**By: S.Justin Raj**

**Data Analysis on Alcoholic Drinks in Russia and Design Promotional strategy.**

**Induction:**

Alcohol consumption in Russia remains among the highest in the world. According to a 2011 report by the World Health Organization, which makes it the best place to start a beverage business. People love their drinks and the company which owns a chain of stores across Russia that sell a variety of alcoholic drinks wants to invest in marketing campaigns. The company recently ran a wine promotion in Saint Petersburg that was very successful. Due to the cost to the business, it isn’t possible to run the promotion in all regions. In this project, we are going to analyses our data, fix missing values, visualize data, train the clustering model, and finally visualize our results. Our main goal is to find the next ten locations similar to Saint Petersburg and where the wine promotion campaign became successful.

# 

# Data:

The marketing team has sourced you with historical sales volumes per capita for several different drinks types.

* "year" - year (1998-2016)
* "region" - the name of a federal subject of Russia. It could be oblast, republic, krai, autonomous okrug, federal city, and a single autonomous oblast
* "wine" - sale of wine in litters by year per capita
* "beer" - sale of beer in litters by year per capita
* "vodka" - sale of vodka in litters by year per capita
* "champagne" - sale of champagne in litters by year per capita
* "brandy" - sale of brandy in litters by year per capital.

## **Loading Dataset:**

We have used pandas to load the .csv dataset and its pretty small dataset containing yearly (1998-2016) alcohol consumption (beer, champagne, brandy, wine, vodka) per region (85). There is 1615 sample which is logical as we have 19 years of data and 85 regions. The beer is leading the game as the mean value of beer is 51.3 litters by year per capita and the second-highest is vodka 11.81 litters by year per capita which is now even close to beer. This means people prefer beer as a go-to beverage. The beer also has the highest standard deviation which means that its demand is not stable and can fluctuate with time whereas champagne and brandy are a pretty safe bet if you want to start a low-risk business with the lowest standard deviation.

## **Correlation:**

There is a high correlation between champagne and brandy which makes it even better. If you promote champagne there will be an increase in the sale of brandy and champagne, which makes it win-win situation.

# Missing Values:

We will be using pandas’ data value has frame background gradient to display several missing values and percentages. It seems like all drink’s columns have missing values and the highest is brandy with 66.

# Alcohol Consumption Past Trend:

# As we can see beer consumption has risen with time up till 2007, then it became steady and its declining since 2011. Wine, champagne, and brandy consumption are lower than beer but they are steady. The vodka demand has increased up till 2002 and it's declining slowly and steadily with time. **The safest bet is to launch a campaign on either brandy or champagne but wine can be profitable as it has more consumption per capita and it is stable.**

### **Social:**

In the early 1980s, an estimated "two-thirds of murders and violent crimes were committed by intoxicated persons; and drunk drivers were responsible for 14,000 traffic deaths and 60,000 serious traffic injuries". In 1995, about three quarters of those arrested for homicide were under the influence of alcohol, and 29% of respondents reported that children beaten within families were the victims of drunks and alcoholics.

A 1997 report published in the Journal of Family Violence found that among male perpetrators of spousal homicide, 60–75% of offenders had been drinking prior to the incident.

**The marketing team has sourced you with historical sales volumes per capita for several different drinks types.**

##"year" - year (1998-2016)

##"region" - the name of a federal subject of Russia. It could be oblast, republic, krai, autonomous okrug, federal city, and a single autonomous oblast

##"wine" - sale of wine in litters by year per capita

##"beer" - sale of beer in litters by year per capita

##"vodka" - sale of vodka in litters by year per capita

##"champagne" - sale of champagne in litters by year per capita

##"brandy" - sale of brandy in litters by year per capita.

