

*DataScience for Development and Social Change, 2015*

---

# Cleaning Data, part i

Getting data into usable  
formats

---

---

# Where we're heading

---

## How to Format your Data for Tableau Public

- Start your data in cell A1
- Have the first row be the column headers
- Have every subsequent row be one piece of data

---

# Data Cleaning

---

- ❖ The right format (e.g. CSV)
- ❖ The right shape (e.g. no 'extra' rows on top, one row per piece of data)
- ❖ Consistent labels
- ❖ Consistent “no data” values
- ❖ No junk symbols, whitespace etc.



# First, get some data

[popstats.unhcr.org/PSQ\\_POC.aspx](http://popstats.unhcr.org/PSQ_POC.aspx)

- ❖ 2009 to 2013
- ❖ all countries,
- ❖ all origins
- ❖ all data items
- ❖ click “submit”

popstats.unhcr.org/PSQ\_POC.aspx

Selection criteria

Date range: 2009 to 2013

Country / territory of residence

All countries / territories  
Afghanistan  
Albania  
Algeria  
Angola  
Antigua and Barbuda  
Argentina  
Armenia

Origin / Returned from

All origins  
Afghanistan  
Albania  
Algeria  
Andorra  
Angola  
Antigua and Barbuda  
Argentina

Data items to display

- |  |  |
|--|--|
| <input checked="" type="checkbox"/> Country / territory of residence | <input checked="" type="checkbox"/> Refugees                     |
| <input checked="" type="checkbox"/> Origin / Returned from           | <input checked="" type="checkbox"/> Asylum-seekers               |
|  | <input checked="" type="checkbox"/> Returned refugees            |
|  | <input checked="" type="checkbox"/> Internally displaced persons |
|  | <input checked="" type="checkbox"/> Returned IDPs                |
|  | <input checked="" type="checkbox"/> Stateless persons            |
|  | <input checked="" type="checkbox"/> Others of concern            |
|  | <input checked="" type="checkbox"/> Total population             |

Submit



# Gets you this file:

Extracted from the UNHCR Population Statistics Reference Database, United Nations High Commissioner for Refugees.										
Date extracted: 2015-04-15 17:02 +00:00										
Overview – Persons of concern to UNHCR										
Year	Country/ territory of residence	Origin / Returned from	Refugees	Asylum seekers	Returne d refugee s	IDPs	Returne d IDPs	Stateles s	Others of concern	Total populati on
2013	Afghanistan	Afghanistan				631286	21830		275486	928602
2013	Afghanistan	Azerbaijan			17					17
2013	Afghanistan	India			117					117
2013	Afghanistan	Iraq	*	*						*
2013	Afghanistan	Islamic Republic of Iran	36	18	8247					8301
2013	Afghanistan	Kazakhstan			*					*
2013	Afghanistan	Kyrgyzstan		*						*
2013	Afghanistan	Pakistan	16825	45	31224					48094
2013	Afghanistan	Palestinian	*							*
2013	Afghanistan	Russian Federation			19					19
2013	Afghanistan	Tajikistan		*	41					41
2013	Albania	China	12							12



---

# Cleaning with Python

---

- ❖ It's okay opening and cleaning up 1 file by hand... what if you have 100 files?
- ❖ Let's think about what we did, as “pseudocode”:
  - ❖ Read in each row of the csv file (e.g. “in.csv”).
  - ❖ Ignore the first 5 rows
  - ❖ Write the next row to another csv file (e.g. “out.csv”)
  - ❖ For each row after that:
    - ❖ Remove “\*” from all columns
    - ❖ Write the cleaned row to out.csv

# Cleaning with Python

```
import csv

csvinfile = '../Data_examples/UNHCR_popstats/2009_2013_popstats_PSQ_P0C.csv';
csvoutfile = 'cleaned_popstats.csv';
fin = open(csvinfile, "rb");
fout = open(csvoutfile, "wb");
csvin = csv.reader(fin);
csvout = csv.writer(fout, quoting=csv.QUOTE_NONNUMERIC);

print("Headers:");
for i in range(0,6):
    header = csvin.next();
    print("Header {}".format(",".join(header)))
    csvout.writerow(header);

for row in csvin:
    print("Data {}".format(",".join(row)))
    cleanedrow = ['' if value == '*' else value for value in row]
    csvout.writerow(cleanedrow);

fin.close();
fout.close();
```



