Aplicación de Machine Learning para Alojamientos Turísticos: Predicción de Cancelación de Reservas Hoteleras y de la Tarifa Media Diaria

TRABAJO DE FIN DE GRADO

GRADO EN TURISMO

FACULTAD DE COMERCIO Y TURISMO

UNIVERSIDAD COMPLUTENSE MADRID

AUTOR: Álvaro Ángel García Sánchez

TUTOR: José Carlos Soto Gómez

AÑO ACADÉMICO: 2020/2021

1. RESUMEN

En las últimas décadas el crecimiento del sector turístico ha sido notable, hasta el punto de convertirse en uno de los principales motores económicos de muchas economías de todo el mundo. Paralelamente, las nuevas tecnologías han crecido de manera aún más espectacular de forma que todas las industrias han podido beneficiarse de ellas, y la industria turística no ha sido una excepción.

En la nueva era tecnológica, dos conceptos resuenan fuertemente en la industria turística: el Big Data y la Inteligencia Artificial. Y entre ambas, surge como nexo la herramienta del Aprendizaje Automático para poder exprimir al máximo las nuevas oportunidades que se presentan de cara al futuro.

Sin embargo, en la vida real estas oportunidades parecen quedarse solo en la teoría y muchas empresas del sector aún no tienen claro cómo sacar rendimiento a los infinitos datos que están recolectando.

En este documento abordamos un proyecto que pretende acercar la aplicación del Aprendizaje Automático al sector turístico. En concreto, veremos, entre otras cosas, una aplicación sencilla de un algoritmo que nos permitirá predecir la cancelación de las reservas de hotel con la esperanza de mostrar algunas de las capacidades que nos brinda esta herramienta.

Palabras Clave: Aprendizaje Automático, Turismo, Predicción de Cancelación, Aprendizaje Supervisado, Reservas Hoteleras, Modelos Predictivos

2. ABSTRACT

In recent decades the growth of the tourism sector has been remarkable, to the point of becoming one of the main economic engines of many economies around the world. At the same time, new technologies have grown even more spectacularly so that all industries have been able to benefit from them, and the tourism industry has not been an exception.

In the new technological age, two concepts are strongly popular in the tourism industry: Big Data and Artificial Intelligence. And among both, the Machine Learning tool emerges as a nexus to be able to squeeze out the new opportunities that are presented for the future.

However, in real life these opportunities seem to remain only in the theory and many companies in this sector are still not sure how to get the most out of the massive data they are collecting.

In this document we address a project that aims to bring the application of Machine Learning closer to the tourism sector. Concretely, we will see a simple application of an algorithm that will allow us to predict the cancellation of hotel reservations in the hope of showing some capabilities that this tool offers to us.

Key Words: Machine Learning, Tourism, Cancellation Prediction, Supervised Learning, Hotel Reservations, Predictive Models.

3. INTRODUCCIÓN

Hasta comienzos del año 2020 el turismo ha sido sin duda una de las industrias que más crecimiento ha experimentado, lo que le ha posicionado como una de las industrias más importantes a nivel mundial.

Según las últimas estadísticas publicadas por la Organización Mundial del Turismo (OMT) en el año 2019 se registraron aproximadamente 1460 millones de llegadas de turistas en todo el mundo. El crecimiento se hace palpable cuando miramos 10 años atrás. En el año 2009 la OMT registró 898 millones de llegadas, lo que supone un aumento de más de 500 millones de turistas (Figura 3.1.).

Con estas mismas estadísticas a nivel europeo percibimos algo similar. En el año 2019 se registraron 744 millones de movimientos mientras en 2009 se registraron casi 300 millones menos, cerca de 480 millones (Figura 3.1.).

Por su parte, España, como ya hizo en el año 2018, volvió a liderar el ranking de países más visitados en 2019, registrando un total de casi 84 millones de turistas.

La comparación con el año 2009 también muestra un crecimiento notable, ya que se registraron 52 millones de turistas (figura 3.1.).

En los tres casos que hemos visto, tomando como referencia los últimos 10 años y los datos de la OMT, se ha experimentado un crecimiento medio anual del 5%.

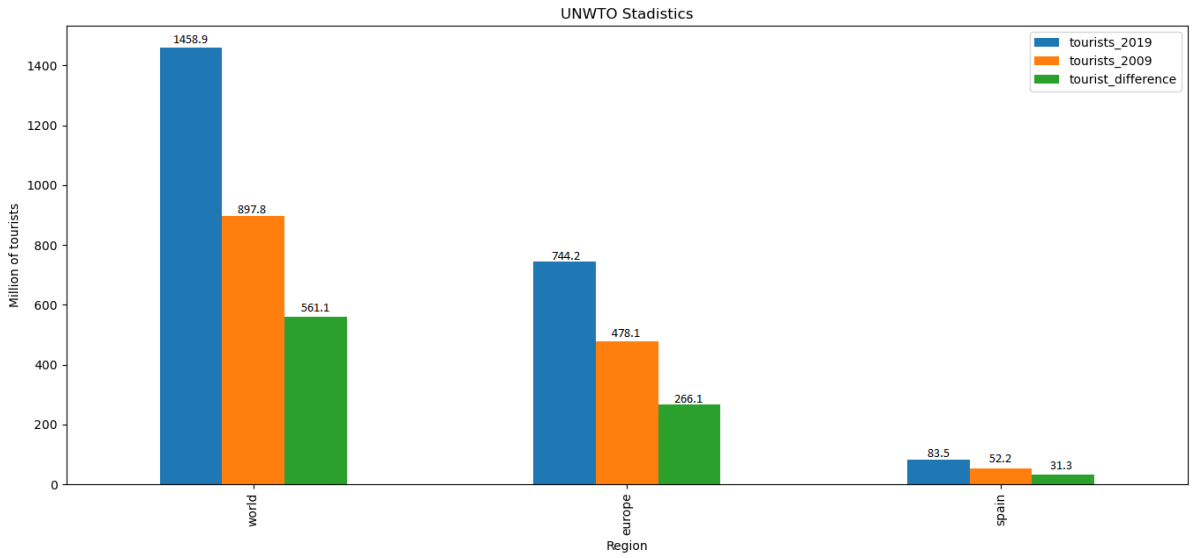


Figura 3.1. Llegada de turistas internacionales en el mundo, Europa y España en 2019

(Fuente: Elaboración propia a partir de los datos de la OMT)

A nivel económico el crecimiento también es notable. Según datos obtenidos por la investigación anual realizada por el Consejo Mundial de Viajes y Turismo (WTTC) en conjunto con Oxford Economics la industria experimentó un crecimiento del 3,5% en 2019, superando así por noveno año consecutivo el crecimiento de la economía global en ese mismo año que fue del 2,5%.

Esto supuso de manera directa e indirecta una aportación al PIB mundial del 10,3% y un total de 330 millones de puestos de trabajo alrededor del mundo (figura 3.2.).

En Europa el crecimiento fue más bajo, del 2,4%, con una aportación al PIB europeo del 9,1% (figura 3.2.).

En España, teniendo en cuenta que en el 2018 era el país que más turistas recibía, el crecimiento es aún más bajo, del 1,8%. Sin embargo, es reseñable que la

aportación del turismo al PIB del país fue más elevada, del 14,3%, así como el número de puestos de trabajo que generó, que suponía casi un 15% del total del empleo español (figura 3.2.).

Aunque lógicamente en el año 2020 las cifras serán presumiblemente diferentes debido a la pandemia del Covid-19, estos últimos datos de la OMT y la WTTC demuestran que el turismo es sin duda uno de los sectores más relevantes hoy en día en todo el mundo.

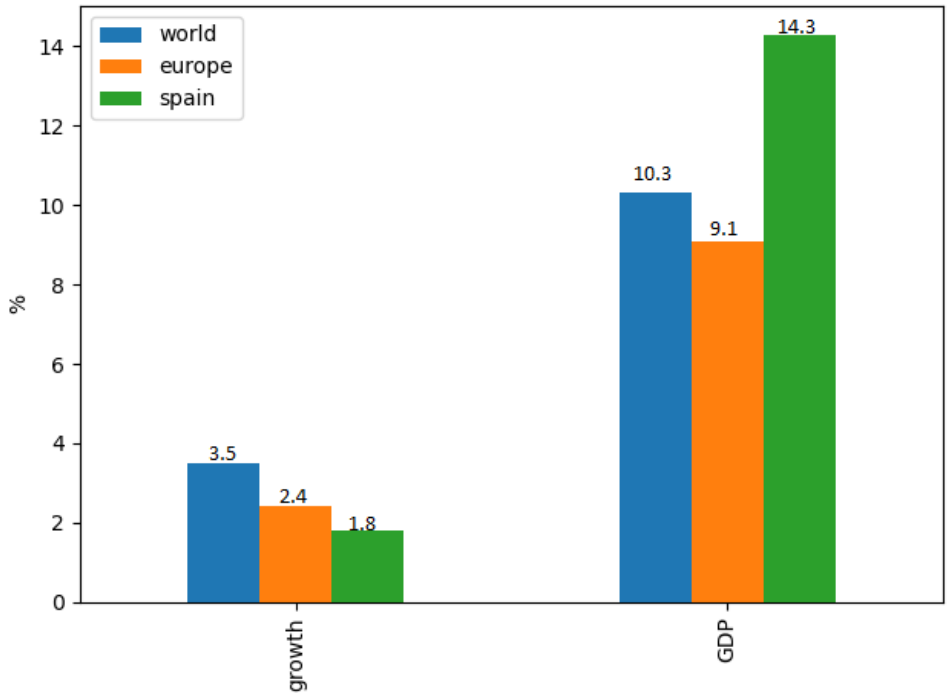


Figura 3.2. Impacto económico del Turismo en el mundo, Europa y España en 2019

(Fuente: Elaboración propia basada en los datos de WTTC)

Según el consultor estratégico de destinos García, R. (2020) una de las razones que explican este crecimiento de la industria turística es el surgimiento y el establecimiento de internet, así como la aparición de nuevas tecnologías que están llamadas, si es que no lo están haciendo ya, a transformar la industria turística, entre otros sectores de actividad.

Entre otras, García destaca los dispositivos móviles, los cuales la gran mayoría de personas hoy en día utilizan para una infinidad de tareas. Además, gracias a ellos las agencias de viajes online (OTAs) como por ejemplo Booking o Expedia han

adquirido una gran fuerza y han cambiado completamente el turismo en unos pocos años, relegando a las agencias de viajes clásicas a un segundo plano.

García también menciona el internet de las cosas, Blockchain, Big Data y la Inteligencia Artificial como las tecnologías más revolucionarias que se están avecinando en la última década.

Ante este nuevo paradigma de digitalización y nuevas tecnologías surge el Turismo 4.0, también conocido como Turismo Inteligente o Neoturismo. El término Turismo 4.0 tiene su origen en la industria 4.0 y tiene como objetivo mejorar el valor añadido del turismo a través de la innovación, el conocimiento y la tecnología (Urbančič et al, 2019).

En este nuevo escenario en el que el turismo se encuentra, Vázquez, A. (2016) enumera una serie de beneficios que las nuevas tecnologías han traído al sector turístico, entre los cuales se encuentran la agilización de procesos gracias a la automatización, la reducción de costes que ello implica y las experiencias exclusivas y la personalización del servicio al cliente.

Estos beneficios son algunos de los que concretamente el tratamiento del Big Data y la Inteligencia Artificial han traído al sector, y son estas dos tecnologías en las cuales se centrará nuestro proyecto.

La adopción del Big Data en el sector turístico ya ha sido asumido como un medio indispensable para enfrentar los desafíos del crecimiento inteligente de los destinos turísticos y de las empresas (Ardito et al, 2019). Por lo tanto sería lógico pensar que muchas de las empresas ya están aprovechando las ventajas que ofrece el uso del Big Data.

Sin embargo, al observar la implementación real de estos modelos basados en Big Data en el sector turístico, ésta suele estar más limitada a la teoría, con la excepción de unas pocas empresas que realmente lo están aplicando correctamente (Ardito et al, 2019).

De hecho, para Gretzel (et al, 2015), el concepto de Turismo Inteligente es un concepto actualmente aún mal definido y que requiere de un mayor conocimiento por parte de las empresas antes de cualquier tipo de implementación.

Llegados a este punto surgen varias preguntas que es indispensable responder: ¿Cómo aprovechar las oportunidades que nos brinda el Big Data? ¿Cómo usar el Big Data para mejorar los objetivos de negocio en Turismo?

La respuesta a estas dos preguntas es la Inteligencia Artificial, y más concretamente, el Aprendizaje Automático o Machine Learning. El Machine Learning es una potente herramienta que es capaz de extraer valor de estos datos masivos a través del análisis de datos y debido a esto es fundamental para aprender de estos datos y proporcionar información, decisiones y predicciones basadas en ellos (L'Heureux et al, 2017).

Algunas de las aplicaciones del Machine Learning en el ámbito turístico son la previsión de llegadas de turistas (Sun et al, 2018), el análisis automatizado de sentimientos de críticas o reseñas de alojamientos o destinos (Stepchenkova et al., 2017) y la predicción de las cancelaciones de reservas de hotel (Almeida et al, 2017).

Ahora que ya hemos visto la importancia del sector turístico en los últimos años y las oportunidades que se presentan gracias a las nuevas tecnologías y concretamente al Machine Learning, es momento de elegir el tema de trabajo para este proyecto.

Sin lugar a dudas, es apasionante pensar en la evolución tecnológica que hemos vivido en estas últimas décadas, y la Inteligencia Artificial es, bajo mi punto de vista, la tecnología que más expectativas está generando en todos los sectores. Por esta razón he decidido que el tema de mi trabajo de fin de grado trate de acercar las opciones que brinda el Machine Learning a la industria turística.

Dado que, como hemos visto, aún muchas empresas del sector no parecen tener claro cómo extraer todas las posibilidades que nos ofrece el Big Data, y dado que es un trabajo orientado al sector turístico, he pensado que la opción adecuada es elegir una aplicación que no requiera de una metodología que se complique en exceso.

Por este motivo, de las tres opciones que he nombrado anteriormente, que son probablemente las tres más interesantes, he optado por centrarme, entre otras cosas, en la predicción de cancelaciones de reservas de hotel, aplicación con la que

intentaremos predecir de manera exitosa si una reserva será cancelada o no, antes de que ocurra.

En los siguientes bloques del trabajo describiré los objetivos, los resultados esperados, definiré todos los conceptos que acabo de nombrar relevantes para nuestro caso, describiré la metodología utilizada e interpretaré los resultados obtenidos.

4. OBJETIVOS

El propósito principal de este proyecto es mostrar algunas de las aplicaciones del Machine Learning en el sector turístico, concretamente en el caso de los alojamientos turísticos.

Para alcanzar este objetivo nos centraremos primeramente en predecir de manera exitosa la cancelación de reservas de dos hoteles.

Se pretende que, dado un nuevo conjunto de reservas hoteleras con unas características determinadas, el algoritmo diseñado detecte automáticamente si prevé que cada esas reservas serán canceladas o no. En este paso, trataremos de realizar esta predicción para la totalidad del conjunto de datos y, de manera similar, trataremos de realizar esta predicción de manera independiente para cada tipo de hotel. De esta manera podremos observar si el tipo de hotel es determinante para predecir una cancelación, dato que nos puede resultar interesante.

De manera complementaria y para diseñar un proyecto más completo también diseñaremos un algoritmo que a partir del mismo conjunto de datos logre predecir la tasa media diaria que gasta cada una de las reservas.

Por último, con el objetivo de utilizar un modelo de aprendizaje no supervisado, vamos a realizar un análisis tanto bivariado como multivariado de cada uno de los tipos de segmento de mercado que tenemos registrados en nuestro conjunto de datos y su relación con el resto de variables.

Además de los modelos, en este proyecto también será fundamental el preprocesado de los datos: limpieza, filtrado y selección de variables. El objetivo de

esta parte, además de aumentar la eficiencia de nuestros modelos, es obtener comparaciones entre variables e información acerca de ellas que creemos que nos puede ser de gran importancia de cara a comprender el abanico de posibilidades que nos ofrece el Aprendizaje Automático.

En resumen, los objetivos del trabajo son:

- Encontrar un dataset de reservas hoteleras con un número aceptable de muestras y de características.
- Analizar y tratar los datos mediante su preprocesado.
- Aplicar el lenguaje de programación Python y algunas de sus librerías para idear los modelos para cada caso.
- Realizar el Análisis Exploratorio de los Datos preprocesados para sacar conclusiones sobre ellos.
- Realizar un análisis multivariado y bivariado entre variables mediante un modelo de aprendizaje no supervisado y diferentes tipos de gráficos.
- Analizar algunos de los algoritmos de aprendizaje supervisado potencialmente elegibles de cara a aplicar en cada uno de nuestros ejemplos.
- Elegir los modelos apropiados a utilizar en cada caso.
- Interpretar los resultados obtenidos.

Se ha elaborado un diagrama del proyecto que será desarrollado más adelante (figura 4.1.). Para ello se ha utilizado el Modelo y Notación de Procesos de Negocio (BPMN). El BPMN es una notación gráfica estandarizada que representa los pasos en un proceso de negocio y facilita su comprensión.

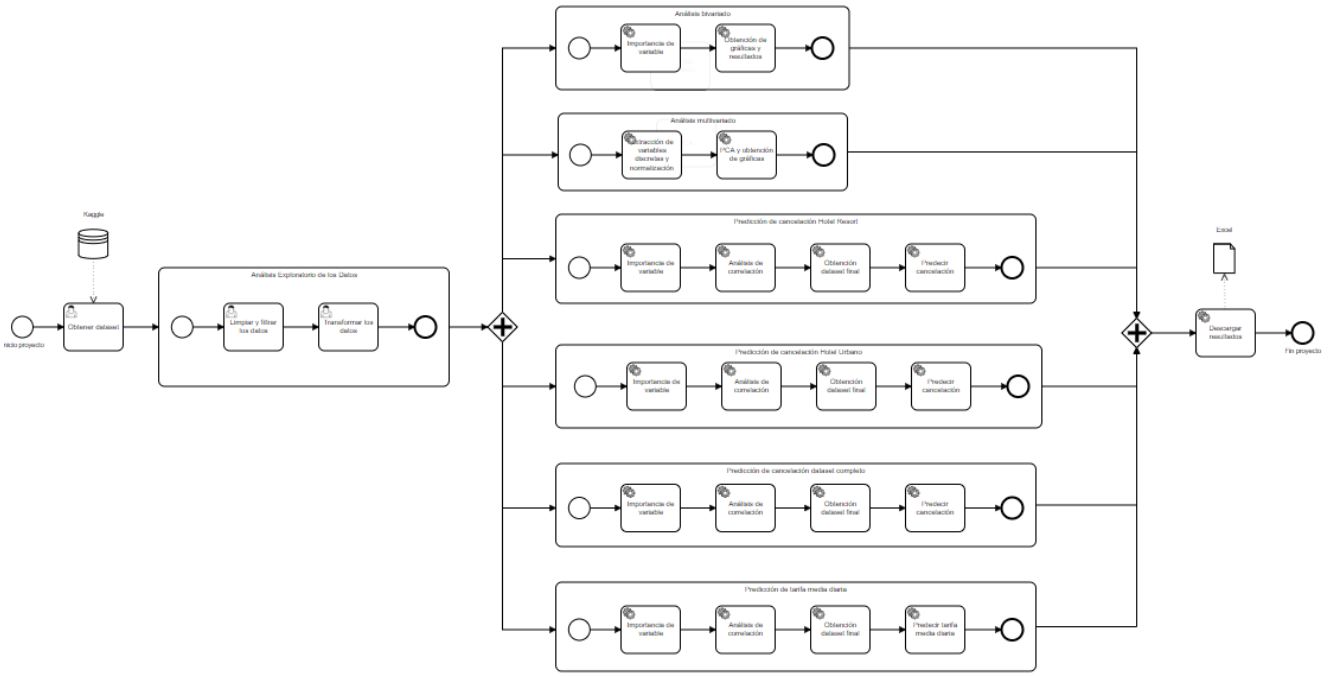


Figura 4.1. Diagrama BPMN del proyecto

5. RESULTADOS ESPERADOS

Tras la finalización de este proyecto se esperan obtener los siguientes resultados:

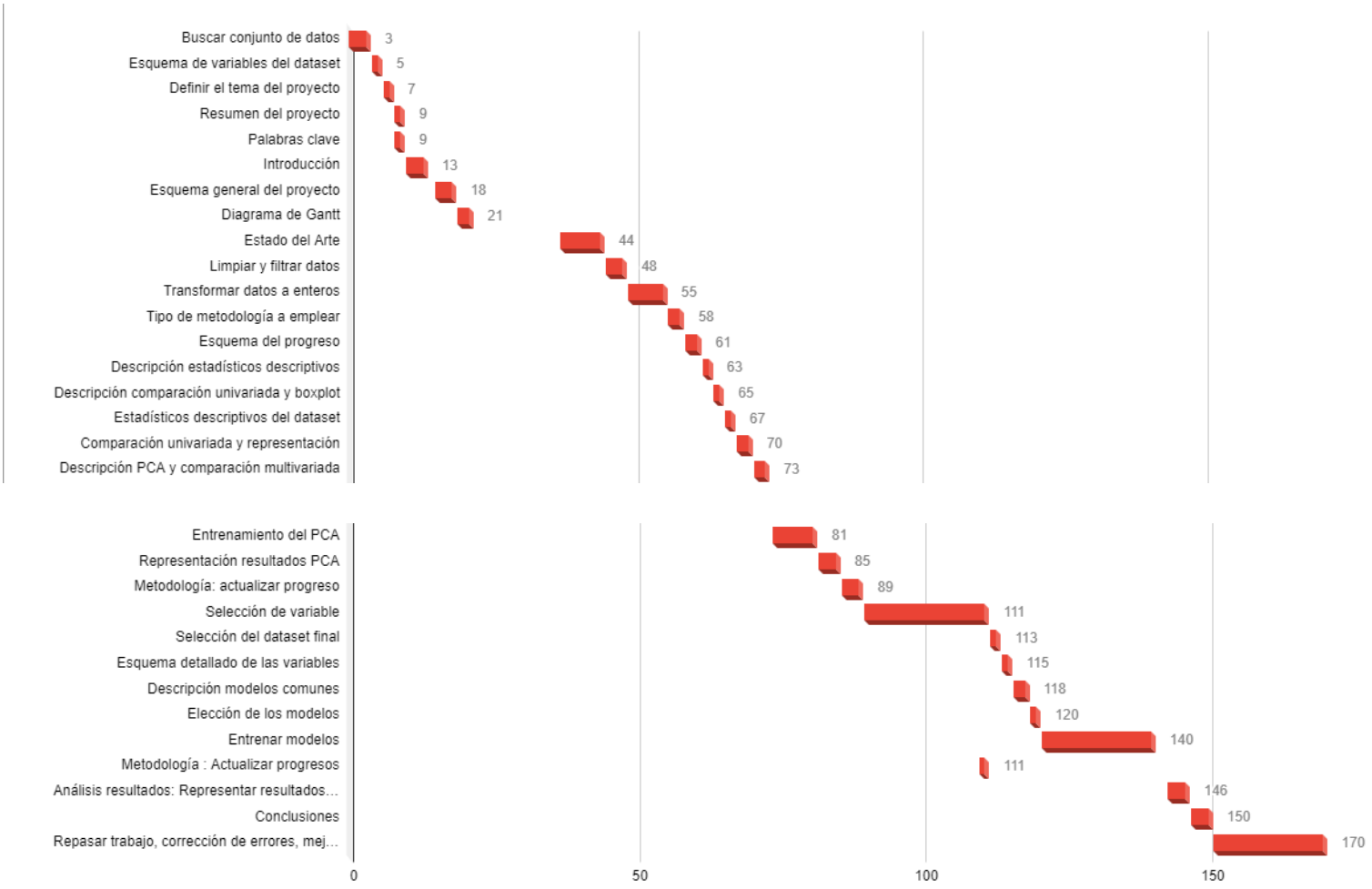
- Comprender el conjunto de datos encontrado.
- Obtener un esquema de los datos preprocesados.
- Encontrar las relaciones entre las variables de los datos preprocesados.
- Aplicar modelos de Aprendizaje Supervisado de clasificación y de regresión y de aprendizaje no supervisado.
- Lograr una efectividad aceptable en las predicciones. Esto es obtener un porcentaje superior al 85% en la cancelación de reservas y un error inferior al 15% en la tasa media diaria. Estos valores son los mínimos exigibles en un entorno laboral y por tanto el objetivo principal es alcanzar este mínimo.
- Visualizar los datos obtenidos.
- Un análisis de los resultados obtenidos.

Se espera por tanto conseguir un resultado positivo que demuestre la utilidad de aplicar esta tecnología al sector turístico.

6. PLANIFICACIÓN

La planificación del proyecto se ha representado con un diagrama de Gantt que refleja la estimación del tiempo que será necesario emplear para cada una de las tareas.

En concreto, el diagrama representa un total de 29 tareas que se realizarán en un período de tiempo comprendido entre el 1 de diciembre de 2020 y el 20 de mayo de 2021. Por lo tanto, se estima que el desarrollo del proyecto tendrá una duración algo superior a cinco meses (Figuras 6.1. y 6.2.).



Figuras 6.1 y 6.2. Diagrama de Gantt

7. MARCO TEÓRICO

7.1. Concepto de turismo

Como hemos visto en puntos anteriores el turismo es un sector de vital importancia en la sociedad actual, tanto a nivel social como a nivel económico. Para entenderlo un poco mejor conviene definirlo.

Al ser un término amplio, definirlo puede ser algo complejo y la prueba de ello es que los distintos organismos nacionales e internacionales lo definen de maneras muy diferentes.

Si consultamos un diccionario en línea oficial de lengua española el turismo es definido como la actividad o el hecho de viajar por placer, así como el conjunto de los medios conducentes a facilitar los viajes turísticos y el conjunto de personas que realizan dichos viajes (Real Academia Española, d.f., definiciones 1, 2 y 3).

No obstante, esta definición parece demasiado simple, escueta y anticuada; por lo que es interesante complementar con las definiciones oficiales ofrecidas por otros organismos.

El diccionario en línea francés El Tesoro de la Lengua Francesa define el término como la actividad de una persona que viaja por placer, visita una región, un país, un continente diferente al suyo, para satisfacer su curiosidad, su gusto por la aventura y el descubrimiento, su deseo de enriquecer su experiencia y su cultura. (Trésor de la Langue Française informatisé, s.f., definición 1). Las siguientes acepciones también relacionan al turismo con el conjunto de actividades turísticas, la industria hotelera y los medios de transporte.

Esta definición parece mucho más completa que la anterior, pero para acabar de comprender el concepto de turismo lo mejor es estudiar la definición que nos brinda el organismo mundial más importante del turismo, que lo define como el fenómeno social, cultural y económico que supone el desplazamiento de personas a países o lugares fuera de su entorno habitual por motivos personales, profesionales o de negocios. (Organización Mundial del Turismo, s.f.). La OMT denomina a las personas que viajan como viajeros, que pueden ser turistas o excursionistas y

residentes o no residentes. También aclaran que el turismo abarca sus actividades, algunas de las cuales suponen un gasto turístico.

Tras estas definiciones podemos comprender un poco mejor qué es el turismo, y podemos observar que está inexorablemente relacionado con las personas y con los viajes, algo que es conocido popularmente, pero además observamos que el término también está relacionado con el impacto económico, social y cultural que estos desplazamientos y estas personas suponen para el destino, así como el conjunto de medios y actividades que participan en el proceso.

