Introduction:

This write-up is about different ways we can visualize UN Migrant Stock data. We will be conducting exploratory data analysis as our first step in analysis in order to find some useful patterns and insights. Visualizing data would also help us in better understanding our dataset.

Description:

UN Dataset contains data on trends of international migrant stock. It is a revised version for year 2015. It has data for every five years starting from 1990. After applying tidy data principles on this dataset, we have 19 columns and 4176 rows. Details are shared on the screenshot below.

	country_code	country	major_area	major_area_code	region	region_code	sub_saharan_africa	IMS	year	sex	total_population
0	4	Afghanistan	Asia	935	Southern Asia	5501	No	57686	1990	both	12067.57
1	4	Afghanistan	Asia	935	Southern Asia	5501	No	71522	1995	both	16772.52
2	4	Afghanistan	Asia	935	Southern Asia	5501	No	75917	2000	both	19701.94
3	4	Afghanistan	Asia	935	Southern Asia	5501	No	87300	2005	both	24399.95
4	4	Afghanistan	Asia	935	Southern Asia	5501	No	102246.0	2010	both	27962.21

total_population	IMS_total_population	female_IMS	RMS	RMS_IMS	foreign_pop_status	refugee_incl	imputation	country_classification
12067.57	0.48	NaN	25	0.04	born	no	no	least
16772.52	0.43	NaN	19605	27.41	born	no	no	least
19701.94	0.39	NaN	0	0	born	no	no	least
24399.95	0.36	NaN	32	0.04	born	no	no	least
27962.21	0.37	NaN	6434.0	6.29	born	no	no	least

As our first step, we will be conducting Exploratory Data Analysis and for that purpose we will be following Tufte's principles on data visualization.

Principle 1: Sort

When we want to simply explore our dataset and we are clueless on where to start from we can always sort the data in ascending or descending in order to find patterns which might help in explaining our dataset better.

Here, I have sorted the column for International Migrant Stock (IMS) in descending order so that I can see the highest value in the column.

	country_code	country	major_area	major_area_code	region	region_code	sub_saharan_africa	IMS	year	sex	total_populatio
3965	840	United States of America	Northern America	905	Northern America	905	No	46627102.0	2015	both	321773.6
3964	840	United States of America	Northern America	905	Northern America	905	No	44183643.0	2010	both	309876.1
3963	840	United States of America	Northern America	905	Northern America	905	No	39258293	2005	both	296139.6
3962	840	United States of America	Northern America	905	Northern America	905	No	34814053	2000	both	282895.7
3961	840	United States of America	Northern America	905	Northern America	905	No	28451053	1995	both	266275.5

In the dataframe above, we can see that United States of America has the highest amount of international migrant stock in year 2015. This also shows that United States of America has been dominating for the last 25 years as the all the top rows are of United States of America.

	America									
840	United States of America	Northern America	905	Northern America	905	No	34814053	2000	both	
840	United States of America	Northern America	905	Northern America	905	No	28451053	1995	both	
840	United States of America	Northern America	905	Northern America	905	No	23802795.0	2015	female	
840	United States of America	Northern America	905	Northern America	905	No	23251026	1990	both	
840	United States of America	Northern America	905	Northern America	905	No	22824307.0	2015	male	
	United									

By exploring a bit further, we see that maximum amount of international migrant stock constitutes of majority of female population.

Similarly, we can also sort in ascending order to see which country has the smallest number of international migrant stock.

	country_code	country	major_area	major_area_code	region	region_code	sub_saharan_africa	IMS	year	sex
3869	798	Tuvalu	Oceania	909	Polynesia	957	No	63.0	2015	female
3868	798	Tuvalu	Oceania	909	Polynesia	957	No	69.0	2010	female
3863	798	Tuvalu	Oceania	909	Polynesia	957	No	78.0	2015	male
3867	798	Tuvalu	Oceania	909	Polynesia	957	No	82	2005	female
3862	798	Tuvalu	Oceania	909	Polynesia	957	No	85.0	2010	male

Here, we can see that Tuvalu, a country in Oceania had the lowest quantity of IMS in 2015.

Principle 2: Group

By grouping our data, we can look at our dataset in different dimensions. We can also call aggregate function in order to derive some findings.

I grouped data according to region, sex and year to compare the trends of IMS over the span of last 25 years.

	year	1990	1995	2000	2005	2010	2015
region	sex						
Australia and New Zealand	**both**	4473260	4741947	5065063	5717982	6.83e+06	7.80e+06
	female	2201841	2359437	2545236	2891102	3.45e+06	3.96e+06
	male	2271419	2382510	2519827	2826880	3.38e+06	3.84e+06
Caribbean	**both**	1056555	1154700	1255647	1328740	1.35e+06	1.37e+06
	female	506889	558008	610543	648026	6.57e+05	6.66e+05
...	**...**	...	...	...	...	...	...
Western Asia	**female**	6317950	6960145	7496995	8294747	1.07e+07	1.36e+07
	male	8921324	9996494	11062014	13363900	1.99e+07	2.46e+07
Western Europe	**both**	16237829	18318040	20425378	22776284	2.52e+07	2.74e+07
	female	7731651	8965592	10212028	11613877	1.32e+07	1.42e+07
	male	8506178	9352448	10213350	11162407	1.21e+07	1.32e+07

