Tendencias y patrones del turismo interno en Argentina: un análisis exploratorio con métodos de *Machine Learning*

Propuesta de investigación

Ciencia de datos

Lic. en Ciencias del Comportamiento

Tutorial 1
Grupo 5

Magdalena Cobb
Pedro García Vassallo
Marcos Olavarría

Profesores:
María Noelia Romero
Ignacio Spiousas

7 de Diciembre, 2024

1. Introducción

El turismo en la Argentina es una fuente significativa de empleo, causa alternativa de crecimiento económico y crecimiento regional (Porto, 1999; Oliva & Schejer, 2006). El entendimiento de los distintos patrones y tendencias en turismo puede ser relevante para tomar decisiones tanto desde el ámbito privado como público. Por ejemplo, conocer los factores influyentes en los comportamientos turísticos de las personas podría aportar a la toma de decisiones a la hora de esbozar un presupuesto de inversión en publicidad turística de cada localidad.

La siguiente propuesta busca responder a la pregunta de investigación: ¿cuáles son los principales predictores del comportamiento turístico interno en Argentina? Además, se busca construir diferentes modelos estadísticos para poder predecir distintas variables de respuesta, como las principales localidades de destino, el gasto incurrido durante el viaje y la cantidad de noches de estadía, según las características del turista y del viaje realizado.

2. Revisión bibliográfica

Dentro del área de investigación de tendencias y comportamiento turístico, se ha destacado la aplicación de metodologías de *clusters* para segmentar al tipo de turistas según la población. Por ejemplo, el equipo de Ramirez y colaboradores (2018) observa cambios de flujo y de turistas que generó la clasificación cultural del *World Heritage City* en Portugal. Utilizan el método de *clusters* para comprender las motivaciones y valoraciones que tienen los turistas los cuales denominan como "turistas culturales" y, de esta manera, lograron clasificar tres segmentos principales con características diferentes. En otro ejemplo, el equipo de Disegnaa y colaboradores (2018) buscó generar una clasificación de turistas a partir de la identificación de patrones relevantes. Hicieron extracciones de muestras equivalentes a lo largo del tiempo y utilizaron un método de *matching* que permite realizar análisis de *clusters* inter temporales, evitando recurrir a estudios longitudinales que son más costosos.

Por otro lado, se han generado modelos predictivos a partir de datos provenientes del turismo. Por ejemplo, la investigación llevada a cabo por Georgieva-Trifonova y Mancheva-Ali (2024) buscó predecir la cantidad de noches de estadía en la ciudad de Veliko Tarnovo, Bulgaria a partir de datos del distrito obtenidos a partir de una encuesta nacional de turismo. Adicionalmente, complementaron esta información con datos de Google Trends para generar un modelo polinómico para predecir la cantidad de noches de estadía en Veliko Tarnovo.

Si bien la investigación acerca de patrones turísticos ha estado creciendo en los últimos años, se destaca la falta de bibliografía referida a investigación turística en Argentina. Se han encontrado varias referencias al estudio y predicción de comportamientos turísticos en diferentes países como en China (Wang et al., 2021) y en Bulgaria (Georgieva-Trifonova et al., 2024), pero la falta de investigación en la región plantea una oportunidad de estudio interesante. El área de estudio de tendencias y comportamientos turísticos tiene una gran relevancia para una serie de diferentes actores sociales, tales como instituciones privadas y públicas, que buscan maximizar beneficios a la hora de proveer servicios en este sector.

3. Base de datos

Se utilizará la base de datos proveniente de la Encuesta de Viajes y Turismo de los Hogares, disponible en el Sistema de Información Turística de la Argentina de la Subsecretaría de Turismo (EVyTH; Subsecretaría de Turismo, 2012–2023). Los datos se recopilaron mediante un sistema de encuestas telefónicas a una muestra probabilística de 5.000 hogares, entre el 2012 y el 2023.

