Sinéad Dunne

s\_dunne1@hotmail.com

UCD Certificate in Introductory Data Analytics

A picture containing clipart

Description automatically generated

global consumption

for wine, beer & spirits

Project Report

# GitHub URL: <https://github.com/sdunne1/UCDPA_SineadDunne>

# Abstract

My data analysis project investigates data figures published in 2010 by the Global Information System on Alcohol and Health (WHO GISAH), and manipulated by Mona Chalabi (data editor at Guardian U.S.), as part of her 2014 article for website FiveThirtyEight: “Where Do People Drink the Most Beer, Wine and Spirits?”

WHO originally published this data, in total litres of pure alcohol consumed per person in each country. However, for the purpose of this article, the average alcohol content and average serving size for each beverage was taken and those numbers converted into standard serving sizes. The data produced shows units of wine by glass, units of beer by cans and units of spirits by shots, all consumed per person in each listed country. This data was retrieved from Kaggle as referenced below and is the dataset I have used for my data analysis report.

# Introduction

Ireland as a small country on the outskirts of Europe is widely renowned as being the top country in the world, if not one of them, for the highest level of alcohol consumption. I wanted to explore alcohol stereotypes, using real data figures published. The data description is outlined in the Abstract of this report.

There are some limitations with this dataset that should be highlighted. As it was published over ten years ago, it can give us good insights. However, they will not be the most up to date. The dataset only deals with three alcohol types: Wine, Beer and Spirits. The dataset only contains 193 countries of the total 195 in the world. For some countries listed the serving size figures are set to 0. These are mostly for countries with strong religious influences. However, I believe there to be an inaccuracy for these rows, as there may still be a small minority in these countries who consume alcohol. In which case, these figures will not be represented.

With more time, I would have liked to extend my project to explore the latest figures available from Global Information System on Alcohol and Health (GISAH). This would allow me to demonstrate more current and up to date global insights. It would also facilitate data visualisations of a ten-year comparison. Demonstrating increases and reductions in total litres of alcohol, and the types of alcohol by country / continent, linking in possible trends that might explain them. For example, campaigns to increase sales for alcohol types or brands.

# Dataset

The data is a csv file which contains 5 column headers: Country, Beer\_Servings, Wine\_Servings, Spirit\_Servings, total\_litres\_of\_pure\_alcohol, and 193 rows of data. The source of the data is [www.kaggle.com](http://www.kaggle.com), and the target URL is listed in ‘References’ of this report.

I chose this dataset as it was in line with the area that I wanted to review. It was also a good size for me to work with, being new to data analytics. The dataset contained several data types - categorical and numerical, which I could use to make good data visualisations. Further, I was able to build on the data provided by creating my own dataset to include ‘Continent’. This was done by merging an additional dataframe and allowed me to aggregate the data by continent.

# Implementation Process

For the implementation of my data analysis, I am using the Pycharm IDE and GitHub for my version control repository. I used Anaconda Jupyter notebook, to assist with checking my code syntax and visualisations.

I started by importing and cleaning my dataset. I imported the Kaggle API package to assist with downloading the file directly, as opposed to manually downloading the file. This method requires the use of a json file to be stored in a local ‘.kaggle’ folder. It contains an API token associated with my Kaggle account. Once this was set up, I was able to call the authenticate function to verify my credentials. The package also allowed me to call the ‘dataset\_download\_file’ function, to retrieve the data file by passing the target user and dataset name I require. To read in the csv file, I used the pandas ‘read\_csv’ function, and I also renamed my columns headers at this point. This was to assist with making the column names easier to work with later.

