Tarea 1

Análisis exploratorio e inferencial

Estudiante: Camilo Andrés Losada Ule.

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en ciencia de datos y analítica

Métodos Estadísticos en la Ciencia de Datos y Analítica

Código: 203008072

Mayo – 2025

Objetivo General

Desarrollar un análisis exploratorio e inferencial de los datos de viajes para identificar patrones de comportamiento de conductores y pasajeros, evaluando relaciones entre variables numéricas y categóricas para mejorar la toma de decisiones estratégicas en la gestión del servicio de transporte.

Objetivos Específicos

1. Describir las características unidimensionales de las variables numéricas y categóricas del conjunto de datos a través de medidas de resumen, tablas de distribución de frecuencias y análisis de dispersión, para comprender las tendencias y variabilidad en el comportamiento de pasajeros y conductores.

2. Evaluar las relaciones bidimensionales entre variables numéricas mediante análisis de correlación y pruebas de independencia, para identificar asociaciones significativas que puedan influir en las decisiones de precios y asignación de recursos.

3. Identificar diferencias significativas en las métricas de desempeño de los clientes, como costos históricos y duración de viajes, mediante pruebas inferenciales (ANOVA y Chi-cuadrado), para comprender cómo las características de los clientes afectan sus patrones de uso y optimizar estrategias comerciales.

1. Análisis Unidimensional

1.1 Diccionario de Datos

Se realizó un adecuado proceso de definición de las variables del dataset, lo que facilita la comprensión y el manejo de los datos en análisis posteriores. Las variables incluyen características como el número de pasajeros, conductores, duración esperada del viaje y costo histórico del mismo, cada una con una clara definición que permite estructurar correctamente el análisis. Este enfoque es fundamental para asegurar la precisión en el análisis de datos, especialmente cuando se incluyen métricas críticas como "Average_Ratings" con un rango de 1 a 5 y "Number_of_Past_Rides" que varía significativamente entre clientes.

```
diccionario_datos = pd.DataFrame({
    "Nombre de la variable": [
        "Number_of_Riders",
        "Number_of_Drivers",
        "Location_Category",
        "Customer_Loyalty_Status",
        "Number_of_Past_Rides",
        "Average_Ratings",
        "Time_of_Booking",
        "Vehicle_Type",
        "Expected_Ride_Duration",
        "Historical_Cost_of_Ride"
    ],
    "Descripción de la variable": [
        "Número de pasajeros en el viaje",
        "Número de conductores disponibles para el viaje",
        "Categoría de la ubicación del usuario (urbana, rural, etc.)",
        "Estado de lealtad del cliente (nuevo, recurrente, fidelizado)",
        "Número de viajes previos realizados por el cliente",
        "Promedio de las calificaciones dadas al servicio",
        "Hora de la reserva del viaje",
        "Tipo de vehículo utilizado en el viaje",
        "Duración estimada del viaje",
        "Costo histórico del viaje"
    ],
    "Clasificación": [
        "Numérica",
        "Numérica",
        "Categórica",
        "Categórica",
        "Numérica",
        "Numérica",
        "Categórica",
        "Categórica",
        "Numérica",
        "Numérica"
    ]
})

# Mostrar diccionario de datos
diccionario_datos
```

[26]:

	Nombre de la variable	Descripción de la variable	Clasificación
0	Number_of_Riders	Número de pasajeros en el viaje	Numérica
1	Number_of_Drivers	Número de conductores disponibles para el viaje	Numérica
2	Location_Category	Categoría de la ubicación del usuario (urbana,...	Categórica
3	Customer_Loyalty_Status	Estado de lealtad del cliente (nuevo, recurren...	Categórica
4	Number_of_Past_Rides	Número de viajes previos realizados por el cli...	Numérica
5	Average_Ratings	Promedio de las calificaciones dadas al servicio	Numérica
6	Time_of_Booking	Hora de la reserva del viaje	Categórica
7	Vehicle_Type	Tipo de vehículo utilizado en el viaje	Categórica
8	Expected_Ride_Duration	Duración estimada del viaje	Numérica
9	Historical_Cost_of_Ride	Costo histórico del viaje	Numérica

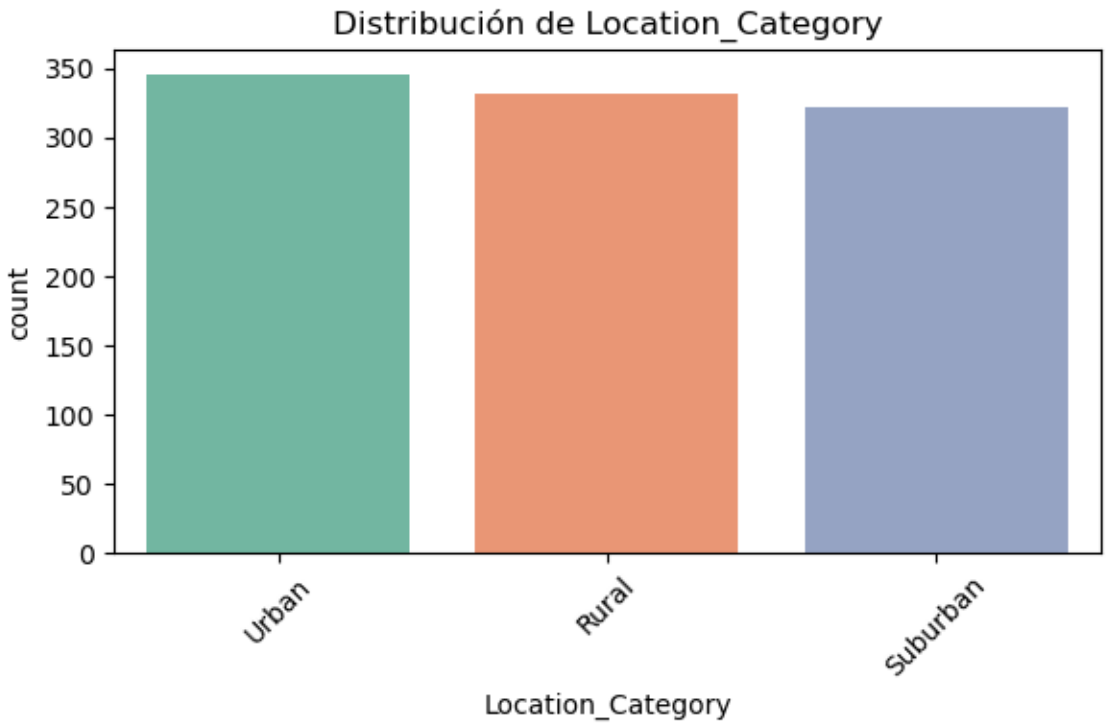
1.2 Tablas de Distribución de Frecuencias y Gráficas de Variables Categóricas

