Porto Business School
/ University of Porto

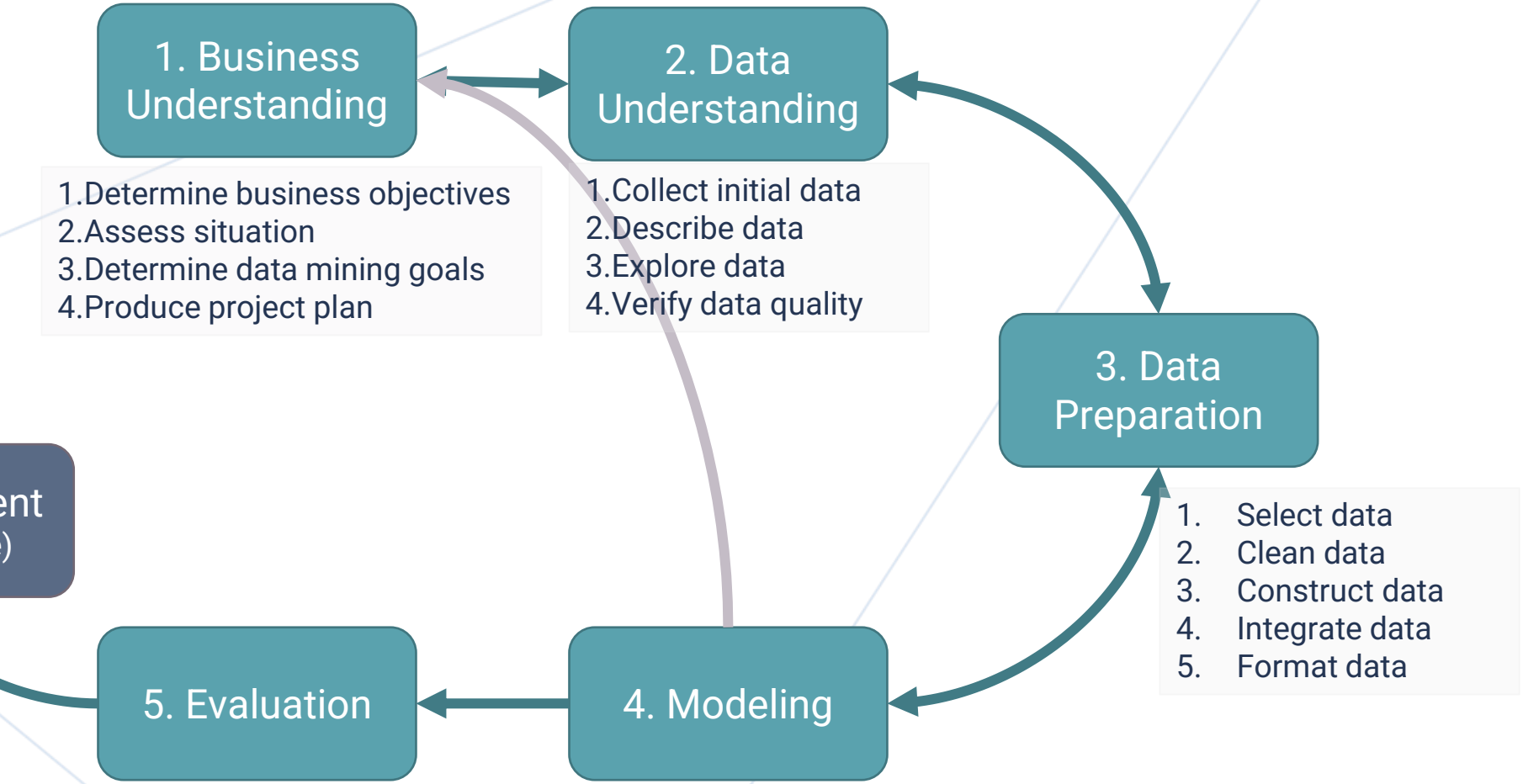
Landing.Jobs

# Projeto 1

## PGBIA 13 – Grupo 4


Barbara Correia
Diogo Miranda
Joana Silva
José Alexandre

# CRISP-DM: Etapas e tarefas



# 1. Business Understanding

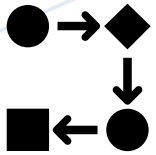
# Business Understanding

A sua proposta de valor passa por ligar os **utilizadores** e as **empresas** para o preenchimento de **ofertas de trabalho** na **área da tecnológica**


A Landing Jobs pretende **otimizar** o **match** entre **utilizadores** e **empresas** usando automatização, previsão e classificação

Porto Business School
/ University of Porto

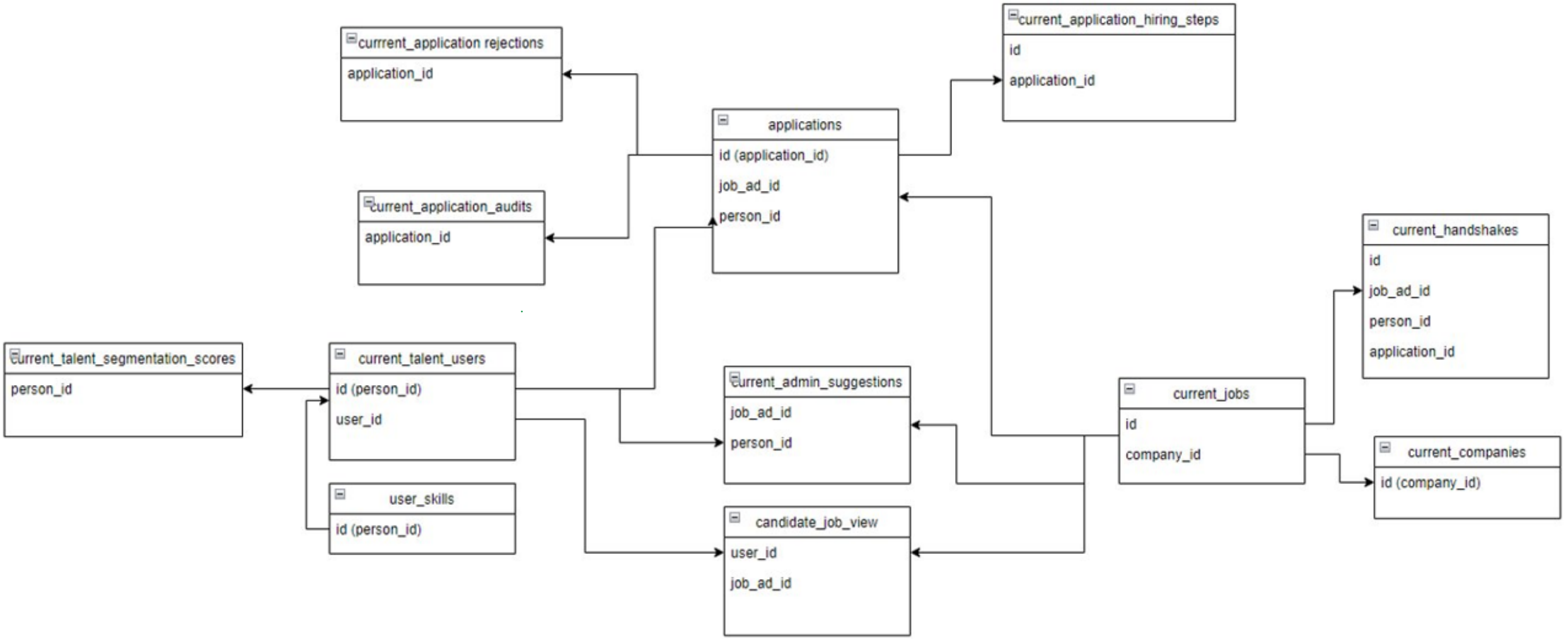
# Objetivos do Data Mining:



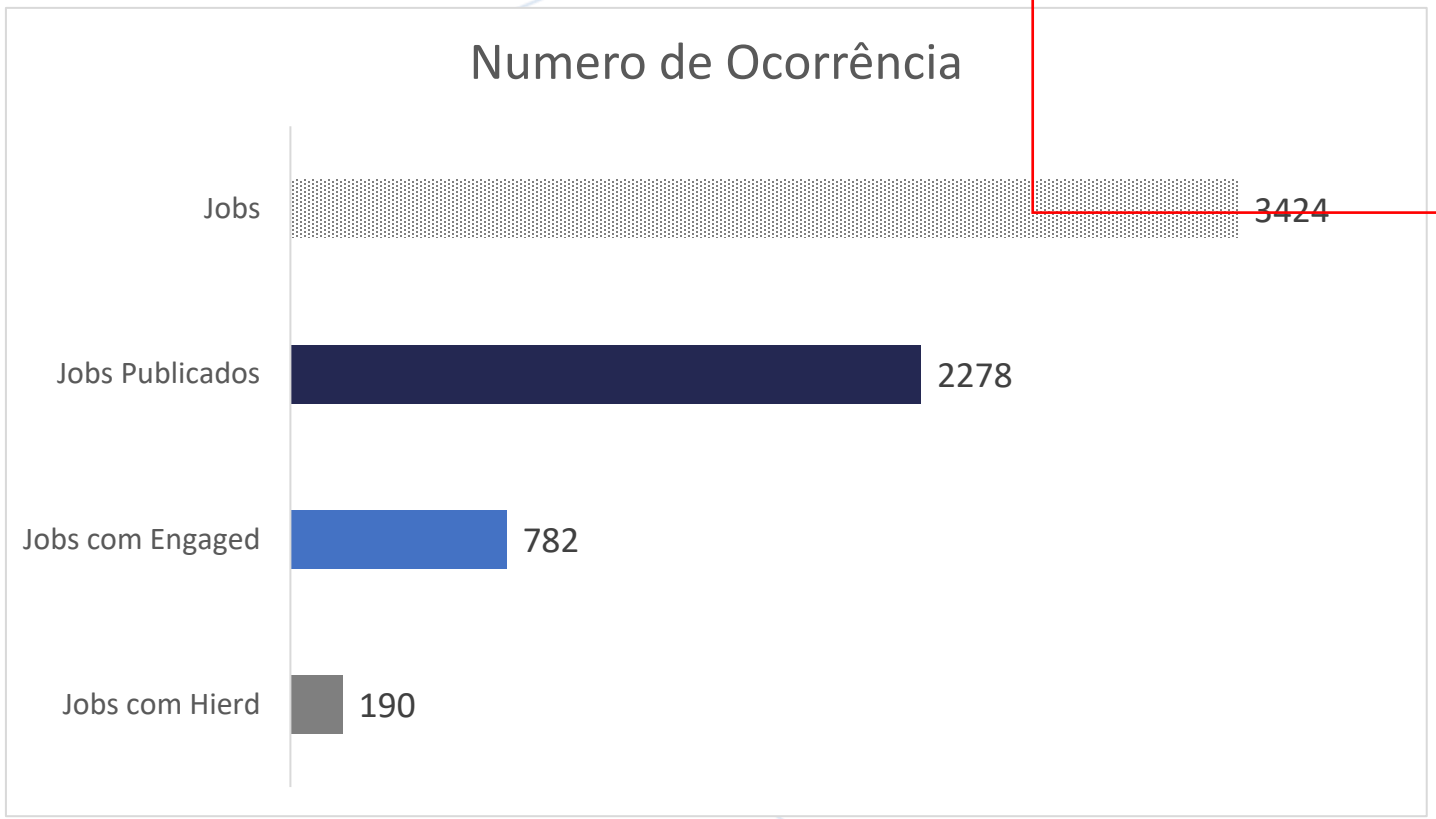
- Perceber que tipos de jobs existem na Plataforma (Clustering)
- Prevêr quanto tempo um job vai demorar a ter um engaged (Regressão)
- Prevêr se um job vai ter um Engaged (Classificação)

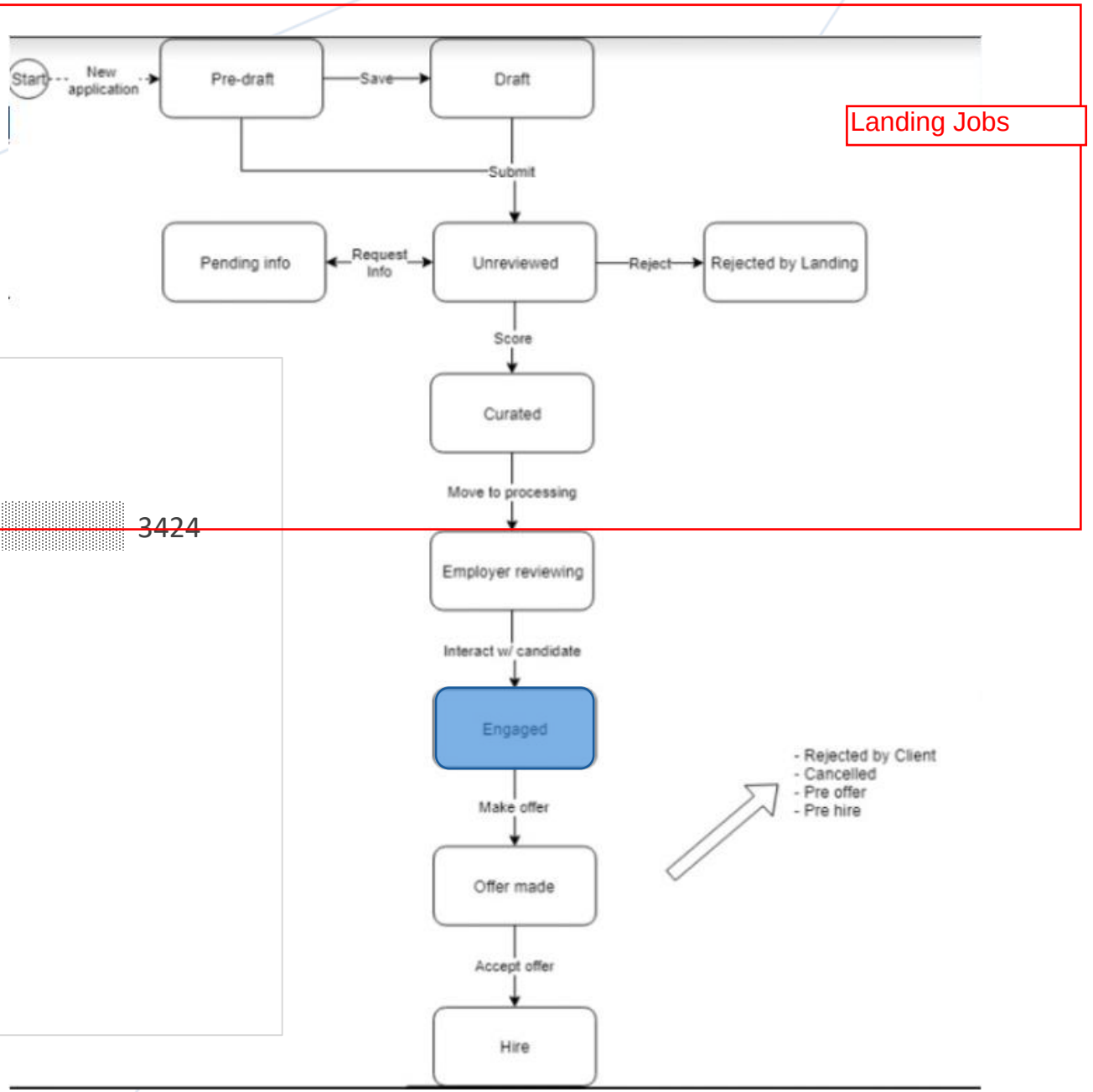
# 2. Data Understanding

# Data Understanding: Modelo Relacional



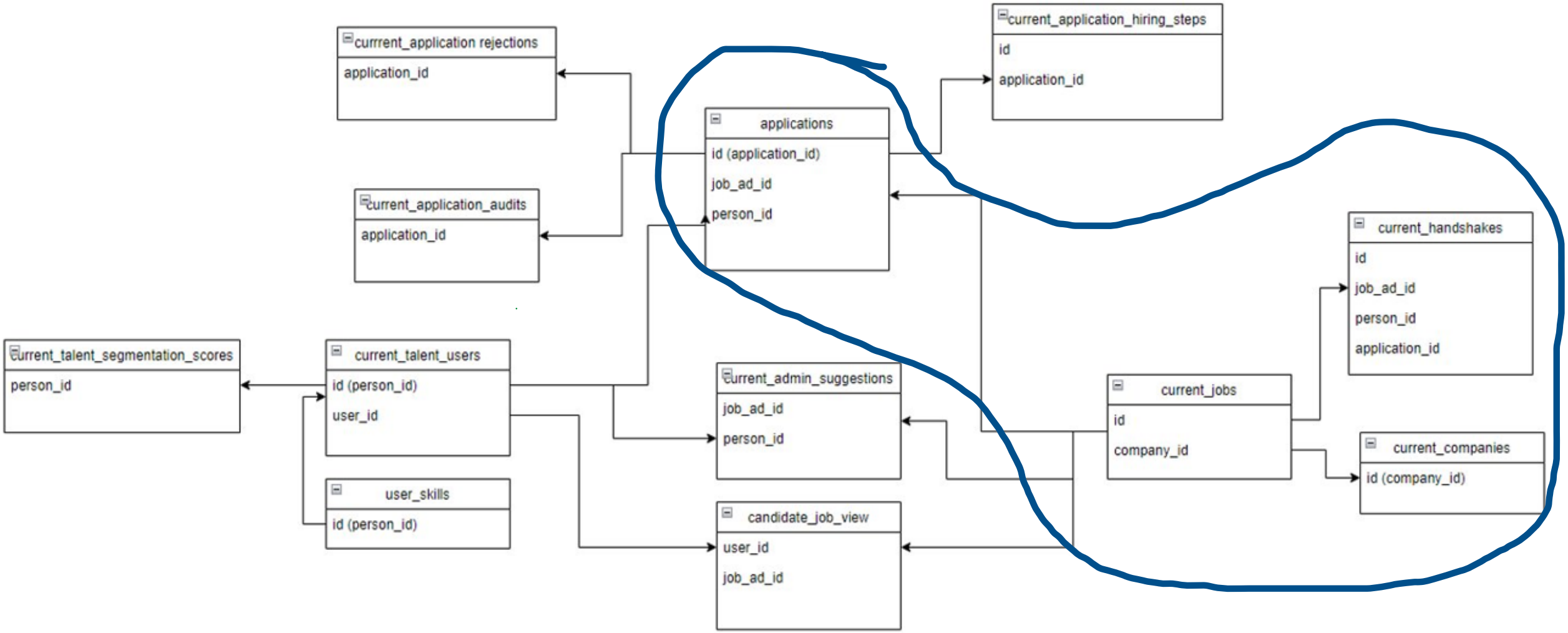
# Data Understanding




Dados de 01/05/2021 a 15/03/2022 (update_at)

# Data Preparation: Modelo Relacional




Dados Usados

# Summary Application

Porto Business School
/ University of Porto

```
summary(applications)
```

```
       id            job_ad_id       person_id       created_at         updated_at         starts_on         has_work_permit  seen_by_employer_at current_hiring_step_id tracking_codes     match_score
 Min.   :293878   Min.   : 2586   Min.   :    11   Length:151167      Length:151167      Length:151167      Min.   : NA      Length:151167       Min.   :24729          Min.   : NA      Min.   :0.00
 1st Qu.:308979   1st Qu.:12035   1st Qu.:112418   Class :character   Class :character   Class :character   1st Qu.: NA      Class :character    1st Qu.:26706          1st Qu.: NA      1st Qu.:0.26
 Median :324630   Median :12820   Median :165532   Mode  :character   Mode  :character   Mode  :character   Median : NA      Mode  :character    Median :28690          Median : NA      Median :0.55
 Mean   :324624   Mean   :12660   Mean   :141667                                                         Mean   :NaN                          Mean   :28712          Mean   :NaN      Mean   :0.60
 3rd Qu.:339730   3rd Qu.:13620   3rd Qu.:177672                                                         3rd Qu.: NA                          3rd Qu.:30666          3rd Qu.: NA      3rd Qu.:0.80
 Max.   :357966   Max.   :15072   Max.   :196096                                                         Max.   : NA                          Max.   :32790          Max.   : NA      Max.   :2.00
                                                                                                         NA's   :151167                       NA's   :142613         NA's   :151167   NA's   :48245

 submitted_at       availability    availability_detail last_state_change_at gross_annual_salary currency_code       deleted_at       handpicked_requested_at handpicked_state     state
 Length:151167      Min.   :0.00    Length:151167       Length:151167        Min.   : 1200       Length:151167      Length:151167      Length:151167           Min.   :0.00     Length:151167
 Class :character   1st Qu.:0.00    Class :character    Class :character     1st Qu.:28842       Class :character   Class :character   Class :character        1st Qu.:1.00     Class :character
 Mode  :character   Median :0.00    Mode  :character    Mode  :character     Median :33570       Mode  :character   Mode  :character   Mode  :character        Median :1.00     Mode  :character
                    Mean   :0.55                                            Mean   :34409                                                                  Mean   :0.82
                    3rd Qu.:1.00                                            3rd Qu.:40450                                                                  3rd Qu.:1.00
                    Max.   :2.00                                            Max.   :70500                                                                  Max.   :2.00
                    NA's   :150838                                          NA's   :151133                                                                 NA's   :149421
   initiator      initiator_label
 Min.   :0.0000   Length:151167
 1st Qu.:0.0000   Class :character
 Median :0.0000   Mode  :character
 Mean   :0.2106
 3rd Qu.:0.0000
 Max.   :3.0000
```

# Summary Current_jobs

Porto Business School
/ University of Porto

```
summary(current_jobs)
```

```
      id        job_country_code    job_region          company_id     job_title         company_persona    company_segment      job_city         full_remote        partial_remote     job_category      
 Min.   :11470   Length:3424        Length:3424        Min.   :   8   Length:3424        Length:3424        Length:3424        Length:3424        Length:3424        Length:3424        Length:3424       
 1st Qu.:12358   Class :character   Class :character   1st Qu.:2170   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Median :13276   Mode  :character   Mode  :character   Median :4399   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :13271                                         Mean   :3736                                                                                                                                  
 3rd Qu.:14164                                         3rd Qu.:5247                                                                                                                                  
 Max.   :15084                                         Max.   :6070                                                                                                                                  

  job_skills            team          active_state       contractor         consultancy        created_at         updated_at         first_published_at last_published_at   education        
 Length:3424        Min.   : NA     Length:3424        Length:3424        Length:3424        Length:3424        Length:3424        Length:3424        Length:3424        Length:3424       
 Class :character   1st Qu.: NA     Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Median : NA     Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                    Mean   :NaN                                                                                                                                                                
                    3rd Qu.: NA                                                                                                                                                                
                    Max.   : NA                                                                                                                                                                
                    NA's   :3424                                                                                                                                                               

  permanent          preferred_language  show_salary         show_rate          relocation_paid    total_visits       visa_support       work_from_home     experience_level_label   job_type        
 Length:3424        Length:3424        Length:3424        Length:3424        Length:3424        Min.   :    0.0   Length:3424        Length:3424        Length:3424        Length:3424       
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:    0.0   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :  843.5   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                             Mean   : 1629.7                                                                                                
                                                                             3rd Qu.: 2030.5                                                                                                
                                                                             Max.   :30539.0                                                                                                

 timezone_utc_max  timezone_utc_min     state          
 Min.   :-12.000   Min.   :-12.000   Length:3424       
 1st Qu.:  1.000   1st Qu.: -2.000   Class :character  
 Median :  1.000   Median :  0.000   Mode  :character  
 Mean   :  2.216   Mean   : -1.324                     
 3rd Qu.:  3.000   3rd Qu.:  0.000                     
 Max.   : 14.000   Max.   : 10.000                     
 NA's   :2079      NA's   :2078                        
```

# Summary Current_handshakes

Porto Business School
/ University of Porto

```
summary(current_handshakes)
```

```
       id          created_at         accepted_at        rejected_at          company_id    company_country_code candidate_country_code user_active_state    job_ad_id       person_id
 Min.   : 53708   Length:83297       Length:83297       Length:83297       Min.   :   8   Length:83297         Length:83297          Length:83297       Min.   : 2586   Min.   :     1
 1st Qu.: 74564   Class :character   Class :character   Class :character   1st Qu.:2092   Class :character     Class :character      Class :character   1st Qu.:12322   1st Qu.: 41399
 Median : 95410   Mode  :character   Mode  :character   Mode  :character   Median :3853   Mode  :character     Mode  :character      Mode  :character   Median :13012   Median : 95781
 Mean   : 95395                                                           Mean   :3462                                                              Mean   :12841   Mean   : 92813
 3rd Qu.:116222                                                           3rd Qu.:5080                                                              3rd Qu.:13513   3rd Qu.:144117
 Max.   :137094                                                           Max.   :6014                                                              Max.   :15011   Max.   :195857
                                                                                                                                                    NA's   :4763
 state_labels       suggestion_score application_id     expired          company_persona    candidate_persona  candidate_newness_state created_by_id
 Length:83297       Min.   :2.39     Min.   :285758   Length:83297       Length:83297       Length:83297       Length:83297            Min.   :   194
 Class :character   1st Qu.:2.39     1st Qu.:312875   Class :character   Class :character   Class :character   Class :character        1st Qu.:   439
 Mode  :character   Median :2.39     Median :329803   Mode  :character   Mode  :character   Mode  :character   Mode  :character        Median :   542
                    Mean   :2.39     Mean   :326873                                                                                  Mean   : 74664
                    3rd Qu.:2.39     3rd Qu.:338970                                                                                  3rd Qu.:173300
                    Max.   :2.39     Max.   :357947                                                                                  Max.   :202230
                    NA's   :83296    NA's   :74889
```

# Summary Current_companies

Porto Business School
/ University of Porto

```
summary(current_companies)
```

```
       id          perks              VALUES          time_to_review   time_to_engage   time_to_reject      actions            score        engagement_level   seg_engagement  seg_strategic_fit
 Min.   :   1   Length:4917        Length:4917        Min.   : 0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :-4.473   Length:4917        Min.   :0.000   Min.   :0.000
 1st Qu.:1584   Class :character   Class :character   1st Qu.: 0.000   1st Qu.:  1.75   1st Qu.:  5.00   1st Qu.:  0.000   1st Qu.:-0.412   Class :character   1st Qu.:1.000   1st Qu.:1.000
 Median :3129   Mode  :character   Mode  :character   Median : 2.000   Median :  5.00   Median : 10.00   Median :  0.000   Median :-0.350   Mode  :character   Median :2.000   Median :2.000
 Mean   :3095                                         Mean   : 4.988   Mean   : 10.68   Mean   : 14.79   Mean   :  8.463   Mean   :-0.252                      Mean   :1.923   Mean   :1.785
 3rd Qu.:4588                                         3rd Qu.: 5.000   3rd Qu.: 13.00   3rd Qu.: 18.25   3rd Qu.:  0.000   3rd Qu.:-0.261                      3rd Qu.:3.000   3rd Qu.:2.000
 Max.   :6081                                         Max.   :68.000   Max.   :106.00   Max.   :110.00   Max.   :452.000   Max.   : 1.428                      Max.   :3.000   Max.   :3.000
                                                      NA's   :4832     NA's   :4861     NA's   :4829     NA's   :4191      NA's   :4433                        NA's   :3725    NA's   :3726
 seg_brand_awareness   seg_score     has_traas         has_account_manager seg_business_potential     city           country_code       job_types          full_remote     persona
 Min.   :0.000       Min.   :0.00   Length:4917        Length:4917         Min.   :0.000           Length:4917        Length:4917        Length:4917        Min.   :0      Length:4917
 1st Qu.:1.000       1st Qu.:1.25   Class :character   Class :character    1st Qu.:1.000           Class :character   Class :character   Class :character   1st Qu.:0      Class :character
 Median :1.000       Median :1.50   Mode  :character   Mode  :character    Median :1.000           Mode  :character   Mode  :character   Mode  :character   Median :0      Mode  :character
 Mean   :1.377       Mean   :1.59                                          Mean   :1.359                                                                    Mean   :0
 3rd Qu.:2.000       3rd Qu.:1.95                                          3rd Qu.:2.000                                                                    3rd Qu.:0
 Max.   :3.000       Max.   :3.00                                          Max.   :3.000                                                                    Max.   :0
 NA's   :3725        NA's   :3725                                          NA's   :3725                                                                     NA's   :3970
 active_state       newness_state        category          industry         company_size        numbers            segment           updated_at        short_pitch_replaced
 Length:4917        Length:4917        Length:4917        Length:4917        Length:4917        Length:4917        Length:4917        Length:4917        Length:4917
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

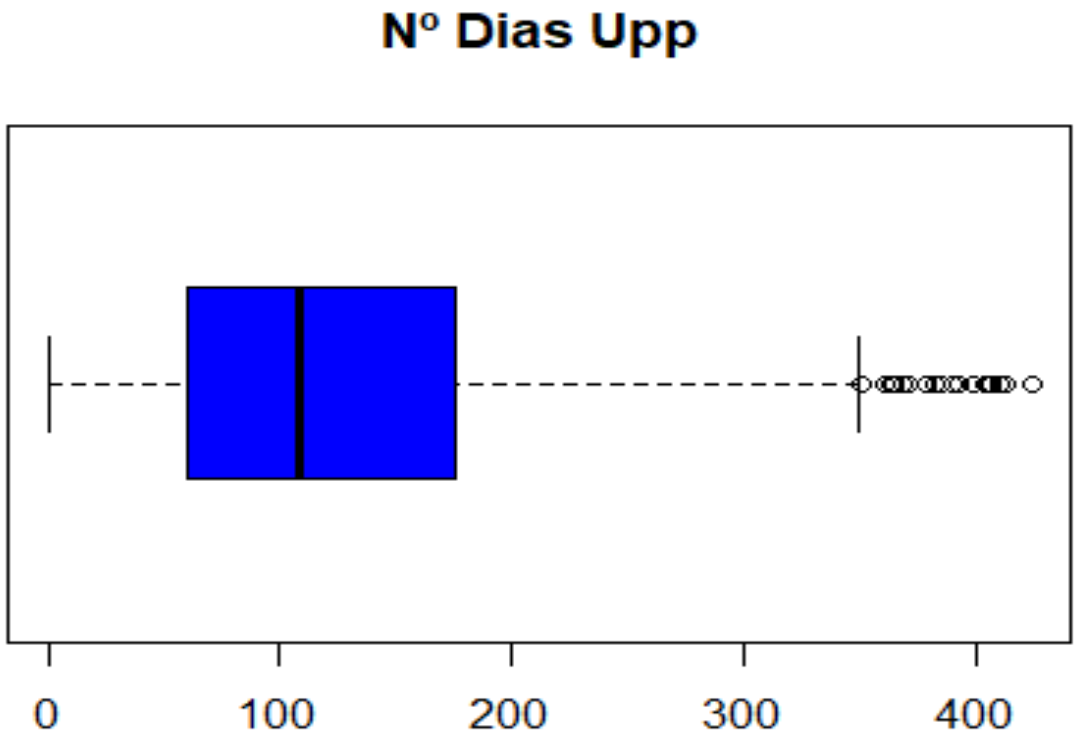
## #Retrive Dataset de MySQL

```
library(RMySQL)

drv <- RMySQL::MySQL()

lan <- dbConnect(drv,
                 host = "169.51.29.197",
                 user ="g4u1",
                 password = "5In4Pf",
                 dbname = "landing")
#criar ligação para a base de dados onde que
remos ler está com o user e pass da Bárbara


Jobs_apps <- dbGetQuery(lan,"
select jobs.id,
jobs.job_region,
jobs.job_category,
jobs.contractor,
jobs.consultancy,
jobs.education,
jobs.permanent,
jobs.preferred_language,
jobs.full_remote,
jobs.partial_remote,
jobs.show_salary,
jobs.show_rate,
jobs.relocation_paid,
jobs.visa_support,
jobs.experience_level_label,
jobs.job_type,
jobs.total_visits,
comps.score comp_score,
comps.persona comp_type,
case
 when (jobs.state = 'Closed' or jobs.state =
'Unpublished')
 then datediff(jobs.updated_at,jobs.first_pu
blished_at)
 else datediff('2022-03-15',jobs.first_publi
shed_at)
 end  dias_up,
coalesce(jobs_hands.n_hands,0) n_hands,
coalesce(count_apps.n_apps,0) n_apps,
coalesce(count_eng.n_eng,0) n_eng,
coalesce(count_hir.n_hired,0) n_hired,
count_apps.min_app,
date_eng.min_engaged
from landing.current_jobs jobs
inner join landing.current_companies comps o
n comps.id=jobs.company_id
 left join
(select jobs.id, count(distinct( apps.id)) n
_eng
from landing.current_jobs jobs
join landing.applications apps on jobs.id=ap
ps.job_ad_id
where jobs.first_published_at is not null
and apps.state in ('Engaged')
group by jobs.id) count_eng on count_eng.id=
jobs.id
left join
(select jobs.id, count(distinct( apps.id)) n
_hired
from landing.current_jobs jobs
join landing.applications apps on jobs.id=ap
ps.job_ad_id
where jobs.first_published_at is not null
and apps.state in ('Hired')
group by jobs.id) count_hir on count_hir.id=
jobs.id
left join
(select jobs.id, count(distinct( apps.id)) n
_apps, min(datediff(apps.last_state_change_a
t,jobs.first_published_at)) min_app
from landing.current_jobs jobs
join landing.applications apps on jobs.id=ap
ps.job_ad_id
where jobs.first_published_at is not null
and apps.state not in ('Pre-draft','Draft')
group by jobs.id) count_apps on count_apps.i
d=jobs.id
left join
 (Select jobs.id,
 min(datediff(apps.last_state_change_at,jobs
.first_published_at)) min_engaged
From landing.current_jobs as jobs
 join landing.applications as apps on jobs.i
d=apps.job_ad_id
where jobs.first_published_at<>''
and apps.state='Engaged'
group by jobs.id ) date_eng on date_eng.id=j
obs.id
left join
(select jobs.id id, count(distinct hands.id)
n_hands
 from landing.current_jobs jobs
 left join landing.current_handshakes hands
on hands.job_ad_id=jobs.id
 where jobs.first_published_at <>''
 and hands.state_labels not in ('Requested',
'Rejected')
 group by jobs.id )jobs_hands on jobs_hands.
id=jobs.id
 where jobs.first_published_at <>''

