Testy

Krzysztof Tabeau

12/05/2020

1. Wstęp

W poniższych testach jest zastosowany następujący schemat. Dla każdej funkcji agregującej, dla jednej z 3 metryk (1,2,INF) dla jednego z trzech zbiorów danych (affairs, auto_ord, glass) jest oblicznana ramka danych w postaci błędów dla różnych k (1,3,5,7,9,11,13,15,17,19). Do tego jest wykres obrazujący dane. Dodatkowo jest obliczana próba 1-nn, gdzie dla k=1 próba ucząca i testowa są te same. Wnioski znajdują się na końcu.

2. Potrzebne dane

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
source("knn_pomocnicze.R")
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

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

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

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

```
source("funkcje_agregujace.R")
source("knn.R")
source("bledy.R")
affairs <- read.csv("https://www.gagolewski.com/resources/data/ordinal-regression/affairs.csv")
auto_ord <- read.csv("https://www.gagolewski.com/resources/data/ordinal-regression/auto_ord.csv")
glass <- read.csv("https://www.gagolewski.com/resources/data/ordinal-regression/glass.csv")
```

3. Srednia (L2, affairs)

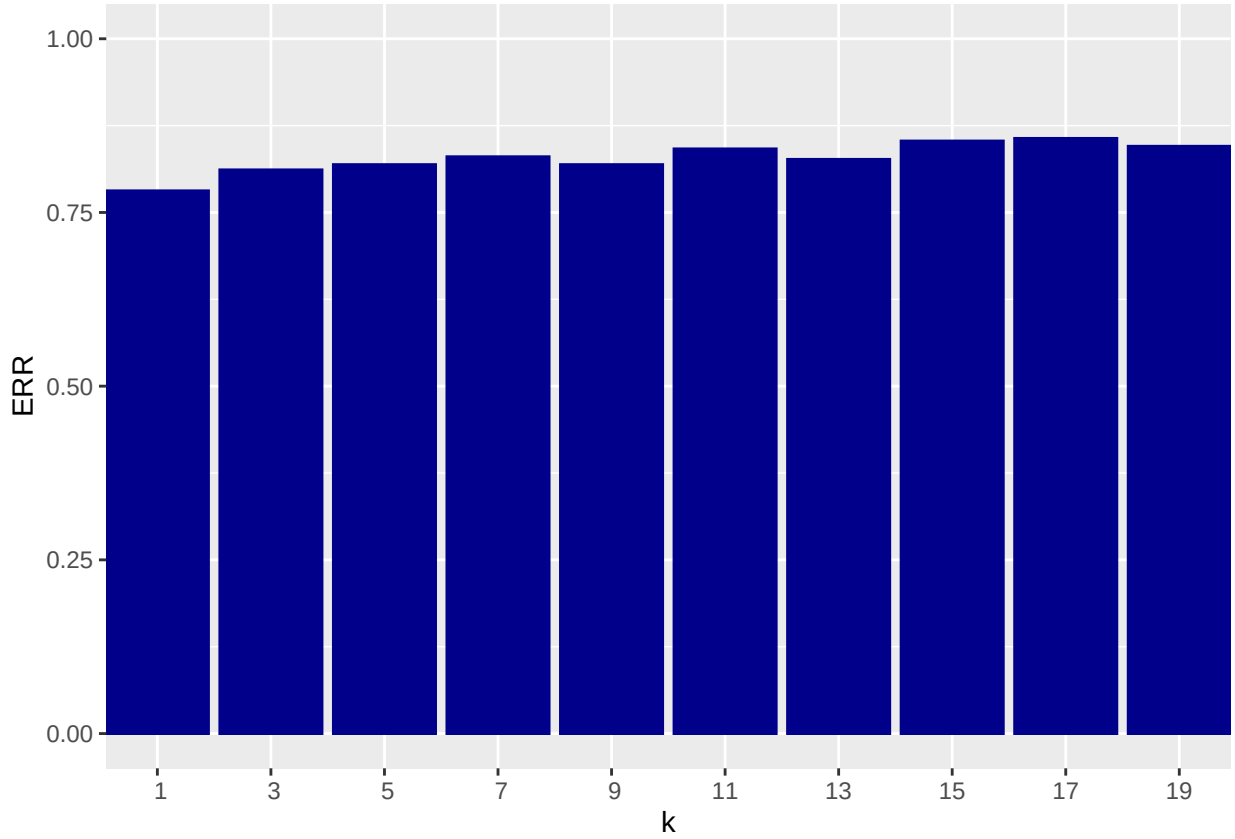
```
tab <- przetworz_test(affairs,2, FN = srednia_a)
print(tab)
```

```
##      k       ERR      MAD      MSE
## 1    1 0.7811321 2.101887 7.550943
## 2    3 0.8113208 1.739623 4.788679
## 3    5 0.8188679 1.671698 4.328302
## 4    7 0.8301887 1.626415 4.018868
## 5    9 0.8188679 1.528302 3.626415
## 6   11 0.8415094 1.615094 3.871698
## 7   13 0.8264151 1.562264 3.660377
## 8   15 0.8528302 1.667925 4.083019
## 9   17 0.8566038 1.645283 3.811321
## 10 19 0.8452830 1.600000 3.713208
```

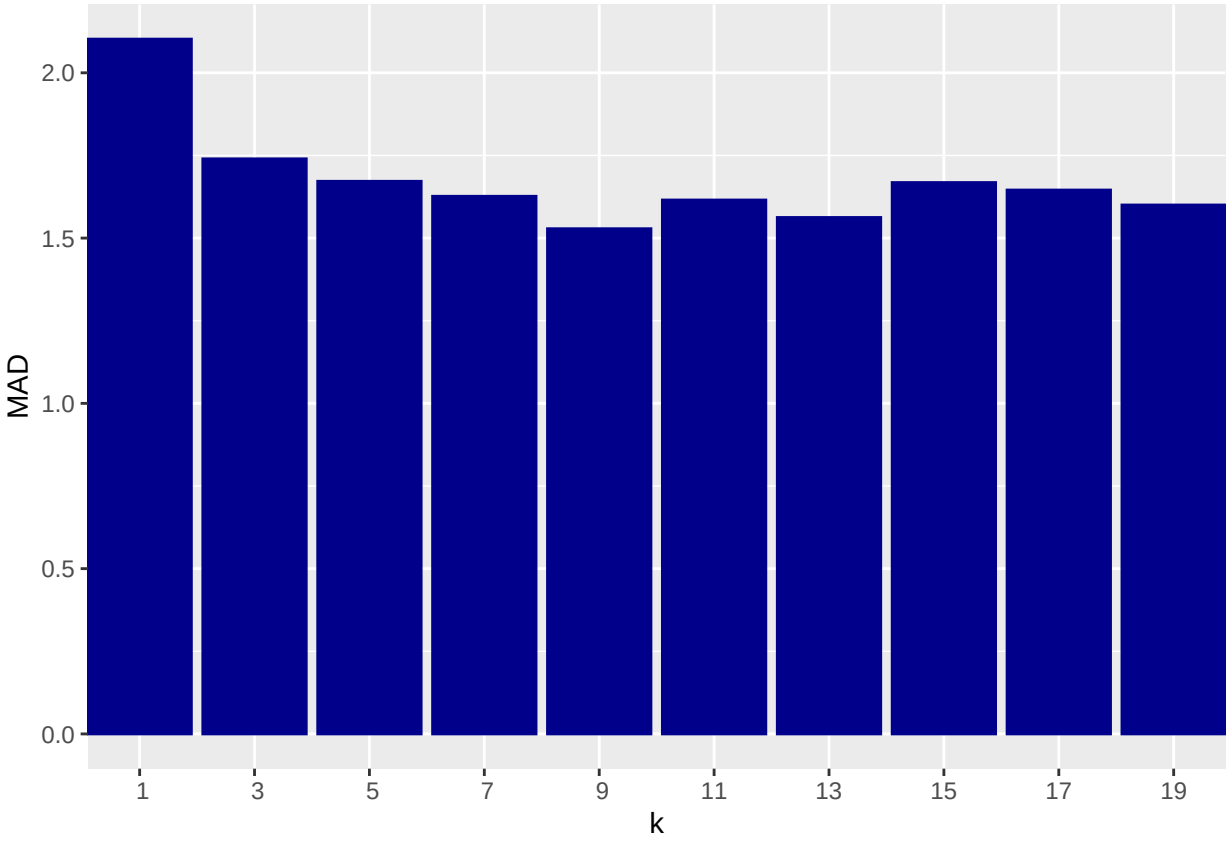
```
przetworz_1nn(affairs,2, FN = srednia_a )
```

```
##             ERR      MAD      MSE
## [1,] 0.4226415 1.154717 4.158491
```

