A MINI PROJECT REPORT OF BIG DATA ANALYTICS
ON
“Alcohol and Happiness”

Submitted in partial fulfilment of the

BE in Computer Engineering

(Semester-VII)

By

Saurabh Rajendra Jadhav	**24**
Sanket Chandrashekhar Harvande	**19**
Shreya Chandrakant Malavade	**35**

Department of Computer Engineering

GHARDA FOUNDATION

GHARDA INSTITUTE OF TECHNOLOGY, LAVEL

2022-2023

CERTIFICATE

This is to certify that the Machine Learning Mini Project Report Entitled

“Alcohol and Happiness”

Submitted by

Sanket Chandrashekhar Harvande	**19**
Saurabh Rajendra Jadhav	**24**
Shreya Chandrakant Malavade	**35**

is a record of bonafide work carried out by them, under our guidance, in partial fulfilment of the requirement for Bachelors of Engineering (Computer Engineering) Sem-VII at GIT, Lavel under the University of Mumbai. This work is done during July 2022- October 2022 of Academic year 2022-23.

Date:

Place: GIT, Lavel

Prof. D. N. Londhe	**Prof. R. R. Bane**	**Dr. S. K. Patil**
(Project Guide) Dept.CE,	(HOD) Dept.CE, GIT	(Principal) GIT L

CONTENTS

INTRODUCTION

The workplace provides several opportunities for implementing prevention strategies to reduce the harm done by alcohol, since the majority of adults are employed and spend a significant proportion of their time at work. The workplace can also be a risk factor for harmful alcohol use. Many studies have found significant associations between stress in the workplace and elevated levels of alcohol consumption, an increased risk of problem drinking and alcohol dependence. Evidence has found that alcohol, and in particular heavy drinking, increases the risk of unemployment and, for those in work, absenteeism. Alcohol, especially episodic heavy drinking, has also been found to increase the risk of arriving late at work and leaving early or disciplinary suspension, resulting in loss of productivity; a higher turnover due to premature death; disciplinary problems or low productivity from the use of alcohol; inappropriate behaviour (such as behaviour resulting in disciplinary procedures); theft and other crime; poor co-worker relations and low company morale. Studies suggest that alcohol consumption may have more effect on productivity on the job than on the number of workdays missed. Overall, the costs of lost productivity feature as the dominant element in studies of the social costs arising from the harm done by alcohol, being about half of the total social cost of alcohol in the EU. Despite the evidence of the negative impact of alcohol on the workplace, there are surprisingly few good-quality scientific studies to inform policy and practice, and of those that have been undertaken, it is not always possible to convincingly conclude the best approaches. Increasingly, and as an alternative, evidence suggests that prevention activities at the workplace to reduce the harm done by alcohol should be embedded in broader workplace health promotion and wellbeing at work initiatives.

LITERATURE SURVEY

Based on the data of Chinese Family Panel Studies (CFPS), the research uses an ordered probit model to test whether mobile Internet is positively related with residents' happiness. And it finds that mobile Internet has a significant positive impact on residents' happiness. The heterogeneity analysis shows that residents in rural areas with an undergraduate degree are more likely to obtain happiness from mobile Internet. And the further study denotes that mobile Internet significantly reduces the contribution of income to residents' perception of happiness.[1]

The daily human action understanding system based on HAPPINESS factors is presented for happiness promotion using human action recognition. Recognition and understanding of human behaviour is a popular research topic in computer vision. The proposed system extracts the features of skeleton information using the entire body and arm movement. There are two contributions in this work, 1) A suitable template selection among different subjects regarding as a representative human HAPPINESS action. And Non-Static Sequence Segmentation is proposed for action recognition. The system classifies the corresponding HAPPINESS factors from the results of action recognition. Experiments were performed on an online test and the results show that the accuracy is 84.81%.[2]

The study aims to combine the theories in Happiness Informatics with advanced trends of social science, in order to interpret and define the meaning of happiness. Functional Magnetic Resonance Imaging (fMRI) measures blood oxygenation level-dependent (BOLD) signals to infer the association between brain functions and neural activities. Experiment results are obtained from static activation maps generated. We conclude from this work that the neural mechanisms of happiness do exist.[3]

People are gradually accustomed to sharing everything in their lives and managing interpersonal relationships on the social network.The main purpose of this research is to take users of social networks as the research object, and to explore the behaviour intentions of the sharers and viewers will be affected by each person's different real feelings, and feel a sense of happiness. Therefore, this study used a questionnaire survey to systematically sort and analyse all the collected data. The results pointed out that the higher the degree of social Network-Instagram users through self-disclosure or emotional sharing, the more happiness they can enhance through social presence.[4]

PROBLEM STATEMENT AND OBJECTIVE

Problem Statement :

To identify the happiness level of the particular region based on the dataset

Objective :

Each region has different levels of alcohol intake capacity and their percentages. The dataset contains percentages of different levels of alcohol intakes and we have to identify the happiness levels based on their intakes and calculate the accuracy of their levels of happiness.

