.UBAeconómicas posgrado
ENAP Escuela de Negocios y Administración Pública

2023

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Identificación de patrones en la demanda de turismo interno en la región norte de Argentina.

-Implementación de análisis de Clúster en Python-

AUTORA: ROMINA SILVIA LUCCHETTI

MENTORA: SILVIA VIETRI

[SEPTIEMBRE-2023]

Resumen

La cultura data-driven y los grandes volúmenes de datos presentan múltiples desafíos y oportunidades a las organizaciones turísticas. El análisis de datos y la aplicación de diversas tecnología de información se transforman en un factor que impulsa el crecimiento de las organizaciones . Esto genera la necesidad de adaptarse a nuevos escenarios, integrar tecnologías y adecuar la cultura de gestión organizacional de modo de posibilitar el aprovechamiento de las oportunidades que los nuevos retos plantean.

Dentro del contexto definido en el párrafo anterior, el presente trabajo tiene como objetivo identificar patrones en la demanda de turismo interno de la región norte de Argentina. Los mismos serán utilizados como marco para definir perfiles de consumidores que impulsen el diseño de propuestas de productos turísticos personalizados en los canales comerciales de la agencia de viajes “Turi Nor”.

Para ello se implementarán técnicas multivariantes de análisis de clúster de tipo jerárquico y no jerárquicos. Se utilizará como base de datos la encuesta de viajes y turismo de los hogares (EVyTH), elaborada por La Dirección Nacional de Mercados y Estadísticas a cargo del Ministerio de Turismo y Deportes de la Nación para la región norte del país en el periodo 2022.

Como resultado, se logra identificar 5 clústeres bien definidos respecto a cada una de las variables consideradas . Ello posibilita segmentar la demanda e identificar perfiles de turistas , sobre los cuales se elaboran tres opciones de productos turísticos personalizados en función de los gustos y necesidades de los consumidores.

Palabras clave

Demanda turística. Turismo inteligente. Análisis de Clúster. Segmentación de clientes.

ÍNDICE

Introducción

El turismo es una de las actividades económicas que ha experimentado mayor crecimiento en las últimas décadas. El Consejo Mundial de Viajes y Turismo, a través de su Informe anual de Impacto Económico estima que para el término de 2023 el sector turístico representará el 11,60% de la economía mundial (World Travel & Tourism Council, 2023). Este impulso está directamente relacionado al proceso de globalización y el surgimiento del contexto Big Data, donde los datos adquieren un papel preponderante en la toma de decisiones estratégicas de las organizaciones.

La cultura del data-driven ha alterado las estructuras tradicionales de los negocios turísticos, la ola de innovaciones que impulsa al sector a escala mundial modifica el modo de consumo, influye en la elección de los viajeros y altera las formas de competir en el sector. (Más Ferrando, 2020). Esto genera la necesidad de adaptarse a nuevos escenarios, integrar tecnologías y adecuar la cultura de gestión organizacional de modo de posibilitar el aprovechamiento de las oportunidades que los nuevos desafíos plantean.

Así la información se convierte en uno de los más activos más valiosos en las organizaciones turísticas porque permite predecir el comportamiento del mercado y su satisfacción, permitiendo conocer al cliente y personalizar su experiencia. También encontrar nuevas oportunidades no explotadas, generando ventajas comerciales. Las agencias de viajes que logren entender las necesidades del mercado y traducirlas en propuesta de producto turístico serán las que obtengan su ventaja comercial. Se puede observar que la gestión y capitalización adecuada los datos se convertirá en la clave de su éxito.

De esta manera en el presente trabajo se aplicarán técnicas multivariantes para explorar y analizar datos provenientes de la encuesta de viajes y turismo de los hogares (EvyTH) en un entorno de desarrollo usando el lenguaje Python[1] a fin de integrarlos al funcionamiento y definición de la estrategia comercial de la agencia de turismo “Turi Nor”. Es importante aclarar que dicho nombre es ficticio y los datos referentes a la organización se anonimizan a fin de mantener su confidencialidad.

La agencia de viajes mencionada cuenta con una oficina física donde ofrece atención personalizada a sus clientes prestando servicios de forma tradicional. A la vez a desarrollado una sucursal virtual a través de su página web, en la cual ofrece su producto turístico de manera de adaptarse al nuevo mercado digital, con el objetivo de acercar y facilitar la toma de

[1] Acerca de Python: https://www.python.org/

decisiones al potencial cliente que utiliza ese medio, fomentando la interacción y acompañamiento mediante canales de comunicación complementarios a la web.

Dado la importancia de considerar a la gestión de los datos e información como actividad que agrega valor en las organizaciones, se buscará identificar patrones en la demanda de turismo de la región norte del país que puedan dar lugar a la segmentación y personalización de paquetes turísticos acordes a diferentes necesidades, para generar valor a través del uso de las tecnologías de la información como ventaja competitiva para la agencia de viaje "Turi Nor".

El trayecto del presente desarrollo estará dirigido a identificar las principales características y relaciones entre los turistas internos región Norte de Argentina en el periodo 2022, explorando la encuesta de viajes y turismo de los hogares (EVyTH). Se analizarán características demográficas y preferencias, buscando similitudes y diferencias entre las elecciones de los turistas y sus decisiones de consumo.

En efecto, el presente trabajo tiene como objetivo general identificar patrones en la demanda de turismo interno de la región norte de Argentina para el período 2022. Para ello, se planten los siguientes objetivos específicos:

- Analizar y explorar datos referidos a características y preferencias de los turistas internos del país
- Identificar patrones en la demanda turística
- Definir segmentos de clientes con rasgos distintivos comunes
- Proponer productos turísticos personalizados para satisfacer los requerimientos y necesidades de los consumidores

En un primer apartado se analiza la gestión de datos en contextos de Big Data, destacando sus características y los retos que impone para las organizaciones. Ese ámbito afecta fuertemente al sector turístico, el cual con la integración de los datos generados en diversas fuentes a través de las tecnologías de la información da lugar al nacimiento de un nuevo concepto: el turismo inteligente. Por último, se plantea la situación del sector en Argentina.

En un segundo apartado se describe la base de datos seleccionada y se presenta un análisis descriptivo de la demanda de turismo interno de la región norte de Argentina para el período 2022. En el mismo se aplican procedimiento de limpieza y procesamiento de datos para que los mismo estén presentados de forma que sea posible aplicar técnicas de análisis multivariantes.

En un tercer apartado se desarrolla un análisis de clúster de tipo jerárquicos y no jerárquico. A partir de sus resultados, se identifican segmentos de consumidores con

características diferenciales, las cuales permiten definir propuestas de paquetes turísticos adaptados a sus preferencias y orientados a cubrir sus demandas.

Por último, se extraen y plasman conclusiones sobre todos los conceptos y análisis desarrollados en el trabajo. A continuación, en la *Figura 1.1: Nube de texto* se pueden observar las principales palabras que conforman el contenido del presente trabajo.

Figura 1.1: Nube de texto

Fuente: Elaboración propia con Python

1.Gestión de datos en organizaciones hacia un turismo inteligente

El objetivo de este apartado es analizar la gestión de datos en organizaciones en contextos de Big Data y la importancia y la relación en la aplicación al sector turístico. El mismo inicia explorando el contexto actual de grandes volúmenes de datos en las organizaciones para luego dirigirse al sector de turismo donde se analizan el concepto de turismo inteligente. Por último, propone analizar el sector turístico en Argentina, articulando todos los conceptos antes indicados.

1.1 Gestión de datos en organizaciones en contextos Big Data

Los datos atraviesan todos los circuitos y actividades dentro de las organizaciones. Su captura, procesamiento y análisis permitirán convertirlos en información útil para la toma de decisiones. Las decisiones controladas por los datos tienden a ser mejores decisiones. (McAfee, 2012).

En este marco global la búsqueda de las organizaciones estará orientada a transformar la inmensa cantidad de eventos en datos inteligentes, aquellos que no solo brindan información del pasado, sino que permiten diseñar escenarios futuros y accionar de forma temprana ante ellos. Esto se traducirá en generar ventajas comerciales (McAfee, 2012) e impulsará la creación de valor en contextos de Big Data.

Los grandes volúmenes de datos, macrodatos o Big Data han ido incrementándose de forma acelerada e irrumpiendo en la vida de las personas, en el funcionamiento de las organizaciones, y en la sociedad, en general. En ella los cambios se pueden apreciar en forma transversal desde aspectos que afectan la movilidad, las redes sociales, el aumento de la conectividad y reducción de sus costos, el uso de internet, internet de las cosas o la geolocalización (Aguilar, 2016).

Big Data refiere a grandes conjuntos de datos que tienen tres características principales: volumen, velocidad y variedad, normalmente denominados como las 3 Vs. La definición puede variar según las características de las organizaciones que operen con ellos. Para algunas será más relevante el volumen, ya que están interesadas en capturar los datos, guardarlos, actualizarlos e incorporados a sus procesos. Otras, en cambio, aunque tengan muchos volúmenes de transacciones buscaran velocidad en su procesamiento, es decir, trabajar en tiempo real y a gran celeridad. Otras estarán interesadas en gestionar diferentes tipos de datos. (Aguilar, 2016).

El concepto de grandes volúmenes de datos no solo hace referencia a los problemas relacionados con las 3 Vs que lo caracterizan, sino que también incluye un amplio espectro de

técnicas, tecnologías, métodos y paradigmas no convencionales que apoyan la solución de problemas relacionados con datos de una forma diferente y, generalmente, más adecuada que los métodos tradicionales. (Tabares, 2014). La consultora Gartner define al big data como "Los activos de información de gran volumen, alta velocidad y gran variedad que exigen formas rentables e innovadoras de procesamiento de información que permitan una mejor comprensión, toma de decisiones y automatización de procesos" (Gartner).

En este contexto es un elemento clave la tecnología e infraestructura que es el medio que permite y facilita la búsqueda, almacenamiento y procesamiento de los datos. La estructura tecnológica hace que la información sea más accesible y, por lo tanto, más valiosa. (Shapiro, 1999, p. 8). De esta manera, a los desafíos planteados se adiciona la arquitectura de Big Data la cual debe incorporar las nuevas tecnologías y herramientas de grandes volúmenes de datos y su integración con datos tradiciones. (Aguilar, 2016).

Otro de los importantes desafíos que impone la gestión de grandes volúmenes de datos e información para las organizaciones es su gestión. La consultora Gartner define a la gobernanza de la información como "la especificación de los derechos de decisión y de una estructura de responsabilidades y control, con objeto de fomentar el comportamiento deseable en la valoración, creación, almacenamiento, uso, archivo y eliminación de información".

La gobernanza de la información abarcará múltiples aspectos en la organización referidos a especificaciones de calidad y seguridad, minimizando los riesgos asociados a su uso. La misma incluirá procesos, roles, estándares y medidas que aseguren el uso efectivo y eficiente de la información que permitan a una organización conseguir sus objetivos de negocio. (Gartner, n.d.). Se puede observar que, desde esta perspectiva, la gobernanza de datos se ubicaría en un plano político. Representaría como un paraguas necesario para dar cohesión y gobierno al flujo de información en el marco de los procesos de negocio (García-Morales, 2012).

Estas definiciones contienen el cumplimiento de las legislaciones o normas regulatorias vigentes ya sea internas de la organización o bien en el país o región geográfica en la cual la misma se asiente referidas a la protección de datos, a información de carácter sensible o documentación de tipo confidencial. También abarcará aspectos referidos a la disponibilidad, conservación, integridad o eliminación de la información.

La adopción de Big Data supone desafíos en diversas áreas de las organizaciones e implica mucho más que procesar datos por medio de tecnologías de información. Es necesario efectuar cambios en la estructura, en las estrategias expresadas en diferentes horizontes temporales, así como en recursos físicos y humanos dentro de la organización. La explotación

de los flujos de información en contextos Big Data pueden mejorar radicalmente el desempeño de las organizaciones, pero es necesario cambiar la cultura de la toma de decisiones (McAfee, 2012).

1.2 Turismo inteligente

En contextos de Big Data el concepto de valor de los datos hace referencia a los beneficios extraídos de su uso. El dato en sí no tiene valor, el valor se obtiene cuando es accionable, es decir, sirve para tomar mejores decisiones (Lamelas, 2017). En este entorno resulta fundamental el uso de herramientas y algoritmos de machine learning e inteligencia artificial, que permiten gestionar y potenciar el análisis de datos y su transformación en información.

El turismo esta caracterizado por ser una actividad con fuerte utilización de datos. Es un sector intensivo en información (Benckendorff, 2018) . Cada turista planificará su viaje en función a sus gustos, necesidades, deseos, motivaciones y presupuesto. Cada instancia, dará lugar a un proceso de viaje diferente, el cual es altamente probable que quede reflejado en una huella digital. Gran parte de estos datos ofrecen información sobre las actividades humanas. Los humanos dejamos un rastro digital, de forma voluntaria o involuntaria, cuando realizamos actividades (Puebla, 2018).

En una sociedad altamente digitalizada, los consumidores de turismo son cada vez más exigentes, buscas nuevas emociones y realizan un mayor número de viajes con menor duración. Sus hábitos de consumo se van modificando desde la búsqueda de información, la forma de realizar compras o el modo y destinos de viaje. Los nuevos turistas apuestan por una gran variedad de experiencias creando, de esta forma, nuevas tendencias de consumo. (Muñoz & Sánchez, 2015).

