Analisis_Exploratorio_Datos_Limpios

Análisis Exploratorio de los Datos

En la sección anterior de este documento, todos los archivos proporcionados por la compañÍa de telecomunicaciones *ABCD* fueron sometidos a una limpieza de datos con el fin de estructurarlos de tal manera que su manipulación se pueda ejecutar de forma más sencilla.

Una vez que ya se tienen los datos en un formato homogeneizado se procede a hacer un Análisis Exploratorio de los Datos, análisis que consiste en examinar los datos previamente a la aplicación de cualquier modelo que proporcione una solución a la problemática de la empresa. La finalidad de dicho análisis es extraer información relevante de los datos en crudo que permita adquirir un entendimiento general de estos e indagar en el comportamiento de las variables en cuestión y su relación en conjunto.

A continuación se presenta el Análisis Exploratorio de los Datos de Venta de la compañía *ABCD*.

Los Datos

La compañía de telecommunicaciones *ABCD* posee un registro de sus ventas que se ve de la siguiente manera:

```
#Lectura de datos de la tabla con variables relacionadas con el nivel socioeconomico por estado
tabla_final_2 <- read_csv("DATOS_LIMPIOS_PARA_EDA.csv")
```

```
#head(tabla_final_2)
```

El registro consta de 932,963 observaciones y 13 variables.

```
dim(tabla_final_2)
```

```
## [1] 932963     13
```

```
names(tabla_final_2)
```

```
##  [1] "punto_de_venta" "fecha"          "mes"            "anio"
##  [5] "sku"            "marca"          "gamma"          "zona"
##  [9] "estado"         "ciudad"         "latitud"        "longitud"
## [13] "ventas_diarias"
```

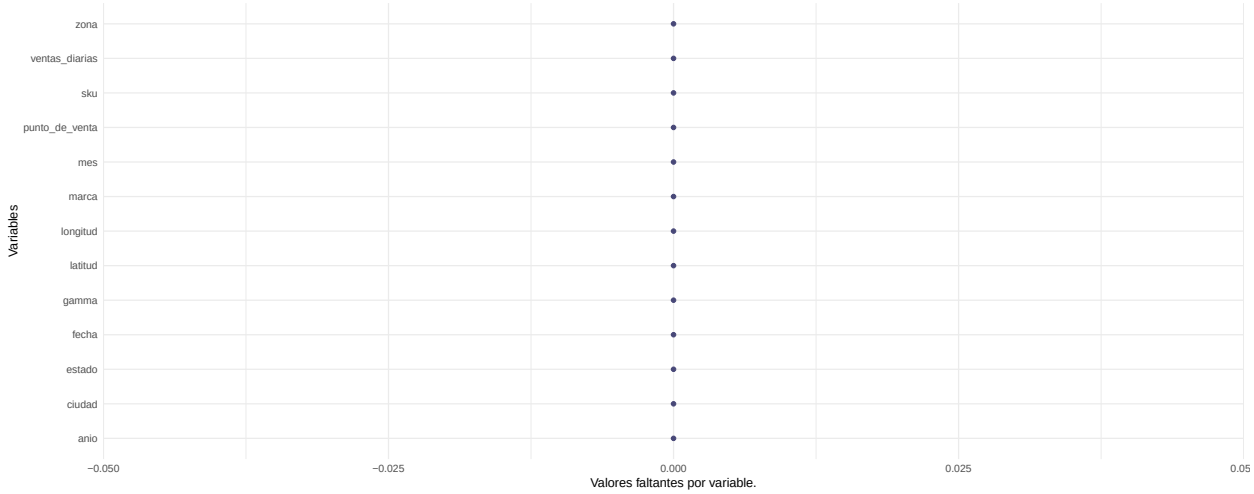
Con esta visualización general de los datos se pueden estructurar preguntas que permitan entender mejor el contenido dentro de este documento. A continuación se enlistan las preguntas base para comenzar con la comprensión de los datos.

Preguntas Base

1. ¿Hay campos vacíos?

Tras la limpieza de datos, a estas alturas los registros **no** contienen campos vacíos en ninguna variable.

```
p1
```



2. ¿Cuántos puntos de venta distintos hay?

Hay **1,909** puntos de venta distintos.

```
nrow(p2)
```

```
## [1] 1909
```

3. ¿Rango de fechas de los datos?

El conjunto de observaciones abarcan 10 meses en total, del primero de junio del 2018 al 31 de marzo del 2019.

```
p3[c(1,301),]
```

```
## # A tibble: 2 x 1
##   fecha
##   <date>
## 1 2018-06-01
## 2 2019-03-31
```

4. ¿Cuántos productos distintos hay?

Hay **455** productos distintos, cada uno identificado por un SKU único, que la compañía vende.

```
p4 <- tabla_final_2%>%select(sku)%>%group_by(sku)%>%unique()
nrow(p4)
```

```
## [1] 455
```

5. ¿Cuántas marcas vende la compañía y cuáles son?

La compañia *ABCD* cuenta con productos de **12** marcas distintas.

```r
nrow(p5)
```

```
## [1] 12
```

```r
p5%>%arrange(marca)
```

```
## # A tibble: 12 x 1
## # Groups:   marca [12]
##    marca
##    <chr>
##  1 affix
##  2 alcatel
##  3 apple
##  4 hisense
##  5 huawei
##  6 lanix
##  7 lenovo
##  8 lg
##  9 motorola
## 10 samsung
## 11 sony
## 12 zte
```

6. ¿Cuántas gammas consideran los datos?

Son **4** las gammas en las que se agrupan los productos.

```r
p6
```

```
## # A tibble: 4 x 1
## # Groups:   gamma [4]
##   gamma
##   <chr>
## 1 baja
## 2 media
## 3 premium
## 4 alta
```

8. ¿En cuántas zonas esta dividido el territorio?

Son **8** las zonas en las que esta dividido el territorio.

```r
p8
```

```
## # A tibble: 8 x 1
## # Groups:   zona [8]
##   zona
##   <chr>
## 1 centro sur
## 2 centro occidente
## 3 golfo de mexico
## 4 norte
```

```
## 5 pacifico sur
## 6 peninsula de yucatan
## 7 noreste
## 8 noroeste
```

A continuación se presenta la división territorial de la República Mexicana en las 8 zonas delimidas por la *Comisión Nacional para el Conocimiento y Uso de la Biodiversidad.*

```
knitr::include_graphics("mapa_nuevo.png")
```

9. ¿En cuántos estados tiene presencia la compañía?

