INF 1340

Jessica Wang

1005230439

December 12th, 2022

Final Project of INF 1340

Introduction

For this final project, I used Exploratory Data Analysis to investigate the data that I have cleaned from my midterm project and summarize the key insights by using different types of graphs. By performing EDA, it helps me to do the data visualization so that I can see the data clearer and have a better understanding of the data. Also, it can help me to make the contrast easier by using the graphs such as scatter, box, bar, density and correlation plots. From the datasets of Union Nations's migrants stock that I have cleaned in my midterm project, I used Tufte's principle to use the visual to represent the data.

Methods

For this project, I used five types of graph in total.

Multi-set Bar chart: this type of graph is also known as grouped bar chart or clustered bar chart. It can be used when two or more data series need to be plotted all on the same axis and grouped into parent categories. It is useful to compare across categories that contain the same sub-categorical variables between them.

Bar chart: the bar charts usually show discrete, numerical comparisons across categories. The bar charts do not display continuous developments over an interval. For example, the answers should be the questions of "how many?", not "how much?"

Scatter plot/Point plot: The scatterplot places points on a Cartesian Coordinates system to display all the values between two variables. By observing the patterns, we can see if there is a relationship or correlation between two variables. There are positive correlation (values increase together), negative correlation (one value decreases when another increases) and null (no correlation). Also, the shapes of the correlations are linear, exponential and U-shaped. By using scatter plot, we can quickly determine the relationship by how closely packed the points are to each other on the graph. The outliers are the points which are far outside the general cluster of points. Furthermore, the lines or curves can be displayed over the graph to aid in the analysis, in order to show how it would look if all the points were condensed together into a single line.

Box and whisker plot: This type of graph is good to visualize the data distribution through the quartiles. By observing the box and whisker plot, we can know

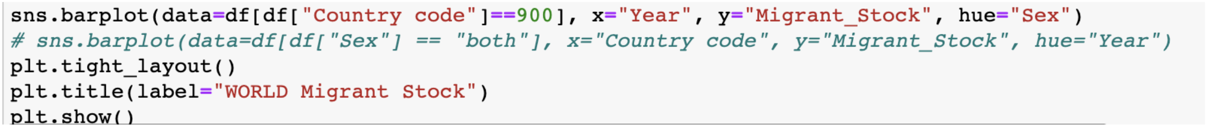
- key values, such as the average, median, 25th percentile and 75th percentile;
- Outliers: if there are any outliers and the values of them
- if the data is symmetrical or not
- How tightly is the data grouped
- if the data is skewed and the direction of skew.

Therefore, the box and whisker plot is a good way to compare distributions between many groups or datasets. It takes up less space compared to other types of graphs.

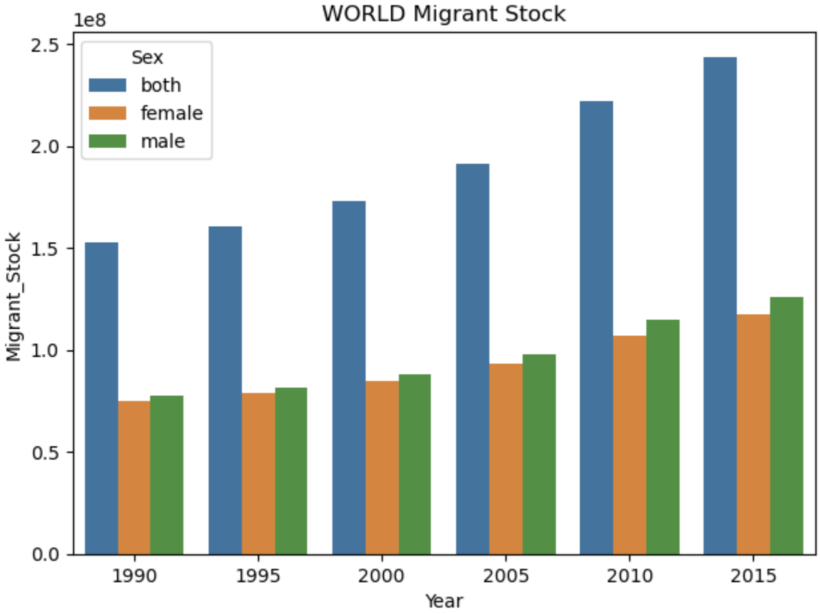
Line graph: Different from bar charts, this type of chart is used to display quantitative values of a continuous interval or time period. It is a good way to show the trends and analyze how the data has changed over time. The direction of the lines on the graph play the role as a metaphor for the data. The upward slope indicates where the values have increased and a downward slope indicates where values have decreased. The line can create patterns that reveal trends in a dataset.

