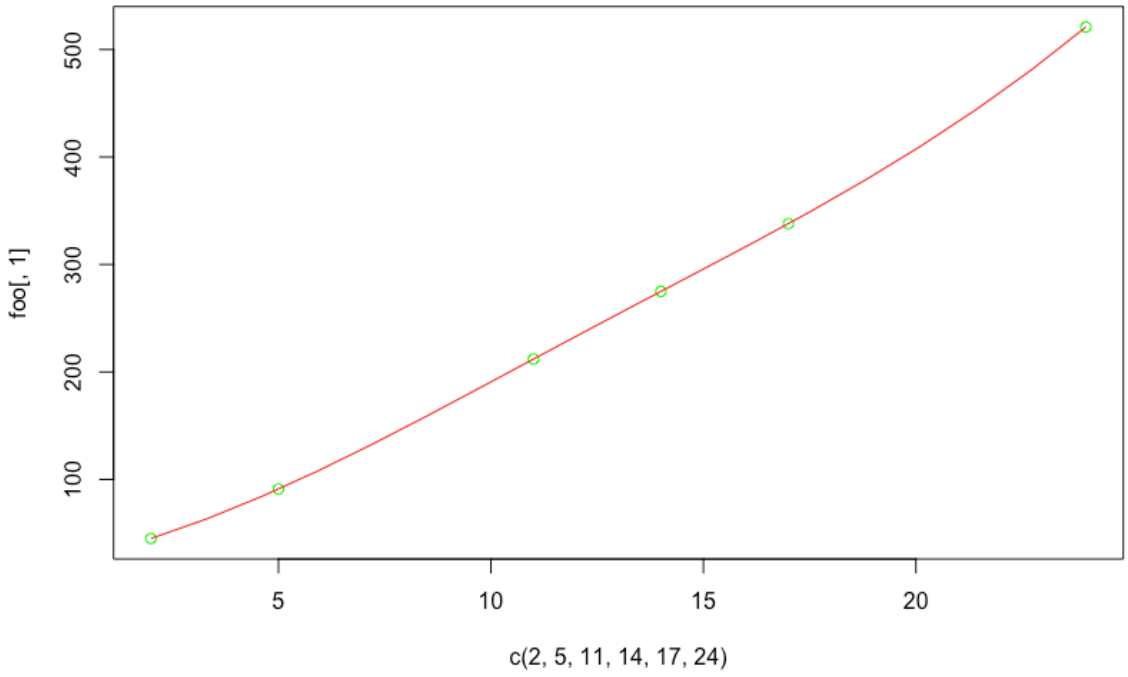
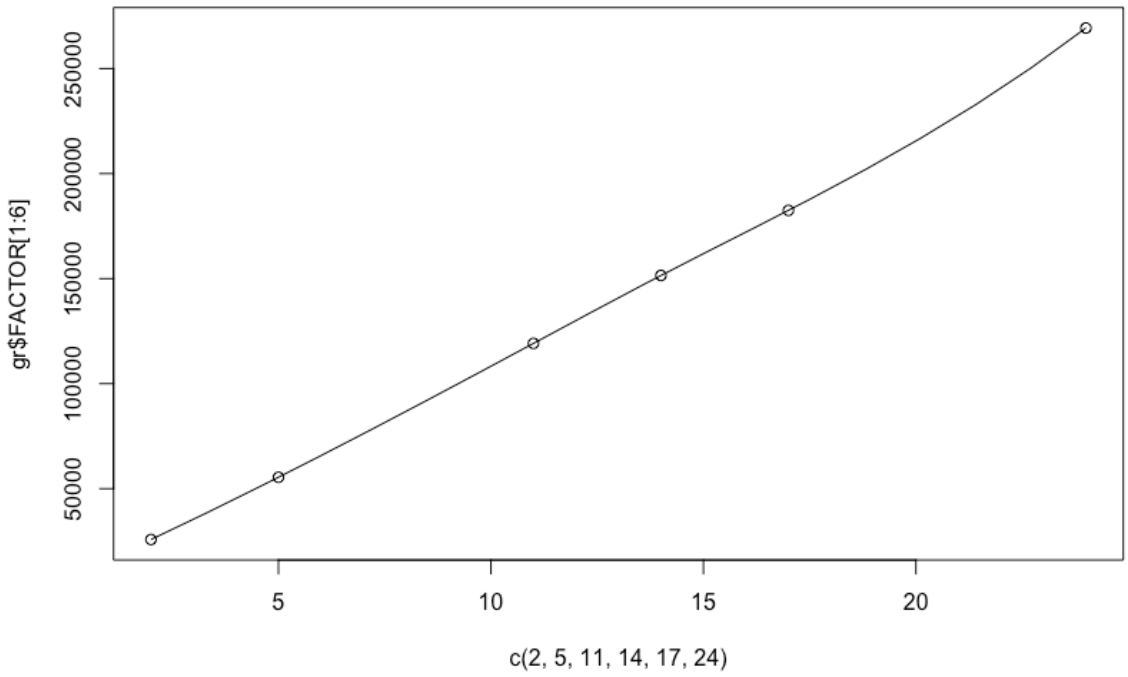
Sección 1: Bebés (Obligatorio) 1. Para cada AGEB de la delegación Álvaro Obregón estima cuántos bebés de 0 a 6 meses de edad habitan ahí el día de hoy. Explica tu razonamiento en menos de 300 palabras. Enlista tus fuentes y presenta los resultados. (Hint: revisa el CPV 2010 , puede ser útil)

Respuesta: Hay dos dimensiones a considerar en este problema. La primera es que la tasa de crecimiento entre grupos de edad no es igual. Es decir, no hay el doble de personas de 0 a dos años que de 0 a un año. La segunda es que tenemos la dimensión temporal. Hay que estimar qué tanto crece del 2010 (dato censal) al 2017 (estimación). Para ello, tomé los datos del censo 2010 por grupos de edad por ageb. Al tener la cantidad de personas para ciertos grupos de edad, puedo interpolar cuántos bebés de 0 a seis meses había en 2010 a nivel ageb, para la delegación Álvaro Obregón. Los resultados son los siguientes para una la suma acumulada de años en una ageb cualquiera:



Una vez hecho esto, necesitamos estimar cuánto creció este sector de la población en años siguientes. Para esto utilizo la fuente de mayor confianza disponible: la encuesta intercensal de 2015. Esta encuesta me permite, utilizando la ponderación correcta, estimar los mismos grupos de edad usados en la interpolación censal. Así, estimo el número de bebés de 0 a 6 meses en Álvaro Obregón. La interpolación municipal luce de la siguiente manera:



En tercer lugar, saco la tasa de crecimiento de estas dos poblaciones. Esto lo hice de dos formas: estimación logarítmica y estimación geométrica. Ambos métodos arrojan que, dados los dos puntos en el tiempo, la estimación de 2017 será la estimación de 2010 multiplicada por un factor de 0.94. Es decir, hay menos bebés con esa edad que en aquel año. El supuesto más grande esta estimación es que las tasas de crecimiento de las agebs individuales no están lejos de la tasa de crecimiento del municipio total. Las fuentes utilizadas son las siguientes: https://www.researchgate.net/publication/26438106_Spline_Interpolation_for_Demographic_Variables_The_Monotonicity_Problem: que justifica la interpolación entre grupos de edad.

Código:

```
rm(list=(ls()))
wd<-"~/Documents/WD"
setwd(wd)
##########DESCARGAR CENSO (2010) Y ENCUESTA INTERCENSAL (2015)
url<-
c("http://www.beta.inegi.org.mx/contenidos/proyectos/ccpv/2010/microdatos/ageb_y_manzana/resageburb_09_2010_dbf.zip",

"http://www.beta.inegi.org.mx/contenidos/proyectos/enchogares/especiales/intercensal/2015/microdatos/eic2015_09_csv.zip")
temp<-c(tempfile(), tempfile())
lapply(1:length(url), function(x) download.file(url[x], temp[[x]], mode="wb"))
lapply(temp, unzip)
library(foreign)
```

