CAPÍTULO III

ANÁLISIS EXPLORATORIO DE DATOS A PARTIR DEL ACERCAMIENTO AL OBJETO EN LA REALIDAD

El presente capítulo tiene como propósito analizar los datos disponibles recopilados a través de la aplicación del cuestionario CPSTC17-18, para luego aplicar las técnicas de análisis estadístico multivariante de clúster y correlacional. En la primera parte, se lleva a cabo un análisis exploratorio de los datos y se presenta la información de acuerdo con las variables de estudio para su mayor compresión. En la segunda parte, se analizan de los datos ausentes, en donde se estima la aleatoriedad de la pérdida de los datos y se define que técnicas se emplearán para su imputación.

3.1 Análisis exploratorio por variables de interés y variables de caracterización

En esta sección se examina cada variable implicada en el análisis con relación a los datos obtenidos y se parte de ahí para conocer en detalle las características de distribución de cada una de ellas.

3.1.1. Perfil del turista.

La variable de interés explorada y analizada en esta parte es el perfil del turista y ésta se caracterizó a través de las variables: perfil demográfico y preferencias de viaje, las cuales tienen una serie de indicadores para su medición, como se muestra en la tabla 1.

Tabla 4
Variables e indicadores para el definir el perfil del turista

Variable de interés	Variables de caracterización	Indicadores
Perfil del turista.	Perfil demográfico.	Sexo, edad, país de residencia, ciudad donde vive y situación laboral.
	Preferencias de viaje.	Nombre del crucero, motivo de viaje, viaje en compañía, número de veces viajando en cruceros y duración del crucero.

Nota: Elaboración propia de los autores.

Perfil sociodemográfico

En la tabla 2 se presenta el resumen de los casos válidos y los casos perdidos, así como su participación porcentual para cada indicador.

Tabla 2
Resumen de procesamiento de casos: perfil sociodemográfico

Perfil Sociodemográfico	Casos						
	Válido		Perdidos		Total		
	N	%	N	%	N	%	
Edad	562	95.7%	25	4.3%	587	100.0%	
Sexo	539	91.8%	48	8.2%	587	100.0%	
País de residencia	579	98.6%	8	1.4%	587	100.0%	
Ciudad dónde vive	509	86.7%	78	13.3%	587	100.0%	
Situación laboral	582	99.1%	5	0.9%	587	100.0%	

Nota: Elaboración propia de los autores a través de SPSS.

De los indicadores que conforman la variable perfil demográfico, se tiene que cada uno de ellos presentan casos perdidos, los cuales oscilan entre 0.9% y 13.3%. El indicador situación laboral y edad son los que tienen un porcentaje por debajo del 5% de casos perdidos, lo que estadísticamente no influiría a la hora de realizar inferencias. En el caso de la edad, por tener un comportamiento natural cuantitativo, fue necesario realizar pruebas de normalidad para determinar si los datos tienen una distribución normal. Los resultados de las pruebas de normalidad se presentan en la tabla 3.

Tabla 3
Pruebas de normalidad: edad

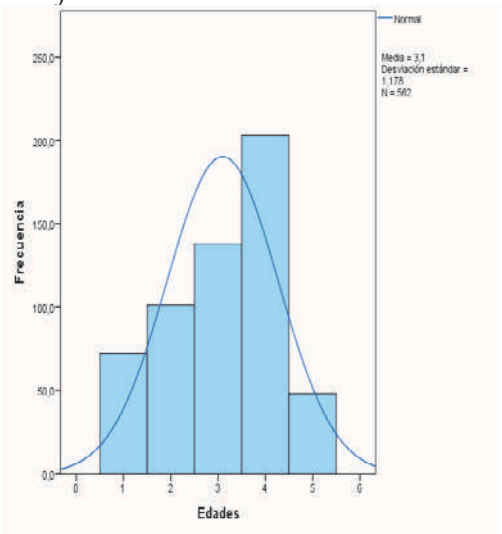
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Edad	.225	562	.000	.893	562	.000

