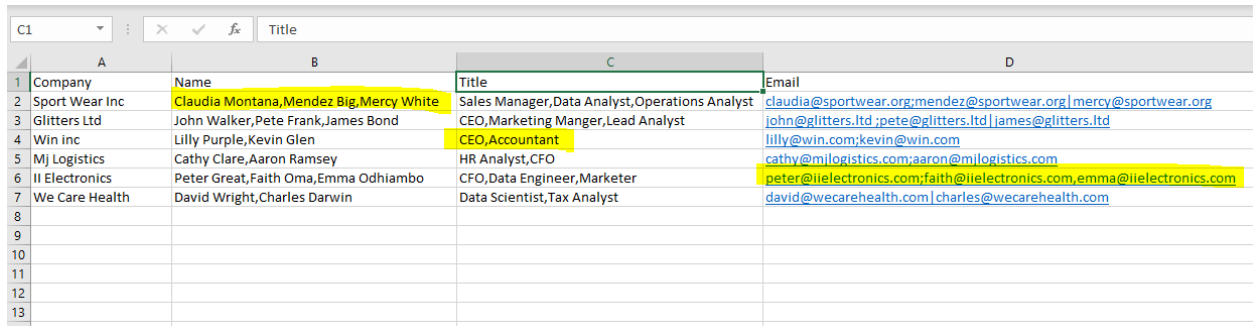


# Contact Cleaning

Maangi Josiah

It's common to receive contact data formatted as in the screenshot below whereby 2 or more names, titles, and emails are in one cell.. Each client may have multiple emails. To load this data into your database's contact table, you must separate each contact into its own record.



	A	B	C	D
1	Company	Name	Title	Email
2	Sport Wear Inc	Claudia Montana,Mendez Big,Mercy White	Sales Manager,Data Analyst,Operations Analyst	claudia@sportwear.org;mendez@sportwear.org mercy@sportwear.org
3	Glitters Ltd	John Walker,Pete Frank,James Bond	CEO,Marketing Manger,Lead Analyst	john@glitters.ltd;pete@glitters.ltd james@glitters.ltd
4	Win inc	Lilly Purple,Kevin Glen	CEO,Accountant	lilly@win.com;kevin@win.com
5	MJ Logistics	Cathy Clare,Aaron Ramsey	HR Analyst,CFO	cathy@mjlogistics.com;aaron@mjlogistics.com
6	II Electronics	Peter Great,Faith Oma,Emma Odhiambo	CFO,Data Engineer,Marketer	peter@iielectronics.com;faith@iielectronics.com,emma@iielectronics.com
7	We Care Health	David Wright,Charles Darwin	Data Scientist,Tax Analyst	david@wecarehealth.com charles@wecarehealth.com
8				
9				
10				
11				
12				
13				

Figure 1: contacts to clean screenshot

## Steps to follow

- Load the required libraries.
- Import the contacts dataset to clean.

A quick preview of the file shows that names are separated by a comma, ditto titles. Emails are separated by either a comma, pipe or semi colon.

- To get all the delimiters without missing any, use gsub with regex.
- Call the separate\_rows function from tidyr.
- We will then pass in our data, columns to separate, and delimiters into this function. Good news, you can pass multiple delimiters into sep and that is what we do.
- Lastly,separate Name into First\_Name and Last\_Name.

**NB: The separate\_rows function has been superseded by separate\_longer\_delim.**

## Loading Required Libraries

```
## [1] "C:/Users/MaangiJ"
```

## Import Contacts Dataset

```
contacts_to_clean <- read_excel("emails_to_clean.xlsx")
```

## Extract all Unique Delimiters

Create function to extract delimiters, exclude letters,numbers,space,., and @.

```
extract_delimiters <- function(text) {  
  return(gsub("[[:alnum:]]@. ]+", "", text))  
}
```

Apply the function to our data to extract all delimiters

```
# Extracting all delimiters  
email_delimiters <- c(  
  unlist(lapply(contacts_to_clean$Name, extract_delimiters)),  
  unlist(lapply(contacts_to_clean$Title, extract_delimiters)),  
  unlist(lapply(contacts_to_clean$Email, extract_delimiters))  
) %>%  
  str_trim() %>% str_c(collapse = "")
```

Get unique delimiters

```
# Extracting the unique delimiters  
unique_delimiters <- unlist(strsplit(email_delimiters, "")) %>% str_unique()  
unique_delimiters
```

```
## [1] " , " ; " | "
```

## Separate each Contacts into it's own record

Pass the unique delimiters to sep as in the code below.

```
cleaned_contacts <- contacts_to_clean %>%  
  separate_rows(c(Name, Title, Email), sep = '[,;+|]')  
  
cleaned_contacts %>% gt()
```

Company	Name	Title	Email
Sport Wear Inc	Claudia Montana	Sales Manager	claudia@sportwear.org
Sport Wear Inc	Mendez Big	Data Analyst	mendez@sportwear.org
Sport Wear Inc	Mercy White	Operations Analyst	mercy@sportwear.org
Glitters Ltd	John Walker	CEO	john@glitters.ltd
Glitters Ltd	Pete Frank	Marketing Manger	pete@glitters.ltd
Glitters Ltd	James Bond	Lead Analyst	james@glitters.ltd
Win inc	Lilly Purple	CEO	lilly@win.com
Win inc	Kevin Glen	Accountant	kevin@win.com
Mj Logistics	Cathy Clare	HR Analyst	cathy@mjlogistics.com
Mj Logistics	Aaron Ramsey	CFO	aaron@mjlogistics.com
II Electronics	Peter Great	CFO	peter@iielectronics.com
II Electronics	Faith Oma	Data Engineer	faith@iielectronics.com
II Electronics	Emma Odhiambo	Marketer	emma@iielectronics.com
We Care Health	David Wright	Data Scientist	david@wecarehealth.com
We Care Health	Charles Darwin	Tax Analyst	charles@wecarehealth.com

## Separate name into first and last name

```
cleaned_contacts %>%  
  separate(Name, c('First_Name', 'Last_Name'), sep = ' ') %>%  
  gt()
```

Company	First_Name	Last_Name	Title	Email
Sport Wear Inc	Claudia	Montana	Sales Manager	claudia@sportwear.org
Sport Wear Inc	Mendez	Big	Data Analyst	mendez@sportwear.org
Sport Wear Inc	Mercy	White	Operations Analyst	mercy@sportwear.org
Glitters Ltd	John	Walker	CEO	john@glitters.ltd
Glitters Ltd	Pete	Frank	Marketing Manger	pete@glitters.ltd
Glitters Ltd	James	Bond	Lead Analyst	james@glitters.ltd
Win inc	Lilly	Purple	CEO	lilly@win.com
Win inc	Kevin	Glen	Accountant	kevin@win.com
Mj Logistics	Cathy	Clare	HR Analyst	cathy@mjlogistics.com
Mj Logistics	Aaron	Ramsey	CFO	aaron@mjlogistics.com
II Electronics	Peter	Great	CFO	peter@iielectronics.com
II Electronics	Faith	Oma	Data Engineer	faith@iielectronics.com
II Electronics	Emma	Odhiambo	Marketer	emma@iielectronics.com
We Care Health	David	Wright	Data Scientist	david@wecarehealth.com
We Care Health	Charles	Darwin	Tax Analyst	charles@wecarehealth.com