Project notes

Basic info

Research question:

What are the factors shaping the negative attitude towards homsexuals in France?

Hypothesis:

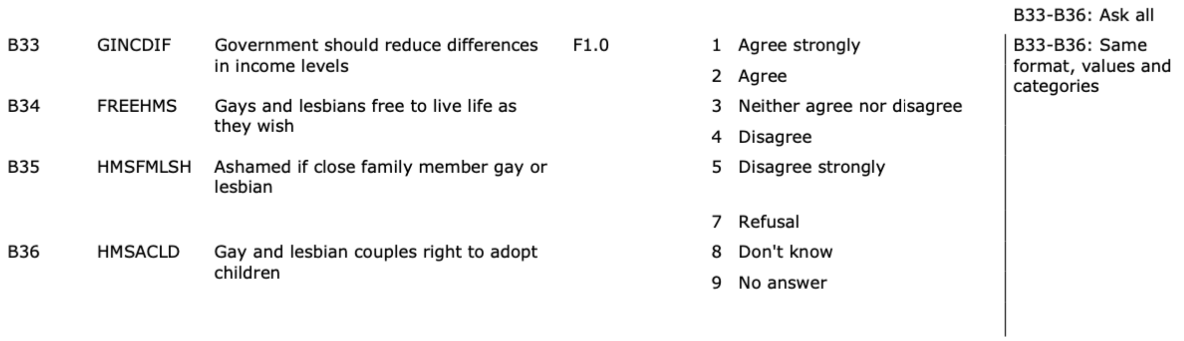
The religiosity of a person stimulates his/her negative attitude towards homosexuals.

Main variables

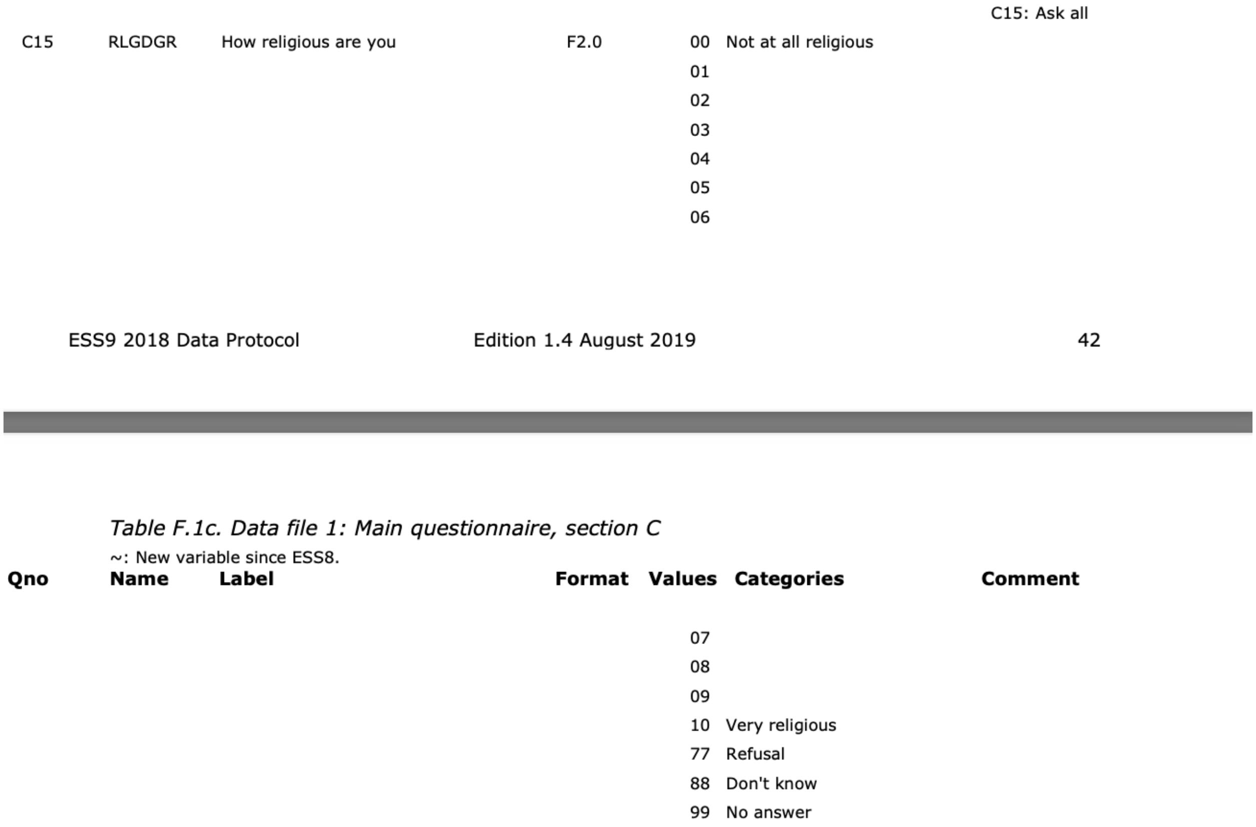
Target - FREEHMS (to answer why we chose this one among other variables related to homosexuality. In my opinion we can say that's it's the most general one, bc we can imagine a situation that a person in general can accept homosexuals, but has something against them in his/her family or has something against them having children.