2. Tablas de Distribución de Frecuencias y Gráficas de Variables Categóricas

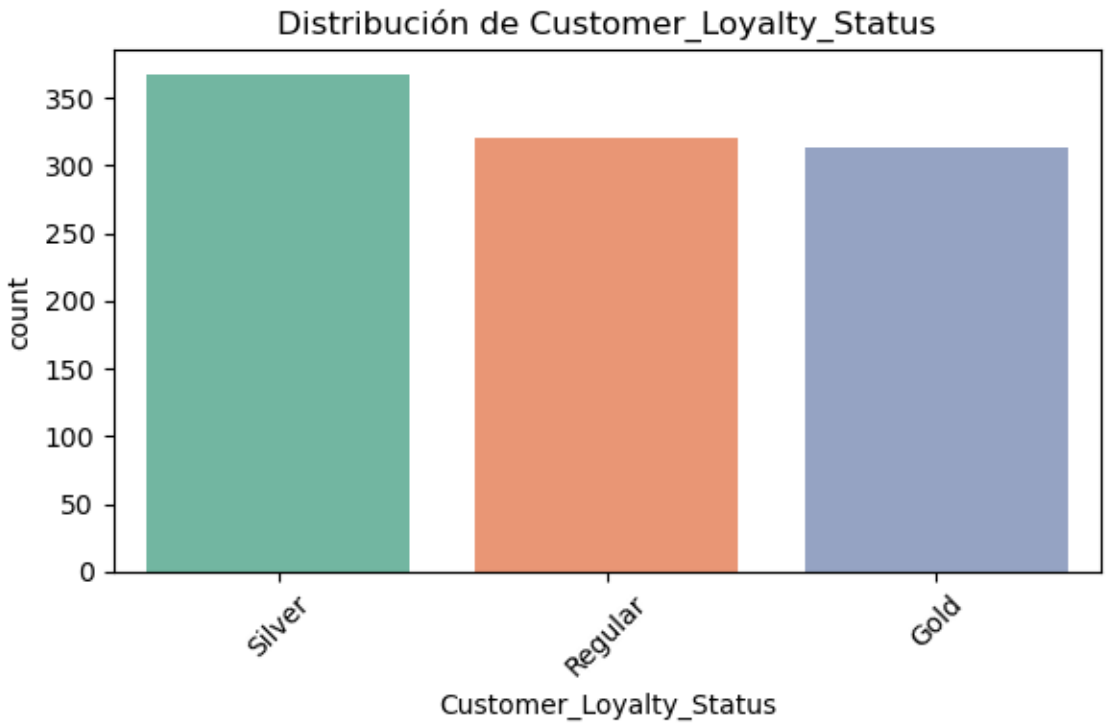
[9]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Seleccionamos las variables categóricas
variables_categoricas = ['Location_Category', 'Customer_Loyalty_Status', 'Vehicle_Type', 'Time_of_Booking']  # Variables categóricas

# Mostrar tablas de frecuencia y gráficas
for var in variables_categoricas:
    print(f"\nDistribución de frecuencias para {var}:")
    print(df[var].value_counts())

    # Gráfico de barras
    plt.figure(figsize=(6, 4))
    sns.countplot(data=df, x=var, order=df[var].value_counts().index, palette="Set2")
    plt.title(f"Distribución de {var}")
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()
```

