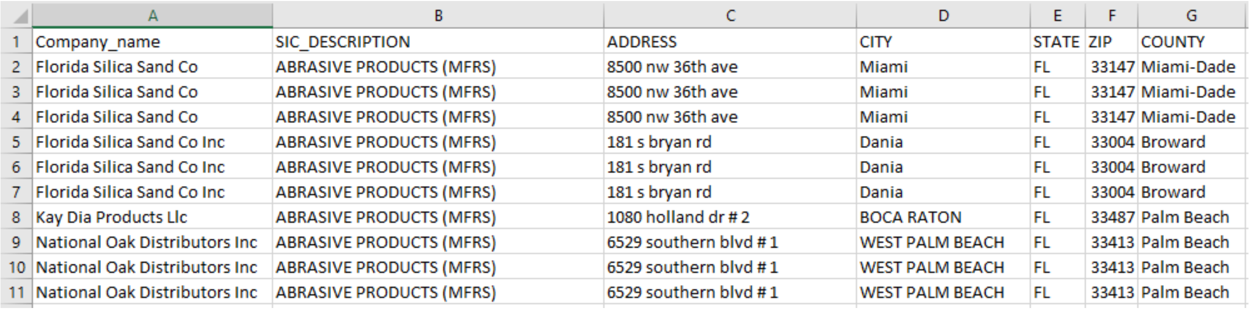
Excel Data Cleaning Demo

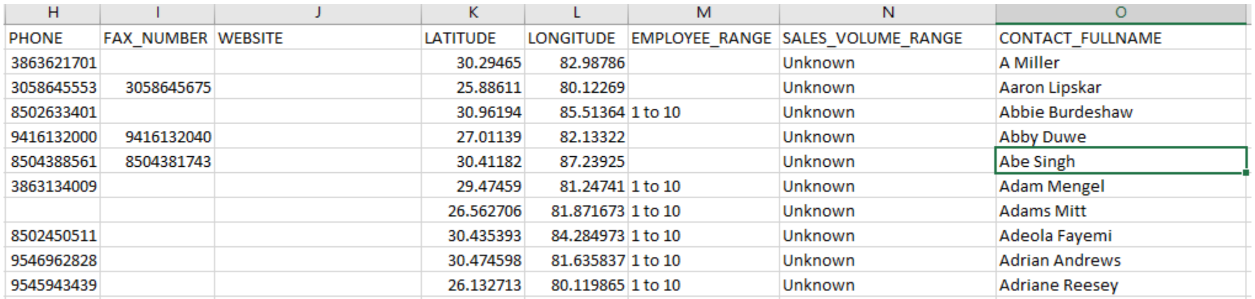
The following workflow shows an example of data cleaning using Excel Visual Basic for Applications (VBA) code.

The file to clean up is an excerpt of a larger business information dataset. This file has about 10,000 records.

Initial Dataset Look:



	A	B	C	D	E	F	G
1	Company_name	SIC_DESCRIPTION	ADDRESS	CITY	STATE	ZIP	COUNTY
2	Florida Silica Sand Co	ABRASIVE PRODUCTS (MFRS)	8500 nw 36th ave	Miami	FL	33147	Miami-Dade
3	Florida Silica Sand Co	ABRASIVE PRODUCTS (MFRS)	8500 nw 36th ave	Miami	FL	33147	Miami-Dade
4	Florida Silica Sand Co	ABRASIVE PRODUCTS (MFRS)	8500 nw 36th ave	Miami	FL	33147	Miami-Dade
5	Florida Silica Sand Co Inc	ABRASIVE PRODUCTS (MFRS)	181 s bryan rd	Dania	FL	33004	Broward
6	Florida Silica Sand Co Inc	ABRASIVE PRODUCTS (MFRS)	181 s bryan rd	Dania	FL	33004	Broward
7	Florida Silica Sand Co Inc	ABRASIVE PRODUCTS (MFRS)	181 s bryan rd	Dania	FL	33004	Broward
8	Kay Dia Products Llc	ABRASIVE PRODUCTS (MFRS)	1080 holland dr # 2	BOCA RATON	FL	33487	Palm Beach
9	National Oak Distributors Inc	ABRASIVE PRODUCTS (MFRS)	6529 southern blvd # 1	WEST PALM BEACH	FL	33413	Palm Beach
10	National Oak Distributors Inc	ABRASIVE PRODUCTS (MFRS)	6529 southern blvd # 1	WEST PALM BEACH	FL	33413	Palm Beach
11	National Oak Distributors Inc	ABRASIVE PRODUCTS (MFRS)	6529 southern blvd # 1	WEST PALM BEACH	FL	33413	Palm Beach

