Data Analytics Project: Religion

1 Introduction: why this data set?

While looking for various data sets, we wanted to find original data sets which can fit the size constraints we had. We oriented our research on social and governemental subjects. This data set is constructed from a survey about religion. We chose this one because it has many types of questions so we can practice data cleaning and make analyses on people's perspective, from different religions, regions, gender etc., on public exhibition of religious behaviors.

2 Description of the data set

The dimension of the data set is (1040,48) but we have only 47 questions and 1039 people who answered the survey. These numbers are different due to data cleaning operations. The persons who answered the survey are:

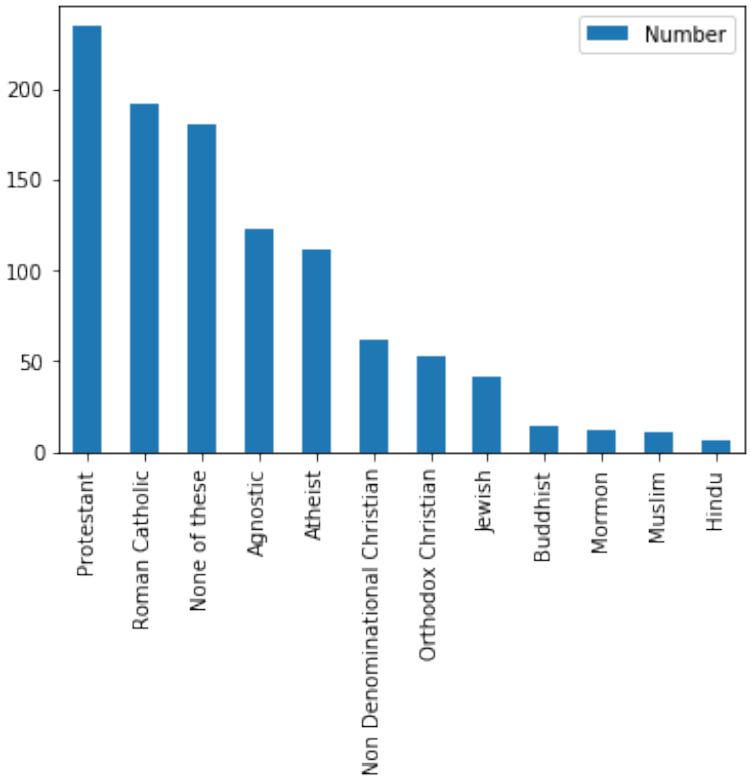

Figure 1: Distribution of religions

This data set is only representative of the USA population. For some religions like islam, there are not enough people who participated to the survey to properly represent them.

There are no open questions in this survey, except for the first question where people can write their religion if it's not in the list. So they have to choose among different types of answers, such as:

- Multiple choices
- Yes or No
- Range-based

The survey has five categories:

- Religeous affiliation
- Demographics
- Personnal religious practice
- Confort with public faith
- Being witnessed and witnessing public faith

3 Data cleaning

3.1 Remove irrelevant lines

In the data set, the first line was only constituted of "Response" in every column, which was irrelevant for our study. That's why we had 1040 rows but only 1039 people who answered the survey.

3.2 Creating new religion entries

As told before, the first question of this survey asks about the religion. We noticed that 6% of people who answered "None of these religions" responded "Non Denominational Christian". For this reason, we decided to create a new religion option. Then the other answers of the question were deleted, because the "None of these" option covers the rest. That explains why there was 48 columns in the data set but only 47 questions.

3.3 Use numerical values

We cannot properly use the data we have with the provided answers, that are in the string format. For this reason, the strings are transformed into numerical values. These numerical values can be of several types:

- Binary
  For the gender, or the answer for a yes/no question
- Scaled integers
  For different kinds of questions, which could be: "How often ...", "How comfortable do you feel when ...", age, revenue, etc.
- -1
  For question that does not apply to one's religion.

4 Exploratory analysis

For the analysis, we had a lot of options since our data set contains a lot of information. We chose to begin with a basic analysis and then to do a specific analysis of our data set: the standard deviation and a demographic analysis (on age, gender and income).

