Objective:

To design a recommendation system that can suggest personalized news articles or blog posts to users based on their reading history and interests.The system should be able to learn and adapt to the user's preferences overtime and provide relevant and engaging content.

About Dataset:

The datasets consists of variables like category, headline, authors, link, short_description, date.

Variables Description:

- category: The category of the news.
- headline: The title of news.
- authors: Author of that particular news.
- link: The web address of that particular news.
- short_desceiption: In short about the news published.
- date: Day on which the news was published.

Tasks carried out:

- Data Extraction
- Data Exploration and EDA.

1) Importing The Required Libraries:

In [139…
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import datetime
import warnings
warnings.filterwarnings("ignore")
```

2) Loading The Dataset:

In [140…
```
news_articles = pd.read_csv(r"C:\Users\sksho\Desktop\ZenteiQ\Work\Data\News category dat
```

In [141…
```
news_articles.head()
```

Out[141]:

	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69caad89	She left her husband. He killed their children. Just another day in America.	2018-05-26
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2018 World Cup's Official Song	Andy McDonald	https://www.huffingtonpost.com/entry/will-smith-joins-diplo-and-nicky-jam-for-the-official-2018-world-cup-song_us_5b09726fe4b0fdb2aa541201	Of course it has a song.	2018-05-26
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-grant-marries_us_5b09212ce4b0568a880b9a8c	The actor and his longtime girlfriend Anna Eberstein tied the knot in a civil ceremony.	2018-05-26
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And Democrats In New Artwork	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carrey-adam-schiff-democrats_us_5b0950e8e4b0fdb2aa53e675	The actor gives Dems an ass-kicking for not fighting hard enough against Donald Trump.	2018-05-26
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags To Pick Up After Her Dog	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-margulies-trump-poop-bag_us_5b093ec2e4b0fdb2aa53df70	The "Dietland" actress said using the bags is a "really cathartic, therapeutic moment."	2018-05-26

In [142…