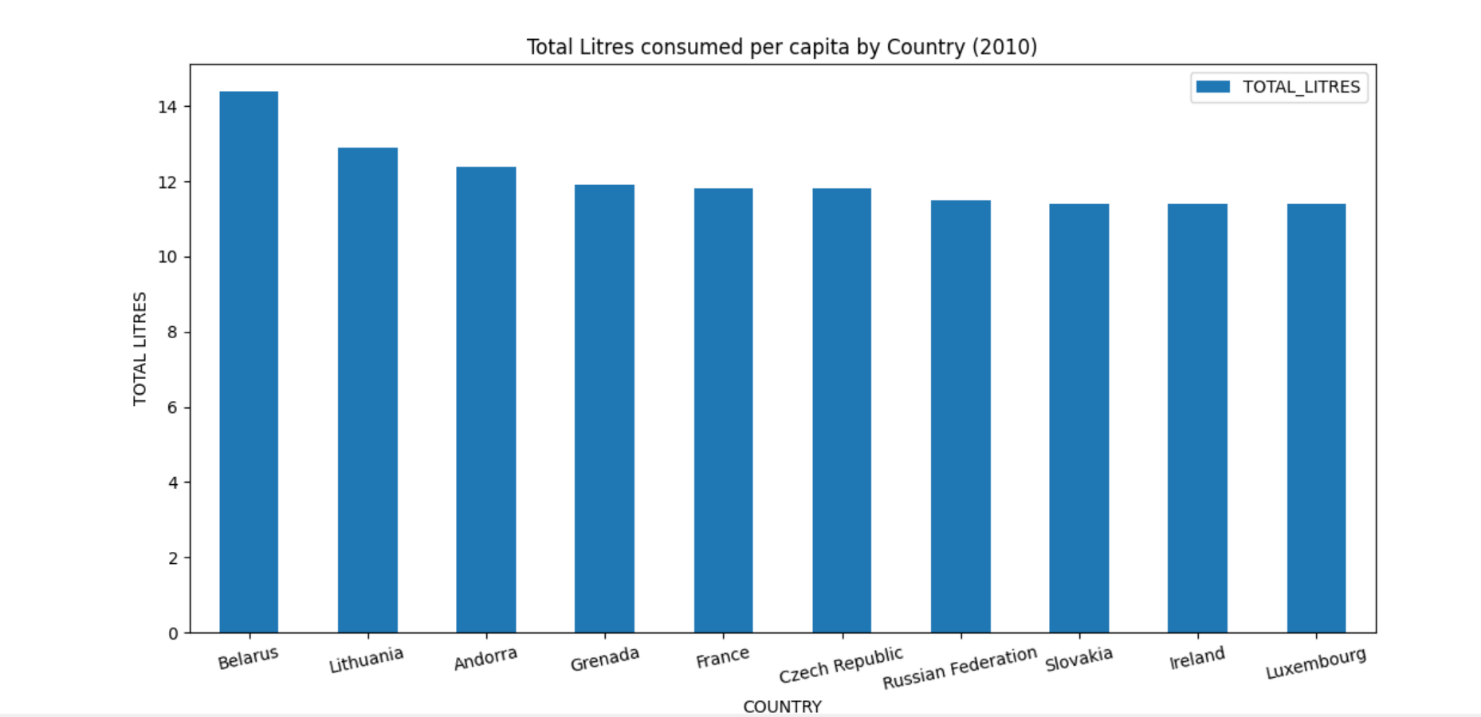
I completed several checks to get an initial understanding of the dataset:

Checking for null values, checking for duplication etc. As I knew the dataset usability was scored high from the Kaggle information page, I manually removed zeros from the csv file. This allowed me to demonstrate the use of numpy replace function to replace Nan values with zero again.

Initially, I also wanted to see where Ireland factored in the order of countries for highest alcohol volumes. I sorted the dataframe by ‘TOTAL\_LITRES’ and set the index as the ‘COUNTRY’ column. I thought it would be interesting to add another dimension to my dataset, by adding the column ‘CONTINENT’. I created a csv file using the country list from ‘drinks.csv’ and added the continent name for each. I read the file in as csv and called this dataframe ‘df\_drinks\_con’. Next, I merged my two dataframes using the merge function, joining on the country column. I wanted to keep code for cleaning and manipulating the dataset in one file. I downloaded the merged dataframe as a .csv file so that I could re-import into a new .py file, specifically for data visualisations.

