Crowfunding en el Mercado de Capitales Colombiano - Proyecto de Estadística para la Analítica de Datos

Sergio Castañeda, Guillermo Castillo, Alexander Guecha

18/05/2022

Justificación: el presente analisis estadístico propone estimar el monto a invertir por un inversionista, teniendo en cuenta las variables que componen la base con la información recopilada de su interacción con el producto.

Con la anterior información se puede ser mas especifico con el portafolio que se le ofrezca a un inversionistas actual y potencial, de cara a generar una diversificación del uso de las alternativas de inversión en personas que han tomado cercanía con la oferta de valor de a2censo.

knitr::opts\_chunk$set(echo = T)  
knitr::opts\_chunk$set(echo = T)  
library(readxl)  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(fBasics)

## Loading required package: timeDate

## Loading required package: timeSeries

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(corrgram)  
library(gclus)

## Loading required package: cluster

library(descr)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(agricolae)

##   
## Attaching package: 'agricolae'

## The following objects are masked from 'package:timeDate':  
##   
## kurtosis, skewness

library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:MASS':  
##   
## select

## The following object is masked from 'package:timeSeries':  
##   
## filter

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(fdth)

##   
## Attaching package: 'fdth'

## The following objects are masked from 'package:stats':  
##   
## sd, var

library(caTools)  
library(ROCR)  
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following object is masked from 'package:fdth':  
##   
## var

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:fBasics':  
##   
## densityPlot

## The following object is masked from 'package:dplyr':  
##   
## recode

library(foreign)  
library(apaTables)  
library(PerformanceAnalytics)

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following object is masked from 'package:timeSeries':  
##   
## time<-

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':  
##   
## first, last

##   
## Attaching package: 'PerformanceAnalytics'

## The following objects are masked from 'package:agricolae':  
##   
## kurtosis, skewness

## The following objects are masked from 'package:timeDate':  
##   
## kurtosis, skewness

## The following object is masked from 'package:graphics':  
##   
## legend

library(psych)

##   
## Attaching package: 'psych'

## The following object is masked from 'package:car':  
##   
## logit

## The following object is masked from 'package:fBasics':  
##   
## tr

## The following object is masked from 'package:timeSeries':  
##   
## outlier

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(corrr)  
library(igraph)

##   
## Attaching package: 'igraph'

## The following object is masked from 'package:plotly':  
##   
## groups

## The following object is masked from 'package:agricolae':  
##   
## similarity

## The following objects are masked from 'package:gclus':  
##   
## diameter, star

## The following objects are masked from 'package:dplyr':  
##   
## as\_data\_frame, groups, union

## The following objects are masked from 'package:stats':  
##   
## decompose, spectrum

## The following object is masked from 'package:base':  
##   
## union

library(corrgram)  
library(gclus)  
library(GGally)  
library(gvlma)

library(readxl)  
BDa2censo <- read\_excel("BD inversionistas depurada V6 anonimizada.xlsx")

#library(readxl)  
#BDa2censo <- read\_excel("maestria analitica de datos/Estadistica/Proyecto General/Base de datos #inversionistas depurada v3.xlsx",   
# col\_types = c("text", "text", "text",   
# "text", "text", "text", "text", "text",   
# "text", "text", "text", "numeric",   
# "numeric", "numeric", "text", "numeric",   
# "date", "numeric", "numeric", "numeric",   
# "numeric", "text", "text", "date",   
# "text", "text", "text", "numeric",   
# "text", "numeric"))  
#View(BDa2censo)

# ANÁLISIS ESTADÍSTICO BASE DE DATOS INVERSIONISTAS DE A2CENSO

## Resumen univariado de variables cualitativas

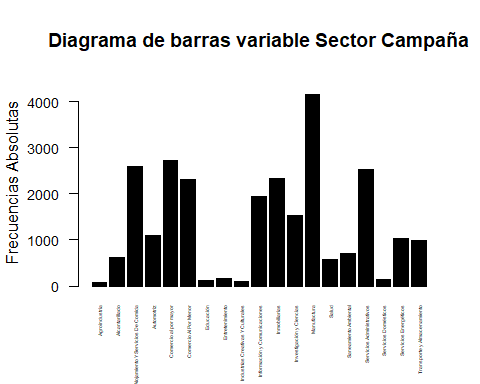
### Variable 1: SectorCampaña

Esta variable agrupa el conjunto de actividades productivas o comerciales que reúnen una serie de características similares en relación con una o más campañas, es decir, son negocios que cuentan con una naturaleza común. Mediante el siguiente análisis de frecuencias, se busca determinar cual sector de campaña cuenta con una mayor aceptación por parte de los inversionista

SectorCampaña\_= BDa2censo$SectorCampaña  
SCtabla=data.frame(table(SectorCampaña\_))  
porcentaje=prop.table(SCtabla[,2])  
SCtabla2= cbind(SCtabla, porcentaje)  
cum\_frequencia=cumsum(SCtabla2[,2])  
SCtabla3= cbind(SCtabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(SCtabla3[,3])  
SCtabla4= cbind(SCtabla3, cum\_porcentaje)  
knitr::kable(  
 (SCtabla4)  
)

| SectorCampaña\_ | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- |
| Agroindustria | 77 | 0.0029938 | 77 | 0.0029938 |
| Alcantarillado | 610 | 0.0237170 | 687 | 0.0267107 |
| Alojamiento Y Servicios De Comida | 2593 | 0.1008165 | 3280 | 0.1275272 |
| Automotriz | 1092 | 0.0424572 | 4372 | 0.1699844 |
| Comercio al por mayor | 2723 | 0.1058709 | 7095 | 0.2758554 |
| Comercio Al Por Menor | 2307 | 0.0896967 | 9402 | 0.3655521 |
| Educación | 125 | 0.0048600 | 9527 | 0.3704121 |
| Entretenimiento | 155 | 0.0060264 | 9682 | 0.3764386 |
| Industrías Creativas Y Culturales | 104 | 0.0040435 | 9786 | 0.3804821 |
| Información y Comunicaciones | 1941 | 0.0754666 | 11727 | 0.4559487 |
| Inmobiliarias | 2324 | 0.0903577 | 14051 | 0.5463064 |
| Investigación y Ciencias | 1531 | 0.0595257 | 15582 | 0.6058320 |
| Manufactura | 4166 | 0.1619751 | 19748 | 0.7678072 |
| Salud | 569 | 0.0221229 | 20317 | 0.7899300 |
| Saneamiento Ambiental | 714 | 0.0277605 | 21031 | 0.8176905 |
| Servicios Administrativos | 2532 | 0.0984448 | 23563 | 0.9161353 |
| Servicios Domésticos | 142 | 0.0055210 | 23705 | 0.9216563 |
| Servicios Energéticos | 1033 | 0.0401633 | 24738 | 0.9618196 |
| Transporte y Almacenamiento | 982 | 0.0381804 | 25720 | 1.0000000 |

barplot(table(BDa2censo$SectorCampaña), col = "black", main="Diagrama de barras variable Sector Campaña",ylim=c(0,4500),las=2, cex.names = 0.32, lwd = 1.5, cex.axis=0.9 , ylab = "Frecuencias Absolutas")



t <- list(  
family = "Ariel",  
size = 7,  
color = 'black')  
  
#plot\_ly(SCtabla4, labels = SCtabla4[,1], values = SCtabla4[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores por Sector Campaña",autosize = T, font = t)

Podemos evidenciar que las campañas de inversión que mayor tracción de inversionistas tuvieron, se encontraban asociadas con los sectores económicos de Manufactura (n= 4166), Comercio al por mayor (n= 2724)y Alojamiento y servicios de comida (n = 2593), por otro lado las campañas asociadas a sectores económicos de:Agroindustria (n=77), industrias creativas (n=104) y educación (n=124)fueron los 3 sectores económicos que menor tracción de numero de inversionistas tuvieron.

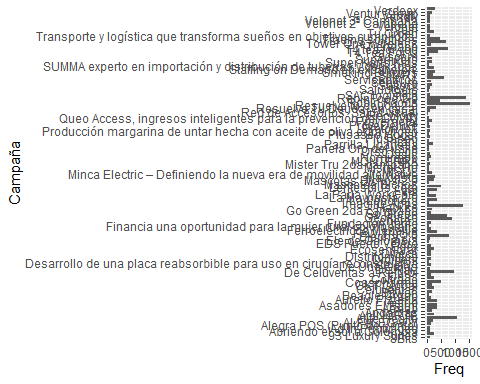
### Variable 2: Campaña

Esta variable está relacionada con las empresas o compañías que solicitan financiación para la ejecución de proyectos dentro de un sector económico en particular. Con la siguiente exploración se busca determinar la frecuencia de los lanzamientos de financiación que realizan cada una de las campañas inscritas al crowfunding de a2censo.

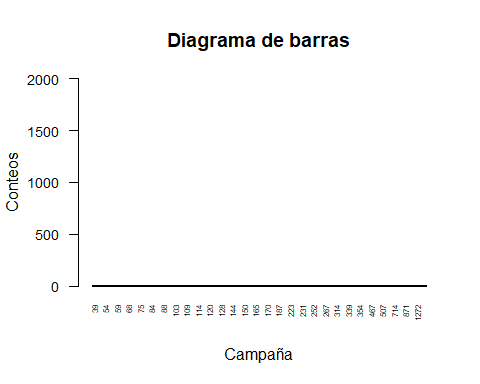
Campaña= BDa2censo$Campaña  
CPtabla=data.frame(table(Campaña))  
porcentaje=prop.table(CPtabla[,2])  
CPtabla2= cbind(CPtabla, porcentaje)  
cum\_frequencia=cumsum(CPtabla2[,2])  
CPtabla3= cbind(CPtabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(CPtabla3[,3])  
CPtabla4= cbind(CPtabla3, cum\_porcentaje)  
CPtabla4o <- CPtabla4[order(-CPtabla4$Freq),]  
knitr::kable(  
 (CPtabla4o)  
)

|  | Campaña | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- | --- |
| 72 | Resuelve tu deuda 2.0 | 1531 | 0.0595257 | 17802 | 0.6921462 |
| 74 | RobinFood 2.0 | 1383 | 0.0537714 | 19656 | 0.7642302 |
| 41 | Habi.co | 1272 | 0.0494557 | 11866 | 0.4613530 |
| 7 | Alife Health | 1062 | 0.0412908 | 1932 | 0.0751166 |
| 21 | De Celuventas a Refurbi | 947 | 0.0368196 | 6023 | 0.2341757 |
| 37 | Genersa | 871 | 0.0338647 | 9583 | 0.3725894 |
| 32 | Elepha 2.0 | 777 | 0.0302100 | 8034 | 0.3123639 |
| 89 | T4 tea for you | 742 | 0.0288491 | 23167 | 0.9007387 |
| 38 | Geofuturo | 714 | 0.0277605 | 10297 | 0.4003499 |
| 91 | Tower One Wireless | 674 | 0.0262053 | 23990 | 0.9327372 |
| 80 | Servisépticos | 610 | 0.0237170 | 20879 | 0.8117807 |
| 18 | Coaspharma | 507 | 0.0197123 | 4724 | 0.1836703 |
| 47 | Lentesplus.com | 502 | 0.0195179 | 13508 | 0.5251944 |
| 73 | Robin Foods | 471 | 0.0183126 | 18273 | 0.7104588 |
| 92 | Trading solutions | 467 | 0.0181571 | 24457 | 0.9508942 |
| 88 | T4 Tea For U | 370 | 0.0143857 | 22425 | 0.8718896 |
| 10 | Aoxlab | 366 | 0.0142302 | 2733 | 0.1062597 |
| 33 | FAW Trucks | 354 | 0.0137636 | 8388 | 0.3261275 |
| 11 | Asadores El Barril | 348 | 0.0135303 | 3081 | 0.1197900 |
| 46 | Leal | 342 | 0.0132970 | 13006 | 0.5056765 |
| 43 | La Lonchera | 339 | 0.0131804 | 12262 | 0.4767496 |
| 12 | Asfrio | 333 | 0.0129471 | 3414 | 0.1327372 |
| 44 | La Mayorista 2.0 | 329 | 0.0127916 | 12591 | 0.4895412 |
| 71 | Resuelve Tu Deuda en casa. | 314 | 0.0122084 | 16271 | 0.6326205 |
| 82 | Smoking Burgers | 275 | 0.0106921 | 21351 | 0.8301322 |
| 101 | Verdeex | 274 | 0.0106532 | 25720 | 1.0000000 |
| 50 | Mayorista | 267 | 0.0103810 | 14061 | 0.5466952 |
| 8 | ALINORTE | 259 | 0.0100700 | 2191 | 0.0851866 |
| 28 | Ecosembrar | 259 | 0.0100700 | 6985 | 0.2715785 |
| 16 | Celuventas | 257 | 0.0099922 | 4030 | 0.1566874 |
| 83 | Staffing on Demand en expansión. | 252 | 0.0097978 | 21603 | 0.8399300 |
| 94 | Tu Orden | 233 | 0.0090591 | 24783 | 0.9635692 |
| 49 | Mascotas Bichos 2.0 | 232 | 0.0090202 | 13794 | 0.5363142 |
| 19 | Coltrade | 231 | 0.0089813 | 4955 | 0.1926516 |
| 2 | 93 Luxury Suites | 228 | 0.0088647 | 337 | 0.0131026 |
| 60 | Parrilla Libanesa | 224 | 0.0087092 | 15120 | 0.5878694 |
| 6 | ALFALEGACY | 223 | 0.0086703 | 870 | 0.0338258 |
| 81 | SMART | 197 | 0.0076594 | 21076 | 0.8194401 |
| 78 | Satlock | 189 | 0.0073484 | 20125 | 0.7824650 |
| 17 | CKT Global | 187 | 0.0072706 | 4217 | 0.1639580 |
| 40 | Go Green 2da campaña | 178 | 0.0069207 | 10594 | 0.4118974 |
| 9 | Andamas | 176 | 0.0068429 | 2367 | 0.0920295 |
| 51 | Minca Electric – Definiendo la nueva era de movilidad alternativa | 170 | 0.0066096 | 14231 | 0.5533048 |
| 15 | Bitwan | 169 | 0.0065708 | 3773 | 0.1466952 |
| 100 | Ventur Group | 167 | 0.0064930 | 25446 | 0.9893468 |
| 86 | Superlikers | 165 | 0.0064152 | 21998 | 0.8552877 |
| 67 | Proyemetal | 164 | 0.0063764 | 15765 | 0.6129471 |
| 85 | Super Karts Kids | 155 | 0.0060264 | 21833 | 0.8488725 |
| 23 | De Una Grúas | 150 | 0.0058320 | 6257 | 0.2432737 |
| 90 | Tiendapp | 149 | 0.0057932 | 23316 | 0.9065319 |
| 26 | Distritornillos | 146 | 0.0056765 | 6598 | 0.2565319 |
| 79 | Sensum | 144 | 0.0055988 | 20269 | 0.7880638 |
| 34 | Ferroeléctricos Medellín | 139 | 0.0054044 | 8527 | 0.3315319 |
| 56 | Nominapp | 139 | 0.0054044 | 14683 | 0.5708787 |
| 25 | Digident | 136 | 0.0052877 | 6452 | 0.2508554 |
| 27 | Diveco | 128 | 0.0049767 | 6726 | 0.2615086 |
| 29 | Edex | 125 | 0.0048600 | 7110 | 0.2764386 |
| 20 | Coss | 121 | 0.0047045 | 5076 | 0.1973561 |
| 36 | Fundación Junfe | 121 | 0.0047045 | 8712 | 0.3387247 |
| 13 | Aurelio Pizzería | 120 | 0.0046656 | 3534 | 0.1374028 |
| 5 | Alegra POS (Punto de Venta) | 119 | 0.0046267 | 647 | 0.0251555 |
| 39 | Go Green | 119 | 0.0046267 | 10416 | 0.4049767 |
| 64 | Producción margarina de untar hecha con aceite de oliva extra virgen | 118 | 0.0045879 | 15446 | 0.6005443 |
| 54 | Mister Tru 2da campaña | 114 | 0.0044323 | 14486 | 0.5632193 |
| 98 | Velonet 3ª Campaña | 113 | 0.0043935 | 25196 | 0.9796267 |
| 75 | SAC logística | 112 | 0.0043546 | 19768 | 0.7685848 |
| 1 | 8Bits | 109 | 0.0042379 | 109 | 0.0042379 |
| 77 | Saludtools | 109 | 0.0042379 | 19936 | 0.7751166 |
| 96 | Velonet | 109 | 0.0042379 | 24975 | 0.9710342 |
| 97 | Velonet 2ª Campaña | 108 | 0.0041991 | 25083 | 0.9752333 |
| 4 | Ad In Publicidad | 104 | 0.0040435 | 528 | 0.0205288 |
| 65 | Proton IoT | 103 | 0.0040047 | 15549 | 0.6045490 |
| 93 | Transporte y logística que transforma sueños en objetivos cumplidos. | 93 | 0.0036159 | 24550 | 0.9545101 |
| 62 | PlusAseo | 91 | 0.0035381 | 15277 | 0.5939736 |
| 31 | Ele de Colombia | 88 | 0.0034215 | 7257 | 0.2821540 |
| 3 | Abriendo el sol a Colombia | 87 | 0.0033826 | 424 | 0.0164852 |
| 70 | Red de Accesorios. Salud visual | 85 | 0.0033048 | 15957 | 0.6204121 |
| 22 | De Raíz | 84 | 0.0032659 | 6107 | 0.2374417 |
| 57 | Oasis | 84 | 0.0032659 | 14767 | 0.5741446 |
| 95 | Vain | 83 | 0.0032271 | 24866 | 0.9667963 |
| 99 | Vendty | 83 | 0.0032271 | 25279 | 0.9828538 |
| 59 | Panela Oro del Valle | 77 | 0.0029938 | 14896 | 0.5791602 |
| 52 | MISOL | 75 | 0.0029160 | 14306 | 0.5562208 |
| 84 | SUMMA experto en importación y distribución de tuberías y derivados | 75 | 0.0029160 | 21678 | 0.8428460 |
| 45 | La Parla WorkCafé | 73 | 0.0028383 | 12664 | 0.4923795 |
| 14 | Beagle Group | 70 | 0.0027216 | 3604 | 0.1401244 |
| 68 | Queo Access, ingresos inteligentes para la prevención del COVID | 68 | 0.0026439 | 15833 | 0.6155910 |
| 53 | Mister Tru | 66 | 0.0025661 | 14372 | 0.5587869 |
| 61 | Pirani | 66 | 0.0025661 | 15186 | 0.5904355 |
| 35 | Financia una oportunidad para la mujer rural colombiana | 64 | 0.0024883 | 8591 | 0.3340202 |
| 24 | Desarrollo de una placa reabsorbible para uso en cirugía reconstructiva | 59 | 0.0022939 | 6316 | 0.2455677 |
| 30 | EDS Puerto Velero | 59 | 0.0022939 | 7169 | 0.2787325 |
| 76 | Sajú | 59 | 0.0022939 | 19827 | 0.7708787 |
| 55 | MONTERRA | 58 | 0.0022551 | 14544 | 0.5654743 |
| 42 | Imagine Apps | 57 | 0.0022162 | 11923 | 0.4635692 |
| 87 | Syscom | 57 | 0.0022162 | 22055 | 0.8575039 |
| 48 | Mascotas Bichos | 54 | 0.0020995 | 13562 | 0.5272939 |
| 58 | Orso to go | 52 | 0.0020218 | 14819 | 0.5761664 |
| 66 | ProveDeluxe | 52 | 0.0020218 | 15601 | 0.6065708 |
| 63 | Plusaseo Hogar | 51 | 0.0019829 | 15328 | 0.5959565 |
| 69 | Recology | 39 | 0.0015163 | 15872 | 0.6171073 |

ggplot(CPtabla4o, aes(x=Campaña, y=Freq)) +   
 geom\_bar(stat = "identity") + coord\_flip()



barplot(table(CPtabla4o$Freq), col = "blue", main="Diagrama de barras",ylim=c(0,2000),las=2, cex.names = 0.5, lwd = 1.5, cex.axis=0.9, xlab="Campaña", ylab = "Conteos")



t <- list(  
family = "Ariel",  
size = 7,  
color = 'black')  
  
#plot\_ly(CPtabla4, labels = CPtabla4o[,1], values = CPtabla4o[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores por Campaña", font = t)

Si bien esta variable no es tan relevante para el analisis de la investigación puesto que la cantidad de campañas y empresas financiadas es muy amplia y cada uno tiene un comportamiento específico, podemos encontrar tan solo hay 4 campañas que lograron reunir más de 1000 inversionistas, entre ellas se encuentra a) Resuelve tu deuda 2.0(n=1531), b) RobinFood 2.0 (n=1383), c) Habi.co, (n=1135) y d) Alife Health, (n=1062) y por tanto podríamos concluír que son las 4 campañas con mayor tracción, aspectos que pueden estar relacionados con la tasa ofertada, y otras variables no medidas como la inversion en publicidad que hicieron estas compañías de su campaña

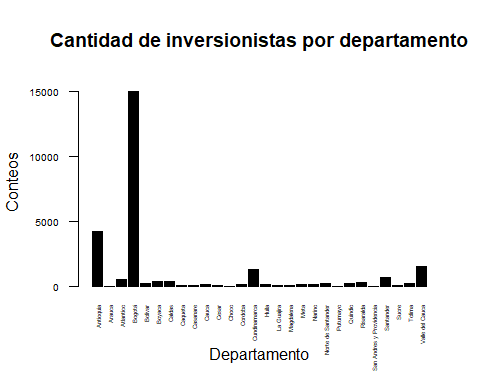
### Variable 3: Departamento

Con esta variable se buscar conocer el grado de participación de los inversionistas así como de las campañas en el territorio colombiano.

