kickstarter_project

Grecia

12/3/2020

Proyecto de kickstarter

Utilizaremos la base del proyecto [http://saumyaray.me/files/Kickstarter%20Data%20Analysis.pdf] para realizar un analisis de lo siguiente:

1.¿La duración de la campaña afecta la tasa de éxito?

2.¿Cuál es la relación entre el objetivo ($) y el éxito de la campaña? y

2a.¿Cuál es la relación entre la distribución de objetivo y el monto recaudado?

Empezaremos por fijar nuestro directorio de trabajo y las librerias utilizadas

```
library(magrittr)
library(lubridate)
library(dplyr)
library(utils)
library(zoo)
library(TTR)
library(forecast)
library(ggplot2)
library(scales)
library(ggrepel)
```

```
#Directorio de trabajo
setwd("~/Desktop/programacion/Python/kickstarter/ks-projects-201801.csv")
ks<-read.csv("ks-projects-201801.csv",sep = ",",na.strings = "")
```

Vemos algunas características de unas variables con la función summary

```
summary(ks)
```

```
##        ID                                 name                 category
##  Min.   :5.971e+03   New EP/Music Development:    41   Product Design: 22314
##  1st Qu.:5.383e+08   Canceled (Canceled)     :    13   Documentary   : 16139
##  Median :1.075e+09   Music Video             :    11   Music         : 15727
##  Mean   :1.075e+09   N/A (Canceled)          :    11   Tabletop Games: 14180
##  3rd Qu.:1.610e+09   Cancelled (Canceled)    :    10   Shorts        : 12357
##  Max.   :2.147e+09   (Other)                 :378571   Video Games   : 11830
##                      NA's                    :     4   (Other)       :286114
##       main_category       currency            deadline           goal
##  Film & Video: 63585   USD    :295365   2014-08-08:   705   Min.   :        0
##  Music       : 51918   GBP    : 34132   2014-08-10:   558   1st Qu.:     2000
##  Publishing  : 39874   EUR    : 17405   2014-08-07:   541   Median :     5200
##  Games       : 35231   CAD    : 14962   2015-05-01:   489   Mean   :    49081
```

```
##  Technology  : 32569   AUD    :  7950   2014-08-09:   477   3rd Qu.:    16000
##  Design      : 30070   SEK    :  1788   2015-07-01:   449   Max.   :100000000
##  (Other)     :125414   (Other):  7059   (Other)   :375442
##                  launched         pledged               state
##  1970-01-01 01:00:00:     7   Min.   :       0   canceled  : 38779
##  2009-09-15 05:56:28:     2   1st Qu.:      30   failed    :197719
##  2010-06-30 17:29:43:     2   Median :     620   live      :  2799
##  2011-02-08 04:29:48:     2   Mean   :    9683   successful:133956
##  2011-02-25 09:58:36:     2   3rd Qu.:    4076   suspended :  1846
##  2011-03-03 17:55:38:     2   Max.   :20338986   undefined :  3562
##  (Other)            :378644
##     backers            country        usd.pledged       usd_pledged_real
##  Min.   :     0.0   US     :292627   Min.   :       0   Min.   :       0
##  1st Qu.:     2.0   GB     : 33672   1st Qu.:      17   1st Qu.:      31
##  Median :    12.0   CA     : 14756   Median :     395   Median :     624
##  Mean   :   105.6   AU     :  7839   Mean   :    7037   Mean   :    9059
##  3rd Qu.:    56.0   DE     :  4171   3rd Qu.:    3034   3rd Qu.:    4050
##  Max.   :219382.0   N,0"   :  3797   Max.   :20338986   Max.   :20338986
##                     (Other): 21799   NA's   :3797
##  usd_goal_real
##  Min.   :        0
##  1st Qu.:     2000
##  Median :     5500
##  Mean   :    45454
##  3rd Qu.:    15500
##  Max.   :166361391
##
```

Limpieza de la base

Inconsistencias de la base:

1. En la base se encuentran unas fechas con formato inconsistente del año de 1970

2.La columna deadline y launched no están en formato de fecha

3.La columna de launched contiene la hora de la fecha, por lo que se removerá la hora para un mejor manejo de la base.

Formato de fecha Cambiamos el formato de la columna launch a fecha y le quitamos la hora

```
#Seleccionamos primero las variables de interés para u nuevo dataframe
ks1<-ks %>%
  select(ID,deadline,launched, state, currency, goal,usd.pledged,usd_pledged_real,usd_goal_real)
#La base se verá de la siguiente manera
head(ks1)
```

```
##           ID   deadline            launched      state currency  goal
## 1 1000002330 2015-10-09 2015-08-11 12:12:28     failed      GBP  1000
## 2 1000003930 2017-11-01 2017-09-02 04:43:57     failed      USD 30000
## 3 1000004038 2013-02-26 2013-01-12 00:20:50     failed      USD 45000
## 4 1000007540 2012-04-16 2012-03-17 03:24:11     failed      USD  5000
## 5 1000011046 2015-08-29 2015-07-04 08:35:03   canceled      USD 19500
## 6 1000014025 2016-04-01 2016-02-26 13:38:27 successful      USD 50000
##   usd.pledged usd_pledged_real usd_goal_real
```

```
## 1           0                 0       1533.95
## 2         100              2421      30000.00
## 3         220               220      45000.00
## 4           1                 1       5000.00
## 5        1283              1283      19500.00
## 6       52375             52375      50000.00
```

```
#cambiamos el formato de las columnas a fecha
ks1$deadline %<>% ymd()
#summary(ks1)

#La columna launched la convertimos a
Lan<- as.POSIXlt(ks1$launched)
#Quitamos la hora
Lan1<-strftime(Lan,format="%Y-%m-%d")
#Convertimos a dataframe el formato
Launch2<-data.frame(Lan1)
#head(Launch2)
#summary(Launch2)
#Convertimos  a formato de fecha la columna de launched
Launch2$Lan1%<>% ymd()
#summary(Launch2)
```

```
#Unimos la columna de launched ya en su formato de fecha  y sin la hora a nuestro dataframe ks1
Launch3<-ks1 %>%
  bind_cols(Launch2)

#Eliminamos las fechas con formato inconsistente y calculamos la diferencia de días de las campaña
ks4<-Launch3%>%
  mutate(Lenght_campaing=difftime(deadline,Lan1)) %>%
  filter(launched!="1970-01-01")

#Convertimos la nueva columna a formato numérico para trabajar con los datos.
ks4$Lenght_campaing<-as.numeric(ks4$Lenght_campaing)
#head(ks4)
#summary(ks4)
```

Categorizamos los dias de campaña por grupos Grupo 1: 0-7 días

Grupo 2: 8-22 días

Grupo 3: 23-37 días

Grupo 4: 38-52 días

Grupo 5: 53-68 días

Grupo 6: 69-82 días

Grupo 7: 83-92 días

```
#categorizar por rangos de acuerdo a los días de campaña
ks5<-ks4 %>%
  filter(Lenght_campaing<=92)

ks5$Lenght_campaing<-cut(ks5$Lenght_campaing,breaks=c(0,8,23,38,53,69,83,92),labels=c("1","2","3","4","5
#La base se ve de la siguiente manera
```

```
head(ks5)
```

```
##           ID   deadline            launched      state currency  goal
## 1 1000002330 2015-10-09 2015-08-11 12:12:28     failed      GBP  1000
## 2 1000003930 2017-11-01 2017-09-02 04:43:57     failed      USD 30000
## 3 1000004038 2013-02-26 2013-01-12 00:20:50     failed      USD 45000
## 4 1000007540 2012-04-16 2012-03-17 03:24:11     failed      USD  5000
## 5 1000011046 2015-08-29 2015-07-04 08:35:03   canceled      USD 19500
## 6 1000014025 2016-04-01 2016-02-26 13:38:27 successful      USD 50000
##   usd.pledged usd_pledged_real usd_goal_real       Lan1 Lenght_campaing
## 1           0                0       1533.95 2015-08-11               5
## 2         100             2421      30000.00 2017-09-02               5
## 3         220              220      45000.00 2013-01-12               4
## 4           1                1       5000.00 2012-03-17               3
## 5        1283             1283      19500.00 2015-07-04               5
## 6       52375            52375      50000.00 2016-02-26               3
```

```
summary(ks5$Lenght_campaing)
```

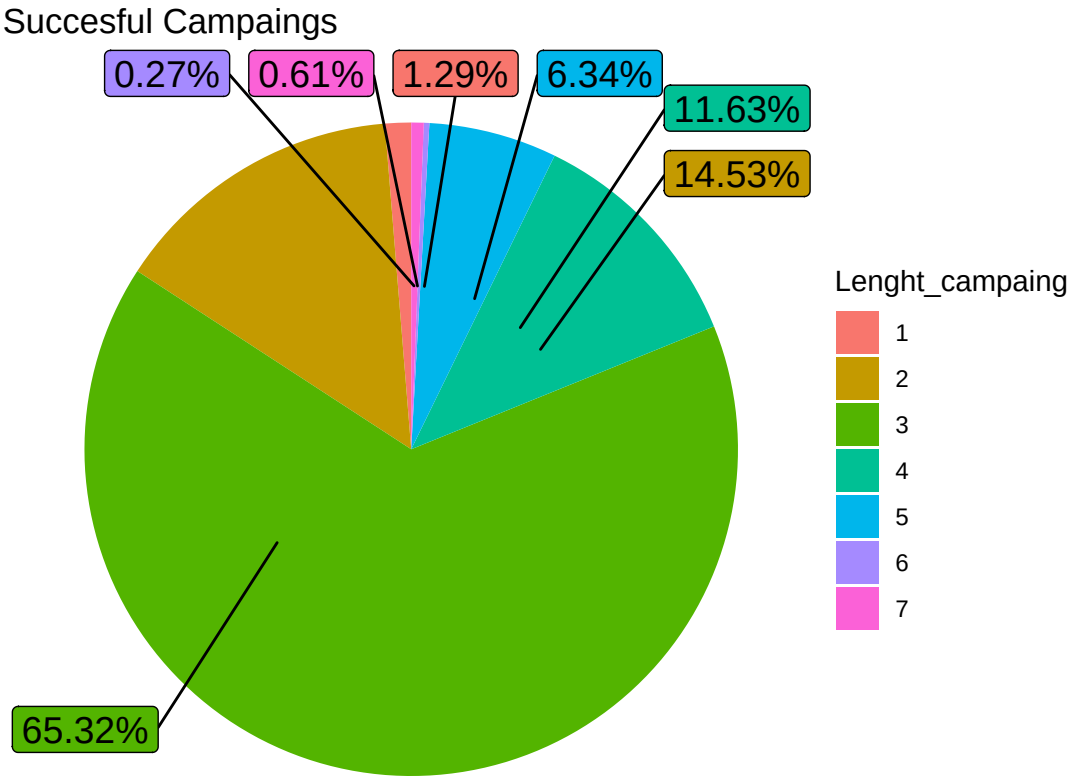
```
##      1      2      3      4      5      6      7
##   3951  39693 243344  46633  41229    941   2863
```

La mayoría de las campañas tienen una duración de 23 a 37 días, mientras que las menos frecuentes, tienen una duración de 69 a 82 días.

¿La duración de la campaña afecta la tasa de éxito?

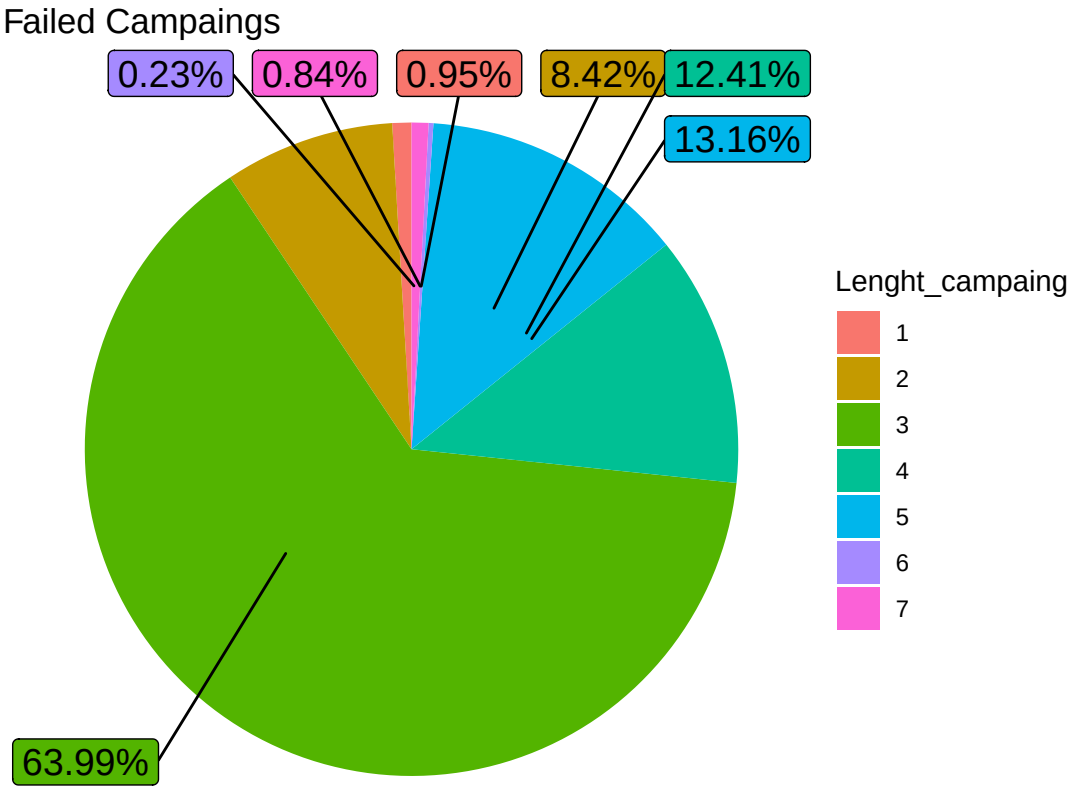
Campañas exitosas

```
piedf<-ks5%>%
  filter(state=="successful")

piedf%>%
  count(Lenght_campaing) %>%
  mutate(prop = percent(n / sum(n)))%>%
ggplot( aes(x="", y=n, fill=Lenght_campaing))+
geom_bar(width = 1, stat = "identity")+
ggtitle("Succesful Campaings") + theme_void() +
coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```



Campañas fallidas

```
pieFailed<-ks5%>%
  filter(state=="failed")

pieFailed%>%
  count(Lenght_campaing) %>%
  mutate(prop = percent(n / sum(n)))%>%
ggplot( aes(x="", y=n, fill=Lenght_campaing))+
geom_bar(width = 1, stat = "identity")+
ggtitle("Failed Campaings") +
theme_void() + coord_polar("y", start=0)+
geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

