Student Name: Zaid Abdulaziz

Student Number: 1008397888

Professor: Shion Guha

Final Project: Visualizations

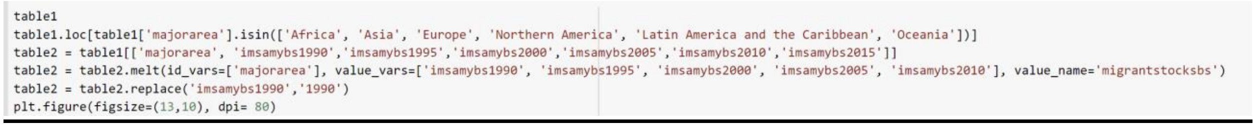
Introduction

The purpose of this final project is a continuation of the midterm regarding data cleaning and analysis, is to now create visualizations with the data frames/tables that were created to effectively communicate and display desired information. The method I used to create the tables was to merge headers, skip the needed rows and rename the columns appropriately as to save space. I then used either country code or major areas as the primary keys to then transpose/melt the migrant stock data for table 1. One chart will be made for each table using both sexes columns of migrant stocks as they are just the combination of the male and female columns. The primary key and column we will be suing to show how migrant stocks are distributed is major areas, specifically per continent as there are too many countries to list them on a single chart so this is the most optimal solution. All charts will be discussed in depth with information such as where the information was pulled from and what valuable information the chart is showing. All of this will be in the methods and results section making the bulk of the writeup with a discussion section right after to go over what was done and the introspects observed in the methods and results that was implemented. Finally, a conclusion part of the writeup will summarize everything that was done, and lessons knowledge gained from this project.

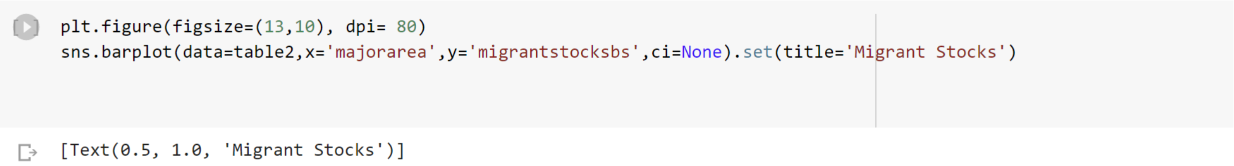
Tukey principles

All of the following charts will have the sort, group, select, and compare principles while also conducting exploratory data analysis and creating the appropriate charts.

Methods and Results



```
table1
table1.loc[table1['majorarea'].isin(['Africa', 'Asia', 'Europe', 'Northern America', 'Latin America and the Caribbean', 'Oceania'])]
table2 = table1[['majorarea', 'imsamybs1990','imsamybs1995','imsamybs2000','imsamybs2005','imsamybs2010','imsamybs2015']]
table2 = table2.melt(id_vars=['majorarea'], value_vars=['imsamybs1990', 'imsamybs1995', 'imsamybs2000', 'imsamybs2005', 'imsamybs2010'], value_name='migrantstocksbs')
table2 = table2.replace('imsamybs1990','1990')
plt.figure(figsize=(13,10), dpi= 80)
```



```
plt.figure(figsize=(13,10), dpi= 80)
sns.barplot(data=table2,x='majorarea',y='migrantstocksbs',ci=None).set(title='Migrant Stocks')
```

```
[Text(0.5, 1.0, 'Migrant Stocks')]
```