# The cleaned file

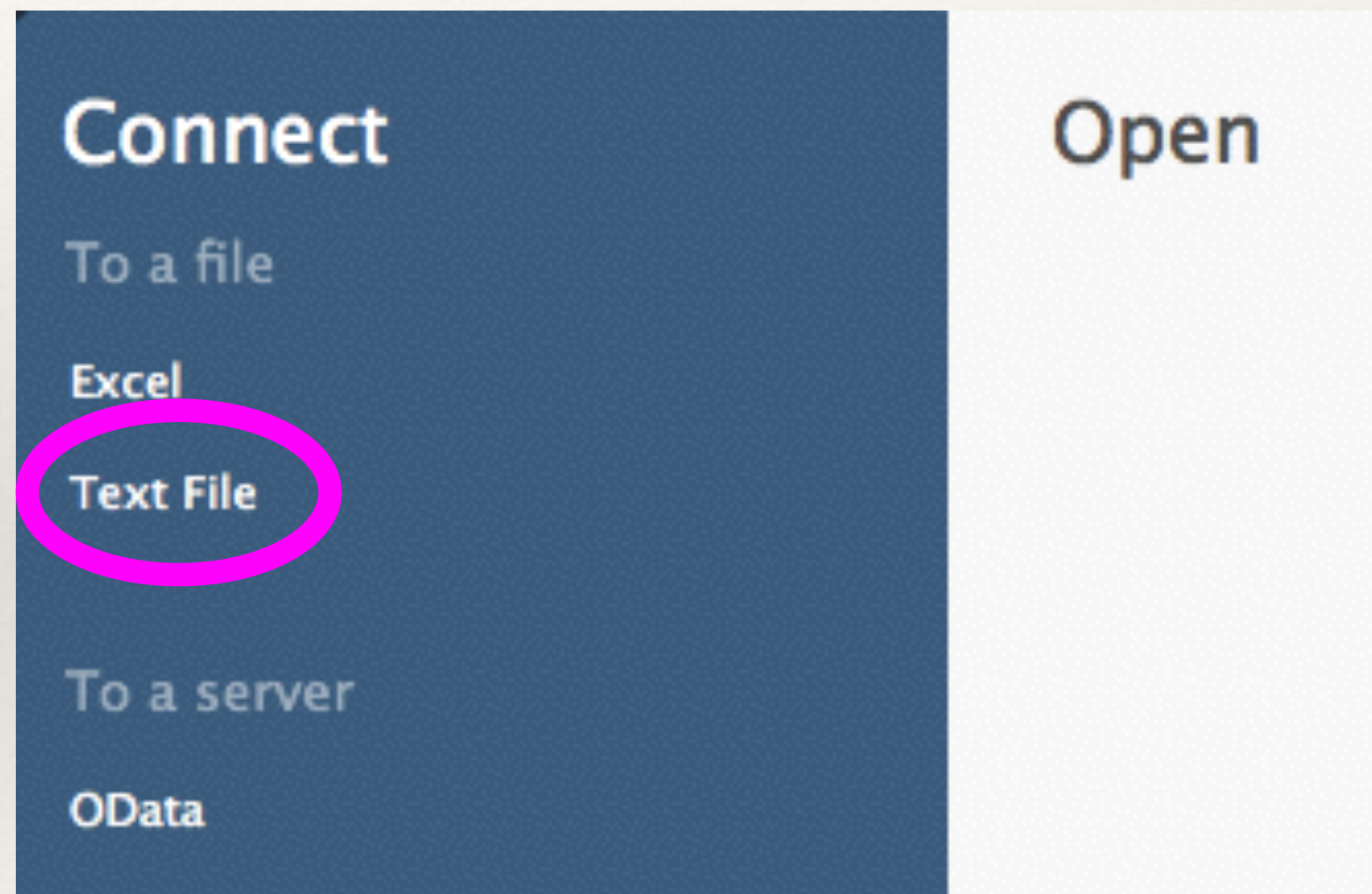
Year	Country/ territory of residence	Origin / Returned from	Refugees	Asylum seekers	Returne d refugee s	IDPs	Returne d IDPs	Stateles s	Others of concern	Total populati on
2013	Afghanistan	Afghanistan				631286	21830		275486	928602
2013	Afghanistan	Azerbaijan			17					17
2013	Afghanistan	India			117					117
2013	Afghanistan	Iraq								
2013	Afghanistan	Islamic Republic of Iran	36	18	8247					8301
2013	Afghanistan	Kazakhstan								
2013	Afghanistan	Kyrgyzstan								
2013	Afghanistan	Pakistan	16825	45	31224					48094
2013	Afghanistan	Palestinian								
2013	Afghanistan	Russian Federation			19					19
2013	Afghanistan	Tajikistan			41					41
2013	Albania	China	12							12

