Proyecto 1

Ecobici en CDMX

Elements of Machine Learning

Daniel Hidalgo y Denisse Bolaños

2019
Enero - Diciembre

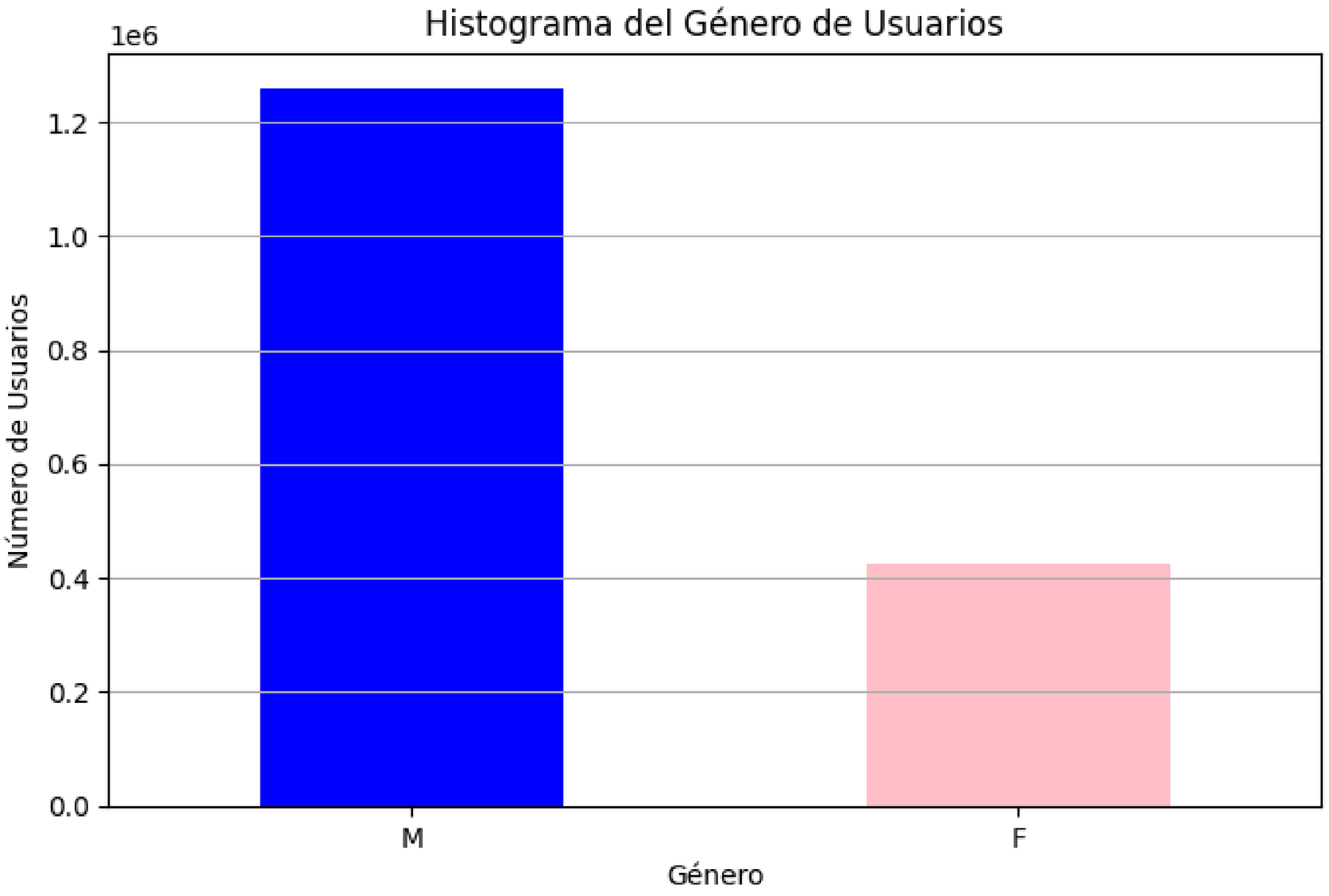
Muestras del 20% de cada CSV

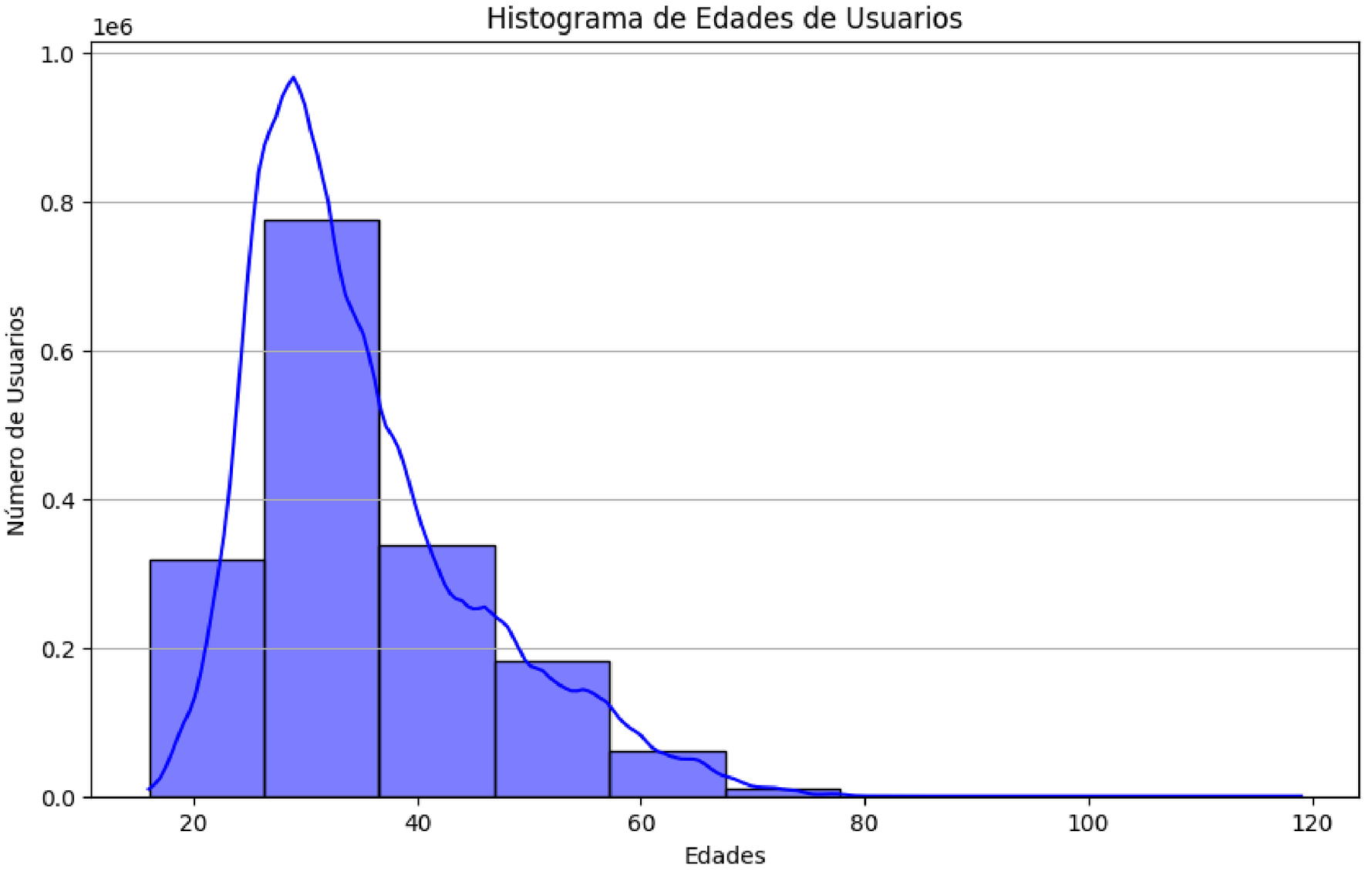
Dataset

- *+1,600,000 registros*
- *Edad de Usuario, Género de Usuario, Bici, Ciclo Estación Retiro, Fecha Retiro, Hora Retiro, Ciclo Estación Arribo, Fecha Arribo, Hora Arribo.*
- *Stations: Id, nombre, latitud y longitud.*

EDA

1. *No habían datos nulos*
