FC7P01- MSc Project Summer 2022

Predictive analytics for life expectancy of European countries- a machine learning approach

Student Name- Jiji Ajitha Kumari Venugopalan Assari

London Met ID- 20033188

Course Name- MSc. Data Analytics

Supervisor- Dr.Elaheh Homayounvala

Abstract

Life expectancy is defined as a number of years an individual is expected to live. It is based in the estimation of average age of age group, which can be different in distinct countries. In this concern, the given research study aims at analysing the life expectancy of European countries. The study will be done with the use of quantitative research method following which an experiment will be conducted to perform predictive analytics for life expectancy of European countries. In addition, a machine learning based approach will be introduced for performing predictive analytics for automatically analysing the life expectancy in more accurate way. The whole analysis in this research study has been done based on secondary data in which significant indicators will be determined using machine learning approach for the purpose of predicting life expectancy mainly for European countries. The analysis also shows that population, mortality rate and GDP are the major factors that have a huge impact on changing life expectancy in European countries. It has also been evaluated that random forest is one of the effective machine learning approaches for performing predictive analytics in getting clear insights regarding life expectancy in more accurate manner.

Acknowledgement

I would like to express my deepest thanks of gratitude to my supervisor **Dr.Elaheh Homayounvala** who made the entire research project possible for me. Her guidance carried me throughout all the stages of the given project.

I would also like to thank my family for letting my work be an enjoyable moment and my classmates for all your help in finishing every stage of the research project.

Finally, I would also like to thank God to let me through all the challenges and problems during the project. He is the one who also be with me to finish my work successfully.

Table of contents

Table of Figures

Chapter 1. Introduction

This research dissertation mainly focuses on the development of an efficient machine learning model which can be used for predicting life expectancy in the European countries with high accuracy. Life expectancy is a statistical measure of the average life of a living organism including human beings till they are expected to live. Various factors are considered while predicting life expectancy of living beings such as their current age, year of birth, gender, other demographic factors, and cultural factors. Machine learning due to its predictive analytics abilities are widely used ibn this research domain. The current study also tends to explore ML model which can help in giving higher predictive accuracies for predicting life expectancies of the citizens in the European countries. An exploratory analysis will be performed on the identified research domain, which starts with the introduction of the identified research area. In the below sections, details information regarding context of the current research, problem statement, aims and objectives, research question, significance of the research, method undertaken to complete this project and outline of the research dissertation is presented.

1.1 Background/Motivation

Life expectancy refers to the number of years an individual is expected to live based on statistical average. It varies by geographical area and the life expectancy of an individual or group of population mainly depends on different variables such as diet, access to healthcare, economic status, lifestyle etc. In simple words, life expectancy is defined as the calculation of the total number of years a person is expected to live, or it can be said that it is the count of remaining years of a person's life. In addition, counts associated with life expectancy are considered as the widely used summary indicators of the overall health of a person. The prediction or estimation of life expectancy and life limit has become significant in these days as it can effectively support existing theories that presume the availability of biological limit for the life of human being. In this concern, (Adetunji, 2020) highlighted that prediction of life expectancy is significant as it allows health service providers to make people more concerned regarding their health and motivate them to adopt healthy lifestyle choices. Additionally, Governments can also use these predictions to allocate and limit resources and services among public (Adetunji, 2020).

On an average, the life expectancy of white females who were born in 2003 in the United States is 80.4 years. In addition, life expectancy at birth in the poorly developed countries is low as compared to the well-developed countries. The main reasons behind the low life expectancies in the less developed countries at age 1 is the higher infant mortality rates which caused due to lack of water sanitization and infectious diseases (Bezy, 2022). There are several factors behind the rise of predicting life expectancies such as changing living standards, infant mortality, better education, and advances in the healthcare and medicine sectors, improved lifestyles. Apart from this, there are several other factors which impacts life expectancies in males and females such as race and ethnicity, risky lifestyle choices, and medical history of the family (Gupta, 2020).

Recent reports of Statista revealed that average life expectancy of the world for males and females is 71 years and 75 years respectively (Statista, 2022). As per the reports of 2021, life expectancy of males in the Europe is 75, whereas 81 for females. In addition, life expectancies in the EU in the year 2020 have declined in 23 out of all 27 member states. As per the current report, life expectancy at birth in males was 77.5, whereas 83.2 for females (Eurostat, 2022). It has also been identified that females have higher life expectancies than males because of biological differences in both (Gupta, 2020). As per the existing studies done on life expectancy, alcohol consumption, smoking obesity, etc. are also considered as major factors impacting the life expectancy in the EU countries. In this context, (Janssen, Bardoutsos, El Gewily and De Beer, 2021) conducted a study on 18 EU countries under the impact of these factors and found that on an average, life expectancy of females in EU was 83.4 and 78.3 for males in the year 2014, which is expected to increase significantly by the year 2065, in which life expectancy for females in expected to be 92.8 and 90.5 for males (Janssen, Bardoutsos, El Gewily and De Beer, 2021).

After going through existing studies, it has been found that accurate prediction of life expectancies is a challenging task because of the changes in the cultural, demographic, and geographical factors. Therefore, it is important to have an efficient method for making future predictions about life expectancies at a given age of the people belonging to a particular population. Also, a huge change has been found after Covid-19 pandemic, which has also become a major factor for the accurate prediction of life expectancy. In this concern, the current study will focus on performing predictive and descriptive analytics for measuring the life expectancy in Europe before and after Covid-19 pandemic for which different factors are taken into the consideration such as GDP population, infant death, adult mortality, development

status etc. In addition, the analysis on life expectancy before and after Covid-19 Pandemic is also done to attain accurate insights in the chosen domain, which can be considered by governmental authorities to make proper policies and procedures to handle situations to be occurred with the increase or decrease of life expectancy rate.

1.2 Research Aims

This research study basically aims at introducing an efficient model based on machine learning to calculate the life expectancy of European countries based on different factors such as GDP, infant death, population etc. In addition, the life expectancy of European countries before and after the spread of Covid-19 will also be analysed through descriptive analytics for the purpose of getting clear insights regarding the impact of pandemic on life expectancy rate so that accurate strategies could made to manage it in future in case of any kind of pandemic.

1.3 Research Objectives

Below are some objectives given which helps in accomplishing the main aims of this research

- To predict life expectancy of European countries according to GDP and other factors using predictive analytics
- To analyse life expectancy of European countries before and after the spread of Covid-19 using descriptive analysis with the help of Python and Tableau.

1.4 Research Questions

To accomplish the aims and objectives of the current research dissertation, below mentioned research questions will be taken into consideration.

RQ1: What are the major factors that can be used to measure life expectancy of European countries before Covid-19?

RQ2: To what extent proposed novel machine learning approach can be accurate in calculating and predicting life expectancy of EU citizens?

1.5 Significance of Research

Life expectancy is a hypothetical measure helpful to assume the age specific death rates of the individuals for a given period in which a particular age-group is born. As per age, sex, race,

ethnicity, and other cultural and geographical factors, the rate of life expectancy can differ. Life expectancy determination plays an important role in making suitable decisions for ensuring the good health of the individuals. It provides many benefits to the individuals, healthcare sector professionals and service providers as well as governments to develop suitable health interventions for protecting the public health. Therefore, the current research study is mainly focused to develop a machine learning model for the purpose of predicting life expectancy regarding European countries based on different factors such as GDP, infant death, population etc. The study will also include descriptive analysis regarding the life expectancy of European countries before and after the spread of Covid-19, which can be considered as very significant insights for government, healthcare providers and others in terms of taking efficient decisions to manage the life expectancy rate in case of causing similar type of pandemic situations in future. The main purpose of focusing on European countries is the major changes found in these countries in terms of GDP rate and other major factors, but very few studies have been done in this area on some common factors. Due to this, it has been found as significant to understand the reasons behind the continuous changes in life expectancy in European countries and get insights to assist governments for taking actions to manage these changes efficiently to maintain the expectancy level efficiently.

Chapter 2. Literature review

This chapter of dissertation involves a review-based analysis which will be carried out in context of the chosen research topic. Thematic analysis approach has been selected for conducting this comprehensive review. As a matter of fact, thematic analysis will help in examining themes relevant to chosen research topic. It will help to review existing research which has been done on selected research topic in an efficient manner such that users can interpret issue which has been discussed in it. Relevant themes which are formalized in context of chosen research context are entailed below-

2.1 Review of work done

2.1.1 Theme 1- Need of predicting Life expectancy

Life expectancy can be viewed as an average count of number of years of life lived by an individual who has reached a certain age. It is specified that areas or countries that have relatively low life expectancy rate must have adequate health development programs and social programs in place; especially those which lay major emphasis on factors such as environmental health, nutritional and calories adequacy. In this context, (Zamzamy Sormin et al., 2019) described key benefits of knowing life expectancy as it helps in evaluation of performance of government bodies in enhancing welfare of public and improving health status. The researchers have outlined that there is a need for estimation of life expectancy because it can help government authorities of countries such as Indonesia to determine relevant policies and other strategies so that life expectancy rate which is falling each day can be increased. Based on life expectancy of 38 countries in the world, which are calculated after every 5 years, it was found out that Japan is one such country which has the highest life expectancy at 83.5 years which is followed by Hong Kong with 83.3 years. Indonesia, however, managed to rank itself at the lowest with life expectancy of 70.1 years which is relatively low.

Therefore, it is suggested that government must estimate life expectancy in the following years so that they have clear references and are able to formulate policies and strategies with which high life expectancy rate can be assured. Similar ideologies have been explained by (Bali, Aggarwal, Singh and Shukla, 2021) in their article showing relevance and need of predicting life expectancy rate as it can help government bodies to improve healthcare systems and decide certain strategies with which better life expectancy rate can be achieved. Further, authors have

explained about a technique i.e., cyclical order weigh which according to authors is a proficient technique used for estimating life expectancy. It was determined that this technique makes extensive use of artificial neural network because it is suitable for handling problems related to estimation, grouping and pattern recognition. This method can be applied to calculate life expectancy by dividing data related to world population life expectations into training and testing parts and accordingly applying ANN algorithm for computations.

In similar research context, (Bali, Aggarwal, Singh and Shukla, 2021) undertook an investigative study to highlight the need and purpose of life expectancy since it has vast effects on social and financial positions of different countries around the globe. The researchers opined that determining life expectancy is essential because it helps in analyzing social aspects as well as healthcare system management around the world. There is a need for development of certain models with which life expectancy can be calculated because these models can provide certain ways using which healthcare and care planning mechanism related to society can be improved to a large extent. It is stated that the current models for estimation of life expectancy are not appropriate enough to predict longevity of generic set of population. There is no denying the fact that life expectancy observations have been extensively used in different areas such as medical, healthcare planning and other pension-related services. Similar idea had been proposed by (Zamzamy Sormin et al., 2019) as researchers outlined that government could decide its healthcare initiatives of it is aware of life expectancy rate which is reported in current times. These observations had been used by government authorities and other private bodies for improving healthcare system management in their country because these observations provide valuable insights to certain determinants that affect life expectancy rate. In different countries, it has become a controversial issue as to which age must be regarded as retirement age and how to comprehend and manage financial issues in relation to public matter. It has been clarified that prediction of life expectancy is vital for government authorities for policy formulation which will have substantial impact on healthcare facilities. As better healthcare facilities, education, and positive change in lifestyle of people can be brought if accurate life expectancy rates are calculated as it will help government bodies and other private agencies to comprehend GDP (gross-domestic product) related and other factors that significantly impact life-expectancy such as poverty rate of that economy.

Conversely, researchers of (Beeksma et al., 2019) have provided details about relevance of predicting life-expectancy has been explained. Indubitably, life expectancy is one of those critical factors in end-of-life decision making. It has been clarified that prognostication could

be an instrumental tool in deciding the course of treatment that one must take. Not just this, it can help to decide procurement of healthcare services and other facilities with which better course of treatment can be anticipated. As per researchers, an effective prognostication could facilitate advance care planning which would substantially enhance final phase of life. As a matter of fact, it would help doctors to investigate preferences for end-of-life care for their patients. Oftentimes, it is seen that physicians tend to overestimate life whereby missing a window of opportunity to facilitate advanced care planning. The researchers have provided a very limited discussion on the need and relevance of life expectancy prognostication which is therefore marked as its one key limitation. In the article, majority of emphasis has been laid on machine learning models which could be used for development of an improvised model. The study highlighted that the existing life expectancy estimation models are highly inefficient because they tend to overestimate life expectancy which thereby shows inaccurate prognoses. This subsequently affects decision making ability of government authorities as they fail to understand key determinants that significantly impact life expectancy rate. Lastly, it has been concluded that machine learning and NLP (i.e., natural language processing) techniques are two promising techniques which can help to develop a promising approach using life expectancy prognoses can be done efficiently. In fact, these techniques have the potential to be used in real-time applications for estimation of life expectancy rate.

2.1.2 Theme 2- Effect of various factors of life expectancy

(Li et al., 2018) undertook an empirical study for showing the impact of healthy lifestyle factors on life expectancies in relation to US population. Americans have relatively less life expectancy in comparison to residents of almost all other high-income countries. In the conducted study, researchers have outlined some low risk factors such as never smoking, body mass index of 18.5- 24.9 kg/m2, physical activity involved, moderate alcohol intake and diet quality score. These are some of the pertinent low risk lifestyle factors which have been taken for estimation of hazard ratios. It was identified that 95% people got subjected to cardiovascular disease mortality. Not just this, it was discovered that life expectancy at the age of 50 years was found to be 29.0 years for women and as for men, it was recorded to be 25.5 years. One important fact which must not be overlooked here is that records for men was taken for those who adopted zero-risk factors. Further, it was found out that key prominent factors which significantly impact life-expectancy rate of people typically includes factors such as their diet, BMI, physical activities involved, alcohol usage and smoking status.

(Mackenbach et al., 2019) have also identified several determinants of inequalities in life expectancy. It has been clarified that inequality in morality between socioeconomic groups is relatively high and this is what translates into high inequality which could be identified in life expectancy. The research study highlights certain factors which significantly contribute to morality inequalities such as- material living conditions, childhood conditions, behavioural risk factors and psychosocial factors. Similarly, (Li et al., 2018) have also described certain health related factors and psychological factors which impact the life span of individuals. The study highlights broad range of risk factors which are highly persist in European countries including Norway, Sweden, Finland, England and Wales, Belgium. Switzerland, France, and Spain. The overall study results showed that life expectancy was found to be shorter amongst people who have relatively low levels of education. There is also a table provided in the study which indicates life expectancy and gaps differed between different countries. Some of these factors responsible behind expectancy are- low income, few social contacts, smoking, high alcohol consumption, high bodyweight, and less physical activity. It can be seen as following figure1. From the figure 1, it can be analyzed that men with high level of education had partial life expectancy which ranged between 39.4 years – 42 years in Lithuania and Switzerland, respectively. Another major discovery is that majority of factors were highly prevalent amongst people with low level of education and who had high alcohol consumption. It was identified that largest inequalities were identified for people who were indulged in smoking and had low income. Similar views have been endorsed by (Miladinov, 2020) undertook an investigative study for analyzing the effects of socioeconomic development on life expectancy at birth. This has been taken as a key indicator of mortality or longevity in five EU accession candidate countries which mainly include Serbia, Macedonia, Bosnia, and Albania). The researchers pointed out that income significantly influences the conditions of people's living. Income is one of the main determinants of health which has a direct and position impact on the demographic changes i.e., positive relation with individual morality or to life expectancy. Further, analyzing results from Preston's article (1975), researchers pointed out that life expectancy is profoundly less for economies which have relatively less per capita income. From this, it can be articulated that views endorsed by researchers of present article are quite similar with ideas endorsed by (Mackenbach et al., 2019).

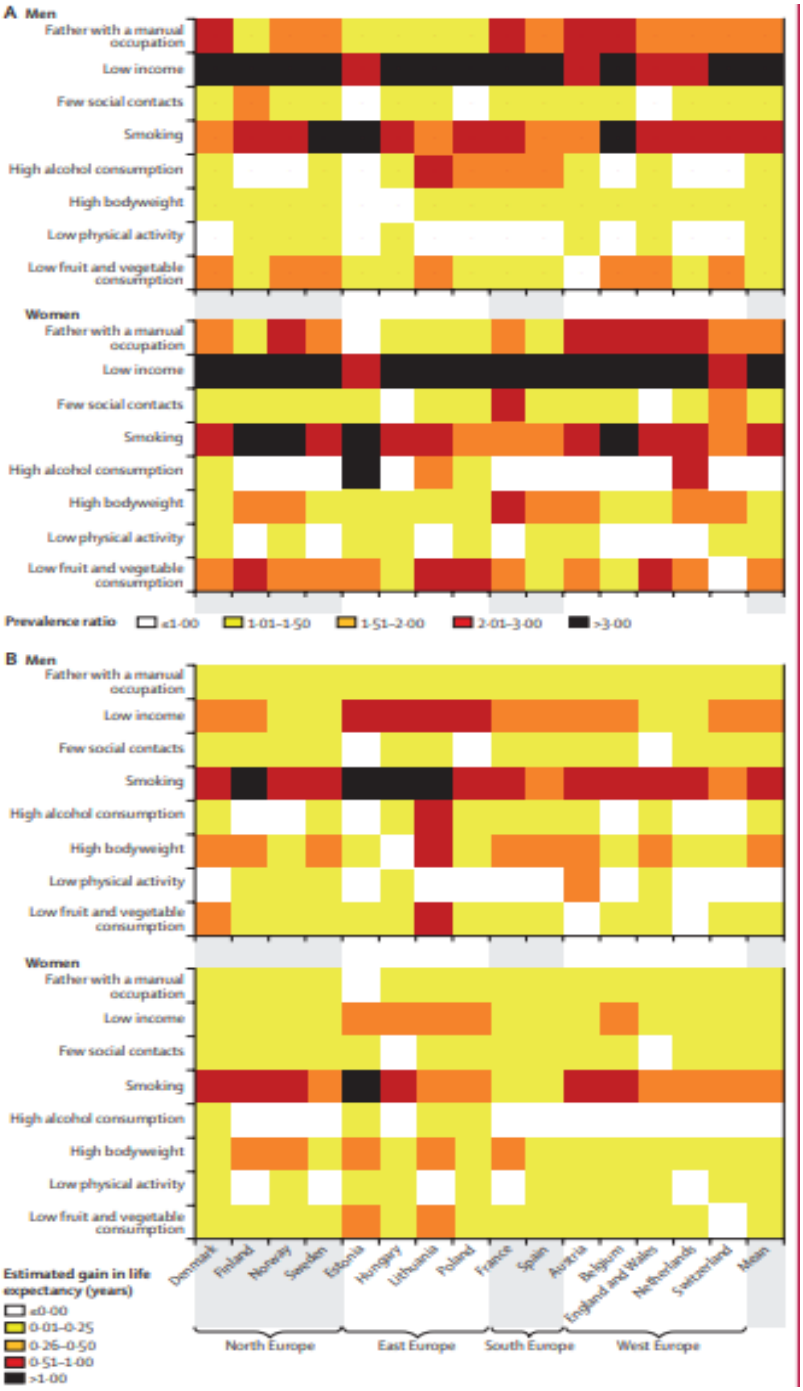

Figure 1.Factors responsible behind expectancy (Li et al., 2018)

Another major finding is that there lies a positive correlation between factors such as income and health as low-income people tend to live shorter lives than those who have high income. Aside from income factor, there is a pertinent factor i.e., infant mortality factor which basically reflects children's socioeconomic development and well-being. There are considerably many socioeconomic determinants which impact child mortality. Apart from this, researchers have identified certain factors which are more likely to influence health status of whole populations. Some of these factors majorly include- economic growth, general living conditions, and social

well-beings, quality of environment and rates of illness. Further, it has been clarified that technological advancement has also brought a positive change in the lifestyle of individuals and subsequently it has increased longevity and life expectancy of world population to a significant magnitude. Similarly, (Aburto et al., 2020) in which authors have attempted to explain dynamics of life expectancy and life span quality. Measuring life expectancy has now become a necessity for every country because it acts as a useful measure of average life spans. It is generally seen that although life expectancy is monitored in every country, yet few countries have initiated monitoring and acknowledgement of importance of disparities in age at death.

Certainly, in many counties and different communities within a country, a major downfall could be seen in life span equality, although a major rise was noted in average life span. Another major discovery in this research article showed that main causes of death which had significantly contributed to increasing life expectancy are different from those factors which significantly increased overall quality in life spans especially in developed economies. The authors of (Bilas, Franc and Bonjak, 2022) have provided some relevant details related to key determinant factors of life expectancy at birth and that too in context of European nations. There were a few variables which were shortlisted for exploration of key determinants of life expectancy in EU nations. Some of these variables defined are- gross domestic product i.e. GDP rate. Level of education, population growth rate, education enrolment, GDP per capita income and life expectancy. Here, it is important to note that the aforementioned factors all sum up to those key factors which significantly influence life expectancy rate of whole population. The study results also indicate that factors such as GDP per capita income and attained education level together show differences of around 72.6% and 82.6% in life expectancy at birth. From this, it can be anticipated that these are two most prominent factors which significantly impact life expectancy. The empirical data gathered in the study shown that public health care expenditures and other lifestyle factors which mainly include- high alcohol intake, tobacco, food, education, environmental pollution, and income are some other health determinants. The study results indicate that these factors have a significant impact on life expectancy and declining premature mortality.

(Knauss, 2022) is yet another study which has been performed in this research regard showing as to how health physical and diet factors could increase life expectancy in future. The research is relevant to the chosen dissertation topic because it provides credible information related to physical and diet factors that could influence life expectancy rate. Key differences in factors

such education, employment opportunities, social mobility, lifestyle behaviour and wider local environment are some of those pertinent factors which tend to have major impact on males as well as female longevity. (Stenholm et al., 2016) have also presented similar ideas to this research showing factors such as smoking, tobacco, physical inactivity, and their role in minimizing life expectancy rate of individuals. Further, it has been clarified that life expectancy at birth has exponentially increased in the last century and there are many factors which have driven this change. Some of the most important factors which have led to this dramatic change in life expectancy rate include- reductions in infant mortality, improving living standards, improved lifestyles, better education opportunities as well as improved healthcare and medicine facilities owing to notable technological advancements which have taken place in the last century. Not just this, study also revealed that economic growth and enhancements in environmental conditions such as improved lifestyles, advancements in healthcare and clinical setting have also resulted to a continuous improvement in life expectancy rate in the last century. Contributing to the previous studies, (Stenholm et al., 2016) have undertook an investigative study for explaining role of factors such as smoking, physical inactivity, and obesity as predictors of healthy and disease-free life expectancy between an age group of 50-75. In this multi-cohort study, researchers outlined several risk factors such as smoking, physical inactivity and obesity stating that these are some of the modifiable risk factors which lead to morbidity and mortality.

A comparative technique has been employed in this study in which comparison was performed for demographics who had least two behaviours related risk factors to those who had zero behaviour related risk factors. Overall results drawn from this study revealed that people who showed zero behaviour related risk factors are more likely to live on average 8 years longer in excellent health conditions and 6 years longer free of chronic diseases during age between 50 and 75. The study results are like those of (Li et al., 2018) showing health related psychological factors and how they tend to impact life span conditions of individuals. It was found that identification of any single risk factor could have resulted in bringing reduction in healthy years (Stenholm et al., 2016). The overall study results indicate that people residing in European countries have better short at healthy life expectancy and are less likely to get subjected to any chronic disease if they do not show any co-occurring behaviour-related risk factors. Another major finding of the study was that population level reductions in physical inactivity, obesity or smoking is highly likely to increase life years lived in excellent health conditions.

2.1.3 Theme 3- Life expectancies of European Countries

Mortality projects are crucial for forecasting as to how long will people live on average, predictions are necessary for analyzing the future extent of population ageing and for analyzing sustainability of social security mechanisms and pension schemes (Janssen, Bardoutsos, El Gewily and De Beer, 2021). This is primarily needed to set life insurance premiums and for assisting government plan for addressing increasing demand for services such as healthcare services.

