Infografía del viajero

Índice

BASES DE ESTUDIO

Introducción

En los últimos años, el sector del alquiler vacacional en España ha experimentado una evolución notablemente positiva, marcada por un crecimiento sostenido y una diversificación en su oferta. Esta transformación no solo refleja un cambio en las preferencias de los consumidores, sino que también evidencia la adaptabilidad y la innovación dentro del sector turístico.

Ante la gran variedad y cantidad de datos que se generan en este tipo de negocio, realizar un análisis en profundidad es fundamental. Comprender patrones y tendencias de comportamiento no solo es esencial para conocer el mercado actual, sino que permite la creación de acciones comerciales más efectivas y adaptadas a los diferentes tipos de viajeros.

En el presente EDA (Análisis Exploratorio de Datos), se estudia el caso particular de FeelFree, una empresa con varios años de trayectoria, especializada en la gestión de apartamentos en dos regiones clave de España: San Sebastián y Baqueira.

San Sebastián, una ciudad conocida por su belleza y su oferta cultural, presenta una temporada turística más amplia con períodos muy diferenciados de alta y baja demanda. Los huéspedes en esta región son tanto nacionales como internacionales, con una notable presencia de visitantes de fuera de España. Esto plantea un desafío único en términos de marketing y gestión de servicios, requiriendo un enfoque flexible y multicultural.

Por otro lado, Baqueira, conocida por sus estaciones de esquí, atrae principalmente a un perfil de viajero nacional. La demanda en esta región se concentra en los meses de invierno, especialmente durante la temporada de nieve. Esto supone retos diferentes en cuanto a logística, mantenimiento y promoción, enfocándose en atraer y satisfacer a un mercado más localizado y estacional.

Este análisis se ha enfocado con más peso en el destino de San Sebastián por varias razones significativas. Primero, hay una mayor riqueza de datos disponibles para este destino, lo que permite un análisis más detallado y profundo. Segundo, San Sebastián representa una porción más significativa del negocio de FeelFree, lo que hace que sea crucial entender y optimizar su rendimiento. Finalmente, se ha detectado que existe un mayor potencial de oportunidades de mejora en este destino, lo que podría llevar a estrategias más efectivas y a un incremento en la rentabilidad.

Objetivos

1. **Encontrar segmentos clave con alto impacto en los ingresos.**

   La metodología incluye un análisis detallado de los datos de reservas, segmentando por variables como origen, género, temporada, y otras características de los huéspedes. El análisis buscará patrones y tendencias que indiquen los segmentos que contribuyen de manera más significativa al ingreso total, permitiendo así enfocar esfuerzos y recursos en los grupos más rentables.

2. **Encontrar segmentos clave por crecimiento interanual**

   Aquí, el enfoque estará en identificar aquellos segmentos que han mostrado un crecimiento interanual sustancial. Se compararán datos históricos para evaluar el crecimiento en términos de reservas. Este análisis permitirá detectar tendencias emergentes y segmentos en crecimiento.

3. **Detectar fortalezas en el posicionamiento del canal de venta propio**

   Este objetivo implica analizar el rendimiento del canal de ventas propio de la empresa (la web) comparado con otros canales (como agencias de viajes online o plataformas de reservas). El análisis se enfocará en identificar segmentos de mercado o países donde el canal propio tiene una

ventaja competitiva en términos de reservas. Esta información es clave para desarrollar estrategias que potencien el canal propio, aprovechando sus fortalezas en mercados específicos.

4. **Mejorar el conocimiento demográfico y geográfico de los viajeros**

   El último objetivo busca establecer una segmentación clara tanto geográfica como demográfica para afinar acciones de comunicación y marketing. Esto implica analizar los datos disponibles para crear perfiles de huéspedes basados en su ubicación geográfica, edad, género, intereses, y comportamientos de reserva. Esta segmentación ayudará a identificar oportunidades de mercado no explotadas y a ajustar la oferta de productos o servicios para satisfacer mejor las necesidades de diferentes grupos de clientes.

CONCLUSIONES RELEVANTES

En este apartado se presentan las conclusiones más relevantes del estudio. Las diferentes líneas de análisis seguidas y sus principales resultados se pueden encontrar en el anexo 3.

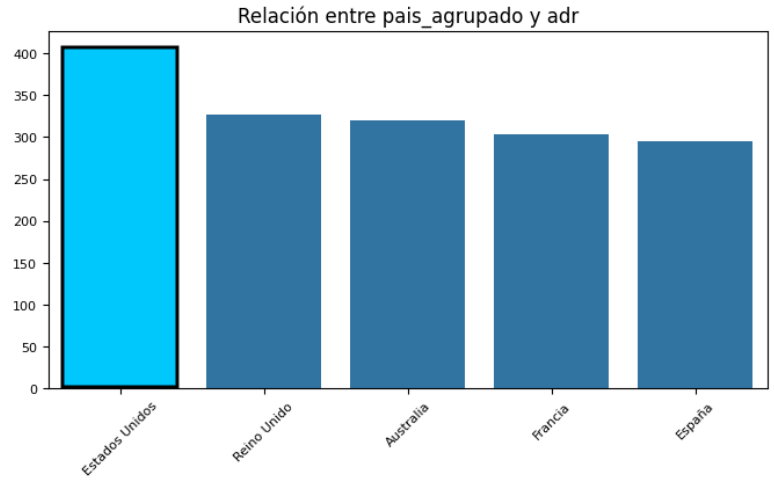
Para este estudio se analizan reservas de 2019, 2022 y 2023.

Segmentos clave por alto impacto en revenue

Para medir aquellos grupos que tienen un alto impacto en los ingresos se han establecido 2 indicadores: adr (precio medio por noche reservada) y estancia media de noches por reserva. A mayores valores de estos dos indicadores mayores ingresos.

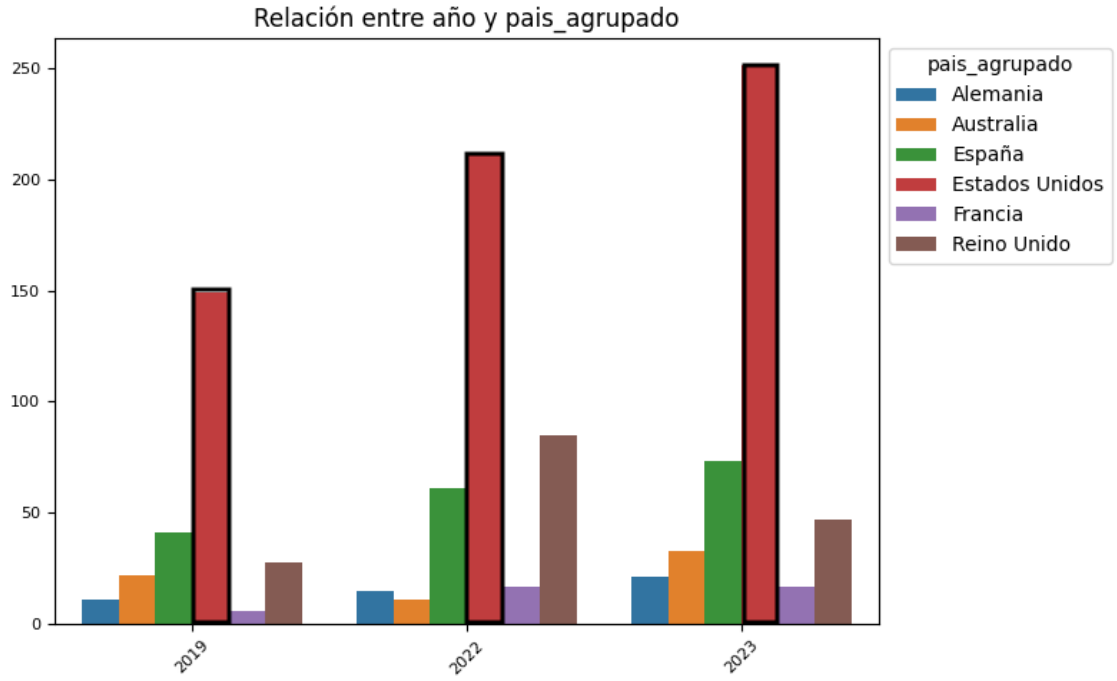
Los siguientes resultados son para el destino **San Sebastián de alquiler vacacional.**

Después de agrupar los viajeros por nacionalidad y analizar su *adr* observamos mayor precio por noche en los estadounidenses.



Analizando aquellas reservas de más de 2.500 euros vemos que el mayor porcentaje y el mayor crecimiento se encuentra dentro de EE.UU.

Reservas de más de 2.500 euros por nacionalidad y año:



Los diferentes datos obtenidos en el análisis marcan EE.UU como la nacionalidad con mayor impacto en revenue por tener un **adr mayor** que el resto. Por otro lado también se ha observado que esta nacionalidad también tiene una **estancia media mayor** que el resto.

Como se detalla en análisis posteriores, este segmento tiene un peso clave en **temporada alta** y también en **temporada media** siendo su impacto en temporada baja reducido. El perfil del viajero es de mediana edad con cierto crecimiento del perfil senior. También se ha analizado las reservas por género viendo una ligera mayoría del grupo femenino. Todo esto desarrollado en el análisis del anexo 3.

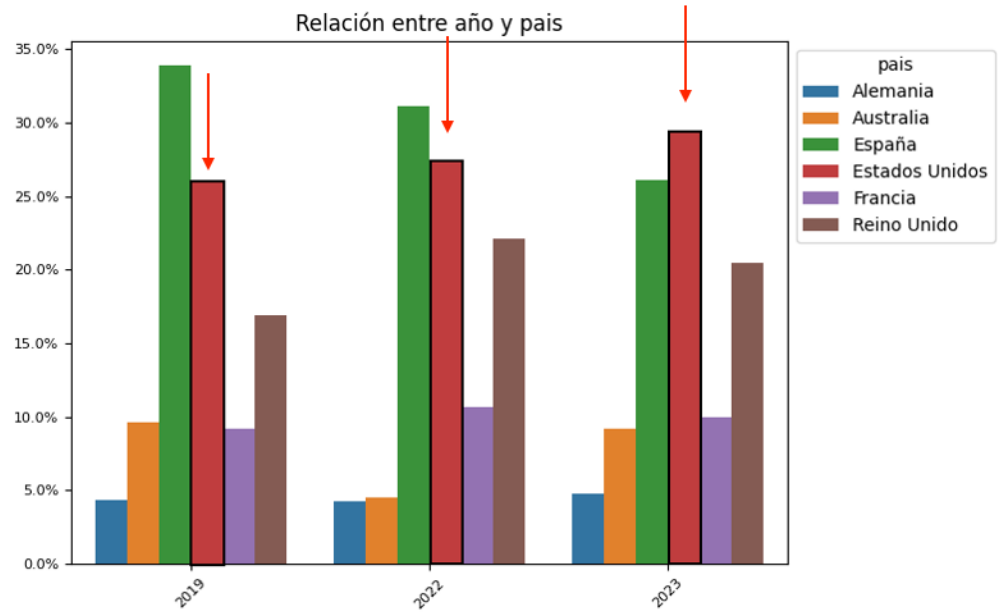
Segmentos clave por alto crecimiento interanual

En el siguiente apartado se estudiarán aquellos grupos de viajeros en los que se observa mayor crecimiento interanual.

Los siguientes resultados son para el destino **San Sebastián de alquiler vacacional.**

Después de agrupar los viajeros por nacionalidad se detecta un crecimiento significativo en EE.UU. y un decrecimiento en el turista nacional.

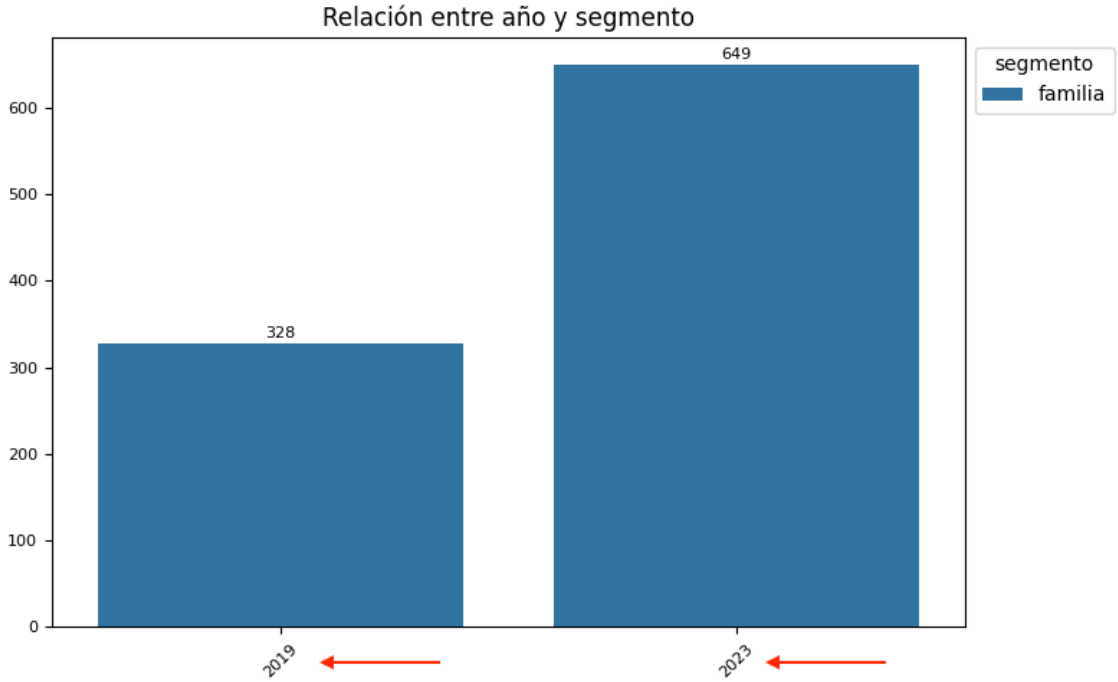
Porcentaje de reservas por año y nacionalidad en San Sebastián vacacional



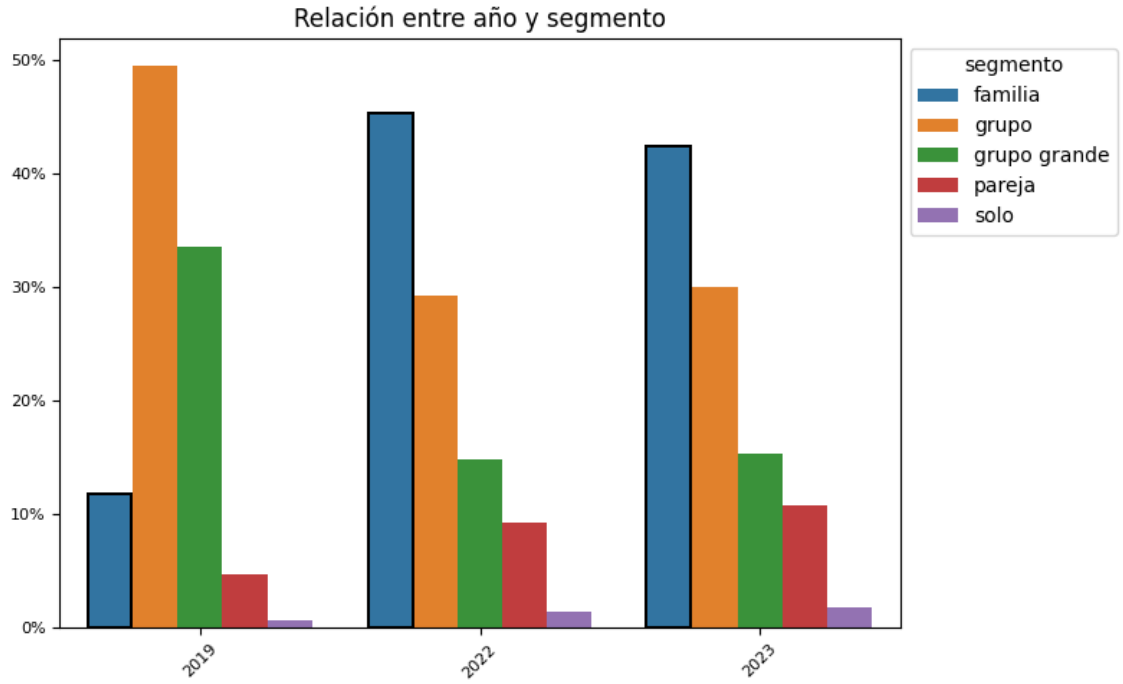
Por otro lado, para un mejor conocimiento de los viajeros se crean diferentes segmentos con esta clasificación:

- Solo: reservas de una sola persona.
- Pareja: reservas de dos personas.
- Familia: reservas de dos personas y al menos un niño.
- Grupo: más de tres personas.
- Grupo grande: más de cinco personas.

Después de diversos análisis puede observarse un **crecimiento mayor en el segmento familias**. El siguiente gráfico muestra las reservas de este segmento en 2019 y su evolución en 2023. Datos para Sebastián vacacional.



El crecimiento del segmento familias no es algo exclusivo del destino San Sebastián, el siguiente gráfico muestra el porcentaje de reservas por segmento y año en el **destino Baqueira.** Podemos ver un crecimiento muy importante. En el anexo 3 pueden verse todos los resultados de esta línea de análisis.



Los datos analizados muestran que por un lado que el grupo de viajeros de **origen norteamericano** crece en la ciudad y por otro el **segmento de familias en ambos destinos**.

