Informe de Exploración y Limpieza de Datos de Estudiantes

Introducción: En este informe, se aborda la exploración y limpieza de datos del conjunto "students_dirty.csv". El propósito fundamental es garantizar la calidad y coherencia de la información para su posterior análisis. Se seguirá un proceso estructurado para identificar y corregir posibles inconsistencias en los datos, asegurando que estén listos para análisis detallados y conclusiones significativas.

Carga de Datos: Comenzamos importando los módulos necesarios (NumPy, Pandas, os) y obteniendo el directorio actual. A continuación, se verifica la existencia del archivo CSV "students_dirty.csv" en el directorio y se carga en un DataFrame de Pandas para su análisis.

```
import numpy as np

import pandas as pd

import os


# Obtiene el directorio actual

directorio_actual = os.getcwd()


# Nombre del archivo CSV que se desea abrir

nombre_archivo = "students_dirty.csv"
```

```
# Se construye la ruta completa del archivo utilizando os.path.join
ruta_completa = os.path.join(directorio_actual, nombre_archivo)

# Verifica si el archivo existe antes de intentar abrirlo
if os.path.exists(ruta_completa):
    # Abre el archivo CSV utilizando pandas
    df = pd.read_csv(ruta_completa)
else:
    print(f"El archivo {nombre_archivo} no existe en el directorio actual.")
```

Exploración Inicial: Realizamos una exploración inicial del DataFrame mostrando las primeras y últimas filas, así como las dimensiones y la información detallada.

```
# Muestro el DataFrame
print(df)

# Muestro las 5 primeras filas
print(df.head())

# Muestro las últimas 5 filas
print(df.tail())

# Muestro las dimensiones del DataFrame
print(df.shape)
```

```
# Muestro la información detallada del DataFrame
print(df.info())
```

Limpieza de Datos: En esta sección, llevamos a cabo la limpieza de datos para garantizar la coherencia y calidad del DataFrame. Mostramos máscaras booleanas indicando las posiciones de los valores nulos y eliminamos las filas que contienen valores nulos. También verificamos la existencia de filas duplicadas en el DataFrame.

```
# Muestro una máscara booleana indicando las posiciones de los valores nulos
print(df.isna())

# Elimino filas con valores nulos
df.dropna(inplace=True)

# Verifico si hay filas duplicadas en todo el DataFrame
print(df.duplicated().any())
```

Guardado del Archivo Limpio: Guardamos el DataFrame limpio en un nuevo archivo CSV llamado "students_clean.csv".

```
# Guardo el archivo limpio
df.to_csv('students_clean.csv', index=False)
```

Conclusión: El proceso de exploración y limpieza de datos se ha llevado a cabo de manera exitosa, asegurando la calidad de los datos y preparándolos para análisis más detallados. El archivo limpio "students_clean.csv" está listo para su uso en futuras investigaciones.

Este informe proporciona una visión general de las acciones realizadas, los pasos específicos seguidos y asegura una presentación clara de los resultados obtenidos durante la exploración y limpieza de datos.

Exploración Integral del Rendimiento Académico: Un Análisis Detallado de Variables Clave en un Conjunto de Datos Educativos.

Objetivos del Análisis:

- **Explorar la distribución de clases y streams.**
- **Evaluar posibles correlaciones entre el rendimiento académico y variables como género, edad y tarifas.**
- **Visualizar tendencias temporales en el rendimiento académico.**
- **Identificar patrones relacionados con direcciones de correo electrónico y otros aspectos que podrían estar vinculados al rendimiento académico.**

Dimensiones del Conjunto de Datos

El conjunto de datos contiene 961 registros y 12 columnas.

(961, 12)

Tipos de Datos y Valores No Nulos

Se observa que no hay valores nulos en ninguna de las columnas y se ha asignado correctamente el tipo de dato apropiado a cada columna.