(Leon, 2011) have provided a salutary view for highlighting major trends in European life expectancy. In the article, it has been clarified that the fluctuations in overall life expectancy in Easter Europe is the main cause behind increasing variations in mortality that could be seen in working ages. This phenomenon is striking because it in times of social upheaval and change, it is usually speculated that people who are either too young or too old are more likely to be highly vulnerable. Having said that, it is important to understand that the most important influences on life expectancy trends in former soviet countries is found to be hazardous alcohol consumption. Heavy and hazardous alcohol consumption is what truly accounts for a larger fraction of circulatory disease related deaths than it was previously presumed. Also, in terms of population level, it was found that hazardous alcohol consumption is found to be greater amongst men and women and this scenario holds true value in context of former Soviet Union. Conversely, (Kolasa-Więcek and Suszanowicz, 2019) has undertook an investigative study to highlight correlation between air pollution and life expectancy rate in European nations. A neural network based novel methodology has been developed for understanding the correlation between these two factors. There were a few input variables which have been designed to understand this association and it is important to note that these are a few key determinants of air pollution which significantly impact life expectancy of individuals in European countries. Some of these variables or air pollutants are- polycyclic aromatic hydrocarbons, non-methane volatile organic compounds. The overall findings of the study show that each of these air pollutants tend to affect life expectancy of individuals residing in European countries.

Another significant contribution in this research regard has been done by (Mäki et al., 2013) in which researchers have outlined educational differences in disability free life expectancy. Healthy life expectancy can be marked as a composite measure that is used to determine length as well as quality of life that are bother important health indicators in aging populations. In the study, it had been clarified that life expectancy and disability-free life expectancy both have a positive correlation with each other i.e., are directly related to the level of education. However,

it was found that the educational differences are relatively higher in the latter and that too in majority of European countries. The study results revealed that highly education Europeans can expect to live relatively longer and that tool in better health than those who had lower education. Further, it was analysed that the smallest and largest disability free life expectancies were found to be in southern Europe and northern Europe.

Similarly, thoughts are presented by (Khouri, Klaudia and Cehlar, 2022) in which researchers have attempted to identify key determinants of expected life expectancy rate in selected European nations. To explore the determinants that significantly affect life expectancy rate, researchers have quantified the implications of selected determinants on the overall lifestyle of people in European regions. After drawing discussions from the undertaken study, unexpected results were obtained as it showed that GDP per capita income is one key determinant which affects life expectancy rate of individuals. An existing study undertook by (Zamzamy Sormin et al., 2019) also produced similar results showing association between GDP per capita income factor and health outcomes which substantially affect the life expectancy rate of world's population. Equivalent to (Khouri, Klaudia and Cehlar, 2022), another pertinent study had been undertaken by (Raleigh and Fund, n.d.) in which authors have highlighted key trends in life expectancy in European nation and other OCED countries. It was analyzed that the pace of mortality improvement is much lower in European countries than other OCED countries. It was analyzed that this trend has gone wrong since 2011 when diseases of older ages were found to a large extent, and this is what has led to a major slowdown in mortality improvements in economies such as European countries and Australia. A major decline could be witnessed in the mortality improvement in countries like USA and UK and the cornerstone behind this scenario is the intermittently increasing number of deaths of working age adults owing to drug related accidental poisoning. There are several health-related factors i.e., risk factors which are quite prevalent in European countries such as obesity, diabetes, smoking, excessive alcohol consumption, high blood pressures and cholesterol levels.

Similar research has been conducted by (Robine, 2022) in which authors have studied about health life expectancy in context of European countries. There are three measures which are commonly used in European countries for measuring health expectancy. Each of these factors reflect a specific dimension of health and the first factor is that of perceived health which us a subjective indicator which is associated with mortality risk and healthcare consumption. Like the research outcomes presented by (Stenholm et al., 2016) in which health indicators also included factors such as individual lifestyle or perceived health treatment, the authors of current

article also provide a similar view; stating that perceived health helps in determining life expectancy of people in European countries. Self-reported chronic morbidity is yet another factor which is based upon user's level of knowledge of their state of health. This will help in analyzing any sort of disability that the respondent might have which may influence the decision outcomes related to his life expectancy. Third important factor is that of interpreting estimates as well as key differences between countries wherein it is said that these cross-country comparisons or changes over time. Further, researchers have shown prevalence of health-related problems in European Union-27 in the year 2010 by sex and age. A tabular representation will be provided to show predicted life expectancy at age 65 in EU countries by health status (as per 2008-2010) in the figure2.

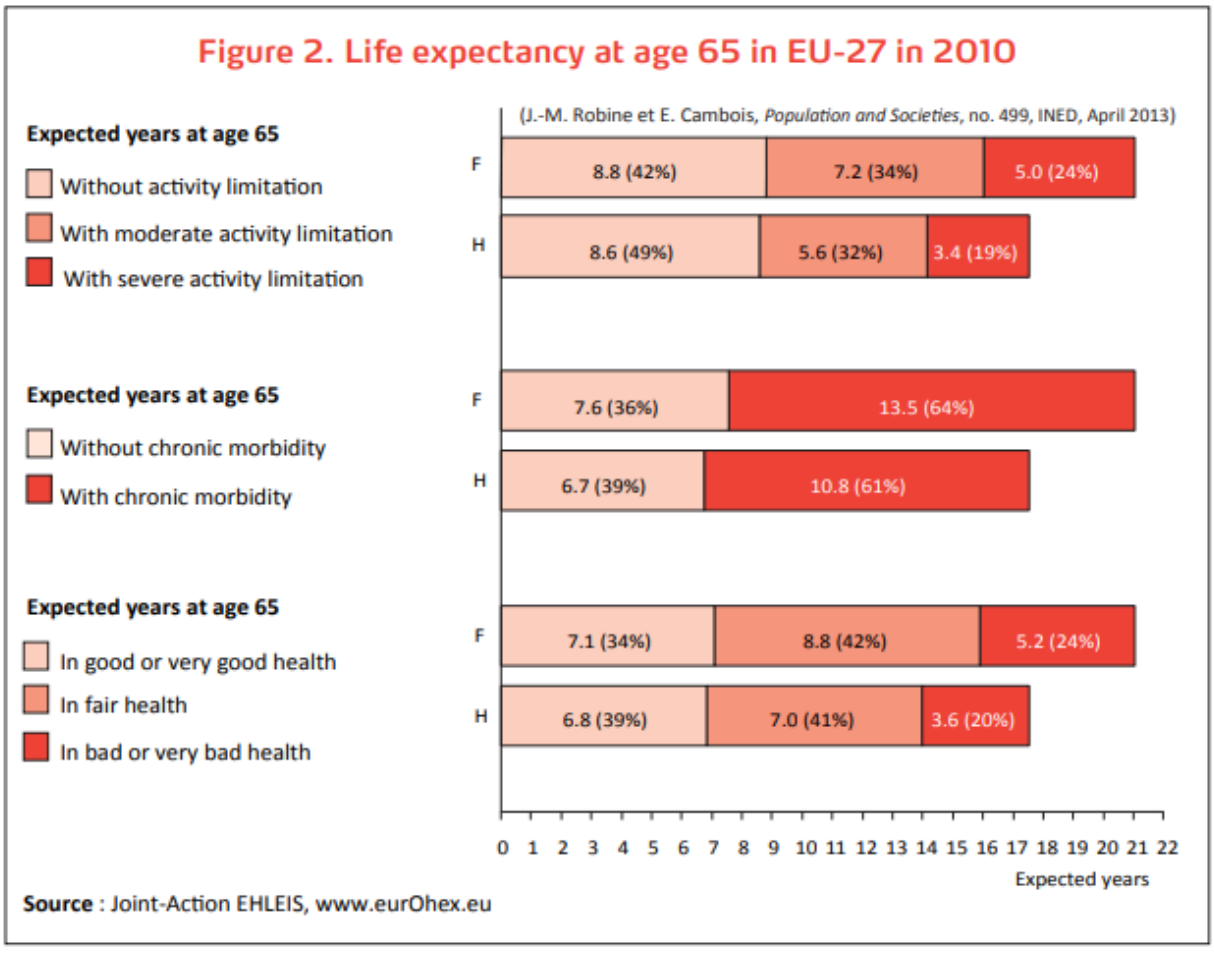

Figure 2. Life expectancy at the age of 65 in EU (Stenholm et al., 2016)

Another prevalent study in this domain has been performed by (OECD, 2018) in which authors have described several trends in life expectancy. There is no denying the fact that life expectancy has increased in EU over past few decades particularly in Western Europe. It is

analyzed that life expectancy at birth has now reached up to 81 years in almost 28 European Union member states in 2016. It was analyzed that EU countries such as Spain and Italy were found to be one of those countries which have relatively high life expectancy amongst other EU economies, with life expectancy reaching up to 83%, as it was found in year 2016. In almost 2/3rd of EU countries, life expectancy of up to 80 years could be found, and some of these EU countries include Bulgaria, Latvia, Romania, and Lithuania. Further, researchers have talked about some austere measures that have been implemented by countries such United Kingdom for effectively dealing with potential implications of austerity measures on health and other public spending. Nonetheless, in EU countries, some of EU nations have begun to adopt more severe austerity measures to increase its life expectancy rate wherein a notable exception of year 2015 during which a major downfall; was reported in life expectancy in both these nations. Therefore, researchers articulate that there is a need to conduct further research in this domain to gain better understanding of how recent slowdown in life expectancy gains in majority of EU nations can be effectively dealt with.

2.1.4 Prediction of life expectancy using machine learning algorithms

In this regard, (Beeksma, et al., 2019) conducted a study to highlight the effectiveness of using machine learning in the prediction of the functional health of elderly population. The author stated that in recent times, human lives are longer, and their life expectancies have also increased. In general, the lifespan of a human being is supposed to be around 125 years and the records show growth in old age deaths. It has also been analyzed that there is no conventional method which could explain this disagreement. This could be supported by the prediction or estimation of the life limit of humans using current theories made regarding this subject matter and it is assumed based on the biological life limit of humans. Therefore, the author in this analysis proposed a system which is expected to help in predicting the life expectancy of humans by taking some parameters such as BMI, AIDS, HIV, etc. These parameters are taken into consideration in model training so that the accuracy or the proposed system can be enhanced. For this purpose, CNN is adopted which is an artificial neural networks-based technique and uses perceptron as well as machine learning unit algorithms for analyzing data. Furthermore, the author explained major steps involved in the proposed algorithm such as importing libraries and data sets, avoiding the dummy variables trap, adding hidden layers to the network, splitting datasets into two sets as training and testing, fitting the model and predicting the test results (Beeksma, et al., 2019). In this analysis the author has also used some independent variables such as time spent on web and applications, email address, gender,

length of membership and salary, whereas dependent variables include yearly amount spent. From the findings of the study, it is seen that the proposed CNN based model predicts the life expectancy of the individuals. It has also been found that the proposed method and accuracy of the system can be enhanced by adding some new parameters and variables. In the future, the author wants to extend this system by implementing it on mobile phones as well as on desktops.

In a similar context, (Beeksma, et al., 2019) conducted a research study to predict the human life span using machine learning algorithms and techniques. Further, the author mentioned that human life span depends on various factors and features such as financial development and well-being of the nation's people. To conduct this study and accomplish the objectives of this research the author has conducted experiments on the dataset of WHO life expectancy which is taken from kaggle.com. This dataset contains previous records related to human life span and consists of 22 features and 2938 rows. Due to the predictive analytics capabilities of machine learning, it is being used in several domains of research and has grown exponentially in recent years. Additionally, the author mentioned that selection of appropriate data for designing a machine learning model is a challenging task. In addition, the study expounded that anticipation of life can be determined by analyzing the normal endurance time of a person, which shows the average of the population, when someone from the same age group dies or lives till that point. Prognosis, prediction, and classification are the three major aspects in this research area which are used to predict the life expectancy of individuals. Prediction of diseases depends on various factors including social, economic, cultural, and other factors, which impact the growth of the country, GDP, illiteracy rate or literacy rate, birth, and population awareness etc. For this purpose, machine learning models can be used to provide accurate results for disease prediction or predicting the lifespan of individuals. Furthermore, the author described the working of machine learning models in a diagrammatic manner as illustrated below figure 3.

The author has applied various regression models on this dataset which include mean squared error (MSE), root MSE, and mean absolute error, etc. to perform classifications and regression tasks in this analysis, the author used decision tree algorithms. Also, a random forest regressor is used for predicting the final output generated by the proposed algorithm using several decision trees. By applying backward elimination, regressions models, classification algorithms (such as cross validation, random forest, KNN, etc.) the proposed model is evaluated for determining the life expectancy.

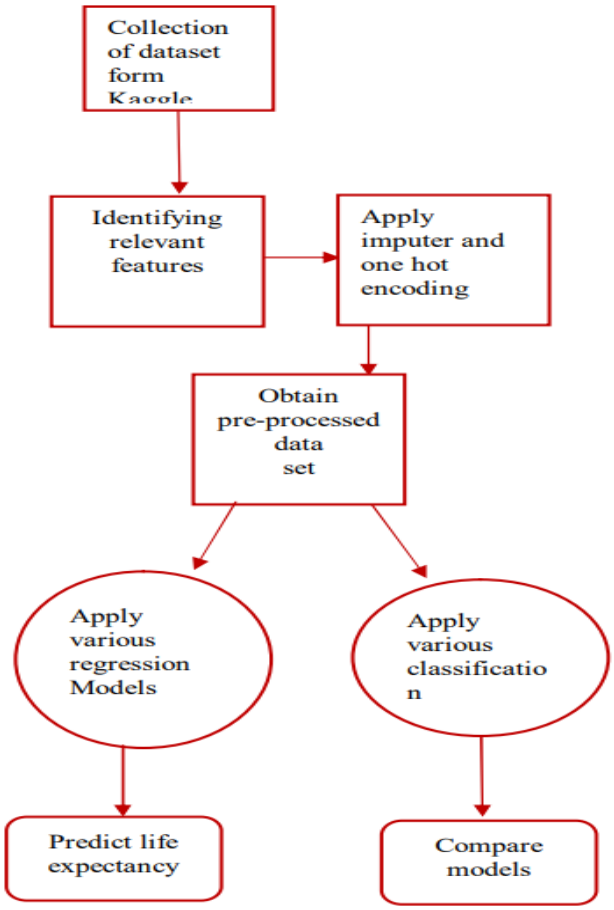

Figure 3: Illustration of the proposed model (Beeksma, et al., 2019)

Results of this study indicated that the r-squared value is better obtained with random forest regressor. Furthermore, the backward elimination method provides information regarding the features such as adult mortality, total expenditures, GDP, etc. which impacts life expectancy. The overall results of this study indicated that random forest gives best accuracy of around 88% among other classification methods for predicting life expectancy.

Furthermore, (Svensson, 2018) also sheds light on the prediction of human life expectancy in future by taking insights from the previous data using machine learning algorithms. Machine learning algorithms are used in this prediction work because it provides various techniques which can be used to draw inferences about a given problem based on the selected dataset and provide predictions accordingly. Further, this study illustrated that prediction of life expectancy in any country is very important so that respective nations can analyze the requirements for the development and growth in future. The primary reason behind selecting a machine learning model for conducting this analysis is that it is the widely accepted application of AI and able to learn and improve systems automatically without requiring any explicit programming. Machine learning models are mainly used in the development of computer

programs that access and learn data by themselves. Apart from this, the author illustrated various types of machine learning algorithms such as unsupervised learning that consists of clustering approaches and supervised learning including classification and regression algorithms. To predict the life expectancy of the people from different countries, these algorithms of machine learning are used which take input from the previous data and provide outputs as predicted lifespans. Apart from this, (Beeksma, et al., 2019) provides insights regarding the effectiveness of using LSTM (long short-term memory) in predicting life expectancy of individuals.

In this study, a user interface is also developed using node RED so that general users can also access this system. The proposed regression algorithm is extended by adding more features and parameters to enhance the accuracy of the system so that future insight can be predicted using previous data. This system is also implemented on mobile phones and desktops to predict the average predicted years of life of an individual. Various tests are such as feasibility analysis, unit testing, integration testing, software testing, black box testing, security testing, etc. to ensure the proper working of the proposed system as well as to determine if the integrated software system meets its all requirements or not. In this system, input is given as certain features such as country and year to analyze its impact on the outcomes of predicted life expectancy (Svensson, 2018). By performing a comparative analysis on several machine learning models, it has been found that random forest gives best performance results and factors such as schooling, adult mortality, HIV/AIDS, BMI, etc. mostly impacts the life expectancy of individuals. In addition, factors such as schooling, BMI and income composition are found to be positively associated with the outcomes produced using machine learning models. On the other hand, impact of features such as infant deaths, GDP and total expenditures do not lead to any significant impact on the results.

2.1.5 Summary and Research gap

The above analysis illustrates the importance and need of predicting life expectancy of individuals so that suitable measures can be put in place to foster the development and growth of the nations. Herein, all the major findings obtained by performing literature-based analysis are described so that strengths and weaknesses in the existing studies can be determined. On analysing these findings, it is found that covid-19 pandemic caused mortalities highly impacted the prediction of life expectancy of individuals across the globe, due to the seriousness and widespread infection. On the other hand, this analysis also sheds light on some machine learning algorithms that can be used to predict the life expectancy of people by considering the

social, economic, environmental, and cultural factors. As per the above analysis, in most of the studies, it has been found that many studies have already been focused on evaluating the life expectancy based on GDP and population, but these studies lacked in focusing the impact of Covid-19, which could have a huge impact on changing life expectancy rate. Therefore, the given research study will be conducted for performing predictive and descriptive analysis to measure life expectancy based on different factors such as GDP, population, income composition, infant deaths, adult mortality etc. along with Covid-19. This type of studies will also provide practitioners or researchers with an efficient area of research on the basis of which they can conduct further studies for the purpose of identifying the suitable strategies for the EU countries and their governments that can be deployed for maintaining and controlling the life expectancy rate in an efficient even in different types of circumstances such as increase or decrease in population, increase or decrease in mortality rate, different types of pandemics or lockdowns in cities, increase or decrease in infant deaths and many more. The existing literature demonstrated that due to the increased prevalence of covid-19, the number of deaths has also increased significantly worldwide. It is expected that if this prevalence of deaths is increasing at the same pace, then it may impact the life expectancy in an indirect manner. Therefore, it is important to consider Covid-19 as a major factor for the purpose of evaluating its impact on life expectancy so that effective strategies could be prepared to manage them in a more appropriate manner.

Chapter 3. Methodology

This section of the research dissertation provides in- depth information about the methods or techniques used for successfully completing the research study. It also illustrates deep insights about an appropriate structure which is to be followed for achieving the designed research objectives. Herein, an appropriate methodology to predict life expectancy of European countries and continents using a Machine Learning approach is explained in an appropriate manner. Research methodology, data collection, data analysis and ethical considerations are the main sections involved in this chapter of dissertation providing knowledge regarding a selected methodology either qualitative, quantitative, or mixed for providing a solution to the defined problem.

3.1 Selected Research Methodology

The main aim of conducting this research study is to predict the life expectancy of European countries according to GDP, income composition, infant deaths, adult mortality etc. and make individuals aware about their general health and its improvement over time. In addition to this, analysis is also performed on life expectancy of European countries before the spread of coronavirus as well as different factors are identified that play a key role in impacting life expectancy. This can be performed in an appropriate manner with the help of a research methodology or a research design which helps in providing a systematic plan or structure for achieving the research aims and objectives. To complete this research study, quantitative research methodology is implemented in this research study. The main agenda of selecting this research methodology is its ability to emphasize on figures and numbers in data collection and analysis related to predictive analysis for life expectancy of European countries. To perform this research study, literature review and experimental analysis approach are used that helps in successfully completing the research study.

In this research study, literature review is performed to understand the perceptions and opinions of existing researchers about the change in life expectancy of European countries, factor impacting the life expectancy based on GDP in European countries and life expectancy before the spread of coronavirus of European countries. The performed literature review also helps in the identification of research gaps, open questions left, improves knowledge regarding the topic under consideration and allows the identification of innovative ways to interpret the prior

research. To perform this, multiple peer reviewed research articles, academic journals and conference papers are used from online available digital repositories. Moreover, to collect accurate information that helps in answering all the defined research questions as well as achieving all the defined research objectives, there is a need to implement a data collection strategy as well as an appropriate technique is used for analysing the collected data. A brief description of the selected methods for collecting and analysing the data are elaborated in the following section of the research methodology chapter.

Along with this, experimental analysis is performed in this research study to analyze the life expectancy of European continents by selecting an appropriate dataset. The life expectancy is further analyzed with the identification of different factors based on the selected dataset. The overall analysis is done based on predictive and descriptive analysis where predictive analysis is done for income composition, infant deaths, adult mortality etc. and descriptive analysis is done for Covid-19 factor to measure their impact on life expectancy of Europe.

As mentioned above, that research methodology provides a systematic structure for completing the research study. The first step is to identify a research problem, a problem related to the prediction of life expectancy is considered in this research study. It has been determined that there is no appropriate approach to predict life expectancy because of which the current research study is introduced. Following this, aims, objectives and questions are formulated to successfully complete the research and derive conclusions. Following this, literature review is performed to understand the methods used by existing researchers to predict life expectancy of European countries and continents. Also, it helps in determining a brief research gap on selected topics which is further addressed for creating the research questions/ hypotheses. The next step is to demonstrate an appropriate methodology and methods of data collection and analysis whereas quantitative research methodology is implemented in this current research study followed with an experimental analysis approach for obtaining valuable insights. Further, data collection and analysis are performed to achieve and answer the formulated research objectives and questions, respectively. The research methodology, data collection and analysis are the main steps involved in a research dissertation as they ensure the quality and validity of the information generated. The results obtained are highlighted separately in reporting the results section to make the readers easily understand the main motive of conducting this research study. Based on results obtained, a conclusion is derived that ensures the completion of research by providing a solution to the designed research problem which is further followed with future work in a similar research domain.

3.2 Data collection Method

To carry out the research study on prediction of life expectancy of European continents according to GDP and various factors, secondary data is collected. The main reason behind the selection of secondary data collection is to generate new insights from already available information on the internet. This secondary data for the same is collected from various secondary data sources including peer reviewed research articles, conference papers, academic journals, governmental publications, official documents, reports, and other relevant sources. To ensure the quality of data to be generated, the majority of the information is collected from online digital repositories as it holds a wide range of information for a variety of purposes such as learning and research. Some of the digital repositories include ScienceDirect, Mdpi, Emerald, SpringerLink, IEEE and Taylor & Francis. These repositories include immense data; therefore, it becomes difficult to select an appropriate one that helps in achieving the defined research objectives. To overcome this issue, there is a need to implement a relevant data collection method whereas keyword-based strategy is the one that is used for collecting valuable information for conducting the literature review.

In addition to this, only those research articles are considered that were published after 2015 to ensure the overall quality of the literature review. Also, experimental analysis is performed using a dataset that include various factors and are helpful in predicting life expectancy of European continents. This dataset is further trained to obtain valuable statistical information regarding the concerned research problem. This collected data is further required to be analysed for achieving and answering the designed research objectives and questions in this research study.

3.3 Data Analysis techniques

Data analysis is performed for presenting accurate and reliable data for successfully completing a research study. In this current research study, secondary data is collected from online databases and this type of information is further analyzed with the help of thematic analysis. Thematic analysis is the one that allows the selection of different patterns and themes from collected data so that accurate and valuable information is used for achieving the research objectives. The current research study is based upon the identification of Machine learning based approaches for predicting life expectancy of European countries and continents;

information is collected for the same and required to be analyzed by creating different themes. Some of the themes may include, “Importance of predicting life expectancy”, “Methods to predict life expectancy in European countries” and “Machine Learning approaches to predict life expectancy”. The information collected using these themes help in providing in-depth knowledge regarding the research topic which further ensures the successful achievement of designed research aims and objectives.

In addition to this, experimental analysis is performed on python programming environment using jupyter notebook and Anaconda software as well as two Machine Learning approaches are used including Linear Regression and Random Forest regressor. The performance of both the approaches are evaluated using accuracy and comparison among two are made to recommend the best one for future predictions. In addition to this, Tableau is also used for creating visualizations of chosen dataset and achieve the desired findings for the given research study.

To complete this research study different steps are followed below which include importing libraries, dataset loading, exploratory data analysis, data wrangling, data visualization, Label Encoding, Feature Engineering and Model implementation. Also, the results obtained from experimental analysis are supported by the existing literature to increase the overall quality of presented content in the research study.

3.3.1 Importing Libraries

The first and most important step is to import libraries which include bundles of code to be used repeatedly in different programs. They also help in providing flexibility and functionality for any task as well as help in structuring the code in an efficient manner. numpy and pandas are the two commonly used libraries while working with data. The other visualization libraries also imported for the analysis process are plotly.express, seaborn and matplotlib.pyplot is shown in the below figure 4.

