Reporte Técnico - Análisis Estratégico de Reservaciones Hoteleras Mediante ML Clustering y BI

Profesores:

Daniel Otero Fadul — José Luis Gerónimo

Luis Antonio García — Rafael Muñoz

Marco Otilio Peña — Monserrat Pineda Rasgado

Oscar Alejandro López

Equipo:

Diego Elián Rodríguez Cantú — Ivan Karlo González Barreda

Axel Antonio Maldonado del Bosque — Elias Garza Valdés

Óscar Antonio Banderas Alvarez — Taurino López González

Monterrey, México

September 26, 2024

Contenido

Abstract

Este reporte explora la implementación de un sistema avanzado de análisis de datos para el análisis de la gestión de reservaciones en hoteles. Utilizando técnicas de clustering, en particular el modelo Gaussian Mixture, el proyecto clasifica las reservaciones en diversos segmentos para identificar patrones de comportamiento del cliente. Los resultados son integrados en dashboards interactivos de Power BI, ofreciendo una herramienta esencial para la toma de decisiones estratégicas. Este enfoque permite adaptaciones ágiles a las fluctuaciones del mercado y mejora la personalización de servicios, incrementando la rentabilidad y la satisfacción del cliente.

Palabras clave: Clustering, Gaussian Mixture, Power BI, gestión de reservaciones, análisis de datos.

Código Fuente

Este proyecto fue desarrollado principalmente en `python` y `SQL`. El código fuente se puede acceder en el siguiente repositorio:

`https://github.com/DiegoElian02/multicripto_omega`

1 Introducción

1.1 Introducción y Contexto

En la actualidad, la industria hotelera se encuentra ante un panorama competitivo y dinámico, impulsado por cambios tecnológicos y una demanda de servicios personalizados. Las expectativas de los clientes están en constante evolución, lo que obliga a los hoteles a adaptar sus estrategias para ofrecer experiencias más atractivas y eficientes. La utilización de grandes volúmenes de datos y la ciencia de datos emergen como herramientas fundamentales en este contexto, permitiendo a los hoteles no solo comprender mejor las necesidades de sus clientes, sino también prever tendencias y optimizar recursos. Según un estudio de Deloitte, la implementación de soluciones basadas en datos puede resultar en un aumento significativo de la rentabilidad para las empresas de hospitalidad, al mejorar la toma de decisiones y personalizar la experiencia del cliente [Deloitte, 2024].

En este entorno, el Tecnológico de Monterrey y TCA Software Solutions, una empresa líder en el desarrollo de aplicaciones de software con presencia en 18 países de Latinoamérica, han

formado una alianza estratégica para enfrentar los retos actuales del sector hotelero mediante la aplicación de técnicas avanzadas de análisis de datos. TCA Software Solutions, desde su fundación en 1982, ha estado al frente de la innovación tecnológica en los sectores de Comercio, Salud y Hospitalidad, y reconoce la importancia de la transformación digital como un pilar para el crecimiento y la competitividad en la industria. La colaboración con instituciones educativas como el Tecnológico de Monterrey no solo facilita el acceso a problemas industriales reales para los estudiantes, sino que también promueve la integración de conocimientos teóricos y prácticos que son esenciales en la formación de futuros profesionales [Solutions, 2023].

El proyecto que abordaremos como parte de este esfuerzo conjunto se centra en la aplicación de técnicas de clustering para segmentar el mercado de reservaciones de hotel. Esta metodología permitirá identificar patrones en las preferencias y comportamientos de los huéspedes, lo cual es crucial para personalizar servicios y maximizar la eficiencia operativa. A través de este enfoque, esperamos no solo incrementar la satisfacción y fidelización del cliente, sino también potenciar la gestión de ingresos y optimización de costos en los establecimientos asociados a TCA. Al enfrentar estos desafíos con herramientas de ciencia de datos, el proyecto no solo busca resolver un problema específico, sino también contribuir al desarrollo de competencias analíticas avanzadas entre los estudiantes participantes, preparándolos para los desafíos tecnológicos y empresariales del futuro.

1.2 Socioformador: TCA Software Solutions - Tecnológico de Monterrey

TCA Software Solutions se ha consolidado como un socio tecnológico clave para diversas industrias en Latinoamérica desde su fundación en 1982. Con presencia en 18 países, la empresa se especializa en el desarrollo de soluciones de software que abordan necesidades críticas en los sectores de Comercio, Salud y Hospitalidad. Esta experiencia extensa y diversificada coloca a TCA en una posición ideal para ofrecer desafíos reales y relevantes a los estudiantes, permitiéndoles aplicar y expandir sus conocimientos en un contexto práctico. En el ámbito de la hospitalidad, específicamente, TCA ha liderado la transformación digital mediante la implementación de sistemas avanzados que facilitan la gestión y operación hotelera, resaltando la importancia de la adaptación tecnológica en un sector tradicionalmente dependiente de la interacción humana directa.

La colaboración con el Tecnológico de Monterrey se inscribe en el marco de la educación integral que la universidad promueve, donde el aprendizaje basado en proyectos y la exposi-

ción a situaciones reales forman la columna vertebral de la formación académica. A través de esta alianza, los estudiantes no solo enfrentan desafíos que les permiten aplicar teorías y técnicas aprendidas en clases, sino que también desarrollan habilidades blandas como el trabajo en equipo, la resolución de problemas y la comunicación efectiva. Esta sinergia entre TCA y el Tecnológico de Monterrey es un testimonio del compromiso de ambas instituciones con la innovación y la excelencia educativa, proporcionando una plataforma sólida para que los estudiantes se transformen en profesionales capaces y conscientes de las demandas del mercado laboral actual.

1.3 Metas

El objetivo principal de este proyecto es aplicar la técnica de clustering para segmentar eficazmente el mercado de reservaciones de hotel, identificando patrones de comportamiento que permitan optimizar tanto la ocupación como los ingresos en diferentes temporadas. Con esto, se espera no solo mejorar la eficiencia operativa de los hoteles asociados a TCA Software Solutions, sino también incrementar la satisfacción y lealtad del cliente mediante estrategias de marketing más efectivas y personalizadas.

2 Desarrollo

2.1 Gobernanza de Datos

2.1.1 Business Case

El enfoque de este proyecto se centra en la aplicación de técnicas de clustering para segmentar el mercado de reservaciones en el sector hotelero. Esta segmentación permite identificar grupos de clientes con patrones de comportamiento similares, facilitando así la personalización de servicios y la optimización de la oferta según las necesidades y preferencias específicas de cada segmento. Este enfoque estratégico no solo mejora la experiencia del cliente, sino que también aumenta la eficiencia operativa, lo que se traduce en un incremento sustancial en la ocupación y los ingresos, especialmente durante temporadas bajas.

1. Valor Estratégico del Clustering en la Industria Hotelera: La aplicación de clustering en la industria hotelera se ha mostrado eficaz para entender mejor la dinámica del mercado. Por ejemplo, mediante el análisis de los datos de reservación, se puede determinar cuáles caracterís-

ticas de los clientes correlacionan con mayores niveles de gasto o preferencias por ciertos servicios. Esto permite a los hoteles ajustar sus estrategias de marketing y operaciones para atraer y retener a los segmentos de mercado más lucrativos. Adicionalmente, las insights derivadas del clustering pueden ayudar a prever tendencias y adaptar las ofertas en tiempo real, lo que es crucial en una industria tan susceptible a las fluctuaciones estacionales y económicas.

2. Impacto Económico: Desde una perspectiva económica, la segmentación efectiva del mercado puede llevar a un aumento en la ocupación hotelera al identificar y atraer segmentos de clientes potencialmente desatendidos o emergentes. Esto se refleja en un incremento en los ingresos directos de alojamiento y en ingresos secundarios derivados de servicios adicionales, como la gastronomía, el ocio y las conferencias. Además, la eficiencia operativa mejorada reduce costos y mejora la rentabilidad general del hotel. Según un informe de McKinsey & Company, las empresas que implementan estrategias de análisis de datos pueden ver mejoras de hasta un 10% en su rentabilidad (McKinsey & Company, 2020).

3. Beneficios Cualitativos: Más allá de los beneficios cuantitativos, la segmentación efectiva mejora la percepción del cliente respecto a la marca, ya que una oferta personalizada aumenta la satisfacción y fidelización del cliente. Esto no solo mejora la imagen del hotel, sino que también fortalece la posición competitiva en el mercado. La capacidad para anticipar las necesidades del cliente y adaptar rápidamente los servicios en consecuencia es una ventaja competitiva significativa en la industria hotelera.

En resumen, este Business Case destaca cómo el uso de técnicas avanzadas de ciencia de datos, como el clustering, puede ser transformador para la industria hotelera, proporcionando no solo beneficios económicos directos, sino también mejorando la experiencia del cliente y la imagen de marca. Este proyecto ofrece una oportunidad valiosa para que los estudiantes apliquen y expandan sus habilidades analíticas en un contexto real, lo que resultará en soluciones innovadoras que benefician tanto a la institución educativa como al socio formador.

2.1.2 Recolección de Datos

En el marco del proyecto de clusterización que estamos llevando a cabo, la gobernanza de datos se establece como un pilar fundamental para asegurar la calidad y la eficiencia de las operaciones realizadas. Este proceso comienza con la conexión segura a través de una VPN al servidor del socioformador, TCA Software Solutions. Una vez establecida la conexión, se accede a la base de datos del servidor donde se ejecuta una consulta SQL específica. Esta

consulta está diseñada para extraer los datos necesarios desde la tabla 'iar_Reservaciones', que contiene información vital para nuestro análisis.

Realizar estas tareas garantiza que manejemos los datos con la máxima seguridad y eficiencia, preparándonos adecuadamente para la fase subsiguiente de análisis de calidad de datos. Este análisis es crucial, ya que nos permite evaluar la integridad y la precisión de los datos recopilados, asegurando que la base sobre la cual se construirán los modelos de clusterización sea sólida y confiable. Este enfoque metódico no solo refleja nuestro compromiso con la excelencia en la gestión de datos, sino que también establece una fundación robusta para las fases de análisis y modelado que siguen.

2.1.3 Análisis de calidad de datos

El análisis de calidad de datos en este proyecto de clusterización se lleva a cabo con una meticulosa revisión de la base de datos `iar_Reservaciones`. Esta base contiene una amplia gama de variables que ofrecen un detallado panorama sobre las reservaciones de hoteles y sus modificaciones. La calidad de los datos se evalúa como excepcional debido a la riqueza de la información proporcionada, que incluye desde datos demográficos básicos hasta detalles específicos sobre cada reserva.

- **Relevancia de las Variables**: Cada columna de la base de datos aporta elementos críticos para el análisis. Por ejemplo, variables como `h_num_per` (número de personas por reserva), `h_num_noc` (número de noches) o `h_tot_hab` (total de habitaciones reservadas) son directamente relevantes para entender patrones de consumo y preferencias de los clientes. La existencia de campos que registran cambios en la información (`aa_h_num_per`, `aa_h_num_adu`, etc.) aporta un nivel adicional de profundidad al análisis, permitiendo seguir la evolución de las reservas a lo largo del tiempo.
- **Integridad de los Datos**: La integridad se refiere a la completitud y exactitud de los datos. En nuestro proyecto, la base de datos `iar_Reservaciones` ha demostrado tener una alta integridad. Las entradas están bien documentadas y los campos relevantes están completos, lo que facilita un análisis robusto sin la necesidad de imputaciones significativas o correcciones de datos erróneos.
- **Aplicaciones para el Clustering**: Esta calidad de datos es fundamental para los modelos de clustering, ya que permite segmentar a los clientes de manera efectiva según diversas características y comportamientos de reserva. La segmentación se basa en patrones de-

tectados en los datos, y la precisión de estos patrones depende directamente de la calidad de los datos analizados.

