Coderhouse – Proyecto Final - Data Science - Comision 49150

Consumo de Sustancias Argentina 2022

Autor: Daniel Perco

Abstracto con Motivación y Audiencia

1.1 Motivación

Este proyecto busca aplicar técnicas avanzadas de machine learning para analizar los datos de la encuesta realizada por el INDEC sobre el consumo de sustancias como tabaco, alcohol, medicamentos, marihuana y cocaína en Argentina. El objetivo es identificar patrones de consumo, factores de riesgo y desarrollar modelos predictivos que puedan informar políticas públicas y estrategias de intervención.

1.2 Audiencia

La audiencia a la que está dirigido este informe incluye:

- Ministerio de Salud y Entidades públicas de Salud.
- Organizaciones no gubernamentales de Ayuda Social.
- Instituciones de Educación e Investigación.
- Gobiernos Locales y Municipales.
- Departamentos de Policía y Justicia.
- Organizaciones Internacionales de Salud y de Políticas Antidrogas.

1.3 Abstract

El dataset utilizado proviene del INDEC Argentina y contiene datos de una encuesta realizada en 2022 sobre el consumo de sustancias/drogas que pueden afectar la salud humana, incluyendo alcohol, medicamentos, marihuana y cocaína. Este análisis busca mostrar las posibles causas, herramientas de

A	B	C	D	E	F	G	H
ID_PER	WPER	CANT_MIEMBROS_HOGAR	CANT_PERSONAS0A17	CLIMA_EDUCATIVO	J_SEXO	J_EDAD	J_NIVEL_EDUCATIVO
336578	124	1	0	2	1	27	4
305909	781	1	0	3	1	33	6
358892	34193	4	1	2	1	43	4
342664	968	1	0	3	2	51	6
394688	11509	2	0	3	2	59	5
370155	13159	2	0	3	2	44	5
308312	1440	1					
354770	258	1					
394258	14378	4					
390628	315	3					
309538	18560	3					

W	X	Y	Z	AA	AB	AC	AD	AE
SA_07__4	SA_07__5	SA_07__6	SA_07__99	consumo_de_med_sin_receta	AL_01	AL_02	edad_consumo_bebidas_alcohólicas	consumo_alcohol
0	0	0	0	3	1	3	17	1
0	0	0	0	3	1	3	12	1
							15	1
							30	1
							15	1
							14	1
							20	1
							15	1
							14	1
							16	1
							15	1
							18	1
							16	1
							15	1
							17	1
							998	1
							50	1
							998	1
							15	1
							998	1
							998	1

BZ	CL	CM	CN	CO	CP	CQ	CR	CS
_99	riesgos_consumo_alcohol	AL_28	AL_29	creencias_alcohol_accidentes	hol_proble	as_alcoho	creencias_alcohol_dificultades_laborales	s_alcohol_	as_alco
0	0	2	1	1	1	1	1	1	
0	0	2	1	2	1	2	1	1	
0	0	1	1	2	2	2	2	2	
0	0	2	1	1	1	1	1	1	
0	1	2	1	2	2	3	3	2	
0	0	2	2	1	2	2	1	1	
0	0	2	1	1	2	2	2	1	
0	0	2	1	1	1	2	2	1	
0	0	2	1	1	1	1	1	1	
0	0	2	3	2	2	2	2	2	
0	0	2	3	1	1	2	1	1	
0	0	2	3	1	1	2	1	2	
0	0	2	3	1	1	1	1	1	
0	0	1	2	1	1	1	1	1	
0	0	2	1	1	1	1	1	1	
0	98	2	3	98	98	2	98	98	
0	0	2	2	1	1	1	1	1	
0	0	2	3	2	2	2	2	1	
0	0	1	2	1	2	2	2	1	
0	98	2	3	1	1	1	1	1	
0	0	2	3	1	1	1	1	1	

Sheet1

2. Preguntas/Hipótesis

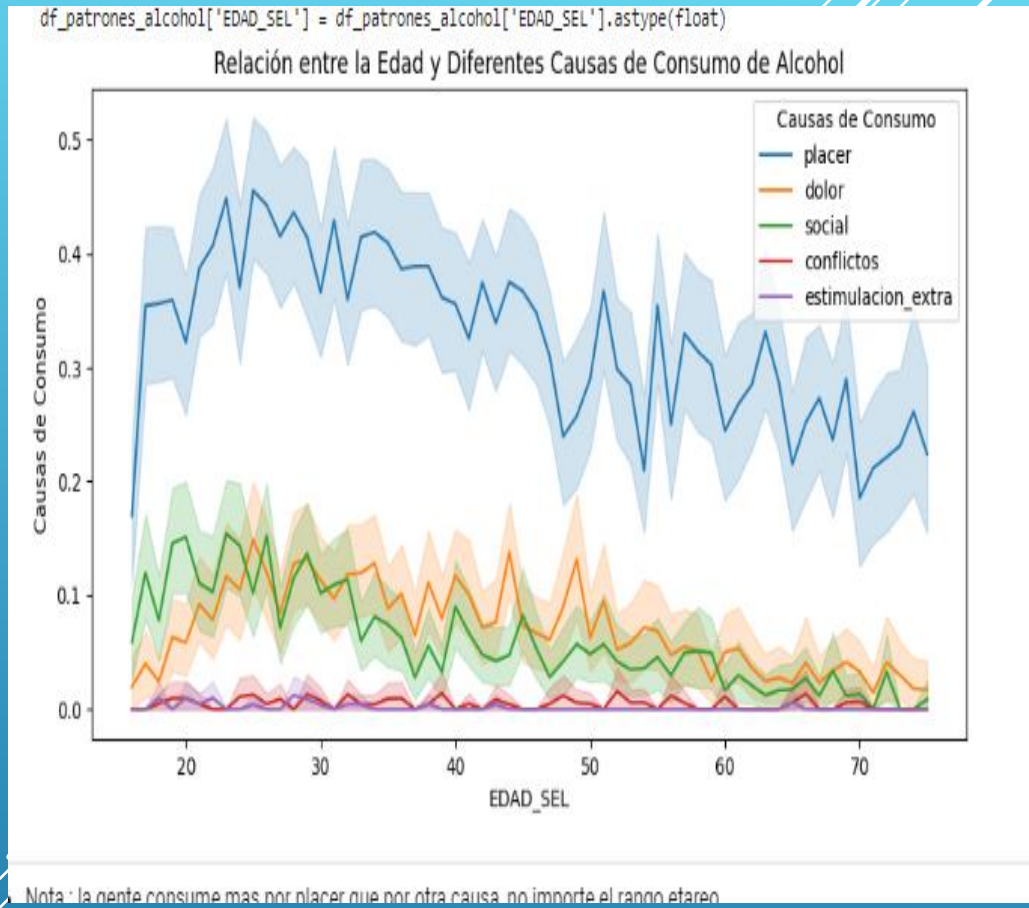
2.1 Definición de Objetivo

2.1.1 Objetivo Principal

"¿Cuáles son los principales factores predictivos del consumo problemático de sustancias en Argentina, y cómo se pueden utilizar estos factores para predecir las tendencias futuras y los posibles impactos en la salud pública y la sociedad?“

