Data Analysis and Visualization using Python

Alip Yalikun

Virginia Tech

CS5764 Information Visualization

web link https://dashapp-473886336048.us-east1.run.app/

Table of Content

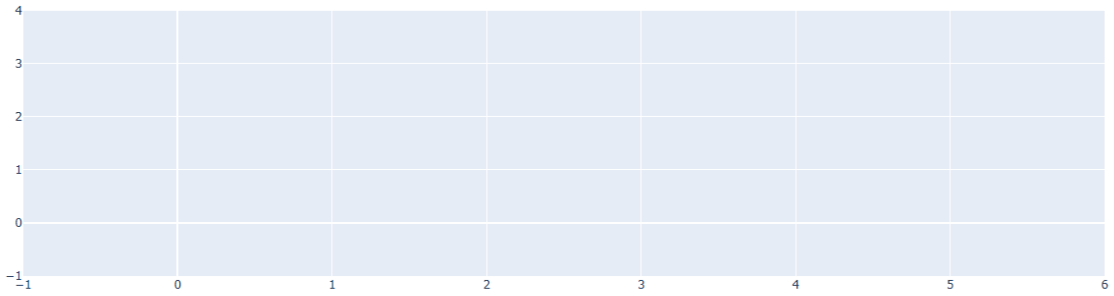
Table of Figures

Data Cleaning | Outlier Detection and Removal | Dimensionality Reduction | Normality Tests | Data Transformation | Statistics | Dynamic Plotting

Data Cleaning Methods

Dataset contains 15778 duplicate rows and 0 missing values.

CLEAN DATA



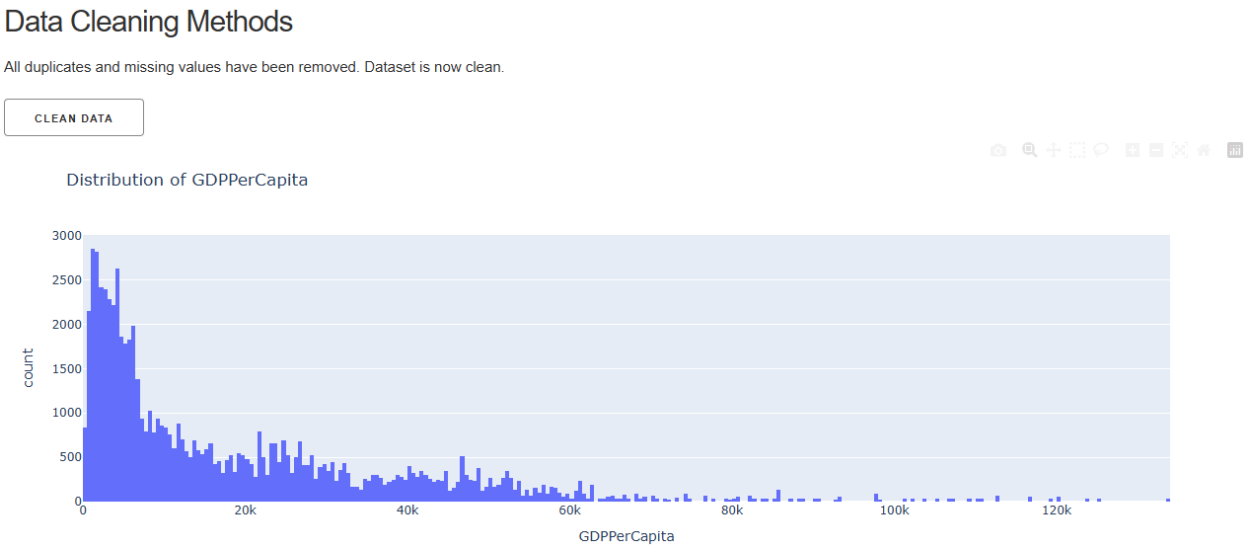


Figure 2.

Outlier detection and removal.

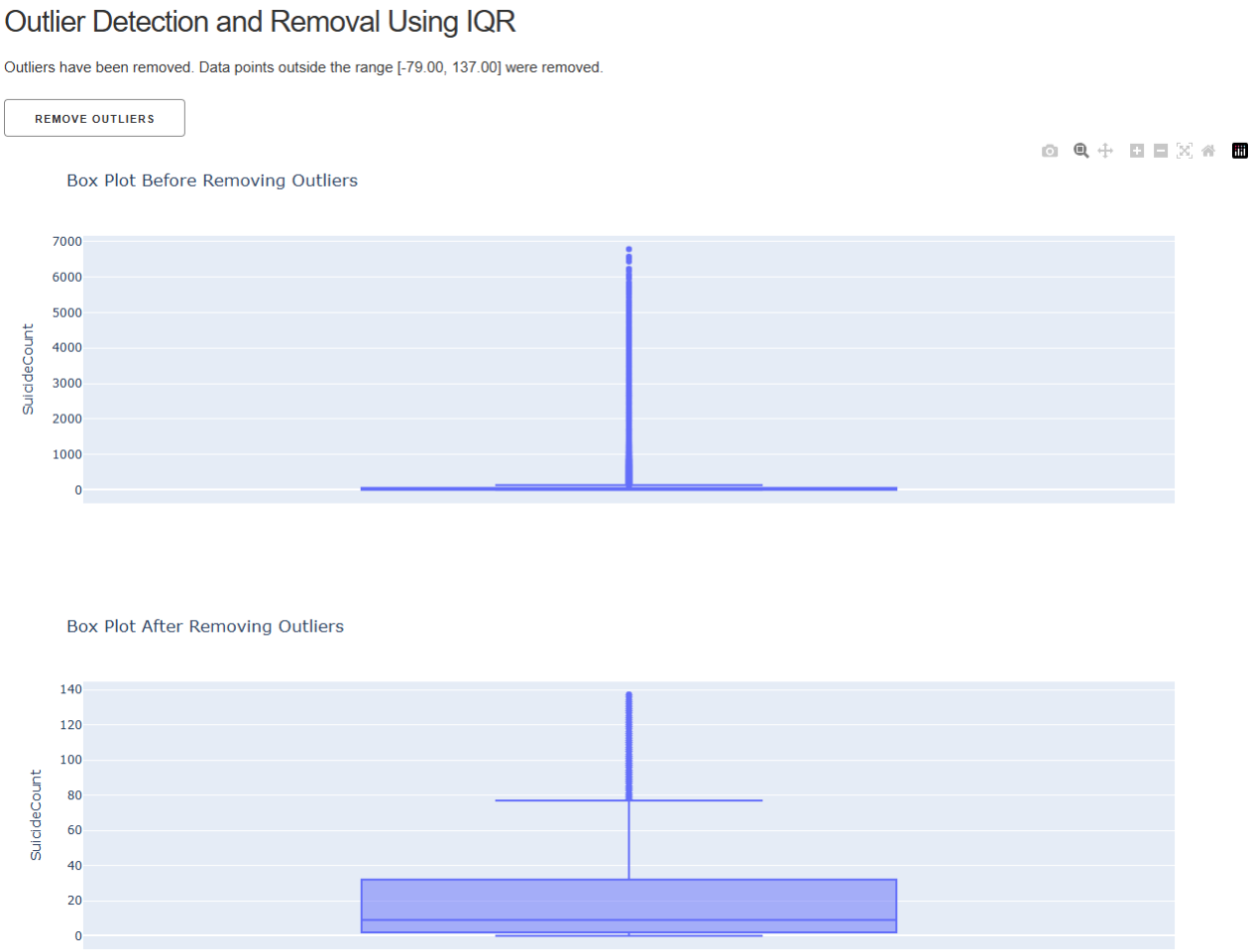

Outlier Detection and Removal Using IQR
Outliers have been removed. Data points outside the range [-79.00, 137.00] were removed.
REMOVE OUTLIERS
Box Plot Before Removing Outliers
SuicideCount
7000
6000
5000
4000
3000
2000
1000
0
Box Plot After Removing Outliers
SuicideCount
140
120
100
80
60
40
20
0

Figure 3.

PCA

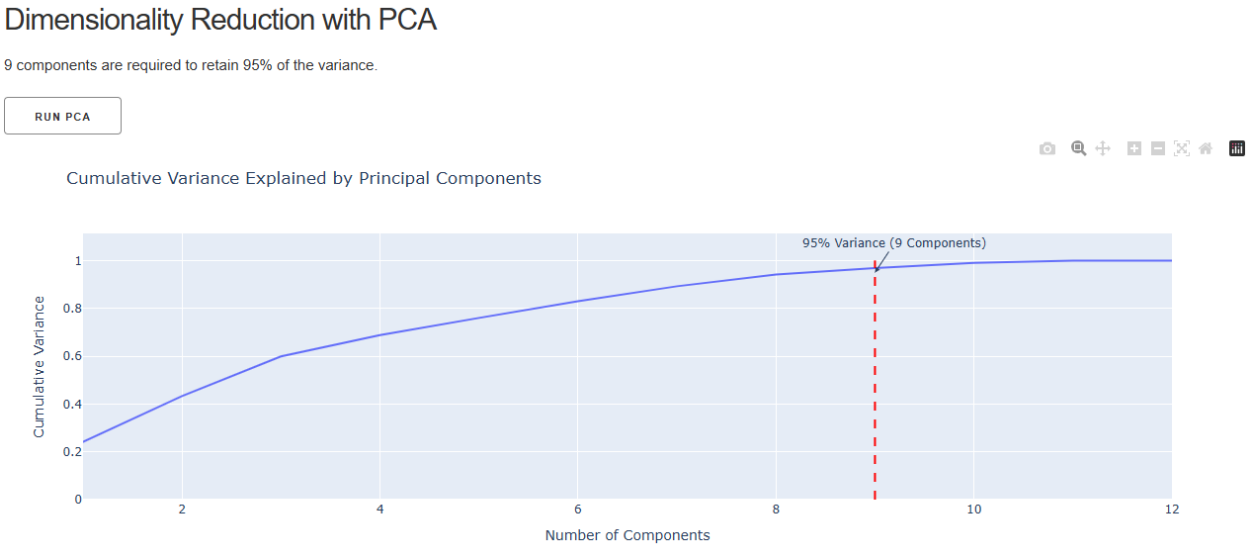


Figure 4.

Normality test.

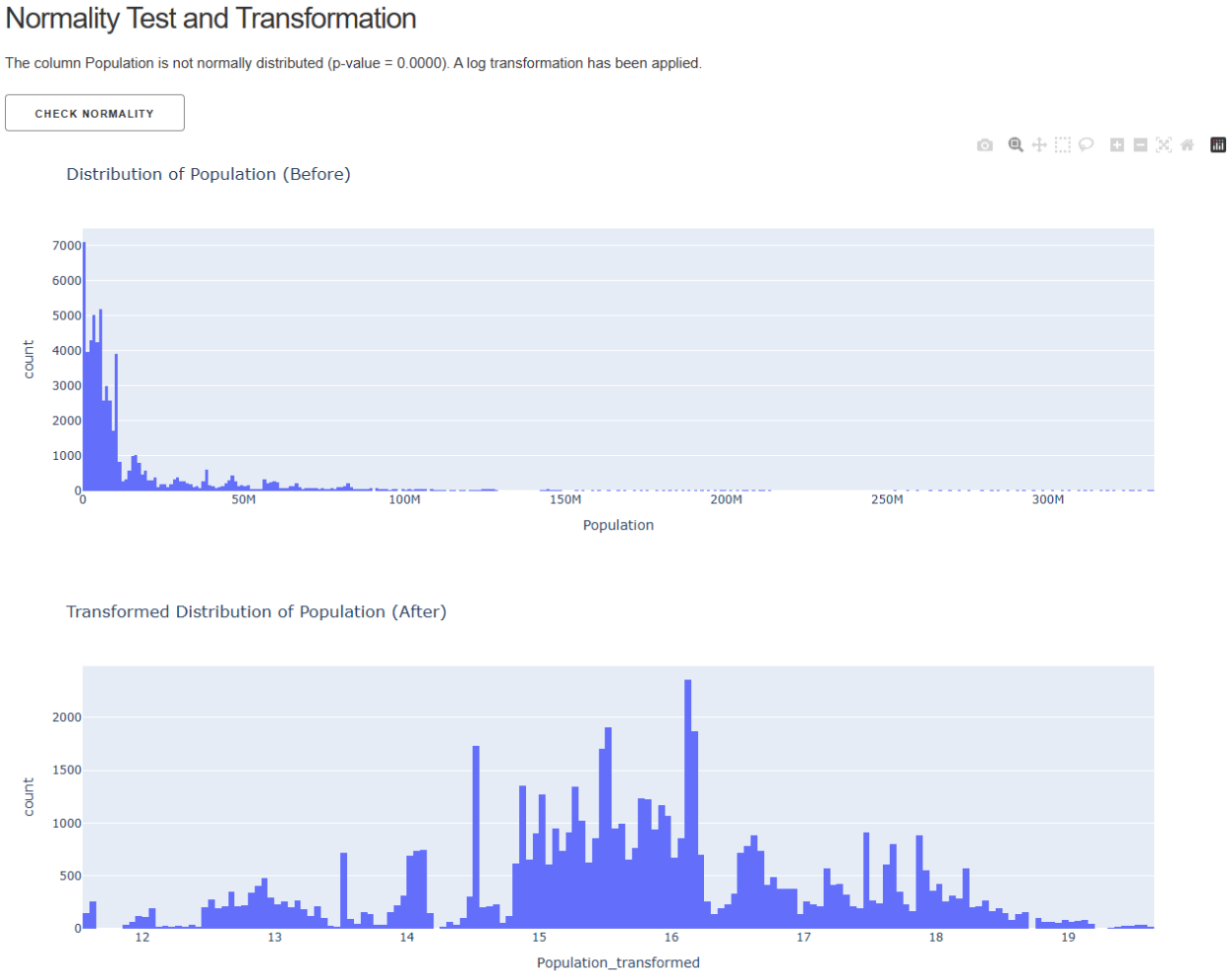


Figure 5.

