Midterm Project Write up

Haoyu Zhang

Introduction

The data used in this project shows the trends in international migrant stock between 1990 and 2015 by country or area and gender. The data includes one content, 6 tables, one annex and one notes. Table 1 shows the international migrant stock at mid-year between 1990 and 2015. Tables 2 contains the total population at mid-year from 1990 to 2015. Table 3 includes the international migrant stock as a percentage of the total population 1990 - 2015. Table 4 shows the percentage of the female migrants in the international migrant stock from 1990 to 2015. Table 5 has the annual rate of change of the migrant stock between 1990 and 2015. Table 6 includes the estimated refugee stock at mid-year 1990 - 2015. In the following paragraphs, I will clearly describe how I clean this dataset using tidy data principles.

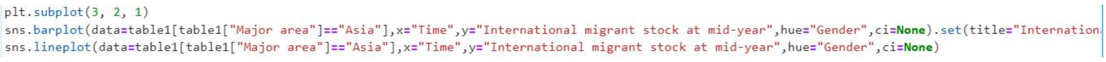
Method

Mainly, I use seaborn and motplotlib for visualization. Thus, I import them at the beginning. I also set the style for seaborn, which can show the scale for the graphs.

Table 1 (International migrant stock)

For the table1 (International migrant stock), I used barplot and lineplot to make the tables. I follow the Tufte's principle about data ink, which make the tables only includes necessary elements for visualization. I make the one table for each major areas (Asia, Africa, Northern America, Europe, Oceania and Latin America and the Caribbean), which shows the number of international migrant stock for that area between 1990 and 2015 by sex. These tables are easy to understand and can show the trend of the amount and the difference between male and female.

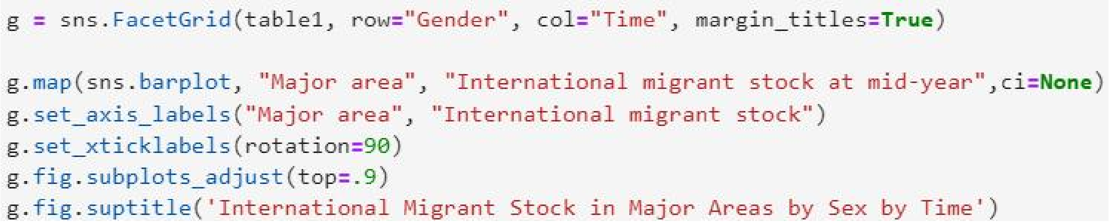
Code example:



```
plt.subplot(3, 2, 1)
sns.barplot(data=table1[table1["Major area"]=="Asia"],x="Time",y="International migrant stock at mid-year",hue="Gender",ci=None).set(title="Internation
sns.lineplot(data=table1[table1["Major area"]=="Asia"],x="Time",y="International migrant stock at mid-year",hue="Gender",ci=None)
```

Then, I create multiple small graphs according to Tufte's principle about small multiples. These graphs are under same scales and easy to compare the difference between different time and different gender. Because of the Tufte's principle about data density, I label the amount for each bars. Each graph show the amount of the specific gender's international migrant stock in major areas in specific year.

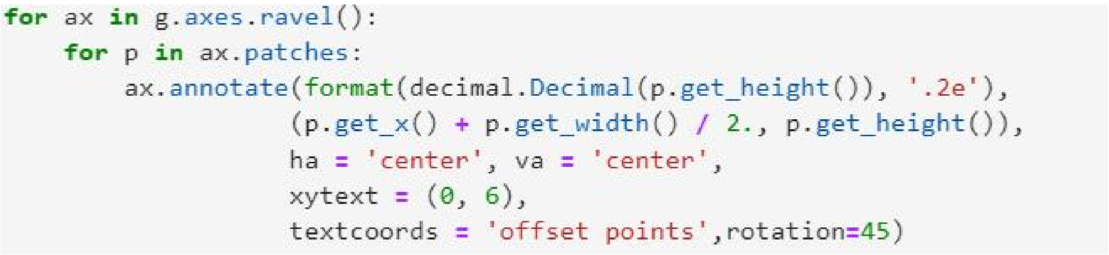
Codes for small multiples:



```
g = sns.FacetGrid(table1, row="Gender", col="Time", margin_titles=True)

g.map(sns.barplot, "Major area", "International migrant stock at mid-year",ci=None)
g.set_axis_labels("Major area", "International migrant stock")
g.set_xticklabels(rotation=90)
g.fig.subplots_adjust(top=.9)
g.fig.suptitle('International Migrant Stock in Major Areas by Sex by Time')
```

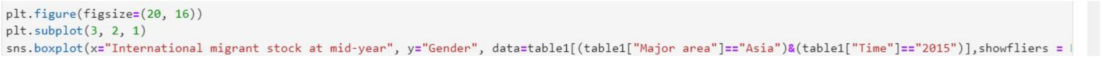
Codes for labeling:



```
for ax in g.axes.ravel():
    for p in ax.patches:
        ax.annotate(format(decimal.Decimal(p.get_height()), '.2e'),
                    (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha = 'center', va = 'center',
                    xytext = (0, 6),
                    textcoords = 'offset points',rotation=45)
```

After this, I create the box chart for each areas about their international migrant stock in 2015. I used sns.boxplot and plt.subplot to create the graphs. The graphs are pretty straight and clean, which follows the principle of data-ink and the size is properly, which follows the principle of data density. The graph shows 2015 international migrant stock in one major areas for each gender.

Codes example:

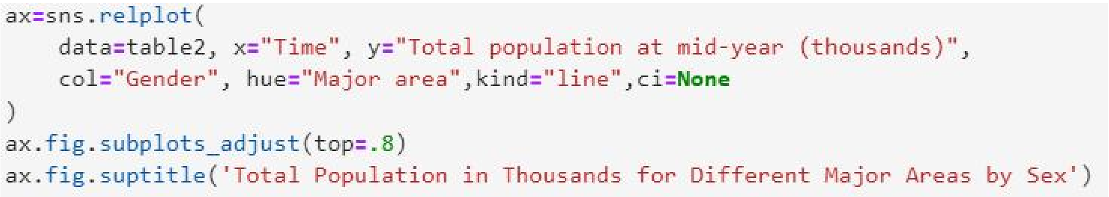


```
plt.figure(figsize=(20, 16))
plt.subplot(3, 2, 1)
sns.boxplot(x="International migrant stock at mid-year", y="Gender", data=table1[(table1["Major area"]=="Asia")&(table1["Time"]=="2015")],showfliers =
```

Table 2 (Total population)

I follow the principle of small multiples to create the line graph to show the trend of total population in major areas between 1990 and 2015. I used sns.relpot to build the graphs. The clear graphs with suitable size is also following the Tufte's principles about data-ink and data density.

