United States and Science Values 2006 - 2011

IS606 Final Project

Daniel Dittenhafer

December 7, 2015

Part 1 - Introduction:

How have the values of people from the United States changed over time with regard to science and technology and its positive/negative impact on the world? The World Values Survey includes a question where in the survey respondent is asked to characterize their view of science and the state of the world because of it (World Values Survey Association, 2014). Using the World Values Survey longitudinal data from 2006 and 2011, this project investigates changes in the views of United States respondents associated with science as well as correlations with educational level and the importance of religion.

Part 2 - Data:

The World Values Survey is an observational study, and this data project will be analyzing a subset of the observational study conducted by the World Values Survey Association. The data was collected and made available by the World Values Survey Association via their website. (World Values Survey Association, 2014).

The original data set in its entirety is available from the World Values Survey website: http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp

Although the World Values Survey was conducted in the United States starting in 1995, the question regarding science and the world (005_203) was not introducted until 2006. As such, only the cases from 2006 and beyond, 3481 cases across 2006 and 2011, are considered in this study.

Scope of Inferencce

Based on the data, collection methods and type of study, we review the scope of inference this data project can claim.

Generalizability: The population associated with the survey in this data project is the population of the United States between the ages of 18 and 85. Any results or conclusions from this project should not be generalized outside this population.

Causality: The data is derived from an observational survey and not an experiment. As such, no causal relationship can be established from any conclusion.

Variables Included in Data Subset

The following table lists the variables extracted from the original World Values Survey dataset which will be considered in this data project.

ID	Variable	Description
005_203	E234	The world is better off, or worse off, because of science and technology
010_023	S020	Year survey
010_028	S024	Country wave

ID	Variable	Description
010_004	S003	Country/region
014_003	X003	Age
014_030	X025	Highest educational level attained
001_006	A006	Important in life: Religion

Response Variable

The response variable is the answer provided to the key question, "Is the world is better off, or worse off, because of science and technology?"

The answers are categorical in nature, but ordinal in their degree of support for the better/worse outcome. The distinct answer values and the description of each value are shown in the table below. Note that negative values are variations on missing data. These will be eliminated during the exploratory data phase.

Value	Description
1	A lot worse off
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	A lot better off
-5	Missing; Unknown
-4	Not asked in survey
-3	Not applicable
-2	No answer
-1	Don´t know

Explanatory Variable(s)

The explanatory variables considered for analysis are the "Highest educational level attained" value and the "Important in life: Religion" variable. Again, these are categorical variables, but ordinal in terms of level of education and degree of importance of religion. The values and descriptions for each variable follow:

Highest educational level attained

Value	Description
1	Inadequately completed elementary education
2	Completed (compulsory) elementary education
3	Incomplete secondary school: technical/vocational type/elementary education, basic vocational qual.
4	Complete secondary school: technical/vocational type/Secondary, intermediate vocational qualification
5	Incomplete secondary: university-preparatory type/Secondary, intermediate general qualification
6	Complete secondary: university-preparatory type/Full secondary, maturity level certificate
7	Some university without degree/Higher education - lower-level tertiary certificate
8	University with degree/Higher education - upper-level tertiary certificate
-5	Missing; Unknown
-4	Not asked in survey

Value	Description
-3	Not applicable; No formal education
-2	No answer
-1	Don´t know

Important in life: Religion

Value	Description
-5	Missing; Unknown
-4	Not asked in survey
-3	Not applicable
-2	No answer
-1	Don´t know
1	Very important
2	Rather important
3	Not very important
4	Not at all important

Part 3 - Exploratory data analysis:

Let us explore the variables within the data subset. First, some simple summary statistics:

```
##   KeyQuestion          Year       CountryWave      Country
##  Min.   :-5.000   Min.   :1995   Min.   :8403   Min.   :840
##  1st Qu.:-4.000   1st Qu.:1999   1st Qu.:8404   1st Qu.:840
##  Median : 5.000   Median :2006   Median :8405   Median :840
##  Mean   : 2.236   Mean   :2004   Mean   :8405   Mean   :840
##  3rd Qu.: 8.000   3rd Qu.:2011   3rd Qu.:8406   3rd Qu.:840
##  Max.   :10.000   Max.   :2011   Max.   :8406   Max.   :840
##       Age        HighestEducation ReligionImportant CountryName
##  Min.   :-1.00   Min.   :-3.000   Min.   :-2.0      Length:6223
##  1st Qu.:33.00   1st Qu.: 4.000   1st Qu.: 1.0      Class :character
##  Median :46.00   Median : 6.000   Median : 1.0      Mode  :character
##  Mean   :47.15   Mean   : 5.771   Mean   : 1.8
##  3rd Qu.:61.00   3rd Qu.: 8.000   3rd Qu.: 2.0
##  Max.   :94.00   Max.   : 8.000   Max.   : 4.0
```

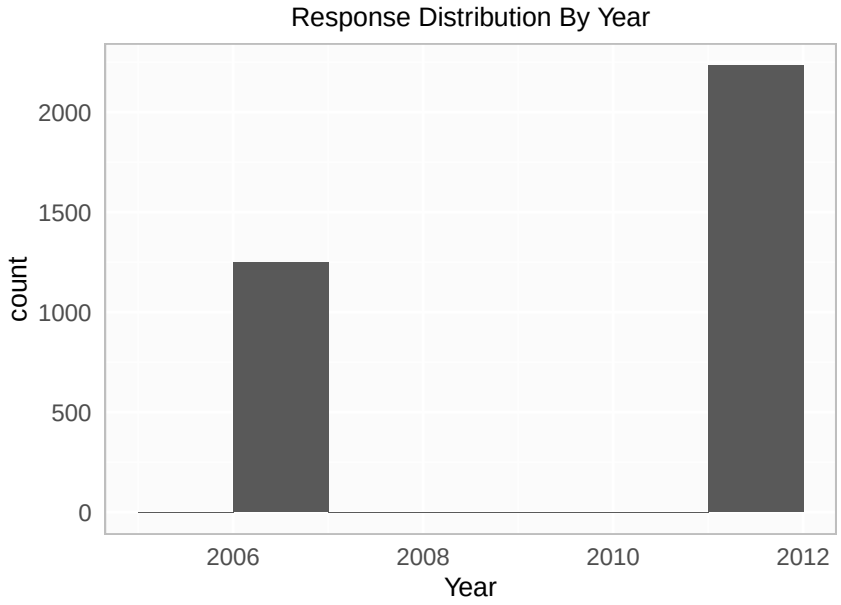
The following table shows the mean and standard deviation of the response variable for the various years for which the survey was conducted. Unfortunately, in 1995 and 1999 the question regarding science was not asked as evidenced by the mean of -4 ("Not asked in survey") and standard deviation of 0 (no variation). As a result, the focus will be on differences between 2006 and 2011.