Departamento= BDa2censo$Departamento  
Dotabla=data.frame(table(Departamento))  
porcentaje=prop.table(Dotabla[,2])  
Dotabla2= cbind(Dotabla, porcentaje)  
cum\_frequencia=cumsum(Dotabla2[,2])  
Dotabla3= cbind(Dotabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(Dotabla3[,3])  
Dotabla4= cbind(Dotabla3, cum\_porcentaje)  
Dotabla4o <- Dotabla4[order(-Dotabla4$Freq),]  
knitr::kable(  
 (Dotabla4o)  
)

|  | Departamento | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- | --- |
| 4 | Bogotá | 15026 | 0.5842146 | 19715 | 0.7665241 |
| 1 | Antioquia | 4207 | 0.1635692 | 4207 | 0.1635692 |
| 28 | Valle del Cauca | 1534 | 0.0596423 | 25720 | 1.0000000 |
| 14 | Cundinamarca | 1260 | 0.0489891 | 22207 | 0.8634137 |
| 25 | Santander | 688 | 0.0267496 | 23957 | 0.9314541 |
| 3 | Atlantico | 477 | 0.0185459 | 4689 | 0.1823095 |
| 7 | Caldas | 346 | 0.0134526 | 20589 | 0.8005054 |
| 6 | Boyaca | 323 | 0.0125583 | 20243 | 0.7870529 |
| 23 | Risaralda | 295 | 0.0114697 | 23266 | 0.9045879 |
| 22 | Quindio | 230 | 0.0089425 | 22971 | 0.8931182 |
| 5 | Bolivar | 205 | 0.0079705 | 19920 | 0.7744946 |
| 27 | Tolima | 182 | 0.0070762 | 24186 | 0.9403577 |
| 20 | Norte de Santander | 164 | 0.0063764 | 22736 | 0.8839813 |
| 19 | Narino | 114 | 0.0044323 | 22572 | 0.8776050 |
| 15 | Huila | 113 | 0.0043935 | 22320 | 0.8678072 |
| 10 | Cauca | 102 | 0.0039658 | 20771 | 0.8075816 |
| 18 | Meta | 100 | 0.0038880 | 22458 | 0.8731726 |
| 13 | Cordoba | 97 | 0.0037714 | 20947 | 0.8144246 |
| 11 | Cesar | 78 | 0.0030327 | 20849 | 0.8106143 |
| 9 | Casanare | 55 | 0.0021384 | 20669 | 0.8036159 |
| 26 | Sucre | 47 | 0.0018274 | 24004 | 0.9332815 |
| 17 | Magdalena | 30 | 0.0011664 | 22358 | 0.8692846 |
| 8 | Caqueta | 25 | 0.0009720 | 20614 | 0.8014774 |
| 16 | La Guajira | 8 | 0.0003110 | 22328 | 0.8681182 |
| 2 | Arauca | 5 | 0.0001944 | 4212 | 0.1637636 |
| 21 | Putumayo | 5 | 0.0001944 | 22741 | 0.8841757 |
| 24 | San Andres y Providencia | 3 | 0.0001166 | 23269 | 0.9047045 |
| 12 | Choco | 1 | 0.0000389 | 20850 | 0.8106532 |

barplot(table(BDa2censo$Departamento), col = "black", main="Cantidad de inversionistas por departamento",ylim=c(0,16000),las=2, cex.names = 0.38, lwd = 1, cex.axis=0.6, xlab="Departamento", ylab = "Conteos")



t <- list(  
family = "Ariel",  
size = 7,  
color = 'black')  
  
#plot\_ly(Dotabla4, labels = Dotabla4[,1], values = Dotabla4[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores por Departamento", font = t)

De otro lado, si bien es esperado que la mayor participación de inversionistas se encuentren ubicados en Bogotá (n= 15032), resulta llamativo observar las participaciones que se tienen en otros departamentos como Antioquia (n= 4207) , Valle del cauca (n= 1534), cundinamarca (n= 1260) santander (n= 688) y atlantico (n= 477).

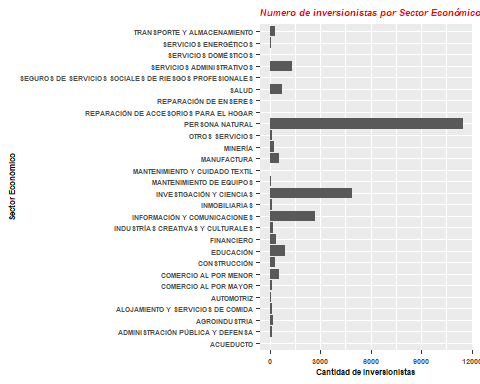
### Variable 4: Sector Económico del inversionista

Esta variable categórica es muy relevante para nuestro análisis estadístico, toda vez que nos permite conocer de primera mano los sectores económicos de los inversionistas interesados en las campañas publicitarias que ofrece a2censo, así como la capacidad de inversión para financiar los distintos proyectos que ofrece la compañía.

SectorEconomico= BDa2censo$SectorEconomico  
SEtabla=data.frame(table(SectorEconomico))  
porcentaje=prop.table(SEtabla[,2])  
SEtabla2= cbind(SEtabla, porcentaje)  
cum\_frequencia=cumsum(SEtabla2[,2])  
SEtabla3= cbind(SEtabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(SEtabla3[,3])  
SEtabla4= cbind(SEtabla3, cum\_porcentaje)  
SEtabla4o <- SEtabla4[order(-SEtabla4$Freq),]  
knitr::kable(  
 (SEtabla4o)  
)

|  | SectorEconomico | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- | --- |
| 20 | PERSONA NATURAL | 11467 | 0.4458398 | 23261 | 0.9043935 |
| 14 | INVESTIGACIÓN Y CIENCIAS | 4893 | 0.1902411 | 10784 | 0.4192846 |
| 12 | INFORMACIÓN Y COMUNICACIONES | 2672 | 0.1038880 | 5729 | 0.2227449 |
| 25 | SERVICIOS ADMINISTRATIVOS | 1337 | 0.0519829 | 25314 | 0.9842146 |
| 9 | EDUCACIÓN | 938 | 0.0364697 | 2484 | 0.0965785 |
| 23 | SALUD | 711 | 0.0276439 | 23976 | 0.9321928 |
| 7 | COMERCIO AL POR MENOR | 583 | 0.0226672 | 1236 | 0.0480560 |
| 17 | MANUFACTURA | 552 | 0.0214619 | 11398 | 0.4431571 |
| 10 | FINANCIERO | 403 | 0.0156687 | 2887 | 0.1122473 |
| 28 | TRANSPORTE Y ALMACENAMIENTO | 318 | 0.0123639 | 25720 | 1.0000000 |
| 8 | CONSTRUCCIÓN | 310 | 0.0120529 | 1546 | 0.0601089 |
| 18 | MINERÍA | 277 | 0.0107698 | 11675 | 0.4539269 |
| 3 | AGROINDUSTRIA | 170 | 0.0066096 | 325 | 0.0126361 |
| 11 | INDUSTRÍAS CREATIVAS Y CULTURALES | 170 | 0.0066096 | 3057 | 0.1188569 |
| 13 | INMOBILIARIAS | 162 | 0.0062986 | 5891 | 0.2290435 |
| 6 | COMERCIO AL POR MAYOR | 157 | 0.0061042 | 653 | 0.0253888 |
| 2 | ADMINISTRACIÓN PÚBLICA Y DEFENSA | 148 | 0.0057543 | 155 | 0.0060264 |
| 19 | OTROS SERVICIOS | 119 | 0.0046267 | 11794 | 0.4585537 |
| 4 | ALOJAMIENTO Y SERVICIOS DE COMIDA | 114 | 0.0044323 | 439 | 0.0170684 |
| 27 | SERVICIOS ENERGÉTICOS | 66 | 0.0025661 | 25402 | 0.9876361 |
| 5 | AUTOMOTRIZ | 57 | 0.0022162 | 496 | 0.0192846 |
| 15 | MANTENIMIENTO DE EQUIPOS | 57 | 0.0022162 | 10841 | 0.4215008 |
| 26 | SERVICIOS DOMÉSTICOS | 22 | 0.0008554 | 25336 | 0.9850700 |
| 1 | ACUEDUCTO | 7 | 0.0002722 | 7 | 0.0002722 |
| 16 | MANTENIMIENTO Y CUIDADO TEXTIL | 5 | 0.0001944 | 10846 | 0.4216952 |
| 21 | REPARACIÓN DE ACCESORIOS PARA EL HOGAR | 2 | 0.0000778 | 23263 | 0.9044712 |
| 22 | REPARACIÓN DE ENSERES | 2 | 0.0000778 | 23265 | 0.9045490 |
| 24 | SEGUROS DE SERVICIOS SOCIALES DE RIESGOS PROFESIONALES | 1 | 0.0000389 | 23977 | 0.9322317 |

ggplot(SEtabla4o, aes(x=SectorEconomico, y=Freq)) +   
 geom\_bar(stat = "identity") + ggtitle("Numero de inversionistas por Sector Económico") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic")) +   
 xlab("Sector Económico")+ ylab("Cantidad de inversionistas") + coord\_flip()+theme(axis.text=element\_text(size=5,face="bold"),  
 axis.title=element\_text(size=6,face="bold"))



t <- list(  
family = "Ariel",  
size = 5,  
color = 'black')  
  
#plot\_ly(SEtabla4, labels = SEtabla4[,1], values = SEtabla4[,3], type = "pie",autosize = T) %>%  
# layout(title = "Diagrama de sectores por Sector Económico del Inversionista",autosize = T, font=t)#, width = 450, height = #450, margin = m)

La mayor proporsion de inversionistas autodenominan su actividad económica como persona natural (n= 11471), siendo que es la opcion mas entendible para usuarios que no conocen estas clasificaciones económicas, no obtante resulta llamativo observar la participacion de personas que se dedican a actividades de investigación y ciencias (n= 4893), información y comunicaciones (n= 2672) y servicios administrativos (n= 1337) en donde podemos encontrar posiblemente un proporsion de inversionistas asalariados que usan sus excedentes de dinero en inversiones en a2censo.

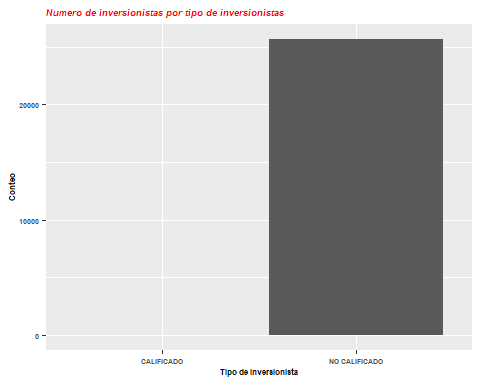
### Variable 5: Tipo Inversionista

Esta variable categórica es usada por a2censo para segmentar la población que aplica a las distintas campañas de inversión en función del interés y nivel de experticia en los actuales ecosistemas de inversión y financiación.

TipoInversionista= BDa2censo$TipoInversionista  
TItabla=data.frame(table(TipoInversionista))  
porcentaje=prop.table(TItabla[,2])  
TItabla2= cbind(TItabla, porcentaje)  
cum\_frequencia=cumsum(TItabla2[,2])  
TItabla3= cbind(TItabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(TItabla3[,3])  
TItabla4= cbind(TItabla3, cum\_porcentaje)  
TItabla4 <- TItabla4[order(-TItabla4$Freq),]  
knitr::kable(  
 (TItabla4)  
)

|  | TipoInversionista | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- | --- |
| 2 | NO CALIFICADO | 25674 | 0.9982115 | 25720 | 1.0000000 |
| 1 | CALIFICADO | 46 | 0.0017885 | 46 | 0.0017885 |

ggplot(TItabla4, aes(x=TipoInversionista, y=Freq)) +   
 geom\_bar(stat = "identity") + ggtitle("Numero de inversionistas por tipo de inversionistas") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic")) +   
 xlab("Tipo de inversionista")+ ylab("Conteo") +theme(axis.text=element\_text(size=5,face="bold"),  
 axis.title=element\_text(size=6,face="bold"))



t <- list(  
family = "Ariel",  
size = 10,  
color = 'black')  
  
#plot\_ly(TItabla4, labels = TItabla4[,1], values = TItabla4[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores por Tipo Inversionista", font = t)

Por otro lado, podemos encontrar que la proporsion de inversionistas que se autodenominan inversionistas calificados y que han invertido en a2censo es muy baja (n= 46), respecto a los inversionistas no calificados (n= 25680) representando una oportunidad de abordar más a este nicho de inversionistas calificados con este producto y hacer un mayor detalle a su comportamiento de inversión.

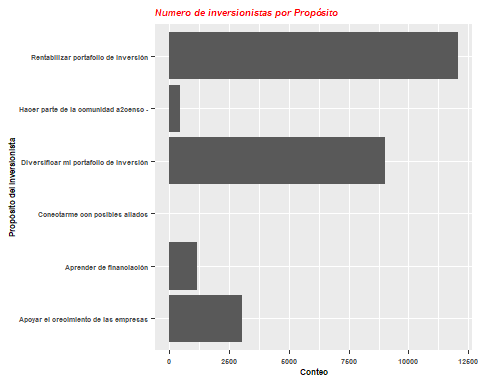
### Variable 6: Propósito del inversionista en A2censo

Mediante esta variable categórica, a2censo busca conocer la finalidad o el objetivo de inversión de cada uno de los inversionistas, así como la correlación que pueda existir entre estos objetivos y las campañas realizadas, para brindar mejores opciones por parte de la compañía y por consiguiente acrecentar la motivación y confianza de los clientes.

Proposito = BDa2censo$Propósito  
PTtabla=data.frame(table(Proposito))  
porcentaje=prop.table(PTtabla[,2])  
PTtabla2= cbind(PTtabla, porcentaje)  
cum\_frequencia=cumsum(PTtabla2[,2])  
PTtabla3= cbind(PTtabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(PTtabla3[,3])  
PTtabla4= cbind(PTtabla3, cum\_porcentaje)  
PTtabla4 <- PTtabla4[order(-PTtabla4$Freq),]  
knitr::kable(  
 (PTtabla4)  
)

|  | Proposito | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- | --- |
| 6 | Rentabilizar portafolio de inversión | 12063 | 0.4690124 | 25720 | 1.0000000 |
| 4 | Diversificar mi portafolio de inversión | 9032 | 0.3511664 | 13198 | 0.5131415 |
| 1 | Apoyar el crecimiento de las empresas | 3027 | 0.1176905 | 3027 | 0.1176905 |
| 2 | Aprender de financiación | 1138 | 0.0442457 | 4165 | 0.1619362 |
| 5 | Hacer parte de la comunidad a2censo - | 459 | 0.0178460 | 13657 | 0.5309876 |
| 3 | Conectarme con posibles aliados | 1 | 0.0000389 | 4166 | 0.1619751 |

ggplot(PTtabla4, aes(x=Proposito, y=Freq)) +   
 geom\_bar(stat = "identity") + ggtitle("Numero de inversionistas por Propósito") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic")) + coord\_flip() +  
 xlab("Propósito del inversionista")+ ylab("Conteo") +theme(axis.text=element\_text(size=5,face="bold"),  
 axis.title=element\_text(size=6,face="bold"))



t <- list(  
family = "Ariel",  
size = 10,  
color = 'black')  
  
  
#plot\_ly(PTtabla4, labels = PTtabla4[,1], values = PTtabla4[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores por propósito del inversionista en A2censo", font = t)

De acuerdo a los anteriores analisis podemos evidenciar como la mayoria de inversionistas actuales siguen propósitos de inversión asociados al crecimiento y rentabilidad del dinero y el portafolio de inversión actual (n = 12066); y por otro lado a la diversificación del portafolio de inversion ( n= 9034). No obstante podemos encontrar que hay un 11.8% de la muestra que tienen fines filantrópicos de apoyar el emprendimiento Colombiano (n = 3027).

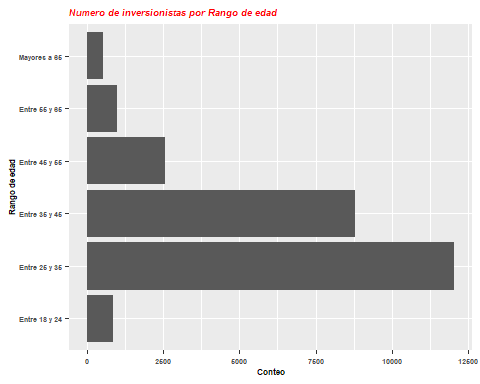
### Variable 7: Grupo de Edad

Mediante esta variable categórica, a2censo realizad una segmentación de los clientes por grupo etario, el cual brinda información de gran importancia a la compañía al momento de lanzar un proyecto de inversión, toda vez que le permite predecir de antemano la población objetivo en función de la edad.

GrupoEdad = BDa2censo$GrupoEdad  
GEtabla=data.frame(table(GrupoEdad))  
porcentaje=prop.table(GEtabla[,2])  
GEtabla2= cbind(GEtabla, porcentaje)  
cum\_frequencia=cumsum(GEtabla2[,2])  
GEtabla3= cbind(GEtabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(GEtabla3[,3])  
GEtabla4= cbind(GEtabla3, cum\_porcentaje)  
GEtabla4 <- GEtabla4[order(-GEtabla4$Freq),]  
knitr::kable(  
 (GEtabla4)  
)

|  | GrupoEdad | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- | --- |
| 2 | Entre 25 y 35 | 12048 | 0.4684292 | 12886 | 0.5010109 |
| 3 | Entre 35 y 45 | 8786 | 0.3416019 | 21672 | 0.8426128 |
| 4 | Entre 45 y 55 | 2556 | 0.0993779 | 24228 | 0.9419907 |
| 5 | Entre 55 y 65 | 970 | 0.0377138 | 25198 | 0.9797045 |
| 1 | Entre 18 y 24 | 838 | 0.0325816 | 838 | 0.0325816 |
| 6 | Mayores a 65 | 522 | 0.0202955 | 25720 | 1.0000000 |

ggplot(GEtabla4, aes(x=GrupoEdad, y=Freq)) +   
 geom\_bar(stat = "identity") + coord\_flip() + ggtitle("Titulo grafico") + ggtitle("Numero de inversionistas por Rango de edad") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic")) +  
 xlab("Rango de edad")+ ylab("Conteo") +theme(axis.text=element\_text(size=5,face="bold"),  
 axis.title=element\_text(size=6,face="bold"))



t <- list(  
family = "Ariel",  
size = 10,  
color = 'black')  
  
#plot\_ly(GEtabla4, labels = GEtabla4[,1], values = GEtabla4[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores por rango de Edad del inversionista", font = t)

De acuerdo al anterior analisis, resulta interesante observar la importante participación del 47% de inversionistas jovenes entre 25 y 35 años (n = 12049) y tambien de inversionistas entre 35 y 45 años (n = 8787). No obstante la participacion de inversionistas entre 45 y 55 años (9,9%, n = 2558) no es una participación despreciable dentro del grupo total de inversionistas.

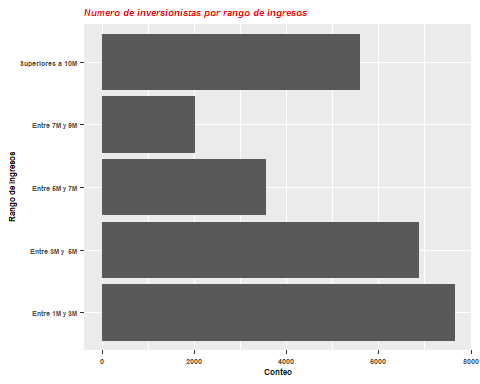
### Variable 8: Grupo de Ingresos

Mediante esta variable categórica se busca agrupar a los clientes inversionistas en razón al nivel de ingresos mensuales y de esta forma tener un proyección del capital disponible al momento de realizar una campaña en particular y, del grupo de interés al cual se puede brindar el lanzamiento del proyecto. De otra parte, para fines de este análisis se proyecta conocer si existe correlación de esta variable con la proyección del monto de inversión.

GrupoIngresos = BDa2censo$GrupoIngresos  
GItabla=data.frame(table(GrupoIngresos))  
porcentaje=prop.table(GItabla[,2])  
GItabla2= cbind(GItabla, porcentaje)  
cum\_frequencia=cumsum(GItabla2[,2])  
GItabla3= cbind(GItabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(GItabla3[,3])  
GItabla4= cbind(GItabla3, cum\_porcentaje)  
GItabla4 <- GItabla4[order(-GItabla4$Freq),]  
knitr::kable(  
 (GItabla4)  
)

|  | GrupoIngresos | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- | --- |
| 1 | Entre 1M y 3M | 7652 | 0.2975117 | 7652 | 0.2975117 |
| 2 | Entre 3M y 5M | 6873 | 0.2672240 | 14525 | 0.5647356 |
| 5 | Superiores a 10M | 5609 | 0.2180793 | 25720 | 1.0000000 |
| 3 | Entre 5M y 7M | 3563 | 0.1385303 | 18088 | 0.7032659 |
| 4 | Entre 7M y 9M | 2023 | 0.0786547 | 20111 | 0.7819207 |

ggplot(GItabla4, aes(x=GrupoIngresos, y=Freq)) +   
 geom\_bar(stat = "identity") + coord\_flip() + ggtitle("Numero de inversionistas por rango de ingresos") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic")) +  
 xlab("Rango de ingresos")+ ylab("Conteo") +theme(axis.text=element\_text(size=5,face="bold"),  
 axis.title=element\_text(size=6,face="bold"))



t <- list(  
family = "Ariel",  
size = 10,  
color = 'black')  
  
  
#plot\_ly(GItabla4, labels = GItabla4[,1], values = GItabla4[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores por rango de Ingresos", font = t)

Dentro de los inversionistas de a2censo encontramos inversionistas de tres grupos de rangos salariales predominantes, dentro de los cuales se encuentran en primer lugar una mayor proporsión de inversionistas de ingresos bajos (menor a 1 millon mensual; n= 7652), Inversionistas de ingresos medios - bajos (entre 3 y 5 millones mensuales; n= 6873), Inversionistas de ingresos altos (Superiores a 10 Millones mensuales; n = 5615).No obstante si agruparamos, podemos observar que mas del 50% de inversionistas suelen ser personas con ingresos inferiores a 5 millones de pesos.

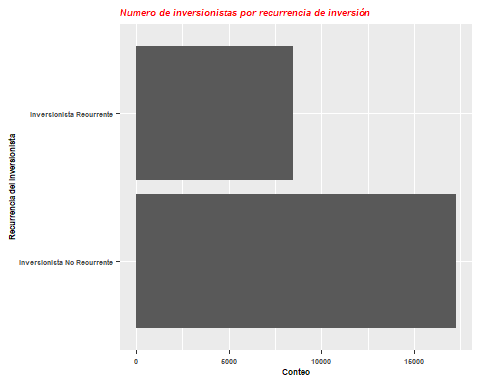
### Variable 9: Inversionistas recurrentes

Esta variable categórica está con la frecuencia de participación de cada uno de los inversionistas de a2censo en las campañas ofrecidas, el cual se categoriza a cada cliente entre frecuente o no, partiendo de un parámetro de medición previamente establecido por la compañía, ayudando en la toma de decisiones, toda vez que brinda información relevante para predecir el monto de inversión que puede llegar a tener una campaña en particular.

InvRecurr = BDa2censo$CodEsInversionistaRecurrente  
Rtabla=data.frame(table(InvRecurr))  
porcentaje=prop.table(Rtabla[,2])  
Rtabla2= cbind(Rtabla, porcentaje)  
cum\_frequencia=cumsum(Rtabla2[,2])  
Rtabla3= cbind(Rtabla2, cum\_frequencia)  
cum\_porcentaje=cumsum(Rtabla3[,3])  
Rtabla4= cbind(Rtabla3, cum\_porcentaje)  
Rtabla4 <- Rtabla4[order(-Rtabla4$Freq),]  
knitr::kable(  
 (Rtabla4)  
)

| InvRecurr | Freq | porcentaje | cum\_frequencia | cum\_porcentaje |
| --- | --- | --- | --- | --- |
| Inversionista No Recurrente | 17274 | 0.6716174 | 17274 | 0.6716174 |
| Inversionista Recurrente | 8446 | 0.3283826 | 25720 | 1.0000000 |

ggplot(Rtabla4, aes(x=InvRecurr, y=Freq)) +   
 geom\_bar(stat = "identity") + coord\_flip() + ggtitle("Numero de inversionistas por recurrencia de inversión") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic")) +  
 xlab("Recurrencia del inversionista")+ ylab("Conteo") +theme(axis.text=element\_text(size=5,face="bold"),  
 axis.title=element\_text(size=6,face="bold"))



t <- list(  
family = "Ariel",  
size = 10,  
color = 'black')  
  
  
  
#plot\_ly(Rtabla4, labels = Rtabla4[,1], values = Rtabla4[,3], type = "pie") %>%  
# layout(title = "Diagrama de sectores ¿Es inversionista recurrente", font = t)

Podemos evidenciar que la mayoría (67%) de inversionistas no son recurrentes ( n = 17280), Sin embargo existe una alta proporsión de inversionistas recurrentes (33%; n= 8446) que pueden ser de interés en el analisis realizado

## Resumen univariado de variables cuantitativas

A continuación se presenta el analisis variable por variable en lo que respecta a información cuantitativa.

Vale la pena resaltar que dichos analisis se hicieron siguiendo la regla de establecimiento de clases de Sturges que incluye la funcion fdt del paquete fdth.