La base contiene más de 80 variables y más de 450.000 observaciones. Principalmente, incluye datos sobre el motivo del viaje, el tipo de alojamiento y medio de transporte usado, la fecha y duración del viaje, los gastos incurridos, entre otras. Además, contiene información sobre los viajeros: cantidad de viajeros, tipo de cobertura médica,

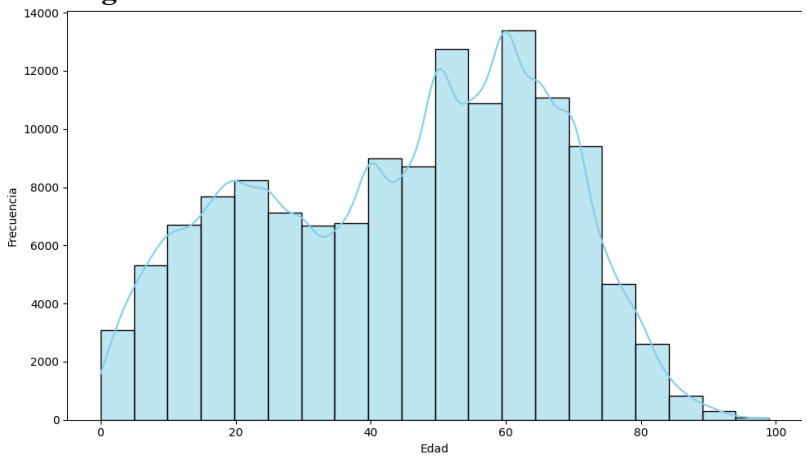
localidad de origen, y otros datos demográficos (como el sexo, la edad, el nivel educativo alcanzado, la condición de actividad, etc.).

La Tabla 1 presenta estadísticas descriptivas preliminares (previo a la limpieza de la base) para algunas características demográficas y de los viajes realizados. Como se observa en la Figura 1, la mayoría de los encuestados tienen entre 40 y 70 años de edad (Media=44,6, SD=21,59) y son de género masculino. Además, los grupos de viajeros fueron de aproximadamente 3 personas (Media=2,91, SD=1,45) y los viajes duraron 3,11 noches en promedio, aunque con un desvío estándar de 5,58.

Tabla 1.
Estadísticas descriptivas preliminares

	Edad (años)	Sexo	Gasto incurrido	Noches de estadía	Personas por grupo
Media	44,60	1,48	7518,99	3,11	2,91
SD	21,59	0,49	36385,94	5,58	1,45

Figura 1. *Distribución de edad de los encuestados*



En términos generales, las principales localidades de destino fueron Villa Carlos Paz, la Ciudad de Buenos Aires, y Mar del Plata (Figura 2); incluso, su predominancia por sobre las demás localidades se observa a través de los años, como refleja la Figura 3. Además, se observa un patrón de alta coincidencia entre las regiones de origen y las regiones de destino (Figura 4), lo cual puede indicar una tendencia al turismo intrarregional.

