INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES

DE OCCIDENTE

ANALISIS ESTADISTICO MULTIVARIADO

Examen 2

Presentan:
Diego Canales Morales

Profesor: Mtro. Alan Topete
Fecha: 6/05/2024

Análisis Exploratorio

Para el análisis Exploratorio escogí las columnas “latitude”, “longitude” y “ocean_proximity” busque valores nulos (no había), use un .describe(), un histograma y un gráfico de pares.

Histograma de “latitude” y “longtude”

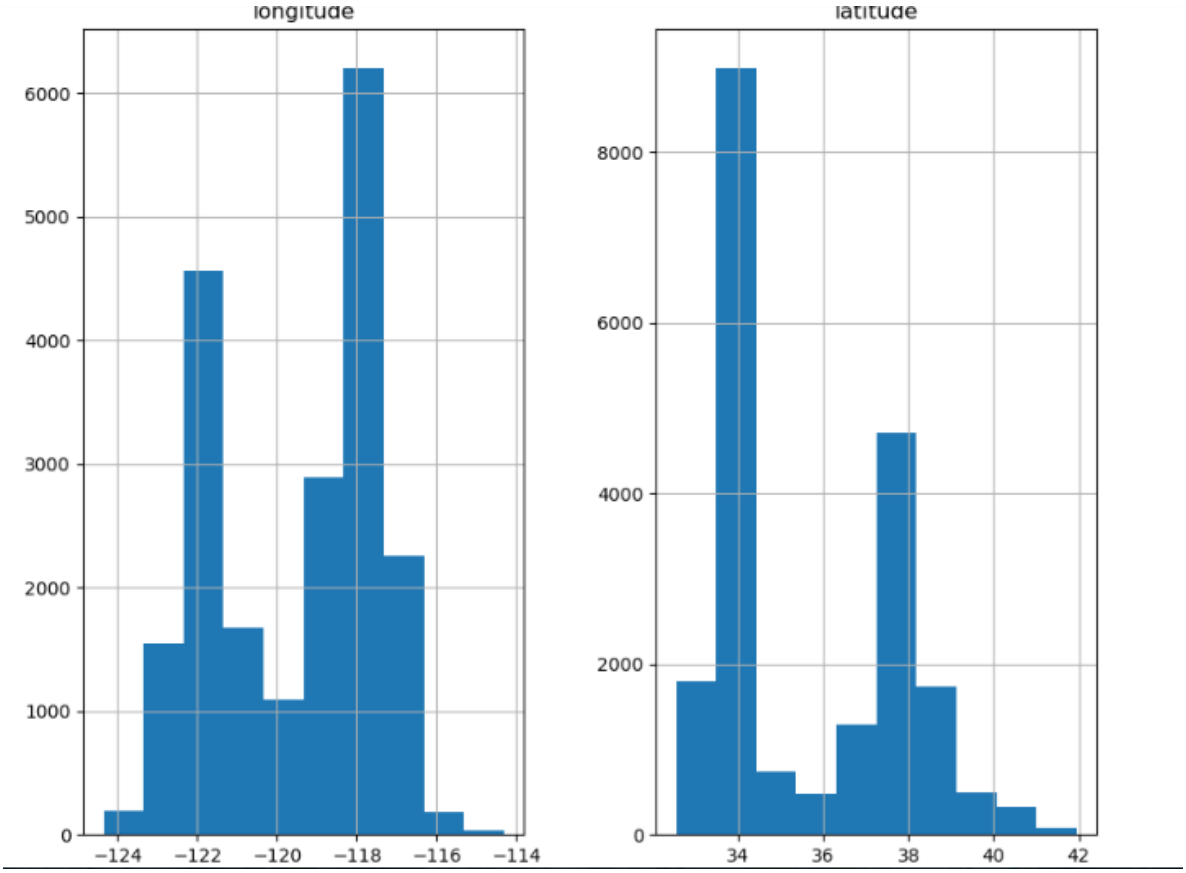
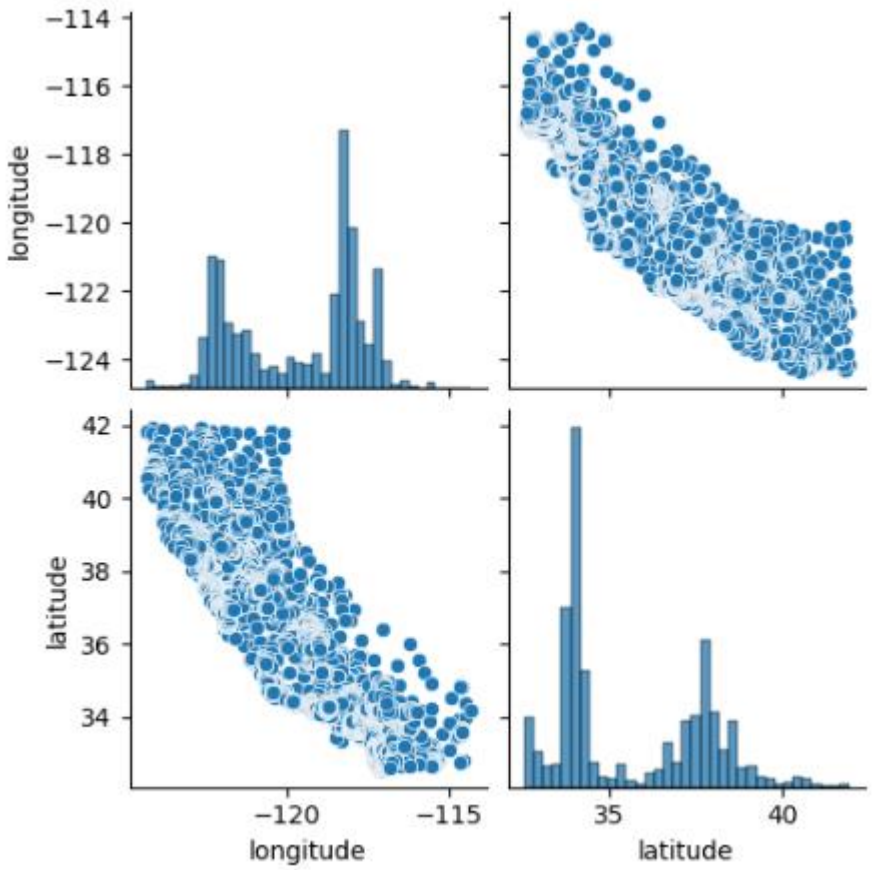