 ")
dbDisconnect(lan)


#write.csv2(Jobs_apps,paste0("C:/Users/barba
/Associação Porto Business School/PGBIA13P1G
04 - General/03_Data Preparation/",format(Sy
s.time(), "%Y%m%d-%H%M"),"jobsApps.csv"))
```

# Summary DataSet

```
      id          job_region        job_category        contractor        consultancy         education          permanent         preferred_language full_remote        partial_remote
 Min.   :11474   Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278
 1st Qu.:12270   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :13127   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :13181
 3rd Qu.:14072
 Max.   :15072

 show_salary         show_rate         relocation_paid    visa_support       experience_level_label   job_type          total_visits       comp_score          dias_up          n_hands
 Length:2278        Length:2278        Length:2278        Length:2278        Length:2278              Length:2278        Min.   :    1.0   Min.   :-4.47263   Min.   :  0.0   Min.   :   0.00
 Class :character   Class :character   Class :character   Class :character   Class :character         Class :character   1st Qu.:  847.8   1st Qu.:-0.39795   1st Qu.: 60.0   1st Qu.:   0.00
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character         Mode  :character   Median : 1519.0   Median :-0.13100   Median :108.0   Median :   0.00
                                                                                                                       Mean   : 2449.6   Mean   :-0.03382   Mean   :124.7   Mean   :  27.36
                                                                                                                       3rd Qu.: 2859.5   3rd Qu.: 0.37811   3rd Qu.:176.0   3rd Qu.:  26.00
                                                                                                                       Max.   :30539.0   Max.   : 1.42838   Max.   :424.0   Max.   :2439.00
                                                                                                                                         NA's   :260

    n_apps            n_eng            n_hired          min_engaged
 Min.   :  0.00   Min.   :  0.000   Min.   :0.00000   Min.   :  0.00
 1st Qu.:  2.00   1st Qu.:  0.000   1st Qu.:0.00000   1st Qu.:  8.00
 Median :  7.00   Median :  0.000   Median :0.00000   Median : 21.00
 Mean   : 15.13   Mean   :  1.944   Mean   :0.09702   Mean   : 37.15
 3rd Qu.: 19.00   3rd Qu.:  1.000   3rd Qu.:0.00000   3rd Qu.: 46.75
 Max.   :430.00   Max.   :134.000   Max.   :4.00000   Max.   :341.00
                                                      NA's   :1496
```

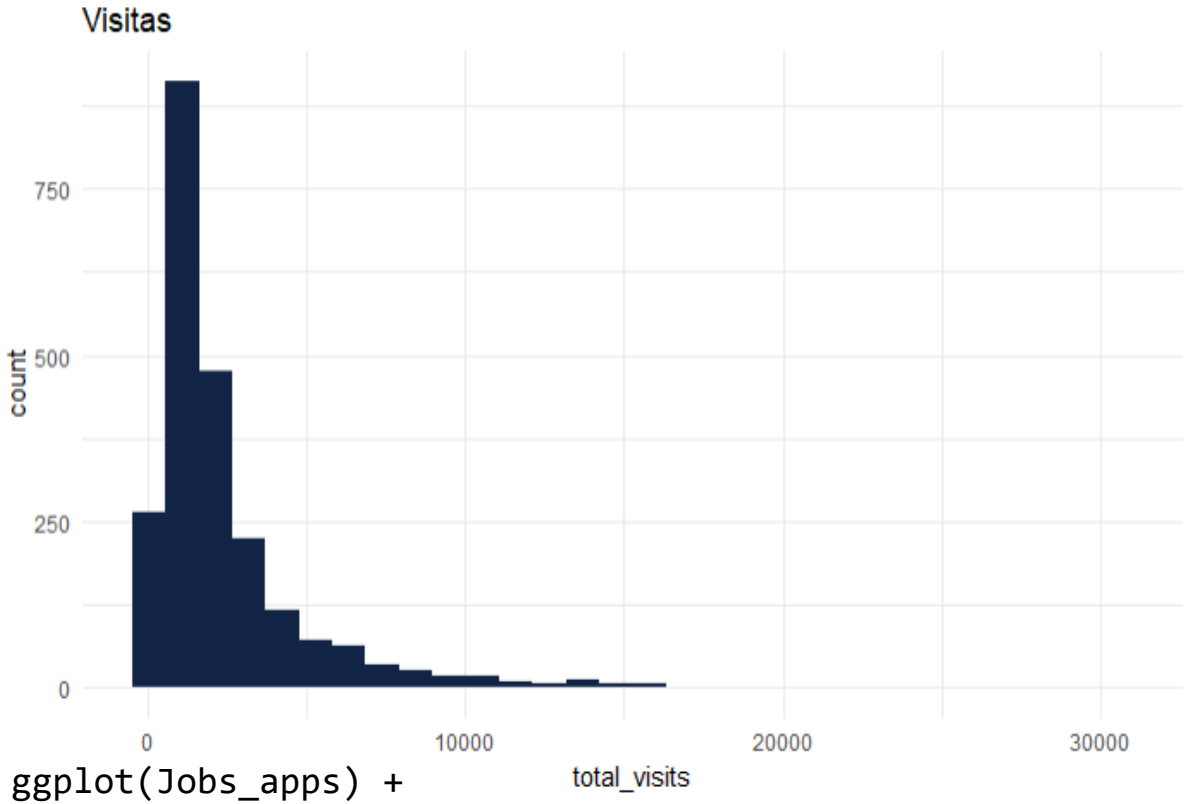
# Data Understanding



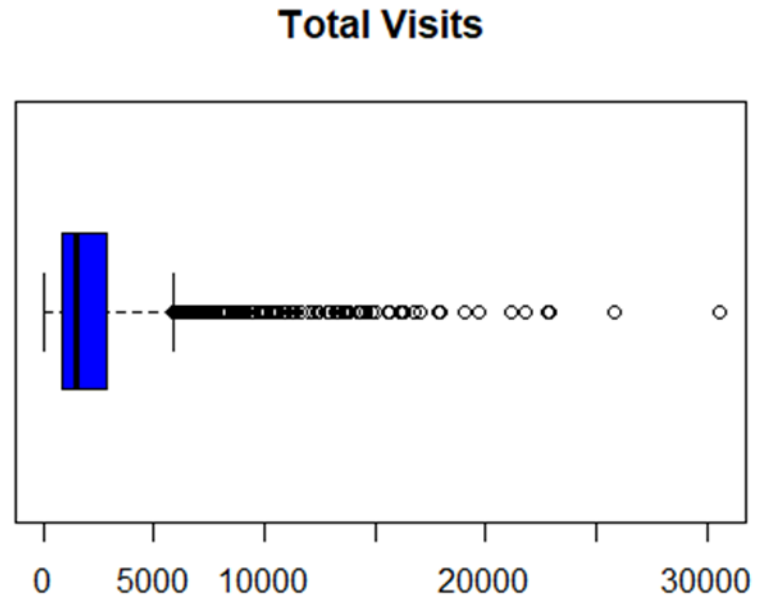

Boxplot do nº de dias que cada job esteve online

# Data Understanding

## Análise univariada – Número de Visitas por Job



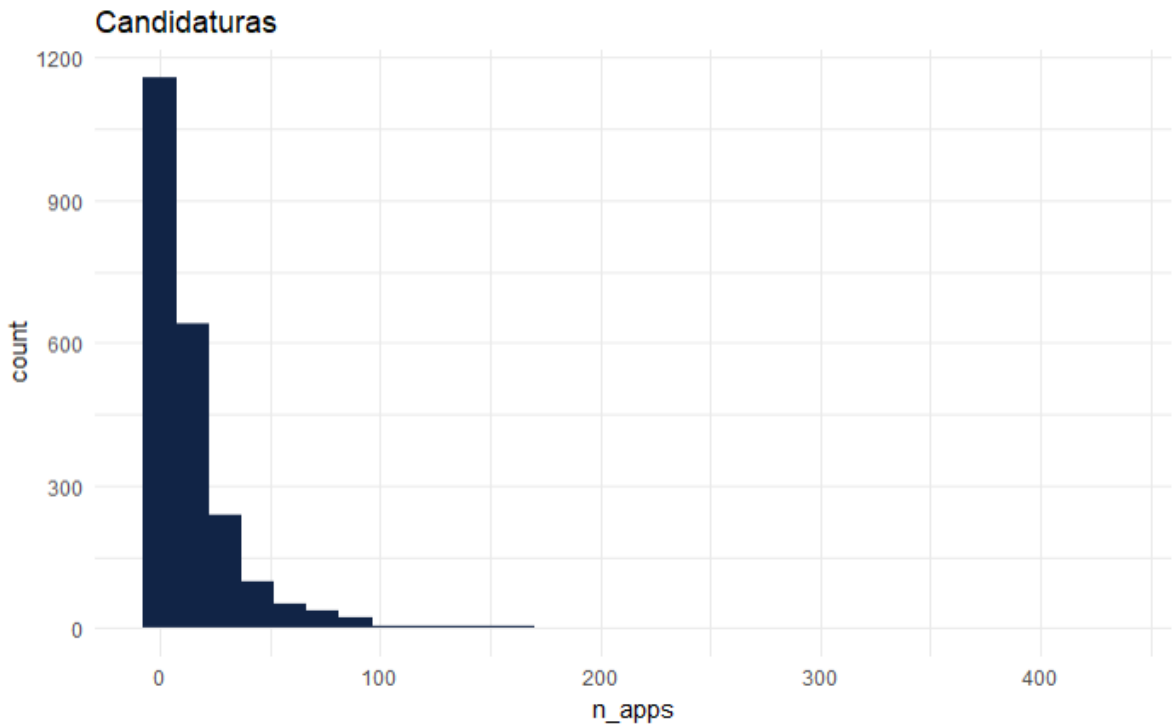

```
ggplot(Jobs_apps) +
  aes(x = n_apps) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(title = "Candidaturas") +
  theme_minimal()
```



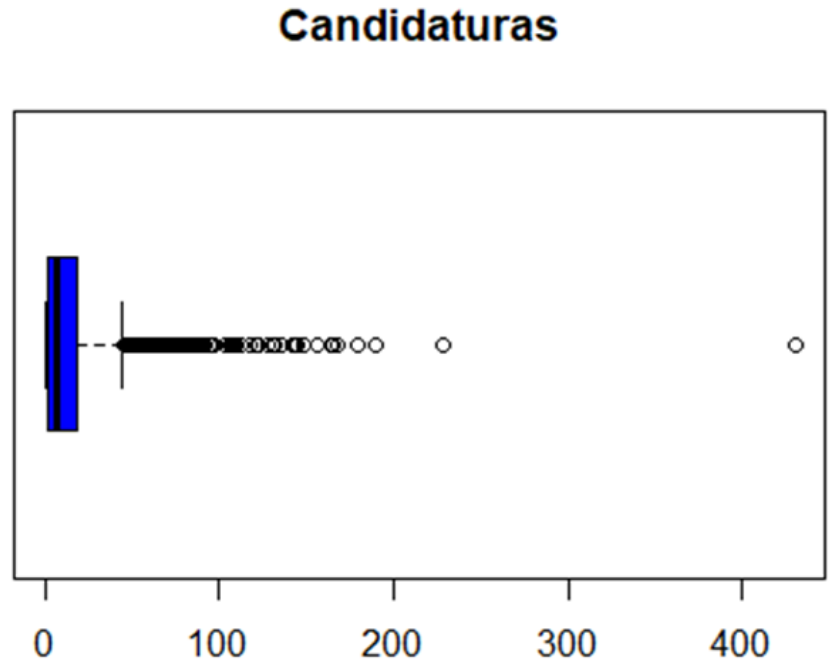

```
boxplot(Jobs_apps$total_visits,
main = "Total Visits",
col = "blue",
border = "black",
horizontal = TRUE
#notch = TRUE
)
```

# Data Understanding

## Análise univariada – Número de Candidaturas por Job



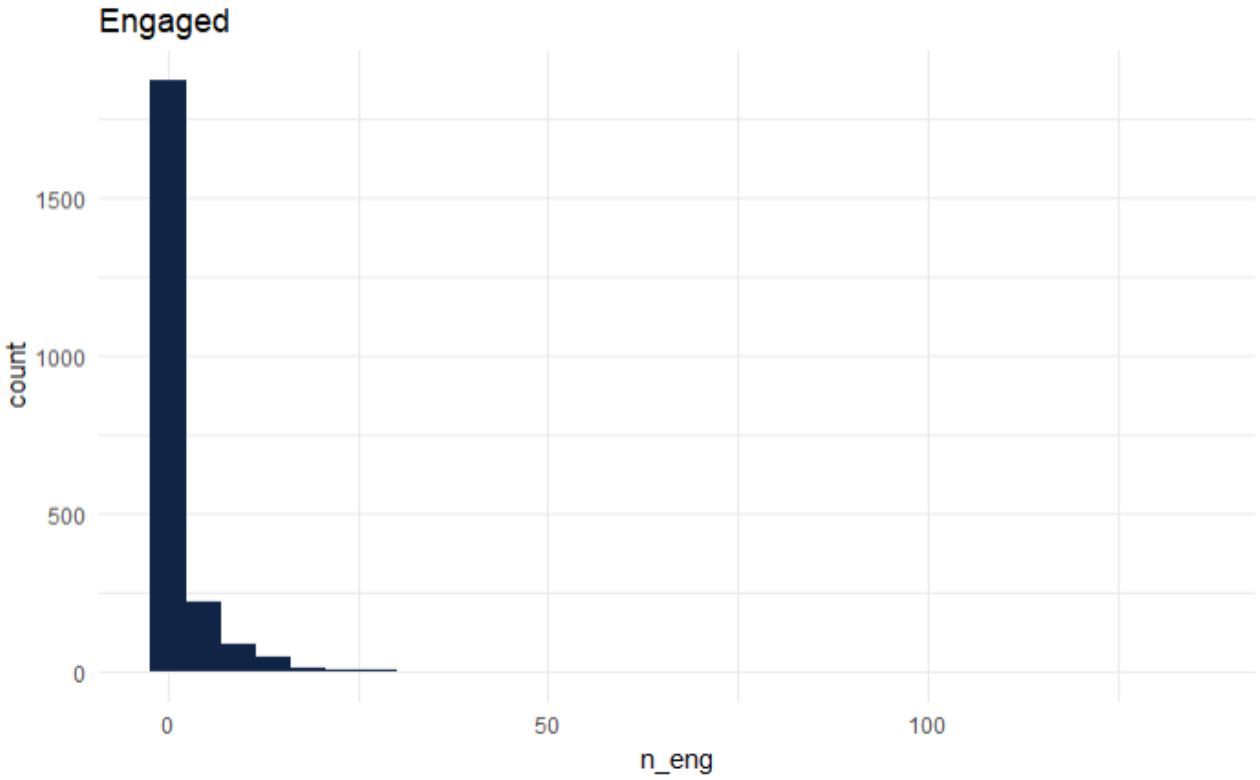

```
ggplot(Jobs_apps) +
  aes(x = n_eng) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(title = "Engaged") +
  theme_minimal()
```



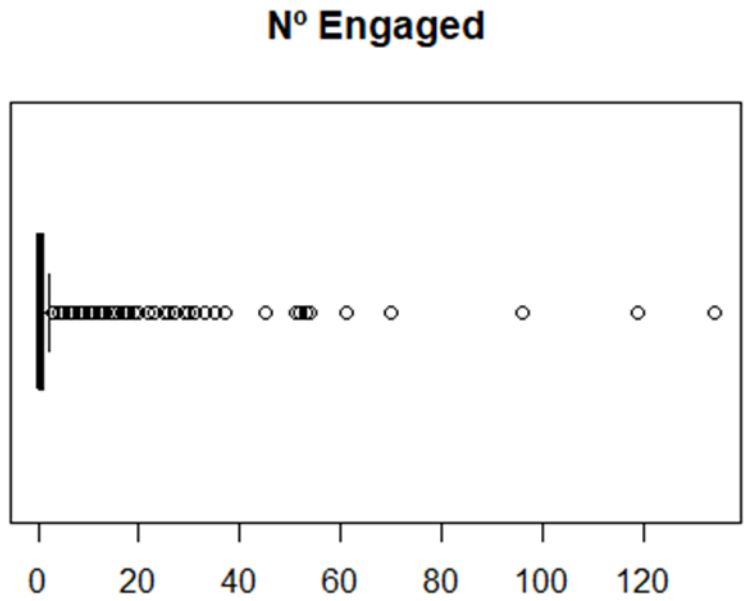

```
boxplot(Jobs_apps$n_eng,
main = "Nº Engaged",

col = "blue",
border = "black",
horizontal = TRUE
#notch = TRUE
)
```

# Data Understanding

## Análise univariada – Número de Candidaturas que chegaram a Engaged por Job



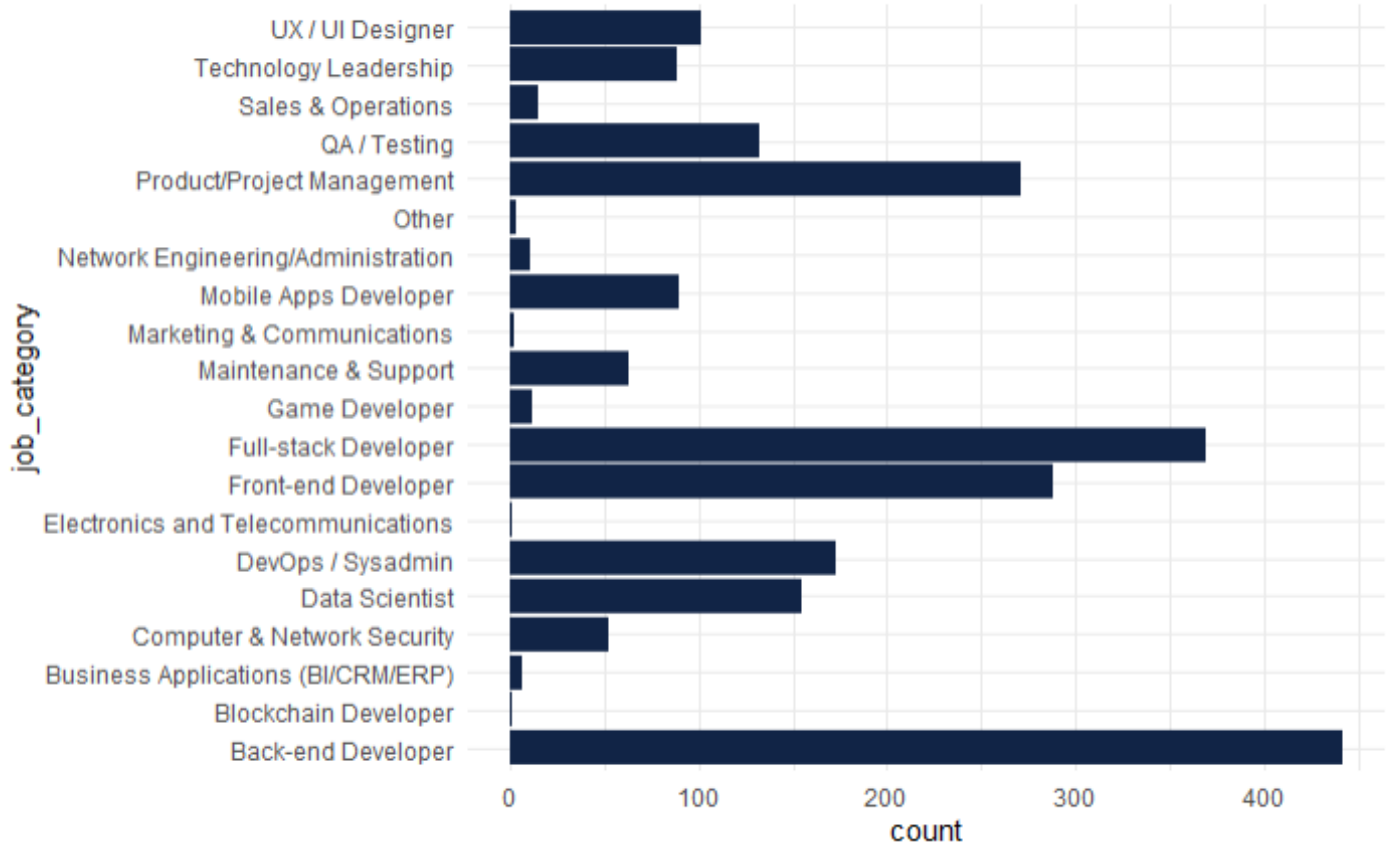

```
ggplot(Jobs_apps) +
  aes(x = n_apps) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(title = "Candidaturas") +
  theme_minimal()
```




```
boxplot(Jobs_apps$n_eng,
main = "Nº Engaged",

col = "blue",
border = "black",
horizontal = TRUE
#notch = TRUE
)
```

# Data Understanding

## Gráfico de Barras Ofertas por Categoria de trabalho

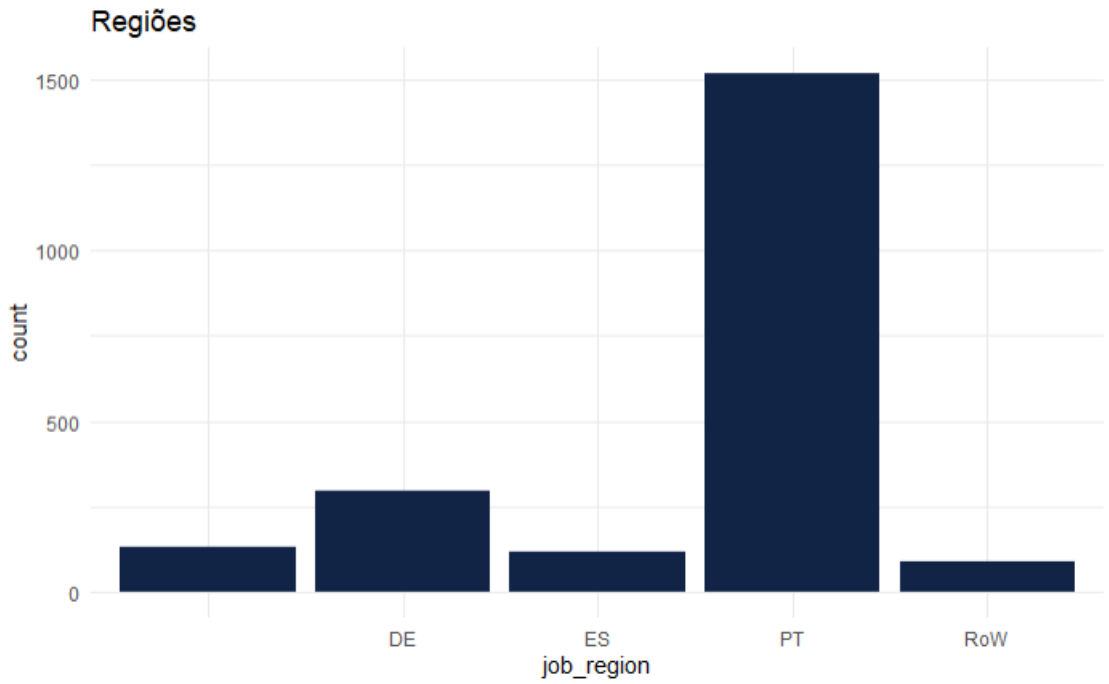


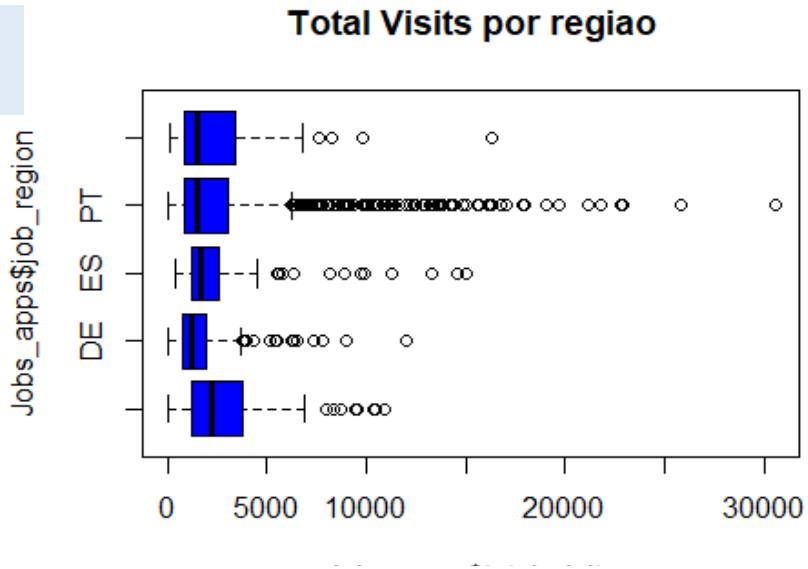

```
library(ggplot2)

ggplot(Jobs_apps) +
  aes( y= job_category) +
  geom_bar(fill =
"#112446") +
  theme_minimal()
```

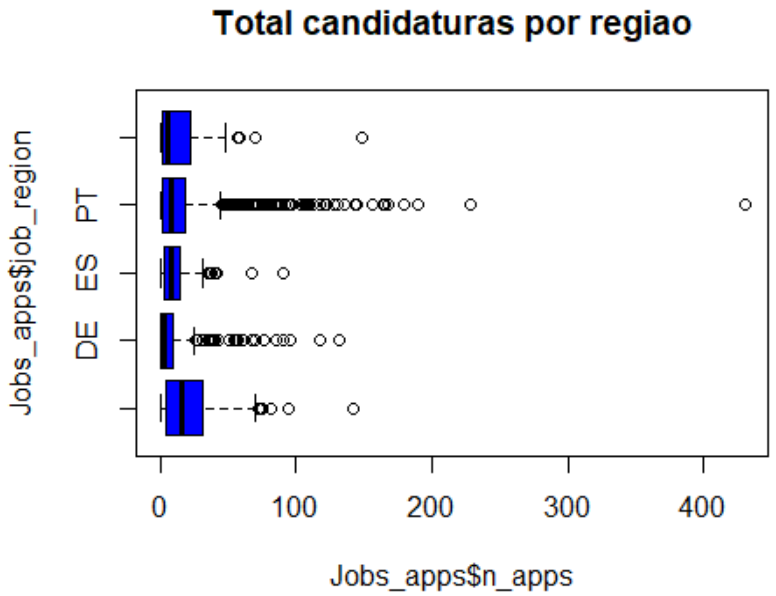
# Data Understanding

## Regiões

**Comentário:**
A maior parte dos Jobs publicados é em Portugal



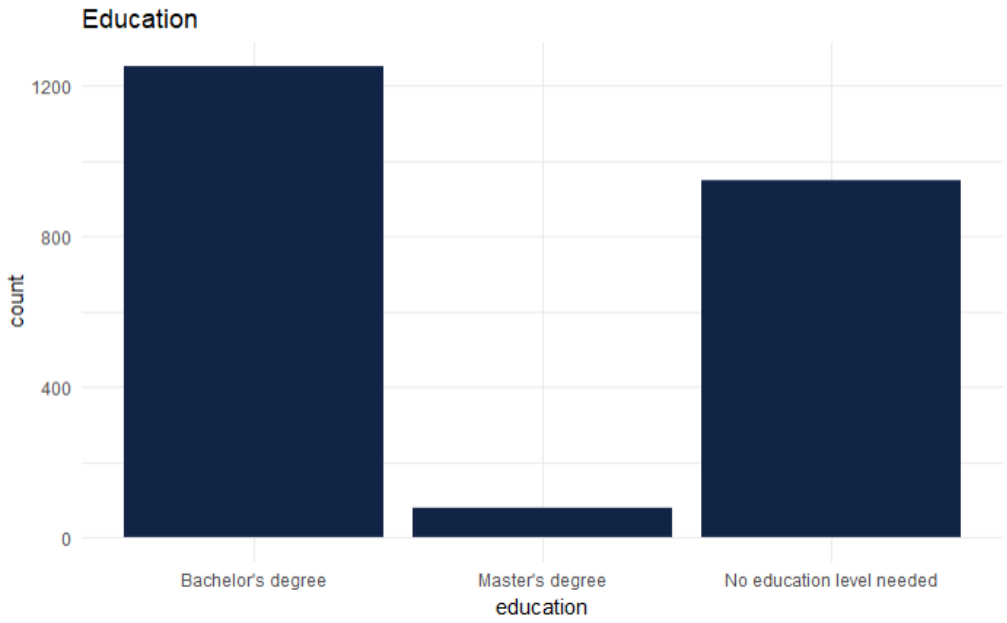



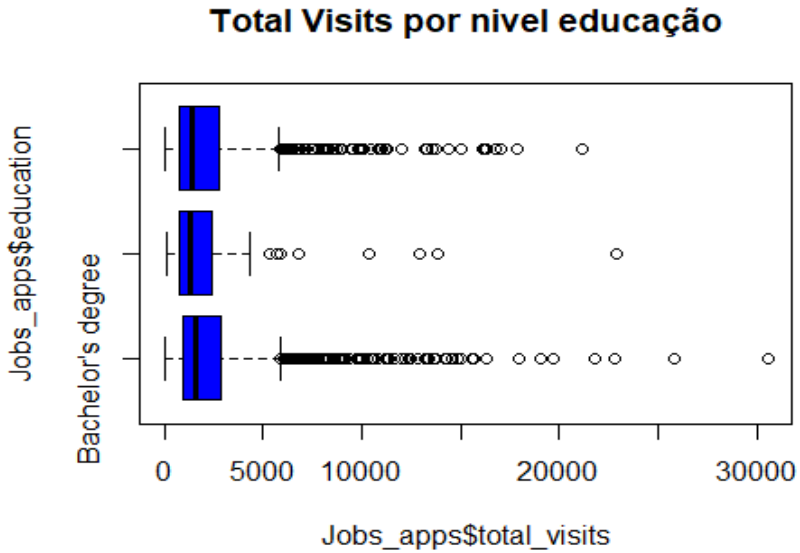



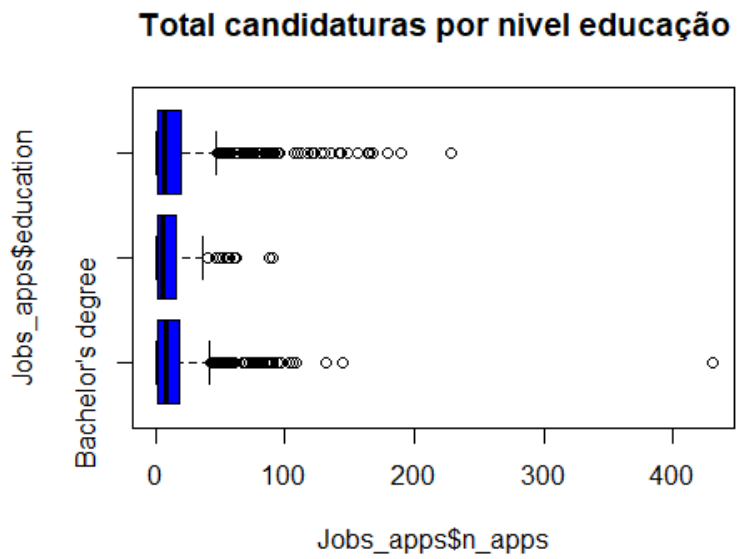