Codes:



```
ax=sns.relplot(
    data=table2, x="Time", y="Total population at mid-year (thousands)",
    col="Gender", hue="Major area",kind="line",ci=None
)
ax.fig.subplots_adjust(top=.8)
ax.fig.suptitle('Total Population in Thousands for Different Major Areas by Sex')
```

Table 3 (international migrant stock as percentage of the total population

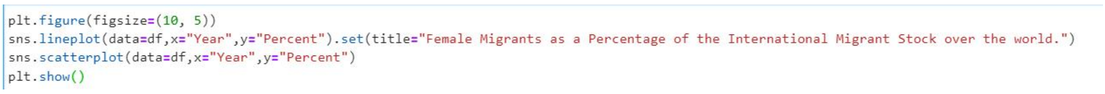
)

Similar to above, I use same ways to create the small multiples for international migrant stock as percentage of the total population, which shows the international migrant stock as percentage of the total population in major areas between 1990 and 2015 for each gender. (Three in total, one for male, one for female and one for both) Then I use similar methods to create bar chart and line graph in one table and create one table for each major area, which can show the amount of international migrant stock as percentage of the total population in that major area and how this number change from 1990 to 2015 by gender.

Table4(Percentage of female in international migrant stock)

Because the number is pretty closed. I used a line graph to show how the percentage of female change in worldwide from 1990 to 2015. I also use scatterplot to point out the amount for each year.

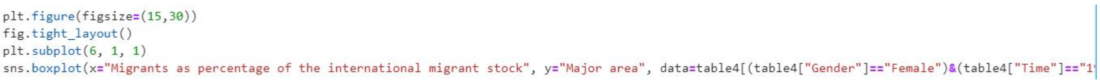
Codes:



```
plt.figure(figsize=(10, 5))
sns.lineplot(data=df,x="Year",y="Percent").set(title="Female Migrants as a Percentage of the International Migrant Stock over the world.")
sns.scatterplot(data=df,x="Year",y="Percent")
plt.show()
```

Then I create box plot to detailed show the percentage of female in international migrant stock for major areas in specific year (1990, 1995, 2000, 2005, 2010, 2015)

Codes:



```
plt.figure(figsize=(15,30))
fig.tight_layout()
plt.subplot(6, 1, 1)
sns.boxplot(x="Migrants as percentage of the international migrant stock", y="Major area", data=table4[(table4["Gender"]=="Female")&(table4["Time"]=="1
```

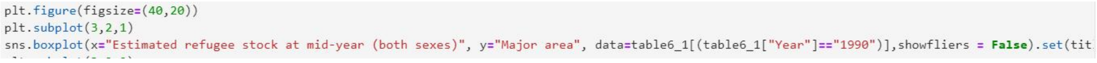
Table 5(Annual rate of change of the migrant stock)

Same method used in table 1 and 3, using sns.lineplot and sns.barplot to create one table for each major ares. Each tables shows the annual rate of change of the migrant stock in that major area 1990 - 2015.

Table6 (Estimated refugee stock, Annual rate of change of refugee stock and Refugee as percentage of the international migrant stock)

I use box plot to show the amount of estimated refugee stock of major areas in each year.

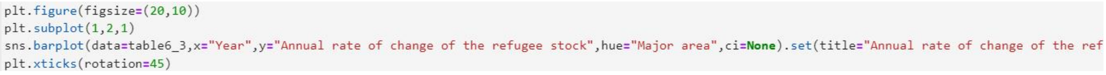
Codes example:



```
plt.figure(figsize=(40,20))
plt.subplot(3,2,1)
sns.boxplot(x="Estimated refugee stock at mid-year (both sexes)", y="Major area", data=table6_1[(table6_1["Year"]=="1990")],showfliers = False).set(tit
```

Then, I use bar plot to create a bar chart to show the annual rate of change of refugee stock in major areas from 1990 to 2015.

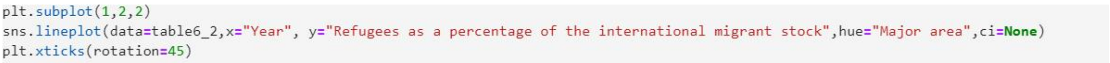
Codes:



```
plt.figure(figsize=(20,10))
plt.subplot(1,2,1)
sns.barplot(data=table6_3,x="Year",y="Annual rate of change of the refugee stock",hue="Major area",ci=None).set(title="Annual rate of change of the ref
plt.xticks(rotation=45)
```

Final, create a line table, where each line shows the changes of refugee as percentage of the international migrant stock for one major area between 1990 and 2015.