Figura 2. *Principales 20 localidades de destino*

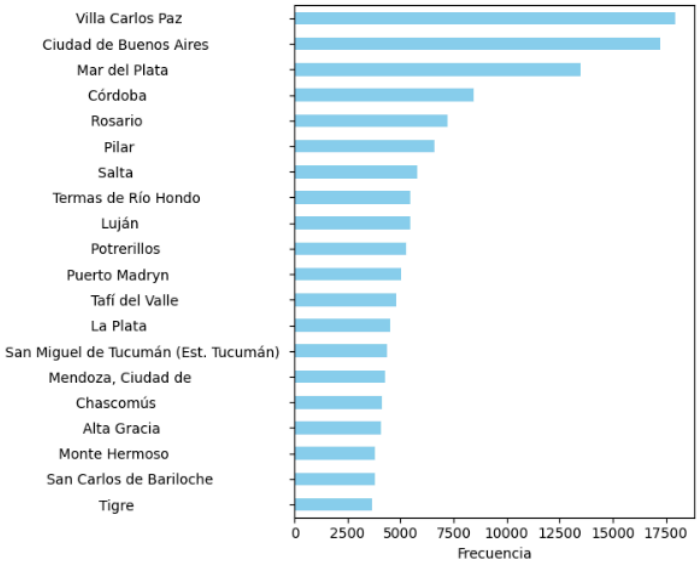


Figura 3. *Localidades de destino en el tiempo*

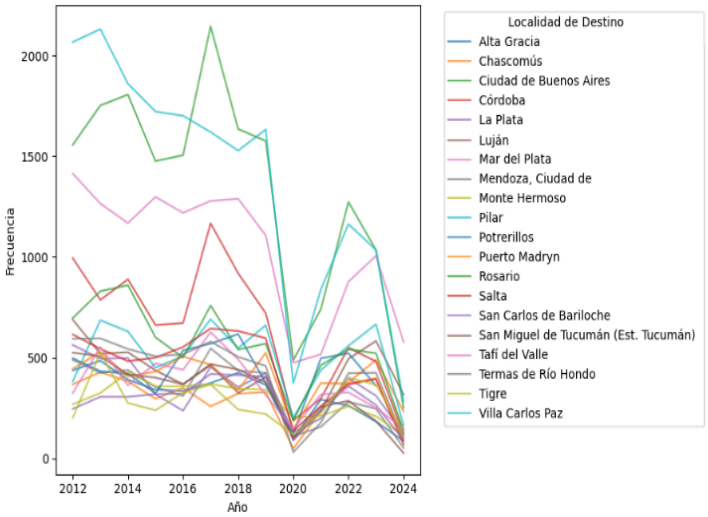
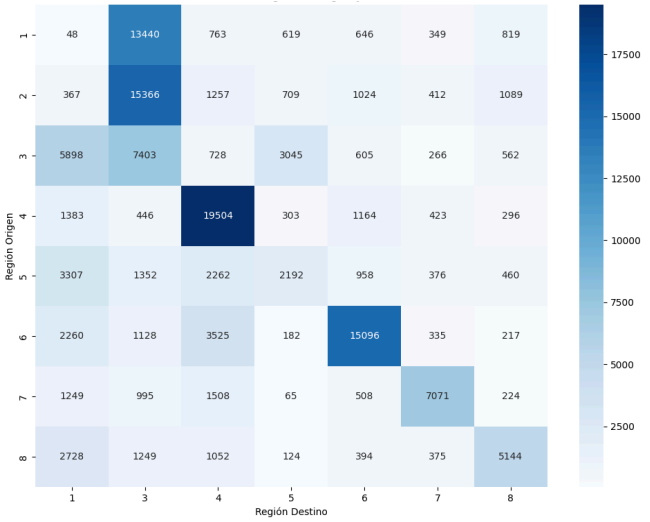


Figura 4. *Relación entre regiones de origen y destino*



Entre las principales características de los viajes realizados, se identifica a los viajes de esparcimiento, ocio o recreación como el principal motivo, seguido de visitas a familiares o amigos (Figura 5). No obstante, esto varía según la localidad de destino: por ejemplo, en Luján, Salta y Alta Gracia, se observa una mayor proporción de turistas por motivos religiosos que en el resto de las localidades. Adicionalmente, la mayoría de los encuestados reportó alojarse en viviendas de familiares y amigos, por sobre hoteles o viviendas alquiladas (Figura 6). Además, el medio de transporte más frecuente para los viajes de turismo interno fue el automóvil propio, seguido del ómnibus (Figura 7).

