Olivya Pastis

Professor Levkoff

ECON 494

25 October 2020

Project I Written Report

Through Gapminder, I chose to use the variables of child mortality, population, median age, life expectancy, and the employment rate for individuals 15+ within the year 2019 in order to compare how income affects them within every country. I began collecting my data by compiling the six different data sources into one Excel sheet as a means of constructing a cross-sectional dataset. I organized the columns by using the six variables and year as the headers, and organized the rows for each country alphabetically along with the respective data from each variable. Then, I proceeded to open the Excel file through RStudio, through which the data was cleanly inputted into its respective columns and rows.

Despite a lack of obvious indicators of disorganized data, I took steps to ensure it was optimally clean by coding NA for missing values and utilizing deletion when necessary. I created a new dataset, Project\_1, to compare any changes made from assigning NA to invalid/outlier values, but the dimensions of the new dataset were exactly the same. Since I ensured the data was cleanly compiled in Microsoft Excel onto one sheet, I did not need to rename column headers or recode any data.

Furthermore, I created several visualizations in order to depict and analyze unique types of graphs in order to display any relationships among the variables. I used income per person as my primary variable for comparison against the five other variables as a means of determining some kind of relation.

![Chart, histogram

Description automatically generated]()

As displayed above, I made a frequency histogram to illustrate a summary of the distribution of income per person among all the countries using a density scale for the vertical axis. Given the miniscule density percentages, the data varies significantly and showcases a positively skewed distribution of data for the income per person in terms of GDP/capita. According to the data, global income mainly comprises lower values and has a median of $12,000, which makes sense given the poor economic conditions of numerous countries.

Chart, scatter chart

Description automatically generated

I also used a scatterplot to illustrate the weak, indirectly proportional relationship between income per person and child mortality. Along with many outliers, there appears to be a lack of a sufficient relationship between these two variables, given the cluster towards low rates of child mortality regardless of income per person. Overall, an indirect relationship cannot be established between income per person and child mortality based on the data and graph.

![Chart

Description automatically generated]()

Through a violin plot, I compared the employment rate of individuals fifteen and older against income per person in order to see if there was a correlation. The graph portrays what appears to be a normal distribution, but with slight negative skewness. From using summary statistics, I determined the median and mean of the employment rate are both approximately .58, which conveys a slight positive relationship between the variables of employment rate of individuals fifteen and older and income per person.

![Chart, histogram

Description automatically generated]() ![Chart, histogram

Description automatically generated]()

The median age for all the countries is 29.35, which falls under a general category of population’s mainly consisting of individuals employed in the work force, which ties into the employment rate being above half the population for most countries, with a median of .58. The vast differences in global economic situations is a major factor in impacting the data, and explains the significant outliers within the boxplot model, where the maximum income greatly exceeds the median.

![Chart, scatter chart

Description automatically generated]()![Chart, histogram

Description automatically generated]()

Moreover, as shown in the scatterplot, life expectancy has one of the strongest, direct relationships with income, through which the best fit line slightly resembles the shape of a parabola. In the histogram on the right, the data is negatively skewed, displaying a tendency towards larger values. The two graphs illustrate the apparent direct connection between older ages and higher income per person.

![Chart, histogram

Description automatically generated]()

Similar to the income density distribution, the total population distribution is very positively skewed, with an emphasis on lower values. The depiction of graphs comparing income per person with population exemplified the lack of a relationship between the two variables, along with overbearing and inaccurate confidence intervals with the best fit line. Income per person and population shared the least in common, in terms of a relationship.

From scatterplots to histograms, I made sure to portray the unique data comparisons in different ways that best functioned with quantitative, numerical data. A majority of the variables had some kind of relation with income, but they were primarily skewed or were fairly weak. However, income per person and life expectancy had the strongest, directly proportional relationship to one another, which explained the negative skewness of the data. Ultimately, the lack of strong, directly proportional relationships among the variables with income surprised me, but also taught me the importance of data visualizations for optimal comprehension.