Data set (Import):

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.graph\_objects as go

import plotly.express as px

import os

import warnings

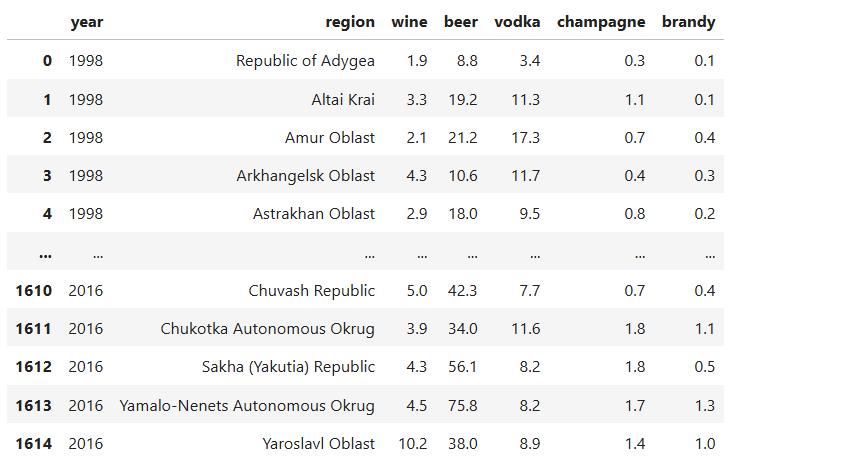
warnings.filterwarnings('ignore')

np.random.seed(2021)

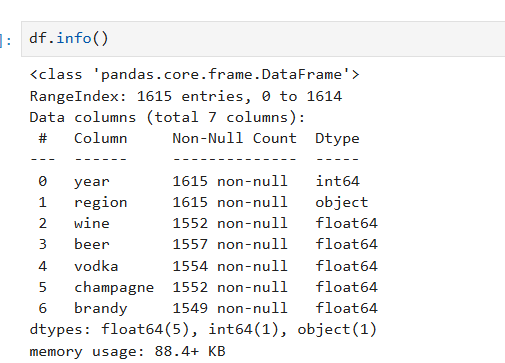
Reading data:

df=pd.read\_csv("russia\_alcohol.csv")

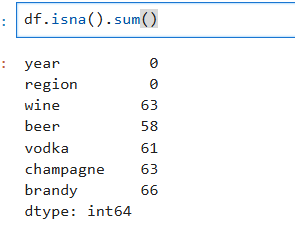
df



df.info():



df.isna().sum():



**Missing Values data set:**

NA = pd.DataFrame(

data=[

df.isna().sum().tolist(),

[

"{:.2f}".format(i) + "%"

for i in (df.isna().sum() / df.shape[0] \* 100).tolist()

],

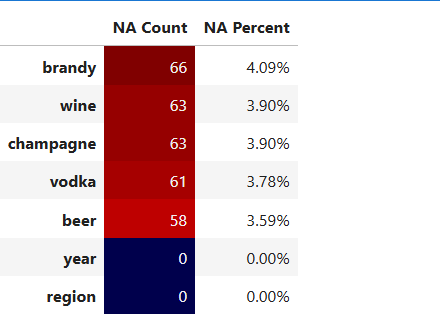
],

columns=df.columns,

index=["NA Count", "NA Percent"],

).T.sort\_values(by="NA Count", ascending=False)

NA.style.background\_gradient(cmap="seismic", subset=["NA Count"])

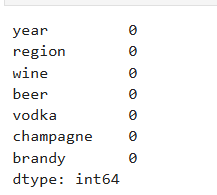
****

**df.fillna(method="pad", inplace=True)**

**df.isna().sum()**

## We are going to use the fill function and method pad to fill missing values with previous values in a column.

## As we can see that there are no missing values in our dataset.



df.head():



df.describe():