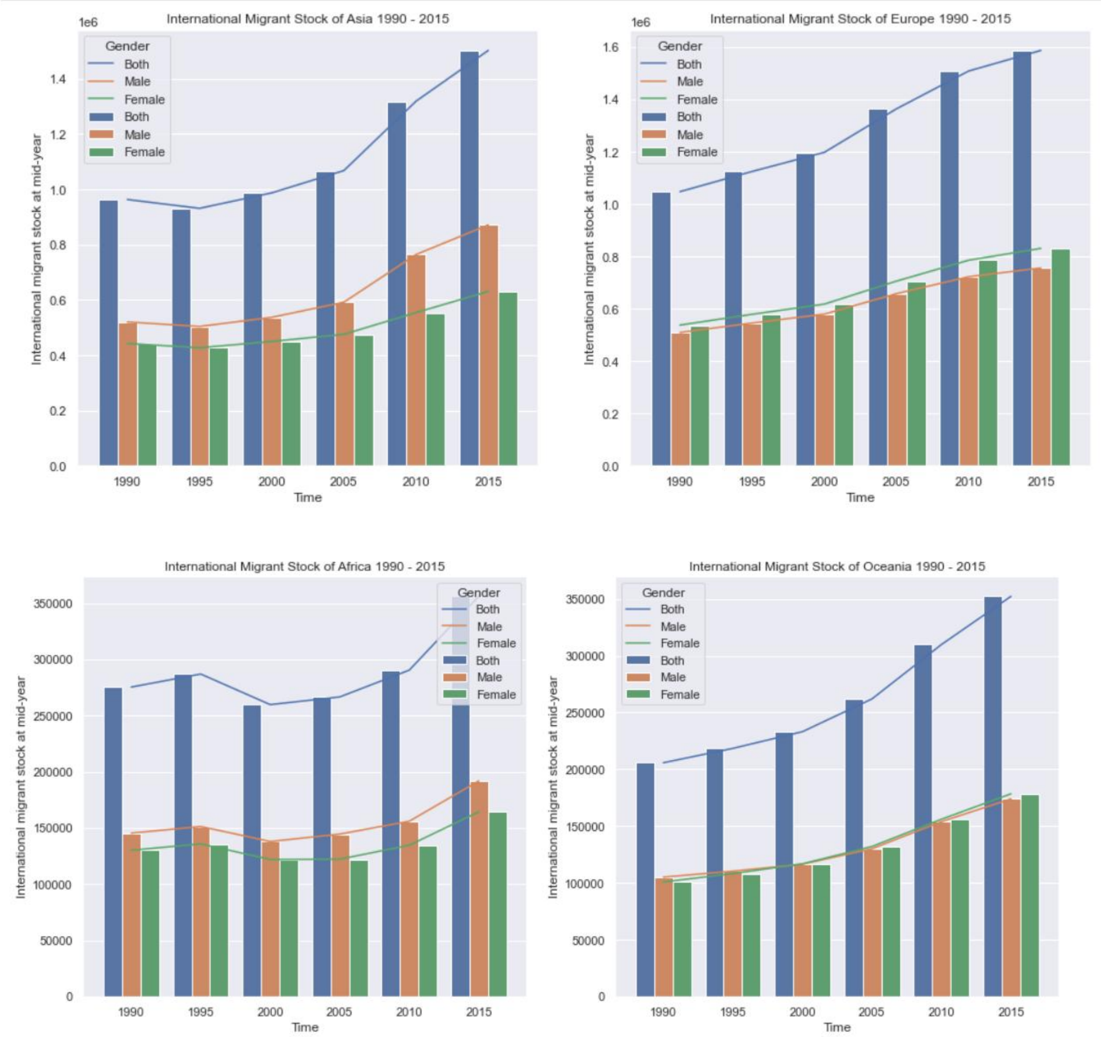
Codes:

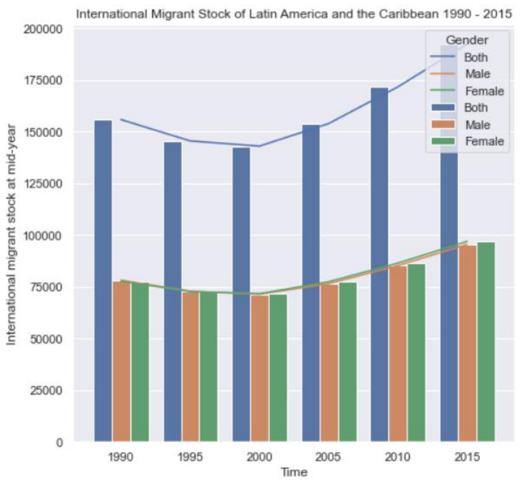


```
plt.subplot(1,2,2)
sns.lineplot(data=table6_2,x="Year", y="Refugees as a percentage of the international migrant stock",hue="Major area",ci=None)
plt.xticks(rotation=45)
```

Result

Table1 (line + bar)





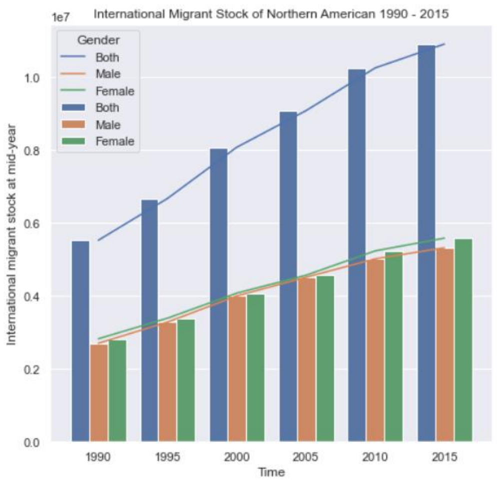


Table1(small multiples: bar)

