

Statistics Term Project

David Berberena

03-02-2024

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 separated 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 `purrr` 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 each 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 character 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.

The Story of the Data

Introduction

Life expectancy is one of those grim subjects that people tend to shy away from as humans typically do not like to think of the lack of time they have on this planet. While people do their best to exercise, sleep a certain number of hours, eat right, and be happy, they still are not sure if those things are enough to truly prolong the amount of time we have to be alive. Society deems a lot of the factors just stated as those that could potentially maximize one’s life expectancy, yet when it comes to entire countries, data science demands more concrete topics that encompass the majority of the people living there. With data science being used to help businesses and customers alike, the scenario of whether data science can shed light on those significant factors affecting life expectancy is born.

Problem Statement Summary

Now that the topic of data science aiding in the search for important factors pertaining to a life expectancy increase has been created, the generation of a testable problem statement is needed. Using data science concepts like correlation and regression, I generated multiple questions that led me to the creation of a consolidated problem statement. That singular question is this:

What are the most highly correlated factors to life expectancy, and do those factors realize a significant increase in life expectancy when applied through linear regression?

Addressing the Problem Statement

With the problem statement at hand and data science skills ready to tackle the scenario, the process of unraveling the answer can begin. First, the gathering of appropriate and reliable data must take place. Kaggle is a well-known and reputable public database among the data science industry, so I acquired three datasets that looked complete and had the life expectancy variable within each one. These datasets went through a data cleaning process to fill in any missing values with averages and/or to drop observations with missing values. Once the data was cleaned, every variable in one of the datasets was extracted and placed into a correlation matrix to visualize the three most positively correlated variables in relation to the life expectancy variable. This process was repeated with all of the other datasets to obtain those most positively correlated variables as well. Once found, those variables were then placed into three multiple linear regression models (one for each dataset) as predictors with the life expectancy variable in each dataset as the outcome variable. Multiple linear regression is the route needed for this type of data and problem scenario as I am looking at a discrete quantitative variable as the dependent variable, and not count data or categorical classifier data. Viewing the summary of each regression model and looking at the coefficient table for the p-values and the significance codes will define whether these positively correlated variables significantly increase life expectancy.

Analysis of Findings

After performing the methodology explained above, I found that within the datasets, some variables that were positively correlated were not those that I first expected. Running the correlation tests on the three datasets yielded the following most positively correlated variables to the life expectancy variable present in each dataset:

life_expectancy (Top 3 positively correlated variables)

- Polio (Percentage of coverage of immunization for one-year old infants)
- Diphtheria (Percentage of coverage of immunization for one-year old infants)
- Schooling (Average years that people aged 25+ spent in formal education)

life_two (Top 3 positively correlated variables)

- Schooling (Average number of years of schooling received by the population)
- BMI (Average Body Mass Index (BMI) of the population)
- Income Composition of Resources (Composite index reflecting the income distribution and access to resources in the country)

healthy_life (Top 2 positively correlated variables)

This dataset only had two variables that were positively correlated, while all of the other variables were negatively correlated, so the analysis showed that only two variables are available for use in the multiple linear regression model.

- Happiness Levels (Country) (Rated on a scale of 1 through 10)
- Number of Take Out Places (City) (Total number of foodservice locations that have take-out services)