La compañía tiene presencia en los **32** estados de la república.

```
nrow(p9)
```

```
## [1] 32
```

10. ¿En cuántas ciudades tiene presencia la compañía?

son **228** las ciudades en las que la compañía tiene presencia.

```
nrow(p10)
```

```
## [1] 228
```

Información más detallada

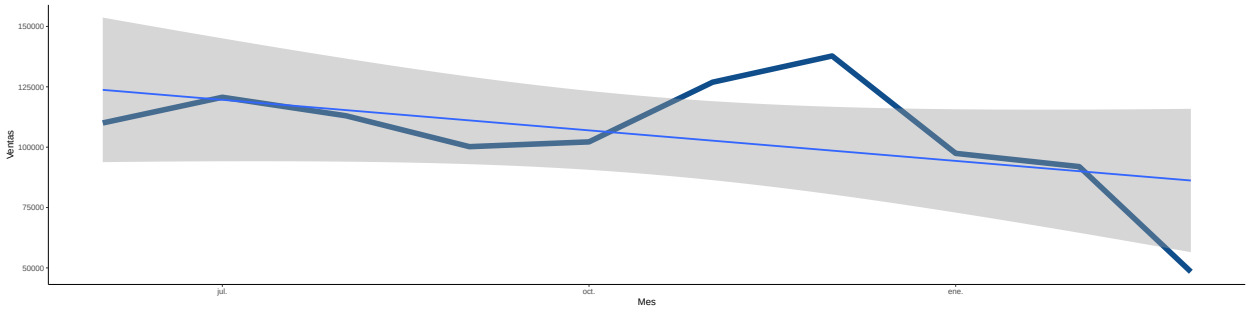
```
tabla_final_2
```

```
## # A tibble: 932,963 x 13
##    punto_de_venta fecha        mes  anio sku   marca gamma zona  estado
##    <chr>          <date>     <dbl> <dbl> <chr> <chr> <chr> <chr> <chr>
##  1 1 poniente     2018-06-01     6  2018 N.HU~ huaw~ baja  cent~ puebla
##  2 1 poniente     2018-06-01     6  2018 N.MZ~ moto~ baja  cent~ puebla
##  3 1 poniente     2018-06-01     6  2018 N.ZV~ zte   baja  cent~ puebla
##  4 1 poniente     2018-06-02     6  2018 N.HU~ huaw~ media cent~ puebla
##  5 1 poniente     2018-06-02     6  2018 N.HU~ huaw~ baja  cent~ puebla
##  6 1 poniente     2018-06-04     6  2018 N.HU~ huaw~ baja  cent~ puebla
##  7 1 poniente     2018-06-04     6  2018 N.MO~ moto~ baja  cent~ puebla
##  8 1 poniente     2018-06-04     6  2018 N.ZV~ zte   baja  cent~ puebla
##  9 1 poniente     2018-06-05     6  2018 N.HU~ huaw~ baja  cent~ puebla
## 10 1 poniente     2018-06-05     6  2018 N.MO~ moto~ baja  cent~ puebla
## # ... with 932,953 more rows, and 4 more variables: ciudad <chr>,
## #   latitud <dbl>, longitud <dbl>, ventas_diarias <dbl>
```

13. ¿Cómo se comportan las ventas totales de la compañía por mes?

Con relación a las ventas totales de la compañía *ABCD* en los 10 meses de registro, se puede observar un incremento de ventas en los meses de noviembre y diciembre, seguido por una caida drástica en los 3 meses siguientes.

```
ggplot(p13, aes(x=month, y = amount))+geom_line(color='dodgerblue4', size=3)+theme_classic() + geom_smo
```



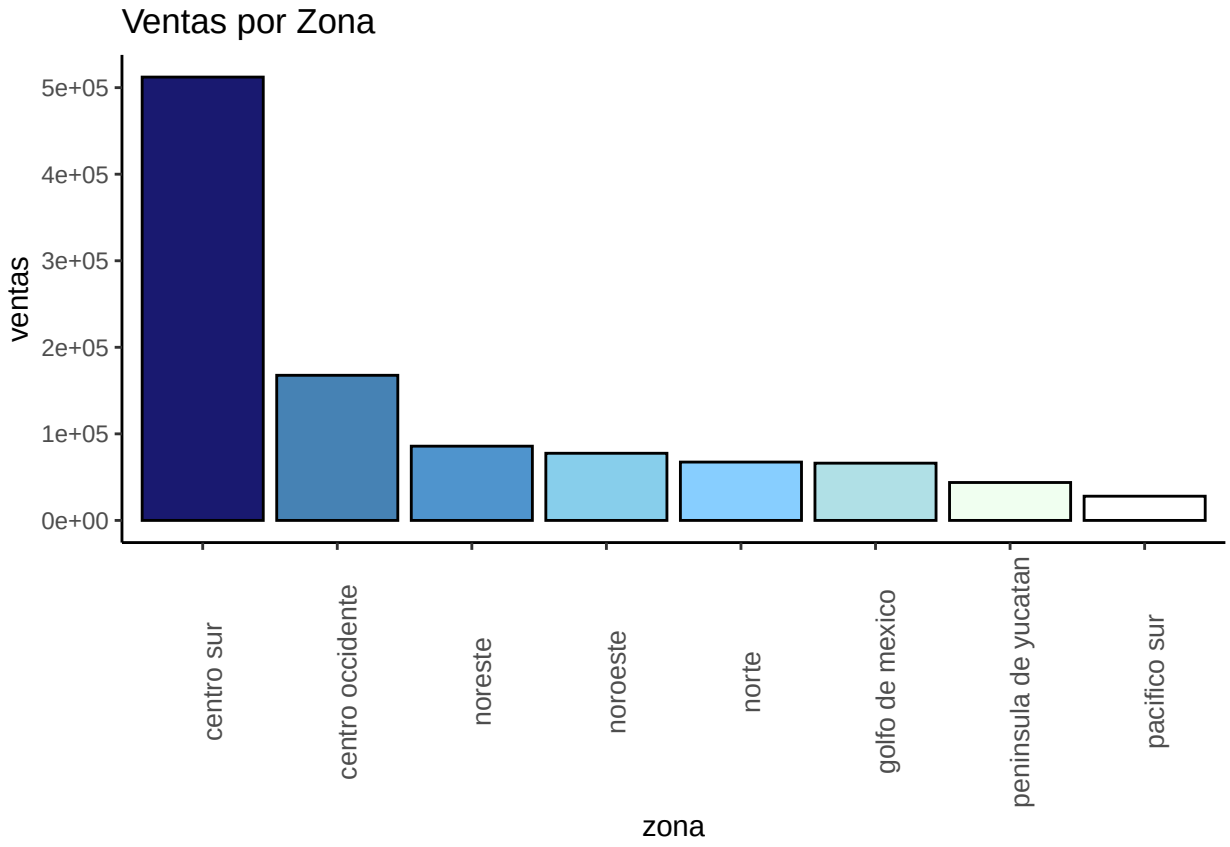
14. ¿Cuál es la región con mayor número de ventas en los meses de registro?

La región con mayor número de ventas es la zona **centro sur** con un total de **512,223** ventas en los 10 meses de registro, equivalente al **48.85%** de las ventas registradas. Seguido se encuentra la zona **Centro occidente** con **167,461** ventas, equivalente al **15.97%** de las ventas registradas.