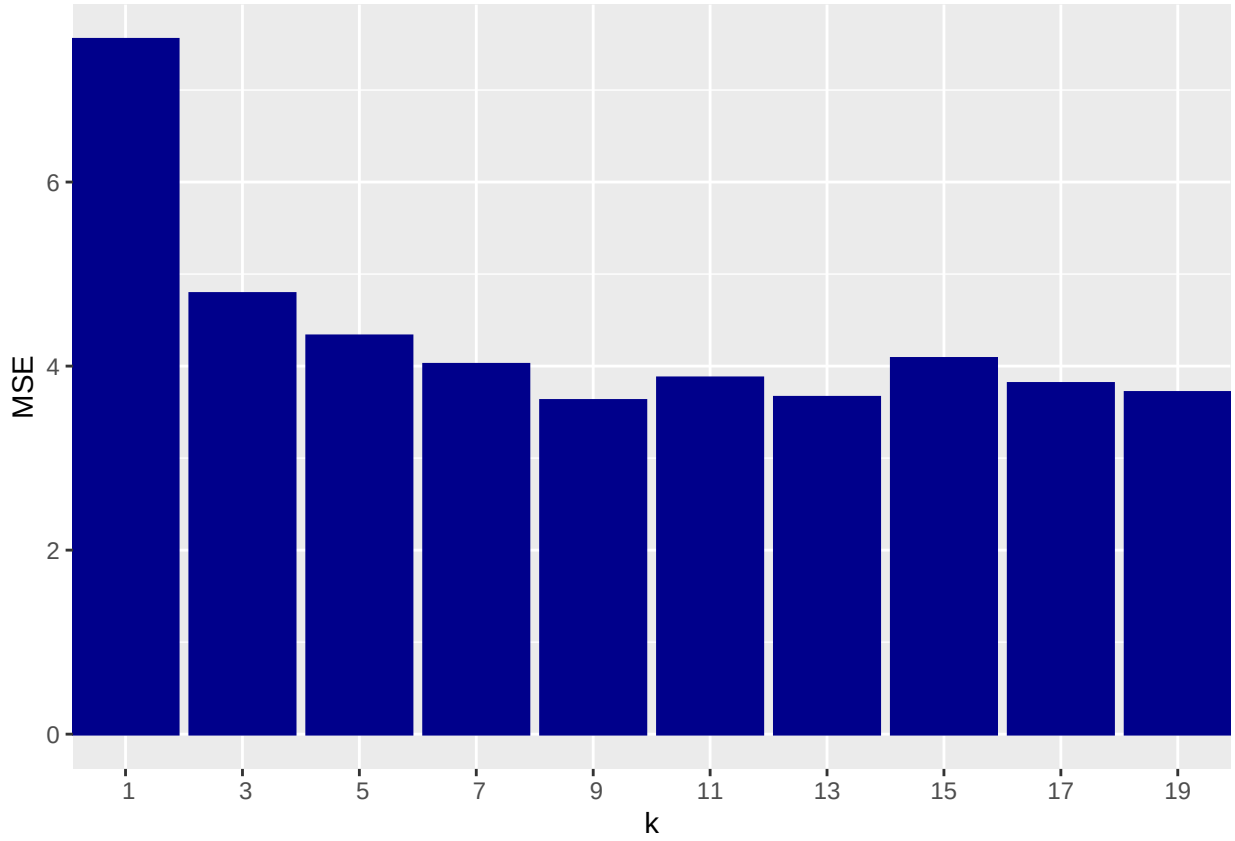
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```



```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



4. Moda (L(INF), glass)

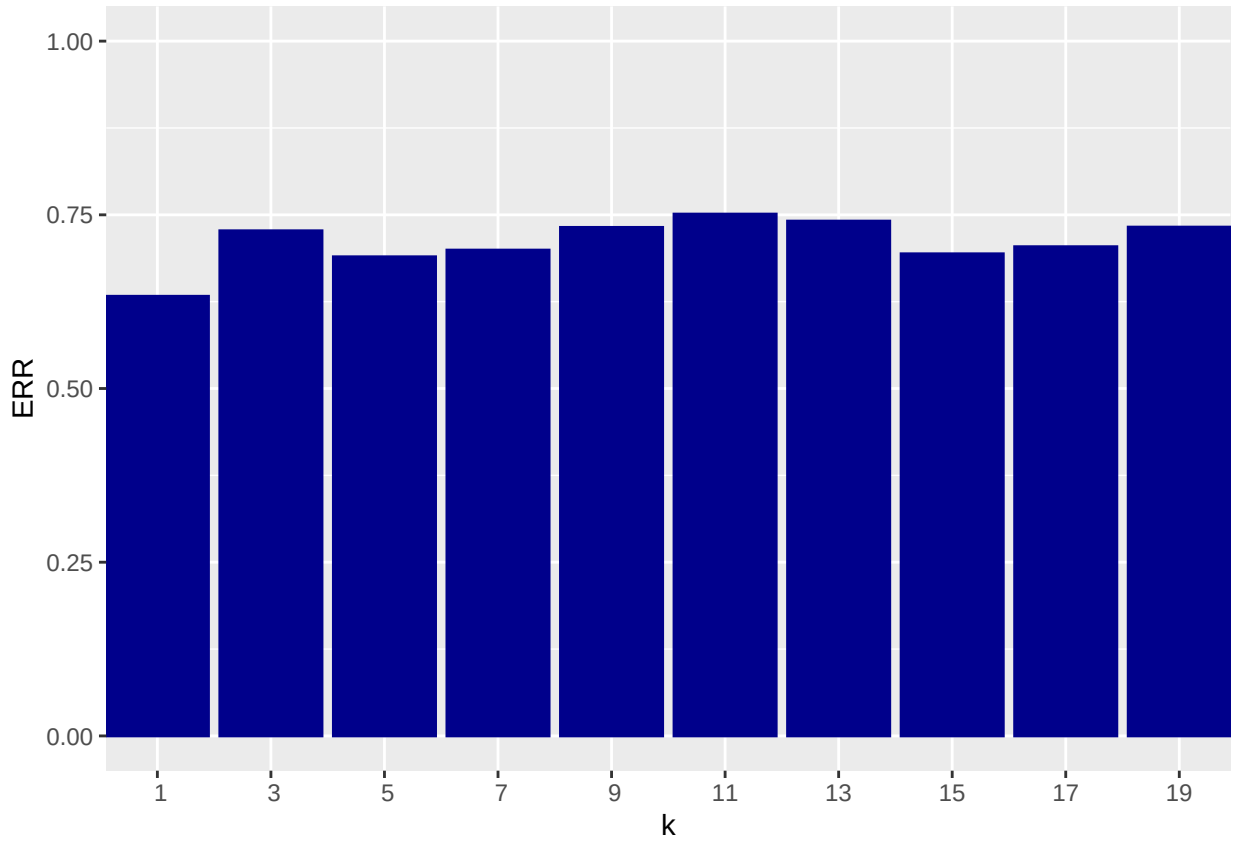
```
tab <- przetworz_test(glass,Inf, FN = moda)
print(tab)
```

```
##      k       ERR      MAD      MSE
## 1    1 0.6327796 1.221927 2.976966
## 2    3 0.7270210 1.474308 4.299779
## 3    5 0.6895903 1.450609 4.387265
## 4    7 0.6992248 1.465781 4.358140
## 5    9 0.7318937 1.474640 4.272425
## 6   11 0.7510520 1.550166 4.619048
## 7   13 0.7409745 1.530565 4.581063
## 8   15 0.6939092 1.450720 4.341750
## 9   17 0.7040975 1.479623 4.501107
## 10  19 0.7323367 1.488815 4.491030
```

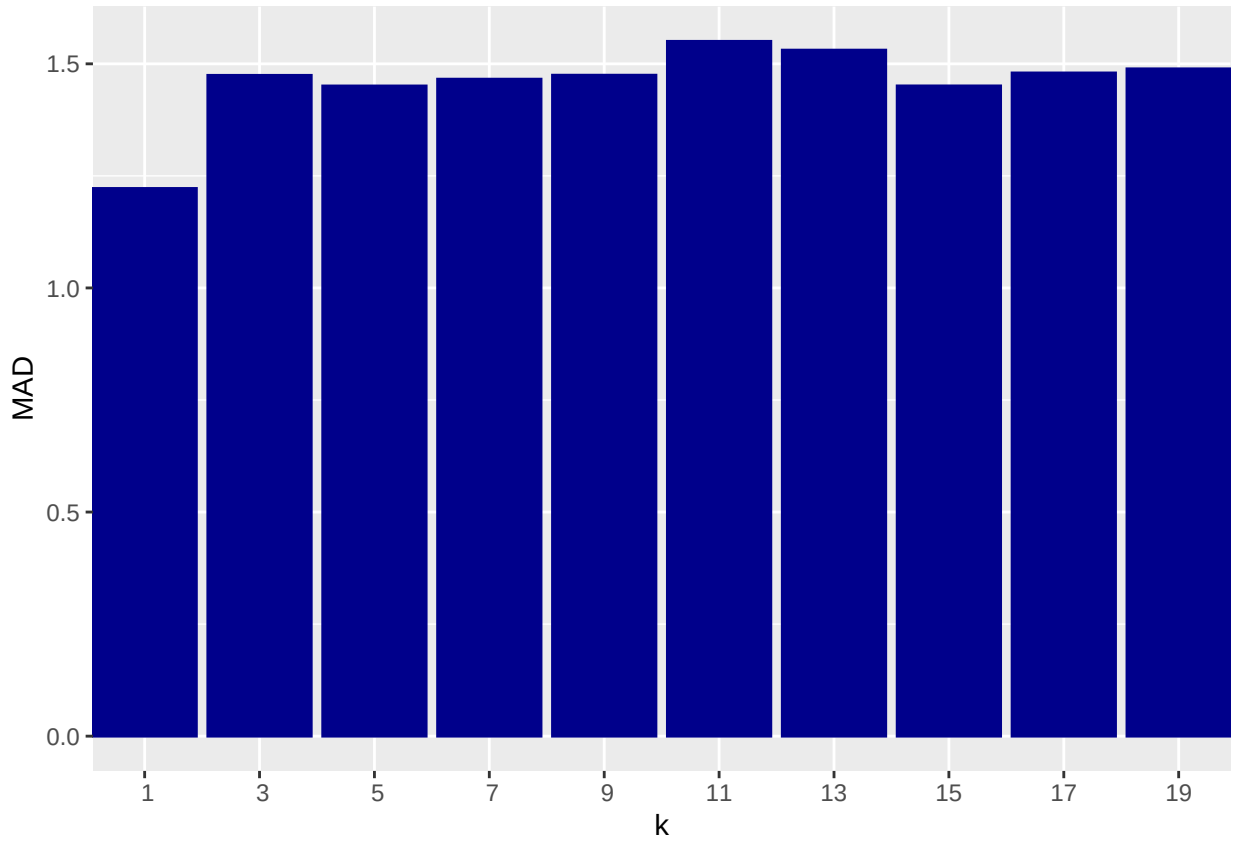
```
przetworz_1nn(glass,Inf, FN = moda )
```

```
##      ERR MAD MSE
## [1,]   0   0   0
```

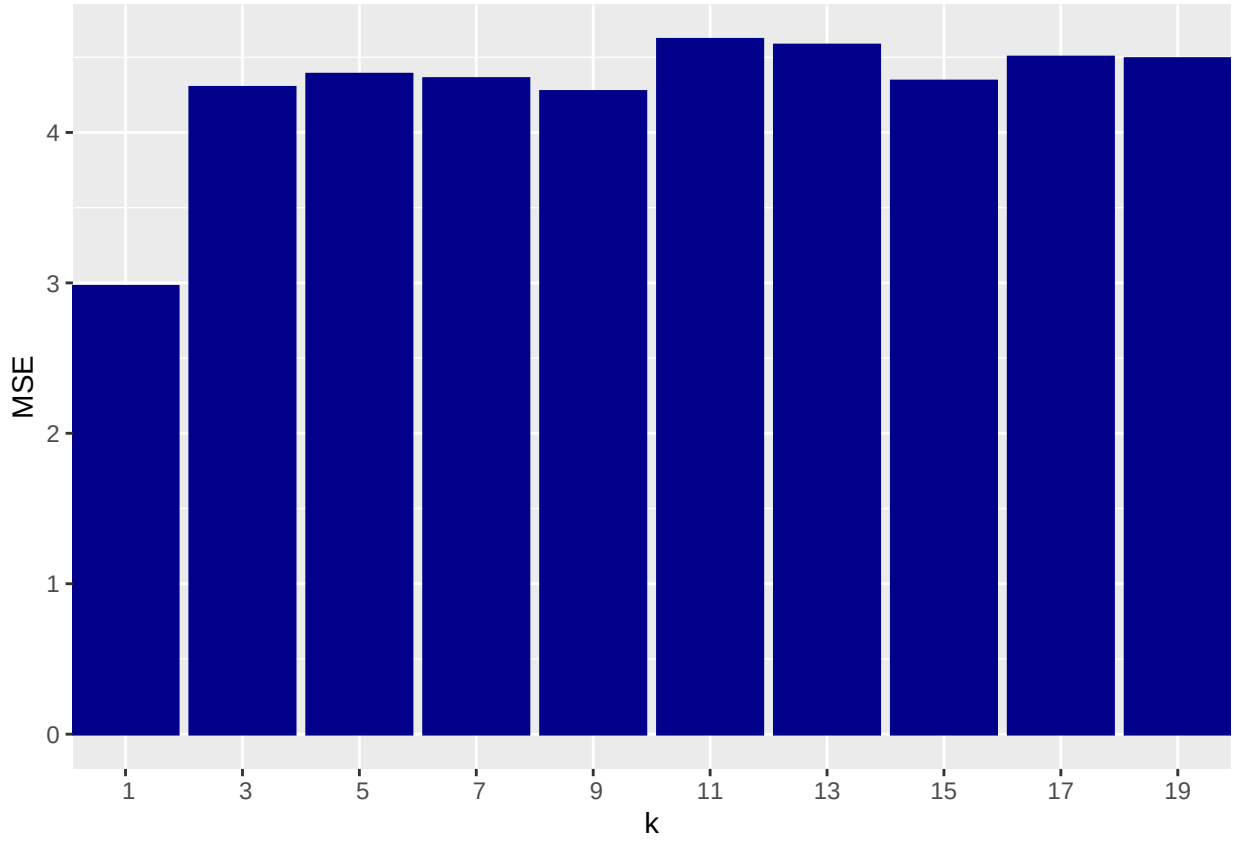
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```