4.1 Boxplot

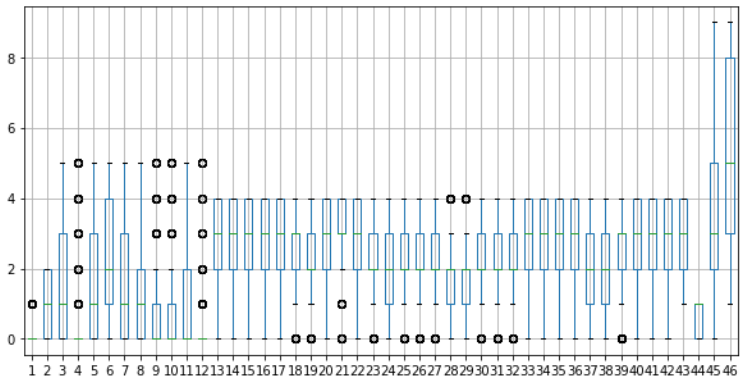

Figure 2: Boxplot

The box-plot allows us to have a really quick overall view of our data-set. Excluiding the binary answers, we can see that most of the questions are well distributed and spread among all possible values. This is a good fact for the analysis, the response are correclty spread along all possible answers values without extreme variations. The only exessively spreaded value are the location and the money earn by the surveyed person. In our case, it's good because it means that the survey represent a general population.

4.2 Correlation Matrix

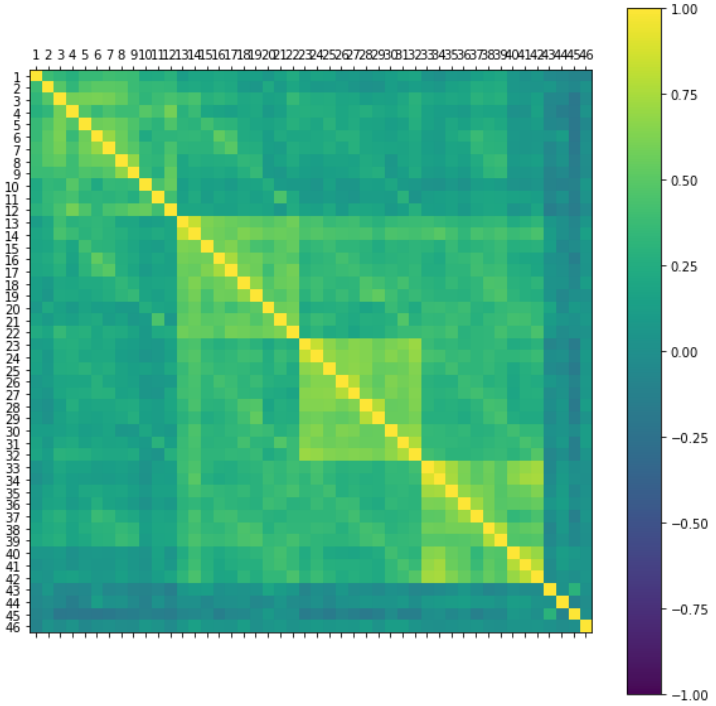

Figure 3: Correlation Matrix

Thanks to the correlation matrix, we can see that the answers to questions influence each other in four main groups: "How often do you do it?", "How comfortable are you when you do it?", "How comfortable do you think others are when you do it ?" and "How comfortable do you feel when others do it?".

The rest is not really relevant like age / income or age / location.

This helps us to understand the four main axis of questions, sort of a base space of all the dataset.

4.3 Biplot

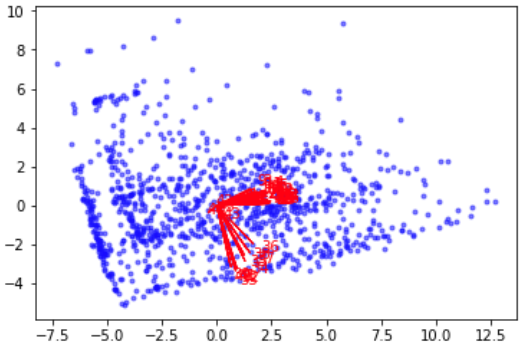

Figure 4: Biplot

Thanks to the PCA analysis, we can see that all questions have their own importance (we need at least 20 questions to reach 90% of the cumulative variance). That means the dataset is still complex and cannot be resumed in only few simple questions.

However, by doing the 2D-biplot, we can see that all questions group into two axes. We can then think that two tendencies exist in this

dataset and compose a base space allowing to represent it (loosing some information but giving a good base). If we link that to the correlation matrix, we can understand that "How comfortable are you when you do it ?" and "How comfortable do you feel when others do it?" are surely the two underlying vectors which allows to contruct the answer space, which we retrieve with the pca and biplot.

4.4 Standard deviation

We searched the question with the greatest standard deviation because a high standard deviation means that the answers are more spreaded, so it might be because of the different religions' practices. That's why it was relevant to study this question which is the third one: How often do you pray in public with visible motions (sign of the cross, bowing, prostration, shokeling, etc)?

