**Research Question: Using a predictor variable, can we predict the average Canadian Life Expectancy?**

**Life Expectancy**: <https://www.kaggle.com/kumarajarshi/life-expectancy-who> (could control for predictor variables like Developing status, adult mortality, hepatitis to find mean life expectancy of different countries)

**2000-2015**

**1. Life expectancy vs Infant Deaths**

**2. Life expectancy vs Alcohol**

**3. Life expectancy vs Health Expenditure Percentage**

**4. Life expectancy vs BMI**

**5. Life expectancy vs Schooling**

**Instructions**

Each group is expected to prepare a 1 page (max 500 words) written proposal that identifies the dataset they plan to work on, as well as the question they would like to answer using that dataset for their group project. The proposal should be done in a Jupyter notebook, and then submitted **both**as an .html file (File -> Download As -> HTML) and **an .ipynb file that is reproducible (i.e. works and runs without any additional files.)**

**Proposal Draft**

**Title**

**Research Question: Using a predictor variable, can we predict the average Canadian Life Expectancy?**

**Introduction**

* + **Provide some relevant background information on the topic so that someone unfamiliar with it will be prepared to understand the rest of your proposal**
  + **Clearly state the question you will try to answer with your project**
  + **Identify and describe the dataset that will be used to answer the**
  + **(Also explain why we chose the 6 predictor variables that we chose)**

**How long does people live is one of the necessary indicators of the living standards of a nation. As economics developed, the average life expectancy has increased for most countries, however, we interested in the key factors that contribute to the increasing trend. This research examines the possible driving factors of average life expectancy by decomposing the effects of each parameter. Using the latest data from WHO, we retrieve the nations’ panel health data, including infants death rate, usage of alcohol, BMI, age of schooling and consumption for all countries form 2000 to 2015. By applying regression analysis, we expect to find to correlation between life expectancy and these factors for shed some light on predicting future life expectancy in Canada.(Hongwei)**

**Preliminary data analysis(check R document on github, and interpret them)**

* + Demonstrate that the dataset can be read from the web into R
  + Clean and wrangle your data into a tidy format
  + Using only training data, summarize the data in at least one **table** (this is exploratory data analysis). An example of a useful table could be one that reports the number of observations in each class, the means of the predictor variables you plan to use in your analysis and how many rows have missing data.
  + Using only training data, visualize the data with at least one plot relevant to the analysis you plan to do (this is exploratory data analysis). An example of a useful visualization could be one that compares the distributions of each of the predictor variables you plan to use in your analysis

**Methods**

* + **Explain how you will conduct either your data analysis and which variables/columns you will use. *Note - you do not need to use all variables/columns that exist in the raw data set. In fact, that's often not a good idea. For each variable think: is this a useful variable for prediction?***
  + **Describe at least one way that you will visualize the results**

**Expected outcomes and significance**

* + **What do you expect to find?**
  + **What impact could such findings have?**
  + **What future questions could this lead to?**

**====================================**

**Predictor Variables We Chose:**

**Life Expectancy**(in age)

**Infant Deaths** (Number of Infant Deaths per 1000 population)**,**

**Alcohol** (recorded per capita (15+) consumption (in litres of pure alcohol))

**BMI (**Average Body Mass Index of entire population)**,**

**Percentage Expenditure** (Expenditure on health as a percentage of Gross Domestic Product per capita(%))

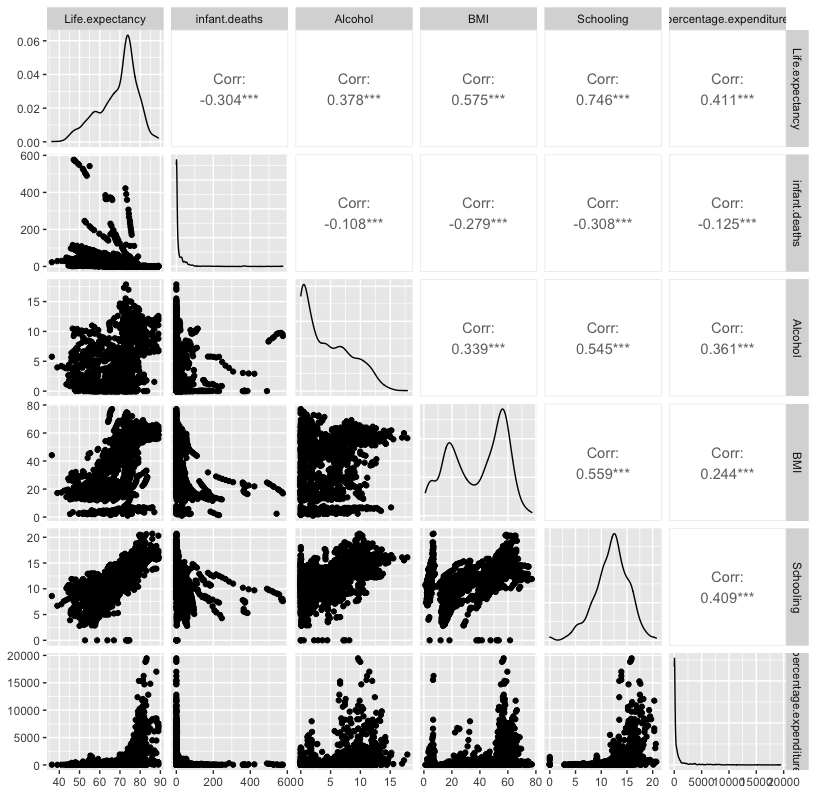
**Schooling**(Number of years of Schooling(years))

**Preliminary data analysis:**

* Our data set is the Life Expectancy data set (WHO) kaggle data set ([here](https://www.kaggle.com/kumarajarshi/life-expectancy-who))
* **Removing Missing data, we remove the whole observation row**
* **we eliminate data from India because the data of "infant death" in India does not make sense(they are over 1000)**

|  |  |
| --- | --- |
| Total number of observations | 2938 |
| Rows including missing data | 363 |
|  | |
| variables | mean |
| **Life Expectancy**(in age) | 69.37 |
| **Infant Deaths** | 21.58 |
| **Alcohol** | 4.54 |
| **BMI** | 38.23 |
| **Percentage Expenditure** | 847.65 |
| **Schooling** | 12.05 |

**Preliminary Data Visualization:**

****

(far left column shows predictors vs Life Expectancy )

(diagonal shows the distribution of the variables)

**Outline of the Methods:**

The method that we using to answer the research question is:

* After selecting for the 6 chosen predictor variables and filtering NAs, split the data into training and testing data(R file line 28)
* Make a plot of all the 6 predictor variables versus Life expectancy (See Preliminary data visualization above)
* Determine which predictor variable has the strongest linear fit with life expectancy: this will be the designated predictor variable
* Make a linear regression model with this predictor model and Life Expectancy (using the training data)
* Look up the average Canadian statistic for the predictor (like average BMI or average Schooling), and predict the resultant life expectancy
* Use the model built on the training data and predict the test data life expectancy, and compare with the actual test data life expectancy.

**Final data visualization:**

Scatterplot of the final 1-predictor model with the linear regression line shown as a solid line.

(i.e. Life Expectancy v.s. Schooling with the regression line)