DATA SET DETAILS

Total alcohol per capita consumption is defined as the total (sum of recorded and unrecorded alcohol) amount of alcohol consumed per person (15 years of age or older) over a calendar year, in litres of pure alcohol, adjusted for tourist consumption.

Statistical concept and methodology: The estimates for the total alcohol consumption are produced by summing up the 3-year average per capita (15+) recorded alcohol consumption and an estimate of per capita (15+) unrecorded alcohol consumption for a calendar year. Tourist consumption takes into account tourists visiting the country and inhabitants visiting other countries.

Variable time span 2000 – 2018

Link: https://ourworldindata.org/alcohol-consumption

ALGORITHMS AND LIBRARIES USED

Clustering Algorithm :

Clustering or cluster analysis is an unsupervised learning problem. It is often used as a data analysis technique for discovering interesting patterns in data, such as groups of customers based on their behaviour. There are many clustering algorithms to choose from and no single best clustering algorithm for all cases.

There are different types of clustering algorithms used in this mini project.

- K-means Clustering Algorithm :
  K-Means Clustering is the most widely known clustering algorithm and involves assigning examples to clusters in an effort to minimise the variance within each cluster.

- BIRCH :
  BIRCH Clustering (BIRCH is short for Balanced Iterative Reducing and Clustering using
  Hierarchies) involves constructing a tree structure from which cluster centroids are extracted.

- DBSCAN :
  DBSCAN Clustering (where DBSCAN is short for Density-Based Spatial Clustering of Applications with Noise) involves finding high-density areas in the domain and expanding those areas of the feature space around them as clusters.

- Mini-Batch K-Means :
  Mini-Batch K-Means is a modified version of k-means that makes updates to the cluster centroids using mini-batches of samples rather than the entire dataset, which can make it faster for large datasets, and perhaps more robust to statistical noise.

Libraries Used :

- Numpy -
  NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning.
- Pandas -
  Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and a wide variety of tools for data analysis. It provides many inbuilt methods for grouping, combining and filtering data.

- Scikit-learn (sklearn) -

  Scikit-learn is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great tool for those starting out with ML.

- Matplotlib -

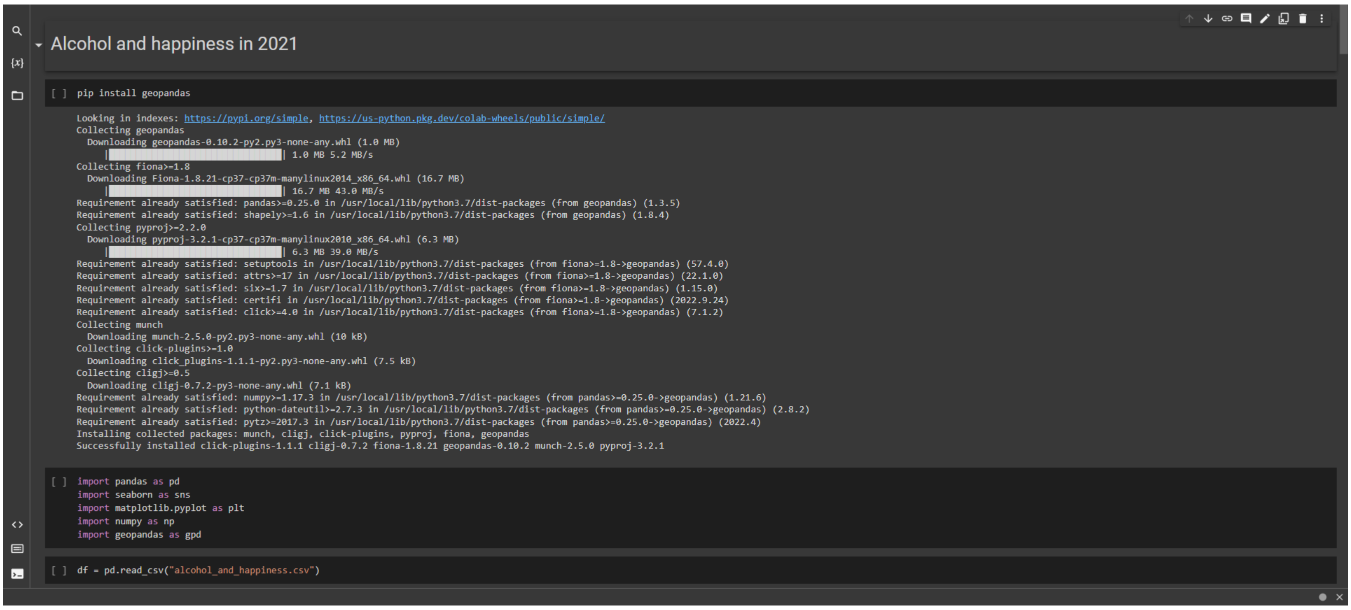
  Matplotlib is a very popular Python library for data visualisation. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualise the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualisation, viz., histogram, error charts, bar charts, etc .

