CA-2

EDA PROJECT

* UNIVARIATE ANALYSIS
* BIVARIATE ANALYSIS
* STATISTICAL INTUITION

SUBMITED BY: AMARDEEP SINGH

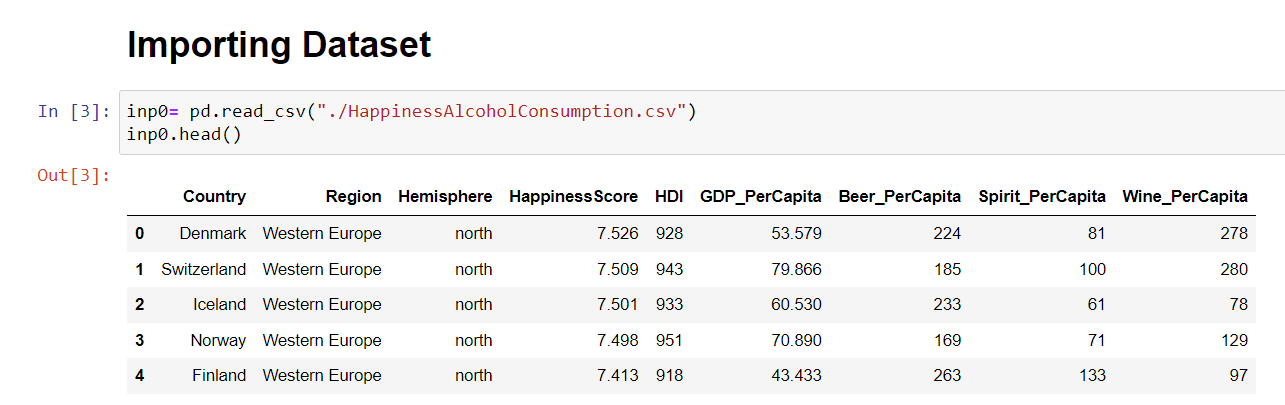
REG: 12006933

ROLL NUM: 52

SECTION: K20CH

**Data Preparation**

Importing Data:



Disclamier**:**

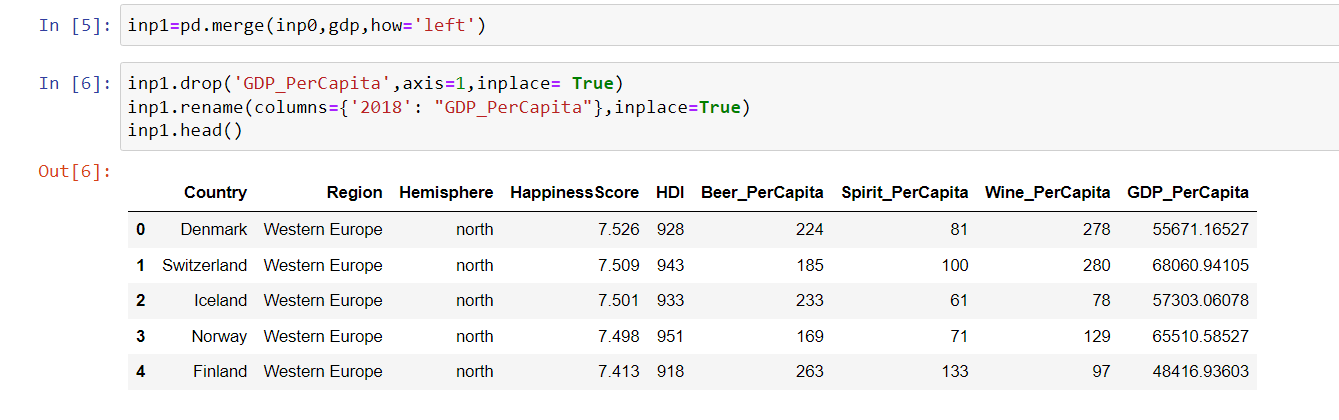
*Column GDP\_PerCapita in our dataset is incorrect. So we will be replacing that column with another dataset's column GDP.*

Importing another dataset for gdp values from 2018

**Graphical user interface, text, application, email

Description automatically generated**

**Merging and renaming**

****

**Cleaning Dataset**

**Text, table

Description automatically generated**

We have got a few null values. Let’s check where they are.

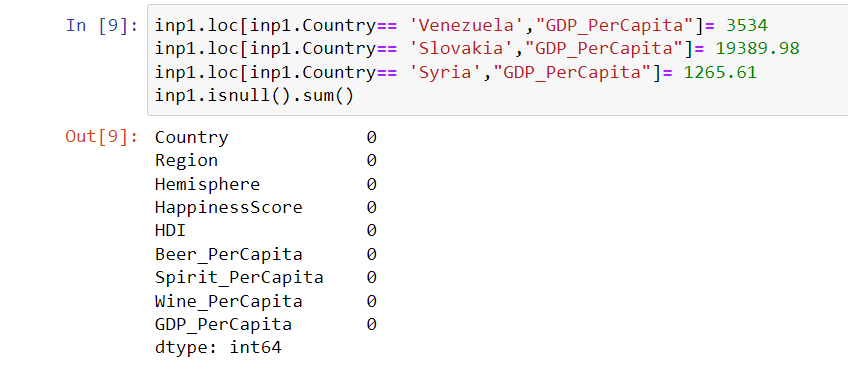
Graphical user interface, text, application

Description automatically generated

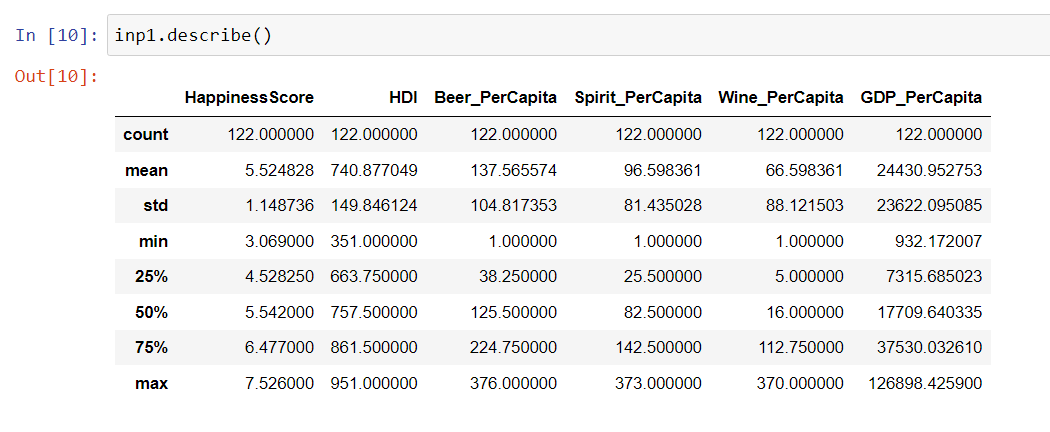
Now after finding these values from google

1. Venezuela=3534
2. Slovakia= 19389.98
3. Syria= 1265.61[¶](http://localhost:8888/notebooks/PycharmProjects/Machine%20Learning/EDA%20Project/cse%20353%20ca2.ipynb#Syria=-1265.61)