Results

For table 1, this dataset is about the international migrant stock at mid-year for both sexes, male and female. I used the multi-set bar chart to do the world migrant stock graph. For the horizontal axis, it represents the year, and the vertical axis represents the migrant stock. Since there are three groups of data, the multi-set bar chart is better than the bar chart to see the data of different groups with parent categories.



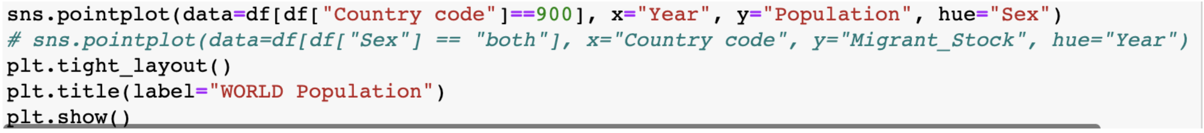
```
sns.barplot(data=df[df["Country code"]==900], x="Year", y="Migrant_Stock", hue="Sex")
# sns.barplot(data=df[df["Sex"] == "both"], x="Country code", y="Migrant_Stock", hue="Year")
plt.tight_layout()
plt.title(label="WORLD Migrant Stock")
plt.show()
```



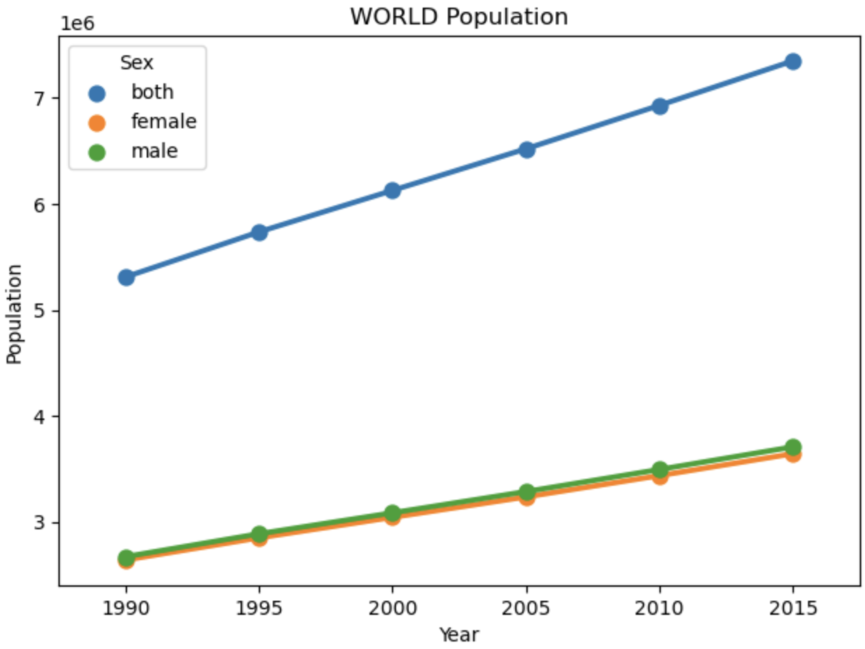
From this graph, we can easily see the migrant stock from 1990 to 2015, and the difference between male, female and both sexes.

For table 2, the data is about the total population of both sexes / male / female at mid-year. I used the point plot to visualize this dataset because it is easy to see the trends of population over a period of time. The horizontal axis is the year and the vertical axis is the population.

By applying the code



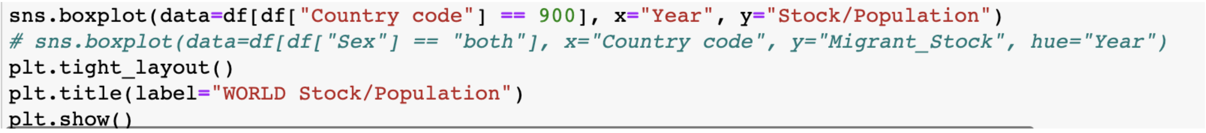
```
sns.pointplot(data=df[df["Country code"]==900], x="Year", y="Population", hue="Sex")
# sns.pointplot(data=df[df["Sex"] == "both"], x="Country code", y="Migrant_Stock", hue="Year")
plt.tight_layout()
plt.title(label="WORLD Population")
plt.show()
```