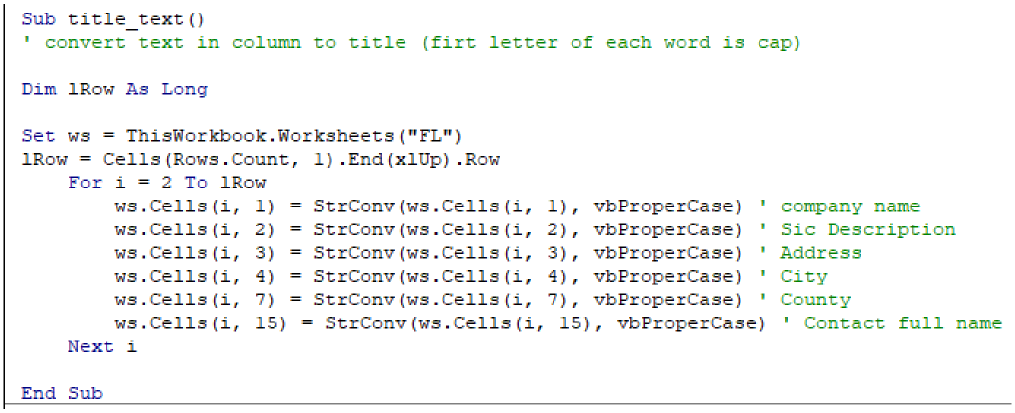


H	I	J	K	L	M	N	O
PHONE	FAX_NUMBER	WEBSITE	LATITUDE	LONGITUDE	EMPLOYEE_RANGE	SALES_VOLUME_RANGE	CONTACT_FULLNAME
3863621701			30.29465	82.98786		Unknown	A Miller
3058645553	3058645675		25.88611	80.12269		Unknown	Aaron Lipskar
8502633401			30.96194	85.51364	1 to 10	Unknown	Abbie Burdeshaw
9416132000	9416132040		27.01139	82.13322		Unknown	Abby Duwe
8504388561	8504381743		30.41182	87.23925		Unknown	Abe Singh
3863134009			29.47459	81.24741	1 to 10	Unknown	Adam Mengel
			26.562706	81.871673	1 to 10	Unknown	Adams Mitt
8502450511			30.435393	84.284973	1 to 10	Unknown	Adeola Fayemi
9546962828			30.474598	81.635837	1 to 10	Unknown	Adrian Andrews
9545943439			26.132713	80.119865	1 to 10	Unknown	Adriane Reesey

You can download a working version of this workbook here (must enable macros). The file will load with 'Dirty" data. Run the macro named "Main" to run all the subroutines. All code runs on the worksheet named "FL". You can rename the other worksheets to "FL" to see the code run again.

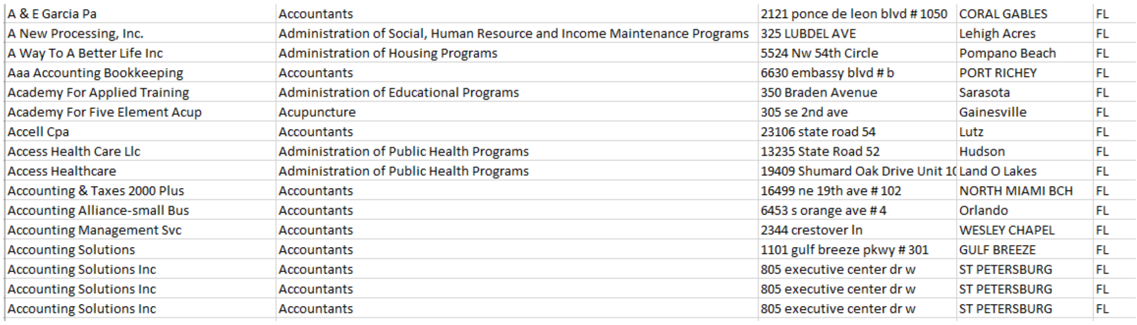
After an initial look at the data we can see that we can do quite a bit of cleanup:

1) Standardize text capitalization across and within columns.

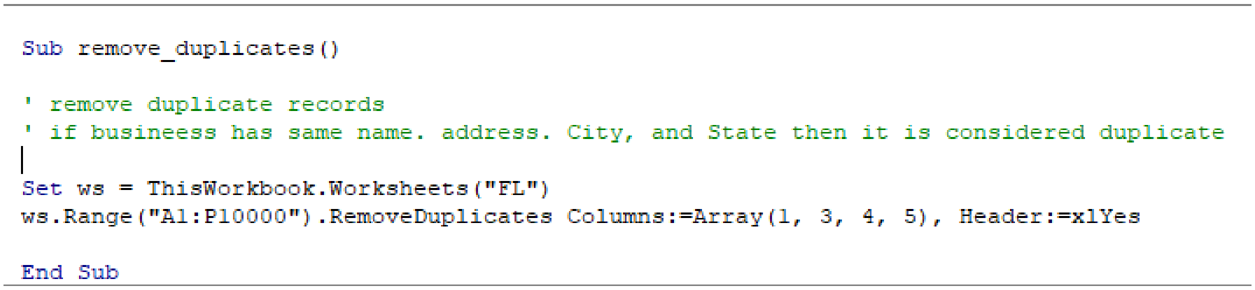


```
Sub title_text()
' convert text in column to title (firt letter of each word is cap)

Dim lRow As Long

Set ws = ThisWorkbook.Worksheets("FL")
lRow = Cells(Rows.Count, 1).End(xlUp).Row
    For i = 2 To lRow
        ws.Cells(i, 1) = StrConv(ws.Cells(i, 1), vbProperCase) ' company name
        ws.Cells(i, 2) = StrConv(ws.Cells(i, 2), vbProperCase) ' Sic Description
        ws.Cells(i, 3) = StrConv(ws.Cells(i, 3), vbProperCase) ' Address
        ws.Cells(i, 4) = StrConv(ws.Cells(i, 4), vbProperCase) ' City
        ws.Cells(i, 7) = StrConv(ws.Cells(i, 7), vbProperCase) ' County
        ws.Cells(i, 15) = StrConv(ws.Cells(i, 15), vbProperCase) ' Contact full name
    Next i

End Sub
```

2) Remove Duplicate entries for the same business.