Se pone de manifiesto la importancia de comprender los cambios e implementar acciones orientadas a evolucionar, generando oportunidades de crecimiento. El uso de algoritmos de procesamiento y machine learning, son herramientas que permiten extraer, modelizar y predecir comportamientos en base a los datos relevados. Los métodos más eficientes de recopilación de datos son necesarios para que la industria del turismo trabaje con los datos más fácilmente. (Önder, 2016). Es el punto de partida al concepto de turismo inteligente.

El turismo inteligente, es un turismo apoyado por esfuerzos integrados en un destino para encontrar formas innovadoras de recopilar y aprovechar datos derivados de datos de infraestructura, conexiones sociales, fuentes gubernamentales, organizacionales y

> recursos humanos en combinación con el uso de tecnologías avanzadas para transformar esos datos en experiencias mejoradas y propuestas de valor comercial con un claro enfoque en eficiencia, sostenibilidad y experiencias enriquecidas durante el viaje.(Gretzel, 2015, p. 181)

El concepto de valor que añade el turismo inteligente se asienta en colocar al turista como centro de éste, facilitando la generación de información inteligente orientada a mejorar la experiencia antes, durante y después del viaje, fomentando una mejor interacción e integración con el destino, agilizando la toma de decisiones e incrementando la calidad de la experiencia vacacional y de ocio (Muñoz & Sánchez, 2015).

El sector turismo deberá evolucionar hacia nuevas formas de satisfacer la demanda ofreciendo productos turísticos más flexibles y personalizados que acompañen la experiencia del turista en todo el ciclo de vida del viaje. El mismo inicia antes de la aventura, por medio de internet, el uso de redes y canales de comunicación para obtener información y reservar o comprar los servicios; durante el viaje generando una mejora en la experiencia y contribuyendo a satisfacer las expectativas de los visitantes. Por último, después del viaje, el reto principal de las empresas y los destinos es saber dónde, cómo y quién habla de sus productos y servicios. De esta manera, será posible conocer el grado de satisfacción real de los turistas y poder aplicar sistemas de mejora continua, así como desarrollar nuevos sistemas de fidelización. (Muñoz & Sánchez, 2015)

La aplicación del concepto de turismo inteligente ha sido utilizada en muchos países y ciudades como base para crear y fomentar el uso de nuevas infraestructuras tecnológicas por medio del desarrollo de programas y aplicaciones inteligentes para el usuario final. Permite centrarse en la innovación para enriquecer las experiencias turísticas y aumentar la competitividad y el atractivo de sus destinos (J. Hwang, 2015).

Las diferentes plataformas y herramientas a través de la inteligencia artificial son capaces de obtener un mejor conocimiento de los mercados y las necesidades de los turistas por sus algoritmos para predecir, recomendar y optimizar ingresos (Más Ferrando, 2020). Ahí es donde entran en juego las tecnologías disruptivas que continuarán impactando en la industria del turismo (…) para recrear cualquier tipo de entorno vacacional, detectando y creando nuevas formas de ocio y oportunidades. (Más Ferrando, 2020).

Todo lo planteado, hace notar la importancia de desarrollar y aplicar nuevas tecnologías en vistas de obtener información de valor referentes al área de turismo. Esto hace suponer la

necesidad de armonizar las estadísticas existentes con las nuevas disciplinas. Así se ha plasmado en la Comisión Económica para América Latina y el Caribe (CEPAL), la cual plantea que hay interrogantes que los datos sobre la huella digital no podrían contestar por sí solos, ese tipo de datos no reemplazarán por completo las estadísticas existentes y las tareas de investigación sobre el terreno. Más bien, las complementarán. (CEPAL, N, 2020).

1.3 Sector turístico en Argentina

Durante décadas, el sector turístico ha experimenta un gran crecimiento y expansión a nivel global, sumando año tras años nuevos destinos y diversificación de alternativas en su consumo. "El turismo es un 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". (OMT, 2008).

Este fenómeno de crecimiento del sector turístico comenzó a desarrollarse en la década de los 50 finalizada la segunda guerra mundial, cuando diversos factores políticos, económicos y sociales desembocaron en el fenómeno de la democratización del turismo, impulsado por el consumo de viajes estandarizados, comercializados a gran escala, económicos y con todos los servicios incluidos. (Espelt, 2000).

La Argentina no ha sido una excepción dentro de este fenómeno. Diversas razones dentro de las cuales se puede identificar su basta extensión, la diversidad de climas, culturas y regiones han posicionado al país como un destino turístico muy atractivo y con pujante crecimiento. El Consejo Mundial de Viajes y Turismo, a través de su Informe anual de Impacto Económico estima que para el término de 2023 el sector turístico del país representará el 8,90% de su economía total (World Travel & Tourism Council, 2023).

El desarrollo de la actividad turística suele asociarse con un impacto positivo a nivel económico, impulsando mercados locales como regionales, así como un importante estímulo para el intercambio cultural. Ello implica una relación iterativa, donde el desarrollo de la actividad turística genera crecimiento regional y ese impulso junto con las atracciones ofrecidas en el destino, atraen turistas. A ello se puede sumar la capacidad del sector turístico para superar crisis y el estancamiento de los lugares o impulsar y alcanzar su desarrollo socioeconómico (Almirón, 2008).

Es interesante observar el efecto transversal que tiene la actividad turística sobre el resto de la economía. Su desarrollo implica múltiples circuitos, donde intervienen y se interrelacionan diversos agentes económicos. Esto responde a una característica particular del sector turístico, donde la clasificación de un bien como producto turístico no responde al producto ofrecido, sino que depende de quién lo consume. Por lo tanto, el "producto turístico" se compone de

múltiples bienes y servicios que se destinan a satisfacer las necesidades de los turistas (Fernández, 2017). Esto se debe a la propia naturaleza de la actividad turística, la cual no satisface la demanda de un bien o servicio en particular, sino que se compone de diversos bienes y servicios que el viajero consume mientras se encuentra fuera de su lugar usual de residencia (Sturzenegger & Porto, 2008).

En ese contexto, el sector turístico en Argentina es una actividad dinamizadora de la economía, su evolución no depende exclusivamente de la coyuntura interna, contribuyendo a amortiguar las situaciones de recesión. La misma está configurada como un entramado de diversos agentes económicos que forman circuitos interrelacionados para satisfacer la demanda del producto turístico. En general, las actividades turísticas son intensivas en mano de obra y tienen un efecto positivo sobre el empleo. En países con fuerte heterogeneidad estructural en el desarrollo social de sus regiones, el turismo permite el desarrollo y la creación de empleo en zona periféricas de los circuitos productivos (Oliva & Schejer, 2006).

La estructura del sector empresarial en Argentina está conformada en su mayoría por empresas PyMES de carácter familiar. La Secretaría de la Pequeña y Mediana Empresa y los Emprendedores dependiente del Ministerio de Producción y Trabajo de la Nación las identifica como aquellas micro, pequeña o mediana empresa que realiza en el país sus actividades en alguno de estos sectores: servicios, comercial, industrial, agropecuario, construcción o minero. (Argentina.gob.ar). Son diferenciadas por tramos según sus ventas, personal ocupado y actividades realizadas.

Se calcula que más del 70% de los puestos de trabajo en la argentina son generados por PyMES familiares con un tipo de organización informal donde las tareas y responsabilidades no están asignadas en forma estricta. (Barreto & Azeglio, 2013) . Dado su tamaño los miembros de la familia se encuentran con tareas de dirección y ejecución lo cuales muchas veces carecen de formación en innovación tecnológica y no se adaptan con facilidad a los cambios.

La estacionalidad es otro factor que influye fuertemente en el sector turístico y lo caracteriza. Esta refiere un fenómeno que está asociado a la concentración de la demanda turística, de manera desproporcionada, en ciertos periodos del año. La misma expresa la variación de la demanda a través de las estaciones del año (Corvo et al., 2012). Las causas de la estacionalidad podrían agruparse en naturales e institucionales (Lee et al., 2018).

Las causas naturales de estacionalidad se relacionan con fenómenos físicos e involucran las variaciones temporales de la naturaleza (Lee et al., 2018). Las condiciones climáticas del destino de turismo elegido pueden resultar fundamentales para posibilitar o limitar la realización de actividades específicas como el turismo de sol y playa, el basado en deportes de

invierno o el turismo de salud. Dentro de ellas también podemos resaltar eventos propios de la naturaleza que solo pueden apreciarse en determinadas épocas de año.

Las causas institucionales se identificas con aspectos culturales y sociales, como la época en la cual son establecidas el periodo de receso escolar y laboral, así como las festividades nacionales o religiosas. También es destacable nombrar a la realización de eventos exposiciones, convenciones y congresos destinadas a la difusión e intercambio de información con relación a una actividad productiva específica o a un área del conocimiento científico que suelen repetirse en las mismas fechas en los sucesivos años.

El turismo en Argentina está fuertemente afectado por la estacionalidad natural. Se puede identificar un evento propio de la naturaleza como el ritual de reproducción de la ballena franca cuyo avistaje es fuente de atracción turística por excelencia en las costas del sur del país. También en épocas estivales el turismo de playa y mar toma relevancia en las costas balnearias, así como en el extremo opuesto, en época invernales los centros de esquí y snowboard tienen alta demanda para la práctica de deportes de nieve. A nivel institucional, se pueden encontrar más de 300 fiestas y festivales nacionales distribuidas en los 12 meses del año a lo largo del país (Fiestas Argentinas).

Los desarrollos tecnológicos traen consigo una gran oportunidad de crecimiento para el ámbito del turismo, pero a la vez supone un gran desafío de adaptación de los agentes más tradicionales que lo integran. Se ha podido evidenciar que el sector ha mostrado extrema sensibilidad a las tecnologías digitales, ajustando rápidamente la cadena de valor ante las nuevas tendencias (Sigala, 2018). Pero esa adaptación ha estado protagonizada por agentes distintos a los tradicionales de la industria, son actores innovadores que ganan ventaja al enfocarse en detectar necesidades que no han sido satisfechas y crear nuevos mercados donde no existían (Christensen et. al, 2015).

El éxito del mercado turístico actual pasa por adaptarse a la nueva era de los algoritmos para ser competitivos en los mercados. Sin innovación no hay evolución, pero también es necesaria la disrupción (Más Ferrando, 2020). Es en ese lugar donde toman protagonismo las tecnologías como la inteligencia artificial, el aprendizaje automático, aprendizaje profundo y realidad virtual, las cuales, a través del análisis de los datos, generan valor personalizando la oferta al viajero, detectado nuevas oportunidades y creando nuevas formas de turismo y recreación.

Se concluye que la aplicación de tecnologías de machine learning a la actividad del turismo permite avanzar hacia una idea de turismo inteligente, el cual posibilitaría el mejor aprovechamiento del turismo interno del país. La creación de valor a través de la exploración,

análisis y medición de datos que permita comprender y tener conocimiento más profundo del sector e identificar y gestionar nuevas oportunidades, complementando de forma armónica las estadísticas existentes.

2. Demanda Turismo interno en la región norte de Argentina

Este segundo apartado presenta la base de datos seleccionado para trabajar. Por medio de la realización de tareas de análisis y exploración de datos se buscará investigar, aprender y obtener un adecuado entendimiento del conjunto de datos. Luego se llevan a cabo actividades de limpieza, procesamiento e ingeniería de atributos que permitan obtener un conjunto de datos de calidad adecuado para aplicar métodos de análisis multivariantes de forma posterior.

2.1. Encuesta de viajes y turismo de los hogares (EVyTH)

La base de datos utilizada en el presente trabajo es la encuesta de viajes y turismo de los hogares (EVyTH) elaborada por La Dirección Nacional de Mercados y Estadísticas a cargo del Ministerio de Turismo y Deportes de la Nación para el periodo 2022. Su objetivo es medir la evolución del turismo interno, es decir los viajes realizados por los turistas argentinos dentro de Argentina, sus características y los aspectos sociodemográficos de los mismos. La información es recolectada por medio del llamadas telefónicas o sistema CATI (Computer Assisted Telephone Interview), donde cada hogar es contactado para realizar consultas sobre los viajes turísticos realizados por sus integrantes en los dos meses anteriores al mes de realización de la encuesta y se solicita información sociodemográfica y económica que permite caracterizar a los turistas. Esto incluye datos estructurados por región de destino, región de origen, edad, sexo, quintil de ingresos, tipo de transporte, tipo de alojamiento y pernoctaciones.

La EVyTH se relevó por única vez para el año 2006, luego se realizó en el primer trimestre de 2011 y a partir de inicios de 2012 se estableció como un operativo de carácter continuo. Los resultados se referencian de manera trimestral surgiendo de una muestra de alrededor de cinco mil hogares residentes en las capitales de todas las provincias argentinas y en los aglomerados urbanos cuya población es superior a los cien mil habitantes las que concentran actualmente casi dos tercios de la población total del país. El universo contemplado replica la cobertura de la Encuesta Permanente de Hogares (EPH), relevada por el INDEC, Instituto Nacional de Estadistas y Censos.

La información de la EVyTH es puesta a disposición de forma pública a través del portal de datos abiertos de turismo (https://datos.yvera.gob.ar/). La misma está resguardada bajo

protocolo de seguridad informática y técnicas de anonimizarían, encriptación y gobernanza de datos que siguen los organismos que son parte del Sistema Estadístico Nacional bajo la coordinación del INDEC. Los mismos están resguardados por el secreto estadístico definido en la Ley N° 17622 (Marco legal de las estadísticas oficiales, 1968) la cual establece que la información se publica siempre en compilaciones de conjunto, sin identificar a las personas u organismos respondientes.

2.2. Análisis exploratorio y descriptivo

La base de datos original consta de 5000 registros y 88 variables las cuales abarcan aspectos demográficos y socioeconómicos de los turistas, detalle de la ubicación geográfica tanto de origen como de destino elegido, así como de las actividades realizadas en el mismo. La misma está conformada por los aglomerados pertenecientes a las provincias de Jujuy, Salta, Tucumán, Santiago del Estero, Catamarca y La Rioja para el año 2022. Para iniciar el análisis, se seleccionan solo las observaciones de personas en un rango etario entre 18 a 80 años, cuyo motivo de viaje fue la recreación u ocio con destino la región norte del país y las siguientes 9 variables:

1. Edad: Es una variable cuantitativa expresada en números enteros que especifica la edad en años de los turistas muestreados.
2. Integrantes: Es una variable cuantitativa expresada en números enteros que especifica la cantidad de integrantes del hogar que efectuaron el viaje.
3. Estadía: Es una variable cuantitativa expresada en números enteros que especifica las noches de estadía de los turistas muestreados en el destino elegido.
4. Alojamiento: Es una variable cuantitativa expresada en números enteros del 1 al 10 que califica según la opinión del entrevistado el nivel de importancia que asigna al alojamiento y sus comodidades en la planificación del viaje. En efecto, una puntuación igual a 1 reflejaría que le asigna muy poca importancia y, por el contrario, una calificación igual a 10 implica que lo considera muy importante valorando las características, así como los servicios y comodidades ofrecidas por los establecimientos de alojamiento como puede ser áreas de recreación, servicios en las habitaciones, masajes, spa e instalaciones.
5. Gastronomía: Es una variable cuantitativa expresada en números enteros del 1 al 10 que califica según la opinión del entrevistado el nivel de importancia que asigna a la gastronomía en la planificación del viaje. En efecto, una puntuación igual a 1 reflejaría que le asigna muy poca importancia y, por el contrario, una calificación igual a 10

implica que lo considera muy importante valorando las características de los servicios gastronómicos del destino elegido, la oferta disponible y variedad.

6. Circuito turístico: Es una variable cuantitativa expresada en números enteros del 1 al 10 que califica según la opinión del entrevistado el nivel de importancia que asigna a los circuitos turísticos en la planificación del viaje. En efecto, una puntuación igual a 1 reflejaría que le asigna muy poca importancia y, por el contrario, una calificación igual a 10 implica que lo considera muy importante, convirtiéndose las atracciones de tipo natural, cultural e históricas, así como su accesibilidad, en elementos fundamentales en la elección del destino de viaje.
7. Jurisdicción de origen: Es una variable cualitativa que describe la ubicación geográfica donde residen de forma habitual los turistas que visitaron la región norte de Argentina.
8. Provincia de destino: Es una variable cualitativa que identifica la provincia elegida como destino turístico por los visitantes de la muestra.
9. Localidad de destino: Es una variable cualitativa que identifica la localidad elegida como destino turístico por los visitantes.

Esta selección inicial da lugar a un conjunto de datos de 889 observaciones y 9 variables, que no poseen valores faltantes.

A continuación, se analizan cada una de las variables del conjunto de datos. Para el análisis y manejo de datos se utilizará el módulo pandas y nunpy y para graficar los módulos seaborn y matplotlib.pyplot de Python. En la *Tabla 1: Descripción Estadística* comenzamos por observar los principales indicadores estadísticos para las variables cuantitativas.

Tabla 1: Descripción Estadística

Medida\ Variable	Edad	Integrantes	Estadía	Alojamiento	Gastronomía	Cir. Turístico
Media	55,53	3,21	4,73	8,85	8,33	4,83
Desvío estándar	16,52	1,54	3,41	1,38	2,13	4,02
Moda	70	2	3	10	10	1
Mínimo	18	1	1	1	1	1
Máximo	80	10	21	10	10	10
Rango	62	9	20	9	9	9
25%	45	2	2	8	8	1
Mediana (50%)	60	3	4	9	9	1
75%	69	4	7	10	10	9
Coef. Variación	0,30	0,48	0,72	0,16	0,26	0,83

Fuente: Elaboración propia

1. Edad

En la *Tabla 1. Descripción Estadística* podemos observar que la variable edad cuenta con un rango de 62, un mínimo de 18 y un máximo de 80 años. El promedio de edad de los turistas de la base es de 55 años, donde el 50 % de los entrevistados están comprendidos entre los 45 a 69 años. Presenta un desvío estándar de 16 años y un coeficiente de variación de 0.33, lo que podría sugerir que la variable es heterogénea y los datos se encontrarían relativamente dispersos respecto de su media. En el *Gráfico 1: Histograma edad* se puede observar la distribución de frecuencias de la variable edad donde el valor más repetido o moda es de 70 años. Con frecuencias menores encontramos al promedio o media de 55 años indicada con la línea violeta y la mediana de 60 años resaltada con la línea amarilla . En el *Gráfico 2: Diagrama de caja edad* se ve claramente como los turistas con edades inferiores a 55 (representada por el círculo blanco dentro de la caja), presentan una distribución más dispersa con un bigote más alargado que aquellos con un rango etario superior.

Gráfico 1: Histograma edad

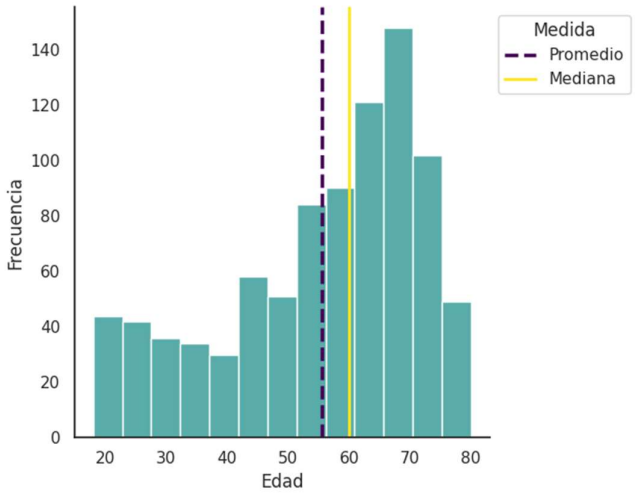
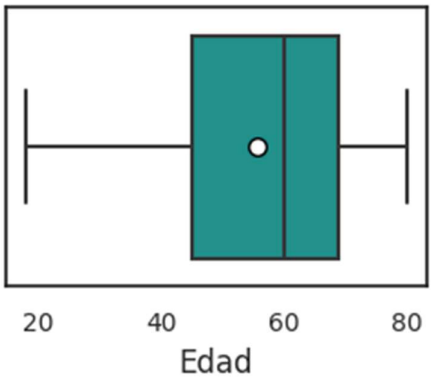


Gráfico 2: Diagrama de caja edad



Fuente: Elaboración propia con Python

2. Integrantes

En la *Tabla 1. Descripción Estadística* podemos observar que la variable presenta un rango de 9, con un mínimo de 1 y un máximo de 10 individuos por hogar. El promedio de integrantes por hogar es de 3, donde el 50 % de todos hogares encuestados están constituidos entre 2 y 4 integrantes. Presenta un desvío estándar de 1 y un coeficiente de variación de 0,48 lo que sugeriría que la variable es relativamente heterogénea y datos se encontrarían dispersos en torno de su media. Se puede observar la distribución de frecuencias de la variable integrantes

en el *Gráfico 3: Histograma integrantes* donde el número de integrantes más repetido es de 2 individuos por hogar. Con una frecuencia menor se encuentra el valor promedio y la mediana indicados con la línea violeta y línea amarilla respectivamente, ambos valores muy próximos entre sí. En el *Gráfico 4: Diagrama de caja integrantes* se puede observar que los valores están más concentrados en una cantidad menor de integrantes, presentando valores atípicos en el extremo superior de la caja, para cantidades mayores a 7 integrantes. El círculo blanco dentro de la caja representa el promedio de 3 integrantes por hogar.

Gráfico3: Histogramas integrantes

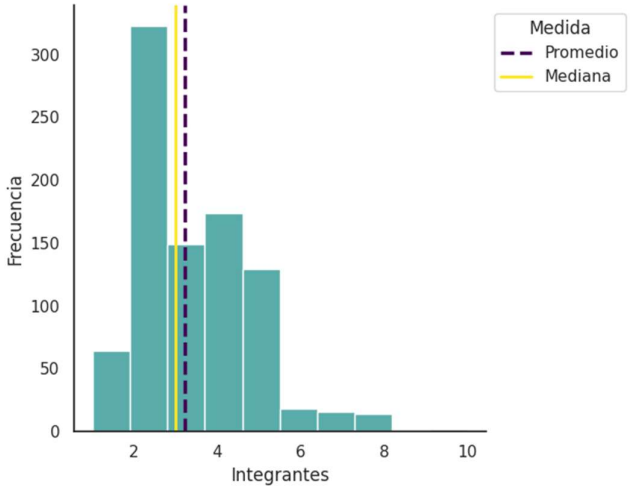
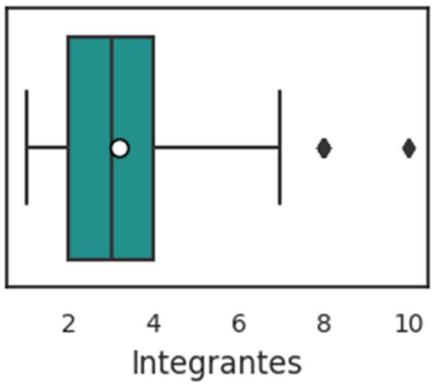


Gráfico 4: Diagrama de caja integrantes



Fuente: Elaboración propia con Python

3. Estadía

En la *Tabla 1. Descripción Estadística* podemos observar que la estadía presenta un rango de 20, con un mínimo de 1 y un máximo de 21 noches de estadía. El promedio de pernoctes es de 4 noches, donde el 50 % de los turistas disfrutan de entre 2 y 7 noches de estadía. Presenta un desvío estándar de 3 noches y un coeficiente de variación de 0,72 lo que sugeriría que la variable es heterogénea , encontrándose los datos dispersos en torno de su media. En el *Gráfico 5: Histograma estadía* se puede ver la distribución de frecuencia, donde los valores están más concentrados en una cantidad inferior de pernoctes. La línea violeta representa el valor promedio y la amarilla la mediana, siendo la estadía más usual elegida por los turistas de 3 noches. En el *Gráfico 6: Diagrama de caja estadía* se puede ver claramente como las estadías superiores presentan una distribución considerablemente más dispersa, encontrando valores atípicos en pernoctes superiores a 14 noches. El círculo blanco dentro de la caja representa el promedio de 4 noches de estadía por hogar.

Gráfico 5: Histograma estadía

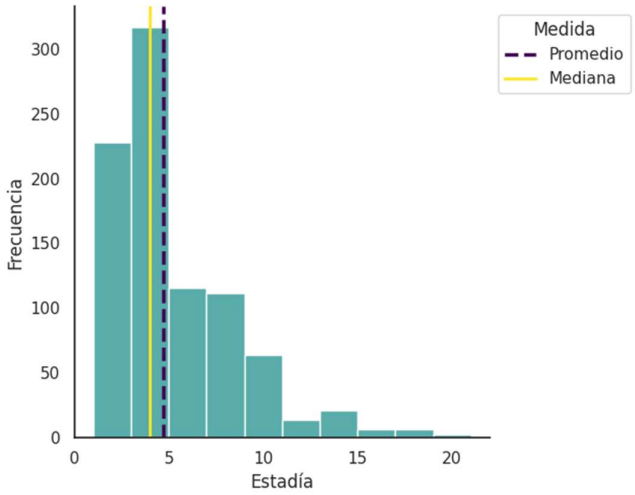
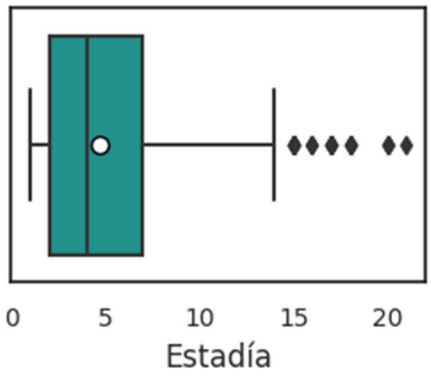


Gráfico 6: Diagrama de caja estadía



Fuente: Elaboración propia con Python

4. Alojamiento

En la *Tabla 1: Descripción Estadística* podemos observar que presenta un rango de 9, con un mínimo de 1 y un máximo de 10 en las respuestas relevadas. El promedio de calificación es de 8,85 puntos, donde el 50 % de los turistas asignaron entre 8 y 10 de puntaje. Presenta un desvío estándar de 1,38 puntos con un coeficiente de variación de 0,16, por lo que los datos se podrían encontrar relativamente agrupados en torno a su media. En el *Gráfico 7: Histograma alojamiento* se puede ver la distribución de frecuencia, donde los valores están más concentrados en las calificaciones más altas. La línea violeta representa el valor promedio y la amarilla la mediana, siendo 10 la calificación más elegida por los turistas.