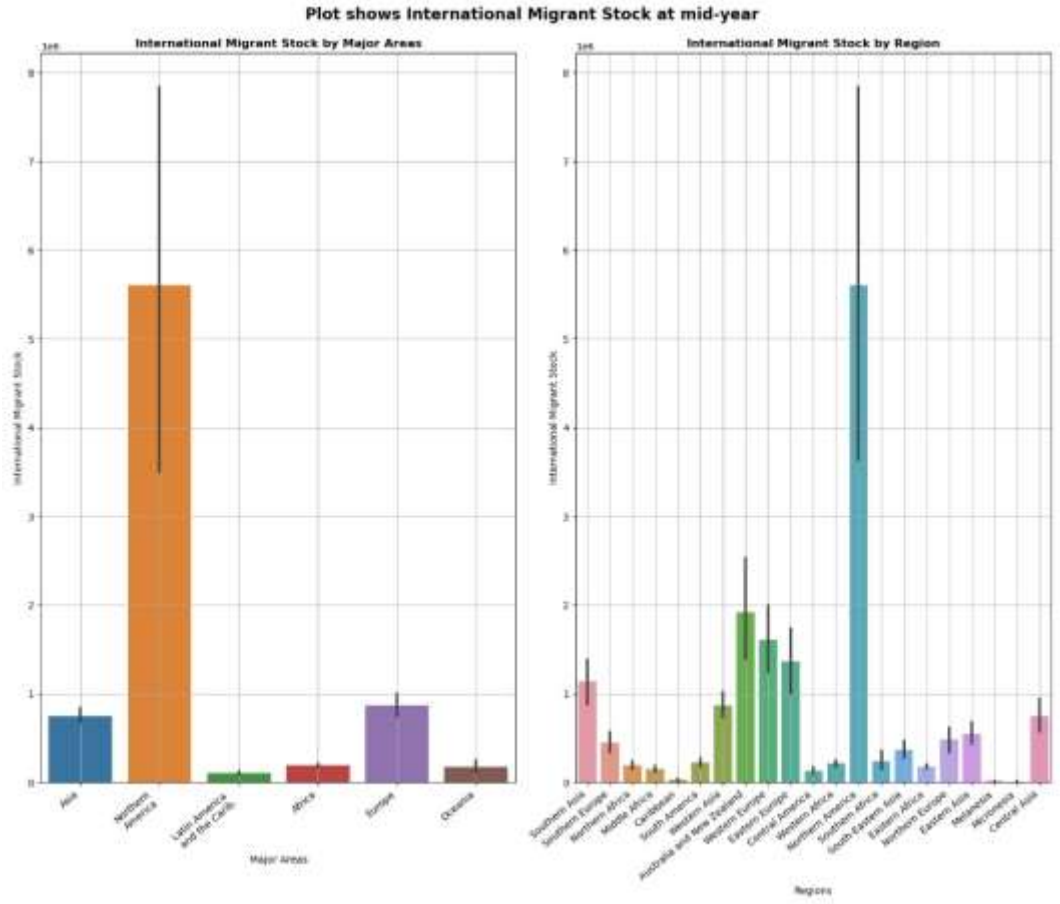
66 rows x 6 columns

I also grouped Refugee migrant stock to see patterns on how it evolved over time.

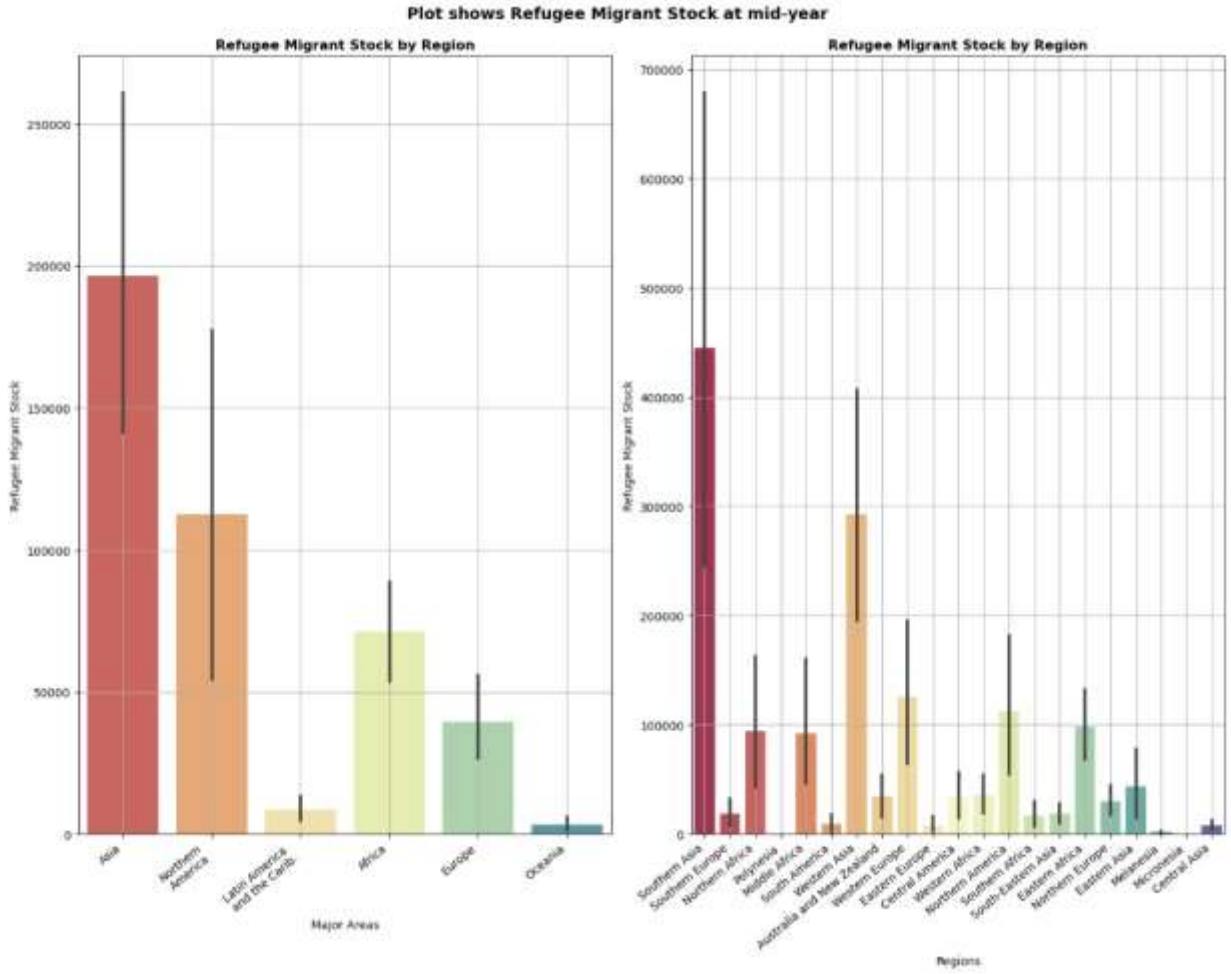
	year	1990	1995	2000	2005	2010	2015
major_area							
Africa		5687352	5949953	3609138	2750644	2.41e+06	4.02e+06
Asia		9937007	7937682	9175210	8109615	1.05e+07	1.32e+07
Europe		1321884	2746090	2283959	1760748	1.59e+06	1.48e+06
Latin America and the Caribbean		1197198	352256	44088	47186	3.76e+05	3.84e+05
Northern America		583450	775419	633376	526511	4.30e+05	4.24e+05
Oceania		109680	92440	82032	82029	3.25e+04	5.46e+04