2. *Conversión de Horas y Fechas a un formato manejable*
3. *Exploración de distribuciones (Edad y Género)*
4. *Manejo de Outliers en Edad*


Histograma del Género de Usuarios
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Número de Usuarios
M
F
Género

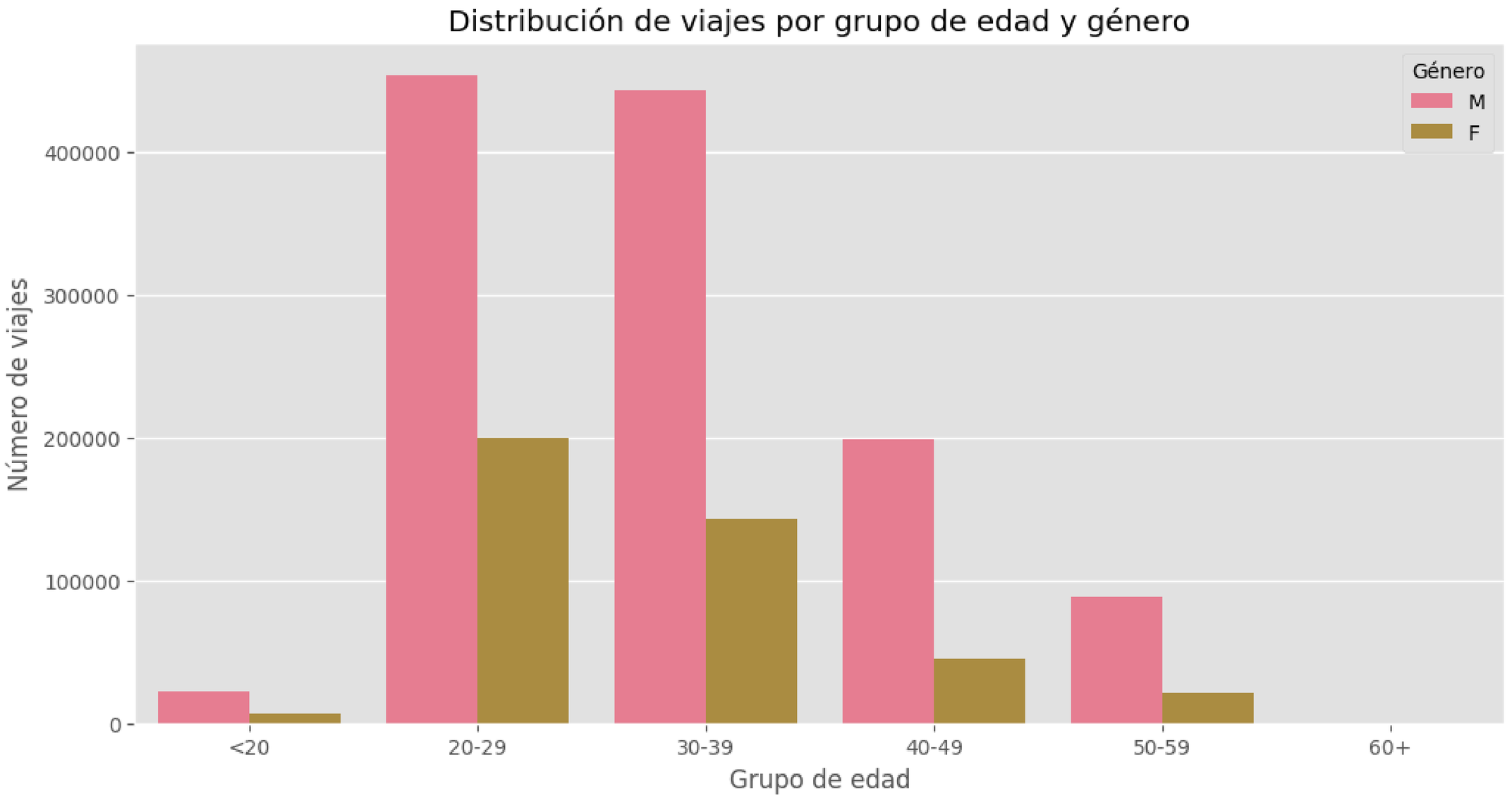

Histograma de Edades de Usuarios
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Número de Usuarios
20
40
60
80
100
120
Edades

Estrategia

Uso de rango intercuartílico para definir límites y detectar outliers.

