

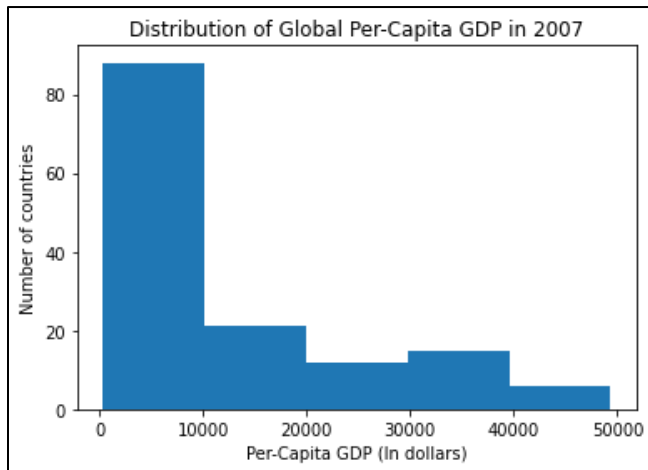
CO 544 Machine Learning and Data Mining

Lab 02

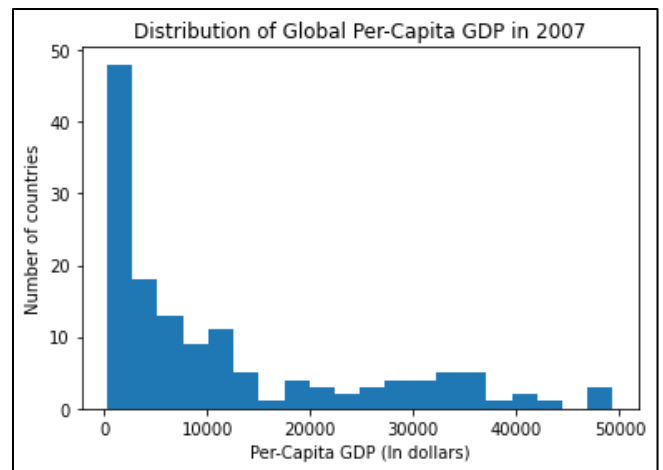
E/18/323

SEEKKUBADU H.D.

