Exploratory analysis on the variable "Nativism"

The "nativism" variable is a 5-items variable with the questions like:

MW_Q9_1. [Immigrants take jobs away from real ...] Do you agree or disagree with the following statements?
MW_Q9_2. [Immigrants take important social services away from real ...] Do you agree or disagree with the following statements?
MW_Q9_3. [When jobs are scarce, employers should prioritize hiring people of this country over immigrants] Do you agree or disagree with the following statements?
MW_Q9_4. [... would be better off if we let in all immigrants who wanted to come here] Do you agree or disagree with the following statements?
MW_Q9_5. [... would be stronger if we stopped immigration] Do you agree or disagree with the following statements?

I explored this variable on two levels: **country level and individual-country level**
For the country level, I firstly recode the variables into numerical responses and then take the average as the numerical representation of that country.

One caveat: question Q9_4 is actually in the reverse direction with the other four questions. Therefore, in this analysis and the last report, I did NOT include the question 4 to average the nativism score. So:

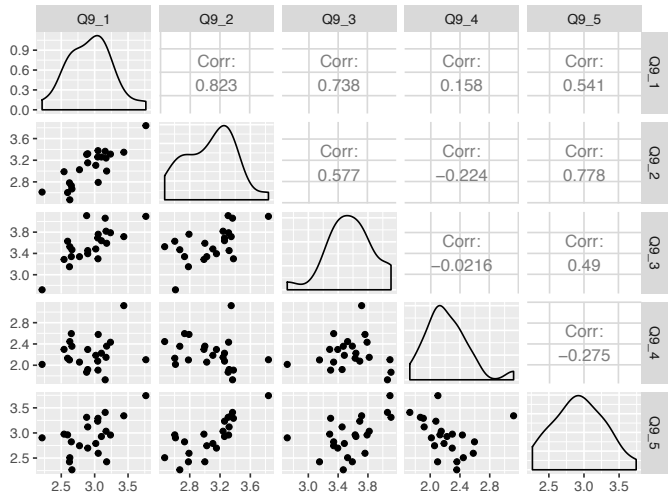
nativism score = (MW_Q9_1 + MW_Q9_2 + MW_Q9_3 + MW_Q9_5) / 4

However, I did include the Q9_4 to explore the full picture of the nativism variable (5-items).

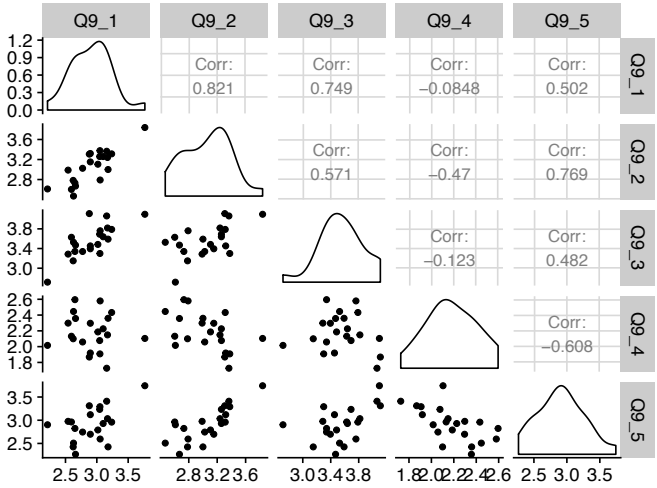
First, I compiled the results from a **country-level**.