But as you can see, it become really hard to compare these values especially when the number are huge. We need visualizations so we can see which region or area had the highest amount in which year at one glance.

For this purpose, I created bar charts for International Migrant Stock according to major areas and Regions. As they are easier to read when comparing many different areas or regions with each other.



In the bar graph above, we can instantly tell that ‘Orange’ bar which is for Northern America has the highest number of international migrant stock and we look at region wise even then Northern America beats the rest by a huge difference.



The plot above show refugee migrant stock at mid-year. Here we can tell that red bar has the highest number, which is approximately 250,000, especially Southern Asia has the highest density. Therefore, by comparing the two plots we can say that Northern America dominates in International Migrant Stock whereas Southern Asia dominates in Refugee Migrant Stock.

Principle 3: Subset

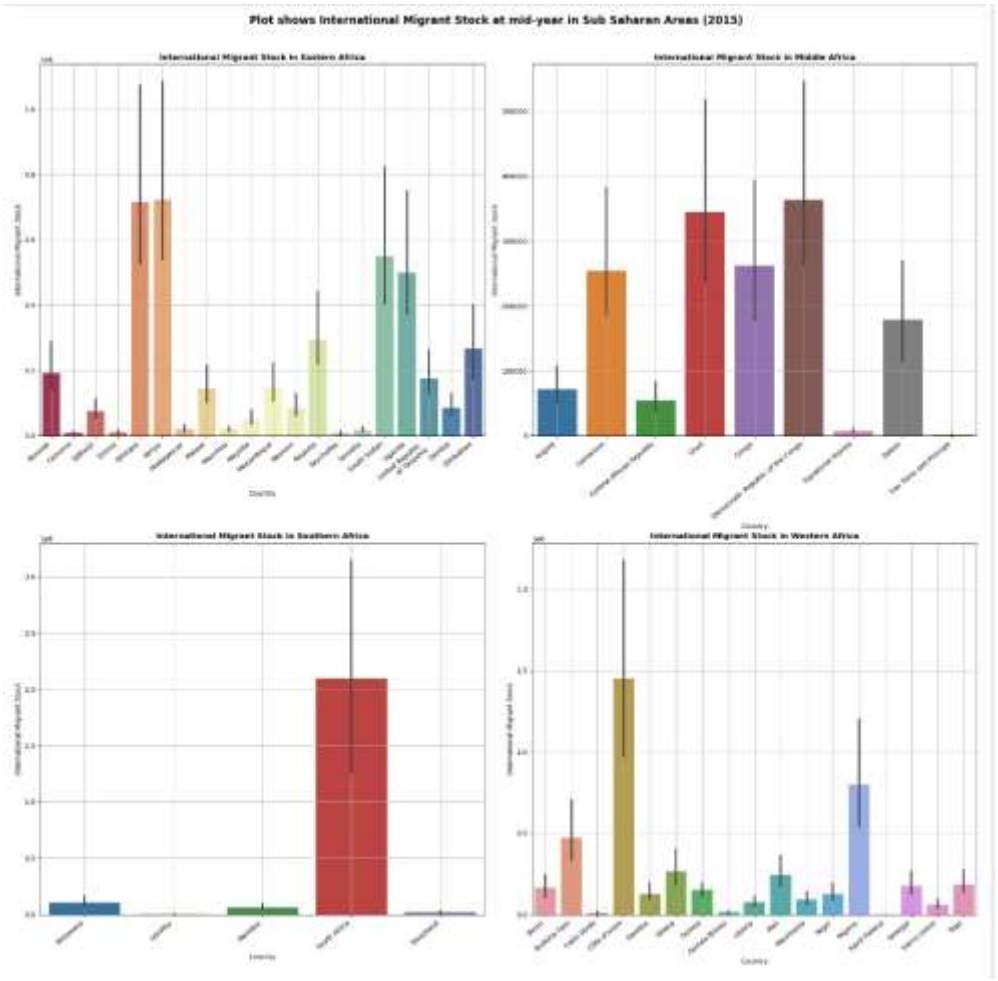
This bring us to Tufte's third principle which is to create subset of the dataset in order to study it in depth. For this principle, I decided to create a subset for Sub-Saharan Africa. I ran a condition on the dataset for sub-Saharan regions and created a separate dataset for them.

	country_code	country	major_area	major_area_code	region	region_code	sub_saharan_africa	IMS	year	sex	total_population	IMS_total_populat
90	24	Angola	Africa	903	Middle Africa	911	Yes	33517	1990	both	11127.87	
91	24	Angola	Africa	903	Middle Africa	911	Yes	39813	1995	both	13042.67	0
92	24	Angola	Africa	903	Middle Africa	911	Yes	46108	2000	both	15058.64	0
93	24	Angola	Africa	903	Middle Africa	911	Yes	61329	2005	both	17912.94	0
94	24	Angola	Africa	903	Middle Africa	911	Yes	76549.0	2010	both	21219.95	0