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
!pip install plotly
import plotly.express as px
```

Figure 4.Screenshot of python code for importing libraries

3.3.2 Loading dataset

```
pd.set_option('display.max_columns', None)
df=pd.read_csv("LIFE_EXPECTANCY.csv")
```

Figure 5. Screen shot of python code for loading dataset

After importing the required libraries, we import our data set from source file to data frame with pandas **read_csv ( )** method to read the dataset csv file.

3.3.2 Exploratory data analysis

Further exploratory data analysis is performed with the use of different functions including head (for showing top five rows of the dataset), info (to provide general information regarding the dataset), shape (for providing numbers to rows and columns), columns (to extract all the columns of the dataset) and describe (to calculate mean, standard deviation, min, and max values for every identified attribute of the dataset) are explained in chapter 4.

3.3.3 Data wrangling

The next step is to perform data wrangling also known as data cleaning for transforming the raw data into easily understandable format. To do so, different functions are used such as isna (to check the null values), drop (to eliminate the unwanted columns from the dataset) and corr (to show the correlation).

3.3.4 Data visualization

Further, data visualization is performed by presenting a heatmap, graphs, scatter plots and box plots regarding the selected dataset values. In addition to this, tableau is also used for creating visualizations using which clear insights are found for the chosen dataset with the use of different types of graphs.

3.3.5 Label encoding

The next step is to perform label encoding in which the labels identified from the dataset are converted into numerical values to make them into a machine-readable form.

3.3.6 Feature Engineering

Further, feature selection is performed based on which the results are obtained using relevant and accurate data values. The dataset selected in this experiment includes various factors that help in predicting life expectancy of the European continent including infant Deaths, adult Mortality, Alcohol, Hepatitis B, Measles, Diphtheria, Polio, and various others. Splitting of data is performed in which the dataset is divided into training and testing data for the purpose of training the model and evaluating the data, respectively.

3.3.7 Model Implementation

Following this, model implementation is performed in which two different Machine Learning based models are developed, Linear Regression and Random Forest regressor.

3.3.8 (1) Linear Regression

Linear regression is one of the supervised algorithms and is an effective Machine learning model to predict the targeted values based on independent variables using a correlation coefficient. Linear regression fitted a linear model among the explanatory variables and target variable by minimize the residual sum of squares.

The simplest form of Equation of a linear regression is given by the formula $y = mx + c$, where c is the intercept, y is the independent variable and x is the dependent variable and m is the regression coefficient. The general plot for a linear regression model is depicted in the below figure 6.

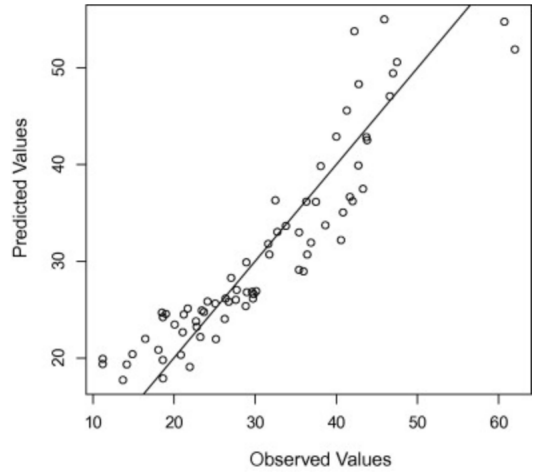

Figure 6.plot for linear regression (google-https://www.sciencedirect.com/)

3.3.8 (2) Random Forest Regression Model

Random Forest Regression is a supervised machine learning technique using ensemble of decision trees algorithms for predictive analytics. This regression is used in both prediction and classification task using the technique known as bagging. The general plot of a random forest regressor model is shown in the figure 7.

In a random forest regressor, prediction is the average of individual prediction of each trees produced, which increase models' accuracy and reduce overfitting. Moreover, it works with missing data by creating estimates, that's why random forest produces more accurate results than other regression models. It is easy to use and scales well when we added new features to the data set.

The Random Forest regressor Model algorithm consists of two step process. The first process is built decision tree estimators and the second process takes the average of prediction as final output.

We can improve the performance of Random Forest regressor by specify maximum number of features to be included in each node, specify the maximum depth of the trees and or increase or decrease no of estimators.

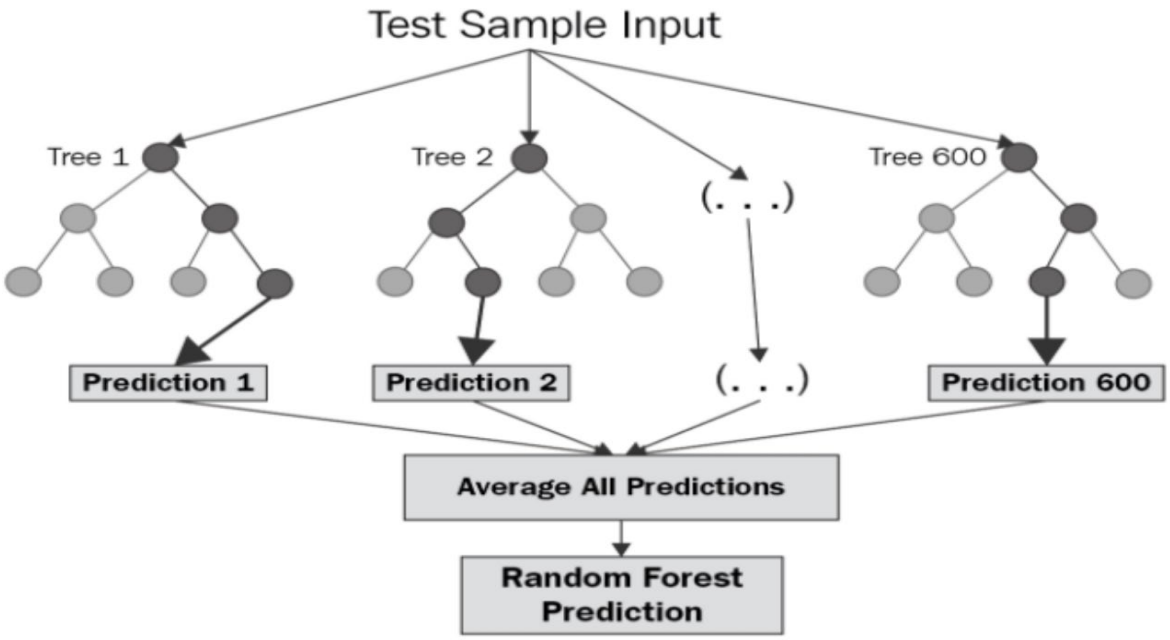

Figure 7.general plot for random forest regressor(google-https://levelup.gitconnected.com/)

3.3.9 Model Evaluation

These models are further evaluated using different performance metrics to understand performance of a model. The one with highest accuracy is recommended for further predictions regarding life expectancy. Evaluation Metrics for a regression model are Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared or Coefficient of Determination.

These steps when followed in a sequential manner helps in successfully completing the experimental analysis by providing accurate results.

Chapter 4. Analysis and findings

The analysis performed below has been majorly focused on the prediction on life expectancy rate of European Countries according to their GDP and some other explanatory variables such as population, income composition, infant deaths, adult mortality etc., Based on these factors, the life expectancy of the European Countries has been analysed using machine learning algorithms. During the analysis process the major processes includes the data understanding and exploration, data cleaning, data visualization, and model implementation. using which the aim of the analysis process to predict the life expectancy based on explanatory variables to analyse the impact of Covid-19 Pandemic on the life expectance of the European Countries, has been fulfilled.

4.1 Data Understanding and Exploration

4.1.1 Dataset collection

The data set for life expectancy was extracted from the data science related environment website Kaggle (https://www.kaggle.com/datasets), where health factors for 193 countries has been collected from the WHO data repositories website and economic data collected from the United Nations website.

Some of the details, that has been added in the data set, is gathered from various sources and then updated and sorted in the given data set such as the life expectancy rates of the 35 European countries from year 2016 to 2020 and the GDP and Population of these countries for last five years (2016-2020) has been gathered from external sources. Finally formed a new dataset of life expectancies of European countries.

4.1.2 Data Understanding

The data set used for the analysis process contains the data of various European countries from the year 2000 to 2020 containing 716 rows and 22 columns providing various stats of these countries such as- name of the country, year for which the stats has been provided, development status, life expectancy rate, adult mortality rate, infant death rate, alcohol, expense percentage, Hepatitis B cases, measles cases, BMI, under five deaths, Polio cases, total expenditure, Diphtheria, HIV/AIDS, GDP, population, thinness 1-19 years, thinness 5-9 years, income composition of resources and schooling.

4.1.3 Dataset Description

Attributes	**Description**
Country	Name of the countries
Year	Year
Status	Development status (currently being developed or already developed)
Life expectancy	number of years an individual is expected to live
Adult Mortality	Mortality rate for adults of both sexes
Infant death	The number of infant deaths per 1000 population
Alcohol	Per capita Alcohol consumption (in litres of pure alcohol)
Percentage Expenditure	Healthcare expenditure as a percentage of Gross Domestic Product per capita (%)
Hepatitis B	Immunization coverage against hepatitis B against one-year-old children (%)
Measles	Number of reported cases per thousand population
BMI	The average body weight of the entire population
Under-five-deaths	The number of deaths under the age of five per 1000 population
Polio	Anti-polio-coating among one-year-old children (%)
Total Expenditure	National health expenditure as a percentage of total public expenditure (%)
Diphtheria	Coverage by immunoprophylaxis against tetanus pertussis among one-year-old children (%)
HIV/AIDS	Deaths per 1000 live births HIV-AIDS (0-4 years)
GDP	Gross Domestic Product per Capita (In US dollars)
Population	The population of a country
Thinness 1-19 years	Prevalence of weight loss among children and adolescents aged 10 to 19 years (%)
Thinness 5-9 years	Prevalence of weight loss among children aged 5 to 9 years (%)
Schooling	Numbers of years of Schooling

Table 1.Variables description

The above table.1 describes all the observed attributes in the data set and its corresponding descriptions.

4.2 Exploratory Data Analysis

It is an important part of the data analysis process which includes summarizing the characteristics of the data using visualization tool and to explore these characteristics to prepare the data for fitting model by eliminating or replacing the null values or missing values, evaluating the statistical distribution using statistical graphs. In this analysis process, the procedures included in the EDA process are checking the number of columns and rows, extraction of column names, calculation of the statistical values (mean, median, mode, standard deviation, minimum, maximum, 25%, 50% and 75% values), data wrangling (null values checking, checking the skewness, and replacing the null values).

All the procedures discussed above has been described below with the relevant evidence of the code and its output. The major aim of the process is to predict the life expectancy of all the European Countries whose data has been considered based on various factors.

From the analysis of over 35 European countries, it is effectively represented that how the life expectancy has been changed over the years based on all these factors discussed above. The step-by-step analysis procedure has been given below.

4.2.1 Exploratory data analysis for information of data

a) In the exploratory data analysis process, the general information regarding the data set has been displayed such as the top five rows using the head () method shown in Figure 8.

```
df.head()
```

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure
0	Albania	2016	Developing	78.194	96.726103	1.309524	9.553424	2335.831812	83.585434	550.584249	52.874816	1.564103	93.553704	7.225361
1	Albania	2017	Developing	78.333	96.726103	1.309524	9.553424	2335.831812	83.585434	550.584249	52.874816	1.564103	93.553704	7.225361
2	Albania	2018	Developing	78.458	96.726103	1.309524	9.553424	2335.831812	83.585434	550.584249	52.874816	1.564103	93.553704	7.225361
3	Albania	2019	Developing	78.573	96.726103	1.309524	9.553424	2335.831812	83.585434	550.584249	52.874816	1.564103	93.553704	7.225361
4	Albania	2020	Developing	78.686	96.726103	1.309524	9.553424	2335.831812	83.585434	550.584249	52.874816	1.564103	93.553704	7.225361

Figure 8.the top five rows of a dataset

b) check the shape of our data using df.shape property shown Figure.9, which determine the dimension of the dataset containing 716 rows and 22 columns.

```
df.shape
```

```
(716, 22)
```

Figure 9.Shape of a dataset

c) data columns, not null count of each column along with the details of the data type of each attributes using info () is shown in Figure 10. there are 14 columns with null values and the remaining 8 columns has no null values, while the data type of two columns are object, for one column it is integer while the remaining has float64 data type. After checking this general information, the number of rows, columns and the name of the columns are also displayed as a part of data exploration process

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 716 entries, 0 to 715
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          716 non-null    object
 1   Year                             716 non-null    int64
 2   Status                           716 non-null    object
 3   Life expectancy                  714 non-null    float64
 4   Adult Mortality                  714 non-null    float64
 5   infant deaths                    716 non-null    float64
 6   Alcohol                          684 non-null    float64
 7   percentage expenditure           716 non-null    float64
 8   Hepatitis B                      527 non-null    float64
 9   Measles                          716 non-null    float64
 10   BMI                             714 non-null    float64
 11  under-five deaths                716 non-null    float64
 12  Polio                            710 non-null    float64
 13  Total expenditure                683 non-null    float64
 14  Diphtheria                       710 non-null    float64
 15   HIV/AIDS                        716 non-null    float64
 16  GDP                              682 non-null    float64
 17  Population                       682 non-null    float64
 18   thinness  1-19 years            714 non-null    float64
 19   thinness 5-9 years              714 non-null    float64
 20  Income composition of resources  698 non-null    float64
 21  Schooling                        699 non-null    float64
dtypes: float64(19), int64(1), object(2)
memory usage: 123.2+ KB
```

Figure 10.Summary of a data frame using info ()

4.2.2 Statistical Exploration

The statistical values are also evaluated using the describe () function. In this step, we can determine the value of count, mean, standard deviation, maximum, minimum etc of all the features. From the below figure11. we can identify some attributes have outliers by analysing through the difference between their mean and maximum values. Using boxplot visualization technique, we can determine outliers of each dependent variables, its range and distribution.

```
df.describe()
```

	Year	lifeExpectancy	adultMortality	infantDeaths	Alcohol	percentageExpense	Hepatitis B	Measles	BMI
count	716.000000	714.000000	714.000000	716.000000	684.000000	716.000000	527.000000	716.000000	714.000000
mean	2010.008380	78.158716	96.726103	1.309524	9.553424	2335.831812	83.585434	550.584249	52.874816
std	6.053139	4.586313	57.114120	2.629231	2.784504	3159.041975	20.651400	2255.025888	13.191692
min	2000.000000	64.600000	1.000000	0.000000	0.010000	0.000000	2.000000	0.000000	5.100000
25%	2005.000000	75.000000	68.000000	0.000000	9.145000	193.418201	83.585434	1.000000	52.874816
50%	2010.000000	78.339671	96.726103	1.000000	9.553424	1536.066875	86.000000	39.000000	55.200000
75%	2015.000000	81.545732	113.000000	1.309524	11.200000	2335.831812	96.000000	550.584249	58.500000
max	2020.000000	89.000000	327.000000	22.000000	17.870000	19479.911610	99.000000	42724.000000	69.600000

Figure 11.calculate the mean, std, min, max and count of every attributes

4.3 Data Wrangling

4.3.1 Check for null values

It is an important part of the EDA process where the errors and complex data sets are simplified and prepared for the analysis process. In the given process, the null values have been checked for each column of the data set and the results is shown the figure 12.

As displayed in the results, 14 columns (adult mortality, infant deaths, alcohol, Hepatitis B, BMI, Polio, Total Expenditure, Diphtheria, GDP, population, thinness 1-19 years, thinness 5-9 years, income composition of resources and schooling) in our data frame has null values, where Hepatitis B column has the highest number of null values. Null values can impact the out or the prediction accuracy while the model implementation process as it represents the mission or errors values in the data set which needs to be removed or replaced for better results.

```
# check for null values
df.isna().sum()
```

```
Country                             0
Year                                0
Status                              0
lifeExpectancy                      2
adultMortality                      2
infantDeaths                        0
Alcohol                            32
percentageExpense                   0
Hepatitis B                       189
Measles                             0
BMI                                 2
under-five deaths                   0
Polio                               6
Total expenditure                  33
Diphtheria                          6
HIV/AIDS                            0
GDP                                34
Population                         34
thinness  1-19 years                2
thinness 5-9 years                  2
Income composition of resources    18
Schooling                          17
dtype: int64
```

Figure 12.Result after checking null values in the data set

4.3.2 Checking Skewness

```
# Checking skewness
visDf=df.loc[:,'Country':'Schooling']
visDf=visDf.select_dtypes([np.int, np.float])
for i, col in enumerate(visDf.columns):
    print(f"\nSkewness of {col} is {df[col].skew()}")
```

Figure 13.Screenshot of python code for checking Skewness

Skewness is the measure of deviation of distortion from its normal distribution. It is the distribution of the data which defines the direction of the outliers whether positive, negative or zero. The right skewed distribution is positive, the left skewed distribution is negative while the central skewed distribution is zero skew. As per this, the skewed values, or the level of skewness of all the columns has been checked in this step to determine how to replace the null values in the data set. The result of the skewed values of all the columns are given below figure.14

```
Skewness of Year is -0.0038233216246554014

Skewness of lifeExpectancy is -0.2444477529508443

Skewness of adultMortality is 1.0881046630478037

Skewness of infantDeaths is 5.0227871287689485

Skewness of Alcohol is -1.1725770733064136

Skewness of percentageExpense is 2.51095586944433

Skewness of Hepatitis B is -2.5901454157726316

Skewness of Measles is 12.037273807060107

Skewness of BMI is -2.9346997785500446

Skewness of under-five deaths is 5.0384905313231805

Skewness of Polio is -6.634406516316393

Skewness of Diphtheria is -6.042545615036075

Skewness of HIV/AIDS is 7.907971931487196

Skewness of GDP is 1.463095197642917

Skewness of Population is 3.4275556226546833

Skewness of thinness  1-19 years is 0.5250265994625504

Skewness of thinness 5-9 years is 0.564087807056684

Skewness of Income composition of resources is -5.19377780414259

Skewness of Schooling is -3.5883618760353744
```

Figure 14.Result of Skewness of attributes

Based on the skewness values of each column, the null values are replaced with its mean and median using the code given in the appendix. all the null values are replaced with mean and median. The null values in the population and GDP columns are replaced with their median values while the others are replaced with their mean values evaluated during the EDA process. As a result of this replacement, the null values in all the columns are eliminated as displayed below figure15.

```
# check for null values
df.isna().sum()
```

```
Country                            0
Year                               0
Status                             0
lifeExpectancy                     0
adultMortality                     0
infantDeaths                       0
Alcohol                            0
percentageExpense                  0
Hepatitis B                        0
Measles                            0
BMI                                0
under-five deaths                  0
Polio                              0
Diphtheria                         0
HIV/AIDS                           0
GDP                                0
Population                         0
thinness  1-19 years               0
thinness 5-9 years                 0
Income composition of resources    0
Schooling                          0
dtype: int64
```

Figure 15.Replacement of null values

4.4 Data visualization

4.4.1 Data visualization using Correlation matrix

Correlation Matrix is a two-dimensional plot using heatmaps in seaborn visualization tool, which tells correlation between pairs of variables in each dataset. The correlation values close to one is highly correlated, zero has no linear relationship between the variables, less than zero is negatively correlated.

We can identify from the figure, under-five-deaths with infant deaths and thinness 1-19 years with thinness 5-9 years are strongly correlated with each other. We can analyse how much all the attributes are correlated to each other using df.corr () in the below depicted correlation plot in the figure 16.

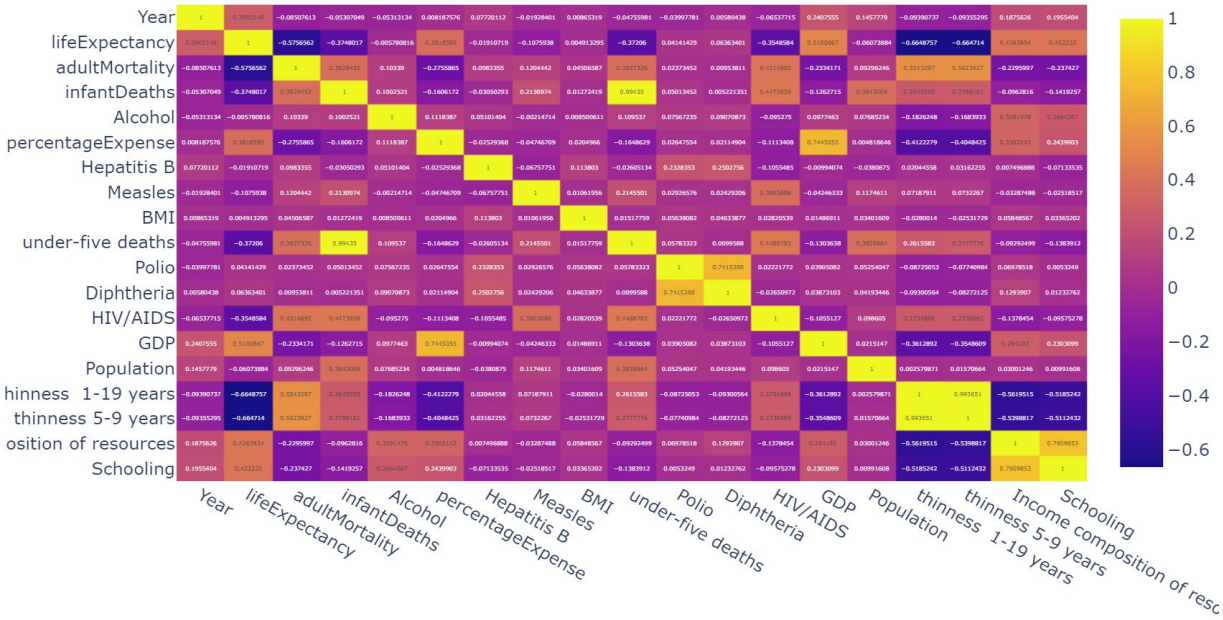

Figure 16.Correlation plot using heatmaps

4.4.2 Data visualization using Box plot

A box plot or a box-and-whisker diagram displaying variables distribution using five-number-theory. The purpose of the box plot is to identify the outliers and is removed from the attributes before fit to the model, because outliers which affects the accuracy of the result. The basic box-plot diagram is shown in the below figure 17.

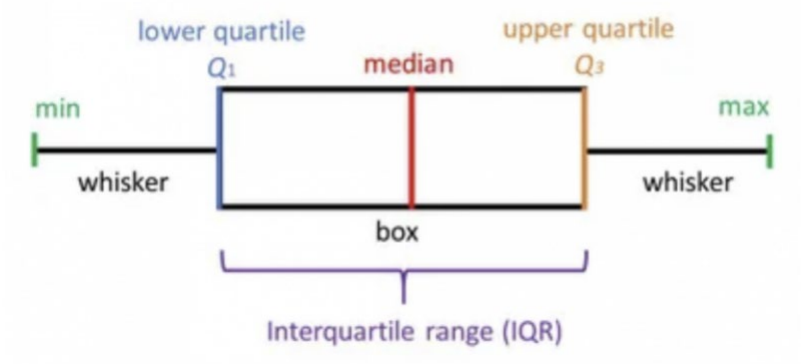

Figure 17.The basic box-plot diagram (https://www.isixsigma.com/)

Min is the minimum value in the dataset, max is the highest value in the dataset. lower quartile Q1 is the first quartile value, which is the median of the lower half of the dataset. The median

Q2 is the middle value of the dataset. Upper quartile Q3 is the upper half of the dataset. Interquartile range is the difference between upper quartile and lower quartile.

The box plot of every attribute showing the minimum, maximum, mean, median, 25%, 50% and 75% values are shown in the below figures.

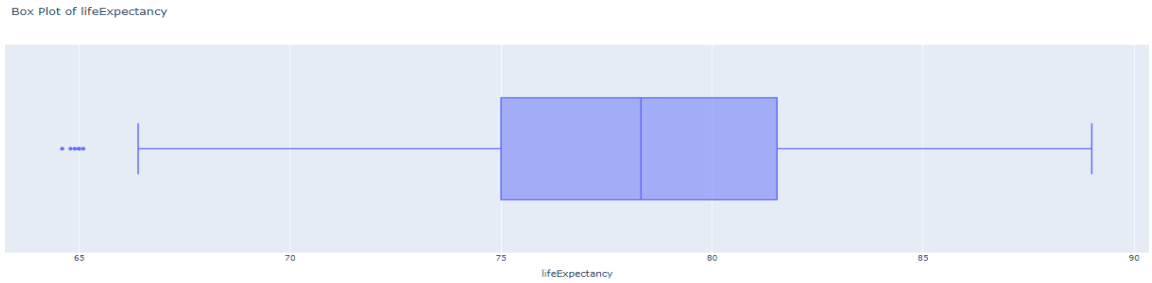

Figure 18: Life expectancy box plot

The box plot for the life expectancy rate in European countries shows the ideal life expectancy rate in these countries which is between 75 to 85 as most of the countries has their life expectancy between thus ranges which is an effective rate.

