MNGUNI ZULU

Analysis of Life Expectancy

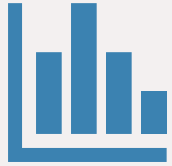
Mnguni.msimang@gmail.com

+49 0160 595 1998

TOOLS & LIBRARIES

Tools

- Python for data manipulation
- Compiled in Jupyter Notebook
- Pandas, Numpy, Seaborn, Matplotlib and Sklearn libraries
- supervised & unsupervised machine learning



Data

- open source data
- +2,000rows
- 17 numeric and categorical variables
- 119 countries

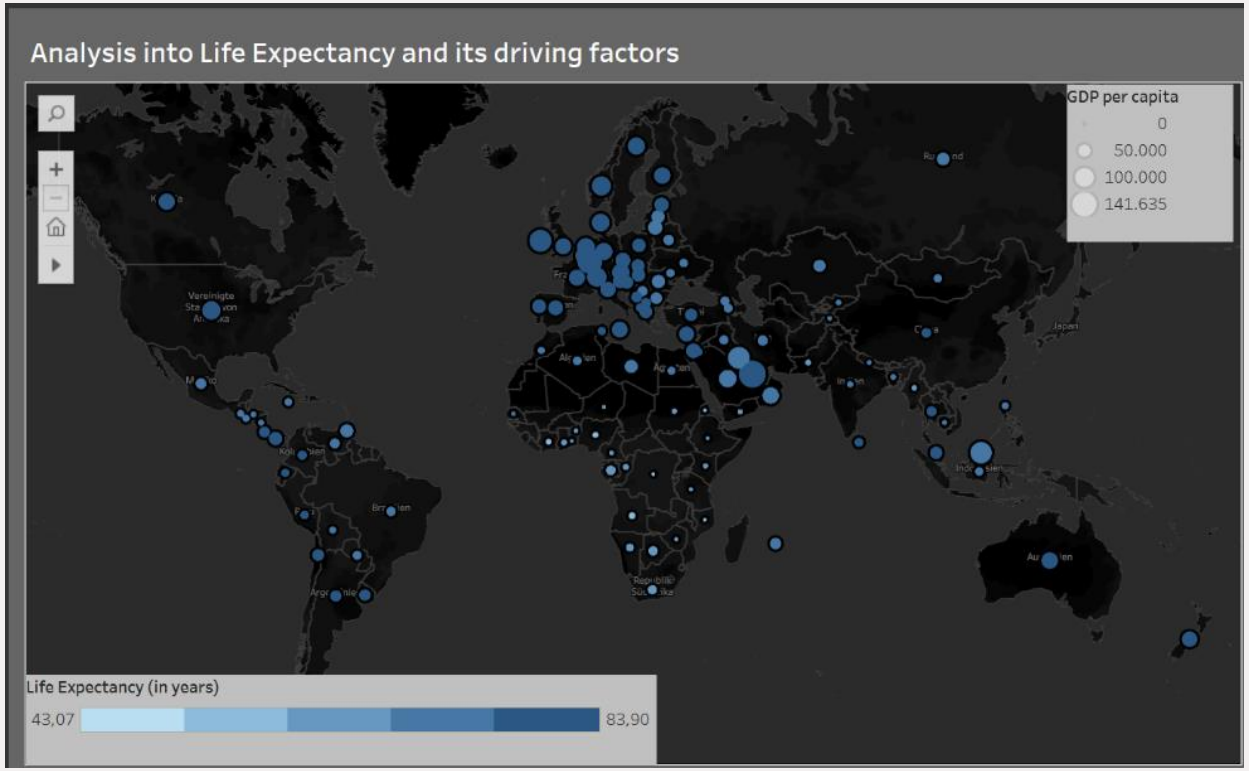
Recommendations

- Dashboard in Tableau
- Presentation of Notebooks

THE PROBLEM

Life expectancy is one of the most important and interesting measurements of human welfare. Although a long life does not guarantee a happy one, it does seem to be important to most people. There are so many variables which may or may not influence life expectancy, and the objective of this self-chosen project was to look at life expectancy in some 119 countries as well as 16 variables and their possible relationship with life expectancy.

Understanding what influences life expectancy, or at least the correlation between certain variables and life expectancy would be extremely useful for both internal and external development agencies.



