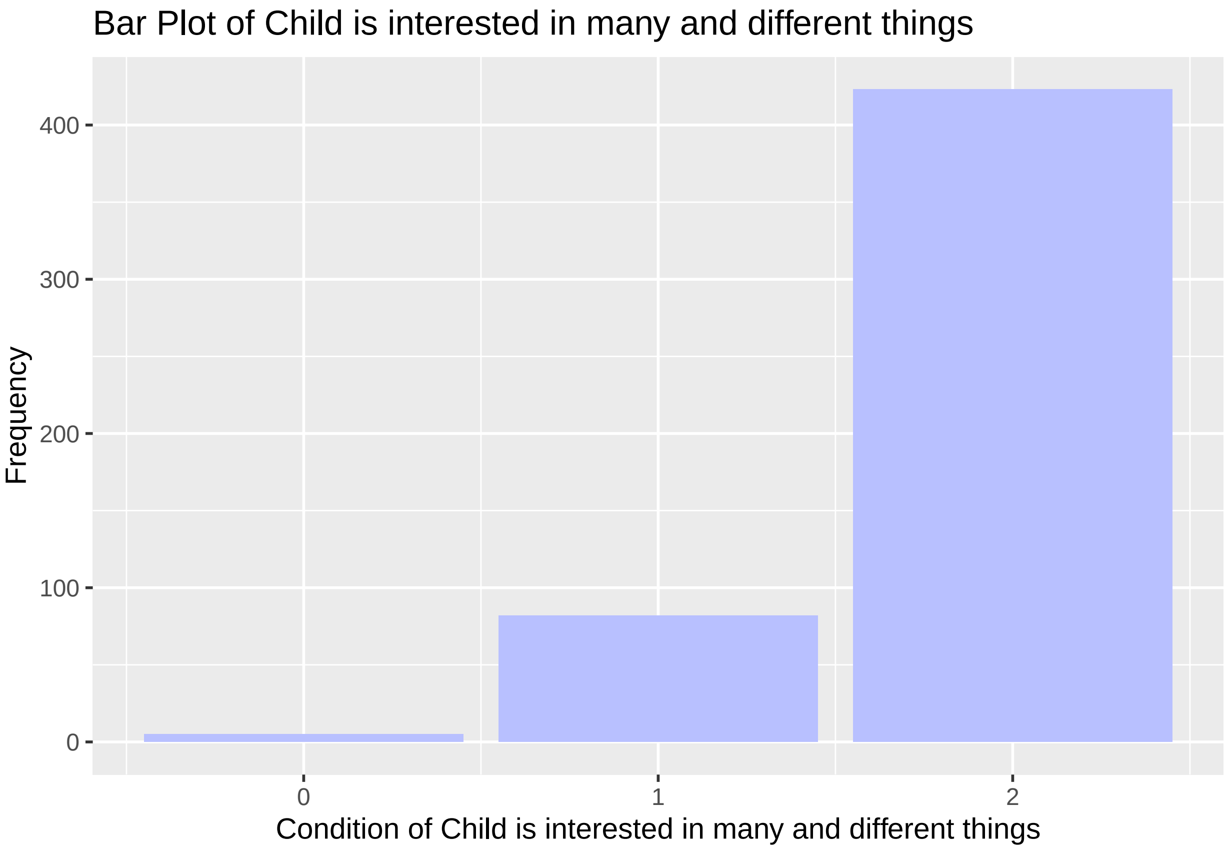


```
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$k5d1e), main = "Mosaic Plot of Success Score vs.
