Tarea 2

Autor

Juan Román Uribe

Sebastián Mena Aliaga

Profesor

Christian Araya Muñoz

Modelación Estadística Aplicaciones Multidisciplinaria

Magister en Estadística, PUCV

Agosto 2021

1 Pregunta 1. AED

El Análisis Exploratorio de Datos (AED) se realiza en base a las variables escogidas de Sexo, Rango de edad (llamado Rango para efectos prácticos), Frecuencia de visita a la clínica (llamado Frecuencia para efectos prácticos) y las variables de tipo Likert X5, X8, X14 y X18.

La base de datos cuenta con 320 registros. Tal como se observa en la tabla resumen 1, el 54.7% de los encuestados son mujeres y el resto hombres, mayoritariamente (56.2%) entre un rango de edad de 30 a 60 años, y asisten principalmente (51.5%) una vez al mes a la clínica, además, la variable de frecuencia cuenta con un NA, dato mal registrado que no pudo ser agrupado en alguna de las 4 categorías.

Sexo	Frecuencia
Femenino	176
Masculino	144

Rango	Frecuencia
<18	48
[18, 30[	53
[30, 60[	180
>= 60	39

Frecuencia Visita	Frecuencia
Primera visita	37
Una vez al año	89
Una vez al mes	165
Una vez por semana	28
NA's	1

Table 1: Resumen de Sexo, Rango y Frecuencia

Se observa del gráfico de barras Sexo/Rango, a la izquierda de la figura 1 que, mayoritariamente tanto mujeres como hombres encuestados tienen un rango de edad entre 30 y 60 años, para los demás rangos de edad se observa más o menos equiparada la frecuencia de consulta. Respecto al gráfico Rango/Frecuencia (derecha de la figura), se observa una clara preferencia entre personas de 30 y 60 años, y más de 60 años, a asistir una vez al mes a la clínica, por otro lado, mayoritariamente el rango entre 18 y 30 años de los encuestados afirma asistir una vez al año.

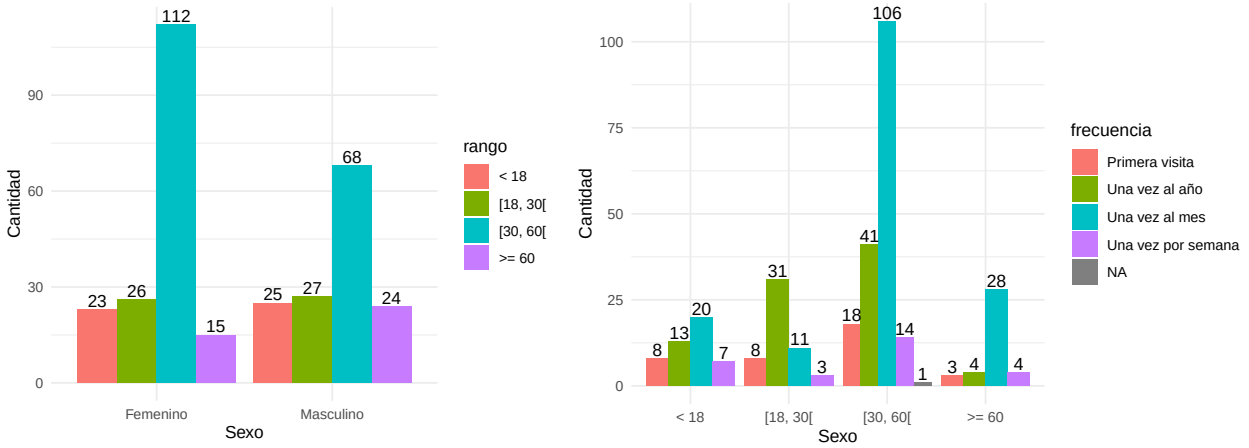

Figure 1: Gráfico de barras de Sexo/Rango y Rango/Frecuencia

Al comenzar el análisis de las variables X5, X8, X14 y X18, se observa una alta presencia de valores perdidos con, 13, 10, 96 y 75 NA's registrados respectivamente. Para explorar los datos likerts, se opta por el gráfico de barras de likert y, tal como se observa en la figura 2, se apertura por rango de edad, dado que al tratarse de una encuesta de calidad de servicio de la clínica, se hace interesante analizar las respuesta de los clientes según su grupo etario, entendiendo la posibilidad que personas de un mismo grupo etario responden de forma similar a las preguntas realizadas. Desde una perspectiva general, se observan puntuaciones bastante altas para las 4 variables, siendo superior al 60% las puntuaciones positivas, entendiendo al 5 como centro neutro.

De la variable X5, sobre la gama de convenios y descuento que ofrece la clínica, son los grupos etarios entre 18 a 30 y, 30 a 60 años, en demostrarse un poco más disconformes respecto a este punto. De la variable X8, el grupo de mayores de 60 años afirman sentirse más satisfechos con la calidad de las prestaciones respecto al precio que pagan, y los menores de 18 años se observan un poco más neutrales. Respecto a la variable X14, se observa un claro desacuerdo entre los grupos entre 18 a 30 y, 30 a 60 años, sobre la cobertura para afiliados a FONASA, siendo este el item de menor puntuación entre las 4 variables. De X18, contra intuitivamente se observa que, además de los menores de 18 años, los mayores a 60 consideran que la *website* de la clínica esta bien constituida, y, un desacuerdo un poco más pronunciado para el grupo entre 18 a 30 años sobre el mismo punto, sin embargo, en general las puntuaciones son bastante altas para este item respecto a las otras variables.

