Statistics Term Project

David Berberena

02-18-2024

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. Life insurance companies like AARP could benefit from a data science study relating to life expectancy. Recurring monthly subscription services for the elderly like Life Alert could realize an increase in revenue if the company were to see the factors involved in a longer life, as their customers could then be notified of them so they would do their best to live longer using Life Alert services while the company earns more money. The finance world would also potentially be affected with the identification of factors that increase life expectancy. Depending on the prevalence of these factors in countries, the financial sector may wish to reevaluate the age at which Social Security, pensions, and 401k retirement plans are distributed so that the recipients will have a greater chance of not running out of funds before their lives end due to the prominence of factors that positively affect life expectancy. With all of the potential effects that this data-oriented issue could have on business, 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. Taking several variables from each dataset to test against life expectancy would be computationally expensive and time consuming, so three variables from each dataset should be sufficient to visualize the various factors involved in life expectancy and how each predictor weighs on life expectancy. Continuing to flood the multiple regression models with an increasing number of variables saturates the data and leads to the model not being a good fit for the data with the inclusion of so many data points and the resulting residuals. 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 significantly positive 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)
- purrr (more data manipulation functions to reveal new information)

  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.

Cleaning of the Data

To clean the datasets being used, we must import the datasets that have been downloaded to the local machine into RStudio. We can do this by loading (or installing if it has not yet been installed on the local machine) the readr package, setting the working directory to ensure smooth file importing with the setwd() function, and reading the comma sepearated dataset files into RStudio with readr's read_csv() function.

Once the datasets have been imported, we can see that two of the datasets have already been properly cleaned and one has a few missing values. We will focus on cleaning this one dataset named healthy_life. The missing values will be filled in with the average of the column in which the missing values are present. To do this, we must check the dataset for columns with missing values and identify them as a variable so RStudio can work with those specific columns. Next, we need to make sure that RStudio recognizes the values in the specified columns as numeric. With this particular dataset, the missing values are denoted with a dash, so these dashes need to be converted into actual NA values. With all of this now complete, the averages of the columns can be found and can be appended to the columns with the NA value present. Now that the dataset is cleaned, it can be used like the other two with no issues.

Here are the datasets in their cleaned forms:

life_expectancy

```
## # A tibble: 6 x 21
##   Country    Region         Year Infant_deaths Under_five_deaths Adult_mortality
##   <chr>      <chr>         <dbl>         <dbl>             <dbl>           <dbl>
## 1 Turkiye    Middle East    2015          11.1              13             106.
## 2 Spain      European Uni~  2015           2.7               3.3            57.9
## 3 India      Asia           2007          51.5              67.9           201.
## 4 Guyana     South America  2006          32.8              40.5           222.
## 5 Israel     Middle East    2012           3.4               4.3            58.0
## 6 Costa Rica Central Amer~  2006           9.8              11.2            95.2
## # i 15 more variables: Alcohol_consumption <dbl>, Hepatitis_B <dbl>,
## #   Measles <dbl>, BMI <dbl>, Polio <dbl>, Diphtheria <dbl>,
## #   Incidents_HIV <dbl>, GDP_per_capita <dbl>, Population_mln <dbl>,
## #   Thinness_ten_nineteen_years <dbl>, Thinness_five_nine_years <dbl>,
## #   Schooling <dbl>, Economy_status_Developed <dbl>,
## #   Economy_status_Developing <dbl>, Life_expectancy <dbl>
```