Gráfico de pares

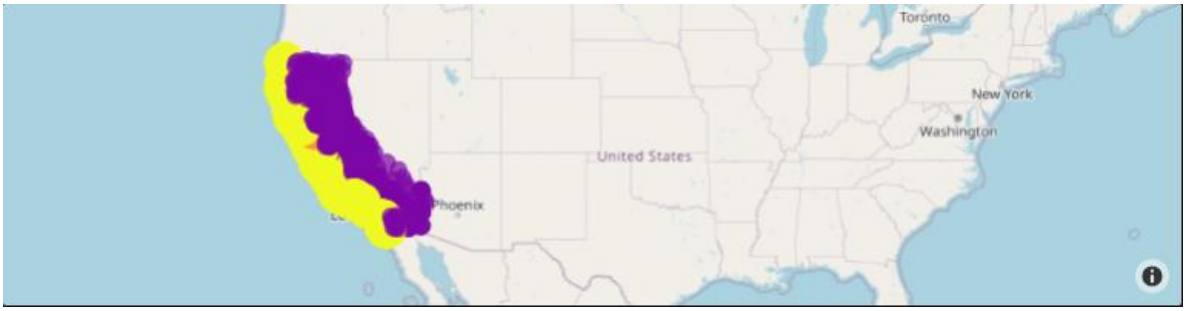


Después vi las definiciones de “ocean_proximity” para saber de que estoy viendo la categoría del punto geográfico.

```
df["ocean_proximity"].unique()
```

```
array(['NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND'],
      dtype=object)
```