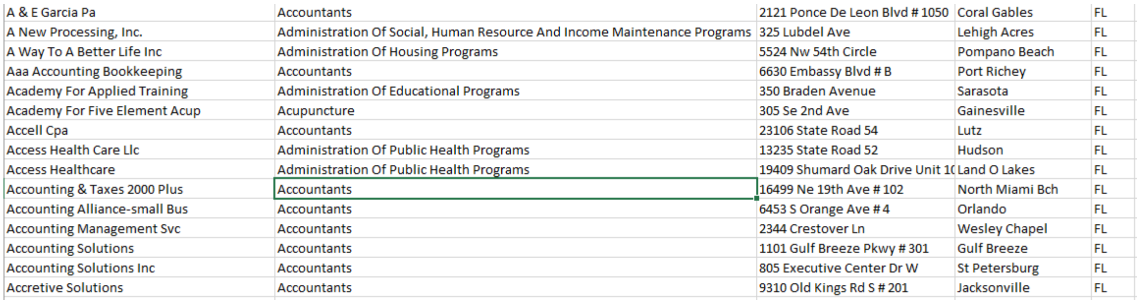


A & E Garcia Pa	Accountants	2121 ponce de leon blvd # 1050	CORAL GABLES	FL
A New Processing, Inc.	Administration of Social, Human Resource and Income Maintenance Programs	325 LUBDEL AVE	Lehigh Acres	FL
A Way To A Better Life Inc	Administration of Housing Programs	5524 Nw 54th Circle	Pompano Beach	FL
Aaa Accounting Bookkeeping	Accountants	6630 embassy blvd # b	PORT RICHEY	FL
Academy For Applied Training	Administration of Educational Programs	350 Braden Avenue	Sarasota	FL
Academy For Five Element Acup	Acupuncture	305 se 2nd ave	Gainesville	FL
Accell Cpa	Accountants	23106 state road 54	Lutz	FL
Access Health Care Llc	Administration of Public Health Programs	13235 State Road 52	Hudson	FL
Access Healthcare	Administration of Public Health Programs	19409 Shumard Oak Drive Unit 1(	Land O Lakes	FL
Accounting & Taxes 2000 Plus	Accountants	16499 ne 19th ave # 102	NORTH MIAMI BCH	FL
Accounting Alliance-small Bus	Accountants	6453 s orange ave # 4	Orlando	FL
Accounting Management Svc	Accountants	2344 crestover ln	WESLEY CHAPEL	FL
Accounting Solutions	Accountants	1101 gulf breeze pkwy # 301	GULF BREEZE	FL
Accounting Solutions Inc	Accountants	805 executive center dr w	ST PETERSBURG	FL
Accounting Solutions Inc	Accountants	805 executive center dr w	ST PETERSBURG	FL
Accounting Solutions Inc	Accountants	805 executive center dr w	ST PETERSBURG	FL



```
Sub remove_duplicates()

' remove duplicate records
' if busineess has same name. address. City, and State then it is considered duplicate

Set ws = ThisWorkbook.Worksheets("FL")
ws.Range("A1:P10000").RemoveDuplicates Columns:=Array(1, 3, 4, 5), Header:=xlYes

End Sub
```

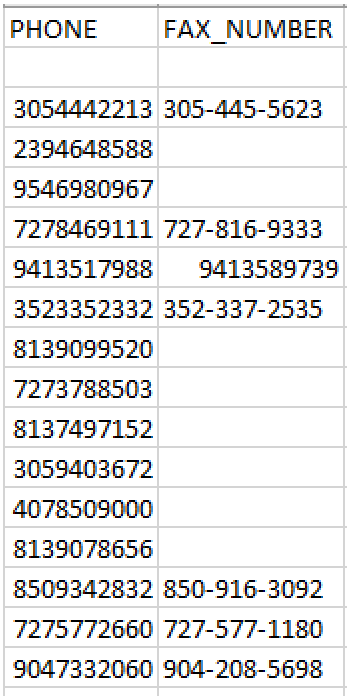
Dataset after removal of duplicates:



A & E Garcia Pa	Accountants	2121 Ponce De Leon Blvd # 1050	Coral Gables	FL
A New Processing, Inc.	Administration Of Social, Human Resource And Income Maintenance Programs	325 Lubdel Ave	Lehigh Acres	FL
A Way To A Better Life Inc	Administration Of Housing Programs	5524 Nw 54th Circle	Pompano Beach	FL
Aaa Accounting Bookkeeping	Accountants	6630 Embassy Blvd # B	Port Richey	FL
Academy For Applied Training	Administration Of Educational Programs	350 Braden Avenue	Sarasota	FL
Academy For Five Element Acup	Acupuncture	305 Se 2nd Ave	Gainesville	FL
Accell Cpa	Accountants	23106 State Road 54	Lutz	FL
Access Health Care Llc	Administration Of Public Health Programs	13235 State Road 52	Hudson	FL
Access Healthcare	Administration Of Public Health Programs	19409 Shumard Oak Drive Unit 1(	Land O Lakes	FL
Accounting & Taxes 2000 Plus	Accountants	16499 Ne 19th Ave # 102	North Miami Bch	FL
Accounting Alliance-small Bus	Accountants	6453 S Orange Ave # 4	Orlando	FL
Accounting Management Svc	Accountants	2344 Crestover Ln	Wesley Chapel	FL
Accounting Solutions	Accountants	1101 Gulf Breeze Pkwy # 301	Gulf Breeze	FL
Accounting Solutions Inc	Accountants	805 Executive Center Dr W	St Petersburg	FL
Accretive Solutions	Accountants	9310 Old Kings Rd S # 201	Jacksonville	FL

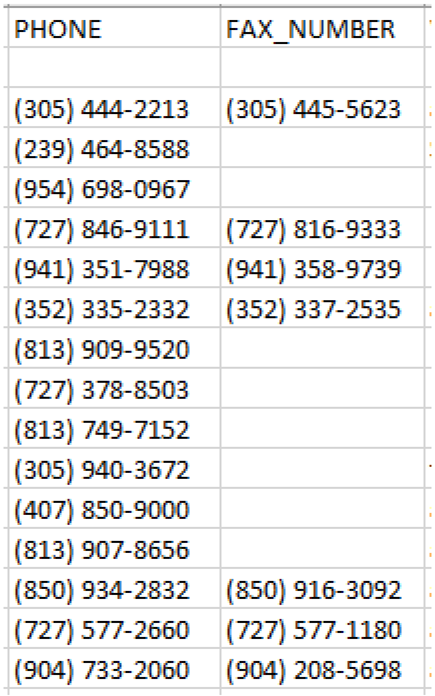
3) Format **Phone** and **Fax Number** Column.