Main explanatory variable - RLGDGR (scale 1-10, 1 - not at all religious, 10 - very religious)

Scales of main variables, and meaning of responses:



						B33-B36: Ask all
B33	GINCDIF	Government should reduce differences in income levels	F1.0	1	Agree strongly	B33-B36: Same format, values and categories
				2	Agree	
B34	FREEHMS	Gays and lesbians free to live life as they wish		3	Neither agree nor disagree	
				4	Disagree	
B35	HMSFMLSH	Ashamed if close family member gay or lesbian		5	Disagree strongly	
				7	Refusal	
B36	HMSACLD	Gay and lesbian couples right to adopt children		8	Don't know	
				9	No answer	



						C15: Ask all
C15	RLGDGR	How religious are you	F2.0	00	Not at all religious	
				01		
				02		
				03		
				04		
				05		
				06		

ESS9 2018 Data Protocol Edition 1.4 August 2019 42

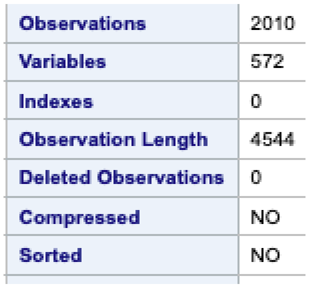
Table F.1c. Data file 1: Main questionnaire, section C

~: New variable since ESS8.

Qno	Name	Label	Format	Values	Categories	Comment
				07		
				08		
				09		
				10	Very religious	
				77	Refusal	
				88	Don't know	
				99	No answer	

In introduction, add an info that our alpha in this analysis is 0.05 (threshold of statistical significancy)

Analysis:



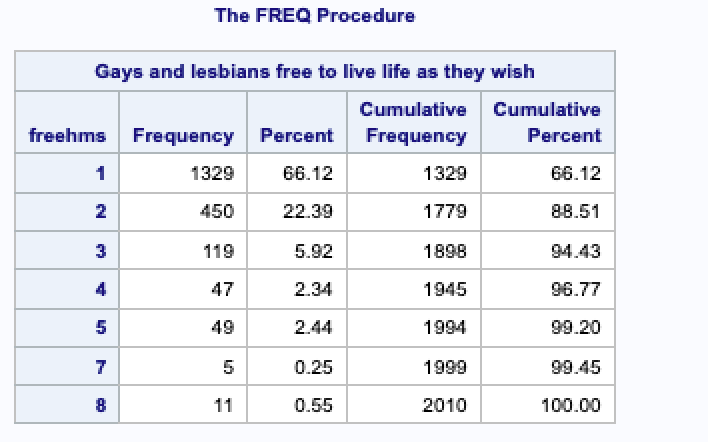
Observations	2010
Variables	572
Indexes	0
Observation Length	4544
Deleted Observations	0
Compressed	NO
Sorted	NO

At the begging of our analysis, we create a dataset containing only information about France, as we are interested only in this one country. Resulting dataset consists of 2010 rows, and 572 columns.