Nota: Elaboración propia de los autores a través de SPSS.

a. Corrección de significación de Lilliefors

De acuerdo con las pruebas de normalidad efectuada, se precisa que la edad no sigue una distribución normal, debido a que se rechaza la hipótesis nula al nivel α cuando p-valor es menor que α, y se acepta en caso contrario (Pérez, 2004). Es decir, los valores de p-valor arrojados por las pruebas de Kolmogorov-Smirnoy y Shapiro-Wilk fue 0.00 y es menor con relación al valor α de 0.05 (nivel significancia predeterminado). Lo anterior, puede constatarse por medio del gráfico 10.

Gráfico 10. Distribución indicador edad.



Elaboración propia de los autores a través de SPSS.

Preferencias de viaje

Como se muestra en la tabla 4, los indicadores que conforman la variable preferencias de viajes tienen un porcentaje de casos válidos entre 90.1% y 96.1% y de casos perdidos entre 3.1% y 9.9%.

Tabla 4
Resumen de procesamiento de casos: preferencias de viaje

Preferencias de viaje	Casos						
	Válido		Perdidos		Total		
	N	%	N	%	N	%	
Nombre del crucero	561	95.6%	26	4.4%	587	100.0%	
Razón o motivo de viaje	563	95.9%	24	4.1%	587	100.0%	
¿Con quién está viajando?	569	96.9%	18	3.1%	587	100.0%	
¿Cuántas veces ha tomado un crucero?	529	90.1%	58	9.9%	587	100.0%	
Duración del crucero	543	92.5%	44	7.5%	587	100.0%	

Nota: Elaboración propia de los autores a través de SPSS.

Como se mencionó anteriormente, los indicadores que representarían inconvenientes al momento de realizar inferencias son aquellos que tienen un porcentaje de casos perdidos menor o igual al 5% y, para este caso, se evidencia que el nombre del crucero, la razón de viaje y la persona o las personas con las que se viaja no tendrían incidencias negativas al momento de analizar el universo de estudio. A los indicadores de tipo cuantitativo, se les realizaron las pruebas respectivas de normalidad para determinar si los datos están distribuidos normalmente. Los resultados se presentan en la tabla 5.

Tabla 5

Pruebas de normalidad: duración y veces que ha tomado un crucero

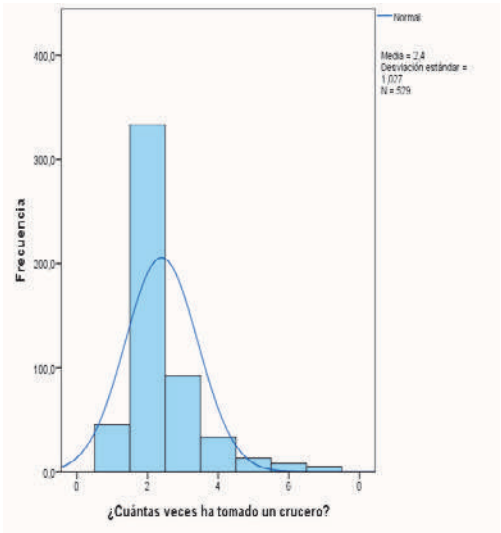
Preferencias de viaje Estadístico	Kolmogorov-Smirnov[a]		Shapiro-Wilk			
	gl	Sig.	Estadístico	gl	Sig.	
¿Cuántas veces ha tomado un crucero?	.364	529	.000	.720	529	.000
Duración del crucero	.340	543	.000	.816	543	.000

Nota: Elaboración propia de los autores a través de SPSS.

