# Exercises Class 11-12-19 - Pipeline

Lais Silva Almeida Zanchetta, Daniel Ferreira Zanchetta

18/11/2019

## 1) Let's play a bit with pipes. Using the pipeline operator perform the following operations:

### a) Compute the squared root of the squared of any number.

```
library(magrittr)

var_quad <- function(x) x^2

num_pipe_rsquared <- . %>% var_quad() %>% sqrt()
num_pipe_rsquared(2)
```

```
## [1] 2
```

### b) Sample 1000 individuals from a normal distribution (mean = 5 , sd = 3), standardize the sample (subtract the mean and divide by the standard deviation, i.e., scale) and compute the max value.

```
rnorm(n=1000,mean=5,sd=3) %>% scale(.,center=TRUE, scale=TRUE) %>% max(.)
```

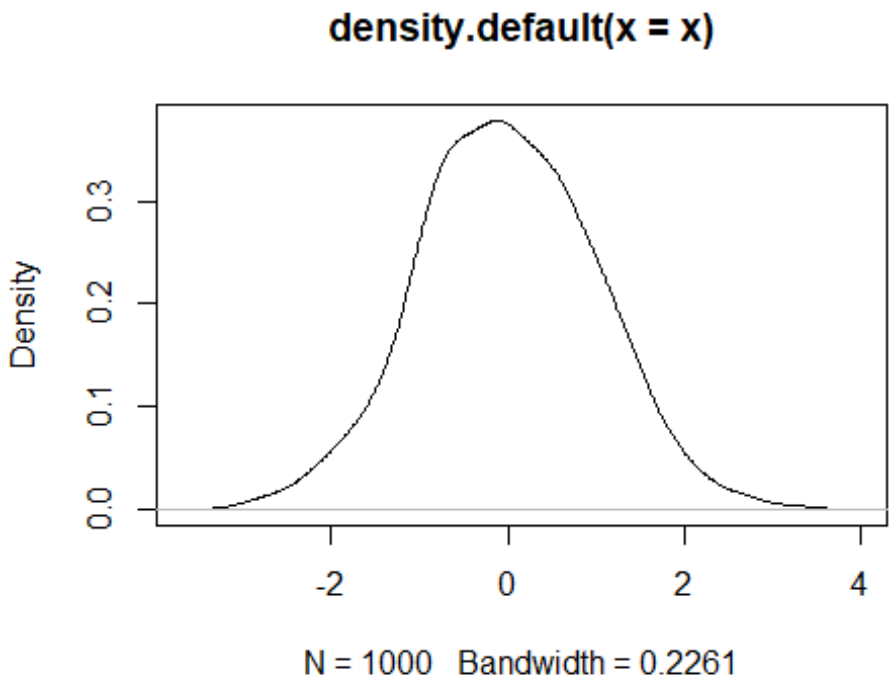
```
## [1] 3.2098
```

### c) Same as b) but plotting the density function before computing the max value.

#P.S.: It is not possible to use the plot density inside the pipeline.

```
plot_dens <- function(x) plot(density(x))

rnorm(n=1000,mean=5,sd=3) %>% scale(.,center=TRUE, scale=TRUE) %T>%
plot_dens(.) %>% max(.)
```




```
## [1] 3.314092
```

## 2. With the pisos dataset and using an only pipeline, compute the following transformations:

### a) Drop the duplicated individuals and compute the mean value of the flats ("Valor") by district ("Dist").

```
bcnpisos <- read.table(file.choose(), header=TRUE) #to read files in mac chosing the file you want from the fold

library(dplyr)

##
## 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

 distinct(bcnpisos) %>% group_by(Dist) %>% summarise(mean_flats =
mean(Valor))
```

```
## # A tibble: 10 x 2
##    Dist         mean_flats
##    <fct>             <dbl>
##  1 Ciutat_Vella  11523408.
##  2 Eixample      21861301.
##  3 Gracia        17401416
##  4 Horta         15769832.
##  5 Les_Corts     28754475.
##  6 Nou_Barris    13352474.
##  7 Sant_Andreu   15365746.
##  8 Sant_Marti    15117212.
##  9 Sants         15598655.
## 10 Sarria        33022471.
```