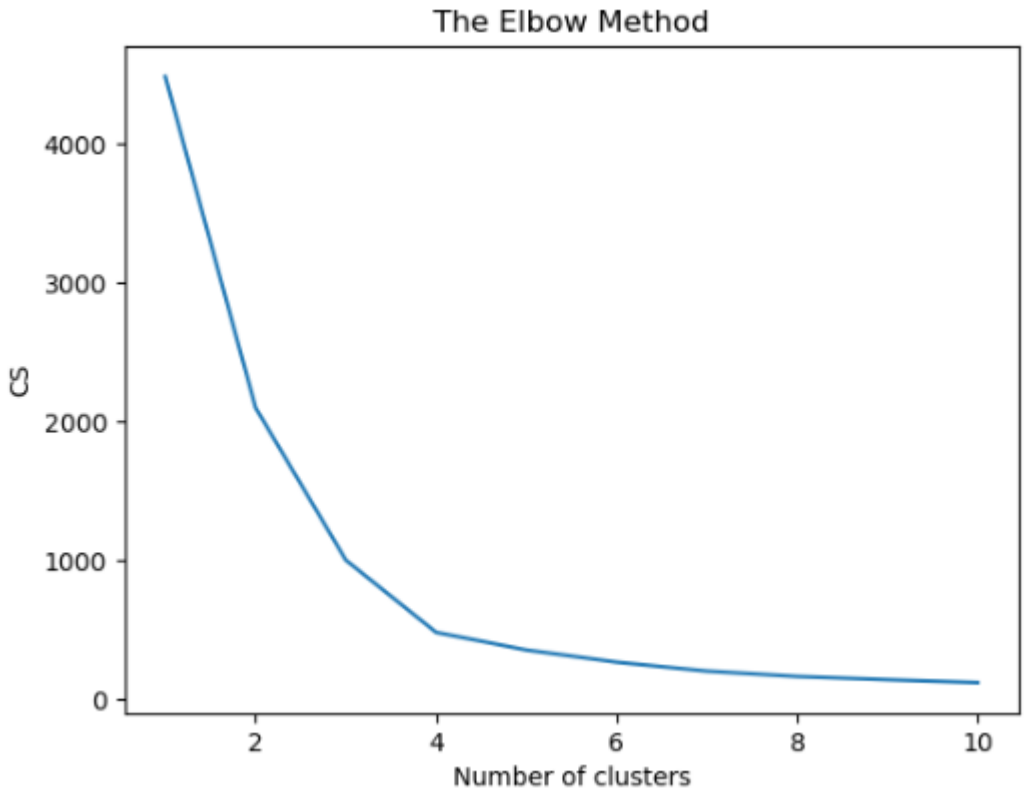
Después usé una visualización de un mapa para ver los datos y vi que los datos son sobre la costa del estado de california.



Análisis de K-means

K-means es un método de clustering particional que divide los datos en K clusters. Cada observación pertenece al cluster con la media más cercana.

Para saber el número mas optimo de clusters hice la prueba de codo o elbow method, que me dio el resultado de que la cantidad de K debe de ser 4



Después de definir el modelo con el numero de clusters correcto, sacamos varios valores como la inercia, etiquetas y las coordenadas de los centroides.

```
kmeans.cluster_centers_
```

```
array([[0.6307601 , 0.15647541, 0.06059012],
       [0.20198864, 0.55956895, 0.8163298 ],
       [0.64187325, 0.08581514, 1.        ],
       [0.30235266, 0.5810959 , 0.18131676]])
```

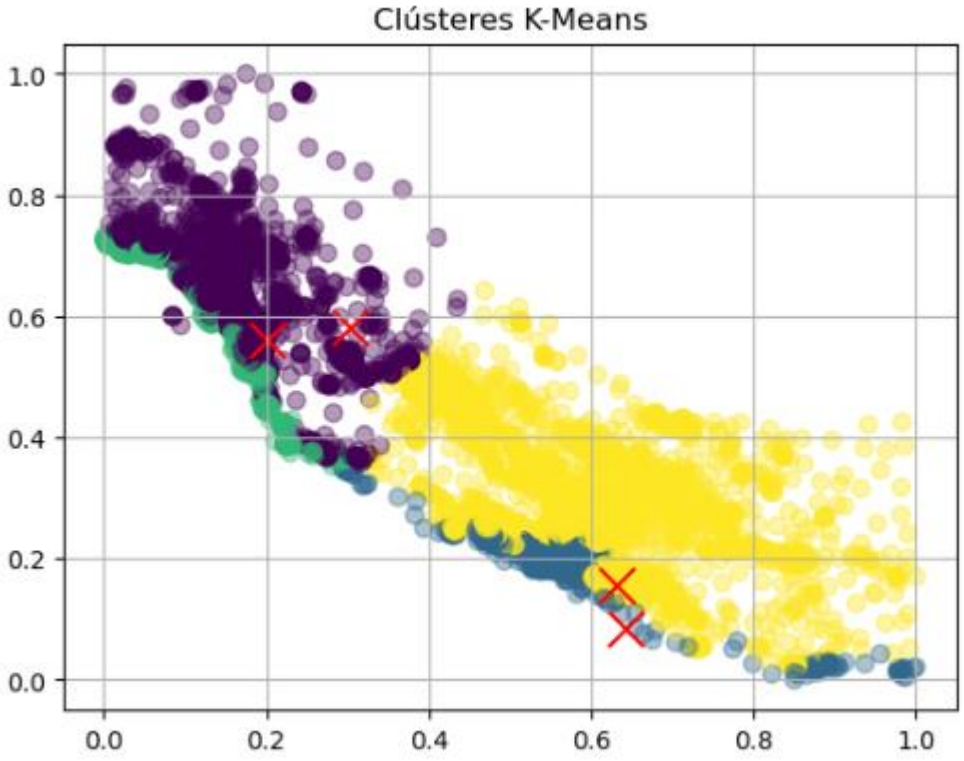
```
kmeans.inertia_
```

```
479.8641600758018
```

```
kmeans.labels_
```

```
array([1, 1, 1, ..., 3, 3, 3])
```

Con esto ya terminamos nuestro método de K-Means, solo falta observar los resultados gráficamente para ver como quedó.