Once these variables were found to be the most positively correlated predictors among all three of the datasets, the multiple regression models were created. Each multiple regression model had a small range of residuals, showing that each model was a good fit for its corresponding dataset. The residual standard error for the regression models ranged between 5.8 and 3.7, backing the claim that the models were all good fits for the data. This fact is further justified by the adjusted R-squared values ranging from 0.61 to 0.53, which is favorable as adjusted R-squared operates on a scale of 0 to 1. Looking at the p-values and significance codes, every variable in each model is considered significant in the increase of the life expectancy outcome variables present except for the Number of Take Out Places (City) variable present in the healthy_life dataset. The model code and output below will support this assessment.

```
# Correlation matrix for life_expectancy dataset

numeric_vars1 <- sapply(life_expectancy, is.numeric)

numeric_data1 <- life_expectancy[, numeric_vars1]

correlation_matrix1 <- cor(numeric_data1)

correlation_matrix1
```

```
##              Year Infant_deaths Under_five_deaths
## Year          1.0000000000 -0.17240170 -0.176392621
## Infant_deaths -0.1724016970  1.00000000  0.985651346
## Under_five_deaths -0.1763926213  0.98565135  1.000000000
## Adult_mortality -0.1586595781  0.79466086  0.802361123
## Alcohol_consumption -0.0006105222 -0.45452615 -0.409367397
## Hepatitis_B      0.1768240714 -0.51256224 -0.507427407
## Measles          0.0859447214 -0.52628201 -0.512971742
## BMI              0.1614224541 -0.66198827 -0.665255042
## Polio            0.1398583960 -0.74079046 -0.742983474
## Diphtheria       0.1451429275 -0.72187465 -0.725355032
## Incidents_HIV    -0.0817425731  0.34945826  0.369617726
## GDP_per_capita   0.0409981721 -0.51228611 -0.469681668
## Population_mln   0.0151576184  0.00762199 -0.005234231
## Thinness_ten_nineteen_years -0.0449005325  0.49119174  0.466978458
## Thinness_five_nine_years -0.0480377469  0.47763934  0.450755699
## Schooling        0.1505393684 -0.78851253 -0.773195983
## Economy_status_Developed 0.0000000000 -0.47586620 -0.427134181
## Economy_status_Developing 0.0000000000  0.47586620  0.427134181
## Life_expectancy  0.1743589433 -0.92003192 -0.920419134
##              Adult_mortality Alcohol_consumption Hepatitis_B
## Year          -0.15865958 -0.0006105222  0.17682407
## Infant_deaths  0.79466086 -0.4545261472 -0.51256224
## Under_five_deaths 0.80236112 -0.4093673971 -0.50742741
## Adult_mortality  1.00000000 -0.2447937555 -0.34488221
## Alcohol_consumption -0.24479376  1.0000000000  0.16843582
## Hepatitis_B     -0.34488221  0.1684358238  1.00000000
## Measles         -0.41615254  0.3186029309  0.42916779
## BMI             -0.52286551  0.2840319455  0.34542091
## Polio           -0.52422554  0.3019262324  0.72434526
## Diphtheria      -0.51380270  0.2990159210  0.76178009
## Incidents_HIV    0.69911938 -0.0341180147 -0.07578195
## GDP_per_capita  -0.51012141  0.4439659537  0.15937504
## Population_mln  -0.05384768 -0.0391186595 -0.08239640
```


