UNIVERSIDAD ICESI | A OTRO NIVEL

Análisis Exploratorio

Vacunaciones COVID-19

Laura Daniela Espinosa

Carlos Enrique Jaramillo

Agenda

- CONTEXTO
- PREGUNTA SMART
- Diccionario de datos
- Análisis Exploratorio
- Limpieza de Datos
- Análisis Estadístico (2)

Contexto

La pandemia de COVID-19 ha sido un reto sin precedentes para la salud pública global. A nivel mundial se han reportado “más de 767 millones de casos confirmados y más de 6,9 millones de muertes” hasta junio de 2023. La región de las Américas ha sido una de las más golpeadas con el 43% de las defunciones reportadas globalmente y más de 2.9 millones de defunciones hasta dicha fecha.

En Colombia la vacunación contra COVID-19 inició en febrero de 2021 dirigida inicialmente a grupos de riesgo y paulatinamente a la población general.

DATASET

Subconjunto de datos con información de 500.000 individuos vacunados de 4 ciudades de Colombia (Bogota, Cali, Medellin, Monteria).

NECESIDAD

Por recomendación de OMS, se busca estudiar diferencias epidemiológicas a nivel subnacional y las variaciones de la implementación de estrategias y políticas de vacunación.

Pregunta SMART

¿Existen diferencias significativas entre los individuos vacunados contra COVID-19 en 4 ciudades colombianas durante el periodo de 2021 a 2022 y sus desenlaces?

Specific: Se quiere saber si existen diferencias entre las personas que se vacunaron en las 4 ciudades colombianas.

Measurable: Se mide con pruebas estadísticas.

Action – oriented: Alcanzable con la información disponible y se motiva a realizar la investigación en otras ciudades o países.

Relevant: Se contribuye a la investigación con relación a las vacunas del COVID.

Time – bound: Se pretende estudiar en un plazo de 1 año.

Diccionario Datos

PersonaBasicaID: Se refiere al identificador de la persona vacunada, está totalmente anonimizado.

Sexo: El sexo de la persona vacuna.

Edad: Edad de la persona vacunada al momento de recibir la dosis.

FechaNacimiento: Corresponde a la fecha de nacimiento del individuo vacunado.

DepartamentoAplicacion: Departamento en donde recibo la dosis el individuo vacunado.

MunicipioAplicacion: Municipio en donde recibo la dosis el individuo vacunado.

Biológico: El biológico que se le aplico a la persona vacunada en cada una de las dosis.

FechaApliacion: Fecha cuando recibió la correspondiente dosis de la vacuna.

NumDosis: Indica cual dosis le fue suministrada a la persona vacunada.

CAC_VIH: Indica 1 si la persona vacunada tiene VIH, en caso contrario 0.

CAC_HTA: Indica 1 si la persona vacunada tiene Hipertensión Arterial, en caso contrario 0.

CAC_Diabetes: Indica 1 si la persona vacunada tiene Diabetes, en caso contrario 0.

CAC_PEH: Indica 1 si la persona vacunada tiene Enfermedades huérfanas, en caso contrario 0.

CAC_Cancer: Indica 1 si la persona vacunada tiene Cáncer, en caso contrario 0.

CAC_Artritis: Indica 1 si la persona vacunada tiene Artritis, en caso contrario 0.

Confirmado: Indica 1 si el individuo vacunado resulto positivo para COVID-19, en caso contrario 0.

FechaInicioSintomas: Es la fecha en la que el individuo resulto positivo.

ServicioMayorComplejidad: Indica 1 si la persona requirió internación en una entidad hospitalaria, 0 en caso contrario.

FechaIngresoServicioMayorComplejidad: Fecha en la que la persona fue internada en una entidad hospitalaria.

NDEstadoVital: Indica 1 si la persona falleció a causa del COVID-19, 0 en caso contrario.

NDFechaDefuncion: Fecha en la que la persona falleció por COVID-19.

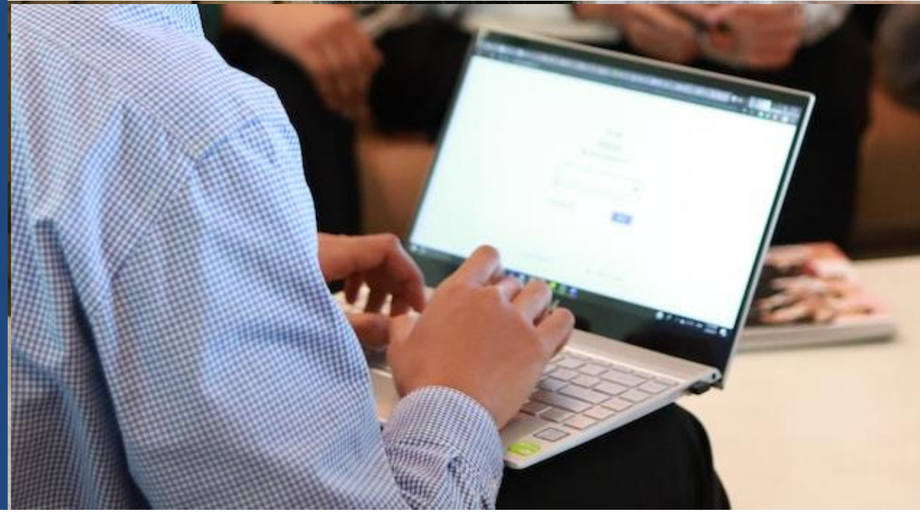
EstadoAfiliacion: Es el estado de afiliación del individuo vacunado al Sistema General de Seguridad Social en Salud.

Regimen: Es el régimen del individuo vacunado.

Análisis Exploratorio

- Estructura
- Variables categoricas
- Outliers

Estructura



Estandarización Fechas

DD-MM-YYYY

Valores Perdidos

Sexo 522
FechaNacimiento 68

* Edad (Calculo y Discretizacion)

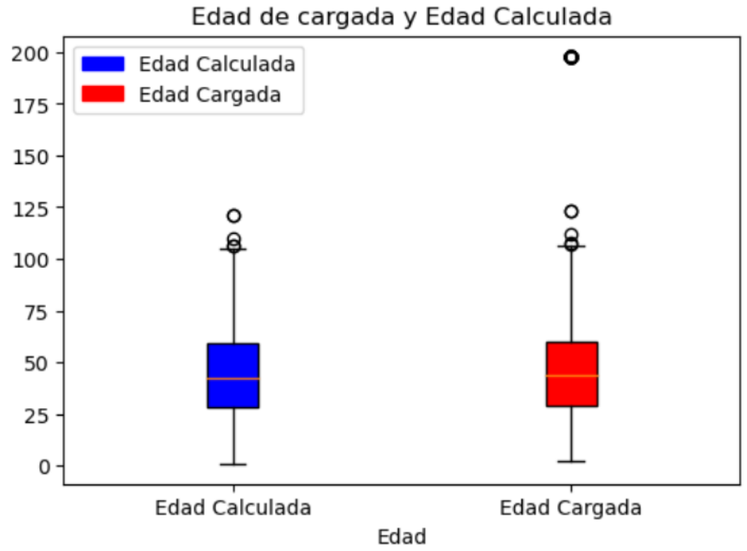
. random.choice()
.mean()

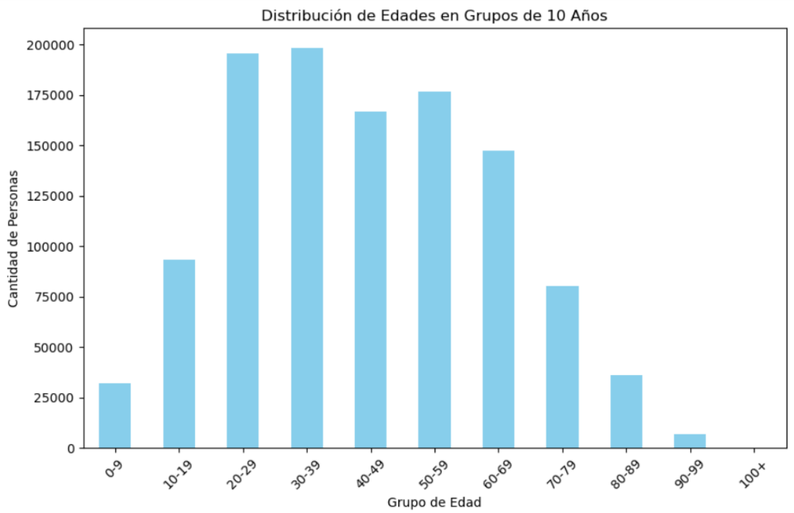
Normalización

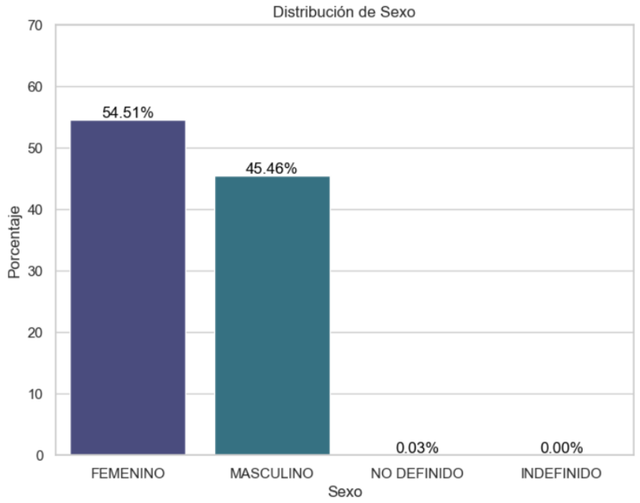
05001 – Medellín
05 - Antioquia

```
Value_count = df.isna().sum()
Value_Count_Mayor = Value_count[Value_count > 0]
print(f'Resumen:\r\n',Value_Count_Mayor)
```

```
Resumen:
 Sexo          522
GrupoEdad        3
dtype: int64
```





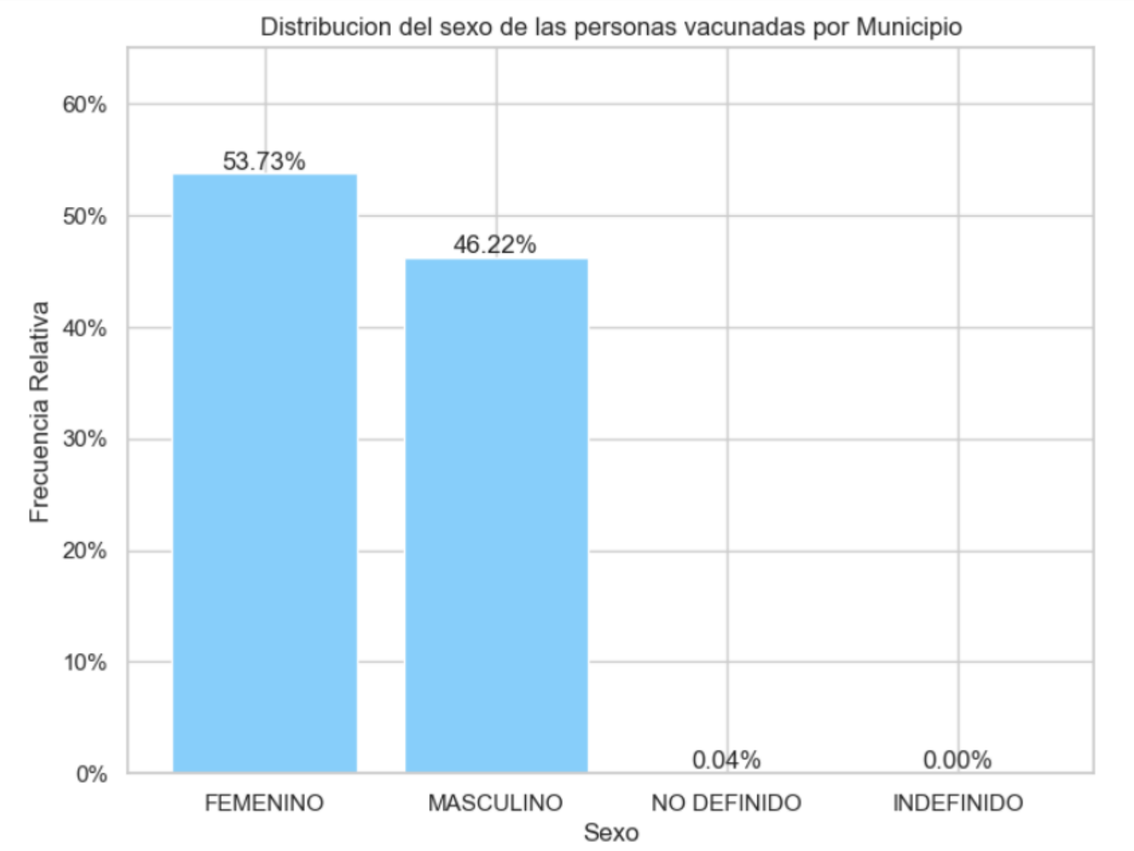


Análisis Exploratorio

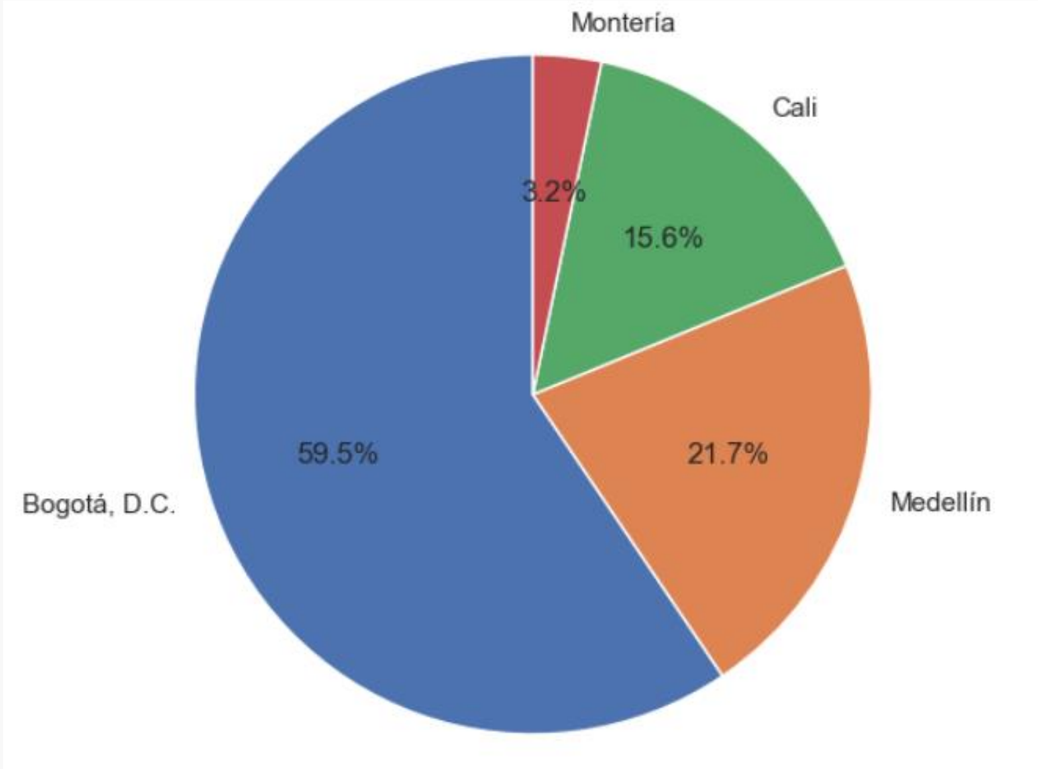
- Estructura
- Variables categoricas
- Outliers

Variables categóricas

Sexo
Requiere Limpieza



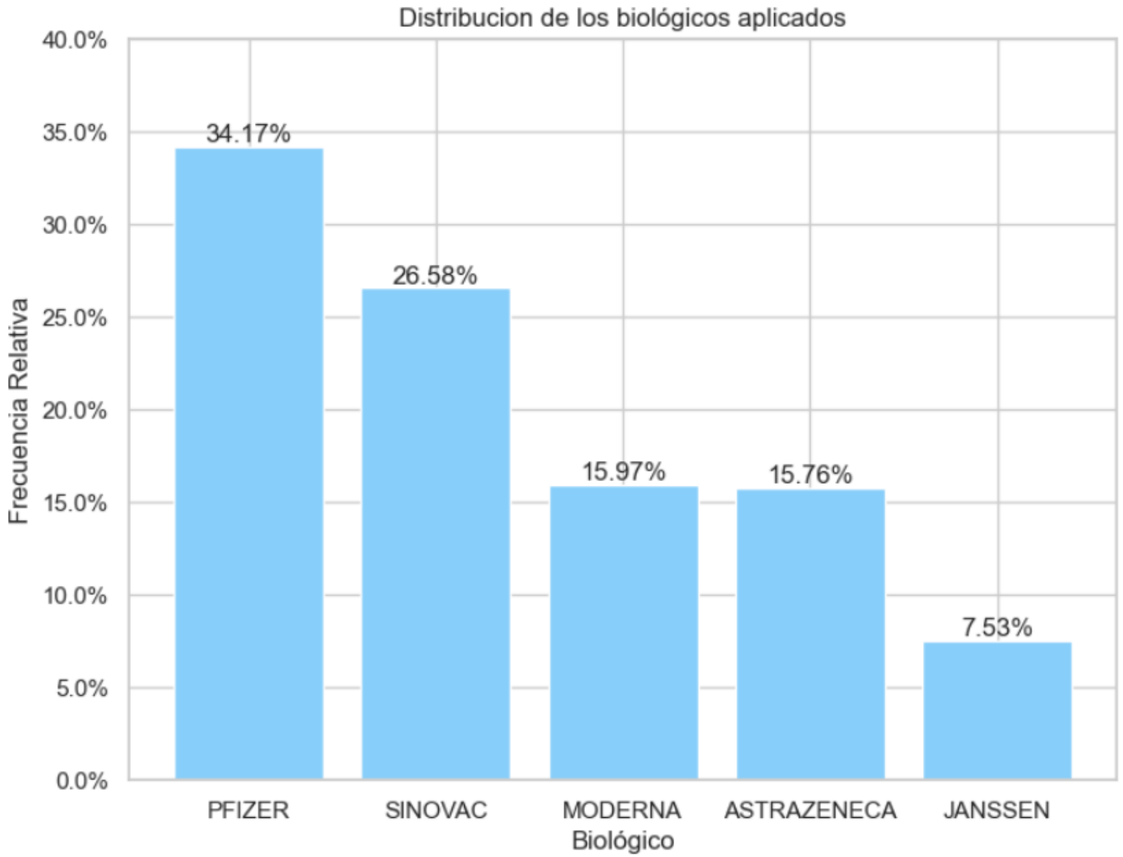
Municipio Aplicación
Por naturaleza de los datos se omite Departamento



