knitr::opts\_chunk$set(echo = TRUE)  
library(causaldata)  
library(ggplot2)  
library(dplyr)

##   
## Присоединяю пакет: 'dplyr'

## Следующие объекты скрыты от 'package:stats':  
##   
## filter, lag

## Следующие объекты скрыты от 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ComplexUpset)  
library(ggstatsplot)

## Registered S3 method overwritten by 'parameters':  
## method from   
## format.parameters\_distribution datawizard

## You can cite this package as:  
## Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.  
## Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

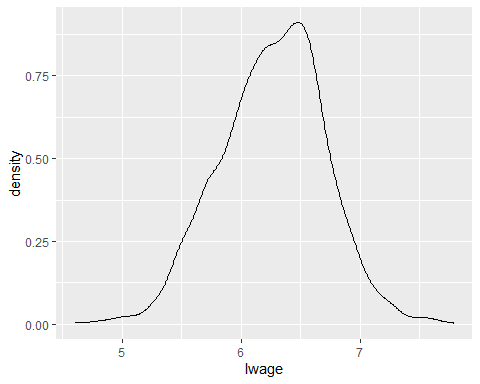
## ✓ tibble 3.1.6 ✓ purrr 0.3.4  
## ✓ tidyr 1.2.0 ✓ stringr 1.4.0  
## ✓ readr 2.1.2 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

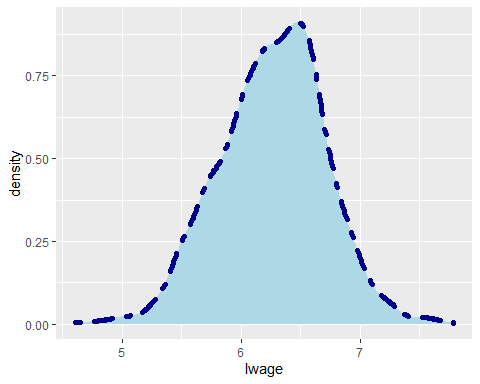
data(close\_college)  
close\_college

## # A tibble: 3,010 × 8  
## nearc4 educ black smsa south married exper lwage  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 7 1 1 0 1 16 6.31  
## 2 0 12 0 1 0 1 9 6.18  
## 3 0 12 0 1 0 1 16 6.58  
## 4 1 11 0 1 0 1 10 5.52  
## 5 1 12 0 1 0 1 16 6.59  
## 6 1 12 0 1 0 1 8 6.21  
## 7 1 18 0 1 0 1 9 6.34  
## 8 1 14 0 1 0 1 9 6.41  
## 9 1 12 0 1 0 4 10 6.05  
## 10 1 12 0 1 0 1 11 6.24  
## # … with 3,000 more rows

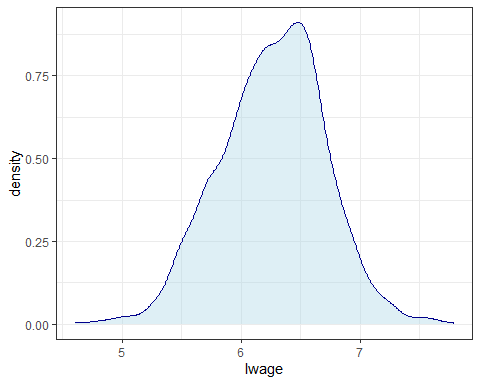
#lwage - логарифм заработной платы   
#educ - сколько лет учился   
#exper - сколько лет работал   
#black - 1 - афроамериканец   
#south - 1 - живет в южной части США   
#married - 1 - женат  
  
#РАБОТА С ПАКЕТОМ GGPLOT2  
  
#график №1 - график плотности распределения  
ggplot(data = close\_college, aes(x = lwage))+geom\_density()



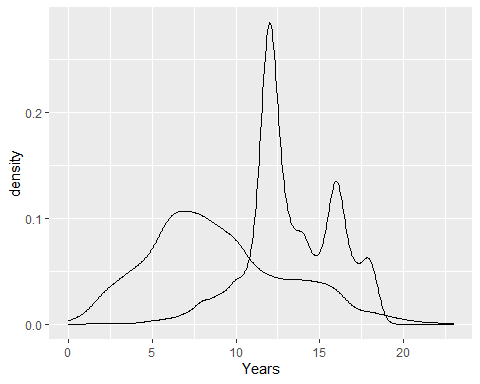
#настройка линии  
ggplot(data = close\_college, aes(x = lwage))+geom\_density(color = "darkblue", size = 2, linetype="dotdash", fill = "lightblue")