```
Jobs_apps %>%
  filter(total_visits >= 1L & total_visits <=
7473L) %>%
  ggplot() +
    aes(x = job_region) +
    geom_bar(fill = "#112446") +
    labs(title = "Regiões") +
    theme_minimal()
```

# Data Understanding

## Educação

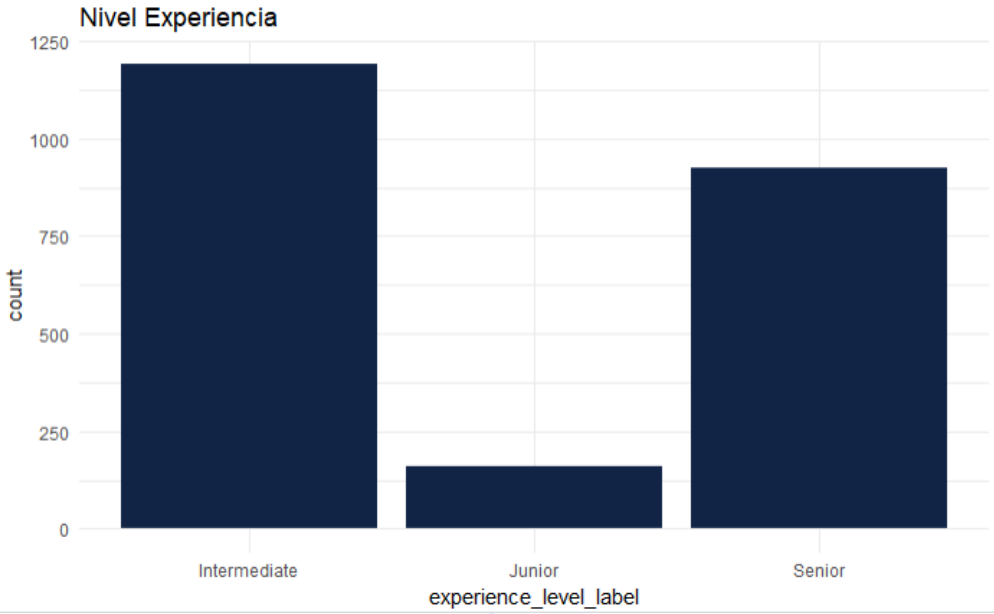


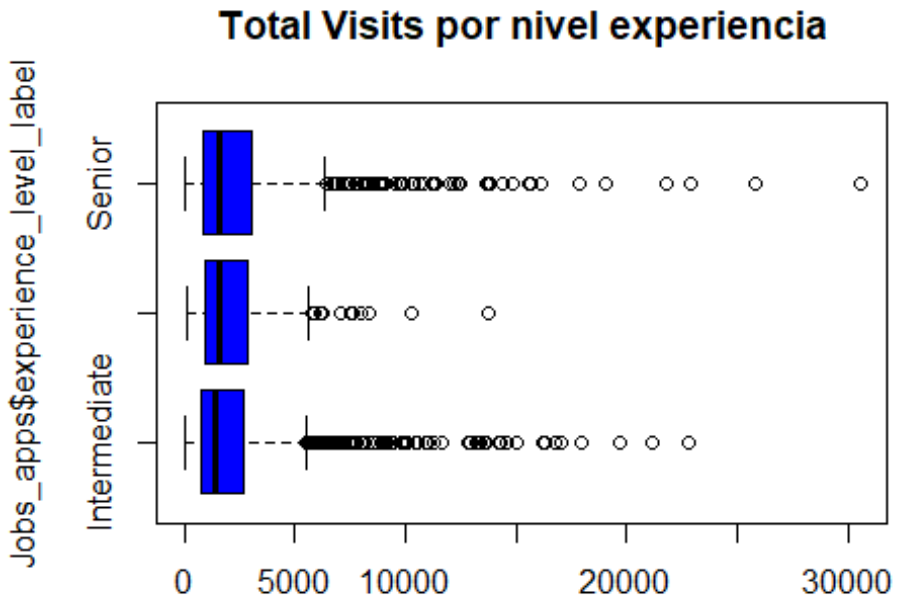



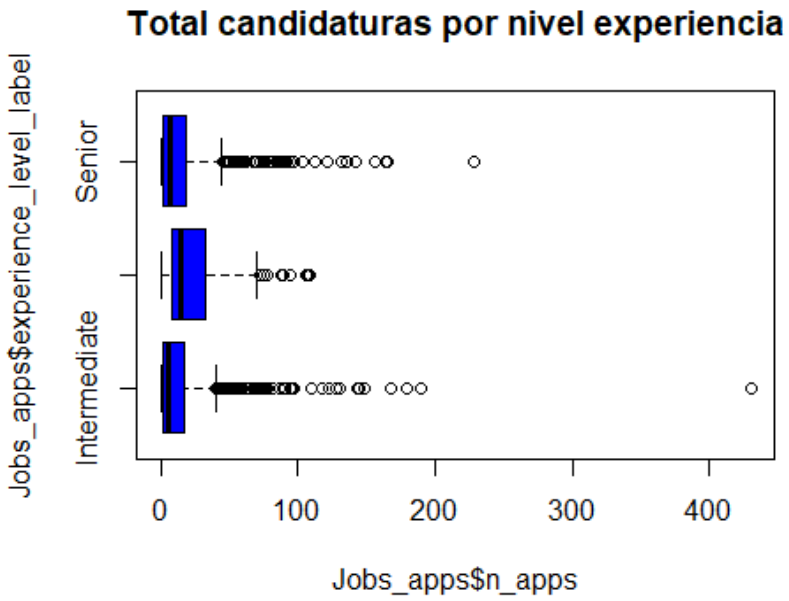



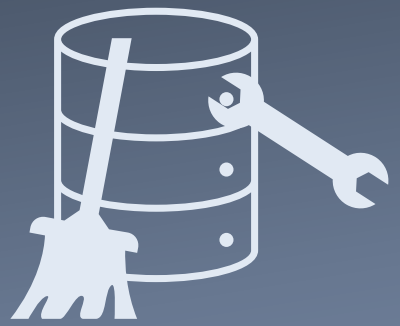
# Data Understanding

## Experiência





# 3. Data Preparation

# Feature Engineering

1. Transformação de dados do tipo True e False em 0 e 1
2. Agregação de variáveis como full remote e partial remote
3. Transformação do nível de experiência num rank (1, 2, 3)
4. Criação de variáveis de visitas por dia, candidaturas por dia, e handshake por dia
5. Transformação da variável education numa variável binária (1, 0)
6. Criação de um categoria de trabalho “outros” para categorias com menos de 100 ofertas
7. One-hot-encoding na variável categoria e comp_type

## #Preparar dados

```
DataSet <-
  DataSet %>%
  mutate(show_sal_rate=if_else((show_salary == "True"| show_rate=="True") ,1, 0,missing = NULL))%>%
  mutate(home_work=if_else((full_remote == "True"| partial_remote=="True") ,1, 0,missing = NULL))%>%
  mutate(education_req=if_else((education == "Bachelor's degree"| education=="Master's degree") ,1,
0,missing = NULL))%>%
  mutate(exp_nivel=if_else(experience_level_label== "Senior" ,3,if_else(experience_level_label==
"Intermediate",2,1) ))%>%
  mutate(languagePT=if_else(preferred_language== "pt",1,0 ))%>%
  mutate(regionPT=if_else(job_region=="PT",1,0))%>%
  mutate(contractor=if_else(contractor=="True",1,0))%>%
  mutate(consultancy=if_else(consultancy=="True",1,0))%>%
  mutate(permanent=if_else(permanent=="True",1,0))%>%
  mutate(visa_support=if_else(visa_support=="True",1,0))%>%
  mutate(relocation_paid=if_else(relocation_paid=="True",1,0))%>%
  mutate(full_time=if_else(job_type=="Full-time",1,0))%>%
  mutate(vist_dia=if_else(dias_up==0 & total_visits!=0,total_visits,total_visits/dias_up))%>%
  mutate(apps_dia=if_else(dias_up==0 & n_apps!=0,n_apps,n_apps/dias_up))%>%
  mutate(hands_dia=if_else(dias_up==0 & n_hands!=0,n_hands,n_hands/dias_up))%>%
  mutate(hierd=if_else(n_hired>0,1,0))%>%
  mutate(engaged=if_else(n_eng>0,1,0))
```

#Limpar dataset

```
DataSet<-
DataSet %>%
select(-job_region,-education,-preferred_language,-full_remote,-partial_remote,-show_salar
y,-show_rate,-experience_level_label,-job_type,-cout,-total_visits,-n_apps)
```

#Categoty

```
DataSet <-
  DataSet %>%
  group_by(job_category)%>%
  mutate(cout=n())%>%
  ungroup%>%
  mutate(job_category=if_else(cout<100,"outros",job_category))
```

##Defenir factos

```
DataSet$job_category<-as.factor(DataSet$job_category)
DataSet$comp_type<-as.factor(DataSet$comp_type)
```

##One Hot Encoding

```
DataSet<-one_hot(as.data.table(DataSet))
```

# Data Understanding

```
       id          job_region        job_category         contractor        consultancy         education          permanent         preferred_language full_remote        partial_remote
 Min.   :11474   Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278        Length:2278
 1st Qu.:12270   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :13127   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :13181
 3rd Qu.:14072
 Max.   :15072

 show_salary         show_rate         relocation_paid    visa_support       experience_level_label   job_type          total_visits      comp_score          dias_up          n_hands
 Length:2278        Length:2278        Length:2278        Length:2278        Length:2278            Length:2278        Min.   :    1.0   Min.   :-4.47263   Min.   :  0.0   Min.   :   0.00
 Class :character   Class :character   Class :character   Class :character   Class :character       Class :character   1st Qu.:  847.8   1st Qu.:-0.39795   1st Qu.: 60.0   1st Qu.:   0.00
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character       Mode  :character   Median : 1519.0   Median :-0.13100   Median :108.0   Median :   0.00
                                                                                                                     Mean   : 2449.6   Mean   :-0.03382   Mean   :124.7   Mean   :  27.36
                                                                                                                     3rd Qu.: 2859.5   3rd Qu.: 0.37811   3rd Qu.:176.0   3rd Qu.:  26.00
                                                                                                                     Max.   :30539.0   Max.   : 1.42838   Max.   :424.0   Max.   :2439.00
                                                                                                                                       NA's   :260