Country-level results

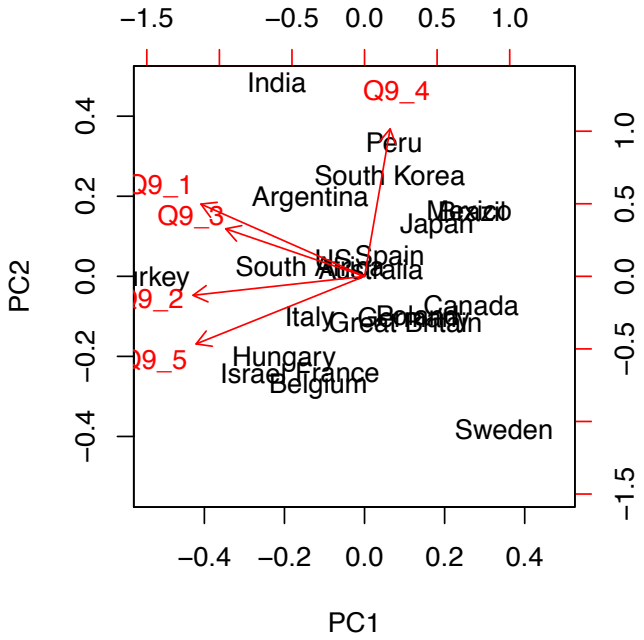
First is the pair-plot to view the relationship and the distribution of the 5 items:



After some deeper look, I found that India is a serious outlier that influences the data, so a pair-plot WITHOUT India is as below:

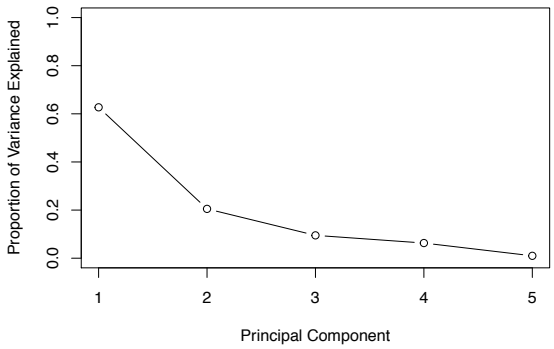


Next I did a principle component analysis to view the relationships between the variables and the observations (countries):

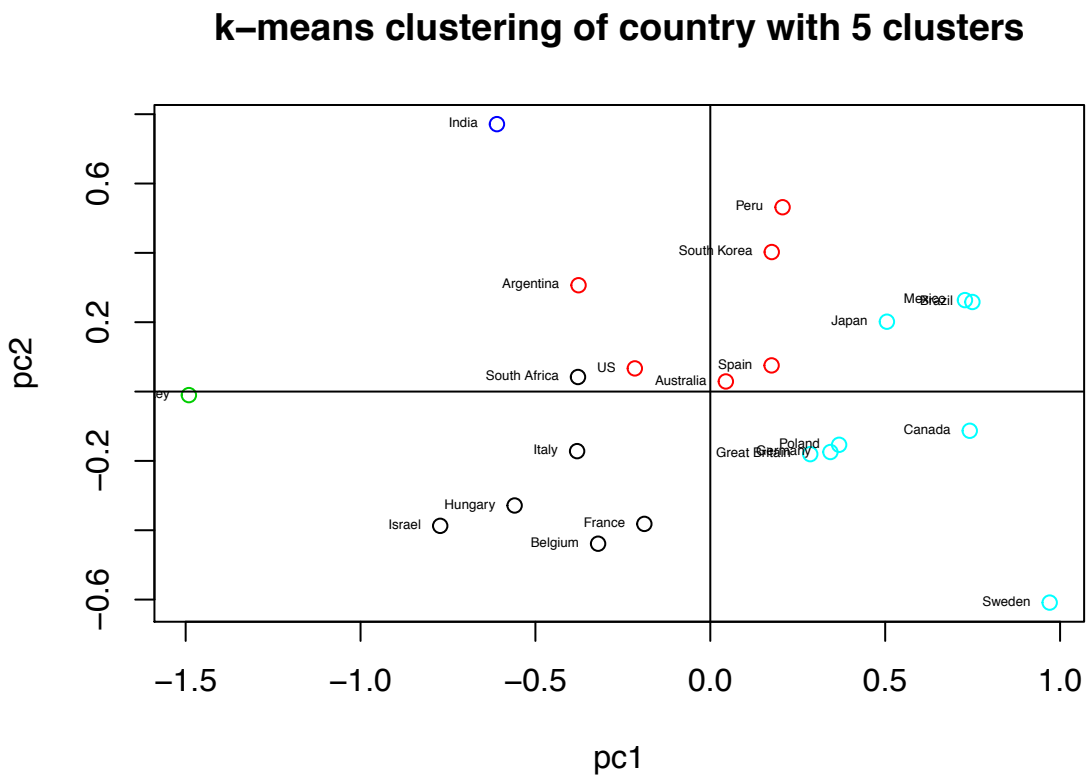


Q4 might be distant from other variables, and Sweden, India, and Turkey are probably the most outliers from other countries.