- PyMongo -

  PyMongo is a Python distribution containing tools for working with MongoDB, and is the recommended way to work with MongoDB from Python.

CODE, RESULTS, OUTPUTS

```
package com.saurabhjadhav.miniprojects.BDA;

import androidx.appcompat.app.AppCompatActivity;

import android.annotation.SuppressLint;
import android.os.Bundle;
import android.webkit.WebView;
import android.webkit.WebViewClient;

import com.saurabhjadhav.miniprojects.R;

public class BdaProject extends AppCompatActivity {

   WebView webViewOfficial;

   @SuppressLint("SetJavaScriptEnabled")
   @Override
   protected void onCreate(Bundle savedInstanceState) {
      super.onCreate(savedInstanceState);
      setContentView(R.layout.activity_bda_project);
      webViewOfficial = findViewById(R.id.WebViewOfficial);

      webViewOfficial.setWebViewClient(new WebViewClient());
      webViewOfficial.getSettings().setJavaScriptEnabled(true);

webViewOfficial.loadUrl("https://colab.research.google.com/drive/1jMxrcwp0x3uCtB9_-_
VYRVJDazkNj662#scrollTo=CplXLsDz9IUo");
   }
}
```



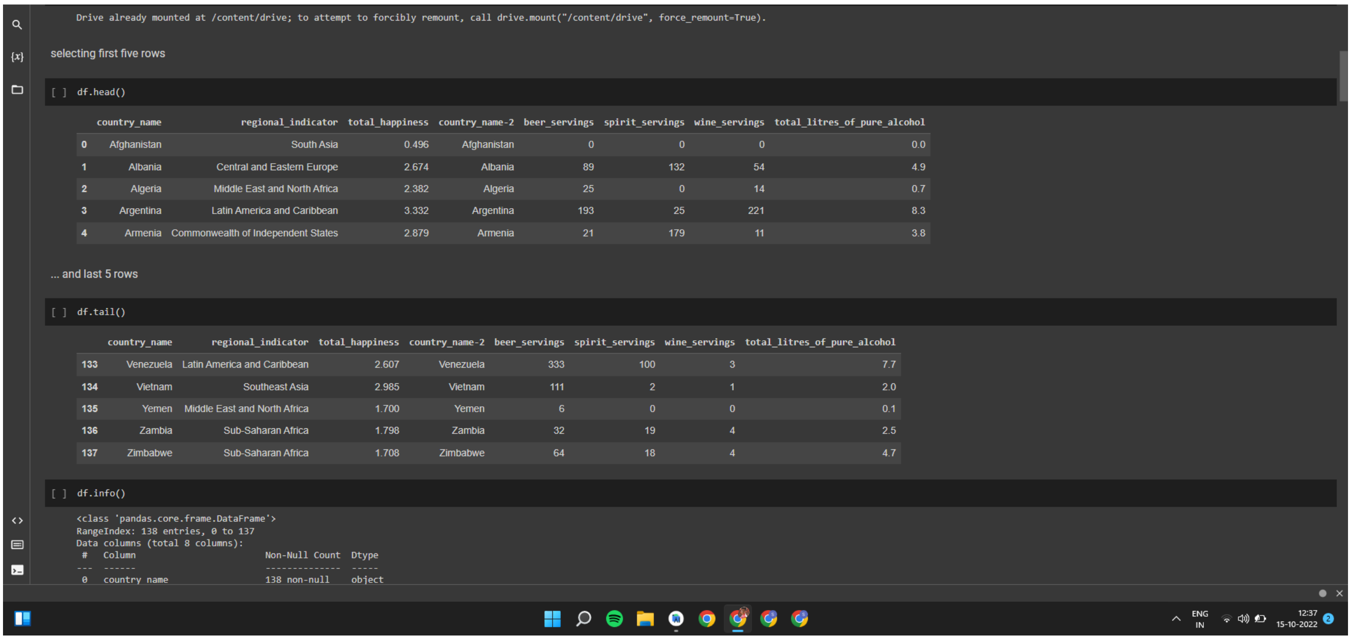
Alcohol and happiness in 2021

```
[ ] pip install geopandas
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting geopandas
  Downloading geopandas-0.10.2-py2.py3-none-any.whl (1.0 MB)
     |████████████████████████████████| 1.0 MB 5.2 MB/s
Collecting fiona>=1.8
  Downloading Fiona-1.8.21-cp37-cp37m-manylinux2014_x86_64.whl (16.7 MB)
     |████████████████████████████████| 16.7 MB 43.0 MB/s
Requirement already satisfied: pandas>=0.25.0 in /usr/local/lib/python3.7/dist-packages (from geopandas) (1.3.5)
Requirement already satisfied: shapely>=1.6 in /usr/local/lib/python3.7/dist-packages (from geopandas) (1.8.4)