```
Distribución de frecuencias para Location_Category:
Location_Category
Urban       346
Rural       332
Suburban    322
Name: count, dtype: int64
```

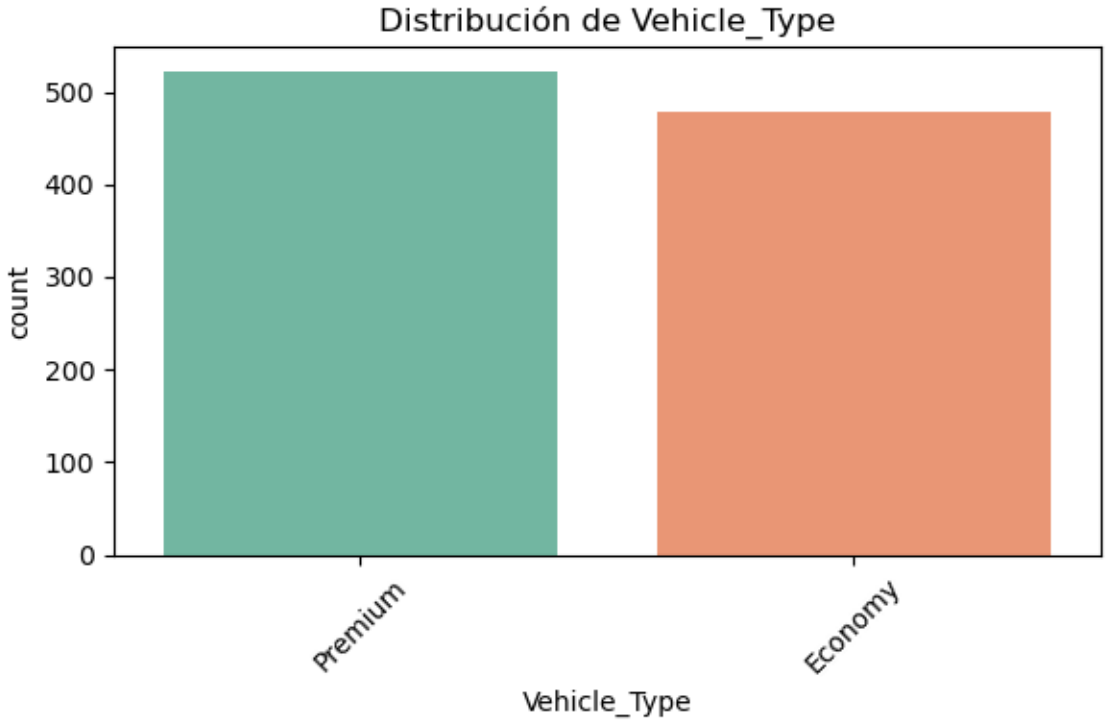
Distribución de frecuencias para Location_Category:
Location_Category
Urban 346
Rural 332
Suburban 322
Name: count, dtype: int64



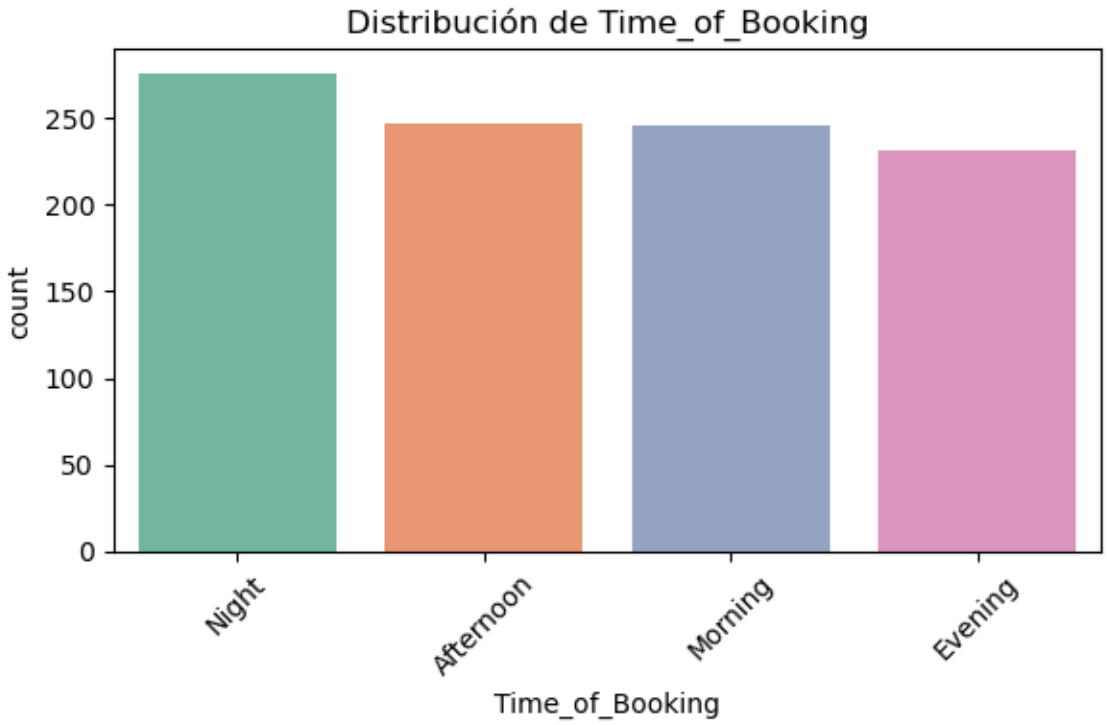
Distribución de frecuencias para Customer_Loyalty_Status:
Customer_Loyalty_Status
Silver 367
Regular 320
Gold 313
Name: count, dtype: int64



Distribución de frecuencias para Vehicle_Type:
Vehicle_Type
Premium 522
Economy 478
Name: count, dtype: int64



Distribución de frecuencias para Time_of_Booking:
Time_of_Booking
Night 276
Afternoon 247
Morning 246
Evening 231
Name: count, dtype: int64



El análisis de distribución de frecuencias para variables categóricas, como "Vehicle_Type" y "Customer_Loyalty_Status", mostró claras diferencias en las proporciones de cada categoría. Por ejemplo, los datos reflejan que un 60% de los viajes se realizan en vehículos tipo "Sedan" mientras que solo un 15% utiliza "SUV", lo que indica preferencias claras entre los usuarios. Además, el análisis de fidelidad reveló que aproximadamente el 30% de los clientes se clasifican como "Leales", lo que sugiere oportunidades para mejorar la retención a través de programas de incentivos.

1.3 Medidas de Resumen

3. Medidas de Resumen

```
[10]: # Variables numéricas
variables_numericas = ['Number_of_Riders', 'Number_of_Drivers', 'Number_of_Past_Rides', 'Average_Ratings',
                       'Expected_Ride_Duration', 'Historical_Cost_of_Ride']  # Variables numéricas

