SYMBIOSIS
STATISTICAL INSTITUTE
॥वसुधैव कुटुम्बकम्॥

A project report on -

MULTIPLE LINEAR REGRESSION ANALYSIS ON

Happiness Index

Under the Guidance of
Dr. Priya Deshpande
Symbiosis Statistical Institute, Pune

Submitted by
Puspal Singha – 23060641075
Aman Kumar – 23060641049

Introduction

In recent years, the pursuit of happiness has emerged as a central focus in societal development, prompting researchers and policymakers to explore its determinants across various domains. Understanding what factors contribute to the overall happiness of a nation is crucial for devising effective policies aimed at enhancing the well-being of its citizens. This project seeks to develop a predictive model capable of estimating the happiness index of a country based on a multitude of socio-economic and cultural indicators.

The happiness index, often measured through surveys such as the World Happiness Report, encompasses a range of dimensions including economic prosperity, social support systems, health outcomes, personal freedoms, and perceptions of corruption. By leveraging data on these factors, collected over different years and across diverse countries, we aim to construct a comprehensive model that can elucidate the intricate relationship between these variables and the subjective well-being of individuals within a society.

By developing an accurate predictive model, we aim to gain insights into the key drivers of happiness and provide valuable information for policymakers, researchers, and organizations working towards enhancing well-being and quality of life worldwide. This project contributes to the growing body of knowledge in the intersection of data science, social sciences, and public policy, ultimately striving towards a happier and more prosperous global community.

Problem Statement

Predicting National Happiness Index: A Comprehensive Analysis of Socio-Economic Factors Using Multiple Linear Regression and various other Machine Learning Techniques.

In recent years, the measurement of national happiness has gained significant attention as a crucial indicator of societal well-being and progress. This study aims to develop a robust predictive model for estimating the happiness index of countries based on various socio-economic factors. The primary objective is to utilize multiple linear regression as a baseline model and compare its performance with other advanced machine learning algorithms.

Objective

- The primary objective of this project is to employ statistical and machine learning techniques to discern patterns and trends within the data, thereby uncovering the key determinants of happiness across nations.
- Initially, we will utilize multiple linear regression to establish a baseline model, which will provide insights into the linear relationships between predictor variables and the happiness index.
- Subsequently, we will explore the efficacy of alternative machine learning algorithms, such as decision trees, random forests, and support vector machines, in capturing non-linear interactions and improving predictive accuracy.
- To conduct model selection and optimization techniques such as cross-validation, hyperparameter tuning, and feature selection to enhance the predictive performance of the machine learning models.
- To provide recommendations for policymakers and stakeholders based on the predictive model's results to foster improvements in societal well-being and happiness.

Data Description

The project relies on accuracy of data. The data has been collected from World Happiness Report (WHR) 2024. This report used the data which was collected on the basis of Gallup World Poll (GWP). The survey measure of SWB (Subjective Well-being) is from the February 15, release of the Gallup World Poll (GWP). The Gallup World Poll (GWP) is indeed a widely recognized survey conducted by Gallup, an American analytics and advisory company. It measures various aspects of people's lives, including subjective well-being (SWB), across different countries and regions around the world. Subjective well-being typically encompasses factors such as happiness, life satisfaction, and overall quality of life, as perceived by individuals themselves. The GWP provides valuable insights into global trends and disparities in well-being, helping policymakers and researchers understand the factors that contribute to people's happiness and satisfaction.

The dataset contains observations covering years from 2005/06 to 2023. It consists of various socio-economic variables. Observations has been noted for almost all the countries.

Feature	Description	GWP Questions
Country Name	This variable represents the name of the country for each data entry. It's a categorical variable that identifies the geographic location or nation to which other variables correspond.	-
Year (2005 – 2023)	Year represents the time period or year in which the data was collected. It's a temporal variable that allows for the analysis of trends or changes over time.	-
Life Ladder (Happiness Index) (Target Variable) (0 – 10)	The Life Ladder, also known as the happiness index, is a measure of subjective well-being or life satisfaction. It is the national average response to the question of life evaluations.	On which step of the ladder would you say you personally feel you stand at this time?

	(0-worst possible life,10-best possible life)	
Log GDP per Capita	This variable represents the logarithm (typically natural logarithm) of the Gross Domestic Product (GDP) per capita of a country. GDP per capita is a measure of the economic output per person and is used to assess the economic well-being of a nation's residents. The data has been collected from World Development Indicators (WDI).	-
Social Support (0 – 1)	Social support measures the extent to which individuals perceive the availability of help and support from family, friends, and social networks. It's a psychological and social variable that influences overall well-being. It is the national average of the binary responses (either 0 or 1).	"If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
Healthy Life Expectancy at Birth	Healthy life expectancy at birth is an indicator of the number of years a person is expected to live in good health, without major illnesses or disabilities. It reflects the quality of life and healthcare access in a population. The data has been collected from WHO.	-
Freedom to Make Life Choices (0 – 1)	This variable assesses the perceived level of freedom individuals have to make life choices, such as decisions about their careers, relationships, and personal development. It's a measure of personal autonomy and agency. It is the national average of the binary responses (either 0 or 1).	Are you satisfied or dissatisfied with your freedom to choose what you do with your life?

Generosity (-1 – 1)	Generosity refers to the inclination or tendency of individuals to engage in charitable or altruistic behaviors, such as donating time, money, or resources to help others. It is the residual of regressing national average of the responses to the GWP question on GDP per capita.	Have you donated money to a charity in the past month?
Perceptions of Corruption (0 – 1)	Perceptions of corruption gauge how individuals perceive the level of corruption in their society or country. It reflects trust in institutions, transparency, and ethical standards within the public and private sectors. This measure is the national average of the binary responses to two questions in the GWP.	"Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?"
Positive Affect (0 – 1)	Positive affect measures the frequency or intensity of positive emotions and experiences individuals have, such as joy, contentment, and enthusiasm. It's a psychological variable linked to happiness and well-being. It is the national average of the binary responses (either 0 or 1) to three questions.	"Did you smile or laugh a lot yesterday?", "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?", "Did you learn or do something interesting yesterday?"
Negative Affect (0 – 1)	Negative affect represents the frequency or intensity of negative emotions and experiences, including sadness, worry, and anger. It's a psychological variable that can influence overall emotional well-being and life satisfaction. It is the national average of the binary	"Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Worry?", "Did you experience the following feelings during A LOT OF THE DAY yesterday?

	responses (either 0 or 1) to various questions.	How about Sadness?", and "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Anger?

The link of the source from where the dataset has been collected is given below:

https://worldhappiness.report/ed/2024/#appendices-and-data

The link for the entire dataset is given below:

https://docs.google.com/spreadsheets/d/146d1lWLZBB_e2ITZEusrBwjtbQe0yNwi/edit?usp=drive_link&ouid=114123193263949191412&rtpof=true&sd=true

Code File Description

In this report, we showcase an analysis performed using Python, utilizing a range of libraries tailored for visualization and data analysis tasks. Python stands out as a flexible programming language, boasting an array of libraries that streamline data manipulation, visualization, and statistical computations. The analysis aimed to explore and derive insights from a given dataset, utilizing Python's capabilities and the power of its libraries.

Analysis Process:

1. Data Loading: The analysis initiates by importing the dataset into a Pandas DataFrame for subsequent examination and manipulation.
2. Exploratory Data Analysis: Utilizing various statistical and graphical methods to uncover insights into the dataset's structure and attributes. This encompasses summary metrics, distribution analysis, and visualization of key variables.
3. Statistical Analysis: Leveraging the Statsmodels library to conduct advanced statistical procedures like regression analysis, time series analysis, and hypothesis testing.
4. Visualization: Employing Matplotlib and Seaborn to generate visual depictions of the data, including histograms, scatter plots, box plots, heatmaps, among others. These visualizations assist in discerning patterns, trends, and associations within the dataset.

Imported Libraries:

NumPy (np): NumPy is a fundamental package for scientific computing with Python, providing support for arrays, matrices, and mathematical functions.

Pandas (pd): Pandas is a powerful library for data manipulation and analysis, offering data structures and operations for manipulating numerical tables and time series.

Matplotlib.pyplot (plt): Matplotlib is a comprehensive library for creating static, interactive, and animated visualizations in Python. The "pyplot" module provides a MATLAB-like interface for plotting.

Seaborn (sns): Seaborn is a statistical data visualization library built on top of Matplotlib, providing high-level functions for drawing attractive and informative statistical graphics.

Statsmodels.api (sm): Statsmodels is a Python module that provides classes and functions for estimating many different statistical models, as well as for conducting statistical tests and exploring data.

scipy.stats: This module provides a wide range of statistical functions and distributions for scientific computing and statistical analysis.

sklearn: This library, also known as scikit-learn, is a comprehensive machine learning library for Python. It includes various tools for data preprocessing, model selection, and evaluation, making it a fundamental component of many machine learning workflows.

colorcet is a Python library that provides a collection of perceptually uniform colormaps for use in data visualization.

The link for the entire python notebook (code + output) is given below:

HappinessIndex_WHR_Final.ipynb - Colab (google.com)

Results and Interpretations

- **Data Understanding:**
  The first step involves for any statistical analysis is to understand the data by exploration and validation of data. This includes checking type of data, null values, missing values, calculating descriptive statistics, etc.
  After performing these techniques, we found out that we have 1 object data type, 9 float type and 1 integer data type variables in our dataset.

```
dtypes: float64(9), int64(1), object(1)
```

  Then we checked for null values in our dataset and found out 386 null values among all the variables.

```
Country name                        0
year                                0
Life Ladder                         0
Log GDP per capita                 28
Social support                     13
Healthy life expectancy at birth   63
Freedom to make life choices       36
Generosity                         81
Perceptions of corruption         125
Positive affect                    24
Negative affect                    16
```

  We handled these null values by dropping the entire row containing null values since we had large number of observations.

  There are 155 countries in total in our dataset.

