Statistics Term Project

David Berberena

02-11-2024

Introduction

Looking at today's world, we have advanced far in many aspects of life. From technology to medicine, the human race has managed to exceed its own limits to craft some things made to make the world a better place for those living in it. We can also look at exceeding our own lifespan by using these advances we have made to realize a longer life, one that could easily be spent enjoying what others have created to make life more convenient. With that, I'd like to take a deeper look into what factors are relevant to increasing one's life expectancy. Now I'm not referring to the Fountain of Youth, but the presence of exterior factors can affect the ability to survive up to a certain age. Of course if you're interested in living a longer life, this may be something that you would at least give a quick looks towards. Data science can also play a huge role in this topic of life expectancy and could be considered an industry issue due to the fact that in order to know how long a customer could be eligible for certain projects, advertisements and/or events, the life expectancy of a sample derived from the dataset could be used in targeted advertisements that now fit the increased age of life expectancy in humans. With more people surviving to be able to use a certain product or service in question, Data Science can help companies perform better in higher age demographic sales. How would one go about providing insight on the topic of increased life expectancy factors? First, we can begin by brainstorming meaningful and concise questions.

Research Questions

With the factors of life expectancy being looked at to see which ones prove to have a significantly positive effect on the life expectancy, we can ask the following questions:

1. What factors (variables) are evident in datasets concerning life expectancy?
2. How will these variables seen today affect life expectancy in the future?
3. Are there any variables that have a negative correlation with life expectancy?
4. Can the factors that affect life expectancy be displayed by a linear regression model or another popular form of regression?
5. How many years gained (0.5 years, 1 year, 3 years, etc.) is considered to have a "significantly positive" effect on life expectancy in the topic statement?

Approach

My plan to address the issue of discovering what variables are significantly relevant to increasing a person's life expectancy is rather straightforward regarding the work needed to be done on datasets obtained for the purpose of answering this question. I would take each dataset and extract each variable within the dataset to then bind them into a matrix and create a correlation matrix. Only the top three variables with the highest correlation coefficients in relation to the life expectancy variable would then be used to create a multiple regression model. I would have one multiple regression model for each dataset and compare the p-values of each of the three predictor variables to see if any of the variables chosen have a significant effect on the life expectancy variable.

How The Approach Addresses The Problem

To realize the answer to the problem at hand, I would need to find the variables with statistically significant p-values and correlation coefficients that are close to one after creating both the correlation matrix and the multiple regression models. To do this, the creation of the correlation matrix would show the correlation coefficients for all variables present in each dataset involved, and would also show which variables are the most positively correlated to the life expectancy variable. Taking the top three variables from each dataset and making multiple regression models for them addresses the problem from the statistically significant aspect, since the F-statistic containing the p-value shows whether the effect on the life expectancy variable is statistically significant.

Data

There are three datasets that I will be implementing on my journey to address the topic statement on what factors have a significantly positive effect on life expectancy. They are as follows:

1. KANAWATTANACHAI, P. (n.d.). Healthy Lifestyle Cities Report 2021. Kaggle. Retrieved February 7, 2024, from https://www.kaggle.com/datasets/prasertk/healthy-lifestyle-cities-report-2021

This dataset contains statistics regarding healthy lifestyle metrics in forty-four popular cities. This data was collected in 2021, and the original dataset contains twelve variables (city and rank included) related to a healthy lifestyle. In the original dataset, there was no attempt to fill in missing values, so in order for me to use this dataset, I will have to transform the data to fill in the missing observations.

2. THARMALINGAM, L. (n.d.). Health and Demographics Dataset. Kaggle. Retrieved February 7, 2024, from https://www.kaggle.com/datasets/uom190346a/health-and-demographics-dataset

This dataset here was created specifically for the purpose of exploratory data analysis using life expectancy data. The data collected within the dataset was generated between 2000 and 2015, as evident by the year variable within the data. 22 variables make up this dataset, and there were no missing or strange values input into the file, so it is ready to use.

3. Gochiashvili, L. (n.d.). Life Expectancy (WHO) Fixed. Kaggle. Retrieved February 7, 2024, from https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated?resource=download

The intention of exploratory data analysis is evident in this set of data as well. However, this dataset is the most involved in terms of effort expended to create it. As the data was gathered by Kaggle Source, this dataset was pieced together by collecting data on each variable separately from other websites and placing it all into one readable format. The observations collected here also were from 2000 to 2015 and involved 179 countries, ultimately culminating with 21 variables in the final dataset. Missing values in the original dataset for any year were fixed by inputting the closest three-year average. For countries missing values for all years within the dataset, the fix was to fill it in with the average Region values.

These sets of data will require the usage of specific packages within R to manipulate their contents to yield the results needed to answer the topic statement.

Required Packages

The required packages that should be brought to the table are the following:

- readr (load datasets into the R environment)
- ggplot2 (plotting scatter plots to show linear correlation)
- dplyr (various data manipulation functions and pipe operators)

These libraries along with the built-in base R functions will allow the proper extraction of the dataset variables for the correlation matrices and the subsequent creation of the multiple regression models.

Plots And Table Needs

As stated earlier, a correlation matrix would be needed for each dataset to identify the three most positively correlated variables to life expectancy. These variables then could be plotted via the ggplot2 library to visualize their linear correlation with life expectancy. QQ plots can be used to plot the residuals of the multiple regression models to view whether or not the models created are a good fit for the datasets that they represent.

Questions For Future Steps

Looking at the topic statement and the accompanying brainstorming questions, I would like to see if I can figure out how to realize the predicted increase in life expectancy in years given the input of a variable positively correlated to life expectancy. This would certainly involve testing the regression models to output a prediction based on a known input, yet I am not quite sure how to do that in this context. In conclusion, using the variables highly and positively correlated to life expectancy to build multiple regression models and compare p-values for statistical significance looks to be the best way to address the topic statement of what factors are significantly relevant to the increase in a person's life expectancy.