2.1.2 Objetivos Secundarios

- Identificar Tendencias y Patrones.
- Evaluar Factores de Riesgo y Causalidad.
- Impacto en Salud y Sociedad.
- Desarrollo de Modelos Predictivos.
- Informar Políticas Públicas y Estrategias de Intervención.

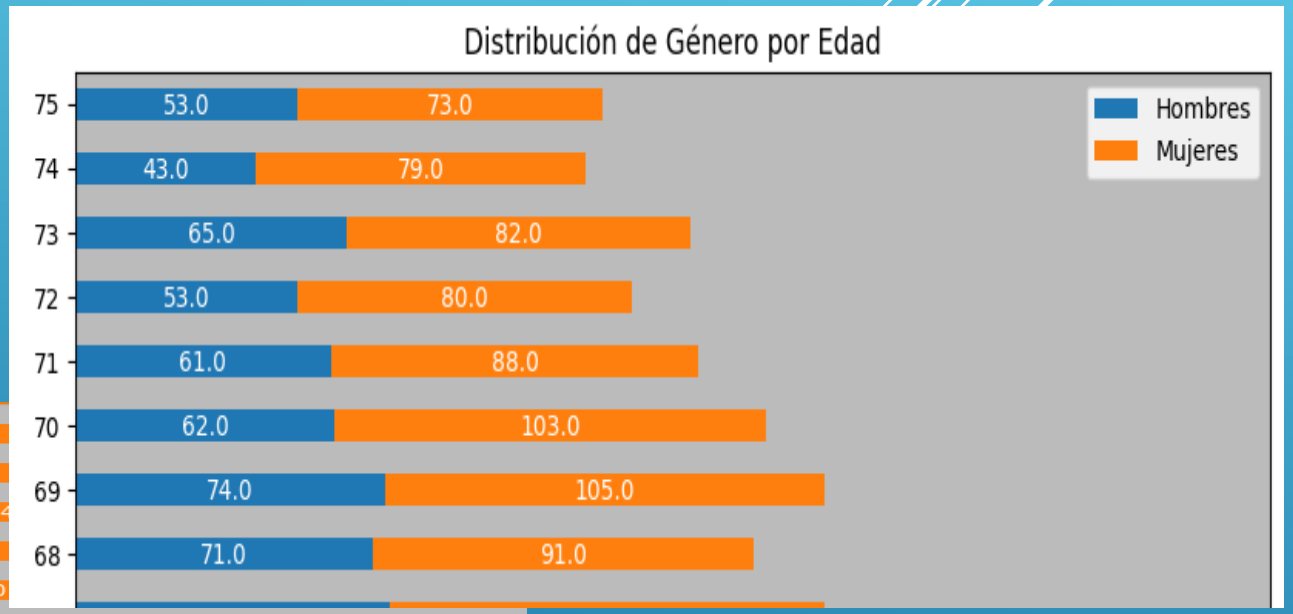


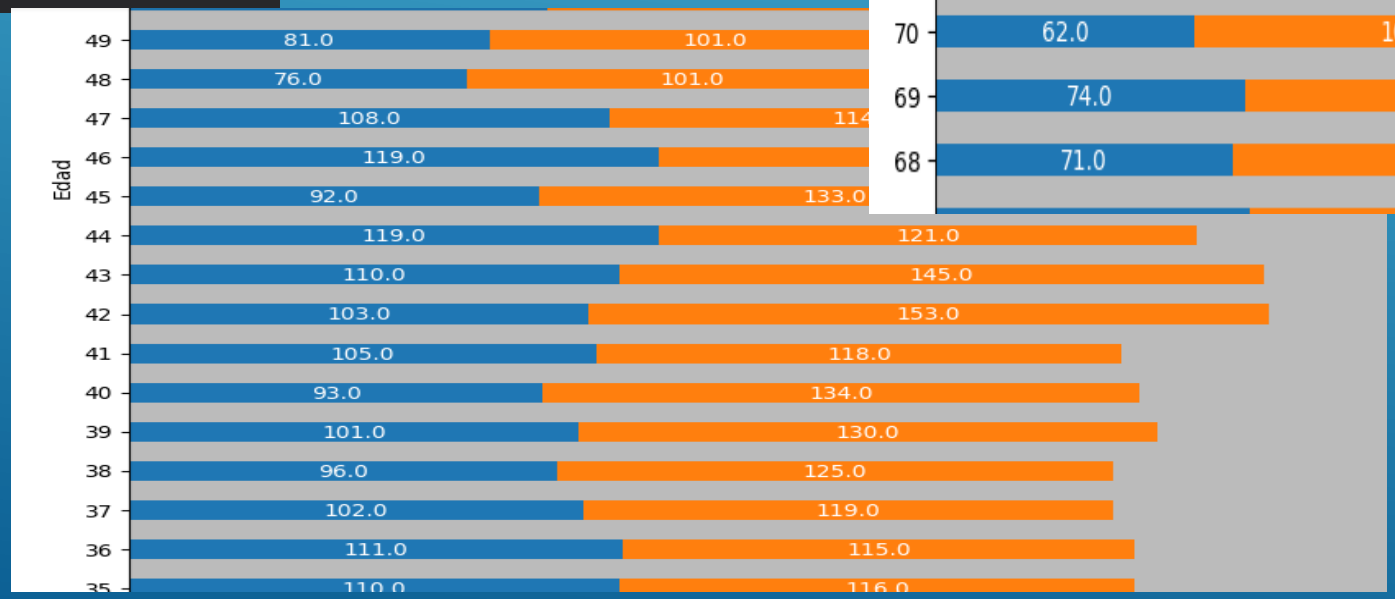
. Contexto Comercial

3.1 Problema Comercial

El análisis de los datos de consumo de sustancias puede responder a varias incógnitas y proporcionar resultados valiosos, tales como:

- Patrones de Consumo.
- Evolución del Consumo.
- Factores de Riesgo.
- Impacto en Salud y Sociedad.
- Intervenciones Efectivas.

