Data Mining

Alumnos:

- Alcántara Luna Diego Alexis
- Islas Diez de Sollano Brandon Jovani
- Torres Barajas Bryan Oswaldo

Proyecto – Limpieza de datos.

3CV2

Ocampo Botello Fabiola

Objetivo

Tomar los datos de la encuesta de Museos de México y sus visitantes durante el año 2019 para realizar un tratamiento de selección y limpieza de datos.

Esta base de datos fue proporcionada por el INEGI y se puede encontrar en el siguiente enlace
https://www.inegi.org.mx/programas/museos/default.html#Microdatos

El propósito de tomar estos datos es encontrar la relación que estos ocultan mediante la selección de 52 categorías, las cuales recibirán un tratamiento para los datos atípicos que se presenten, en este caso se tratarán valores nulos representados de la forma NaN.
Nuestro propósito es aplicar ciertas técnicas de limpieza y además una imputación de datos para poder producir un análisis eficiente de la información, ya que a lo largo del despliegue de la información de la base de datos, se producirán ciertas anomalías que pueden interferir con nuestros resultados y nuestro objetivo es evitar esto a toda costa

Tratamiento y Selección de los datos

Dentro de la base de datos seleccionada los datos con inconsistencias que se presentaron fueron del tipo nulo, básicamente campos que presentan información faltante.

Limpieza de Datos de visitas19

Librerias

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

Lectura del archivo visitas.csv

In [2]:
```
dataframe_general= pd.read_csv('visita19.csv')
```

Selección de columnas ¶

In [3]:
```
predataframe1=dataframe_general.iloc [:, [1,4,5,11,13,17,18,19,21]]
predaraframe2=dataframe_general.iloc [:, 23:66]
dataframe_seleccion= pd.concat([predataframe1,predaraframe2], axis=1)
```

Columnas que presentan inconsistencias

In [17]:
```
print(dataframe_seleccion.isnull().sum())
```

Figura 1. Librerías y selección de datos.

Para realizar el tratamiento de los datos se hizo uso del lenguaje de programación Python y su librería pandas.

Posteriormente, seleccionamos nuestra base de datos de donde se extraerá la información.

Y en nuestra entrada 3, seleccionamos las columnas de la base de datos que utilizaremos durante todo el proceso de limpieza.
Estas columnas las elegimos porque consideramos que son los datos que nos proporcionarán mayor información acerca de las visitas dentro del museo.

Y posteriormente hacemos la consulta para saber de nuestras columnas elegidas, cuales son las que presentan valores nulos, obteniendo como resultado lo siguiente:

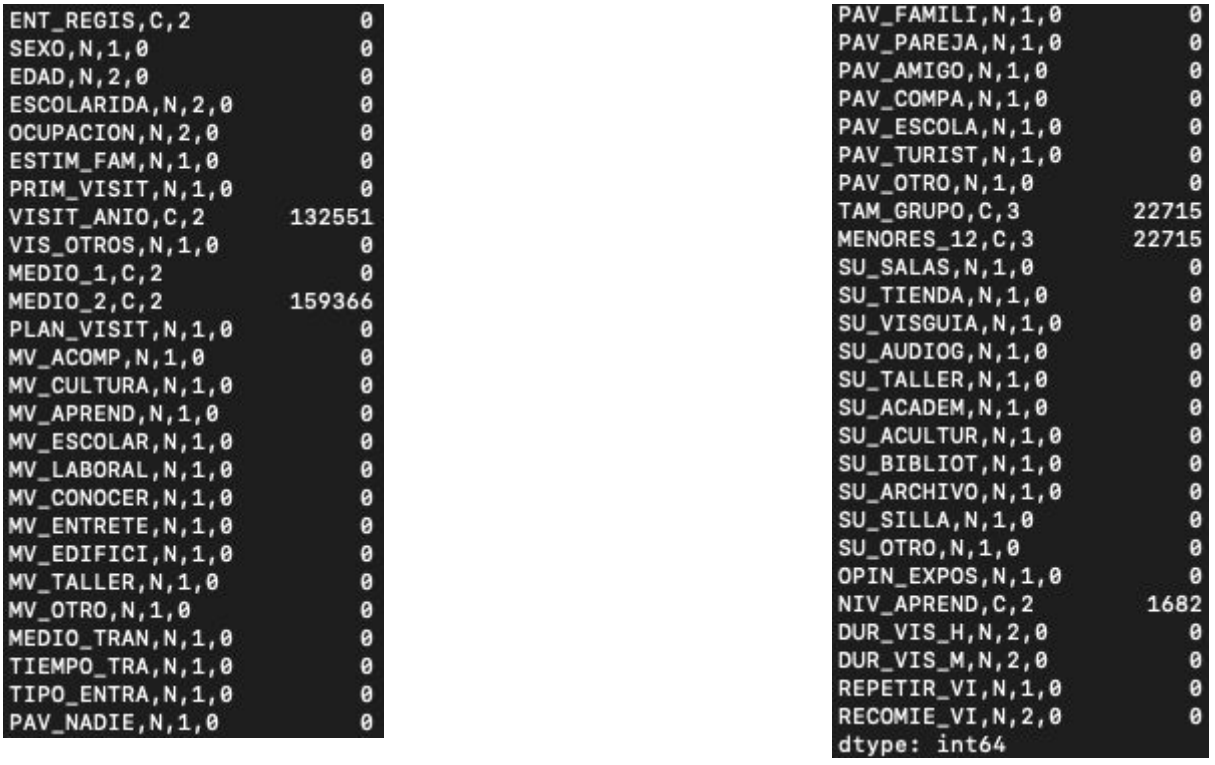


Figura 2. Consulta de valores nulos

Una vez que conocemos cuales son nuestras columnas procedemos a realizar el tratamiento de los datos nulos para nuestras columnas , '*VISIT_ANIO,C,2* ', '*MEDIO_2,C,2*', '*TAM_GRUPO,C,3*', *'MENORES_12,C,3'* y *' NIV_APREND,C,2 '* respectivamente.

Lo que se realizó fue, borrar cada uno de los datos nulos (representados con na) y posteriormente hallar la moda de los valores de cada columna antes mencionada y con ese valor rellenar las celdas vacías, esto con la finalidad de no perder la mayor parte de nuestra base de datos y poder realizar un mejor análisis.

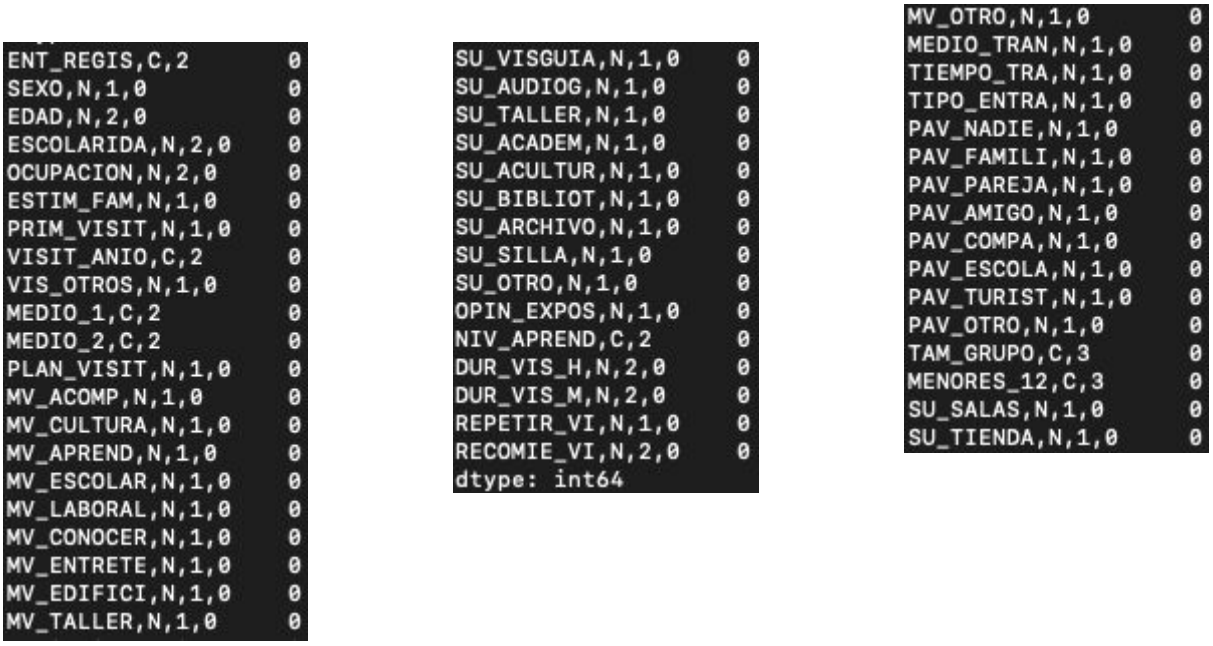
Este proceso se aplica individualmente a cada columna como se muestra en la siguiente figura.

Proceso de limpieza de datos

```
In [13]: #Para VISIT_ANIO
aux=dataframe_seleccion['VISIT_ANIO,C,2'].dropna()
dataframe_seleccion['VISIT_ANIO,C,2'].fillna(aux.mode(), inplace=True)
#Para MEDIO_2
aux=dataframe_seleccion['MEDIO_2,C,2'].dropna()
dataframe_seleccion['MEDIO_2,C,2'].fillna(aux.mode(), inplace=True)
#Para TAM_GRUPO
aux=dataframe_seleccion['TAM_GRUPO,C,3'].dropna()
dataframe_seleccion['TAM_GRUPO,C,3'].fillna(aux.mode(), inplace=True)
#Para MENORES
aux=dataframe_seleccion['MENORES_12,C,3'].dropna()
dataframe_seleccion['MENORES_12,C,3'].fillna(aux.mode(), inplace=True)
#Para NIV_APREND
aux=dataframe_seleccion['NIV_APREND,C,2'].dropna()
dataframe_seleccion['NIV_APREND,C,2'].fillna(aux.mode(), inplace=True)
```

Figura 3. Proceso de limpieza de datos

Una vez realizada la limpieza de los valores nulos, hacemos nuevamente la consulta para verificar que ya no tengamos dichos valores dentro de nuestra base, obteniendo como resultado lo siguiente:



```
ENT_REGIS,C,2        0
SEXO,N,1,0           0
EDAD,N,2,0           0
ESCOLARIDA,N,2,0     0
OCUPACION,N,2,0      0
ESTIM_FAM,N,1,0      0
PRIM_VISIT,N,1,0     0
VISIT_ANIO,C,2       0
VIS_OTROS,N,1,0      0
MEDIO_1,C,2          0
MEDIO_2,C,2          0
PLAN_VISIT,N,1,0     0
MV_ACOMP,N,1,0       0
MV_CULTURA,N,1,0     0
MV_APREND,N,1,0      0
MV_ESCOLAR,N,1,0     0
MV_LABORAL,N,1,0     0
MV_CONOCER,N,1,0     0
MV_ENTRETE,N,1,0     0
MV_EDIFICI,N,1,0     0
MV_TALLER,N,1,0      0
SU_VISGUIA,N,1,0     0
SU_AUDIOG,N,1,0      0
SU_TALLER,N,1,0      0
SU_ACADEM,N,1,0      0
SU_ACULTUR,N,1,0     0
SU_BIBLIOT,N,1,0     0
SU_ARCHIVO,N,1,0     0
SU_SILLA,N,1,0       0
SU_OTRO,N,1,0        0
OPIN_EXPOS,N,1,0     0
NIV_APREND,C,2       0
DUR_VIS_H,N,2,0      0
DUR_VIS_M,N,2,0      0
REPETIR_VI,N,1,0     0
RECOMIE_VI,N,2,0     0
dtype: int64
MV_OTRO,N,1,0        0
MEDIO_TRAN,N,1,0     0
TIEMPO_TRA,N,1,0     0
TIPO_ENTRA,N,1,0     0
PAV_NADIE,N,1,0      0
PAV_FAMILI,N,1,0     0
PAV_PAREJA,N,1,0     0
PAV_AMIGO,N,1,0      0
PAV_COMPA,N,1,0      0
PAV_ESCOLA,N,1,0     0
PAV_TURIST,N,1,0     0
PAV_OTRO,N,1,0       0
TAM_GRUPO,C,3        0
MENORES_12,C,3       0
SU_SALAS,N,1,0       0
SU_TIENDA,N,1,0      0
```

Figura 4 . Limpieza de datos completada

Como podemos observar, no queda rastro de los valores nulos.

