The findings of the World Risk Poll are based on over 150,000 interviews by Gallup in 142 countries including places where little or no official data exists, giving a voice for the first time to millions of the world's citizens.

The poll provides us with the first ever complete picture of the differences between people's thoughts about, and experiences of risk. This insight will be used by regulators, businesses, governments and researchers to develop relevant and relatable policies that empower people to take action, that saves lives and help them feel safer.

Background

The above is provided by the organization that performed the poll. I wanted to look at the data myself and see what techniques I know to research the relationship between what people feel puts them at risk and their personal experiences with harm from those risks.

Questions to be Answered

I also wanted to look at whether an experience of someone they know (either family or co-worker) experiencing harm during the past few years increases their level of worry.

In addition, I hoped to compare the financial factors (both personal and countrywide financial issues) which increase a feeling of being at "at risk.."

The process that I used to try see if any particularly worries or past experiences contribute to current feeling of risk were a bit convoluted. I started with some EDA, then moved to modeling the data. But that modeling found no particularly predictive questions. So I returned to more EDA, and will run further models before I feel like I have finished exploring the possibilities in this data.

Variables

A variety of demographic data was gathered, including:

- Country and Global Region
- Country income level
- Age, gender, and education of individual
- Family income level and individual's financial comfort level
- Household size and location (urban/rural)

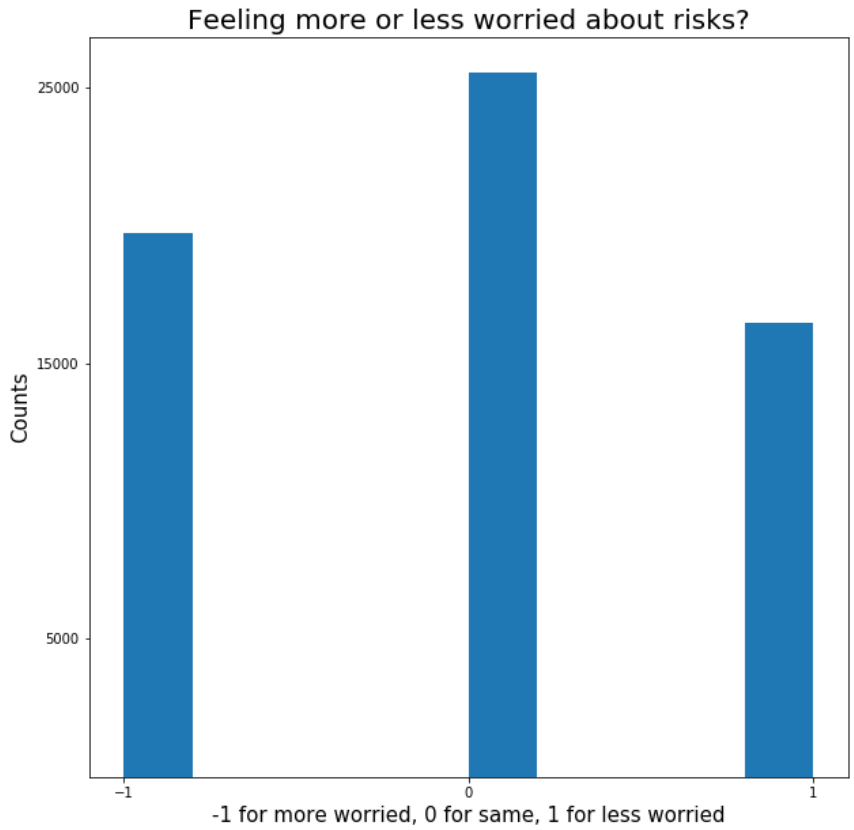
The question that became my dependent variable is:
Compared to 5 years ago, do you feel more safe, less safe, or just as safe?

