Limpieza del dato

Tipos y ciclo de vida del dato

CONTENIDO

Objetivos

- Conocer la estructura de los datos.
- Determinar la presencia de datos ausentes.
- Conocer técnicas específicas utilizadas para la transformación del dato.

Generalidades

La limpieza de los datos es sin duda una parte crucial del procesamiento de los datos para un correcto análisis de los mismos, en los diferentes módulos que integran esta asignatura quedará completamente demostrado que los datos deberán poseer una integridad estructura que darán lugar a la posibilidad y fiabilidad en la realización de las diferentes evaluaciones estadísticas a las que estos puedan ser sometidos para estudiar su comportamiento y para la posible evolución de comportamiento futuros, así como sus posibles relacionamientos.

Lo primero que debe realizarse cuando nos encontramos ante un conjunto de datos, es una exploración exhaustiva de los mismo, a través de métodos numéricos y visuales que nos permitan tener una aproximación a su forma, estructura e integridad.

Entenderemos por forma la característica de cada uno de los datos, aquí puede mencionarse si es numérico, entero, decimal, un porcentaje o cualquier otro tipo de presentación, así mismo podrá ser categórico, dicotómico o nominal.

En la estructura mostrara si las categorías son correctas y consistentes y, finalmente, se evaluará la integridad para ver la carencia de datos y posibles errores de captura a través de *outlier*.

Para la comprensión de este tema se hará hincapié en ejemplos prácticos que realizarán con R, ya que es la manera más clara de comprender que se tiene en un conjunto de datos, sobre todo en conjuntos de datos medianos a grandes donde sería imposible visualizar una tabla de datos y detectar errores de forma o estructura a través de la observación de los mismos.

Para ayudar en el proceso de iniciación en R, se darán códigos genéricos para la exploración de datos que serán fácilmente adaptables a conjuntos alternativos de datos diferentes a los usados en los ejemplos de este módulo, así como tips, para su rápida adaptación, si ya estas familiarizado con R, sabrás que este lenguaje ofrece una serie de recursos amplísima para optimizar los datos.

Los datos son hoy día considerados unos de los más valiosos activos. Por lo tanto, es indispensable que te asegures de que tengan la calidad más alta posible y presenten el servicio más relevante a tu negocio *(tabla 1)*.

Para mantener todo en orden en tus bases de datos, puedes aplicar una estrategia de *data cleansing* o limpieza de datos. Si no sabes a qué nos referimos, a continuación, te hablamos acerca de su función y cómo realizarla.

¿Qué es la limpieza de datos o *data cleansing*?

Es, como su nombre lo indica, la depuración de datos erróneos en una tabla o base de datos. Esta acción permite identificar datos incorrectos, incompletos o poco relevantes para un estudio en concreto. Después de la limpieza, se sustituyen, modifican o eliminan por completo los datos inservibles. Para realizar esta acción siempre es necesario conocer el contexto de los datos, ya que el analista debe analizar de forma estéril los datos, pero esto no quiere decir que deba estar desinformado sobre los mismos, por el contrario, los analistas deben estar lo más documentados posible con relación a los datos que trabajan.

La depuración de datos asegura que los datos con los que se cuenta sean confiables. Esto te da la seguridad de que cualquier información que obtengas de ellos será mucho más precisa y útil.

Importancia de la limpieza de datos

La limpieza de datos es vital para garantizar una alta integridad de los datos en la empresa. Si toda la información con la que se cuenta es confiable, entonces puede estar seguro de que las conclusiones que presentes también lo serán y, por tanto, las decisiones que se tomen con base en ella.

Característica	Descripción
Estandarizados	Deben estar definidos para ser de fácil interpretación
Completos	Se deben poseer todos los datos posibles
Precisos	Deben ser exactos y fiables
Validables	Deben corresponderse con la magnitud y/o característica que pretenden presentar
Únicos	Las observaciones deben corresponderse con único suceso o individuo

Tabla 1. Características principales de los datos. Tomado de: Elaboración propia, 2021.