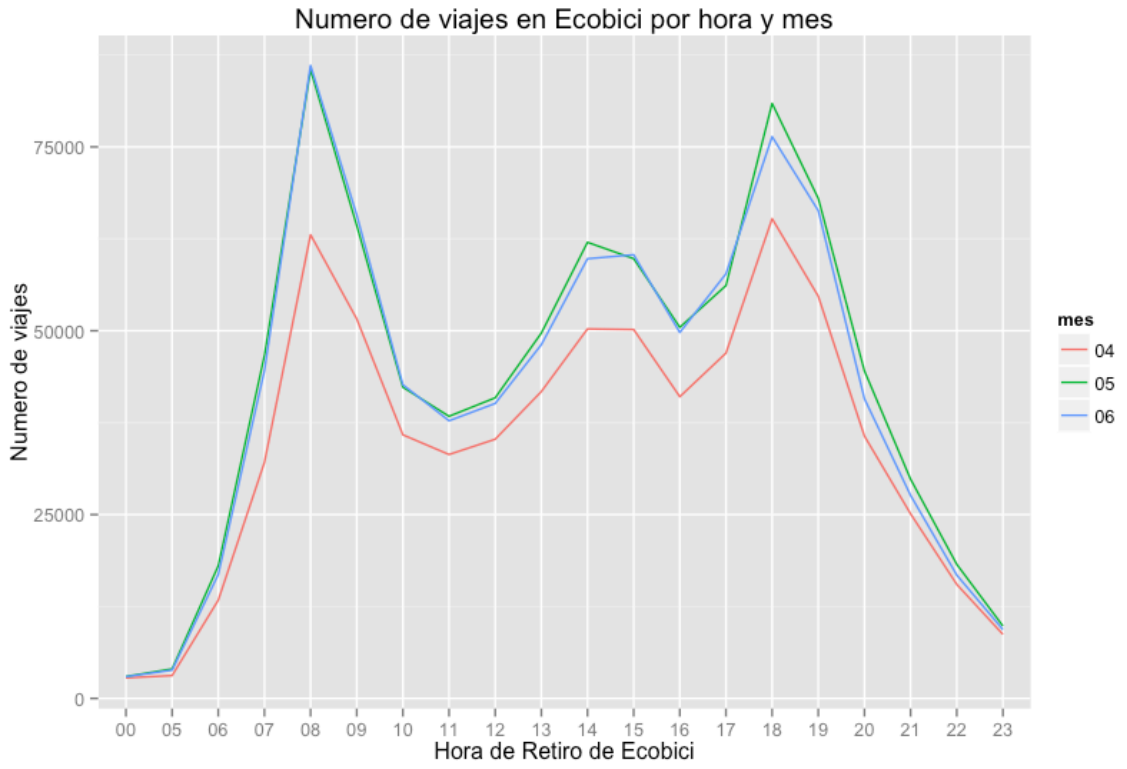
```
ccpv<-read.dbf("RESAGEBURB_09DBF10.dbf")
ein15<-read.csv("TR_PERSONA09.CSV")

##########PROCESAR CENSO 2010############
ids<-c("ENTIDAD", "MUN", "LOC", "AGEB")
edades<-c("P_0A2", "P_3A5", "P_6A11", "P_12A14", "P_15A17", "P_18A24")
#Datos ageb
ccpv<-ccpv[ccpv$AGEB!="0000",]
ccpv<-ccpv[ccpv$NOM_LOC=="Total AGEB urbana",]
#Delegacion Alvaro Obregon
ccpv$NOM_MUN<-iconv(as.character(ccpv$NOM_MUN), 'utf-8', 'ascii', sub="")
ccpv<-ccpv[ccpv$NOM_MUN=="lvaro Obregn",]

#Cambiar variables a class adecuada
ccpv<-cbind(ccpv[, ids], ccpv[edades], ccpv[, "POBTOT"])
ccpv[, ids]<-apply(ccpv[, ids], 2, as.character)
ccpv[,edades]<-apply(ccpv[,edades], 2, function(x) as.numeric(as.character(x)))
names(ccpv)[names(ccpv)=="ccpv[, \"POBTOT\"]"]<-"POBTOT"
ccpv$POBTOT<-as.numeric(as.character(ccpv$POBTOT))

#Suma acumulada de grupos de edad
foo<-apply(t(ccpv[, edades]), 2, cumsum)
#Prueba de interpolacion
plot(c(2, 5, 11, 14, 17, 24),  foo[,1], col="green")
lines(spline(c(2, 5, 11, 14, 17, 24), foo[,1], method="hyman"), col="red")
# interpolar suma acumulada de 0 a 0.5 (0.5 anios=6 meses)
#ref:
https://www.researchgate.net/publication/26438106_Spline_Interpolation_for_Dem
ographic_Variables_The_Monotonicity_Problem
ccpv$P_0A6m<-round(apply(foo, 2, function(x) spline(c(2, 5, 11, 14, 17, 24), x,
method="hyman", xout=0.5)$y))

##########PROCESAR INTERCENSAL 2015##########
#Delegacion Alvaro Obregon
ein15$NOM_MUN<-iconv(as.character(ein15$NOM_MUN), 'utf-8', 'ascii', sub="")
ein15<-ein15[ein15$NOM_MUN=="lvaro Obregn",]

#recodificar edad
library(car)
ein15$EDAD<-recode(ein15$EDAD, "0:2=1; 3:5=2; 6:11=3; 12:14=4; 15:17=5;
18:24=6; 25:999=7")

#colapsar factor de expansion por edad, para obtener los mismos cortes que en la
interpolacion 2010
#Solo podemos sacar resultado agregado porque la intercensal no es
representativa por ageb
gr<-aggregate(FACTOR~EDAD, data=ein15, sum)
```

```
gr$FACTOR<-cumsum(gr$FACTOR)
sum(ein15$FACTOR)

##########ESTIMACION############
plot(c(2, 5, 11, 14, 17, 24), gr$FACTOR[1:6])
lines(spline(c(2, 5, 11, 14, 17, 24), gr$FACTOR[1:6], method = "hyman"))

#obtener interpolacion 2015
bbs15<-round(spline(c(2, 5, 11, 14, 17, 24), gr$FACTOR[1:6], method = "hyman",
xout=0.5)$y)

#dato censal agregado 2010
bbs10<-sum(ccpv$P_0A6m, na.rm = T)

#comparar metodos log y geometrico para estimar el factor correspondiente a
2017
#REF
factor17.log<-(bbs15/bbs10)^((2017-2015)/(2015-2010))
r<-(bbs15/bbs10)^(1/(2015-2010))-1
factor17.geom<-(1+r)^(2017-2015)
#media (aunque son iguales)
factor.17<-mean(factor17.log, factor17.geom)

#Estimacion final 2017. SUPUESTO: EL FACTOR DE 2017 AGREGADO APLICA
PARA EL DATO POR AGEB.Las agebs no son muy
#diferentes de la grand mean.
out<-cbind(ccpv[ids],  ccpv[, "P_0A6m"], round(ccpv[, "P_0A6m"]*factor.17))
names(out)<-c(ids, "pob.0a6meses.2010", "pob.esp.0a6meses.2017")
summary(out)
out$pob.0a6meses.2010[out$pob.0a6meses.2010<0] <- NA
out$pob.esp.0a6meses.2017[out$pob.esp.0a6meses.2017<0] <- NA

plot(out$pob.0a6meses.2010, out$pob.esp.0a6meses.2017,
main="Comparacion estimacion 2017 v. Censo 2010",
xlab="Estimacion Censo 2010", ylab="Estimacion 2017", sub="linea=x=y")
lines(0:250, 0:250, col="red")
View(out)
```

Pregunta 2: ECOBICI

1. ¿En qué horarios hay mayor afluencia y en qué estaciones? Da una breve descripción de por qué crees que es así

Los horarios de mayor afluencia son las ocho de la mañana y las seis de la tarde, con un pequeño pico entre la 1:30 y las 4 pm. Esto coincide con el horario en que la gente entra y sale de trabajar. Además de la hora de comida



Esta es una lista con las 10 estaciones con más viajes en los útimos tres meses:

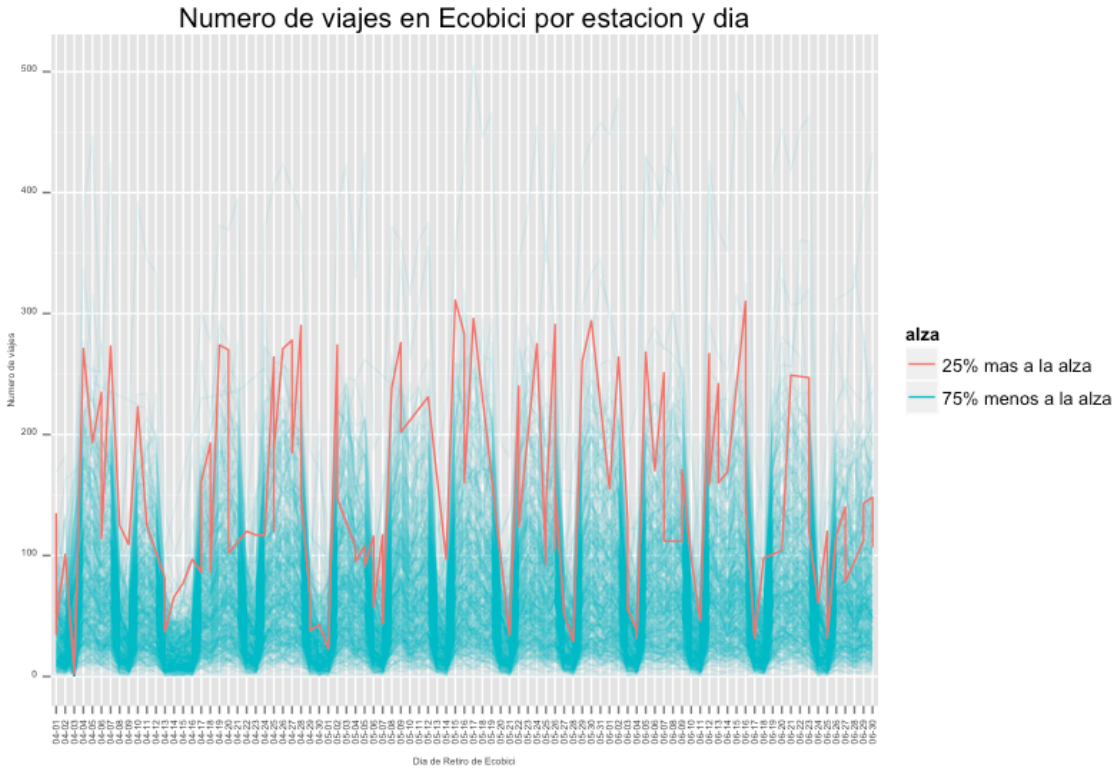
Estación	Frecuencia
27	28989
271	19294
1	18616
18	17833
15	15278
21	15270
64	15156
41	15131
25	15127
36	15018

2. A partir de un análisis temporal:

a. ¿En qué estaciones puedes observar una tendencia de uso a la alta?

En esta gráfica se aplicó un modelo lineal para cada cicloestación. La variable dependiente es el número de viajes realizados, mientras que las variables independientes son el día y una variable dicotómica que toma el valor de uno si es fin de semana y cero si no lo es, esto para controlar por la estacionalidad. Después, extraje los coeficientes de regresión de la variable "día" y obtuve las estaciones donde la beta está en el percentil 75 de la distribución de las betas. Estas fueron definidas como estaciones 75% a la alza. Después, graficamos los resultados (promedio de viajes del 75% a la alza y todas las curvas de las demás estaciones). Si bien podemos notar que los viajes de las estaciones a la alza son ligeramente más alto que los de los demás, no se observa una tendencia claramente a la alza.

De este modo, las estaciones resultantes son las número 1 2 3 4 6 7 8 9 10 13 14 15 16 18 19 20 22 23 25 26 27 28 32 36 37 38 39 42 44 45 46 47 52 54 56 59 60 63 64 66 68 71 73 74 77 79 82 84 86 96 107 108 115 119 120 124125 126 141 142 145 146 154 155 158 160 162 167 174 175 178 180 182 185 191 193 197 205 207 208 209 211 212 217 219 226 235 236 237 238 242 245 247 248 250 253 254 255 257 258 261 266 267 270 271 272 291 295 305 390 396 449 452



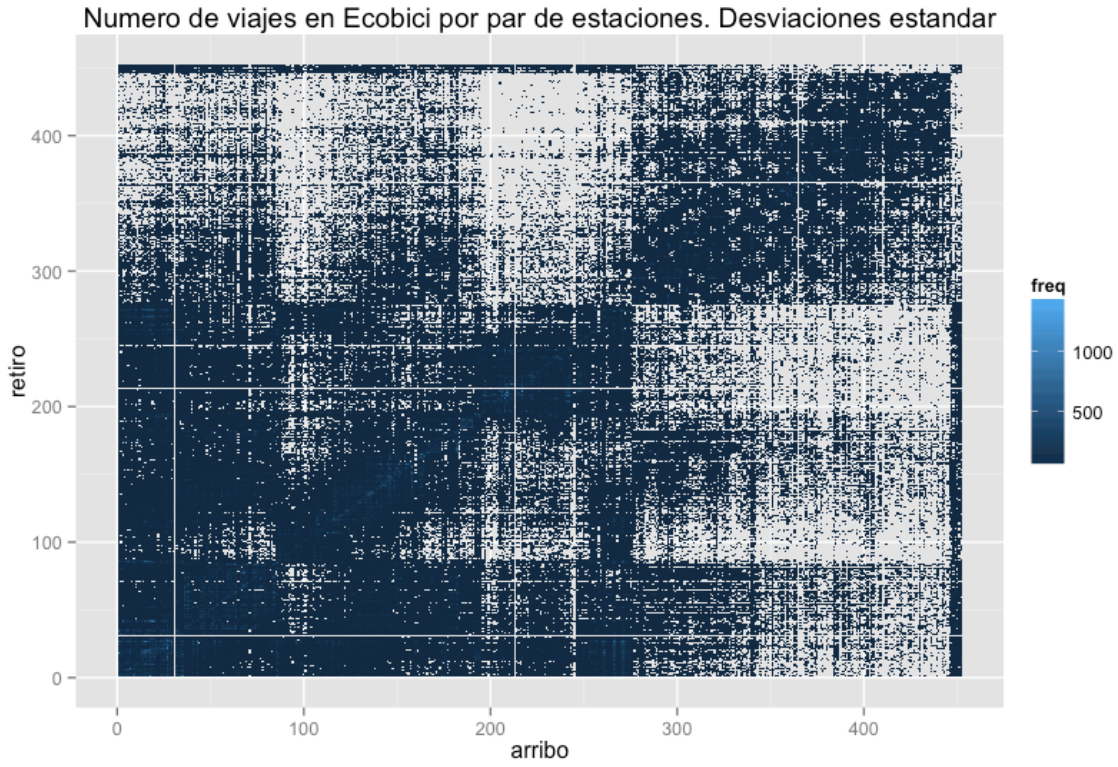
b. ¿Puedes categorizar las estaciones con base en su tendencia de uso?