Este proceso se realizó de esta manera puesto que al realizar los analisis bajo un numero entre 3 y 7 intervalos de clase que fueron probados en cada una de las variables, la distribución de la información no permitía desagregarla en grupos que contribuyeran al analisis y que se representaran de mejor manera, en todo caso la distribución en clases bajo la regla de sturges permitió la mejor visualización de la información cuantitativa.

### Variable 10: Ingresos Mes

Mediante esta variable cuantitativa, a2cendo lleva el registro del ingreso mensual reportado por el inversionista en pesos colombianos. Para el presente análisis estadístico fue seleccionada esta variable, con el propósito de establecer el nivel de correlación que existe entre el nivel de ingresos mensuales de cada inversionista con la proyección del monto de inversión.

basicStats(BDa2censo$IngresosMes)

## X..BDa2censo.IngresosMes  
## nobs 2.572000e+04  
## NAs 0.000000e+00  
## Minimum 1.010000e+06  
## Maximum 1.540000e+08  
## 1. Quartile 3.000000e+06  
## 3. Quartile 8.000000e+06  
## Mean 7.590004e+06  
## Median 4.878000e+06  
## Sum 1.952149e+11  
## SE Mean 7.490181e+04  
## LCL Mean 7.443193e+06  
## UCL Mean 7.736816e+06  
## Variance 1.442964e+14  
## Stdev 1.201235e+07  
## Skewness 7.140657e+00  
## Kurtosis 6.439523e+01

IMtabla\_Sturges <- fdt(BDa2censo$IngresosMes,breaks="Sturges",right=F, na.rm=TRUE)  
  
  
knitr::kable(  
 (IMtabla\_Sturges)  
)

| Class limits | f | rf | rf(%) | cf | cf(%) |
| --- | --- | --- | --- | --- | --- |
| [999900,10658656.2) | 21389 | 0.8316096 | 83.1609642 | 21389 | 83.16096 |
| [10658656.2,20317412.5) | 3345 | 0.1300544 | 13.0054432 | 24734 | 96.16641 |
| [20317412.5,29976168.8) | 383 | 0.0148911 | 1.4891135 | 25117 | 97.65552 |
| [29976168.8,39634925) | 155 | 0.0060264 | 0.6026439 | 25272 | 98.25816 |
| [39634925,49293681.2) | 115 | 0.0044712 | 0.4471229 | 25387 | 98.70529 |
| [49293681.2,58952437.5) | 48 | 0.0018663 | 0.1866252 | 25435 | 98.89191 |
| [58952437.5,68611193.8) | 39 | 0.0015163 | 0.1516330 | 25474 | 99.04355 |
| [68611193.8,78269950) | 13 | 0.0005054 | 0.0505443 | 25487 | 99.09409 |
| [78269950,87928706.2) | 37 | 0.0014386 | 0.1438569 | 25524 | 99.23795 |
| [87928706.2,97587462.5) | 29 | 0.0011275 | 0.1127527 | 25553 | 99.35070 |
| [97587462.5,107246219) | 68 | 0.0026439 | 0.2643857 | 25621 | 99.61509 |
| [107246219,116904975) | 4 | 0.0001555 | 0.0155521 | 25625 | 99.63064 |
| [116904975,126563731) | 34 | 0.0013219 | 0.1321928 | 25659 | 99.76283 |
| [126563731,136222488) | 15 | 0.0005832 | 0.0583204 | 25674 | 99.82115 |
| [136222488,145881244) | 12 | 0.0004666 | 0.0466563 | 25686 | 99.86781 |
| [145881244,155540000) | 34 | 0.0013219 | 0.1321928 | 25720 | 100.00000 |

|  | x |
| --- | --- |
| start | 999900 |
| end | 155540000 |
| h | 9658756 |
| right | 0 |

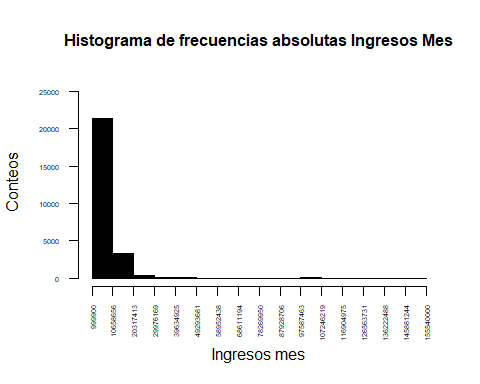
plot(IMtabla\_Sturges,type='fh',main="Histograma de frecuencias absolutas Ingresos Mes", cex.main = 1 ,las=2, cex.names = 0.35, lwd = 1, cex.axis=0.45,  
 xlab="Ingresos mes",ylab="Conteos",col="black", las=2)

## Warning in title(main = main, xlab = xlab, ylab = ylab, ...): "cex.names" is not  
## a graphical parameter

## Warning in axis(2, ...): "cex.names" is not a graphical parameter

## Warning in rect(brk[-length(brk)], 0, brk[-1], y, col = col, ...): "cex.names"  
## is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



plot(IMtabla\_Sturges,type='fp',main="Poligono de frecuencias absolutas Ingresos Mes",cex.main = 1, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab="Ingresos mes",ylab="Conteos",col="black",lwd=2,las=2)

## Warning in plot.window(...): "cex.names" is not a graphical parameter

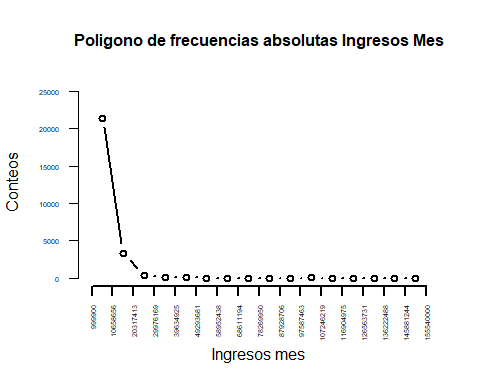
## Warning in plot.xy(xy, type, ...): "cex.names" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter

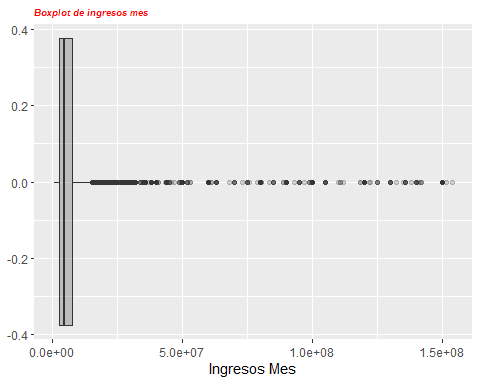
## Warning in box(...): "cex.names" is not a graphical parameter

## Warning in title(...): "cex.names" is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



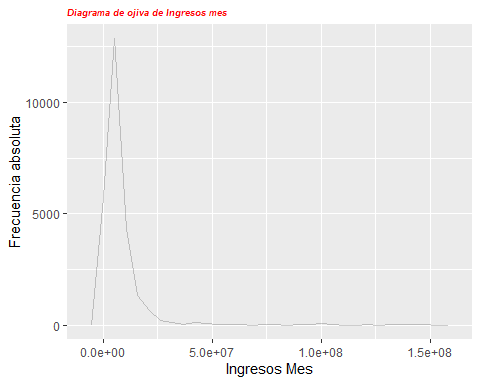
ggplot(BDa2censo, aes( y=IngresosMes)) +   
 geom\_boxplot(fill="black", alpha=0.2) +   
 ylab("Ingresos Mes") + coord\_flip() + ggtitle("Boxplot de ingresos mes") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))



ggplot(BDa2censo, aes( y=IngresosMes)) +   
 geom\_freqpoly(fill="black", alpha=0.2) +   
 ylab("Ingresos Mes")+xlab("Frecuencia absoluta") + coord\_flip() + ggtitle("Diagrama de ojiva de Ingresos mes") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Ignoring unknown parameters: fill

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 De acuerdo al analisis de frecuencia podemos encontrar que un total de 21.389 que corresponde al 83% de los inversionistas se agrupan en salarios entre $990. 000 y 10.658.656 pesos siendo en rango salarial predominante. No obstante el grupo de inversionistas con ingresos entre 10.658.656 y 20.317.412 de pesos representa el 13% (n = 3345) de la muestra de inversionsitas.

### Variable 11: Patrimonio

Mediante esta variable cuantitativa, a2cendo lleva el registro del valor en pesos colombianos de los activos que reporta tener el inversionista. Para el presente análisis estadístico fue seleccionada esta variable, con el propósito de establecer el nivel de correlación que existe entre el valor del patrimonio de cada inversionista con la proyección del monto de inversión.

basicStats(BDa2censo$Patrimonio)

## X..BDa2censo.Patrimonio  
## nobs 2.572000e+04  
## NAs 0.000000e+00  
## Minimum 1.000000e+06  
## Maximum 8.064210e+08  
## 1. Quartile 2.000000e+07  
## 3. Quartile 2.158680e+08  
## Mean 1.570082e+08  
## Median 1.000000e+08  
## Sum 4.038252e+12  
## SE Mean 1.120823e+06  
## LCL Mean 1.548114e+08  
## UCL Mean 1.592051e+08  
## Variance 3.231059e+16  
## Stdev 1.797515e+08  
## Skewness 1.613582e+00  
## Kurtosis 2.098108e+00

POtabla\_St <- fdt(BDa2censo$Patrimonio,breaks="Sturges",right=F)  
  
knitr::kable(  
 (POtabla\_St)  
)

| Class limits | f | rf | rf(%) | cf | cf(%) |
| --- | --- | --- | --- | --- | --- |
| [990000,51833450.6) | 9512 | 0.3698289 | 36.9828927 | 9512 | 36.98289 |
| [51833450.6,102676901) | 5510 | 0.2142302 | 21.4230171 | 15022 | 58.40591 |
| [102676901,153520352) | 2476 | 0.0962675 | 9.6267496 | 17498 | 68.03266 |
| [153520352,204363803) | 1597 | 0.0620918 | 6.2091757 | 19095 | 74.24184 |
| [204363803,255207253) | 1136 | 0.0441680 | 4.4167963 | 20231 | 78.65863 |
| [255207253,306050704) | 1381 | 0.0536936 | 5.3693624 | 21612 | 84.02799 |
| [306050704,356894154) | 650 | 0.0252722 | 2.5272162 | 22262 | 86.55521 |
| [356894154,407737605) | 873 | 0.0339425 | 3.3942457 | 23135 | 89.94946 |
| [407737605,458581056) | 403 | 0.0156687 | 1.5668740 | 23538 | 91.51633 |
| [458581056,509424506) | 521 | 0.0202566 | 2.0256610 | 24059 | 93.54199 |
| [509424506,560267957) | 315 | 0.0122473 | 1.2247278 | 24374 | 94.76672 |
| [560267957,611111408) | 389 | 0.0151244 | 1.5124417 | 24763 | 96.27916 |
| [611111408,661954858) | 245 | 0.0095257 | 0.9525661 | 25008 | 97.23173 |
| [661954858,712798309) | 271 | 0.0105365 | 1.0536547 | 25279 | 98.28538 |
| [712798309,763641759) | 190 | 0.0073872 | 0.7387247 | 25469 | 99.02411 |
| [763641759,814485210) | 251 | 0.0097589 | 0.9758942 | 25720 | 100.00000 |

|  | x |
| --- | --- |
| start | 990000 |
| end | 814485210 |
| h | 50843451 |
| right | 0 |

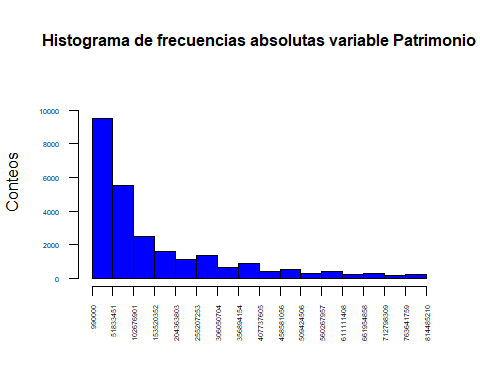
plot(POtabla\_St,type='fh',main="Histograma de frecuencias absolutas variable Patrimonio",cex.main = 1 ,las=2, cex.names = 0.35, lwd = 1, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="blue", las=2)

## Warning in title(main = main, xlab = xlab, ylab = ylab, ...): "cex.names" is not  
## a graphical parameter

## Warning in axis(2, ...): "cex.names" is not a graphical parameter

## Warning in rect(brk[-length(brk)], 0, brk[-1], y, col = col, ...): "cex.names"  
## is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



plot(POtabla\_St,type='fp',main="Poligono de frecuencias absolutas de Patrimonio", cex.main = 1, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="black",lwd=2,las=2)

## Warning in plot.window(...): "cex.names" is not a graphical parameter

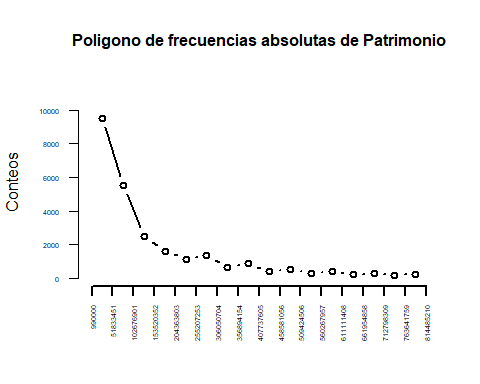
## Warning in plot.xy(xy, type, ...): "cex.names" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter

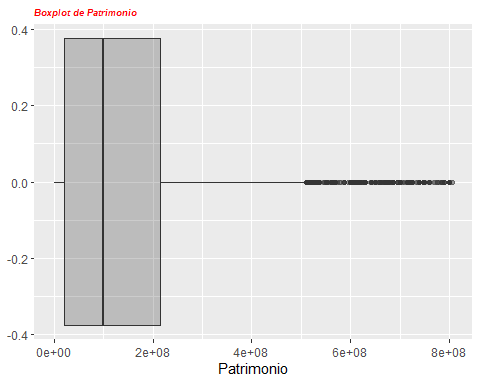
## Warning in box(...): "cex.names" is not a graphical parameter

## Warning in title(...): "cex.names" is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



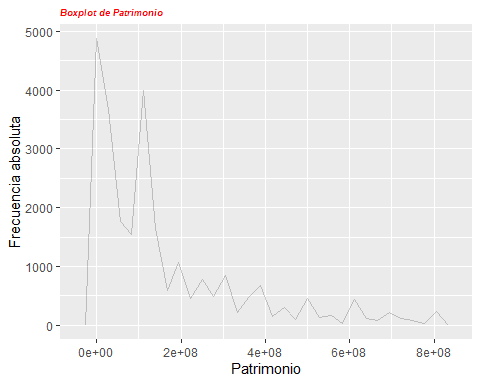
ggplot(BDa2censo, aes( y=Patrimonio)) +   
 geom\_boxplot(fill="black", alpha=0.2) +   
 ylab("Patrimonio") + coord\_flip() + ggtitle("Boxplot de Patrimonio") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))



ggplot(BDa2censo, aes( y=Patrimonio)) +   
 geom\_freqpoly(fill="black", alpha=0.2) +   
 ylab("Patrimonio")+xlab("Frecuencia absoluta") + coord\_flip() + ggtitle("Boxplot de Patrimonio") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Ignoring unknown parameters: fill

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 Por otro lado en lo que respecta al patrimonio de los inversionistas podemos encontrar que el 36% de los inversionistas que representan un total de 9.514 individuos tienen patrimonios entre 990.000 y 51.833.450. Adicionalmente podemos encontrar que el 21% (n= 5510) de los inversionistas tienen patrimonios entre 51.833.450. y 102.676.901 y una proporsión del 9% de los inversionistas ( n= ) tienen patrimonios entre 102.676.901 y 153.520.352

### 14: Monto Inversión

Mediante esta variable cuantitativa, a2cendo lleva el registro del valor en pesos colombianos invertidos por cada inversionista en cada una de las campañas disponibles. Para el presente análisis estadístico, fue seleccionada como variable dependiente, con el propósito de desarrollar el mejor modelo que permita a la compañía hacer una proyección de la demanda con base en las variables explicativas seleccionadas para este estudio en particular.

basicStats(BDa2censo$MontoInversion)

## X..BDa2censo.MontoInversion  
## nobs 2.572000e+04  
## NAs 0.000000e+00  
## Minimum 2.000000e+05  
## Maximum 5.300000e+06  
## 1. Quartile 3.000000e+05  
## 3. Quartile 1.000000e+06  
## Mean 9.673676e+05  
## Median 5.000000e+05  
## Sum 2.488069e+10  
## SE Mean 6.673247e+03  
## LCL Mean 9.542876e+05  
## UCL Mean 9.804475e+05  
## Variance 1.145369e+12  
## Stdev 1.070219e+06  
## Skewness 2.150895e+00  
## Kurtosis 4.542472e+00

MItabla\_st <- fdt(BDa2censo$MontoInversion,breaks="Sturges",right=F)  
  
  
knitr::kable(  
 (MItabla\_st)  
)

| Class limits | f | rf | rf(%) | cf | cf(%) |
| --- | --- | --- | --- | --- | --- |
| [198000,520187.5) | 13900 | 0.5404355 | 54.0435459 | 13900 | 54.04355 |
| [520187.5,842375) | 1842 | 0.0716174 | 7.1617418 | 15742 | 61.20529 |
| [842375,1164563) | 4175 | 0.1623250 | 16.2325039 | 19917 | 77.43779 |
| [1164563,1486750) | 568 | 0.0220840 | 2.2083981 | 20485 | 79.64619 |
| [1486750,1808938) | 791 | 0.0307543 | 3.0754277 | 21276 | 82.72162 |
| [1808938,2131125) | 1899 | 0.0738336 | 7.3833593 | 23175 | 90.10498 |
| [2131125,2453313) | 183 | 0.0071151 | 0.7115086 | 23358 | 90.81649 |
| [2453313,2775500) | 284 | 0.0110420 | 1.1041991 | 23642 | 91.92068 |
| [2775500,3097688) | 755 | 0.0293546 | 2.9354588 | 24397 | 94.85614 |
| [3097688,3419875) | 87 | 0.0033826 | 0.3382582 | 24484 | 95.19440 |
| [3419875,3742063) | 108 | 0.0041991 | 0.4199067 | 24592 | 95.61431 |
| [3742063,4064250) | 339 | 0.0131804 | 1.3180404 | 24931 | 96.93235 |
| [4064250,4386438) | 34 | 0.0013219 | 0.1321928 | 24965 | 97.06454 |
| [4386438,4708625) | 56 | 0.0021773 | 0.2177294 | 25021 | 97.28227 |
| [4708625,5030813) | 661 | 0.0256998 | 2.5699844 | 25682 | 99.85226 |
| [5030813,5353000) | 38 | 0.0014774 | 0.1477449 | 25720 | 100.00000 |

|  | x |
| --- | --- |
| start | 198000.0 |
| end | 5353000.0 |
| h | 322187.5 |
| right | 0.0 |

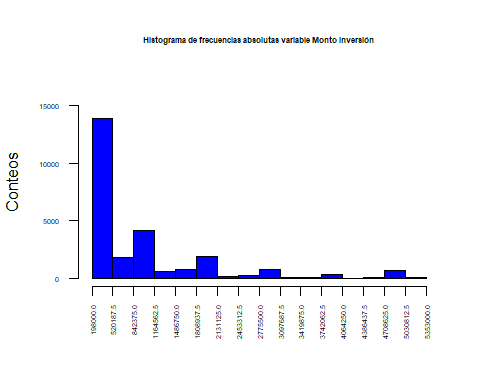
plot(MItabla\_st,type='fh',main="Histograma de frecuencias absolutas variable Monto Inversión",cex.main = 0.5, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="blue", las=2)

## Warning in title(main = main, xlab = xlab, ylab = ylab, ...): "cex.names" is not  
## a graphical parameter

## Warning in axis(2, ...): "cex.names" is not a graphical parameter

## Warning in rect(brk[-length(brk)], 0, brk[-1], y, col = col, ...): "cex.names"  
## is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



plot(MItabla\_st,type='fp',main="Poligono de frecuencias absolutas",cex.main = 1, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="black",lwd=2,las=2)

## Warning in plot.window(...): "cex.names" is not a graphical parameter

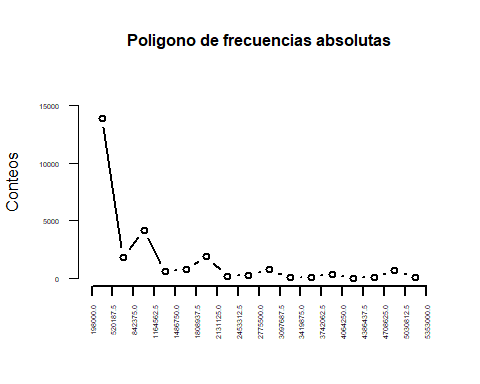
## Warning in plot.xy(xy, type, ...): "cex.names" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter

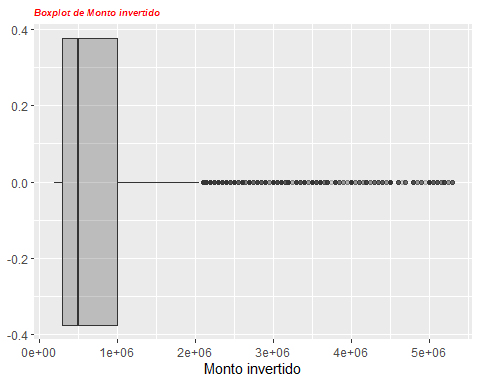
## Warning in box(...): "cex.names" is not a graphical parameter

## Warning in title(...): "cex.names" is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



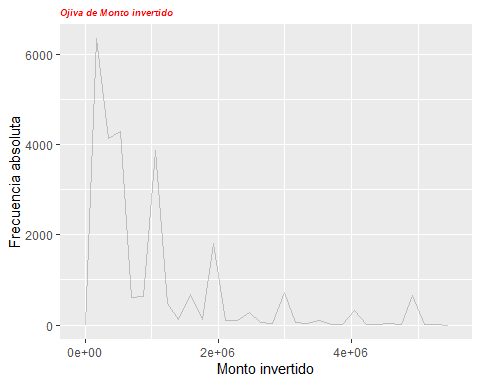
ggplot(BDa2censo, aes( y=MontoInversion)) +   
 geom\_boxplot(fill="black", alpha=0.2) +   
 ylab("Monto invertido") + coord\_flip() + ggtitle("Boxplot de Monto invertido") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))



ggplot(BDa2censo, aes( y=MontoInversion)) +   
 geom\_freqpoly(fill="black", alpha=0.2) +   
 ylab("Monto invertido")+xlab("Frecuencia absoluta") + coord\_flip() + ggtitle("Ojiva de Monto invertido") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Ignoring unknown parameters: fill

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 Respecto al monto invertido podemos encontrar que un 54% de los inversionistas (n= 13.904) invierten en un rango entre 198.000 pesos y 520.187 pesos,encontrandose que este es rango ded ticket mayoritatio de inversión actual. No obstante se encuentra una proporsion del 16% de los inversionistas ( n= 4175) que invierten rangos entre 842.375 pesos y 1.164.563 pesos. Por lo tanto más del 60% de los inversionistas invierten entre 198.000 pesos y 842.375 pesos

### Variable: Edad

Mediante esta variable cuantitativa, a2cendo lleva el registro de la edad de cada inversionista, lo cual es un dato que se debe considerar para el presente análisis estadístico, toda vez que este dato puede afectar el cálculo de la proyección del monto de inversión.

basicStats(BDa2censo$Edad)

