Project Report: Exploring the Relationship Between Alcohol Consumption and Life Expectancy

1. Objectives

The objective of this project is to explore and quantify the relationship between Life Expectancy and Alcohol Consumption. The research questions addressed are:

- Primary Research Question: Is there a statistically significant relationship between alcohol consumption and life expectancy?
- Secondary Research Questions:
  • Can a linear regression model effectively predict life expectancy based on alcohol consumption?
  • Does a non-linear regression model offer a better fit than a linear model?

Motivation: Understanding the impact of lifestyle factors such as alcohol consumption on life expectancy is critical for informing public health policies and personal health decisions.

2. Method

Data Source

- Dataset: Life Expectancy Data.csv (sourced from Kaggle)
- Variables Considered: Life.expectancy (dependent variable) and Alcohol (independent variable)

Data Preprocessing

- Removed records with missing values.
- Selected only the relevant variables for analysis.
- Randomly split the dataset into 80% training and 20% testing subsets.

Analytical Techniques

- Correlation Analysis: Pearson correlation coefficient manually and using ggcorr().
- Regression Models:
  • Simple Linear Regression (manual and lm() function).

• Non-linear Regression (custom exponential function and polynomial regression).
- Model Evaluation Metrics: R-squared (R^2) and RMSE.

Correlational Analysis:

The correlation coefficient (ρ) is a measure that determines the degree to which the movement of two different variables is associated. The most common correlation coefficient, generated by the Pearson product-moment correlation, is used to measure the linear relationship between two variables. Correlation coefficients indicate the strength of the linear relationship between two different variables, x, and y.

A linear correlation coefficient greater than zero indicates a positive relationship. The value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y.

Linear Regression analysis:

Linear regression is a basic and commonly used type of predictive analysis. It is a machine learning algorithm based on supervised learning. It performs the task to predict a dependent variable value (Y) based on a given independent variable (x). This regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

In simple linear regression, we have only two variables:

- Dependent variable (also called Response), usually denoted as Y.
- Independent variable (alternatively called Regressor), usually denoted as x.

The linear relationship between the dependent variable(Y) and the independent variable(x) is represented in the form of $Y=\alpha+\beta x$. The concept of regression analysis deals with finding the best relationship between Y and x.

Non-linear regression analysis:

When the regression equation is in terms of r-degree, $r>1$, then it is called a nonlinear regression model. When more than one independent variable is there, then it is called the Multiple Non-linear Regression model. It is also alternatively termed as a polynomial regression model. In general, it takes the form as shown

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r + \epsilon$$

The polynomial model can be transformed into a general linear regression model setting $x_1 = x, x_2 = x^2$, …, $x_n = x^r$. Thus, the equation assumes the form:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x^r + \epsilon_i$$

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r + e_i$$

This model then can be solved using the procedure followed for multiple linear regression model.

The goal of the model is to make the **Sum of the Squares (SSE)** as small as possible. The sum of squares is a measure that tracks how far the Y observations vary from the nonlinear (curved) function that is used to predict Y.

Coefficient of Determination (R-Square)

Calculation of R2 for both linear and nonlinear regression quantity R2 is called the coefficient of determination which is used to measure the proportion of the variability of the fitted model.

let us define the total corrected sum of squares, defined as

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- SST represents the variation in the response values. The R^2 is

$$R^2 = 1 - \frac{SSE}{SST}$$

Note:

- If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)

Assumptions Verified

- Linearity: Inspected through scatterplots.
- Independence: Assumed due to random sampling.
- Homoscedasticity and Normality: Checked through residual visualization.

3. Interpretation

Correlation Findings

The computed correlation coefficient between Alcohol and Life Expectancy is approximately 0.402.
Interpretation: A moderate positive correlation exists, indicating that higher alcohol consumption is moderately associated with increased life expectancy.

Simple Linear Regression Results

R^2 Value: Approximately 0.1668.
Interpretation: Approximately 16.68% of the variance in life expectancy is explained by alcohol consumption.