- Distribución de Viajes por Edad y Género
- Variabilidad Y Distribución entre Grupos
- Estaciones con Tiempos Mayores de viaje

Comportamiento de Usuario


Distribución de viajes por grupo de edad y género
Género
M
F
Número de viajes
400000
300000
200000
100000
0
<20
20-29
30-39
40-49
50-59
60+
Grupo de edad

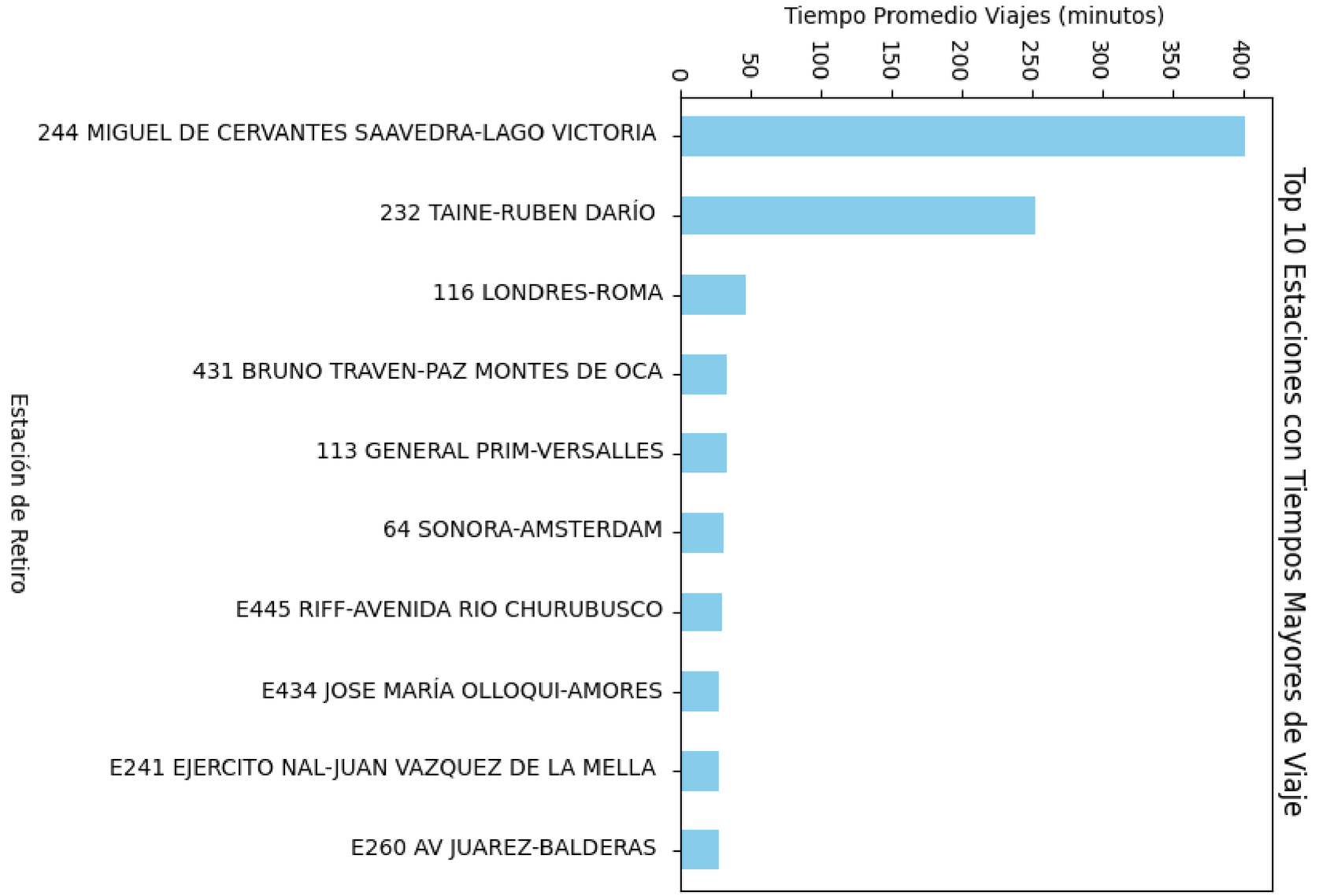
Distribución

chi2=17699.97
La distribución de género no es uniforme entre los grupos de edad.

```
Tabla de contingencia edad-género:
Genero_Usuario        F       M
grupo_edad
<20                6763   22465
20-29            199725  453633
30-39            142986  442802
40-49             45322  199102
50-59             21082   88480
```

Variabilidad

ANOVA: F = 0.39
Sí hay diferencias significativas en la duración promedio del viaje entre los grupos de edad.