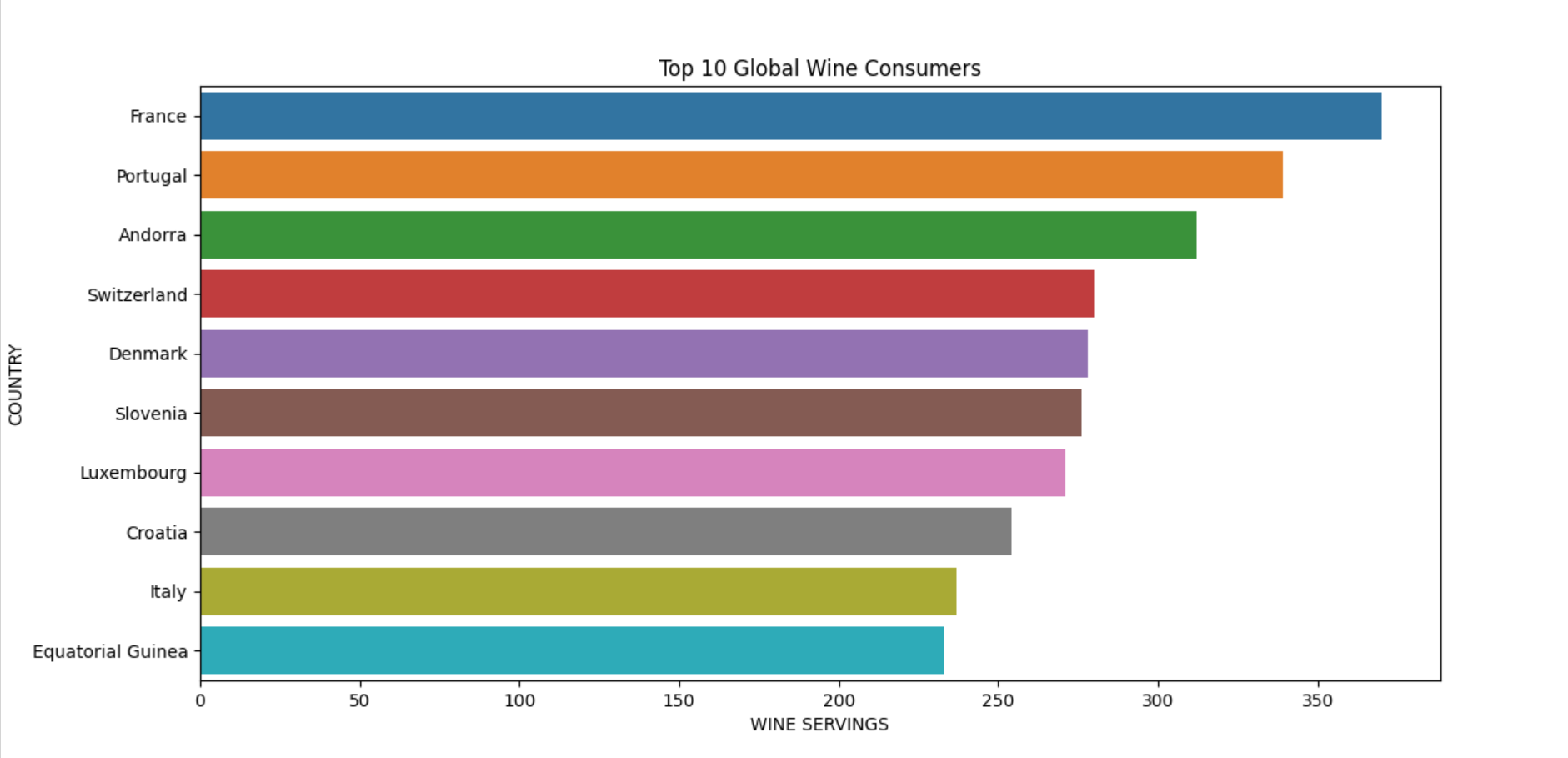
In drinks\_viz.py, I imported my downloaded dataframe and the necessary import packages I needed, to create the data visualisation: Pandas, matplotlib and Seaborn.