```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



5. Mediana (L1, auto_ord)

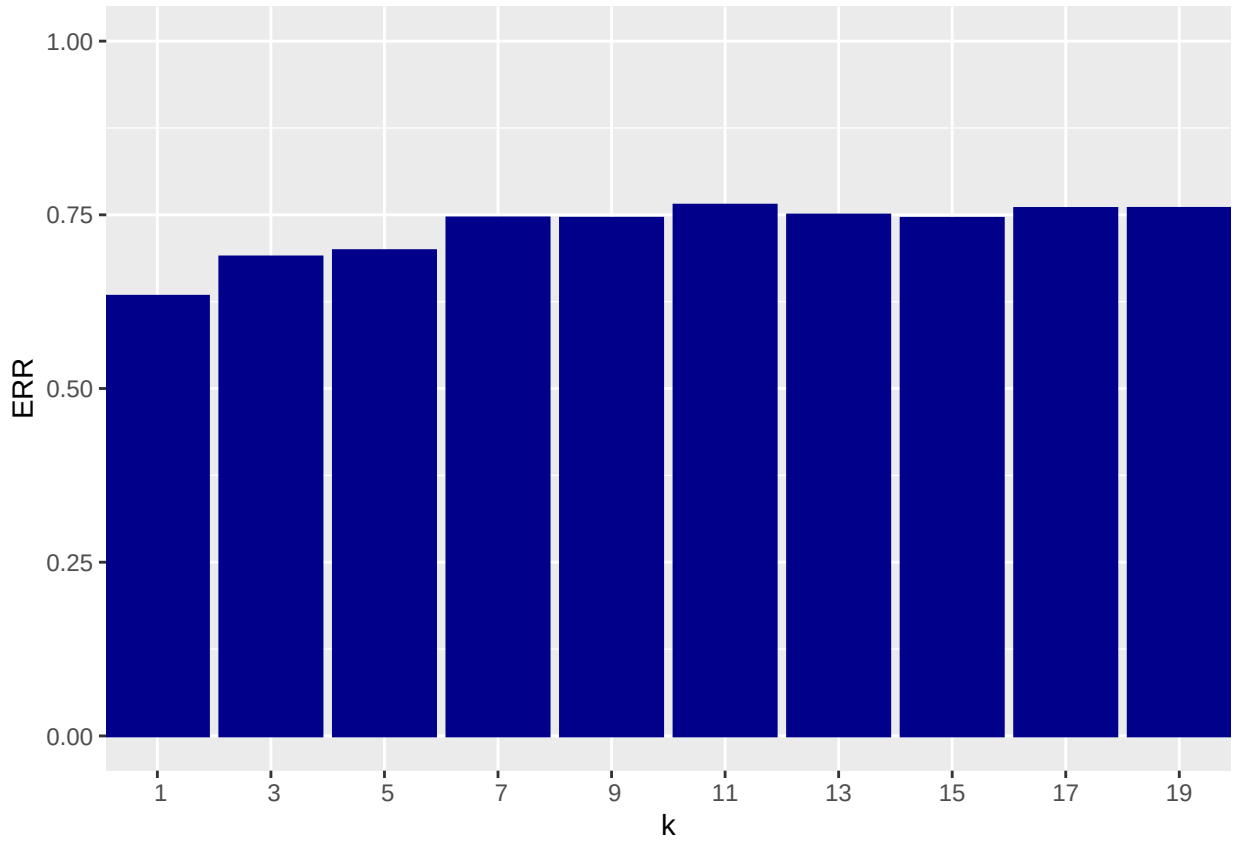
```
tab <- przetworz_test(glass,Inf, FN = mediana)
print(tab)
```

```
##     k       ERR      MAD      MSE
## 1   1 0.6327796 1.221927 2.976966
## 2   3 0.6893688 1.333998 3.488594
## 3   5 0.6984496 1.422481 4.006312
## 4   7 0.7455150 1.436877 3.890033
## 5   9 0.7450720 1.436656 3.853267
## 6  11 0.7638981 1.492913 4.012071
## 7  13 0.7497231 1.464563 3.955371
## 8  15 0.7449612 1.468992 3.997010
## 9  17 0.7591362 1.455039 3.861683
## 10 19 0.7592470 1.450388 3.866113
```

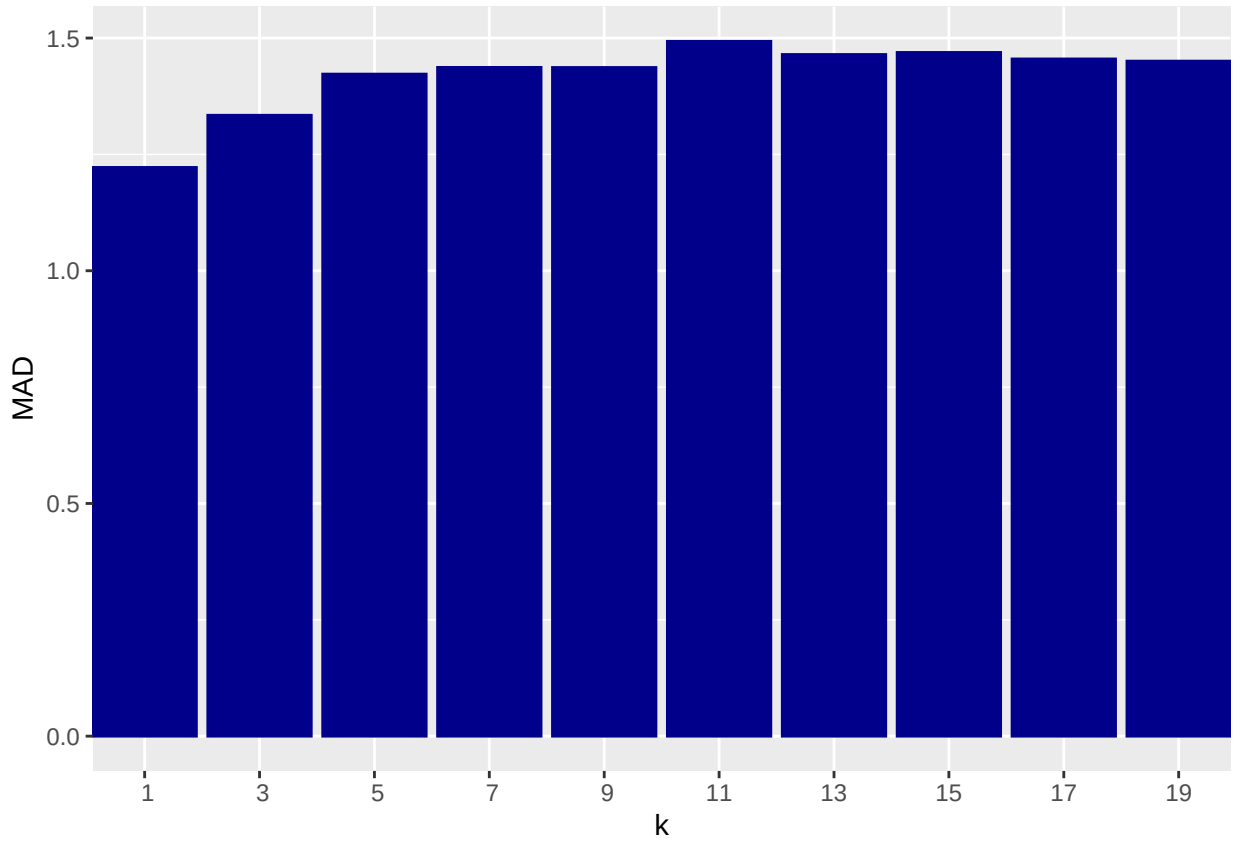
```
przetworz_1nn(glass,Inf, FN = mediana )
```

```
##      ERR MAD MSE
## [1,]   0   0   0
```