```


c Plot of Success Score vs. Time on chat with friends on the computer

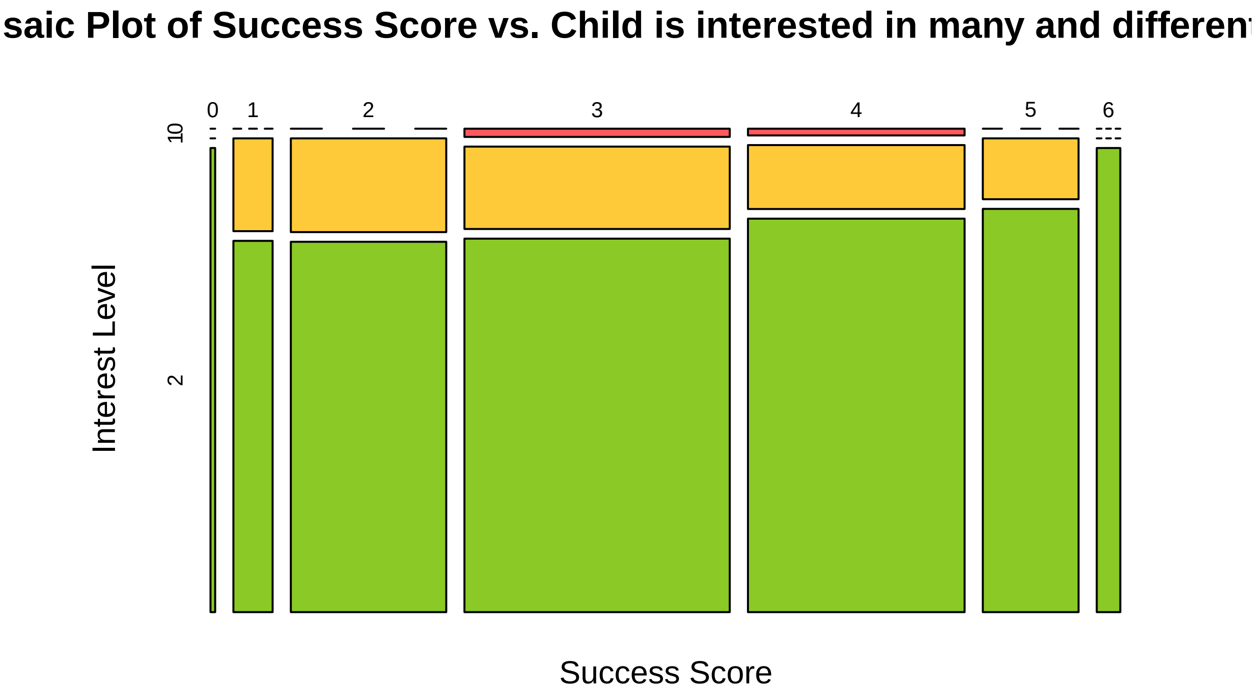
'p4l63', 'p4l59', 'p4d1c', 'm4d1b', 'f4b8d' p4l63: Child is interested in many and different things p4l59: Child threatens people p4d1c: Last 12M, couldn't afford to eat balanced meals m4d1b: You can trust father to take good care of child f4b8d: Does person/agency give money/voucher/scholarship to help pay for program?

```