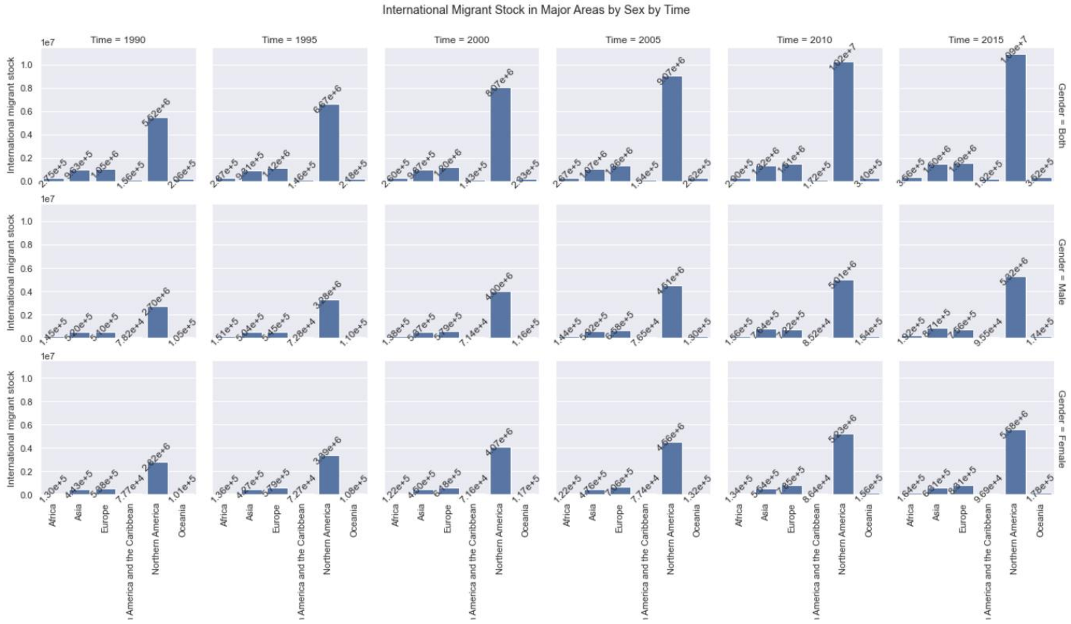
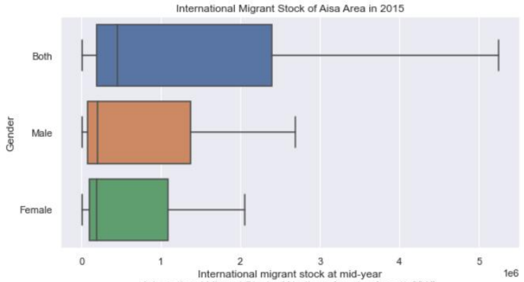
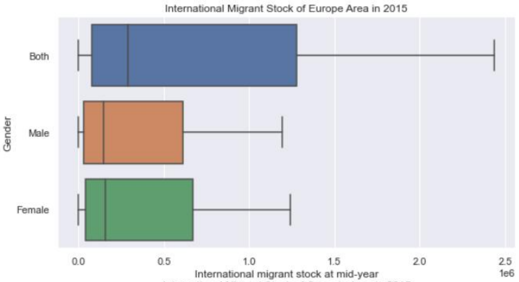
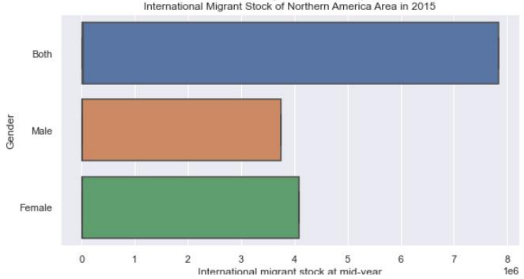
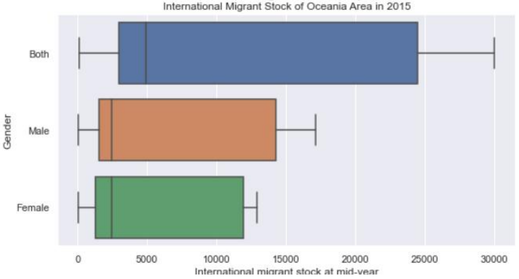


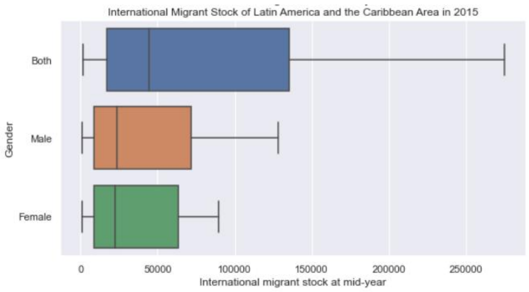
Table1(Box)











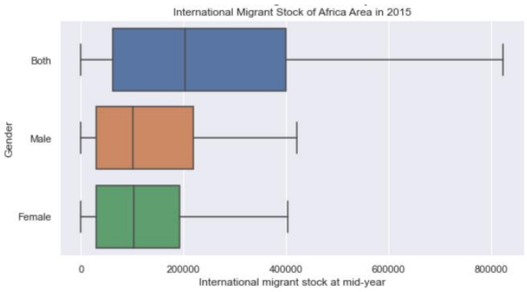


Table2(small multiples: line)

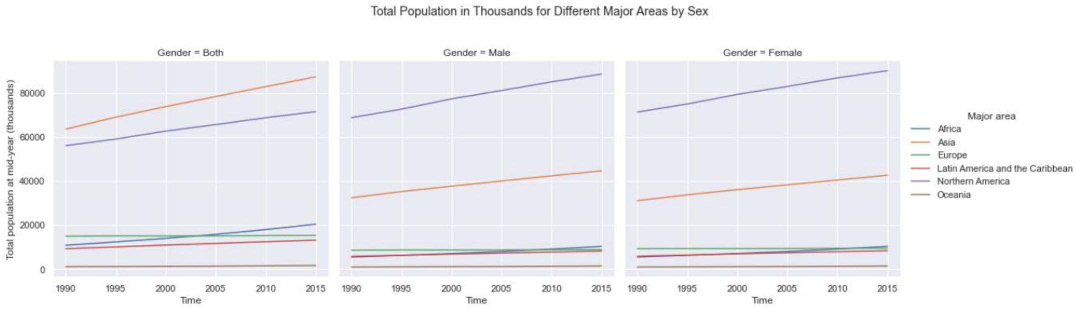


Table3(small multiples: line)