**Lets fix them**

****

**Lets look for logically null values or outliers**

****

**Chart, box and whisker chart

Description automatically generated**

**Table

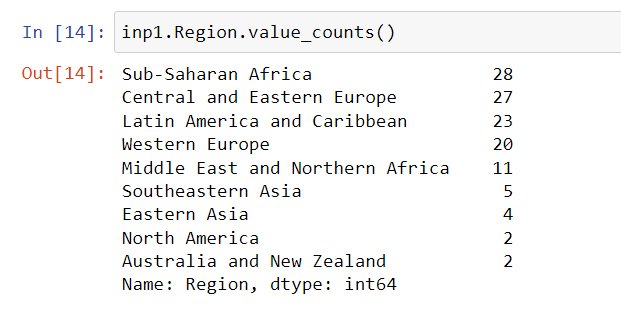
Description automatically generated**

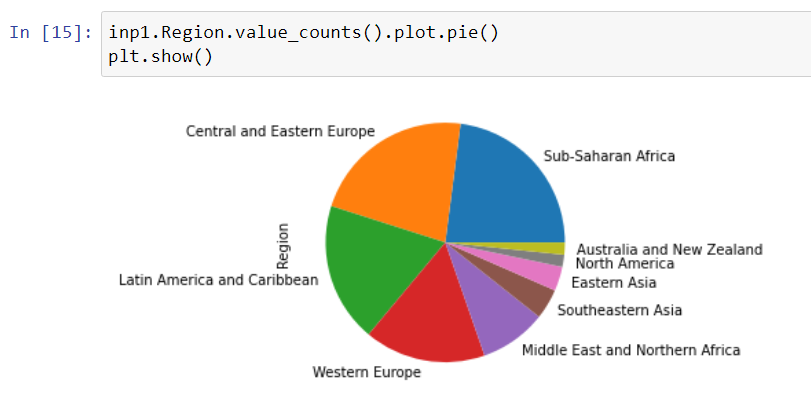
As all of these countries' data is important, we can’t rule out them.

To avoid further issues we will work mostly on percentiles and medians

**Univariate Analysis**

1. **Region**



****

1. **Hemisphere**

**Table

Description automatically generated with medium confidence**

Some Values are incorrect, lets fix them.

Graphical user interface, application

Description automatically generated

Chart, pie chart

Description automatically generated

1. **Happiness Score**

**Graphical user interface, text, application, email

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

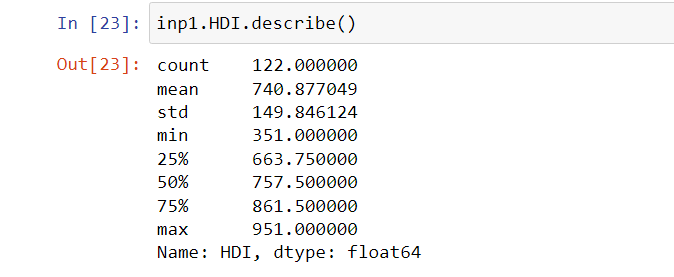
**Graphical user interface, text, application

Description automatically generated** **Graphical user interface, text, application, email

Description automatically generated**

1. **HDI (Human Development Index)**

Checking for data distribution

****

Boxplot to check outliers and data distribution

**Chart, box and whisker chart

Description automatically generated**

Highest Human Development Index

**Graphical user interface, text, application, email

Description automatically generated**

Lowest Human Development Index

**Graphical user interface, text, application, email

Description automatically generated**

1. **Beer , Spirit, Wine per capita**

Beer:

Checking for data abnormalities and other statistical data

**Graphical user interface, text

Description automatically generated**

Spirit:

Checking for data abnormalities and other statistical data

Table

Description automatically generated with medium confidence

Wine:

Checking for data abnormalities and other statistical data

Table

Description automatically generated

Lowest Consumption of these beverages

Graphical user interface, text, application

Description automatically generated

Heighest Consumption of these beverages

Text

Description automatically generated with medium confidence

1. **GDP per capita**

**Chart, box and whisker chart

Description automatically generated**

**Heighest:**

**Graphical user interface, application

Description automatically generated**

**Lowest:**

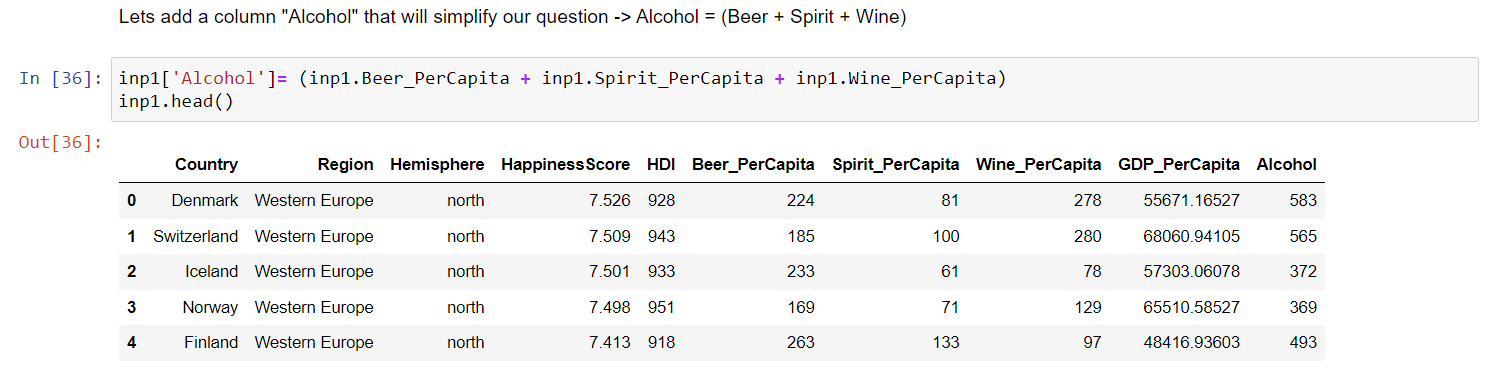
**Text

Description automatically generated**

1. **Overall Alcohol Consumption**

Let’s add a column "Alcohol" that will simplify our analysis

Alcohol = (Beer + Spirit + Wine)

****

These countries drink alcohol the most.

**Graphical user interface, text

Description automatically generated with medium confidence**

These countries drink alcohol the least.

**Graphical user interface

Description automatically generated**

1. **Happiness Category**

To simplify our analysis, we will make a new category HappinessCat, and divide happiness score based on persentiles.

Table

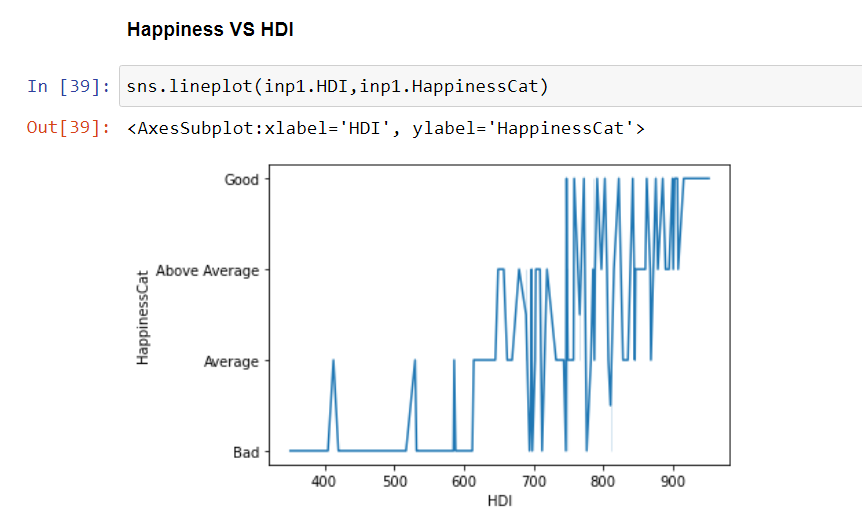
Description automatically generatedGraphical user interface, chart, application, pie chart