Ver inciso A

c. Demuestra tus conclusiones gráficamente

Ver inciso A

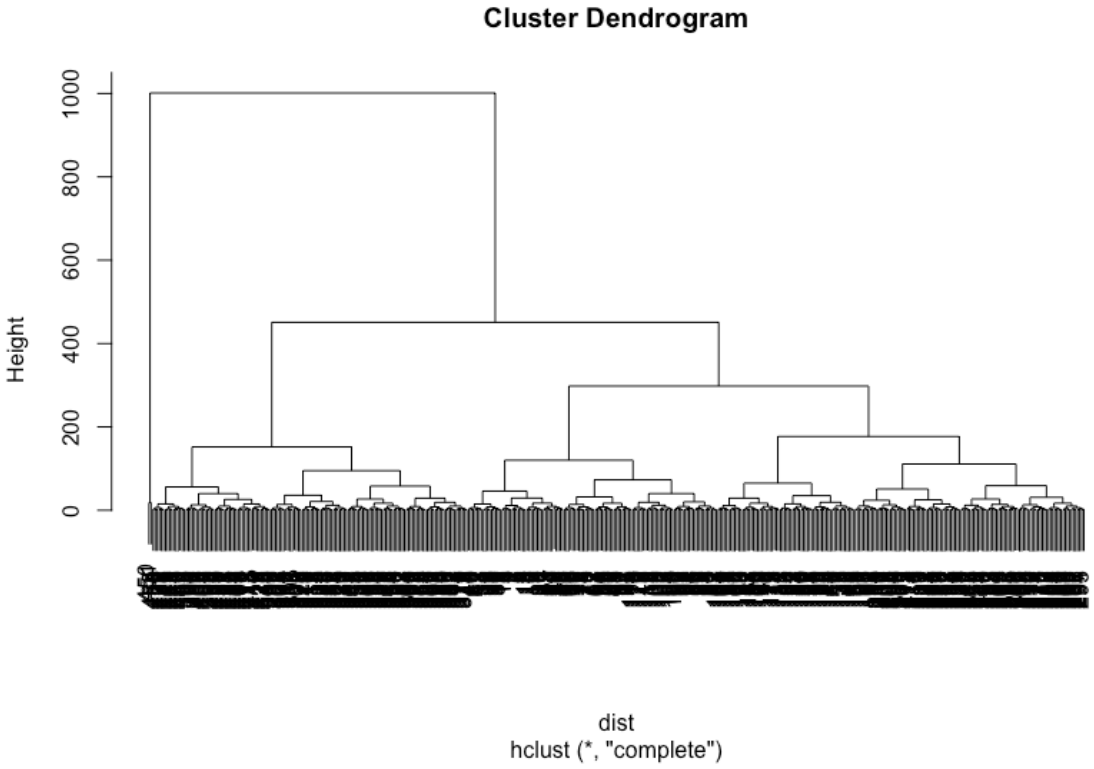
3. Por cada estación de Ecobici, identifica cómo están correlacionadas las entradas-salidas entre las otras estaciones (Hint: Puedes usar un heatmap para mostrar la correlación o matrices de origen-destino).



En esta gráfica se aprecia que la numeración de las estaciones debe tener relación con la proximidad de las mismas. Esto debido a que las que están más cera tienen una frecuencia de viajes entre sí más alta (colores más oscuros). También, se puede observar que un buen número de gente deja la bicicleta en la misa estación donde la tomó (representada por la diagonal).

4. Usa un método de aprendizaje no supervisado para encontrar “perfiles de uso” de las estaciones. Lo que debes de hacer es categorizar a las estaciones en diferentes grupos a partir de su comportamiento de entradas y salidas. Explica qué método usaste y por qué. De los grupos que encontraste describe las características que puedes inferir de estos a partir de lo descubierto en el inciso anterior.

   Se utilizó un método de conglomerados jerárquico. Esto es porque no sabemos a priori el posible número de clusters. Al analizar en dendrograma, podemos notar que los clusters donde la altura es mayor entre clusters es 6. Es decir, ahí la distancia promedio entre los clusters es mayor a que si escogiéramos 5 o 7.



Las variables elegidas para el algoritmo fueron: número de viajes, duración promedio del viaje, edad promedio del usuario, y porcentaje de viajes hechos en fin de semana

