novikovPrac7.R

Новиков Денис Владиславович ИНБО-05-22

2024-04-26

# Установка библиотек --------------------------------------------  
  
if("ggplot2" %in% rownames(installed.packages()) == FALSE) {  
 install.packages("ggplot2") }  
library(ggplot2)  
  
if("readxl" %in% rownames(installed.packages()) == FALSE) {  
 install.packages("readxl") }  
library(readxl)  
  
if("dplyr" %in% rownames(installed.packages()) == FALSE) {  
 install.packages("dplyr") }  
library(dplyr)

##   
## Присоединяю пакет: 'dplyr'

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

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

# Часть 1 -----------------------------------------------------------------  
  
# Задание 1 ---------------------------------------------------------------  
  
# 1 -----------------------------------------------------------------------  
  
demography = read.csv("https://raw.githubusercontent.com/allatambov/R-programming-3/master/seminars/sem8-09-02/demography.csv", sep=",")  
head(demography)

## region district  
## 1 Белгородская область Алексеевский муниципальный район и город Алексеевка  
## 2 Белгородская область Белгородский муниципальный район  
## 3 Белгородская область Борисовский муниципальный район  
## 4 Белгородская область Город Валуйки и Валуйский муниципальный район  
## 5 Белгородская область Вейделевский муниципальный район  
## 6 Белгородская область Волоконовский муниципальный район  
## empl\_total A\_agro B\_fish C\_fossil D\_proc E\_energy F\_build G\_merc H\_host  
## 1 19027 1824 NA NA 6194 474 242 1634 80  
## 2 19276 4807 NA NA 2132 1073 591 1828 166  
## 3 6558 1073 NA NA 1584 191 35 323 NA  
## 4 16222 1150 NA NA 5123 660 NA 345 50  
## 5 4030 1309 NA NA 19 210 35 111 20  
## 6 5450 1702 NA NA 413 320 NA 116 NA  
## I\_trans J\_fin K\_prop L\_gov M\_educ N\_heal O\_comm popul\_total urban\_total  
## 1 1809 149 817 1245 2390 1554 615 62338 38566  
## 2 335 NA 481 1238 3801 1557 931 116297 35900  
## 3 239 24 102 485 890 1303 280 25818 13743  
## 4 1580 126 389 1838 2357 1678 885 66655 41039  
## 5 NA 29 174 436 931 520 238 19604 6412  
## 6 228 42 144 556 1013 669 237 31130 15189  
## rural\_total wa\_total wa\_female wa\_male ret\_total ret\_female ret\_male  
## 1 23772 35735 16758 18977 15944 10966 4978  
## 2 80397 64714 30848 33866 31827 21714 10113  
## 3 12075 14103 6459 7644 7560 5203 2357  
## 4 25616 36651 16715 19936 19342 13215 6127  
## 5 13192 10560 4863 5697 5731 3933 1798  
## 6 15941 16747 7924 8823 9319 6539 2780  
## young\_total young\_female young\_male X18\_19 X20\_24 X25\_29 X30\_34 X35\_39 X40\_44  
## 1 10659 5151 5508 932 3501 5380 4849 4413 4378  
## 2 19756 9621 10135 1709 5856 9295 9654 8848 7895  
## 3 4155 1979 2176 385 1273 2016 1899 1743 1717  
## 4 10662 5069 5593 979 2943 4570 5203 5038 4950  
## 5 3313 1576 1737 350 668 1055 1189 1191 1474  
## 6 5064 2532 2532 594 1281 2095 2096 1994 2120  
## X45\_49 X50\_54 X55\_59 X60\_64 X65\_69 X70\_plus  
## 1 4047 4729 4910 3890 3026 6426  
## 2 6952 8356 9776 8890 6874 10602  
## 3 1670 1994 2164 1859 1478 3050  
## 4 4309 4957 5474 4759 3951 7703  
## 5 1536 1845 1823 1397 1149 2229  
## 6 2094 2567 2629 2118 1804 4048

str(demography)