Los datos de calidad pueden variar dependiendo de cuál sea su cualidad, entre las principales se encuentran:

- **Exactitud:** todos los datos dentro de la empresa deben ser precisos. Una forma de comprobar su exactitud es comparándolos con otras fuentes. Si esta fuente no existe o es inexacta, entonces la información también lo será.
- **Coherencia:** la coherencia de los datos te permite saber si la información de contacto que tienes de una persona u organización es la misma en diferentes bases de datos, tablas o aplicaciones que utilices.
- **Validez:** todos los datos deben cumplir con reglas o restricciones definidas. De igual forma, cada información debe poder ser validada para comprobar si son correctos o no.
- **Uniformidad:** es importante que todos los datos dentro de tus bases tengan los mismos valores o unidades. Este es un elemento realmente indispensable a la hora de hacer *data cleansing*, pues de no tener todo en orden, el proceso se vuelve complejo.

Otro aspecto relevante es la metodología empleada para la captura y registro de nuevos datos, ya que mientras más fina sea la configuración utilizada para la captura de datos estos serán siempre de mayor cantidad y albergaran menor cantidad de errores. Por ejemplo, en el llenado del campo de numero de móvil, condicionare a que el dato egresado sea numérico, con una longitud de 9 dígitos y en el caso de España que el mismo inicie los dígitos 6 o 7, o el ingreso de DNI que debe poseer una estructura obligatoria de 8 dígitos numéricos y terminado en una letra, que por defecto será mayúscula.

Unos datos de poca calidad entregaran análisis de baja calidad y predicciones muy poco fiables.

Sin embargo, es importante entender que el análisis predictivo no es magia y, aunque el algoritmo aprende, solo puede extraer sentido de los datos que le proporcionamos. Los algoritmos no tienen capacidad de intuición como los humanos, sea esto bueno o malo, por lo que el éxito del sistema depende principalmente de los datos de entrada.

De los datos a los algoritmos

Es habitual enfrentarse a un proyecto de análisis datos con muchos datos. La primera tarea es recopilarlos todos. Normalmente están en diferentes repositorios:

- En el CRM de la empresa.
- En bases de datos SQL (o noSQL).
- En hojas de cálculo.
- En las redes sociales.
- En el programa de facturación empresarial.
- En el programa de gestión de las listas de correo electrónico.
- En los informes de transacciones bancarias.

En algunos de estos casos, los datos fueron recopilados en estos repositorios sin considerar su uso posterior en análisis específicos y se registraron con fines meramente documentales, sin embargo, hoy en día todos los registros de datos son susceptibles de análisis, aunque para ello debemos en algunos casos hacer mayor manipulación de los mismos para optimizar su usabilidad para realizar análisis a partir de estos.

Los algoritmos son importantes, pero no son lo más importante. La fase previa de recolección y preparación de datos requiere un esfuerzo y conocimientos mínimos para poder llevar un proyecto con éxito. Esta fase puede tomar entre un 80 % y un 90 % del tiempo del proyecto.

Factores como la experiencia, la intuición, el conocimiento del negocio y de los clientes son básicos. Estas competencias puedes tenerlas en tu empresa, o quizás necesites contratarlas fuera

Para llevar a cabo la limpieza del dato deben considerarse una serie de pasos:

- Exploración de los datos.
- Determinación de datos ausentes.
- Determinar si existen datos atípicos o errores.
- Transformar los datos.

¿QUÉ PAQUETES SE NECESITARÁN?

Librerías o paquetes de R para la exploración y limpieza de datos.

En todo proyecto de ciencia de datos, la limpieza y el análisis exploratorio de estos resultan de vital importancia para las siguientes etapas del mismo. De hecho, solo el preprocesado de los datos representa entre un 70 y un 80 % de todo el trabajo. Se habla de tareas como la corrección de codificaciones y tipos de datos erróneos, trasformaciones de los datos o la detección y manejo adecuado de datos anómalos y faltantes.

Una vez se han llevado a cabo los procedimientos de limpieza de datos y estos son consistentes, resulta necesario realizar un análisis de los mismos por medio de distintos estadísticos y gráficos que permitan describir las variables con las que se trabaja y determinar las posibles relaciones entre ellas. La calidad del conjunto de datos finalmente obtenido determinará la robustez y fiabilidad de los resultados del proyecto.