Process/Plan

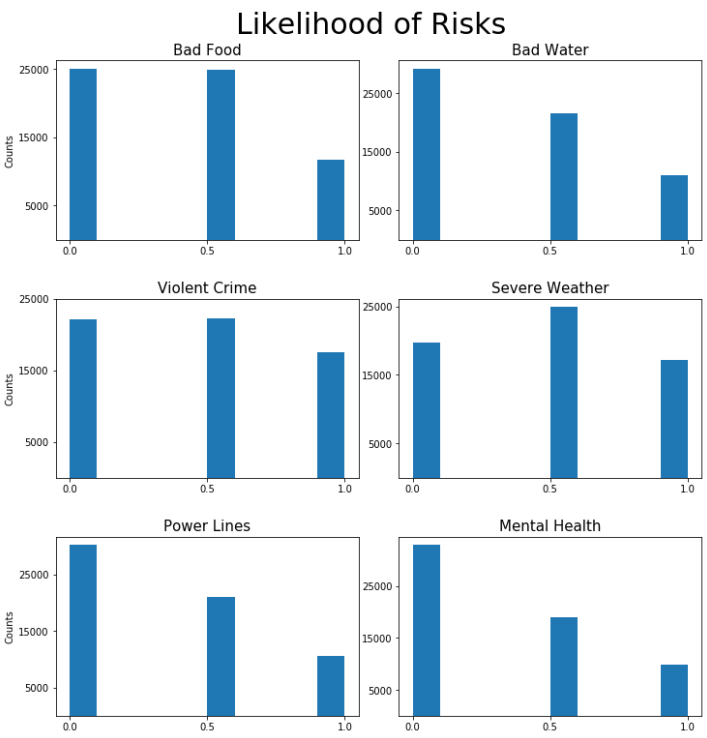
First, I will load and clean the data, removing lines with invalid or missing responses. Then I will convert the responses collected into formats that will be easily handled by the analyses that I will run.

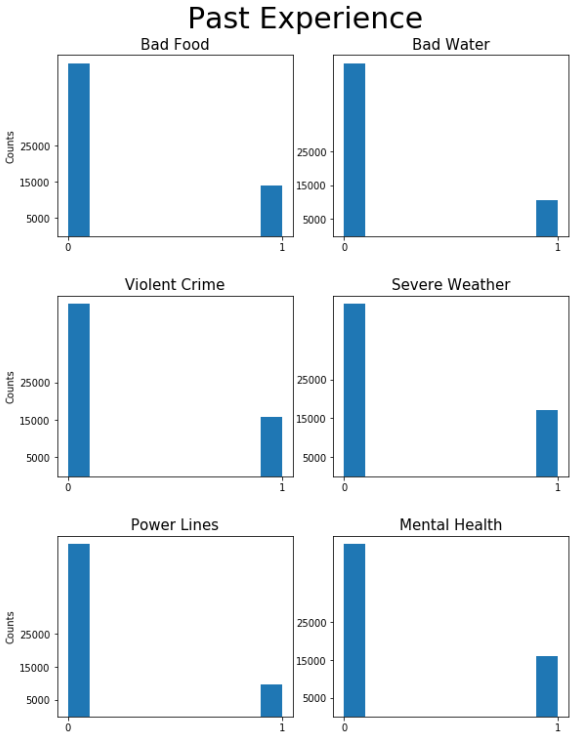
My **graph analysis** will include the following:

- How many people feel safer now than 5 years ago (histograph)?