Before Formatting

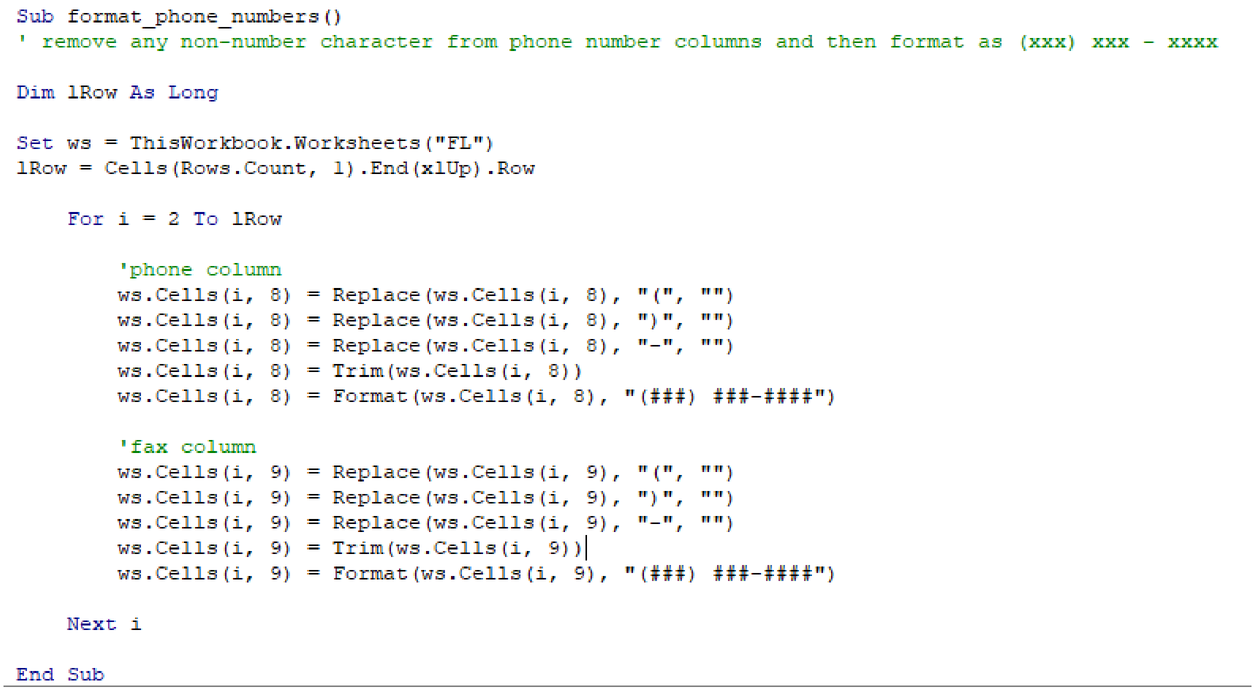


PHONE	FAX_NUMBER
3054442213	305-445-5623
2394648588	
9546980967	
7278469111	727-816-9333
9413517988	9413589739
3523352332	352-337-2535
8139099520	
7273788503	
8137497152	
3059403672	
4078509000	
8139078656	
8509342832	850-916-3092
7275772660	727-577-1180
9047332060	904-208-5698

After Formatting



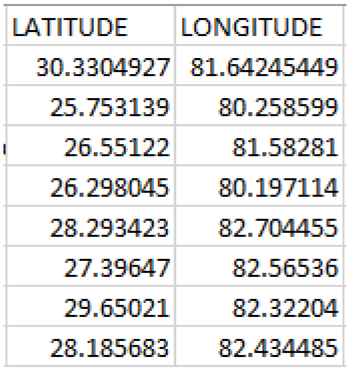
PHONE	FAX_NUMBER
(305) 444-2213	(305) 445-5623
(239) 464-8588	
(954) 698-0967	
(727) 846-9111	(727) 816-9333
(941) 351-7988	(941) 358-9739
(352) 335-2332	(352) 337-2535
(813) 909-9520	
(727) 378-8503	
(813) 749-7152	
(305) 940-3672	
(407) 850-9000	
(813) 907-8656	
(850) 934-2832	(850) 916-3092
(727) 577-2660	(727) 577-1180
(904) 733-2060	(904) 208-5698



```
Sub format_phone_numbers()
' remove any non-number character from phone number columns and then format as (xxx) xxx - xxxx

Dim lRow As Long

Set ws = ThisWorkbook.Worksheets("FL")
lRow = Cells(Rows.Count, 1).End(xlUp).Row

    For i = 2 To lRow

        'phone column
        ws.Cells(i, 8) = Replace(ws.Cells(i, 8), "(", "")
        ws.Cells(i, 8) = Replace(ws.Cells(i, 8), ")", "")
        ws.Cells(i, 8) = Replace(ws.Cells(i, 8), "-", "")
        ws.Cells(i, 8) = Trim(ws.Cells(i, 8))
        ws.Cells(i, 8) = Format(ws.Cells(i, 8), "(###) ###-####")

        'fax column
        ws.Cells(i, 9) = Replace(ws.Cells(i, 9), "(", "")
        ws.Cells(i, 9) = Replace(ws.Cells(i, 9), ")", "")
        ws.Cells(i, 9) = Replace(ws.Cells(i, 9), "-", "")
        ws.Cells(i, 9) = Trim(ws.Cells(i, 9))
        ws.Cells(i, 9) = Format(ws.Cells(i, 9), "(###) ###-####")

    Next i

End Sub
```

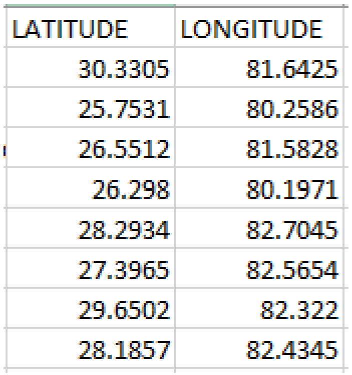
4) Round to a standard length the **Latitude** and **Longitude** columns.