En R se disponen de diversas librerías que facilitan todas estas tareas. Se describirán algunas de las más útiles. Las mencionadas aquí no son todas las disponibles, ya que como vimos en temas anteriores R dispone de más de 18 000 paquetes esta es una de las baterías de paquetes de manipulación y transformación de data más usada.

Tidyverse

En la *(figura 1)*, se presentan los paquetes R agrupados en paquete Tidyverse.

Ggplot2

ggplot2 es un sistema para crear gráficos de forma declarativa, basado en *the grammar of graphics*. Tú proporcionas los datos, le dices a ggplot2 cómo asignar variables a la estética, qué primitivas gráficas usar y él se encarga de los detalles.

Dplyr

Proporciona una gramática de manipulación de datos, proporcionando un conjunto consistente de verbos que resuelven los desafíos de manipulación de datos más comunes. Ir a documentos, etc.

Tidyr

Tidyr proporciona un conjunto de funciones que le ayudarán a ordenar los datos. Los datos ordenados son datos con una forma coherente: en resumen, cada variable va en una columna y cada columna es una variable. Ir a documentos, etc.

Readr

Readr proporciona una forma rápida y sencilla de leer datos rectangulares (como csv, tsv y fwf). Está diseñado para analizar de manera flexible muchos tipos de datos que se encuentran en la naturaleza, mientras sigue fallando cuando los datos cambian inesperadamente. Ir a documentos, etc.

Purrr

Purrr mejora el kit de herramientas de programación funcional (FP) de R al proporcionar un conjunto completo y consistente de herramientas para trabajar con funciones y vectores. Una vez que domine los conceptos básicos, purrr le permite reemplazar muchos bucles *for* con un código que es más fácil de escribir y más expresivo. Ir a documentos, etc

Tibble

Tibble es una reimaginación moderna del marco de datos, manteniendo el tiempo que ha demostrado ser efectivo y descartando lo que no. Tibbles son *data. frames* que son perezosos y hoscos: hacen menos y se quejan más, lo que lo obliga a enfrentar los problemas antes, lo que generalmente conduce a un código más limpio y expresivo. Ir a documentos, etc.

Stringr

Stringr proporciona un conjunto cohesivo de funciones diseñadas para hacer que trabajar con cadenas sea lo más fácil posible. Está construido sobre Stringr, que usa la biblioteca ICU C para proporcionar implementaciones rápidas y correctas de manipulaciones comunes de cadenas. Ir a documentos, etc.

Forcats

Forcats proporciona un conjunto de herramientas útiles que resuelven problemas comunes con factores. R usa factores para manejar variables categóricas, variables que tienen un conjunto fijo y conocido de valores posibles.