7.2. Orígenes y evolución del turismo

El término “turismo” que acabamos de definir está formado por el sufijo -ismo, que define una acción o proceso, y por la raíz “tur”, que proviene del francés “tour” y actualmente se define como un movimiento o desplazamiento aproximadamente circular donde se vuelve al punto de partida. (Trésor de la Langue Française informatisé, s.f., definición 3.B.1.).

Esta acepción del término “tour” procede a su vez de los vocablos del latín “tornare” y “tornus” (Quesada, R., 2007), que tenían una definición similar y que implican siempre una vuelta o regreso al lugar de partida.

Según Leiper N. (1983) antes del siglo XVI no existía este término, sino que se usaban únicamente palabras como “viaje” o “peregrinaje”. De hecho, el término “turismo” no fue incluído por el diccionario inglés Oxford hasta el año 1811. Por su parte. la palabra “tour” se tomó primeramente del griego y su traducción era “torre”.

La principal teoría sobre el origen etimológico del turismo se remonta al siglo XVII, concretamente al año 1670, época en la que los jóvenes aristócratas franceses terminaban sus estudios y solían viajar por Francia, Italia o Alemania para completar sus conocimientos y obtener experiencia personal y cultural. Esto fue popularmente conocido como el Grand Tour, que duraba aproximadamente 2 años y tenía como destinos principales ciudades como Florencia, París o Nápoles.

Entre los años 1790 y 1913 se establecieron las bases del turismo tal y como lo conocemos con la Revolución Industrial, que propició el acceso a la clase media al aumentar la riqueza de la población y además consiguió estipular y regular tiempos

de trabajo y descanso, lo que permitió a la población gozar de tiempo libre, aunque este no era apenas utilizado para viajar. Sin embargo, el desarrollo tecnológico de este período, sobre todo en el transporte con el ferrocarril, facilitó los viajes, en especial los viajes comerciales.

Fue el británico Thomas Cook quien en el año 1841 vio como una oportunidad que la gente aprovechara este tiempo libre para viajar, y tras organizar un viaje para 540 personas entre dos ciudades inglesas, creó su propia empresa para organizar viajes en la que se conoce como el primer turoperador de la historia. Henry Wells y Thomas Fargo en el año 1859 y Cesar Ritz en el 1872 siguieron los pasos de Cook y emprendieron sus propios negocios de viajes.

El prometedor auge de este primer turismo fue frenado en seco en el año 1914 con el estallido de la Primera Guerra Mundial, las secuelas posteriores como la crisis de 1929, y con el comienzo de la Segunda Guerra Mundial en 1936 cuyas consecuencias fueron notables hasta el año 1949.

A partir de este momento el turismo tuvo un despegue significativo. Algunos de los factores que provocaron este crecimiento fueron la puesta en el mercado de automóviles de marca Renault y Citroën a precios económicos en 1947 y la aparición de paquetes de viajes en avión como el de Córcega en 1949 (Gordon, B.M., 2002). De esta manera, la normalización del uso del automóvil unido a la nueva estabilidad política y al crecimiento económico propició este desarrollo que se vio reforzado con los aviones Jumbo de los años 60. Por esta razón la fase comprendida entre los años 1950 y 1979 es conocida como el Boom Turístico o Turismo de masas.

En 1973, con la crisis del petróleo, la industria turística vuelve a resentirse de nuevo y para compensarlo se apuesta por aplicar una reducción de precios. Esta medida, a pesar de reducir la calidad del servicio, tiene éxito y provoca el nacimiento de nuevas infraestructuras como son los casos de Benidorm y de Cancún y se comienza a legislar exclusivamente sobre el sector turístico. La aparición de grandes multinacionales en la industria convierten al turismo en el motor económico de países como España o Italia y el crecimiento del mismo no deja de aumentar de forma exponencial hasta los años 90. En la última década del siglo XX el desarrollo

continúa pero de forma más moderada y el turismo empieza a diversificarse con el objetivo de sustituir el turismo de masas por un turismo más sostenible.

La aparición de internet en esos años, cuyo uso empieza a popularizarse y extenderse por Europa a comienzos del siglo XXI, supone el inicio de una gran revolución tecnológica que algunos autores como Schwab, K. (2016) consideran la cuarta revolución industrial. Es el comienzo de la era digital que tiene un gran impacto en todos los sectores, y el turístico no es una excepción. El aumento de la velocidad, la reducción de intermediarios y la personalización del servicio son algunos de los cambios que se observan en esta fase conocida como turismo inteligente.

La palabra “inteligente” describe los desarrollos tecnológicos, económicos y sociales impulsados por tecnologías que dependen del Big Data, del Open Data y de nuevas formas de conectividad e intercambio de información como La Tecnología de Comunicación Inalámbrica (NFC) o el Internet de las Cosas (IoT).

El turismo inteligente se aplicaría por tanto para describir la creciente dependencia de los destinos turísticos, sus industrias y sus turistas en las formas emergentes de tecnologías de la información y la comunicación que permiten transformar cantidades masivas de datos en propuestas de valor (Gretzel et al, 2015).

7.3. Turismo y tecnología

El impacto en el turismo a partir de la aparición de internet fue evidente tanto para los turistas como para los proveedores de servicios turísticos. Entre los años 2007 y 2012 las investigaciones relacionadas con el internet y el turismo habían aumentado considerablemente siendo la búsqueda de información, el análisis web y el marketing digital las tres principales áreas de investigación (Tang-Taye et al, 2012).

Otra de las consecuencias que trajo consigo la aparición de internet en la industria turística fueron las agencias de viajes online (OTAs), que son aquellas agencias que operan solamente en internet sin oficina física. La primera OTA fue creada en 1995 con el nombre de Internet Travel Network. Un año más tarde Microsoft lanzó Expedia, que en 2001 ya era líder en el mercado. Otras importantes

plataformas de viajes online que nacieron posteriormente y que también fueron líderes de mercado son Booking (1996), Tripadvisor (2000) y AirBnb (2008).

Estas plataformas online revolucionaron el mercado turístico sobre todo debido a la sencillez y velocidad de gestionar las reservas, a los precios bajos y a la posibilidad de leer y compartir valoraciones de las experiencias. Su interconexión con la tecnología móvil también fue clave de su éxito.

A pesar de todos estos avances ya conseguidos, la tecnología sigue evolucionando cada día y en la actualidad se está trabajando para encontrar la próxima tecnología llamada a revolucionar el mercado turístico.

Algunos autores como Kournavis (2012) creían que esta tecnología revolucionaria sería la Realidad Aumentada (AR). Con aplicaciones tales como tours virtuales personalizados, la AR podría ayudar significativamente a los museos, las ciudades y los profesionales del turismo en general porque la información se puede organizar y transmitir en capas o bajo petición. Además, sería especialmente útil para aquellos lugares protegidos y de aforos muy limitados, como La Alhambra, o inaccesibles, como las Cuevas de Altamira.

Por otro lado, autores como Kaur R. y Kaur K. (2016) investigaron acerca del Internet de las Cosas (IoT). IoT es un campo en el que cualquier dispositivo u objeto puede hacerse inteligente e identificable a través de etiquetas de radiofrecuencia, por esta razón tiene una relación directa con las ciudades inteligentes y, por supuesto, con el turismo inteligente. Sin embargo, para poder aplicar el Internet de las Cosas al mercado turístico sería necesario apoyarse en el Big Data y la Inteligencia Artificial, que son sin duda las dos tecnologías que forman parte del verdadero reto al que se enfrenta el turismo de hoy en día y en las que a continuación nos adentramos.

7.4. Inteligencia Artificial y Big Data

El diccionario de Oxford define la Inteligencia Artificial como la teoría y el desarrollo de sistemas informáticos capaces de realizar tareas que normalmente requieren inteligencia humana, como la percepción visual, el reconocimiento de voz, la toma de decisiones y la traducción entre idiomas.

Aunque es un tema muy actual, el término fue utilizado por primera vez en el año 1956 en la Conferencia de Darmouth y es en el año 1957 cuando se diseña la primera red neuronal artificial. A partir de este momento evolucionó hasta lograr hitos tales como chatbots, conducción autónoma o reconocimiento facial.

Para lograr muchos de los objetivos de la Inteligencia Artificial son necesarios los datos, y es ahí donde entra en juego el Big Data.

El Big Data puede definirse como una cantidad masiva de datos que se acumulan con el tiempo y que son difíciles de analizar y manejar utilizando herramientas comunes de gestión de bases de datos.

Un subconjunto de la Inteligencia Artificial muy ligado a estos datos masivos es el Aprendizaje Automático o Machine Learning.

El científico de la computación Tom Mitchell (1997) define el Machine Learning como el estudio de algoritmos informáticos que permiten que los programas informáticos mejoren automáticamente a través de la experiencia. En otras palabras, es la capacidad de las máquinas para recibir un conjunto de datos y aprender por sí mismos, adaptando los algoritmos a medida que aprenden más sobre la información que procesan.

Como habíamos mencionado anteriormente, el Machine Learning es, por tanto, la herramienta que tiene la capacidad de sacar provecho del Big Data. La industria turística solo es una más de las que puede beneficiarse de sus aplicaciones.

Un ejemplo de aplicación general del Machine Learning es la predicción del precio de una casa dado un set de datos con información de diversas características de un número amplio de casas y el precio al que actualmente se venden en el mercado. Un algoritmo de Machine Learning sería capaz de predecir el precio ideal al que se podría vender una casa dadas sus características.

Otro ejemplo más simple es la detección y clasificación de correos electrónicos no deseados o spam, siendo capaz de diferenciarlos de los correos comunes.

Los modelos de Machine Learning pueden clasificarse en Métodos de aprendizaje supervisado y métodos de aprendizaje no supervisado.

El aprendizaje supervisado trabaja con estructuras de datos conocidas a priori para entrenar patrones y reglas para predecir nuevos datos, por lo que el algoritmo

se utiliza para aprender la función desde las variables de entrada (X) a las de salida (Y). Los métodos supervisados se agrupan en modelos de regresión y de clasificación. En las tareas de clasificación, la salida es una variable categórica mientras que en los problemas de regresión la salida es una variable continua.

Por el contrario, el aprendizaje no supervisado establece similitudes y diferencias desconocidas en los datos de entrada para modelar la estructura o distribución de los datos con el fin de aprender más sobre ellos.

En el siguiente apartado veremos algunas de las aplicaciones del Machine Learning que se están realizando actualmente en el sector turístico.

8. ESTADO DEL ARTE

Actualmente la mayor parte de las aplicaciones de Machine Learning están orientadas a temas médicos como detección de cáncer; al reconocimiento de imágenes y reconocimiento facial y al procesamiento del lenguaje natural.

No obstante, debido a la importancia que ha tomado el sector turístico en la economía global, en los últimos años han aumentado las investigaciones acerca de cómo aplicar el Aprendizaje Automático en esta industria.

Las aplicaciones actuales más populares pueden dividirse principalmente en cuatro áreas que describimos a continuación:

- Pronóstico de la demanda turística.

El análisis de la demanda turística y lidiar con las fluctuaciones estacionales de los datos turísticos, aunque complejo, siempre ha sido un tema importante para prever con antelación los movimientos de mercado. Tras el año 2020 y la pandemia de covid-19, este tema ha cobrado si cabe aún más importancia debido a los inesperados impactos que ha tenido en el sector y en la propia población a nivel mundial.

Algunos investigadores como Cankurt y Subasi (2015), Xie et al. (2021) Claveria et al (2015) han abordado este tema de manera exitosa tomando la estacionalidad como un proceso estadístico estocástico y aplicando redes neuronales artificiales (ANN) y Soportes de Vectores Regresión (SVR).

- Sistemas de recomendación de destinos turísticos.

Los sistemas de recomendación surgieron en el dominio del comercio electrónico y fueron desarrollados para recomendar activamente los elementos idóneos a los usuarios en internet. Popularizados por plataformas como Facebook, Youtube, Netflix o Amazon, también puede aplicarse para recomendar a turistas potenciales productos turísticos que puedan ser de su interés.

En este área los autores utilizan calificaciones, valoraciones y otras retroalimentaciones de los usuarios para entrenar el sistema de recomendación (Nilashi et al, 2017) o bien utilizan su información demográfica en caso de carecer de otros datos adicionales (Wang et al, 2012).

- Análisis de sentimiento de reseñas hoteleras.

En relación con el punto anterior, el análisis de sentimiento de reseñas hoteleras puede ayudar a entrenar a los sistemas de recomendación y además permite a los alojamientos turísticos recibir retroalimentación de sus clientes de manera directa y automática.

La aparición de diversas plataformas turísticas en internet ha dado lugar a una enorme cantidad de reseñas personales con información relacionada con viajes en la web, información muy valiosa tanto para los viajeros como para los proveedores de servicios turísticos. Sin embargo, esta información es abrumadora e inabarcable para el ojo humano, por lo que, mientras los buscadores pueden ayudar a los viajeros a obtener la información que necesitan, los proveedores necesitan de la ayuda de la inteligencia artificial para manejar toda esta información e interpretar los resultados.

Los primeros autores emplearon algoritmos como las Máquinas de Vector Soporte (SVM) o el algoritmo Naïve Bayes para analizar reseñas extraídas de blogs de viaje (Ye et al, 2009).

Sin embargo en la actualidad muchos autores optan primero por extraer las reseñas de plataformas web como TripAdvisor o Booking.com. Esto se consigue gracias a las técnicas de Web Crawling y Web Scraping. Posteriormente se analiza el contenido de los textos recolectados mediante los algoritmos de Procesamiento del Lenguaje Natural (PLN). A partir de ahí, se utilizan algoritmos de Machine

Learning como SVM para obtener información concreta como la detección de perfiles de usuario en el dominio turístico (Torres, S., 2017).

- Pronóstico de cancelación de reservas hoteleras.

Las cancelaciones de reservas tienen un impacto significativo en las decisiones que se toman para gestionar la demanda en la industria hotelera y por esta razón los hoteles suelen implementar unas políticas de cancelación estrictas y unas tácticas arriesgadas de overbooking. Por lo tanto, ser capaz de conocer la probabilidad de que una reserva se cancele, o no, es una de las prioridades de algunos investigadores los últimos años.

Algunos de los autores de trabajos que tienen como objetivo predecir la cancelación de reservas hoteleras son Almeida et al (2017), Falk y Vieru (2018), Sánchez-Medina y Sánchez (2020) y Alotaibi (2020). Como se puede observar todos estos trabajos son muy recientes y buscan maximizar el porcentaje de acierto de la predicción comprendiendo las relaciones entre las variables, estudiando la importancia de cada una de ellas, utilizando series temporales y aplicando modelos de Machine Learning como redes neuronales artificiales, bosques aleatorios o SVM, entre otros. El experto en Ciencia de Datos Grogan (2020) implementó un modelo SVM que consiguió predecir cancelaciones con un 76% de acierto, superando el rendimiento de otros modelos como KNN, Naive-Bayes y XGBoost.

De las cuatro áreas de aplicación más comunes en el sector que hemos visto, la más asequible a priori es esta última. A pesar de ello, se puede obtener mucha información interesante a partir de la aplicación de modelos de Machine Learning a un conjunto de datos formado por características de reservas hoteleras y en este proyecto vamos a tratar de mostrar algunas de estas aplicaciones.

9. METODOLOGÍA

9.1. Tipo de metodología

Para el proyecto he elegido utilizar una metodología ágil. La metodología ágil consiste en dividir el proyecto en pequeñas partes que pueden completarse de manera individual en relativamente poco tiempo. Esta metodología permite estructurar el trabajo de tal forma que se perciban avances de tareas concretas en

cortos espacios de tiempo, manteniendo así una alta motivación entre las partes que participan en el proyecto. Otra de las ventajas de la metodología ágil es que está abierta a los cambios que puedan surgir en las diferentes fases del proyecto. Por lo tanto, permite una gran flexibilidad para modificar ciertas partes del mismo sin perjudicar necesariamente al resto de etapas.

De entre todas las metodologías ágiles he seleccionado basarme en la metodología Scrum, ya que consiste en realizar entregas parciales y regulares del proyecto que permiten a todas las partes ser conscientes de los avances y a su vez corregir o mejorar el mismo gracias a la retroalimentación para así cumplir con sus expectativas. También está indicada para proyectos con requisitos cambiantes o poco definidos y para proyectos que necesitan resultados en cortos períodos de tiempo. Por tanto, esta metodología, de una manera adaptada a estas condiciones, es ideal para proyectos como este que está abierto a modificaciones debido a complicaciones que puedan surgir y permite tener un control absoluto sobre la planificación del proyecto, lo cual es muy útil para saber si puede ser necesaria una de estas modificaciones para tener el proyecto listo en el tiempo previsto.

9.2. Herramientas utilizadas

Para la realización de este proyecto he utilizado el lenguaje de programación Python, en su versión 3.8.3. Python es un lenguaje de programación interpretado, de alto nivel y que se caracteriza por tener una sintaxis sencilla y legible. Debido a sus características, Python es frecuentemente utilizado en ámbitos educativos, para el desarrollo web y en el ámbito científico y numérico, por lo que es el lenguaje más utilizado en inteligencia artificial y es el ideal para este proyecto.

La aplicación Anaconda, con el intérprete Spyder, es la que he elegido para escribir el código necesario para el proyecto. Los paquetes y librerías que incluye son especialmente útiles para proyectos como este. Una librería es una interfaz que ofrece una funcionalidad concreta que suele ser compleja o de uso recurrente y tiene como objetivo simplificar nuestro código, ya que evita que tengamos que programar estas funciones nosotros mismos.

Las librerías de Python más importantes que he utilizado en este proyecto son *Pandas*, para el tratamiento de datos, *Matplotlib* y *Seaborn*, para la visualización de

datos, *Numpy*, para el cálculo algebraico y *Scikit-Learn*, para la implementación de algoritmos de Machine Learning.

10. DESARROLLO DEL PROYECTO E INTERPRETACIÓN DE LOS RESULTADOS

10.1. Obtención del conjunto de datos

El conjunto de datos que vamos a utilizar en este proyecto se ha obtenido gracias a la investigación de los autores Antonio, Almeida y Nunes (2019) y se puede acceder a ellos de forma gratuita a través de la plataforma Kaggle.

El conjunto de datos está formado por dos subconjuntos que comparten la misma estructura. El primer subconjunto está formado por 40.060 reservas hoteleras y hace referencia a un hotel resort situado en Algarve (Portugal). El segundo está formado por 79.330 reservas hoteleras que hacen referencia a un hotel urbano situado en Lisboa (Portugal). En total son 119.390 reservas que contienen 31 variables que las describen (32 si incluimos el tipo de hotel).

Los datos proceden de dos hoteles portugueses reales y por ello los datos relacionados con la identificación del hotel y de los clientes ha sido eliminada por los propios autores con el fin de preservar la privacidad.

A continuación se muestra un esquema de las variables del conjunto de datos (Figura 10.1.).

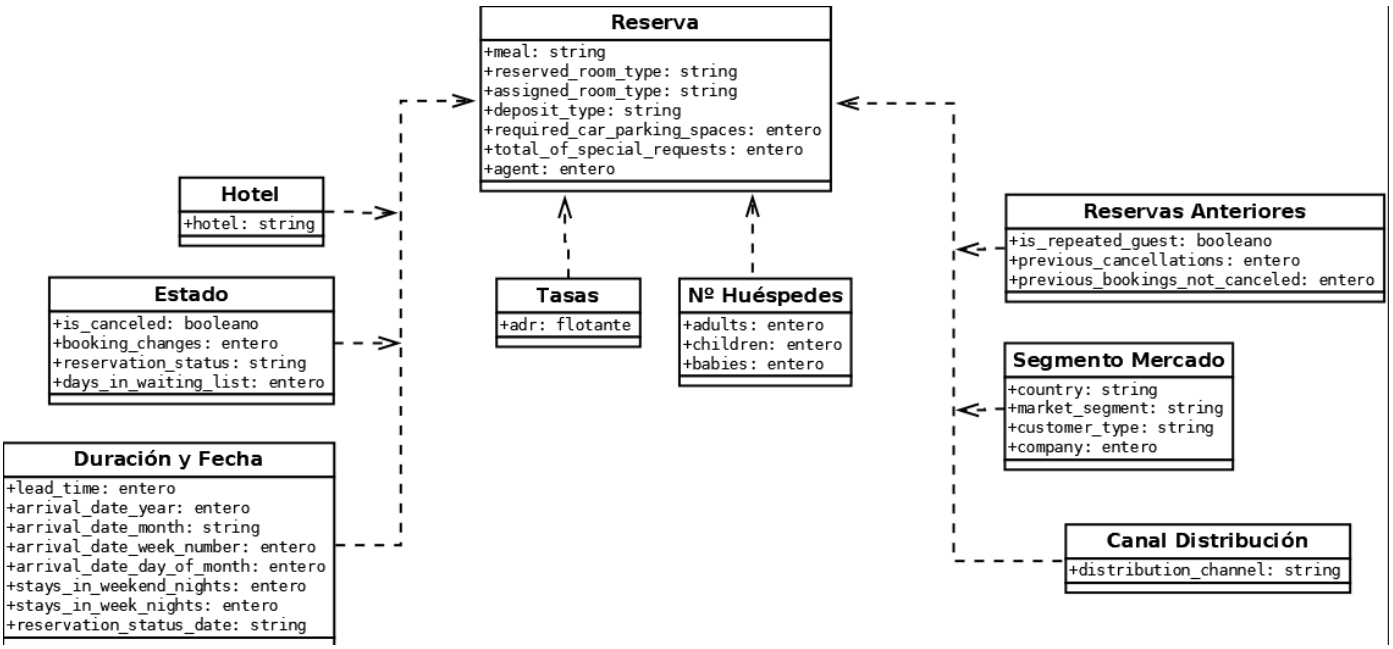


Figura 10.1. Esquema relacional de las variables del conjunto de datos

10.2. Análisis Exploratorio de Datos

El Análisis Exploratorio de Datos (EDA) consiste en el tratamiento de los datos obtenidos como muestra y es de vital importancia para cualquier proyecto de Aprendizaje Automático. Algunos de los objetivos del EDA son:

- Comprobar cuántos registros hay, es decir, el número de filas del conjunto.
- Comprobar cuántas variables hay, es decir, el número de columnas.
- Conocer el tipo de datos de cada variable, si son continuas, discretas, enteros o caracteres.
- Encontrar valores corruptos, nulos o mal escritos y decidir cómo tratarlos.
- Entender si los datos dependen del tiempo. De ser así estamos ante una serie temporal.
- Encontrar patrones y distribuciones estadísticas de las variables tales como correlación entre características, importancia de variable o análisis de importancia.

A partir de este estudio, dependiendo de cuál sea el objetivo, podrán seleccionarse las variables y los registros adecuados para cada caso, pudiendo desechar algunos de ellos.

Este proyecto está dividido en seis partes diferenciadas: el análisis bivariado, el análisis multivariado, la predicción de la tasa media diaria y la predicción de

cancelación para todo el dataset, para el hotel resort y para el hotel urbano. Por esta razón no podemos realizar un único EDA para todo el proyecto. En su lugar, en el siguiente punto realizaré un pequeño análisis general para obtener un set de datos preparado para ser tratado posteriormente para cada parte específica del trabajo.

10.3. Primer análisis de los datos

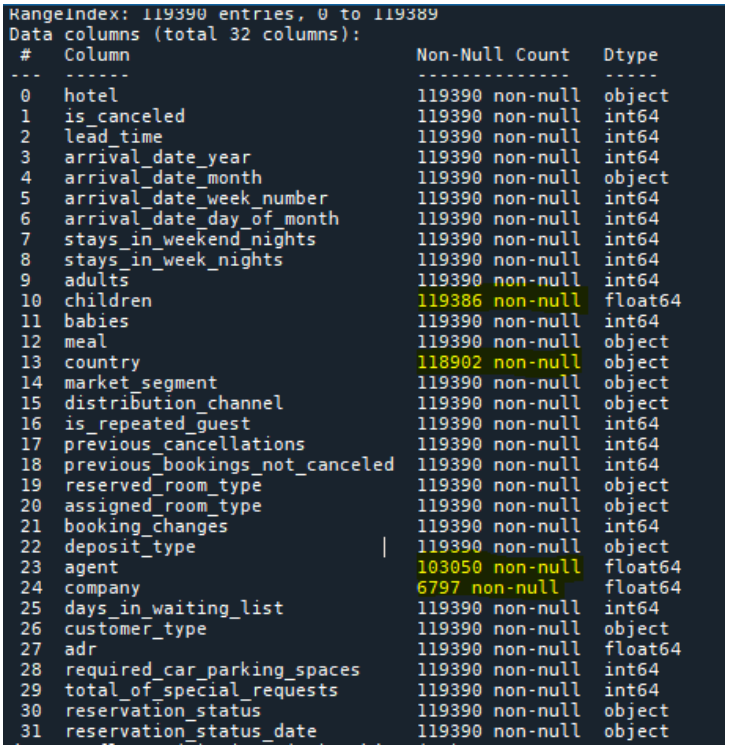
Este primer análisis consiste en conocer las características principales de nuestro conjunto de datos, encontrar y tratar los valores nulos o corruptos y transformar los datos a enteros o bien prepararlos para que puedan ser analizados como una serie temporal.

Características principales

Como hemos mencionado anteriormente, el dataset original está compuesto por 119.390 registros y 32 variables, de las cuales 16 son numéricas de tipo entero, 4 numéricas de tipo decimal y 12 son cadenas de texto. Una descripción detallada de cada una de las variables que incluye el contenido de todos sus valores puede encontrarse en el Anexo A.1.

Tratamiento de los valores nulos

En la siguiente tabla (Figura 10.2.) se observa en primera instancia que cuatro de las treinta y dos variables contienen valores nulos: *children*, *country*, *agent* y *company*, aunque la cantidad de los mismos varía considerablemente siendo 4, 488, 16.340 y 112.543 respectivamente.



```
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
```

Figura 10.2. Variables y sus valores no nulos

Nota: object=cadena de caracteres, int64=entero, float64=decimal

Una vez identificados, procedemos a su tratamiento. En primer lugar nos centramos en las variables *agent* y *company*, que identifican el agente y la empresa de cada reserva, respectivamente, y tienen un número muy elevado de valores nulos. Al analizar sus valores (Anexo A.1) observamos que cada dígito corresponde al código de un agente o de una empresa, de lo cual podemos inferir que los valores nulos corresponden a aquellas reservas que no están relacionadas con ningún agente o empresa. En este caso, el tratamiento más inteligente es sustituir los valores nulos por el valor 0, que no aparece en los datos originales y que por tanto puede indicar la ausencia de agente o de empresa. De esta forma, los datos de estas dos variables ya no son nulos y no hemos perdido la información de sus registros.

Sin embargo, el tratamiento de las variables *children* y *country* será diferente. La primera contiene cuatro valores nulos 'NA', pero entre sus valores también encontramos el valor 0, lo que quiere decir que el hecho de que el dato no esté disponible no garantiza que no hubiera niños en dicho registro, pues podría tratarse de un error. Además, perder cuatro registros de un total de más de 100.000 a priori

perjudica menos que la posibilidad de introducir datos incorrectos. En el caso de la nacionalidad, un valor nulo nos impide conocer la nacionalidad del turista, que podría ser cualquiera de las existentes. Por lo tanto, la decisión que tomamos respecto a estas dos variables es borrar todos los registros que contengan valores nulos en dichas columnas, esto corresponde a un total de 492 registros.