Estadísticas Descriptivas

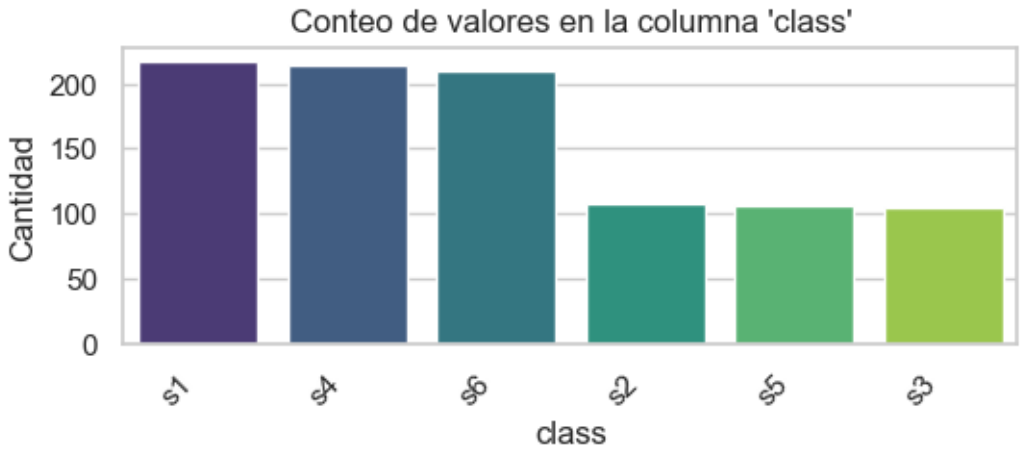
Se proporciona un resumen estadístico para las columnas numéricas, destacando la media, desviación estándar y otros estadísticos importantes.

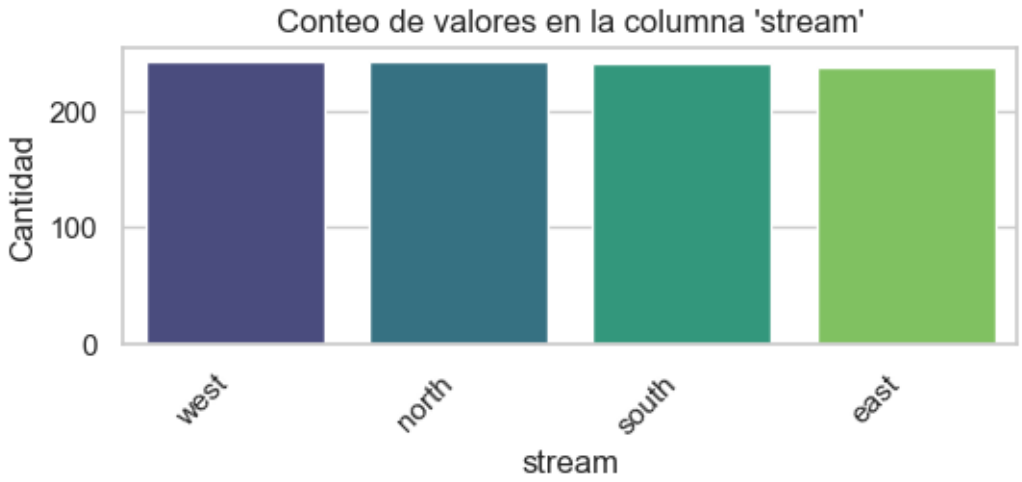
	contact	parent_id	fees
count	9.610000e+02	961.000000	961.000000
mean	2.567072e+11	504.778356	19951.092612
std	2.839295e+02	283.929521	623.081516
min	2.567072e+11	1.000000	10000.000000
25%	2.567072e+11	262.000000	20000.000000
50%	2.567072e+11	509.000000	20000.000000
75%	2.567072e+11	749.000000	20000.000000
max	2.567072e+11	1000.000000	20000.000000