1. Analysis of main variables (descriptive and discriminatory performance)

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/3f98a866-dc06-4c72-b73e-173b589a3ae7/SAS_project_ordinal_mw117894_117408-results(1).html

1. freehms



The FREQ Procedure

Gays and lesbians free to live life as they wish				
freehms	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1329	66.12	1329	66.12
2	450	22.39	1779	88.51
3	119	5.92	1898	94.43
4	47	2.34	1945	96.77
5	49	2.44	1994	99.20
7	5	0.25	1999	99.45
8	11	0.55	2010	100.00

As wee can see more than 88% of observations contains a value 1 or 2 of the target variable. We can also see a gradual decrease of number of observations among further categories. Also, there is insufficient number of observations among categories 4 and 5. Therefore, we decided to reduce the number of categories to 3 of them:

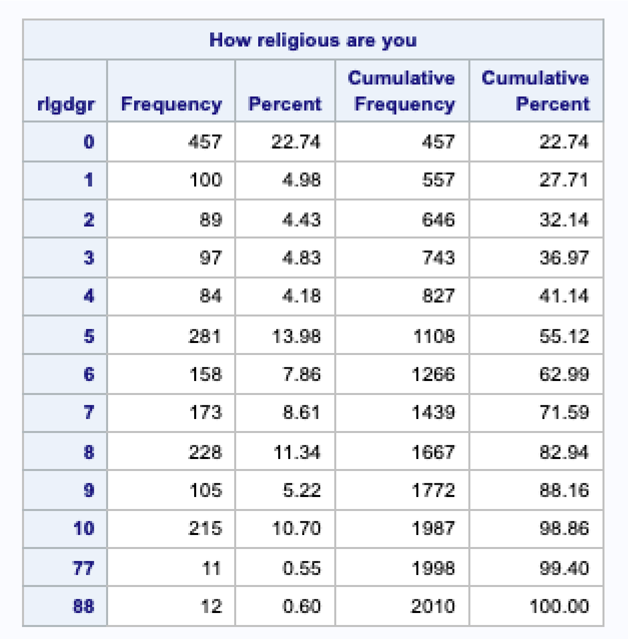
a. 1. Strongly agree (group 1)

b. 2. Agree (group 2)

c. 3. Neutral or disagree (groups 3, 4 and 5)

We know, that there are different combinations possible. For example, We could leave groups 1,2 and 3 and combine only 4 and 5. It would be probably more informative division, however combining only groups 4 and 5 wouldn't result in sufficient number of observations in this category (we assume that lowest number that a category should contain is 100). Oversampling might help here, however it would create a question how much the data we work with after oversampling is similar do data without oversampling, which would make our verification of the hypothesis doubtful.

