

Alcohol Consumption in Russia

Introduction

Exploratory Data Analysis

The data contains sales of different alcohols (wine, beer, brandy, champagne, vodka) in liters by year per capita for 85 different regions throughout Russia. The data tracks yearly sales from 1998 to 2016. There were only four regions that had missing values and were dropped from the dataset. The shape of the data after this was (1539, 7).

Summary Statistics for the data:

	wine	beer	vodka	champagne	brandy
count	1539	1539	1539	1539	1539
mean	5.64	51.72	11.90	1.32	0.52
std	2.81	25.12	5.08	0.80	0.40
min	0.10	1.00	0.40	0.10	0.00
25%	3.55	32.70	8.40	0.80	0.20
50%	5.40	50.30	11.60	1.20	0.40
75%	7.40	67.50	15.00	1.66	0.70
max	18.10	207.30	40.60	5.56	2.30

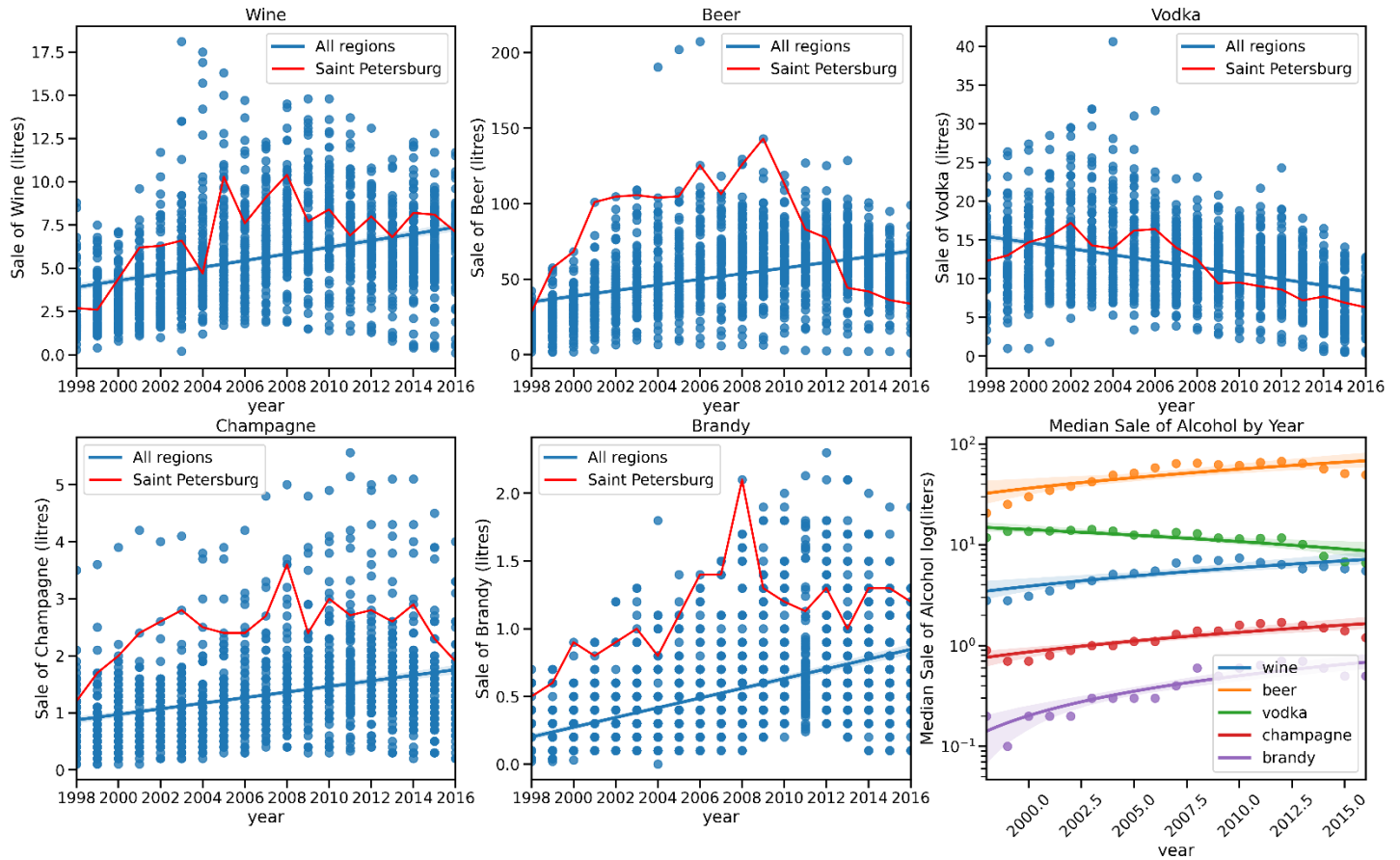
There is a lot of variances in sales in each region over the years, moreover, it doesn't seem like the sales of alcohol follow a specific pattern.