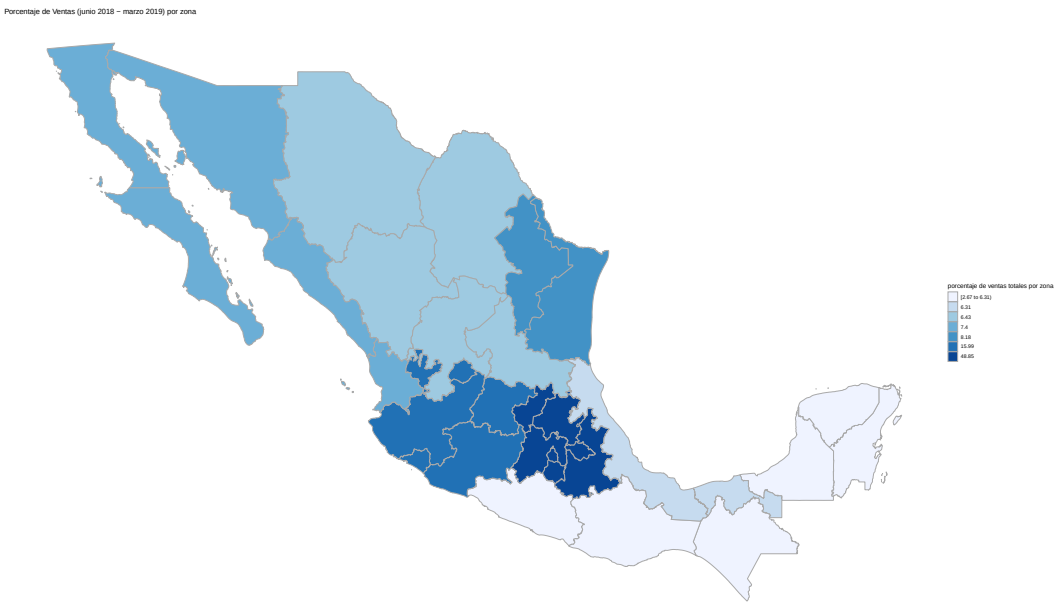
```
p14 <- p14 %>% mutate(porcentaje = ventas_totales_zona*100/1048575)
p14
```

```
## # A tibble: 8 x 3
##   zona                 ventas_totales_zona porcentaje
##   <chr>                              <dbl>      <dbl>
## 1 centro sur                        512223      48.8
## 2 centro occidente                  167655      16.0
## 3 noreste                            85779       8.18
## 4 noroeste                           77573       7.40
## 5 norte                              67410       6.43
## 6 golfo de mexico                    66143       6.31
## 7 peninsula de yucatan               43813       4.18
## 8 pacifico sur                       27979       2.67
```

```
q <- ggplot(data=p14, aes(x=reorder(zona,-ventas_totales_zona), y=ventas_totales_zona, fill=zona)) +
    geom_bar(colour="black", stat="identity") +
    guides(fill=FALSE)+
    scale_fill_manual(values=c("steelblue", "midnightblue","powderblue","steelblue3","skyblue","skyblue
q+labs(x="zona", y="ventas",title="Ventas por Zona")+theme_classic()+theme(axis.text.x = element_text(a
```




```
mxstate_choropleth(mapa1, num_colors = 8, legend="porcentaje de ventas totales por zona", title="Porcen
```



15. ¿Cuáles son los estados con mayor número de ventas en los 10 meses de registro?

Los tres estados que registraron mayor número de ventas en los 10 meses de registro con: Cdmx, estado de mexico y jalisco; y los estados que registraron menor nùmero de ventas en los ùltimos 10 meses son: durango, baja california sur y zacatecas

```
p15%>%arrange(estado)
```

```
##    region             estado  value
## 1      01      aguascalientes  10796
## 2      02     baja california  33550
## 3      03 baja california sur   4781
## 4      04            campeche   8818
## 5      09                cdmx 207187
## 6      07             chiapas  12533
## 7      08           chihuahua  19988
## 8      05            coahuila  18734
## 9      06              colima   9358
## 10     10             durango   5213
## 11     15    estado de mexico 174189
## 12     11          guanajuato  60194
## 13     12            guerrero   9187
## 14     13             hidalgo  19360
## 15     14              jalisco  71879
## 16     16           michoacan  15428
## 17     17             morelos  29704
## 18     18             nayarit   5244
## 19     19          nuevo leon  56557
## 20     20              oaxaca   6259
## 21     21              puebla  48496
## 22     22           queretaro  24212
## 23     23        quintana roo  17932
## 24     24     san luis potosi  18842
## 25     25             sinaloa  15210
```

```
## 26     26              sonora  18788
## 27     27             tabasco  11543
## 28     28          tamaulipas  29222
## 29     29            tlaxcala   9075
## 30     30            veracruz  54600
## 31     31             yucatan  17063
## 32     32           zacatecas   4633
```

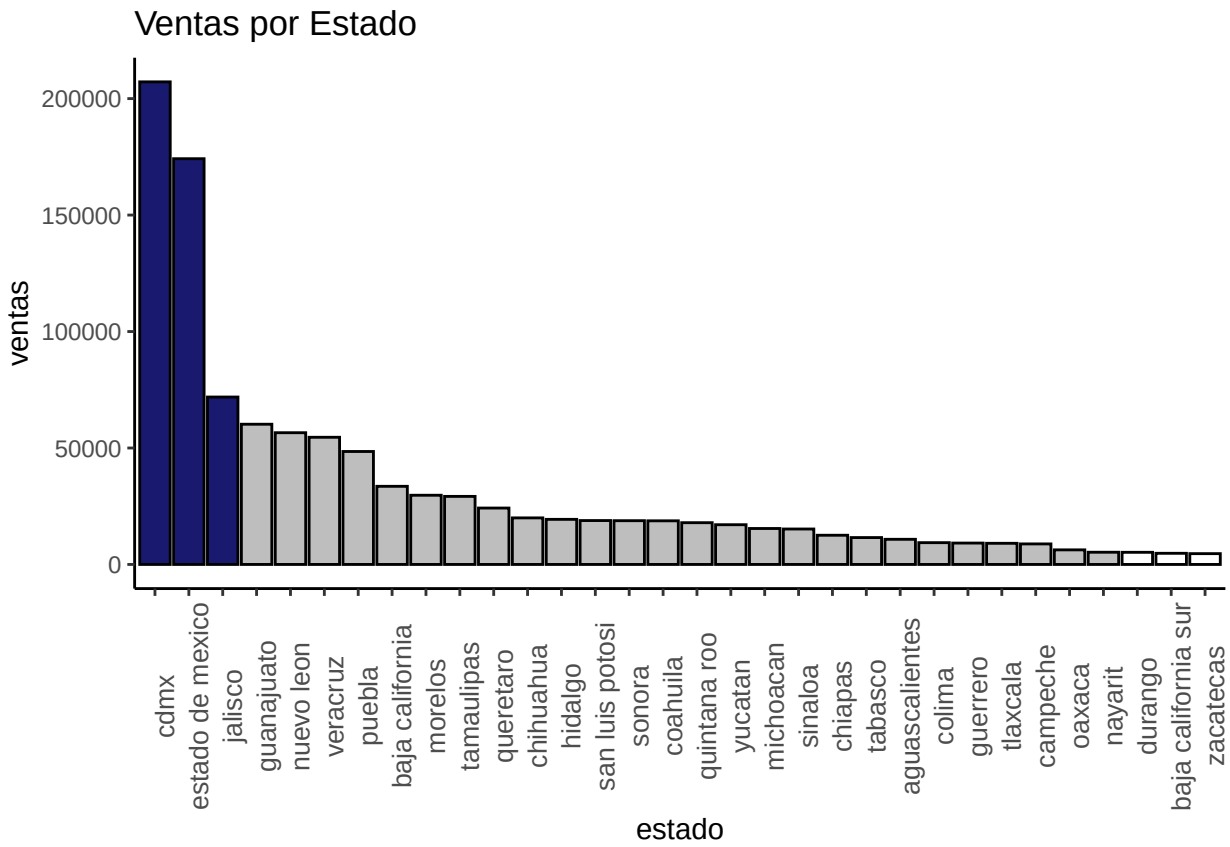
```
head(p15[,2:3],3)
```

```
##             estado  value
## 1              cdmx 207187
## 2 estado de mexico 174189
## 3          jalisco  71879
```