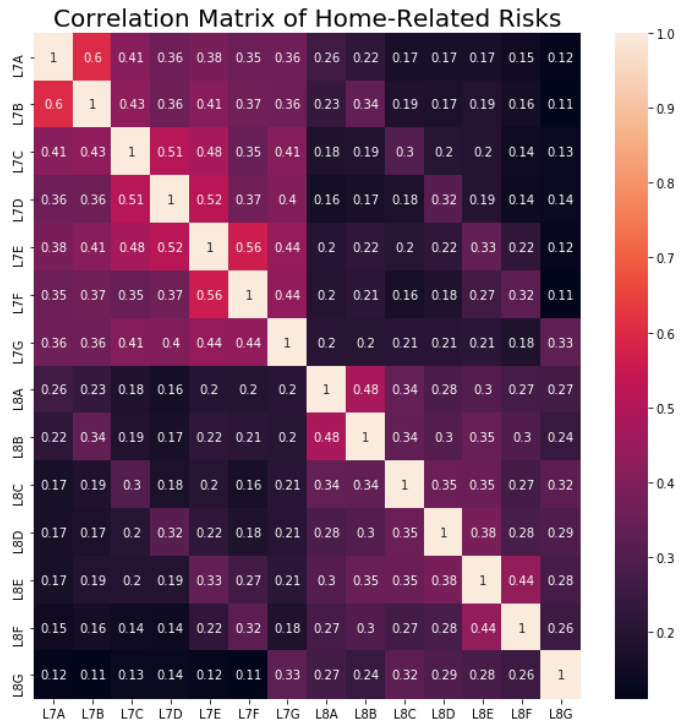


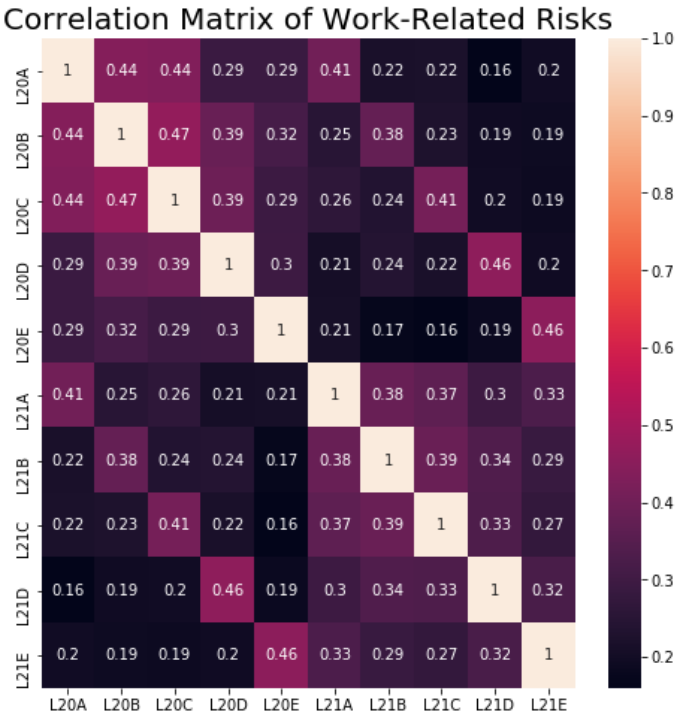
- Which factors have the highest number of people who feel that they are at risk from them (histograph) and what is their experience with people actually being harmed from those risks (histograph)?





- Is thinking a risk is likely correlated with having had someone near them harmed by it (correlation matrix)?





Evaluating the features for **dimensionality reduction** via PCA, LDA and recursive feature elimination using cross-validation (RFECV) showed that there was little opportunity for dimension reduction in this dataset.

Before running any further analysis on the data, I wanted to look deeper into the types of questions asked. They're all about risks, but they look at those risks in three different ways. Each question fits one classification in each category.

Time

Past – event already happened

Future – event may or may not happen

Source

Work – risks at work (falling, working with heavy equipment, etc.)

Home – risks at home (appliances, bad food or water, etc.)

External – risks not at home or work (car accident, lightning strike, etc.)