Log and min max transformation.

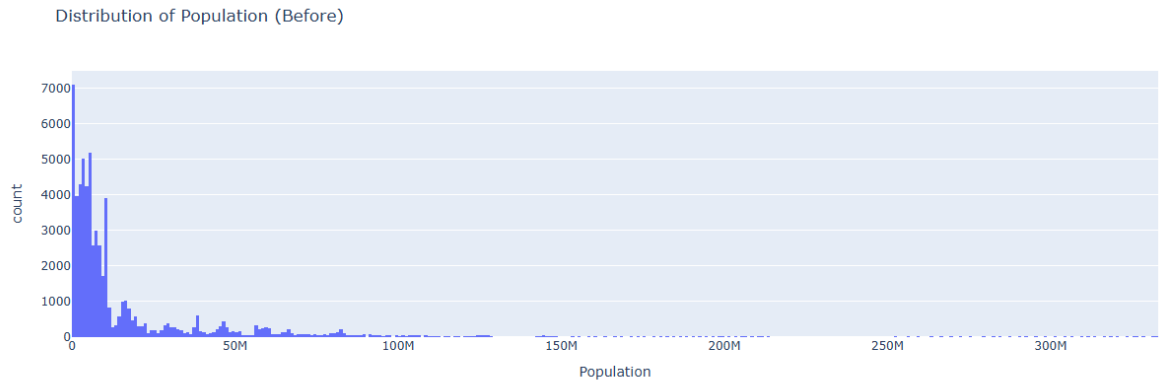
Apply Data Transformation

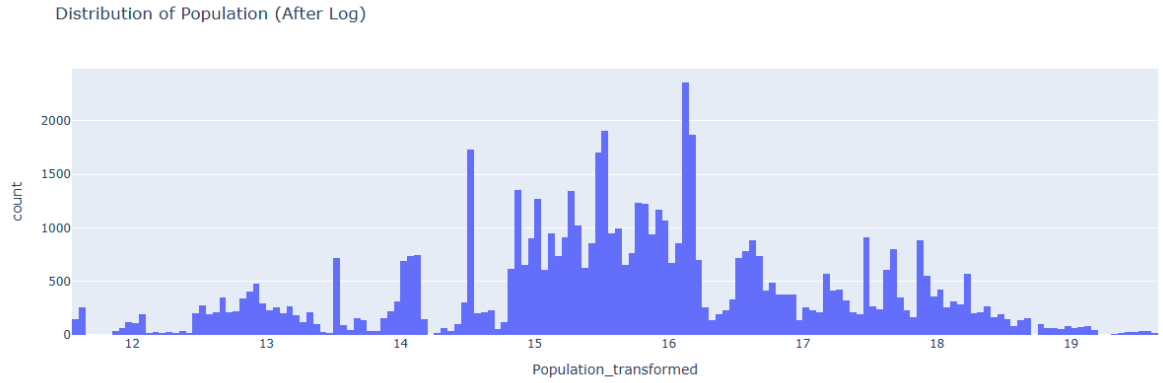
Select Transformation Type:

Log Transformation

Log transformation applied. Transformation applied to Population.

APPLY TRANSFORMATION





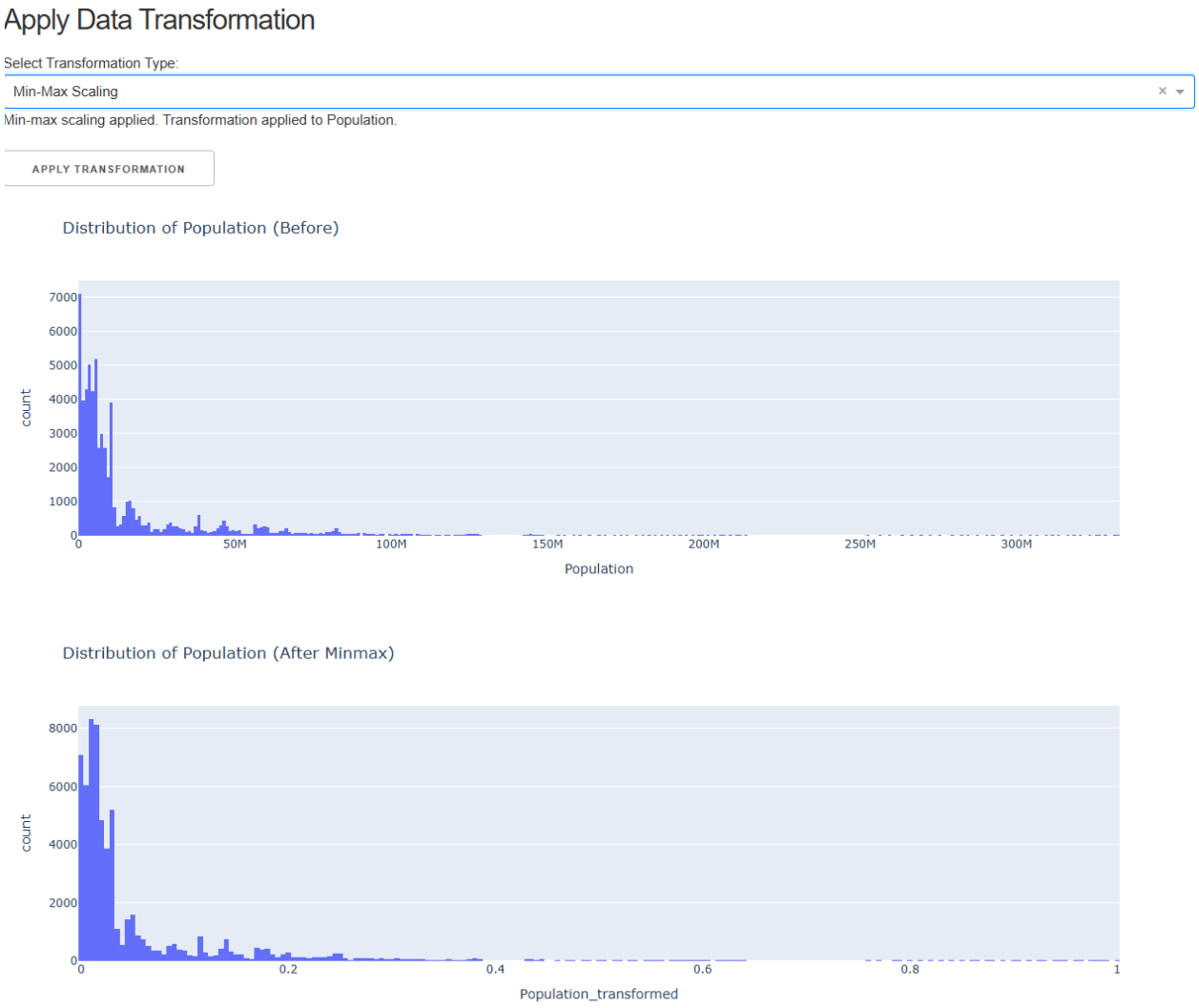


Figure 6.

Statistics.

Figure 7.

Line plot.

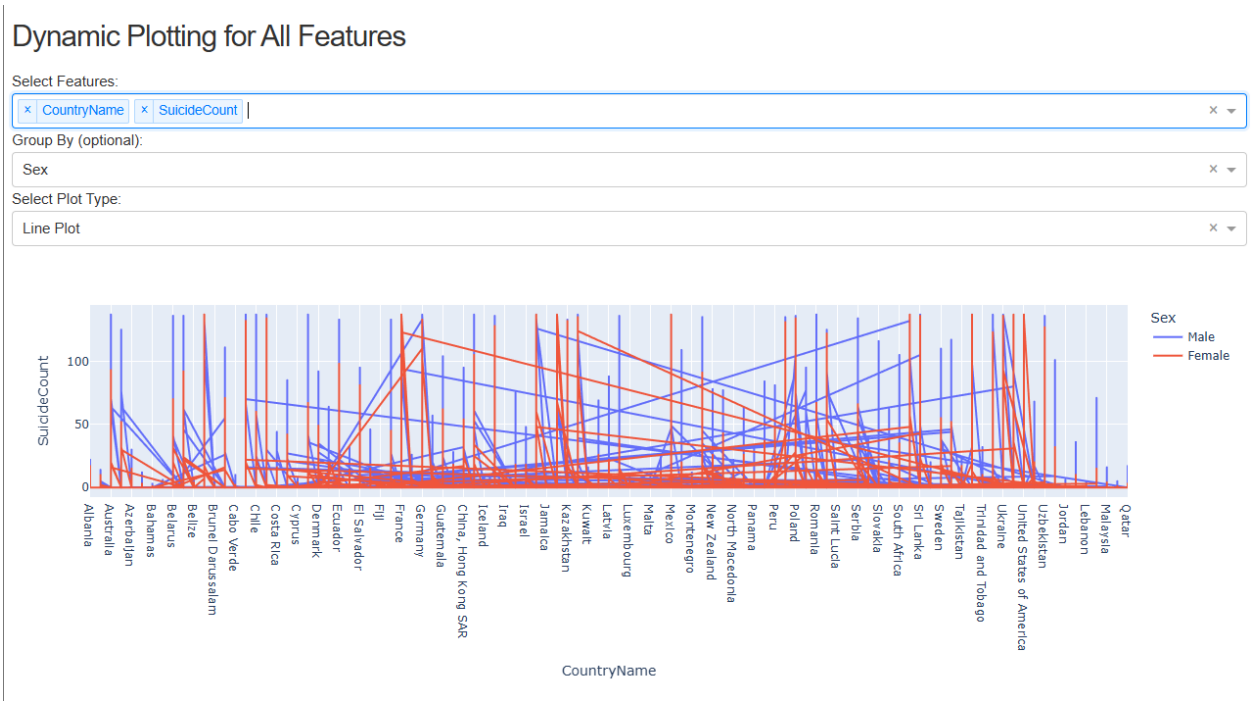


Figure 8.

Grouped bar plot.

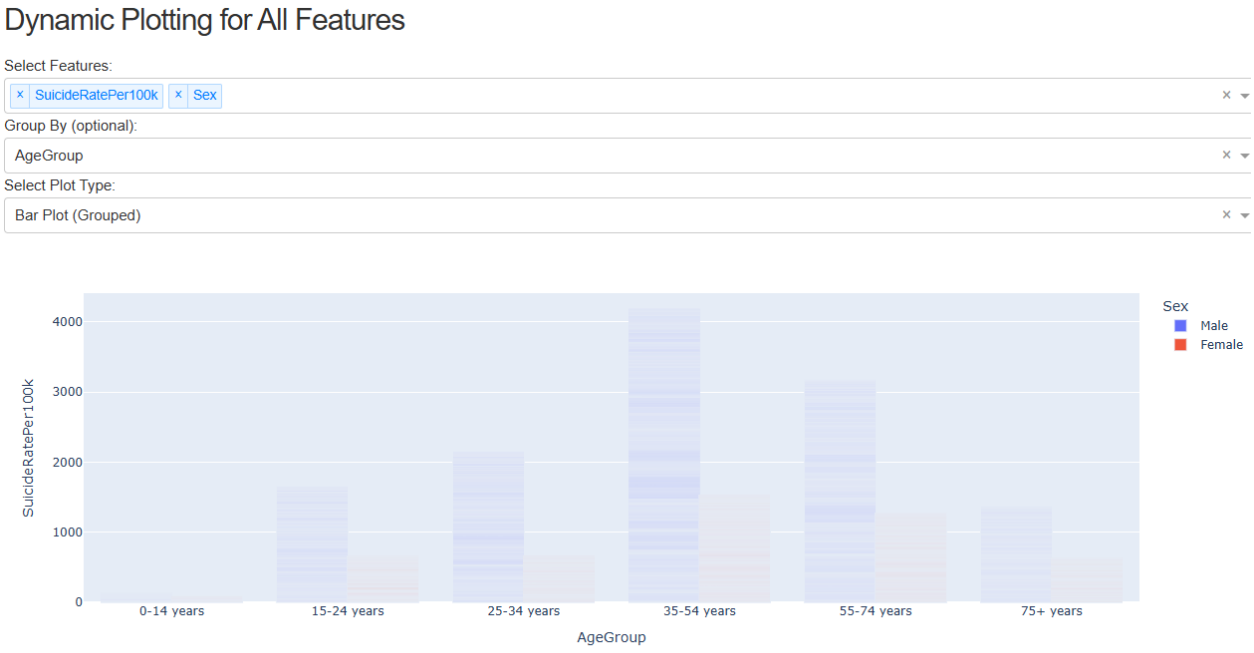


Figure 9.

Count plot.

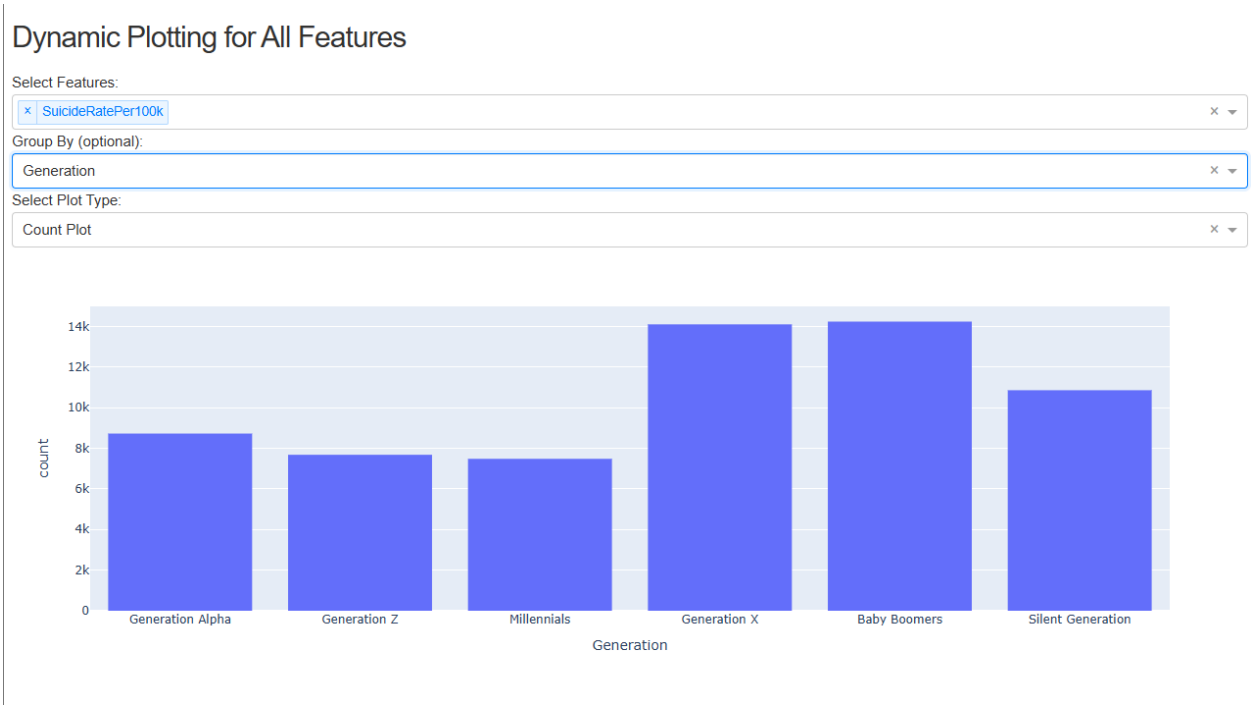


Figure 10.

Pie chart.

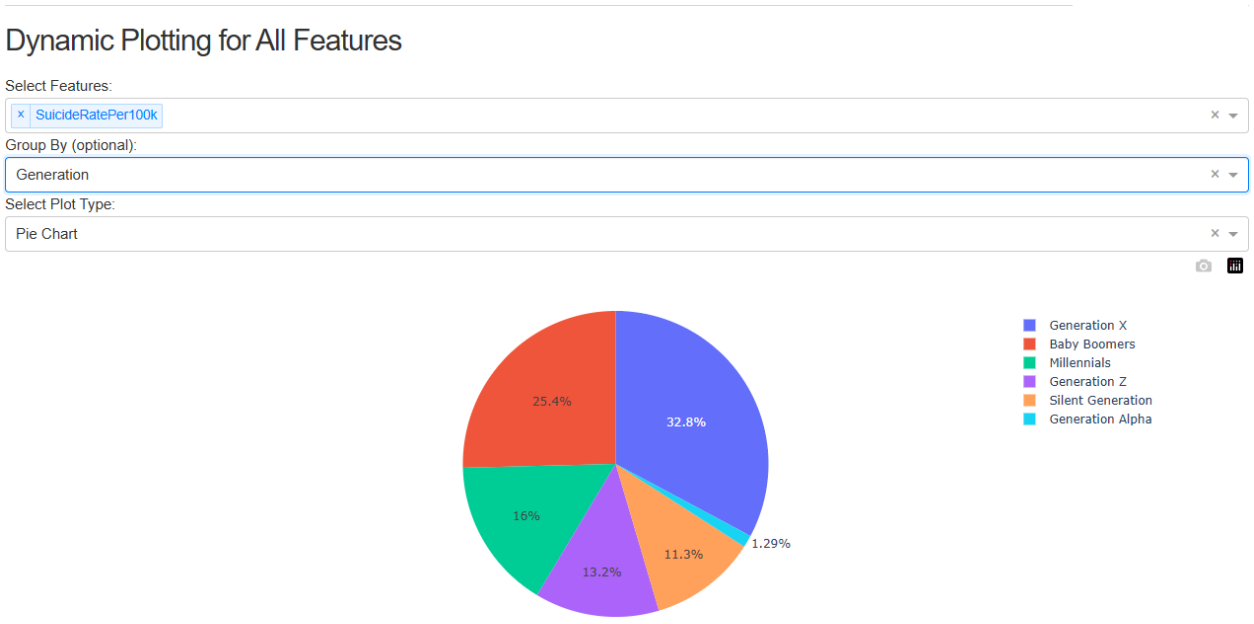


Figure 11.

Dist. Plot.

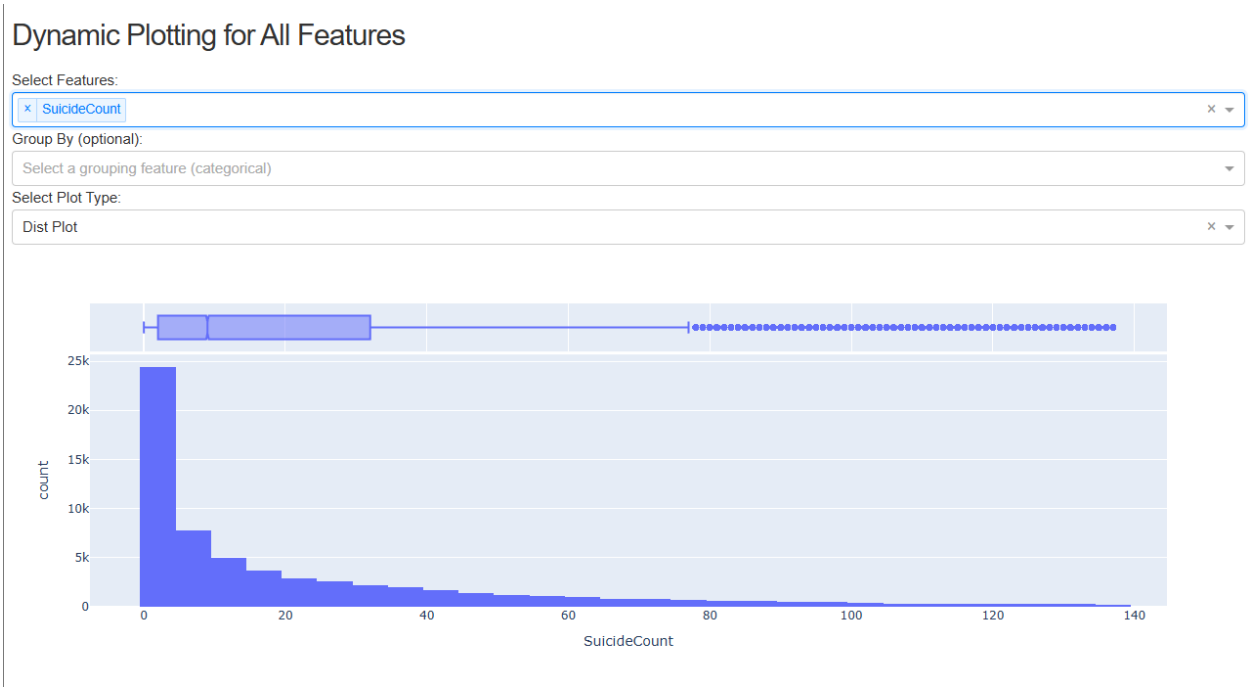


Figure 12.

Pair plot.

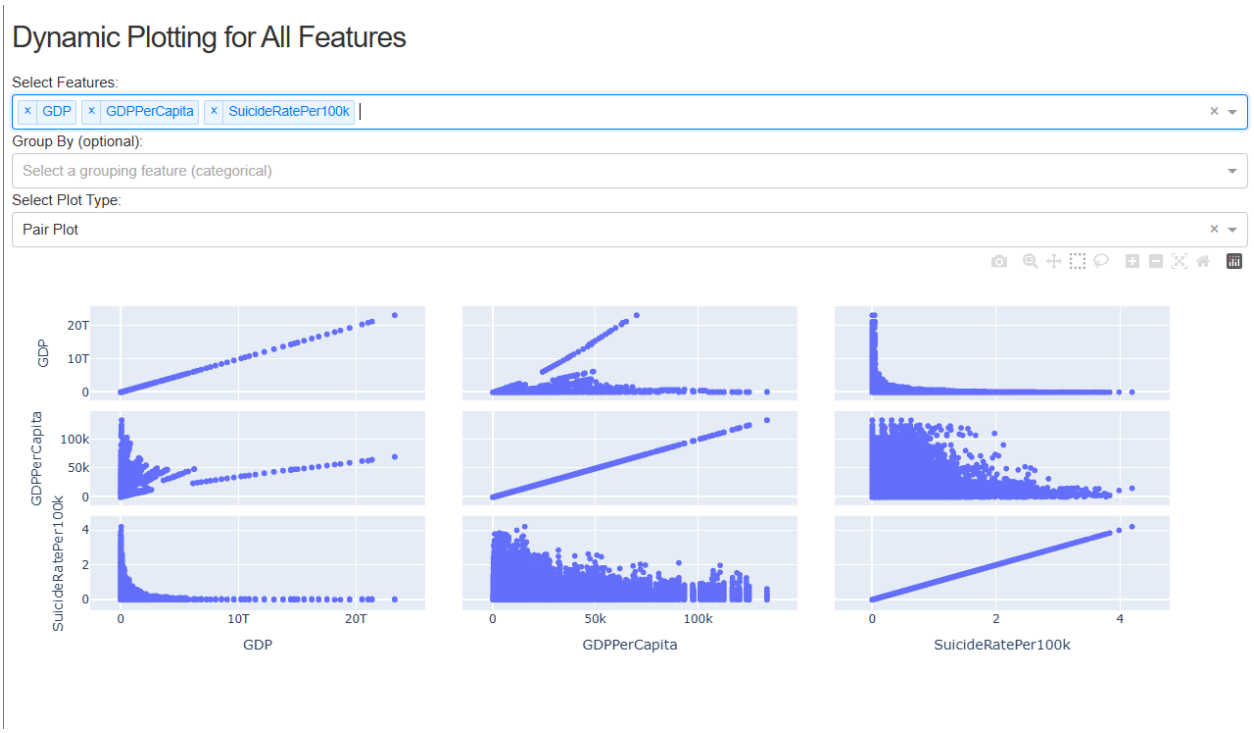


Figure 13.

Heat map.

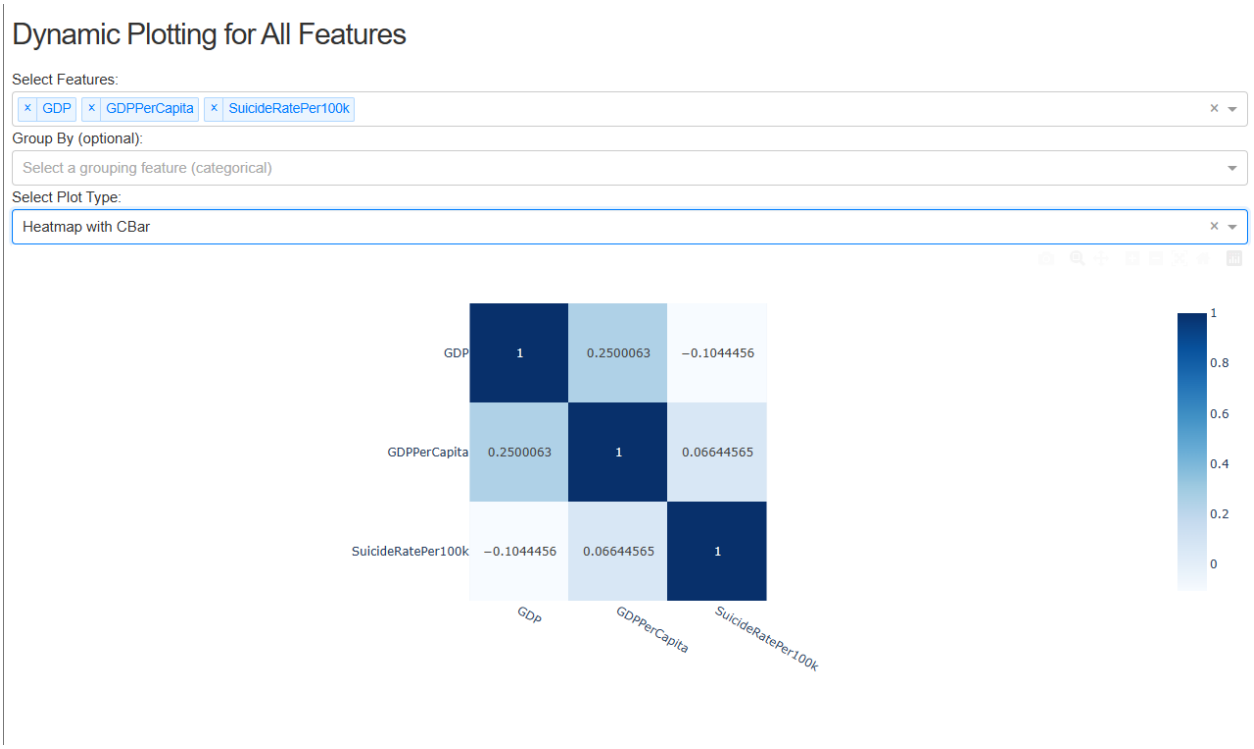


Figure 14.

Qq plot.

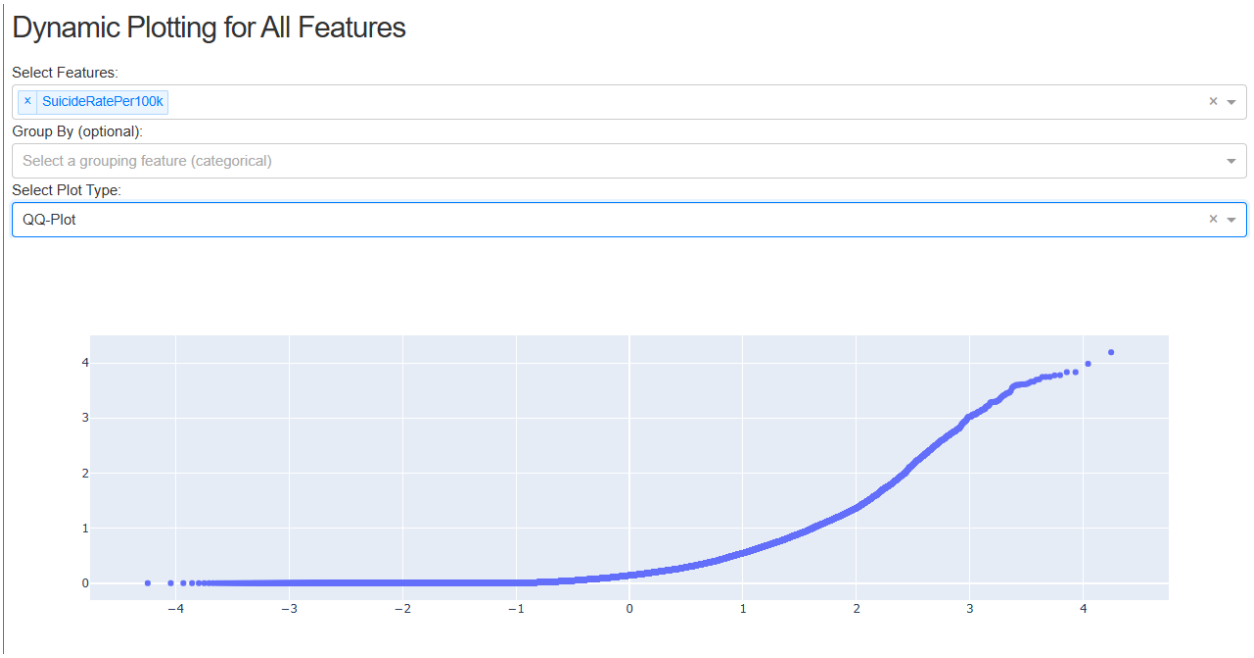


Figure 15.

KDE plot.

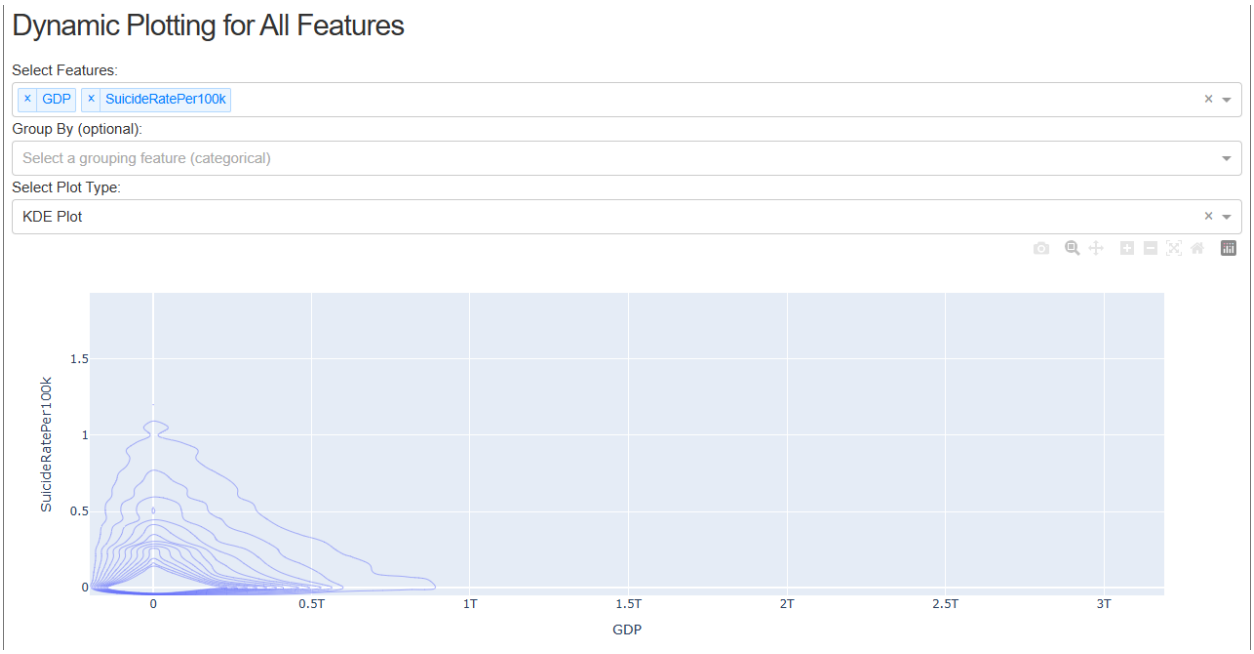


Figure 16.

Box plot.

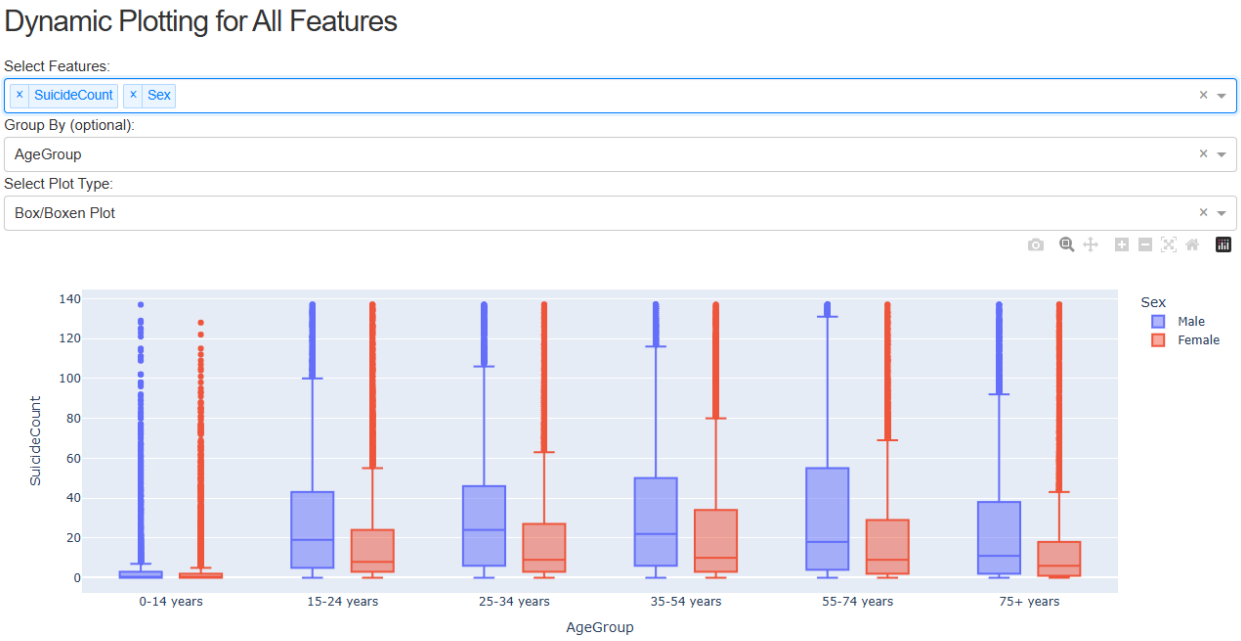


Figure 17.

Area plot.

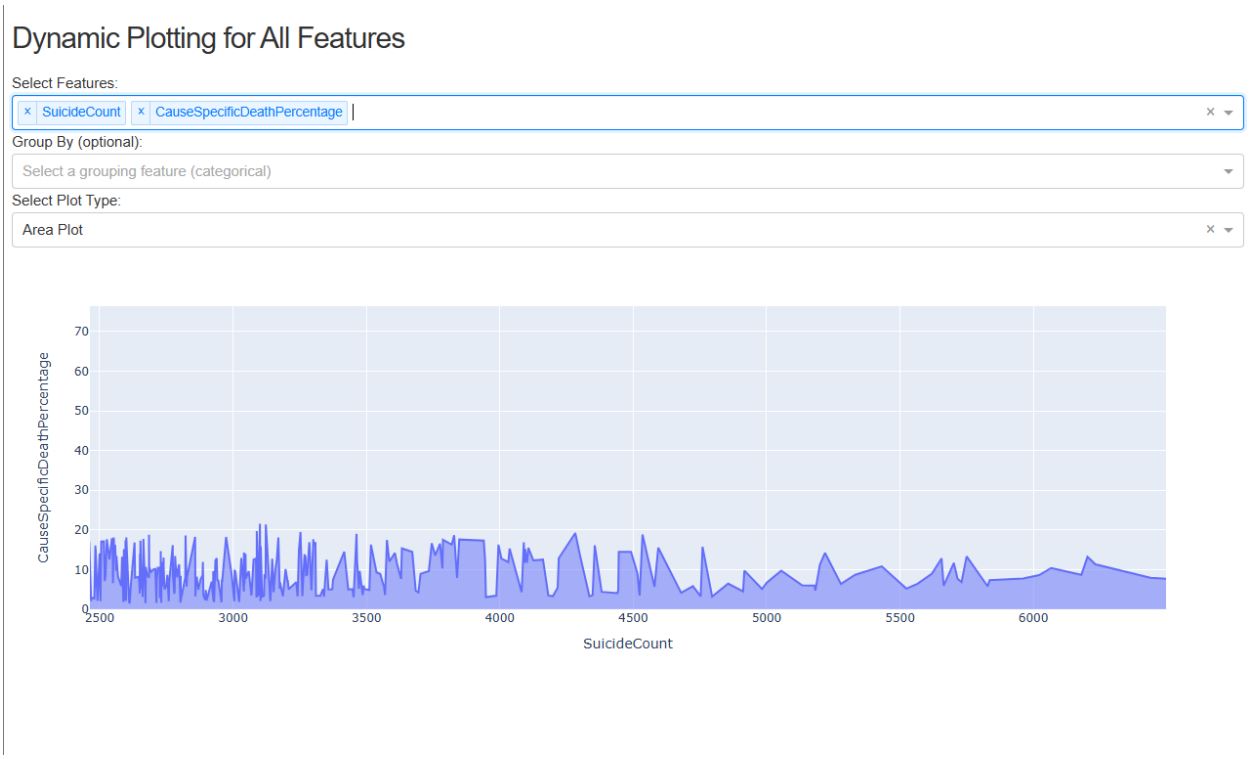


Figure 18.

Violin plot.

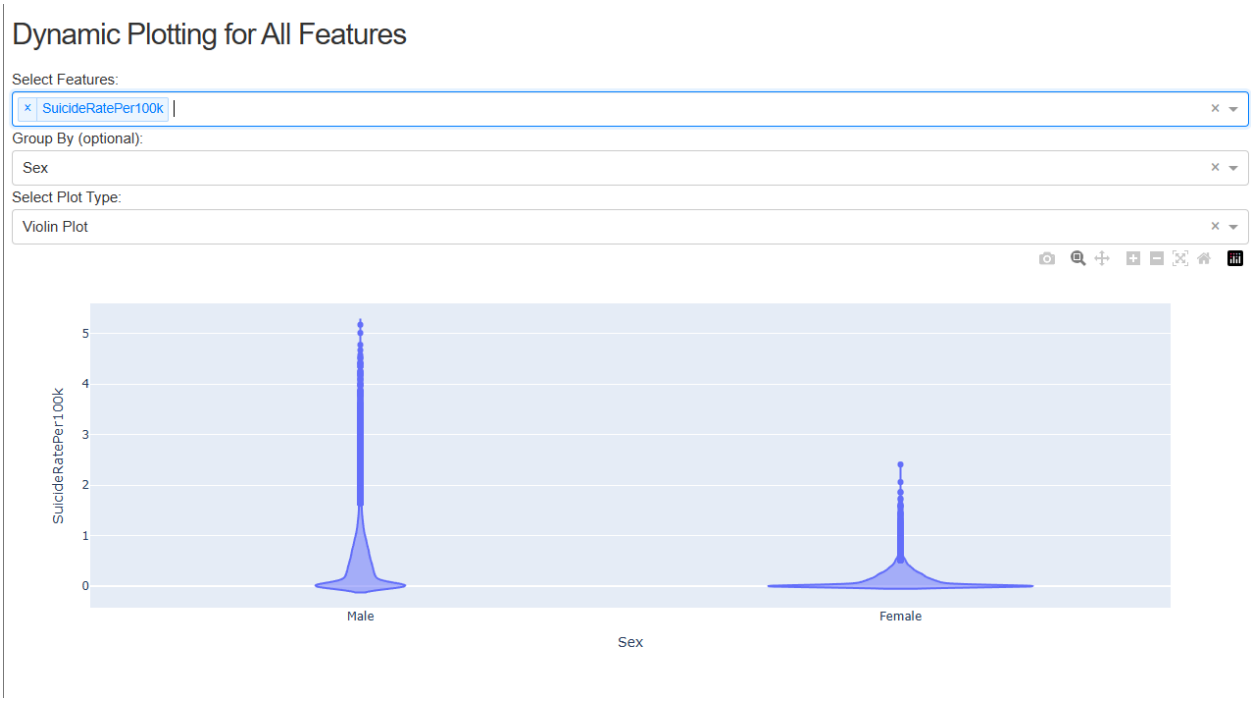


Figure 19.

Rug plot.

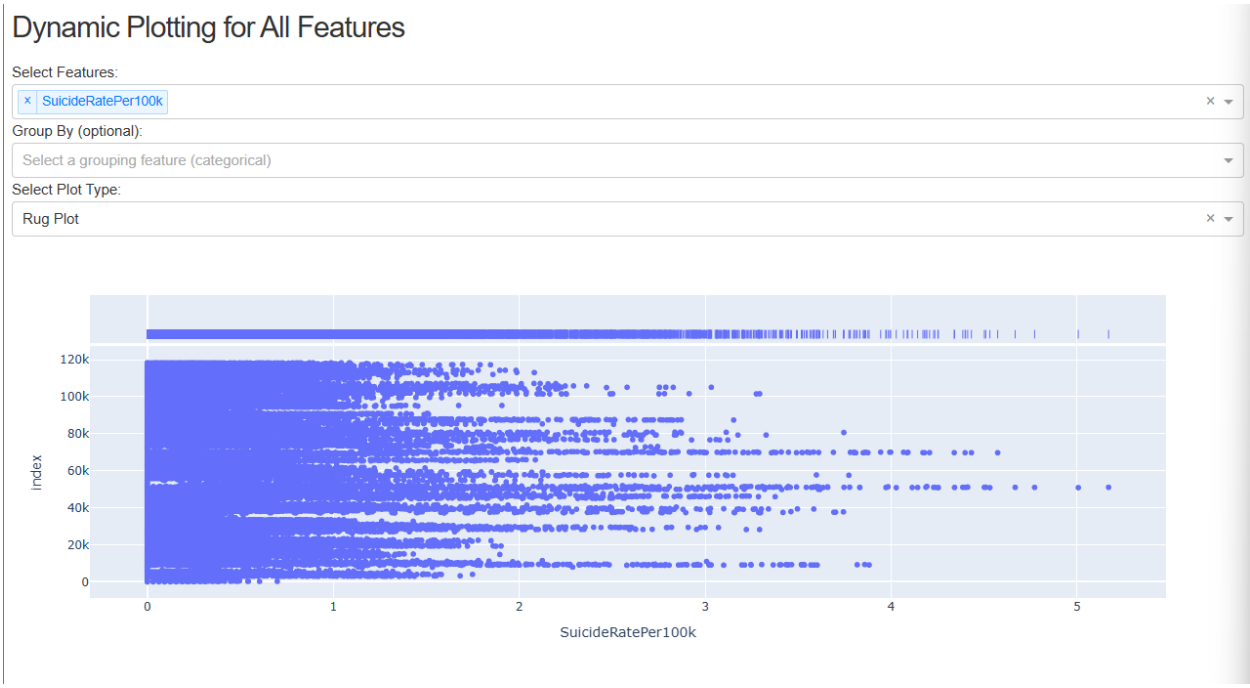


Figure 20.

3D plot.

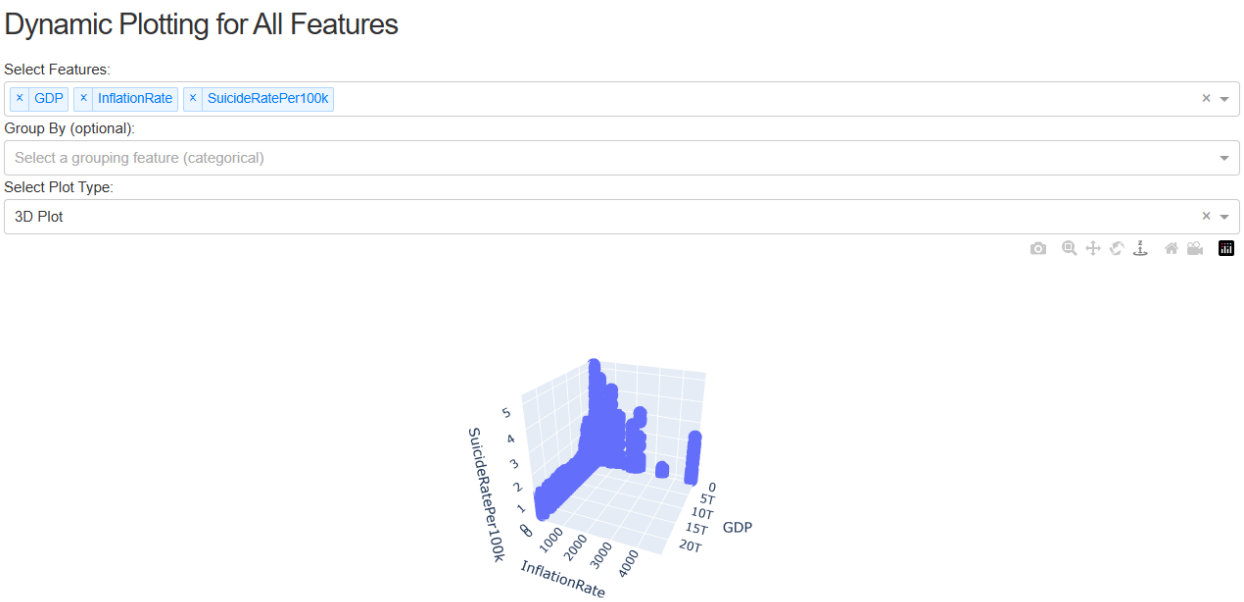


Figure 21.

Contour plot.

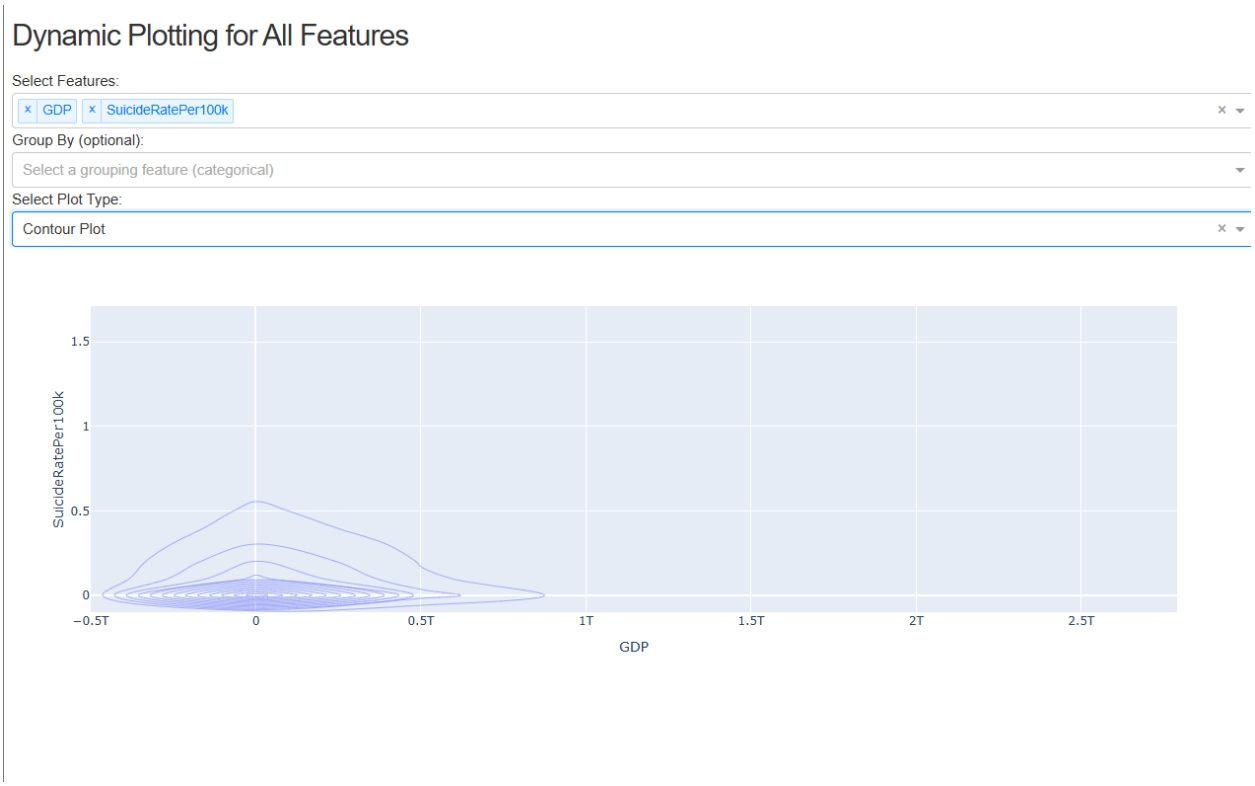


Figure 22.

Hexbin plot.

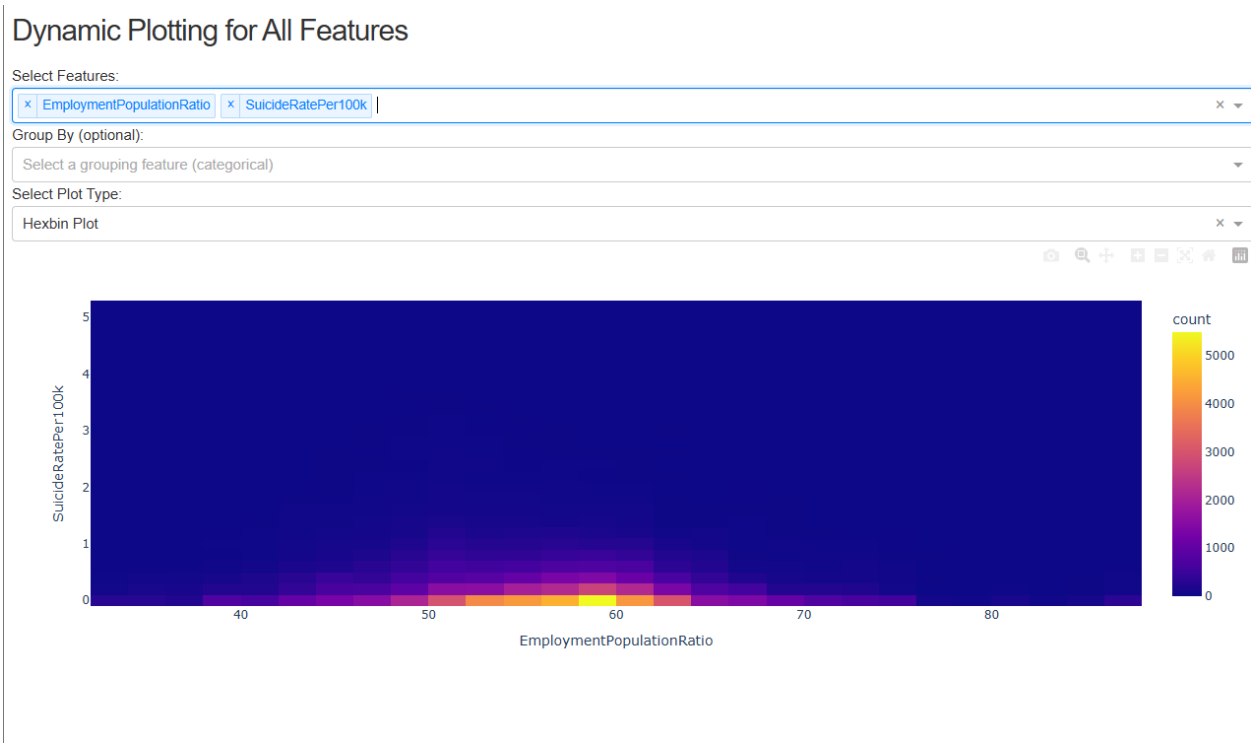


Figure 23.

Strip plot.

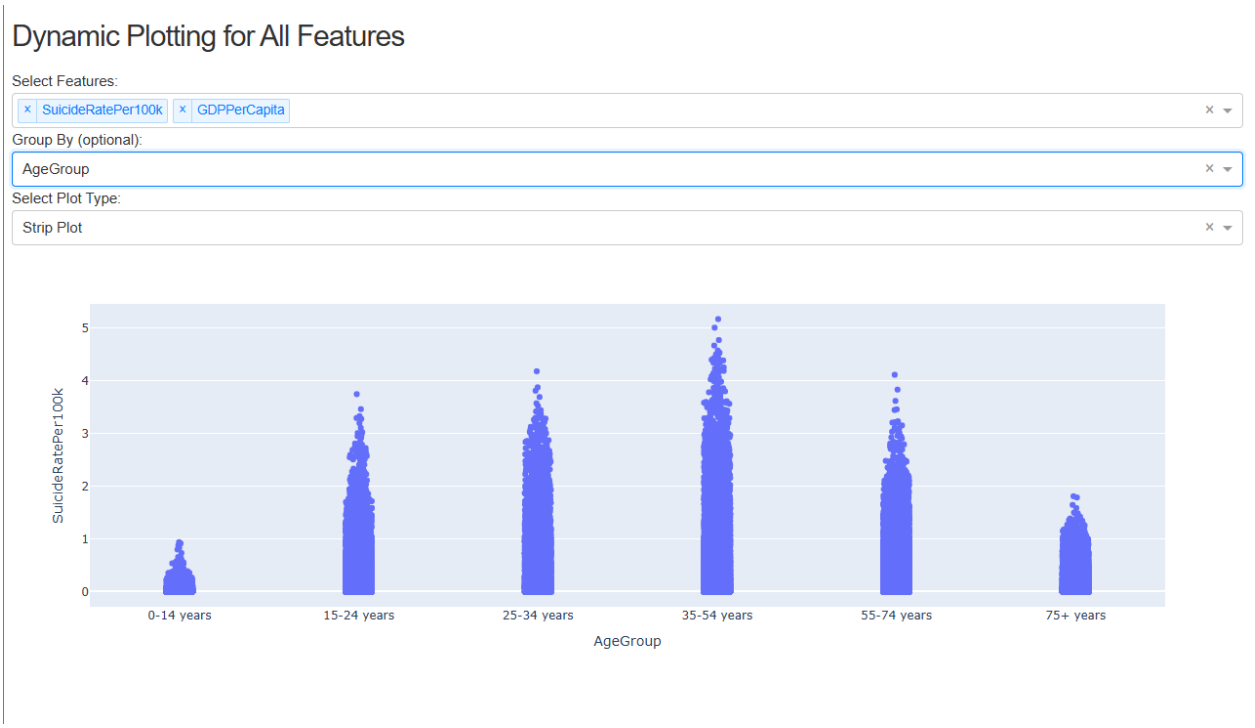
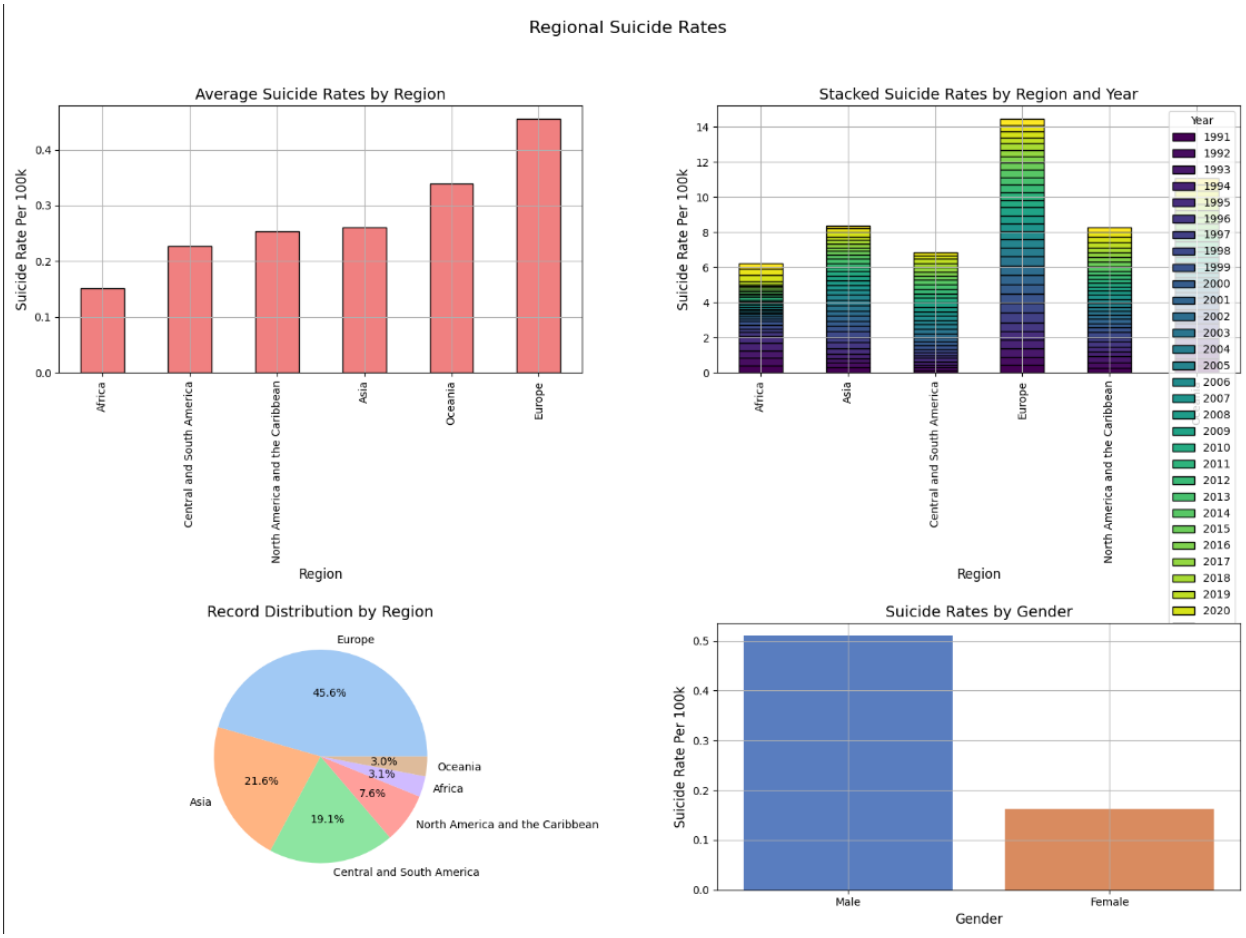


Figure 24.

Subplot.



Abstract

Suicide is a serious topic that has been talked about throughout the history of time. This paper explores the world of data visualization using python coding language. It demonstrates the process of visualizing selected dataset in various visualization techniques. Applying methods such as data cleaning, outlier detection and deletion, feature selection using PCA, statistical analysis, and finally displaying all that information on a user-friendly dashboard.

Keywords: information visualization, outlier detection, PCA, statistics.

Introduction

This study shows the process of understanding and visualizing a selected data set using python coding language. This study goes through various techniques, including exploratory data analysis, data processing, statistical analysis and visualizing data using an interactive dashboard made by Dash python library.

Description of the Dataset

The dataset contains 118,560 rows of observation and 18 columns of features. The dependent variable of this data set is SuicideCountPer100k. The independent variables include GDP, population, socioeconomic, employment rate and inflation rates. This data set is relevant in industry because of its importance in understanding the trends in public health and socioeconomic factors. The independent variable often time do play a significant role in prediction the trends in real world application, like mental health interventions.

Data pre-processing

This dataset was nowhere near ready for further analysis, the pre-processing contains using various techniques included in Python library Pandas. Non existing elements and duplicated observations were deleted (f*igure 1*). The plot shows one of the features after the dataset was cleaned.

Outlier detection and removal

In this phase, outliers were detected using the IQR method and visualized with a box plot (*figure 2*). The after removal of outlier plot shows that values for Suicide Count beyond the dedicated IQR range were removed to better the quality of the dataset for future processing and analyzing.

Principal Component Analysis

After removing the outliers in previous phase. PCA was conducted, the functionality of PCA was to reduce feature dimensions while retaining 95% of the variance. 9 out of 12 features suffice the variance threshold, and 3 did not (*figure 3*). The conditional number and singular values were within the acceptable range.

Normality Test

The normality test was done by Shapiro-Wilk test. The feature Population was not gaussian. This resulted in a log transformation of the feature (*figure 4*).

Data Transformation

Data transformation was done by log transformation and min max transformation on the feature Population (*figure 5*). After transformation, the normality was improved, and this enhanced the dataset's ability to further analysis.

Heatmap

The heatmap (*figure 13*), highlighted the relationship between three variables, GDP, GDP per capita and suicide rate per 100k. Revealing minor correlation between suicide count and GDP per capita.

Statistics

The statistical portion of the dashboard (*figure 6*) includes, mean, median, and standard deviation. They were calculated for key features selected from the dropdown menu.

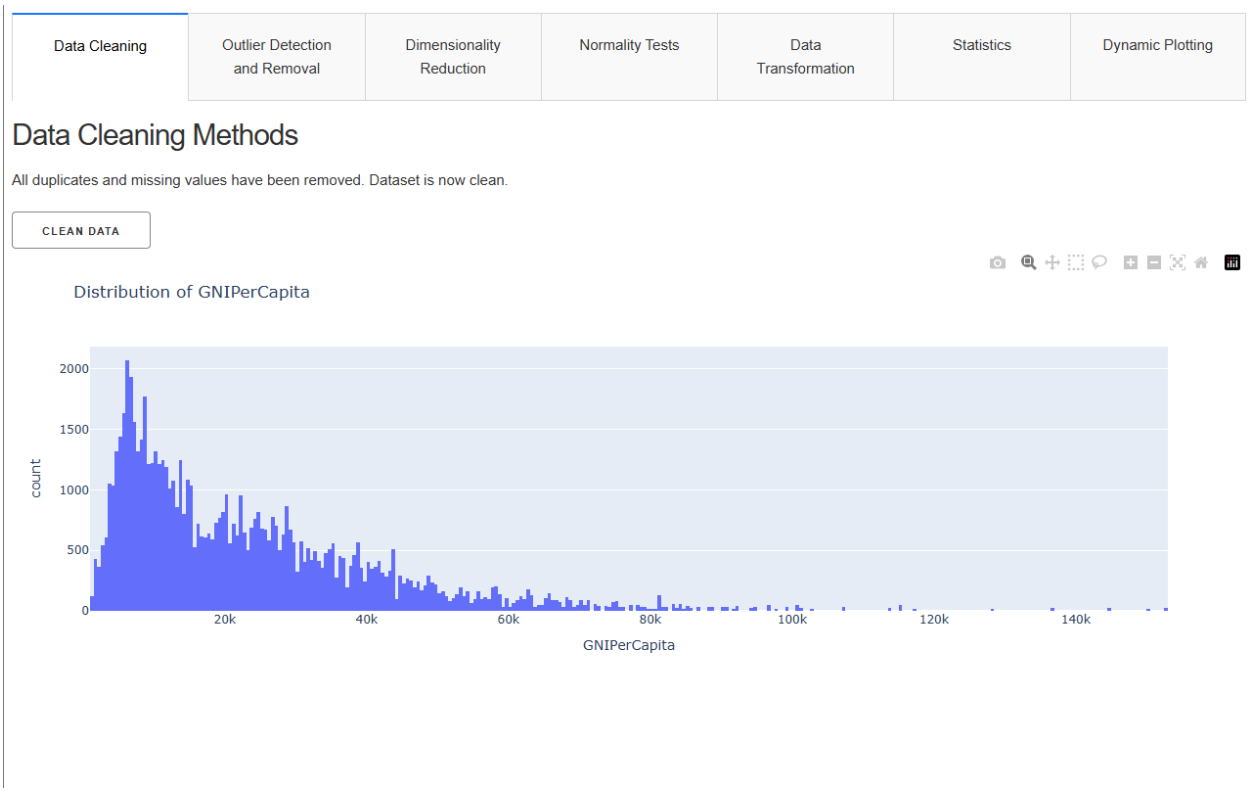
Data Visualization

The visualizations were listed in this paper through figure 7 to figure 23. Each observation will be covered in the **Observation** section.

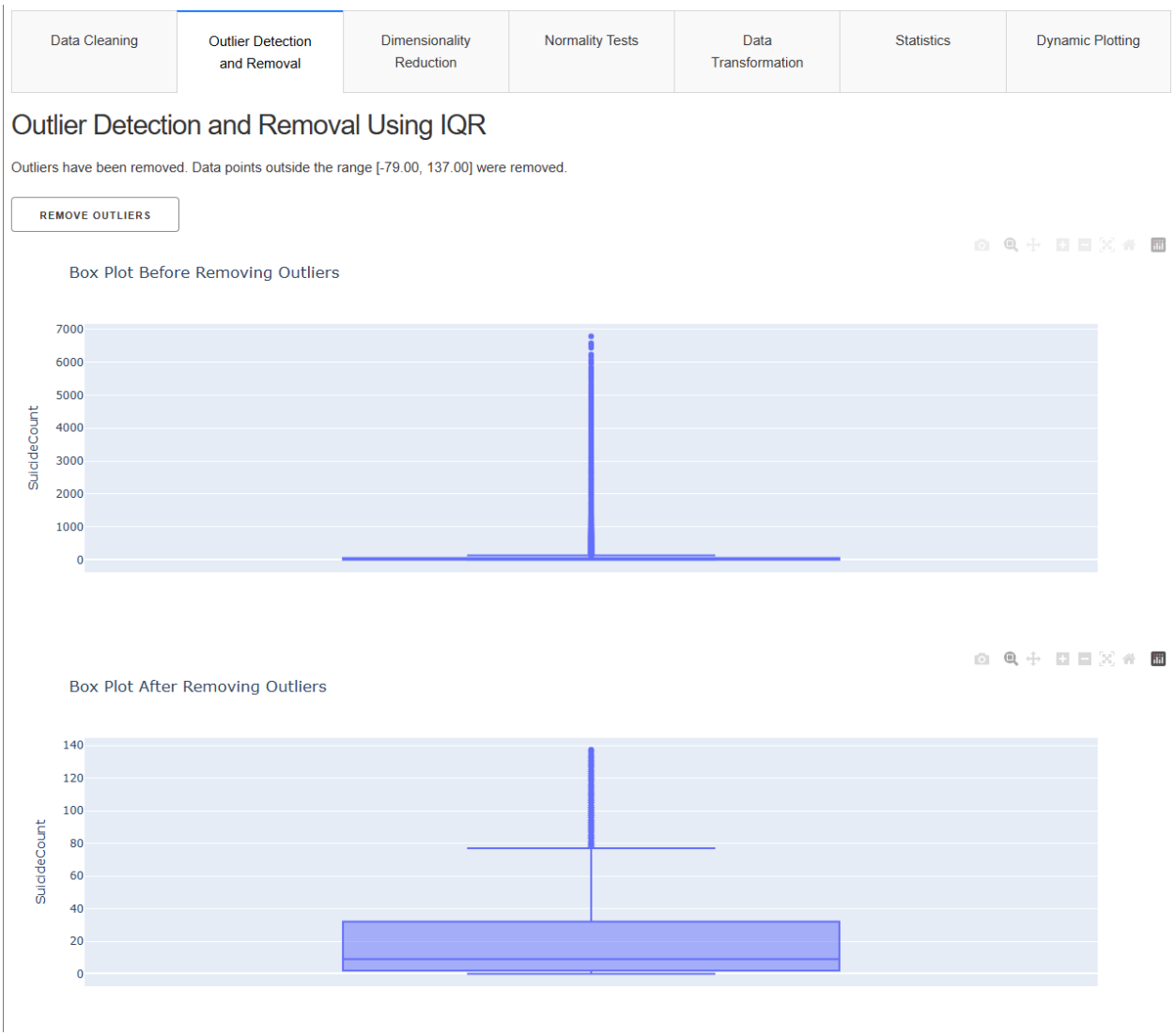
Subplot

The subplot in *figure 23* shows the global picture of suicide rates. From the subplot, Europe has higher rates compared to other regions (*plot 1 in figure 23*). Suicide rates have fluctuated across years, but still with noticeable peaks in certain regions (*plot 2 in figure 23*). The record of suicide was mostly distributed in the Europe region (*plot 3 in figure 23*). The final plot (*plot 4 in figure 23*) shows that male has much higher suicide records than female counterparts.

Dashboard



Data cleaning



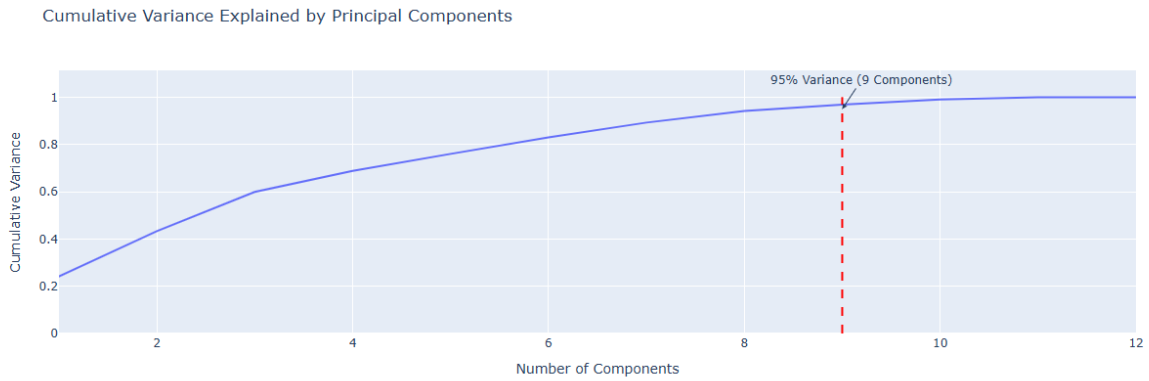
Outlier detection and removal

Data Cleaning | Outlier Detection and Removal | Dimensionality Reduction | Normality Tests | Data Transformation | Statistics | Dynamic Plotting

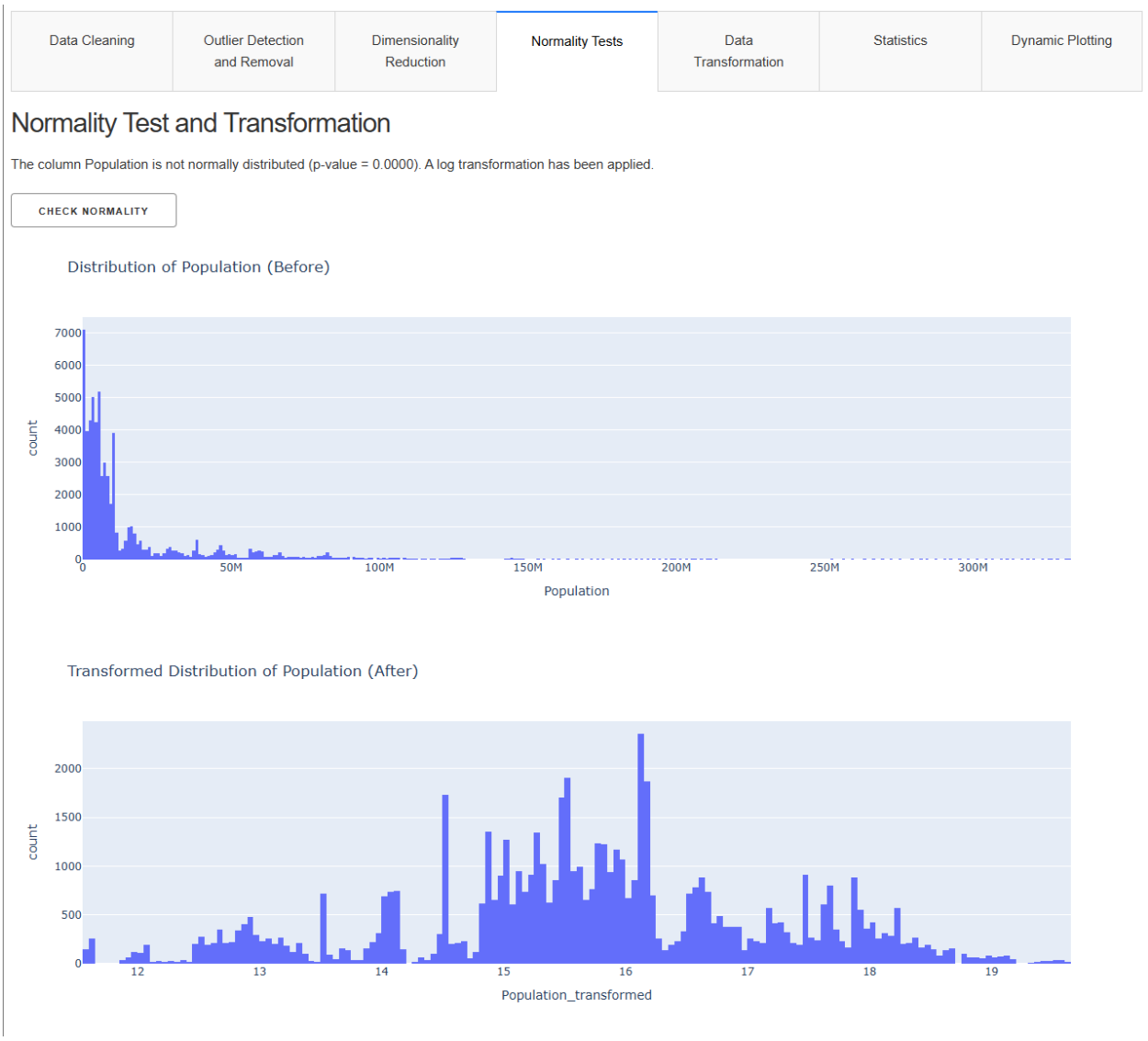
Dimensionality Reduction with PCA

9 components are required to retain 95% of the variance.

