An Attempt on Exploratory Data Analysis using Tableau

Reka Supraja Thasma Lakshmanan Balajibabu

Computer Science Graduate Student, California State University, Sacramento.
rekasuprajathasmala@csus.edu

Abstract

Exploratory data analysis (EDA) involves selecting the data set, wrangling/ cleaning it and then start analyzing it by forming initial analysis questions. Then construct graphics/visualizations to address the questions using any of the data visualization tools such as tableau to see whether we will able to get the answers to all the initial review questions. Again, inspect the answers, modify the graphics if needed and assess the new questions by repeating the process.

Keywords: *Data Visualization, Data Cleaning, Data Analysis, Exploring the data, EDA, Data Wrangling, Tableau, Iterative exploration.*

1. Introduction

Exploratory Data Analysis or EDA is the first step in any of the data science project, which is exploring the data [1]. After we select the data we need to preprocess or clean it or we may not be able to get the correct outcome for example, if our data set has any null values then it will affect the visualization and so we need to cleanup/preprocess the data. This step is part of data wrangling. Then we have to come up with the initial analysis questions and start creating visualizations using any of the tools to answer the questions. EDA is an iterative process of creating graphics, assessing the answer and iterating the same for the next questions. So that we can transform the data appropriately to show the data variation.

2. Data Wrangling and Data quality assessment

Data wrangling is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. We have to check the Data quality on the data set we selected because there may be

- Missing Data such as no measurements, or redacted,
- Erroneous Values such as misspelling, outliers
- Type Conversion e.g., zip code to lat-Lon
- Entity Resolution e.g., different values for the same thing
- Data Integration effort/errors when combining data

Therefore, we have to preprocess/clean the data, assess the quality and then integrate in order to avoid data quality issues by doing manual manipulation in spreadsheets. There are also a lot of data wrangling tools available such as Custom code (e.g., dplyr in R, Pandas in Python), Trifacta Wrangler http://www.trifacta.com/products/ wrangler/ Open Refine (old Google's) http://openrefine.org/, etc…

3. Tableau

I used tableau as a data visualization tool for the EDA example in this article. Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry [2]. It can simultaneously specify both database queries and visualization in Tableau and users can easily make statements via drag-and-drop and we can easily add filters and see the changes immediately [3].

4. Example

4.1. Data set

I used The World bank data - by indicators masters. The World Bank has tracked global human development by indicators such as climate change, economy, education, environment, gender equality, health, and science and technology since 1960. I am interested to do data analytics on areas such as education, population, climate control etc… In addition, this data set is easy to download and clean. The data set is open source and available online [3]. I used Microsoft excel for cleaning this data set such as removing or filtering null values.

4.2. Initial Analysis Questions

I had the following initial questions

- Is this data set clean?
- How am I going to analyze it?
- How/What, am I going to select which part/section of this data set to start with?
- Am I going to learn and use Tableau to create visualizations?
- Am I going to use an interactive visualization?
-

After I decided to use the World Bank data – education data set, and after glancing the data I added these following specific questions

- What is the progression to Secondary school, male verses female?
- Whether the data set is clean?
- What are the Population growth rates in various countries?
- Which country has most graduate rates?
- Is there any relationship between mortality rate and population growth?

4.3. Discoveries and Insights

I imported the education csv into tableau and stared creating visualizations to get answers to the questions, which I framed.

Figure.1 and Figure.2 – answer these questions:

- What is the progression to Secondary school, male verses female from 1960 to 2020?
- Do you think the data after 2015 is correct?

I went back and looked at the data set after 2015, all the entries on progression to secondary school is zero. The data set does not say what the zero means whether it is null or numeric zero. Since I saw many zeros after 2014 in the data set I wanted to use filter in tableau to see only the progression to secondary school from 1996 to 2014.(Figure.2) Also I noticed that both male and female are about the same numbers for the below 2 visualizations.