    n_apps            n_eng            n_hired          min_engaged
 Min.   :  0.00   Min.   :  0.000   Min.   :0.00000   Min.   :  0.00
 1st Qu.:  2.00   1st Qu.:  0.000   1st Qu.:0.00000   1st Qu.:  8.00
 Median :  7.00   Median :  0.000   Median :0.00000   Median : 21.00
 Mean   : 15.13   Mean   :  1.944   Mean   :0.09702   Mean   : 37.15
 3rd Qu.: 19.00   3rd Qu.:  1.000   3rd Qu.:0.00000   3rd Qu.: 46.75
 Max.   :430.00   Max.   :134.000   Max.   :4.00000   Max.   :341.00
                                                      NA's   :1496
```

# Dicionário de DataSet

| Coluna | Descrição | Tipo |
|---|---|---|
| id | Id's das ofertas que têm data de 1ª publicação | Nominal |
| job_category_Back-end Developer | Categoria da Oferta 1=True; 0=False | Binário |
| job_category_Data Scientist | Categoria da Oferta 1=True; 0=False | Binário |
| job_category_DevOps / Sysadmin | Categoria da Oferta 1=True; 0=False | Binário |
| job_category_Front-end Developer | Categoria da Oferta 1=True; 0=False | Binario |
| job_category_Full-stack Developer | Categoria da Oferta 1=True; 0=False | Binario |
| job_category_outros | Categoria da Oferta 1=True; 0=False | Binario |
| job_category_Product/Project Management | Categoria da Oferta 1=True; 0=False | Binario |
| job_category_QA / Testing | Categoria da Oferta 1=True; 0=False | Binario |
| job_category_UX / UI Designer | Categoria da Oferta 1=True; 0=False | Binário |
| contractor | Vinculo tipo contractor; 1=True; 0=False | Binario |
| consultancy | Vinculo tipo consultancy 1=True; 0=False | Binario |
| permanent | Vinculo tipo permanente 1=True; 0=False | Binário |
| relocation_paid | Relocalização paga 1=True; 0=False | Binário |
| dias_up | Dias desde que a oferta foi publicada a 1ª vez até passar a "Closed" ou a "Unpublished" ou até ao dia 2022-03-15 | Nominal |
| n_hands | Nº de handshakes associados à oferta que não estão como 'Requested' ou 'Rejected' | Nominal |
| n_eng | Nº de candidaturas, à oferta, que passaram à fase Engaged | Nominal |
| n_hired | Nº de candidaturas, à oferta, que passaram à fase Hierd | Nominal |
| min_engaged | Dias desde que a oferta foi publicada a 1ª vez até à primeira candidatura chegar à fase de Engaged | Nominal |
| regionPT | É em Portugal; 1=True; 0=False | Binário |
| full_time | 1=True; 0=False | Binário |
| vist_dia | Nº de visitas por dia (total_visits/dias_up) | Nominal |
| apps_dia | Nº de candidaturas por dia (n_apps/dias_up) | Nominal |
| hands_dia | Nº de hankshakes por dia (n_hands/dias_up) | Nominal |

# Dicionário de DataSet

| Coluna | Descrição | Tipo |
|---|---|---|
| hierd | Oferta teve pelo menos uma candidatura como Hierd; 1=True; 0=False | Binário |
| engaged | Oferta teve pelo menos uma candidatura como Engaged; 1=True; 0=False | Binário |
| total_visits | Numero de visitas total à oferta | Nominal |
| n_apps | Nº de candidaturas (distintas) à oferta | Nominal |
| min_app | Dias desde que a oferta foi publicada a 1ª vez até à primeira candidatura | Nominal |
| comp_type_Boutique | Tipo de empresa 1=True; 0=False | Binário |
| comp_type_Corporate | Tipo de empresa 1=True; 0=False | Binário |
| comp_type_Not classified | Tipo de empresa 1=True; 0=False | Binário |
| comp_type_Scale-up | Tipo de empresa 1=True; 0=False | Binário |
| comp_type_SME | Tipo de empresa 1=True; 0=False | Binário |

# 3. Modeling | Clustering

# Modeling: Clustering

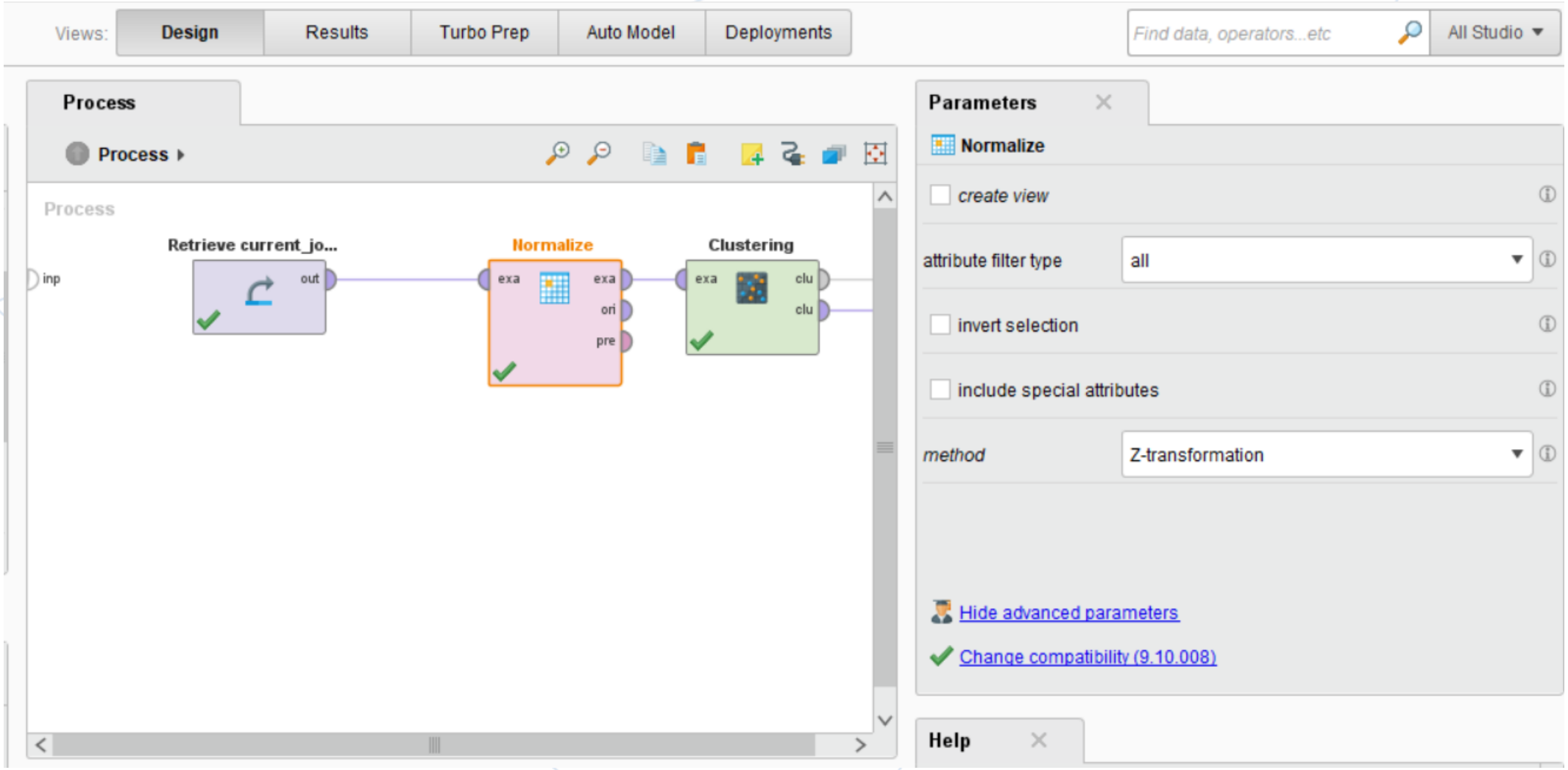
Porto Business School / University of Porto

Objetivo: **Segmentação dos jobs**


Algoritmos usados: K-means & DBSCAN & K-medoids

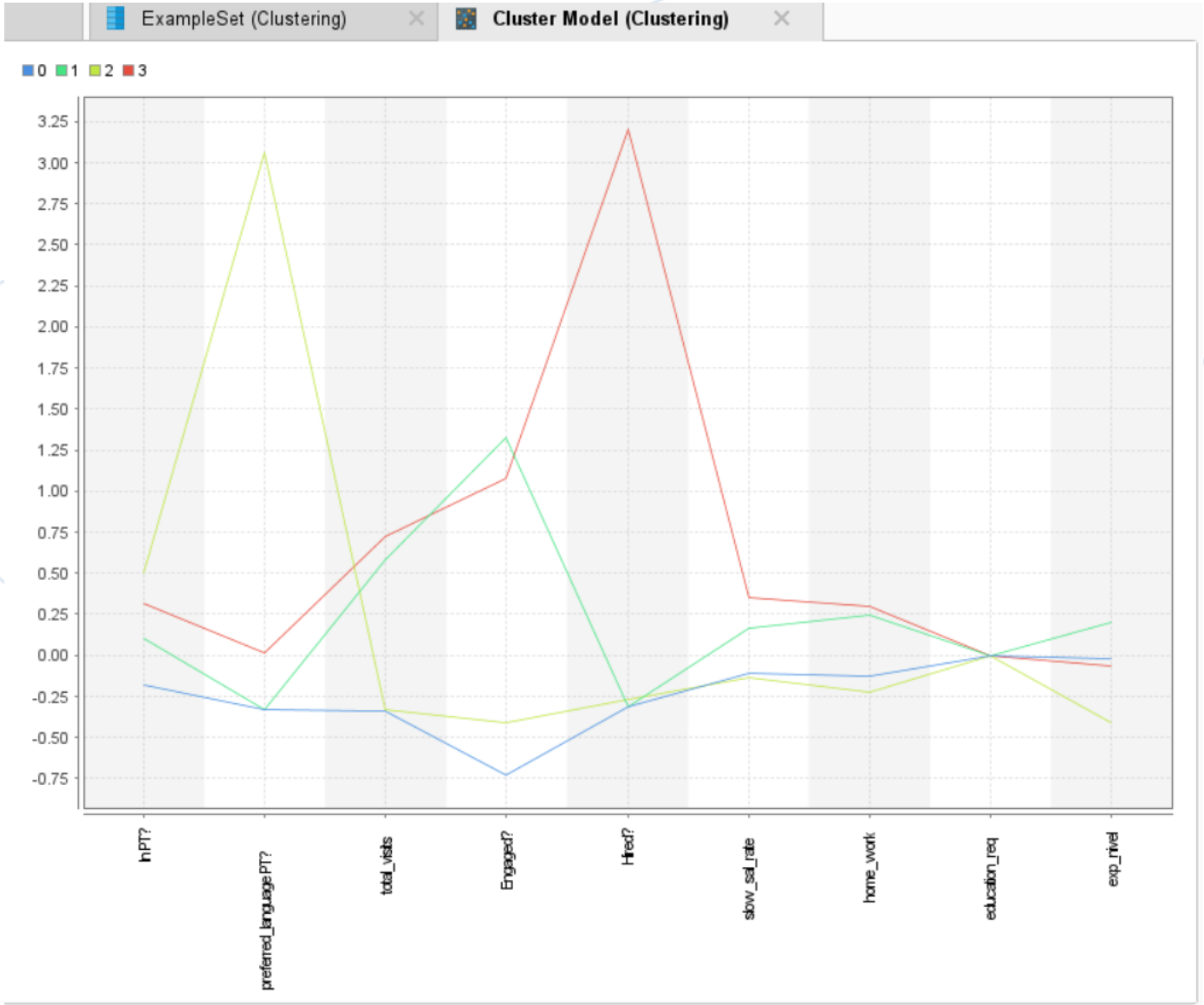
# Modeling: Clustering -1ª iteração – K-means

- **Pipeline**



# Modeling: Clustering -1ª iteração – K-means

- **Resultados**



# Modeling: Clustering -1ª iteração – K-means

- **Resultados**

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 |
|---|---|---|---|---|
| In PT? | -0.177 | 0.102 | 0.506 | 0.317 |
| preferred_language PT? | -0.326 | -0.326 | 3.061 | 0.017 |
| total_visits | -0.339 | 0.584 | -0.334 | 0.725 |
| Engaged? | -0.732 | 1.322 | -0.413 | 1.074 |
| Hired? | -0.312 | -0.312 | -0.267 | 3.206 |
| slow_sal_rate | -0.111 | 0.162 | -0.139 | 0.356 |
| home_work | -0.127 | 0.242 | -0.228 | 0.302 |
| education_req | 0 | 0 | 0 | 0 |
| exp_nivel | -0.020 | 0.199 | -0.412 | -0.066 |

# Modeling: Clustering -1ª iteração – K-means

## Conclusões

- Inicialmente o objetivo do nosso modelo clustering era descobrir o que torna o job apelativo. E por isso consideramos variáveis como “total views” e “nº aplicações”. Mas percebemos que o cluster por si só não nos iria responder a essa pergunta. Seria como se estivessemos a tentar forçar o algoritmo a definir um grupo de jobs apelativos. Por isso mudamos o objetivo do clustering para “segmentação dos jobs”. Nesse sentido, decidimos retirar as variáveis mencionadas anteriormente e focomo-nos em variáveis caracteristicas do job.
- Outro erro que cometemos foi na normalização das variáveis. Normalizamos todas as variáveis através duma Z-transformation. Neste caso não faz sentido normalizar variáveis binárias. Vamos corrigir este ponto na iteração 2.

# Modeling: Clustering – 2ª iteração – K-means

- **Dados de entrada:**

| Variáveis de entrada |
|---|
| VISA_SUPPORT |
| SHOW_SALARY |
| REMOTE |
| EDUCATION_REQ |
| EXP_LEVEL |
| REGION_PT |
| FULL_TIME |

Result History | ExampleSet (//Landing Jobs/Dados/2022053 [...] rmodels_Cluster_RevD_JOBSCHARACTERISTICS)

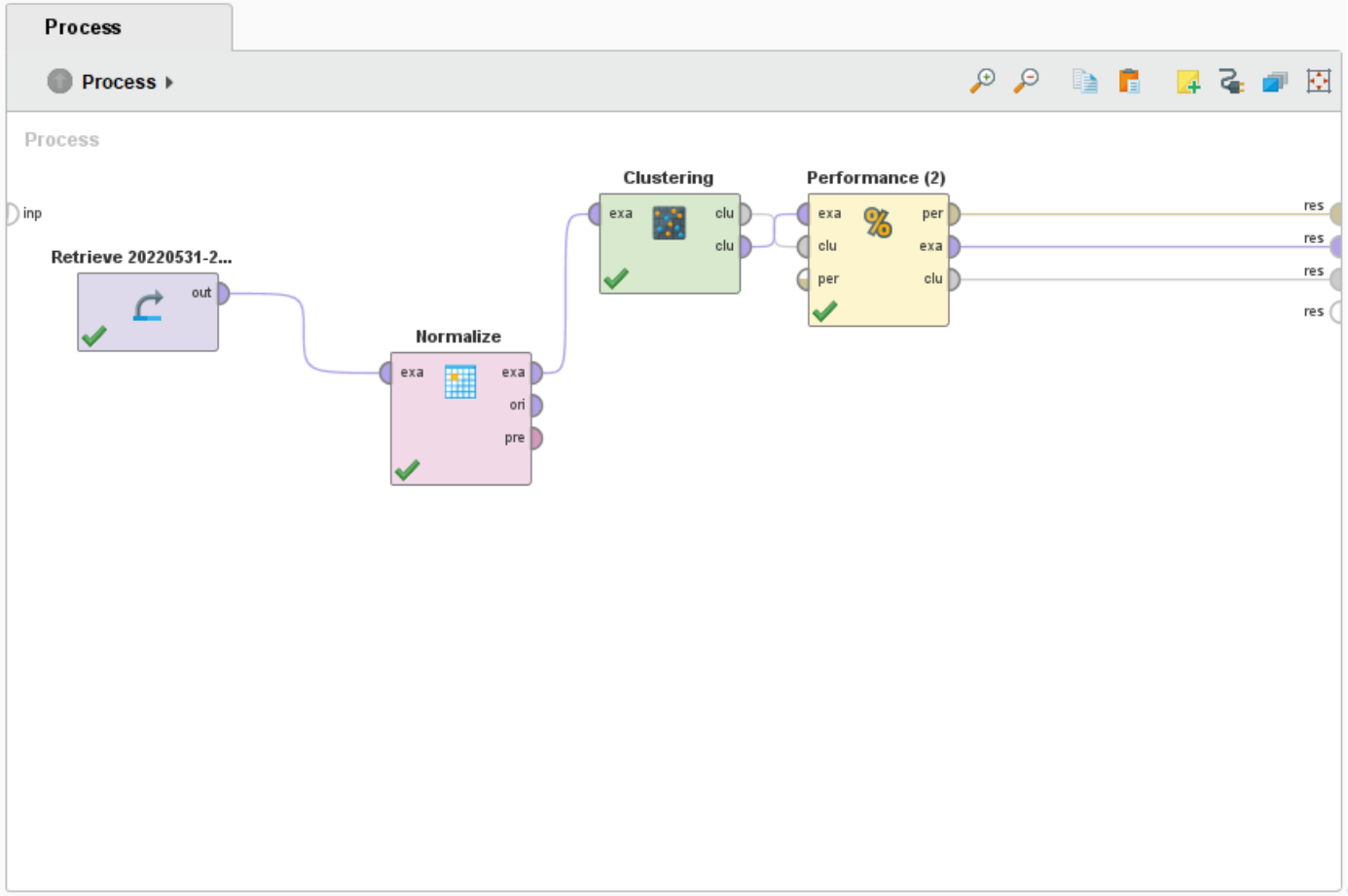
Data | Statistics | Visualizations | Annotations

Open in | Turbo Prep | Auto Model | Filter (

| Row No. | VISA_SUPP... | SHOW_SAL... | REMOTE | EDUCATION... | EXP_LEVEL | REGION_PT | FULL_TIME |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 3 | 1 | 1 |
| 2 | 0 | 1 | 1 | 1 | 2 | 1 | 1 |
| 3 | 0 | 1 | 1 | 1 | 3 | 1 | 1 |
| 4 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5 | 0 | 1 | 1 | 1 | 2 | 1 | 1 |
| 6 | 0 | 1 | 1 | 1 | 3 | 1 | 1 |
| 7 | 0 | 1 | 1 | 0 | 3 | 1 | 1 |
| 8 | 0 | 1 | 1 | 0 | 3 | 1 | 1 |
| 9 | 0 | 1 | 1 | 0 | 2 | 1 | 1 |
| 10 | 0 | 1 | 1 | 0 | 2 | 1 | 1 |
| 11 | 0 | 0 | 1 | 0 | 3 | 1 | 1 |
| 12 | 0 | 1 | 1 | 1 | 3 | 0 | 0 |
| 13 | 0 | 1 | 1 | 0 | 3 | 1 | 0 |
| 14 | 0 | 1 | 1 | 0 | 3 | 1 | 1 |
| 15 | 0 | 0 | 1 | 0 | 3 | 1 | 1 |
| 16 | 0 | 0 | 1 | 0 | 3 | 1 | 1 |
| 17 | 0 | 1 | 1 | 1 | 3 | 1 | 1 |
| 18 | 0 | 0 | 1 | 0 | 2 | 1 | 1 |
| 19 | 0 | 1 | 1 | 1 | 3 | 1 | 1 |
| 20 | 0 | 1 | 1 | 1 | 2 | 1 | 1 |
| 21 | 0 | 0 | 1 | 1 | 2 | 1 | 1 |
| 22 | 0 | 0 | 1 | 1 | 2 | 1 | 1 |
| 23 | 0 | 0 | 1 | 1 | 2 | 1 | 1 |

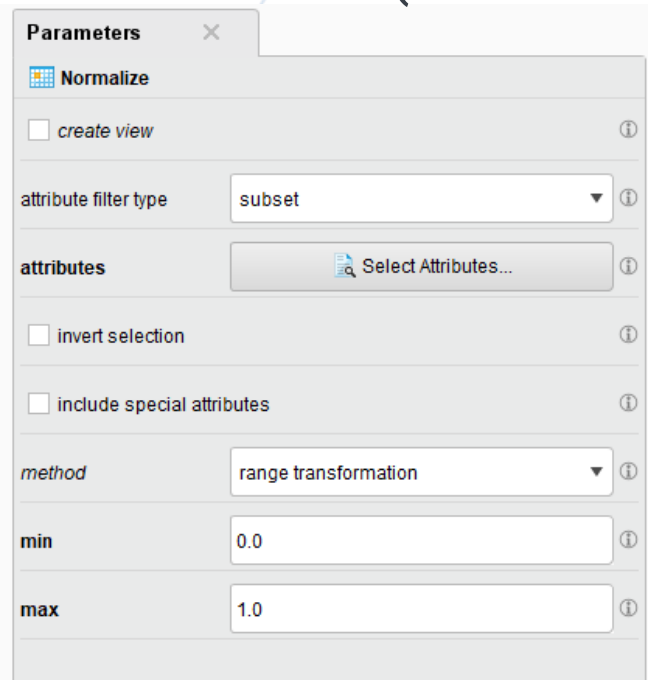
# Modeling: Clustering – 2ª iteração – K-means

- **Dados de entrada:**



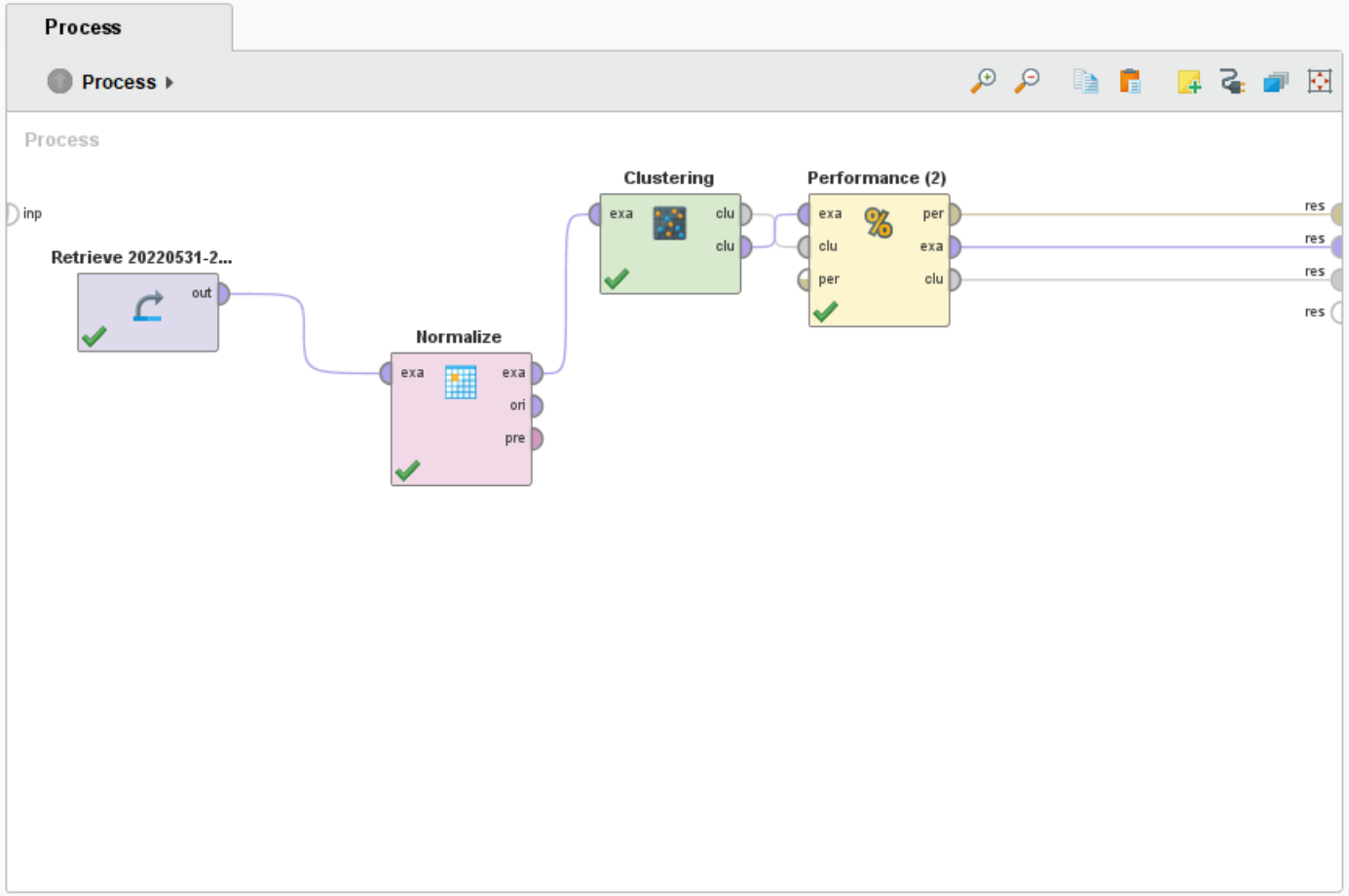

Normalização:

- EXP_LEVEL(não binária)

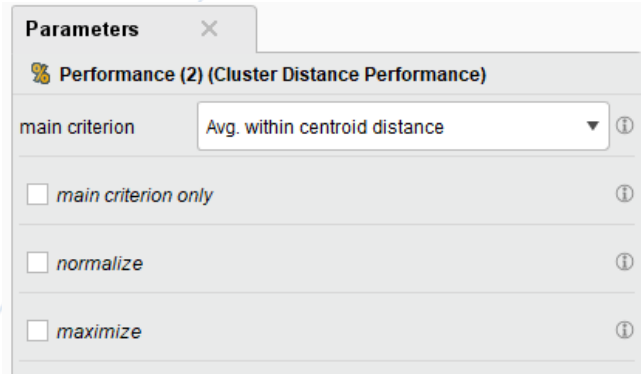


# Modeling: Clustering – 2ª iteração – K-means

- **Dados de entrada:**



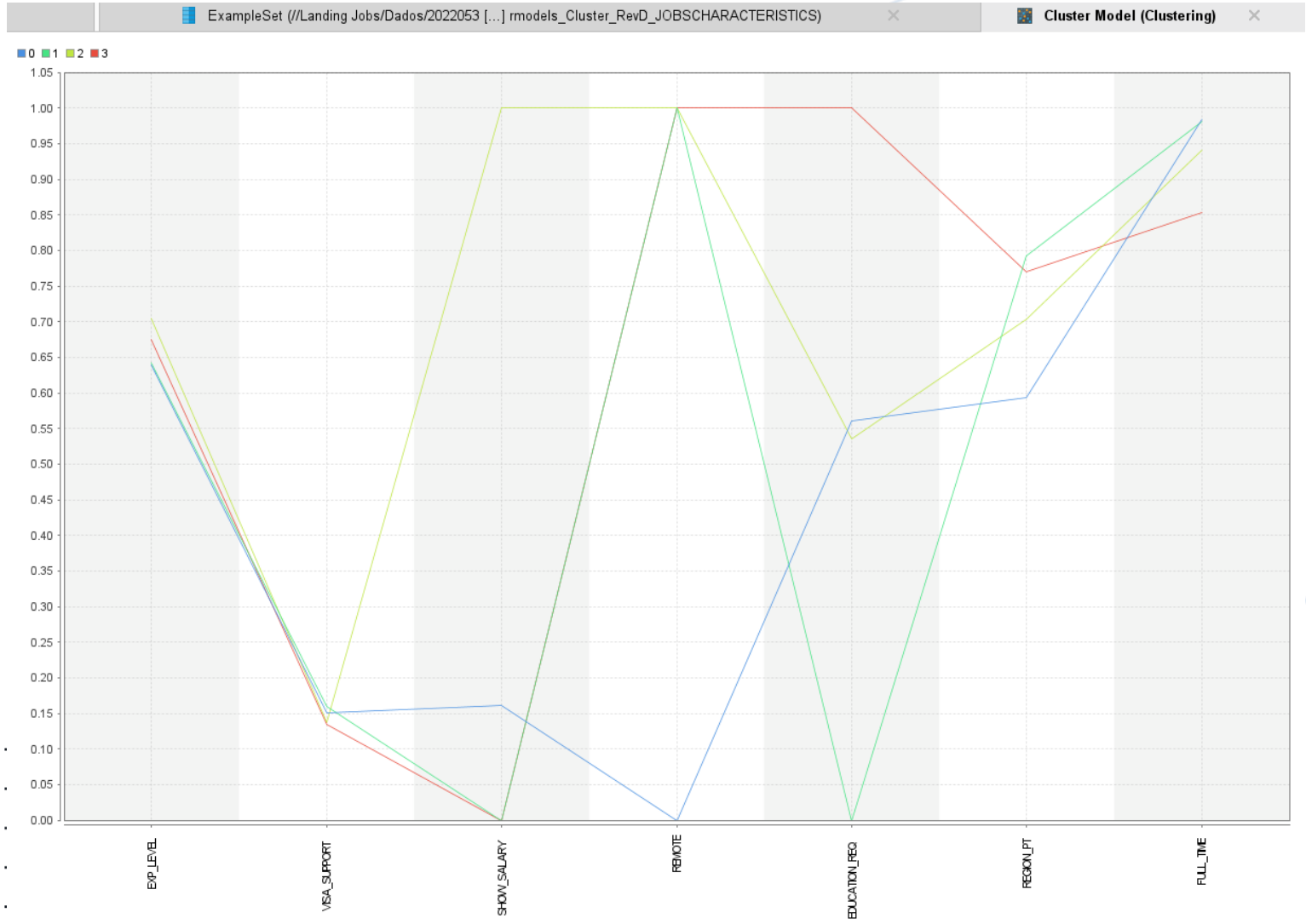

Nº de cluster determinado pelo algoritmo “Cluster Distance Performance”




Nº de clusters:4

# Modeling: Clustering – 2ª iteração – K-means

- **Análise de resultados:**




- As variáveis EXP_LEVEL e VISA_SUPPORT não se distinguem bem entre clusters;
- As variáveis SHOW_SALARY, REMOTE e EDUCATION_REQ estão bem definidas nos diferentes clusters;

# Modeling: Clustering – 2ª iteração– K-means

- **Análise de resultados:**

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 |
|---|---|---|---|---|
| EXP_LEVEL | 0.639 | 0.643 | 0.705 | 0.675 |
| VISA_SUPPORT | 0.151 | 0.160 | 0.138 | 0.135 |
| SHOW_SALARY | 0.161 | 0 | 1 | 0 |
| REMOTE | 0 | 1 | 1 | 1 |
| EDUCATION_REQ | 0.561 | 0 | 0.535 | 1 |
| REGION_PT | 0.593 | 0.793 | 0.703 | 0.770 |
| FULL_TIME | 0.983 | 0.981 | 0.941 | 0.854 |

- Cluster_0:
  - Nenhum job é remote/híbrido;
- Cluster_1:
  - Todos os jobs são remote/híbrido;
  - Nenhum job apresenta o salário;
  - Nenhum job exige grau académico;
- Cluster_2:
  - Todos os jobs são remote/híbrido;
  - Todos os jobs apresentam o salário;
- Cluster_3:
  - Todos os jobs são remote/híbrido;
  - Nenhum job apresenta o salário;
  - Todos os job exigem grau académico;

**Cluster Model**

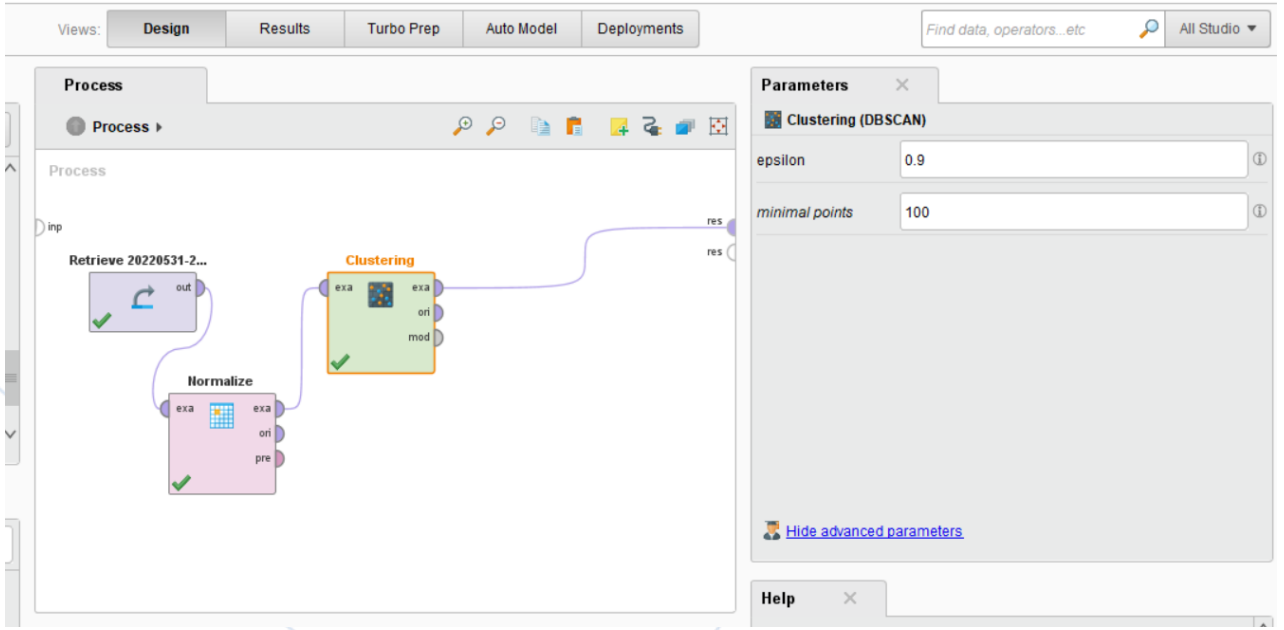
```
Cluster 0: 595 items
Cluster 1: 420 items
Cluster 2: 572 items
Cluster 3: 691 items
Total number of items: 2278
```

## Trabalhos Futuros

- Temos os jobs segmentados, seria interessante analisar o impacto de cada um dos clusters no número de aplicações e no número de “engaged”;
- Utilizar o nosso modelo de regressão (que iremos apresentar no próximo capítulo) para perceber se existem diferenças significativas nos tempos que um job demora a ter um “engage” entre os diferentes clusters. Esta análise poderia ajudar a landingjob a oferecer recomendações as empresas no sentido de obterem um engage mais rápido;
- Inclusão de mais variáveis quantitativas;
- Avaliação do cluster com o “silhouette method”;

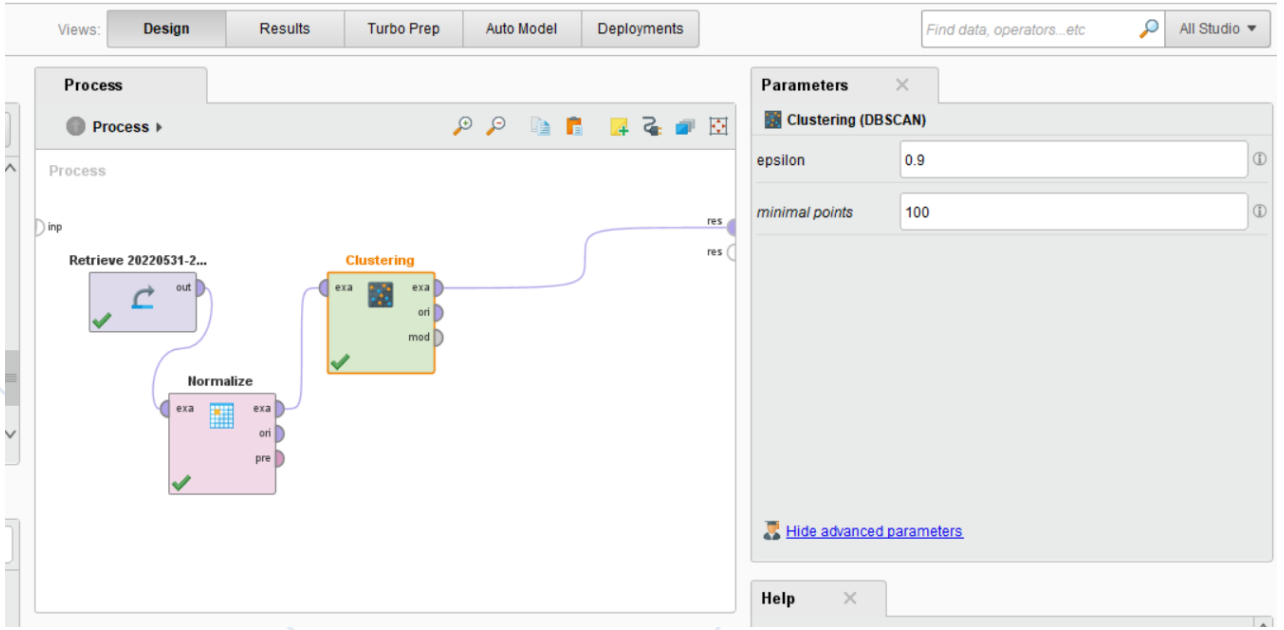
# Data Preparation: Clustering – 3ª iteração

- Na tentativa de obter novos clusters recorrendo a outro algoritmo que não o K-means, usamos o DBSCAN;
- Uma vez que as variáveis estão todas normalizadas para valores entre 0 e 1, o valor de epsilon deve ser inferior a 1;
- Como o DSBCAN estava a identificar muitos clusters decidimos aumentar o valor minimal points para obtermos um número de clusters que considerássemos aceitável;

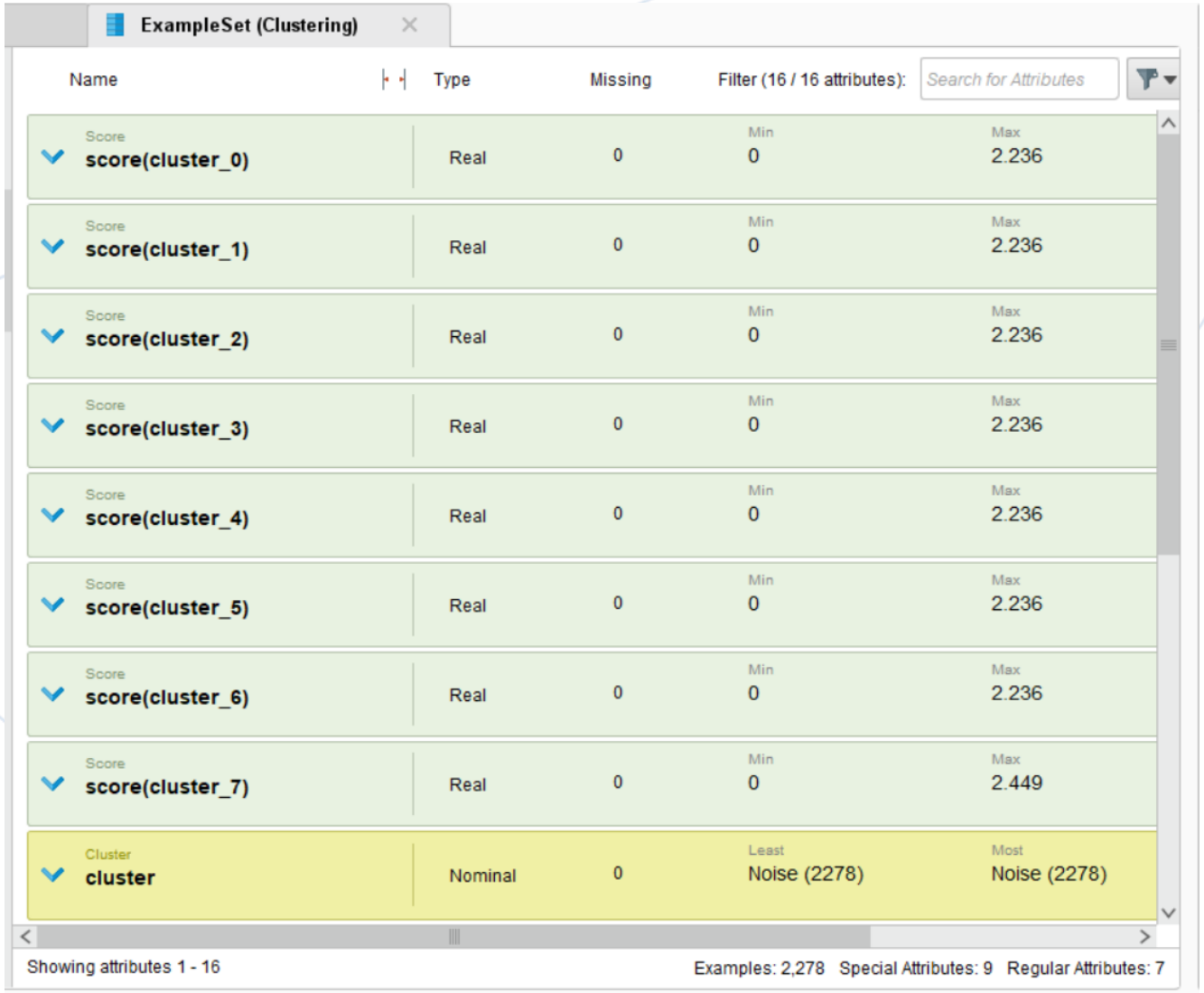


# Data Preparation: Clustering – 3ª iteração

- Na tentativa de obter novos clusters recorrendo a outro algoritmo que não o K-means, usamos o DBSCAN;
- Uma vez que as variáveis estão todas normalizadas para valores entre 0 e 1, o valor de epsilon deve ser inferior a 1;
- Como o DSBCAN estava a identificar muitos clusters decidimos aumentar o valor minimal points para obtermos um número de clusters que considerássemos aceitável;



# Data Preparation: Clustering – 3ª iteração



ExampleSet (Clustering)

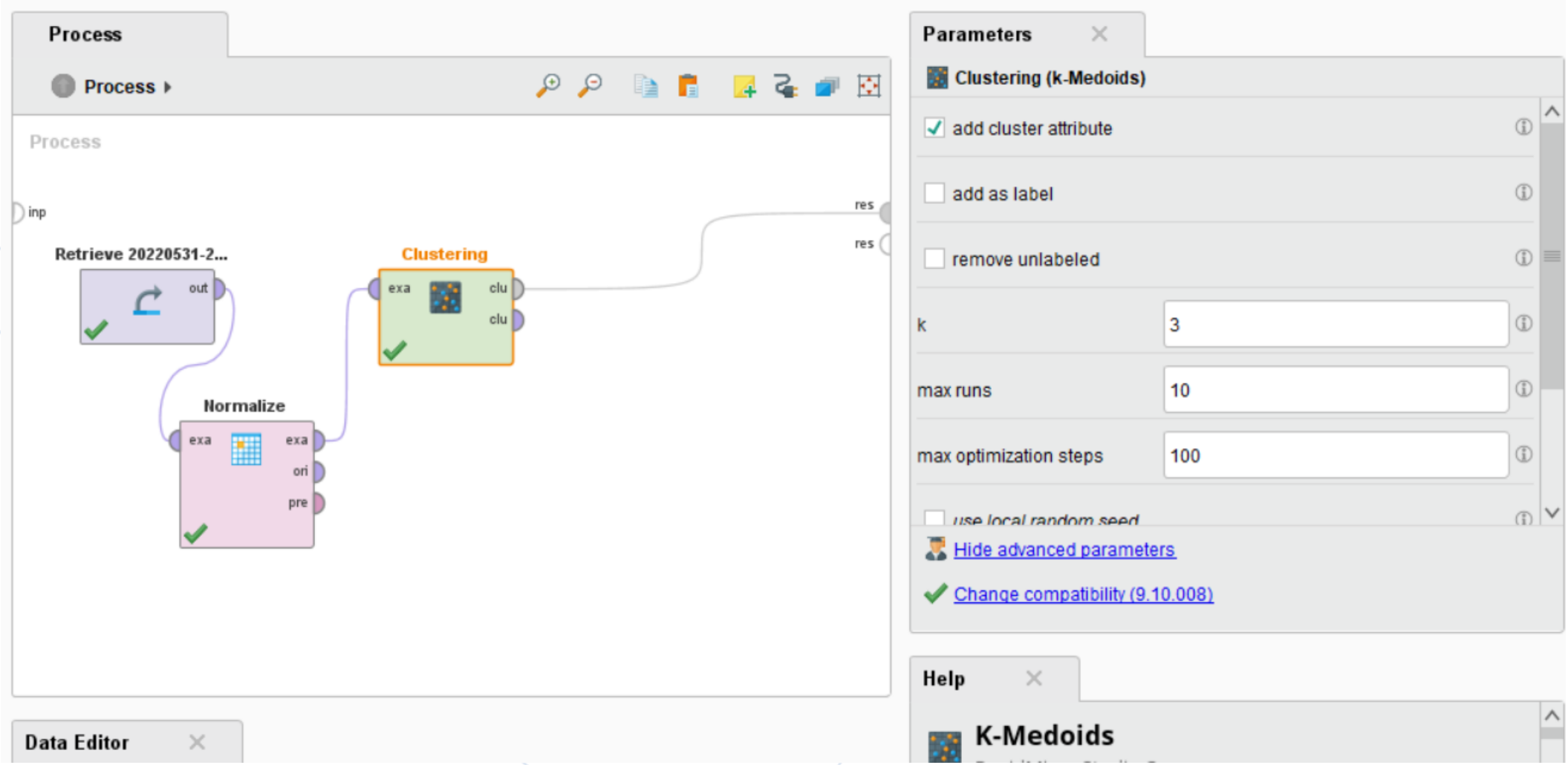
Filter (16 / 16 attributes): Search for Attributes

| Name | Type | Missing | Min | Max |
|---|---|---|---|---|
| Score score(cluster_0) | Real | 0 | 0 | 2.236 |
| Score score(cluster_1) | Real | 0 | 0 | 2.236 |
| Score score(cluster_2) | Real | 0 | 0 | 2.236 |
| Score score(cluster_3) | Real | 0 | 0 | 2.236 |
| Score score(cluster_4) | Real | 0 | 0 | 2.236 |
| Score score(cluster_5) | Real | 0 | 0 | 2.236 |
| Score score(cluster_6) | Real | 0 | 0 | 2.236 |
| Score score(cluster_7) | Real | 0 | 0 | 2.449 |
| Cluster cluster | Nominal | 0 | Least: Noise (2278) | Most: Noise (2278) |

Showing attributes 1 - 16 Examples: 2,278 Special Attributes: 9 Regular Attributes: 7

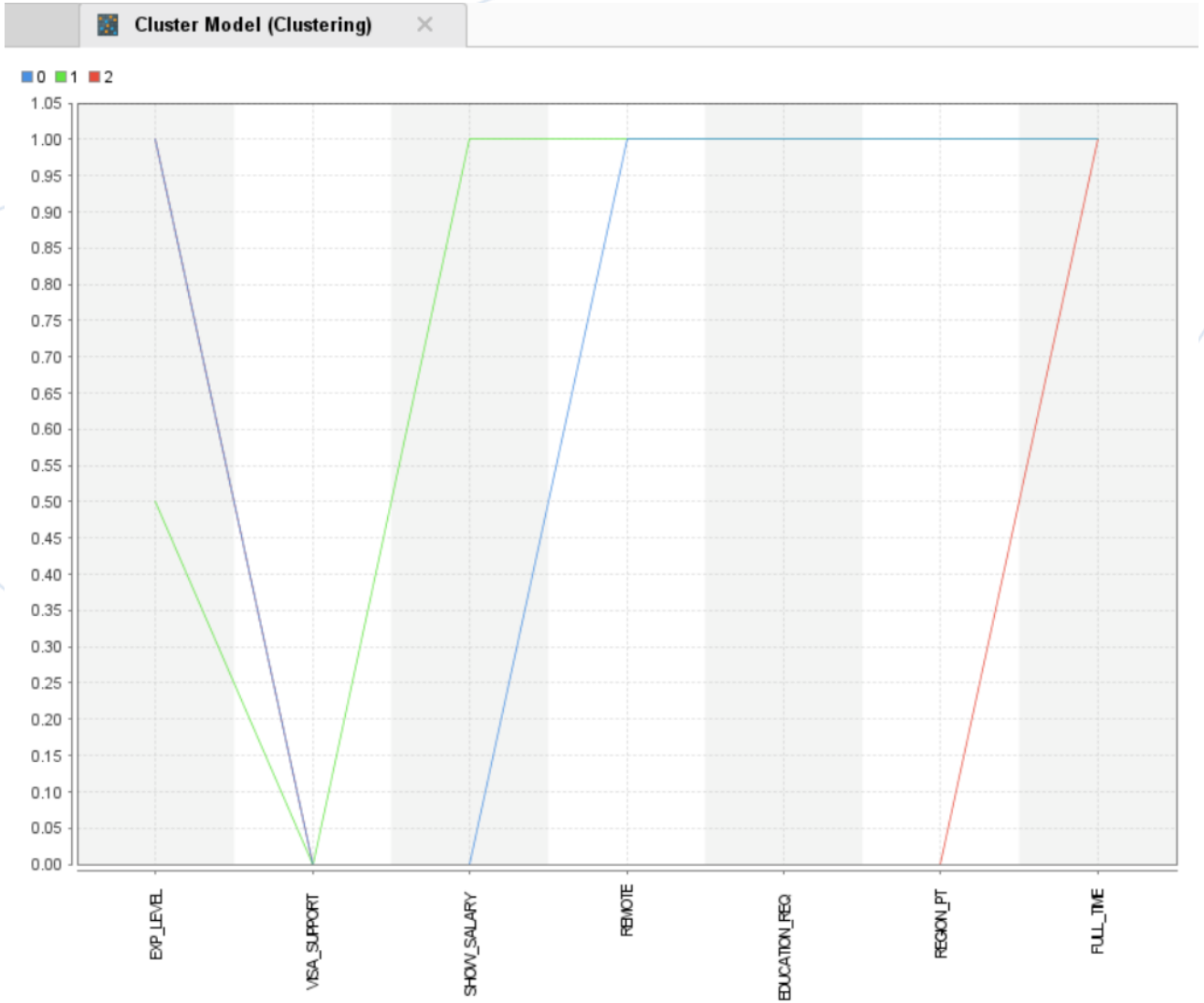
# Data Preparation: Clustering – 4ª iteração

- K-medoids



# Data Preparation: Clustering – 4ª iteração

- K-medoids



# Data Preparation: Clustering – 4ª iteração

- K-medoids

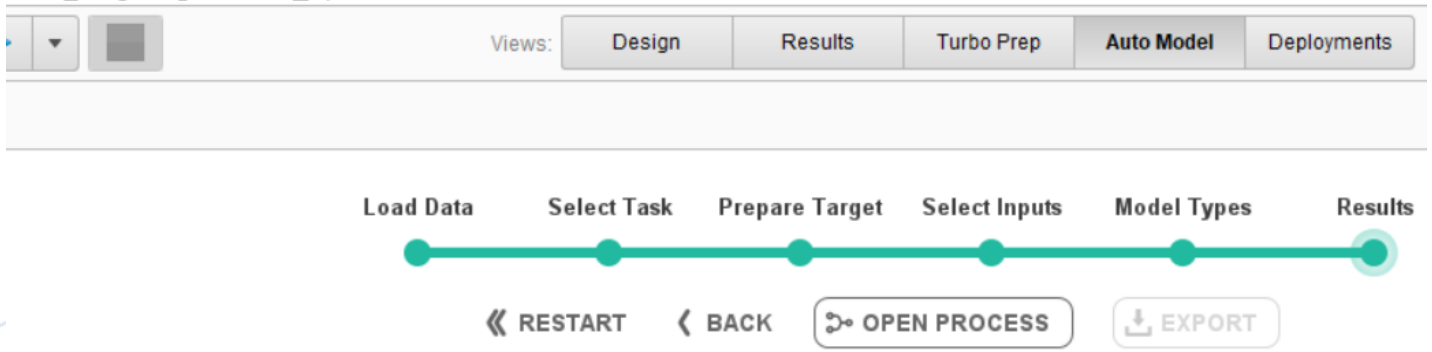
| Attribute | cluster_0 | cluster_1 | cluster_2 |
|---|---|---|---|
| EXP_LEVEL | 1 | 0.500 | 1 |
| VISA_SUPPORT | 0 | 0 | 0 |
| SHOW_SALARY | 0 | 1 | 0 |
| REMOTE | 1 | 1 | 0 |
| EDUCATION_REQ | 1 | 1 | 0 |
| REGION_PT | 1 | 1 | 0 |
| FULL_TIME | 1 | 1 | 1 |

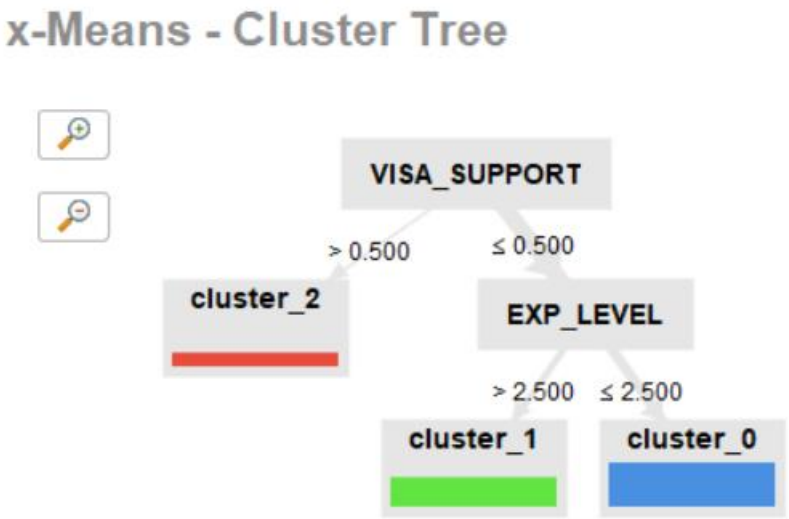
**Cluster Model**

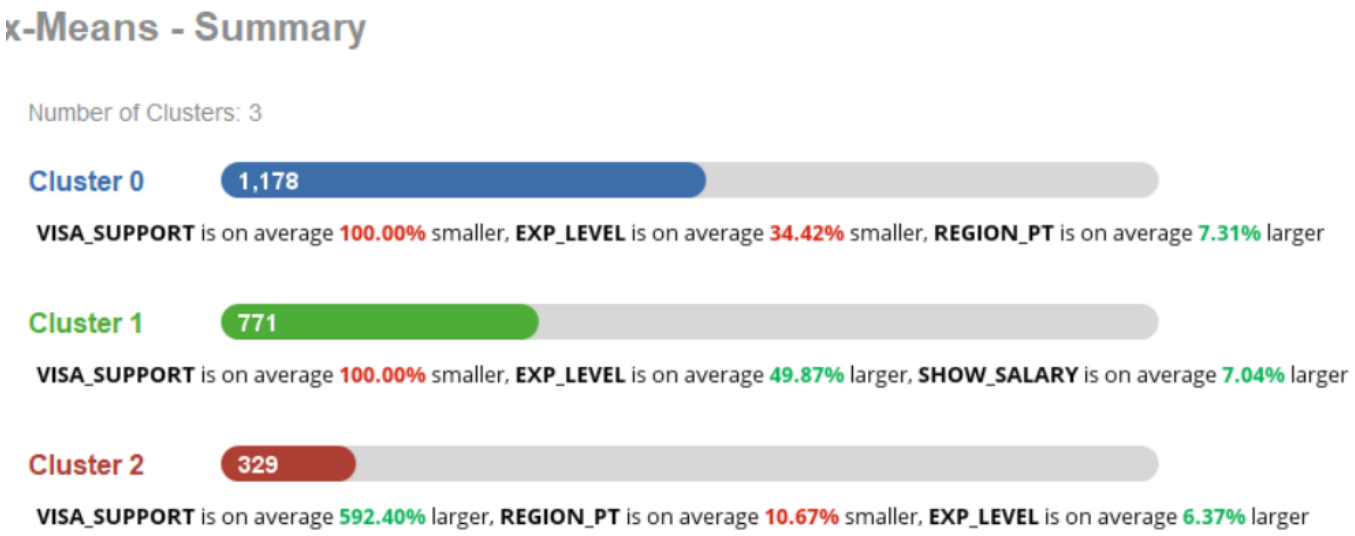
```
Cluster 0: 1157 items
Cluster 1: 608 items
Cluster 2: 513 items
Total number of items: 2278
```

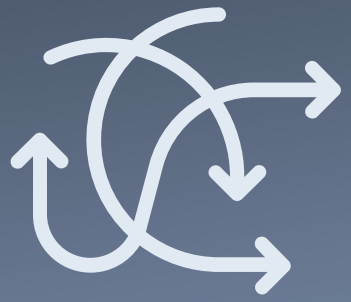
# Data Preparation: Clustering – 5ª iteração

- X-means (auto-model result)





# 3. Modelling | Regressão Linear

# Regressão Linear

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots$$

Problema: **Quanto tempo até um "job" ter, pelo menos, um "engaged"?**


# Ponto de Partida

Intervalo de tempo!

$$y = \text{Min\_Engaged}$$

Dataset:

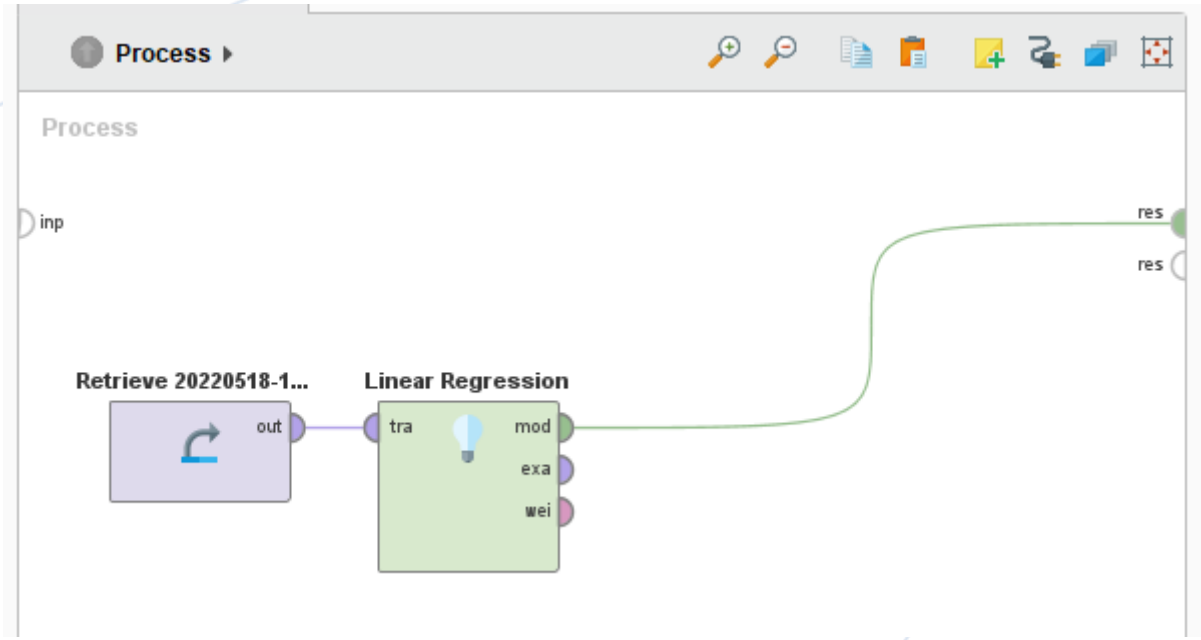
```
 [1] "job_category_Back.end.Developer"          "job_category_Data.Scientist"
 [3] "job_category_DevOps...Sysadmin"           "job_category_Front.end.Developer"
 [5] "job_category_Full.stack.Developer"        "job_category_outros"
 [7] "job_category_Product.Project.Management" "job_category_QA...Testing"
 [9] "job_category_UX...UI.Designer"            "contractor"
[11] "permanent"                                "relocation_paid"
[13] "visa_support"                             "min_engaged"
[15] "show_sal_rate"                            "home_work"
[17] "education_req"                            "exp_nivel"
[19] "languagePT"                               "regionPT"
[21] "full_time"                                "vist_dia"
[23] "apps_dia"                                 "hands_dia"
```

**Erro absoluto do modelo trivial: 37.20 dias**

# 3. Modelling | Regressão Linear

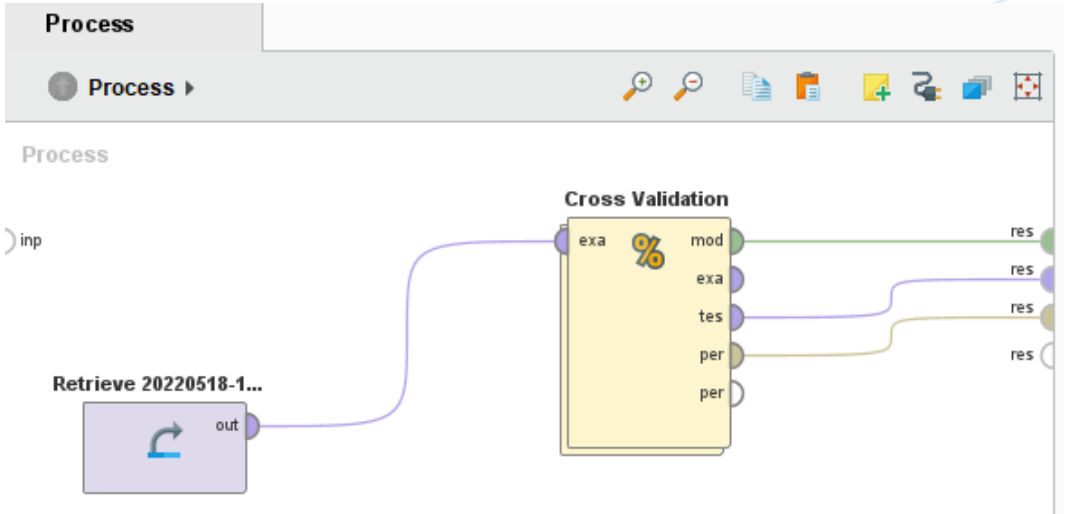
# Análise em Rapid Miner

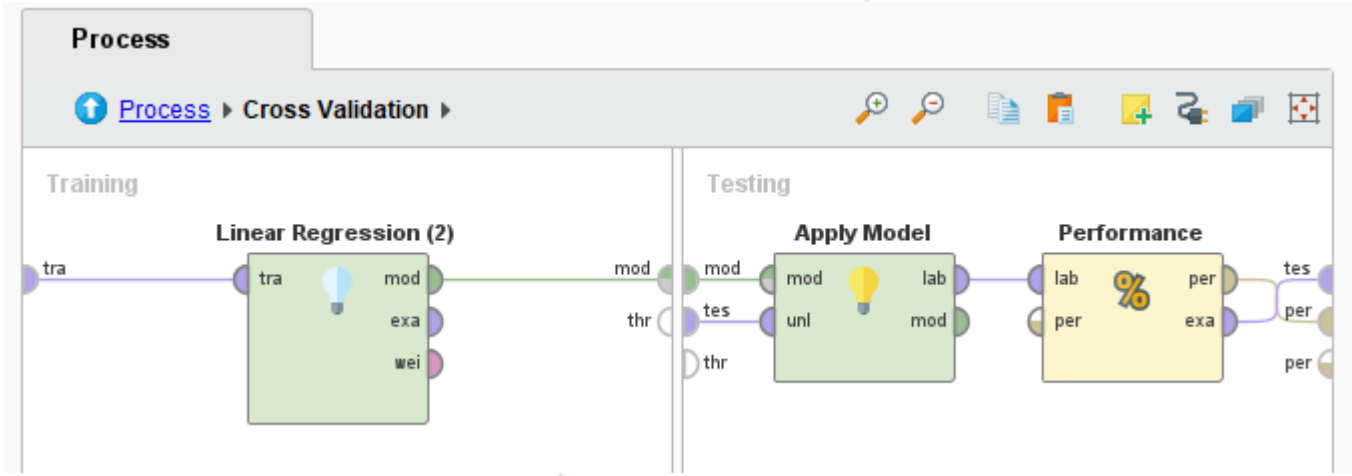
## Regressão Linear Simples



# Análise em Rapid Miner

## Regressão Linear Simples com Cross Validation



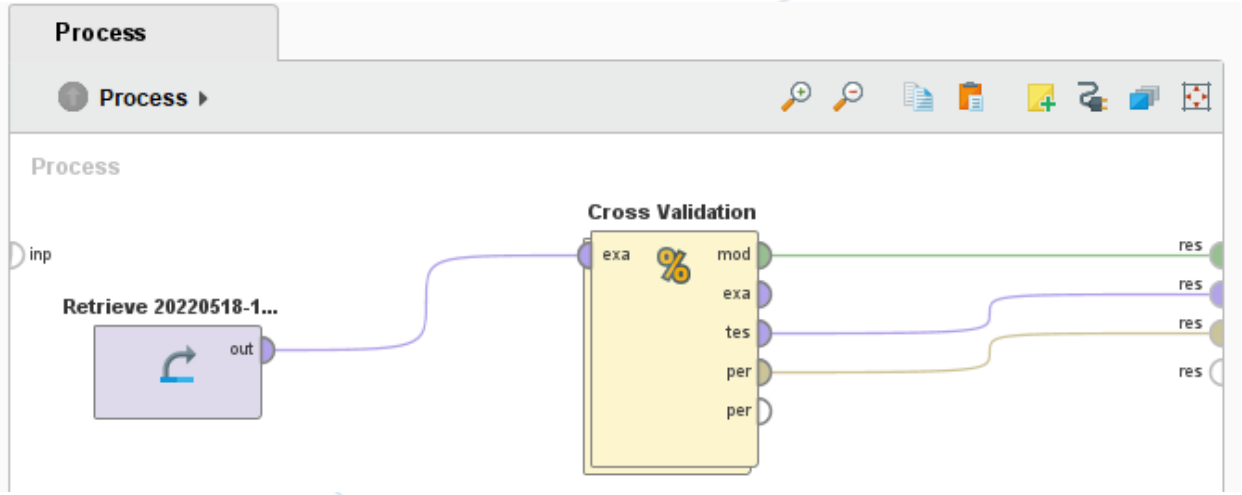



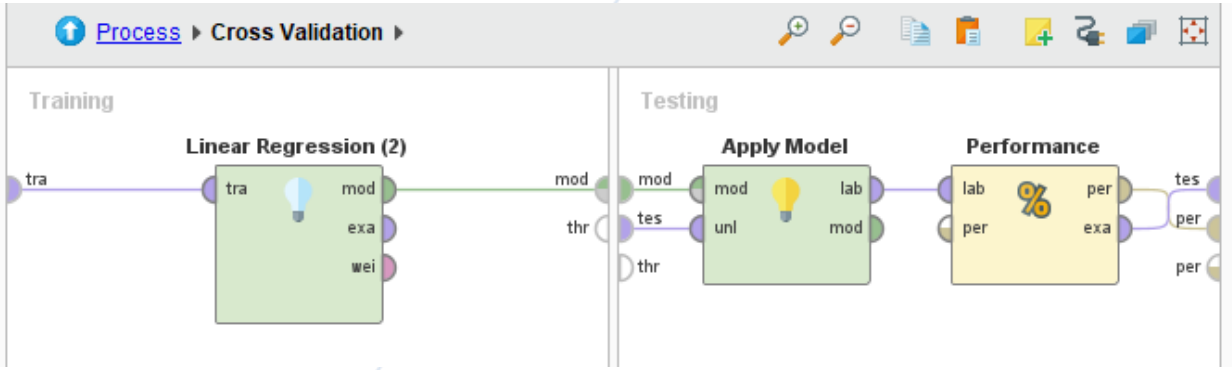

**Comentário:**
Optamos pelo cross validation, é mais rigoroso na seleção dos dados de treino e de teste.
Inicialmente começamos por selecionar o número de folds como sendo 10, mas o dataset é pequeno alteramos para 5.

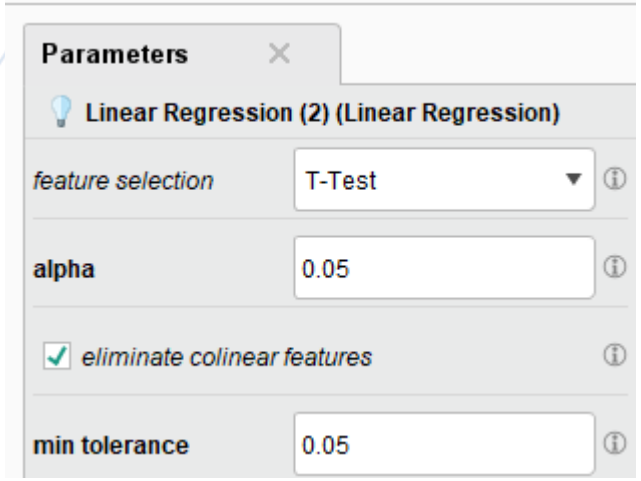
# Análise em Rapid Miner

## Regressão Linear Simples com cross validation e feature selection

Optamos por escolher o T-Test para um alfa de 0.05.



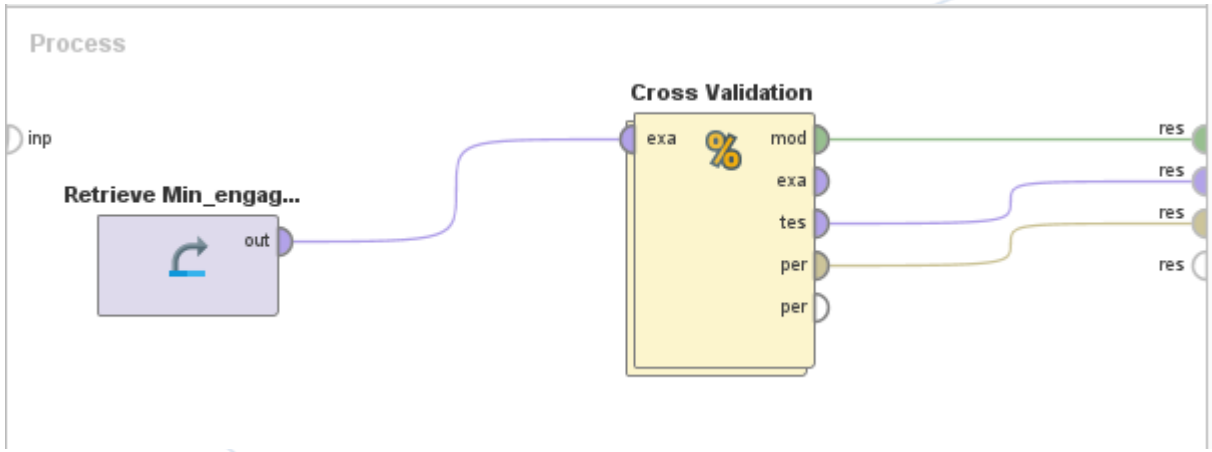



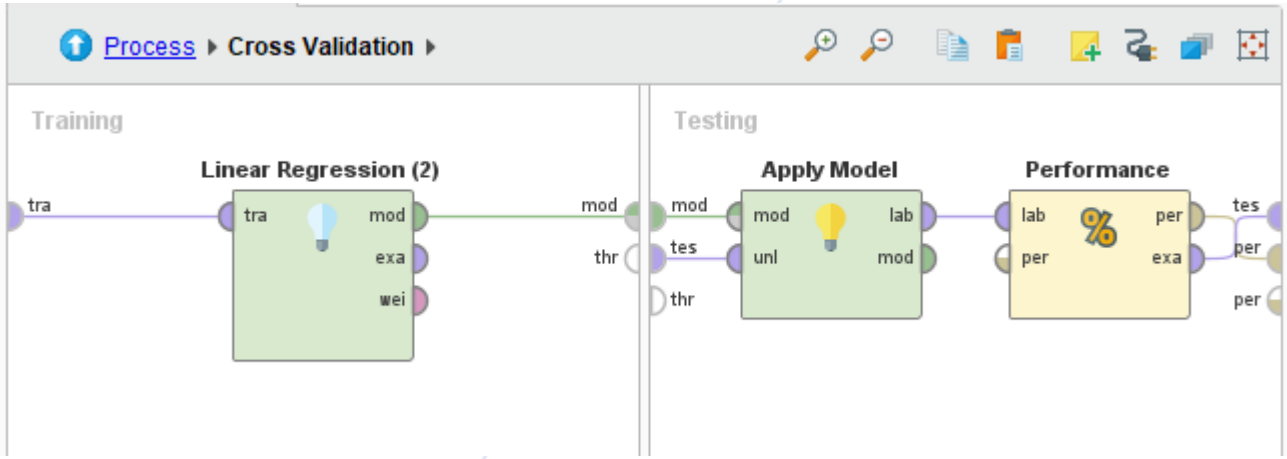
# Análise em Rapid Miner

- Na análise que fizemos em R percebemos que existem outliers que poderão estar a enviesar o nosso dataset.
- Removemos os outliers no R (explicação detalhada mais à frente).
- Exportamos o novo dataset e analisamos os resultados no Rapid Miner.

# Análise em Rapid Miner

Regressão Linear Simples (Sem Outliers) com cross validation

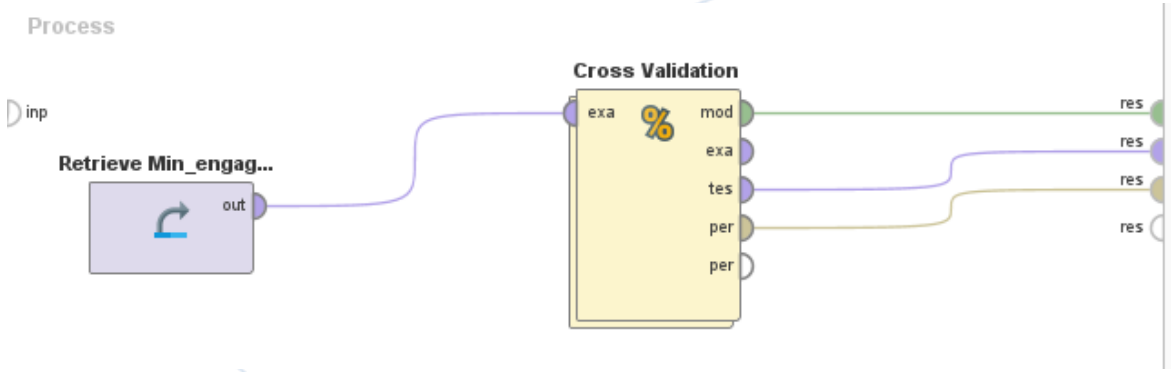


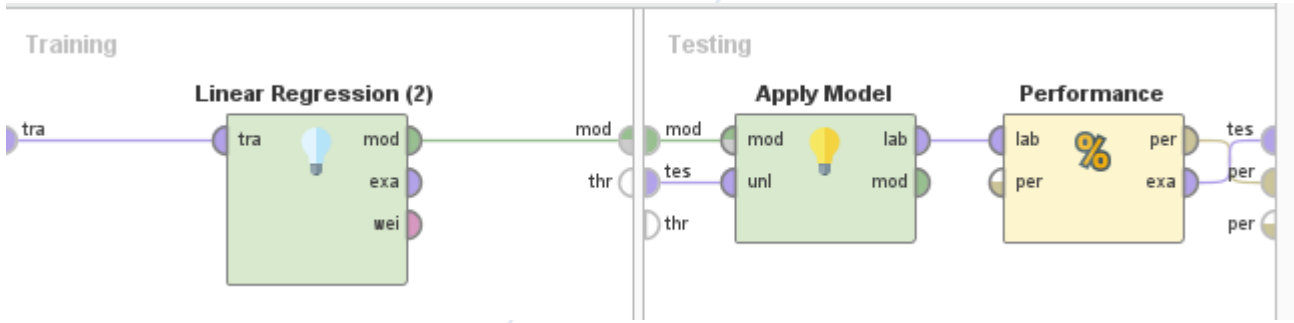



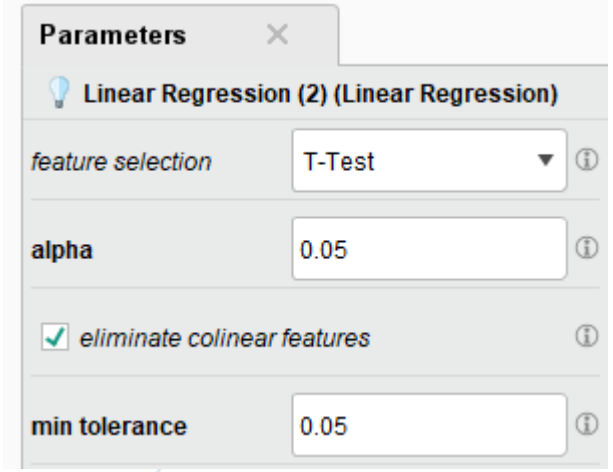
# Análise em Rapid Miner

## Regressão Linear Simples (Sem Outliers) com cross validation e feature selection

Optamos por escolher o T-Test para um alfa de 0.05.

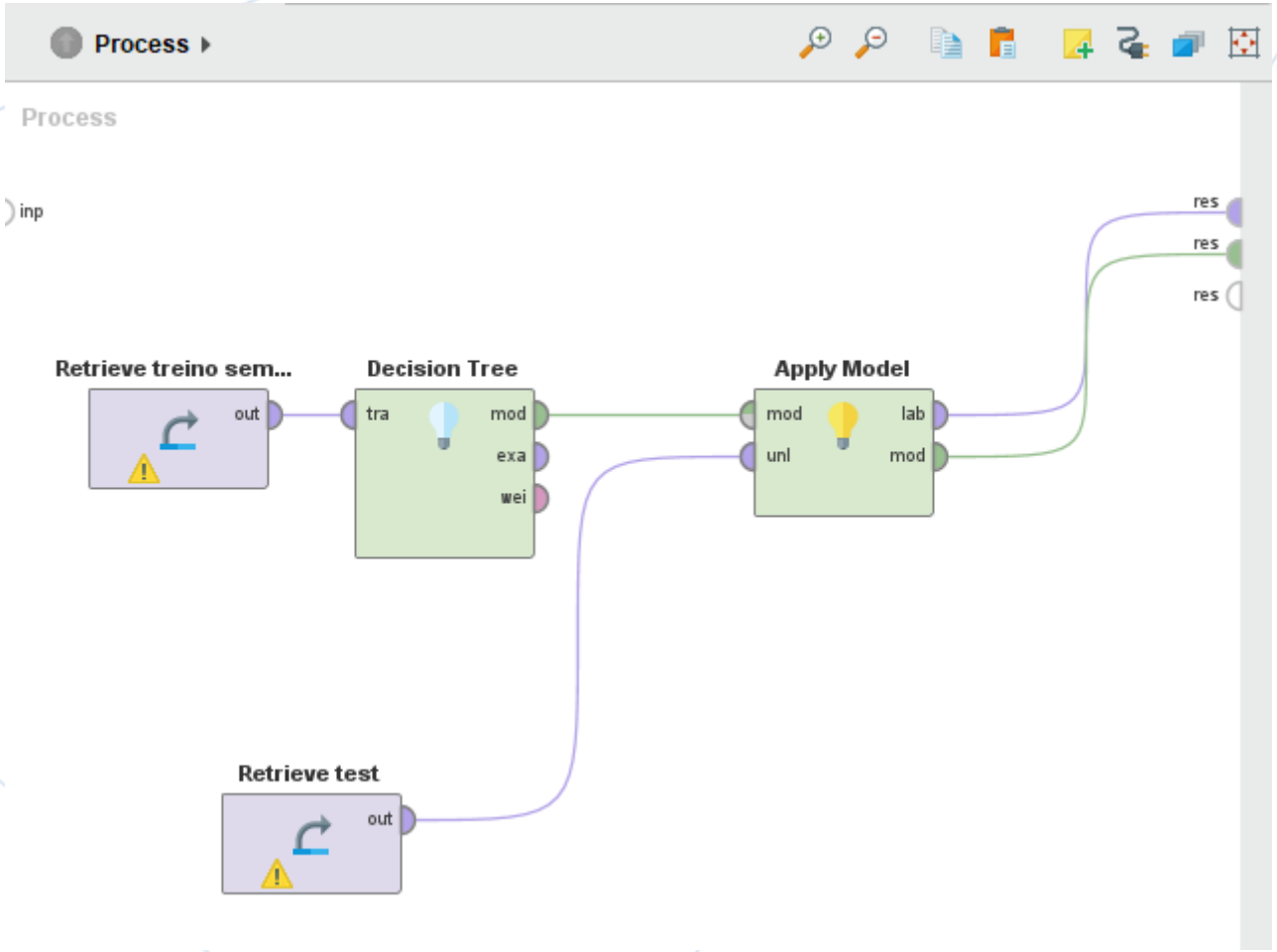


# Análise em Rapid Miner

## Árvore de decisão



# Análise em Rapid Miner

## Resumo

| | Que variáveis ficam? | Influência |
|---|---|---|
| Regressão Simples Com Feature Selection (T-test, alfa=0.05) | apps_dia | - |
| | vist_dia | - |
| | hands_dia | - |
| Regressão Simples Sem Outliers Com Feature Selection (T-Test, alfa =0.05) | vist_dia | - |
| | apps_dia | - |
| | contractor | + |
| | relocation_paid | + |

# Análise em Rapid Miner

## Resumo

| | Erro absoluto |
|---|---|
| Regressão Linear Simples Sem Feature Selection | 30.60 |
| Regressão Linear Sem Outliers Sem Feature Selection | 18.45 |
| Regressão Linear Simples Com Feature Selection (T-test, alfa=0.05) | 30.28 |
| Regressão Linear Sem Outliers Com Feature Selection (T-Test, alfa =0.05) | 18.29 |

Erro absoluto do modelo trivial: **37.20 dias**

# 3. Modelling | Regressão Linear

# Análise em R

| Etapas no R |
| --- |
| Análise Estatística |
| Divisão dos dados em Treino e Teste |
| Remoção Outliers |
| Regressão Linear Simples |
| Análise Fatorial |
| Feature Selection com Teste ANOVA |
| Feature Selection - Técnica Step Wise Forward and Backward Selection |
| Feature Selection - Técnica Boruta |
| Análise global dos modelos |
| Árvore de decisão |

# Análise em R

## Carregar as Library

```
library(tidyverse)
library(caret)
library(ggplot2)
library(lattice)
library(DAAG)
library(DMwR2)
library(Boruta)
library(Metrics)
library(TH.data)
library(caret)
library(Boruta)
library(rpart)
library(AICcmodavg)
```

## Importação e Renomear os dados

```
setwd("C:/Users/jofis/Associação Porto Business School/PGBIA13P1G04 - General/04_Modeling/Regressão Linear")
#importação dos dados
`20220528.1530NAjobsApps_formodels` <- read.csv2("C:/Users/jofis/Associação Porto Business School/PGBIA13P1G04 - General/03_Data
Preparation/20220528-1530NAjobsApps_formodels.csv")

#simplificação dos nomes
L<-`20220528.1530NAjobsApps_formodels`
```

# Análise em R

```
#análise geral dos dados
summary(L)
```

```
##        X             id        job_category_Back.end.Developer
##  Min.   :  1   Min.   :11477   Min.   :0.0000
##  1st Qu.:196   1st Qu.:12444   1st Qu.:0.0000
##  Median :391   Median :13175   Median :0.0000
##  Mean   :391   Mean   :13215   Mean   :0.1869
##  3rd Qu.:586   3rd Qu.:13984   3rd Qu.:0.0000
##  Max.   :781   Max.   :15007   Max.   :1.0000
##  job_category_Data.Scientist job_category_DevOps...Sysadmin
##  Min.   :0.00000             Min.   :0.00000
##  1st Qu.:0.00000             1st Qu.:0.00000
##  Median :0.00000             Median :0.00000
##  Mean   :0.04866             Mean   :0.08195
##  3rd Qu.:0.00000             3rd Qu.:0.00000
##  Max.   :1.00000             Max.   :1.00000
##  job_category_Front.end.Developer job_category_Full.stack.Developer
##  Min.   :0.0000                   Min.   :0.0000
##  1st Qu.:0.0000                   1st Qu.:0.0000
##  Median :0.0000                   Median :0.0000
##  Mean   :0.1549                   Mean   :0.1933
##  3rd Qu.:0.0000                   3rd Qu.:0.0000
##  Max.   :1.0000                   Max.   :1.0000
##  job_category_outros job_category_Product.Project.Management
##  Min.   :0.0000      Min.   :0.00000
##  1st Qu.:0.0000      1st Qu.:0.00000
##  Median :0.0000      Median :0.00000
##  Mean   :0.1268      Mean   :0.09091
##  3rd Qu.:0.0000      3rd Qu.:0.00000
##  Max.   :1.0000      Max.   :1.00000
##  job_category_QA...Testing job_category_UX...UI.Designer   contractor
##  Min.   :0.0000            Min.   :0.00000               Min.   :0.00000
##  1st Qu.:0.0000            1st Qu.:0.00000               1st Qu.:0.00000
##  Median :0.0000            Median :0.00000               Median :0.00000
##  Mean   :0.0653            Mean   :0.05122               Mean   :0.05122
##  3rd Qu.:0.0000            3rd Qu.:0.00000               3rd Qu.:0.00000
##  Max.   :1.0000            Max.   :1.00000               Max.   :1.00000
##   consultancy   permanent      relocation_paid   visa_support
##  Min.   :0    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0    1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0    Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0    Mean   :0.9731   Mean   :0.1613   Mean   :0.1613
##  3rd Qu.:0    3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :0    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   min_engaged    show_sal_rate     home_work      education_req
##  Min.   :  0.0   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  8.0   1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median : 21.0   Median :0.000   Median :1.0000   Median :1.0000
##  Mean   : 37.2   Mean   :0.356   Mean   :0.8105   Mean   :0.5749
##  3rd Qu.: 47.0   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :341.0   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##    exp_nivel      languagePT       regionPT       full_time
##  Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.0000
##  Median :2.000   Median :0.0000   Median :1.0000   Median :1.0000
##  Mean   :2.356   Mean   :0.1165   Mean   :0.7977   Mean   :0.9731
##  3rd Qu.:3.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     vist_dia          apps_dia          hands_dia
##  Min.   :  2.473   Min.   :0.005263   Min.   : 0.00000
##  1st Qu.: 10.575   1st Qu.:0.066116   1st Qu.: 0.02308
##  Median : 19.400   Median :0.142180   Median : 0.17949
##  Mean   : 32.993   Mean   :0.241688   Mean   : 0.41180
##  3rd Qu.: 43.244   3rd Qu.:0.282609   3rd Qu.: 0.46342
##  Max.   :191.000   Max.   :3.609756   Max.   :10.84000
```

# Análise em R

```
#nome das colunas
colnames(L)
```

```
##  [1] "X"
##  [2] "id"
##  [3] "job_category_Back.end.Developer"
##  [4] "job_category_Data.Scientist"
##  [5] "job_category_DevOps...Sysadmin"
##  [6] "job_category_Front.end.Developer"
##  [7] "job_category_Full.stack.Developer"
##  [8] "job_category_outros"
##  [9] "job_category_Product.Project.Management"
## [10] "job_category_QA...Testing"
## [11] "job_category_UX...UI.Designer"
## [12] "contractor"
## [13] "consultancy"
## [14] "permanent"
## [15] "relocation_paid"
## [16] "visa_support"
## [17] "min_engaged"
## [18] "show_sal_rate"
## [19] "home_work"
## [20] "education_req"
## [21] "exp_nivel"
## [22] "languagePT"
## [23] "regionPT"
## [24] "full_time"
## [25] "vist_dia"
## [26] "apps_dia"
## [27] "hands_dia"
```

# Análise em R

```
#número de colunas
ncol(L)
```

```
## [1] 27
```

```
#uma vez que "consultancy" e  só assumem valor Zero, serão para
retirar da análise (irei também retirar o ID inicial e o ID)
L1<-L[,-c(1,2,13)]
colnames(L1)
```

```
##  [1] "job_category_Back.end.Developer"
##  [2] "job_category_Data.Scientist"
##  [3] "job_category_DevOps...Sysadmin"
##  [4] "job_category_Front.end.Developer"
##  [5] "job_category_Full.stack.Developer"
##  [6] "job_category_outros"
##  [7] "job_category_Product.Project.Management"
##  [8] "job_category_QA...Testing"
##  [9] "job_category_UX...UI.Designer"
## [10] "contractor"
## [11] "permanent"
## [12] "relocation_paid"
## [13] "visa_support"
## [14] "min_engaged"
## [15] "show_sal_rate"
## [16] "home_work"
## [17] "education_req"
## [18] "exp_nivel"
## [19] "languagePT"
## [20] "regionPT"
## [21] "full_time"
## [22] "vist_dia"
## [23] "apps_dia"
## [24] "hands_dia"
```

# Análise em R

```
#vamos confirmar que estão as colunas
head(L1)
```

```
##   job_category_Back.end.Developer job_category_Data.Scientist
## 1                               0                           0
## 2                               1                           0
## 3                               0                           0
## 4                               0                           0
## 5                               0                           0
## 6                               0                           0
##   job_category_DevOps...Sysadmin job_category_Front.end.Developer
## 1                              0                                1
## 2                              0                                0
## 3                              0                                0
## 4                              0                                1
## 5                              0                                0
## 6                              0                                0
##   job_category_Full.stack.Developer job_category_outros
## 1                                 0                   0
## 2                                 0                   0
## 3                                 0                   0
## 4                                 0                   0
## 5                                 0                   0
## 6                                 1                   0
##   job_category_Product.Project.Management job_category_QA...Testing
## 1                                       0                        0
## 2                                       0                        0
## 3                                       1                        0
## 4                                       0                        0
## 5                                       0                        0
## 6                                       0                        0
##   job_category_UX...UI.Designer contractor permanent relocation_paid
## 1                             0          0         1               1
## 2                             0          0         1               0
## 3                             0          0         1               0
## 4                             0          0         1               0
## 5                             1          0         1               1
## 6                             0          0         1               1
##   visa_support min_engaged show_sal_rate home_work education_req exp_nivel
## 1            0          12             1         1             1         3
## 2            0           6             1         1             0         2
## 3            0          15             0         1             1         2
## 4            0          18             0         1             0         1
## 5            1           4             0         1             0         3
## 6            0         219             0         1             0         3
##   languagePT regionPT full_time  vist_dia   apps_dia   hands_dia
## 1          0        1         1 27.741935 0.06048387 0.000000000
## 2          1        1         1 47.342541 0.23756906 0.077348066
## 3          0        1         1 13.432558 0.20000000 0.539534884
## 4          0        1         1  7.380645 0.14838710 0.006451613
## 5          0        0         1 77.888889 0.92592593 1.518518519
## 6          0        1         1 11.819372 0.03664921 0.068062827
```

# Análise em R

## Análise Estatistica