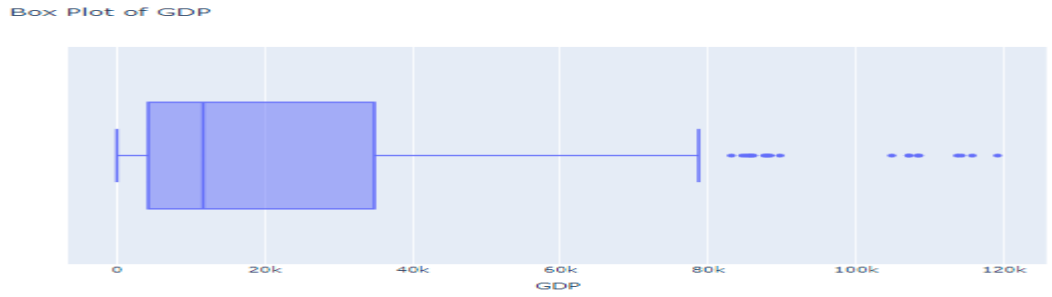

Figure 19: GDP box plot

The box plot of GDP displays that the highest GDP of the European countries are above 100k but most of the countries has the GDP between 20k to 40k. Life expectancy has a major impact on the GDP of the country as the higher will be the life expectance higher will be the workable human capital availability in the country resulting in higher returns.

Similarly, the box plots have been created for all the attributes of the data set defining their minimum, maximum and the median values. In the next step, the outliers in the box plots (the values or data that completely differs from other observations) are removed using the code given below in the figure20.

```
# Removing outliers
for i, col in enumerate(visDf.columns):
  Q1 = df[col].quantile(0.25)
  Q2 = df[col].quantile(0.50)
  Q3 = df[col].quantile(0.75)

  IQR = Q3 - Q1
  print('\nInterquartile range of',col, 'is: %.2f'% IQR)
  whisker_width = 1.5
  low_lim = Q1 - (whisker_width * IQR)
  up_lim = Q3 + (whisker_width * IQR)
  print('Lower Limit of',col,'is: %.2f'%low_lim)
  print('Upper Limit of',col,'is: %.2f'%up_lim,'\n')

  try:
    df[col]=np.where(df[col]>up_lim,up_lim,np.where(df[col]<low_lim,low_lim,df[col]))
  except:
    print('\t\nUnable to remove an outlier for:', col)
```

Figure 20.Code for removing outliers

The interquartile, lower, and upper range of all the data attributes has been evaluated and then the outliers are eliminated.

Interquartile range of Year is: 10.00
Lower Limit of Year is: 1990.00
Upper Limit of Year is: 2030.00

Interquartile range of life Expectancy is: 6.53
Lower Limit of life Expectancy is: 65.20
Upper Limit of life Expectancy is: 91.34
Interquartile range of adult Mortality is: 45.00
Lower Limit of adult Mortality is: 0.50
Upper Limit of adult Mortality is: 180.50
Interquartile range of infant Deaths is: 1.31
Lower Limit of infant Deaths is: -1.96
Upper Limit of infant Deaths is: 3.27
Interquartile range of Alcohol is: 1.82
Lower Limit of Alcohol is: 6.55
Upper Limit of Alcohol is: 13.84
Interquartile range of percentage Expense is: 2142.41
Lower Limit of percentage Expense is: -3020.20

Upper Limit of percentage Expense is: 5549.45

Interquartile range of Hepatitis B is: 10.66

Lower Limit of Hepatitis B is: 67.59

Upper Limit of Hepatitis B is: 110.25

Interquartile range of Measles is: 549.58

Lower Limit of Measles is: -823.38

Upper Limit of Measles is: 1374.96

Interquartile range of BMI is: 5.63

Lower Limit of BMI is: 44.44

Upper Limit of BMI is: 66.94

Interquartile range of under-five deaths is: 1.56

Lower Limit of under-five deaths is: -2.35

Upper Limit of under-five deaths is: 3.91

Interquartile range of Polio is: 3.45

Lower Limit of Polio is: 88.38

Upper Limit of Polio is: 102.17

Interquartile range of Diphtheria is: 4.00

Lower Limit of Diphtheria is: 87.00

Upper Limit of Diphtheria is: 103.00

Interquartile range of HIV/AIDS is: 0.02

Lower Limit of HIV/AIDS is: 0.07

Upper Limit of HIV/AIDS is: 0.14

Interquartile range of GDP is: 30583.46

Lower Limit of GDP is: -41628.14

Upper Limit of GDP is: 80705.72

Interquartile range of Population is: 7875407.00

Lower Limit of Population is: -11310930.25

Upper Limit of Population is: 20190697.75

Interquartile range of thinness 1-19 years is: 1.00

Lower Limit of thinness 1-19 years is: -0.60

Upper Limit of thinness 1-19 years is: 3.40

Interquartile range of thinness 5-9 years is: 1.10

Lower Limit of thinness 5-9 years is: -0.75

Upper Limit of thinness 5-9 years is: 3.65
Interquartile range of Income composition of resources is: 0.07
Lower Limit of Income composition of resources is: 0.69
Upper Limit of Income composition of resources is: 0.98
Interquartile range of Schooling is: 1.30
Lower Limit of Schooling is: 12.85
Upper Limit of Schooling is: 18.05

The outliers are determined by calculating the interquartile range for each attribute and then the identified outliers were removed. After removing the outliers, the box plots are recreated with all the variable values are eliminated.

4.4.3. Data visualization using Scatterplot (life expectancy with Explanatory variables)

Scatterplot is a diagram in which values are depicted using cartesian coordinates. According to our dataset, Life Expectancy is considered as the independent variable and all the other parameters such as population, GDP, adult mortality rate, infant deaths etc. Considered to be dependant variable. Based on these visualizations, the impact on life expectancy rate of various factors and how it has been changed over the years in all the 35 European Countries has been determined and helped in predicting the life expectancy rate of these countries for the coming years based on the trends of last 20 years.

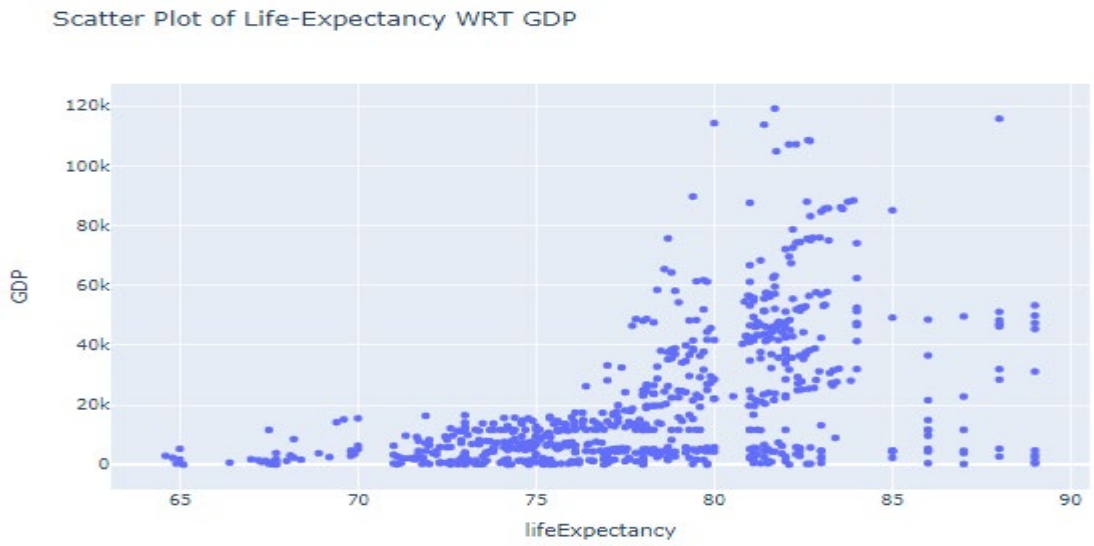

Figure 21. Plot showing ideal life expectancy rate for countries with highest GDP

The figure 21. above illustrates a scatter plot of life expectancy in European countries with respect to their GDP per capita income. As per the scatter plot, the GDP and life expectancy are somehow interdependent as decrease in the GDP, the less will be the life expectancy rate or in other words, the less life expectancy can result in lower GDP. Also, the plot given above shows that the ideal life expectancy rate for country with highest GDP is between 80-85. This shows that, the country with a higher or satisfactory GDP will have a life expectancy rate between 80-85 or the life expectancy rate of the country between 80-85 will have a better GDP. Higher the life expectancy, lower will be the mortality rate, higher will be the human capital availability and higher will be the productive returns to the country and thus higher will be the GDP per capita income of the country.

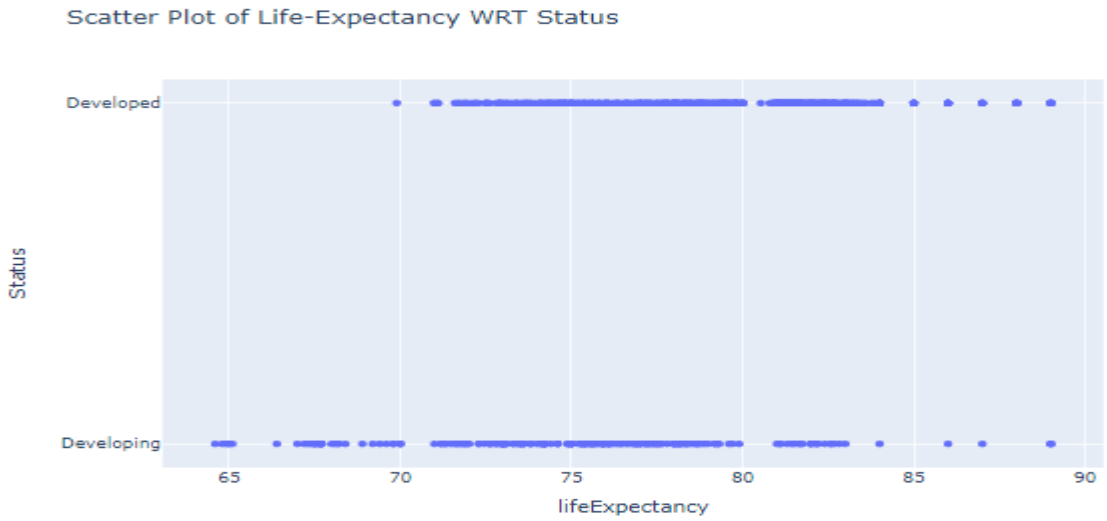

Figure 22. Scattered plot for life expectancy of developed and developing countries

Another factor that can be impactful for the life expectancy rate of the European country that has been considered in this process is the development status of the countries. It is obvious that the developed countries have more sources, established structure and workflow helping them in providing better services or facilities to their citizens while the developing countries are emerging with new steps taken every day to make the better. Based on the scatter plot given above, the developed countries have life expectancy higher than 70 while the number of developing countries with below 70 life expectancy rate is also high. This can be considered as a factor for the life expectancy rate because of the services, facilities sources, and availability of resources.

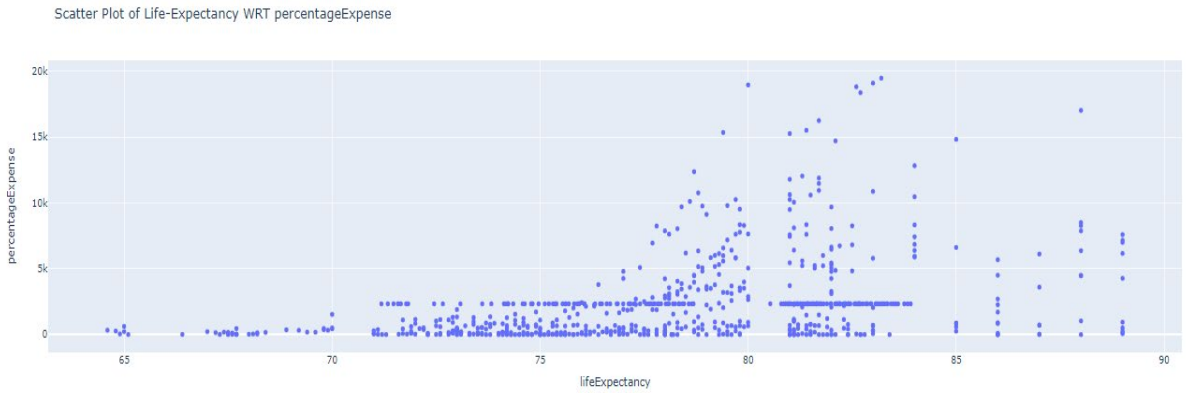

Figure 23. Scattered plot to define expense % of EU countries

The scatter plot in the figure 23. illustrates the expense percentage of the European countries and its impact on their life expectancy rate. The countries with highest rate of expenses percentage have the life expectancy rate between 80-85 showing that the higher rate of expense percent can be a factor that impacts the life expectancy rate of the country or life expectancy rate of any country can result in higher expense percent. The more will be the availability of workable human capital, the more will be the income and more will be the expenses.

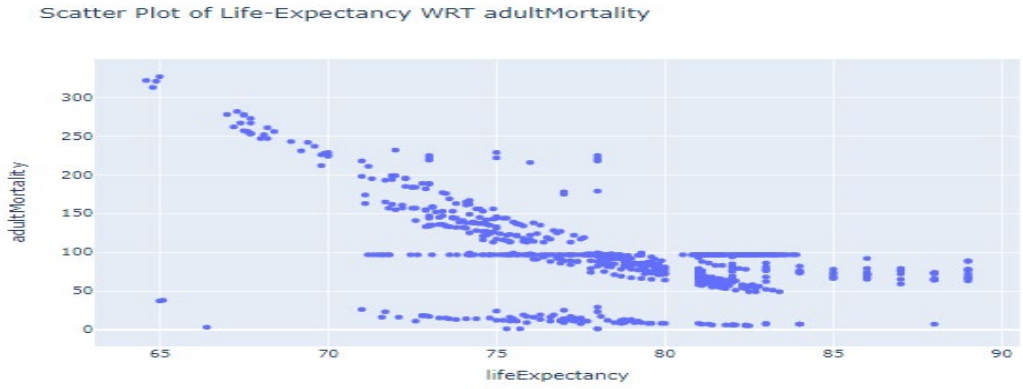

Figure 24. Life expectancy and adult morality scatter plot

Adult mortality is an important factor on which the life expectance of the country depends and the details of the 35 countries that were analysed shows that the least adult mortality rate can result in a better or higher life expectancy rate of the country, and it will be helpful for them to have the adults or workable human capital in large numbers to get higher returns and more GDP of the country.

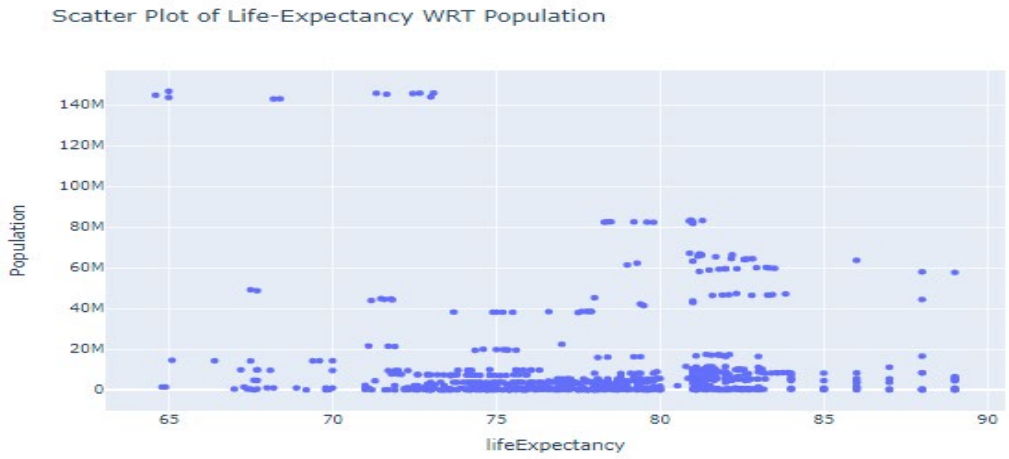

Figure 25. Life expectancy and population scatter plot

Based on the population of the countries, the life expectancy rate can be changed as displayed in the scatter plot in the above figure 25, which illustrates that the country with lower population has the higher or ideal life expectancy rate. The increase in population will also be increasing the demands, the diseases, the problems and thus impacting the way of living of the people. This can be an impacting factor for the life expectancy rate of the country.

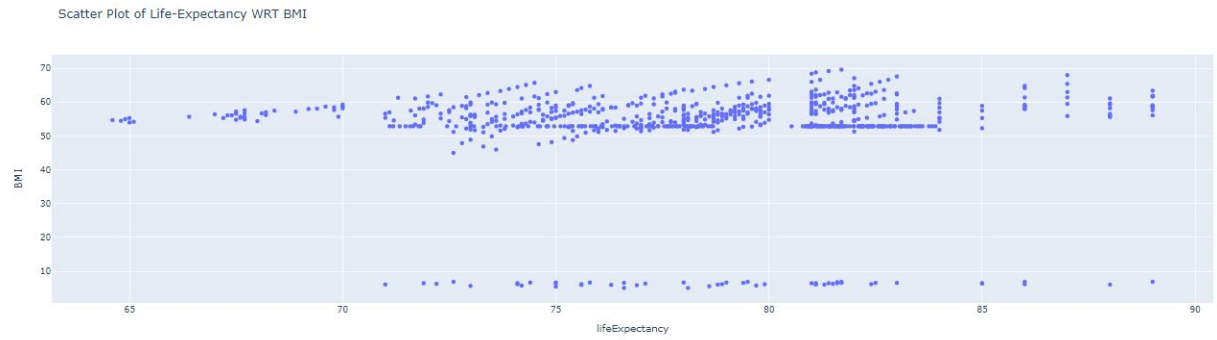

Figure 26. Scatter plot of EU countries with different BMI

The Body mass index can also be a factor that might impact the life expectancy in countries as it shows how much the citizens of the country are overweight and underweight, increasing their chances of being affected by any disease or other health related problem. Thus, this can be a factor on which the life expectancy of the country depends. As per the scatter plot of European countries given above, most of the countries has higher BMI but still can maintain the life expectancy rate above 80 that shows that it is a factor based on which life expectancy can be predicted but it is not entirely dependent on it.

4.4.4. Data visualization using Scatterplot (life expectancy with year-Country wise)

In the next step, country wise scatter plots have been created for life expectancy rate over the course of 20 years from 2000 to 2020.

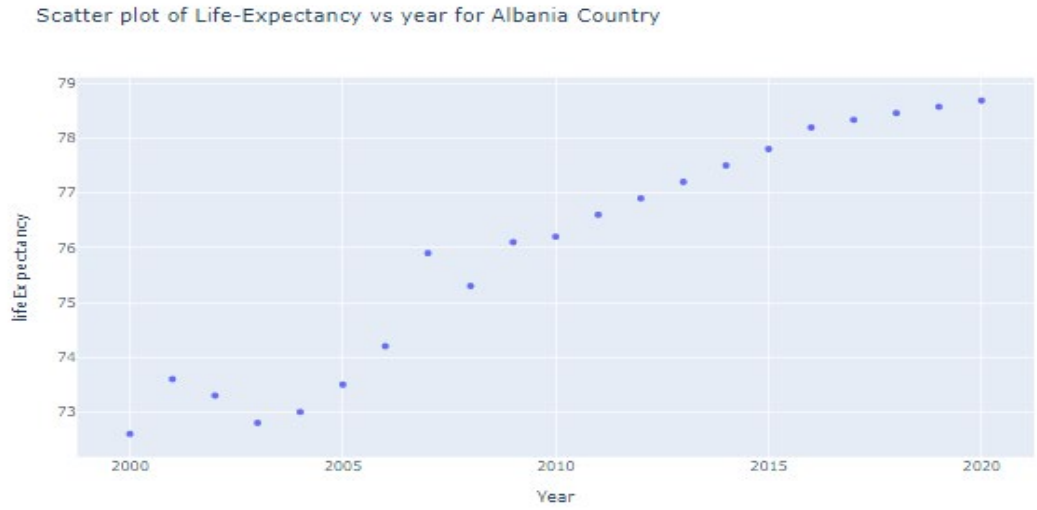

Figure 27. Scatter plot of life expectancy vs year for Albania country

Life expectancy rate of Albania country has been visualized using the scatter plot given above figure 27 from the year 2000 to 2020. It shows that the life expectancy of the country has been increasing since 2005 with a slight dip in the 2008 but after that it has effectively increased even in the year 2019 and 2020 when the Covid pandemic has impacted the whole world.

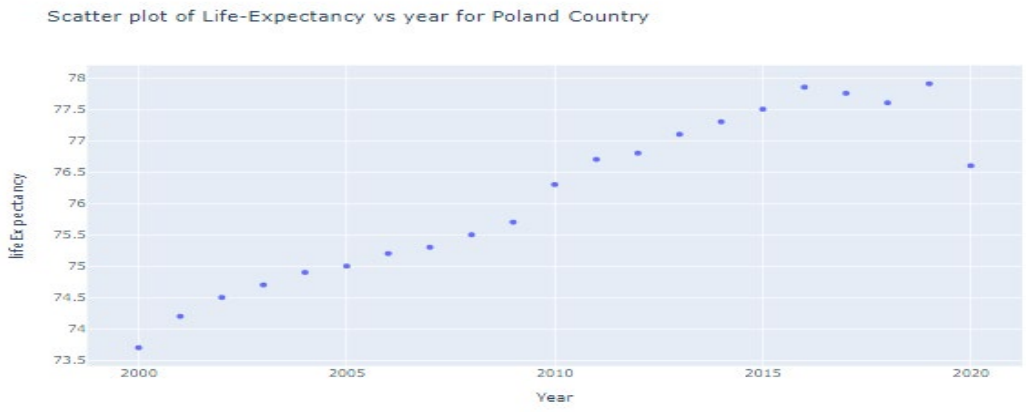

Figure 28. Scatter plot of life expectancy vs year for Poland country

The life expectancy rate of Poland has been increasing over the years, but a drastic fall has been observed in the year 2020 as its rate falls from 78 to 76. The major cause could be the

Covid Pandemic which has impacted the country severely during the period and thus impacted their life expectancy rate. In similar concern, (Marois, Muttarak, and Scherbov, 2020) envisioned that if the death counts or mortalities caused due to covid-19 continues to grow, it would directly impact the period of life expectancy. It can be evident from the previous epidemics such as the 1918 influenza and Ebola virus outbreak in the year 2014, because these catastrophic events have decreased the life expectancy at birth in the USA and Liberia. Also, reports of mortalities caused due to covid-19 have shown that this unprecedented rise in the mortalities may result in the loss of significant years of life.

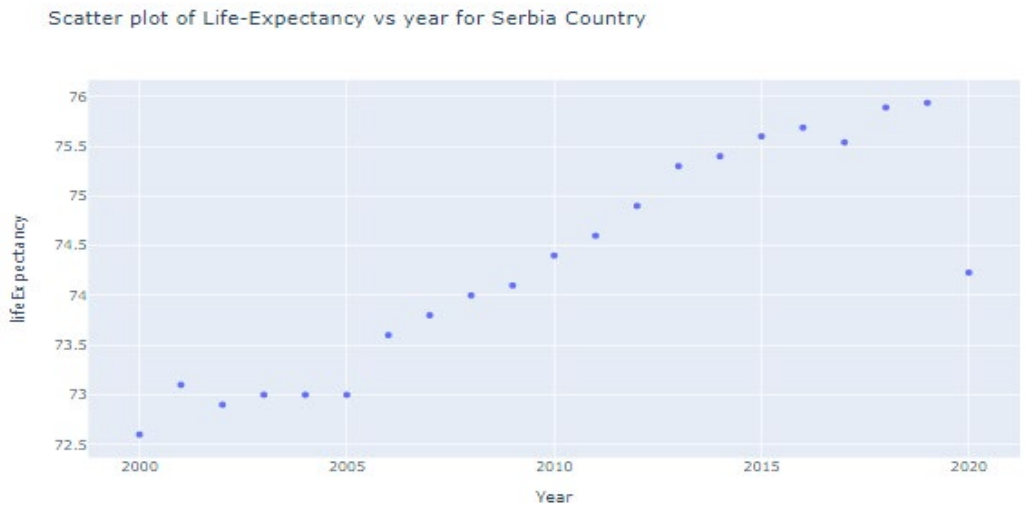

Figure 29: Scatter plot of life expectancy vs year for Serbia country

Serbia's life expectancy rate in the year 2000 was 72.5 and in the year 2019 it has reached an all-time high for the country with 76, but it has drastically fallen to 74.5 in the year 2020 showing the effect of Covid pandemic which caused sever health problems to several people across the country resulting in this drastic decrease in their life expectancy rate.