Variables categóricas

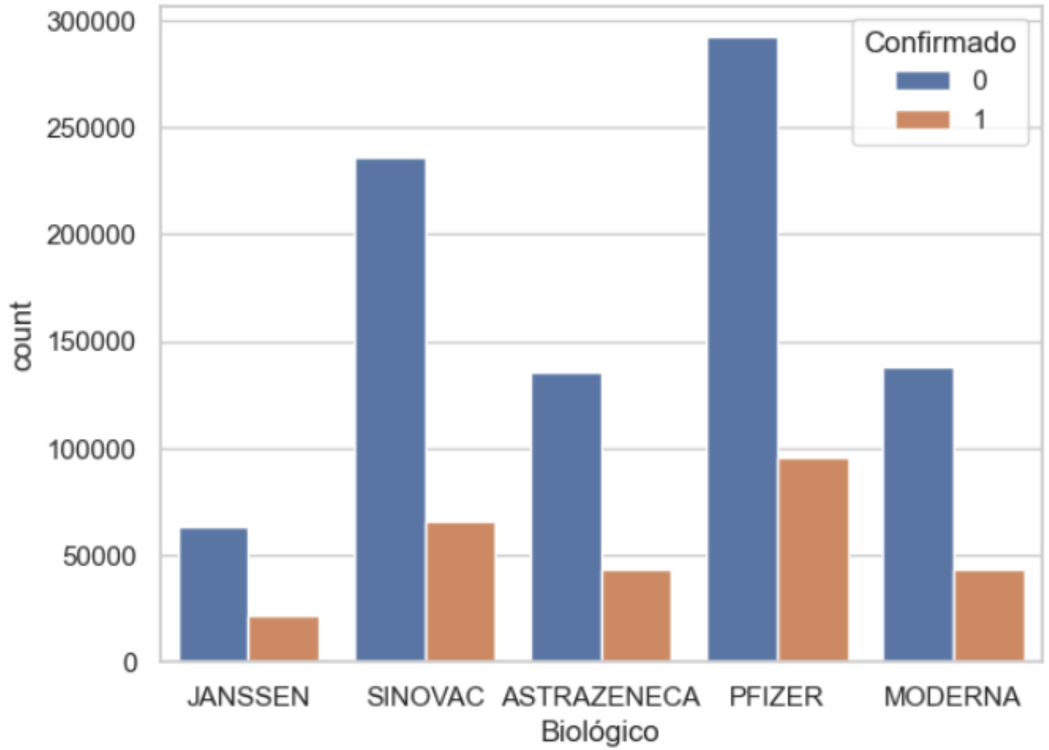
Biólogico

Class BiologicoTransformer - Estandarizar



Biológico

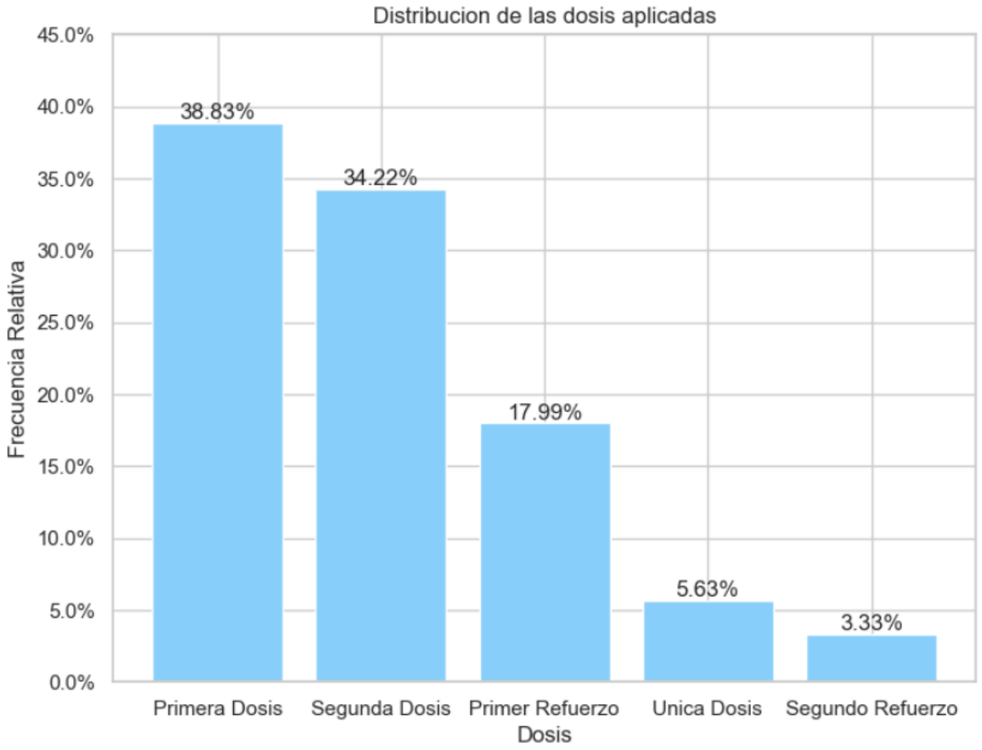
Casos confirmados que tienen Biológico



Variables categóricas

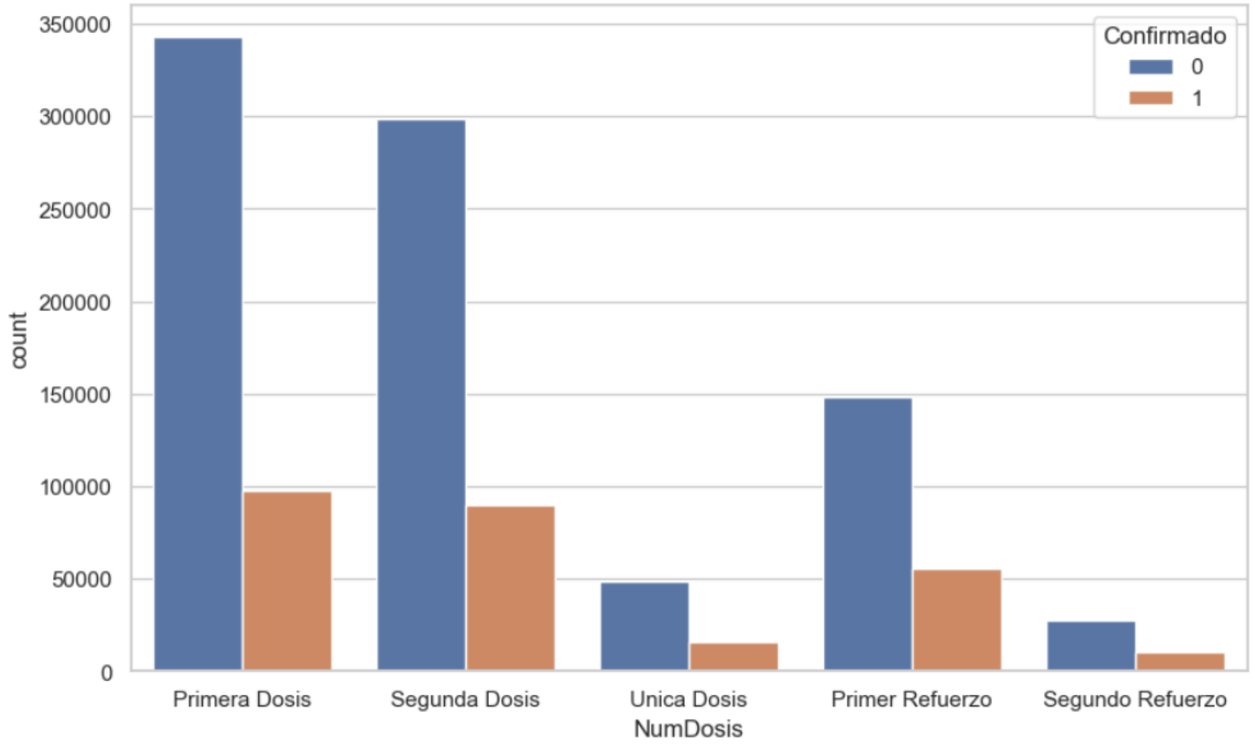
NumDosis

Class DosisTransformer - Estandarizar



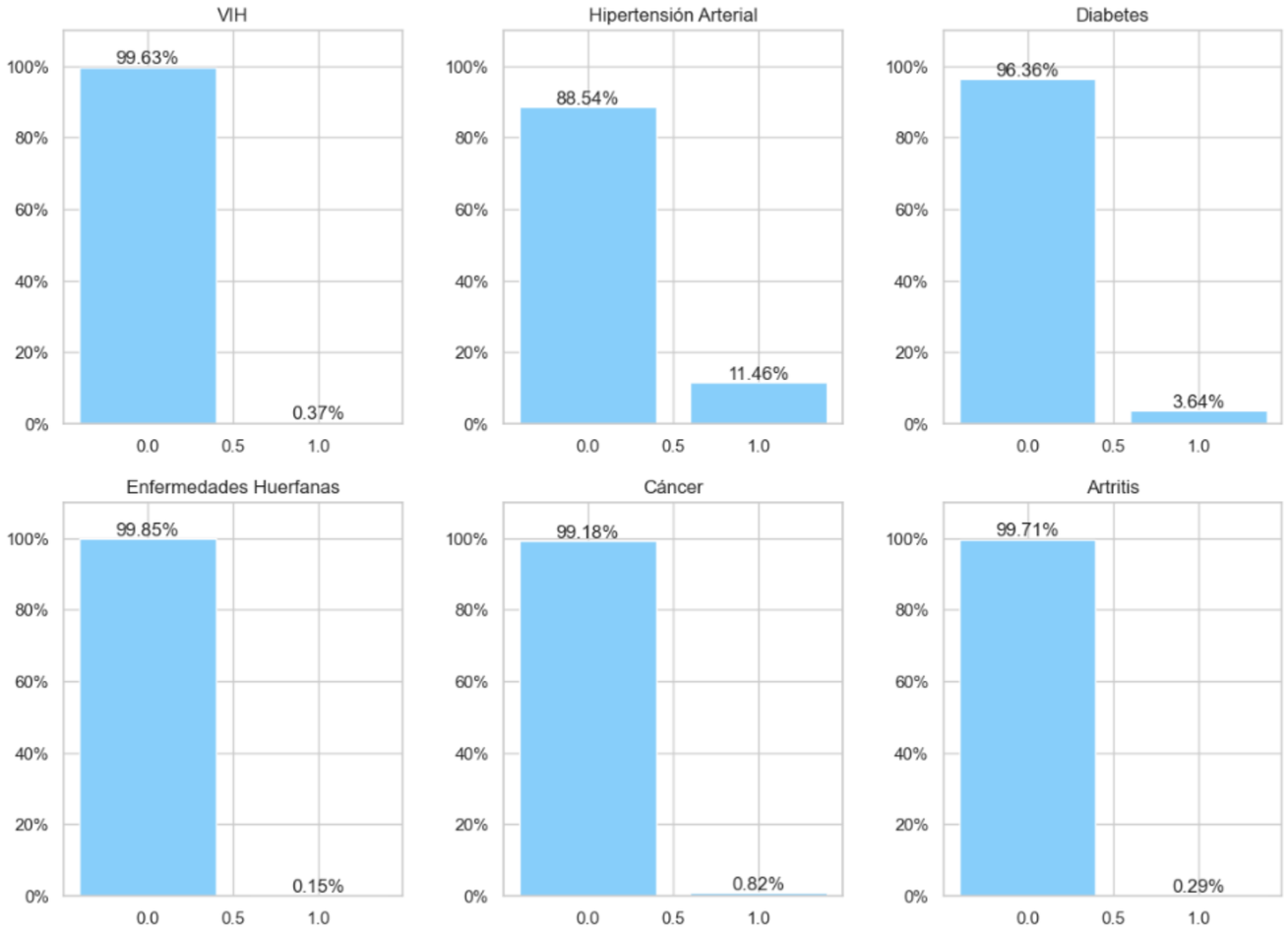
NumDosis

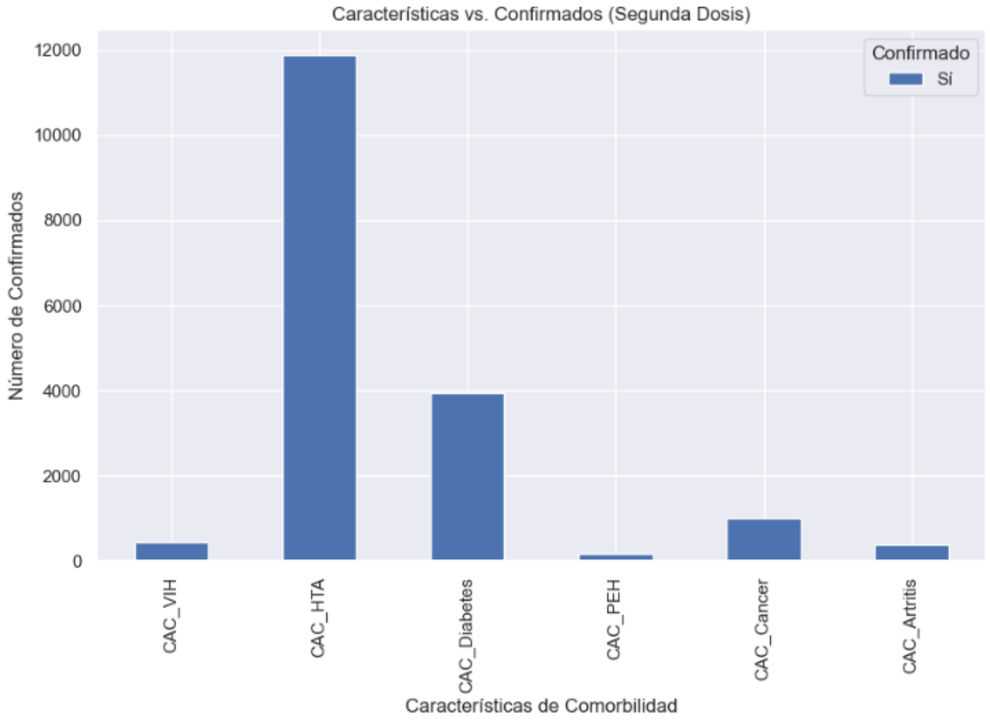
Casos confirmados con vacunación



Variables categóricas

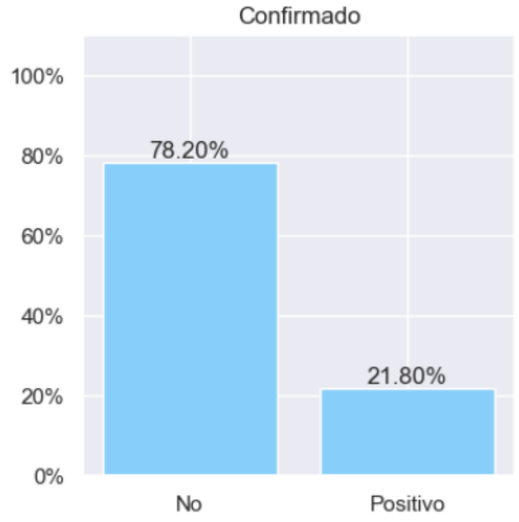
Comorbilidades

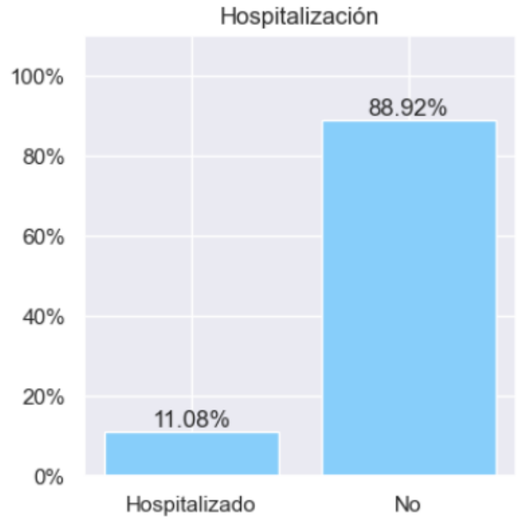




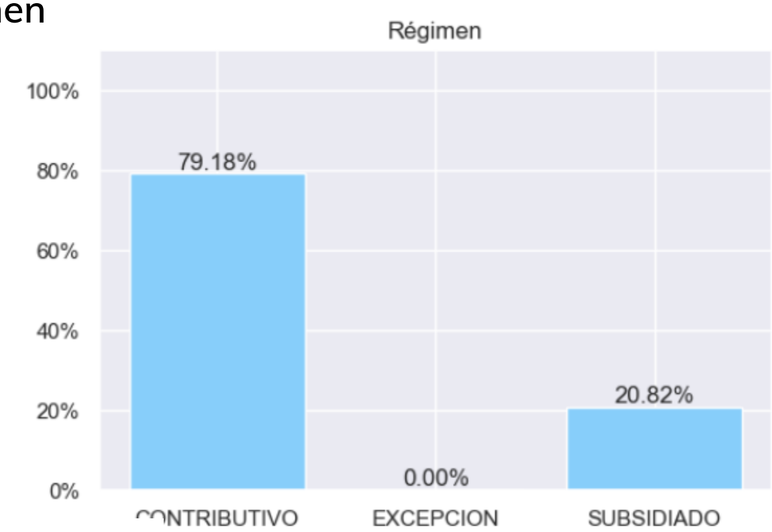
Variables categóricas

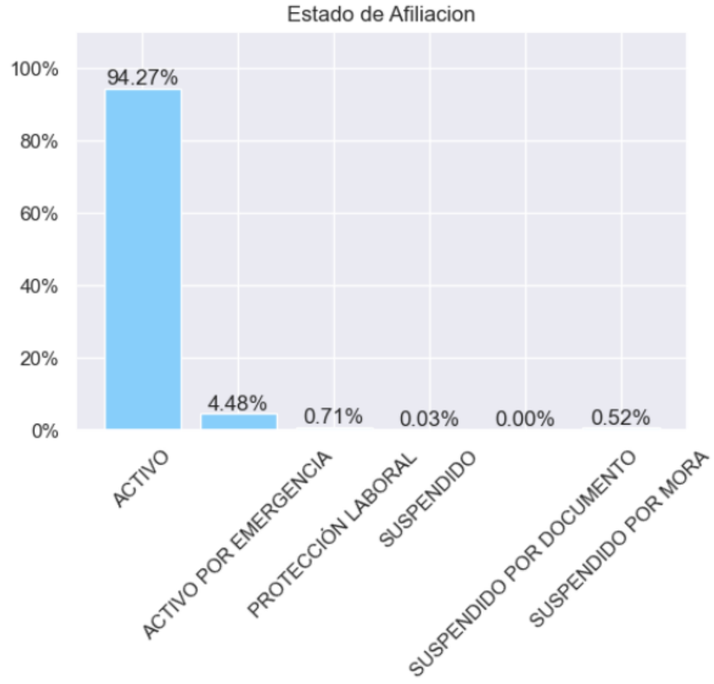
Desenlaces





Afiliación y Régimen





Análisis Exploratorio

Estructura

Variables categoricas

Outliers

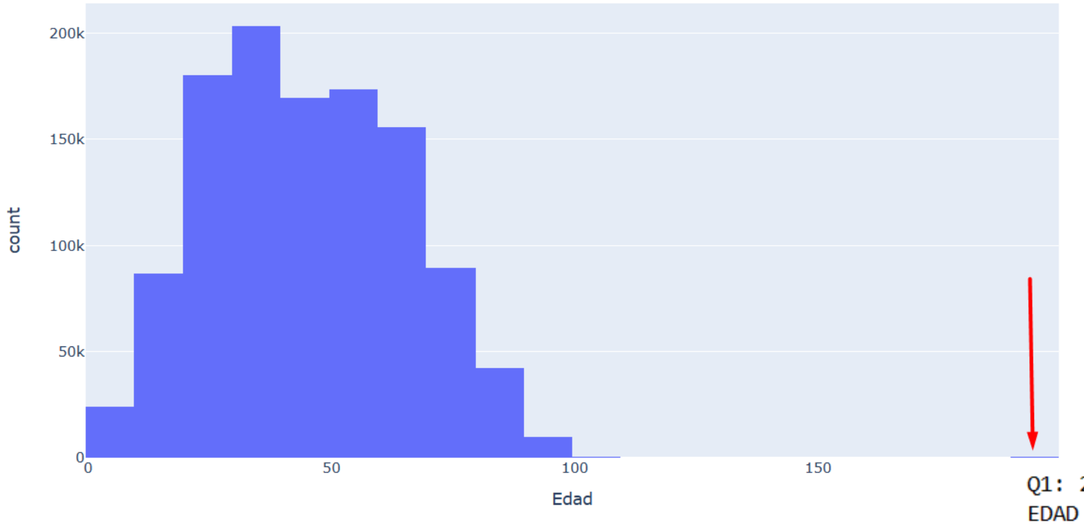
Outliers

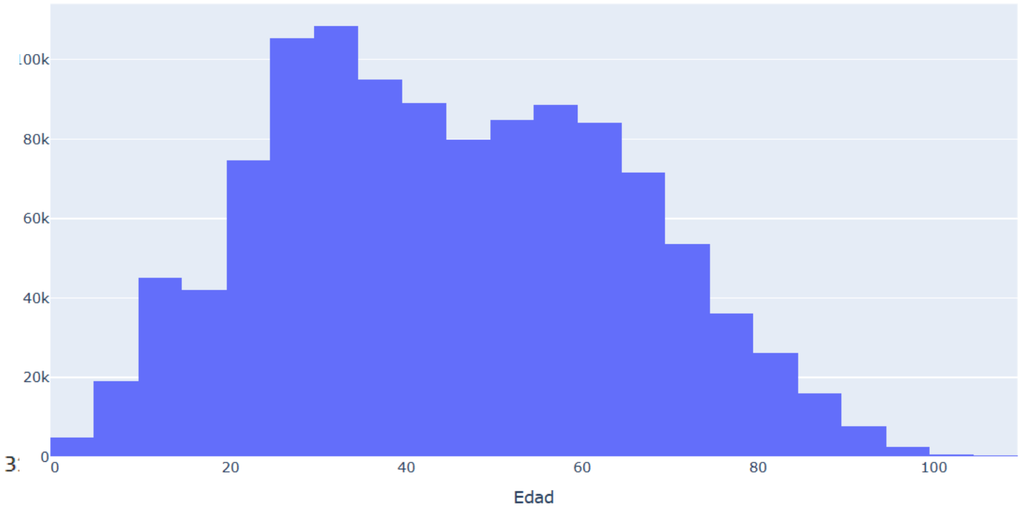
:

	PersonaBasicalD	Edad	NumDosis	CAC_VIH	CAC_HTA	CAC_Diabetes	CAC_PEH	CAC_Cancer	CAC_Artritis	EdadAplicacion
count	1,134,483.00	1,134,483.00	1,134,483.00	1,134,483.00	1,134,483.00	1,134,483.00	1,134,483.00	1,134,483.00	1,134,483.00	1,134,483.00
mean	55,942,567.09	44.91	0.83	0.00	0.15	0.05	0.00	0.01	0.00	43.49
std	43,540,300.43	20.12	1.18	0.07	0.35	0.21	0.04	0.10	0.06	19.99
min	2.00	2.00	-2.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
25%	20,773,274.50	29.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.00
50%	41,547,199.00	44.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	42.00
75%	100,005,058.00	60.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	59.00
max	146,087,548.00	198.00	2.00	1.00	1.00	1.00	1.00	1.00	1.00	121.00

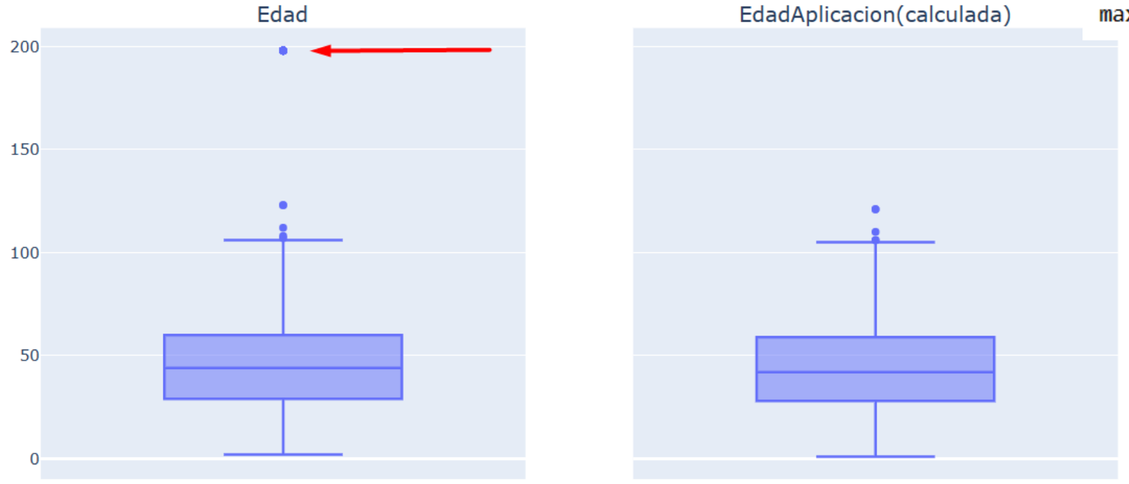
```python
#funcion para encontrar IQR (rango intercuartil)
def find_outlier_IQR(df):
    q1=df.quantile(0.25)