Posteriormente procedemos a eliminar valores que nos pueden generar inconsistencias al momento de realizar la descripción de los datos, en este caso eliminamos los valores “99”, “999” que significan “no especificado” y y “98” que hace referencia a un valor más alto del que se puede especificar.
Se tomó la decisión de eliminar estos datos, ya que al momento de realizar descripciones como el promedio, estos valores numéricos nos afectan seriamente el resultado esperado.

Tratamiento de inconsistencias.

In [14]:
```
#Eliminar inconsistencias en Duracion Visitas al año
condicion_visit=(dataframe_seleccion['VISIT_ANIO,C,2']!=98) &(dataframe_seleccion['VISIT_ANIO,C,2']!=99)
dataframe_seleccion=dataframe_seleccion[condicion_visit]
#Eliminar inconsistencias en Tamaño del grupo
condicion_tam=(dataframe_seleccion['TAM_GRUPO,C,3']!=999)
dataframe_seleccion=dataframe_seleccion[condicion_tam]
#Eliminar inconsistencias en Tamaño del grupo menores de 12
condicion_menores=(dataframe_seleccion['MENORES_12,C,3']!=999)
dataframe_seleccion=dataframe_seleccion[condicion_menores]
#Eliminar inconsistencias en nivel de aprendizaje
condicion_ap=(dataframe_seleccion['NIV_APREND,C,2']!=99)
dataframe_seleccion=dataframe_seleccion[condicion_ap]
#Eliminar inconsistencias en Duracion Visitas en horas
condicion_vish=(dataframe_seleccion['DUR_VIS_H,N,2,0']!=98) &(dataframe_seleccion['DUR_VIS_H,N,2,0']!=99)
dataframe_seleccion=dataframe_seleccion[condicion_vish]
#Eliminar inconsistencias en Duracion Visitas en minutos
condicion_vism=(dataframe_seleccion['DUR_VIS_M,N,2,0']!=99)
dataframe_seleccion=dataframe_seleccion[condicion_vism]
#Eliminar datos inconsistentes en el campo Sexo
condicion_edad=(dataframe_seleccion['EDAD,N,2,0']!=98) &(dataframe_seleccion['EDAD,N,2,0']!=99)
dataframe_seleccion=dataframe_seleccion[condicion_edad]
```

Figura 5 . Tratamiento de inconsistencias

Diccionario de Datos

En la presente gráfica se muestran los distintos tipos de variables y cuales son su significado dentro de la base de datos.

En donde se ubica la clasificación “Generico” nos referimos a valores que suelen aparecer con cierta regularidad dentro de nuestro registro de datos

```
sexo= {1:"Hombre", 2:"Mujer"}

escolaridad= {1:"Ninguna",2:"Preescolar",3:"Primaria", 4:"Secuandaria",
              5:"Estudios técnicos con secundaria terminada", 6:"Normal básica",
              7:"Preparatoria o bachillerato", 8:"Estudios técnicos con preparatoria terminada",
              9:"Licenciatura", 10:"Maestría o doctorado", 99:"No especifica"}

generico={1:"Si", 2:"No", 99:"No especifica"}

ocupacion={1:"Funcionarios, directores y jefes", 2:"Profesionistas y técnicos",
           3:"Trabajadores auxiliares en actividades administrativas",
           4:"Comerciantes, empleados en ventas y agentes de ventas",
           5:"Trabajadores en servicios personales y vigilancia",
           6:"Trabajadores en actividades agrícolas, ganaderas, forestales, caza y pesca",
           7:"Trabajadores artesanales",
           8:"Operadores de maquinaria industrial, ensambladores, choferes y conductores de transporte",
           9:"Trabajadores en actividades elementales y de apoyo",
          10:"Busca trabajo", 11:"No trabaja", 98:"Insuficientemente especificada", 99:"No especificada"}

medio_e={1:"Maestro, compañeros de estudio o libros de texto",
         2:"Conoce desde siempre este lugar",3:"Amigos, familiares o conocidos",
         4:"Por la televisión", 5:"Folleto, espectacular, anuncio o volante",
         6:"Internet", 7:"Oficina turística o viaje turístico", 8:"Por la radio",
         9:"Periódico, revista o libro", 10:"Redes sociales", 11:"Por casualidad", 12:"Otro", 99:"No especificado"}

medio_t={1:"Vehículo particular", 2:"Transporte público", 3:"Transporte turístico", 4:"Taxi",
         5:"Bicicleta",6:"Caminando", 7:"Otro", 9:"No especificado"}

timpo_t={1:"De 1 a 30 min", 2:"De 31 min a 1h",
         3:"De 1.01 a 1.30 h", 4:"De 1.31 a 2 h",
         5:"De 2.01 a 3 h", 6:"De 3.01 a 4 h",
         7:"De 4.01 a 5 h",
         8:"De 5.01 y más", 9: "No especificado"}

ent_regis = {1:"Aguascalientes", 2:"Baja California",3:"Baja California Sur", 4:"Campeche",
             5:"Campeche", 6:"Colima", 7:"Colima", 8:"Chihuahua", 9:"Ciudad de México",
            10:"Durango", 11:"Guanajuato", 12:"Guerrero", 13:"Hidalgo", 14:"Jalisco",
             15:"México", 16:"Michoacán de Ocampo", 17:"Morelos", 18:"Nayarit", 19:"Nuevo León",
             20:"Oaxaca",21:"Puebla", 22:"Querétaro", 23:"Quintana Roo", 24:"San Luis Potosí",
             25:"Sinaloa", 26:"Sonora", 27:"Sonora", 28:"Tamaulipas", 29:"Tlaxcala",
             30:"Veracruz de Ignacio de la Llave", 31:"Yucatán", 32:"Zacatecas", 33:"Extranjera",
             99:"No especificada"}
```

Figura 6 . Diccionario de Datos

Descripción de los Datos

Posteriormente, para visualizar las características deseadas, realizamos la siguiente consulta para cada una de las columnas, obteniendo como resultado lo siguiente:

Visualización de algunos datos estadisticos

```
In [16]: print(dataframe_seleccion['EDAD,N,2,0'].describe())
         print(dataframe_seleccion['VISIT_ANIO,C,2'].describe())
         print(dataframe_seleccion['NIV_APREND,C,2'].describe())
         print(dataframe_seleccion['TAM_GRUPO,C,3'].describe())
         print(dataframe_seleccion['MENORES_12,C,3'].describe())
         print(dataframe_seleccion['DUR_VIS_H,N,2,0'].describe())
         print(dataframe_seleccion['DUR_VIS_M,N,2,0'].describe())
```

```
count    158443.000000
mean          5.476588
std          13.966455
min           1.000000
25%           1.000000
50%           3.000000
75%           4.000000
max         990.000000
Name: TAM_GRUPO,C,3, dtype: float64
count    158443.000000
mean          0.753571
std           3.873936
min           0.000000
25%           0.000000
50%           0.000000
75%           1.000000
max         490.000000
Name: MENORES_12,C,3, dtype: float64
count    180747.000000
mean          0.666960
std           0.954106
min           0.000000
25%           0.000000
50%           0.000000
75%           1.000000
max           8.000000
Name: DUR_VIS_H,N,2,0, dtype: float64
count    180747.000000
mean         22.411703
std          16.018873
min           0.000000
25%           8.000000
50%          25.000000
75%          30.000000
max          59.000000
Name: DUR_VIS_M,N,2,0, dtype: float64
```

```
count    180747.000000
mean         34.110823
std          14.847694
min          12.000000
25%          22.000000
50%          31.000000
75%          43.000000
max          97.000000
Name: EDAD,N,2,0, dtype: float64
count    50824.000000
mean         2.268987
std          3.405377
min          0.000000
25%          0.000000
50%          1.000000
75%          3.000000
max         30.000000
Name: VISIT_ANIO,C,2, dtype: float64
count    179389.000000
mean          8.990802
std           1.366201
min           0.000000
25%           8.000000
50%           9.000000
75%          10.000000
max          10.000000
Name: NIV_APREND,C,2, dtype: float64
```

Figura 7 . Descripción de los datos

La información desplegada con esta consulta es la totalidad de los datos (‘Count’‘), media(‘Mean’), la desviación estándar (‘Std’), el valor mínimo (‘Min’), el valor máximo (‘Max’) y los respectivos cuartiles (“25, 50, 75”).

El tipo de dato que en estas consultas podemos hallar son de tipo Ordinal numérico, ya que conforme aumenta el valor numérico, implica un valor de mayor importancia.

Y lo que estos datos nos dicen, por ejemplo, dentro de TAM_GRUPO podemos observar que el tamaño promedio fue de 5 personas, el tamaño máximo fue de 990, el tamaño mínimo fue de 1.

Cálculo de la mediana y la varianza

Posteriormente, para hallar el valor de la mediana, procedemos a hacer la siguiente consulta, la cual nos proporcionará el dato que se encuentre en medio dentro de todos los datos que contiene cada columna.

```
varianza=dataframe_seleccion.var()
moda=dataframe_seleccion.mode()
mediana=dataframe_seleccion.median()
diccionario={'Categorias':list(dataframe_seleccion),'Mediana':list(mediana), 'Varianza':list(varianza)}
#print(diccionario)
datos= pd.DataFrame(diccionario)
datos
```

Figura 8 . Consulta de la Mediana y varianza

	Categorias	Mediana	Varianza
0	ENT_REGIS,C,2	14.0	61.177618
1	SEXO,N,1,0	2.0	0.248953
2	EDAD,N,2,0	31.0	220.454002
3	ESCOLARIDA,N,2,0	9.0	58.017930
4	OCUPACION,N,2,0	11.0	696.658308
5	ESTIM_FAM,N,1,0	1.0	0.332346
6	PRIM_VISIT,N,1,0	1.0	0.202123
7	VISIT_ANIO,C,2	1.0	11.596594
8	VIS_OTROS,N,1,0	1.0	0.483813
9	MEDIO_1,C,2	3.0	24.005381
10	MEDIO_2,C,2	6.0	10.001199
11	PLAN_VISIT,N,1,0	1.0	0.541757
12	MV_ACOMP,N,1,0	0.0	0.179683
13	MV_CULTURA,N,1,0	0.0	0.236976
14	MV_APREND,N,1,0	0.0	0.206305
15	MV_ESCOLAR,N,1,0	0.0	0.116627
16	MV_LABORAL,N,1,0	0.0	0.026092
17	MV_CONOCER,N,1,0	0.0	0.191111
18	MV_ENTRETE,N,1,0	0.0	0.149229
19	MV_EDIFICI,N,1,0	0.0	0.106777
20	MV_TALLER,N,1,0	0.0	0.021979
21	MV_OTRO,N,1,0	0.0	0.005950
22	MEDIO_TRAN,N,1,0	2.0	3.322012
23	TIEMPO_TRA,N,1,0	1.0	3.298891
24	TIPO_ENTRA,N,1,0	2.0	0.274770
25	PAV_NADIE,N,1,0	0.0	0.108172
26	PAV_FAMILI,N,1,0	0.0	0.248623
27	PAV_PAREJA,N,1,0	0.0	0.122856
28	PAV_AMIGO,N,1,0	0.0	0.163191
29	PAV_COMPA,N,1,0	0.0	0.024587
30	PAV_ESCOLA,N,1,0	0.0	0.061225
31	PAV_TURIST,N,1,0	0.0	0.015514
32	PAV_OTRO,N,1,0	0.0	0.000382
33	TAM_GRUPO,C,3	3.0	195.061852
34	MENORES_12,C,3	0.0	15.007379
35	SU_SALAS,N,1,0	1.0	0.077341
36	SU_TIENDA,N,1,0	0.0	0.137715
37	SU_VISGUIA,N,1,0	0.0	0.219194
38	SU_AUDIOG,N,1,0	0.0	0.027290
39	SU_TALLER,N,1,0	0.0	0.050099
40	SU_ACADEM,N,1,0	0.0	0.017220
41	SU_ACULTUR,N,1,0	0.0	0.045136
42	SU_BIBLIOT,N,1,0	0.0	0.035361
43	SU_ARCHIVO,N,1,0	0.0	0.023409
44	SU_SILLA,N,1,0	0.0	0.002632
45	SU_OTRO,N,1,0	0.0	0.022512
46	OPIN_EXPOS,N,1,0	1.0	0.631761
47	NIV_APREND,C,2	9.0	1.866506
48	DUR_VIS_H,N,2,0	0.0	0.910318
49	DUR_VIS_M,N,2,0	25.0	256.604304
50	REPETIR_VI,N,1,0	1.0	0.176497
51	RECOMIE_VI,N,2,0	13.0	18.847979

Figura 9 . Resultados consulta Mediana y Varianza

En este caso, para los resultados obtenidos, se puede conocer el significado consultando el diccionario de datos, por ejemplo, para “ENT_REGIS”, que es la columna que nos proporciona la información de la entidad de registro, el valor 12 corresponde a Jalisco, lo que nos dice que este es el valor que se halla a la mitad de los datos.
En el caso de la columna “SEXO”, la mediana es 2, cuyo valor es Mujer.