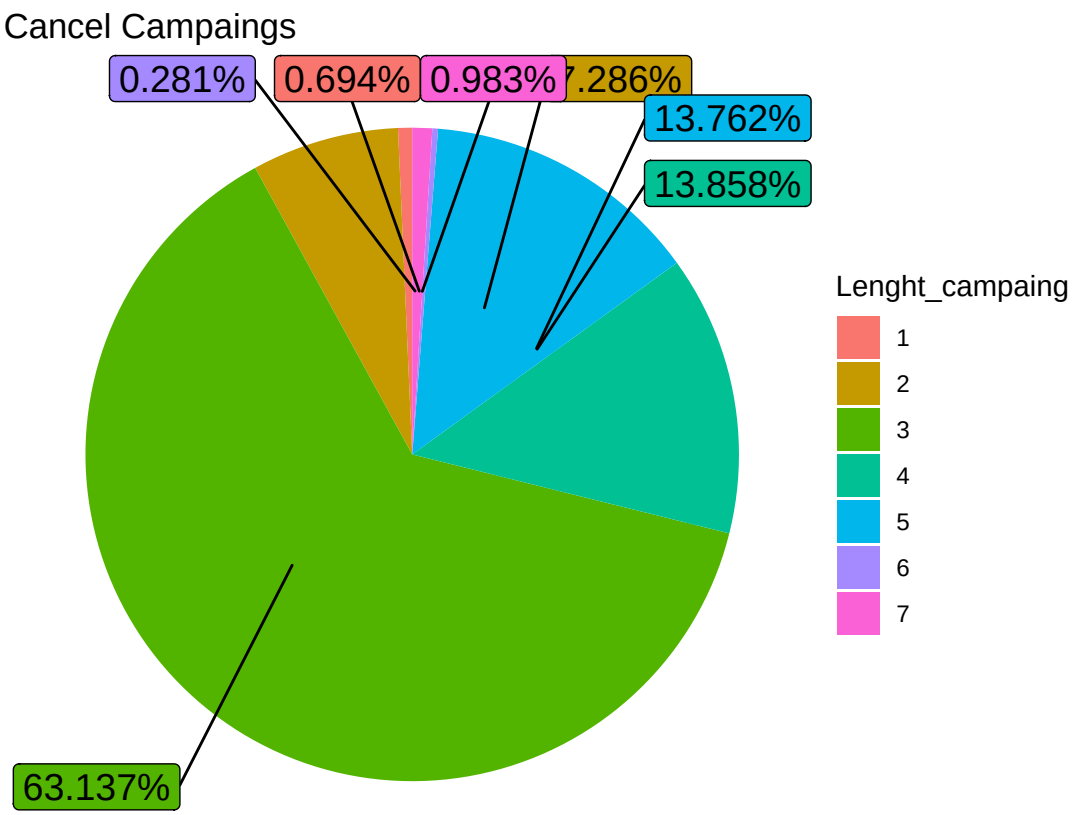


Campañas canceladas

```
piecancel<-ks5%>%
  filter(state=="canceled")
head(piecancel)
```

```
##           ID   deadline            launched    state currency   goal
## 1 1000011046 2015-08-29 2015-07-04 08:35:03 canceled      USD  19500
## 2 1000034518 2014-05-29 2014-04-24 18:14:43 canceled      USD 125000
## 3  100004195 2014-08-10 2014-07-11 21:55:48 canceled      USD  65000
## 4 1000256760 2015-08-07 2015-07-08 21:46:53 canceled      CAD  15000
## 5 1000260691 2016-03-25 2016-02-29 20:30:27 canceled      USD  87000
## 6 1000278154 2015-04-10 2015-03-10 13:19:18 canceled      USD  13000
##   usd.pledged usd_pledged_real usd_goal_real       Lan1 Lenght_campaing
## 1     1283.00          1283.00      19500.00 2015-07-04               5
## 2     8233.00          8233.00     125000.00 2014-04-24               3
## 3     6240.57          6240.57      65000.00 2014-07-11               3
## 4      553.32           535.09      11466.14 2015-07-08               3
## 5     2030.00          2030.00      87000.00 2016-02-29               3
## 6     2453.00          2453.00      13000.00 2015-03-10               3
```

```
piecancel%>%
  count(Lenght_campaing) %>%
  mutate(prop = percent(n / sum(n)))%>%
  ggplot( aes(x="", y=n, fill=Lenght_campaing))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("Cancel Campaings") +
  theme_void() + coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

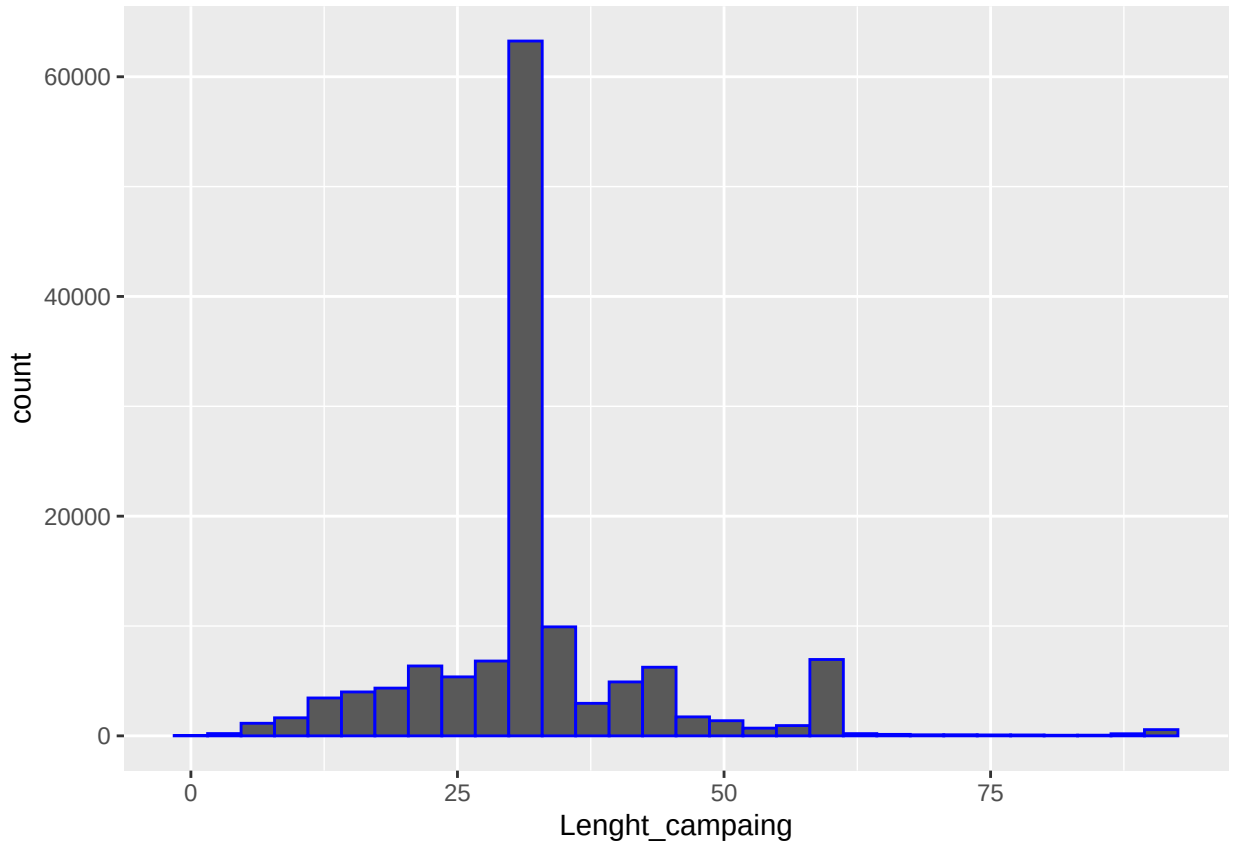


Frecuencia de las campañas y Funcion empirica

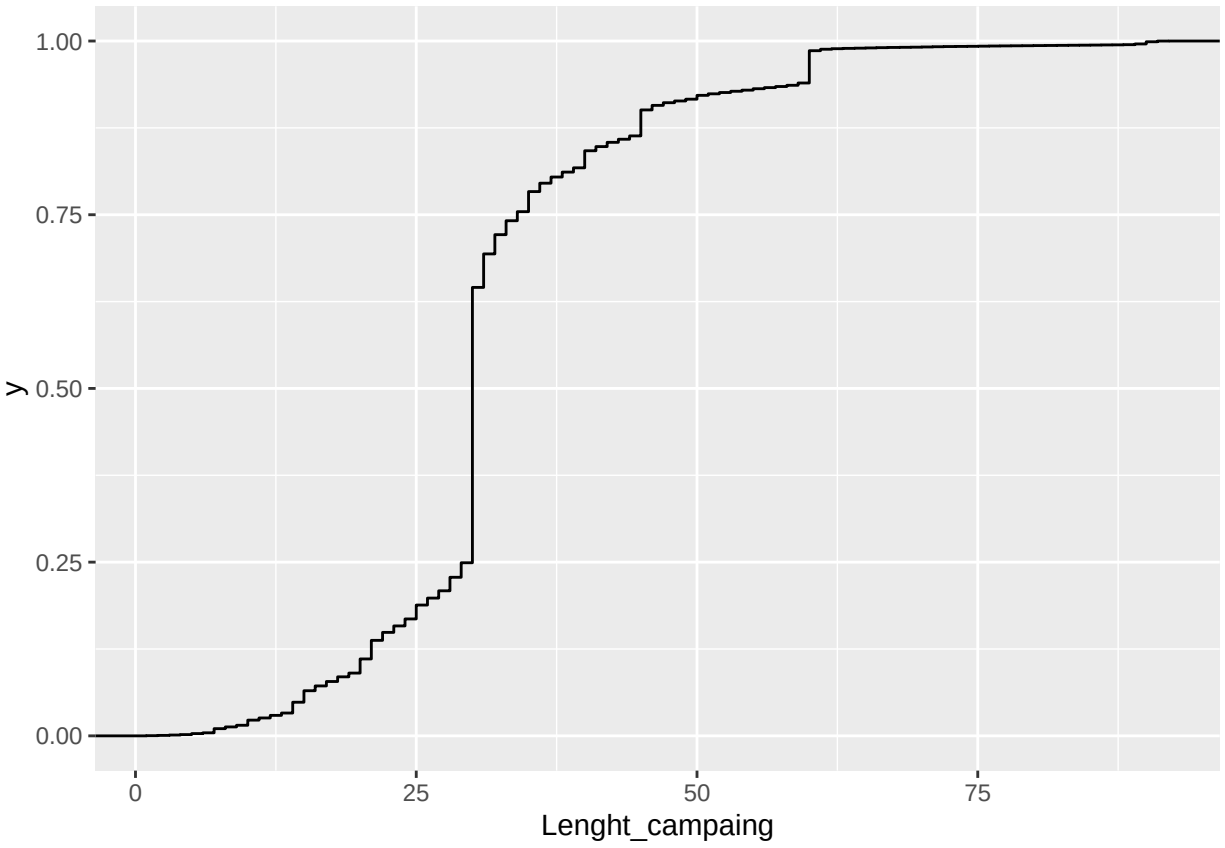
```
#Histograma
histSuces<-ks4%>%
  filter(state=="successful")

ggplot(data=histSuces, aes(x=Lenght_campaing)) +
  geom_histogram(color="blue")
```

campañas exitosas

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



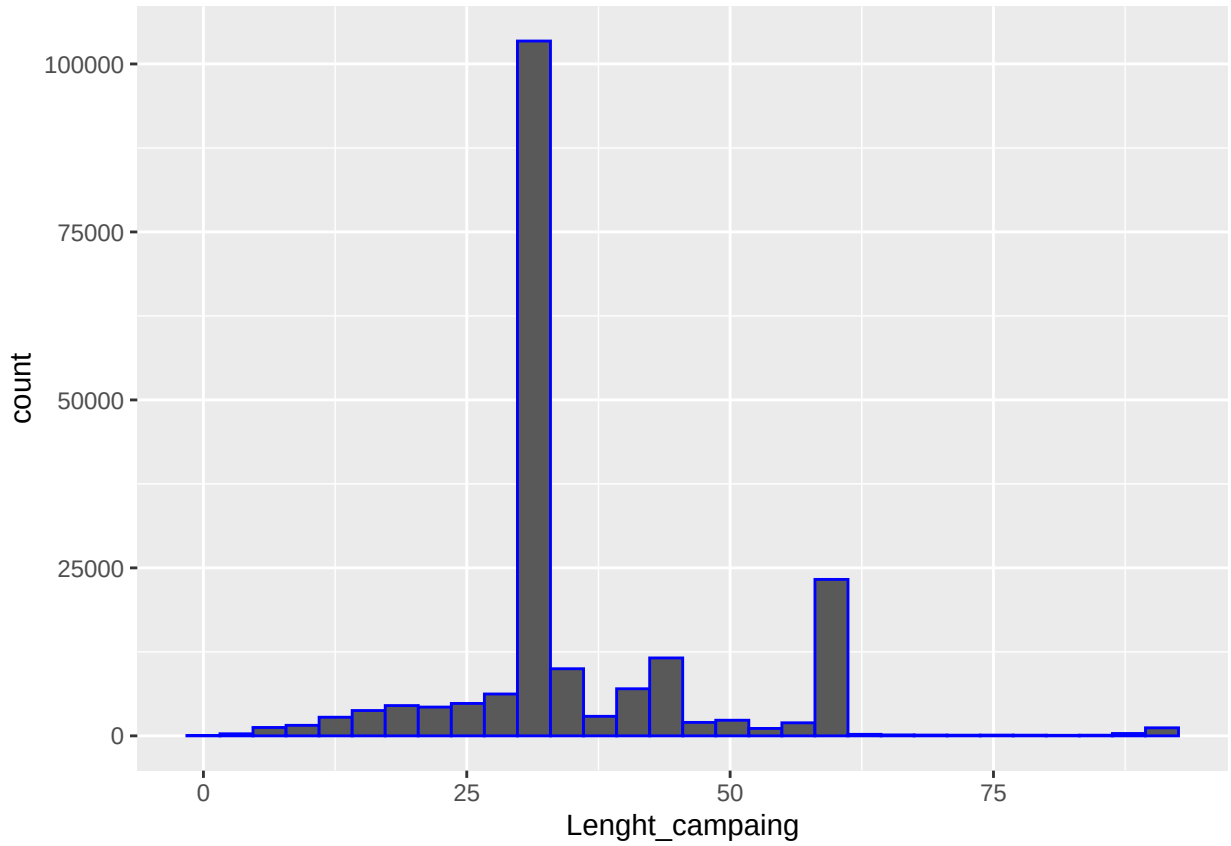

```
#Empirical distribution
ggplot(histSuces,aes(Lenght_campaing)) + stat_ecdf(geom = "step")
```



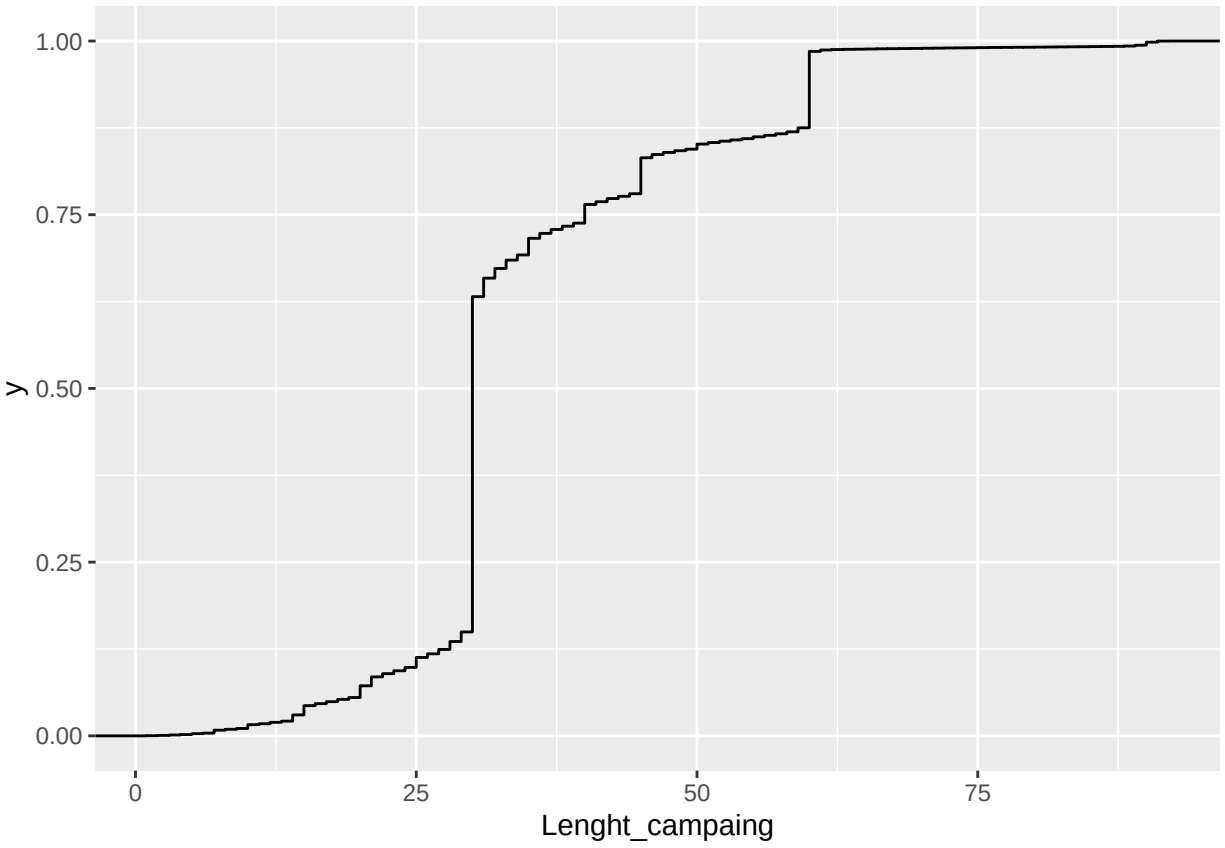
Frecuencia de campañas fallidas

```
histFail<-ks4%>%
  filter(state=="failed")
#head(histFail)

ggplot(data=histFail, aes(x=Lenght_campaing)) +
  geom_histogram(color="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