a. Corrección de significación de Lilliefors

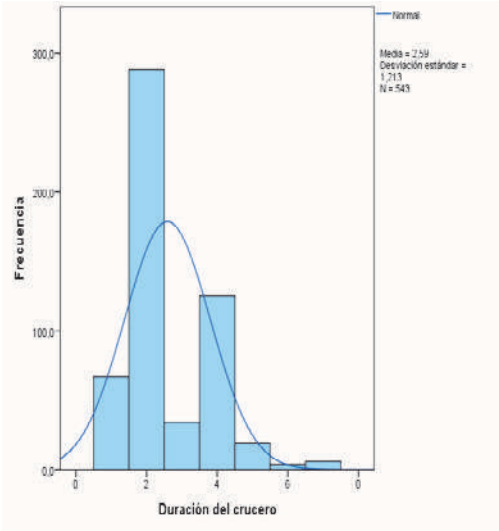
Las dos pruebas de normalidad para los indicadores arrojaron que ninguno de ellos se distribuye normalmente porque p-valor es menor que 0.05, entonces, se rechaza la hipótesis nula de las pruebas. Los gráficos 11 y 12 muestran el comportamiento de la distribución normal para los indicadores, respectivamente.

Gráfico 11. Distribución indicador ¿cuántas veces ha tomado un crucero?



Elaboración propia de los autores a través de SPSS.

Gráfico 12. Distribución indicador duración del crucero.



Elaboración propia de los autores a través de SPSS.

Con los gráficos anteriores se demuestra que no existe simetría entre los datos obtenidos con ambos indicadores, es decir, no se distribuyen normalmente.

3.1.2. Satisfacción del turista.

La variable de interés satisfacción del turista se evaluó teniendo en cuenta tres variables de caracterización: preferencia por el destino, gasto y satisfacción con el destino/visita. Para cada variable se determinaron una serie de indicadores, como se muestra en la tabla 6.

Tabla 6
Variables e indicadores para el medir la satisfacción del turista

Variable de interés	Variables de caracterización	Indicadores
Satisfacción del turista.	Preferencia por el destino.	Visita al destino con anterioridad, planificación de la visita al destino, medio para encontrar la excursión, servicios incluidos en la excursión, actividades realizadas en tierra.
	Gasto.	Gasto en tierra, distribución del gasto.
	Satisfacción con el destino/visita.	Satisfacción con visita al destino, calificación atributos del destino, recomendación del destino, volver a visitar el destino, calificación al destino.

Nota: Elaboración propia.

Preferencias por el destino

De acuerdo con la tabla 7, en los indicadores de tipo cualitativo se evidencia que cada uno contiene casos perdidos. Por ende, el porcentaje de casos válidos oscila entre el 84.7% y 94.4%.

Tabla 7
Resumen de procesamiento de casos: preferencia por el destino – I

Preferencia por el destino	Casos						
	Válido		Perdidos		Total		
	N	%	N	%	N	%	
¿Ha visitado anteriormente Cabo San Lucas?	540	92.0%	47	8.0%	587	100.0%	
¿La visita a Cabo San Lucas fue pre-planeada?	554	94.4%	33	5.6%	587	100.0%	
¿Cómo encontró la excursión?	497	84.7%	90	15.3%	587	100.0%	

Nota: Elaboración propia a través de SPSS.

Teniendo en cuenta que el porcentaje de casos perdidos es mayor del 5%, para cada indicador, entonces, es probable que su interpretación lleve a conclusiones no significativas sobre el universo de estudio. Para el caso de los indicadores de servicios que incluyen la excursión y actividades realizadas en tierra, estos son de tipo cualitativo, pero con múltiples valores finales, por esa razón se agruparon en la tabla 11, en la cual se muestran los casos válidos y perdidos.