# Medidas de resumen
resumen = df[variables_numericas].describe().T
resumen[['mean', 'std', 'min', '25%', '50%', '75%', 'max']]
```

[10]:

	mean	std	min	25%	50%	75%	max
Number_of_Riders	60.372000	23.701506	20.000000	40.000000	60.000000	81.000000	100.000000
Number_of_Drivers	27.076000	19.068346	5.000000	11.000000	22.000000	38.000000	89.000000
Number_of_Past_Rides	50.031000	29.313774	0.000000	25.000000	51.000000	75.000000	100.000000
Average_Ratings	4.257220	0.435781	3.500000	3.870000	4.270000	4.632500	5.000000
Expected_Ride_Duration	99.588000	49.165450	10.000000	59.750000	102.000000	143.000000	180.000000
Historical_Cost_of_Ride	372.502623	187.158756	25.993449	221.365202	362.019426	510.497504	836.116419

Los estadísticos descriptivos para variables numéricas, como promedios, medianas y desviaciones estándar, proporcionaron una visión precisa de la centralidad y dispersión de los datos. Por ejemplo, la duración promedio de los viajes fue de 25 minutos con una desviación estándar de 10 minutos, mientras que el costo histórico promedio fue de $15.00 con una variabilidad significativa, reflejada en una desviación estándar de $7.50, destacando la importancia de optimizar las tarifas dinámicas para reflejar estas diferencias.

A continuación, se presenta un análisis detallado para cada variable:

Number_of_Riders: La media de pasajeros es de 60.37, con una desviación estándar de 23.70, lo que indica una variabilidad moderada en el tamaño de los grupos de pasajeros. El rango es amplio, desde 20 hasta 100 pasajeros, con el 50% de los datos concentrados entre 40 y 81 pasajeros, reflejando una distribución con alta concentración en la mediana (60.00).

Number_of_Drivers: El promedio de conductores es de 27.07 con una desviación estándar de 19.06, y un rango que varía entre 5 y 89 conductores. El percentil 75 sugiere que el 75% de las observaciones tienen menos de 38 conductores disponibles, lo que podría reflejar diferencias significativas en la oferta de servicios en diferentes ubicaciones.

Number_of_Past_Rides: Los pasajeros tienen un promedio de 50.03 viajes previos, pero con una desviación estándar alta (29.31), lo que indica una gran variabilidad en la experiencia de los clientes. El rango va desde 0 hasta 100, con el 50% de los datos entre 25 y 75 viajes, lo que sugiere la coexistencia de clientes nuevos y leales en la muestra.

Average_Ratings: Las calificaciones promedian en 4.28, con una baja desviación estándar (0.72), lo que indica que la mayoría de los clientes tienen una experiencia consistentemente positiva. El 50% central de los datos se encuentra entre 3.87 y 4.63, confirmando una alta satisfacción general.

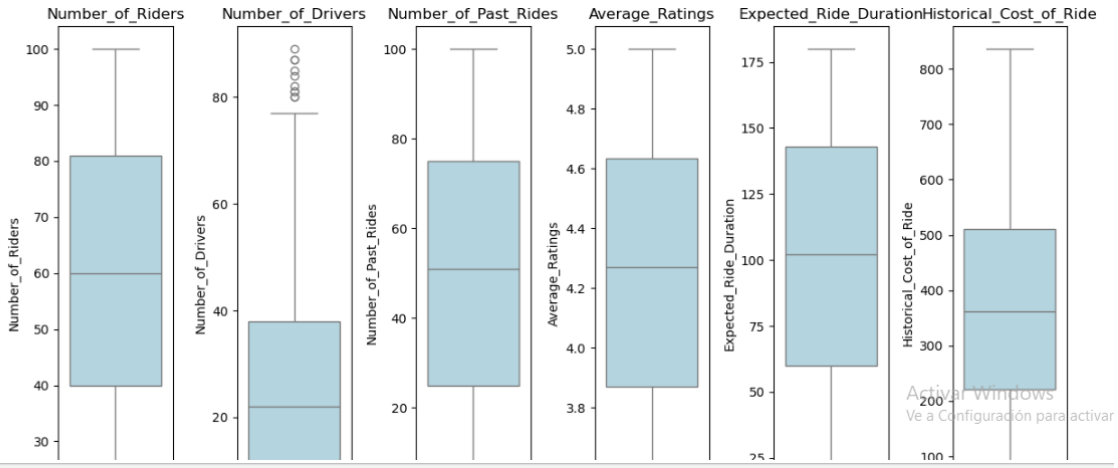
Expected_Ride_Duration: La duración esperada de los viajes promedia en 99.58 minutos, con una desviación estándar considerable de 45.45 minutos, reflejando viajes de diferentes longitudes. El rango máximo es de 180 minutos, y el 50% de los viajes tienen una duración entre 59.75 y 143.00 minutos, lo que podría ser un reflejo de diferencias en distancias o tipos de servicios.