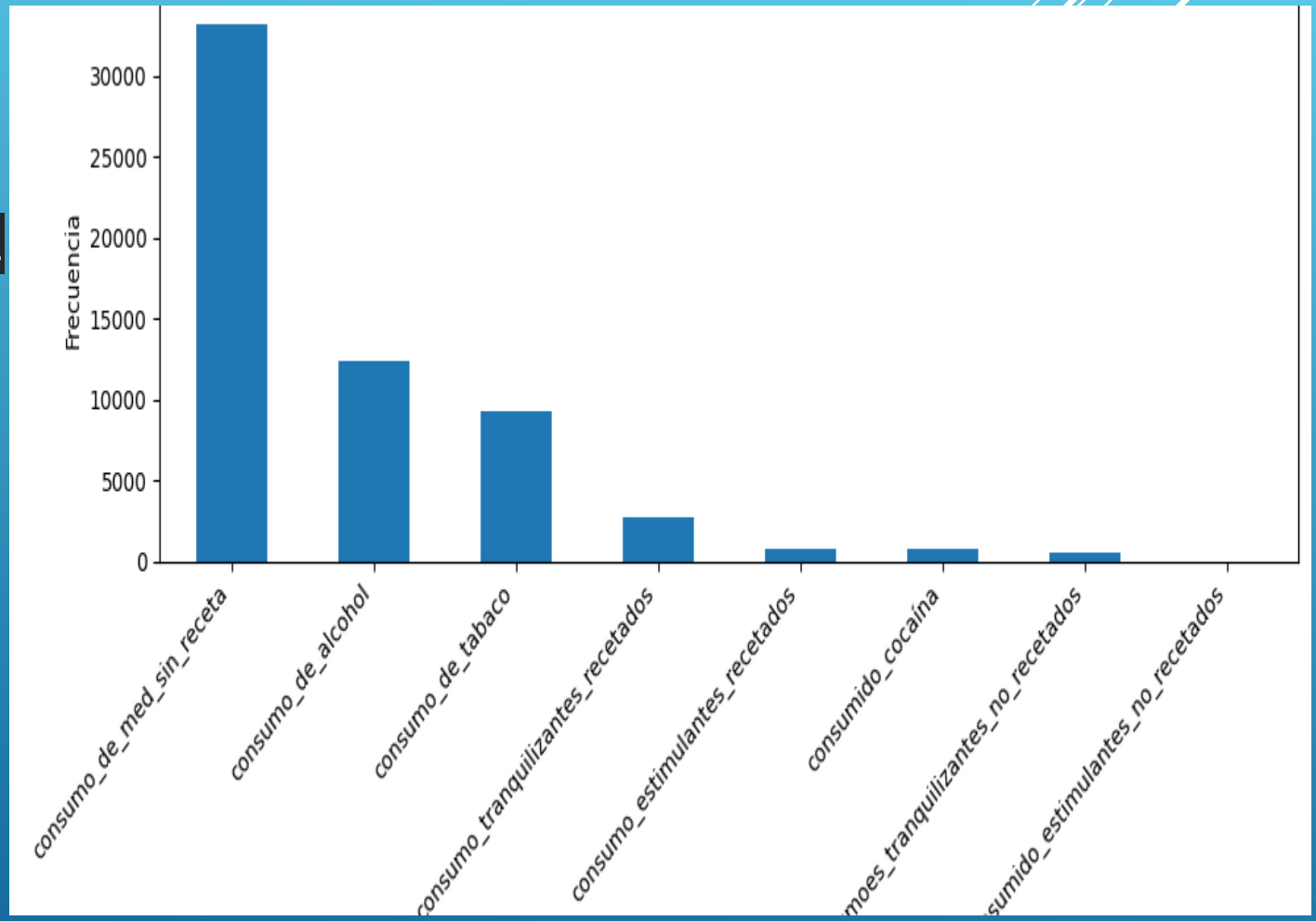




4. Contexto Analítico

4.1 Hipótesis del Problema Analítico

- Patrones de Consumo.
- Factores de Riesgo.
- Tendencias Temporales.
- Impacto en la Salud y la Sociedad.
- Efectividad de Intervenciones.
- Policonsumo.
- Accesibilidad y Disponibilidad.
- Prevención y Educación.