## X..BDa2censo.Edad  
## nobs 2.572000e+04  
## NAs 0.000000e+00  
## Minimum 1.800000e+01  
## Maximum 8.600000e+01  
## 1. Quartile 3.000000e+01  
## 3. Quartile 4.200000e+01  
## Mean 3.718176e+01  
## Median 3.500000e+01  
## Sum 9.563150e+05  
## SE Mean 6.133800e-02  
## LCL Mean 3.706154e+01  
## UCL Mean 3.730199e+01  
## Variance 9.676859e+01  
## Stdev 9.837102e+00  
## Skewness 1.263190e+00  
## Kurtosis 2.000467e+00

Edtabla\_st <- fdt(BDa2censo$Edad,breaks="Sturges",right=F)  
  
  
knitr::kable(  
 (Edtabla\_st)  
)

| Class limits | f | rf | rf(%) | cf | cf(%) |
| --- | --- | --- | --- | --- | --- |
| [17.82,22.135) | 216 | 0.0083981 | 0.8398134 | 216 | 0.8398134 |
| [22.135,26.45) | 2083 | 0.0809876 | 8.0987558 | 2299 | 8.9385692 |
| [26.45,30.765) | 4741 | 0.1843313 | 18.4331260 | 7040 | 27.3716952 |
| [30.765,35.08) | 5846 | 0.2272939 | 22.7293935 | 12886 | 50.1010886 |
| [35.08,39.395) | 4610 | 0.1792379 | 17.9237947 | 17496 | 68.0248834 |
| [39.395,43.71) | 3092 | 0.1202177 | 12.0217729 | 20588 | 80.0466563 |
| [43.71,48.025) | 2228 | 0.0866252 | 8.6625194 | 22816 | 88.7091757 |
| [48.025,52.34) | 898 | 0.0349145 | 3.4914463 | 23714 | 92.2006221 |
| [52.34,56.655) | 673 | 0.0261664 | 2.6166407 | 24387 | 94.8172628 |
| [56.655,60.97) | 366 | 0.0142302 | 1.4230171 | 24753 | 96.2402799 |
| [60.97,65.285) | 445 | 0.0173017 | 1.7301711 | 25198 | 97.9704510 |
| [65.285,69.6) | 283 | 0.0110031 | 1.1003110 | 25481 | 99.0707621 |
| [69.6,73.915) | 112 | 0.0043546 | 0.4354588 | 25593 | 99.5062208 |
| [73.915,78.23) | 91 | 0.0035381 | 0.3538103 | 25684 | 99.8600311 |
| [78.23,82.545) | 21 | 0.0008165 | 0.0816485 | 25705 | 99.9416796 |
| [82.545,86.86) | 15 | 0.0005832 | 0.0583204 | 25720 | 100.0000000 |

|  | x |
| --- | --- |
| start | 17.820 |
| end | 86.860 |
| h | 4.315 |
| right | 0.000 |

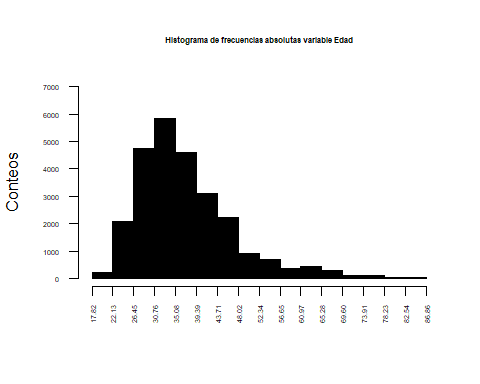
plot(Edtabla\_st,type='fh',main="Histograma de frecuencias absolutas variable Edad",cex.main = 0.5, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="black", las=2)

## Warning in title(main = main, xlab = xlab, ylab = ylab, ...): "cex.names" is not  
## a graphical parameter

## Warning in axis(2, ...): "cex.names" is not a graphical parameter

## Warning in rect(brk[-length(brk)], 0, brk[-1], y, col = col, ...): "cex.names"  
## is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



plot(Edtabla\_st,type='fp',main="Poligono de frecuencias absolutas Variable Edad",cex.main = 0.5, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="black",lwd=2,las=2)

## Warning in plot.window(...): "cex.names" is not a graphical parameter

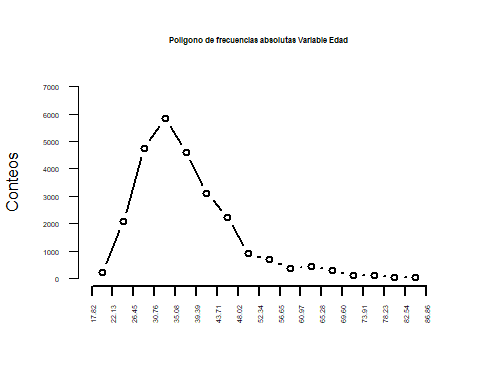
## Warning in plot.xy(xy, type, ...): "cex.names" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter

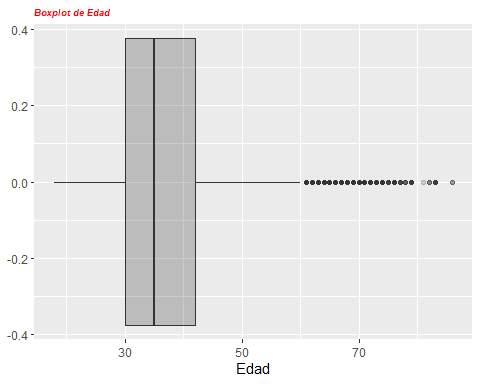
## Warning in box(...): "cex.names" is not a graphical parameter

## Warning in title(...): "cex.names" is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



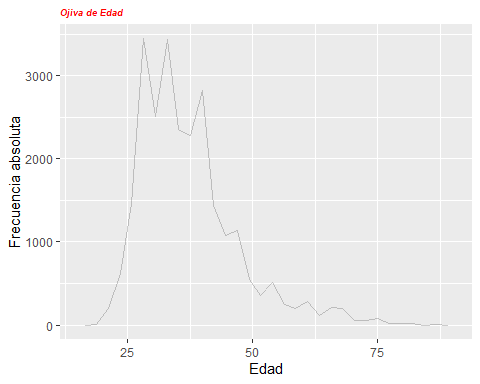
ggplot(BDa2censo, aes( y=Edad)) +   
 geom\_boxplot(fill="black", alpha=0.2) +   
 ylab("Edad") + coord\_flip() + ggtitle("Boxplot de Edad") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))



ggplot(BDa2censo, aes( y=Edad)) +   
 geom\_freqpoly(fill="black", alpha=0.2) +   
 ylab("Edad")+xlab("Frecuencia absoluta") + coord\_flip() + ggtitle("Ojiva de Edad") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Ignoring unknown parameters: fill

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 De acuerdo al analisis cuantitativo de la variable edad podemos encontrar que la mayor cantidad de inversionistas (22%, n = 5847) se encuentran en un rango de edad entre 31 años y 35 años. No obstante la participación de inversionistas entre los 26 años y 31 años tambien es relevante con un 18% de participación (n= 4741) y tambien el rango comprendido entre los 35 años y 39 años con un 18% de participación (n = 4611) y el rango entre 39 años y 44 años con un 12% de la participación total (n = 3092). Dicho lo anterior los inversionistas entre los 26 años y 45 años suman un total del 59% de participacion de los inversionistas siendo la edad media 35 años.

### Variable 15 : Cantidad Inversiones

Mediante esta variable cuantitativa, a2cendo lleva el registro de la edad de cada inversionista, lo cual es un dato que se debe considerar para el presente análisis estadístico, toda vez que este dato puede afectar el cálculo de la proyección del monto de inversión.

basicStats(BDa2censo$CantidadInversiones)

## X..BDa2censo.CantidadInversiones  
## nobs 2.572000e+04  
## NAs 0.000000e+00  
## Minimum 1.000000e+00  
## Maximum 9.100000e+01  
## 1. Quartile 3.000000e+00  
## 3. Quartile 1.700000e+01  
## Mean 1.219810e+01  
## Median 8.000000e+00  
## Sum 3.137350e+05  
## SE Mean 7.678000e-02  
## LCL Mean 1.204760e+01  
## UCL Mean 1.234859e+01  
## Variance 1.516242e+02  
## Stdev 1.231358e+01  
## Skewness 1.748056e+00  
## Kurtosis 3.334008e+00

CItabla\_st <- fdt(BDa2censo$CantidadInversiones,breaks="Sturges",right=F)  
  
  
knitr::kable(  
 (CItabla\_st)  
)

| Class limits | f | rf | rf(%) | cf | cf(%) |
| --- | --- | --- | --- | --- | --- |
| [0.99,6.6725) | 11094 | 0.4313375 | 43.1337481 | 11094 | 43.13375 |
| [6.6725,12.355) | 5411 | 0.2103810 | 21.0381026 | 16505 | 64.17185 |
| [12.355,18.037) | 3623 | 0.1408631 | 14.0863142 | 20128 | 78.25816 |
| [18.037,23.72) | 1649 | 0.0641135 | 6.4113530 | 21777 | 84.66952 |
| [23.72,29.402) | 1566 | 0.0608865 | 6.0886470 | 23343 | 90.75816 |
| [29.402,35.085) | 761 | 0.0295879 | 2.9587869 | 24104 | 93.71695 |
| [35.085,40.768) | 538 | 0.0209176 | 2.0917574 | 24642 | 95.80871 |
| [40.768,46.45) | 359 | 0.0139580 | 1.3958009 | 25001 | 97.20451 |
| [46.45,52.133) | 263 | 0.0102255 | 1.0225505 | 25264 | 98.22706 |
| [52.133,57.815) | 212 | 0.0082426 | 0.8242613 | 25476 | 99.05132 |
| [57.815,63.498) | 115 | 0.0044712 | 0.4471229 | 25591 | 99.49844 |
| [63.498,69.18) | 128 | 0.0049767 | 0.4976672 | 25719 | 99.99611 |
| [69.18,74.862) | 0 | 0.0000000 | 0.0000000 | 25719 | 99.99611 |
| [74.862,80.545) | 0 | 0.0000000 | 0.0000000 | 25719 | 99.99611 |
| [80.545,86.227) | 0 | 0.0000000 | 0.0000000 | 25719 | 99.99611 |
| [86.227,91.91) | 1 | 0.0000389 | 0.0038880 | 25720 | 100.00000 |

|  | x |
| --- | --- |
| start | 0.9900 |
| end | 91.9100 |
| h | 5.6825 |
| right | 0.0000 |

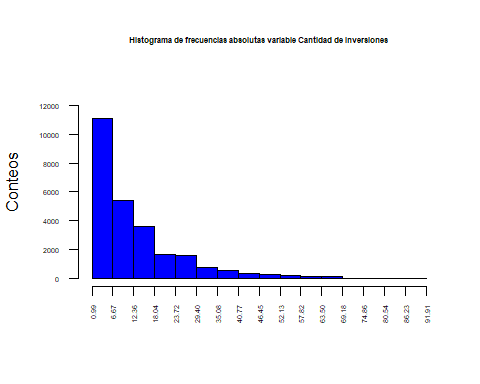
plot(CItabla\_st,type='fh',main="Histograma de frecuencias absolutas variable Cantidad de inversiones",cex.main = 0.5, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="blue", las=2)

## Warning in title(main = main, xlab = xlab, ylab = ylab, ...): "cex.names" is not  
## a graphical parameter

## Warning in axis(2, ...): "cex.names" is not a graphical parameter

## Warning in rect(brk[-length(brk)], 0, brk[-1], y, col = col, ...): "cex.names"  
## is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



plot(CItabla\_st,type='fp',main="Poligono de frecuencias absolutas Variable Cantidad de inversiones",cex.main = 0.5, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="black",lwd=2,las=2)

## Warning in plot.window(...): "cex.names" is not a graphical parameter

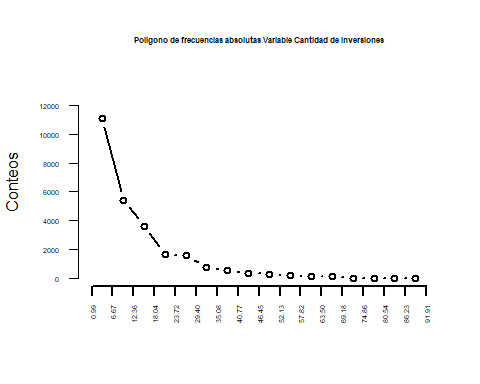
## Warning in plot.xy(xy, type, ...): "cex.names" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter

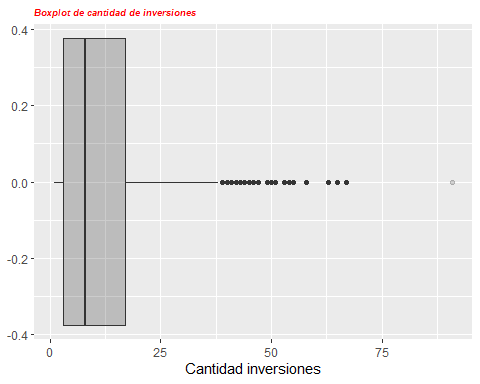
## Warning in box(...): "cex.names" is not a graphical parameter

## Warning in title(...): "cex.names" is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



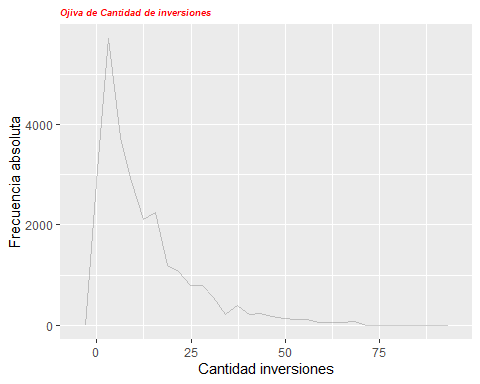
ggplot(BDa2censo, aes( y=CantidadInversiones)) +   
 geom\_boxplot(fill="black", alpha=0.2) +   
 ylab("Cantidad inversiones") + coord\_flip() + ggtitle("Boxplot de cantidad de inversiones") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))



ggplot(BDa2censo, aes( y=CantidadInversiones)) +   
 geom\_freqpoly(fill="black", alpha=0.2) +   
 ylab("Cantidad inversiones")+xlab("Frecuencia absoluta") + coord\_flip() + ggtitle("Ojiva de Cantidad de inversiones") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Ignoring unknown parameters: fill

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 Respecto a la variable Cantidad de inversiones podemos encontrar que el promedio de inversiones es 12 inversiones. No obstante un 43% (n = 11.099) de los inversionistas han realizado entre 1 y 7 inversiones. un 21% (n = 5.412) han realizado entre 7 y 12 inversiones, un 14% (n= 3.623) han realizado entre 12 y 18 inversiones y un 6% (n = 1.649) han realizado entre 18 y 24 inversiones en la plataforma de a2censo.

basicStats(BDa2censo$Tasa)

## X..BDa2censo.Tasa  
## nobs 25720.000000  
## NAs 0.000000  
## Minimum 0.080000  
## Maximum 0.130000  
## 1. Quartile 0.100000  
## 3. Quartile 0.110000  
## Mean 0.104405  
## Median 0.100000  
## Sum 2685.293400  
## SE Mean 0.000064  
## LCL Mean 0.104280  
## UCL Mean 0.104530  
## Variance 0.000105  
## Stdev 0.010254  
## Skewness 0.436234  
## Kurtosis -0.255612

CItabla\_st <- fdt(BDa2censo$Tasa,breaks="Sturges",right=F)  
  
  
knitr::kable(  
 (CItabla\_st)  
)

| Class limits | f | rf | rf(%) | cf | cf(%) |
| --- | --- | --- | --- | --- | --- |
| [0.0792,0.08246) | 187 | 0.0072706 | 0.7270607 | 187 | 0.7270607 |
| [0.08246,0.08571) | 370 | 0.0143857 | 1.4385692 | 557 | 2.1656299 |
| [0.08571,0.08897) | 77 | 0.0029938 | 0.2993779 | 634 | 2.4650078 |
| [0.08897,0.09223) | 2636 | 0.1024883 | 10.2488336 | 3270 | 12.7138414 |
| [0.09223,0.09548) | 1581 | 0.0614697 | 6.1469673 | 4851 | 18.8608087 |
| [0.09548,0.09874) | 1289 | 0.0501166 | 5.0116641 | 6140 | 23.8724728 |
| [0.09874,0.102) | 7593 | 0.2952177 | 29.5217729 | 13733 | 53.3942457 |
| [0.102,0.1053) | 2749 | 0.1068818 | 10.6881804 | 16482 | 64.0824261 |
| [0.1053,0.1085) | 213 | 0.0082815 | 0.8281493 | 16695 | 64.9105754 |
| [0.1085,0.1118) | 3067 | 0.1192457 | 11.9245723 | 19762 | 76.8351477 |
| [0.1118,0.115) | 2284 | 0.0888025 | 8.8802488 | 22046 | 85.7153966 |
| [0.115,0.1183) | 51 | 0.0019829 | 0.1982893 | 22097 | 85.9136858 |
| [0.1183,0.1215) | 2765 | 0.1075039 | 10.7503888 | 24862 | 96.6640747 |
| [0.1215,0.1248) | 0 | 0.0000000 | 0.0000000 | 24862 | 96.6640747 |
| [0.1248,0.128) | 119 | 0.0046267 | 0.4626750 | 24981 | 97.1267496 |
| [0.128,0.1313) | 739 | 0.0287325 | 2.8732504 | 25720 | 100.0000000 |

|  | x |
| --- | --- |
| start | 0.0792000 |
| end | 0.1313000 |
| h | 0.0032562 |
| right | 0.0000000 |

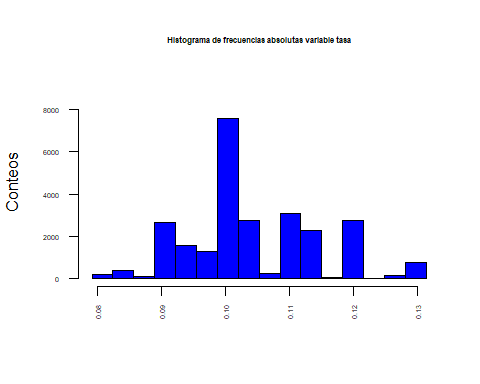
plot(CItabla\_st,type='fh',main="Histograma de frecuencias absolutas variable tasa",cex.main = 0.5, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="blue", las=2)

## Warning in title(main = main, xlab = xlab, ylab = ylab, ...): "cex.names" is not  
## a graphical parameter

## Warning in axis(2, ...): "cex.names" is not a graphical parameter

## Warning in rect(brk[-length(brk)], 0, brk[-1], y, col = col, ...): "cex.names"  
## is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



plot(CItabla\_st,type='fp',main="Poligono de frecuencias absolutas Variable tasa",cex.main = 0.5, las=2, cex.names = 0.35, cex.axis=0.45,  
 xlab=" ",ylab="Conteos",col="black",lwd=2,las=2)

## Warning in plot.window(...): "cex.names" is not a graphical parameter

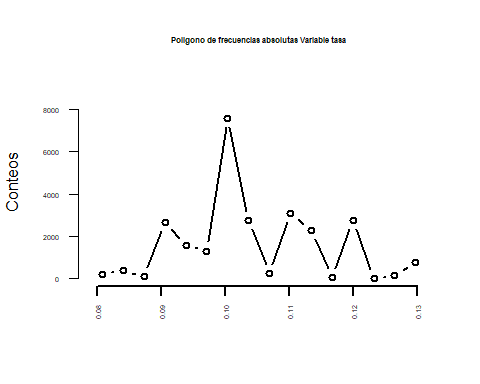
## Warning in plot.xy(xy, type, ...): "cex.names" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "cex.names" is not  
## a graphical parameter

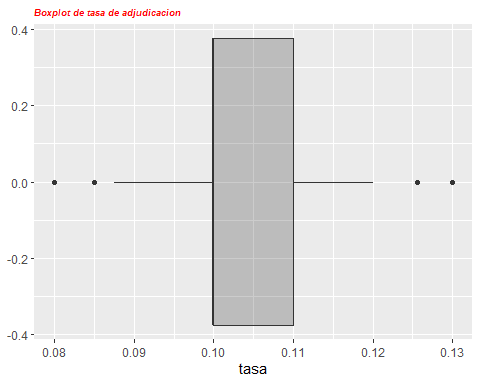
## Warning in box(...): "cex.names" is not a graphical parameter

## Warning in title(...): "cex.names" is not a graphical parameter

## Warning in axis(1, at = round(brk, x.round), las = xlas, ...): "cex.names" is  
## not a graphical parameter



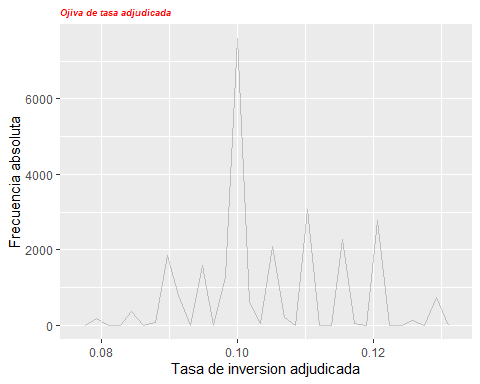
ggplot(BDa2censo, aes( y=Tasa)) +   
 geom\_boxplot(fill="black", alpha=0.2) +   
 ylab("tasa") + coord\_flip() + ggtitle("Boxplot de tasa de adjudicacion") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))



ggplot(BDa2censo, aes( y=Tasa)) +   
 geom\_freqpoly(fill="black", alpha=0.2) +   
 ylab("Tasa de inversion adjudicada")+xlab("Frecuencia absoluta") + coord\_flip() + ggtitle("Ojiva de tasa adjudicada") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Ignoring unknown parameters: fill

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 De acuerdo al analisis de la variable cuantitativa de tasa de adjudicación de la campaña podemos encontrar que hay dos rangos de tasas predominantes en cantidad de inversionistas, el 29.52% (n = 7595) de los inversionistas lograron tasas de adjudicacion de la inversion entre los 0.098 y el 0.102. En siguiente lugar podemos encontrar que el 12% de los inversionistas (n= 3068 )lograron tasas entre el 0.1085 y el 0.1118

## Resumen de medidas de localización

# Media Geométrica  
#Función  
geometric<-function(x) exp(sum(log(x))/length(x))  
  
#-------------------------------------------------------------------------------  
  
# Media Armónica  
#Función  
armonic<-function(x) 1/mean(1/x)  
  
  
#-------------------------------------------------------------------------------  
  
#Media Recortada en 100α% :  
#Función:  
recortada<-function(x) mean(x, trim=0.1)  
  
  
#-------------------------------------------------------------------------------  
  
#Media Winsorizada:  
library(psych)#Paquete útil para calcular la media Winsorizada:  
#Funcion  
winsorizada<-function(x) winsor.mean(x, trim= 0.1, na.rm = TRUE)  
  
