

Task1

Sokolikov Jack

07 10 2019

1. Выбор данных

<https://vincentarelbundock.github.io/Rdatasets/datasets.html> (<https://vincentarelbundock.github.io/Rdatasets/datasets.html>)
- здесь есть множество разных датасетов с описанием. При этом можно предварительно посмотреть на количество наблюдений, количество признаков, сколько из них являются факторами и т.д.

2. Считывание и просмотр

```
df <- read.csv(file = "Angell.csv", header = TRUE) #Скачиваем csv, сохраняем в папку, где лежит r markdown и считываем данные.  
head(df) #Head() Показывает первые шесть строк данных.
```

```
##           X moral hetero mobility region  
## 1 Rochester 19.0    20.6    15.0      E  
## 2 Syracuse 17.0    15.6    20.2      E  
## 3 Worcester 16.4    22.1    13.6      E  
## 4      Erie 16.2    14.0    14.8      E  
## 5 Milwaukee 15.8    17.4    17.6     MW  
## 6 Bridgeport 15.3    27.9    17.5      E
```

3. Описание данных

Рассматриваемые данные содержат различную информацию о 43 городах в Америке около 1950г. moral - Moral integration - Composite of crime rate and welfare expenditures hetero - Ethnic Heterogeneity - From percentages of nonwhite and foreign-born white residents mobility - Geographic Mobility - From percentages of residents moving into and out of the city region - A factor with levels: E Northeast; MW Midwest; S Southeast; W West.

4. Типы признаков

moral - количественный hetero - количественный mobility - количественный region - качественный

```
library(plyr) #подключает библиотеку plyr, в которой есть функция count, строящая таблицу частот  
sapply(df[,2:5], function(x) max(count(x)$freq)) #считает сколько раз встречается мода у каждого признака
```

```
##      moral      hetero mobility      region  
##          2          1          2          14
```

Из таблицы выше можем заключить, что hetero - абсолютно непрерывный признак. Кроме того, можем считать, что moral и mobility ближе к непрерывным признакам.

5. Порядковые признаки

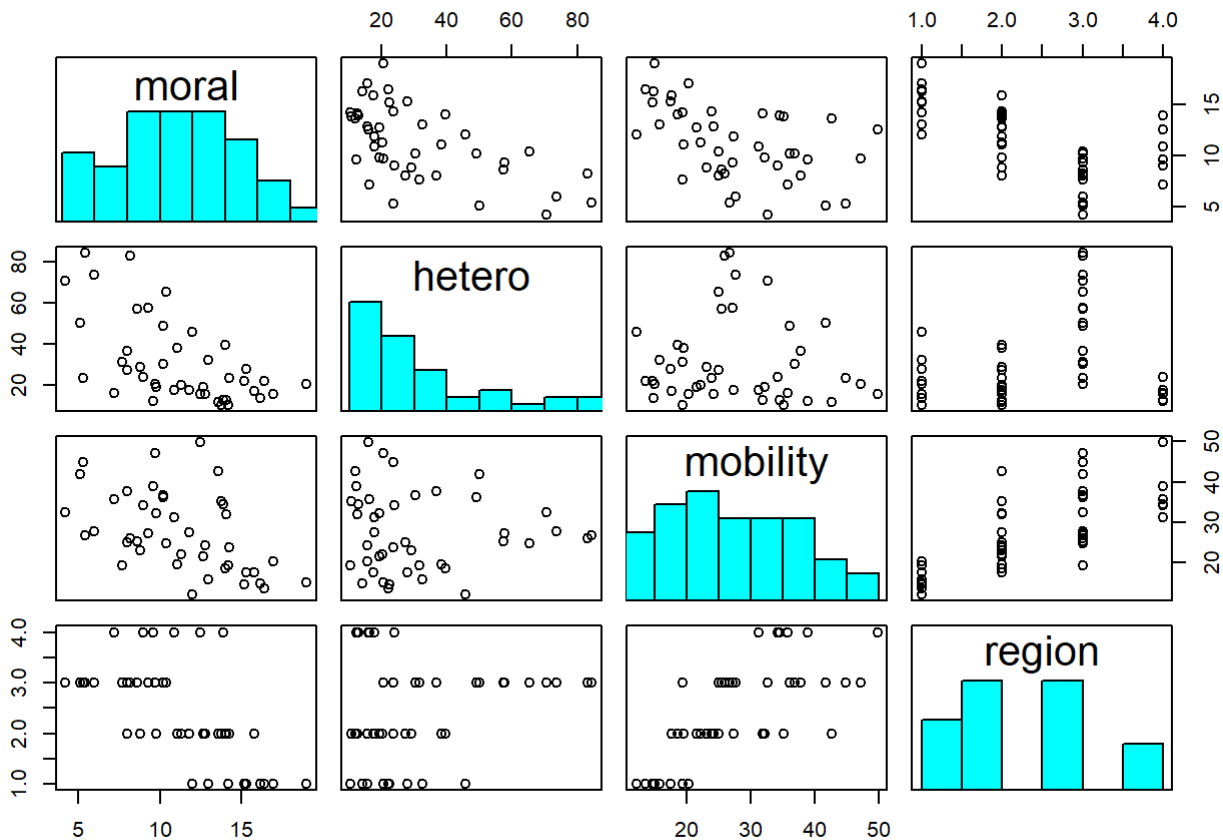
Здесь необходимо проверить соответствие текстовых меток порядкового признака (если такие использованы) их естественному порядку.

В данном датасете таких признаков нет.