6. Respuestas a Preguntas e Hipótesis

6.1 Patrones de Consumo

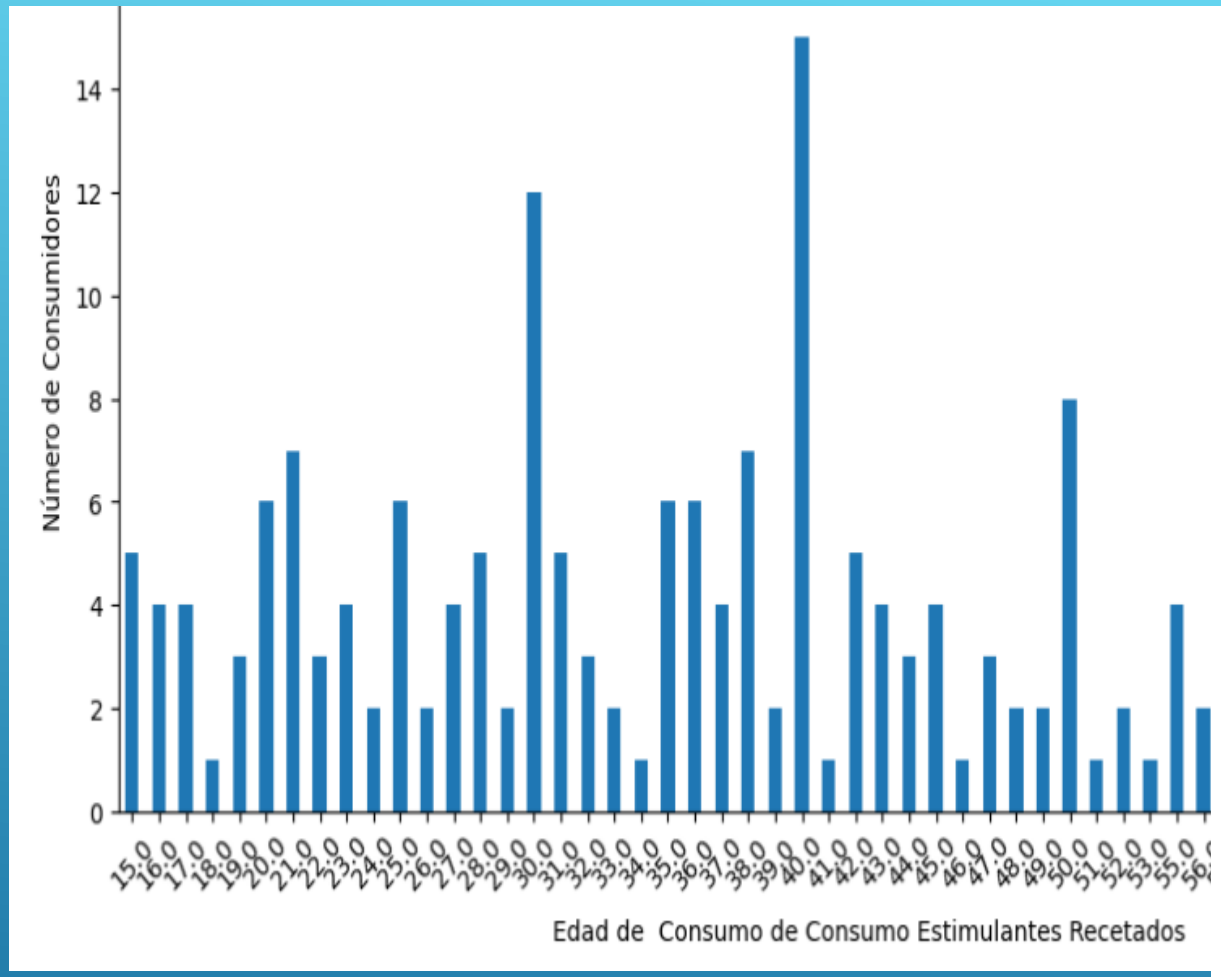
Pregunta: ¿Cuáles son los patrones de consumo? Respuesta: Se identificaron patrones de consumo específicos para diferentes sustancias.

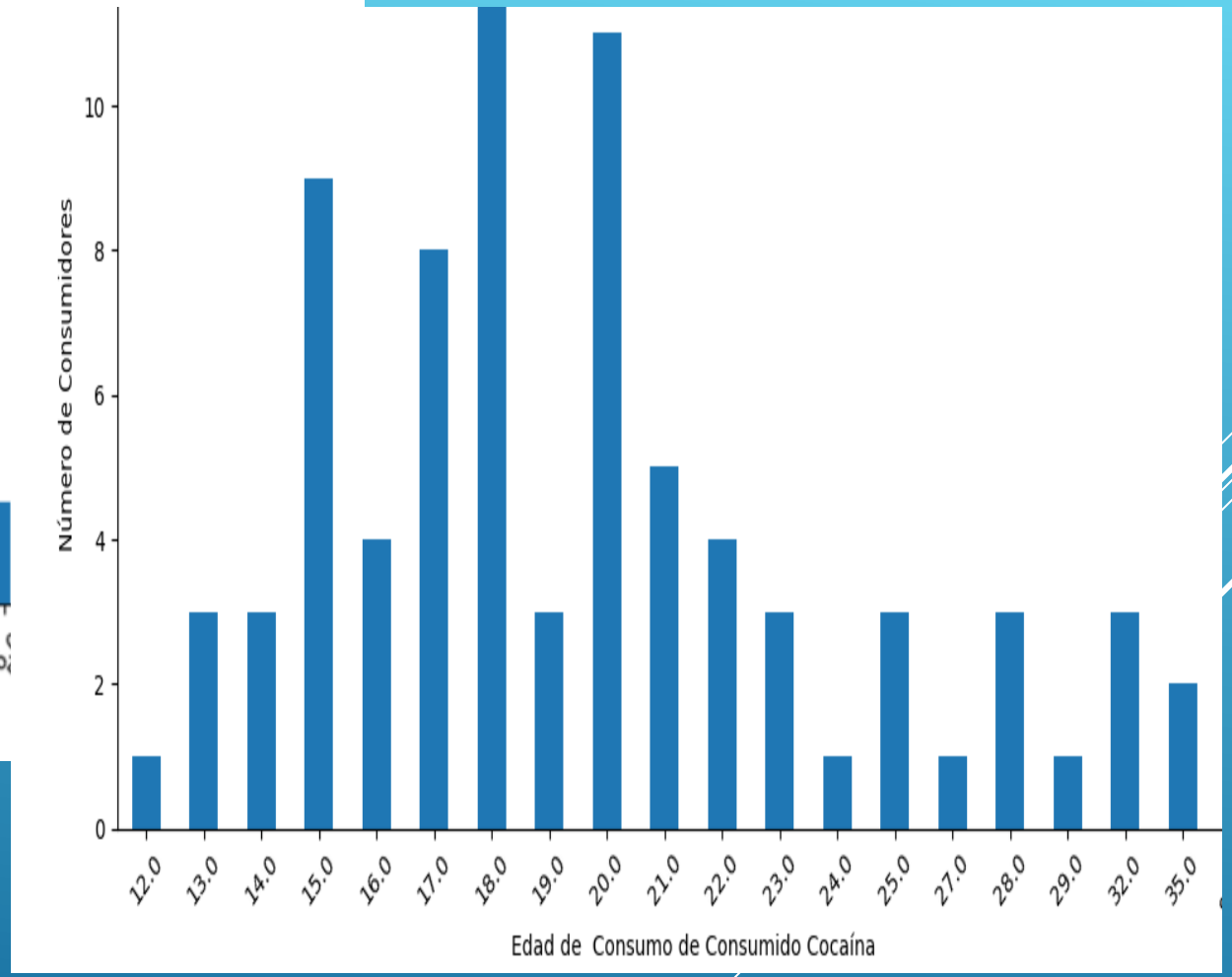
6.2 Factores Demográficos

Pregunta: ¿Qué sustancias se consumen más y por qué grupos demográficos? Respuesta: El análisis mostró que el consumo de marihuana es más alto entre los hombres jóvenes (16-24 años).