#--------------------------------------------------------------------------------  
#Cálculo Resumen de medidas de localización  
  
options(scipen = 999)  
ResumenVaribles<- c("Edad","Tasa","Ingresos Mes", "Patrimonio","Monto Inversion","Cantidad Inversiones" )  
ResumenPromedio<- c(mean(BDa2censo$Edad),mean(BDa2censo$Tasa),mean(BDa2censo$IngresosMes),mean(BDa2censo$Patrimonio), mean(BDa2censo$MontoInversion),mean(BDa2censo$CantidadInversiones))  
ResumenMediana<- c(median(BDa2censo$Edad),median(BDa2censo$Tasa),median(BDa2censo$IngresosMes),median(BDa2censo$Patrimonio), median(BDa2censo$MontoInversion),median(BDa2censo$CantidadInversiones))  
ResumenGeometrica<- c(geometric(BDa2censo$Edad),geometric(BDa2censo$Tasa),geometric(BDa2censo$IngresosMes),geometric(BDa2censo$Patrimonio), geometric(BDa2censo$MontoInversion),geometric(BDa2censo$CantidadInversiones))  
ResumenArmonica<- c(armonic(BDa2censo$Edad),armonic(BDa2censo$Tasa),armonic(BDa2censo$IngresosMes),armonic(BDa2censo$Patrimonio), armonic(BDa2censo$MontoInversion),armonic(BDa2censo$CantidadInversiones))  
ResumenRecortada<- c(recortada(BDa2censo$Edad),recortada(BDa2censo$Tasa),recortada(BDa2censo$IngresosMes),recortada(BDa2censo$Patrimonio), recortada(BDa2censo$MontoInversion),recortada(BDa2censo$CantidadInversiones))  
ResumenWinsorizada<- c(winsorizada(BDa2censo$Edad),winsorizada(BDa2censo$Tasa),winsorizada(BDa2censo$IngresosMes),winsorizada(BDa2censo$Patrimonio), winsorizada(BDa2censo$MontoInversion),winsorizada(BDa2censo$CantidadInversiones))  
  
resumenMedia <- data.frame(ResumenVaribles ,ResumenPromedio,ResumenMediana,ResumenGeometrica)  
  
resumenMedia1 <- data.frame(ResumenVaribles ,ResumenArmonica,ResumenRecortada,ResumenWinsorizada)  
  
  
knitr::kable(  
 (resumenMedia)  
)

| ResumenVaribles | ResumenPromedio | ResumenMediana | ResumenGeometrica |
| --- | --- | --- | --- |
| Edad | 37.1817652 | 35.0 | 36.0365311 |
| Tasa | 0.1044049 | 0.1 | 0.1039105 |
| Ingresos Mes | 7590004.2873250 | 4878000.0 | 5017014.3233656 |
| Patrimonio | 157008240.3223561 | 100000000.0 | 64676964.5259231 |
| Monto Inversion | 967367.5738725 | 500000.0 | 611385.6725858 |
| Cantidad Inversiones | 12.1980949 | 8.0 | 7.1994345 |

knitr::kable(  
 (resumenMedia1)  
)

| ResumenVaribles | ResumenArmonica | ResumenRecortada | ResumenWinsorizada |
| --- | --- | --- | --- |
| Edad | 35.0149957 | 35.9849339 | 36.4879471 |
| Tasa | 0.1034249 | 0.1040474 | 0.1044279 |
| Ingresos Mes | 3793266.2022598 | 5523128.3514288 | 6008502.6811431 |
| Patrimonio | 14267648.1084715 | 122361003.4177683 | 139643202.7342146 |
| Monto Inversion | 430745.4551011 | 728999.5139969 | 813199.6111975 |
| Cantidad Inversiones | 3.9411647 | 10.0013122 | 11.0010498 |

## Análisis Bivariado Cualitativo

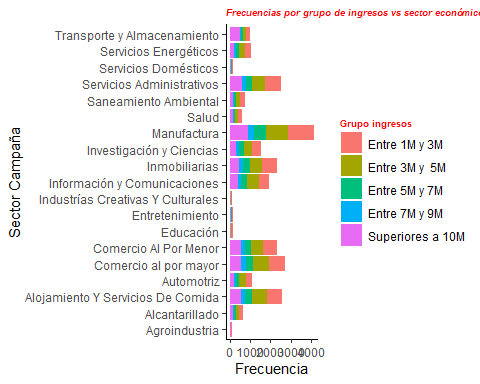
#Visualización de las variables cualitativas desde la perspectiva de grupo de ingresos

A continuación el analisis bivariado cualitativo se realizará comparando las variables cualitativas contra la variable de grupo ingresos

ggplot(BDa2censo,aes(x=BDa2censo$SectorCampaña,)) + geom\_bar(aes(fill= BDa2censo$GrupoIngresos)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por grupo de ingresos vs sector económico") + ylab("Frecuencia") + xlab("Sector Campaña") + theme(legend.title = element\_text(colour="red", size=7,face="bold")) + labs(fill = "Grupo ingresos") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Use of `BDa2censo$GrupoIngresos` is discouraged. Use `GrupoIngresos`  
## instead.

## Warning: Use of `BDa2censo$SectorCampaña` is discouraged. Use `SectorCampaña`  
## instead.



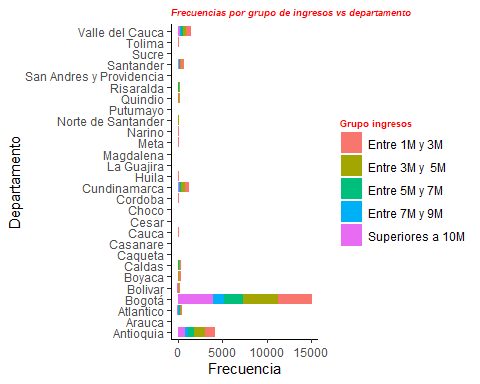
A partir del anterior grafico podemos observar la concentración de inversionistas con ingresos superiores a 10M de pesos en sectores económicos de manufactura y servicios administrativos, en los cuales la pacticipacion de personas con ingresos entre 3M y 5M también es relevante.

Resulta interesante observar que en general las empresas financiadas en a2censo que pertenecen al sector de manufactura Servicios administrativos e inmobiliarias, demuestran mayor traccion de inversionistas de alto capital y medio-bajo capital caraterizandose por tener inversionistas con ingresos superiories a 10 millones, y entre 3 y 5 MM

ggplot(BDa2censo,aes(x=BDa2censo$Departamento,)) + geom\_bar(aes(fill= BDa2censo$GrupoIngresos)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por grupo de ingresos vs departamento") + ylab("Frecuencia") + xlab("Departamento") + theme(legend.title = element\_text(colour="red", size=7,face="bold")) + labs(fill = "Grupo ingresos") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Use of `BDa2censo$GrupoIngresos` is discouraged. Use `GrupoIngresos`  
## instead.

## Warning: Use of `BDa2censo$Departamento` is discouraged. Use `Departamento`  
## instead.

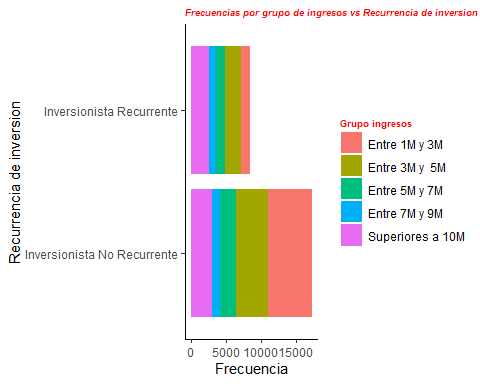


Lejos de encontrar relevante sacar alguna conclusión de acuerdo al rango de ingresos, de manera general podemos observar que departamentos como valle del cauca, Santander, Antioquia son regiones potenciales para atracción de inversionistas en donde a2censo no ha tenido una alta participacion a nivel de marketing o acercamiento con inversionistas. Para el caso de Bogotá podemos evidenciar como la proporsion de inversionistas con rangos de ingresos medios- bajos (3 a 5 millones) y medios altos (superior a 10 millones) es predominante

ggplot(BDa2censo,aes(x=BDa2censo$CodEsInversionistaRecurrente,)) + geom\_bar(aes(fill= BDa2censo$GrupoIngresos)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por grupo de ingresos vs Recurrencia de inversion") + ylab("Frecuencia") + xlab("Recurrencia de inversion") + theme(legend.title = element\_text(colour="red", size=7,face="bold")) + labs(fill = "Grupo ingresos") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Use of `BDa2censo$GrupoIngresos` is discouraged. Use `GrupoIngresos`  
## instead.

## Warning: Use of `BDa2censo$CodEsInversionistaRecurrente` is discouraged. Use  
## `CodEsInversionistaRecurrente` instead.



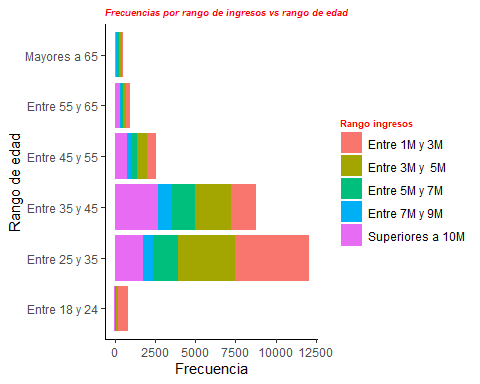
De acuerdo a la anterior grafica podemos observar como los inversionistas recurrentes son en mayor medida personas con ingresos superiores a 10 M, aun cuando hay una proporsion importante de inversionistas recurrentes con ingresos entre 3M y 5M que no son necesariamente de un alto rango de ingresos y aun asi estan viendo una oportunidad recurrente de invertir en el mercado

En el caso de los inversionistas no recurrentes podemos encontrar como una importante proporsion de inversionistas son personas con ingresos entre 1 M a 5 M

ggplot(BDa2censo,aes(x=BDa2censo$GrupoEdad,)) + geom\_bar(aes(fill= BDa2censo$GrupoIngresos)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por rango de ingresos vs rango de edad") + ylab("Frecuencia") + xlab("Rango de edad") + theme(legend.title = element\_text(colour="red", size=7,face="bold")) + labs(fill = "Rango ingresos") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Use of `BDa2censo$GrupoIngresos` is discouraged. Use `GrupoIngresos`  
## instead.

## Warning: Use of `BDa2censo$GrupoEdad` is discouraged. Use `GrupoEdad` instead.



Realizando un analisis bivariado entre el grupo de edad y el rango de ingresos resulta interesante observar la participacion que tienen personas entre rangos de edad de 25 a 35 con ingresos entre 1 a 5 millones de pesos los cuales representan un volumen importante, no obstante volumenes similares se observan el personas con edades entre 35 y 45 años suelen ser en mayor medida personas con ingresos superiores a 10 millones de pesos.

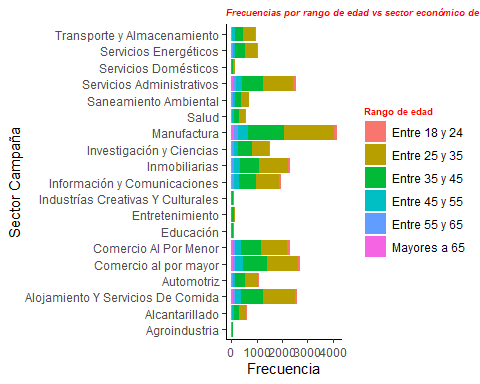
#Visualización de las variables cualitativas desde la perspectiva de grupo de edad

A continuación el analisis bivariado cualitativo se realizará comparando las variables contra la variable de grupo de edad.

ggplot(BDa2censo,aes(x=BDa2censo$SectorCampaña,)) + geom\_bar(aes(fill= BDa2censo$GrupoEdad)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por rango de edad vs sector económico de la campaña") + ylab("Frecuencia") + xlab("Sector Campaña") + theme(legend.title = element\_text(colour="red", size=7,face="bold")) + labs(fill = "Rango de edad") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Use of `BDa2censo$GrupoEdad` is discouraged. Use `GrupoEdad` instead.

## Warning: Use of `BDa2censo$SectorCampaña` is discouraged. Use `SectorCampaña`  
## instead.

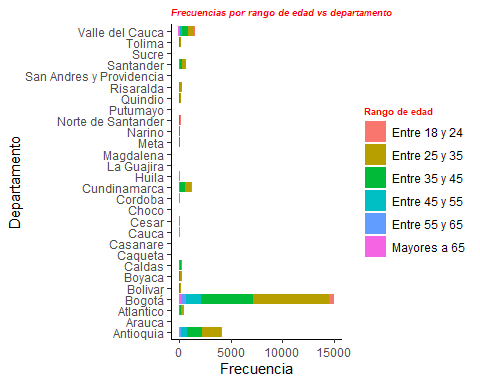


Podemos observar que de manera general independiente al sector economico de la empresa financiada, los grupos poblacionales que más invierten en a2censo se encuentran entre 25 y 45 años , no obstante se observa una interesante participación de personas mayores de 65 años en las campañas de tecnología , servicios alimenticios y salud

ggplot(BDa2censo,aes(x=BDa2censo$Departamento,)) + geom\_bar(aes(fill= BDa2censo$GrupoEdad)) + coord\_flip() + theme\_classic() + theme\_classic() + ggtitle("Frecuencias por rango de edad vs departamento") + ylab("Frecuencia") + xlab("Departamento") + theme(legend.title = element\_text(colour="red", size=7,face="bold")) + labs(fill = "Rango de edad") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Use of `BDa2censo$GrupoEdad` is discouraged. Use `GrupoEdad` instead.

## Warning: Use of `BDa2censo$Departamento` is discouraged. Use `Departamento`  
## instead.

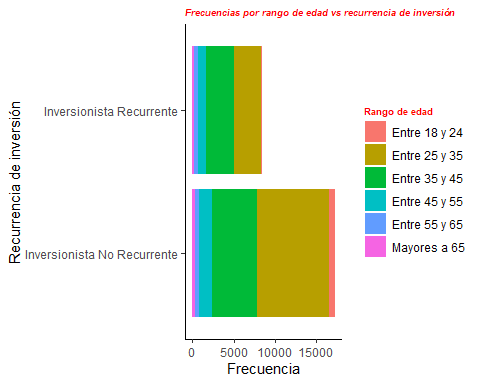


No es el grupo poblacional mas grande dentro de la comunidad de inversionistas, no obstante puede ser un grupo de atención las personas entre 45 y 55 años los cuales pueden ser un segmento a desarrollar en Bogotá, Antioquia y valle del cauca

ggplot(BDa2censo,aes(x=BDa2censo$CodEsInversionistaRecurrente,)) + geom\_bar(aes(fill= BDa2censo$GrupoEdad)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por rango de edad vs recurrencia de inversión") + ylab("Frecuencia") + xlab("Recurrencia de inversión") + theme(legend.title = element\_text(colour="red", size=7,face="bold")) + labs(fill = "Rango de edad") + theme(plot.title = element\_text(color="red", size=7, face="bold.italic"))

## Warning: Use of `BDa2censo$GrupoEdad` is discouraged. Use `GrupoEdad` instead.

## Warning: Use of `BDa2censo$CodEsInversionistaRecurrente` is discouraged. Use  
## `CodEsInversionistaRecurrente` instead.



Del anterior grafico podemos concluir que para el caso de los inversionistas recurrentes hay una leve proporsion mayor de inversionistas en rangos de edades entre 35 y 45 años mientras que esta proporsion mayor de inversionistas no recurrentes de centra en personas entre 25 y 35 años, aun cuando ambos grupos de edades son predominantes estos dos grupos de edades

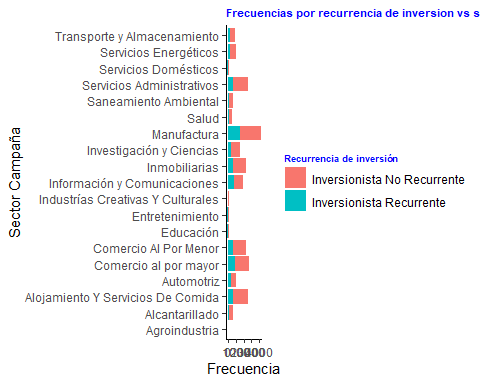
#Visualización de las variables cualitativas desde la perspectiva de inversionista recurrente

A continuación el analisis bivariado cualitativo se realizará comparando las variables contra la variable de categorización si es un inversionista recurrente o no lo es.

ggplot(BDa2censo,aes(x=BDa2censo$SectorCampaña,)) + geom\_bar(aes(fill= BDa2censo$CodEsInversionistaRecurrente)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por recurrencia de inversion vs sector campaña") + ylab("Frecuencia") + xlab("Sector Campaña") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + labs(fill = "Recurrencia de inversión") + theme(plot.title = element\_text(color="blue", size=8, face="bold"))

## Warning: Use of `BDa2censo$CodEsInversionistaRecurrente` is discouraged. Use  
## `CodEsInversionistaRecurrente` instead.

## Warning: Use of `BDa2censo$SectorCampaña` is discouraged. Use `SectorCampaña`  
## instead.

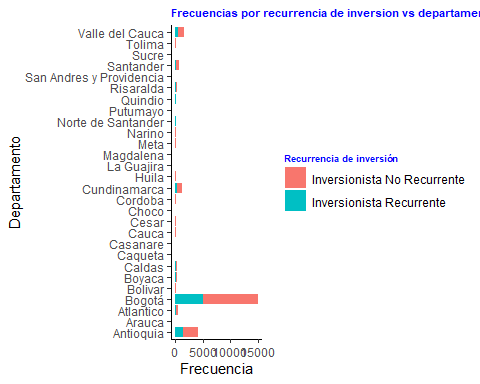


De manera general se puede observar que la cantidad de inversionistas NO recurrentes es mayor para todos los sectores economicos de las empresas financiadas, es una distribución muy acorde al numero de inversionistas sin considerar si es recurrente o no. No obstante llama la atención la proporsion grande que mantienen los inversionistas recurrentes en sector económico de manufactura

ggplot(BDa2censo,aes(x=BDa2censo$Departamento,)) + geom\_bar(aes(fill= BDa2censo$CodEsInversionistaRecurrente)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por recurrencia de inversion vs departamento") + ylab("Frecuencia") + xlab("Departamento") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + labs(fill = "Recurrencia de inversión") + theme(plot.title = element\_text(color="blue", size=8, face="bold"))

## Warning: Use of `BDa2censo$CodEsInversionistaRecurrente` is discouraged. Use  
## `CodEsInversionistaRecurrente` instead.

## Warning: Use of `BDa2censo$Departamento` is discouraged. Use `Departamento`  
## instead.

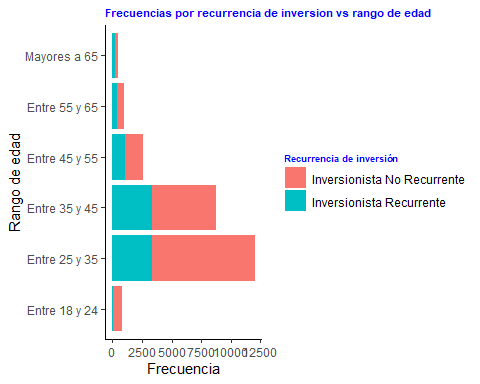


Se puede evidencia la alta proporsión de inversionistas recurrentes ubicados en Bogotá

ggplot(BDa2censo,aes(x=BDa2censo$GrupoEdad,)) + geom\_bar(aes(fill= BDa2censo$CodEsInversionistaRecurrente)) + coord\_flip() + theme\_classic() + ggtitle("Frecuencias por recurrencia de inversion vs rango de edad") + ylab("Frecuencia") + xlab("Rango de edad") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + labs(fill = "Recurrencia de inversión") + theme(plot.title = element\_text(color="blue", size=8, face="bold"))

## Warning: Use of `BDa2censo$CodEsInversionistaRecurrente` is discouraged. Use  
## `CodEsInversionistaRecurrente` instead.

## Warning: Use of `BDa2censo$GrupoEdad` is discouraged. Use `GrupoEdad` instead.



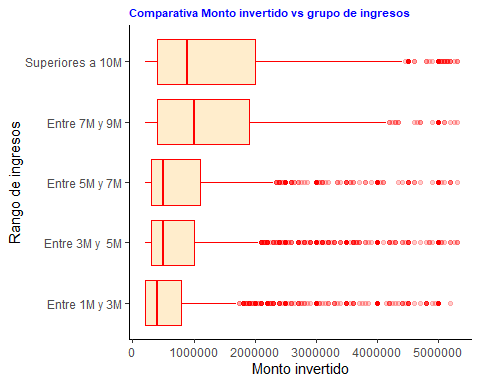
De la anterior grafica resulta interesante observar que actualmente la cantidad de inversionistas recurrentes son mas en personas entre 25 y 45 años, en donde podemos observar que para este rango de edad entre 25 y 35 años la mayor cantidad de inversionistas son no recurrentes

## 2. Resumen Bivariado entre variables Cualitativas y cuantitativas

A continuación se realizará un analisis Bivariado que contemple una de las variables cualitativas de mayor interés del estudio en cuestión, respecto a una de las variables cuantitativas evaluadas.