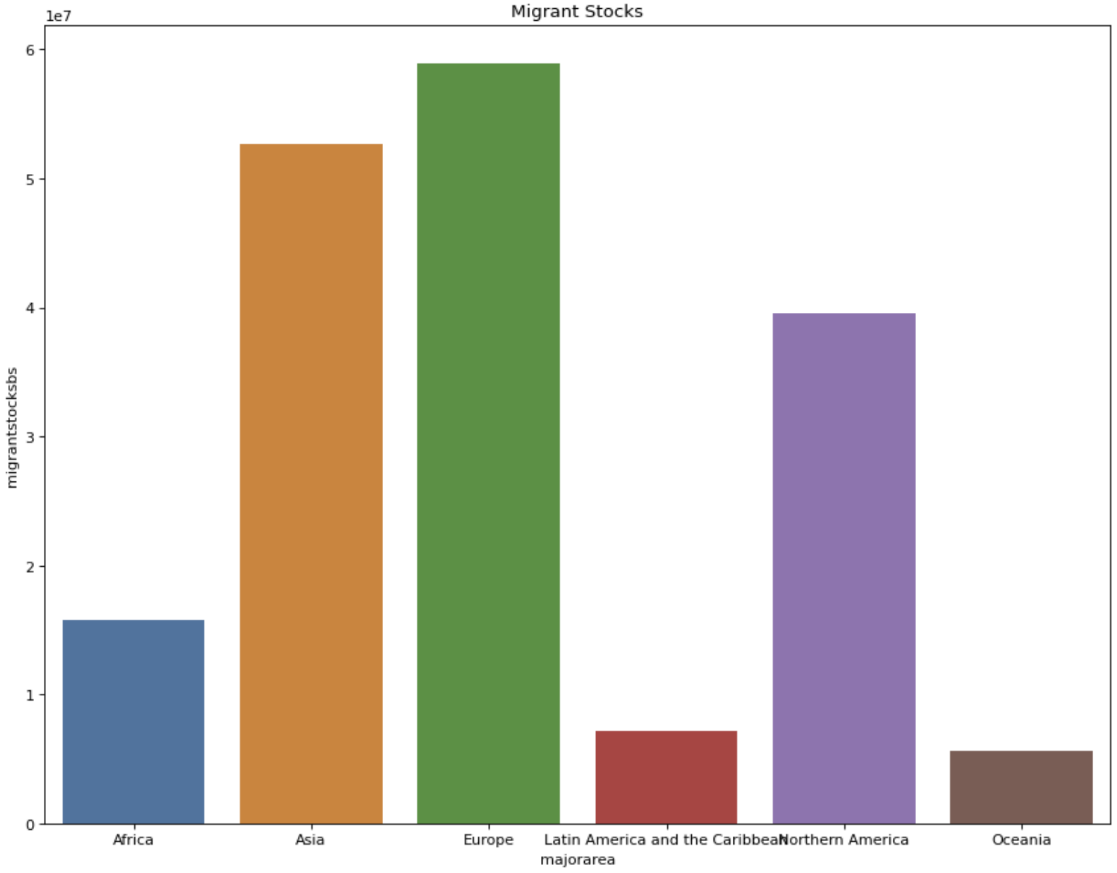
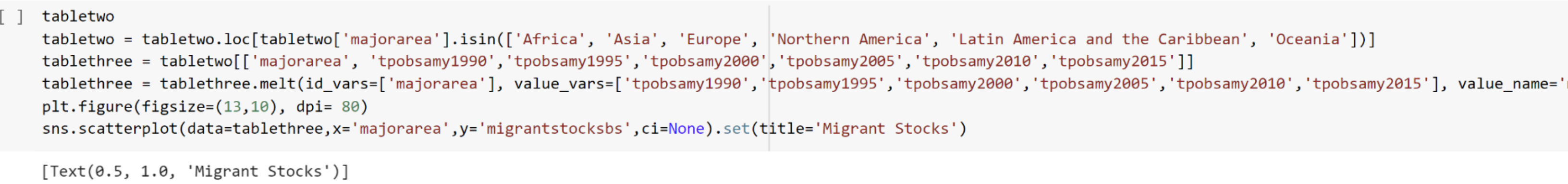
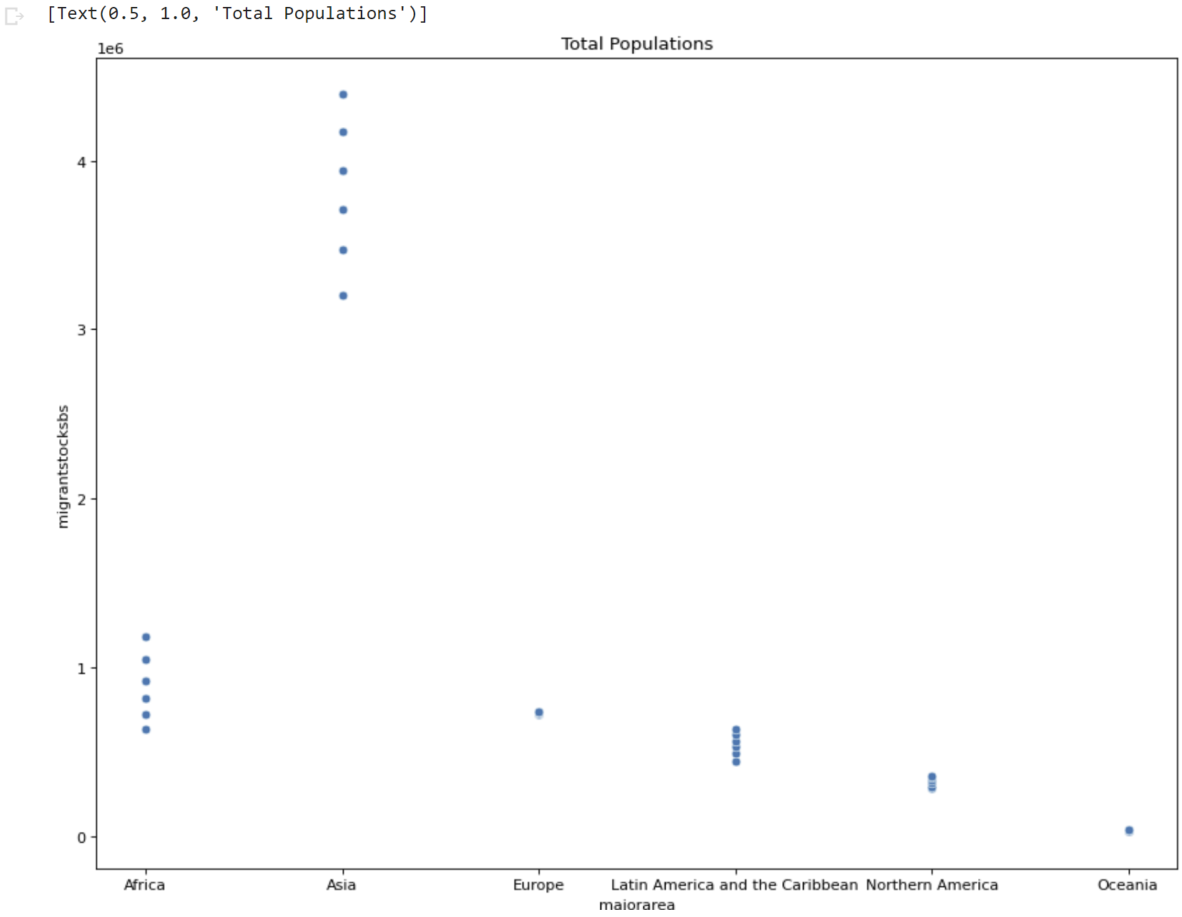