```
#análise global
summary(L1)
```

```
##  job_category_Back.end.Developer job_category_Data.Scientist
##  Min.   :0.0000                  Min.   :0.00000
##  1st Qu.:0.0000                  1st Qu.:0.00000
##  Median :0.0000                  Median :0.00000
##  Mean   :0.1869                  Mean   :0.04866
##  3rd Qu.:0.0000                  3rd Qu.:0.00000
##  Max.   :1.0000                  Max.   :1.00000
##  job_category_DevOps...Sysadmin job_category_Front.end.Developer
##  Min.   :0.00000                Min.   :0.0000
##  1st Qu.:0.00000                1st Qu.:0.0000
##  Median :0.00000                Median :0.0000
##  Mean   :0.08195                Mean   :0.1549
##  3rd Qu.:0.00000                3rd Qu.:0.0000
##  Max.   :1.00000                Max.   :1.0000
##  job_category_Full.stack.Developer job_category_outros
##  Min.   :0.0000                    Min.   :0.0000
##  1st Qu.:0.0000                    1st Qu.:0.0000
##  Median :0.0000                    Median :0.0000
##  Mean   :0.1933                    Mean   :0.1268
##  3rd Qu.:0.0000                    3rd Qu.:0.0000
##  Max.   :1.0000                    Max.   :1.0000
##  job_category_Product.Project.Management job_category_QA...Testing
##  Min.   :0.00000                         Min.   :0.0000
##  1st Qu.:0.00000                         1st Qu.:0.0000
##  Median :0.00000                         Median :0.0000
##  Mean   :0.09091                         Mean   :0.0653
##  3rd Qu.:0.00000                         3rd Qu.:0.0000
##  Max.   :1.00000                         Max.   :1.0000
##  job_category_UX...UI.Designer   contractor        permanent
##  Min.   :0.00000               Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000               1st Qu.:0.00000   1st Qu.:1.0000
##  Median :0.00000               Median :0.00000   Median :1.0000
##  Mean   :0.05122               Mean   :0.05122   Mean   :0.9731
##  3rd Qu.:0.00000               3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000               Max.   :1.00000   Max.   :1.0000
##  relocation_paid   visa_support     min_engaged     show_sal_rate
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.0   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  8.0   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median : 21.0   Median :0.000
##  Mean   :0.1613   Mean   :0.1613   Mean   : 37.2   Mean   :0.356
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 47.0   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :341.0   Max.   :1.000
##    home_work      education_req      exp_nivel       languagePT
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :2.000   Median :0.0000
##  Mean   :0.8105   Mean   :0.5749   Mean   :2.356   Mean   :0.1165
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :3.000   Max.   :1.0000
##     regionPT        full_time         vist_dia          apps_dia
##  Min.   :0.0000   Min.   :0.0000   Min.   :  2.473   Min.   :0.005263
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.: 10.575   1st Qu.:0.066116
##  Median :1.0000   Median :1.0000   Median : 19.400   Median :0.142180
##  Mean   :0.7977   Mean   :0.9731   Mean   : 32.993   Mean   :0.241688
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 43.244   3rd Qu.:0.282609
##  Max.   :1.0000   Max.   :1.0000   Max.   :191.000   Max.   :3.609756
##    hands_dia
##  Min.   : 0.00000
##  1st Qu.: 0.02308
##  Median : 0.17949
##  Mean   : 0.41180
##  3rd Qu.: 0.46342
##  Max.   :10.84000
```

# Análise em R

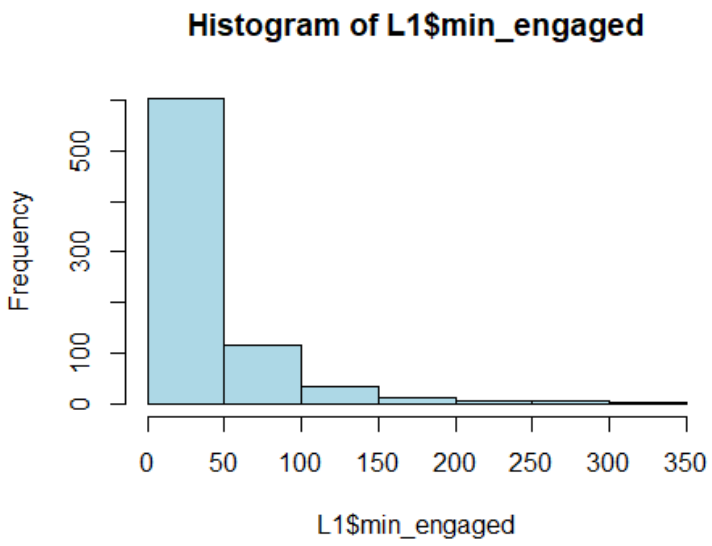
```
#ver a correlação das variáveis
dim(cor(L1))
```

```
## [1] 24 24
```

```
#erro modelo trivial (unidade: dias)
mean(L1$min_engaged)
```

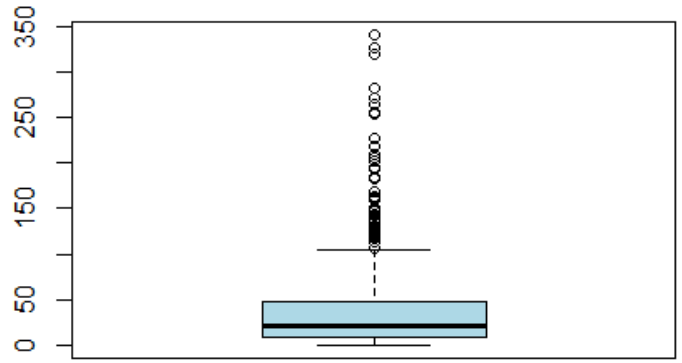
```
## [1] 37.20102
```

```
#vamos ver a distribuição da nossa variável dependente:
hist(L1$min_engaged, col = "lightblue" )
```



# Análise em R

```
boxplot(L1$min_engaged, col="lightblue")
```




```
#se precisar de exportar a tab de correlações (atenção diretoria)
#corre<-cor(L1)
#write.csv(corre,"./Tabela de Correlações.csv")
```

**Comentário:**
A nossa variável dependente tem uma distribuição com cauda alongada à direita ( assimétrica positiva).
Erro do modelo trivial: 37.20 dias

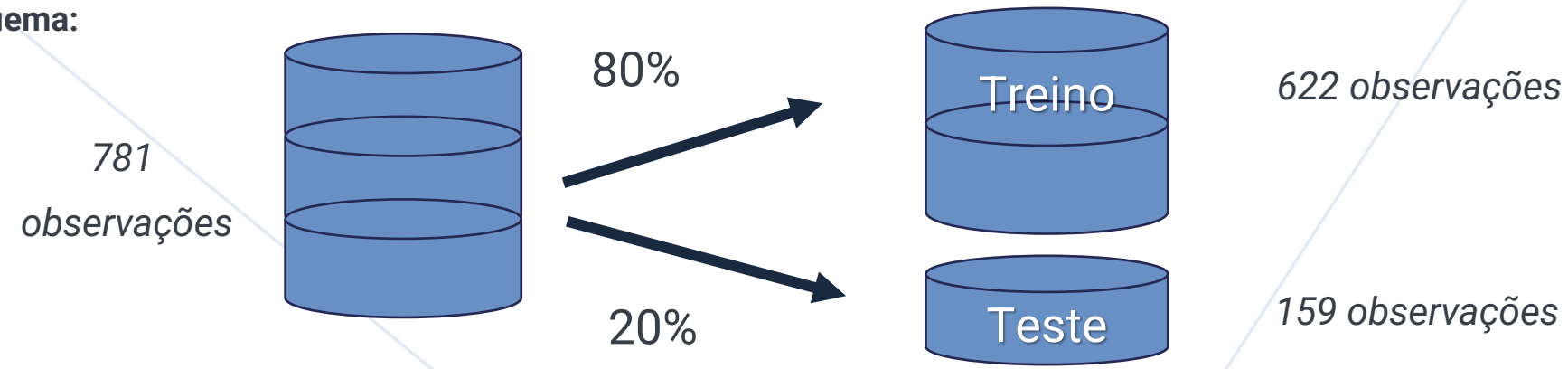
# Análise em R

## Divisão dos dados em Treino e Teste

```
#para tornar a amostra reproduzivel (fixar samples)
set.seed(123)

#vamos criar a amostra
sample<-sample(c(TRUE, FALSE), nrow(L1), replace=TRUE, prob=c(0.8,0.2))
train<-L1[sample, ]
test<-L1[!sample, ]
nrow(test)
```

```
## [1] 159
```

**Esquema:**




**Comentário:**
Uma vez que temos um dataset pequeno, optamos por uma divisão dos dados em 80% em treino e 20% em teste.

# Análise em R

## Retirar Outliers

```
#Quartis
quantile(train$min_engaged)
```

```
##   0%  25%  50%  75% 100%
##    0    8   21   48  341
```

```
#amplitude interquartil
IQR(train$min_engaged)
```

```
## [1] 40
```

```
Q<-quantile(train$min_engaged,probs=c(.25,.75),na.rm=TRUE)
up<-Q[2]+1.5*IQR(train$min_engaged) # Upper Range

#quantos ouliers vamos remover?
o<-nrow(train[train$min_engaged>up, ])
```

```
#qual a proporção no nosso dataset que vamos remover?
o/nrow(L1)*100
```

```
## [1] 5.633803
```

```
#vamos criar o dataset sem os outliers
train<-subset(train,train$min_engaged<up)
nrow(train)
```

```
## [1] 578
```

```
#write.csv(train,"./treino sem outliers.csv")
#write.csv(test,"./test.csv")
```

**Comentário:**
Apenas removemos outliers no dataset de treino.

# Análise em R

## Regressão Linear Simples

```
#Regressão Linear do TREINO:
lm(min_engaged~.,train)
```

```
##
## Call:
## lm(formula = min_engaged ~ ., data = train)
##
## Coefficients:
##                                (Intercept)
##                                    52.3376
##         job_category_Back.end.Developer
##                                    -1.4599
##             job_category_Data.Scientist
##                                    -4.3710
##          job_category_DevOps...Sysadmin
##                                    -8.5893
##        job_category_Front.end.Developer
##                                     1.0518
##       job_category_Full.stack.Developer
##                                    -6.6206
##                      job_category_outros
##                                    -4.2555
## job_category_Product.Project.Management
##                                    -2.7008
##                job_category_QA...Testing
##                                    -7.5983
##           job_category_UX...UI.Designer
##                                        NA
##                                contractor
##                                    9.4625
##                                 permanent
##                                    -3.2836
##                           relocation_paid
##                                    5.9275
##                              visa_support
##                                    -3.7087
##                             show_sal_rate
##                                    -2.5959
##                                 home_work
##                                    -2.0479
##                             education_req
##                                    -0.6272
##                                 exp_nivel
##                                    -2.5490
##                                languagePT
##                                    -2.0532
##                                  regionPT
##                                     0.3759
##                                 full_time
##                                        NA
##                                 vist_dia
##                                    -0.1702
##                                 apps_dia
##                                    -13.5177
##                                hands_dia
##                                    -2.0448
```

# Análise em R

```
summary(lm(train$min_engaged~.,train))

## Call:
## lm(formula = train$min_engaged ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.112 -15.712  -6.869  11.098  80.923
##
## Coefficients: (2 not defined because of singularities)
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              52.33762   11.43895   4.575 5.87e-06
## job_category_Back.end.Developer          -1.45987    5.21936  -0.280 0.779809
## job_category_Data.Scientist              -4.37104    6.58316  -0.664 0.506983
## job_category_DevOps...Sysadmin           -8.58934    5.78832  -1.484 0.138399
## job_category_Front.end.Developer          1.05182    5.31417   0.198 0.843174
## job_category_Full.stack.Developer        -6.62058    5.19979  -1.273 0.203465
## job_category_outros                      -4.25549    5.45455  -0.780 0.435622
## job_category_Product.Project.Management  -2.70081    5.71385  -0.473 0.636628
## job_category_QA...Testing                -7.59826    6.29737  -1.207 0.228108
## job_category_UX...UI.Designer                  NA         NA      NA       NA
## contractor                                9.46248    6.39066   1.481 0.139260
## permanent                                -3.28365    8.69546  -0.378 0.705851
## relocation_paid                           5.92755    3.21763   1.842 0.065977
## visa_support                             -3.70870    3.31982  -1.117 0.264418
## show_sal_rate                            -2.59586    2.18599  -1.187 0.235538
## home_work                                -2.04792    2.81766  -0.727 0.467645
## education_req                            -0.62716    2.14006  -0.293 0.769590
## exp_nivel                                -2.54896    1.71424  -1.487 0.137600
## languagePT                               -2.05322    3.31394  -0.620 0.535795
## regionPT                                  0.37594    2.61005   0.144 0.885524
## full_time                                      NA         NA      NA       NA
## vist_dia                                 -0.17021    0.03489  -4.879 1.39e-06
## apps_dia                                -13.51771    3.60664  -3.748 0.000197
## hands_dia                                -2.04484    1.19846  -1.706 0.088524
##
## (Intercept)                              ***
## job_category_Back.end.Developer
## job_category_Data.Scientist
## job_category_DevOps...Sysadmin
## job_category_Front.end.Developer
## job_category_Full.stack.Developer
## job_category_outros
## job_category_Product.Project.Management
## job_category_QA...Testing
## job_category_UX...UI.Designer
## contractor
## permanent
## relocation_paid                          .
## visa_support
## show_sal_rate
## home_work
## education_req
## exp_nivel
## languagePT
## regionPT
## full_time
## vist_dia                                 ***
## apps_dia                                 ***
## hands_dia                                .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.03 on 556 degrees of freedom
## Multiple R-squared:  0.1462, Adjusted R-squared:  0.1139
## F-statistic: 4.533 on 21 and 556 DF,  p-value: 2.092e-10
```

# Análise em R

```
model<-lm(min_engaged~.,data=train)

#categorias UX_UI Designer e Full time estão NA

#Estisticas importantes da regressão
coefficients(model) # model coefficients
```

```
##                              (Intercept)         job_category_Back.end.Developer
##                                52.3376205                              -1.4598732
##             job_category_Data.Scientist          job_category_DevOps...Sysadmin
##                                -4.3710386                              -8.5893431
##       job_category_Front.end.Developer       job_category_Full.stack.Developer
##                                 1.0518226                              -6.6205849
##                    job_category_outros job_category_Product.Project.Management
##                                -4.2554882                              -2.7008110
##               job_category_QA...Testing           job_category_UX...UI.Designer
##                                -7.5982620                                      NA
##                                contractor                               permanent
##                                 9.4624805                              -3.2836451
##                           relocation_paid                            visa_support
##                                 5.9275454                              -3.7087022
##                             show_sal_rate                               home_work
##                                -2.5958623                              -2.0479161
##                            education_req                               exp_nivel
##                                -0.6271553                              -2.5489639
##                                languagePT                                regionPT
##                                -2.0532190                               0.3759411
##                                 full_time                                vist_dia
##                                        NA                              -0.1702113
##                                  apps_dia                               hands_dia
##                               -13.5177129                              -2.0448370
```

# Análise em R

Porto Business School / University of Porto

```
fitted(model) # predicted values, os valores que prevê
```

```
##          1          2          3          7          9         10
##  37.95186363 24.74732635 32.86292919 33.28192708 27.08473097 32.36465422
##         12         13         14         15         17         18
##  26.31614838 31.14599630 34.98863683 -6.81468134 24.97767623 28.86536290
##         19         22         23         26         27         28
##  31.22277905 28.84013032 28.52957810 33.89177744 41.58300709 42.24787707
##         29         30         33         34         35         36
##  28.40632608 26.46442813 28.73735639 19.48046872 26.32875350 16.41115714
##         37         38         39         40         41         42
##  22.68973106 15.50787595 20.47449560 29.52298150 32.36591424 30.53254768
##         43         44         45         46         47         48
##  32.59369460 23.60342458  5.63760519 20.21571237 24.44443088 17.06035250
##         49         52         53         54         55         56
##  37.91974725 28.17809985 21.30979341 20.03417028 32.48493203 42.40166166
##         57         58         60         61         62         63
##  31.70962591 29.48252484  4.75259315 33.40523339  6.48280111 39.51313413
##         64         66         69         70         71         72
##  25.31323598 24.70491840 22.44091103 34.17147925 38.24421419 39.05162409
##         73         74         75         76         77         78
##  28.03579605 29.41207185 -17.32293821 28.68480075 17.57092318 -6.87482238
##         79         80         81         82         83         84
##  27.52472194 10.87505281 36.73300779 30.81064620 18.36998357 23.85303874
##         85         86         91         92         93         94
##  34.02811227  5.20484211  3.51331420 14.86188412 18.71083763 25.10028212
##         95         96         97         98         99        100
##  23.54617984 32.03806368 26.56918656 33.74563229 31.23463599 16.72131199
##        101        102        103        105        108        109
##  26.47782469 22.56034812 35.80691501 30.84800736 34.58559587 36.42915320
##        110        112        113        115        116        117
##  28.90311273 28.60140384 12.23388453 31.29639835 33.94334826 27.03638383
##        119        120        121        122        123        124
##  32.74383272 24.16285506 35.71376554 34.70539426 26.27737053 14.89986124
##        125        127        128        129        130        133
##  33.48007107 26.99419871 28.92216753 33.05736567 34.37420932 13.80604147
##        134        136        138        140        141        142
##  34.76172106 29.09902480 14.71676225 -28.73031448 28.11254035 40.62354168
##        146        147        148        149        150        152
##  26.70353670 27.49146658 33.96078359 35.67546778 31.95023555 32.81717624
##        153        154        156        157        158        159
##  24.05347909 34.62518987 33.93178522 34.50453369 11.12552091 24.66395411
##        160        161        162        164        165        166
##  34.65698495 19.33202792 31.78974808 32.60786870 29.33729963  7.29939821
##        167        168        169        170        171        172
##   9.40352773 17.82582287 32.94886868 37.42384618 36.44655607 38.19702468
##        174        175        176        177        178        182
##  30.79659747 28.51324417 28.69012657 27.36200380 38.54537359 33.52754982
##        184        185        186        187        191        194
##  38.01174105 34.77486660 36.87567278  8.17450829 20.13587224 35.91529742
##        196        197        198        199        201        203
##  35.20276877 29.77576329 40.45578013 22.57542790 32.01442024 16.79344227
##        204        205        207        208        209        211
##  25.48447475 43.19683682 35.73860359 35.21716744 30.71569352 30.47514094
##        212        213        214        216        217        218
##  31.69828433 32.78382119 23.77485389 33.22335341 34.09449709 24.15133006
##        221        223        224        226        227        228
##  30.76484947 26.03920755 24.18049261 38.66817573 37.37082745 30.90348565
##        229        231        232        233        234        235
##  37.91304530 20.89636750 26.54119239 29.88243514 30.82855927 31.87013144
##        236        237        239        241        242        243
##  30.03280826 12.86446262 17.08099313 30.23417520 17.86053575 22.37134037
##        244        245        246        247        250        252
##  16.83597535 31.25435063 28.47328672 24.90158512 25.69305128 30.00809543
##        254        255        256        257        258        259
##  31.34985424 24.88535852 26.91850469 30.03600570 26.45787533 29.86562918
##        263        265        266        267        269        270
##  35.94457186 27.07410440 31.39622502 21.85063791 27.35579651 29.91126070
##        272        273        274        275        276        278
##  39.19685490 38.68017937 30.92985819 36.52212133 35.97847208 35.49961719
##        279        281        282        285        286        287
##  30.05176322 37.86975921 22.66214488 27.02551393 28.91030262 40.03782529
##        288        289        290        291        293        295
##  23.37787047 30.41888215 29.75581443 15.94898014 12.61434995 31.22152782
##        298        299        301        302        303        304
##  26.30308189 19.35559652 29.58882435 26.16890024 14.54963803 34.35269707
##        305        306        307        308        309        310
##  17.04812952 13.61403082 29.28836815 30.58043982 28.47068420 28.83313389
##        311        312        313        314        315        318
##  10.22415811 24.17228358 13.05885542  6.48766364 13.32486645 -6.41608419
##        319        322        323        324        326        328
##   7.93667906 32.76958859 33.71127348 34.70877474 28.99849773 25.31197089
##        329        331        332        333        335        336
##  35.12004324 28.20906831 30.16628144 29.35428698 25.13179820 40.53656219
##        337        338        339        341        342        343
##  26.66414498 30.90430333 26.01882327 29.69828004 21.55614579 29.68538547
##        344        345        346        348        349        351
##  24.91834144 20.61869910 22.54471953 32.12559737 29.95342760 34.04371989
##        353        354        355        357        358        359
##  34.95689898 38.55360115 31.28436158 29.35627999 24.01418334 40.95392993
##        361        362        364        365        366        367
##  34.13440403 35.60189921 17.39507313 33.76192295 39.59311462 19.30615018
##        368        369        370        371        372        374
##  16.24073594 40.90959184 42.13805437 17.31722604 30.51547038 44.87700686
##        375        377        378        379        381        382
##  39.21496607 35.20718492 36.23180133 33.85346763 34.74559734 24.36332078
##        383        384        385        387        389        392
##  26.09773858  0.69779525 34.25061561 35.60306441 23.61776948 30.27689617
##        393        394        397        399        402        404
##  30.70480880 26.65790177 30.32185288 38.30388077 46.54430740 19.23208508
##        405        406        407        408        409        410
##  27.29452918 24.22308441 32.17164104 36.02483829 27.39434825 20.76703897
##        411        414        415        416        418        419
##  31.47728804 36.28059672 39.93610387 27.30860721 41.87937387 25.35198306
##        420        421        422        423        424        426
##  29.08816885 32.56981072 32.96437048 30.68404174 19.24810250 29.92689709
##        427        428        429        432        433        435
##  26.37860228 35.97474166 26.57773602 30.71647590 29.57962123 33.64662872
##        436        437        438        439        440        441
##  39.97115925 29.18963875 31.08482869 13.53441335 31.52129290 34.60434096
##        442        443        444        446        447        448
##  29.98632776 33.38360240 23.40730880 35.82325052 34.84453437 22.62461907
##        449        450        451        453        454        455
##  30.74866563 46.43525159  2.09116217 33.87213639 30.31757681 32.48459019
##        458        459        460        462        463        464
##  29.43176257 21.96425502 19.37414143 29.50129123 22.08252454 31.75163032
##        465        466        467        468        469        471
##  27.80805985  5.60087654 27.70046923 27.32188157 11.10286100 35.40717189
##        473        474        475        476        477        478
##  27.58627999 29.76261574 24.89795598 47.13941140 34.90478432 16.63206254
##        479        480        481        483        484        486
## -10.07053051 -0.49825983 37.50022878 35.64957795 31.10048638 33.03122929
##        487        488        489        492        493        494
##  -3.85400121 25.63201222 23.33968752 29.07526845 23.85431124 22.10340317
##        495        497        498        499        500        501
##  33.66255048 29.76602214 30.48940295 -0.05427364  2.39495088 30.28236330
##        502        503        504        505        506        507
##  25.11502866 31.61413797 30.56644324  7.39463425 24.97514355 32.77075320
##        508        510        511        512        514        515
##  36.13517816  5.32034266 26.68473261 10.76892602 10.95278026 24.00851408
##        516        517        518        519        521        522
##  11.42472051 20.81372868 17.73263965 42.25006536 27.68749302 40.83333150
##        523        524        525        526        528        530
##  18.81300386 37.87454475 35.89618780 16.06709521 33.84730426 22.79255872
##        532        533        535        536        537        539
##  25.44738593 20.99905425 15.02413491  3.11960503  8.00823441 38.11401032
##        540        542        544        545        547        550
##  40.13943620 22.10659268 27.31783059 19.66268643 29.06688842 28.57466450
##        552        553        555        556        557        558
##  36.27941649 28.21766700 22.34886757 20.34522829 33.25621659 31.41952619
##        559        560        561        563        564        565
##  27.43715220 37.38481899 28.05394243 -0.84692039  0.22383588 33.61966983
##        566        567        571        573        576        577
##  27.37209490 28.92269887 17.80238432 16.69305564 25.42500442 29.29911024
##        578        579        580        584        586        587
##  21.98077169 41.44998043 23.93371446 29.62797761 30.80389703 28.16655823
##        588        590        591        592        594        595
##  18.11871250 27.88349518 24.53781890 27.27602071 20.78156865 27.43981923
##        597        598        600        601        602        603
##  27.90288514 29.64649538 27.75109692 29.72924188 29.95949247 13.60991823
##        604        605        607        609        610        611
##  29.94469385 39.95949643 39.32294758  6.34214637 36.10263314 30.80962563
##        612        613        615        616        618        620
##  29.08855628 41.11156205 23.07712717 32.86295568 24.76594543 24.62021491
##        623        624        625        626        627        629
##  27.00165071 33.76147063 25.62427931 29.89073272 33.96129368 35.22934038
##        630        631        633        635        636        638
##  29.85933185 32.44010822 31.09296818 18.68784720 28.21074584 18.67856565
##        640        641        642        644        645        646
##  34.02254626 42.54239703 41.43252784 45.89883980 35.83371530 19.30360506
##        647        648        649        650        651        653
##  36.99882889 16.63508844 36.77806232 21.56033296 39.48822416 30.92429350
##        654        655        658        659        660        663
##  30.94317339 33.03908524 27.01393941 29.20067914 14.55502737 26.98238993
##        664        665        666        668        669        670
##  39.36798019 22.87713603 33.43352250 35.84498914 27.30623149 27.75012544
##        671        672        673        674        675        676
##  26.06540784 32.85482517 21.32923173 27.52001865 30.09067514 41.68757514
##        678        679        680        682        686        687
##  18.81603261 35.66367799 40.62623802 38.14514931 34.43468145 36.62581339
##        688        689        690        691        693        694
##  18.05201443 21.52187185 17.33520865 33.45317907 15.72284151 14.12094935
##        695        696        697        698        699        700
##  19.97652989 38.22025414 19.58380130 25.73167005 28.64551351 28.47324927
##        702        703        705        706        707        708
##  16.14312883 14.17640692 17.53692542 25.76359829 25.41049702 28.46497526
##        709        710        711        712        713        714
##  21.09888625 38.65847741 26.95833277 28.19288113 38.63118633 34.83494965
##        715        716        717        718        720        721
##  38.33008274 36.37723251 39.62835417 36.97666451 25.28087804 20.45108371
##        722        723        725        726        728        729
##  28.98083353  9.11971258 41.91888092 27.17371559 29.65457964 13.20004029
##        730        731        733        735        736        739
##  30.09822236 53.07553597 34.54148334 28.54668880 20.63778602 40.96990370
##        740        742        743        744        746        747
##  37.88226956 33.22697926 35.76089652 28.68641991 32.55510709 40.44234273
##        748        749        751        752        753        754
##  26.77038846 29.65149240 29.03358014 14.33649027 26.63480400 41.07955218
##        755        756        757        759        760        762
##  27.35365047  8.82989027 33.33160547 21.33974489 37.49876439 34.55973352
##        763        764        765        766        767        768
##  28.14896754 15.13590929 29.68235501 33.12667453 29.10441531 26.61604093
##        769        770        773        775        777        778
##  29.58357880 23.15502520 23.83487423 35.14222771  7.27658656 24.32532678
##        779        781
##  30.82428591 40.06049597
```

# Análise em R

```
residuals(model) # residuals
```

```
##          1          2          3          7          9         10
## -25.9518636 -18.7473263 -17.8629292 -24.2819271   4.9152690 -24.3646542
##         12         13         14         15         17         18
## -24.3161484 -20.1459963  36.0113632  11.8146813 -22.9776762  51.1346371
##         19         22         23         26         27         28
##  11.7772209  16.1598697  10.4704219  22.1082226 -20.5830071  14.7521229
##         29         30         33         34         35         36
## -21.4063261 -12.4644281 -13.7373564   2.5195313 -18.3287535  -9.4111571
##         37         38         39         40         41         42
## -20.6897311 -10.5078759 -19.4744956 -28.5229815  -4.3659142  -2.5325477
##         43         44         45         46         47         48
##  36.4063054 -17.6034246  -1.6376052  -8.2157124 -22.4444309 -13.0603525
##         49         52         53         54         55         56
## -25.9197472 -14.1780998  37.6902066  -4.0341703 -26.4849320  14.5983383
##         57         58         60         61         62         63
## -16.7096259  45.5174752  -2.7525932  58.5947666   3.5171989  42.4868659
##         64         66         69         70         71         72
## -18.3132360 -21.7049184  -7.4409110 -25.1714792 -26.2442142 -25.0516241
##         73         74         75         76         77         78
##  21.9642039 -17.4120718  22.3229382 -22.6848007  -4.5709232  17.8748224
##         79         80         81         82         83         84
## -13.5247219  -6.8750528  10.2669922 -11.8106462 -12.3699836 -17.8530387
##         85         86         91         92         93         94
##  -6.0281123  -4.2048421  -0.5133142  -6.8618841  79.2891624  -4.1002821
##         95         96         97         98         99        100
##   3.4538202 -13.0380637 -24.5691866  29.2543677  -4.2346360  -6.7213120
##        101        102        103        105        108        109
##  15.5221753 -10.5603481   5.1930850   5.1519926   8.4144041  15.5708468
##        110        112        113        115        116        117
##  23.0968873  25.3985962  -5.2338845  16.7036017  42.0566517  70.9636162
##        119        120        121        122        123        124
## -16.7438327   8.8371449 -30.7137655 -16.7053943 -25.2773705  -9.8998612
##        125        127        128        129        130        133
##  18.5199289  12.0058013  -9.9221675  10.9426343  41.6257907  -4.8060415
##        134        136        138        140        141        142
##   8.2382789 -15.0990248  -7.7167622  29.7303145  53.8874597  40.3764583
##        146        147        148        149        150        152
##   9.2964633 -14.4914666 -33.9607836  62.3245322 -27.9502356  35.1828238
##        153        154        156        157        158        159
##  13.9465209   0.3748101 -12.9317852  13.4954663 -10.1255209 -13.6639541
##        160        161        162        164        165        166
## -26.6569849   2.6679721  23.2102519   9.3921313  -2.3372996  -6.2993982
##        167        168        169        170        171        172
##  -8.4035277  17.1741771 -12.9488687   1.5761538 -12.4465561  -2.1970247
##        174        175        176        177        178        182
##  58.2034025  60.4867558  75.3098734  76.6379962  51.4546264 -11.5275498
##        184        185        186        187        191        194
## -12.0117411 -33.7748666  -9.8756728   4.8254917 -11.1358722  -9.9152974
##        196        197        198        199        201        203
## -11.2027688   4.2242367 -32.4557801  14.4245721 -23.0144202 -14.7934423
##        204        205        207        208        209        211
##  10.5155253  -3.1968368 -26.7386036 -34.2171674  -9.7156935 -16.4751409
##        212        213        214        216        217        218
##   3.3017157  53.2161788 -20.7748539 -21.2233534 -29.0944971 -15.1513301
##        221        223        224        226        227        228
## -27.7648495 -20.0392075  -6.1804926  36.3318243 -11.3708274  -4.9034856
##        229        231        232        233        234        235
##  50.0869547 -14.8963675  -1.5411924 -29.8824351  28.1714407 -11.8701314
##        236        237        239        241        242        243
## -28.0328083  -9.8644626  27.9190069  -6.2341752  -7.8605358 -12.3713404
##        244        245        246        247        250        252
##  12.1640246  25.7456494  -8.4732867 -17.9015851 -13.6930513  -9.0080954
##        254        255        256        257        258        259
##  -8.3498542 -17.8853585 -21.9185047  -2.0360057  -9.4578753  -9.8656292
##        263        265        266        267        269        270
## -35.9445719  -8.0741044 -30.3962250  11.1493621  19.6442035  12.0887393
##        272        273        274        275        276        278
##  17.8031451  47.3198206  70.0701418 -20.5221213  48.0215279 -15.4996172
##        279        281        282        285        286        287
##  -3.0517632  31.1302408  42.3378551  50.9744861 -14.9103026  37.9621747
##        288        289        290        291        293        295
## -18.3778705  -9.4188821  -4.7558144 -13.9489801  10.3856500  12.7784722
##        298        299        301        302        303        304
##  -4.3030819  -8.3555965 -14.5888244 -11.1689002  -8.5496380  -3.3526971
##        305        306        307        308        309        310
##   1.9518705  -4.6140308 -14.2883681  -9.5804398 -21.4706842  -7.8331339
##        311        312        313        314        315        318
##  -4.2241581 -21.1722836 -13.0588554  -6.4876636   7.6751335   7.4160842
##        319        322        323        324        326        328
##  -3.9366791 -15.7695886  32.2887265 -20.7087747  22.0015023   0.6880291
##        329        331        332        333        335        336
##  -1.1200432 -23.2090683 -20.1662814 -14.3542870   4.8682018  15.4634378
##        337        338        339        341        342        343
##  21.3358550  -6.9043033  -2.0188233 -12.6982800   2.4438542  -5.6853855
##        344        345        346        348        349        351
##  -7.9183414   3.3813009  -2.5447195  -9.1255974  25.0465724 -23.0437199
##        353        354        355        357        358        359
##  15.0431010 -21.5536011 -30.2843616 -25.3562800  17.9858167 -14.9539299
##        361        362        364        365        366        367
## -28.1344040 -13.6018992  -8.3950731  26.2380770 -19.5931146 -15.3061502
##        368        369        370        371        372        374
##  -13.2407359  24.0904082 -26.1380544   3.6827740 -24.5154704  41.1229931
##        375        377        378        379        381        382
##  33.7850339 -21.2071849 -17.2318013 -15.8534676 -33.7455973  37.6366792
##        383        384        385        387        389        392
##  10.9022614  32.3022048  68.7493844 -34.6030644   5.3822305 -14.2768962
##        393        394        397        399        402        404
## -24.7048088 -20.6579018 -20.3218529 -11.3038808  43.4556926  -8.2320851
##        405        406        407        408        409        410
##  -5.2945292 -18.2230844  46.8283590 -30.0248383 -25.3943482  54.2329610
##        411        414        415        416        418        419
##  51.5227120 -27.2805967 -21.9361039 -25.3086072   5.1206261   2.6480169
##        420        421        422        423        424        426
## -26.0881689  -7.5698107  -7.9643705 -13.6840417 -14.2481025   1.0731029
##        427        428        429        432        433        435
## -11.3786023 -23.9747417   0.4222640   9.2835241  23.4203788  42.3533713
##        436        437        438        439        440        441
## -11.9711593  39.8103612  -4.0848287  -2.5344133  45.4787071   9.3956590
##        442        443        444        446        447        448
##  40.0136722  -6.3836024 -19.4073088  47.1767495 -14.8445344  26.3753809
##        449        450        451        453        454        455
## -23.7486656  17.5647484  -2.0911622 -14.8721364 -23.3175768 -27.4845902
##        458        459        460        462        463        464
## -20.4317626   4.0357450 -16.3741414 -22.5012912  -9.0825245  39.2483697
##        465        466        467        468        469        471
##  42.1919402   0.3991235   8.2995308  49.6781184  -3.1028610  19.5928281
##        473        474        475        476        477        478
## -24.5862800 -23.7626157 -11.8979560   5.8605886  -6.9047843   9.3679375
##        479        480        481        483        484        486
##  14.0705305   6.4982598  48.4997712  34.3504221  38.8995136 -17.0312293
##        487        488        489        492        493        494
##  60.8540012   8.3679878 -10.3396875  17.9247316  -1.8543112 -14.1034032
##        495        497        498        499        500        501
## -23.6625505   8.2339779 -23.4894029  12.0542736   6.6050491  -3.2823633
##        502        503        504        505        506        507
## -10.1150287 -22.6141380 -15.5664432   0.6053657  -0.9751435  -5.7707532
##        508        510        511        512        514        515
##  -9.1351782  -4.3203427   8.3152674  21.2310740  -3.9527803  38.9914859
##        516        517        518        519        521        522
##  -5.4247205 -14.8137287 -10.7326397  -8.2500654 -22.6874930 -34.8333315
##        523        524        525        526        528        530
##  71.1869961 -30.8745447 -19.8961878 -14.0670952 -12.8473043   3.2074413
##        532        533        535        536        537        539
##  -9.4473859  -9.9990542 -10.0241349  -0.1196050  -5.0082344  20.8859897
##        540        542        544        545        547        550
##  29.8605638 -20.1065927 -26.3178306 -11.6626864  -2.0668884  -8.5746645
##        552        553        555        556        557        558
## -30.2794165 -12.2176670  -7.3488676  -7.3452283  25.7437834   1.5804738
##        559        560        561        563        564        565
## -14.4371522  13.6151810  22.9460576   1.8469204   4.7761641  -6.6196698
##        566        567        571        573        576        577
##   0.6279051  22.0773011 -12.8023843  -8.6930556  34.5749956 -24.2991102
##        578        579        580        584        586        587
## -14.9807717  -6.4499804 -20.9337145 -22.6279776 -16.8038970 -21.1665582
##        588        590        591        592        594        595
## -11.1187125 -12.8834952  22.4621811  -8.2760207  19.2184314 -22.4398192
##        597        598        600        601        602        603
##  19.0971149   4.3535046  75.2489031 -10.7292419  14.0405075   6.3900818
##        604        605        607        609        610        611
##  47.0553061  -2.9594964   8.6770524  27.6578536 -22.1026331  25.1903744
##        612        613        615        616        618        620
## -28.0885563 -39.1115621  80.9228728  -6.8629557   0.2340546 -17.6202149
##        623        624        625        626        627        629
##   7.9983493 -11.7614706 -13.6242793   0.1092673  67.0387063 -15.2293404
##        630        631        633        635        636        638
##  -6.8593318  -3.4401082  37.9070318 -16.6878472  13.7892542  -5.6785657
##        640        641        642        644        645        646
##  -9.0225463  64.4576030  48.5674722 -25.8988398 -25.8337153  17.6963949
##        647        648        649        650        651        653
##  28.0011711  -0.6350884  61.2219377  -1.5603330  29.5117758  -5.9242935
##        654        655        658        659        660        663
## -14.9431734 -22.0390852 -13.0139394 -27.2006791  -0.5550274 -12.9823899
##        664        665        666        668        669        670
## -31.3679802  -8.8771360 -25.4335225 -35.8449891 -18.3062315  -7.7501254
##        671        672        673        674        675        676
## -16.0654078  -4.8548252 -17.3292317   0.4799814 -22.0906751   6.3124249
##        678        679        680        682        686        687
## -11.8160326  20.3363220  -9.6262380  16.8548507 -11.4346814 -31.6258134
##        688        689        690        691        693        694
##   1.9479856  -1.5218719  -9.3352086 -10.4531791 -14.7228415 -10.1209493
##        695        696        697        698        699        700
## -13.9765299 -19.2202541 -10.5838013 -15.7316701  55.3544865  -3.4732493
##        702        703        705        706        707        708
##  -3.1431288 -10.1764069 -15.5369254 -23.7635983 -11.4104970 -13.4649753
##        709        710        711        712        713        714
## -20.0988863 -34.6584774 -23.9583328 -22.1928811 -22.6311863 -15.8349496
##        715        716        717        718        720        721
##  24.6699173   7.6227675  61.3716458   7.0233355 -12.2808780 -15.4510837
##        722        723        725        726        728        729
##  -8.9808335  -9.1197126  57.0811191   4.8262844  -7.6545796   0.7999597
##        730        731        733        735        736        739
##   4.9017776 -32.0755360 -32.5414833 -18.5466888 -12.6377860 -17.9699037
##        740        742        743        744        746        747
## -17.8822696  40.7730207  33.2391035 -11.6864199  36.4448929  43.5576573
##        748        749        751        752        753        754
##   1.2296115 -15.6514924  -9.0335801  13.6635097 -17.6348040  28.9204478
##        755        756        757        759        760        762
##   7.6463495  -1.8298903  -4.3316055 -11.3397449   1.5012356  -6.5597335
##        763        764        765        766        767        768
## -20.1489675 -12.1359093  20.3176450   8.8733255  12.8955847   4.3839591
##        769        770        773        775        777        778
##  19.4164212 -16.1550252  21.1651258  34.8577723   0.7234134  -7.3253268
##        779        781
##  14.1757141 -14.0604960
```

# Análise em R