Before Formatting

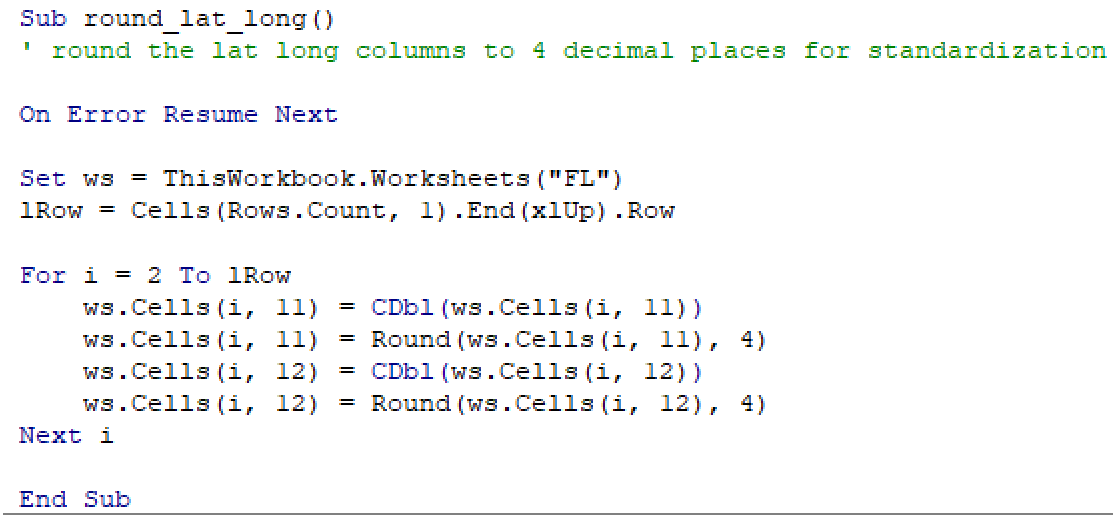


LATITUDE	LONGITUDE
30.3304927	81.64245449
25.753139	80.258599
26.55122	81.58281
26.298045	80.197114
28.293423	82.704455
27.39647	82.56536
29.65021	82.32204
28.185683	82.434485

After Formatting



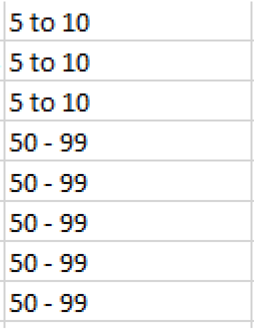
LATITUDE	LONGITUDE
30.3305	81.6425
25.7531	80.2586
26.5512	81.5828
26.298	80.1971
28.2934	82.7045
27.3965	82.5654
29.6502	82.322
28.1857	82.4345



```
Sub round_lat_long()
' round the lat long columns to 4 decimal places for standardization

On Error Resume Next

Set ws = ThisWorkbook.Worksheets("FL")
lRow = Cells(Rows.Count, 1).End(xlUp).Row

For i = 2 To lRow
    ws.Cells(i, 11) = CDbl(ws.Cells(i, 11))
    ws.Cells(i, 11) = Round(ws.Cells(i, 11), 4)
    ws.Cells(i, 12) = CDbl(ws.Cells(i, 12))
    ws.Cells(i, 12) = Round(ws.Cells(i, 12), 4)
Next i

End Sub
```

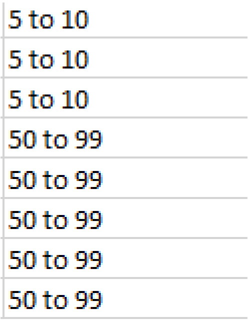
5) **Employee Range** column uses “ – “ and “ to “ as separator. Standardize to use “ to “.

Before Formatting

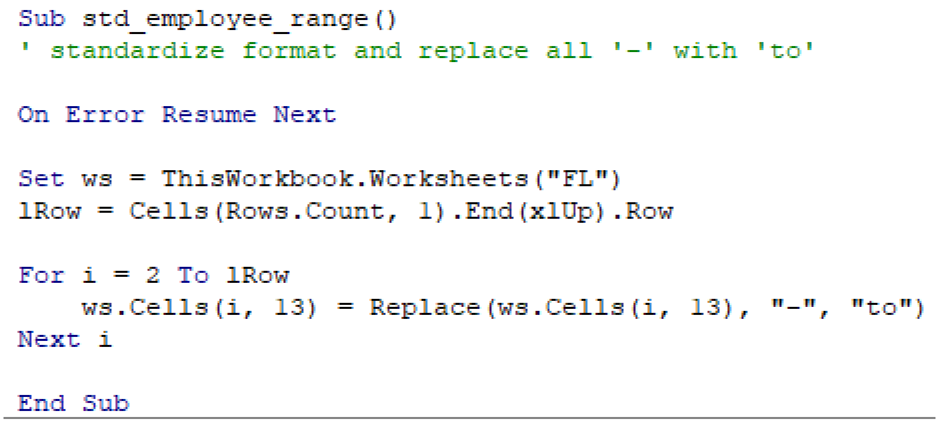


5 to 10
5 to 10
5 to 10
50 - 99
50 - 99
50 - 99
50 - 99
50 - 99

After Formatting



5 to 10
5 to 10
5 to 10
50 to 99
50 to 99
50 to 99
50 to 99
50 to 99

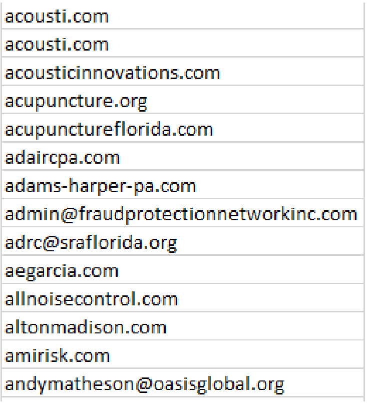


```
Sub std_employee_range()
' standardize format and replace all '-' with 'to'

On Error Resume Next

Set ws = ThisWorkbook.Worksheets("FL")
lRow = Cells(Rows.Count, 1).End(xlUp).Row

For i = 2 To lRow
    ws.Cells(i, 13) = Replace(ws.Cells(i, 13), "-", "to")
Next i

End Sub
```

