Методы машинного обучения

Рубежный контроль №1

Пряхин Владимир Геннадьевич
ИУ5-24М

Вариант №9

Задача №9

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения "хвостом распределения".

Возьмем датасет - https://www.kaggle.com/mmattson/who-national-life-expectancy

Функция на языке R для вычисления процента прощенных значений в колонках набора данных.

```
infoNA <- function(datas){
  for(c in names(datas)) {
    sumNa <- sum(is.na(datas[c]))
    col <- sum(!is.na(datas[c]))
    if(sumNa>0) {
      propusk <- round(sumNa / NROW(datas[c]) * 100, 3)
      print(paste(c,' - Пропущено',sumNa,'значений, это',propusk,'% от набора данных'))
    }
    else print(paste(c ,' - Full column data - ',col,'не нулевых значений'))
  }
}
```

Загрузим датасет

```
datas <- read.csv("/home/hino/life.csv")
head(datas)
```

```
> head(datas)
  country country_code region year life_expect life_exp60 adult_mortality infant_mort age1.4mort alcohol  bmi age5.19thinness age5.19obesity hepatitis
1  Angola          AGO Africa 2000    47.33730   14.73400        383.5583    0.137985   0.025695 1.47439 21.7           11.0           0.5        NA
2  Angola          AGO Africa 2001    48.19789   14.95963        372.3876    0.133675   0.024500 1.94025 21.8           10.9           0.5        NA
3  Angola          AGO Africa 2002    49.42569   15.20010        354.5147    0.128320   0.023260 2.07512 21.9           10.7           0.6        NA
4  Angola          AGO Africa 2003    50.50266   15.39144        343.2169    0.122040   0.021925 2.20275 22.0           10.5           0.7        NA
5  Angola          AGO Africa 2004    51.52863   15.56860        333.8711    0.115700   0.020545 2.41274 22.2           10.3           0.8        NA
6  Angola          AGO Africa 2005    52.72512   15.75107        322.7077    0.109205   0.018945 3.48640 22.3           10.2           0.8        NA
  measles polio diphtheria basic_water doctors hospitals gni_capita  gghe.d che_gdp  une_pop une_infant une_life une_hiv une_gni une_poverty
1      32    21         31    41.14431      NA        NA       2190 1.11099 1.90860 16395.47      122.2   46.522     1.0    2530        32.3
2      60    28         42    42.25467      NA        NA       2290 2.04631 4.48352 16945.75      118.9   47.059     1.1    2630          NA
3      59    22         47    43.37680      NA        NA       2690 1.30863 3.32946 17519.42      115.1   47.702     1.2    3180          NA
4      44    21         46    44.36387      NA        NA       2820 1.46560 3.54797 18121.48      110.8   48.440     1.3    3260          NA
5      43    18         47    45.35134   0.621        NA       3080 1.68663 3.96720 18758.15      106.2   49.263     1.3    3560          NA
6      21    14         47    46.33602      NA        NA       3570 1.27876 2.85220 19433.60      101.3   50.165     1.4    4060          NA
  une_edu_spend une_literacy une_school
1       2.60753           NA         NA
2            NA     67.40542         NA
3            NA           NA         NA
4            NA           NA         NA
5            NA           NA         NA
6       2.12011           NA         NA
```