Figure 1

Figure 1 shows us the migrant stocks for both sexes from from years 1990 to 2015 at mid year with a coloured bar chart showing migrant stocks for both sexes from each continent. This data was pulled from table 1 sheet from the excel file using major area as primary key and using melt function to get total migrant stocks. This answers the question as which parts of the world have the highest migrant stocks.



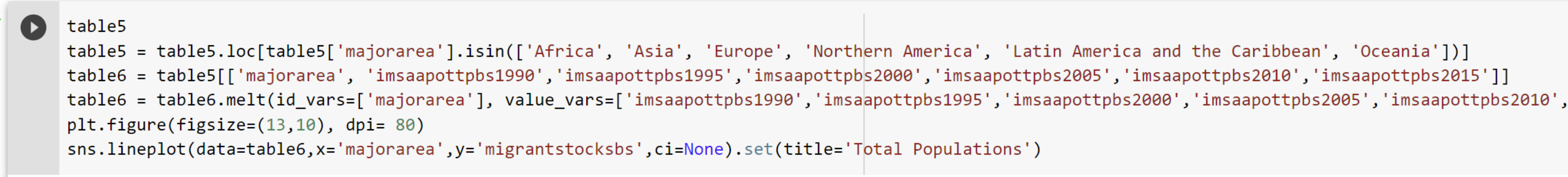
```
[ ]  tabletwo
     tabletwo = tabletwo.loc[tabletwo['majorarea'].isin(['Africa', 'Asia', 'Europe', 'Northern America', 'Latin America and the Caribbean', 'Oceania'])]
     tablethree = tabletwo[['majorarea', 'tpobsamy1990','tpobsamy1995','tpobsamy2000','tpobsamy2005','tpobsamy2010','tpobsamy2015']]
     tablethree = tablethree.melt(id_vars=['majorarea'], value_vars=['tpobsamy1990','tpobsamy1995','tpobsamy2000','tpobsamy2005','tpobsamy2010','tpobsamy2015'], value_name='
     plt.figure(figsize=(13,10), dpi= 80)
     sns.scatterplot(data=tablethree,x='majorarea',y='migrantstocksbs',ci=None).set(title='Migrant Stocks')

     [Text(0.5, 1.0, 'Migrant Stocks')]
```



```
[Text(0.5, 1.0, 'Total Populations')]
```

Figure 2

Figure 2 shows us the total populations for both sexes from from years 1990 to 2015 at mid year with a scatterplot chart showing total populations for both sexes from each continent. This data was pulled from table 2 sheet from the excel file using major area as primary key and using melt function to get total population. This answers the question of how closely certain population numbers are to each other based on the distance of the dots per continent.



```
table5
table5 = table5.loc[table5['majorarea'].isin(['Africa', 'Asia', 'Europe', 'Northern America', 'Latin America and the Caribbean', 'Oceania'])]
table6 = table5[['majorarea', 'imsaapottpbs1990','imsaapottpbs1995','imsaapottpbs2000','imsaapottpbs2005','imsaapottpbs2010','imsaapottpbs2015']]
table6 = table6.melt(id_vars=['majorarea'], value_vars=['imsaapottpbs1990','imsaapottpbs1995','imsaapottpbs2000','imsaapottpbs2005','imsaapottpbs2010',
plt.figure(figsize=(13,10), dpi= 80)
sns.lineplot(data=table6,x='majorarea',y='migrantstocksbs',ci=None).set(title='Total Populations')
```

```
[Text(0.5, 1.0, 'Total Populations as %')]
```

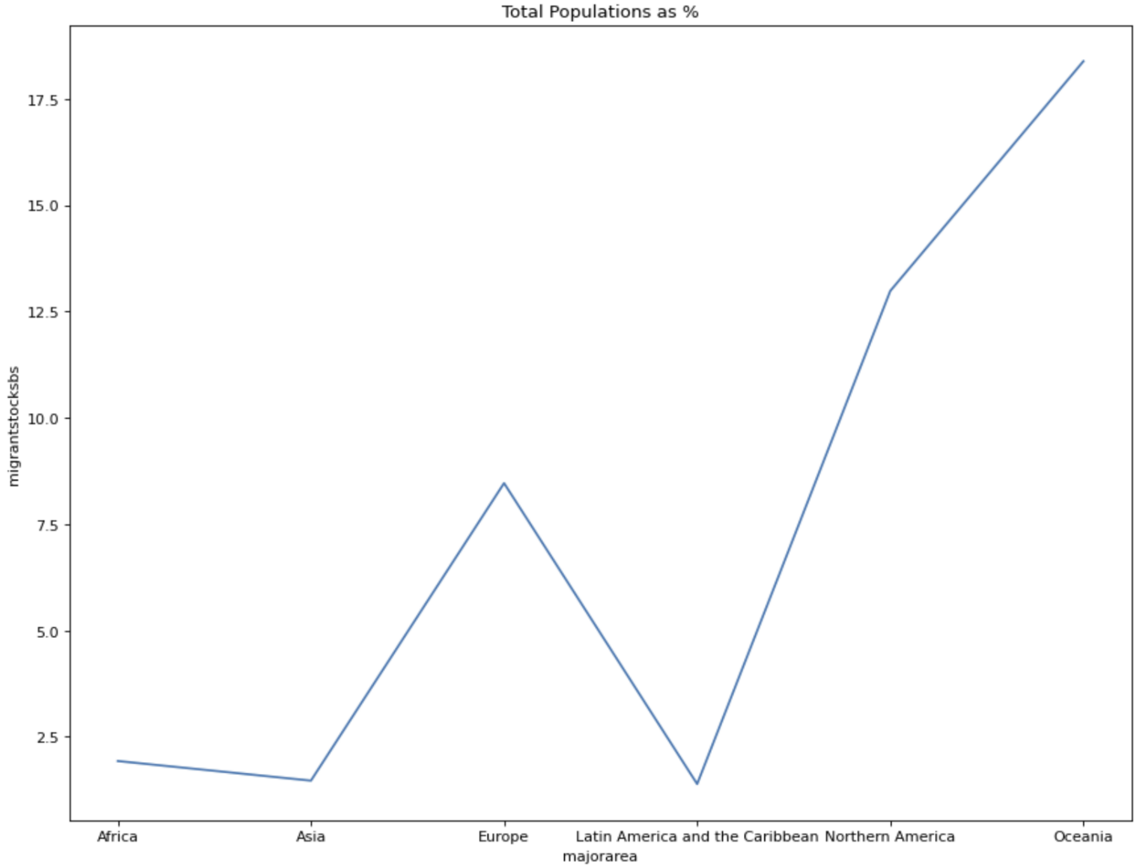
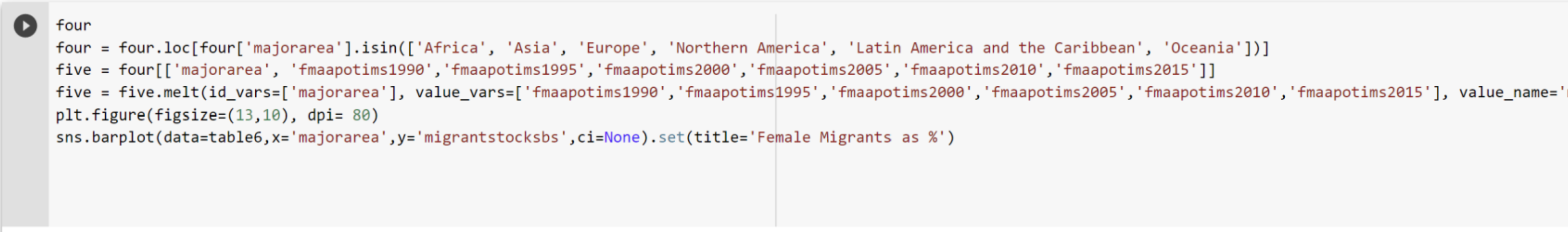


Figure 3

Figure 3 shows us the total populations for both sexes as a percentage from from years 1990 to 2015 with a line plot total populations for both sexes as a percentage from each continent. This data was pulled from table 3 sheet from the excel file using major area as primary key and using melt function to get total population. This answers the question of how differently populations numbers and percentages can be shown differently as Oceania having a smaller population overall can make the percentage number seem like they have the highest number of migrants when they do not.



```
four
four = four.loc[four['majorarea'].isin(['Africa', 'Asia', 'Europe', 'Northern America', 'Latin America and the Caribbean', 'Oceania'])]
five = four[['majorarea', 'fmaapotims1990','fmaapotims1995','fmaapotims2000','fmaapotims2005','fmaapotims2010','fmaapotims2015']]
five = five.melt(id_vars=['majorarea'], value_vars=['fmaapotims1990','fmaapotims1995','fmaapotims2000','fmaapotims2005','fmaapotims2010','fmaapotims2015'], value_name='
plt.figure(figsize=(13,10), dpi= 80)
sns.barplot(data=table6,x='majorarea',y='migrantstocksbs',ci=None).set(title='Female Migrants as %')
```

```
[Text(0.5, 1.0, 'Female Migrants as %')]
```

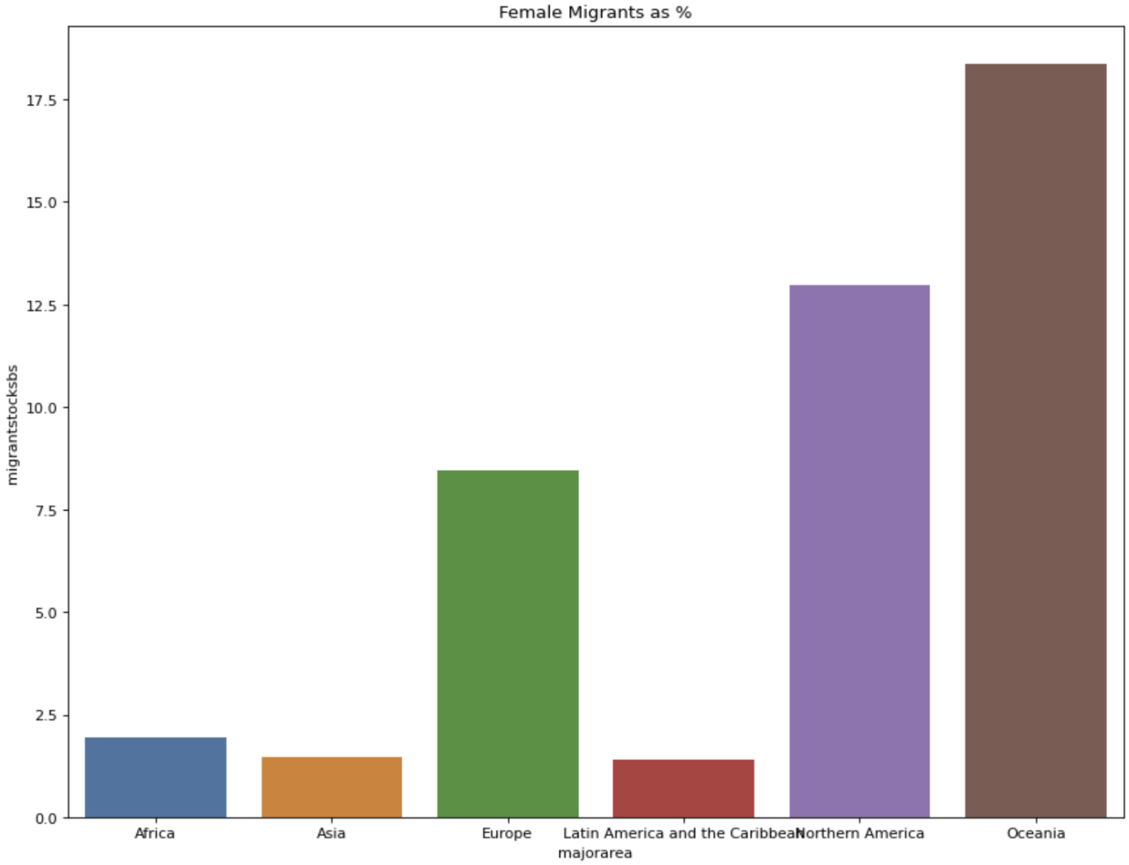
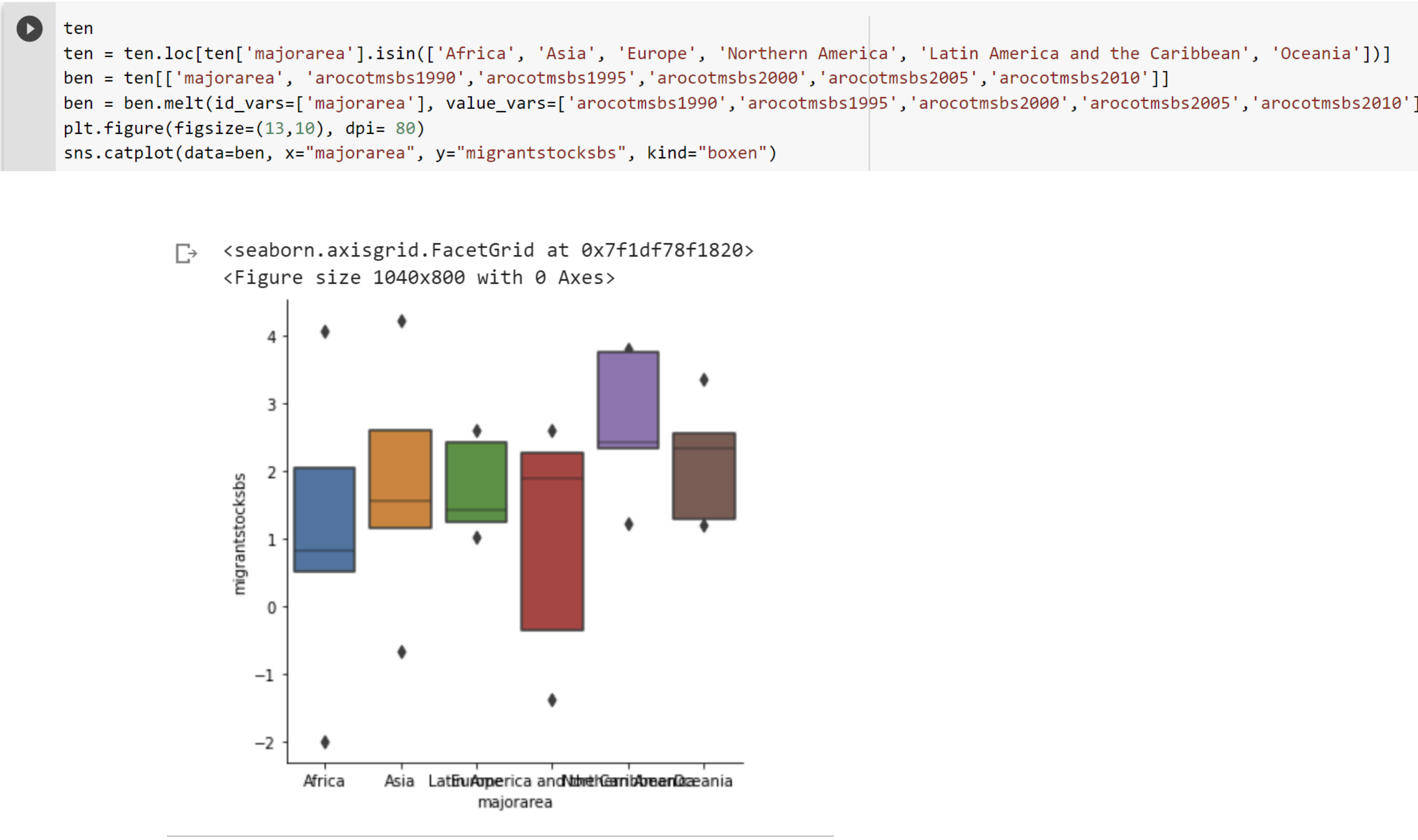