Con esta base sólida, el siguiente paso es aplicar técnicas estadísticas para verificar la consistencia de los datos y realizar pruebas adicionales que confirmen su validez para los análisis subsiguientes. Este proceso asegura que las conclusiones derivadas de los modelos de clustering serán confiables y de gran valor para las decisiones estratégicas en la gestión hotelera.

2.1.4 Limpieza y Preparación de Datos

El proceso de **data wrangling** en nuestro proyecto ha revelado áreas significativas que requieren atención para asegurar la calidad y la utilidad de los datos antes de proceder al análisis y modelado. Se llevó a cabo un análisis exhaustivo de las columnas de la base de datos `'iar_Reservaciones'`, la cual fue enriquecida con datos de varias tablas relacionadas como `'iar_Tipos_Habitaciones'`, `'iar_paquetes'`, `'iar_canales'`, `'iar_Agencias'`, `'iar_empresas_demo'`, y `'iar_estatus_reservaciones'`. Este enriquecimiento se realizó mediante consultas SQL que utilizan el ID de la reservación como llave primaria.

Hallazgos Clave en el Análisis de Datos:

1. **Valores Anómalos y Ceros:**
    - La columna “Capacidad_hotel” contiene exclusivamente valores cero, lo que sugiere un error en la carga de datos o un desuso de este campo.
    - Cantidades significativas de ceros en columnas críticas como “Numero_personas”, “Numero_adultos”, “Numero_men”, “Numero_noches”, y “Numero_habitaciones”. Estos ceros pueden indicar reservaciones incompletas o errores en la entrada de datos.
    - La columna “IngresoMto” también muestra un alto número de ceros, lo que podría afectar cualquier análisis relacionado con los ingresos.
2. **Datos Únicos y Valores Incorrectos:**
    - Las columnas “Hotel” y “Cupo” contienen datos únicos o irrelevantes que no contribuyen al análisis y podrían ser eliminados o reconsiderados en la estructura de la base de datos.
    - Se encontraron 10 valores “|” en la columna “Cupo”, los cuales fueron reemplazados

por ceros para mantener la consistencia numérica.

Acciones de Limpieza:

En respuesta a estos hallazgos, se ha decidido incorporar una limpieza de datos en la query final de extracción. Esta limpieza incluye:

- Eliminación o imputación de valores cero en las columnas donde estos no son lógicos, como el número de personas o habitaciones, a menos que se confirme que representan un valor legítimo (por ejemplo, cancelaciones o no-shows).
- Revisión de campos con datos únicos o valores incorrectos para determinar su relevancia o necesidad de corrección.
- Establecimiento de reglas de validación más estrictas para futuras entradas de datos para evitar la acumulación de errores y valores atípicos.

Transformación de Datos y Cálculo de Nuevas Variables

En la etapa de preparación de nuestros datos, implementamos una serie de funciones en Python para asegurar la calidad y estructura adecuada de la información antes de proceder al análisis y modelado. Estas funciones nos permiten realizar operaciones cruciales de limpieza, transformación y carga de datos (ETL), fundamentales para el manejo eficiente de grandes volúmenes de datos.

1. **Eliminación de Valores Atípicos:** Utilizamos la función remove_outliers para identificar y remover valores atípicos en columnas específicas del DataFrame. Esta función calcula la media y la desviación estándar de la columna objetivo y excluye aquellos datos que se encuentran a más de un número especificado de desviaciones estándar de la media, lo cual es crucial para evitar distorsiones en el análisis.
2. **Determinación de Temporada Alta:** La función is_high_season clasifica las fechas en 'temporada alta' o 'temporada baja' basándose en períodos predefinidos que corresponden a las temporadas de mayor demanda. Esta clasificación es vital para análisis relacionados con la planificación de capacidad y estrategias de precios en la industria hotelera.
3. **Mapeo de Tipos de Datos para SQL**: Con la función get_sql_type, transformamos los tipos de datos de pandas a los equivalentes de SQL, facilitando la creación dinámica de esquemas de tablas en bases de datos SQL, lo que permite una integración fluida de los

datos procesados en sistemas de gestión de bases de datos.

4. **Creación de Tablas en SQL:** Utilizando create_table_sql, generamos instrucciones SQL para crear nuevas tablas que reflejen el esquema de los DataFrames de pandas. Esta función es esencial para preparar la infraestructura de datos necesaria para análisis más complejos.
5. **Carga de Datos al Servidor SQL:** La función load_df_to_sql_server maneja la carga de datos desde un DataFrame de pandas hacia un servidor SQL Server. Esta función verifica la existencia de la tabla, inserta los datos y, si se requiere, sobrescribe los existentes, asegurando que los datos estén siempre actualizados y disponibles para su análisis.

2.1.5 Proteccion de Datos

En el desarrollo de nuestro proyecto, la protección de datos es una prioridad absoluta debido a la sensibilidad y la importancia de la información manejada. Se ha tomado la precaución de crear una copia de seguridad de la base de datos del socioformador para mantener la integridad y la seguridad de los datos originales. Este respaldo se ha importado a nuestro propio servidor en Microsoft Azure, configurando un entorno controlado y seguro para la manipulación y análisis de los datos.

Para garantizar un acceso seguro y controlado, los detalles críticos del servidor como el nombre del servidor, la base de datos, el usuario y la contraseña se mantienen como variables de entorno. Esto significa que dicha información esencial se almacena de manera segura y sólo es accesible a través de procesos que requieren su utilización, evitando exposiciones accidentales o accesos no autorizados. Este enfoque no solo cumple con las mejores prácticas de seguridad informática sino que también se alinea con normativas de protección de datos como GDPR y otras leyes locales dependiendo de la ubicación del servidor y la operación.

2.1.6 Trazabilidad y Linaje de Datos

En el desarrollo de nuestro proyecto, la protección de datos es una prioridad absoluta debido a la sensibilidad y la importancia de la información manejada. Se ha tomado la precaución de crear una copia de seguridad de la base de datos del socioformador para mantener la integridad y la seguridad de los datos originales. Este respaldo se ha importado a nuestro propio servidor en Microsoft Azure, configurando un entorno controlado y seguro para la manipulación y análisis de los datos.

Para garantizar un acceso seguro y controlado, los detalles críticos del servidor como el nombre del servidor, la base de datos, el usuario y la contraseña se mantienen como variables de entorno. Esto significa que dicha información esencial se almacena de manera segura y sólo es accesible a través de procesos que requieren su utilización, evitando exposiciones accidentales o accesos no autorizados. Este enfoque no solo cumple con las mejores prácticas de seguridad informática sino que también se alinea con normativas de protección de datos como GDPR y otras leyes locales dependiendo de la ubicación del servidor y la operación.

2.2 EDA

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) es un proceso que tiene como objetivo limpiar, procesar y obtener un resumen de las características más importnates de una base de datos, utilizando gráficas y diferentes estadísticas. Lo anterior permitirá comprender mejor la estructura de los datos, así como identificar patrones, valores atípicos y generar hipótesis. A continuación, se explicarán las siguientes etapas llevas a cabo en este proyecto.

2.2.1 Limpieza de datos

La limpieza de datos es la base del EDA, pues es la parte del proceso en la que los datos serán homologados y solamente se mantendrán aquellos registros que proporcionen un valor significativo al análisis que se realizará. En este caso, se ejecutaron las siguientes acciones:

- **Corrección del tipo de dato de cada columna:** Se actualizó el tipo de dato que cada columna debe almacenar para su respectivo análisis, obteniendo tres diferentes tipos de columnas (numéricas, categóricas y fechas), como se observa en la figura 1.

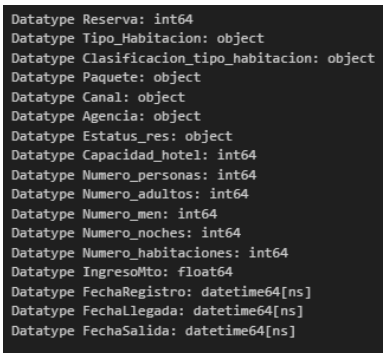

Figure 1: Tipo de dato de cada columna en la base.

- **Manejo de valores faltantes:** Se realizó la búsqueda de valores nulos en la base de datos,

realizando un conteo de los mismos. Los resultados obtenidos se muestran en la figura 2, donde es posible observar que no existían NaN.

```
Reserva                          0
Tipo_Habitacion                  0
Clasificacion_tipo_habitacion    0
Paquete                          0
Canal                            0
Agencia                          0
Estatus_res                      0
Capacidad_hotel                  0
Numero_personas                  0
Numero_adultos                   0
Numero_men                       0
Numero_noches                    0
Numero_habitaciones              0
IngresoMto                       0
FechaRegistro                    0
FechaLlegada                     0
FechaSalida                      0
dtype: int64
```

Figure 2: Conteo de valores nulos en las columnas de la base de datos.

- **Manejo de outliers:** Primeramente, para identificar valores atípicos, es importante detectar visualmente si existen algunos, por lo que se generaron las gráficas de caja para cada una de las columnas numéricas, obteniendo los siguientes resultados:

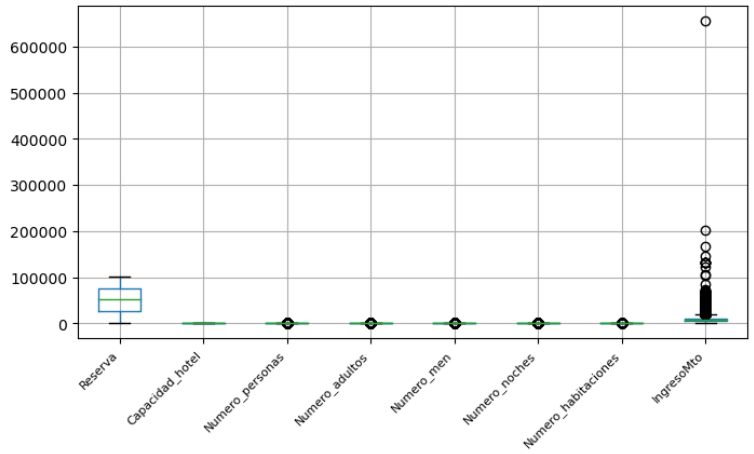

Figure 3: Vista completa del boxplot.

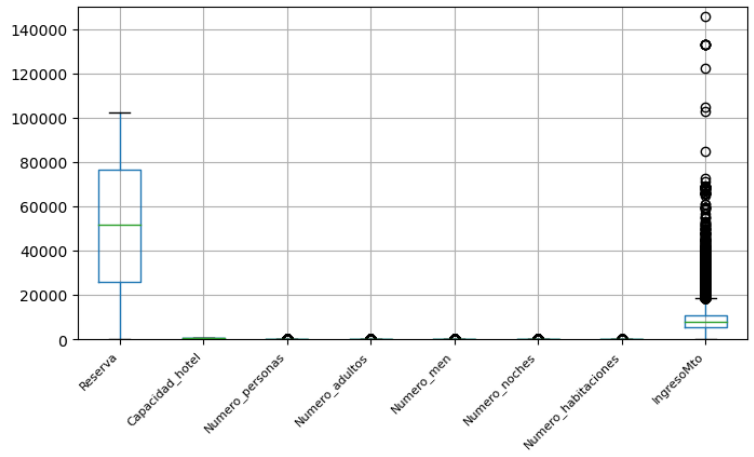

Figure 4: Vista cercana del boxplot.