6. Matrix plot, outliers, etc.

```
panel.hist <- function(x, ...) #функция рисующая гистограмму, предлагаемая в help'е функции pairs
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

pairs(df[,2:5], diag.panel = panel.hist) #рисует matrix plot для 2-5 столбца датасета df с гистограммами на диагонали
```

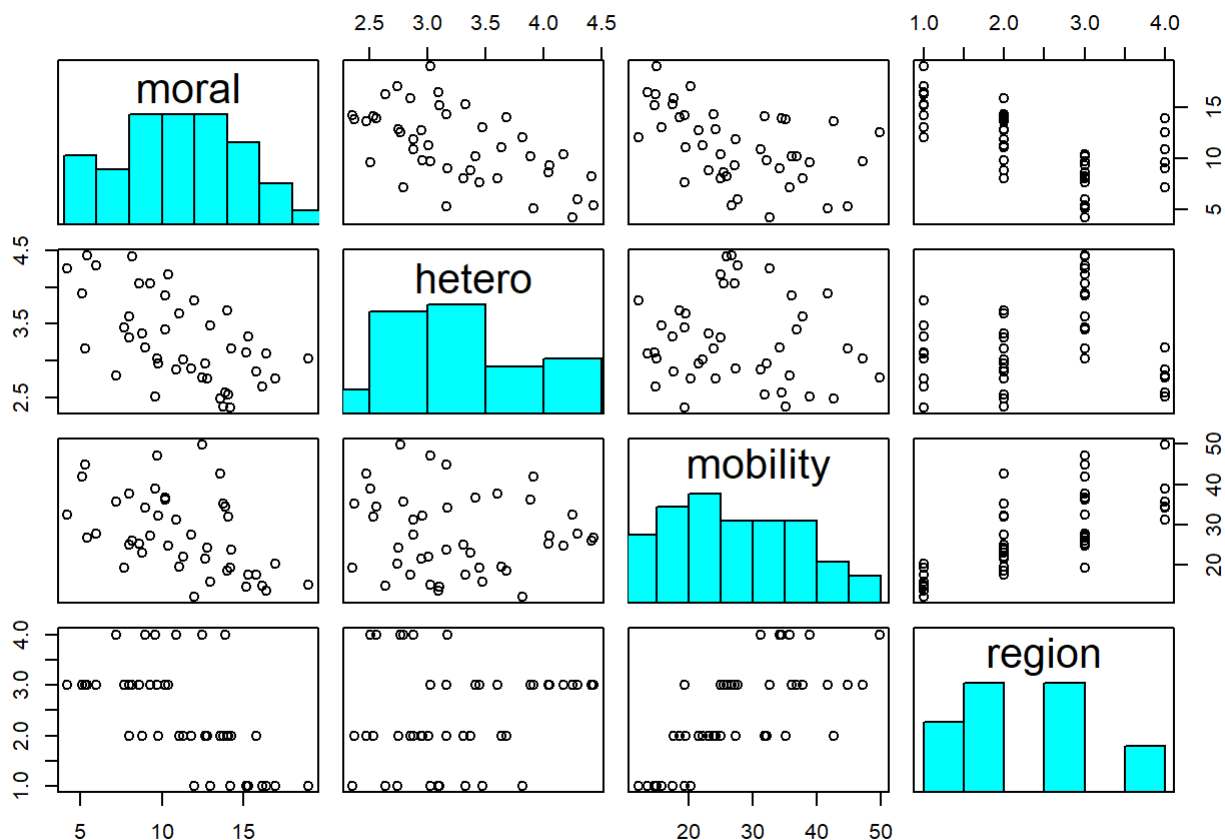


7. Симметричность распределений.

Из матрикс плота видно, что распределение hetero - сильно несимметричное с хвостом вправо, поэтому прологарифмируем его и построим заного матрикс плот.

```
df1 <- df #Сохраняем датасет
df1[, 3] <- log(df[3]) #Логарифмируем признак, стоящий в третьем столбце

pairs(df1[,2:5], diag.panel = panel.hist) #Перерисовываем матрикс плот
```



8. Аутлаеры

В данном случае на одном скаттерплоте (hetero & moral) видна линейная зависимость, не имеющая аутлаеров относительно нее. На скаттерплоте (mobility & hetero) отсутствует зависимость и, соответственно, аутлаеры. На (moral & mobility) также линейная зависимость, но вместо аутлаеров наблюдается скорее неоднородность.