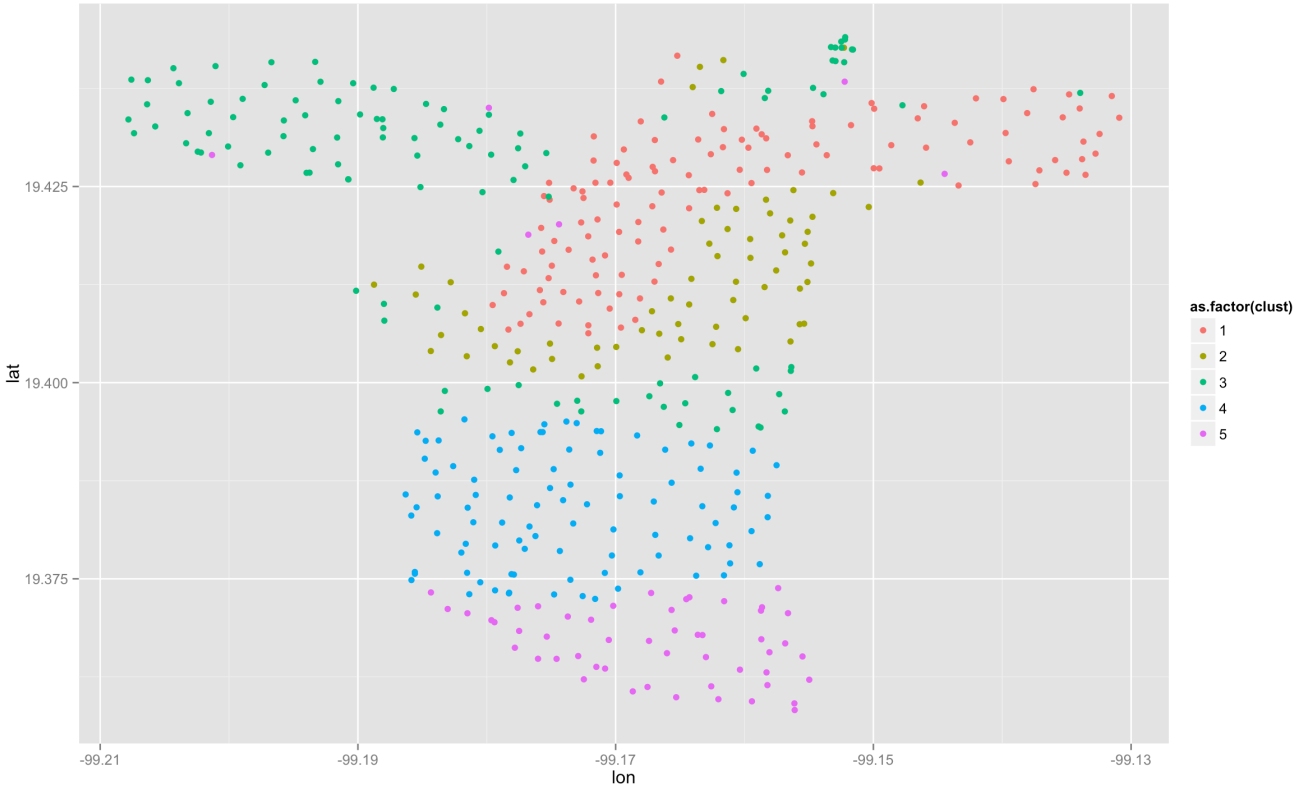
Estas fueron las medias estandarizadas por cluster para cada variable

cluster		viajes	duracion.prom	edad.prom	fines.sh	Estaciones
1	1	0.79893312	-0.1225678	0.6117275	-0.08002107	120
2	2	0.3013218	-0.2709614	-0.15192513	-0.22387284	66
3	3	-0.01163374	0.2568497	-0.37964092	0.42245305	111
4	4	-0.64360346	-0.1389099	0.14690581	-0.01928906	95
5	5	-0.88469813	-0.1614586	-0.03947304	-0.24025357	57
6	6	-1.44887302	13.2404277	-2.51704198	-3.49361917	2

Lo primero que noto es que el cluster más pequeño (6) está compuesto por la gente que hace viajes más largos, sin embargo, al ser tan pocos, no sabemos qué tan confiable es el resultado. Este cluster permanece aún si redujera el número de clusters. El cluster uno se caracteriza por tener muchos viajes. Así como el 4 por tener pocos. El cluster 3 se caracteriza por tener los usuarios más jóvenes y en fines de semana.

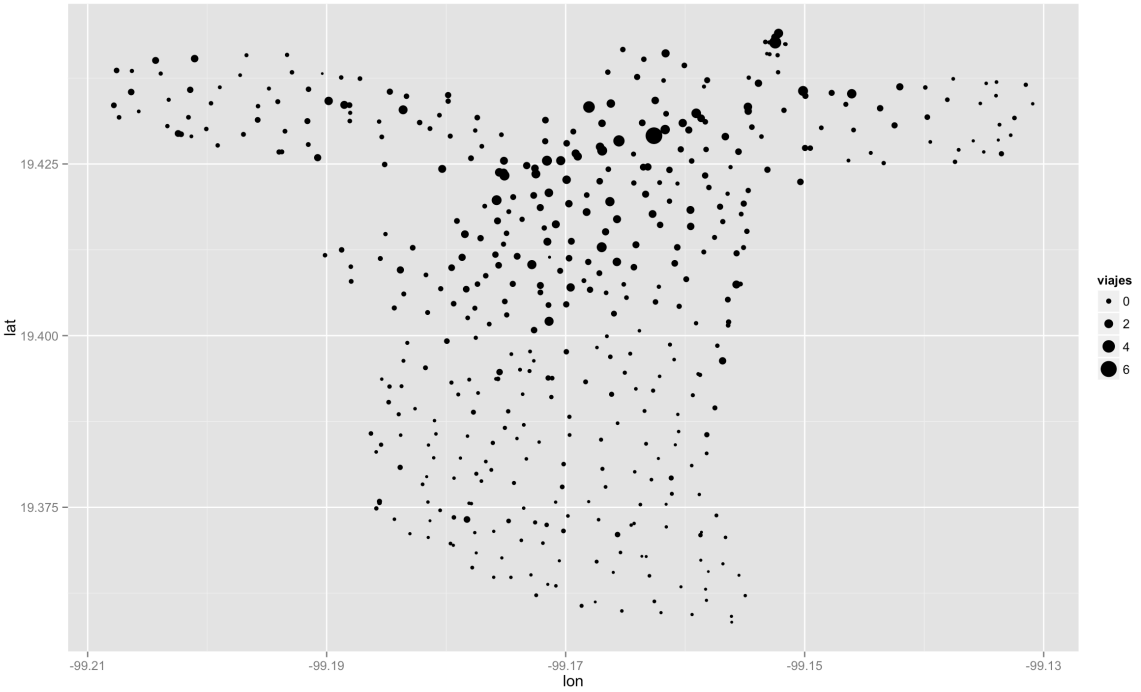
5. BONUS: En el sitio de Ecobici te puedes registrar para obtener URLs que regresan información sobre cada estación (Número de Slots, Latitud, Longitud). Usa la información de algunas estaciones para explicar el comportamiento de la relaciones que encontraste en la pregunta 3. Explica cómo los atributos geográficos te pueden ayudar a entender las relaciones.(O puedes bajar un Json de aqu í)

En esta gráfica se pueden ver los puntos de cicloestaciones en coordenadas geográficas. El color representa los clusters del algoritmo. Esto me da confianza en el método de clasificación, ya que nos permiten observar que las cicloestaciones cercanas comparten características. Esto puede deberse a que es el mismo perfil de gente que las usa.