Then I grouped this dataset according to different regions within the Sub-Saharan Africa.

region	year	1990	1995	2000	2005	2010	2015
	country						
Eastern Africa	Burundi	666220	509706	251256	345748	4.71e+05	5.74e+05
	Comoros	28158	27878	27598	26418	2.52e+04	2.51e+04
	Djibouti	244442	199548	201014	184182	2.03e+05	2.25e+05
	Eritrea	23696	24800	25904	28628	3.14e+04	3.19e+04
	Ethiopia	2310780	1613808	1222768	1028484	1.14e+06	2.15e+06
	Kenya	594584	1237490	1398278	1513788	1.85e+06	2.17e+06
	Madagascar	47834	42354	47082	52116	5.78e+04	6.42e+04
	Malawi	2255448	483248	465240	443322	4.35e+05	4.30e+05
	Mauritius	7226	14986	31086	39294	4.97e+04	5.72e+04
	Mayotte	30458	52632	90948	126352	1.46e+05	1.54e+05

Now, to compare the number of International Migrant Stock within the Sub-Saharan Region I ran a condition to plot only for year 2015.



This plot shows international migrant stock at mid-year in 2015 in Eastern Africa, Middle Africa, Southern Africa and Western Africa. Creating subplots makes it easier to compare the different regions. We can see from the plot that Southern Africa has only one country which is South Africa that still has International Migrant Stock, all the others have negligible numbers.

Principle 4: Compare

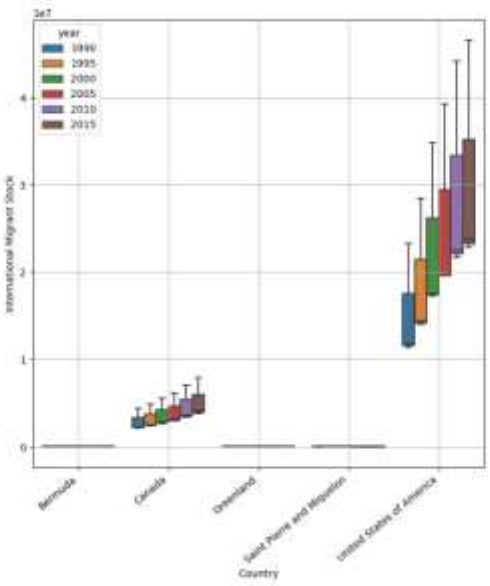
Tufte's fourth principle says to compare the subsets with each other in order to explain the data in a better way.

For this principle we will be looking at Northern America. Again, I created a subset for Northern America.

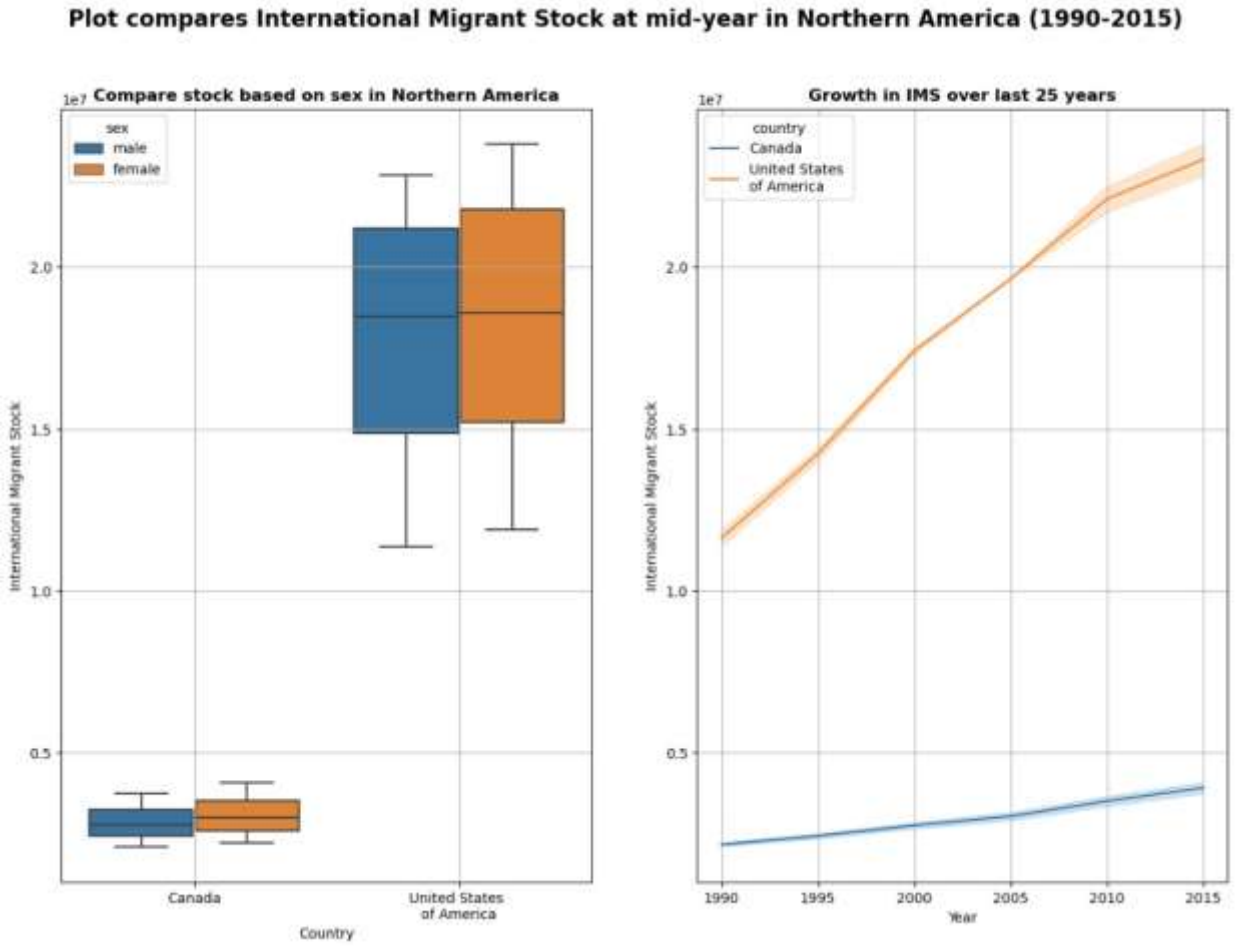
	country_code	country	major_area	major_area_code	region	region_code	sub_saharan_africa	IMS	year
396	60	Bermuda	Northern America	905	Northern America	905	No	15683	1990
397	60	Bermuda	Northern America	905	Northern America	905	No	16676	1995
398	60	Bermuda	Northern America	905	Northern America	905	No	17668	2000
399	60	Bermuda	Northern America	905	Northern America	905	No	18276	2005
400	60	Bermuda	Northern America	905	Northern America	905	No	18884.0	2010
...	...	...	...	...	...	...	...	...	...
		United	Northern		Northern				