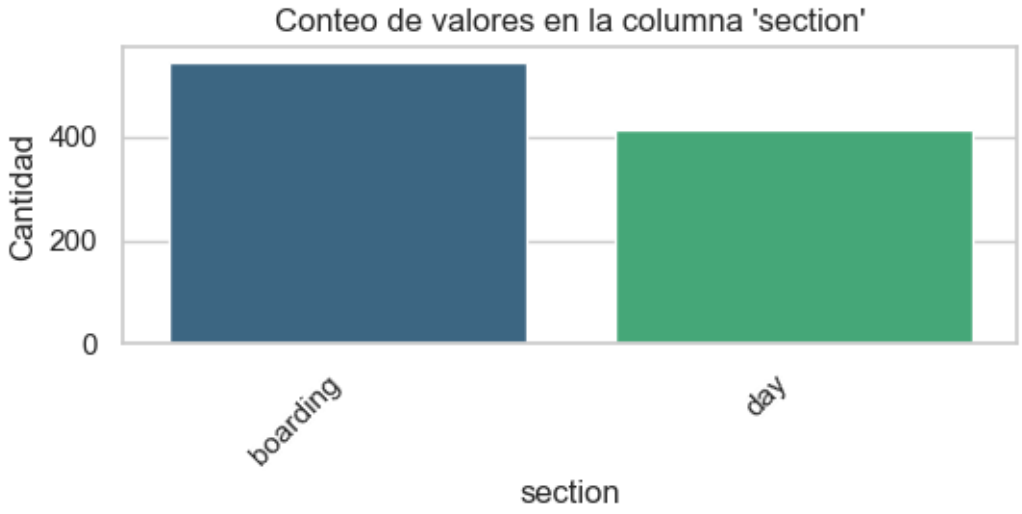
Exploración de Variables Clave

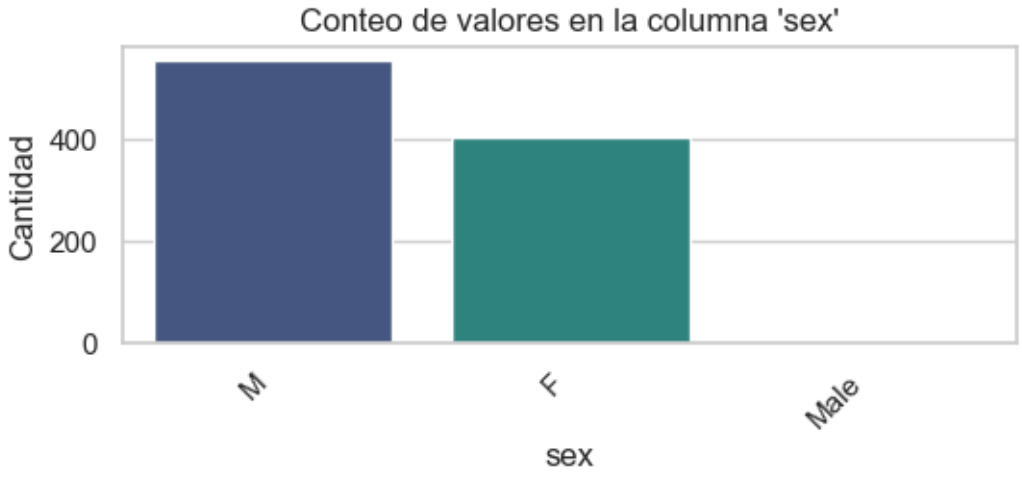
Distribución de Clases,Streams,Section,Sex

Se han creado gráficos de barras para visualizar la distribución de estudiantes en las clases,streams,section,sex, destacando posibles desequilibrios en la cantidad de estudiantes por categoría.




Conteo de valores en la columna 'stream'
Cantidad
200
100
0
west
north
south
east
stream


Conteo de valores en la columna 'section'
Cantidad
400
200
0
boarding
day
section


Conteo de valores en la columna 'sex'
Cantidad
400
200
0
M
F
Male
sex

Corrección de Valores en la Columna 'Sex'