# Results

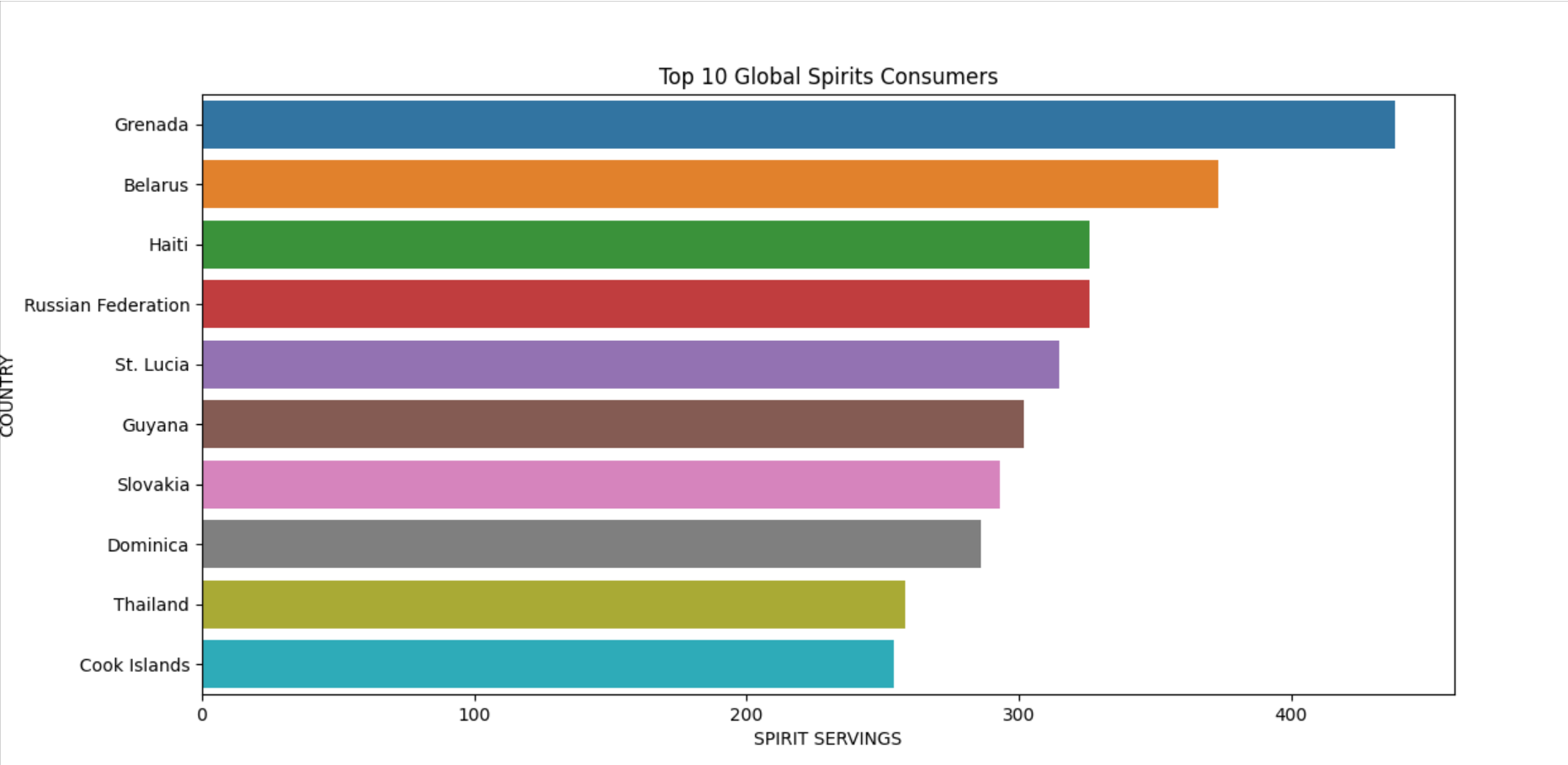
Figure 1.0 is my first visualisation which showcases per capita, the top ten countries in the world who consume the most alcohol. To create this visualisation, I selected the ‘Country’ and ‘Total Litres’ columns from my dataframe and grouped by country. I summed the ‘Total Litres’ and used ‘.head(10)’ to get the first ten rows. This is a bar chart using ‘Matplotlib’.