Collecting pyproj>=2.2.0
  Downloading pyproj-3.2.1-cp37-cp37m-manylinux2010_x86_64.whl (6.3 MB)
     |████████████████████████████████| 6.3 MB 39.0 MB/s
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from fiona>=1.8->geopandas) (57.4.0)
Requirement already satisfied: attrs>=17 in /usr/local/lib/python3.7/dist-packages (from fiona>=1.8->geopandas) (22.1.0)
Requirement already satisfied: six>=1.7 in /usr/local/lib/python3.7/dist-packages (from fiona>=1.8->geopandas) (1.15.0)
Requirement already satisfied: certifi in /usr/local/lib/python3.7/dist-packages (from fiona>=1.8->geopandas) (2022.9.24)
Requirement already satisfied: click>=4.0 in /usr/local/lib/python3.7/dist-packages (from fiona>=1.8->geopandas) (7.1.2)
Collecting munch
  Downloading munch-2.5.0-py2.py3-none-any.whl (10 kB)
Collecting click-plugins>=1.0
  Downloading click_plugins-1.1.1-py2.py3-none-any.whl (7.5 kB)
Collecting cligj>=0.5
  Downloading cligj-0.7.2-py3-none-any.whl (7.1 kB)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.25.0->geopandas) (1.21.6)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.25.0->geopandas) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.25.0->geopandas) (2022.4)
Installing collected packages: munch, cligj, click-plugins, pyproj, fiona, geopandas
Successfully installed click-plugins-1.1.1 cligj-0.7.2 fiona-1.8.21 geopandas-0.10.2 munch-2.5.0 pyproj-3.2.1
```

```
[ ] import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    import numpy as np
    import geopandas as gpd
```

```
[ ] df = pd.read_csv("alcohol_and_happiness.csv")
```



```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

selecting first five rows

