Alcohol Drinking Habits in Russia

Dustin Bryant

July 2021

1 Abstract

Here we predict the top ten regions in Russia with drinking habits most similar to Saint Petersburg. Drinking volume per capita per year was collected from 1998 to 2016 for consumption of five alcoholic beverages: beer, wine, brandy, champagne, and vodka.

2 Method

Finding the ten regions most similar to St. Petersburg in drinking habits is nothing more than finding the top ten regions with the minimal *distance* to St. Petersburg after the correct measure of distance is established. Such a distance is not merely geographical distance but rather distance as measured in drinking habits. To motivate our method we argue by analogy. Suppose we want to know the distance between point A and B with coordinates (x_1, y_1, z_1) and (x_2, y_2, z_2), respectively. The distance, $d(A, B)$, from A to B is calculated as

$$d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

In our case, rather than a point being in geometric space, we consider points to be measures of possible drinking values. Thus if the average person of St. Petersburg drinked 7.1l of wine, 33.7l of beer, 6.3l of vodka, 1.9l of champagne, and 1.2l of brandy, we could represent this by the vector equation $(w_{sp}, b_{sp}, v_{sp}, c_{sp}, br_{sp}) = (7.1, 33.7, 6.3, 1.9, 1.2)$, where the subscript sp specifies St. Petersburg. Likewise we could compute the "distance" bewteen St. Petersburg and Region r by the equation

$$d(sp, r) = \sqrt{(w_{sp} - w_r)^2 + (b_{sp} - b_r)^2 + (v_{sp} - v_r)^2 + (c_{sp} - c_r)^2 + (br_{sp} - br_r)^2}$$

In an obvious sense, an ounce of beer is very different from an ounce of vodka both in its intoxicating effects as well as price; to account for this we *normalize* our data. We compute noramlized x as $\hat{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$ where min [resp. max] denotes the minimum [resp. maximum] value of all actual values x obtains from the dataset.

If we were only given data from 2016 then we would use this equation above to compute the distance from St. Petersburug to each region r then we would sort the results and enumerate the top ten regions. Fortunately, we have data over a 19 year period. Suppose that $d_{2016}(sp, r)$ denotes the distance to region r as measured in 2016 and $d_{2015}(sp, r)$ denoted the distance as measure in 2015, we might try combining this data to make a prediction. A natural first guess for how to relate these distances is to make a third distance $d(sp, r)$ which is the sum of these distances, so that $d(sp, r) = d_{2016}(sp, r) + d_{2015}(sp, r)$. This works well, but it also assumes that information from 2016 is just as relevant as information from 2015. To account for the more recent year being more relevant for predicting drinking habits in the year 2017, we *may* wish to weigh the distance in 2015 less. To that end, we may define $d(sp, r) = d_{2016}(sp, r) + 0.9 * d_{2015}(sp, r)$. The use of 0.9 is completely arbitrary but it encodes the fact that trends care more about recent facts than more distant facts. Continuing in this fashion, we can compute the total distance across all years as

$$d(sp, r) = d_{2016}(sp, r) + 0.9 * d_{2015}(sp, r) + (0.9)^2 * d_{2014}(sp, r) + (0.9)^3 * d_{2013}(sp, r) + \cdots$$

Finally, to perform the computations quicker we can use the square of the distance from each year since the predictions will be equivalent.

3 Results

First we remove the Chechen Republic, the Republic of Crimea, the Republic of Ingushetia, and Sevastopol from our dataset since there is limited data on these regions due most likely to their disputed status as part of Russia. Using the above mentioned method we arrive at the following

Ranking	Region
1	Leningrad Oblast
2	Tyumen Oblast
3	Murmansk Oblast
4	Khanty–Mansi Autonomous Okrug – Yugra
5	Sverdlovsk Oblast
6	Kaliningrad Oblast
7	Yamalo-Nenets Autonomous Okrug
8	Chelyabinsk Oblast
9	Moscow Oblast
10	Khabarovsk Krai
...	...
77	Tuva Republic
78	Karachay-Cherkess Republic
79	Republic of Dagestan
80	Republic of North Ossetia-Alania