El número de clusters de K = 4 me dio un accuracy del 37%.

Análisis de DBSCAN

DBSCAN es un método de clustering basado en densidad que define los clusters como áreas de alta densidad separadas por áreas de baja densidad.

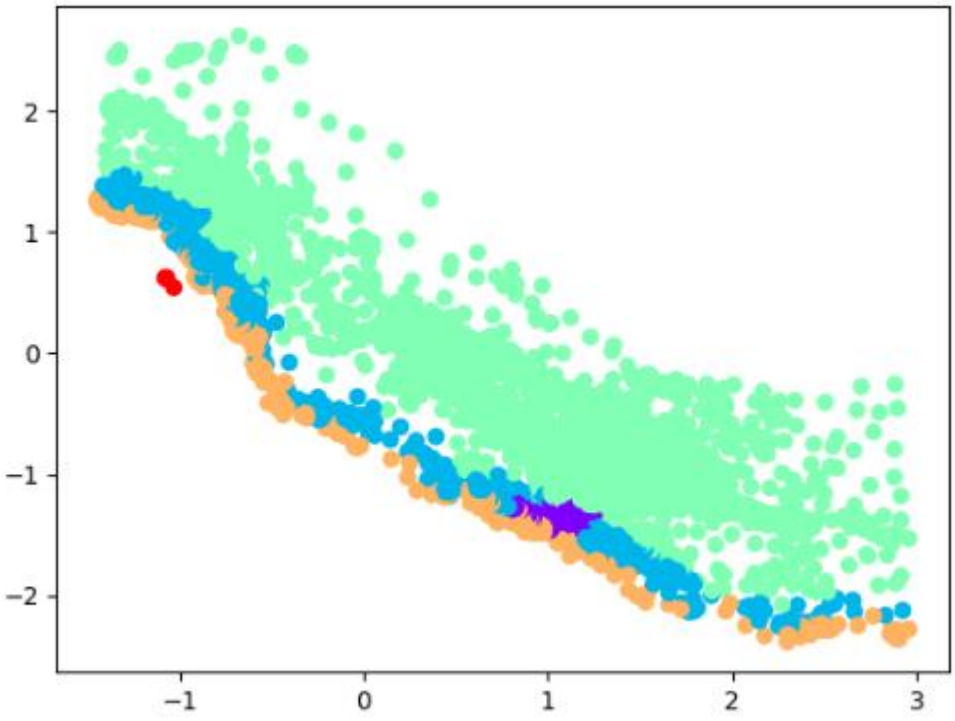
El método de DB Scan se trata mucho sobre prueba y error, no hay una forma en sí de saber si lo que haces esta siendo lo correcto, por eso debes de tener un contexto de los datos.

Las etiquetas de los cluster fueron estos:

```
labels = db.labels_
labels
```

```
array([0, 0, 0, ..., 2, 2, 2], dtype=int64)
```

El resultado de los clusters es el siguiente:



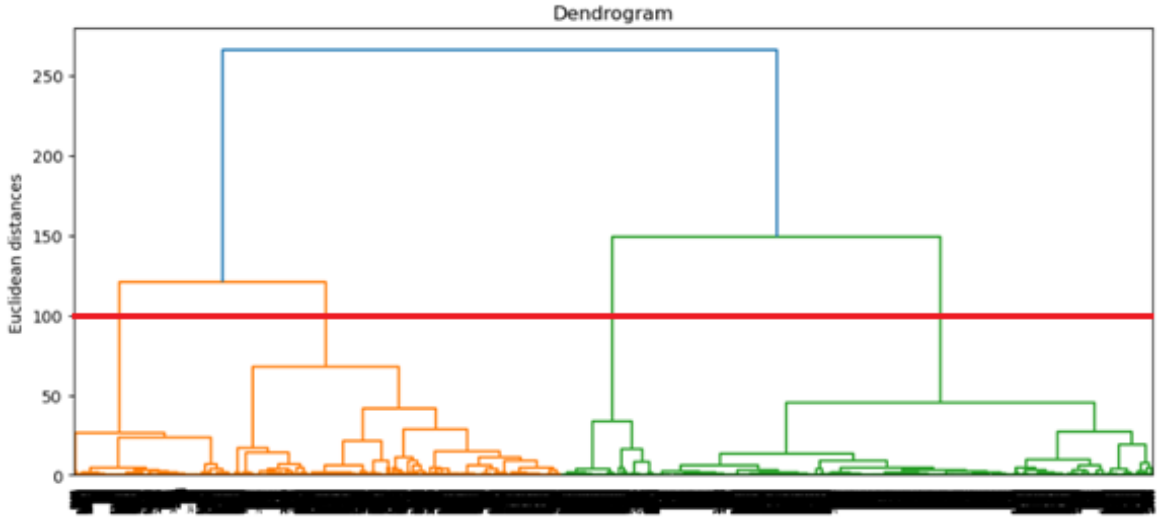
Defini el radio épsilon en 0.4 y con un numero de muestras de mínimo 3, o sea que cada muestra debía tener 3 muestras dentro de su radio para ser tomada como del mismo cluster.