Year	Mean	Std Dev
1995	-4.000000	0.000000
1999	-4.000000	0.000000
2006	6.885508	2.671164
2011	7.294803	2.483114

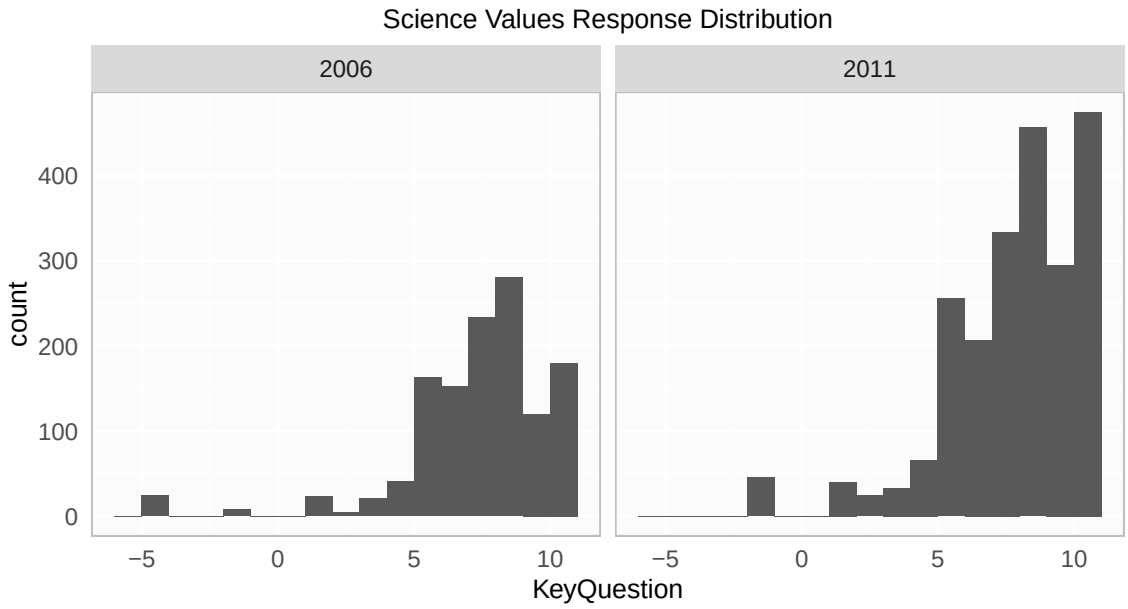
After eliminating the 1995 and 1999 data rows, 3481 cases remain. Of those, 35.88% are in 2006, and 64.12% are in 2011. The number of cases in the sample is less than 10% of the United States population in 2006 299,398,484 (United States Census Bureau, 2006). The same holds true for the 2011 population at 311,591,917 (United States Census Bureau, 2011).

Year	Cases	Ratio
2006	1249	35.88049
2011	2232	64.11951
Total	3481	100.00000

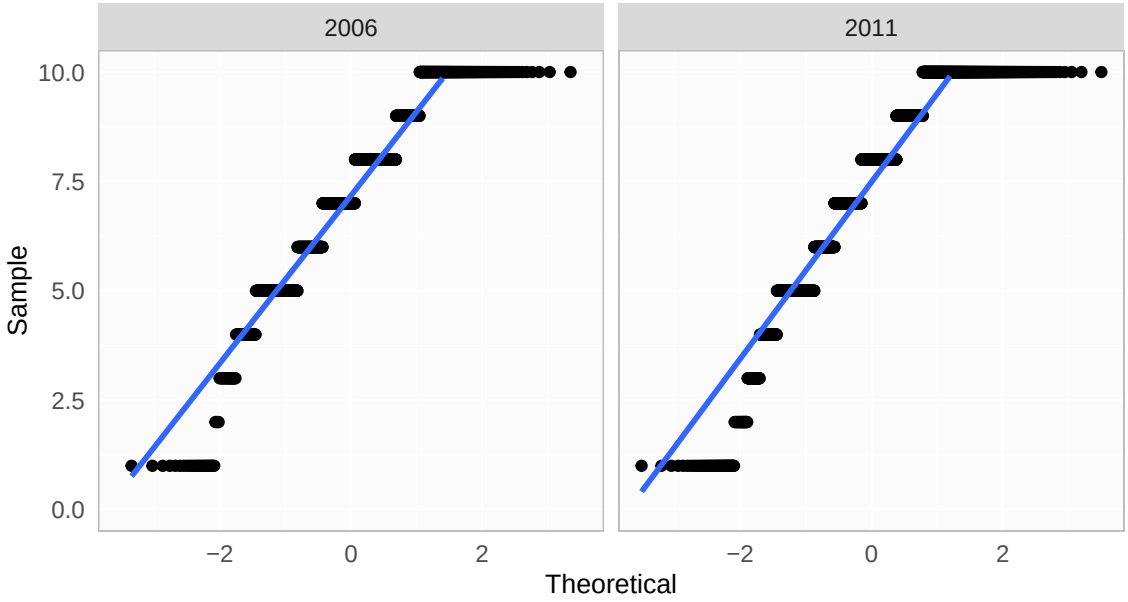
```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.
```



When viewing the data distribution of answer values for our `KeyQuestion` in each year, we see different distributions, as shown in the following Science Values Response Distribution histograms. Both years are skewed left, but 2006 appears a bit more normal where 2011 could be characterized as bi-modal at values 8 and 10. We see also that ~2.56% of 2006 answers and ~2.02% of 2011 answers are negative and constitute missing values. These data rows will be removed shortly.