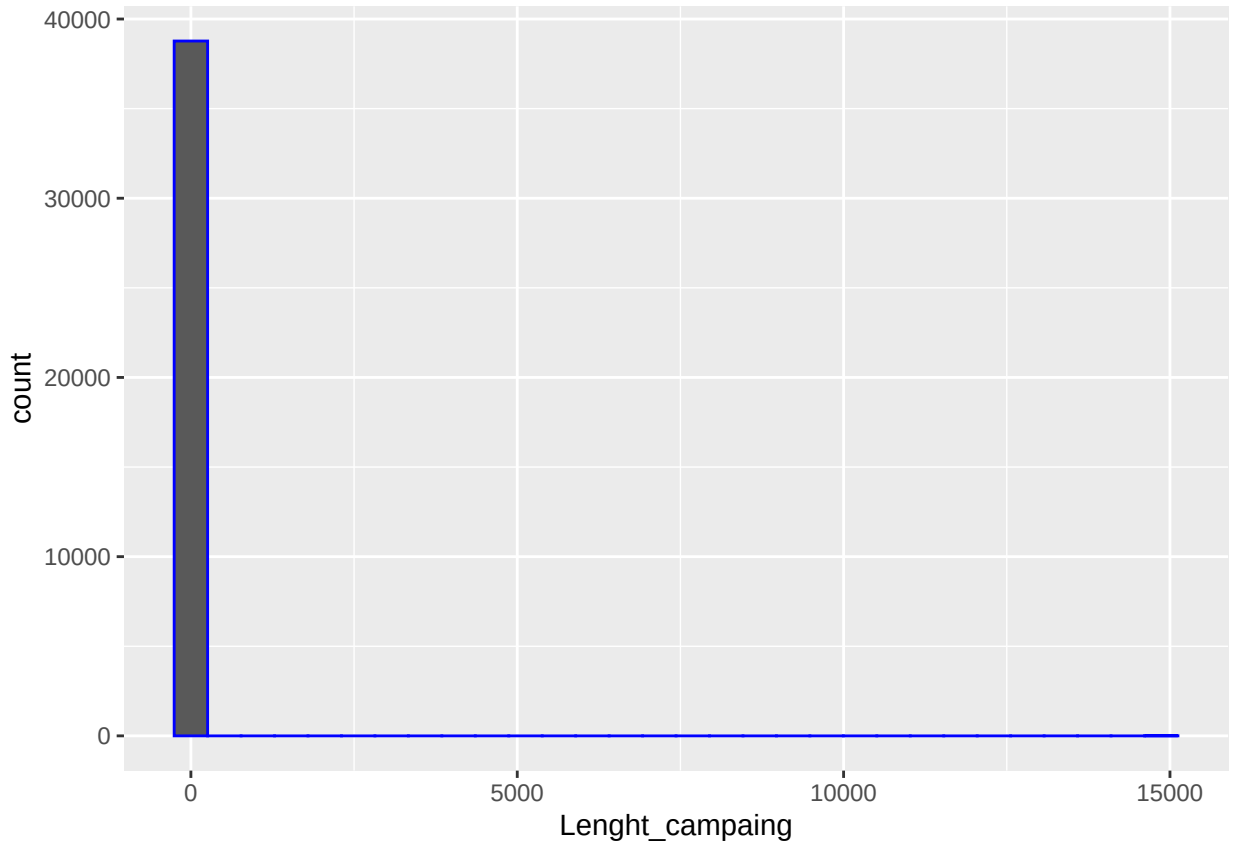

```
#Empirical distribution
ggplot(histFail,aes(Lenght_campaing)) + stat_ecdf(geom = "step")
```



Frecuencia de campañas canceladas

```
histcanceled<-ks4%>%
  filter(state=="canceled")
#head(histcanceled)

ggplot(data=histcanceled, aes(x=Lenght_campaing)) +
  geom_histogram(color="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Empirical distribution
ggplot(histcanceled,aes(Lenght_campaing)) + stat_ecdf(geom = "step")
```

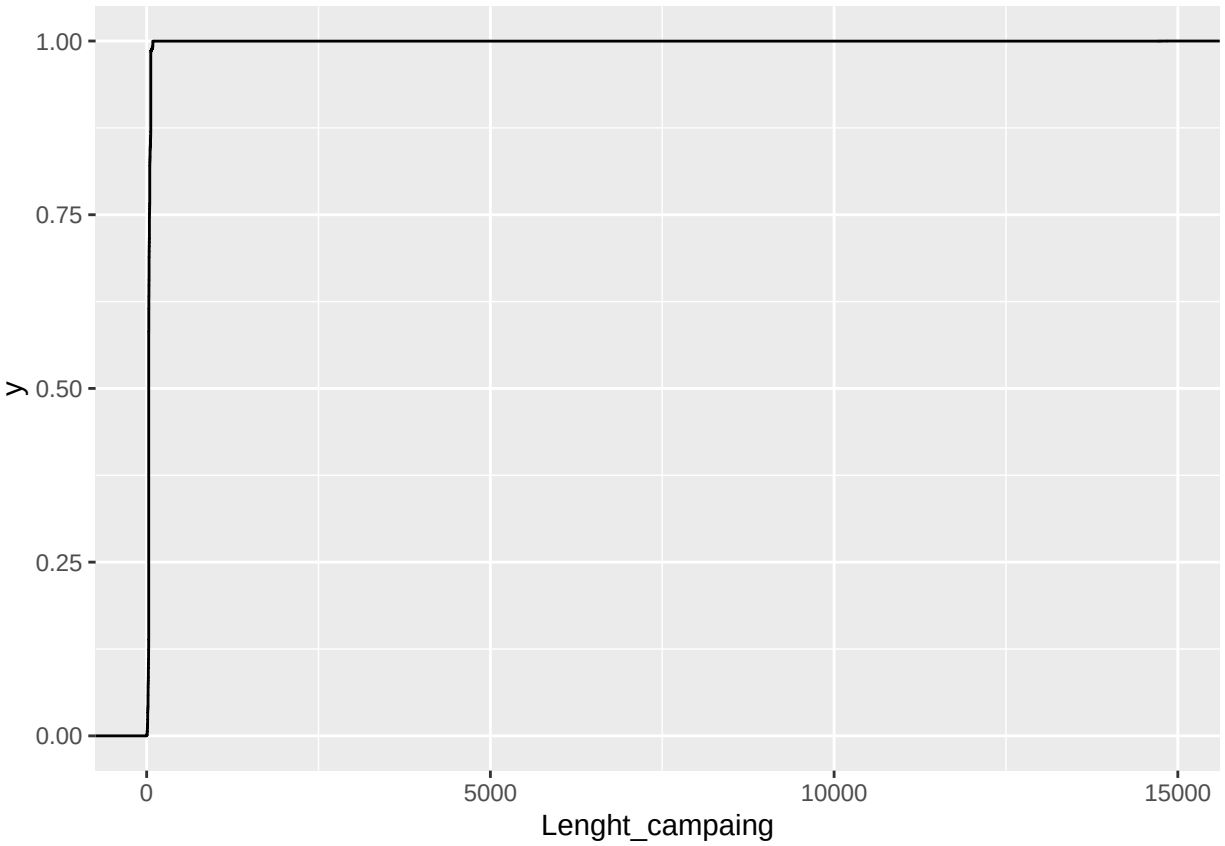
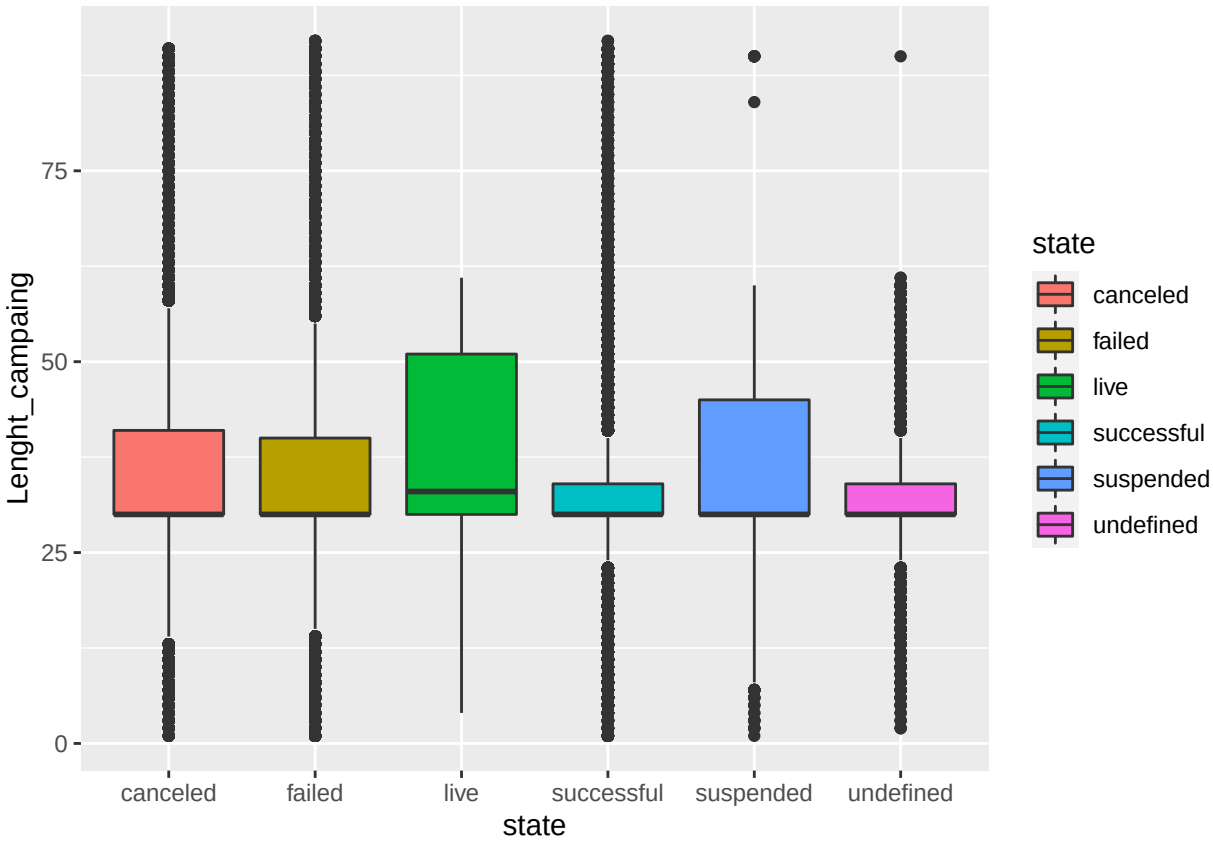


Diagrama de caja

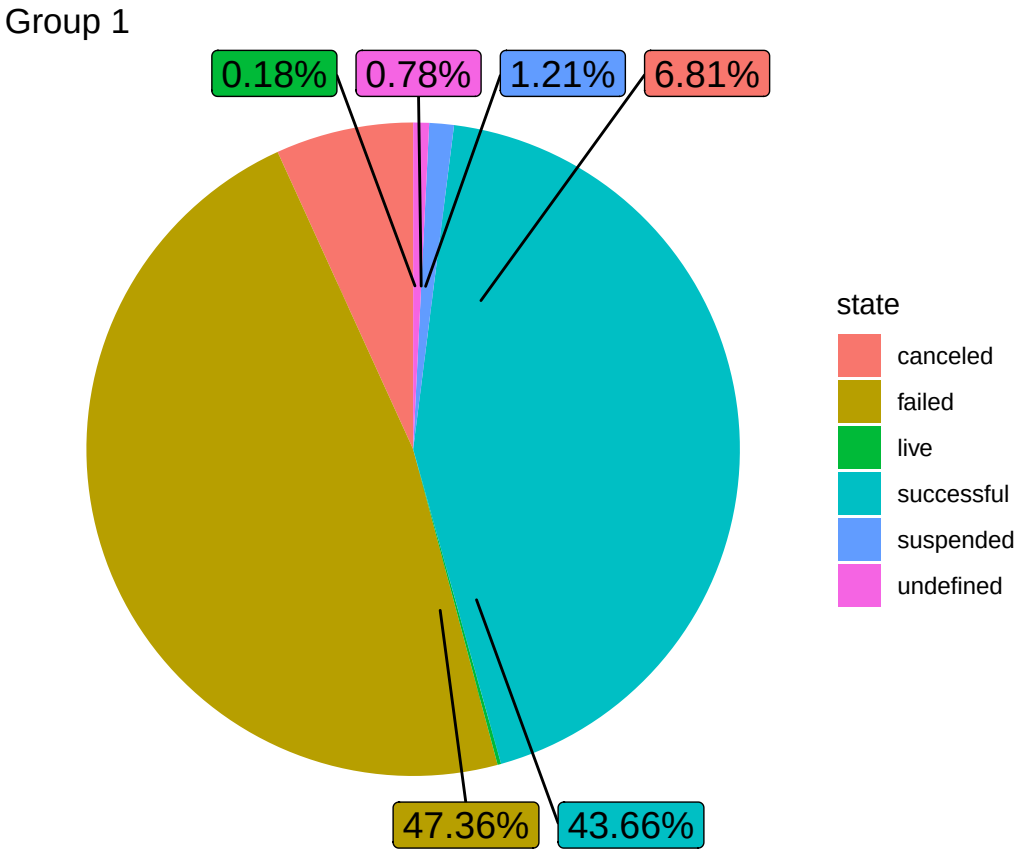
Dias de campaña

```
ks11<-ks4%>%
  filter(Lenght_campaing<=92)
ggplot(ks11,aes(x=state,y=Lenght_campaing, fill=state)) +
  geom_boxplot()
```



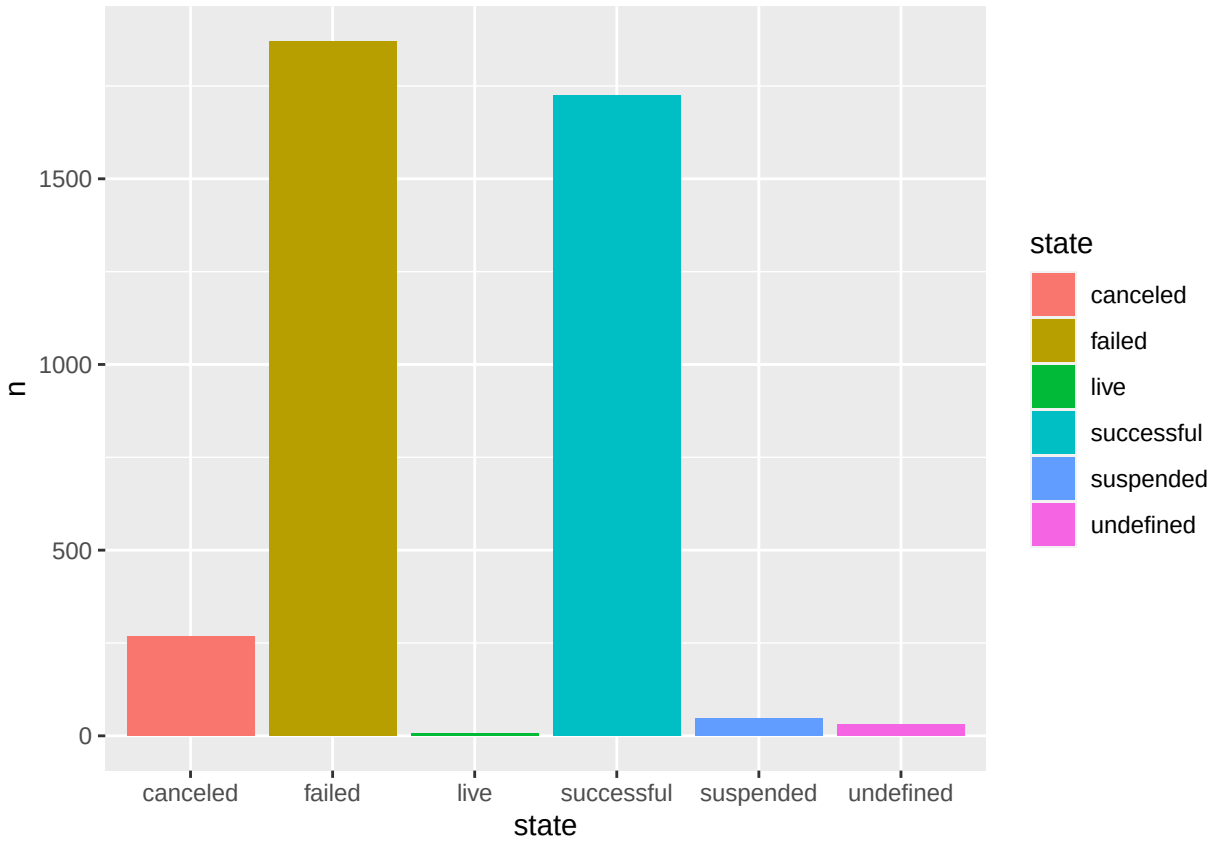
Campañas por dias

```
pie1<-ks5%>%
  filter(Lenght_campaing=="1")

pie1%>%
  count(state)%>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot( aes(x="", y=n, fill=state))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("Group 1") +
  theme_void() + coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

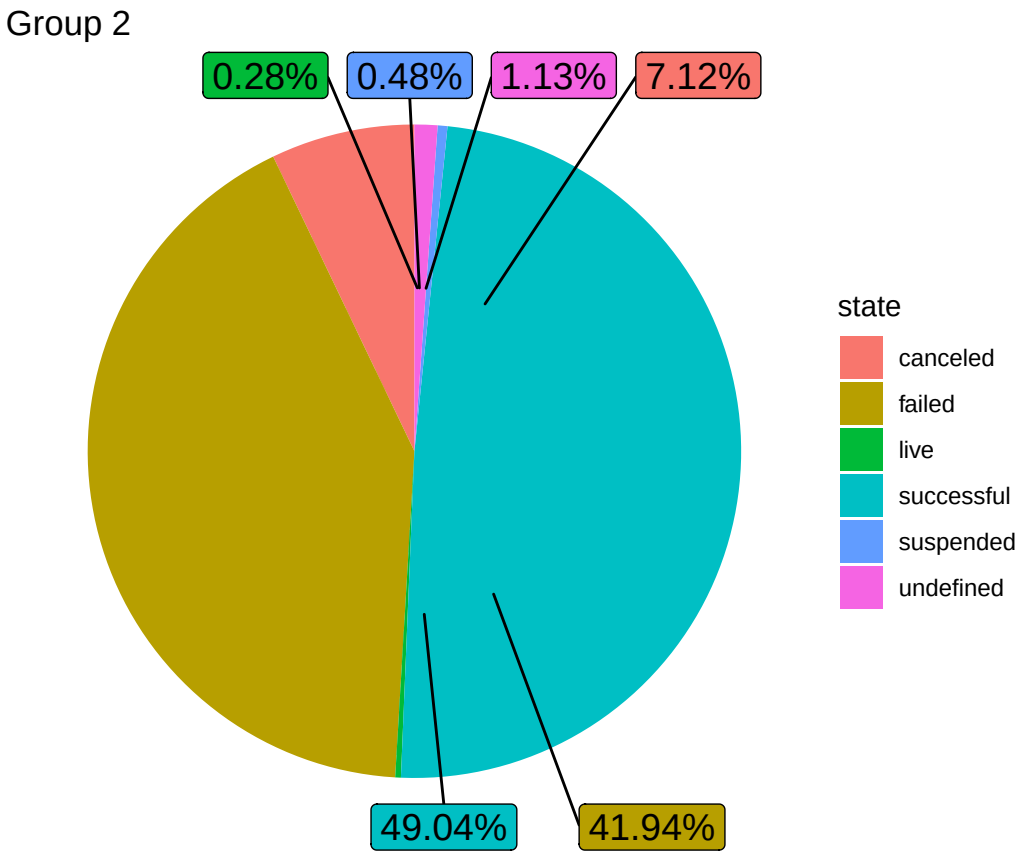


Grupo 1 (1-7 días)