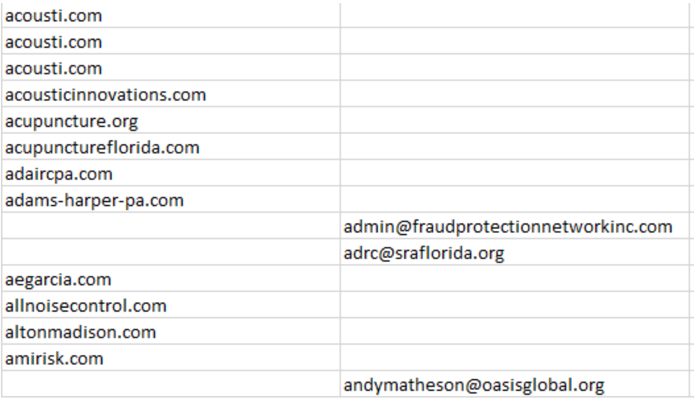
6) **Website** column contains both websites and emails. Create an **Email** column and migrate emails accordingly.

Before Email Extraction

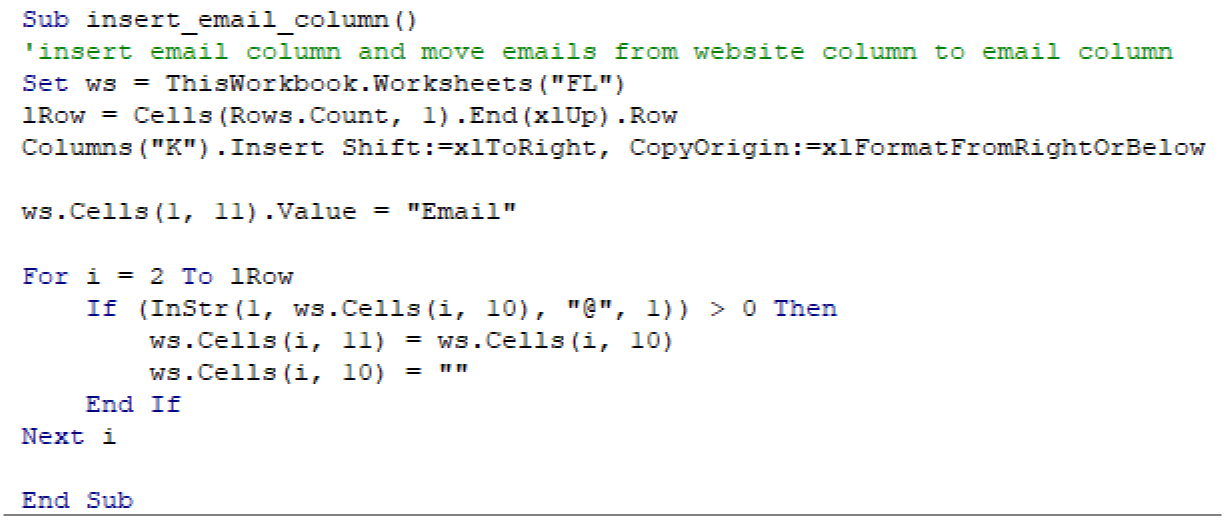
After Email Extraction



acousti.com
acousti.com
acousticinnovations.com
acupuncture.org
acupunctureflorida.com
adaircpa.com
adams-harper-pa.com
admin@fraudprotectionnetworkinc.com
adrc@sraflorida.org
aegarcia.com
allnoisecontrol.com
altonmadison.com
amirisk.com
andymatheson@oasisglobal.org



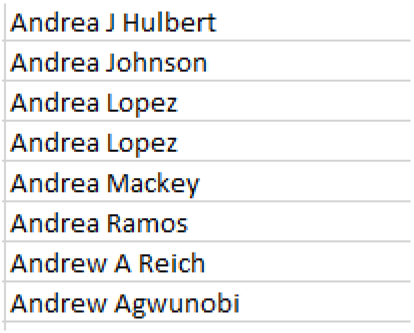
acousti.com	
acousti.com	
acousti.com	
acousticinnovations.com	
acupuncture.org	
acupunctureflorida.com	
adaircpa.com	
adams-harper-pa.com	
	admin@fraudprotectionnetworkinc.com
	adrc@sraflorida.org
aegarcia.com	
allnoisecontrol.com	
altonmadison.com	
amirisk.com	
	andymatheson@oasisglobal.org



```
Sub insert_email_column()
'insert email column and move emails from website column to email column
Set ws = ThisWorkbook.Worksheets("FL")
lRow = Cells(Rows.Count, 1).End(xlUp).Row
Columns("K").Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromRightOrBelow

ws.Cells(1, 11).Value = "Email"

For i = 2 To lRow
    If (InStr(1, ws.Cells(i, 10), "@", 1)) > 0 Then
        ws.Cells(i, 11) = ws.Cells(i, 10)
        ws.Cells(i, 10) = ""
    End If
Next i

End Sub
```

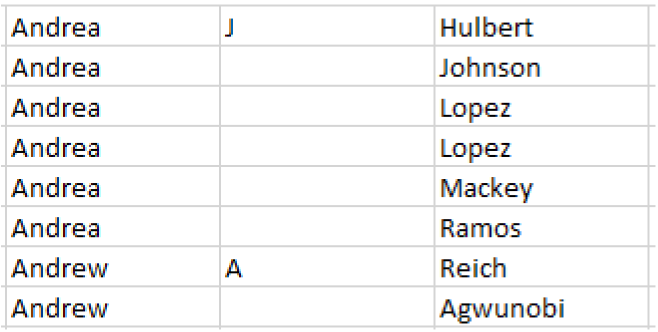
7) **Contact Names** are in a single column. Separate names in to **First, Middle,** and **Last Name** columns.