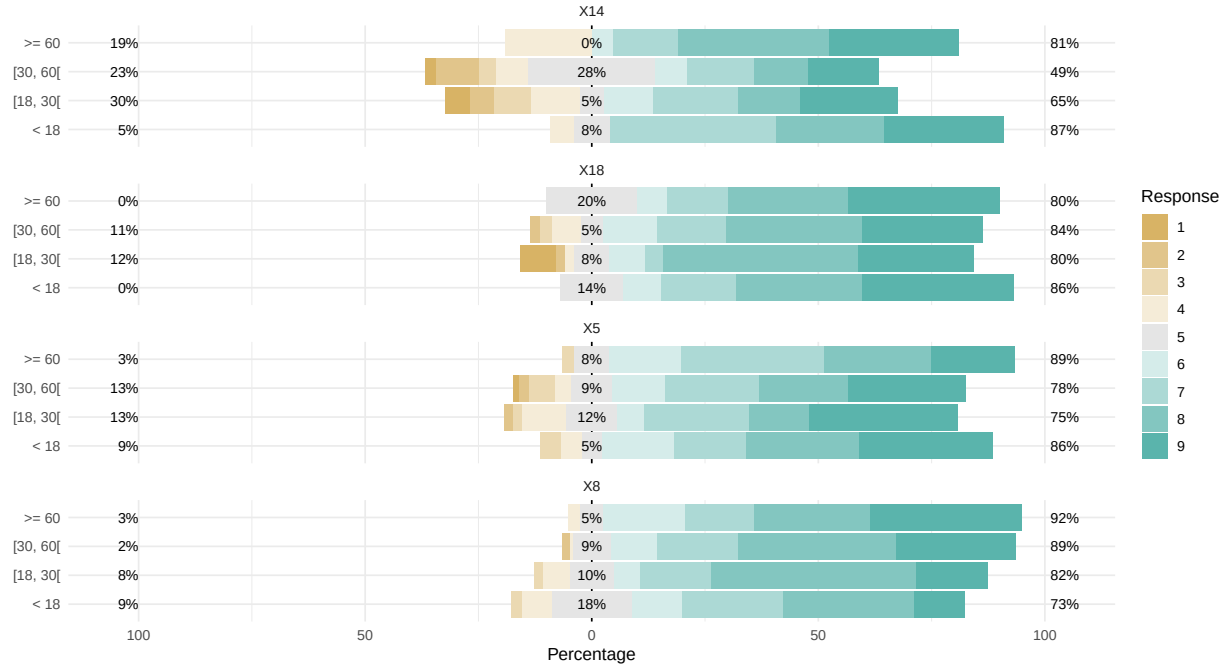

Figure 2: Gráfico barras de Likert aperturado por rango de edad

2 Pregunta 2

Tal como se señala en la pregunta 1 y se observa a la izquierda de la figura 3, las variables de tipo likert X5, X8, x14 Y x18, las variables cuenta con 13 (4%), 10 (3%), 96 (30%) y 75 (23.4%) NA's registrados respectivamente, además, se puede observar 1 variable perdida en frecuencia (0.3%). Observando las combinaciones de datos perdidos de las variables X5, X8, X14, X18, frecuencia, sexo y rango, a l lado derecho de la figura 3, se puede apreciar que aproximadamente el 57% de los datos no posee valores perdidos, o dicho de otro modo, el 46% de los datos posee al menos una variable perdida entre sus registros. El 14% posee solo X14 como valor perdido, el 13.4% posee la combinación de X14 y X18 como perdidos en sus registros, y el 9.4% de los registro tiene a X18 como perdidos.

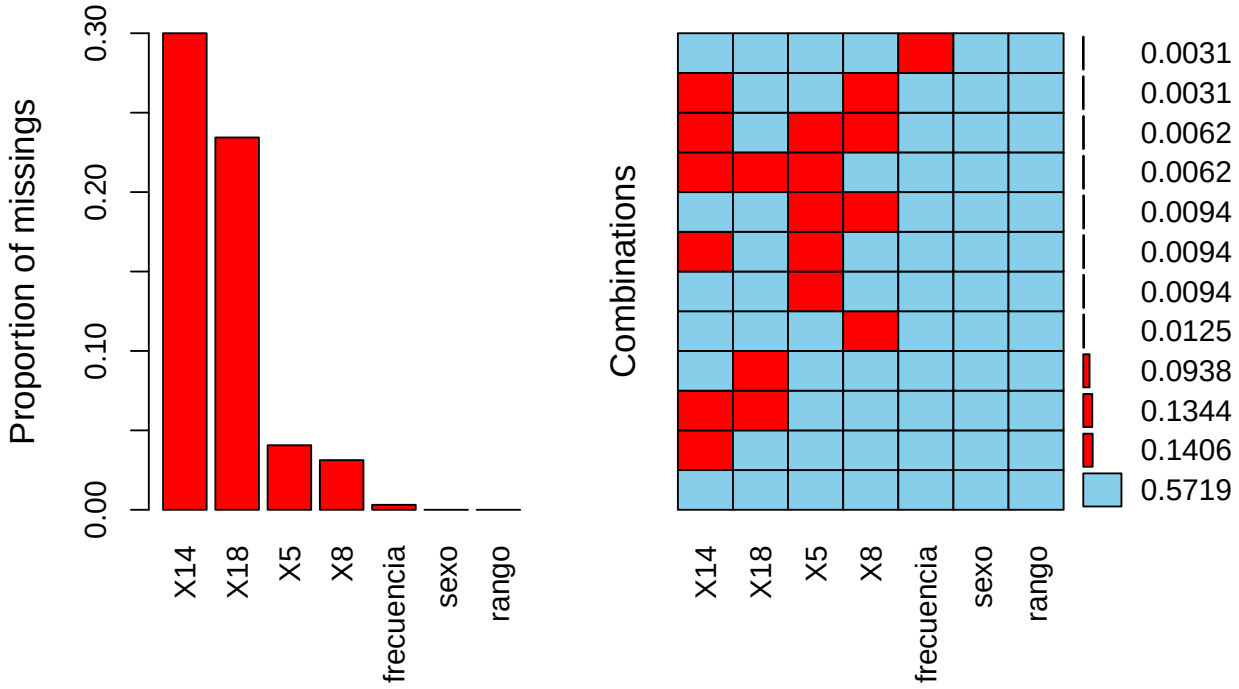

Figure 3: Gráfico agregado de valores perdidos

Luego de observar las combinaciones de datos perdidos entre las variables likerts, es natural sospechar si las perdidas de una variable depende de la presencia o ausencia de las otras variables implicadas, o si es debido netamente a procesos aleatorios. Es por ello que se definen tipos de observaciones perdidas: Las variables perdidas son dadas por conceptos completamentes del azar al momento de recolectar el dato *Missing Completely at Random*, MCAR), al darse MCAR cualquier observación tiene igual probabilidad de perderse; Los datos faltantes pueden estar condicionadas por otras variables explicativas del conjunto (*Missing at Random*, MAR) osea, MAR no se distribuyen al azar para un subconjunto de variables; Los datos perdidos dependen completamente de la presencia o ausencia de otra variable (*Missing Not at Random*, MNAR), no hay azar en la ausencia del dato (Mandeville, 2010).