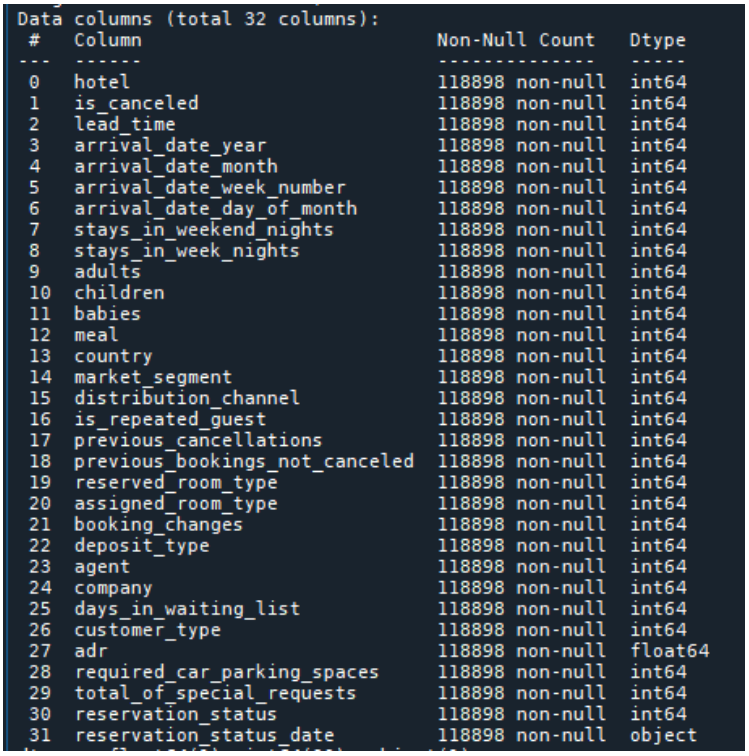
Como pequeña observación antes de avanzar al siguiente paso, en el esquema detallado de las variables hemos observado que las variables *children*, *agent* y *company* eran decimales, a pesar de que todos sus valores son siempre enteros. Por esta razón, en este paso también han sido transformados para evitar confusiones.

Tras este primer paso tenemos un dataset con 118.898 registros y podemos observar los cambios realizados con éxito (Anexo A.2).

Transformación de datos

Otra modificación interesante en este primer análisis es transformar los datos que son cadenas de texto en números enteros. Esta transformación se puede realizar de dos maneras, de forma automática o de forma manual. En el primer caso, las variables se mapean secuencialmente por orden alfabético empezando por el 0. Convertimos cuatro variables de forma automática: *country, reserved_room_type, assigned_room_type* y *market_segment*. En la conversión de forma manual nosotros seleccionamos el valor entero que le asignamos a cada variable. En este caso seleccionamos las siguientes variables: *hotel*, *arrival_date_month*, *meal, distribution_channel, deposit_type, customer_type, reservation_status.*

De esta manera, en la siguiente tabla (Figura 10.3.) observamos que los datos se han transformado correctamente:



```
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118898 non-null  int64
 1   is_canceled                     118898 non-null  int64
 2   lead_time                       118898 non-null  int64
 3   arrival_date_year               118898 non-null  int64
 4   arrival_date_month              118898 non-null  int64
 5   arrival_date_week_number        118898 non-null  int64
 6   arrival_date_day_of_month       118898 non-null  int64
 7   stays_in_weekend_nights         118898 non-null  int64
 8   stays_in_week_nights            118898 non-null  int64
 9   adults                          118898 non-null  int64
 10  children                        118898 non-null  int64
 11  babies                          118898 non-null  int64
 12  meal                            118898 non-null  int64
 13  country                         118898 non-null  int64
 14  market_segment                  118898 non-null  int64
 15  distribution_channel            118898 non-null  int64
 16  is_repeated_guest               118898 non-null  int64
 17  previous_cancellations          118898 non-null  int64
 18  previous_bookings_not_canceled  118898 non-null  int64
 19  reserved_room_type              118898 non-null  int64
 20  assigned_room_type              118898 non-null  int64
 21  booking_changes                 118898 non-null  int64
 22  deposit_type                    118898 non-null  int64
 23  agent                           118898 non-null  int64
 24  company                         118898 non-null  int64
 25  days_in_waiting_list            118898 non-null  int64
 26  customer_type                   118898 non-null  int64
 27  adr                             118898 non-null  float64
 28  required_car_parking_spaces     118898 non-null  int64
 29  total_of_special_requests       118898 non-null  int64
 30  reservation_status              118898 non-null  int64
 31  reservation_status_date         118898 non-null  object
```

Figura 10.3. Tabla de variables tras la transformación a números enteros

También está disponible la tabla detallada de asignaciones de cada variable (Anexo A.3).

Fechas

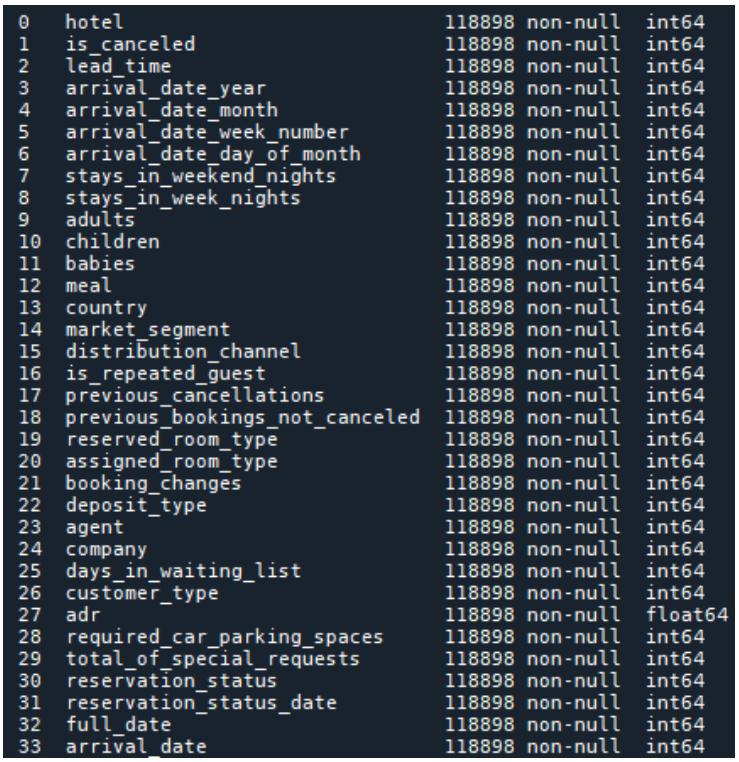
Al analizar las variables anteriormente, hemos visto que algunas de ellas son temporales, ya que nos dan información acerca del día, mes o año en el que se actualizó el estado de la reserva o de la llegada. No obstante, podemos observar que el día, la semana, el mes y el año de llegada están distribuidos en variables diferentes, por lo que nos puede resultar de utilidad crear dos nuevas variables a partir de ellas para establecer la estacionalidad con claridad: *arrival_date* y *full_date*.

arrival_date: Combina las variables *arrival_date_year*, *arrival_date_month* y *arrival_date_day_of_month* en una única variable que muestra la fecha de llegada, sin guiones, y será de tipo entero. Ejemplo de valor de esta variable: ‘20171001’ (equivalente a ‘01-10-2017’).

full_date: Combina las variables *arrival_date_year* y *arrival_date_week_number* en una única variable de tipo entero que nos indica la semana del año en que cada cliente llegó. Ejemplo de valor: ‘201750’ (semana 50 del año 2017).

De manera paralela transformamos la variable *reservation_status_date*, que era una cadena de caracteres, a número entero, eliminando los guiones que separan los años, meses y días. De no hacer esto, podríamos tener problemas al ordenar los datos, ya que el orden de una cadena de texto no coincide con el de una fecha.

De esta manera hemos obtenido un set de datos con 34 variables listo para ser tratado de manera individual para cada apartado específico (Figura 10.4.)



0	hotel	118898 non-null	int64
1	is_canceled	118898 non-null	int64
2	lead_time	118898 non-null	int64
3	arrival_date_year	118898 non-null	int64
4	arrival_date_month	118898 non-null	int64
5	arrival_date_week_number	118898 non-null	int64
6	arrival_date_day_of_month	118898 non-null	int64
7	stays_in_weekend_nights	118898 non-null	int64
8	stays_in_week_nights	118898 non-null	int64
9	adults	118898 non-null	int64
10	children	118898 non-null	int64
11	babies	118898 non-null	int64
12	meal	118898 non-null	int64
13	country	118898 non-null	int64
14	market_segment	118898 non-null	int64
15	distribution_channel	118898 non-null	int64
16	is_repeated_guest	118898 non-null	int64
17	previous_cancellations	118898 non-null	int64
18	previous_bookings_not_canceled	118898 non-null	int64
19	reserved_room_type	118898 non-null	int64
20	assigned_room_type	118898 non-null	int64
21	booking_changes	118898 non-null	int64
22	deposit_type	118898 non-null	int64
23	agent	118898 non-null	int64
24	company	118898 non-null	int64
25	days_in_waiting_list	118898 non-null	int64
26	customer_type	118898 non-null	int64
27	adr	118898 non-null	float64
28	required_car_parking_spaces	118898 non-null	int64
29	total_of_special_requests	118898 non-null	int64
30	reservation_status	118898 non-null	int64
31	reservation_status_date	118898 non-null	int64
32	full_date	118898 non-null	int64
33	arrival_date	118898 non-null	int64

Figura 10.4. Tabla final tras el primer análisis

10.4. Análisis bivariado

El análisis bivariado consiste en un estudio de dos variables, analizando algunas de las relaciones que existen entre ellas. En este apartado voy a enfocarme en una variable importante como es el segmento de mercado, *market_segment*, y su relación con algunas de las otras variables. Para seleccionar las variables más interesantes a comparar voy a aplicar un algoritmo llamado Importancia de Variable.

La Importancia de Variable (VI) representa la significancia estadística de todas las variables x_i que utilizamos para predecir la variable y. Para entender esta definición, debemos saber que la significancia es la probabilidad de que la relación entre dos variables no sea casual, sino que es debida a cierto factor F.

Para ello, vamos a utilizar un modelo llamado *Extra Trees Classifier*, disponible en la librería *Scikit-Learn*. Este modelo implementa un estimador que aplica una serie de árboles aleatorios de decisión en varias muestras de nuestro conjunto de datos utilizando algunos estadísticos para mejorar la precisión en la predicción.

En nuestro proyecto, esto se traduce en que suponiendo que queremos predecir la variable *market_segment*, al aplicar el modelo *Extra Trees Classifier* obtendremos la importancia que tiene cada una de las variables para predecir *market_segment*. Es decir, obtenemos las que están más conectadas a nuestra variable. La importancia viene dada por un número entre 0 y 1, siendo 1 muy alta y 0 muy baja o nula.

Tras la aplicación del modelo, obtenemos que las variables con mayor significancia en relación con el segmento de mercado son *distribution_channel*, *agent* y *customer_type* (figura 10.5). Mientras que *babies* es la que tiene la menor significancia. (Tabla completa de significancias en Anexo A.4). De esto podemos deducir que el canal de distribución es clave para conocer el segmento de mercado, mientras que conocer el número de bebés es prácticamente irrelevante.

Index	feature	v_importance
0	distribution_channel	0.179422
1	agent	0.129067
2	customer_type	0.0734242
3	deposit_type	0.0598706
4	lead_time	0.0443011
5	adr	0.0429902

Figura 10.5. Importancia de Variable

Para realizar el análisis bivariado nos ayudaremos de las siguientes gráficas:

Gráfico de Barras de Conjunto Múltiple: la longitud de cada barra se utiliza para mostrar comparaciones numéricas discretas entre variables. A cada segmento de datos de cada grupo de barras se le asigna un color que representa a su clase (Figura 10.6.). Lo utilizaremos para representar variables discretas como *distribution_channel*.

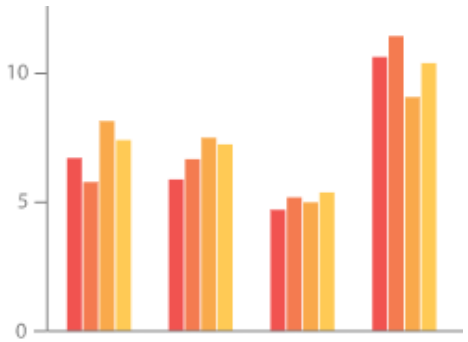


Figura 10.6. Gráfico de barras de conjunto múltiple

Diagrama de Caja y Bigotes: También conocido como boxplot, representa la distribución de puntuaciones de una variable y señala los valores atípicos a través de varios estadísticos descriptivos (Figura 10.7.). La mediana o segundo cuartil (Q2) divide en dos partes iguales la distribución, de tal forma que el 50% de los valores son menores o iguales a este valor. El primer y el tercer cuartil (Q1 y Q3) son similares a la mediana, salvo que representan el 25% y el 75% de los valores, respectivamente. Por otra parte, los bigotes marcan la frontera superior e inferior entre los datos anómalos o atípicos y el resto de datos.

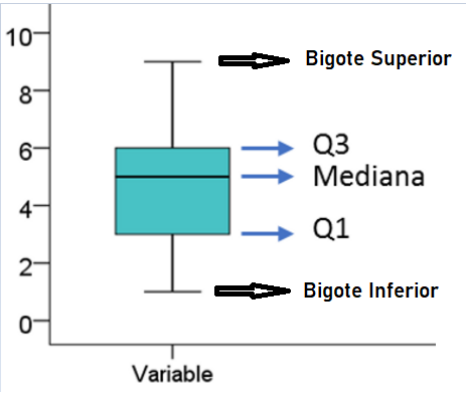


Figura 10.7. Diagrama de caja y bigotes

He aplicado el análisis bivariado a las siguientes variables: *lead_time, adr, distribution_channel, agent, customer_type, total_of_special_requests, meal, hotel, arrival_date, reservation_status_date, full_date* y *arrival_date_week_number*. Los resultados de todas ellas pueden encontrarse en el Anexo A.5.

Vamos a ver tres de estas gráficas obtenidas tras el análisis. En primer lugar, para la variable *lead_time* hemos aplicado un boxplot. Para una mejor visibilidad,

elimino aquellos datos excepcionalmente atípicos. Los datos atípicos aparecen representados como puntos en el exterior de las líneas que delimitan los bigotes. En la representación (figura 10.8) podemos observar cómo varía el tiempo de espera para cada uno de los diferentes segmentos de mercado, observando información interesante como que los grupos suelen reservar con mucho tiempo de antelación mientras que los segmentos de aviación y el complementario reservan muy cerca del día de llegada.

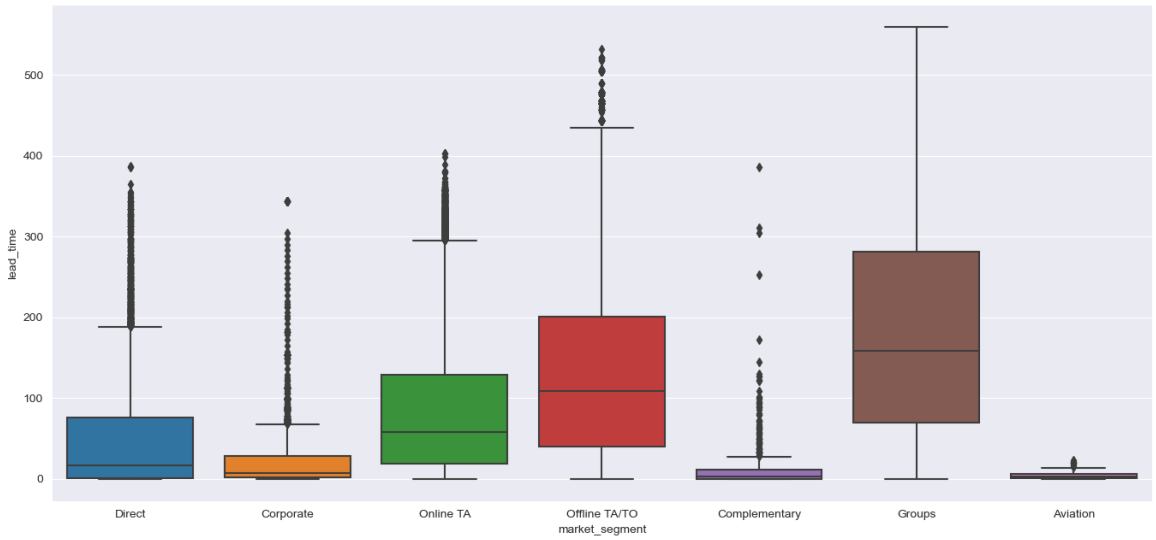


Figura 10.8. Diagrama de caja y bigotes: lead_time y market_segment

En segundo lugar, para la variable total_of_special_requests he usado un diagrama de barras. En él podemos observar el número de solicitudes especiales totales que cada segmento de mercado solicita. Se puede observar (figura 10.9) que aunque lo normal es que no haya ninguna solicitud, en el caso de las reservas que llegan a través de las OTAs lo normal es que existan peticiones especiales.

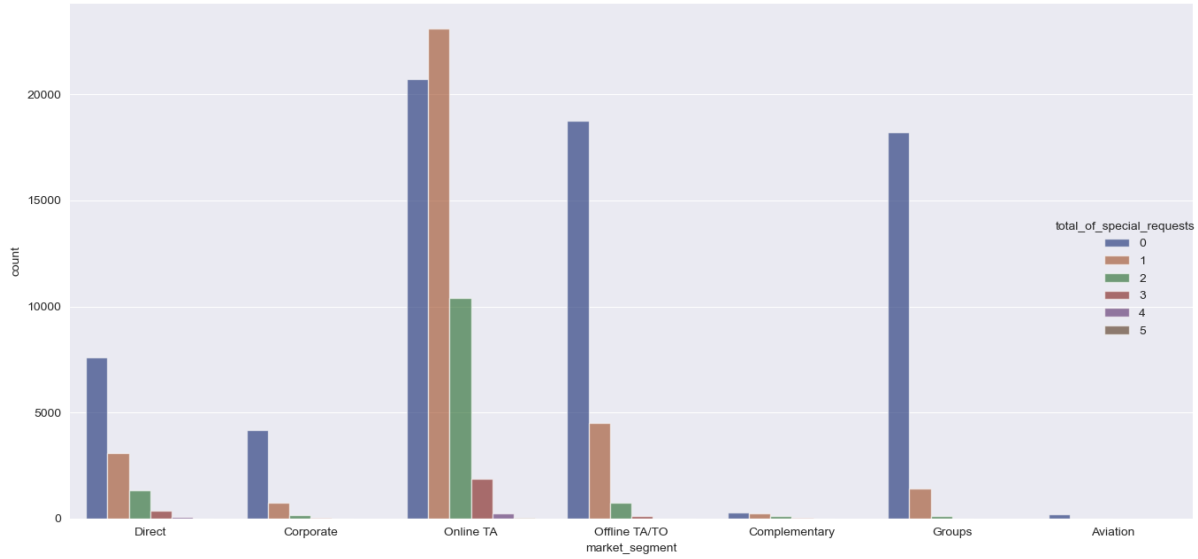


Figura 10.9. Diagrama de barras de conjunto múltiple: market_segment y total_of_special_requests

Por último, he representado *arrival_date* con un gráfico de barras. En la representación (figura 10.10) vemos algunos meses de 2017 que muestran la fecha de llegada en función del segmento de mercado y como era de esperar, se observa estacionalidad, ya que aumenta considerablemente en los meses de verano.

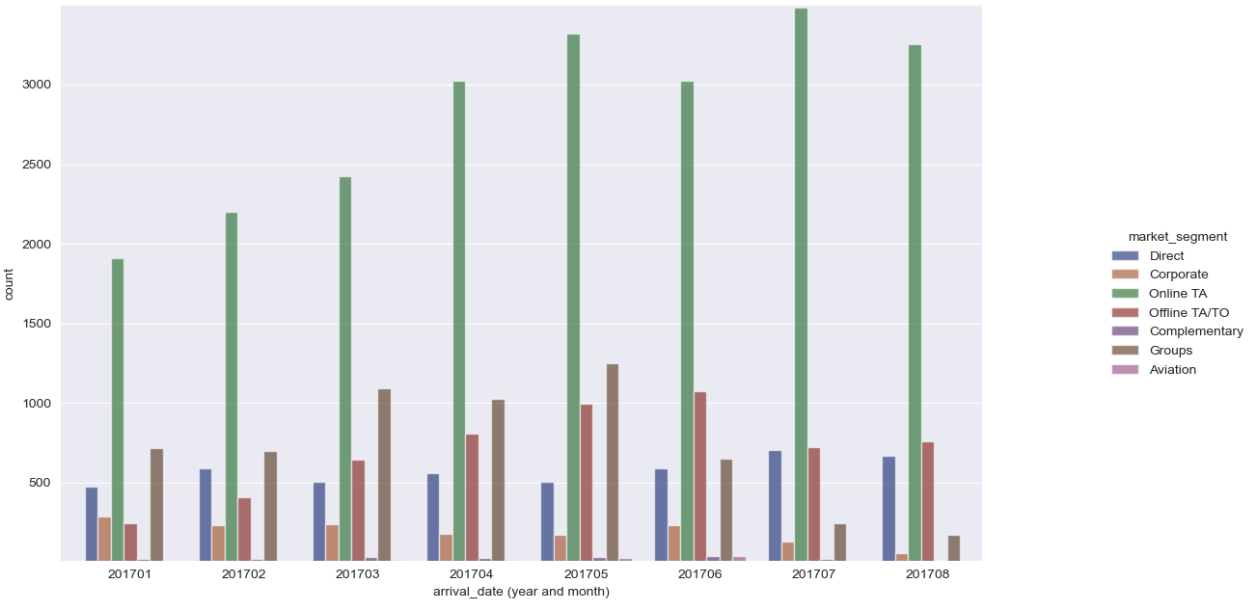


Figura 10.10. Diagrama de barras de conjunto múltiple: market_segment y arrival_date

10.5. Análisis multivariado

El análisis multivariado consiste en analizar las relaciones entre tres o más variables. Al igual que en el apartado anterior, la variable clave en este punto será de nuevo *market_segment*. Para realizar el análisis multivariado nos ayudaremos de las siguientes gráficas:

Gráfico de Barras de Conjunto Múltiple: Ya definido en el punto anterior.

Diagrama de dispersión: También conocido como Scattergram (figura 10.11), representa los valores de dos variables mediante puntos colocados en coordenadas cartesianas y permite detectar la correlación entre ambas variables, si es que existe.

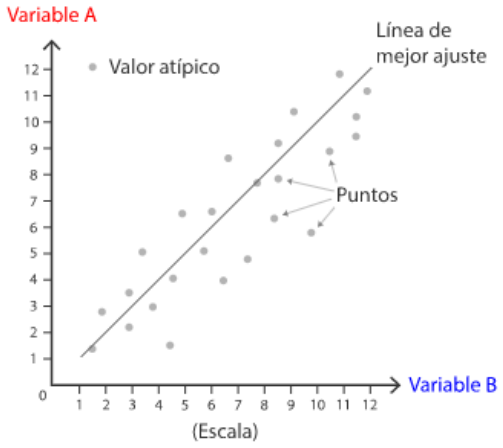


Figura 10.11. Diagrama de dispersión

Otra herramienta que vamos a aplicar en este análisis multivariado es un algoritmo de Machine Learning de aprendizaje no supervisado llamado Análisis de Componentes Principales (PCA).

El PCA es una técnica de extracción de características que se utiliza entre otras cosas para reducir la dimensionalidad de un conjunto de datos, es decir, reducir su número de variables de tal forma que las nuevas variables obtenidas sean independientes entre sí. Esto, además, lo hace tratando de mantener la parte más importante de todas las variables originales (retención de varianza), por lo que la información del conjunto de datos no varía excesivamente.

Aplicaremos el PCA a nuestro conjunto de datos extrayendo anteriormente la variable market_segment y las variables numéricas discretas. De esta forma

podemos representar el segmento de mercado en dos dimensiones tras la aplicar la reducción de dimensionalidad y, por otra parte, compararlo de manera similar al análisis bivariado con cada una de las nuevas variables, que contendrán información de un grupo de las variables originales, es por eso que este análisis es multivariado.

Para ello, primero he extraído del dataset las variables categóricas y discretas, así como la variable objetivo market_segment. Después he normalizado el resto de variables y he aplicado el modelo de PCA. La normalización es la transformación de escala de la distribución de una variable con el objetivo de poder hacer comparaciones respecto a conjuntos de elementos, a la media y a la varianza (Rodó, P., 2019). Este paso es obligatorio para aplicar el PCA.

A continuación se muestra en un diagrama de dispersión la representación en dos dimensiones del segmento de mercado en función de sus dos componentes principales obtenida gracias a nuestro modelo entrenado (figura 10.12). También adjunto 5 diagramas de caja y bigotes que comparan cada una de las componentes principales obtenidas con el segmento de mercado, reteniendo el 51% de la varianza (anexo A.6).

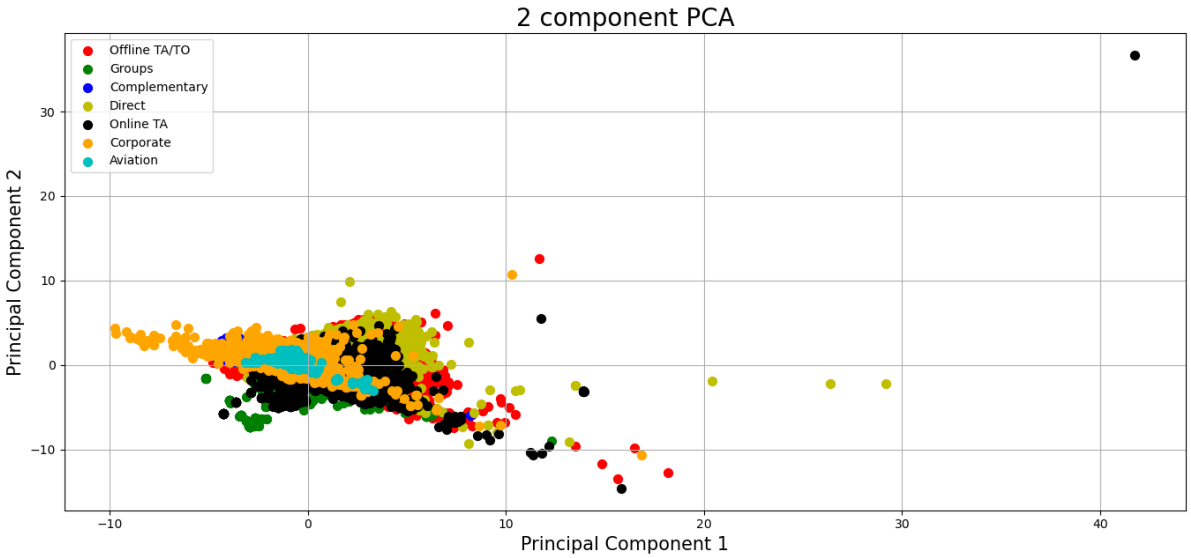


Figura 10.12. Representación del segmento de mercado en función de sus dos componentes principales (Scattergram)

Como era de esperar, se observa que cada tipo de segmento de mercado aparece agrupado de manera similar, aunque diferenciable. No obstante, hay que

tener en cuenta que la retención de varianza obtenida (51%) es bastante baja. Esto quiere decir que el 49% de la información se ha perdido tras la reducción de dimensionalidad, por lo que lamentablemente estos resultados no aportan información del todo fiable y, por tanto, relevante.

10.6. Predicción de cancelación

El objetivo de este apartado es crear un algoritmo que pueda predecir si una reserva está cancelada o no, en función del resto de variables. Dado que contamos con dos conjuntos de datos que hacen referencia a un hotel urbano y a un hotel resort, puede ser interesante crear tres modelos diferentes. El primero incluirá los datos de ambos hoteles, y los otros dos solamente incluirán los datos de uno de los hoteles. De esta manera podemos analizar si existen diferencias notables en la predicción.

Antes de diseñar nuestros modelos debemos decidir si vamos a utilizar todas las variables para nuestra predicción. Para elegir nuestro dataset nos ayudaremos de la importancia de variable (VI) y de la matriz de correlación.

La matriz de correlación establece cómo se relacionan las variables entre sí. Para que sea más visual, nos ayudaremos de un mapa de calor usando la librería de Python *Seaborn*. Por otro lado, tal y como hicimos en el punto 10.4, para la importancia de variable usaremos el modelo *Extra Trees Classifier*. Tanto este modelo como la importancia de variable fueron explicados previamente en ese punto.

Para realizar la predicción se van a utilizar dos de los algoritmos de aprendizaje supervisado que suelen dar mejores resultados para problemas de clasificación:

Máquinas de Vectores de Soporte (SVM): Clasificador lineal que se basa en el principio de maximización de márgenes con el objetivo de separar los datos en dos categorías. El SVM puede ser ampliado también para resolver problemas de regresión, en este caso se le conoce como *Regresión de Vectores de Soporte (SVR)*

Bosques Aleatorios (Random Forest): Un metaestimador que utiliza una serie de clasificadores de árboles de decisión en varias submuestras del conjunto de datos y

utiliza promedios para mejorar la precisión predictiva. Además de clasificación, también puede ampliarse a problemas de regresión.

Por otra parte, para optimizar estos modelos utilizaremos otro metaestimador llamado *GridSearchcv*, que aplicando la validación cruzada o cross validation (cv) encuentra la mejor selección de parámetros en cada modelo para que éstos obtengan los resultados más óptimos.

Normalmente se suele dividir el set de datos en dos partes de cara a evaluar su precisión. Un 80% se utiliza para entrenar el modelo, aprendiendo a predecir en nuestro caso la cancelación o la tasa media diaria. Para ello la máquina tendrá acceso al resultado para poder así sacar conclusiones y una precisión aproximada. La efectividad del modelo se pondrá a prueba en el 20% restante, en este caso ya sin saber los resultados de antemano. Los resultados los tendremos nosotros para poder así valorar la efectividad real del modelo. En este contexto, la validación cruzada consiste en fragmentar el 20% de test n veces tomando partes diferentes y aleatorias del dataset. La precisión media obtenida tras los n testeos, que en modelos bien implementados ha de tener valores similares, se acerca mucho más a la precisión real del entrenamiento, y en el caso del Grid, le permite elegir los mejores parámetros con una mayor fiabilidad. Para vislumbrar la mejora que se consigue con la optimización voy a realizar una comparativa.

El último detalle a tener en cuenta en nuestros modelos de clasificación es que los datos están desbalanceados. Esto ocurre debido a que en nuestro set de datos tenemos una clase mayoritaria, las reservas no canceladas, y una clase minoritaria, las reservas que sí fueron canceladas. Ignorar este hecho provocaría que el modelo predijera siempre que la reserva no se va a cancelar para obtener un rendimiento con un aceptable, incluso elevado, número de aciertos, algo que no sería de utilidad real, ya que en este caso una alta precisión no garantiza una buena predicción.

Con el objetivo de vislumbrar que independientemente de la precisión nuestro modelo funciona de manera adecuada utilizaremos la matriz de confusión (figura 10.13) y las métricas precisión, recall y el coeficiente F1 (F1 score).

La matriz de confusión nos muestra el total de aciertos (true positive) y los fallos (false positive) de cada una de las clases.

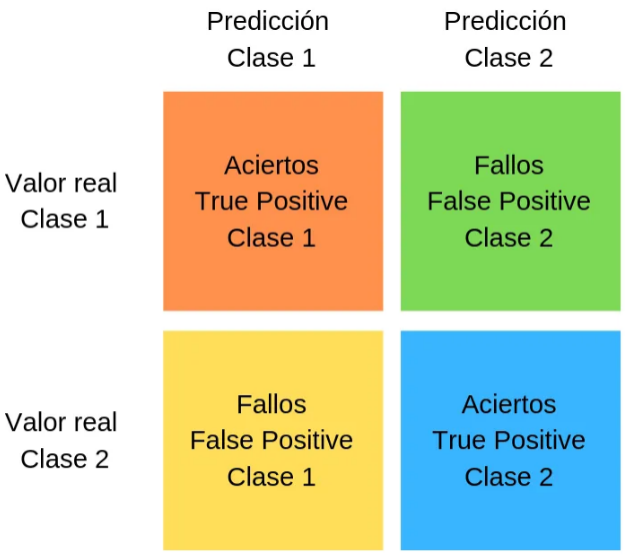


Figura 10.13. Matriz de confusión

La *eficiencia o accuracy* del modelo es el total de predicciones acertadas dividido por el total de predicciones.

La *precisión* de una clase nos dice la capacidad del modelo para responder correctamente si un dato corresponde a esa clase.

El *recall* de una clase nos dice la capacidad del modelo para identificar esa clase.

El coeficiente F1 de una clase combina la precisión y el recall en una misma métrica con la fórmula (2 x precision x recall / (precision+recall)).

Lo ideal en cualquier modelo es que todas las métricas anteriormente mencionadas estén lo más cerca posible del 100% para cada una de las clases.

Sin embargo, en un dataset desequilibrado lo normal es que la precisión en la clase mayoritaria sea muy alta y el recall de la clase minoritaria sea muy bajo. Para solucionar esto equilibraremos ambas clases penalizando a la clase mayoritaria mediante el ajuste de los parámetros del modelo.

10.6.1. Hotel urbano

En la figura 10.14 podemos observar algunas de las variables más importantes a la hora de predecir si una reserva está cancelada. La fecha en la que se actualizó el estado de la reserva, el tipo de depósito y la nacionalidad parecen ser claves para la predicción de cancelaciones en un hotel urbano. Por el contrario, el número de

bebés y datos de reservas anteriores no parecen ser apenas relevantes. Algunas de las variables menos importantes podrían ser eliminadas. ya que no son útiles para nuestra predicción. En el anexo A.7 podemos ver la tabla completa.

0	reservation_status_date	0.14665	25	children	0.00575011
1	deposit_type	0.13282	26	is_repeated_guest	0.00403039
2	country	0.0915002	27	company	0.0034386
3	lead_time	0.0798705	28	days_in_waiting_list	0.00327531
4	total_of_special_requests	0.060382	29	previous_bookings_not_canceled	0.00128719
5	full_date	0.0389473	30	babies	0.000745825

Figura 10.14. Importancia de variable respecto is_canceled para hotel urbano

En la figura 10.15 observamos un mapa de calor que muestra cómo se relacionan las variables entre ellas. En color granate vemos aquellas variables que tienen una alta correlación. Por ejemplo, is_canceled está estrechamente relacionada con reservation_status. Un alto nivel de correlación podría indicar que dos o más variables están aportando una información muy similar, por lo que podría ser útil eliminar una de las dos de nuestro dataset.

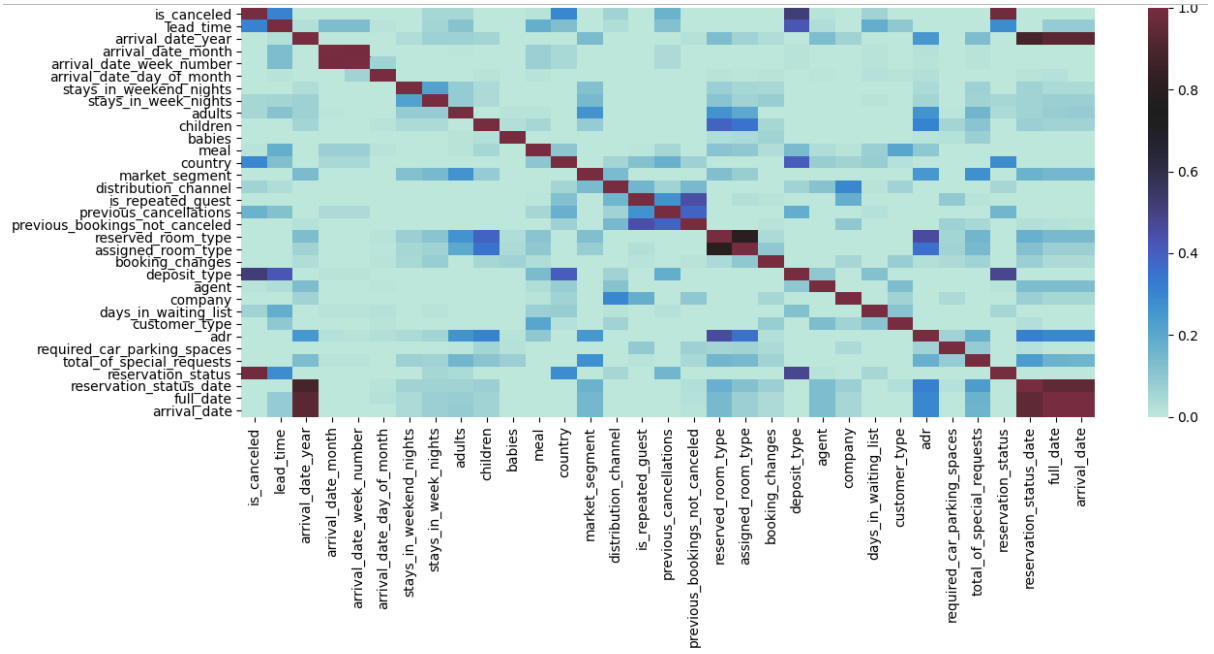
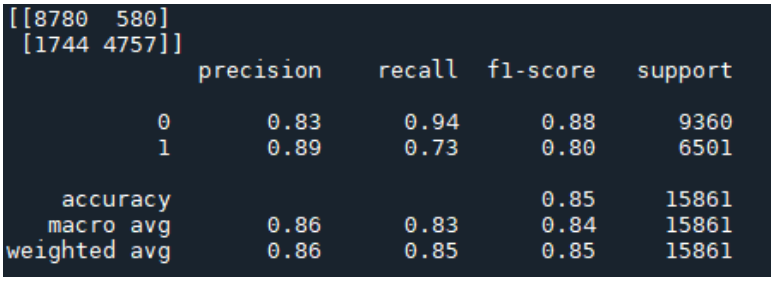


Figura 10.15. Matriz de correlación, hotel urbano

Teniendo en cuenta estos datos, extraigo algunos datos para obtener nuestro dataset final. En primer lugar vamos a extraer los registros correspondientes al hotel resort y después extraemos la propia columna *hotel.* En segundo lugar, extraemos cinco variables que hemos visto que tienen una importancia pequeña, que son *babies*, *previous_bookings_not_canceled*, *company*, *days_in_waiting_list*, *is_repeated_guest.* Por último, eliminamos cuatro variables debido a la alta correlación con *is_canceled*: *reservation_status*, *arrival_date_month*, *arrival_date_year* y *arrival_date_of_month.*

Aplico ahora el modelo SVM sin optimizar parámetros. Como he explicado anteriormente, en todos los casos entreno los modelos con el 80% de los datos y lo testeo en el 20% restante, para ver su precisión. Estos son los resultados obtenidos:



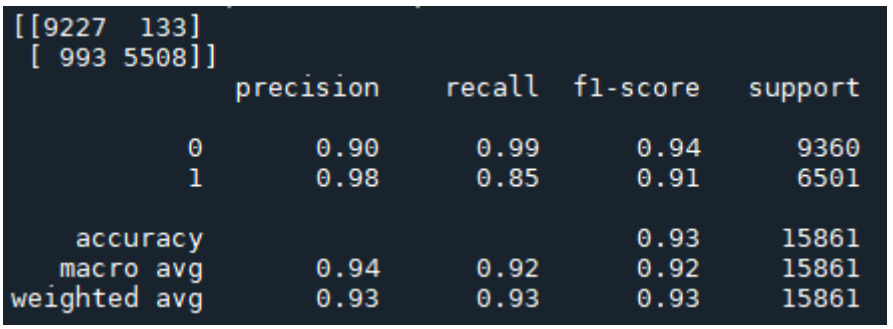
```
[[8780  580]
 [1744 4757]]
              precision    recall  f1-score   support

           0       0.83      0.94      0.88      9360
           1       0.89      0.73      0.80      6501

    accuracy                           0.85     15861
   macro avg       0.86      0.83      0.84     15861
weighted avg       0.86      0.85      0.85     15861
```

Figura 10.16. Resultados del modelo SVM sin optimizar para Hotel Urbano

El modelo tiene un 85% de precisión media, aunque si tenemos en cuenta que lo que nos interesa principalmente es predecir las reservas que van a cancelarse (predecir exitosamente cuándo *is_canceled = 1*), la precisión del modelo, en ese caso, desciende al 80%. Es un resultado a priori positivo, aunque observando la matriz de confusión (arriba a la izquierda) el modelo yerra en más de 2300 reservas. Aplicando la optimización estos son los resultados de este mismo modelo SVM:



```
[[9227  133]
 [ 993 5508]]
              precision    recall  f1-score   support

           0       0.90      0.99      0.94      9360
           1       0.98      0.85      0.91      6501

    accuracy                           0.93     15861
   macro avg       0.94      0.92      0.92     15861
weighted avg       0.93      0.93      0.93     15861
```

Figura 10.17. Resultados del modelo SVM optimizado para Hotel Urbano

La precisión media aumenta hasta el 93% y la predicción de reservas canceladas aumenta hasta el 91%, reduciendo los errores de 2324 a 1126, más de la mitad.

A continuación se muestran los resultados del modelo de Bosques Aleatorios, primero sin optimizar y posteriormente optimizados:

```
[[9131  229]
 [ 722 5779]]
              precision    recall  f1-score   support

           0       0.93      0.98      0.95      9360
           1       0.96      0.89      0.92      6501

    accuracy                           0.94     15861
   macro avg       0.94      0.93      0.94     15861
weighted avg       0.94      0.94      0.94     15861
```

Figura 10.18. Resultados del modelo de Bosques Aleatorios sin optimizar

```
[[9094  266]
 [ 676 5825]]
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      9360
           1       0.96      0.90      0.93      6501

    accuracy                           0.94     15861
   macro avg       0.94      0.93      0.94     15861
weighted avg       0.94      0.94      0.94     15861
```

Figura 10.19. Resultados del modelo de Bosques Aleatorios optimizado

En este caso el modelo ya tenía una precisión media del 94% sin optimizar y no mejora demasiado al optimizar, aunque comete 9 errores menos (942 frente a 951) y mejora en un 1% la precisión de reservas canceladas (93%).

En ambos casos hemos obtenido una buena predicción, aunque los resultados han sido ligeramente superiores utilizando el modelo de los Bosques Aleatorios.

A partir de ahora se mostrarán solo los resultados optimizados, ya que en mayor o menor medida todos han conseguido mejorar el modelo original. No obstante, el resto estarán disponibles en los anexos.

10.6.2. Hotel resort

Aunque los resultados obtenidos en un hotel resort son similares a los obtenidos en un hotel urbano, podemos observar algunas diferencias (figura 10.20). Por ejemplo, se aprecia una ligera mayor importancia de la nacionalidad y el tiempo de espera, por delante del tipo de depósito. Por otro lado, los días en la lista de espera

parece carecer de importancia a la hora de predecir una cancelación en este tipo de hoteles. En el anexo A.8. podemos ver la tabla completa.

0	reservation_status_date	0.143034
1	country	0.101421
2	lead_time	0.0744809
3	deposit_type	0.0564227
4	required_car_parking_spaces	0.0497723
5	full_date	0.0434706

25	children	0.010116
26	is_repeated_guest	0.00726938
27	company	0.00645666
28	previous_bookings_not_canceled	0.00259681
29	babies	0.00200427
30	days_in_waiting_list	0.00120434

Figura 10.20. Importancia de variable respecto is_canceled para hotel resort

En la matriz de correlación no se observan apenas diferencias entre un hotel urbano y un hotel resort (figura 10.21).

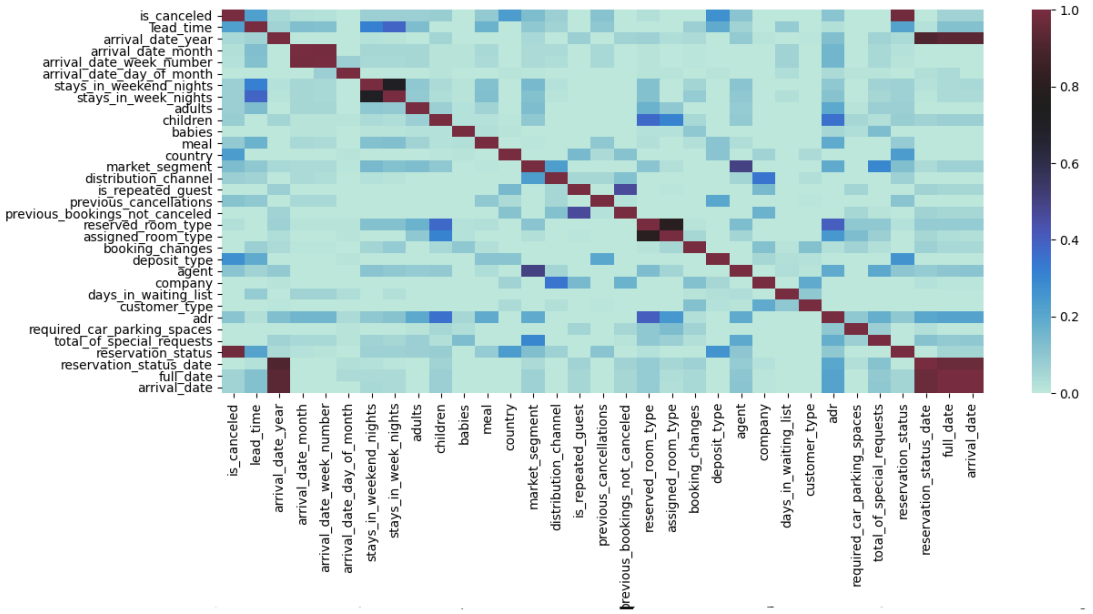


Figura 10.21. Matriz de correlación, hotel resort

En este caso, eliminamos los siguientes datos para obtener nuestro dataset final:

En primer lugar, eliminamos los registros del hotel urbano y la variable *hotel*. En segundo lugar, eliminamos siete variables con poca importancia: *babies*, *previous_bookings_not_canceled*, *days_in_waiting_list*, *is_repeated_guest*, *children*,

required_car_parking_spaces, *company*. Y debido a la alta correlación eliminamos cuatro variables: *reservation_status*, *arrival_date_month*, *arrival_date_year* y *arrival_date_of_month.* Aunque este dataset final es similar al anterior, observamos que hay ligeras diferencias.

Estos son los resultados obtenidos con el modelo de SVM optimizado:

```
[[5439  248]
 [ 277 1956]]
              precision    recall  f1-score   support

           0       0.95      0.96      0.95      5687
           1       0.89      0.88      0.88      2233

    accuracy                           0.93      7920
   macro avg       0.92      0.92      0.92      7920
weighted avg       0.93      0.93      0.93      7920
```

Figura 10.22. Resultados del modelo de SVM optimizado para Hotel Resort

Y estos son los resultados aplicando los Bosques Aleatorios optimizados:

```
[[5560  127]
 [ 332 1901]]
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      5687
           1       0.94      0.85      0.89      2233

    accuracy                           0.94      7920
   macro avg       0.94      0.91      0.93      7920
weighted avg       0.94      0.94      0.94      7920
```

Figura 10.23. Resultados del modelo de Bosques Aleatorios optimizado para Hotel Resort

En el segundo modelo se obtiene un resultado ligeramente superior, con un 94% de precisión media respecto al 93%. Además en un 89% de las veces predice una cancelación correctamente, respecto al 88% de SVM y yerra tan solo en 459 reservas respecto a las 525 del primero, de un total de 7920.

Por tanto nuevamente el Bosque Aleatorio arroja una mayor precisión, aunque ambos modelos superan el 90%.

10.6.3. Hotel urbano y hotel resort

Tomando nuestro dataset completo, los resultados que nos muestra la importancia de variable (figura 10.24) y la matriz de correlación (figura 10.25) son similares a los del hotel urbano. Podemos visualizar la tabla completa en el anexo A.9.

0	reservation_status_date	0.144045	26	children	0.00675653
1	deposit_type	0.108553	27	is_repeated_guest	0.00623503
2	country	0.0907956	28	company	0.0046337
3	lead_time	0.077154	29	days_in_waiting_list	0.00270911
4	total_of_special_requests	0.0492539	30	previous_bookings_not_canceled	0.00160368
5	full_date	0.0385076	31	babies	0.00106378

Figura 10.24. Importancia de variable respecto is_canceled para todo el dataset

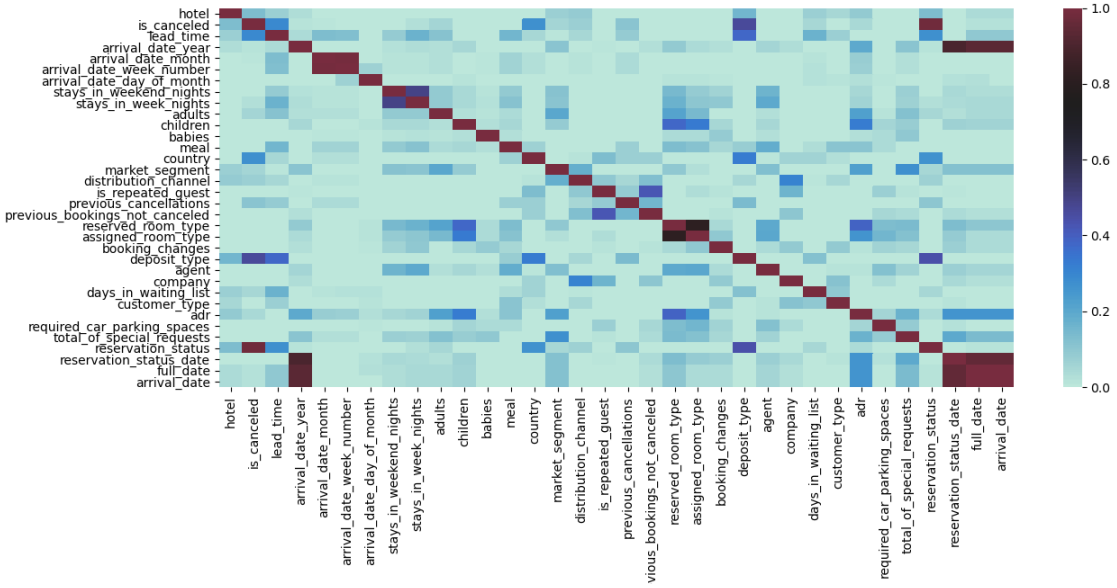


Figura 10.25. Matriz de correlación, hotel resort y hotel urbano

Con estos resultados, realizamos los cambios para obtener el dataset final. En este caso no es necesario eliminar ningún registro ni la variable *hotel*. Borramos también seis variables por la importancia de variable: *babies*, *previous_bookings_not_canceled*, *days_in_waiting_list*, *is_repeated_guest, children, company*. Y debido a la alta correlación con *is_canceled* eliminamos también *reservation_status*, *arrival_date_month*, *arrival_date_year* y *arrival_date_of_month.*

Estos son los resultados obtenidos aplicando los modelos optimizados de SVM y Random Forest a este dataset final:

```
[[14488   357]
 [ 1216  7719]]
              precision    recall  f1-score   support

           0       0.92      0.98      0.95     14845
           1       0.96      0.86      0.91      8935

    accuracy                           0.93     23780
   macro avg       0.94      0.92      0.93     23780
weighted avg       0.94      0.93      0.93     23780
```

Figura 10.26. Resultados del modelo de SVM optimizado para todo el dataset

```
[[14420   425]
 [  898  8037]]
              precision    recall  f1-score   support

           0       0.94      0.97      0.96     14845
           1       0.95      0.90      0.92      8935

    accuracy                           0.94     23780
   macro avg       0.95      0.94      0.94     23780
weighted avg       0.94      0.94      0.94     23780
```

Figura 10.27. Resultados del modelo de Bosques Aleatorios optimizado para todo el dataset

El primer modelo ha errado en 1573 reservas del total de 23780, mientras que el segundo erró en 1323. Tanto la precisión media como la precisión en las reservas canceladas solo varía en un 1%, siendo Random Forest con un 94% y un 92% respectivamente el que mejores resultados ha obtenido.

Tras analizar los modelos en dos hoteles diferentes por separado y también de forma conjunta, llegamos a la conclusión de que el Bosque Aleatorio arroja unos mejores resultados, y en todos los casos, bastante similares. Aunque bien es cierto que en cada caso aplicamos un tratamiento de los datos ligeramente diferente.

10.7. Predicción de la tarifa media diaria

El objetivo de este apartado es crear un algoritmo que pueda predecir la tarifa media diaria de una reserva, en función del resto de variables. En este caso lo haremos con todo el conjunto de datos.

De la misma manera que en la predicción de cancelación, en este apartado también vamos a aplicar la importancia de variable con el modelo *Extra Trees Classifier* y la matriz de correlación con un mapa de calor y la librería *Seaborn*.

Para esta predicción usaremos los ya mencionados algoritmos de regresión, la *Regresión de Vectores de Soporte (SVR)* y los *Bosques Aleatorios de Regresión*.

Al tratarse de un algoritmo de regresión en el que tenemos que predecir valores continuos ya no tiene sentido evaluar los resultados con la matriz de confusión ni con métricas como la precisión o el recall. Al tener que predecir un valor decimal concreto, lo normal es encontrar valores aproximados pero no exactos como ocurría en la predicción de cancelación, donde los valores eran discretos. Para este caso lo habitual es usar las métricas de regresión: el *error medio absoluto* (MAE) y el *error cuadrático medio* (RMSE).

El RMSE es la raíz cuadrada del error al cuadrado y mide la diferencia entre las predicciones y el valor objetivo y luego promedia esos valores. Es sensible a valores atípicos, por lo que de haberlos el error podría aumentar con facilidad.

El MAE es un promedio de diferencias absolutas entre los valores objetivo y las predicciones. Tiende a ignorar los valores atípicos, por lo que puede omitir errores que podrían ser relevantes.

Cuanto mayores sean estos valores, peor será el modelo, por lo que nos interesa acercar ambos valores lo máximo posible a 0.

Aplicamos la importancia de variable y en la figura 10.28 se observa que a la hora de predecir la tarifa media diaria el tipo de habitación reservada, el mes y la semana de llegada y el hotel son las variables más importantes. La tabla completa se puede visualizar en el anexo A.10.

