Data Analysis And Predictive Modelling For Llyod World Risk Poll

Project Report-1
(Group 1)

MBA652 – Statistical Modelling for Business Analytics

Submitted By:
Abhinav Pratap Singh(231140001)
Aditya Kumar Singh(231130002)
Aman Ingle(231140003)
Samyak Srivatsa(231140020)
Yash Mishra(231140026)

Submitted To:
Prof. Devlina Chatterjee
IME Department
IIT Kanpur

Table of Contents

Introduction

The 2021 Lloyd's Register Foundation World Risk Poll represents the second wave of data collection aimed at understanding people's attitudes, experiences, and behaviors concerning various aspects of risk and safety. This global survey was conducted against the backdrop of the Covid-19 pandemic and sought to explore how risk-related experiences and perceptions have evolved since the 2019 World Risk Poll. The survey was embedded within the Gallup World Poll, a long-running initiative that regularly surveys residents in more than 140 countries, providing nationally representative samples of the world's population aged 15 years and older.

Objective

1. **Analyzing Global Trends Impacting Resilience Indices Across Diverse Nations:** This objective aimed to identify and understand trends in resilience indices across various countries of a region, shedding light on the factors contributing to resilience.
2. **Utilizing Linear Regression for Predictive Modeling:** Linear regression was employed to develop models that predict resilience indices based on the answers to the survey questions. This predictive modeling aimed to uncover relationships between independent variables and the Resilience Index.
3. **Comparative Analysis of Higher-Order Models Versus Linear Models:** To enhance the modeling process, higher-order regression models were explored and compared against the linear models. This analysis aimed to determine the model that best fits the dataset.

Methodology

1. **Data Acquisition**: Acquired the 2021 Lloyd's Register Foundation World Risk Poll dataset from 143 countries, divided into 15 regions.
2. **Feature Selection**: Applied systematic removal of columns with >50% null values, eliminated low-correlation columns, and reviewed for contextual relevance, ensuring a streamlined dataset.
3. **Assumption Validation**: Validated OLS regression assumptions, including homoscedasticity, normal residuals, independence, and no multicollinearity.

4. **Linear Regression**: Utilized linear regression to model the impact of 17 independent variables on the Resilience Index, minimizing errors with OLS.
5. **Polynomial Regression**: Explored polynomial regression, observing improved training data fit but higher test error beyond the second degree.
6. **Regularization**: Applied L1 regularization to minimize model complexity, penalize loss function, and reduce overfitting, followed by hyperparameter tuning.
7. **Outcome**: This methodology enabled analysis of global resilience trends, selection of suitable modeling techniques, and the enhancement of predictive accuracy, contributing to a deeper understanding of resilience factors worldwide

Data Description

Dataset Composition

- The dataset comprises data from 143 countries categorized into 15 regions based on geographical location.
- Over 70 questions were posed to more than 280,000 residents worldwide.
- We considered the data of Region 10 for the year 2021 for our analysis as we have the most contextual and behavioural understanding of the same.
- Independent Variables: These are the questions asked in the survey.
- Dependent Variable: Resilience Index

Steps in Regression Analysis:

1. Selection of Dependent and Independent Variables

- Dependent Variable: **Resilience Index**

  The Resilience Index is a measure of the capacity of individuals, communities, and systems to survive, adapt, and grow in the face of stress and shocks and even transform when conditions require it.

- Independent variables: There were more than 70 independent features (questions) available in the 2021 Lloyd's Register Foundation dataset. Out

of these 17 main features were selected using following feature selection methods:

- **Correlation**: Furthermore, columns exhibiting a correlation coefficient (r) with the dependent variable below 5% were thoughtfully eliminated.
- **Context**: In addition, a meticulous review of the dataset led to the removal of certain columns based on contextual understanding and relevance.
- **Redundancy**: Redundant features were also meticulously identified and subsequently excluded from the dataset.