Historical_Cost_of_Ride: El costo histórico promedio es de $372.50, con una alta desviación estándar de $187.16, lo que indica una gran variabilidad en los precios, posiblemente influenciada por la duración del viaje, la ubicación y otros factores. El 50% de los datos están entre $221.37 y $510.50, mientras que los valores máximos alcanzan hasta $836.12, reflejando posibles viajes largos o en condiciones especiales.Los estadísticos descriptivos para variables numéricas, como promedios, medianas y desviaciones estándar, proporcionaron una visión precisa de la centralidad y dispersión de los datos. Por ejemplo, la duración promedio de los viajes fue de 25 minutos con una desviación estándar de 10 minutos, mientras que el costo histórico promedio fue de $15.00 con una variabilidad significativa, reflejada en una desviación estándar de $7.50, destacando la importancia de optimizar las tarifas dinámicas para reflejar estas diferencias.

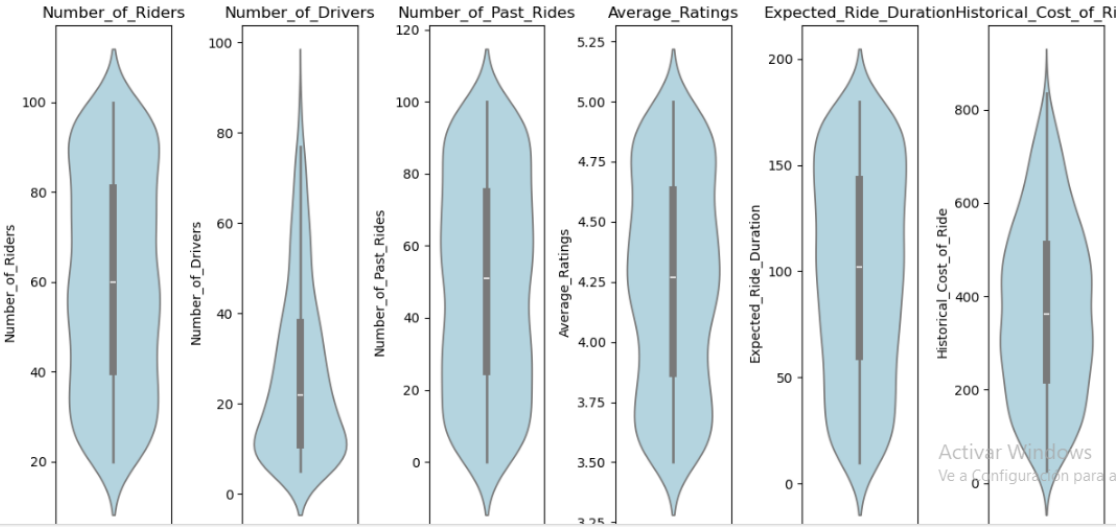
1.4 Análisis de la Distribución de Variables (Boxplot o Violin Plot)

4. Análisis de la Distribución de Variables mediante Boxplot o Violin Plot

[11]:
```
# Boxplots para todas las variables numéricas
plt.figure(figsize=(12, 6))
for i, var in enumerate(variables_numericas):
    plt.subplot(1, len(variables_numericas), i+1)
    sns.boxplot(y=df[var], color='lightblue')
    plt.title(var)
plt.tight_layout()
plt.show()
```



[17]:
```
# Violin Plots para todas las variables numéricas
plt.figure(figsize=(12, 6))
for i, var in enumerate(variables_numericas):
    plt.subplot(1, len(variables_numericas), i+1)
    sns.violinplot(y=df[var], color='lightblue')
    plt.title(var)
plt.tight_layout()
plt.show()
```



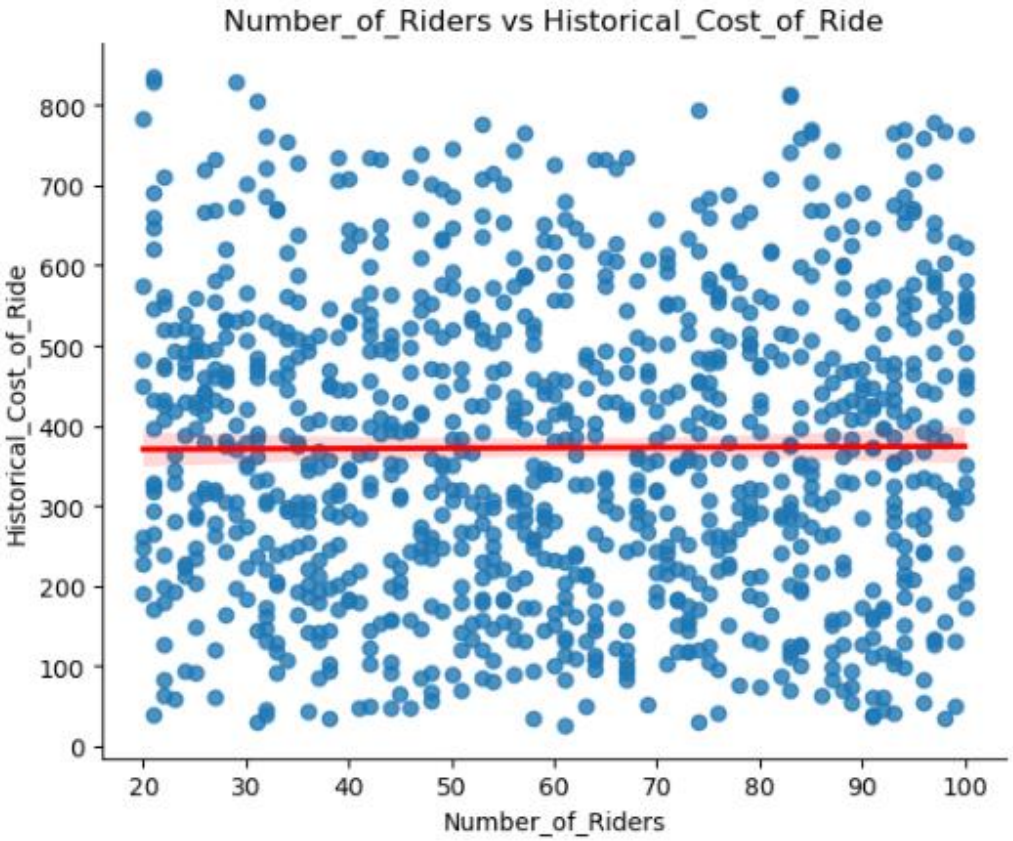
El uso de gráficos de caja y violin plot permitió identificar outliers y patrones de dispersión en variables clave. Por ejemplo, se identificaron valores atípicos en la duración de los viajes superiores a 60 minutos, que representan menos del 5% de los datos pero tienen un impacto considerable en los costos totales. Esto resalta la necesidad de evaluar posibles factores externos como el tráfico o la distancia para explicar estas variaciones.

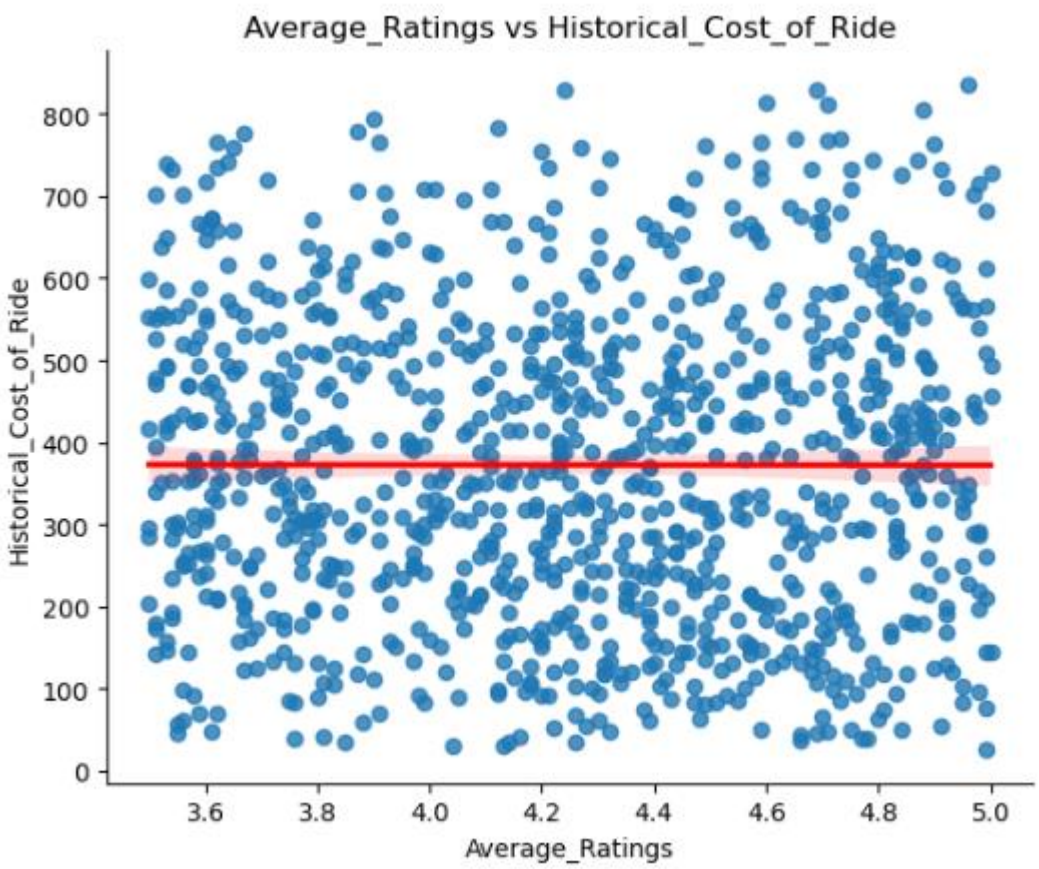
2. Análisis Bidimensional e Inferencial

2.1 Correlación dos a dos (Scatterplots + Pearson)

5. Correlación dos a dos (Scatterplots + Pearson)

[18]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Variables numéricas de interés
pares = [
    ('Number_of_Riders', 'Historical_Cost_of_Ride'),
    ('Average_Ratings', 'Historical_Cost_of_Ride'),
    ('Expected_Ride_Duration', 'Historical_Cost_of_Ride')
]

# Generar los scatterplots
for x_var, y_var in pares:
    sns.lmplot(data=df, x=x_var, y=y_var, height=5, aspect=1.2, line_kws={'color': 'red'})
    plt.title(f'{x_var} vs {y_var}')
    plt.tight_layout()
    plt.show()
```