2. RLGDGR

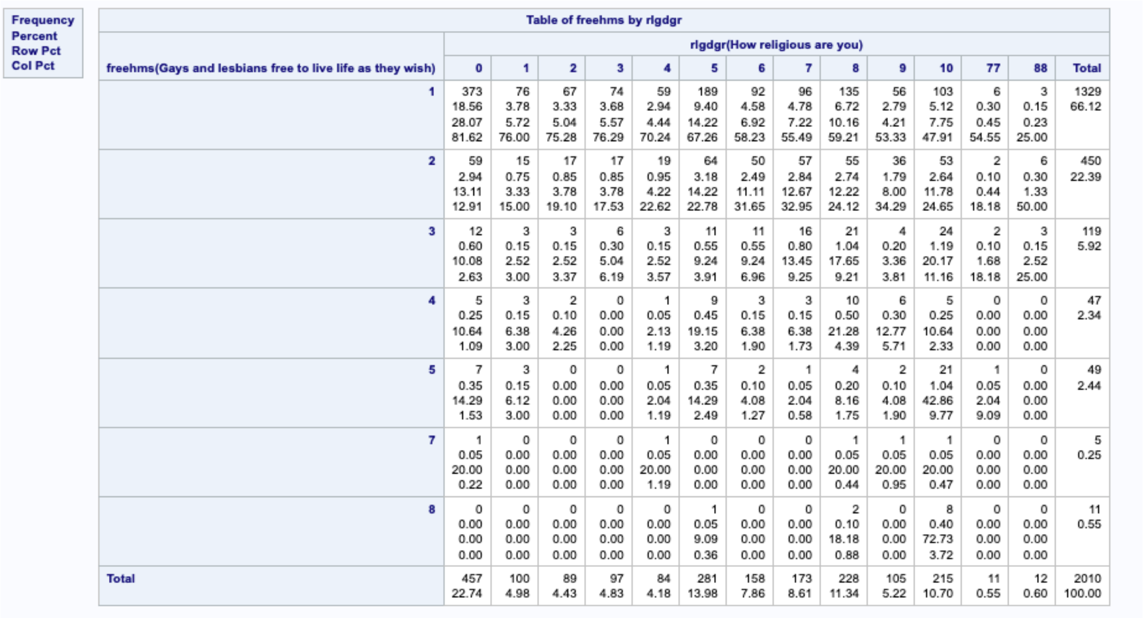


How religious are you

rlgdgr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	457	22.74	457	22.74
1	100	4.98	557	27.71
2	89	4.43	646	32.14
3	97	4.83	743	36.97
4	84	4.18	827	41.14
5	281	13.98	1108	55.12
6	158	7.86	1266	62.99
7	173	8.61	1439	71.59
8	228	11.34	1667	82.94
9	105	5.22	1772	88.16
10	215	10.70	1987	98.86
77	11	0.55	1998	99.40
88	12	0.60	2010	100.00

As we can see the biggest fraction, over 22% declared themselves as not religious at all, however the disproportion is not that big as in case of the target variable. Here, we find it reasonable to reduce the dimensionality, but not to such an extent as in case of the target. We're going to combine categories 2,3 and 4 into category 1. In result we will receive an ordinal variable with 9 degrees. (0 - not religious at all, 8 - very religious).

3. FREEHMS * RLGDGR



Frequency
Percent
Row Pct
Col Pct

Table of freehms by rlgdgr

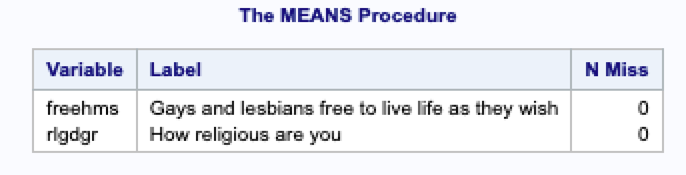
	rlgdgr(How religious are you)													
freehms(Gays and lesbians free to live life as they wish)	0	1	2	3	4	5	6	7	8	9	10	77	88	Total
1	373 18.56 28.07 81.62	76 3.78 5.72 76.00	67 3.33 5.04 75.28	74 3.68 5.57 76.29	59 2.94 4.44 70.24	189 9.40 14.22 67.26	92 4.58 6.92 58.23	96 4.78 7.22 55.49	135 6.72 10.16 59.21	56 2.79 4.21 53.33	103 5.12 7.75 47.91	6 0.30 0.45 54.55	3 0.15 0.23 25.00	1329 66.12
2	59 2.94 13.11 12.91	15 0.75 3.33 15.00	17 0.85 3.78 19.10	17 0.85 3.78 17.53	19 0.95 4.22 22.62	64 3.18 14.22 22.78	50 2.49 11.11 31.65	57 2.84 12.67 32.95	55 2.74 12.22 24.12	36 1.79 8.00 34.29	53 2.64 11.78 24.65	2 0.10 0.44 18.18	6 0.30 1.33 50.00	450 22.39
3	12 0.60 10.08 2.63	3 0.15 2.52 3.00	3 0.15 2.52 3.37	6 0.30 5.04 6.19	3 0.15 2.52 3.57	11 0.55 9.24 3.91	11 0.55 9.24 6.96	16 0.80 13.45 9.25	21 1.04 17.65 9.21	4 0.20 3.36 3.81	24 1.19 20.17 11.16	2 0.10 1.68 18.18	3 0.15 2.52 25.00	119 5.92
4	5 0.25 10.64 1.09	3 0.15 6.38 3.00	2 0.10 4.26 2.25	0 0.00 0.00 0.00	1 0.05 2.13 1.19	9 0.45 19.15 3.20	3 0.15 6.38 1.90	3 0.15 6.38 1.73	10 0.50 21.28 4.39	6 0.30 12.77 5.71	5 0.25 10.64 2.33	0 0.00 0.00 0.00	0 0.00 0.00 0.00	47 2.34
5	7 0.35 14.29 1.53	3 0.15 6.12 3.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.05 2.04 1.19	7 0.35 14.29 2.49	2 0.10 4.08 1.27	1 0.05 2.04 0.58	4 0.20 8.16 1.75	2 0.10 4.08 1.90	21 1.04 42.86 9.77	1 0.05 2.04 9.09	0 0.00 0.00 0.00	49 2.44
7	1 0.05 20.00 0.22	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.05 20.00 1.19	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.05 20.00 0.44	1 0.05 20.00 0.95	1 0.05 20.00 0.47	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 0.25
8	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.05 9.09 0.36	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 0.10 18.18 0.88	0 0.00 0.00 0.00	8 0.40 72.73 3.72	0 0.00 0.00 0.00	0 0.00 0.00 0.00	11 0.55
Total	457 22.74	100 4.98	89 4.43	97 4.83	84 4.18	281 13.98	158 7.86	173 8.61	228 11.34	105 5.22	215 10.70	11 0.55	12 0.60	2010 100.00

