Sazykin\_data\_visualization

Sazykin Georgie

2023-10-31

library(ggplot2)  
library(ggpubr)

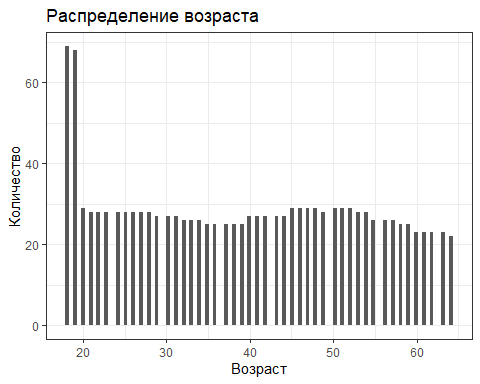
df <- read.csv('insurance\_cost.csv', stringsAsFactors = TRUE)  
#df

## 2. Гистограммы всех нумерических переменных.

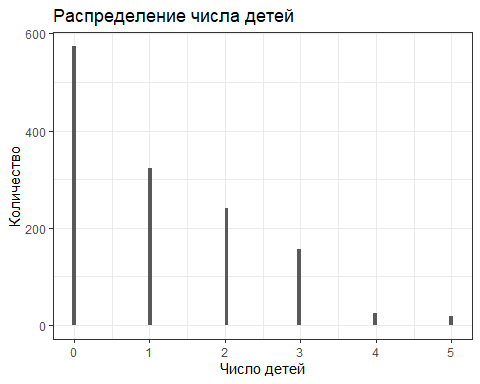
str(df)

## 'data.frame': 1338 obs. of 7 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ children: int 0 1 3 0 0 0 1 3 2 0 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ charges : num 16885 1726 4449 21984 3867 ...

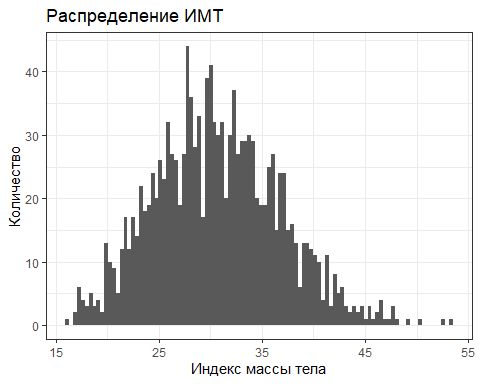
ggplot(data = df,   
 aes(x =age)) +  
 geom\_histogram(bins = 100) +theme\_bw()+labs(x = "Возраст",y = "Количество")+ggtitle('Распределение возраста')



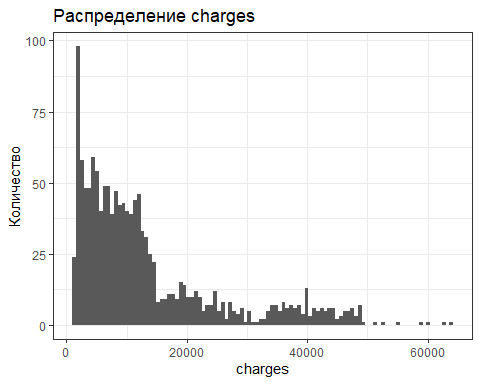
ggplot(data = df,   
 aes(x =children)) +  
 geom\_histogram(bins = 100) +theme\_bw()+labs(x = "Число детей",y = "Количество")+ggtitle('Распределение числа детей')



ggplot(data = df,   
 aes(x =bmi)) +  
 geom\_histogram(bins = 100) +theme\_bw()+labs(x = "Индекс массы тела",y = "Количество")+ggtitle('Распределение ИМТ')

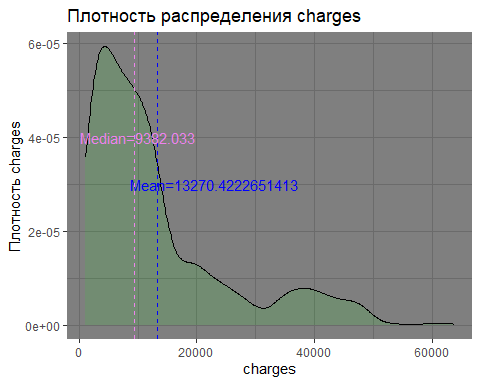


ggplot(data = df,   
 aes(x =charges)) +  
 geom\_histogram(bins = 100) +theme\_bw()+labs(x = "charges",y = "Количество")+ggtitle('Распределение charges')