**## list(df['region'].unique()):**

**fig, axes = plt.subplots(1, 5, figsize=(15, 6), sharey=True)**

**for x in range(1, 6, 1):**

**column\_name = df.columns[x + 1]**

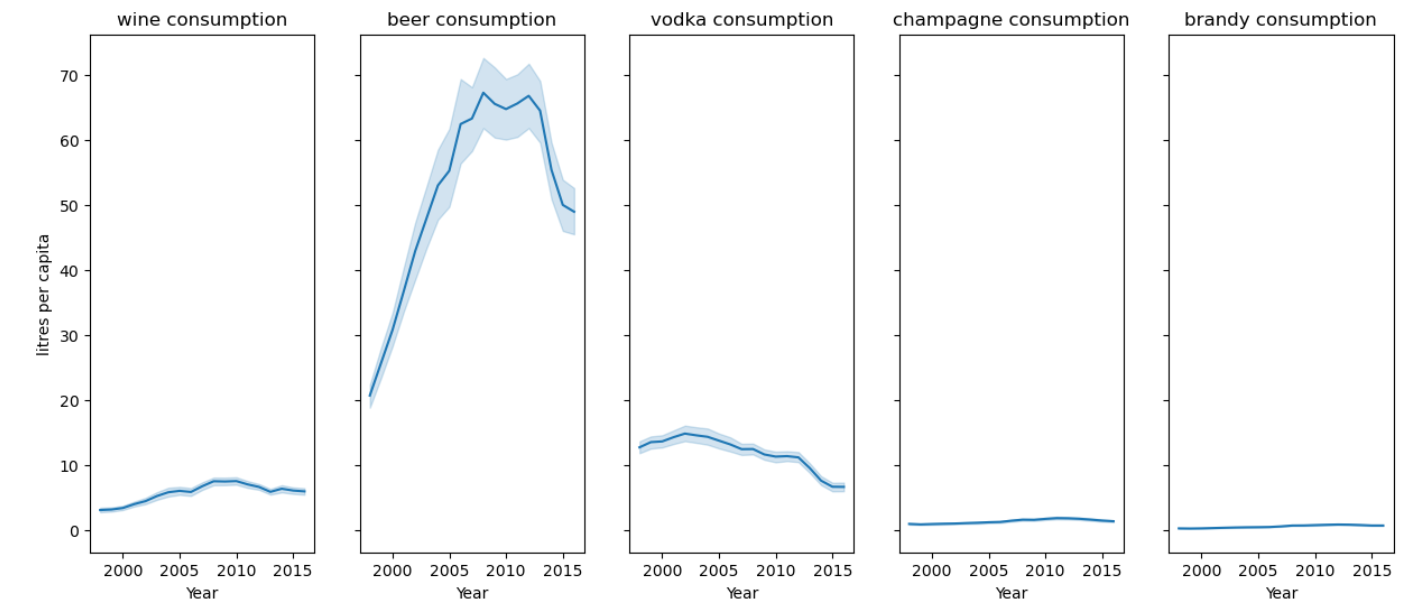
**sns.lineplot(data=df, x="year", y=f"{column\_name}", ax=axes[x - 1])**

**axes[x - 1].set\_title(f"{column\_name} consumption")**

**axes[x - 1].set\_xlabel("Year")**

**axes[x - 1].set\_ylabel("litres per capita")**

## The safest bet is to launch a campaign on either brandy or champagne but wine can be profitable as it has more consumption per capita and it is stable.