STEP 1: INITIAL EXPLORATION OF DATA

4.2. Missing values

```
In [42]: # counting null values in dataset
         df_life.isnull().sum()
```

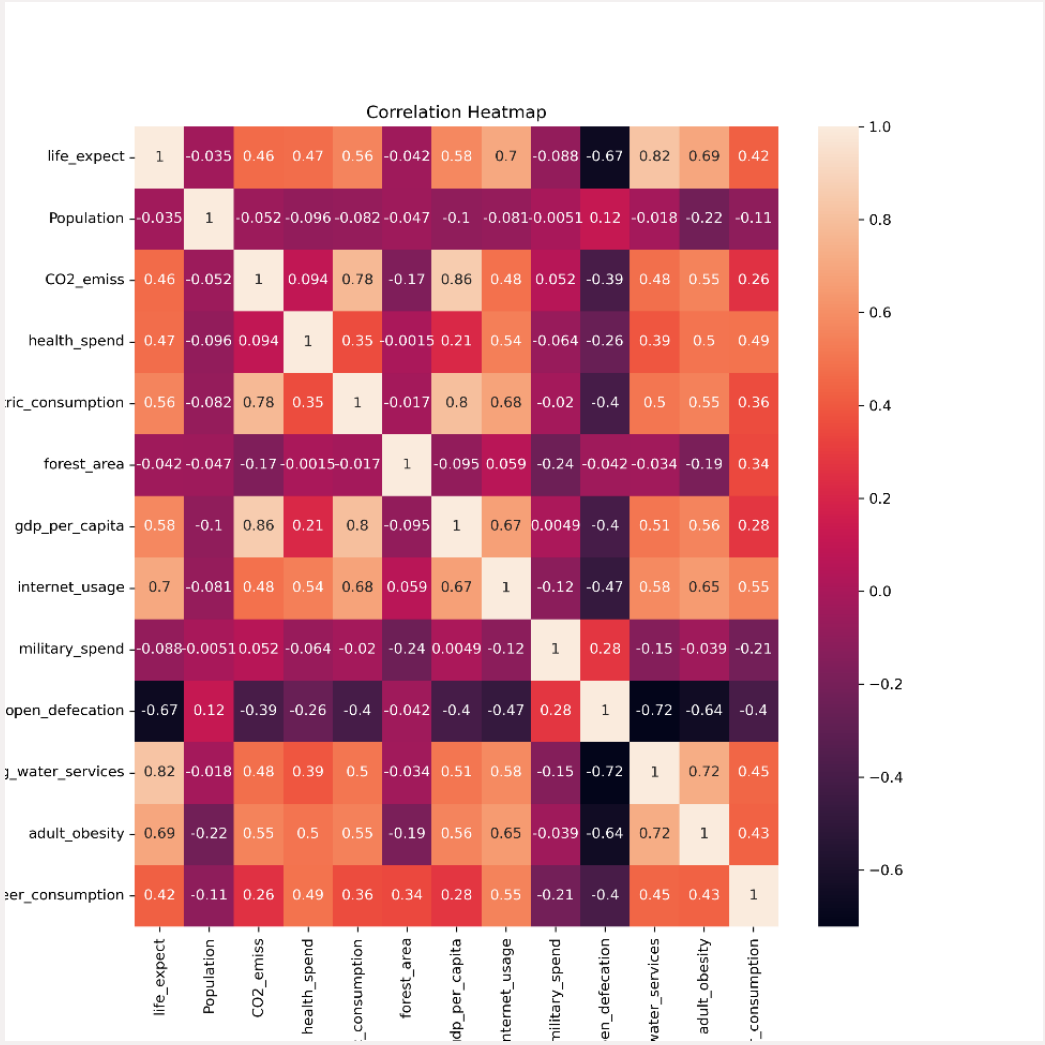
Out[42]:

Country	0
Year	0
Continent	0
Least Developed	0
life_expect	0
Population	0
CO2_emiss	0
health_spend	9
Electric_consumption	565
forest_area	0
gdp_per_capita	0
internet_usage	0
military_spend	0
open_defecation	0
drinking_water_services	0
adult_obesity	452
beer_consumption	117
dtype: int64	

The dataset was open-source data from the Worldbank. The dataset had a patchwork of missing values, which I uncovered during my initial exploratory data analysis. After identifying where these missing values were prevalent, I chose to limit the range of years from 2000 to 2015. Additionally, omitted 6 countries from the dataset, because they had far too many missing values. After this I remained with a dataset which that had less then 3% missing values. I used averages for each variable, for each country, to fill in the missing values respectively.