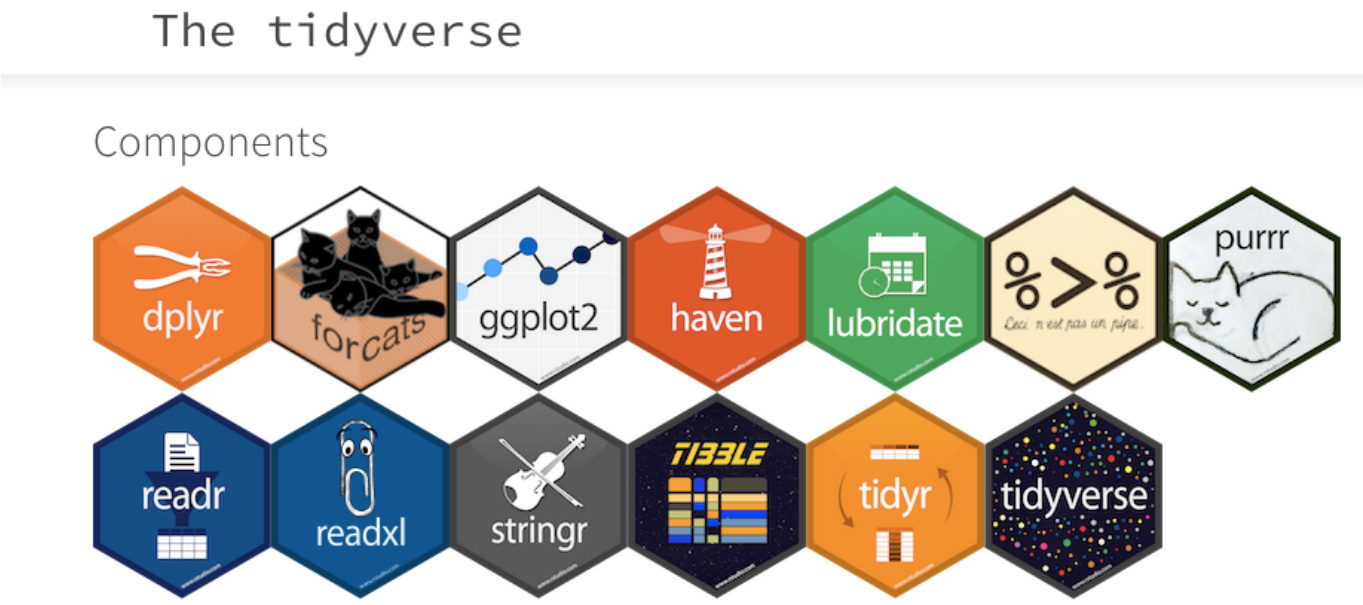


Figura 1. Paquetes de R agrupados en paquete tidyverse. Tomado de: Beatrizmilz, 2019, https://beatrizmilz.github.io/2019-02-R-Interm-R-LadiesSP/img/Tidyverse_packages.png.

Stringr

Contiene funciones con las que manipular cadenas de caracteres que sirven, por ejemplo, para resolver problemas en la codificación de variables categóricas. Entre sus muchas funciones, algunas de ellas nos permiten: Eliminar espacios en blanco en cualquiera de los lados de las cadenas.

Lubridate

El paquete Lubridate para R, creado por Garrett Grolemund y Hadley Wickham. Según los autores del paquete, "el lubridate tiene una sintaxis consistente y memorable, que hace que trabajar con las fechas sea divertido en lugar de frustrante" [1]. Sin este paquete, manipular las fechas en R es muy complejo.

Existen diferentes representaciones de fecha y hora, con este paquete podrá determinarse cuál es el formato en el que viene presentado nuestros datos de fecha y si se considera pertinente hacer una transformación, este mismo paquete ofrece la posibilidad de completar esta tarea.

para ver nuestra configuración regional usamos la siguiente función:

Sys.getlocale("LC_TIME")

"*Spanish*_Mexico.1252"

extraer información a partir de una fecha

y <- *year*(fecha)

m <- *month*(fecha)

d <- *day*(fecha)

extraer información a partir de una hora

hr <- *hour*(hora)

minu <- *minute*(hora)

sec <- *second*(hora). [7]

Detección de datos faltantes

La detección de datos faltantes en un conjunto de datos es un proceso ineludible, ya que la aplicación de métodos y modelos estadísticos a los datos en muchos casos exige que estos se encuentren íntegros, debe conocerse si los métodos que pretendemos aplicar para analizar los datos presentan esta exigencia. Por otro lado, si en un conjunto de datos registramos una falencia muy significativa de datos, es decir, ausencia de estos estaríamos frente a la posibilidad de generar conclusiones asociadas al estudio poco fiables *(figura 2)*.

Outliers con *boxplot*

El gráfico de caja (*boxplot*) constituye una primera opción al momento de analizar e identificar datos atípicos, el mismo presenta la mediana, el primer y tercer cuartil, además del 1.5*iqr o rango intercuartílico. En el caso de R, se puede verificar que la opción *boxplot. stats(x)*, *out* permite identificar los valores considerados como atípico y los valores utilizados para representar el *boxplot (figura 3).*

Rango intercuartílico

En el ejemplo que se trae entre manos, se ve que *Q*1 está en 353 pasos y que *Q*3 está en 1947 pasos. Entre *Q*1 y *Q*3 se sabe que están el 50 % de los valores obtenidos en el estudio. A esta distancia se le llama rango intercuartílico, *interquantile range* (IQR):

$IQR = Q3 - Q1 = 1947.5 - 353.0 = 1594.5$ *pasos*

Se define como valor atípico leve aquel que dista 1,5 veces el rango intercuartílico por debajo de Q1 o por encima de Q3:

$q < Q1 - 1{,}5 \cdot IQR$ *o bien* $q > Q3 + 1{,}5 \cdot IQR$

Y valor atípico extremo aquel que dista 3 veces el rango intercuartílico por debajo de Q1 o por encima de Q3:

$q < Q1 - 3 \cdot IQR$ *o bien* $q > Q2 + 3 \cdot IQR$

Desarrollando estas fórmulas en R se será capaz de determinar tanto el umbral superior como el inferior a partir de los cuales consideraremos que los valores son atípicos [2].

Métodos de imputación

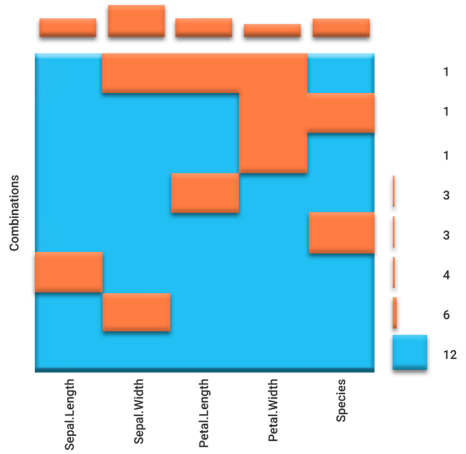
Una vez conocidos los principales tipos de datos perdidos, se revisarán algunas vías o métodos de imputación de datos para *datasets* transversales, sin profundizar en series de tiempo (dado que este tipo de datos requiere otro tratamiento).

No hacer nada

- Este método es el más fácil y rápido. Al no hacer nada, el algoritmo que se entrene manejará los datos perdidos.
- Algunos algoritmos pueden tener en cuenta los valores perdidos y aprender los mejores valores de imputación (o incluso tratarlos como una categoría aislada) para los datos faltantes en base a la función de pérdida calculada durante el entrenamiento (tales como los modelos basados en árboles). Otros algoritmos tienen la opción de simplemente ignorar estos casos (tales como los modelos de regresión). Dependiendo de la librería utilizada, algunos algoritmos simplemente no se ejecutarán y nos solicitarán revisar los valores perdidos.

Dataframe con datos faltantes				
sapply(df, function(x)(sum(is.na(x)))) # NA counts				
Sepal.Length	*Sepal.Width*	*Petal.Length*	*Petal.Width*	*Species*
4	7	4	3	4

aggr(iris, prop=FALSE, numbers=TRUE, border=NA, combine=TRUE)



Dataframe con datos faltantes				
sapply(df, function(x)(sum(is.na(x)))) # NA counts				
Sepal.Length	*Sepal.Width*	*Petal.Length*	*Petal.Width*	*Species*
0	0	0	0	0

aggr(df, prop=FALSE, numbers=TRUE, border=NA, combine=TRUE)

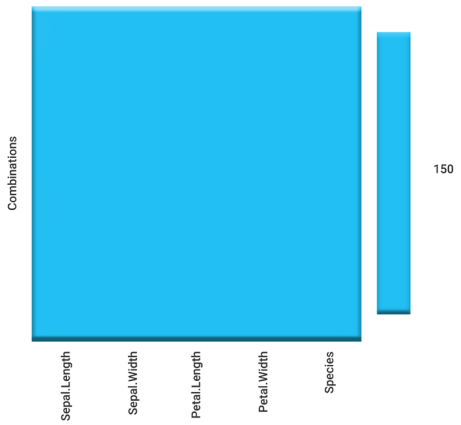


Figura 2. Dataframe con datos faltates. Tomado de: Elaboración propia, 2021.