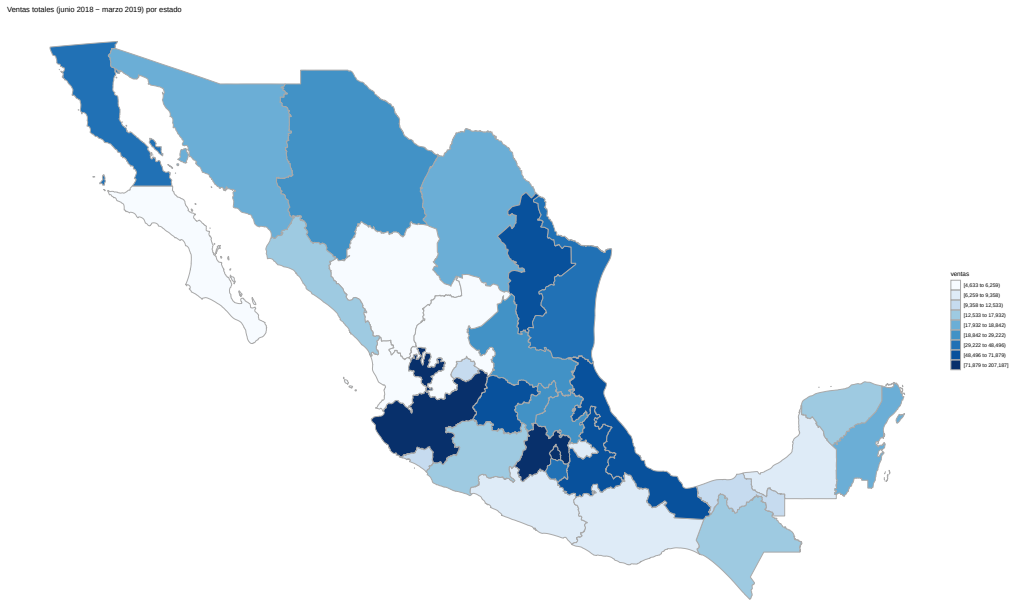
```
tail(p15[,2:3],3)
```

```
##                  estado value
## 30             durango  5213
## 31 baja california sur  4781
## 32           zacatecas  4633
```

```
a <- ggplot(data=p15, aes(x=reorder(estado,-value), y=value, fill=estado)) +
    geom_bar(colour="black", stat="identity") +
    guides(fill=FALSE)+
    scale_fill_manual(values=c("grey", "grey","white","grey","midnightblue","grey","grey","grey","grey"
a+labs(x="estado", y="ventas",title="Ventas por Estado")+theme_classic()+theme(axis.text.x = element_te
```




```
mxstate_choropleth(mapa2, num_colors = 9, legend="ventas", title="Ventas totales (junio 2018 - marzo 20
```



16. Identificar puntos de venta-

Son **1,909** puntos de venta en toda la República Mexicana.

```
nrow(p16)
```

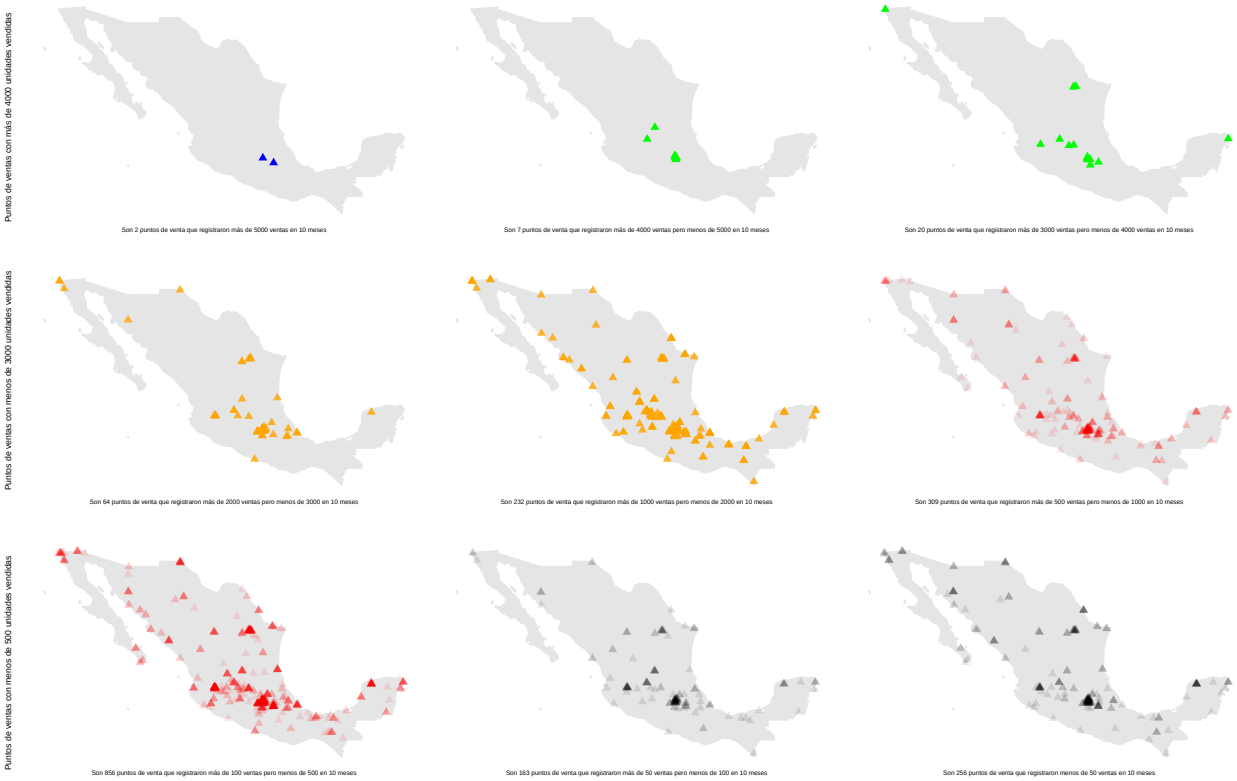
```
## [1] 1909
```

```
mapa3<- mapa3_pdv +
  geom_point(data=p16,
             aes(x=longitud, y=latitud),colour="blue",
             fill="blue",
             size=2,pch=17, alpha=I(0.7))+ggtitle("Localización de puntos de vent")
```

```
mapa3
```

17. Identificar puntos de venta con mayor unidades vendidas en los 10 meses de registro.

```
mapa4 <- ggarrange(mapa12, mapa11, mapa10, mapa9, mapa8, mapa7, mapa6, mapa5, mapa4, ncol=3, nrow=3)
annotate_figure(mapa4, bottom = text_grob("Puntos de venta de la compañía de telecomunicaciones ABCD dep
```



