**TAREA # 1: ANÁLISIS DE COMPONENETES PRINCIPALES (PCA)**

**DATOS:** Los datos originales son tomados de la base de datos del gobierno nacional de Colombia (https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data) la cual está actualizada al 19 de marzo del año 2021. En ella se encuentra los registros de todos los casos reportados de covid-19 por departamento, municipio, sexo y edad. En total son 2,294,617 registros y un total de 23 variables.

**Codigo:**

Lectura de datos:

datos<-read.csv("Casos\_positivos\_de\_COVID-19\_en\_Colombia.csv",header=T)

**LIMPIEZA DE DATOS:** Para facilidades de interpretación, solo se tomaron los registros correspondientes al año 2021, se eliminaron la gran mayoría de variables dejando solo el nombre del departamento, nombre del municipio, edad, sexo y recuperado. Por último se eliminaron aquellas columnas con vacios o NA y se creó un archivo csv con los datos filtrados.

**Codigo:**

Cargar datos solo del año 2021:

datos <- datos[1642775:2294657,]

Eliminación de variables:

eliminar <- c(1:4, 6, 9, 11:15, 17:23)

datos <- datos[, -eliminar]

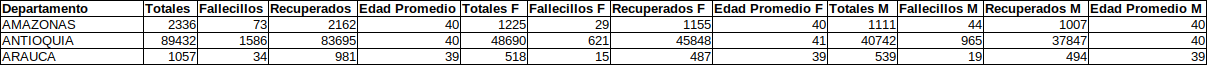
Suprimir registros con NA’s:

datos <- na.omit(datos)

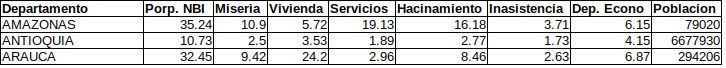
Creación de archivo csv:

write.csv(df, "covid\_colombia\_2021")

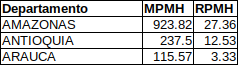
Luego en excel se procede a crear una tabla dinámica filtrando las variables de interés por departamento, teniendo como resultado la siguiente tabla:



Se agregan nuevas variables tomadas de la pagina del DANE (https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018), tomando indices que miden la pobreza por departamentos, agregando las variables Prop. NBI (Proporcion de la población con las Necesidades Basicas Insatisfechas), Miseria (proporcion de personas que viven en la miseria), Vivienda (proporcion de personas que no tienen una vivienda digna), Servicios (proporcion de la poblacion sin servicios públicos), Hacinamiento (proporcion de personas que viven en hacinamiento), Inasistencia (proporcion de la población que no tienen una asistencia básica), Dep. Econo (proporcion de personas que no poseen una asistencia económica), población total de cada departamento



Por último se calculan dos indicadores, Muertos por millon de habitantes (MPMH) y Recuperados por millon de habitantes (RPMH).



Se guarta todo lo anterior en un archivo csv con el nombre de **departamentos.csv.**

**LECTURA DEL NUEVO DATA FRAME:** S e procede a la lectura de los nuevos datos, eliminamos la primera columna de los departamentos y asignamos los nombres a las filas y por último se determina como variable numérica la variable MPMH, ya que no la cargó como tal.

**Codigo:**

deptos <- read.csv('departamentos.csv')

df<- deptos[,-1] # Eliminamos la primera columna

dimnames(df)[[1]] <- deptos[,1] # Asignar los nombres a las filas

df$MPMH <- as.numeric(df$MPMH) # Determinar variable como numérica

Procedemos a hacer un resumen estadistico e identificamos variables con valores NA.

**Codigo:**

summary(df)

Eliminamos registros con valores nulos, calculamos correlaciones y graficamos.

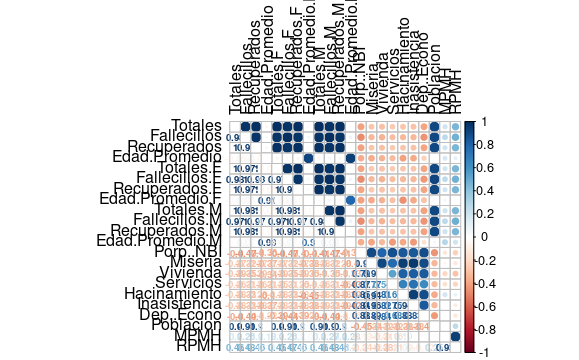
**Codigo:**

df <- na.omit(df)

correlaciones<-cor(df)

corrplot.mixed(correlaciones, tl.pos = 'lt', tl.col = 'black', number.cex = 0.6)

**CORRELACIONES:**

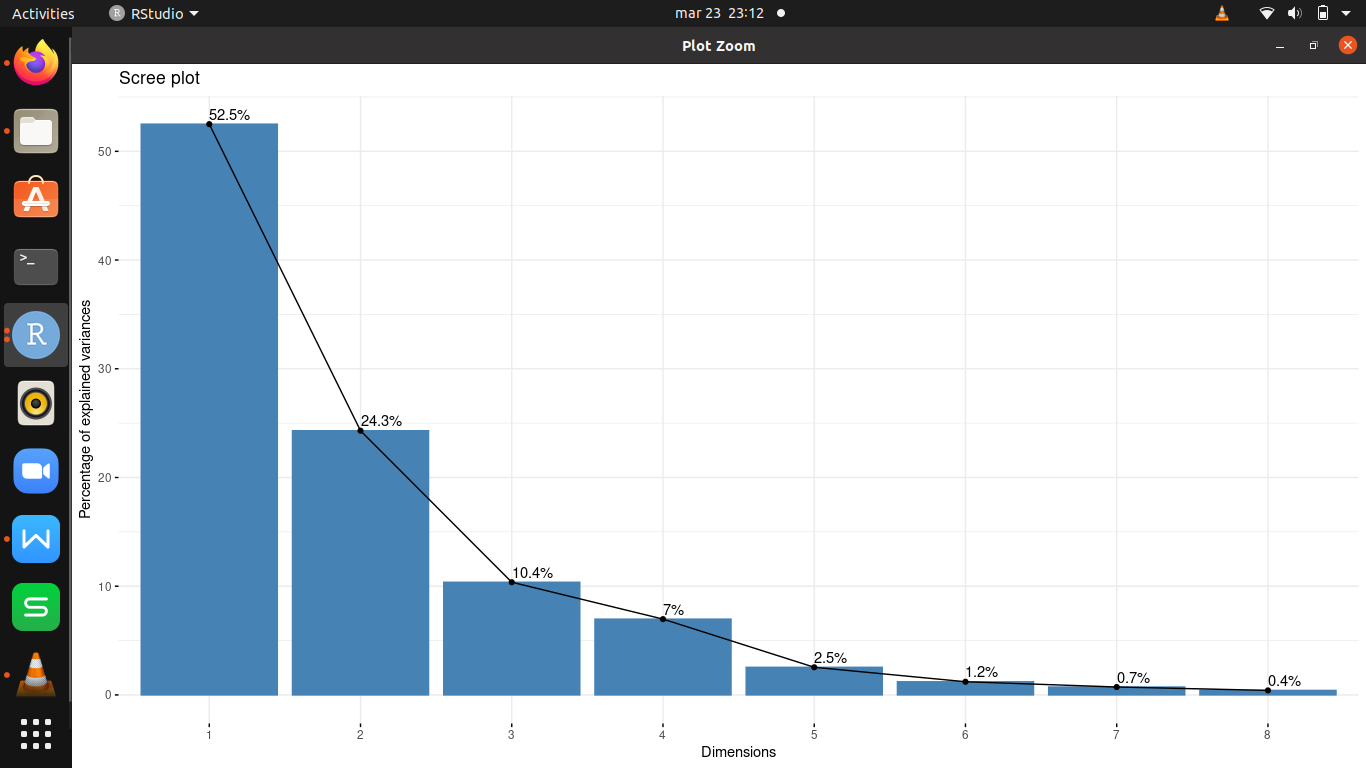


Se observa una alta correlación entre las variables, lo cual es un buen indicador de realizar Análisis de Componentes Principales (PCA) y para las predicciones no es adecuado utilizar una regresión múltiple, ya que se estaría violando el supuesto de no autocorrelación entre las variables.