## b) Drop the duplicated individuals, get the numeric features of the dataset and standardize it.

```
bcn_numeric <- distinct(bcnpisos) %>% .[,unlist(lapply(., is.numeric))]
%>% scale(.,center = TRUE, scale = TRUE)
head(bcn_numeric)
```

```
##           Valor     Superf       Dorm      Banys      Edat     ValSol
## [1,] -1.3060214 -1.6157433 -0.9749402 -0.5875153  1.036136 -0.4258923
## [2,] -1.0983483 -1.2884824 -2.0628190 -0.5875153  2.097013 -1.0909849
## [3,] -0.9677313 -1.2736490 -0.9749402 -0.5875153  1.389761 -0.1672452
## [4,] -0.7676190 -0.7013287  0.1129386 -0.5875153  2.097013 -0.6475898
## [5,]  0.1816739  0.3802701 -0.9749402  1.2517340 -1.439244 -1.1648841
## [6,] -0.8361111 -0.2285155 -0.9749402 -0.5875153  2.556726 -1.5343800
```

## c) Drop the duplicated individuals, add a new factor to the dataset ”Greater than is mean” with values (Y,N) indicating if the Value (”Valor”) of the flat is greater or not than the mean of the flats in the district.

```
dist_mean <- distinct(bcnpisos) %>% group_by(Dist) %>%
summarise(mean_flats = mean(Valor))
dist_greater <- distinct(bcnpisos) %>% left_join(y=dist_mean, by="Dist")
%>% mutate("GREATER_THAN_IS_MEAN"=ifelse(Valor>mean_flats,"Y","N"))
head(dist_greater)
```

```
##      Valor Superf Dorm Banys Edat Estat Planta         Dist    ValSol
## 1  4962780  31.41    2     1   70  1_MM Planta Ciutat_Vella 113322.15
## 2  7001400  42.00    1     1  100   2_M Planta Ciutat_Vella  89407.53
## 3  8283600  42.48    2     1   80   2_M Planta Ciutat_Vella 122622.28
## 4 10248000  61.00    3     1  100  1_MM Planta Ciutat_Vella 105350.61
## 5 19566720  96.00    2     2    0  5_MB Planta Ciutat_Vella  86750.35
## 6  9575648  76.30    2     1  113  1_MM   Atic Ciutat_Vella  73464.45
##   Tipus Ascens ExtInt   Reforma mean_flats GREATER_THAN_IS_MEAN
## 1  MANZ     NO    EXT  REF15-20   11523408                    N
## 2  MANZ     NO    EXT    REF1A5   11523408                    N
## 3  MANZ     NO    EXT RECIENREF   11523408                    N
## 4  MANZ     SI    EXT RECIENREF   11523408                    N
```

```
## 5  MANZ     SI     EXT OBRANUEVA   11523408                    Y
## 6  MANZ     NO     EXT  REF10-15   11523408                    N
```

## 3) Finally, you are asked to do a complete transformation of the pisos dataset. We want to analyse and visualize some general features of the districts of the city, characterizing a sample of flats.

