Mentoría: Coronavirus en Argentina y el mundo

Matías A. Bettera Marcat - Lucas C. Didoné - Daniel Peralta

Práctico 2: Análisis y Curación

Los datos analizados y donde hemos probado los códigos de notebook son hasta el domingo 26 de Julio.

Valores Faltantes

Análisis de algunos faltantes del Dataset 1 - *'Argentina-covid19.csv'*:

En las variables *'%mujer'*, *'%varon'*, *'mujer_total'*, *'varon_total'*, *'tests_realizados_total'*, *'tests_realizados_nuevos'* y *'test_por_millon_hab'* hay faltantes de los primeros 20 días donde no se reportaba el género de los infectados, y patir de ahí no vuelven a haber faltantes hasta el día 20 de Junio (fila 106) donde no se publicó el informe diario matutino (el vespertino si se publicó).

En la variable '*comunitario_nuevos'* faltan los primeros 28 días. Puede ser debido a no hubieran casos comunitarios al principio o a que no se reportaba por el ministerio. Se podrían reemplazar por 0.

Las variables *'alta_total'* y *'alta_nuevos'* presentan el faltante del día 20 de Junio, y varios faltantes durante los primeros 20 días. Se reporta el primer alta el día 12 de Marzo (fila 7) y hasta el 22 de Marzo no hay datos. Es posible no se reportara durante esos días. No sería adecuado completar los faltantes sin recurrir a fuentes alternativas. Además la variable *'alta_nuevos'* tiene faltantes a desde el día 20 de Junio hasta el final, seguramente, se calculaba tomando la diferencia de altas totales de cada día con el anterior. Este faltante se puede completar para los días posteriores al 21 de Junio.

La variable *'alta_definitiva'* tiene más faltantes que datos. Se reporta el dato 0 desde el comienzo hasta el 20 de Marzo, donde comienzan los faltantes. Luego, solo entre los días 25 de Marzo y 4 de Abril, esta variable tiene datos. Probablemente fue el único periodo donde se informó cuántas de las altas eran definitivas. Esta variable no se puede completar sin información de otras fuentes.

Dataset 2 - "Argentina-covid19-por-provincia.csv"

Las únicas variables que presentan datos faltantes en este dataset son 'muertes_total', 'muertes_nuevos' y 'observaciones'.

Para la variable 'muertes_total' se observan faltantes en distintas provincias para los días 24, 25, 26 y 27 en el mes de marzo, y para el día 27 de junio. A partir de los reportes oficiales, se pudo comprobar que los valores reales son cero para esos registros, por lo cual estos datos podrían corregirse haciendo un reemplazo de NaN por "0".

En cuanto a la variable 'muertes_nuevos', se corroboró que ocurre lo mismo que lo observado para la variable 'muertes_total', pudiendo corregirse a través de un reemplazo de NaN por "0".

Por último, la variable 'observaciones' contiene muy pocos datos (menos del 5% de los registros tienen valor). En este caso, es una variable que podría descartarse.

Observación general: este dataset no se sincroniza desde el día 26/06.

Análisis valores faltantes Dataset 3

Variable	Cantidad de valores faltantes	Porcentaje de valores faltantes
fecha	0	0 %
provincia	12	0.505 %
num_caso	0	0 %
genero	16	0.674 %
edad	104	4.381 %
tipo_caso	2342	98.652 %
comorbilidades	2330	98.147 %
viajes	2352	99.073 %
observaciones	2218	93.429 %

Tras ver los valores de datos faltantes de todas las columnas del dataset, podemos dividirlos en 3 grupos: **Sin datos faltantes**, **Con pocos faltantes**, **Con muchos faltantes**.

- **Sin datos faltantes:** dentro de este grupo podemos incluir a las columnas que no presentan datos faltantes. Las columnas que pertenecen a este grupo son: **fecha** y **num_caso**.
- **Con pocos faltantes:** dentro de este grupo incluimos a las columnas que que cuentan con aproximadamente un 5% de datos faltantes. Las columnas que pertenecen a este grupo son: **provincia**, **genero** y **edad**.
- **Con muchos faltantes:** dentro de este grupo incluimos a las columnas que que cuentan con más del 90% de datos faltantes. Las columnas que pertenecen a este grupo son: **tipo_caso**, **comorbilidades**, **viajes** y **observaciones**.