```
#head(pie1)

 pie1%>% count(state)%>%
  ggplot( aes(x=state,y=n,fill=state)) +
  geom_bar(stat="identity")
```

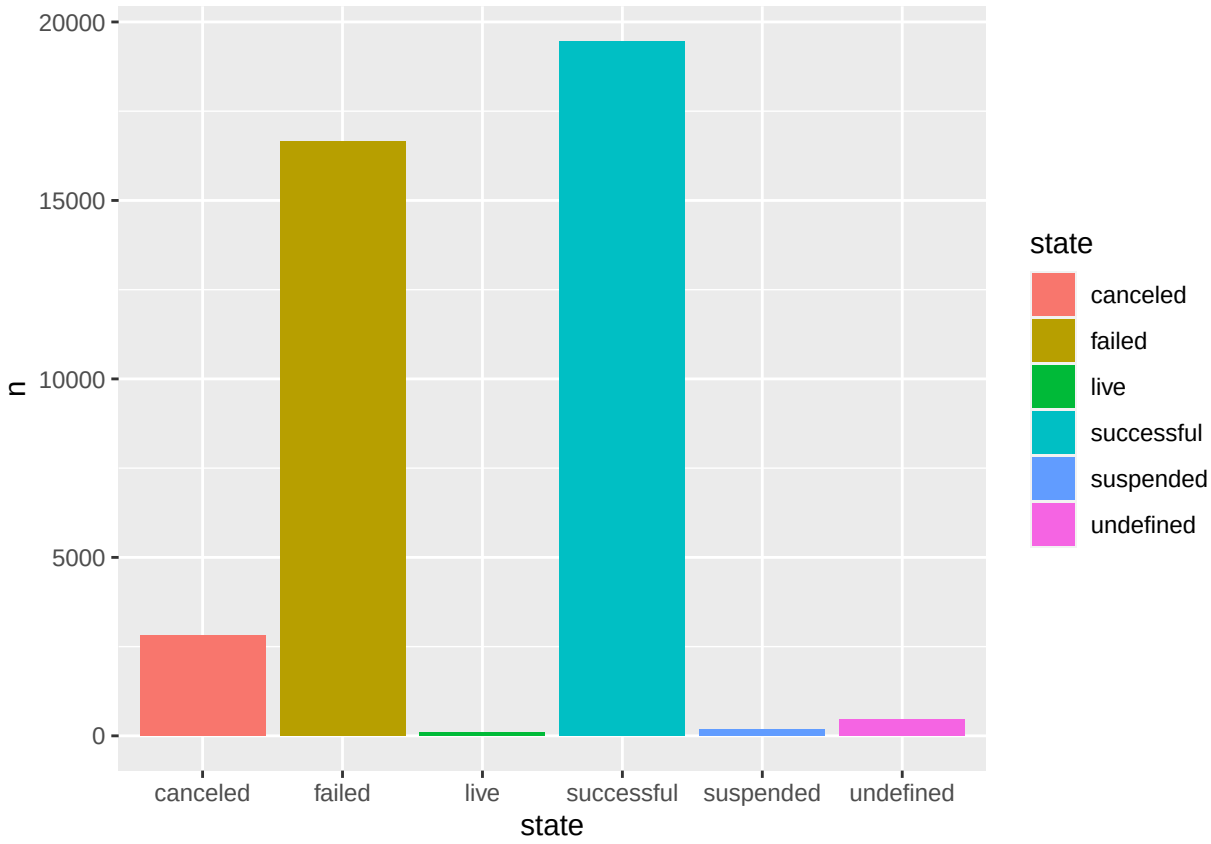



```
pie2<-ks5%>%
  filter(Lenght_campaing=="2")

pie2%>%
  count(state)%>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot( aes(x="", y=n, fill=state))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("Group 2") +
  theme_void() + coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

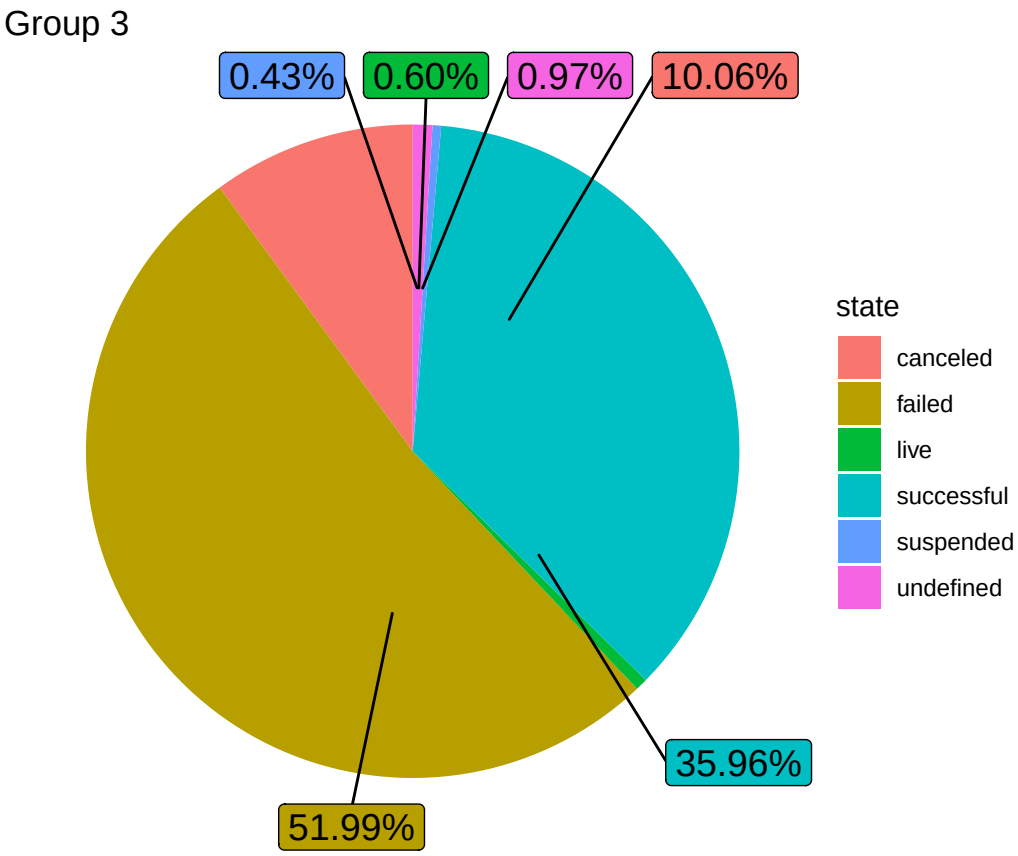


Grupo 2 (8-22 días)

```
pie2%>% count(state)%>%
  ggplot( aes(x=state,y=n,fill=state)) +
  geom_bar(stat="identity")
```

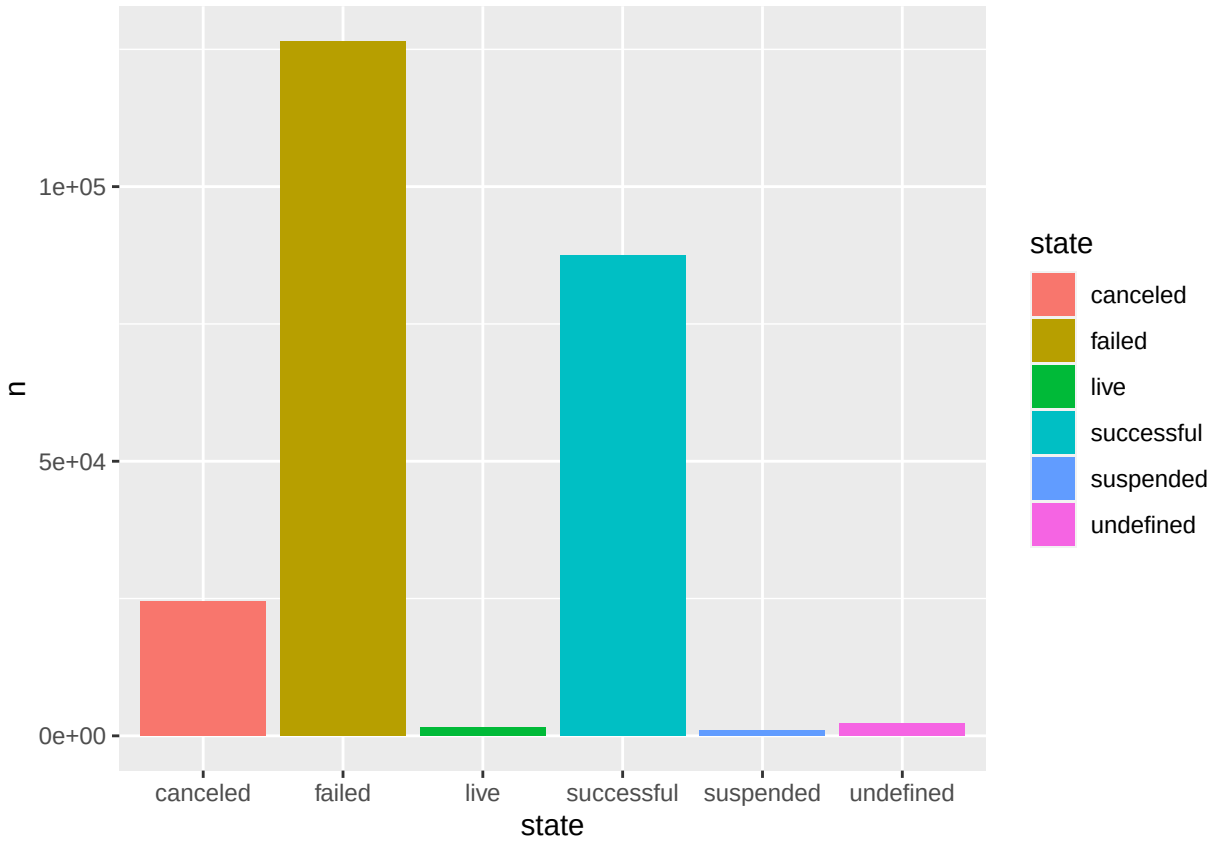



```
pie3<-ks5%>%
  filter(Lenght_campaing=="3")

pie3%>%
  count(state)%>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot( aes(x="", y=n, fill=state))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("Group 3") +
  theme_void() + coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

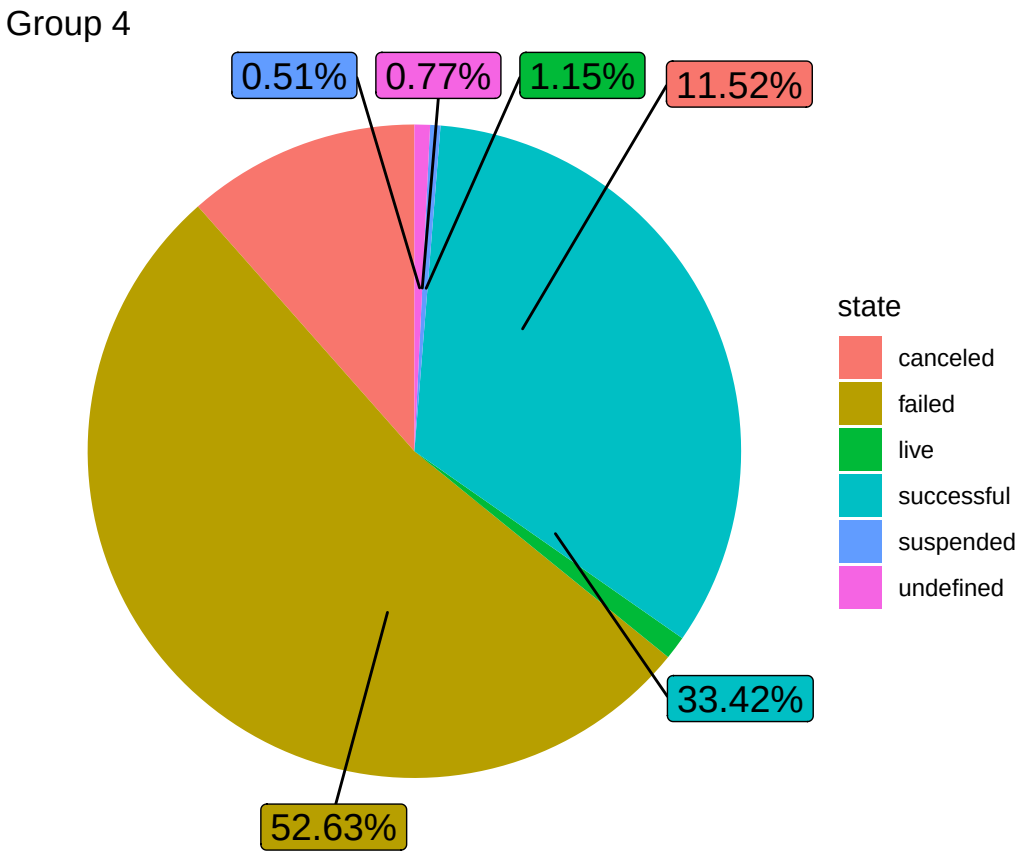


Grupo 3 (23-37 días)

```
pie3%>% count(state)%>%
  ggplot( aes(x=state,y=n,fill=state)) +
  geom_bar(stat="identity")
```

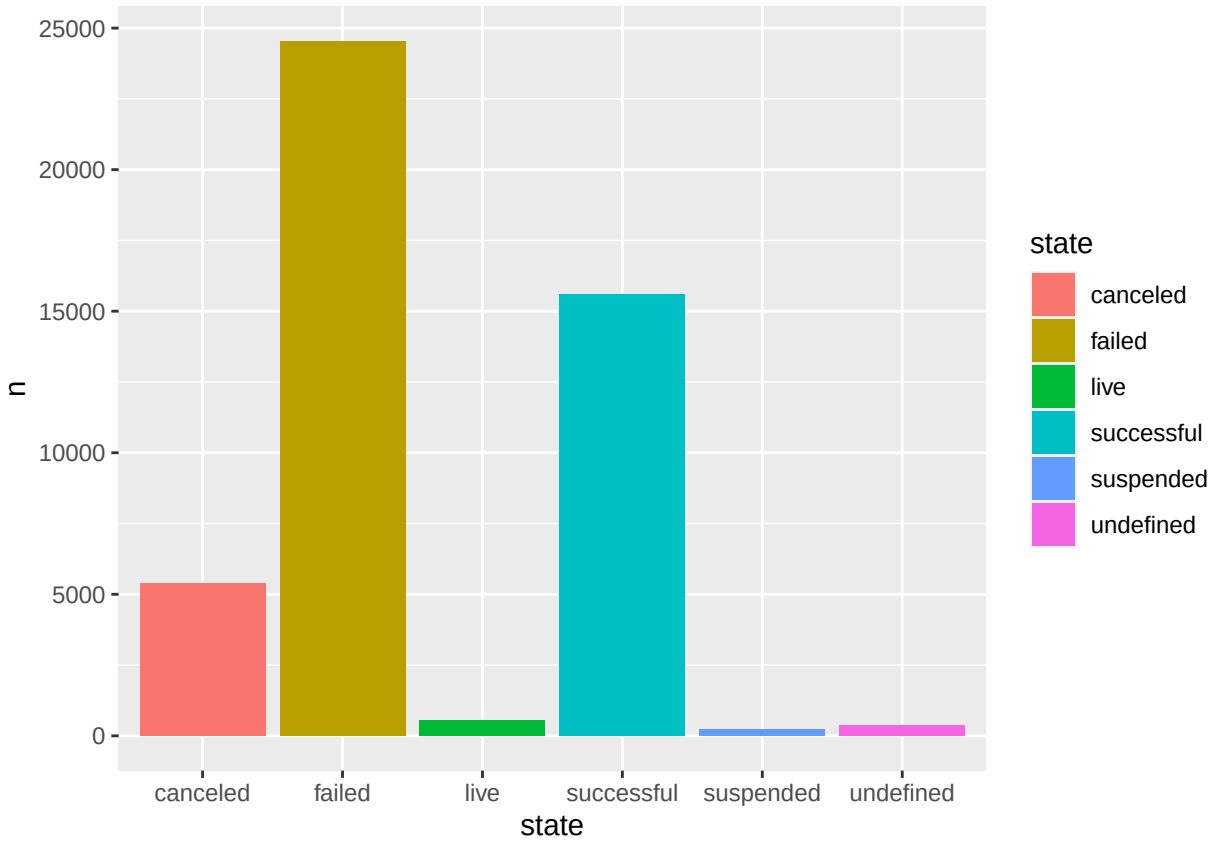