2.1 Impacto en la distribución de X5 por ausencia de variables

Se estudiará el impacto que tienen la presencia o ausencia de las variables X8, X14 y X18 en la distribución de X5, con la finalidad de establecer si la falta de X8, X14 o X18 es explicada por algún comportamiento de X5 o si se debe solo por azar, en otras palabras, se busca evidencia para sustentar si hay mecanismos MCAR o MAR en dichas variables.

Al analizar la figura 4, que contiene *boxplots* pareados de X5 con información de datos faltantes e imputados en X8, X14 o X18 se observa, en primer lugar, que la presencia de X8, X14 y X18 (*boxplots* azules) siguen distribuciones muy similares al total de X5 (*boxplot* blanco). Al observar los casos con variables perdidas en X8, X14 y X18 (*boxplots* rojos), para X14 se observa que a respuestas más altas de X5 es más probable la ausencia de X14 implicando la posibilidad de MAR, en el caso de X18 y los pocos datos en X8, se observa una distribución más similar a la de contraste por lo que se puede señalar un mecanismo MAR para estas dos variables. Es prudente señalar que para un análisis estadístico más detallado se hace necesario realizar test de hipótesis que contrasten la diferencia entre las medidas de posición entre los grupos para fortalecer las conclusiones de MCAR y MAR, sin embargo, dicho estudio queda fuera de los alcances en la realización de este trabajo.

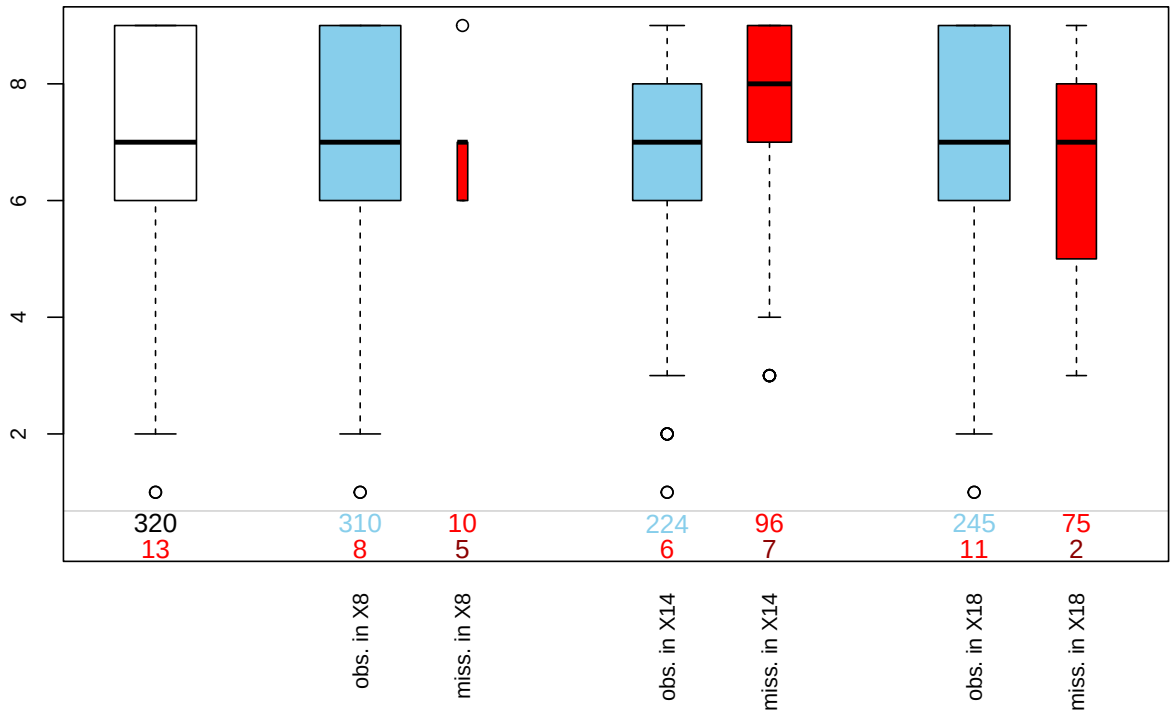

Figure 4: *Boxplots* pareados con valores perdidos/imputados en relación a X5

2.2 Impacto en la distribución de X8 por ausencia de variables

Al observar la figura 5, los grupos de *boxplots* de X8 donde existen variables perdidas en X5, X14 y X18 (rojo), si bien presentan una misma mediana, los cuartiles se notan ligeramente diferentes en contraste con el *boxplot* completo de X8 (blanco), sin embargo, visualmente no se aprecia una diferencia significativa para sustentar la posibilidad de MAR, por lo que se establece las perdidas tipo MCAR, osea, que las perdidas de X8, X14 y X18 son completamente por razones de azar.

