494 ECO PROJ 1

DATA ANALYSIS ON FACTORS AFFECTING INFLATION

University of San Diego

Dr. Levkoff

October 25, 2020

1. Data Collection

This project is intended to analyze factors that may affect inflation. The data set is

searched and downloaded from the Gapminder website. All of the variables are formed by a “mix-and-match format”, with each variable downloaded separately and combined afterwards. Based on pervious knowledge, 7 variables and 175 observations are selected and included in the data set. The 7 variables are annual inflation, GDP per capita growth, merchandise trade, income per person, Gini coefficient, annual population growth, and developed country. Among all variables, all preceding 6 variables are numerical variables, while the last one is categorical variable, in which all developed country are filled in with number 1; the rest are 0. All of the 175 observations used are collected among countries all over the world in 2019. For each country, the 7 variables are filled in with values for that particular country. With the data set uploaded onto GitHub, it can be extracted and uploaded into R Studio for further analyzation.

Figure 1 shows line 1-28 from the R script, which gives the basic structure of the data set, including the dimension, head, tail, view, class and summary of the data set. By running the lines, the data set matches the result from data extraction and proves that the data set is uploaded successfully with no missing data.

Graphical user interface, text, application, email

Description automatically generated

Figure 1 R Script on Data Structure

1. Data Cleaning

Since the data is “mixed-and-matched”, it was organized by eliminating all empty values before uploading to GitHub. Before cleaning the data, variables of the data set is too long to be used in R studio. In Figure 2, line 35-36 is created to change the titles of variables. In order to help plot the graphs, line 41-46 are created to convert all percentage numbers into decimals. With the convenient-extracting variables, the data set can be cleaned with narrowing the ranges for each variable, which are included in line 49-54. The cleaned data is included in a new variable called “cleandata”. Since the uploading data process added 2 extra NA columns on the data set, line 56-58 are created to delete those columns. After the cleaning process, there are no data lost but there are some data turning into NA in order to reduce some outliers for plotting. It can be checked with line 61 and 62 resulting in the same-dimension data. After all those processes, the data set is ready for further analyzation.

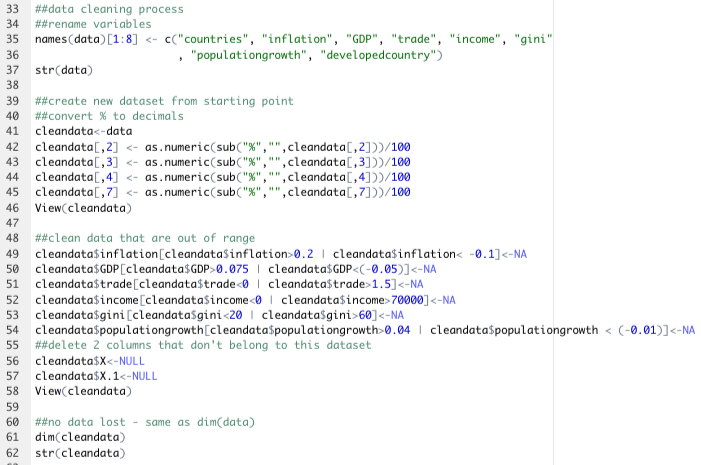


Figure 2 Data Cleaning Process

1. Data Analysis

The most important step in this project is to use analysis to build visualization. Package

ggplot2 and the library have to be installed first, which is reflected in line 71 in Figure 3. Lines 74-85 are used to load libraries, load preinstalled data from ggplot2 library, and check to see if there are any lost data. In order to support the plots, 3 other packages need to be uploaded: plyr, MASS, and mgcv.

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure 3 Package Installed

Before plotting graphs to understand the relationship between inflation and 5 other variables, graphs of density curves and histograms are plotted to help understand the distributions of each variable. The codes are included through lines 89-112 with the resulted graphs placed in Figure 5 and 6. As shown in those figures, all 6 distributions can be classified as normal distribution with some skewness.

Text

Description automatically generated

Figure 4 Coding for Distribution Graphs

Chart

Description automatically generated

Figure 5 Resulted Graphs for Distribution Graphs (inflation, GDP, Income, Population Growth)

Chart, histogram

Description automatically generated

Figure 6 Histograms for Gini and Trade

Next, in order to find factors that influence inflation most with the data set, a new data set was created and new limits were set for eliminating the outliers in order to have a better plot for the relationship. Lines 116-119 shows the cleaning step to limit on the boundaries. Lines 121-126 are codes used to graph the scatter plots between inflation and GDP and between inflation and income. Figure 8 shows the results of those 2 graphs. According to the graphs among the sampled countries, GDP and inflation has a positive relationship. In other words, as GDP goes up, inflation will go up. While there is a negative relationship between income and inflation. In other words, as income goes up, inflation will go down.

Text

Description automatically generated

Figure 7 Code for Scatter Plots

Chart, scatter chart

Description automatically generated

Figure 8 Scatter Plots between Inflation and GDP & between Inflation and Income

Then, in order to determine how inflation changes with population growth among developed countries and developing countries, a scatter plot with 2 subset of data is shown in Figure 10. According to the graph, the relationship between population growth and inflation is positive among developed countries, but the relationship is slightly negative among developing countries. It can be inferred that among developed countries, as population grow, inflation will grow according. The opposite situation applies for developing countries. Lines 136-138 shows coding for violin plots. Figure 11 shows a visualization of two violin charts of the relationship between trade and population growth with inflation separately. The graphs tells that inflation is more spread out among countries when trade is low; inflation is spread out almost evenly no matter how the population grow among countries.

Text

Description automatically generated

Figure 9 Another Scatter Plot and Other Graphs

Chart, scatter chart

Description automatically generated

Figure 10 Scatter Plot with Developing Countries and Developed Countries

A picture containing chart

Description automatically generated

Figure 11 Violin Plots

1. Conclusion

This project is designed to find the factors that influence inflation and find relationships

between factors and inflation. Based on the coding and visualization shown, there is a relationship between inflation and factors, such as income, GDP and population growth. In this project, whether a country is developing country or developed country also take into account in finding the relationship; and being as a developing country or developed country does have different impact on inflation. Furthermore, all relationships in this report can be identified toward a “flat” relationship instead of a “steep” relationship, which indicates that those relationship are not too strong. So there is need to involve more data from different countries and time period in order to further analyze the relationship between inflation and other factors.