Gráfico 7: Histograma alojamiento

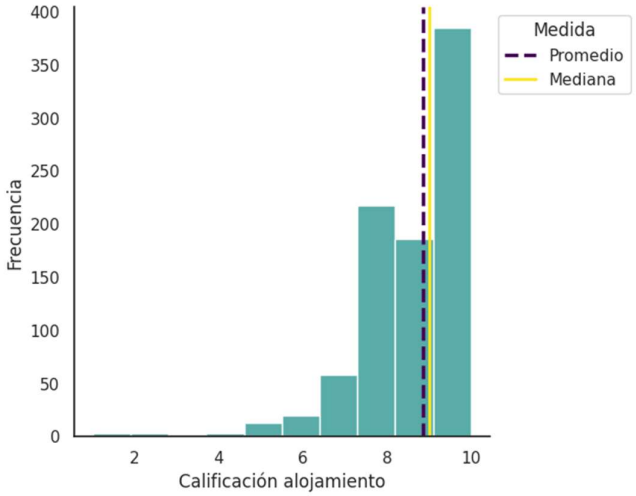
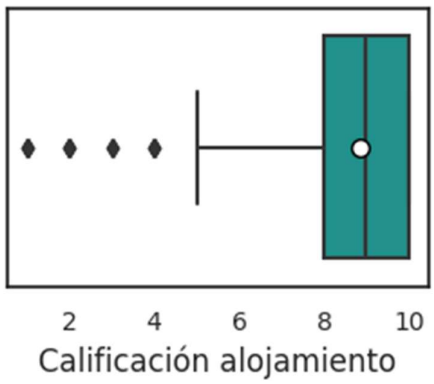


Gráfico 8: Diagrama de caja alojamiento



Fuente: Elaboración propia con Python

En el *Gráfico 8: Diagrama de caja alojamiento* se puede ver claramente como las calificaciones más bajas son relativamente más infrecuentes, encontrando valores atípicos en puntuaciones inferiores a 5. El círculo blanco dentro de la caja representa el promedio de las clasificaciones de los alojamientos.

5. *Gastronomía*

En la *Tabla 1. Descripción Estadística* podemos observar que presenta un rango de 9, con un mínimo de 1 y un máximo de 10 en las respuestas relevadas. El promedio de calificación de toda la base de datos es 8,33 puntos, donde el 50 % de los turistas asignaron entre 8 y 10 de puntaje. Presenta un desvío estándar de 2,13 puntos con un coeficiente de variación de 0,26 por lo que los datos se podrían encontrar relativamente dispersos en torno a la media. En el *Gráfico 9: Histograma gastronomía* se puede ver la distribución de frecuencia, donde los valores están más concentrados en las calificaciones más altas. La línea violeta representa el valor promedio y la amarilla la mediana, siendo 10 la calificación más elegida por los turistas. En el *Gráfico 10: Diagrama de caja alojamiento* se puede ver claramente como las calificaciones más bajas son relativamente más infrecuentes, encontrando valores atípicos en puntuaciones inferiores a 5. El círculo blanco dentro de la caja representa el promedio de las clasificaciones de los servicios de gastronomía.

Gráfico 9: Histograma gastronomía

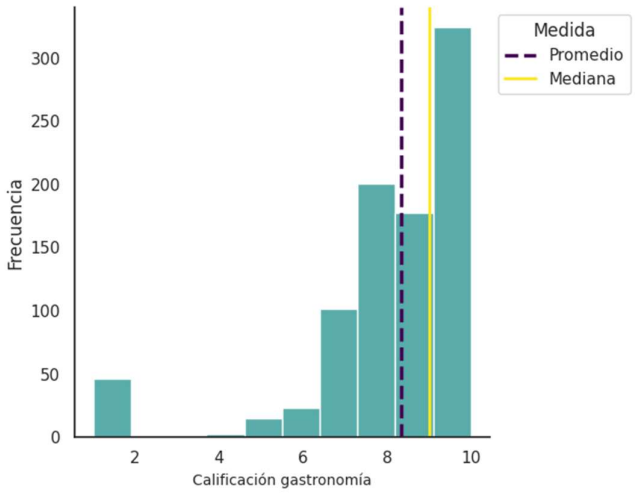
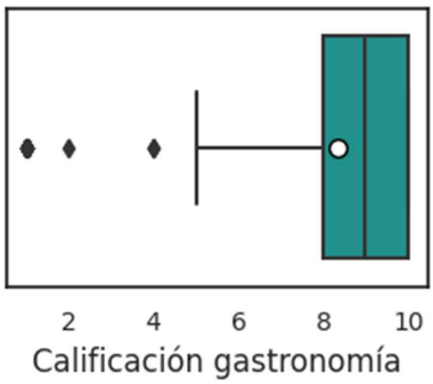


Gráfico 10: Diagrama de caja gastronomía



Fuente: Elaboración propia con Python

6. *Circuitos turísticos*

En la *Tabla 1. Descripción Estadística* podemos observar que presenta un rango de 9, con un mínimo de 1 y un máximo de 10 en las respuestas relevadas. El promedio de calificación de toda la base de datos es 4,83 puntos, donde el 50 % de los turistas asignaron entre 1 y 9 de

puntaje. Presenta un desvío estándar de 4 puntos con un coeficiente de variación de 0,83 por lo que los datos se podrían encontrar relativamente dispersos en torno a la media. En el *Gráfico 11: Histograma circuito turístico* se puede ver la distribución de frecuencia, donde los valores están distribuidos de forma muy dispar a lo largo del rango de la variable. En efecto, están concentrados en la calificación más baja, pero presentes y más distribuidos en las más altas. La línea violeta representa el valor promedio y la amarilla la mediana, siendo 1 la calificación más elegida por los turistas. En el *Gráfico 12: Diagrama de caja circuito turístico* se puede ver claramente, la dispersión de los valores, no encontrando valores atípicos. El círculo blanco dentro de la caja representa el promedio de las clasificaciones.

Gráfico 11: Histograma circuito turístico

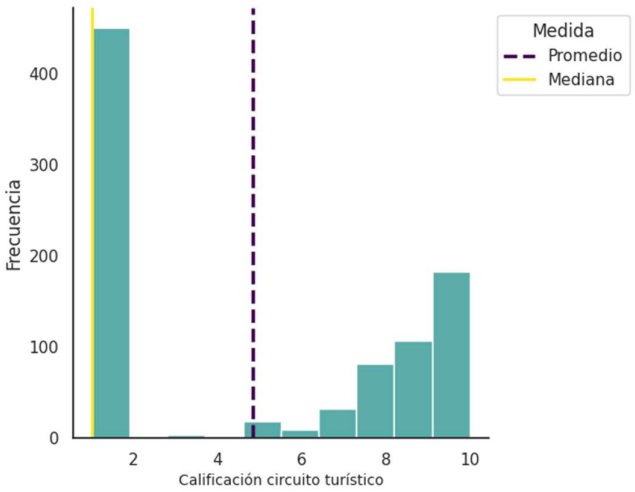
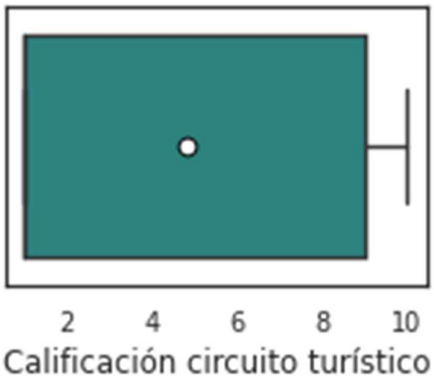


Gráfico 12: Diagrama de caja circuito turístico



Fuente: Elaboración propia con Python

En la *Tabla II: Descripción Categóricas* se puede observar los principales marcadores para las variables cualitativas.

Tabla II: Descripción Categóricas

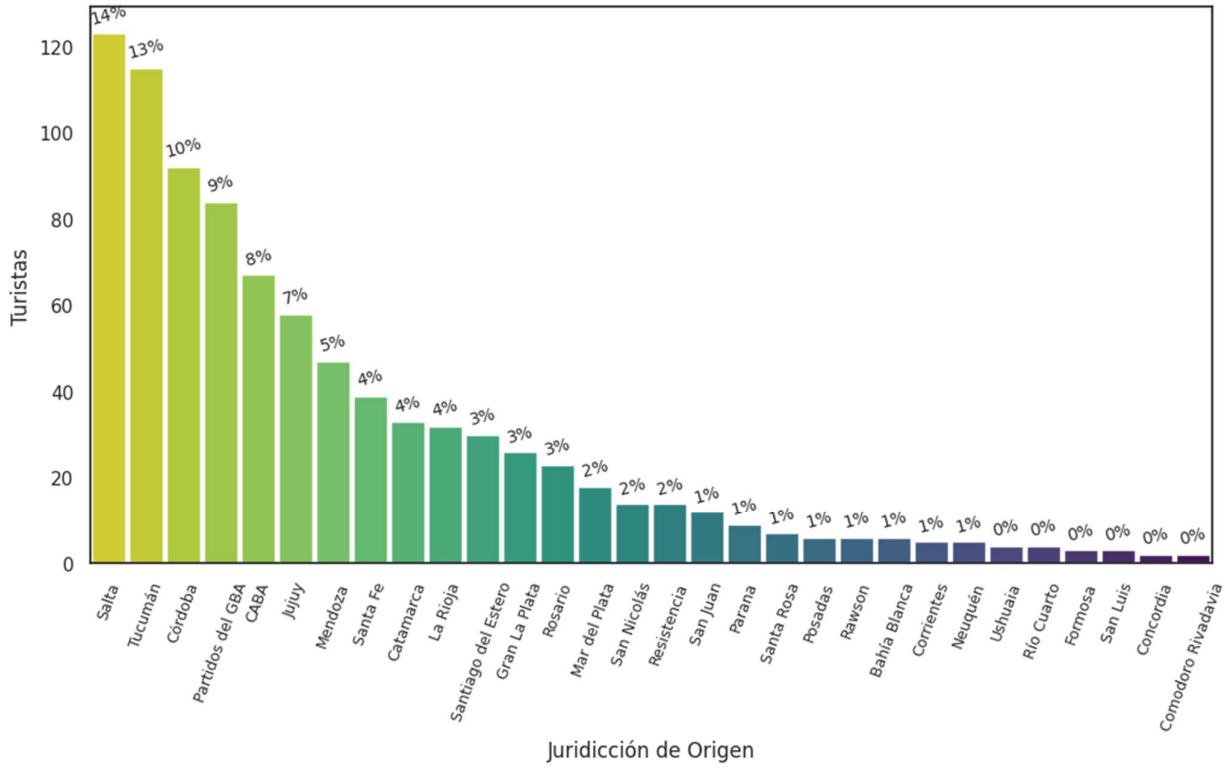
	Localidad origen	**Destino**	**Localidad destino**
Valores únicos	30	6	65
Top	Salta	Salta	Salta
Frecuencia	123	314	189

Fuente: Elaboración propia

7. *Jurisdicción de Origen*

Como se puede observar en la *Tabla II: Descripción Categóricas* las mismas se corresponden con 30 aglomerados, de los cuales la jurisdicción de Salta representa el valor modal con una frecuencia de aparición de 123 repeticiones. En el *Gráfico 13: Jurisdicción de origen* podemos observar todos los aglomerados con su incidencia en la base de datos seleccionada. La provincia de Salta es la jurisdicción donde tienen residencia permanente el 14% de todos los turistas que visitaron la región norte en el año 2022 para la muestra seleccionada. Luego encontramos a Tucumán con el 13%, Córdoba con el 10% y Gran Buenos Aires con el 9% y Ciudad Autónoma de Buenos Aires con el 8%, acumulando entre ellos el 54% de todos los turistas que visitaron la región. Luego con un porcentaje menor encontramos las provincias de Jujuy con el 7%, Mendoza 5% y Santa Fe, Catamarca y La Rioja con el 4% respectivamente. Se puede notar que gran parte de los turistas visitantes del norte tienen residencia permanente en la misma región o bien en localidades aledañas.