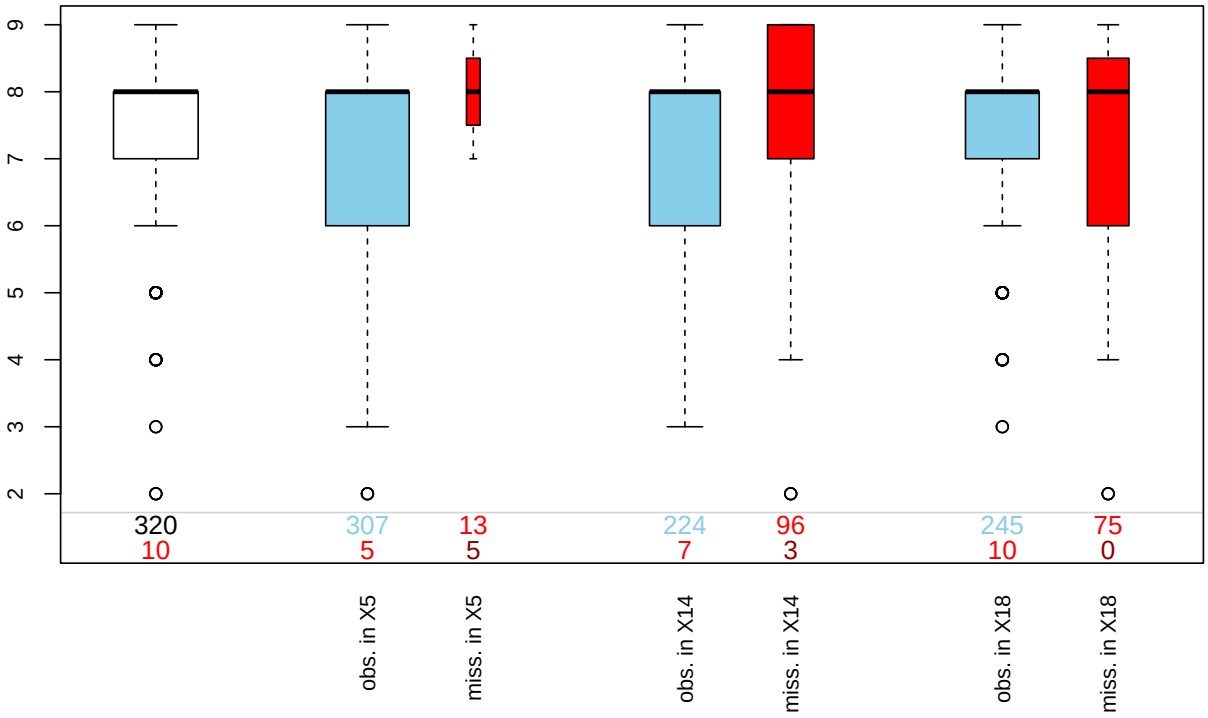

Figure 5: *Boxplots* pareados con valores perdidos/imputados en relación a X8

2.3 Impacto en la distribución de X14 por ausencia de variables

Respecto a la muestra de X14, se puede observar en la figura 6 claras diferencias significativas en los grupos de X14 con ausencia en X5, X8 y X18, en el caso del grupo con ausencia en X8 se puede concluir la posibilidad de MAR en X14 y X18 cuando la puntuación de la encuesta es inferior a la mediana de X8. Por otro lado, en el grupo con ausencia en X5 se puede observar un comportamiento MAR cuando la puntuación es superior a la mediana de X14. Cabe destacar que en los grupos con ausencia en X5 y X8 poseen muestras de 10 y 13 observaciones, valores bastante bajos, por lo que se hace necesario realiza un test no perimétrico para evaluar las conclusiones anteriores.

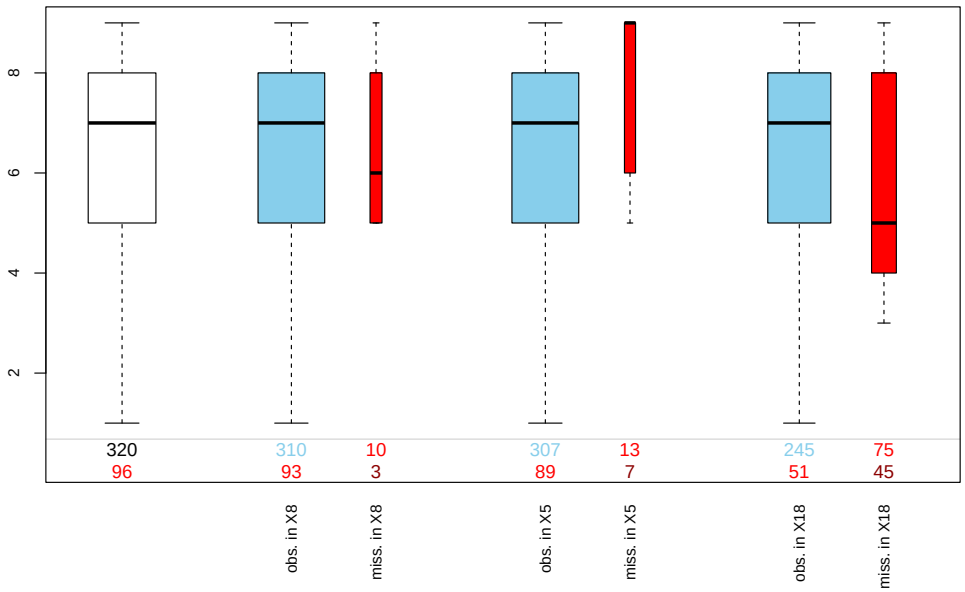

Figure 6: *Boxplots* pareados con valores perdidos/imputados en relación a X14

2.4 Impacto en la distribución de X18 por ausencia de variables

Para el caso de X18, al observar al figura 7, da un caso muy similar en X8 con medianas idénticas entre grupos pero con diferencias en sus cuartiles, indicando un probable comportamiento MCAR para las variables X5, X8 y X14, concluyendo que la ausencia de estas variables respecto a X18 serían completamente al azar.