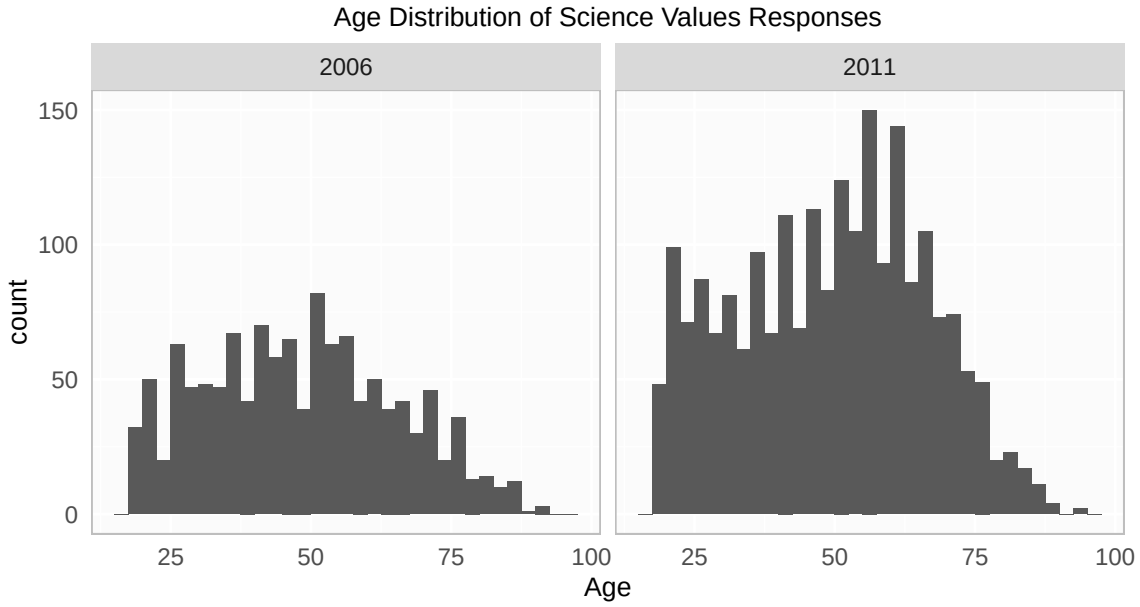


The missing/unknown response data rows have now been removed and we further explore the data. A Quartile-Quartile plot will be used to check the normalness of these distributions in more detail. As shown in the following charts, the distributions are not strictly normal, but for our purposes we will accept them as nearly normal. This allows us to proceed with the difference of two means analysis in Part 4 - Inference.



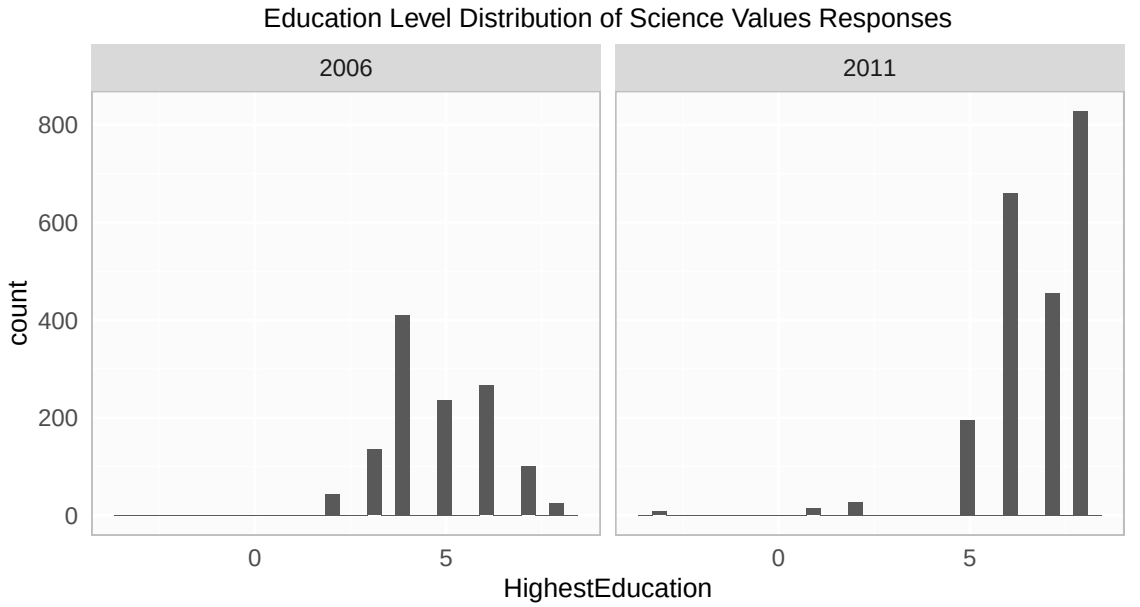
Age

In the following charts we see the age distribution of respondents to the science question. Both years are somewhat normally distributed but with a right skew. Older people are less represented in the data particularlly over the age of 75. Obviously we have more responses in 2011 (which matches the Response by Year result seen earlier).