Selected Independent Features:

Feature	Question asked
Urbanicity	Urban/rural -- recoded into 2 groups
Q4E	Worried Traffic or Roadside Accident Could Cause Serious Harm
Age	Age
Q1	Feel More, Less or About as Safe Compared With Five Years Ago
Q9	Artificial Intelligence Will Help or Harm People in Next 20 Years
Q4F	Worried Mental Health Issues Could Cause Serious Harm
EMP_2010	Employment Status
Q4C	Worried Violent Crime Could cause serious harm
Gender	Gender
HouseholdSize	Total number of people in household
Q10Q11Recode	how long would you be abl to cover your basic needs
ChildrenInHouseHold	Children under 15 in the household
Q6	Used Internet, Including Social Media, in Past 30 Days
CountryIncomeLevel2021	World Bank Country Income Classification corresponding to 2021 data
IncomeFeeling	Feelings About Household Income
Education	Education Level

2. Encoding:

In this project, a robust encoding methodology was employed to effectively handle categorical features present in the dataset. Categorical variables often pose a challenge in predictive modeling, and the choice of encoding technique plays a pivotal role in enhancing model performance and interpretability.

For nominal categorical features, the project implemented the One-Hot Encoding technique. This method transforms categorical variables into binary vectors, creating new binary columns for each category and assigning a '1' to the corresponding category while setting '0' for all others. One-Hot Encoding ensures that no ordinal relationship is assumed between categories, preserving the independence of each category.

On the other hand, ordinal categorical features were subjected to Ordinal Encoding. Ordinal Encoding assigns numerical values to categories based on their inherent order or ranking. This technique captures the ordinal relationships between categories, allowing the model to understand and leverage the inherent hierarchy within the data.

Feature	Correlation (r) with dependent variable	Type of Encoding
Urbanicity	0.019438	One Hot
Q4E	0.025087	Ordinal
Age	0.066773	Ordinal
Q1	0.086451	Ordinal
Q9	0.096583	One Hot
Q4F	0.102080	Ordinal
EMP_2010	0.136254	One Hot
Q4C	0.139095	Ordinal
Gender	0.143974	One Hot
HouseholdSize	0.221710	One Hot
Q10Q11Recode	0.253197	One Hot
ChildrenInHouseHold	0.271989	One Hot
Q6	0.331593	One Hot
CountryIncomeLevel2021	0.364206	Ordinal
IncomeFeeling	0.371784	Ordinal
Education	0.438818	Ordinal

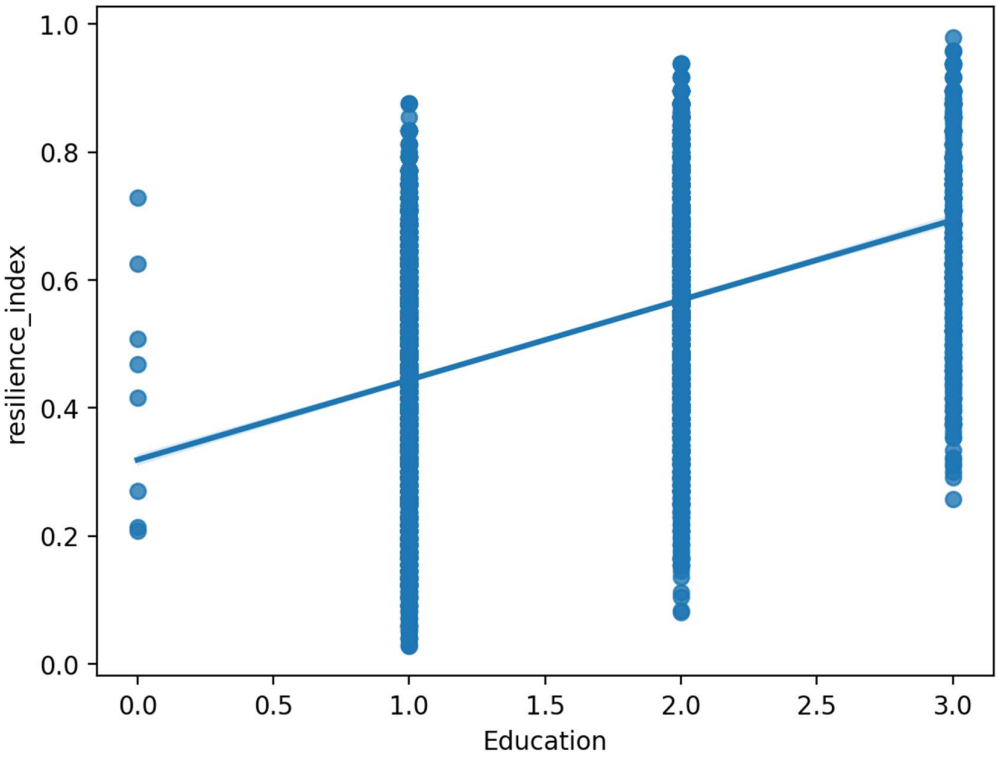
The careful selection and application of these encoding techniques contributed to the project's ability to handle diverse categorical data, enabling accurate and meaningful insights into the factors influencing resilience indices across nations and regions. This encoding methodology significantly enhanced the modeling process's effectiveness