summary(datas) #описательная статистика

```
> summary(datas)
             country      country_code                  region         year       life_expect      life_exp60    adult_mortality  infant_mort
 Afghanistan     :  17   AFG    :  17   Africa              :799   Min.   :2000   Min.   :36.23   Min.   :10.73   Min.   : 49.2   Min.   :0.001470
 Albania         :  17   AGO    :  17   Americas            :561   1st Qu.:2004   1st Qu.:63.20   1st Qu.:16.62   1st Qu.:108.3   1st Qu.:0.008255
 Algeria         :  17   ALB    :  17   Eastern Mediterranean:357   Median :2008   Median :71.60   Median :18.51   Median :164.8   Median :0.019995
 Angola          :  17   ARE    :  17   Europe              :850   Mean   :2008   Mean   :69.15   Mean   :18.91   Mean   :193.5   Mean   :0.032496
 Antigua and Barbuda:  17   ARG    :  17   South-East Asia     :187   3rd Qu.:2012   3rd Qu.:75.54   3rd Qu.:21.10   3rd Qu.:250.8   3rd Qu.:0.051720
 Argentina       :  17   ARM    :  17   Western Pacific     :357   Max.   :2016   Max.   :84.17   Max.   :26.39   Max.   :696.9   Max.   :0.164515
 (Other)         :3009   (Other):3009
   age1.4mort        alcohol            bmi         age5.19thinness  age5.19obesity     hepatitis         measles          polio          diphtheria
 Min.   :0.000065   Min.   : 0.000   Min.   :19.80   Min.   : 0.100   Min.   : 0.100   Min.   : 2.00   Min.   :16.00   Min.   : 8.00   Min.   :19.00
 1st Qu.:0.000355   1st Qu.: 1.198   1st Qu.:23.30   1st Qu.: 1.800   1st Qu.: 2.000   1st Qu.:81.00   1st Qu.:79.00   1st Qu.:81.00   1st Qu.:82.00
 Median :0.000895   Median : 3.994   Median :25.50   Median : 3.800   Median : 5.200   Median :92.00   Median :92.00   Median :93.00   Median :93.00
 Mean   :0.003489   Mean   : 4.835   Mean   :25.05   Mean   : 5.312   Mean   : 5.972   Mean   :85.44   Mean   :85.54   Mean   :86.61   Mean   :86.42
 3rd Qu.:0.004877   3rd Qu.: 7.723   3rd Qu.:26.50   3rd Qu.: 7.800   3rd Qu.: 8.900   3rd Qu.:97.00   3rd Qu.:96.00   3rd Qu.:97.00   3rd Qu.:97.00
 Max.   :0.039095   Max.   :20.182   Max.   :32.20   Max.   :28.100   Max.   :26.700   Max.   :99.00   Max.   :99.00   Max.   :99.00   Max.   :99.00
                    NA's   :50       NA's   :34      NA's   :34       NA's   :34       NA's   :569     NA's   :19      NA's   :19      NA's   :19
  basic_water       doctors         hospitals         gni_capita        gghe.d          che_gdp          une_pop          une_infant
 Min.   : 18.70   Min.   : 0.128   Min.   : 0.0000   Min.   :   250   Min.   : 0.06236   Min.   : 1.025   Min.   :     76   Min.   :  1.60
 1st Qu.: 71.66   1st Qu.: 6.391   1st Qu.: 0.5352   1st Qu.:  2540   1st Qu.: 1.53344   1st Qu.: 4.239   1st Qu.:   2195   1st Qu.:  8.00
 Median : 91.99   Median :20.523   Median : 1.0727   Median :  7460   Median : 2.60130   Median : 5.758   Median :   8544   Median : 19.50
 Mean   : 83.33   Mean   :19.866   Mean   : 2.0449   Mean   : 13397   Mean   : 3.12293   Mean   : 6.110   Mean   :  37076   Mean   : 30.49
 3rd Qu.: 98.55   3rd Qu.:30.982   3rd Qu.: 2.1048   3rd Qu.: 18250   3rd Qu.: 4.27811   3rd Qu.: 7.850   3rd Qu.:  25096   3rd Qu.: 48.05
 Max.   :100.00   Max.   :79.541   Max.   :56.4470   Max.   :123860   Max.   :12.06273   Max.   :20.413   Max.   :1414049   Max.   :142.40
 NA's   :32       NA's   :1331     NA's   :2981      NA's   :682      NA's   :100        NA's   :117      NA's   :37
   une_life         une_hiv          une_gni        une_poverty     une_edu_spend     une_literacy      une_school
 Min.   :39.44   Min.   : 0.100   Min.   :   420   Min.   : 0.10   Min.   : 0.7874   Min.   : 14.38   Min.   : 0.5593
 1st Qu.:62.84   1st Qu.: 0.100   1st Qu.:  2970   1st Qu.: 0.60   1st Qu.: 3.2628   1st Qu.: 72.70   1st Qu.: 7.7359
 Median :71.41   Median : 0.400   Median :  8340   Median : 3.10   Median : 4.4254   Median : 90.95   Median :10.2704
 Mean   :68.96   Mean   : 2.038   Mean   : 14965   Mean   :10.85   Mean   : 4.5329   Mean   : 81.98   Mean   : 9.7122
 3rd Qu.:75.57   3rd Qu.: 1.500   3rd Qu.: 20482   3rd Qu.:12.40   3rd Qu.: 5.4950   3rd Qu.: 95.79   3rd Qu.:12.0706
 Max.   :83.98   Max.   :28.200   Max.   :122670   Max.   :94.10   Max.   :14.0591   Max.   :100.00   Max.   :14.3788
                 NA's   :741      NA's   :117      NA's   :2198    NA's   :1286      NA's   :2540     NA's   :2306
```