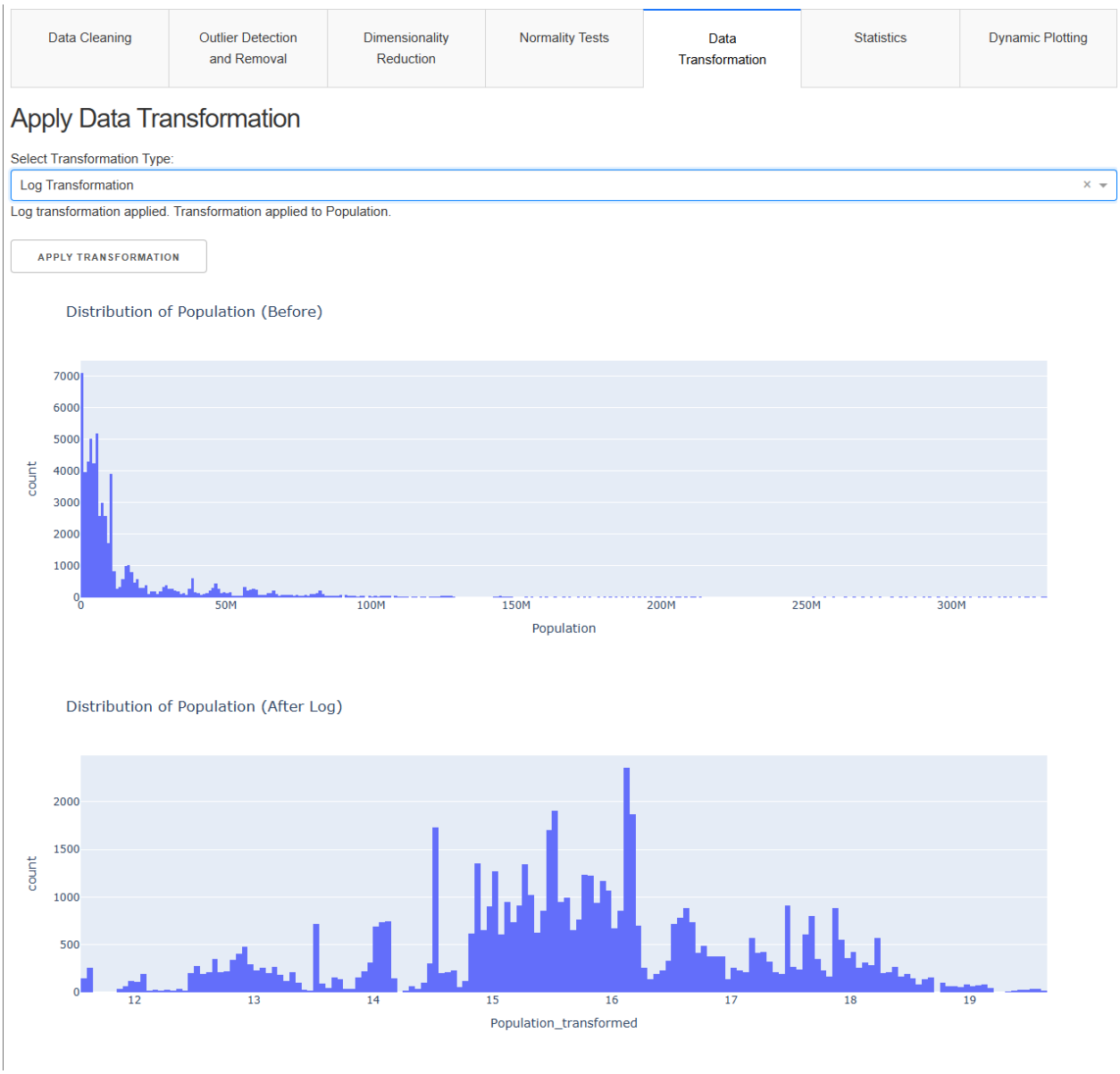
RUN PCA



PCA



Normality test



Data transformation

Data Cleaning | Outlier Detection and Removal | Dimensionality Reduction | Normality Tests | Data Transformation | Statistics | Dynamic Plotting

Dataset Statistics

Select Feature for Statistics:

EmploymentPopulationRatio

Mean: 56.15

Median: 56.60

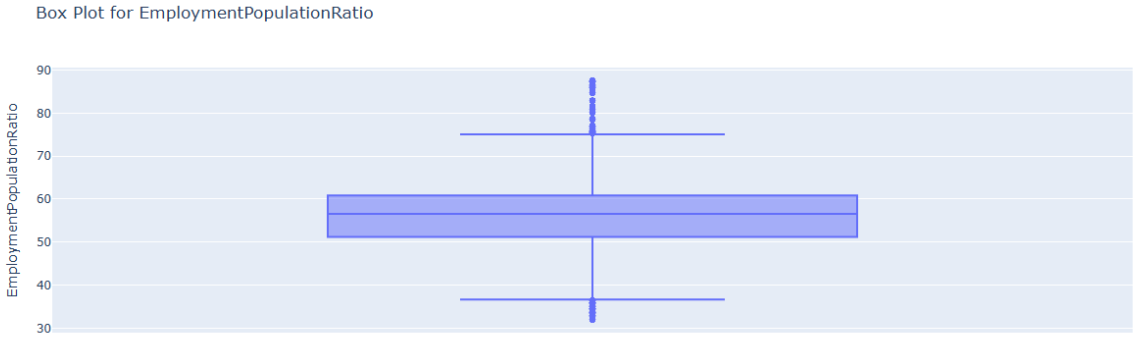
Standard Deviation: 7.96

Minimum: 32.03

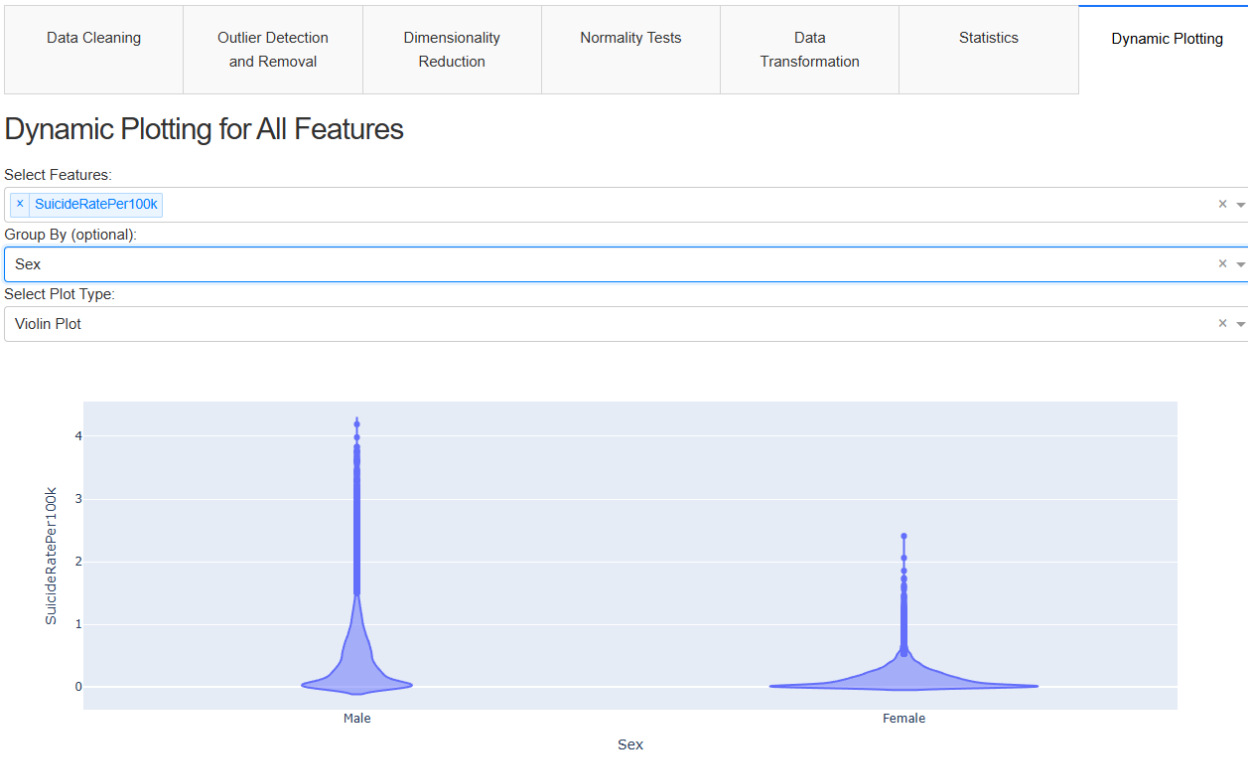
Maximum: 87.52

25th Percentile: 51.19

75th Percentile: 60.87



Statistic



Dynamic plotting

Observation

Line plot (*figure 7*), this plot shows major variation in suicide count across different countries, there is a clear distinction between suicide records for males and females, with male having higher records. Some countries show significant spike meaning there may be some cultural aspects for that to happen.

Grouped bar plot (*figure 8*), this plot shows that suicide rate per 100k population increases significantly in older age groups, the peak is in the 35-54 years age group category. Male suicide still is significantly higher than female record counts in all age group.

Count plot (*figure 9*), this plot shows Generation X and Baby boomers have the highest count of suicide rate, and generation like Silent generation and generation Alpha have the lower suicide rate records.

Pie chart (*figure 10*), this plot shows the generation X and baby boomers combined for most suicide records, at 33 percent and 25 percent.

Dist. Plot (*figure 11*), this plot shows the distribution of suicide count is highly skewed to the left, with most data points at lower values, this might indicate the need for data transformation.

Pair plot (*figure 12*), this plot shows no strong liner correlation with suicide count and GDP, GDP per capita, suggesting that suicide rates might depend on other features.

Heat map (*figure 13*), this plot indicates a weak correlation between suicide rates per 100k and GPD per capita, suggesting that suicide rates might depend on other features.

QQ plot (*figure 14*), this plot shows deviation from the diagonal line, indicating that suicide rate per 100k does not follow a normal distribution.

KDE plot (*figure 15*), this plot shows clustering of data point at lower GDP and suicide rate values, the density decreases as GDP increases, suggesting the wealthier country might have lower suicide rate records.

Box plot (*figure 16*), this plot shows male suicide counts are consistently higher across all the age groups, reinforcing earlier observation.

Area plot (*figure 17*), this plot shows a decreasing trend of cause specific percentages is visible as suicide count increases, indicating that higher suicide counts are associated with a bigger range of contributing factors.

Violin plot (*figure 18*), this plot shows males exhibit a wider range and higher density of suicide rates compared to females.

Rug plot (*figure 19),* this plot shows a dense clustering of data points at lower suicide rates, some high outliers are visible, meaning there might be some extreme cases.

3D plot (*figure 20*), this plot shows no linear relationship between suicide rates, GDP, and inflation rates, but higher GDP values are associated with lower suicide rates.

strip plot (*figure 23),* this plot shows with the 35-54 years age group having the widest spread, and higher GDP per capita values are associated with lower suicide rates.

Conclusion

From this study, all the graphs show important insights. The correlation between GPD and suicide rates are not as clear, but the graphs suggested that lower GDP valued countries have higher suicide rates compared to countries with higher GDP values. The dashboard designed with python was user friendly, but sometimes it takes a while to render the graph, suggesting a need for higher computational power. In conclusion, suicide is a serious topic, no amount of data observed can help to reduce the cases of suicide, make sure you are being vigilant and help anyone that is close to you if they are having a hard time.

Appendix

The python files, one for phase I, and another for phase II will be submitted on canvas separate from this paper.

References

Python Libraries: *Pandas, NumPy, plotty and other various ML libraries*.

Various passed information visualization assignments and in-class coding samples.

Dataset Source:

“Suicide Rates & Socioeconomic Factors (1990 - 22).” Onyango, Kaggle, 1 Mar. 2024,

https://www.kaggle.com/datasets/ronaldonyango/global-suicide-rates-1990-to-2022/.