Fortalezas del canal de venta propio

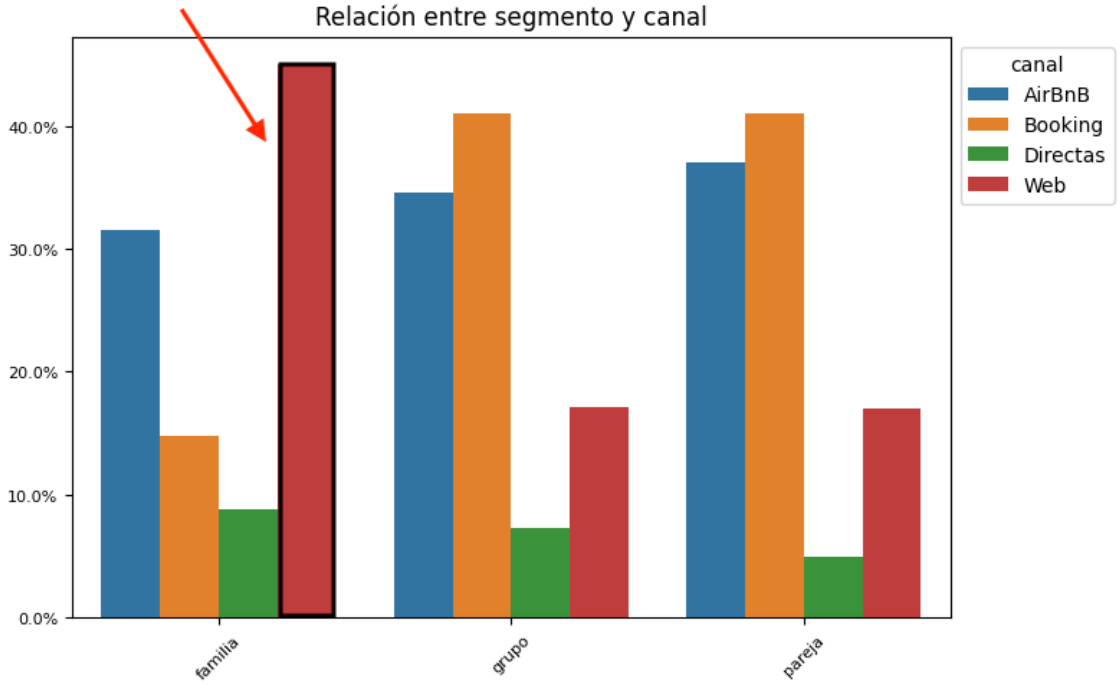
En el ámbito del alquiler de apartamentos turísticos, la dimensión del comercio electrónico juega un papel crucial, con la gran mayoría de las ventas realizándose online. En este contexto digital, diversas agencias especializadas desempeñan un rol esencial atendiendo a segmentos específicos de clientes y contribuyendo significativamente al volumen de ventas. Sin embargo, es importante destacar que estas reservas a través de agencias conllevan el pago de comisiones como honorarios por sus servicios.

Dada esta realidad, cobra especial importancia la potenciación del canal de venta directa, en este caso, a través del sitio web propio de la empresa. La venta directa ofrece una ventaja clave en términos de rentabilidad, ya que se evitan las comisiones asociadas a las reservas intermediadas por agencias. Al fomentar este canal, se busca no solo incrementar los márgenes de beneficio sino también fortalecer la relación directa con los clientes, permitiendo una mayor flexibilidad y personalización en la oferta de servicios.

En este estudio se han seguido diversas líneas de análisis tratando de detectar grupos de viajeros con una mayor preferencia por la reserva web para tratar de potenciarlo.

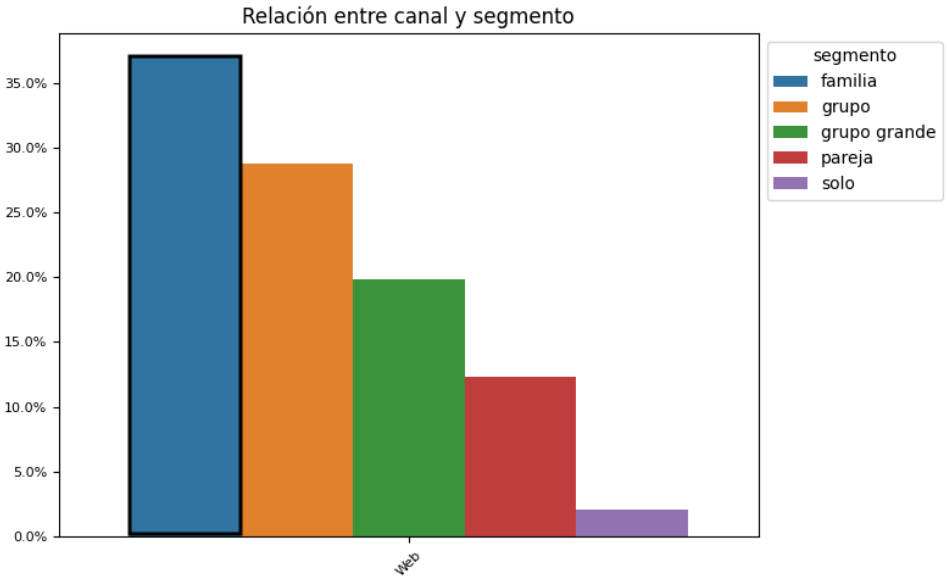
Los siguientes resultados son para el destino **San Sebastián de alquiler vacacional.**

El siguiente gráfico muestra las reservas agrupadas de los principales segmentos y canales de venta. Como puede verse hay una clara **preferencia de las familias** por el canal de venta propio, siendo casi el 50% de las reservas de este segmento **por la web propia**.



Si analizamos las reservas de más de 2.500 euros hechas por la web propia vemos que el mayor porcentaje corresponde al segmento familias. Además, como muestran resultados obtenidos en las líneas de análisis del anexo 3, **las familias tienen un mayor adr** en las reservas de apartamentos.

Reservas de más de 2.500 euros hechas en web por segmento en San Sebastián vacacional

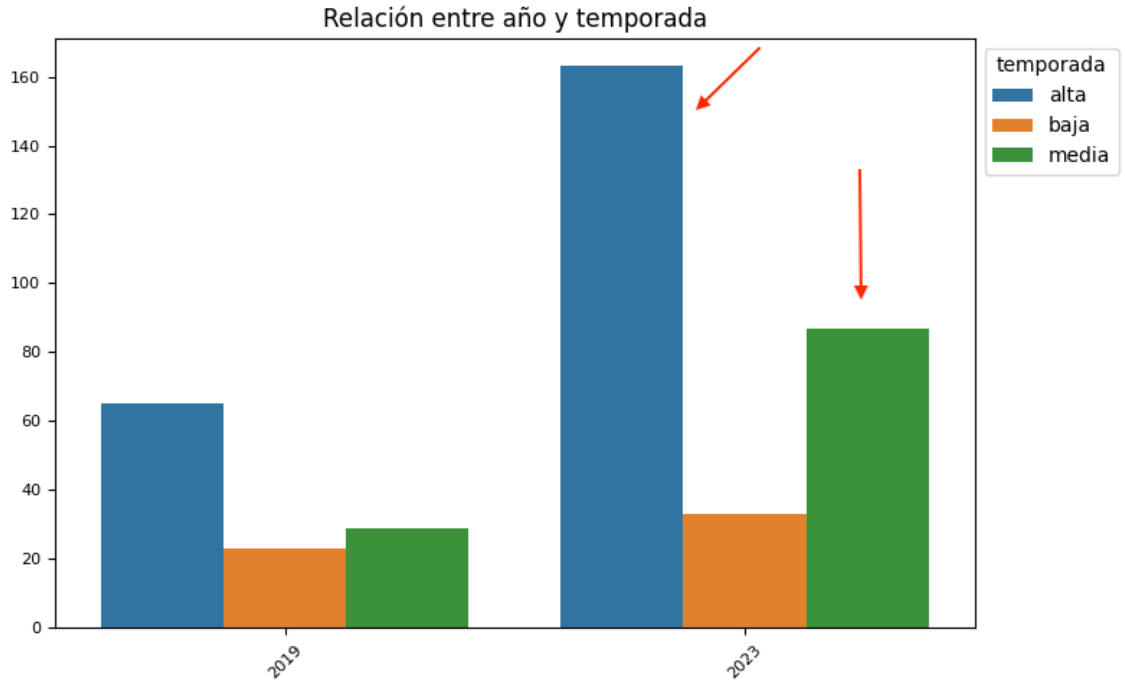


Además de lo anterior, puede verse un **crecimiento considerable del segmento familias**.
En el siguiente gráfico podemos ver las reservas del segmento familias hechas por la web propia por temporada en 2019 y en 2023.

Para estudiar las diferentes temporadas del año he realizado esta clasificación:

- Alta: junio a septiembre (incluidos).
- Media: octubre, mayo y abril.
- Baja: resto de meses.

Reservas de familias de web por temporada



A modo de conclusión podemos ver una **mayor preferencia de las familias por el canal propio**. Es lógico considerar que los viajeros con niños quieren tener todos los detalles de su viaje bien organizados y por ello prefieren la reserva directa con el gestor del alojamiento.

Segmentación por destino y modelo de negocio

Por último, este estudio pretende arrojar más datos sobre el perfil más común en los diferentes destinos y ambos modelos de negocio, el alquiler vacacional y el alquiler por temporadas. Como esta información es muy extensa, pueden consultarse todos los detalle en el anexo 3.

ANEXO 1: Data cleaning y Data mining

En mi estudio comienzo con un proceso de limpieza y procesado de los datos orientado a mejorar la calidad de la información que proporcionan e identificar errores.

Herramientas

Pandas: Pandas es fundamental para el manejo de datos. Utilizo esta biblioteca para cargar, manipular y analizar los datos en formato DataFrame. Es especialmente útil para la limpieza de datos, como la eliminación de duplicados, el manejo de valores nulos, y la transformación de formatos de datos. Funciones como read_excel, drop_duplicates, y fillna son algunas de las que empleo para estas tareas.

NumPy: NumPy es una biblioteca que me permite realizar operaciones matemáticas y estadísticas avanzadas. La utilizo para manipulaciones numéricas como cálculos estadísticos, operaciones con matrices y generación de números aleatorios. Funciones como mean, median, y std de NumPy me ayudan en el análisis estadístico de los datos.

Matplotlib y Seaborn: Para la visualización de datos, Matplotlib y Seaborn son mis herramientas preferidas. Estas bibliotecas me permiten crear gráficos y diagramas que facilitan la interpretación y presentación de los datos. Utilizo Matplotlib para personalizar gráficos detalladamente y Seaborn para generar visualizaciones más complejas de manera sencilla, como histogramas, diagramas de caja, y mapas de calor.

Funciones Personalizadas: Además de estas bibliotecas, utilizo funciones personalizadas, posiblemente almacenadas en un módulo denominado funciones. Estas funciones me permiten automatizar tareas repetitivas y específicas de mi proyecto, como transformaciones de datos particulares o cálculos estadísticos específicos.

Fecha y Hora (datetime): Para manejar datos de fecha y hora, empleo la biblioteca datetime. Es útil para convertir y manipular columnas de fecha y hora, lo que es crucial en muchos conjuntos de datos, especialmente para identificar tendencias a lo largo del tiempo.

Gender Guesser: Biblioteca de Python que se utiliza para predecir el género de una persona basado en su primer nombre. Esta herramienta se basa en un conjunto de reglas y datos que asocian nombres específicos con un género particular.

Procesos

Garantizar la Unicidad de los Datos: Verifico y elimino registros duplicados para asegurar la integridad y precisión de los análisis posteriores. Utilizo funciones de Pandas para identificar duplicados y los elimino del conjunto de datos.

```
df.duplicated().sum()
```

```
0
```

Revisión Exhaustiva de la Estructura de Datos: Analizo los tipos de datos presentes y reviso la existencia de valores nulos o faltantes. Esta tarea me ayuda a entender el contexto y la calidad general de los datos. Realizo una inspección detallada de cada columna para comprender su naturaleza y cómo deben ser tratados los datos.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36839 entries, 0 to 36838
Data columns (total 25 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   RESERVA_CODIGO        36839 non-null  int64
 1   Destino               36839 non-null  object
 2   Tipo alquiler         36839 non-null  object
 3   Apartamento           36839 non-null  object
 4   Nº habitaciones       36839 non-null  int64
 5   Zona                  36839 non-null  object
 6   Categoria             36839 non-null  object
 7   Tour Operador         36839 non-null  object
 8   Fecha reserva         36839 non-null  datetime64[ns]
 9   Fecha entrada         36839 non-null  datetime64[ns]
 10  Fecha salida          36839 non-null  datetime64[ns]
 11  date_cancel           36839 non-null  object
 12  PAIS                  36612 non-null  object
 13  Estado                36839 non-null  object
 14  noches                36839 non-null  int64
 15  revenue               36839 non-null  float64
 16  adr                   36839 non-null  float64
 17  sex                   31868 non-null  object
 18  num_people            36839 non-null  int64
 19  birth_date            36839 non-null  object
...
 23  name                  36834 non-null  object
 24  comunications_acepted  36839 non-null  object
dtypes: datetime64[ns](3), float64(2), int64(4), object(16)
memory usage: 7.0+ MB
```

Evaluación Estadística de Variables Numéricas: Realizo un análisis estadístico de las variables numéricas. Esto incluye calcular medidas de tendencia central como la media y la mediana, así como medidas de dispersión como el rango, la desviación estándar y los cuartiles.

```
df.describe()
```

	RESERVA_CODIGO	Nº habitaciones	Fecha reserva	Fecha entrada
count	36839.000000	36839.000000	36839	36839
mean	71381.538723	2.420505	2021-06-23 12:15:13.510138880	2021-08-20 16:55:46.019164416
min	45446.000000	0.000000	2018-02-08 00:00:00	2018-08-19 00:00:00
25%	61926.500000	2.000000	2019-12-31 00:00:00	2020-02-15 00:00:00
50%	71378.000000	2.000000	2021-09-28 00:00:00	2021-12-03 00:00:00
75%	80800.500000	3.000000	2022-10-24 00:00:00	2022-12-28 00:00:00
max	91725.000000	6.000000	2023-12-12 00:00:00	2023-12-31 00:00:00
std	11022.101444	0.962979	NaN	NaN

Identificación de Inconsistencias y Limpieza Detallada: Basándome en el análisis estadístico, identifico áreas que requieren una limpieza más profunda. Esto incluye el manejo de valores atípicos y la imputación o eliminación de valores nulos, según sea adecuado. En algunos casos, transformo los datos para mejorar su utilidad. Por ejemplo, normalizo algunas variables numéricas y convierto variables categóricas en un formato más adecuado para el análisis.

Trabajo con fechas, solo quiero aquellas reservas con entrada en 2019,2022 y 2023

```
filtro_fecha_1 = df['Fecha entrada'] < '2019-01-01'
fecha_inicio = '2019-12-31'
fecha_fin = '2022-01-01'
filtro_fecha_2 = (df['Fecha entrada'] > fecha_inicio) & (df['Fecha entrada'] < fecha_fin)
filtro_combinado = filtro_fecha_1 | filtro_fecha_2
fecha_limite = datetime(2023, 11, 30)
filtro_fecha_limite = df['Fecha entrada'] <= fecha_limite
filtro_final = ~filtro_combinado & filtro_fecha_limite
df = df[filtro_final]
```

Trabajo con la columna “sex”, uso gender guesser para estimar el género de las filas con datos nulos en función del nombre del viajero

```
df['nombre_2'] = df['name'].str.split().str[0]
mask = (df["sex"].isna()) & (df["Estado"] == "COMPLETADO")
import gender_guesser.detector as gender
d = gender.Detector()
df.loc[mask, 'genero'] = df.loc[mask, 'nombre_2'].apply(d.get_gender)
df['sex'] = df['sex'].fillna(df['genero'])
df.loc[df["Estado"]=="COMPLETADO"].sex.value_counts(dropna=False)
```

```
sex
M              9497
F              8100
unknown          54
male             18
female            4
-                 2
mostly_male       2
andy              1
Name: count, dtype: int64
```

Aplico la misma nomenclatura que los valores originales y donde no hay dato aplico “sin dato”

```
#aplico la misma nomenclatura que la columna original sex
sex_mapping = {"M":"M","F":"F",'male': 'M', 'mostly_male': 'M', 'female': 'F','andy': 'M'}
df['sex'] = df['sex'].map(sex_mapping)
#los que ya no puedo cambiar los relleno
df['sex'].fillna('sin dato', inplace=True)
df['sex'].replace('-', 'sin dato', inplace=True)
```

Detecto valores erróneos en el campo país para eliminarlos posteriormente

```
valores_menos_4_digitos = df[df['PAIS'].str.len() < 4]['PAIS'].unique()
valores_menos_4_digitos
```

```
array(['0', 'lv', 'ru', 'ec', 'dz', 'ua', 'vn', 'za', 'uy', 'cz', 'si',
       'ad', 'qa', 'my', 'jp', 'cl', 'do', 'tr', 'kh', 'eg', 'hk', 'bo',
       'cr', 'lu', 'hu', 'tw', 'rs', 'gu', 'cn'], dtype=object)
```

Transformo las fechas a un formato manejable

```
df['birth_date'] = pd.to_numeric(df['birth_date'], errors='coerce')
fecha_referencia = pd.to_datetime('1900-01-01')
df['fecha_nacimiento'] = fecha_referencia + pd.to_timedelta(df['birth_date'], unit='D')
df.fecha_nacimiento
```

```
158           NaT
159    1968-08-05
160           NaT
161    1969-12-08
162    1985-11-09
          ...
36141  1970-01-03
36142         NaT
36143  1965-11-01
36144  1993-02-07
36145  1989-04-14
Name: fecha_nacimiento, Length: 25284, dtype: datetime64[ns]
```

Preparación para Análisis Posteriores: Tras completar estos pasos, me aseguro de que los datos estén limpios, estructurados y listos para análisis más avanzados. Además añado información necesaria para posteriores anális mediante procesamiento de los datos existentes

Añado una columna con el campo edad en función de la fecha de nacimiento y el año de estancia