Description automatically generated

**Bivariate Analysis**

1. **Happiness vs HDI**

**Line plot**



**Bar plot**

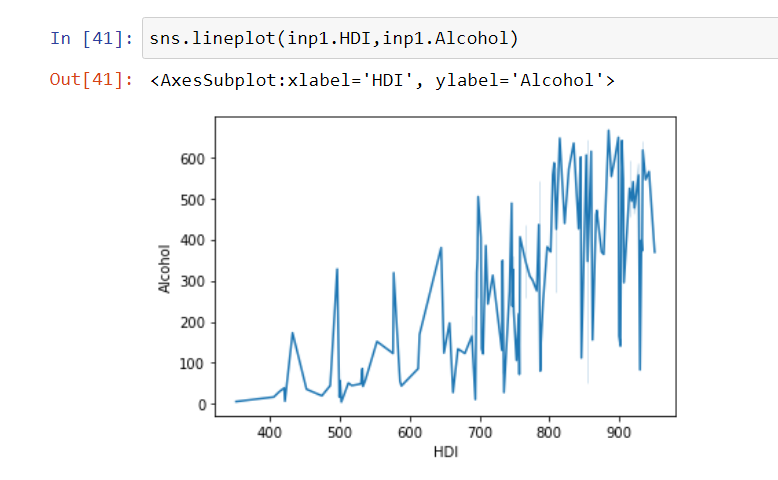
Chart, bar chart

Description automatically generated

This graph somewhat tells developed countries are mostly happy. Whereas mostly underdeveloped countries have low happiness score.

1. **HDI VS Alcohol**

**Line plot**



**Scatter plot**

Chart, scatter chart

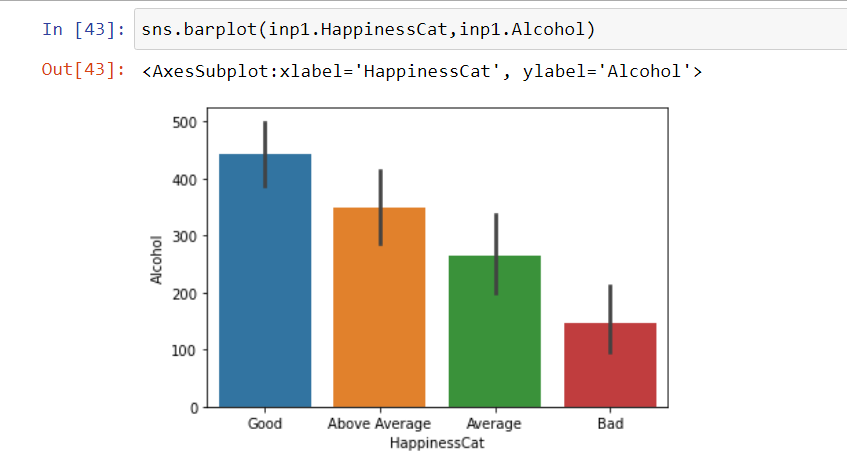
Description automatically generated

Not much is clear here, but we can see, developed countries consumes more alcohol than underdeveloped ones[**.**](http://localhost:8888/notebooks/PycharmProjects/Machine%20Learning/EDA%20Project/cse%20353%20ca2.ipynb#Not-much-is-clear-here,-but-we-can-see,-developed-countries-consumes-more-alcohol-than-underdeveloped-ones)

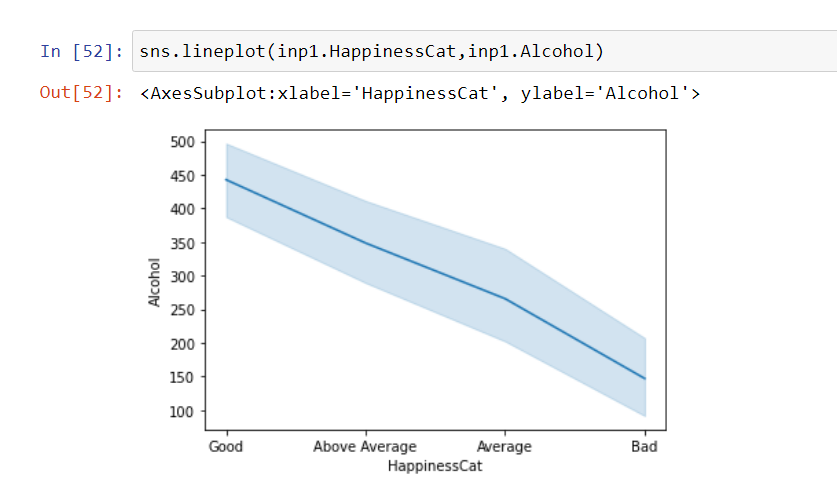
More will be revealed in further multivariate analysis.