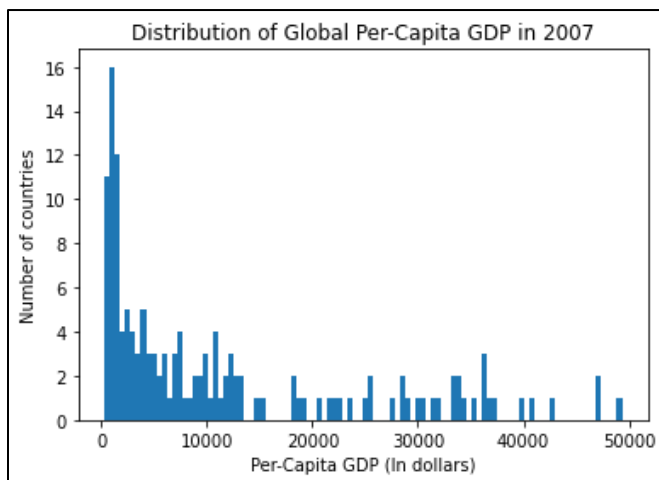
TODO 1



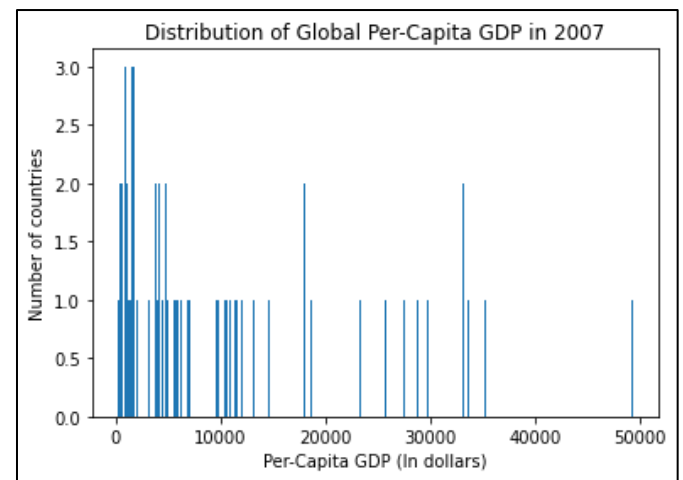
Number of bins = 5



Number of bins = 20



Number of bins = 100



Number of bins = 1000

In here Each bar in the histogram represents a bin. The height of the line represents the number of countries within the range of values spanned by the bin.

For the smaller bin sizes, it showed a somewhat smooth distribution of the per-capita GDP, which made it difficult to see any detailed patterns or features. For the large bin sizes, it provided a more detailed view of the distribution, with more distinct peaks and valleys. This histogram made it easier to see that there were a few countries with very high GDPs and many more with lower GDPs.

Therefore, when using a small number of bins, the histogram may appear too smooth and miss important features of the distribution, such as peaks and valleys. As the number of bins increases, the histogram may become more detailed and reveal more information about the distribution.

Mechanism for calculating the optimal number of bins in a histogram:

There are numerous methods that can be used. Sturges' rule, Freedman-Diaconis, and Kernel density estimation are examples.

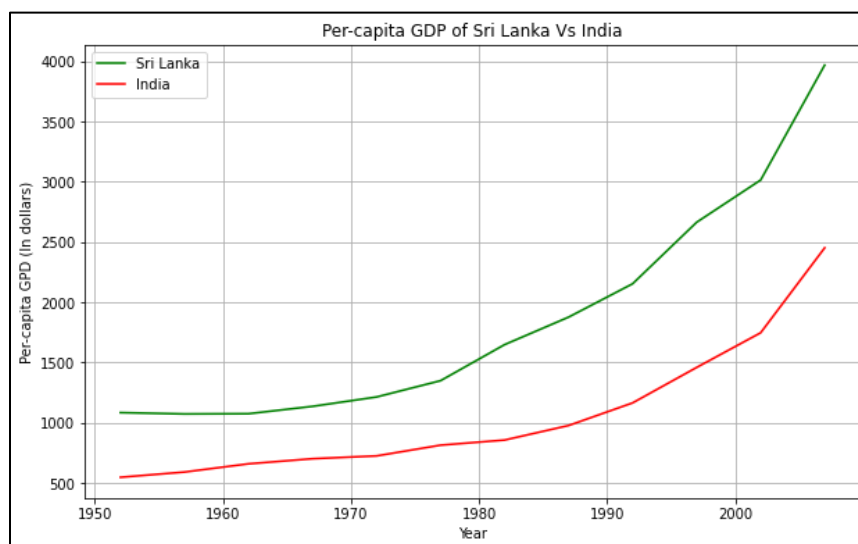
Sturges' rule: The number of bins should be approximately equal to the square root of the number of data points.

Freedman-Diaconis: this considers the data's range and distribution to determine an appropriate bin size.

Estimating kernel density: To estimate the density of the underlying distribution and generate a smooth, continuous density plot without specifying bins, use this function. This method works by estimating the probability density function (PDF) of the data using the kernel function and bandwidth that the user specifies.

Using the methods described above, we can determine the optimal number of bins based on the context and goals of the analysis.

TODO 2



This plot shows that Sri Lanka's per-capita GDP is greater than India's throughout the period. But the growth of the function/graph is similar for both countries.