Figure 4

Figure 4 shows us the female migrants as a percentage from from years 1990 to 2015 with a bar plot showing the female migrants as a percentage from each continent. This data was pulled from table 4 sheet from the excel file using major area as primary key and using melt function to get total population. This chart like the previous one shows us that large percentages of migrants are under the Oceania category regardless of sex.



```
ten
ten = ten.loc[ten['majorarea'].isin(['Africa', 'Asia', 'Europe', 'Northern America', 'Latin America and the Caribbean', 'Oceania'])]
ben = ten[['majorarea', 'arocotmsbs1990','arocotmsbs1995','arocotmsbs2000','arocotmsbs2005','arocotmsbs2010']]
ben = ben.melt(id_vars=['majorarea'], value_vars=['arocotmsbs1990','arocotmsbs1995','arocotmsbs2000','arocotmsbs2005','arocotmsbs2010']
plt.figure(figsize=(13,10), dpi= 80)
sns.catplot(data=ben, x="majorarea", y="migrantstocksbs", kind="boxen")
```

```
<seaborn.axisgrid.FacetGrid at 0x7f1df78f1820>
<Figure size 1040x800 with 0 Axes>
```

Figure 5

Figure 5 shows us the annual rate of change of migrant stocks for both sexes from years 1990 to 2015 with a box plot showing the annual rate of change of migrant stocks for both sexes from each continent. This data was pulled from table 5 sheet from the excel file using major area as primary key and using melt function to get total population. This chart helps answer questions like min and max values, interquartile ranges, and medians for each continent as well as display outliers to see which continents have the most amount of movement.

Discussion

The way I cleaned the data set in the midterm I have discovered is not the most optimal as all major areas would be displayed with all countries, regions and continents with the melt function being used to get the first 5 columns of populations. I had discovered that the charts would not have enough room to display all that information, so I decided to take a different approach. I would only use the continents as my x axis and population numbers as my y axis so that I could summarize as much information as possible while still having accurate visualizations for my data. The different charts used all answered different questions and displayed information in a way that help better understand averages, outliers, distributions percentages

Conclusion

In conclusion I have done data cleaning and visualizations from a rather complex data set with multiple sheets that were not organized using the the tidy data principles in the midterm then the visualization techniques we have learned for this assignment. I was able to adjust accordingly and make the data cleaning for this project a bit different than the midterm to make the charts legible for viewing. I believe the skills learned in this class will serve as solid foundation to expand my data science skills and assist me in my growth within the tech and business industry.