Tratamiento valores faltantes Dataset 3

Con pocos faltantes

Como se puede observar, si filtramos el dataset para visualizar las filas que tienen datos faltantes en todas las columnas del grupo Con pocos faltantes, podemos observar que todas las filas resultantes tienen completa la columna observaciones. En esta columna se indica que estas filas pertenecen a muertes faltantes en los reportes o bien corresponden a datos de reportes no publicados. Es por este motivo que la mayoría de las columnas se encuentran incompletas. Una alternativa de limpieza de estos datos, sería eliminar las filas que no cuentan con datos en estas columnas.

Con muchos faltantes

Teniendo en cuenta la gran cantidad de datos faltantes que contienen estas columnas y que los datos de las mismas no corresponden a valores numéricos calculados. Una alternativa para limpiar el conjunto de datos, sería eliminar estas columnas.

Inconsistencias

Inconsistencias de 'muertes_nuevos' y 'muertes_totales' - dataset1

Inconsistencia del día 30 de Marzo (fila 25): Habían reportados 20 fallecimientos en total en el reporte vespertino del día 29, el reporte matutino del día 30 suma un fallecimiento, y el reporte vespertino del mismo día suma cuatro fallecimientos nuevos, pero reporta un total de 24 muertes totales. En este caso la inconsistencia está en la fuente de los datos. En los informes del día siguiente, no se rectifica esta inconsistencia, y se considera que el número de fallecidos es efectivamente 24. Esta inconsistencia puede deberse a que el fallecimiento del reporte matutino del día 30 con datos de edad y/o provincia incorrectos y por eso no coincide con ningún fallecimiento del vespertino, aunque esté contabilizado. Se podría restar un caso al día 29, para corregir la inconsistencia..

Los días 2 y 3 de Abril (filas 28 y 29) sucede que hay muertes que se están contando de manera errónea al cargarlas a la base de datos. El criterio parece ser que incluye a las muertes del reporte matutino en el dato de muertes nuevas del día anterior. Por lo que las muertes acumuluadas pueden no coincidir con el dato del reporte vespertino del día. El reporte vespertino del 2 de Abril incluye las dos muertes del reporte matutino del mismo día, se debería corregir el número de muertes nuevas de 4 a 3 (2 del vespertino del 2/04 y 1 del matutino 3/04). Esto hace corrige la inconsistencia.

El día 3 de Abril sucede de forma distinta al día anterior. En el informe matutino se reporta una muerte, y luego el vespertino suma 5 muertes sin contar la anterior. En el siguiente reporte matutino no hay muertes nuevas, por lo que las muertes de 3 de Abril deben ser 5 en lugar de las 6 que figuran en la base de datos, debido a que la muerte del reporte matutino del mismo día ya había sido contabilizada en las 37 totales del día anterior.

El en día 30 de Abril (fila 56) las muertes acumuladas coinciden con las reportadas en el informe vespertino de ese día, y las muertes nuevas en la base son 3, aunque en el reporte matutino se informa una muerte y en el vespertino 3, por lo que el total es 4 en lugar de 3. Con el valor de 4 las es consistente con las muertes totales del día anterior y del día 30.

El caso del día 22 de Mayo (fila 78) presenta una inconsistencia. Las muertes totales del mismo día y del anterior coinciden con los reportes, pero las muertes nuevas no coinciden, se informan 3 en el reporte matutino y 14 en el vespertino, que suman 17, mientras en la base de datos figuran 18. Con el valor de 17 se corrige la inconsistencia.

Inconsistencias en la columna 'casos_nuevos' entre dataset 1 y dataset 2

Para una fecha dada el valor de la columna **casos_nuevos** del dataset 1 debería ser igual a la sumatoria de valores que toma esta columna para la misma fecha en el dataset 2. Esto se debe a que mientras en el dataset 1 hay una fila por fecha, para el dataset 2 puede que haya más de una fila para la misma fecha (tantas filas como provincias que presenten casos para esa fecha). Teniendo esto en cuenta se procede analizar si se cumple o no esta igualdad. Luego de realizar el análisis se encontraron las siguientes inconsistencias:

INCONSISTENCIAS		
FECHA	**casos_nuevos dataset 1**	**casos_nuevos dataset 2**
15/03/2020	11	12
27/03/2020	101	117
28/03/2020	55	56

29/03/2020	75	73
15/04/2020	128	127
05/05/2020	134	133
09/05/2020	165	164
29/05/2020	717	718
06/06/2020	983	573
30/06/2020	2262	2263
18/07/2020	3223	3212

Una posible alternativa de solución a esta inconsistencia sería reemplazar el valor de la columna 'casos_nuevos' del dataset 1 por el calculado a partir del dataset 2 para aquellos casos en donde se presente la inconsistencia, el problema de realizar esto es que otras columnas podrían quedar inconsistentes (como la columna casos_total).

