Project 1 Summary

Team “Los Cracks”

Statistical Correlation of Life Expectancies and Possible Causes

Members:

Alan Vilchis

Aldo G. Sancininea

Elizabeth Garcia

Miguel A. Bonachea

Pablo Barrios

As our first project, we decided to ask the following question:

What characteristics of countries have the greatest impact on their life expectancy?

So, our project was to discover correlations between various characteristics of countries around the world, and their average life expectancy.

The first step in our analysis was to determine the different characteristics or behaviors that might influence life expectancy.

We examined the relationship between the following trends:

Body Mass Index

Tobacco Consumption

Fertilizer Application Rates per Country

Sugar Consumption

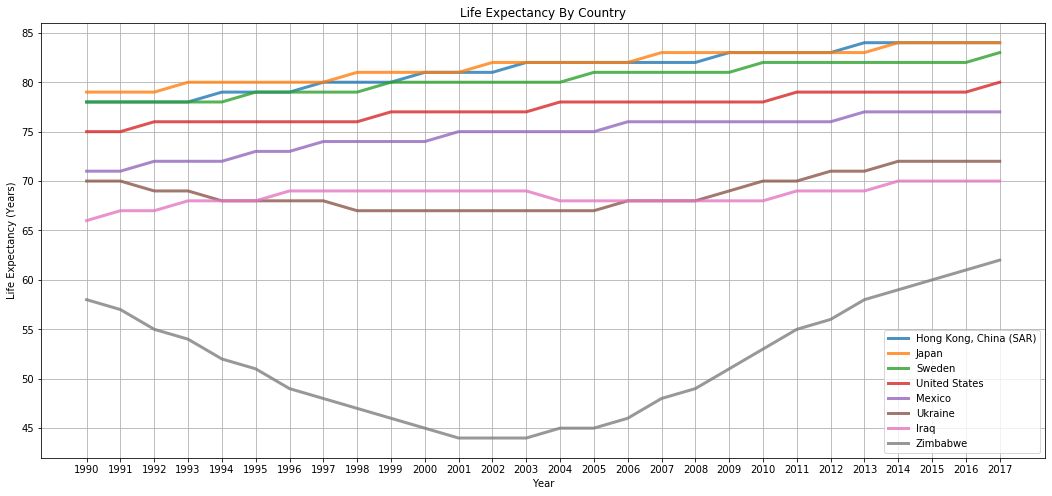
Alcohol Consumption per Capita

Access to Healthy Drinking Water

CO2 Emissions

Urban Population %

Working Hours per Week



Once our data was cleaned up, we could start with some basic plots to search for patterns.

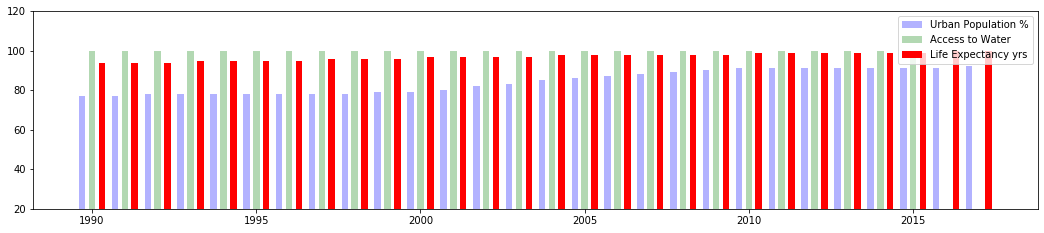
In this line plot we can clearly see some of the countries at the top and bottom range of life expectancies.

Hong Kong, Japan, and Sweden all have very high life expectancies (between 75 and 84 years)

United States and Mexico are in the intermediate range (between 70 and 80 years)

And Ukrane, Iraq & Zimbabwe are in the lower ranges (less than 73 years of age all the way down to about 44 years in 2002 for Zimbabwe)

We know Iraq and Ukrane have been negatively impacted by war and political unrest in the past few years, however it would be interesting to research the possible cause of such a drastic drop in life expectancy for Zimbabwe between 1990 and 2002, and also what changes took place to return the life expectancy back to around 60 years in 2017.



Our first attempt at trying to find causes and effects turned out to be a little disappointing and a bit misleading.

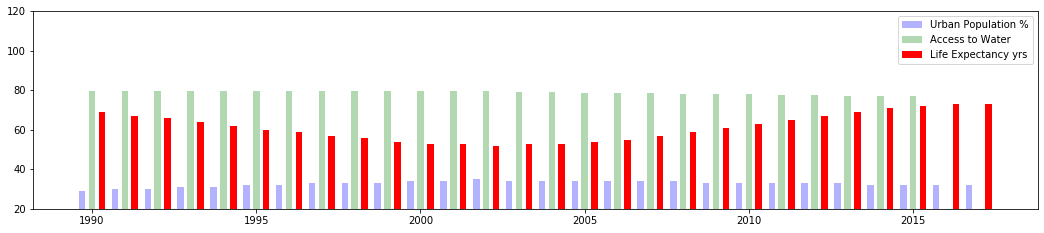
For example, this plot shows how life expectancy in Japan has increased slightly in the past 25 years,

At the same time, Urban population increased significantly, from around 79% to 99%, showing some probability that urban population % might increase life expectancy. Having a longer time frame might improve the accuracy of this assumption.