6.3 Evolución del Consumo

Pregunta: ¿Cómo ha evolucionado el consumo de sustancias a lo largo del tiempo? Respuesta: Se observó una tendencia creciente en el consumo de marihuana y una disminución en el consumo de tabaco.


Número de Consumidores
0
2
4
6
8
10
12
14
15.0 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0 26.0 27.0 28.0 29.0 30.0 31.0 32.0 33.0 34.0 35.0 36.0 37.0 38.0 39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0 49.0 50.0 51.0 52.0 53.0 55.0
Edad de Consumo de Consumo Estimulantes Recetados


Número de Consumidores
0
2
4
6
8
10
12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0 27.0 28.0 29.0 32.0 35.0
Edad de Consumo de Consumido Cocaína

EDA

5. Análisis Exploratorio de Datos (EDA)

5.1 Carga y Limpieza de Datos

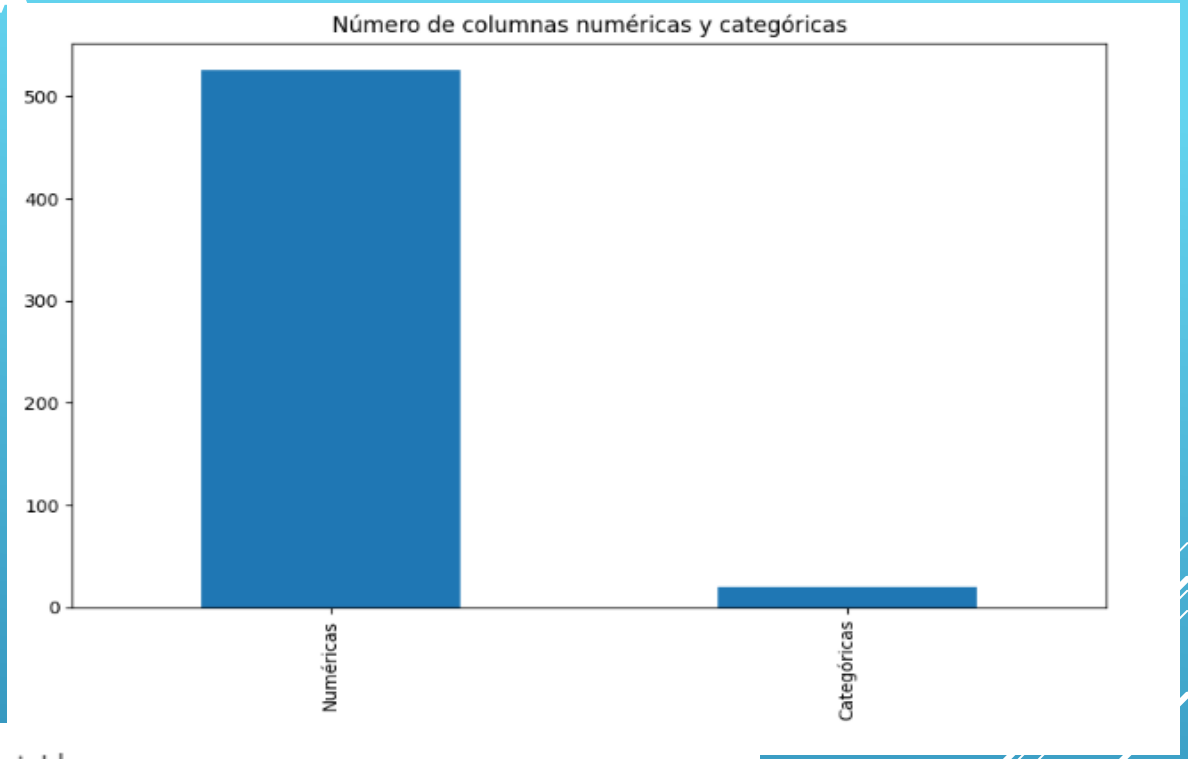
PYTHON

```
import pandas as pd url =
'https://drive.google.com/uc?id=1MwJRQi1BC82ZMu50HL1Q9_rDRZxP6N3D'
df = pd.read_csv(url, sep=';', on_bad_lines='skip')
 df_parcial = df.loc[:,
(df != 0).any(axis=0)]
df_parcial.dropna(inplace=True)
```

5.2 Visualización de Valores Nulos

PYTHON

```
import matplotlib.pyplot as plt
 import seaborn as sns plt.figure(figsize=(10, 6))
sns.heatmap(df_parcial.isnull(), cbar=False)
plt.title('Valores nulos en cada columna')
plt.savefig('valores_nulos.pdf')
```



5.3
PY

```
pr
pr
```

5_5 Tipo de Datos

```
[ ] # Resumen de los datos
    print('Resumen de los datos:')
    print(df_parcial.describe(include='all'))
```

```
25%     325680.000000    394.000000            2.000000          0.000000
50%     350751.000000    805.000000            3.000000          0.000000
75%     375176.750000   1662.000000            4.000000          1.000000
max     399979.000000  34394.000000           17.000000          8.000000

       CLIMA_EDUCATIVO        J_SEXO        J_EDAD  J_NIVEL_EDUCATIVO  \
count     12062.000000  12062.000000  12062.000000       12062.000000
unique             NaN           NaN           NaN                NaN
top                NaN           NaN           NaN                NaN
freq               NaN           NaN           NaN                NaN
mean          1.890317      1.461118     50.507793           3.889736
std           0.791081      0.498507     15.463877           1.567362
min           0.000000      1.000000     16.000000           1.000000
25%           1.000000      1.000000     39.000000           3.000000
50%           2.000000      1.000000     50.000000           4.000000
75%           3.000000      2.000000     63.000000           5.000000
max           3.000000      2.000000     99.000000           9.000000

            SEXO_SEL      EDAD_SEL  ...        ID_06B        ID_06C  \
count   12062.000000  12062.000000  ...  12062.000000  12062.000000
unique           NaN           NaN  ...           NaN           NaN
top              NaN           NaN  ...           NaN           NaN
freq             NaN           NaN  ...           NaN           NaN
mean        1.547339     43.478030  ...      9.363455     10.992373
std         0.497775     16.498695  ...     24.988618     27.818175
```

5_4 Tipo de Variables

```
[ ] # Tipos de variables
    print('Tipos de variables:')
    print(df_parcial.dtypes)
```

```
Tipos de variables:
ID_PER                    int64
WPER                      int64
CANT_MIEMBROS_HOGAR       int64
CANT_PERSONAS0A17         int64
CLIMA_EDUCATIVO           int64
                          ...
ID_06G                    int64
ID_06H                    int64
ID_06I                    int64
ID_07                     int64
ID_08                   float64
Length: 546, dtype: object
```