I created box plot, because they are the best way to visualize the central tendency of data. I wanted to compare International Migrant Stock in different countries within Northern America.

Plot compares International Migrant Stock at mid-year in Northern America (1990-2015)

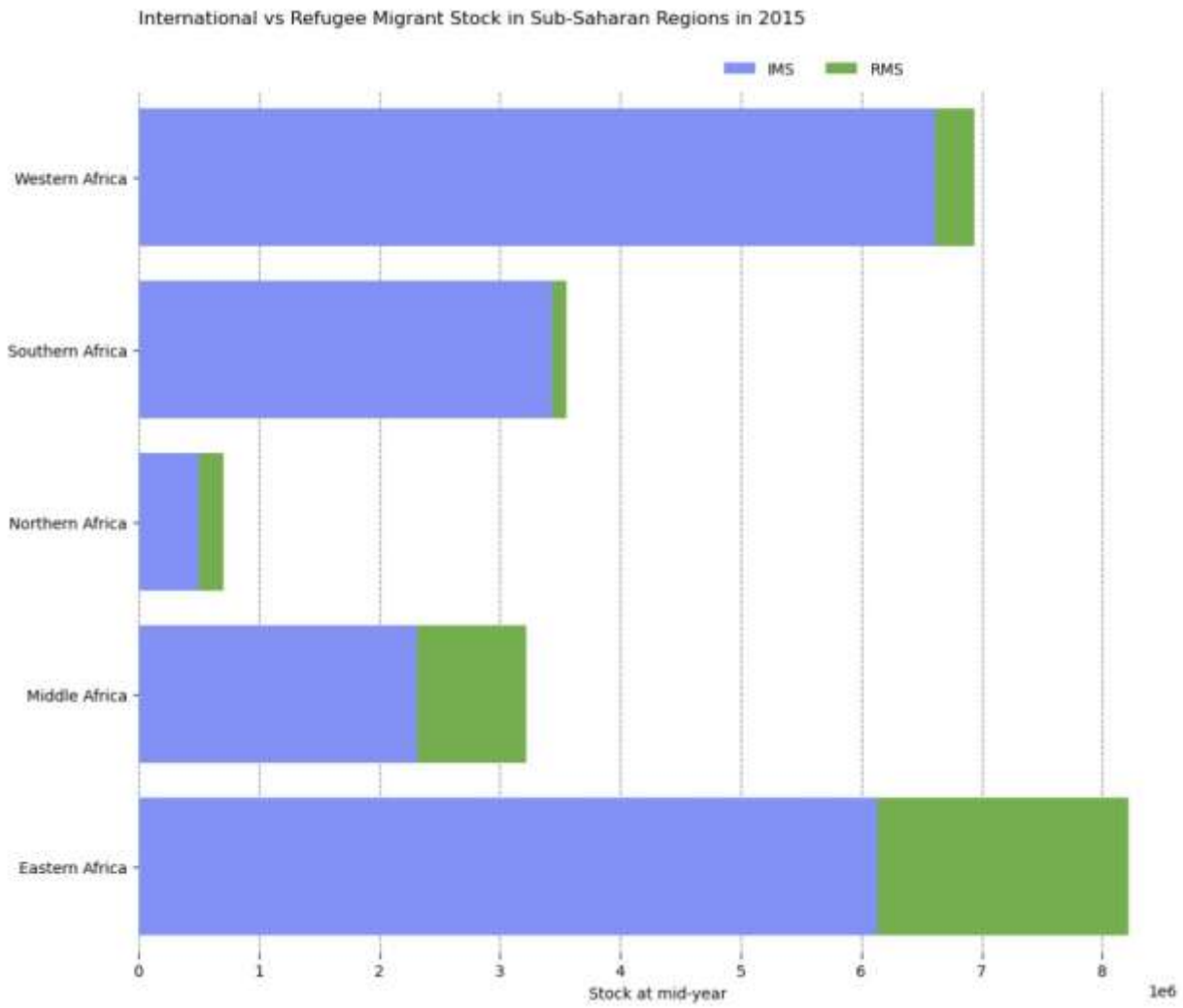


In the above plot, we can see that some of the countries have missing data so we will remove those and plot again. So that our box plots actually make sense.



Here, I created a box plot for two countries in Northern America for 1990-2015. Plot show the average for IMS over the last 25 years. Canada is much lower than Northern America. Male and Female averages are almost the same. Female population is slightly higher. Second plot shows the growth over the span of 25 years. The line plot for United States of America is way steeper than Canada. United States of America is growing in numbers whereas Canada did show some positive increase, but it is not as steep. Rate of change is much lower in Canada as compared to United States of America.

I created another plot to compare International Migrant Stock and Refugee Migrant Stock in Sub Saharan Region in the year 2015.