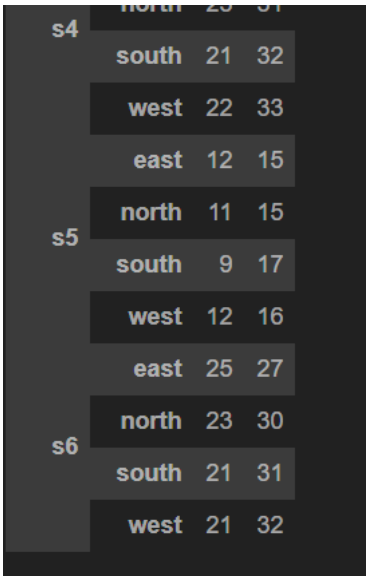
Se realizó la corrección de valores en la columna 'sex', reemplazando 'Male' con 'M' para mejorar la consistencia.

```
M    557
F    404
Name: sex, dtype: int64
```

Distribución de Estudiantes por Género en Cada Clase y Stream

Se proporciona una tabla que muestra la distribución de estudiantes por género en cada clase y stream.

	sex	F	M
class	stream		
s1	east	18	36
	north	24	30
	south	25	30
	west	21	33
s2	east	11	17
	north	11	16
	south	11	16
	west	14	12
s3	east	11	13
	north	10	18
	south	11	16
	west	12	14
s4	east	25	27
	north	23	31



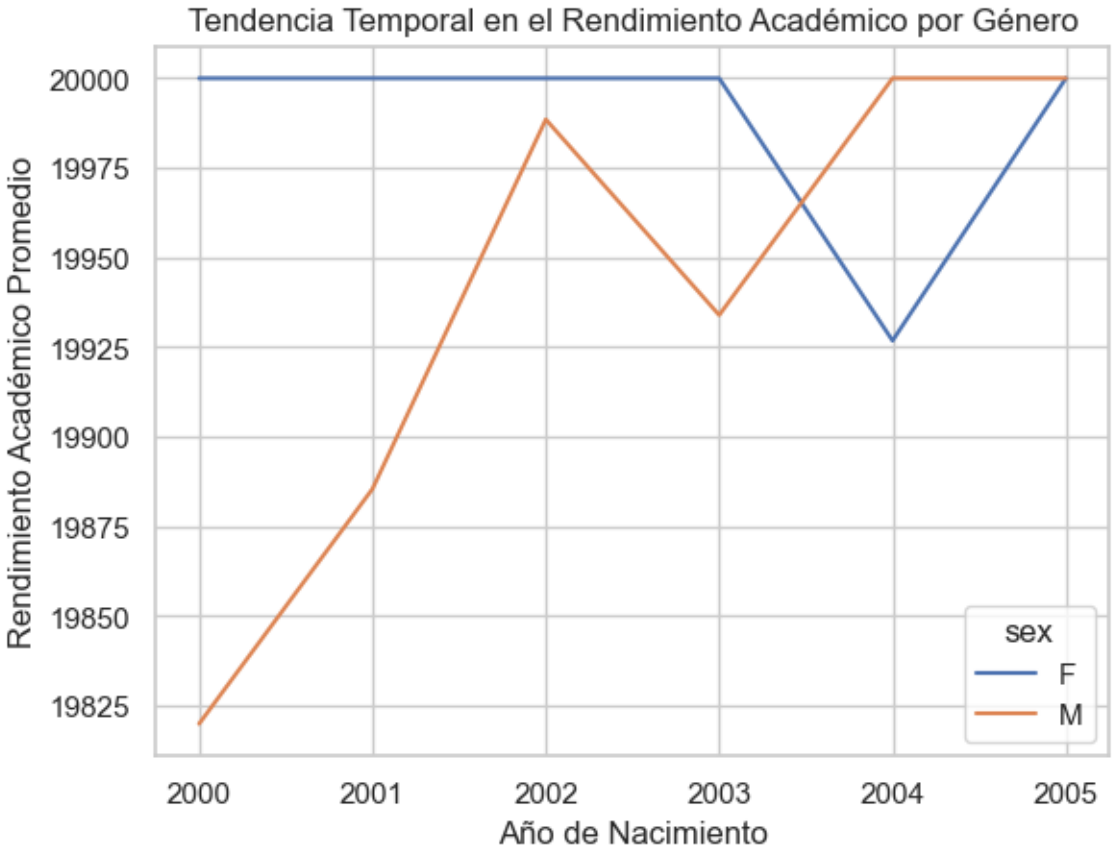
Diferencia en Rendimiento Académico entre Géneros