Before Name Split

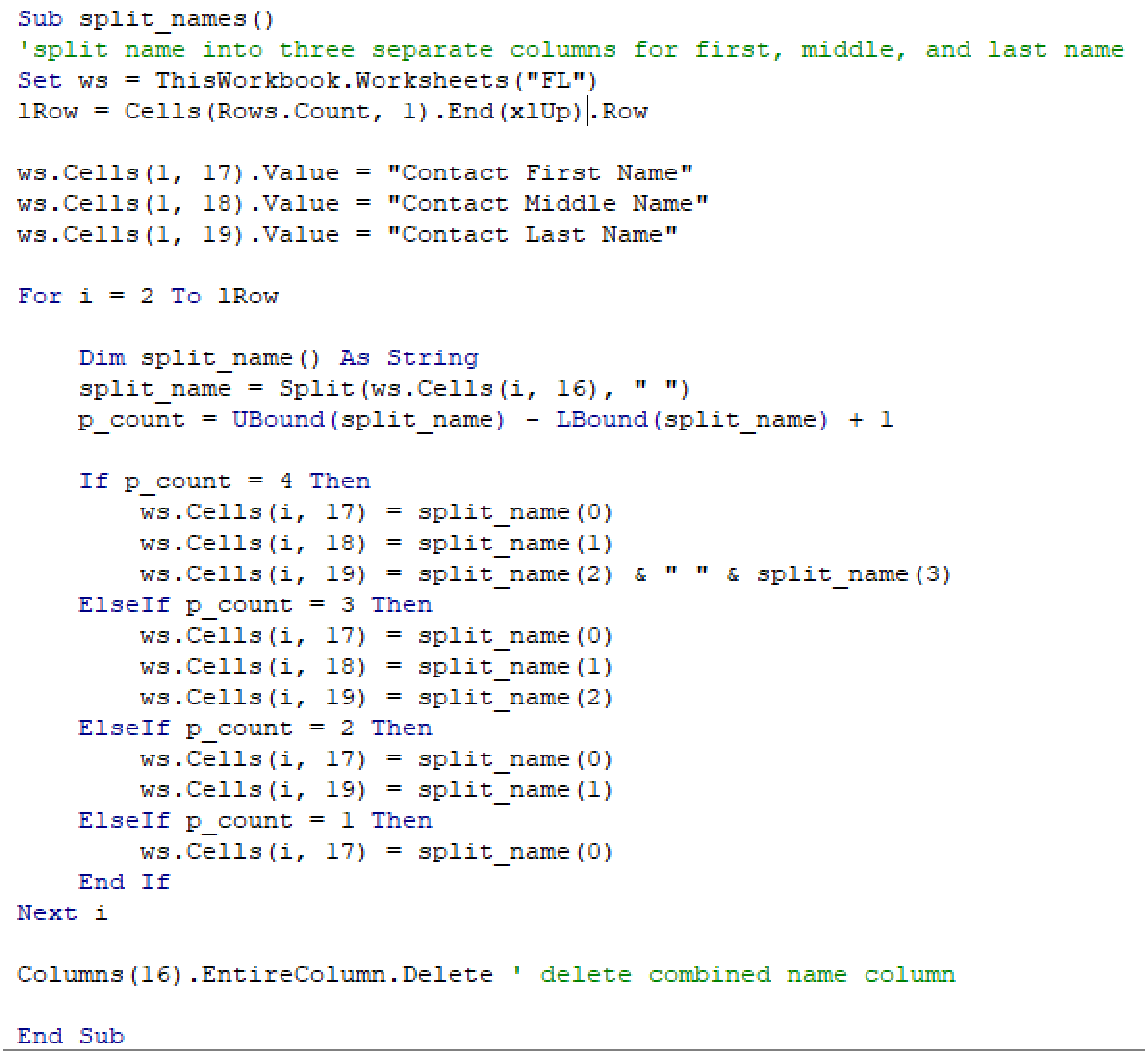


Andrea J Hulbert
Andrea Johnson
Andrea Lopez
Andrea Lopez
Andrea Mackey
Andrea Ramos
Andrew A Reich
Andrew Agwunobi