5_8 Estadisticas Descriptivas

```
df_seleccionado.describe()
```

	ID_PER	J_EDAD	J_NIVEL_EDUCATIVO	EDAD_SEL	NIVEL_EDUCATIVO_SEL	COBERTURA_SEL	SA_06	AL_01	edad_consumo_bebidas_alcoholicas	consumo_de_alcoh
count	12062.000000	12062.000000	12062.000000	12062.000000	12062.000000	12062.000000	12062.000000	12062.000000	12062.000000	12062.0000
mean	350395.476372	50.507793	3.889736	43.478030	4.020395	1.297961	3.497098	1.204029	1.179655	1.0242
std	28625.603078	15.463877	1.567362	16.498695	1.458175	0.482092	2.811689	1.820729	1.109153	1.0749
min	300003.000000	16.000000	1.000000	16.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.0000
25%	325680.000000	39.000000	3.000000	30.000000	3.000000	1.000000	2.000000	1.000000	0.000000	1.0000
50%	350751.000000	50.000000	4.000000	42.000000	4.000000	1.000000	4.000000	1.000000	1.000000	1.0000
75%	375176.750000	63.000000	5.000000	57.000000	5.000000	2.000000	5.000000	1.000000	2.000000	1.0000
max	399979.000000	99.000000	9.000000	75.000000	7.000000	9.000000	99.000000	99.000000	4.000000	99.0000

5_6 Dataframe Resultante

```
print(df_parcial.columns.to_list())
```

```
['ID_PER', 'WPER', 'CANT_MIEMBROS_HOGAR', 'CANT_PERSONAS0A17', 'CLIMA_EDUCATIVO', 'J_SEXO', 'J_EDAD',
```

```
print(df_parcial.head(10))
```

```
   ID_PER   WPER  CANT_MIEMBROS_HOGAR  CANT_PERSONAS0A17  CLIMA_EDUCATIVO  \
0  336578    124                    1                  0                2
1  305909    781                    1                  0                3
2  358892  34193                    4                  1                2
3  342664    968                    1                  0                3
4  394688  11509                    2                  0                3
5  370155  13159                    2                  0                3
6  308312   1440                    1                  0                2
7  354770    258                    1                  0                2
8  394258  14378                    4                  0                2
9  390628    315                    3                  0                2

   J_SEXO  J_EDAD  J_NIVEL_EDUCATIVO  SEXO_SEL  EDAD_SEL  ...  ID_06B  ID_06C  \
0       1      27                  4         1        27  ...       4       2
1       1      33                  6         1        33  ...       2       2
2       1      43                  4         2        19  ...       1       2
3       2      51                  6         2        51  ...       1       4
4       2      59                  5         1        31  ...       2       1
5       2      44                  5         2        44  ...       1      99
6       2      33                  4         2        34  ...       3       2
7       1      36                  4         1        36  ...       3       2
8       2      46                  6         1        42  ...       2       3
9       2      56                  3         1        24  ...       2       2

   ID_06D  ID_06E  ID_06F  ID_06G  ID_06H  ID_06I  ID_07  ID_08
0       2       1       1      99       2       3      1   98.0
1       1       3       2       3       2       4      2    0.0
```

Modelos y Cluster

Algoritmo Elegido

Para segmentar a los encuestados en grupos por tipos de consumo, se consideraron varios algoritmos de clustering, incluyendo K-means, DBSCAN y Agglomerative Clustering. Cada uno de estos algoritmos tiene sus propias ventajas y desventajas, y se eligió K-means por las siguientes razones:

- **Simplicidad y Eficiencia**: K-means es un algoritmo sencillo y eficiente que es fácil de implementar y entender. Es particularmente útil cuando se tiene una idea aproximada del número de clusters que se desea identificar.

- **Escalabilidad**: K-means es altamente escalable y puede manejar grandes conjuntos de datos de manera eficiente. Dado que el dataset de la encuesta contiene más de 12,000 filas, la escalabilidad del algoritmo fue un factor importante en su elección.

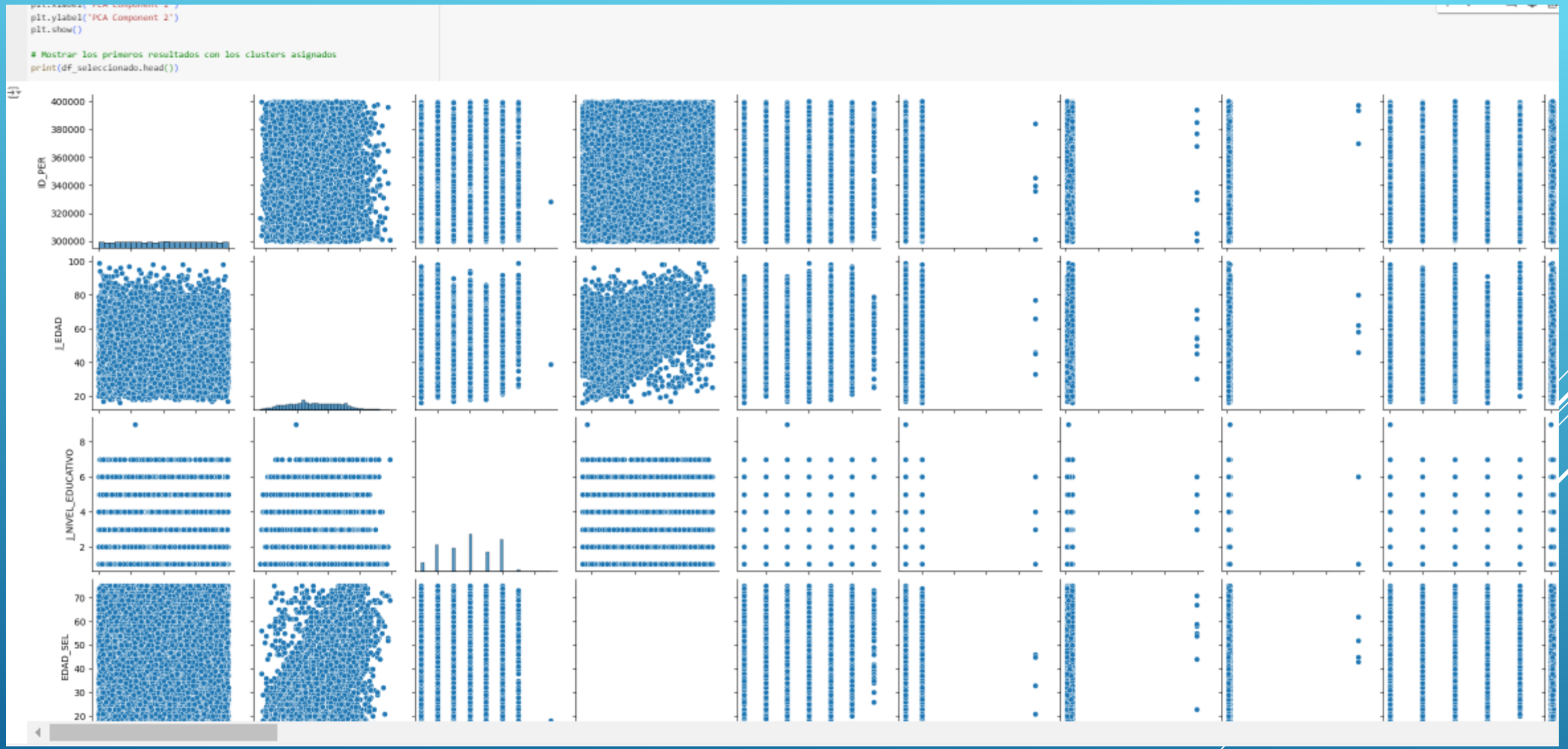
- **Interpretabilidad**: Los resultados de K-means son fáciles de interpretar. Cada encuestado se asigna a un cluster específico, y las características de cada cluster pueden ser analizadas para identificar patrones y tendencias.