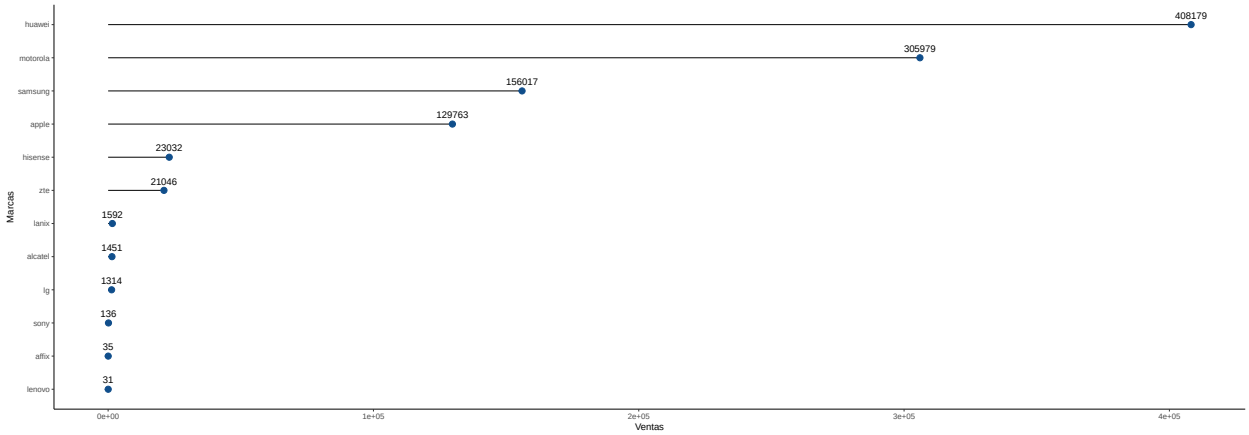
Puntos de venta de la compañía de telecomunicaciones ABCD dependiendo del número de unidades vendidas

19. Ventas por marca durante 10 meses

Los registros indican que en los últimos 10 meses la marca cuyas ventas son las mayores es Huawei, seguidas por Motorola y Samsung. De la misma manera se pude observar que hay 9 registros que poseen NA en el campo de Marca y que Lenovo y Affixx son las marcas con menos ventas.

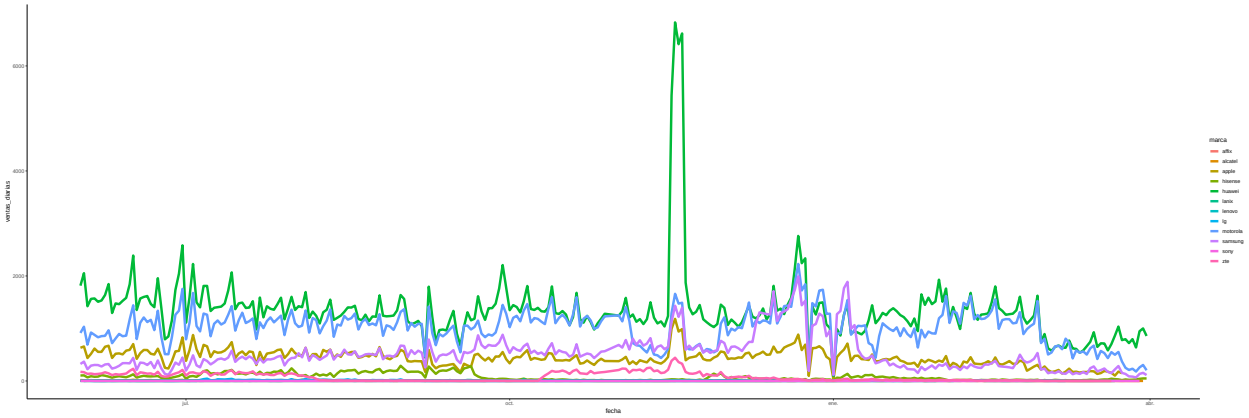
```
p19 <- tabla_final_2 %>% group_by(marca) %>% summarise(ventas_totales = sum(ventas_diarias)) %>% arrang
```

```
p19 %>% ggplot( aes(x=marca, y=ventas_totales, label=round(ventas_totales,1))) +
  geom_segment( aes(xend=marca, yend=3)) +
  geom_point( size=3, color="dodgerblue4") +
  coord_flip() +
  theme_classic() +
  xlab("Marcas")+
  ylab("Ventas")+
  geom_text(nudge_x = .28)
```



20. *Ventaspormarcapordía*

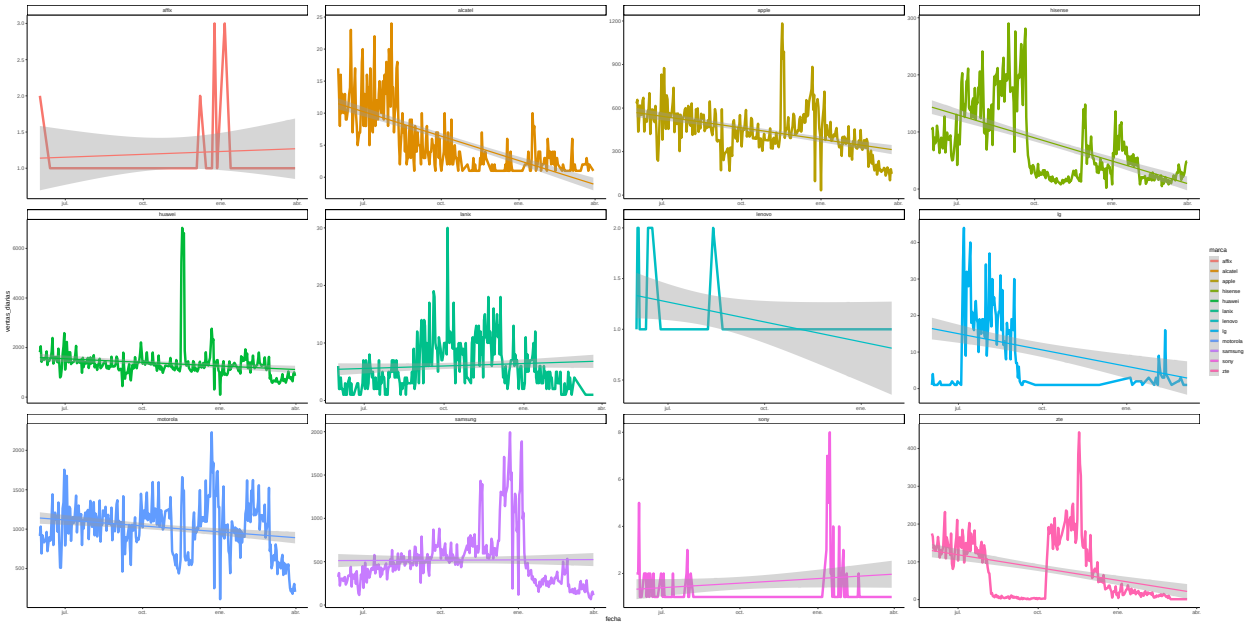
```
por_marca_por_dia<- tabla_final_2 %>%
  group_by(marca, fecha) %>%
  summarise(ventas_diarias = sum(ventas_diarias))

ggplot(por_marca_por_dia, aes(x=fecha, y = ventas_diarias, color=marca))+
  geom_line(size=2) +
  theme_classic()
```



Lo que se pude observar en esta gráfica es que, sin importar la marca, hay periodos en los que las ventas se disparan, por ejemplo: Diciembre.