```
fecha_actual = datetime.now()
df['edad'] = ((fecha_actual - df['fecha_nacimiento']).dt.days) / 365.25
df['edad'] = df['edad'] - (2023 - df['año'])
df.edad=round(df.edad)
df['edad'] = pd.to_numeric(df['edad'], errors='coerce')
df.edad
```

```
158      NaN
159     51.0
160      NaN
161     50.0
162     34.0
        ...
36141   54.0
36142    NaN
36143   58.0
36144   31.0
36145   35.0
Name: edad, Length: 25284, dtype: float64
```

Añado nueva información

```
df['date_cancel'] = pd.to_datetime(df['date_cancel'], errors='coerce')#cambio fecha de cancelación a formato fecha
df["duración"]=df["Fecha salida"]-df["Fecha entrada"]
df["antelación_reserva"]=df["Fecha entrada"]-df["Fecha reserva"]
df["antelación_cancelación"]=df["Fecha entrada"]-df["date_cancel"]
df['día'] = df['Fecha entrada'].dt.day_name()
df['mes'] = df['Fecha entrada'].dt.month_name()
df['año'] = df['Fecha entrada'].dt.year
df.comunications_acepted=pd.to_numeric(df.comunications_acepted,errors='coerce')
```

Creo una segmentación de clientes en función del tipo y cantidad de viajeros en cada reserva

```
df.loc[(df['num_people'] == 1) & (df['children'] == 0) & (df['babies'] >= 0) , 'segmento'] = 'solo'
df.loc[(df['num_people'] == 2) & (df['children'] == 0) & (df['babies'] == 0) , 'segmento'] = 'pareja'
df.loc[(df['num_people'] >= 2) & ((df['children'] > 0) | (df['babies'] > 0)), 'segmento'] = 'familia'
df.loc[(df['num_people'] > 2) & (df['children'] == 0) & (df['babies'] == 0), 'segmento'] = 'grupo'
df.loc[(df['num_people'] > 5) & (df['children'] == 0) & (df['babies'] == 0), 'segmento'] = 'grupo grande'
```

Segmento clientes en diferentes grupos de edad

```
bins = [0, 25, 35, 45,55,65, 150]
etiquetas = ['0-24', '25-34', '35-44','45-54','55-65','66+']
df['edad_agrupada'] = pd.cut(df['edad'], bins=bins, labels=etiquetas, right=False)
```

Segmento por temporadas dentro del año

```
df['mes'] = pd.to_datetime(df['mes'], format='%B').dt.month
condiciones = [
    (df['mes'].isin([6, 7, 8, 9])),  # Junio a septiembre (alta)
    (df['mes'].isin([4, 5, 10])),     # Octubre, mayo, abril (media)
    (~df['mes'].isin([4, 5, 6, 7, 8, 9, 10]))  # Resto de meses (baja)
]
etiquetas = ['alta', 'media', 'baja']
df['temporada'] = np.select(condiciones, etiquetas, default='baja')
```

Agrupo los datos de Touroperador y País para poder manejar mejor esta información

```
#TOUR OPERADOR
frecuencias_tour_operador = df['Tour Operador'].value_counts()
top6_tour_operador = frecuencias_tour_operador.head(6).index
df['canal'] = df['Tour Operador'].where(df['Tour Operador'].isin(top6_tour_operador), 'Otros')
#PAIS
frecuencias_pais = df['pais'].value_counts()
top6pais= frecuencias_pais.head(6).index
df['pais_agrupado'] = df['pais'].where(df['pais'].isin(top6pais), 'Otros')
```

Por tratarse de datos erróneos para la modalidad de alquiler temporal elimino aquellas reservas inferiores a 15 noches

```
df = df.loc[~((df['Tipo alquiler'] == "Temporada") & (df['estancia_media'] < 16))]
```

Por tratarse de datos erróneos elimino aquellas reservas con antelación de reserva negativa

```
df = df[df['antelación_reserva'] >0]
```

Por último, creo subdataframes para un mejor análisis posterior de los datos

```
# DataFrame de alquiler vacacional y estado COMPLETADO
df_vacacional_confirmado = df[(df['Tipo alquiler'] == 'Turistico') & (df['Estado'] == 'COMPLETADO')]
df_vacacional_confirmado_BQ = df[(df['Tipo alquiler'] == 'Turistico') & (df['Estado'] == 'COMPLETADO')& (df['Destino'] == 'Baqueira')]
df_vacacional_confirmado_SS = df[(df['Tipo alquiler'] == 'Turistico') & (df['Estado'] == 'COMPLETADO')& (df['Destino'] == 'San Sebastián')]

# DataFrame de alquiler vacacional y estado CANCELADO
df_vacacional_cancelado = df[(df['Tipo alquiler'] == 'Turistico') & (df['Estado'] == 'CANCELADO')]

# DataFrame de alquiler temporal y estado COMPLETADO
df_temporal_confirmado = df[(df['Tipo alquiler'] == 'Temporada') & (df['Estado'] == 'COMPLETADO')]

# DataFrame de alquiler temporal y estado CANCELADO
df_temporal_cancelado = df[(df['Tipo alquiler'] == 'Temporada') & (df['Estado'] == 'CANCELADO')]
```

Desarrollo de una Tabla Descriptiva de Variables: Elaboro una tabla descriptiva que resume las características clave de las variables. En esta tabla, clasifico cada variable por tipo (numérica, binaria, categórica) y les asigno una importancia inicial basada en mi comprensión de los datos. Esta tabla descriptiva facilita la identificación rápida de las características clave de cada variable y sirve como una guía para análisis más detallados.

Muestra de la tabla

Columna/Variable	Descripción	Tipo de Variable	Importancia Inicial	Nota
RESERVA_CODIGO	Código de reserva	Numérica Continua	Baja	Identificador único de cada reserva.
Destino	Destino del alquiler	Binaria	Media	Indica el destino de la reserva, Baqueira o San Sebastián.
Tipo alquiler	Tipo de alquiler realizado	Binaria	Media	Diferencia entre dos tipos de alquiler, corto y largo plazo.
Apartamento	Número identificativo del apartamento	Numérica Discreta	Baja	Identificador único de cada apartamento.
Nº habitaciones	Número de habitaciones del apartamento	Categórica	Media	Importante para clasificar los apartamentos por tamaño.
Zona	Zona geográfica del apartamento	Categórica	Media	Puede ser relevante para análisis geográficos.
Categoria	Categoría de calidad	Categórica	Media	Útil para segmentación de mercado.
Tour Operador	Identificador del canal de venta	Numérica Discreta	Alta	Relevante para análisis de colaboradores.
Fecha reserva	Fecha en que se realizó la reserva	Numérica Discreta	Alta	Crucial para análisis de tendencias y temporadas.
Fecha entrada	Fecha de entrada al apartamento	Numérica Discreta	Alta	Crucial para análisis de tendencias y temporadas.
Fecha salida	Fecha de salida del apartamento	Numérica Discreta	Alta	Útil para calcular la estancia media
fecha_cancelación	Fecha en que se canceló la reserva	Numérica Discreta	Media	Útil para análisis de cancelaciones y políticas de retención.
pais	País de origen del cliente	Numérica Discreta	Alta	Valioso para marketing y adaptación cultural.
Estado	Estado de la reserva	Binaria	Media	Indica si la reserva está activa o cancelada.
noches	Número de noches de la estancia	Numérica Discreta	Alta	Fundamental para cálculos de ocupación y facturación.
revenue	Ingresos generados por la reserva	Numérica Continua	Alta	Directamente relacionado con el rendimiento financiero.
adr	Tarifa media diaria	Numérica Discreta	Alta	Indicador clave de rendimiento.
sex	Sexo del cliente	Categórica	Alta	Puede ser relevante para análisis de mercado.
personas	Número de personas en la reserva	Numérica Discreta	Media	Importante para segmentación.
niños	Número de niños en la reserva	Categórica	Media	Importante para segmentación.
bebes	Número de bebés en la reserva	Categórica	Media	Importante para segmentación.
repetidor	Si el cliente es repetidor o no	Categórica	Alta	Útil para fidelización y estrategias de marketing.
comunications_acepted	Si el cliente aceptó comunicaciones	Binaria	Media	Importante para estrategias de marketing y comunicación.

ANEXO 2: Funciones de análisis

Estas funciones son herramientas de análisis de datos que facilitan la visualización y comprensión de datos categóricos y numéricos en un DataFrame. Estas funciones incluyen:

1. distribucion_categoricas: Visualiza la distribución de variables categóricas, puede mostrar la frecuencia relativa o absoluta según se requiera.
2. grouped_boxplots: Compara distribuciones numéricas en diferentes categorías utilizando boxplots agrupados. Esto es útil cuando se necesita analizar cómo varían los datos numéricos en diferentes grupos definidos por una variable categórica, lo que facilita la identificación de tendencias y patrones en los datos.
3. plot_categorical_numerical_relationship: Analiza la relación entre variables categóricas y numéricas. Puede calcular tanto la media como la mediana de la variable numérica para cada categoría categórica y representar esta relación en gráficos de barras.
4. plot_categorical: Visualiza la relación entre dos variables categóricas. Esto es valioso para comprender cómo se distribuyen las categorías de una variable en función de otra
5. cardinalidad: Evalúa la cantidad de valores únicos en las columnas y las clasifica en categorías. Esto es esencial para comprender la naturaleza de las columnas en un conjunto de datos, lo que puede guiar la preparación y el análisis de datos.

Estas funciones son útiles para explorar datos, identificar patrones y relaciones, y tomar decisiones informadas en el análisis de datos.

Ejemplo de la función distribucion_categoricas

```
def distribucion_categoricas(df, columnas_categoricas, relativa=False,
mostrar_valores=False):
    num_columnas = len(columnas_categoricas)
    num_filas = (num_columnas // 2) + (num_columnas % 2)

    fig, axes = plt.subplots(num_filas, 2, figsize=(12, 4 * num_filas))
    axes = axes.flatten()

    for i, col in enumerate(columnas_categoricas):
        ax = axes[i]
        if relativa:
            total = df[col].value_counts().sum()
            serie = df[col].value_counts().apply(lambda x: x / total)
            sns.barplot(x=serie.index, y=serie, ax=ax, palette='viridis',
hue=serie.index, legend=False)

            # Configura el eje Y en formato de porcentaje
            ax.yaxis.set_major_formatter(mtick.PercentFormatter(xmax=1,
decimals=0))

            if mostrar_valores:
                for p in ax.patches:
                    height = p.get_height()
                    ax.annotate(f'{height * 100:.0f}%', (p.get_x() +
p.get_width() / 2., height),
                                ha='center', va='center', fontsize=8,
xytext=(0, 5), textcoords='offset points')
```

```
        else:
            serie = df[col].value_counts()
            sns.barplot(x=serie.index, y=serie, ax=ax, palette='viridis',
hue=serie.index, legend=False)
            ax.set_ylabel('')

            if mostrar_valores:
                for p in ax.patches:
                    height = p.get_height()
                    ax.annotate(f'{height:.0f}', (p.get_x() + p.get_width() /
2., height),
                                ha='center', va='center', fontsize=8,
xytext=(0, 5), textcoords='offset points')

        ax.set_title(f'Distribución de {col}')
        ax.set_xlabel('', fontsize=5)
        ax.tick_params(axis='x', rotation=45)
        ax.set_ylabel('', fontsize=5)
        ax.tick_params(axis='both', labelsize=8)

    for j in range(i + 1, num_filas * 2):
        axes[j].axis('off')