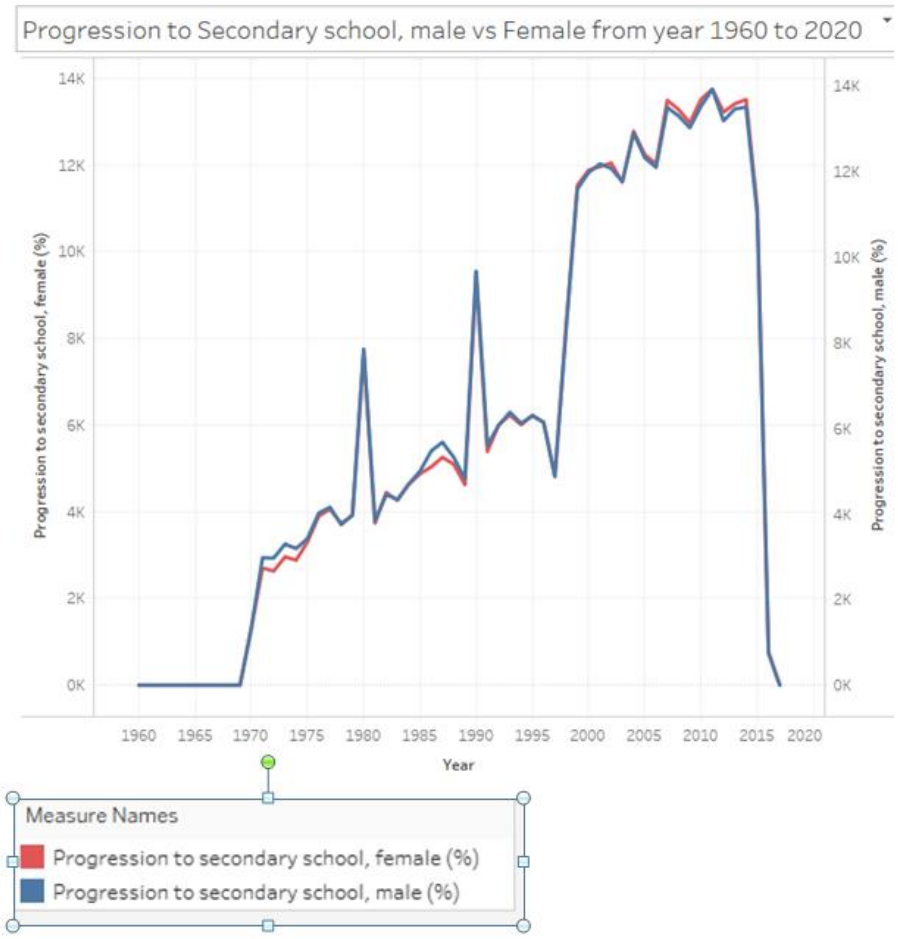

Figure 1. Visualization1

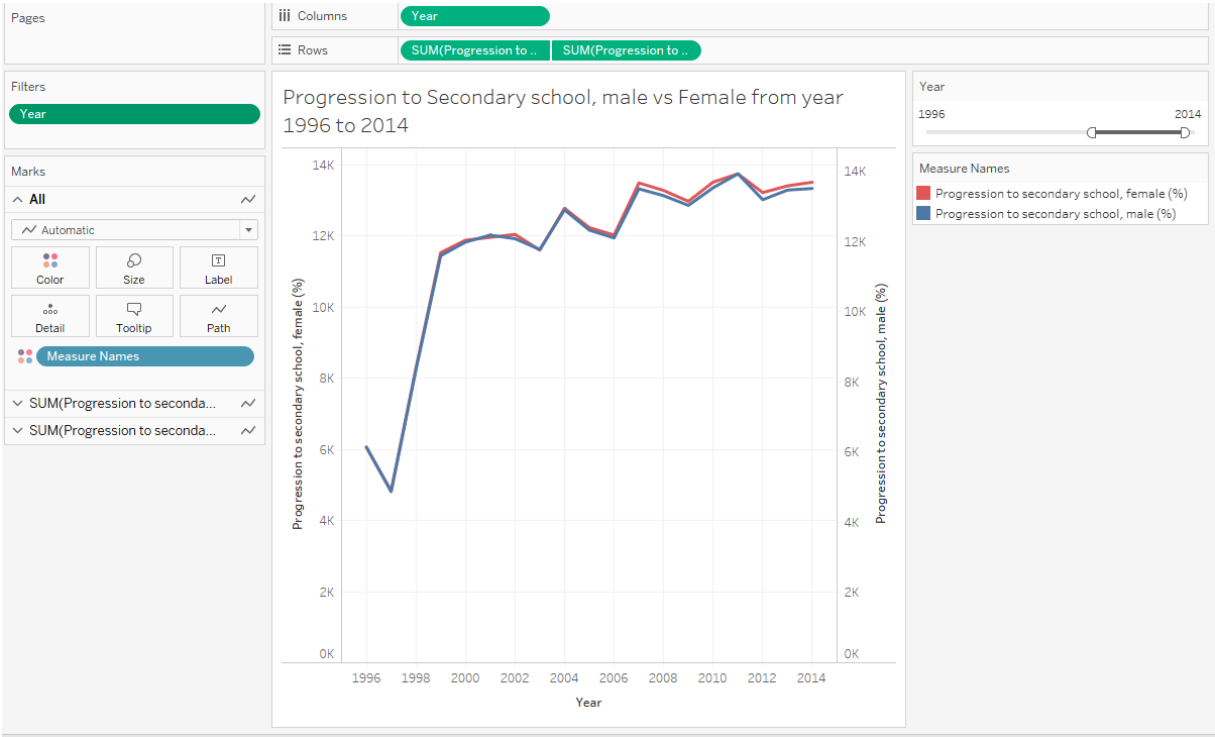

Figure 2. Visualization2

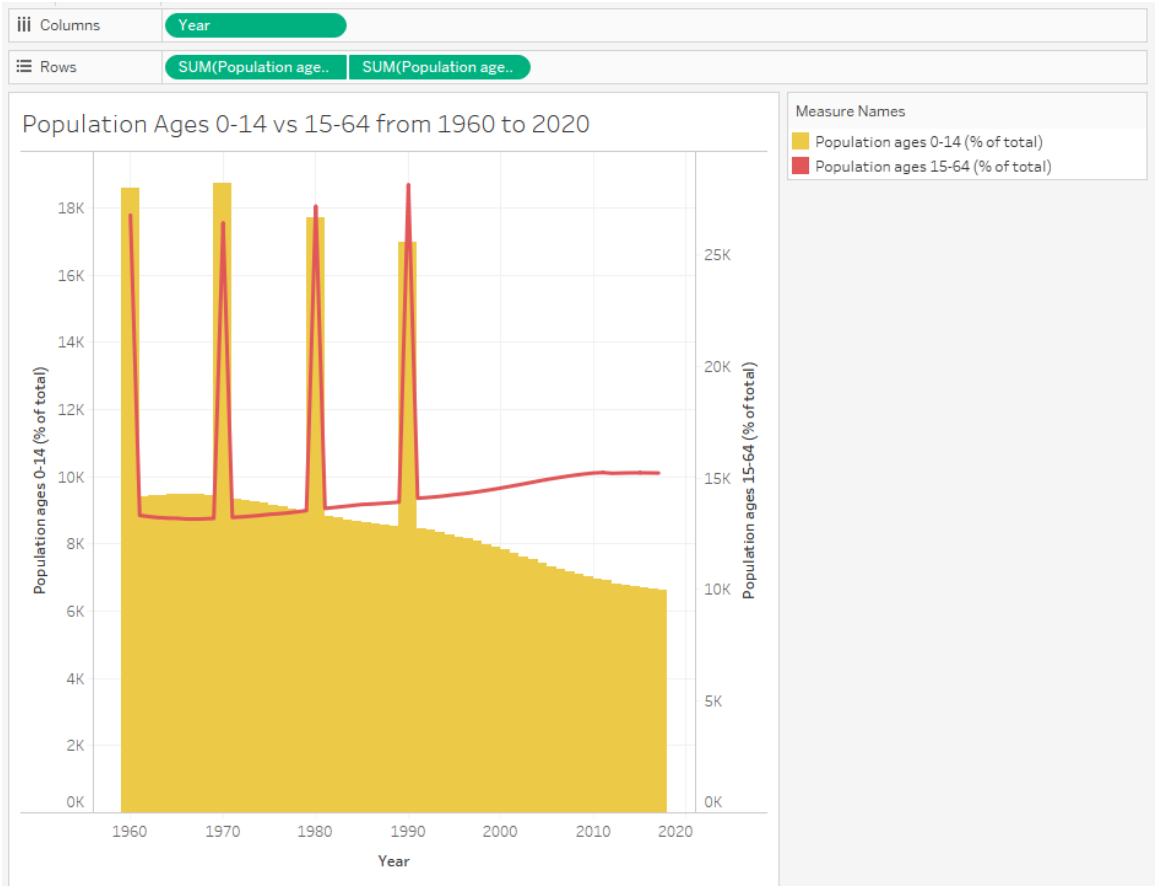

Figure 3. Visualization3

In Figure 3. I compared the population of ages 0-14 and 15-64 from 1960 to 2020. I verified in the