**Figure 1.0**

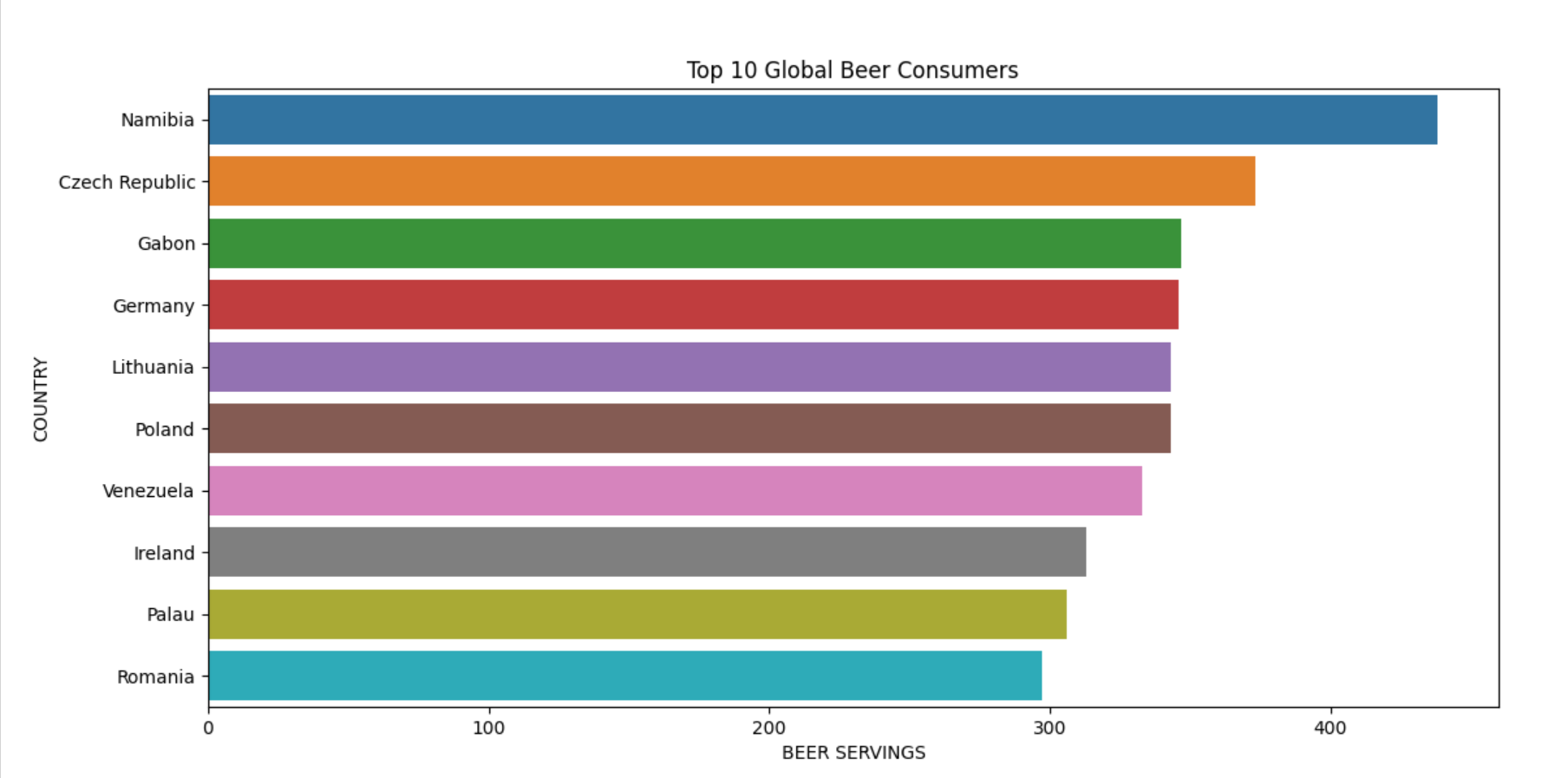
Figures 1.1 to Figure 1.3 are visualisation that showcase the top ten consumers of each alcohol type consumed. To create these visualisations, I selected the ‘Country’ and alcohol type columns, i.e. Wine, Spirit, Beer - from my dataframe and grouped by country. I summed each alcohol column for the individual charts and used ‘.head(10)’ or sliced dataset [:10] to get the first ten rows. The figures are horizontal bar charts using ‘Seaborn’.