After Name Split

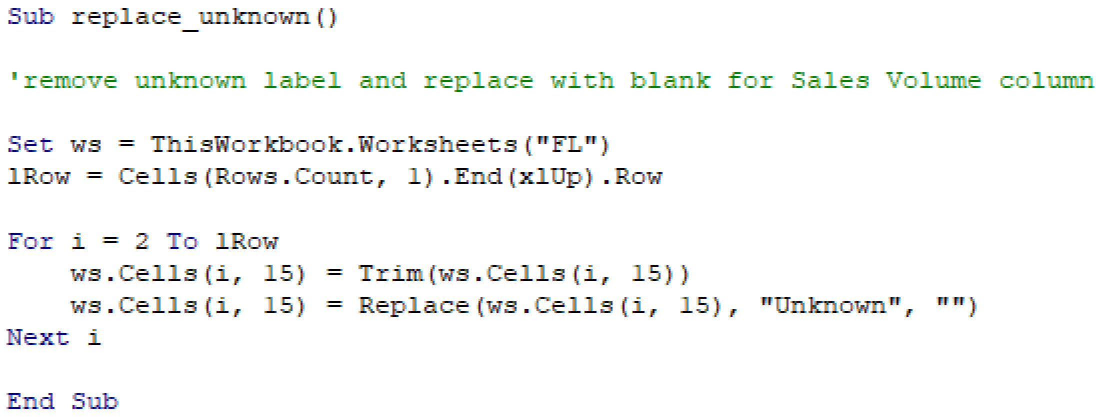


Andrea	J	Hulbert
Andrea		Johnson
Andrea		Lopez
Andrea		Lopez
Andrea		Mackey
Andrea		Ramos
Andrew	A	Reich
Andrew		Agwunobi



```
Sub split_names()
'split name into three separate columns for first, middle, and last name
Set ws = ThisWorkbook.Worksheets("FL")
lRow = Cells(Rows.Count, 1).End(xlUp).Row

ws.Cells(1, 17).Value = "Contact First Name"
ws.Cells(1, 18).Value = "Contact Middle Name"
ws.Cells(1, 19).Value = "Contact Last Name"

For i = 2 To lRow

    Dim split_name() As String
    split_name = Split(ws.Cells(i, 16), " ")
    p_count = UBound(split_name) - LBound(split_name) + 1

    If p_count = 4 Then
        ws.Cells(i, 17) = split_name(0)
        ws.Cells(i, 18) = split_name(1)
        ws.Cells(i, 19) = split_name(2) & " " & split_name(3)
    ElseIf p_count = 3 Then
        ws.Cells(i, 17) = split_name(0)
        ws.Cells(i, 18) = split_name(1)
        ws.Cells(i, 19) = split_name(2)
    ElseIf p_count = 2 Then
        ws.Cells(i, 17) = split_name(0)
        ws.Cells(i, 19) = split_name(1)
    ElseIf p_count = 1 Then
        ws.Cells(i, 17) = split_name(0)
    End If
Next i

Columns(16).EntireColumn.Delete ' delete combined name column

End Sub
```

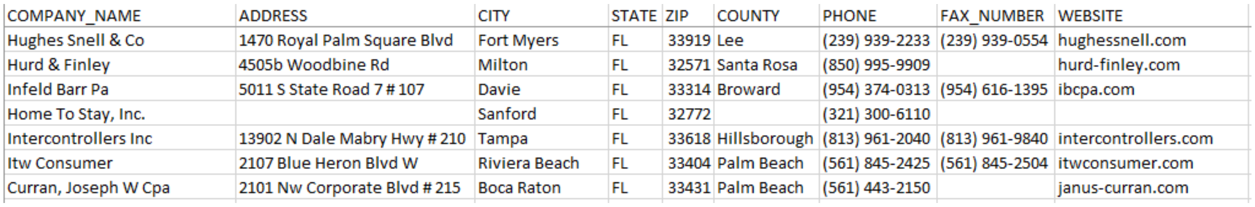
8) For the **Sales Volume Range** column, remove the Unknown label and replace with blank.



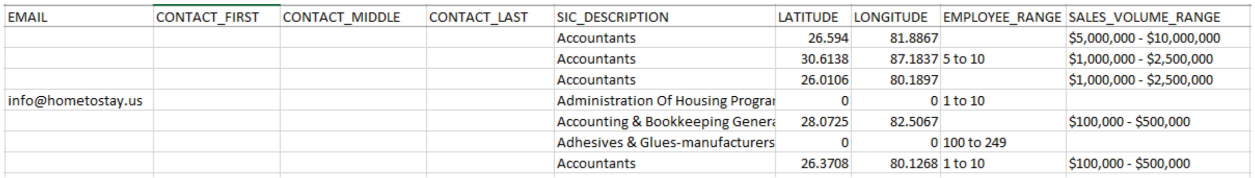
```
Sub replace_unknown()

'remove unknown label and replace with blank for Sales Volume column

Set ws = ThisWorkbook.Worksheets("FL")
lRow = Cells(Rows.Count, 1).End(xlUp).Row

For i = 2 To lRow
    ws.Cells(i, 15) = Trim(ws.Cells(i, 15))
    ws.Cells(i, 15) = Replace(ws.Cells(i, 15), "Unknown", "")
Next i

End Sub
```

<u>Sample of Final Deliverable*</u>

Our final deliverable show consistent text formating across the dataset, phone data is properly formated, emails were collected on the appropate columns and collapsed data (names) was expanded for increased data granulity.



COMPANY_NAME	ADDRESS	CITY	STATE	ZIP	COUNTY	PHONE	FAX_NUMBER	WEBSITE
Hughes Snell & Co	1470 Royal Palm Square Blvd	Fort Myers	FL	33919	Lee	(239) 939-2233	(239) 939-0554	hughessnell.com
Hurd & Finley	4505b Woodbine Rd	Milton	FL	32571	Santa Rosa	(850) 995-9909		hurd-finley.com
Infeld Barr Pa	5011 S State Road 7 # 107	Davie	FL	33314	Broward	(954) 374-0313	(954) 616-1395	ibcpa.com
Home To Stay, Inc.		Sanford	FL	32772		(321) 300-6110		
Intercontrollers Inc	13902 N Dale Mabry Hwy # 210	Tampa	FL	33618	Hillsborough	(813) 961-2040	(813) 961-9840	intercontrollers.com
Itw Consumer	2107 Blue Heron Blvd W	Riviera Beach	FL	33404	Palm Beach	(561) 845-2425	(561) 845-2504	itwconsumer.com
Curran, Joseph W Cpa	2101 Nw Corporate Blvd # 215	Boca Raton	FL	33431	Palm Beach	(561) 443-2150		janus-curran.com



EMAIL	CONTACT_FIRST	CONTACT_MIDDLE	CONTACT_LAST	SIC_DESCRIPTION	LATITUDE	LONGITUDE	EMPLOYEE_RANGE	SALES_VOLUME_RANGE
				Accountants	26.594	81.8867		$5,000,000 - $10,000,000
				Accountants	30.6138	87.1837	5 to 10	$1,000,000 - $2,500,000
				Accountants	26.0106	80.1897		$1,000,000 - $2,500,000
info@hometostay.us				Administration Of Housing Prograr	0	0	1 to 10	
				Accounting & Bookkeeping Genera	28.0725	82.5067		$100,000 - $500,000
				Adhesives & Glues-manufacturers	0	0	100 to 249	
				Accountants	26.3708	80.1268	1 to 10	$100,000 - $500,000

* there are still some clean up that can be done. The Employee Range and Sales Volume Range can be split into min-max columns for each. Additionally, some names included tittles such as Dr., Mr., etc. The name splitting code needs to be updated to capture those.