Gráfico 13: Jurisdicción de origen

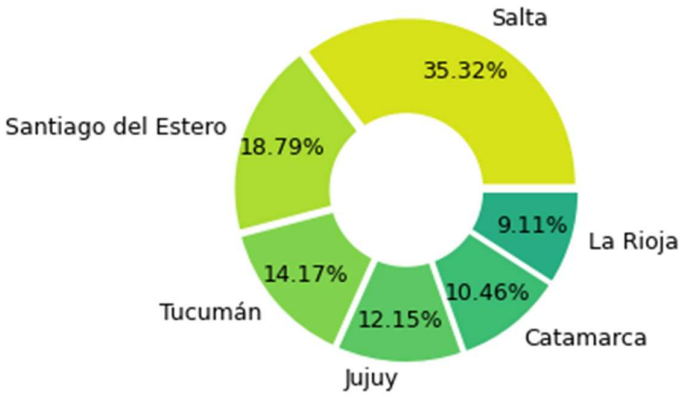


Fuente: Elaboración propia con Python

5. *Provincia de destino*

Como se puede observar en la *Tabla II: Descripción Categóricas* la región Norte está conformada por 6 provincias: Jujuy, Salta, Tucumán, Santiago del Estero, Catamarca y La Rioja. En el *Gráfico 14: Destino* podemos observar que el principal destino elegido fue la provincia de Salta quien acaparó el 32,69% de todos los visitantes. Luego se puede encontrar a Santiago del Estero con el 21,34%, Tucumán 12,31%, Jujuy 11,76%, Catamarca 11,08% y La Rioja con el 10,81%.

Gráfico 14: Destino



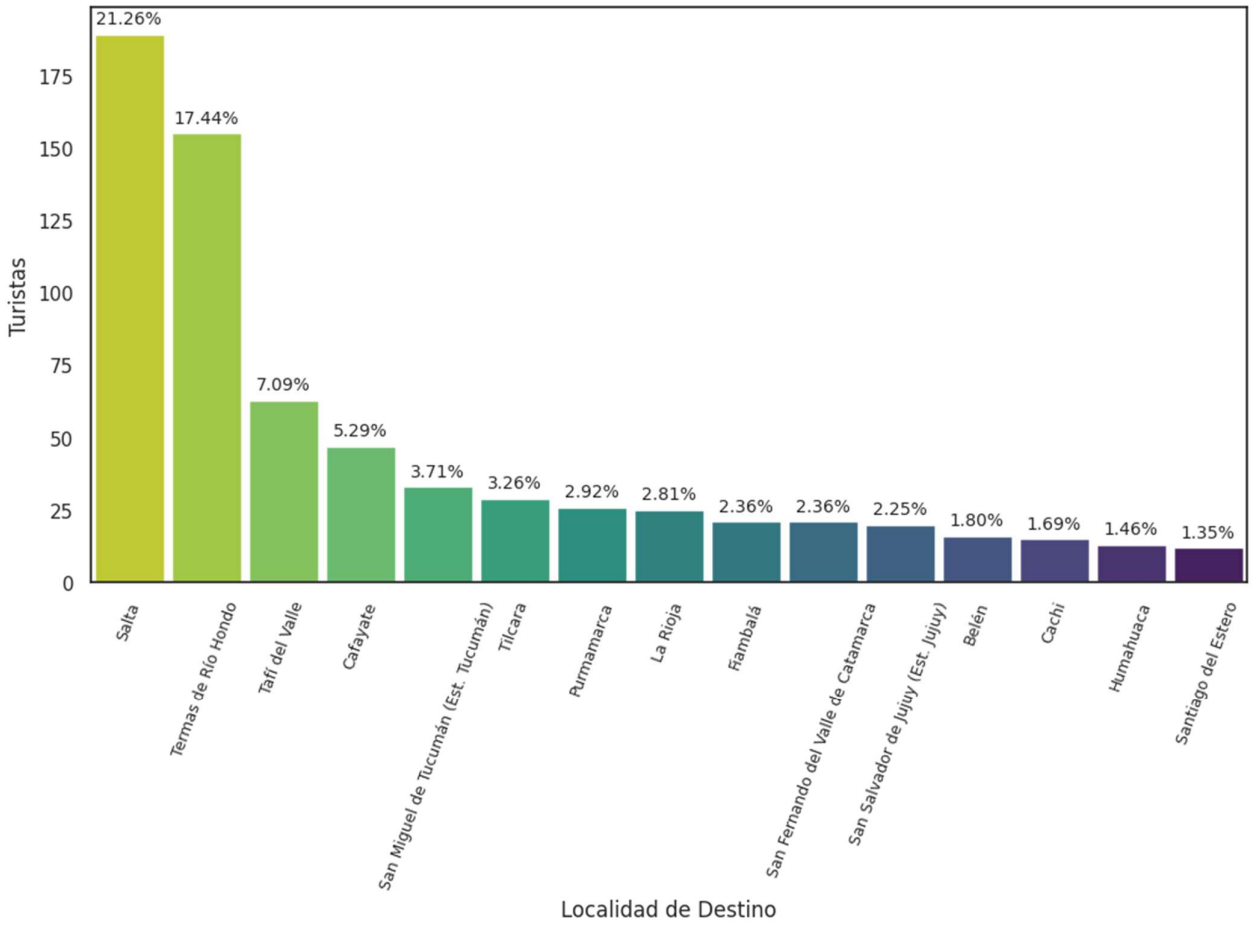
Fuente: Elaboración propia con Python

6. *Localidad de destino*

La base está conformada por 65 localidades que se encuentran dentro de las 6 provincias que conforman la región norte, tal como se puede observar en la *Tabla II: Descripción Categóricas.* En el *Gráfico 15: Top 15 de localidad elegidas como destino* podemos encontrar los 15 destinos principales elegidos por los turistas con sus respectivos porcentajes sobre el total. Como se puede observar, los dos destinos más elegidos por los visitantes fueron ciudad de Salta que ostenta el 21,26% y Termas de Río Hondo con el 17,44% las cuales suman entre las dos el 38,70% de todos los turistas. Luego, con marcada diferencia podemos encontrar a Tafí de Valle, Cafayate y Tilcara. Asimismo, podemos notar que, dentro de las 15 más elegidas, la Provincia de Salta presenta dos localidades la ciudad de Salta y Cafayate. En cuanto a la Provincia de Santiago del Estero solo se puede encontrar a la localidad de Termas de Rio Hondo.

Las localidades de Tilcara, Purmamarca, San Salvados de Jujuy, y Humahuaca perteneces a la provincia de Jujuy, mientras que La Rioja y Santa Teresita a la provincia de La Rioja. En Catamarca se ubican las localidades de Fiambalá, San Fernando del Valle, Belén y Tinogasta. Por último, en la Provincia de Tucumán se ubican las localidades de Tafí del Valle y San Miguel de Tucumán.

Gráfico 15: Top 15 de localidades elegidas como destino



Fuente: Elaboración propia con Python

2.3. Limpieza y preparación datos

Una vez analizado en conjunto de datos inicial, se procede a seleccionar 6 de las 9 variables de la base buscando identificar patrones en la demanda de los turistas de la región norte del país para el año 2022 que den lineamientos sobre las cuales trazar recomendaciones para estrategias comerciales de la empresa “Turi Nor”.

De esta manera resulta un conjunto de datos de 889 observaciones y 6 variables: edad, integrantes, estadía, alojamiento, gastronomía y circuito turístico. No se seleccionaron las variables relacionadas a la ubicación, porque se busca definir un perfil de visitante en función a aspectos demográficos y preferencias de consumo del producto turístico.

La gestión de datos en la organización incluye numerosos aspectos a considerar, uno de los más importantes refiere a su calidad. Si los datos carecen de ese atributo en su estructura, será complejo llevar a cabo su análisis y extraer conclusiones con un nivel de fiabilidad aceptable. Esto puede conducir a tomar malas decisiones, generar conflictos operacionales o bien trazar estrategias empresariales no adecuadas, afectado no solo en desempeño de la organización sino su estructura de costos. En efecto, la calidad de los datos puede ser analizada en diferentes aspectos dentro de los cuales podemos encontrar su completitud, unicidad y validez.

Del análisis efectuado sobre la base de datos se ha podido constatar que no existen valores nulos o inexistentes, por lo que los datos comprendidos en ella se encuentran completos. Asimismo, respecto de su unicidad se ha corroborado la inexistencia de registros duplicados. En cuanto a su validez se ha constatado que los rangos de valores de cada una de las variables de estudio son razonables respecto de los sucesos que las mismas representan.

Otro punto para tener en cuenta es el análisis y criterio respecto del tratamiento de los valores atípicos o outliers. Las variables estadía, integrantes, alojamiento y gastronomía tal como se ha analizado en la exploración de datos presentan este tipo de valores. En efecto, teniendo en cuenta la naturaleza de las variables y el objetivo del caso de estudio se ha optado por seguir los siguientes lineamientos. Para las variables estadía e integrantes se ha aplicado el método univariado del cálculo del rango intercuartílico (IQR) determinando como valor atípico a aquellas observaciones que cuyo valor sea a 1.5 veces mayor o menor del IQR. Una vez obtenidos los límites, se aplicó el método del recorte o capping por el cual se reemplazan los valores atípicos por los límites previamente indicados. En forma particular se reemplazan por el límite superior de 15 pernoctes y 7 miembros del hogar en la variable estadía e integrantes respectivamente ya que teniendo en cuenta la naturaleza de los paquetes o productos turísticos ofrecidos por la empresa se consideró que podrían no ser adecuados y generar distorsión en el análisis posterior.

En el caso de las variables alojamiento y gastronomía se decide no efectuar ningún tratamiento ya que al ser variables que expresan una preferencia del consumidor del producto turístico, los valores atípicos pueden permitir identificar elementos distintivos y diferenciales importantes de las conductas de consumo. Resulta interesante no solo encontrar las similitudes

en los comportamientos de la demanda, sino hallar elementos que en un conjunto de población aparentemente similar genere comportamientos de consumo diferentes.

Por último, como las variables seleccionadas están expresadas en diferentes unidades de medida se procede a estandarizar la base de datos restando la media a cada variable y dividiendo la misma por su desvío estándar. Para ello se utilizó el comando StandardScaler de la librería Sklearn de Python.

Las tareas de exploración y análisis descriptas en este apartado han permitido profundizar el conocimiento de los datos y establecer metodologías y criterios de trabajo para avanzar con su desarrollo. A partir de ellos, se ha procesado y preparado el conjunto de datos cumplimentando pautas de calidad, dando tratamiento a los valores atípicos y estandarizando la base de datos para aplicar los métodos de análisis multivariado que se exponen en el siguiente apartado.

3. Aplicación de métodos de análisis multivariantes

Es este apartado se aplicarán dos técnicas multivariantes sobre el conjunto de datos previamente procesados. Primero se realizará un análisis de clúster aplicando 4 métodos jerárquicos para definir los números posibles de agrupamientos. Luego partiendo de ese primer acercamiento, se aplicará el método de partición K-means con el fin de agrupar a los turistas en función de los atributos. Por último, se presentan los resultados de las técnicas aplicadas y se establecen lineamientos sobre posibles estrategias comerciales.

3.1. Análisis de clúster

El análisis de clúster también llamado de conglomerado tiene por fin agrupar elementos en grupos homogéneos en función de las similitudes entre ellos (Peña, 2002). Aplicando el mismo al conjunto de datos elegido se buscará agrupar a los turistas que eligieron destino de viaje el norte de Argentina, identificando patrones en función de su rango etario, integrantes del hogar, estadía y preferencias sobre alojamiento, gastronomía y circuitos turísticos. Esto involucra agruparlos y dividirlos en un número de grupos o clúster de manera que cada visitante solo pertenezca a un grupo, encontrándose todos clasificados en algún grupo y buscado que cada agrupamiento sea lo más homogéneo posible. (Peña, 2002). De esta manera se buscará la mayor separación externa, esto es, que los clústeres estén lo más alejados entre sí posible, pero con la mayor cohesión interna, es decir, que, dentro de cada uno de ellos, las observaciones sean lo más parecidas posible.

3.2. Aplicación de Métodos Jerárquicos

Los métodos jerárquicos parten de una matriz de distancias o similitud entre los elementos de la muestra y construyen una jerarquía basada en una distancia**.** (Peña, 2002). Se pueden utilizar diferentes métricas de distancia, en el presente desarrollo se aplicó la distancia euclídea. Partiendo de la mencionada matriz se buscará clasificar a los elementos en jerarquías aplicando diferentes métodos de asignación de forma de repartir las observaciones a los grupos. El proceso general inicia tomando todas las observaciones de la base y selecciona dos elementos según el criterio de distancia fijado con el método jerárquico elegido. Con ese par de elementos se forma una clase que agrupa y sustituye el par de elementos individuales iniciales. Se vuelve a calcular la distancia de esa clase con otro elemento del conjunto. Este es un proceso iterativo que finalizaría cuando se encuentren todos los elementos agrupados en una clase única.

A continuación, se describen e implementan en el conjunto de datos los siguientes 4 diferentes métodos jerárquicos: vecino más cercano o single linkage, vecino más lejano o complete linkage, vinculación promedio o average linkage y el método de Ward. Para ello se utilizó el módulo Scipy.cluster comando hierarchy de la librería Sklearn y módulo Yellowbrick.cluster comando KElbowVisualizer de Python.