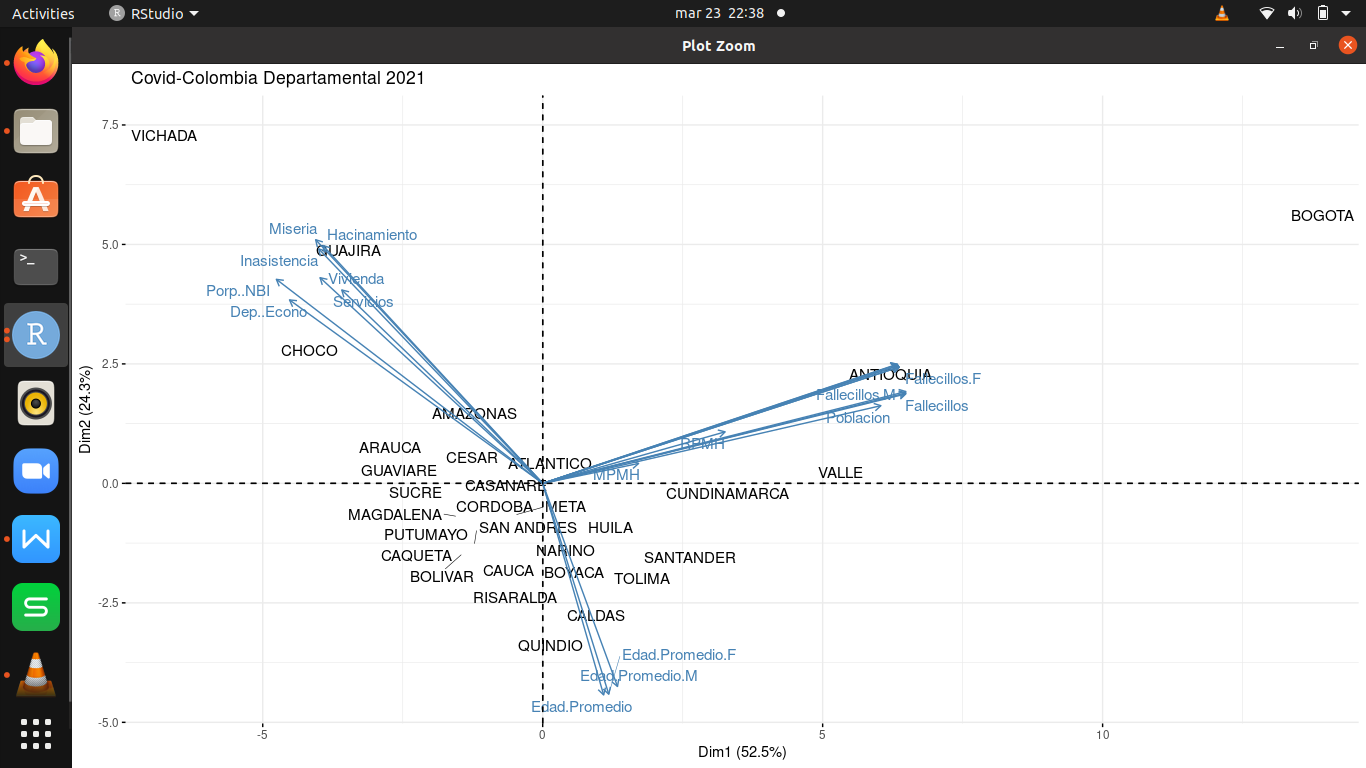
**ANÁLISIS DE COMPONENETES PRINCIPALES (PCA)**