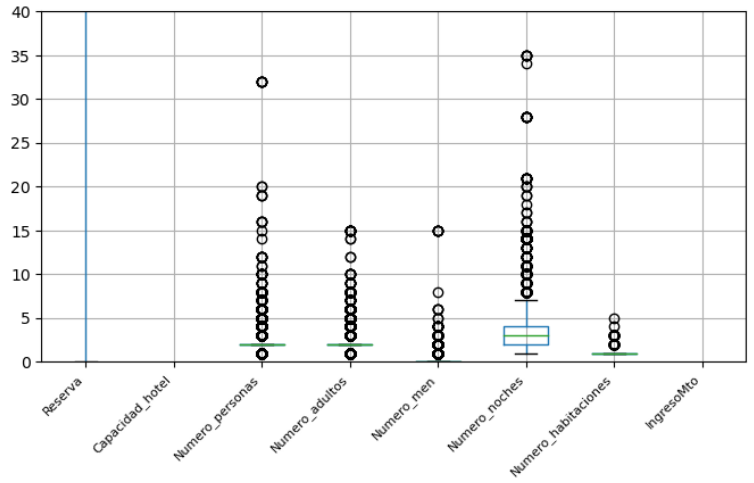

Figure 5: Zoom total del boxplot

En la figura 3 es posible observar cómo la variable que cuenta con valores más alejados de su media, es 'IngresoMto'. Después, se realizaron dos acercamientos para observar el comportamiento del resto de variables y también se encontró que el resto de variables contienen valores atípicos pero en menor volumen, como se observa en la figura 5.

Para identificar los valores atípicos y truncarlos de la base de datos se hizo uso de la métrica z-score, misma que identifica aquellos registros que se alejan de la media por más de una determinada cantidad de desviaciones estándar (en este caso, 3). Una vez identificados dichos *outliers*, se borran por completo los registros de la base de datos, pasando de tener 79986 a tener 74621, identificando un total de 5365 valores atípicos.

- **Agregando columnas de temporada alta:** Para el presente caso de estudio, se ha con-

siderado importante identificar si la fecha de llegada pertenece a una temporada de alta demanda o una temporada de baja demanda, por lo que se ha creado una columna adicional llamada 'Tipo_temporada' que almacenará un 1 si es temporada alta o un 0 si no lo es.

2.2.2 Visualización de datos

La parte de visualización tiene el objetivo de facilitar la comprensión de la distribución de los datos y las relaciones que existen entre las variables. En este caso, las gráficas que se realizaron son las siguientes:

- **Gráficas de barras:** Para comprender la distribución de las variables categóricas. Algunos de los resultados obtenidos fueron:

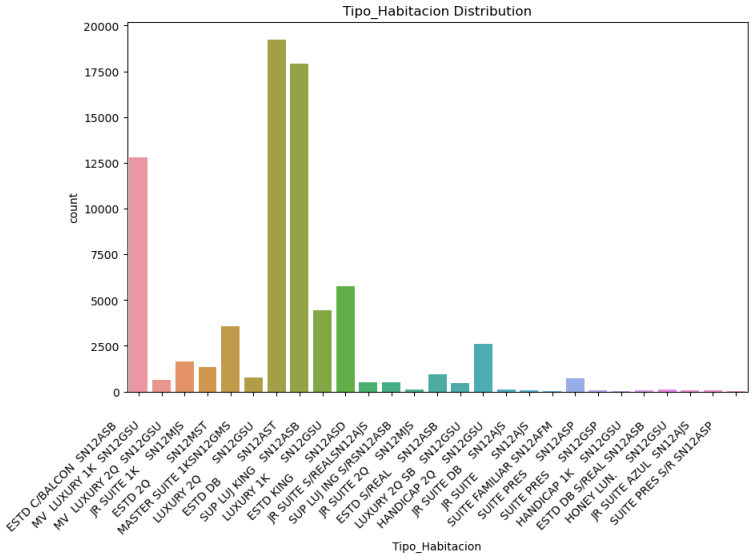

Figure 6: Tipo de habitación.

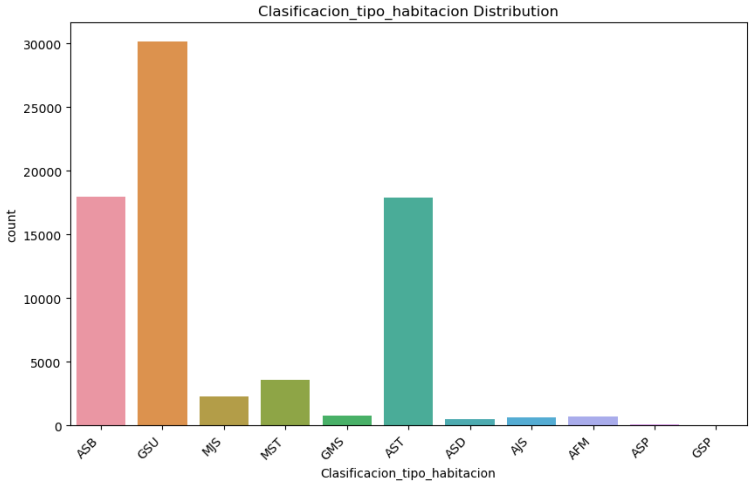

Figure 7: Clasificación de las habitaciones.

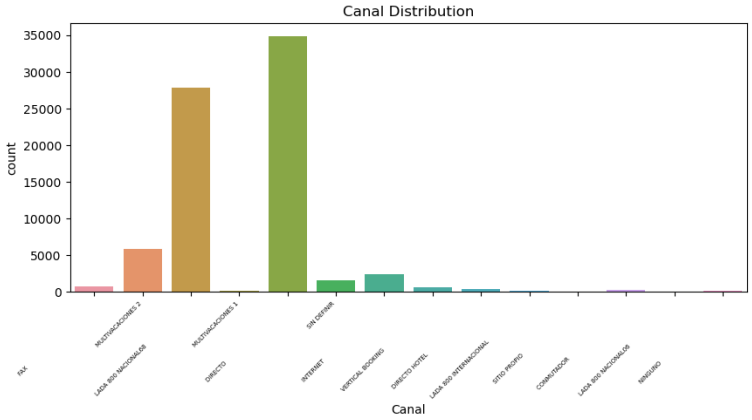

Figure 8: Canal por el cual se realizó el registro.

- **Histogramas:** Se crearon para observar la distribución de las variables numéricas. A continuación se muestran algunas de las gráficas resultantes:

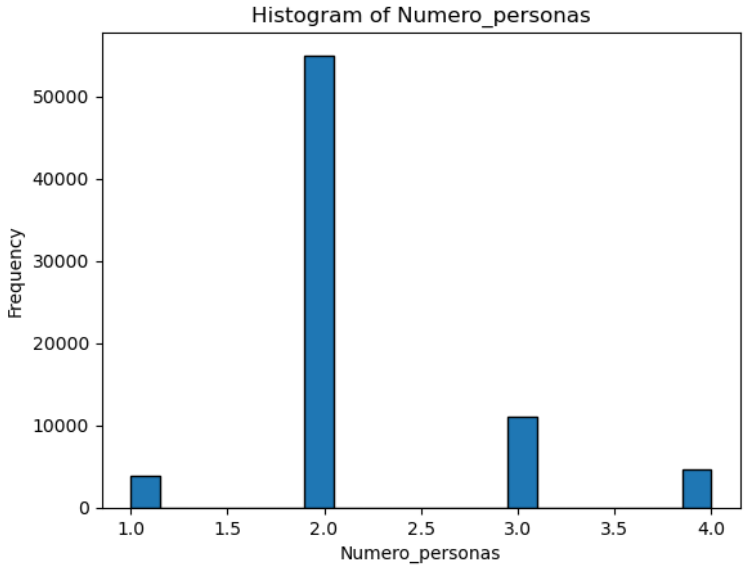

Figure 9: Frecuencia de cada cantidad de personas.

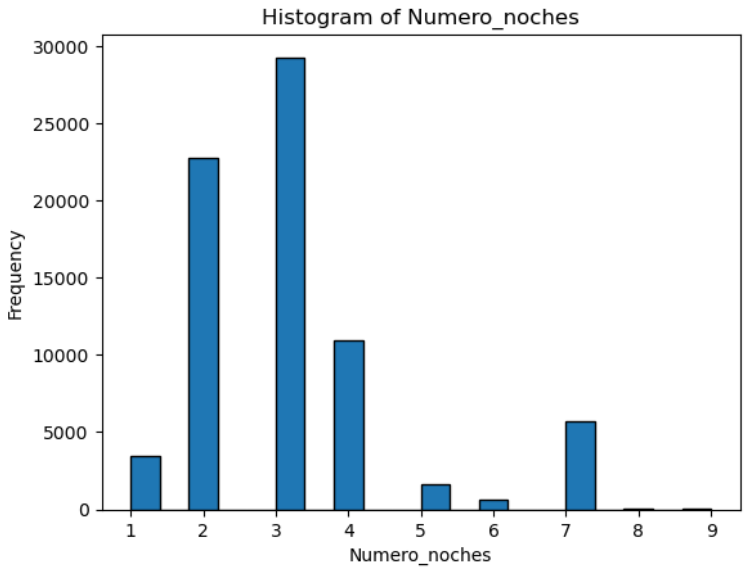

Figure 10: Frecuencia por cada cantidad de noches.

- **Time series:** Se hizo el intento de realizar una serie de tiempo de los ingresos, aunque la gráfica no es completamente visible debido al largo periodo de tiempo que se tiene. De igual manera, en la figura **??** se muestra el resultado de dicho grafo.

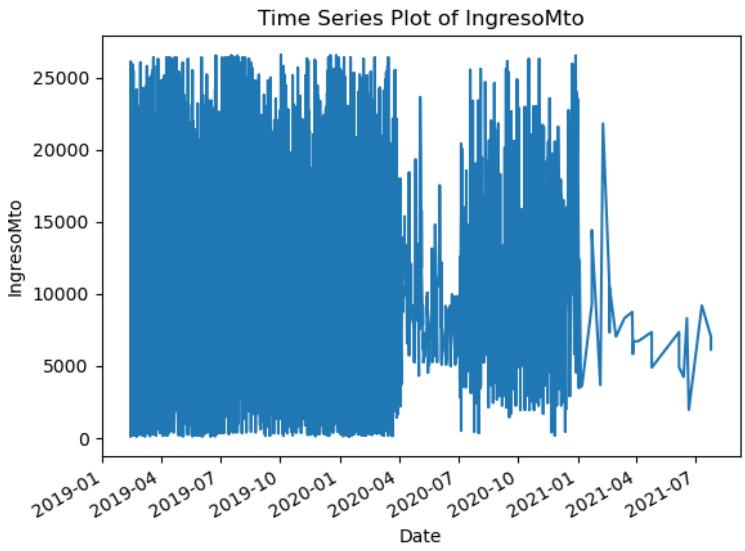

Figure 11: Serie de tiempo.

2.2.3 Estadísticas

Se obtuvieron diversas estadísticas para analizar las características de cada variable, así como las relaciones que existen entre ellas. Los procesos que se realizaron fueron:

- **Matriz de correlación:** Con el objetivo de determinar si existía algún tipo de relación entre las variables, se generó una matriz de correlaciones que nos permitiera observar dichos valores. En ella, es posible observar que existe una correlación alta entre el número de personas y el número de adultos o entre el número de noches y el monto de ingreso, donde ambas relaciones hacen sentido debido a la lógica detrás de ellas.

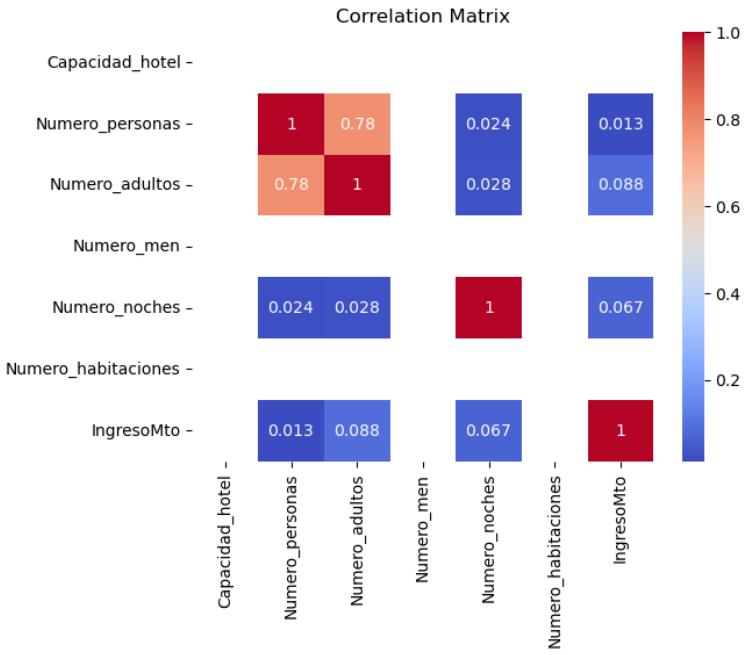

Figure 12: Matriz de correlación.

- **Resumen de datos:** Con ayuda de la función *describe* se obtuvo un resumen de las variables numéricas, mismo que presenta la desviación estándar, la media, los cuartiles, el valor mínimo y el valor máximo de cada columna. En esta tabla, es posible observar que las columnas de Capacidad_hotel, Num_habitaciones y Num_men almacenan únicamente un valor, así como podemos observar el ingreso de la venta más grande y de la más pequeña, entre otras características del modelo.

	Reserva	Capacidad_hotel	Numero_personas	Numero_adultos	Numero_men	Numero_noches	Numero_habitaciones	IngresoMto	Tipo_temporada
count	74621.00	74621.00	74621.00	74621.00	74621.00	74621.00	74621.00	74621.00	74621.00
mean	51583.13	735.00	2.22	2.10	0.00	3.13	1.00	8318.86	0.18
std	29451.38	0.00	0.63	0.50	0.00	1.43	0.00	4546.34	0.39
min	0.00	735.00	1.00	1.00	0.00	1.00	1.00	101.60	0.00
25%	26308.00	735.00	2.00	2.00	0.00	2.00	1.00	5400.00	0.00
50%	52294.00	735.00	2.00	2.00	0.00	3.00	1.00	7878.00	0.00
75%	76926.00	735.00	2.00	2.00	0.00	4.00	1.00	10644.00	0.00
max	102198.00	735.00	4.00	4.00	0.00	9.00	1.00	26572.00	1.00

Figure 13: Resumen de los datos.

2.2.4 Pruebas de hipótesis

Finalmente, se realizó una prueba de hipótesis para validar la suposición de que en temporada alta existe un mayor ingreso y un mayor movimiento de mercado que en temporada baja. Para ello, realizamos el *Two sample T-test*, que analiza si existe una diferencia significativa en el ingreso promedio de dos diferentes temporadas.

Con ayuda de la librería *stats* de python, obtuvimos el t-statistic y el p-value de dicha hipótesis, obteniendo los siguientes valores:

- **t-statistic: 41.390896652245715**. Este valor es considerado alto y representa que la diferencia entre el promedio de ingresos por temporada es muy alta, significando así que no sería posible obtener este valor en la métrica de haber escogido dos variables aleatorias.
- **p-value: 0.0**. Esta métrica indica la probabilidad de observar dicha diferencia de promedios al azar, la cual es 0, lo que significa que la hipótesis nula es rechaza y, por ende, es posible tomar a consideración la diferencia existente en el ingreso por temporadas para la toma de decisiones.

x

2.3 Modelos de Clustering

Una vez que hemos asegurado que los datos están limpios y transformados adecuadamente, avanzamos hacia la etapa crucial de nuestro pipeline: la modelización. En esta fase, se implementarán técnicas de clustering para descubrir patrones subyacentes en los datos, lo que nos permitirá identificar segmentos de clientes y comportamientos de manera más efectiva.

Para este proyecto, hemos decidido implementar cuatro modelos de clustering distintos: K-Means, DBSCAN, HDBSCAN y Gaussian Mixture Model. La elección de estos modelos está motivada por sus características únicas y su aplicabilidad en diferentes escenarios de agrupación, lo que nos brinda una oportunidad robusta para comparar resultados y determinar el enfoque más efectivo para nuestro conjunto de datos.

1. **K-Means:** Este es uno de los algoritmos de clustering más conocidos y utilizados. K-Means es eficaz para identificar grupos bien separados y es relativamente fácil de entender y aplicar. Dado que es necesario especificar el número de clusters de antemano, se emplearán métodos como el método del codo y el análisis de silueta para determinar el número óptimo de clusters.
2. **DBSCAN:** Este modelo es excelente para identificar clusters con formas irregulares y tamaños variados, y es particularmente útil porque no requiere que se especifique el número de clusters de antemano. DBSCAN minimiza el impacto de los valores atípicos, lo que es ventajoso dado que algunos ruidos persisten incluso después del proceso de limpieza de datos.

3. **HDBSCAN:** Como una extensión de DBSCAN, HDBSCAN permite una mayor flexibilidad en la densidad de clusters, lo que es ideal para nuestros datos que pueden variar significativamente en densidad. Este modelo es adecuado para datos de alta dimensionalidad como los nuestros, donde los patrones de agrupación no son uniformes
4. Gaussian Mixture Model (GMM): Este modelo ofrece un enfoque probabilístico para el clustering, asumiendo que los datos están compuestos por varias distribuciones gaussianas. GMM es particularmente útil cuando los clusters son asimétricos o más complejos en su estructura interna, proporcionando una estimación suave de la pertenencia al cluster.

Cada uno de estos modelos tiene sus fortalezas y debilidades en contextos específicos. Al aplicar todos ellos, no solo podemos abordar diferentes tipos de estructuras de datos, sino también mejorar la confianza en los resultados de clustering al comparar la coherencia entre los diferentes métodos. Esto nos permite una evaluación comprensiva y detallada de segmentaciones potenciales y sus implicaciones prácticas para la estrategia de negocio.

2.3.1 Implementacion de K-Means

Una vez que los datos están preparados y transformados adecuadamente, se procede a la aplicación de técnicas de clustering para identificar patrones y segmentos dentro del conjunto de datos. En este contexto, el modelo K-Means ha sido seleccionado por su eficiencia y efectividad en la formación de clusters claramente diferenciados.

K-Means es adecuado para los datos debido a su capacidad para manejar grandes volúmenes de forma eficiente, lo cual es crucial dada la magnitud y complejidad del dataset. Además, se favorece por su facilidad de implementación y la interpretación intuitiva de sus resultados, importantes para la comunicación efectiva con los stakeholders.

1. **Preprocesamiento:** Se utiliza ‘OneHotEncoder‘ para las variables categóricas y ‘StandardScaler‘ para las numéricas, asegurando que todas las características influyan equitativamente en el modelo sin ser sesgadas por su escala original.
2. **Determinación del Número de Clusters:**
    - Se realiza un análisis de codo para determinar el número óptimo de clusters, variando el número de clusters (k) y calculando la suma de las distancias cuadradas dentro de cada cluster.
    - ‘KneeLocator‘ se emplea para identificar el punto donde la curva de la suma de

distancias cuadradas comienza a aplanarse, lo cual indica un balance óptimo entre la compactación de los clusters y el número de estos. Para la implementación con los datos a la fecha de realización del proyecto se detectaron 4 clusters (14).

3. **Ajuste del Modelo:**

   - Con el número óptimo de clusters determinado, se ajusta el modelo K-Means a los datos transformados. Este modelo agrupa los datos intentando minimizar la varianza interna de cada cluster.
   - Se aplica el modelo para segmentar los datos y asignar cada punto a un cluster.

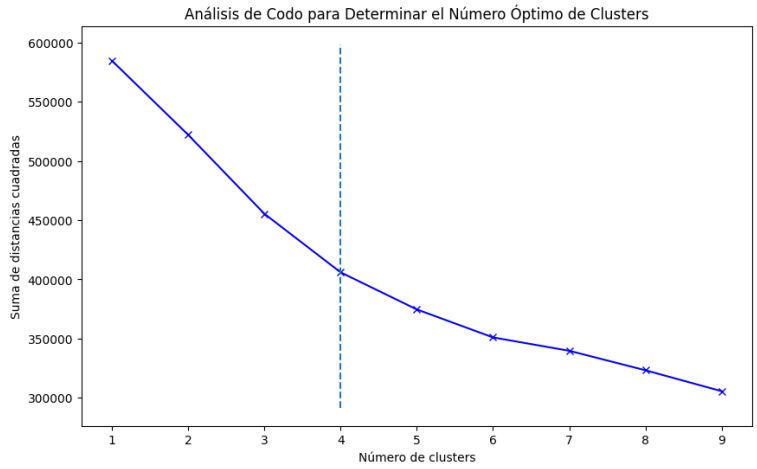

Figure 14: Resultados del Knee Locator para el K means

Visualización y Análisis de Resultados: Se visualizan los resultados del análisis de codo para confirmar la elección del número de clusters y se muestra la distribución de los tamaños de los clusters finales para evaluar la distribución de las observaciones entre los grupos identificados (14). Estas visualizaciones validan la metodología empleada y proporcionan insights sobre la estructura de los datos, esenciales para la implementación de estrategias de marketing dirigidas y la personalización de servicios. Se detectaron 4 grupos con distintas cantidades de reservaciones (15)

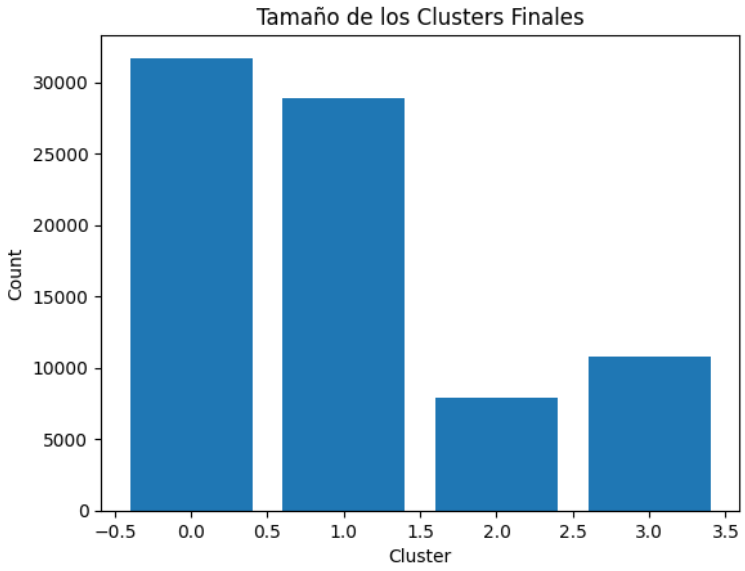

Figure 15: Cuenta por cluster del K means

2.3.2 Implementación de Gaussian-Mixture:

1. **Preprocesamiento:** Se utiliza OneHotEncoder para las variables categóricas y StandardScaler para las variables numéricas. Este paso es crucial para manejar adecuadamente las variables de diferentes escalas y facilitar la convergencia del modelo
2. **Determinación del Número de Clusters:**
   - Se analizan los Criterios de Información de Bayesiano (BIC) y Akaike (AIC) para determinar el número óptimo de componentes gaussianos. Estos criterios equilibran la complejidad del modelo contra el ajuste a los datos, evitando el sobreajuste.
   - Se emplea KneeLocator para identificar el codo en las curvas de BIC y AIC, lo que sugiere un número óptimo de componentes para el modelo. Para la implementación con los datos a la fecha de realización del proyecto se detectaron 4 clusters (16).
3. **Ajuste del Modelo:**
   - Con el número óptimo de componentes establecido, se ajusta el modelo Gaussian Mixture a los datos transformados utilizando el número de componentes identificado como el más adecuado.
   - Este paso implica la maximización de la expectativa para estimar los parámetros que mejor explican los datos observados en términos de mezclas de distribuciones gaussianas.

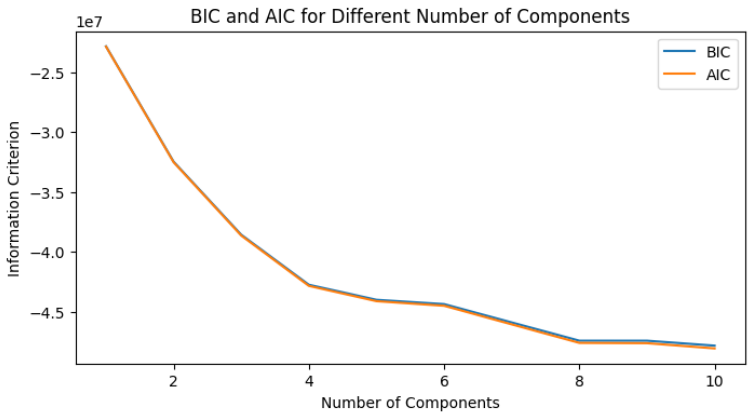

Figure 16: Resultados del Knee Locator para el Gaussian-Mixture

Visualización y Análisis de Resultados: Se visualizan los resultados del análisis de BIC y AIC para confirmar la elección del número de componentes. Además, se muestra la distribución de los tamaños de los clusters finales para entender cómo se distribuyen las observaciones entre los grupos identificados. (16). Estas visualizaciones son cruciales para validar la efectividad del modelo y proporcionan insights importantes sobre la estructura subyacente de los datos, que son esenciales para la toma de decisiones informadas en las estrategias de negocio (17).

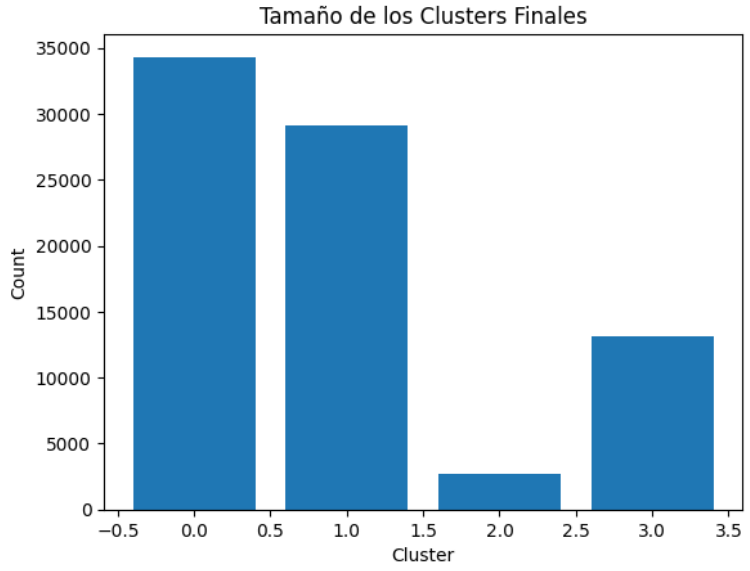

Figure 17: Cuenta por cluster del Gaussian-Mixture

Este enfoque con el Gaussian Mixture Model proporciona una perspectiva sofisticada y detallada del agrupamiento en nuestros datos, permitiendo descubrimientos analíticos que son fundamentales para intervenciones estratégicas y personalizadas basadas en patrones de datos complejos.

2.3.3 Implementación de HDBSCAN:

HDBSCAN es una elección excelente para nuestro análisis debido a su capacidad para identificar clusters basados en la densidad sin necesidad de especificar el número de clusters a priori. Este modelo es particularmente útil para datos con variaciones significativas en la densidad de los puntos, permitiendo detectar clusters que otros métodos podrían no identificar. Además, HDBSCAN gestiona eficazmente el ruido, clasificando los puntos dispersos como 'ruido' en lugar de forzarlos a pertenecer a un cluster inapropiado.

1. **Preprocesamiento:** Similar a otros modelos, se utiliza OneHotEncoder para las variables categóricas y StandardScaler para las numéricas, preparando los datos para un análisis de clustering efectivo y equitativo.
2. **Aplicación del Modelo:**
    - HDBSCAN se aplica a una muestra de los datos transformados para asegurar eficiencia dado el tamaño extenso del dataset completo. Esto permite una exploración preliminar eficaz de la estructura de los datos.
    - Se determina el tamaño mínimo de cluster, un parámetro crucial en HDBSCAN, para controlar la granularidad de la agrupación.
3. **Visualización de Clusters:**
    - Se utiliza t-SNE, una técnica de reducción de dimensionalidad, para visualizar los clusters en dos dimensiones, proporcionando una representación intuitiva de cómo los datos están organizados espacialmente
    - La visualización ayuda a interpretar los resultados del clustering y a validar la calidad de los clusters formados (18).

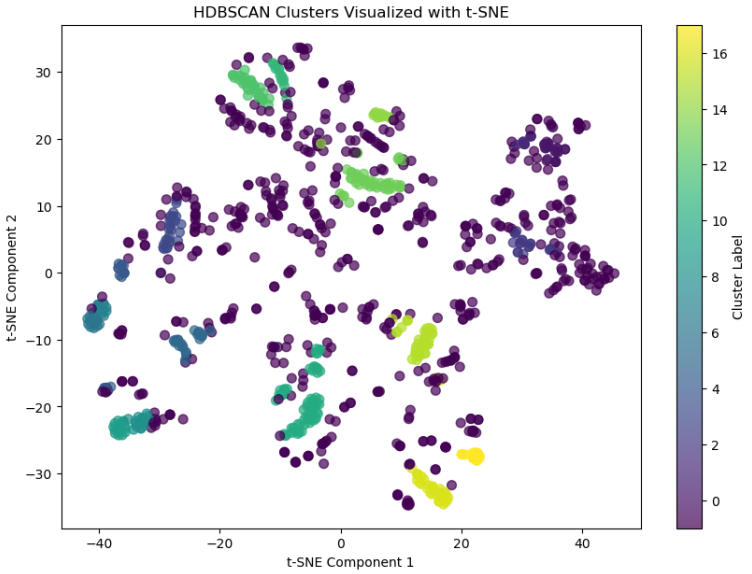

Figure 18: Resultados de los grupos generados por el HDBSCAN

Visualización y Análisis de Resultados: El número de clusters identificados por HDBSCAN es de 18, destacando la capacidad del algoritmo para discernir una variedad de agrupaciones basadas en la estructura intrínseca de los datos. La distribución de los clusters, junto con la visualización t-SNE, ofrece insights valiosos sobre las relaciones subyacentes en el dataset y cómo diferentes atributos influyen en la formación de grupos.

2.3.4 Implementación de K-means con gower:

En la exploración avanzada del clustering, se implementa el modelo K-Means utilizando la distancia de Gower para manejar eficazmente la naturaleza mixta de los datos (numéricos y categóricos). La distancia de Gower permite medir similitudes entre registros que incluyen diferentes tipos de variables, adaptándose bien a los conjuntos de datos complejos como el nuestro.

1. **Preparación de datos:**
    - Se selecciona una muestra de 20,000 registros del conjunto de datos para garantizar una operación eficiente y representativa del algoritmo sobre los datos.
    - Se calcula la matriz de distancias utilizando la distancia de Gower, que ofrece un enfoque comprensivo al considerar múltiples tipos de datos simultáneamente.
2. **Aplicación de K-means con gower:**
    - Se aplica el algoritmo K-Means sobre la matriz de distancias generada. Aquí, K-

Means opera basándose en las distancias precalculadas, agrupando los datos en cinco clusters, como determinado por análisis preliminares.

3. **Evaluación de la Cohesión del Cluster:**

   - Se utilizan métricas como el índice Silhouette y Davies-Bouldin para evaluar la calidad de los clusters formados. El índice Silhouette proporciona una medida de cuán similar es un objeto a su propio cluster comparado a otros clusters, mientras que el índice Davies-Bouldin evalúa la separación entre los clusters.

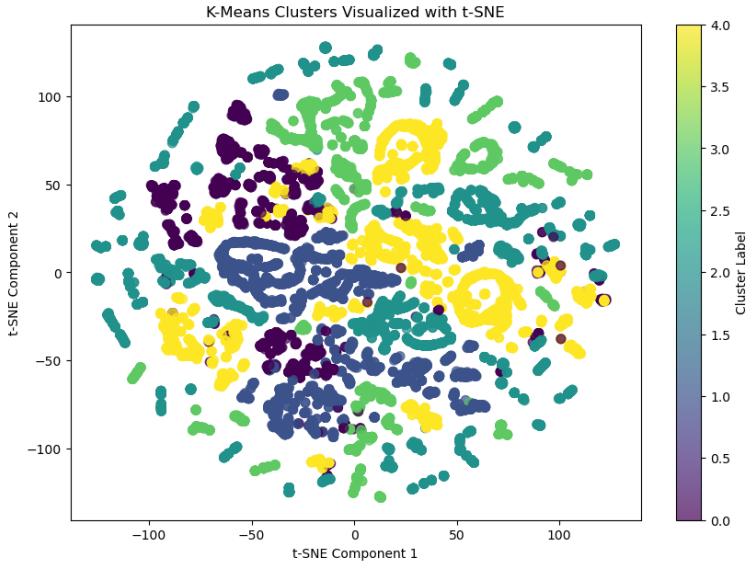

Figure 19: Resultados de los grupos generados por el K-means con gower

Visualización y Análisis de Resultados: Para una interpretación visual de cómo los clusters están distribuidos en el espacio, se emplea t-SNE (t-Distributed Stochastic Neighbor Embedding). t-SNE reduce la dimensionalidad de la matriz de distancias, permitiendo visualizar los clusters en un plano bidimensional (19). Se detectaron 5 grupos distintos, la visualización con t-SNE revela cómo los clusters están distribuidos, mostrando agrupaciones distintas que reflejan patrones subyacentes en los datos.

2.3.5 Selección de modelo

En el pipeline completo del proyecto, tras el procesamiento y la preparación de los datos, se lleva a cabo la ejecución de los cuatro modelos de clustering seleccionados: K-Means, DBSCAN, HDBSCAN y Gaussian Mixture Model. Cada uno de estos modelos se aplica bajo las mismas condiciones de entrada para garantizar una comparación equitativa y objetiva de sus

desempeños.

La selección del modelo más adecuado se basa en la evaluación de métricas clave de desempeño: el índice Silhouette y el índice Davies-Bouldin. Estas métricas son críticas ya que proporcionan una medida de la cohesión y la separación de los clusters, respectivamente. El índice Silhouette mide cuán similares son los objetos dentro de su propio cluster comparados con los de otros clusters, idealmente buscando valores más altos que indican una agrupación adecuada. Por otro lado, un menor valor de Davies-Bouldin sugiere una mejor separación entre los clusters. La combinación de ambas métricas ayuda a identificar el modelo que efectivamente logra un equilibrio entre la compactación interna de los clusters y la separación entre ellos.