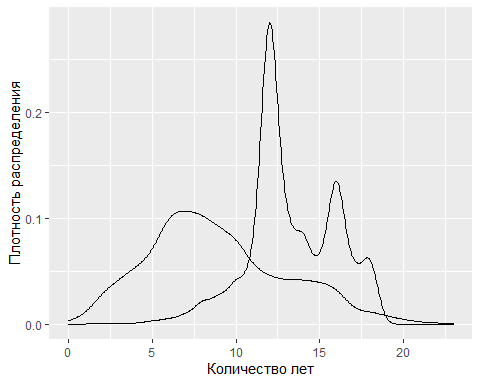
#color - цвет линии  
#size - размер линии  
#linetype - тип линии (blank, solid, dashed, dotted, dotdash, longdash, twodash)  
#fill - заливка под линией плотности распределения  
  
#настройка темы  
ggplot(data = close\_college, aes(x = lwage))+geom\_density(color = "darkblue", fill = "lightblue", alpha=0.4) + theme\_bw()



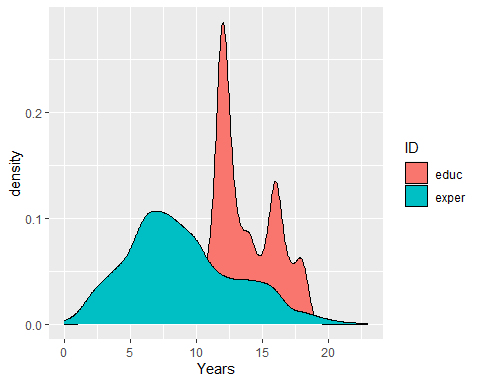
#дополнительные темы: theme\_gray, theme\_linedraw, theme\_light, theme\_dark, theme\_minimal, theme\_classic, theme\_void, theme\_test  
#alpha - прозрачность  
  
#график №2 - распределение двух параметров: опыта и образования  
#предварительно нужно сделать pivot  
data <- close\_college %>% pivot\_longer(cols = c("educ", "exper"), values\_to = "Years", names\_to = "ID")  
ggplot(data = data, aes(x = Years, group = ID)) + geom\_density()



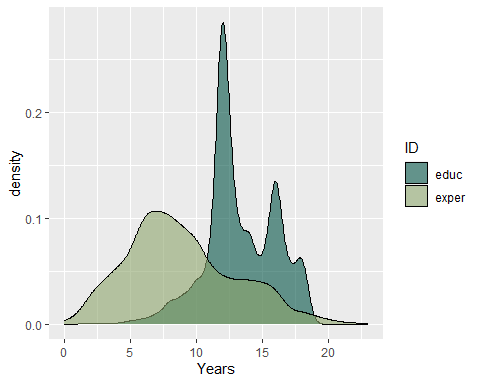
#подпись осей   
ggplot(data = data, aes(x = Years, group = ID)) + geom\_density()+ xlab("Количество лет") + ylab("Плотность распределения")



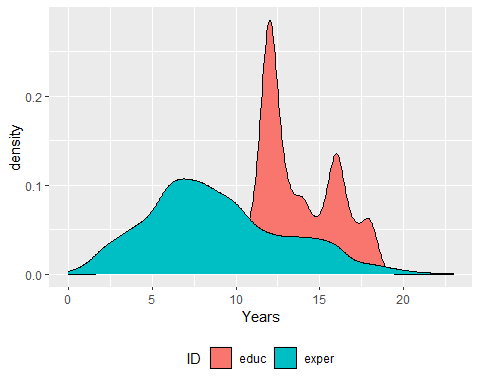
#раскраска  
ggplot(data = data, aes(x = Years, group = ID, fill = ID)) + geom\_density()



