گزارش پروژه اول

دانیال سعیدی

بخش اول

کد های این بخش در فایل part1.r نوشته شده است.

نتیجه اجرای برنامه:

امتحان اول ویژگی های زیر را دارد:

- Max: 99
- Median: 77.5
- Mean: 79.25
- Max: 99
- Variance: 102.3026

امتحان دوم ویژگی های زیر را دارد:

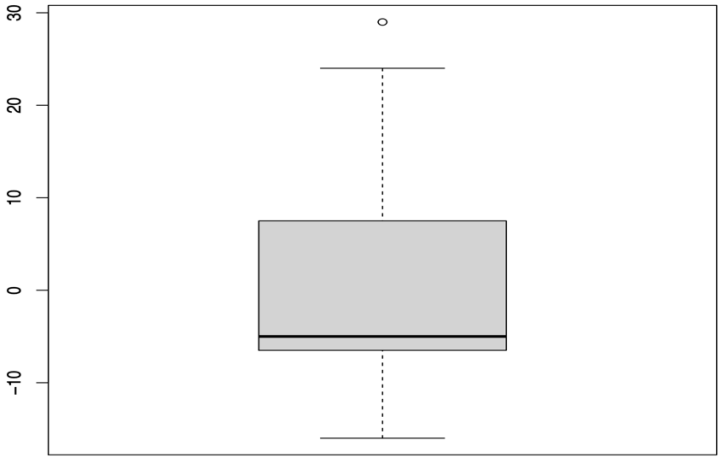
- Max: 100
- Median: 75
- Mean: 80.55
- Max: 100
- Variance: 160.7868

یک وکتور به نام diffVector تعریف کردم که اختلاف نمرات افراد را ذخیره می کند.

مقدار چارک های diffVector با استفاده از تابع quantile به دست می آید:

0%	25%	50%	75%	100%
-16.00	-6.25	-5.00	6.75	29.00

نمودار جعبه diffVecotr



بخش دوم

کد های این بخش در فایل part1.r نوشته شده است.

نتیجه اجرای برنامه:

```
During startup - Warning messages:
1: Setting LC_COLLATE failed, using "C"
2: Setting LC_TIME failed, using "C"
3: Setting LC_MESSAGES failed, using "C"
4: Setting LC_MONETARY failed, using "C"
 [1] "age"           "workclass"      "education"      "martial.status"
 [5] "occupation"    "relationship"   "race"           "sex"
 [9] "nativecountry"  "income"
   age         workclass  education         martial.status         occupation
1   39         State-gov  Bachelors          Never-married       Adm-clerical
2   50  Self-emp-not-inc  Bachelors     Married-civ-spouse    Exec-managerial
3   38           Private    HS-grad               Divorced  Handlers-cleaners
4   53           Private       11th     Married-civ-spouse  Handlers-cleaners
5   28           Private  Bachelors     Married-civ-spouse     Prof-specialty
6   37           Private    Masters     Married-civ-spouse    Exec-managerial
7   49           Private        9th  Married-spouse-absent      Other-service
8   52  Self-emp-not-inc    HS-grad     Married-civ-spouse    Exec-managerial
9   31           Private    Masters          Never-married     Prof-specialty
10  42           Private  Bachelors     Married-civ-spouse    Exec-managerial
     relationship   race     sex  nativecountry income
1   Not-in-family  White    Male  United-States  <=50K
2         Husband  White    Male  United-States  <=50K
3   Not-in-family  White    Male  United-States  <=50K
4         Husband  Black    Male  United-States  <=50K
5            Wife  Black  Female           Cuba  <=50K
6            Wife  White  Female  United-States  <=50K
7   Not-in-family  Black  Female        Jamaica  <=50K
8         Husband  White    Male  United-States   >50K
9   Not-in-family  White  Female  United-States   >50K
10        Husband  White    Male  United-States   >50K

 Amer-Indian-Eskimo  Asian-Pac-Islander              Black              Other
                311                1039               3124                271
              White
              27816
Warning message:
In yinch(0.1) : y log scale:  yinch() is nonsense
(base) daniel@Daniels-MacBook-Pro CA1 %
```

مطابق زیر فایل outcome.csv را می خوانیم و در یک dataframe ذخیره می کنیم.

```
elements <- read.csv("./outcome.csv", head = TRUE,",")

df <- data.frame(elements)

df[df==""] <- NA
```