Figura 5. *Principales motivos de viaje según localidad de destino*

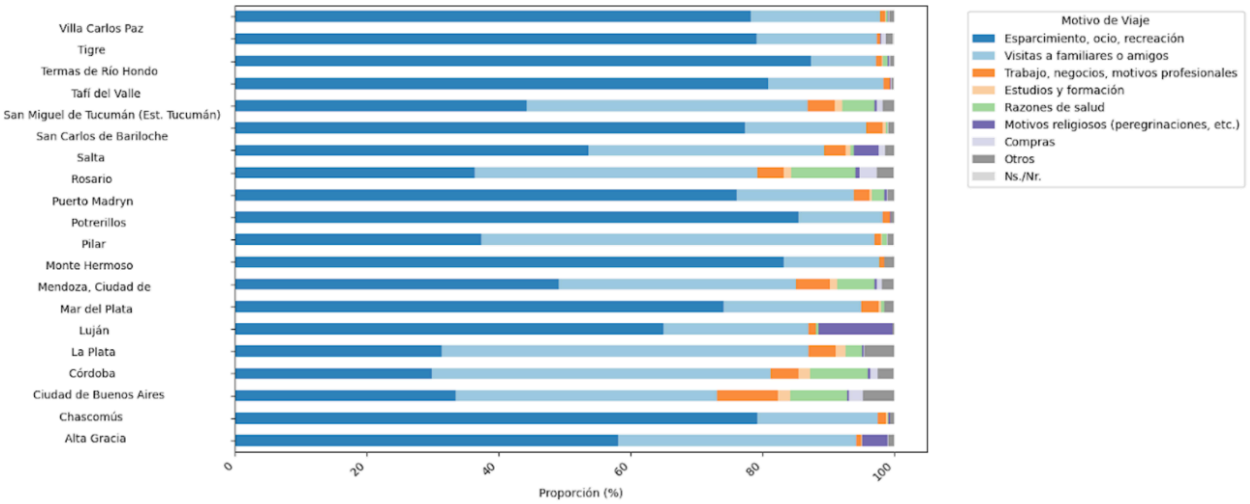


Figura 6.
Proporción de encuestados por tipo de alojamiento

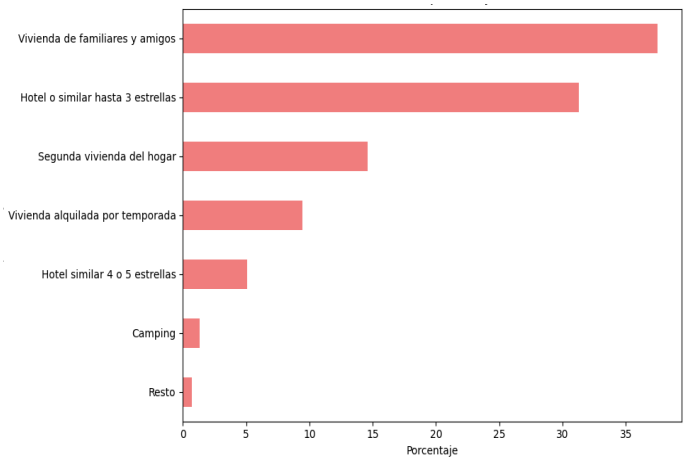
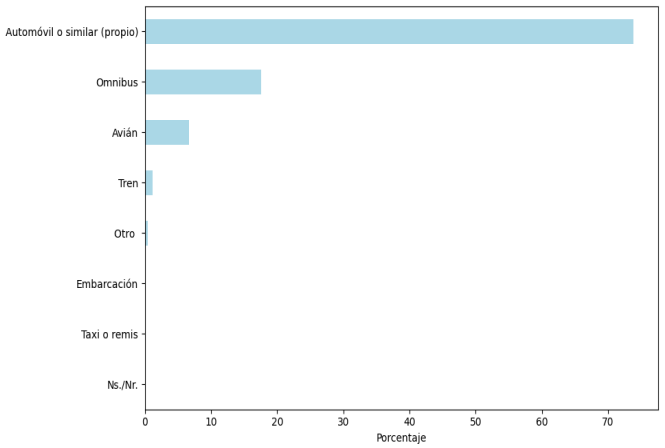


Figura 7.
Proporción de encuestados por medio de transporte



4. Metodología

Previo a la implementación de las técnicas estadísticas, se realizará una limpieza profunda de la base de datos para asegurar el mejor funcionamiento de los modelos en las etapas siguientes. En esta instancia, se eliminarán observaciones que no hagan sentido (por ejemplo, valores negativos para variables como edad o ingreso), se tratará a los valores faltantes y los *outliers*, y se crearán *dummies* para las variables categóricas para poder utilizar ciertos métodos.

En la etapa exploratoria, se emplearán histogramas, gráficos de barra y de dispersión para visualizar la distribución de distintas variables demográficas a modo de descripción de la muestra (como complemento a aquellas presentadas: Figuras 1 a 7). Además, se harán

matrices de correlación para comprender la relación entre las variables para detectar posibles colinealidades entre los potenciales predictores. De este modo, se podrán identificar patrones iniciales para poder obtener un pantallazo general de las variables de interés.

Adicionalmente, se llevará a cabo un análisis de clústers de k-medias para generar una segmentación de los distintos perfiles de turistas. Este método permitirá identificar k grupos homogéneos sin superposición, según su similitud en cuanto a distintas variables sociodemográficas y elecciones de consumo turístico (como el motivo o la cantidad de participantes). Por ejemplo, se podrían visualizar los grupos de turistas que realizan viajes familiares o viajes de negocio, entre otros. Obtener esta segmentación podría facilitar la identificación de patrones en el tipo de turismo ejercitado en Argentina, y así proporcionar información relevante a la hora de diseñar publicidades, paquetes de viaje y ofertas más atinadas.

Por otra parte, se emplearán distintos métodos para predecir dos variables de respuesta de interés a partir de las características del turista y el contexto del viaje: localidad de destino turístico, gasto incurrido durante el viaje y cantidad de noches de estadía. Estos modelos de clasificación y predicción pueden resultar útiles a la hora de identificar los factores relevantes a la promoción de distintos destinos turísticos y a también para generar herramientas que permitan a los distintos actores involucrados (turistas, organizaciones estatales, agencias de viaje) poder estimar el gasto a incurrir en un determinado viaje previo a su realización.

Por un lado, se utilizará el método de ensamble *Random Forest* para obtener información sobre la importancia de las distintas características del turista y del viaje a la hora de predecir la localidad de destino turístico. Este método genera árboles de clasificación de manera aleatoria a partir de muestras *bootstrappeadas*, usando en cada árbol m predictores aleatorios ($m < p$); se seleccionará el valor del hiperparámetro m usando *Cross-Validation*. De

esta manera, se obtiene una medida robusta de la importancia de las variables; por ejemplo, este método podría indicar que la variable de nivel educativo es más importante para predecir localidad de destino que la edad, o que el género (o vice versa). Además, este método permitirá clasificar a las observaciones en una localidad de destino según sus características (sociodemográficas, por ejemplo). Para evaluar su desempeño con respecto a esto último, se utilizará la medida de error *Out-of-the-bag* (OOB), como proxy del error de predicción del modelo por fuera de la muestra. Esta medida utiliza las observaciones que no fueron usadas en cada una de las muestras *bootstrappeadas* como "muestra de validación" para estimar el error de clasificación del *Random Forest.*

Por otro lado, se emplearán dos modelos de regresión con mínimos cuadrados parciales (PLS) para predecir el gasto incurrido en el viaje y las noches de estadía. Este método de aprendizaje supervisado permitirá proyectar los p predictores iniciales (más de 80 predictores de la base total) a un subespacio de m dimensiones para poder reducir la dimensionalidad y así generar un modelo con menor varianza para evitar el overfitting. En particular, resulta relevante tomar medidas de reducción de la dimensionalidad, ya que la base cuenta con una gran cantidad de predictores que podrían estar altamente correlacionados. Se transformará a los predictores en distintos componentes principales según la variabilidad de la información explicada, teniendo en cuenta la relación de estos con la variable de respuesta (gasto incurrido). Una vez obtenidos los m componentes, se los usará para ajustar el modelo de regresión de mínimos cuadrados parciales. Nuevamente, el hiperparámetro m (cantidad de componentes principales) se seleccionará por *Cross-Validation* para cada uno de los modelos.