Non-linear Regression Results

R^2 Value: Approximately 0.1668 (similar to linear model).
Interpretation: Non-linear modeling does not significantly improve predictive performance.

Overall Insights

Alcohol consumption shows a weak yet positive association with life expectancy. Neither model (linear or non-linear) offers strong predictive capability, suggesting additional variables are necessary for robust prediction.

4. Figures and Tables

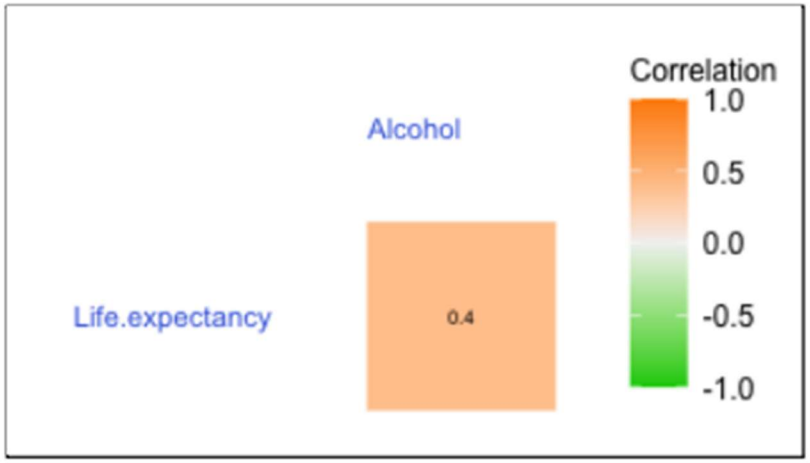
Figures and tables should visually represent correlation heatmaps, regression scatterplots, non-linear fit plots, and model evaluation summaries.

Analysis Step	**Visual**	**Description**
Correlation Analysis	Correlation Heatmap	Visualizes positive association between Alcohol and Life Expectancy.
Linear Regression	Scatterplot with Line of Best Fit	Depicts the linear relationship observed in training data.
Non-linear Regression	Non-linear Fit Plot	Displays the curve fit applied to the training data.
Model Evaluation	Metric Summary Table	Summarizes R^2 and RMSE values for comparison.

Model Performance Metrics:

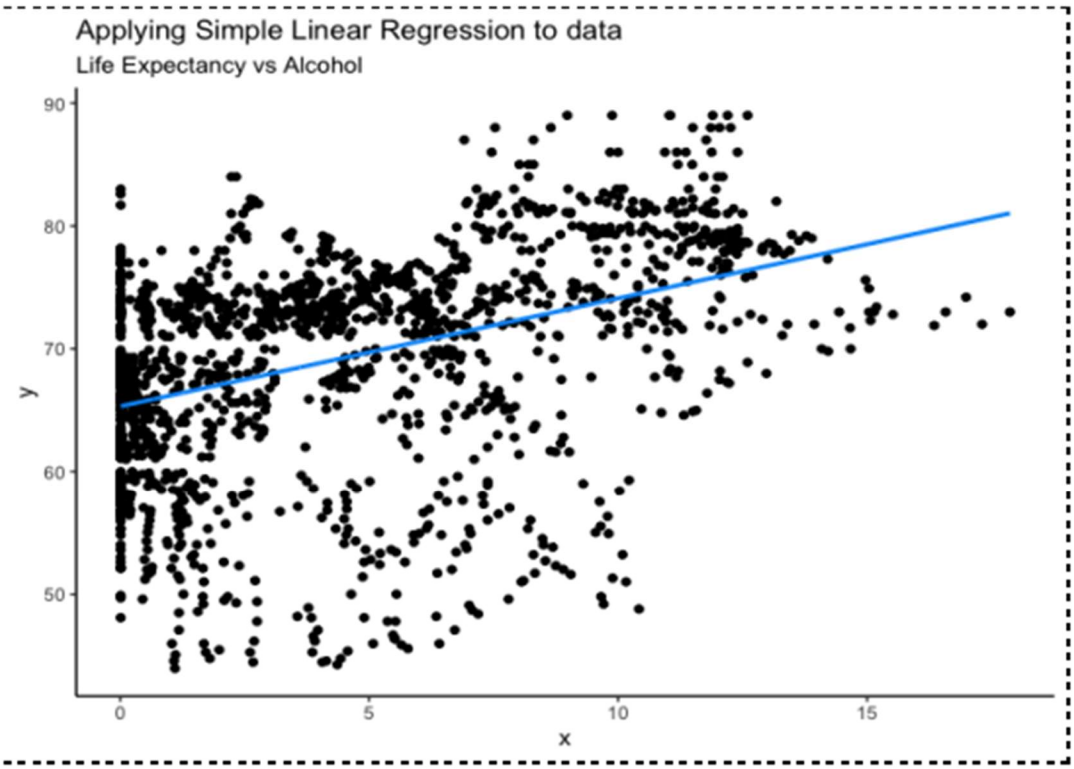
Model Type	R^2 (Manual Calculation)	R^2 (Library Calculation)	RMSE
Simple Linear Regression	0.166774	0.1668495	Computed from test set
Non-linear Regression	0.166551	0.1668495	Computed from test set