    plt.tight_layout()
    plt.show()
```

ANEXO 3: Análisis de datos

Impacto en Revenue

Exploración inicial de datos

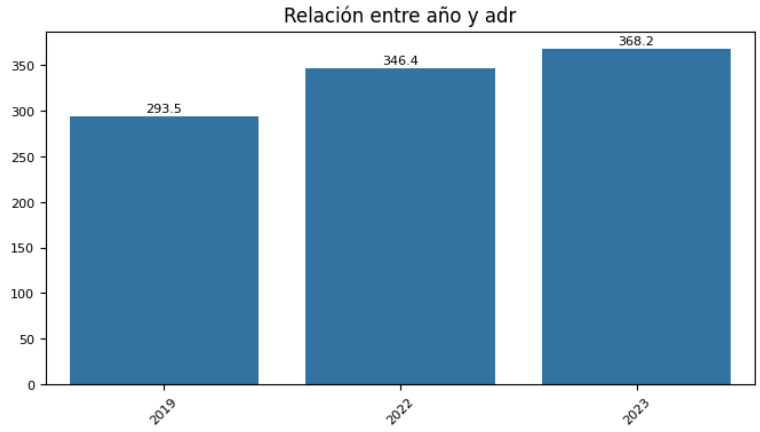
Para medir aquellas agrupaciones con un alto impacto en los ingresos vamos a usar 2 indicadores.

- ADR: Ingresos totales de la reserva / noches de la reserva, precio medio por noche reservada.
- Estancia media: número de noches por reserva.

Estos 2 indicadores representan muy bien el valor de cada reserva desde la perspectiva de ingresos y en general a mayores valores mayor revenue.

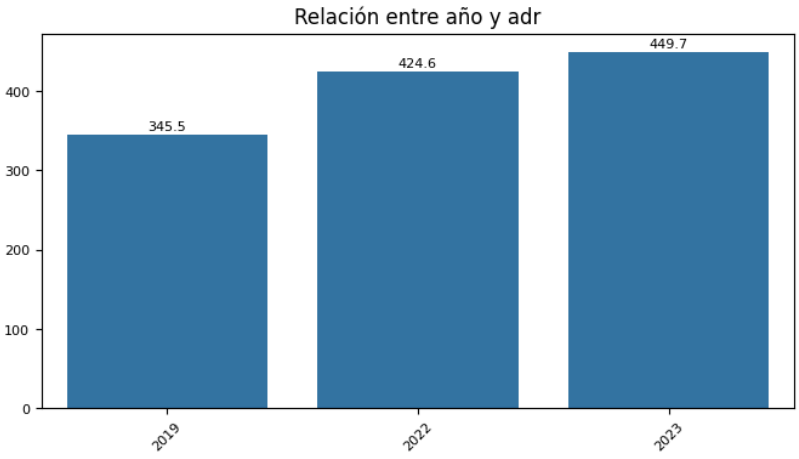
Para este estudio se analizan reservas de 2019, 2022 y 2023.

EVOLUCIÓN ADR anual
Por destino, San Sebastián, vemos una evolución muy positiva del precio medio, 26% en 2023 vs 2019

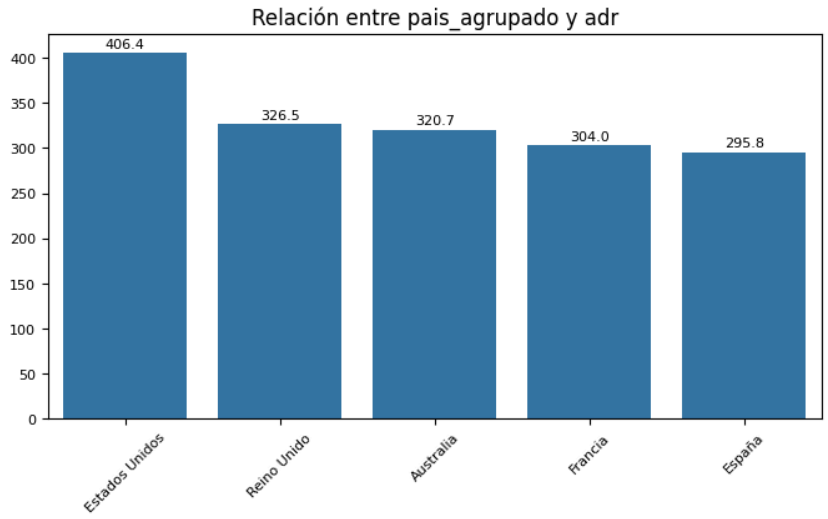


Analizando por temporada vemos que el mayor crecimiento es en temporada alta, no tanto en temporada media y casi no es significativo en temporada baja. Temporada alta comprende desde principio de junio a final de septiembre.

Precio medio por año en temporada alta (crece un 30% en 2023 vs 2019)



ADR y PAÍS
Por destino, San Sebastián, el mayor precio medio podemos observarlo en EEUU, el más bajo en España.

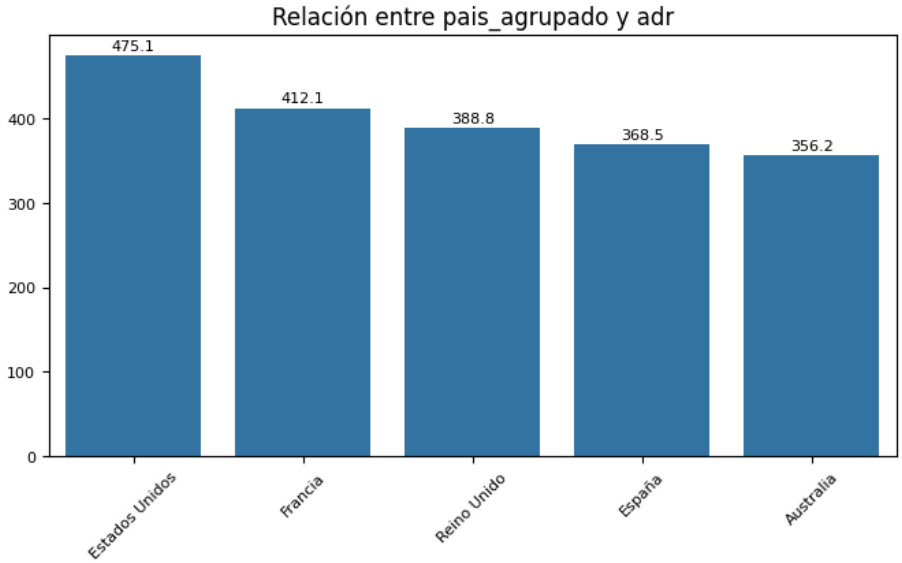


ADR y PAÍS por temporada
Por destino, San Sebastián.

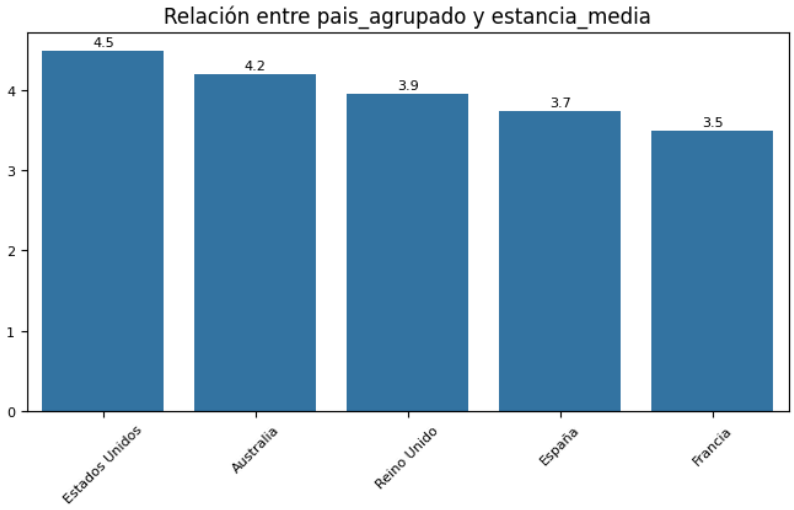
Hay una gran variación entre las diferentes temporadas del año en la ciudad, es por ello que para realizar un análisis más detallado sobre el precio medio agrupado por país debemos ir al detalle de la temporada y así descartar que el mayor precio medio de una nacionalidad solo se deba a su mayor peso en temporada alta donde los precios son más elevados.

Solo analizando las reservas de temporada alta vemos que EEUU sigue teniendo una diferencia significativa respecto al resto, por lo tanto el mayor adr global no solo viene de su mayor volumen de reservas en temporada alta

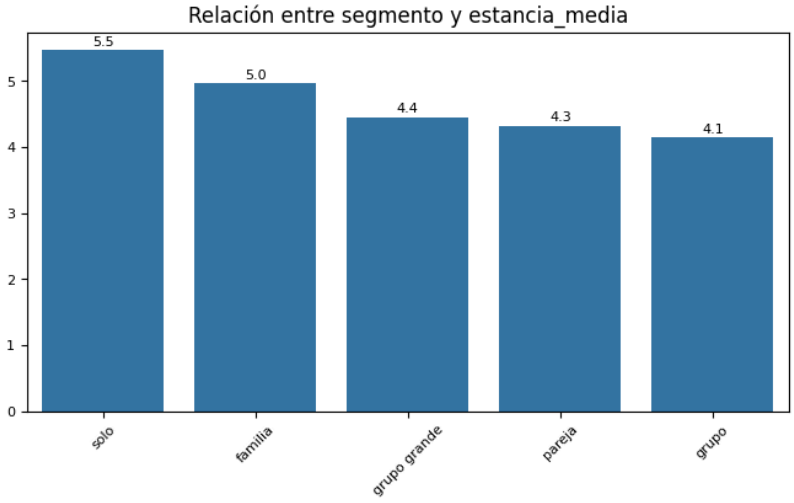
Adr por nacionalidad en temporada alta



ESTANCIA MEDIA y PAÍS
Por destino, San Sebastián vemos una mayor estancia media en Estados Unidos y Australia.

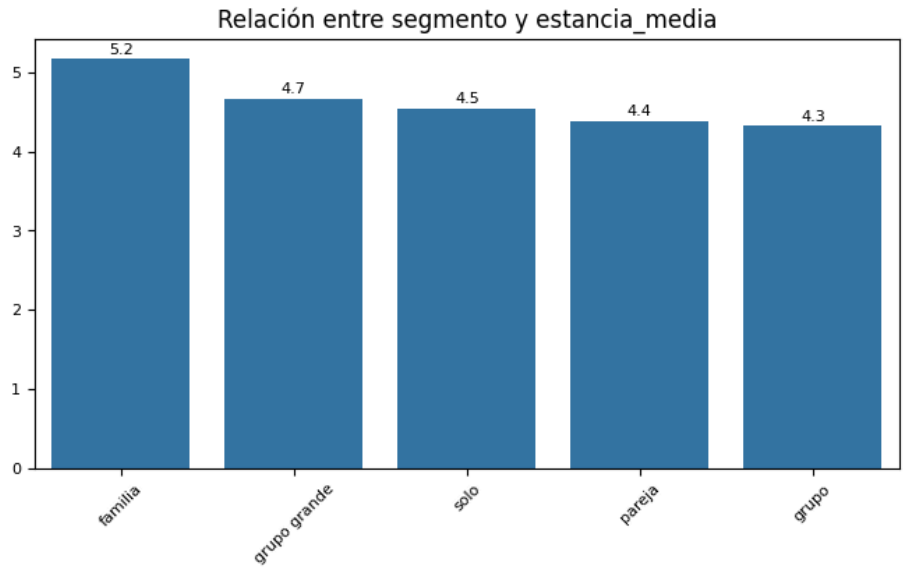


ESTANCIA MEDIA y SEGMENTO
Por destino, San Sebastián vemos una mayor estancia media en los viajeros solo y en las familias.



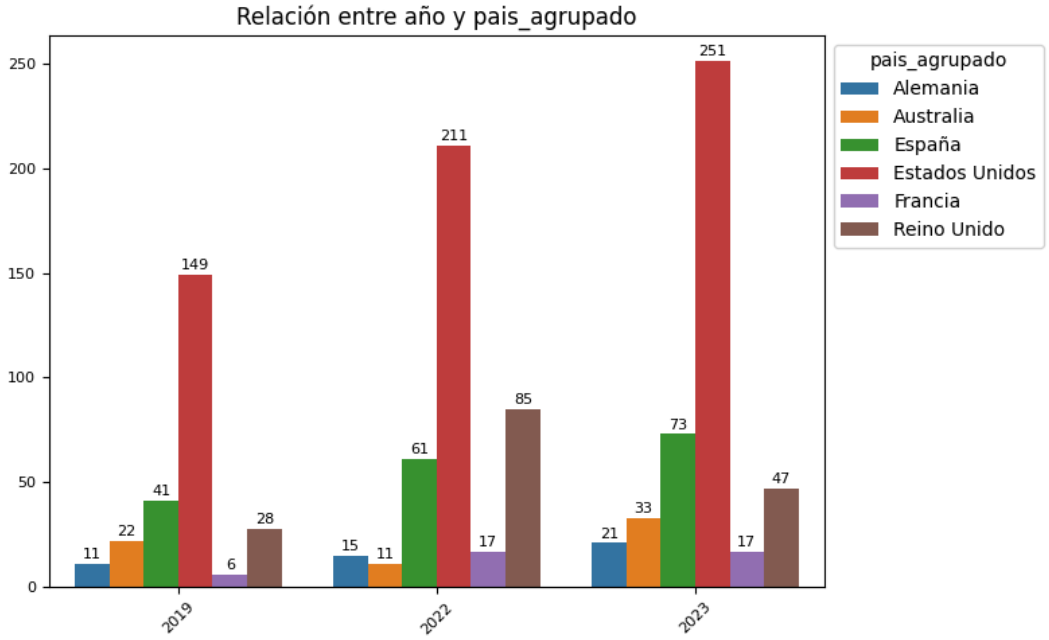
¿Cómo es en EEUU, el canal con mayor adr y estancia media?
La mayor estancia media la encontramos en el segmento familias

Estancia media por segmento de EE.UU.



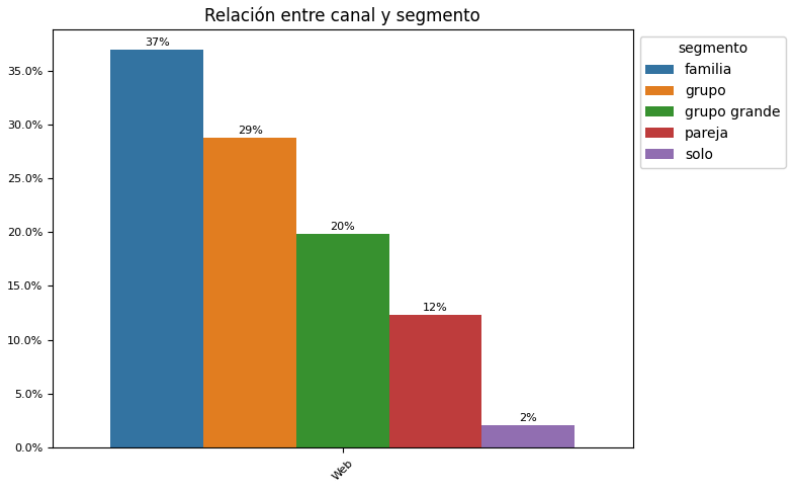
Con todo lo anterior puede verse que EEUU es la nacionalidad con el mayor impacto en revenue. Realizando un análisis solo de aquellas reservas de más de 2.500 euros vemos una clara mayoría de EEUU además de un crecimiento muy positivo

Reservas de importe alto por año y nacionalidad



Si además analizamos estas reservas de revenue alto y muy alto de EEUU por el canal Web vemos que el mayor porcentaje es del segmento familias

Reservas de importe alto por segmento en canal WEB

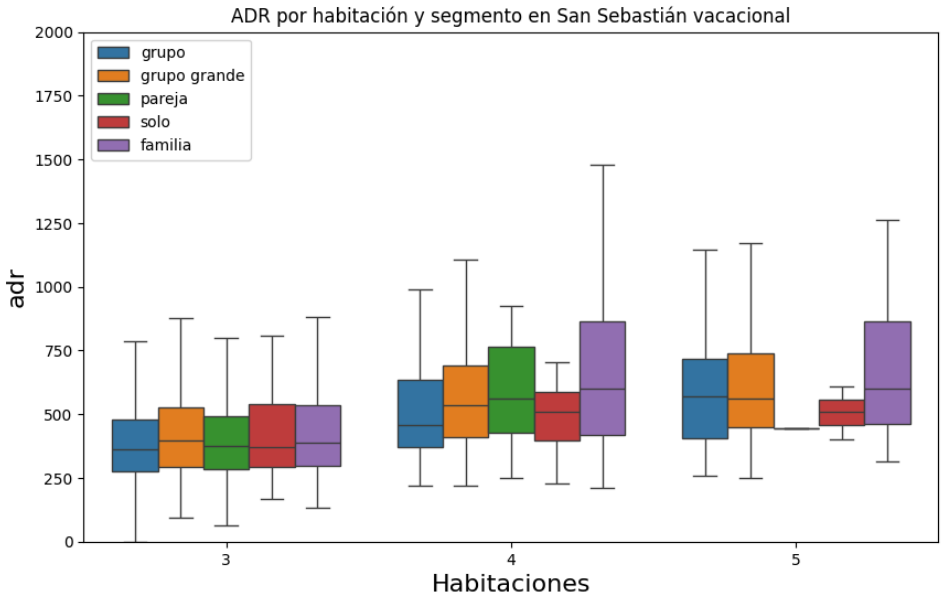


SEGMENTO Y ADR

Por destino, San Sebastián vemos que el segmento familia es el que tiene una distribución de valores en adr más alta para cada una de las diferentes tipologías de habitaciones, por encima de los grupos grandes. Parece verse una mayor preferencia en las familias por los apartamentos más exclusivos sin importar el precio.

