Linear Regression Final Project

Overview

This project centers on the application of linear regression, a powerful statistical modeling technique, to address a specific problem or research question. Linear regression allows us to examine the relationships between one or more independent variables and a dependent variable, offering valuable insights and predictive capabilities.

The key steps of this project encompass data collection, preprocessing, exploration, and model development. We will carefully curate and clean the dataset, select relevant features, and transform the data when necessary. Afterward, we will explore the dataset visually and statistically to gain a deeper understanding of its characteristics.

Our model selection will involve choosing the appropriate model for linear regression with multiple predictors. We will split the data into training and testing sets for model training and evaluation.

The project's success will be determined by the model's performance, assessed through a variety of metrics. Interpretation of the model's coefficients and visualization of its predictions will enable us to draw valuable insights from the analysis.

Ultimately, this project aims to deliver a comprehensive understanding of the relationships between variables, enabling us to make informed predictions and conclusions. The results will be communicated effectively, and the model may be deployed for practical applications if required.

In summary, this linear regression project will leverage statistical analysis to solve a specific problem, providing valuable insights and predictive capabilities that can be applied in various domains.

Description of Dataset

Source:
https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016/data

The above dataset contains data pulled from four other datasets linked by time and place and was built to find signals correlated to increased suicide rates among different cohorts globally across the socio-economic spectrum. The reference for these datasets can be found in the above link. The dataset contains 27820 rows and 12 columns. A detailed description of each variable can be found below:

Sr No.	Column	Non-Null Count	Datatype	Description
1	country	27820 non-null	Object	Country of incident
2	year	27820 non-null	Integer	Year of incident (1985-2016)
3	sex	27820 non-null	Object	Sex of individual (Male/Female)
4	age	27820 non-null	Object	Age category of individuals (6 categories)
5	suicides_no	27820 non-null	Integer	Number of suicides
6	population	27820 non-null	Integer	Population of the country for that year
7	suicides/100k pop	27820 non-null	Float	Number of suicides per 100,000 population
8	country-year	27820 non-null	Object	String of the combination of country and year columns
9	HDI for year	8364 non-null	Float	The Human Development Index for the year
10	gdp_for_year ($)	27820 non-null	String	Gross Domestic Product of the country for the year
11	gdp_per_capita ($)	27820 non-null	Integer	Per capita GDP for the year
12	generation	27820 non-null	Object	Generation of the individual (6 categories)

Statement of the Research Problem

The research problem at hand involves predicting the number of suicides per 100,000 population based on a dataset comprising various socio-economic and demographic variables. The dataset encompasses information from multiple countries, spanning different years, and includes factors such as gender, age group, economic indicators, and human development index (HDI). The primary research question is: Can we develop a predictive model that effectively estimates the number of suicides based on these diverse predictors, providing valuable insights into the determinants of suicide rates?

Summary of Methods

To address the research problem and answer the research question, the following methods and steps will be employed:

Data Cleaning: Handling missing values and outliers.

Feature Selection: Identifying relevant predictors.

Data Transformation: If required, scaling or normalizing variables.

*Data Visualization:*Creating visual representations of the data to better understand the relationships and distributions of variables. Exploring how variables like gender, age group, and economic factors relate to suicide rates.

Modelling: Choosing an appropriate regression model for predicting the number of suicides per 100k population. Given the nature of the problem, multiple linear regression may be a suitable choice.

Interpretation of Model Coefficients: Analyzing the coefficients of the regression equation to understand the impact of different predictors on suicide rates.

Evaluating the model: The model is evaluated using metrics such as R squared, Adjusted R squared, AIC and BIC and checked for problems such as Multicollinearity, Heteroscedasticty and Influential points. The various models will then be compared to select the best model.

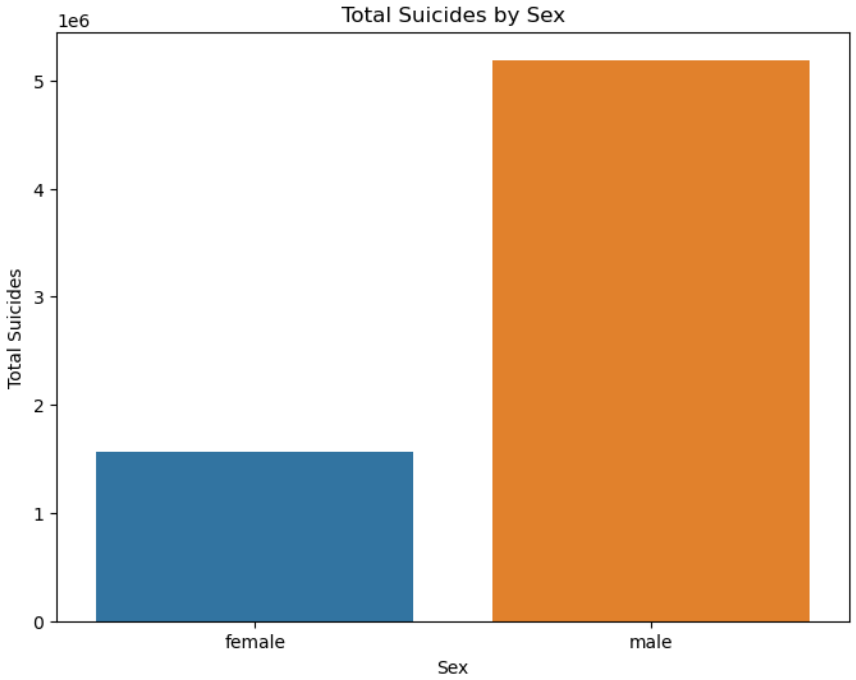
Communication of Findings: A summary of the results elucidating the significant inferences and how the model explains our research questions. We will also discuss the potential issues and future scope for this project.

Overview of the Data

Exploratory Data Analysis (EDA) is a fundamental phase in our research process. It serves as the cornerstone of our analysis, enabling us to delve into the dataset's intricacies, uncover patterns, and gain a deeper understanding of the variables at our disposal. In the context of our research problem, which aims to predict the number of suicides based on various socio-economic and demographic predictors, EDA plays a crucial role in revealing the insights necessary to inform our subsequent modeling efforts. We have already given an overview of what our columns look

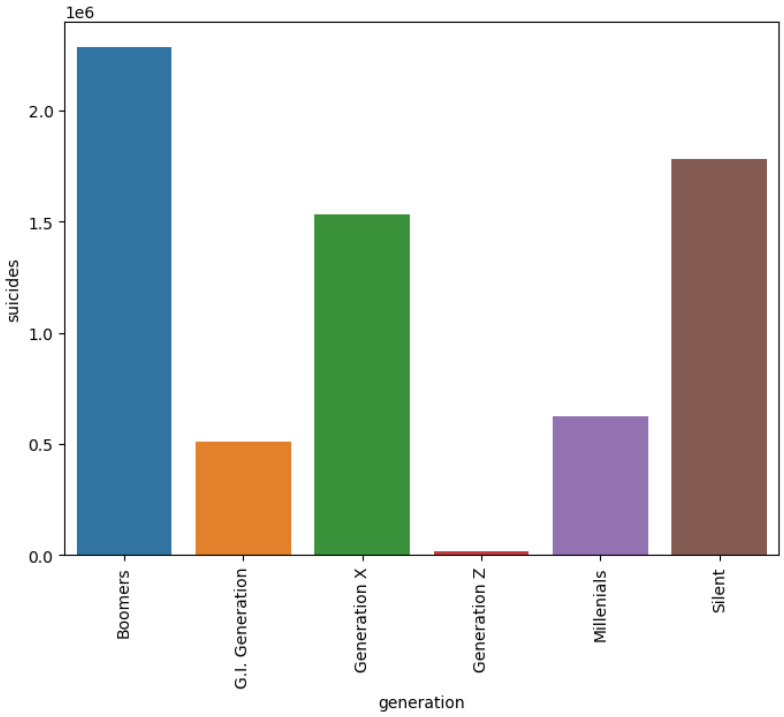
like and what the size of our dataset is; we will now dive deeper into how the various factors affect suicide.

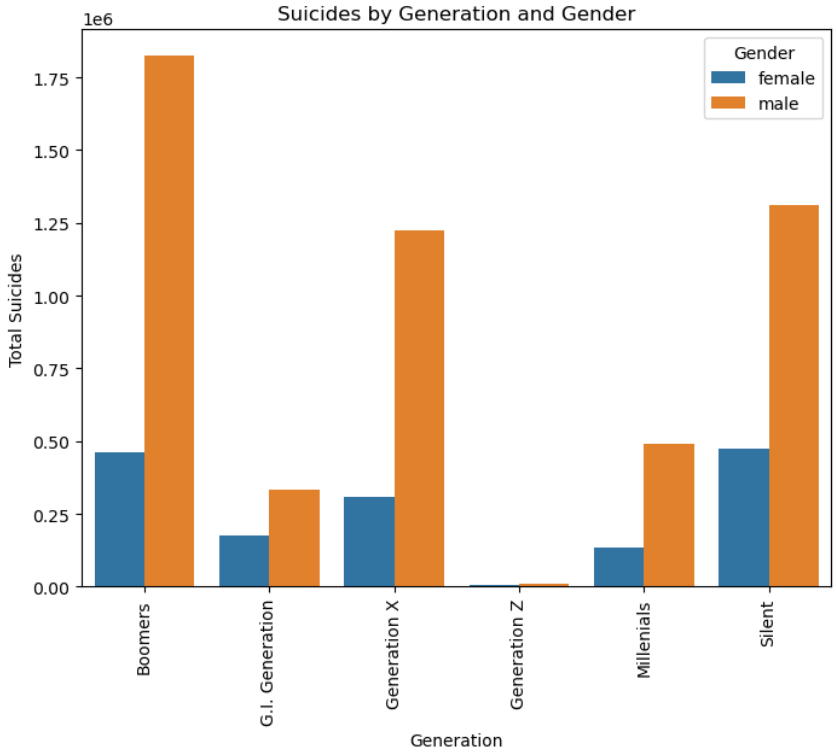
First, we wanted to examine the disparities of suicide numbers by sex.



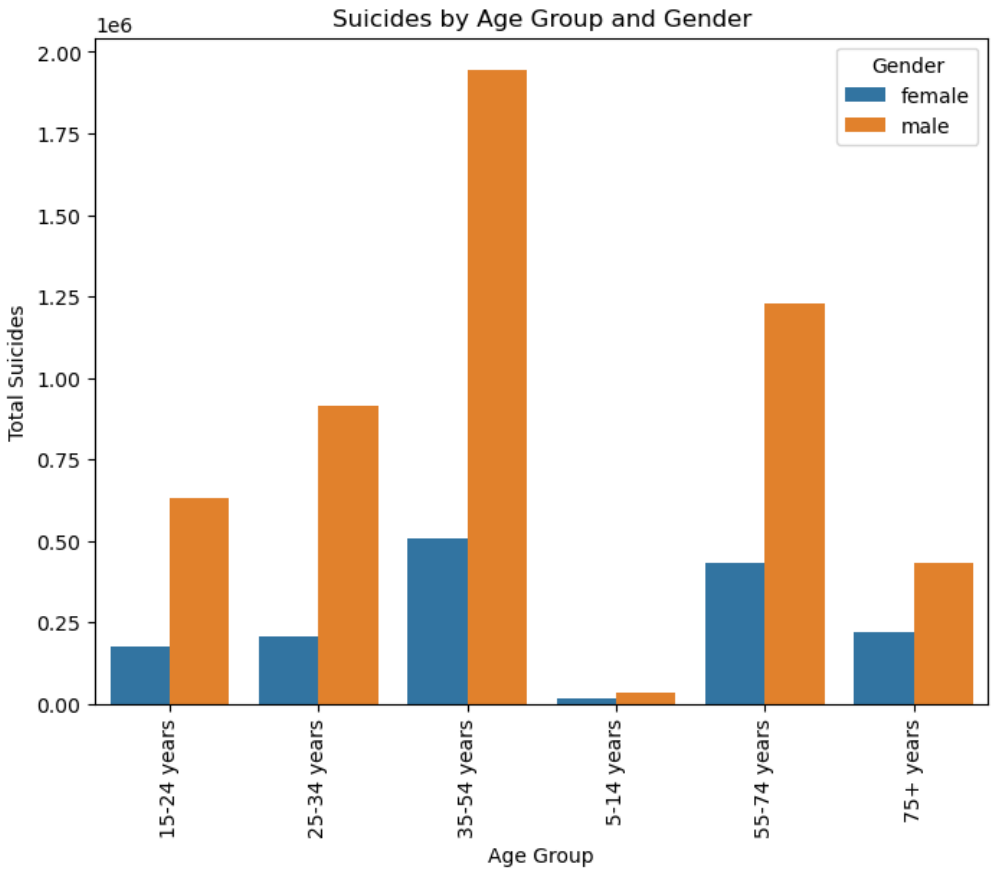
As seen in the plot above, males are disproportionately more likely to be the victims of suicide.

Next, we wanted to examine the trends across each generation.



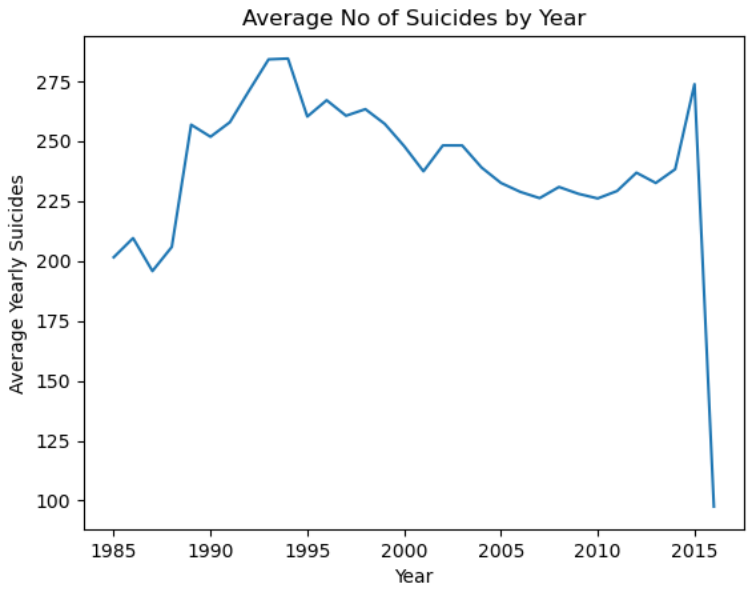


As seen above, the older generations have higher numbers of suicide. A notable exception to this is the GI generation which seems to have rather lower numbers. Another key detail in these plots is that the trend of males dying at higher rates by suicide is consistent across all the generations We then examined how age affected suicide numbers.



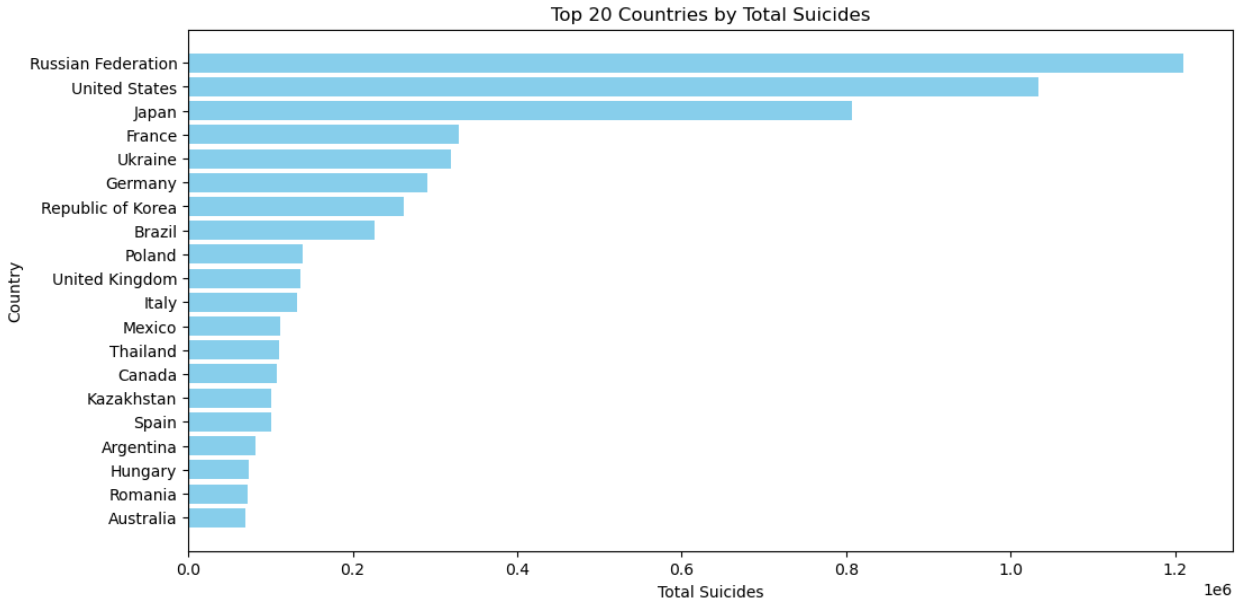
Middle aged people were at a higher risk of dying by suicide, and once again, males are more likely to be victims.

Next we have a look at the numbers across the years.

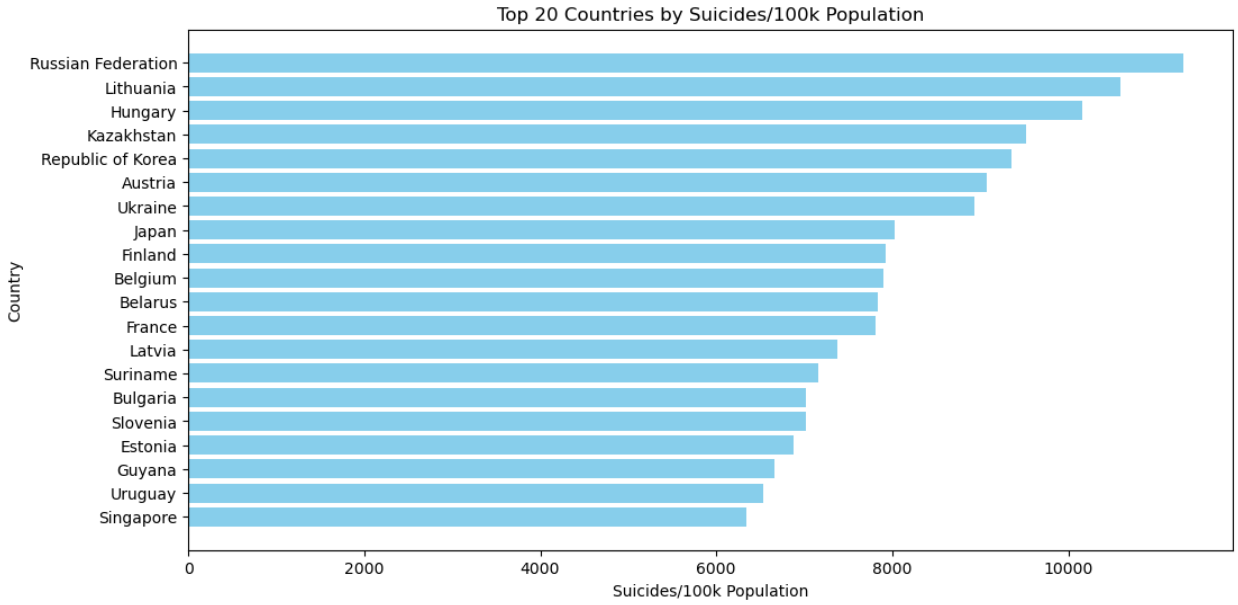


As seen in the plot, the numbers seem to be fairly consistent across time. There is a decrease around the late 2000s, which can be a little shocking considering there was the global housing crisis around that time. The drop after 2015 is due to the end of data collection.

Finally, we also wanted to understand what countries had higher suicide rates.



This plot shows a lot of countries that have high populations, so perhaps it isn't the best metric to judge. We will, therefore, explore rates of suicide per 100k population by country.



One interesting finding here is that a lot of the top 20 countries include nations that are very wealthy, indicating that perhaps poverty might not be the main reason for suicide and there might be other factors that influence this.

Regression Analysis:

In the regression model,

- we've fitted the variable `Suicides per 100k population` as the dependent variable, and
- [Sex, Age, population, gdp_per_capita, generation, HDI for year] as independent variables.
    - Here Sex, Age, Generation are Categorical variable and accordingly treated as C(column) in the initial model
    - Rest as numeric variables.

Initial Model:

- *Suicides_per_100k_pop ~ C(country)*
  + *Year*
  + *C(sex)*
  + *C(age_group)*
  + *population*
  + *gdp_year*
  + *gdp_per_capita*
  + *C(generation)*

The model fitted on the data is a Least squares method since the target column suicides_per_100k_population is a numeric variable.

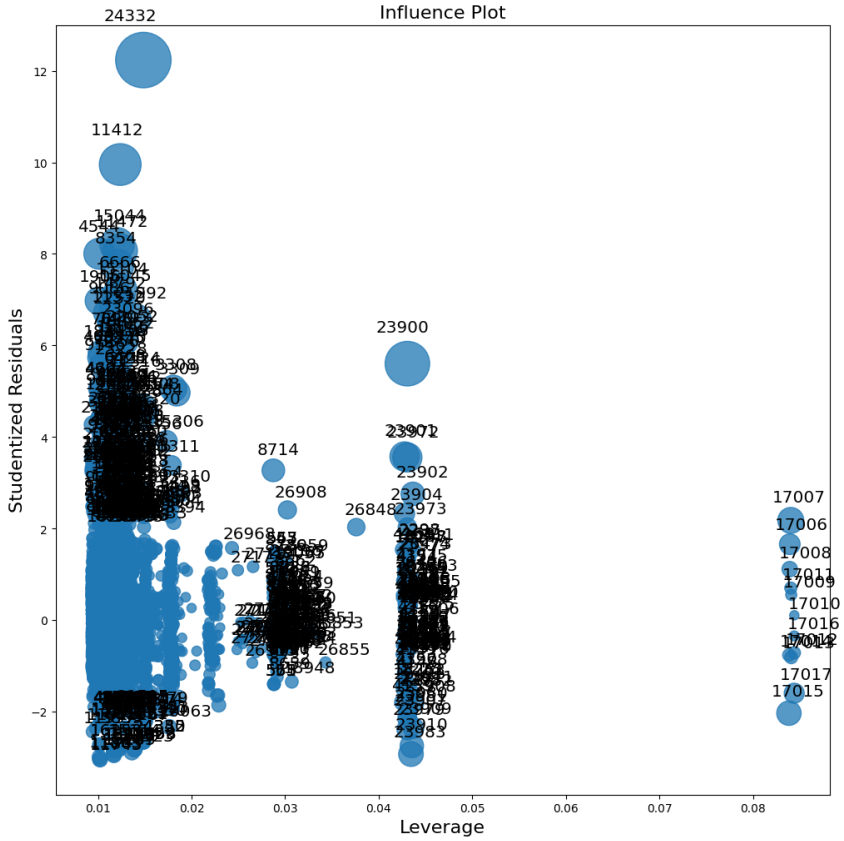
```
OLS Regression Results
==============================================================================
Dep. Variable:     suicides_per_100k_pop   R-squared:                       0.520
Model:                               OLS   Adj. R-squared:                  0.518
Method:                    Least Squares   F-statistic:                     288.5
Date:                   Thu, 12 Oct 2023   Prob (F-statistic):               0.00
Time:                           20:48:16   Log-Likelihood:            -1.1113e+05
No. Observations:                  27820   AIC:                         2.225e+05
Df Residuals:                      27715   BIC:                         2.233e+05
Df Model:                            104
Covariance Type:               nonrobust
```

In the initial Regression model, looking at the R-squared and Adj. R-squared metrics which show the goodness of fit:

- The observed R2 metrics suggest that the model was able to explain ~52 % of the variance in the target column: suicides per population of 100k.
- We observe that with R-square measured on a range of (0,1), there is scope for further improvement and we'll work on this by doing regression analysis.
-

Model Diagnosis

1. Influential points



The influence diagram provides valuable feedback about our regression analysis. On the vertical axis, studentized residuals point out discrepancies between predicted and actual values, with distinct outliers such as "24332" and "11412" indicating that the model might miss specific trends. Circle sizes, reflecting Cook's distance, shed light on the weight of each data point on the regression. Observations like "24332", "11412", and "23900" with larger circles suggest a potential skew in our model interpretation. While the dense grouping on the left signifies a good fit for many observations, the notable outliers with heightened leverage and residuals could be skewing key model metrics, risking misinterpretations.

2. Multicollinearity

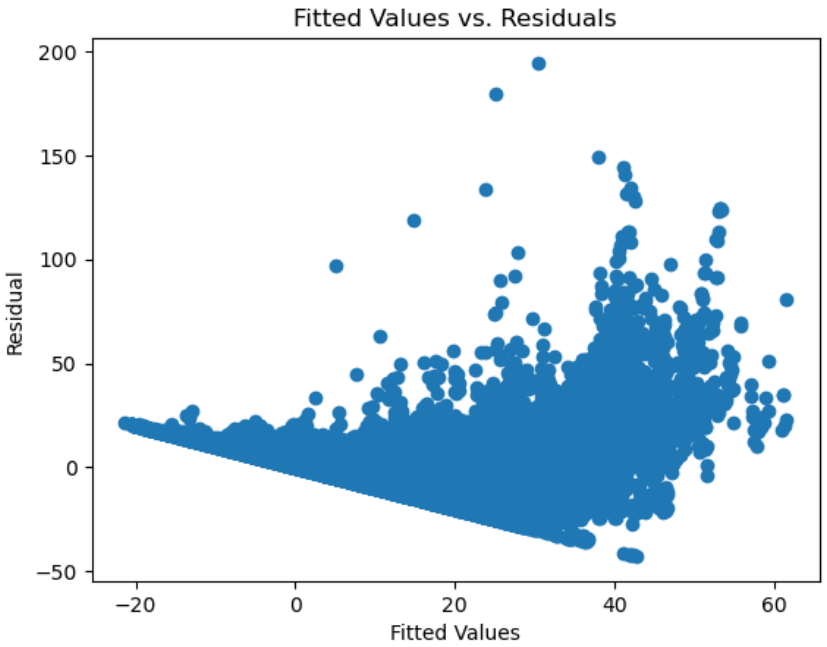
VIF Factor	Features
13.437677	C(age_group)[T.55-74 years]
18.157737	C(age_group)[T.75+ years]

These figures confirm the existence of multicollinearity. Nevertheless, a closer examination reveals that these variables are subsets of the broader "age_group" category. Their close

correlation is an expected outcome due to their categorical nature, reflecting distinct age segments. Based on this understanding, it was deemed essential to keep these variables intact. Excluding or merging them could deprive the model of critical age-specific insights. To counteract the effects of multicollinearity, future iterations of this model might explore methods such as principal component analysis or even regularization techniques.

3. Heteroscedasticity

3.1 Fitted values vs. Residuals

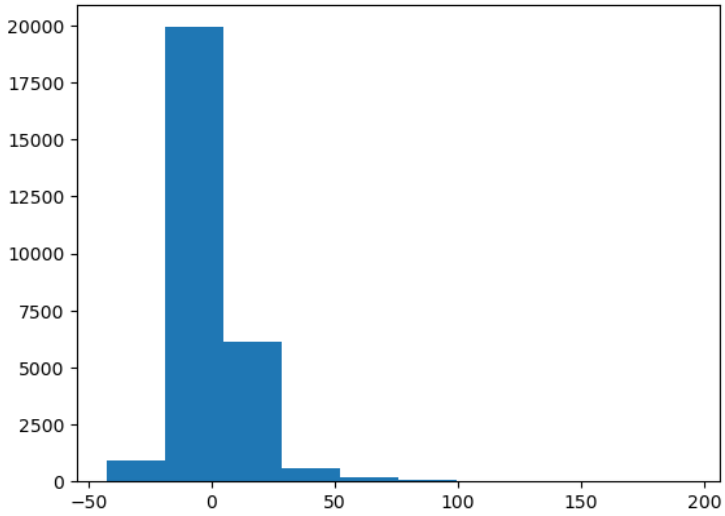


The scatter plot indicates potential heteroscedasticity within the model. There seems to be an inconsistency in the spread of residuals across different fitted values. Such variations in residuals can affect the accuracy of some standard statistical analyses. It would be advisable to validate these findings with a more formalized test like the Breusch-Pagan test. If heteroscedasticity is confirmed, exploring possible solutions would be essential.

3.2 Breusch-Pagan test

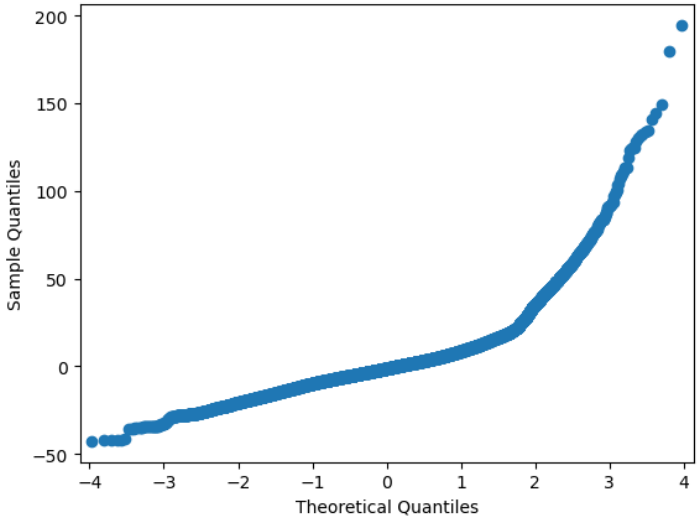
The results from the Breusch-Pagan test indicate an LM Statistic of 3665.41. With a p-value standing at 0.0, which falls below the standard significance level of 0.05, we can reject the null hypothesis that suggests homoscedasticity. This points towards the existence of heteroscedasticity within the model. It's crucial to explore solutions for this heteroscedasticity to guarantee the model's accuracy and dependability.

3.3 Histogram of Residuals



The residuals predominantly cluster around zero, and there's a swift reduction in their count as they move away from this central point. This histogram suggests a scenario consistent with homoscedasticity, where the variation in the residuals remains consistent across the spectrum of independent variables. While this graphical representation hints at homoscedasticity, it's still advisable to employ formal tests, such as the Breusch-Pagan, to ascertain the nature of variance definitively.

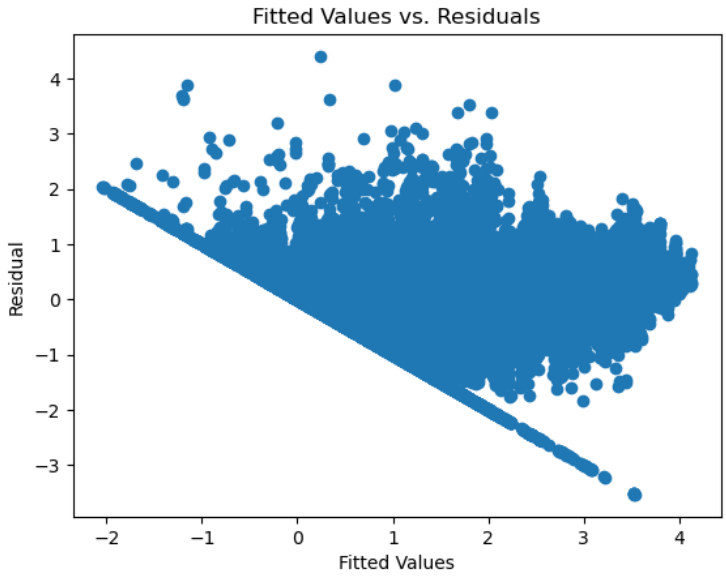
4. Normality



The QQ-Plot indicates that the residuals largely follow a normal distribution, yet the tail behavior raises questions about complete normality. To confirm the exact nature of these residuals, additional analyses or tests would be beneficial.

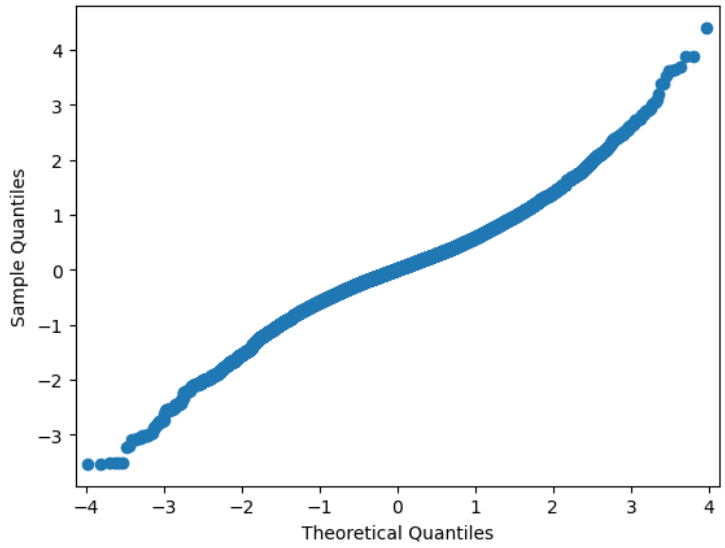
Model Diagnosis After Log transformation

1.Heteroscedasticity



After implementing the log transformation, the graph depicting residuals versus fitted values displays a marked reduction in signs of heteroscedasticity. There's still a slight variation in constant variance, but it's less evident than previously. This suggests that the transformation has alleviated some aspects of the heteroscedasticity challenge. Nonetheless, a few inconsistent patterns are still visible, implying that while improvements have been made, the issue hasn't been fully rectified. Additional model tweaks or strategies might be worth exploring to further tackle this concern.

2. Normality



Following the logarithmic adjustment, the QQ-Plot demonstrates a marked enhancement in the normality of the residuals. Most data points align closely with the expected line, hinting at a

closer fit to a normal distribution. Yet, there are minor departures at the distribution's extremities that suggest it hasn't achieved perfect normality. The transformation has undeniably bettered the distribution, but there are still slight imperfections. It might be prudent to consider additional tests or tweaks to the model to further refine these nuances.

Model Selection

When determining the best-fitting model, it's essential to consider four primary indicators: adjusted R-squared, Mallow's Cp, AIC, and BIC. Together, these metrics offer a comprehensive view of the model's alignment, intricacy, and forecasting ability. After evaluating all options and selecting models that exhibit both the highest adjusted R-squared and the lowest AIC and BIC values, I've identified six top-performing models.

Model 1

log_suicides_per_100k_1 ~ C(country) + C(sex) + C(age_group) + population + gdp_year + gdp_per_capita

Model 2

log_suicides_per_100k_1 ~ C(country) + C(sex) + C(age_group) + gdp_year + gdp_per_capita + C(generation)

Model 3

log_suicides_per_100k_1 ~ C(country) + year + C(sex) + C(age_group) + population + gdp_year + gdp_per_capita

Model 4

log_suicides_per_100k_1 ~ C(country) + year + C(sex) + C(age_group) + gdp_year + gdp_per_capita + C(generation)

Model 5

log_suicides_per_100k_1 ~ C(country) + C(sex) + C(age_group) + population + gdp_year + gdp_per_capita + C(generation)

Model 6

log_suicides_per_100k_1 ~ C(country) + year + C(sex) + C(age_group) + population + gdp_year + gdp_per_capita + C(generation)

Model	AIC	Adj_R2	Adj_R2	Cp	Num_Prediction
Model 1	57862.81	58694.39	0.72	101	100
Model 2	57837.62	58702.14	0.72	105	104
Model 3	57859.63	58691.22	0.72	101	100
Model 4	57839.24	58703.76	0.72	105	104
Model 5	57822.94	58687.46	0.72	105	104
Model 6	57823.87	58688.39	0.72	105	104

When the adjusted R-squared figures for all six models closely align, it becomes pivotal to prioritize models with the minimal AIC and BIC scores. This shift in focus is warranted as the variation in Mallow's Cp across each model doesn't notably deviate due to the count of predictors.

Final model of choice

If the adjusted R2 and Cp values don't distinctly point to the optimal model, we then prioritize models based on the lowest AIC and BIC scores. For the best model, we pick model number 5 that relevant metrics are: AIC at 57822.94, BIC at 58687.46, Adj_R2 at 0.7235, Cp at 105.0, and the model incorporates 104 predictors.

Model 5

log_suicides_per_100k_1 ~ C(country) + C(sex) + C(age_group) + population + gdp_year + gdp_per_capita + C(generation)

OLS Regression

Dep. Variable:	log_suicides_per_100k_1	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.724
Method:	Least Squares	F-statistic:	701.0
Date:	Tue, 10 Oct 2023	Prob (F-statistic):	0.00
Time:	14:14:05	Log-Likelihood:	-28806.
No. Observations:	27820	AIC:	5.782e+04
Df Residuals:	27715	BIC:	5.869e+04
Df Model:	104		
Covariance Type:	nonrobust		

Observation:

- The model exhibits a strong overall fit.
- As evidenced by the high R2 value of 0.725, indicating that approximately 72.5% of the variance in the dependent variable is explained by the independent variables.
- The adjusted R2 value of 0.724 confirms the model's robustness, accounting for the number of predictors.

Summary:

Preliminary Results:

- Omnibus Test: Omnibus: 16590.500 indicates significant deviation from normality (Prob(Omnibus): 0.000). The high value suggests substantial non-normality in the residuals.
- Jarque-Bera Test: JB statistic of 328026.789 further confirms non-normality in the data (Prob(JB): 0.00).
- Skewness: A high positive skewness of 2.497 suggests a long tail on the right side of the distribution.
- Kurtosis: High kurtosis (19.064) indicates heavy tails and potential outliers.

Final Model Results:

- Omnibus Test: Omnibus: 1272.306 indicates significant deviation from normality (Prob(Omnibus): 0.000). Although still non-normal, the value is considerably lower than the preliminary results, indicating an improvement.
- Jarque-Bera Test: JB statistic of 4309.273, while still confirming non-normality, is notably lower than the preliminary model (Prob(JB): 0.00).
- Skewness: The skewness reduced significantly to -0.097, suggesting a more balanced distribution.
- Kurtosis: Kurtosis decreased to 4.918, indicating a reduction in the presence of extreme outliers.

Improvements in Final Model:

- *Reduced Non-Normality*: The Omnibus and Jarque-Bera tests indicate a notable reduction in non-normality in the final model's residuals compared to the preliminary model. Although the final model still deviates from normality, the final model shows improvement.
- *Skewness Improvement*: The skewness reduced from 2.497 to -0.097, indicating a more symmetric distribution in the final model. This suggests a better fit to the data.

- *Kurtosis Reduction*: Kurtosis decreased from 19.064 to 4.918, indicating a reduction in the presence of extreme outliers in the final model. The distribution is closer to a normal distribution.
- *Moderate Reduction in Multicollinearity*: While both models show signs of multicollinearity, the final model indicates a moderate reduction in the severity of multicollinearity among predictors.

Regularization: Lasso (L1) and Ridge (L2)

- Further exploring ways to improve, we implemented approaches, such as Lasso and Ridge regression, in an effort to improve the performance of our model.
    - Ridge regression: Using L2 regularization, penalizes the large coefficients.
      `R^2 Score: 0.72; Adjusted R^2 Score: 0.714`
    - Lasso regression: Using L1 regularization, penalizes the large coefficients and pushes them to almost Zero.
      `R^2 Score: 0.72; Adjusted R^2 Score: 0.7149`
- Despite our exhaustive efforts, there was little change in the R2 measurements.
- This result highlights a key point in data analysis: regularization strategies do not always produce the desired results for all datasets.
- Lasso and Ridge regression are essential methods, particularly when dealing with multicollinearity or high-dimensional data. In order to reduce potential overfitting and improve our model, we made an effort to use these strategies.
- However, regularization's basic restrictions were resisted by our data's inherent complexity, which is defined by complicated interactions and a wide range of variables.

Conclusion: our exploration of regression analysis has been insightful. We now know how to explore data to discover connections between various phenomena; we also put it into practice. We've uncovered how things affect one another by applying different regression algorithms and regularization techniques. This report demonstrates what we have learned and how we might use it in our future endeavors.