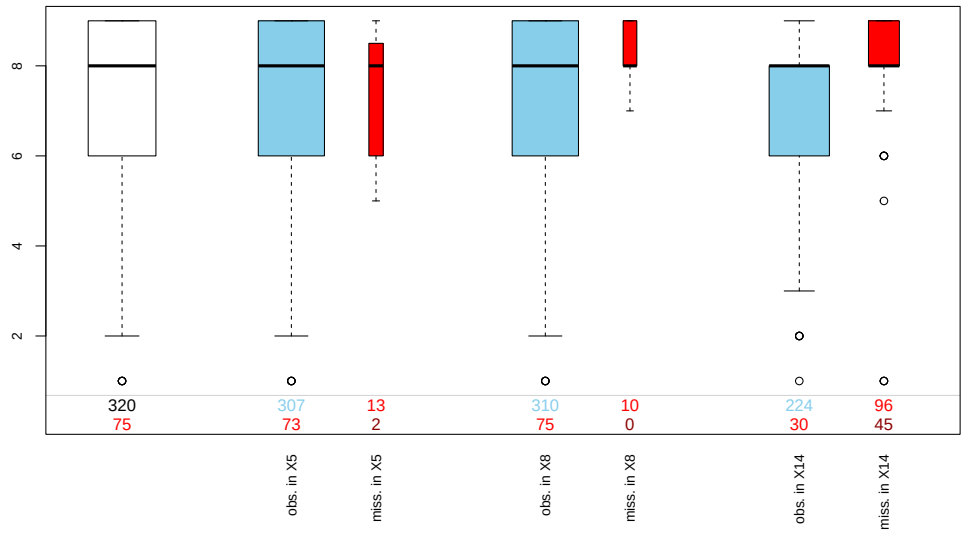

Figure 7: *Boxplots* pareados con valores perdidos/imputados en relación a X18

3 Pregunta 3

Los métodos de imputación de datos son una tarea esencial cuando se desea realizar inferencia sobre los mismos pero la ausencia de estos datos puede provocar ruido al momento de su análisis, sin embargo, sería imprudente aplicar un método de imputación inapropiado, dado que estos podrían generar más problemas de los que intenta resolver, induciendo a sesgos y en el peor de los casos provocando conclusiones totalmente erróneas en la investigación (Medina, Galván 2007). Dicho lo anterior, se realizarán las técnicas de imputación de datos: a partir de la mediana, mediana condicionada por grupos, de tipo *hot deck* y de regresión lineal, con la finalidad de poner a prueba el rendimiento de cada técnica y escoger aquella que sopese la menor perdida en al calidad de la muestra para cada variable.

3.1 Imputación por mediana

Es conocido como un método de imputación simple y a su vez muy poco apropiada dado que se asume que todos los datos siguen un patrón MCAR. Para las variables 5X, 8X, X14 y X18 se reemplazan por los valores de sus medianas: 7, 8, 7 y 8 respectivamente. En los diagramas de 8 a 11 (en color rojo antes de imputar y color azul luego de imputar) se observa que, para la variables X5 y X8 (figuras 8 y 9) la distribución no ha cambiado lo suficiente luego de imputar, esto se debe principalmente a que ambas variables presentaban pocos casos de datos perdidos. Para las variables X14 y X18 (figuras 10 y 11) se observa claramente la falencia de este método al concentrar la imputación en un solo valor, la mediana.

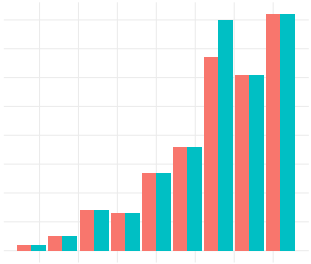

Figure 8: X5 c/s imputación por mediana

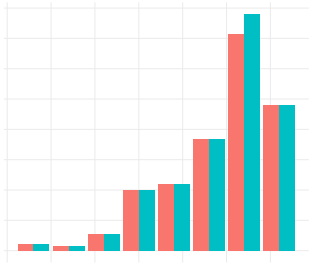

Figure 9: X8 c/s imputación por mediana

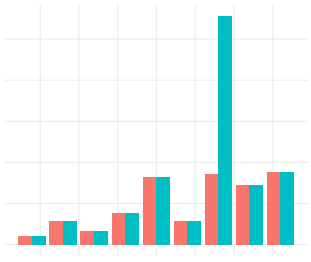

Figure 10: X14 c/s imputación por mediana

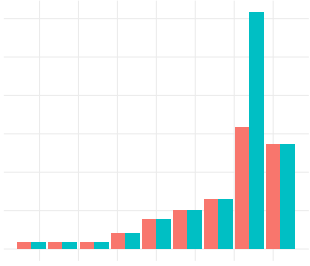

Figure 11: X18 c/s imputación por mediana

3.2 Imputación por mediana condicionada por frecuencia

A modo de sofisticar el método de imputación simple de la mediana, se aplica el método agrupando los datos por la variable de frecuencia de visita, este método permite evitar la concentración de la imputación en un solo valor. Luego de imputar, al observar los diagramas de 12 a 15 en comparación con los diagramas anteriores, se observar que hubo un claro cambio en la distribución de las gráficos 12 y 14, en este último además se observa que los datos se han logrado redistribuir mejor respecto al los datos sin imputar.