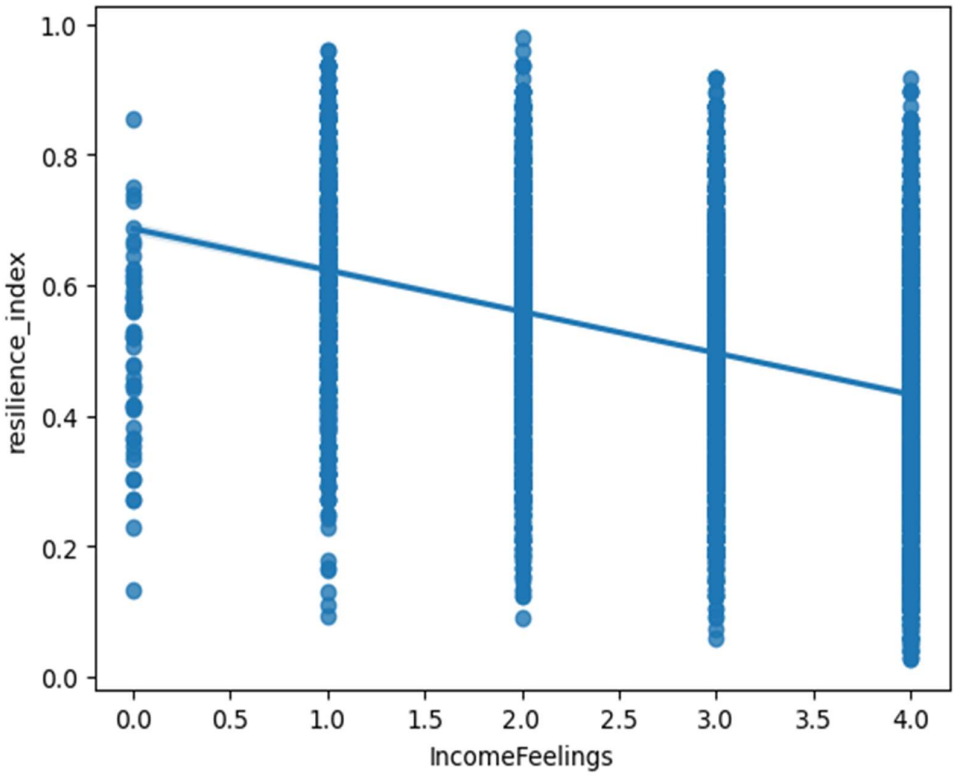
and interpretability, ultimately leading to more robust predictive models and a comprehensive analysis of global resilience trends.