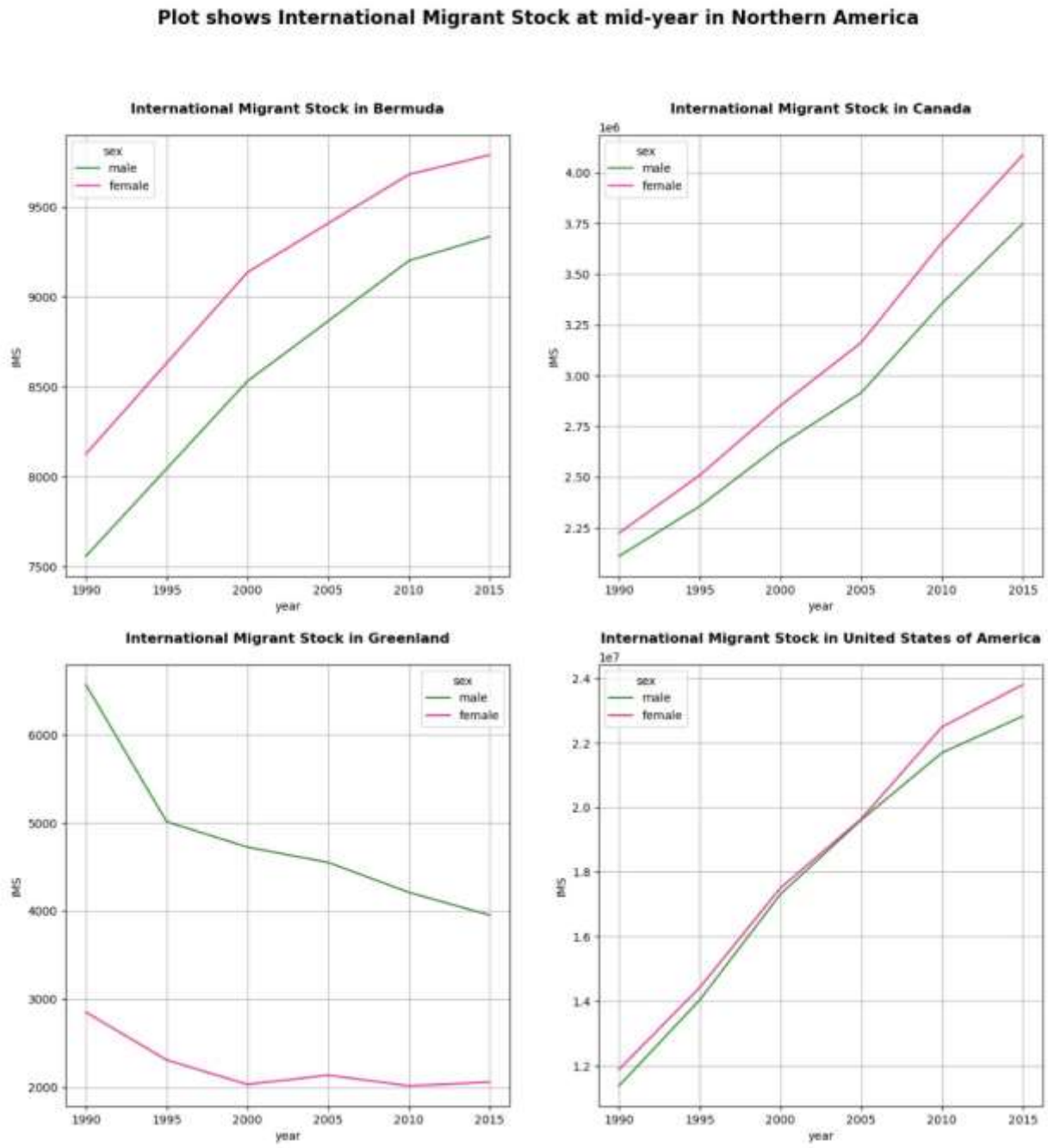


These horizontal stacked bars depict the difference in numbers for International Migrant Stock versus Refugee Migrant Stock. RMS is much lower in number than IMS. This plot also shows that Eastern Africa has the maximum number of RMS in 2015. But IMS is still three times higher in number. It is also a great way to compare different categories of data.