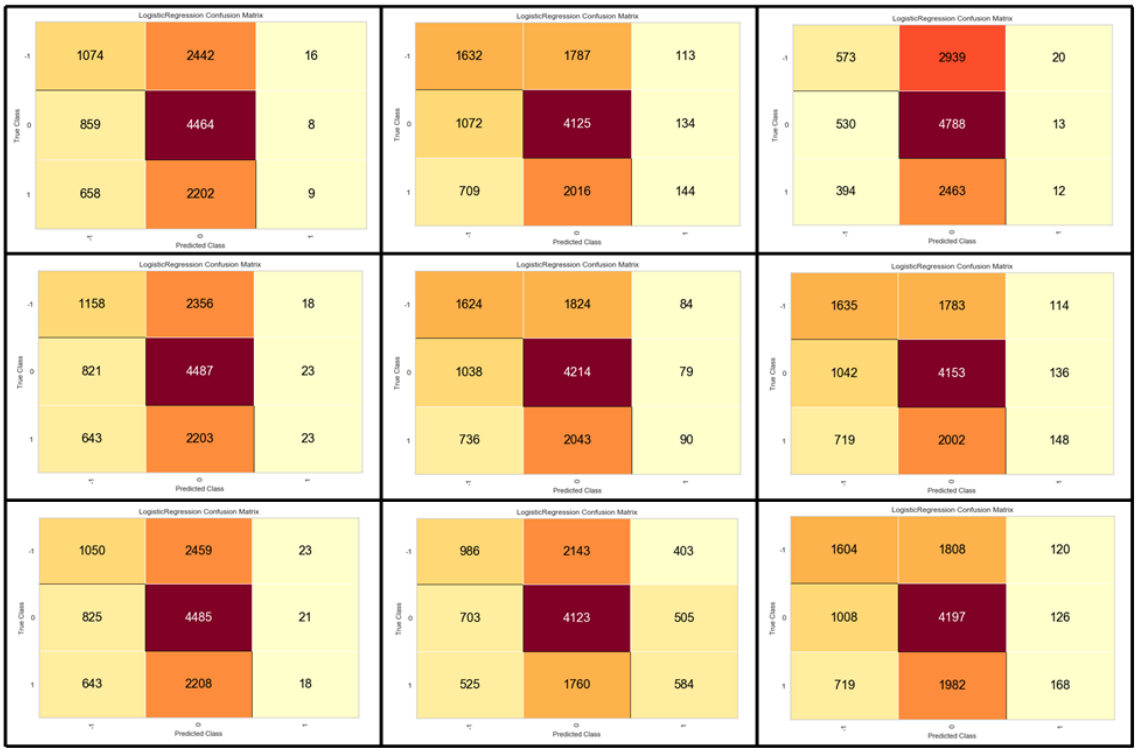
Target

Self – person hurt or at risk is person answering the survey

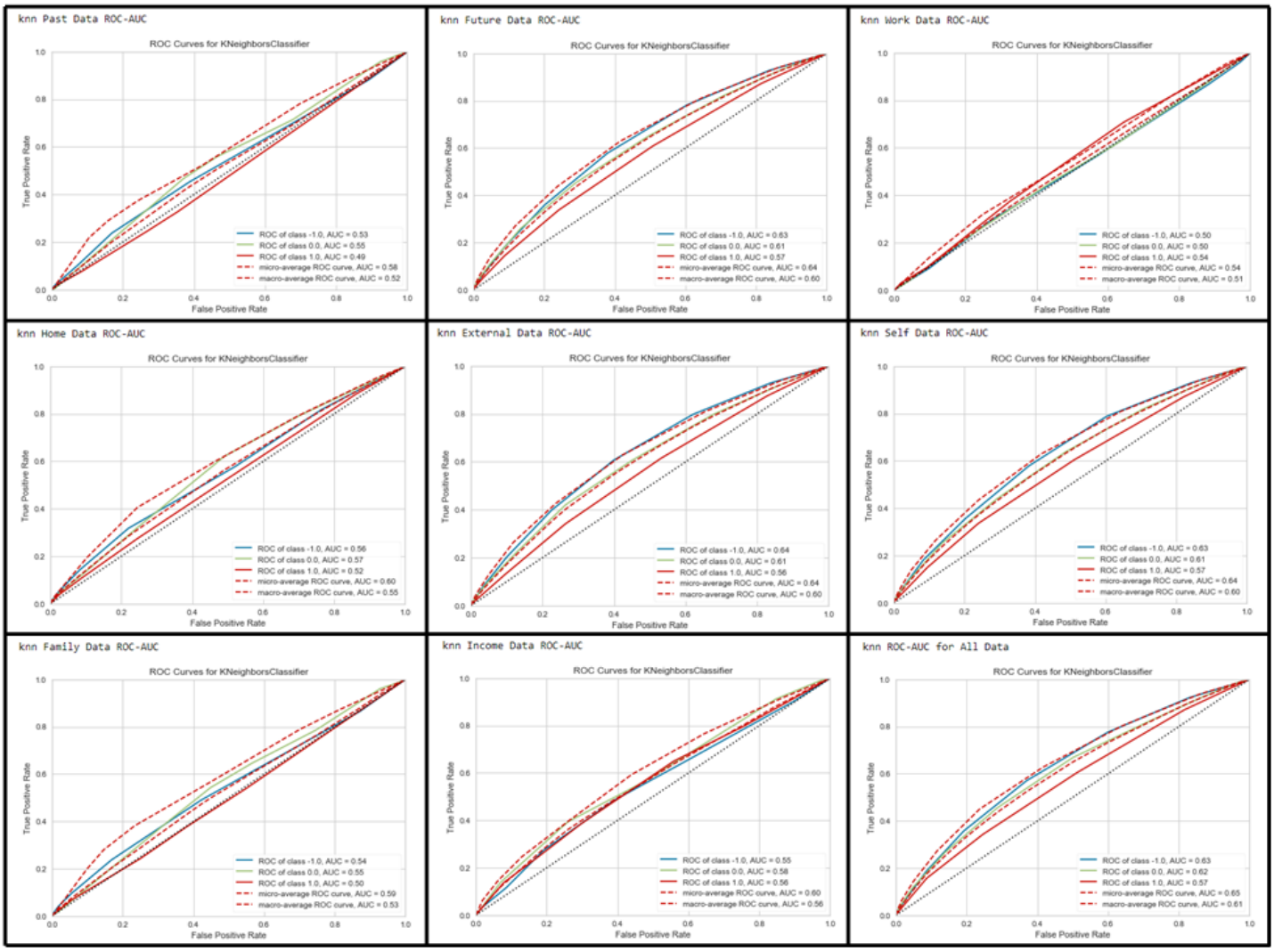
Family or Co-Worker – person hurt or at risk is known to person answering the survey