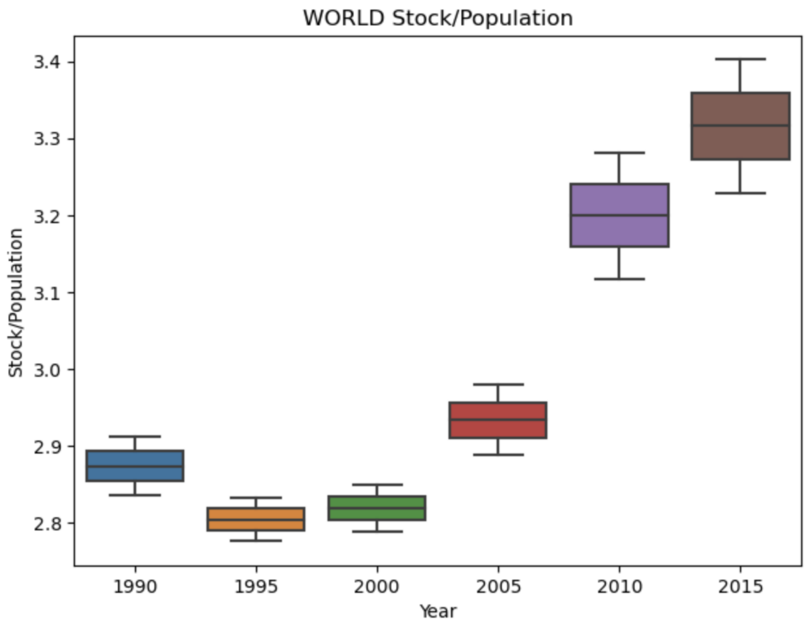
This graph shows the world population that there are all positive correlations between the year and population. From 1990 to 2015, the population of both sexes, male and female groups were continually increasing.

For table 3, the dataset is about the international migrant stock as a percentage of the total population for both sexes, male and female groups. For this table, I made the box and whisker plot. The horizontal axis is the year and the vertical axis is the percentage of stock/population. This graph has a sight difference from the above two because the graph does not show the data of three groups separately. We can look at the data as a whole and see the percentage from 1990 to 2015.

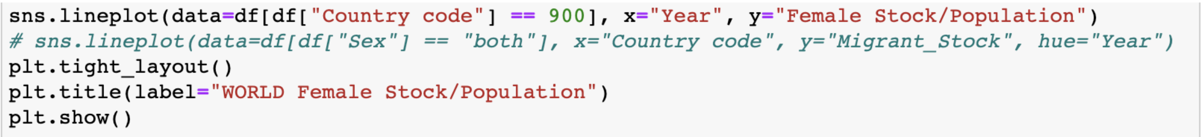
By applying the code of box and whisker plot



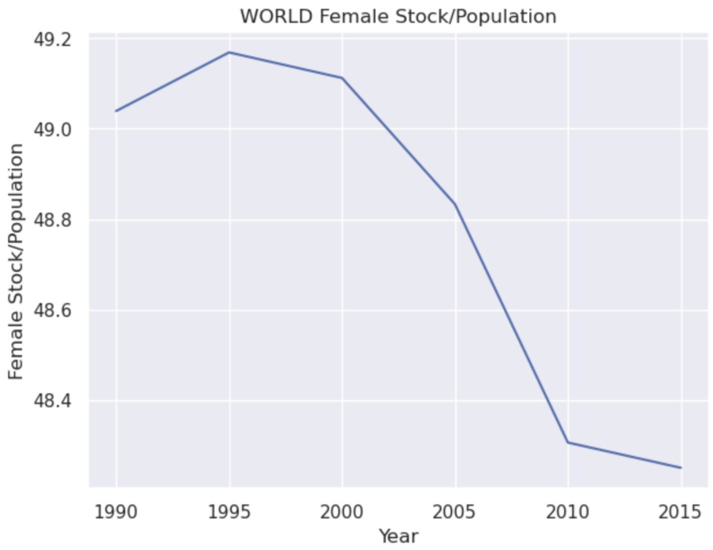
```
sns.boxplot(data=df[df["Country code"] == 900], x="Year", y="Stock/Population")
# sns.boxplot(data=df[df["Sex"] == "both"], x="Country code", y="Migrant_Stock", hue="Year")
plt.tight_layout()
plt.title(label="WORLD Stock/Population")
plt.show()
```



For table 4, the data is about the female migrants as a percentage of the international migrant stock. For this graph, I used the line graph to visualize the data. The line graph is the best way to show the trend over a period of time.