I also confirmed that were no duplicate rows in the dataset or mixed type values in the columns, which might skew the results of the analysis.

STEP 2: EXPLORING CORRELATIONS

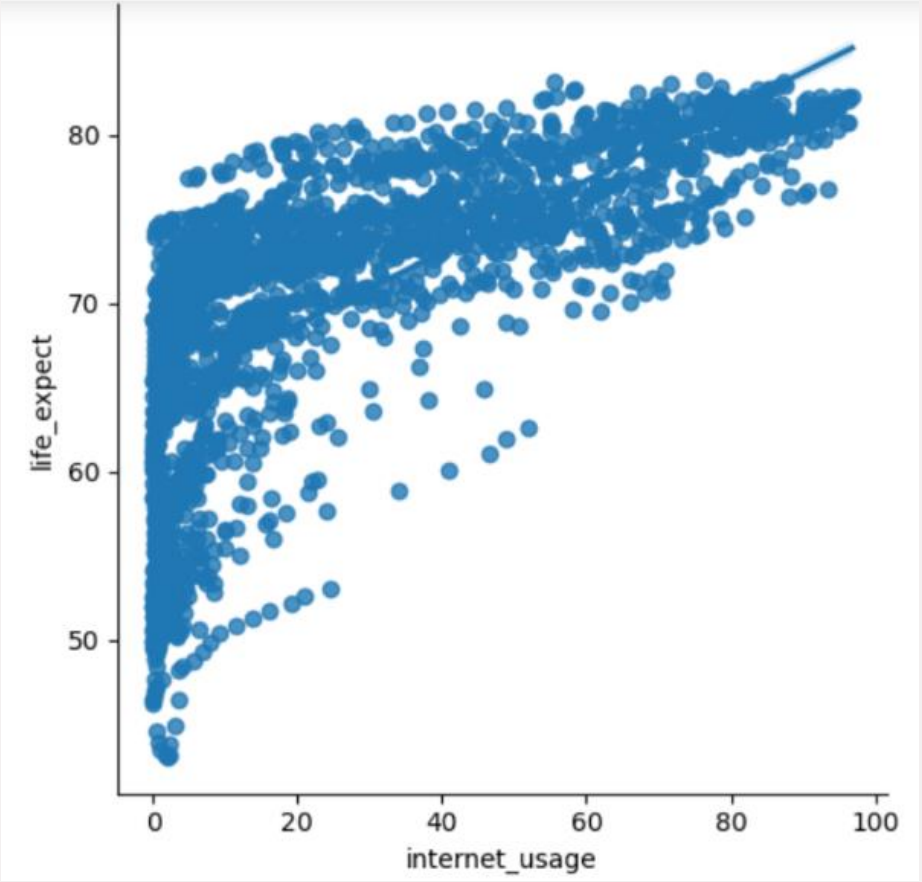


Now the data was explored for any interesting relationships between the 13 numerical variables. First a pair plot and heatmap were used to visualise the correlation across all the numeric variables, ranging from internet usage to the practice of open defecation.

