МОДУЛЬНОЕ ДОМАШНЕЕ ЗАДАНИЕ №2
подготовила **Сланова Айгерим**
группы **MCA201**

Задание 1

Данные из файла "задание1.xlsx".

Значения наименований атрибутов содержится в таблице Tab.1.

```
> str(data)
'data.frame':   440 obs. of  15 variables:
 $ id : int  96 97 99 101 103 104 106 108 109 111 ...
 $ x1 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ x2 : num  61 55 69.1 53.5 61.6 ...
 $ x3 : num  2.68 7.42 51.5 64.43 56.86 ...
 $ x4 : num  186.8 11.3 57 59.1 96.7 ...
 $ x5 : num  0.67 2.18 23.53 1.34 19.57 ...
 $ x6 : num  4.02 4.36 28.41 2.68 20.31 ...
 $ x7 : num  45.1 89.6 129.3 86.7 88.3 ...
 $ x8 : num  5.93 3.29 16.78 9.09 13.87 ...
 $ x9 : num  0.67 1.31 9.92 2.68 1.85 4.45 9.27 2.8 9.7 0 ...
 $ x10: num  706 569 806 676 705 ...
 $ x11: num  0 2 18 0 6 14 6 25 55 14 ...
 $ x12: num  42 30 29.2 21.3 18.8 ...
 $ x13: num  2.68 4.72 11.24 8.51 3.14 ...
 $ x14: num  100 100 41.7 32.2 75 ...
```

Уберем переменную id и рассмотрим корреляцию между атрибутами x1-x14:

```
> data2 <- select(data, -id)
> cor(data2)
             x1          x2          x3          x4          x5          x6         x7
x1  1.000000000  0.22627206  0.38916828  0.40166200  0.581799752 0.61015889 0.5565576
x2  0.226272058  1.00000000  0.13125322  0.16948977  0.249363500 0.23581220 0.1848333
x3  0.389168277  0.13125322  1.00000000  0.92072417  0.794319640 0.77478346 0.4989194
x4  0.401661998  0.16948977  0.92072417  1.00000000  0.722199454 0.74494927 0.4756570
x5  0.581799752  0.24936350  0.79431964  0.72219945  1.000000000 0.95099419 0.6011616
x6  0.610158893  0.23581220  0.77478346  0.74494927  0.950994186 1.00000000 0.6506310
x7  0.556557597  0.18483329  0.49891939  0.47565699  0.601161578 0.65063103 1.0000000
x8  0.362890982  0.10543383  0.23802093  0.22024778  0.356282738 0.38736229 0.3661828
x9  0.182499306 -0.05733381  0.32350102  0.27441446  0.313272134 0.30120398 0.2113126
x10 0.292270945  0.40313459  0.19198852  0.18036012  0.300582870 0.33341125 0.4478129
x11 0.421595754  0.26614824  0.56292304  0.46667766  0.637678007 0.59577836 0.3571765
x12 0.083967243  0.33240237  0.02196341  0.02198431  0.098230326 0.11837489 0.2152434
x13 0.519306760  0.08644460  0.49140149  0.51497314  0.581709429 0.62730779 0.7687515
x14 0.008816082 -0.09717176 -0.03521508 -0.03839485 -0.009981722 0.00988818 0.2402921
             x8          x9         x10         x11         x12         x13
x1  0.3628910  0.18249931  0.29227094  0.42159575  0.08396724  0.5193067602
x2  0.1054338 -0.05733381  0.40313459  0.26614824  0.33240237  0.0864445970
x3  0.2380209  0.32350102  0.19198852  0.56292304  0.02196341  0.4914014866
x4  0.2202478  0.27441446  0.18036012  0.46667766  0.02198431  0.5149731408
x5  0.3562827  0.31327213  0.30058287  0.63767801  0.09823033  0.5817094290
x6  0.3873623  0.30120398  0.33341125  0.59577836  0.11837489  0.6273077898
x7  0.3661828  0.21131261  0.44781292  0.35717652  0.21524342  0.7687514753
x8  1.0000000  0.27630611  0.18818998  0.27356410  0.26643145  0.3542949638
x9  0.2763061  1.00000000 -0.01157611  0.24016471 -0.00860175  0.2393677341
x10 0.1881900 -0.01157611  1.00000000  0.20680559  0.65838405  0.2076824085
x11 0.2735641  0.24016471  0.20680559  1.00000000  0.08187249  0.3941612413
x12 0.2664315 -0.00860175  0.65838405  0.08187249  1.00000000  0.1100320617
x13 0.3542950  0.23936773  0.20768241  0.39416124  0.11003206  1.0000000000
x14 0.2011547  0.02850109  0.14975760 -0.07509848  0.36546612 -0.0005366207
              x14
x1   0.0088160820
x2  -0.0971717577
x3  -0.0352150754
x4  -0.0383948497
x5  -0.0099817218
x6   0.0098881797
x7   0.2402921186
x8   0.2011546580
x9   0.0285010873
x10  0.1497576012
x11 -0.0750984758
x12  0.3654661248
x13 -0.0005366207
x14  1.0000000000
```

Высокие корреляции заметны между

- x3 и x4 (= 0.92072417)
- x3 и x5 (= 0.79431964)
- x3 и x6 (= 0.77478346)