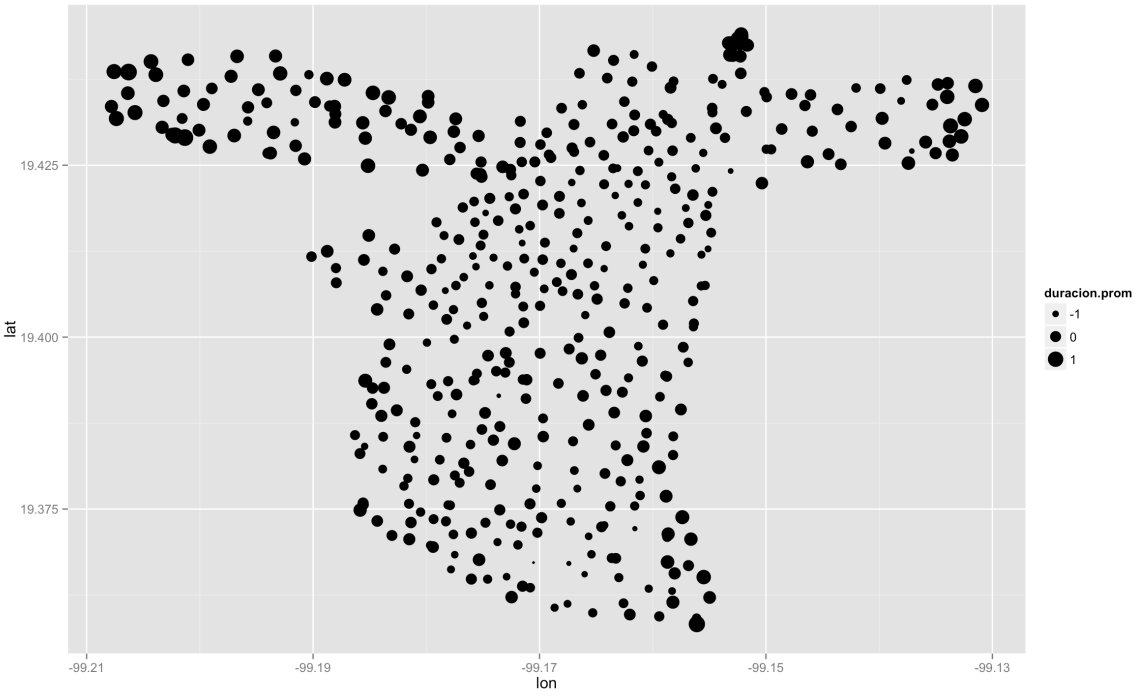


Aquí podemos ver que las cicloestaciones con más viajes se encuentran sobre todo en el centro/norte de la ciudad, que es donde están las estaciones superiores.

Tamaño proporcional al número de viajes



Tamaño proporcional a la duración del viaje



Por útlimo, vemos que las areas más lejanas al centro/sur de la ciudad es donde se hacen los viajes más largos. Esto puede deberse a la falta de infraestructura de transporte público además de las cicloestaciones, que hace que la gente tenga que recorrer grandes tramos.