0	reserved_room_type	0.141388	27	company	0.0031234
1	arrival_date_month	0.0805516	28	is_canceled	0.00300713
2	arrival_date_week_number	0.0798686	29	days_in_waiting_list	0.00281663
3	full_date	0.0746737	30	previous_cancellations	0.00148169
4	hotel	0.0734031	31	previous_bookings_not_canceled	0.00144035
5	arrival_date	0.0687751	32	babies	0.00083533

Figura 10.28. Importancia de variable respecto adr para todo el dataset

En la matriz de correlación (figura 10.29) observamos que la variable adr no tiene una correlación muy elevada con ninguna de las variables.

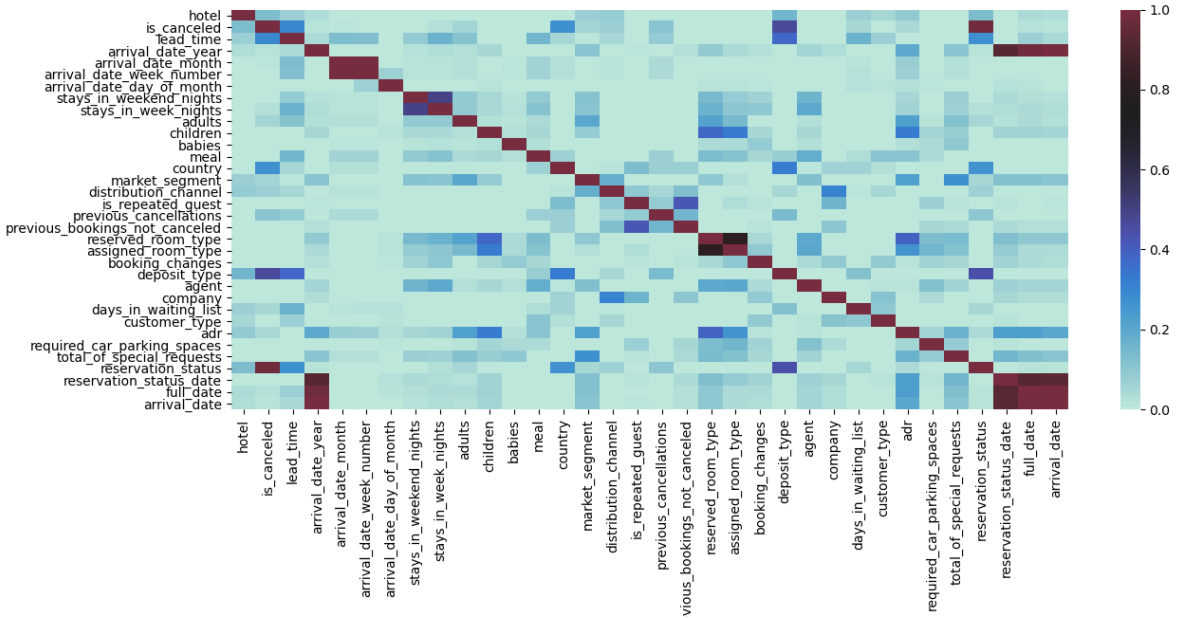


Figura 10.29. Matriz de correlación, hotel resort y hotel urbano

Para obtener el dataset final eliminamos siete variables que tienen poca importancia en relación con *adr: babies, previous_bookings_not_canceled, previous_cancellations, days_in_waiting_list, is_canceled, reservation_status,* y *company.* También eliminamos tres variables que tienen una alta correlación con otras variables del dataset: *reservation_status_date, arrival_date_year* y *arrival_date_day_of_month.*

Tras entrenar los modelos de SVR y Bosques Aleatorios y optimizar sus parámetros, estos son los resultados obtenidos:

```
Mean Absolute Error: 26.3345951006891 67
Mean Squared Error: 38.75355089722858
```

Figura 10.30. Resultados del modelo SVR optimizado para la predicción de adr

```
Mean Absolute Error: 7.197857226334377
Mean Squared Error: 15.646068789056578
```

Figura 10.31. Resultados del modelo Bosques Aleatorios optimizado para la predicción de adr

SVR aún tras ser optimizado nos ofrece un error bastante elevado, cerca del 26% de error absoluto y 39% del error cuadrático. Pero analizando los resultados, si consideramos que una predicción que está 10 euros por encima o por debajo del

valor real es una predicción correcta, tan solo hemos obtenido 7342, por 16438 incorrectas, lo cual corresponde a un pobre 30,87% de acierto.

Por su parte los Bosques Aleatorios nos devuelven un error más bajo, 7,2% el error absoluto y 15,6% el error cuadrático. En este caso hemos obtenido 18565 predicciones aproximadas por 5215 predicciones erradas, lo que supone un 78% de acierto..

Nuevamente los Bosques Aleatorios devuelven un mejor resultado, esta vez para nuestro problema de regresión. Teniendo en cuenta que en este caso no buscamos acertar un número exacto, sino acercarnos a un rango cercano, se podría decir que, al contrario que SVR, los resultados de los Bosques Aleatorios son bastante aceptables.

11. CONCLUSIONES

En este documento hemos aplicado técnicas de Ingeniería de Variables y algoritmos de Machine Learning a un conjunto de datos de reservas hoteleras con el fin de comprenderlo y realizar predicciones de manera exitosa.

De este trabajo pueden extraerse las siguientes conclusiones:

- El Análisis Exploratorio de Datos es de vital importancia para conocer en profundidad los campos y los registros de nuestro conjunto de datos y es estrictamente necesario para evitar que datos nulos o erróneos perjudiquen a las futuras predicciones.

- El Análisis Bivariado es de utilidad para conocer la distribución y los valores de cada variable y permite vislumbrar posibles relaciones entre variables que puedan ser de ayuda a la hora de obtener el conjunto de datos final.

- El Análisis Multivariado permite observar datos con un gran número de campos en dos dimensiones, aunque no es de gran utilidad si su retención de varianza es poca porque pierde mucha información en el proceso.

- La Importancia de Variable y la Matriz de Correlación son imprescindibles de cara a descartar aquellas variables que pueden afectar a la precisión de las predicciones debido a que aporten información demasiado similar (alta correlación) o que aporten información poco relevante (importancia baja).

- Se puede predecir la cancelación de reservas hoteleras con un porcentaje bastante elevado (94%) independientemente del tipo de hotel.

- De los modelos de clasificación seleccionados, tanto SVM como Bosques Aleatorios han ofrecido unos resultados superiores a las expectativas iniciales, mejorandolos en un 9%. Por tanto, podemos concluir que la elección ha sido positiva.

- Es posible predecir la Tasa Media Diaria con un error que oscila entre el 7% y el 15%, dentro de los baremos que habíamos marcado como aceptables en los objetivos.

- De los modelos de regresión seleccionados Bosques Aleatorios ha alcanzado las expectativas iniciales, mientras que SVR no ha alcanzado los objetivos mínimos marcados, quedándose a un 11% del error máximo que se pretendía obtener. No obstante, habría que reintentar la predicción con algunos modelos diferentes antes de concluir si SVR no era una buena opción.

En conclusión, los resultados obtenidos pueden calificarse de muy positivos, en especial en lo que a predicción de la cancelación se refiere y muestran la capacidad que tiene el Machine Learning para ser aplicado al sector turístico. Aunque superar el 90% de precisión es un excelente puntaje, cabe la posibilidad de que aún pueda mejorarse. Algunas formas de hacerlo podrían ser variar el número de variables utilizadas o utilizar otros modelos diferentes como Naive Bayes. No obstante, debido

a la extensión del trabajo, estas posibles mejoras quedaban fuera del alcance del mismo.

Esta ha sido mi primera experiencia en la aplicación real de Machine Learning y el proceso ha supuesto un enorme aprendizaje a nivel personal tanto del tratamiento de datos, como del aprendizaje automático, y por supuesto del lenguaje Python. Por esta razón, puede que aún sea posible optimizar los modelos aquí utilizados y conseguir mejorar los resultados, pero dado que estos son mis primeros pasos considero que los resultados son muy satisfactorios y por ello estoy muy satisfecho con el trabajo realizado.

12. REFERENCIAS BIBLIOGRÁFICAS

Introducción:

World Tourism Organization (2020). *Country Profile - Inbound Tourism*. Recuperado de:https://www.unwto.org/country-profile-inbound-tourism

World Travel & Tourism Council (2020). *Economic Impact Reports.* Recuperado de:

https://wttc.org/Research/Economic-Impact

Peceny, U.S., Urbančič, J., Mokorel, S., Kuralt V., e Ilijaš, T. (2019). Tourism 4.0: Challenges in Marketing a Paradigm Shift. *Consumer Behavior and Marketing*. DOI: 10.5772/intechopen.84762. Disponible en: https://www.intechopen.com/

García López, R. (3 febrero de 2020). Evolución del turismo mundial según la Organización Mundial del Turismo [Entrada en blog]. *Aprende de Turismo*. Recuperado de: https://www.aprendedeturismo.org/

Vázquez, A. (26 de septiembre de 2016). ¿Qué beneficios han traído las nuevas tecnologías al sector turístico? *Hosteltur*. Recuperado de: https://www.hosteltur.com/

García, R. (19 de febrero de 2018). 12 tendencias tecnológicas que transformarán el turismo en 2018 [Entrada en blog]. *Aprende de Turismo*.

Recuperado de: https://www.aprendedeturismo.org/

Ardito, L., Cerchione R., Del Vecchio, P., Raguseo, E. (2019). Big data in smart tourism: challenges, *issues and opportunities*. Current Issues in Tourism. 22(15). 1805-1809. DOI: 10.1080/13683500.2019.1612860. Disponible en: https://www.tandfonline.com/

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets*, 25(3), 179–188. DOI: 10.1007/s12525-015-0196-8. Disponible en: https://link.springer.com

L'Heureux, A., Elyamany, H. F., Capretz, M. A. M., (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*. 5. 7776 - 7797. DOI: 10.1109/ACCESS.2017.2696365. Disponible en: https://ieeexplore.ieee.org/

Sun, S., Wei, Y., Tsui, K.L., Wang, S. (2019). *Tourism Management*. 70. 1-10. DOI: 10.1016/j.tourman.2018.07.010. Disponible en: https://www.sciencedirect.com/

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., Li, X. (2017), Automated Sentiment Analysis in Tourism: Comparison of Approaches. *Journal of Travel Research.* 57(8). DOI: 10.1177/0047287517729757. Disponible en: https://journals.sagepub.com/

Antonio, N., Almeida, A., Nunes, L. (2017). Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).* DOI: 10.1109/ICMLA.2017.00-11. Disponible en: https://ieeexplore.ieee.org/

Objetivos:

Object Management Group (s.f.). *Business Process Model and Notation*. Recuperado el 30 de abril de 2021, de https://www.bpmn.org/

Marco teórico:

Real Academia Española (s.f.). *Diccionario de la Lengua Española*. Recuperado el 12 de enero de 2021, de https://dle.rae.es/turismo

Trésor de la Langue Française (s.f.). *Trésor de la Langue Française informatisé*. Recuperado el 12 de enero de 2021, de http://stella.atilf.fr/

Organización Mundial del Turismo (s.f.) *Glosario de Términos de Turismo*. Recuperado de https://www.unwto.org/es/glosario-terminos-turisticos

Trésor de la Langue Française (s.f.). *Trésor de la Langue Française informatisé*. Recuperado el 12 de enero de 2021, de http://stella.atilf.fr/

Quesada R. (2007) *Elementos del Turismo.* San José, Costa Rica. EUNED. Recuperado de: https://books.google.es/

Leiper, N. (1983). An Etymology of Tourism. *Annals of Tourism Research, 10, 277-281.* Recuperado de https://www.sciencedirect.com/

Brodsky-Porges, E. (1981). The grand tour travel as an educational device 1600–1800. *Annals of Tourism Research,* 8(2), 171-186. Recuperado de: https://www.sciencedirect.com

Guerrero, P. y Ramos, J.R. (2014). *Introducción al Turismo*. México. Grupo Editorial Patria. Recuperado de https://books.google.es/

Acerenza M.A. (1980). Agencias de Viajes: Características Especiales. Estudios Turísticos, 66. 131-249. Recuperado de: https://d1wqtxts1xzle7.cloudfront.net/

Gordon B.M. (2002). El Turismo de Masas: Un Concepto Problemático en la Historia del Siglo XX. *Historia Contemporánea, 25*. 125-156.

Recuperado de: https://addi.ehu.es/

Schwab, K. (2016). *La Cuarta Revolución Industrial*. Penguin Random House. Recuperado de: https://books.google.es/

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets*, 25(3), 179–188. DOI: 10.1007/s12525-015-0196-8. Disponible en: https://link.springer.com

Standing C., Tang-Taye, J.P. y Boyer M. (2014). The Impact of the Internet in Travel and Tourism: A Research Review 2001–2010. *Journal of Travel & Tourism Marketing*, 1. DOI: 10.1080/10548408.2014.861724.

Recuperado de: https://www.tandfonline.com/

Flecha M.D. (2016). El papel de las OTAs en el proceso de distribución de las cadenas hoteleras españolas. *Esic Market Economics and Business Journal, 47(3)*. 479-504. Disponible en: https://www.esic.edu/

Booking.com. (s.f.). Visible body: sobre Booking.com. Amsterdam, Países Bajos. Web Oficial de Booking.com. Recuperado de: https://www.booking.com/

Tripadvisor (s.f.). Visible body: Información sobre Tripadvisor. Web oficial de Tripadvisor. Disponible en: https://tripadvisor.mediaroom.com/

AirBnb. (s.f.) Visible body: Sobre nosotros. Web Oficial de AirBnb. Recuperado de https://news.airbnb.com/es/about-us/

Kounavis C.D., Kasimati A.E., Zamani E.D. (2012). Enhancing the Tourism Experience through Mobile Augmented Reality: Challenges and Prospects. International Journal of Engineering Business Management, 4. DOI: 10.5772/51644. Recuperado de https://journals.sagepub.com/doi/

Kaur K., Kaur, R. (2016). Internet of Things to promote Tourism: An insight into Smart Tourism. *International Journal of Recent Trends in Engineering & Research, 2(4)*. Recuperado de https://www.ijrter.com/

Lexico & Oxford Dictionary (s.f.). *Lexico*.

Disponible en: https://www.lexico.com/definition/artificial_intelligence

National Geographic España (2020). Visible body: Breve historia visual de la inteligencia artificial. Web Oficial National Geographic Channel España. Disponible en https://www.nationalgeographic.com.es/ciencia

Camargo-Vega, J.J., Camargo-Ortega, J.F., Joyanes-Aguilar, L. (2015). Conociendo Big Data. *Facultad de Ingeniería, 24(38)*, 63-77.

Disponible en https://www.redalyc.org/

Ng, A. (2015). Machine Learning. What Is Machine Learning? [Vídeo Online]. Obtenido de https://www.coursera.org/learn/machine-learning/

Estado del Arte:

Cankurt, S. y Subasi, A. (2015). Developing Tourism Demand Forecasting Models Using Machine Learning Techniques with Trend, Seasonal, and Cyclic Components. *Balkan Journal of Electrical & Computer Engineering, 3(1)*. 42-49. Recuperado de: https://www.researchgate.net

Xie, G., Qian, Y. y Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management, 82*. DOI: 2020.104208. Recuperado de: https://www.sciencedirect.com

Claveria, O., Monte, E., Torra, S. (2018). Modelling tourism demand to Spain with machine learning techniques. The impact of forecast horizon on model selection. *Revista de Economía Aplicada, 24 (72)*, 109-132. Recuperado de: https://arxiv.org/

Claveria, O., Monte, E., Torra, S. (2016). Combination forecasts of tourism demand with machine learning models. *Applied Economic Letters, 23(6)*, 428-431. Recuperado de : https://www.tandfonline.com

Nilashi, M., Bagherifard, K., Rahmani, M., Rafe, V. (2017). A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Computers & Industrial Engineering, 109*. 357-368.

Recuperado de: https://www.sciencedirect.com

Wang, Y., Chi-Fai, S., Ngai, G. (2012). Applicability of Demographic Recommender System to Tourist Attractions: A Case Study on Trip Advisor. *IEEE WIC ACM International Conference on Web Intelligence (WI).* DOI: 10.1109/WI-IAT.2012.133. Recuperado de: https://ieeexplore.ieee.org/

Ye, Q., Zhang, Z., Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36(3)*. 6527-6535. Recuperado de: https://www.sciencedirect.com

Torres, S. (2017). Detección de perfiles de usuario en el dominio turístico. Universidad de Jaén. [Trabajo de Fin de Grado].

Recuperado de: http://tauja.ujaen.es/

Antonio, N., Almeida, A., Nunes, L. (2019). Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hospitality Quarterly*. DOI: 10.1177/1938965519851466.

Recuperado de: https://journals.sagepub.com

Falk, M., Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management*.

Recuperado de: https://www.emerald.com

Sánchez-Medina, A.J., Sánchez, E. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management, 89*. Recuperado de: https://www.sciencedirect.com

Alotaibi, E. (2020). Application of Machine Learning in the Hotel Industry: A Critical Review. *Journal of Association of Arab Universities for Tourism and Hospitality, 18(3). 78-96.* Recuperado de: https://jaauth.journals.ekb.eg/

Grogan, M. (2020). Imbalanced Classes: Predicting Hotel Cancellations with Support Vector Machines. *Towards Data Science*. Recuperado de: https://towardsdatascience.com

Metodología:

Tena, M. (2020), Visible Body: ¿Qué es la metodología 'agile'?. *Web oficial BBVA*. Recuperado de: https://www.bbva.com/es

Proyectos Ágiles (s.f.). Visible Body: Qué es SCRUM. Web de Proyectos Ágiles. Recuperado de: https://proyectosagiles.org/que-es-scrum/

Python (s.f.). Visible body: What is Python?. *Web oficial de Python*.

Recuperado de: https://www.python.org/doc

Desarrollo del proyecto:

Antonio, N., Almeida, A., Nunes, L. (2019). Hotel Booking Demand Datasets. *Data in Brief, 22. 41-49*. Recuperado de: https://www.sciencedirect.com/

Bagnato, J.I. (2020). Aprende Machine Learning en Español (1ª ed.). [EPub], La Coruña, España: LeanPub.

Ávila, H.L. (s.f.). Introducción a la Metodología de la Investigación. *Ciencia y Técnica Administrativa*. Recuperado de: http://www.cyta.com.ar/

Chauhan, A. (2015). What Is Variable Importance and How Is It Calculated?. *AI Zone, 17*. Recuperado de: https://dzone.com/

Scikit-Learn (s.f.). Visible body: Extra Trees Classifier. *Documentación de Scikit-Learn*. Recuperado de: https://scikit-learn.org

Ribecca, S. (s.f.). Gráfica de Barras de Conjunto Múltiple [Blog]. *The Data Visualisation Catalogue*. Recuperado de: https://datavizcatalogue.com/

Almela, M. (s.f.) ¿Cómo se interpreta un diagrama de caja y bigotes? [Blog]. *Análisis de Datos*. Recuperado de: https://www.analisisdedatos.org

Ribecca, S. (s.f.). Diagrama de Dispersión [Blog]. *The Data Visualisation Catalogue*. Recuperado de: https://datavizcatalogue.com/

Rodó, P. (2019). Normalización estadística. *Economipedia.com*. Recuperado de: https://economipedia.com/definiciones/normalizacion-estadistica.html

Shaikh, R. (2018). Feature Selection Techniques in Machine Learning with Python. *Towards Data Science Inc*. Recuperado de: https://towardsdatascience.com/

Scikit-Learn (s.f.). Visible body: Random Forest Classifier. *Documentación de Scikit-Learn*. Recuperado de: https://scikit-learn.org

Adankon M., Cheriet, M. (2009). Support Vector Machine. *Encyclopedia of Biometrics*. Springer. Recuperado de:. https://doi.org/10.1007

Sitio Big Data (2018). Aprendizaje Automático y las Métricas de Regresión [Blog]. Sitiobigdata.com. Recuperado de: https://sitiobigdata.com/

13.ANEXOS

13.1. Anexo A: Tablas y figuras

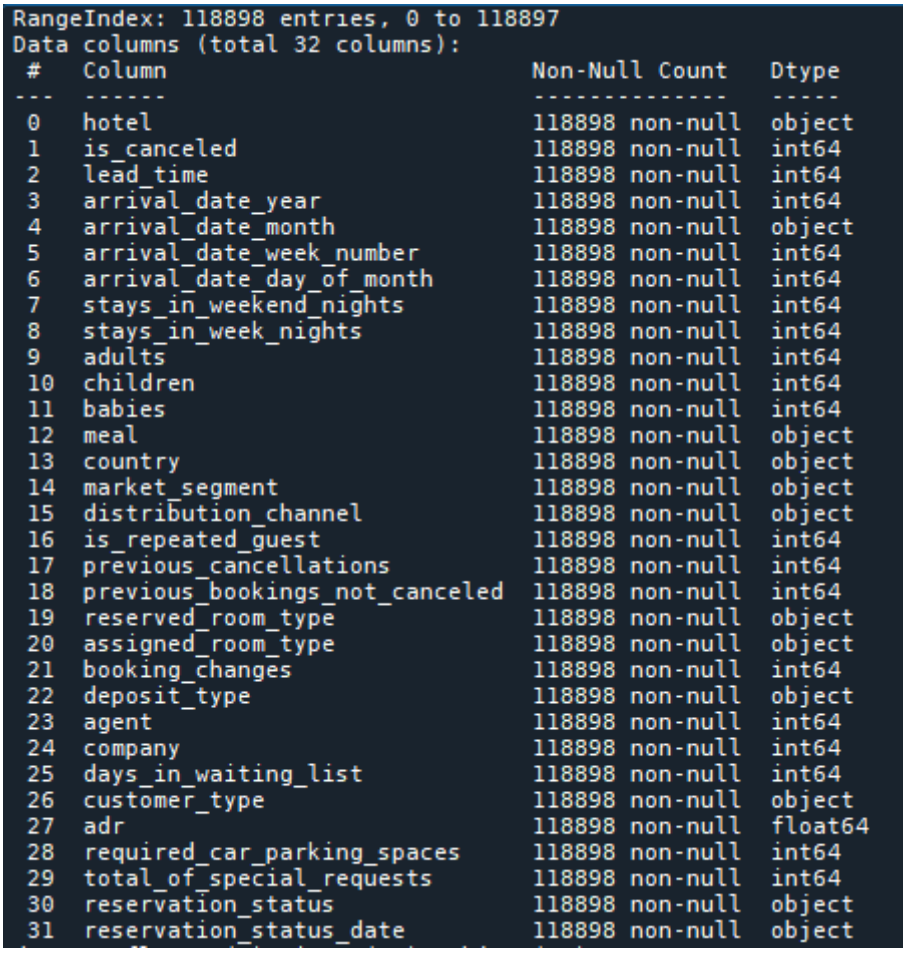
Anexo A.1. Descripción detallada del conjunto de datos

VARIABLE	TIPO DE DATO	DESCRIPCIÓN	VALORES
hotel	Cadena de caracteres	Tipo de hotel	2 valores: 'City Hotel', 'Resort Hotel'
arrival_date_month	Cadena de caracteres	Mes de llegada	12 valores: de 'January' a 'December'
meal	Cadena de caracteres	Tipo de comida reservada	4 valores: ‘SC/Undefined’ (No meal) ‘BB’: Bed & Breakfast ‘HB’: Half Board ‘FB’: Full Board
country	Cadena de caracteres	Nacionalidad	178 valores, incluyendo el valor nulo ‘NULL’. Se representan en formato ISO 3155-3:2013
market_segment	Cadena de caracteres	Segmento de mercado	8 valores: ‘Offline TA’: Agencia viajes ‘Online TA’: OTA ‘Undefined’: Indefinido ‘Groups’: Grupos ‘Direct’: Directo ‘Corporate’: Corporativo ‘Complementary’ ‘Aviation’
distribution_channel	Cadena de caracteres	Canal de distribución	5 valores: ‘Undefined’: indefinido ‘TA/TO’: Agencia de viajes ‘GDS’ ‘Direct’: Directa ‘Corporate’: Corporativo
reserved_room_type	Cadena de caracteres	Código que representa el tipo de habitación reservada	10 valores: ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’, ‘H’, ‘L’, ‘P’
assigned_room_type	Cadena de caracteres	Código que representa el tipo de habitación asignada	12 valores: ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’, ‘H’, ‘I’, ‘K’, ‘L’, ‘P’
deposit_type	Cadena de caracteres	Depósito que pagó el cliente para garantizar su reserva	3 valores: ‘No Deposit’: Sin depósito ‘Non Refund’: No reembolsable ‘Refundable’: Reembolsable
customer_type	Cadena de caracteres	Tipo de reserva	4 valores: ‘Contract’: Tiene un contrato asociado. ‘Group’: Asociada a un grupo. ‘Transient’: No forma parte de un contrato o un grupo, ni tampoco a otra reserva

			transitoria. ‘Transient-Party’: Como transient, pero sí está asociada a otra reserva transitoria.
reservation_status	Cadena de caracteres	Estado de la reserva	3 valores: ‘Check-Out’: El cliente hizo check-in y posteriormente check-out. ‘No-Show’: No hizo check-in e informó de las razones. ‘Canceled’: Reserva cancelada por el cliente.
reservation_status_date	Cadena de caracteres	Fecha en la que se actualizó por última vez el estado de la reserva	926 fechas: desde el ‘2014-10-17’ al ‘2017-09-14’
is_canceled	Entero	Indica si la reserva fue cancelada o no	0: No cancelada 1: Cancelada
is_repeated_guest	Entero	Indica si el cliente había reservado anteriormente o no.	0: Cliente no repite 1: Cliente repite
lead_time	Entero	Días que han pasado desde que se hizo la reserva hasta la llegada del cliente.	Máximo: 737 Mínimo: 0
arrival_date_year	Entero	Año de llegada	‘2015´, ‘2016’, ‘2017’
arrival_date_week_number	Entero	Semana del año de llegada	Entre ‘1’ y ‘53’
arrival_date_day_of_month	Entero	Día del mes de llegada	Entre ‘1’ y ‘31’
stays_in_weekend_nights	Entero	Número de noches de fin de semana en el hotel	Entre ‘0’ y ‘19’
stays_in_week_nights	Entero	Número de noches de entre semana en el hotel	Entre ‘0’ y ‘50’
adults	Entero	Número de adultos	Entre ‘0’ y ‘55’
babies	Entero	Número de bebés	Entre ‘0’ y ‘10’
previous_cancellations	Entero	Cancelaciones de reservas previas del cliente	Entre ‘0’ y ‘26’
previous_bookings_not_canceled	Entero	Reservas previas que no fueron canceladas de este cliente	Entre ‘0’ y ‘72’
booking_changes	Entero	Cambios en la reserva	Entre ‘0’ y ‘21’
days_in_waiting_list	Entero	Días en la lista de espera	Entre ‘0’ y ‘391’
required_car_parking_spaces	Entero	Plaza de aparcamiento solicitadas	Entre ‘0’ y ‘8’
total_of_special_requests	Entero	Número de solicitudes especiales	Entre ‘0’ y ‘5’

children	Decimal	Número de niños	Entre ‘0.0’ y ‘10.0’. También contiene ‘NA’: No disponible
agent	Decimal	Código que indica el agente	Entre ‘1.0’ y ‘535.0’. Contiene valores nulos.
company	Decimal	Código que indica la empresa	Entre ‘10.0’ y ‘543.0’. Incluye valores nulos.
adr	Decimal	Tarifa media diaria	Mínimo: ‘-6.38’ Máximo: ‘5400.0’