Gráfico de caja: representación gráfica que muestra la distribución de un conjunto de datos. En él se visualizan el rango intercuartílico, la mediana, los valores extremos y los posibles valores atípicos.

Distribución de adr en los diferentes segmentos para 3h,4h y 5h (de 6h no hay registros)



Conclusiones

Los análisis realizados sacan a la luz que dentro de la mejora experimentada en precio medio en los últimos años en la ciudad de San Sebastián, la mayor aportación viene de la nacionalidad EEUU tanto por mayor adr como por mayor estancia media y por el segmento familias. La mayor apuesta para rentabilizar acciones comerciales sería apostar por el target familias y turistas americanos.

Perfil demográfico y geográfico: variación interanual

Exploración inicial de datos

Para este estudio se analizan reservas de 2019, 2022 y 2023.

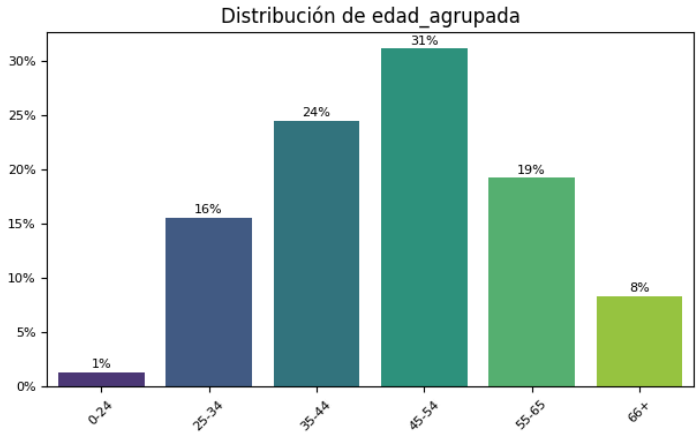
Reservas vacacionales por destino, vemos una amplia mayoría para la ciudad, San Sebastián. La temporada es más larga con un periodo de 4 meses de alta demanda (junio, julio, agosto y septiembre), 3 meses de temporada media (abril, mayo y octubre) y el resto del calendario temporada baja.

Por otro lado, en Baqueira la temporada alta apenas comprende 3 meses, diciembre, enero y febrero; marzo y agosto podrían considerarse temporada media y el resto prácticamente inexistente en cuanto a viajeros se refiere.

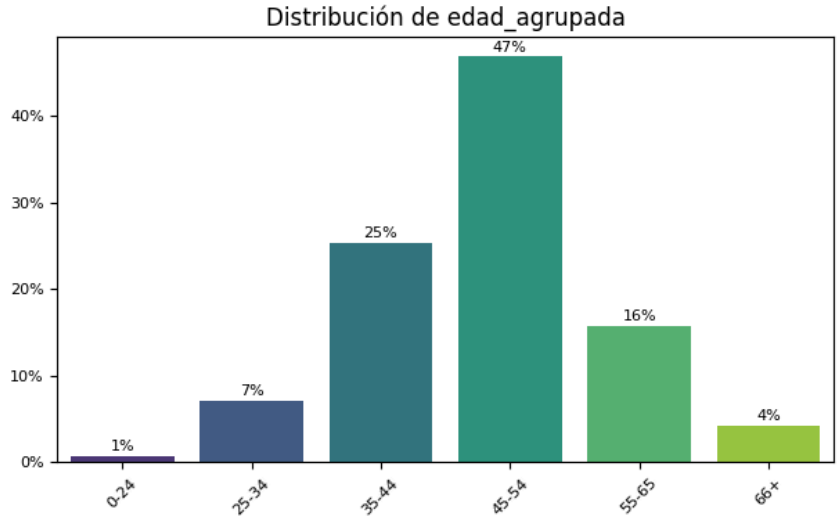
Con lo anterior puede verse una distribución de las reservas vacacionales desigual en los dos destinos teniendo la ciudad un peso mayor.

EDAD

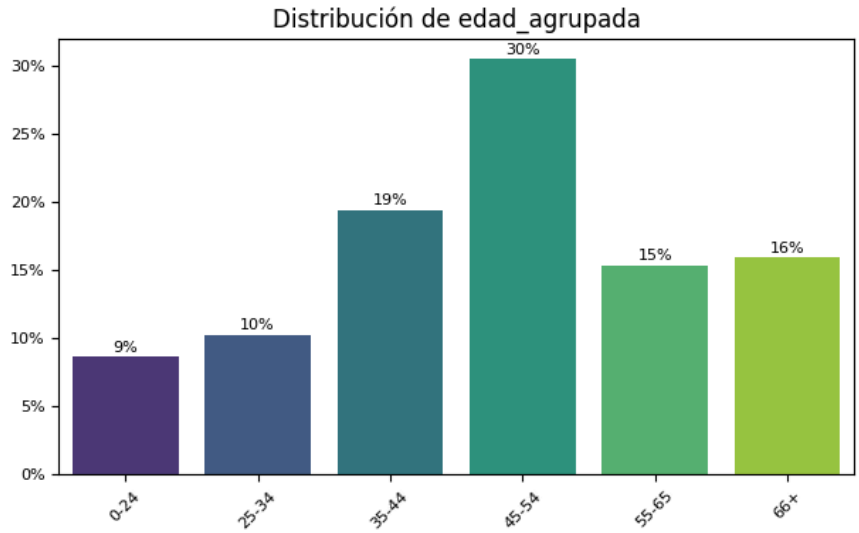
Vamos a analizar la distribución de las reservas en función de los diferentes rangos de edad.
Por destinos: San Sebastián



Por destinos: Baqueira



Por destinos: Temporal

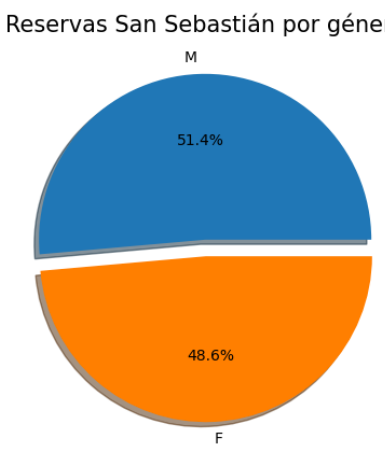


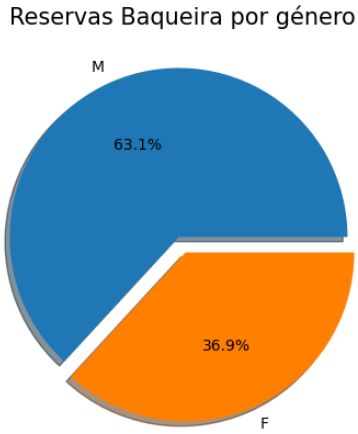
En la ciudad se observa una distribución más repartida entre los diferentes rangos de edad.
En Baqueira por el contrario muy concentrado en el rango 45-54 y poca presencia en rangos de edad jóvenes y senior.
En el modelo temporal destaca un mayor peso del perfil senior.

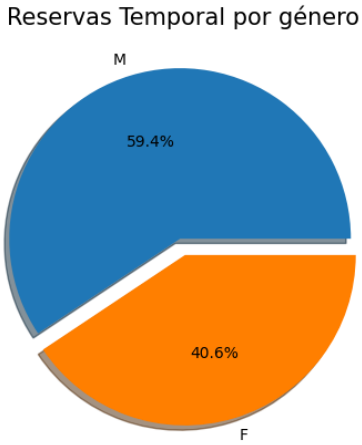
EDAD Y EVOLUCIÓN

Al analizar la evolución temporal en la distribución de las edades en los diferentes años no se aprecia una evolución marcada de ningún grupo en un análisis general.

GÉNERO







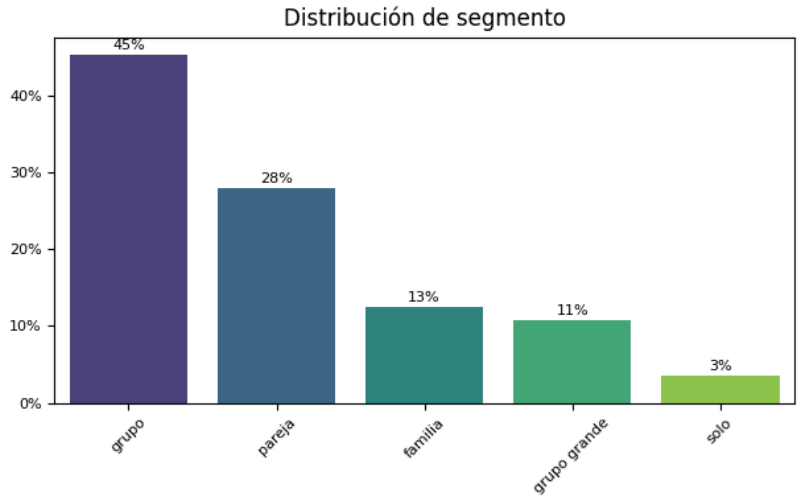
Tanto en las reservas de Baqueira como en las reservas temporales se observa una mayor proporción de reservas masculinas. Analizando la evolución temporal no se observan variaciones significativas.
Para posteriores análisis buscaremos relación de esta variable con otras como segmento o país.

SEGMENTO

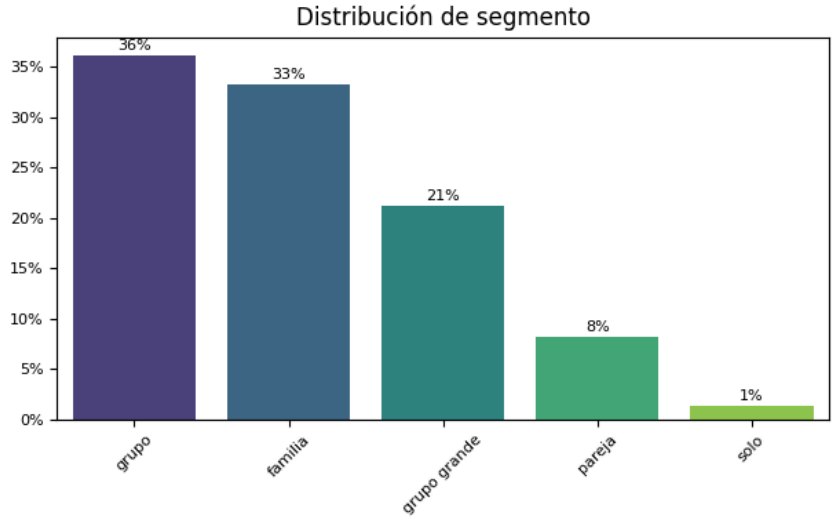
Para estudiar los segmentos se han creado los siguientes grupos:

- Solo: reservas de una sola persona.
- Pareja: reservas de dos personas.
- Familia: reservas de dos personas y al menos un niño.
- Grupo: más de tres personas.
- Grupo grande: más de cinco personas.

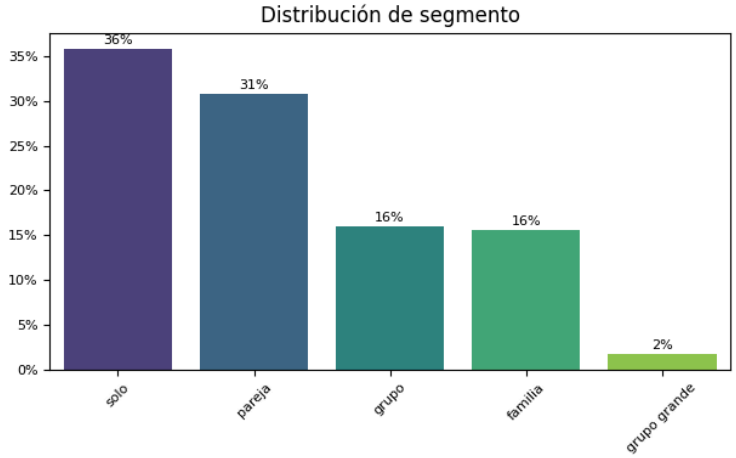
Por destinos: San Sebastián, destaca grupos y pareja.



Por destinos: Baqueira, destaca grupos y en segundo puesto las familias.

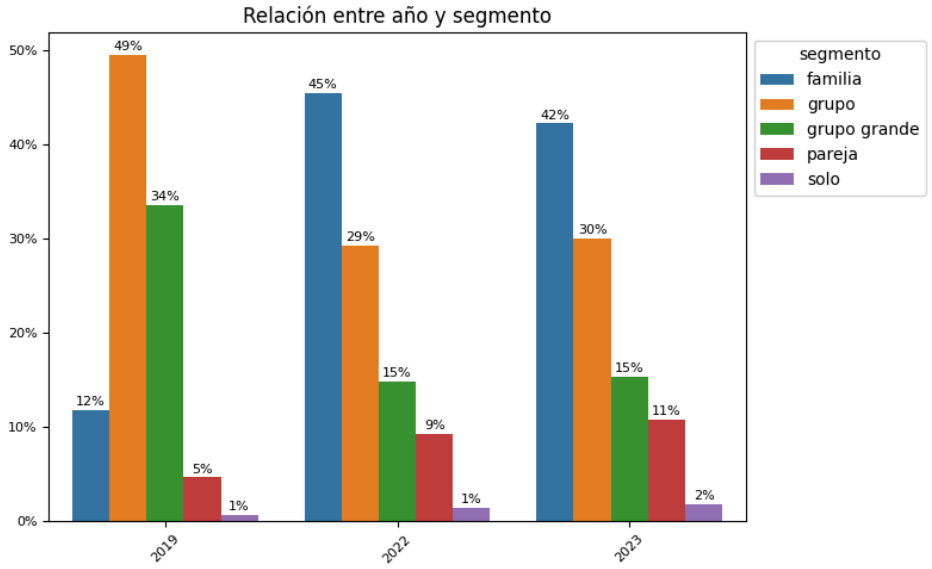


Por destinos: Temporal, destaca el viajero único y las parejas.

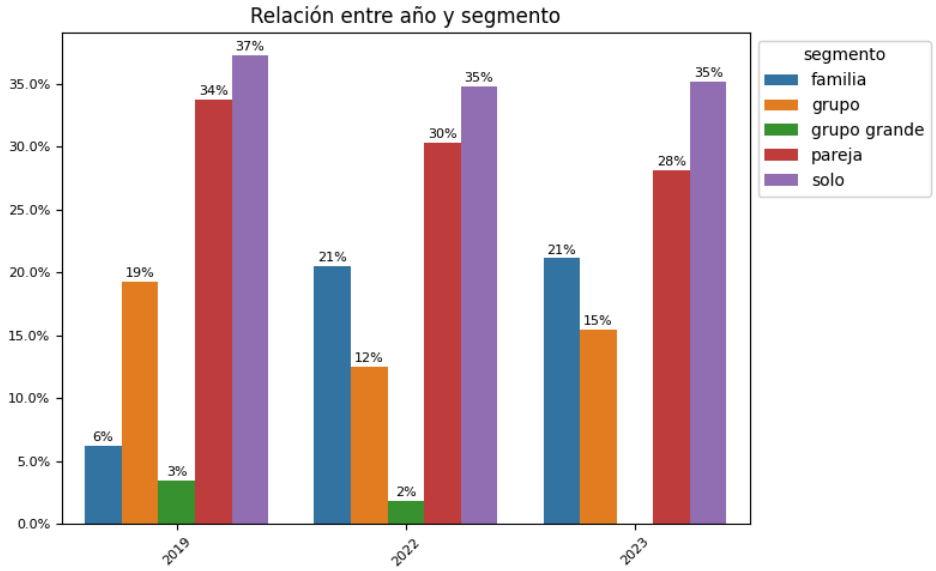


SEGMENTO Y EVOLUCIÓN

Por destinos: Baqueira

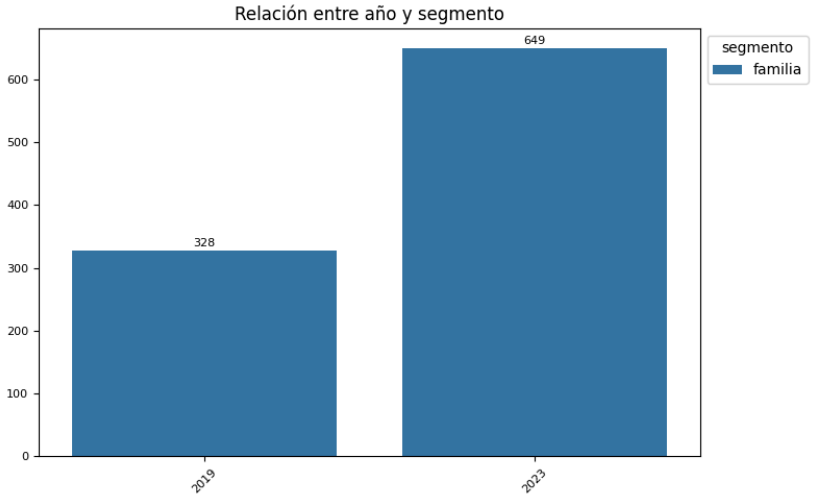