```
news_articles.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200853 entries, 0 to 200852
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype 
---  ------             --------------   ----- 
 0   category           200853 non-null  object
 1   headline           200847 non-null  object
 2   authors            164233 non-null  object
 3   link               200853 non-null  object
 4   short_description  181141 non-null  object
 5   date               200853 non-null  object
dtypes: object(6)
memory usage: 9.2+ MB
```

To check the Number of missing values in the data and the contribution of missing values in the data.

In [143…
```
missing_values_contribution = pd.DataFrame({'No.of Missing Values': news_articles.isna()
                                             '% of Missing Values': (news_articles.isna()
```

In [144…
```
missing_values_contribution
```

Out[144]:

	No.of Missing Values	% of Missing Values
category	0	0.000000
headline	6	0.002987
authors	36620	18.232239
link	0	0.000000
short_description	19712	9.814143
date	0	0.000000

Note: There are total 200853 records in the data with 6 object type features. Also there are missing values in the data.

- We can observe that the feature authors have maximum number of missing values followed by short_description and headlines which contributes about 18.23%, 9.81%, and 0.0029% of data respectively.
- We will drop this null values as this will not contribute much in our analysis of data set.

In [145…
```
news_articles.dropna(inplace=True)
```

In [146…
```
news_articles.isna().sum()
```

Out[146]:
```
category             0
headline             0
authors              0
link                 0
short_description    0
date                 0
dtype: int64
```

In [147…
```
news_articles.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 148983 entries, 0 to 200848
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype 
---  ------             --------------   ----- 
 0   category           148983 non-null  object
 1   headline           148983 non-null  object
 2   authors            148983 non-null  object
 3   link               148983 non-null  object
 4   short_description  148983 non-null  object
 5   date               148983 non-null  object
dtypes: object(6)
memory usage: 8.0+ MB
```

Note: we can observe that all null values have be dropped and we are left with 148983 non null values.

In [148…
```
news_articles.shape
```

Out[148]:
```
(148983, 6)
```

Note: From the dataset we can note that the columns category, authors, and short description can have same values. hence we have to note that if there are any duplicate values in the headlines column. and eliminate them.

In [149…
```
news_articles['headline'].count()
```

Out[149]:
```
148983
```

In [150…
```
news_articles
```

Out[150]:

	category	headline	authors	link	short_
0	CRIME	There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69caad89	husbar their c an
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2018 World Cup's Official Song	Andy McDonald	https://www.huffingtonpost.com/entry/will-smith-joins-diplo-and-nicky-jam-for-the-official-2018-world-cup-song_us_5b09726fe4b0fdb2aa541201	Of cou
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-grant-marries_us_5b09212ce4b0568a880b9a8c	The a longti Ann tied t civ
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And Democrats In New Artwork	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carrey-adam-schiff-democrats_us_5b0950e8e4b0fdb2aa53e675	The D ki f enc Do
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags To Pick Up After Her Dog	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-margulies-trump-poop-bag_us_5b093ec2e4b0fdb2aa53df70	Th actres t "rea
...	...	...	...	...	
200843	TECH	Good Games -- Is It possible?	Mateo Gutierrez, Contributor\nArtist	https://www.huffingtonpost.com/entry/games-for-change_us_5bb34b30e4b0fa920b95bab3	I don't who gam Jus think t in V They
200844	TECH	Google+ Now Open for Teens With Some Safeguards	Larry Magid, Contributor\nTechnology journalist	https://www.huffingtonpost.com/entry/google-plus-safety_us_5bb34b7de4b0fa920b95c2d7	For th teens on Goo just lik the special for use Goog any m or freedo default
200845	TECH	Web Wars	John Giacobbi, Contributor\nTales from	https://www.huffingtonpost.com/entry/congress-sopa_us_5bb34b8be4b0fa920b95c47f	These thre

	category	headline	authors	link	short_
			the Interweb by The Web Sheriff		consi yet -- whe tries to SOPA door even on
200847	TECH	Watch The Top 9 YouTube Videos Of The Week	Catharine Smith	https://www.huffingtonpost.com/entry/watch-top-youtube-videos_us_5bb34b88e4b0fa920b95c42a	If you' s popu v w fu bring
200848	TECH	RIM CEO Thorsten Heins' 'Significant' Plans For BlackBerry	Reuters, Reuters	https://www.huffingtonpost.com/entry/rim-ceo-thorsten-heins_us_5bb34b8ce4b0fa920b95c4e1	Veriz ar alread smart tablets rivals

148983 rows × 6 columns

In [151…
```
news_articles.sort_values('headline',inplace=True, ascending=False)
duplicated_articles_series = news_articles.duplicated('headline', keep = False)
news_articles = news_articles[~duplicated_articles_series]
print("Total number of articles after removing duplicates:", news_articles['headline'].n
```