positive_data <- positive_data %>%
  filter(p4l63 >= 0, p4l59 >= 0,p4d1c>=0,m4d1b>=0,f4b8d>=0)

#p4l63: Child is interested in many and different things: 0 not true; 1 somewhat true; 2 very true
ggplot(positive_data, aes(x = p4l63)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of Child is interested in many and different things") +
  xlab("Condition of Child is interested in many and different things") +
  ylab("Frequency")
```



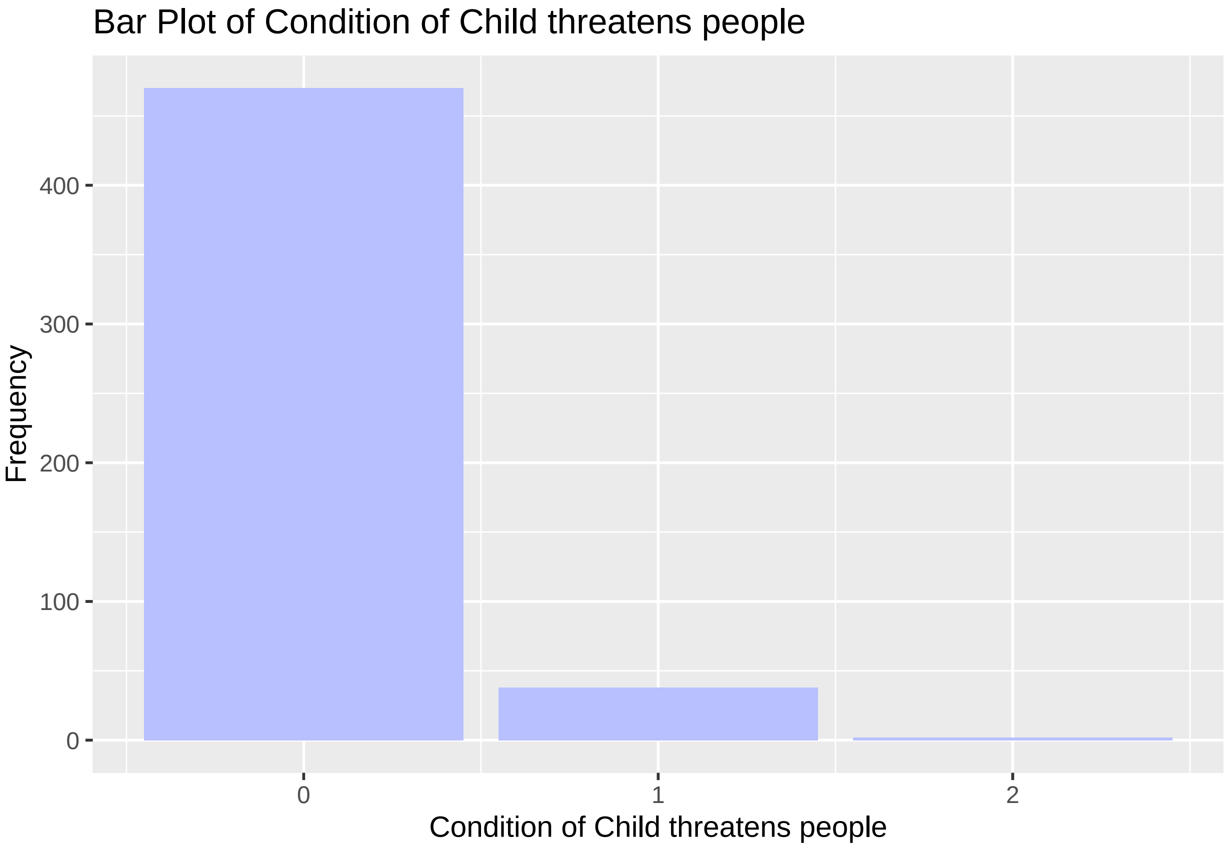