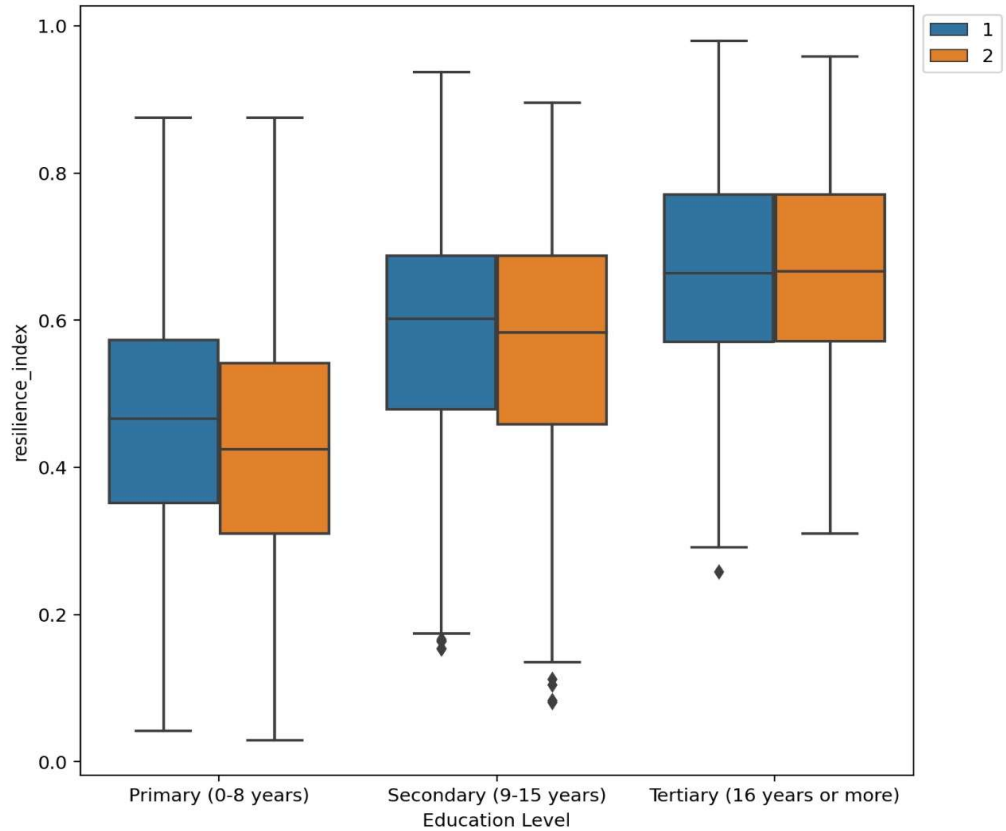
3. **EDA**:

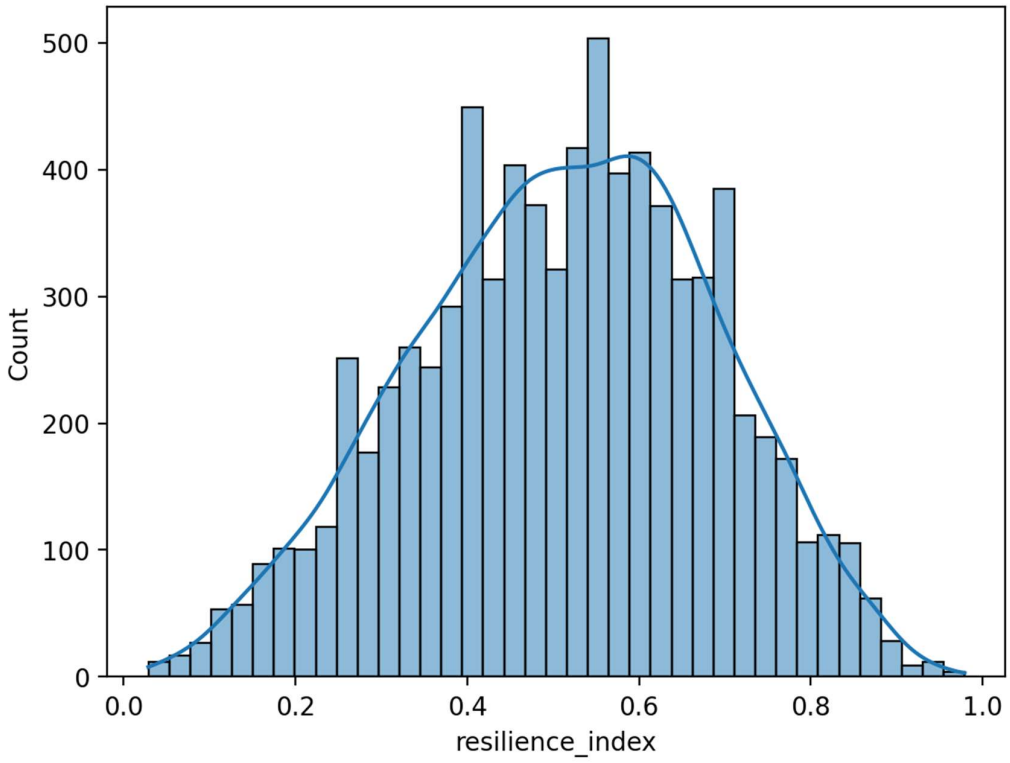
Exploratory Data Analysis played a pivotal role in this project by uncovering initial insights, identifying data patterns, and revealing potential outliers. EDA facilitated a deeper understanding of the dataset's characteristics, laying the foundation for subsequent data preprocessing and modeling steps.