1. Vecino más cercano o single linkage

El mismo inicia tomando todas las observaciones de la base y selecciona los dos elementos más próximos entre sí de forma que la distancia entre dos elementos es aquella que se da entre los dos miembros más cercanos de esos grupos, comenzando el proceso iterativo antes descripto. Los resultados de la aplicación de este método sobre el conjunto de datos se pueden observar en el *Gráfico 16: Dendograma aplicando el método del vecino más cercano* donde se han agrupado las observaciones en 4 clústeres. Dicho dendograma esta truncado mostrando solo las últimas 12 uniones entre elementos para poder visualizarlo en una sola figura, pero se puede apreciar las agrupaciones y la cantidad de elementos de cada una. En efecto, las observaciones ya agrupadas aparecen entre paréntesis mostrando la cantidad de elementos que esa clase tiene. Las observaciones que no se han agrupado hasta esa instancia, se muestran sin paréntesis y el número corresponde a su índice en el conjunto de datos. Está aclaración es válida para todos los dendograma presentados en este trabajo. Así, podemos ver que el primer agrupamiento consta de 2 observaciones, segundo de 3, el tercer agrupamiento consta de 1 observación (índice 292) y el cuarto cuenta con 883 observaciones. También se puede visualizar la distancia en la cual se unen los diferentes grupos, encontrándose todas las observaciones en el extremo final de 2,859.

2. Vecino más lejano o complete linkage

El mismo inicia tomando todas las observaciones de la base y selecciona los dos elementos más alejados entre sí de forma que la distancia entre dos elementos es aquella que se da entre los dos miembros más distanciado de esos grupos. Los resultados de la aplicación de este método sobre el conjunto de datos se pueden observar en el G*ráfico 17: Dendograma aplicando el método del vecino más lejano* donde se han agrupado las observaciones en 8 clústeres. Así, podemos ver que los agrupamientos ordenados del 1 a 8 constan de 6, 39, 54, 43, 285, 244, 174 y 44 observaciones respectivamente. También se puede visualizar la distancia en la cual se unen los diferentes grupos, encontrándose todas las observaciones en el extremo final de 8,941.

3. Vinculación promedio o average linkage

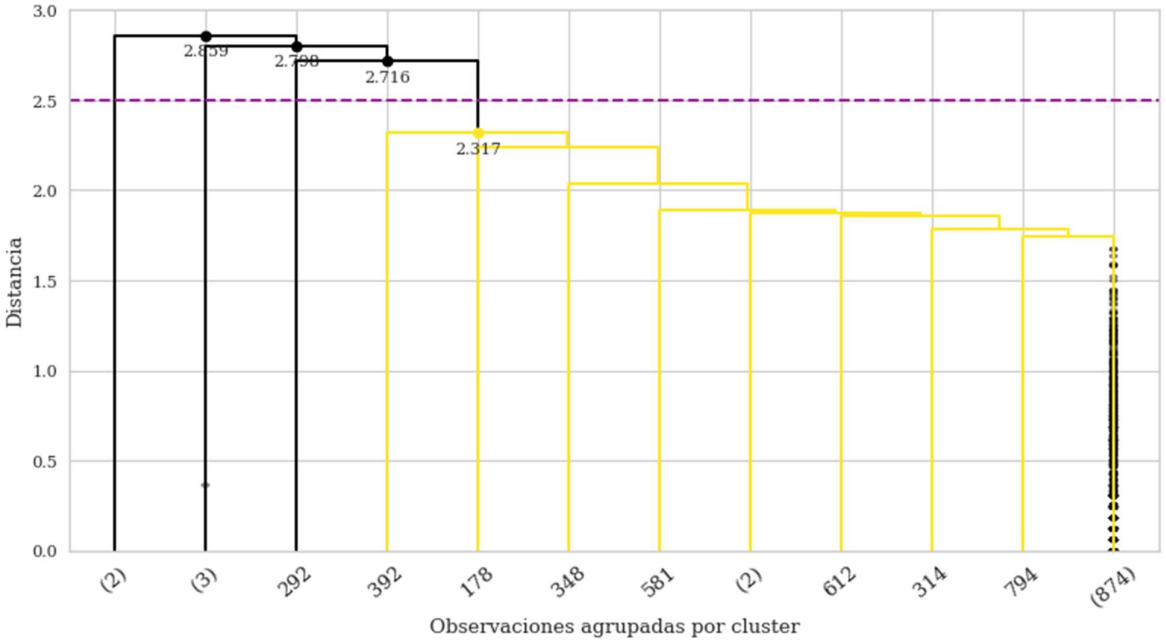
El mismo inicia tomando todas las observaciones de la base calculando la distancia promedio entre todos los pares de observaciones posibles, agrupando un miembro de un grupo con otro miembro del otro grupo. Los resultados de la aplicación de este método sobre el conjunto de datos se pueden observar en el G*ráfico 18: Dendograma aplicando el método vinculación promedio* donde se han agrupado las observaciones en 4 clústeres. Así, podemos ver que los agrupamientos ordenados del 1 al 4 constan de 5, 44, 810 y 30 observaciones respectivamente. También se puede visualizar la distancia en la cual se unen los diferentes grupos, encontrándose todas las observaciones en el extremo final de 5,041.

4. Ward

El mismo considera todas las posibles combinaciones de observaciones, eligiendo la agrupación que maximice la homogeneidad de cada agrupamiento. Los resultados de la aplicación de este método sobre el conjunto de datos se pueden observar en el G*ráfico 19: Dendograma aplicando el método Ward* donde se han agrupado las observaciones en 5 clústeres. Así, podemos ver que los agrupamientos ordenados del 1 al 5 constan de 283, 32, 264, 44 y 266 observaciones respectivamente. También se puede visualizar la distancia en la cual se unen los diferentes grupos, encontrándose todas las observaciones en el extremo final de 42,06.

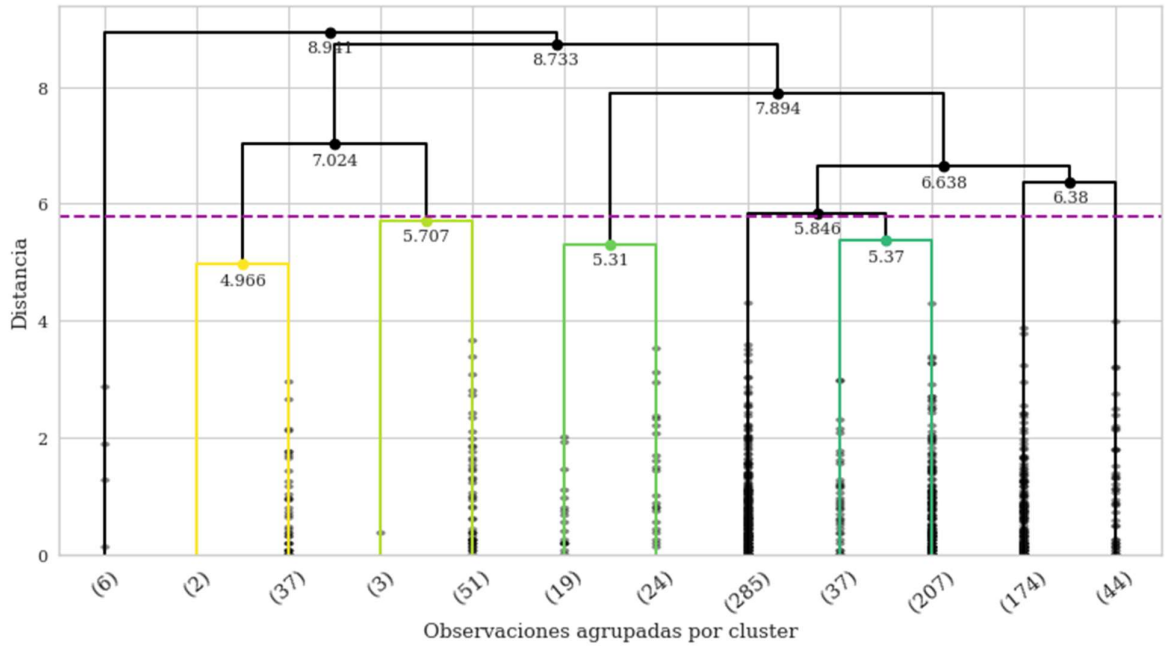
En el G*ráfico20* se complementan los análisis anteriores mediante un gráfico utilizando el método del codo en el cual el óptimo de agrupaciones es coincidente con el análisis del dendograma.