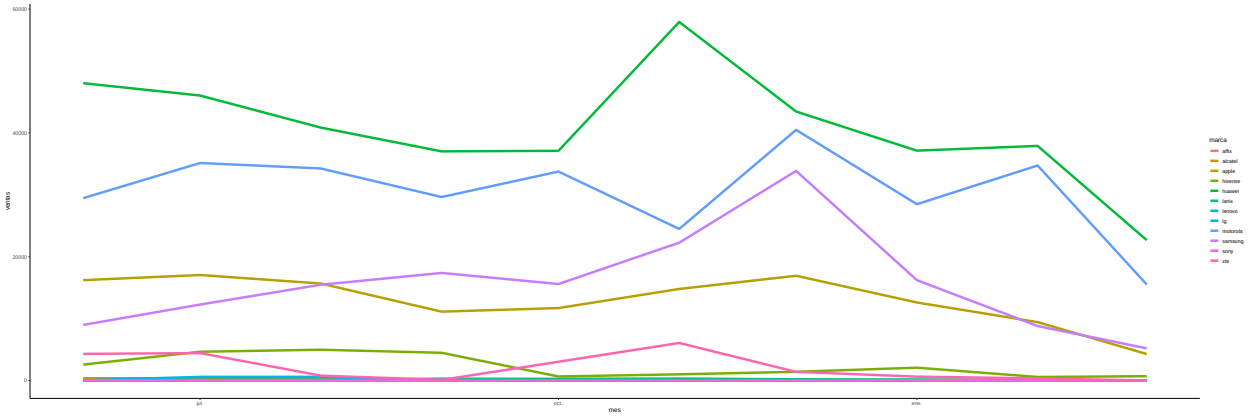
Sin embargo, para tener una visión más clara de lo que pasa en cada escenario de cada marca, se hace la siguiente gráfica:

```
ggplot(por_marca_por_dia, aes(x=fecha, y = ventas_diarias, color=marca))+
  geom_line(size=2) +
  facet_wrap(~marca, scales = "free") +
  geom_smooth(method='lm')+
  theme_classic()
```

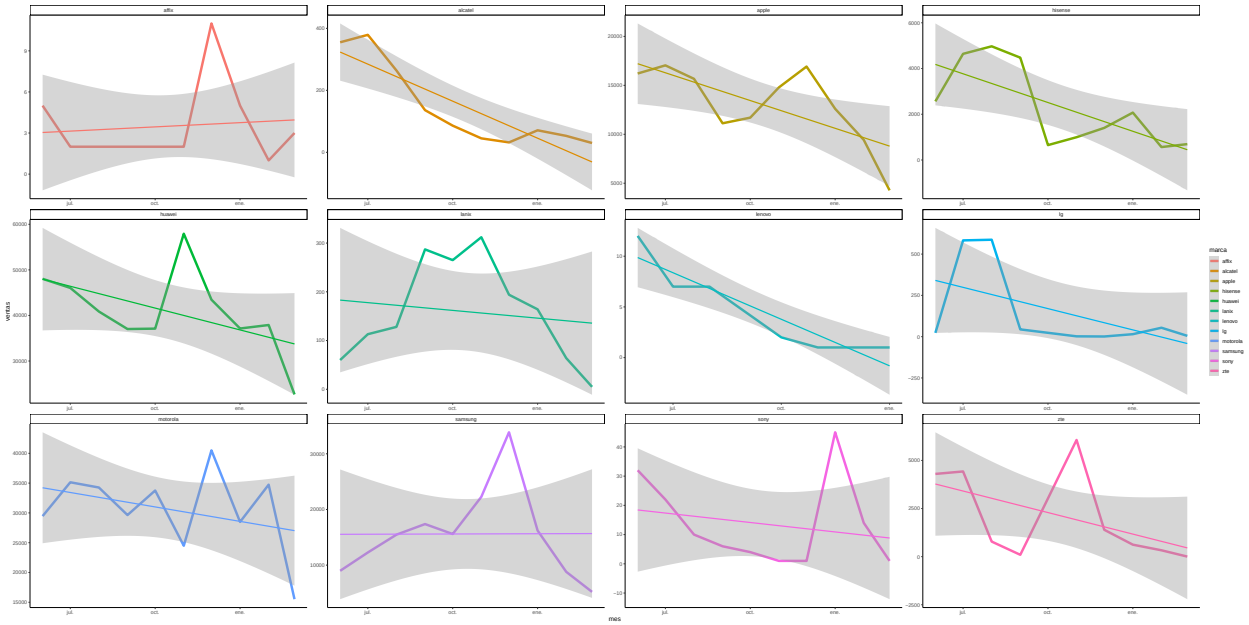


20_B. $Ventas totales por marca por mes $

```
ventas_marca_mes <- tabla_final_2 %>% group_by(marca, mes=floor_date(fecha, "month")) %>%
   summarize(ventas=sum(ventas_diarias))

ggplot(ventas_marca_mes, aes(x=mes, y = ventas, color=marca))+
  geom_line(size=2) +
  theme_classic()
```




```
ggplot(ventas_marca_mes, aes(x=mes, y = ventas, color=marca))+
  geom_line(size=2) +
  facet_wrap(~marca, scales = "free") +
  geom_smooth(method='lm')+
  theme_classic()
```



21. *VentasporSKUspormarca* Comportamiento de productos por marca en los últimos 10 meses

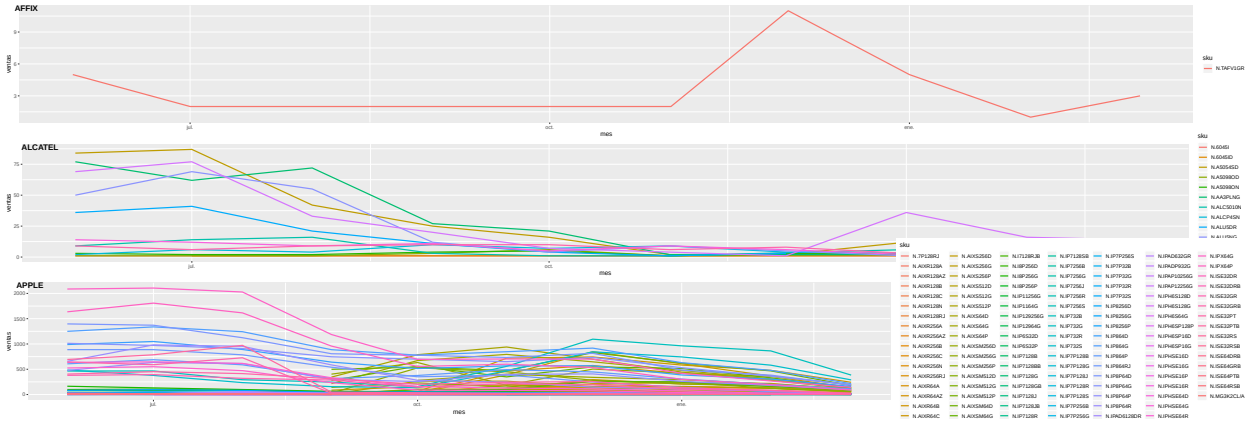
```
sku_marca <- tabla_final_2 %>% group_by(marca, sku, mes=floor_date(fecha, "month")) %>%
   summarise(ventas=sum(ventas_diarias))
#cuantos sku por marca hay
sku_marca%>%select(marca,sku)%>%unique()%>%mutate(tol=1)%>%group_by(marca)%>%summarise(n=sum(tol))%>%ar
```

```
## # A tibble: 12 x 2
##    marca        n
##    <chr>    <dbl>
##  1 apple      129
##  2 huawei      97
##  3 samsung     92
##  4 motorola    60
##  5 hisense     28
##  6 zte         14
##  7 alcatel     13
##  8 sony        10
##  9 lg           7
## 10 lanix        2
## 11 lenovo       2
## 12 affix        1
```