Tabla 8

Resumen de procesamiento de casos: preferencia por el destino – II

Preferencia por el destino	Casos					
	Válidos		Perdidos		Total	
	N	%	N	%	N	%
¿Qué servicios incluye la excursión?[a]	432	73.6%	155	26.4%	587	100.0%
Actividades realizadas en tierra.[a]	549	93.5%	38	6.5%	587	100.0%

Nota: Elaboración propia a través de SPSS.

a. Grupo

Para ambos indicadores existen casos perdidos, siendo el de servicios que incluye la excursión el que mayores casos perdidos tiene: un 26,4% contra un 6,5% para las actividades realizadas en tierra. Ambos indicadores pueden llevar a un análisis no significativo del universo del estudio.

Gastos

Otras de las variables estudiadas y que hacen parte de la satisfacción del turista es el gasto, y como se muestra en la tabla 9, para el indicador de gasto en tierra no hubo ningún caso perdido. De acuerdo con lo anterior, las inferencias y los análisis del universo de estudio hechos a partir de este indicador no están sesgados por la falta de información.

Tabla 9
Resumen de procesamiento de casos

Gasto	Casos						
	Válido		Perdidos		Total		
	N	%	N	%	N	%	
Gasto en tierra		587	100.0%	0	0.0%	587	100.0%

Nota: Elaboración propia a través de SPSS.

Los dos indicadores de la variable gasto son de tipo cuantitativo y para el caso del indicador gasto en tierra, los resultados de las pruebas de normalidad se presentan en la tabla 10.

Tabla 10
Pruebas de normalidad: gasto en tierra

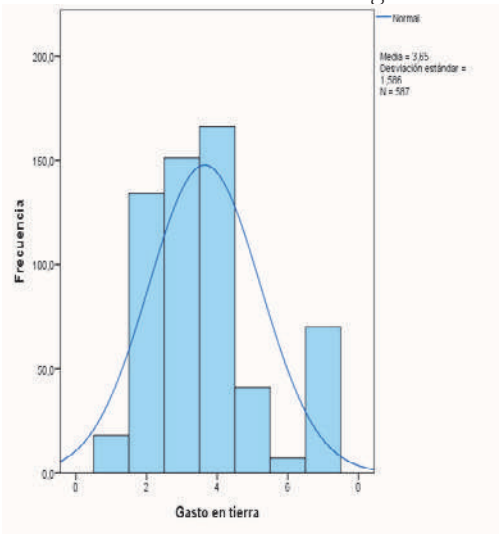
Gasto Estadístico	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	gl	Sig.	Estadístico	gl	Sig.	
Gasto en tierra	.211	587	.000	.877	587	.000

Nota: Fuente: Elaboración propia a través de SPSS.

a. Corrección de significación de Lilliefors

De acuerdo con las pruebas de normalidad, se puede concluir que los datos del indicador gasto en tierra no se distribuyen normalmente, lo cual se visualiza en el gráfico 13.

Gráfico 13. Distribución indicador gasto en tierra.



Elaboración propia a través de SPSS.

En el anterior gráfico se evidencia que los datos para el indicador gasto en tierra no siguen una distribución normal. Aparte de lo anterior, en la tabla 11, se resumen los casos para el indicador distribución del gasto, que es de tipo cuantitativo, pero con múltiples valores finales.

Tabla 11

Resumen de procesamiento de casos: distribución del gasto

Gasto	Casos					
	Válidos		Perdidos		Total	
	N	%	N	%	N	%
Distribución del gasto.[a]	482	82.1%	105	17.9%	587	100.0%

Nota: Elaboración propia de los autores a través de SPSS.

a. Grupo

Para el indicador gasto en tierra se tiene un 17.9% de casos perdidos y, al igual que con los demás indicadores con casos perdidos, un porcentaje mayor a 5% puede llevar a inferencias equívocas del universo del estudio. Como este indicador tiene múltiples valores finales, no fue posible realizar las pruebas de normalidad.

Satisfacción con el destino/visita

La variable satisfacción del destino/visita es la última que conforma el presente estudio y es evaluada por medio de cinco indicadores, de los cuales dos son de tipo cuantitativo y uno de ellos con múltiples valores finales. Los resúmenes de los casos se muestran en la tabla 12.