Por destinos: Temporal



Por destinos: San Sebastián

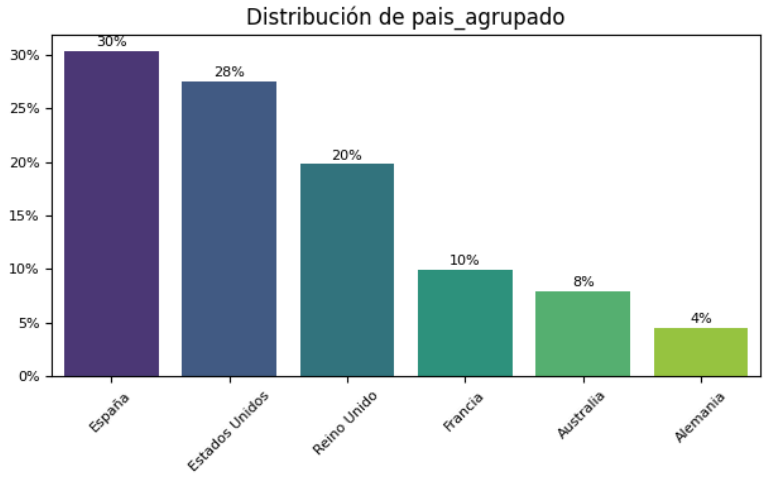
Número de reservas del segmento familia por año en San Sebastián vacacional.



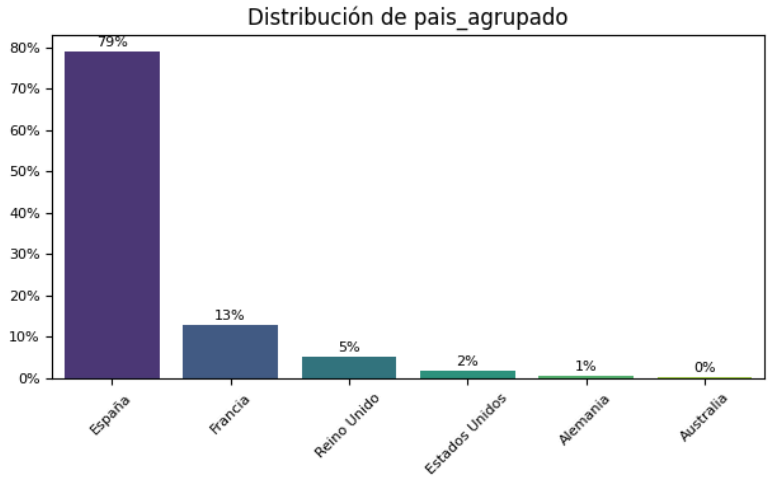
En general se ve un aumento significativo del segmento familia. Para siguientes apartados se analizará el segmento en función del canal y país para tratar de detectar nichos de crecimiento concretos.

PAÍS

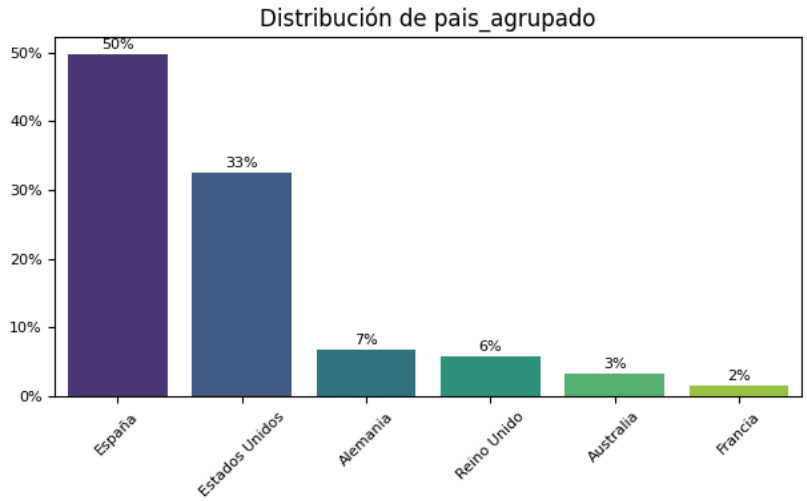
Por destinos: San Sebastián



Por destinos: Baqueira

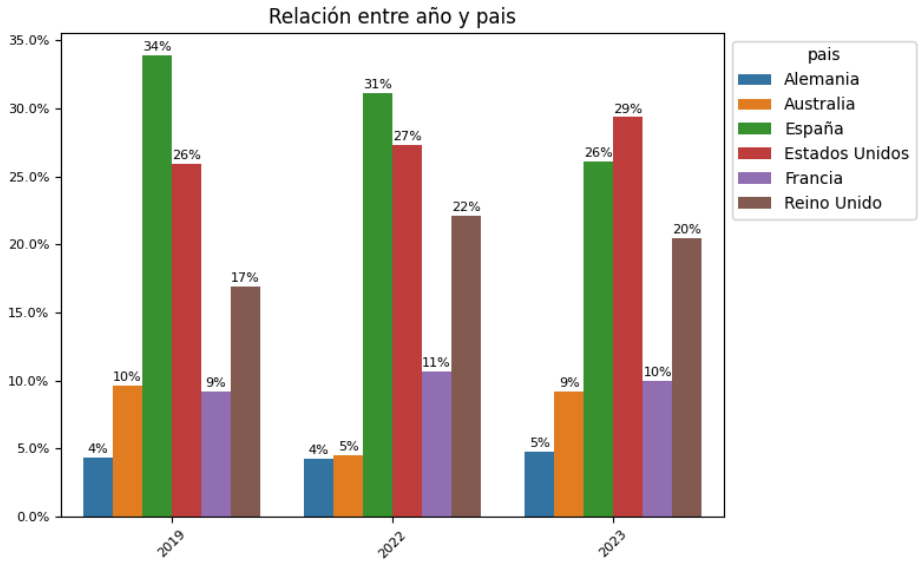


Por destinos: Temporal

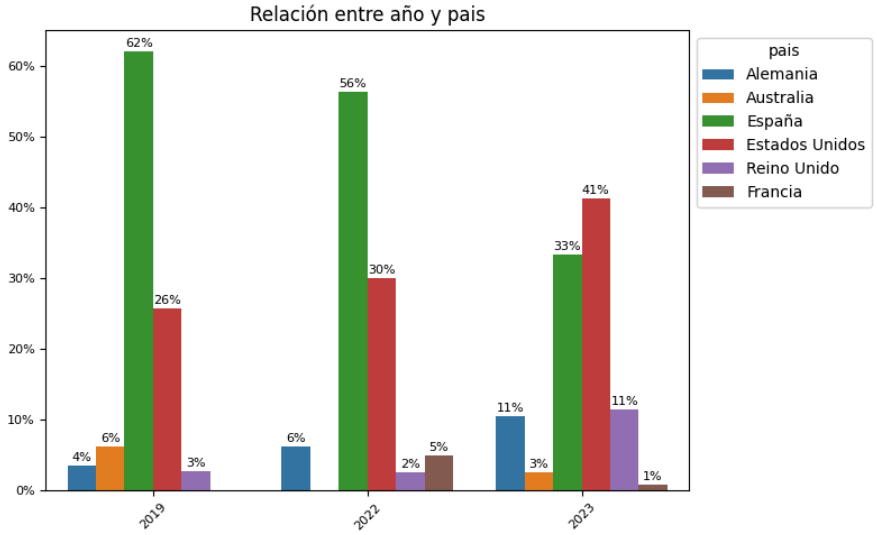


PAÍS Y EVOLUCIÓN

Por destinos: en Baqueira no se observan cambios significativos en la evolución interanual.
Por destinos: San Sebastián se ve una evolución positiva de EEUU y negativa de España

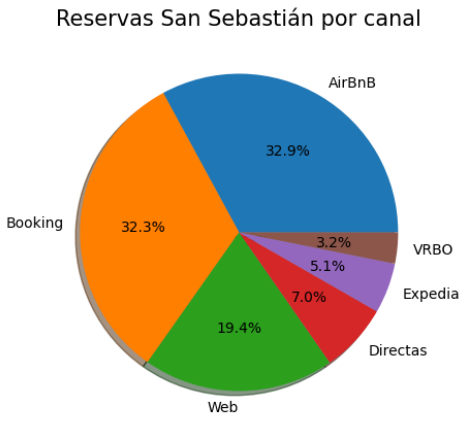


Por destinos: Temporal situación similar al modelo vacacional en San Sebastián

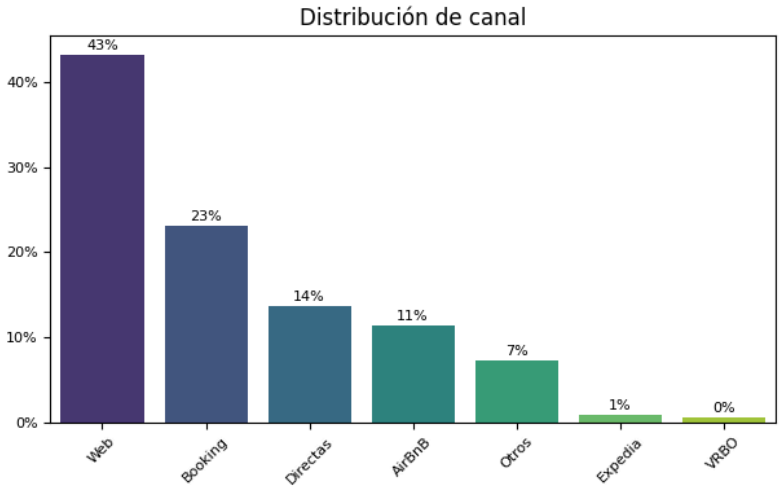


Para siguientes apartados se analizará el país en función del canal y segmento para tratar de detectar nichos de crecimiento concretos.

CANAL

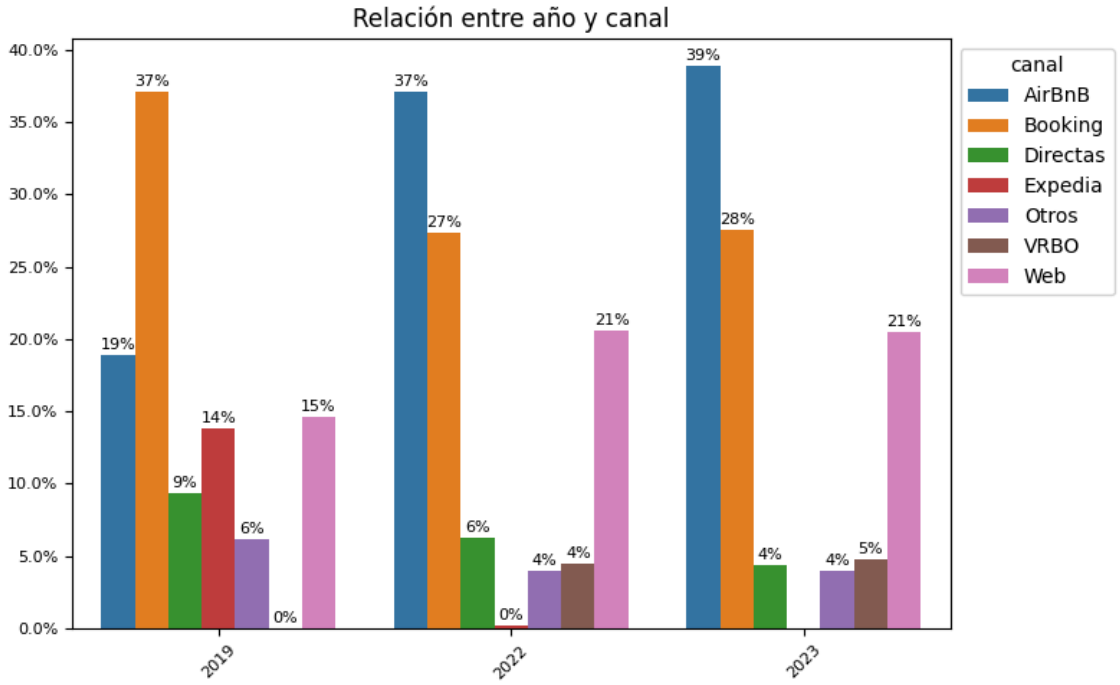


Por destino, Baqueira:



CANAL Y EVOLUCIÓN

Por destinos: San Sebastián, crecimiento de Airbnb y decrecimiento de Booking. Expedia desaparece tras el 2019.



Por destinos: en Baqueira, no se aprecian cambios significativos

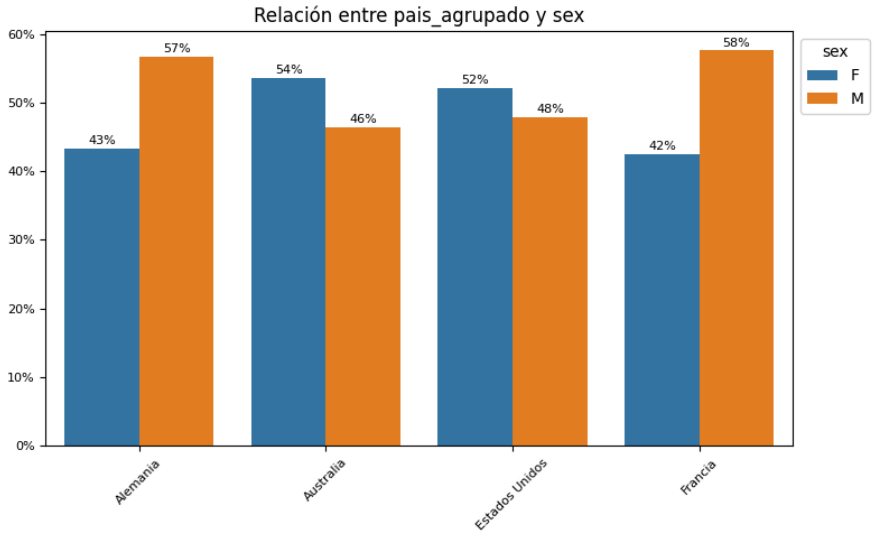
Análisis multivariante

En los siguientes análisis analizo la relación entre diferentes variables y cómo afectan a fenómenos específicos.

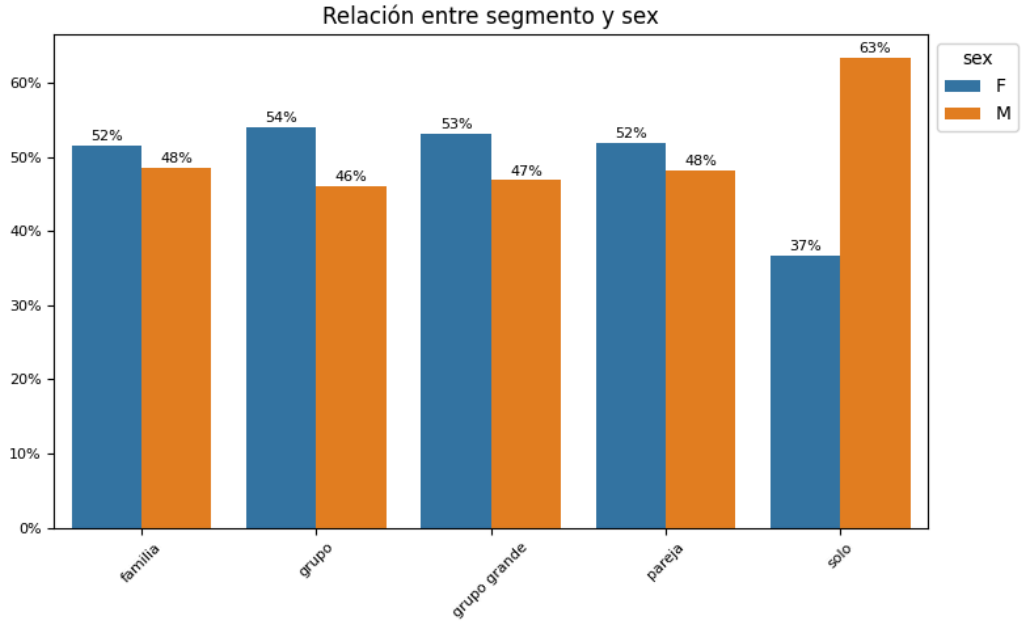
PAÍS Y GÉNERO

Por destinos: San Sebastián vacacional

En la distribución general observamos una cierta igualdad en reservas por género, analizando por país vemos que en algunos se observan desviaciones frente a la tendencia general: en los países europeos parece haber más tendencia a la reserva masculina, frente a origen americano o australiano donde tienen a reservar más las mujeres



¿Y cómo es la distribución por género en los diferentes segmentos para el viajero norteamericano de San Sebastián?



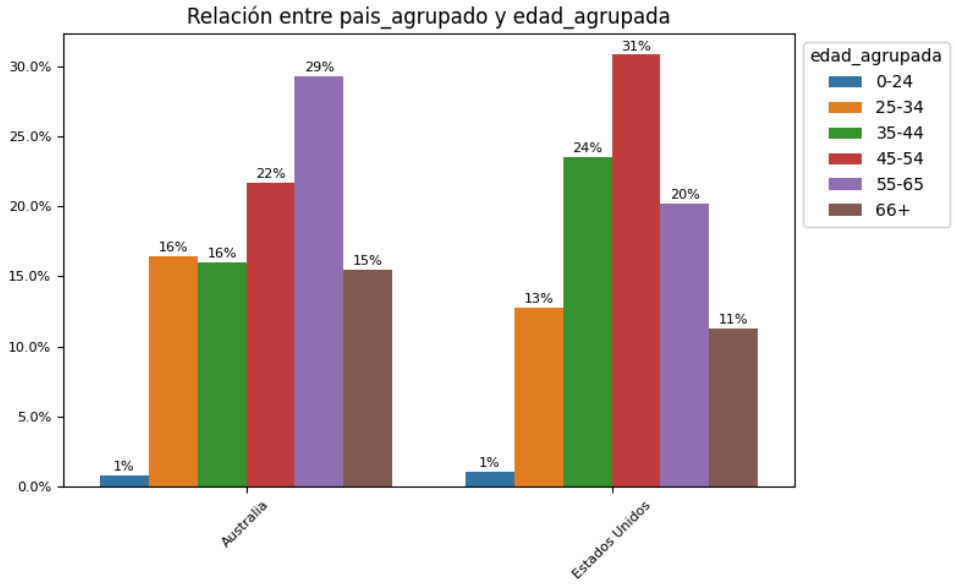
Vemos una distribución igual a la general salvo en los viajeros solo donde se observa más tendencia a la reserva masculina.

PAÍS Y EDAD

Por destinos: San Sebastián vacacional

Las mayores desviaciones frente a la distribución de edades global son en Australia y EEUU.
Australia: destaca por tener un perfil más "senior" de viajeros, donde el grupo de edad predominante es mayor de 55 años

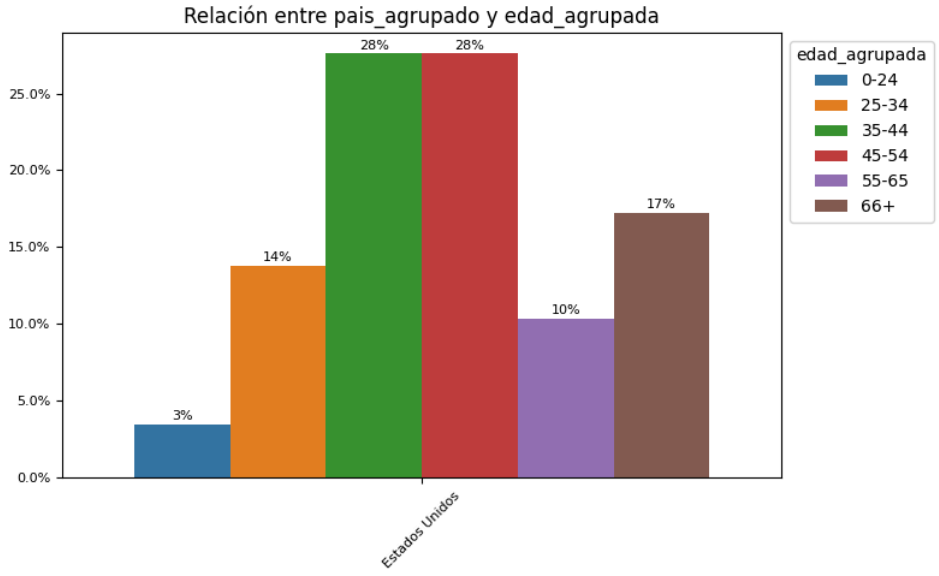
Estados Unidos: similar a Australia, vemos más viajeros de edad adulta, frente a la distribución de edades global



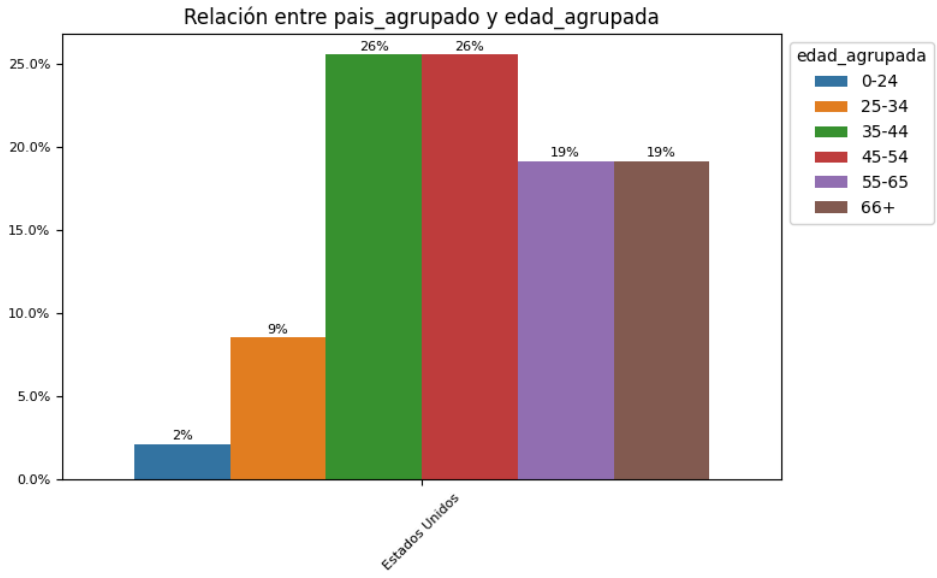
¿Y cómo es la evolución interanual en EE.UU?

Se puede ver una ligera tendencia de crecimiento en perfiles más adultos y decrecimiento en perfiles jóvenes. Estos datos muestran la evolución tanto de vacacional como temporal

EEUU, 2019:



EEUU, 2023:

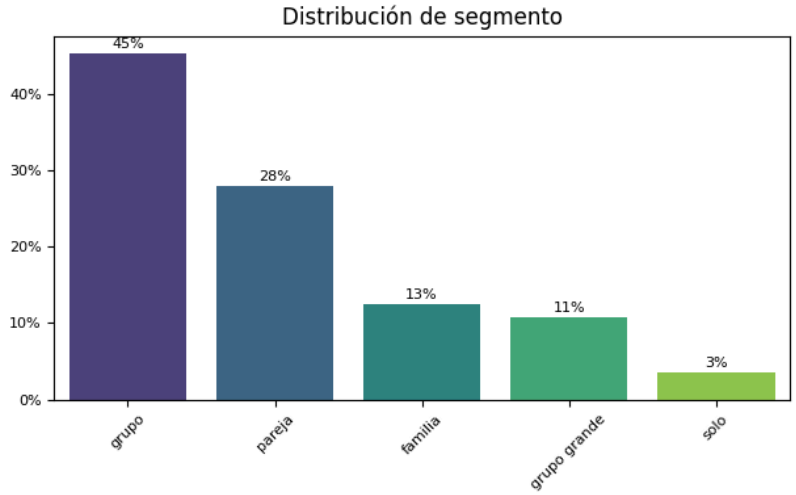


PAÍS Y SEGMENTO

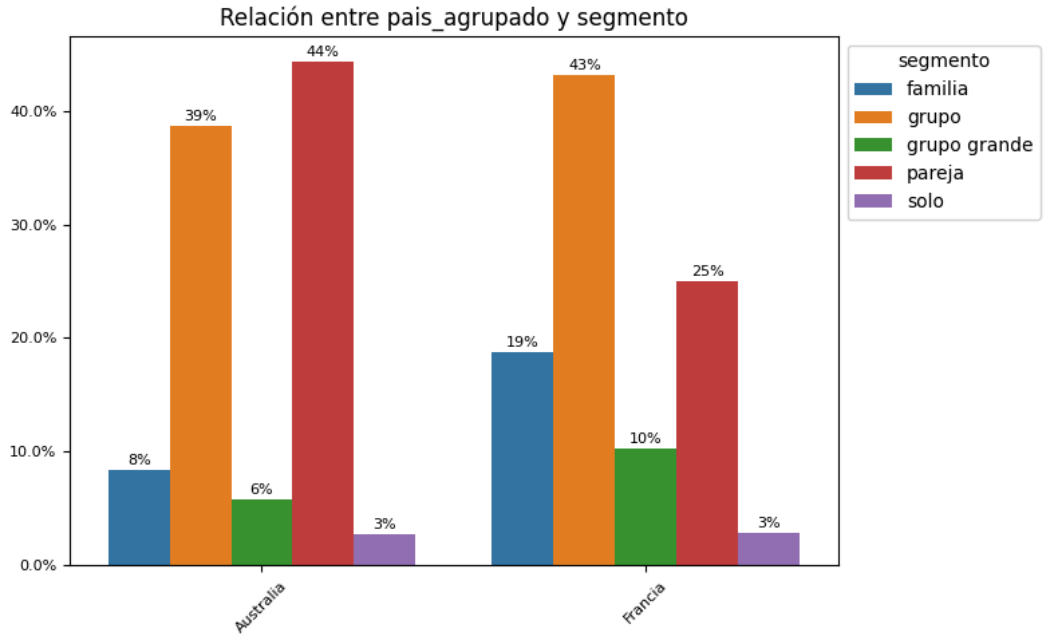
Por destinos: San Sebastián

La distribución general agregada indicaba una mayor presencia de grupos y pareja, veamos por nacionalidad

Distribución por segmento general de San Sebastián vacacional

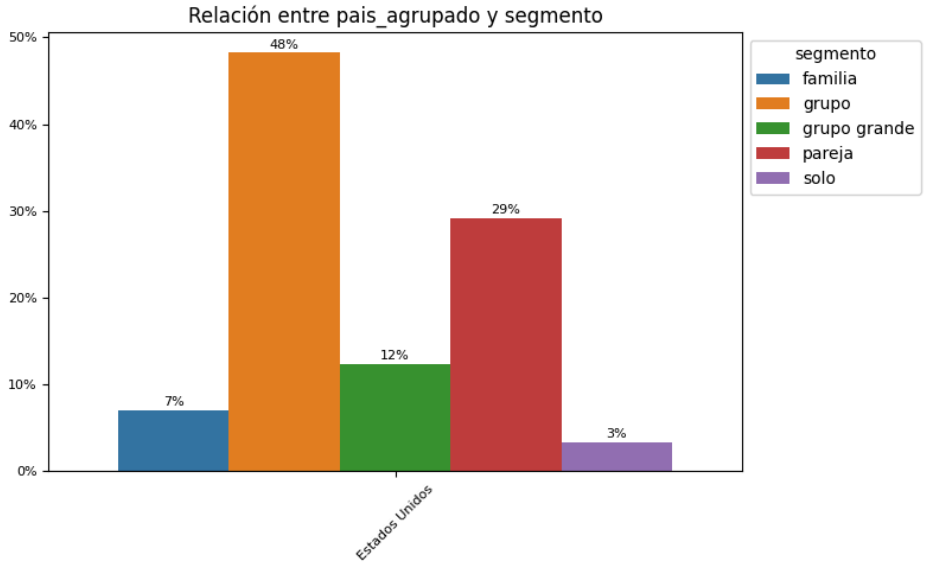


Analizando por país vemos que Australia destaca por mayor porcentaje de parejas y Francia por mayor porcentaje de familias.