First, we can see that none of the Atheists answered because they don't pray. Second, only four religions over the eleven remaining are under the general mean. Especially Buddhist, Agnostic and Jewish people pray in public with visible motion nearly never, at least less than once a year. On the contrary, Muslims do it a lot: many times per week. Then, Hindu, Mormon, Orthodox Christian and Roman Catholic people do it a few times per month. The other religions are in the mean and pray in public with visible motion a few times a year.

4.5 Age

Contrary to what we might think, old people are the ones who are the most discreet about the practice of their religion: they are uncomfortable talking about the other's religion, they wear religious clothes less than young people and don't like to pray in public or to participate in a public religious event on the streets (they do it half as much as young people do).

However, they are more likely to accept food or beverage they should not consume for religious reasons: the older you are, the easier you accept. But, according to the survey, when somebody else from an other religion declines some food, the youngest and the oldest people are both less disturbed by it than the middle aged people are.

The last huge difference between young and old people is about praying for someone else. The young people don't like to say someone they will pray for him or her, they also think that it would disturb the other people as well. Also, that would disturb them if they see someone offering to pray for someone else.
For the other discussions, no obvious trend was detected.

4.6 Gender

Regarding the mean of answers given, men are quite more likely to display some religious behaviours on public than women such as praying with physical objects/ visible

motions, or praying aloud before meals in the presence of people who don't belong to their religion.

However, as an exception of showing signs of religion mentioned, women more often tell someone they'll pray for them.

On the contrary, when it comes to feeling comfort while they see other people displaying public religious behaviours, such as praying with physical objects/ visible motions, or praying aloud before meals in the presence of people who don't belong to that religion or telling someone to pray for them, women seem distinguishably more comfortable than men in such cases.

For the other questions, no obvious difference was detected.

4.7 Income

Due to the mean of answers given, less incomed people seems more evangelical. Although there is no linear obvious relation between all income ranges and attending religous services, we can say that people with smallest income are likely to attend religious services more than people with greatest income. Less-incomed distinguishably pray with visible motions and objects than others, and are more comfortable to pray aloud before meal. Plus, there exists an inverse relation between income and participating public religious event on the streets.

Less-incomed people are more comfortable when they are withnessed while praying in public with visible motion, when bring up their religion or when they are offered to pray for. On the contrary, people with greater income are slightly more comfortable when they decline food or beverage.

For the other questions, no obvious trend was detected.

5 Unsupervised learning

As an unsupervised learning technique, we used silhouette analysis on KMeans clustering which provides a visual representation of clusters with unlabel data. We chose different K variables, from 2 to 20, to define the number of groups in the plot.

Then, we compared them to decide which one was the best representation of our data set. For the comparison, the greatest average silhouette score, which quantifies the quality of clustering achieved, and the similar size of silhouette plots are desired.

The greatest average silhouette score we had was 0.28670681932 with two clusters (n clusters=2). Also, the clusters have the similar thicknesses.However, even the highest silhouette score is pretty low. So, we can say that our data seems to be quite unstructured and nonhomogenous. But in order to reach some theoretical conclusions from our clustering, further researches must be conducted.

6 Supervised learning : neural network

We chose the easiest architecture, a multilayer perceptron, which is represented on the picture bellow. This kind of network is enough to compute a basic classification, and small enough to be easily implemented. For this purpose, we used the framework keras which relies on Tensorflow. These frameworks allow a quick and easy construction of deep learning methods.

For the supervised learning, we wanted to train a neural network over our data set to make it learn the type of religion from the 46 questions. The goal is to answer the 46 questions and let the neural network guess the religion.

For the crossvalidation, Keras does it automatically, at each epoch. It takes apart a small part of the dataset (randomly), trains on the rest and then validates on the small part taken aside.

Finally, we had a dropout layer which delete randomly a node in the network for one epoch. This allows the network to be more robust to over-fitting (because it forces the network to learn by other ways and not "learning by heart").

We could achieve a score of +90% on our dataset, the only limit is induced by the dataset himself because we don't have enough datas on muslims and hindus to correctly train on that.

7 Conclusions

For this project, we started with data cleaning operations to be able to work on it, then we made some exploratory analysis regarding different attributes and representations to gain more insight about the behaviors within the set. Then, we applied an unsupervised learning technique to understand the nature of clustering of the data. Finally, we applied supervised learning which was the most interesting part. The objective was to guess the religion of a new participant by her/his answers to the survey. Although the accuracy was quite high, it is solely representative of US. In order to have a more applicable program, we would need a global and diverse data set.

Moreover, we can modify our program to guess of some other characteristics such as gender, location, income, age. For the further studies, one can use this program to correlate other characteristics to religion such as education level, political opinion, scientific believes.