This is probably due to the ease of access to better medical care and specialized doctors who usually work in big city hospitals.

Access to clean water in Japan shows very little correlation to life expectancy, however the fact that it maintained at 100% for the entire time frame obviously had a positive effect on the population.

So these two factors in combination would most likely create some of the conditions necessary for raising Life Expectancy in the country.



Another example that shows very different results is Zimbabwe.

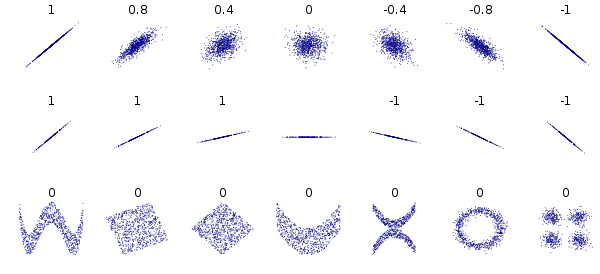
This chart shows access to water maintained at around 80% for the entire period and Urban population increasing slightly from approximately 30% to 39% in 2002 and then back down to 30% by 2015

In this case, neither Access to Water nor Urban Population % can explain the sudden drop in life expectancy between 1990 and 2002.

Furthermore, certain events may not have an immediate effect on life expectancy.

In other words, there may be a delayed effect from some change in behavior or events in the country.

For example, the negative effects of global warming are just now being felt after many years of uncontrolled pollution and burning of fossil fuels. This led us to the theory that we could find correlations from certain events in the past with the resulting effect of life expectancy in recent years. To do that we had to rely on more advanced statistics to create the following formulas and experiments.

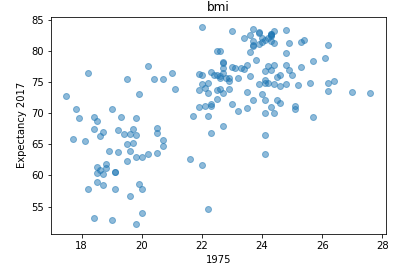


Four our analysis, we relied on two mathematical measurements: The p-value and the Pearson correlation index.

As we have seen in class before, the p-value is a measurement of how significant a relationship is. This tells us whether we can reject the null hypothesis or not. Since we are doing linear regressions for our analysis, the null hypothesis would be that there is a correlation between the variables we are measuring. Like we all know, if we found a p-value more than 0.05, then we would not be able to reject the null hypothesis, meaning that there does exist a relationship between our variables.

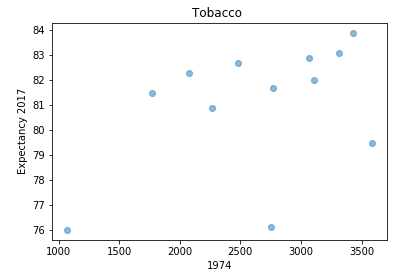
That said, just because there exists a relationship does not mean that we know how strong it is, less so whether the variables are positively or negatively correlated. This is where the Person correlation index comes into play. This coefficient is a measure of how strongly correlated are two variables on a scale that goes from -1 to 1. At either extreme of this scale, we can say that the two variables are perfectly correlated either positively (they both increase at the same time) or negatively (one increases as the other one decreases). A coefficient of zero would indicate that there is no linear correlation between the variables.

It is important to note, however, that this analysis only applies to linear correlations. Just because the correlation coefficient is zero, it does not necessarily mean that the variables are not related. As we can see on this example image, even if there is a clear pattern between the two variables, the correlation coefficient can still be zero. So, in the cases where there is no correlation found, a deeper analysis would be required to rule out other kinds of relationships (like exponential, sinusoidal, etc). Or, it could also mean that there are hidden variables that we are not taking into account

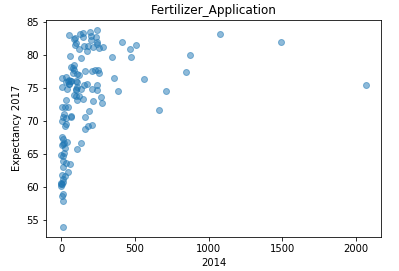


In this example, we could see the body mass index is always positively correlated with life expectancy. This means that the higher the body mass index of a country, the longer the people of that country live (although for this example it would be much more useful to have the percentage of overweight and obesity, to have results that make more sense ).