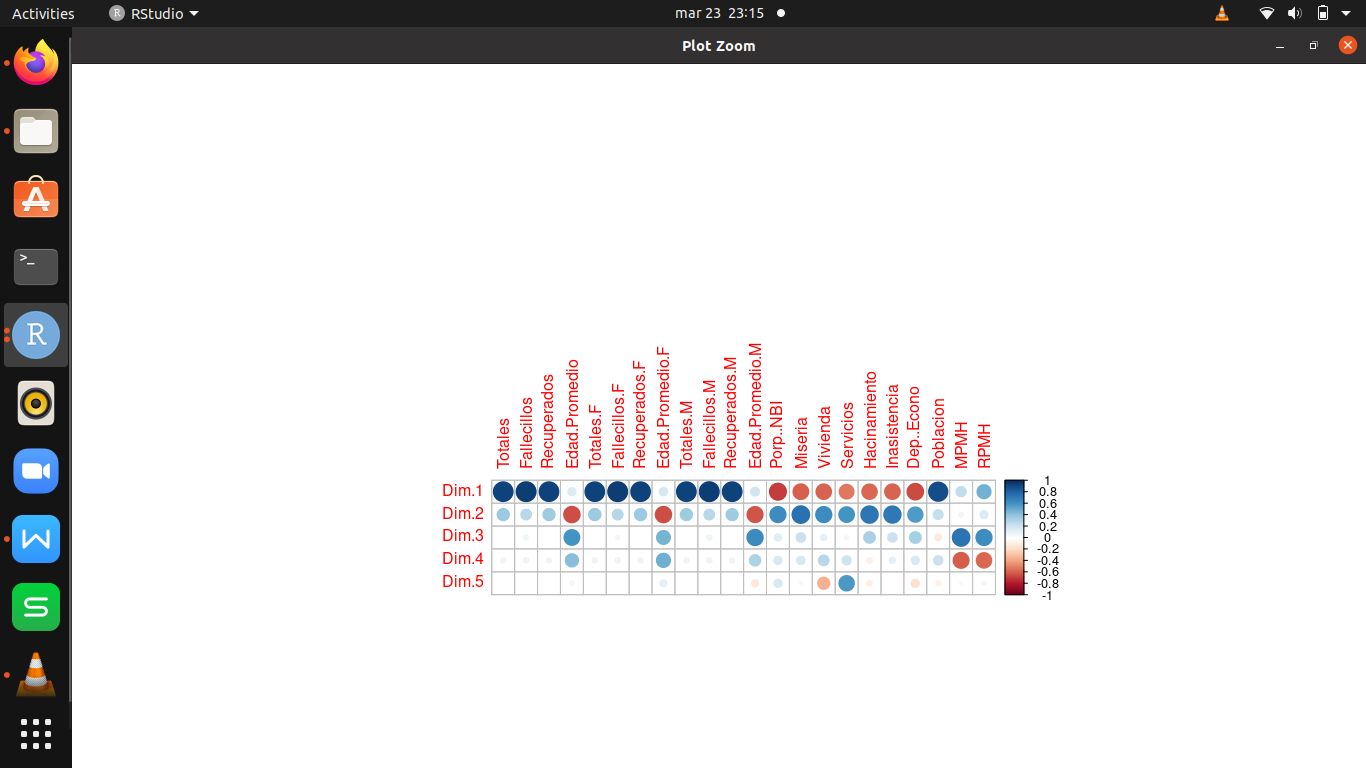
Para el Análisis de Componentes Principales, inicialmente se estandarizaron las variables, ya que tienen escalas muy diferentes y esto afectaría el adecuado análisis de los vectores, ya que al haber variables con valores nominales mas altos, los vectores de estos mismos serían mas relevantes, lo cual daría como resultado una interpretanción incorrecta de los resultados.