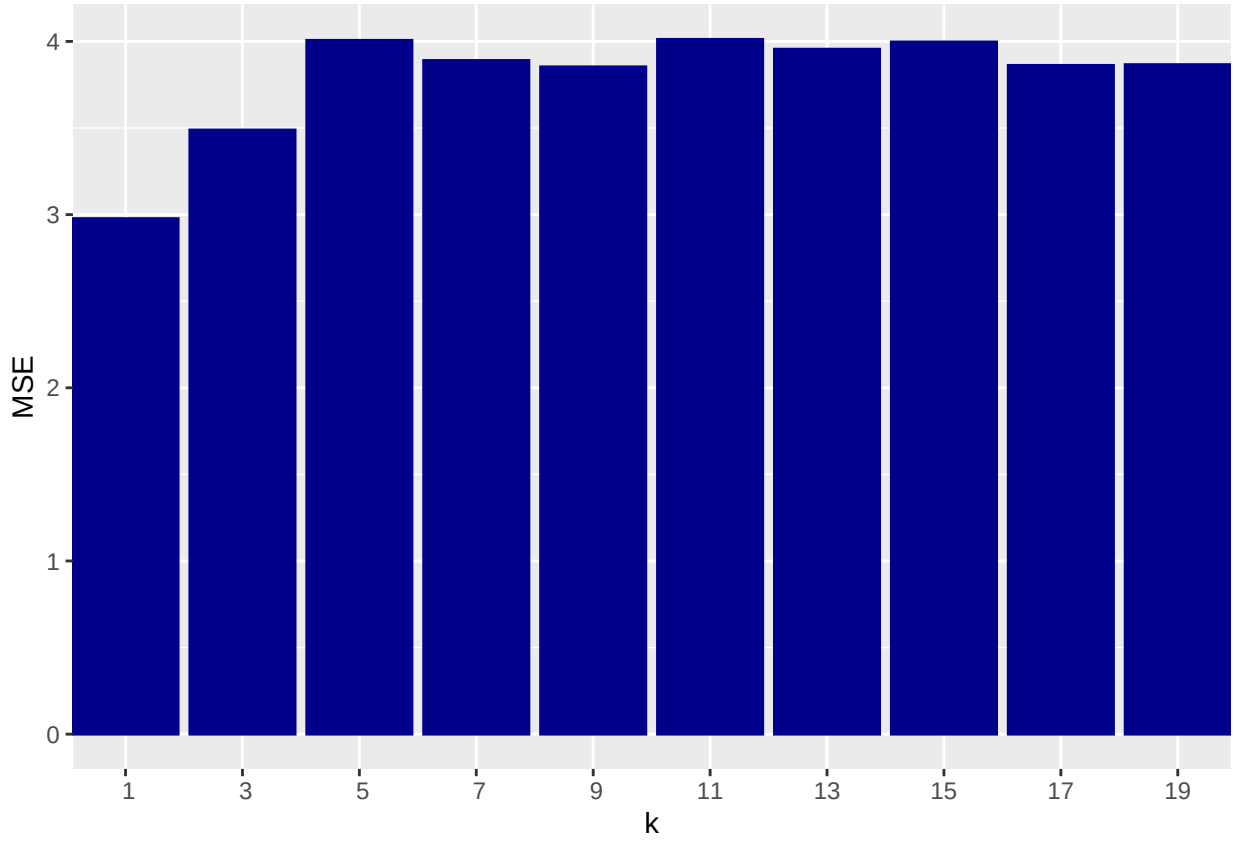
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```




```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



6. Minkara1.5 (L1, affairs)

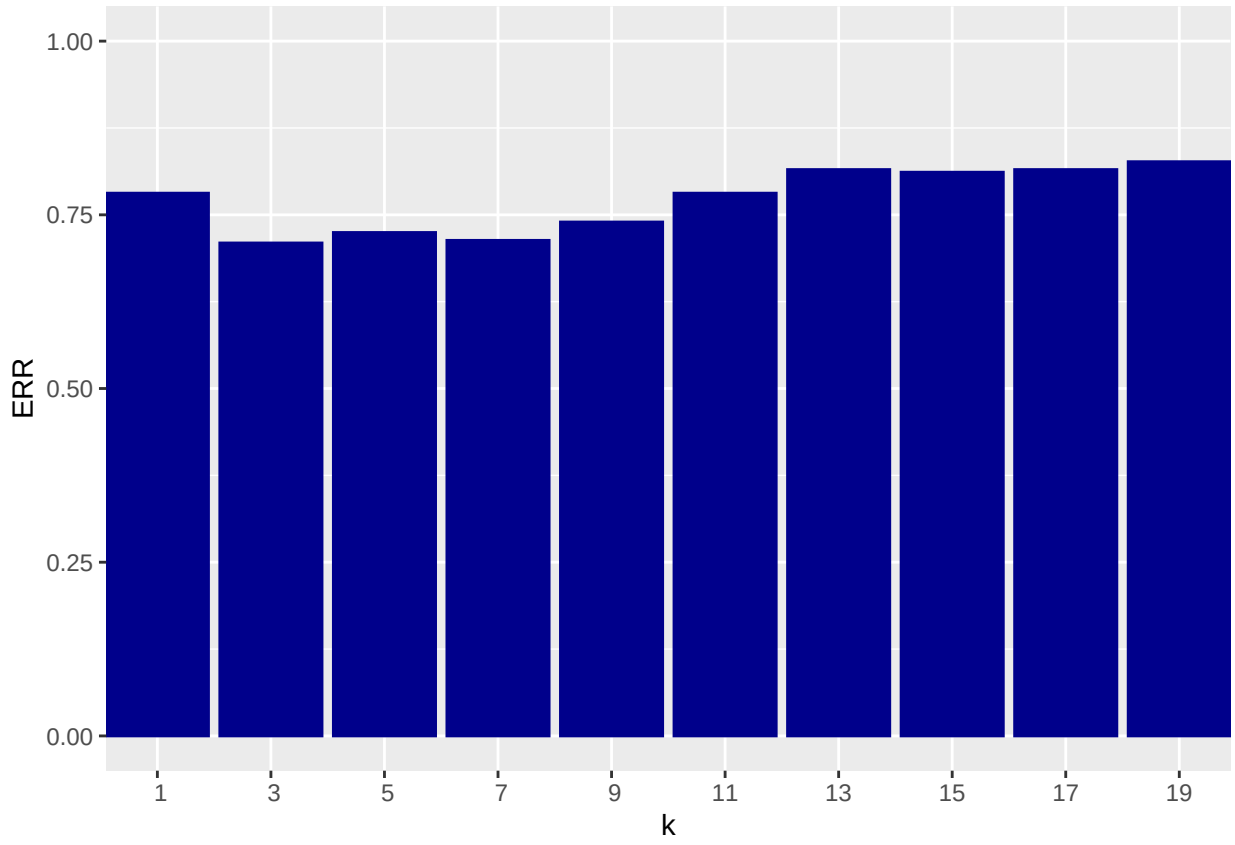
```
tab <- przetworz_test(affairs,1, FN = minkara1.5)
print(tab)
```

```
##     k       ERR      MAD      MSE
## 1   1 0.7811321 2.101887 7.550943
## 2   3 0.7094340 1.849057 6.332075
## 3   5 0.7245283 1.713208 5.501887
## 4   7 0.7132075 1.641509 5.098113
## 5   9 0.7396226 1.566038 4.486792
## 6  11 0.7811321 1.569811 4.211321
## 7  13 0.8150943 1.547170 3.916981
## 8  15 0.8113208 1.524528 3.788679
## 9  17 0.8150943 1.532075 3.735849
## 10 19 0.8264151 1.543396 3.739623
```

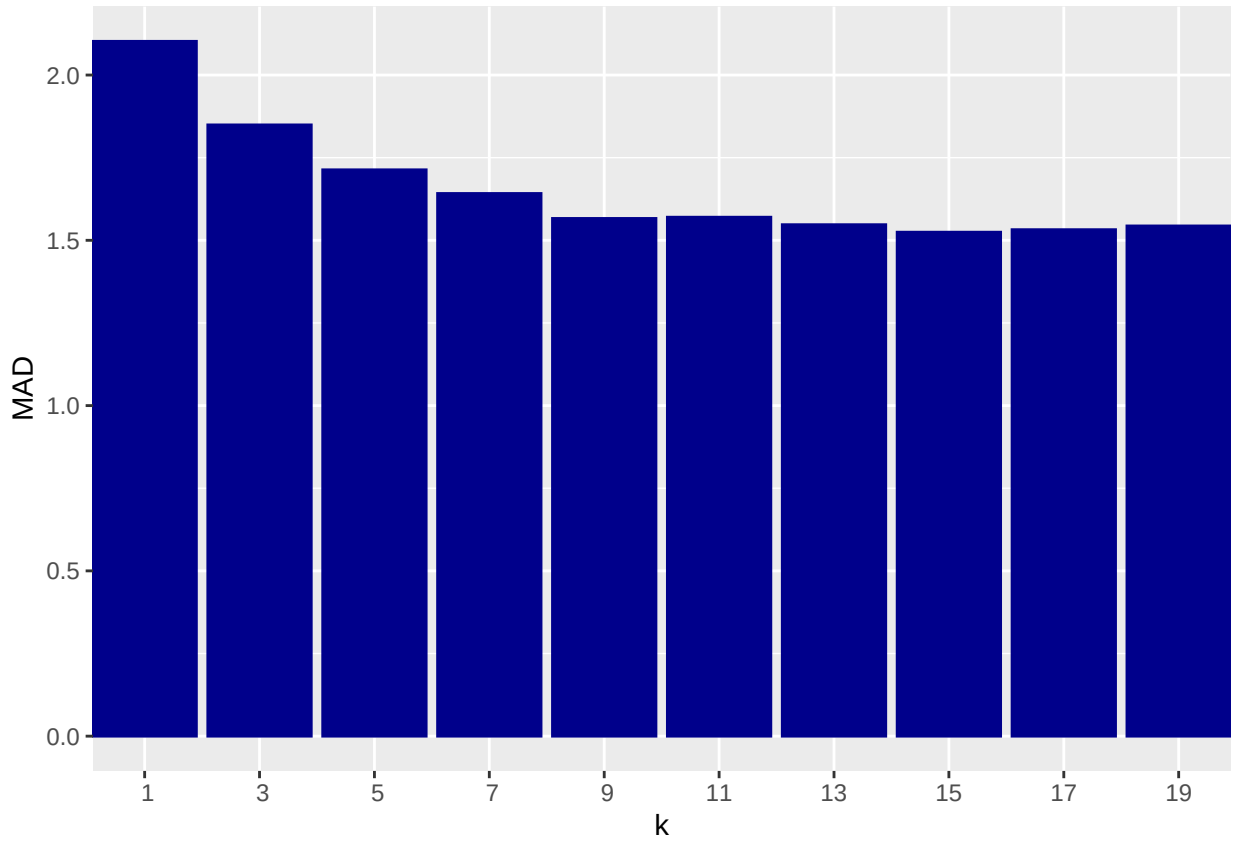
```
przetworz_1nn(affairs,1, FN = minkara1.5 )
```

```
##            ERR      MAD      MSE
## [1,] 0.4226415 1.154717 4.158491
```

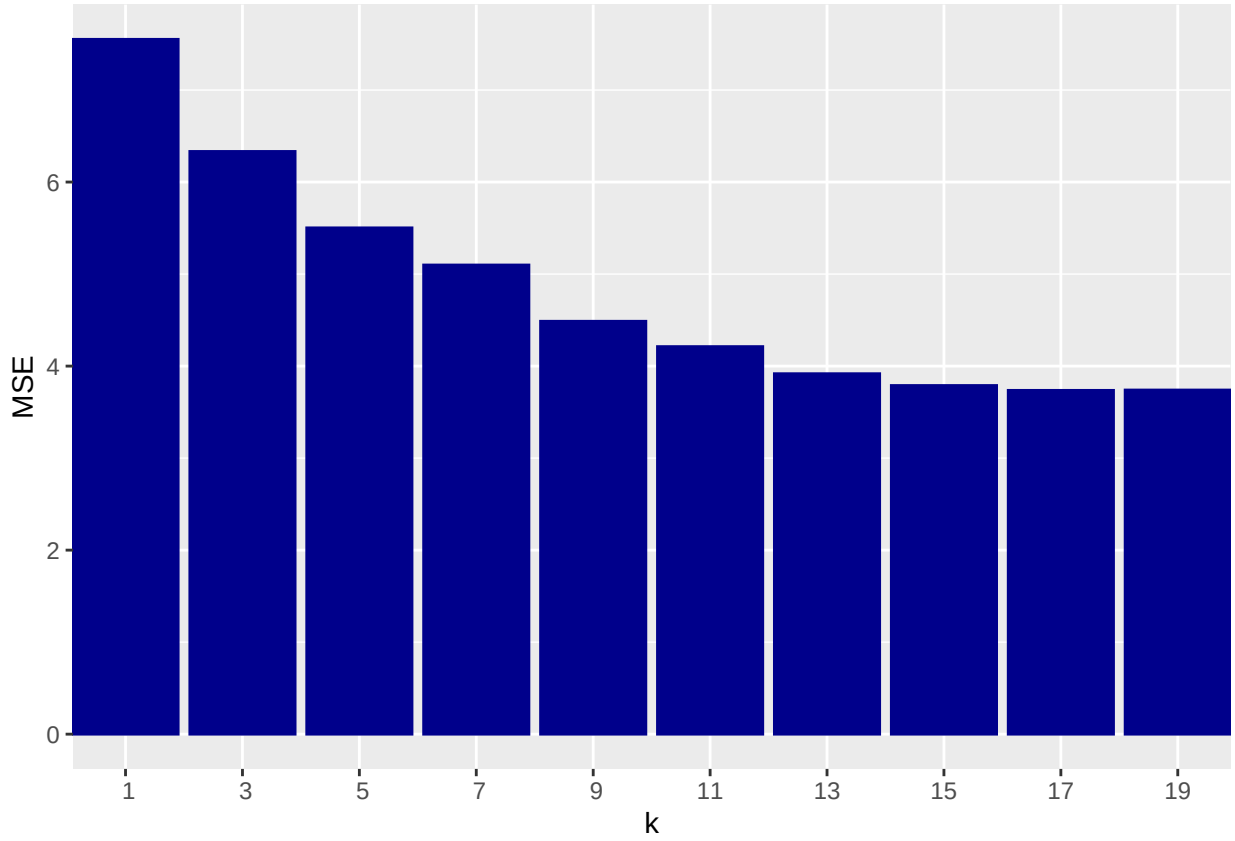
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```