infoNA(datas)

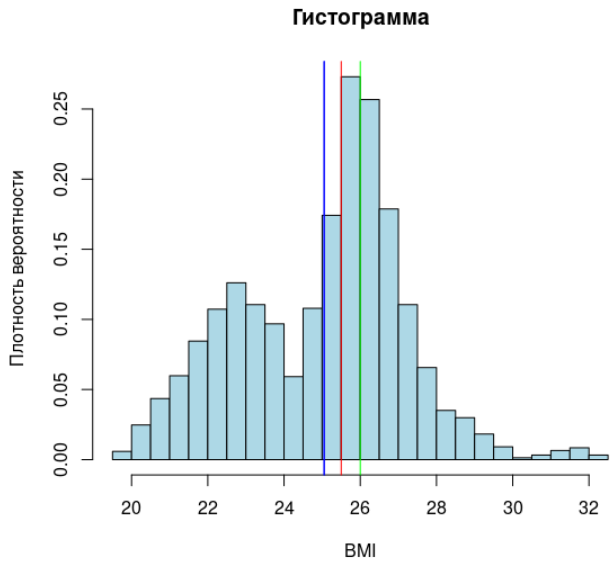
```
> infoNA(datas)
[1] "country  - Full column data -  3111 не нулевых значений"
[1] "country_code  - Full column data -  3111 не нулевых значений"
[1] "region  - Full column data -  3111 не нулевых значений"
[1] "year  - Full column data -  3111 не нулевых значений"
[1] "life_expect  - Full column data -  3111 не нулевых значений"
[1] "life_exp60  - Full column data -  3111 не нулевых значений"
[1] "adult_mortality  - Full column data -  3111 не нулевых значений"
[1] "infant_mort  - Full column data -  3111 не нулевых значений"
[1] "age1.4mort  - Full column data -  3111 не нулевых значений"
[1] "alcohol  - Пропущено 50 значений, это 1.607 % от набора данных"
[1] "bmi  - Full column data -  3111 не нулевых значений"
[1] "age5.19thinness  - Пропущено 34 значений, это 1.093 % от набора данных"
[1] "age5.19obesity  - Пропущено 34 значений, это 1.093 % от набора данных"
[1] "hepatitis  - Пропущено 569 значений, это 18.29 % от набора данных"
[1] "measles  - Пропущено 19 значений, это 0.611 % от набора данных"
[1] "polio  - Пропущено 19 значений, это 0.611 % от набора данных"
[1] "diphtheria  - Пропущено 19 значений, это 0.611 % от набора данных"
[1] "basic_water  - Пропущено 32 значений, это 1.029 % от набора данных"
[1] "doctors  - Пропущено 1331 значений, это 42.784 % от набора данных"
[1] "hospitals  - Пропущено 2981 значений, это 95.821 % от набора данных"
[1] "gni_capita  - Пропущено 682 значений, это 21.922 % от набора данных"
[1] "gghe.d  - Пропущено 100 значений, это 3.214 % от набора данных"
[1] "che_gdp  - Пропущено 117 значений, это 3.761 % от набора данных"
[1] "une_pop  - Пропущено 37 значений, это 1.189 % от набора данных"
[1] "une_infant  - Full column data -  3111 не нулевых значений"
[1] "une_life  - Full column data -  3111 не нулевых значений"
[1] "une_hiv  - Пропущено 741 значений, это 23.819 % от набора данных"
[1] "une_gni  - Пропущено 117 значений, это 3.761 % от набора данных"
[1] "une_poverty  - Пропущено 2198 значений, это 70.653 % от набора данных"
[1] "une_edu_spend  - Пропущено 1286 значений, это 41.337 % от набора данных"
[1] "une_literacy  - Пропущено 2540 значений, это 81.646 % от набора данных"
[1] "une_school  - Пропущено 2306 значений, это 74.124 % от набора данных"
```