```
Total number of articles after removing duplicates: 147706
```

To see the total number of unique records.

In [152…
```
news_articles.nunique()
```

Out[152]:
```
category              41
headline              147706
authors               27055
link                  147706
short_description     146102
date                  2309
dtype: int64
```

3) Exploratory Data Analysis:

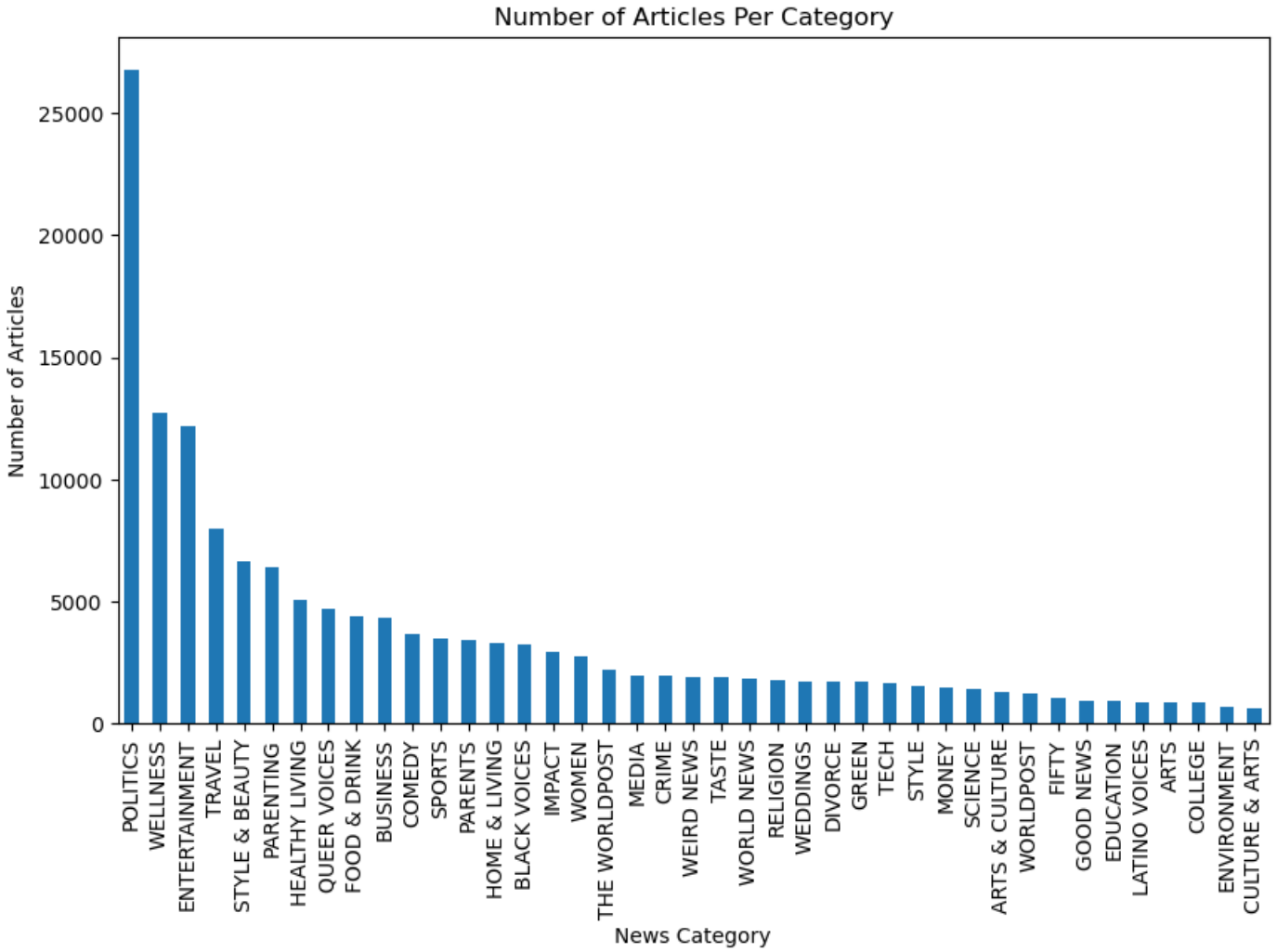
In [153…
```
news_articles["category"].value_counts().head()
```

Out[153]:
```
POLITICS          26779
WELLNESS          12745
ENTERTAINMENT     12171
TRAVEL            7992
STYLE & BEAUTY    6635
Name: category, dtype: int64
```

In [154…
```
news_articles["category"].value_counts().tail()
```

Out[154]:
```
LATINO VOICES     877
ARTS              861
COLLEGE           859
ENVIRONMENT       669
CULTURE & ARTS    611
Name: category, dtype: int64
```

In [155…
```
plt.figure(figsize=(10,6))
news_articles["category"].value_counts().plot.bar(xlabel = 'News Category',
                                    ylabel = 'Number of Articles',