Se calcula la diferencia en el rendimiento académico promedio entre géneros, destacando ligeras variaciones.

```
sex
F    19985.148515
M    19926.391382
Name: fees, dtype: float64
```

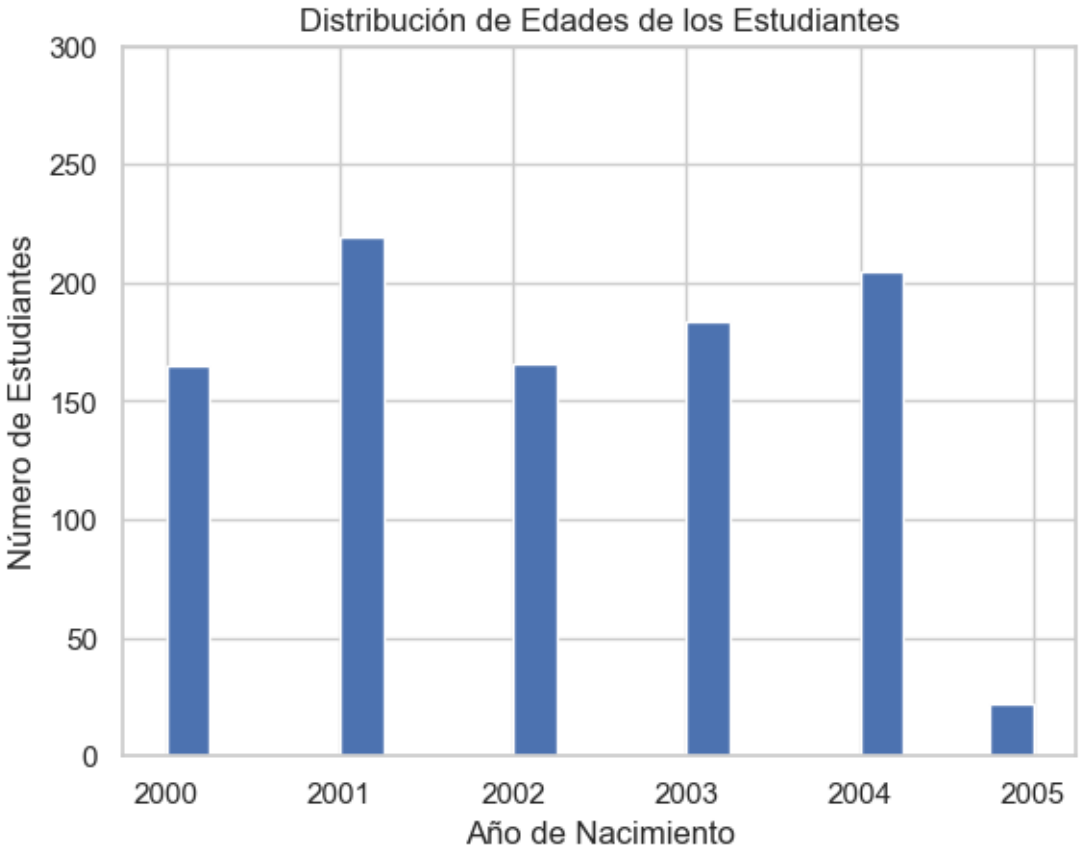
Tendencia Temporal en el Rendimiento Académico por Género

Se presenta un gráfico que visualiza la tendencia temporal en el rendimiento académico promedio de hombres y mujeres a lo largo de los años.