Una vez identificado el modelo óptimo basado en estas métricas, se procede a aplicar este modelo sobre la totalidad de la base de datos. Este paso es crucial para asegurar que las conclusiones y estrategias subsiguientes derivadas del análisis de clustering sean representativas y robustas, reflejando la estructura completa de los datos.

2.4 Implementación técnica del servicio

2.4.1 Computo en la nube: Extracción y Carga de Datos

La extracción y carga de datos constituyen la primera etapa esencial de nuestro pipeline analítico, donde se establece la conexión segura con la base de datos del servidor para recuperar información relevante mediante una consulta SQL bien especificada. Este paso es crucial para asegurar que solo se recolecten datos precisos y pertinentes para el análisis posterior.

Conexión y Consulta SQL: Utilizando credenciales seguras almacenadas en archivos de configuración, se establece una conexión al servidor a través de ODBC, garantizando un acceso controlado y seguro a la base de datos. La consulta SQL diseñada recoge datos de múltiples tablas relacionadas a través de joins, seleccionando atributos específicos como tipo de habitación, detalles del paquete, y datos financieros. Además, se implementan filtros para descartar registros no relevantes como reservaciones canceladas o con ingresos menores a 100, asegurando la calidad y la relevancia del conjunto de datos extraído.

Filtrado y Carga: El filtrado durante la consulta ayuda a optimizar los resultados para el análisis, eliminando datos que podrían introducir sesgos. Los datos son cargados en un DataFrame de Pandas, facilitando la manipulación y transformación en pasos subsecuentes del pipeline.

2.4.2 Base de datos

El filtrado de datos y la integración efectiva de información a través de funciones agregadas y joins son componentes cruciales de nuestro pipeline de análisis de datos. Estos procesos mejoran significativamente la calidad y relevancia del conjunto de datos utilizado para el análisis posterior, facilitando la identificación de patrones y la generación de insights más precisos.

Filtrado de Datos: El filtrado de datos en nuestra consulta SQL asegura que solo los registros relevantes y de calidad sean seleccionados para el análisis. Este proceso incluye:

- **Exclusión de Reservaciones Canceladas y Anomalías:** Mediante condiciones específicas en la cláusula WHERE de la consulta SQL, eliminamos reservaciones que están marcadas como canceladas o aquellas con ingresos menores a 100, lo que nos permite concentrarnos en datos financieramente significativos.
- **Filtrado por Noches de Estancia:** Solo se consideran reservaciones con una o más noches reservadas, garantizando que el análisis se enfoque en estadías efectivas que contribuyen a la ocupación del hotel.
- **Limpieza de Datos Corruptos:** Se excluyen registros con tipos de habitación no válidos (e.g., campos con 'l'), asegurando que los datos utilizados sean precisos y coherentes.

Uso de Funciones Agregadas y Joins: Las funciones agregadas y los joins son utilizados para ensamblar un conjunto de datos completo y multifacético:

- **Joins de Tablas Relacionadas:** Se combinan datos de tablas relacionadas como iar_Tipos_Habitaciones, iar_paquetes, iar_canales, iar_Agencias, y iar_empresas, para proporcionar una visión detallada y completa de cada reservación.
- **Funciones Agregadas:**
  - **Cálculo de Temporada Alta/Baja:** La función is_high_season se utiliza para determinar si una fecha de llegada cae dentro de temporada alta o baja, basándose en periodos predefinidos de alta demanda, lo que es crucial para análisis relacionados con la planificación de capacidad y estrategias de precios.
  - **Porcentaje de Ocupación:** Se calcula el porcentaje de ocupación entre las fechas de llegada y salida para cada reservación, proporcionando una métrica directa de la utilización del hotel.

El filtrado riguroso y el uso de funciones agregadas aseguran que solo los datos más rele-

vantes y precisos sean considerados para análisis posteriores. Esto no solo aumenta la eficiencia de los procesos analíticos sino también la precisión de los insights generados, permitiendo decisiones basadas en datos sólidos y confiables.

Este enfoque meticuloso para preparar los datos subraya la importancia de una base de datos robusta y bien curada, estableciendo un fundamento sólido para análisis avanzados y la toma de decisiones estratégicas en la gestión hotelera.

2.4.3 Montado de Pipeline con Dagster

Dagster es una herramienta que permite orquestrar un flujo de scripts y tener una agenda de ejecucion. En este caso tenemos montado una carpeta dentro del repositorio en donde guardamos todo lo de esta implementación.

Dentro de esta carpeta tenemos creado un `venv` el cual tiene instalado todas las librerias necesarias para ejecutar los códigos. Una vez activado esto debemos ejecutar el comando `dagster dev` en shell para activar el deamon que mantiene activo la pipeline. Una vez hecho esto tendremos una interfaz en nuestro servidor local en el puerto 3000 (`localhost:3000`) en donde podremos ver la UI del orquestrador.

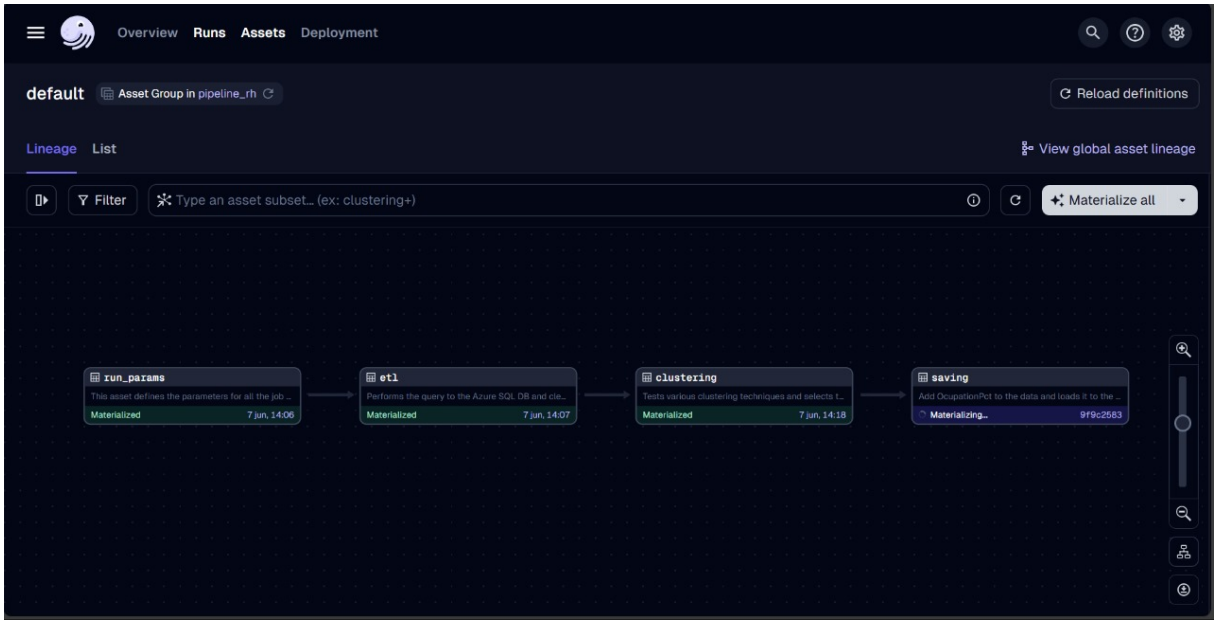

Figure 20: Pantalla con los assets creados para la pipeline.

En nuestra pipeline tenemos 4 scripts principales los cuales son:

1. `0_run_parameters`: Este recurso define los parámetros para toda la ejecución del trabajo.
2. `1_etl`: Realiza la consulta a la base de datos Azure SQL y limpia los datos.
3. 2_clustering: Prueba varias técnicas de clustering y selecciona la más adecuada.
4. `3_saving`: Añade OcupationPct a los datos y los carga en la base de datos de Azure.

Con esto configurado podemos ejecutar todos los scripts cuando queramos en orden y tener un estatus en tiempo real de como va la ejecucion.

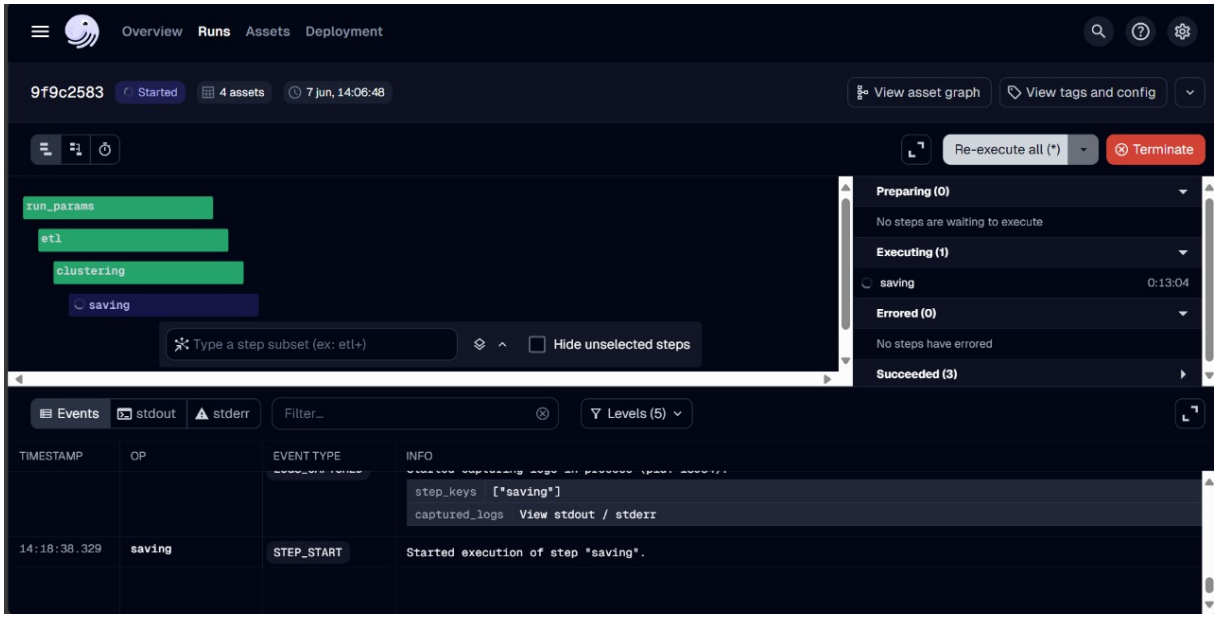

Figure 21: Pantalla con los registros de una ejecución.

Por último, también se tiene programada que esta pipeline se ejecute de manera automática cada quincena a media noche.

Figure 22: Parámetros del scheduling del job

2.4.4 Pipeline de Análisis y Procesamiento

El pipeline de análisis y procesamiento de datos de este proyecto está diseñado para gestionar y transformar grandes volúmenes de datos de reservaciones de manera eficiente y sistemática. El flujo del pipeline se estructura en varias etapas clave, cada una con un propósito específico que contribuye al análisis final y a la generación de modelos predictivos.

Inicialización y Parametrización:

El proceso comienza con la inicialización de parámetros de ejecución. En esta etapa, se generan identificadores únicos y timestamps que serán utilizados para rastrear y registrar cada sesión de análisis de manera única. Este paso es crucial para mantener la integridad de los datos y facilitar la auditoría de los procesos en fases posteriores.