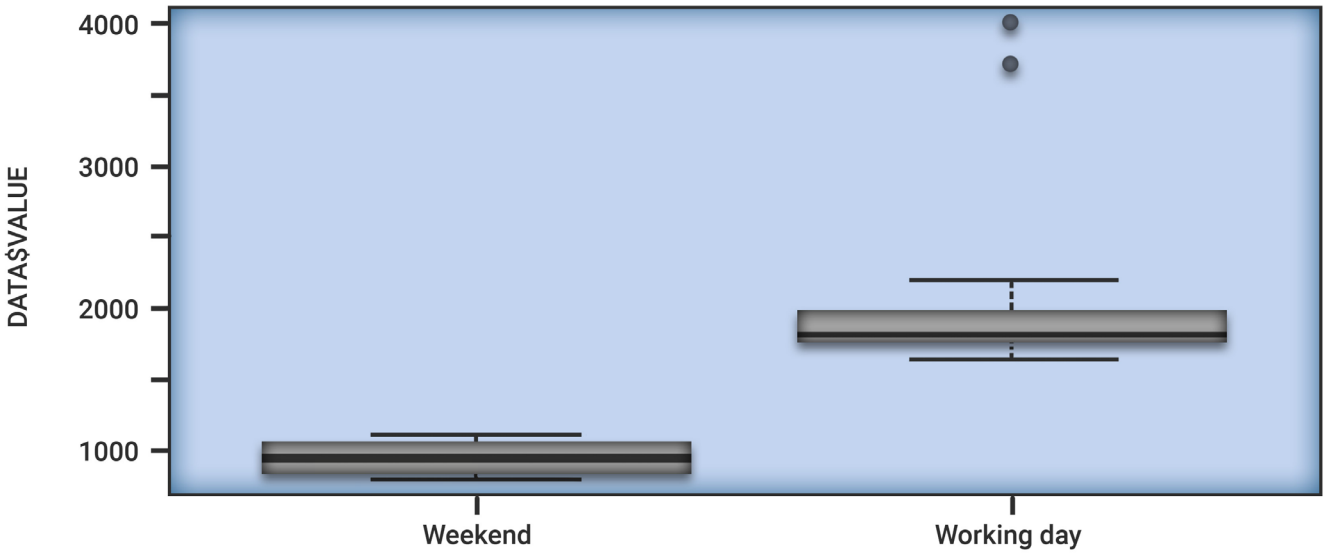


Figura 3. Visualización de datos atípicos con una gráfica de cajas. Tomado de: elaboración propia, 2021.

Media y mediana

- Este método solo puede ser utilizado con datos numéricos.
- Para imputar con la media o la mediana deben seguirse los pasos descritos a continuación:
  - Identificar valores perdidos.
  - Calcular la media / mediana de los valores no perdidos en una columna.
  - Reemplazar los valores perdidos con la media/ mediana en la misma columna.
- Esta imputación dependerá de la simetría de la distribución de los datos existentes. Si la distribución de los datos es simétrica, la media y la mediana son iguales, entonces se puede usar cualquiera de los dos valores. Si la distribución es sesgada hacia la derecha, la media será mayor a la mediana, y es preferible utilizar esta última para no alterar la distribución de los datos. Finalmente, si la distribución es sesgada hacia la izquierda, la media será menor a la mediana, y por tanto es la medida más conveniente.

Sus ventajas son las siguientes:

- Fácil y rápido.
- Funciona bien con *datasets* numéricos pequeños.

Sus desventajas son las siguientes:

- No tiene en cuenta las correlaciones entre las variables. Solo funciona a nivel de columna.
- No se recomienda usarlo con variables categóricas, ya que ofrece malos resultados trabajando con variables categóricas codificadas.
- No es muy exacto y le falta precisión con datos dispersos.
- No tiene en cuenta la incertidumbre en las imputaciones.

Moda

Un método estadístico similar al anterior, útil para imputar datos perdidos es usar los valores más frecuentes (o moda). Esta técnica puede trabajar con variables categóricas y consiste en lo siguiente:

- Identificar valores perdidos.
- Calcular el valor más frecuente por columna.
- Reemplazar los valores perdidos con el valor más frecuente dentro de cada columna.

Sus ventajas son las siguientes:

- Funciona bien con variables categóricas.

Sus desventajas son las siguientes:

- No factoriza las correlaciones entre variables.
- Puede introducir sesgo en los datos.

Ceros o constantes

- Este es un método muy sencillo. Tal como sugiere su nombre, reemplaza los valores faltantes con cero o cualquier valor constante que especifiquemos.
- Este método se utiliza cuando el valor perdido en realidad corresponde a un cero, por ejemplo.