### a) Propose R code for the transformation of this dataset. You are free to use any technique explained during the course (and others) but the use of some pipes will be valued positively (7 points).

```
library(dplyr)
new_bcnpisos <- bcnpisos %>% distinct %>% group_by(Dist) %>%
rename(DistrictName = Dist) %>% summarise('1Dorm' = sum(Dorm==1),'2Dorm'
= sum(Dorm==2),'3Dorm' = sum(Dorm == 3), '4Dorm' = sum(Dorm == 4),
'5Dorm' = sum(Dorm == 5), 'Valor' = mean(Valor, na.rm = TRUE), 'AscS' =
sum(Ascens == 'SI'), 'AscN' = sum(Ascens == 'NO'), 'Atic' = sum(Planta ==
'Atic'), 'Bajos' = sum(Planta == 'Bajos'), 'Planta' = sum(Planta ==
'Planta'), 'Nous' = sum(Edat <= 10), 'SemiNous' = sum(Edat >=11 && Edat
<=20), 'Vells' = sum(Edat >=21 && Edat<=50), 'MoltVells' = sum(Edat >=
51), 'Superf' = mean(Superf, na.rm = TRUE)
)

arrange(new_bcnpisos, DistrictName)
```

```
## # A tibble: 10 x 17
##    DistrictName `1Dorm` `2Dorm` `3Dorm` `4Dorm` `5Dorm`  Valor  AscS
AscN
##    <fct>          <int>   <int>   <int>   <int>   <int>  <dbl> <int>
<int>
##  1 Ciutat_Vella      51      68      53      13       3 1.15e7    37
151
##  2 Eixample          24      63     126     125      29 2.19e7   283
84
##  3 Gracia            13      41      68      34       6 1.74e7    81
81
##  4 Horta             12      52     139      45       1 1.58e7   111
138
##  5 Les_Corts          1      13      33      31       5 2.88e7    68
15
##  6 Nou_Barris         6      65     127      25       0 1.34e7   106
117
##  7 Sant_Andreu       11      32     106      35       1 1.54e7   121
64
##  8 Sant_Marti        16      70     187      51       1 1.51e7   209
116
##  9 Sants             23      71     165      59       1 1.56e7   223
96
```

```
## 10 Sarria            15      21      45      50      22 3.30e7   131
22
## # ... with 8 more variables: Atic <int>, Bajos <int>, Planta <int>,
## #   Nous <int>, SemiNous <int>, Vells <int>, MoltVells <int>, Superf
<dbl>
```

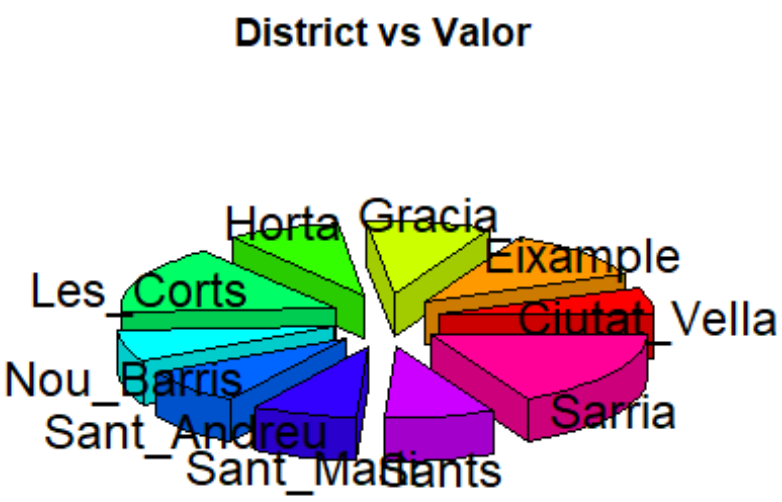
## b) Propose nice visualizations of this new dataset (3 points).

```
library(ggplot2)
library(plotrix)
library(dplyr)
library(tidyr)

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract

slices <- new_bcnpisos$Valor
labels <- new_bcnpisos$DistrictName
pie3D(slices,labels = labels, explode=0.25, main="District vs Valor")
```




```
new_bcnpisos %>% arrange(desc(DistrictName)) %>%
ggplot(aes(x=Superf,y=DistrictName, size=Valor))+
  geom_point(alpha=0.5) +
  scale_size(range = c(.1, 24), name="Valor (M)") +
```

```
  theme(legend.position="bottom") +
  ylab("Nombre Districto") +
  xlab("Superficie") +
  theme(axis.title.y = element_text(angle = 1))
```