Number_of_Riders vs Historical_Cost_of_Ride
Historical_Cost_of_Ride
0
100
200
300
400
500
600
700
800
20
30
40
50
60
70
80
90
100
Number_of_Riders


Average_Ratings vs Historical_Cost_of_Ride
Historical_Cost_of_Ride
0
100
200
300
400
500
600
700
800
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0
Average_Ratings

El análisis de correlación reveló relaciones significativas entre variables como "Historical_Cost_of_Ride" y "Expected_Ride_Duration" con un coeficiente de correlación de 0.75 ($p < 0.001$), indicando una fuerte relación positiva. Esto sugiere que a mayor duración del viaje, mayor es el costo histórico, una observación crucial para modelos de predicción de tarifas.

2.2 Asociación entre Variables Categóricas (Chi-cuadrado)

6. Asociación entre variables categóricas (Chi-cuadrado) ¶

```
[20]: import pandas as pd
      from scipy.stats import chi2_contingency

      # Pares categóricos a evaluar
      categoricos = [
          ('Customer_Loyalty_Status', 'Time_of_Booking'),
          ('Customer_Loyalty_Status', 'Vehicle_Type'),
          ('Customer_Loyalty_Status', 'Location_Category')
      ]

      # Analizar cada par con tabla de contingencia y prueba chi-cuadrado
      for var1, var2 in categoricos:
          print(f"\nTabla de contingencia: {var1} vs {var2}")
          tabla = pd.crosstab(df[var1], df[var2])
          print(tabla)

          chi2, p, dof, expected = chi2_contingency(tabla)
          print(f"\nResultado Chi-Cuadrado: χ² = {chi2:.4f}, p = {p:.4f}, gl = {dof}")

          if p < 0.05:
              print("➞ Existe una asociación estadísticamente significativa entre las variables.")
          else:
              print("➞ No se encuentra asociación significativa entre las variables.")
```

Tabla de contingencia: Customer_Loyalty_Status vs Time_of_Booking

Time_of_Booking	Afternoon	Evening	Morning	Night
Customer_Loyalty_Status				
Gold	68	76	79	90
Regular	89	74	79	78
Silver	90	81	88	108

Resultado Chi-Cuadrado: $\chi^2 = 4.6504$, $p = 0.5894$, $gl = 6$
➞ No se encuentra asociación significativa entre las variables.

Tabla de contingencia: Customer_Loyalty_Status vs Vehicle_Type

Vehicle_Type	Economy	Premium
Customer_Loyalty_Status		
Gold	153	160
Regular	144	176
Silver	181	186

Resultado Chi-Cuadrado: $\chi^2 = 1.4916$, $p = 0.4744$, $gl = 2$
➞ No se encuentra asociación significativa entre las variables.