```
pie4<-ks5%>%
  filter(Lenght_campaing=="4")

pie4%>%
  count(state)%>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot( aes(x="", y=n, fill=state))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("Group 4") + theme_void() +
  coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

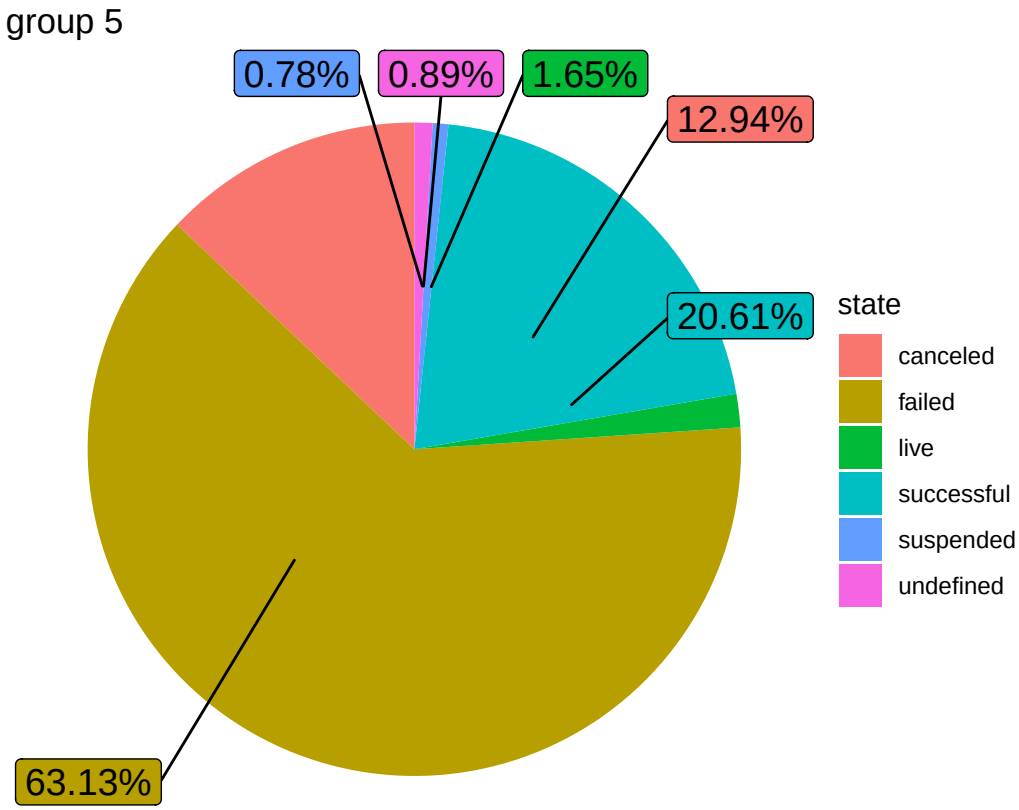


Grupo 4 (38-52)

```
pie4%>% count(state)%>%
  ggplot( aes(x=state,y=n,fill=state)) +
  geom_bar(stat="identity")
```

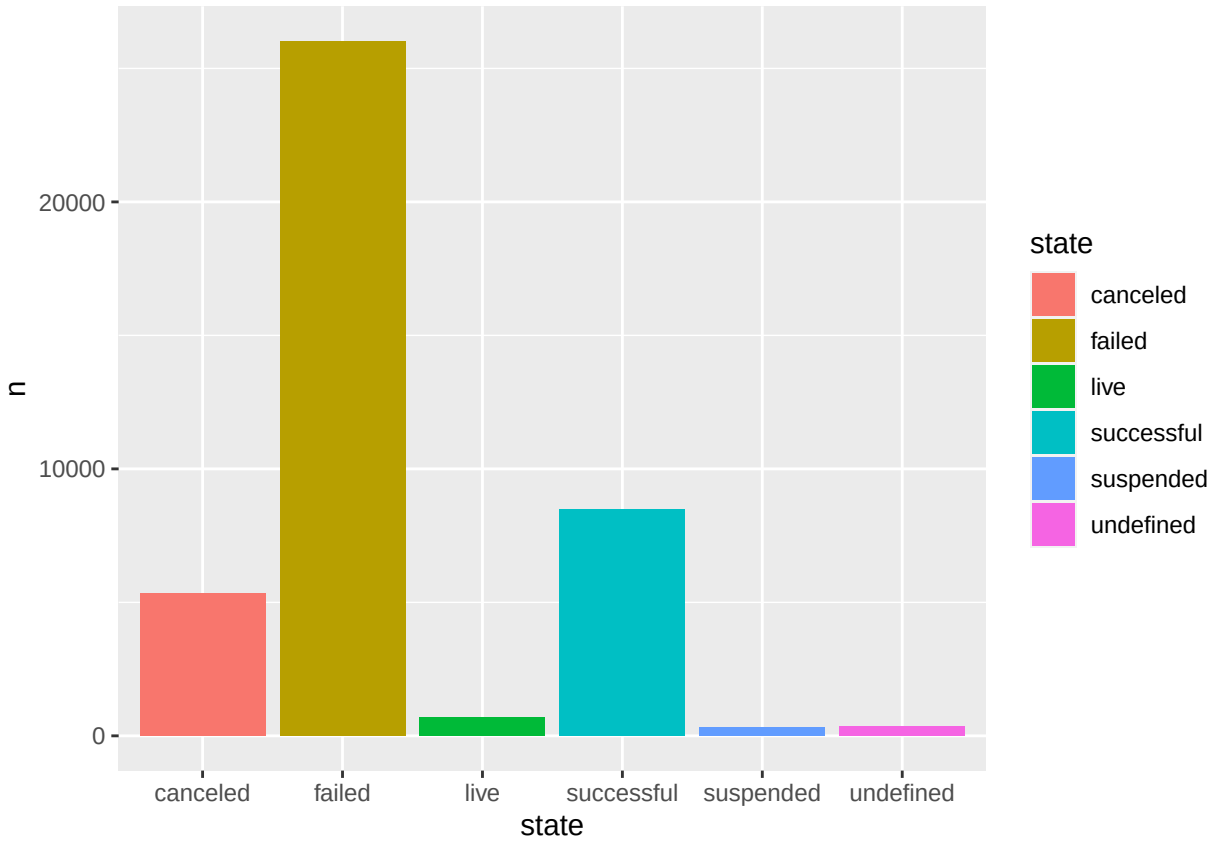



```
pie5<-ks5%>%
  filter(Lenght_campaing=="5")

pie5%>%
  count(state)%>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot( aes(x="", y=n, fill=state))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("group 5") + theme_void() +
  coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

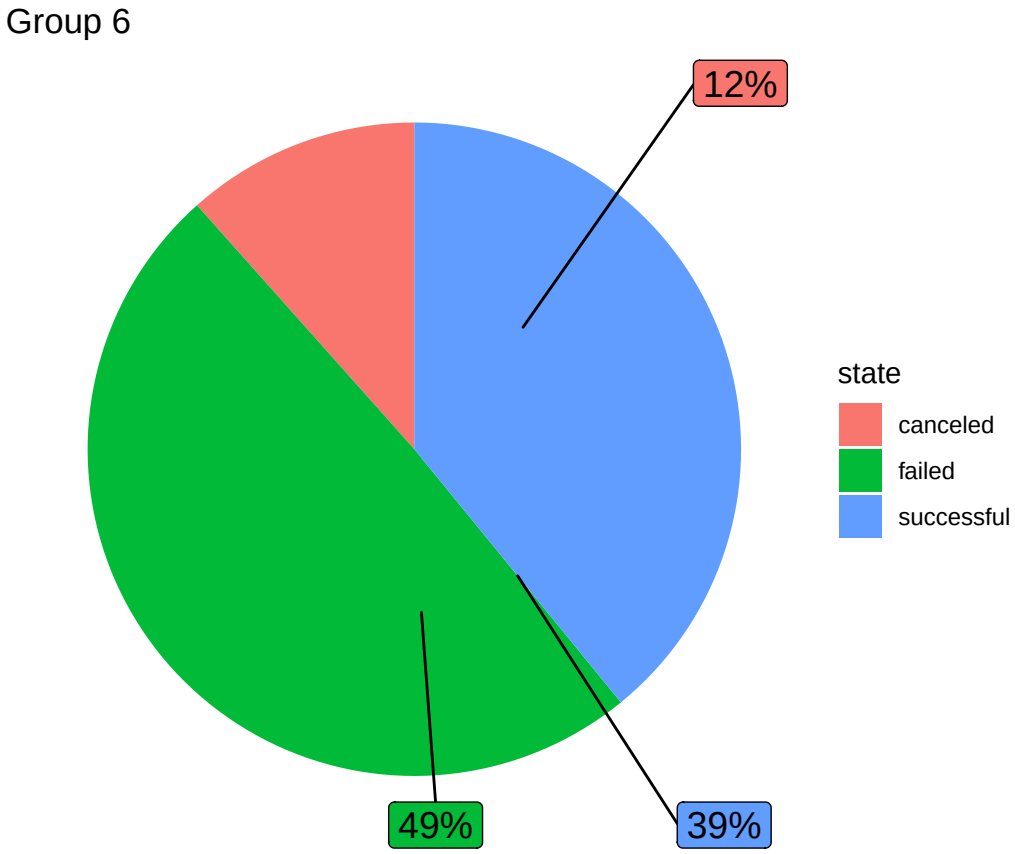


Grupo 5 (53-67)

```
pie5%>% count(state)%>%
  ggplot( aes(x=state,y=n,fill=state)) +
  geom_bar(stat="identity")
```

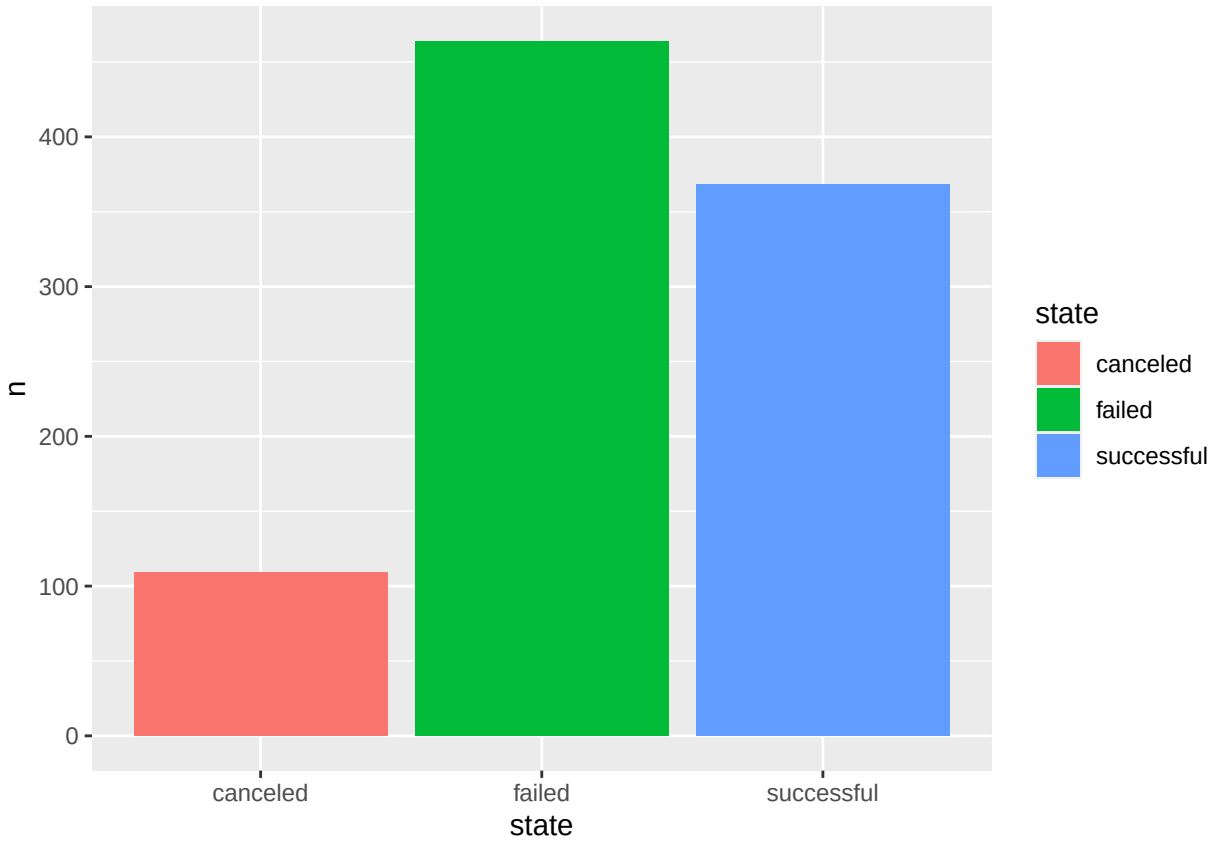



```
pie6<-ks5%>%
  filter(Lenght_campaing=="6")

pie6%>%
  count(state)%>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot( aes(x="", y=n, fill=state))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("Group 6") +
  theme_void() + coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

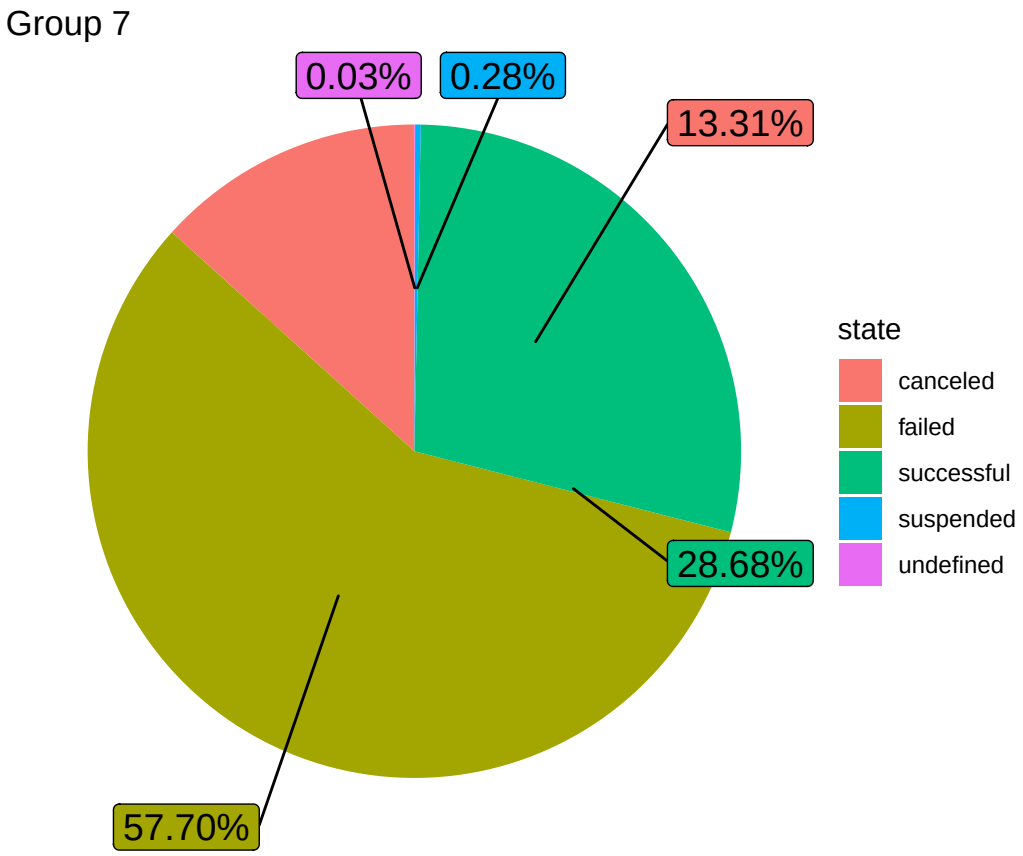


Grupo 6 (68-82)

```
pie6%>% count(state)%>%
  ggplot( aes(x=state,y=n,fill=state)) +
  geom_bar(stat="identity")
```