The scatter plot for all the 35 countries has been created to show their life expectancy rate from the year 2000 to 2020 is shown in the appendix section. Most of the countries has seen a drastic fall in their life expectancy rate in the year 2020, due to various damages the Covid-19 pandemic has caused such as adult mortalities, infant deaths, least expenditure due to pandemic and many others. But this could not be the only factor as some countries has also showed a stable life expectancy rate even at the time of pandemic because of factors such as availability of resources, better services, better management etc. Similarly, (Marois, Muttarak, and Scherbov, 2020) found that with 50% of prevalence of covid-19, life expectancy rate decreased to 3-9 years in the Southeastern Asia and 1-4 years in sub-Saharan Africa. Overall findings

revealed that if the infection prevalence rate remains under 1 or 2%, then the life expectancy would not be affected by the covid-19 pandemic.

4.4.5 Data visualization using Count plot

Count plots are used to count the no of observation in each categorical attributes have. Here, we are depicting count plot of different attributes using seaborn.countplot()

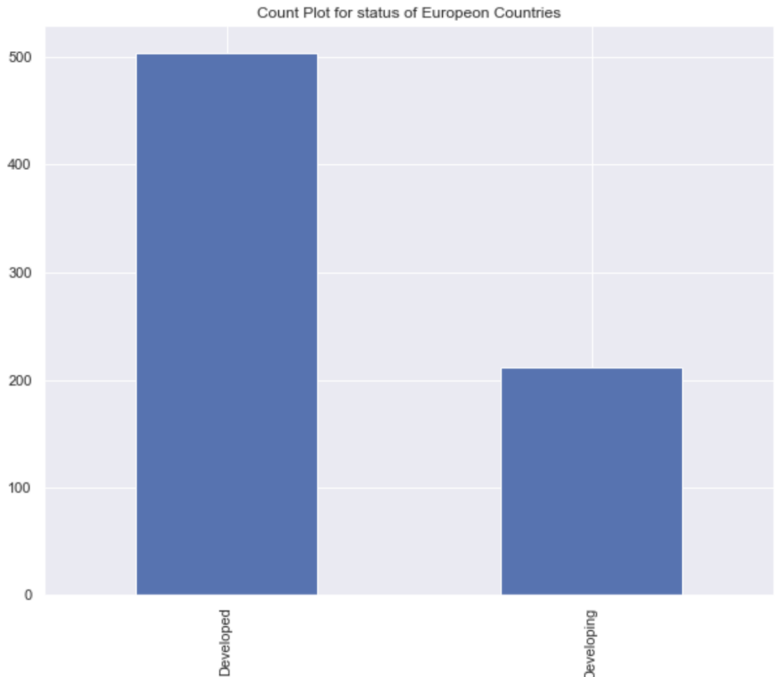

Figure 30.Count plot for status

4.4.6 Data visualization using Tableau

For doing the analysis based on chosen dataset and creating visualizations, Tableau is used. It is a data visualization tool used for converting numerical and textual information to visual dashboards with the use of which users can easily see and understand their data. In the given analysis, Tableau is used to create visualizations regard life expectancy in European countries before and after the Covid-19 pandemic appropriately using graphs.

All the visualizations created with the use of this tool are as follows:

1. **GDP per Country**

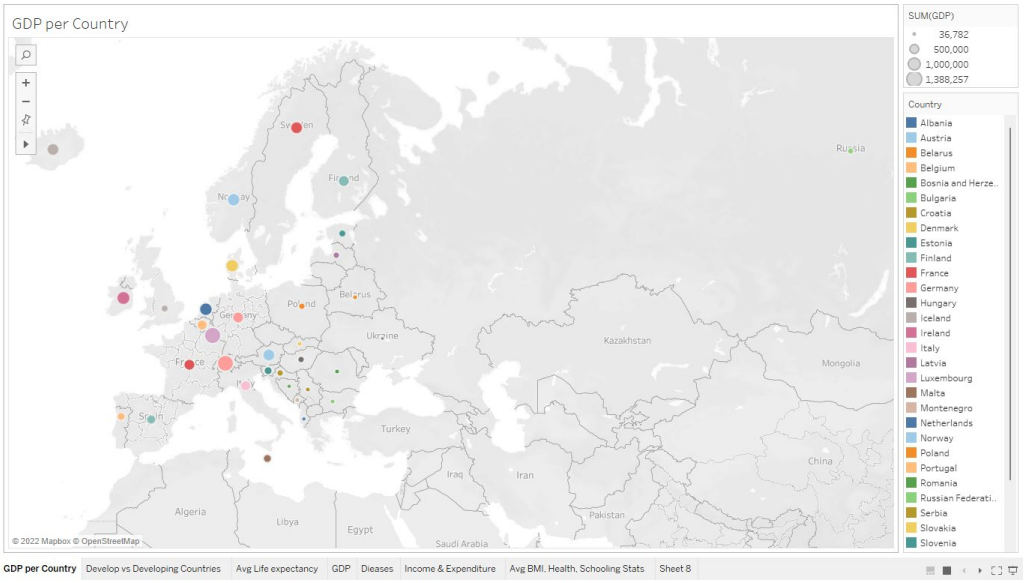

Figure 31. GDP per Country

The above given figure31. shows the visualization based on GDP per country according to which GDP of countries named as Germany, Luxembourg, Ireland, and Albania is high as compared to other countries.

2. **Develop vs Developing Countries**

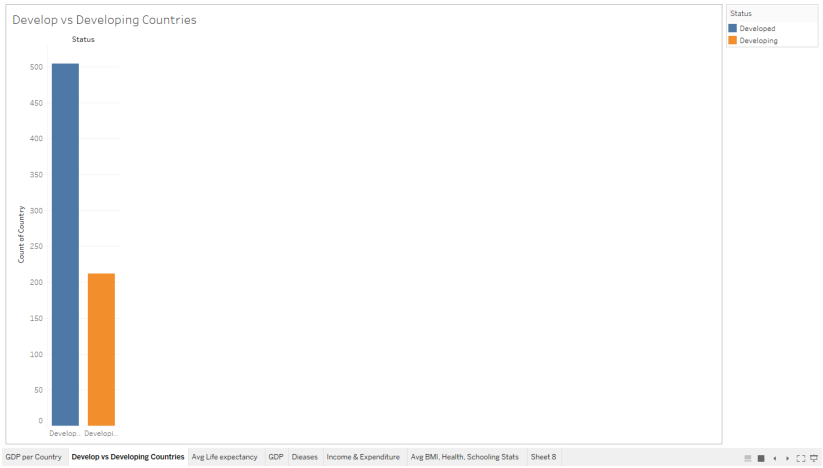

Figure 32. Develop vs Developing Countries

The above figure32. shows that developed countries are more in Europe as compared to developing countries.

3. **Average Life expectancy**

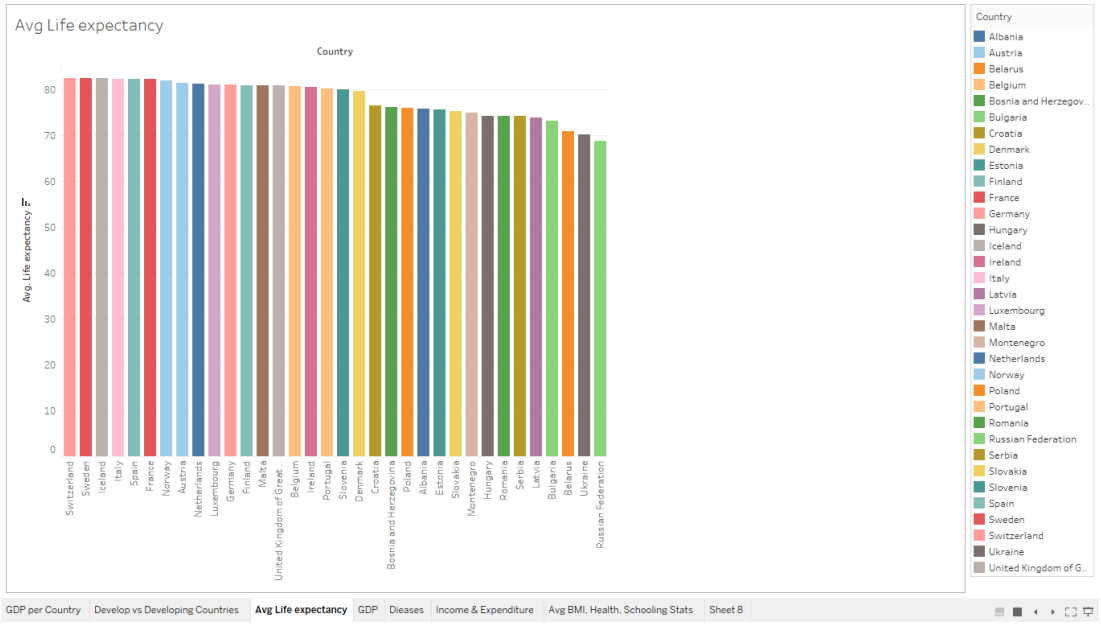

Figure 33. Average Life expectancy

The above figure33. shows the average life expectancy according to which average life expectancy of Switzerland is high whereas life expectancy of Russian Federation is less as comparative to all European countries.

4. **GDP**

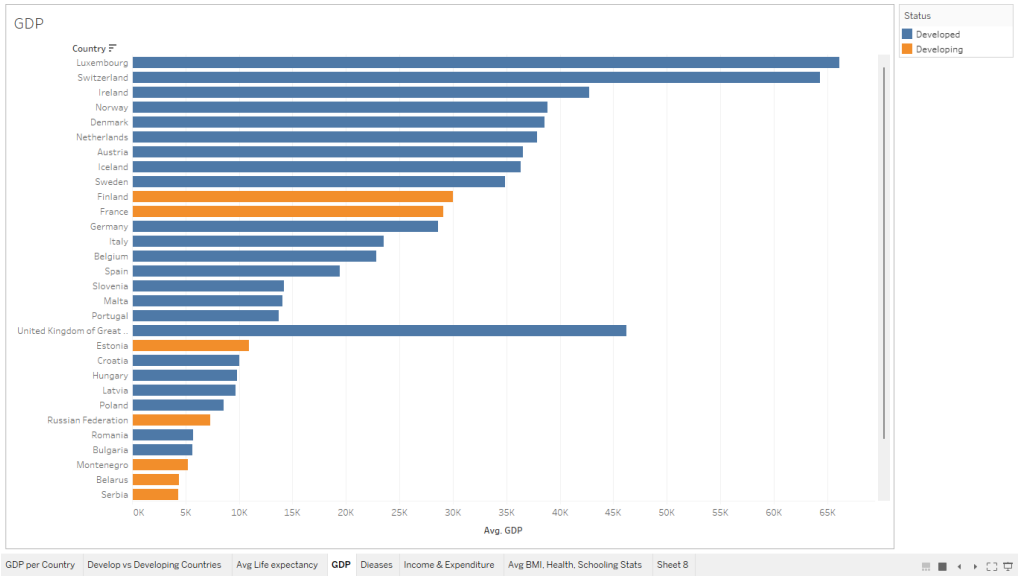

Figure 34.GDP of developed and developing countries

The above figure 34 shows the GDP of developed and developing countries according to which the UK is a developed country having GDPP more than 45K whereas Luxembourg and Switzerland are two other developed countries having GDP more than UK.

5. Diseases

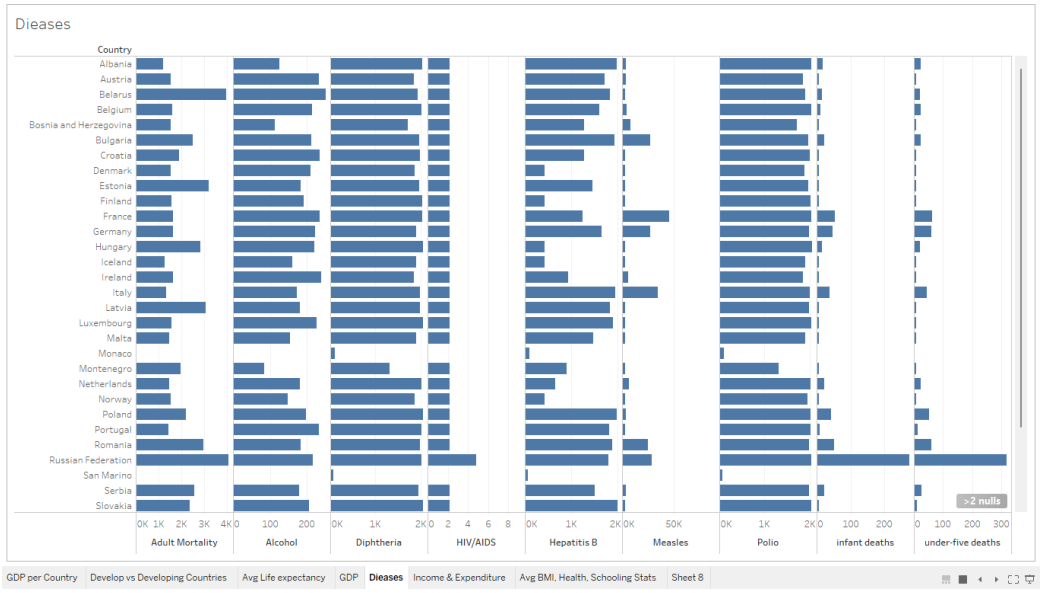

Figure 35. Diseases Analysis

The above given graph in the fig.35 shows different diseases such as adult mortality, measles, polio, infant deaths, under- five deaths, alcohol, and HIV/AIDS. According to this, it has been shown that Russian federation is one of the countries with high rate of under- five deaths

6. **Income & Expenditure**

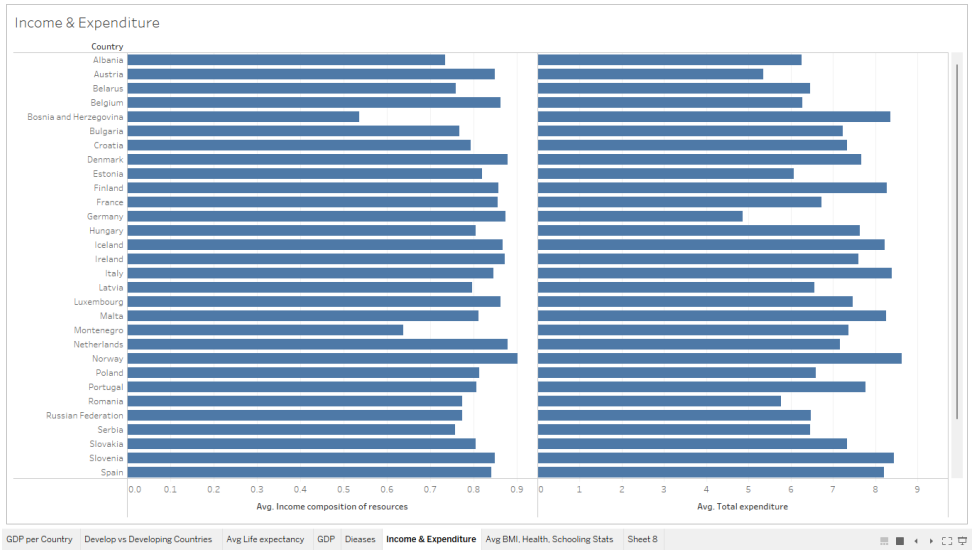

Figure 36. Income & Expenditure

This image shows the income and expenditure rate of every organization in Europe.

7. **Average BMI, Health, Schooling Stats**

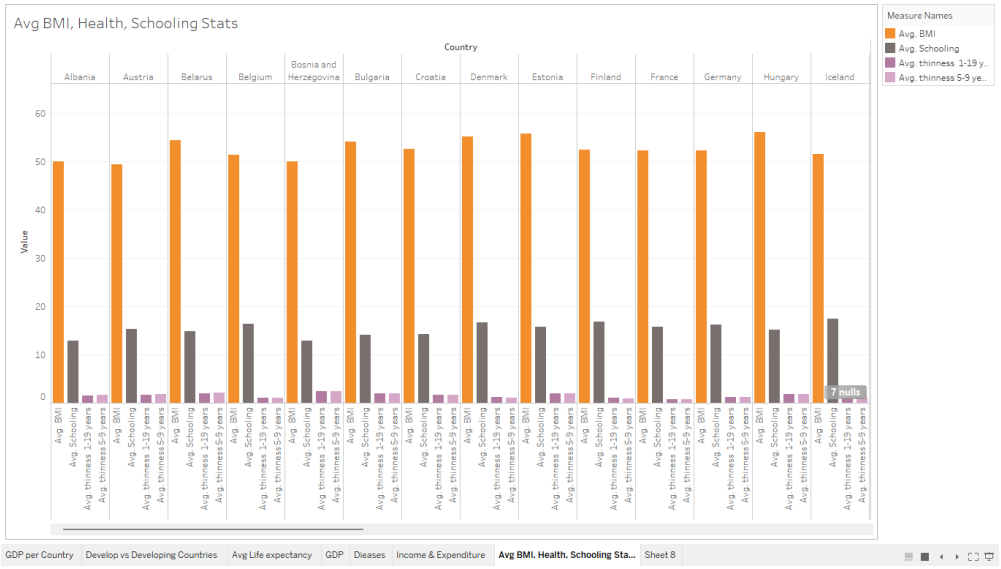

Figure 37. Avg BMI, Health, Schooling Stats

This fig.37 shows the stats regarding BMI, Health, Schooling of every country with the help of which it is easy to understand what areas need improvements in European continents for which effective strategies required to be developed by government of different organizations.

8. **GDP and Life expectancy**

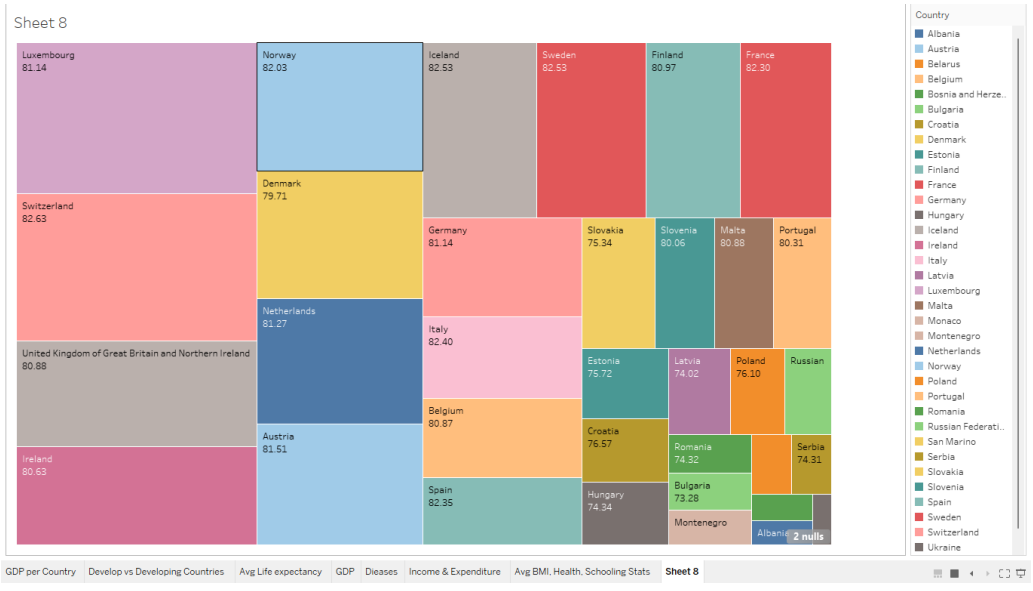

Figure 38.GDP and Life expectancy

The above figure38. shows the GDP and life expectancy of different European countries with the help of which it can be easy to understand the impact of GDP on life expectancy easily.

4.5 Feature engineering

Feature engineering is the processing of input dataset with mathematical notations, compatible with the machine learning algorithms requirements. After the data visualization process, the major process that has been performed is the feature engineering procedure for which first the StandardScalar and train_test_split libraries has been imported from the scikit-learn library, is one of the widely used machine learning library in python.

Before going to the feature engineering, I did label encoding, the python code associated with this is seen in the appendix part. For developed countries assigned the value 0 and for developing countries assigned value to 1.

The next step is feature selection, we need to separate the target and the feature variables. for which attributes the target and features to the y and X variables. the X variable defines the dropped variables which will not have any impact on the solution and the "y" variable defines the targeted value of the solution. Here, life expectancy is the target and all the other attributes except life expectancy, status, country is considered as features.

The dataset is splitting using train_test_split () method, which is setting the train size test size arguments such as X_train, X_test, Y_train, and Y_test. After that, I performed to split the data into two parts- 80% for training and 20% for testing i.e., 572 rows and 19 columns for training and 144 rows and 19 columns for testing. The following figure 39. shows the code associated with the feature selection.

```
# imorting libraries
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
```

```
# feature selection
X = df.drop(["lifeExpectancy", "Status", "Country"], axis=1)
y = df['lifeExpectancy']
```

```
# scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scaled,y, test_size = 0.2,  random_state=3)
```

```
print(X_train.shape, X_test.shape)
```

```
(572, 19) (144, 19)
```

Figure 39. Screenshot of python code for feature Engineering

4.6 Model implementation

Two models have been implemented during this process and their evaluations are also performed as defined in the following steps.

4.6.1 Linear regression Model

In this project linear regression are going to implement in python using Scikit-learn library. we divided the data into two parts, life expectancy is the target variable (Y variable) and other variables such as year, adult mortality, infant deaths, alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Diphtheria, HIV/AIDS, GDP, population, thinness 1-19 years, thinness 5-9 years, Income composition of resources and schooling are features (X variables). The problem is to fit a model to predict life expectancy of European countries with respect to the features.

Training a linear model

For Training the linear model to the train data set, imported linearRegression from scikit-learn library. Using the fit( ) method along with X_train, Y_train, we can fit the model to the train data set. After that, the regressor fitting the best line, through which we can determine the intercept, the coefficient of features, test data score and train data score.

Predicting the linear model

To predicting the result, assign the variable linpred. We pass the X_test values using predict() method. Now the linpred contains all the predicted values for the input values in the X_test values. Then we can compare predicted values with X_test values.

The coding used to implement linear regression Model is shown in the below figure 40.

```
Linregr = linear_model.LinearRegression()

# Fitting the model on the train dataset
Linregr.fit (X_train, Y_train)

# Predicting for the X_test points
linPred =Linregr.predict(X_test)

print('Coefficients - ', Linregr.coef_)
print('Intercept - ',Linregr.intercept_)

print("Train Data Score - ",Linregr.score(X_train,Y_train))
print("Test Data Score - ",Linregr.score(X_test,Y_test))
```

```
Coefficients -  [ 1.22420074 -0.64044909 -0.11070123 -0.61397652 -0.46127632 -0.29825697
  0.04315656 -0.13375446  0.72000303  0.41379004 -0.52540021 -1.28345285
  1.27351035 -0.08608596 -0.58320429 -0.54569745  1.17778934 -0.21713255
 -0.59765621]
Intercept -  78.17372126369578
Train Data Score -  0.7079482447652736
Test Data Score -  0.6749212502993451
```

Figure 40.Code for implementing linear regression

As per the output displayed in the image above, the testing data score is approximately 67% which is very low and not ideal, the training data score is around 71%, the value of intercept is 78.173 and the value of coefficients with corresponding features are shown in the figure 40.

According to the value of coefficient, we can understand when GDP, Income composition of country with coefficient 1.2735, 1.711 has biggest impact on life expectancy of European countries. Adult mortality of the country with coefficient -0.6404 has an adverse effect on life expectancy.

	Coefficient value
Year	1.224201
adultMortality	-0.640449
infantDeaths	-0.110701
Alcohol	-0.613977
percentageExpense	-0.461276
Hepatitis B	-0.298257
Measles	0.043157
BMI	-0.133754
under-five deaths	0.720003
Polio	0.413790
Diphtheria	-0.525400
HIV/AIDS	-1.283453
GDP	1.273510
Population	-0.086086
thinness 1-19 years	-0.583204
thinness 5-9 years	-0.545697
Income composition of resources	1.177789
Schooling	-0.217133
StatusLabel	-0.597656

Figure 41.Value of coefficients

4.6.2 Random Forest Regression Model

Random forest algorithm is a powerful prediction model, in which multiple algorithms of decision trees combines to form a Random Forest. Random forest algorithms can be used in both regression and classification problem.

In this section, Random Forest algorithm implemented in python using Scikit-learn library.