- x5 b x6 (= 0.950994186)
- и тд

x1	Ведущий вуз
x2	Средний балл ЕГЭ студентов, принятых по результатам ЕГЭ на обучение по очной форме по программам бакалавриата и специалитета за счет средств соответствующих бюджетов бюджетной системы РФ
x3	Количество цитирований публикаций, изданных за последние 5 лет, индексируемых в информационно-аналитической системе научного цитирования Web of Science в расчете на 100 НПР
x4	Количество цитирований публикаций, изданных за последние 5 лет, индексируемых в информационно-аналитической системе научного цитирования Scopus в расчете на 100 НПР
x5	Число публикаций организации, индексируемых в информационно-аналитической системе научного цитирования Web of Science, в расчете на 100 НПР
x6	Число публикаций организации, индексируемых в информационно-аналитической системе научного цитирования Scopus, в расчете на 100 НПР
x7	Доходы от НИОКР (за исключением средств бюджетов бюджетной системы Российской Федерации, государственных фондов поддержки науки) в расчете на одного НПР
x8	Удельный вес численности НПР без ученой степени – до 30 лет, кандидатов наук – до 35 лет, докторов наук – до 40 лет, в общей численности НПР
x9	Количество полученных грантов за отчетный год в расчете на 100 НПР
x10	Доходы образовательной организации из средств от приносящей доход деятельности в расчете на одного НПР
x11	Число статей, подготовленных совместно с зарубежными организациями
x12	Доля доходов вуза из внебюджетных источников
x13	Доля доходов вуза от научных исследований и разработок в общих доходах вуза
x14	Доля внебюджетных средств в доходах от научных исследований и разработок

Tab. 1: Значения атрибутов.

Метод главных компонент.

Применяем метод главных компонент к исходным переменным и печатаем сами главные компоненты:

```
> data2.pca <- prcomp(data2,scale = TRUE)
> pc <- data2.pca$x
> head(pc)
            PC1         PC2          PC3         PC4         PC5         PC6          PC7          PC8
[1,] -1.4965876  0.3151939 -0.43999541 -0.2395978  1.2856789  0.4468463 -0.64006204  0.08645641
[2,] -1.6526855 -0.4180577 -0.72676583  0.1863014  1.4564744  0.2613117 -0.99776951  0.24388149
[3,] -0.0722358 -0.6903870 -0.25390118 -0.1305109 -0.9425785 -0.3899267  0.41296437 -0.27838296
[4,] -1.3995789 -1.4311570 -0.08171096  0.6532300  0.1597397 -0.3968516  0.33036880  0.61627273
[5,] -0.9513883 -0.8031789 -0.45814802  0.1167132  0.2716925  0.7905054 -0.05430427 -0.12535458
[6,] -1.1565221 -0.3049966 -0.32053911 -0.2823398  0.4768330  0.2639206 -0.49986362 -0.22935512
             PC9         PC10         PC11         PC12         PC13          PC14
[1,] -0.02013414  0.62767164  0.12206408 -0.16415577 -0.16212061  0.041022117
[2,]  0.04988232  0.31200486  0.03178291  0.01298064 -0.01888097 -0.001622828
[3,]  0.25500215 -0.26925955  0.28520174  0.15368388  0.06093750  0.050403247
[4,]  0.04875078  0.07550318 -0.29733150 -0.03680930  0.03386413  0.028772767
[5,] -0.23832049 -0.63082907 -0.18040863  0.11945344  0.01076175  0.128972799
[6,]  0.04085784  0.20176516  0.02786997 -0.07552283 -0.04988124 -0.045331958
```

a) Покажем относительные вклады каждого компонента в общий разброс данных:

```
> summary(data2.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.4059 1.4092 1.14417 0.96359 0.94776 0.82369 0.74918 0.72619 0.69719 0.56762 0.53358
Proportion of Variance 0.4134 0.1419 0.09351 0.06632 0.06416 0.04846 0.04009 0.03767 0.03472 0.02301 0.02034
Cumulative Proportion  0.4134 0.5553 0.64880 0.71513 0.77929 0.82775 0.86784 0.90551 0.94023 0.96324 0.98358
                          PC12    PC13    PC14
Standard deviation     0.35872 0.2592 0.18459
Proportion of Variance 0.00919 0.0048 0.00243
Cumulative Proportion  0.99277 0.9976 1.00000
```

Ответ на вопрос в подпункте a) о минимальном количестве компонент, которые необходимо использовать для сохранения 75% первоначальной информации: **5**

b) Выведем первые шесть весов исходных переменных в главных компонентах:

```
> head(data2.pca$rotation)
         PC1         PC2          PC3          PC4         PC5          PC6
x1 0.2860774  0.02481524 -0.006826135  0.394031346 -0.28151793  0.130036785
x2 0.1309980  0.28601704  0.552065419 -0.089233427 -0.25887789  0.079215678
x3 0.3411948 -0.22033491  0.067549833 -0.257805906  0.31404117  0.021332910
x4 0.3287936 -0.20743165  0.083679503 -0.195291296  0.33994000 -0.009045589
```

```
x5 0.3779381 -0.10949390  0.071679815 -0.065173350  0.05997512  0.117606524
x6 0.3846081 -0.08495492  0.037909166 -0.002734889  0.07193408  0.080666361
            PC7          PC8          PC9         PC10          PC11         PC12
x1 -0.345542618  0.04642600 -0.58657182  0.43039161 -0.095876113 -0.04075329
x2 -0.164811484 -0.66785601  0.18195012 -0.02361931  0.002860096  0.05099161
x3  0.165264013 -0.05386199 -0.05408528  0.10981445 -0.229121695 -0.22100264
x4  0.240761830 -0.21846266 -0.11125036  0.30541929 -0.278467538  0.16830430
x5 -0.050483736  0.06113523 -0.11923142 -0.30688260  0.482238403 -0.02107039
x6  0.005708695  0.03130751 -0.15371144 -0.28543367  0.440588758  0.08755241
           PC13         PC14
x1  0.06390926 -0.01669658
x2  0.05783168 -0.04680847
x3  0.60500433 -0.39034334
x4 -0.49914697  0.34887955
x5  0.31001238  0.61202557
x6 -0.43061467 -0.58172838
```

Формулы зависимости главных компонент из пункта а) от первоначальных данных:

$PC1 = 0.2860774 \cdot x1 + 0.1309980 \cdot x2 + 0.3411948 \cdot x3 + 0.3287936 \cdot x4 + 0.3779381 \cdot x5 + 0.3846081 \cdot x6 + 0.32176089 \cdot x7 + 0.20206610 \cdot x8 + 0.15501420 \cdot x9 + 0.18264310 \cdot x10 + 0.28062705 \cdot x11 + 0.09893056 \cdot x12 + 0.30772848 \cdot x13 + 0.02711555 \cdot x14$

$PC2 = 0.02481524 \cdot x1 + 0.28601704 \cdot x2 - 0.22033491 \cdot x3 - 0.20743165 \cdot x4 - 0.10949390 \cdot x5 - 0.08495492 \cdot x6 + 0.13651277 \cdot x7 + 0.16075792 \cdot x8 - 0.16134255 \cdot x9 + 0.49870019 \cdot x10 - 0.09218763 \cdot x11 + 0.59425588 \cdot x12 - 0.03930942 \cdot x13 + 0.35270505 \cdot x14$

$PC3 = -0.006826135 \cdot x1 + 0.552065419 \cdot x2 + 0.067549833 \cdot x3 + 0.083679503 \cdot x4 + 0.071679815 \cdot x5 + 0.037909166 \cdot x6 - 0.193201958 \cdot x7 - 0.355388519 \cdot x8 - 0.346102966 \cdot x9 + 0.209974986 \cdot x10 + 0.186320464 \cdot x11 + 0.012686508 \cdot x12 - 0.170180509 \cdot x13 + -0.534503896 \cdot x14$

$PC4 = 0.394031346 \cdot x1 - 0.089233427 \cdot x2 - 0.257805906 \cdot x3 - 0.195291296 \cdot x4 - 0.065173350 \cdot x5 + -0.002734889 \cdot x6 + 0.383669548 \cdot x7 - 0.073756200 \cdot x8 - 0.514354997 \cdot x9 - 0.023377115 \cdot x10 - 0.188699215 \cdot x11 - 0.251929909 \cdot x12 + 0.435781133 \cdot x13 - 0.145560825 \cdot x14$

$PC5 = -0.28151793 \cdot x1 - 0.25887789 \cdot x2 + 0.31404117 \cdot x3 + 0.33994000 \cdot x4 + 0.05997512 \cdot x5 + 0.07193408 \cdot x6 + 0.11159878 \cdot x7 - 0.51233973 \cdot x8 - 0.40257112 \cdot x9 + 0.09227097 \cdot x10 - 0.18320373 \cdot x11 + 0.06123391 \cdot x12 - 0.02174976 \cdot x13 + 0.38401476 \cdot x14$

Задание 2

Задана случайная величина X, распределение которой неизвестно. Из данной случайной величины была получена выборка x = 3,5,7,9,9,10,11,13,17,19. Найдите оценки f(x) в точках 5 и 10, используя:

1. Ядро Епанечникова

$$K(r) = E(r) = \frac{3}{4}(1 - r^2)[|r| \leq 1]$$

2. Гауссово ядро

$$K(r) = G(r) = (2\pi)^{-\frac{1}{2}} \exp^{-\frac{1}{2}r^2}$$

Используемая программа для вычислений и построений графиков: **Excel**

```
x: 3,5,7,9,9,10,11,13,17,19
h: 0,05
Points: [1,20]
```

1) Ядро Епанечникова Вычисляем точки с помощью Ядра Епанечникова, пример кода:

```
=IF((1-(($A4-H$3)/$F$1)^2)>0;(0,75*(1-(($A4-H$3)/$F$1)^2))/$F$1;0)
```

Далее вычисляем среднее значения(Density), пример кода:

```
=AVERAGE(H4:H13)
```

В точках 5 и 10 равно 0,05243 и 0,07152 соответственно.

2) Ядро Гаусса Вычисляем точки с помощью Ядра Гаусса, пример кода:

```
=EXP(-(((($A4-H$3)/$F$1)^2)/2)/SQRT(2*PI())/$F$1
```

Далее вычисляем среднее значения(Density), пример кода:

```
=AVERAGE(H4:H13)
```

В точках 5 и 10 равно 0,15957 и 0,15973 соответственно.