Bar Plot of Child is interested in many and different things

```
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$p4163), main = "Mosaic Plot of Success Score vs. C
```

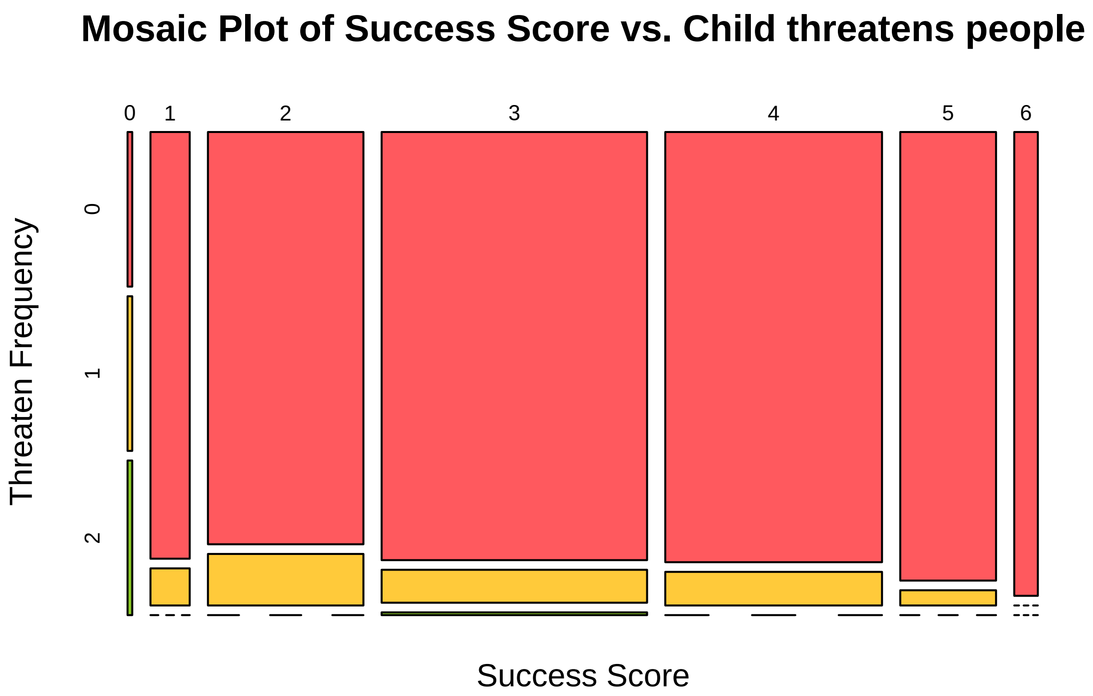


saic Plot of Success Score vs. Child is interested in many and differen

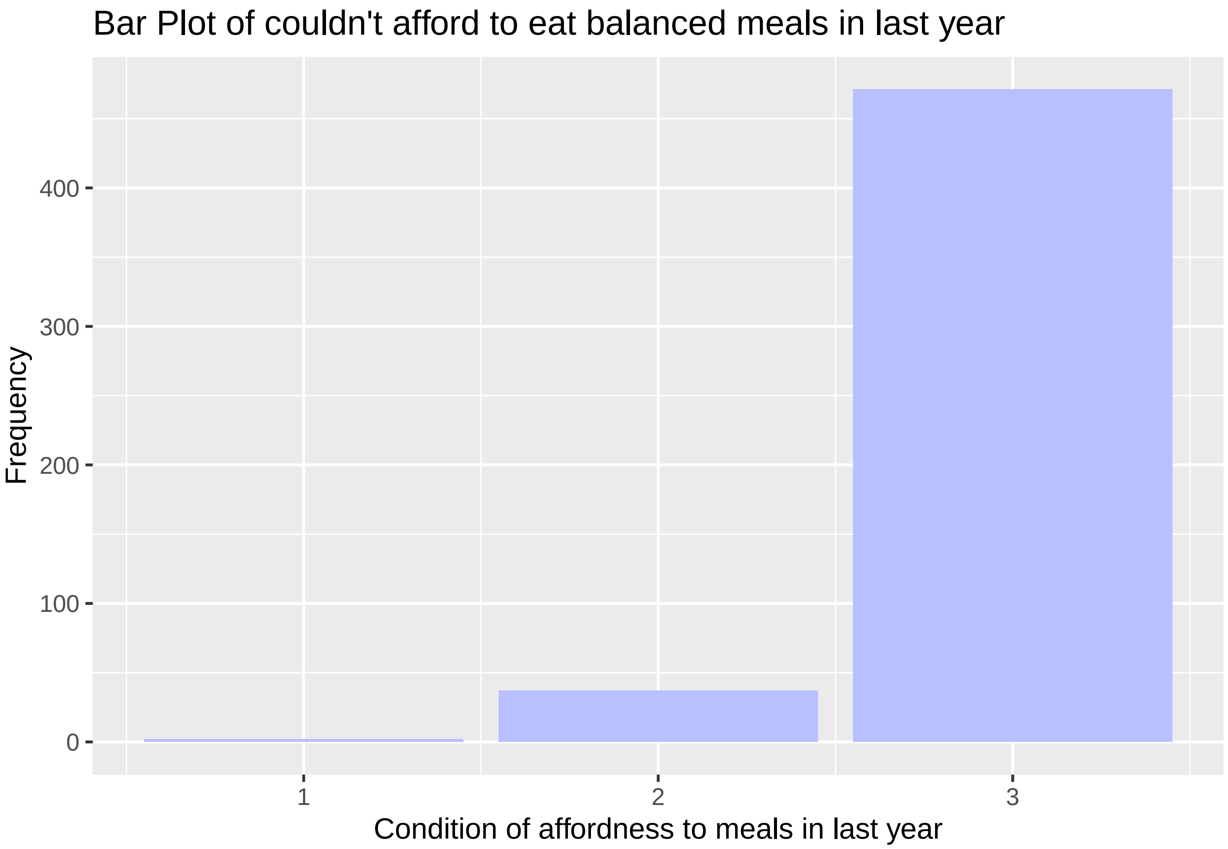
```
#p4159: Child threatens people: 0 not true; 1 somewhat true; 2 very true