## Thinness_ten_nineteen_years	0.38214030	-0.4463661789	-0.20845350
## Thinness_five_nine_years	0.37979229	-0.4330297154	-0.21379442
## Schooling	-0.58103548	0.6157280403	0.34764345
## Economy_status_Developed	-0.42937477	0.6703660889	0.11353405
## Economy_status_Developing	0.42937477	-0.6703660889	-0.11353405
## Life_expectancy	-0.94536036	0.3991591076	0.41780443
##	Measles	BMI	Polio Diphtheria
## Year	0.08594472	0.1614225	0.13985840 0.14514293
## Infant_deaths	-0.52628201	-0.6619883	-0.74079046 -0.72187465
## Under_five_deaths	-0.51297174	-0.6652550	-0.74298347 -0.72535503
## Adult_mortality	-0.41615254	-0.5228655	-0.52422554 -0.51380270
## Alcohol_consumption	0.31860293	0.2840319	0.30192623 0.29901592
## Hepatitis_B	0.42916779	0.3454209	0.72434526 0.76178009
## Measles	1.00000000	0.4163214	0.51409629 0.49405877
## BMI	0.41632141	1.00000000	0.45720604 0.42650090
## Polio	0.51409629	0.4572060	1.00000000 0.95317790
## Diphtheria	0.49405877	0.4265009	0.95317790 1.00000000
## Incidents_HIV	-0.15058000	-0.1611421	-0.14795220 -0.14693191
## GDP_per_capita	0.31372372	0.3361796	0.31378567 0.31332094
## Population_mln	-0.09822189	-0.1664820	-0.03348589 -0.02733598
## Thinness_ten_nineteen_years	-0.34070533	-0.5964833	-0.31268545 -0.30446625
## Thinness_five_nine_years	-0.36696995	-0.5991122	-0.30699811 -0.29559745
## Schooling	0.49839128	0.6354752	0.55276511 0.53562097
## Economy_status_Developed	0.29869329	0.2432870	0.28326012 0.28941718
## Economy_status_Developing	-0.29869329	-0.2432870	-0.28326012 -0.28941718
## Life_expectancy	0.49001859	0.5984233	0.64121746 0.62754139
##	Incidents_HIV	GDP_per_capita	Population_mln
## Year	-0.08174257	0.04099817	0.015157618
## Infant_deaths	0.34945826	-0.51228611	0.007621990
## Under_five_deaths	0.36961773	-0.46968167	-0.005234231
## Adult_mortality	0.69911938	-0.51012141	-0.053847680
## Alcohol_consumption	-0.03411801	0.44396595	-0.039118659
## Hepatitis_B	-0.07578195	0.15937504	-0.082396398
## Measles	-0.15058000	0.31372372	-0.098221891
## BMI	-0.16114208	0.33617960	-0.166482004
## Polio	-0.14795220	0.31378567	-0.033485888
## Diphtheria	-0.14693191	0.31332094	-0.027335977
## Incidents_HIV	1.00000000	-0.16958972	-0.058039708
## GDP_per_capita	-0.16958972	1.00000000	-0.040838867
## Population_mln	-0.05803971	-0.04083887	1.000000000
## Thinness_ten_nineteen_years	0.18876454	-0.37526974	0.256322009
## Thinness_five_nine_years	0.19384734	-0.38103211	0.258485836
## Schooling	-0.20124620	0.58062592	-0.033561816
## Economy_status_Developed	-0.17563524	0.66754691	-0.035301833
## Economy_status_Developing	0.17563524	-0.66754691	0.035301833
## Life_expectancy	-0.55302746	0.58308972	0.026297880
##	Thinness_ten_nineteen_years		
## Year		-0.04490053	
## Infant_deaths		0.49119174	
## Under_five_deaths		0.46697846	
## Adult_mortality		0.38214030	
## Alcohol_consumption		-0.44636618	
## Hepatitis_B		-0.20845350	
## Measles		-0.34070533	