data set that there is no zero until 2017. This picture shows that the population diminishes while lifespan increases with years.

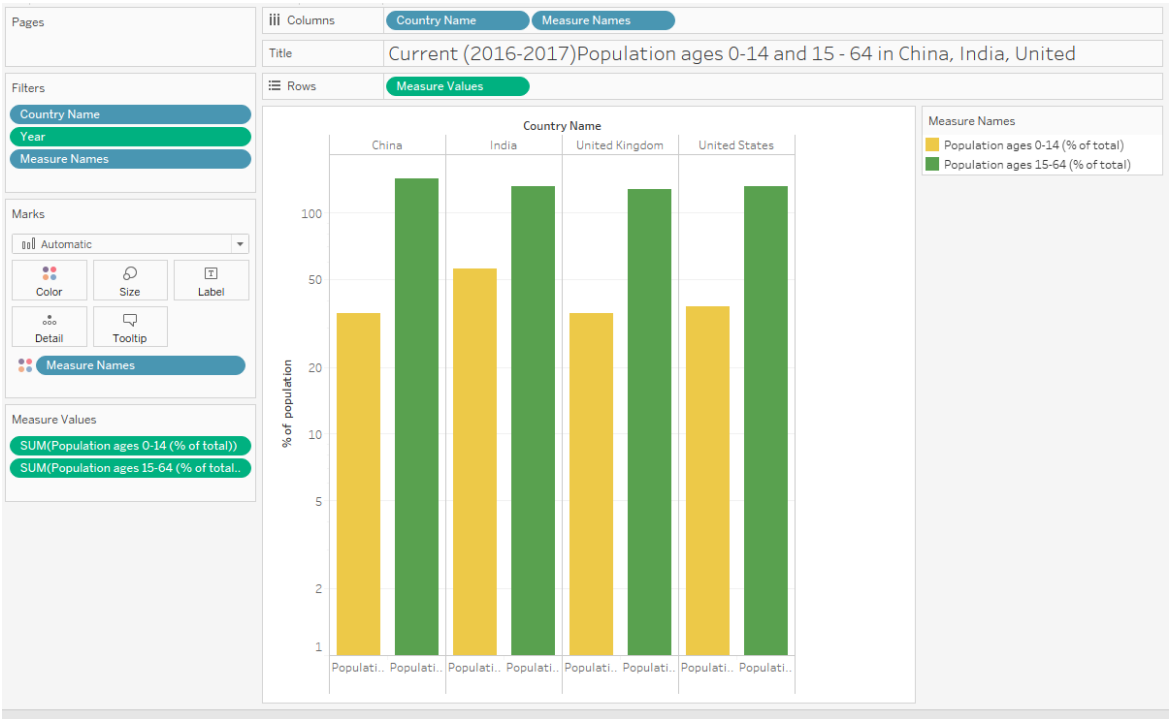

Figure 4. Visualization4

In Figure 4, I applied filter on countries and took only India, China, United Kingdom, and the United States. I also applied filter on year (2016-2017) to learn about current population. Here I took population ages and country in x-axis and percentage of population in y-axis. The current population of ages 15-64 is about the same for all four countries while the newborn population is higher in India.