```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



7. Minkara3 (L(Inf), affairs)

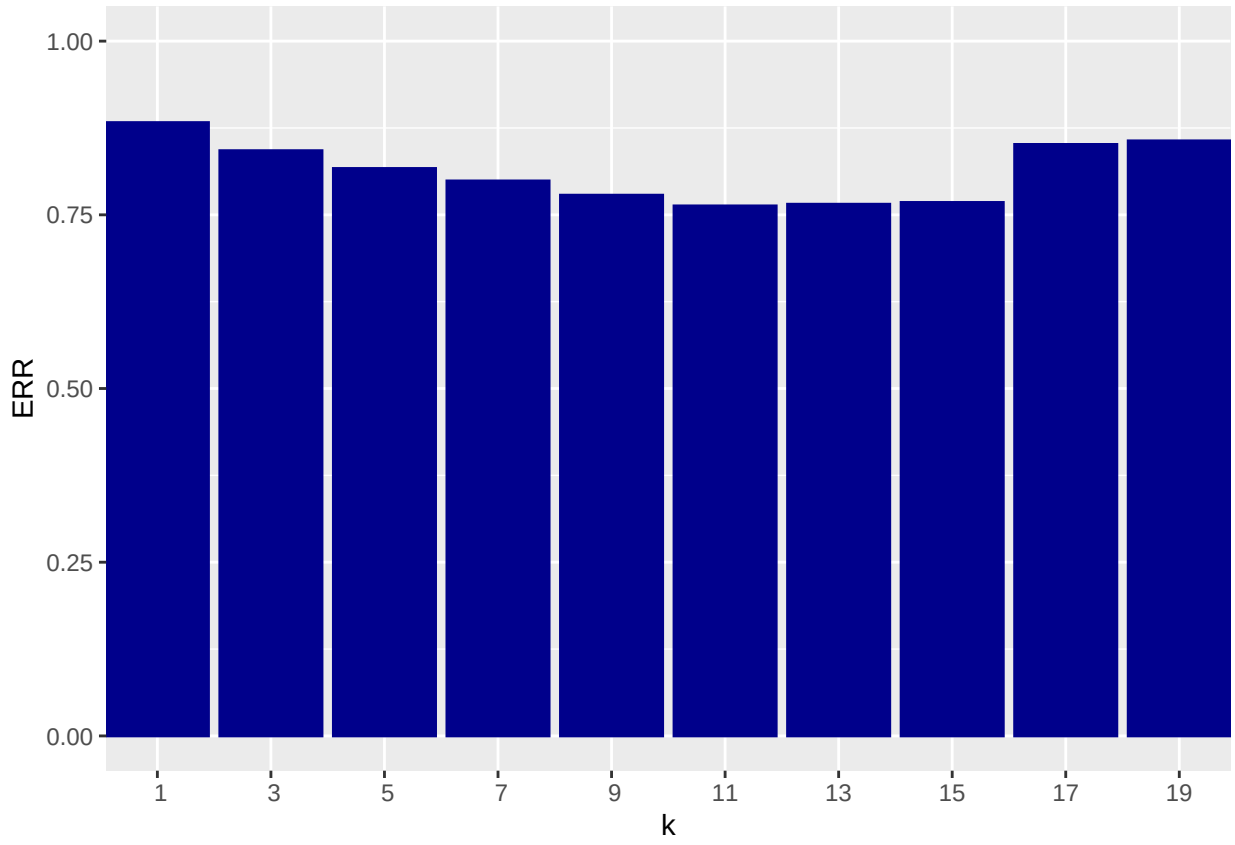
```
tab <- przetworz_test(auto_ord,Inf, FN = minkara3)
print(tab)
```

```
##     k       ERR      MAD      MSE
## 1   1 0.8827329 1.456086 3.167932
## 2   3 0.8423239 1.305648 2.633268
## 3   5 0.8166829 1.292730 2.650860
## 4   7 0.7987666 1.285070 2.638007
## 5   9 0.7782538 1.267056 2.660565
## 6  11 0.7628367 1.231353 2.559039
## 7  13 0.7652386 1.223401 2.520058
## 8  15 0.7677702 1.228497 2.540409
## 9  17 0.8513145 1.324473 2.656248
## 10 19 0.8564427 1.349951 2.727426
```

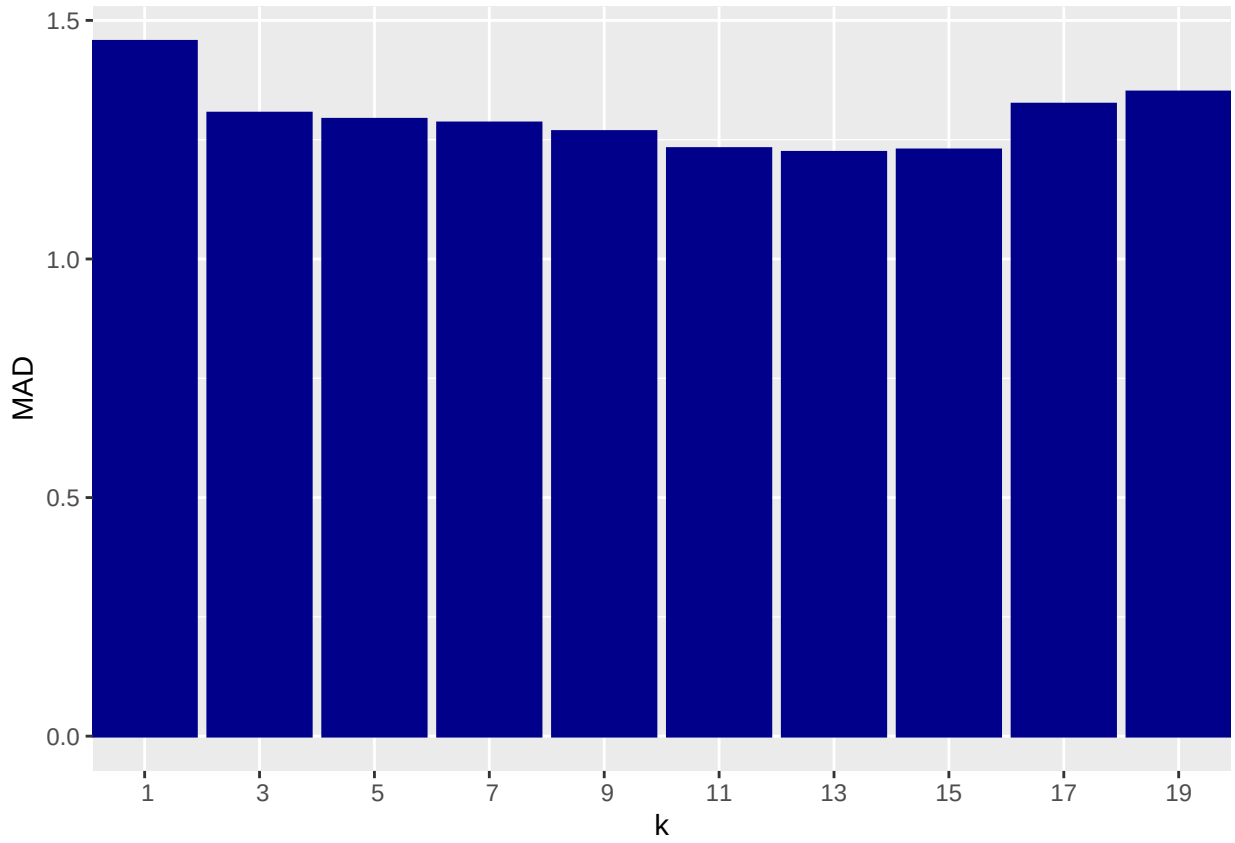
```
przetworz_1nn(auto_ord,Inf, FN = minkara3 )
```

```
##      ERR MAD MSE
## [1,]   0   0   0
```

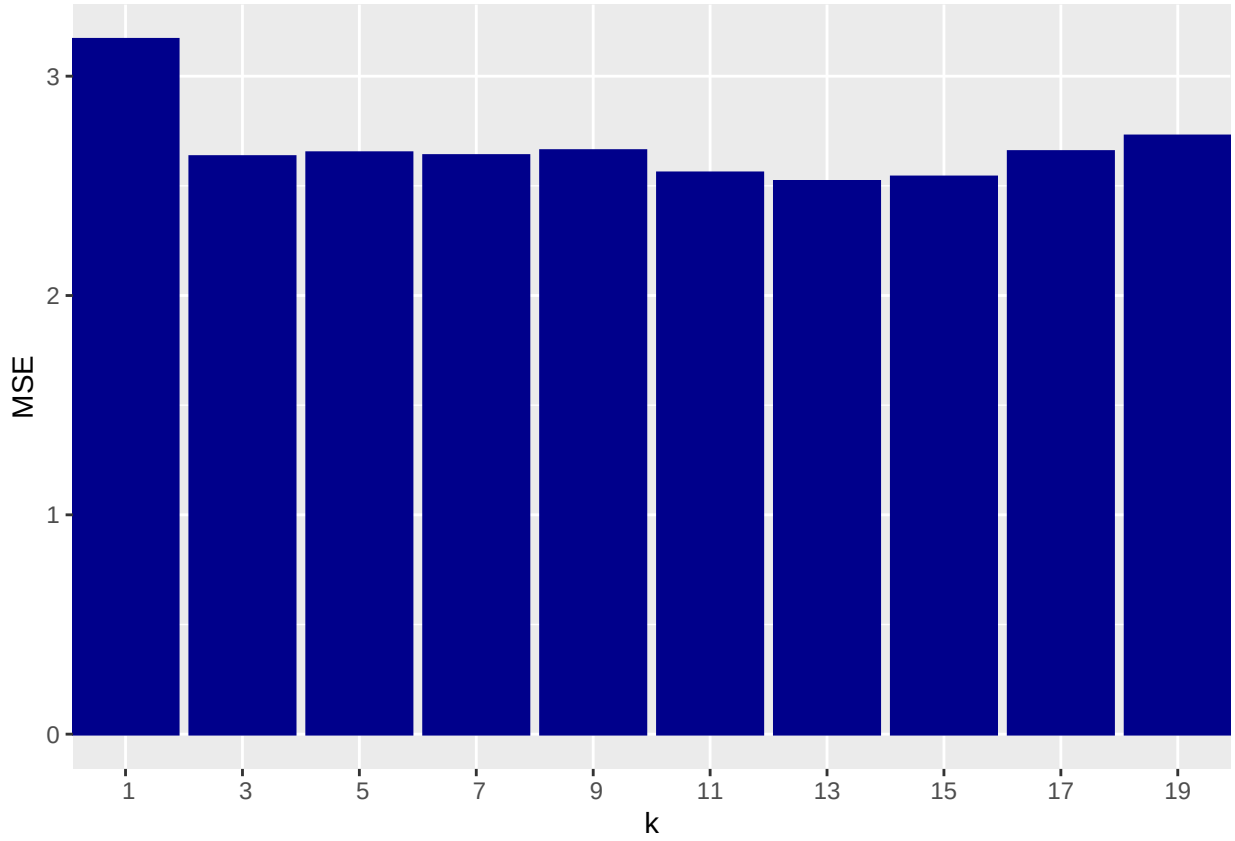
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```




```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



8. Minkara10 (L2, auto_ord)

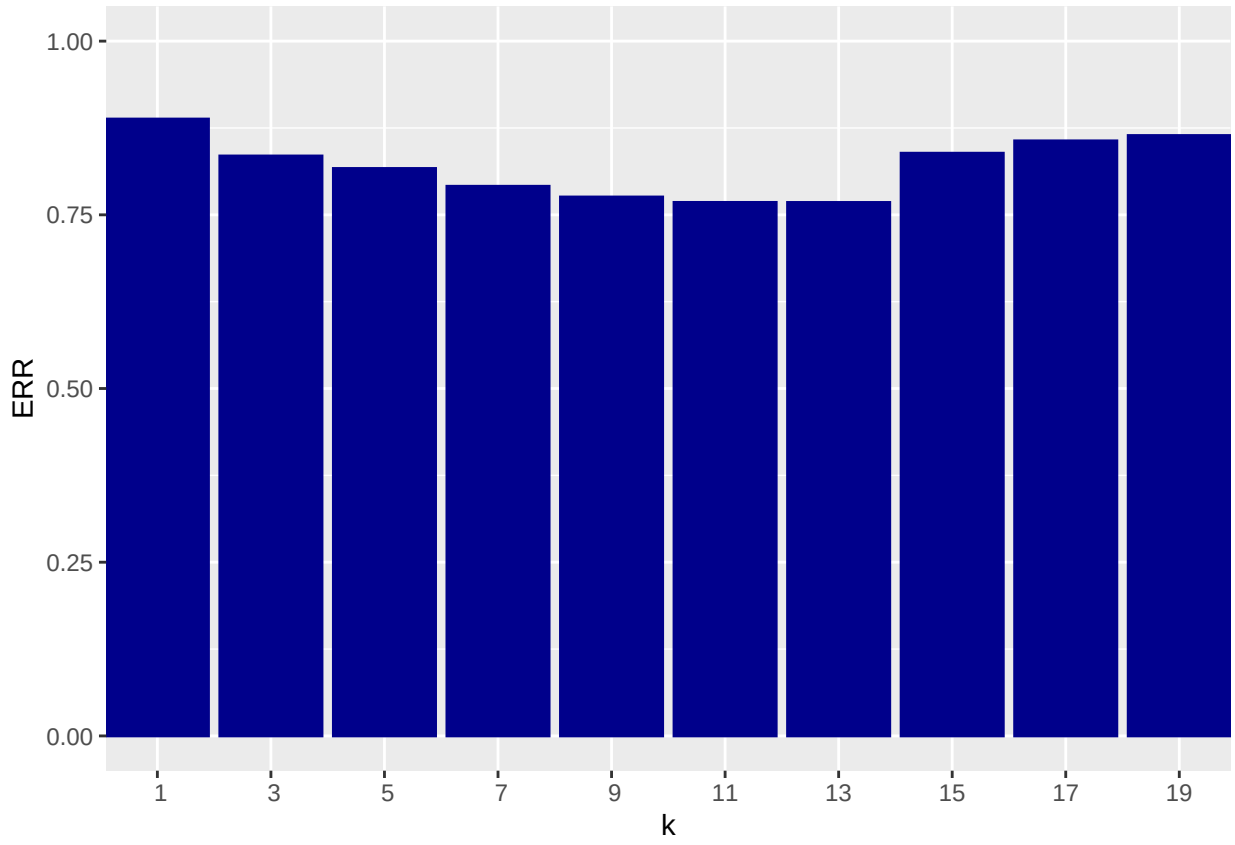
```
tab <- przetworz_test(auto_ord,2, FN = minkara10)
print(tab)
```