**SCREE PLOT**

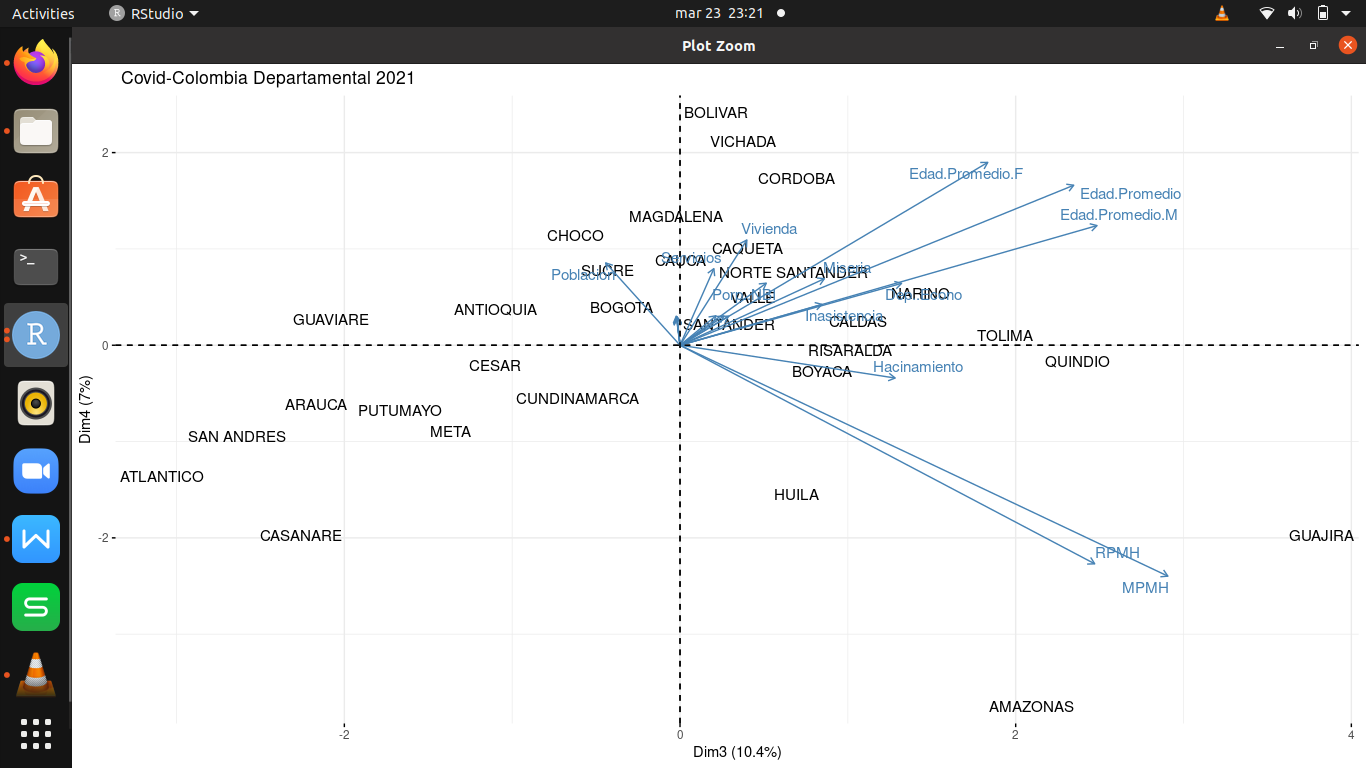
****

**BIPLOT**

****

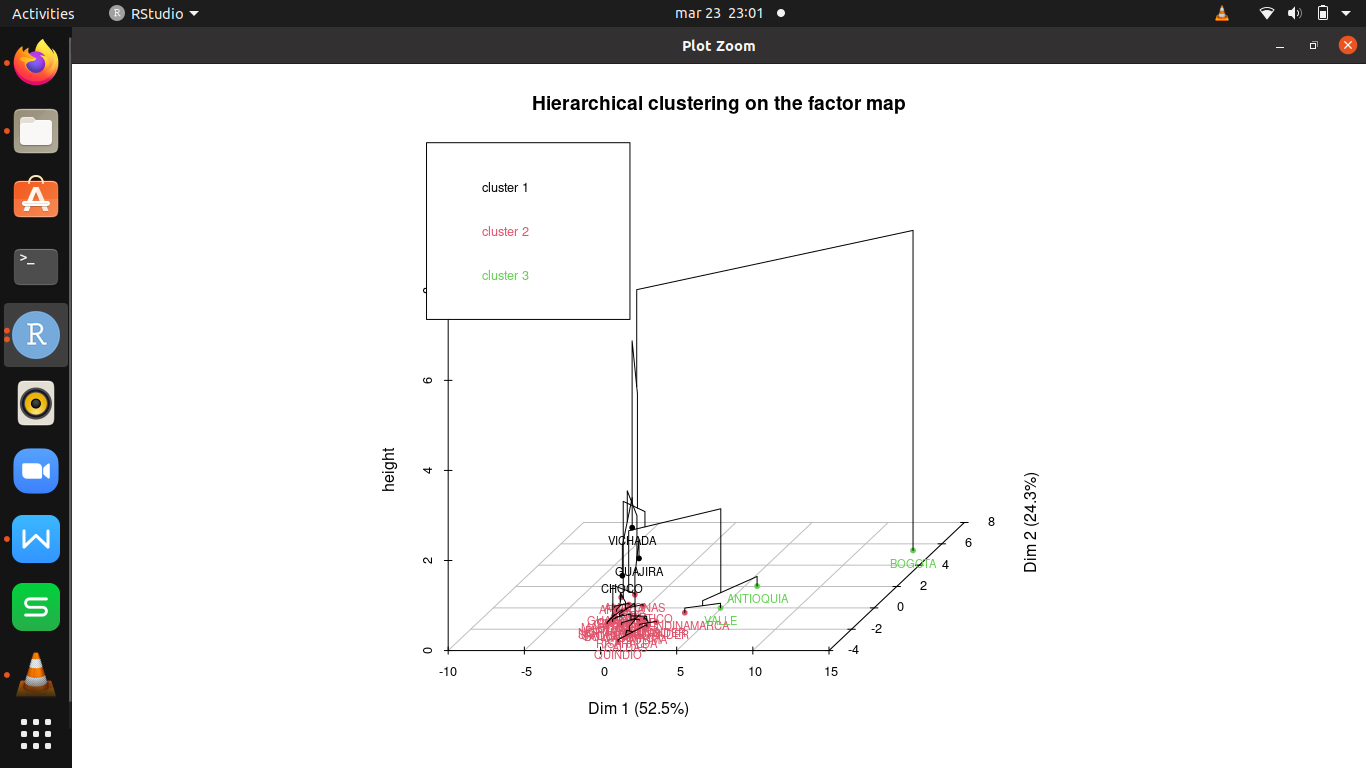
****

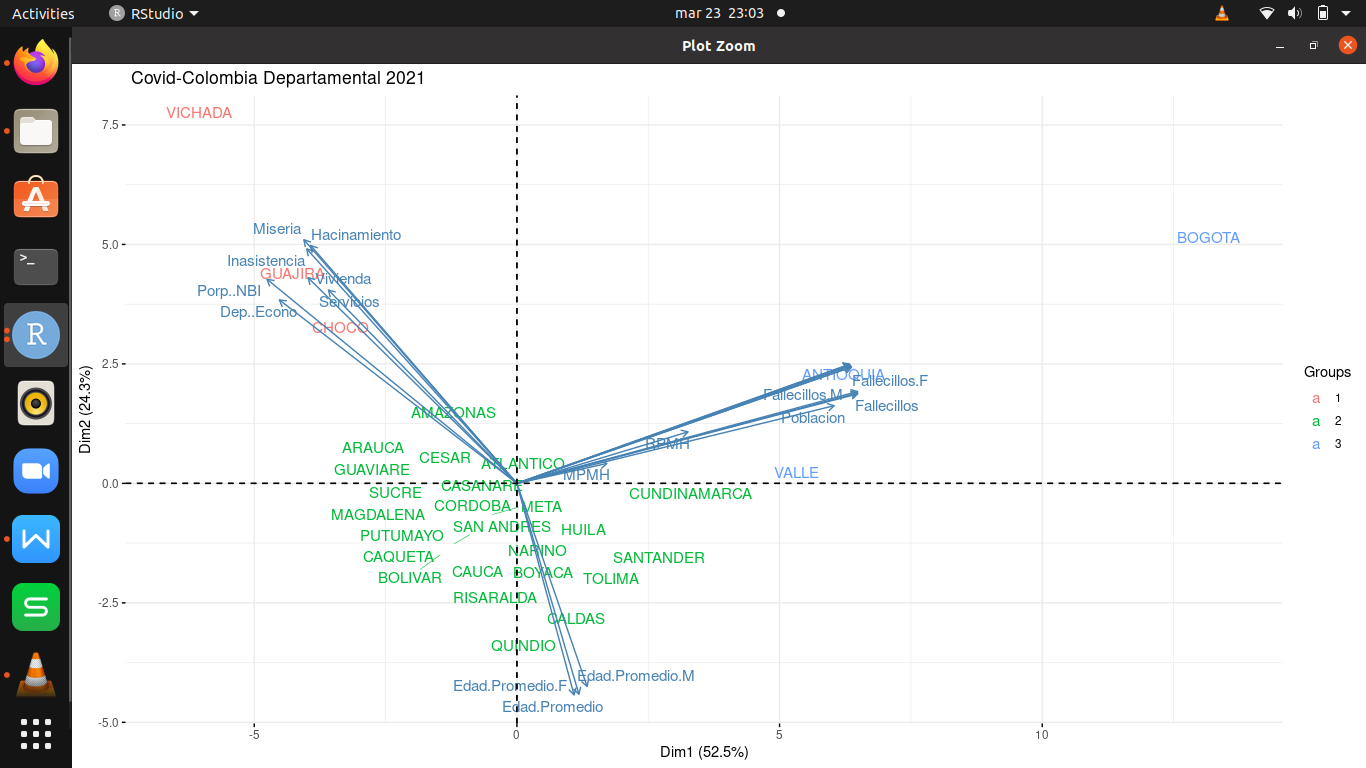
Se puede observar que la 1 y 2 dimensión representan la variabilidad del modelo en un 76,8%, lo cual es un gran porcentaje y muy cercano a la meta del 80%. Las variables muertos y recuperados por millon de habitantes (RPMH y MPMH), cantidad de fallecidos y poblacion tienen una gran fuerza en los departamentos de Antioquia, Valle del Cauca y Bogotá, lo cual es de esperarse ya que son los municipios con mayor densidad poblacional, contrario a poblaciones como Putumayo, Caqueta y San Andrés que tienen un número muy reducido de habitantes. La variable edad es muy significativa en departamentos como Risaralda, Caldas y Quindio, es decir que estos municipios