Trabajo Práctico N° 6
Clustering

Presentado en la fecha: 23/11/2019

Hecho por: Andrada Alexander

Encina Guadalupe

Huarca Brian

Contents

Resumen

El presente informe detalla un proceso de análisis del desarrollo de la actividad ganadera en las diferentes provincias de Argentina, mediante el uso de la función de clustering en R, para hacer comparaciones con datos oficiales publicados en la web del gobierno nacional, sobre mediciones de la emisión de Gases de Efecto Invernadero por el sector agroganadero para cada provincia.

Palabras clave: *Ganadería, Ambiente, Argentina, Gases de Efecto Invernadero, Clustering, Minería de Datos.*

Introducción

En este trabajo analizamos, a través del uso de **clustering**, el desarrollo de la actividad ganadera en el país, para comparar los resultados con la información publicada en el informe del Inventario Nacional de Gases de Efecto invernadero.

De acuerdo al informe publicado en 2017 (donde se incluyen datos recolectados hasta el año 2014), la ganadería es la principal fuente de emisión de Gases de Efecto Invernadero (GEI). Esto puede observarse en la **Figura 1**, que detalla los porcentajes de emisión por sector de actividad humana, y en la **Figura 2**, que detalla las emisiones separadas por organismo de aplicación.

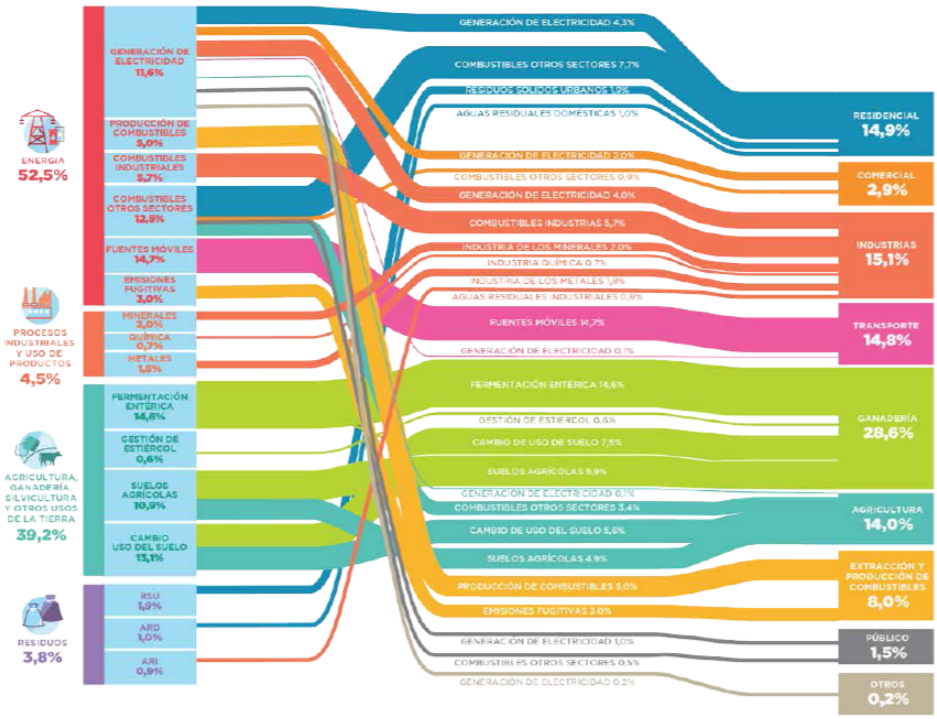

Figure 1: Emisiones de GEI por sector de actividad.

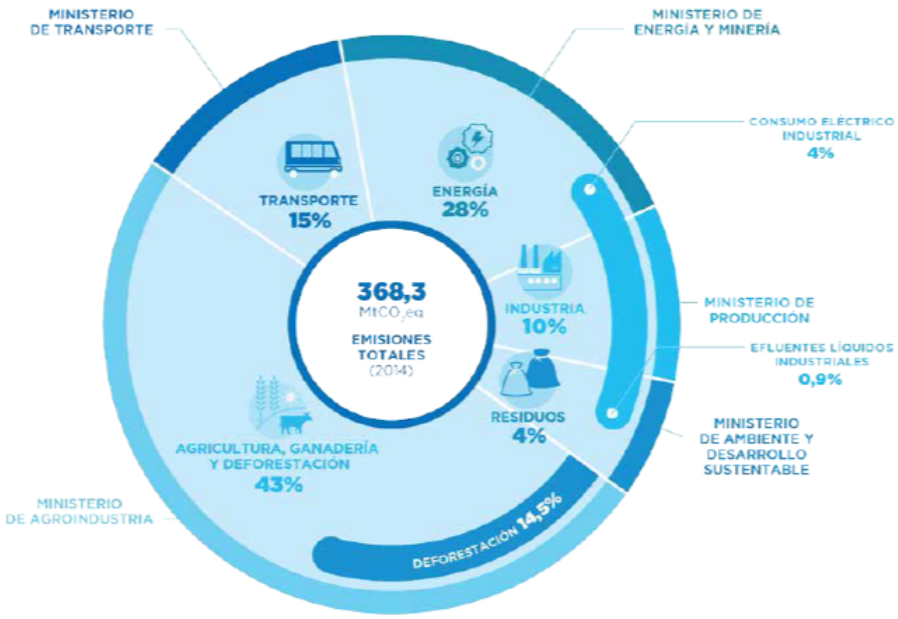

Figure 2: Emisiones de GEI por organismo.

Esta información corresponde sólo al territorio argentino, que es lo que trabajamos en este informe. Pero, según publicaciones de la FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura), la ganadería es la mayor fuente de emisión de gases a la atmósfera a nivel mundial, incluso superando al sistema de transporte y a la industria del petróleo.

Además, en la **Figura 3** se detallan los valores de emisión de GEI del sector ganadero, separados por diferentes características.