Some important plots:



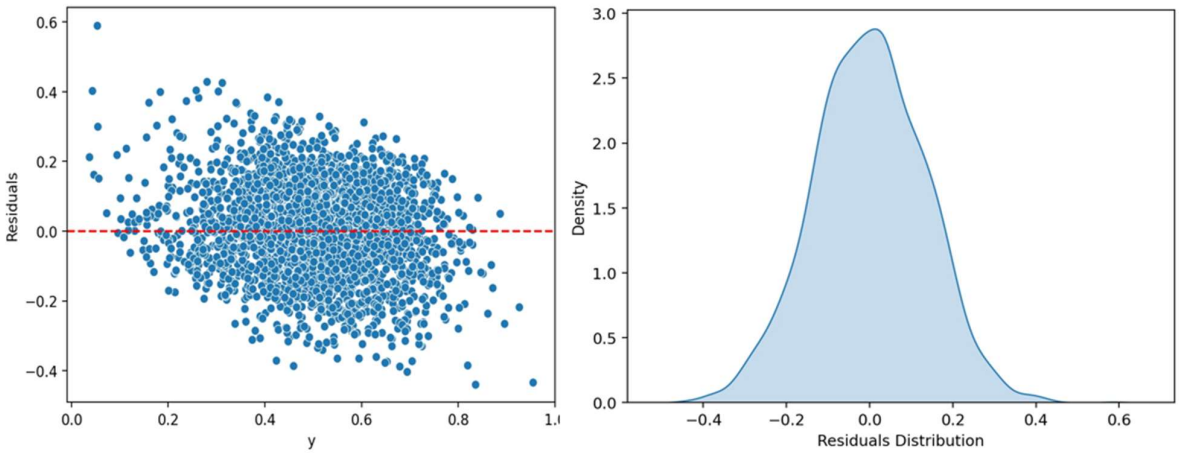

resilience_index
IncomeFeelings
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


resilience_index
Education Level
1
2
Primary (0-8 years)
Secondary (9-15 years)
Tertiary (16 years or more)
0.0
0.2
0.4
0.6
0.8
1.0



4. Least Square Assumptions:

1. The conditional distribution of u given X has mean zero, that is, $E(u|X = x) = 0$.



2. (X, Y), i=1,...,n, are i.i.d.

