Trabajo Práctico Final Laboratorio de Datos

Grupo 3 Manuela Cerdeiro Martín Estévez

Introducción

La Ilusión de los Mapas

Cuando uno observa un mapa político de cualquier país, en particular en este caso la Argentina, todo parece estar distribuido de manera más o menos equitativa.

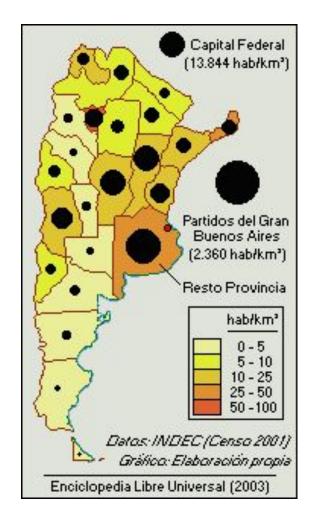
Una provincia aquí, otra provincia allá. Aún viendo los departamentos y partidos, cualquier persona que lo mira sin pensarlo mucho podría verse tentada a pensar que las divisiones siguen un criterio balanceado.

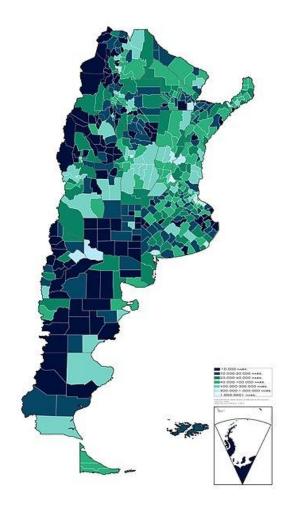




La Ilusión de los Mapas

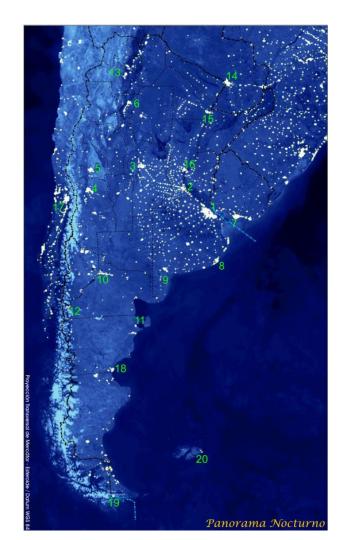
La realidad, sin embargo, es otra. Por diversas razones históricas y económicas, nuestro país está entre lo más asimétricos en lo que a densidad de población se refiere. Más allá de lo esperado, alrededor del 30% de la población se concentra en un solo centro urbano: otro 30% en un radio de no más de 500 km alrededor de éste. Y con la población concentrada, es esperable que el resto del país se vea olvidado en muchas cuestiones de infraestructura.





Las Luces del Estadio

Basta con ver un mapa nocturno. La Cuenca del Plata y las capitales provinciales concentran la mayor parte de la población y por ende de la infraestructura. Además, suelen ser los centros de cultura, educación y, en concreto para este trabajo, salud de muchas provincias.

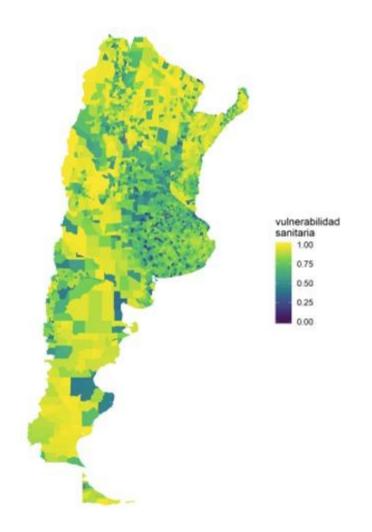


La Cuestión de la Salud

Sobre esto último, utilizaremos los datos de la Fundación Bunge y Born y del INDEC. En concreto, la fundación elabora un "Índice de Vulnerabilidad Sanitaria" (IVS), que representa la vulnerabilidad de una población en un área censal dada.

En el informe de donde se sacó este mapa señalan como notable la disparidad que hay entre la región central del país y el resto de las zonas periféricas. Sin embargo, nos interesaba poner la lupa en un punto en concreto.

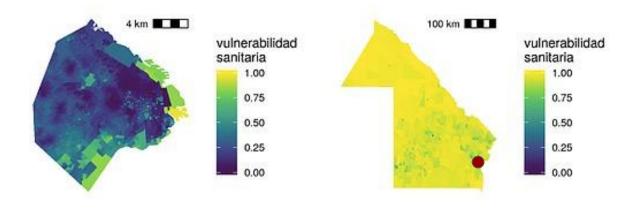
Notamos que muchas provincias del a veces llamado "interior profundo" tienen sus capitales muy alejadas del resto de la provincia y mucha población concentrada en ellas. Nuestra hipótesis inicial fue, entonces, que la lejanía de una ciudad a su capital provincial afecta negativamente su IVS.



Diferencias entre Jurisdicciones

Hay algunos casos, como los de la imagen, que son extremos y que pueden entenderse si se comparan las situaciones de las provincias en general. Sin embargo, usando a Chaco como ejemplo, se puede ver claramente que los centros urbanos tienen un mejor puntaje que las zonas rurales y en particular su capital.

Nos pareció interesante entonces, estudiar cada provincia por separado y comparar los resultados tratando de ver cuánto incidía la distancia comparativamente, no en términos absolutos.



CABA

Chaco (en rojo Resistencia)

Hipótesis Inicial

Nuestra hipótesis inicial, como se mencionó, es que la distancia a la capital es de hecho un factor que está correlacionado positivamente con el empeoramiento de la calidad de vida en lo que a salud respecta.

Además, vamos a analizar algunos otros factores como la distancia a hospitales de alta complejidad y a centros universitarios a modo exploratorio para testear algunos otros factores y además nos interesa ver *qué tanto* las capitales actuales se alejan de un teórica "ideal" y si ésto es algo notorio.

Desarrollo

Obtención de la Información



- Datos censales INDEC + CEUR
- IVS Bunge y Born.
- Distancias API de Google Maps.
- Wikipedia



Centro de Estudios Urbanos y Regionales

Consejo Nacional de Investigaciones Científicas y Técnicas







Distancia Geométrica vs. Distancia en Ruta vs. Tiempo de Viaje ¿Qué usar para medir la distancia entre capitales?



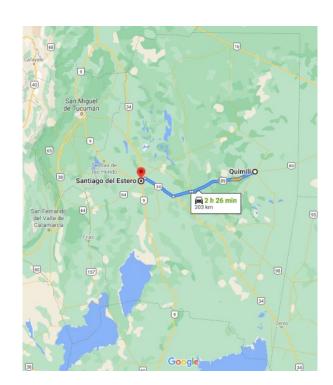
Distancia Geométrica vs. Distancia en Ruta vs. Tiempo de Viaje ¿Qué usar para medir la distancia entre capitales?

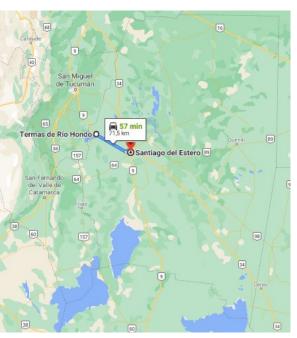


Procedimiento

 Analizar para cada provincia por separado las distancias entre poblaciones y la capital.

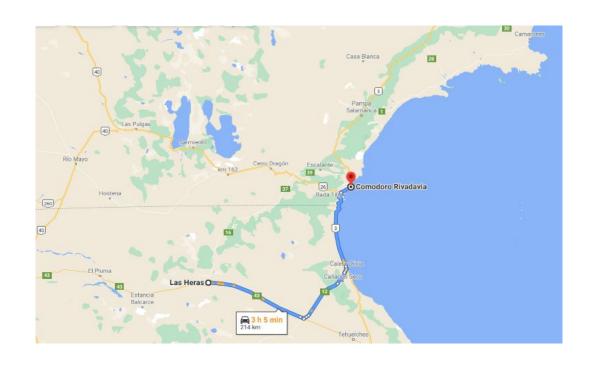
 Contrastar valores y ver qué tanto la capital se aleja de las otras ciudades





Algunos Problemas

Al dividir todo por provincias, no nos permitimos considerar la distancia a centros regionales que pueden llegar a ser más relevantes que una capital pero están fuera de los límites de la provincia.



Armado de diccionario{provincia: capital}guardado en formato .json

Armado de lista de diccionarios con:

```
{"departamento": " ",
  "provincia": " ",
  "cabecera": " ",
  "poblacion_dep": " ",
  "tipo_ciudad": " "}
```

```
"departamento": "12 de Octubre".
                                                                                                        Viedma-San_Carlos_de_Bariloche
"provincia": "Chaco",
                                                                                                        Viedma-General Conesa
"cabecera": "General Pinedo",
                                                                                                        Viedma-El_Cuy
"poblacion dep": "22225",
"tipo ciudad": "Ciudad"
                                                                                                        Viedma-General Roca
                                                                                                        Viedma-Sierra Colorada
"departamento": "25 de Mayo",
                                                                                                        Viedma-Rio_Colorado
"provincia": "Chaco",
                                                                                                        Viedma-Pilcaniveu
"cabecera": "Machagai",
"poblacion dep": "29315",
                                                                                                        Viedma-San Antonio Oest
"tipo ciudad": "Ciudad"
"departamento": "25 de Mayo",
"provincia": "Misiones".
"cabecera": "25 de Mayo",
"poblacion_dep": "27754", 
"tipo ciudad": "Ciudad"
"departamento": "9 de Julio",
"provincia": "Chaco",
"cabecera": "Las Brenas"
"poblacion dep": "30153",
                                                     "Rio Negro" : "Viedma",
"tipo ciudad": "Ciudad"
                                                     "Catamarca": "San Fernando del Valle de Catamarca",
                                                     "Cordoba" : "Cordoba",
                                                     "Corrientes": "Corrientes",
                                                     "La Pampa" : "Santa Rosa",
"departamento": "9 de Julio",
"provincia": "Santa Fe",
                                                     "La Rioja" : "La Rioja",
"cabecera": "Tostado",
                                                     "Mendoza" : "Mendoza",
"poblacion_dep": "29810",
                                                     "Misiones" : "Posadas",
"tipo_ciudad": "Ciudad"
                                                     "Salta" : "Salta".
                                                     "San Juan" : "San Juan",
                                                     "Santiago del Estero" : "Santiago del Estero",
                                                     "Tucuman" : "San Miguel de Tucuman",
                                                     "Neuguen": "Neuguen".
                                                     "Formosa": "Formosa".
                                                     "Santa Cruz" : "Rio Gallegos",
                                                     "San Luis" : "San Luis",
                                                     "Santa Fe" : "Santa Fe de la Vera Cruz",
                                                     "Buenos Aires" : "La Plata",
                                                     "Entre Rios" : "Parana",
                                                     "Chubut" : "Rawson",
                                                     "Chaco" : "Resistencia"
                                                     "Jujuy" : "San Pedro de Jujuy",
                                                     "Tierra del Fuego, Antartida e Islas del Atlantico Sur" : "Ushuaia"
```

duracion 3 hours 24 mins

distancia 293 km duracion 3 hours 36 mins distancia 37.6 km

Con estos datos hicimos el primer intento. Hubo algunos problemas técnicos ya que al usar los nombres de las ciudades google tendía a confundirse.

Además, mucha información no estaba disponible con ese nivel de granularidad por lo que cambiamos el set. Nos fue útil para algunas cosas que mostraremos al final.

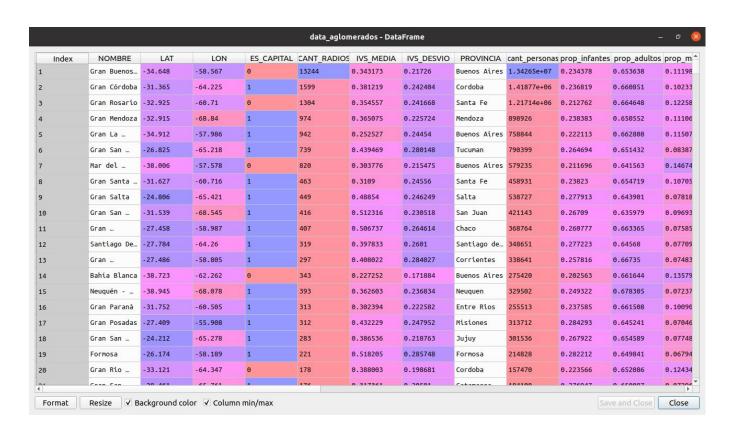
```
urllib.request
departamentos = ison.loads(file departamentos.read())
file_capitales = open("capitales.json", "r")
capitales = json.loads(file_capitales.read())
provincias = capitales.keys()
    cap in provincias:
     distancias[cap] = []
         capital = capitales[dep["provincia"]].replace(" ", "_")
         ciudad = dep["cabecera"].replace(" ",
         crudud - dep[ 'provincia"].replace(" ", "_")
url = "https://maps.googleapis.com/maps./api/distancematrix/json?origins=" + ciudad
         dato de google = urllib.request.urlopen(url).read()
         datos_json = json.loads(dato_de_google)
          data = datos_json["rows"][0]["elements"][0]
                                                                                                                        Cordoba/ciu Cordoba/distancia
                                                                    Catamarca/c Catamarca/distancia
                                                                                                  Catamarca/tiempo
                                                                                                                                                       Cordoba/tiempo
                                         19499 477296.25 477296.25 La Puerta 10770.8
                                                                                           10770,8 199846.666( 199846,67 San_Agustin 9225.88
                 San Carlos (22686,6666) 22686,667 508162,916 508162,92 Andalgala 12410,8666 12410,867 256819,266 256819,27 Jesus Maria 8460,08
                                                                                                                                                8460,08 190883.12
                 General Cor 15888.1666i 15888,167 380651.333 380651,33 Antofagasta 29873.2
                                                                                                                                                10933,56 228381.2
                          16908.0833: 16908,083 365774.416( 365774,42 Belen
                 General Roc 17269.9166 17269.917 386389.666 386389.67 Huillapima 8930.73333 8930,7333 184350.2
                                                                                                                                                9917,04 218284.88
                 Sierra Color 13610.6666 13610,667 305274.5 305274.50 San Fernand 8881.73333 8881,7333 171648.4
                                                                                                                                                  9762 221550 44
                                                                                                                                                                   221550 44
                 norquinco 25589.6666i 25589,667 506844.916i 506844,92 El_Alto
                                                                             12386.66661 12386,667 203117.2
                                                                                                                                               13354,28 298863.32
file write
                 Rio Colorad 19300.5833 19300,583 459451.833 459451,83 Fray Mamer 10071.6666 10071,667 184483.533 184483,53 Marcos Juar 13131.28
                                                                                                                                               13131,28 318148.76
                                       20763.5 474554.333: 474554,33 Recreo
                                                                            12896 26661 12896 267 276612 6
                                                                                                                                               12352 08 265367 04
                 San Antonio 14609.8333 14609,833 345852.416 345852,42 La Merced 9428.06666 9428,0667 190878.866 190878,87 Salsacate 12187.68
                                                                                                                                               12187 68 263603 36
              3 Valcheta 13960.6666 13960,667 318098.916 318098,92 Saujil
                                                                             11170.7333: 11170,733 232796.2
                                                           339862 50 Sente Maria 17149 1333 17149 133 315618 866 315618 87 Cosquin 9310 84
                                                                                                                                                9310.84 197482.52
                                                                    Banado_de_ 11005.5333: 11005,533 222175.4
                                                                                                                                                12082 36 264946 32
                                                                                                                                                11822.2 274261.64
                                                                                                                         Villa del Ro 9281.68
                                                                                                                                                9281.68 199813.6
                                                                                                                                                                    199813.6
                                                                                                                         Villa Cura B 11812.6
                                                                                                                                                11812.6 248193.3
                                                                                                                       Villa_Dolore 13330.24
                                                                                                                                                13330.24 271041.52
                                                                                                                                                                   271041.52
                                                                                                                         San Francisc 13949.4
                                                                                                                                                13949.4 302501.04
                                                                                                                                                                   302501.04
                                                                                                                                                8367.96 179458.48
                                                                                                                                                9015.64 198228.04
                                                                                                                                                                   198228.04
```

Luego juntamos los datos de las fuentes utilizando (de Pandas).

- df con aglomerados y sus centroides (en lat-lon)
- df con info de qué radios censales pertenecen a cada aglomerado
- df con info demográfica por cada radio censal
- df con info de ivs por radio censal
- diccionario de capitales de provincia

Juntamos todo en un data frame

seleccionamos aglomerados con al menos 4000 personas

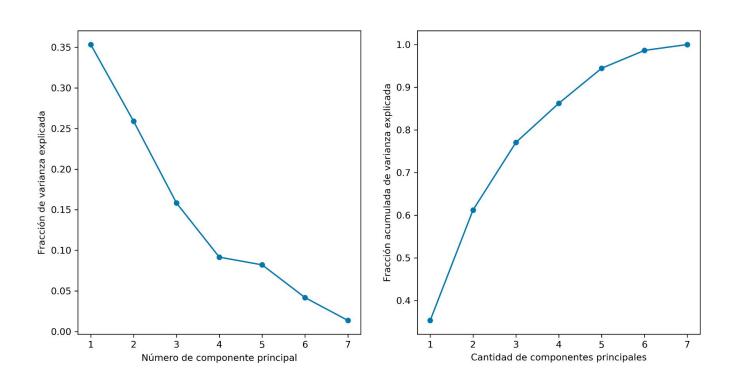


Con esta nueva base de datos fue mucho más sencillo utilizar la api de google maps. También sirvió para expandir notablemente la cantidad de sitios a considerar.

Debido a esto último, se crearon algunos diccionarios auxiliares para optimizar el proceso de búsqueda.

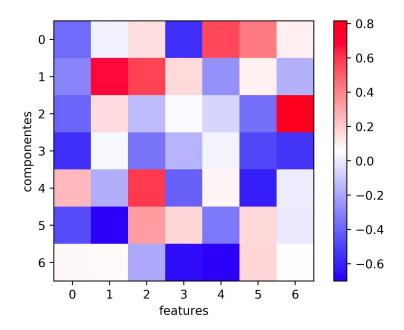
```
aglomerado in aglomerados:
  aglomerado["ES CAPITAL"] == 0:
    capital = capitales[aglomerado["PROVINCIA"]].replace(" ", " ")
    lat lon = str(aglomerado["LAT"]) + "," + str(aglomerado["LON"])
   url = "https://maps.googleapis.com/maps/api/distancematrix/
        json?origins=" + lat lon + "&destinations=" + capital + "+"
           aglomerado["PROVINCIA"].replace(" ", "_") + "+Argentina"
           "%region=.ar" + "%key=" + API key
   dato de google = urllib.request.urlopen(url).read()
   datos json = json.loads(dato de google)
    data = datos json["rows"][0]["elements"][0]
    if data["status"] == "OK" :
        aglomerado["DIST CAP PROV"] =
                                      data["distance"]
        aglomerado["TIEMPO CAP PROV"]
                                        data["duration"]
```

- + agregamos log(poblacion)
- + seleccionamos features a usar
- + escalamos datos
- + realizamos un PCA

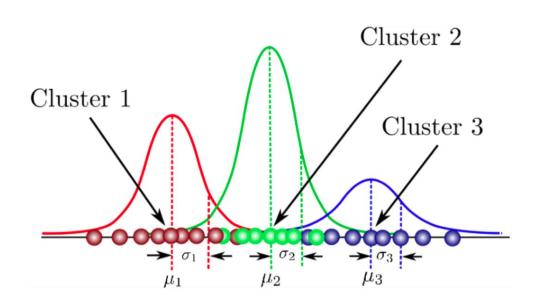


Vemos cómo se componen las primeras componentes principales.

```
In [182]: d
Out[182]:
{0: 'IVS_MEDIA',
   1: 'IVS_DESVIO',
   2: 'log_pob',
   3: 'prop_infantes',
   4: 'prop_mayores',
   5: 'prop_mujeres',
   6: 'TIEMPO_CAP_PROV'}
```



Gaussian Mixture Model



Generalización de K-Means.

Es una familia de modelos de aprendizaje no supervisado basados en sumas de normales, o sea de gaussianas. Tiene varios parámetros que dan lugar a distintos clasificadores.

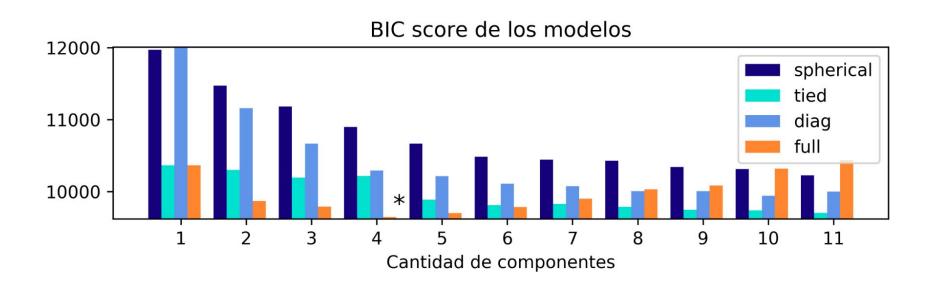
Tiene que ver con la forma y cantidad de las gaussianas.

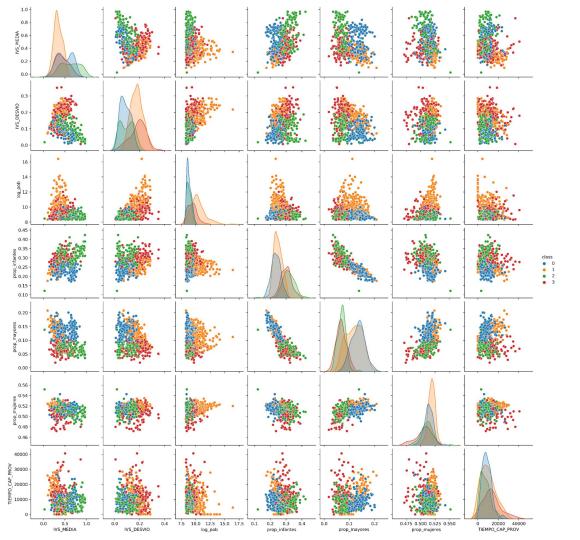
Gaussian Mixture Model

```
from sklearn import mixture
#GMM Gaussian Mixture Model
lowest_bic = np.infty
bic = []
n_components_range = range(1, 12)
cv_types = ['spherical', 'tied', 'diag', 'full']
for cv_type in cv_types:
    for n components in n components range:
        gmm = mixture.GaussianMixture(n_components=n_components,
                                       covariance_type=cv_type)
        gmm.fit(X)
        bic.append(gmm.bic(X))
        if bic[-1] < lowest_bic:</pre>
            lowest bic = bic[-1]
            best gmm = gmm
bic = np.array(bic)
```

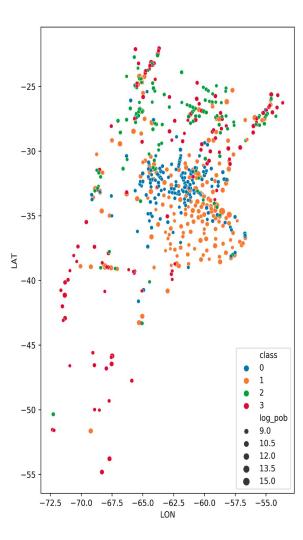
Probamos con distintas combinaciones de cantidad de gaussianas y de formas de las mismas, seleccionando el modelo con mejor bic.

Gaussian Mixture Model



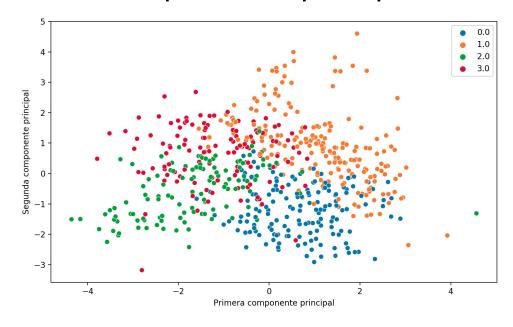


- La clase naranja es la de aglomerados con más población. Ahí hay menor IVS.
- → Las clases verde y azul tienen menos población, y mayor IVS media, con bajos desvíos.
- → La proporción de mujeres es alta en lugares de mayor población.
- → Las clases roja y verde tienen baja proporción de adultos mayores, las clases azul y naranja tiene alta proporción.
- → Algo obvio: Las capitales tienen alta población. Las proporciones de infantes y de mayores tienen una relación inversa.



Dónde se ubican las clases

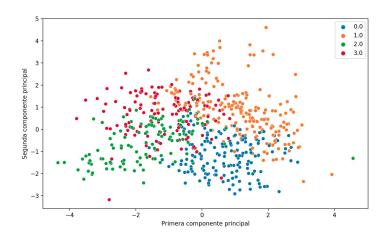
Proyección a 2D componentes principales

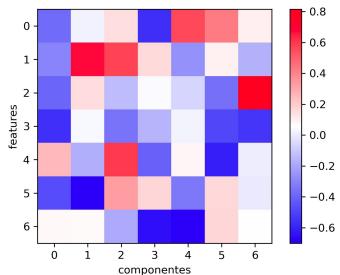


Conclusiones

Pudimos ver, después de analizar varias combinaciones de variables, que parece haber alguna correlación clara entre lo alejada que una población esté de su capital y del IVS, aún con el modelo más simple.

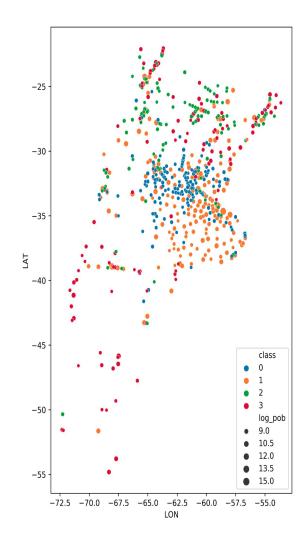
Existen outliers, por supuesto, pero pareciera notarse un patrón.



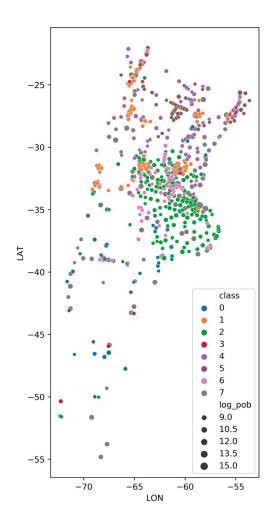


En el ejemplo, sólo utilizamos la distancia a la capital como variable a tomar en cuenta y parecieran haberse formado 4 grupos bastante definidos:

- Naranja: aglomerados grandes con mucha población.
- Azul: eje Santa Fé Córdoba
- Zonas más extremas N y S del país (verde principalmente al norte).

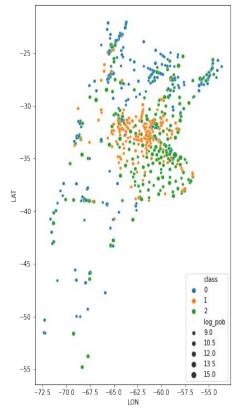


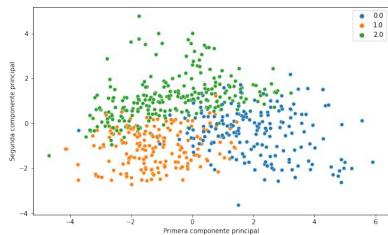
Otro experimento considerando más features, en el que nos dieron 8 clases.



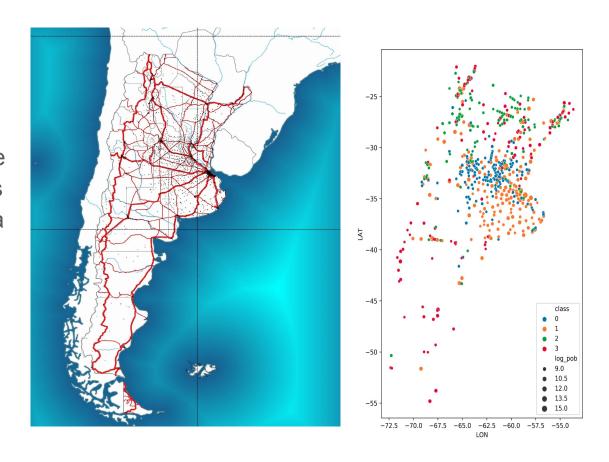
Otro caso interesante fue el que nos dio considerando la proporción de gente que no asistió a la escuela.

Parecen formarse 3 grupos aún más definidos en las regiones del interior.





Es interesante notar cómo muchos de los lugares que parecen estar mejor en las zonas periféricas van de la mano con los sitios por donde pasan las rutas nacionales (las líneas gruesas a la izquierda).

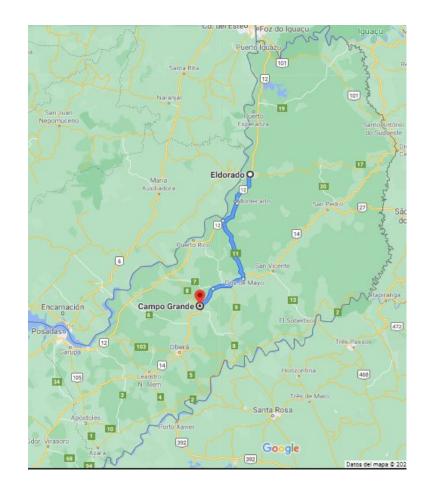


Entre las cosas que se pudieron hacer con los datos iniciales, fue un cálculo muy sencillo para ver cuál sería una capital "ideal".

La definición utilizada fue que la ciudad elegida debía ser la que esté más cerca de todas las otras en promedio.

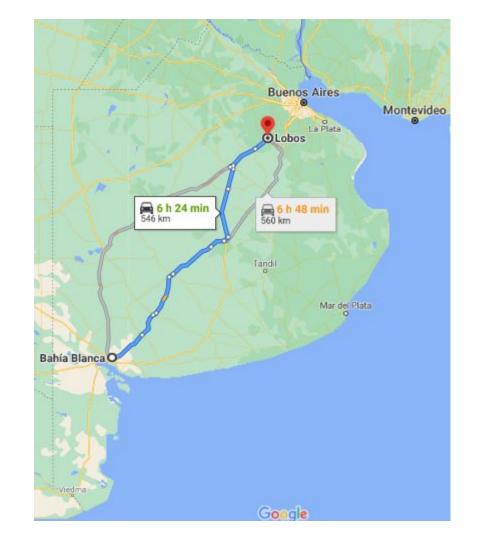
De las 23 provincias, 7 de ellas tienen una capital que cumple con esta condición.

Y de las 16 restantes, hay 5 de las cuales la ciudad que se encontró está al menos un 40% más cerca al resto que la capital actual.



Es importante recalcar el hecho de que algunas de las potenciales capitales usando nuestro criterio pueden no ser ideales desde un punto de vista práctico.

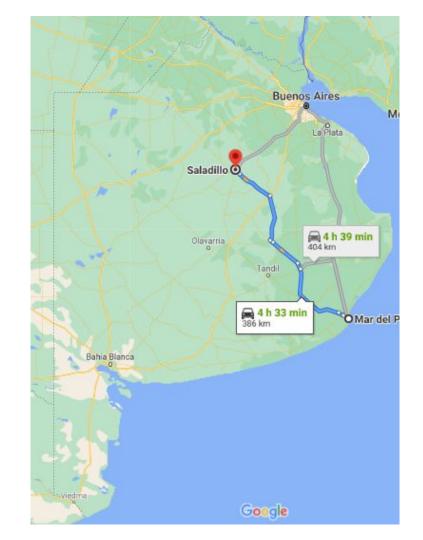
Además hay un caso notable en el cual, por las particularidades demográficas de la región, nuestra forma de analizarlo pierde un poco el sentido: Buenos Aires



En el caso particular, se debe a la inmensa cantidad de municipios en el conurbano.

Si tomamos al conurbano como una sola ciudad, la situación cambia un poco.

Hay que tener en cuenta que sigue habiendo más densidad poblacional en el norte.



FIN

Fuentes:

- INDEC https://www.indec.gob.ar/
- Datos del INDEC condensados -https://dump.jazzido.com/CNPHV2010-RADIO/
- Fundación Bunge y Born https://www.fundacionbyb.org/vulnerabilidad-sanitaria
- Mapas https://mapamundi.online/
- Mapas https://lageografiaweb.blogspot.com/
- Información General https://es.wikipedia.org