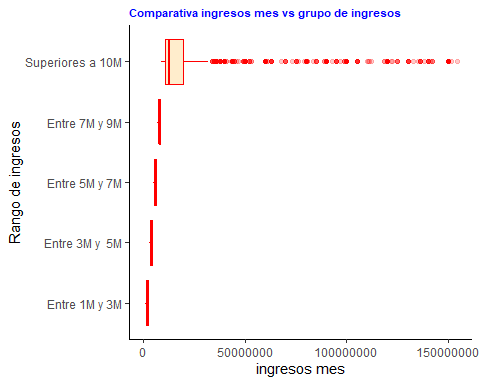
# 1. Analisis de las variables desde el monto invertido

ggplot(data = BDa2censo, mapping = aes(x = MontoInversion , y = GrupoIngresos)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa Monto invertido vs grupo de ingresos") + ylab("Rango de ingresos") + xlab("Monto invertido") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



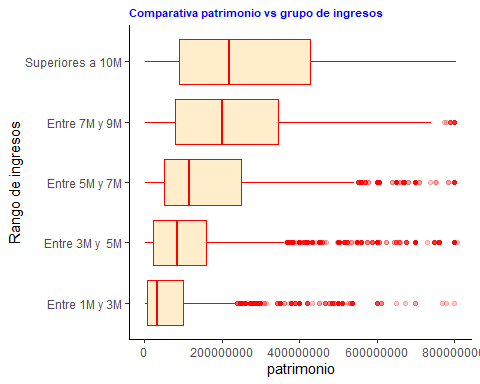
Del anterior diagrama de cajas y bigotes podemos concluir que aunque desde la correlacion lineal simple no se encuentra correlacion dada la alta dispersion de los datos que se tienen, a nivel del analisis de la media del monto invertido parece comportarse de manera acorde al rango de ingresos en donde podemos observar que a mayor rango de ingresos la media de monto invertido va incrementando proporsionalmente. El unico rango de ingresos que muestra un comportamiento atipico es el de los inversionistas con rangos de ingresos entre 7 y 9 millones mensuales, quienes muestran tener un promedio de inversión mayor incluso que las personas con ingresos mayores a 10 millones.

ggplot(data = BDa2censo, mapping = aes(x = IngresosMes , y = GrupoIngresos)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa ingresos mes vs grupo de ingresos") + ylab("Rango de ingresos") + xlab("ingresos mes") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



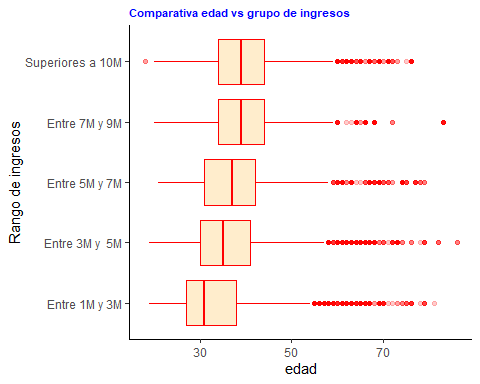
De manera similar al anterior grafico podemos encontrar que a) la mayor dispersión de datos respecto a ingresos mes se encuentra en las personas con ingresos superiores a 10 millones de pesos, no obstante desde el analisis de medidas de tendencia central y dispersión podemos observar menores desviaciones estandar en lo que respecto a ingresos menores a 10 millones de pesos

ggplot(data = BDa2censo, mapping = aes(x = Patrimonio , y = GrupoIngresos)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa patrimonio vs grupo de ingresos") + ylab("Rango de ingresos") + xlab("patrimonio") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



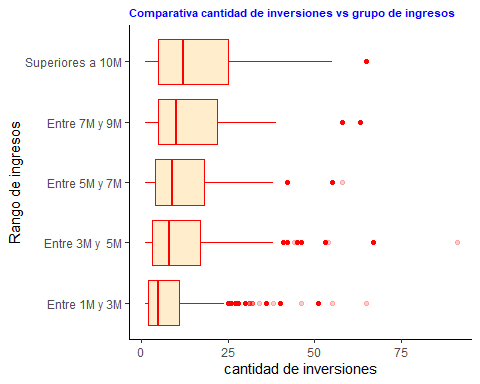
De manera similar, la medida de tendencia central es acorde a la conclusión de que un mayor promedio de patrimonio reportado es acorde a un mayor rango de ingresos, no obstante la dispersión no permitiría generalizar esta observación para todos los sujetos encontrandose sujetos con alto patrimonio en ingresos entre 1 y 7 millones, lo cual puede hablar de la calidad de la información que están diligenciando los inversionistas

ggplot(data = BDa2censo, mapping = aes(x = Edad , y = GrupoIngresos)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa edad vs grupo de ingresos") + ylab("Rango de ingresos") + xlab("edad") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



De acuerdo con el comportamiento que muchas veces se ha observado en las estadisticas Nacionales podemos observar que a mayor edad el promedio de ingresos es mayor encontrandose que el promedio de edad de nuestros inversionistas se encuentra en mayor medida entre los 30 y 40 ños

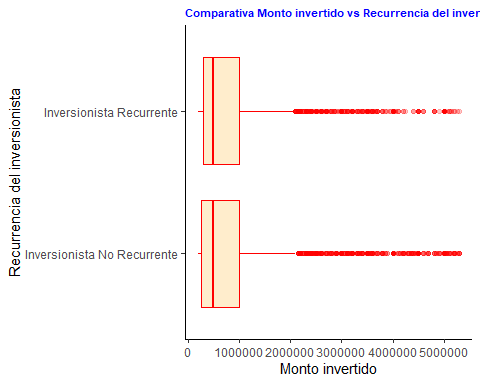
ggplot(data = BDa2censo, mapping = aes(x = CantidadInversiones , y = GrupoIngresos)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa cantidad de inversiones vs grupo de ingresos") + ylab("Rango de ingresos") + xlab("cantidad de inversiones") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



Si bien hay outliers en todos los rangos de ingresos, desde la tendencia central podemos ver que el mayor foco de inversionistas recurrentes se encuentra en personas con ingresos entre 7 a 10 millones de pesos

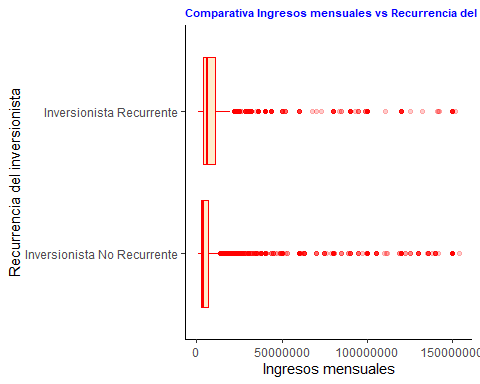
#2. Analisis desde el tipo de inversionista (Recurrente o no Recurrente)

ggplot(data = BDa2censo, mapping = aes(x = MontoInversion , y = CodEsInversionistaRecurrente)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa Monto invertido vs Recurrencia del inversionista") + ylab("Recurrencia del inversionista") + xlab("Monto invertido") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



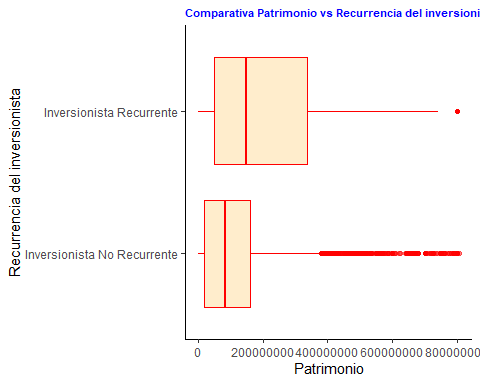
Desde el promedio podemos concluir que los inversionistas NO recurrentes invierten en promedio los mismos montos que inversionistas recurrentes, no obtante la dispersión de los datos no permite encontrar dicha relacion

ggplot(data = BDa2censo, mapping = aes(x = IngresosMes , y = CodEsInversionistaRecurrente)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa Ingresos mensuales vs Recurrencia del inversionista") + ylab("Recurrencia del inversionista") + xlab("Ingresos mensuales") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



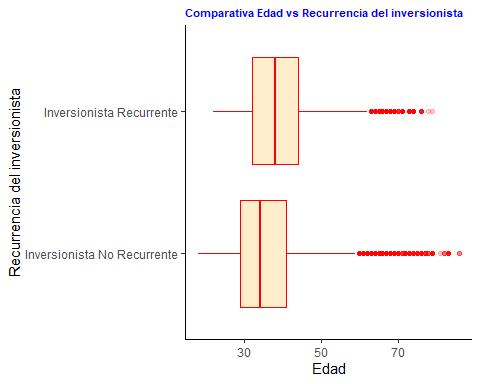
Los inversionistas recurrentes pueden llegar a tener un promedio de ingresos levemente mayor a los inversionistas no recurrentes

ggplot(data = BDa2censo, mapping = aes(x = Patrimonio , y = CodEsInversionistaRecurrente)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa Patrimonio vs Recurrencia del inversionista") + ylab("Recurrencia del inversionista") + xlab("Patrimonio") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



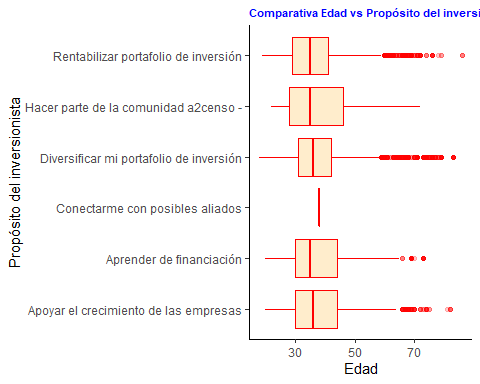
Los inversionistas Recurrentes llegan a tener un promedio de patrimonio superior, aun cuando hay personas consideradas inversionistas no recurrentes con un muy alto nivel de patrimonio

ggplot(data = BDa2censo, mapping = aes(x = Edad , y = CodEsInversionistaRecurrente)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa Edad vs Recurrencia del inversionista") + ylab("Recurrencia del inversionista") + xlab("Edad") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



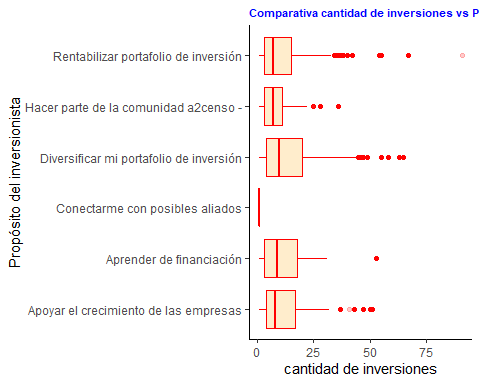
Los inversionistas recurrentes tiene un promedio de edad mayor, lo cual puede estar asociado con la siguiente grafica en donde observamos que en este rango de edad dichas inversiones recurrentes pueden estar asociadas con propositos de inversión como diversificar portafolio, aprender de financiación o sencillamente rentabilidad

ggplot(data = BDa2censo, mapping = aes(x = Edad , y = Propósito)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa Edad vs Propósito del inversionista") + ylab("Propósito del inversionista") + xlab("Edad") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



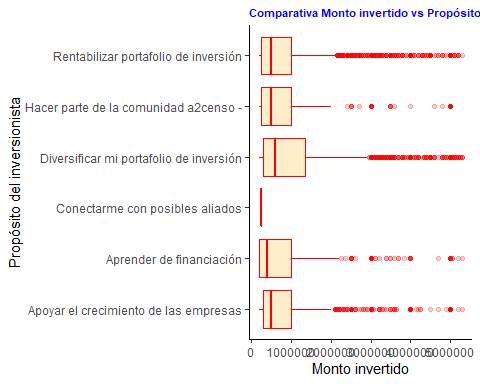
Aqui podemos observar que no se evidencia alguna relación entre la edad y el propósito, es decir que tanto personas jovenes como viejas se distribuyen entre los diferentes propósitos de inversión.

ggplot(data = BDa2censo, mapping = aes(x = CantidadInversiones , y = Propósito)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa cantidad de inversiones vs Propósito del inversionista") + ylab("Propósito del inversionista") + xlab("cantidad de inversiones") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))



Resulta interesante identificar que el promedio de numero de inversiones es superior respecto a propósitos de inversion asociados a la diversificación, por el contrario aquellos que han invertido con objetivos de conexión con otros posibles aliados son en promedio menos cantidades de inversiones

ggplot(data = BDa2censo, mapping = aes(x = MontoInversion , y = Propósito)) + geom\_boxplot(color="red", fill="orange", alpha=0.2) + theme\_classic() + ggtitle("Comparativa Monto invertido vs Propósito del inversionista") + ylab("Propósito del inversionista") + xlab("Monto invertido") + theme(legend.title = element\_text(colour="blue", size=7,face="bold")) + theme(plot.title = element\_text(color="blue", size=8, face="bold"))

 Llama la atención observar que en general el promedio del monto invertido es mayor en personas con propósitos de diversificar el portafolio de inversión, en donde puede haber una mayor posibilidad de profesionalizar al inversionista.

## 3. Resumen Multivariado entre variables Cualitativas y cuantitativas

A continuación se realizará un analisis Multivariado que contemple dos de las variables cualitativas de mayor interés del estudio en cuestión, respecto a las variables cuantitativas evaluadas que permiten entender el comportamiento de inversión. Dichas variables son el monto invertido y la recurrencia de inversión

# Analisis del propósito de inversión a la luz de las variables de interés cualitativas y cuantitativas

library(ggplot2)  
library(dplyr)  
library(tidyr)

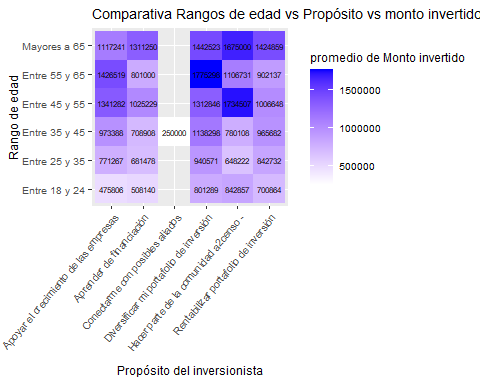
##   
## Attaching package: 'tidyr'

## The following object is masked from 'package:igraph':  
##   
## crossing

df1 <- BDa2censo %>%  
 dplyr::group\_by(Propósito, GrupoEdad)%>%  
 dplyr::summarize(Mean = round(mean(MontoInversion, na.rm=TRUE)))

## `summarise()` has grouped output by 'Propósito'. You can override using the  
## `.groups` argument.

ggplot(df1, aes(Propósito, GrupoEdad, fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=50, hjust=1)) + ggtitle("Comparativa Rangos de edad vs Propósito vs monto invertido") + ylab("Rango de edad") + xlab("Propósito del inversionista") + labs(fill = "promedio de Monto invertido")+  
 geom\_text(aes(label = Mean), color = "black", size = 2)



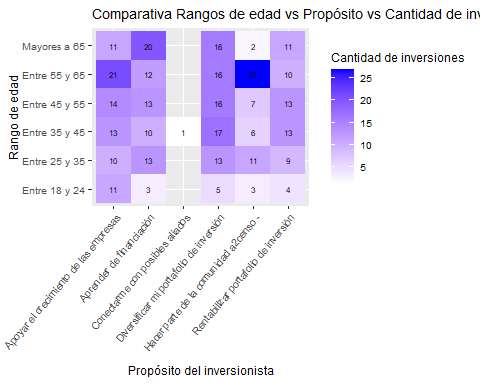
Del anterior grafico podemos observar como de manera general el promedio de monto invertido incrementa con el rango de edad, en donde podemos observar que los mayores montos promedios invertidos concuerdan con un interés de diversificar el portafolio de inversión en personas mayores de 55 años

# Analisis de la cantidad de inversiones realizadas respecto al grupo de edad y el propósito de inversión

df2 <- BDa2censo %>%  
 dplyr::group\_by(Propósito, GrupoEdad)%>%  
 dplyr::summarize(Mean = round(mean(CantidadInversiones, na.rm=TRUE)))

## `summarise()` has grouped output by 'Propósito'. You can override using the  
## `.groups` argument.

ggplot(df2, aes(Propósito, GrupoEdad, fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=50, hjust=1))+ ggtitle("Comparativa Rangos de edad vs Propósito vs Cantidad de inversiones") + ylab("Rango de edad") + xlab("Propósito del inversionista") + labs(fill = "Cantidad de inversiones") +  
 geom\_text(aes(label = Mean), color = "black", size = 2)



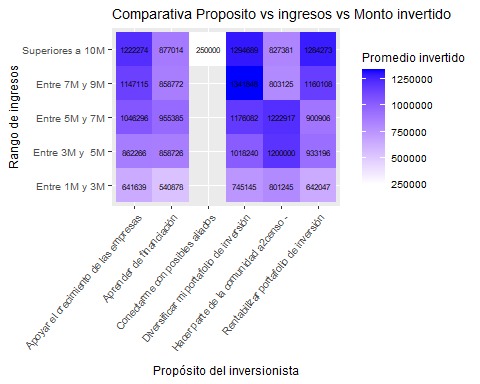
Respecto a la cantidad de inversiones realizadas resulta llamativo observar que personas entre 55 y 65 años tienen un proposito de afiliación a la comunidad de a2censo y en general son las personas con mayor recurrencia de inversiones desde los datos promedio de inversiones realizadas

# Analisis de la cantidad de inversiones realizadas respecto al sector económico de la campaña y el propósito de inversión

df3 <- BDa2censo %>%  
 dplyr::group\_by(Propósito, GrupoIngresos)%>%  
 dplyr::summarize(Mean = round(mean(MontoInversion, na.rm=TRUE)))

## `summarise()` has grouped output by 'Propósito'. You can override using the  
## `.groups` argument.

ggplot(df3, aes(Propósito, GrupoIngresos , fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=50, hjust=1))+ ggtitle("Comparativa Proposito vs ingresos vs Monto invertido") + ylab("Rango de ingresos") + xlab("Propósito del inversionista") + labs(fill = "Promedio invertido") +  
 geom\_text(aes(label = Mean), color = "black", size = 2)

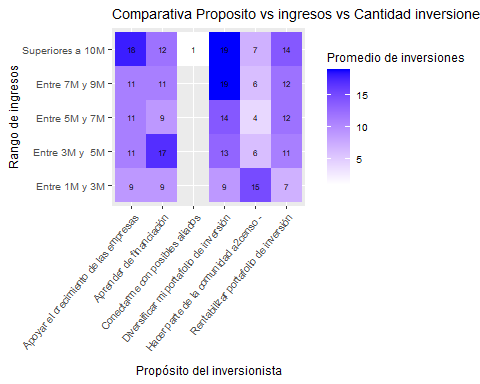


Desde los valores promedio de monto invertido resulta llamativo observar que los montos de inversión incrementan a medida que los rangos salariales tambien incrementan

df4 <- BDa2censo %>%  
 dplyr::group\_by(Propósito, GrupoIngresos)%>%  
 dplyr::summarize(Mean = round(mean(CantidadInversiones, na.rm=TRUE)))

## `summarise()` has grouped output by 'Propósito'. You can override using the  
## `.groups` argument.

ggplot(df4, aes(Propósito, GrupoIngresos , fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=50, hjust=1))+ ggtitle("Comparativa Proposito vs ingresos vs Cantidad inversiones") + ylab("Rango de ingresos") + xlab("Propósito del inversionista") + labs(fill = "Promedio de inversiones") +  
 geom\_text(aes(label = Mean), color = "black", size = 2)

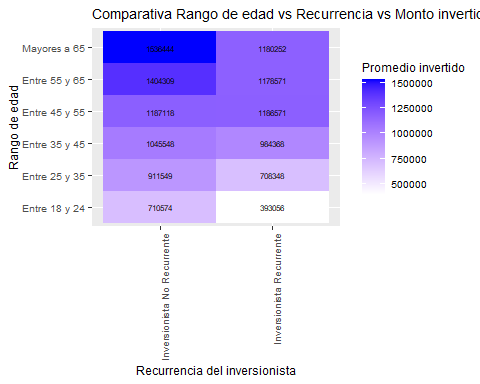
 Por otro lado si se analizan los rangos de ingresos y propósitos a la luz del promedio de inversiones realizadas podemos ver como personas con mayores ingresos y con interés de diversificar el portafolio de inversion son quienes mas inversiones recurrentes realizan. No obstante hay un grupo de personas de bajos ingresos entre 1 y3 millones mensualees con interéses de afiliación a la comunidad de a2censo que estan invirtiendo en promedio 15 veces, demostrando una recurrencia de inversiones de bajos montos

#Analisis desde la recurrencia de inversión a la luz de las demás variables de interés cualitativas y cuantitativas

df5 <- BDa2censo %>%  
 dplyr::group\_by(CodEsInversionistaRecurrente, GrupoEdad)%>%  
 dplyr::summarize(Mean = round(mean(MontoInversion, na.rm=TRUE)))

## `summarise()` has grouped output by 'CodEsInversionistaRecurrente'. You can  
## override using the `.groups` argument.

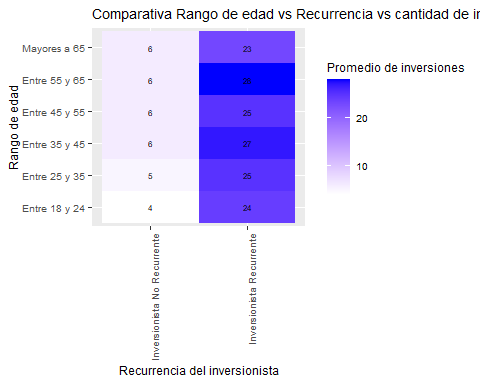
ggplot(df5, aes(CodEsInversionistaRecurrente, GrupoEdad, fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=90, hjust=1)) + ggtitle("Comparativa Rango de edad vs Recurrencia vs Monto invertido") + ylab("Rango de edad") + xlab("Recurrencia del inversionista") + labs(fill = "Promedio invertido") +  
 geom\_text(aes(label = Mean), color = "black", size = 2)

 Podemos encontrar como los montos promedios invertidos suelen ser mayores en inversionistas no recurrentes con rangos de edad superiores a los 55 años

df6 <- BDa2censo %>%  
 dplyr::group\_by(CodEsInversionistaRecurrente, GrupoEdad)%>%  
 dplyr::summarize(Mean = round(mean(CantidadInversiones, na.rm=TRUE)))

## `summarise()` has grouped output by 'CodEsInversionistaRecurrente'. You can  
## override using the `.groups` argument.

ggplot(df6, aes(CodEsInversionistaRecurrente, GrupoEdad, fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=90, hjust=1)) + ggtitle("Comparativa Rango de edad vs Recurrencia vs cantidad de inversiones") + ylab("Rango de edad") + xlab("Recurrencia del inversionista") + labs(fill = "Promedio de inversiones") +  
 geom\_text(aes(label = Mean), color = "black", size = 2)

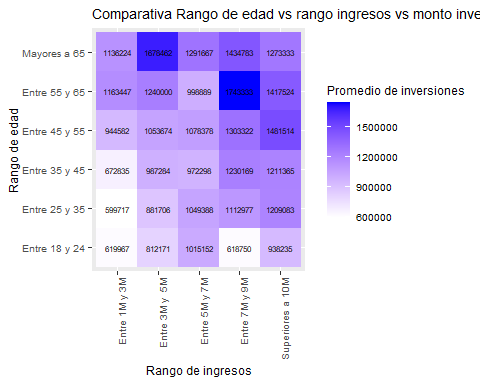
 Podemos encontrar como en inversionistas recurrentes el promedio de inversiones es 6 inversiones respecto a mas o menos 20 inversiones que suelen realizar los inversionistas recurrentes

#Analisis desde los rangos de edad y rangos de ingresos a la luz de las variables cuantitativas de interés

df7 <- BDa2censo %>%  
 dplyr::group\_by(GrupoIngresos, GrupoEdad)%>%  
 dplyr::summarize(Mean = round(mean(MontoInversion, na.rm=TRUE)))

## `summarise()` has grouped output by 'GrupoIngresos'. You can override using the  
## `.groups` argument.

ggplot(df7, aes(GrupoIngresos, GrupoEdad, fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=90, hjust=1)) + ggtitle("Comparativa Rango de edad vs rango ingresos vs monto invertido") + ylab("Rango de edad") + xlab("Rango de ingresos") + labs(fill = "Promedio de inversiones") +  
 geom\_text(aes(label = Mean), color = "black", size = 2)

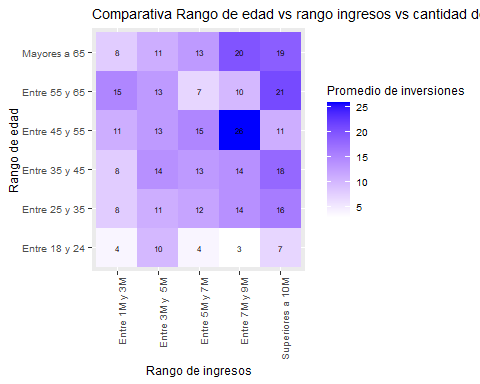
 ahora bien, al realizar el analisis entre rangos de edad y de ingresos podemos observar como los mayores valores promedio de inversión se encuentran en personas con ingresos entre 3y 5 millones de mas de 65 años. personas con ingresos entre 7 y 9 millones con rangos de eedad entre 55 a 65 y las personas con ingresos de mas de 10 millones de pesos.

Analisis

df8 <- BDa2censo %>%  
 dplyr::group\_by(GrupoIngresos, GrupoEdad)%>%  
 dplyr::summarize(Mean = round(mean(CantidadInversiones, na.rm=TRUE)))

## `summarise()` has grouped output by 'GrupoIngresos'. You can override using the  
## `.groups` argument.

ggplot(df8, aes(GrupoIngresos, GrupoEdad, fill= Mean)) + geom\_tile() +  
 scale\_fill\_gradient(low="white", high="blue") + theme(text = element\_text(size=9), axis.text.x = element\_text(angle=90, hjust=1)) + ggtitle("Comparativa Rango de edad vs rango ingresos vs cantidad de inversiones") + ylab("Rango de edad") + xlab("Rango de ingresos") + labs(fill = "Promedio de inversiones") +  
 geom\_text(aes(label = Mean), color = "black", size = 2)



Finalmente podemos observar que de acuerdo al valor promedio de inversiones realizads, los grupos con mayor recurrencia de inversión se encuentra en las edades entre 45 y 55 años con ingresos entre 7 y 9 millones. Por otro lado personas con ingresos superiores a 10 MM suelen invertir de manera más recurrente.