To predict the problem, the first step is to train our random forest algorithm using RandomForestRegressor from the class of Sklearn.ensemble. Here, n_estimators is the number of trees in the random forest. In this model, the training and testing data scores are evaluated after implementing the model. The python code associated with the random forest regressor is shown in the figure below.

```
# random-forest regressor model
rf = RandomForestRegressor(n_estimators=150, random_state=1, min_samples_leaf=2)
rf.fit (X_train, Y_train)
# Predicting for the X_test points
rfPred =rf.predict(X_test)

print("Train Data Score - ",rf.score(X_train,Y_train))
print("Test Data Score - ",rf.score(X_test,Y_test))
```

```
Train Data Score -  0.9708204105918486
Test Data Score -  0.8617428617565025
```

Figure 42.Screenshot of python code for implementing Random Forest Regression model

The training and testing data score are comparatively higher in random forest model than the linear regression model as the testing data score of random forest is approximately 86% which is much higher than 67% score of linear regression model and the train data score is approximately 97%. According to the train data score and test data score, this model works well to this regression problem.

Chapter 5. Result and discussion

5.1 Actual and Predicted value comparison

The actual and predicted values for linear regression model are compared with the plotted graph displaying the change in the life expectancy rate. In the figure 43, predicted values are represented by blue colour and actual values are represented by red colour. From the plot, predicted value seems to be highly diverging from actual value.

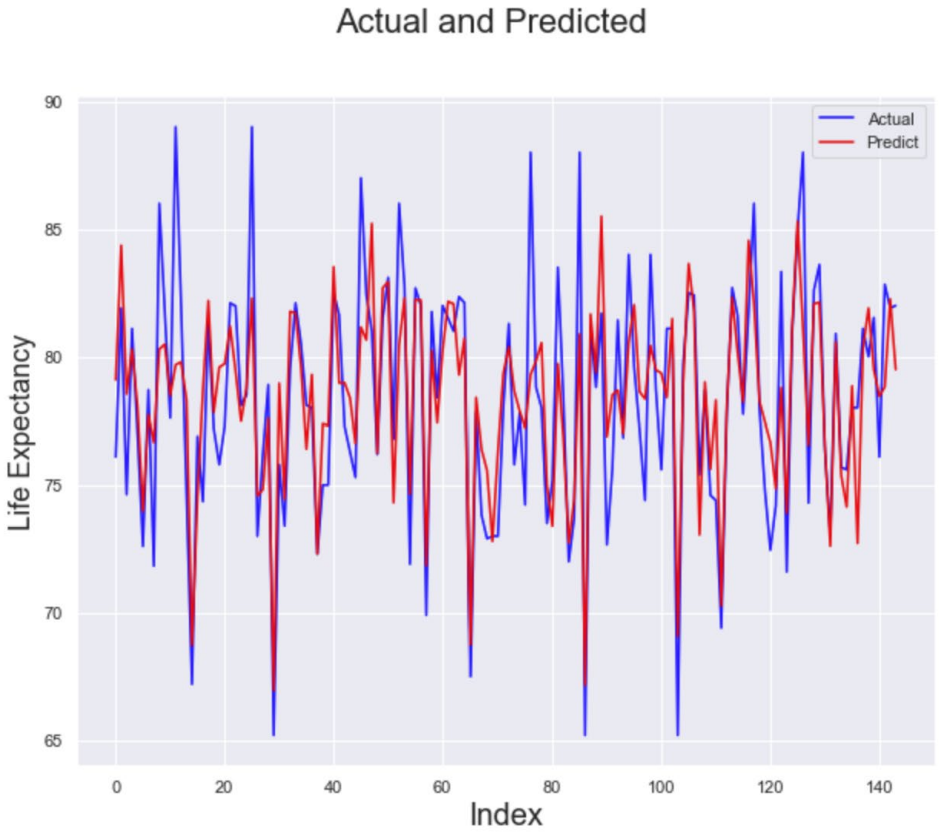

Figure 43.Plot for actual and predicted values in a linear regression model

Actual values and predicted values for random regression model are compared using the plot given in the figure 44. In this plot, it is seen that most of the predicted values are fitted well with the actual values.

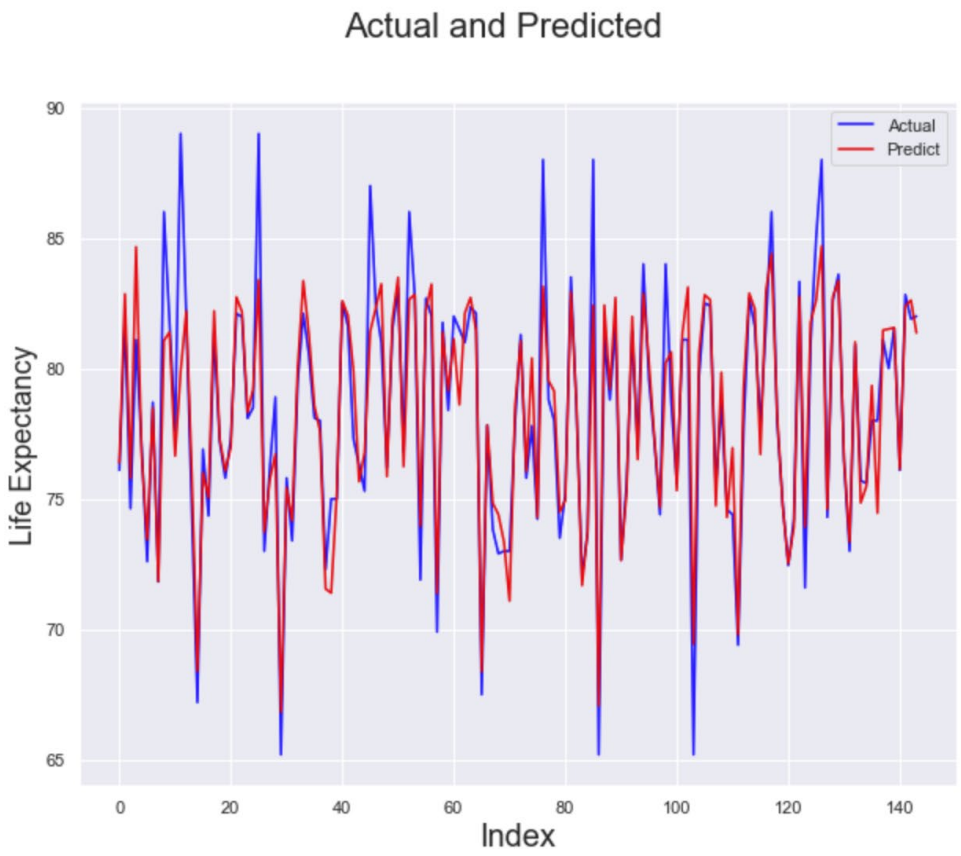

Figure 44.plot for actual and predicted values in random forest regression model

5.2 Model Evaluation

This is the last step of the data analysis part, in which implemented models are evaluated using some metrics calculations. In this project for evaluating models, mean absolute error, Residual sum of squares (MSE) and value of R2 are considered.

For linear regression model, the value of Mean absolute error is 2.06, the value of mean squared error is 7.30 and the value of R2 score is 42%. The screenshot of python code for calculating metrics is shown in the figure 45.

```
Y_test = np.asanyarray(Y_test)

print("Mean absolute error -  %.2f" % np.mean(np.absolute(linPred - Y_test)))
print("Residual sum of squares (MSE) -  %.2f" % np.mean((linPred - Y_test) ** 2))
print("R2-score -  %.2f" % r2_score(linPred,Y_test))
```

```
Mean absolute error -  2.06
Residual sum of squares (MSE) -  7.30
R2-score -  0.42
```

Figure 45.Screen shot for calculating regression metrics, for linear regression model

From the metrics calculation, linear regression Model can be able to predict only 42 percentage of variance in the outcome based on the features.

In random forest regression Model, the value of Mean absolute error is 1.10, the value of mean squared error is 3.10 and the value of R2 score is 82%. The screenshot of python code for calculating metrics in random forest regressor is shown in the figure 46.

```
Y_test = np.asanyarray(Y_test)

print("Mean absolute error -  %.2f" % np.mean(np.absolute(rfPred - Y_test)))
print("Residual sum of squares (MSE) -  %.2f" % np.mean((rfPred - Y_test) ** 2))
print("R2-score -  %.2f" % r2_score(rfPred,Y_test))
```

```
Mean absolute error -  1.10
Residual sum of squares (MSE) -  3.10
R2-score -  0.82
```

Figure 46.screenshot of python code for calculating error in random forest regression

From the result, we can see random regressor Model can be able to Predict 82 percentage of variance in the outcome based on the features. Moreover, value of absolute error and mean squared error are low compared to linear regression model.

When I plotted the graph for comparing actual, predicted values for random regression model and linear model it is seen that plot for random regression model is fitted well than the linear model. So as a researcher, we can say Random Forest regression model is the best model for explaining life expectancy of European countries.

Chapter 6. Conclusion and recommendation for future work

6.1 Conclusion

At the end of the study, it has been asserted that the given research study has focused on conducting prediction regarding life expectancy of European countries according to the parameters such as GDP, adult mortality, infant deaths, alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Diphtheria, HIV/AIDS, GDP, population, thinness 1-19 years, thinness 5-9 years, Income composition of resources and schooling before the spread of covid-19. The complete study has been done with the use of quantitative research approach following which descriptive and predictive analytics have been performed using machine learning approach to achieve the research objectives appropriately.

As per the above given study, it has been found that life expectancy is the calculation of the average number of additional years that a person of a specific age group is expected to live. In this similar concern, the given research study mainly has focused on the development of a machine learning model for the purpose of predicting life expectancy regarding European countries based on different factors such as GDP, infant death, population etc., For the purpose of conducting this research study, quantitative research design has been conducted with the help of which a novel Machine learning approach has been introduced, which gives higher prediction accuracies on predicting EU citizens' life expectancy.

Along with this, it has been demonstrated that the performance of Random Forest regression Machine Learning approach in terms of accuracy in predicting life expectancy is higher than the Linear Regression Machine learning based model. Moreover, Random forests easily overfit than Linear regression, which helps in the identification of optimized weight resulting in quick and efficient outcomes.

As per the study done through this analysis, it has been analysed that Norway, Latvia, Finland etc. are some of the countries that successfully maintain their life expectancy rate. The scatter plots showing the year wise life expectancy rate of all 35 European countries has evidently shown a drastic fall in the life expectancy rate of several countries such as Poland, Serbia, Slovakia, Spain, and many others.

The predictive analytics performed in this research study demonstrated that Life expectancy has a major impact on the GDP of the country as the higher will be the life expectance higher

will be the workable human capital availability in the country resulting in higher returns. In addition to this, it has also been evaluated that the increase in population will also be increasing the demands, the diseases, the problems and thus impacting the way of living of the people. Therefore, population can also be considered as one of the major factors to be considered by government when developing strategies to maintain life expectancy rate.

At the end of the given study, it has been found that Covid-19 pandemic has a huge impact on the life expectancy rate of the European countries as the business processes were disrupted, several people lost their lives and various unfavourable circumstances were created. Some countries were able to maintain their life expectancy rate at that time of pandemic such as Norway, Latvia, Finland, and some others. This shows that, the Covid-19 pandemic has impacted the life expectancy, but it is not entirely dependent on one factor as various factors are responsible for its decrease and increase such as- population, mortality rate, GDP, and various other factors. In addition to this, tableau has also been used for the purpose of analysing the chosen data and creating accurate visualizations. After doing the analysis, it has been analysed that GDP is one of the major factors having a huge impact on changing ratio of life expectancy in European continents.

6.2 Recommendation for future work

After doing this analysis, it has been evaluated that current study is limited to European countries only based on some factors such as GDP, population, mortality rate etc. In future study, the concern of Brexit can be taken into the consideration as one of the major factors for evaluating the life expectancy. In addition, the comparison between life expectancy before and after Brexit can also be done in future to get clear insights in the chosen domain.

References

- Aburto, J., Villavicencio, F., Basellini, U., Kjærgaard, S. and Vaupel, J., 2020. Dynamics of life expectancy and life span equality. *Proceedings of the National Academy of Sciences*, 117(10), pp.5250-5259.
- Adetunji, J., 2020. Don't die wondering: apps may soon be able to predict your life expectancy, but do you want to know?. [online] The Conversation. Available at: <https://theconversation.com/dont-die-wondering-apps-may-soon-be-able-to-predict-your-life-expectancy-but-do-you-want-to-know-129068.> [Accessed 19 August 2022].
- Bali, V., Aggarwal, D., Singh, S. and Shukla, A., 2021. Life Expectancy: Prediction & Analysis using ML. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*,.
- Beeksma, M., Verberne, S., van den Bosch, A., Das, E., Hendrickx, I. and Groenewoud, S., 2019. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC medical informatics and decision making, 19(1), pp.1-15.
- Bezy, J., 2022. life expectancy | Definition & Facts. [online] Encyclopedia Britannica. Available at: <https://www.britannica.com/science/life-expectancy> [Accessed 19 August 2022].
- Bilas, V., Franc, S. and Bonjak, M., 2022. Determinant Factors of Life Expectancy at Birth in the European Union Countries. [online] Available at: <https://hrcak.srce.hr/file/178666> [Accessed 22 August 2022].
- Eurostat, 2022. Mortality and life expectancy statistics - Statistics Explained. [online] Ec.europa.eu. Available at: <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Mortality_and_life_expectancy_statistics> [Accessed 19 August 2022].
- Gupta, S., 2020. Life Expectancy Prediction Using Linear Regression. [online] Enjoyalgorithms.com. Available at: <https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression> [Accessed 19 August 2022].
- Janssen, F., Bardoutsos, A., El Gewily, S. and De Beer, J., 2021. Future life expectancy in Europe taking into account the impact of smoking, obesity, and alcohol. eLife, 10.
- Khouri, Klaudia and Cehlar, 2022. EXPECTED LIFE EXPECTANCY AND ITS DETERMINANTS IN SELECTED EUROPEAN COUNTRIES. [online] Available at: <https://web.p.ebscohost.com/abstract? > [Accessed 22 August 2022].

- Knauss, S., 2022. Longevity: Extending Life Span Expectancy. [online] Available at: <https://www.disabled-world.com/fitness/longevity/> [Accessed 22 August 2022].
- Kolasa-Więcek, A. and Suszanowicz, D., 2019. Air pollution in European countries and life expectancy—modelling with the use of neural network. *Air Quality, Atmosphere & Health*, 12(11), pp.1335-1345.
- Leon, D., 2011. Trends in European life expectancy: a salutary view. *International Journal of Epidemiology*, 40(2), pp.271-277.
- Li, Y., Pan, A., Wang, D., Liu, X., Dhana, K., Franco, O., Kaptoge, S., Di Angelantonio, E., Stampfer, M., Willett, W. and Hu, F., 2018. Impact of Healthy Lifestyle Factors on Life Expectancies in the US Population. *Circulation*, 138(4), pp.345-355.
- Mackenbach, J., Valverde, J., Bopp, M., Brønnum-Hansen, H., Deboosere, P., Kalediene, R., Kovács, K., Leinsalu, M., Martikainen, P., Menvielle, G., Regidor, E. and Nusselder, W., 2019. Determinants of inequalities in life expectancy: an international comparative study of eight risk factors. *The Lancet Public Health*, 4(10), pp.e529-e537.
- Mäki, N., Martikainen, P., Eikemo, T., Menvielle, G., Lundberg, O., Östergren, O., Jasilionis, D. and Mackenbach, J., 2013. Educational differences in disability-free life expectancy: a comparative study of long-standing activity limitation in eight European countries. *Social Science & Medicine*, 94, pp.1-8.
- McDonnell, C., 2018. What really drives higher life expectancy? [online] Medium. Available at: <https://towardsdatascience.com/what-really-drives-higher-life-expectancy-e1c1ec22f6e1> [Accessed 19 August 2022].
- Miladinov, G., 2020. Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries. *Genus*, 76(1).
- Miladinov, G., 2020. Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries. Genus, 76(1), pp.1-20.
- OECD, 2018. Health at a Glance: Europe 2018. *Health at a Glance: Europe*,.
- Raleigh, V. and Fund, T., n.d. OECD Health Working Papers.
- Robine, J., 2022. Healthy life expectancy in Europe. [online] Available at: <https://www.researchgate.net/publication/289381207_Healthy_life_expectancy_in_Europe> [Accessed 22 August 2022].

- Statista, 2022. Life expectancy by continent 2021 | Statista. [online] Statista. Available at: <https://www.statista.com/statistics/270861/life-expectancy-by-continent/> [Accessed 19 August 2022].
- Stenholm, S., Head, J., Kivimäki, M., Kawachi, I., Aalto, V., Zins, M., Goldberg, M., Zaninotto, P., Magnuson Hanson, L., Westerlund, H. and Vahtera, J., 2016. Smoking, physical inactivity and obesity as predictors of healthy and disease-free life expectancy between ages 50 and 75: a multicohort study. *International Journal of Epidemiology*, 45(4), pp.1260-1270.
- Svensson, K., 2018. Predicting Life Expectancy Using Machine Learning.
- Zamzamy Sormin, M., Sihombing, P., Amalia, A., Wanto, A., Hartama, D. and Chan, D., 2019. Predictions of World Population Life Expectancy Using Cyclical Order Weight / Bias. *Journal of Physics: Conference Series*, 1255(1), p.012017.
- G. Marois, R. Muttarak and S. Scherbov, "Assessing the potential impact of COVID-19 on life expectancy", PLOS ONE, vol. 15, no. 9, p. e0238678, 2020. Available: 10.1371/journal.pone.0238678 [Accessed 6 August 2022].
- Comparative Analysis of Regression Regularization Methods for Life Expectancy Prediction-Nataliya Boyko and Olena Moroz **(***Lviv Polytechnic National University, Profesorska Street 1, Lviv, 79013, Ukraine)*
- E. Ortiz-Ospina, "“Life Expectancy” – What does this actually mean?", Our World in Data, 2017. [Online]. Available: https://ourworldindata.org/life-expectancy-how-is-it-calculated-and-how-should-it-be-interpreted. [Accessed: 06- Aug- 2022].
- H. Beltrán-Sánchez, S. Soneji and E. Crimmins, "Past, Present, and Future of Healthy Life Expectancy: Figure 1.", Cold Spring Harbor Perspectives in Medicine, vol. 5, no. 11, p. a025957, 2015. Available: 10.1101/cshperspect.a025957 [Accessed 6 August 2022].
- M. Luy, P. Di Giulio, V. Di Lego, P. Lazarevič and M. Sauerberg, "Life Expectancy: Frequently Used, but Hardly Understood", Gerontology, vol. 66, no. 1, pp. 95-104, 2019. Available: 10.1159/000500955 [Accessed 6 August 2022].
- E. Arias, B. Tejada-Vera, F. Ahmad and K. Kochanek, "Provisional Life Expectancy Estimates for 2020", Vital Statistics Surveillance Report, 2021. Available: https://www.cdc.gov/nchs/products/index.htm. [Accessed 4 August 2022].
- D. Narayana, S. Pallavi, S. kiran, V. Prathyusha and T. Mahesh, "LIFE EXPECTANCY PREDICTION USING MACHINE LEARNING", Journal of Interdisciplinary Cycle

Research, vol. 7, no. 5, 2020. Available: http://www.jicrjournal.com/gallery/119-jicr-may-2743.pdf. [Accessed 6 August 2022].

- K. Vydehi, K. Manchikanti, T. Kumari and S. Shah, "Machine Learning Techniques for Life Expectancy Prediction", International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 4, 2020. Available: http://warse.org/IJATCSE/static/pdf/file/ijatcse45942020.pdf. [Accessed 6 August 2022].
- R. S, "Predicting Life Expectancy Using Machine Learning", Science, Technology and Development, vol. 11, no. 7, 2022. Available: http://journalstd.com/gallery/38-july2022.pdf. [Accessed 6 August 2022].
- https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
- https://www.isixsigma.com/
- -https://www.sciencedirect.com/
- https://levelup.gitconnected.com/

Appendix

*Python code script used for the project

```
import sys
import warnings
if not sys.warnoptions:
  warnings.simplefilter("ignore")
warnings.filterwarnings("ignore")
```

Importing Libraries

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
!pip install plotly
import plotly.express as px
```

Loading Dataset

```
pd.set_option('display.max_columns', None)
df=pd.read_csv("LIFE_EXPECTANCY.csv")
```

Exploratory Data Analysis

```
# show top five rows
df.head()
```

```
# general information
df.info()
```

```
# give the number of rows and columns
df.shape
```

```
# rename the column
df.rename(columns = {'Life expectancy ':'lifeExpectancy', 'Adult Mortality':'adultMortality','infant deaths':'infantDeaths','perc
```

```
# extract all columns of the dataset
df.columns
```

```
# calculate the mean , std, min, max and count of every attributes
df.describe()
```

```
# check for null values
df.isna().sum()
```

```
# Checking skewness
visDf=df.loc[:,'Country':'Schooling']
visDf=visDf.select_dtypes([np.int, np.float])
for i, col in enumerate(visDf.columns):
    print(f"\nSkewness of {col} is {df[col].skew()}")
```

```
# replace the null values with its mean and median as per skewness values.
df['lifeExpectancy']=df['lifeExpectancy'].fillna(df['lifeExpectancy'].mean())
df['adultMortality']=df['adultMortality'].fillna(df['adultMortality'].mean())
df['Population'] = df['Population'].fillna(df['Population'].median())
df['GDP'] = df['GDP'].fillna(df['GDP'].median())
df['BMI'] = df['BMI'].fillna(df['BMI'].mean())
df['Income composition of resources'] = df['Income composition of resources'].fillna(df['Income composition of resources'].mean()
df['Polio'] = df['Polio'].fillna(df['Polio'].mean())
df['thinness  1-19 years'] = df['thinness  1-19 years'].fillna(df['thinness  1-19 years'].mean())
df['thinness 5-9 years'] = df['thinness 5-9 years'].fillna(df['thinness 5-9 years'].mean())
df['Hepatitis B'] = df['Hepatitis B'].fillna(df['Hepatitis B'].mean())
df['Diphtheria'] = df['Diphtheria'].fillna(df['Diphtheria'].mean())
df['Alcohol'] = df['Alcohol'].fillna(df['Alcohol'].mean())
df['Schooling'] = df['Schooling'].fillna(df['Schooling'].mean())
```

```
# check for null values
df.isna().sum()
```

```
# show the correlation
df.corr()
```

Data visualization

```
# Heatmap
dCorr=df.corr()
sns.set(rc={'figure.figsize':(10,8)})
fig = px.imshow(dCorr, text_auto=True, aspect="auto")
fig.show()
```

```
#Scatter plot
column=['Status','adultMortality',
       'infantDeaths', 'Alcohol', 'percentageExpense', 'Hepatitis B',
       'Measles', 'BMI', 'under-five deaths', 'Polio',
       'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness  1-19 years',
       'thinness 5-9 years', 'Income composition of resources', 'Schooling']
for i in column:
  plot=px.scatter(df,x='lifeExpectancy',y=i ,title=f"Scatter Plot of Life-Expectancy WRT {i} ")
  plot.show()
```

```
# count plot for developing and developed country of Europe
sns.set(rc={'figure.figsize':(10,8)})
plt.title("Count Plot for status of Europeon Countries")
df['Status'].value_counts().plot.bar()
```

```
for i, col in enumerate(df.columns):
    fig = px.histogram(df, x=col, title="Plot Distribution of "+col)
    fig.update_layout(bargap=0.2)
    fig.show()
```

```
# Box-Plot for all attributes
for i, col in enumerate(df.columns):
    fig = px.box(df, x=col, title="Box Plot of "+col)
    fig.show()
```

```
# Removing outliers
for i, col in enumerate(visDf.columns):
  Q1 = df[col].quantile(0.25)
  Q2 = df[col].quantile(0.50)
  Q3 = df[col].quantile(0.75)

  IQR = Q3 - Q1
  print('\nInterquartile range of',col, 'is: %.2f'% IQR)
  whisker_width = 1.5
  low_lim = Q1 - (whisker_width * IQR)
  up_lim = Q3 + (whisker_width * IQR)
  print('Lower Limit of',col,'is: %.2f'%low_lim)
  print('Upper Limit of',col,'is: %.2f'%up_lim,'\n')

  try:
    df[col]=np.where(df[col]>up_lim,up_lim,np.where(df[col]<low_lim,low_lim,df[col]))
  except:
    print('\t\nUnable to remove an outlier for:', col)
```