    q3=df.quantile(0.75)
    IQR = q3-q1
    print(f'Q1: {q1}, Q3: {q3}, IQR: {IQR}')
    outliers = df[ ((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR))) ]
    return outliers
```

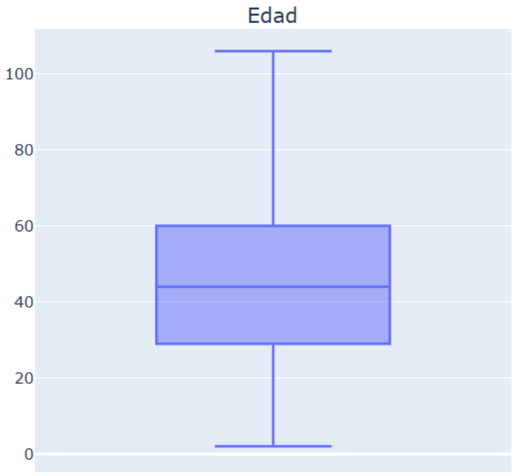
Outliers

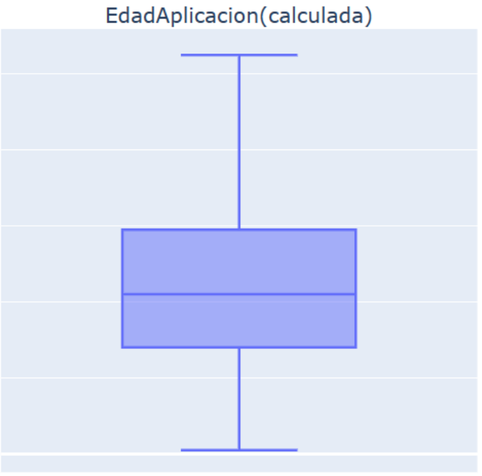





```
Q1: 29.0, Q3: 60.0, IQR: 3
EDAD
número de outliers: 73
min valor outlier: 107
max valor outlier: 198
Q1: 28.0, Q3: 59.0, IQR: 31.0
EDAD_APLICACION
número de outliers: 5
min valor outlier: 106
max valor outlier: 121
```



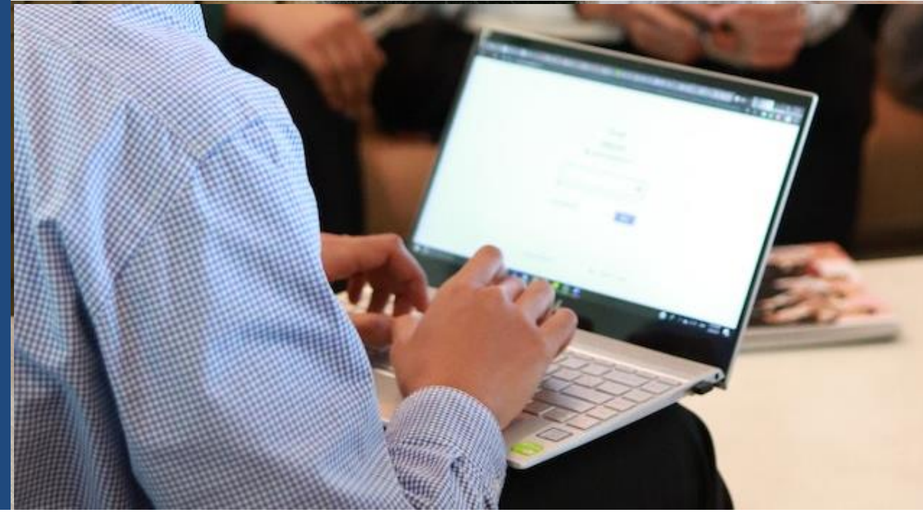




Limpieza de Datos

- Estandarización
- Imputación

Limpieza de Datos



Estandarización

.str.replace()

```
Sexo
FEMENINO        54.51%
MASCULINO       45.47%
NO DEFINIDO      0.03%
INDEFINIDO       0.0%
dtype: object
```

```
Sexo
FEMENINO        54.51%
MASCULINO       45.47%
INDEFINIDO       0.03%
dtype: object
```