ggplot(positive_data, aes(x = p4159)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of Condition of Child threatens people") +
  xlab("Condition of Child threatens people") +
  ylab("Frequency")
```



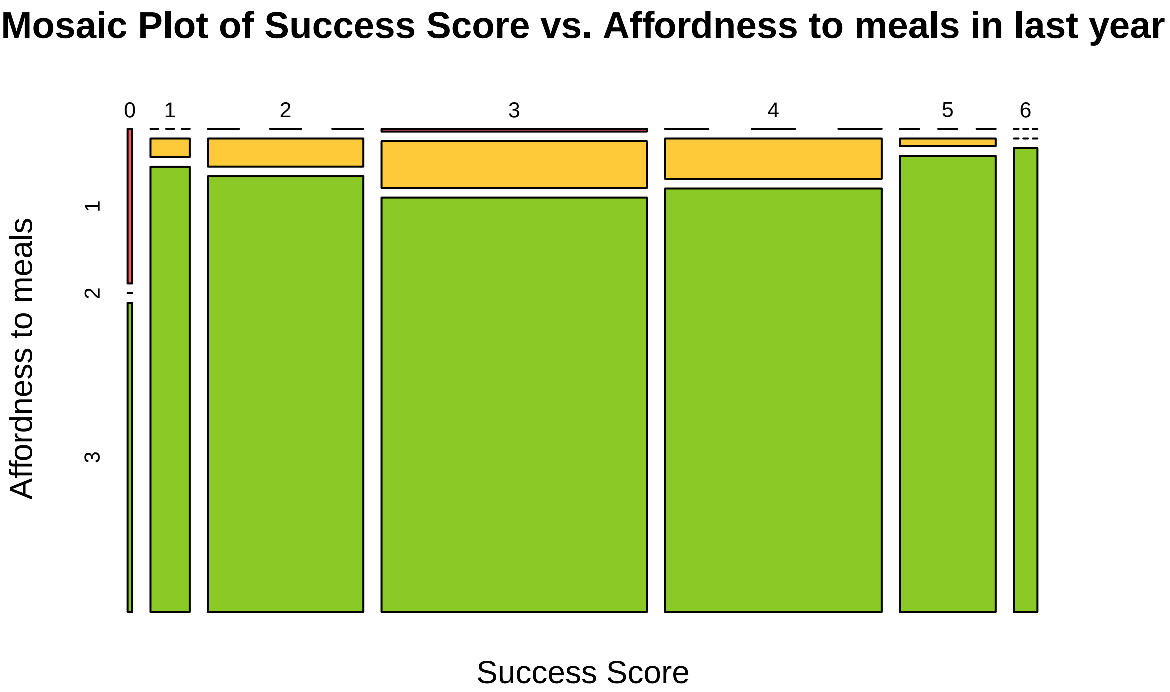
```
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$p4159), main = "Mosaic Plot of Success Score vs. (
```



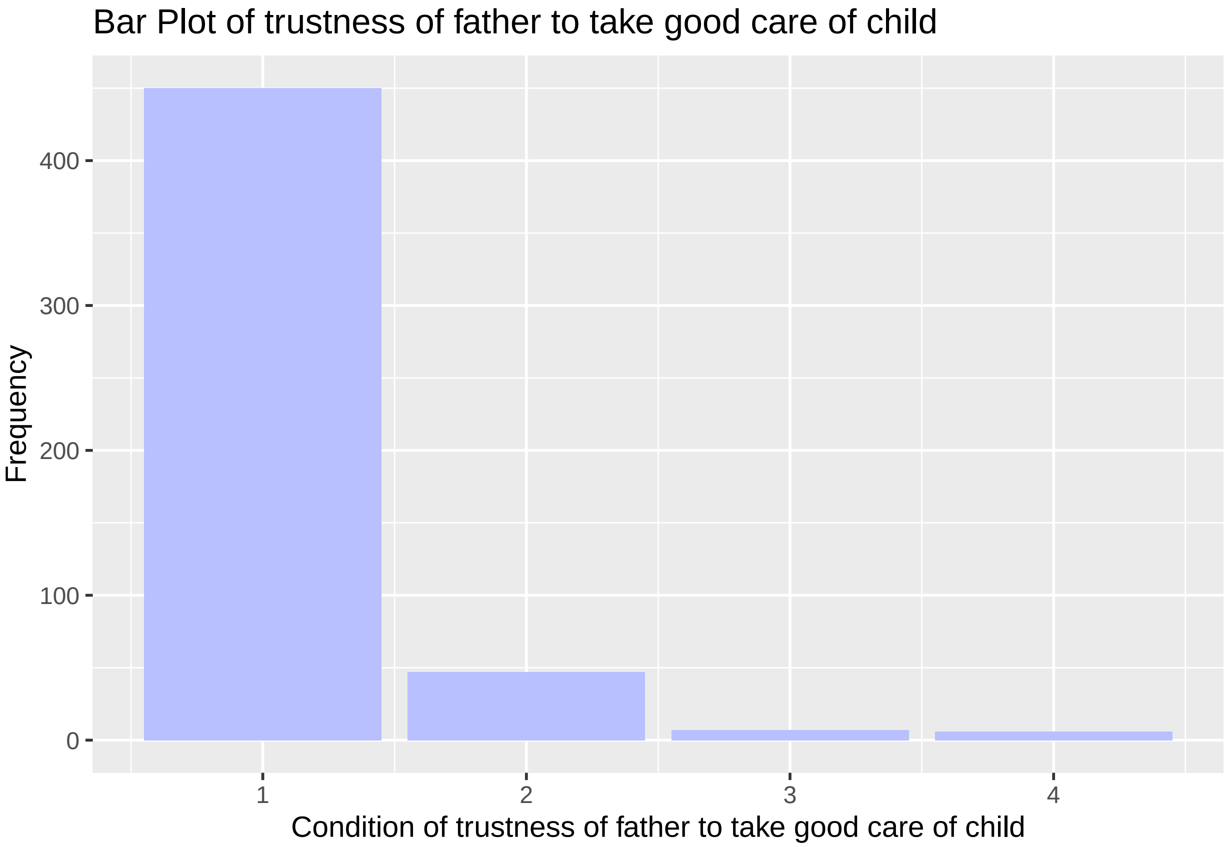
```r
#p4d1c:  Last 12M, couldn't afford to eat balanced meals: 1 often true; 2 sometimes true; 3 never treu