- **Exploratory Data Analysis (EDA):**

```
Mean Life Ladder by Country name:
Country name
Afghanistan    3.505506
Albania        5.072808
Algeria        5.268920
Angola         4.420299
Argentina      6.289722
                 ...
Venezuela      6.042505
Vietnam        5.524966
Yemen          3.932129
Zambia         4.365957
Zimbabwe       3.792418
Name: Life Ladder, Length: 155, dtype: float64
```

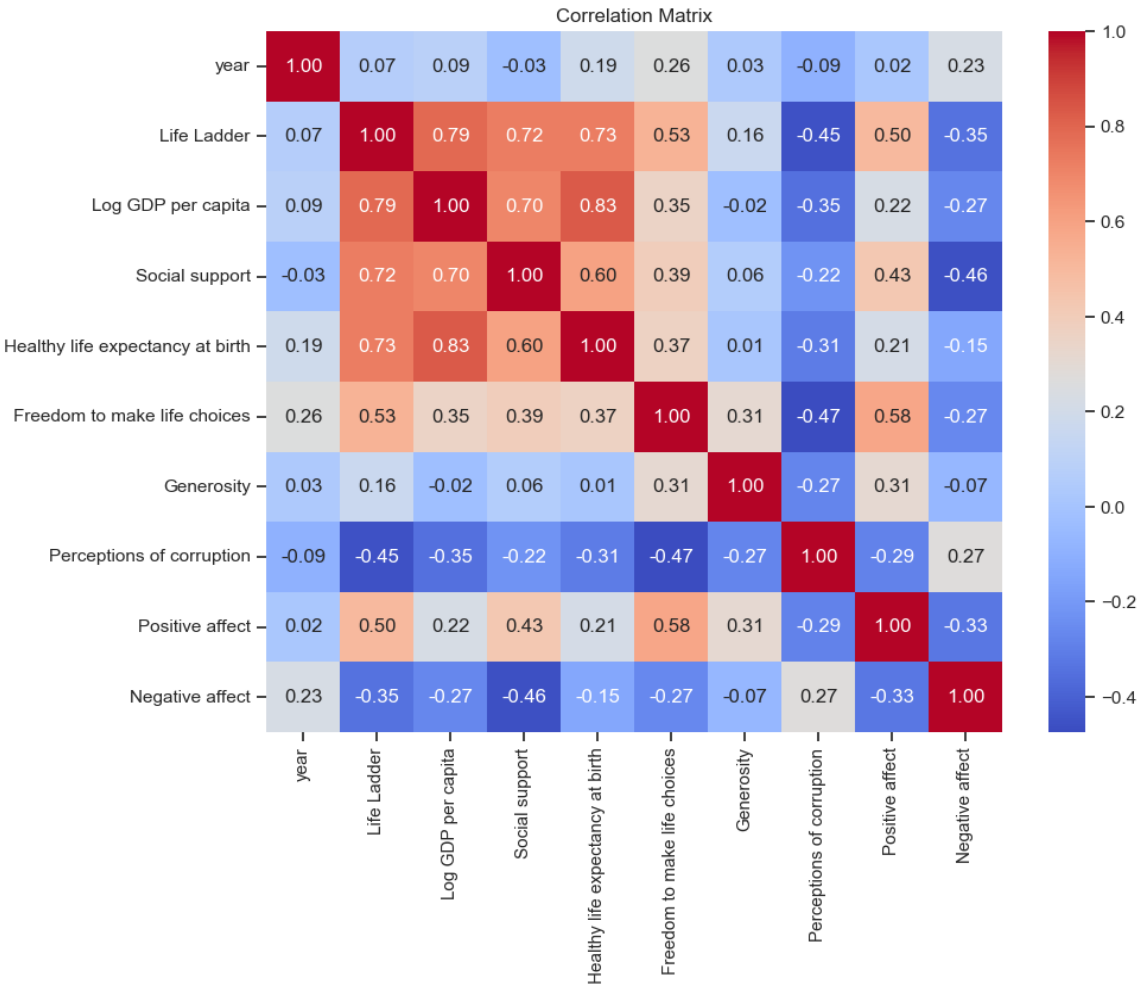
We calculated the mean “Life ladder” (Happiness Index) for each and every country.

- **Correlation Analysis:**

  While performing any type of regression analysis the foremost step is to check the correlation between the dependent variable and other independent variables and also between each and every independent variable.

  Generally, it is preferable that the dependent variable and other independent variables should possess high magnitude of correlation.

  To check this, we found out the correlation matrix and made a correlation plot (Heat map) based on this matrix which is shown below:



It is observable that the “**Life Ladder**” (target variable) is significantly correlated with almost all the variables excluding some such as “Generosity” and “Negative affect”. “Life Ladder” is

highly correlated (positively) with “Log GDP per capita”, “Social support” and “Healthy life expectancy at birth”.
Examining the correlation heat map suggests the potential presence of multicollinearity. Multicollinearity arises when independent variables within a regression model exhibit significant correlation, potentially distorting coefficient estimates and complicating the model's interpretation. The presence of strong correlations suggests that alterations in one variable may be tightly associated with changes in another. Detecting and addressing multicollinearity is crucial for refining our comprehension of data relationships and improving the accuracy of predictive modelling.
We identified that “Log GDP per capita” is highly correlated with “Social support” and “Healthy life expectancy at birth”.
So, it would be better to omit these correlated independent variables while fitting model to avoid multicollinearity to get better estimates.

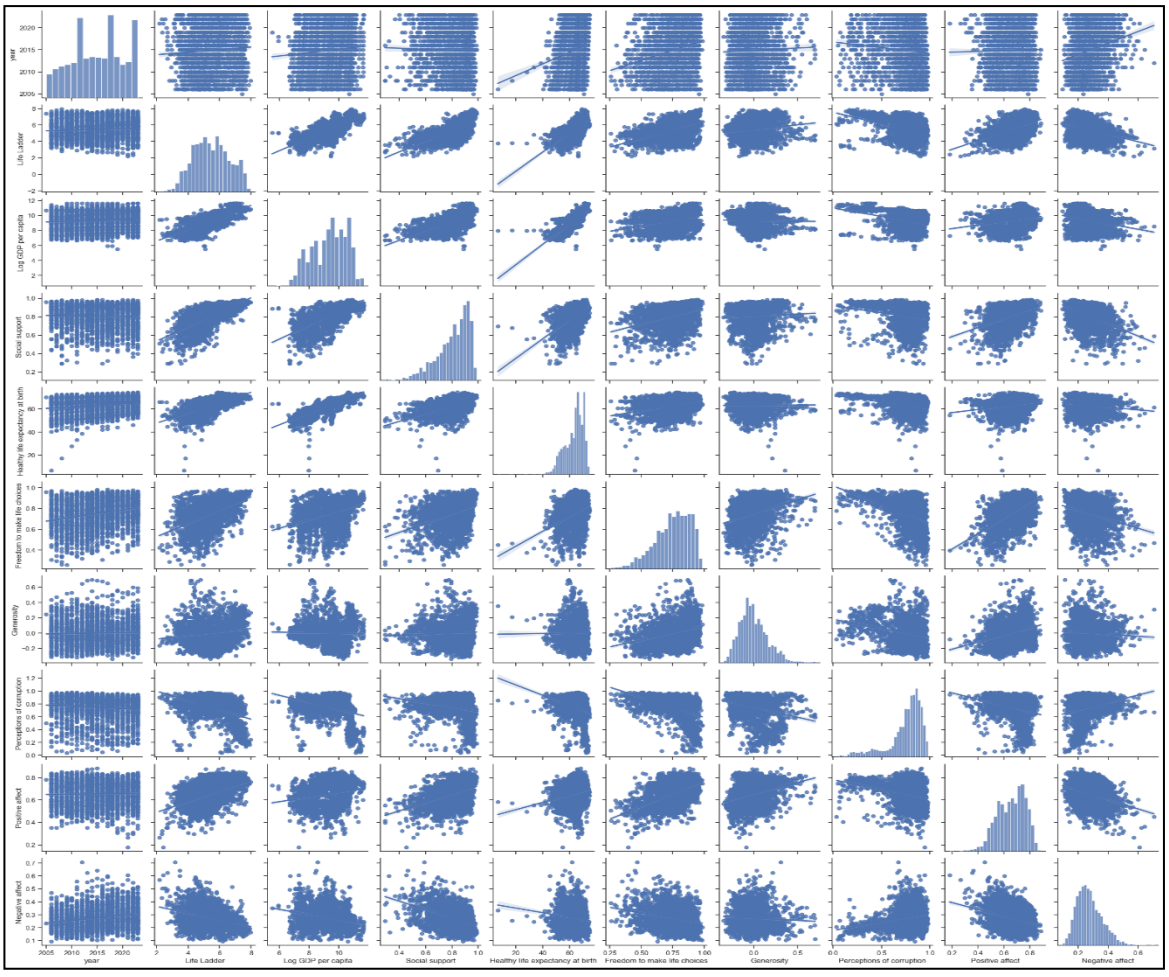
- **Multicollinearity Check:**
  Since, we suspected the presence of multicollinearity from the correlation heat map, so Variance Inflation Factor (VIF) which is a statistical measure used to detect multicollinearity in regression analysis. The higher the VIF value, the stronger the multicollinearity. Generally, a VIF value exceeding 5 or 10 is considered problematic, although the threshold can vary depending on the context.
  For this dataset, we found out the VIF scores for each and every independent variable which is given below:

	Feature	VIF
0	year	0.195137
1	Life Ladder	0.779484
2	Log GDP per capita	0.803062
3	Social support	0.656717
4	Healthy life expectancy at birth	0.728264
5	Freedom to make life choices	0.527391
6	Generosity	0.185193
7	Perceptions of corruption	0.361063
8	Positive affect	0.492837
9	Negative affect	0.338200

We can note that the VIF scores for all the independent variables are less than 1 indicating **no presence of multicollinearity**.

Basically, when the VIF is less than 1, it implies that the variance of the estimated regression coefficients is not inflated due to multicollinearity. Each independent variable provides unique information to the regression model, and there is no redundancy or overlap among them. If multicollinearity would have been detected through high VIF values, it's important to address it to improve the reliability of the regression model. This may involve removing highly correlated variables, transforming variables, or using regularization techniques.

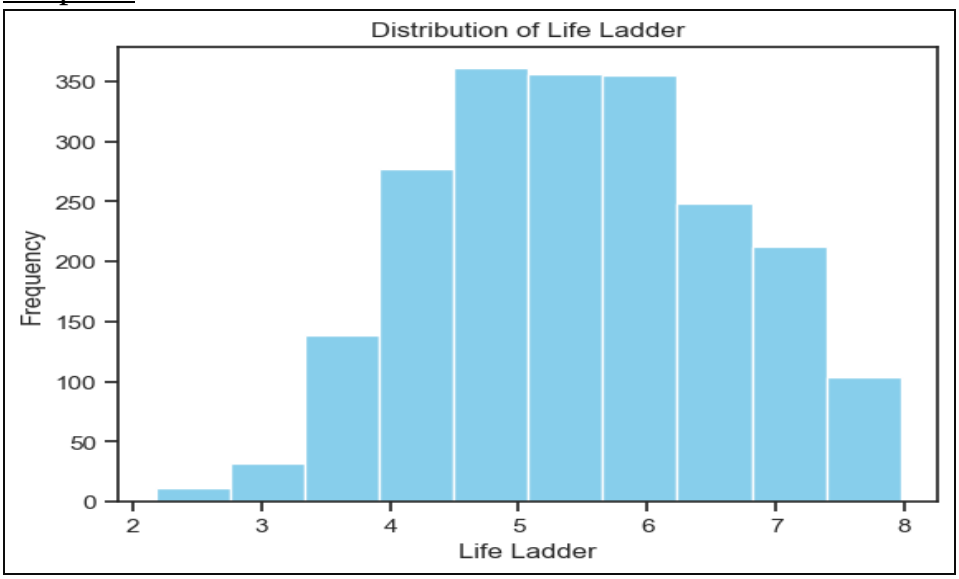
- **Linearity Check:**
  After checking the correlation matrix, our next step is to check the linearity assumption between the dependent variable and the independent variables.
  So, to check the validity of this assumption we made scatter plots between the variables.