---


# Let's take it for a spin in Tableau

---

- ❖ Open Tableau Public
- ❖ Click on “text file”
- ❖ Open your newly-cleaned file



# You've got data

 cleaned\_popstats

Connected to Text File

Directory



/Users/sara/Dropbox\_Personal/DO...

Files

Enter file name


cleaned\_popstats.csv


cleaned\_popstats.csv






Copy

☐ Show aliases ☐ Show hidden fields Rows 10,000 →

Year #	Country/territory of... 	Origin / Returned f... Abc	Refugees #	Asylum seekers #	Returned refugees #	ID P #
2013	Afghanistan	Afghanistan	null	null	null	
2013	Afghanistan	Azerbaijan	null	null	17	
2013	Afghanistan	India	null	null	117	
2013	Afghanistan	Iraq	null	null	null	
2013	Afghanistan	Islamic Republic o...	36	18	8,247	
2013	Afghanistan	Kazakhstan	null	null	null	

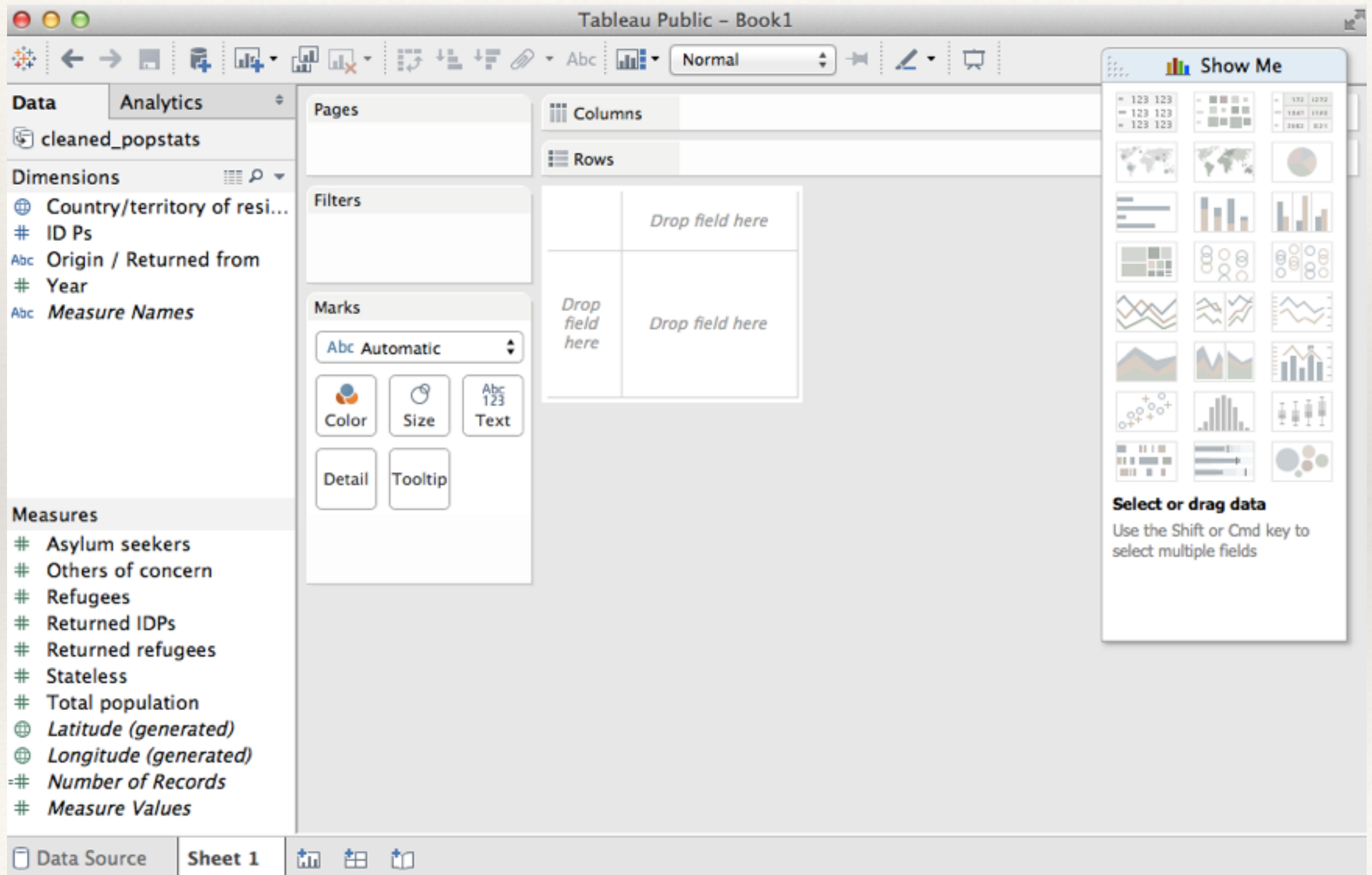
 Go to Worksheet

Sheet 1

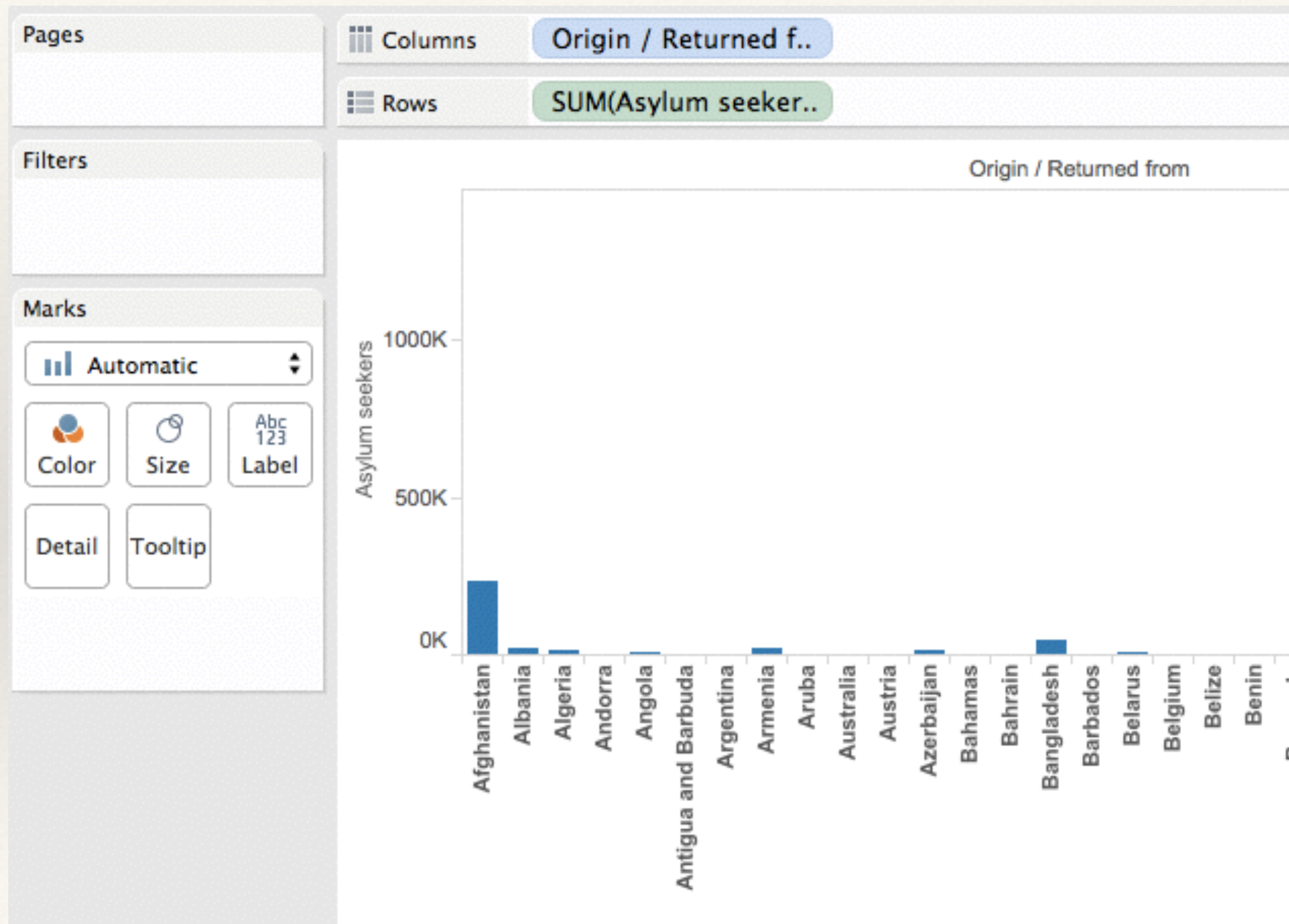




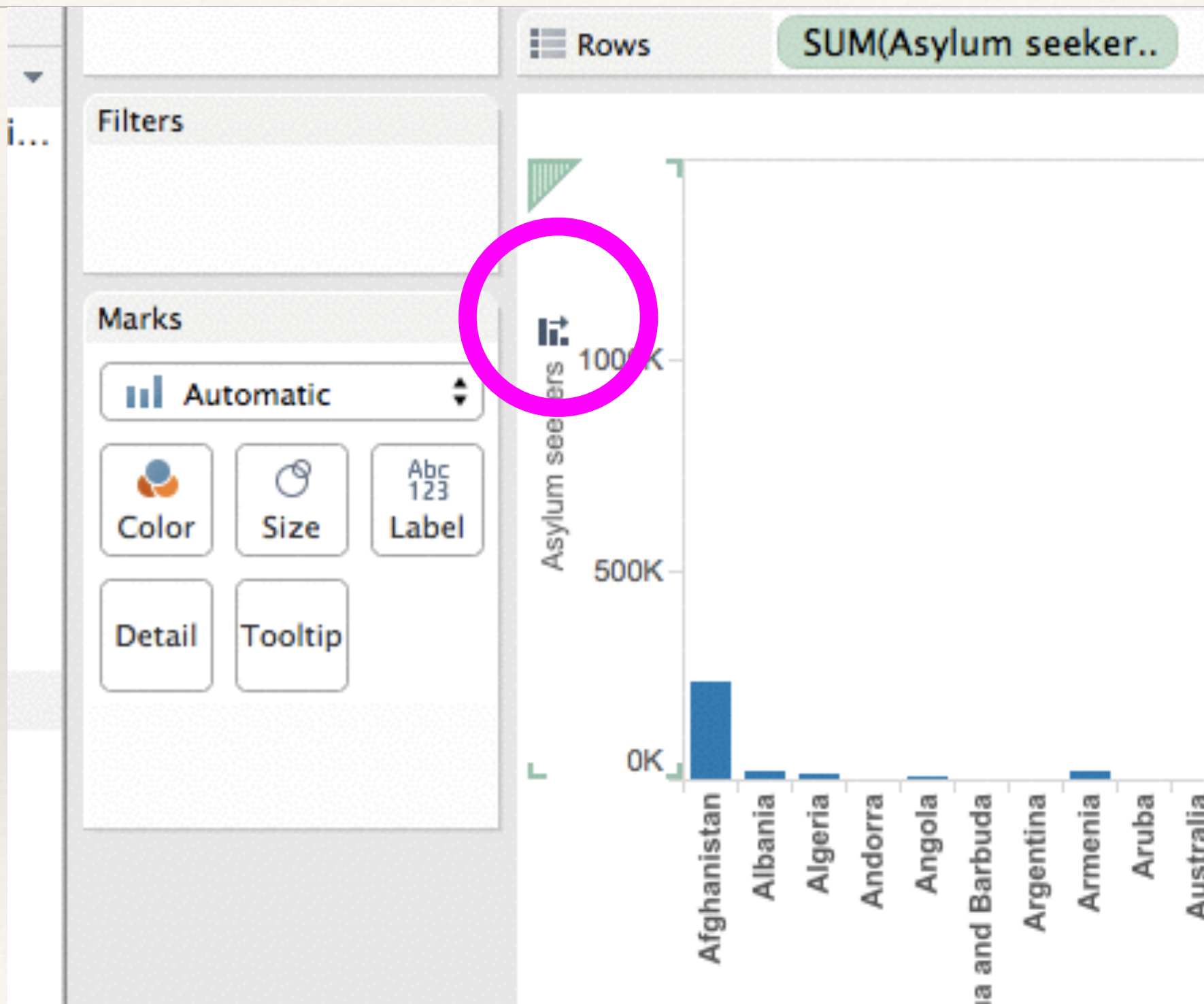
# Loaded into Tableau



# Let's make a column chart

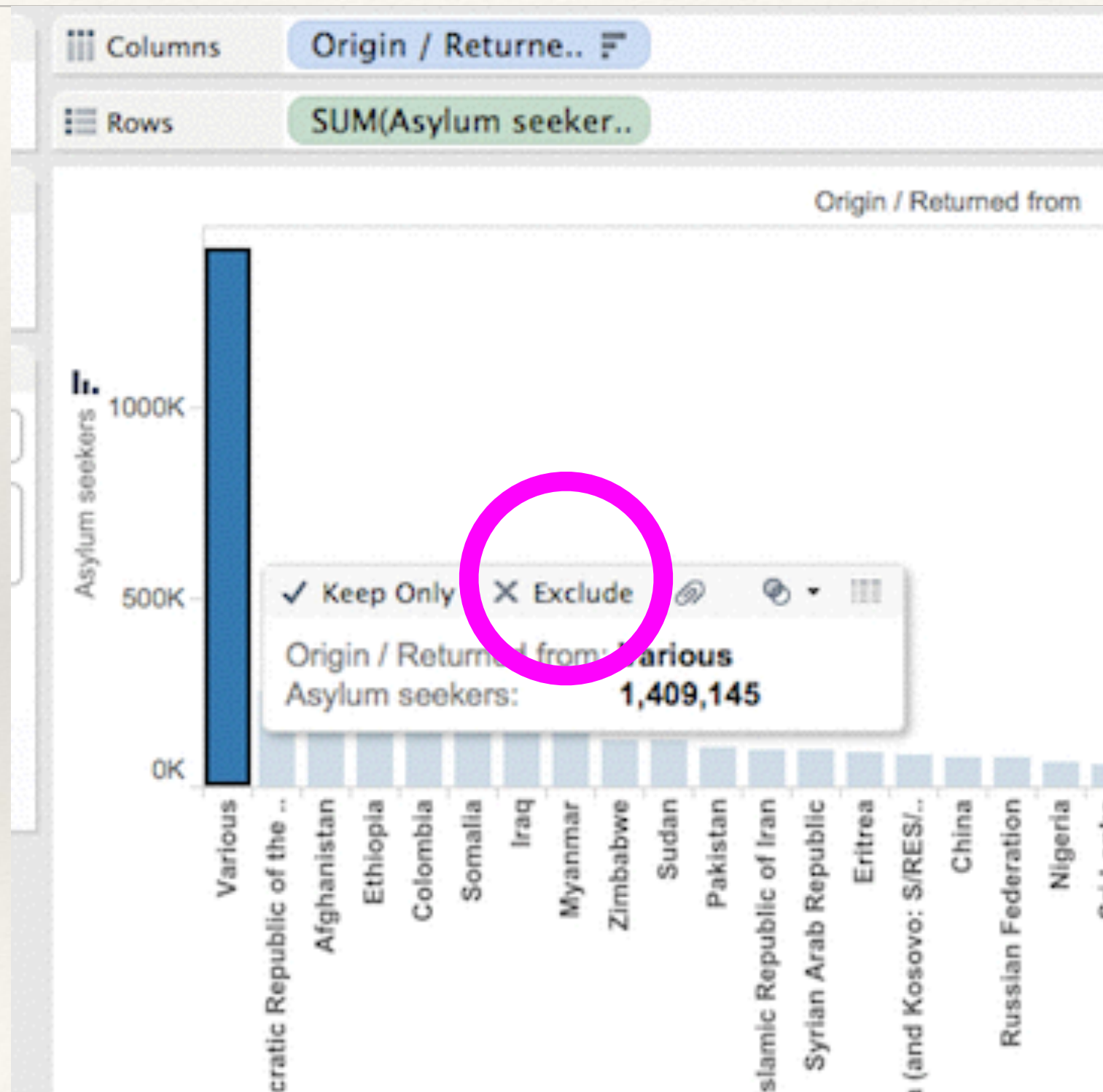