So we might have a question like, "Is working with heavy machinery a source of risk to your personal safety at work?" which would fall into the future, work, and self categories.

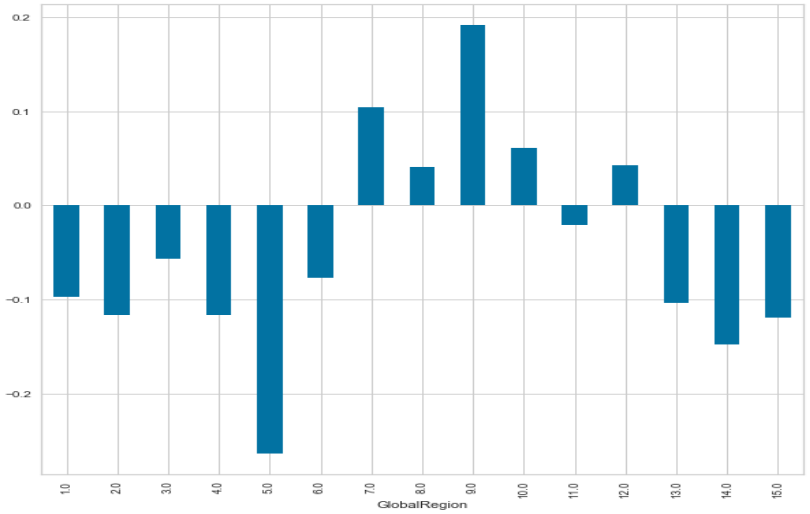
With a dependent variable that is effectively three numerical classes, I felt that a linear regression might give me some information. I ran it on the data sets separated for all of the groups above, as well as one for income and one using all of the data. Then I ran confusion matrices to see if the linear regression was effective. It was not.



Then I ran a k-nearest neighbors model, using a variety of values for k. Once I had found the best possible value, I used ROC curves to evaluate the performance of the models, and found again that this model was not effective.