Mean variance in region for each alcohol

Alcohol	Mean Variance
Wine	4.57
Beer	412.07
Vodka	13.43
Champagne	0.29
Brandy	0.09

Saint Petersburg on average has higher sales for all alcohols except vodka compared to the other regions. However, the sales were in general down in more recent years.

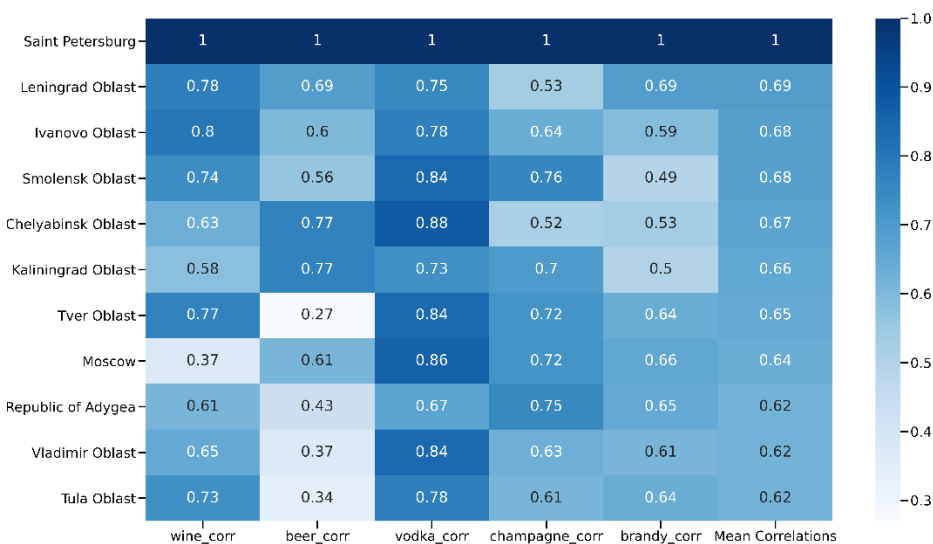
Sale of Alcohol in Russia



Methodology

Because of the erratic nature of the sales of alcohols over the years along with the limited size of the dataset, it is not possible to do a time series analysis. I used correlation and k-means clustering to find regions similar to Saint Petersburg.

First, I calculated the correlation of each region with Saint Petersburg for each alcohol and then found the mean correlation. The correlations would capture the trend of the data to some extent. The following is a heatmap of the 10 regions with the highest correlations:



Then I prepared the data for k-means clustering by scaling it using min max scaler with feature range 0 to 1. I added the correlations to the data frame so that k-means picks up on the trend and not just cluster the closest values.