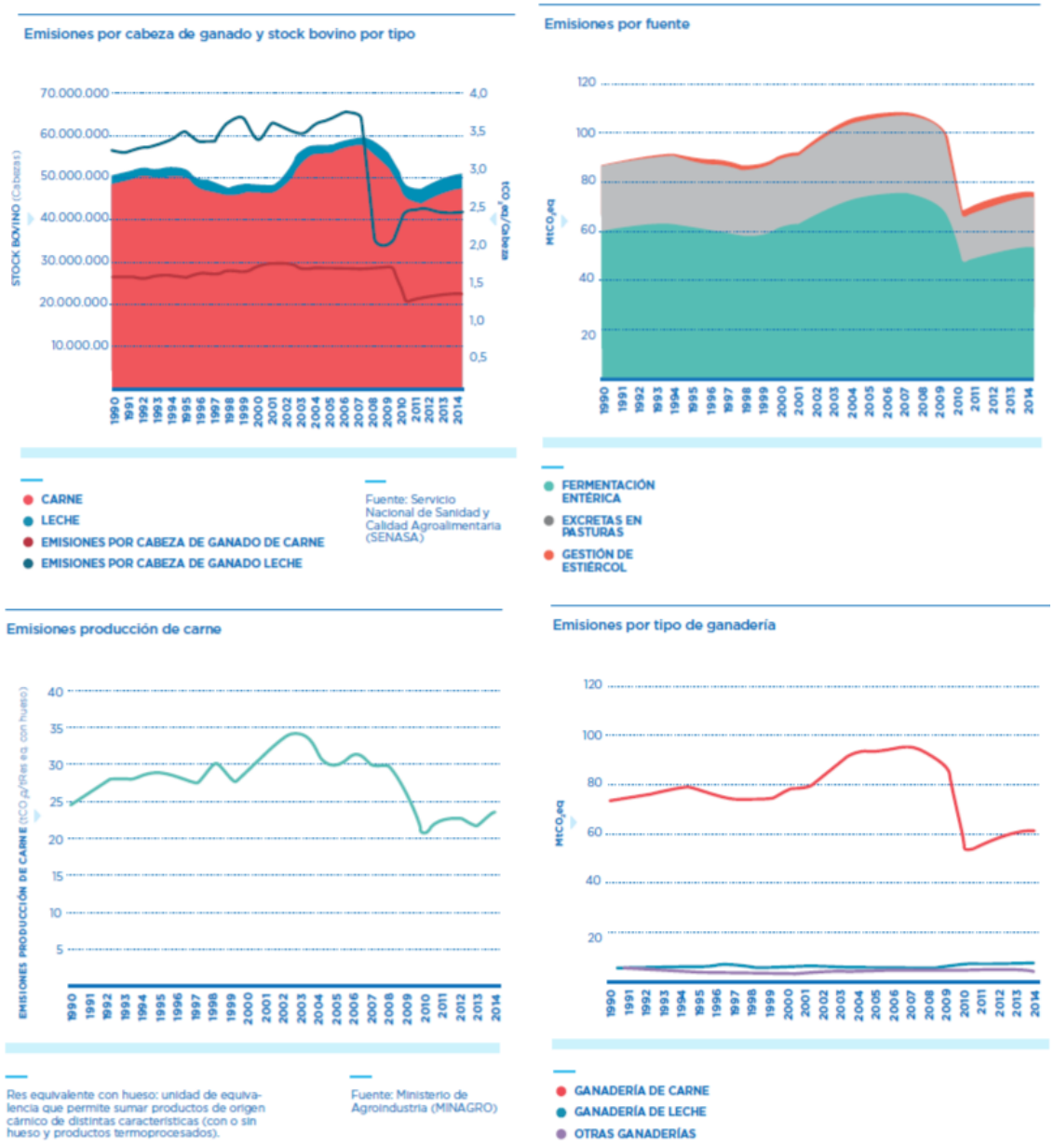

Figure 3: Fuente: Inventario Nacional de GEI

A partir de esto, analizaremos la distribución del ganado en nuestro país, para observar cómo se desarrolla la actividad en las provincias y cuáles son las que presentan mayor nivel de concentración de ganado, comparando estos valores con los niveles de emisión de cada provincia que se pueden ver en el sitio del Inventario GEI.

Mediante el clustering, utilizando la herramienta R, intentaremos observar qué provincias tienen similitudes en la actividad ganadera, a partir de un dataset obtenido del sitio del Ministerio de Agricultura, Ganadería y Pesca.

Objetivo

El objetivo de este trabajo es lograr observar similitudes en la actividad ganadera en las diferentes provincias de la Argentina, y poder establecer una relación de esta con los niveles de emision de GEI del país.

Normalizacion del dataset

1.1 Limpieza y Transformacion de Datos

Los datos que utilizamos para el análisis fueron extraídos de la página del Ministerio de Agricultura, Ganadería y Pesca. El dataset elegido corresponde a las mediciones de stock bovino por especie, para cada provincia argentina. En la figura que se muestra a continuación puede visualizarse el set de datos:

	Provincia	Vacas	Vaquillonas	Novillos	Novillitos	Terneros	Terneras	Toros	Toritos	Bueyes
1	BUENOS AIRES	8375806	2321588	607346	1109019	3050149	3231435	343212	88349	1758
2	CATAMARCA	112173	44422	10827	23447	38503	26099	8102	619	455
3	CHACO	1238340	335503	97725	171788	303250	314897	65307	23499	281
4	CHUBUT	99392	28717	8268	11483	34952	37729	6035	860	459
5	CÓRDOBA	1916491	779266	250556	478549	635382	654093	77342	20872	111
6	CORRIENTES	2125695	723397	223386	290444	458840	525898	112042	2175	8
7	ENTRE RÍOS	1831495	561222	312498	399081	528932	536947	80547	20990	3
8	FORMOSA	797878	273939	68857	147238	209810	195624	41745	6234	335
9	JUJUY	48265	22693	8699	12010	13192	10806	3871	360	25
10	LA PAMPA	1440733	382811	270030	327852	407307	403547	52164	13980	109
11	LA RIOJA	88789	29080	4923	9082	22771	22849	5347	440	32
12	MENDOZA	259512	54220	13035	19261	54506	56703	15810	1982	55
13	MISIONES	179718	66500	17658	32046	39604	42291	9903	757	4744
14	NEUQUÉN	107871	26079	3401	9688	20301	28436	5989	1139	433
15	RÍO NEGRO	346151	65819	10841	24072	88132	110555	16214	6595	87
16	SALTA	530376	256589	104837	141786	163500	148645	30600	18872	225
17	SAN JUAN	22022	6952	2363	2500	7416	6457	1662	19	0
18	SAN LUIS	772606	244136	90351	125089	179675	190806	36857	11127	103
19	SANTA CRUZ	57866	14427	3415	4329	9008	13980	3041	596	8