Education Level

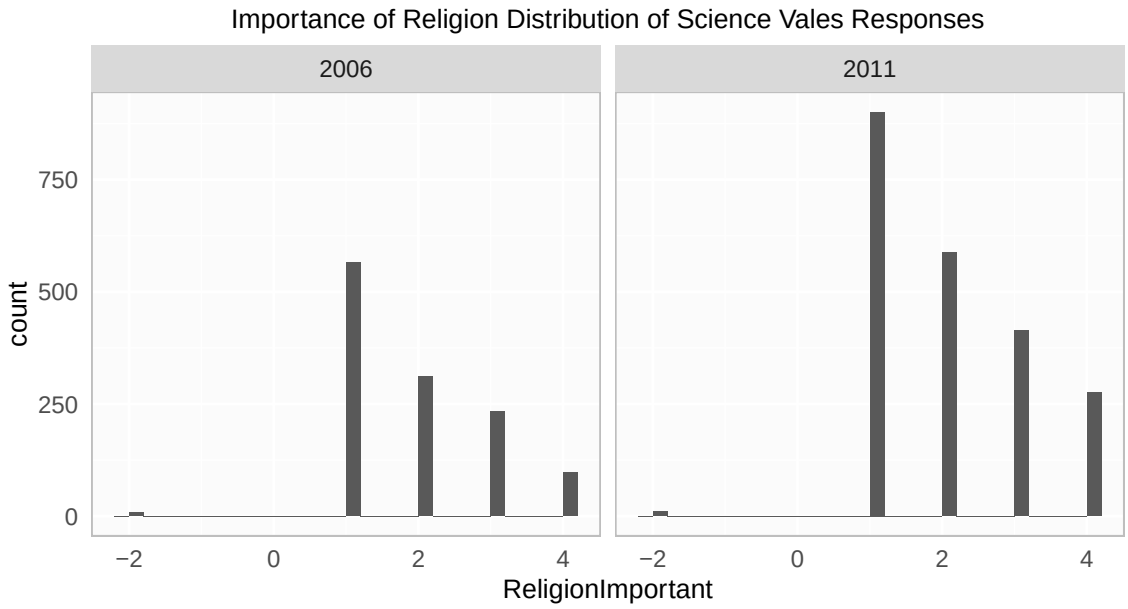
Now lets explore the education level of the science values respondents. Recall that the highest level in this survey is 8 - "University with degree/Higher education - upper-level tertiary certificate"", and -3 ="Not applicable; No formal education".



The distributions shown in the previous charts raise some questions. Did the education level distribution change that much in 5 years? What bias's might be introduced with these education levels? Unfortunately, these questions are out of scope for this data project.

Importance of Religion

Now lets explore the importance of religion to the science values respondents. As shown in the following charts, the distributions are visually very similar, but certainly not normally shaped.



Part 4 - Inference:

Our initial interest relates to differences between years for the response variable. Is there a statistically significant difference between 2006 and 2011 for the response to the science and the world question?

The following table lists the mean and standard deviation for our cleansed data set per our exploratory data analysis (Part 3).

Year	Mean	Std Dev	n
2006	7.178307	1.982304	1217
2011	7.486054	2.116065	2187

Our hypothesis test is set up as follows. The null hypothesis states that there is no difference between the population means, whereas the alternative hypothesis is that there is a difference.

$H_0 : \mu_{2006} - \mu_{2011} = 0$

$H_a : \mu_{2006} - \mu_{2011} \neq 0$

$\alpha = 0.05$

The point estimate for the difference in means is -0.3077466.

$$\bar{x}_{2006} - \bar{x}_{2011} = -0.3077466$$

The standard error of the point estimate becomes:

$$SE_{\bar{x}_{2011} - \bar{x}_{2006}} = \sqrt{\frac{\sigma_{2011}^2}{n_{2011}} + \frac{\sigma_{2006}^2}{n_{2006}}} = 0.0726381$$

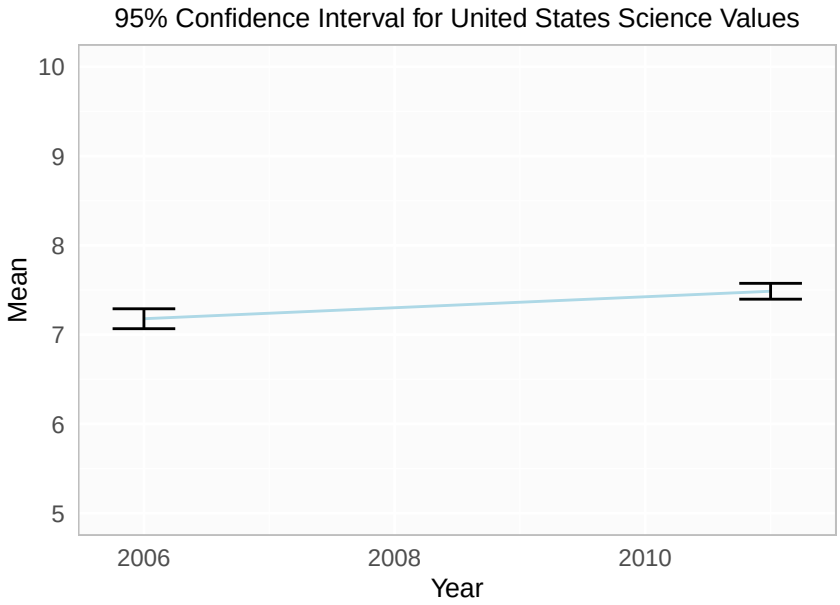
The T-score is computed as:

$$T = \frac{-0.3077466 - 0}{0.0726381} = -4.2367097$$

Using degrees of freedom based on the smaller of $n_{2006} - 1 = 1216$ vs $n_{2011} - 1 = 2186$: df=1216 we get a p-value $\approx 0 < 0.05$. Therefore, we reject the null hypothesis and conclude that the United States' mean view of the world being better off because of science and technology has *increased* from 2006 to 2011.

We can construct a 95% confidence interval around the response variable's means and visualize it to get a sense of the change and the range of probable population values.