## BMI	-0.59648328	
## Polio	-0.31268545	
## Diphtheria	-0.30446625	
## Incidents_HIV	0.18876454	
## GDP_per_capita	-0.37526974	
## Population_mln	0.25632201	
## Thinness_ten_nineteen_years	1.00000000	
## Thinness_five_nine_years	0.93875710	
## Schooling	-0.57148516	
## Economy_status_Developed	-0.41609766	
## Economy_status_Developing	0.41609766	
## Life_expectancy	-0.46782450	
##	Thinness_five_nine_years	Schooling
## Year	-0.04803775	0.15053937
## Infant_deaths	0.47763934	-0.78851253
## Under_five_deaths	0.45075570	-0.77319598
## Adult_mortality	0.37979229	-0.58103548
## Alcohol_consumption	-0.43302972	0.61572804
## Hepatitis_B	-0.21379442	0.34764345
## Measles	-0.36696995	0.49839128
## BMI	-0.59911219	0.63547517
## Polio	-0.30699811	0.55276511
## Diphtheria	-0.29559745	0.53562097
## Incidents_HIV	0.19384734	-0.20124620
## GDP_per_capita	-0.38103211	0.58062592
## Population_mln	0.25848584	-0.03356182
## Thinness_ten_nineteen_years	0.93875710	-0.57148516
## Thinness_five_nine_years	1.00000000	-0.55137635
## Schooling	-0.55137635	1.00000000
## Economy_status_Developed	-0.41486734	0.59943940
## Economy_status_Developing	0.41486734	-0.59943940
## Life_expectancy	-0.45816623	0.73248447
##	Economy_status_Developed	Economy_status_Developing
## Year	0.00000000	0.00000000
## Infant_deaths	-0.47586620	0.47586620
## Under_five_deaths	-0.42713418	0.42713418
## Adult_mortality	-0.42937477	0.42937477
## Alcohol_consumption	0.67036609	-0.67036609
## Hepatitis_B	0.11353405	-0.11353405
## Measles	0.29869329	-0.29869329
## BMI	0.24328705	-0.24328705
## Polio	0.28326012	-0.28326012
## Diphtheria	0.28941718	-0.28941718
## Incidents_HIV	-0.17563524	0.17563524
## GDP_per_capita	0.66754691	-0.66754691
## Population_mln	-0.03530183	0.03530183
## Thinness_ten_nineteen_years	-0.41609766	0.41609766
## Thinness_five_nine_years	-0.41486734	0.41486734
## Schooling	0.59943940	-0.59943940
## Economy_status_Developed	1.00000000	-1.00000000
## Economy_status_Developing	-1.00000000	1.00000000
## Life_expectancy	0.52379098	-0.52379098
##	Life_expectancy	
## Year	0.17435894	

```
## Infant_deaths -0.92003192
## Under_five_deaths -0.92041913
## Adult_mortality -0.94536036
## Alcohol_consumption 0.39915911
## Hepatitis_B 0.41780443
## Measles 0.49001859
## BMI 0.59842332
## Polio 0.64121746
## Diphtheria 0.62754139
## Incidents_HIV -0.55302746
## GDP_per_capita 0.58308972
## Population_mln 0.02629788
## Thinness_ten_nineteen_years -0.46782450
## Thinness_five_nine_years -0.45816623
## Schooling 0.73248447
## Economy_status_Developed 0.52379098
## Economy_status_Developing -0.52379098
## Life_expectancy 1.00000000
```

```
# Correlation matrix for life_two dataset
```

```
numeric_vars2 <- sapply(life_two, is.numeric)

numeric_data2 <- life_two[, numeric_vars2]

correlation_matrix2 <- cor(numeric_data2)

correlation_matrix2
```

```
##
## Year 1.000000000 0.05077103 -0.037091782
## Life expectancy 0.050771035 1.00000000 -0.702523062
## Adult Mortality -0.037091782 -0.70252306 1.000000000
## infant deaths 0.008029128 -0.16907380 0.042450237
## Alcohol -0.113364764 0.40271832 -0.175535086
## percentage expenditure 0.069553468 0.40963082 -0.237609890
## Hepatitis B 0.114897092 0.19993528 -0.105225443
## Measles -0.053822046 -0.06888122 -0.003966685
## BMI 0.005739061 0.54204159 -0.351542478
## under-five deaths 0.010478594 -0.19226530 0.060365026
## Polio -0.016698803 0.32729440 -0.199853000
## Total expenditure 0.059492777 0.17471764 -0.085226535
## Diphtheria 0.029640586 0.34133123 -0.191428759
## HIV/AIDS -0.123404990 -0.59223629 0.550690745
## GDP 0.096421485 0.44132181 -0.255034733
## Population 0.012566893 -0.02230498 -0.015011838
## thinness 1-19 years 0.019756611 -0.45783819 0.272230044
## thinness 5-9 years 0.014122422 -0.45750829 0.286722882
## Income composition of resources 0.122891780 0.72108259 -0.442203288
## Schooling 0.088731787 0.72763003 -0.421170523
## infant deaths Alcohol
## Year 0.008029128 -0.11336476
## Life expectancy -0.169073804 0.40271832
## Adult Mortality 0.042450237 -0.17553509
```