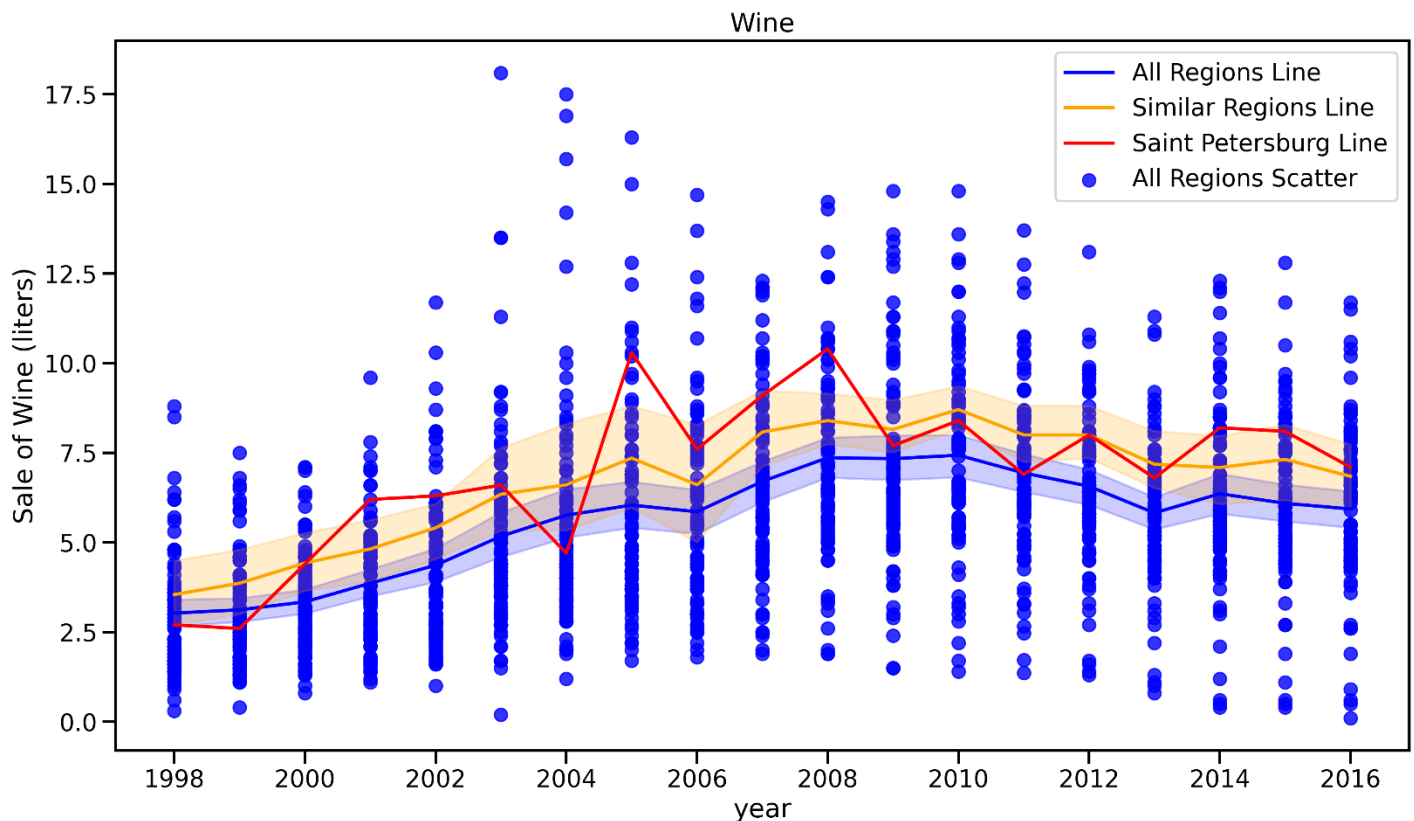
Using the elbow method, I determined that the data should be clustered into 3 clusters. I extracted the data that was in the same cluster as Saint Petersburg and selected the ten regions with the highest correlations.