ggplot(positive_data, aes(x = p4d1c)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of couldn't afford to eat balanced meals in last year") +
  xlab("Condition of affordness to meals in last year") +
  ylab("Frequency")
```



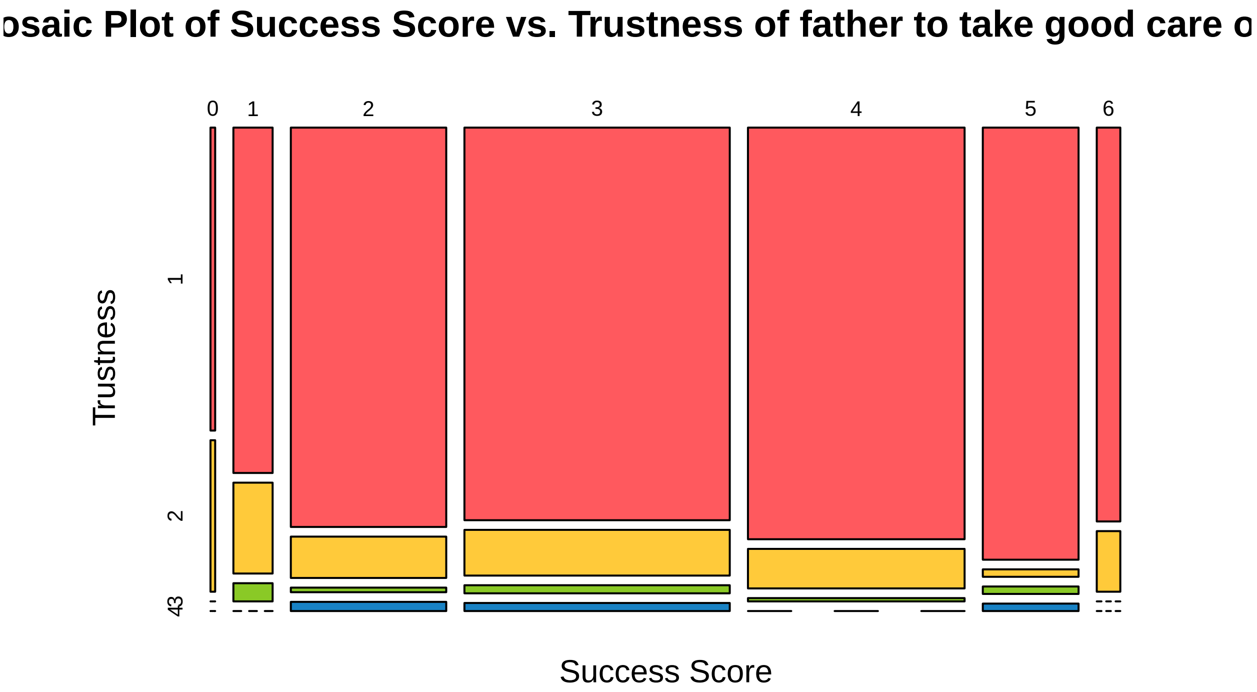
```
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$p4d1c), main = "Mosaic Plot of Success Score vs. 
```



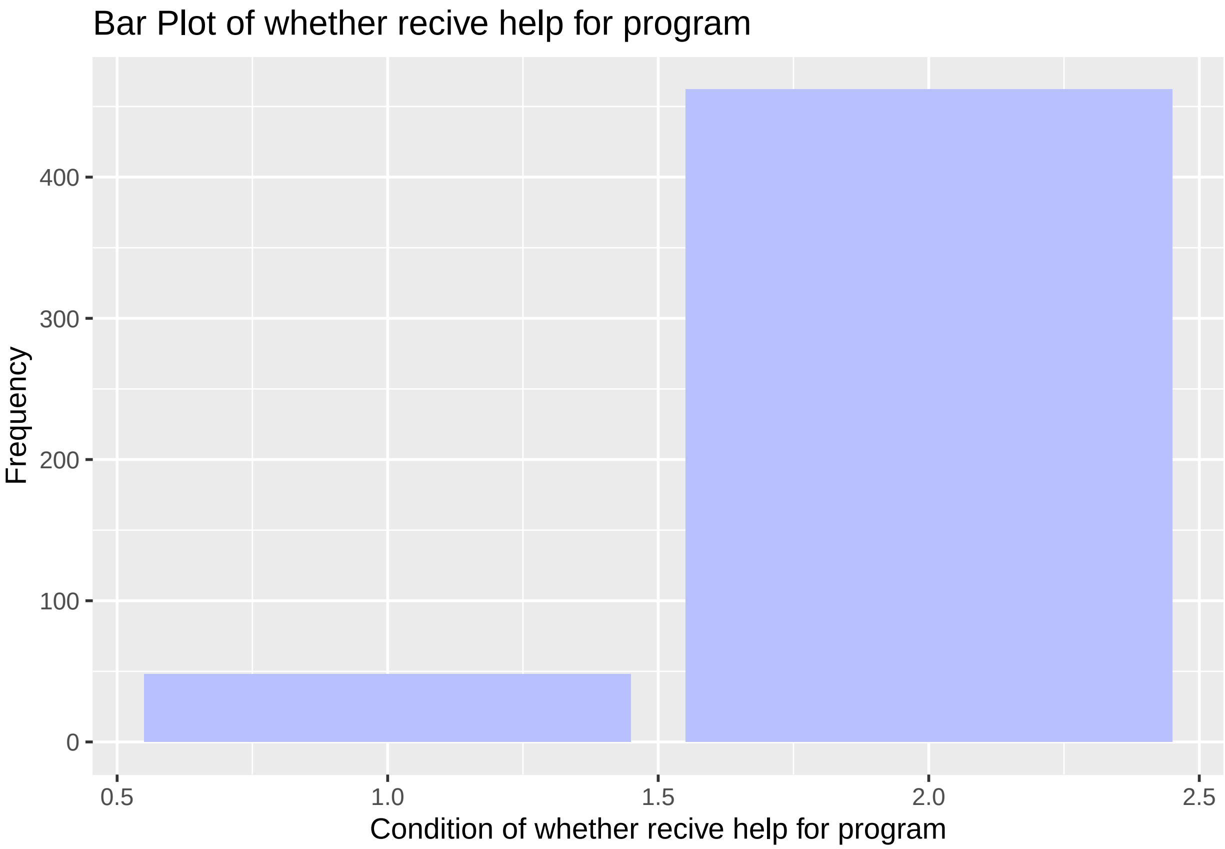
```