```
##     k       ERR      MAD      MSE
## 1   1 0.8878286 1.484258 3.278156
## 2   3 0.8346641 1.280331 2.557384
## 3   5 0.8166829 1.287861 2.626063
## 4   7 0.7910743 1.267218 2.640734
## 5   9 0.7756573 1.238883 2.581435
## 6  11 0.7678676 1.251542 2.650243
## 7  13 0.7678027 1.243817 2.627264
## 8  15 0.8386563 1.332392 2.736092
## 9  17 0.8564427 1.362772 2.801785
## 10 19 0.8641350 1.398410 2.928822
```

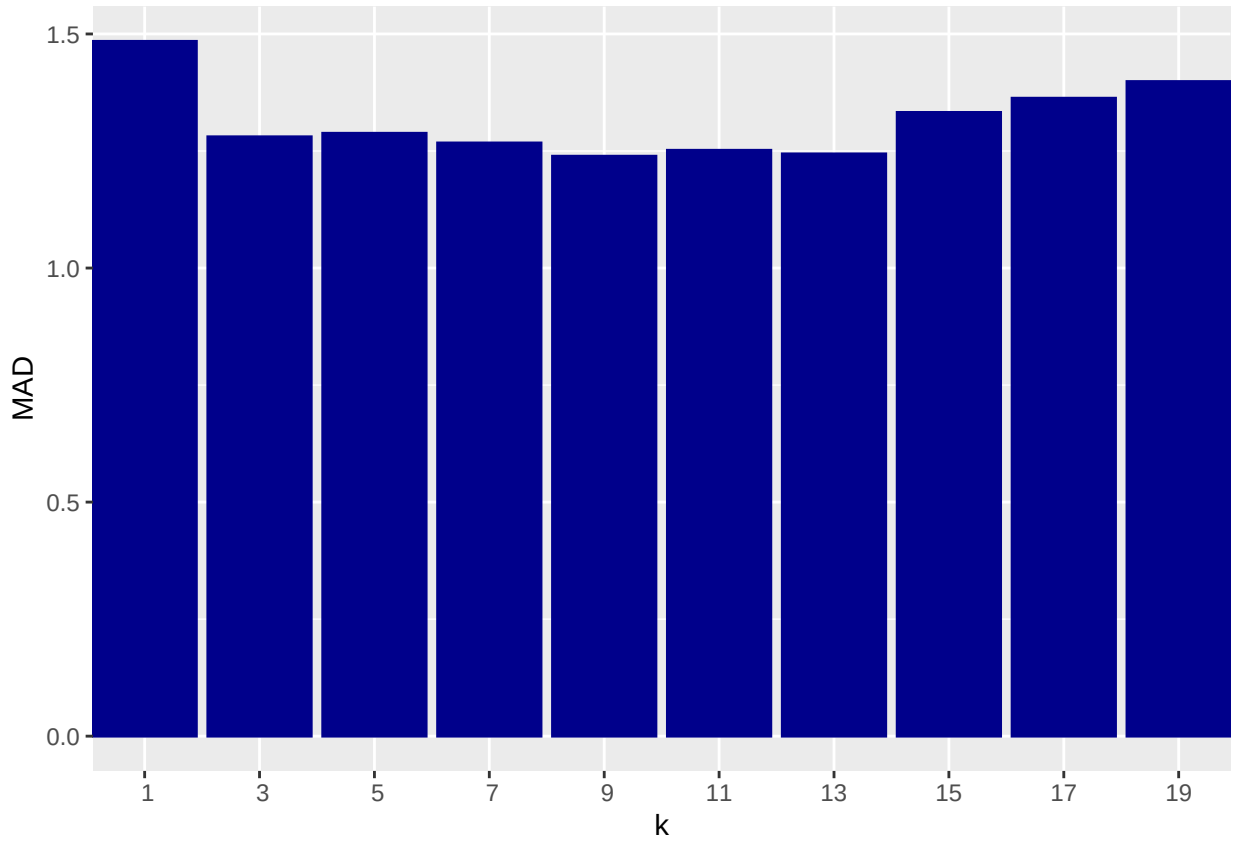
```
przetworz_1nn(auto_ord,2, FN = minkara10 )
```

```
##      ERR MAD MSE
## [1,]   0   0   0
```

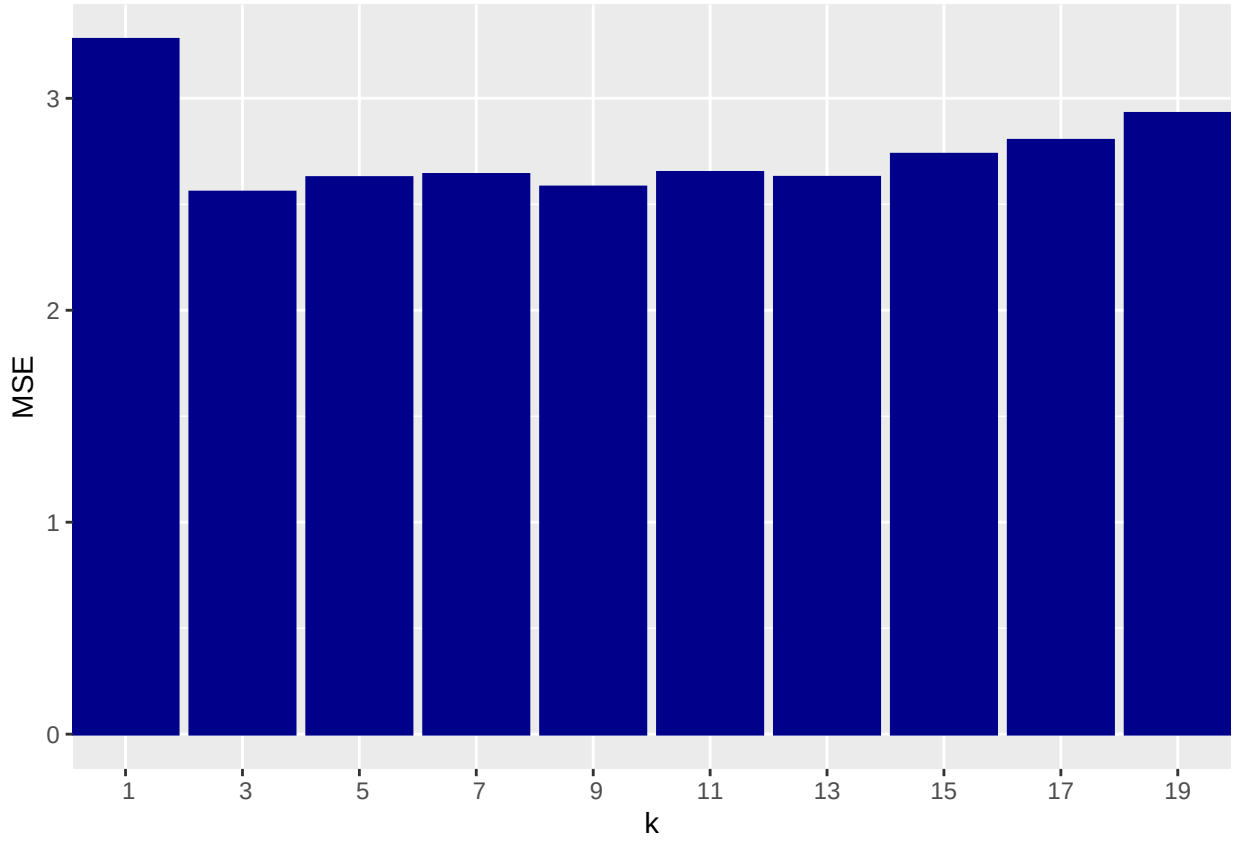
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```