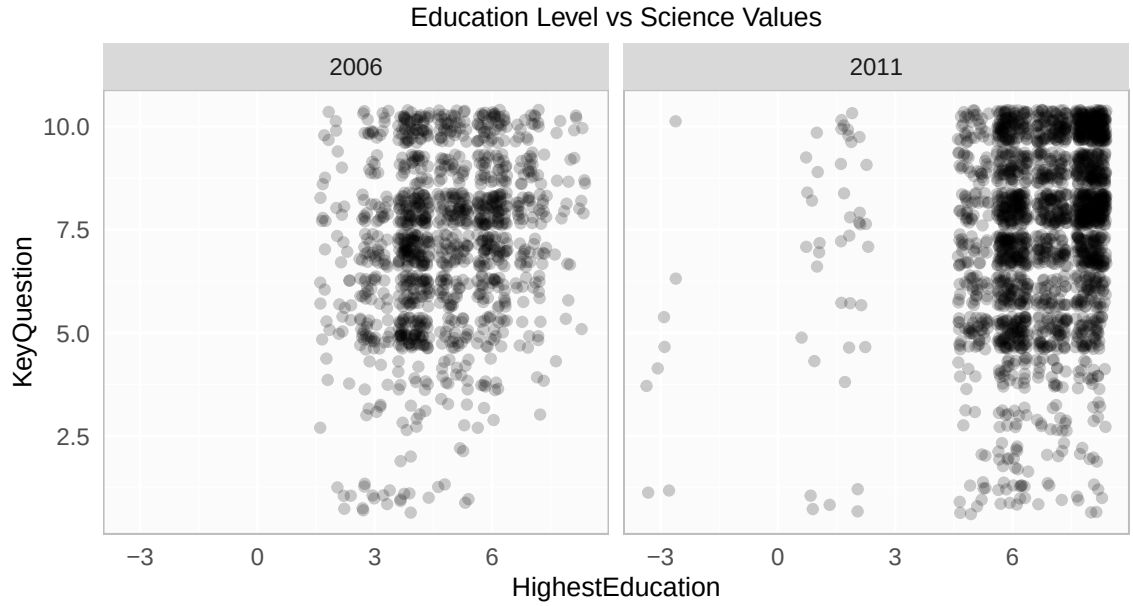
Year	Mean	Std Dev	n	LowerBound	UpperBound
2006	7.178307	1.982304	1217	7.066825	7.289790
2011	7.486054	2.116065	2187	7.397319	7.574789



The chart above helps illustrate the change in science values between 2006 and 2011. While subtle, the change is distinct and significant.

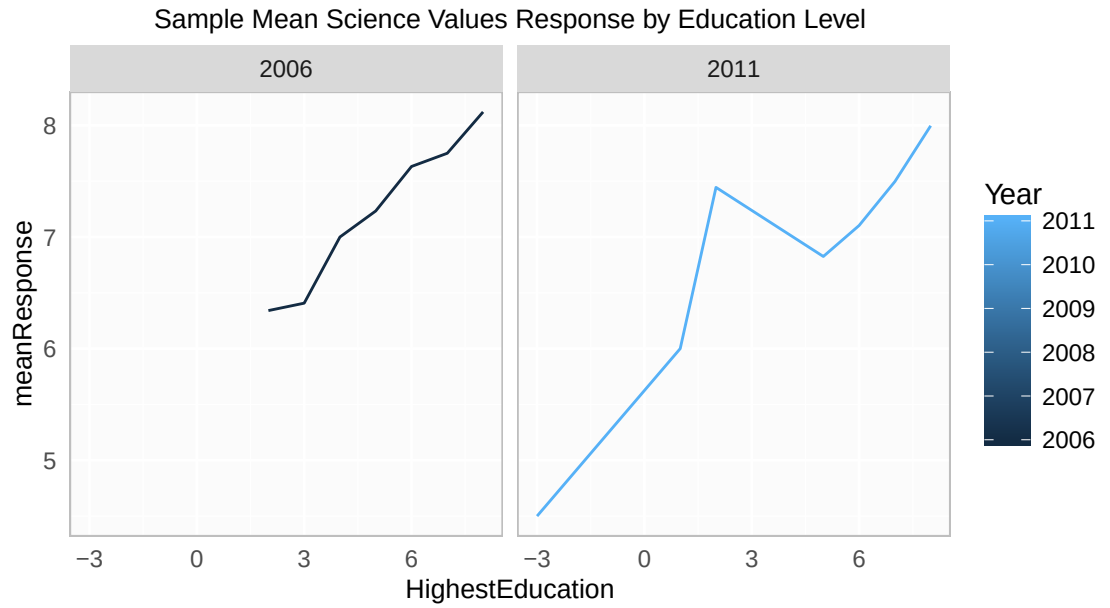
Education Level Correlation

Now lets focus on what factors might be related to the belief that science is making the world better off. We'll start out looking at education level through a simple scatter plot. As can be seen in the plots, there doesn't appear to be any dominate relationship though we can see evidence of the data distribution we observed during the exploratory phase.



What if we summarize the response variable by the education level as in the table below? When visualizing these means by educational level, there does appear to be a positive relationship.

HighestEducation	Year	meanResponse	sdResponse	nResponse
2	2006	6.340909	2.514545	44
3	2006	6.407407	2.312511	135
4	2006	7.000000	1.929236	411
5	2006	7.233051	2.044382	236
6	2006	7.631579	1.650494	266
7	2006	7.750000	1.604130	100
8	2006	8.120000	1.715615	25
-3	2011	4.500000	2.878492	8
1	2011	6.000000	3.113247	14
2	2011	7.444444	2.577019	27
5	2011	6.825641	2.322094	195
6	2011	7.103030	2.160128	660
7	2011	7.496703	2.108168	455
8	2011	7.996377	1.837731	828



Our hypotheses for an analysis of variance (ANOVA) test as to whether the means for each education level are the same or different for a given year are as follows:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8$ (The mean science values response is the same across all education levels.)

H_a : At least one mean is different.

$\alpha = 0.05$

First we will start with the 2006 data. The ANOVA results follow:

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## HighestEducation    1    223  223.45   59.61 2.41e-14 ***
## Residuals        1215   4555    3.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the F statistic for `HighestEducation`, 59.61, and the p-value $\approx 0 < 0.05$, we reject the null hypothesis and conclude the response means are not the same across education levels for 2006.

Next, the 2011 data ANOVA results:

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## HighestEducation    1    411   411.3   95.84 <2e-16 ***
## Residuals        2185   9377     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the F statistic for `HighestEducation`, 95.84, and the p-value $\approx 0 < 0.05$, we reject the null hypothesis and conclude the response means are not the same across education levels for 2011 either.

An extension of these results would be to perform a multiple comparisions test on each of the education level response means. For the time being, we will defer this, and instead look at differences at each education level across years.