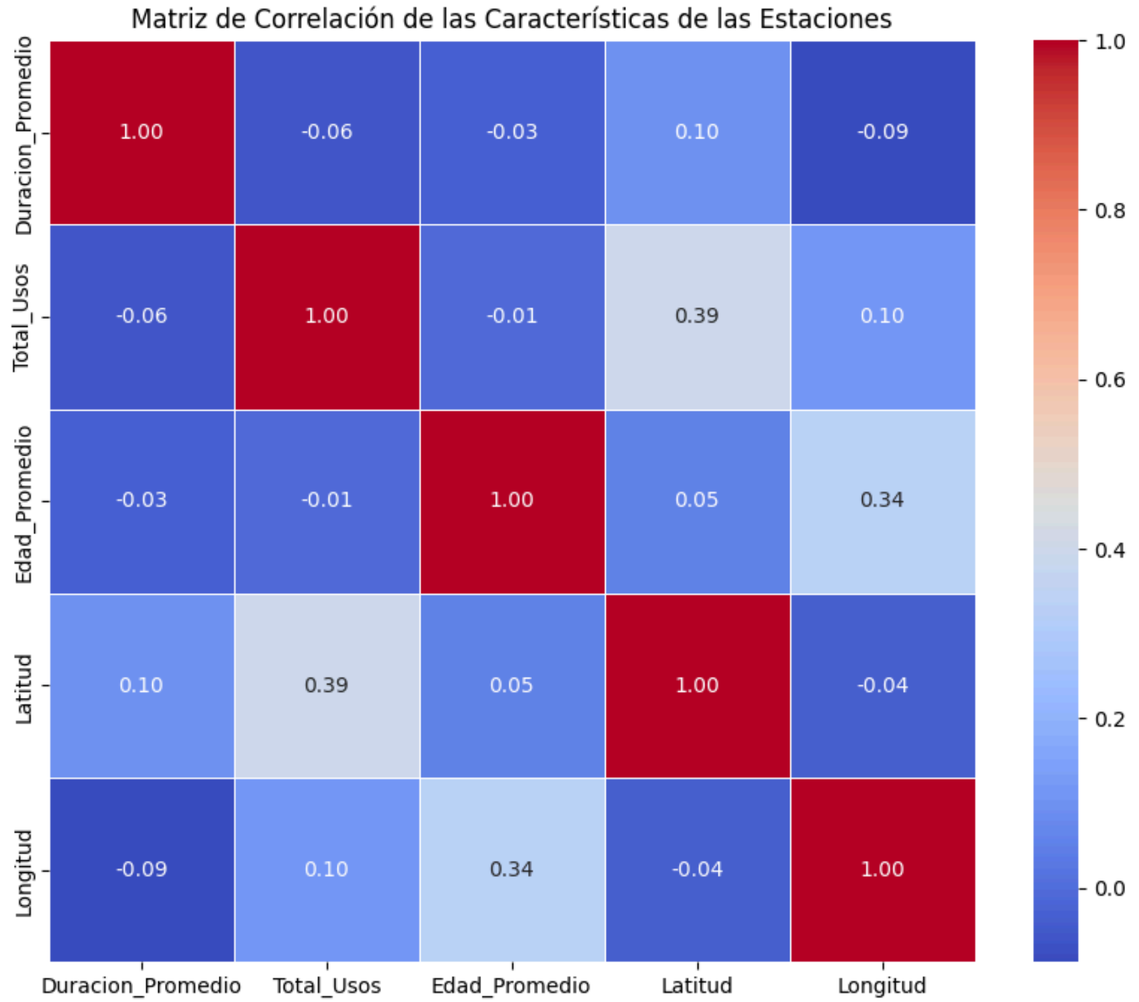

Top 10 Estaciones con Tiempos Mayores de Viaje
Tiempo Promedio Viajes (minutos)
0
50
100
150
200
250
300
350
400
Estación de Retiro
244 MIGUEL DE CERVANTES SAAVEDRA-LAGO VICTORIA
232 TAINE-RUBEN DARÍO
116 LONDRES-ROMA
431 BRUNO TRAVEN-PAZ MONTES DE OCA
113 GENERAL PRIM-VERSALLES
64 SONORA-AMSTERDAM
E445 RIFF-AVENIDA RIO CHURUBUSCO
E434 JOSE MARÍA OLLOQUI-AMORES
E241 EJERCITO NAL-JUAN VAZQUEZ DE LA MELLA
E260 AV JUAREZ-BALDERAS