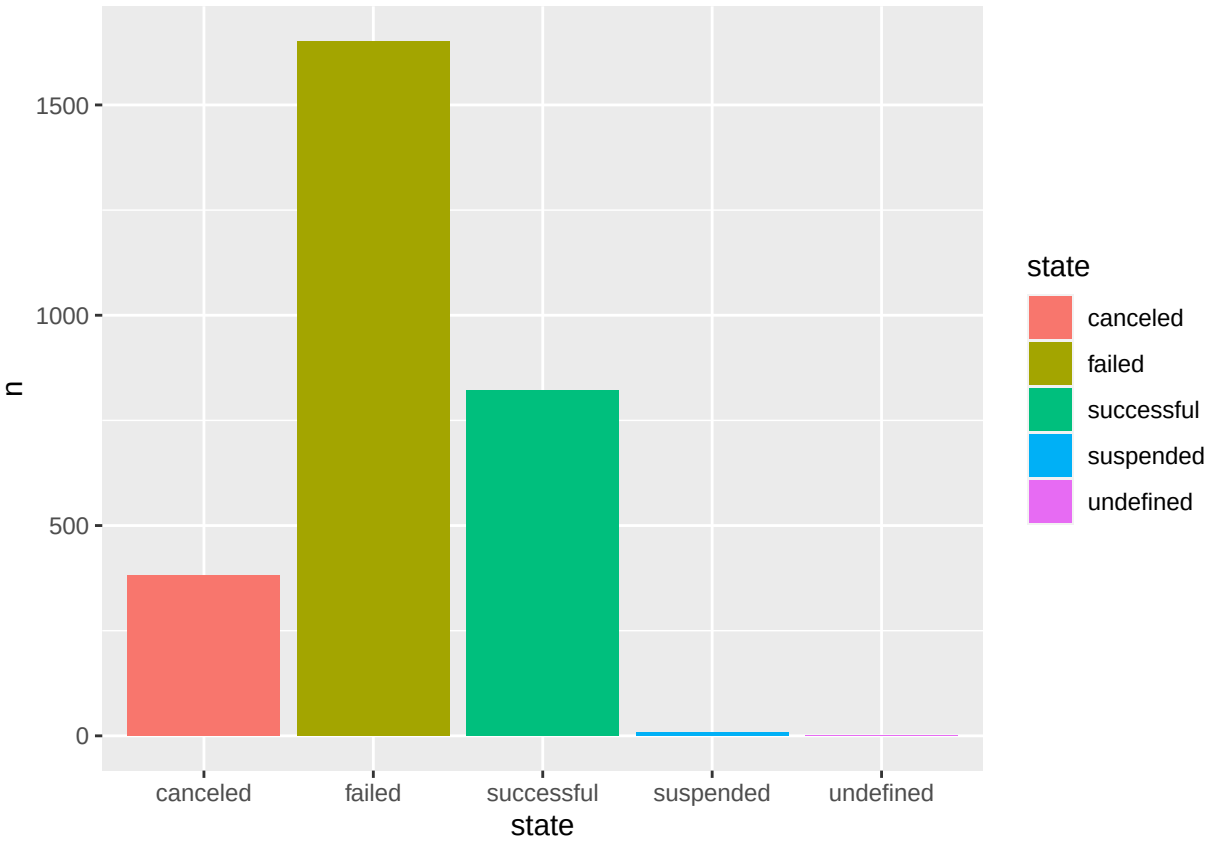



```
pie7<-ks5%>%
  filter(Lenght_campaing=="7")

pie7%>%
  count(state)%>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot( aes(x="", y=n, fill=state))+
  geom_bar(width = 1, stat = "identity")+ ggtitle("Group 7") +
  theme_void() + coord_polar("y", start=0)+
  geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```



Group 7 (83-92)

```
pie7%>% count(state)%>%
  ggplot( aes(x=state,y=n,fill=state)) +
  geom_bar(stat="identity")
```



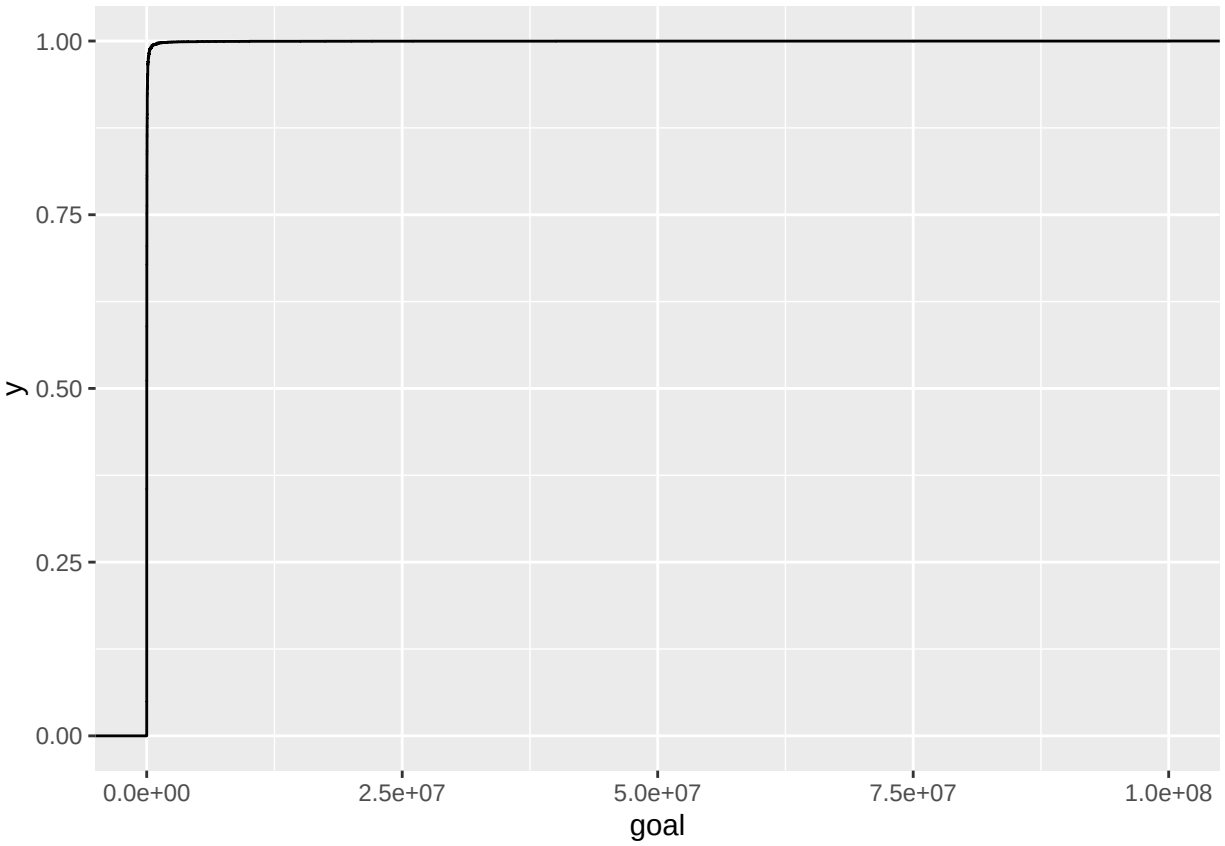
La duración de campaña y el éxito de campaña están relacionadas, pues en las gráficas del grupo 1 (0-7 días), grupo 2 (8-22 días) y grupo 3 (23-37 días) se observa que el porcentaje de éxito de campaña está entre el 35%-44%. Conforme pasan los días de periodo de campaña, el porcentaje de éxito disminuye, entre más larga la duración de campaña, las campañas tienden más al fracaso. En el grupo 5 y 7, el éxito de campaña tiene un menor porcentaje, entre 20% y 28%.

Relación entre el objetivo y el éxito de la campaña y el objetivo y la cantidad recaudada

Se analizará la relación entre los montos de las metas y el éxito de las campañas, así como también las diferencias entre las cantidades prometidas y los objetivos de estas campañas.

función de distribución empírica del objetivo de las campañas exitosas

```
#Objetivo de campaña
ggplot(ks5, aes(goal)) + stat_ecdf(geom = "step")
```

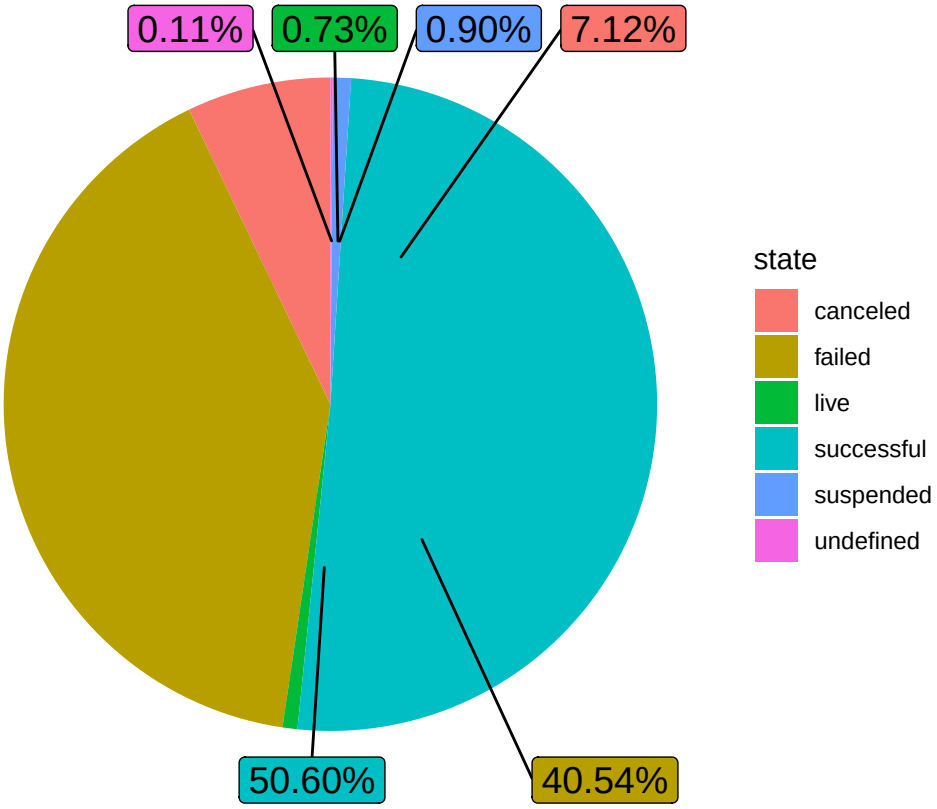


Analisis de disitntos rangos de metas para la campaña

Grupo 1

```
#distribucion acumulativa empirica

emp1<-ks5 %>% filter(goal<1000)

 emp1%>% count(state) %>%
   mutate(prop = percent(n / sum(n))) %>%
   ggplot(aes(x="", y=n, fill=state))+ geom_bar(width = 1, stat = "identity")+
   ggtitle("Group 1: <= 1000 (USD)") +
   theme_void() + coord_polar("y", start=0)+
   geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

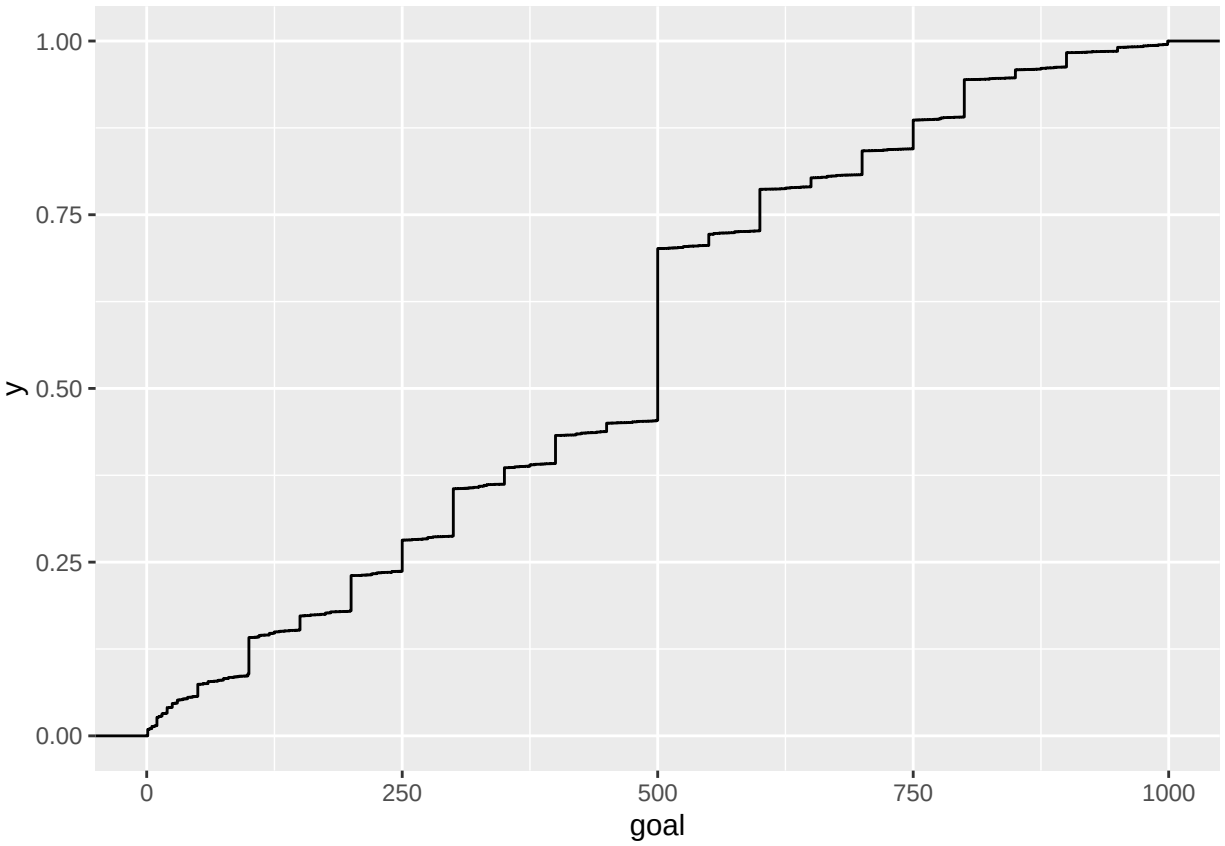
Group 1: <= 1000 (USD)




```
summary(emp1$goal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.01  250.00  500.00  438.36  600.00  999.99
```

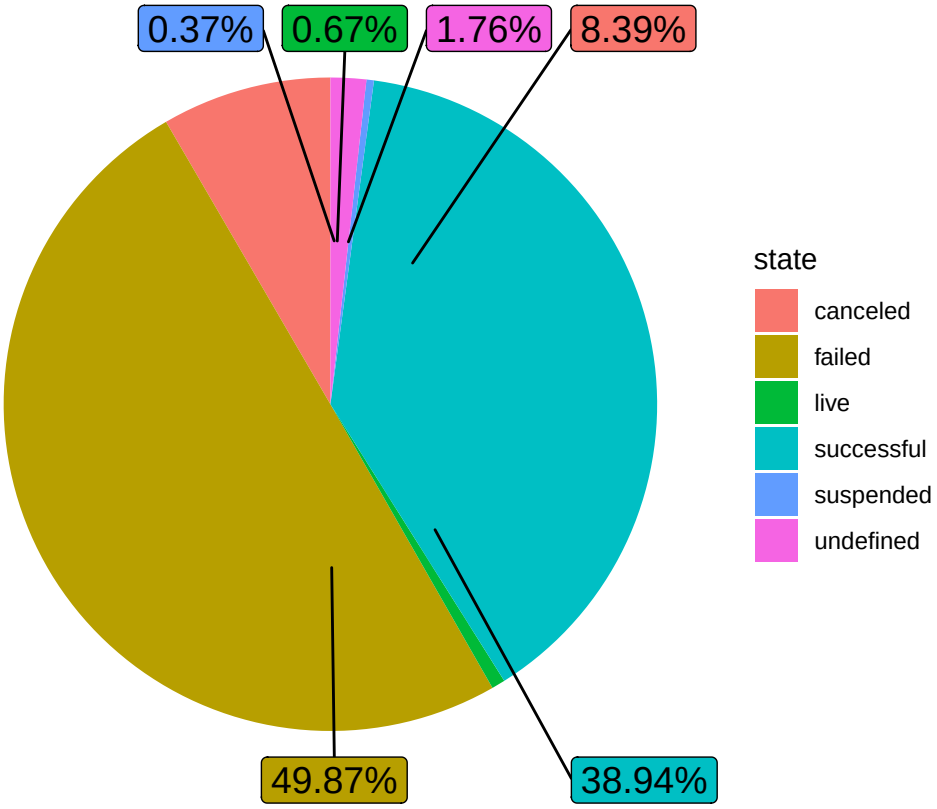
```
emp1%>%
ggplot(aes(goal)) + stat_ecdf(geom = "step")
```