Código:

```
####################pregunta 2a####################
rm(list=(ls()))
wd<-"~/Documents/WD"
#tmp<-list(tempfile(), tempfile(), tempfile())
setwd(wd)
#lapply(c(1:length(tmp))+3, function(x)
download.file(paste0("https://www.ecobici.cdmx.gob.mx/sites/default/files/data/usages/2017-0", x, ".csv"), tmp[[x-3]], mode="wb"))
#data<-lapply(tmp, read.csv)

#saveRDS(do.call(rbind, data), "ecobici")
#data<-do.call(rbind, data)
data<-readRDS("ecobici")
head(data)
data[, grep("Fecha", names(data))]<-apply(data[, grep("Fecha", names(data))], 2, function(x) as.POSIXlt(as.character(x), format="%d/%m/%Y") )

data[, grep("Hora", names(data))]<-apply(data[, grep("Hora", names(data))], 2, function(x) as.POSIXlt(as.character(x), format="%H:%M:%S") )
data<-data[format(data$Fecha_Arribo, '%m') %in% c("04", "05", "06"),]
data<-data[format(data$Fecha_Retiro, '%m') %in% c("04", "05", "06"),]
##########1. en que horarios y estaciones hay mas afluencia##########
horas<-aggregate(rep(1, nrow(data)), by=list(format(data$Fecha_Retiro, '%m'), format(data$Hora_Retiro, '%H')), sum)
names(horas)<-c("mes", "hora", "freq")


library(ggplot2)
ggplot(data=horas, aes(x=hora, y=freq))+
  geom_line(data=horas, aes(x=hora, y=freq, group=mes, color=mes))+
  ggtitle("Numero de viajes en Ecobici por hora y mes")+
  xlab("Hora de Retiro de Ecobici")+
  ylab("Numero de viajes")
dev.copy(png, "horas.png")
dev.off()
estaciones<-aggregate(rep(1, nrow(data)), by=list(data$Ciclo_Estacion_Retiro), sum)
names(estaciones)<-c("Estacion", "freq")
estaciones<-estaciones[order(estaciones$freq, decreasing = T),]
estaciones

######analisis temporal##################

dias<-aggregate(rep(1, nrow(data)), by=list(format(data$Fecha_Retiro, "%m-%d"), data$Ciclo_Estacion_Retiro), sum)
```

```
names(dias)<-c("dia", "estacion", "viajes")
dias$dia_num<-as.numeric(factor(dias$dia, unique(dias$dia)))

fines<-seq(12, 91, 7)
fines<-c(fines, fines+1, 5, 6)
dias$fin<-1
dias$fin[dias$dia_num %in% fines]<-2
dias$fin<-factor(dias$fin, levels=c("no fin", "fin"))

#prueba<-lm(viajes~dia_num, data=dias)$coefficients["dia_num"]
cor<-unlist(lapply(split(dias, dias$estacion), function(x)
tryCatch(lm(viajes~dia_num+fin, data=x)$coefficients["dia_num"], error=function(e)
lm(viajes~dia_num, data=x)$coefficients["dia_num"])))
a.la.alza<-cor[cor>quantile(cor, 0.75, na.rm = T)]
foo<-as.numeric(gsub(".dia_num", "", names(a.la.alza)))
dias$alza<-"75% menos a la alza"
dias$alza[foo]<-"25% mas a la alza"

ggplot(data=dias, aes(x=dia, y=viajes))+
  geom_line(data=dias[dias$alza=="75% menos a la alza",], aes(x=dia, y=viajes,
group=estacion, color=alza), alpha=1/10)+
  geom_line(data=dias[dias$alza=="25% mas a la alza",], aes(x=dia, y=viajes,
group=alza, color=alza), alpha=1)+
  ggtitle("Numero de viajes en Ecobici por estacion y dia")+
  xlab("Dia de Retiro de Ecobici")+
  ylab("Numero de viajes")+
  theme(axis.text.x =
element_text(colour="grey20",size=5,angle=90,hjust=.5,vjust=.5,face="plain"),
        axis.text.y =
element_text(colour="grey20",size=5,angle=0,hjust=1,vjust=0,face="plain"),
        axis.title.x =
element_text(colour="grey20",size=5,angle=0,hjust=.5,vjust=0,face="plain"),
        axis.title.y =
element_text(colour="grey20",size=5,angle=90,hjust=.5,vjust=.5,face="plain"))
dev.copy(png, "a.la.alza.png")
dev.off()

########relaciones entre estaciones
library(reshape2)
est<-aggregate(rep(1, nrow(data)), by=list(data$Ciclo_Estacion_Arribo,
data$Ciclo_Estacion_Retiro), sum)
names(est)<-c("arribo", "retiro", "freq")
summary(est)
est<-est[est$arribo<999,]
est<-est[est$retiro<999,]
est$freq<-est$freq
ggplot(data = est, aes(x=arribo, y=retiro, fill=freq)) +
```

```
  geom_tile()+ggtitle("Numero de viajes en Ecobici por par de estaciones.
Desviaciones estandar")
dev.copy(png, "heat.png")
dev.off()

#######clusters#######
for.clust<-aggregate(rep(1, nrow(data)), by=list(data$Ciclo_Estacion_Retiro), sum)
data$dia_num<-as.numeric(factor(format(data$Fecha_Retiro, "%m-%d"),
unique(format(data$Fecha_Retiro, "%m-%d"))))
fines<-seq(12, 91, 7)
fines<-c(fines, fines+1, 5, 6)
data$fin<-0
data$fin[data$dia_num %in% fines]<-1
summary(data$fin)
for.clust.fines<-aggregate(data$fin, by=list(data$Ciclo_Estacion_Retiro), mean)
data$dura<-as.numeric(data$Hora_Arribo-data$Hora_Retiro)
for.clust.dura<-aggregate(data$dura, by=list(data$Ciclo_Estacion_Retiro), mean)
data$hora_ret_num<-as.numeric((format(data$Hora_Retiro, '%H')))
data$hora_ret_num[data$hora_ret_num==0]<-24
for.clust.hora<-aggregate(data$hora_ret_num,
by=list(data$Ciclo_Estacion_Retiro), mean)
for.clust.edad<-aggregate(data$Edad_Usuario,
by=list(data$Ciclo_Estacion_Retiro), mean)

for.clust<-merge(for.clust, for.clust.dura, by="Group.1")
for.clust<-merge(for.clust, for.clust.edad, by="Group.1")
for.clust<-merge(for.clust, for.clust.fines, by="Group.1")
names(for.clust)<-c("stations.id", "viajes", "duracion.prom", "edad.prom", "fines.sh")
#estandarizar
for.clust[,-1]<-apply(for.clust[,-1], 2, scale)

dist<-dist(for.clust)
clust<-hclust(dist, method="complete" )
plot(clust)
clust <- cutree(clust, 6)
clusters<-cbind(clust, for.clust)
table(clusters$clust)
sum<-aggregate(cbind(clusters$viajes, clusters$duracion.prom,
clusters$edad.prom, clusters$fines.sh), by=list(clusters$clust), mean)
names(sum)<-c("cluster","viajes", "duracion.prom", "edad.prom", "fines.sh")

#bonus
library(jsonlite)

data.estaciones<-as.data.frame(fromJSON("estaciones.json"))
loc<-cbind(data.estaciones$stations.location, data.estaciones$stations.id)
names(loc)<-c("lat", "lon", "stations.id")
```

```
clusters<-merge(data.estaciones, clusters, by="stations.id")
clusters<-cbind(clusters$stations.location, clusters)

ggplot(data=clusters, aes(x=lon, y=lat))+
  geom_point(data=clusters, aes(x=lon, y=lat, size=viajes))
dev.copy(png, "mapa_clust.png")
dev.off()

ggplot(data=clusters, aes(x=lon, y=lat))+
  geom_point(data=clusters, aes(x=lon, y=lat, size=duracion.prom))
dev.copy(png, "mapa_dura.png")


ggplot(data=clusters, aes(x=lon, y=lat))+
  geom_point(data=clusters, aes(x=lon, y=lat, size=viajes))
dev.copy(png, "mapa_viajes.png")
```