la edad promedio poblacional es mayos a la de otros departamentos y Amazonas, Chocó y La Guajira poseen poblaciones relativamente mas jóvenes en promedio. Los indicadores de pobreza son bastante relevantes en departamentos como Vichada, Guajira, y Choco, lo cual demuestra que son aquellas zonas donde hay una mayor vulnerabilidad en sus habitantes y poseen unos indices de pobreza y con necesidades básicas como vivienda y servicios públicos insatisfechas y altos indices de hacinamiento y miseria. También se puede observar que los casos tanto de fallecidos, como de personas recuperadas no tienen relacion alguna con la edad promedio y los indices de pobresa de su población, además de una relación inversa entre los niveles de pobreza y la edad de sus habitantes. También se observa una posible agrupación (clusters) muy definidas en tres zonas, para verificar lo anterior se procede a realizar un análisis de cluster. Por último se puede observar que la gran mayoría de las variables estan muy bien representadas por las dimensiones 1 y 2, mientras que la dimension 3 representa de una mejor manera las variables recuperados y muertos por millon.



Como se mensionó con anterioridad las variables RPMH y MPMH son mejore representadas en la 3 y 4 dimensión y nos dice que en regiones como Amazonas y Guajira los niveles de fallecidos por cantidad de habitantes son mayores a la de los demás departamentos.

**CLUSTER**





Efectivamente se ve claramente una segmentación muy definida en los departamentos principales (Antioquia, Valle y Bogotá), los departamentos con población mas vulnerable (Vichada, Guajira y Chocó) y aquellos departamentos intermedios.