Figure 4: Set de datos utilizado

Con el objeto de normalizar la tabla para una mejor comprensión y desempeño de los gráficos se eliminó todo tipo de dato que no sea numérico del set de datos, sin embargo, el identificador de Provincia se asignó al numero de Observación correspondiente para cada caso.

	Vacas	Vaquillonas	Novillos	Novillitos	Terneros	Terneras	Toros	Toritos	Bueyes
BUENOS AIRES	8375806	2321588	607346	1109019	3050149	3231435	343212	88349	1758
CATAMARCA	112173	44422	10827	23447	38503	26099	8102	619	455
CHACO	1238340	335503	97725	171788	303250	314897	65307	23499	281
CHUBUT	99392	28717	8268	11483	34952	37729	6035	860	459
CÓRDOBA	1916491	779266	250556	478549	635382	654093	77342	20872	111
CORRIENTES	2125695	723397	223386	290444	458840	525898	112042	2175	8
ENTRE RÍOS	1831495	561222	312498	399081	528932	536947	80547	20990	3
FORMOSA	797878	273939	68857	147238	209810	195624	41745	6234	335
JUJUY	48265	22693	8699	12010	13192	10806	3871	360	25
LA PAMPA	1440733	382811	270030	327852	407307	403547	52164	13980	109
LA RIOJA	88789	29080	4923	9082	22771	22849	5347	440	32
MENDOZA	259512	54220	13035	19261	54506	56703	15810	1982	55
MISIONES	179718	66500	17658	32046	39604	42291	9903	757	4744
NEUQUÉN	107871	26079	3401	9688	20301	28436	5989	1139	433
RÍO NEGRO	346151	65819	10841	24072	88132	110555	16214	6595	87
SALTA	530376	256589	104837	141786	163500	148645	30600	18872	225
SAN JUAN	22022	6952	2363	2500	7416	6457	1662	19	0
SAN LUIS	772606	244136	90351	125089	179675	190806	36857	11127	103
SANTA CRUZ	57866	14427	3415	4329	9008	13980	3041	596	8

Figure 5: Set de datos Modificado

Una vez estructurada la tabla, se normalizo escalando los datos.

	Vacas	Vaquillonas	Novillos	Novillitos	Terneros	Terneras	Toros	Toritos	Bueyes
BUENOS AIRES	4.1316281	3.89872890	3.10618199	3.44364550	4.30753980	4.31194153	4.06909238	4.097041509	1.319144449
CATAMARCA	-0.5146764	-0.54254570	-0.64437454	-0.58783203	-0.43480092	-0.44853077	-0.51794151	-0.560623828	0.028467335
CHACO	0.1185215	0.02516462	-0.09801162	-0.03693959	-0.01791246	-0.01961633	0.26508876	0.654096142	-0.143887091
CHUBUT	-0.5218626	-0.57317597	-0.66046401	-0.63226261	-0.44039257	-0.43125823	-0.54623490	-0.547828919	0.032429506
CÓRDOBA	0.4998182	0.89065855	0.86289877	1.10227557	0.50508497	0.48414775	0.42982557	0.514626327	-0.312279347
CORRIENTES	0.6174451	0.78169435	0.69206998	0.40371198	0.22709004	0.29375625	0.90480416	-0.478014375	-0.414305243
ENTRE RÍOS	0.4520284	0.46539606	1.25235321	0.80715608	0.33746163	0.31016590	0.47369607	0.520891054	-0.419257957
FORMOSA	-0.1291324	-0.09490683	-0.27951643	-0.12811067	-0.16504938	-0.19675714	-0.05743129	-0.262518381	-0.090397786
JUJUY	-0.5506092	-0.58492489	-0.65775414	-0.63030550	-0.47465733	-0.47124349	-0.57585605	-0.574374373	-0.397466018
LA PAMPA	0.2323188	0.11743186	0.98533969	0.54263368	0.14594271	0.11204408	0.08518549	0.148723791	-0.314260432
LA RIOJA	-0.5278242	-0.57246800	-0.68149538	-0.64117918	-0.45957359	-0.45335757	-0.55565235	-0.570127100	-0.390532219
MENDOZA	-0.4318336	-0.52343615	-0.63049195	-0.60337753	-0.40960152	-0.40307859	-0.41243330	-0.488260921	-0.367749737
MISIONES	-0.4766986	-0.49948584	-0.60142528	-0.55589801	-0.43306721	-0.42448288	-0.49328916	-0.553297282	4.276904897
NEUQUÉN	-0.5170952	-0.57832100	-0.69106481	-0.63892868	-0.46346302	-0.44505992	-0.54686456	-0.533016556	0.006675396
RÍO NEGRO	-0.3831201	-0.50081403	-0.64428652	-0.58551097	-0.35665176	-0.32309916	-0.40690329	-0.243352563	-0.336052371
SALTA	-0.2795378	-0.12874543	-0.05329560	-0.14835770	-0.23797222	-0.26652899	-0.20998565	0.408444511	-0.199357481
SAN JUAN	-0.5653646	-0.61562538	-0.69759114	-0.66562268	-0.48375261	-0.47770250	-0.60609316	-0.592478372	-0.422229585
SAN LUIS	-0.1433418	-0.15303316	-0.14437495	-0.21036518	-0.21250198	-0.20391270	-0.12433893	-0.002744569	-0.320203689

Figure 6: Set de datos Escalado

Analisis Exploratorio

2.1 Analisis Preliminar de Datos

Mediante una primera observación, podemos ver que los valores de stock de ganado bovino en Buenos Aires son enormemente mayores a los del resto, lo cual coincide con este gráfico obtenido de la misma página de la que sacamos el set de datos:

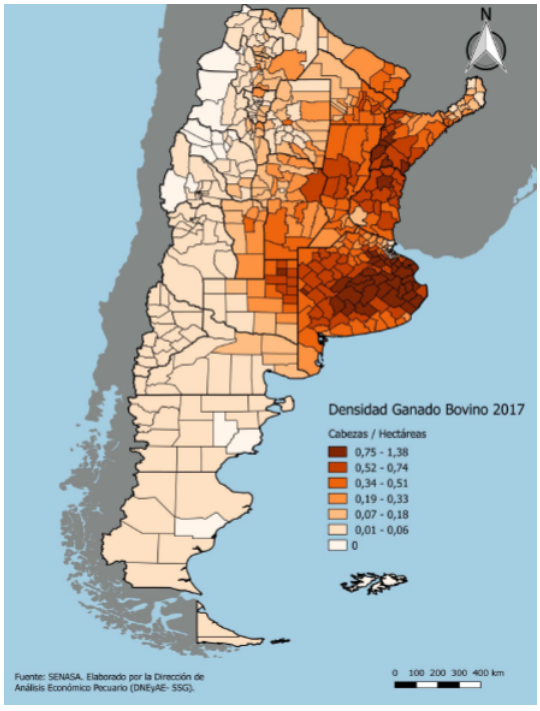

Figure 7: Mapa de densidad bovina 2017

2.1.1 Analisis de similitud en distancia de Variables

Como primera observación de Análisis, mediante una Matriz en donde vemos la distancia mas proxima entre variables. Buscamos observar la correlación existente o similitud de los datos entre las distintas regiones.

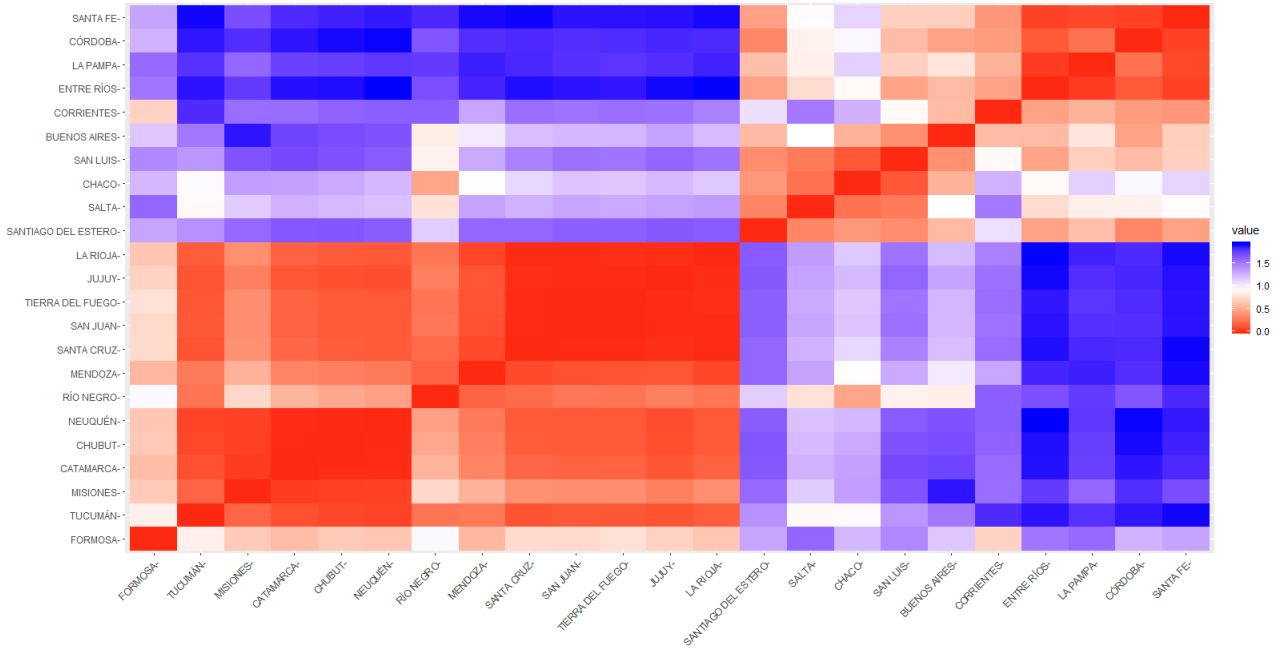

Figure 8: Matriz de distancia

2.1.2 Metodo de K-means

Kmeans es un método de agrupación de casos que se basa en la distancia existente entre ellos en un conjunto de variables. Por lo tanto, vamos a usar este mótodo para poder agrupar los casos en donde la distancia existente entre los objetos sea mínima.

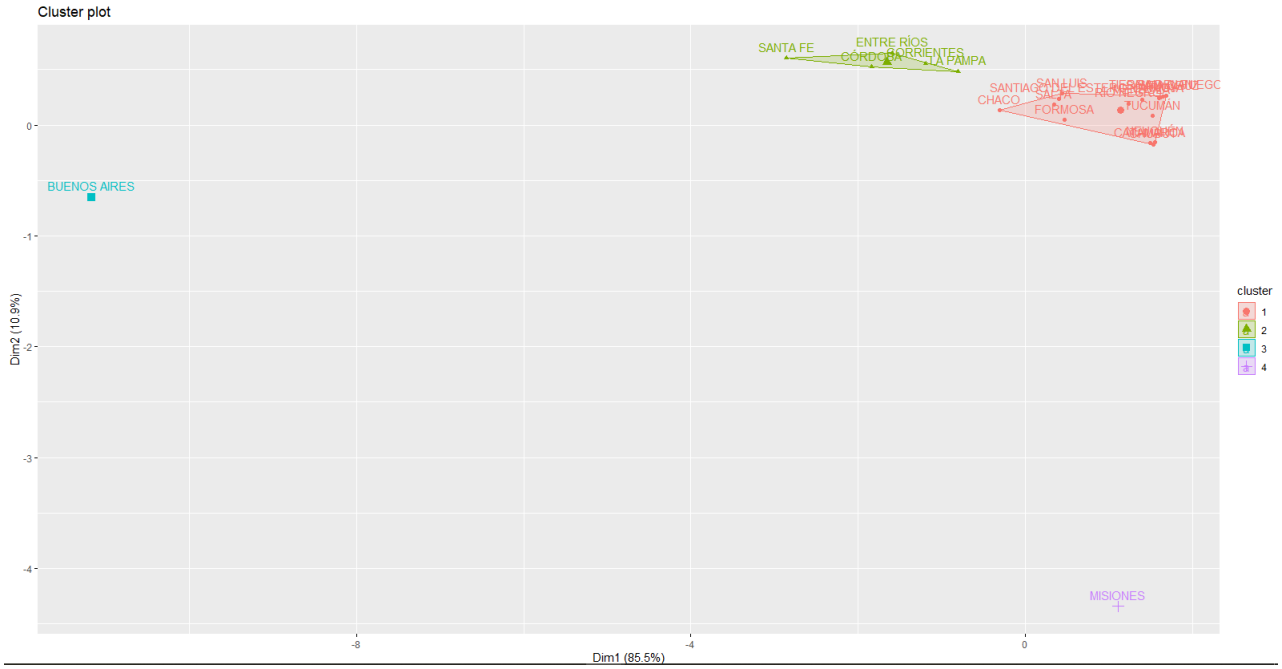

Figure 9: Gráfico K-Means

```
> scale_ganado_km
K-means clustering with 4 clusters of sizes 16, 5, 1, 1

Cluster means:
       Vacas Vaquillonas   Novillos Novillitos   Terneros   Terneras      Toros    Toritos      Bueyes
1 -0.3932502  -0.4232190 -0.5050696 -0.4681147 -0.3586550 -0.3593097 -0.3885466 -0.3063615 -0.2452320
2  0.5274149   0.6744523  1.1152714  0.9204177  0.3728016  0.3722992  0.5281886  0.2716080 -0.3344675
3  4.1316281   3.8987289  3.1061820  3.4436455  4.3075398  4.3119415  4.0690924  4.0970415  1.3191444
4 -0.4766986  -0.4994858 -0.6014253 -0.5558980 -0.4330672 -0.4244829 -0.4932892 -0.5532973  4.2769049

Clustering vector:
       BUENOS AIRES          CATAMARCA              CHACO             CHUBUT             CÓRDOBA
                  3                  1                  1                  1                  2
         CORRIENTES          ENTRE RÍOS            FORMOSA              JUJUY          LA PAMPA
                  2                  2                  1                  1                  2
          LA RIOJA            MENDOZA           MISIONES           NEUQUÉN          RÍO NEGRO
                  1                  1                  4                  1                  1
              SALTA          SAN JUAN          SAN LUIS         SANTA CRUZ          SANTA FE
                  1                  1                  1                  1                  2
SANTIAGO DEL ESTERO    TIERRA DEL FUEGO            TUCUMÁN
                  1                  1                  1

Within cluster sum of squares by cluster:
[1] 7.592569 4.279885 0.000000 0.000000
 (between_SS / total_SS =  94.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
>
```

Figure 10: Call K-Means

2.1.3 Aplicación de Dendograma y Correlación

Este Gráfico de tipo dendograma, con el agregado de un mapa de correlación, nos permite determinar no sólo la cantidad de agrupamientos y la distancia inherente entre estos sino que también nos permite determinar que objeto o región es mas propensa a una cierta densidad de distribución por especies ganado bovino.

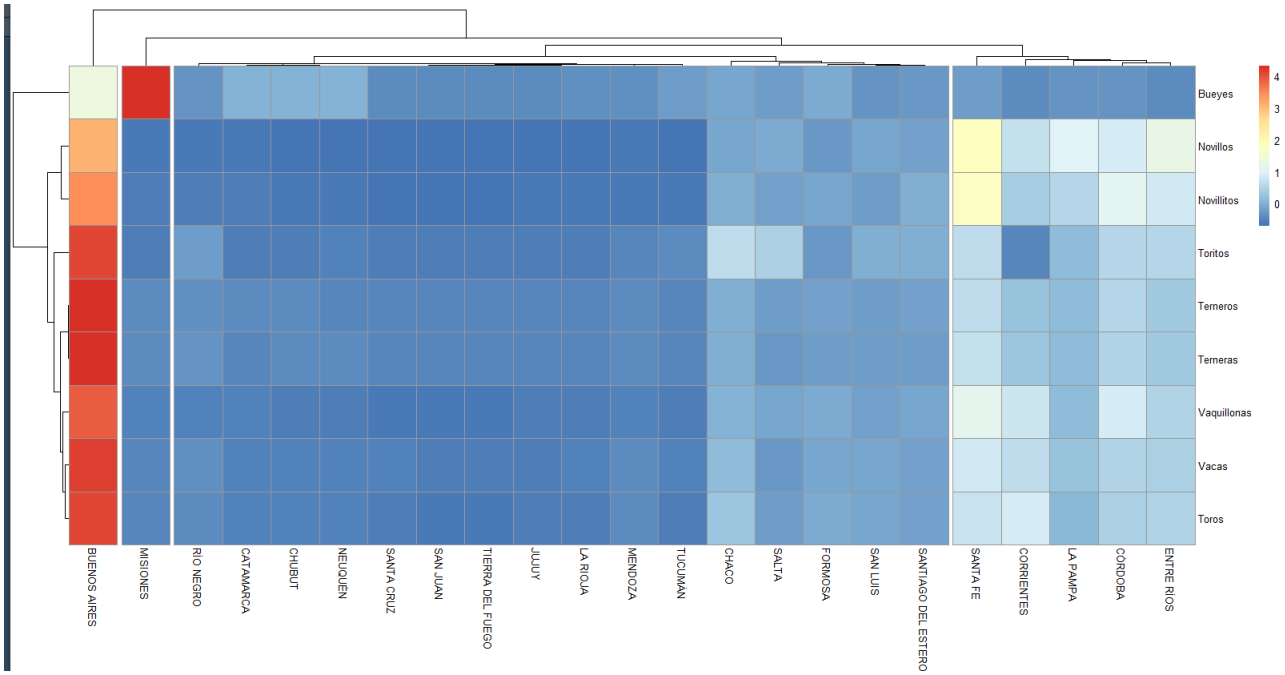

Figure 11: Gráfico DendoGrama y Correlación

Analsis cantidad de Kmeans Optimos

Mediante un Gráfico del metodo wss vamos a poder determinar cuantos Cluster son óptimos para nuestro modelo de Agrupamiento.

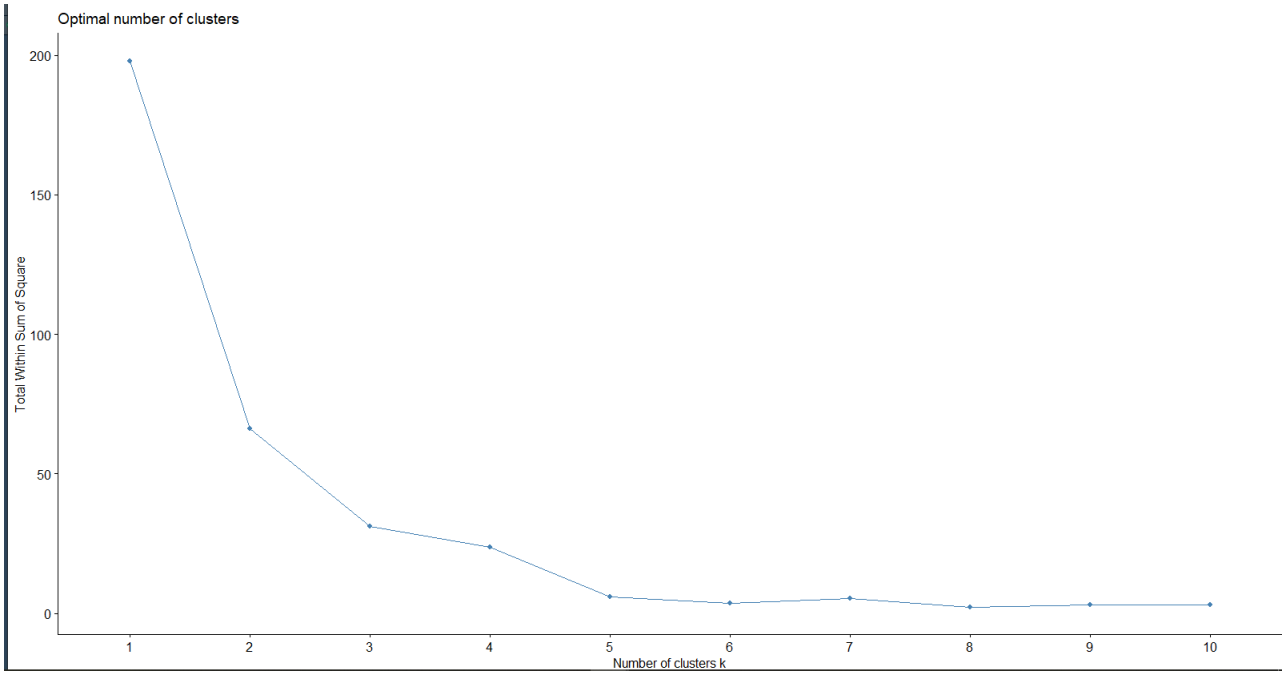

Figure 12: Grafico de Metodo wss

3.0.1 Análisis de 3 Cluster

Agrupamiento de distancia mínima para 3 Cluster

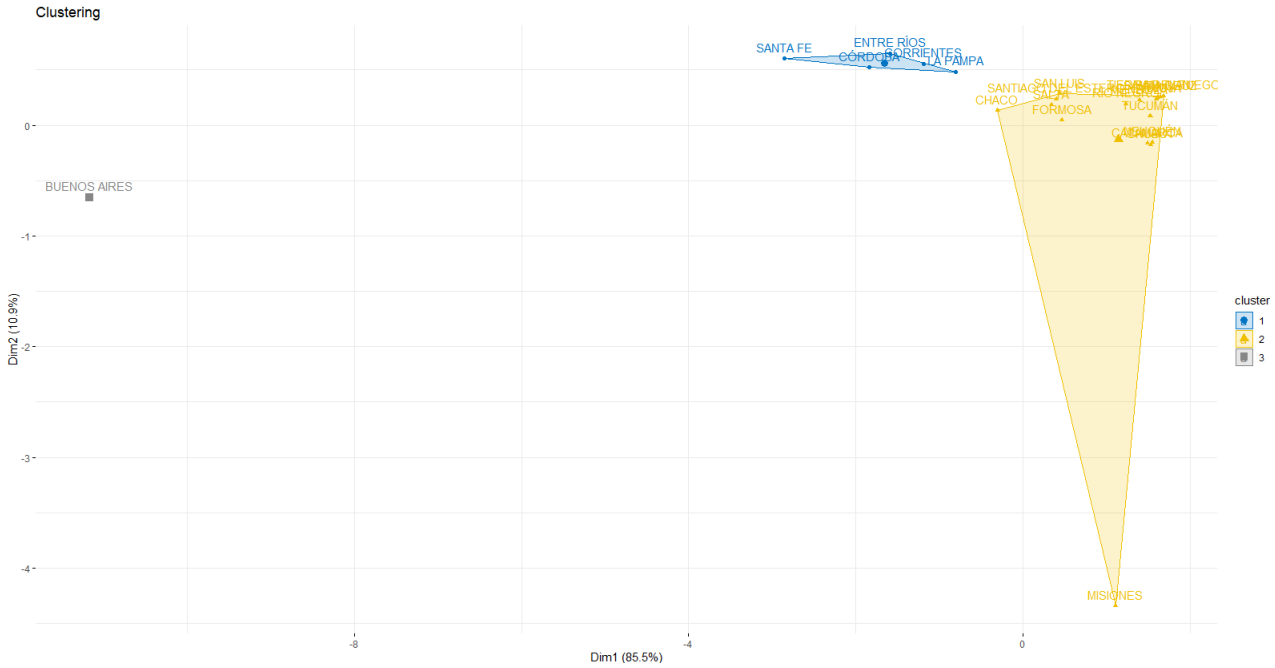

Figure 13: Grafico K-Means 3 Cluster

3.0.2 Analsis de 2 Cluster

Agrupamiento de distancia minima para 2 Cluster

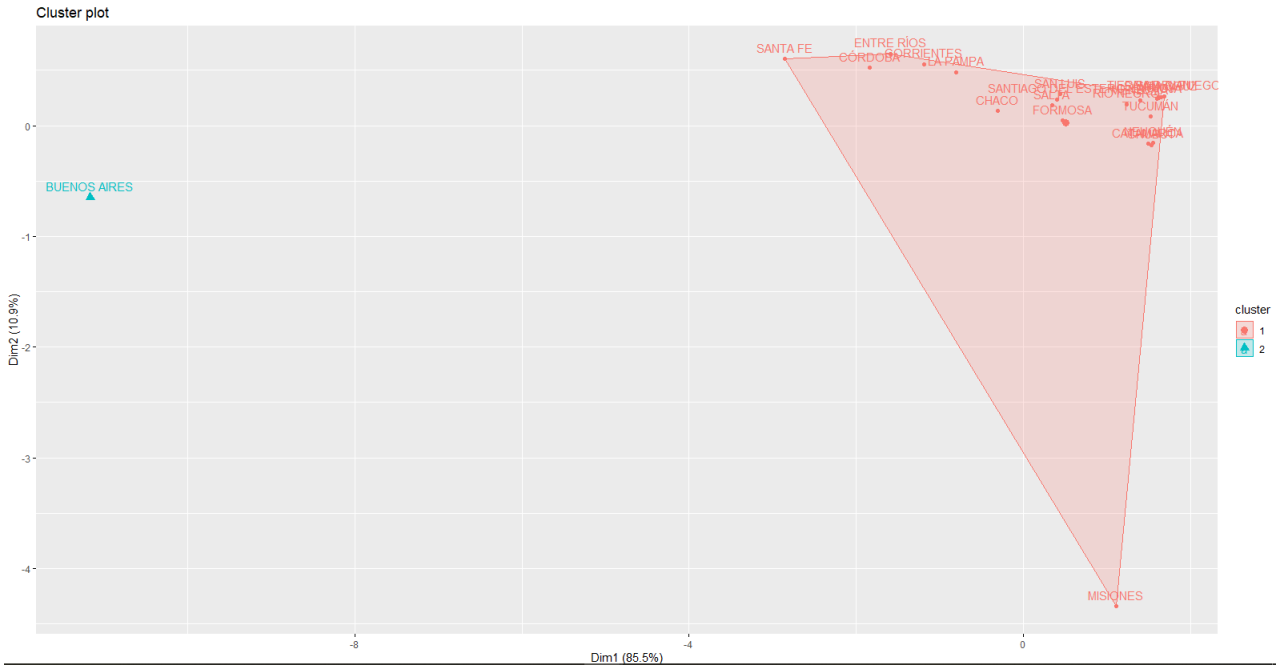

Figure 14: Grafico K-Means 2 Cluster

Análisis de Dendograma

El gráfico de Dendograma nos permitirá observar los grupos que se unen basándose en una distancia entre los dos miembros más cercanos. Ward es un método que tiende a formar clusters más compactos y de igual tamaño y forma.

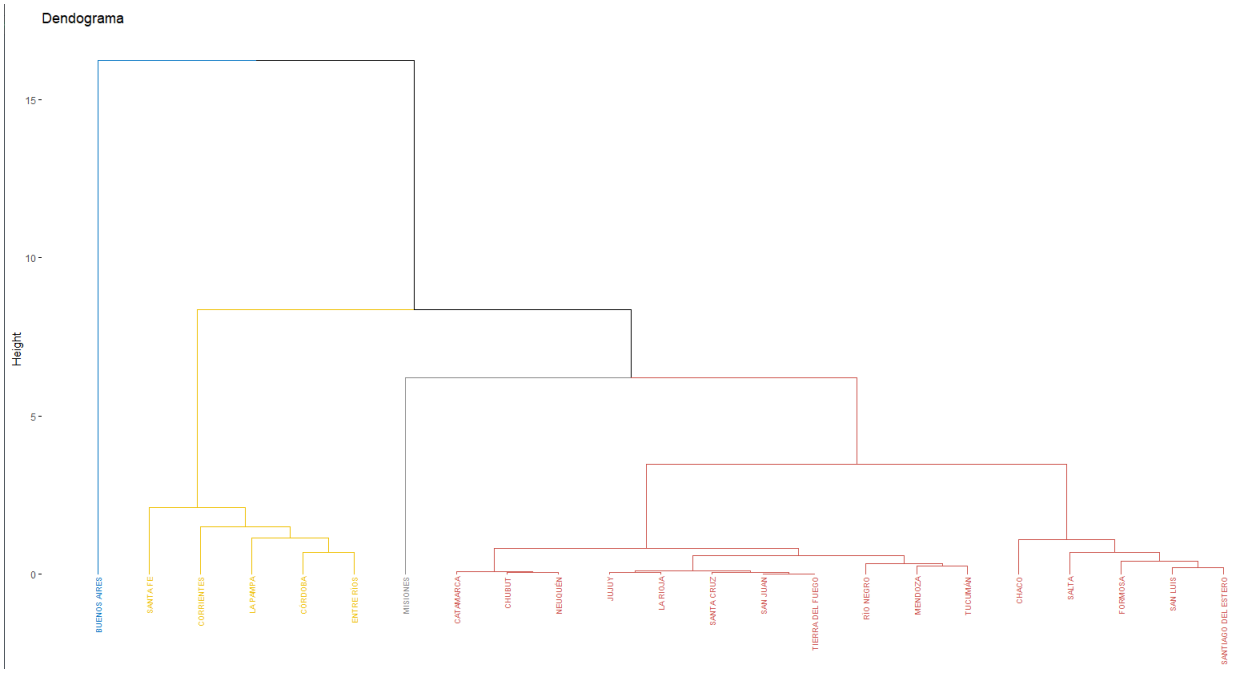

Figure 15: Grafico Dendograma

A partir de la observación de los gráficos, podemos establecer una coincidencia entre los valores de distribución de ganado vistas en el análisis exploratorio de los datos y los obtenidos mediante el clustering: Buenos Aires se encuentra tan alejada del resto de las provincias que incluso ocupa todo un cluster para sí misma. Luego tenemos agrupadas las provincias de la región pampeana más Corrientes, y por último se agrupan las restantes. También observamos que la provincia de misiones se encuentra bastante alejada de las demás, pero hacia el lado contrario al de Buenos Aires (En el caso de 4 clusters, incluso ocupa también un cluster completo). Podemos inferir que esta distribución del ganado

tiene que ver con las características geográficas de cada región, que son factores determinantes al momento de decidir qué actividad podrá desarrollarse allí.
Cabe señalar también que, según los datos del Inventario de GEI, Buenos Aires presenta la mayor emisión de gases del país en cuanto al sector agro, y Misiones es la de menor valor.

Conclusión

Luego del análisis realizado, determinamos que la principal provincia en actividad ganadera es, por gran diferencia, buenos Aires, seguida de la Región Pampeana y luego en tercer lugar todo el resto del país que, probablemente por cuestiones geográficas como relieve, clima y demás, no son muy óptimas para el desarrollo de este sector. En último lugar hallamos a Misiones que, en parte por ser una provincia pequeña y también por sus características en general, ha sido de las menos explotadoras de la industria del ganado.
Por otro lado, podemos suponer que existe una relación notable entre las distribuciones de ganado y las diferentes ramas del sector, con los valores de emisión de Gases de Efecto Invernadero que, curiosamente, tienen a Buenos Aires en el extremo superior con un valor de 24,91 MtCO2eq (Millones de toneladas de carbono equivalente), y Misiones en el extremo inferior con -4,86 MtCO2eq. Observando la distribución de las provincias en los clusters, hallamos coincidencia en este sentido (No sucede lo mismo con el resto de las provincias, en las cuales encontramos poca correlación entre los valores de GEI y las distribuciones del ganado). Esto tal vez tenga que ver con que los datos de emisiones de gases incluyen a todo el sector agro, y no sólo a ganadería, y también con el tipo de ganadería que predomine en cada provincia, que no se haya contemplado en nuestro set de datos.
No obstante, podemos afirmar con seguridad que las emisiones de gases producto de la ganadería son un tema preocupante en cuanto al riesgo ambiental que ello implica, principalmente en Buenos Aires, que está a la cabeza de esta actividad.

Anexo

```r
library(tidyr)
library(readxl)
library(dplyr)
library(forecast)
library(ggvis)
library("highcharter")
library(devtools)
library(factoextra)
library(cluster)
library(pheatmap)

DF_ganado = read_excel("C:/Users/brian/Desktop/Explotacion de Datos/TP6/GANADO_BOBINO/Dataset/Dataset_Clustering.xlsx")
View(DF_ganado)
str(DF_ganado)


############ Normalizacion de los Datos
ganado2=as.data.frame(DF_ganado)
row.names(ganado2)=ganado2$Provincia
ganado2$Provincia=NULL

scale_ganado2 = scale(ganado2)
head(scale_ganado2)


########### Matriz de Correlacion
ganado_hc = dist(scale_ganado2, method = "euclidean")
fviz_dist(ganado_hc)

##Con al matriz se busca explicar que tan corrlacionados estan los datos en niveles de densidad

ganado_dist.cor <- get_dist(scale_ganado2, method = "pearson")
fviz_dist(ganado_dist.cor)

ganado_dd <- daisy(scale_ganado2)
fviz_dist(ganado_dd)


############ K-means
scale_ganado_km = kmeans(scale_ganado2,4)
```

```r
aggregate(scale_ganado2, by=list(cluster=scale_ganado_km$cluster), mean)
fviz_cluster(scale_ganado_km, data=scale_ganado2)

########### Dendograma de Correlaciones
pheatmap(t(scale_ganado2), cutree_cols = 4)

########### Dendograma
ganado2_res.hc <- hclust(dist(scale_ganado2), method = "ward.D2")
fviz_dend(ganado2_res.hc, cex = 0.5, k = 4, palette = "jco", main = 'Dendograma')

########### Numero Optimo de Clusters
fviz_nbclust(scale_ganado2, kmeans, method = "gap_stat")
fviz_nbclust(scale_ganado2, kmeans, method="wss")

################## Dendograma de Correlacion
pheatmap(t(scale_ganado2), cutree_cols = 4)


################## Kmeans con 2 Cluster
scale_ganado3_km = kmeans(scale_ganado2,2)
aggregate(scale_ganado2, by=list(cluster=scale_ganado3_km$cluster), mean)
fviz_cluster(scale_ganado3_km, data=scale_ganado2)
```