# 4. Resumen Bivariado Cuantitativas - Regresiones Lineales simples, multiples y regresión logistica multiple

## Correlaciones Lineales simples entre las variables cuantitativas y su influencia en el monto invertido

cor(x = BDa2censo$Edad, y = BDa2censo$MontoInversion, method = "pearson")

## [1] 0.144479

cor(x = BDa2censo$IngresosMes, y = BDa2censo$MontoInversion, method = "pearson")

## [1] 0.09333871

cor(x = BDa2censo$Patrimonio, y = BDa2censo$MontoInversion, method = "pearson")

## [1] 0.2583145

cor(x = BDa2censo$CantidadInversiones, y = BDa2censo$MontoInversion, method = "pearson")

## [1] -0.02984937

#cor(x = BDa2censo$CodProposito, y = BDa2censo$MontoInversion, method = "pearson")  
  
#cor(x = BDa2censo$CodProposito, y = BDa2censo$CantidadInversiones, method = "pearson")  
  
#cor(x = BDa2censo$PotencialP2, y = BDa2censo$Edad, method = "pearson")

Se encuentra que desde una correlacion lineal simple no hay correlación entre ninguna de las variables a la luz de la variable monto inversión.

## Regresion Lineales simples entre las variables cuantitativas y su influencia en la cantidad de inversiones

cor(x = BDa2censo$Edad, y = BDa2censo$CantidadInversiones, method = "pearson")

## [1] 0.1549182

cor(x = BDa2censo$IngresosMes, y = BDa2censo$CantidadInversiones, method = "pearson")

## [1] 0.09968228

cor(x = BDa2censo$Patrimonio, y = BDa2censo$CantidadInversiones, method = "pearson")

## [1] 0.2861183

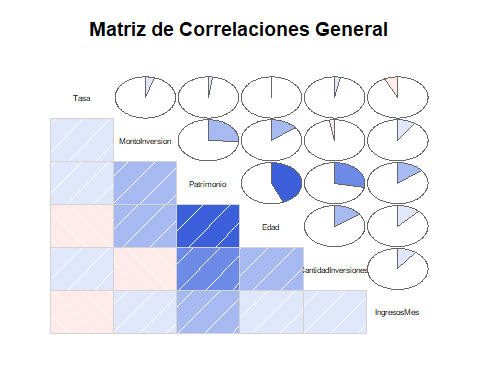
cor(x = BDa2censo$MontoInversion, y = BDa2censo$CantidadInversiones, method = "pearson")

## [1] -0.02984937

Se encuentra que desde una correlacion lineal simple no hay correlación entre ninguna de las variables a la luz de la variable cantidad de inversiones.

#Grafico General de Correlaciones lineales

corrgram(BDa2censo, order=TRUE, lower.panel=panel.shade,  
 upper.panel=panel.pie, text.panel=panel.txt,  
 main="Matriz de Correlaciones General")

 En el anterior grafico podemos observar que aunque dichas correlaciones no son significativas la relación entre las variables Ingresos mes y patrimonio y la relación entre las variables patrimonio y edad son más fuertes respecto a las demás.

# Paso Cero:

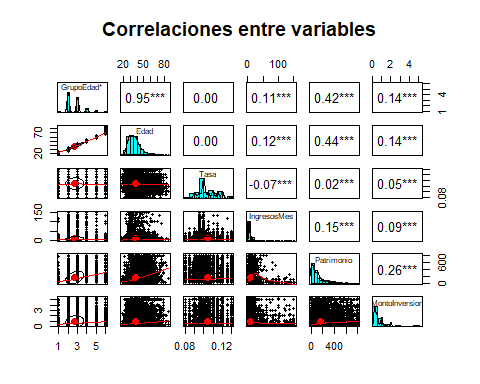
Para facilidad de interpretación de las gráficas se pretende manejar una escala lo más similar posible por lo tanto se transformarán las variables de: ingreso mes , monto de inversión y patrimonio a millones dividiendo en un millón.

BDa2censo$IngresosMes <- BDa2censo$IngresosMes / 1000000  
BDa2censo$Patrimonio <- BDa2censo$Patrimonio / 1000000  
BDa2censo$MontoInversion <- BDa2censo$MontoInversion / 1000000

# Paso 1 :

Realizamos un diagrama de las variables cuantitativas con sus distribuciones y correlaciones, para identificar magnitud y sentido de la relación entre ellas

pairs.panels(BDa2censo[,15:20],pch = 20,stars = TRUE, main="Correlaciones entre variables")



Luego de analizar las diferentes correlaciones entre variables cuantitativas identificamos que la mayor correlación se presenta entre el patrimonio y la edad (0.44) le sigue la correlación que se da entre la cantidad de inversiones con el patrimonio (0.29 ) y la siguiente es la correlación que se da entre monto de inversión y patrimonio con un 0.26

Dentro de los objetivos que se han planteado en este proyecto se pretende encontrar un modelo que pueda predecir el valor del monto de inversión dado las variables dependientes con las que se cuenta. Viendo que el monto de inversión presenta una correlación alrededor del 0.26 (aunque es una correlación moderada), plantearemos preliminarmente un modelo de regresión lineal simple que relacione precisamente estas dos variables patrimonio y monto de inversión.

# Paso 2:

Modelo de Regresion Lineal Simple

modeloSimple <- lm(MontoInversion ~ Patrimonio, data = BDa2censo)  
summary(modeloSimple)

##   
## Call:  
## lm(formula = MontoInversion ~ Patrimonio, data = BDa2censo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7563 -0.5720 -0.3720 0.2126 4.5592   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.72589296 0.00856002 84.80 <0.0000000000000002 \*\*\*  
## Patrimonio 0.00153797 0.00003587 42.88 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.034 on 25718 degrees of freedom  
## Multiple R-squared: 0.06673, Adjusted R-squared: 0.06669   
## F-statistic: 1839 on 1 and 25718 DF, p-value: < 0.00000000000000022

modeloSimple$coefficients

## (Intercept) Patrimonio   
## 0.725892957 0.001537974

El modelos obtenido tiene la forma:

MontoInversion = 0.72 + Patrimonio\*0.0015 Evidenciamos que aunque tanto el coeficiente como el intercepto resultan ser significativos dentro del análisis, el R2 ajustado que nos arroja este modelo es muy bajo (6.6%). Es decir que solo el 6.6% de porcentaje de variación en la variable de respuesta ( monto inversion) es explicado por su relación con la variable predictora (Patrimonio)

Dado a este bajo desempeño del modelo de regresión lineal simple, buscaremos alternativas con un modelo de regresión múltiple incluyendo las otras variables o información disponible dentro de nuestra data. Intentaremos encontrar o predecir el monto de la inversión incluyendo la relación que puede existir con la edad del inversionista, con los ingresos mensuales, con la tasa a la cual se pacta el retorno de su inversión y otras variables que están bajo el análisis. Esperamos que con esto el desempeño del modelo sea mucho mejor.

# paso 3:

Modelo de Regresión Lienal Multiple

modeloMultiple1 <- lm(MontoInversion ~ Edad +Tasa+IngresosMes+Patrimonio+CantidadInversiones, data = BDa2censo )  
summary(modeloMultiple1)

##   
## Call:  
## lm(formula = MontoInversion ~ Edad + Tasa + IngresosMes + Patrimonio +   
## CantidadInversiones, data = BDa2censo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1659 -0.5821 -0.3297 0.2066 4.5504   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.08695033 0.07034580 1.236 0.216   
## Edad 0.00433078 0.00072268 5.993 0.00000000209 \*\*\*  
## Tasa 5.31433292 0.62444368 8.511 < 0.0000000000000002 \*\*\*  
## IngresosMes 0.00584545 0.00054045 10.816 < 0.0000000000000002 \*\*\*  
## Patrimonio 0.00157547 0.00004086 38.558 < 0.0000000000000002 \*\*\*  
## CantidadInversiones -0.01042623 0.00054229 -19.226 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.023 on 25714 degrees of freedom  
## Multiple R-squared: 0.08625, Adjusted R-squared: 0.08607   
## F-statistic: 485.4 on 5 and 25714 DF, p-value: < 0.00000000000000022

modeloMultiple1$coefficients

## (Intercept) Edad Tasa IngresosMes   
## 0.086950331 0.004330777 5.314332923 0.005845453   
## Patrimonio CantidadInversiones   
## 0.001575472 -0.010426229

El modelo de regresion multiple obtenido tiene la forma:

MontoInversion = 0.086 + Edad*0.004 + Tasa*5.31 + IngresosMes *0.0058 + Patrimonio*0.0015 - CantidadInversiones\*0.0104

Luego del análisis, podemos evidenciar que tanto el intercepto como los coeficientes de las variables predictoras: edad, tasa.ingresos mes, patrimonio, cantidad de inversiones son significativos ( a un nivel de significancia del 0.5%) la medida de desempeño R2ajustado sigue siendo muy bajo. Para este caso del modelo de regresión múltiple nos dio un 8.6% aproximadamente ( Solo el 8.6% de la variabilidad en los datos se puede explicar por el modelo). Es una leve mejoría con el modelo presentado anteriormente, pero va en contra del principio de parsimonia ya que hemos agregado y cuatro variables adicionales y la mejora fue prácticamente insignificante.

Propendiendo por el principio de parsimonia aplicaremos el algoritmo step para encontrar quizás una mejor combinación de variables que nos permitan un mejor índice de pérdida de información o AIC para un modelo de regresión múltiple

Luego de correr el algoritmo evidenciamos que el modelo que nos propone con menor pérdida de información AIC 1182.07 es el modelo que previamente hemos definido (modeloMultiple1), que incluye todas las variables predictoras cuantitativas

step(object = modeloMultiple1, direction = "both", trace = 1)

## Start: AIC=1182.07  
## MontoInversion ~ Edad + Tasa + IngresosMes + Patrimonio + CantidadInversiones  
##   
## Df Sum of Sq RSS AIC  
## <none> 26917 1182.1  
## - Edad 1 37.59 26955 1216.0  
## - Tasa 1 75.82 26993 1252.4  
## - IngresosMes 1 122.46 27040 1296.8  
## - CantidadInversiones 1 386.95 27304 1547.2  
## - Patrimonio 1 1556.24 28473 2625.7

##   
## Call:  
## lm(formula = MontoInversion ~ Edad + Tasa + IngresosMes + Patrimonio +   
## CantidadInversiones, data = BDa2censo)  
##   
## Coefficients:  
## (Intercept) Edad Tasa   
## 0.086950 0.004331 5.314333   
## IngresosMes Patrimonio CantidadInversiones   
## 0.005845 0.001575 -0.010426

#Paso4 Validacion de Supuestos Es muy importante al momento de realizar un modelo de regresión lineal validar el cumplimiento de los supuestos ya que podemos ya que sin esta validación podemos caer en la estimación de variables no confiables

## 1. Multicolinealidad

Los predictores deben ser independientes,no debe de haber colinialidad entre ellos , para validar esto calculamos FACTOR DE INFLACION DE LA VARIANZA(VIF)

vif(modeloMultiple1)

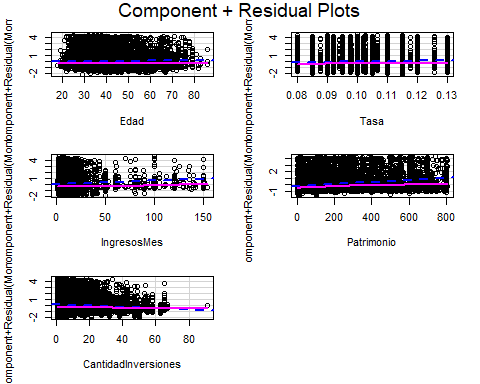
## Edad Tasa IngresosMes Patrimonio   
## 1.241702 1.007372 1.035539 1.325390   
## CantidadInversiones   
## 1.095531

Este resultado para todas las variables predciotres no esta dando ligeramente mayor a 1 , por lo tanto hay un indicio de una posible colinialidad. Es relvante seguir evaluando los demas supuestos, par atebner la certeza q no se cumple y q no se podria usar un modelo de regresion como herramienta predicora.

Para evaluar la influencia que tienen los predictores sobre la variable dependiente

## 2. Relacion Lineal entre predictores y variable respuesta

crPlots(modeloMultiple1)

 Luego de graficar los residuos para cada variable predictora podemos identificar comportamientos aleatorios sobre el valor cero pero de alguna manera desiguales en los sentidos positivo y negativos. La gran mayoría de los residuos para cada variable predictora se encuentran por encima cero, No se cumple que se distribuyan en igual proporción alrededor del cero esto indicaría que este supuesto no se cumpliría

##3. Distribuciónnormaldelosresiduos

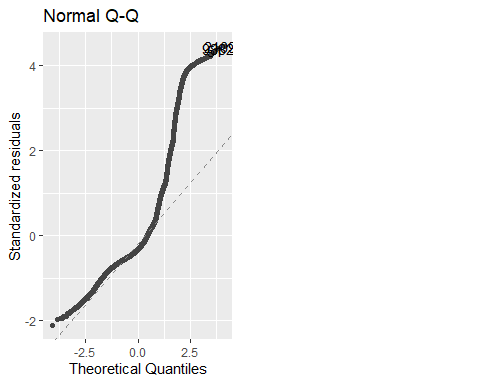
La media de los residuos nos da aproxidamente cero (-1.30e-16) , mas sin embargo la distribucion de los residuos estudentizados presentean una comportamiento anomalo ( gran parte de los residuos superan el limite de 2). Este comporamtiento lo podemos evidenciar en la grafica Normal QQplot donde es calro que los residus no se ajustan sobre la diagonal de los Quantiles teoricos , por lo tanto este supuesto tampoco se satisface.

mean(modeloMultiple1$residuals)

## [1] -0.0000000000000001302145

# Si la relación es lineal,los residuos se distribuyen de forma aleatoria entorno a cero

library(ggfortify)  
autoplot(modeloMultiple1,2) ## Qqplot



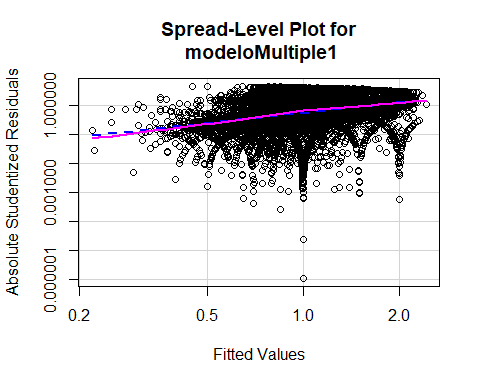
##4.Homocedasticidad

Para valdiar este supuesto acudiremos a realizar el test de Breusch-Pagan, el cual nos arrojo un p-value de 2.2e-16 , lo cual nso permite rechazr la hipotesis nula q existe varibilidad constante de los residuos. Es decir q no se presenta homcedasticidad, imcumpliendo asi el supuesto . Adcionalmente en la grafica podemos observar un incremnto en valor absoluto de los residuos estudentizados, confirmando la heterocedasticidad de los residuos.

library(lmtest)  
bptest(modeloMultiple1) #Breusch-Pagan Test For Homoscedasticity

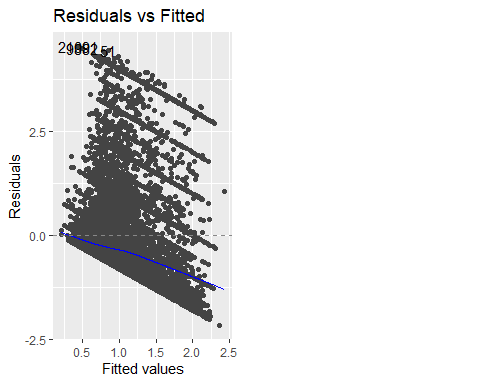
##   
## studentized Breusch-Pagan test  
##   
## data: modeloMultiple1  
## BP = 1108.5, df = 5, p-value < 0.00000000000000022

par(mfrow=c(1,1))  
spreadLevelPlot(modeloMultiple1)



##   
## Suggested power transformation: -0.2003072

autoplot(modeloMultiple1,1)



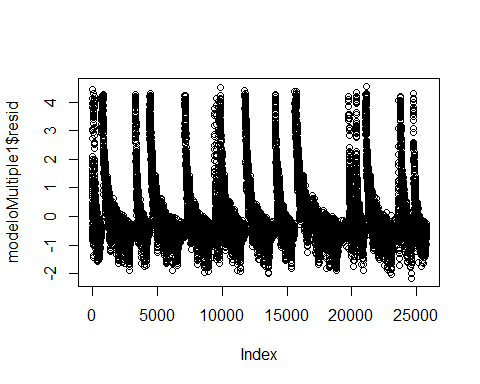
## 5.Independencia

Para validar esto supuesto , realizamos prueba Durbin Watson y evidecmoa q el p-value nos da ceor por lo tanto se rechaza la hipotesis nula , tambien podemos ver graficamente la que si existe autocorrelacion

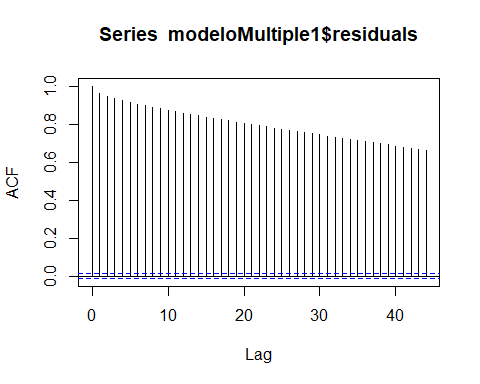
#independencia ()  
durbinWatsonTest(modeloMultiple1) # no cumple test de independencia

## lag Autocorrelation D-W Statistic p-value  
## 1 0.9619126 0.07553598 0  
## Alternative hypothesis: rho != 0

plot(modeloMultiple1$resid) # existe un patron , significa q no pasa la prueba , no son independintes



acf(modeloMultiple1$residuals) # Si usamos una gracica de autocorrelacion envidenciamos q los residuos no se establizand de manera alterna al interio del intervalo de confianza , por lo tanto se peude inferir que existe autocorrelacion

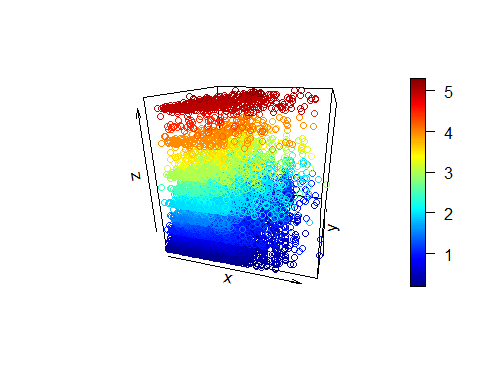


Con animo ilustrativo y para confirmar se realiza una representación 3D de la regresión con solo dos variables predictoras ( edad y patrimonio)

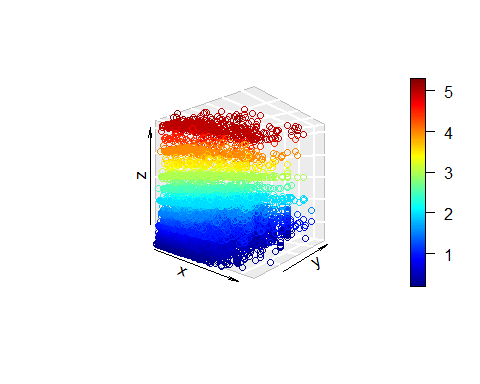
library(plot3D)  
library(plot3Drgl)

## Loading required package: rgl

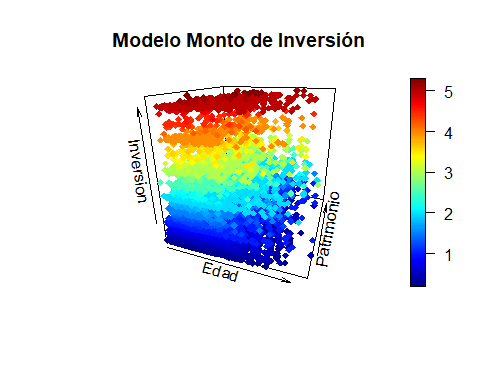
z<-BDa2censo$MontoInversion  
y<-BDa2censo$Patrimonio  
x<-BDa2censo$Edad  
  
scatter3D(x, y, z, theta = 15, phi = 20)



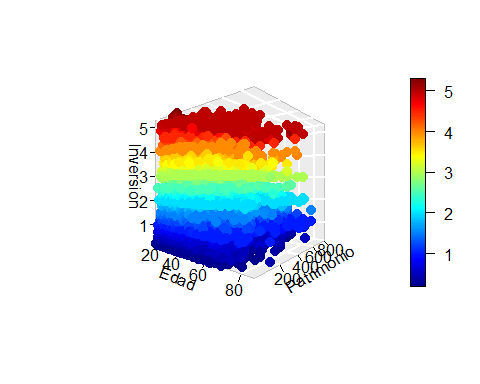
scatter3D(x, y, z, phi = 0, bty ="g")



scatter3D(x, y, z, pch = 18, theta = 20, phi = 20,  
 main = "Modelo Monto de Inversión", xlab = "Edad",  
 ylab ="Patrimonio", zlab = "Inversion")



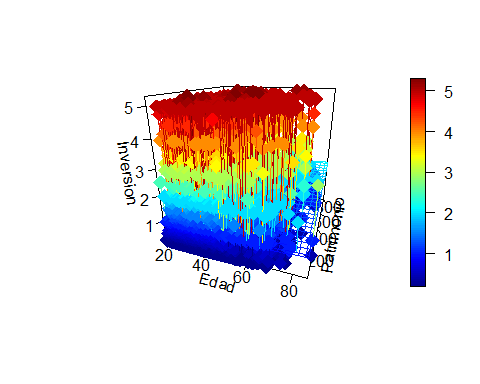
scatter3D(x, y, z, phi = 0, bty = "g",  
 pch = 20, cex = 2, ticktype = "detailed",xlab = "Edad",  
 ylab ="Patrimonio", zlab = "Inversion")



objr<-lm(z ~ x+y)  
objr

##   
## Call:  
## lm(formula = z ~ x + y)  
##   
## Coefficients:  
## (Intercept) x y   
## 0.583237 0.004267 0.001436

grid.lines = 26  
x.pred <- seq(min(x), max(x), length.out = grid.lines)  
y.pred <- seq(min(y), max(y), length.out = grid.lines)  
xy <- expand.grid( x = x.pred, y = y.pred)  
z.pred <- matrix(predict(objr, newdata = xy),   
 nrow = grid.lines, ncol = grid.lines)  
  
# Marcamos las líneas de iteracción para que busquen la recta de regresión  
fitpoints <- predict(objr)  
#ploteamos la gráfica en 3d con recta de regresión  
scatter3D(x, y, z, pch = 18, cex = 2,   
 theta = 20, phi = 20, ticktype = "detailed",  
 xlab = "Edad", ylab = "Patrimonio", zlab = "Inversion",   
 surf = list(x = x.pred, y = y.pred, z = z.pred,   
 facets = NA, fit = fitpoints), main = "")



plotrgl() ### ojo interactivo

## Regresión logistica

De acuerdo a los objetivos planteados para el presente estudio existe un alto interés por parte de la compañía en conocer si existe alguna variable que nos permita predecir la fidelización de los clientes con el producto, entendiendo que la mejor variable que nos puede hablar de fidelización es la recurrencia de inversión. Por esto se analizarán las variables para identificar si alguna de ellas o un conjunto de ellas permiten predecir el comportamiento de inversión referente a la recurrencia.Siendo asi plantearemos un modelo de regresion logisitica que no permita estimar la probabilidad de recurrencia de un inversionista.