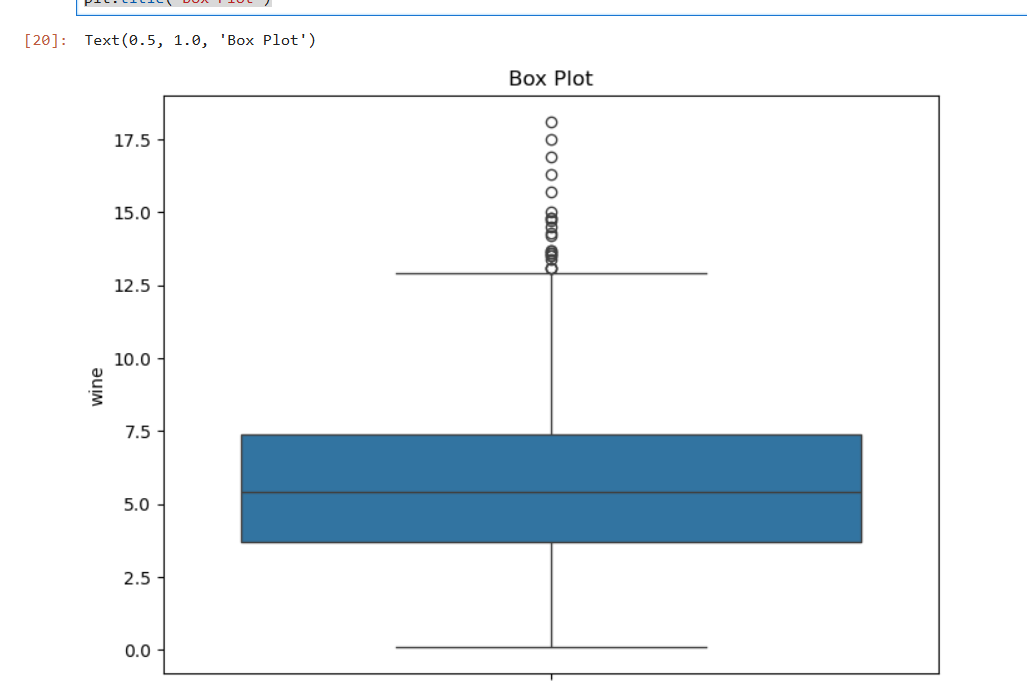


**Box plot:**

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

sns.boxplot(data=df['wine'])

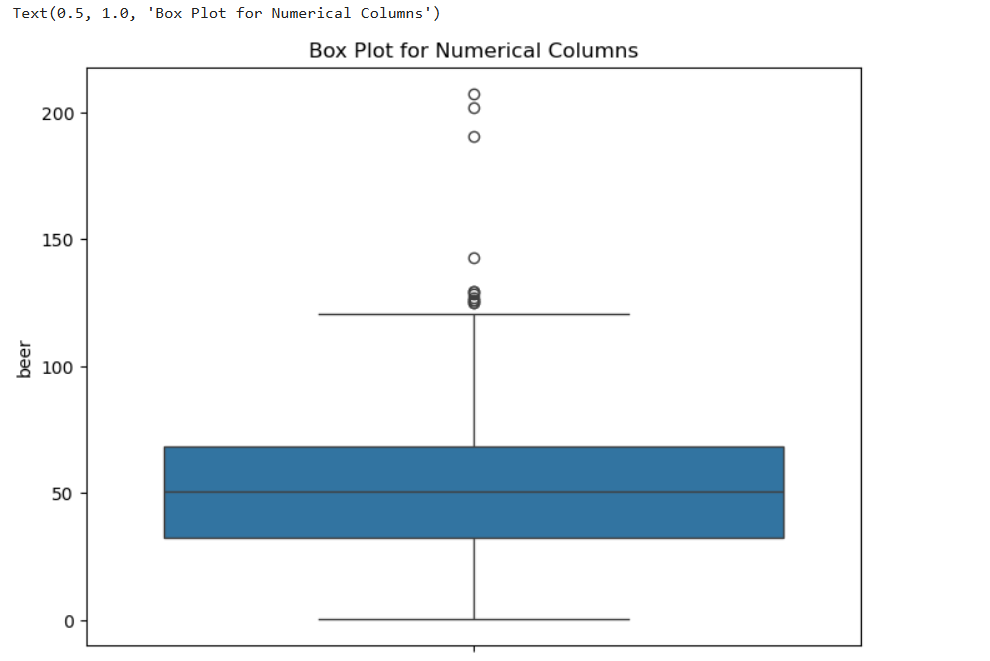
plt.title('Box Plot')



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

sns.boxplot(data=df['beer'])

