Limpieza y adecuación de datos en R

Introducción a la Ciencia de Datos
2c-2023

Características deseables de los datos

- Relevantes para el problema
- Sin datos duplicados
- Sin datos faltantes o incorrectos
- Que tengan una estructura *tidy*

Pasos generales

1. Familiarizarse con el *data set* y el dominio del problema
2. Chequear por errores estructurales (p. ej. tipos de datos de las columnas)
3. Chequear que sean *tidy* o que tengan una estructura que nos sirva
4. Chequear por irregularidades en los datos (ver si los *outliers* en realidad son datos mal ingresados, ver si hay datos faltantes)
5. Si hay datos faltantes o mal ingresados, decidir qué hacer
6. Documentar las versiones y los cambios realizados

Familiarizarse con el *data set* y el dominio del problema

- Tipos de datos de cada columna (`int, dbl, chr, lgl, char, fct`)
- Del dominio del problema ya hablamos: conocer el *negocio,* googlear
- Funciones útiles: `glimpse(...)`

```
> df_seguros <- read.csv('~/Downloads/Insurance.csv')
> glimpse(df_seguros)
Observations: 1,338
Variables: 7
$ age      <int> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56, 27, 19, 52,
$ sex      <fct> female, male, male, male, male, female, female, female, male, femal
$ bmi      <fct> "27,90", "33,77", "33,00", "22,71", "28,88", "25,74", "33,44", "27,
$ children <int> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0
$ smoker   <fct> yes, no, no, no, no, no, no, no, no, no, no, yes, no, no, yes, no,
$ region   <fct> southwest, southeast, southeast, northwest, northwest, southeast, s
$ charges  <fct> "16884,92", "1725,55", "4449,46", "21984,47", "3866,86", "3756,62",
```

¿Se acuerdan que pasaba cuando usaban read.csv() en vez de read_csv ()?

Errores estructurales

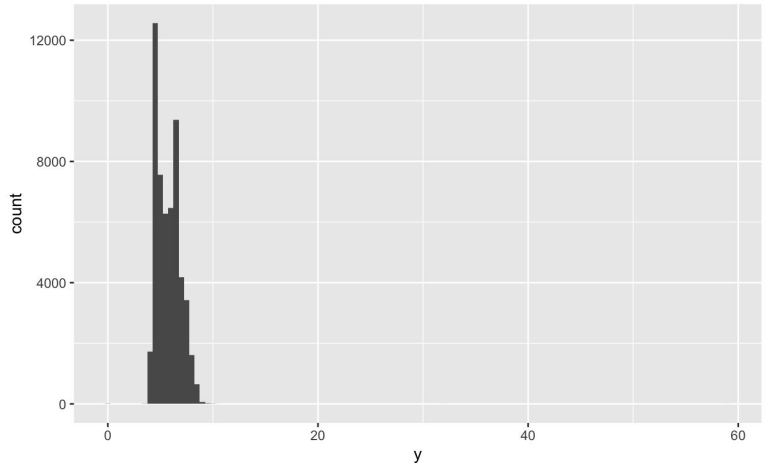
- Verificar si hubo errores en la importación de los datos
- Los tipos de datos de cada columna ¿Son consistentes con lo que deberían tener?
- Funciones útiles: `problems(...)`, `as.factor(...)`, `as.integer(...)`, `as.numeric(...)`, `as.character(...)` , `as.logical(...)`