Grupo 2

```
emp2<-ks5 %>% filter(goal>1000 &goal<=10000)

 emp2%>% count(state) %>%
   mutate(prop = percent(n / sum(n))) %>%
   ggplot(aes(x="", y=n, fill=state))+ geom_bar(width = 1, stat = "identity")+
   ggtitle("Group 2: <1000<x<10000") + theme_void() +
   coord_polar("y", start=0)+
   geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

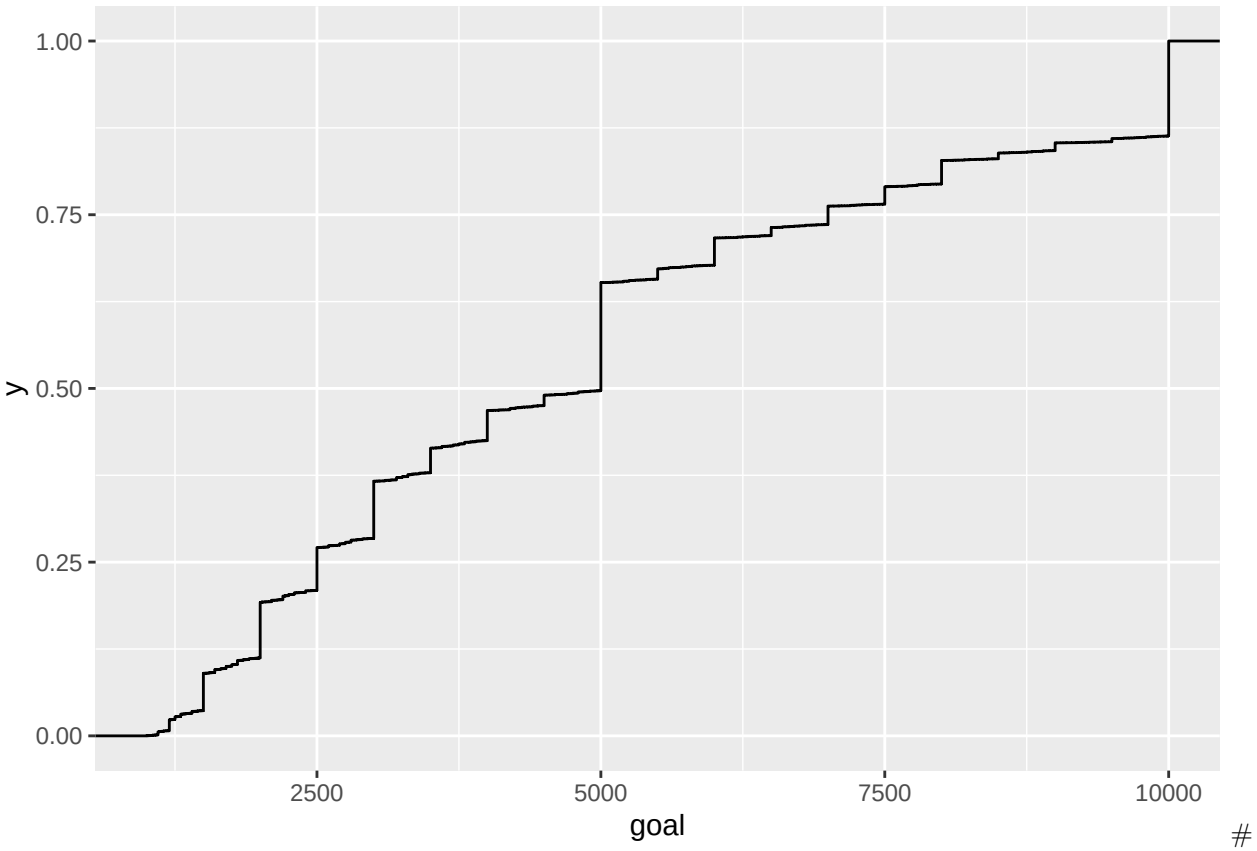
Group 2: <1000<x<10000




```
summary(emp2$goal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1001    2500    5000    4960    7000   10000
```

```
emp2%>%
ggplot(aes(goal)) + stat_ecdf(geom = "step")
```

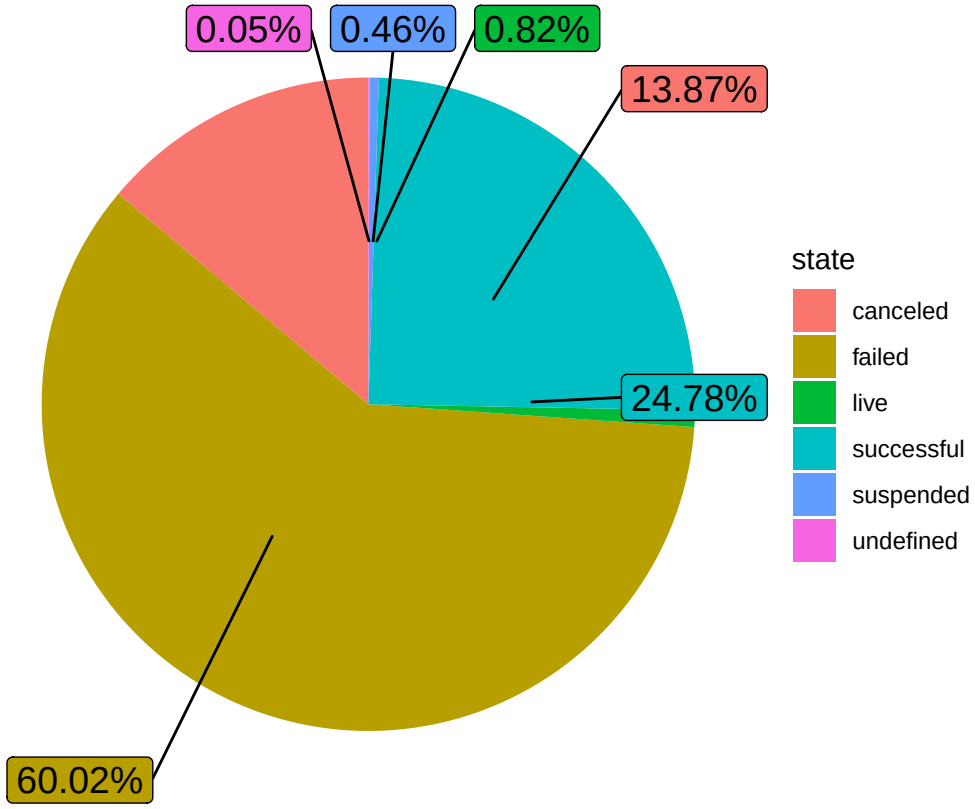


#

Grupo 3

```
emp3<-ks5 %>% filter(goal>10000 &goal<=100000)

 emp3%>% count(state) %>%
   mutate(prop = percent(n / sum(n))) %>%
   ggplot(aes(x="", y=n, fill=state))+ geom_bar(width = 1, stat = "identity")+
   ggtitle("Group 3: 10000<x<100000 (USD)") +
   theme_void() + coord_polar("y", start=0)+
   geom_label_repel(aes(label = prop), size=5, show.legend = F, nudge_x = 1)
```

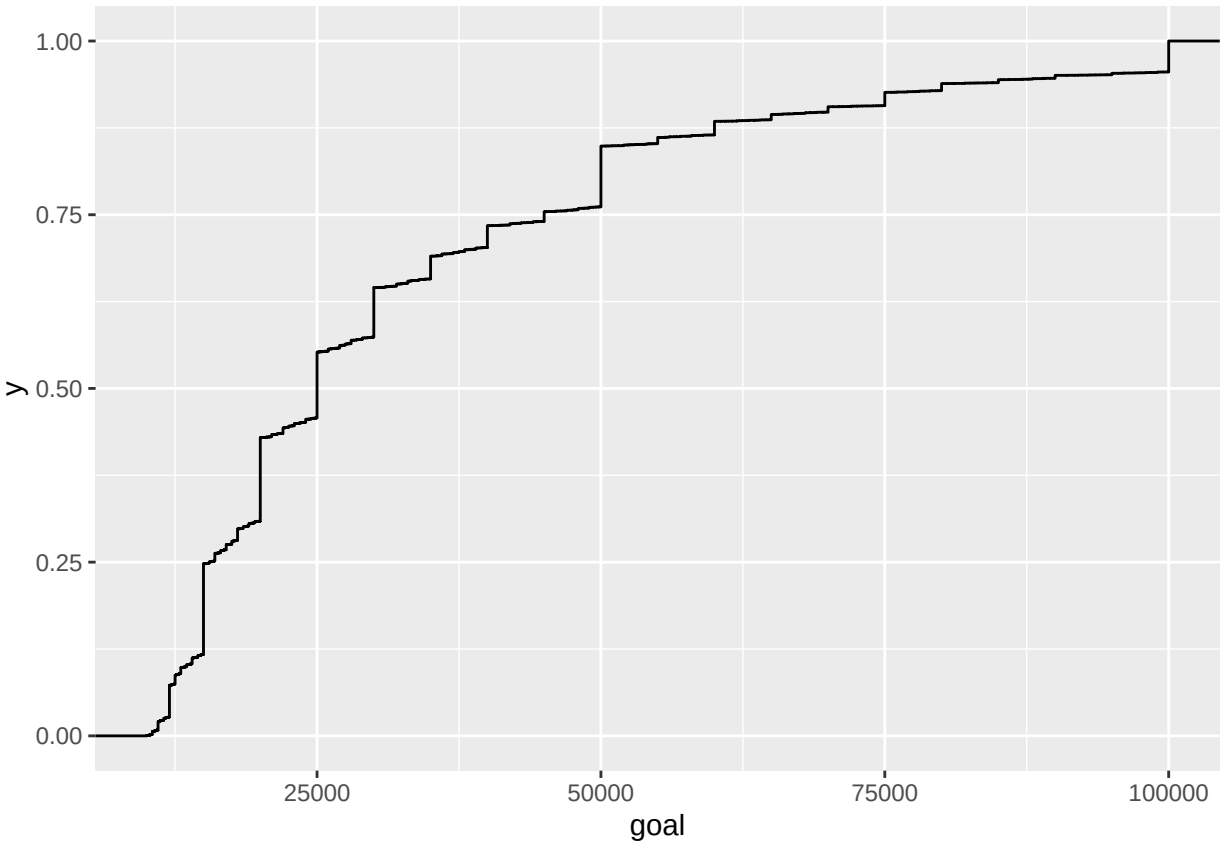
Group 3: 10000<x<100000 (USD)




```
summary(emp3$goal)
```

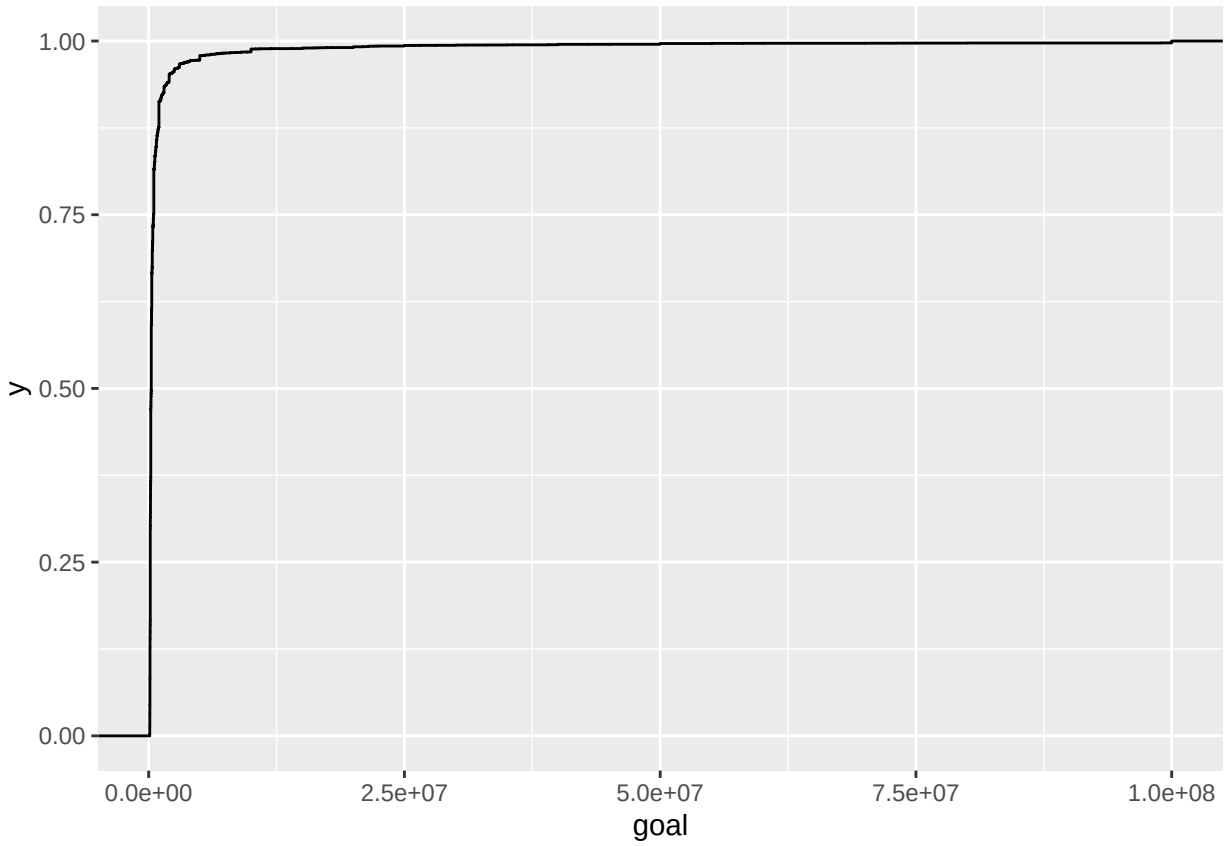
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10001   15500   25000   33441   45000  100000
```

```
emp3%>%
ggplot(aes(goal)) + stat_ecdf(geom = "step")
```



Grupo 4

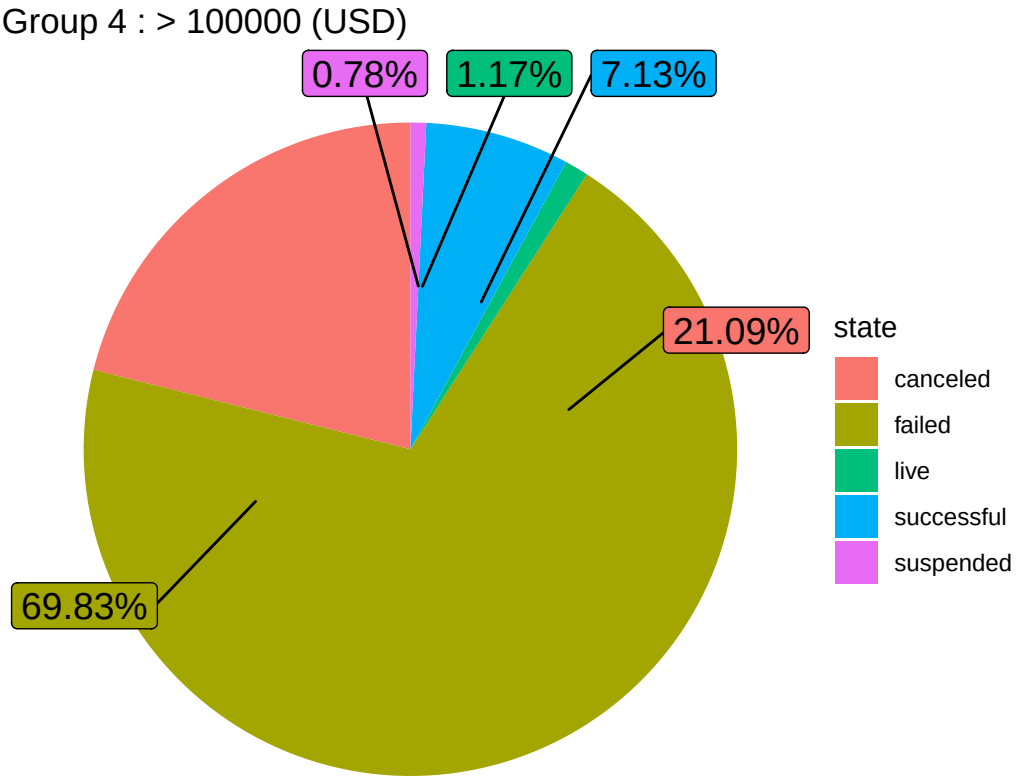
```
emp4<-ks5 %>% filter(goal>100000)

emp4%>%
ggplot(aes(goal)) + stat_ecdf(geom = "step")
```




```
summary(emp4$goal)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 1.00e+05 1.50e+05 2.50e+05 1.05e+06 4.90e+05 1.00e+08
```

```
emp4%>% count(state) %>%
  mutate(prop = percent(n / sum(n))) %>%
  ggplot(aes(x="", y=n, fill=state))+ geom_bar(width = 1, stat = "identity")+
  ggtitle("Group 4 : > 100000 (USD)") +
  theme_void() + coord_polar("y", start=0)+geom_label_repel(aes(label = prop), size=5, show.legend = F,
```