Transformación del conjunto de datos

De una las acciones que se suele necesitar utilizar con bastante frecuencia, son tareas de transformación del conjunto de datos, que pueden incluir filtrado ordenación renombrado de columnas, asignación de valores a datos ausentes, eliminar columnas, incluir columnas a partir de los datos existentes, establecer el tipo de dato que presenta cada columna y si el descrito en la data original no se corresponde con el tipo de dato transformarlo, para darle la forma correcta, dar formato a fechas a números a categorías donde pueden describir categorías o simplificarlas con la utilización de números coma, cuando los volúmenes de datos son importantes puede interesar cambiar las categorías con etiquetas clásicas alfanuméricas a un único dígito que las represente, pero esta es una decisión que queda en manos del analista.

Algunas de las acciones habituales en la limpieza de datos son las siguientes:

- Igualar formatos.
- Descartar campos.
- Corregir errores ortográficos.
- Dar formato a fechas.
- Eliminar columnas duplicadas.
- Borrar registros no útiles.

Con los datos "limpios" ya se puede empezar a hacer una selección de los que serán útiles para hacer las predicciones.

La transformación de los datos también permite generar nuevos campos basados en los que ya se tienen. El conocimiento del dominio del entorno de los datos es fundamental para abordar esta fase. Esta fase de conocimiento de los datos es quizá la que más tiempo ocupará, ya que tendremos que familiarizarse con diversos temas que estarán asociados a diversos problemas o casos de estudio.

Bibliografía

[1] Santos D. Data cleansing: qué es la limpieza de datos y cómo realizarla, 2021 [En línea]. Disponible en: https://blog.hubspot.es/marketing/limpieza-de-datos. [Accedido: 05-nov-2021]

[2] *Tutorial: limpieza y análisis de datos*, H Fallas, 2015 [En línea]. Disponible en: https://ladatacuenta.com/wp-content/uploads/2015/07/tutorial-limpieza-y-anc3a1lisis-de-datos-bc3a1sico-hassel-fallas1.pdf. [Accedido: 05-nov-2021]

[3] "Praxis IR, de Big Data G. La importancia de la limpieza de datos en Big Data", Praxis, 2020 [En línea]. Disponible en: https://mexico.praxisglobe.com/recursos/diseminaciones/BIG_DATA/BDA-WP-01-2020-ENTR.pdf. [Accedido: 05-nov-2021]

[4] G. Salvador, S. Ramírez, J. Luengo y F. Herrera, "Big data: preprocesamiento y calidad de datos", *Novática*, 237, pp.17-23, 2016 [En línea]. Disponible en: https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf. [Accedido: 05-nov-2021]

[5] "8 techniques for efficient data cleaning", Codemotion, 2020 [En línea]. Disponible en: https://www.codemotion.com/magazine/dev-hub/big-data-analyst/data-cleaning/. [Accedido: 05-nov-2021]

[6] "Perfil V. Minería de texto: limpiar, estandarizar y "tokenizar" datos de texto en R", Beta Economía, 2019 [En línea]. Disponible en: http://betaeconomia.blogspot.com/2019/03/mineria-de-texto-limpiar-estandarizar-y.html. [Accedido: 05-nov-2021]

[7] "Clase 6 - Limpieza de datos", RPubs, 2021 [En línea]. Disponible en: https://rpubs.com/camilamila/limpieza. [Accedido: 05-nov-2021]

[8] "Guía Rápida de Lubridate", RStudio, 2017 [En línea]. Disponible en: https://rstudio-pubsstatic.3.amazonaws.282491_53d8767462f84a4e9f3b928890cc1854.html. [Accedido: 05-nov-2021]

[9] "Perfil V. 5 Métodos para la identificación de valores atípicos en R", Blogspot, 2018 [En línea]. Disponible en: http://betaeconomia.blogspot.com/2018/10/metodos-para-la-identificacion-de.html. [Accedido: 05-nov-2021]

[10] "Transformación de datos", Análisis de datos con R, 2019 [En línea]. Disponible en: https://intro-r-analisis.netlify.app/transformacion-de-datos.html. [Accedido: 05-nov-2021]