As we can see, for all religiosity levels from 0 to 4 90% of observations contain a value 1 or 2 of freehms variable. We can se a decrease in percentage among higher religiosity levels, - the least amount of people responding 1 or 2 is among value 10 of religiosity - 72.56%. The remaining question however is if this difference is statistically significant.

When we look at observations 3,4 and 5 of freehms and their correlation with rlgdgr - it seems as there is an opposite phenomena. Number of people that decided to answer 3, 4 or 5 for freehms looks to be rising with religiosity. However, these differences seem not to be either purely linear, nor significant.

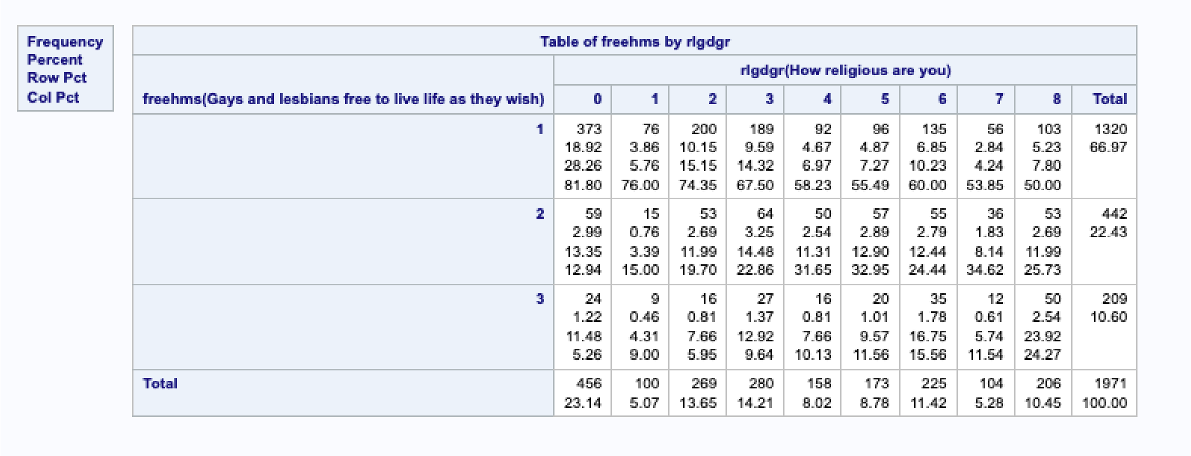
To conclude, after analysis of FREEHMS and RLGDGR we come to a conclusion, that we should reduce the number of degrees in both of them. While doing so, we will also exclude observations, with values 7 and 8 of FREEHMS and 77, 88 of RLGDGR, as these answers indicate that person either didn't know the answer or refused to answer, which virtually means missing information. It is also worth to notice, that neither of these variables contains missing data (2010 observations of both of them).

