7PAM2000 Applied Data Science 1

Assignment 1: Visualisation

**Name**: Uriti Srikanth

**Student ID**: 22028166

**Dataset:**

**Jupyter code:**

# Program Overview

# Program 1 is the Overview of the BBC News Summary dataset

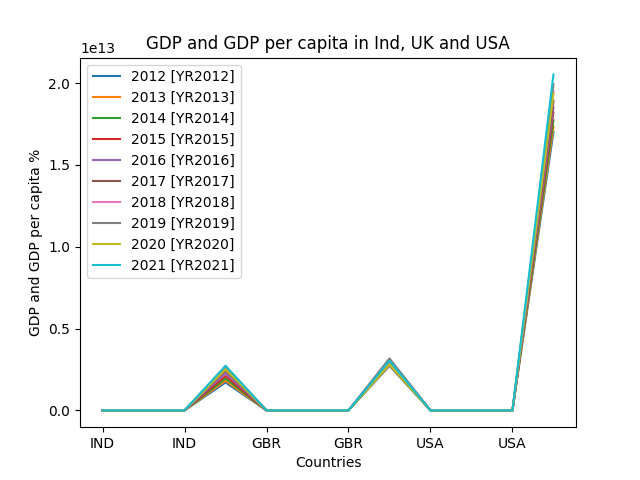
The given code loads a dataset from the BBC website using the Pandas library and creates a Pandas DataFrame object called "uk\_df". The dataset contains news articles and summaries from the BBC News website. The load\_dataset() function loads the data from the CSV file and returns the DataFrame.

Next, a new column called length is added to the DataFrame using the panda’s method apply (), which applies the len () function to the text column of the DataFrame to calculate the length of each article.

The code defines functions to create different types of plots using Matplotlib and Seaborn. The first function, line plot (), creates a line plot showing the number of items in each category.

Overall, the code provides a simple way to load a dataset, add a new column, and create some kind of visualization to explore and analyze the data.

# 1st Visualisation: GDP criterion visualization using a Line Plot.



***Figure 1: GDP criterion visualisation using a Line Plot***

A line chart displays information as a series of data points connected by straight lines. It is used to show trends or changes over time. Its advantages include simplicity and ease of interpretation. The first plot consists of a line chart in which each GDP criterion in the three countries is plotted for ten consecutive years. The graph shows that of the three countries, the US has the highest GDP and India has the lowest. In the ten years, India had the lowest GDP (constant 2015 US$) (1704595760877.17) and GDP per capita (constant 2015 US$) (1337.47) in 2012. The USA had the lowest GDP (constant 2015 US$) (17016393933000) and GDP per capita (constant 2015 US$) (54213.4595516389) in 2012 and the UK had the lowest GDP (constant 2015 US$) (2727755400457.14) and GDP per capita (constant 2015 US$) (42821.7612838063) in 2012 and 2020.

# Program 2 is the Overview of the BBC News Summary dataset

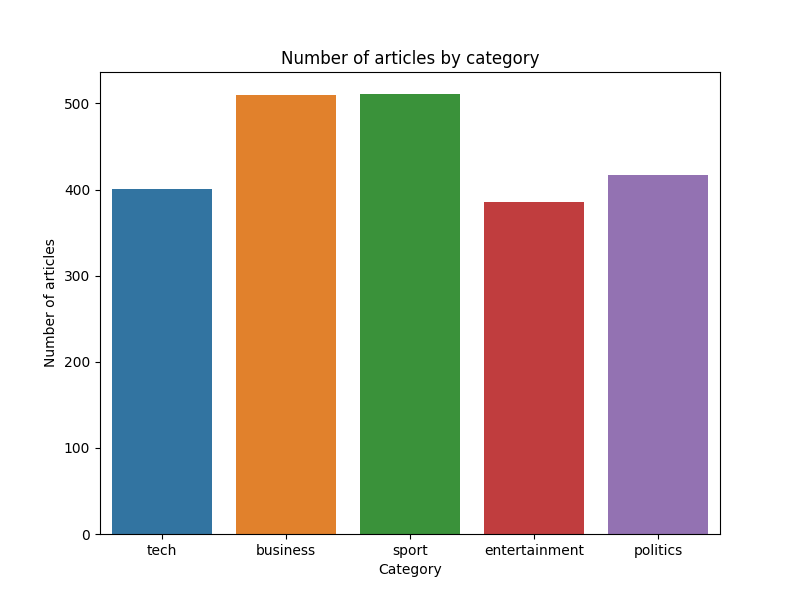
The given code loads a dataset from the BBC website using the Pandas library and creates a Pandas DataFrame object called "uk\_df". The dataset contains news articles and summaries from the BBC News website. The function load\_dataset() loads the data from the CSV file and returns the DataFrame.

Next, a new column called "length" is added to the DataFrame using the Pandas apply () method, which applies the Len() function to the "text" column of the DataFrame to calculate the length of each article.

The code defines three functions for creating different types of plots using Matplotlib and Seaborn. The second function, bar\_chart(), creates a bar chart that shows the number of articles in each category. The third function, scatter\_plot(), creates a scatter plot that shows the relationship between article length and category.

Finally, the code calls all two functions to display the plots one after the other using the show() method of the pyplot module in Matplotlib. Overall, the code provides a simple way to load a dataset, add a new column, and create two different types of visualizations to explore and analyse the data.