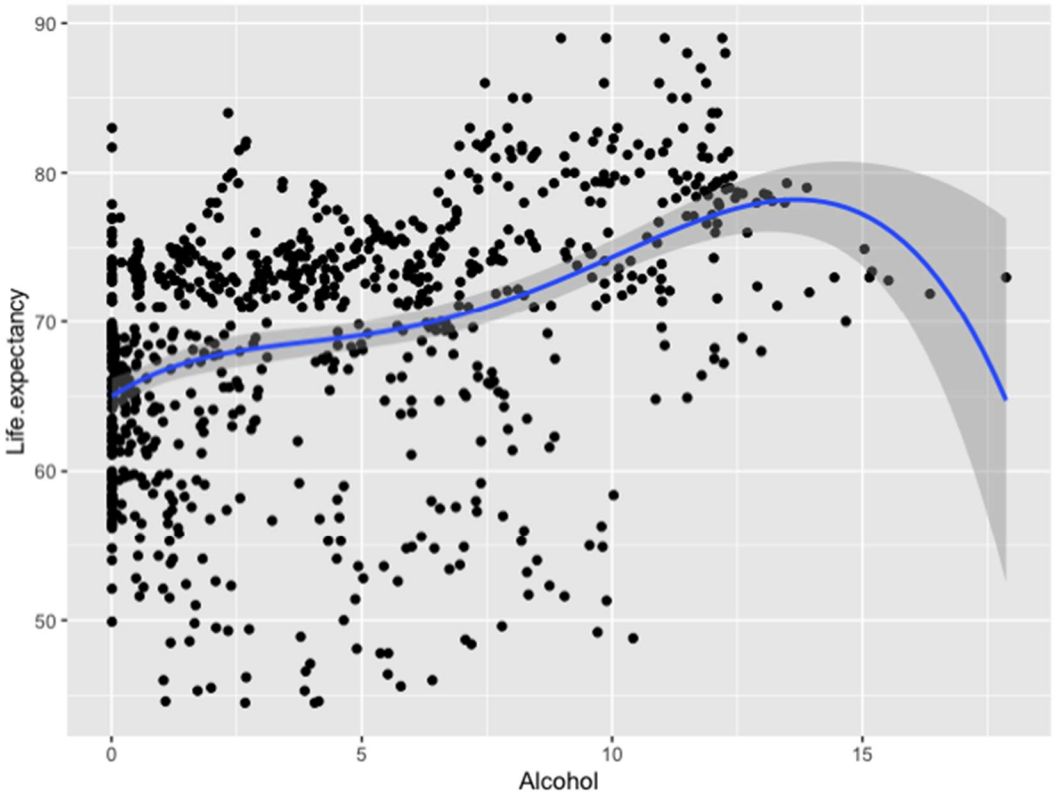
Correlation between Life Expectancy and Alcohol