The above Principle component 1 (PC1) and PC2 explained about 80% of the variance of the whole data, which I think are pretty good indicators of the variable nativism.

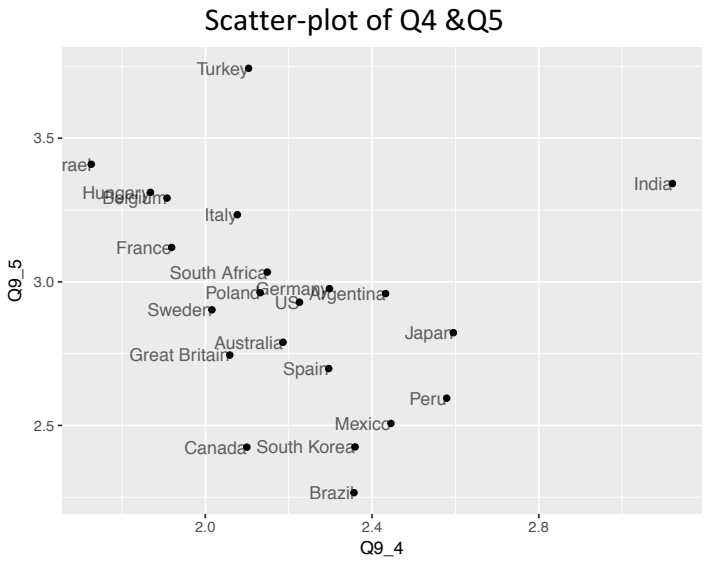


The K-means clustering results plotted on the PC1 and PC2 axis.

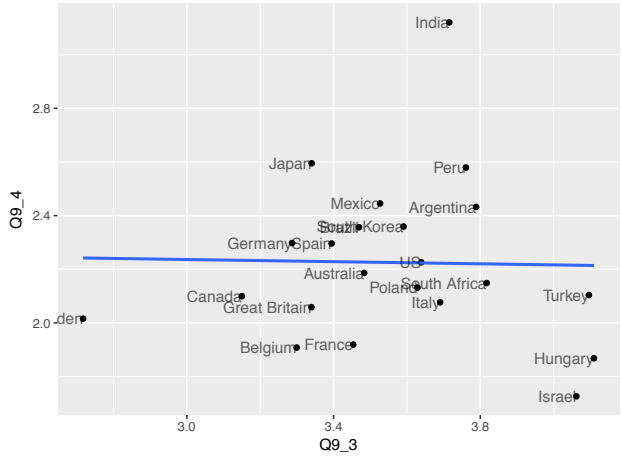


We can see that in the middle of the plot, there are three major clusters (colored in black, bright blue and red). India, Turkey, and Sweden are the farthest from the central clusters.

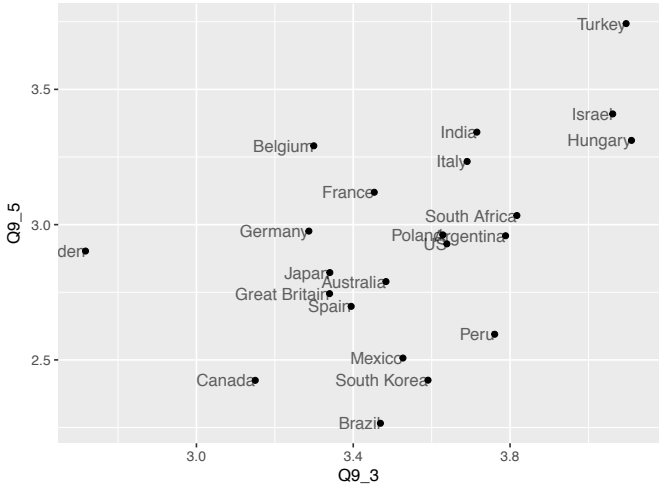
Next, I “cross-tabulated” the different questions to view deeper about the relationships among the question items and countries.