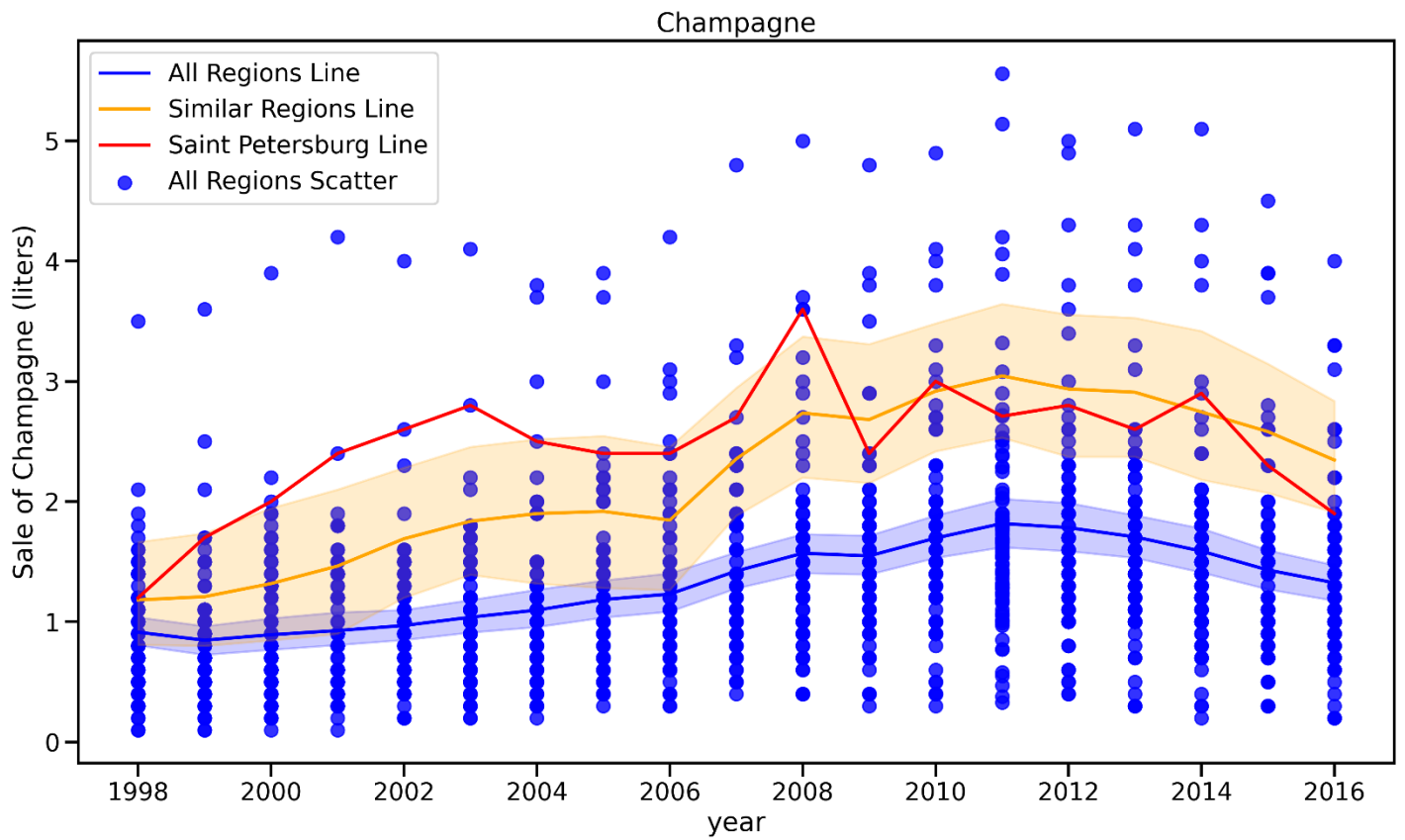
Results

The regions most similar to Saint Petersburg are the following:

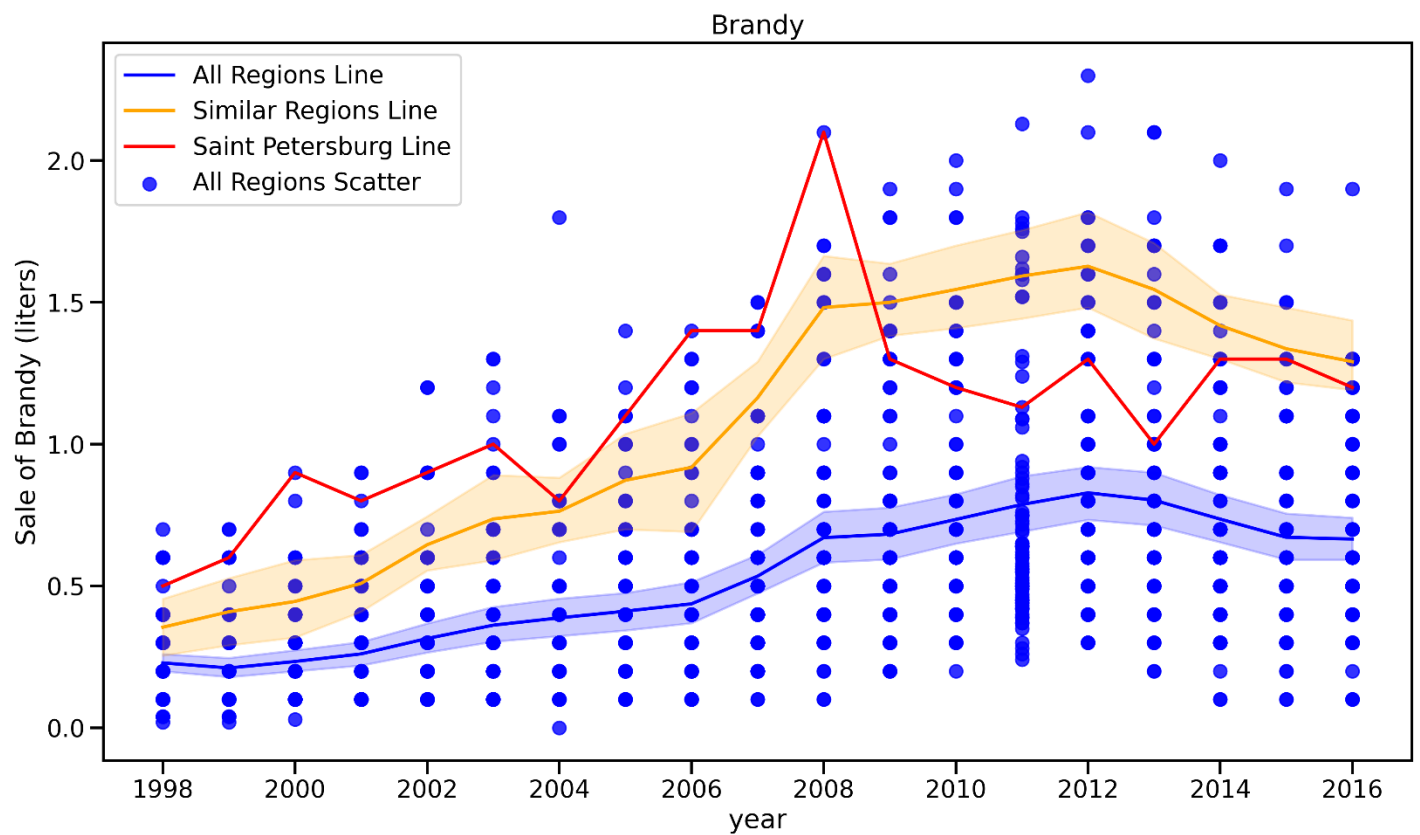
1. Leningrad oblast
2. Kaliningrad oblast
3. Moscow
4. Tyumen oblast
5. Khanty–Mansi autonomous okrug – yugra
6. Murmansk oblast
7. Moscow oblast
8. Yamalo-nenets autonomous okrug
9. Sakhalin oblast
10. Kamchatka krai

For the wine and champagne data, the values of the ten regions are very close to Saint Petersburg, and they have a similar trend as well.