The 2006 education levels range from 2 - 8, *Completed (compulsory) elementary education* through *University with degree/Higher education - upper-level tertiary certificate*, while the 2011 education levels start at 1, but

skip 3 and 4. For the levels in common between the years, 2 & 5-8, we will analyze the difference in means to uncover true differences or whether the differences might occur as a result of chance.

The common education level response means are shown in the data below.

	HighestEducation	Year	meanResponse	sdResponse	nResponse
3	2	2006	6.340909	2.514545	44
4	2	2011	7.444444	2.577019	27
7	5	2006	7.233051	2.044382	236
8	5	2011	6.825641	2.322094	195
9	6	2006	7.631579	1.650494	266
10	6	2011	7.103030	2.160128	660
11	7	2006	7.750000	1.604130	100
12	7	2011	7.496703	2.108168	455
13	8	2006	8.120000	1.715615	25
14	8	2011	7.996377	1.837731	828

Our generalized hypothesis test is set up as follows, where i is the educational level:

$H_0 : \mu_{i,2006} - \mu_{i,2011} = 0$

$H_a : \mu_{i,2006} - \mu_{i,2011} \neq 0$

$\alpha = 0.05$

We will do this hypothesis test following the same process used for the mean science response test above, but applied enmasse to the education levels using our statistical software. The results are shown below.

HighestEducation	DiffMean	StdErr	Tscore	DegFdm	Pval
2	-1.1035354	0.6242332	-1.767826	26	0.0888254
5	0.4074098	0.2129827	1.912877	194	0.0572365
6	0.5285486	0.1315714	4.017201	265	0.0000768
7	0.2532967	0.1884149	1.344356	99	0.1819057
8	0.1236232	0.3490160	0.354205	24	0.7262784

We see some interesting results summarized in the following table.

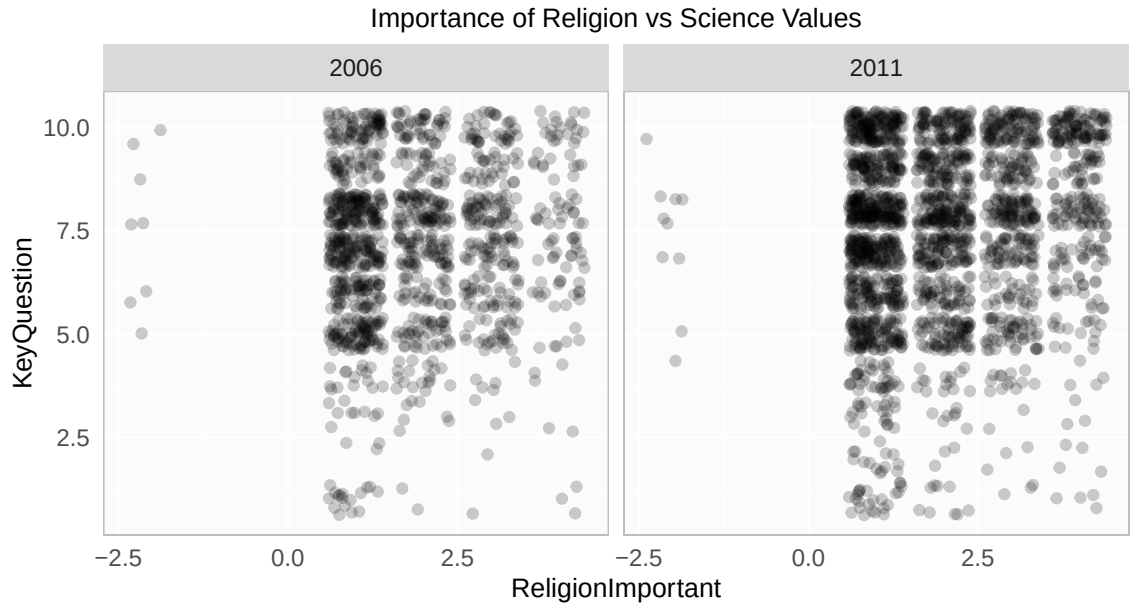
HighestEducation	DiffMean	Pvalue	Conclusion
2	-1.1035354	0.0888254	Failed to Reject Null Hypothesis
5	0.4074098	0.0572365	Failed to Reject Null Hypothesis
6	0.5285486	0.0000768	Reject Null Hypothesis
7	0.2532967	0.1819057	Failed to Reject Null Hypothesis
8	0.1236232	0.7262784	Failed to Reject Null Hypothesis

Basically, the mean responses to the science question for education level 6 changed significantly from 2006 to 2011, while the mean responses for the other education levels did not. Based on the sign of the Difference of Means (DiffMean), those responses in education level 6 reduced their view of science and the world by some degree.

Finally, if we compute the correlation coefficient for education level and science values, we find $R = 0.2113878$, not a strong correlation.