## 'data.frame': 45 obs. of 42 variables:  
## $ region : chr "Белгородская область" "Белгородская область" "Белгородская область" "Белгородская область" ...  
## $ district : chr "Алексеевский муниципальный район и город Алексеевка" "Белгородский муниципальный район" "Борисовский муниципальный район" "Город Валуйки и Валуйский муниципальный район" ...  
## $ empl\_total : int 19027 19276 6558 16222 4030 5450 5532 4761 9547 2395 ...  
## $ A\_agro : int 1824 4807 1073 1150 1309 1702 1354 1629 1568 690 ...  
## $ B\_fish : logi NA NA NA NA NA NA ...  
## $ C\_fossil : int NA NA NA NA NA NA NA NA NA NA ...  
## $ D\_proc : int 6194 2132 1584 5123 19 413 295 558 3144 NA ...  
## $ E\_energy : int 474 1073 191 660 210 320 176 185 285 166 ...  
## $ F\_build : int 242 591 35 NA 35 NA NA NA 393 NA ...  
## $ G\_merc : int 1634 1828 323 345 111 116 64 53 135 NA ...  
## $ H\_host : int 80 166 NA 50 20 NA NA NA 21 NA ...  
## $ I\_trans : int 1809 335 239 1580 NA 228 37 22 404 NA ...  
## $ J\_fin : int 149 NA 24 126 29 42 31 27 30 12 ...  
## $ K\_prop : int 817 481 102 389 174 144 173 54 210 138 ...  
## $ L\_gov : int 1245 1238 485 1838 436 556 925 474 688 322 ...  
## $ M\_educ : int 2390 3801 890 2357 931 1013 1008 881 1300 480 ...  
## $ N\_heal : int 1554 1557 1303 1678 520 669 1187 588 986 385 ...  
## $ O\_comm : int 615 931 280 885 238 237 189 288 383 185 ...  
## $ popul\_total : int 62338 116297 25818 66655 19604 31130 29740 22299 39213 12237 ...  
## $ urban\_total : int 38566 35900 13743 41039 6412 15189 6506 7506 5890 NA ...  
## $ rural\_total : int 23772 80397 12075 25616 13192 15941 23234 14793 33323 12237 ...  
## $ wa\_total : int 35735 64714 14103 36651 10560 16747 16372 12174 20955 6418 ...  
## $ wa\_female : int 16758 30848 6459 16715 4863 7924 7507 5567 9751 2907 ...  
## $ wa\_male : int 18977 33866 7644 19936 5697 8823 8865 6607 11204 3511 ...  
## $ ret\_total : int 15944 31827 7560 19342 5731 9319 8351 6397 12107 3914 ...  
## $ ret\_female : int 10966 21714 5203 13215 3933 6539 5855 4515 8414 2733 ...  
## $ ret\_male : int 4978 10113 2357 6127 1798 2780 2496 1882 3693 1181 ...  
## $ young\_total : int 10659 19756 4155 10662 3313 5064 5017 3728 6151 1905 ...  
## $ young\_female: int 5151 9621 1979 5069 1576 2532 2360 1811 2984 922 ...  
## $ young\_male : int 5508 10135 2176 5593 1737 2532 2657 1917 3167 983 ...  
## $ X18\_19 : int 932 1709 385 979 350 594 475 393 632 229 ...  
## $ X20\_24 : int 3501 5856 1273 2943 668 1281 1381 869 1946 507 ...  
## $ X25\_29 : int 5380 9295 2016 4570 1055 2095 2468 1511 2894 813 ...  
## $ X30\_34 : int 4849 9654 1899 5203 1189 2096 2162 1552 2579 697 ...  
## $ X35\_39 : int 4413 8848 1743 5038 1191 1994 1961 1357 2414 719 ...  
## $ X40\_44 : int 4378 7895 1717 4950 1474 2120 2006 1494 2637 853 ...  
## $ X45\_49 : int 4047 6952 1670 4309 1536 2094 1977 1607 2622 888 ...  
## $ X50\_54 : int 4729 8356 1994 4957 1845 2567 2276 1958 3084 987 ...  
## $ X55\_59 : int 4910 9776 2164 5474 1823 2629 2459 2037 3197 996 ...  
## $ X60\_64 : int 3890 8890 1859 4759 1397 2118 1915 1569 2879 691 ...  
## $ X65\_69 : int 3026 6874 1478 3951 1149 1804 1581 1206 2278 570 ...  
## $ X70\_plus : int 6426 10602 3050 7703 2229 4048 3534 2569 5257 2160 ...