```
# Country-wise Scatter plot of " lifeExpectancy vs year "
cntry=['Albania', 'Austria', 'Belarus', 'Belgium',
       'Bosnia and Herzegovina', 'Bulgaria', 'Croatia', 'Denmark',
       'Estonia', 'Finland', 'France', 'Germany', 'Hungary', 'Iceland',
       'Ireland', 'Italy', 'Latvia', 'Luxembourg', 'Malta', 'Monaco',
       'Montenegro', 'Netherlands', 'Norway', 'Poland', 'Portugal',
       'Romania', 'Russian Federation', 'San Marino', 'Serbia',
       'Slovakia', 'Slovenia', 'Spain', 'Sweden', 'Switzerland',
       'Ukraine', 'United Kingdom of Great Britain and Northern Ireland']
for i in cntry:
  cc=df.loc[df['Country']==i]
  plot=px.scatter(cc,x='Year',y='lifeExpectancy',title=f"Scatter plot of Life-Expectancy vs year for {i} Country")
  plot.show()
  print("\n\n\n")
```

```
# label encoding and assign in new variable
from sklearn import preprocessing
Lab_encode = preprocessing.LabelEncoder()
df['StatusLabel'] = Lab_encode.fit_transform(df['Status'].values)
```

```
# check assigned values
st = df.groupby('Status')
st = st['StatusLabel']
st.first()
```

Feature Engineering

```
# imorting libraries
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
```

```
# feature selection
X = df.drop(["lifeExpectancy", "Status", "Country"], axis=1)
y = df['lifeExpectancy']
```

```
# scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scaled,y, test_size = 0.2,  random_state=3)
```

```
print(X_train.shape, X_test.shape)
```

Model Implementation

```
from sklearn.metrics import accuracy_score
from sklearn.metrics import r2_score
from sklearn.ensemble import RandomForestRegressor
from sklearn import linear_model
from sklearn.linear_model import LinearRegression
```

Linear Regression

```
Linregr = linear_model.LinearRegression()

# Fitting the model on the train dataset
Linregr.fit (X_train, Y_train)

# Predicting for the X_test points
linPred =Linregr.predict(X_test)

print('Coefficients - ', Linregr.coef_)
print('Intercept - ',Linregr.intercept_)

print("Train Data Score - ",Linregr.score(X_train,Y_train))
print("Test Data Score - ",Linregr.score(X_test,Y_test))
```

```
feature_names = X.columns
model_coefficients = Linregr.coef_

coefficients_df = pd.DataFrame(data = model_coefficients,
                               index = feature_names,
                               columns = ['Coefficient value'])
print(coefficients_df)
```

```
linPred =Linregr.predict(X_test)
results = pd.DataFrame({'Actual': Y_test, 'Predicted': linPred})
print(results)
```

```
Y_test = np.asanyarray(Y_test)

print("Mean absolute error -  %.2f" % np.mean(np.absolute(linPred - Y_test)))
print("Residual sum of squares (MSE) -  %.2f" % np.mean((linPred - Y_test) ** 2))
print("R2-score -  %.2f" % r2_score(linPred,Y_test))
```

```
loop = [i for i in range (0,144)]
fig=plt.figure()
plt.plot(loop,Y_test, color="blue", linewidth=1.5, linestyle="-")
plt.plot(loop,linPred, color="red", linewidth=1.5, linestyle="-")
fig.suptitle('Actual and Predicted', fontsize=22)
plt.xlabel('Index', fontsize=20)
plt.ylabel('Life Expectancy',fontsize=18)
```

random-forest regressor model

```
# random-forest regressor model
rf = RandomForestRegressor(n_estimators=150, random_state=1, min_samples_leaf=2)
rf.fit (X_train, Y_train)
# Predicting for the X_test points
rfPred =rf.predict(X_test)

print("Train Data Score - ",rf.score(X_train,Y_train))
print("Test Data Score - ",rf.score(X_test,Y_test))
```

```
Y_test = np.asanyarray(Y_test)

print("Mean absolute error -  %.2f" % np.mean(np.absolute(rfPred - Y_test)))
print("Residual sum of squares (MSE) -  %.2f" % np.mean((rfPred - Y_test) ** 2))
print("R2-score -  %.2f" % r2_score(rfPred,Y_test))
```