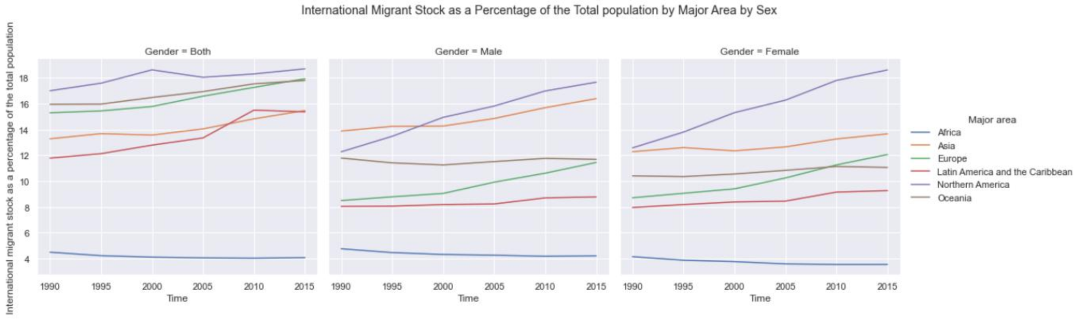
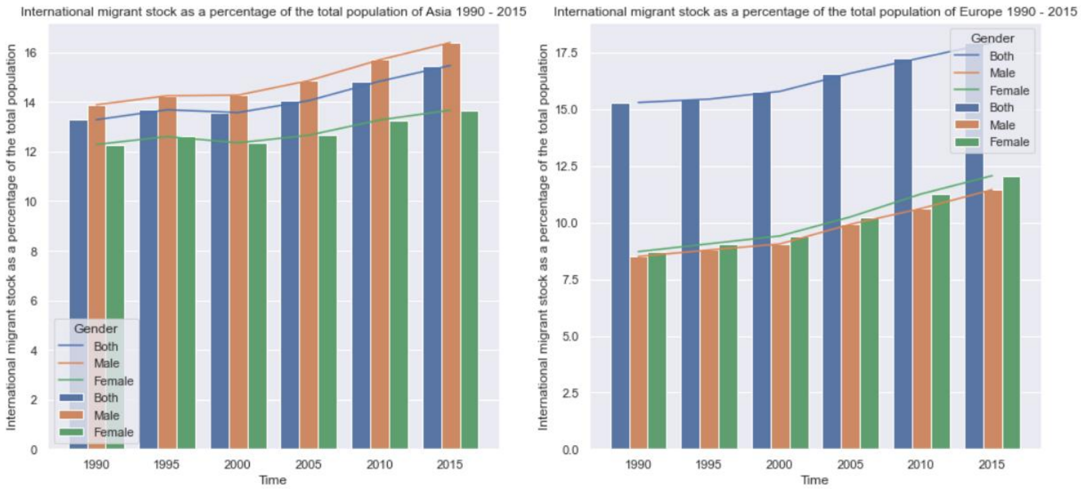


Table3(bar + line)



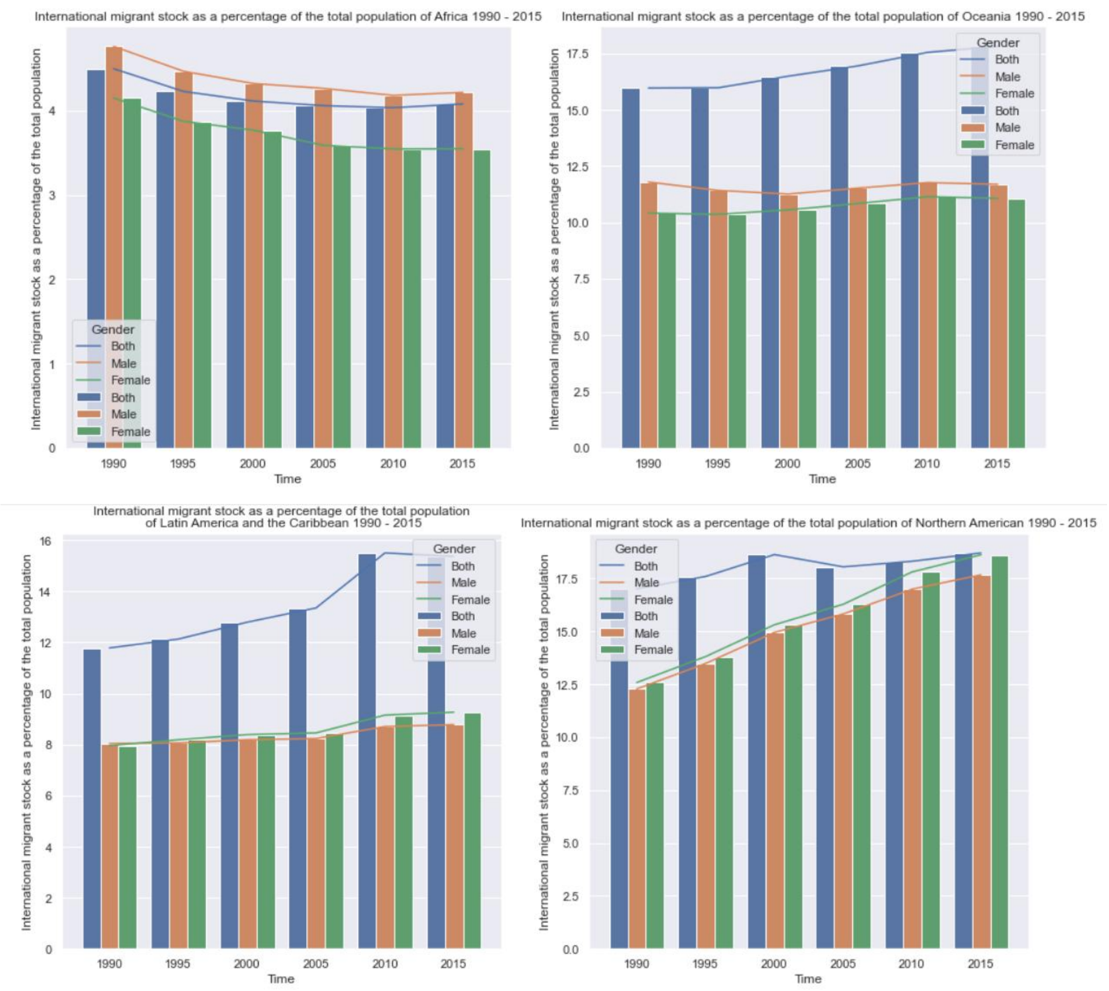
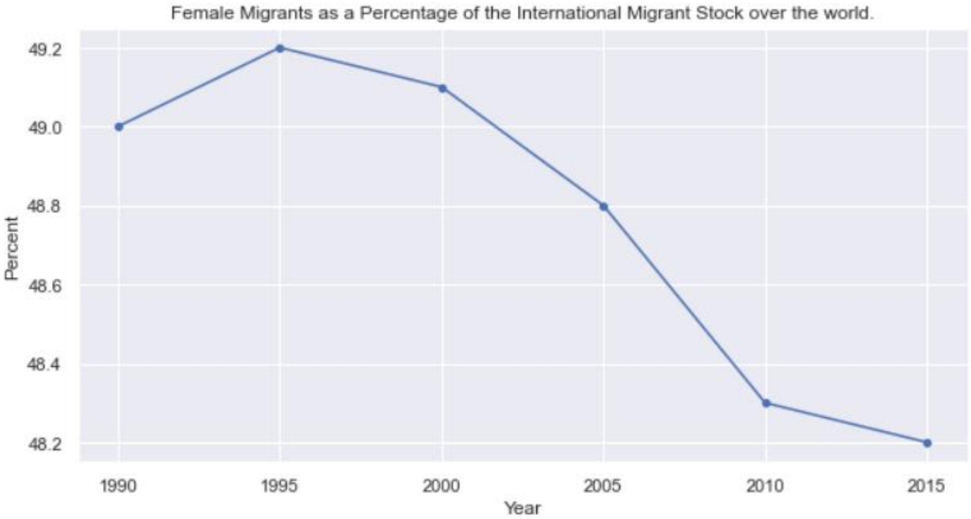
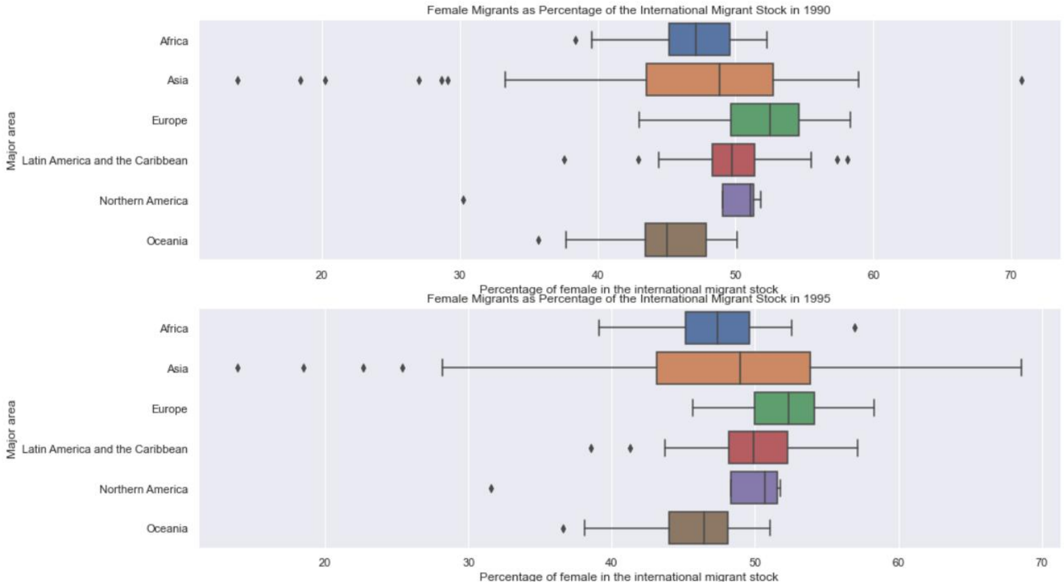
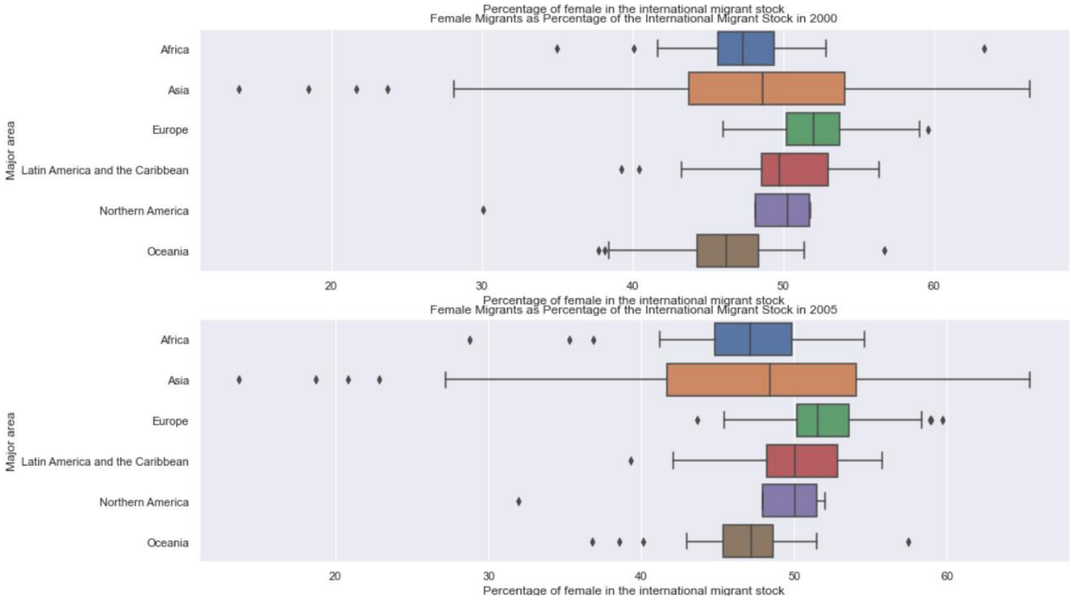


Table4(line)



Tbale4(box)


Female Migrants as Percentage of the International Migrant Stock in 1990
Major area
Africa
Asia
Europe
Latin America and the Caribbean
Northern America
Oceania
20
30
40
50
60
70
Percentage of female in the international migrant stock
Female Migrants as Percentage of the International Migrant Stock in 1995
Major area
Africa
Asia
Europe
Latin America and the Caribbean
Northern America
Oceania
20
30
40
50
60
70
Percentage of female in the international migrant stock


Percentage of female in the international migrant stock
Female Migrants as Percentage of the International Migrant Stock in 2000
Major area
Africa
Asia
Europe
Latin America and the Caribbean
Northern America
Oceania
20
30
40
50
60
Percentage of female in the international migrant stock
Female Migrants as Percentage of the International Migrant Stock in 2005
Major area
Africa
Asia
Europe
Latin America and the Caribbean
Northern America
Oceania
20
30
40
50
60
Percentage of female in the international migrant stock

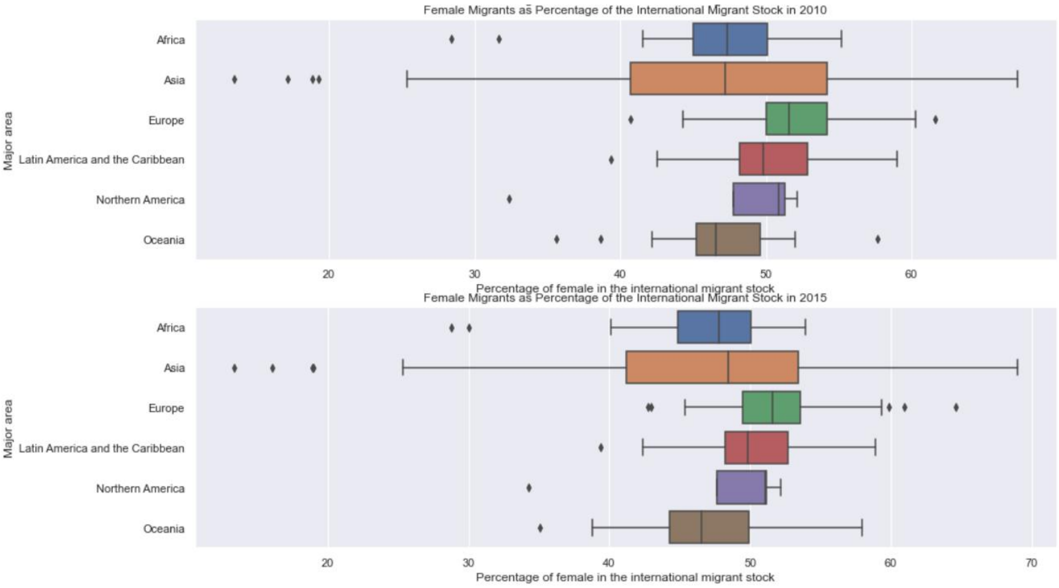

Female Migrants as Percentage of the International Migrant Stock in 2010
Major area
Africa
Asia
Europe
Latin America and the Caribbean
Northern America
Oceania
20
30
40
50
60
Percentage of female in the international migrant stock
Female Migrants as Percentage of the International Migrant Stock in 2015
Major area
Africa
Asia
Europe
Latin America and the Caribbean
Northern America
Oceania
20
30
40
50
60
70
Percentage of female in the international migrant stock

Table5(Bar + line)

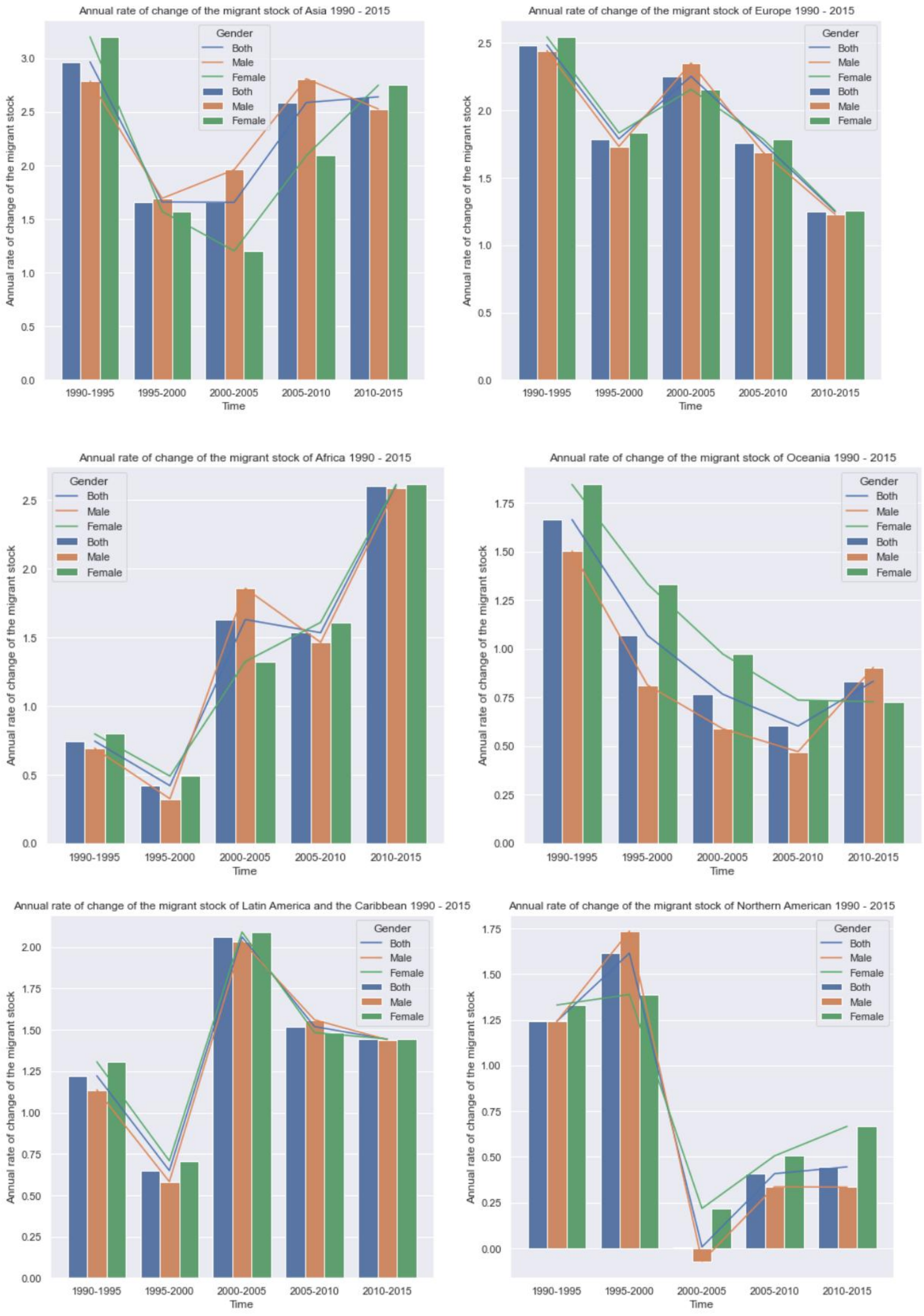


Table6(box)