```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



9. Srednia_parzytych (L2, glass)

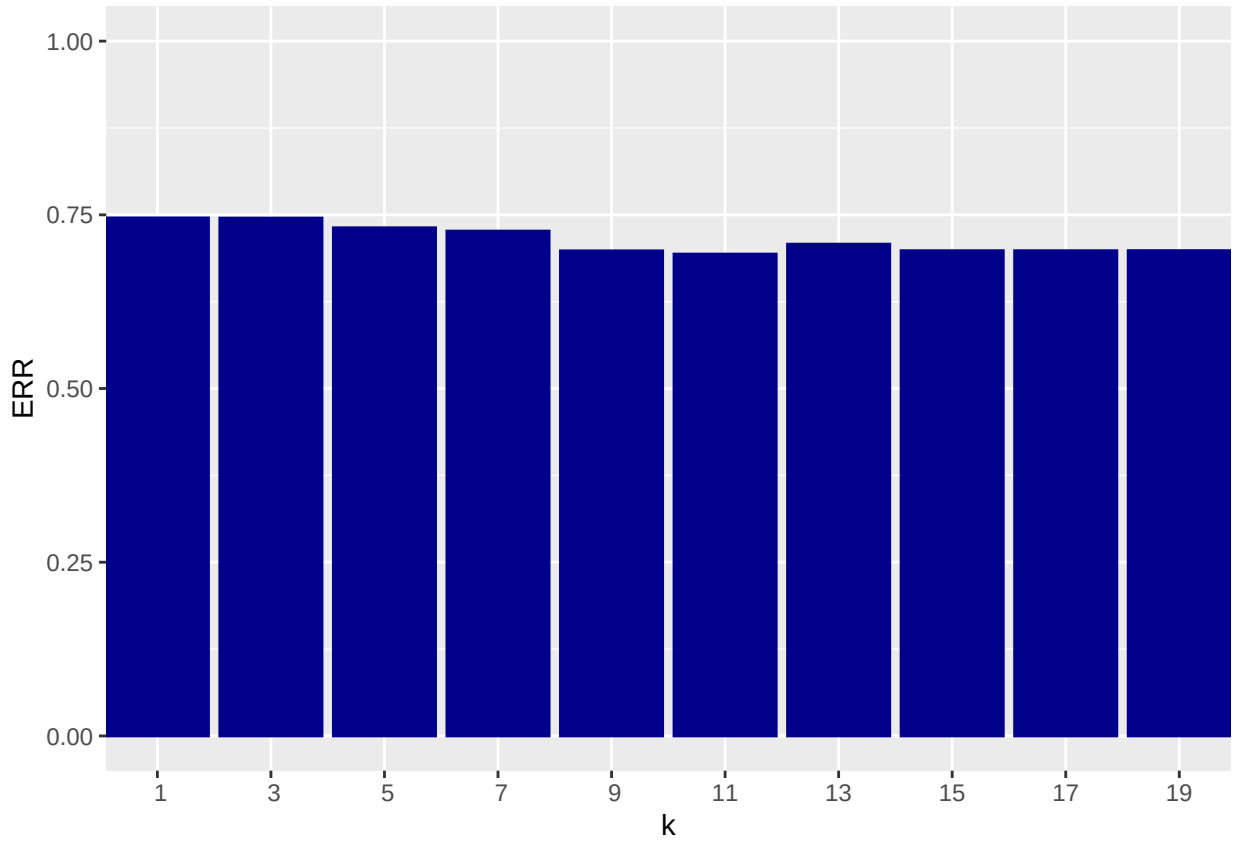
```
tab <- przetworz_test(glass,2, FN = srednia_parzytych)
print(tab)
```

```
##      k       ERR      MAD      MSE
## 1    1 0.7455150 1.401661 3.387486
## 2    3 0.7452935 1.404430 3.610188
## 3    5 0.7313400 1.296899 3.110188
## 4    7 0.7265781 1.380731 3.510078
## 5    9 0.6981174 1.352602 3.604430
## 6   11 0.6934662 1.347951 3.469324
## 7   13 0.7077519 1.352713 3.407863
## 8   15 0.6983389 1.306202 3.174197
## 9   17 0.6983389 1.347951 3.429679
## 10 19 0.6984496 1.281506 3.164341
```

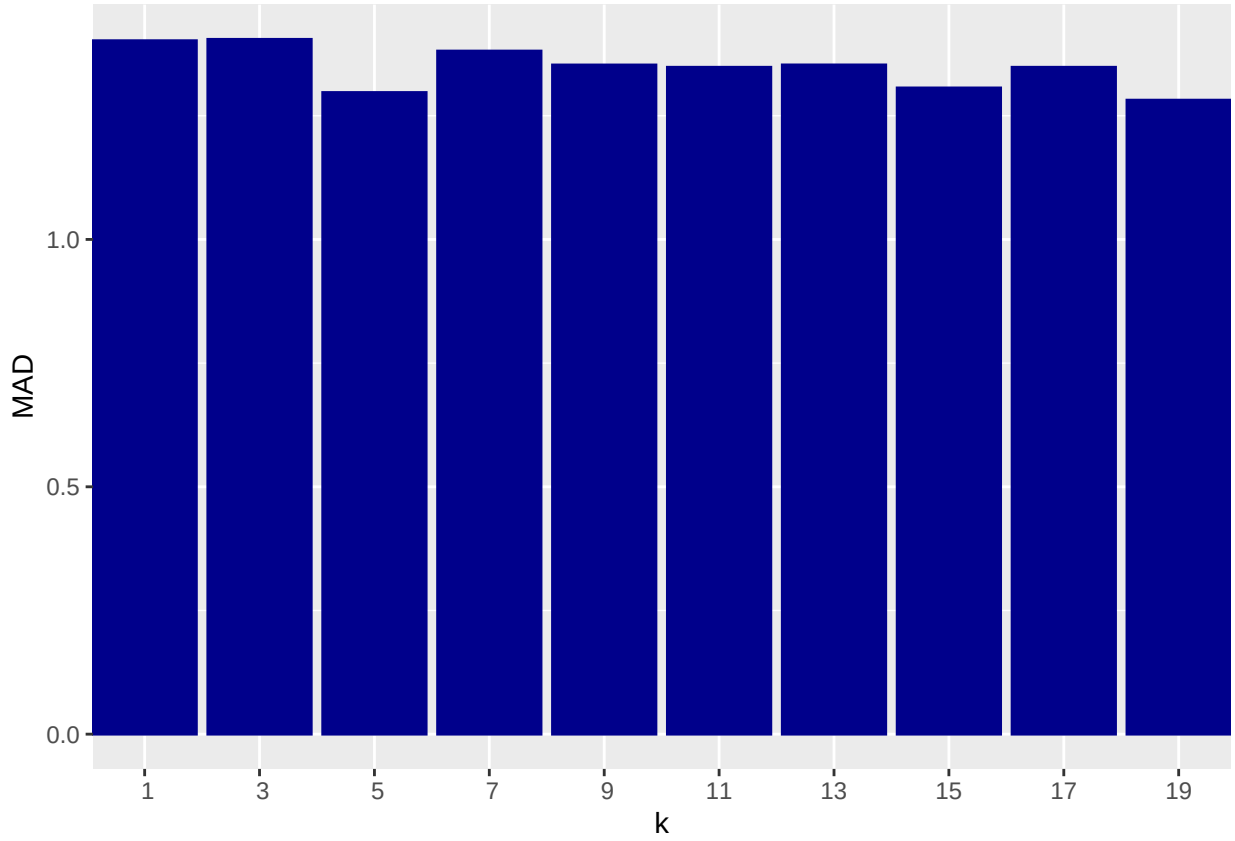
```
przetworz_1nn(glass,2, FN = srednia_parzytych )
```

```
##            ERR       MAD      MSE
## [1,] 0.4460094 0.7746479 2.098592
```

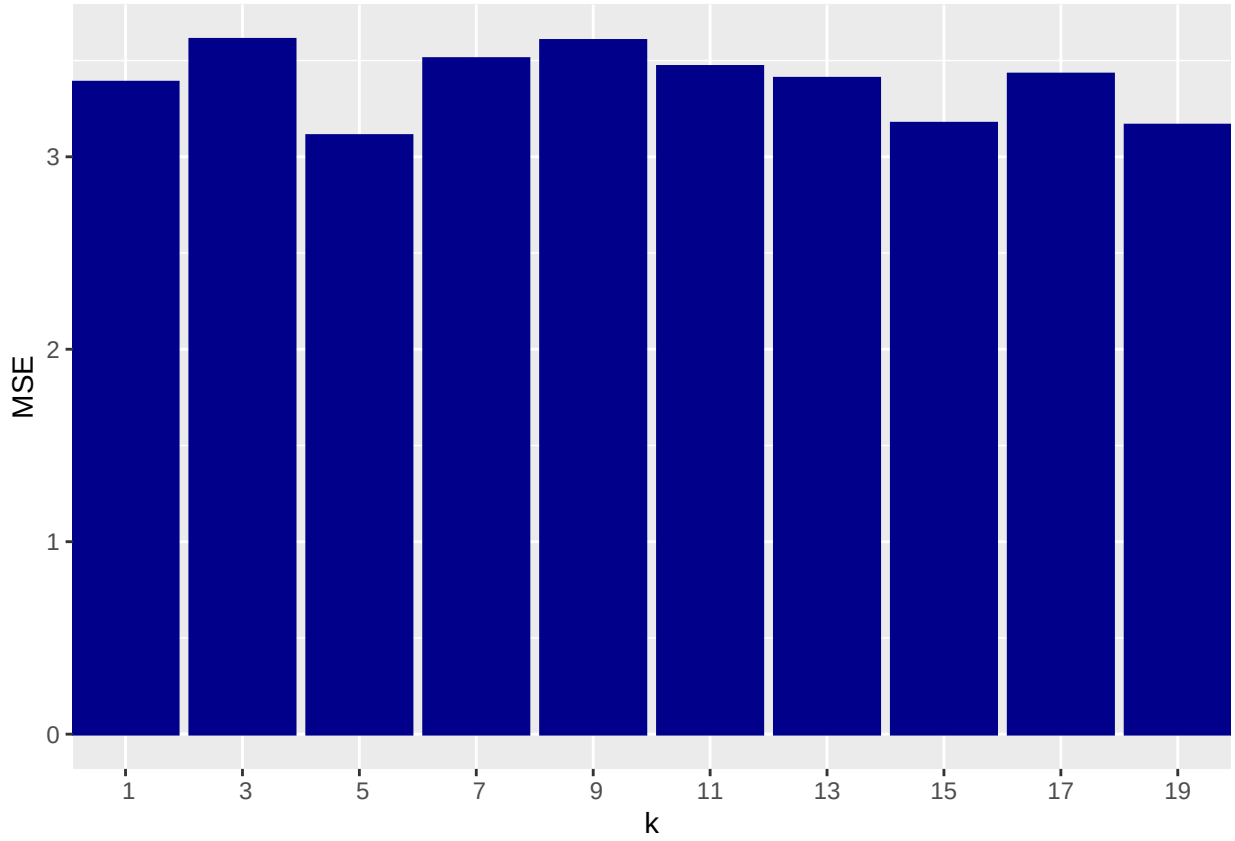
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```




```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



10. Srednia_nieparzytych (L1, glass)

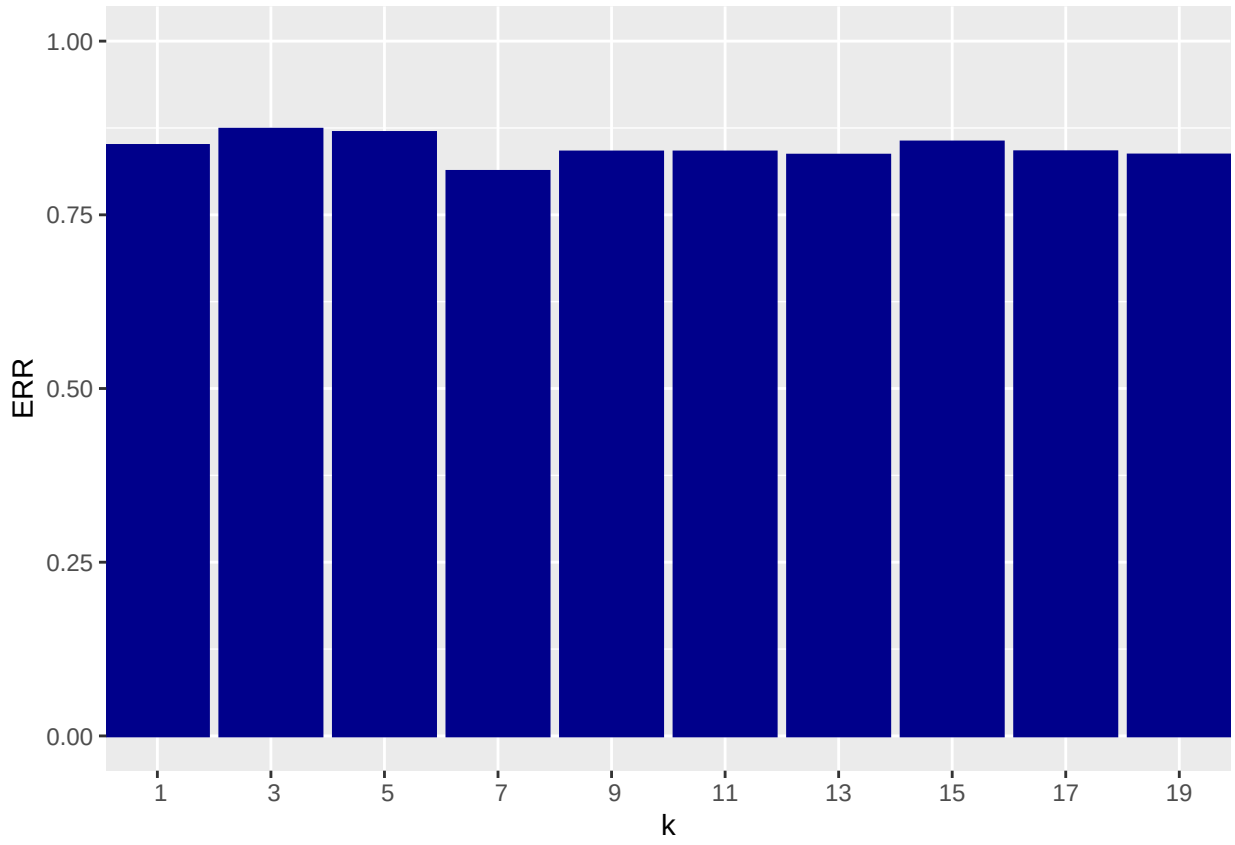
```
tab <- przetworz_test(glass,1, FN = srednia_nieparzytych)
print(tab)
```