Tabla 12

Resumen de procesamiento de casos: satisfacción con el destino/visita

Satisfacción con el destino/visita	Casos						
	Válido		Perdidos		Total		
	N	%	N	%	N	%	
La experiencia en Cabo San Lucas fue	587	100.0%	0	0.0%	587	100.0%	
¿Recomendaría este destino?	572	97.4%	15	2.6%	587	100.0%	
¿Volvería a Cabo San Lucas?	587	100.0%	0	0.0%	587	100.0%	
Nivel de satisfacción con Cabo San Lucas	587	100.0%	0	0.0%	587	100.0%	

Nota: Elaboración propia de los autores a través de SPSS.

De los indicadores que hacen parte de la variable mencionada, el segundo de ellos, ¿recomendaría este destino? es el único que tiene un porcentaje de casos perdidos de 2,6% y como es menor que el 5% estadísticamente aceptable, el análisis del universo de estudio efectuado a partir de él, no estaría sesgado por la falta de información. Uno de los indicadores cuantitativos es el nivel de satisfacción y de acuerdo con eso, los resultados de las pruebas de normalidad se presentan en la tabla 13.

Tabla 13
Pruebas de normalidad: nivel de satisfacción

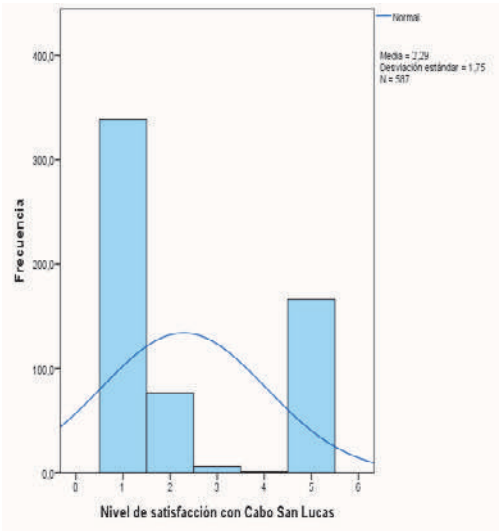
Satisfacción con el destino/visita Estadístico	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	gl	Sig.	Estadístico	gl	Sig.	
Nivel de satisfacción con Cabo San Lucas	.345	587	.000	.654	587	.000

Nota: Elaboración propia de los autores a través de SPSS.

a. Corrección de significación de Lilliefors

De acuerdo con los resultados de las dos pruebas, se tiene que los datos del indicador nivel de satisfacción no se distribuyen normalmente. En el gráfico 14 se muestra la asimetría en la distribución de los datos.

Gráfico 14. Distribución indicador nivel de satisfacción con Cabo San Lucas.



Elaboración propia de los autores a través de SPSS.

El indicador de calificación del destino es de múltiples valores finales y, por esa razón, no se realizaron las pruebas de normalidad. Sin embargo, el resumen de casos se presenta en la tabla 14.

Tabla 14
Resumen de procesamiento de casos: calificación atributos del destino

Satisfacción con el destino/visita	Casos					
	Válidos		Perdidos		Total	
	N	%	N	%	N	%
Calificación a tributos del destino.[a]	555	94.5%	32	5.5%	587	100.0%

Nota: Elaboración propia de los autores a través de SPSS.

a. Grupo

El indicador de calificación del destino posee un 5,5% de valores perdidos por lo que estadísticamente puede sesgar el análisis que se efectúe del universo de estudio.

3.2. Análisis de datos ausentes

A partir del análisis exploratorio de datos, se pudo evidenciar la presencia de datos ausentes y en relación con ellos, se debe determinar el efecto que tienen y si se distribuyen aleatoriamente; por eso, es necesario realizar pruebas formales de aleatoriedad (Pérez, 2004). La proporción de los datos perdidos se presenta en la tabla 15.