4. FREEHMS_REDUCED & RLGDGR _REDUCED



The MEANS Procedure

Variable	Label	N Miss
freehms	Gays and lesbians free to live life as they wish	0
rlgdgr	How religious are you	0

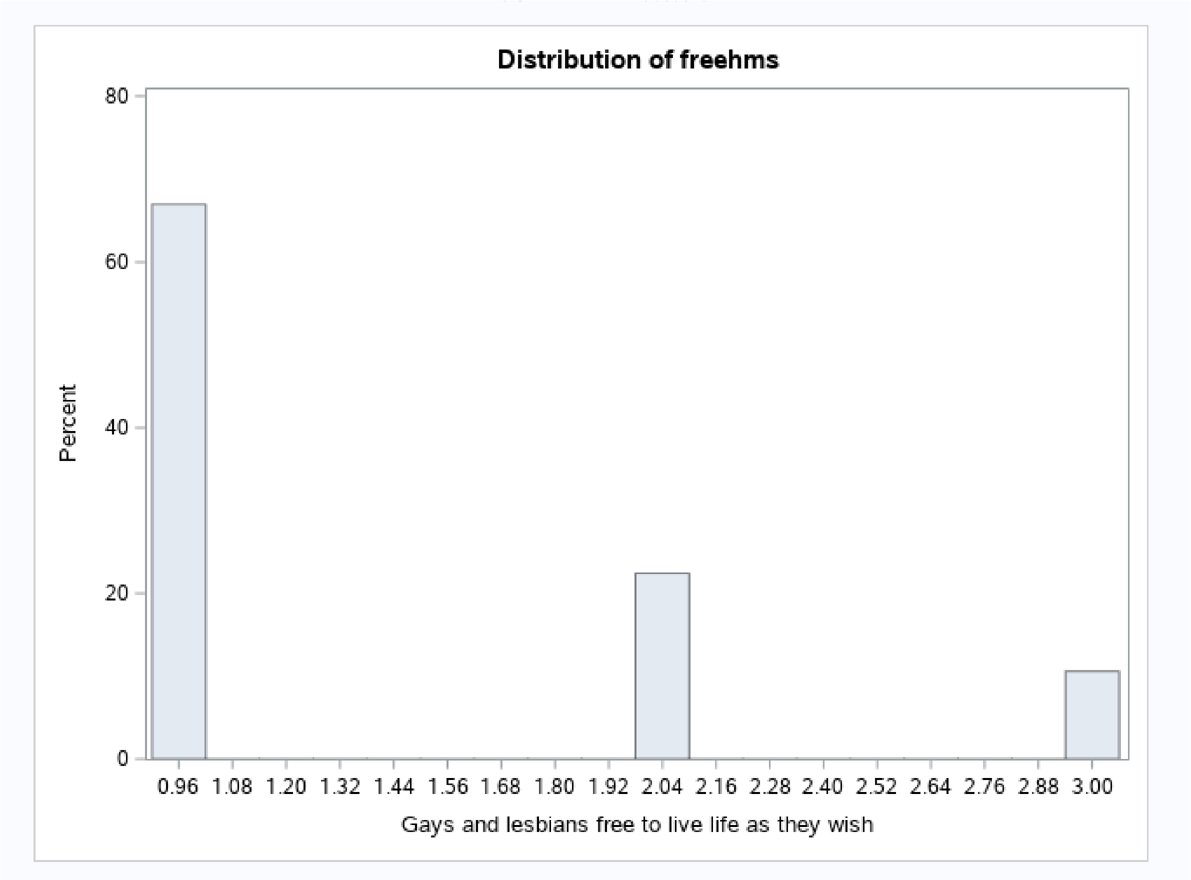


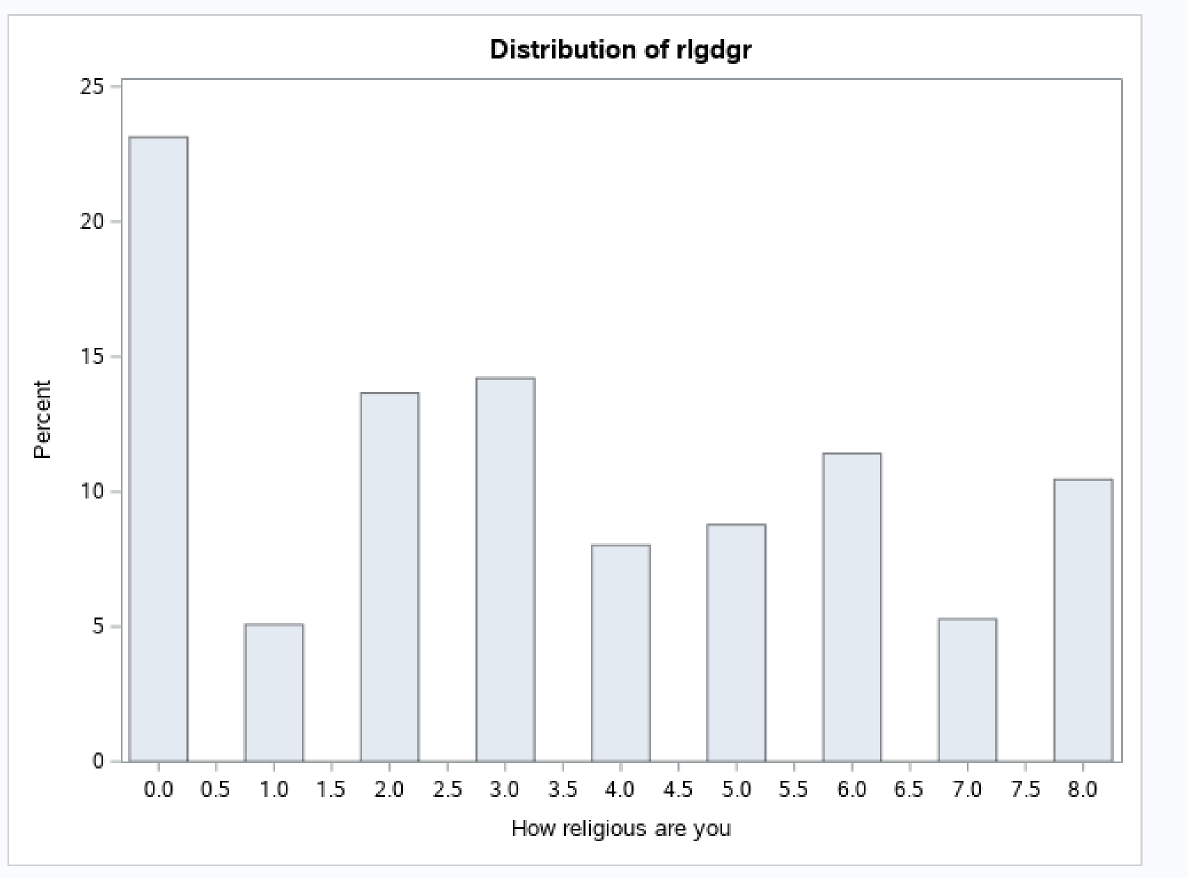
Frequency
Percent
Row Pct
Col Pct

Table of freehms by rlgdgr										
	rlgdgr(How religious are you)									
freehms(Gays and lesbians free to live life as they wish)	0	1	2	3	4	5	6	7	8	Total
1	373 18.92 28.26 81.80	76 3.86 5.76 76.00	200 10.15 15.15 74.35	189 9.59 14.32 67.50	92 4.67 6.97 58.23	96 4.87 7.27 55.49	135 6.85 10.23 60.00	56 2.84 4.24 53.85	103 5.23 7.80 50.00	1320 66.97
2	59 2.99 13.35 12.94	15 0.76 3.39 15.00	53 2.69 11.99 19.70	64 3.25 14.48 22.86	50 2.54 11.31 31.65	57 2.89 12.90 32.95	55 2.79 12.44 24.44	36 1.83 8.14 34.62	53 2.69 11.99 25.73	442 22.43
3	24 1.22 11.48 5.26	9 0.46 4.31 9.00	16 0.81 7.66 5.95	27 1.37 12.92 9.64	16 0.81 7.66 10.13	20 1.01 9.57 11.56	35 1.78 16.75 15.56	12 0.61 5.74 11.54	50 2.54 23.92 24.27	209 10.60
Total	456 23.14	100 5.07	269 13.65	280 14.21	158 8.02	173 8.78	225 11.42	104 5.28	206 10.45	1971 100.00

As we can see number of observations of the 1st category decreased by 9. This is due to the fact, that 9 observations contained a value 77 or 88 in RLGDGR variable. This is also the reason why number observations in category 2 decreased by 8 and number of observations in category 3 (sum of categories 3, 4 and 5) is lower by 6. Besides that, there were also 16 occurrences of categories 7 and 8. When we sum all of these missing values and subtract them from original number of rows (2010) we result with 1971, which is exactly the number of rows in our new dataset. Fortunately, number of missing values is not big enough (and there is also no not labeled missing values) to constitute a significant concern (as well the total number as the number of missing values in each category for both variables).