Проанализируем данные на наличие пропусков. Выберем параметр Индекса Массы Тела (BMI), как подлежащий заполнению. Для этого параметра количество пропущенных значений не превышает 5% - 1,093%.

Отобразим гистограмму и покажем на ней среднее, медиану и моду.

```
x <- datas$bmi  #анализируемое значение (BMI).
hist(x, breaks = 20, freq = FALSE, col = "lightblue",
    xlab = "BMI",
    ylab = "Плотность вероятности",
    main = "Гистограмма")
abline(v=mean(x,na.rm = TRUE),col='blue')
abline(v=median(x,na.rm = TRUE),col='red')
```

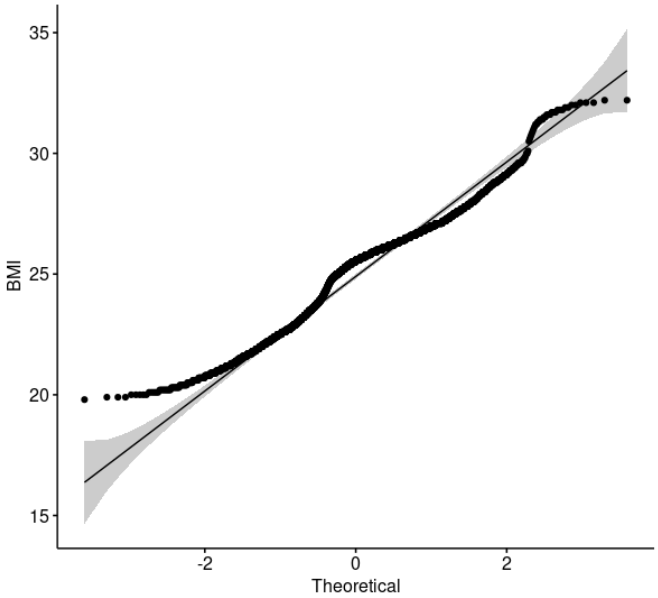
```
library(modeest)
abline(v=mfv(x,na_rm = TRUE),col='green')
```



Среднее, медиана и мода не совпадают, что позволяет считать данное распределение асимметричным.

Воспользуемся графиком квантиль-квантиль, который показывает корреляцию между выбранным нами параметром (BMI) и нормальным распределением.

```
library(ggpubr)
ggqqplot(datas$bmi, ylab = "BMI")
```



Визуальная проверка нормальности данных с помощью QQ графика также показала, что речь идет об асимметричном распределении.

Заполним пропущенные данные с помощью "хвоста распределения" выбрав формулу для ассиметричного распределения.