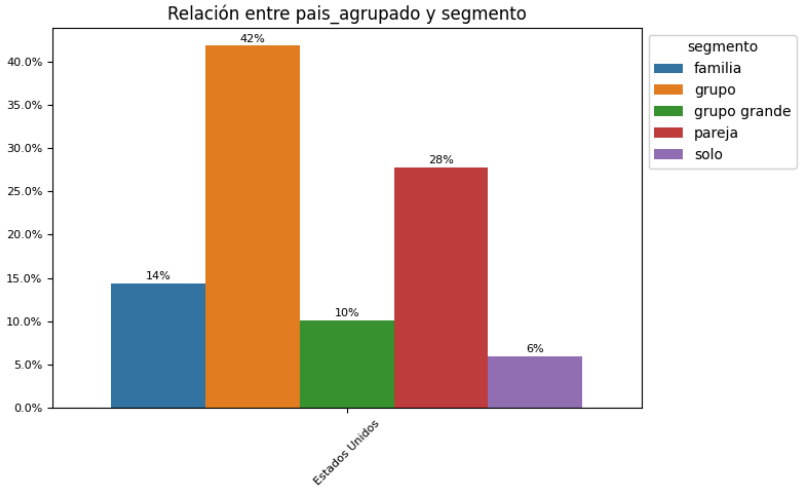


Al analizar la evolución interanual por país vimos que EEUU crecía y España decrecía.
Analizando por segmento, en EEUU vemos un ligero crecimiento de las familias en 2023 vs 2019

EEUU, 2019:



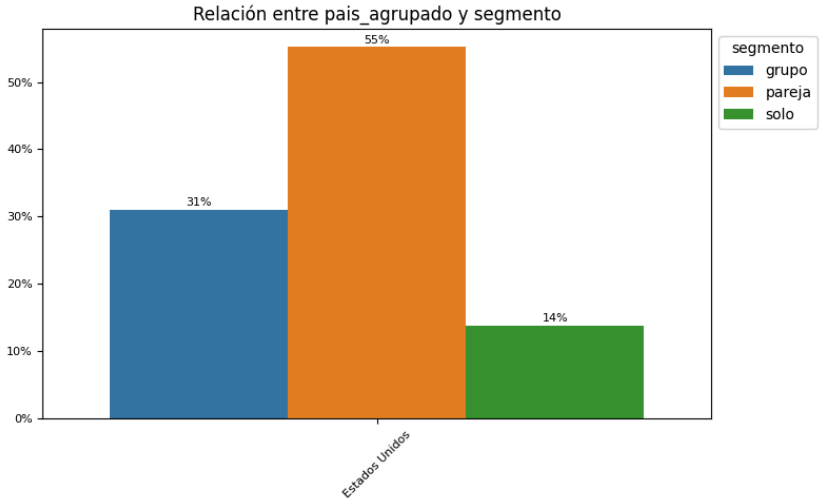
EEUU, 2023:



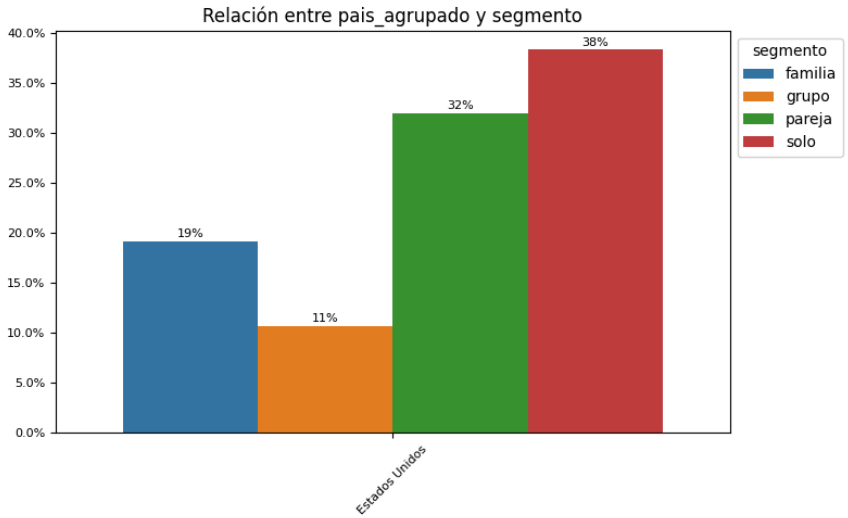
Por destinos: Temporal

En EEUU vemos como en 2023 crece el viajero solo frente a 2019

2019:



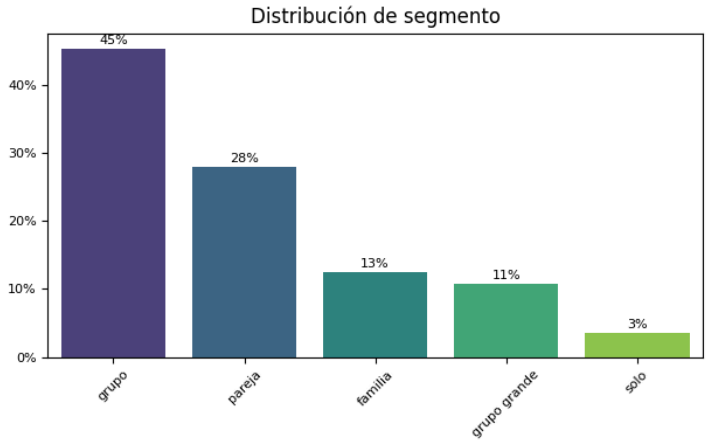
2023:



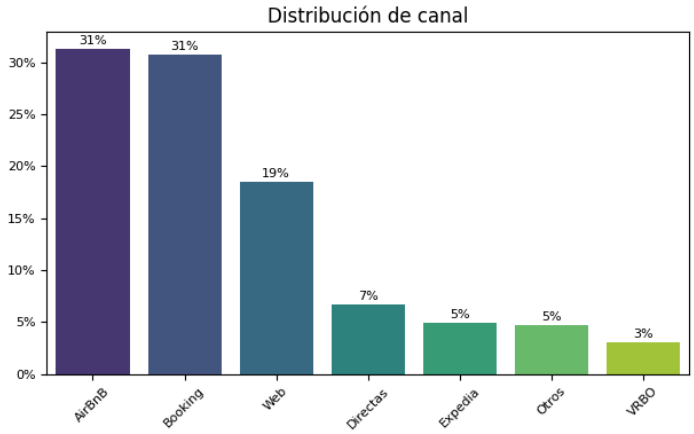
SEGMENTO Y CANAL

¿Los diferentes segmentos de viajeros reservan por igual en todos los canales?

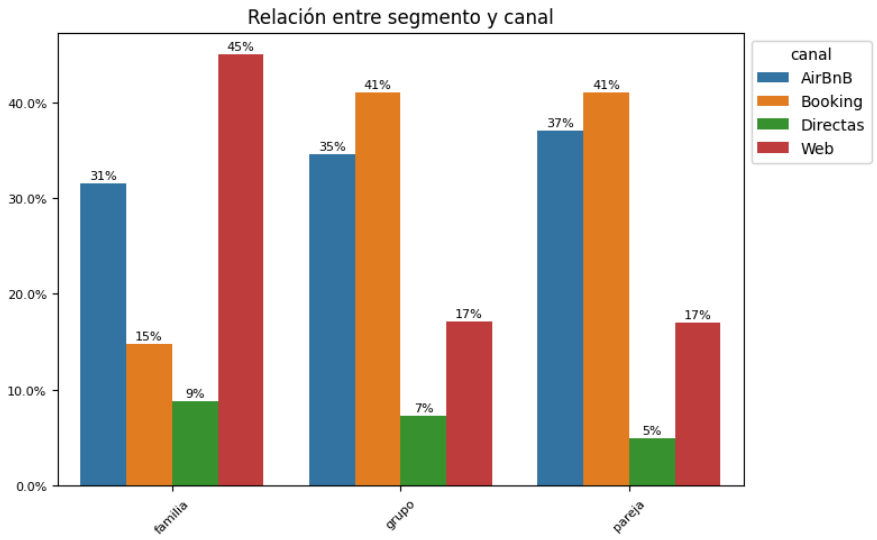
Esta es la distribución por segmento general para San Sebastián



Esta es la distribución por canal general para San Sebastián



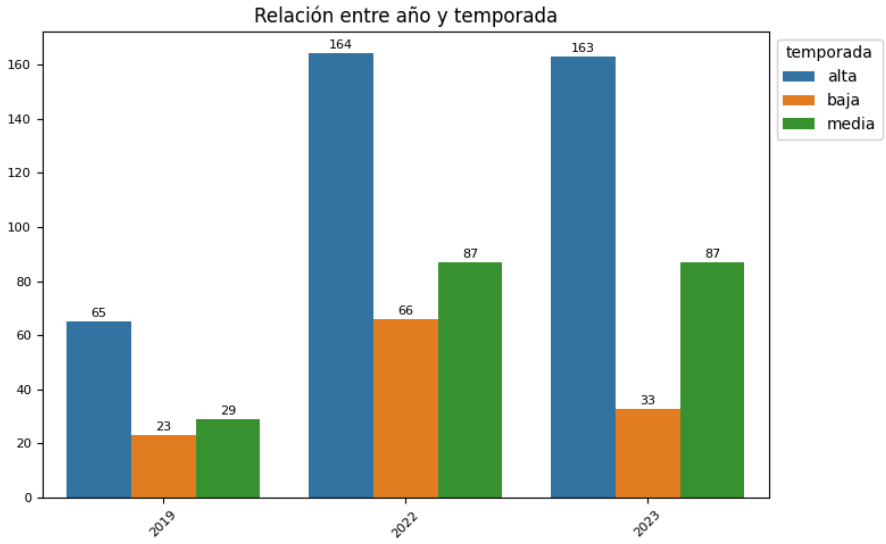
Al analizar por canal los 3 principales segmentos por los principales canales vemos lo siguiente:



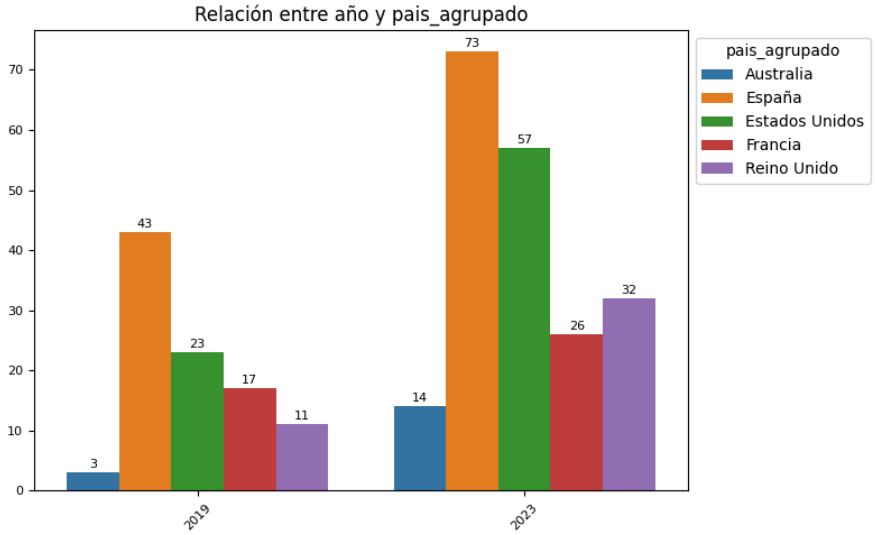
En el segmento familia las reservas Web suponen casi el 50% del total, parece verse una preferencia de este tipo de segmento por el canal Web.

Además si analizamos las reservas de familias por web en evolución temporal vemos un crecimiento considerable en el año 2023 vs 2019 sobre todo en temporada media y alta

Número de reservas web de familias por temporada



Analizando las reservas web de familias en los 4 principales países vemos que la evolución es igualmente positiva, siendo el que más aporta España.



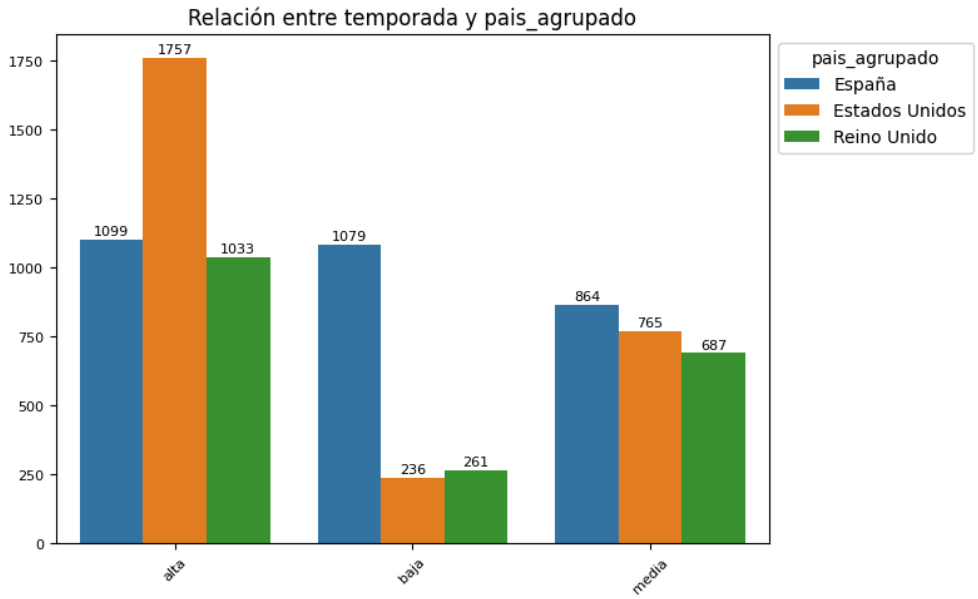
<u>TEMPORADA</u>

En San Sebastián podemos observar temporadas muy diferenciadas, teniendo meses con un nivel muy alto de reservas e ingresos y otros muy flojos. Para este estudio creamos las siguientes temporadas:

- Alta: junio a septiembre (incluidos).
- Media: octubre, mayo y abril.
- Baja: resto de meses.

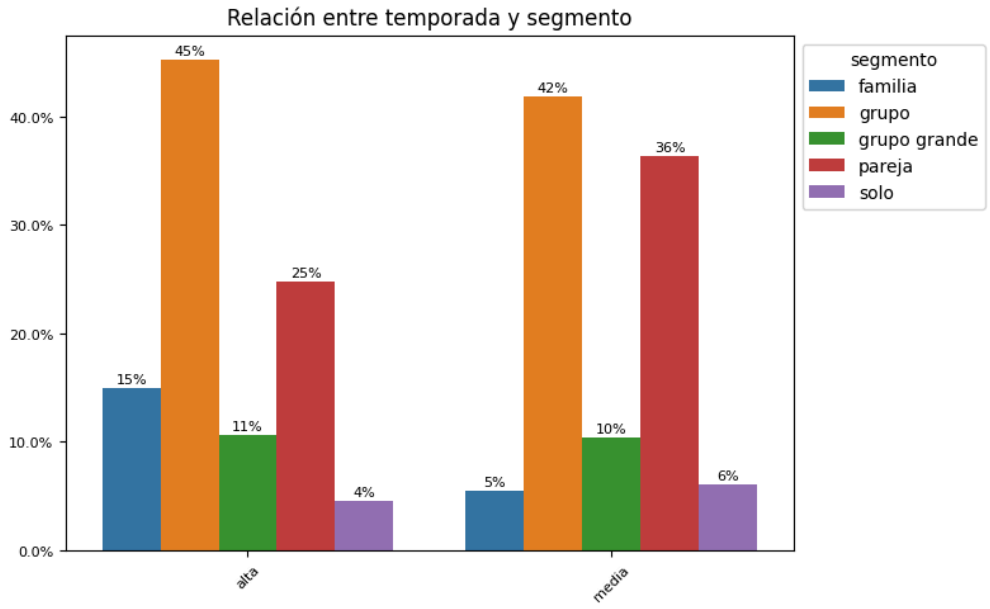
¿Cómo es la distribución de los principales mercados por temporada?, en temporada baja España es el principal emisor de reservas, en temporada alta Estados Unidos supone casi el 50% de las reservas y en temporada media está repartido por igual entre las tres:

Reservas por país y temporada del año



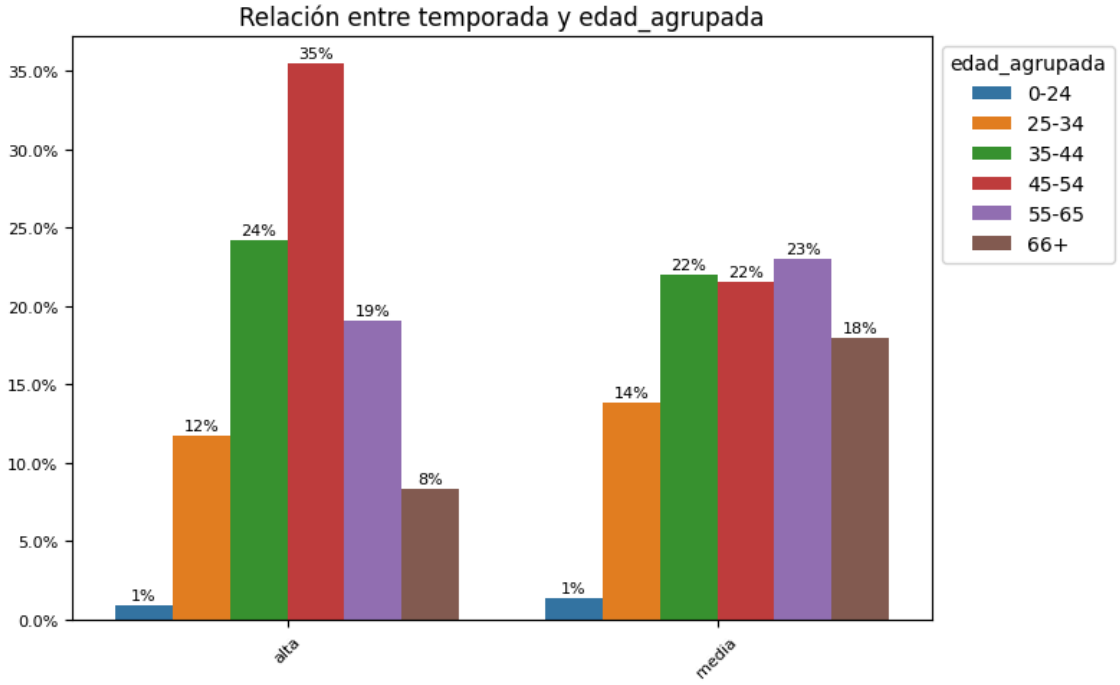
¿Hay diferencias significativas entre el tipo de segmento en EEUU por temporada?, los datos muestran que en temporada media hay una reducción de familias y aumento de parejas frente a temporada alta.

Reservas de EEUU en temporada alta y media por segmento



¿Hay diferencias en EEUU en la edad en diferentes temporadas?, los datos muestran en temporada media un aumento del perfil senior frente a temporada alta

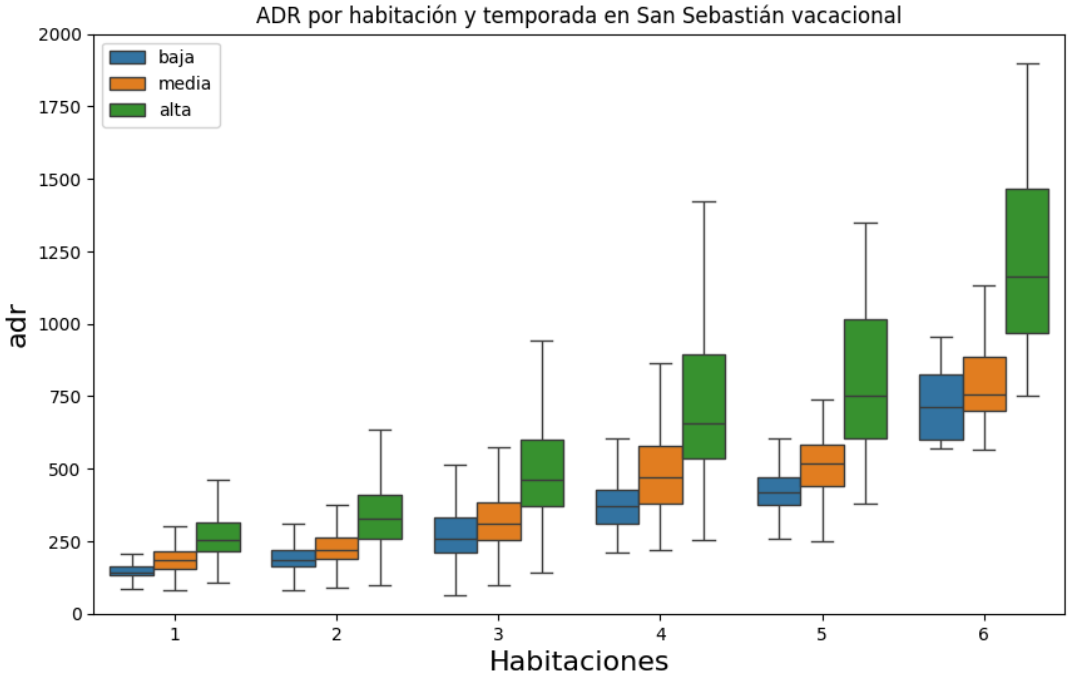
Reservas de EEUU en temporada alta y media por edad



TEMPORADA Y ADR

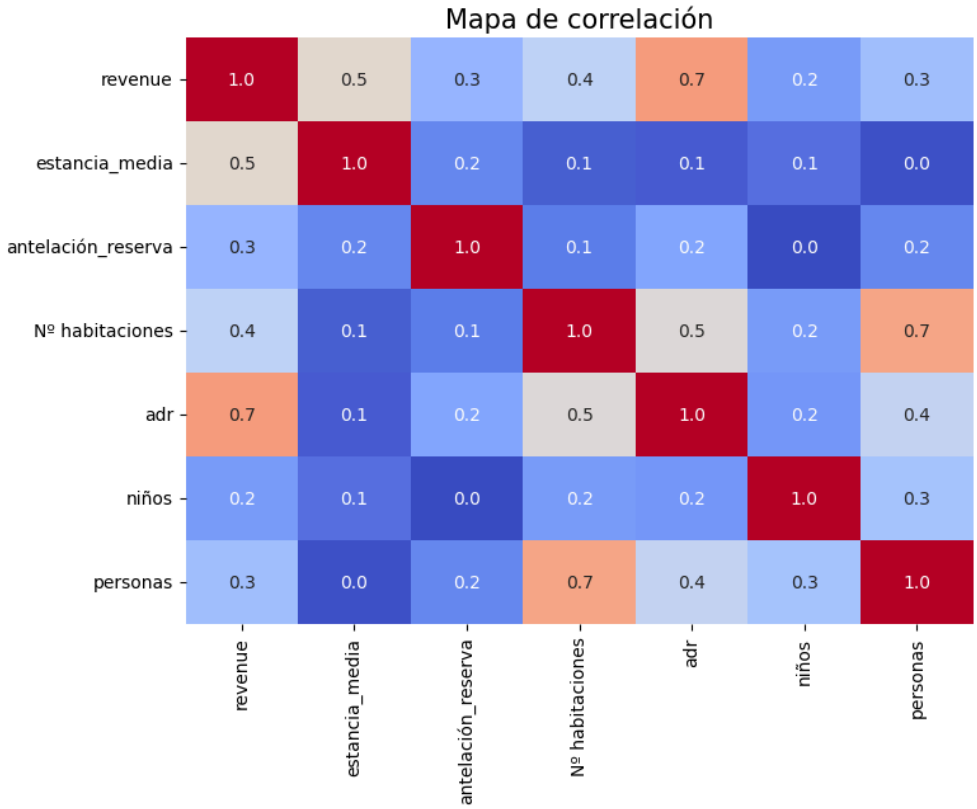
En San Sebastián, tras un análisis de la distribución de adr por tipología y por temporada, vemos que en apartamentos más grandes hay una tendencia a haber una mayor diferencia entre el precio medio de temporada baja y el precio medio de temporada alta.

En el siguiente gráfico de caja podemos ver la distribución de adr con el rango intercuartílico, la mediana, los valores extremos y los posibles valores atípicos.



RELACIÓN ENTRE VARIABLES NUMÉRICAS