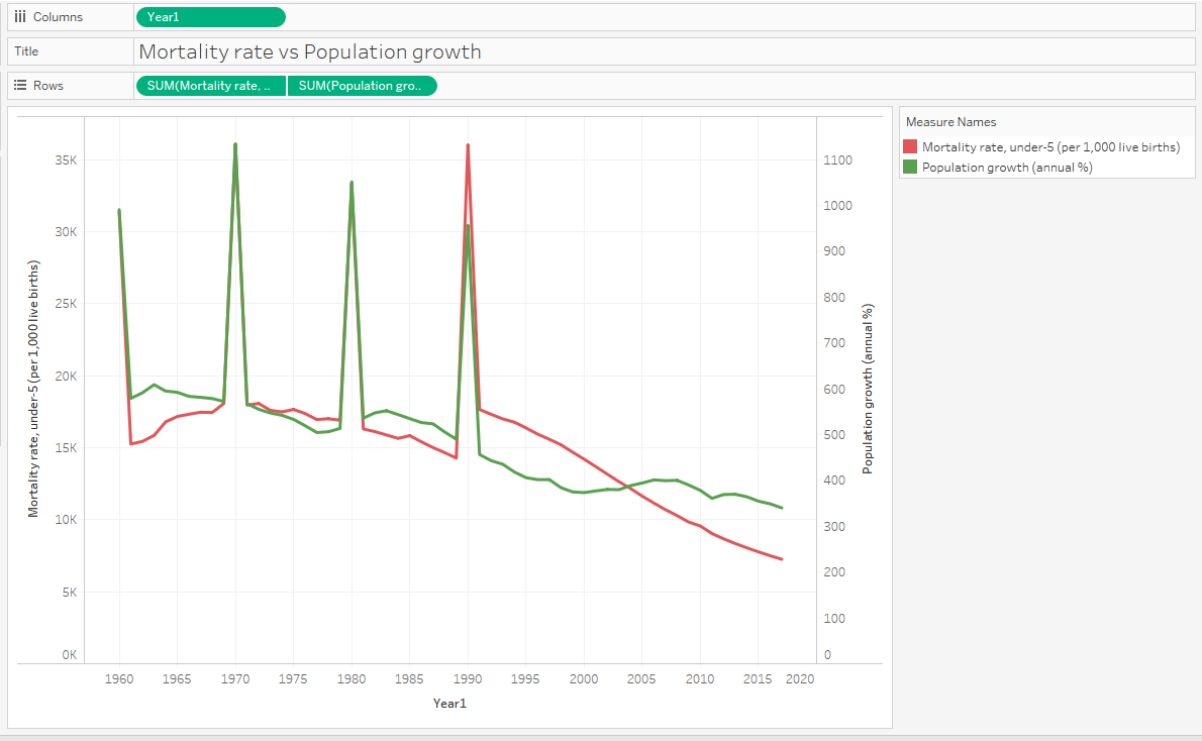

Figure 5. Visualization5

I analyzed the data set more to see if I can find less zero columns/rows and yes, I found. In the climate-change data set, I saw Population growth again which has very less zero entries and so I imported it in tableau and started visualizing. Here I took Mortality rate verses population growth. I surprised here that mortality rate and population growth are rather related and both the pattern is similar until 1995 and changed after 2005, as medical and technology field is keep advancing.

4.4. Conclusion

Even though I did spend a lot of time on this EDA, I have a feeling now that it is just a starting point to the exploration. I felt that choosing a data set took me more time. I first started downloading the Yelp dataset, but it was huge and I could not download and so I took World Bank data and started with education data. Again, I realized that this data is not very clean as I see many zeros in various fields and then I took climate-change data set and felt that this data set is cleaner than education as it has very few zeros. I thought that I wanted to do data analytics on education like which country has most graduate rates, in which year people graduated more etc but I could not find many details on this data set regarding that. Moreover, it is interesting that I took diversion from education to analyze population growth unknowingly and continued on that. I first thought that I have to filter the data in excel but then I realized that there is filter option in tableau which is great! The more you play around tableau the more you learn.

5. References

[1] https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch01.html.
[2] https://www.guru99.com/what-is-tableau.html.
[3] https://github.com/ZeningQu/World-Bank-Data-by-Indicators.
[4] https://www.tableau.com/.