life_two

```
## # A tibble: 6 x 22
##   Country      Year Status   `Life expectancy` `Adult Mortality` `infant deaths`
##   <chr>       <dbl> <chr>                <dbl>             <dbl>           <dbl>
## 1 Afghanistan  2015 Develop~                65               263              62
## 2 Afghanistan  2014 Develop~                59.9             271              64
## 3 Afghanistan  2013 Develop~                59.9             268              66
## 4 Afghanistan  2012 Develop~                59.5             272              69
## 5 Afghanistan  2011 Develop~                59.2             275              71
## 6 Afghanistan  2010 Develop~                58.8             279              74
## # i 16 more variables: Alcohol <dbl>, `percentage expenditure` <dbl>,
## #   `Hepatitis B` <dbl>, Measles <dbl>, BMI <dbl>, `under-five deaths` <dbl>,
## #   Polio <dbl>, `Total expenditure` <dbl>, Diphtheria <dbl>, `HIV/AIDS` <dbl>,
## #   GDP <dbl>, Population <dbl>, `thinness  1-19 years` <dbl>,
## #   `thinness 5-9 years` <dbl>, `Income composition of resources` <dbl>,
## #   Schooling <dbl>
```

healthy_life

```
## # A tibble: 6 x 12
##   City        Rank `Sunshine hours(City)` `Cost of a bottle of water(City)`
##   <chr>      <dbl>                  <dbl> <chr>
## 1 Amsterdam      1                   1858 £1.92
## 2 Sydney         2                   2636 £1.48
## 3 Vienna         3                   1884 £1.94
## 4 Stockholm      4                   1821 £1.72
## 5 Copenhagen     5                   1630 £2.19
## 6 Helsinki       6                   1662 £1.60
## # i 8 more variables: `Obesity levels(Country)` <chr>,
## #   `Life expectancy(years) (Country)` <dbl>,
## #   `Pollution(Index score) (City)` <dbl>, `Annual avg. hours worked` <dbl>,
## #   `Happiness levels(Country)` <dbl>, `Outdoor activities(City)` <dbl>,
## #   `Number of take out places(City)` <dbl>,
## #   `Cost of a monthly gym membership(City)` <chr>
```

Looking at the final cleaned datasets, I do not believe there is anything that I can do to clean and/or condense the data any more than I have. I have chosen not to merge the datasets into one dataset for the cleaning

process as each dataset functions well enough on its own and there are different metrics being measured related to the main topic of factors affecting life expectancy.

Data Manipulation

For data that is not self-evident, which is quite a lot, there are many things that could be done to showcase the data in different ways to reveal new information. For example, if I wanted to know the average values of all columns in each dataset, I could use the purr package's map_dbl() function using a pipe operator for each dataset. This would output the column name and the respective average for numerical columns, and for character columns the output would be the column name and NA as the average. If I wished to list all of the factors involved in life expectancy computation across all datasets, I could use the colnames() function for each dataset and take the unique names that are outputted (the datasets do have column names that are the same) and show them as all of the variables that play a role in life expectancy, hereby answering the first of my five questions listed above. To go even further, I could create a correlation matrix of all the variables within one dataset to show the correlation between each variable and life expectancy. This could be done by storing an entire column of data into a variable using the $ operator and using the cbind() function to create a matrix. Once the matrix is formed, the correlation matrix can be created using the cor() function. Repeating these steps for wach dataset would yield three correlation matrices. Viewing the correlation matrices would answer the third of my five questions listed above, as the negative numbers in the matrices would indicate which variables have a negative correlation to life expectancy.

Different ways to view the data involve the use of the dplyr package, as the functions within this package allow for the datasets to be seen in newer and more specific ways. Two of my datasets involve data recorded each year for a certain time, so I could see the average observations by year for each dataset by using dplyr's group_by() and summarize() functions in conjunction with pipe operators. In the healthy_life dataset, there are variables that represent the country and the city, yet only cities are listed within the dataset. I could separate this one dataset into two datasets by adding a Country variable that lists the country that each city resides in by using the mutate() function in dplyr and then split the dataset into two new datasets via the select() or filter() functions in dplyr. I would then be able to realize correlation coefficients based on geographical locations on the same scale (city correlations and country correlations would be viewed separately and statistics would not be combined). I can even compare the values from the same variables in different datasets to see how they differ. For example, the life expectancy variable is featured in all three datasets, with life_expectancy and life_two containing this statistic for many of the same countries and years. Focusing on these two specific datasets, I could take the yearly averages of the countries each dataset has in common and place those into a new column using the group_by(), summarize(), and mutate() function. Using the pull() function in dplyr, I could then isolate that newly created variable in each dataset as two charatcer vectors and then I could create a new dataset to view the average yearly life expectancy by country for both datasets side-by-side and see the difference values.