## infant deaths	1.000000000	-0.10621692	
## Alcohol	-0.106216917	1.000000000	
## percentage expenditure	-0.090764632	0.41704736	
## Hepatitis B	-0.231768937	0.10988939	
## Measles	0.532679832	-0.05011023	
## BMI	-0.234425154	0.35339621	
## under-five deaths	0.996905622	-0.10108216	
## Polio	-0.156928805	0.24031453	
## Total expenditure	-0.146951117	0.21488509	
## Diphtheria	-0.161871004	0.24295143	
## HIV/AIDS	0.007711547	-0.02711264	
## GDP	-0.098092020	0.44343279	
## Population	0.671758310	-0.02888023	
## thinness 1-19 years	0.463415256	-0.40375499	
## thinness 5-9 years	0.461907925	-0.38620819	
## Income composition of resources	-0.134753863	0.56107433	
## Schooling	-0.214371900	0.61697481	
##	percentage expenditure	Hepatitis B	Measles
## Year	0.06955347	0.11489709	-0.053822046
## Life expectancy	0.40963082	0.19993528	-0.068881222
## Adult Mortality	-0.23760989	-0.10522544	-0.003966685
## infant deaths	-0.09076463	-0.23176894	0.532679832
## Alcohol	0.41704736	0.10988939	-0.050110235
## percentage expenditure	1.000000000	0.01676017	-0.063070789
## Hepatitis B	0.01676017	1.000000000	-0.124799993
## Measles	-0.06307079	-0.12479999	1.000000000
## BMI	0.24273824	0.14330179	-0.153245464
## under-five deaths	-0.09215806	-0.24076603	0.517505563
## Polio	0.12862605	0.46333080	-0.057850133
## Total expenditure	0.18387236	0.11332668	-0.113582738
## Diphtheria	0.13481324	0.58898993	-0.058605907
## HIV/AIDS	-0.09508499	-0.09480197	-0.003521854
## GDP	0.95929886	0.04184950	-0.064767590
## Population	-0.01679214	-0.12972265	0.321946377
## thinness 1-19 years	-0.25503460	-0.12940595	0.180641506
## thinness 5-9 years	-0.25563544	-0.13325099	0.174946217
## Income composition of resources	0.40216974	0.18492097	-0.058277256
## Schooling	0.42208845	0.21518159	-0.115660481
##	BMI	under-five deaths	Polio
## Year	0.005739061	0.01047859	-0.01669880
## Life expectancy	0.542041588	-0.19226530	0.32729440
## Adult Mortality	-0.351542478	0.06036503	-0.19985300
## infant deaths	-0.234425154	0.99690562	-0.15692881
## Alcohol	0.353396205	-0.10108216	0.24031453
## percentage expenditure	0.242738243	-0.09215806	0.12862605
## Hepatitis B	0.143301786	-0.24076603	0.46333080
## Measles	-0.153245464	0.51750556	-0.05785013
## BMI	1.000000000	-0.24213740	0.18626797
## under-five deaths	-0.242137398	1.000000000	-0.17116419
## Polio	0.186267965	-0.17116419	1.000000000
## Total expenditure	0.189468964	-0.14580310	0.11976798
## Diphtheria	0.176294503	-0.17844819	0.60924547
## HIV/AIDS	-0.210896746	0.01947593	-0.10788547
## GDP	0.266113973	-0.10033126	0.15680869