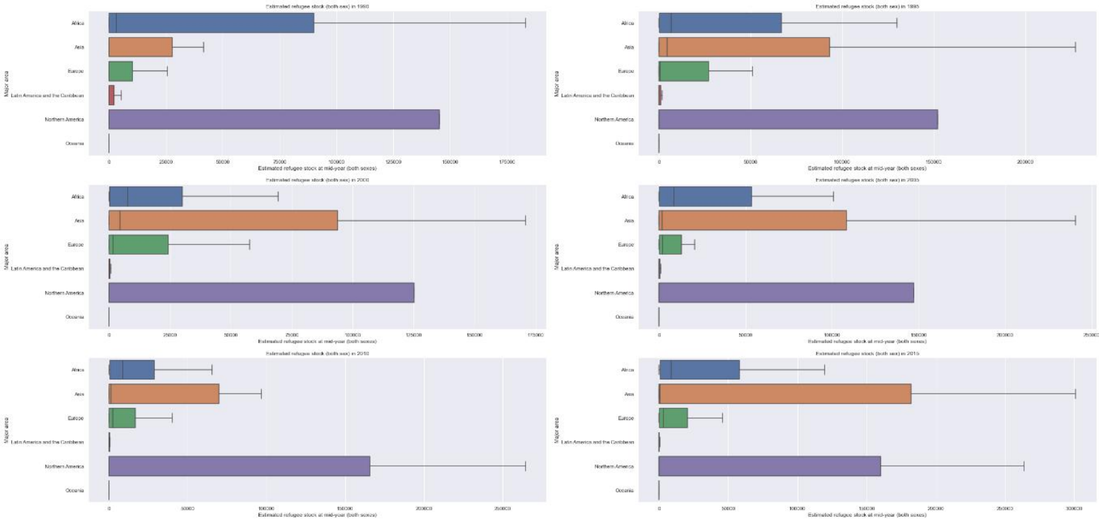
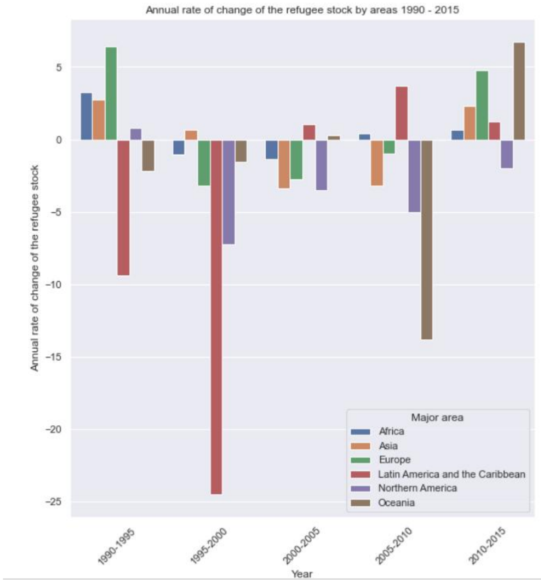
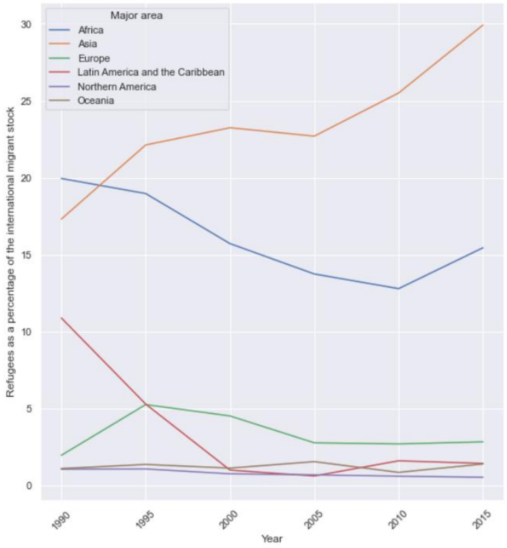
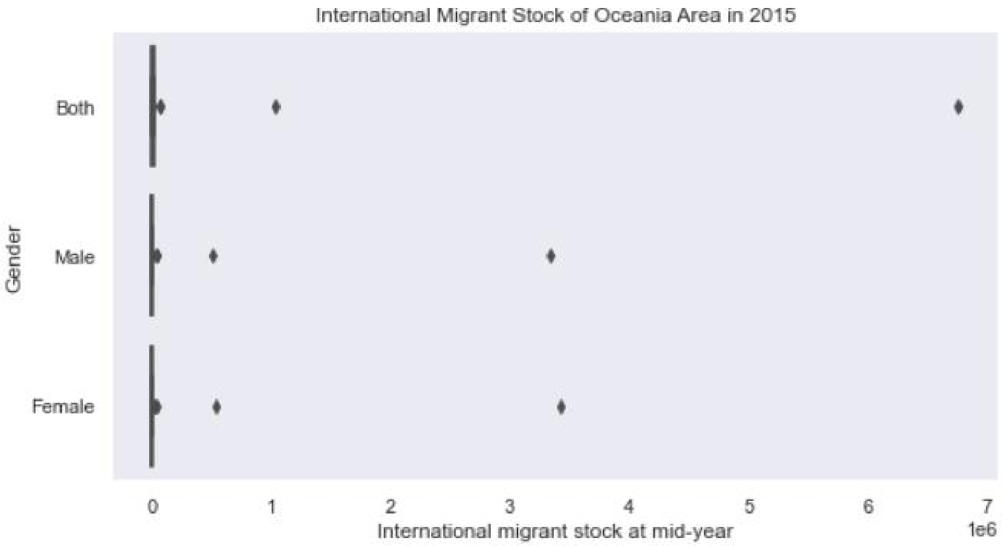


Table6(bar) and table6(line)





Discussion:



When drawing the bar, line and box graphs or charts, there are many outliers that will make the graph unreadable. Thus, I hide the outliers in the graphs in most of case, which follows the Tufte's principles of graphical integrity.

Through all the graphs and tables I created in this project, I follow the Tufte's principles about graphical integrity, data-ink, chartjunk and data density. I only label the necessary elements to show the information in the graphs to make it clear, easy to read and without any unnecessary graphical effects like shadow or set beautiful color. All the graphs are straight to the core information it contains. I also create small multiples for some of tables, which is also one of the Tufte's principle and easy to compare. Because these graphs are under same scale and unit. To follow the principle about data density is pretty hard. Because I have to find a proper size for figures to make it clear enough and as small as possible.