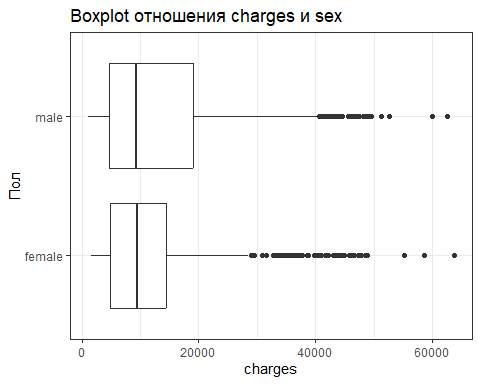
## 3. График плотности по колонке charges.

mean\_charges <- mean(df$charges)  
median\_charges <- median(df$charges)  
density\_plot <- ggplot(data = df,   
 aes(x =charges)) +  
 geom\_density(fill = "green", alpha = 0.1)+  
 labs(x = "charges",y = "Плотность charges")+  
 ggtitle("Плотность распределения charges")+  
 geom\_vline(aes(xintercept =mean\_charges),linetype = "dashed", colour = "blue") +  
 annotate("text",   
 x= 23000,   
 y=3e-5,   
 label=paste0("Mean=", mean\_charges), colour = "blue")+   
 geom\_vline(aes(xintercept =median\_charges), linetype = "dashed", colour = "violet") +  
 annotate("text",   
 x= 10000,   
 y=4e-5,   
 label=paste0("Median=", median\_charges), colour = "violet")+ theme\_dark()  
print(density\_plot)

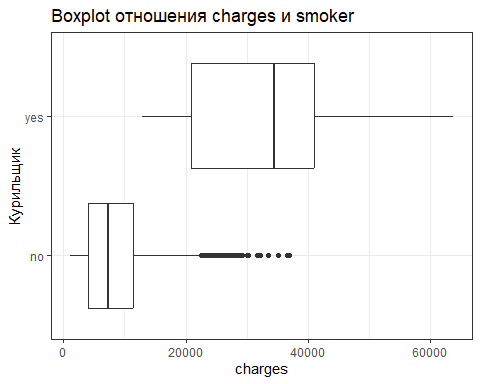


## 4. Boxplots

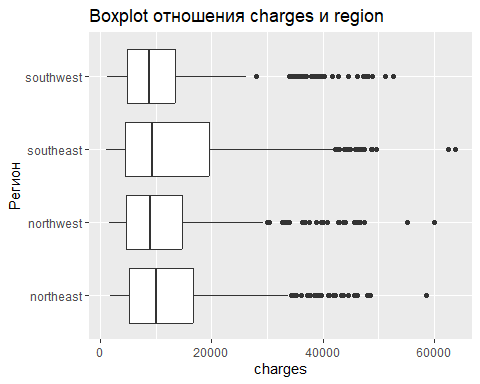
boxplot1 = ggplot(data = df,   
 aes(x = charges, y = sex)) +  
 geom\_boxplot()+ theme\_bw()+ggtitle("Boxplot отношения charges и sex")+  
 labs(x = "charges",y = "Пол")  
  
boxplot2 =ggplot(data = df,   
 aes(x = charges, y = smoker)) +  
 geom\_boxplot()+ theme\_bw()+ggtitle("Boxplot отношения charges и smoker")+  
 labs(x = "charges",y = "Курильщик")  
  
boxplot3 =ggplot(data = df,   
 aes(x = charges, y = region)) +  
 geom\_boxplot()+ theme\_gray()+ggtitle("Boxplot отношения charges и region")+  
labs(x = "charges",y = "Регион")  
print(boxplot1)



print(boxplot2)

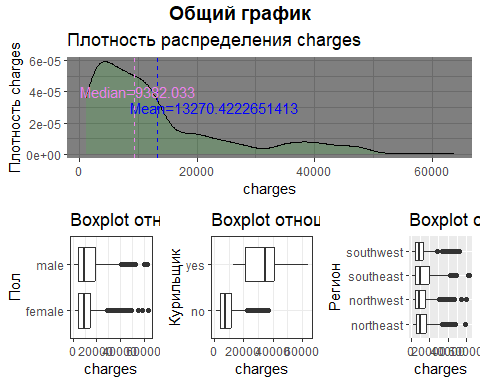


print(boxplot3)