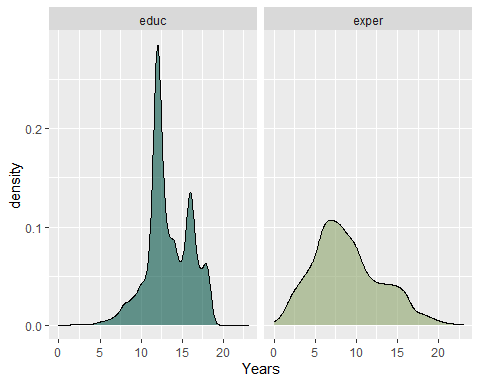
#настройка собственных цветов  
ggplot(data = data, aes(x = Years, group = ID, fill = ID)) + geom\_density(alpha = 0.6) + scale\_fill\_manual(values = c("#045346", "#8DA56C"))



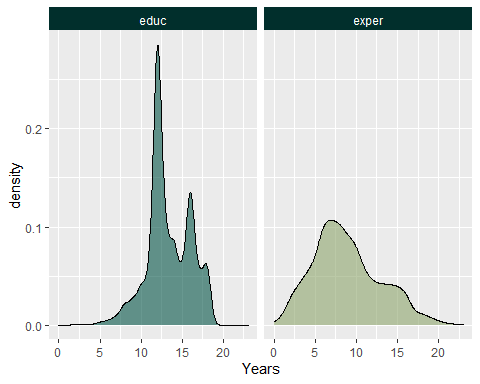
#цвета можно брать отсюда: https://color.romanuke.com/   
#alpha - прозрачность цвета  
#перемещение легенды вниз  
ggplot(data = data, aes(x = Years, group = ID, fill = ID)) + geom\_density() + theme(legend.position = "bottom")



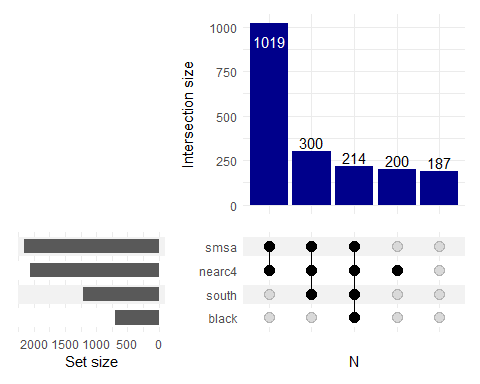
#можно сделать facet  
ggplot(data = data, aes(x = Years, group = ID, fill = ID)) + geom\_density(alpha = 0.6) + scale\_fill\_manual(values = c("#045346", "#8DA56C")) + facet\_grid(~ID) + theme(legend.position = "none")



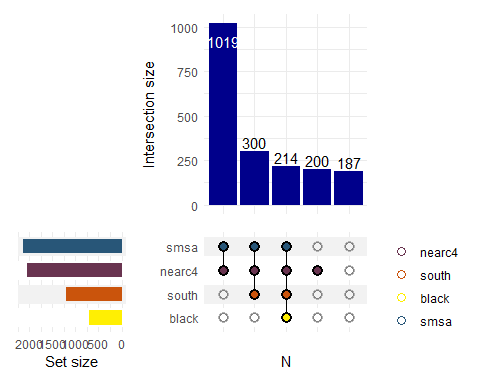
#раскрасить цвета facet (strip.background), цвет текста в facet (strip.text)  
ggplot(data = data, aes(x = Years, group = ID, fill = ID)) + geom\_density(alpha = 0.6) + scale\_fill\_manual(values = c("#045346", "#8DA56C")) + facet\_grid(~ID) + theme(legend.position = "none", strip.background =element\_rect(fill="#012F2C"), strip.text = element\_text(colour = 'white'))



#график №3 - распределение параметров по upset  
#строится по дамми-переменным black, south, married, smsa, nearc4  
dummy <- data.frame(close\_college$black, close\_college$south,   
 close\_college$smsa, close\_college$nearc4)  
colnames(dummy) <- c("black", "south", "smsa", "nearc4")  
N <- names(dummy)  
ComplexUpset::upset(dummy, N, name='N', width\_ratio = 0.4, n\_intersections = 5,   
 base\_annotations = list('Intersection size'=intersection\_size(fill='darkblue')))



#настройка цветов в UpSet  
ComplexUpset::upset(dummy, N, name='N', width\_ratio = 0.4, n\_intersections = 5,   
 base\_annotations = list('Intersection size'=intersection\_size(fill='darkblue')),   
 matrix=(intersection\_matrix(geom=geom\_point(shape='circle filled', size=3))  
 + scale\_color\_manual(values=c('nearc4'='#693450', 'south'='#ca540c', 'black'='#ffef03',  
 'smsa'='#285678'))),   
 queries=list(upset\_query(set='nearc4', fill='#693450'),  
 upset\_query(set='south', fill='#ca540c'),  
 upset\_query(set='black', fill='#ffef03'),   
 upset\_query(set='smsa', fill='#285678')))