**Figure 1.1 –** *Shows wine serving units by the glass****.***



**Figure 1.2 –** *Shows spirit serving units by a standard measure****.***



**Figure 1.3** – *Shows beer serving units by can.*

# Insights

From my analysis, I can confirm that global alcohol stereotypes are inconsistent.

1. The first point I wanted to investigate, ‘Is Ireland the biggest consumers of alcohol in the world?’. Many of the world believe our nation to be. However, in ‘Results – Figure 1.0’, we can see that Ireland places in ninth position on the top ten countries bar chart. What is particularly interesting with this chart is that countries such as Andorra and Grenada, have very small populations in comparison to Ireland and yet they have placed inside the top five.
2. For spirit drinkers, there is a wide presumption that eastern Europe countries are the biggest drinkers of spirits as there is mass production of vodka and other spirits in this region. While there are two countries from this region on the chart that fall within the top five, in ‘Results – Figure 1.2 we can see the first position is Grenada, located in the Caribbean.
3. Some stereotypes can be quite accurate. For example, majority of people would believe France to be the biggest wine drinkers in the world. Mainly because the country is one of the biggest producers in the world. However, there population is smaller than other countries who would also be mass producers, like Australia or Chile. In ‘Results - Figure 1.1’ we can see that for Wine, France are the biggest consumers.
4. preference
5. Type of alcohol mostly consumed by highest total litres of alcohol.

Convert csv to list and update code.

# References

|  |  |
| --- | --- |
| **KAGGLE** | <https://www.kaggle.com/codebreaker619/alcohol-comsumption-around-the-world> |
| **WORLD HEALTH ORGANISATION** | https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-recorded-per-capita-(15-)-consumption-(in-litres-of-pure-alcohol) |
| **DATACAMP** | <https://learn.datacamp.com/custom-tracks/custom-certificate-in-introductory-data-analytics-6859d143-be7b-49e0-ad09-007f207ad33c> |
| **FIVETHIRTYEIGHT** | https://fivethirtyeight.com/features/dear-mona-followup-where-do-people-drink-the-most-beer-wine-and-spirits/ |
| **PYTHON DOCS** | <https://docs.python.org/3/> |
| **PANDAS DOCS** | <https://pandas.pydata.org/docs/> |
| **SHANELYNN.IE** | https://www.shanelynn.ie |
| **STACKOVERFLOW** | https://stackoverflow.com/questions |

Visualisations:

* Top 10 countries for total volume (completed)
* Most popular of each drink in each continent; of the three listed
* Scatter plot for sum of servings for each alcohol by continent

To Do:

* LOC
* LISTS
* Looping / iterrows