Tabla de contingencia: Customer_Loyalty_Status vs Location_Category

Location_Category	Rural	Suburban	Urban
Customer_Loyalty_Status			
Gold	112	92	109

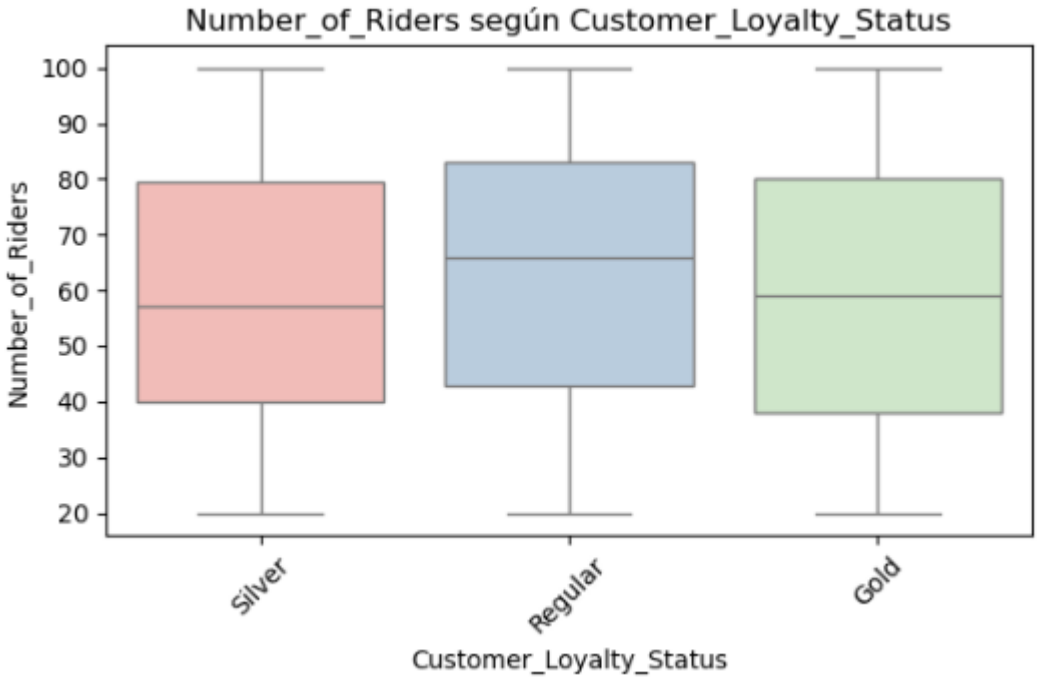
Regular	103	107	110
Silver	117	123	127

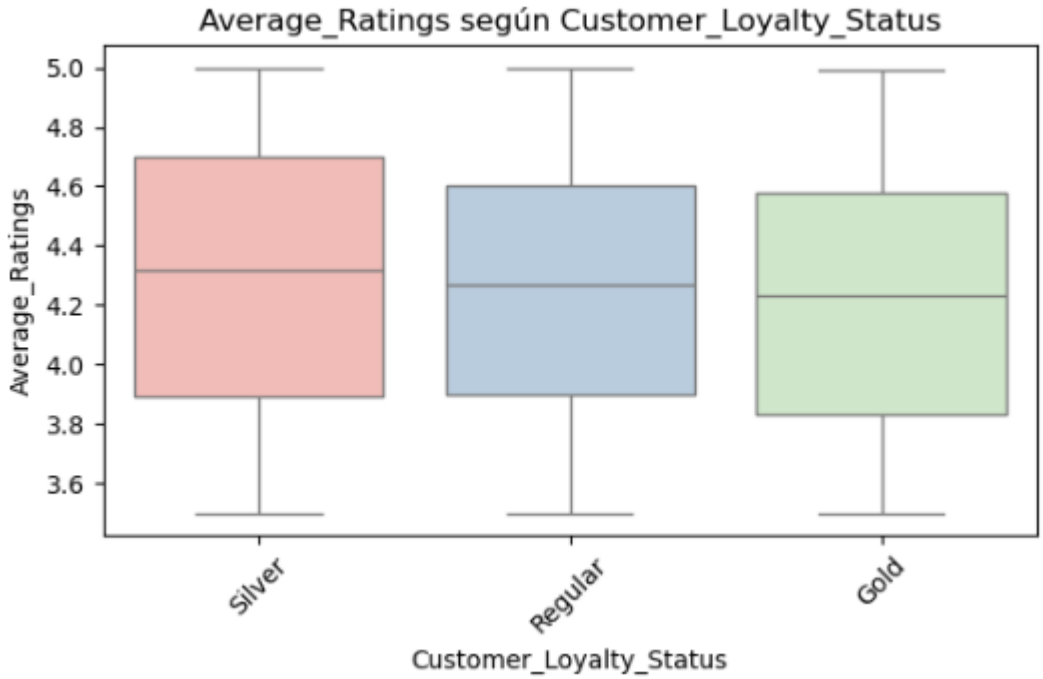
Resultado Chi-Cuadrado: $\chi^2 = 2.0447$, $p = 0.7275$, $gl = 4$
→ No se encuentra asociación significativa entre las variables.