Imputación

.random.choice()

```
Sexo
FEMENINO       618081
INDEFINIDO        291
MASCULINO      515550
Name: Sexo, dtype: int64
```

```
Sexo
FEMENINO       618349
INDEFINIDO        291
MASCULINO      515770
Name: Sexo, dtype: int64
```

```
df['Sexo'].unique()
```

```
array(['MASCULINO', 'FEMENINO', nan, 'NO DEFINIDO', 'INDEFINIDO'],
      dtype=object)
```

```
valores_nulos_por_columna = df.isna().sum()
columnas_con_valores_no_nulos = valores_nulos_por_columna[valores_nulos_por_columna != 0]
print(columnas_con_valores_no_nulos)
```

```
Sexo    488
dtype: int64
```

```
# Calcula la distribución de los valores existentes en la columna 'Sexo'
sexo_distribution = df['Sexo'].value_counts(normalize=True)

# Imputa los valores NaN en función de la distribución
df['Sexo'].fillna(pd.Series(np.random.choice(sexo_distribution.index, p=sexo_distribution.values, size=len(df))), inplace=True)

# Verifica que los valores NaN se hayan reemplazado correctamente
print(df['Sexo'].unique())
```

```
['MASCULINO' 'FEMENINO' 'INDEFINIDO']
```

Análisis Estadístico

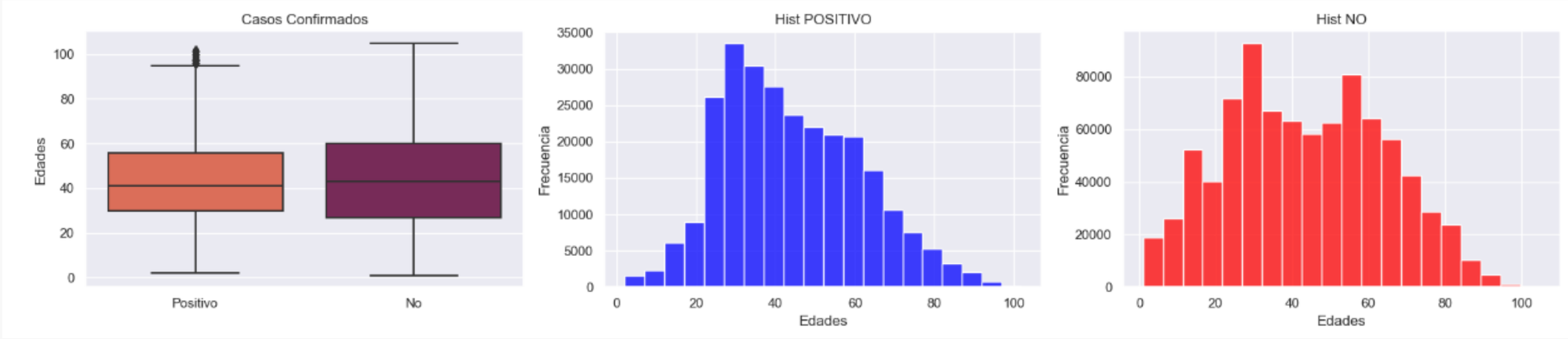
- Caso 1 : Relación entre edad y desenlace.

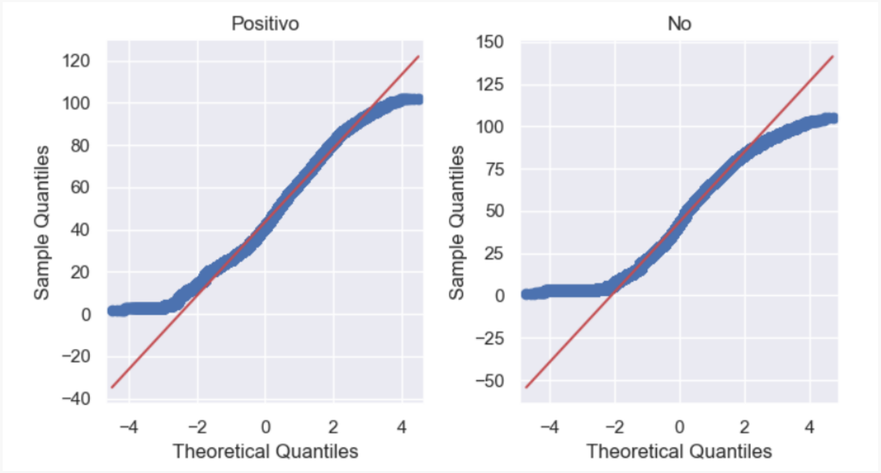
- Caso 2: Relación entre desenlace y sexo.

Caso 1 : Relación edad - desenlace

CONFIRMADO

CONFIRMADO






```
Población: Confirmado Positivo
Estadístico SW= 0.9806710481643677, Valor-p= 0.0
Estadístico AD= 1803.0124239556608, Valor crítico (sign. 5%)= 0.787

Población: No Confirmado positivo
Estadístico SW= 0.985482931137085, Valor-p= 0.0
Estadístico AD= 4039.2869605114684, Valor crítico (sign. 5%)= 0.787
```

Normalidad: Rechazo Hipotesis de normalidad

```
Estadístico W= 7592.293172863293, Valor-p= 0.0
```

Homogeneidad de Varianza: Rechazo Hipotesis de homogeneidad

Prueba U de Mann-Whitney

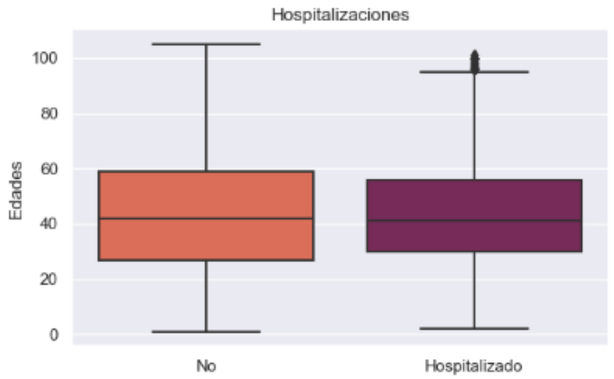
```
Estadístico W= 22637133770.5, Valor-p= 3.740289196574841e-217
```

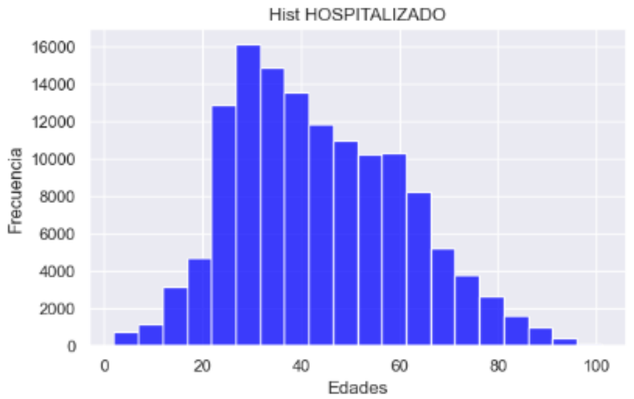
Se Rechaza la hipótesis, por tanto hay diferencia suficiente evidencia para decir que hay diferencia entre confirmados y no confirmados.

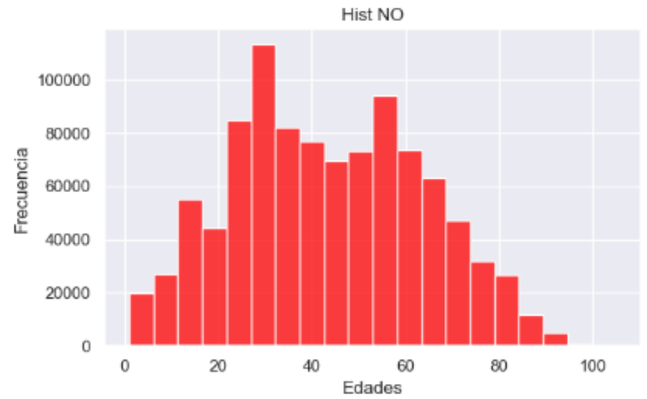
Caso 1 : Relación edad - desenlace

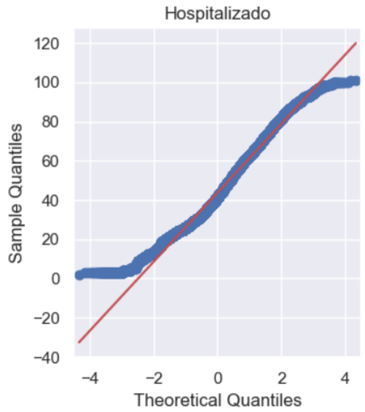
HOSPITALIZADO

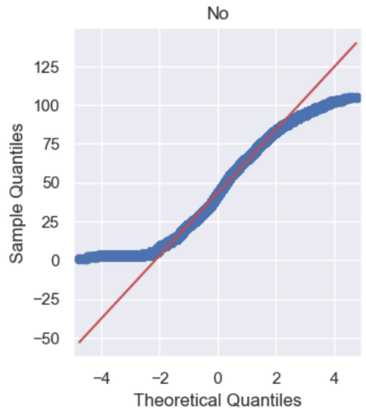
HOSPITALZIADO












```
Población: Hospitalizado
Estadístico SW= 0.9816091060638428, Valor-p= 0.0
Estadístico AD= 837.409109200089, Valor crítico (sign. 5%)= 0.787

Población: No Hospitalizado
Estadístico SW= 0.9861602783203125, Valor-p= 0.0
Estadístico AD= 4321.834139822866, Valor crítico (sign. 5%)= 0.787
```