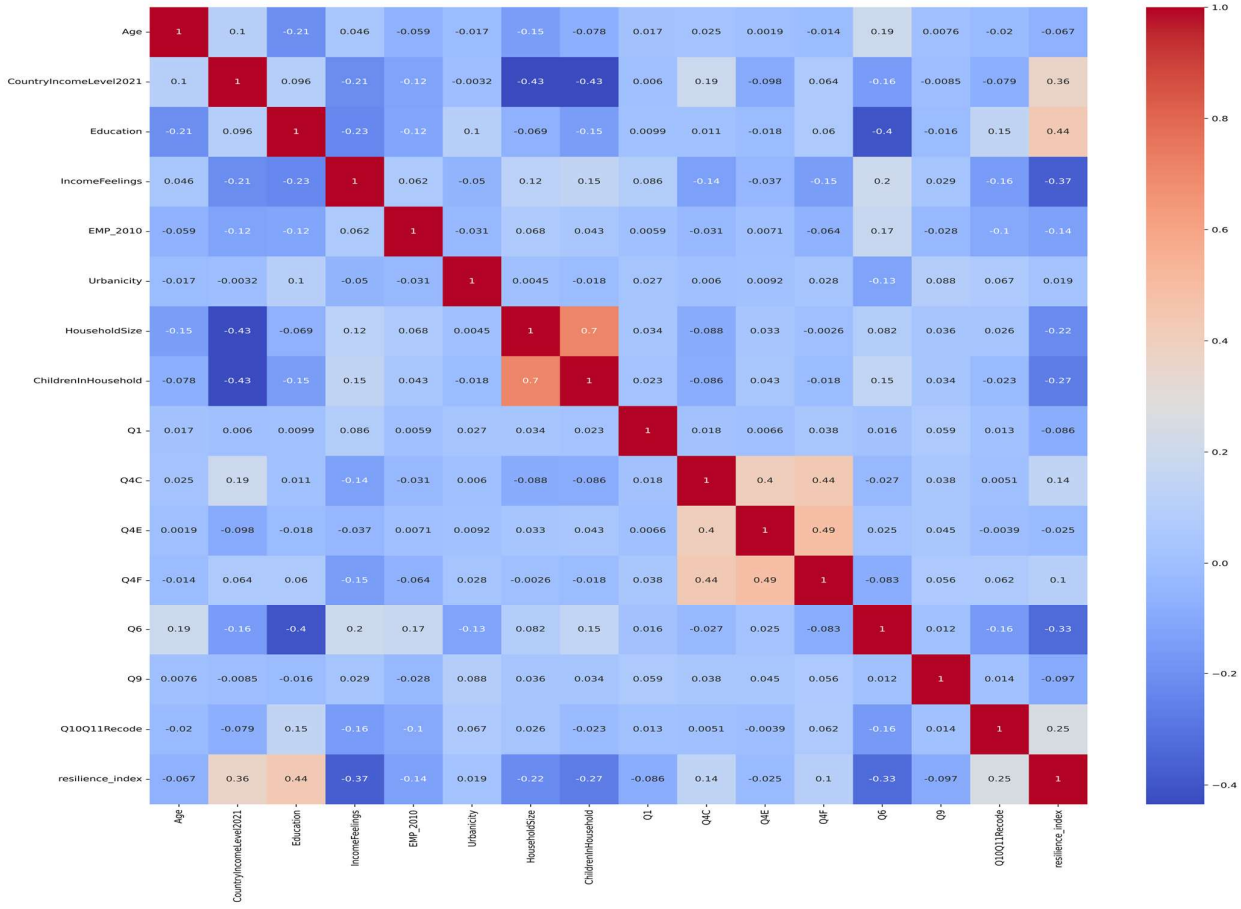
   The sampling was random and same types of questions were asked to the population sample. Hence, we can approximate the distribution as i.i.d.

3. Large outliers in X and/or Y are rare.

Most of the data was categorical and presence of outliers were negligible.

4. There is no perfect multicollinearity.

   This assumption was checked using the correlation matrix.



5. Predictive Modelling

- We have used Linear regression to model the effect of 17 independent variables on our dependent variable i.e. resilience index.
- For the loss/cost function we have used ordinary least squares.
- We have also explored the possibility of higher order regression model and compare it with the Linear model.