```
[ ] df.head()
```

	country_name	regional_indicator	total_happiness	country_name-2	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_alcohol
0	Afghanistan	South Asia	0.496	Afghanistan	0	0	0	0.0
1	Albania	Central and Eastern Europe	2.674	Albania	89	132	54	4.9
2	Algeria	Middle East and North Africa	2.382	Algeria	25	0	14	0.7
3	Argentina	Latin America and Caribbean	3.332	Argentina	193	25	221	8.3
4	Armenia	Commonwealth of Independent States	2.879	Armenia	21	179	11	3.8

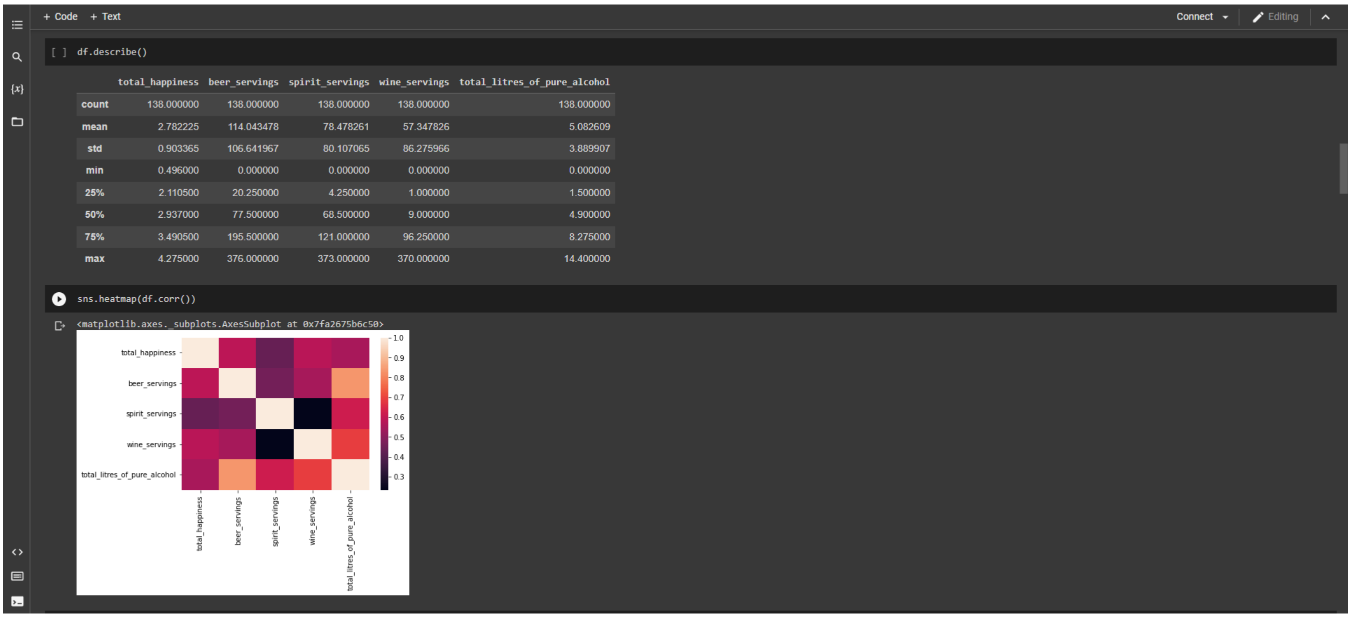
... and last 5 rows

```
[ ] df.tail()
```

	country_name	regional_indicator	total_happiness	country_name-2	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_alcohol
133	Venezuela	Latin America and Caribbean	2.607	Venezuela	333	100	3	7.7
134	Vietnam	Southeast Asia	2.985	Vietnam	111	2	1	2.0
135	Yemen	Middle East and North Africa	1.700	Yemen	6	0	0	0.1
136	Zambia	Sub-Saharan Africa	1.798	Zambia	32	19	4	2.5
137	Zimbabwe	Sub-Saharan Africa	1.708	Zimbabwe	64	18	4	4.7

```
[ ] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 138 entries, 0 to 137
Data columns (total 8 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   country_name                  138 non-null    object
```

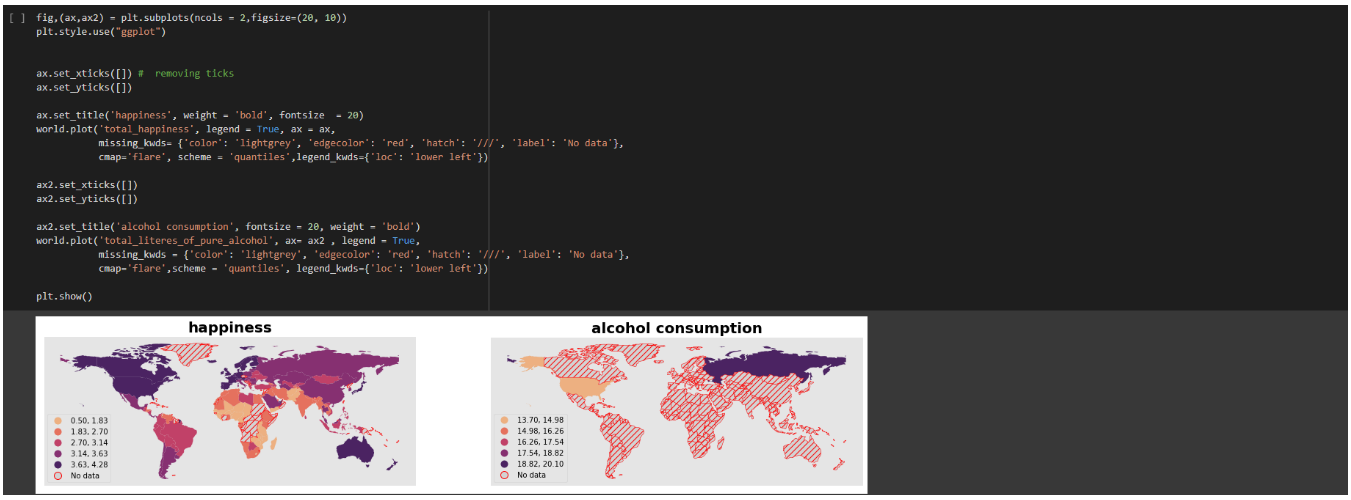