Análisis de Clustering Jerárquico

El clustering jerárquico es un método de clustering que construye una jerarquía de clusters ya sea mediante un enfoque de abajo hacia arriba (aglomerativo) o de arriba hacia abajo (divisivo).

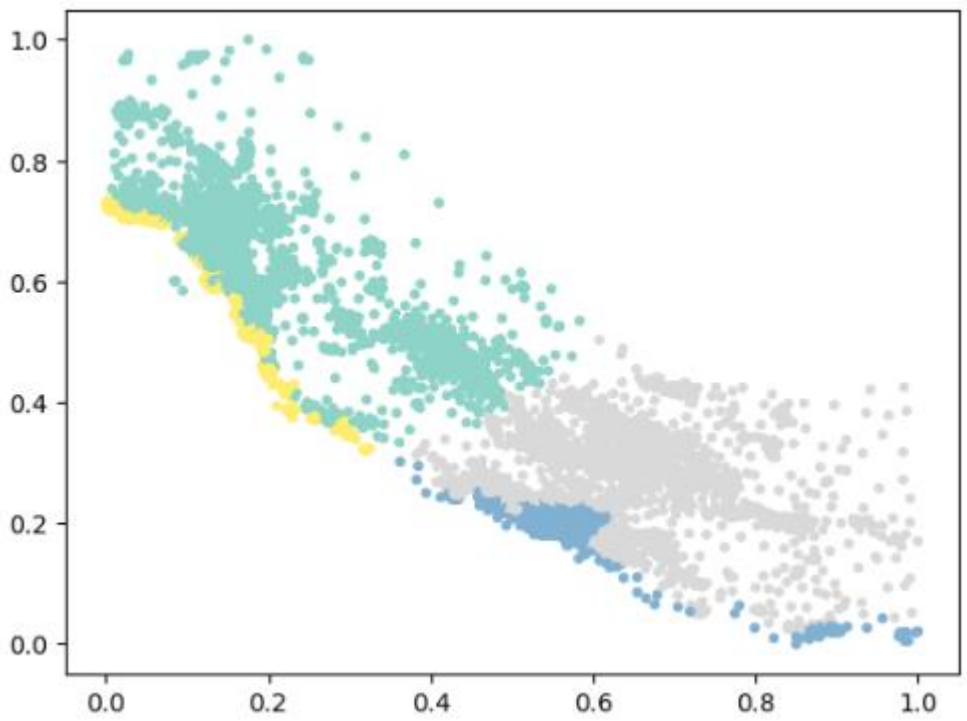
El método de enlace que utilice fue el Ward.

Para el clustering Jerárquico usamos un dendograma para saber la cantidad de clusters que debemos usar, se puede determinar el número óptimo de clusters trazando una línea horizontal que pase por la longitud más larga. El número de líneas verticales que cruza es el número óptimo de clusters.



Según lo que tengo entendido trace la línea horizontal (rojo) y me dio un número de 4 clusters (líneas verticales cruzadas).

Después de definir los clusters hice la gráfica para poder visualizarlos.



Comparación

Después de ver y comparar los diferentes métodos, veo que el método de DB Scan fue el mas acertado porque cuenta con 5 clusters, y son justo las 5 categorias de la variables “Ocean_proximity”. También se pudo haber conseguido 5 clusters en K-Means o en el clustering jerárquico, pero me quedaba corto en temas de efectividad entonces me quedé en 4 con esos.

También afecta que en el DB Scan puedes escoger los puntos que deben de estar dentro de un cluster de manera manual por así decirlo, porque tu mueves el radio y el mínimo de muestras a lo que tu entiendes de los datos.