Scatter-plot of Q3 &Q4



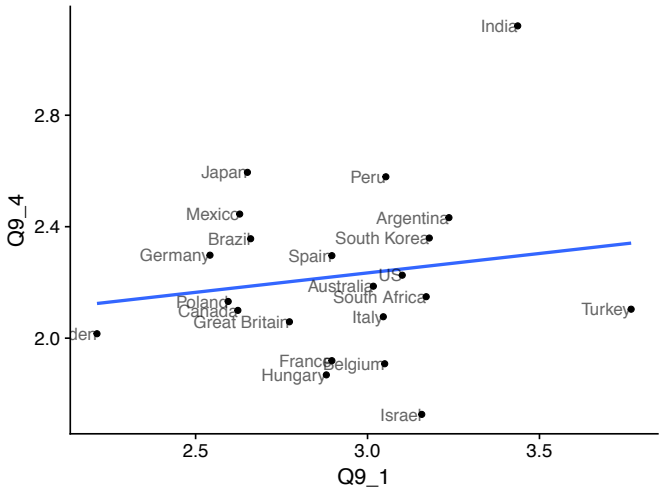
Scatter-plot of Q3 &Q5

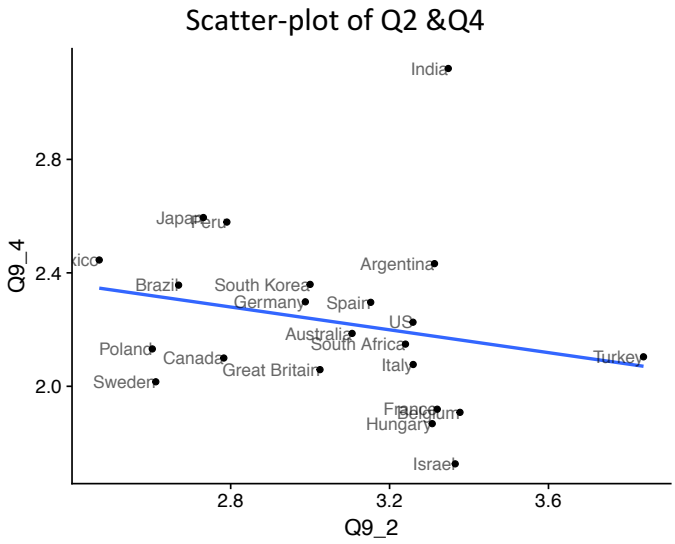


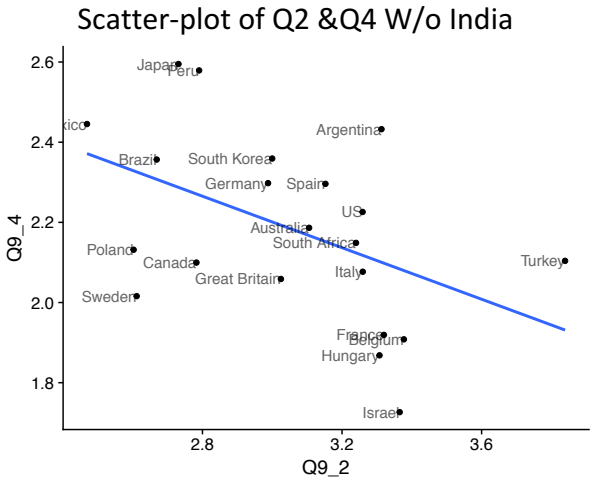
Sweden is an outlier on Q9_3:

```
> compare = c(worldmean = mean(data_nativism$Q9_3),
+             Sweden = data_nativism["Sweden", "Q9_3"])
> compare
worldmean    Sweden
 3.561859  2.717131
```