Normalidad: Rechazo Hipotesis de normalidad

```
Estadístico W= 6599.015596648111, Valor-p= 0.0
```

```
Estadístico t= 3.925659380989586, Valor-p= 8.65238266075685e-05
```

T-Test

Homogeneidad de Varianza: Rechazo Hipotesis de homogeneidad

Prueba U de Mann-Whitney

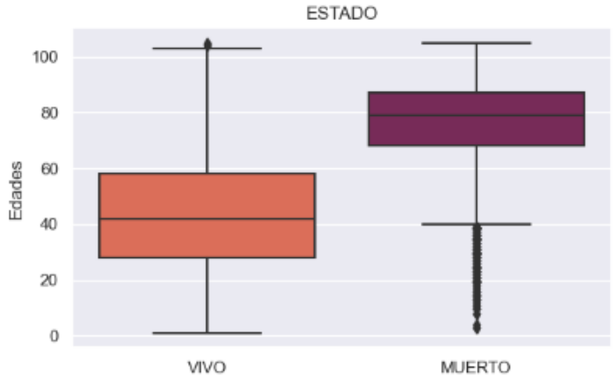
```
Estadístico W= 13029501818.0, Valor-p= 2.85443841538703e-110
```

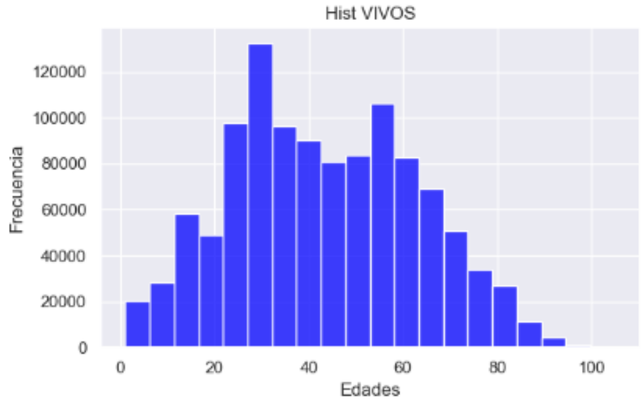
Se Rechaza la hipótesis , por tanto, hay diferencia significativa entre hospitalizados y no hospitalziados

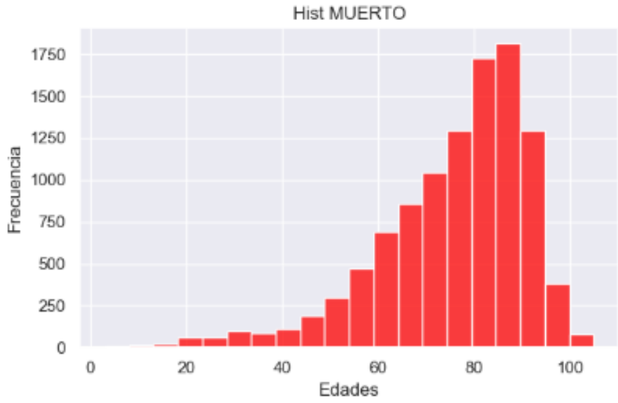
Caso 1 : Relación edad - desenlace

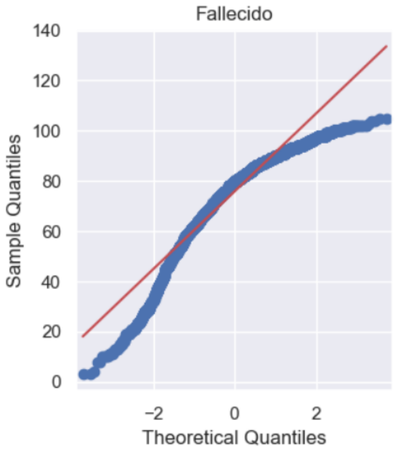
MUERTE

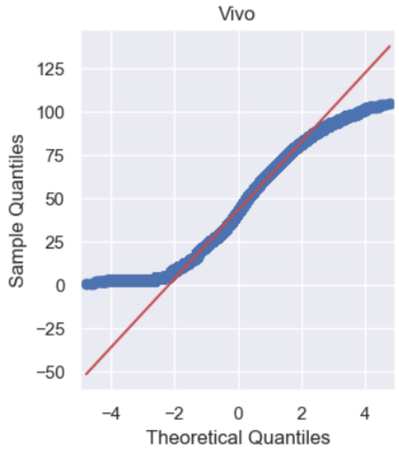
MUERTE












```
Población: Fallecido
Estadístico SW= 0.9159862995147705, Valor-p= 0.0
Estadístico AD= 222.8300510199515, Valor crítico (sign. 5%)= 0.787

Población: Vivo
Estadístico SW= 0.9839651584625244, Valor-p= 0.0
Estadístico AD= 4658.138550783973, Valor crítico (sign. 5%)= 0.787
```

Normalidad: Rechazo Hipotesis de normalidad

```
Estadístico W= 2044.7263386218026, Valor-p= 0.0
```

Levene

```
Estadístico t= 215.46345152659362, Valor-p= 0.0
```

T-Test

Homogeneidad de Varianza: Rechazo Hipotesis de homogeneidad

Prueba U de Mann-Whitney

```
Estadístico W= 10688816539.0, Valor-p= 0.0
```

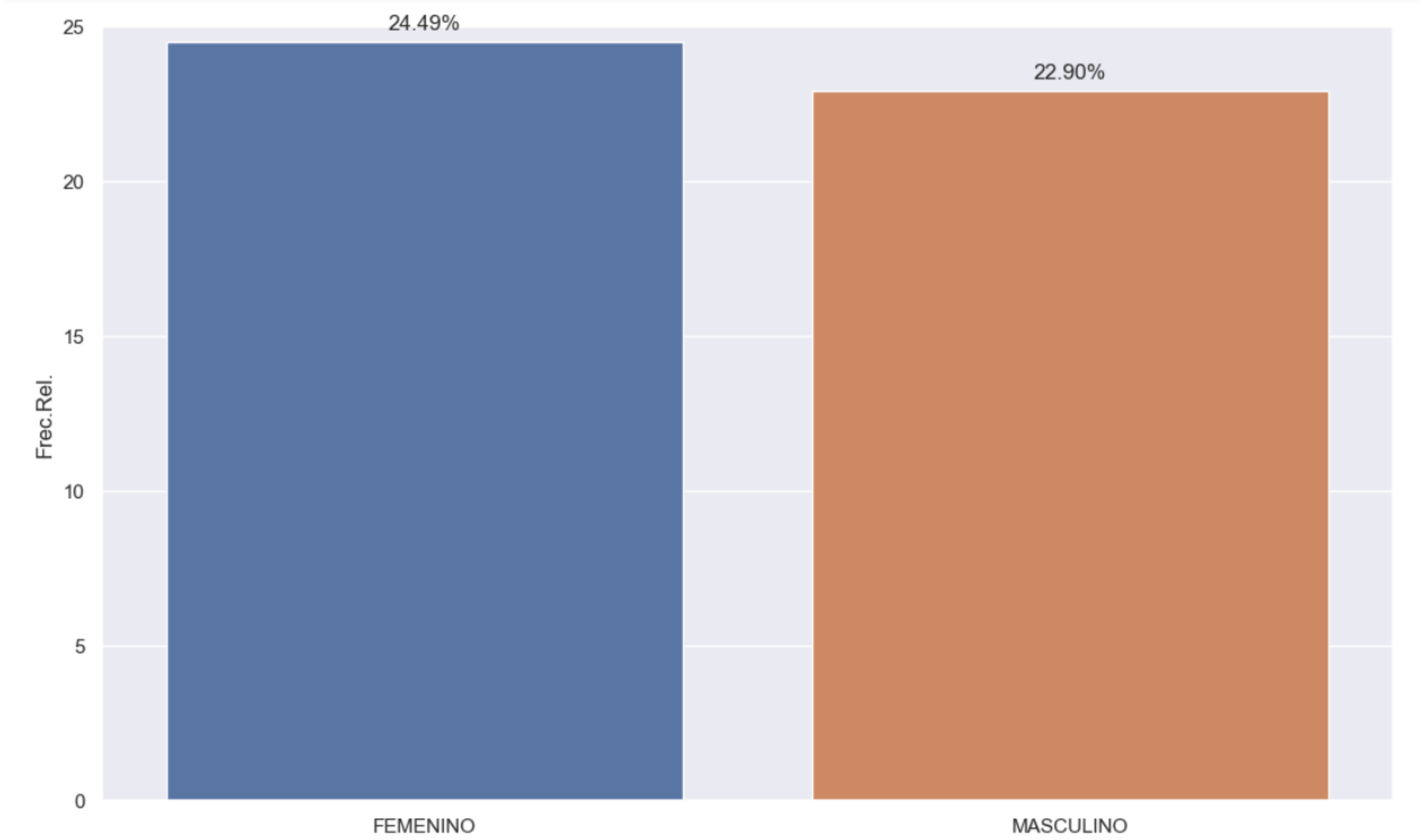
Se Rechaza la hipótesis , por tanto, hay diferencia significativa entre Fallecidos y no Fallecidos.