9. Неоднородности

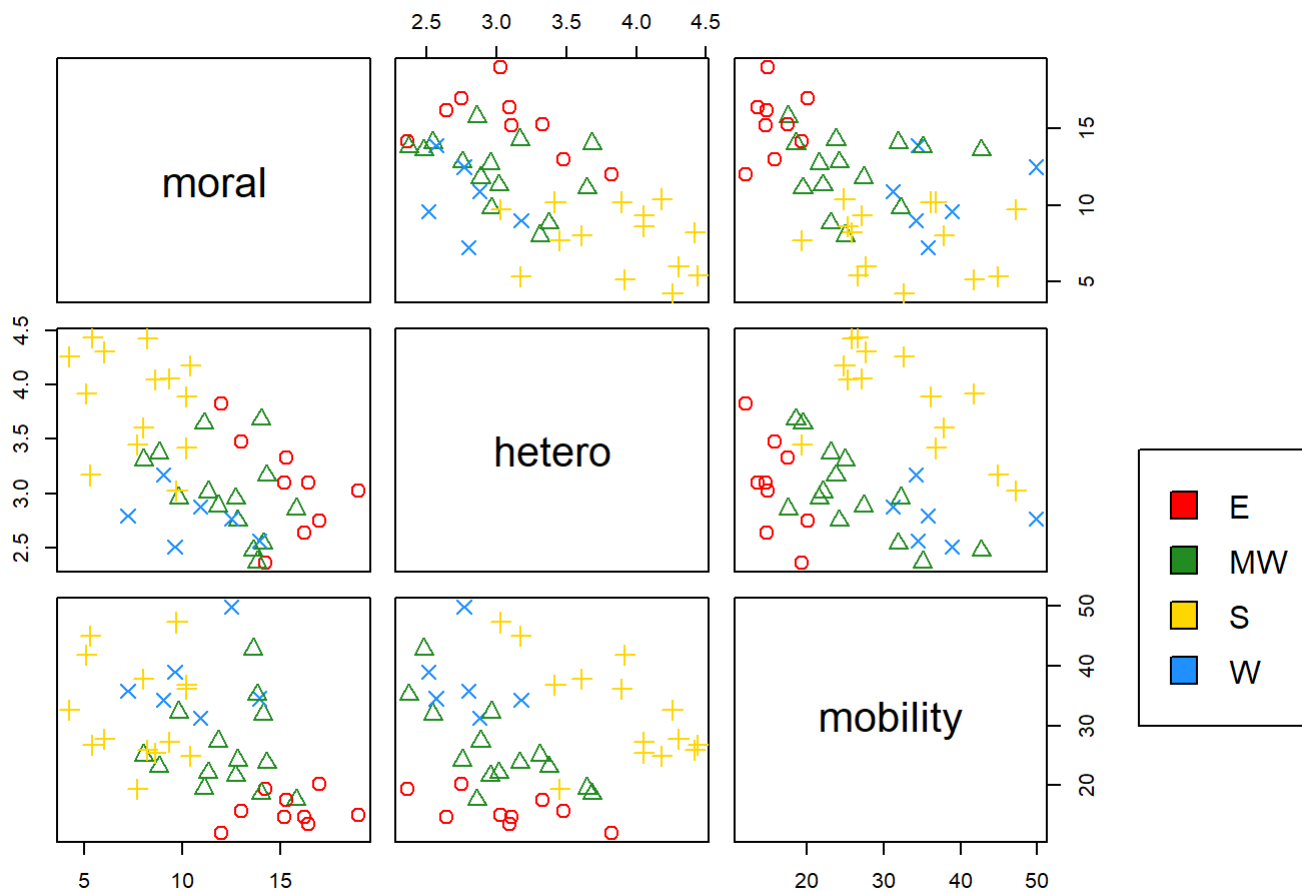
```
mycol <- c("red", "forestgreen", "gold", "dodgerblue") #Задаем раскраску
```

```
pairs(dfl[,2:4], col = mycol[dfl$region], pch = c(1:4)[dfl$region], oma = c(3,3,3,10), cex = 1.5)
```

#col - отвечает за раскраску, pch - за форму точек, oma - за отступы от краев, cex - за размер точек.

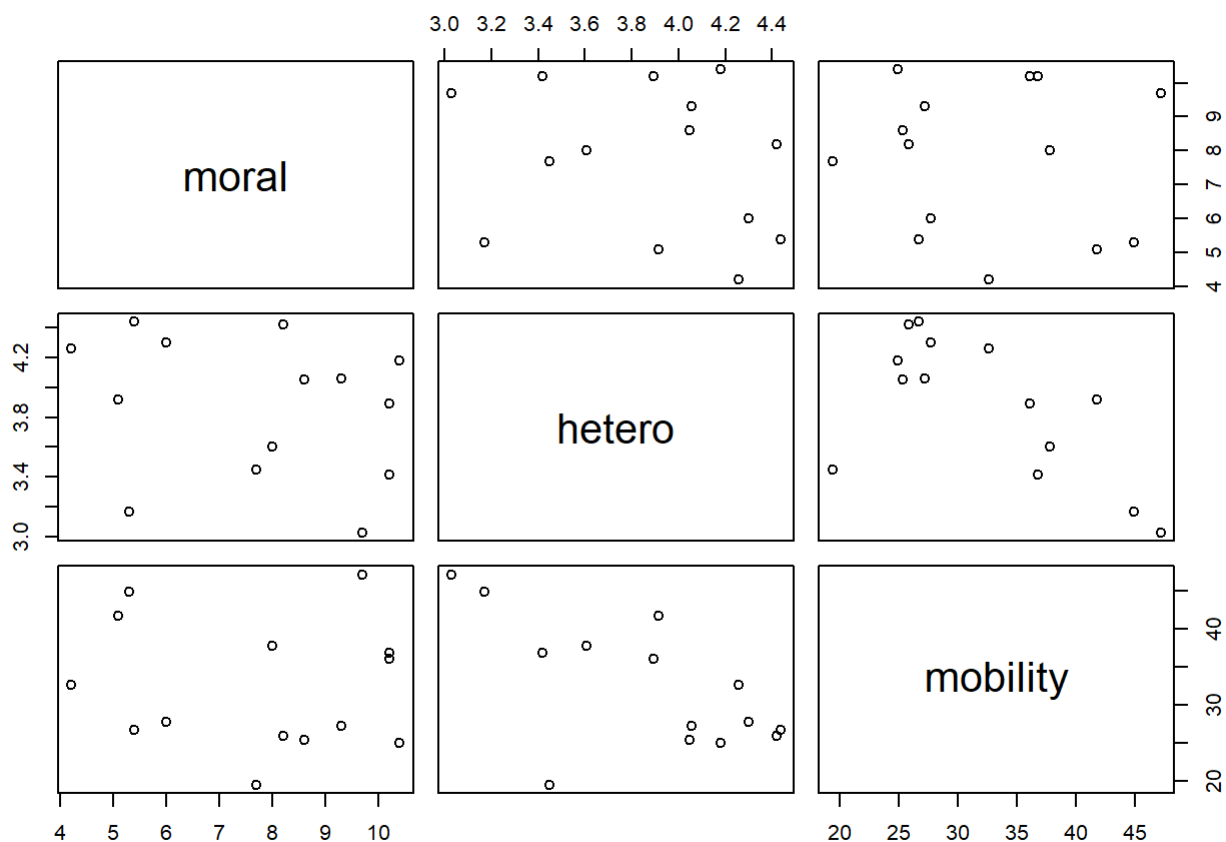
```
par(xpd = TRUE) #Устанавливает параметры, позволяющие нарисовать справа от картинки легенду обо значениях
```

```
legend(0.9, 0.5, levels(dfl$region), fill = mycol) #Показывает легенду справа от матрикс плота
```