Tabla15
Relación de indicadores con datos perdidos

	N	Media	Desviación estándar	Perdidos		Número de extremosa	
				Recuento	Porcentaje	Menor	Mayor
Sexo	539	1.47	.500	48	8.2	0	0
País	579	26.05	7.263	8	1.4	.	.
Ciudad	509	7.37	4.430	78	13.3	0	0
Sitlab	582	2.11	.882	5	.9	.	.
Nomcru	561	3.37	2.648	26	4.4	0	0
Motvia	563	1.97	1,215	24	4.1	0	0
Viacom	569	2.62	.883	18	3.1	0	15
Medencexc	497	1.99	1.266	90	15.3	0	0
Planvisdes	554	1.22	.430	33	5.6	.	.
Visdesant	540	1.57	.662	47	8.0	0	6

	N	Media	Desviación estándar	Perdidos		Número de extremosa	
				Recuento	Porcentaje	Menor	Mayor
Calatrdes1	514	1.25	.567	73	12.4	.	.
Calatrdes2	542	1.28	.555	45	7.7	.	.
Calatrdes3	541	1.26	.572	46	7.8	.	.
Calatrdes4	532	1.39	.674	55	9.4	0	4
Calatrdes5	521	1.32	.626	66	11.2	.	.
Calatrdes6	520	1.49	.750	67	11.4	0	11
Calatrdes7	451	1.27	.558	136	23.2	.	.
Calatrdes8	441	1.37	.718	146	24.9	0	11
Calatrdes9	432	1.30	.595	155	26.4	.	.
Calatrdes10	441	1.26	.570	146	24.9	.	.
alatrdes11	449	1.25	.577	138	23.5	.	.
Calatrdes12	495	1.32	.602	92	15.7	.	.
Calatrdes13	495	1.28	.559	92	15.7	.	.
Calatrdes14	485	1.30	.615	102	17.4	.	.
Calatrdes15	501	1.33	.647	86	14.7	0	8
Calatrdes16	479	1.28	.581	108	18.4	.	.
Calatrdes17	494	1.33	.613	93	15.8	0	2
Calatrdes18	484	1.35	.627	103	17.5	0	4
Recdes	572	1.43	1.305	15	2.6	.	.
Edad	562			25	4.3		
Numvecviacru	529			58	9.9		
Dracru	543			44	7.5		

Nota: Elaboración propia de los autores a través de SPSS.

a. Número de casos fuera del rango (Q1 - 1,5*IQR, Q3 + 1,5*IQR).

Los datos perdidos pueden clasificarse de acuerdo con la relación existente entre los datos perdidos y los datos, y a partir de ahí surgen los siguientes tres supuestos: primero, datos perdidos complemente al azar (MCAR, por sus siglas en inglés) cuando los datos no se relacionan con ninguna variable presente; el segundo, datos perdido al azar (MAR, por sus siglas en inglés) cuando se supone que la pérdida de los datos obedece a una razón predecible; y el tercero, datos perdidos no aleatorios (NMAR, por sus siglas en inglés) cuando los datos perdidos subyacen en la variable en sí misma (Montegro, Oh, & Chesnut, 2015; Rubin, 1976; Little, Jorgesen, Lang, &

Moore, 2014). Dicho lo anterior, se procedió con el análisis de los valores perdidos y se utilizó la estimación EM (expectation-maximization) con la cual se hace una evaluación de los valores perdidos mediante un proceso iterativo. A su vez, cada iteración tiene un paso E, en el cual se calculan los valores esperados y un paso M, para las estimaciones verosímiles.