From this scatter plot, we can see that there is a linear relationship between "Life Ladder" and "Log GDP per capita", "Social support" and "Healthy life expectancy at birth". But for the other variables linearity assumptions is not valid. Since, some of the independent variables are in a linear relationship with the dependent variable, so we can perform **Multiple Linear Regression analysis** on this dataset.

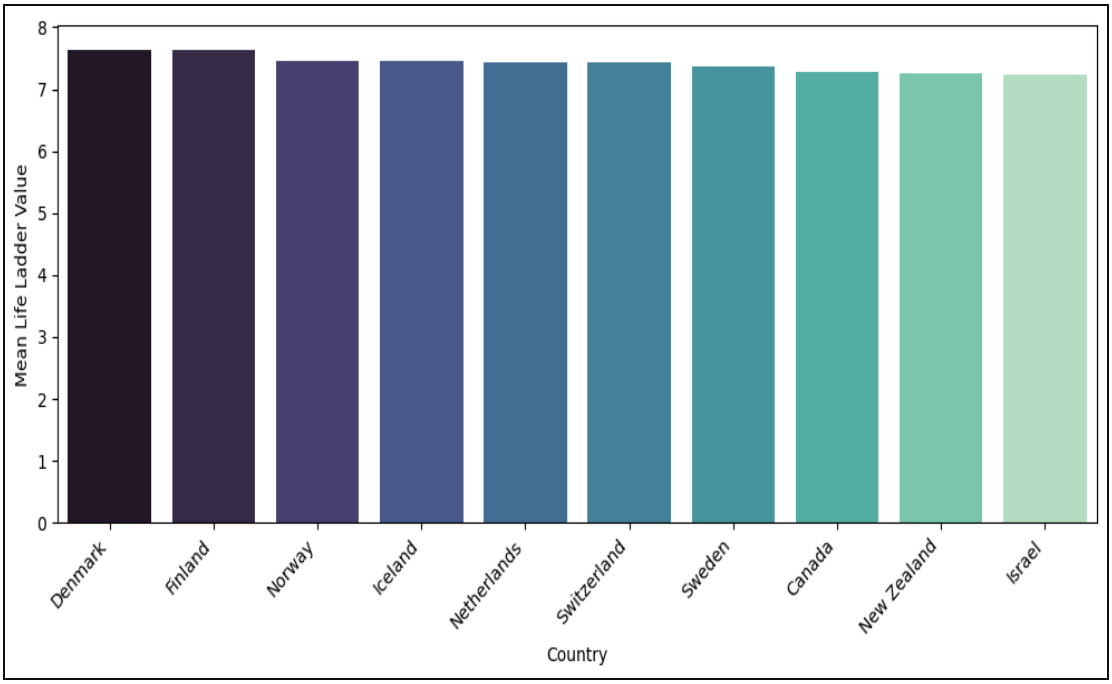
- **Data Visualization:**

Graph 1:



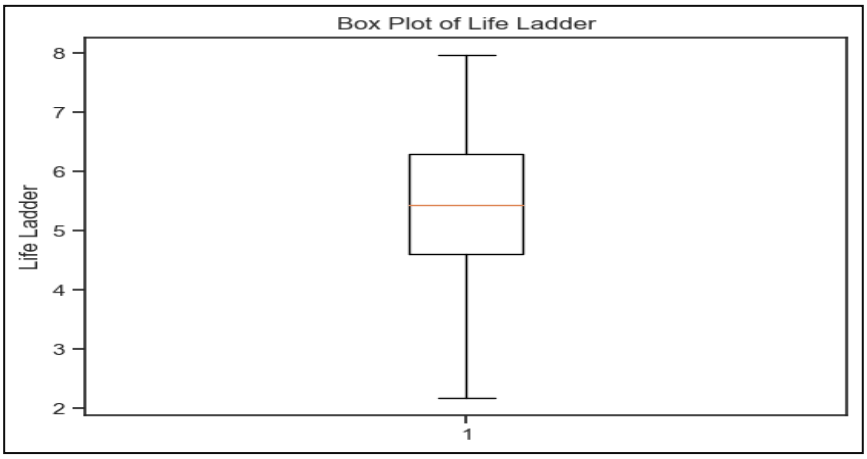
We created histogram visualization of "Life Ladder" to see the distribution of "Life Ladder". It shows how frequently certain happiness score occur in the dataset. The histogram is divided into 10 intervals along the x-axis each representing a range of happiness score. By looking at the histogram we can claim that the frequencies of happiness score follows Normal distribution.

Graph 2:



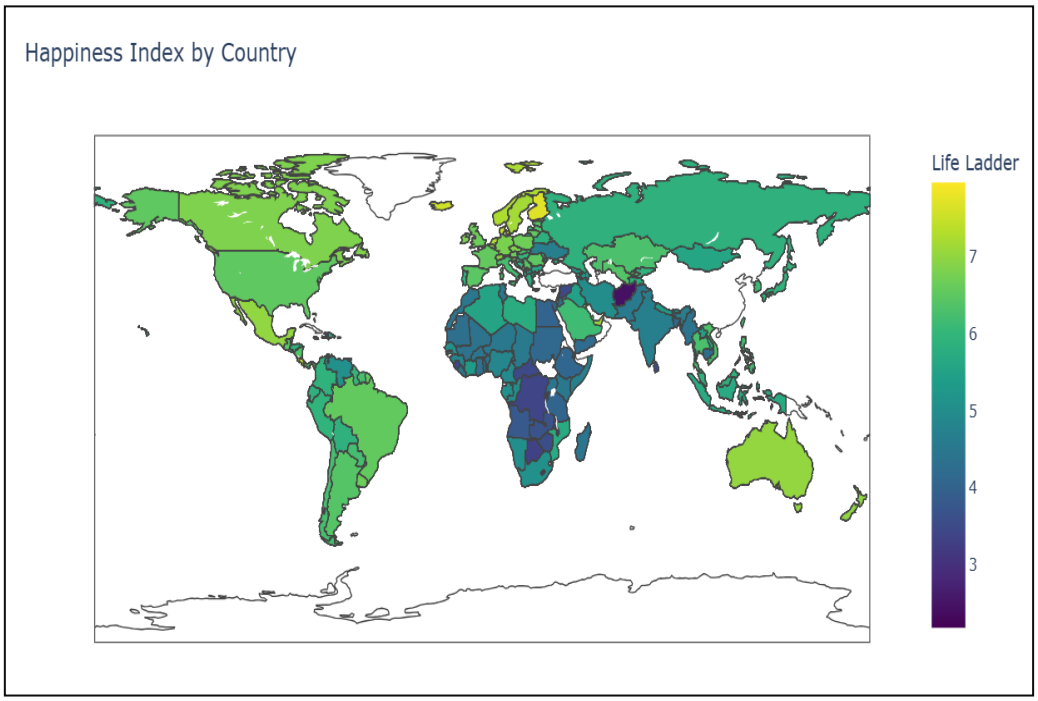
From the above plot we can observe that "Denmark" has the highest mean life ladder score (i.e., happiness score) around **8** further "Finland" has almost equal mean life ladder score i.e., around 7.5.

Graph 3:



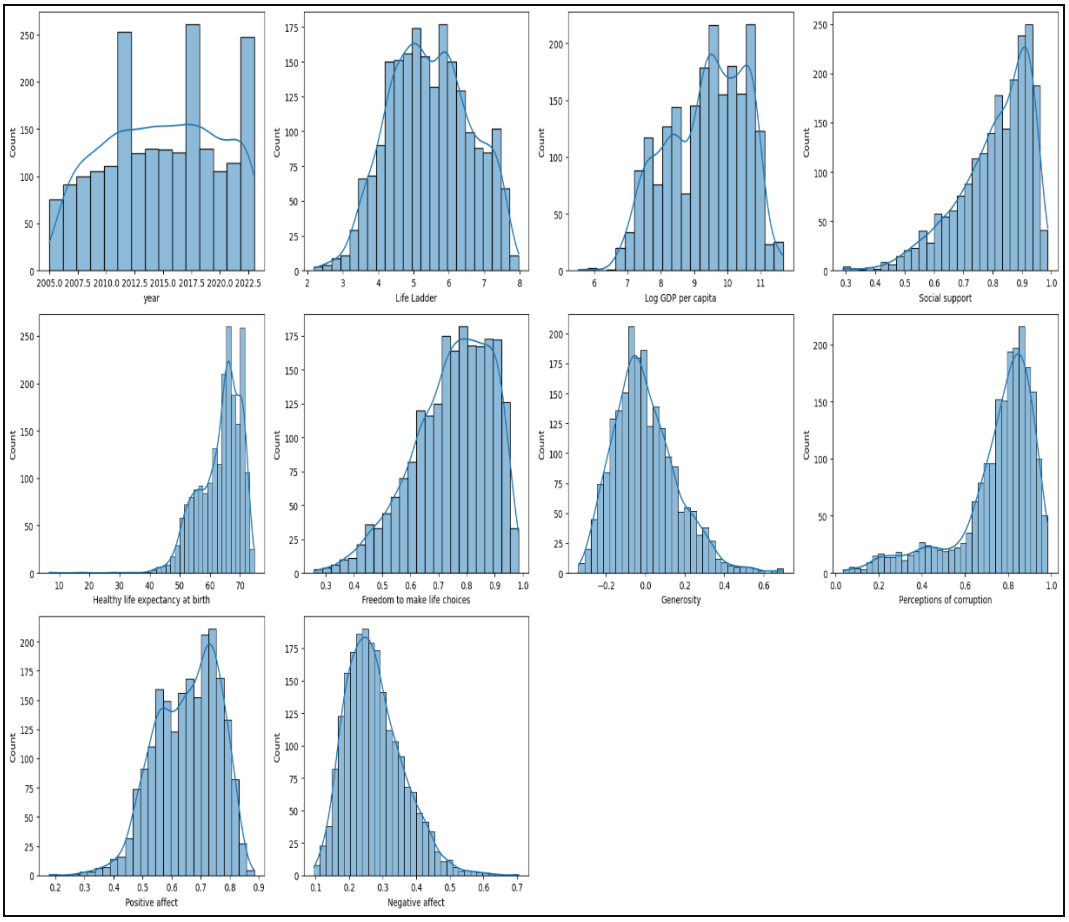
We created Box plot visualization that source the distribution of "Life Ladder". It is particularly useful for visualizing the central tendency, dispersion, and skewness of the data.
From all the Box plot it is evident that there are no outliers present in the "Life ladder" variable. It has shorter box and longer whiskers suggesting greater variability of "Life Ladder".