Most of the above methods of viewing the data in different ways involves slicing and dicing the data. Creating new variables to see statistics not seen in the original data and creating new data frames to see the comparison of some of the same variables in different datasets having varying observations shows that even data of the same nature is collected differently. In theory, the life expectancy of the people in Turkey in 2014 (for example) from one dataset should be the same as this same statistic in the second dataset, but that is not the case, and seeing this discrepancy helps data scientists create more questions and ponder on why data reveals seemingly puzzling information.

Summarizing the data that has been manipulated as seen above is mainly done through the summarize() function in dplyr. The correlation matrices are summarized already as they share the one correlation coefficient needed for the relationship between different variables. So far two of the five questions have been given pathways to be answered by using the manipulation strategies I have employed. The other three questions' answers lie in regression analysis. The summary() function in conjunction with multiple linear regression and time series analysis would list the F-statistic with p-values needed to see how significantly positive a highly correlated variable would affect life expectancy. Once we have realized the regression models needed,

we could look at plotting the model data points with the actual dataset data points using a scatter plot and adding a line of best fit. Residuals could also be plotted with the model values to see the variance of the data.

Plots And Table Needs

As stated earlier, a correlation matrix using the cbind() and cor() functions 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 using the geom_point() function with labels such as the title, the x-axis featuring the predictor variable, and the y-axis being life expectancy, as this is the dependent variable whose results we are attempting to see. In total, nine scatter plots would be needed to visualize the three most positively correlated variables from the three datasets. 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. The creation of the multiple regression models answers the fourth of the five research questions, as the model shows the linear relationship between the predictor variables and the life expectancy outcome variable. The QQ plot can be created with the qqnorm() function and the addition of a title using the title() function. A line of best fit can be added to the QQ plot with the help of the qqline() function. Scatter plots can be made of the model data points against the actual dataset observations by implementing the same geom_point() function as the correlation scatter plots, and the line of best fit for the scatter plots here can be added with the geom_smooth() function with the "method" argument being set to "lm." These tables and plots help to directly answer the second research question stated above, as the model function visualized by these plots can be used to predict the life expectancy of new predictor variable observations.

As seen above, the use of regression models is apparent in the endeavor to answer the questions I have of the datasets. These regression models are considered to be a machine learning technique being utilized. I would be inputting the variables I flesh out of the data with the highest and most positive correlation into a model that returns predicted values of life expectancy. Once the model has outputted the predicted values and we can visualize the relationship between the variables being tested, we would then be able to predict life expectancy values for points not present in the datasets, with machine learning leading the way to that end.

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 through the creation of this research paper, I have an idea of how to accomplish this. What I have an inquiry about now is the fifth question I have stated. Reviewing the question at this point in my research begs this question to be changed a bit. I would like to phrase the question as such:

"What increase in life expectancy (0.5 years, 1 year, 3 years, etc.) is considered to be"significantly positive?"

While this question sounds like a more wholesome inquiry than the previous question, I still find it rather difficult to answer this question with the data. What the data can answer for me is what variable presents the highest statistical significance value to the increase of life expectancy. This would not be able to tell me if a one year increase in life expectancy is considered "significantly positive" as this is something data simply cannot tell. This question is geared more towards the opinion of the person who is answering it. To a person who is on the verge of death, a 0.5 year increase is rather significant. Someone who is relatively young and healthy may not view that 0.5 year increase in the same way and may yearn for a larger increase. 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.