Anexo A.2. Conjunto de datos tras el tratamiento de valores nulos



```
RangeIndex: 118898 entries, 0 to 118897
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           118898 non-null  object
 1   is_canceled                     118898 non-null  int64
 2   lead_time                       118898 non-null  int64
 3   arrival_date_year               118898 non-null  int64
 4   arrival_date_month              118898 non-null  object
 5   arrival_date_week_number        118898 non-null  int64
 6   arrival_date_day_of_month       118898 non-null  int64
 7   stays_in_weekend_nights         118898 non-null  int64
 8   stays_in_week_nights            118898 non-null  int64
 9   adults                          118898 non-null  int64
 10  children                        118898 non-null  int64
 11  babies                          118898 non-null  int64
 12  meal                            118898 non-null  object
 13  country                         118898 non-null  object
 14  market_segment                  118898 non-null  object
 15  distribution_channel            118898 non-null  object
 16  is_repeated_guest               118898 non-null  int64
 17  previous_cancellations          118898 non-null  int64
 18  previous_bookings_not_canceled  118898 non-null  int64
 19  reserved_room_type              118898 non-null  object
 20  assigned_room_type              118898 non-null  object
 21  booking_changes                 118898 non-null  int64
 22  deposit_type                    118898 non-null  object
 23  agent                           118898 non-null  int64
 24  company                         118898 non-null  int64
 25  days_in_waiting_list            118898 non-null  int64
 26  customer_type                   118898 non-null  object
 27  adr                             118898 non-null  float64
 28  required_car_parking_spaces     118898 non-null  int64
 29  total_of_special_requests       118898 non-null  int64
 30  reservation_status              118898 non-null  object
 31  reservation_status_date         118898 non-null  object
```

Anexo A.3. Transformación a números enteros: Tabla de asignaciones

VARIABLE	ASIGNACIÓN

country	‘ABW’ (Aruba): 0 ‘AGO’ (Angola): 1 [...] ‘ZMB’ (Zambia): 175 ‘ZWE’ (Zimbabwe): 176
hotel_room_reserved	‘A’: 0, ‘B’: 1, ‘C’: 2, ‘D’: 3, ‘E’: 4, ‘F’: 5, ‘G’: 6, ‘H’: 7, ‘L’: 8, ‘P: 9
hotel_room_assigned	‘A’: 0, ‘B’: 1, ‘C’: 2, ‘D’: 3, ‘E’: 4, ‘F’: 5, ‘G’: 6, ‘H’: 7, ‘I’: 8, ‘K’: 9, ‘L’: 10, ‘P’: 11
hotel	'Resort Hotel': 0, 'City Hotel': 1
arrival_date_month	'January': 1, 'February': 2, 'March': 3, 'April': 4, 'May': 5, 'June': 6, 'July': 7, 'August': 8, 'September': 9, 'October': 10, 'November': 11, 'December': 12
meal	'Undefined': 0, 'SC': 0, 'BB': 1, 'HB': 2, 'FB': 3
market_segment	'Aviation': 0, 'Complementary': 1 'Corporate': 2, 'Direct': 3, 'Groups': 4, 'Online TA': 5, 'Offline TA/TO': 6
distribution_channel	'Undefined': 0, 'Direct': 1, 'TA/TO': 2, 'Corporate': 3, 'GDS': 4
deposit_type	'No Deposit': 0, 'Non Refund': 1, 'Refundable': 2
customer_type	'Contract': 0, 'Group': 1, 'Transient': 2, 'Transient-Party': 3

reservation_status	'Check-Out': 0, 'Canceled': 1, 'No-Show': 2

Anexo A.4. Importancia de Variable respecto a *market_segment*

Index	feature	v_importance
0	distribution_channel	0.179422
1	agent	0.129067
2	customer_type	0.0734242
3	deposit_type	0.0598706
4	lead_time	0.0443011
5	adr	0.0429902
6	total_of_special_requests	0.0415991
7	meal	0.0335258
8	reservation_status_date	0.0332624
9	hotel	0.0331814
10	country	0.0329889
11	arrival_date	0.0288815
12	company	0.0266879
13	full_date	0.025711
14	arrival_date_week_number	0.0217165
15	arrival_date_day_of_month	0.0207452
16	stays_in_week_nights	0.0195523
17	reserved_room_type	0.0180801

18	arrival_date_month	0.0179046
19	stays_in_weekend_nights	0.0176305
20	is_canceled	0.0161392
21	assigned_room_type	0.0126905
22	adults	0.01231
23	arrival_date_year	0.00966105
24	reservation_status	0.00913059
25	days_in_waiting_list	0.00860229
26	is_repeated_guest	0.00831146
27	booking_changes	0.00638946
28	children	0.0048531
29	previous_cancellations	0.00457571
30	required_car_parking_spaces	0.00350874
31	previous_bookings_not_canceled	0.00247434
32	babies	0.000811926

Anexo A.5. Análisis bivariado: Representaciones gráficas

Diagramas de cajas y bigotes. Se han representado de tres formas diferentes:

1. Utilizando todos los datos.

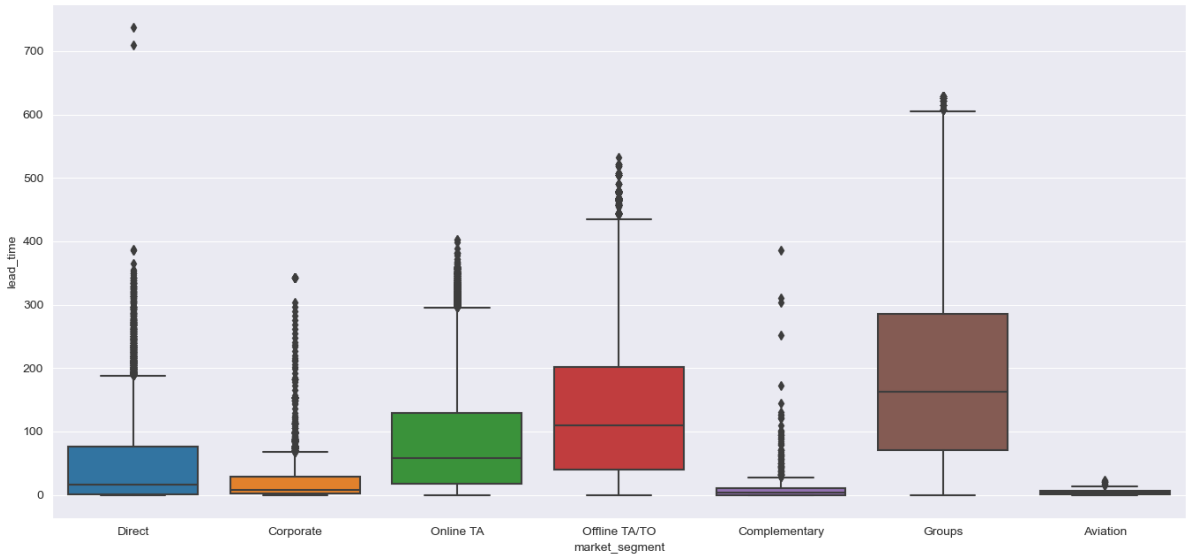
2. Extrayendo los datos excepcionalmente atípicos, que son aquellos que son 1,5 veces mayores o menores que los valores considerados como atípicos (marcados por los bigotes en el boxplot y también llamados outliers).

3. Extrayendo todos estos datos atípicos.

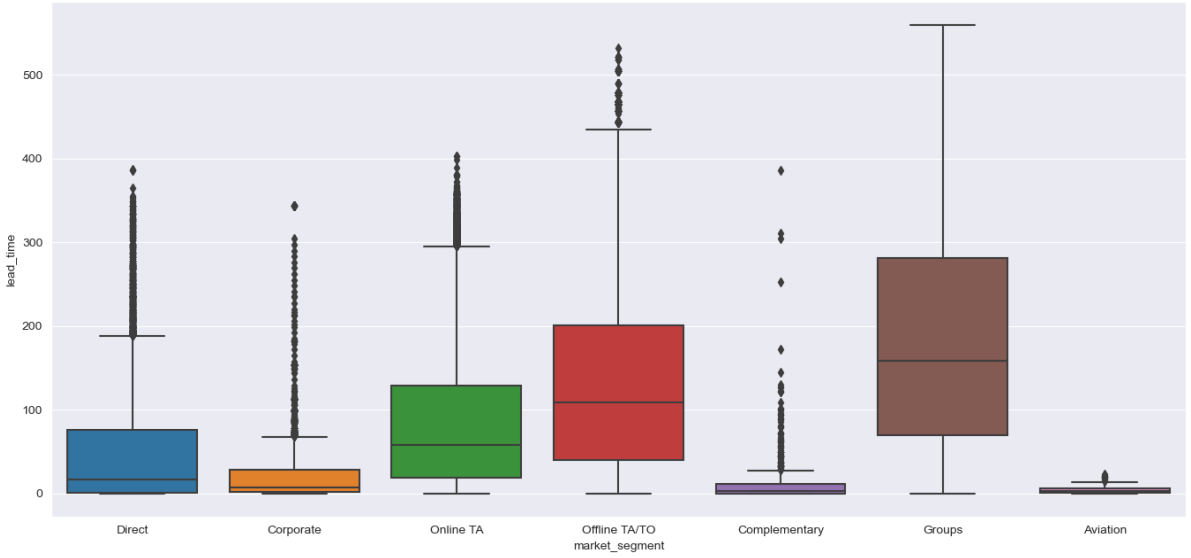
De esta forma podemos ver las gráficas desde diferentes escalas.

Lead_time

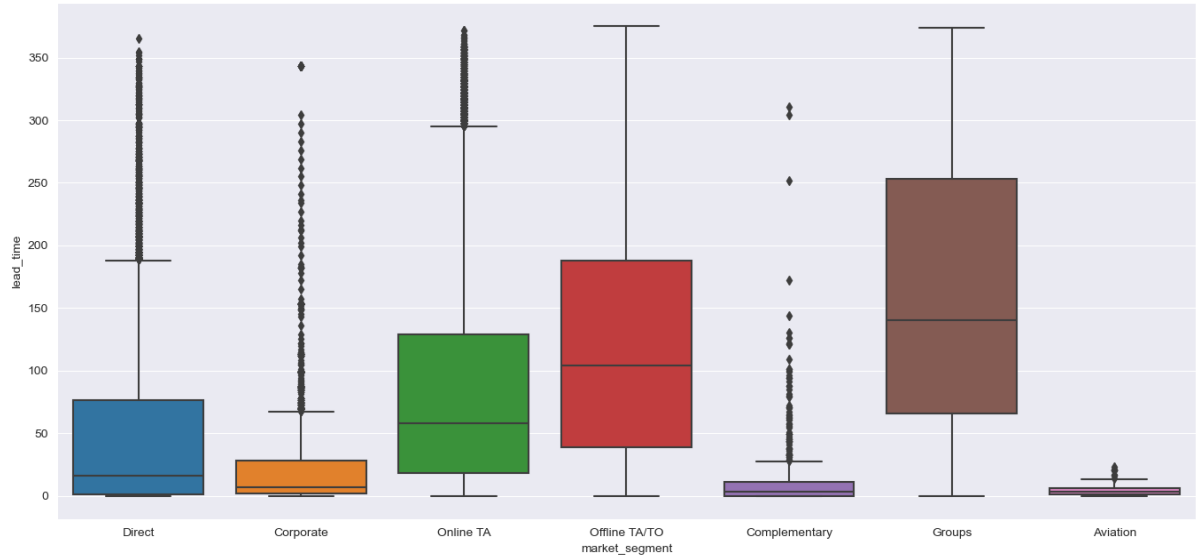
Normal



Sin outliers excepcionalmente atípicos

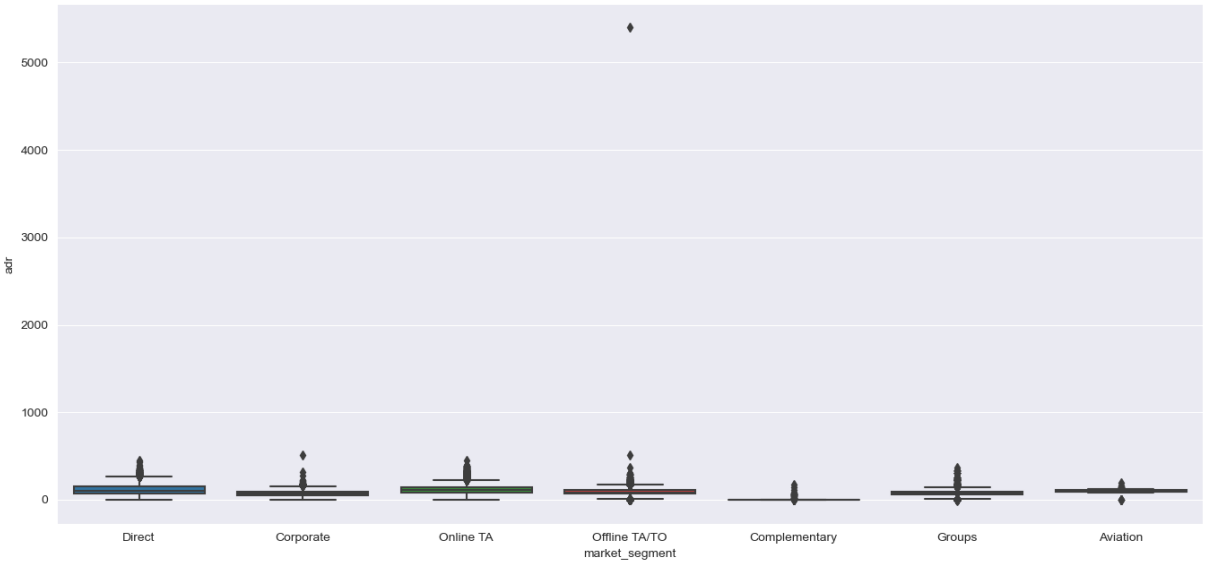


Sin outliers

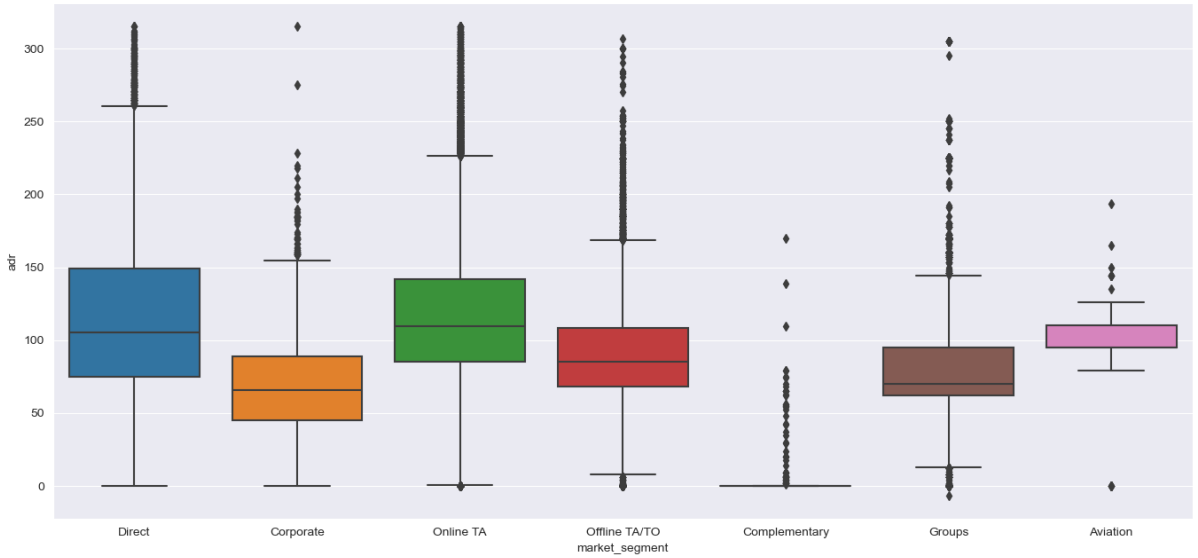


adr

normal



sin outliers excepcionalmente atípicos



sin outliers

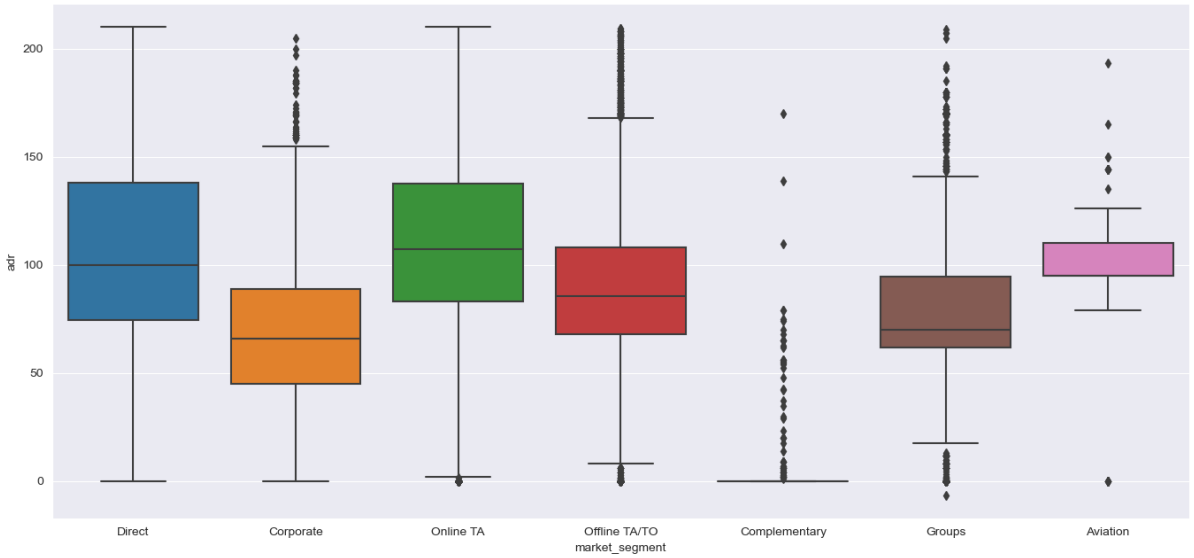
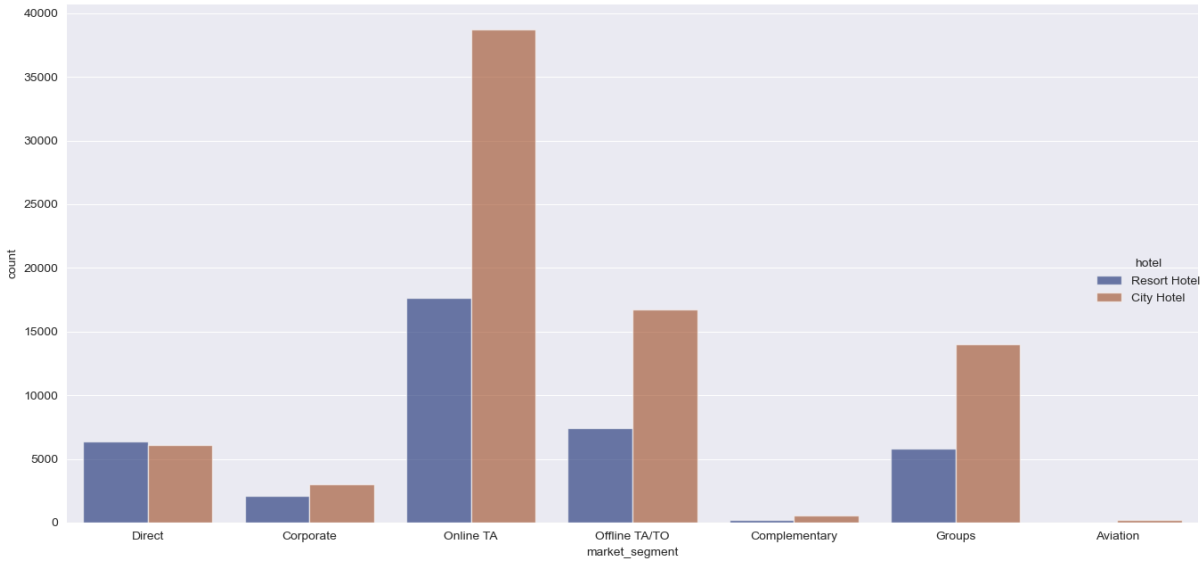
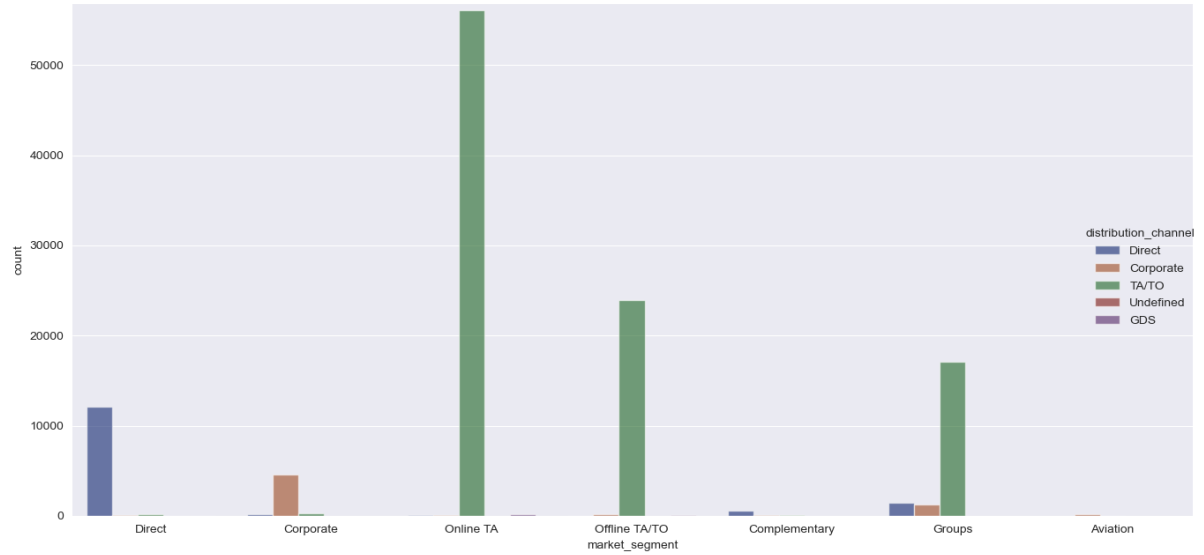


Gráfico de barras de conjunto múltiple. Cada barra representa cada valor de la variable. En el eje horizontal aparecen los segmentos de mercado y en el eje vertical el número total de reservas.