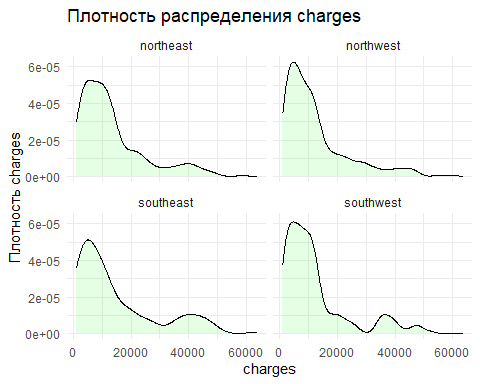
## 5. Merge plots

combine\_plot <- ggarrange(density\_plot , ggarrange(boxplot1,boxplot2, boxplot3, ncol = 3, nrow = 1),   
 ncol = 1, nrow = 2)+ggtitle("Общий график")  
#combine\_plot  
  
annotate\_figure(combine\_plot, top = text\_grob("Общий график", face = "bold", size = 14))



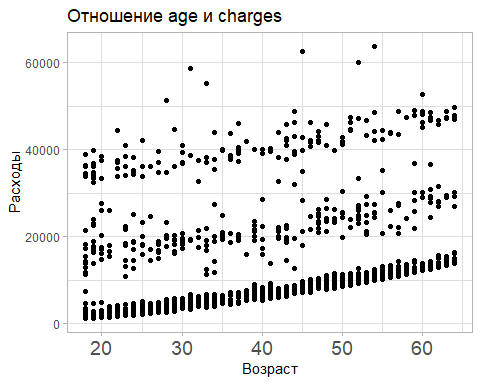
## 6. Фасет графика из задания 3 по колонке region.

ggplot(data = df,   
 aes(x =charges)) +  
 geom\_density(fill = "green", alpha = 0.1)+  
 labs(x = "charges",y = "Плотность charges")+  
 ggtitle("Плотность распределения charges")+  
 facet\_wrap(. ~ region, ncol = 2) +  
 theme\_minimal()



# 7. Scatter plot age and charges

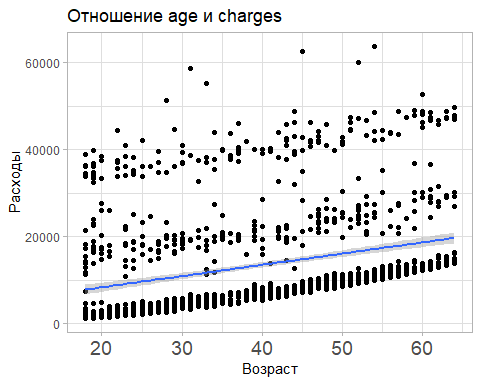
ggplot(data = df,   
 aes(x =age, y = charges))+  
 labs(x = "Возраст",y = "Расходы")+  
 geom\_point()+ggtitle("Отношение age и charges")+ theme\_light()+  
 theme( axis.text.x = element\_text(size=14))



## 8. Линия тренда для Scatter plot age and charges

ggplot(data = df,   
 aes(x =age, y = charges))+  
 labs(x = "Возраст",y = "Расходы")+  
 geom\_point()+ggtitle("Отношение age и charges")+ theme\_light()+  
 geom\_smooth(method=lm)+  
 theme( axis.text.x = element\_text(size=14))

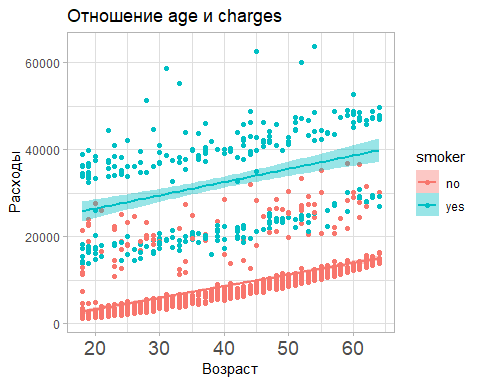
## `geom\_smooth()` using formula = 'y ~ x'