Graph4:



It is a choropleth map, a thematic map where areas are shaded with a colour gradient ranging from light to dark in proportion to the value of "Life Ladder" score being represented. Lighter colour typically representing higher happiness index values while darker colours representing lower happiness index value. By observing the overall pattern of colours across the map, we can identify the regions or cluster of countries with similar levels of happiness.

Graph 5:



This plot generates a grid of histogram for each numeric column in the dataset. Each histogram represents the distribution of values for a specific numeric variable in the dataset. By examining these histograms collectively, we gained insights into the overall distributional characteristics of the numeric variables in the dataset, which is valuable for exploratory data analysis and understanding the dataset's properties.

'Social Support', 'Freedom to make life choices', 'Generosity' and 'Negative affect' are resembling bell-shaped distributions indicating Normal distribution.

- **Data Pre-Processing:**

  Data preprocessing is a crucial step in the data analysis pipeline that involves scaling, transforming, and organizing raw data into a format suitable for analysis or modelling.

  - **Scaling Of Data:**

    Scaling refers to the process of transforming the features of a dataset to a similar scale. This is done to ensure that no particular

feature dominates due to its scale and to make comparisons between different features more meaningful.

For this dataset we performed **Standard scaling** for all the variables excluding the object datatype variable ('Country'). It is also known as z-score normalization. It is a technique used in data preprocessing to standardize the range of independent variables or features in a dataset. It involves transforming the data such that it has a mean of 0 and a standard deviation of 1. This process ensures that the features have the same scale, which can be important for certain machine learning algorithms that are sensitive to the scale of the input features. This will help us to give improved model performance.

- **Transformation of Data:**

  Data transformation refers to the process of altering the original data in order to make it more suitable for analysis or modelling.

  Firstly, we performed Box-Cox transformation in our dataset. Then we chose **Yeo-Johnson transformation** (modification over Box-Cox transformation) over other transformations because initially we noted that our data has high variance and not all the variables was following normal distribution. It is used to stabilize the variance of a dataset and make it more normally distributed.

  Initially, we fitted an OLS model on the original data but the model was overfitted (R^2 value was coming to be 1) and multicollinearity was present to a high order. So, to get a better model and to resolve the issue of multicollinearity, we performed different transformations and Box-Cox transformation (Yeo-Johnson transformation) was giving the better results among others. So, we proceeded with Yeo-Johnson transformation.

- **Splitting Of Data:**

  Splitting of data refers to the process of dividing a dataset into multiple subsets for various purposes such as model training, validation, and testing. The most common type of data splitting is into training, validation, and testing sets.

  Data splitting is important as it allows us to assess the performance of machine learning models accurately. By dividing the dataset into training, validation, and testing sets, we can train models on one subset, tune their parameters on another, and evaluate their performance on yet another, ensuring that the models generalize

well to unseen data. This process helps prevent overfitting, where a model performs well on the training data but poorly on new data, and allows us to build more reliable and robust models for real-world applications.

For this dataset, we split our data into three sets: **training, validation, testing** in the ratio **60:20:20**. Where model has been fitted on 60% of the data, hyperparameter tuning has been done on 20% of data and prediction was done on the remaining 20% of data.

- **Model Fitting:**

  Model fitting, also known as model training or model estimation, is the process of using a machine learning algorithm to learn patterns and relationships within the training data. The goal of model fitting is to create a mathematical representation (the model) that can accurately predict the target variable based on the input features.

  We fitted different supervised learning models like multiple linear regression models (OLS), Decision Tree, Random Forest, Gradient Boosting etc.

  - **Baseline Model:**

    A baseline model, is a simple or naive model that serves as a starting point for comparison with more complex models.

    The purpose of a base model is to establish a benchmark against which the performance of more sophisticated models can be evaluated.

    We fitted the baseline model on our trained data and obtained the following accuracies score:

Accuracy Measure	**Accuracy Score**
Mean Squared Error (MSE)	0.9106224638120071
Mean Absolute Percentage Error (MAPE)	101.41708110762826
Mean Absolute Error (MAE)	0.7968041735734686

    These scores will be used further to compare other models.

  - **Multiple Linear Regression (OLS):**

    Ordinary Least Squares, which is a method used in multiple linear regression analysis to estimate the parameters in a linear regression model. The goal of OLS is to find the line (or hyperplane in higher

dimensions) that minimizes the sum of the squared differences between the observed and predicted values.

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n * X_n + \varepsilon$$

Where,
Y is the dependent variable.
X_1, X_2, ..., X_n are the independent variables.
B_0, β_1, β_2, ..., β_n are the coefficients (parameters) to be estimated.
ε is the error term.

- **Including all the variables:**

OLS Regression Results

Dep. Variable:	Life Ladder	R-squared:	0.800
Model:	OLS	Adj. R-squared:	0.799
Method:	Least Squares	F-statistic:	624.9
Date:	Mon, 22 Apr 2024	Prob (F-statistic):	0.00
Time:	03:51:53	Log-Likelihood:	-795.30
No. Observations:	1257	AIC:	1609.
Df Residuals:	1248	BIC:	1655.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0072	0.013	-0.556	0.578	-0.033	0.018
Log GDP per capita	0.3439	0.030	11.391	0.000	0.285	0.403
Social support	0.2620	0.022	11.839	0.000	0.219	0.305
Healthy life expectancy at birth	0.2037	0.027	7.501	0.000	0.150	0.257
Freedom to make life choices	0.0487	0.019	2.628	0.009	0.012	0.085
Generosity	0.0238	0.014	1.663	0.097	-0.004	0.052
Perceptions of corruption	-0.0883	0.016	-5.679	0.000	-0.119	-0.058
Positive affect	0.2228	0.016	13.510	0.000	0.190	0.255
Negative affect	0.0298	0.015	1.958	0.051	-6.64e-05	0.060

Omnibus:	37.563	Durbin-Watson:	1.936
Prob(Omnibus):	0.000	Jarque-Bera (JB):	50.377
Skew:	-0.317	Prob(JB):	1.15e-11
Kurtosis:	3.748	Cond. No.	5.67

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Interpretation:

- From the above table, it is observed that R^2 value is 0.800 indicating that approximately 80% of the variance in "Life Ladder" (dependent variable) is explained by the

independent variables included in the model. This suggests that the model provides a good fit to the data, capturing a large proportion of the variability in the dependent variable. A high R^2 value like this indicates that the model's predictions align closely with the actual values of the target variable. It indicates a strong relationship between the independent and dependent variables in your model, suggesting that your model has good explanatory and predictive capabilities.

- Adjusted R^2 value coming to be 0.799 value almost same as the R^2. Adjusted R-squared is a modified version of the R-squared value that accounts for the number of predictors in the model. It penalizes the R-squared value for including additional predictors, helping to prevent overfitting. Therefore, the adjusted R-squared value provides a more conservative measure of model fit compared to the regular R-squared. This suggests that the model provides a good fit to the data
- The F-statistic value coming out to be 624.9 suggesting that the regression coefficients are not all equal to zero. In other words, it indicates that at least one of the independent variables in the model has a non-zero coefficient, meaning that the model as a whole is statistically significant meaning that there is strong evidence that the independent variables collectively have an effect on the dependent variable. The model fits the data well and provides valuable information about the relationship between the predictors and the outcome.
- The model with the AIC value of 1609 has a higher penalty for complexity compared to the model with the BIC value of 1655. This suggests that the AIC may favor a more complex model. An AIC value of 1609 suggests that the model provides a good balance between goodness of fit and complexity, making it a strong candidate compared to other models. However, its interpretation should always be considered in the context of model comparison and the specific goals of the analysis.
- Durbin-Watson statistic value coming to be 1.936 indicating that there is likely little or no autocorrelation present in the residuals of your regression model. However, it's essential to

interpret this statistic in conjunction with other diagnostics and consider the specific characteristics of your dataset and research question.

- The condition number is a measure of how sensitive a function (or system of equations) is to changes or errors in the input data. Specifically, it's the ratio of the largest and smallest eigen values of a matrix. For our model condition no. is 5.67 which is pretty low. A low condition number is indicating that no multicollinearity is present, where independent variables are not highly correlated with each other. This will help to estimate the coefficients of the regression model accurately and can lead to unstable or unreliable predictions.

Conclusion:

- Since, from our model it is evident that no multicollinearity is present between the independent variables. Also, the R^2 value is high. So, we can say that this is a good model. But there may be a good model than this. Hence, we will try to fit an enhanced model by excluding different variables. Basically, in model fitting our main motive is to fit a model which is simple to understand without losing any information about the model. So, here also we will try to build a model with a smaller number of variables by eliminating the statistically insignificant variables from our model without reducing the accuracy scores of the model.

- **After excluding statistically insignificant Variable(only "Negative effect"):**

OLS Regression Results

Dep. Variable:	Life Ladder	R-squared:	0.800
Model:	OLS	Adj. R-squared:	0.798
Method:	Least Squares	F-statistic:	712.0
Date:	Mon, 22 Apr 2024	Prob (F-statistic):	0.00
Time:	04:20:30	Log-Likelihood:	-797.23
No. Observations:	1257	AIC:	1610.
Df Residuals:	1249	BIC:	1652.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0080	0.013	-0.619	0.536	-0.033	0.017
Log GDP per capita	0.3424	0.030	11.334	0.000	0.283	0.402
Social support	0.2452	0.020	12.005	0.000	0.205	0.285
Healthy life expectancy at birth	0.2128	0.027	7.944	0.000	0.160	0.265
Freedom to make life choices	0.0484	0.019	2.613	0.009	0.012	0.085
Generosity	0.0252	0.014	1.760	0.079	-0.003	0.053
Perceptions of corruption	-0.0842	0.015	-5.458	0.000	-0.114	-0.054
Positive affect	0.2199	0.016	13.372	0.000	0.188	0.252

Omnibus:	38.683	Durbin-Watson:	1.939
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51.433
Skew:	-0.327	Prob(JB):	6.78e-12
Kurtosis:	3.744	Cond. No.	5.51

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Interpretation:

- Even after removing the statistically insignificant variables there was no change in the value of R^2 and adjusted R^2 suggesting that the model is still a good fit. F-statistic value has been increased by small margin though indicating that the model is still statistically significant. Also, the DW test statistic value has been increased by very small margin but still resulting absence of autocorrelation. Similarly, there is a small change in the condition number which has been reduced by small margin (5.51). Since, the condition number is much smaller than 30 suggesting no multicollinearity is present.

Thus, we will consider this as an improved model over the previous model.

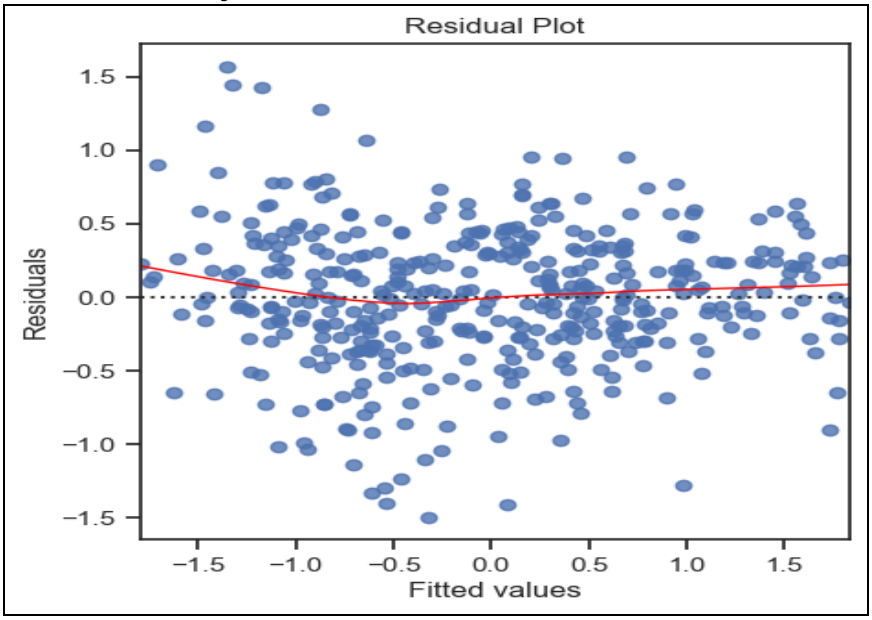
- **Final model after excluding certain variables:**

  Life Ladder = −0.0080 + (0.3424×Log GDP per capita) + (0.2452×Social support) + (0.2128×Healthy life expectancy at birth) + (0.0484×Freedom to make life choices) + (0.0252×Generosity) − (0.0842×Perceptions of corruption) + (0.2199×Positive affect)

➢ **Diagnostic Check:**

In regression analysis, diagnostic checks are essential for ensuring that the model assumptions are met and that the model adequately captures the relationship between the dependent and independent variables.

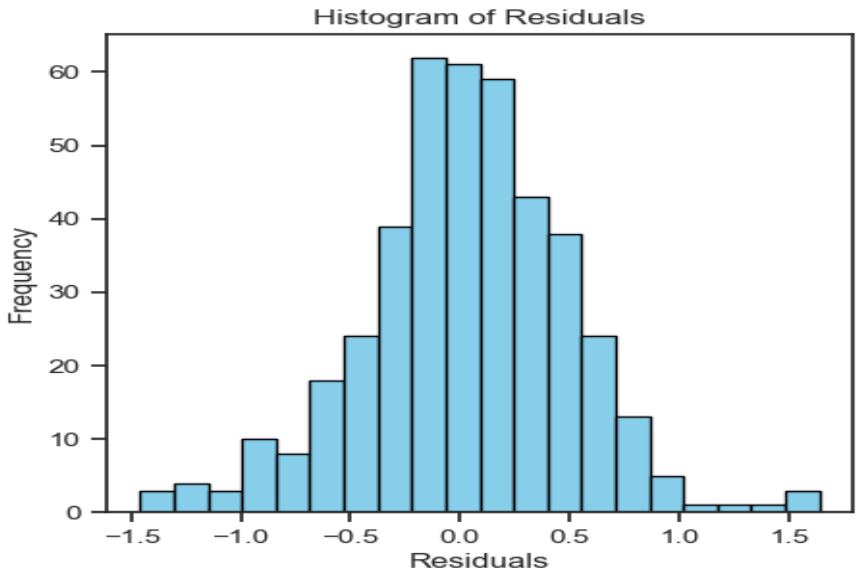
- **Residual analysis:**



Examining the residuals is fundamental after any model fitting. The above residual plot (scatter plot) of residuals against fitted values has been plotted to identify patterns indicating violations of assumptions like linearity, homoscedasticity, and normality. From the plot we can observe that the residuals are randomly scattered around the horizontal line at y = 0, with no discernible pattern, it suggests that the model's assumptions are met, and the regression model is appropriate for the data.

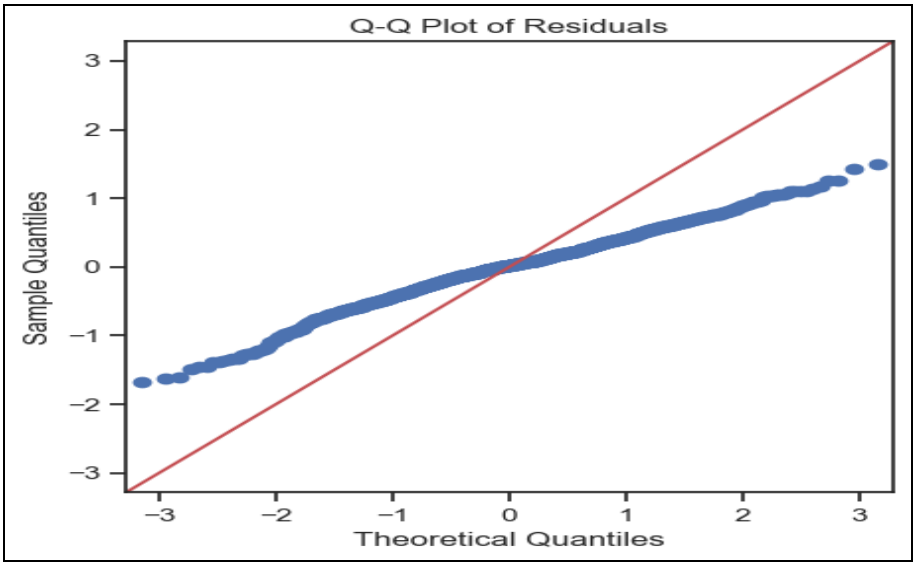
- **Normality of Residuals:**
  We need to check the assumption of normality of residuals by plotting histograms, Q-Q plots, Shapiro-Wilk test.

  Histogram-



  From the histogram, we can see that the residuals are not skewed but also not forming bell-shaped curve and also some spikes can be observed during decaying hence indicating that the residuals are not normally distributed. To cross-check this indication/inference we performed Shapiro-Wilk test for normality and will plot Q-Q plot.

  Q-Q Plot-



A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a dataset follows a particular distribution, typically the normal distribution.

We can see that the data points does not fall approximately along a straight line, it suggests that the sample data follows the normal distribution.

Shapiro-Wilk Test-

The Shapiro-Wilk test is a statistical test used to assess whether a given sample of data comes from a normally distributed population. The null hypothesis of the Shapiro-Wilk test is that the data are normally distributed. For our model, after performing this test, we found the result given below:

```
Shapiro-Wilk Test Statistic: 0.9886190295219421
P-value: 0.002343298401683569
The residuals are not normally distributed (reject null hypothesis)
```

From the above result we can infer that the residuals are not normally distributed since the p-value is much less than 0.05 resulting rejection of null hypothesis.

- **Homoscedasticity:**

  Next step is to Assess whether the variance of the residuals is constant across different levels of the independent variables. This is checked by performing Breusch-Pagan Test.

  Breusch-Pagan test-

The Breusch-Pagan test is a statistical test used to detect heteroscedasticity in a regression model. Heteroscedasticity occurs when the variance of the errors (residuals) is not constant across all levels of the independent variables. After performing this test on the

```
Breusch-Pagan Test Results:
LM Statistic: 27.027244851186474
LM P-value: 0.0003295377860569753
F Statistic: 4.047981419574158
F P-value: 0.0002631459157465393
```

residuals of the model, we got the following result:

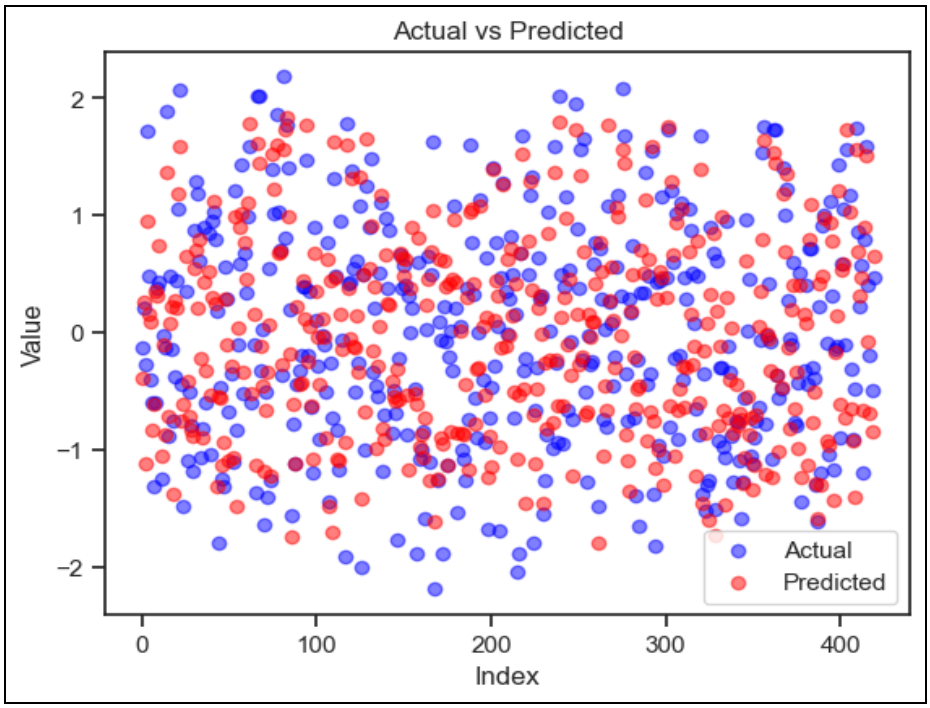
From the above result we can infer that the p-value associated with LM statistic is much less than 0.05, so we rejected our null hypothesis of homoscedasticity. This indicates that there is evidence of heteroscedasticity in the regression model.

- **Autocorrelation:**
  The next step is to check for the autocorrelation between residuals. In simpler words we need to check whether the residuals are independent of each other over time. Durbin Watson statistic is generally used to detect autocorrelation.
  Dubin Watson Test –
  The Durbin-Watson statistic is a measure used to detect the presence of autocorrelation in the residuals of a regression model. The Durbin-Watson statistic ranges in value from 0 to 4, with:
  A value around 2 indicates no autocorrelation, a value significantly less than 2 suggests positive autocorrelation, a value significantly greater than 2 suggests negative autocorrelation.
  After performing this test on the residuals of our model we found that the DW test statistic value is coming to be 1.991313, which is close to 2, suggesting that there is no significant autocorrelation present in the residuals.

- **Comparison between Actual and Predicted Values:**

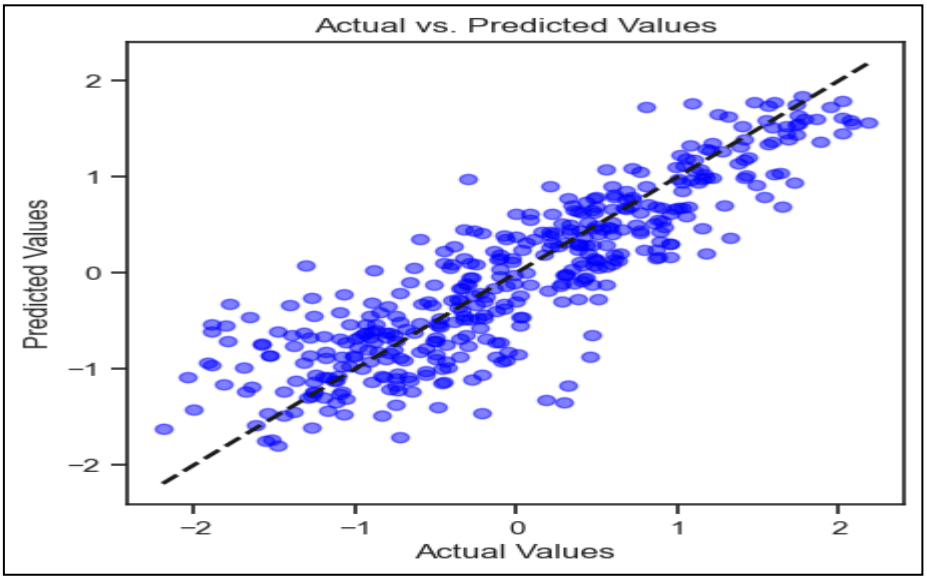
	Actual	Predicted
0	-0.124743	-0.396568
1	0.207329	0.261754
2	-0.276038	-1.115723
3	1.717140	0.945771
4	0.485452	0.155976
...	...	...
415	1.590406	1.512890
416	0.400691	-0.078771
417	-0.189326	-0.685684
418	-0.496508	-0.849991
419	0.468354	0.652136

420 rows × 2 columns



- **Accuracy Checking:**
  Accuracy checking is a crucial step in assessing the performance of a predictive model, particularly in the context of supervised learning tasks. It involves evaluating how well the model's predictions align with the actual outcomes.

Actual VS Predicted Plot-



It is a graphical way to visually assess the performance of a regression model by comparing the actual values of the dependent variable with the predicted values generated by the model.

From the above graph we can conclude that the predicted values (blue dots) are not that much close to the actual value (diagonal line), suggesting that the model's predictions are not that much accurate and there is may not be a good fit between the actual and predicted values.

Coefficient of determination (R^2):

R^2 measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. A higher R-squared value indicates a better fit of the model to the data.
For this model, R^2 value for the 3 different sets are listed below:

Training	Validation	Test
0.7996162275775848	0.7996139964195729	0.7570390995597313

The R-squared score of 0.7996 indicates that our model explains approximately 79.96% of the variance in the dependent variable within the training dataset. This means that the model fits the training data well, capturing a large

portion of the variability in the dependent variable. The performance metric on the training and validation sets is very close, suggesting that the model is not overfitting to the training data and generalizes well to unseen validation data. The performance metric on the test set is slightly lower than on the training and validation sets. This is expected because the test set represents unseen data, and it's common for model performance to decrease slightly when applied to new data.

Mean Absolute Error (MAE):

MAE measures the average absolute difference between the actual and predicted values. It provides a straightforward interpretation of the average prediction error. Lower MAE values indicate better predictive accuracy.

For this model, MAE value for the 3 different sets are listed below:

Training	Validation	Test
0.34646588140047585	0.33530654013413835	0.3614665295046143

The value of MAE of training set is approximately 0.3465. This means that, on average, the predictions of our model are off by around 0.3465 units from the actual values in the training dataThe validation MAE score of approximately 0.335 indicates a similar level of performance to the training phase. This suggests that the model's generalization performance, as evaluated on a separate validation dataset (not used during training), is consistent with its performance on the training dataset. The slightly higher MAE score in the test set compared to the training and validation phases suggests that the model may not generalize as well to new, unseen data. The slightly higher MAE compared to the training and validation phases suggests that the model may not generalize as well to new, unseen data.

Mean Square Error (MSE):

MSE measures the average squared difference between the actual and predicted values. It penalizes large errors more than smaller ones. Like MAE, lower MSE values indicate better predictive accuracy.

For this model, MSE value for the 3 different sets are listed below:

Training	Validation	Test
0.20816799022386595	0.19460409814608098	0.22123860696385525

The training MSE (0.2081) gauges the model's fit to training data, with lower values indicating better fit. The validation MSE (0.1946) assesses generalization to new data, while the test MSE (0.2212) estimates performance on unseen data. Lower MSE values signify better model performance overall, guiding decisions on model refinement and ensuring robustness in real-world scenarios.

Root Mean Square Error (RMSE):

RMSE is the square root of the MSE and provides an interpretation of the average prediction error in the same units as the dependent variable. It is widely used because it is easier to interpret than MSE.

RMSE value for the 3 different sets are listed below:

Training	Validation	Test
0.4562543043346177	0.44113954498104224	0.4703600822389749

A training RMSE of approximately 0.456 suggests that the model's predictions closely match the actual values in the training data. The validation RMSE of around 0.441 indicates that the model generalizes well to unseen data, maintaining similar performance as in the training set. However, the slightly higher test RMSE of approximately 0.470 suggests a slightly larger deviation between predicted and actual values in the test dataset, which could indicate some loss of generalization on completely unseen data but

still maintains reasonable predictive accuracy. Overall, the model demonstrates good performance on both training and validation data, with a slight drop in performance on the test set, indicating a robust but not perfect fit.

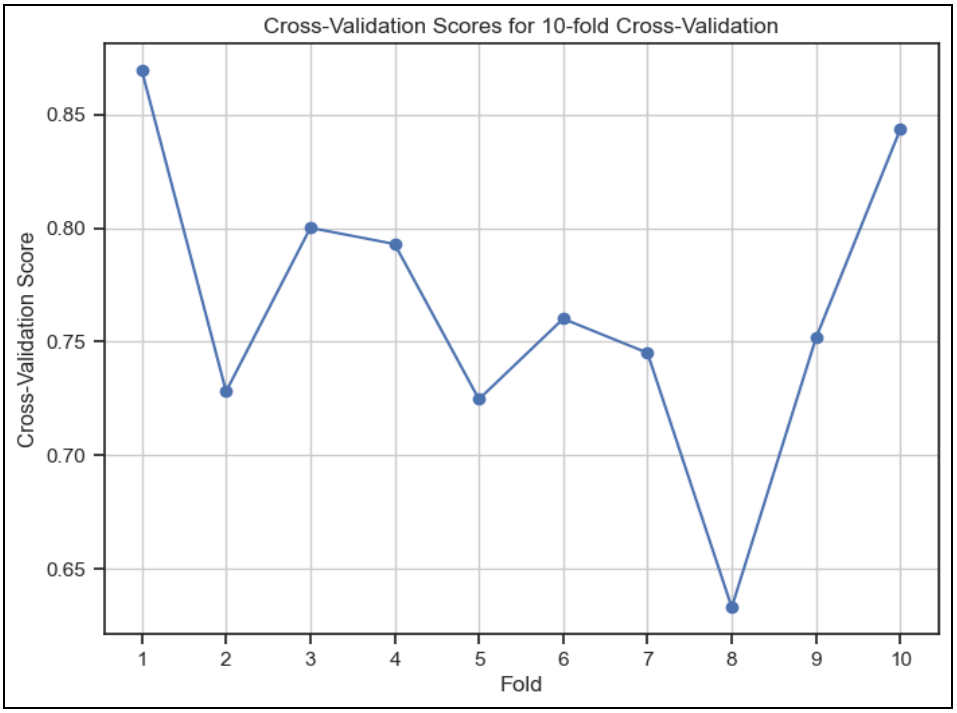
- **Performance Checking using k-fold Cross Validation:**
  K-fold cross-validation is a technique used to evaluate the performance of a machine learning model, especially when the dataset is limited or when there's a need to estimate how the model will perform on unseen data. Provides a more reliable estimate of the model's performance compared to a single train-test split, as it uses multiple splits of the data. Reduces the variance in the performance estimate, as each data point is used in both training and testing.
  We performed k-fold cross validation with linear regression model which involved splitting the dataset into 10-folds (taking k as 10), fitted linear regression model on each training set and then evaluated the model's performance on the correspondence validation set and test set.
  After performing this technique, we found out the cross-validation scores (R^2 value) and other metrics for different folds.

Fold	CV Scores	CV R2 Scores	CV RMSE Scores	CV MSE Scores
1	0.869438	0.869438	0.375927	0.141321
2	0.728107	0.728107	0.553985	0.306899
3	0.800162	0.800162	0.438359	0.192159
4	0.793015	0.793015	0.424094	0.179855
5	0.724677	0.724677	0.427077	0.182395
6	0.760107	0.760107	0.492426	0.242483
7	0.745091	0.745091	0.491723	0.241791
8	0.633019	0.633019	0.492464	0.242520
9	0.751697	0.751697	0.551871	0.304561
10	0.843573	0.843573	0.402361	0.161895
Mean	0.764889	0.764889	0.465029	0.219588

Best Fold Index: 0
Mean R2 Score for Best Fold: 0.869438424066613

The k-fold cross-validation results suggested that the linear regression model performed reasonably well in predicting the target variable, as indicated by the mean R2 score of approximately 0.765. This implies that about 76.5% of the variance in the target variable can be

explained by the independent variables in the model. The best-performing fold achieved an R2 score of 0.869, indicating high predictive accuracy in that particular fold. The RMSE and MSE scores also suggested that the model's predictions are relatively close to the actual values, with lower RMSE and MSE values indicating better performance. Overall, the k-fold cross-validation results provided confidence in the model's generalizability and predictive ability, suggesting that it can effectively capture the relationship between the independent and dependent variables in the dataset. These findings suggest that the model generalizes well to unseen data. However, it's important to note that the model's performance may vary across different folds, with some folds achieving higher R-squared scores than others. Therefore, it's essential to consider the variability in performance when assessing the model's overall effectiveness.



To visualize the difference between cross-validation scores for different folds, we plotted a graph showing the cross-validation scores for different folds where x-axis being the fold index and y-axis being the cross-validation scores. From the graph it is clear that the cross-validation score (R^2 value) for the 1^{st} fold is the highest among the

other folds (almost equals to 0.9) and 8th fold has the lowest among the other folds (nearly equals to 0.6).

We found out the metrics scores for all the three sets (i.e. training, test, validation) to get more insights.

```
Training Set:
MSE_CV: 0.20816799022386595
MAE_CV: 0.34646588140047585
R2_Score_CV: 0.7996162275775848
RMSE_CV: 0.4562543043346177

Test Set:
MSE_CV: 0.22123860696385525
MAE_CV: 0.3614665295046143
R2_Score_CV: 0.7570390995597313
RMSE_CV: 0.4703600822389749

Validation Set:
MSE_CV: 0.19460409814608098
MAE_CV: 0.3353065401341383
R2_Score_CV: 0.7996139964195729
RMSE_CV: 0.44113954498104224
```

These metrics assess the performance of the model in predicting the target variable across different datasets. The model exhibits slightly better performance on the training and validation sets compared to the test set. This observation is supported by lower MSE, MAE, and RMSE values, as well as a higher R-squared score on the training and validation sets. The R-squared score, which measures the proportion of variance in the target variable explained by the model, is approximately 80% on both the training and validation sets. On the test set, the R-squared score is slightly lower, around 76%. This indicates that the model captures a substantial portion of the variability in the target variable across all datasets. Despite minor differences in performance metrics across datasets, the model demonstrates consistency in its predictive ability. Its performance remains robust when generalized to unseen data, suggesting its reliability in making accurate predictions beyond the data it was trained on.

- **Decision Tree Regressor with Randomized Search CV:**
  A decision tree is a non-parametric supervised learning algorithm that splits data into subsets based on input features, minimizing

target variable variance within each subset in regression tasks. Randomized search CV is a method for hyperparameter tuning, randomly sampling hyperparameter space and evaluating model performance using cross-validation, efficient for exploring large hyperparameter spaces. A Decision Tree Regressor with Randomized Search Cross-Validation (CV) is a machine learning model used for regression tasks that combines the Decision Tree algorithm with a randomized search approach to hyperparameter tuning.

We performed the decision tree regressor model on the training set and evaluated its performance on the test and validation set by using metrics like R^2, MSE, RMSE and MAE. It helps assess how well the model generalizes to unseen data and identifies potential overfitting or underfitting issues.

```
Best Hyperparameters: {'min_samples_split': 8, 'min_samples_leaf': 2, 'max_depth': 7}
Training set:
R2 Score_DT: 0.8974256331610259
MSE_DT: 0.10655902688738014
RMSE_DT: 0.32643380169244135
MAE_DT 0.251383497638322

Test set:
R2 Score_DT: 0.7846860172936589
MSE_DT: 0.19606350448763515
RMSE_DT: 0.4427905876231282
MAE_DT: 0.33129874271898

Validation set:
R2 Score_DT: 0.7835014061992409
MSE_DT: 0.2102517782864086
RMSE_DT: 0.45853219983596416
MAE_DT: 1.093485599148656
```

From the above table we can infer that the minimum number of samples required to split an internal node is found out to be 8, meaning that an internal node must have at least 8 samples to be considered for splitting and the minimum number of samples required to be at a leaf node is coming out to 2 indicating that each leaf node must have at least 2 samples. Also, the decision tree is allowed to have a maximum depth of 7 levels.

The model achieved its best performance on the training set, with an R-squared score of approximately 0.897, indicating that around 89.7% of the variance in the target variable is explained by the model. Despite the decrease in performance on the test and validation sets, the model still demonstrates reasonable predictive capability, with R-squared scores around 0.785 and consistent error metrics across both datasets. It may indicate towards potential

overfitting. This suggests that the model generalizes reasonably well to unseen data, although there is room for improvement, particularly in reducing the mean absolute error on the validation set. However, when tested on unseen data (test and validation sets), the model's performance slightly decreases, as reflected in the lower R-squared scores and higher error metrics (MSE, RMSE, MAE).

- **Random Forest:**
  Random Forest is an ensemble learning method based on decision trees. It builds multiple decision trees during training and outputs the mode (for classification) or the average prediction (for regression) of the individual trees. The randomness in Random Forest comes from two main sources: random sampling of data points and random selection of features.

```
Train set:
MSE_RF: 0.01849135021529866
MAE_RF: 0.10454123986035113
R2 Score_RF: 0.9822001139111696
RMSE_RF: 0.13598290412878622

Test set:
MSE_RF: 0.14561460740453194
MAE_RF: 0.2940574878532894
R2 Score_RF: 0.8400882349704848
RMSE_RF: 0.3815948209875652

Validation set:
MSE_RF: 0.14561460740453194
MAE_RF: 0.2940574878532894
R2 Score_RF: 0.8400882349704848
RMSE_RF: 0.3815948209875652
```

The model performed exceptionally well on the training set, with very low MSE, MAE, and RMSE values, and a high R-squared score of approximately 0.982. This indicating that the model explains about 98.2% of the variance in the target variable on the training data.
On the test and validation sets, the model's performance decreased slightly, with higher MSE, MAE, and RMSE values, and a lower R-squared score of approximately 0.840. This indicated that the model's performance is still strong, but it may not generalize as well to unseen data as it does to the training data. The decreased in performance on the test and validation sets suggested that the model may have memorized the training data's noise and specific

patterns, making it less effective at generalizing to new, unseen data. Hence, we can infer that there is a potential overfitting. The consistency in performance between the test and validation sets suggested that the model's performance is stable and not overly influenced by the specific data partitioning. However, the decrease in performance compared to the training set is typical and expected when evaluating on unseen data.

- **Epsilon SVR with Grid Search CV:**
  Support Vector Regression (SVR) is a supervised learning algorithm used for regression tasks. SVR works by mapping input features into a high-dimensional space and finding the optimal hyperplane that maximizes the margin while minimizing errors. Epsilon SVR is a variant of SVR that allows for a certain degree of deviation (epsilon) from the predicted value. Grid Search CV, short for Grid Search Cross-Validation, is a method for hyperparameter tuning that systematically searches through a grid of hyperparameters to find the optimal combination that yields the best performance.
  Epsilon Support Vector Regression (SVR) with Grid Search Cross-Validation (CV) is a machine learning model used for regression tasks, particularly when dealing with non-linear relationships between features and target variables.

```
Fitting 10 folds for each of 36 candidates, totalling 360 fits
Train set:
MSE_SVR: 0.08254369178951754
MAE_SVR: 0.20350306580341576
R2 Score_SVR: 0.9205429406669636
RMSE_SVR: 0.28730417990262086

Validation set:
MSE_SVR: 0.16031954214260724
MAE_SVR: 0.30004889696618453
R2 Score_SVR: 0.8349171849316067
RMSE_SVR: 0.4003992284490659

Test set:
MSE_SVR: 0.16598829415457983
MAE_SVR: 0.305052637658059
R2 Score_SVR: 0.8177141595502381
RMSE_SVR: 0.40741661006220625

Best Parameters for SVR: {'C': 10, 'epsilon': 0.1, 'gamma': 'auto', 'kernel': 'rbf'}
```

The model was trained and evaluated using a grid search cross-validation approach, testing 36 combinations of hyperparameters across 10 folds, resulting in a total of 360 fits. This exhaustive search over the hyperparameter space ensured that the optimal combination of hyperparameters is selected. The model performed well on the training set, explaining approximately 92.1% of the variance in the target variable, with relatively low errors. The model's performance slightly got decreased on the validation set compared to the training set, indicating potential overfitting. However, it still explained approximately 83.5% of the variance in the target variable. The model's performance further decreased on the test set, with an MSE of approximately 0.166, MAE of approximately 0.305, R-squared score of approximately 0.818, and RMSE of approximately 0.407. While the performance was slightly lower compared to the training and validation sets, the model still exhibits a reasonable ability to generalize to new, unseen data.

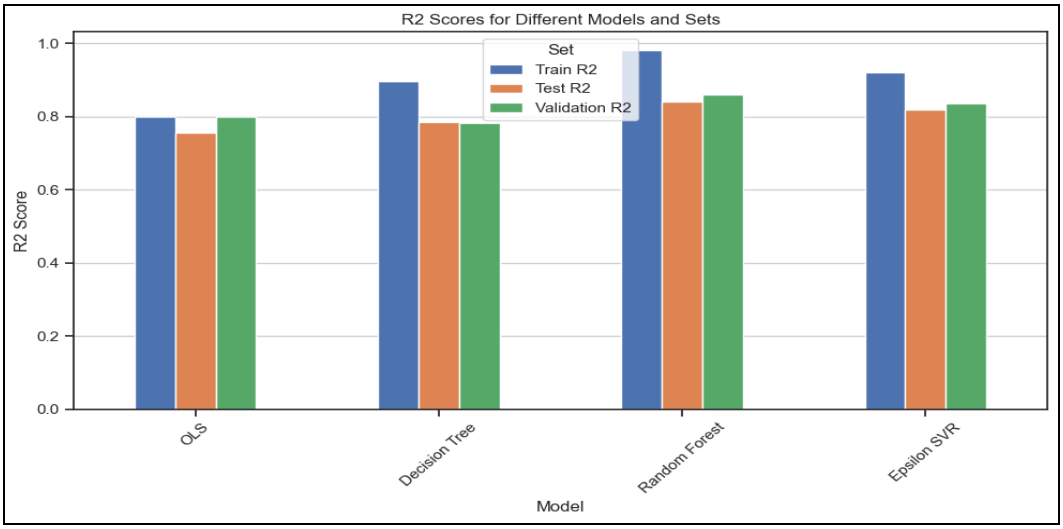
The best-performing SVR model was found with the specified hyperparameters: a regularization parameter (C) of 10, an epsilon value of 0.1, an auto-calculated gamma value, and an RBF kernel function. These hyperparameters were identified through the grid search process as yielding the best performance.

- **Model Comparison:**

  We have fitted various machine learning model on our dataset to predict “Life Ladder” score (happiness index). But we need to find and choose any one model among these which will give us predictions with greater accuracy and less errors. Model comparison typically involves evaluating different models based on their performance metrics, interpretability, computational efficiency, and suitability for a specific task. So, we tried to compare and choose the best model on the basis of performance metrics. Additionally, it's often helpful to experiment with multiple models and compare their performance empirically on a validation dataset.

  - **R^2 comparison:**

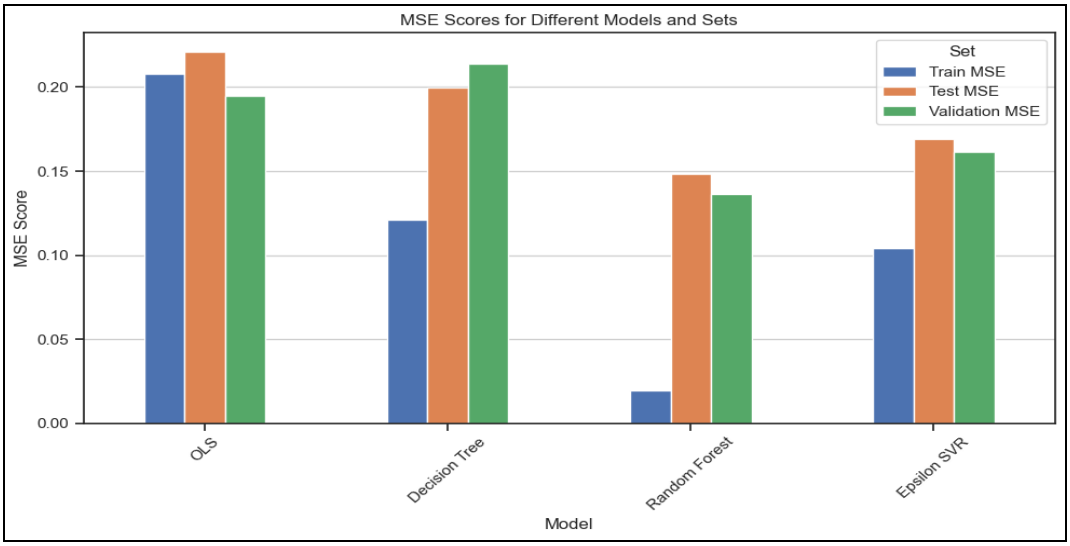
  Comparing models using the R-squared (R^2) value is a common approach in regression analysis. R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression model.



The comparison of regression models based on R-squared (R^2) values reveals distinctive performance patterns. The Ordinary Least Squares (OLS) model shows strong explanatory power but exhibits slight overfitting. The Decision Tree model excels in training but suffers from significant overfitting. In contrast, the Random Forest model demonstrates exceptional generalization, outperforming others across all datasets. The Epsilon Support Vector Regression (SVR) model performs well, with less overfitting compared to the Decision Tree. Overall, the Random Forest emerges as the top performer, offering robustness and accuracy for predictive modeling tasks.

- **MSE comparison:**

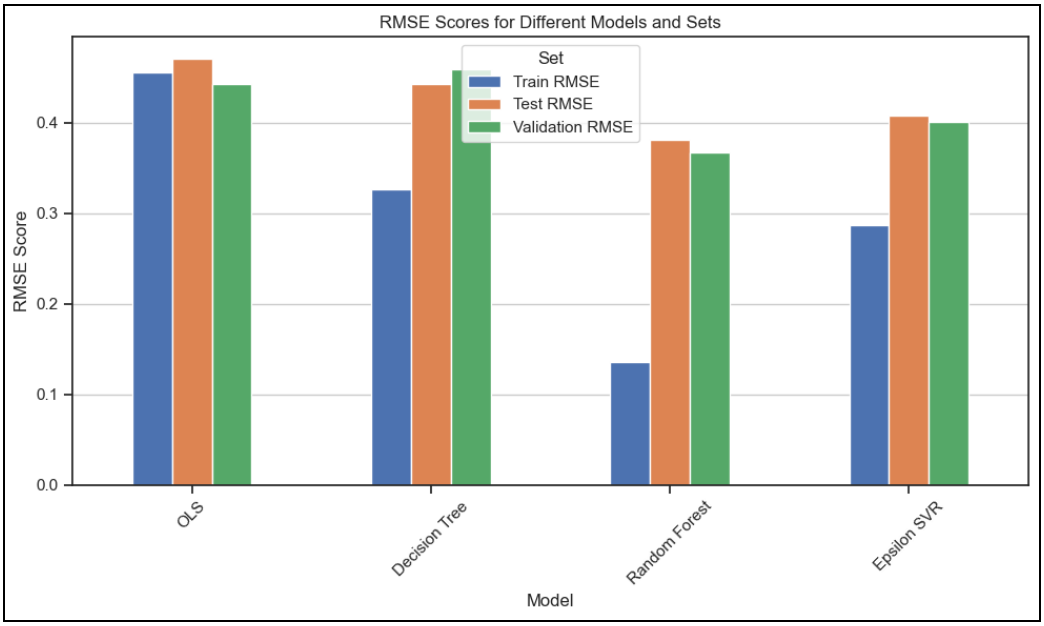
Mean Squared Error (MSE) provides insight into their predictive accuracy.



The Ordinary Least Squares (OLS) model shows moderate performance with slightly higher MSE on the test set, indicating minor overfitting. Conversely, the Decision Tree model exhibits severe overfitting, despite low training MSE, reflected in substantially higher MSE on test and validation sets. In contrast, the Random Forest model performs exceptionally well with consistently low MSE across all datasets, indicating both strong fit and generalization. Similarly, the Epsilon Support Vector Regression (SVR) model demonstrates solid performance with moderate MSE values overall. Overall, the Random Forest model stands out for its remarkable predictive accuracy and generalization capability, making it a preferred choice for regression tasks requiring high performance.

- **RMSE comparison:**

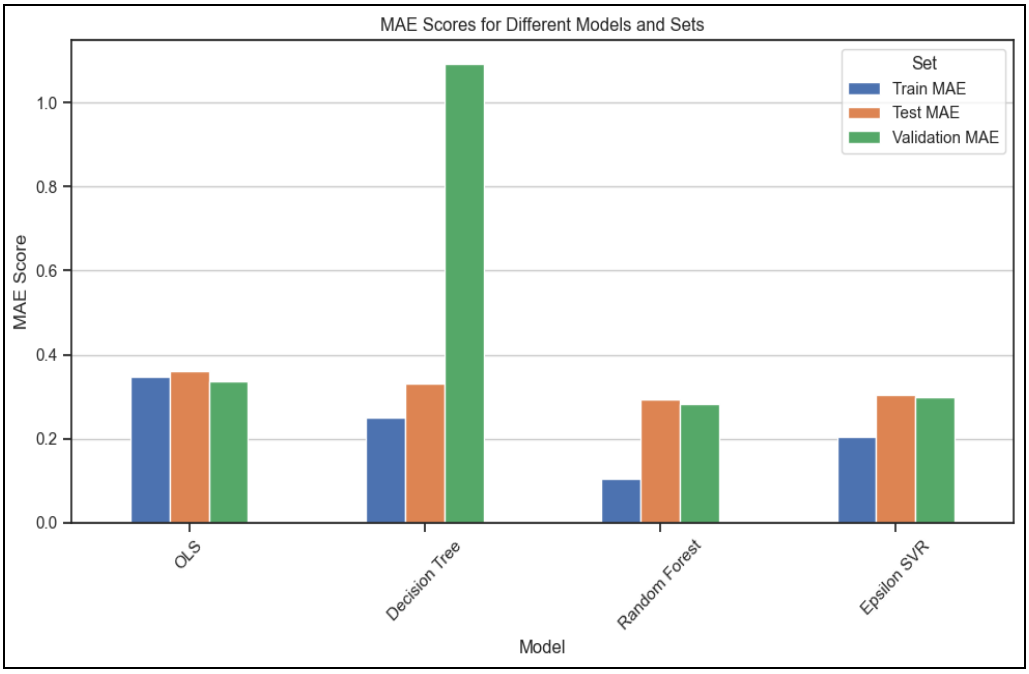
Comparing models using Root Mean Squared Error (RMSE) provides insight into their predictive accuracy, with values closer to zero indicating better performance.



OLS shows moderate RMSE across all datasets. Decision Tree displays low training RMSE but higher values on test and validation sets, indicating overfitting. Random Forest achieves consistently low RMSE, demonstrating superior predictive accuracy and generalization. Epsilon SVR strikes a balance with moderate RMSE across all datasets. Overall, Random Forest stands out with the lowest RMSE, emphasizing its superior performance.

- **MAE comparison:**

Comparing models using Mean Absolute Error (MAE) provides insight into their predictive accuracy, with lower values indicating better performance.



OLS displays moderate MAE across all datasets. Decision Tree shows low training MAE but significantly higher values on test and validation sets, indicating severe overfitting. Random Forest achieves consistently low MAE, signalling superior predictive accuracy and generalization. Epsilon SVR demonstrates moderate MAE, striking a balance between fit and generalization. Overall, Random Forest excels with the lowest MAE values, highlighting its superior performance.

Conclusion

The comprehensive analysis of various regression models sheds light on their performance and suitability for predicting life expectancy based on a myriad of factors. Initially, the Ordinary Least Squares (OLS) regression model, including all variables, demonstrated a promising R^2 value of 0.8 and minimal multicollinearity with no autocorrelation indicating a robust model. However, efforts were made to enhance the model by excluding statistically insignificant variables, resulting in a simpler yet equally effective model. Despite meeting few OLS assumptions, the model's predictive accuracy was moderate, with predicted values showing some deviation from actual values, particularly on the test set.

Further analysis involved k-fold cross-validation, where the linear regression model demonstrated reasonable performance, with variations across folds but maintaining overall accuracy. Employing k-fold cross-validation revealed the linear regression model's reasonable predictive capability, with mean R^2 scores around 0.765 and the best fold achieving an R^2 of 0.869. Despite minor variations across folds, the model demonstrated robustness and reliability in making accurate predictions, albeit with slight decreases in performance on unseen data.

The decision tree model exhibited strong performance on the training set, but its susceptibility to overfitting was evident, as indicated by lower performance on test and validation sets. In contrast, the random forest model showcased exceptional performance on the training set, explaining 98.2% of the variance. While performance slightly decreased on test and validation sets, the model demonstrated robust generalization capabilities. The comparison highlighted random forest as the preferred model due to its superior predictive accuracy and generalization.

The integration of Epsilon Support Vector Regression (SVR) introduced a nuanced approach to capturing nonlinear relationships between features and target variables. While demonstrating solid performance, it grappled with slight overfitting, albeit still offering respectable predictive capabilities. The exhaustive search for optimal hyperparameters underscored the meticulous optimization process essential for achieving the best model performance.

Comparative analysis highlighted the superiority of Random Forest in mitigating overfitting while maintaining superior predictive accuracy and generalization. Its dominance was evident in the consistently low MSE, RMSE, and MAE values, affirming its status as the model of choice for accurate life expectancy predictions. While OLS showed strong explanatory power, it exhibited slight overfitting. Decision trees suffered from severe overfitting, whereas random forest offered exceptional generalization and accuracy. Epsilon SVR struck a balance between fit and generalization. Overall, random forest emerged as the top performer across all metrics, showcasing its superiority for regression tasks.

In summary, the rigorous evaluation of regression models illuminated Random Forest as the pinnacle of predictive modelling for life expectancy. Its unparalleled performance, coupled with robust generalization capabilities, underscores its pivotal role in informing healthcare and policy decisions aimed at improving global life expectancy. Additionally, considerations of simplicity, interpretability, and generalization are crucial for selecting the most suitable regression model for predictive modelling tasks.

References

https://worldhappiness.report/ed/2024/#appendices-and-data

https://happiness-report.s3.amazonaws.com/2024/Ch2+Appendix.pdf

[2] Dixit, Siddharth & Chaudhary, Meghna &Sahni, Niteesh. (2020). Network Learning Approaches to study World Happiness.

[3] Prashanthi, B., and R. Ponnusamy. "Future Prediction of World Countries Emotions Status to Understand Economic Status using Happiness Index and SVM Kernel." Future 6.11 (2019).

[4] Moore, Lisa. "Exploring trends and factors in

the world happiness report." (2020).