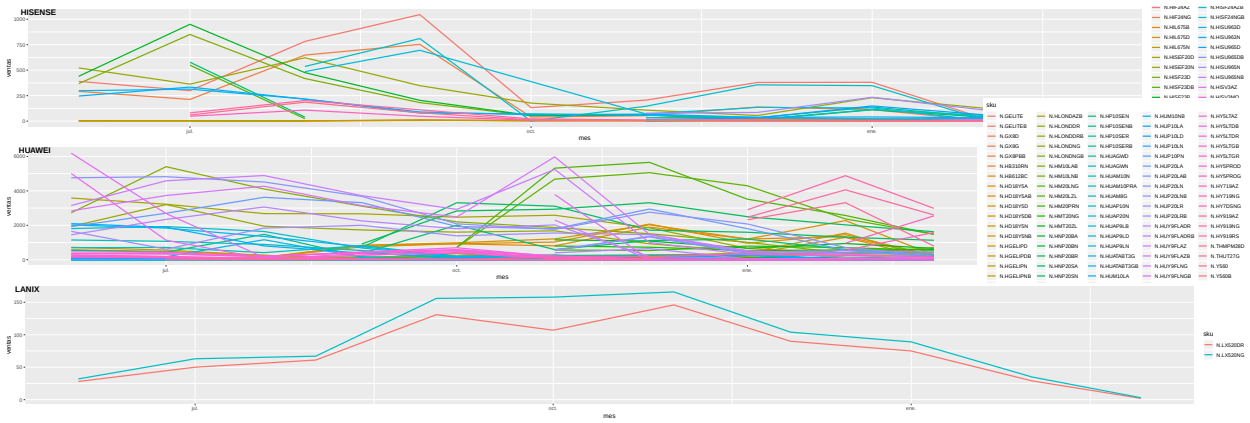
```
affix <- sku_marca%>%filter(marca=="affix")
alcatel <- sku_marca%>%filter(marca=="alcatel")
apple <- sku_marca%>%filter(marca=="apple")
Hisense <- sku_marca%>%filter(marca=="hisense")
Huawei <- sku_marca%>%filter(marca=="huawei")
Lanix <- sku_marca%>%filter(marca=="lanix")
Lenovo <- sku_marca%>%filter(marca=="lenovo")
LG <- sku_marca%>%filter(marca=="lg")
Motorola <- sku_marca%>%filter(marca=="motorola")
Samsung <- sku_marca%>%filter(marca=="samsung")
```

```
Sony <- sku_marca%>%filter(marca=="sony")
ZTE <- sku_marca%>%filter(marca=="zte")
m1<-ggplot(affix, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m2<-ggplot(alcatel, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m3<-ggplot(apple, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m4 <- ggplot(Hisense, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m5 <- ggplot(Huawei, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m6 <- ggplot(Lanix, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m7 <- ggplot(Lenovo, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m8 <- ggplot(LG, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m9 <- ggplot(Motorola, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m10 <- ggplot(Samsung, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m11 <- ggplot(Sony, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
m12 <- ggplot(ZTE, aes(x=mes, y=ventas, color=sku))+geom_line(size=1)
```

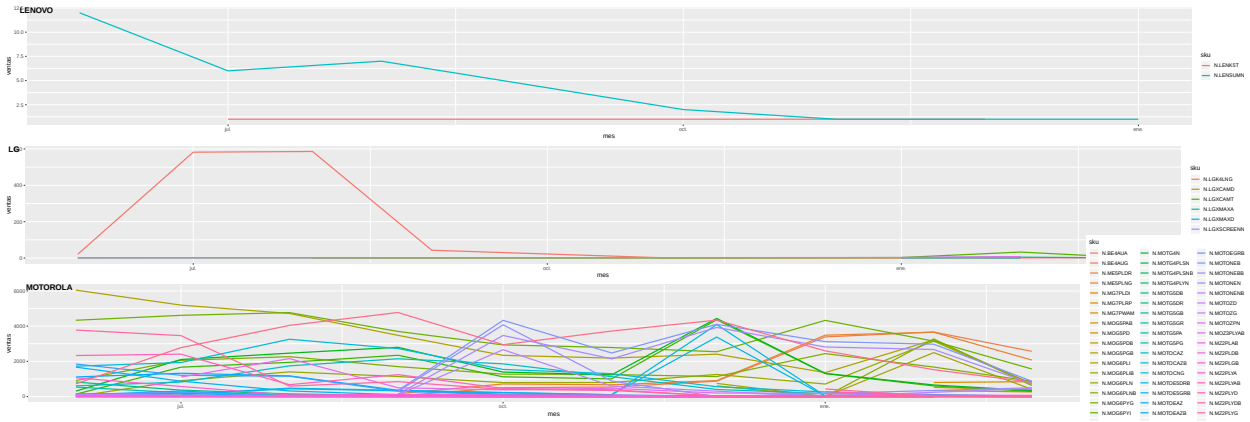
```
ggarrange(m1, m2, m3,
          labels = c("AFFIX", "ALCATEL", "APPLE"),
          ncol = 1, nrow = 3)
```



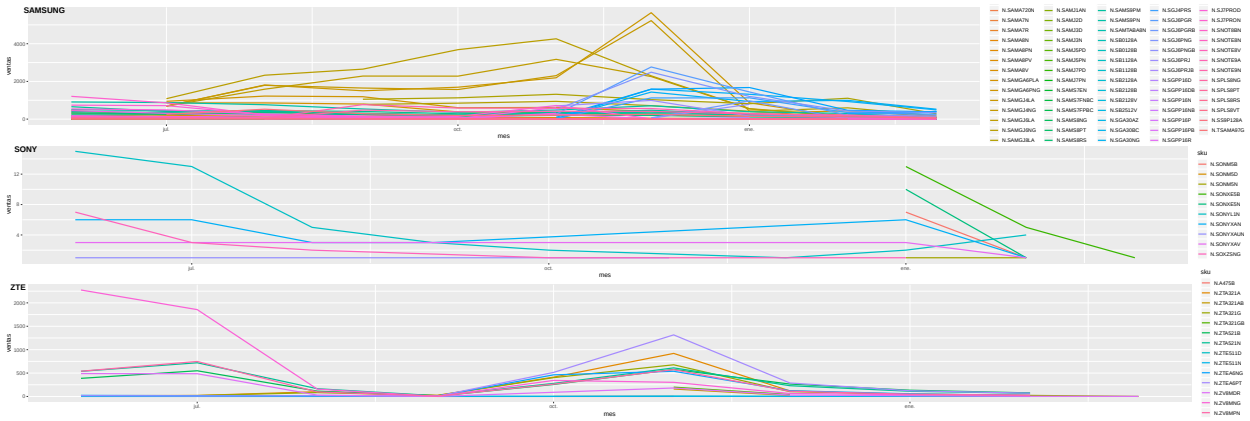

```
ggarrange(m4, m5, m6,
          labels = c("HISENSE", "HUAWEI", "LANIX"),
          ncol = 1, nrow = 3)
```



```
ggarrange(m7, m8, m9,
          labels = c("LENOVO", "LG", "MOTOROLA"),
          ncol = 1, nrow = 3)
```