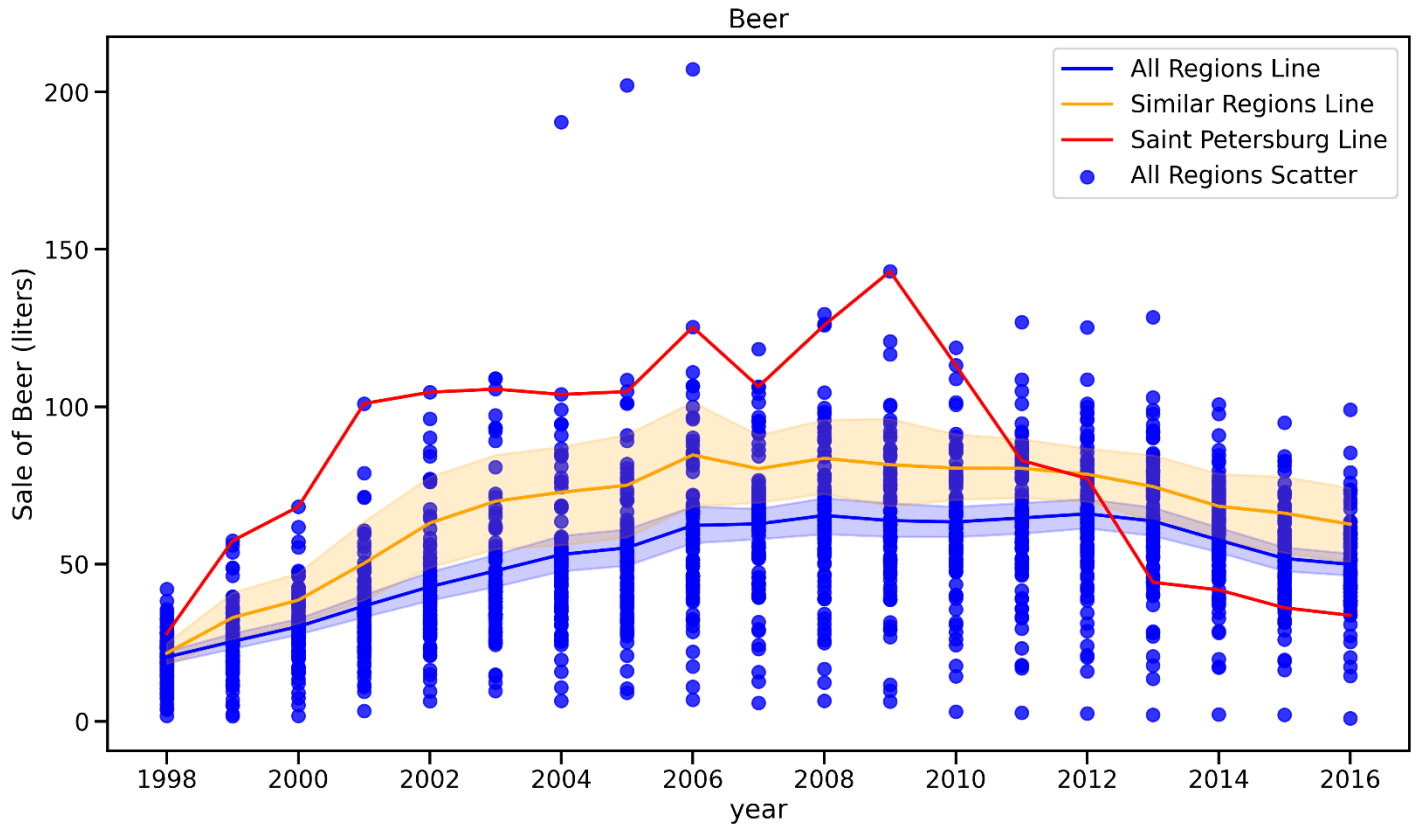




For the brandy data the ten selected regions have similar values and trend to Saint Petersburg but too the same degree as wine or champagne data.



For the beer data the ten regions have above average sales just like Saint Petersburg, however, that is the extent to which they are similar.



Unfortunately, the sales of vodka are too high in the ten regions.