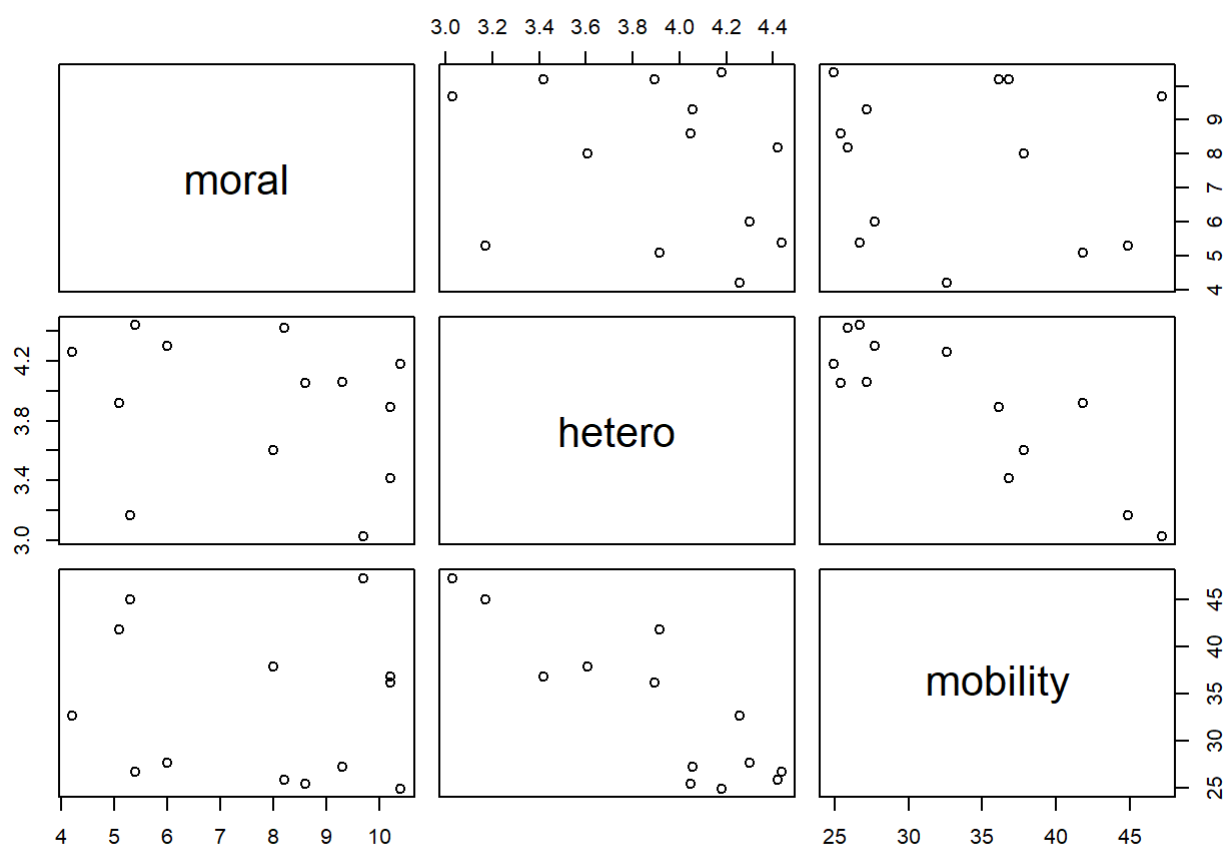


10. Матрикс плот и аутлаеры для отдельной группы.

```
pairs(dfl[df1$region == 'S',2:4])
```



```
dflo <- dfl #сохраняем датасет
dflo[38,] <- NA #удаляем аутлаер относительно линейной зависимости между hetero и mobility
pairs(dflo[dflo$region == 'S',2:4]) #перерисовываем матрикс плот
```



11. Descriptive statistics

```
summary(df)
```

```
##           X          moral          hetero          mobility          region
## Akron      : 1   Min.    : 4.20   Min.    :10.60   Min.    :12.10   E : 9
## Atlanta    : 1   1st Qu.: 8.70   1st Qu.:16.90   1st Qu.:19.45   MW:14
## Baltimore  : 1   Median :11.10   Median :23.70   Median :25.90   S :14
## Birmingham: 1   Mean    :11.20   Mean    :31.37   Mean    :27.60   W : 6
## Bridgeport: 1   3rd Qu.:13.95   3rd Qu.:39.00   3rd Qu.:34.80
## Buffalo    : 1   Max.    :19.00   Max.    :84.50   Max.    :49.80
## (Other)    :37
```

```
summary(dfl)
```

```
##           X          moral          hetero          mobility          region
## Akron      : 1   Min.    : 4.20   Min.    :2.361   Min.    :12.10   E : 9
## Atlanta    : 1   1st Qu.: 8.70   1st Qu.:2.827   1st Qu.:19.45   MW:14
## Baltimore  : 1   Median :11.10   Median :3.165   Median :25.90   S :14
## Birmingham: 1   Mean    :11.20   Mean    :3.267   Mean    :27.60   W : 6
## Bridgeport: 1   3rd Qu.:13.95   3rd Qu.:3.663   3rd Qu.:34.80
## Buffalo    : 1   Max.    :19.00   Max.    :4.437   Max.    :49.80
## (Other)    :37
```

```
library(fBasics)
```

```
## Loading required package: timeDate
```

```
## Loading required package: timeSeries
```

```
colSkewness(dfl[,2:4]) #Коэффициент асимметрии
```

```
##          moral          hetero          mobility  
## -0.01846484  0.41254605  0.39774260
```

```
colKurtosis(dfl[,2:4]) #Коэффициент эксцесса
```

```
##          moral          hetero          mobility  
## -0.8097116 -0.9406462 -0.8028643
```

2.1 Выбор категоризирующей переменной

В качестве категоризирующего признака возьмем region. Сравнивать будем центральные штаты и юго-восточный регион, как имеющие наибольшее количество наблюдений (14 и 13, соответственно).

2.2 Boxplot

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'
```

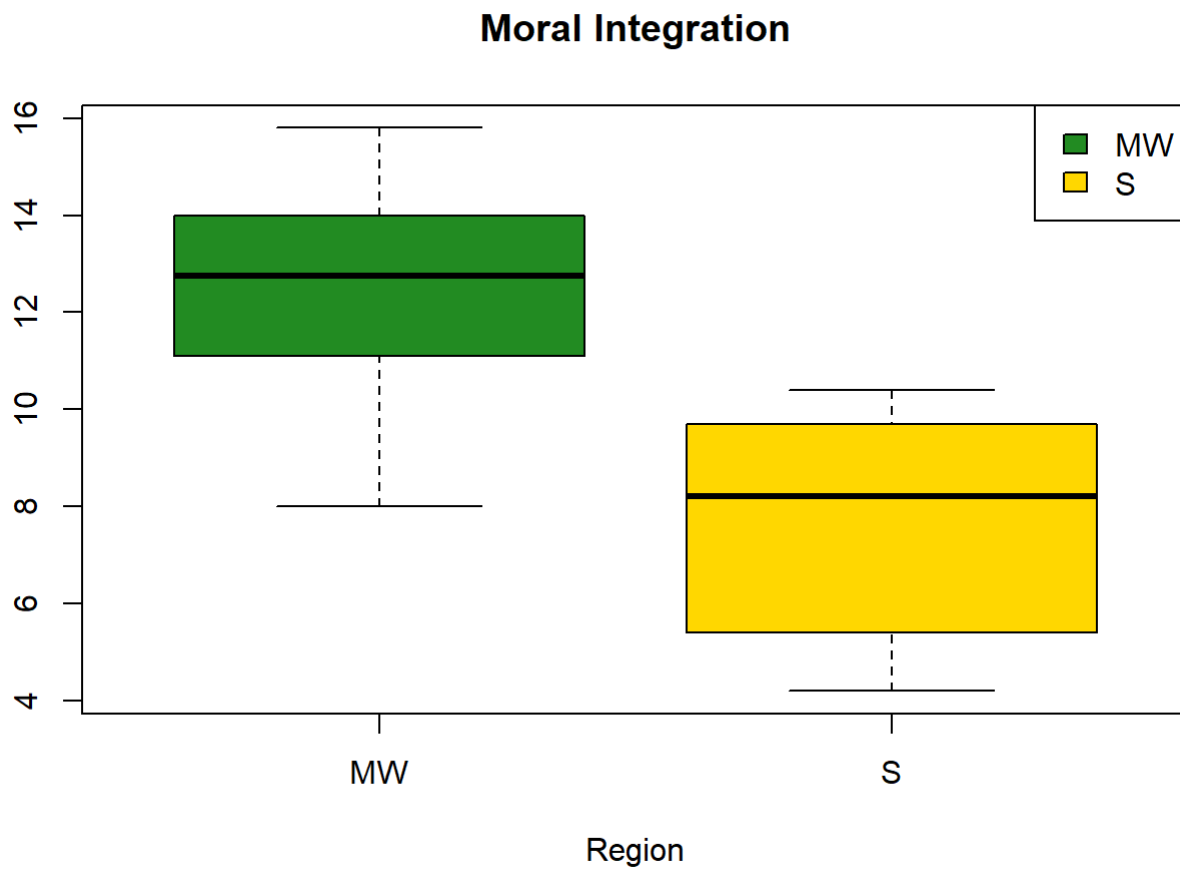
```
## The following objects are masked from 'package:timeSeries':  
##  
##      filter, lag
```

```
## The following objects are masked from 'package:plyr':  
##  
##      arrange, count, desc, failwith, id, mutate, rename, summarise,  
##      summarize
```

```
## The following objects are masked from 'package:stats':  
##  
##      filter, lag
```

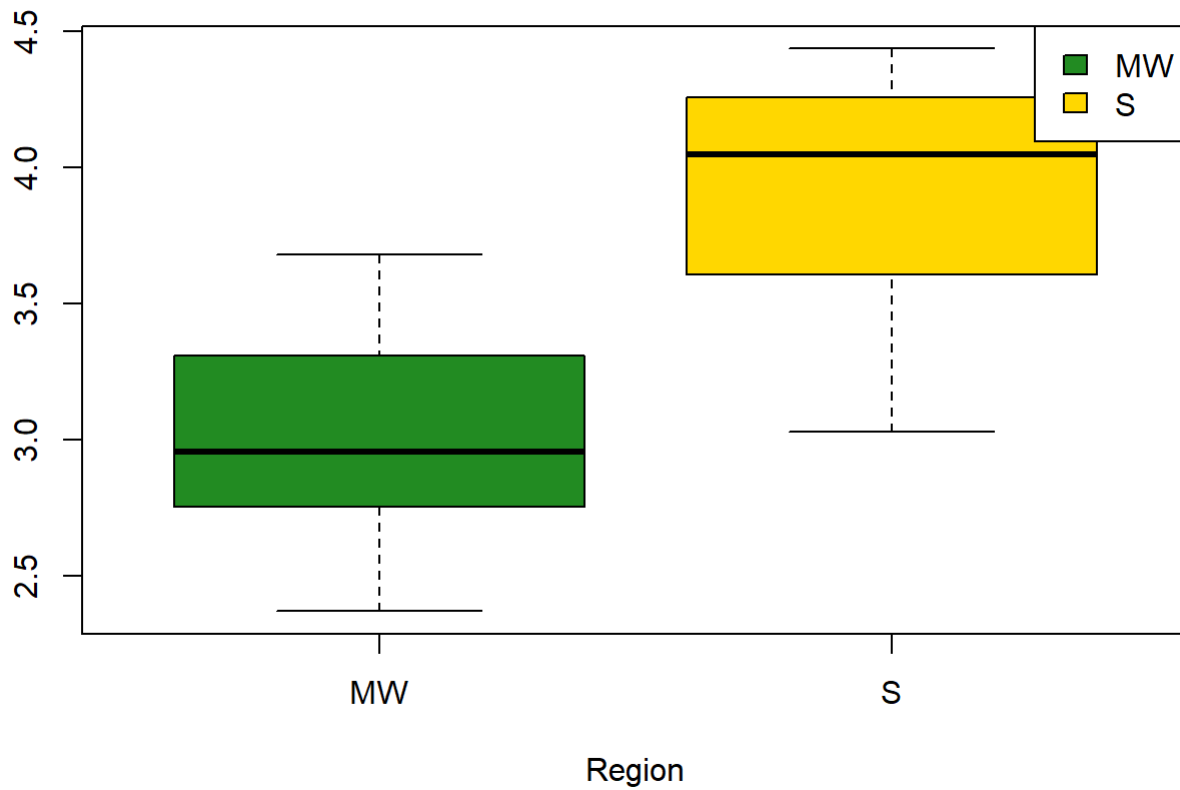
```
## The following objects are masked from 'package:base':  
##  
##      intersect, setdiff, setequal, union
```

```
dfcomp <- dflo %>% filter(region == "S" | region == "MW")
dfcomp$region <- factor(dfcomp$region)
boxplot(moral ~ region, data = dfcomp, col = c("forestgreen", "gold"), main = "Moral Integration", xlab = "Region")
legend("topright", levels(dfcomp$region), fill = c("forestgreen", "gold"))
```



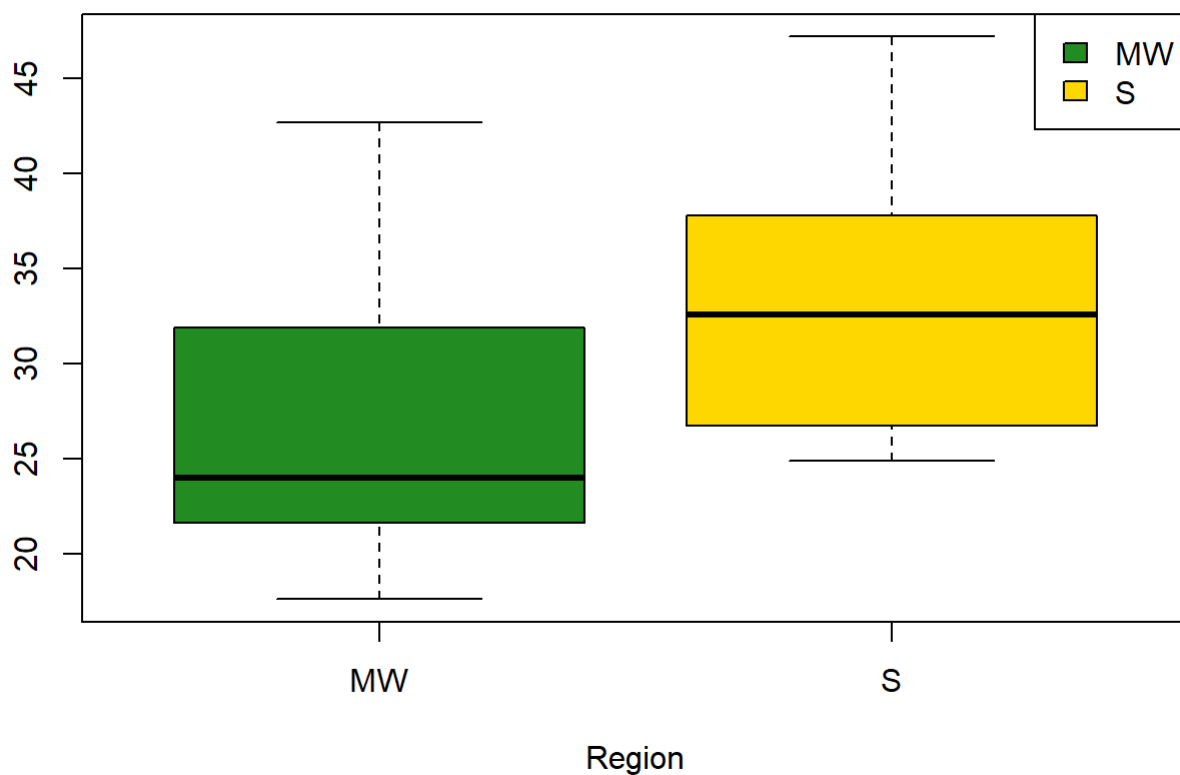
```
boxplot(hetero ~ region, data = dfcomp, col = c("forestgreen", "gold"), main = "Ethnic Heterogeneity", xlab = "Region")
legend("topright", levels(dfcomp$region), fill = c("forestgreen", "gold"))
```

Ethnic Heterogeneity



```
boxplot(mobility ~ region, data = dfcomp, col = c("forestgreen", "gold"), main = "Geographic Mobility", xlab = "Region")
legend("topright", levels(dfcomp$region), fill = c("forestgreen", "gold"))
```

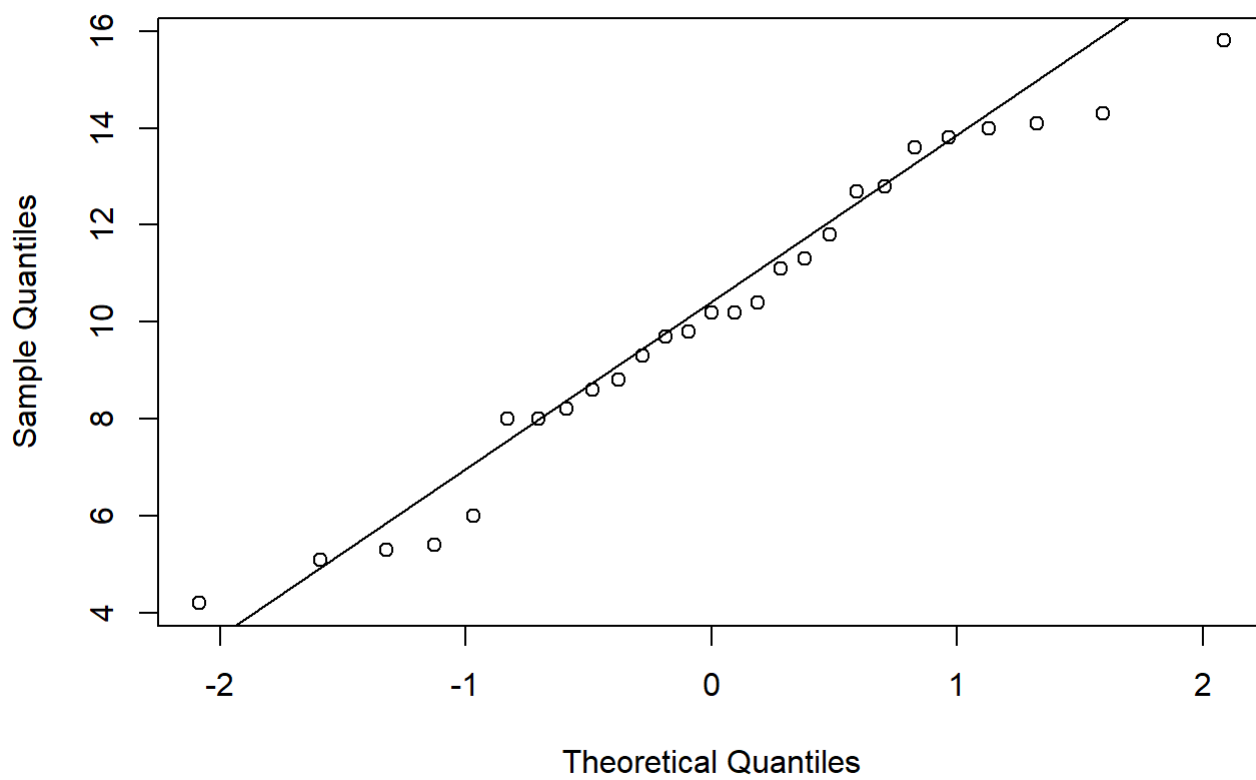
Geographic Mobility



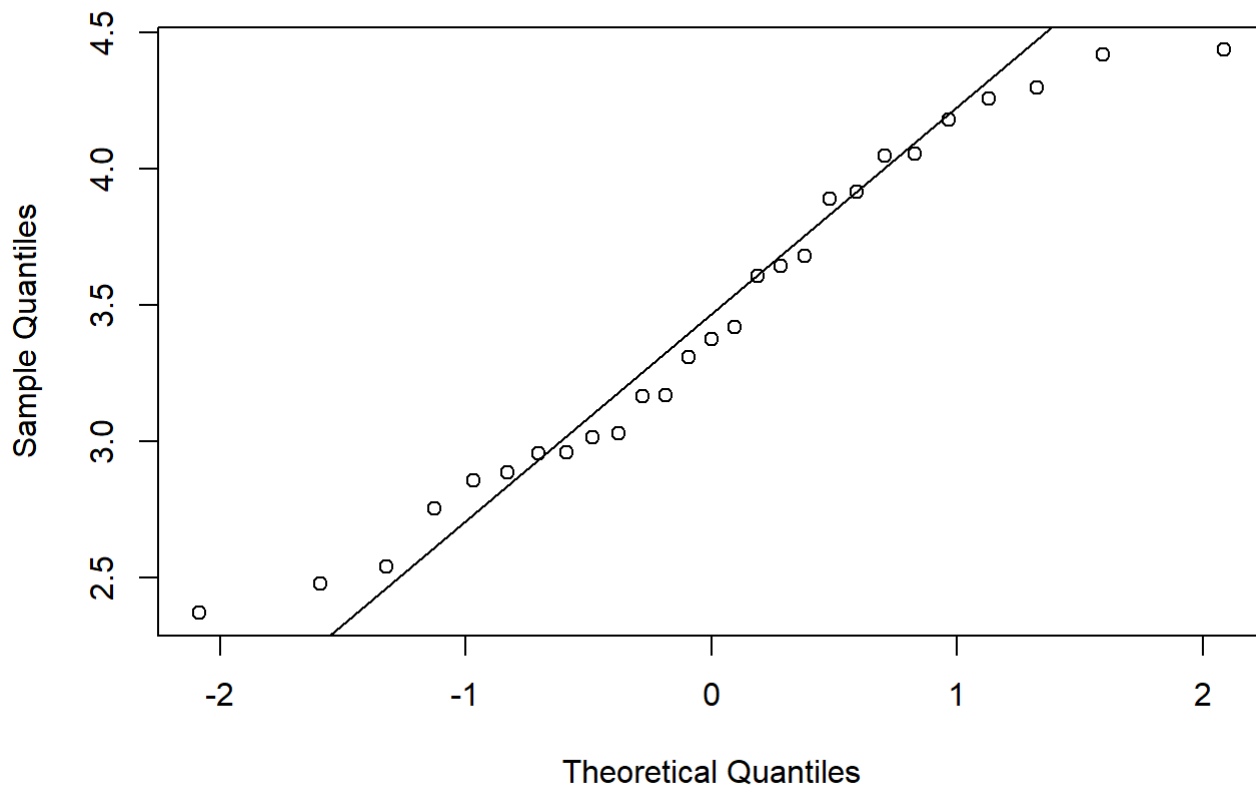
2.3 Нормальность признаков

```
sapply(dfcomp[,2:4], function(x) {  
  qqnorm(x)  
  qqline(x) })
```

Normal Q-Q Plot

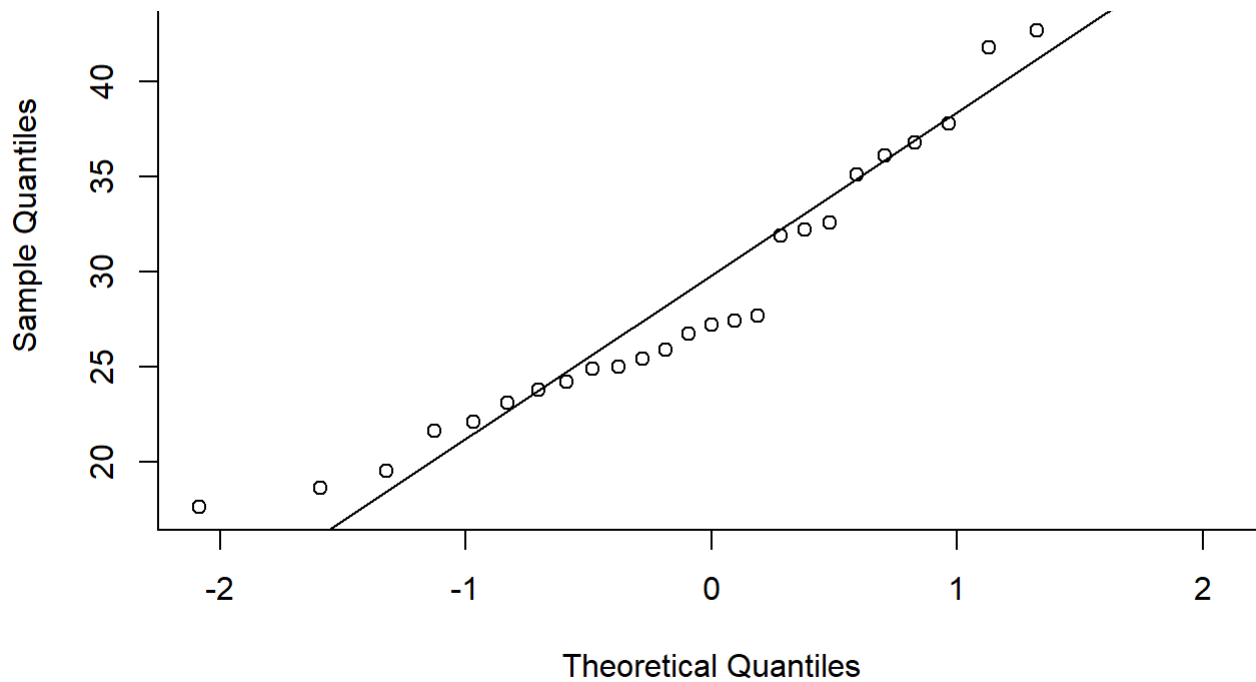


Normal Q-Q Plot



Normal Q-Q Plot





```
## $moral
## NULL
##
## $hetero
## NULL
##
## $mobility
## NULL
```

```
shapiro.test(dfcomp$moral)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dfcomp$moral
## W = 0.96495, p-value = 0.4754
```

```
shapiro.test(dfcomp$mobility)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dfcomp$mobility
## W = 0.93886, p-value = 0.1142
```

2.4 t-test, критерий Манна-Уитни

```
t.test(moral ~ region, data = dfcomp)$p.value #show only p-value
```

```
## [1] 2.062762e-05
```

```
t.test(hetero ~ region, data = dfcomp)
```

```
##
## Welch Two Sample t-test
##
## data: hetero by region
## t = -5.3758, df = 23.917, p-value = 1.628e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.249586 -0.556182
## sample estimates:
## mean in group MW mean in group S
## 2.999347 3.902231
```

```
t.test(mobility ~ region, data = dfcomp)
```

```
##
## Welch Two Sample t-test
##
## data: mobility by region
## t = -2.5619, df = 24.245, p-value = 0.01704
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -13.366225 -1.442566
## sample estimates:
## mean in group MW mean in group S
## 26.05714 33.46154
```

```
wilcox.test(mobility ~ region, data = dfcomp)
```

```
##
## Wilcoxon rank sum test
##
## data: mobility by region
## W = 36, p-value = 0.006618
## alternative hypothesis: true location shift is not equal to 0
```

2.5 Критерий Колмогорова-Смирнова

```
ks.test(dfcomp[dfcomp$region == "S", 4], dfcomp[dfcomp$region == "MW", 4])
```

```
##
## Two-sample Kolmogorov-Smirnov test
##
## data: dfcomp[dfcomp$region == "S", 4] and dfcomp[dfcomp$region == "MW", 4]
## D = 0.57143, p-value = 0.009595
## alternative hypothesis: two-sided
```