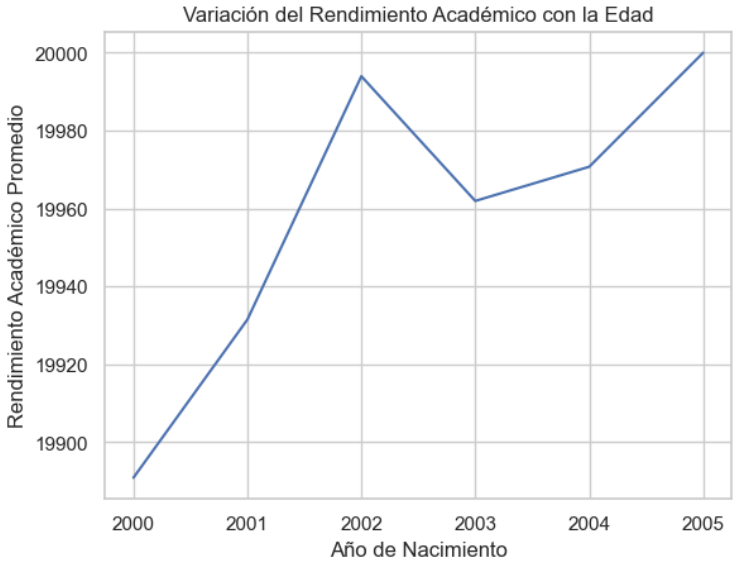
Distribución de Edades de los Estudiantes

Se crea un histograma para explorar la distribución de edades de los estudiantes.



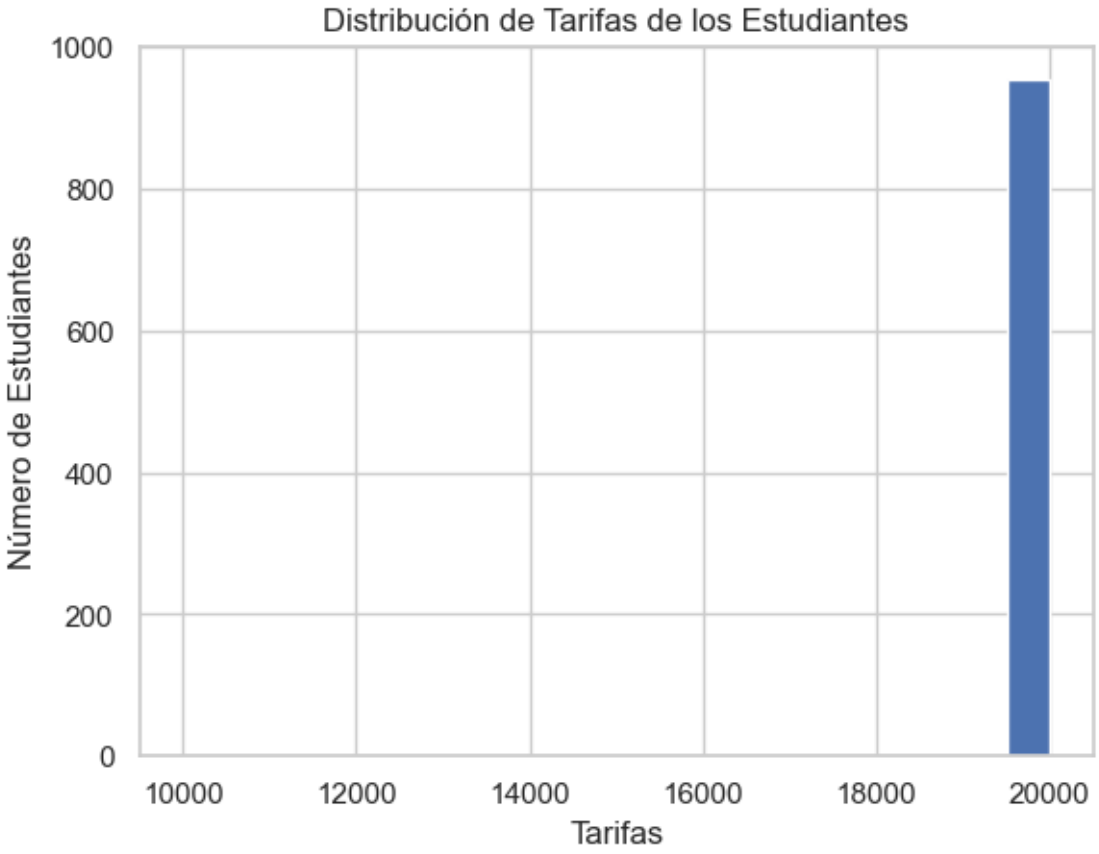
Variación del Rendimiento Académico con la Edad

Se presenta un gráfico que muestra cómo varía el rendimiento académico promedio con la edad de los estudiantes.