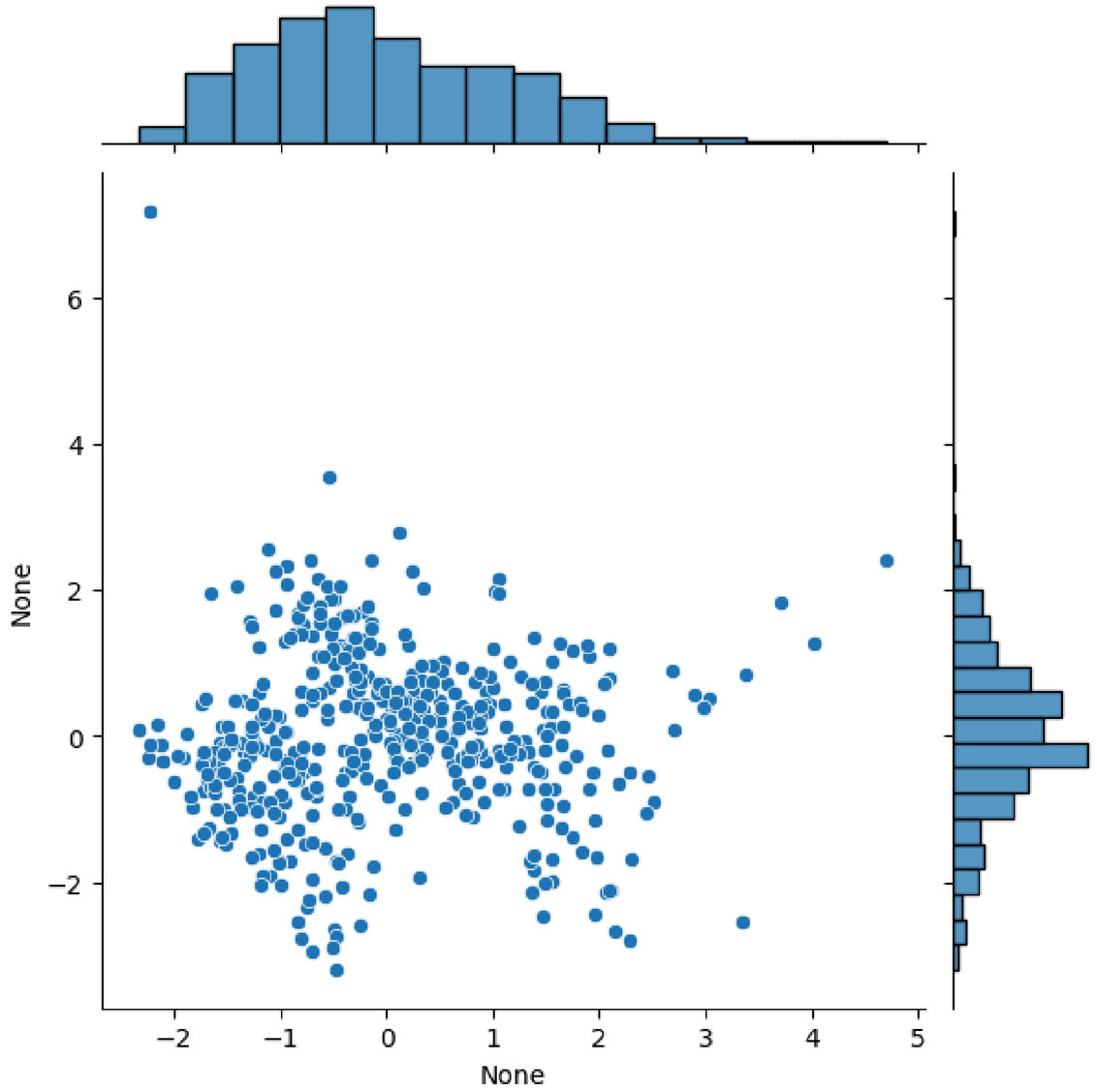
Análisis multivariado

- *PCA*
- *MDS*
- *Mapa*

Feature Matrix

- Duracion_Promedio
- Total_Usos
- Edad_Promedio
- Latitud
- Longitud





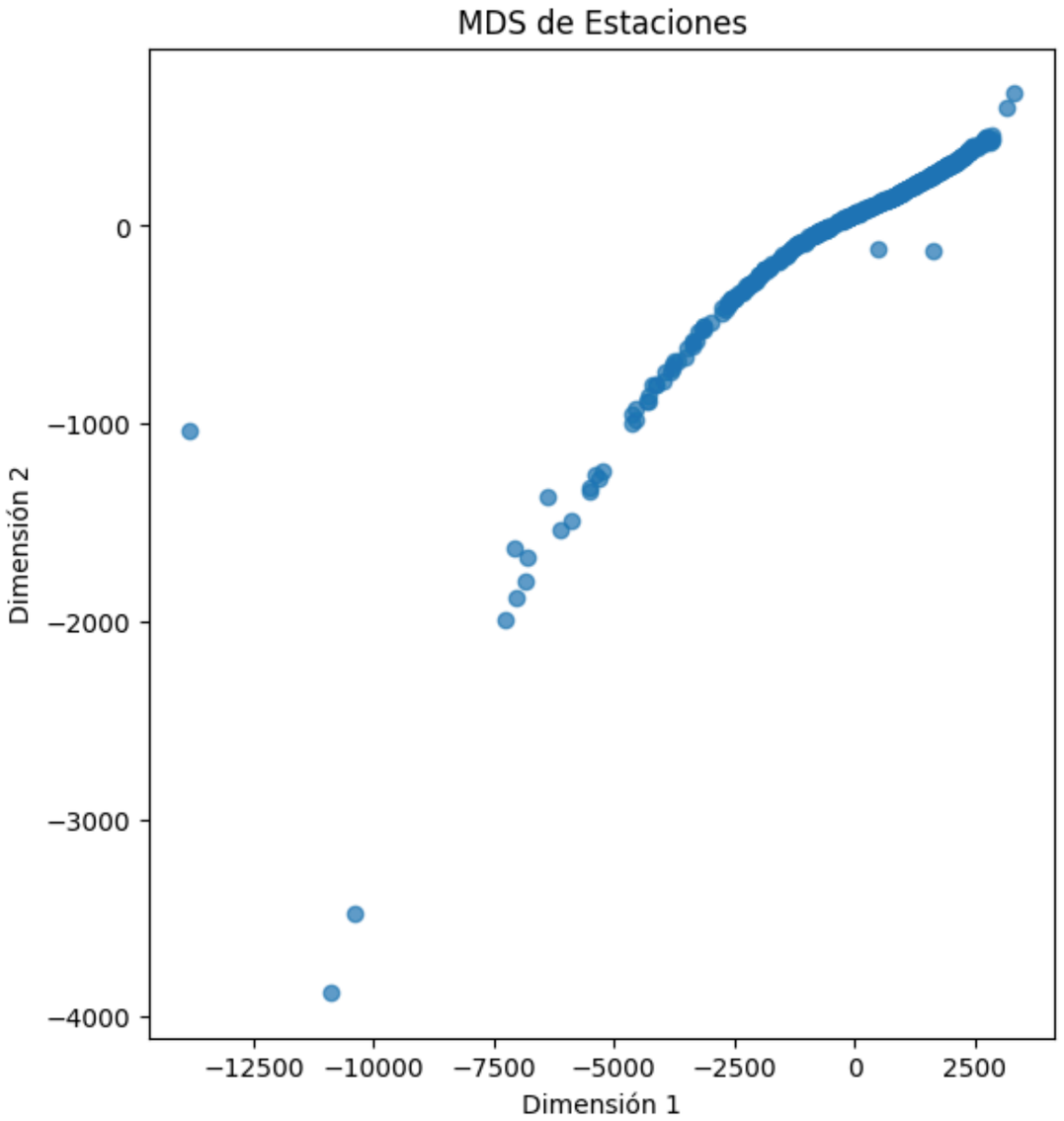
Cargas PCA

CP1:

- Total de usos

CP2:

- Duración promedio



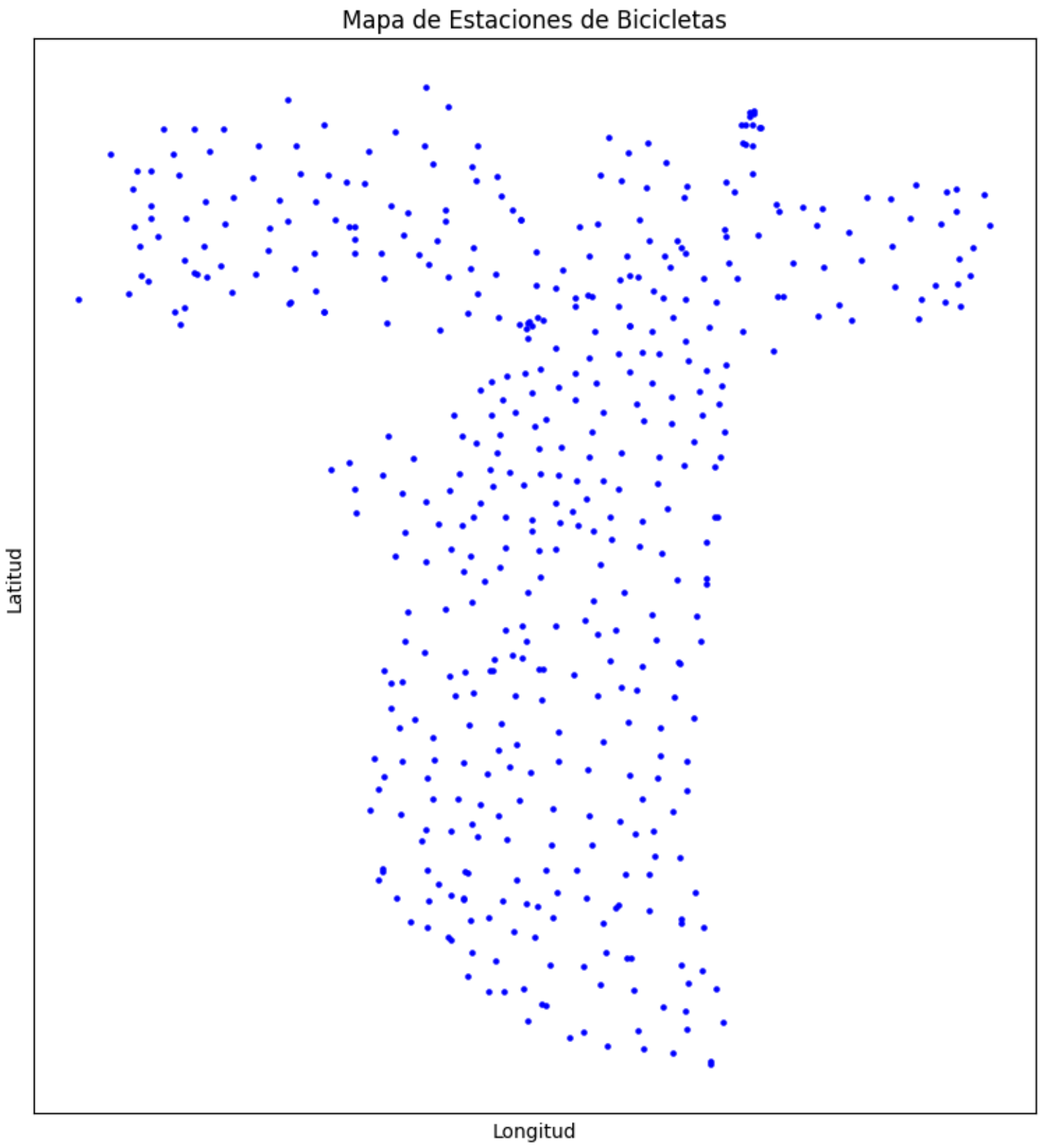
Dimensiones

D1:

- Total de usos / Latitud
- Más al norte = Más uso

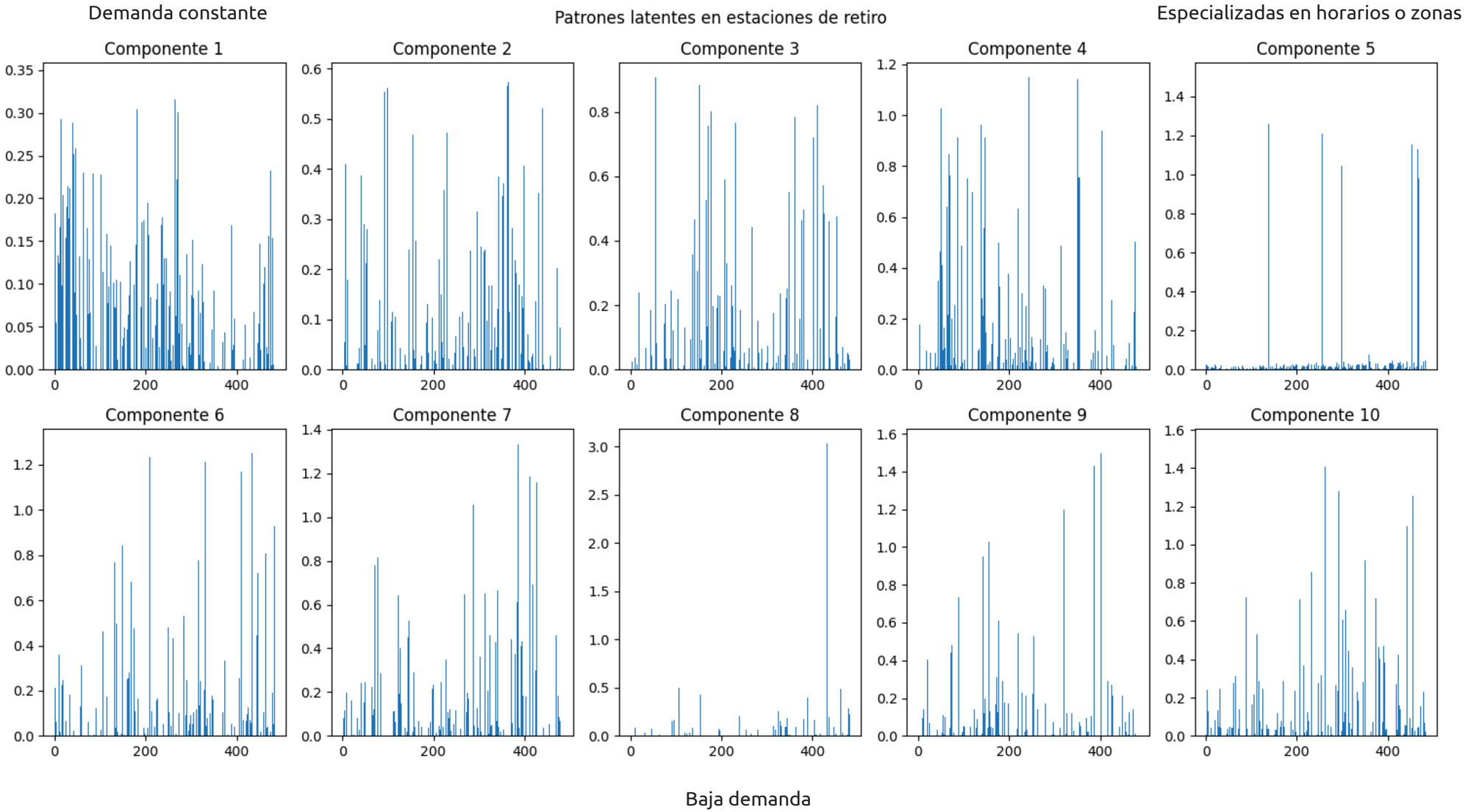
D2:

- Edad promedio / Longitud
- Diferentes áreas (este-oeste) podrían atraer grupos distintos.