Gráfico 16: Dendograma aplicando el método del vecino más cercano



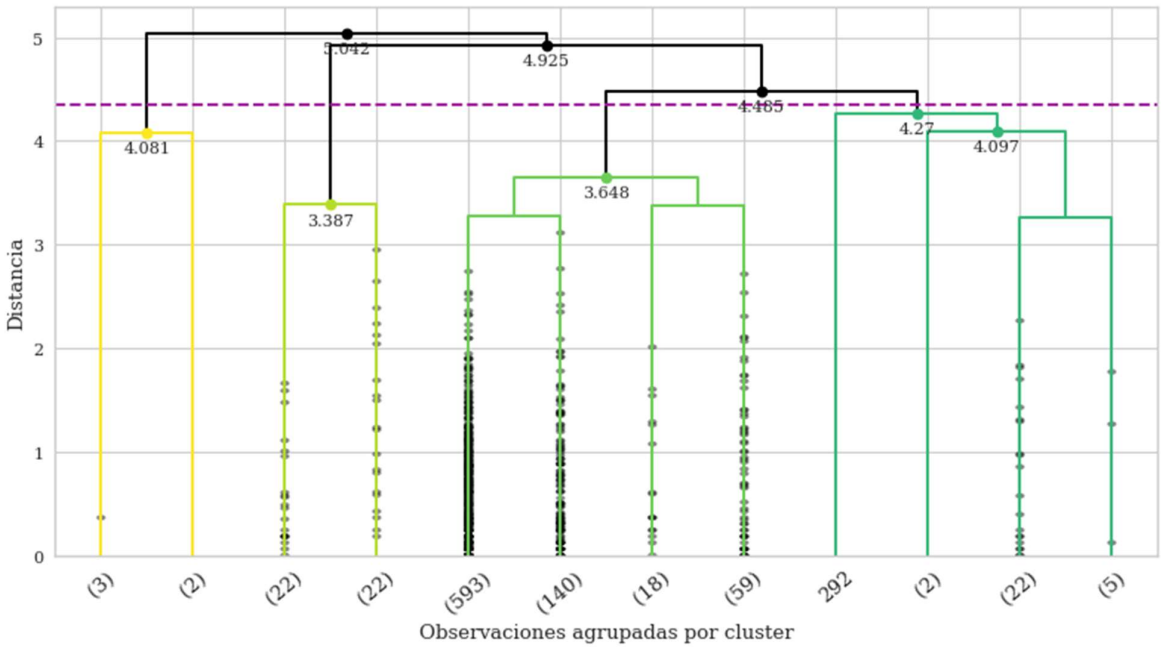
Fuente: Elaboración propia con Python

Gráfico 17: Dendograma aplicando el método del vecino más lejano



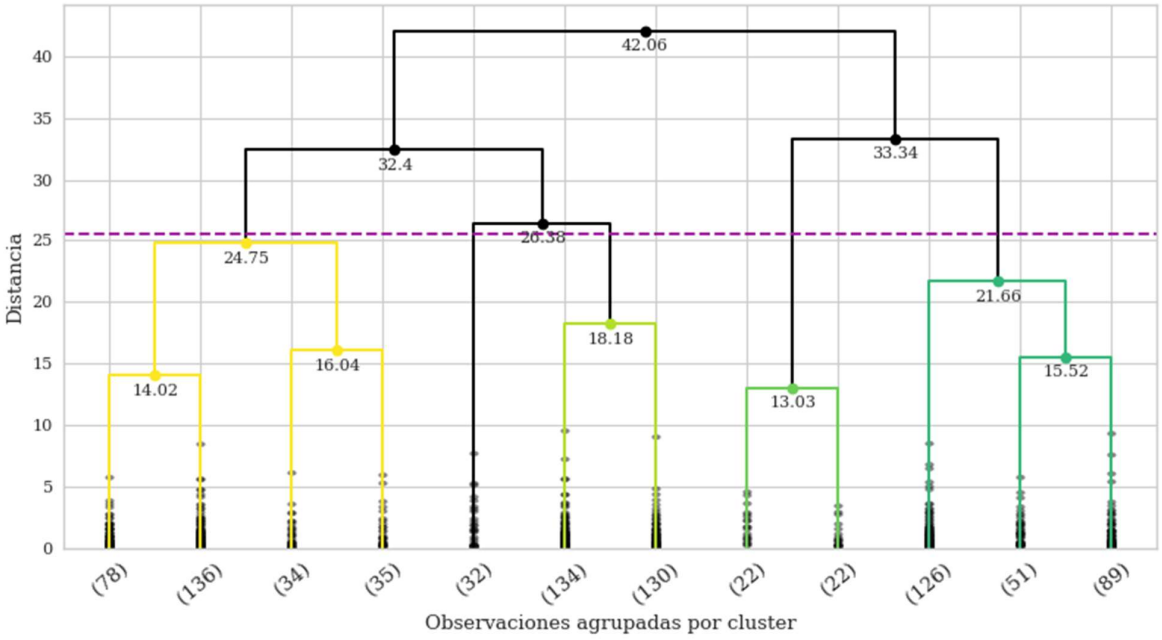
Fuente: Elaboración propia con Python

Gráfico 18: Dendograma aplicando el método vinculación promedio



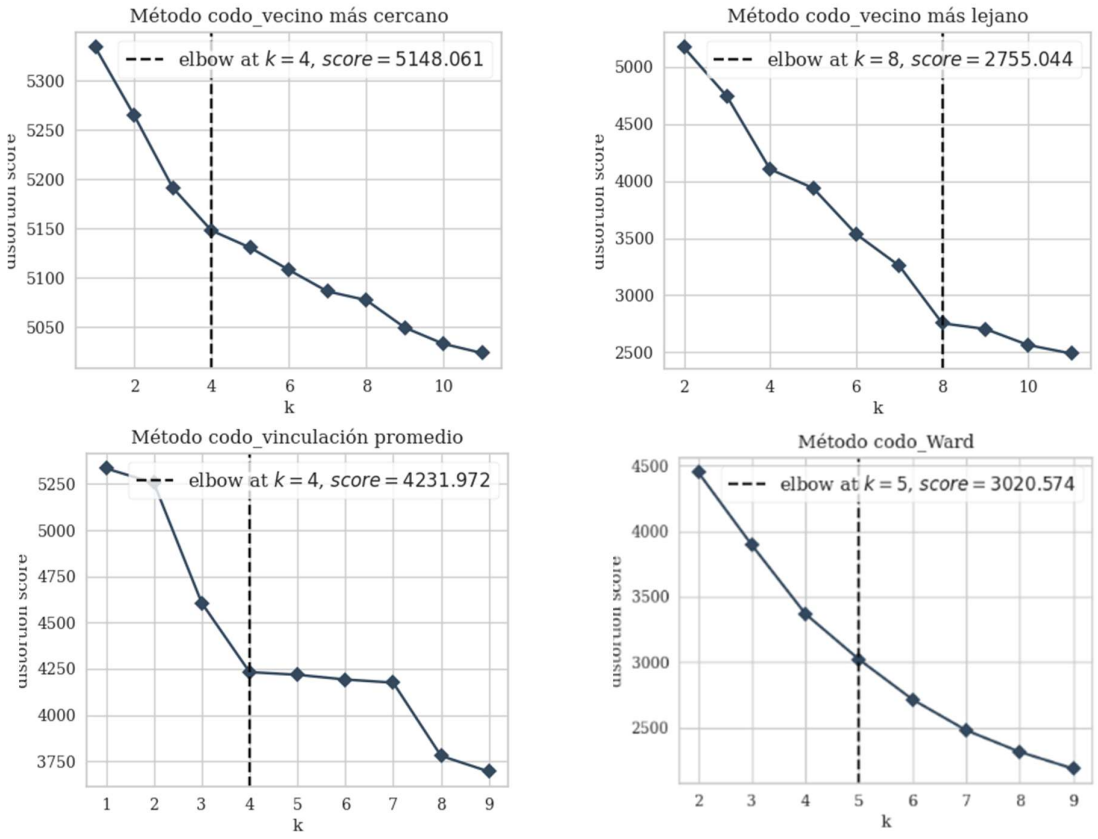
Fuente: Elaboración propia con Python

Gráfico 19: Dendograma aplicando el método Ward



Fuente: Elaboración propia con Python

Gráfico 20: *Gráfico de codo*



Fuente: Elaboración propia con Python

3.3. Aplicación de Métodos no Jerárquicos

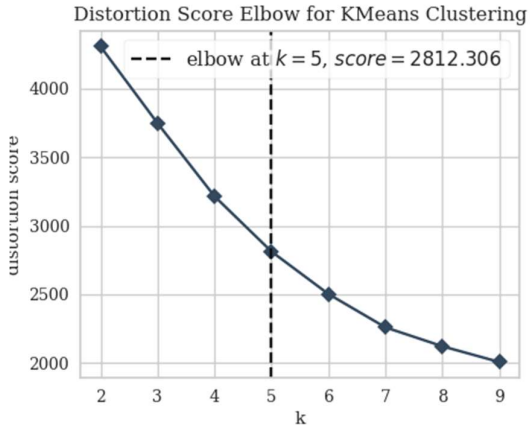
Algoritmo K-Means

En un método no jerárquico que parte de seleccionar n puntos aleatorios como centros de grupo. A partir de ahí calcula las distancias de cada observación al centro y las asigna al grupo más cercano. Es un proceso iterativo que concluye cuando ya no es posible optimizar las asignaciones, alcanzando ese punto cuando se encuentre la mayor homogeneidad dentro de cada grupo, pero con la mayor heterogeneidad entre cada uno de ellos.

En la sección inmediata anterior se ha podido establecer un rango de 4 a 8 como número de clústeres adecuados según los métodos analizados. Para conciliar ese resultado y encontrar el número de k óptimos se aplica el método del codo utilizando el modelo de K-means, cuyos resultados se pueden observar en *la Gráfico 21: Método codo k-means*. Cómo se puede

observar, el resultado es coincidente con lo hallado aplicando el método de Ward donde k=5 sería el número de agrupamiento más adecuado.

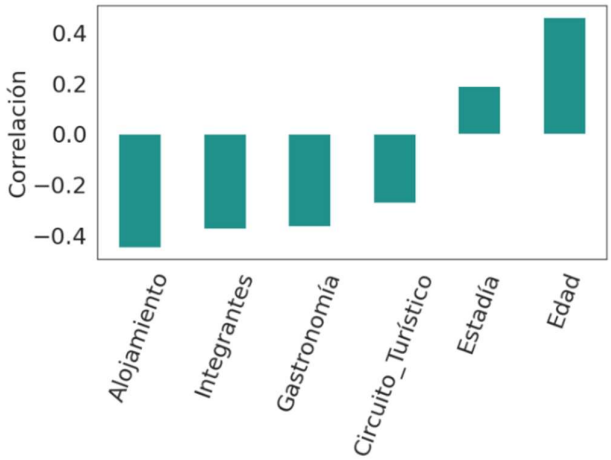
Gráfico 21: Método codo K-means



Fuente: Elaboración propia con Python

A partir de la definición del número óptimo de clústeres se ejecuta el algoritmo K-Means. Para ello se utilizó el módulo sklearn.cluster comando KMeans de la librería Sklearn de Python. Este proceso da como resultado la conformación de los clústeres del 1 al 5 con 234, 254,241, 47 y 117 observaciones cada uno respectivamente. Asimismo, analizando la correlación entre las variables, respecto del clúster podemos establecer el impacto de las variables en el proceso de agrupamiento. En efecto, observando el *Gráfico 22: Impacto en el agrupamiento* se puede evidenciar que las variables con más incidencia son la edad y alojamiento.

Gráfico 22: Impacto en el agrupamiento



Fuente: Elaboración propia con Python

Luego le siguen el número de integrantes, la gastronomía y el circuito turístico. Por último, encontramos la estadía. Es importante resaltar que en términos general mientras que la variable edad y estadía se relacionan positivamente en la formación del agrupamiento, toda la demás lo hacen de forma negativa.

Continuando con el análisis se puede observar la *Tabla 3: Análisis de medias* donde están indicados los valores medios por cada variable en cada uno de los clústeres y en el conjunto de datos general. De esta manera se pueden ver de forma clara que sus elementos diferenciales y característicos, logrando definir las siguientes tipologías:

Tabla 3. Análisis medias

Clúster	Edad	Estadía	Integrantes	Alojamiento	Gastronomía	Cir. Turístico
1	34,68	3,74	4,45	8,98	8,76	4,73
2	64,41	5,04	2,67	9,31	8,91	9,19
3	64,83	4,66	2,73	9,31	9,14	1,05
4	47,55	3,00	4,36	9,28	1,06	1,91
5	62,26	6,60	2,19	6,42	7,46	4,58
Media General	55,53	3,21	4,73	8,85	8,33	4,83

Fuente: Elaboración propia

1. *Clúster N °1:* Con 234 casos, representa el 26,32% de la base. Está conformado por los turistas más jóvenes de la muestra, con pocas noches de estadías, pero con mayor cantidad de integrantes por hogar. En sus preferencias respecto de la planificación del viaje atribuyen importancia al alojamiento y gastronomía, pero en menor medida a las atracciones que ofrece el destino.

2. *Clúster N°2:* Con 254 casos, Representa el 28,57% de la base. Está conformado por turistas de mayor edad, con estadías más prolongadas, pero menor cantidad de integrantes por hogar. En sus preferencias respecto de la planificación del viaje atribuyen elevada importancia tanto al alojamiento como a la gastronomía y destacan como elemento diferencial la importancia en la planificación y elección del destino las atracciones que ofrecen los circuitos turísticos del lugar de destino.

3. *Clúster N°3*: Con 241 casos *r*epresenta el 27,11% de la base. Está conformado por turistas de mayor edad, en promedio levemente superior al clúster N°2. Presentan estadías coincidentes a la media general de toda la base y una menor cantidad de integrantes por hogar. En sus preferencias respecto de la planificación del viaje atribuyen gran importancia tanto al alojamiento como a la gastronomía, pero no a las

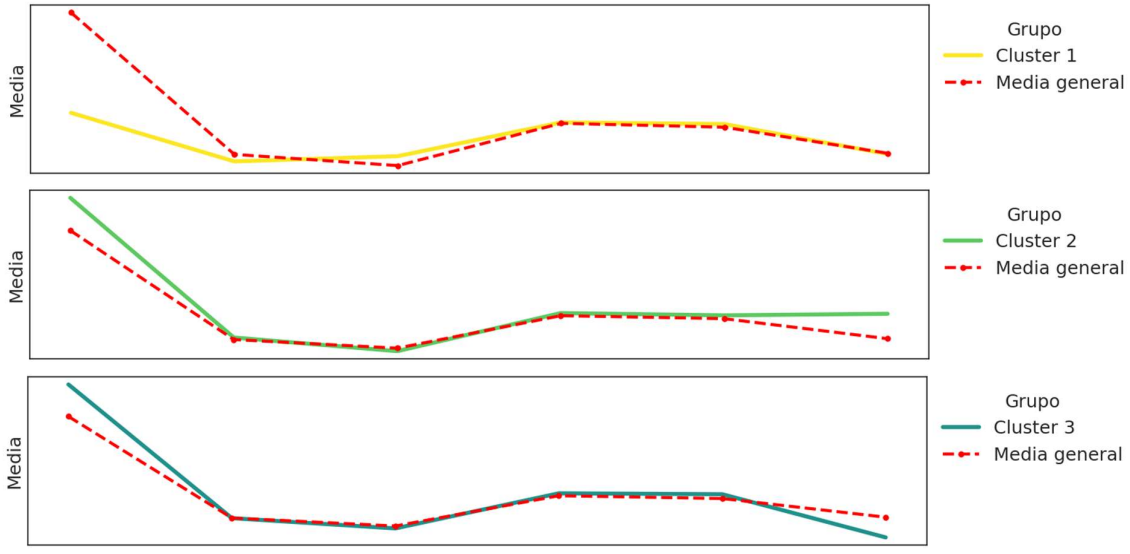
atracciones que ofrecen los circuitos turísticos del lugar de destino. Este agrupamiento pone en el centro de sus preferencias el disfrute de los como los servicios y comodidades ofrecidas por los establecimientos de alojamiento como puede ser áreas de recreación, servicios en las habitaciones, masajes, spa e instalaciones, así como también a la gastronomía ofrecida en el destino.

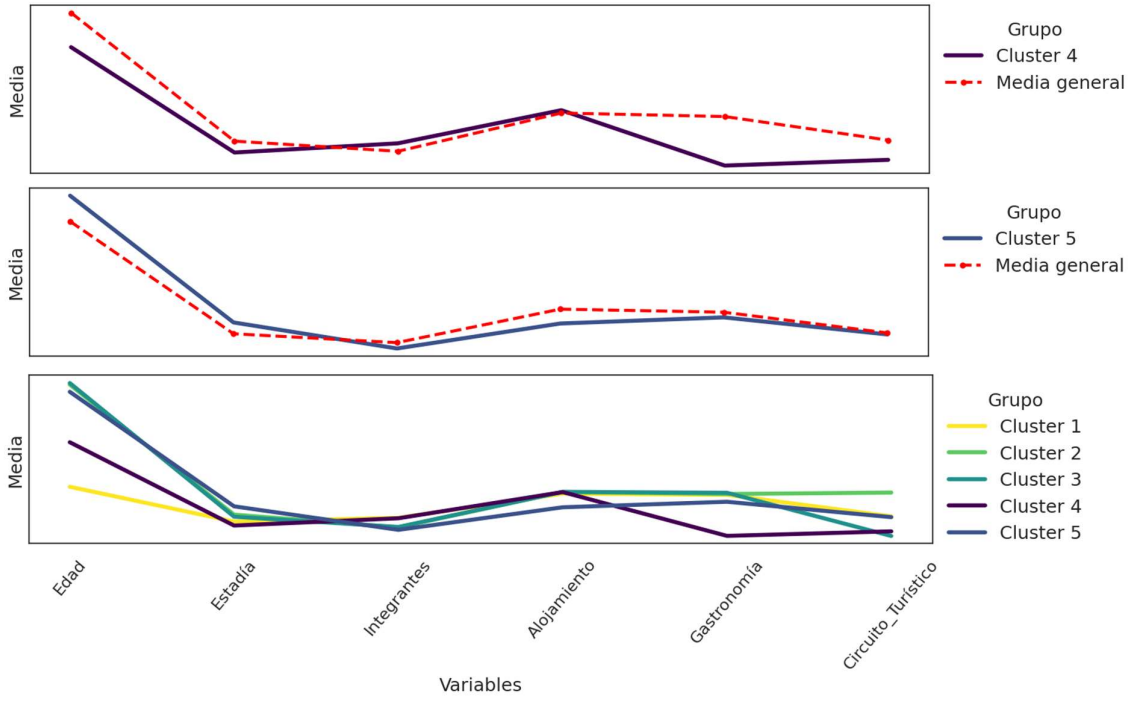
4. *Clúster N°4*: Con 47 casos *r*epresenta el 5,29% de la base. Está conformado por turistas de edad intermedia, levemente inferior al promedio general total. Presentan estadías y cantidad de integrantes coincidentes a la media general. En el centro de sus preferencias respecto de la planificación del viaje ponen al alojamiento y las comodidades que este pueda brindar, atribuyendo muy poca importancia a la gastronomía y los circuitos turísticos.