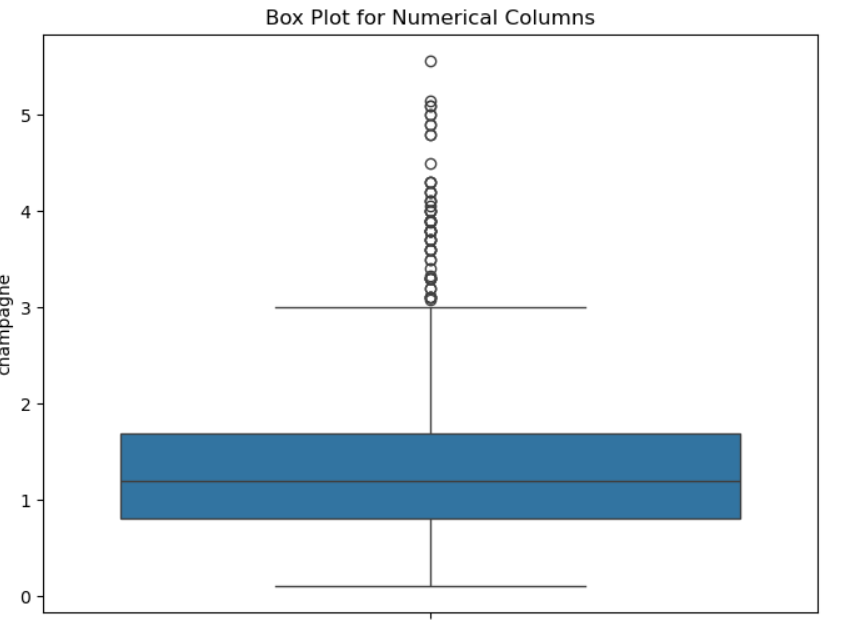
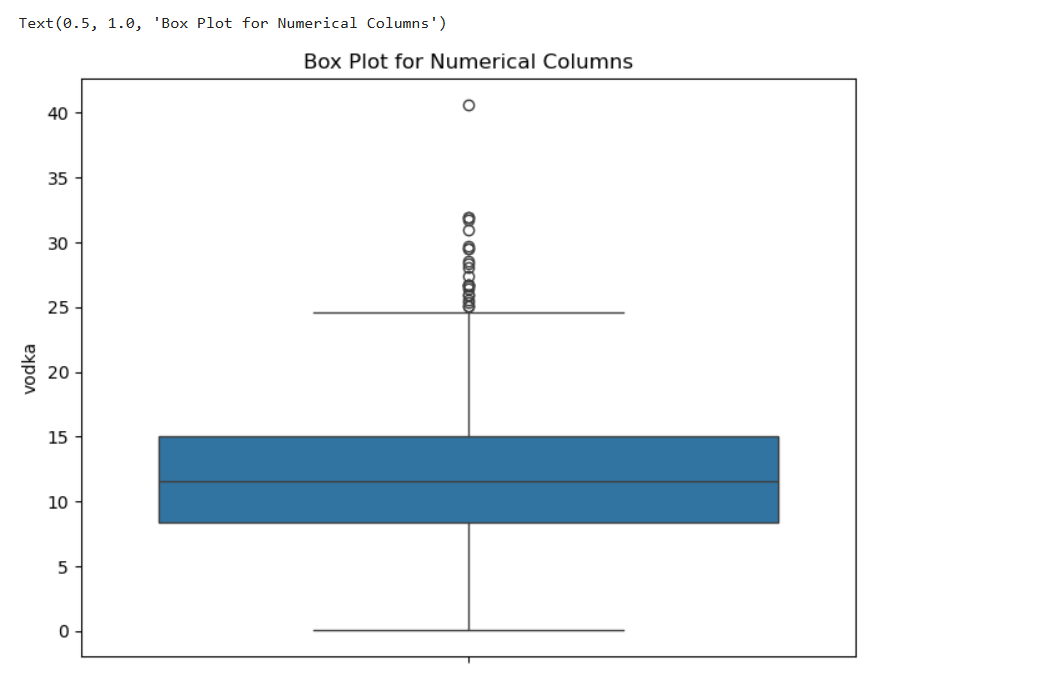
plt.title('Box Plot for Numerical Columns')



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

sns.boxplot(data=df['vodka'])

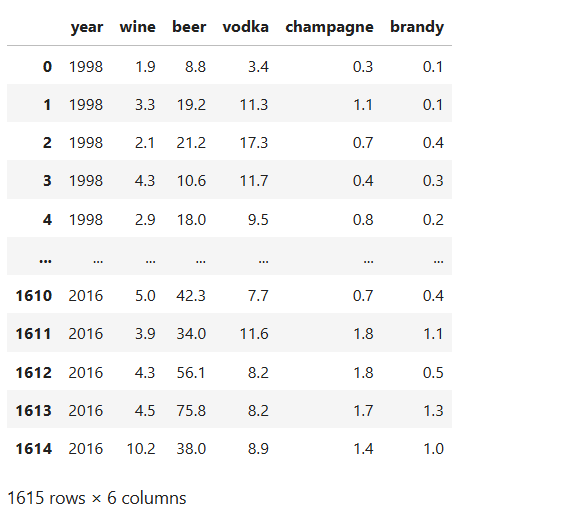
plt.title('Box Plot for Numerical Columns')