لیست نام ستون ها با تابع colnames و به کمک تابع head ۱۰ تا row اول رو را چاپ کردم.

```
colnames(df)

##  [1] "age"            "workclass"      "education"      "martial.status"
##  [5] "occupation"     "relationship"   "race"           "sex"
##  [9] "nativecountry"  "income"

head(df,10)

##    age         workclass  education         martial.status
occupation
## 1   39         State-gov  Bachelors          Never-married       Adm-
clerical
## 2   50  Self-emp-not-inc  Bachelors     Married-civ-spouse    Exec-
managerial
## 3   38           Private    HS-grad               Divorced  Handlers-
cleaners
## 4   53           Private       11th     Married-civ-spouse  Handlers-
cleaners
## 5   28           Private  Bachelors     Married-civ-spouse     Prof-
specialty
## 6   37           Private    Masters     Married-civ-spouse    Exec-
managerial
## 7   49           Private        9th  Married-spouse-absent      Other-
service
## 8   52  Self-emp-not-inc    HS-grad     Married-civ-spouse    Exec-
managerial
## 9   31           Private    Masters          Never-married     Prof-
specialty
## 10  42           Private  Bachelors     Married-civ-spouse    Exec-
managerial
##      relationship   race     sex  nativecountry income
## 1   Not-in-family  White    Male  United-States  <=50K
## 2         Husband  White    Male  United-States  <=50K
## 3   Not-in-family  White    Male  United-States  <=50K
## 4         Husband  Black    Male  United-States  <=50K
## 5            Wife  Black  Female           Cuba  <=50K
## 6            Wife  White  Female  United-States  <=50K
## 7   Not-in-family  Black  Female        Jamaica  <=50K
## 8         Husband  White    Male  United-States   >50K
```

```
## 9   Not-in-family  White  Female  United-States   >50K
## 10        Husband  White    Male  United-States   >50K
```

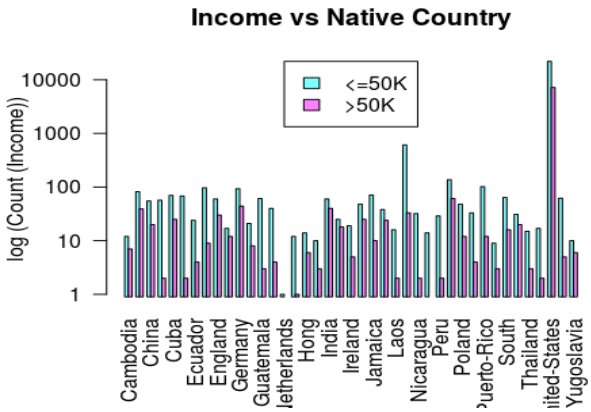
به کمک تابع table مقدار تکرار متغیر های race را چاپ کردم:

```
table(df$race)

##
##  Amer-Indian-Eskimo  Asian-Pac-Islander               Black
Other
##                 311                1039                3124
271
##               White
##               27816
```

رسم نمودار grouped bar plot

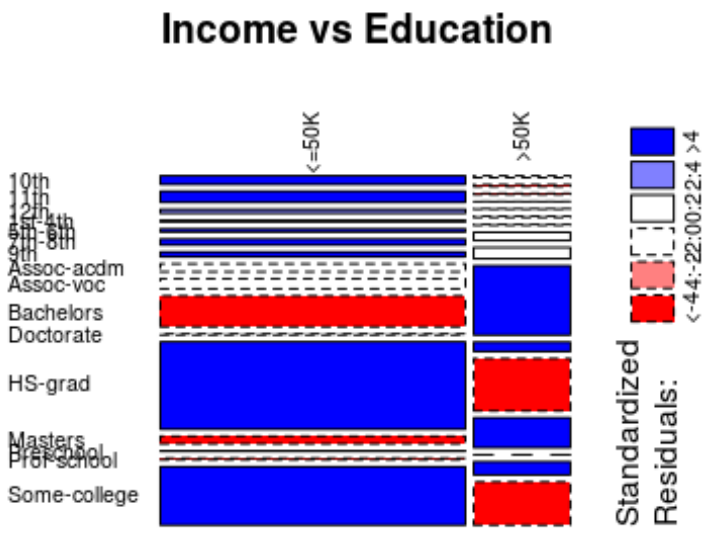
```
counts <- table(df$income, df$nativecountry)

counts[counts==0] <- NA # because log(0) is not defined

barplot(counts, main="Income vs Native Country",
        ylab="log (Count (Income))",
        col = cm.colors(2),                    # set column colors
        legend = rownames(counts),             # legend variables
        args.legend = list(x = "top"),         # legend position
        las = 2,                               # Rotate x labels name 90 degree
        beside=TRUE,                           # columns of a row come beside
        log='y',                               # y axis in logarithm
        )

## Warning in yinch(0.1): y log scale: yinch() is nonsense
```

محور x اسم nativecountry و محور y لگاریتم تعداد تکرار متغیر ها می باشد. با آرگومان las = 2 اسم ها را ۹۰ درجه می چرخانیم تا بشود در یک صفحه جا داد!



نمودار موزاییکی متغیرهای income,education

```
mosaicplot(table(df$income, df$education),
           main = "Income vs Education",
           las = 2,
           shade=TRUE,
           )
```



نمودار تجمعی فراوانی افراد

```
plot(ecdf(df$age),
     main= "cumulative distribution of Age",
     xlab="age",
     ylab="cumulative distribution function",
     col = cm.colors(1),
     )
```