Mapa de Estaciones de Bicicletas
Latitud
Longitud

- Recomendaciones histórico usuario
- Recomendación por distancia

Modelo de Recomendación


Demanda constante
Patrones latentes en estaciones de retiro
Especializadas en horarios o zonas
Componente 1
Componente 2
Componente 3
Componente 4
Componente 5
Componente 6
Componente 7
Componente 8
Componente 9
Componente 10
Baja demanda

```
Estaciones recomendadas para retiro:
      id        lat        lon                                  name
86    87  19.431820 -99.139740         87 GANTE-VENUSTIANO CARRANZA
101  102  19.428210 -99.139490                102 ECHEVESTE-BOLIVAR
103  104  19.427059 -99.137116  104 SAN JERONIMO-ISABEL LA CATOLICA
111  112  19.432819 -99.151734     112 AV. MORELOS-ABRAHAM GONZALEZ
401  402  19.370945 -99.158722            402 UXMAL-MUNICIPIO LIBRE
```

```
Últimas 10 estaciones de retiro del usuario:
      id        lat        lon                                        name
17    18  19.428880 -99.164176                       18 REFORMA-RIO RHIN
31    32  19.422705 -99.169922                        32 LONDRES-SEVILLA
83    84  19.407020 -99.169587                 84 CHILPANCINGO-TLAXCALA
84    85  19.434250 -99.162508                 85 ROSAS MORENO-SULLIVAN
97    98  19.430267 -99.148610           98 EMILIO DONDE-AV. BALDERAS
220  221  19.431799 -99.201565  221 EMILIO CASTELAR-PRESIDENTE MASARYK
```

Problemas

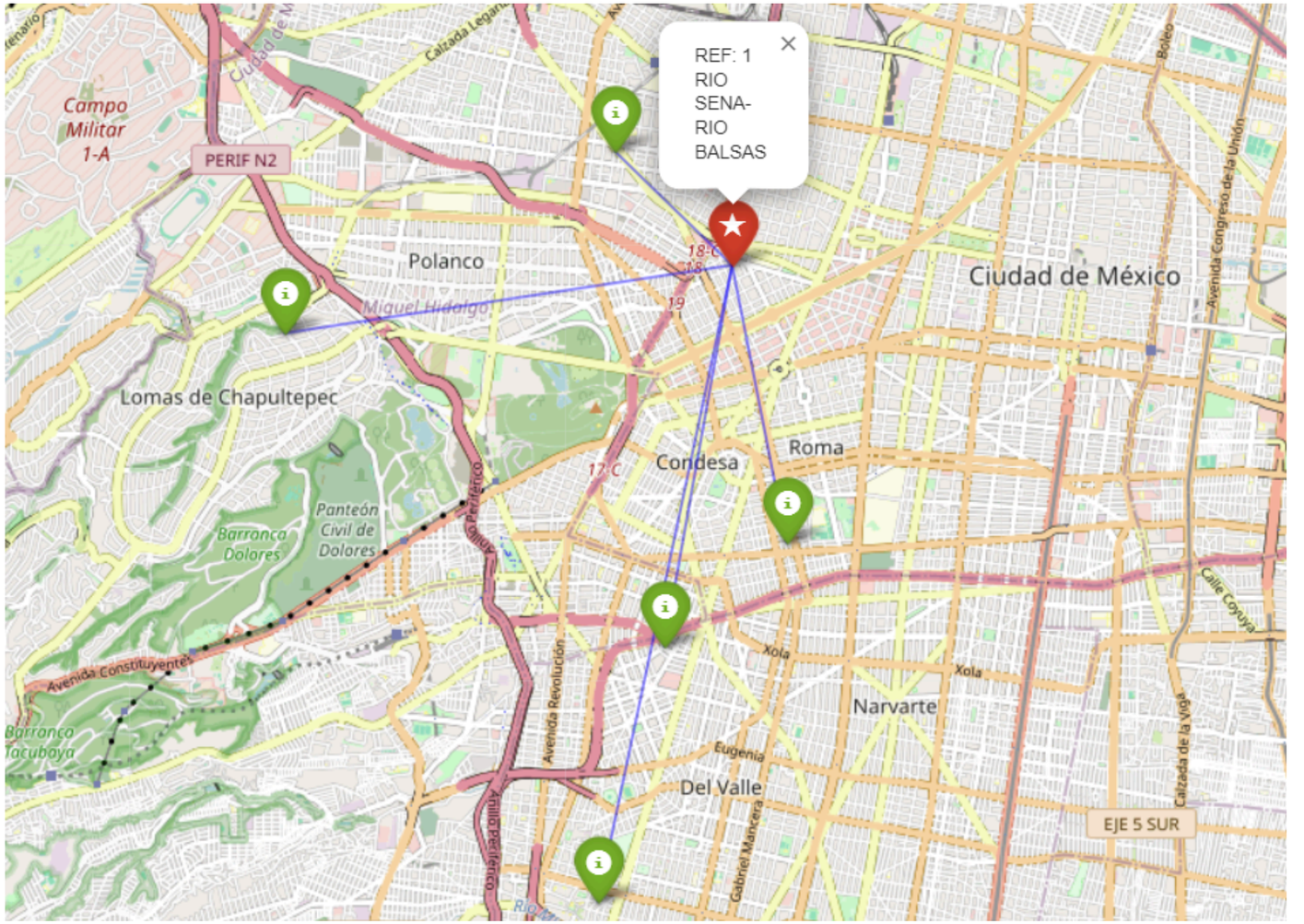
- **ID de usuario poco robusto:** Ya no se toma en cuenta bicicleta
- **Normalización inadecuada:** StandardScaler puede producir valores negativos

```
Estaciones recomendadas para arribo:
      id        lat        lon  \
177  178  19.406833 -99.180457
274  275  19.441062 -99.153164
323  324  19.387247 -99.165655
414  415  19.371498 -99.176020
419  420  19.369723 -99.179640

                                                  name
177                    178 TAMAULIPAS-FRANCISCO MURGUÍA
274  275 HÉROES FERROCARRILEROS-AV. INSURGENTES NORTE
323                           324 AMORES-AVENIDA EUGENIA
414                         415 SAN FRANCISCO-PARROQUÍA
419                           420 MARÍA LUZ BRINGAS-OSO
```

```
Últimas 10 estaciones de arribo del usuario:
      id        lat        lon                                    name
5      6  19.430922 -99.166959                     6 RIO PANUCO-RIO SENA
17    18  19.428880 -99.164176                       18 REFORMA-RIO RHIN
23    24  19.426470 -99.168220                       24 REFORMA-VARSOVIA
60    61  19.413742 -99.169525                  61 AVENIDA MEXICO-SONORA
263  264  19.440985 -99.152935  264 HÉROES FERROCARRILEROS-AV. CENTRAL
464  465  19.445039 -99.194078        465 LAGUNA DE GINEBRA-LAGO WETTER
```