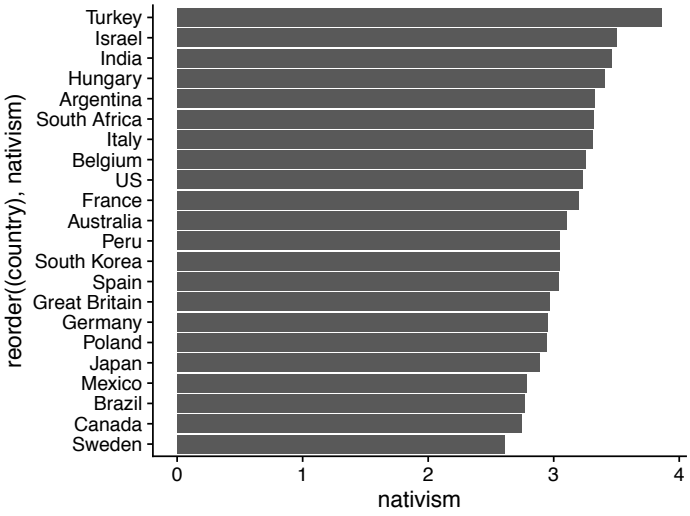
Scatter-plot of Q1 & Q4







Overall, the rankings of the numerical nativism score by countries:

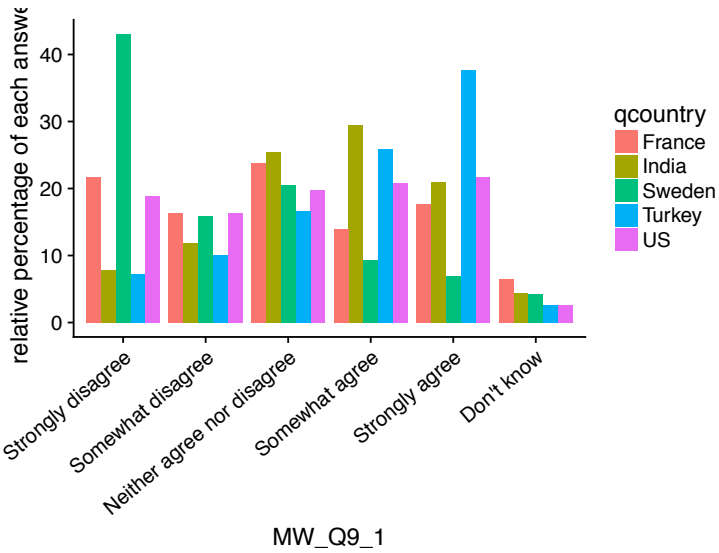


Next, let's move to the individual country level

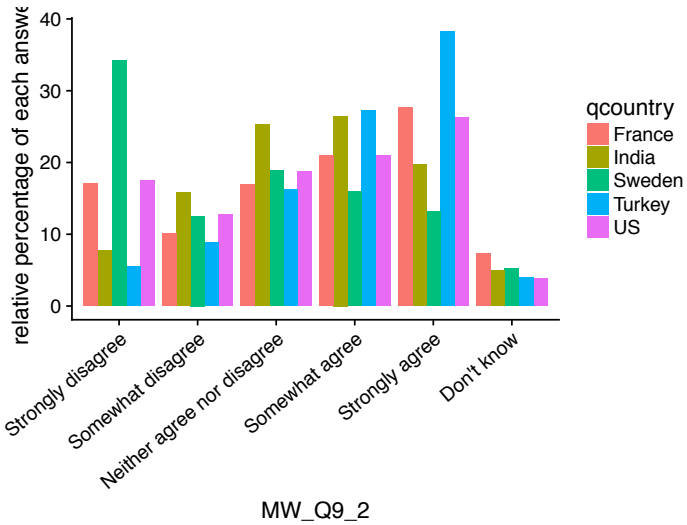
Individual country level

I plotted the results by each question on a bar plot on the scale of **percentage rather than count** (because different countries have different sample numbers) about 5 countries (US, Sweden, India, Turkey & France) with each of them from one cluster.

