PREPROCESSING or DATA CLEANING

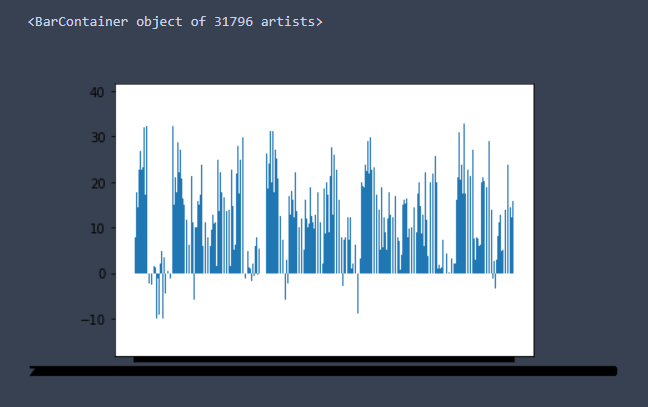
Outlier removal

The outliers can be a result of a mistake during data collection or it can be just an indication of variance in data. Python library seaborn is used to visualise data in box plots using the method sns.boxplot(). A **box plot** is a method for graphically depicting groups of numerical data through their quartiles, they also have vertical lines extending outside the box indicating the range of the data and outliers are plotted as individual points*.* Seaborn uses the interquartile range to remove outliers. Interquartile range is the value of the difference between the third and first quartiles. Data greater or less than 1.5 times the Interquartile range is an outlier which is removed by filtering the dataset. Any conversions of data to int/float has been made when required using the .astype() method for ease of plotting on a box plot.

EXPLORATORY DATA ANALYSIS

Data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs and makes it easier to identify trends, patterns, and outliers within large data sets. For visualisation of our data python library mathplotlib has been used which is used for static, animated and interactive visualisation in python. To plot the bar graphs plt.bar() is used.

We have used bar graphs for visualisation of categorical data in our dataset. In the graph for Temperature vs Date we have tried to study the mean temperature on each date specified and tried to study the patterns.



In the graph for Humidity vs date, similar to the previous one, we studied the trends set by the mean humidity each specified date. We also observed the shape of the curve these graphs display.