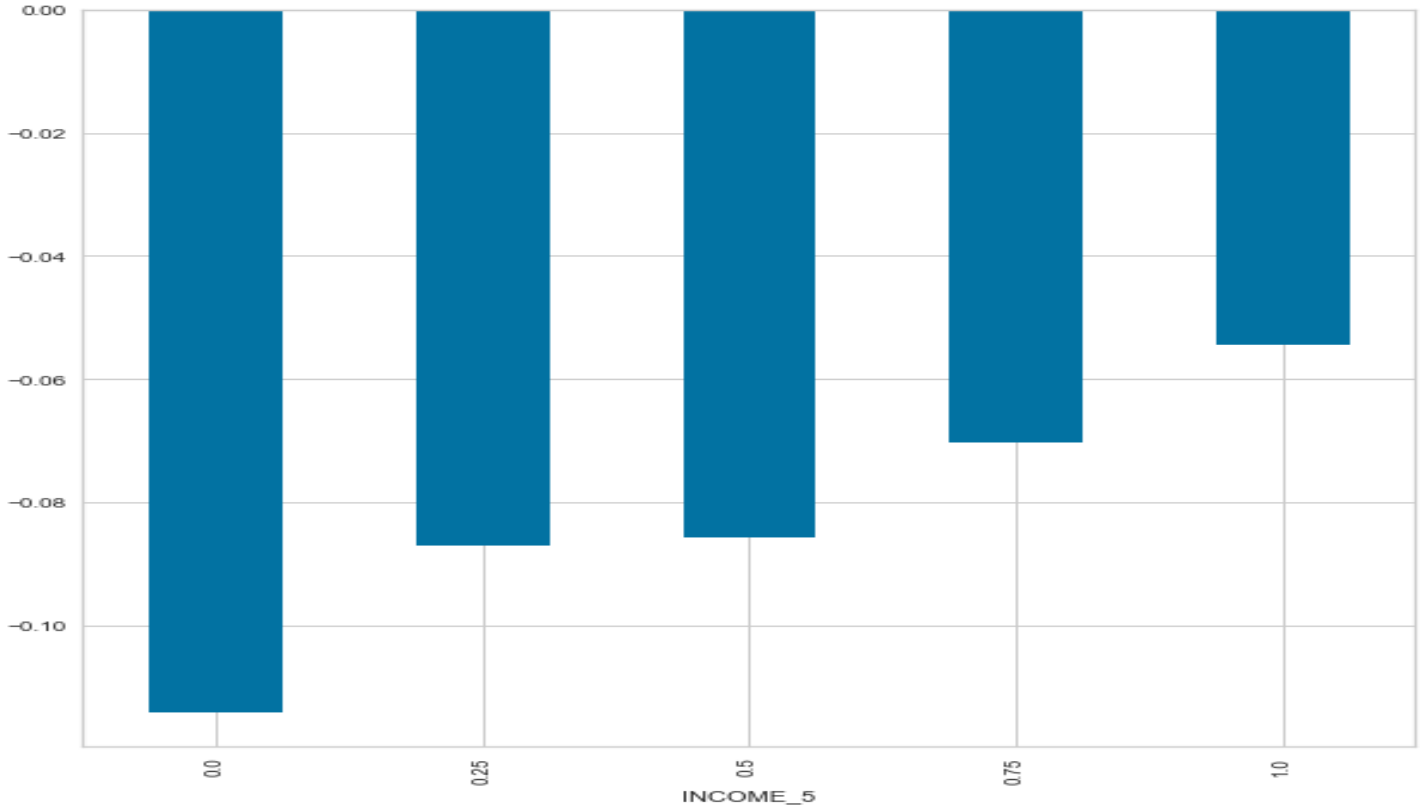
So, I went back to my EDA, and grouped the data based on answers to some of the questions, then ran bar charts comparing those answers to the mean of how "at risk" people felt (noting that –1 is less safe, 0 is as safe, and 1 is more safe than 5 years ago). For example, the region of the world vs feelings of safety looks like this:



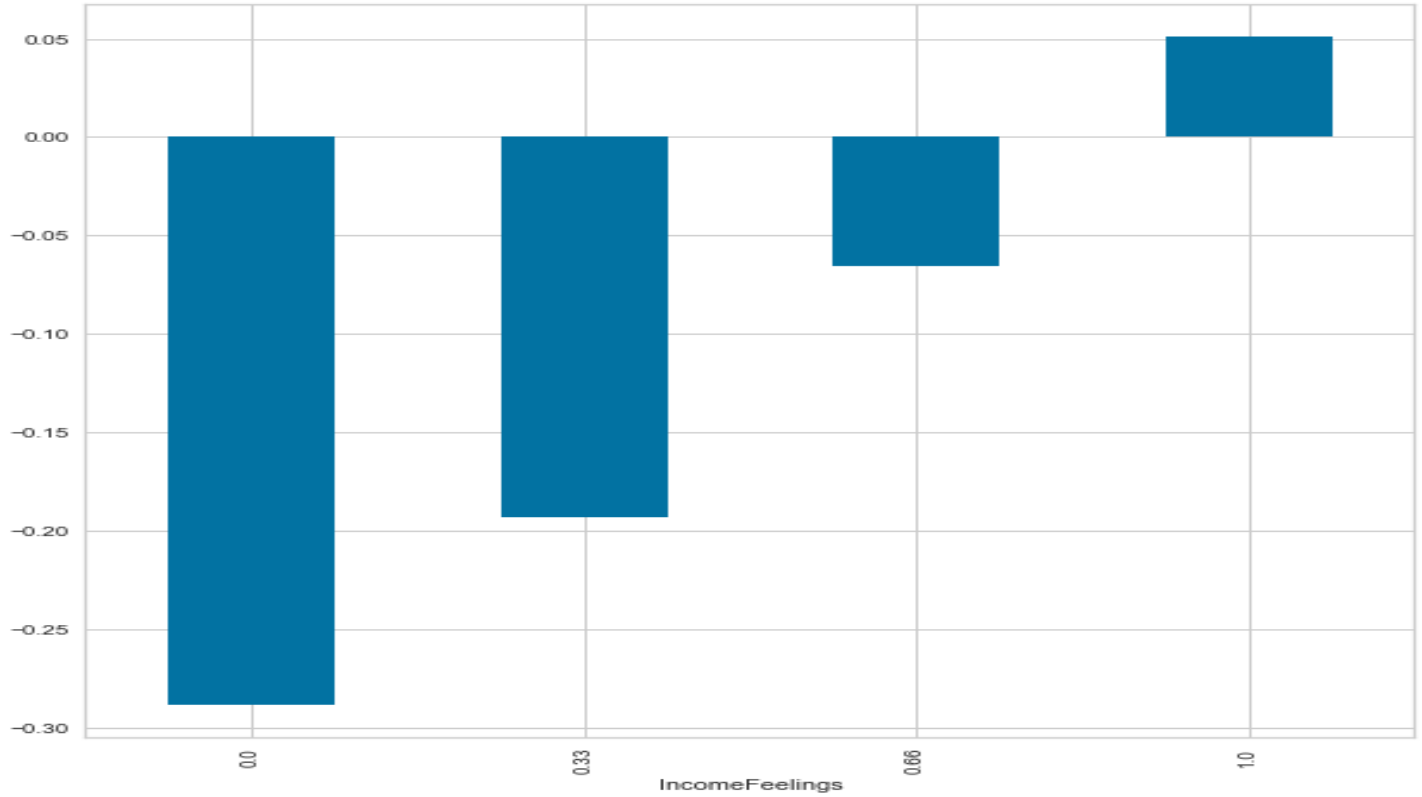
1 Eastern Africa
2 Middle/Western Africa
3 Northern Africa
4 Southern Africa
5 Latin America & Caribbean
6 Northern America
7 Central Asia
8 Eastern Asia
9 South-eastern Asia
10 Southern Asia
11 Middle East
12 Eastern Europe
13 Northern/Western Europe
14 Southern Europe
15 Australia & New Zealand

In looking at these charts, there are a few that appear to have potential, and which I will evaluated in the final step.