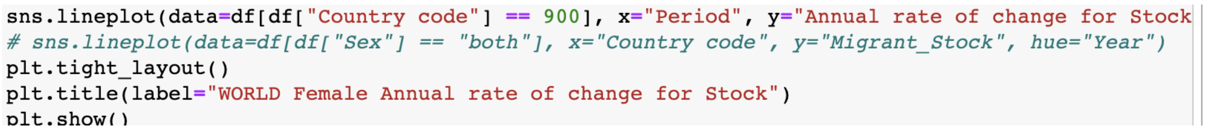


```
sns.lineplot(data=df[df["Country code"] == 900], x="Year", y="Female Stock/Population")
# sns.lineplot(data=df[df["Sex"] == "both"], x="Country code", y="Migrant_Stock", hue="Year")
plt.tight_layout()
plt.title(label="WORLD Female Stock/Population")
plt.show()
```

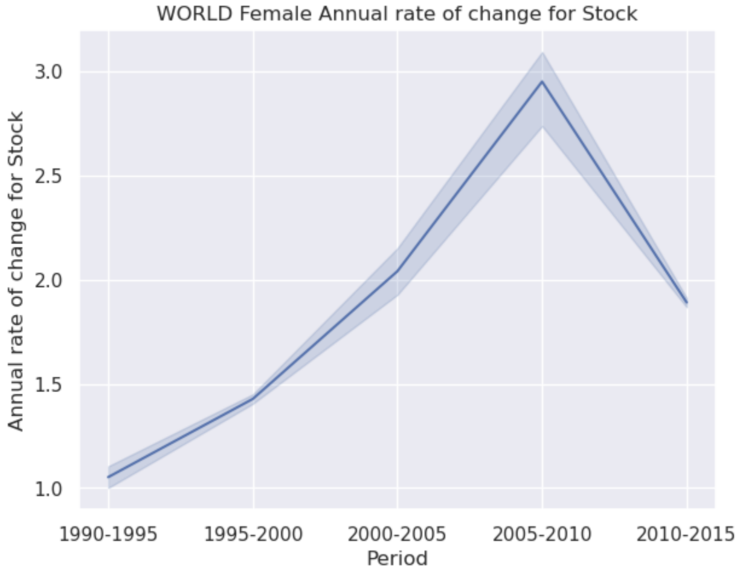


From the line graph, I discovered that the world stock/percentage was increasing from 1990 to 1995, however, it was continually decreasing from 1995 to 2015. The steeper the slope, the more it decreased. The line graph can show all of this information to me right away.

For table 5, the data is about the annual rate of change of the migrant stock for both sexes, male and female groups. I used the line graph to visualize this table.

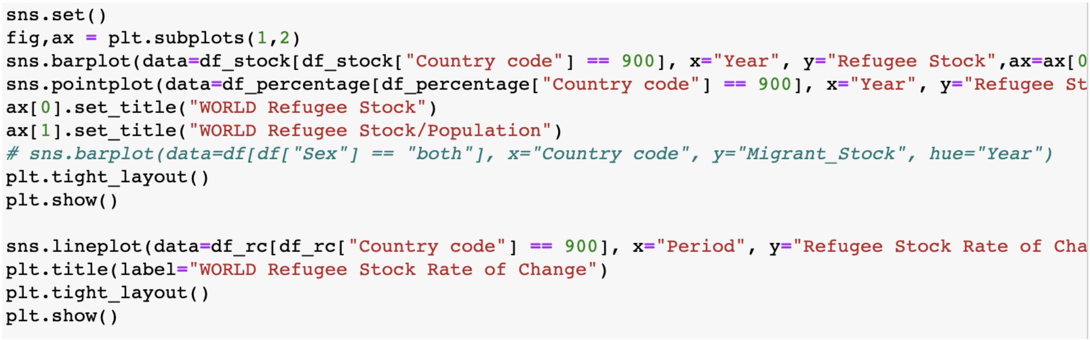


```
sns.lineplot(data=df[df["Country code"] == 900], x="Period", y="Annual rate of change for Stock
# sns.lineplot(data=df[df["Sex"] == "both"], x="Country code", y="Migrant_Stock", hue="Year")
plt.tight_layout()
plt.title(label="WORLD Female Annual rate of change for Stock")
plt.show()
```