1. **Happiness vs Alcohol**

**Bar Plot**

****

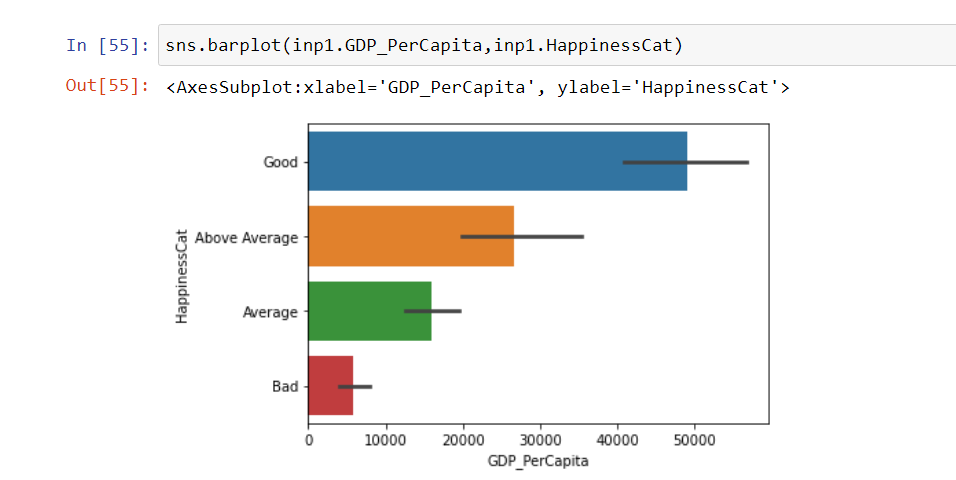
**Line Plot**

****

Here we are getting a bit of hint that happiest countries usually consume more alcohol.

1. **GDP per capita vs Happiness**

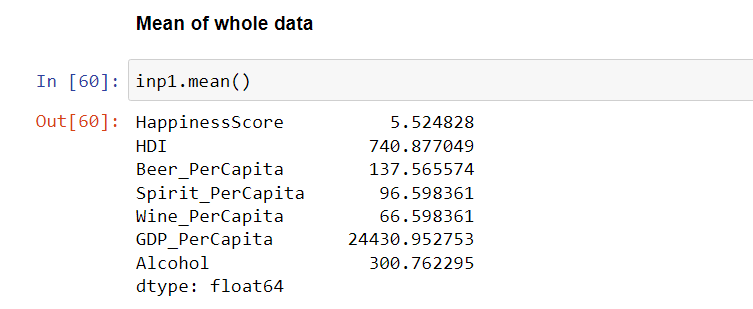
**Bar Plot**

****

Better GDP per capita means happier people, happier nation.

**Statistical Intuition**

**Mean:** Average of all the columns respectively.

****

**Median:** Centre of all the columns respectively.

**Table

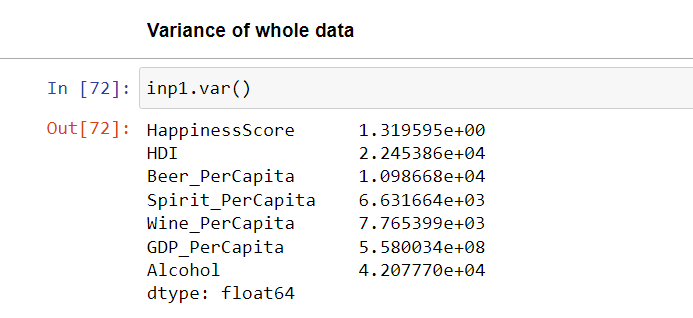
Description automatically generated**

**Standard Deviation:**

**Table

Description automatically generated with low confidence**

**Variance:**

****

**Percentile Division:**

**Graphical user interface

Description automatically generated with medium confidence**