Notably Leningrad Oblast is predicted to be most similar to St. Petersburg which is reassuring since "Leningrad" was the name of St. Petersburg from 1924 until 1991 and "oblast" refers to administrative regions and they often share a name with a nearby city. That is to say, our model predicts that St. Petersburg's oblast is predicted to have drinking habits most similar to St. Petersburg. Notably, our model predicts Moscow Oblast but not Moscow (city), the capital and largest city, in the top ten most similar regions. Note that while Moscow may drink more and share many cultural similarities with St. Petersburg, the level of drinking in Moscow may significantly dwarf the drinking habits of St. Petersburg such that it is very far away under our metric. Recall, we are interested in finding the regions most similar to St. Petersburg and not merely the 10 regions that drink the most, yet it just so happens that St. Petersburg is a cosmopolitan region like Moscow so that it may be easy to conflate finding the regions that drink the most and finding the regions most like St. Petersburg. Lastly, our model predicts Tuva Republic (TR), Karachay-Cherkess Republic (KCR), Republic of Dagestan (RD), and Republic of North Ossetia-Alania (NOA) in last place. This, too, is reassuring since KCR's and RD's population is predominately Muslim, a demographic which tends to avoid alcohol consumption; while TR is predominately populated by Tuvans which are culturally more similar to East Asians.

4   PCA Reveals Two Audiences

The above analysis found the regions with drinking habits most similar to St Petersburg but it did not discriminate the type of drinkers. On the one extreme, we could ask for the regions most similar to one another based on an individual drink, such as wine, and on the other extreme we could consider all drinks as equal as we had in our analysis above. Practically, however, some people prefer hard liquor (such as brandy and vodka) more than beverages with less alcohol (such as wine, beer, and champagne). Advertisements and promotional events need to be able to discriminate these sorts of facts. However, it might be that the aforementioned dichotomy is simply not supported by the data. To this end, we conduct principal component analysis (PCA) to determine the types of drinkers that actually exist in Russia to see if we can more accurately target advertising. Very roughly, PCA will transform this 5 dimensional data set into 5 principal components (PCs). These PCs represent a new coordinate system where the first PC is the dimension of the data with the most variation, the second PC represents the dimension with the second most variation in data, and so on. By considering simply two PCs we may be able to reduce our consumers to two categories rather than five.
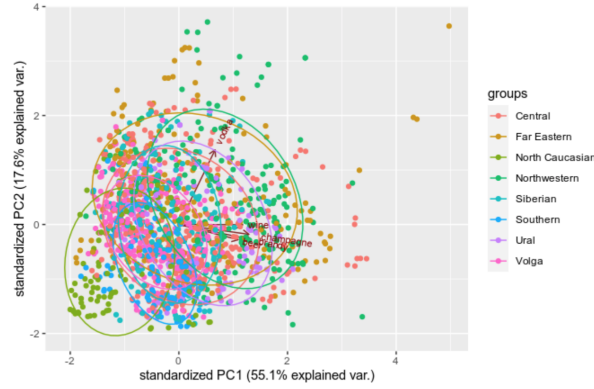
The above graph shows how the various regions across all 19 years can be represented with the first two PCs. Notably, we see that vodka drinkers are approximately represented by the second PC while everyone else's drinking habits are correlated to the first PC. In other words, there are two types of drinkers in Russia: those that prefer drinks other than vodka, and those that prefer vodka. Accordingly we should reperform the previous calculations on the vodka and non-vodka datasets.

The regions with vodka drinking habits most similar to St. Petersburg are as follows

Ranking	Region
1	Chelyabinsk Oblast
2	Vladimir Oblast
3	Republic of Khakassia
4	Ivanovo Oblast
5	Primorsky Krai
6	Udmurt Republic
7	Sverdlovsk Oblast
8	Tomsk Oblast
9	Perm Krai
10	Krasnoyarsk Krai

While the regions with drinking habits most similar to St. Petersburg while ignoring vodka consumption are as follows

Ranking	Region
1	Leningrad Oblast
2	Murmansk Oblast
3	Tyumen Oblast
4	Khanty–Mansi Autonomous Okrug – Yugra
5	Kaliningrad Oblast
6	Sverdlovsk Oblast
7	Yamalo-Nenets Autonomous Okrug
8	Kamchatka Krai
9	Chelyabinsk Oblast
10	Arkhangelsk Oblast

We note that the list of vodka drinkers is significantly different from the original list while the list of non-vodka drinkers introduced Kamchatka Krai and Arkhangelsk Oblast while removing Moscow Oblast and Khabarovsk Krai but otherwise being identical in which regions were in the top ten (though perhaps the orders changed for those still in).

5   Closing Remarks

The above analysis used differences in drinking habits to predict which regions we should target for advertisement. We used all 19 years of our data to make predictions but we put more weight in more recent time periods. Our predictions for regions with drinking habits similar to St. Petersburg were not unexpected as we predict the Leningrad Oblast to be most similar in preference. Finally, we performed PCA to confirm our intuition that vodka drinkers are a different sort of customer and should be target independently. Lastly, we note that when we target non-vodka drinkers we get almost the same results, and notably Moscow Oblast only appears on the initial list when we consider all types of drinkers.