- **Resultados Consistentes**: K-means tiende a producir resultados consistentes y reproducibles, especialmente cuando se utiliza una semilla aleatoria fija (random_state). Esto facilita la comparación de resultados y la validación del modelo.

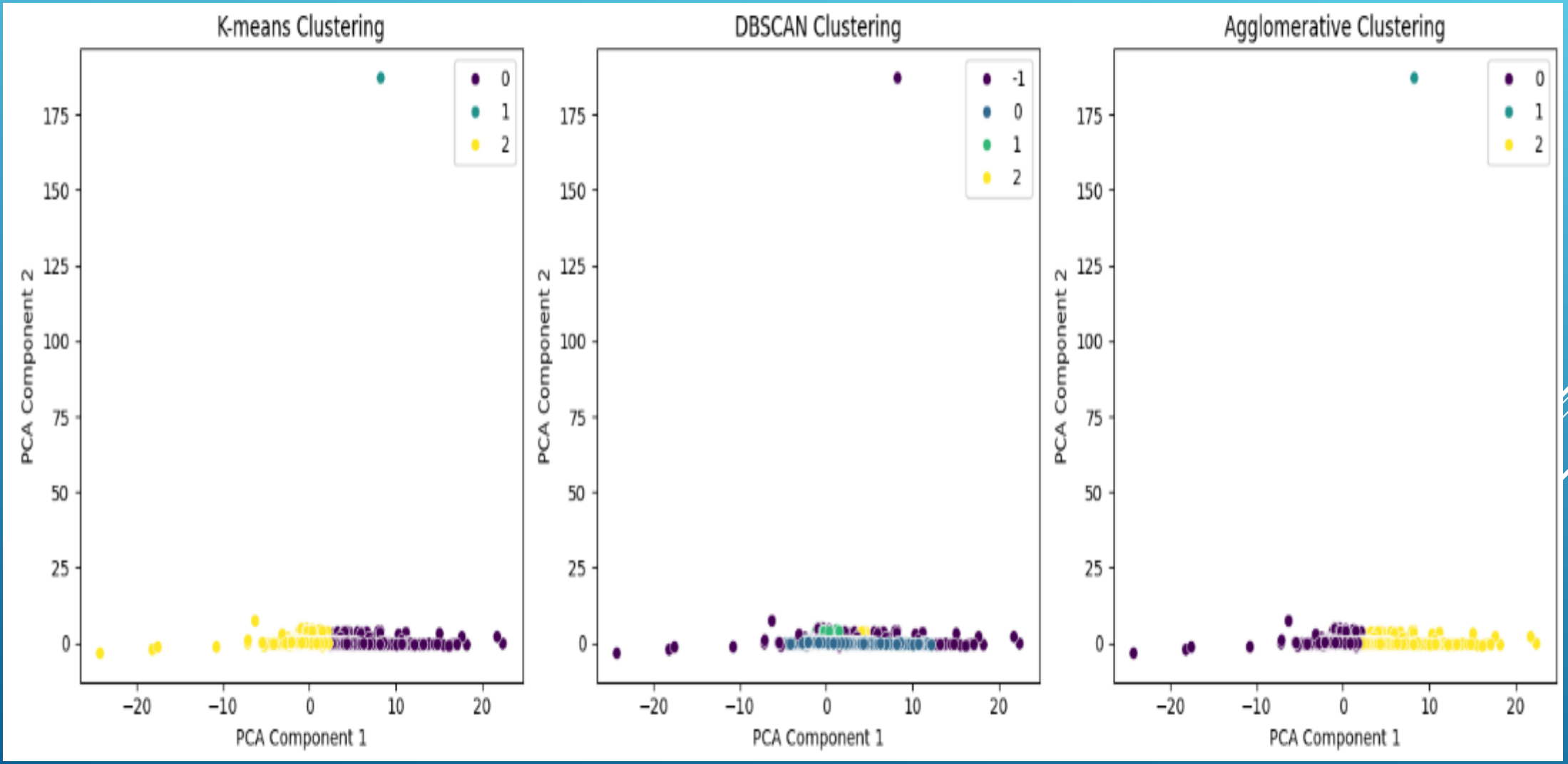
- **Adecuación a los Datos**: En pruebas preliminares, K-means mostró una buena capacidad para identificar grupos significativos en los datos de consumo de sustancias. Los clusters resultantes eran coherentes y proporcionaban información valiosa sobre los patrones de consumo.

```
plt.ylabel('PCA Component 2')
plt.show()

# Mostrar los primeros resultados con los clusters asignados
print(df_seleccionado.head())
```