```
residuals(model) # residuals
```

```
##          1          2          3          7          9         10
## -25.9518636 -18.7473263 -17.8629292 -24.2819271   4.9152690 -24.3646542
##         12         13         14         15         17         18
## -24.3161484 -20.1459963  36.0113632  11.8146813 -22.9776762  51.1346371
##         19         22         23         26         27         28
##  11.7772209  16.1598697  10.4704219  22.1082226 -20.5830071  14.7521229
##         29         30         33         34         35         36
## -21.4063261 -12.4644281 -13.7373564   2.5195313 -18.3287535  -9.4111571
##         37         38         39         40         41         42
## -20.6897311 -10.5078759 -19.4744956 -28.5229815  -4.3659142  -2.5325477
##         43         44         45         46         47         48
##  36.4063054 -17.6034246  -1.6376052  -8.2157124 -22.4444309 -13.0603525
##         49         52         53         54         55         56
## -25.9197472 -14.1780998  37.6902066  -4.0341703 -26.4849320  14.5983383
##         57         58         60         61         62         63
## -16.7096259  45.5174752  -2.7525932  58.5947666   3.5171989  42.4868659
##         64         66         69         70         71         72
## -18.3132360 -21.7049184  -7.4409110 -25.1714792 -26.2442142 -25.0516241
##         73         74         75         76         77         78
##  21.9642039 -17.4120718  22.3229382 -22.6848007  -4.5709232  17.8748224
##         79         80         81         82         83         84
## -13.5247219  -6.8750528  10.2669922 -11.8106462 -12.3699836 -17.8530387
##         85         86         91         92         93         94
##  -6.0281123  -4.2048421  -0.5133142  -6.8618841  79.2891624  -4.1002821
##         95         96         97         98         99        100
##   3.4538202 -13.0380637 -24.5691866  29.2543677  -4.2346360  -6.7213120
##        101        102        103        105        108        109
##  15.5221753 -10.5603481   5.1930850   5.1519926   8.4144041  15.5708468
##        110        112        113        115        116        117
##  23.0968873  25.3985962  -5.2338845  16.7036017  42.0566517  70.9636162
##        119        120        121        122        123        124
## -16.7438327   8.8371449 -30.7137655 -16.7053943 -25.2773705  -9.8998612
##        125        127        128        129        130        133
##  18.5199289  12.0058013  -9.9221675  10.9426343  41.6257907  -4.8060415
##        134        136        138        140        141        142
##   8.2382789 -15.0990248  -7.7167622  29.7303145  53.8874597  40.3764583
##        146        147        148        149        150        152
##   9.2964633 -14.4914666 -33.9607836  62.3245322 -27.9502356  35.1828238
##        153        154        156        157        158        159
##  13.9465209   0.3748101 -12.9317852  13.4954663 -10.1255209 -13.6639541
##        160        161        162        164        165        166
## -26.6569849   2.6679721  23.2102519   9.3921313  -2.3372996  -6.2993982
##        167        168        169        170        171        172
##  -8.4035277  17.1741771 -12.9488687   1.5761538 -12.4465561  -2.1970247
##        174        175        176        177        178        182
##  58.2034025  60.4867558  75.3098734  76.6379962  51.4546264 -11.5275498
##        184        185        186        187        191        194
## -12.0117411 -33.7748666  -9.8756728   4.8254917 -11.1358722  -9.9152974
##        196        197        198        199        201        203
## -11.2027688   4.2242367 -32.4557801  14.4245721 -23.0144202 -14.7934423
##        204        205        207        208        209        211
##  10.5155253  -3.1968368 -26.7386036 -34.2171674  -9.7156935 -16.4751409
##        212        213        214        216        217        218
##   3.3017157  53.2161788 -20.7748539 -21.2233534 -29.0944971 -15.1513301
##        221        223        224        226        227        228
## -27.7648495 -20.0392075  -6.1804926  36.3318243 -11.3708274  -4.9034856
##        229        231        232        233        234        235
##  50.0869547 -14.8963675  -1.5411924 -29.8824351  28.1714407 -11.8701314
##        236        237        239        241        242        243
## -28.0328083  -9.8644626  27.9190069  -6.2341752  -7.8605358 -12.3713404
##        244        245        246        247        250        252
##  12.1640246  25.7456494  -8.4732867 -17.9015851 -13.6930513  -9.0080954
##        254        255        256        257        258        259
##  -8.3498542 -17.8853585 -21.9185047  -2.0360057  -9.4578753  -9.8656292
##        263        265        266        267        269        270
## -35.9445719  -8.0741044 -30.3962250  11.1493621  19.6442035  12.0887393
##        272        273        274        275        276        278
##  17.8031451  47.3198206  70.0701418 -20.5221213  48.0215279 -15.4996172
##        279        281        282        285        286        287
##  -3.0517632  31.1302408  42.3378551  50.9744861 -14.9103026  37.9621747
##        288        289        290        291        293        295
## -18.3778705  -9.4188821  -4.7558144 -13.9489801  10.3856500  12.7784722
##        298        299        301        302        303        304
##  -4.3030819  -8.3555965 -14.5888244 -11.1689002  -8.5496380  -3.3526971
##        305        306        307        308        309        310
##   1.9518705  -4.6140308 -14.2883681  -9.5804398 -21.4706842  -7.8331339
##        311        312        313        314        315        318
##  -4.2241581 -21.1722836 -13.0588554  -6.4876636   7.6751335   7.4160842
##        319        322        323        324        326        328
##  -3.9366791 -15.7695886  32.2887265 -20.7087747  22.0015023   0.6880291
##        329        331        332        333        335        336
##  -1.1200432 -23.2090683 -20.1662814 -14.3542870   4.8682018  15.4634378
##        337        338        339        341        342        343
##  21.3358550  -6.9043033  -2.0188233 -12.6982800   2.4438542  -5.6853855
##        344        345        346        348        349        351
##  -7.9183414   3.3813009  -2.5447195  -9.1255974  25.0465724 -23.0437199
##        353        354        355        357        358        359
##  15.0431010 -21.5536011 -30.2843616 -25.3562800  17.9858167 -14.9539299
##        361        362        364        365        366        367
## -28.1344040 -13.6018992  -8.3950731  26.2380770 -19.5931146 -15.3061502
##        368        369        370        371        372        374
##  -13.2407359  24.0904082 -26.1380544   3.6827740 -24.5154704  41.1229931
##        375        377        378        379        381        382
##  33.7850339 -21.2071849 -17.2318013 -15.8534676 -33.7455973  37.6366792
##        383        384        385        387        389        392
##  10.9022614  32.3022048  68.7493844 -34.6030644   5.3822305 -14.2768962
##        393        394        397        399        402        404
## -24.7048088 -20.6579018 -20.3218529 -11.3038808  43.4556926  -8.2320851
##        405        406        407        408        409        410
##  -5.2945292 -18.2230844  46.8283590 -30.0248383 -25.3943482  54.2329610
##        411        414        415        416        418        419
##  51.5227120 -27.2805967 -21.9361039 -25.3086072   5.1206261   2.6480169
##        420        421        422        423        424        426
## -26.0881689  -7.5698107  -7.9643705 -13.6840417 -14.2481025   1.0731029
##        427        428        429        432        433        435
## -11.3786023 -23.9747417   0.4222640   9.2835241  23.4203788  42.3533713
##        436        437        438        439        440        441
## -11.9711593  39.8103612  -4.0848287  -2.5344133  45.4787071   9.3956590
##        442        443        444        446        447        448
##  40.0136722  -6.3836024 -19.4073088  47.1767495 -14.8445344  26.3753809
##        449        450        451        453        454        455
## -23.7486656  17.5647484  -2.0911622 -14.8721364 -23.3175768 -27.4845902
##        458        459        460        462        463        464
## -20.4317626   4.0357450 -16.3741414 -22.5012912  -9.0825245  39.2483697
##        465        466        467        468        469        471
##  42.1919402   0.3991235   8.2995308  49.6781184  -3.1028610  19.5928281
##        473        474        475        476        477        478
## -24.5862800 -23.7626157 -11.8979560   5.8605886  -6.9047843   9.3679375
##        479        480        481        483        484        486
##  14.0705305   6.4982598  48.4997712  34.3504221  38.8995136 -17.0312293
##        487        488        489        492        493        494
##  60.8540012   8.3679878 -10.3396875  17.9247316  -1.8543112 -14.1034032
##        495        497        498        499        500        501
## -23.6625505   8.2339779 -23.4894029  12.0542736   6.6050491  -3.2823633
##        502        503        504        505        506        507
## -10.1150287 -22.6141380 -15.5664432   0.6053657  -0.9751435  -5.7707532
##        508        510        511        512        514        515
##  -9.1351782  -4.3203427   8.3152674  21.2310740  -3.9527803  38.9914859
##        516        517        518        519        521        522
##  -5.4247205 -14.8137287 -10.7326397  -8.2500654 -22.6874930 -34.8333315
##        523        524        525        526        528        530
##  71.1869961 -30.8745447 -19.8961878 -14.0670952 -12.8473043   3.2074413
##        532        533        535        536        537        539
##  -9.4473859  -9.9990542 -10.0241349  -0.1196050  -5.0082344  20.8859897
##        540        542        544        545        547        550
##  29.8605638 -20.1065927 -26.3178306 -11.6626864  -2.0668884  -8.5746645
##        552        553        555        556        557        558
## -30.2794165 -12.2176670  -7.3488676  -7.3452283  25.7437834   1.5804738
##        559        560        561        563        564        565
## -14.4371522  13.6151810  22.9460576   1.8469204   4.7761641  -6.6196698
##        566        567        571        573        576        577
##   0.6279051  22.0773011 -12.8023843  -8.6930556  34.5749956 -24.2991102
##        578        579        580        584        586        587
## -14.9807717  -6.4499804 -20.9337145 -22.6279776 -16.8038970 -21.1665582
##        588        590        591        592        594        595
## -11.1187125 -12.8834952  22.4621811  -8.2760207  19.2184314 -22.4398192
##        597        598        600        601        602        603
##  19.0971149   4.3535046  75.2489031 -10.7292419  14.0405075   6.3900818
##        604        605        607        609        610        611
##  47.0553061  -2.9594964   8.6770524  27.6578536 -22.1026331  25.1903744
##        612        613        615        616        618        620
## -28.0885563 -39.1115621  80.9228728  -6.8629557   0.2340546 -17.6202149
##        623        624        625        626        627        629
##   7.9983493 -11.7614706 -13.6242793   0.1092673  67.0387063 -15.2293404
##        630        631        633        635        636        638
##  -6.8593318  -3.4401082  37.9070318 -16.6878472  13.7892542  -5.6785657
##        640        641        642        644        645        646
##  -9.0225463  64.4576030  48.5674722 -25.8988398 -25.8337153  17.6963949
##        647        648        649        650        651        653
##  28.0011711  -0.6350884  61.2219377  -1.5603330  29.5117758  -5.9242935
##        654        655        658        659        660        663
## -14.9431734 -22.0390852 -13.0139394 -27.2006791  -0.5550274 -12.9823899
##        664        665        666        668        669        670
## -31.3679802  -8.8771360 -25.4335225 -35.8449891 -18.3062315  -7.7501254
##        671        672        673        674        675        676
## -16.0654078  -4.8548252 -17.3292317   0.4799814 -22.0906751   6.3124249
##        678        679        680        682        686        687
## -11.8160326  20.3363220  -9.6262380  16.8548507 -11.4346814 -31.6258134
##        688        689        690        691        693        694
##   1.9479856  -1.5218719  -9.3352086 -10.4531791 -14.7228415 -10.1209493
##        695        696        697        698        699        700
## -13.9765299 -19.2202541 -10.5838013 -15.7316701  55.3544865  -3.4732493
##        702        703        705        706        707        708
##  -3.1431288 -10.1764069 -15.5369254 -23.7635983 -11.4104970 -13.4649753
##        709        710        711        712        713        714
## -20.0988863 -34.6584774 -23.9583328 -22.1928811 -22.6311863 -15.8349496
##        715        716        717        718        720        721
##  24.6699173   7.6227675  61.3716458   7.0233355 -12.2808780 -15.4510837
##        722        723        725        726        728        729
##  -8.9808335  -9.1197126  57.0811191   4.8262844  -7.6545796   0.7999597
##        730        731        733        735        736        739
##   4.9017776 -32.0755360 -32.5414833 -18.5466888 -12.6377860 -17.9699037
##        740        742        743        744        746        747
## -17.8822696  40.7730207  33.2391035 -11.6864199  36.4448929  43.5576573
##        748        749        751        752        753        754
##   1.2296115 -15.6514924  -9.0335801  13.6635097 -17.6348040  28.9204478
##        755        756        757        759        760        762
##   7.6463495  -1.8298903  -4.3316055 -11.3397449   1.5012356  -6.5597335
##        763        764        765        766        767        768
## -20.1489675 -12.1359093  20.3176450   8.8733255  12.8955847   4.3839591
##        769        770        773        775        777        778
##  19.4164212 -16.1550252  21.1651258  34.8577723   0.7234134  -7.3253268
##        779        781
##  14.1757141 -14.0604960
```

# Análise em R

```
#Regressão Linear do TESTE:


#regressao_teste<-lm(test$min_engaged~.,test)

# predicting the target variable
predictions<-predict(model,test)


summary(predictions)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -19.38   23.44   28.81   27.69   33.77   48.88

head(predictions)

##         4         5         8        11        16        20
##  42.60954  12.69892  22.12850  27.76190 -19.38388  33.47344



predictions_without_negative <- predictions
predictions_without_negative[which(predictions_without_negative < 0)]<-0
```

**Comentário:**
Uma vez que não há intervalos de tempo negativos, criamos uma restrição adicional em que as previsões negativas são renomeadas para 0.

# Análise em R

```
#Absolute Error
mae(predictions,test$min_engaged)
```

```
## [1] 24.42624
```

```
mae(predictions_without_negative,test$min_engaged)
```

```
## [1] 24.20305
```

**Comentário:**
O erro absoluto do nosso modelo, incluindo todas as variáveis, é 24.42 dias. Se considerarmos as previsões da nossa variável dependente (min_engaged) negativas como sendo zero, uma vez que não há tempos negativos o erro é 24.20 dias.
O erro do nosso modelo é inferior ao erro do modelo trivial (37.20 dias).

# Análise em R

## Análise Fatorial

```
fit<-princomp(train, cor=TRUE)
summary(fit)

loadings(fit)

JO<-loadings(fit)


write.csv(JO,"./JO-Analise Factorial.csv")
```

**Comentário:**
Exportamos a tabela de correlações para poder fazer a análise fatorial.
A Análise Fatorial é um tipo de analise exploratória cujo objetivo é reduzir o número de variáveis (atributos dos dados), através da identificação de novos eixos fatoriais, não correlacionados entre si.
Pelo critério de Pearson devem reter15 componentes (ou fatores), que correspondem a um mínimo de 80% de proporção acumulada de variação explicada pelo modelo fatorial.

# Análise em R

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 | Comp.14 | Comp.15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| job_category_Back.end.Developer | 0.064 | 0.025 | 0.116 | 0.367 | 0.199 | 0.285 | 0.006 | **0.662** | 0.135 | 0.088 | 0.059 | 0.199 | 0.076 | 0.051 | 0.006 |
| job_category_Data.Scientist | -0.042 | 0.023 | -0.049 | -0.114 | 0.261 | 0.136 | 0.109 | -0.038 | -0.145 | 0.092 | **-0.648** | -0.166 | -0.533 | 0.203 | -0.064 |
| job_category_DevOps...Sysadmin | -0.060 | -0.099 | -0.031 | -0.040 | 0.144 | 0.017 | -0.135 | 0.053 | **-0.538** | -0.523 | 0.307 | -0.349 | 0.037 | 0.150 | -0.104 |
| job_category_Front.end.Developer | 0.072 | 0.232 | -0.012 | -0.269 | -0.153 | **-0.524** | -0.128 | 0.243 | -0.125 | 0.307 | 0.071 | -0.197 | -0.084 | -0.280 | 0.132 |
| job_category_Full.stack.Developer | 0.022 | -0.032 | -0.174 | -0.077 | **-0.591** | 0.491 | -0.032 | -0.234 | 0.111 | -0.043 | -0.081 | -0.122 | 0.101 | -0.119 | 0.033 |
| job_category_outros | -0.039 | -0.093 | 0.091 | 0.155 | 0.024 | -0.210 | **0.639** | -0.333 | -0.209 | 0.194 | 0.098 | 0.143 | 0.200 | 0.129 | 0.198 |
| job_category_Product.Project.Management | -0.038 | -0.086 | 0.044 | -0.235 | 0.275 | -0.080 | 0.017 | -0.157 | **0.703** | -0.112 | 0.255 | -0.210 | -0.057 | 0.160 | -0.136 |
| job_category_QA...Testing | -0.020 | -0.042 | -0.005 | 0.038 | 0.185 | 0.055 | -0.440 | -0.357 | -0.169 | 0.323 | 0.155 | **0.462** | -0.108 | -0.140 | -0.355 |
| job_category_UX...UI.Designer | -0.025 | 0.025 | 0.031 | 0.138 | -0.061 | -0.313 | -0.169 | -0.084 | 0.143 | **-0.500** | -0.461 | 0.353 | 0.194 | -0.051 | 0.172 |
| contractor | **0.487** | -0.150 | 0.030 | -0.051 | 0.005 | -0.039 | -0.050 | -0.003 | -0.029 | -0.056 | -0.011 | -0.010 | 0.026 | 0.027 | -0.021 |
| permanent | **-0.559** | 0.171 | -0.012 | 0.049 | -0.021 | 0.000 | -0.008 | 0.032 | 0.029 | 0.001 | 0.012 | -0.050 | -0.004 | -0.106 | -0.009 |
| relocation_paid | -0.011 | 0.280 | **0.462** | -0.059 | 0.112 | 0.194 | -0.038 | -0.082 | -0.108 | 0.048 | 0.001 | 0.034 | 0.111 | 0.155 | 0.201 |
| visa_support | 0.012 | 0.308 | **0.438** | -0.214 | 0.040 | 0.227 | -0.016 | -0.100 | -0.058 | -0.061 | -0.001 | 0.007 | 0.188 | 0.060 | -0.005 |
| show_sal_rate | 0.123 | 0.225 | -0.118 | 0.207 | -0.133 | 0.046 | 0.242 | 0.061 | 0.011 | 0.181 | -0.104 | -0.241 | 0.236 | 0.002 | **-0.483** |
| home_work | 0.110 | 0.229 | -0.290 | 0.304 | 0.042 | 0.096 | -0.217 | -0.096 | 0.033 | 0.038 | 0.141 | -0.095 | -0.089 | -0.021 | **0.483** |
| education_req | 0.014 | -0.266 | 0.090 | -0.160 | 0.291 | 0.268 | 0.154 | -0.053 | 0.023 | 0.055 | 0.027 | -0.102 | -0.045 | **-0.562** | 0.338 |
| exp_nivel | 0.069 | 0.087 | 0.218 | 0.357 | -0.092 | -0.133 | 0.256 | -0.066 | 0.103 | -0.242 | 0.174 | 0.028 | **-0.476** | -0.213 | -0.113 |
| languagePT | -0.099 | -0.236 | -0.157 | **-0.318** | 0.071 | 0.047 | 0.197 | 0.283 | -0.088 | -0.099 | -0.098 | 0.228 | 0.201 | -0.250 | -0.102 |
| regionPT | -0.169 | -0.128 | -0.364 | -0.125 | 0.022 | 0.025 | 0.069 | 0.039 | 0.012 | 0.104 | 0.167 | 0.214 | -0.021 | **0.473** | 0.181 |
| full_time | **-0.559** | 0.171 | -0.012 | 0.049 | -0.021 | 0.000 | -0.008 | 0.032 | 0.029 | 0.001 | 0.012 | -0.050 | -0.004 | -0.106 | -0.009 |
| vist_dia | 0.122 | 0.296 | **-0.368** | 0.138 | 0.341 | 0.061 | 0.034 | -0.105 | -0.027 | -0.027 | -0.065 | -0.109 | 0.127 | -0.074 | 0.116 |
| apps_dia | 0.155 | **0.445** | -0.099 | -0.358 | 0.107 | -0.064 | 0.079 | 0.028 | 0.105 | -0.034 | -0.016 | 0.118 | 0.125 | 0.048 | 0.005 |
| hands_dia | 0.083 | 0.228 | -0.052 | -0.253 | -0.276 | 0.153 | 0.185 | 0.154 | -0.068 | -0.155 | 0.216 | 0.340 | **-0.431** | 0.062 | 0.076 |

# Análise em R

**Comentário:**

Fazendo a interpretação dos componentes temos:
**Comp.1** - parece dizer respeito a vagas para emprego a full time e permanentes
**Comp.2** - nº de candidaturas por dia
**Comp.3** - está mais relacionada com as variáveis relocation_paid e visa_support, provavelmente associados a empregos para mão de obra estrangeira
**Comp.4** - está associada a vagas para pessoas que falam língua portuguesa
**Comp.5** - associado a vagas para a categoria de Full stack Developer
**Comp.6** - associado a vagas para a categoria de Frontend Developer
**Comp.7** - associado a vagas para a categoria "outros"
**Comp.8** - associado a vagas para a categoria de Backend Developer
**Comp.9** - associado a Tipologia de job?
**Comp.10** - associado a vagas para a categoria de UX...UI.Designer
**Comp.11** - associado a vagas para a categoria de Data Scientist
**Comp.12** - associado a vagas para a categoria de QA...Testing
**Comp. 13** - associado ao nível de experiência da vaga
**Comp. 14** - associado à localização e educação
**Comp. 15** - associado ao trabalho remoto e visualização do salário

# Análise em R

## Feature Selection com Teste ANOVA

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: min_engaged
##                                          Df Sum Sq Mean Sq F value    Pr(>F)
## job_category_Back.end.Developer           1    573   573.5  0.9935  0.319330
## job_category_Data.Scientist               1    276   275.8  0.4778  0.489689
## job_category_DevOps...Sysadmin            1    623   623.3  1.0798  0.299184
## job_category_Front.end.Developer          1    309   308.7  0.5348  0.464925
## job_category_Full.stack.Developer         1   1290  1290.1  2.2348  0.135501
## job_category_outros                       1     51    51.1  0.0884  0.766270
## job_category_Product.Project.Management   1      0     0.0  0.0000  0.994839
## job_category_QA...Testing                 1    923   923.1  1.5990  0.206569
## contractor                                1    678   677.9  1.1744  0.278975
## permanent                                 1     24    23.7  0.0411  0.839423
## relocation_paid                           1    736   735.8  1.2747  0.259369
## visa_support                              1   2008  2008.4  3.4792  0.062671
## show_sal_rate                             1   4814  4814.0  8.3395  0.004030
## home_work                                 1   4041  4040.5  6.9995  0.008384
## education_req                             1     31    31.5  0.0545  0.815518
## exp_nivel                                 1    419   419.3  0.7264  0.394430
## languagePT                                1     80    79.6  0.1378  0.710608
## regionPT                                  1      4     3.7  0.0064  0.936272
## vist_dia                                  1  25035 25035.4 43.3696 1.051e-10
## apps_dia                                  1  11356 11355.7 19.6717 1.109e-05
## hands_dia                                 1   1681  1680.5  2.9112  0.088524
## Residuals                               556 320955   577.3
##
## job_category_Back.end.Developer
## job_category_Data.Scientist
## job_category_DevOps...Sysadmin
## job_category_Front.end.Developer
## job_category_Full.stack.Developer
## job_category_outros
## job_category_Product.Project.Management
## job_category_QA...Testing
## contractor
## permanent
## relocation_paid
## visa_support                              .
## show_sal_rate                             **
## home_work                                 **
## education_req
## exp_nivel
## languagePT
## regionPT
## vist_dia                                  ***
## apps_dia                                  ***
## hands_dia                                 .
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Análise em R

```
#Regressão Linear com Feature Selection Técnica ANOVA para alfa = 0.05
ANOVA_treino<-lm(min_engaged~show_sal_rate+home_work+vist_dia+apps_dia,train)
ANOVA_treino
```

```
##
## Call:
## lm(formula = min_engaged ~ show_sal_rate + home_work + vist_dia +
##     apps_dia, data = train)
##
## Coefficients:
##   (Intercept)  show_sal_rate      home_work       vist_dia       apps_dia
##       37.9998        -2.3443        -1.4215        -0.1647       -11.9255
```

**Comentário:**

A regressão linear com as variáveis selecionadas através da técnica ANOVA.

Assumimos um alfa igual a 0.05.

- As variáveis selecionadas foram show_sal_rate (coeficiente negativo), apps_dia (coeficiente negativo), vist_dia (coeficiente negativo) e home_work (coeficiente negativo).
- Um coeficiente negativo significa que influencia negativamente o interval de tempo, ou seja, será menor. Logo, mais depressa teremos um "engaged" no job.

# Análise em R

```
# predicting the target variable
predictions_ANOVA<-predict.lm(ANOVA_treino,test)
predictions_ANOVA
```

```
##          4          5          8         11         16         20         21
##  33.592887  12.705014  25.983715  25.044533 -15.029798  32.642196  30.092308
##         24         31         32         50         59         65         67
##  27.187452  25.854806  24.456969  36.202328   5.431750  30.734712  35.579894
##         68         87         88         89        104        106        107
##  29.499990  29.005735  31.813303  34.565085  34.314322  29.206221  25.539539
##        111        114        118        126        132        137        139
##  34.061282  32.386955  28.013020  28.292313  35.627006  26.152695   2.883642
##        145        151        173        179        181        189        190
##  17.883247  33.047673  29.108084  30.421117  36.227759  26.470121  29.697136
##        193        195        202        206        219        220        222
##  27.004023  27.989037  16.171837  31.821654  33.746662  32.018232  32.580580
##        230        238        240        248        249        260        261
##  34.513848  24.281428  19.308990  28.074542  12.441796  11.401046  22.778378
##        262        264        271        277        294        296        297
##  24.898467  34.598087  31.010533  33.111515  29.497035  28.608384  28.669597
##        316        317        320        321        327        330        334
##  31.263391  21.496776 -12.723731  32.535364  28.948082  32.366529  19.541364
##        340        347        352        356        360        363        373
##  19.702499  33.111026  25.036290  25.388898  18.120976  18.618485  32.250998
##        376        380        386        391        400        401        403
##  35.104574  30.047637  30.107369  33.841190  36.959636  36.540976  34.010687
##        412        417        425        430        431        434        445
##  22.113402  32.103878  25.505638  29.253302  37.023741  26.033366  32.112571
##        456        457        461        470        472        482        485
##  23.823821  26.952749  30.512730  29.456017  31.262859  30.354834  33.992775
##        490        491        496        509        513        520        527
##  24.083656   7.300129  27.272168  31.496882  11.005795  36.260688  22.446346
##        529        531        534        538        541        543        546
##  14.404189  12.824616  17.945430  26.951536   9.740852  21.348078  31.997501
##        548        549        554        562        568        570        572
##  31.898213  31.853283  35.000965  33.649872  33.886378  28.335286  34.466727
##        575        581        582        583        585        589        593
##  29.520713  30.841191  17.797564  35.627968  33.577445  30.500364  26.683236
##        596        599        606        608        614        617        619
##  30.369870  32.095392  29.291474  25.506394  30.502877  29.091030  30.827057
##        621        622        628        632        637        639        643
##  31.759535  31.606173  25.343785  33.596497  29.613508  21.872052  33.236070
##        652        656        657        661        662        667        677
##  30.617841  32.837391  28.384375   4.115763  16.471424  33.562938  35.993504
##        683        684        685        701        704        719        724
##  25.692140  20.255385  25.045484  29.631908  22.110992  31.930251  29.122496
##        727        732        734        737        738        741        750
##  34.460528  34.795620  30.843764  31.706889  11.348242  31.068494  32.472587
##        758        771        772        774        780
##  34.645126  23.063752  30.156564  32.680579  34.421967
```

# Análise em R

```
predictions_ANOVA_without_negative <- predictions_ANOVA
predictions_ANOVA_without_negative[which(predictions_ANOVA_without_negative < 0)]<-0

#Absolute Error
mae(predictions_ANOVA,test$min_engaged)
```

```
## [1] 23.40691
```

```
mae(predictions_ANOVA_without_negative,test$min_engaged)
```

```
## [1] 23.23236
```

**Comentário:**
O erro absoluto do nosso modelo, incluindo apenas as variáveis selecionadas pelo teste da Anova (com alfa = 0.05), é 23.41 dias. Se considerarmos as previsões da nossa variável dependente (min_engaged) negativas como sendo zero, uma vez que não há tempos negativos, o erro é 23.23 dias.
O erro do nosso modelo é inferior ao erro do modelo trivial (37.20 dias).

# Análise em R

Porto Business School / University of Porto

## Feature Selection - Técnica Step Wise Forward and Backward Selection

```
#Step 1: Define base intercept only model
base.mod <- lm(min_engaged ~ 1 , data=test)

#Step 2: Full model with all predictors
all.mod <- lm(min_engaged ~ . , data=test)

# Step 3: Perform step-wise algorithm. direction='both' implies both forward and backward stepwise
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction = "both", trace = 0, steps = 1000)