# 2nd Visualisation: Number of articles published in different categories using a Bar Chart.

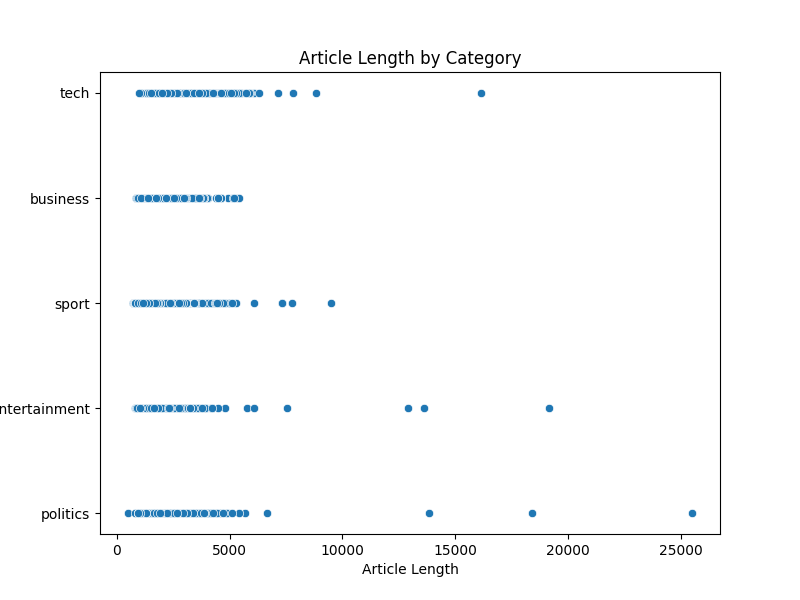


***Figure 2: Line Plot for the number of articles in different categories***

# A bar chart is a graphical representation of data that uses rectangular bars to compare values. Its advantages include easy-to-understand comparisons, clear presentation of data, and visualisation of frequency distributions.

# The visualisation shows the number of articles published by BBC News in different news categories using a pie chart. The bar chart shows that the business and sports categories published the most articles (over 500 each). The entertainment category contains the fewest articles of all five, but also has a high weight of 390 articles. The reason why the BBC has published fewer articles in the entertainment category is that the BBC prioritizes providing articles that complement the public interest rather than the public's opinion. The BBC strives to provide a balanced perspective on all news topics and covers entertainment genres based on ongoing news cycles and editorial decisions.

# 3rd Visualisation: Number of articles published in different categories using Scaler Plot.



***Figure 3: Article Length in different categories with scaler plot***

Scatter plots represent the relationship between two numerical variables. They allow the identification of patterns, trends, and correlations. The advantages of scatter plots include the identification of clusters, outliers, and linear relationships.

The figure shows that the average length of articles published by BBC News is generally around 5000 words. Most articles published by the BBC in any category are less than 5000 words. In terms of articles, most of the words published by BBC News fall into the politics category with over 25000. The BBC does this because it tries to include in-depth features, investigative reports, and journalism instead of standard news articles. This allows for a deeper dive into the subject matter, which is quite complex and requires a higher word length.

**Reference:**

**https://www.kaggle.com/datasets**

#Python script for Visualisation: GDP ::: Line\_plot

#Importing necessary Libraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the data into a DataFrame

df = pd.read\_csv("D:\\Datasets\\New folder\\P\_Data\_Extract\_From\_World\_Development\_Indicators.csv")

df

#Columns

df.columns = ["Country Name", "Country Code", "Series Name", "Series Code", "2012 [2K12]", "2013 [2K13]", "2014 [2K14]", "2015 [2K15]", "2016 [2K16]", "2017 [2K17]", "2018 [2K18]", "2019 [2K19]", "2020 [2K20]", "2021 [2K21]"]

#Line Plot

df.plot(x="Country Code", y=["2012 [2K12]", "2013 [2K13]", "2014 [2K14]","2015 [2K15]", "2016 [2K16]", "2017 [2K17]", "2018 [2K18]", "2019 [2K19]", "2020 [2K20]", "2021 [2K21]"])

plt.title("GDP and GDP per capita in Ind, UK and USA")

plt.xlabel("Countries")

plt.ylabel("GDP and GDP per capita %")

plt.show()

#Python script for Visualisation: BBC ::: Bar plot & Scatter plot

#Importing necessary Libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

#Load the dataset

#Load the dataset

def load\_dataset():

df = pd.read\_csv("D:\\Datasets\\New folder\\BBC\_UK.csv")

return df

uk\_df = load\_dataset() # Add a column for article length

uk\_df['length'] = uk\_df['text'].apply(len)

uk\_df

# Bar chart & # Scatter plot

def bar\_chart():

plt.figure(figsize=(8, 6))

sns.countplot(data=uk\_df, x='category')

plt.xlabel('Category')

plt.ylabel('Number of articles')

plt.title('Number of articles by category')

plt.show()

def scatter\_plot():

plt.figure(figsize=(8, 6))

sns.scatterplot(data=uk\_df, x='length', y='category')

plt.xlabel('Article Length')

plt.ylabel('Category')

plt.title('Article Length by Category')

plt.show()

# Call the functions

bar\_chart()

scatter\_plot()