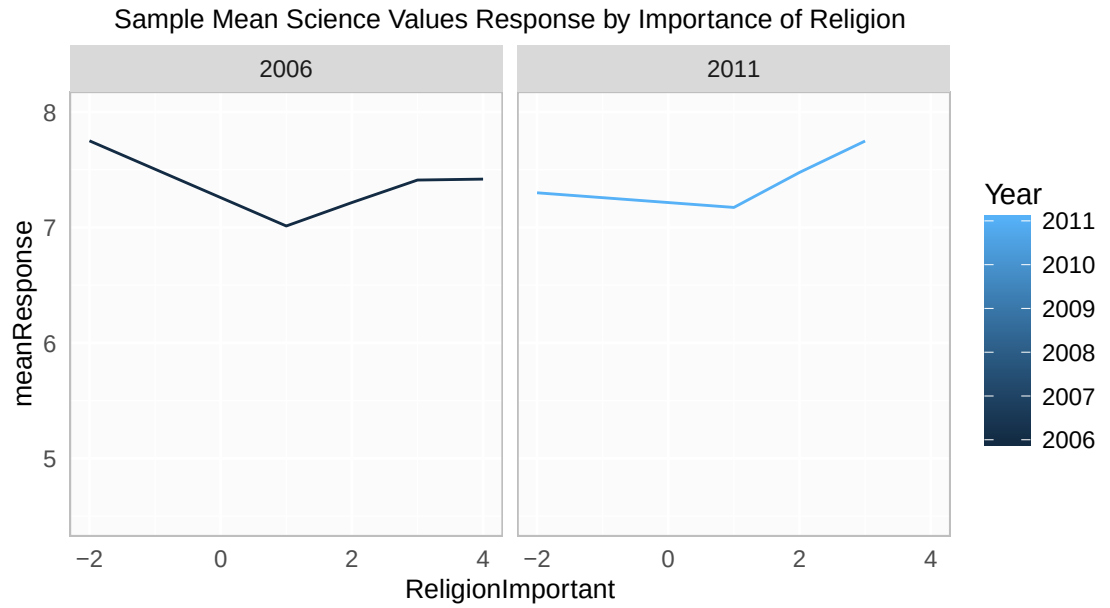
Importance of Religion Correlation

Moving on to the importance of religion and its relationship to the science values responses. As can be seen in the plots, there doesn't appear to be any dominate relationship though we can see evidence of the data distribution we observed during the exploratory phase.



Lets summarize the response variable by the importance of religion, as we did with the education level. When visualizing these means, there isn't an obvious relationship, but we should perform an ANOVA test to confirm.

ReligionImportant	Year	meanResponse	sdResponse	nResponse
-2	2006	7.750000	1.908627	8
1	2006	7.012367	2.058831	566
2	2006	7.215434	1.904634	311
3	2006	7.410256	1.806777	234
4	2006	7.418367	2.119733	98
-2	2011	7.300000	1.702939	10
1	2011	7.173141	2.221859	901
2	2011	7.476191	1.989181	588
3	2011	7.748184	1.973004	413
4	2011	8.145454	2.054177	275



Our hypotheses for an analysis of variance (ANOVA) test as to whether the means for each level of importance of religion are the same or different for a given year are as follows:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (The mean science values response is the same across all levels of importance of religion.)

H_a : At least one mean is different.

$\alpha = 0.05$

First we will start with the 2006 data. The ANOVA results follow:

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## ReligionImportant    1     24  24.146   6.171 0.0131 *
## Residuals         1215   4754   3.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the F statistic for `ReligionImportant`, 6.17, and the p-value $\approx 0.0131 < 0.05$, we reject the null hypothesis and conclude the response means are not the same across religion importance levels for 2006.

Next, the 2011 data ANOVA results:

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## ReligionImportant    1    225  224.60   51.31 1.07e-12 ***
## Residuals         2185   9564    4.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the F statistic for `ReligionImportant`, 51.31, and the p-value $\approx 0 < 0.05$, we reject the null hypothesis and conclude the response means are not the same across religion importance levels for 2011 either.

As with the education level, next we will look at the importance of religion and the science values question across years. Our generalized hypothesis test for the difference of means is set up as follows, where j is the level of importance of religion:

$H_0 : \mu_{j,2006} - \mu_{j,2011} = 0$

$H_a : \mu_{j,2006} - \mu_{j,2011} \neq 0$

$\alpha = 0.05$

We will do this hypothesis test following the same process used for the mean science response test above, but again applied enmasse to the importance of religion using our statistical software. The results are shown below.

ReligionImportant	DiffMean	StdErr	Tscore	DegFdm	Pval
-2	0.4500000	0.8633407	0.5212311	7	0.6182844
1	-0.1607735	0.1138776	-1.4118096	565	0.1585565
2	-0.2607564	0.1356235	-1.9226489	310	0.0554402
3	-0.3379276	0.1528926	-2.2102286	233	0.0280619
4	-0.7270872	0.2473738	-2.9392242	97	0.0041124

Our conclusions are summarized in the following table, but the interpretation is interesting. The population mean associated with those who rated the importance of religion at 3 or higher (3="Not very important" down to 4="Not at all important"), the feeling toward science and the world *increased* from 2006 to 2011 (the world is better off). For the population means associated with -2, 1 and 2 (-2="No answer", 1="Very important" and 2="Rather important"), there was no conclusive change from 2006 to 2011.

ReligionImportant	DiffMean	Pvalue	Conclusion
-2	0.4500000	0.6182844	Failed to Reject Null Hypothesis
1	-0.1607735	0.1585565	Failed to Reject Null Hypothesis
2	-0.2607564	0.0554402	Failed to Reject Null Hypothesis
3	-0.3379276	0.0280619	Reject Null Hypothesis
4	-0.7270872	0.0041124	Reject Null Hypothesis

Finally, if we compute the correlation coefficient for importance of religion and science values, we find $R = 0.1289553$, again not a strong correlation.

Part 5 - Conclusion:

Throughout this data project, we have explored and statistically analyzed World Values Survey data related to responses from the United State regarding feelings on whether the world is better (or worse) off because of science. Statistically significant evidence showed, with a 95% confidence level, that the population mean level related to the world being better off due to science and technology has increased by 0.107529 to 0.507964 from 2006 to 2011.

We have looked at how education level is related to the science/world responses, as well as differences across different education levels. We tested and discovered that in both 2006 and 2011, the science/workd response is different among at least one of the education levels. We also tested and concluded that only one education level, 6 - *Complete secondary: university-preparatory type/Full secondary, maturity level certificate*, had statistically significant changes in the aggregate science and the world responses, while others did not.

We also looked at how the importance of religion is related to the science/world responses and found that the mean science/world response increased for thoses that reported religion as "Not at all Important" and "Not very Important".

Finally, we found the correlation between the science/world response and education level to be stronger than with the importance of religion though both were fairly weak correlations.