Análisis Estadístico

Caso 1 : Relación entre edad y desenlace.

Caso 2: Relación entre desenlace y sexo.

CONFIRMADO



Prueba χ^2

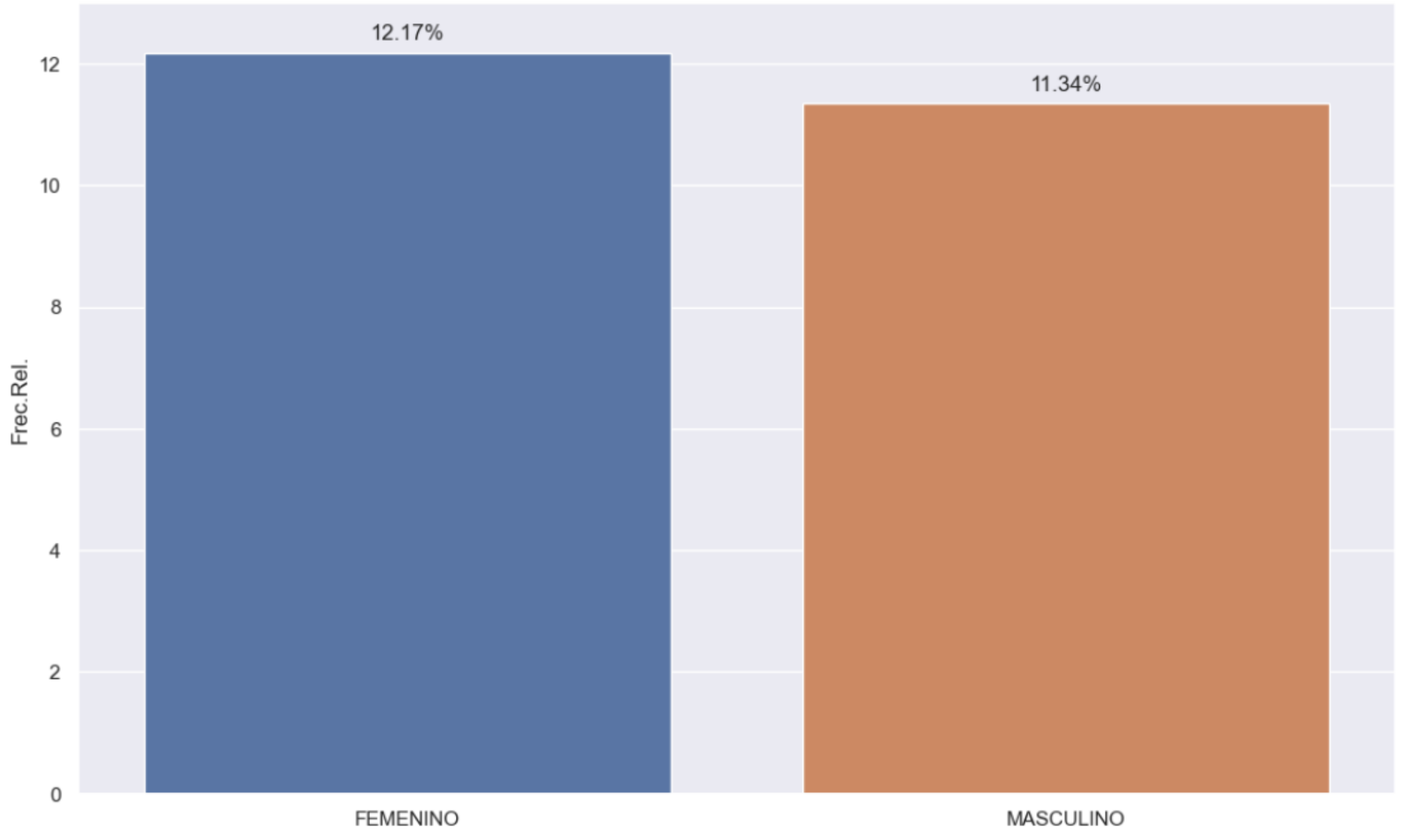
```
Estadístico X^2= 390.79136348682846, Valor-p= 5.566688434622782e-87
```

Se tiene suficiente evidencia para determinar que SI HAY relación entre el desenlace CONFIRMADO y el sexo

Caso 2 : Relación Desenlace - Sexo

CONFIRMADO

HOSPITALIZADO



Prueba χ^2

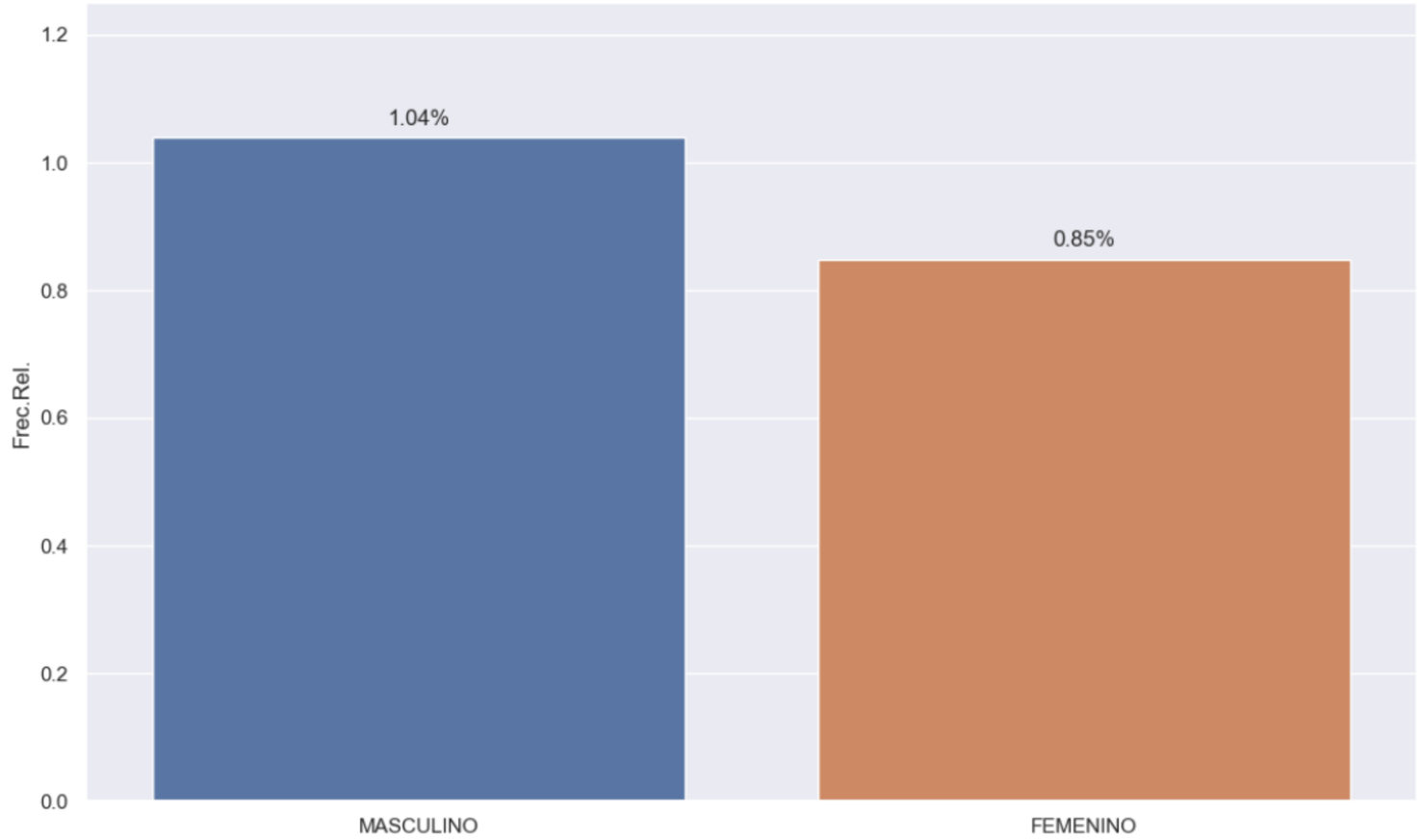
```
Estadístico X^2= 185.82159909819887, Valor-p= 2.59679354035364e-42
```

Se tiene suficiente evidencia para determinar que SI HAY relación entre el desenlace Hospitalizado y el sexo

Caso 2 :
Relación
Desenlace - Sexo

HOSPITALIZADO

FALLECIDO



Prueba χ^2

```
Estadístico X^2= 110.86741449407229, Valor-p= 6.326337054430855e-26
```

Se tiene suficiente evidencia para determinar que SI HAY relación entre el desenlace Fallecido y Sexo

Caso 2 :
Relación
Desenlace - Sexo

FALLECIDO

Conclusión

¿Existen diferencias significativas entre los individuos vacunados contra COVID-19 en 4 ciudades colombianas durante el periodo de 2021 a 2022 y sus desenlaces?

Existe diferencia parcial entre los individuos considerando la edad de aplicación y la edad para el desenlace de Estado (FALLECIDO O VIVO), pero no existe diferencia para los desenlaces Confirmado y Hospitalizado.

UNIVERSIDAD ICESI | A OTRO NIVEL

Análisis Exploratorio

Vacunaciones COVID-19

Laura Daniela Espinosa

Carlos Enrique Jaramillo

¡Muchas Gracias!

UNIVERSIDAD ICESI | A OTRO NIVEL