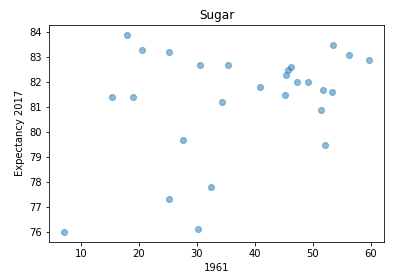
The greatest correlation occurs in 1975, which implies the effects of weight on life expectancy is observed more in the long term. Once this analysis is done for all the factors that decide to include, we can compare have a more immediate effect and specify the greater correlation.



There is a low linear relationship between life expectancy and tobacco consumption, even though we know that it has a significant effect due to studies and P-Value.  This could be due to the fact that we are analyzing the information by country and year instead of locally or regionally.



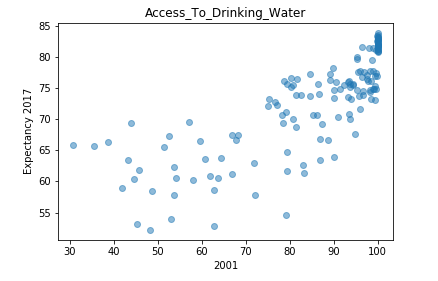
We can see no relation between fertilized application vs life expectancy, nevertheless, we see various outliers, which could mean that if we take those off, then a correlation could be seen.



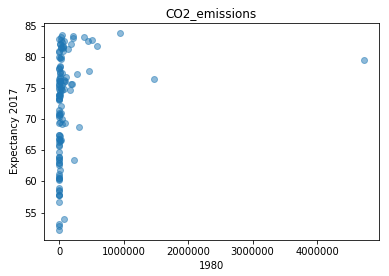
Here we can see a low level of correlation, and a .07 p value.  This means that sugar consumption has a significant effect, but there is no observable correlation.  We would need to compare the consumption with deaths caused by diabetes to have better information .



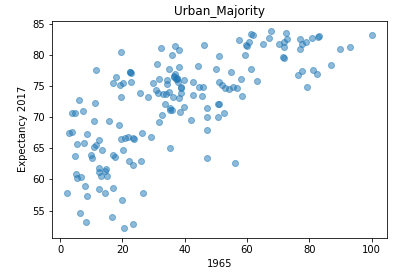
The data shows that there is a low level of correlation, but the p value exceeds .05. Which means that it has a significant effect on life expectancy. We believe that, since the data is by country alcohol consumption could be acting differently by country depending on the year or culture.  Hence, to obtain better info we should compare historical data of each country individually.



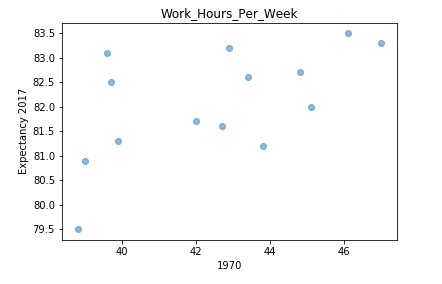
We can see that there is high linear correlation between access to drinking water and life expectancy. Nevertheless, a super low p value is obtained.  Statistically this says that it is not a relevant factor, but we believe it is. We would need to perform a local or regional study to have more info on the topic.



A low level of correlation and pi value are shown.  This means that statistically there is no relation. Nevertheless, we believe that the data could be wrong, that contamination has not had its effect, or since we are analyzing by country it is not very precise.  We should try this by city. We can also see, that outliers also cause a big displacement in the process.



A mid to high level of correlation are showed and a super high p value.  This means that the urban movement of the 60s has had a significant effect on life expectancy.



A mid level correlation is showed, but a low p-value for the year 1970. This means that there is a linear relationship between how working hours and life expectancy bhaved, but there it is not a significant factor.

In summary, we have applied the best practices in data science paradigms to detect trends in our data and find possible solutions.

We asked the questions “What characteristics of countries have the greatest impact on their life expectancy?”

And “What characteristics might influence life expectancy?”

We found data sources from trustworthy organizations such as OECD and Our World in Data

We defined our strategy to measure correlation using Coefficient and Probability Values in passed years for different metrics

We defined a standardized format that we wanted for our data, downloaded the CSV data and assembled our dataframes using Pandas

Once the format was established we cleaned out any Null or NaN values in rows and columns.

We then created plots of various types to analyze for trends, we did this both visually and mathematically.

We acknowledged the limitations or our results and our intention to show how long it takes for the factor to have an effect on life expectancy and the probability of the statistical correlation between the chosen metric and the end result.

We determined, based on our data that Access to Clean Water and Urbanization had the highest correlation and statistical probability of increasing global life expectancy.