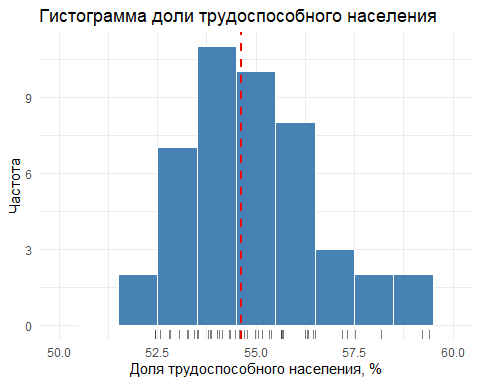
# 2 -----------------------------------------------------------------------  
  
demography$young\_share <- demography$young\_total / demography$popul\_total \* 100  
demography$trud\_share <- demography$wa\_total / demography$popul\_total \* 100  
demography$old\_share <- (demography$popul\_total - demography$young\_total - demography$wa\_total) / demography$popul\_total \* 100  
tail(demography[, c(2, (ncol(demography) - 2):ncol(demography))])

## district young\_share trud\_share old\_share  
## 40 Ульяновский муниципальный район 17.89889 52.57006 29.53105  
## 41 Ферзиковский муниципальный район 17.67294 56.45731 25.86975  
## 42 Хвастовичский муниципальный район 18.20532 54.35740 27.43727  
## 43 Юхновский муниципальный район 15.88188 52.80454 31.31359  
## 44 город Калуга 14.82087 57.52874 27.65039  
## 45 город Обнинск 17.00162 56.30657 26.69181

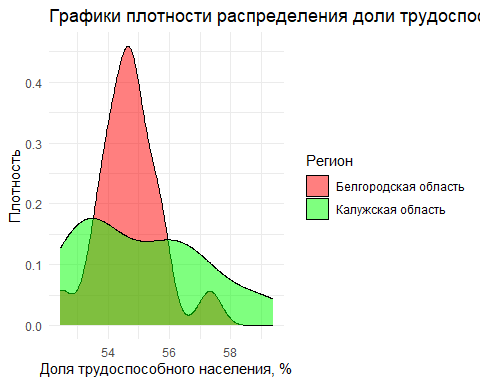
# 3 -----------------------------------------------------------------------  
  
median\_trud\_share <- median(demography$trud\_share)  
ggplot(demography, aes(x = trud\_share)) +  
 geom\_histogram(fill = "steelblue", binwidth = 1, color = "white") +  
 geom\_rug(alpha = 0.5, sides = "b") +  
 geom\_vline(xintercept = median\_trud\_share, color = "red", linetype = "dashed", size = 1) +  
 labs(x = "Доля трудоспособного населения, %", y = "Частота", title = "Гистограмма доли трудоспособного населения") +  
 scale\_x\_continuous(limits = c(50, 60)) +   
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

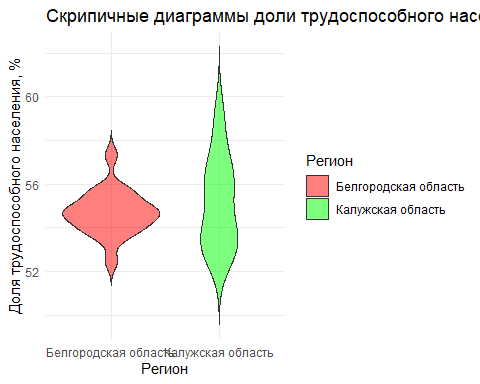
## Warning: Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_bar()`).



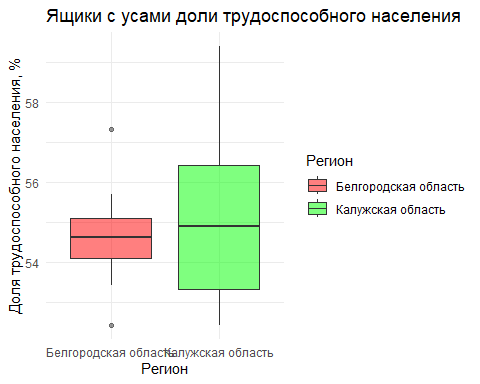
# 4 -----------------------------------------------------------------------  
  
ggplot(demography, aes(x = trud\_share, fill = region, alpha = region)) +  
 geom\_density(color = "black") +  
 scale\_fill\_manual(values = c("red", "green")) +   
 scale\_alpha\_manual(values = c(0.5, 0.5)) +  
 labs(x = "Доля трудоспособного населения, %",  
 y = "Плотность",  
 fill = "Регион",  
 alpha = "Регион",   
 title = "Графики плотности распределения доли трудоспособного населения") +  
 theme\_minimal()



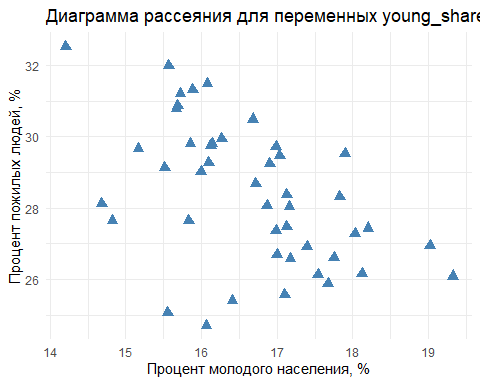
ggplot(demography, aes(x = region, y = trud\_share, fill = region, alpha = region)) +  
 geom\_violin(trim = FALSE) +  
 scale\_fill\_manual(values = c("red", "green")) +  
 scale\_alpha\_manual(values = c(0.5, 0.5)) +  
 labs(x = "Регион",  
 y = "Доля трудоспособного населения, %",  
 fill = "Регион",  
 alpha = "Регион",  
 title = "Скрипичные диаграммы доли трудоспособного населения") +  
 theme\_minimal()



ggplot(demography, aes(x = region, y = trud\_share, fill = region, alpha = region)) +  
 geom\_boxplot() +  
 scale\_fill\_manual(values = c("red", "green")) +  
 scale\_alpha\_manual(values = c(0.5, 0.5)) +  
 labs(x = "Регион",  
 y = "Доля трудоспособного населения, %",  
 fill = "Регион",  
 alpha = "Регион",  
 title = "Ящики с усами доли трудоспособного населения") +  
 theme\_minimal()



# 5 -----------------------------------------------------------------------  
  
ggplot(demography, aes(x = young\_share, y = old\_share)) +  
 geom\_point(color = "steelblue", shape = 17, size = 3) +  
 labs(x = "Процент молодого населения, %",  
 y = "Процент пожилых людей, %",  
 title = "Диаграмма рассеяния для переменных young\_share и old\_share") +  
 theme\_minimal()



cor.test(demography$young\_share,demography$old\_share, method = "spearman")

##   
## Spearman's rank correlation rho  
##   
## data: demography$young\_share and demography$old\_share  
## S = 23018, p-value = 0.0003424  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.5163373

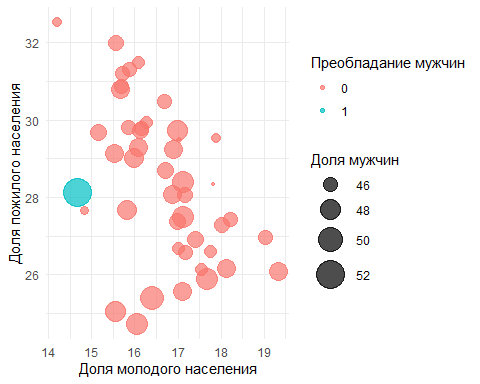
print("Заметна статистически значимая отрицательная корреляция")

## [1] "Заметна статистически значимая отрицательная корреляция"

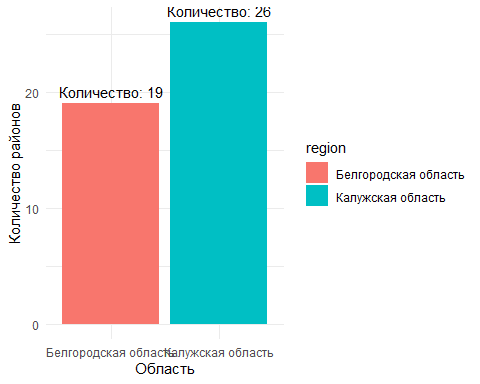
# 6 -----------------------------------------------------------------------  
  
demography$male\_share <- ((demography$ret\_male + demography$wa\_male + demography$young\_male) /   
 demography$popul\_total) \* 100  
demography$male <- ifelse(demography$male\_share > (100 - demography$male\_share), 1, 0)  
head(demography[, c(2, (ncol(demography) - 1):ncol(demography))])

## district male\_share male  
## 1 Алексеевский муниципальный район и город Алексеевка 47.26331 0  
## 2 Белгородский муниципальный район 46.53086 0  
## 3 Борисовский муниципальный район 47.16477 0  
## 4 Город Валуйки и Валуйский муниципальный район 47.49231 0  
## 5 Вейделевский муниципальный район 47.09243 0  
## 6 Волоконовский муниципальный район 45.40636 0

# 7 -----------------------------------------------------------------------  
  
ggplot(demography, aes(x = young\_share, y = old\_share, size = male\_share, color = factor(male))) +  
 geom\_point(alpha = 0.7) +  
 scale\_size\_continuous(range = c(1, 10), name = "Доля мужчин") +  
 labs(x = "Доля молодого населения",  
 y = "Доля пожилого населения",  
 color = "Преобладание мужчин") +  
 theme\_minimal()



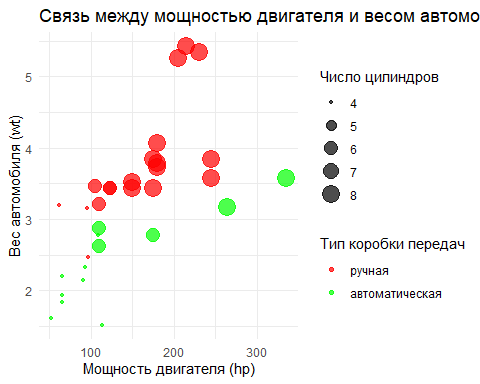
regions <- data.frame(table(demography$region))  
colnames(regions) <- c("region", "count")  
regions$label <- paste0("Количество: ", regions$count)  
  
ggplot(regions, aes(x = region, y = count, fill = region)) +  
 geom\_bar(stat = "identity") +  
 geom\_text(aes(label = label), vjust = -0.5, size = 4) +  
 labs(x = "Область", y = "Количество районов") +  
 theme\_minimal()



# Часть 2 -----------------------------------------------------------------  
  
# Задание 1 ---------------------------------------------------------------  
  
data = mtcars  
head(data)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

data$am\_factor <- factor(mtcars$am, levels = c(0, 1), labels = c("ручная", "автоматическая"))  
  
ggplot(data, aes(x = hp, y = wt, size = cyl, color = am\_factor)) +  
 geom\_point(alpha = 0.7) +  
 scale\_size\_continuous(name = "Число цилиндров") +  
 scale\_color\_manual(name = "Тип коробки передач", values = c("ручная" = "red", "автоматическая" = "green")) +  
 labs(x = "Мощность двигателя (hp)", y = "Вес автомобиля (wt)") +  
 ggtitle("Связь между мощностью двигателя и весом автомобиля") +  
 theme\_minimal()



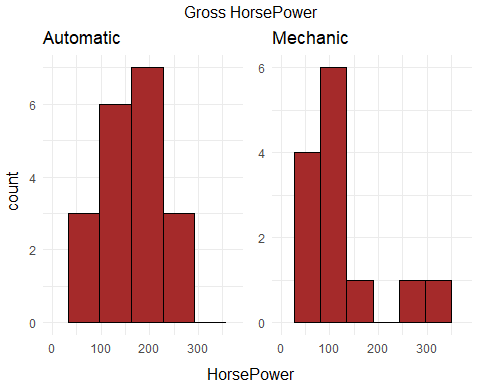
# Задание 2 ---------------------------------------------------------------  
  
if("gridExtra" %in% rownames(installed.packages()) == FALSE) {  
 install.packages("gridExtra") }  
library(gridExtra)

##   
## Присоединяю пакет: 'gridExtra'

## Следующий объект скрыт от 'package:dplyr':  
##   
## combine

p1 = ggplot(subset(data, am == 0), aes(x = hp)) +  
 geom\_histogram(binwidth = 65, fill = "brown", color = "black") +  
 scale\_x\_continuous(limits = c(0, 375)) +  
 labs(title = "Automatic") +   
 theme\_minimal() +   
 theme(axis.title.x = element\_blank(), axis.title.y = element\_blank())  
  
p2 = ggplot(subset(data, am == 1), aes(x = hp)) +  
 geom\_histogram(binwidth = 54, fill = "brown", color = "black") +  
 scale\_x\_continuous(limits = c(0, 375)) +  
 labs(title = "Mechanic") +   
 theme\_minimal() +  
 theme(axis.title.x = element\_blank(), axis.title.y = element\_blank())  
  
grid.arrange(p1, p2, ncol = 2, top = "Gross HorsePower",  
 left = "count", bottom = "HorsePower")

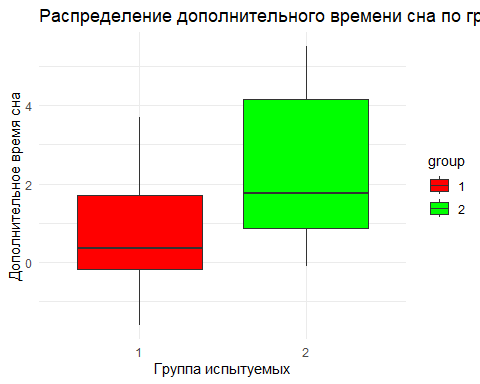
## Warning: Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_bar()`).  
## Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_bar()`).



# Задание 3 ---------------------------------------------------------------  
  
data("sleep")  
sleep

## extra group ID  
## 1 0.7 1 1  
## 2 -1.6 1 2  
## 3 -0.2 1 3  
## 4 -1.2 1 4  
## 5 -0.1 1 5  
## 6 3.4 1 6  
## 7 3.7 1 7  
## 8 0.8 1 8  
## 9 0.0 1 9  
## 10 2.0 1 10  
## 11 1.9 2 1  
## 12 0.8 2 2  
## 13 1.1 2 3  
## 14 0.1 2 4  
## 15 -0.1 2 5  
## 16 4.4 2 6  
## 17 5.5 2 7  
## 18 1.6 2 8  
## 19 4.6 2 9  
## 20 3.4 2 10

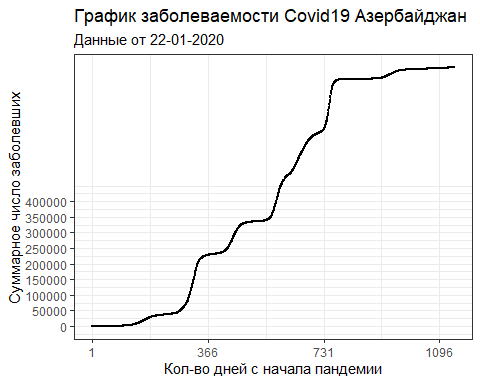
ggplot(sleep, aes(x = group, y = extra, fill = group)) +  
 geom\_boxplot() +  
 scale\_fill\_manual(values = c("red", "green")) +  
 labs(x = "Группа испытуемых", y = "Дополнительное время сна", title = "Распределение дополнительного времени сна по группам испытуемых") +  
 theme\_minimal()



# Часть 3 -----------------------------------------------------------------  
  
# Задание 1 ---------------------------------------------------------------  
  
covid <- read\_excel("covid.xlsx")  
View(covid)  
  
# Азербайджан -------------------------------------------------------------  
  
ggplot(data=covid, mapping = aes(x=1:nrow(covid), y=covid[[20]])) +  
 geom\_line(lwd=I(0.5), lty=1) +  
 geom\_point(cex=0.5) +  
 coord\_cartesian(xlim=c(1,nrow(covid)), ylim=range(covid[[20]])) +  
 labs(title="График заболеваемости Covid19 Азербайджан",  
 subtitle = "Данные от 22-01-2020",  
 x = "Кол-во дней с начала пандемии",  
 y = "Суммарное число заболевших") +  
 scale\_x\_continuous(breaks = seq(1,nrow(covid),365)) +  
 scale\_y\_continuous(breaks = seq(0,400000,50000)) +  
 theme\_bw()

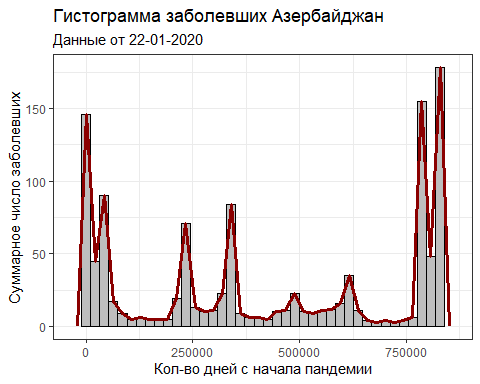
## Warning: Use of `covid[[20]]` is discouraged.  
## ℹ Use `.data[[20]]` instead.

## Warning: Use of `covid[[20]]` is discouraged.  
## ℹ Use `.data[[20]]` instead.



ggplot(data=covid, mapping = aes(covid[[20]])) +  
 geom\_histogram(bins=40, color="black", fill="grey") +  
 geom\_freqpoly(bins=40, color="red4", lwd=I(1.1)) +  
 labs(title="Гистограмма заболевших Азербайджан",  
 subtitle = "Данные от 22-01-2020",  
 x = "Кол-во дней с начала пандемии",  
 y = "Суммарное число заболевших") +  
 theme\_bw()

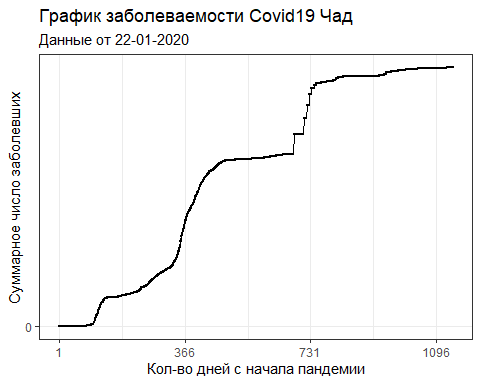
## Warning: Use of `covid[[20]]` is discouraged.  
## ℹ Use `.data[[20]]` instead.  
## Use of `covid[[20]]` is discouraged.  
## ℹ Use `.data[[20]]` instead.



# Чад ----------------------------------------------------------------  
  
ggplot(data=covid, mapping = aes(x=1:nrow(covid), y=covid[[59]])) +  
 geom\_line(lwd=I(0.5), lty=1) +  
 geom\_point(cex=0.5) +  
 coord\_cartesian(xlim=c(1,nrow(covid)), ylim=range(covid[[59]])) +  
 labs(title="График заболеваемости Covid19 Чад",  
 subtitle = "Данные от 22-01-2020",  
 x = "Кол-во дней с начала пандемии",  
 y = "Суммарное число заболевших") +  
 scale\_x\_continuous(breaks = seq(1,nrow(covid),365)) +  
 scale\_y\_continuous(breaks = seq(0,400000,50000)) +  
 theme\_bw()