## Population	-0.081415982	0.65867969	-0.04538657
## thinness 1-19 years	-0.547017514	0.46478470	-0.16406959
## thinness 5-9 years	-0.554093981	0.46228938	-0.17448925
## Income composition of resources	0.510504831	-0.14809728	0.31468159
## Schooling	0.554843903	-0.22601262	0.35014660
##	Total expenditure	Diphtheria	HIV/AIDS
## Year	0.05949278	0.02964059	-0.123404990
## Life expectancy	0.17471764	0.34133123	-0.592236293
## Adult Mortality	-0.08522653	-0.19142876	0.550690745
## infant deaths	-0.14695112	-0.16187100	0.007711547
## Alcohol	0.21488509	0.24295143	-0.027112636
## percentage expenditure	0.18387236	0.13481324	-0.095084991
## Hepatitis B	0.11332668	0.58898993	-0.094801971
## Measles	-0.11358274	-0.05860591	-0.003521854
## BMI	0.18946896	0.17629450	-0.210896746
## under-five deaths	-0.14580310	-0.17844819	0.019475927
## Polio	0.11976798	0.60924547	-0.107885468
## Total expenditure	1.00000000	0.12991481	0.043100657
## Diphtheria	0.12991481	1.00000000	-0.117601074
## HIV/AIDS	0.04310066	-0.11760107	1.000000000
## GDP	0.18037347	0.15843774	-0.108080600
## Population	-0.07996224	-0.03989754	-0.027800562
## thinness 1-19 years	-0.20987232	-0.18724165	0.172591767
## thinness 5-9 years	-0.21786479	-0.18095238	0.183146727
## Income composition of resources	0.18365319	0.34326177	-0.248589855
## Schooling	0.24378345	0.35039793	-0.211840201
##	GDP	Population	thinness 1-19 years
## Year	0.09642148	0.012566893	0.01975661
## Life expectancy	0.44132181	-0.022304978	-0.45783819
## Adult Mortality	-0.25503473	-0.015011838	0.27223004
## infant deaths	-0.09809202	0.671758310	0.46341526
## Alcohol	0.44343279	-0.028880232	-0.40375499
## percentage expenditure	0.95929886	-0.016792141	-0.25503460
## Hepatitis B	0.04184950	-0.129722655	-0.12940595
## Measles	-0.06476759	0.321946377	0.18064151
## BMI	0.26611397	-0.081415982	-0.54701751
## under-five deaths	-0.10033126	0.658679691	0.46478470
## Polio	0.15680869	-0.045386572	-0.16406959
## Total expenditure	0.18037347	-0.079962237	-0.20987232
## Diphtheria	0.15843774	-0.039897537	-0.18724165
## HIV/AIDS	-0.10808060	-0.027800562	0.17259177
## GDP	1.00000000	-0.020368964	-0.27749835
## Population	-0.02036896	1.000000000	0.28252928
## thinness 1-19 years	-0.27749835	0.282529280	1.00000000
## thinness 5-9 years	-0.27795855	0.277913374	0.92791344
## Income composition of resources	0.44685551	-0.008132466	-0.45367885
## Schooling	0.46794697	-0.040312419	-0.49119921
##	thinness 5-9 years		
## Year	0.01412242		
## Life expectancy	-0.45750829		
## Adult Mortality	0.28672288		
## infant deaths	0.46190792		
## Alcohol	-0.38620819		
## percentage expenditure	-0.25563544		

## Hepatitis B	-0.13325099	
## Measles	0.17494622	
## BMI	-0.55409398	
## under-five deaths	0.46228938	
## Polio	-0.17448925	
## Total expenditure	-0.21786479	
## Diphtheria	-0.18095238	
## HIV/AIDS	0.18314673	
## GDP	-0.27795855	
## Population	0.27791337	
## thinness 1-19 years	0.92791344	
## thinness 5-9 years	1.00000000	
## Income composition of resources	-0.43848372	
## Schooling	-0.47248203	
##	Income composition of resources	Schooling
## Year	0.122891780	0.08873179
## Life expectancy	0.721082593	0.72763003
## Adult Mortality	-0.442203288	-0.42117052
## infant deaths	-0.134753863	-0.21437190
## Alcohol	0.561074332	0.61697481
## percentage expenditure	0.402169736	0.42208845
## Hepatitis B	0.184920970	0.21518159
## Measles	-0.058277256	-0.11566048
## BMI	0.510504831	0.55484390
## under-five deaths	-0.148097276	-0.22601262
## Polio	0.314681594	0.35014660
## Total expenditure	0.183653190	0.24378345
## Diphtheria	0.343261772	0.35039793
## HIV/AIDS	-0.248589855	-0.21184020
## GDP	0.446855511	0.46794697
## Population	-0.008132466	-0.04031242
## thinness 1-19 years	-0.453678854	-0.49119921
## thinness 5-9 years	-0.438483721	-0.47248203
## Income composition of resources	1.000000000	0.78474058
## Schooling	0.784740581	1.00000000

Correlation matrix for healthy_life dataset

```

numeric_vars3 <- sapply(healthy_life, is.numeric)

numeric_data3 <- healthy_life[, numeric_vars3]

correlation_matrix3 <- cor(numeric_data3)

correlation_matrix3