OLS Regression

OLS Regression Results

Dep. Variable:	resilience_index	R-squared:	0.471
Model:	OLS	Adj. R-squared:	0.468
Method:	Least Squares	F-statistic:	161.1
Date:	Thu, 28 Sep 2023	Prob (F-statistic):	0.00
Time:	04:35:44	Log-Likelihood:	5147.6
No. Observations:	8004	AIC:	-1.021e+04
Df Residuals:	7959	BIC:	-9891.
Df Model:	44		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1113	0.046	2.394	0.017	0.020	0.202
Country_Afghanistan	0.0374	0.047	0.799	0.424	-0.054	0.129
Country_Bangladesh	0.0857	0.006	13.345	0.000	0.073	0.098
Country_India	0.0170	0.005	3.378	0.001	0.007	0.027
Country_Nepal	0.0057	0.006	0.925	0.355	-0.006	0.018
Country_Pakistan	-0.0649	0.006	-10.165	0.000	-0.077	-0.052
Urbanicity_1	-0.0431	0.128	-0.337	0.736	-0.294	0.207
Urbanicity_2	-0.0453	0.128	-0.355	0.723	-0.296	0.205
Q9_0	0.0198	0.012	1.597	0.110	-0.005	0.044
Q9_1	0.0591	0.012	4.986	0.000	0.036	0.082
Q9_2	0.0304	0.012	2.573	0.010	0.007	0.054
Q9_3	0.0241	0.012	2.011	0.044	0.001	0.047
EMP_2010_1	0.0136	0.004	3.158	0.002	0.005	0.022
EMP_2010_2	0.0090	0.005	1.957	0.050	-1.59e-05	0.018
EMP_2010_3	0.0092	0.007	1.325	0.185	-0.004	0.023
EMP_2010_4	-0.0098	0.005	-1.883	0.060	-0.020	0.000
EMP_2010_5	0.0162	0.006	2.585	0.010	0.004	0.029
HouseholdSize_1	-0.0288	0.009	-3.099	0.002	-0.047	-0.011
HouseholdSize_2	-0.0198	0.008	-2.483	0.013	-0.035	-0.004
HouseholdSize_3	-0.0197	0.007	-2.814	0.005	-0.033	-0.006
HouseholdSize_3	-0.0197	0.007	-2.814	0.005	-0.033	-0.006
ChildrenInHousehold_1	-0.0225	0.052	-0.431	0.667	-0.125	0.080
ChildrenInHousehold_2	-0.0270	0.052	-0.516	0.606	-0.130	0.076
ChildrenInHousehold_3	-0.0335	0.052	-0.640	0.522	-0.136	0.069
ChildrenInHousehold_4	-0.0331	0.052	-0.632	0.527	-0.136	0.070
ChildrenInHousehold_5	-0.0403	0.053	-0.766	0.443	-0.143	0.063
ChildrenInHousehold_6	-0.0307	0.053	-0.583	0.560	-0.134	0.073
Q10Q11Recode_0	-0.0033	0.013	-0.246	0.806	-0.029	0.023
Q10Q11Recode_1	-0.0265	0.013	-2.106	0.035	-0.051	-0.002
Q10Q11Recode_2	-0.0184	0.013	-1.460	0.144	-0.043	0.006
Q10Q11Recode_3	0.0017	0.013	0.134	0.894	-0.024	0.027
Q10Q11Recode_4	-0.0185	0.018	-1.056	0.291	-0.053	0.016
Q10Q11Recode_5	0.0140	0.013	1.069	0.285	-0.012	0.040
Q10Q11Recode_6	0.0330	0.013	2.544	0.011	0.008	0.058
Q10Q11Recode_7	0.0704	0.013	5.267	0.000	0.044	0.097
Q10Q11Recode_8	0.0789	0.013	6.085	0.000	0.053	0.104
Q6_0	-0.0097	0.015	-0.640	0.522	-0.040	0.020
Q6_1	0.0401	0.004	10.955	0.000	0.033	0.047
Gender_1	0.0131	0.003	4.053	0.000	0.007	0.019
Age	-6.99e-05	0.000	-0.645	0.519	-0.000	0.000
CountryIncomeLevel2021	0.1852	0.046	3.984	0.000	0.094	0.276
Education	0.0759	0.003	27.791	0.000	0.071	0.081
IncomeFeelings	-0.0255	0.002	-16.639	0.000	-0.028	-0.022
Q1	-0.0105	0.002	-6.005	0.000	-0.014	-0.007
Q4C	0.0120	0.002	6.395	0.000	0.008	0.016
Q4E	-0.0020	0.002	-0.952	0.341	-0.006	0.002
Q4F	0.0075	0.002	3.703	0.000	0.004	0.011