```
##     k       ERR      MAD      MSE
## 1   1 0.8498339 2.290476 8.674751
## 2   3 0.8732004 2.149391 7.904430
## 3   5 0.8685493 2.064452 7.304540
## 4   7 0.8125138 1.853599 6.374529
## 5   9 0.8404208 1.857807 6.237874
## 6  11 0.8404208 1.806866 5.805759
## 7  13 0.8358804 1.732115 5.246401
## 8  15 0.8548173 1.718715 5.019491
## 9  17 0.8407530 1.699779 5.000111
## 10 19 0.8361019 1.718937 5.048062
```

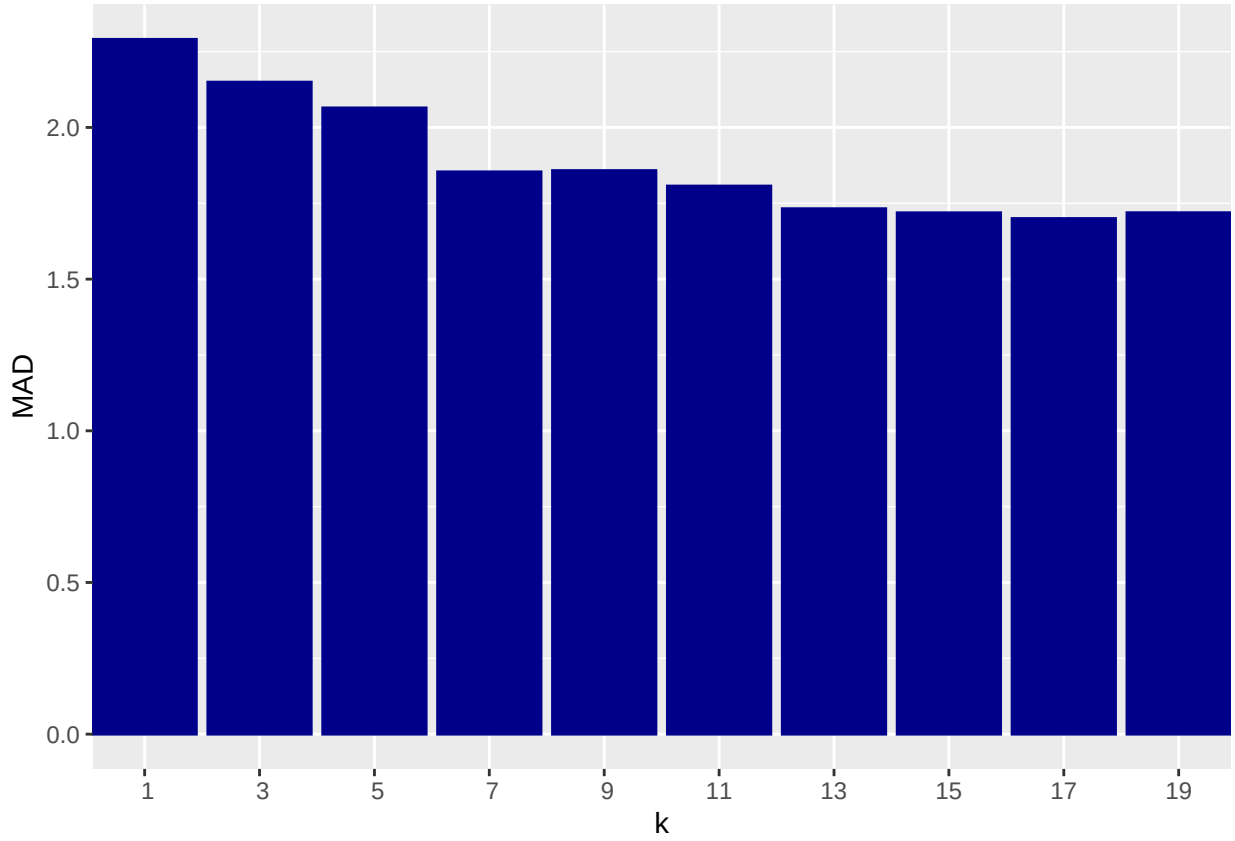
```
przetworz_1nn(glass,1, FN = srednia_nieparzytych )
```

```
##            ERR      MAD      MSE
## [1,] 0.5539906 1.774648 7.305164
```

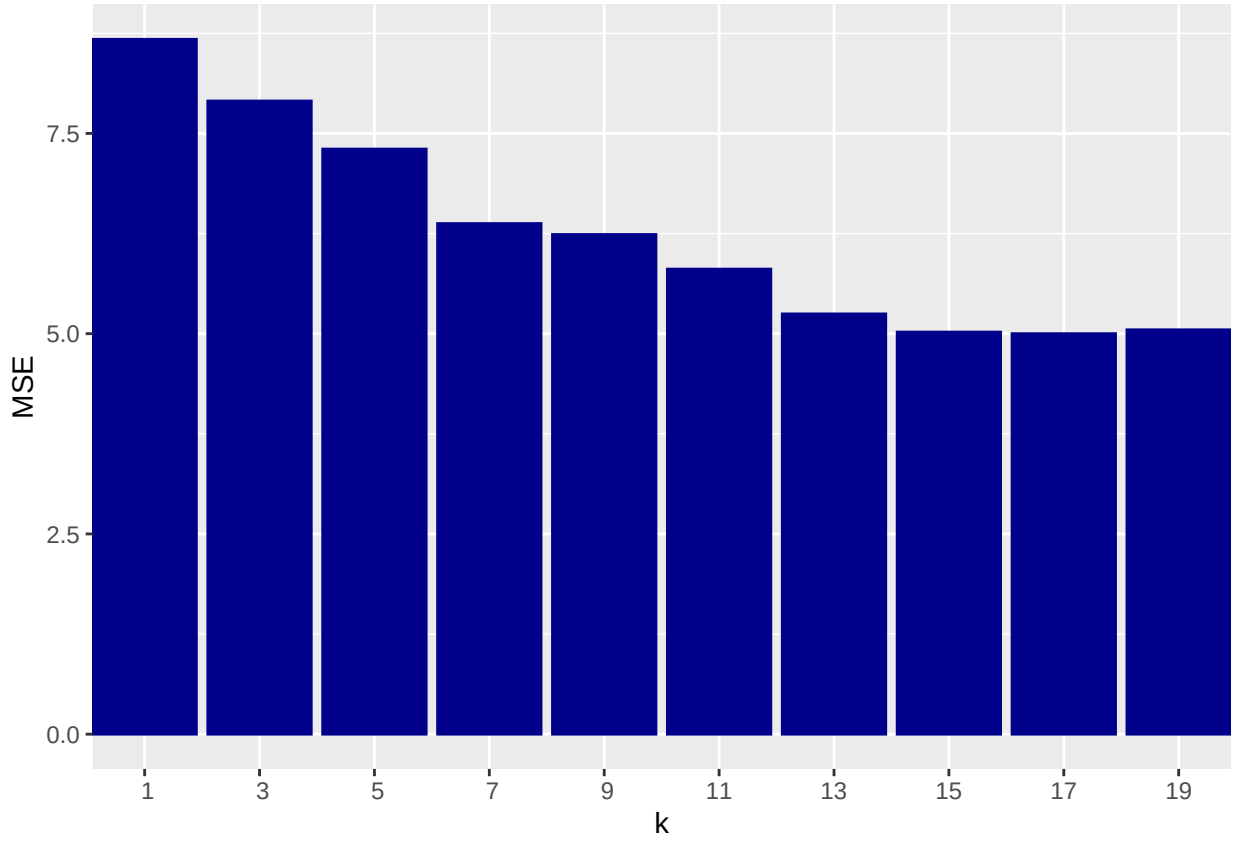
```
ggplot(data = tab, aes(x = k, y = ERR)) + geom_col(color = "darkblue", fill = "darkblue") + ylim(0,1) +
```




```
ggplot(data = tab, aes(x = k, y = MAD)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```




```
ggplot(data = tab, aes(x = k, y = MSE)) + geom_col(color = "darkblue", fill = "darkblue") + scale_x_dis
```



11. Wnioski

W większości przpadków test, gdzie próba testowa i ucząca się pokrywają wychodzi z zerowym błędem. Tak byc powinno zawsze, ale dla pakietu danych affairs zdażają się takie same dane z różnymi etykietami, przez co w pewnym momencie następuje losowanie co nie zawsze daje dobry wynik. Odstepstwo od reguły jest również w testach funkcji srednia_parzystych i srednia_nieparzytych. Jest tak ponieważ poprawne wyniki nie zawsze są odpowiedznio parzyste lub nieparzyste i są pomijane. Dla k = 1 błąd średniokwadratowy i bezwględny są o wiele większe i jest to spodziewany efekt. Tak jak się można spodziewać wyniki typu ERR mieszczą się w granicach 0-100%. Warto zauważyć, że błędy średniokwadradowe i bezwględne są zawyczaj większe od 1. Jest tak dlatego, że każda funkcja agregująca zaokrągla swój wynik do części całkowitej.