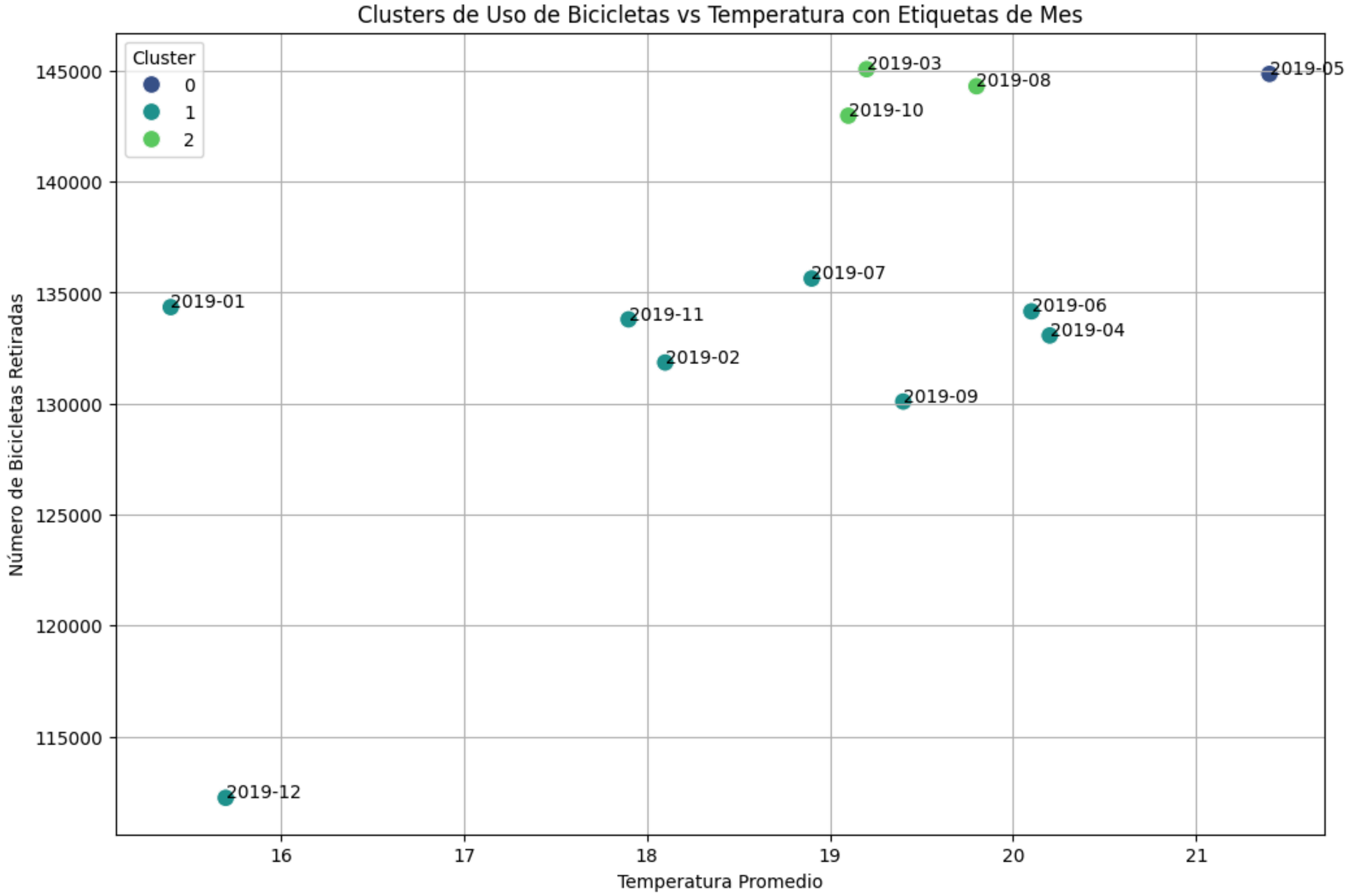
```
Usuario de ejemplo: M_36

Estaciones usadas anteriormente:
         Ciclo_Estacion_Retiro Fecha_Retiro
1565486                   238   2019-12-31
1614321                    11   2019-12-31
1602029                   209   2019-12-31
1601810                    66   2019-12-31
1529667                    84   2019-12-31

Estaciones recomendadas para retiro:
      id         lat         lon                  name
155  156  19.407121 -99.162202  156 TEPIC-AMENALCO
```

```
=== Recomendaciones para usuario: F_23 ===
Género: F
Edad: 23
Total viajes: 11649

Estación de referencia:
ID: 1
Nombre: 1 RIO SENA-RIO BALSAS
Coordenadas: (19.433571, -99.167809)

Recomendaciones cercanas:

1. ID: 479
   Nombre: E479 LAGO MURITZ-AV. MARINA NACIONAL
   Coordenadas: (19.444433, -99.179664)
   Distancia desde referencia: 1.78 km

2. ID: 156
   Nombre: 156 TEPIC-AMENALCO
   Coordenadas: (19.407121, -99.162202)
   Distancia desde referencia: 3.00 km

3. ID: 289
   Nombre: 289 CHICAGO-YOSEMITE
   Coordenadas: (19.397308, -99.174548)
   Distancia desde referencia: 4.09 km
```



Análisis con clima

- *API*
- *Data histórica*


Clusters de Uso de Bicicletas vs Temperatura con Etiquetas de Mes
Cluster
0
1
2
Número de Bicicletas Retiradas
Temperatura Promedio
145000
140000
135000
130000
125000
120000
115000
16
17
18
19
20
21
2019-03
2019-08
2019-05
2019-10
2019-07
2019-01
2019-11
2019-06
2019-04
2019-02
2019-09
2019-12

- Podría ser de valor agregar data más detallas en relación a clima, infraestructura o demografía.
- Búsqueda de features que tengan más impacto para modelos de recomendación.
- Investigación profunda del área donde es la data (CDMX).
- Sería útil que el sistema cree ID de usuario.

Recomendaciones

¿Dudas o comentarios?

Gracias