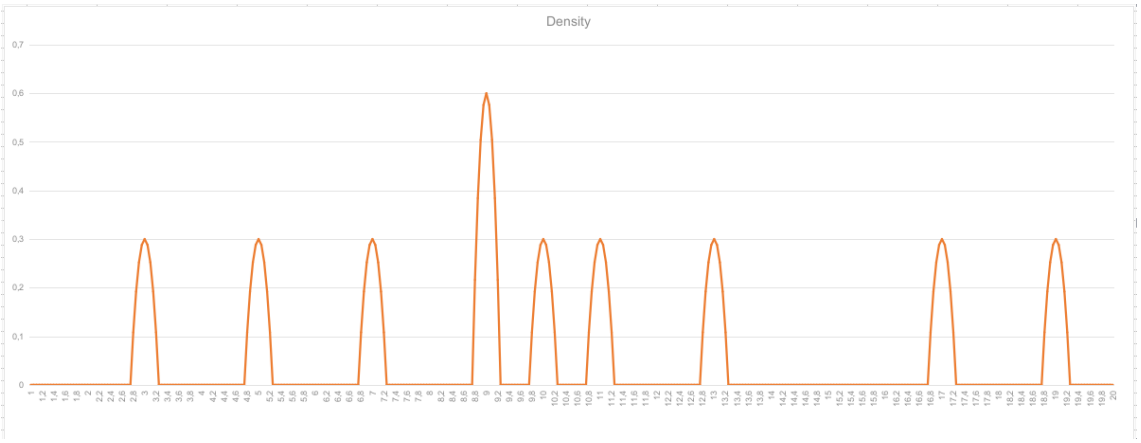

Fig. 1: Ядро Епанечникова для задании №2

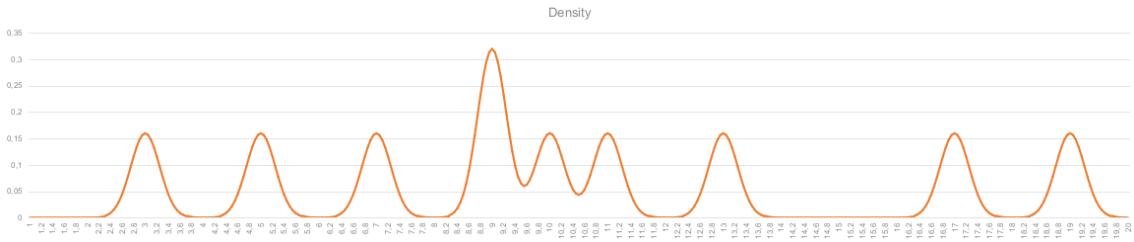

Fig. 2: Ядро Гаусса для задании №2

Задание 3

Используйте набор данных «задание3.xlsx», содержащий основные характеристики социально-экономического развития регионов России. Выберите 3-4 переменные, которые могут использоваться для классификации регионов. Обоснуйте свой выбор. Осуществите классификацию методом к-средних. Попробуйте несколько вариантов классификации с разным количеством кластеров. Определите наилучшее разбиение. Опишите состав кластеров. Реализуйте алгоритм иерархической кластеризации. Сравните результаты с разбиением, полученным на основе метода к-средних.

Для классификации было выбрано 4 переменные:

x1 - Площадь территории, тыс. км2

x2 - Численность населения, тыс. чел.

x4 - Среднедушевые денежные доходы, руб.

x6 - Среднемесячная начисленная заработная плата, руб.,

которые, как мне кажутся отражают соотношение доходности и численности населения на площадь территории.

```
> summary(data)
       x1                x2                x4             x6
 Min.   :   0.90   Min.   :  275.4   Min.   :14730   Min.   :21941
 1st Qu.:  25.70   1st Qu.:  845.5   1st Qu.:22689   1st Qu.:24743
 Median :  49.00   Median : 1246.6   Median :25398   Median :27962
 Mean   :  95.26   Mean   : 1990.6   Mean   :26584   Mean   :30709
 3rd Qu.:  84.20   3rd Qu.: 2521.3   3rd Qu.:28655   3rd Qu.:31637
 Max.   :1464.20   Max.   :12506.5   Max.   :62532   Max.   :73812
```

Используем пакет kmeans ("k-средних") из пакета rattle.data . Перебирая количество классов и итераций, заметно что при взятии больше 5 классов количество классов не меняется. Когда количество класстеров = 3 сравнимо среди всех лучше балансирует :

```
> kmean3_30 = kmeans(data,3,nstart = 30)
> kmean4_30 = kmeans(data,4,nstart = 30)
> kmean5_40 = kmeans(data,5,nstart = 40)
> kmean6_40 = kmeans(data,5,nstart = 40)
> kmean6_30 = kmeans(data,5,nstart = 40)
> kmean6_30 = kmeans(data,5,nstart = 30)
```

```
> kmean7_50 = kmeans(data,5,nstart = 50)
> kmean3_20$size
[1]  5 35 21
> kmean3_30$size
[1] 35  5 21
> kmean4_30$size
[1]  6 20  1 34
> kmean5_40$size
[1]  4 32  4  1 20
> kmean6_40$size
[1]  4 32 20  4  1
> kmean6_30$size
[1]  6 25  1 19 10
> kmean7_50$size
[1]  4  1 20 32  4
> kmean10_20 = kmeans(data,5,nstart = 50)
> kmean10_20$size
[1]  6 10  1 25 19
> kmean2_20 = kmeans(data,2,nstart = 20)
> kmean2_20$size
[1] 54  7
> kmean2_20 = kmeans(data,2,nstart = 50)
> kmean2_20$size
[1] 54  7
> kmean6_20 = kmeans(data,5,nstart = 20)
> kmean6_20$size
[1] 19 25  6  1 10
```

Далее рассмотрим как меняется величина внутри - кластерного расстояния (withinss) при изменении числа кластеров и графически отобразим на рисунке Fig.3.

```
> vec <- numeric(0)
> vec <- c(vec, 1:15)
> for (i in 1:15){ vec[i] =sum(kmeans(data, centers=i)$withinss)}
> plot(vec,type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",pch=16)
```

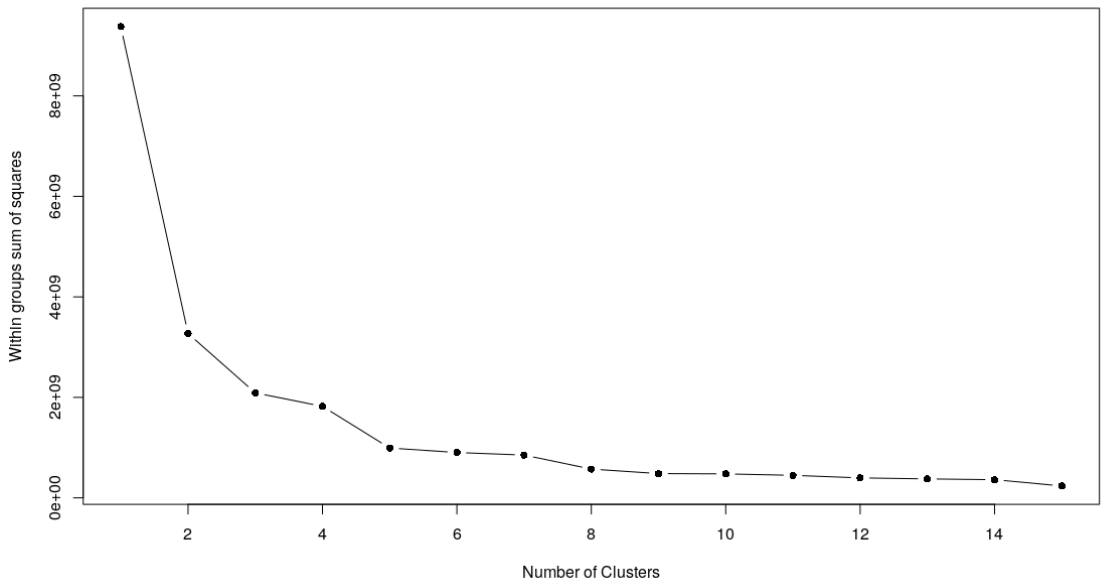

Fig. 3: задание №3

Очевидно, что при увеличении количества класстеров расстояние увеличивается.

Иерархическая кластеризация:

применяем 3 метода: average(расстояние между центроидами), complete(дальние соседи), single(ближние соседи) и выводим графики:

```
> hc.complete = hclust(dist(data),method = "complete")
> hc.average = hclust(dist(data),method = "average")
> hc.single = hclust(dist(data),method = "single")
```

```
> plot(hc.average, main = "Average linkage", xlab ="", sub="",cex = .9)
> plot(hc.complete, main = "Complete linkage", xlab ="", sub="",cex = .9)
> plot(hc.single, main = "Single linkage", xlab ="", sub="",cex = .9)
```

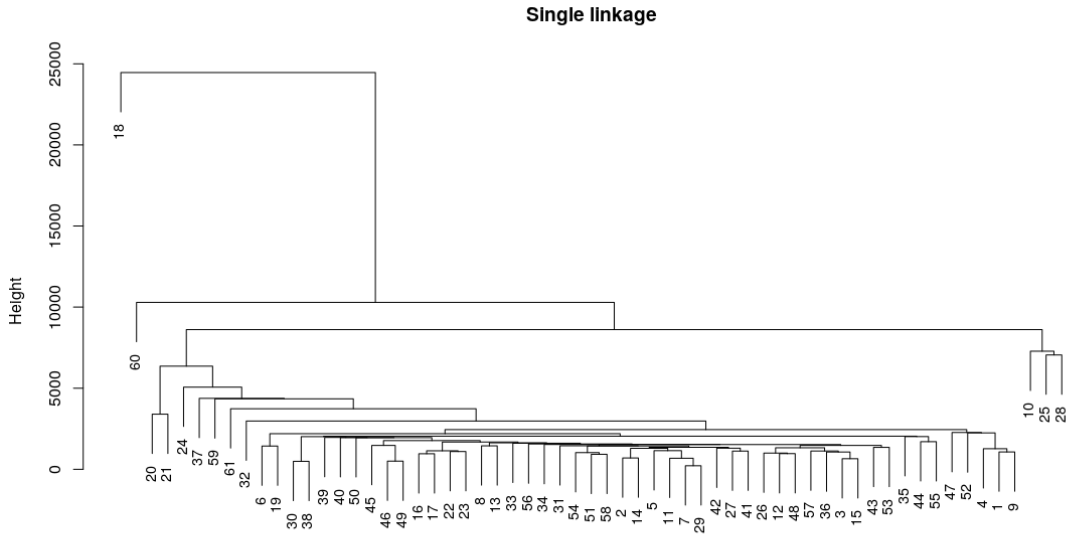

Fig. 4: задание №4

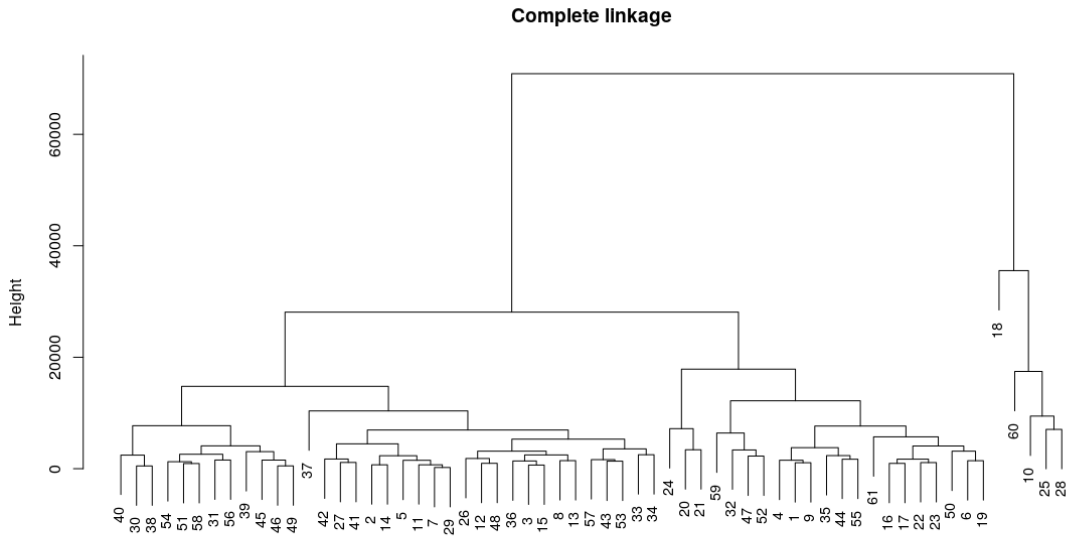

Fig. 5: задание №3

По графикам видно, что Average(Fig.6) и Complete(Fig.5) показывают примерно одно и то же, а Single(Fig.4) показывает похоже на ком, в котором наслаивается класстеры, а это не совсем хорошо. Мне больше понравился графически Average, и чтоб отрезать его на три класстера нам нужно примерно взять h = от 16000 до 18000

Задание 4

Используйте набор данных «задание4.xlsx». Постройте ядерные оценки плотности, используя ядро Гаусса, ядро Епанечникова, а также треугольное ядро. Постройте графики. Проинтерпретируйте результаты.

Выбраны данные под столбцом x2 и для удобства их прологарифмируем.

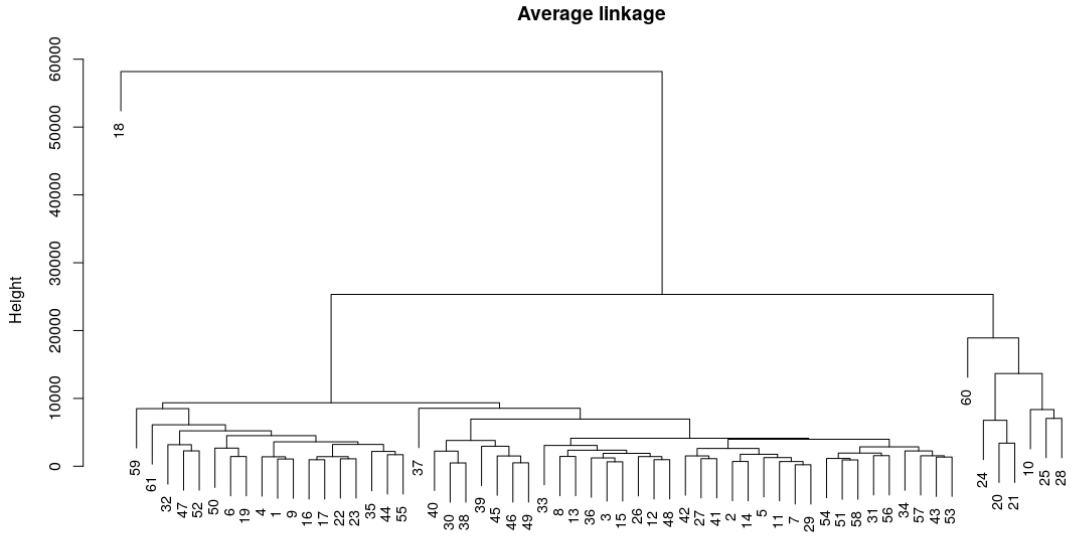

Fig. 6: задание №3

```
h: 0,5
Points: [3,20]
```

Используемая программа для вычислений и построения графиков: **Excel**

1) Ядро Гаусса

Вычисляем точки с помощью Ядра Гаусса, пример кода:

```
=EXP(-((($D2-H$3)/$F$1)^2)/2)/SQRT(2*PI())/$F$1
```

Далее вычисляем среднее значения(Density), пример кода:

```
=AVERAGE(H4:H78)
```

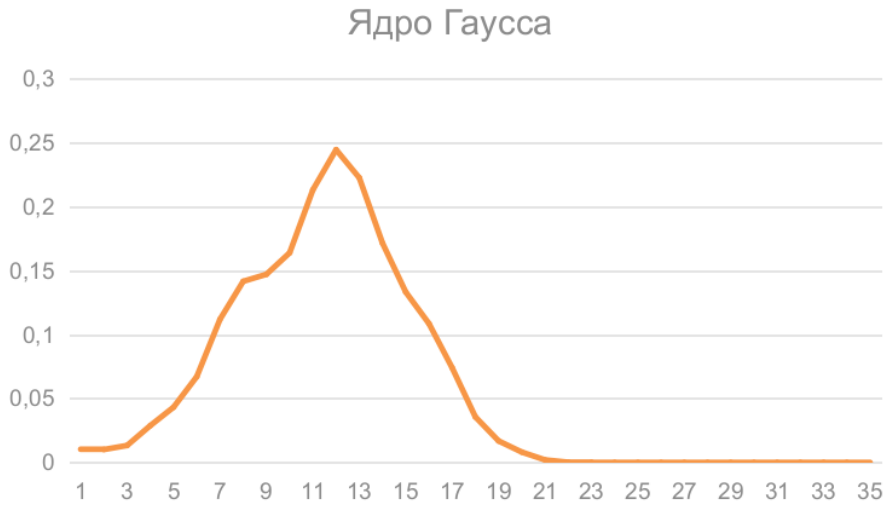

Fig. 7: Ядро Гаусса для задании №4

2) Ядро Епанечникова

Вычисляем точки с помощью Ядра Епанечникова, пример кода:

```
=IF((1-(($A4-H$3)/$F$1)^2)>0;(0,75*(1-((($A4-H$3)/$F$1)^2))/$F$1;0)
```

Далее вычисляем среднее значения(Density), пример кода:

```
=AVERAGE(H3:H77)
```

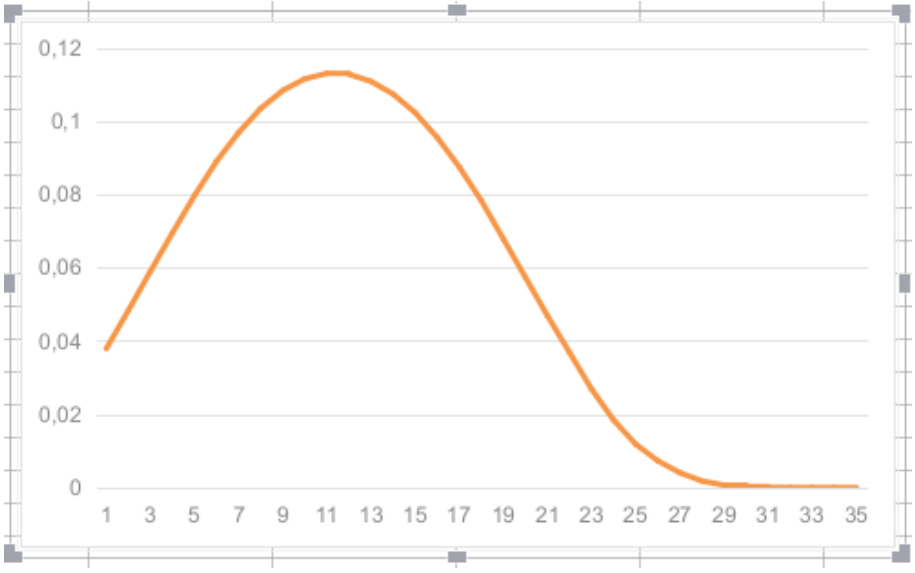

Fig. 8: Ядро Епанечникова для задании №4

3) Треугольное ядро

$$K(r) = T(r) = (1 - |r|) * I\{|r| < 1\}$$

Вычисляем точки с помощью Треугольного ядра, пример кода:

```
=IF(ABS(($D2-H$3)/$F$2)<1;1-ABS(($D2-H$3)/$F$2);0)
```

Далее вычисляем среднее значения(Density), пример кода:

```
=AVERAGE(H4:H78)
```

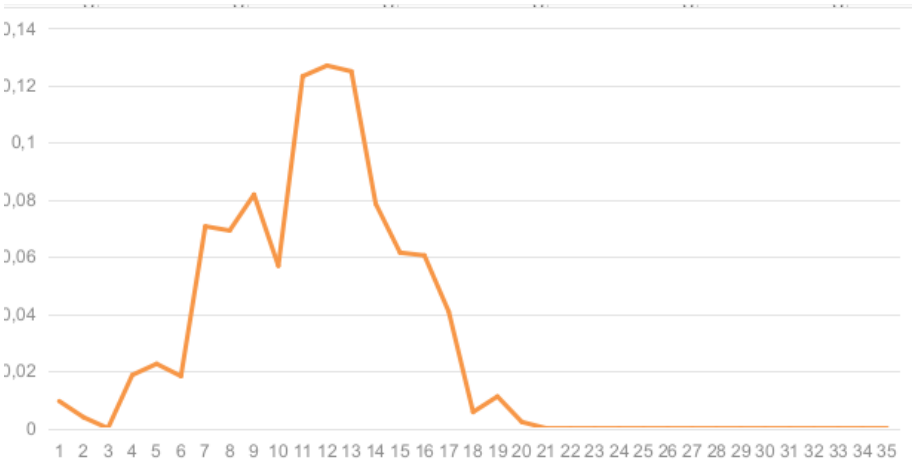

Fig. 9: Треугольное ядро для задании №4

По полученным графикам (Fig.7, Fig.8 и Fig.9), можно сделать несколько замечаний, разместим их по призовым местам:

- И так третье место по подробной информации занимает.... барабанная дробь..... бадам тссс... Треугольное ядро (Fig.9) по сравнению с другими ядрами визуально показывает более рельефную картину и менее подробную
- Остались два претендента и какое ядро займет второе место?..... и снова барабанная дробь.... бадам тссс.... ядро Гаусса (Fig.7). Данный участник намного лучше показал себя, чем предыдущий. Постарался дать нам менее рельефную линию на графике и приблизился к плотности данных. За это он получает честно свое второе место. Возможно он бы догнал участника находящегося на первом месте, если бы шаг h был бы меньше, чем исходный (предложенный мной).

- Лидирующее место достойно занял график ядра Епанечникова (Fig.8) . Он показал себя лучше всех (из предложенных), дал подробную информацию: плавную (непрерывную) линию, что намекает нам на нормальное (гауссово) распределение данных.

Хотелось разнообразить повествование результатов, надеюсь это больно не отразиться на оценке за моё хромающее кхм-кхм чувство юмора :)

Задание 5

Используйте набор данных «Weekly» из пакета «ISLR». Постройте модель линейного дискриминантного анализа, используя в качестве единственного предиктора переменную «Lag2». Проинтерпретируйте полученные результаты.

Используемая программа для вычислений и построения графиков: **Rstudio**
Информация о исходных данных Weekly:

```
Weekly S&P Stock Market Data
Description: Weekly percentage returns for the S&P 500 stock index between 1990 and 2010.
Format: A data frame with 1089 observations on the following 9 variables.

Year: The year that the observation was recorded

Lag1: Percentage return for previous week

Lag2: Percentage return for 2 weeks previous

Lag3: Percentage return for 3 weeks previous

Lag4: Percentage return for 4 weeks previous

Lag5: Percentage return for 5 weeks previous

Volume: Volume of shares traded (average number of daily shares traded in billions)

Today: Percentage return for this week

Direction: A factor with levels Down and Up indicating whether the market
 had a positive or negative return on a given week
```

Подключаем пакет и данные Weekly:

```
> library(ISLR)
> ?Weekly
> head(Weekly)
  Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

Данные делятся на два класса (по Direction), которые растут(Up) и которые падают(Down). Предиктор один: Lag2 (Percentage return for 2 weeks previous)

Подключаем функцию линейного дискриминантного анализа:

```
> lda.fit = lda(Direction ~ Lag2,data = Weekly)
> lda.fit

Call:
lda(Direction ~ Lag2, data = Weekly)

Prior probabilities of groups:
     Down        Up
0.4444444 0.5555556

Group means:
            Lag2
Down -0.04042355
Up    0.30428099

Coefficients of linear discriminants:
           LD1
Lag2 0.4251523
```

```
> lda.pred = predict(lda.fit,Weekly)
> ldahist(data = lda.pred$x[,1], Weekly$Direction )
```

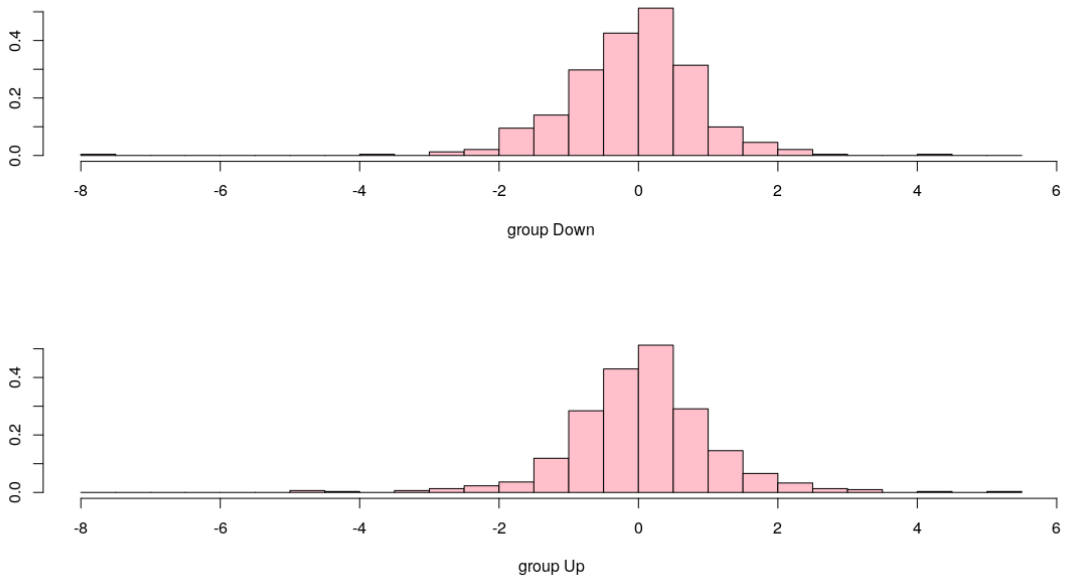

Fig. 10: для задании №5

Дискриминантная функция представляется в виде умножения

$$LD1 = 0.4251523 \cdot Lag2$$

Вероятность группы Up (55%) больше, чем Down (44%) на 11%