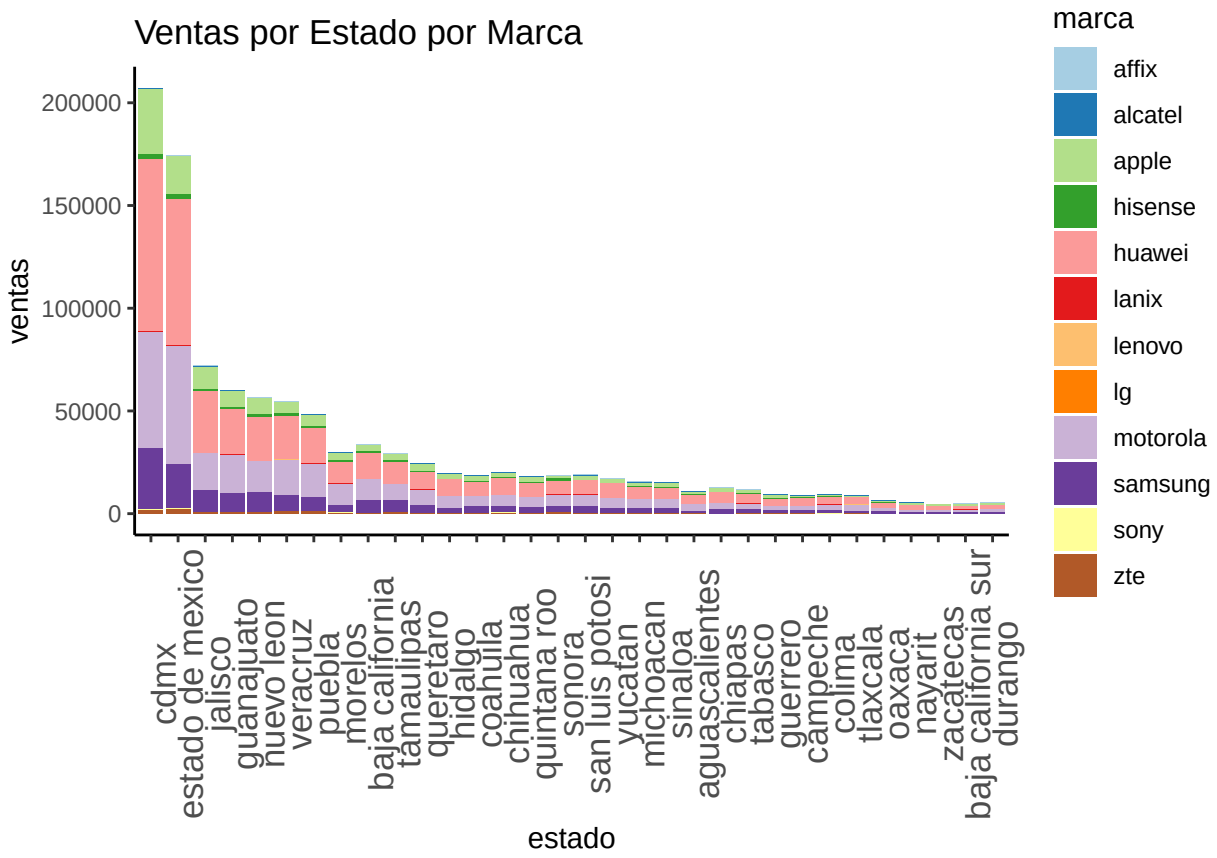

```
ggarrange(m10, m11, m12,
          labels = c("SAMSUNG", "SONY", "ZTE"),
          ncol = 1, nrow = 3)
```



22. Ventaspormarcayestado

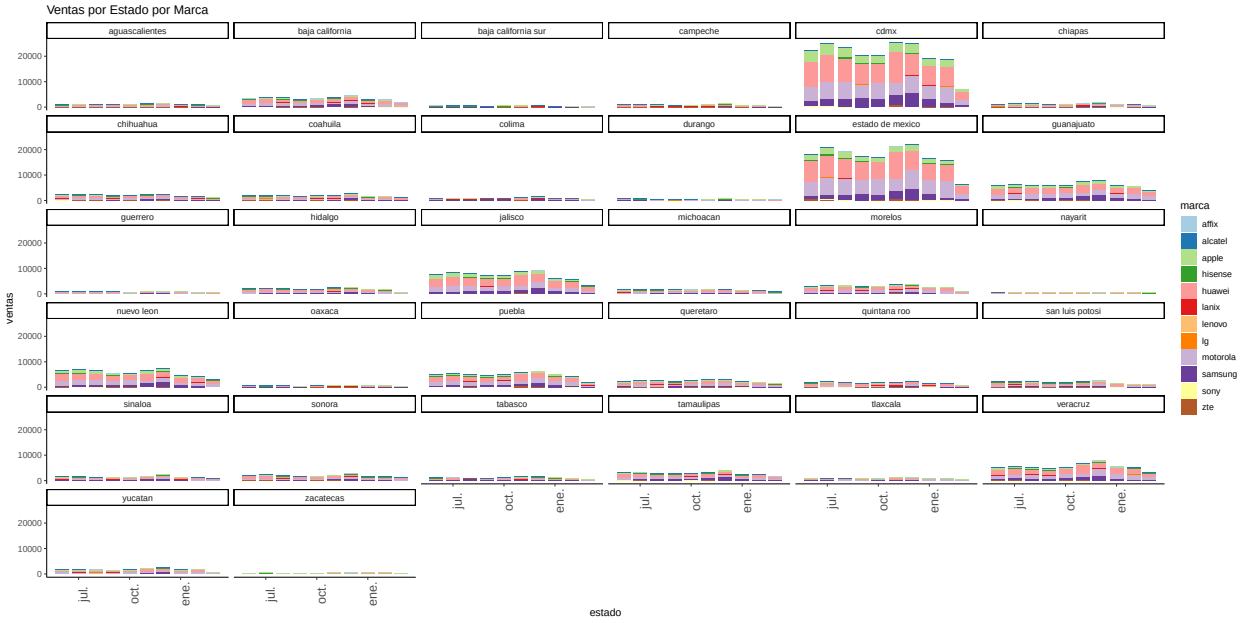
```
marca_estado<- tabla_final_2 %>% group_by(estado, marca) %>%
   summarise(ventas=sum(ventas_diarias))
```

```
library(RColorBrewer)
cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7",
x <- ggplot(marca_estado, aes(x = reorder(estado, -ventas), y = ventas, fill = marca)) +
  geom_bar(stat = "identity")+
  scale_fill_brewer(palette = "Paired")+theme_classic()
x+labs(x="estado", y="ventas",title="Ventas por Estado por Marca")+theme(axis.text.x = element_text(angl
```



23. $Ventaspormarca, estadoymes$

```
marca_estado<- tabla_final_2 %>% group_by(estado, marca, mes=floor_date(fecha, "month")) %>%
   summarise(ventas=sum(ventas_diarias))
ggplot(marca_estado, aes(x = mes, y = ventas, fill = marca)) +
  geom_bar(stat = "identity")+
  scale_fill_brewer(palette = "Paired")+theme_classic()+facet_wrap(~estado)+labs(x="estado", y="ventas"
```


Ventas por Estado por Marca
aguascalientes
baja california
baja california sur
campeche
cdmx
chiapas
chihuahua
coahuila
colima
durango
estado de mexico
guanajuato
guerrero
hidalgo
jalisco
michoacan
morelos
nayarit
nuevo leon
oaxaca
puebla
queretaro
quintana roo
san luis potosi
sinaloa
sonora
tabasco
tamaulipas
tlaxcala
veracruz
yucatan
zacatecas
0
10000
20000
ventas
jul.
oct.
ene.
estado
marca
affix
alcatel
apple
hisense
huawei
lanix
lenovo
lg
motorola
samsung
sony
zte