Principle 5: Small Multiples

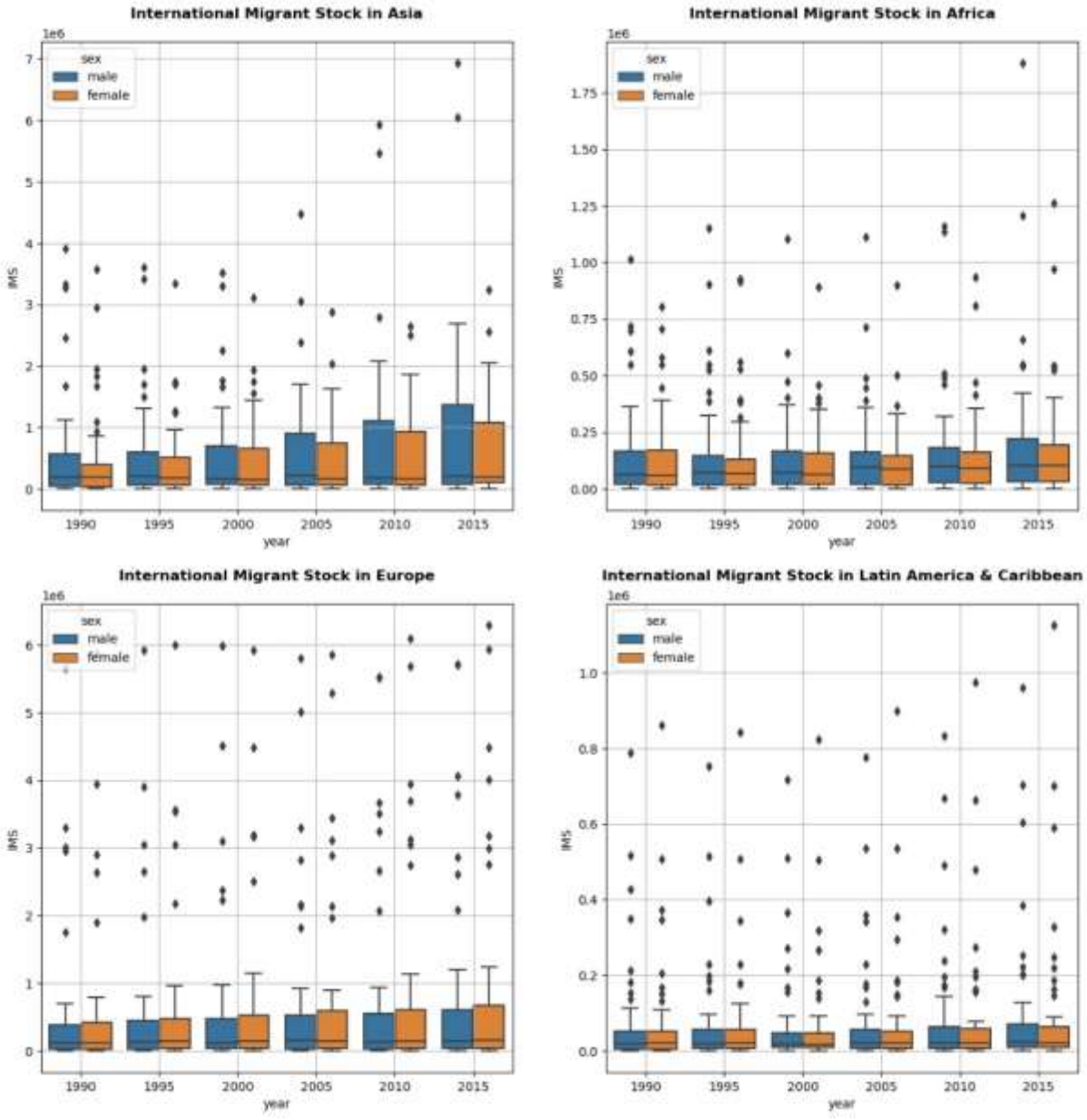
Tufte's fifth principle tells us to create multiple small plots so that its easier to compare data with each other.

For this principle I compared the International Migrant Stock from 1990 to 2015 in terms of male versus female population and how it grew in number.

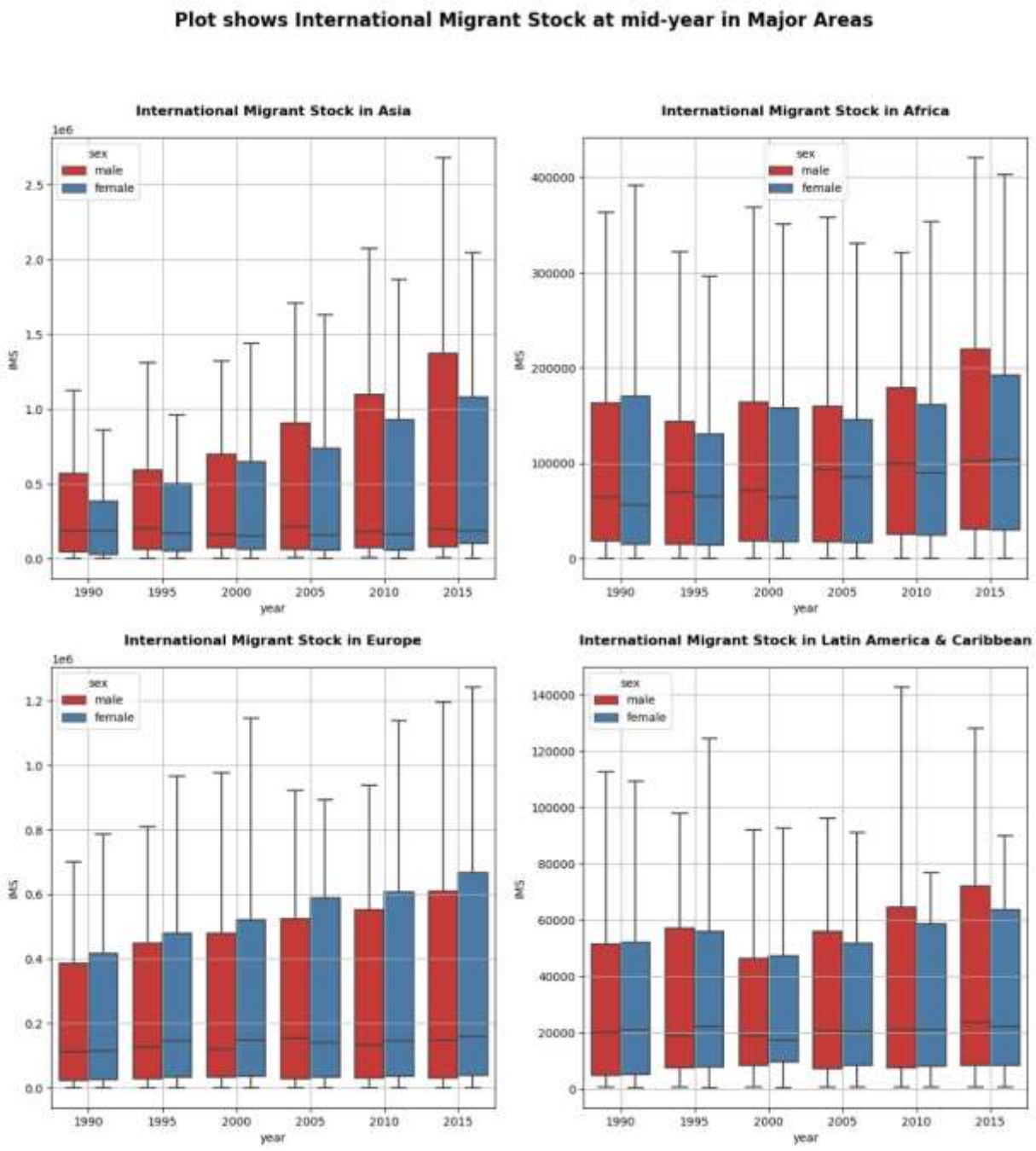


The plot visualizes data from four different countries within Northern America. It shows how male and female population has increased or decreased in number. From this plot, we can tell those United states of America faced an exponential increase in numbers of IMS whereas Greenland faced a decrease as the line plot in going downwards. We can also tell that United States, Bermuda and Canada has the highest rate of change.