#m4d1b: You can trust father to take good care of child: 1 always; 2 sometimes; 3 rarely; 4 never
ggplot(positive_data, aes(x = m4d1b)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of trustness of father to take good care of child") +
  xlab("Condition of trustness of father to take good care of child") +
  ylab("Frequency")
```



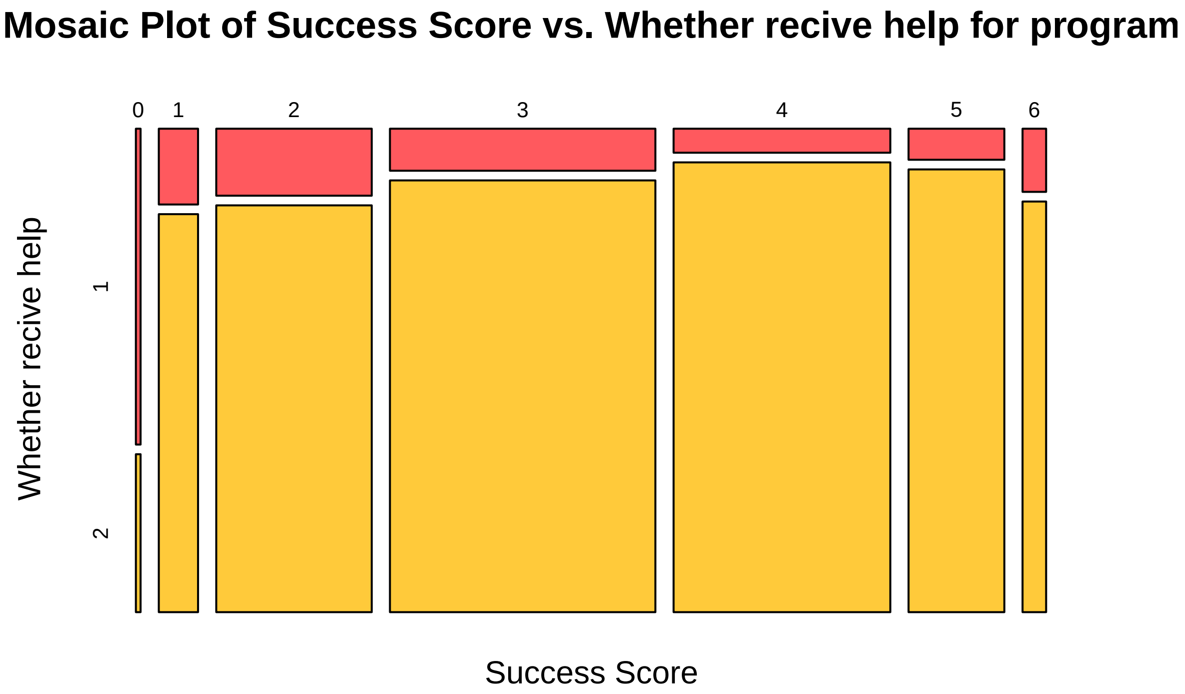
```
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$m4d1b), main = "Mosaic Plot of Success Score vs.