BDa2censo$Inv\_Recurrente<-BDa2censo$CodEsInversionistaRecurrente  
  
BDa2censo$Inv\_Recurrente[BDa2censo$Inv\_Recurrente=="Inversionista No Recurrente"]<-0  
BDa2censo$Inv\_Recurrente[BDa2censo$Inv\_Recurrente=="Inversionista Recurrente"]<-1  
BDa2censo$Inv\_Recurrente<-as.numeric(BDa2censo$Inv\_Recurrente)  
BDa2censo$SectorCampaña<-as.factor(BDa2censo$SectorCampaña)  
BDa2censo$Propósito<-as.factor(BDa2censo$Propósito)  
  
  
library(fdth)  
library(caTools)  
library(ROCR)  
library(pROC)  
library(magrittr)

##   
## Attaching package: 'magrittr'

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

library(InformationValue)  
  
set.seed(88)  
split <- sample.split(BDa2censo$Inv\_Recurrente, SplitRatio = 0.75)  
training <- subset(BDa2censo,split =="TRUE")  
testing <- subset(BDa2censo,split=="FALSE")  
  
  
logMultiple=glm(data=training,Inv\_Recurrente~MontoInversion+Edad+IngresosMes+Patrimonio+SectorCampaña+Propósito, family = binomial)  
summary(logMultiple)

##   
## Call:  
## glm(formula = Inv\_Recurrente ~ MontoInversion + Edad + IngresosMes +   
## Patrimonio + SectorCampaña + Propósito, family = binomial,   
## data = training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0912 -0.8618 -0.6968 1.2166 2.1939   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -1.9246331 0.2979440 -6.460  
## MontoInversion -0.2243309 0.0166163 -13.501  
## Edad 0.0173327 0.0017639 9.826  
## IngresosMes 0.0092231 0.0013452 6.856  
## Patrimonio 0.0022523 0.0001008 22.348  
## SectorCampañaAlcantarillado 0.2878625 0.3045023 0.945  
## SectorCampañaAlojamiento Y Servicios De Comida -0.0295437 0.2920597 -0.101  
## SectorCampañaAutomotriz 0.5099706 0.2968569 1.718  
## SectorCampañaComercio al por mayor 0.3723162 0.2912361 1.278  
## SectorCampañaComercio Al Por Menor 0.0446268 0.2923675 0.153  
## SectorCampañaEducación 1.0029118 0.3505771 2.861  
## SectorCampañaEntretenimiento 0.5914717 0.3454151 1.712  
## SectorCampañaIndustrías Creativas Y Culturales 0.6087700 0.3732196 1.631  
## SectorCampañaInformación y Comunicaciones 0.7096011 0.2925527 2.426  
## SectorCampañaInmobiliarias -0.0060257 0.2926296 -0.021  
## SectorCampañaInvestigación y Ciencias 0.1011636 0.2947185 0.343  
## SectorCampañaManufactura 0.4045251 0.2898151 1.396  
## SectorCampañaSalud 0.4026100 0.3058514 1.316  
## SectorCampañaSaneamiento Ambiental -0.2091516 0.3053564 -0.685  
## SectorCampañaServicios Administrativos -0.2422360 0.2926781 -0.828  
## SectorCampañaServicios Domésticos 0.8118238 0.3515981 2.309  
## SectorCampañaServicios Energéticos 0.2204440 0.2981431 0.739  
## SectorCampañaTransporte y Almacenamiento 0.1203064 0.3000769 0.401  
## PropósitoAprender de financiación 0.1616005 0.0884718 1.827  
## PropósitoDiversificar mi portafolio de inversión 0.3406143 0.0533117 6.389  
## PropósitoHacer parte de la comunidad a2censo - -0.4472824 0.1410949 -3.170  
## PropósitoRentabilizar portafolio de inversión -0.0382445 0.0525907 -0.727  
## Pr(>|z|)   
## (Intercept) 0.00000000010490 \*\*\*  
## MontoInversion < 0.0000000000000002 \*\*\*  
## Edad < 0.0000000000000002 \*\*\*  
## IngresosMes 0.00000000000707 \*\*\*  
## Patrimonio < 0.0000000000000002 \*\*\*  
## SectorCampañaAlcantarillado 0.34448   
## SectorCampañaAlojamiento Y Servicios De Comida 0.91943   
## SectorCampañaAutomotriz 0.08581 .   
## SectorCampañaComercio al por mayor 0.20111   
## SectorCampañaComercio Al Por Menor 0.87868   
## SectorCampañaEducación 0.00423 \*\*   
## SectorCampañaEntretenimiento 0.08683 .   
## SectorCampañaIndustrías Creativas Y Culturales 0.10286   
## SectorCampañaInformación y Comunicaciones 0.01529 \*   
## SectorCampañaInmobiliarias 0.98357   
## SectorCampañaInvestigación y Ciencias 0.73141   
## SectorCampañaManufactura 0.16277   
## SectorCampañaSalud 0.18805   
## SectorCampañaSaneamiento Ambiental 0.49338   
## SectorCampañaServicios Administrativos 0.40787   
## SectorCampañaServicios Domésticos 0.02095 \*   
## SectorCampañaServicios Energéticos 0.45967   
## SectorCampañaTransporte y Almacenamiento 0.68848   
## PropósitoAprender de financiación 0.06776 .   
## PropósitoDiversificar mi portafolio de inversión 0.00000000016686 \*\*\*  
## PropósitoHacer parte de la comunidad a2censo - 0.00152 \*\*   
## PropósitoRentabilizar portafolio de inversión 0.46710   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 24421 on 19289 degrees of freedom  
## Residual deviance: 22772 on 19263 degrees of freedom  
## AIC: 22826  
##   
## Number of Fisher Scoring iterations: 4

Desde el modelo de regresion logistica ajustado se encuentra que el sector económico de la campaña es una variable con un nivel de significancia en la variable respuesta de recurrencia del inversionista. De la misma forma en que muestra significancia respecto a los propósitos de inversión de diversificacion del portafolio y el motivador de hacer crecer el capital.

De otro lado, las variables, monto inversión, edad, patrimonio e ingresos mes tambien muestran significancia respecto a la variable asociada a si el individuo es inversionista recurrente o no.

Probaremos con otros modelos descartando algunas variables con el animo de comparar el criterio de perdida de informacion AIC y escoger el mejor modelo

logMultiple1=glm(data=training,Inv\_Recurrente~MontoInversion+Edad+IngresosMes+Patrimonio, family = binomial)  
summary(logMultiple1)

##   
## Call:  
## glm(formula = Inv\_Recurrente ~ MontoInversion + Edad + IngresosMes +   
## Patrimonio, family = binomial, data = training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8052 -0.8464 -0.7460 1.2711 2.1235   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.57829952 0.06327048 -24.945 < 0.0000000000000002 \*\*\*  
## MontoInversion -0.21313284 0.01633425 -13.048 < 0.0000000000000002 \*\*\*  
## Edad 0.01583860 0.00172735 9.169 < 0.0000000000000002 \*\*\*  
## IngresosMes 0.00831926 0.00125744 6.616 0.0000000000369 \*\*\*  
## Patrimonio 0.00242873 0.00009816 24.744 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 24421 on 19289 degrees of freedom  
## Residual deviance: 23218 on 19285 degrees of freedom  
## AIC: 23228  
##   
## Number of Fisher Scoring iterations: 4

logMultiple2=glm(data=training,Inv\_Recurrente~MontoInversion+Edad+IngresosMes+Patrimonio, family = binomial)  
summary(logMultiple2)

##   
## Call:  
## glm(formula = Inv\_Recurrente ~ MontoInversion + Edad + IngresosMes +   
## Patrimonio, family = binomial, data = training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8052 -0.8464 -0.7460 1.2711 2.1235   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.57829952 0.06327048 -24.945 < 0.0000000000000002 \*\*\*  
## MontoInversion -0.21313284 0.01633425 -13.048 < 0.0000000000000002 \*\*\*  
## Edad 0.01583860 0.00172735 9.169 < 0.0000000000000002 \*\*\*  
## IngresosMes 0.00831926 0.00125744 6.616 0.0000000000369 \*\*\*  
## Patrimonio 0.00242873 0.00009816 24.744 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 24421 on 19289 degrees of freedom  
## Residual deviance: 23218 on 19285 degrees of freedom  
## AIC: 23228  
##   
## Number of Fisher Scoring iterations: 4

De acuerdo a lo anterior encontramos que el modelo en donde menor pérdida de información es el designado con nombre: logMultiple , en donde el AIC es de 22826. Asociado a las variables MontoInversion + Edad + IngresosMes + Patrimonio + SectorCampaña + Propósito . No obstatne dentro de las variables cualitativas no todas sus categorías tienen un nivel de significancia relevante, por ejemplo para la variable sector de campaña, las categorías de variables con mayor nivel de significancia son: Educación, Información y Comunicación, y Servicios Domésticos. Para el caso del propósito se destacan: Diversificar Portafolio y Pertenecer a la comunidad de a2censo.

## Resultando un modelo de la forma:

Inv\_Recurrente = e (-1,925-0,0000002243*MontoInversion+0,01733*Edad+0,000000009223*IngresosMes+0,000000002252*Patrimonio+0,2879*SectorCampañaAlcantarillado-0,02954*SectorCampañaAlojamiento Y Servicios De Comida+0,51*SectorCampañaAutomotriz+0,3723*SectorCampañaComercio al por mayor+0,04463*SectorCampañaComercio Al Por Menor+1,003*SectorCampañaEducación+0,5915*SectorCampañaEntretenimiento+0,6088*SectorCampañaIndustrías Creativas Y Culturales+0,7096*SectorCampañaInformación y Comunicaciones-0,006026*SectorCampañaInmobiliarias+0,1012*SectorCampañaInvestigación y Ciencias+0,4045*SectorCampañaManufactura+0,4026*SectorCampañaSalud-0,2092*SectorCampañaSaneamiento Ambiental-0,2422*SectorCampañaServicios Administrativos+0,8118*SectorCampañaServicios Domésticos+0,2204*SectorCampañaServicios Energéticos+0,1203*SectorCampañaTransporte y Almacenamiento+0,1616*PropósitoAprender de financiación+0,3406*PropósitoDiversificar mi portafolio de inversión-0,4473*PropósitoHacer parte de la comunidad a2censo-0,03824*PropósitoRentabilizar portafolio de inversión) / (1+e(-1,925-0,0000002243*MontoInversion+0,01733*Edad+0,000000009223*IngresosMes+0,000000002252*Patrimonio+0,2879*SectorCampañaAlcantarillado-0,02954*SectorCampañaAlojamiento Y Servicios De Comida+0,51*SectorCampañaAutomotriz+0,3723*SectorCampañaComercio al por mayor+0,04463*SectorCampañaComercio Al Por Menor+1,003*SectorCampañaEducación+0,5915*SectorCampañaEntretenimiento+0,6088*SectorCampañaIndustrías Creativas Y Culturales+0,7096*SectorCampañaInformación y Comunicaciones-0,006026*SectorCampañaInmobiliarias+0,1012*SectorCampañaInvestigación y Ciencias+0,4045*SectorCampañaManufactura+0,4026*SectorCampañaSalud-0,2092*SectorCampañaSaneamiento Ambiental-0,2422*SectorCampañaServicios Administrativos+0,8118*SectorCampañaServicios Domésticos+0,2204*SectorCampañaServicios Energéticos+0,1203*SectorCampañaTransporte y Almacenamiento+0,1616*PropósitoAprender de financiación+0,3406*PropósitoDiversificar mi portafolio de inversión-0,4473*PropósitoHacer parte de la comunidad a2censo-0,03824*PropósitoRentabilizar portafolio de inversión))

## Predicciones

# Elaboramos las respectivas predicciones de acaurdo al modelo escogido para evalaur el desempeño del mismo

predict <- predict(logMultiple, type = 'response')  
predict2 <- as.data.frame(predict(logMultiple, type = 'response'))

Tambien definimos el valor del punto de corte o treshold para determinar la clasificaicon de la variable respuesta, con la funcion optCutOff, q nos permite determinar de manera automatica el punto de corte para la clasificacio de las observaciones del modelo propuesto.

optCutOff <- optimalCutoff(training$Inv\_Recurrente, predict)[1]  
  
optCutOff

## [1] 0.4576869

## matriz de confusion

Por medio de la siguente matriz de confusion podemos evalaur el ajsute del modelo , tomado como criterio de corte el valor estimado 0.457 en el paso anterior.

confusionMatrix(training$Inv\_Recurrente, predict, threshold = optCutOff)

## 0 1  
## 0 11540 4606  
## 1 1416 1728

## Sensibilidad y especificidad

sensitivity(training$Inv\_Recurrente, predict, threshold = optCutOff)

## [1] 0.2728134

Se concluye que el modelo acertó en el 27% de las observaciones de verdaderos positivos ( Grado de Sensibilidad), es decir solo identifica un 27% de los positivos reales

specificity(training$Inv\_Recurrente,predict, threshold = optCutOff)

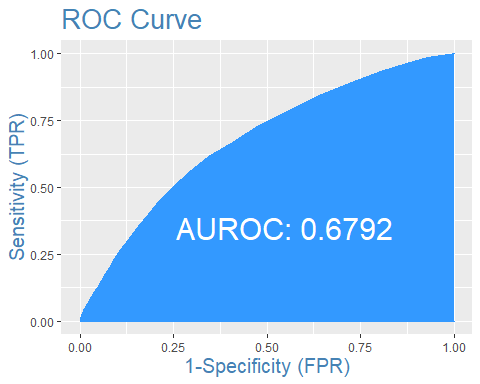
## [1] 0.890707

La especificidad mide la tasa de verdadero negativos, es decir para este caso el modelo identifica el 89% de los negativos reales

## Curva ROC

La curva característica operativa o curva ROC permite identificar el rendimiento de un clasificador, en este caso el modelo de regresión logística. El área bajo la curva mide la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos (1-Especificidad) en varios valores de umbral

plotROC(training$Inv\_Recurrente, predict)



El criterio de precision del modelo concluye que el 68% de las observaciones son acertadas. Y asu vez son evidencia de un modelos q no presta un nivel adecudo para la toma de decisiones.

A pesar de haber escogido un modelos con el menor criterio de pérdida de informacion aún el nivel de precision no es suficiente para tomar decisiones fiables en el futuro.

# Muestreo

N=nrow(BDa2censo)  
n=500  
set.seed(123)  
masBD <- BDa2censo [ sample (N, size = n ),]  
  
  
MediaEdad=round(mean(masBD$Edad),2)  
MediaEdad

## [1] 37.04

VarEdad=var(masBD$Edad)  
VarEdad

## [1] 92.59374

VarEstim=round(((1-n/N)\*VarEdad/n),2)  
VarEstim

## [1] 0.18

EE=round(sqrt(VarEstim),2)  
EE

## [1] 0.42

CV=round((EE/MediaEdad)\*100,2)  
CV

## [1] 1.13

valort=qt(c(0.025),df=(n-1),lower.tail = FALSE)# probabilidad de cola 0.025 equivle a nivel de confianza del 95%, ya q la dist t es simetrica  
  
Lsup=round(MediaEdad+(valort\*EE),2)   
Linf=round(MediaEdad-(valort\*EE),2)  
  
  
round(c(VarEdad,Linf),digit=2)

## [1] 92.59 36.21

resumenMediaEdad <- data.frame(n ,MediaEdad,VarEstim,EE,Linf,Lsup,CV)  
resumenMediaEdad

## n MediaEdad VarEstim EE Linf Lsup CV  
## 1 500 37.04 0.18 0.42 36.21 37.87 1.13

mean(BDa2censo$Edad)

## [1] 37.18177

######  
  
MediaInversion=mean(masBD$MontoInversion)  
VarMuestraInv =var(masBD$MontoInversion)  
VarEstimadaMediaInv=((N-n)/N)\*(VarMuestraInv/n)  
  
  
  
VarEstimInv=round(((1-n/N)\*VarEstimadaMediaInv/n),2)  
VarEstimInv

## [1] 0

EEmediaInv=round(sqrt(VarEstimInv),2)  
EEmediaInv

## [1] 0

CVmediaInv=round((EEmediaInv/MediaInversion)\*100,2)  
CVmediaInv

## [1] 0

valort=qt(c(0.025),df=(n-1),lower.tail = FALSE)# probabilidad de cola 0.025 equivle a nivel de confianza del 95%, ya q la dist t es simetrica  
  
LsupMediaInv=round(MediaInversion+(valort\*EEmediaInv),2)   
LinfMediaInv=round(MediaInversion-(valort\*EEmediaInv),2)  
  
  
  
resumenMediaInversion <- data.frame(n ,MediaInversion,VarEstimInv,EEmediaInv,LinfMediaInv,LsupMediaInv,CVmediaInv)  
resumenMediaInversion

## n MediaInversion VarEstimInv EEmediaInv LinfMediaInv LsupMediaInv  
## 1 500 0.9407 0 0 0.94 0.94  
## CVmediaInv  
## 1 0

mean(BDa2censo$MontoInversion)

## [1] 0.9673676

###  
  
MediaInversion=mean(masBD$MontoInversion)  
TotalInversion=N\*MediaInversion  
TotalInversion

## [1] 24194.8

VarMuestra =var(masBD$MontoInversion)  
VarEstimadaMedia=((N-n)/N)\*(VarMuestra/n)  
  
VarEstimadaTotal=N\*N\*VarEstimadaMedia  
EEtotal=sqrt(VarEstimadaTotal)  
  
  
CVtotal=round((EEtotal/TotalInversion)\*100,2)  
CVtotal

## [1] 4.91

LsupTotal=round(TotalInversion+(valort\*EEtotal),2)   
LinfTotal=round(TotalInversion-(valort\*EEtotal),2)  
  
sum(BDa2censo$MontoInversion)

## [1] 24880.69

resumenTotalInversion <- data.frame(n ,TotalInversion,VarEstimadaTotal,EEtotal,LinfTotal,LsupTotal,CVtotal)  
resumenTotalInversion

## n TotalInversion VarEstimadaTotal EEtotal LinfTotal LsupTotal CVtotal  
## 1 500 24194.8 1413995 1189.115 21858.51 26531.09 4.91

####  
  
library(rgdal)

## Loading required package: sp

## Please note that rgdal will be retired by the end of 2023,  
## plan transition to sf/stars/terra functions using GDAL and PROJ  
## at your earliest convenience.  
##   
## rgdal: version: 1.5-32, (SVN revision 1176)  
## Geospatial Data Abstraction Library extensions to R successfully loaded  
## Loaded GDAL runtime: GDAL 3.3.2, released 2021/09/01  
## Path to GDAL shared files: C:/Program Files/R/R-4.2.0/library/rgdal/gdal  
## GDAL binary built with GEOS: TRUE   
## Loaded PROJ runtime: Rel. 7.2.1, January 1st, 2021, [PJ\_VERSION: 721]  
## Path to PROJ shared files: C:/Program Files/R/R-4.2.0/library/rgdal/proj  
## PROJ CDN enabled: FALSE  
## Linking to sp version:1.4-7  
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,  
## use options("rgdal\_show\_exportToProj4\_warnings"="none") before loading sp or rgdal.

##   
## Attaching package: 'rgdal'

## The following object is masked from 'package:fBasics':  
##   
## getDescription

library(sf)

## Linking to GEOS 3.9.1, GDAL 3.3.2, PROJ 7.2.1; sf\_use\_s2() is TRUE

sp\_df <- readOGR(dsn = "MGN2021\_DPTO\_POLITICO", layer = "MGN\_DPTO\_POLITICO")

## Warning in OGRSpatialRef(dsn, layer, morphFromESRI = morphFromESRI, dumpSRS =  
## dumpSRS, : Discarded datum Marco\_Geocentrico\_Nacional\_de\_Referencia in Proj4  
## definition: +proj=longlat +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +no\_defs

## OGR data source with driver: ESRI Shapefile   
## Source: "D:\Scripts\_SQL\ProyectoEstadistica\ProyectoEstadistica\Sergio\MGN2021\_DPTO\_POLITICO", layer: "MGN\_DPTO\_POLITICO"  
## with 33 features  
## It has 9 fields

#head(sp\_df)  
#fix(sp\_df)  
#as.data.frame(sp\_df)  
  
DPTO\_SH="MGN2021\_DPTO\_POLITICO/MGN\_DPTO\_POLITICO.shp"  
  
DPTO\_SH2 <- st\_read(DPTO\_SH)

## Reading layer `MGN\_DPTO\_POLITICO' from data source   
## `D:\Scripts\_SQL\ProyectoEstadistica\ProyectoEstadistica\Sergio\MGN2021\_DPTO\_POLITICO\MGN\_DPTO\_POLITICO.shp'   
## using driver `ESRI Shapefile'  
## Simple feature collection with 33 features and 9 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: -81.73562 ymin: -4.229406 xmax: -66.84722 ymax: 13.39473  
## Geodetic CRS: MAGNA-SIRGAS

library(raster)

##   
## Attaching package: 'raster'

## The following object is masked from 'package:plotly':  
##   
## select

## The following object is masked from 'package:MASS':  
##   
## select

## The following object is masked from 'package:dplyr':  
##   
## select

library(leaflet)

##   
## Attaching package: 'leaflet'

## The following object is masked from 'package:xts':  
##   
## addLegend

library(dplyr)  
library(sf)  
library(tmap)  
library(tigris)

## To enable caching of data, set `options(tigris\_use\_cache = TRUE)`  
## in your R script or .Rprofile.

##   
## Attaching package: 'tigris'

## The following object is masked from 'package:igraph':  
##   
## blocks

library(dplyr)  
InverDepto <- masBD %>%  
 group\_by(Departamento) %>%  
 summarise(promedioEdad = mean(Edad),promedioInver = mean(MontoInversion),Cod\_Dpto)

## `summarise()` has grouped output by 'Departamento'. You can override using the  
## `.groups` argument.

# Para visualizar la base resumida  
ResumenInv=as.data.frame(InverDepto)  
#ResumenInv  
  
  
library(tidyr)  
Etiquetas=unite(ResumenInv, Etiqueta,c(1,2,3), sep = ": ", remove = TRUE)  
Etiquetas=Etiquetas[,1]  
#Etiquetas  
  
  
  
Resumen3=cbind(ResumenInv, Etiquetas )  
#Resumen3  
  
  
#DPTO\_SH2  
  
DPTO\_JOIN <- geo\_join(DPTO\_SH2, Resumen3,"DPTO\_CCDGO", "Cod\_Dpto")

## Warning: We recommend using the dplyr::\*\_join() family of functions instead.

## Warning: `group\_by\_()` was deprecated in dplyr 0.7.0.  
## Please use `group\_by()` instead.  
## See vignette('programming') for more help  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.

##DPTO\_JOIN = as.data.frame(DPTO\_JOIN)  
#DPTO\_JOIN  
  
  
pal <- colorNumeric( palette = "RdYlBu", domain=DPTO\_JOIN$promedioInver) #palette = "YlGnBu" "RdBu" "RdYlBu" "Spectral" "Paired" "PuRd" "RdYlGn"  
  
popup\_sb <- paste0("Promedio de Monto de Inversión: ", as.character(DPTO\_JOIN$promedioInver))  
  
  
#leaflet(sp\_df) %>%  
# addProviderTiles("CartoDB.Positron") %>%  
# #setView(-98.483330, 38.712046, zoom = 4) %>%   
# addPolygons(data = DPTO\_JOIN ,   
# fillColor = ~pal(DPTO\_JOIN$promedioInver),  
# opacity = 1,  
# color = "black",  
# dashArray = "3",fillOpacity = 0.9,  
# highlight = highlightOptions(  
# weight = 1,  
# color = "#666",  
# dashArray = "",  
# fillOpacity = 1,  
# bringToFront = TRUE),  
# label = DPTO\_JOIN$Etiquetas,  
# labelOptions = labelOptions(  
# style = list("font-weight" = "normal", padding = "3px 8px"),  
# textsize = "15px",  
# direction = "auto"))%>%  
# addLegend(pal = pal, values =DPTO\_JOIN$promedioInver, opacity = 0.7, title = NULL,  
# position = "bottomright")