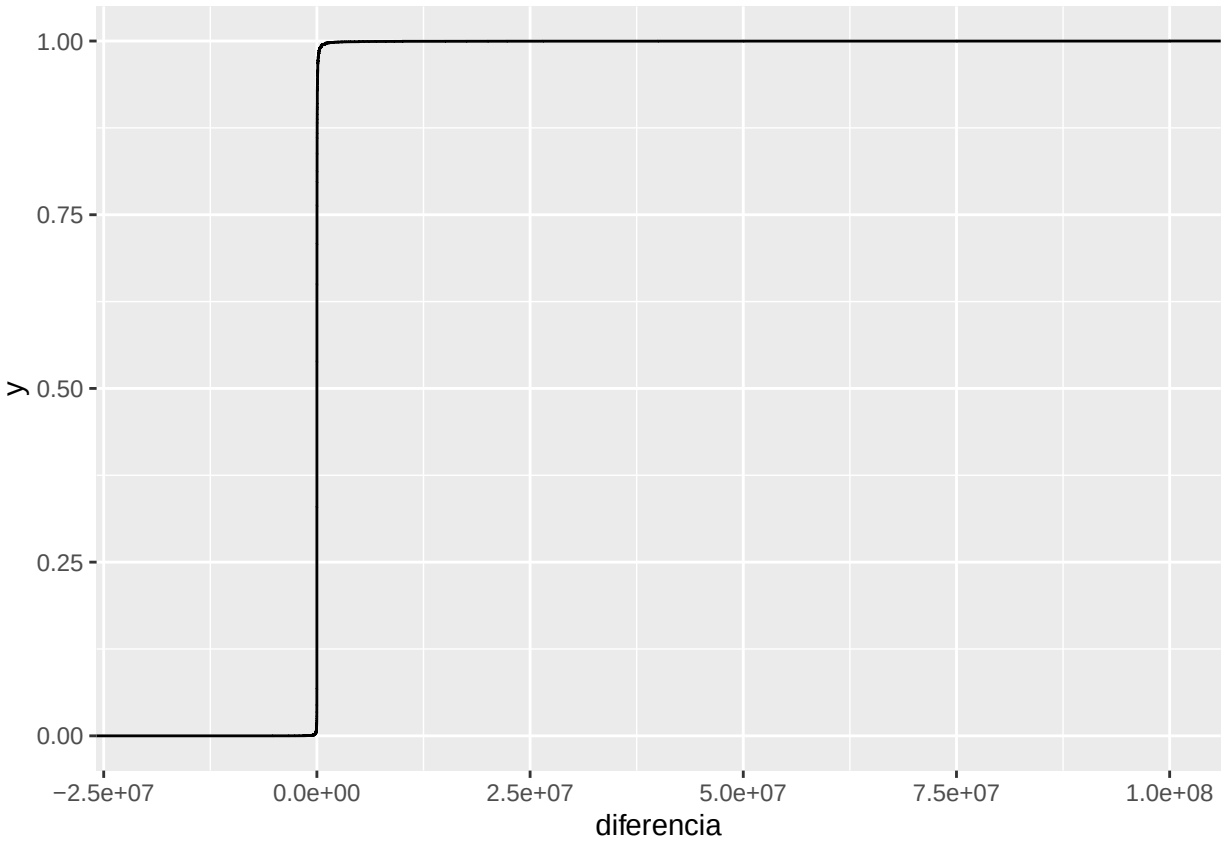
Entre más rande sea el valor de goal, es decir del dinero, mayor es la tendencia a que la campaña fracase. Se observa en el grupo 1 las campañas con menos dinero fueron las que resultaron exitosas. Mientras que en los demás grupos predominaron las campañas que fracasaron.

Diferencia de Goal y pledge

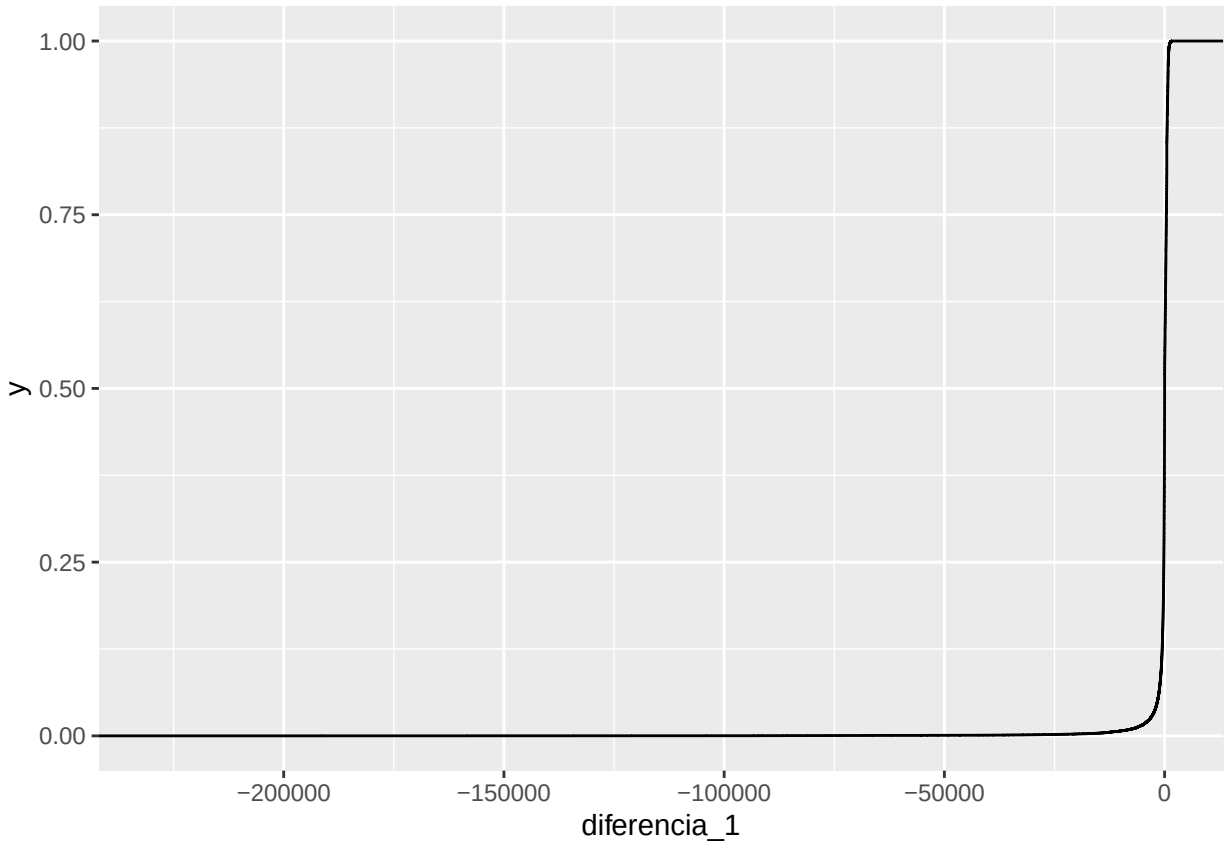
con función de distribución empírica

```
# Empirical de pledged

ks7<-ks4%>%
  mutate(diferencia=goal-usd.pledged)
ggplot(ks7,aes(diferencia)) + stat_ecdf(geom = "step")
```

```
## Warning: Removed 3797 rows containing non-finite values (stat_ecdf).
```

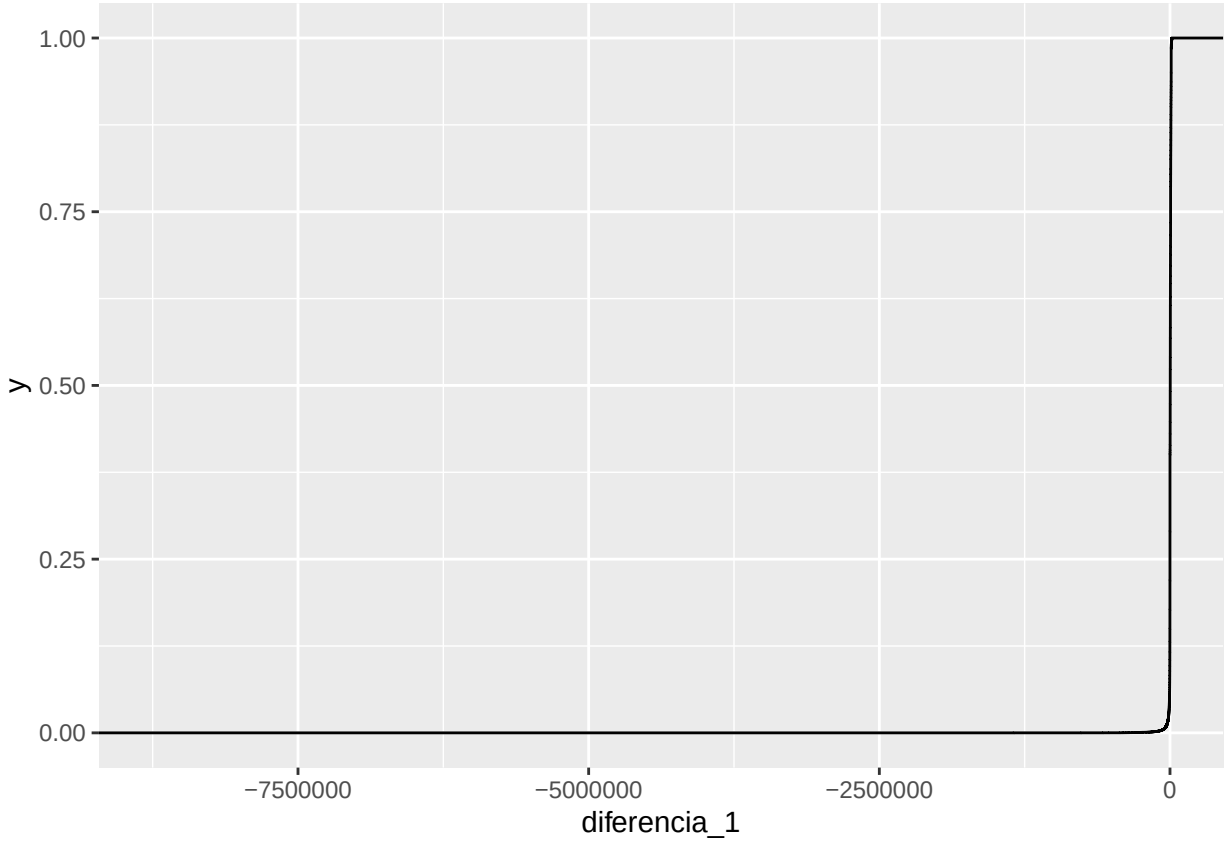



```
emp11<-emp1%>%
   mutate(diferencia_1=usd_goal_real-usd_pledged_real)
ggplot(emp11,aes(diferencia_1)) + stat_ecdf(geom = "step")
```



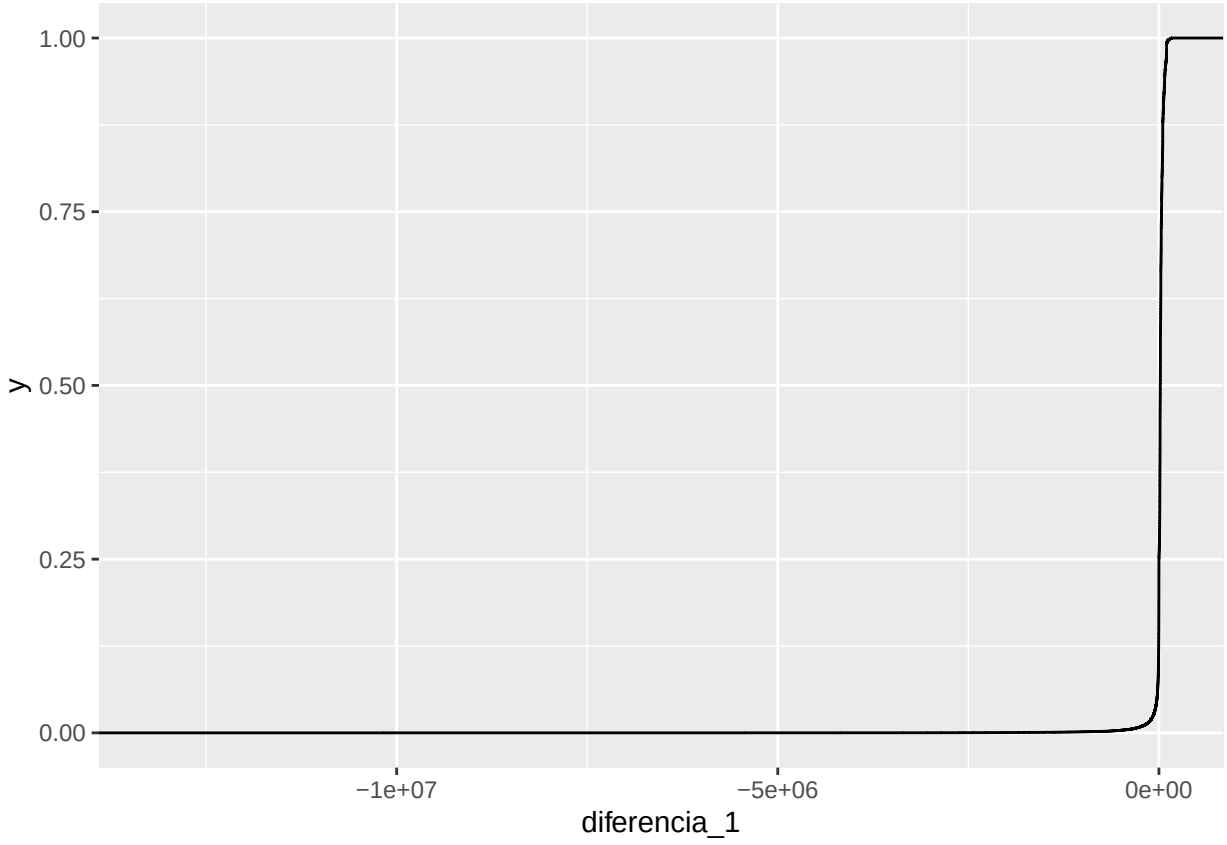
Grupo 1

```
emp22<-emp2%>%
   mutate(diferencia_1=usd_goal_real-usd_pledged_real)
ggplot(emp22,aes(diferencia_1)) + stat_ecdf(geom = "step")
```



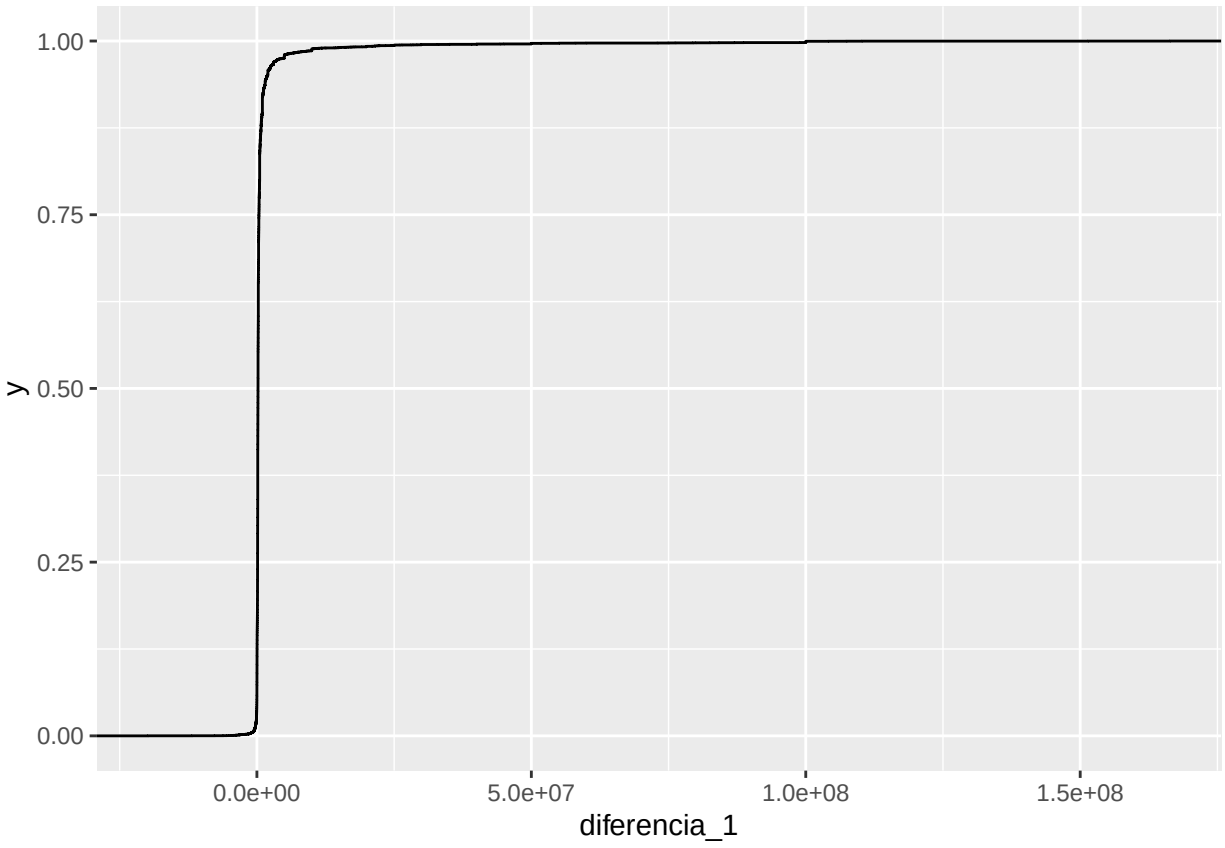
Grupo 2

```
emp33<-emp3%>%
   mutate(diferencia_1=usd_goal_real-usd_pledged_real)
ggplot(emp33,aes(diferencia_1)) + stat_ecdf(geom = "step")
```



Grupo 3

```
emp44<-emp4%>%
   mutate(diferencia_1=usd_goal_real-usd_pledged_real)
ggplot(emp44,aes(diferencia_1)) + stat_ecdf(geom = "step")
```



Grupo 4

Las campañas tienen diferencias positivas.