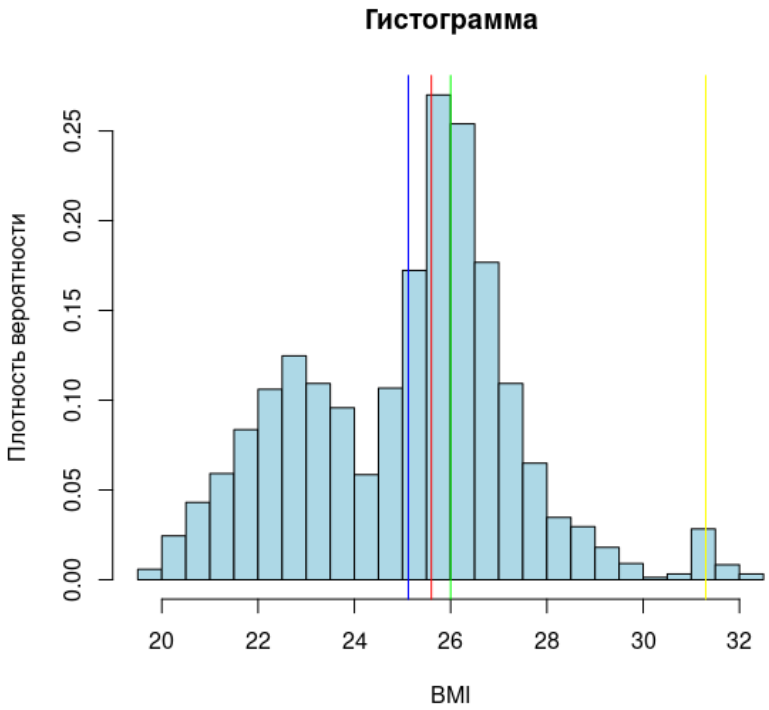
$$IQR = Q3 - Q1$$

$$extreme\_value = Q3 + K \cdot IQR$$

```
q3 <- quantile(x, probs=c(3/4), names = FALSE, na.rm = TRUE)
EX <- q3 + 1.5 * IQR(x,na.rm = TRUE)
datas$bmi[is.na(datas$bmi)] <- EX #заполним недостающие данные в датасете
x <- datas$bmi
```

Покажем на гистограмме полученное нами значение

```
hist(x, breaks = 20, freq = FALSE, col = "lightblue",
    xlab = "BMI",
    ylab = "Плотность вероятности",
    main = "Гистограмма")
abline(v=mean(x,na.rm = TRUE),col='blue')
abline(v=median(x,na.rm = TRUE),col='red')
abline(v=mfv(x,na_rm = TRUE),col='green')
abline(v=EX,col='yellow')
```



Подсчет по формуле для нормального распределения даёт близкое значение.

```
mean(x,na.rm = TRUE) + 3 * sd(x,na.rm = TRUE)
EX
```

```
> mean(x,na.rm = TRUE) + 3 * sd(x,na.rm = TRUE)
[1] 31.63305
> EX
[1] 31.3
```

Задача №29

Для набора данных проведите удаление константных и псевдоконстантных признаков.

Создадим набор данных с константной переменной

```
emp.data <- data.frame(
  emp_id = c (1:5),
  emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
  salary = c(623.3,515.2,611.0,729.0,843.25),
  start_date = as.Date(c("2020-01-01", "2020-09-23", "2020-11-15", "2020-05-11",
                         "2020-03-27")),
  constants = c(1,1.1,1,0.9,1),
  stringsAsFactors = FALSE
)
```

```
  emp_id emp_name constants salary start_date
1      1     Rick       1.0 623.30 2020-01-01
2      2      Dan       1.1 515.20 2020-09-23
3      3 Michelle       1.0 611.00 2020-11-15
4      4     Ryan       0.9 729.00 2020-05-11
5      5     Gary       1.0 843.25 2020-03-27
```

Вычислим дисперсию для параметра «constants»

```
var(emp.data$constants)
```

```
> var(emp.data$constants)
[1] 0.005
```

Удалим столбец из набора данных содержащий константную переменную.