Cálculo de la moda

Para el cálculo de la moda se hace uso de la función que provee python, la moda siempre será el valor que más se repite dentro de nuestro registro, en el caso de la columna “SEXO”, el dato que más se repite es 2, que es el número que representa a dato “Mujer”, por lo que podemos concluir que asisten más mujeres al museo que hombres.

```
In [8]: moda=dataframe_seleccion.mode()
        print(moda)
```

```
   ENT_REGIS,C,2  SEXO,N,1,0  EDAD,N,2,0  ESCOLARIDA,N,2,0  OCUPACION,N,2,0  \
0              9           2          18                 9               11

   ESTIM_FAM,N,1,0  PRIM_VISIT,N,1,0  VISIT_ANIO,C,2  VIS_OTROS,N,1,0  \
0                1                  1             0.0                1

   MEDIO_1,C,2  ...  SU_BIBLIOT,N,1,0  SU_ARCHIVO,N,1,0  SU_SILLA,N,1,0  \
0            3  ...                 0                 0               0

   SU_OTRO,N,1,0  OPIN_EXPOS,N,1,0  NIV_APRENDIZ,C,2  DUR_VIS_H,N,2,0  \
0              0                 1             10.0                0

   DUR_VIS_M,N,2,0  REPETIR_VI,N,1,0  RECOMIE_VI,N,2,0
0               30                 1                13

[1 rows x 52 columns]
```

Figura 10 . Consulta de la Moda

Representaciones gráficas

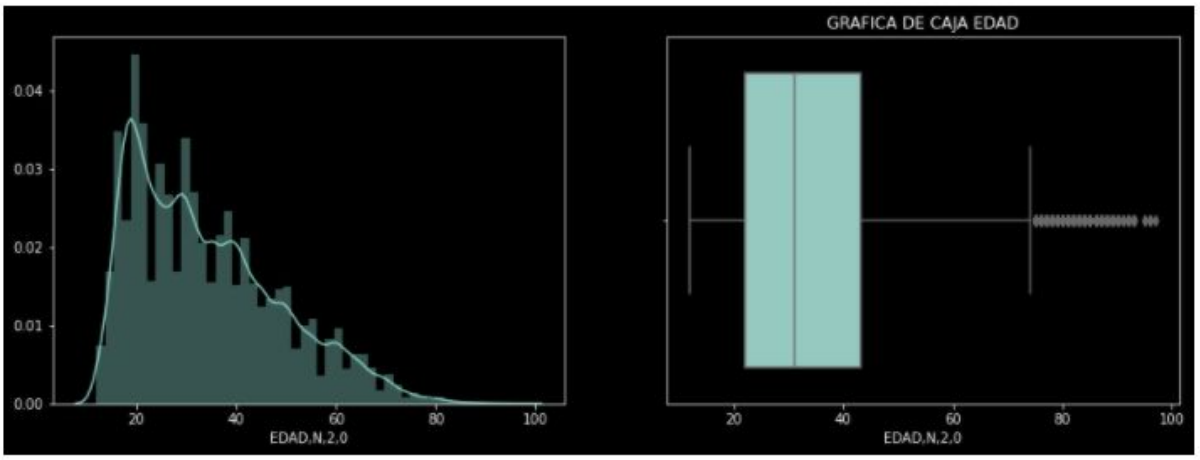


Figura 11 . Gráfica de edad

Según los datos arrojados por la base de datos y con la ayuda del tratamiento de datos eliminando algunas inconsistencias se hizo la gráfica que nos arrojó que el rango principal de edades de la población está entre los 20 y 42 años, eso quiere decir que la mayoría de los datos se encuentra muy cercano al promedio que es de 34 años, además en la gráfica de la izquierda podemos ver la distribución de los datos en un histograma.

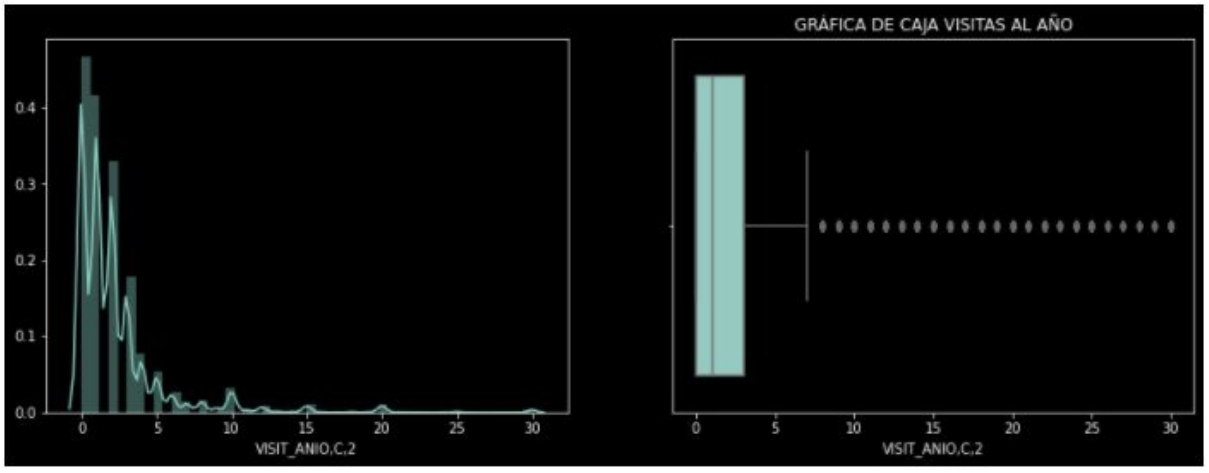


Figura 12. Gráfica de visitas al año

Según los datos arrojados por la base de datos y con la ayuda del tratamiento de datos eliminando algunas inconsistencias se hizo las gráficas correspondientes donde se logra observar una gran cantidad de datos sesgados, que no fueron limpiados por la razón de no perder información importante, pero la mayoría de los encuestados arrojó que solo asisten una vez al año al museo o a lo mucho 5 ocasiones, posteriormente se analiza esta variable en conjunto con demás variables de análisis.

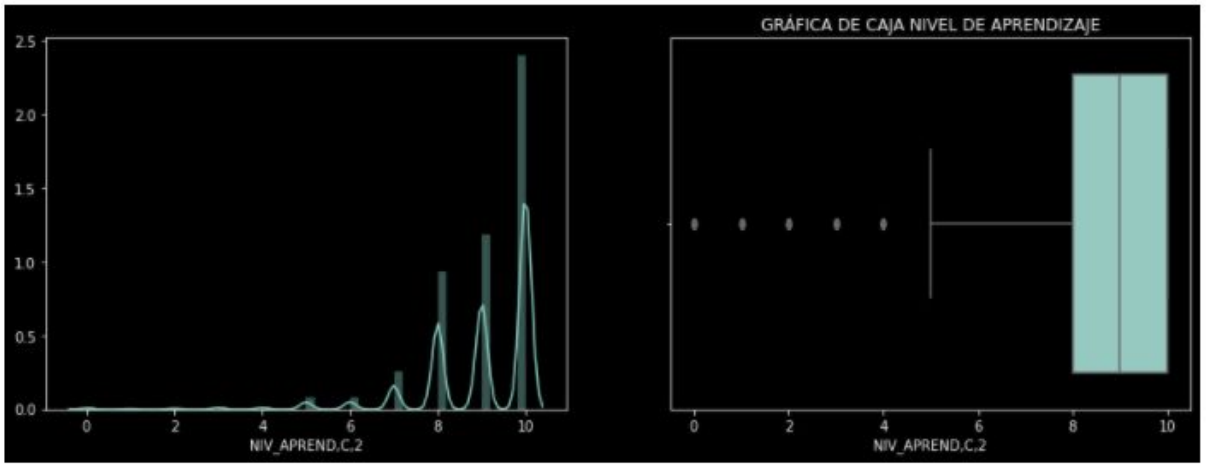


Figura 13. Gráfica de Nivel de aprendizaje

Según los datos arrojados por la base de datos y con la ayuda del tratamiento de datos eliminando algunas inconsistencias se hicieron las gráficas que nos indica que la mayoría de los encuestados que asistieron a un museo aprendieron algo de esa visita la mayoría de los datos se encuentra en calificaciones satisfactorias es decir de 8 a 10 donde el promedio de esta variable es de 9, lo cual nos permite sacar una conclusión de que la mayoría aprende algo al asistir a un museo.

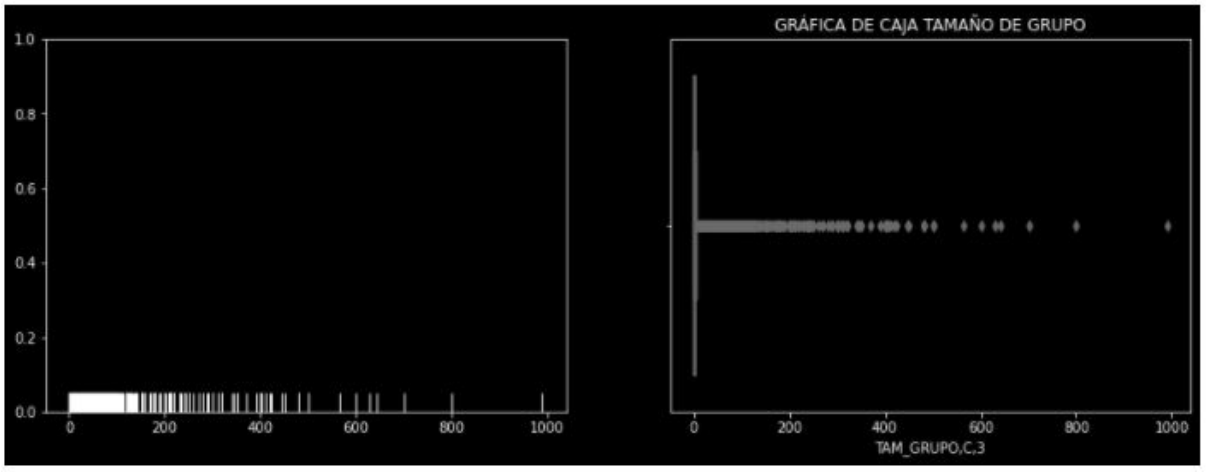


Figura 14. Gráfica de Tamaño de grupo

Según los datos arrojados por la base de datos y con la ayuda del tratamiento de datos eliminando algunas inconsistencias se hicieron las siguientes graficas que estan muy cargados los datos, como sabemos en la educación mexicana normalmente se le solicita a un grupo de alumnos asistir a un museo, por ese motivo el tamaño de grupo oscila demasiado, incluso muchas instituciones hacen excursiones a museos donde siendo conservador asisten 30 personas y exageradamente van grupos de 500 personas, entonces a esto podemos decir que se debe ese sesgo.