Con la estimación EM, entre otros estadísticos, se obtiene el estadístico de Little, el cual sigue una distribución X^2 (ji-cuadrada) con f grados de libertad y tiene como hipótesis nula (Ho) que los datos perdidos siguen un patrón MAR. De acuerdo con la regla de decisión, se rechaza Ho si el valor del estadístico para los datos analizados es menor conforme al nivel de significancia (α) (Medina & Galván, 2007). Al realizar la prueba de Little, los indicadores presentan una p = 0.000; por lo tanto, como p-valor es menor que 0.05 se asume que los datos no están perdidos completamente al azar, sino al azar.

3.2.1.Imputación de datos ausentes.

Constatada la aleatoriedad de los datos ausentes, se toma la decisión de imputar la información que falta para luego iniciar con el análisis estadístico. De acuerdo con Pérez (2004), la imputación es un proceso de estimación de valores para los datos ausentes que se basa en los casos válidos de la muestra. Para efectos de esta investigación, se utiliza el método de imputación múltiple y según Medina y Galván (2007), este método emplea la simulación de Monte Carlo y reemplaza los datos ausentes a partir de un número de simulaciones. En cada simulación se analiza la matriz de los datos con ayuda de los métodos estadísticos convencionales, para luego combinar los datos resultantes y generar estimadores robustos, error estándar e intervalos de confianza. En la tabla 16 se presentan las especificaciones de la imputación realizada.

Tabla 16
Especificaciones de la imputación de datos

Método de imputación	Automático
Número de imputaciones	3
Modelo para variables de escala	Regresión lineal
Interacciones incluidas en modelos	(ninguno)
Porcentaje máximo de valores perdidos	100.0%
Número máximo de parámetros en modelo de imputación	10000

Nota: Elaboración propia a través de SPSS.

Tabla 17
Modelo de imputación para cada indicador

	Valores perdidos	Valores imputados
	Tipo	
Situación laboral	Regresión logística	5
País de residencia	Regresión logística	8
¿Recomendaría este destino?	Regresión logística	15
¿Con quién está viajando?	Regresión logística	18
Razón o motivo de viaje	Regresión logística	24
Edad	Regresión lineal	25
Nombre del crucero	Regresión logística	26
¿La visita a Cabo San Lucas fue pre-planeada?	Regresión logística	33
Duración del crucero	Regresión lineal	44
Gente local	Regresión logística	45
Bienvenida en el puerto	Regresión logística	46
¿Ha visitado anteriormente Cabo San Lucas?	Regresión logística	47
Sexo	Regresión logística	48
Señalización	Regresión logística	55
¿Cuántas veces ha tomado un crucero?	Regresión lineal	58
Información turística	Regresión logística	66
Calidad-precio	Regresión logística	67

	Valores perdidos Tipo	Valores imputados
Destino	Regresión logística	73
Ciudad dónde vive	Regresión logística	78
Instalaciones del puerto	Regresión logística	86
¿Cómo encontró la excursión?	Regresión logística	90
Limpieza de la ciudad	Regresión logística	92
Seguridad	Regresión logística	92
Variedad de locales comerciales	Regresión logística	93
Atractivos	Regresión logística	102
Experiencia de compra	Regresión logística	103
Tiempo de espera en el puerto	Regresión logística	108
Playas	Regresión logística	136
Guías turísticos	Regresión logística	138
Transporte en el lugar	Regresión logística	146
Excursiones	Regresión logística	146
Centro de la ciudad	Regresión logística	155

Nota: Elaboración propia de los autores a través de SPSS.

Como se muestra en la tabla 17, el modelo de imputación fue automático y, por ende, el programa de forma automatizada asignaba a las variables cuantitativas el método de imputación por regresión lineal y a las variables cualitativas el método de imputación por regresión logística. El resumen del método empleado y los valores imputados para cada indicador se presentan en la tabla 20. Se realizaron tres imputaciones, es decir, el programa arrojó tres opciones de imputación de los datos, por eso, teniendo en cuenta que la media y demás estadísticos descriptivos tienen un comportamiento más o menos similar a los casos originales, se escogió la imputación número uno para realizar el análisis estadístico.