I also created box plots to compare IMS over the years in major areas of the world.

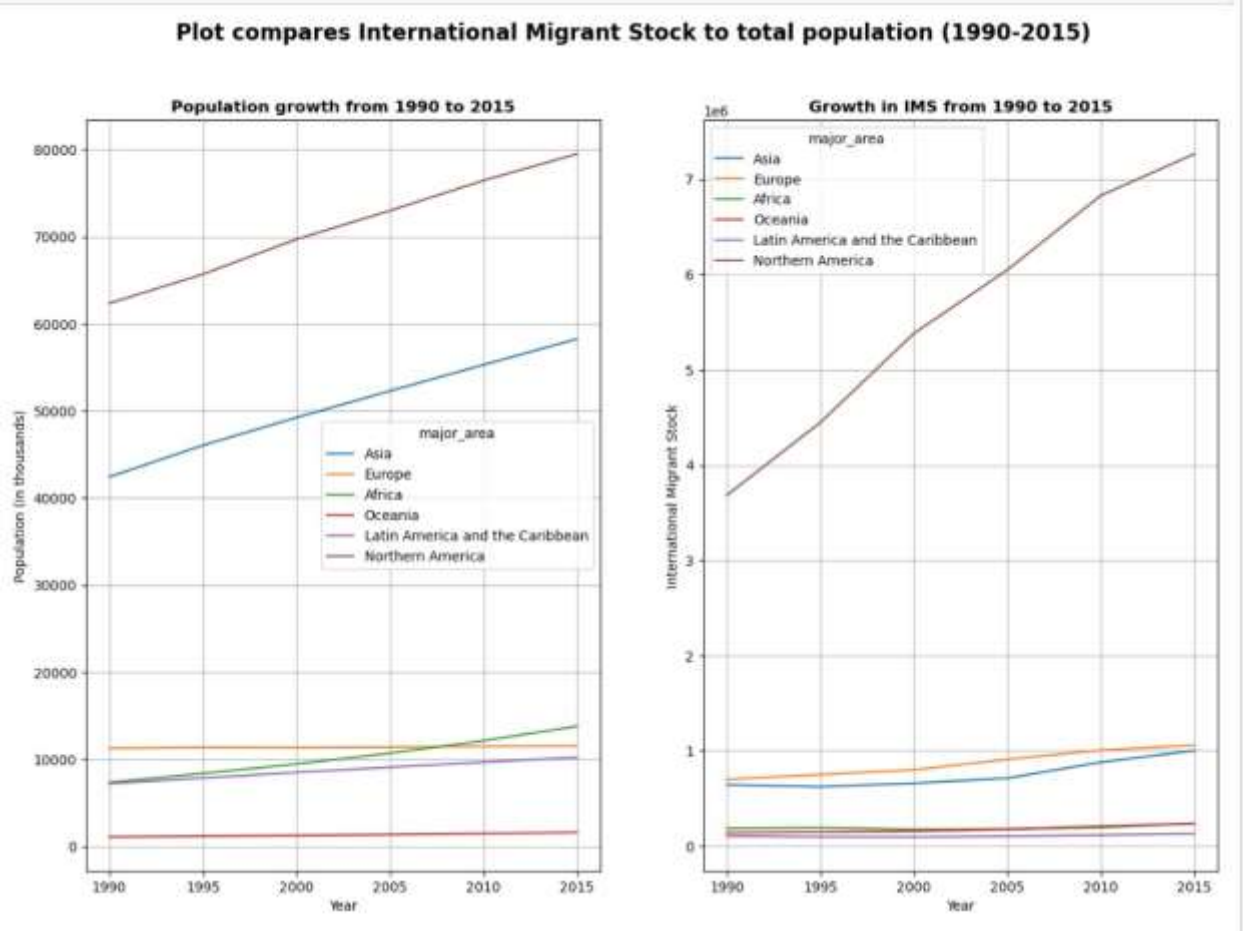


The outliers in this plot were making it hard to read the box plots so I removed them and plotted them again.



This plot shows the central tendency of IMS in Asia, Europe, Africa, and Latin America. We can see that all of them stayed pretty consistent in terms of averages, the change is almost negligible.

Lastly, I created a line plot with population on y-axis and years on x-axis to observe the trends of population growth in various areas of the world. Alongside, I created another plot for IMS to observe it's growth pattern as compared to population growth.



In the above plot, we can see that Northern America has the highest population shortly followed by Asia having second largest population. The steepness of the slope is the same for both of them. Oceania, Africa, and Latin America haven't grown as much since 1990. They have pretty consistent plots. Northern America has shown the maximum increase in International Migrant Stock.

Discussion:

After following these five principles, I have realized that these are the basics of any visualization. In order to create a meaningful visualization, we must go through these five principles every time. Sorting gives our data some order, grouping it together make it easier for the reader to make sense, creating subset makes it easier to work with and comparing using small multiples make it easier to interpret.

Conclusion:

By visualizing UN dataset in different ways, I have gathered that United States of America has the highest number for IMS whereas RMS is highest in Asia. There is very minute difference between male and female populations. Rate of change has been highest observed in Northern America from 1990 to 2015, it almost doubled in number over the span of 25 years.