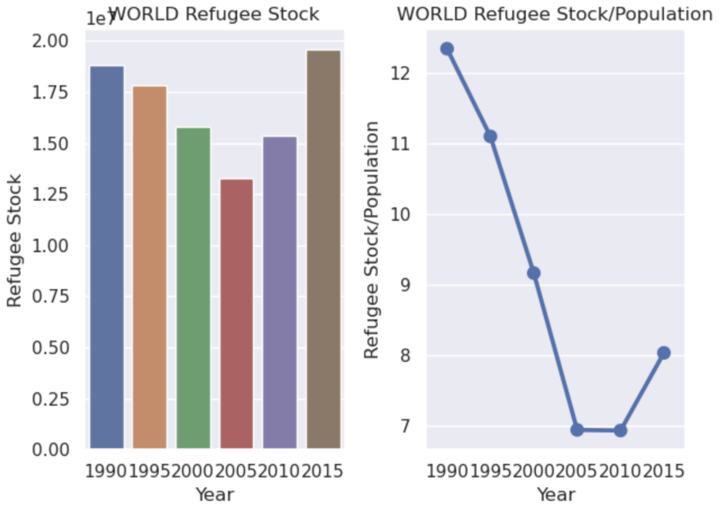
From the graph, we can see that from the period of 1990-1995 to the period of 2005-2010, the rate of change for stock was positive, which means that it was continually increased. After 2010, the annual rate of change for stock decreased.

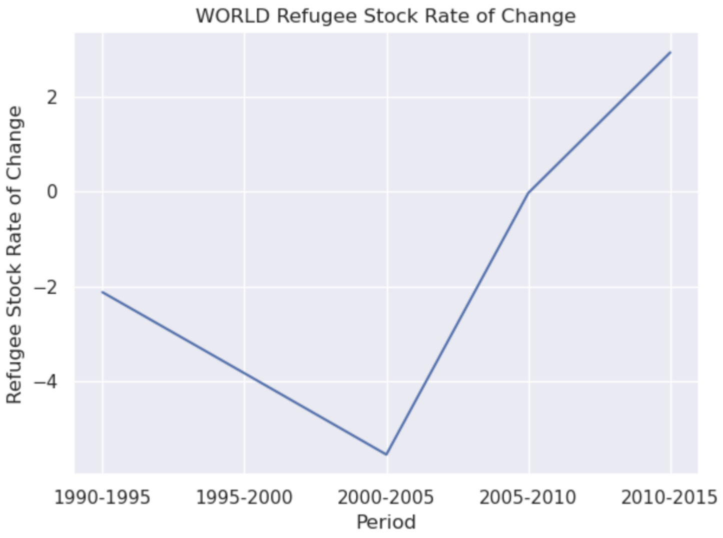
For table 6, the data is about the estimated refugees stock at mid-year, refugees as a percentage of the international migrant stock and the annual rate of change of the refugee stock. Therefore, I used three graphs: the bar chart, the line graph and the point plot. The bar chart represents the world refugee stock because it can easily show the number of the refugee stock by different years. The world refugee stock/population is shown by the point plot that represents the percentage of stock/population from 1990 to 2015. The line graph shows the world refugee stock rate of change from 1990 to 2015 by period of time (1 interval = 5 years), because the line graph is the best way to show the trend over a period of time.



```
sns.set()
fig,ax = plt.subplots(1,2)
sns.barplot(data=df_stock[df_stock["Country code"] == 900], x="Year", y="Refugee Stock",ax=ax[0
sns.pointplot(data=df_percentage[df_percentage["Country code"] == 900], x="Year", y="Refugee St
ax[0].set_title("WORLD Refugee Stock")
ax[1].set_title("WORLD Refugee Stock/Population")
# sns.barplot(data=df[df["Sex"] == "both"], x="Country code", y="Migrant_Stock", hue="Year")
plt.tight_layout()
plt.show()

sns.lineplot(data=df_rc[df_rc["Country code"] == 900], x="Period", y="Refugee Stock Rate of Cha
plt.title(label="WORLD Refugee Stock Rate of Change")
plt.tight_layout()
plt.show()
```





From these graphs, we can see that the world refugee stock was the highest in 2015, and lowest in 2005. The world refugee stock/population decreased from 1990 to 2005, and increased from 2010 to 2015. During the period of 1990-1995 to the period of 2000-2005, the refugee stock rate of change decreased, and it increased from the period of 2000-2005 to 2005-2010, with a flatter increase rate from 2005-2010 to 2010-2015.

Discussion

Data visualization is a good way to do the data analysis as well as to compare the data for different groups. All graphs I have created above only use the "WORLD" data that the country code is 900. The data visualization helps us to easily compare different data. For example, I can change the country code to 935 and 905 in the code, in order to compare the data of Asia and North America. The graphs can help me better understand the difference between two or more sets of data. I think it is a good way to learn and use analysis with comparison research between different sets of groups. Also, the different types of graphs have different characteristics. To better visualize the data, we should select the most appropriate type of graph that matches with the data to do the analysis. From this course, I have learned how to do the data cleaning with tidy data principles. After the data is cleaned, the data visualization helps me to do further research and analysis.