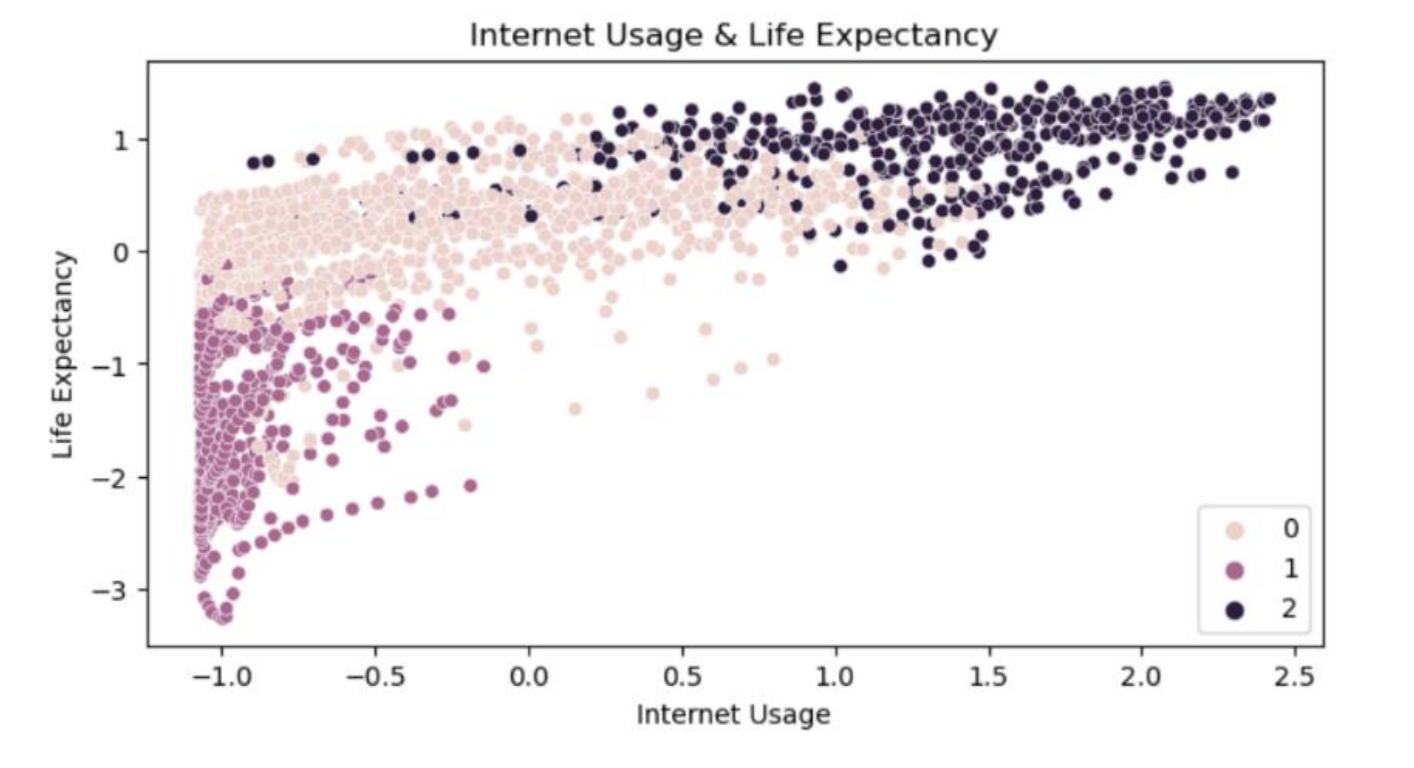
A group of variables which showed some correlation with life expectancy were singled out for further analysis. The three variables which were most interesting to me, were internet usage (as a % of population), access to basic drinking water services and adult obesity. Internet usage seemed like an interesting variable to me, and so I made this the focus of my hypothesis testing and further statistical analysis.

STEP 3 : SIMPLE REGRESSION

- **“Greater internet usage within a country leads to higher life expectancies”.** This was the hypothesis I set out to test using a simple regression. I used test and training subsets of data to determine whether one could use internet usage as a predictor of life expectancy. Unfortunately, the correlation between internet usage and life expectancy was positive but with a great deal of variance around the trendline. In other words the simple model was not successful. This lead me to the conclusion that clustering may provide deeper understanding of the relationship between these two variables.