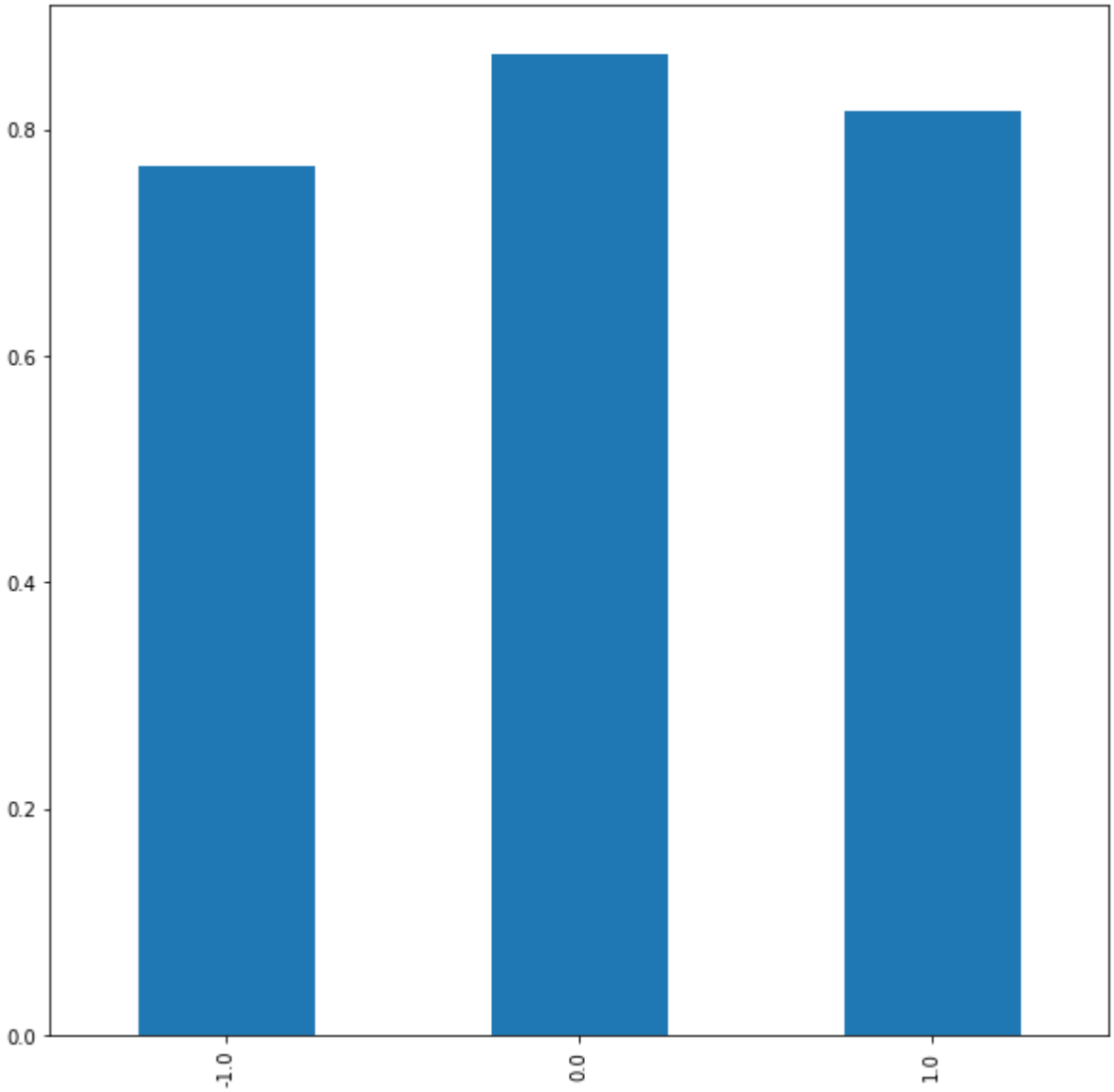
Family's income compared to country (in quartiles):



How comfortable they feel with the income they have:



Sadly, none of the others seemed to have any consistent relationship. For example, it would make sense that if the respondent or someone they know personally (family or co-worker) had been injured in the past 2 years (higher values on the y axis meaning that an injury occurred) should mean that they feel less safe (a negative number on the x axis). But that wasn't the case, since the highest value for injuries shows no change in feelings of safety:



So, in my final evaluation step, I separated the demographic information into three groups: income, non-income, and all. I then ran each group through neural net, logistic regression, k-nearest neighbors, and random forest models. I then checked the accuracy of each:

ACCURACY	Neural Net	Logistic Regression	knn	Random Forest
Income	46.37%	46.24%	41.64%	46.37%
Non-Income	48.11%	46.96%	44.40%	43.45%
All	50.79%	47.67%	45.16%	44.00%

So, in the end, none of these models appear to be very good predictors.

In summary, it appears that what people worry about, their past experiences, and income simply aren't good ways to predict whether people feel safe. Luckily, that prediction isn't what the survey was about. The survey was about finding out what worries people in different areas so that those concerns can be addressed.

In addition to the survey questions which I looked at, respondents were also asked who they would go to for safety information, particularly in regard to immediate dangers. Those questions will help local organizations find ways to get vital information to the people who really need it.

And in the end, the final goal is finding a way to reduce the risks to as many people across the world as possible.

To learn more about the poll, check out the website here: https://wrp.lrfoundation.org.uk/