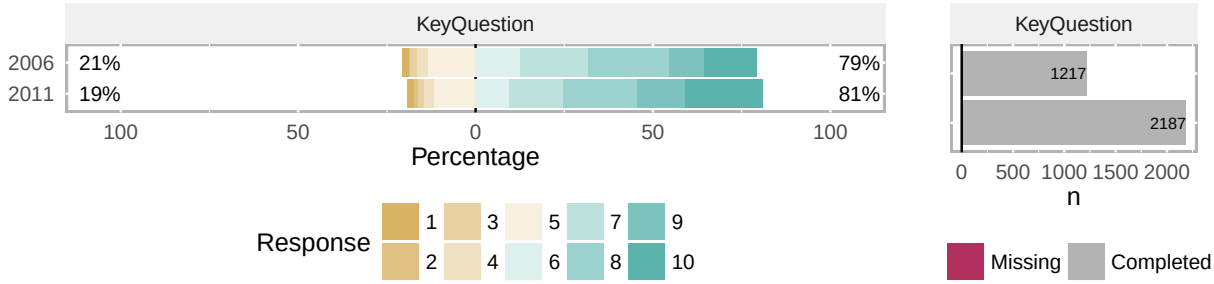
References:

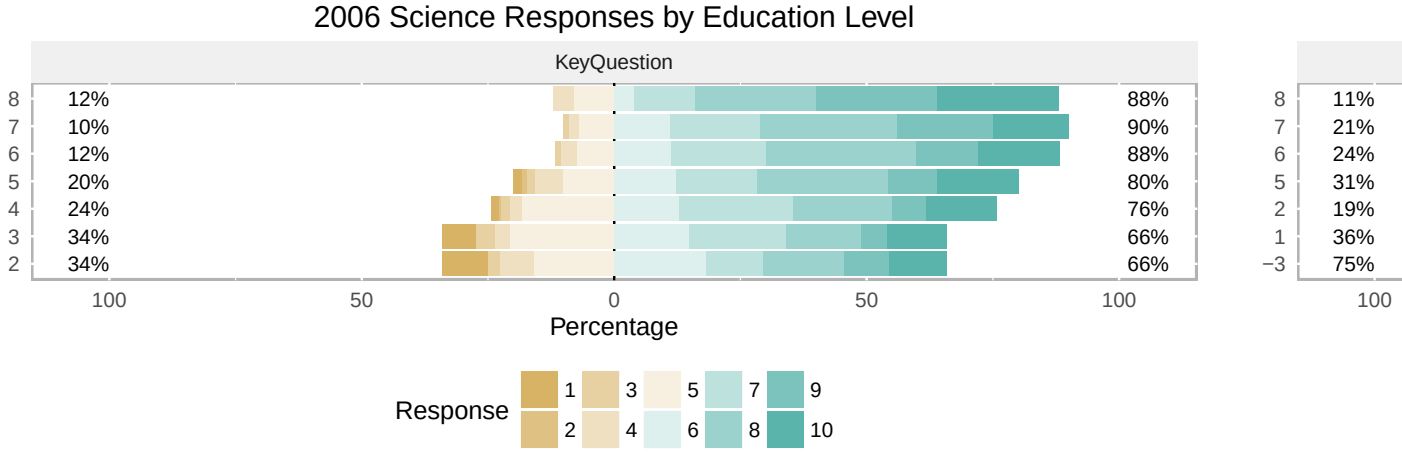
United States Census Bureau. Annual Population Estimates. 2011. URL: http://www.census.gov/popest/data/state/totals/201 EST2011-01.csv.

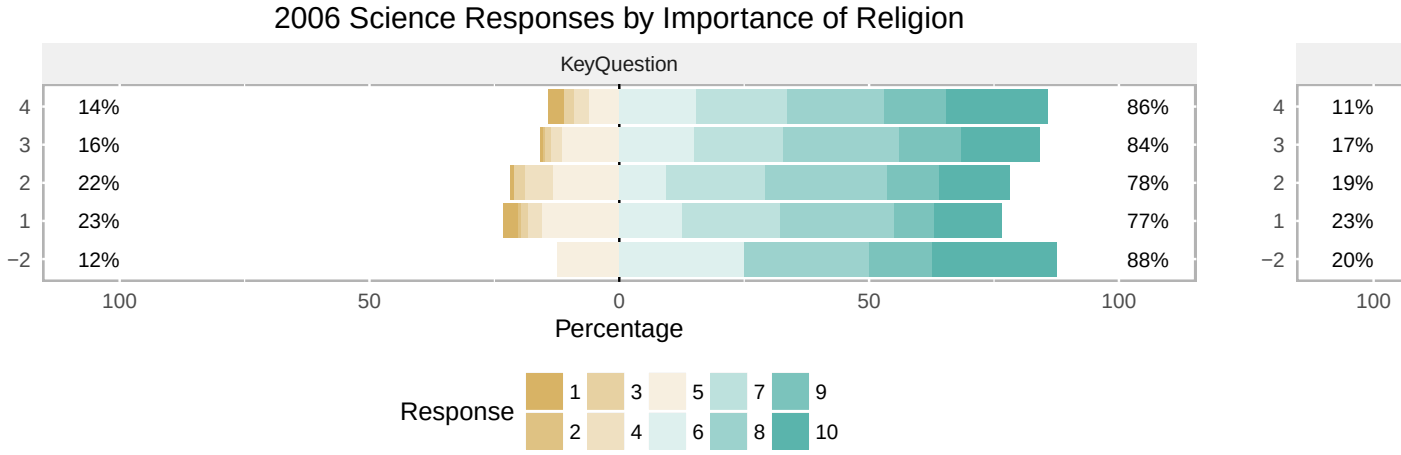
— Annual Population Estimates 2000 - 2006. 2006. URL: http://www.census.gov/popest/data/state/totals/2006/tables/NST-EST2006-01.csv.

World Values Survey Association. WORLD VALUES SURVEY 1981-2014 LONGITUDINAL AGGREGATE v.20150418. Aggregate File Producer: JDSystems. Madrid SPAIN, 2014. URL: http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp.

Appendix - Likert Visualizations:






2006 Science Responses by Importance of Religion
KeyQuestion
4
3
2
1
−2
14%
16%
22%
23%
12%
86%
84%
78%
77%
88%
100
50
0
50
100
Percentage
Response
1
2
3
4
5
6
7
8
9
10
4
3
2
1
−2
11%
17%
19%
23%
20%
100