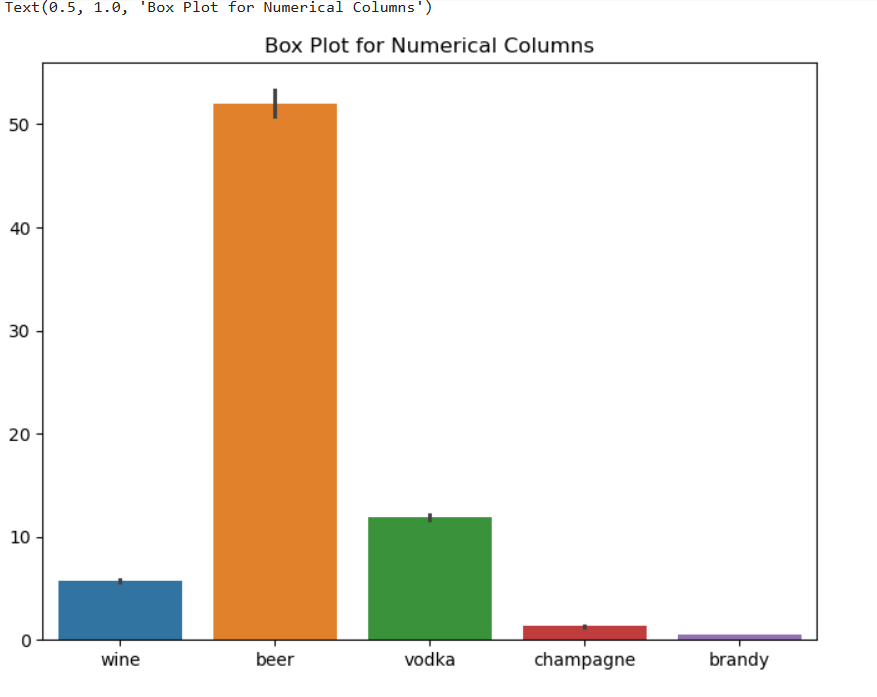
**## PICKING NUMERICAL COLUMNS FROM THE DATASET:**

numrical\_columns=df.select\_dtypes('number')

numrical\_columns



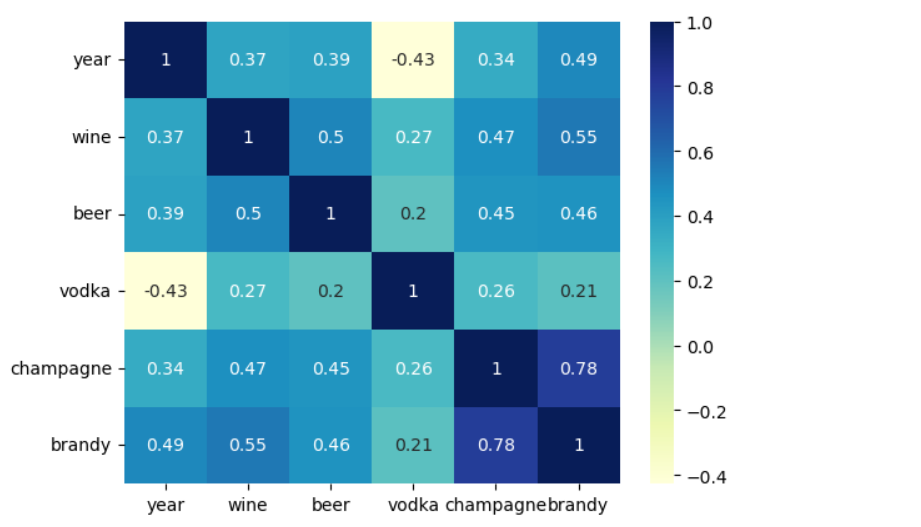
**Bar plots:**

****

**Heat Map:**

dataplot = sns.heatmap(numrical\_columns.corr(), cmap="YlGnBu", annot=True)

plt.show()



numrical\_columns.corr()