TODO 3

2. **Independent** variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
Dependent variables: quality
3. (a) missing values: by checking the check the number of null values in the dataset columns wise using `df.isnull().sum()` command : (shows some columns have missing values)

```
fixed acidity      0
volatile acidity   14
citric acid        0
residual sugar     12
chlorides          0
free sulfur dioxide 0
total sulfur dioxide 0
density           0
pH                5
sulphates          0
alcohol            0
quality            0
dtype: int64
```

(b) ways to handling missing values in ML:

- Removing the rows or columns with missing values (if it's a small number).
- Imputing missing values with the mean, median or mode of the column.
- Using models that can handle missing values (e.g., decision trees, random forests)

(c)

```
# 3.c
# impute the missing values by using the mean of the column
for col in df.columns:
    if df[col].isnull().sum() > 0:
        df[col] = df[col].fillna(df[col].mean())

df.isnull().sum()

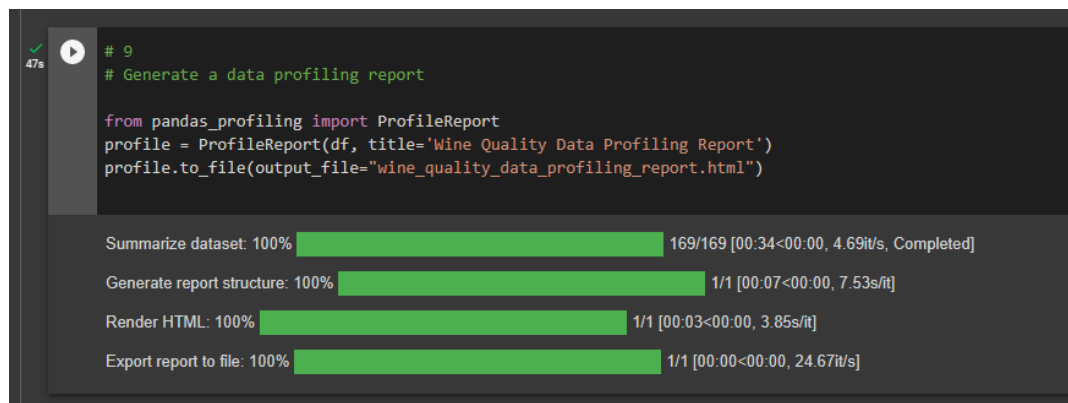
fixed acidity      0
volatile acidity    0
citric acid         0
residual sugar      0
chlorides           0
free sulfur dioxide 0
total sulfur dioxide 0
density            0
pH                 0
sulphates           0
alcohol             0
quality             0
dtype: int64
```

4. Three highly correlated attributes:
Alcohol (0.48)
Volatile acidity (-0.39)
Sulphates (0.25)
5. In the colab file
6. In the colab file

7. By looking at the boxplot, there are outliers in this dataset. (Because there are black circles outside the whiskers of the box plot)
8. (a) Data normalization/standardization is important in ML because it brings all features to a similar scale, which prevents certain features from dominating over others. So that they can be compared and analyzed more easily. This can improve the performance of many machine learning algorithms.

(b) The difference between normalization and standardization is that normalization scales the data to a range between 0 and 1, while standardization scales the data to have a mean of 0 and a standard deviation of 1.

9.



```
# 9
# Generate a data profiling report

from pandas_profiling import ProfileReport
profile = ProfileReport(df, title='Wine Quality Data Profiling Report')
profile.to_file(output_file="wine_quality_data_profiling_report.html")
```

Summarize dataset: 100% 169/169 [00:34<00:00, 4.69it/s, Completed]

Generate report structure: 100% 1/1 [00:07<00:00, 7.53s/it]

Render HTML: 100% 1/1 [00:03<00:00, 3.85s/it]

Export report to file: 100% 1/1 [00:00<00:00, 24.67it/s]

Link for the generated report (.html): (Download and see)

https://drive.google.com/file/d/1mZI2GEXoEcQmOHR84DI6Z5wnQ3q_9TgI/view?usp=share_link

Appendix:

https://colab.research.google.com/drive/1Wb3e9YOSwrkZDY50VFdbTp8qOhai91jk?usp=share_link

(Google colab)