5. *Clúster N°5*: Con 113 casos *r*epresenta el 12,71% de la base. Está conformado por turistas de mayor edad, levemente superior al promedio general. Se caracterizan por presentar estadías considerablemente superiores a la media general con pocos integrantes por hogar. Se puede observar que la importancia asignada a la gastronomía, el alojamiento y las atracciones de los turistas de este agrupamiento son inferiores al promedio general de la base.

La caracterización efectuada de los clústeres puede observarse en forma clara y precisa en el *Gráfico 23: Comparación Clústeres.* Las primeras 5 figuras muestran las medias de cada uno de los variables por clúster contra la media general del conjunto de datos. En la última, en cambio, se pueden observar los 5 agrupamientos.

Gráfico 23: Comparación Clústeres y media general





Fuente: Elaboración propia con Python

3.3. Segmentación de demanda

A través del agrupamiento obtenido es posible diferenciar características que permiten construir perfiles de consumidores de forma de ofrecer propuestas de productos turísticos personalizados en la agencia de viajes "Turi Nor". En efecto,

1. Paquete full: a los turistas perfilados en el clúster 2 se podría ofrecer un "paquetes full" con estadía de aproximadamente una semana donde destaquen los servicios y comodidades de los alojamientos y se combinen con excursiones y circuitos turísticos para conocer las atracciones del destino incluyendo la gastronomía.
2. Paquete bienestar: Para perfiles incluidos en el clúster 3 se podría ofrecer estadías de alrededor de 6 noches con buenas prestaciones en cuanto al alojamiento tal que permita el disfrute de las instalaciones y comodidades de los hospedajes y la posibilidad de realizar tour de tipo gastronómico, para conocer y experimentar platos, comidas y bebidas exclusivas de la región.
3. Paquete pausa: Para perfiles incluidos en el clúster 4, en cambio, se podría ofrecer un "paquete pausa" por un periodo corto de tiempo el cual coloque en el centro las comodidades y servicios ofrecidos por el hospedaje asegurando un buen descanso y relajación.

Conclusión

La importancia del sector turístico como motor de crecimiento y generación de valor en el país es una realidad presente e innegable. De la misma manera lo son las oportunidades que la cultura data-driven y el análisis de grandes volúmenes de dato ofrecen a las organizaciones turísticas que acompañen y promuevan su uso a nivel estratégico y de toma de decisiones. La aplicación del concepto de turismo inteligente permite conectar las preferencias y experiencias a lo largo del ciclo del viaje de los turistas con las tecnologías, mejorando la gestión de los recursos y creando ventajas competitivas a aquellos agentes económicos que lo implementen de forma adecuada.

A través del presente trabajo se ha puesto en evidencia como el análisis de los datos y la aplicación de algoritmos de aprendizaje automático no solo permiten tener un conocimiento más profundo de la demanda de turismo, sino que además permiten acercar las preferencias del consumidor con el producto turístico, por medio de la personalización de los paquetes ofrecidos, acompañando la experiencia turística antes, durante y después del viaje.

En efecto, en el primer apartado se presentan los desafíos que los grandes volúmenes de datos imponen a las organizaciones turísticas. Se manifiesta como el análisis de datos y la aplicación de diversas tecnología de información se transforman en un factor que impulsa el crecimiento de las organizaciones . Ese cambio origina el surgimiento de un turismo inteligente, donde la integración de los diversos elementos descriptos permite brindar una experiencia enriquecedora y con valor agregado al turista, la cual a nivel local debería considerar las particularidades que presenta el sector turístico en Argentina.

En el segundo apartado se explora la base a datos a utilizar, brindando detalles del proceso de recolección inicial de los datos y su accesibilidad . Asimismo, se describe el protocolo de seguridad utilizado en su configuración de modo de asegurar la privacidad y responsabilidad en el manejo de la información. Luego, por medio de análisis de estadística descriptiva se analiza cada una de las variables de estudio , obteniendo así un conocimiento más profundo de la base de datos a utilizar. También se realizan tareas de limpieza y transformación en las variables seleccionadas para asegurar requisitos de calidad en el uso de los datos.

Por último, en el tercer apartado se ejecuta un análisis de clúster de tipo jerárquico y no jerárquico. El análisis efectuado permitió identificar 5 clústeres bien definidos respecto a cada una de las variables consideradas . Ello posibilitó segmentar la demanda e identificar perfiles

de clientes , sobre los cuales se proyectó 3 opciones de productos turisticos adaptados a los requerimientos previamente referidos.

El presente trabajo se concentró en analizar los turistas en el norte Argentina, no considerando otras regiones dentro del país. Sería enriquecedor incorporarlas para futuras líneas de investigación, así como analizar los elementos distintivos de cada una de ellas y su influencia en el sector turístico. En este aspecto, también resultaría un aporte valioso incorporar otras fuente de datos, de manera de poder comparar diversos hallazgos encontrados y nutrir el campo de estudio.

Como se ha podido desarrollar y profundizar a lo largo del presente trabajo, el análisis de datos y aplicación de técnicas de aprendizaje automático en el sector turístico da lugar a un campo de estudio potente, en crecimiento y con muchos caminos y oportunidades por descubrir.

Referencias

Aguilar, L. J. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones.* Alfaomega Grupo Editor.

Almirón, A. V. (2008). El turismo como impulsor del desarrollo en Argentina. Una revisión de los estudios sobre la temática. *Aportes y transferencias, 12(1)*, 57-86.

Argentina, S. d. (n.d.). *https://datos.yvera.gob.ar/*. Retrieved from https://datos.yvera.gob.ar/: https://datos.yvera.gob.ar/

Argentina.gob.ar. (n.d.). *https://www.argentina.gob.ar/*. Retrieved from https://www.argentina.gob.ar/produccion/registrar-una-pyme/que-es-una-pyme

Barreto, A., & Azeglio, A. (2013). La problemática de la gestión deEl capital humano en las MiPyMEs de alojamiento turístico de la Ciudad de Buenos Aires-Argentina. *Estudios y perspectivas en turismo, 22(6)*, 1140-1159.

Benckendorff, P. T. (2018). The role of digital technologies in facilitating intergenerational learning in heritage tourism. . *In Information and Communication Technologies in Tourism 2018: Proceedings of the International Conference in Jönköping, Sweden,Springer International Publishing*, 463-472.

CEPAL, N. (2020). Análisis de la huella digital en América Latina y el Caribe: enseñanzas extraídas del uso de macrodatos (big data) para evaluar la economía digital. 61-65.

Christensen et. al, C. (2015). What is disruptive innovation. *Harvard business review , 93(12)*, 44-53.

Espelt, N. G. (2000). Patrimonio cultural y turismo: nuevos modelos de promoción vía Internet. *Cuadernos de turismo, (6).*, 73-88.

Fernández, R. A. (2017). Estimación del multiplicador Keynesiano del turismo internacional en Argentina. *Estudios y perspectivas en turismo, 26(2)* , 248-266.

Fiestas Argentinas. (n.d.). *https://fiestasargentinas.ar/*. Retrieved from https://fiestasargentinas.ar/

García-Morales, E. (2012). Gobernanza de la información. *Anuario ThinkEpi, 6*, , 100-103.

Gartner. (n.d.). *https://www.gartner.com*. Retrieved from https://www.gartner.com: https://www.gartner.com/en/information-technology/glossary/big-data

Girardin, F. C. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing, 7(4),*, 36-43.

Gretzel, U. S. (2015). Smart tourism: foundations and developments. *Electronic markets, 25,*, 179-188.

J. Hwang, H. P. (2015). Asia Pacific Journal of Information Systems, 25 (1) . 163-178.

Lamelas, J. V. (2017). Revolución Big Data en el turismo: Análisis de las nuevas fuentes de datos para la creación de conocimiento en los Destinos Patrimonio de la Humanidad de España. *International Journal of Information Systems and Tourism (IJIST), 2(2)*, 23-39.

Más Ferrando, A. R.-R. (2020). La revolución digital en el sector turístico. *Oportunidad para el turismo en España.*, 228-249.

McAfee, A. B. (2012). Big data: the management revolution. *Harvard business review, 90(10)*, 60-68.

Muñoz, A. D., & Sánchez, S. G. (2015). Destinos turísticos inteligentes. *Economía industrial, 395*, 61-69.

N°17622, L. (1968). Marco legal de las estadísticas oficiales. *Marco legal de las estadísticas oficiales*.

Oliva, M., & Schejer, C. (2006). El empleo en las ramas características del turismo en Argentina. *Aportes y Transferencias, 10(2)*, 36-68.

OMT, O. M. (2008, 02). *Recomendaciones Internacionales para estadísticas de turismo 2008*. Retrieved from Glosario: https://www.unwto.org/

Önder, I. K.-H. (2016). Tracing tourists by their digital footprints: The case of Austria. *Journal of Travel Research 55(5)*, 566-573.

Peña, D. (2002). *Análisis de datos multivariantes.* McGraw-Hill .

Puebla, J. G. (2018). Big Data y nuevas geografías: la huella digital de las actividades humanas. *Documents d'anàlisi geogràfica, 64(2)*, 195-217.

Shapiro, C. &. (1999). *Information rules: A strategic guide to the network economy.* Harvard Business Press.

Sigala, M. (2018). New technologies in tourism: From multi-disciplinary to anti-disciplinary advances and trajectories. *Tourism management perspectives,25*, 151-155.

Sturzenegger, A., & Porto, N. (2008). *https://www.ahrcc.org.ar/.* Retrieved from https://www.ahrcc.org.ar/_descargas/Informe-de-la-importancia-de-la-act.pdf

Tabares, L. F. (2014). Big Data analytics: Oportunidades, retos y tendencias. *Universidad de San Buenaventura*.

World Travel & Tourism Council, W. (2023, 5). *https://wttc.org/.* Retrieved from https://wttc.org/: https://wttc.org/news-article/sector-de-viajes-y-turismo-argentino-avanza-hacia-su-recuperacion-en-la-contribucion-al-pib-wttc

Reporte Trabajo Final Integrador de Especialización

"IDENTIFICACIÓN DE PATRONES EN LA DEMANDA DE TURISMO INTERNO EN LA REGIÓN NORTE DE ARGENTINA"

-Implementación de análisis de Clúster en Python-

Autora: Romina Silvia Lucchetti

Como mentora del presente Trabajo Final Integrador de Especialización de la alumna Romina Lucchetti, transmito mi opinión sobre la investigación realizada.

Con respecto al tema de análisis, considero que aborda una problemática relacionada con la gestión de datos en organizaciones; en este caso organizaciones relacionadas con el sector turístico de Argentina.

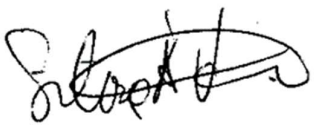
La alumna analiza una base de datos de la encuesta de viajes y turismo de los hogares (EVyTH), elaborada por La Dirección Nacional de Mercados y Estadísticas a cargo del Ministerio de Turismo y Deportes de la Nación para la región norte del país, con el objetivo de identificar patrones en la demanda de turismo interno de la región norte de Argentina.

En la investigación, se enmarca la gestión de datos en contextos de Big data y se detalla el uso de grandes volúmenes de datos para generar turismo inteligente. Se define como turismo inteligente al que a partir de herramientas y algoritmos de machine learning e inteligencia artificial potencia el análisis de datos y brinda un mejor conocimiento de los mercados y de las necesidades de los turistas, optimizando ingresos.

A partir del análisis efectuado se deja en evidencia la importancia del análisis de datos en el área turística para comprender, modelizar y predecir comportamientos de los consumidores e implementar acciones orientadas a generar oportunidades de mejora y crecimiento en el ámbito del turismo.

Se articulan contenidos de las asignaturas Métodos de Análisis Multivariado (con la depuración y estudio descriptivo de la base de datos, y la aplicación de métodos multivariantes como el Método de Conglomerados) y Taller de Programación (con el uso de Python).

La coherencia del enfoque planteado, el uso de algoritmos de Minería de Datos, la pertinencia de las referencias bibliográficas y el correcto análisis de los resultados obtenidos, permiten señalar a este trabajo como un aporte relevante en el área del turismo, evidenciando la posibilidad de emprender futuras líneas de investigación.



Dra. Silvia Vietri