# Step 4: Get the shortlisted variable.
shortlistedVars<-names(unlist(stepMod[[1]]))
shortlistedVars<-shortlistedVars[!shortlistedVars %in% "(Intercept)"] #remover

print(shortlistedVars)


## [1] "apps_dia"                          "home_work"
## [3] "job_category_Back.end.Developer"
```

**Comentário:**
Stepwise regression can be used to select features if the Y variable is a numeric variable. It is particularly used in selecting best linear regression models. It searches for the best possible regression model by iteratively selecting and dropping variables to arrive at a model with the lowest possible AIC. It can be implemented using the step() function and you need to provide it with a lower model, which is the base model from which it won't remove any features and an upper model, which is a full model that has all possible features you want to have.

# Análise em R

```
#Regressão Linear com Feature Selection Técnica SWFBS
SWFBS_treino<-lm(min_engaged~vist_dia+apps_dia+job_category_Back.end.Developer,train)
SWFBS_treino
```

```
## Call:
## lm(formula = min_engaged ~ vist_dia + apps_dia + job_category_Back.end.Developer,
##     data = train)
##
## Coefficients:
##                       (Intercept)                           vist_dia
##                           35.8950                            -0.1782
##                          apps_dia  job_category_Back.end.Developer
##                          -11.8661                             2.8448
```

**Comentário:**
As variáveis selecionadas foram vist_dia (coeficiente negativo), apps_dia (coeficiente negativo) e job_category_Back.end.Developer (coeficiente positivo).
Um coeficiente negativo significa que influencia negativamente o interval de tempo, ou seja, será menor. Logo, mais depressa teremos um "engaged" no job. De notar que para a variável job_category_Back.end.Developer ocorre a situação oposta, já que o coeficiente é positivo.

# Análise em R

```
# predicting the target variable
predictions_SWFBS<-predict(SWFBS_treino,test)
predictions_SWFBS
```

```
##          4          5          8         11         16         20         21
##  32.818894  11.027003  24.868105  24.066963 -15.346950  34.640869  31.654914
##         24         31         32         50         59         65         67
##  31.337267  26.939279  28.625130  36.854402   5.012595  29.616649  37.681452
##         68         87         88         89        104        106        107
##  28.545887  27.854383  30.835993  36.601341  33.465306  30.684871  24.709797
##        111        114        118        126        132        137        139
##  36.090747  31.485845  26.748425  27.005740  33.425834  27.300442   3.164446
##        145        151        173        179        181        189        190
##  17.065132  32.144509  27.869765  29.495940  36.879848  25.043206  29.744919
##        193        195        202        206        219        220        222
##  28.499174  29.769821  13.943930  32.464890  32.960200  33.544188  34.543515
##        230        238        240        248        249        260        261
##  33.698854  22.839305  19.864049  26.731254   9.998135  11.789776  23.804912
##        262        264        271        277        294        296        297
##  23.429589  33.782326  32.448809  35.038646  30.856758  28.695870  30.124290
##        316        317        320        321        327        330        334
##  30.447385  19.795635 -13.152670  31.676532  30.487411  30.075464  18.294990
##        340        347        352        356        360        363        373
##  18.439876  33.263129  23.636522  24.014625  16.451453  17.042804  33.849039
##        376        380        386        391        400        401        403
##  32.921925  34.299231  29.006136  32.971922  34.783490  34.347773  33.228656
##        412        417        425        430        431        434        445
##  23.140166  31.346985  29.342095  26.979639  34.868825  27.409663  32.226529
##        456        457        461        470        472        482        485
##  25.226125  28.294667  31.963380  31.254361  30.351677  32.106837  34.115982
##        490        491        496        509        513        520        527
##  25.666528   6.885972  28.624986  29.264148  11.576557  34.062602  20.972226
##        529        531        534        538        541        543        546
##  15.547434  11.713533  15.910348  28.370166   7.411918  22.770313  32.127507
##        548        549        554        562        568        570        572
##  36.229759  33.690508  32.789986  36.617981  33.052596  27.005611  33.702642
##        575        581        582        583        585        589        593
##  28.486151  29.737185  21.133202  33.397557  32.664545  32.061189  28.172966
##        596        599        606        608        614        617        619
##  29.260795  33.959114  28.051183  29.703174  31.883163  30.610868  28.579021
##        621        622        628        632        637        639        643
##  33.295072  30.651820  23.797368  35.572161  30.913623  22.804032  32.364851
##        652        656        657        661        662        667        677
##  29.480927  31.977394  29.905642   3.608669  16.891659  32.682211  33.758906
##        683        684        685        701        704        719        724
##  24.135672  18.262655  26.349752  28.540602  21.797602  33.471199  33.285195
##        727        732        734        737        738        741        750
##  32.214037  32.564593  32.258356  36.064730  14.592209  30.176466  30.154456
##        758        771        772        774        780
##  32.497223  24.167065  29.274122  34.262314  36.449430
```

# Análise em R

```
predictions_SWFBS_without_negative <- predictions_SWFBS
predictions_SWFBS_without_negative[which(predictions_SWFBS_without_negative < 0)]<-0

#Absolute Error
mae(predictions_SWFBS,test$min_engaged)
```

```
## [1] 23.4778
```

```
mae(predictions_SWFBS_without_negative,test$min_engaged)
```

```
## [1] 23.29856
```

**Comentário:**

Uma vez que não há intervalos de tempo negativos, criamos uma restrição adicional em que as previsões negativas são renomeadas para 0

O erro absoluto do nosso modelo, incluindo apenas as variáveis selecionadas pelo teste SWFBS é 23.47 dias. Se considerarmos as previsões da nossa variável dependente (min_engaged) negativas como sendo zero, uma vez que não há tempos negativos, o erro é 23.29 dias.

O erro do nosso modelo é inferior ao erro do modelo trivial (37.20 dias).

# Análise em R

## Feature Selection - Técnica Boruta

```
# Técnica BORUTA
boruta_output <- Boruta(min_engaged ~ ., data=na.omit(train), doTrace=0)
names(boruta_output)
```

```
##  [1] "finalDecision" "ImpHistory"    "pValue"        "maxRuns"
##  [5] "light"         "mcAdj"         "timeTaken"     "roughfixed"
##  [9] "call"          "impSource"
```

```
# Get significant variables including tentatives
boruta_signif <- getSelectedAttributes(boruta_output, withTentative = TRUE)
print(boruta_signif)
```

```
## [1] "contractor"      "relocation_paid" "visa_support"    "home_work"
## [5] "vist_dia"        "apps_dia"        "hands_dia"
```

**Comentário:**
Boruta is a feature ranking and selection algorithm based on random forests algorithm. The advantage with Boruta is that it clearly decides if a variable is important or not and helps to select variables that are statistically significant. Besides, you can adjust the strictness of the algorithm by adjusting the p values that defaults to 0.01 and the maxRuns. maxRuns is the number of times the algorithm is run. The higher the maxRuns the more selective you get in picking the variables. The default value is 100. In the process of deciding if a feature is important or not, some features may be marked by Boruta as 'Tentative'. Sometimes increasing the maxRuns can help resolve the 'Tentativeness' of the feature.

# Análise em R

```
# Do a tentative rough fix
roughFixMod <- TentativeRoughFix(boruta_output)
boruta_signif <- getSelectedAttributes(roughFixMod)
print(boruta_signif)
```

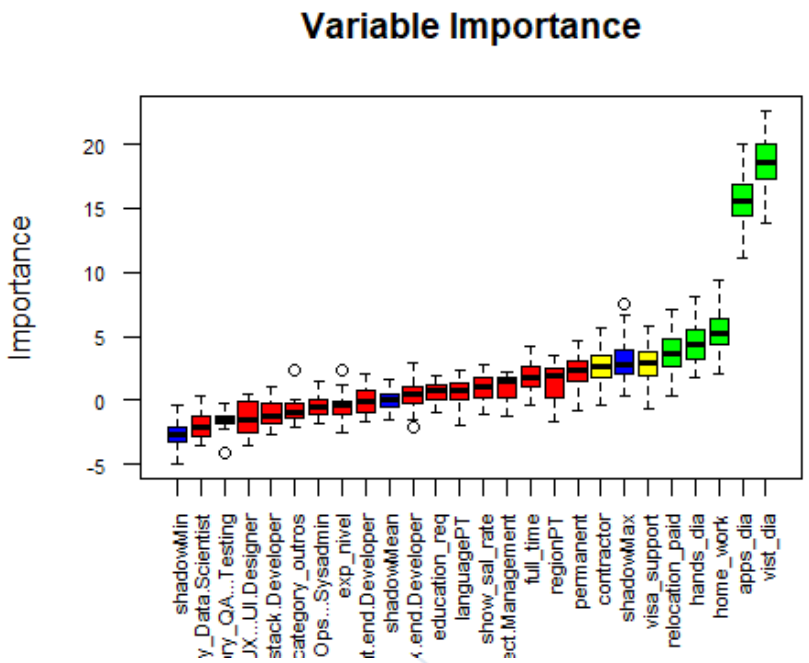
```
## [1] "relocation_paid" "home_work"       "vist_dia"        "apps_dia"
## [5] "hands_dia“
```

```
# Variable Importance Scores
imps <- attStats(roughFixMod)
imps2 <-imps[imps$decision != 'Rejected', c('meanImp', 'decision')]
head(imps2[order(-imps2$meanImp), ])  # descending sort
```

```
##                   meanImp  decision
## vist_dia        18.443570 Confirmed
## apps_dia        15.560795 Confirmed
## home_work        5.337014 Confirmed
## hands_dia        4.472158 Confirmed
## relocation_paid  3.734750 Confirmed
```

# Análise em R

```
# Plot variable importance
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")
```




**Comentário:**
This plot reveals the importance of each of the features. The columns in green are 'confirmed' and the ones in red are not. There are couple of blue bars representing ShadowMax and ShadowMin. They are not actual features, but are used by the boruta algorithm to decide if a variable is important or not.
As variáveis selecionadas com esta técnica são: vist_dia, apps_dia, home_work, hands_dia e relocation_paid.

# Análise em R

```
#Regressão Linear com Feature Selection Técnica Boruta
boruta_treino<-lm(min_engaged~hands_dia+home_work+vist_dia+apps_dia+relocation_paid,train)
boruta_treino

## Call:
## lm(formula = min_engaged ~ hands_dia + home_work + vist_dia +
##     apps_dia + relocation_paid, data = train)
##
## Coefficients:
##    (Intercept)        hands_dia        home_work         vist_dia         apps_dia  relocation_paid
##        37.5882          -2.4570          -1.3510          -0.1755         -10.7647           3.6745
```

**Comentário:**
As variáveis selecionadas foram hands_dia (coeficiente negativo), apps_dia (coeficiente negativo), vist_dia (coeficiente negativo), home_work (coeficiente negativo) e relocation_paid (coeficiente positivo).
Um coeficiente negativo significa que influencia negativamente o interval de tempo, ou seja, será menor. Logo, mais depressa teremos um "engaged" no job. De referir que acontece o contrário com a variavel relocation_paid.

# Análise em R

```
# predicting the target variable
predictions_boruta<-predict(boruta_treino,test)

predictions_boruta_without_negative <- predictions_boruta
predictions_boruta_without_negative[which(predictions_boruta_without_negative < 0)]<-0

#Absolute Error
mae(predictions_boruta,test$min_engaged)
## [1] 23.39883

mae(predictions_boruta_without_negative,test$min_engaged)
## [1] 23.39883
```

**Comentário:**
O erro absoluto do nosso modelo, incluindo apenas as variáveis selecionadas pelo teste do Boruta é 23.40 dias. Se considerarmos as previsões da nossa variável dependente (min_engaged) negativas como sendo zero, uma vez que não há tempos negativos, o erro é 23.40 dias.
O erro do nosso modelo é inferior ao erro do modelo trivial (37.20 dias).

Porto Business School / University of Porto

## Resumo

| Técnica | Que variáveis ficam? | Influência |
|---|---|---|
| Feature Selection com Teste ANOVA | apps_dia | - |
| | vist_dia | - |
| | show_sal_rate | - |
| | home_work | - |
| Feature Selection - Técnica Step Wise Forward and Backward Selection | vist_dia | - |
| | apps_dia | - |
| | job_category_Back.end.Developer | + |
| Feature Selection - Técnica Boruta (based random forests) | home_work | - |
| | hands_dia | - |
| | vist_dia | - |
| | apps_dia | - |
| | relocation_paid | + |

# Análise em R

## Resumo

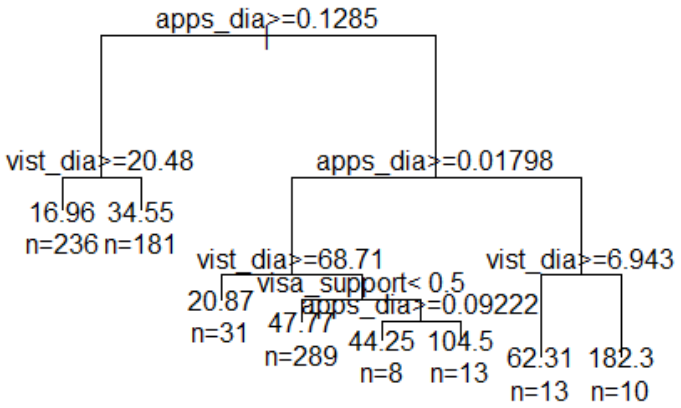
| | Erro absoluto | Erro absoluto (Min_engaged>0) |
|---|---|---|
| Regressão Linear Simples | 24.42 | 24.20 |
| Feature Selection com Teste ANOVA | 23.41 | 23.23 |
| Feature Selection - Técnica Step Wise Forward and Backward Selection | 23.48 | 23.30 |
| Feature Selection - Técnica Boruta (based on random forests) | 23.40 | 23.40 |

Erro absoluto do modelo trivial: **37.20 dias**

# Análise em R

## Modelo de Arvore de decisão

```
{arvore<-rpart(min_engaged ~.,L1)
arvore
par(mfrow = c(1,1), xpd = NA) # otherwise on some devices the text is clipped
plot(arvore)
text(arvore, use.n = TRUE)}
```




**Comentário:**
Uma árvore de decisão é uma ferramenta de suporte à tomada de decisão que usa um gráfico no formato de árvore.

# Análise em R

Porto Business School / University of Porto

```
model.set <- list(model, ANOVA_treino, SWFBS_treino, boruta_treino)
model.names <- c("model", "ANOVA_treino", "SWFBS_treino", "boruta_treino")

aictab(model.set, modnames = model.names)
```

```
##
## Model selection based on AICc:
##
##                K    AICc Delta_AICc AICcWt Cum.Wt       LL
## boruta_treino  6 5327.93       0.00   0.60   0.60 -2657.89
## SWFBS_treino   5 5329.47       1.54   0.28   0.88 -2659.68
## ANOVA_treino   6 5331.13       3.19   0.12   1.00 -2659.49
## model         23 5340.95      13.01   0.00   1.00 -2646.48
```

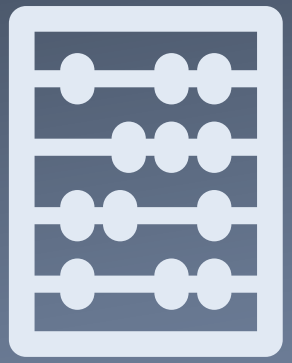
**Comentário:**
O critério de informação de Akaike (AIC) é um método matemático para avaliar quão bem um modelo se ajusta aos dados dos quais foi gerado. Em estatística, o AIC é usado para comparar diferentes modelos possíveis e determinar qual deles é o mais adequado para os dados. O AIC é calculado a partir de:

- o número de variáveis independentes usadas para construir o modelo..
- a estimativa de probabilidade máxima do modelo (quão bem o modelo reproduz os dados).
- O modelo que melhor se ajusta de acordo com o critério AIC é aquele que explica a maior quantidade de variação usando o menor número possível de variáveis independentes.

As variáveis selecionadas com esta técnica são: home_work, hands_dia, vist_dia, apps_dia, relocation_paid

# Análise em R

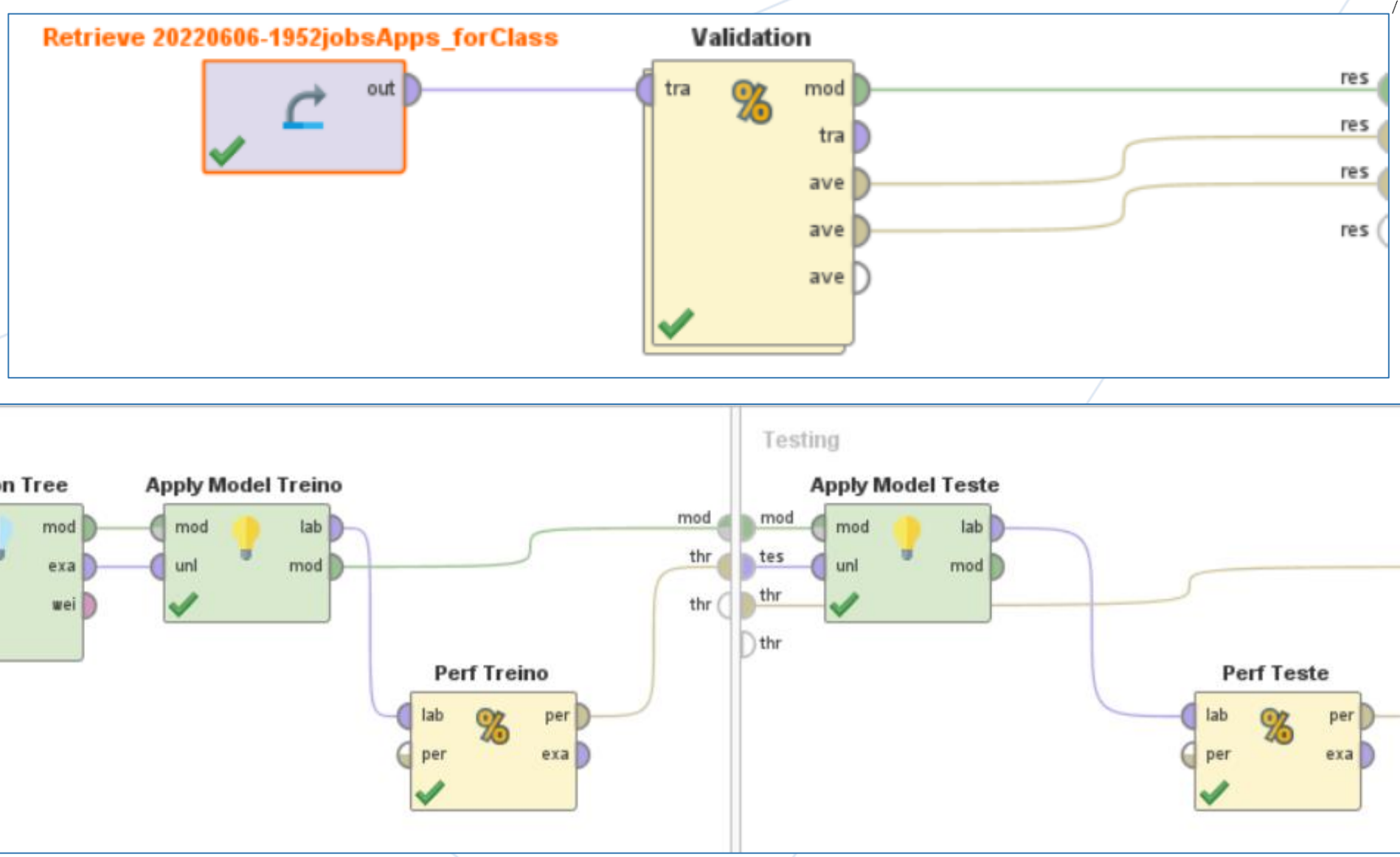
- O melhor modelo é aquele listado em primeiro. A tabela de seleção de modelos inclui informações sobre:
- **K**: O número de parâmetros no modelo. O K padrão é 2, então um modelo com um parâmetro terá um K de 2 + 1 = 3.
- **AICc**: A pontuação de informação do modelo (o 'c' minúsculo indica que o valor foi calculado a partir do teste AIC corrigido para pequenos tamanhos de amostra). Quanto menor o valor de AIC, melhor o ajuste do modelo..
- **Delta_AICc**: A diferença na pontuação AIC entre o melhor modelo e o modelo que está a ser comparado. Nesta tabela, o *next-best model* tem um delta-AIC de 6,33 em comparação com o modelo superior, e o terceiro melhor modelo tem um delta-AIC de 17,57 em comparação com o modelo superior.
- **AICcWt**: peso AICc, que é a proporção da quantidade total de poder preditivo fornecido pelo conjunto completo de modelos contidos no modelo que está a ser avaliado. Nesse caso, o modelo superior contém 96% da explicação total que pode ser encontrado no conjunto completo dos modelos.
- **Cum.Wt**: A soma dos pesos AICc. Aqui, os dois principais modelos contêm 100% do peso acumulado do AICc..
- **LL**: Este é o valor que descreve a probabilidade do modelo, dados os dados. A pontuação AIC é calculada a partir do LL e K.



# 3. Modeling | Classificação

Problema: **O “job” vai ter, pelo menos, um “engaged”?**

# Classificação

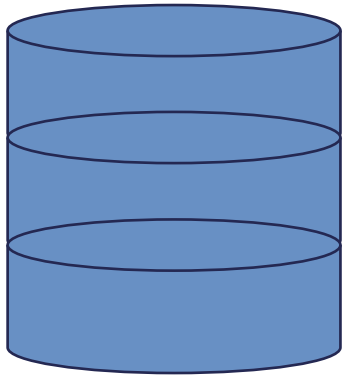


# Divisão do Dataset em treino e teste

***2278 observações***

*Classe positiva=26%*

*Classe negativa=74%*



70%

30%

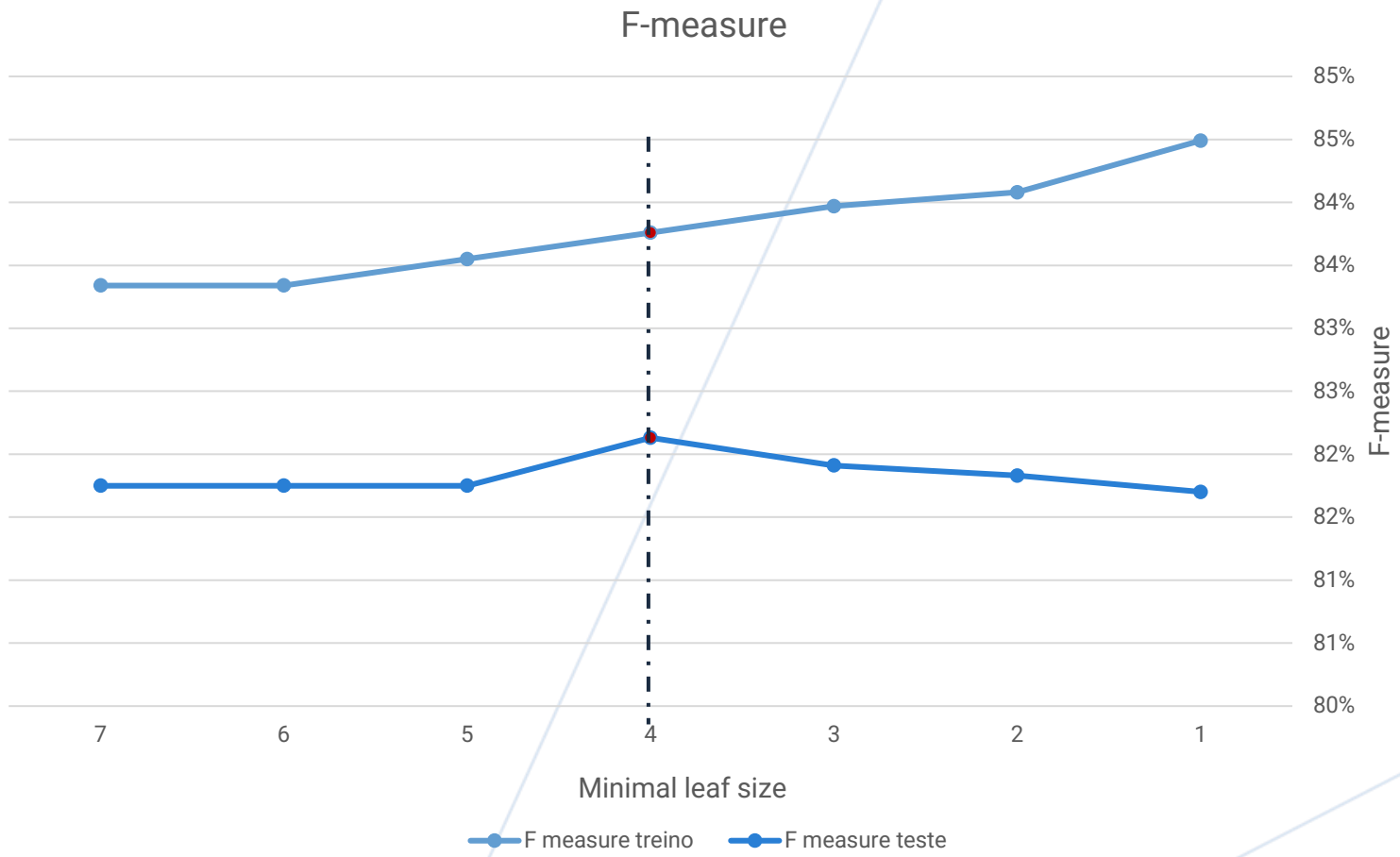
Treino

Teste

*1595 observações*

*683 observações*

# Overfit




Selecionamos a F1-measure como medida de precisão pois é mais útil em dataset desbalanceados, pois permite avaliar em simultâneo o custo de Precision and Recall.

A F1-measure é uma media ponderada de Precision e Recall

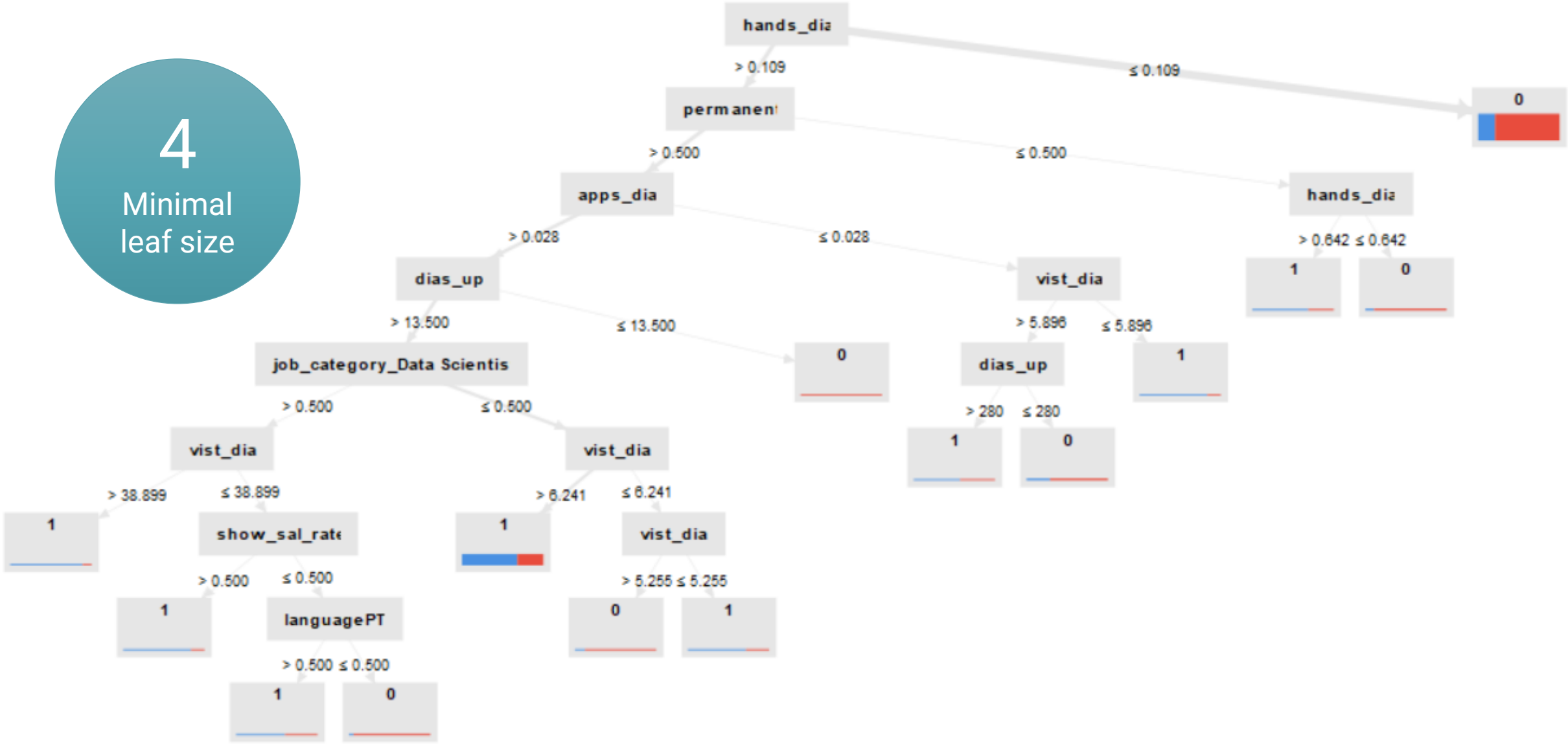
F1-measure=
2*(Recall * Precision) / (Recall + Precision)

- Até a um minimal leaf size de 4 a F1-measure de teste  estava a aumentar, apatir desse momento começou a diminuir pois o modelo estava a  adaptar-se demais aos dados de treino

Maximal depth = 10

# Decision Tree

4
Minimal
leaf size




Maximal depth = 10

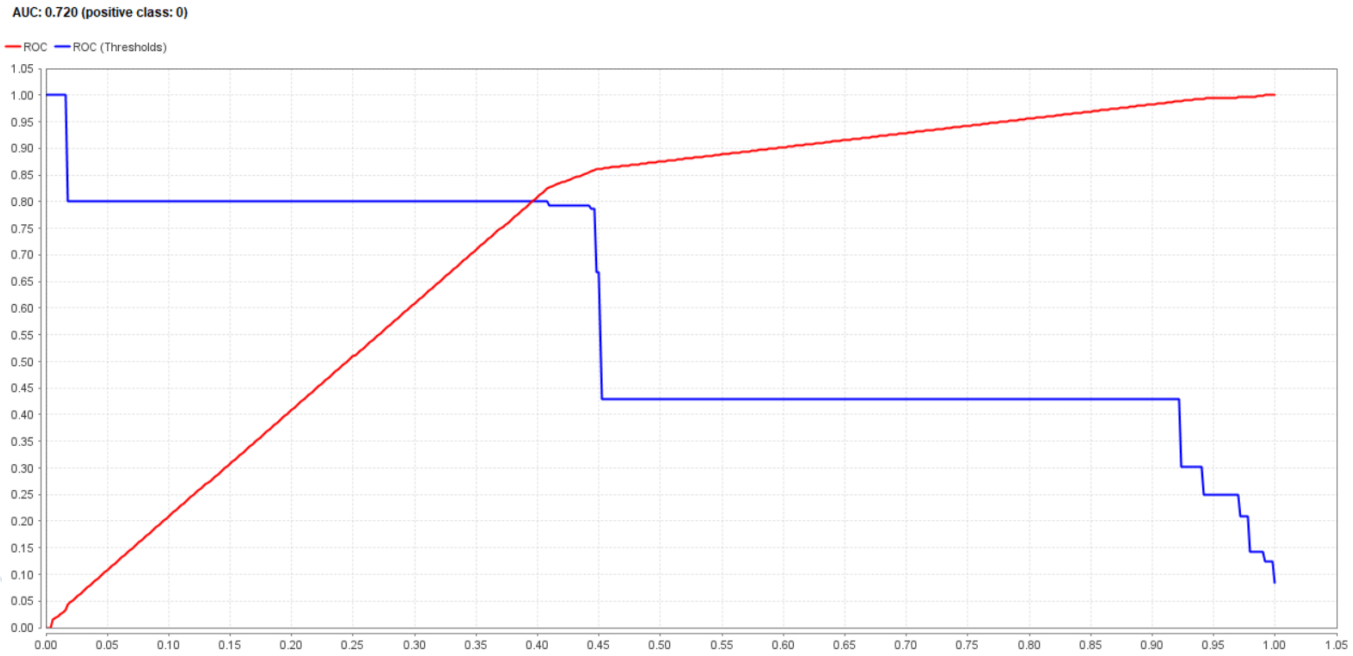
# Matriz de Confusão e Curva ROC: Teste

**f_measure: 82.13% (positive class: 0)**

| | true 1 | true 0 | class precision |
|---|---|---|---|
| pred. 1 | 130 | 63 | 67.36% |
| pred. 0 | 105 | 386 | 78.62% |
| class recall | 55.32% | 85.97% | |

4

Minimal leaf size




Maximal depth = 10

# Dataset

| Coluna | Usado na Classificação |
|---|---|
| id | id |
| job_category_Back-end Developer | |
| job_category_Data Scientist | Sim |
| job_category_DevOps / Sysadmin | |
| job_category_Front-end Developer | |
| job_category_Full-stack Developer | |
| job_category_outros | |
| job_category_Product/Project Management | |
| job_category_QA / Testing | |
| job_category_UX / UI Designer | |
| contractor | |
| consultancy | |
| permanent | Sim |

| Coluna | Usado na Classificação |
|---|---|
| relocation_paid | |
| visa_support | |
| dias_up | Sim |
| show_sal_rate | Sim |
| home_work | |
| education_req | |
| exp_nivel | |
| languagePT | Sim |
| regionPT | |
| full_time | |
| vist_dia | Sim |
| apps_dia | Sim |
| hands_dia | Sim |
| engaged | Label |

# Obrigada

# Make change happen