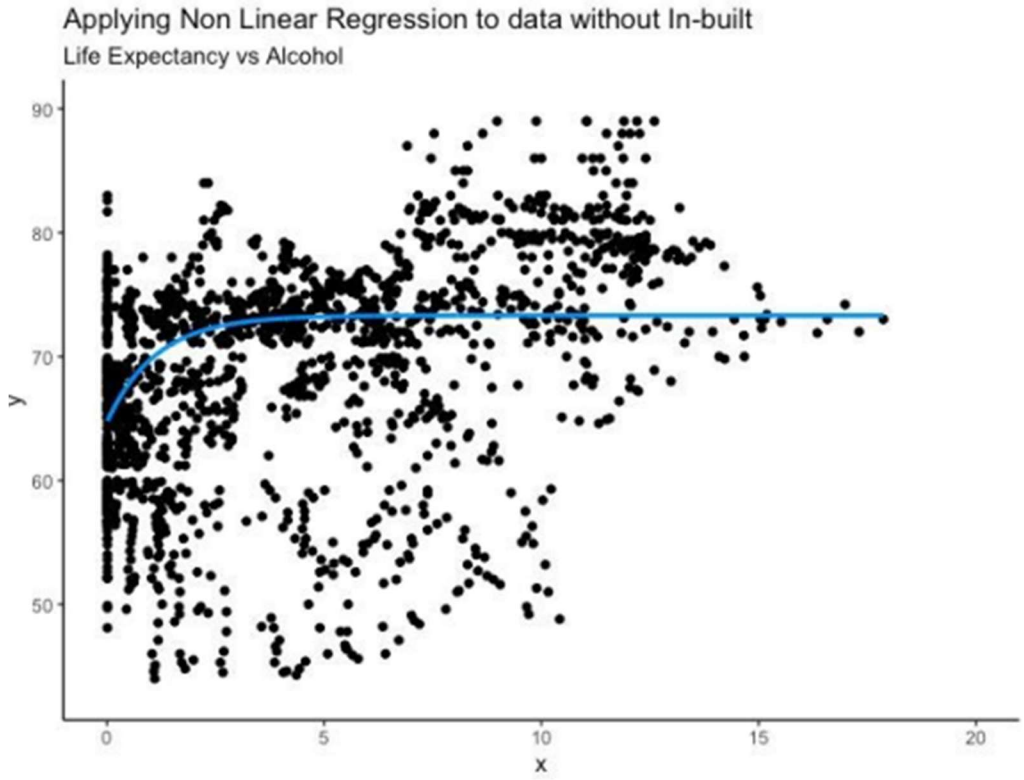


Hence, Life Expectancy has a positive relationship with drinking alcohol.

Linear Regression Output using Formula




90
80
70
60
50
Life.expectancy
0
5
10
15
Alcohol


Applying Non Linear Regression to data without In-built
Life Expectancy vs Alcohol
90
80
70
60
50
y
0
5
10
15
20
x

Final Conclusion

This study concludes that alcohol consumption has a moderate but weak positive relationship with life expectancy. Both simple linear and non-linear regression models exhibited low R^2 values, indicating limited predictive power.

The final results of the sub-tasks are as given below.

(i) Life Expectancy has a positive relationship to Drinking alcohol.
(ii) We have obtained the best fits for Linear Regression Analysis and Non-Linear Regression Analysis, and they were clearly explained in the Experimental Results section.
(iii) The coefficient of determination (R2) values is the best quality of fit in our models.

Recommendations:
- Future research should incorporate multiple health and socio-economic variables.
- Advanced modeling techniques could improve prediction accuracy.