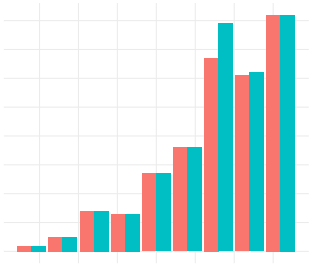

Figure 12: X5 c/s imputación por mediana agrupado

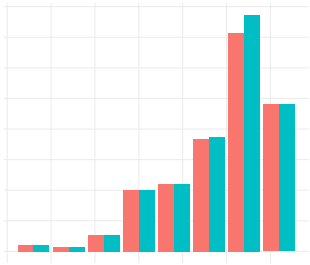

Figure 13: X8 c/s imputación por mediana agrupado

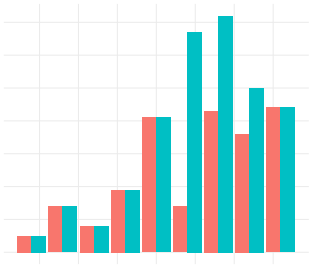

Figure 14: X14 c/s imputación por mediana agrupado

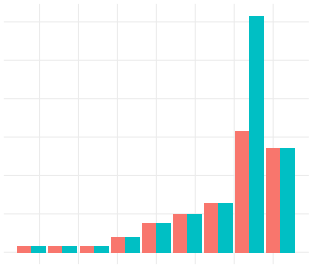

Figure 15: X18 c/s imputación por mediana agrupado

3.3 Método *hot deck*

Este método asigna valores a los datos faltantes con la información existente en la muestra de acuerdo a la celda en la que se encuentra la observación con información faltante. El procedimiento consiste en completar en cada celda las observaciones faltantes utilizando datos de la misma celda, los cuales son seleccionados al azar (Alfaro, Fuenzalida 2009).

Se puede observar de los gráficos del 16 a 19 que las distribuciones según el método se han redistribuidos de mejor forma en comparación a los dos métodos anteriores, obteniendo datos más acordes a su distribución original, siendo el caso más concreto de la variable X14, donde se observa que el crecimiento de cada puntuación de la escala ha sido proporcional luego de imputar los datos faltantes.

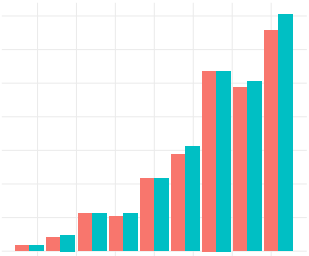

Figure 16: X5 imputador por *hot deck*

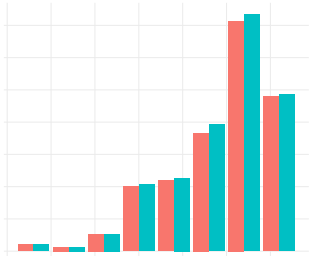

Figure 17: X8 imputador por *hot deck*

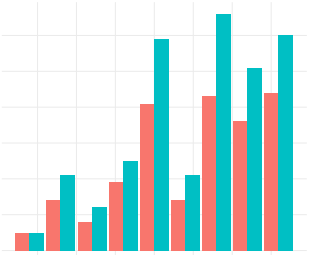

Figure 18: X14 imputador por *hot deck*

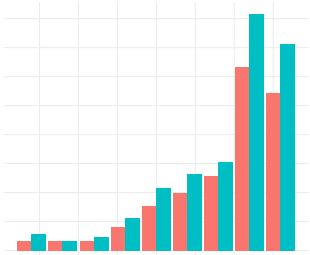

Figure 19: X18 imputador por *hot deck*

3.4 Método por regresión lineal

En la imputación por regresión, los valores faltantes son reemplazado por los valores predichos por el modelo. Dado que el método considera el cálculo a partir de una clásica regresión, se definen como las variables explicadas (la variable queremos predecir e imputar) como X5, X8, X14 y X18, y se define la variable explicativa como las 3 últimas preguntas tipo likert de la encuesta, que son denominadas como A, B y C en la base de datos. Se ha observado que la variable C contiene un valor perdido que ha provocado que para las variables X5, X14 y X18 no se haya podido predecir (imputar) su valor, sin embargo, al tratarse de un solo dato de una muestra grande, se decide no considerar dicho registro en el análisis. Cabe destacar que el método por regresión lineal genera valores decimales, por ello, dado que se trata de imputar una variable tipo likert, se ha decidido que el valor imputador por el método se aproximará a su entero más cercano.

Tal como se observa en las figuras del 20 al 23, la imputación de datos a partir del método por regresión lineal, tomando como variable explicativa a A, B y C, han provocado concentrar los valores al igual que en el caso del método más simple de imputación por la mediana (ver figuras 10 y 11). Sin embargo, solo se ha iterado a partir de las variables anteriormente señaladas, además, en este caso no se ha probado si dichas variables son buenas explicando las variables X5, X8, X14 y X18, análisis importante que se debe realizar a la hora de escoger las variables de una regresión.

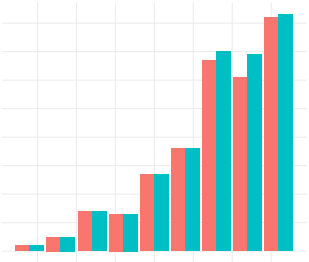

Figure 20: X5 imputados por regresión

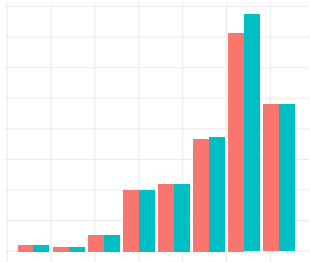

Figure 21: X8 imputador por regresión

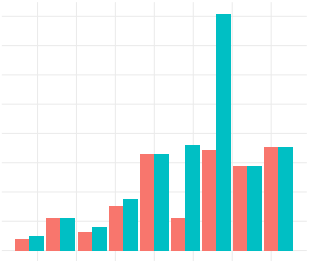

Figure 22: X14 imputador por regresión

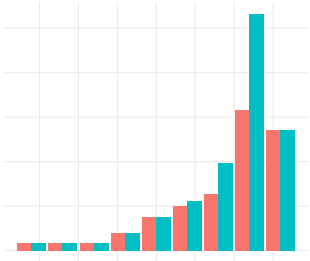

Figure 23: X18 imputador por regresión

3.5 Impresiones

Luego de probar los métodos a partir de la mediana, mediana condicionada por grupos, de tipo *hot deck* y de regresión lineal, se concluye para este caso que el método *hot deck* ha demostrado tener el mejor rendimiento a la hora de imputar los datos, permitiendo obtener una distribución coherente antes y después de la imputación. Cabe señalar que, además de las consideraciones que no se profundizaron para esta pregunta (por ejemplo, test de hipótesis para contrastar las distribuciones antes y después), también existen muchos métodos más y la utilización de un método u otro dependerá del contexto de los datos y estudio realizado.