Comparación de Modelos



Comparación con Otros Algoritmos

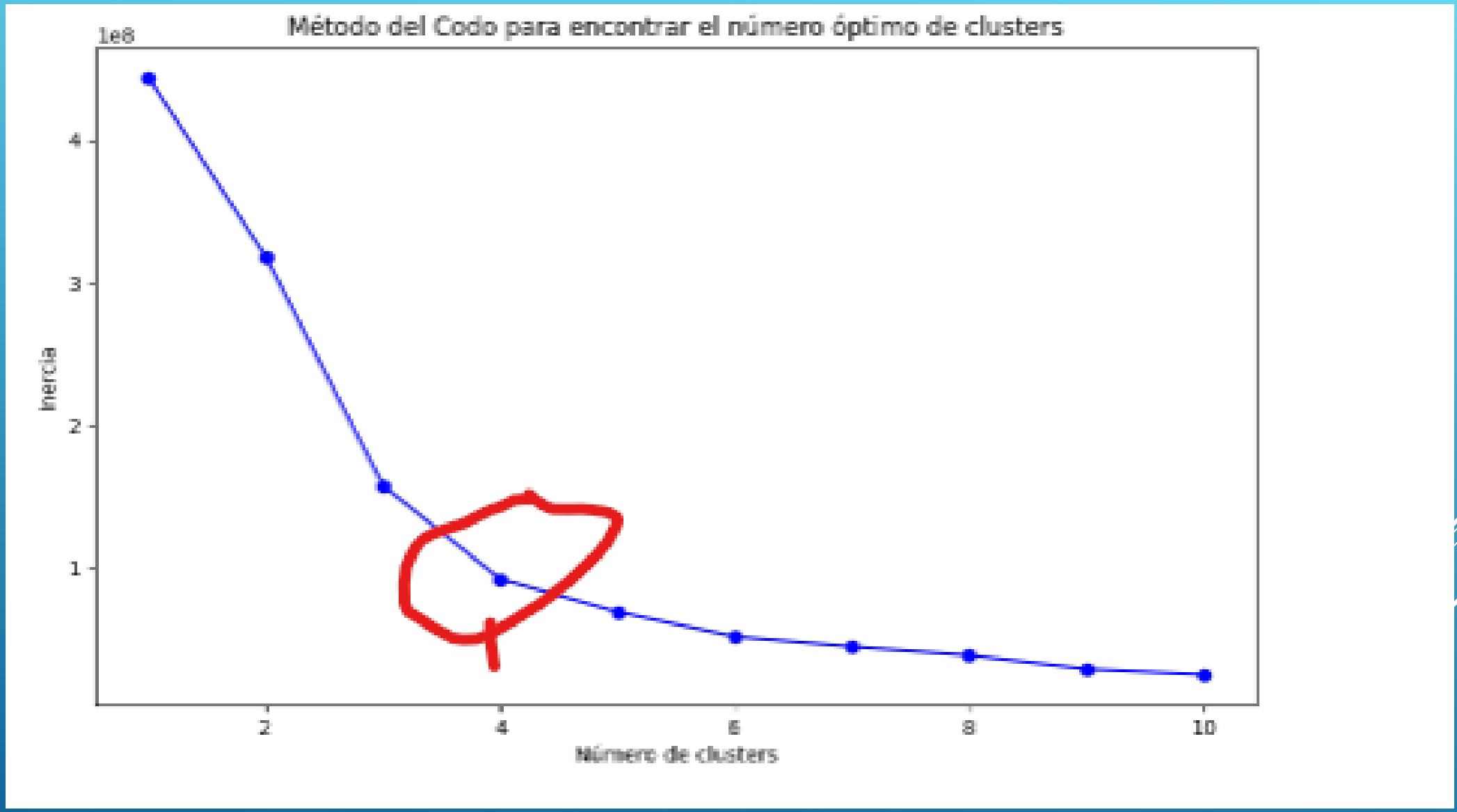
Aunque K-means fue el algoritmo elegido, también se consideraron otros métodos de clustering:

•**DBSCAN**: Este algoritmo es útil para identificar clusters de forma arbitraria y manejar ruido en los datos. Sin embargo, DBSCAN requiere la definición de parámetros como eps y min_samples, que pueden ser difíciles de ajustar correctamente para grandes conjuntos de datos.

•**Agglomerative Clustering**: Este método jerárquico puede proporcionar una visión más detallada de la estructura de los datos, pero es menos escalable y puede ser computacionalmente costoso para grandes datasets.

Conclusión

La elección de K-means se basó en su simplicidad, eficiencia, escalabilidad, interpretabilidad y adecuación a los datos. Aunque otros algoritmos también tienen sus ventajas, K-means proporcionó una solución balanceada y efectiva para segmentar a los encuestados en grupos significativos por tipos de consumo.


Método del Codo para encontrar el número óptimo de clusters
1e8
Inercia
Número de clusters
2
4
6
8
10
1
2
3
4

Interpretación de los Clusters

Los clusters identificados por el algoritmo K-means se interpretaron de la siguiente manera:

- **Cluster 0**: Este grupo incluye principalmente a jóvenes adultos con un nivel educativo medio y una alta prevalencia de consumo de alcohol y tabaco.

- **Cluster 1**: Este grupo está compuesto por adultos de mediana edad con un nivel educativo alto y un consumo moderado de alcohol y tabaco.

- **Cluster 2**: Este grupo incluye a personas mayores con un nivel educativo bajo y un bajo consumo de sustancias.

- **Cluster 3**: Este grupo está compuesto por jóvenes con un nivel educativo alto y un alto consumo de marihuana.

- **Cluster 4**: Este grupo incluye a adultos jóvenes con un nivel educativo medio y un consumo moderado de alcohol y marihuana.

8. Conclusiones y Recomendaciones

8.1 Conclusiones

- Patrones de Consumo: Existen patrones específicos de consumo de sustancias asociados con ciertos grupos demográficos.
- Factores de Riesgo: Los factores socioeconómicos y el acceso a sustancias son predictores significativos del consumo problemático.
- Impacto en Salud y Sociedad: El consumo de sustancias tiene un impacto negativo significativo en la salud y el bienestar social.

8.2 Recomendaciones

- Políticas Públicas: Implementar políticas que restrinjan el acceso a sustancias y promuevan la educación sobre los riesgos asociados.
- Programas de Intervención: Desarrollar programas de intervención basados en la educación y el apoyo comunitario.
- Investigación Continua: Continuar investigando para identificar nuevas tendencias y factores de riesgo asociados con el consumo de sustancias.

FIN