STEP 5: CLUSTERING



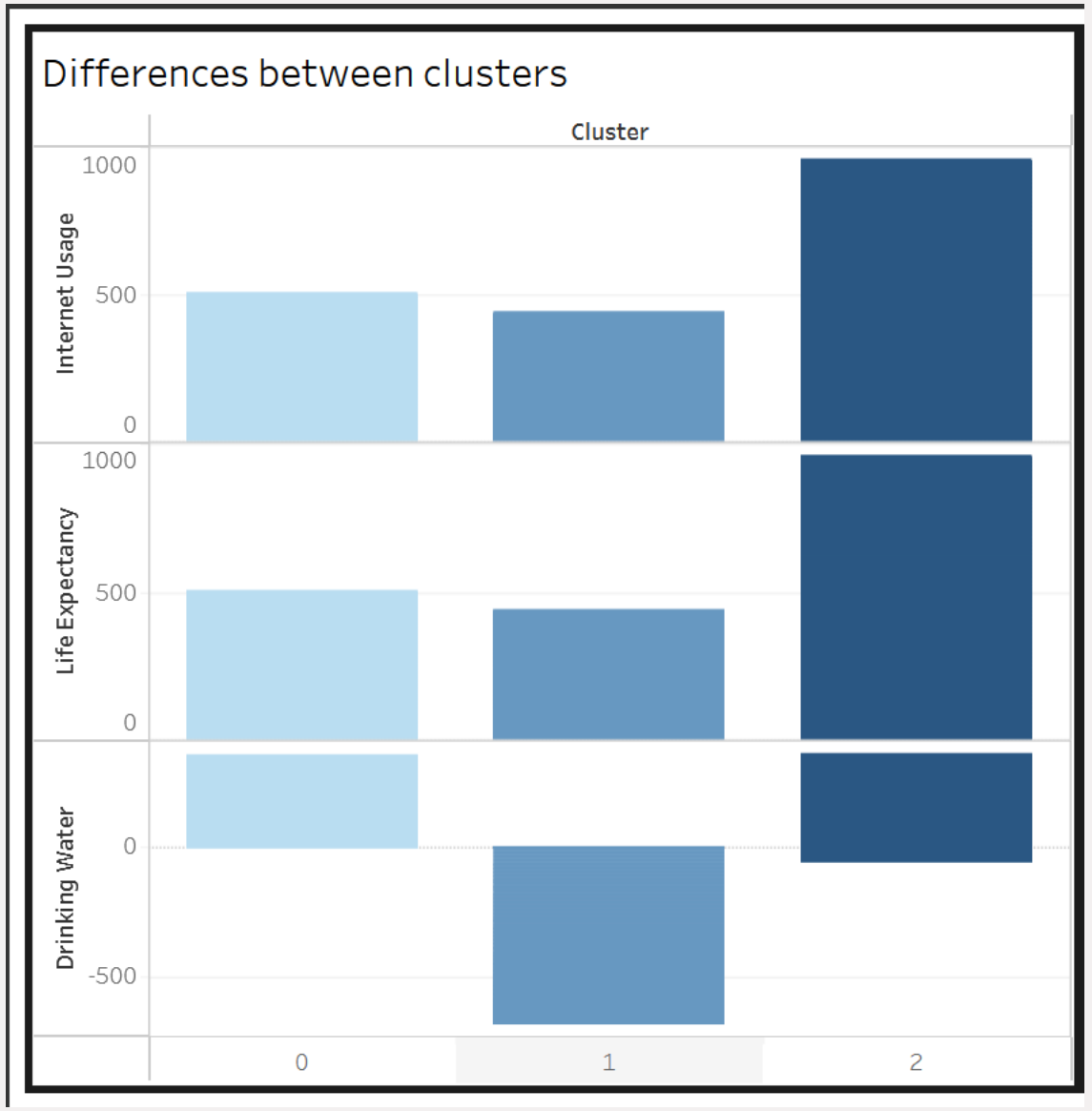
I used machine learning algorithms to perform clustering on the data. Based on scores assigned to each clustering model I chose 3 clusters which far better described the relationship between internet usage and life expectancy.

Before resorting to clustering though, I had to standardise the values so that the algorithm could generate a satisfactory clustering of data points. In addition, I was able to find that countries on different continents exhibited differing correlations.

CONCLUSION

Most points in cluster 1 belonged to Africa(early 2000's), where it seems access to drinking water services was lowest.

At low levels of internet usage, like in Africa during the early 2000's there was practically no correlation between life expectancy and internet usage. From the mid-2000's however there began to be an increasingly stronger relationship between the two variables. In other regions of the world there were very strong positive correlations between life expectancy and internet usage. The conclusion was definite and clear: Higher % of internet usage was in fact linked to higher life expectancies, especially in Asia, Europe and Oceania. A precursor was access to basic drinking water services.



THE WAY FORWARD:

Although I was now able to show that one could in fact use levels of internet usage to predict life expectancy, it would be interesting to see by which means this occurs. Is it that access to the internet helps disseminate knowledge of foods and practices that are unhealthy and thus leads citizens to adopting healthier practices? This is a question that would be interesting to answer. A multiple regression would probably even more interesting results. This is something I have planned for the near future.

Tableau Storyboard:
https://public.tableau.com/views/AnalysisintoLifeExpectancyanditsdrivingfactors/Story1?:language=de-DE&:display_count=n&:origin=viz_share_link

Github Repo:
https://github.com/mngunizulu/mngunizulu.github.io/tree/main/Projects/Life%20Expectancy