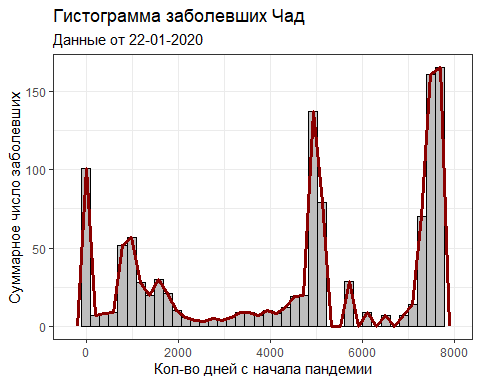
## Warning: Use of `covid[[59]]` is discouraged.  
## ℹ Use `.data[[59]]` instead.

## Warning: Use of `covid[[59]]` is discouraged.  
## ℹ Use `.data[[59]]` instead.



ggplot(data=covid, mapping = aes(covid[[59]])) +  
 geom\_histogram(bins=40, color="black", fill="grey") +  
 geom\_freqpoly(bins=40, color="red4", lwd=I(1.1)) +  
 labs(title="Гистограмма заболевших Чад",  
 subtitle = "Данные от 22-01-2020",  
 x = "Кол-во дней с начала пандемии",  
 y = "Суммарное число заболевших") +  
 theme\_bw()

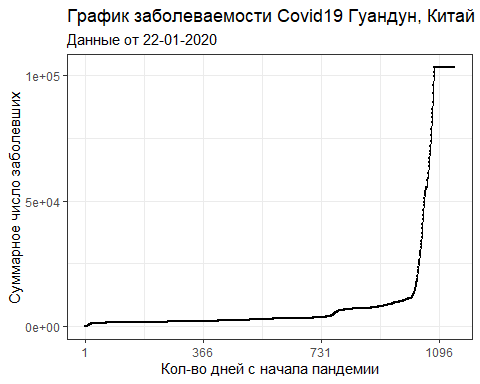
## Warning: Use of `covid[[59]]` is discouraged.  
## ℹ Use `.data[[59]]` instead.  
## Use of `covid[[59]]` is discouraged.  
## ℹ Use `.data[[59]]` instead.



# Гуандун, Китай ----------------------------------------------------------------  
  
ggplot(data=covid, mapping = aes(x=1:nrow(covid), y=covid[[66]])) +  
 geom\_line(lwd=I(0.5), lty=1) +  
 geom\_point(cex=0.5) +  
 coord\_cartesian(xlim=c(1,nrow(covid)), ylim=range(covid[[66]])) +  
 labs(title="График заболеваемости Covid19 Гуандун, Китай",  
 subtitle = "Данные от 22-01-2020",  
 x = "Кол-во дней с начала пандемии",  
 y = "Суммарное число заболевших") +  
 scale\_x\_continuous(breaks = seq(1,nrow(covid),365)) +  
 scale\_y\_continuous(breaks = seq(0,400000,50000)) +  
 theme\_bw()

## Warning: Use of `covid[[66]]` is discouraged.  
## ℹ Use `.data[[66]]` instead.

## Warning: Use of `covid[[66]]` is discouraged.  
## ℹ Use `.data[[66]]` instead.



ggplot(data=covid, mapping = aes(covid[[66]])) +  
 geom\_histogram(bins=40, color="black", fill="grey") +  
 geom\_freqpoly(bins=40, color="red4", lwd=I(1.1)) +  
 labs(title="Гистограмма заболевших Гуандун, Китай",  
 subtitle = "Данные от 22-01-2020",  
 x = "Кол-во дней с начала пандемии",  
 y = "Суммарное число заболевших") +  
 theme\_bw()

## Warning: Use of `covid[[66]]` is discouraged.  
## ℹ Use `.data[[66]]` instead.  
## Use of `covid[[66]]` is discouraged.  
## ℹ Use `.data[[66]]` instead.