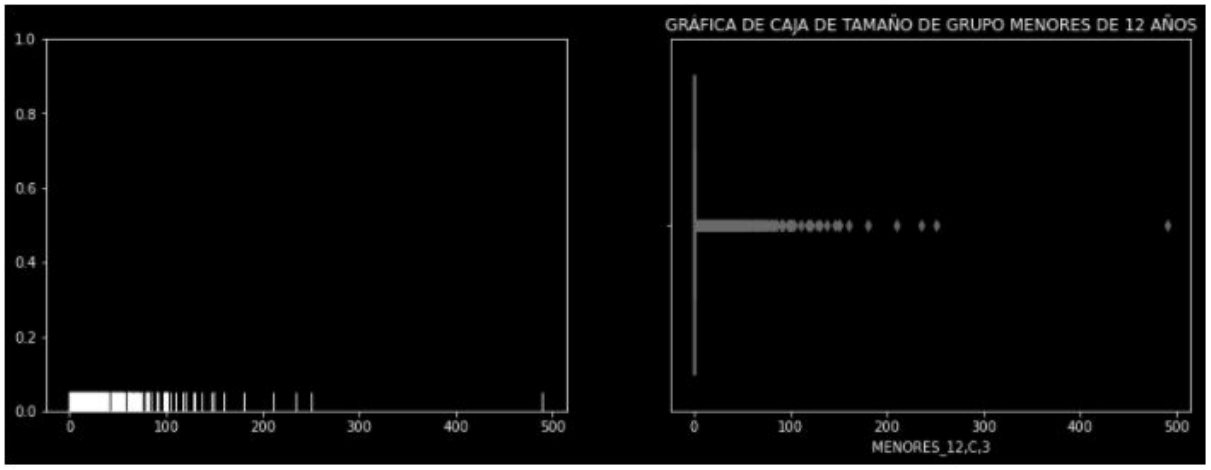


Figura 15. Gráfica de Tamaño de grupo de menores 12 años

De igual manera que el punto anterior los datos están sesgados por que muchas escuelas realizan excursiones, por ejemplo el tamaño máximo en esta categoría fue de 480 personas menores de 12 años donde da mas argumentos a la afirmación de que las escuelas hacen excursiones a los museos, mas adelante en el documento se comparan esta variable con otras mas para poder seguir dando ese argumento de que asisten por excursiones escolares o proyectos estudiantiles.

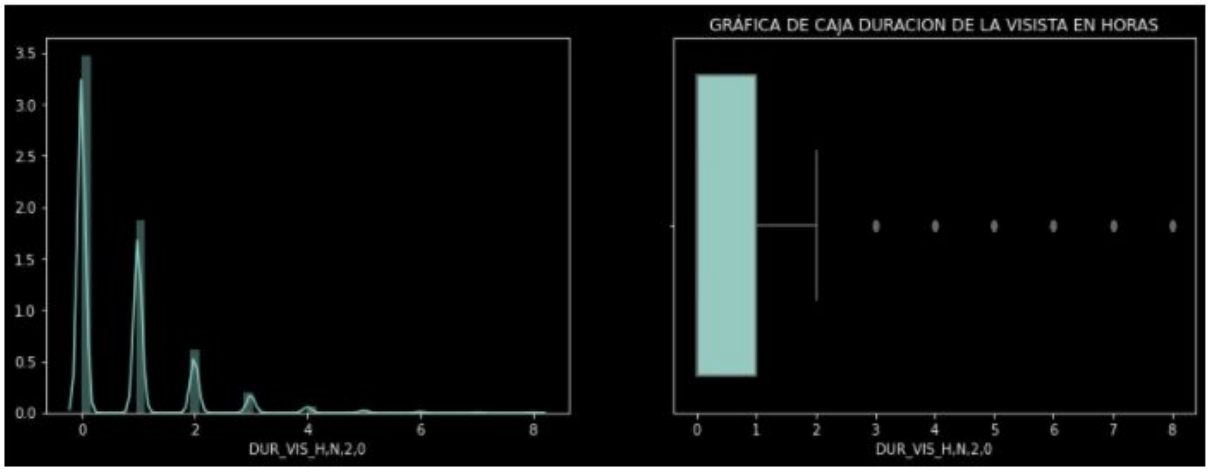


Figura 16. Gráfica de Duración de la visita en horas

Según los datos arrojados por la base de datos y con la ayuda del tratamiento de datos eliminando algunas inconsistencias se hicieron las siguientes gráficas, donde podemos ver que muchas personas no asisten al museo muchas horas lo cual nos puede servir en el futuro para ver qué tipo de museo es y cual debería de ser la estancia ideal en el museo para poder sacarle mayor provecho a esos datos, así mismo como sabemos hay diferentes tipos ya tamaños de museos por lo que muchos aveces son muy pequeños y no tardan mas de una hora en recorrerlos pero de igual manera sabemos que puede ser un museo muy grande o con exposiciones temporales que nos tardamos demasiado en recorrer, hablando de la visita en horas podemos ver que la estancia promedio está entre 0 a 1 hora.¶

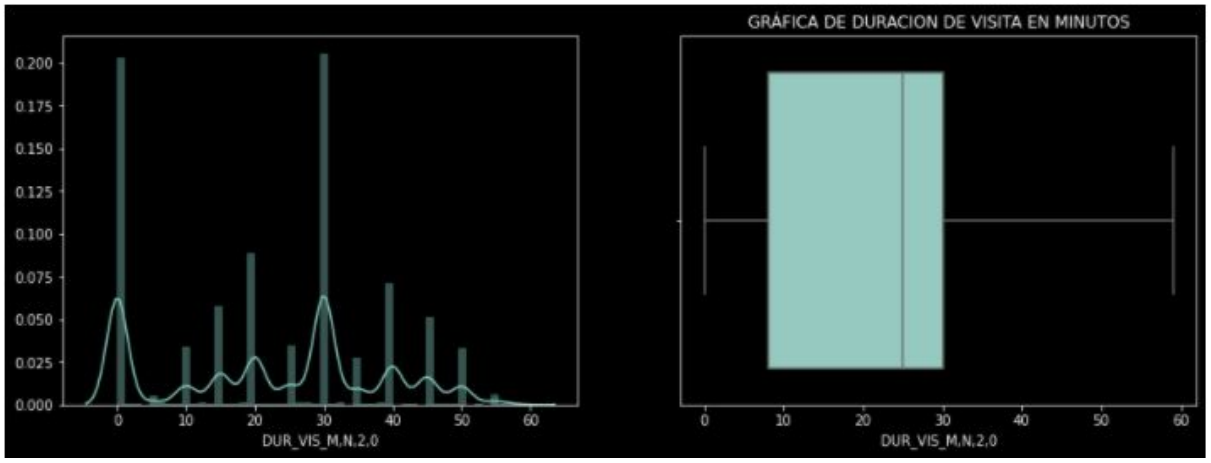


Figura 17. Gráfica de Duración de la visita en minutos

Según los datos arrojados por la base de datos y con la ayuda del tratamiento de datos eliminando algunas inconsistencias se hicieron las siguientes gráficas donde es un como complemento de la categoría anterior donde podríamos decir que en ocasiones la asistencia es un poco obligatoria para asegurar o ver un poco mas a fondo este argumento mas adelante se podrá analizar, hablando de los minutos podemos ver que la estancia promedio va de los 10 minutos a los 30 minutos.

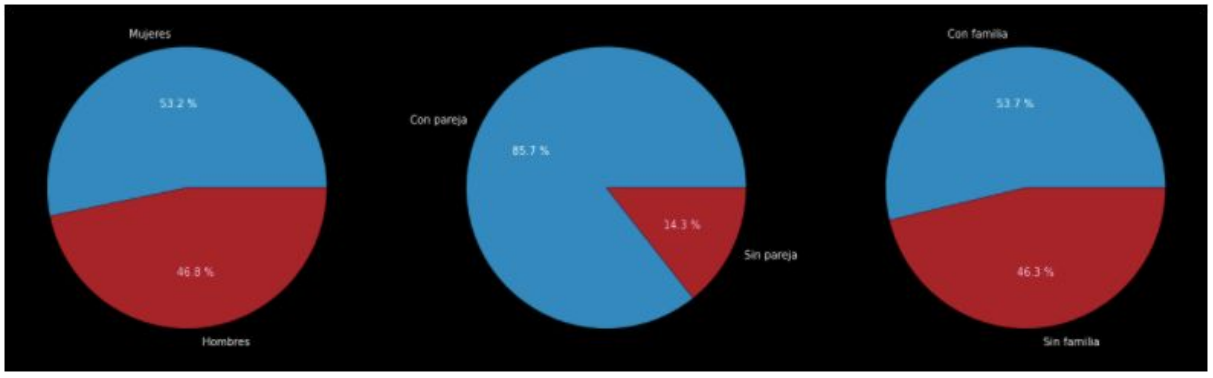


Figura 18. Gráficas de pastel Sexo, acompañante con pareja, acompañañnte con familia

-Gráfica 1

Podemos observar que asisten mas mujeres que hombres al museo por una diferencia de 6.8%

-Gráfica 2

Podemos notar que la mayoría de los asistentes a museos van acompañados con su pareja con un 85.7%

-Gráfica 3

Podemos ver que la mayoria de los asistentes prefiere ir solo o con otras personas que con su familia con un 53%

Figura 19. Gráfica de pastel acompañante con amigo

En esta gráfica podemos ver con que tipo de acompañante van los asistentes donde el 79.6% asiste sin sus amigos.

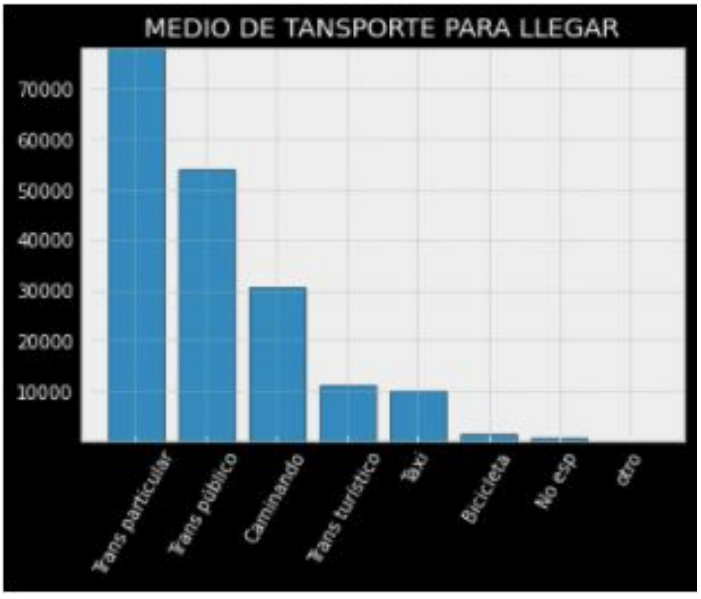


Figura 20. Gráfica de barras medio de transporte para llegar

En esta gráfica podemos ver en qué tipo de transporte se trasladan los visitantes con 80,000 asistentes que van en transporte particular y 55,000 asistentes llegan en transporte público y 30,000 asistentes llegan caminando a los museos, donde podemos ver los accesos de los museos pueden ser un poco difíciles de acceder.

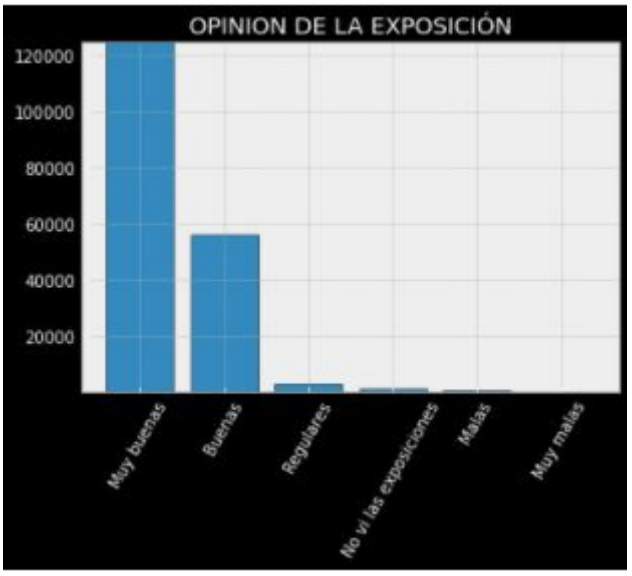


Figura 21. Gráfica de barras opinión de la exposición

En esta gráfica se observa el opinión de las personas respecto a las exposiciones de los museos mas de 120,000 de comentarios que las exposiciones son muy buenas, si consideramos que después de la limpieza quedaron 180,000 registros pues casi el 90 % de las personas aseguró que las exposiciones son muy buenas.

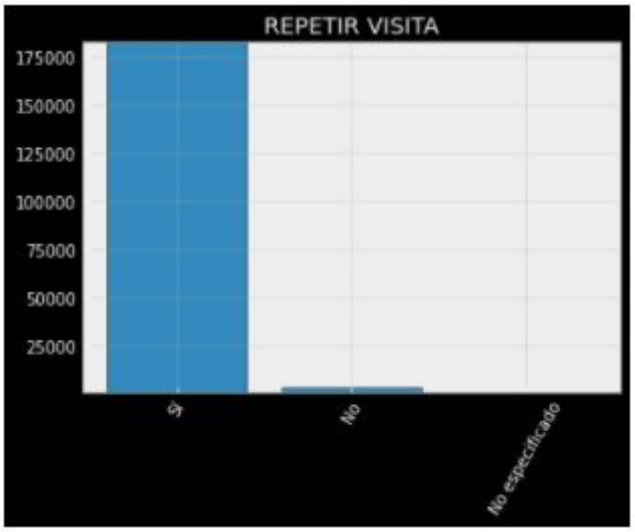


Figura 22. Gráfica de barras repetir visita

Esta gráfica es muy concreta con los datos que arroja donde según 180,000 dice que repetiría la visita al museo donde.

Diagrama de Violín

Un diagrama de violín se utiliza para visualizar la distribución de los datos y su densidad de probabilidad.
Este gráfico es una combinación de un diagrama de cajas y bigotes y un diagrama de densidad girado y colocado a cada lado, para mostrar la forma de distribución de los datos.

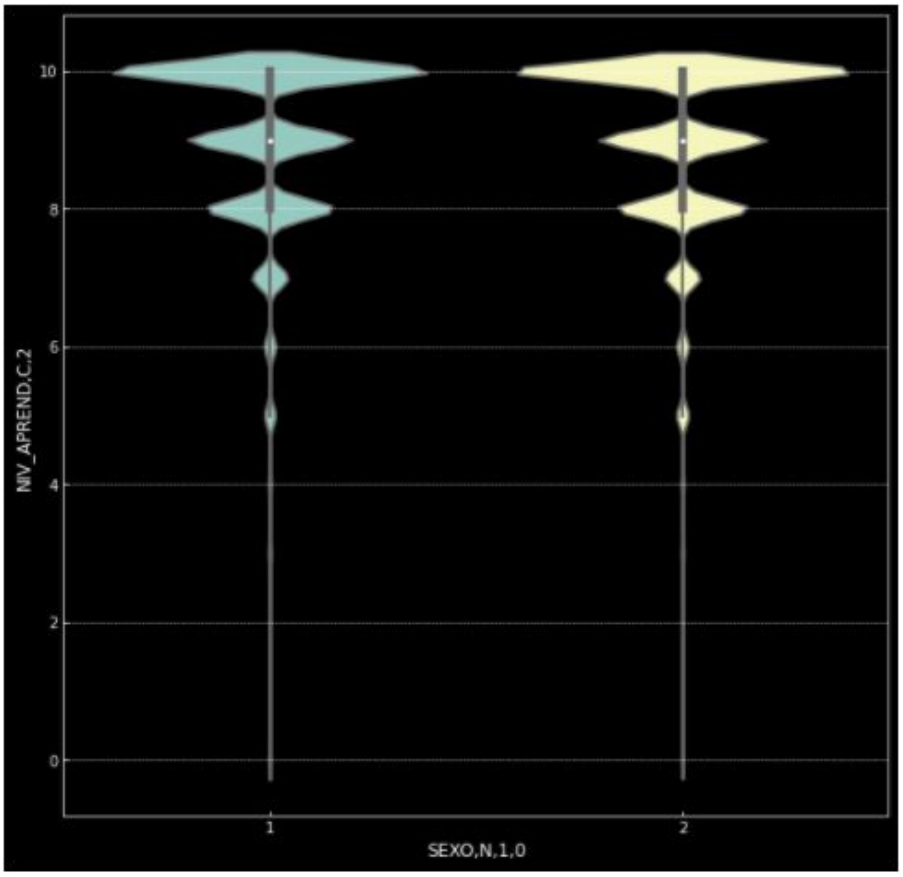


Figura 23. Diagrama de violín sexo y nivel de aprendizaje

En esta grafica de doble entrada analizamos las variebles de sexo de los encuestados y la variable nivel de aprendizaje despues de la asistencia al museo, donde antes de realizar este analisis pensamos que iba a ver una diferencia muy clara o notoria.

Del lado derecho es el registro de los hombres y del lado izquierdo el de mujeres, podemos decir que la condición de género no influye en el nivel de aprendizaje ocurrido en la visita al museo

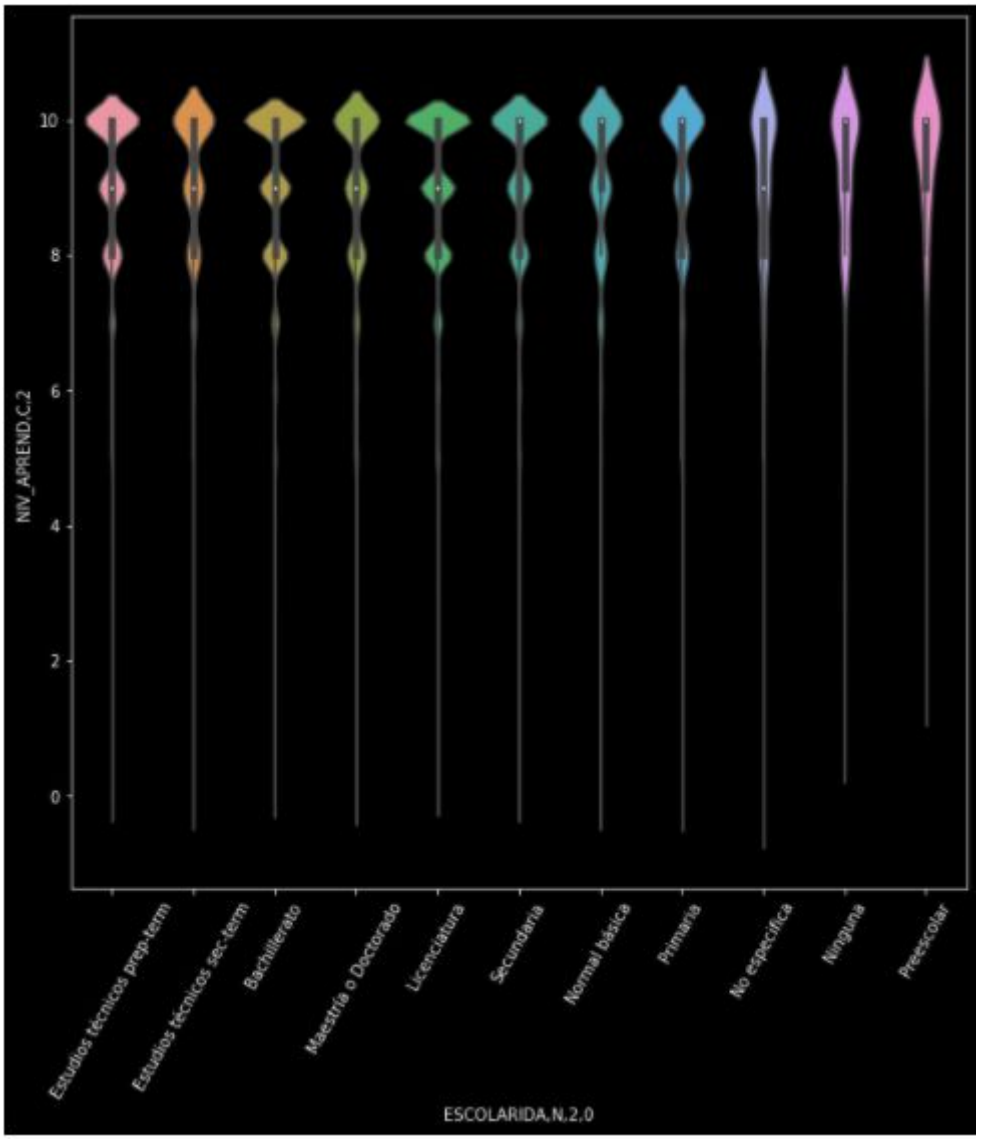


Figura 24. Diagrama de violín escolaridad y nivel de aprendizaje

Con esta grafica podemos decir que la condición de escolaridad no afecta en el aprendizaje de las personas a grandes rasgos, si consideremos lo picos de la gráfica podemos ver que tal vez el nivel de escolaridad hace que el nivel de aprendizaje cambie un poco por la forma de ver la exposición, es decir una persona con nivel de escolaridad de doctorado nota mas detalles en la información brindada por el museo y se hace otro tipo de cuestionamientos que una persona de nivel preescolar, pero a grandes rasgos todos salieron con un grado de aprendizaje alto del museo.

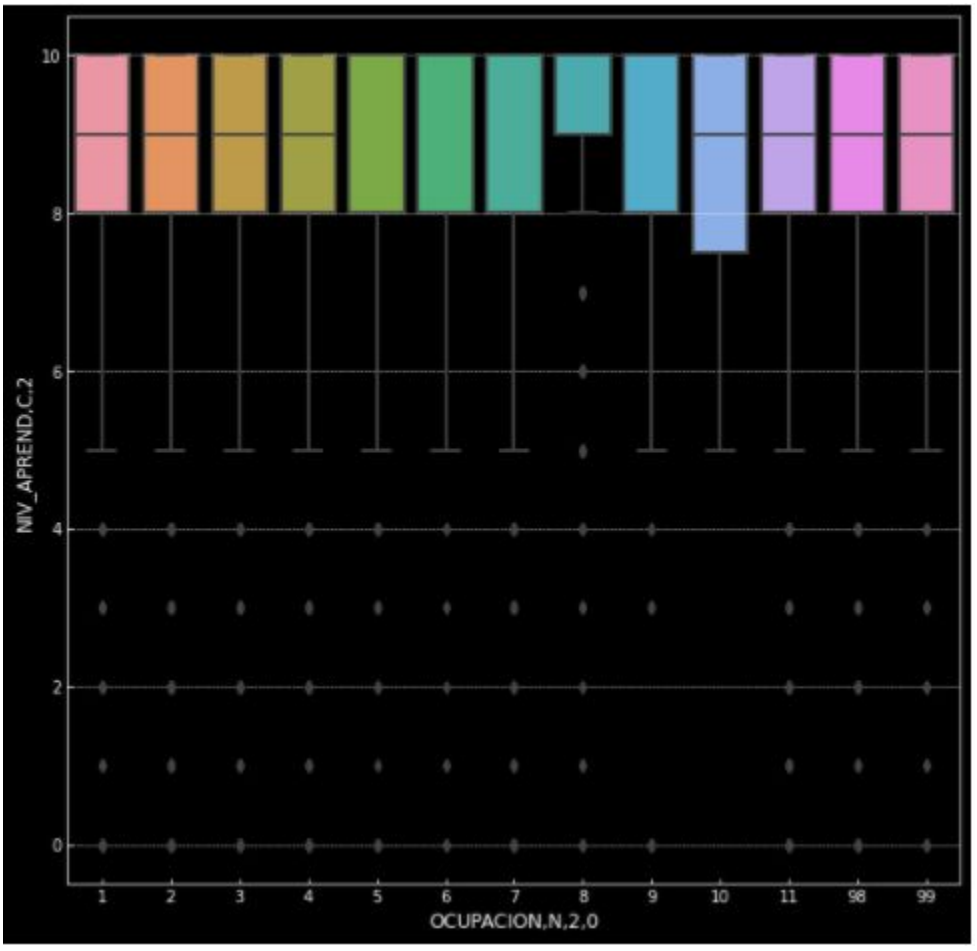


Figura 25. Diagrama de caja ocupación y nivel de aprendizaje

En esta gráfica se puede ver un poco el comportamiento similar que el nivel de escolaridad donde a grandes rasgos dice que la ocupación no influye en el nivel de aprendizaje del museo pero ocurre algo curioso analizando la ocupación "Operadores de maquinas industrial, ensambladores, choferes y conductores de transporte" debido a que su rango de aprendizaje es de 8 a 10 en contrario de las demás ocupaciones que su rango va de 5 a 10 obviamente con muy pocos registros.

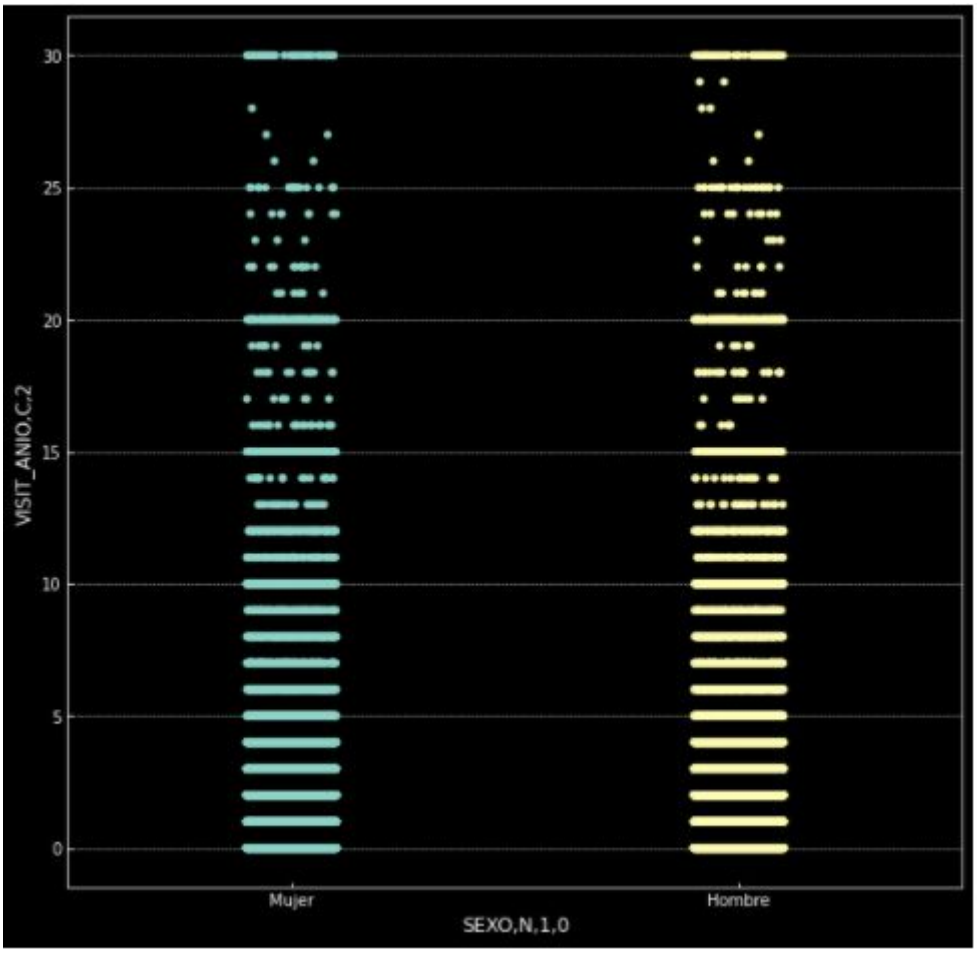


Figura 26. Diagrama puntos Visitas al año y Sexo

En esta grafica de puntos podemos observar dos variables el sexo y las visitas al año que realizaron y en muy pocos rangos de la grafica podemos ver que en ocasiones va mas veces que una mujer o viceversa, pero a grandes rasgos podemos concluir que no existe una distincion de genero para asistir al museo mas veces al año.

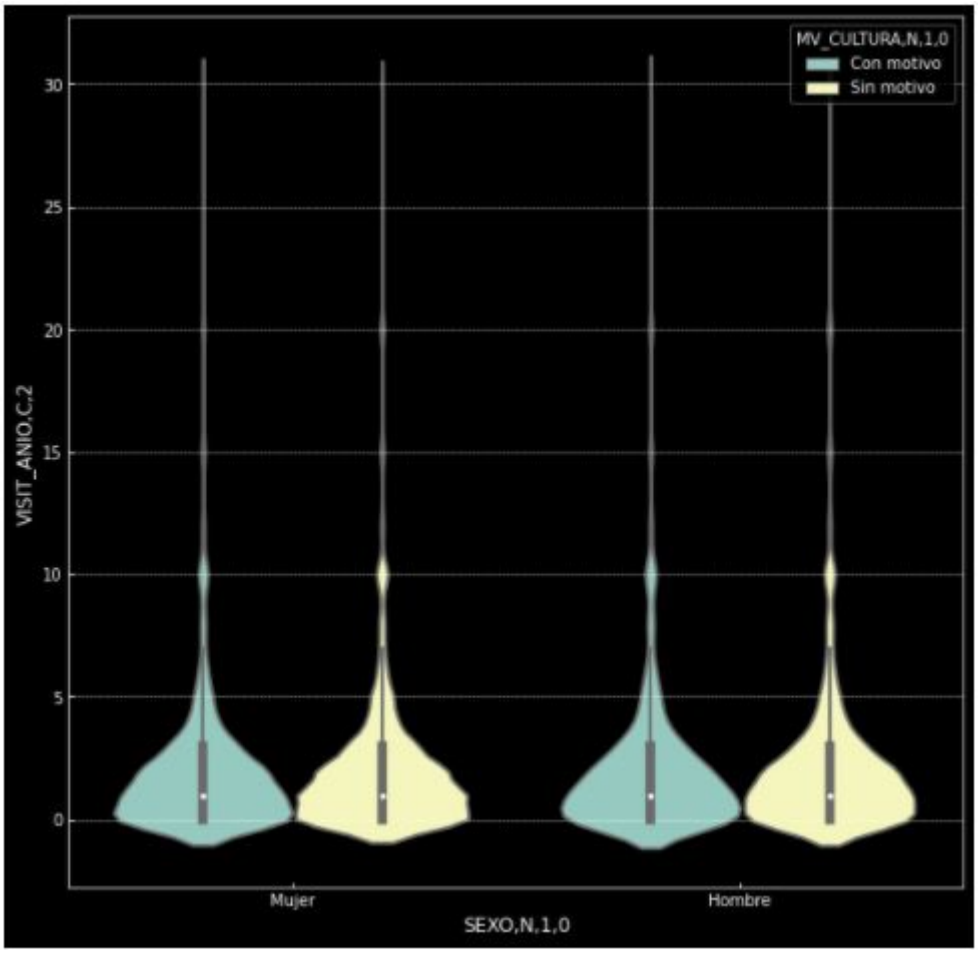


Figura 27. Diagrama violín Visitas al año y Sexo y Motivo de la visita cultura general

En este grafico ya comparamos 3 variables el sexo, el numero de visitas al año y el motivo de visista que sea por cultura general o no, de igual forma podemos ver el comportamiento similar en los generos, donde tanto como hombres y mujeres asisten el mismo o muy cercano numero de veces al museo al año en cambio el motivo de culturra general claramente esta muy clasificado, es decir la mitad va por cultura general y la otra mitad no.

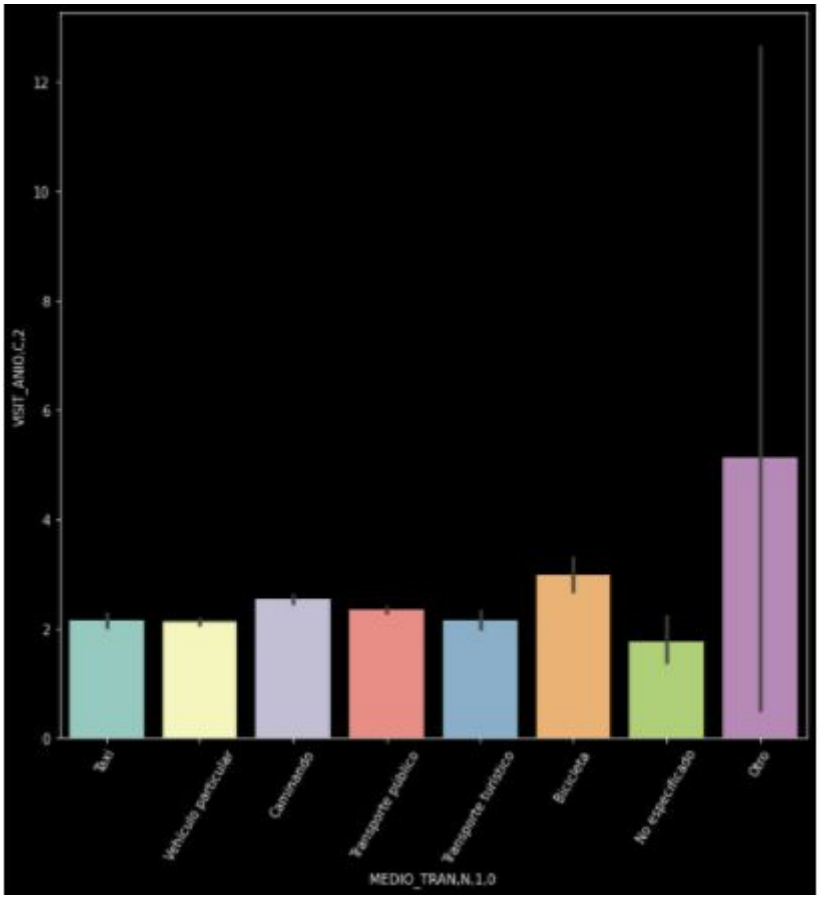


Figura 28. Gráfica de barras Visitas al año y medio de transporte utilizado

En este gráfico comparamos el medio de transporte con el número de visitas al año, podemos observar que varios registrados mencionan que otro transporte lo cual es un poco difícil de predecir cuales pero al considerar esta variable podemos hacer la comparación con con un gráfico de medios de transporte del punto A donde se clasifican diferente, incluso el transporte mas usado era el transporte particular, en cambio al agregar la variable de visitas al año vemos que la mayoría se transporta en bicicleta, a lo que podemos inferir que el museo queda un poco cerca de sus casas.

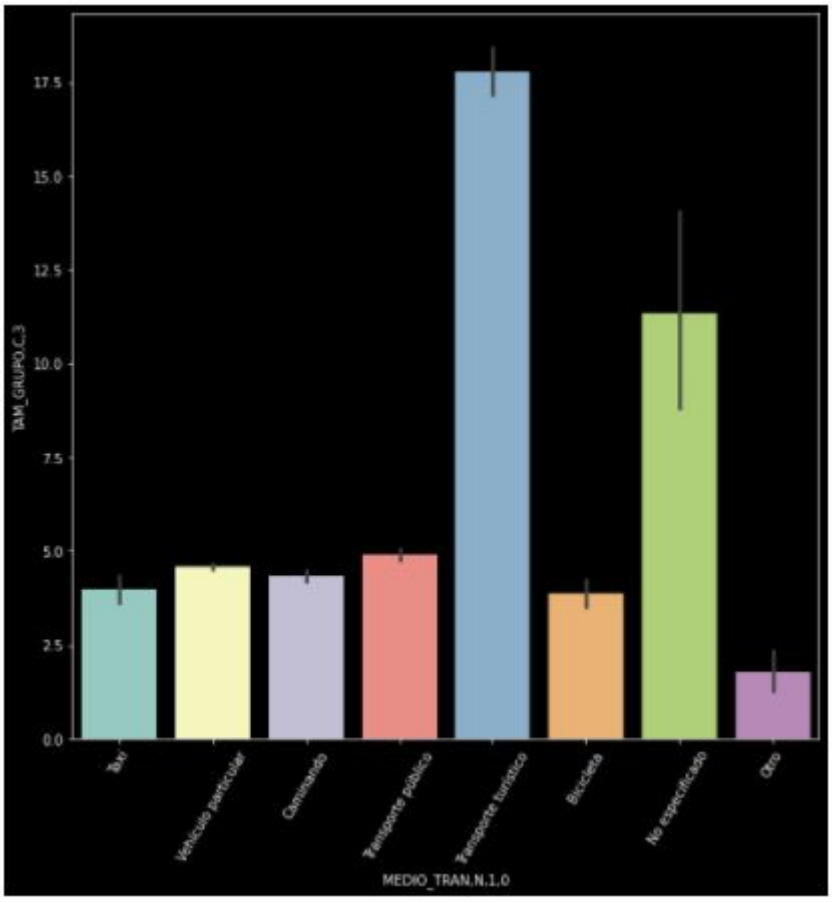


Figura 29. Gráfica de barras Medio de transporte utilizado y visitas al año

Como antes se mencionó la variable de tamaño de grupo, se podría decir que la mayoría llegó al museo en un transporte particular, pero no fue así, incluso podemos considerar que muchos grupos de turistas van al museo esto se debe al analizar la gráfica de transporte utilizado con tamaño de grupo donde vemos que la mayoría llego en transporte turístico, con grupos cercanos a las 18 personas

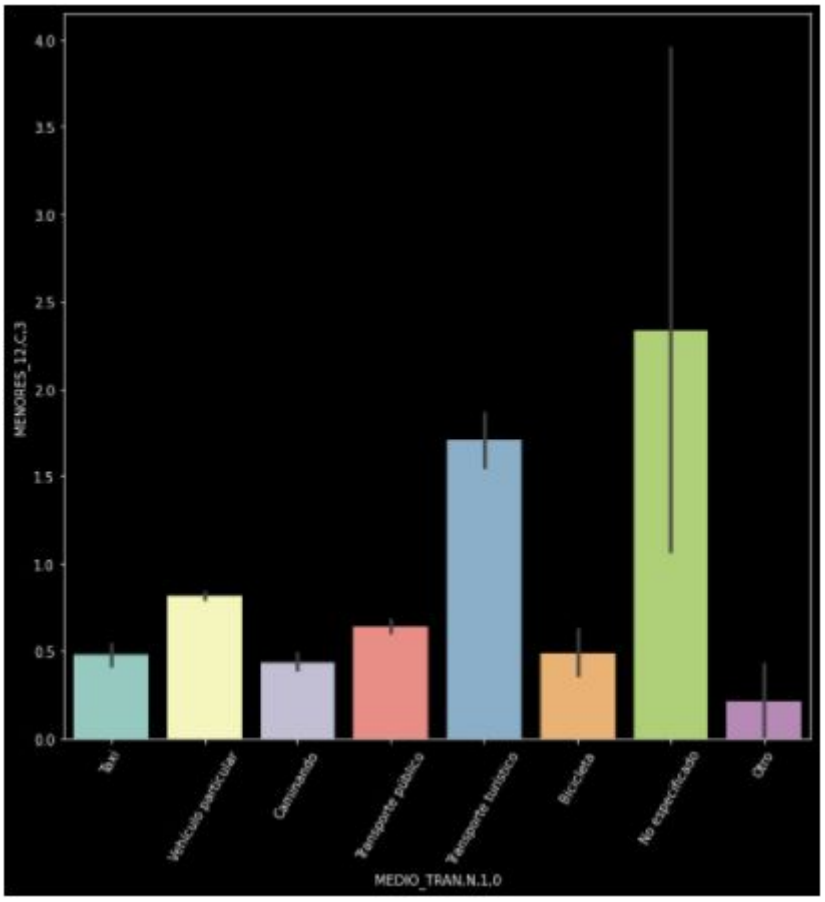


Figura 30. Gráfica de barras Medio de transporte utilizado y tamaño de grupo

En esta gráfica se hace la comparación de las variables de medio de transporte y los grupos menores de 12 años, y el medio de transporte utilizado, y muchos no especificaron, los demás fueron por medio de transporte turistico y otros mas en vehículo particular donde podemos dar un poco de veracidad a la teoría de que las escuelas realizan ese tipo de excursiones por sus propios medios o por parte de cada alumno.

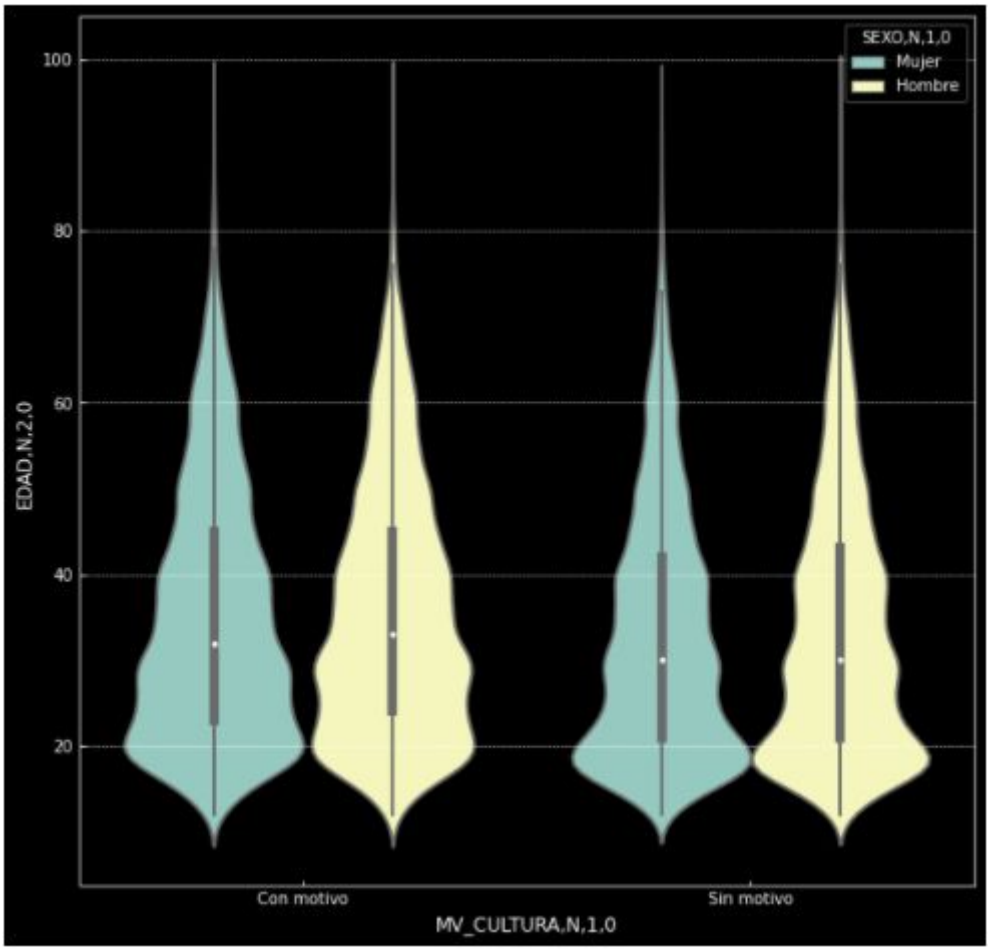


Figura 31. Diagrama violín Motivo de la visita cultura general y Edad

En este grafico comparamos las variables de Motivo de cultura general, las edades y el sexo de las personas, podemos observar que la mayor concentracion de los datos en edades de entre 17 a 21 años aproximadamente, y no hay mucha diferencia con el motivo de cultura general o no, pero si podemos observar el patron de que en un rango de edades de los 20 a los 40 años asisten mas a los museos debido a la facilidad que tienen para poder realizar esa accion de ir a museos.

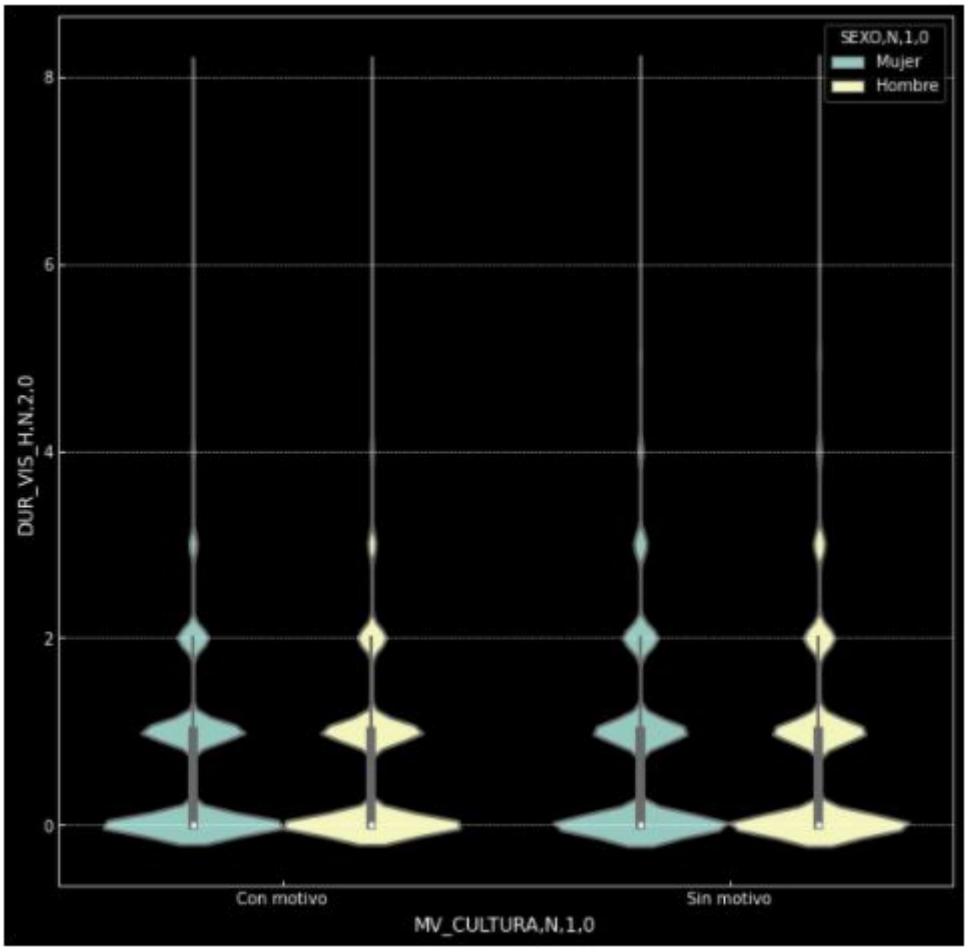


Figura 32. Diagrama violín Motivo de la visita cultura general y duración de la visita en horas

En esta grafica comparamos las variables motivo de cultura general y la duracion de visitas en horas y el sexo y observando muy a detalle podemos ver que algunos registrados que no tienen motivo de cultura general tardaron 4 horas en la visita, esto quiere decir que las personas que lo hacen por cultura general tarda menos horas en el museo (considerando esos detalles especificos) pero si se analiza a grandes rasgos no hay mucha diferencia.

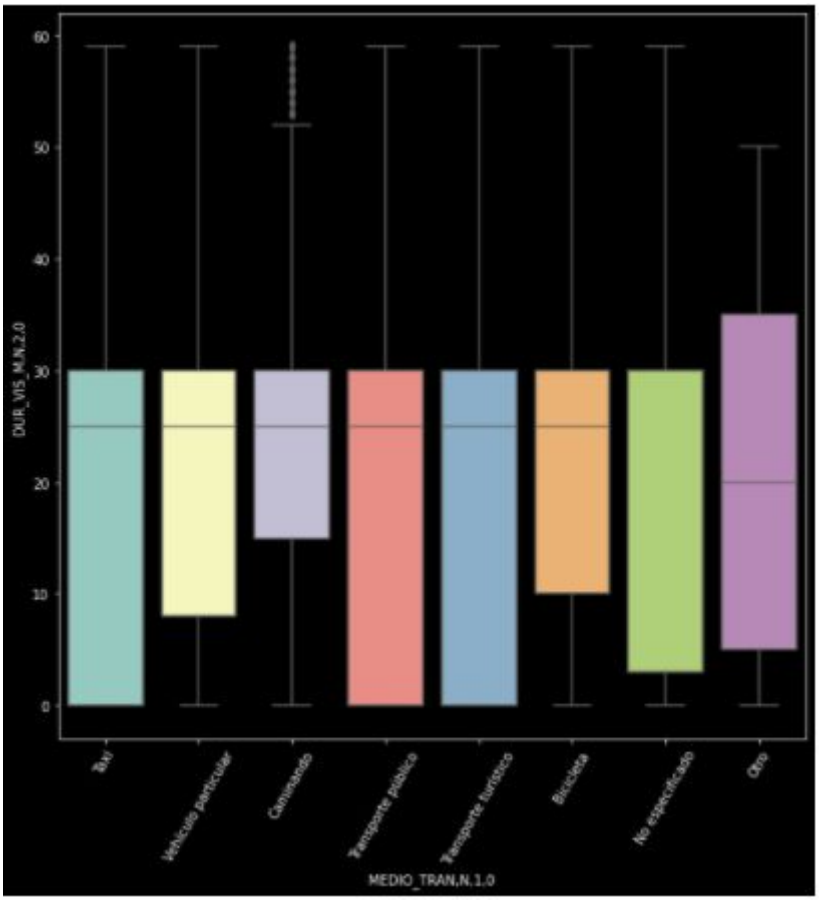


Figura 33. Diagrama de caja Medio de transporte en el que llegaron y duración de la visita en minutos

En esta gráfica analizamos el medio de transporte y la duración de visita en minutos, este gráfico nos brinda de mucha información ya que podemos ver que dependiendo del medio de transporte en el que se llega al museo influye en el tiempo que se está en el museo, por ejemplo los que llegan en transporte público o transporte turistico, van de 0 a 30 minutos, en cambio los que llegan caminando, bicicleta y vehículo particular su rango de estancia es diferente, estan de 10 a 30 minutos, podemos decir que eso influye por la facilidad que tienen para llegar y quedarse cierto tiempo en el museo.

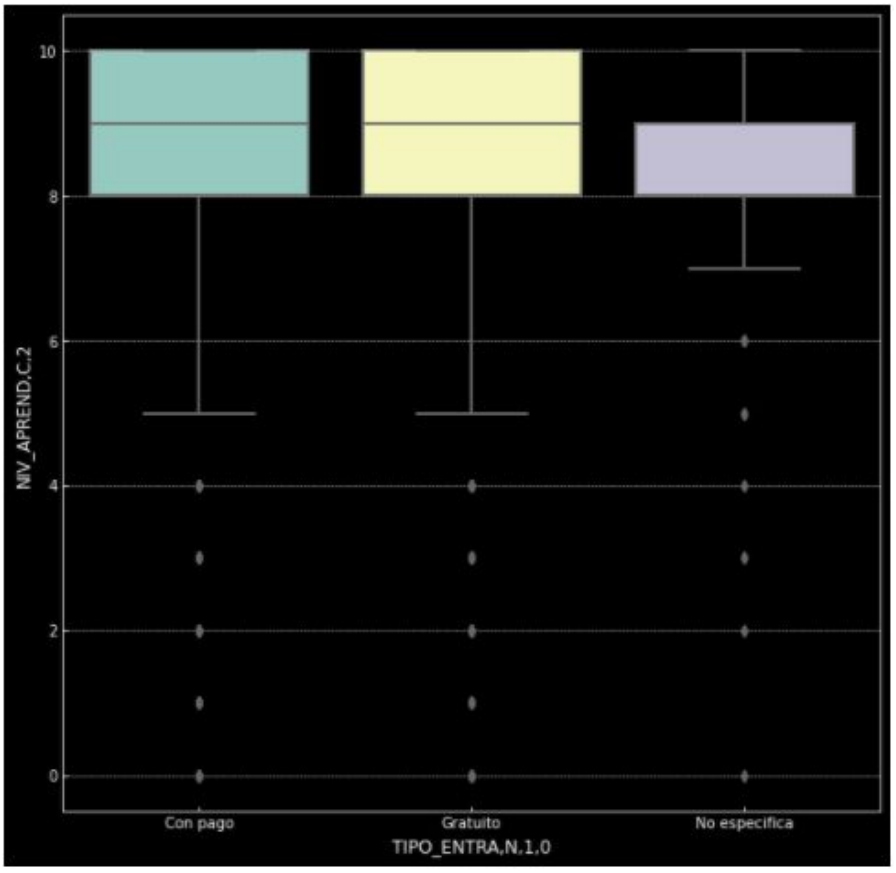


Figura 34. Diagrama de caja Tipo de entrada y Nivel de aprendizaje

Podemos observar que el tipo de entrada no influye en el nivel de aprendizaje

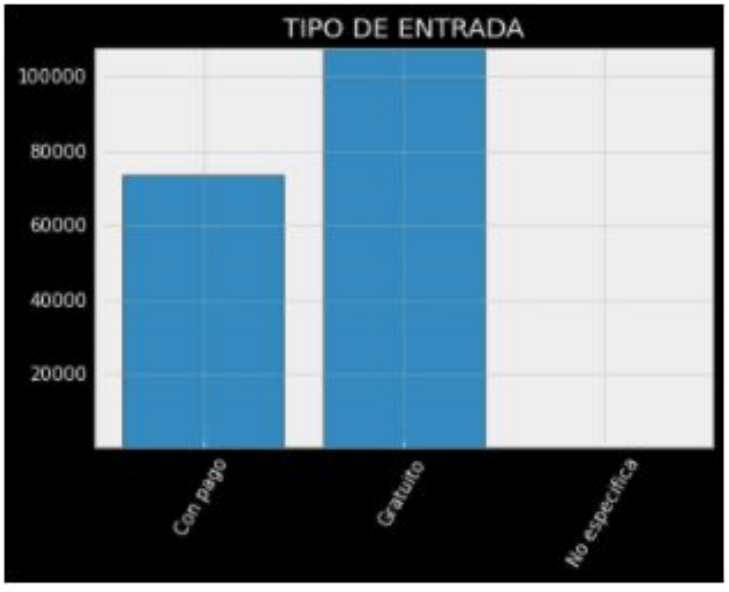


Figura 35. Gráfica de barras Tipo de entrada

En este gráfico podemos observar el tipo de entrada que los asistentes tuvieron para ingresar al museo y más de 100.000 asistentes fueron gratis y casi 80,000 asistentes entraron con pago.

Tablas de frecuencia acumulada

col_0	Frecuencia absoluta	Frecuencia relativa	Frecuencias absolutas acumuladas	Frecuencia relativa acumulada
EDAD,N,2,0				
12	911	1822.0	911	1822.0
13	1382	2764.0	2293	4586.0
14	1962	3924.0	4255	8510.0
15	3207	6414.0	7462	14924.0
16	4737	9474.0	12199	24398.0
...	...	...	...	...
92	9	18.0	180733	361466.0
93	6	12.0	180739	361478.0
95	2	4.0	180741	361482.0
96	5	10.0	180746	361492.0
97	1	2.0	180747	361494.0

Figura 36. Tabla de frecuencia acumulada de Edad.

col_0	Frecuencia absoluta	Frecuencia relativa	Frecuencias absolutas acumuladas	Frecuencia relativa acumulada
NIV_APREND,C,2				
0.0	469	938.0	469	938.0
1.0	183	366.0	652	1304.0
2.0	346	692.0	998	1996.0
3.0	613	1226.0	1611	3222.0
4.0	612	1224.0	2223	4446.0
5.0	2966	5932.0	5189	10378.0
6.0	3016	6032.0	8205	16410.0
7.0	9168	18336.0	17373	34746.0
8.0	33553	67106.0	50926	101852.0
9.0	42467	84934.0	93393	186786.0
10.0	85996	171992.0	179389	358778.0

Figura 37. Tabla de frecuencia acumulada de Nivel de Aprendizaje.

col_0	Frecuencia absoluta	Frecuencia relativa	Frecuencias absolutas acumuladas	Frecuencia relativa acumulada
VISIT_ANIO,C,2				
0.0	14229	28458.0	14229	28458.0
1.0	12717	25434.0	26946	53892.0
2.0	10036	20072.0	36982	73964.0
3.0	5435	10870.0	42417	84834.0
4.0	2367	4734.0	44784	89568.0
5.0	1643	3286.0	46427	92854.0
6.0	837	1674.0	47264	94528.0
7.0	423	846.0	47687	95374.0
8.0	515	1030.0	48202	96404.0
9.0	214	428.0	48416	96832.0
10.0	995	1990.0	49411	98822.0
11.0	114	228.0	49525	99050.0
12.0	251	502.0	49776	99552.0
13.0	50	100.0	49826	99652.0
14.0	34	68.0	49860	99720.0
15.0	317	634.0	50177	100354.0
16.0	25	50.0	50202	100404.0
17.0	14	28.0	50216	100432.0
18.0	32	64.0	50248	100496.0
19.0	16	32.0	50264	100528.0
20.0	323	646.0	50587	101174.0
21.0	16	32.0	50603	101206.0
22.0	18	36.0	50621	101242.0
23.0	7	14.0	50628	101256.0
24.0	17	34.0	50645	101290.0
25.0	40	80.0	50685	101370.0
26.0	4	8.0	50689	101378.0
27.0	3	6.0	50692	101384.0
28.0	3	6.0	50695	101390.0
29.0	2	4.0	50697	101394.0
30.0	127	254.0	50824	101648.0

Figura 38. Tabla de frecuencia acumulada de Visitas por año.

Enunciados (SQL)

1. Listado de personas que recomiendan y que repetirán visita al museo.
2. Listado de personas extranjeras que visitaron el museo.
3. Listado de personas que realizaron una visita por cultura general.
4. Listado de personas que realizaron una visita por motivos escolares.
5. Número de personas que utilizaron las salas de exhibición y la tienda.
6. Número de hombres y de mujeres que visitaron el museo.
7. Número de personas que entraron de manera gratuita.
8. Número de personas que entraron con pago.
9. Número de personas que hicieron una visita con guía.
10. Número de personas que dijeron que las exposiciones son muy buenas.
11. Número de personas que dieron un 10 en nivel de aprendizaje.
12. Número de personas que hicieron más de 8 horas en su visita.
13. Número de personas que evaluaron de manera general con 10.
14. Número de personas que realizaron una visita acompañados con un familiar.
15. Número de personas que realizaron una visita sin acompañante.
16. Número de personas que no recomiendan visitar el museo.
17. Número de personas que llegaron al museo en vehículo particular.
18. Número de personas que tardaron más de 30 minutos en llegar al museo.

Enunciados (Minería de datos)

1. Identificar el número de hombres y de mujeres que visitaron el museo y que recomiendan visitarlo.
2. Identificar de qué estado provienen las personas que les gustaron las exposiciones del museo y harán otra visita.
3. Identificar el número de personas que entraron con pago y que recomiendan visitar la institución museística.
4. Identificar el número de personas que hicieron una visita por cultura general y que no les gustaron las exposiciones.
5. Identificar el número de personas que hicieron una visita con motivo escolar y que evaluaron de manera general con 7.
6. Identificar el número de personas que llegaron en transporte público y que visitaron la exposición por más de 1 hora siendo extranjeros.
7. Identificar a los hombres que visitaron con motivo de la cultura y que se quedaron por 30 minutos en la exposición.

Referencias:

[1]: INEGI (2019) ONLINE Disponible en: https://www.inegi.org.mx/programas/museos/default.html#Microdatos

[2]: Islas Diez de Sollano Brandon (Octubre - 2020) ONLINE Disponible en: https://github.com/BrandonIslas/Proyecto1_Mineria_Datos