4 Pregunta 4

Otro forma de analizar el efecto que tienen los métodos de imputación sobre la muestra es analizando la correlación entre variables antes y después de imputar, es de interés para el investigador que el método escogido posea la menor diferencia significativas entre las correlaciones, de lo contrario se estaría en presencia de un mal método de imputación.

4.1 Datos sin imputar

Se obtienen las correlaciones de los datos sin imputar de las variables X5, X8, X14 y X18, eliminando los registros donde se obtuvieron por lo menos un dato perdido, de este se puede observar en la figura 24 que todas las variables poseen una correlación positiva entre ellas y diferentes a cero, evidencia que se fortalece al realizar el test de correlación de Spearman donde se rechaza la hipótesis nula que las correlaciones sean igual a cero (ver tabla 2).

	X5	**X8**	**X14**	**X18**
X5	1	0.303	0.483	0.389
X8	0.303	1	0.256	0.201
X14	0.483	0.256	1	0.421
X18	0.389	0.201	0.421	1

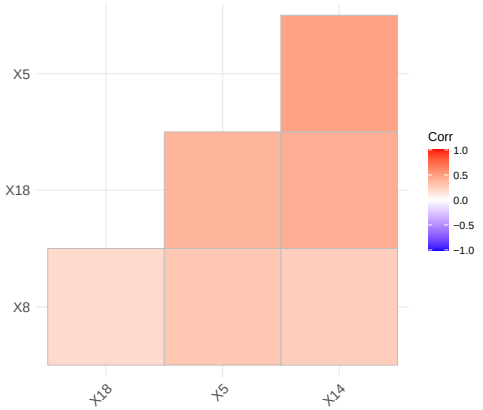

Figure 24: Correlación Spearman, datos sin imputar

Correlación	**Estadístico**	**p-valor**
X5∼X8	2853616	1.027e-11
X5∼X14	839483	4.371e-16
X5∼X18	1270000	1.156e-10
X8∼X14	1225924	2.825e-05
X8∼X18	1580841	2.902e-05
X14∼X18	754334	4.589e-08

Table 2: Test de correlación de Spearman

4.2 Imputadas por mediana

Se observa de la figura 25 que, en general se ha mantenido la tendencia de correlaciones positivas distintas a cero, además, se aprecia un leve aumento en la correlación X5 X8 de 0.06 y una disminución en la correlación de X14 X18 de 0.091. Sin embargo, no se aprecian visualmente diferencias significativas entre las correlaciones antes y después de imputar.

	X5	**X8**	**X14**	**X18**
X5	1	0.363	0.424	0.322
X8	0.363	1	0.229	0.242
X14	0.424	0.229	1	0.330
X18	0.322	0.242	0.330	1

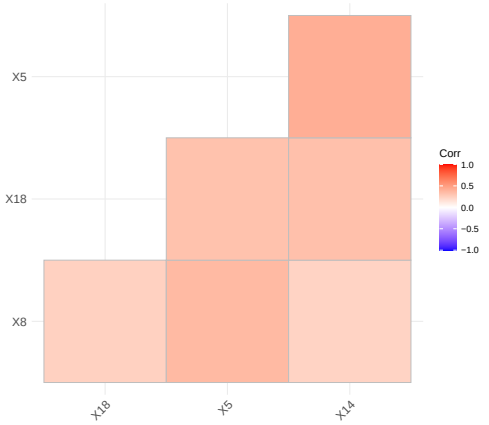

Figure 25: Correlación Spearman, imputación por mediana

4.3 Imputadas por mediana refinada

En general, se observa de la figura 26 una misma tendencia a mantener las correlaciones positivas, además, no se observa una mayor diferencia entre las correlaciones del método imputado por mediana y mediana refinada.

	X5	**X8**	**X14**	**X18**
X5	1	0.364	0.421	0.322
X8	0.364	1	0.204	0.244
X14	0.421	0.204	1	0.318
X18	0.322	0.244	0.318	1

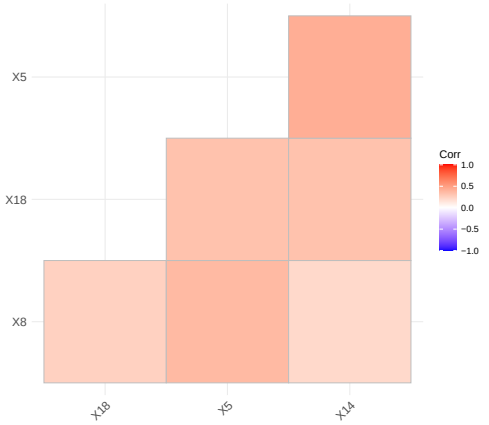

Figure 26: Correlación Spearman, imputación por mediana refinada

4.4 Imputadas por *hot deck*

Respecto a la correlaciones obtenidas por este método, se observa de la figura 27 un claro cambio en las correlaciones antes de imputar y después de aplicar *hot deck*, siendo para este caso la correlación más fuerte la X5 X8 y la más débil la X8 X14. Es interesante señalar que anteriormente se había concluido que el método *hot deck* al comprar las distribuciones, había obtenido un mejor rendimiento que las otras variables, sin embargo, desde esta mirada pareciera cambiar la conclusión a rechazar el uso de este método al momento de imputar.