Extracción y Transformación de Datos:

Posteriormente, se establece una conexión segura con la base de datos corporativa para extraer los datos necesarios. Esta conexión utiliza mecanismos de autenticación y cifrado para asegurar la confidencialidad y la integridad de la información. Una vez establecida la conexión, se ejecuta una consulta SQL predefinida que no solo recupera los datos necesarios sino que también aplica filtros y condiciones para excluir registros no deseados o irrelevantes, como reservaciones canceladas o datos incompletos.

Después de la extracción, los datos pasan por un proceso de limpieza y transformación donde se eliminan valores atípicos y se normalizan formatos. Además, se enriquecen los datos aplicando cálculos como la determinación de la temporada alta o baja basada en las fechas de las reservaciones y el cálculo de diferencias de tiempo entre eventos clave como la reserva y la llegada del cliente.

Modelado y Clustering:

Una vez que los datos están limpios y preparados, se procede a la fase de modelado. En este paso, se aplican algoritmos de clustering para segmentar el conjunto de datos en grupos basados en similitudes en varias dimensiones, como el comportamiento de reserva, las preferencias de habitación, y los patrones de gasto. Se utilizan varios modelos de clustering, incluidos K-Means y Gaussian Mixture Models, HDBSCAN y K-Means con Gower, para explorar diferentes agrupaciones y determinar la estructura óptima de los datos.

Se emplean métricas estadísticas como el índice Silhouette y el índice Davies-Bouldin para evaluar la calidad de los clusters generados por cada modelo. Estas métricas ayudan a identificar el modelo que mejor segmenta los datos, equilibrando efectivamente la cohesión interna del cluster y la separación entre clusters.

Evaluación y Selección del Modelo:

El modelo que demuestra un rendimiento superior según las métricas de evaluación es se-

leccionado para aplicarse sobre el conjunto completo de datos. Esta selección está guiada por criterios objetivos que aseguran que el modelo final no solo es estadísticamente válido sino también relevante para las necesidades del negocio.

Persistencia de Resultados y Uso Operativo:

Finalmente, los resultados del modelo seleccionado se etiquetan en el conjunto de datos y se guardan en un sistema de almacenamiento persistente. Esto incluye guardar las etiquetas de cluster junto con los datos originales para análisis futuros. Además, los resultados se integran en sistemas operativos y plataformas analíticas, como dashboards en Power BI, donde los stakeholders pueden visualizar y explorar los datos para tomar decisiones informadas.

Este pipeline no solo garantiza un procesamiento de datos eficiente y seguro sino que también permite una exploración detallada y fundamentada de los patrones y tendencias dentro de los datos de reservaciones, facilitando la toma de decisiones estratégicas basadas en datos.

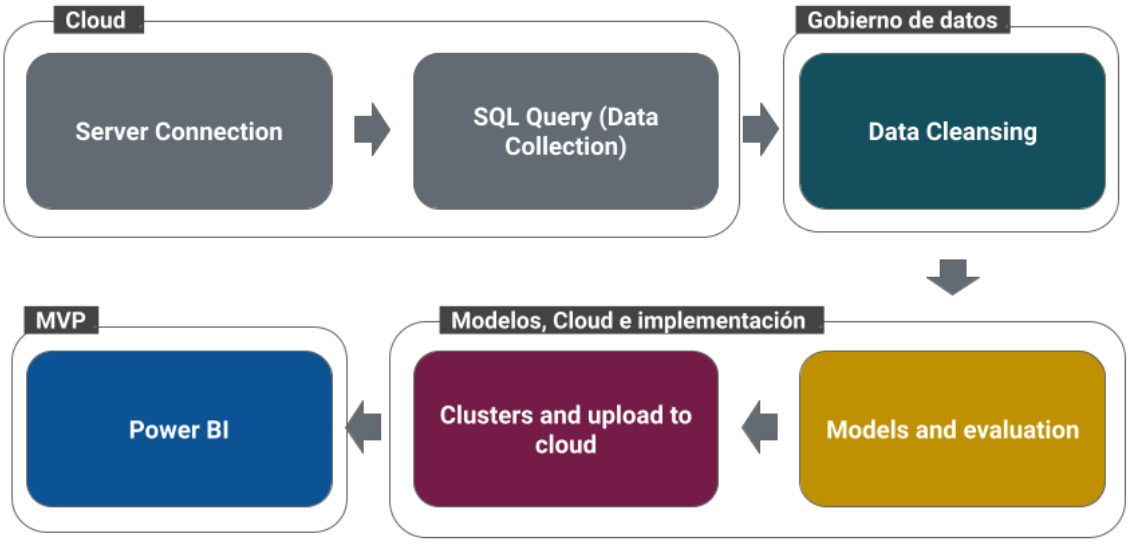

Figure 23: Diagrama de Pipeline

Integración de Resultados en el Servidor y Visualización en Power BI Una vez seleccionado y aplicado el modelo de clustering óptimo, el proceso culmina con la integración y la utilización práctica de los resultados obtenidos. Esta fase del pipeline es esencial para cerrar el ciclo de análisis y asegurar que los insights generados sean accesibles y útiles para la toma de

decisiones estratégicas dentro de la empresa.

Subida de la Base de Datos al Servidor: Tras la asignación de los clusters a cada registro de la base de datos, se genera una nueva tabla que incorpora estos resultados. Esta tabla enriquecida no solo retiene toda la información original de las reservaciones, sino que también incluye una nueva columna que indica el cluster al que cada reserva ha sido asignada según los criterios definidos por el modelo Gaussian Mixture.

Esta tabla actualizada se sube de nuevo al servidor de la empresa utilizando conexiones seguras para garantizar la integridad y la privacidad de los datos. Este paso se realiza a través de scripts automatizados que gestionan la conexión con la base de datos, la ejecución de consultas necesarias y la inserción de los nuevos datos.

Alimentación de Dashboards en Power BI: Con la base de datos actualizada ya en el servidor, se procede a utilizar esta información para alimentar dashboards en Power BI. Estos dashboards están diseñados para visualizar de manera clara y efectiva los resultados del análisis de clustering. La integración con Power BI permite a los usuarios finales, como gerentes y analistas de la empresa, interactuar con los datos a través de interfaces visuales intuitivas.

Los dashboards proporcionan múltiples vistas y segmentaciones de los datos, incluyendo comparativas por tipo de cliente, comportamientos de reserva, y patrones de consumo. Además, se pueden filtrar por cluster para analizar las características específicas de cada grupo, como el número de noches promedio, ingresos generados, y preferencias de paquetes y habitaciones. Esta capacidad de visualización ayuda a los responsables de tomar decisiones a comprender mejor las dinámicas del mercado y a ajustar las estrategias de marketing y operaciones en consecuencia.

Automatización de tablero y publicación: Teniendo el tablero creado, se procede a publicar en Power Bi Service, en un workplace creado por nosotros. La ventaja de publicarlo es para que los demás trabajadores de la organización puedan consultar el tablero desde la web. Para poder publicar se necesita mínimo una licencia de Power Bi Pro, la cual cuesta $10 dólares por usuario, y se necesita también para poder visualizar estos tableros en la web. Otra ventaja que tiene publicarlo es que puedes crear un gateway, la cual te permite dejar programado actualizaciones automáticas de los datos. En nuestro caso el gateway se conecta a la base de datos y programamos que se actualizara automáticamente todos los lunes a las 07:00 am.

3 Conclusiones

3.1 Resultados del Análisis de Clustering con Gaussian Mixture Model

La implementación del modelo Gaussian Mixture para el análisis de clustering en este proyecto ha generado una segmentación en cuatro grupos distintos que revelan patrones específicos y característicos en las reservaciones de un hotel. La interpretación de los clusters se basa en el análisis de variables como el número de noches, el número de personas, los ingresos por reservación, y la elección de paquetes y tipos de habitación. A continuación, se presenta un resumen detallado de cada cluster:

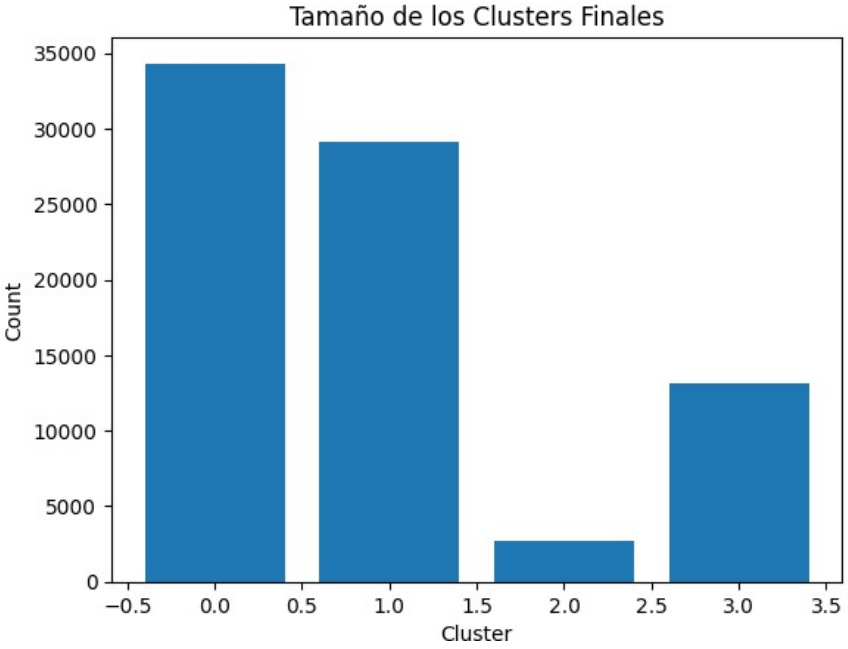

Figure 24: Tamaño de los clusters finales

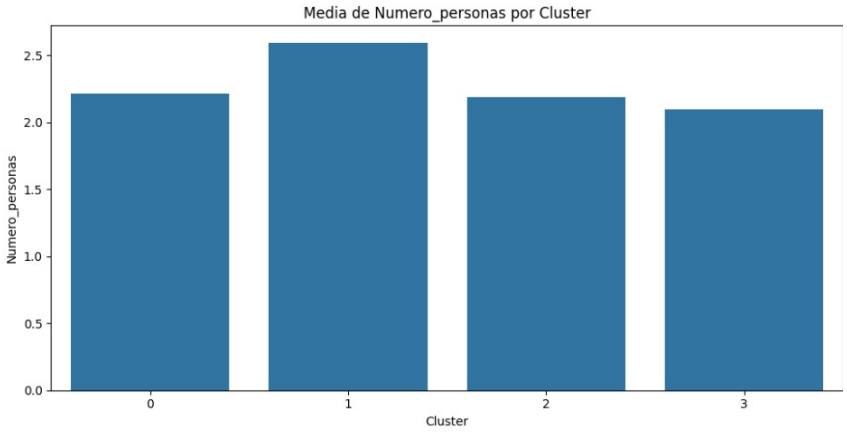

Figure 25: Media de número de personas por reservación por cluster

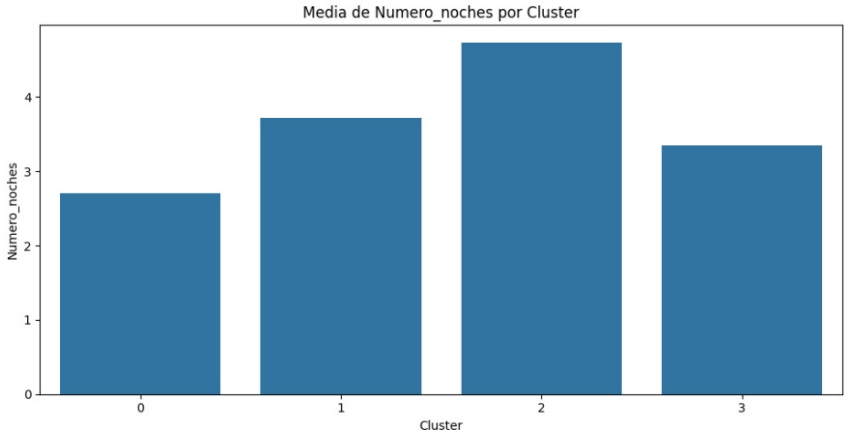

Figure 26: Media de número de noches por reservación por cluster

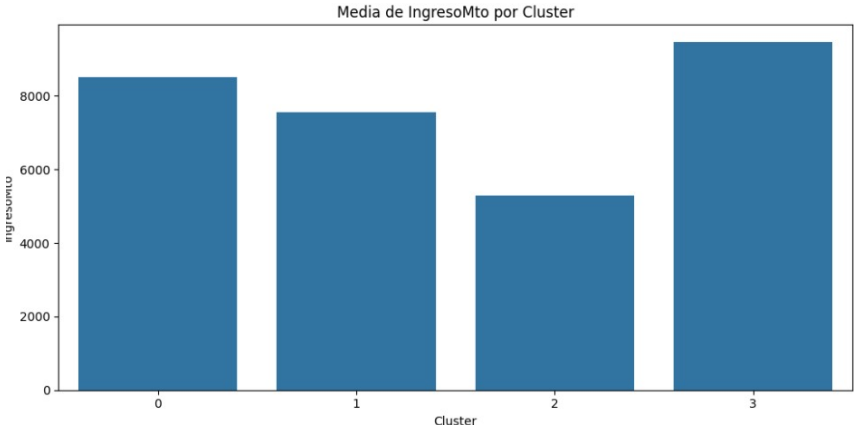

Figure 27: Media de ingreso por reservación por cluster

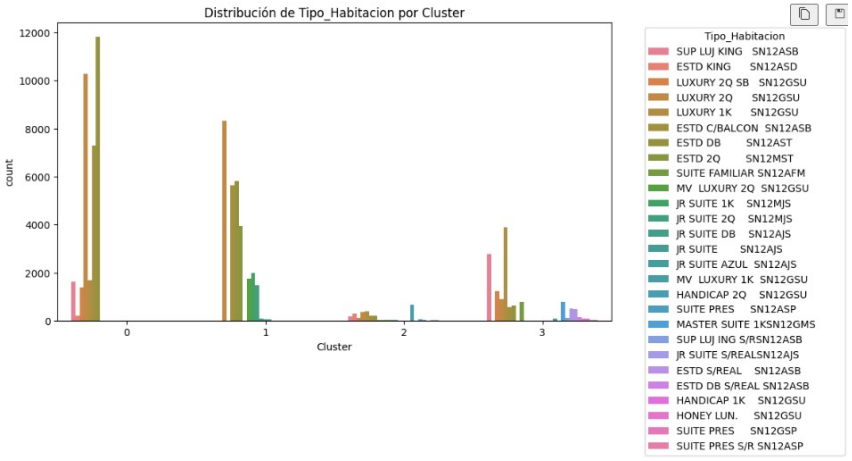

Figure 28: Distribución de habitaciones de reservación por cluster

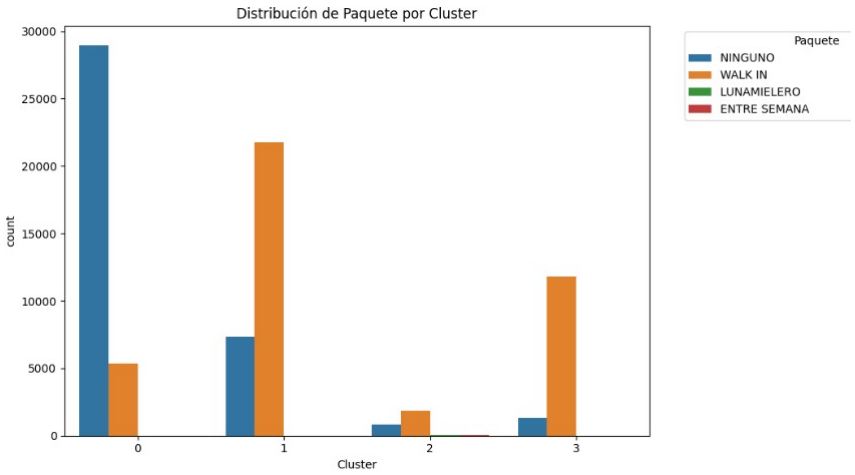

Figure 29: Distribución de paquetes de reservación por cluster

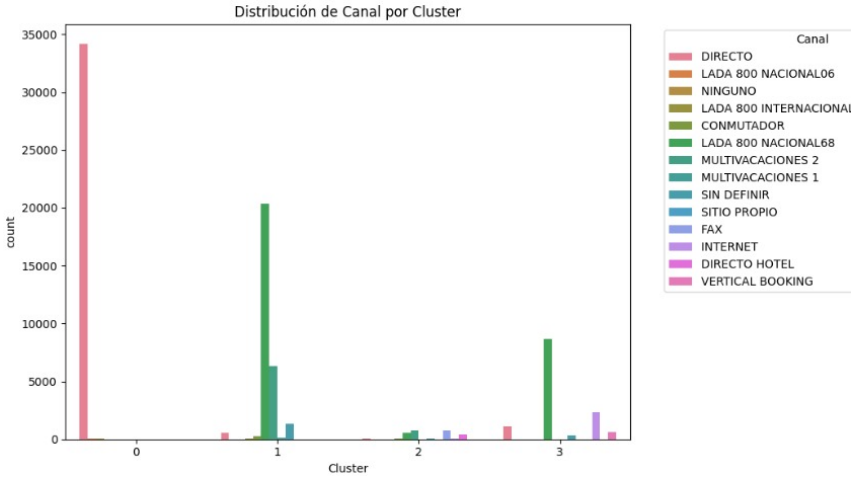

Figure 30: Distribución de canales de reservación por cluster

3.1.1 Cluster 0: Viajes Casuales

- **Características Generales**: Este es el grupo más numeroso, caracterizado por reservaciones hechas directamente, lo que podría indicar una preferencia por la conveniencia o la fidelidad a la marca.
- **Estadía y Consumo**: Tiene el menor número de noches por estadía, lo que sugiere que son viajes cortos o casuales. A pesar de esto, es el segundo en términos de ingreso medio, potencialmente debido a la selección de habitaciones de tipo 'Luxury 2Q' y 'Estándar DB' y posibles consumos adicionales.
- **Comportamiento de Compra**: Predominantemente no adquieren paquetes, lo que puede correlacionar con la naturaleza espontánea o no planificada de sus visitas.

3.1.2 Cluster 1: Vacaciones Familiares

- **Características Generales**: Este cluster muestra el mayor número de personas por reservación, lo cual es indicativo de viajes en grupo o familiares.
- **Estadía y Consumo**: Las reservaciones en este grupo tienden a ser ligeramente más largas, con una duración promedio que supera las tres noches, adecuado para vacaciones cortas de fin de semana.
- **Comportamiento de Compra**: Las reservaciones a menudo se realizan por teléfono y están asociadas con paquetes de multivacaciones, lo que sugiere una planificación previa y la búsqueda de ofertas de paquetes.

3.1.3 Cluster 2: Lunamieleros

- **Características Generales**: Este grupo tiende a incluir parejas, dado el número consistente de dos personas por reservación y una duración de estancia más larga, lo que es típico para lunas de miel o escapadas románticas.
- **Estadía y Consumo**: Seleccionan habitaciones de alto estándar como 'MV Luxury 1K' y 'Luxury 1K', lo que contribuye a mayores ingresos por reservación en comparación con otros clusters.
- **Comportamiento de Compra**: Las reservaciones se hacen frecuentemente a través de agentes de multivacaciones y fax, indicando una planificación cuidadosa y la búsqueda de experiencias exclusivas.

3.1.4 Cluster 3: Viajes de Negocios

- **Características Generales**: Este es el cluster con el ingreso medio más alto, lo que refleja la elección de habitaciones de lujo como 'Luxury 1K' y 'Sup Luj King'.
- **Estadía y Consumo**: La duración de la estancia es moderada, pero consistente, con un promedio de tres noches, lo cual es común en viajes de negocios.
- **Comportamiento de Compra**: La reserva a menudo se realiza a través de canales directos e internet, facilitando reservaciones rápidas y eficientes, típicas de los viajeros de negocios.

Interpretación General

Cada cluster revela patrones únicos en comportamiento de reserva y preferencias, lo que permite al hotel adaptar sus estrategias de marketing y operacionales para mejor atender las necesidades de cada segmento. La implementación de Gaussian Mixture Model ha permitido no solo identificar estos grupos sino también cuantificar y visualizar diferencias clave en comportamiento y preferencias, lo que es esencial para decisiones estratégicas basadas en datos.

3.1.5 Conclusión del Proyecto

Este proyecto ilustra el poder de un pipeline analítico robusto que recopila datos, los procesa, entrena modelos de machine learning y, finalmente, utiliza los resultados para influir en la toma de decisiones empresariales mediante la visualización en dashboards interactivos. El proceso comienza con la extracción de datos del sistema de reservaciones, donde la información es primero limpiada y preprocesada para asegurar su calidad. Posteriormente, se emplean técnicas de clustering avanzadas, específicamente el modelo Gaussian Mixture, para identificar patrones y segmentar las reservaciones en grupos homogéneos.

Los clusters identificados permiten entender profundamente las diversas necesidades y preferencias de los clientes, desde viajes casuales y vacaciones familiares hasta lunas de miel y viajes de negocios. Cada uno de estos grupos es analizado para ajustar las ofertas del hotel, optimizar las estrategias de marketing y mejorar la gestión operativa.

Ejecución Quincenal: La elección de ejecutar este reporte cada 15 días se fundamenta en la dinámica del negocio hotelero, donde las tendencias de reservaciones y ocupación pueden cambiar significativamente en cortos periodos debido a eventos estacionales, festivos o promociones. La actualización quincenal permite a la empresa reaccionar rápidamente a los cambios en el comportamiento del cliente y ajustar las estrategias de manera proactiva, manteniendo la competitividad y maximizando la eficiencia operativa.

Propuestas de Mejora: Para futuras iteraciones del proyecto, se podrían considerar las siguientes mejoras:

1. **Integración de Fuentes de Datos Adicionales**: Incorporar datos de redes sociales y plataformas de reseñas para enriquecer el análisis de sentimientos y preferencias de los clientes.
2. **Modelos de Predicción de Demanda**: Desarrollar modelos predictivos que no solo clasifiquen los tipos de reservaciones, sino que también pronostiquen la demanda futura

basada en tendencias históricas y factores externos.

3. **Automatización y Alertas en Tiempo Real**: Implementar sistemas de alertas automáticas que notifiquen a los gestores sobre cambios significativos en los patrones de reservación o comportamiento del cliente.

4. **Evaluación Continua de Modelos**: Establecer un sistema de revisión continua de la efectividad de los modelos, ajustándolos según la evolución del mercado y la aparición de nuevos datos.

Este pipeline no solo mejora la capacidad de respuesta de la empresa ante las dinámicas del mercado, sino que también fortalece su posición al ofrecer servicios más personalizados y eficientes, fundamentales para el éxito en la industria de la hospitalidad.

4 References

[Deloitte, 2024] Deloitte (2024). 2024 travel and hospitality industry outlook. Online.

[Solutions, 2023] Solutions, T. S. (2023). About tca software solutions. Online.