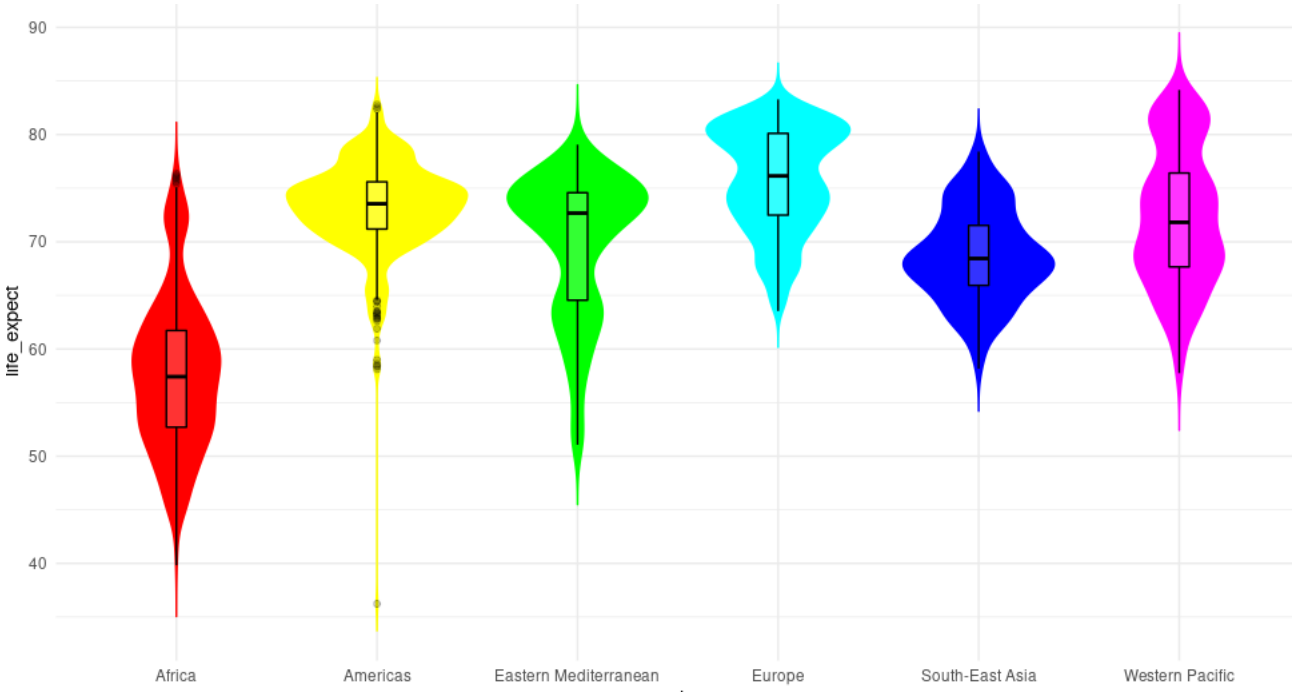
```
  emp_id emp_name salary start_date
1      1     Rick 623.30 2020-01-01
2      2      Dan 515.20 2020-09-23
3      3 Michelle 611.00 2020-11-15
4      4     Ryan 729.00 2020-05-11
5      5     Gary 843.25 2020-03-27
```

Дополнительная задача.

Построить график "Скрипичная диаграмма (violin plot)".

Построим график, показывающий зависимость средней продолжительности жизни от региона.

```
library(ggplot2)
ggplot(datas, aes(x=region, y=life_expect)) +
geom_violin(trim=FALSE, fill=rainbow(3072),col =rainbow(3072)) +
  geom_boxplot(width=0.1, color="black", alpha=0.2) + theme_minimal()
```


90
80
70
60
50
40
life_expect
Africa
Americas
Eastern Mediterranean
Europe
South-East Asia
Western Pacific