	X5	**X8**	**X14**	**X18**
X5	1	0.354	0.312	0.246
X8	0.354	1	0.151	0.186
X14	0.312	0.151	1	0.210
X18	0.246	0.186	0.210	1

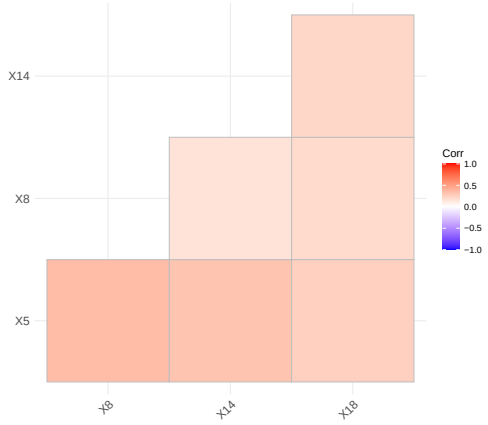

Figure 27: Correlación Spearman con imputación *hot deck*

4.5 Imputadas por regresión lineal

Por el método de regresión, se observa en la figura 28 una similitud en las correlaciones antes y después de imputar, con leves variaciones entre algunas correlaciones, visualmente no se observan diferencias significativas.

	X5	**X8**	**X14**	**X18**
X5	1	0.379	0.477	0.368
X8	0.379	1	0.290	0.289
X14	0.477	0.290	1	0.387
X18	0.368	0.289	0.387	1

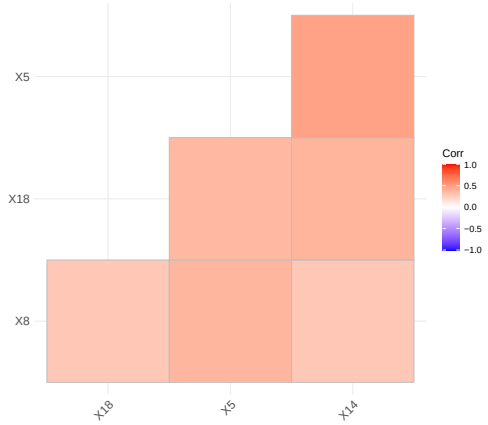

Figure 28: Correlación Spearman con imputación por regresión lineal

5 Pregunta 5

El método de Imputación Múltiple por Ecuaciones Encadenadas (*Multiple Imputation by Chained Equations*, MICE) es una técnica que sigue una serie de pasos, a partir de m conjunto de datos completos, para conseguir imputar los datos de la muestra, su implementación funciona bajo el supuesto de que los datos faltantes son MAR (Romero et al. 2018). Para su implementación con uso de la librería MICE perteneciente a R, se define m = 30 grupos completos, los resultados de imputación se aprecian gráficamente en las distribuciones de las figuras 29 al 32 antes y después de imputar, también se observa las correlaciones obtenidas por este método (ver figura 33).

A partir de la visualización de las distribuciones antes y después de imputar por MICE (figuras del 29 al 32), se puede observar una muy buena respuesta del método, a partir de un análisis visual se puede concluir que las variables X5. X8, X14 y X18 siguen distribuciones muy similares luego de imputar.

Por otro lado, al comparar las correlaciones antes (ver figura 24) y después de imputar (ver figura 33), no se observan cambios demasiado significativos en sus correlaciones, siendo la correlación con mayor variación la de X8 X14 con un aumento de 0.113 luego de imputar.

Luego de aplicar los métodos de imputación por mediana, mediana refinada, tipo *hot deck*, regresión lineal y MICE, para el presente caso de investigación, se concluye que el método MICE aporta mejores resultados en comparación al resto, al evaluar tanto sus distribuciones y correlaciones.

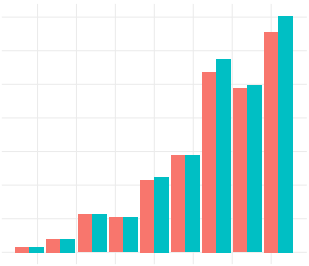

Figure 29: X5 c/s imputación por MICE

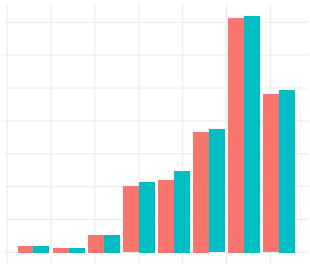

Figure 30: X8 c/s imputación por MICE

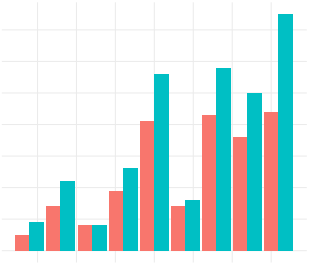

Figure 31: X14 c/s imputación por MICE

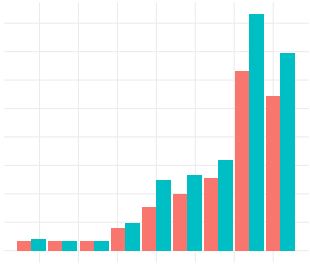

Figure 32: X18 c/s imputación por MICE

	X5	X8	X14	X18
X5	1	0.384	0.553	0.399
X8	0.384	1	0.369	0.259
X14	0.553	0.369	1	0.418
X18	0.399	0.259	0.418	1

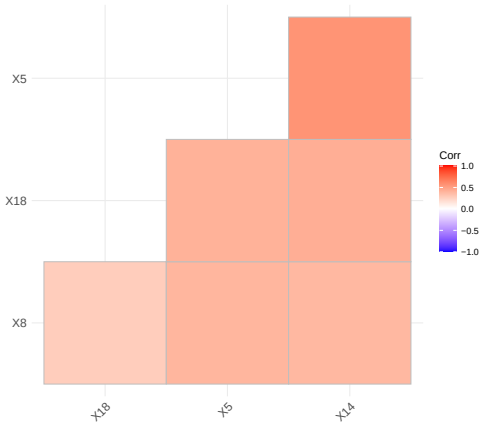

Figure 33: Correlación Spearman, imputación por MICE

References

[1] Peter B. Mandeville (2010) Observaciones perdidas.

[2] F. Medina y M. Galván (2007) Imputación de datos: teoría y práctica. CEPAL.

[3] R. Alfaro, M. Fuenzalida (2009) Imputación Múltiple en Encuestas Microeconómicas.

[4] E. Romero et al. (2018) Evaluación y comparación de métodos de imputación múltiple implementados en el paquete mice de R.