# Sort the columns

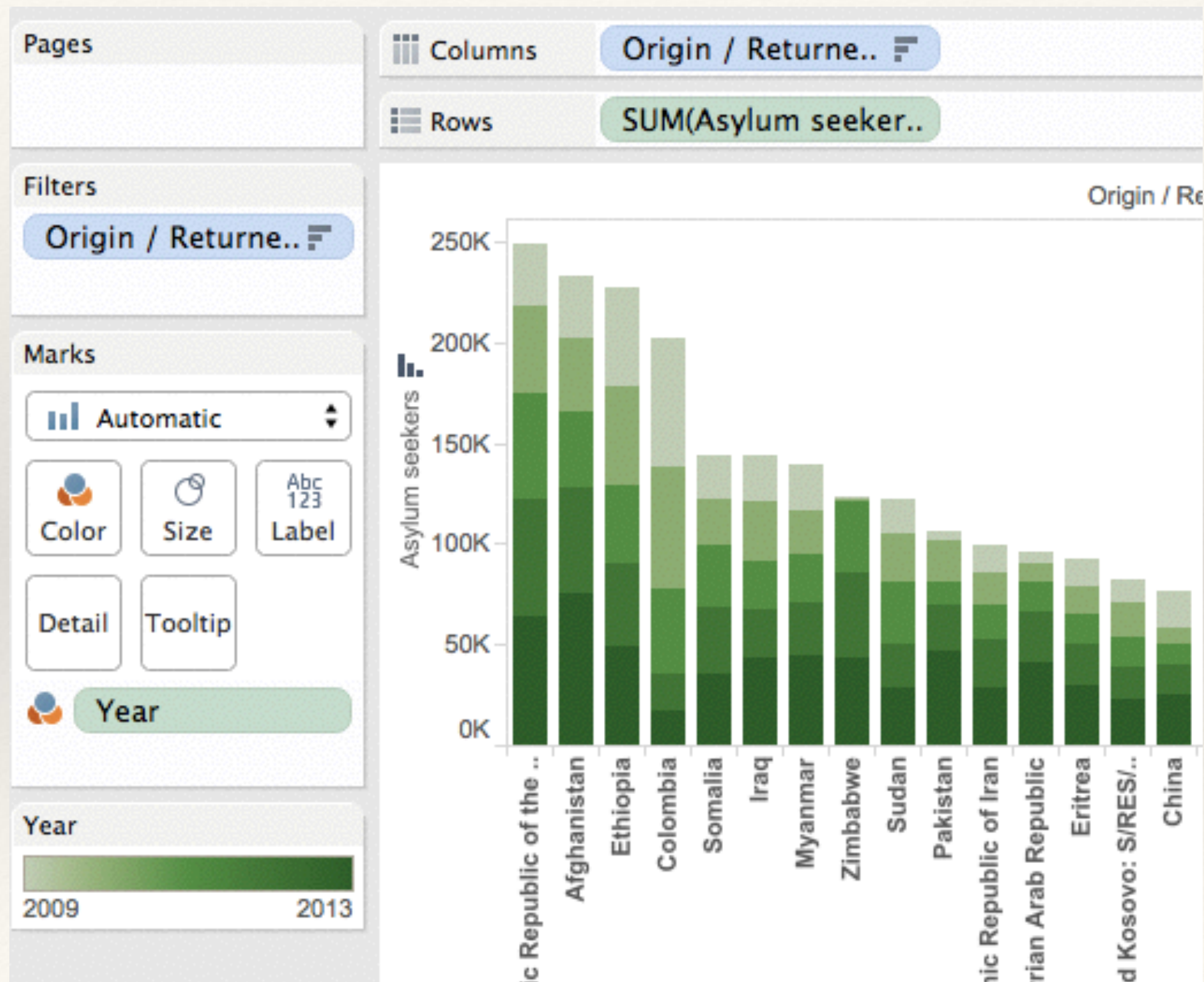




# Get rid of unwanted values

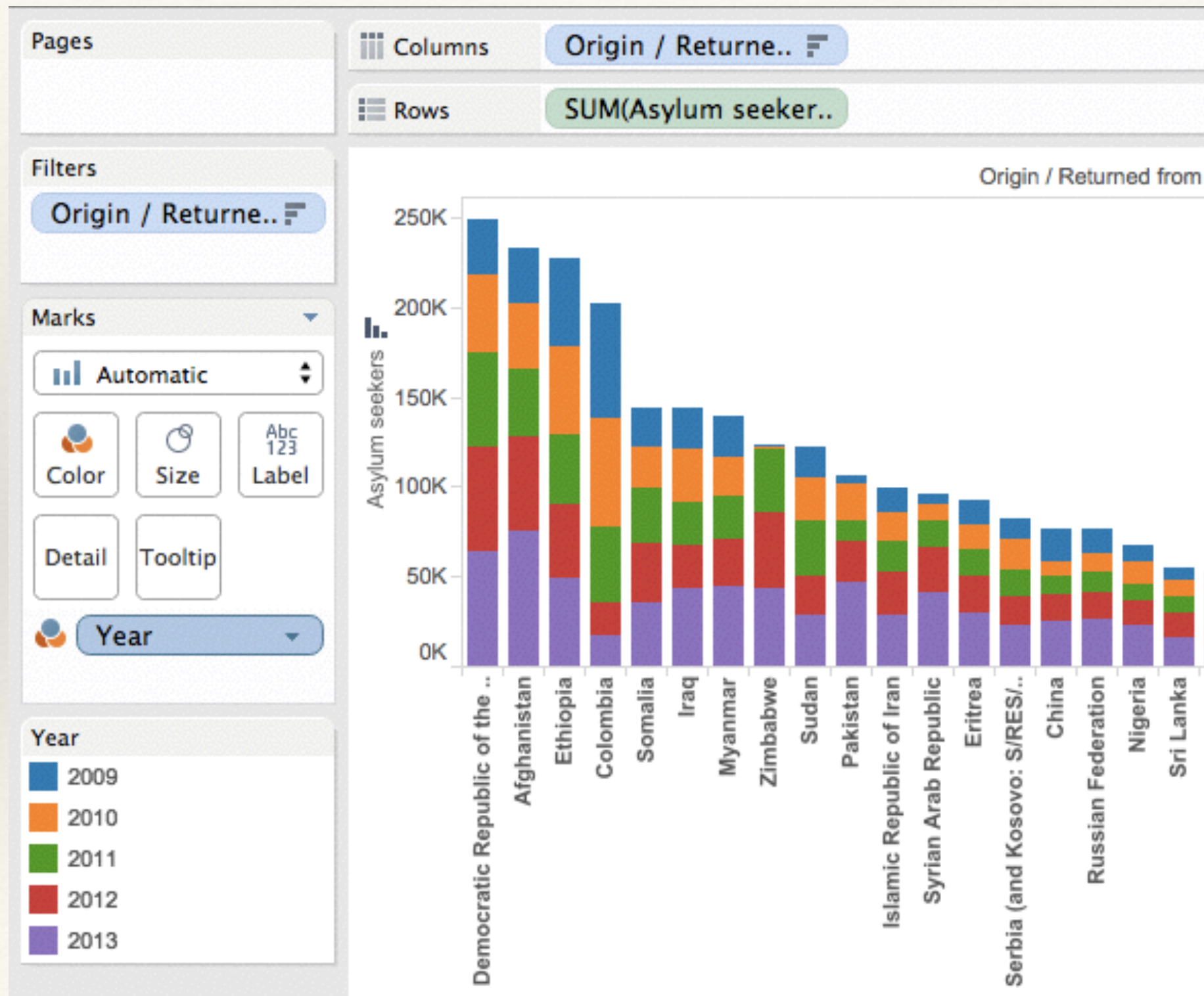


# But wait, there's more





# Clean up the colors





---

# Exercise!

---

- ❖ Think about the cleaning you'd need to do to your data
  - ❖ Does it have missing values? Missing = no data, data marked as missing ("-9999", "-1", "n/a"), missing dates, missing areas etc
  - ❖ Do you have different datasets? Do they code things differently (e.g. 1 / 0 vs m / f vs male / female etc)?