```

##	Rank	Sunshine hours(City)
## Rank	1.000000000	0.007936892
## Sunshine hours(City)	0.007936892	1.000000000
## Life expectancy(years) (Country)	-0.320034317	-0.353387651
## Pollution(Index score) (City)	0.419829163	0.383465687
## Annual avg. hours worked	0.463707935	0.443595612
## Happiness levels(Country)	-0.231969628	-0.328979484
## Outdoor activities(City)	-0.076971885	0.126767333

```

## Number of take out places(City)    0.363058309    -0.210539124
##                                     Life expectancy(years) (Country)
## Rank                               -0.32003432
## Sunshine hours(City)                -0.35338765
## Life expectancy(years) (Country)    1.00000000
## Pollution(Index score) (City)       -0.55271478
## Annual avg. hours worked            -0.21295228
## Happiness levels(Country)           0.72458708
## Outdoor activities(City)            -0.04486383
## Number of take out places(City)     0.12816279
##                                     Pollution(Index score) (City)
## Rank                               0.4198292
## Sunshine hours(City)                0.3834657
## Life expectancy(years) (Country)    -0.5527148
## Pollution(Index score) (City)       1.0000000
## Annual avg. hours worked            0.3954221
## Happiness levels(Country)           -0.7629349
## Outdoor activities(City)            0.2368367
## Number of take out places(City)     0.1862806
##                                     Annual avg. hours worked
## Rank                               0.46370793
## Sunshine hours(City)                0.44359561
## Life expectancy(years) (Country)    -0.21295228
## Pollution(Index score) (City)       0.39542209
## Annual avg. hours worked            1.00000000
## Happiness levels(Country)           -0.29064727
## Outdoor activities(City)            0.07875265
## Number of take out places(City)     -0.03354630
##                                     Happiness levels(Country)
## Rank                               -0.23196963
## Sunshine hours(City)                -0.32897948
## Life expectancy(years) (Country)    0.72458708
## Pollution(Index score) (City)       -0.76293487
## Annual avg. hours worked            -0.29064727
## Happiness levels(Country)           1.00000000
## Outdoor activities(City)            -0.13761226
## Number of take out places(City)     0.03311558
##                                     Outdoor activities(City)
## Rank                               -0.07697188
## Sunshine hours(City)                0.12676733
## Life expectancy(years) (Country)    -0.04486383
## Pollution(Index score) (City)       0.23683670
## Annual avg. hours worked            0.07875265
## Happiness levels(Country)           -0.13761226
## Outdoor activities(City)            1.00000000
## Number of take out places(City)     0.52819990
##                                     Number of take out places(City)
## Rank                               0.36305831
## Sunshine hours(City)                -0.21053912
## Life expectancy(years) (Country)    0.12816279
## Pollution(Index score) (City)       0.18628062
## Annual avg. hours worked            -0.03354630
## Happiness levels(Country)           0.03311558
## Outdoor activities(City)            0.52819990

