**DATA CLEANING**

Los objetivos detrás del “data cleaning” son los siguientes:

- Reorganizar datos que pueden estar “desordenados” (por ejemplo URLs)

-Convertir columnas en distintos tipos de datos

-Trucos para manipular los NULLs.

**LEFT & RIGHT**

Estos statements obtienen los caracteres de la izquierda y derecha respectivamente y los devuelven separados (en un string distinto).

Ejemplo: suponiendo que tenemos phone\_number como un string de 12 caracteres y sabiendo que tiene este formato:

abc – defg – hij

Siendo abc el área y defg – hij el número de teléfono propiamente dicho.

LEFT (phone\_number, 3) AS area\_code

RIGHT(phone\_number, 8) AS phone\_number\_only

**LENGTH**

Devuelve la longitud de un string.

RIGHT(phone\_number, LENGTH(phone\_number – 4) AS phone\_number\_alt

**POSITION, STRPOS, LOWER, UPPER**

POSITION: te devuelve una columna con la posición del carácter buscado en un string.

POSITION(‘,’ IN city\_state)

¿Qué pasa si el carácter está más de 1 vez? ¿Te devuelve la primera o la última posición?

STRPOS: funciona igual que POSITION pero la nomenclatura es distinta.

STRPOS(city\_state, ‘,’)

¡Ojo! Ambos son case sensitive. Para eso nos pueden ayudar “LOWER” o “UPPER”, que nos devuelve una columna poniendo todo el string en mayúsculas o minúsculas.

**CONCAT o ||**

Ambos sirven para unir distintos strings, pudiendo añadir caracteres entre medias de ambos.

CONCAT(first\_name, ‘ ‘, last\_name) AS full\_name

first\_name || ‘ ‘ || last\_name AS full\_name2

https://www.postgresql.org/docs/8.1/functions-string.html

Esta página es útil para trabajar con strings

**CAST**

Hay veces que la fecha viene separado por año, mes y día. Para poder trabajar con el tipo de dato “date” hay que:

1º Concatenar con el formato YYYY-MM-DD

2º Hacer el CAST de esa concatenación para asignar nombre a la nueva columna.

CAST(date\_column AS DATE)

No obstante se pueden hacer otras transformaciones, se puede profundizar aquí: https://www.postgresqltutorial.com/postgresql-tutorial/postgresql-cast/

También se puede usar el CAST usando “::”, un ejemplo:

value::target\_type

**TO\_DATE(month, ‘month\_name’)**

Coge la columna “month\_name” y transforma el carácter en el número asociado de mes.

**COALESCE (columnA, columnB) AS new\_column\_name**

COALESCE rellena la columna B con el valor de la columna A y deja vacía la columna A

Devuelve el primer valor “no nulo” de cada fila.