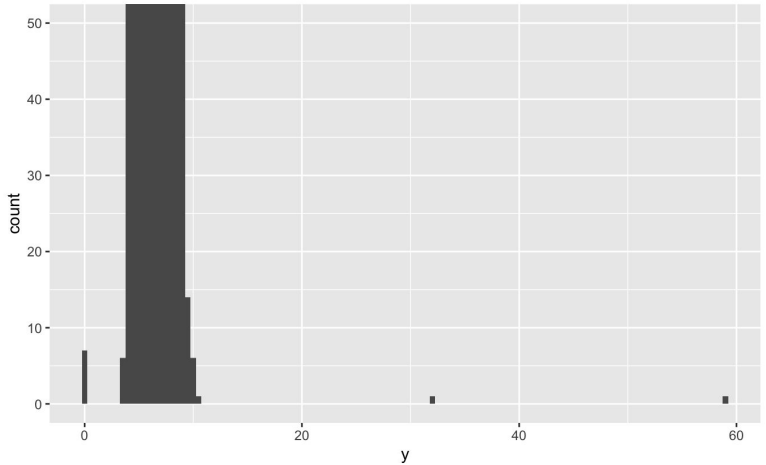
Chequear por irregularidades en los datos

- Los valores de tipo string pueden tener errores de ortografía o capitalización, p. ej. *“yes”*, *“Yes”* o *“YES”* son tres strings diferentes, pero para nuestro análisis podrían significar lo mismo.
- Dependiendo del dominio del problema, identificar valores faltantes o incorrectos en los atributos, p. ej. Un cero en vez de NA, los valores NA o NaN
- Identificar los valores que **claramente** están mal ingresados (dominio del problema). **No confundir con valores extremos de una distribución.**

Chequear por irregularidades en los datos

- P.ej. *Outliers* en diamonds: Histogramas de la variable y (ancho)





Wickham: EDA->Valores faltantes

Chequear por irregularidades en los datos

- Datos faltantes: Si R los identifica como faltantes les pone **NA** (no confundir con **NaN)**
- Dependiendo de la cantidad de datos que tengamos y lo que estamos tratando de analizar quizás nos convenga hacer algo, no se suele recomendar eliminar la fila entera mejor reemplazarlos con un valor, ¿cuál? depende...(dominio del problema).
- Funciones útiles: `is.nan(...)`, `is.na(...)`, `unique(...)`

Repaso datos *tidy*

- Cada variable debe tener su propia columna
- Cada observación debe tener su propia fila
- Cada valor debe tener su propia celda

Datos *anchos* y *largos*

- Datos anchos es cuando en un dataset los nombres de las columnas no representan nombres de variables, sino que representan los *valores* de una variable

pais	anio	casos
Afganistán	1999	745
Afganistán	2000	2666
Brasil	1999	37737
Brasil	2000	80488
China	1999	212258
China	2000	213766

pais	1999	2000
Afganistán	745	2666
Brasil	37737	80488
China	212258	213766

Tabla 4

Datos *anchos* y *largos*

- Datos largos es cuando una observación aparece en múltiples filas

pais	anio	tipo	casos
Afganistán	1999	casos	745
Afganistán	1999	población	19987071
Afganistán	2000	casos	2666
Afganistán	2000	población	20595360
Brasil	1999	casos	37737
Brasil	1999	población	172006362
Brasil	2000	casos	80488
Brasil	2000	población	174504898
China	1999	casos	212258
China	1999	población	1272915272
China	2000	casos	213766
China	2000	población	1280428583

Tabla 2

pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	17504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Algunos errores comunes

- Caracteres especiales (p. ej. Comas o puntos en valores numéricos)
- Caracteres donde debería haber valores numéricos
- Filas duplicadas
- Imprecisiones
- Datos faltantes o con problemas: en blanco, NA, NaN, ceros en vez de valores nulos

Checklist de limpieza y adecuación

- Familiarizarse con el *data set* y el dominio del problema
- Verificar errores de importación
- Verificar tipos de datos de cada columna, que sean acorde a lo que representan.
- Verificar valores de los datos: *valores incorrectos, faltantes, etc.*
- Para los valores problemáticos decidir qué estrategia utilizar para reemplazarlos o borrarlos
- Verificar datos duplicados (ojo, entender si son duplicados o no)
- Todas las decisiones se tienen que **documentar** y justificar