Inconsistencias en la columna 'casos_total' entre dataset 1 y dataset 2

Se procede a realizar un análisis similar al propuesto en el punto anterior, pero para la columna **casos_total**. Luego de realizar el análisis se encontraron las siguientes inconsistencias:

INCONSISTENCIAS		
FECHA	**casos_total dataset 1**	**casos_total dataset 2**
15/03/2020	56	57
18/03/2020	97	98
19/03/2020	128	129
20/03/2020	158	159
21/03/2020	225	226
22/03/2020	266	267
23/03/2020	301	303
24/03/2020	387	389
25/03/2020	502	506
26/03/2020	589	593
27/03/2020	690	710
28/03/2020	745	766
15/04/2020	2571	2570
15/05/2020	7479	7478

24/05/2020	12076	12013
31/05/2020	16851	16823
11/06/2020	27373	27037
12/06/2020	28764	28763

Una alternativa de solución para esta inconsistencia sería la misma que la propuesta para la inconsistencia anterior pero para la columna casos_total (pero se mantendría el mismo problema)

Inconsistencias entre la columna 'casos_total' y ‘casos_nuevos’ en el dataset 2

Se procede a analizar si los valores correspondientes a las columnas **casos_total** y **casos_nuevos** es consistente para todas las provincias, para ellos se realiza un análisis por provincia, donde para cada registro de cada provincia se valida si el valor de la columna **casos_total** es igual a la suma del valor de la columna **casos_nuevos** más el valor de la columna **casos_total** del registro anterior para esa provincia. Luego de realizar el análisis se encuentran las siguientes inconsistencias:

INCONSISTENCIAS	
PROVINCIA	**CANTIDAD DE INCONSISTENCIAS**
CABA	19
Buenos Aires	15
Chaco	0
Río Negro	3
San Luis	0
Córdoba	5
Santa Fe	3
Chubut	1
Tierra del Fuego	5
Entre Ríos	4
Jujuy	7
Salta	4
Santa Cruz	2
Tucumán	5
Corrientes	6
Neuquén	2
Santiago del Estero	1
Mendoza	6

Misiones	12
La Pampa	2
La Rioja	4
San Juan	0
Formosa	2
Catamarca	0

La cantidad total de inconsistencias es de 108.

Una posible alternativa de solución para los casos en que se presenta esta inconsistencia es reemplazar el valor presente en la columna casos_total por el calculado a partir de sumar el valor de la columna casos_nuevos más el valor de la columna casos_total del registro anterior para una provincia dada. El problema de realizar esto, es que se podrían generar otras inconsistencias (por ejemplo la inconsistencia planteada en el punto anterior)

Inconsistencias de columna 'edad' - dataset1

A partir del análisis de outliers, se detectaron inconsistencias en la columna 'edad' del dataset 1. Para corregir los valores de esta columna, resultará necesario recurrir a una fuente externa ya que el resto de los datasets de la solución no cuentan con esa columna.

Outliers

Dataset "Argentina-covid19-fallecidos.csv" - Análisis de edad de personas fallecidas

Teniendo en cuenta la variable 'edad' de las personas fallecidas, se encuentran 19 casos que pueden considerarse outliers, que van desde los 0 a los 26 años. En cuanto al caso de edad "0", se verificó en base al reporte oficial diario emitido por el gobierno nacional que es una inconsistencia en el dataset, ya que no hubo decesos de personas de esa edad registrados hasta el momento. El valor mínimo que sí se pudo verificar con datos oficiales es de 1 año, reportado el día 24/06/2020.

Tratamiento de los outliers

En este caso, dejaríamos estos outliers ya que representan información importante al momento de realizar análisis relacionados a la edad de las personas fallecidas a causa del virus.