```

```
## Number of take out places(City) 1.00000000
```

```
# life_expectancy multiple regression model
```

```
life_expectancy_model <- lm(Life_expectancy ~ Schooling + Diphtheria +  
                             Polio, data = life_expectancy)
```

```
summary(life_expectancy_model)
```

```
##  
## Call:  
## lm(formula = Life_expectancy ~ Schooling + Diphtheria + Polio,  
##     data = life_expectancy)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -23.079  -3.389   1.437   4.174  13.612   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept) 38.06083    0.64127  59.352 < 2e-16 ***  
## Schooling   1.60967    0.04113  39.136 < 2e-16 ***  
## Diphtheria  0.07683    0.02314   3.320 0.000911 ***  
## Polio       0.13736    0.02415   5.687 1.43e-08 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 5.813 on 2860 degrees of freedom  
## Multiple R-squared:  0.6184, Adjusted R-squared:  0.618  
## F-statistic: 1545 on 3 and 2860 DF, p-value: < 2.2e-16
```

```
# Life_two multiple regression model
```

```
names(life_two)[names(life_two) == "Life expectancy"] <- "Life_expectancy"
```

```
names(life_two)[names(life_two) == "Income composition of resources"] <-  
  "income_comp_resources"
```

```
life_two_model <- lm(Life_expectancy ~ Schooling + BMI +  
                     income_comp_resources, data = life_two)
```

```
summary(life_two_model)
```

```
##  
## Call:  
## lm(formula = Life_expectancy ~ Schooling + BMI + income_comp_resources,  
##     data = life_two)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -21.2827  -2.5569   0.6271   3.3629  18.5085   
##  
## Coefficients:
```



```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      42.074540   0.611428  68.814   <2e-16 ***
## Schooling         1.121836   0.082127  13.660   <2e-16 ***
## BMI               0.071637   0.008377   8.551   <2e-16 ***
## income_comp_resources 17.258779   1.213200  14.226   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.53 on 1645 degrees of freedom
## Multiple R-squared:  0.6056, Adjusted R-squared:  0.6049
## F-statistic: 842 on 3 and 1645 DF, p-value: < 2.2e-16
```

```
# healthy_life multiple regression model
```

```
names(healthy_life)[names(healthy_life) ==
                     "Life expectancy(years) (Country)"] <-
  "Life_expectancy_country"

names(healthy_life)[names(healthy_life) ==
                     "Happiness levels(Country)"] <- "happy_levels_country"

names(healthy_life)[names(healthy_life) ==
                     "Number of take out places(City)"] <- "take_out_city"

healthy_life_model <- lm(Life_expectancy_country ~ happy_levels_country +
                        take_out_city, data = healthy_life)

summary(healthy_life_model)
```

```
##
## Call:
## lm(formula = Life_expectancy_country ~ happy_levels_country +
##     take_out_city, data = healthy_life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.225  -1.287   0.351   1.636   7.566
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.277e+01  3.735e+00  14.127 < 2e-16 ***
## happy_levels_country 3.859e+00  5.697e-01   6.774 3.44e-08 ***
## take_out_city     3.983e-04  4.066e-04   0.980  0.333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.701 on 41 degrees of freedom
## Multiple R-squared:  0.5359, Adjusted R-squared:  0.5132
## F-statistic: 23.67 on 2 and 41 DF, p-value: 1.465e-07
```

Implications of Findings

Now that we have visualized the findings of our data story relating to life expectancy, we can see that a lot of the variables that we have found to be significant in the increase of life expectancy are relevant to the medical industry and the education system and less so to normal civilians working in less technical fields. Shifting the focus to these industries regarding the telling of the data story with these results would give them more backing to perpetuate the strong encouragement of some of their practices, like polio and diphtheria immunization and a stronger emphasis on college education after compulsory schooling. Businesses directly or even indirectly related to these fields, such as those in the insurance industry and the banking industry also would benefit from these results. While the insurance companies benefiting from a study such as this has already been fleshed out, the banking industry seems to be a stretch at first glance. However, if prospective college students are having trouble with monetary funds, banks and even governments could benefit by offering those prospects student loans with the message that an increased number of years of education plays a positive role in the longevity of one's life. With this validated information from the data, an increased number of education loans would be seen and the banks would see an increase in their interest revenue in the long run. The negative connotations of these results is that without expert help and guidance in understanding the true meaning behind them, most individuals would only focus on the metrics they could easily understand, potentially leading to missed opportunities in people's lives to increase their life expectancy. In my view, these results (if not carefully and adeptly shared) display a classic case of consumer against business, as businesses have more knowledge and technical prowess than consumers to utilize them in a beneficial way.

Limitations of Findings

While seeing the results and deeming them to be rather promising, some of the variables are definitely in need of further explanation to reach the most amount of people who could potentially benefit from these results. Metrics like happiness levels are rather arbitrary to the individual and do very little for someone trying to incorporate some of these variables into their life in an attempt to increase their longevity. Also, all of the questions that have been posed at the beginning have been answered save for the fifth and final question, which at this point in time seems to be the most important question of all. Looking at the question again, I'd like to rephrase it as such:

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

Even though 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 has answered is which most positively correlated variables are significant to the increase of life expectancy in each dataset. 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. It would be extremely difficult to realize an answer for this question without further collaboration from another industry. Overall, while data can reveal much more than meets the eye, not every single answer is held in the data.

Concluding Remarks

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. As the results have shown, the models created are good fits for the data, and almost every one of the variables that were found to be the most positively correlated were seen to be significant to the increase in life expectancy.

While not every question was answered, we are left with more information than what we first knew and models that were successful in determining factors that are significant to one's increased longevity.