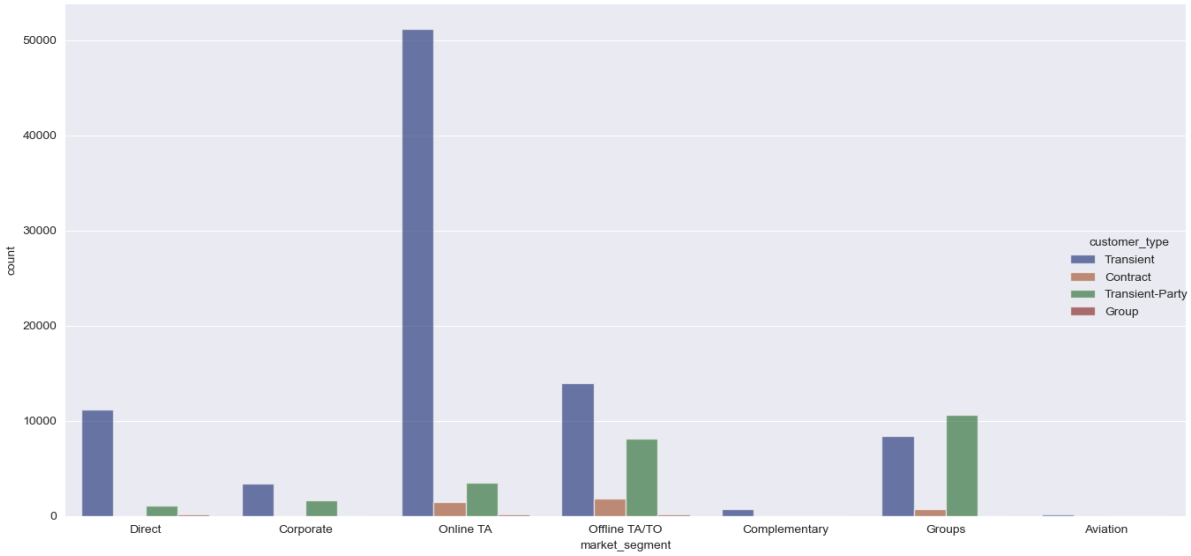
hotel



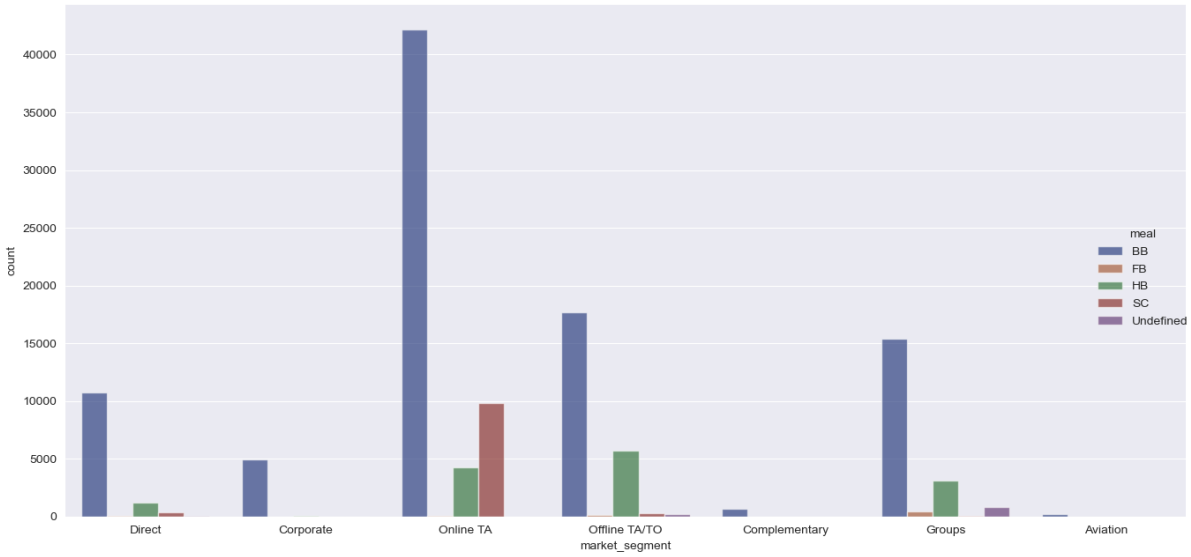
distribution_channel



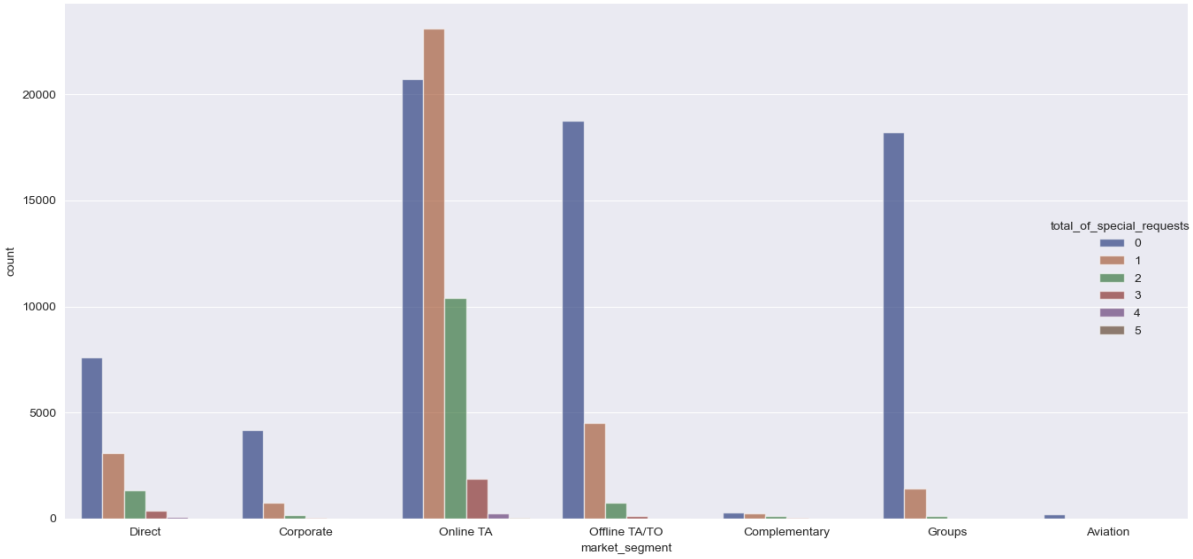
customer_type



meal

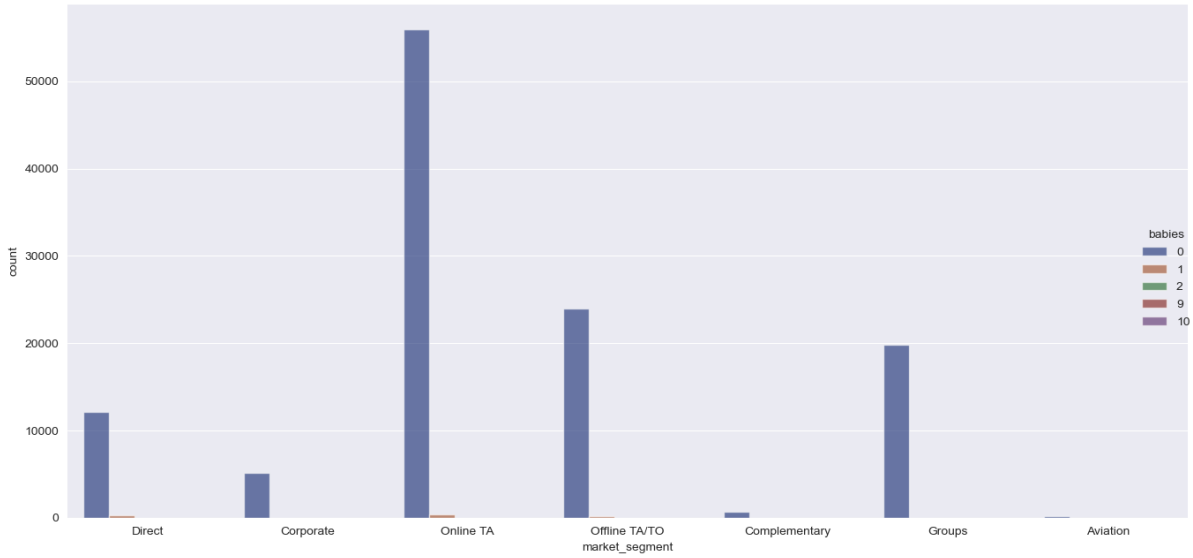


total_of_special_requests



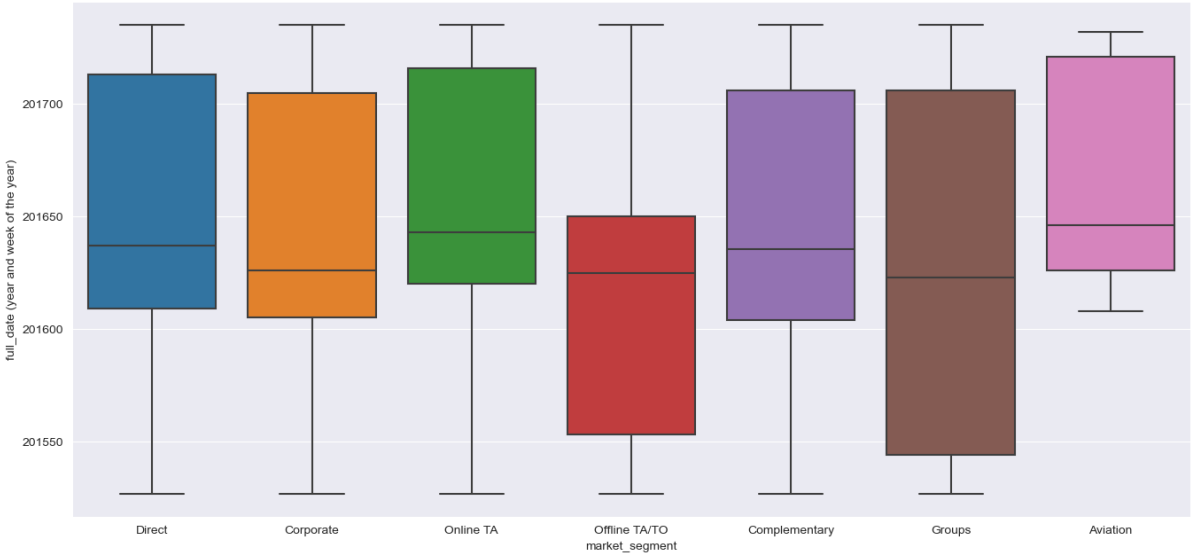
Las variables anteriores eran variables cuya importancia respecto al segmento de mercado era elevada, para ver el contraste también he representado la variable con menor importancia, *babies*, para observar que, efectivamente, no se puede observar información relevante de esta variable.

babies

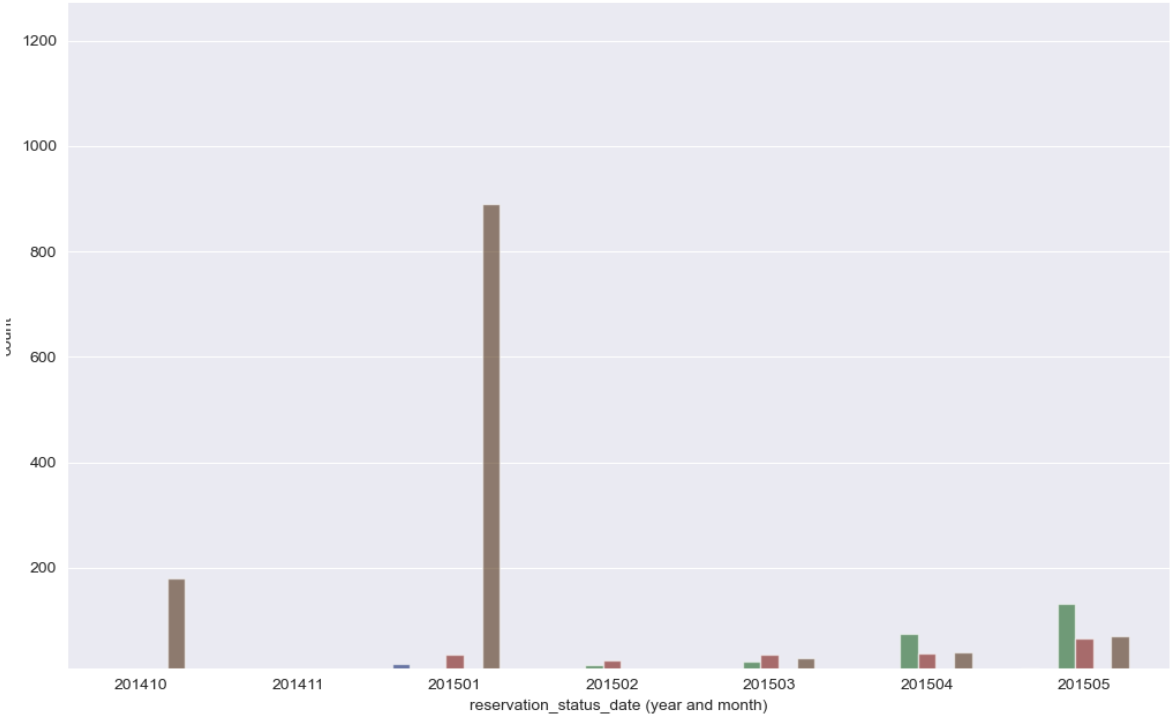


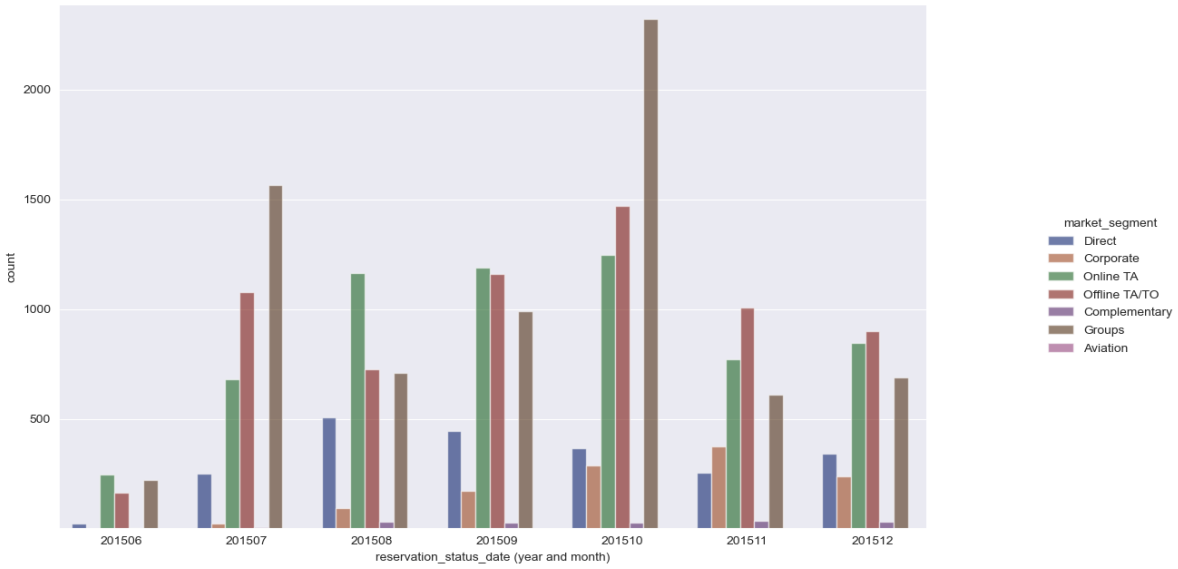
Por último, estas son las representaciones en función de las variables temporales. La mayoría de ellas son diagramas de barras, salvo *full_date*, representada con un boxplot.

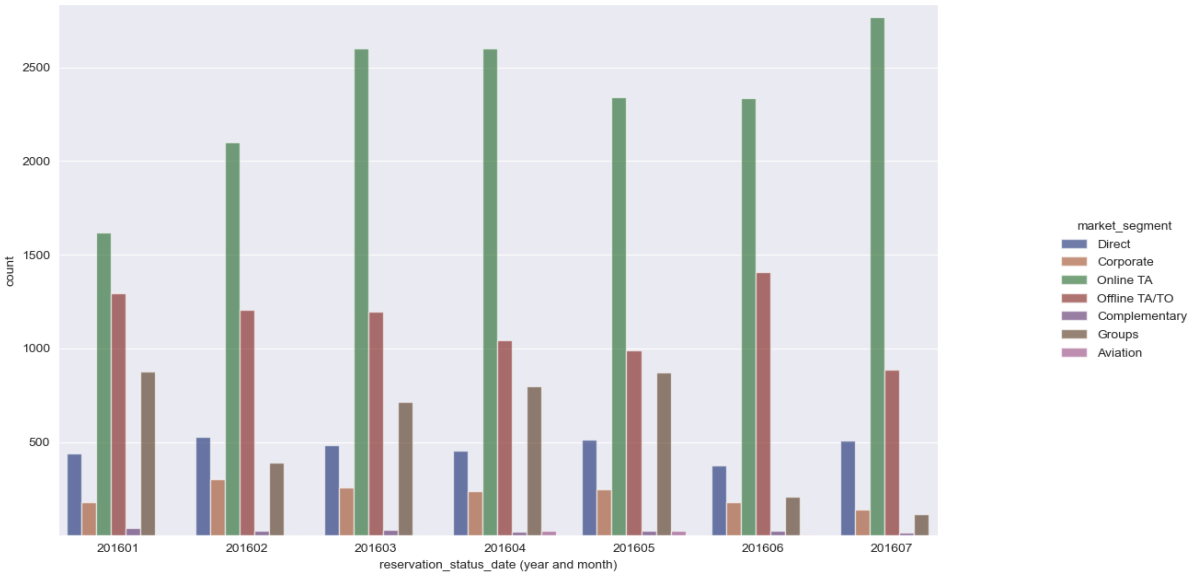
full_date

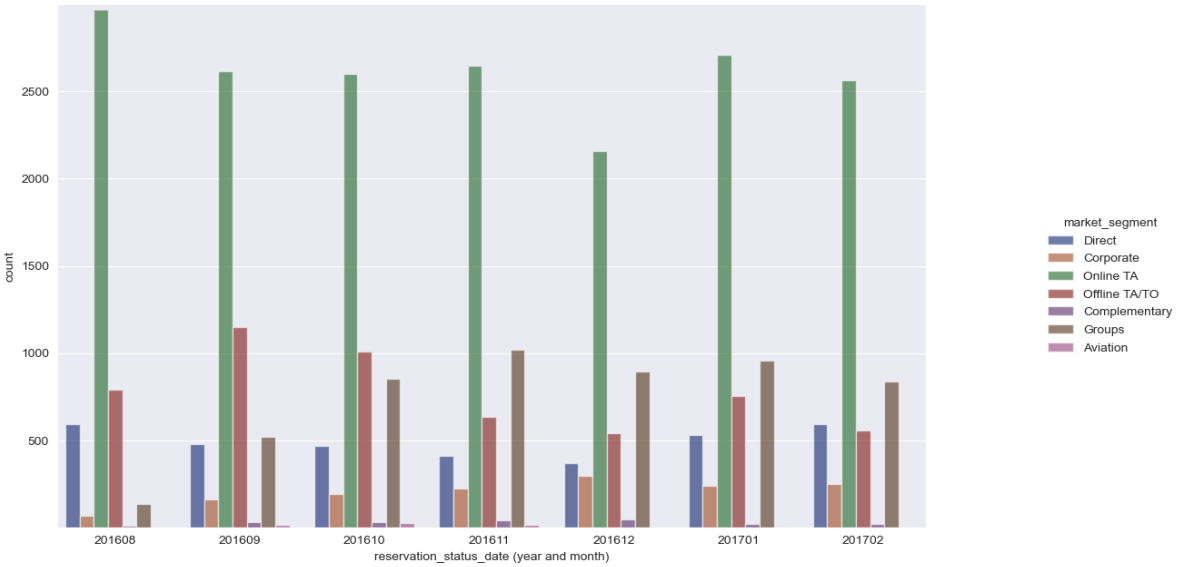


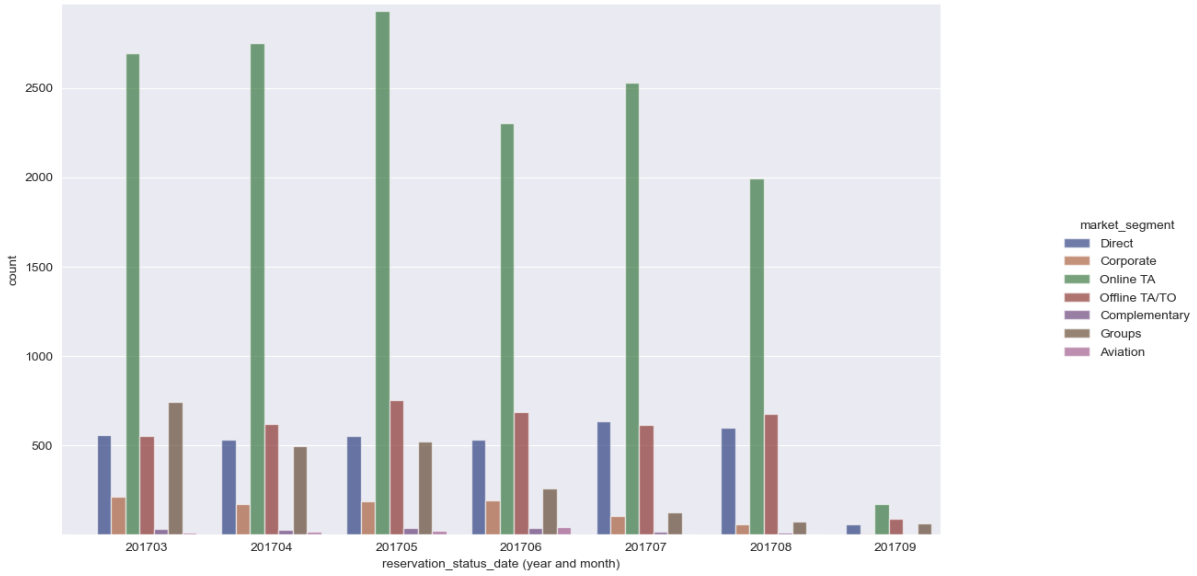
reservation_status_date



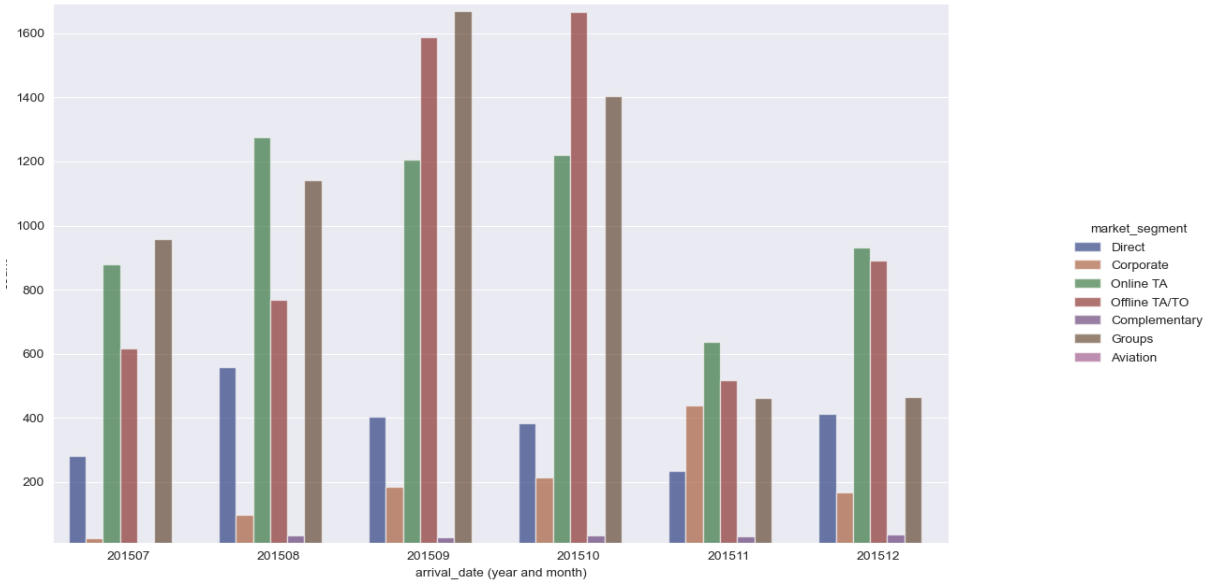

count
2000
1500
1000
500
201506
201507
201508
201509
201510
201511
201512
reservation_status_date (year and month)
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation

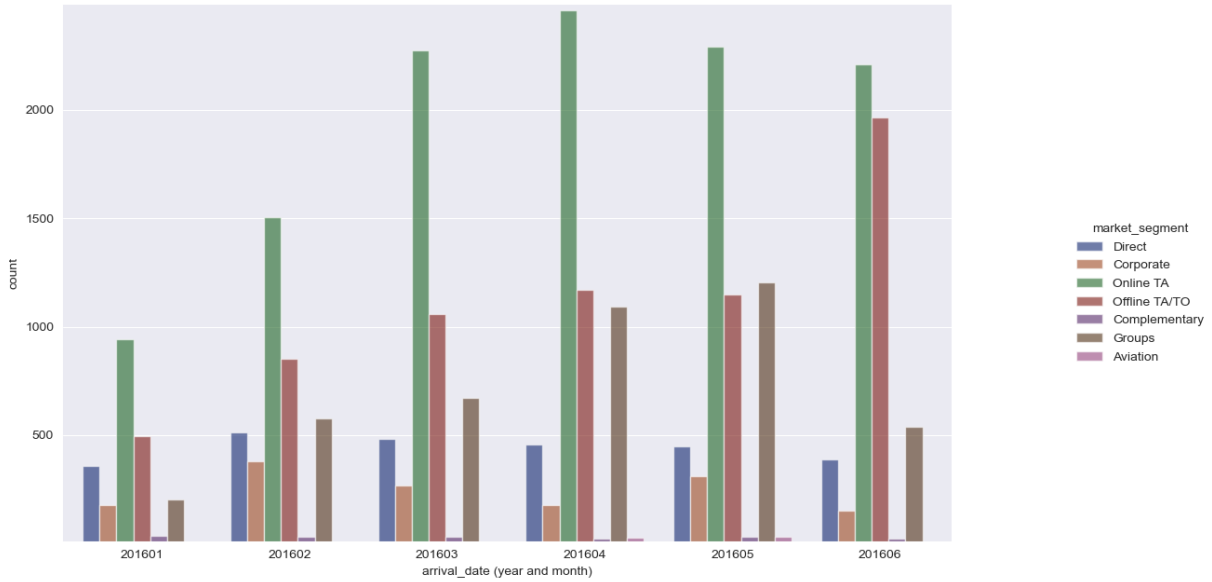

count
2500
2000
1500
1000
500
201601
201602
201603
201604
201605
201606
201607
reservation_status_date (year and month)
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation

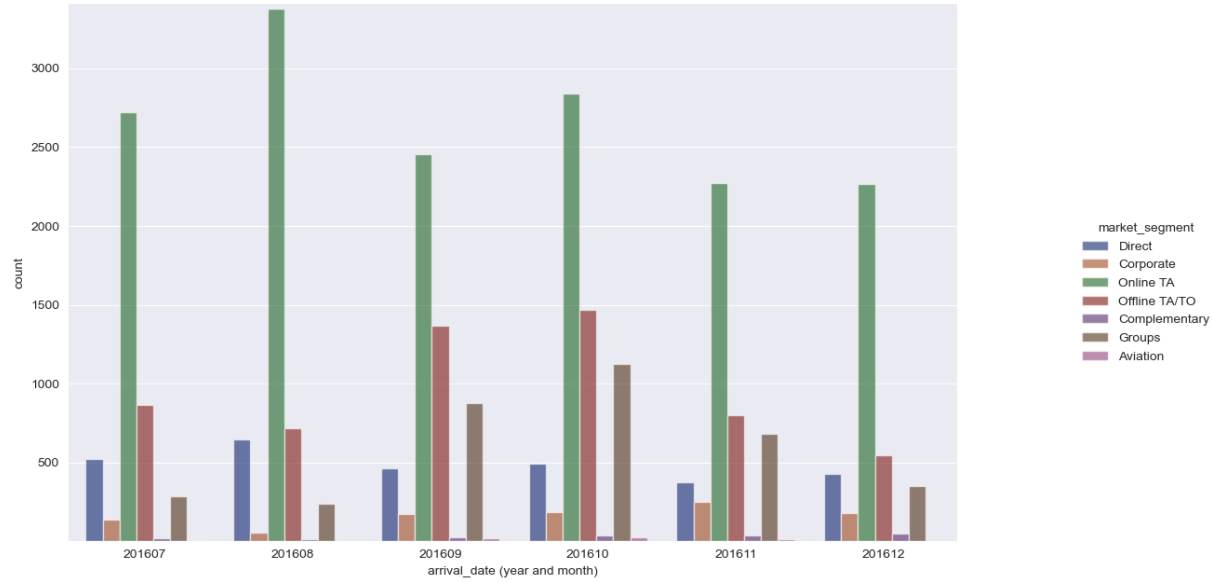

count
2500
2000
1500
1000
500
201608
201609
201610
201611
201612
201701
201702
reservation_status_date (year and month)
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation

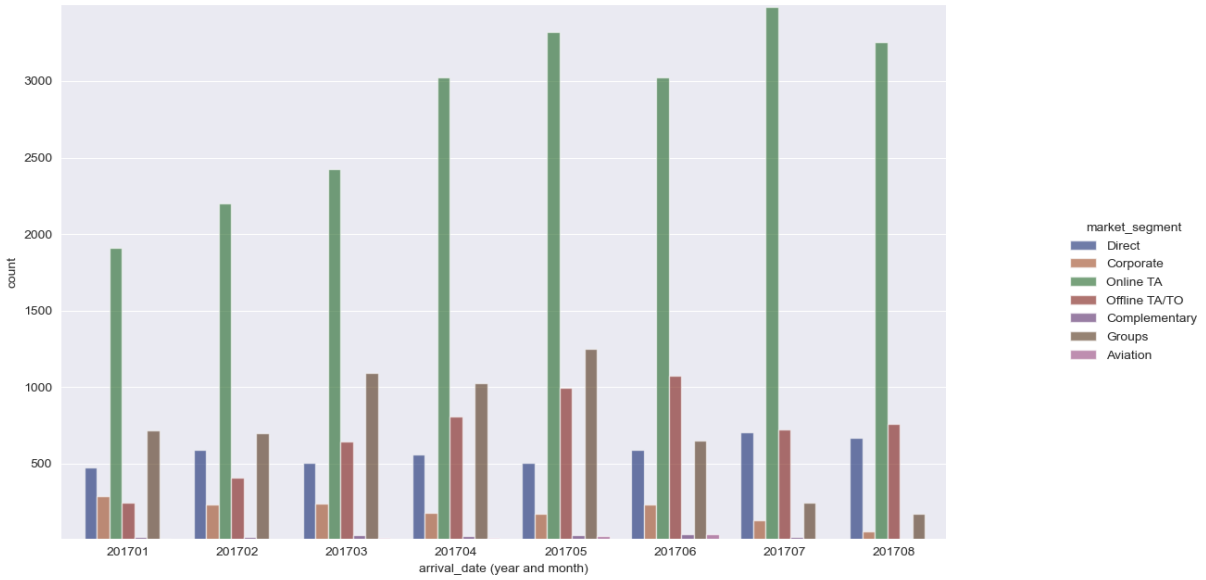


arrival_date

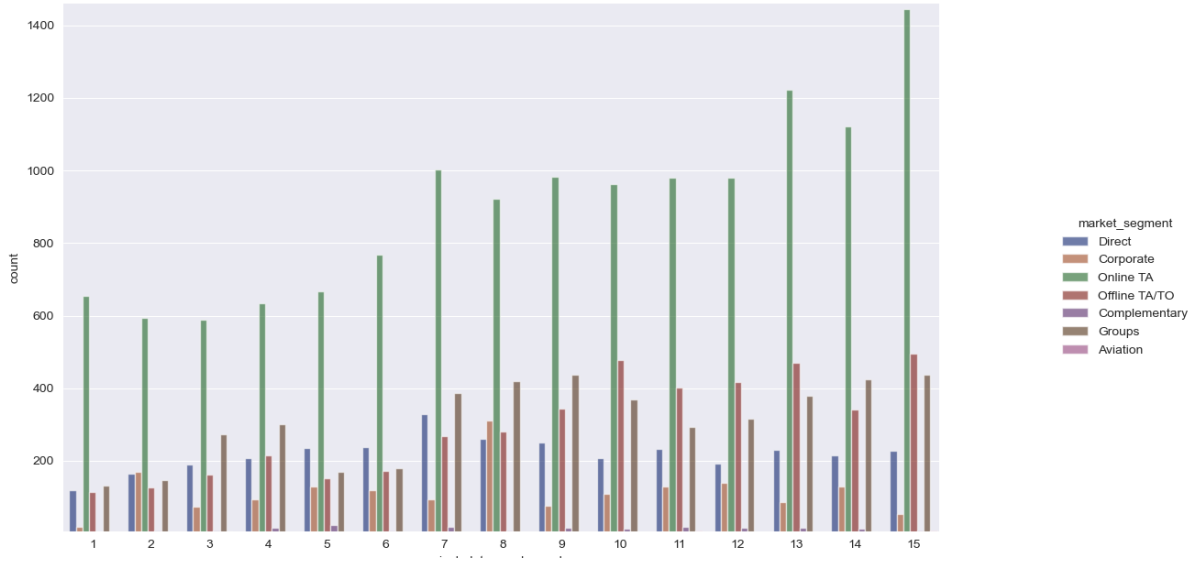


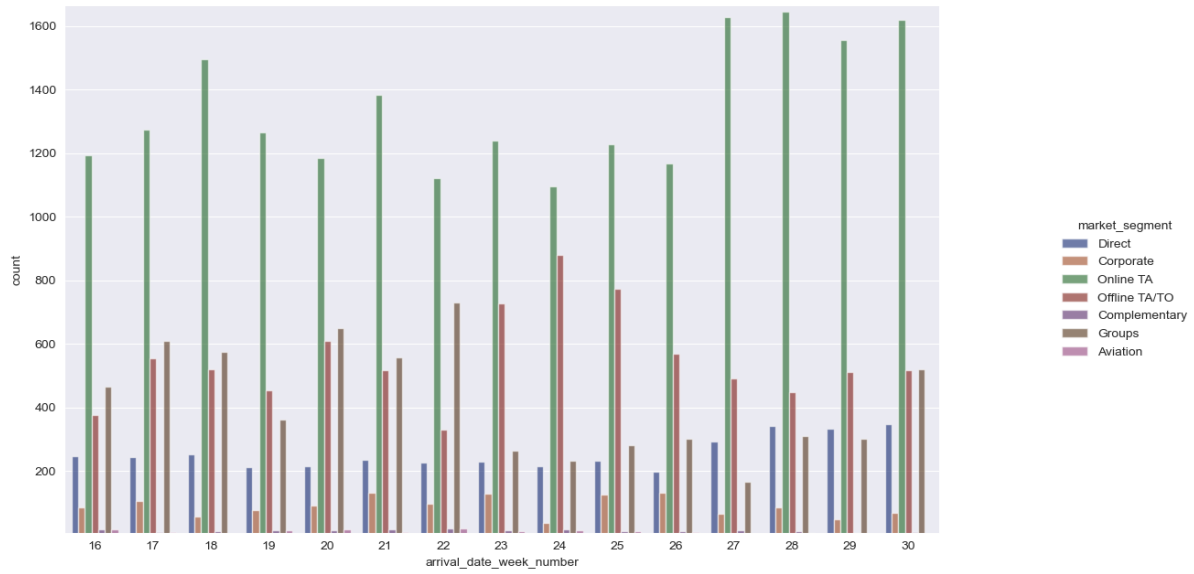

count
2000
1500
1000
500
201601
201602
201603
201604
201605
201606
arrival_date (year and month)
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation

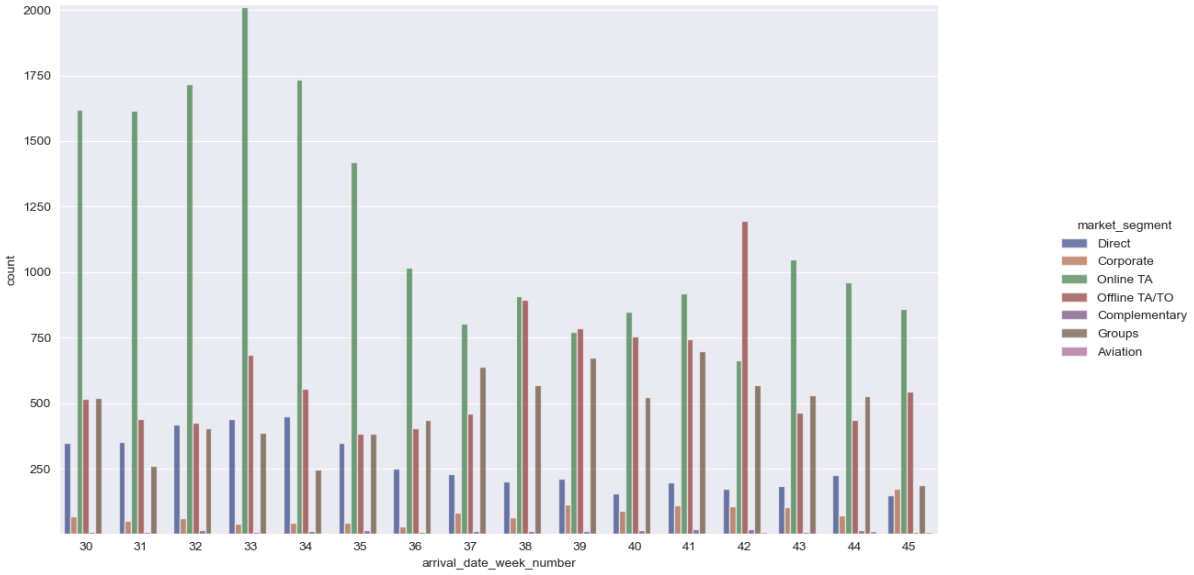

count
3000
2500
2000
1500
1000
500
201607
201608
201609
201610
201611
201612
arrival_date (year and month)
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation

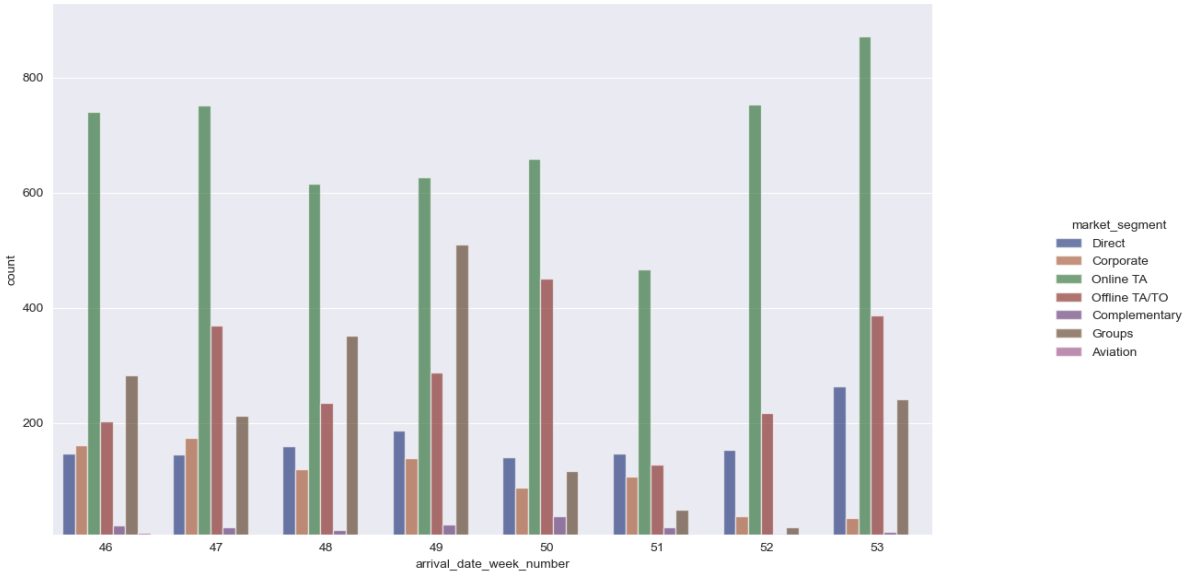


arrival_date_week_number




count
arrival_date_week_number
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation

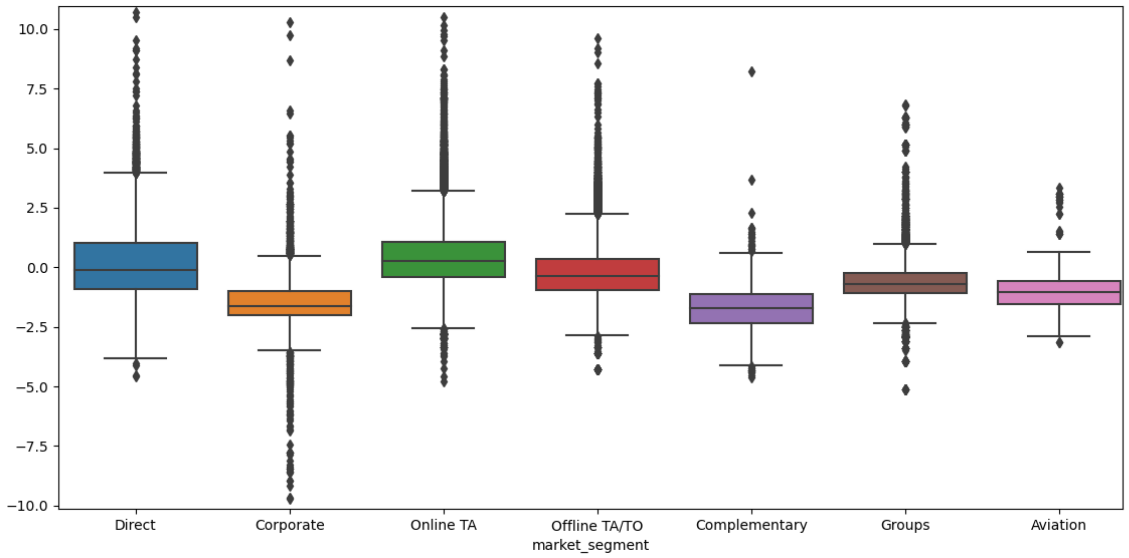

count
arrival_date_week_number
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation


count
arrival_date_week_number
market_segment
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Aviation

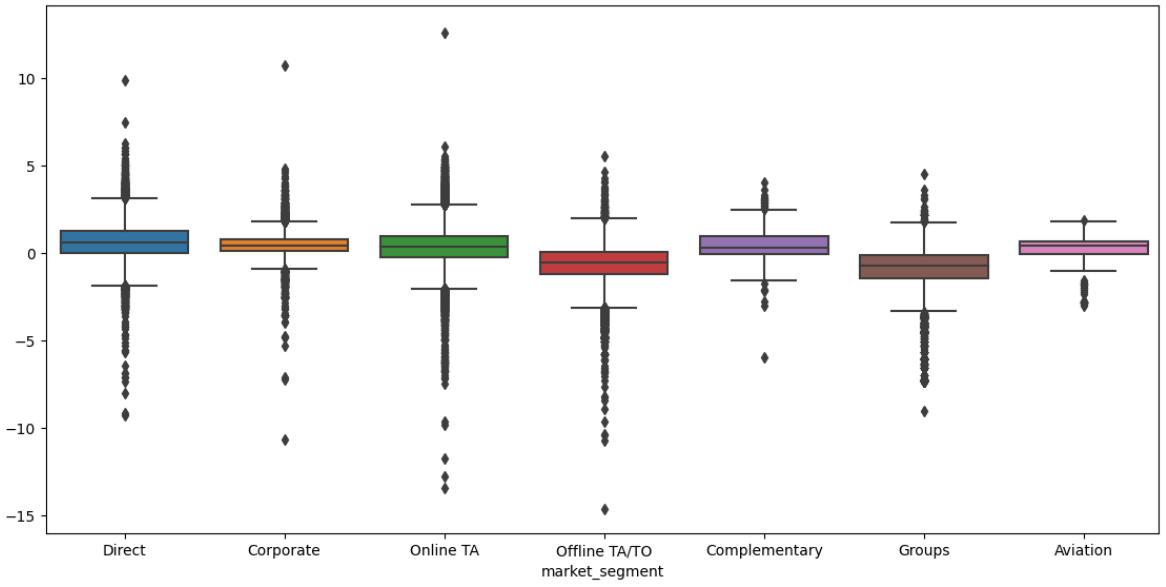
Anexo A.6. Análisis multivariado: Representaciones gráficas

Representa mediante diagramas de caja y bigotes el segmento de mercado respecto de cada una de las 5 componentes principales obtenidas con el PCA (pca1, pca2, pca3, pca4 y pca5) reteniendo para ello un 51% de la varianza.

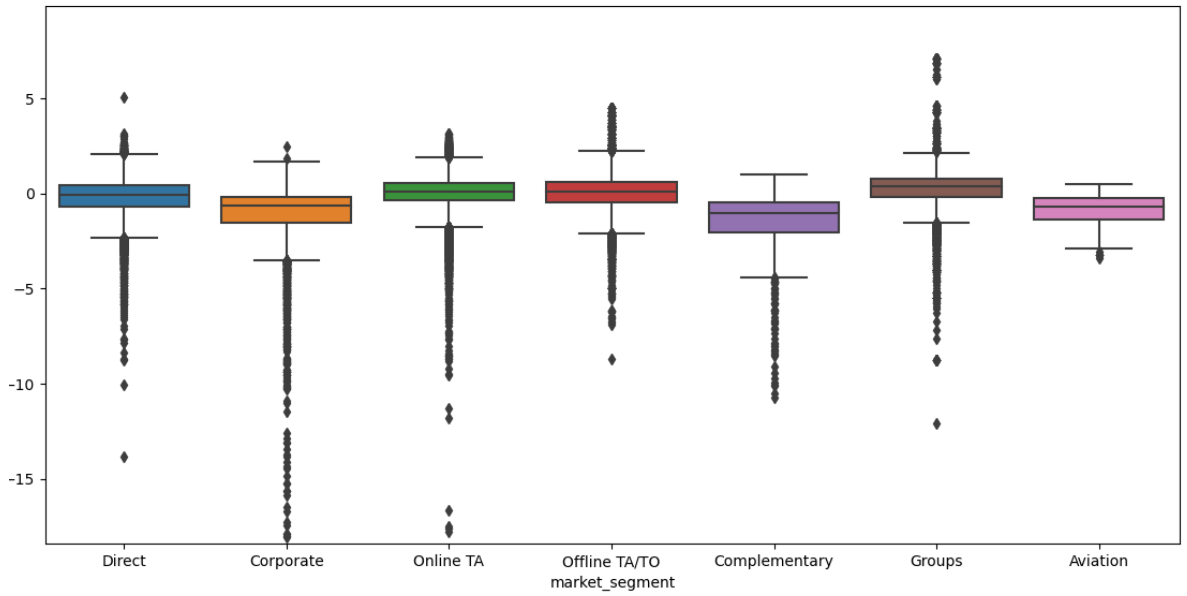
pca1



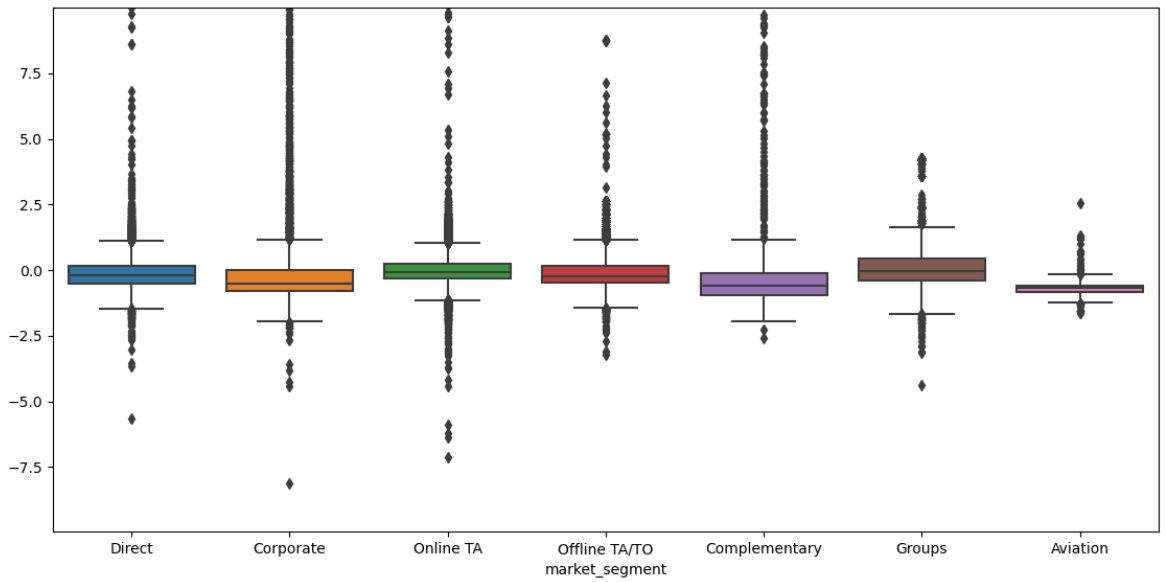
pca2



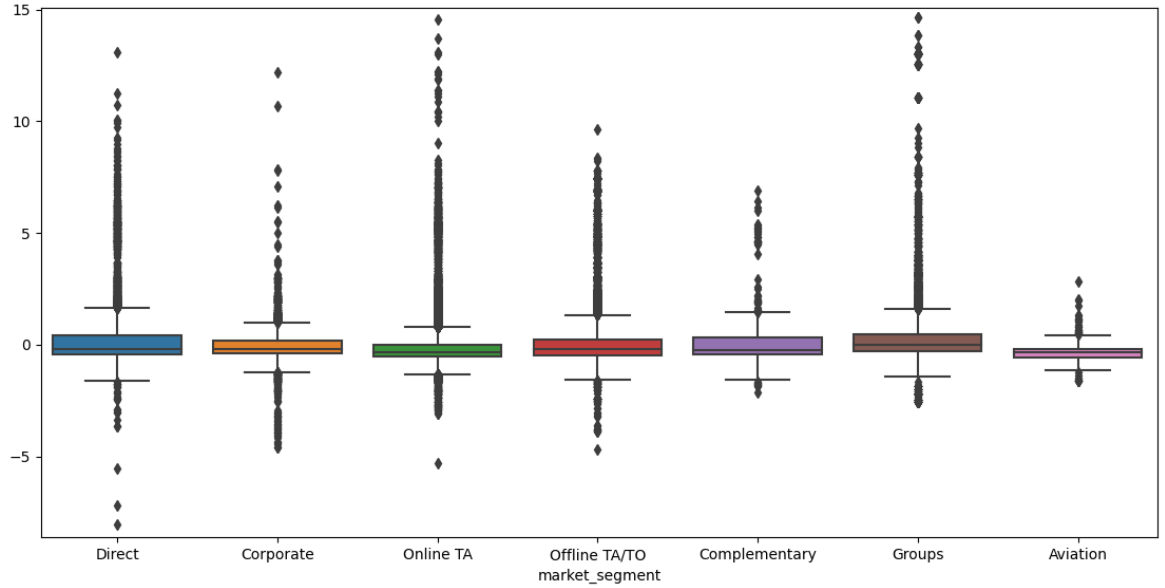
pca3



pca4



pca5



Anexo A.7. Importancia de variable: Predicción de cancelación: Hotel urbano

0	reservation_status_date	0.14665
1	deposit_type	0.13282
2	country	0.0915002
3	lead_time	0.0798705
4	total_of_special_requests	0.060382
5	full_date	0.0389473
6	adr	0.0368419
7	market_segment	0.0351607
8	arrival_date	0.0329495
9	arrival_date_week_number	0.0328876
10	arrival_date_day_of_month	0.0308572
11	stays_in_week_nights	0.0276353
12	arrival_date_month	0.0265708
13	agent	0.0263417
14	previous_cancellations	0.0239833
15	customer_type	0.0239267

16	assigned_room_type	0.0215301
17	stays_in_weekend_nights	0.0213755
18	booking_changes	0.0158916
19	meal	0.0149778
20	adults	0.0149104
21	arrival_date_year	0.0141438
22	distribution_channel	0.0117698
23	reserved_room_type	0.0103515
24	required_car_parking_spaces	0.0085132
25	children	0.00586796
26	is_repeated_guest	0.00427655
27	days_in_waiting_list	0.00352005
28	company	0.00342972
29	previous_bookings_not_canceled	0.00139744
30	babies	0.000719555

Anexo A.8. Importancia de variable: Predicción de cancelación: Hotel resort

0	reservation_status_date	0.143034
1	country	0.101421
2	lead_time	0.0744809
3	deposit_type	0.0564227
4	required_car_parking_spaces	0.0497723
5	full_date	0.0434706
6	adr	0.0418198
7	market_segment	0.0378052
8	arrival_date	0.0350182
9	agent	0.0348838
10	arrival_date_week_number	0.0337773
11	total_of_special_requests	0.0337618
12	arrival_date_day_of_month	0.0321065
13	stays_in_week_nights	0.031104
14	assigned_room_type	0.0289698
15	arrival_date_month	0.0270609
16	customer_type	0.0258746

16	customer_type	0.0258746
17	stays_in_weekend_nights	0.0238454
18	previous_cancellations	0.0201222
19	reserved_room_type	0.0193093
20	meal	0.0181973
21	booking_changes	0.0167717
22	arrival_date_year	0.014477
23	adults	0.0137193
24	distribution_channel	0.0131269
25	children	0.010116
26	is_repeated_guest	0.00726938
27	company	0.00645666
28	previous_bookings_not_canceled	0.00259681
29	babies	0.00200427
30	days_in_waiting_list	0.00120434

Anexo A.9. Importancia de variable: Predicción de cancelación: Hotel urbano y hotel resort

0	reservation_status_date	0.144045
1	deposit_type	0.108553
2	country	0.0907956
3	lead_time	0.077154
4	total_of_special_requests	0.0492539
5	full_date	0.0385076
6	market_segment	0.0383905
7	adr	0.0379141
8	arrival_date	0.0363866
9	arrival_date_week_number	0.0321669
10	arrival_date_day_of_month	0.0304883
11	agent	0.0277688
12	stays_in_week_nights	0.0277335
13	arrival_date_month	0.0266633
14	customer_type	0.024395
15	assigned_room_type	0.0237814
16	required_car_parking_spaces	0.0219639

17	previous_cancellations	0.021851
18	stays_in_weekend_nights	0.0216296
19	booking_changes	0.0157358
20	meal	0.0148606
21	arrival_date_year	0.0144643
22	hotel	0.0139419
23	adults	0.0138226
24	reserved_room_type	0.013212
25	distribution_channel	0.0115195
26	children	0.00675653
27	is_repeated_guest	0.00623503
28	company	0.0046337
29	days_in_waiting_list	0.00270911
30	previous_bookings_not_canceled	0.00160368
31	babies	0.00106378

Anexo A.10. Importancia de variable: Predicción de tarifa media diaria

0	reserved_room_type	0.141388
1	arrival_date_month	0.0805516
2	arrival_date_week_number	0.0798686
3	full_date	0.0746737
4	hotel	0.0734031
5	arrival_date	0.0687751
6	market_segment	0.0557023
7	lead_time	0.0552517
8	agent	0.0397372
9	meal	0.0382429
10	booking_changes	0.0332163
11	children	0.031721
12	adults	0.0294985
13	reservation_status_date	0.0265339
14	arrival_date_day_of_month	0.0260396
15	assigned_room_type	0.0218886
16	stays_in_weekend_nights	0.0159342
17	stays_in_week_nights	0.015892

18	deposit_type	0.0151713
19	distribution_channel	0.0131523
20	customer_type	0.0103879
21	arrival_date_year	0.00884154
22	is_repeated_guest	0.00884073
23	country	0.00803274
24	total_of_special_requests	0.00693826
25	required_car_parking_spaces	0.004102
26	reservation_status	0.00351055
27	company	0.0031234
28	is_canceled	0.00300713
29	days_in_waiting_list	0.00281663
30	previous_cancellations	0.00148169
31	previous_bookings_not_canceled	0.00144035
32	babies	0.00083533

Anexo A.11. Predicción de cancelación: SVM sin optimizar, hotel resort

```
[[4572 1115]
 [ 444 1789]]
              precision    recall  f1-score   support

           0       0.91      0.80      0.85      5687
           1       0.62      0.80      0.70      2233

    accuracy                           0.80      7920
   macro avg       0.76      0.80      0.78      7920
weighted avg       0.83      0.80      0.81      7920
```

Anexo A.12. Predicción de cancelación: Bosques Aleatorios sin optimizar, hotel resort

```
[[5551  136]
 [ 344 1889]]
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      5687
           1       0.93      0.85      0.89      2233

    accuracy                           0.94      7920
   macro avg       0.94      0.91      0.92      7920
weighted avg       0.94      0.94      0.94      7920
```

Anexo A.13. Predicción de cancelación: SVM sin optimizar, dataset completo

```
[[13518  1327]
 [ 2170  6765]]
              precision    recall  f1-score   support

           0       0.86      0.91      0.89     14845
           1       0.84      0.76      0.79      8935

    accuracy                           0.85     23780
   macro avg       0.85      0.83      0.84     23780
weighted avg       0.85      0.85      0.85     23780
```

Anexo A.14. Predicción de cancelación: Bosques Aleatorios sin optimizar, dataset completo

```
[[14487   358]
 [ 1013  7922]]
              precision    recall  f1-score   support

           0       0.93      0.98      0.95     14845
           1       0.96      0.89      0.92      8935

    accuracy                           0.94     23780
   macro avg       0.95      0.93      0.94     23780
weighted avg       0.94      0.94      0.94     23780
```

Anexo A.15. Predicción de tasa media diaria: SVR sin optimizar

```
-------TEST--------
Mean Absolute Error p0: 30.00838454462027
Mean Absolute Error p05: 38.66662879111345
Mean Absolute Error p15: 30.865834499955962
Mean Squared Error p0: 41.94395155891455
Mean Squared Error p05: 48.406366986403356
Mean Squared Error p15: 43.18527633577118
```

En este caso se ha intentado una ligera optimización de parámetros. Para ello, se han tomado 3 valores diferentes para el parámetro epsilon, uno de los parámetros que ayudan a optimizar el modelo SVR. Los valores de epsilon elegidos han sido 0, 0.5 y 1.5. En la solucón optimizada se han elegido automáticamente tanto los mejores valores para el parámetro epsilon como para otros parámetros relevantes en este modelo, como el parámetro C.

Anexo A.16. Predicción de tasa media diaria: Bosques Aleatorios sin optimizar

```
-------TEST--------
Mean Absolute Error: 7.496857957206774
Mean Squared Error: 16.439050761719184
```

13.2. Anexo B: Código y resultados

Tanto el código como los resultados obtenidos se encuentran en el siguiente enlace:

https://github.com/alvgar23/Machine-Learning-for-Tourist-Accommodation