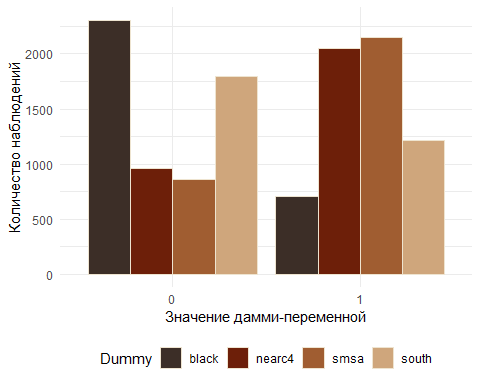
#график №4 - гистограммы отдельные по дамми  
close\_college

## # A tibble: 3,010 × 8  
## nearc4 educ black smsa south married exper lwage  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 7 1 1 0 1 16 6.31  
## 2 0 12 0 1 0 1 9 6.18  
## 3 0 12 0 1 0 1 16 6.58  
## 4 1 11 0 1 0 1 10 5.52  
## 5 1 12 0 1 0 1 16 6.59  
## 6 1 12 0 1 0 1 8 6.21  
## 7 1 18 0 1 0 1 9 6.34  
## 8 1 14 0 1 0 1 9 6.41  
## 9 1 12 0 1 0 4 10 6.05  
## 10 1 12 0 1 0 1 11 6.24  
## # … with 3,000 more rows

summary(close\_college)

## nearc4 educ black smsa   
## Min. :0.0000 Min. : 1.00 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:12.00 1st Qu.:0.0000 1st Qu.:0.000   
## Median :1.0000 Median :13.00 Median :0.0000 Median :1.000   
## Mean :0.6821 Mean :13.26 Mean :0.2336 Mean :0.713   
## 3rd Qu.:1.0000 3rd Qu.:16.00 3rd Qu.:0.0000 3rd Qu.:1.000   
## Max. :1.0000 Max. :18.00 Max. :1.0000 Max. :1.000   
##   
## south married exper lwage   
## Min. :0.0000 Min. :1.000 Min. : 0.000 Min. :4.605   
## 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.: 6.000 1st Qu.:5.977   
## Median :0.0000 Median :1.000 Median : 8.000 Median :6.287   
## Mean :0.4037 Mean :2.271 Mean : 8.856 Mean :6.262   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:11.000 3rd Qu.:6.564   
## Max. :1.0000 Max. :6.000 Max. :23.000 Max. :7.785   
## NA's :7

#сделаем pivot  
data1 <- close\_college %>% pivot\_longer(cols = c("black", "smsa", "south", "nearc4"), values\_to = "Variable", names\_to = "Dummy")  
data1$Variable <- as.character(data1$Variable)  
#распределение по группам, position: dodge, fill  
ggplot(data = data1, aes(x = Variable, fill = Dummy)) + geom\_bar(position="dodge", color = "#eee4d0") + scale\_fill\_manual(values = c("#3c2e27", "#6d1f09", "#a05d31", "#cfa67c")) + xlab("Значение дамми-переменной") + ylab("Количество наблюдений") + theme\_minimal() + theme(legend.position = "bottom")



#ВАЖНО: настройки theme ставить после настройки конкретной темы (например: theme\_minimal)  
  
#график №5 - статистически значимая корреляция  
library(ggside)

## Registered S3 method overwritten by 'ggside':  
## method from   
## +.gg ggplot2

ggscatterstats(  
 data = close\_college,  
 y = lwage,  
 x = exper,  
 ylab = "Логарифм заработной платы",  
 xlab = "Опыт работы в годах",  
 title = "Диаграмма рассеивания опыта работы и логарифма заработной платы",   
 xfill = "#018abd",   
 yfill = "#97cbdc",   
 smooth.line.args = list(size = 1.5, color = "darkblue", method = "lm", formula = y ~ x,  
 na.rm = TRUE),   
 point.args = list(size = 3, alpha = 0.4, color = "#0093b7", stroke = 0, na.rm = TRUE)  
)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