```
[ ] df.describe()
```

	total_happiness	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_alcohol
count	138.000000	138.000000	138.000000	138.000000	138.000000
mean	2.782225	114.043478	78.478261	57.347826	5.082609
std	0.903365	106.641967	80.107065	86.275966	3.889907
min	0.496000	0.000000	0.000000	0.000000	0.000000
25%	2.110500	20.250000	4.250000	1.000000	1.500000
50%	2.937000	77.500000	68.500000	9.000000	4.900000
75%	3.490500	195.500000	121.000000	96.250000	8.275000
max	4.275000	376.000000	373.000000	370.000000	14.400000

```
sns.heatmap(df.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa2675b6c50>
```



```
[ ] fig,(ax,ax2) = plt.subplots(ncols = 2,figsize=(20, 10))
    plt.style.use("ggplot")


    ax.set_xticks([]) #  removing ticks
    ax.set_yticks([])

    ax.set_title('happiness', weight = 'bold', fontsize  = 20)
    world.plot('total_happiness', legend = True, ax = ax,
              missing_kwds= {'color': 'lightgrey', 'edgecolor': 'red', 'hatch': '///', 'label': 'No data'},
              cmap='flare', scheme = 'quantiles',legend_kwds={'loc': 'lower left'})

    ax2.set_xticks([])
    ax2.set_yticks([])

    ax2.set_title('alcohol consumption', fontsize = 20, weight = 'bold')
    world.plot('total_literes_of_pure_alcohol', ax= ax2 , legend = True,
              missing_kwds = {'color': 'lightgrey', 'edgecolor': 'red', 'hatch': '///', 'label': 'No data'},
              cmap='flare',scheme = 'quantiles', legend_kwds={'loc': 'lower left'})

    plt.show()
```

CONCLUSION

As more and more data is generated and collected, data analysis requires scalable, flexible, and high performing tools to provide insights in a timely fashion. Thus we have implemented the analysis of the dataset which contains the happiness percentages of the regions and their intake capacities.

REFERENCES

[1] B. Xiao, "Mobile Internet and Residents' Happiness-Quantitative Analysis Based on CFPS," 2021 2nd International Conference on Big Data Economy and Information Management (BDEIM), 2021, pp. 93-96, doi: 10.1109/BDEIM55082.2021.00027.

[2*] Y. -Y. Ou, A. -C. Tsai, T. -W. Kuan, J. -F. Wang and J. -H. Tian, "Happiness understanding system based on human action recognition," 2016 International Conference on Orange Technologies(ICOT), 2016, pp. 56-59, doi: 10.1109/ICOT.2016.8278978.*

[3] Y. -Y. Ou, D. -R. Yeh, C. -C. Kung, P. -C. Lin, T. -W. Kuan and J. -F. Wang, "Design and implementation of happiness database for fMRI study," 2015 International Conference on Orange Technologies (ICOT), 2015, pp. 127-130, doi: 10.1109/ICOT.2015.7498493.

[4] D. -Y. Liu, C. -T. Lin, K. -C. Wang and M. -Y. Chen, "Real and Virtual Happiness Prediction Model Based on Multiple Regression analysis by Instagram Social Network," 2021 IEEE International Conference on Social Sciences and Intelligent Management (SSIM), 2021, pp. 1-8, doi: 10.1109/SSIM49526.2021.9555216.