Análisis del Precio de Viajes de Tren con Origen/Destino Madrid

Autor: Pedro Durán Porras

Resumen

Este estudio analiza los patrones de precios de los viajes en tren con origen o destino en Madrid, utilizando métodos estadísticos descriptivos y comparativos. Se emplea una base de datos obtenida de Kaggle que contiene información sobre los precios de billetes de tren en diferentes trayectos y tipos de trenes. Se presentan dos enfoques: un análisis descriptivo clásico y un modelo basado en regresión lineal múltiple para predecir el precio del billete. Los resultados muestran que los trayectos hacia Barcelona presentan los precios más altos y que el AVE es el tipo de tren más costoso. Finalmente, se discuten las implicaciones de estos hallazgos y se proponen líneas futuras de investigación.

Palabras clave: Análisis descriptivo, regresión lineal, precio, tren.

Introducción y Estado del Tema

El análisis del precio de los billetes de tren es esencial para comprender la dinámica del transporte ferroviario en España. La variabilidad en los precios puede influir en las decisiones de los usuarios y en la competitividad del sector. Algunos estudios previos han demostrado que el análisis estadístico de tarifas y horarios optimiza la asignación de recursos y mejora la planificación del transporte (de Rus & Inglada , 1993; Álvarez & Ramos ,2011). Por otra parte, artículos más recientes nos incentivan descubrir si los precios están ajustados al mercado o se produce competencia desleal, a favor de las empresas contratadas por el Estado, frente a aquellas que ofrecen su servicio de forma privada (de Souza, L. C. ,2021). Otros artículos como *RENFE y la evolución del precio del transporte* (Cuéllar, D., 2024), destacan la evolución de precios de esta compañía, resaltando causas que intentan explicar la variación del precio.

Es posible conocer información acerca de los precios de los billetes por tipo de tren en el *Observatorio del Transporte y la Logística en España*, donde se encuentra el informe anual de 2023 (Ministerio de Transporte y Movilidad Sostenible, 2024). Este es el más reciente del que se dispone, puesto que los informes de un año se publican entre febrero y mayo del año siguiente. En él se recoge el precio por kilómetro de trayecto y en esta métrica el precio del AVE es superior al de todos los trenes de larga distancia, en los que está el ALVIA. El más barato es claramente el Regional, que se encuentra incluido en los trenes de media distancia, y cuyo precio es de algo más de la mitad del de los otros. En este contexto, esta investigación se centra en evaluar los factores que influyen en los precios de los billetes de algunas líneas actuales, considerando que la ciudad de origen/destino es Madrid y el tipo de tren, así como en predecir el precio del billete.

Metodología y Resultados Obtenidos

Para la investigación, se utiliza un conjunto de datos proveniente de Kaggle, que ha sido depurado, eliminando valores nulos y variables irrelevantes. Se aplican dos enfoques: en primer lugar, vamos a hacer un modelo de análisis descriptivo del dataset, cuyo objetivo es determinar si la probabilidad de que un viaje sea caro o no depende de la ciudad origen/destino o del tipo de tren (es decir, el estudio de los Odds - Ratio por ciudad y por tipo de tren). En segundo lugar, vamos a realizar un Modelo de Regresión Lineal Múltiple con la finalidad de predecir el precio del billete de tren en función de las demás variables.

Análisis Descriptivo:

Las variables analizadas son:

- **Ciudad:** Origen o destino del viaje (que no sea Madrid pues siempre es destino u origen)
- **Precio:** Valor numérico del billete. Es una variable cuantitativa y la variable objetivo
- **Tipo_tren:** AVE, ALVIA o Regional.

Las variables del dataset week_day, hour y month no serán incluidas en la sección de descripción, pues no guardan relación con la variación del precio (hecho que se confirmará más adelante).

La variable objetivo, el precio, es una variable cuantitativa continua (si bien es cierto que se redondea a un número concreto con dos decimales). Las variables Origen y Tipo_tren son variables cualitativas. Podemos ver en la *Ilustración 1* un histograma del comportamiento del precio, parámetro que presenta una gran variación. Podemos destacar que hay dos rangos de precios que se suelen repetir a lo largo del dataset, siendo estos, uno alrededor de los 25 euros y otro entre los 75 y 80 euros.

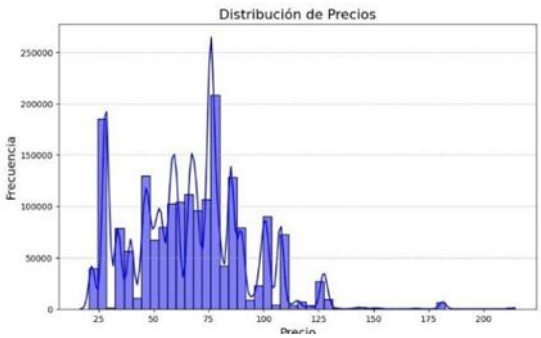

Ilustración 1. Histograma del precio

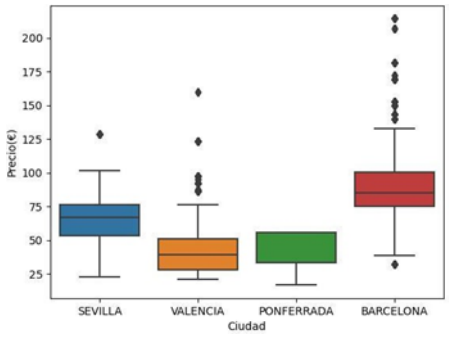

Ilustración 2. Diagrama de cajas del precio según la ciudad

Primeramente, se va a estudiar cómo influye en el precio la ciudad de origen/destino viajando de o desde Madrid, utilizando las medidas de resumen y el diagrama de cajas (ver *Ilustración 2*). Barcelona es la ciudad más cara con un precio mediano de 85.10€, seguida de Sevilla (67,20€), Valencia (39,45€) y Ponferrada (33,50€). Hemos elegido el precio mediano por la gran presencia de atípicos y porque en la ciudad de Ponferrada el primer cuartil de los precios coincide con la mediana, lo que sugiere asimetría en los datos. Valencia, por su parte, tiene los datos desplazados a la izquierda, lo que también refuerza la afirmación.

Análogamente, hemos hecho un análisis descriptivo del precio del tipo de tren. **AVE** es el más caro con un precio mediano de 69.40€, **ALVIA** le sigue con 55.8€ y **Regional** con 28,35€. Observando el histograma en la *Ilustración 3*, podemos ver que los precios de ALVIA tienen la mediana desplazada hacia la derecha y AVE muchos valores atípicos por encima del tercer cuartil, indicando asimetría y justificándose así la elección del precio mediano.

Tras haber hecho los cálculos de Odds Ratios de los precios de los viajes caros respecto a los no caros (suma de precios bajos y normales), hemos obtenido los siguientes valores para cada una de las ciudades. Estableciendo como referencia Barcelona, el Odds Ratio para Sevilla es 0.026, para Valencia 0.0041 (lo que indica asociación negativa media-fuerte por ser cercanas a 0) y para Barcelona 1 (es la referencia). Ponferrada no posee viajes caros y, por lo tanto, su Odds Ratio no se considera. Esto muestra que la probabilidad de que un viaje aleatorio de Barcelona caro es bastante superior a la de las otras ciudades.

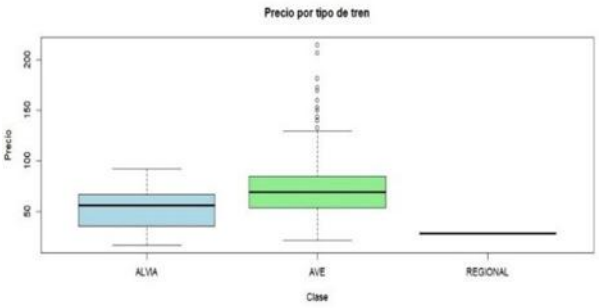

Ilustración 3. Diagrama de cajas del precio según el tipo de tren

Con respecto al tipo de tren hemos calculado también los Odds Ratios de los viajes caros respecto a los no caros, usando como referencia el AVE. Así, se ha obtenido para ALVIA 0.081 (asociación negativa media-fuerte). En el Regional no hay viajes caros, por lo que la comparación no es necesaria. El Odds Ratio del AVE comparado con el Alvia es de 7.80, lo que indica que la probabilidad de que el billete de AVE sea caro es 7.80 veces mayor que la de que el billete de Alvia sea caro.

Modelo de Regresión Lineal Múltiple

En esta sección vamos a analizar si es posible predecir el precio del billete mediante una regresión lineal múltiple. Para ello, lo que se plantea es lo siguiente:

$price = media + a_0\ origin + a_1\ destination + a_2\ train\_type + a_3\ week\_day + a_4\ hour + a_5\ month.$

Para analizar si las variables de entrada tienen relación lineal con el precio, observaremos la matriz de correlación de la *Ilustración 4*. En ella se puede observar que las variables que mayor correlación lineal tienen con el precio son origin y destination, con un coeficiente de -0.45 y -0.46, respectivamente. Es decir, que sugiere una relación de rango medio, y que ciertos orígenes y destinos específicos tienden a estar asociados con precios más bajos. De la misma forma, train_type tiene una relación débil con -0.26, lo que indica que ciertos tipos de tren están asociados con precios más baratos. Así mismo, month, hour y week_day tienen valores muy bajos (0.17, 0.10 y 0.06, respectivamente) lo que sugiere relación inversamente proporcional muy débil entre el mes, día de la semana y la hora con el precio.

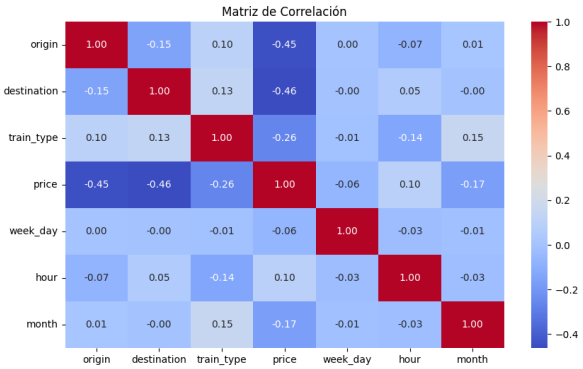

Ilustración 4.Matriz de Correlación

Después de estandarizar las variables y hacer el proceso de One-Hot Encoding, llegamos a que la recta predictora (con los coeficientes estandarizados) de los precios es:

price = 0.0219 hour -0.1430 month + 0.0026 Or_Sevilla -2.1243 Or_Valencia -0.8513 Or_Ponferrada -1.6052 Or_Barcelona - 0.0026 Des_Sevilla -1.7467 Des_Valencia -0.8336 Des_Ponferrada -1.6056 Des_Barcelona -0.1223 AVE + -0.6744 REGIONAL.*

Además, el coeficiente R2R^2R2 es de 0.6136, lo que indica que el modelo explica un 61.36 % de la variación del precio. A partir del análisis de los valores reales y predichos, se observa que existe una amplia dispersión, con un número notable de observaciones alejadas de la tendencia general, lo que sugiere la presencia de valores atípicos. Esto indica que podrían explorarse otros modelos que capten mejor la relación entre las variables y el precio.

Conclusiones y Líneas Futuras

Los resultados confirman que el precio de los viajes en tren depende significativamente de la ciudad de destino y del tipo de tren. Barcelona y el AVE presentan los precios más altos. Además, el análisis de regresión logística refuerza la existencia de una relación entre la ciudad y la probabilidad de que un billete sea caro. Estos resultados se corresponden con los recogidos en el observatorio del transporte y la logística en España, en el informe anual 2023 (Ministerio de Transporte y Movilidad Sostenible, 2024).

Por otro lado, cabe destacar que las variables como “hora”, “día”, y “mes” no afectan a los resultados, por lo que, tanto inicialmente, como tras haberlas eliminado del dataset, los resultados obtenidos no se ven afectados.

Este trabajo abre la posibilidad a su continuación en un futuro, por lo que se van a presentar algunas posibles alternativas a tener en cuenta. Primeramente, el hecho de ajustar aún más el problema, estudiando el precio por kilómetro recorrido (precio / km), para así, dejar de lado el factor distancia que puede alterar en cierta forma el resultado según el origen y el destino se encuentren más o menos lejos. A su vez, se sugiere explorar la influencia de otros factores, como la demanda estacional y las promociones de billetes. Tener en cuenta más variables supone poder ajustar más el modelo a la realidad.

Por último, realizar un estudio con modelos de regresión más avanzados para predecir precios con mayor precisión es algo relevante a tener en cuenta. Por ejemplo, una regresión múltiple polinómica, que, en vez de ser una línea, se ajuste a la forma de la distribución de los datos. También es altamente interesante probar aplicando una red neuronal no demasiado compleja.

Referencias

- de Rus, G., & Inglada, V. (1993). Análisis coste-beneficio del tren de alta velocidad en España. *Revista de Economía Aplicada, 3*(27), 27-48.
- Álvarez, A. G., & Ramos, B. L. (2011). Relación entre el precio básico medio del billete de tren, la velocidad media y la distancia recorrida por el viajero. *Investigación FFE Memoria de artículos, publicaciones y conferencias 2009-2010*, 27
- Cuéllar, D. (2024). RENFE y la evolución del precio del transporte de viajeros durante la segunda mitad del siglo XX. *TST. Transportes, Servicios y Telecomunicaciones*, (54), 38-70.
- de Souza, L. C. (2021). La paradoja de los precios contractuales en mercados anticompetitivos. *Contratación administrativa práctica: revista de la contratación administrativa y de los contratistas*, (175), 46-59.
- Ministerio de Transportes y Movilidad Sostenible. (2023). *Informe anual 2023*. https://otle.transportes.gob.es/inform/es/2023/3competitividad/34-precios-y-costes/346costes-y-precios-en-el-transporteferroviario