## 9. Разбивка по smokers

ggplot(data = df, aes(x =age, y = charges, color = smoker, fill = smoker))+  
 labs(x = "Возраст",y = "Расходы")+  
 geom\_point()+  
 ggtitle("Отношение age и charges")+ theme\_light()+  
 geom\_smooth(method=lm)+  
 theme( axis.text.x = element\_text(size=14))

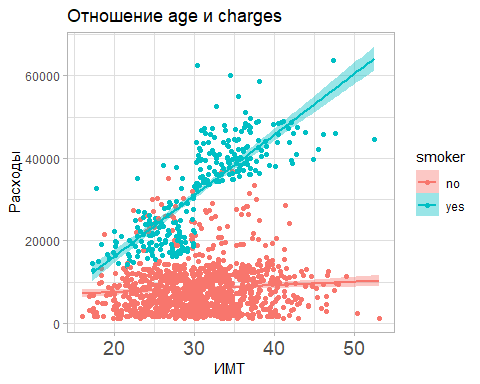
## `geom\_smooth()` using formula = 'y ~ x'



## 10. Scatter plot bmi and charges

ggplot(data = df, aes(x =bmi, y = charges, color = smoker, fill = smoker))+  
 labs(x = "ИМТ",y = "Расходы")+  
 geom\_point()+  
 ggtitle("Отношение age и charges")+ theme\_light()+  
 geom\_smooth(method=lm)+  
 theme( axis.text.x = element\_text(size=14))

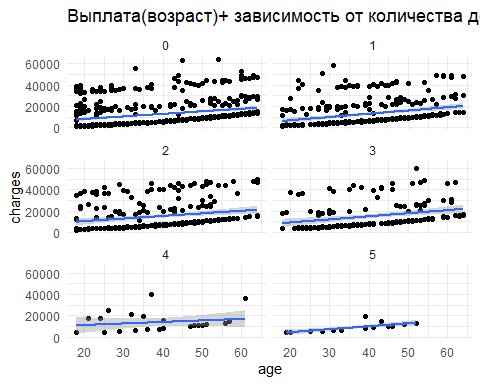
## `geom\_smooth()` using formula = 'y ~ x'



## 11. Построим график (как зависит выплата от региона) зависимости выплаты от возраста

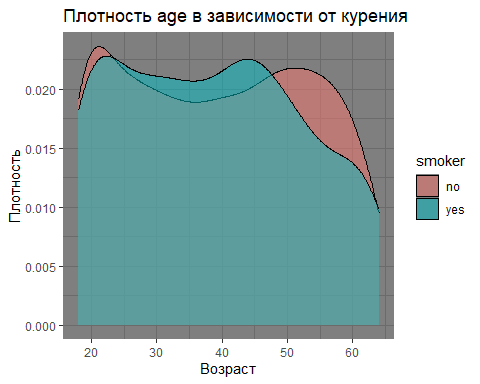
ggplot(data = df,   
 aes( x = age, y=charges)) +  
 geom\_point()+  
 labs(y = "charges", y = "Возраст")+  
 ggtitle("Выплата(возраст)+ зависимость от количества детей")+  
 geom\_smooth(method=lm)+  
 facet\_wrap(. ~ children, ncol = 2) +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



## 12. Плотность age в зависимости от курения

ggplot(data = df,   
 aes(x =age, fill = smoker)) +  
 geom\_density(alpha = 0.5)+  
 labs(x = "Возраст",y = "Плотность")+  
 ggtitle("Плотность age в зависимости от курения")+ theme\_dark()



## 14. Отношение индекса массы тела к логарифму трат по возрастным группам

new\_df <- subset(df, df$age>=21)  
  
new\_df$age\_group = as.factor(ifelse(new\_df$age>=21 & new\_df$age<=34, 'age: 21-34', ifelse(new\_df$age>=35 & new\_df$age<=49, 'age: 35-49', 'age: 50+')))  
  
ggplot(data = new\_df, aes(x =bmi, y = log(charges), color = age\_group))+  
 geom\_point(color= 'darkorchid4', alpha = 0.5)+  
 ggtitle("Отношение индекса массы тела к логарифму трат по возрастным группам")+  
 geom\_smooth(method=lm)+  
 facet\_grid(. ~ age\_group)+ theme\_minimal()+ theme(legend.position="bottom")

## `geom\_smooth()` using formula = 'y ~ x'