Para terminar el análisis se realiza un análisis multivariable numérico para tratar de descubrir relaciones ocultas entre estas variables. Con el siguiente mapa de correlación vemos la correlación existente entre las variables numéricas de los datos estudiados. La correlación mide la relación estadística entre dos variables. Específicamente, indica hasta qué punto los cambios en una variable están asociados con los cambios en otra.



Este análisis muestra la relación evidente entre revenue, adr y estancia media. Como puede observarse el mayor porcentaje de cambio en revenue proviene del adr (0,7) y del número de habitaciones (0,7) , siendo el resto explicado por la estancia media (0,5).

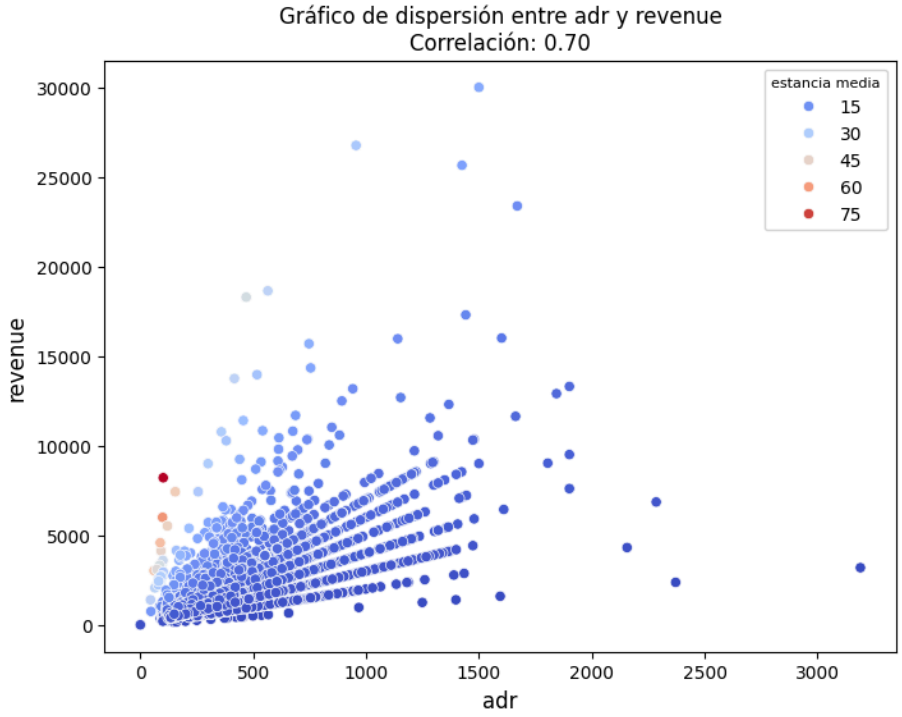
Diagrama de dispersión

Un diagrama de dispersión es una representación gráfica que muestra la relación entre dos variables cuantitativas. Los puntos en el gráfico corresponden a los valores de las variables; uno se traza en el eje X (horizontal) y el otro en el eje Y (vertical). Este tipo de diagrama es útil para identificar tendencias, patrones y posibles correlaciones entre las variables, así como para detectar valores atípicos que no se ajustan a la tendencia general.

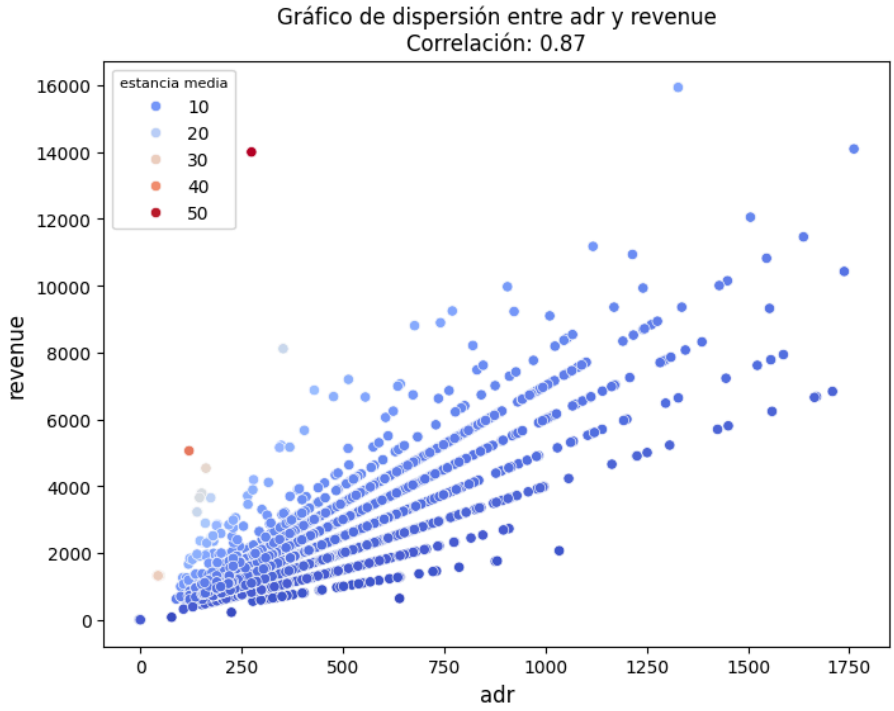
Realizo un diagrama de dispersión para medir la relación entre adr, revenue y estancia media en ambos destinos vacacionales para tratar de visualizar si la relación entre revenue y adr es igual em ambos destinos.

En los gráficos puede verse cómo a media que crece el adr tiende a crecer el revenue, siendo la aportación mayor en las reservas con estancia media superior.

San Sebastián vacacional: correlación 0,7.



Baqueira vacacional: correlación 0,87



Como pude verse la correlación entre ingresos y adr es más fuerte en Baqueira que en San Sebastián, siendo alta en ambos. Esta diferencia se explica por la mayor temporalidad en San Sebastián donde encontramos estaciones muy marcadas con diferentes niveles de precio, por otro lado en Baqueira la temporada prácticamente acaba con el invierno y el nivel de reservas en otras estancias es inexistente.

Futuras líneas de análisis

Tras este primer estudio de reservas han surgido nuevas líneas de estudio que no han podido ser cubiertas en este primer análisis por su extensión limitada.

Las siguientes áreas han sido identificadas como prioritarias para futuras investigaciones:

1. **Análisis de cancelaciones:** Estudiar las causas y patrones de cancelación en relación con variables como la antelación de la reserva, la temporada, y la demografía del cliente. Esto ayudará a identificar factores de riesgo y a desarrollar estrategias para minimizar las cancelaciones y optimizar la ocupación.
2. **Clientes repetidores:** Caracterizar el perfil de los clientes repetidores, su frecuencia de reserva y preferencias. Analizar su comportamiento de reserva en comparación con nuevos clientes para evaluar la lealtad y desarrollar programas de fidelización más efectivos.
3. **Antelación de reserva:** Investigar cómo la antelación con la que se realizan las reservas afecta la ocupación, el ADR y los ingresos. Explorar la posibilidad de crear incentivos para reservas anticipadas o desarrollar estrategias para atraer reservas de última hora sin disminuir la rentabilidad.

Estos estudios proporcionarán una visión más detallada de los comportamientos de los clientes y permitirán mejorar la gestión del inventario y la estrategia de precios. Además, estos análisis pueden revelar oportunidades para mejorar la experiencia del cliente y aumentar la fidelización.