                                    title = 'Number of Articles Per Category')
plt.show()
```



Note:

- We can observe that Top 5 articles that are watched are based on Politics, Wellness, Entertainment, Travel, Style & Beauty.
- We can observe that Bottom 5 articles that are watched are based on Culture & Arts, Environment, College, Arts,Latino Voices.

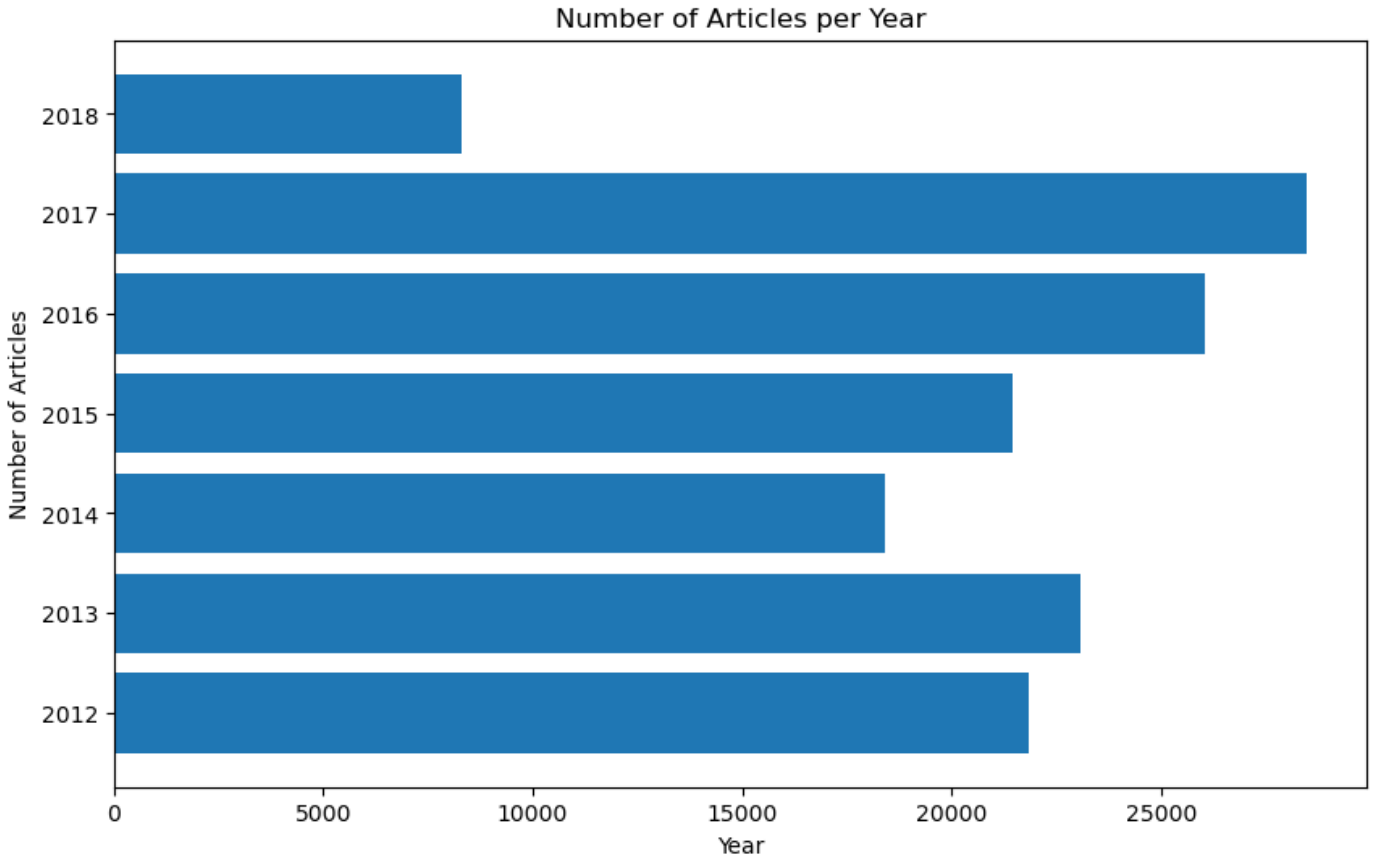
In [156…
```
news_articles['date'] = pd.to_datetime(news_articles['date'])

# Resample the DataFrame by Year and count the number of articles
news_articles_monthly = news_articles.resample('Y', on='date')['headline'].count()

# Create a bar plot of the monthly article counts
plt.figure(figsize = (10,6))
plt.barh(news_articles_monthly.index.strftime('%Y'), news_articles_monthly)
```

```
# Set the plot title and axis labels
plt.title('Number of Articles per Year')
plt.xlabel('Year')
plt.ylabel('Number of Articles')

# Display the plot
plt.show()
```



Note: From above ditribution we can observe the total number of articles on yearly basis.

In [ ]:

In [ ]: