Data Exploration Report

By Ezra Samuel.

Introduction

The goal of this project was to explore a dataset and find useful trends and relationships in the data and then present the findings using appropriate plots and visuals to convey interesting insights in the data. There a variety of data to choose from – either the ones provided or any other one of interest. For me, my data was self-sourced.

The Data

The data I used for this project was gathered from two main sources: gapminder.org (an online data bank) and Kaggle. They came in a set of five. The first set, though originally from gapminder.org was used for an earlier project, so it was retrieved and used for this project. These data were merged together to form a master data for the project.

The data was a set if information gathered from 167 countries between 2000 and 2017. It consisted data of child mortality rate, basic sanitation access, life expectancy, immunization reach, adolescent fertility rate, and so on. The final data had 2987 rows and 11 columns. My aim was to explore the data individually and in pairs to observe trends and see what factors influences child mortality.

Univariate Exploration

I explored the variables (numerical and categorical) variables individually. For the numerical variables, I used a histogram. I also transformed the axes of two of the variables to a log scale to assess the distributions better. I used bar plots to assess the categorical variables of interest – one nominal, the other ordinal. Our variable of interest – child mortality rate had a trimodal distribution.

Bivariate Exploration

I examined pairs of values and saw interesting trends. Child mortality rate had a strong positive relationship with adolescent fertility rate and a strong negative relationship with life expectancy, sanitation, and immunization. For the categorical variables, I observed that income category had a negative relationship with child mortality. Child mortality is high in low income areas and low in high income areas. I also noticed certain regions largely had high child mortality rates and interestingly these areas were low income regions. High income regions experienced the reverse.

Multivariate Exploration

The trends observed with multiple variables were similar to those observed when assessing the variables pairwise. A heat map, scatter plot of all the variables showed trends of all the data not just with the variable of interest, but with one another.

Presentation

My findings was communicated in a set of slides and visuals. Comments were also made to guide on the trends observed generally in the data. The visuals were polished in the way I saw that was best fit for presentation. I endeavored to add titles and labels to the plots.

Conclusion

The data exploration has been a very exciting one. They can strengthen a thought or present surprises altogether. This was just the best way of concluding this exciting journey.