Finalmente, se evaluarán con un método de validación cruzada (o *Cross-Validation*), para estimar la medida de error de predicción por fuera de la muestra. Se partirá a la muestra de entrenamiento usada para ajustar cada modelo en k=10 particiones de manera aleatoria. En cada iteración, se ajustará el modelo y se calculará el error cuadrático medio (ECM) usando

una partición de la muestra como prueba. Finalmente, se computará el promedio de estos errores para obtener una estimación del ECM de los modelos por fuera de la muestra.

5. Conclusiones y limitaciones

A partir de la base de la Encuesta de Viajes y Turismo de los Hogares (EVyTH), este trabajo busca explorar los diferentes factores que influyen en el comportamiento turístico interno en Argentina. En particular, tiene como objetivo desarrollar modelos predictivos para explicar y predecir la localidad de estadía, el gasto incurrido en el viaje y la cantidad de noches de estadía, según las características de los turistas.

Como principales resultados, se espera encontrar los predictores más relevantes para las variables de interés y construir modelos que logre predecirlas fuera de la muestra de entrenamiento. Como primer hallazgo del análisis exploratorio de la base, se observa una fuerte inclinación de los turistas por los viajes intra-región, usando como método de transporte al vehículo propio y como estadía principal las viviendas de familiares o amigos. Asimismo, se observa que los viajes suelen ser de corta estadía (3 noches en promedio), por lo que se puede inferir que se realizan viajes cortos a destinos cercanos. Esto refleja un perfil de turista que resulta interesante, al cual se podría llegar a delimitar y clasificar (según sus características demográficas) en el análisis de clusters.

En cuanto a la predicción del gasto incurrido durante la estadía, este análisis plantea un desafío mayor. La base contiene registros desde el 2012 hasta el 2023, años en los que Argentina ha acumulado niveles altos de inflación. Si no hay un correcto ajuste de esta variable previo a su análisis, los resultados obtenidos pueden estar entorpecidos. De igual manera, se debería tener en cuenta que, al ser una base de datos grande con muchas variables, esta puede ser ruidosa y se debe tomar con cuidado la instancia de limpieza. Finalmente, se identifica como una limitación el posible sesgo en los datos debido al modo de recolección. Al provenir de encuestas telefónicas, pueden existir ciertos sesgos en el muestreo (los

encuestados son en su mayoría hombres de entre 40 y 70 años) que puede hacer que la muestra, y los resultados obtenidos, no sean del todo representativos de la población argentina.

6. Referencias

Andres Artal-Tur, Jose Miguel Navarro-Azorín & Luisa Alamá-Sabater (2022): The role of destination contextual effects in driving the expenditure of tourists: a multilevel spatial modelling approach, Regional Studies, DOI: 10.1080/00343404.2022.2110578

Disegna, M., D'Urso, P., & Massari, R. (2018). Analysing cluster evolution using repeated cross-sectional ordinal data. Tourism Management, 69, 524-536.

Georgieva-Trifonova, T., & Mancheva-Ali, O. (2024). Predicting Tourist Arrivals: A Google Trends-Based Model for Destination Management. TEM Journal, 13(3).

Ramires, A., Brandao, F., & Sousa, A. C. (2018). Motivation-based cluster analysis of international tourists visiting a World Heritage City: The case of Porto, Portugal. Journal of Destination Marketing & Management, 8, 49-60.

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

Porto, N. (1999). El turismo como alternativa de crecimiento. Documentos de Trabajo.

Pulido-Fernández, Juan Ignacio, Jairo Casado-Montilla, and Isabel Carrillo-Hidalgo. (2020). Understanding the Behaviour of Olive Oil Tourists: A Cluster Analysis in Southern Spain. Sustainability 12, no. 17: 6863. https://doi.org/10.3390/su12176863

Subsecretaría de Turismo. (2012-2023). *Encuesta de Viajes y Turismo de los Hogares (EVyTH)*. Sistema de Información Turística de la Argentina (SINTA). Recuperado de https://www.yvera.tur.ar/sinta/.

Wang, L., Wang, S., Yuan, Z., & Peng, L. (2021). Analyzing potential tourist behavior using PCA and modified affinity propagation clustering based on Baidu index: Taking Beijing city as an example. Data Science and Management, 2, 12-19.