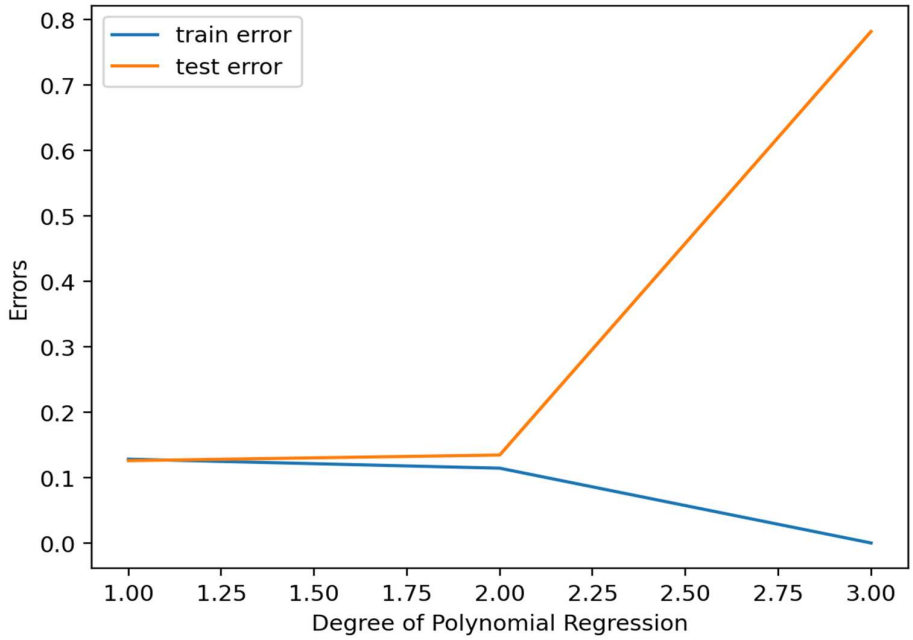
Omnibus:	55.076	Durbin-Watson:	1.647
Prob(Omnibus):	0.000	Jarque-Bera (JB):	38.083
Skew:	-0.016	Prob(JB):	5.38e-09
Kurtosis:	2.664	Cond. No.	8.05e+15

Polynomial Regression

```
In [59]: from sklearn.preprocessing import PolynomialFeatures

In [60]: poly_converter = PolynomialFeatures(degree = 2,include_bias = False)

In [61]: poly_features = poly_converter.fit_transform(X)

In [62]: X_train, X_test, y_train, y_test = train_test_split(poly_features, y, test_size=0.33, random_state=42)

In [63]: model.fit(X_train,y_train)
Out[63]: LinearRegression
         LinearRegression()

In [64]: y_pred = model.predict(X_test)

In [65]: r2_score(y_test,y_pred)
Out[65]: 0.40389010997443486

In [66]: np.sqrt(mean_squared_error(y_test,y_pred))
Out[66]: 0.13444995504029877
```

Graph between error and degrees of polynomial regression

- As we increase the degree of polynomial, we observe that the model gets better on training data but there is a sharp increase in test error beyond the second degree of polynomial.
- R2 decreased to 0.40 (for polynomial degree 2) from earlier value of 0.471.



.

6. Regularization

Regularization techniques were applied for model refinement. The reason for using regularization included minimizing model complexity, penalizing the loss function, and reducing model overfitting.
One important reason to use regularization is that it helps in feature selection by reducing the coefficients of some of the variables to zero.
After hyperparameter tuning, an improved R-squared score of 0.478 was achieved, up from the initial score of 0.471.

Conclusion

In conclusion, the analysis of the Lloyd's World Risk Poll dataset revealed the following insights:

- Polynomial regression did not provide a better fit for the data as increasing the degree of the polynomial led to higher test error.
- OLS regression outperformed polynomial regression, providing a better fit for modeling resilience indices.
- The application of L1 regularization enhanced model accuracy.

This project has significant implications for understanding resilience factors on a global scale and demonstrates the importance of selecting appropriate regression techniques to achieve accurate predictive modeling. The findings can inform decision-making in risk assessment, management, and policy development to promote resilience in diverse regions and communities.

References

1. https://towardsdatascience.com/an-ode-to-r-squared-804d8d0ed22c
2. https://www.researchgate.net/publication/242329609_Small_Is_Beautiful_The_Use_and_Interpretation_of_R2_in_Social_Research
3. Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics*.
4. https://wrp.lrfoundation.org.uk/about-the-lloyds-register-foundation-world-risk-poll/

APPENDIX:

Python Code :
https://drive.google.com/file/d/1O_jWydapegU6XcPe-j4KjS8MD28Hn9y4/view?usp=drive_link

Dataset :
https://docs.google.com/spreadsheets/d/1CiBbanjXIXyNt9hq5ZMyd5e-Ws1ORuVw/edit?usp=sharing&ouid=110105600549839635163&rtpof=true&sd=true