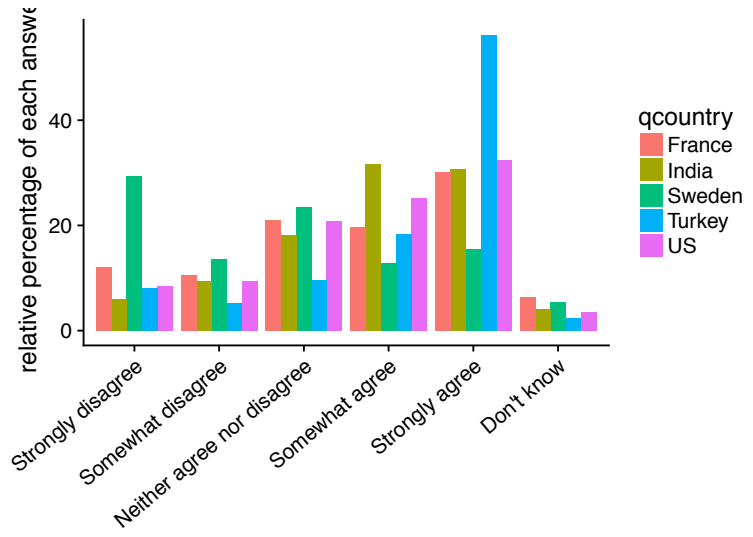
Below are the results. The plots show the relative percentage of each answer within each country and also the comparisons between countries.



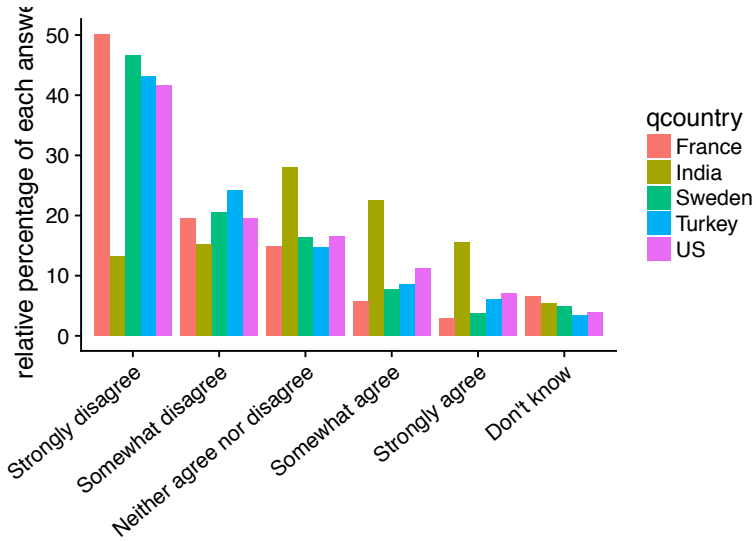
MW_Q9_1



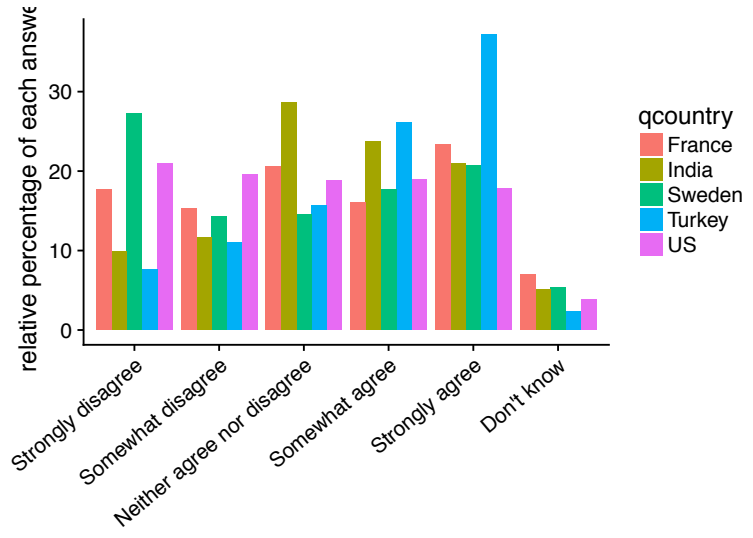
MW_Q9_2



MW_Q9_3



MW_Q9_4



MW_Q9_5