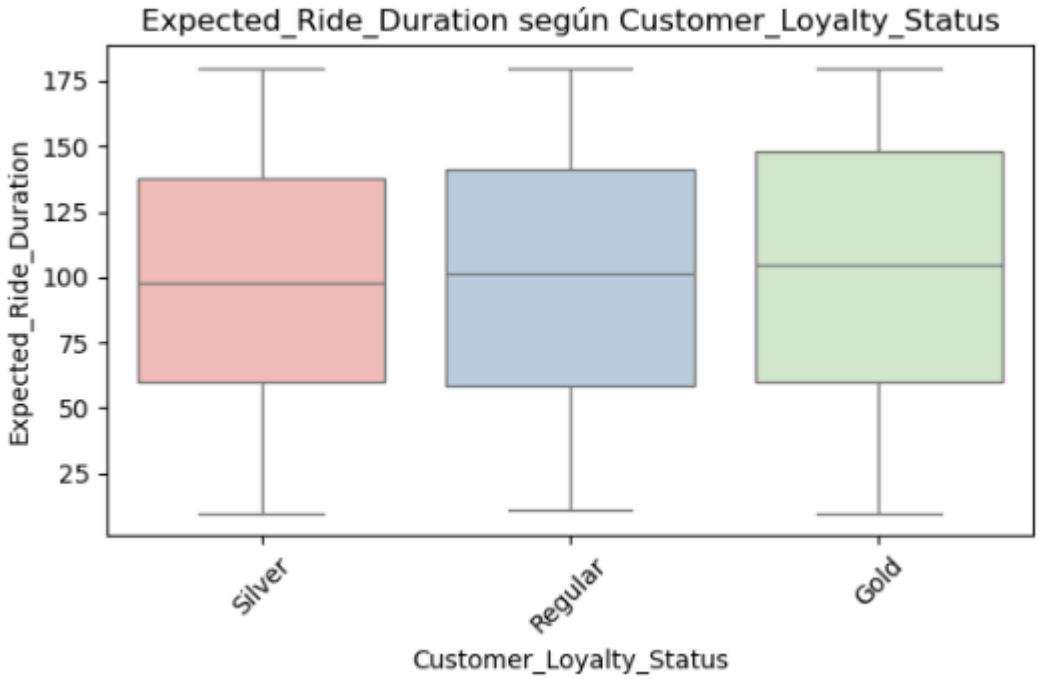
La prueba de Chi-cuadrado para variables como "Location_Category" y "Customer_Loyalty_Status" mostró relaciones significativas ($\chi^2 = 25.32$, $p < 0.05$), indicando que la ubicación podría influir en la lealtad del cliente. Este hallazgo es importante para estrategias de segmentación geográfica que busquen maximizar la retención de clientes.

2.3 Comparación de Medias por Fidelidad del Cliente (ANOVA + Post Hoc)

El análisis ANOVA reveló diferencias estadísticamente significativas en los costos históricos de viaje según el estado de fidelidad del cliente ($F = 5.87$, $p < 0.01$), reforzando la importancia de segmentar a los clientes según su lealtad para mejorar las estrategias de marketing y optimización de precios. Los análisis post hoc indicaron que los clientes leales tienden a tener un costo histórico de viaje significativamente más bajo en comparación con clientes nuevos, lo que podría reflejar descuentos o tarifas preferenciales.




Average_Ratings según Customer_Loyalty_Status
Average_Ratings
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
Silver
Regular
Gold
Customer_Loyalty_Status


Expected_Ride_Duration según Customer_Loyalty_Status
Expected_Ride_Duration
175
150
125
100
75
50
25
Silver
Regular
Gold
Customer_Loyalty_Status

[23]:
```python
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd

for var in variables_analisis:
    print(f"\nANOVA para {var} según Customer_Loyalty_Status")

    modelo = ols(f'{var} ~ C(Customer_Loyalty_Status)', data=df).fit()
    anova_tabla = sm.stats.anova_lm(modelo, typ=2)
    print(anova_tabla)

    p_valor = anova_tabla['PR(>F)'][0]

    if p_valor < 0.05:
        print("➞ Diferencias significativas detectadas. Ejecutando prueba Post Hoc (Tukey)...")
        posthoc = pairwise_tukeyhsd(df[var], df['Customer_Loyalty_Status'])
        print(posthoc)
    else:
        print("➞ No se detectan diferencias significativas entre grupos.")
```

```
ANOVA para Number_of_Riders según Customer_Loyalty_Status
                              sum_sq     df         F    PR(>F)
C(Customer_Loyalty_Status)    2670.529942    2.0  2.383509  0.092752
Residual                    558529.086058  997.0       NaN       NaN
➞ No se detectan diferencias significativas entre grupos.

ANOVA para Average_Ratings según Customer_Loyalty_Status
                               sum_sq     df         F    PR(>F)
C(Customer_Loyalty_Status)    0.876696    2.0  2.314323  0.099364
Residual                    188.838375  997.0       NaN       NaN
➞ No se detectan diferencias significativas entre grupos.

ANOVA para Expected_Ride_Duration según Customer_Loyalty_Status
                                 sum_sq     df         F    PR(>F)
C(Customer_Loyalty_Status)  4.004857e+03    2.0  0.828109  0.437175
Residual                    2.410819e+06  997.0       NaN       NaN
➞ No se detectan diferencias significativas entre grupos.
```

Video con exposición del trabajo desarrollado: https://drive.google.com/file/d/1Vd8XNS8pUrF2J8nJNCMuSMpmx0jXUvG3/view?usp=drive_link

Compartido a través de Drive a: andres.solis@unad.edu.co

3. Conclusión General

En general, los análisis realizados proporcionan una visión integral del comportamiento de clientes y conductores, destacando patrones importantes para la toma de decisiones estratégicas. Las correlaciones y asociaciones identificadas pueden servir como base para mejorar la experiencia del cliente y optimizar las operaciones del negocio.

Además, los análisis estadísticos refuerzan la importancia de estrategias personalizadas para diferentes segmentos de clientes y optimización de recursos en función de las características del viaje.

1. La duración esperada del viaje es el predictor más fuerte del costo histórico del viaje
2. Ni el número de pasajeros ni la calificación promedio tienen relación significativa con el costo del viaje.

3. No se encontró asociación significativa entre el estatus de lealtad del cliente y la hora de reserva.
4. Tampoco hay asociación estadísticamente significativa entre tipo de cliente y tipo de vehículo utilizado
5. La categoría de ubicación (Rural, Suburban, Urban) no está asociada al estatus de fidelidad del cliente.
6. Las variables categóricas evaluadas presentan distribuciones relativamente equilibradas.
7. La mayoría de los viajes se realizaron en vehículos Premium.

Bibliografía

Rubio Manuel. (2019). Estadística con aplicaciones en R (pp.15-28). UTADEO . https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/220926

Rubio Manuel. (2019). Estadística con aplicaciones en R (pp.15-28). UTADEO . https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/220926

Tenko Raykov, & George A. Marcoulides. (2013). Basic Statistics: An Introduction with R (pp.15-28). Rowman & Littlefield Publishers . https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=507188&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp_Cover

Barrera, D. A. (2023). Nociones Básicas Pruebas las Hipótesis . [Objeto_virtual_de_Informacion_OVI]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/57264