Distribución de Tarifas de los Estudiantes

Se proporciona un histograma que muestra la distribución de las tarifas pagadas por los estudiantes.



Diferencias Significativas en Tarifas entre Clases y Streams

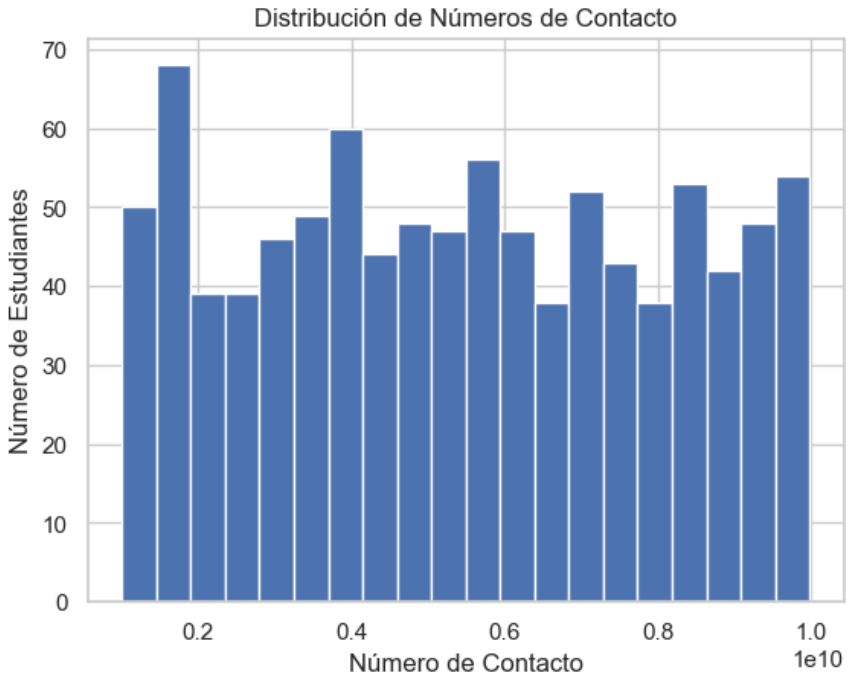
Se calcula la diferencia en tarifas promedio entre clases y streams, destacando posibles disparidades.

Diferencias en tarifas entre clases y streams:

class	east	north	south	west
s1	20000.000000	19814.814815	20000.000000	19907.407407
s2	19785.714286	20000.000000	20000.000000	20000.000000
s3	20000.000000	20000.000000	20000.000000	20000.000000
s4	20000.000000	20000.000000	20000.000000	19981.818182
s5	20000.000000	19730.769231	20000.000000	20000.000000
s6	20000.000000	20000.000000	19653.846154	20000.000000

Distribución de Números de Contacto

Se genera un conjunto de números de contacto aleatorios y se muestra la distribución resultante.



Diferencias en Rendimiento Académico entre Estudiantes Internos y Diarios

Se calcula la diferencia en el rendimiento académico promedio entre estudiantes internos y diarios.

```
section
boarding    19941.391941
day         19963.855422
Name: fees, dtype: float64
```

Diferencias en Género, Edad o Tarifas entre Estudiantes Internos y Diarios

Se presenta una tabla que muestra las diferencias en género, edad y tarifas entre estudiantes internos y diarios.

```
Diferencias en género, edad o tarifas entre estudiantes internos y diarios:
           Year_of_birth          fees
section
boarding    2002.219780  19941.391941
day         2001.978313  19963.855422
```

Conclusión y Almacenamiento del Conjunto de Datos Mejorado

Se ha concluido la exploración integral del rendimiento académico en este conjunto de datos. Se ha mejorado y almacenado el conjunto de datos final como 'Students_clean_mejorado.csv'.