```
loop = [i for i in range (0,144)]
fig=plt.figure()
plt.plot(loop,Y_test, color="blue", linewidth=1.5, linestyle="-")
plt.plot(loop,rfPred, color="red", linewidth=1.5, linestyle="-")
fig.suptitle('Actual and Predicted', fontsize=22)
plt.xlabel('Index', fontsize=20)
plt.ylabel('Life Expectancy',fontsize=18)
```

*Additional figures (Scatterplot of Life Expectancy-Countrywise)

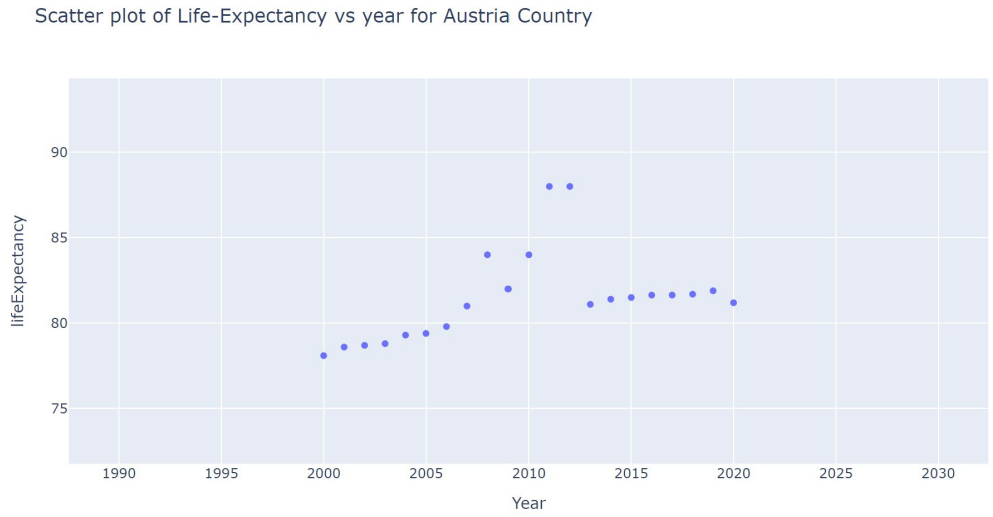

Figure 47.Scatter plot of Austria

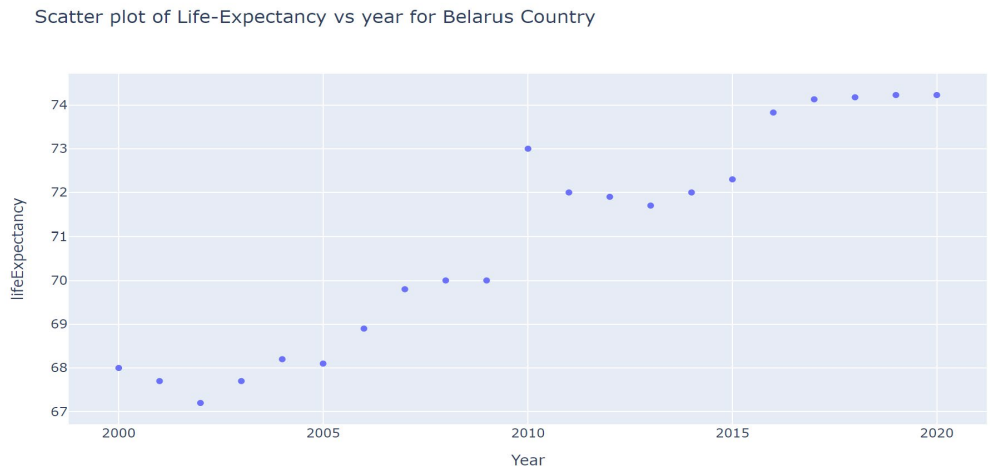

Figure 48.Scatter plot of Belarus

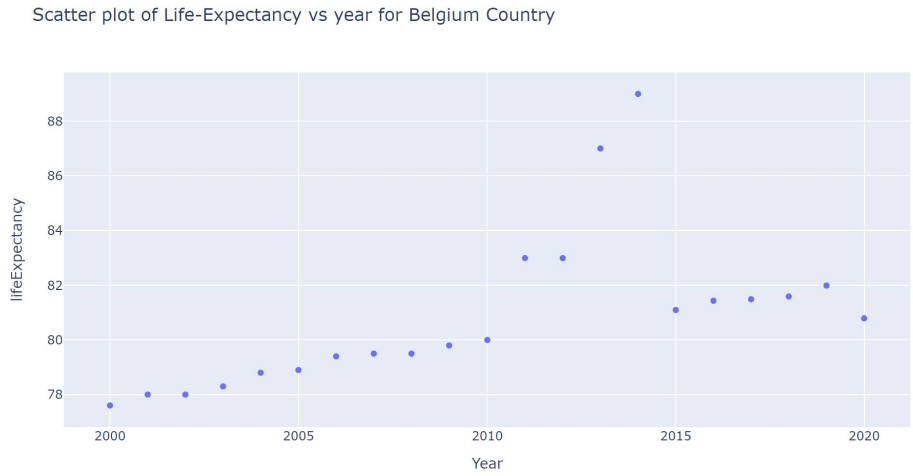

Figure 49.Scatter plot of Belgium

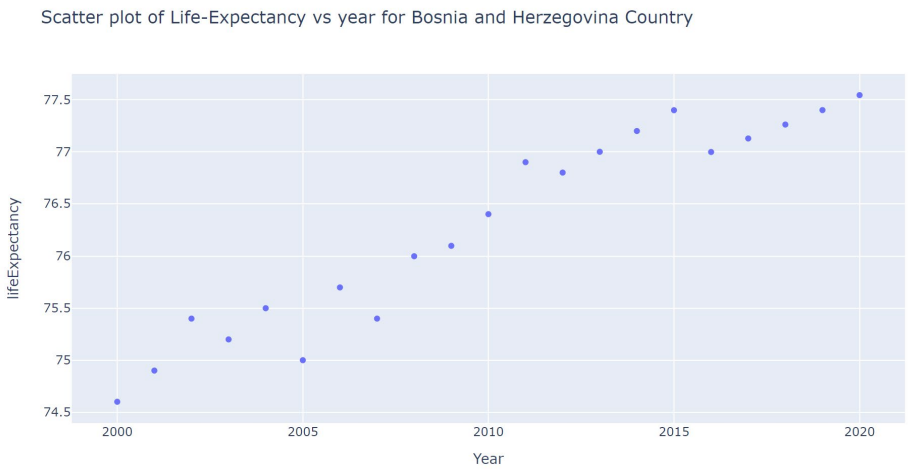

Figure 50.Scatter plot of Bosnia and Herzegovina

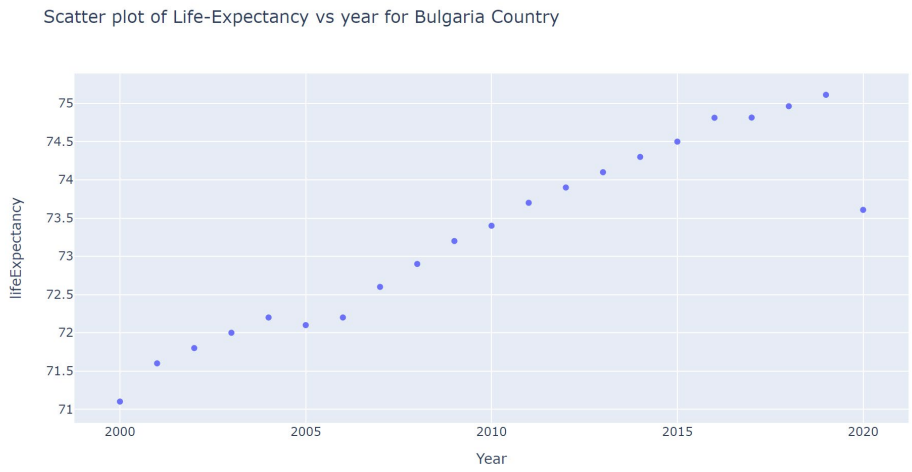

Figure 51.Scatter plot of Bulgaria

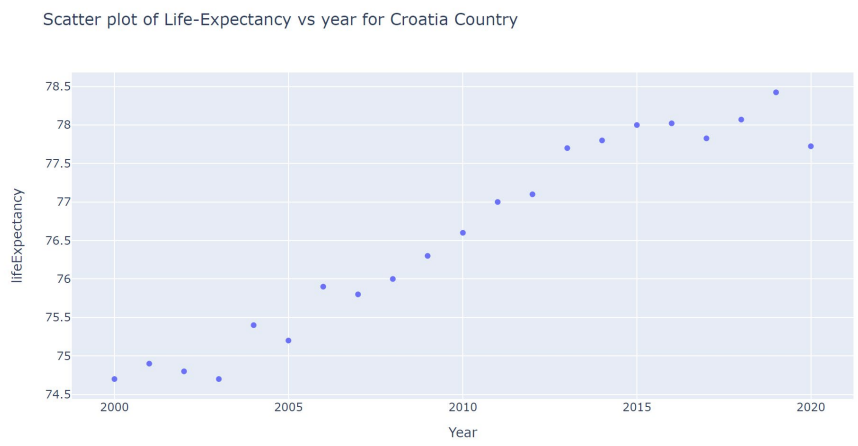

Figure 52.Scatter plot of Croatia

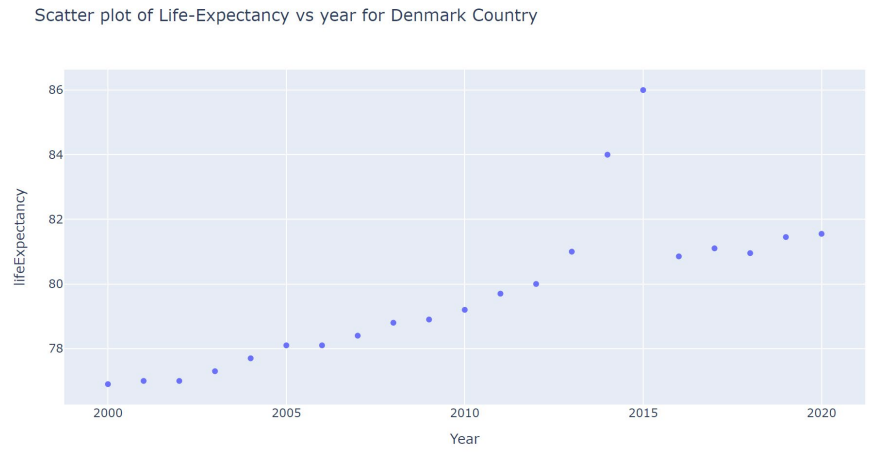

Figure 53.Scatter plot of Denmark

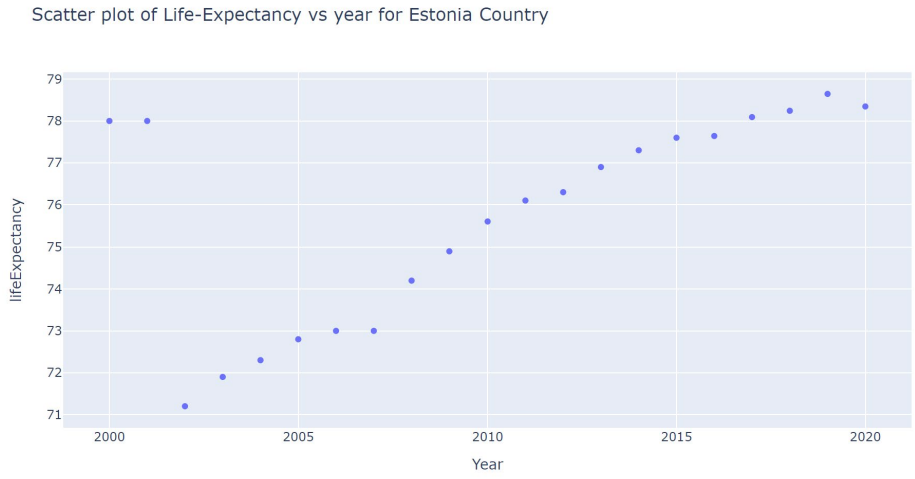

Figure 54.Scatter plot of Estonia

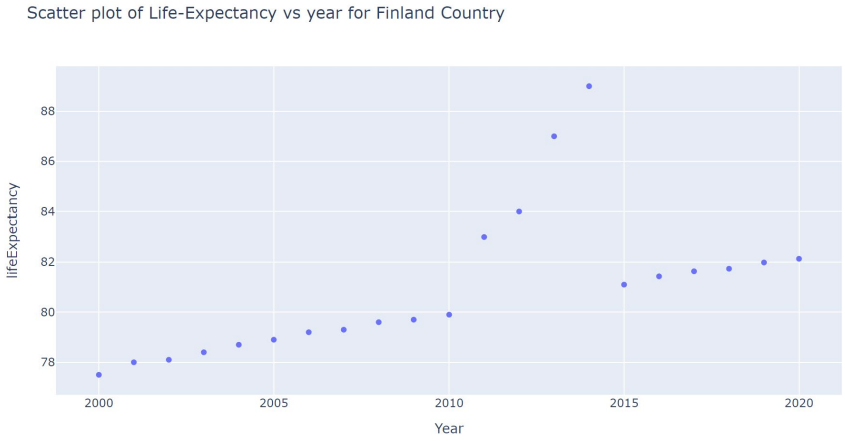

Figure 55.Scatter plot of Finland

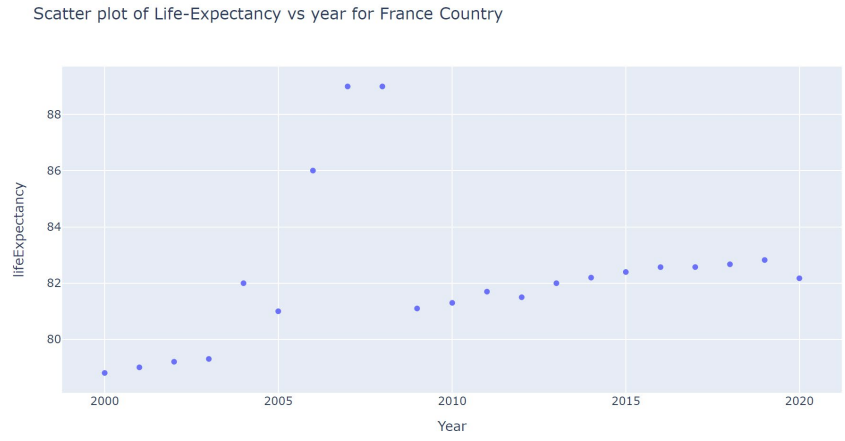

Figure 56.Scatter plot of France

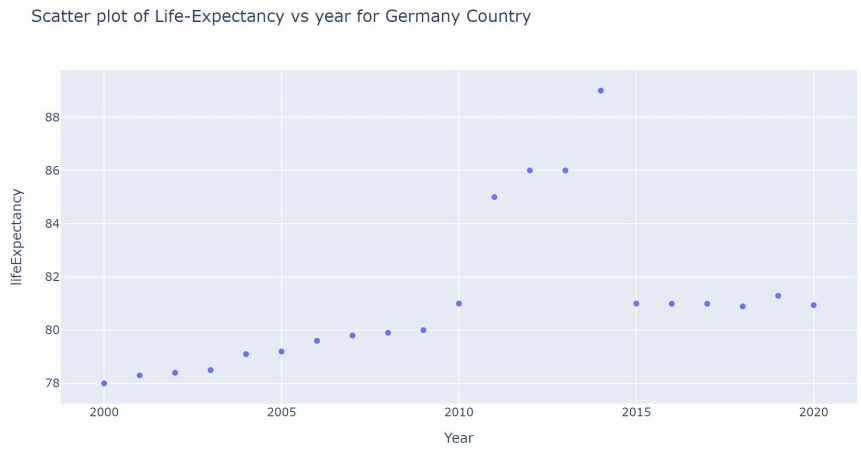

Figure 57.Scatter plot of Germany

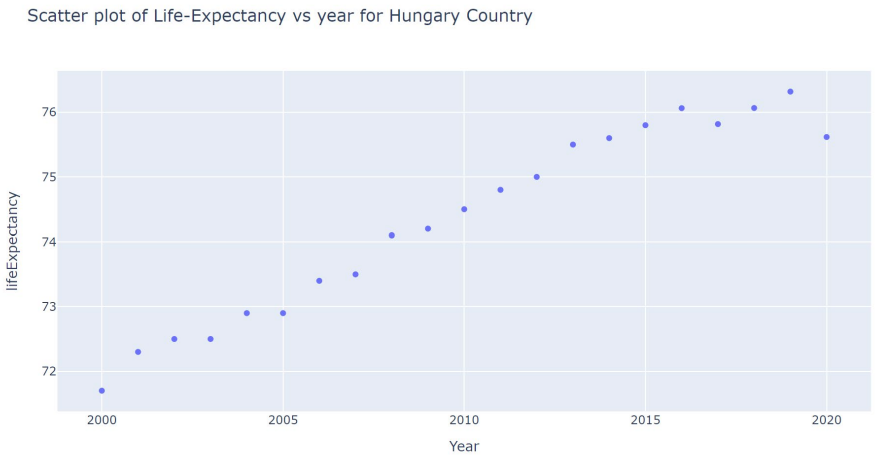

Figure 58.Scatter plot of Hungary

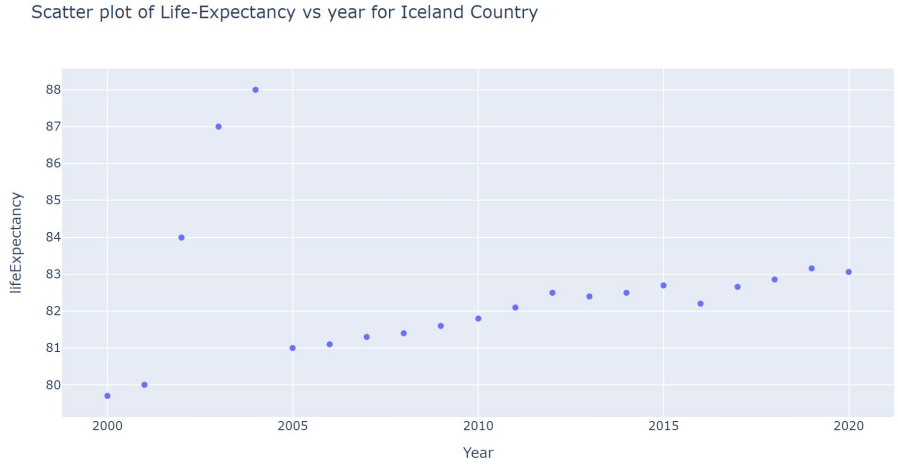

Figure 59.Scatter plot of Iceland

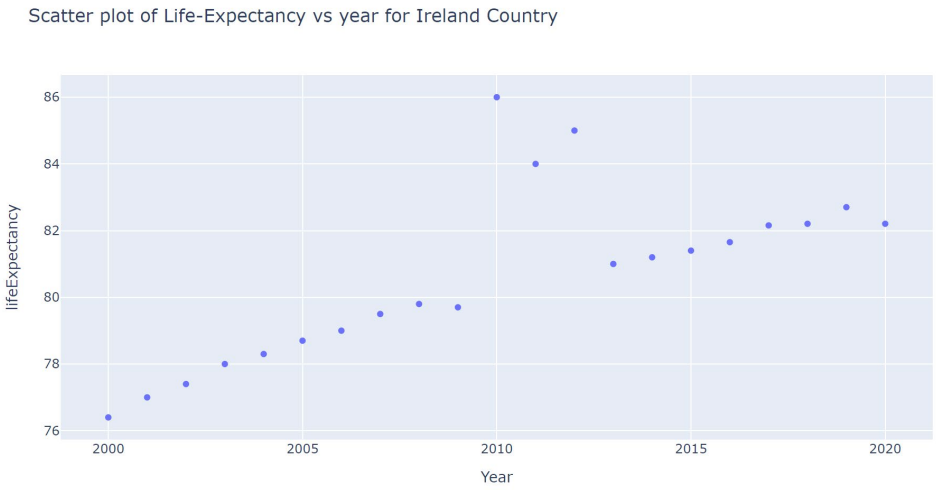

Figure 60.Scatter plot of Ireland

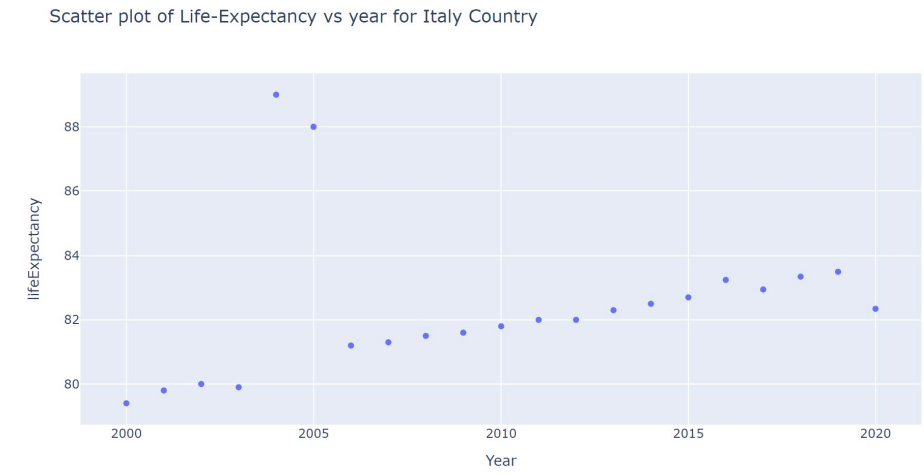

Figure 61.Scatter plot of Italy

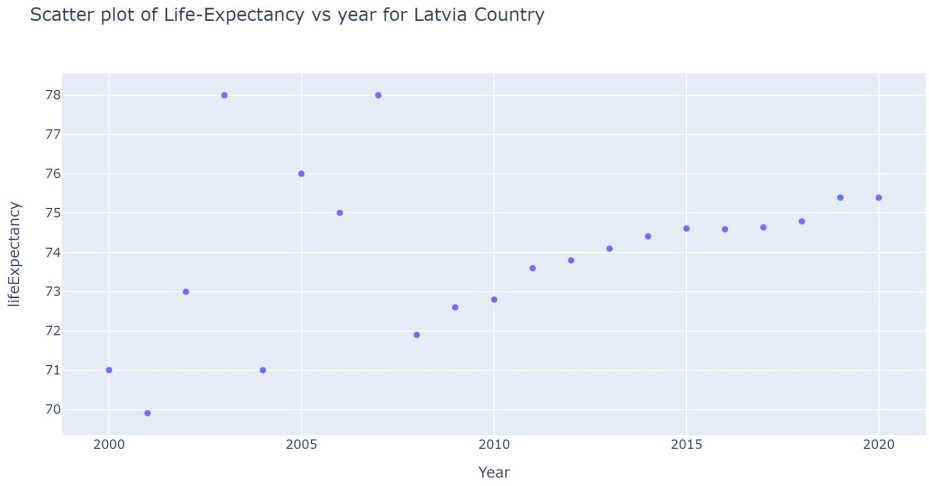

Figure 62.Scatter plot of Latvia

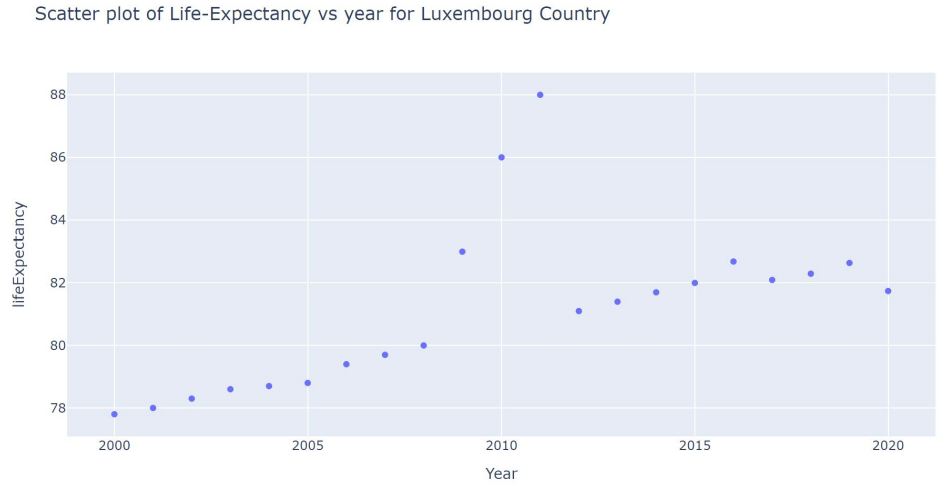

Figure 63.Scatter plot of Luxembourg

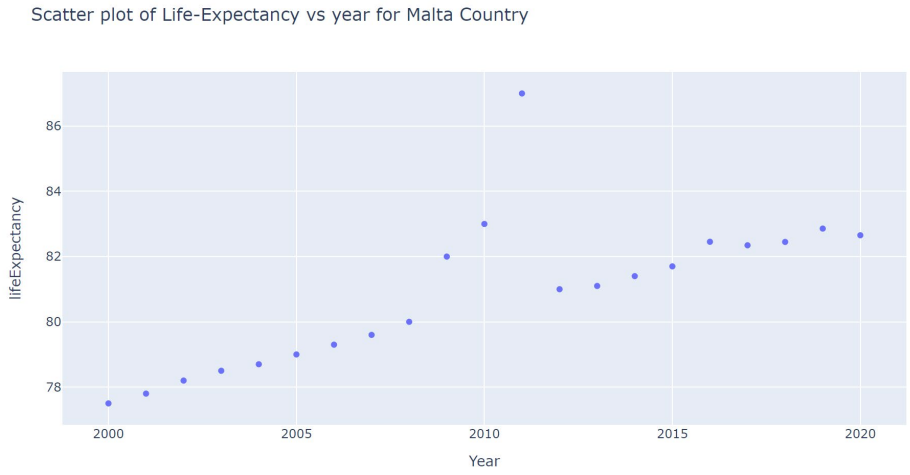

Figure 64.Scatter plot of Malta

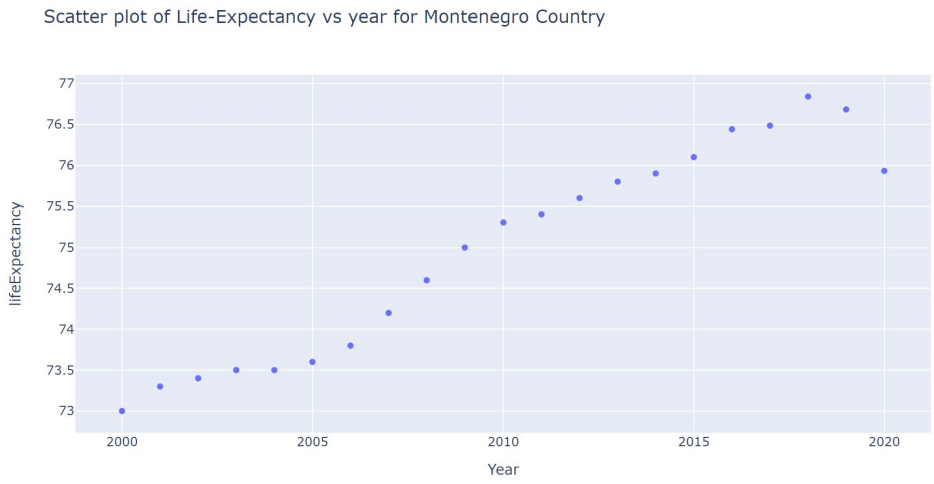

Figure 65.Scatter plot of Montenegro

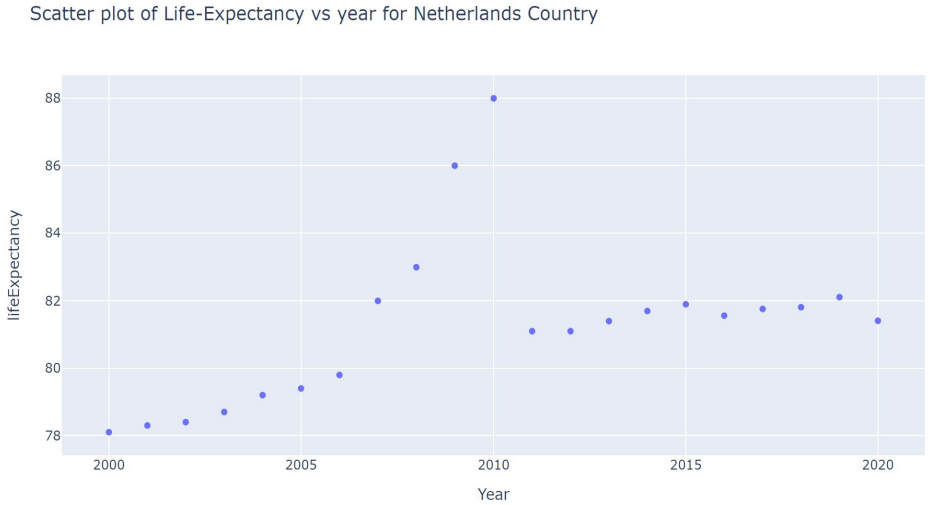

Figure 66.Scatter plot of Netherland

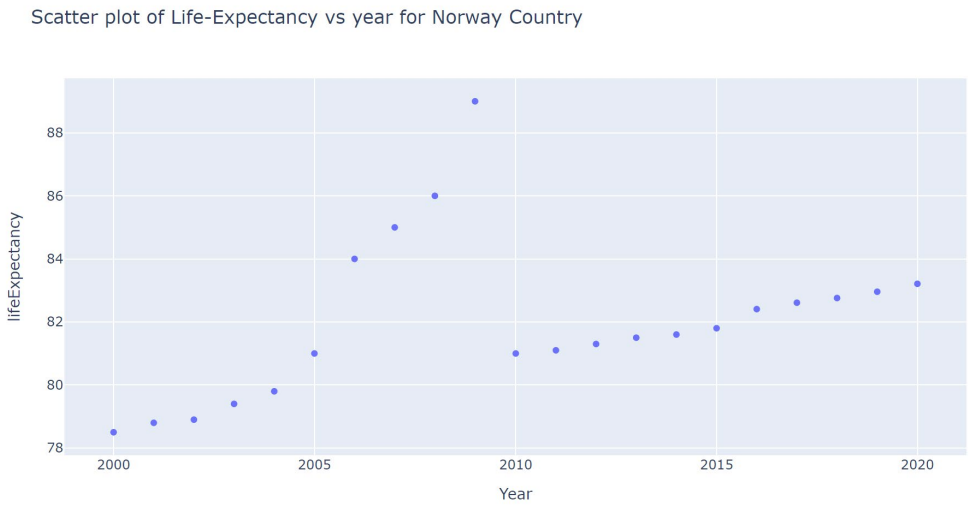

Figure 67.Scatter plot of Norway

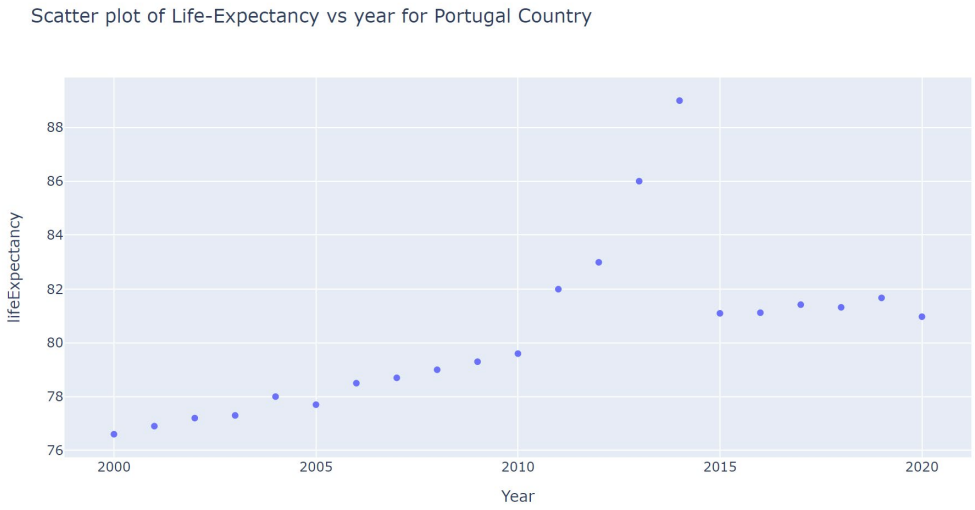

Figure 68.Scatter plot of Portugal

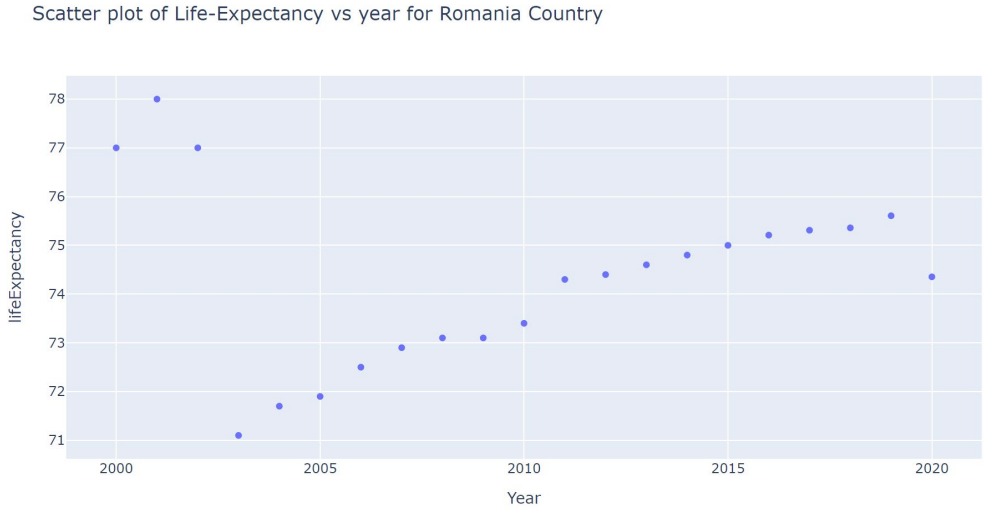

Figure 69.Scatter plot of Romania

Figure 70.Scatter plot of Russian Federation

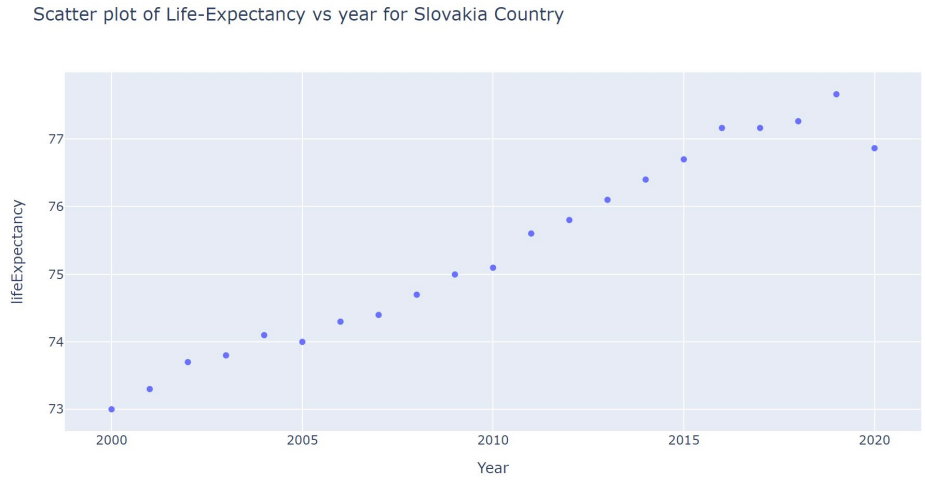

Figure 71.Scatter plot of Slovakia

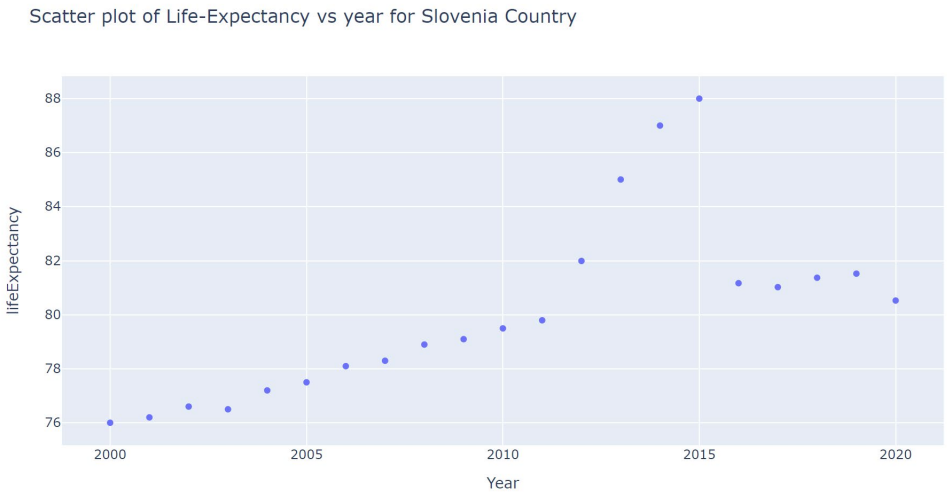

Figure 72.Scatter plot of Slovenia

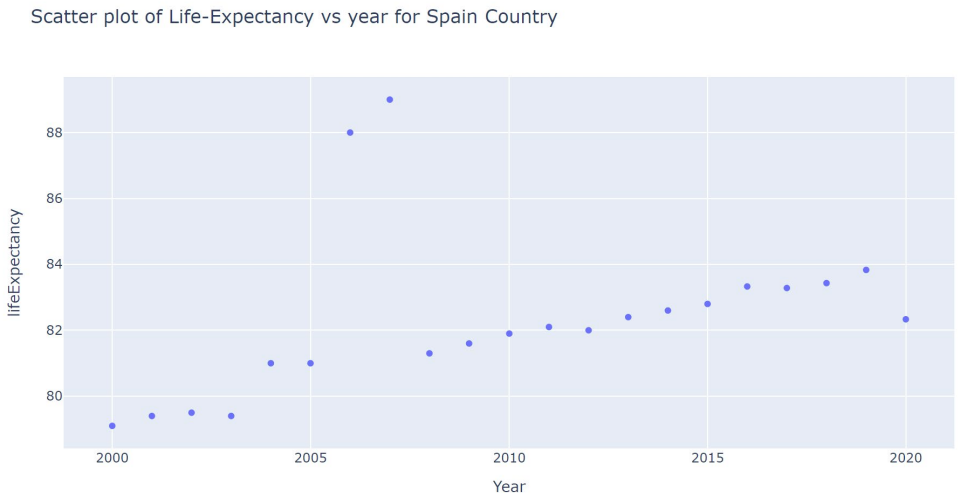

Figure 73.Scatter plot of Spain

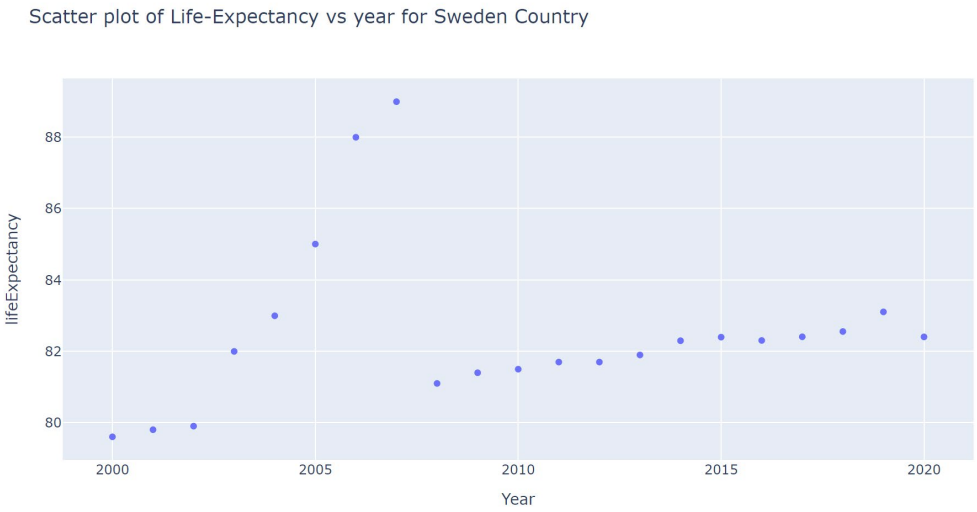

Figure 74.Scatter plot of Sweden

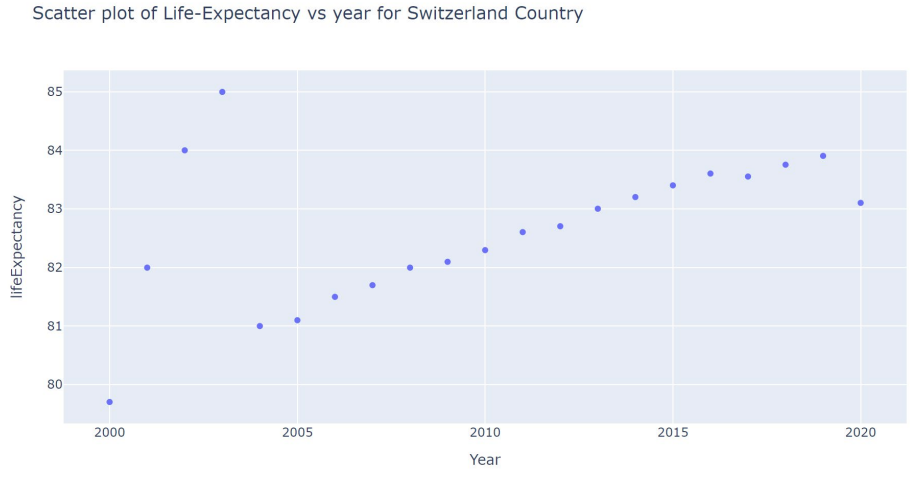

Figure 75.Scatter plot of Switzerland

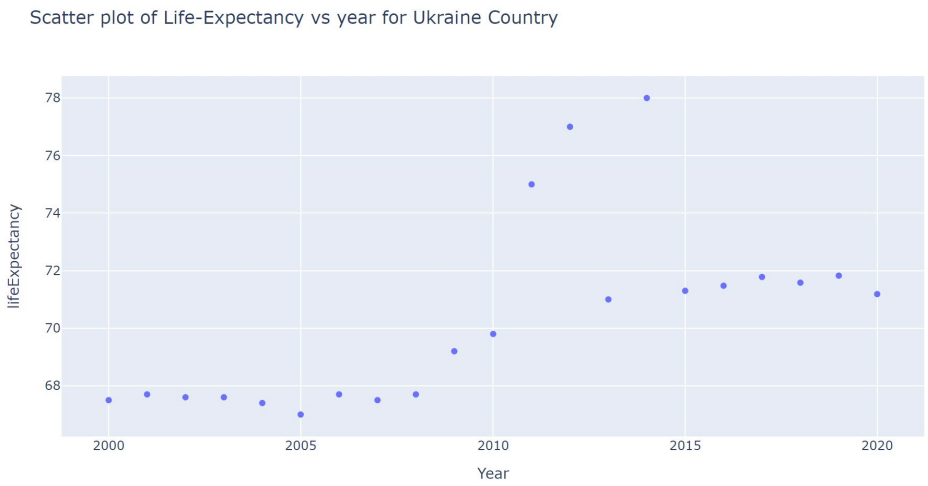

Figure 76.Scatter plot of Ukraine

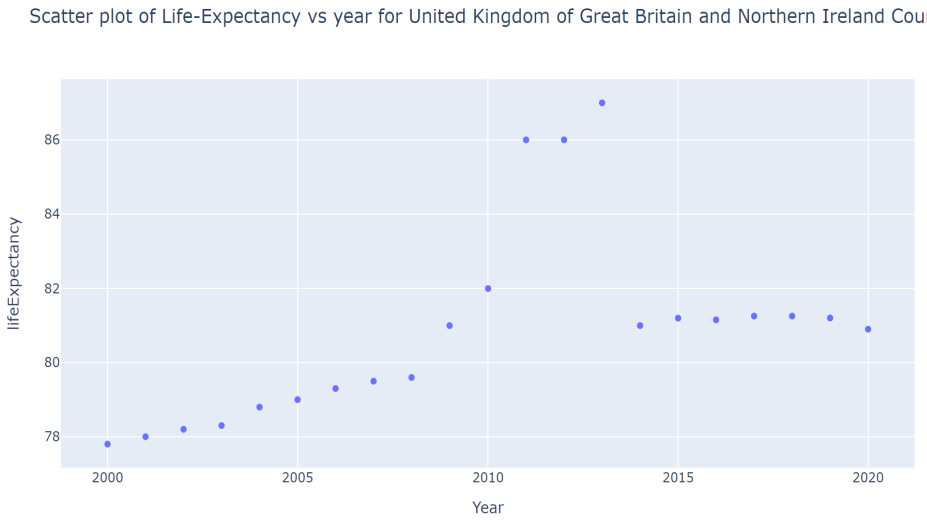

Figure 77.Scatter plot of United Kingdom of Great Britain and Northern Ireland