```



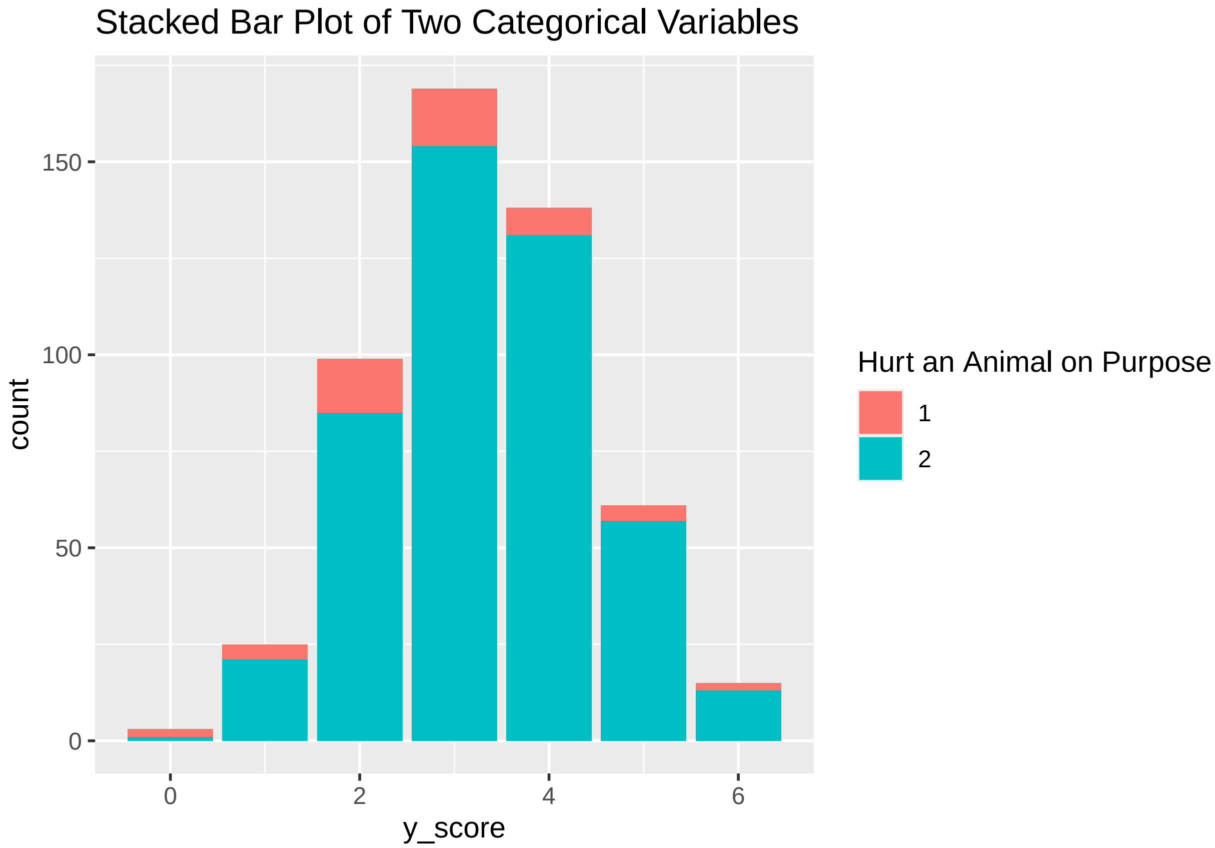
```
#f4b8d: Does person/agency give money/voucher/scholarship to help pay for program? 1 yes; 2 no
ggplot(positive_data, aes(x = f4b8d)) +
  geom_bar(fill = "#b8c0ff") +
  ggtitle("Bar Plot of whether recive help for program") +
  xlab("Condition of whether recive help for program") +
  ylab("Frequency")
```



```
colors <- c("#ff595e", "#ffca3a", "#8ac926","#1982c4","#6a4c93")
mosaicplot(table( positive_data$y_score,positive_data$f4b8d), main = "Mosaic Plot of Success Score vs. 
```



```
ggplot(positive_data, aes(x = y_score, fill = as.factor(f4b8d))) +
  geom_bar(position = "stack") +
  ggtitle("Stacked Bar Plot of Two Categorical Variables")+
  labs(fill = "Hurt an Animal on Purpose")
```


Stacked Bar Plot of Two Categorical Variables
count
y_score
Hurt an Animal on Purpose