Also, as we can see now, the number of observations in each of the categories for both variables after reduction is now sufficient for building more reliable logistic regression model.

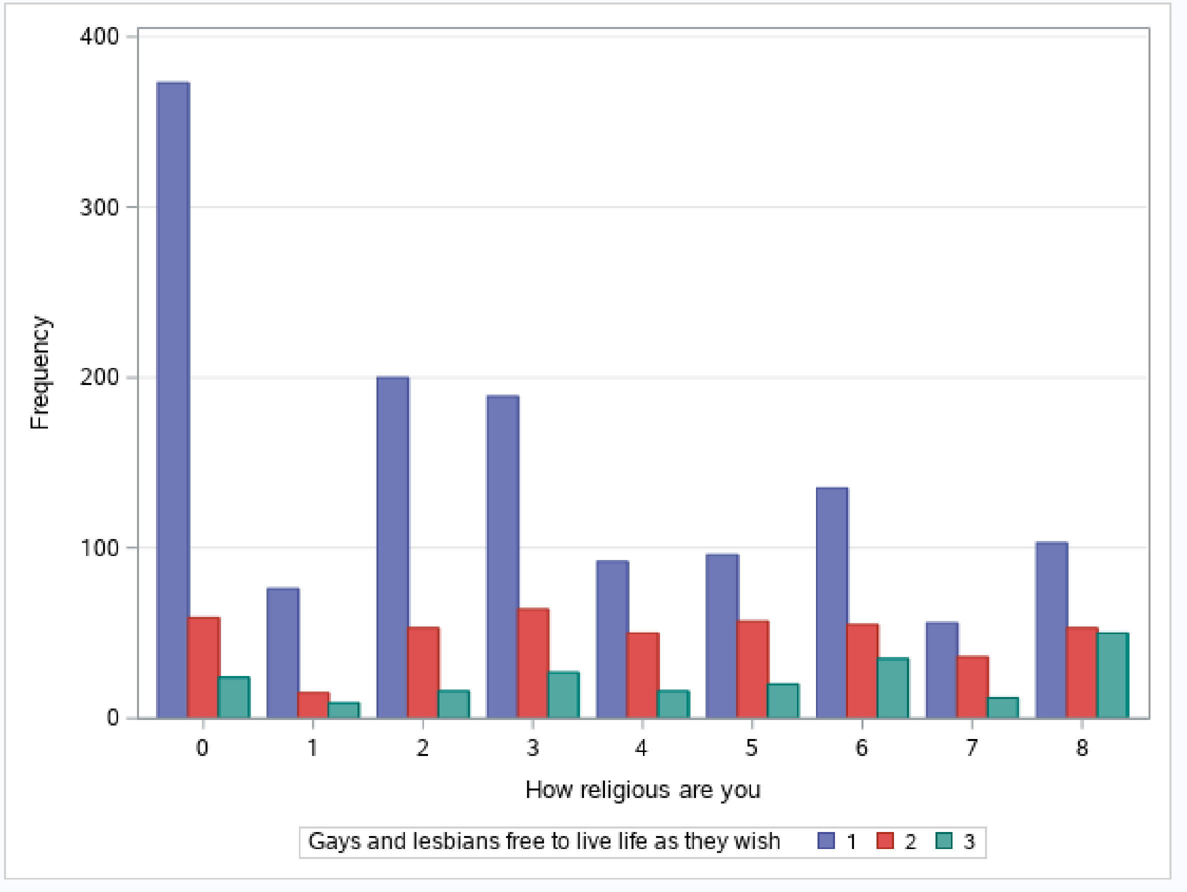

Distribution of freehms
Percent
80
60
40
20
0
0.96
1.08
1.20
1.32
1.44
1.56
1.68
1.80
1.92
2.04
2.16
2.28
2.40
2.52
2.64
2.76
2.88
3.00
Gays and lesbians free to live life as they wish

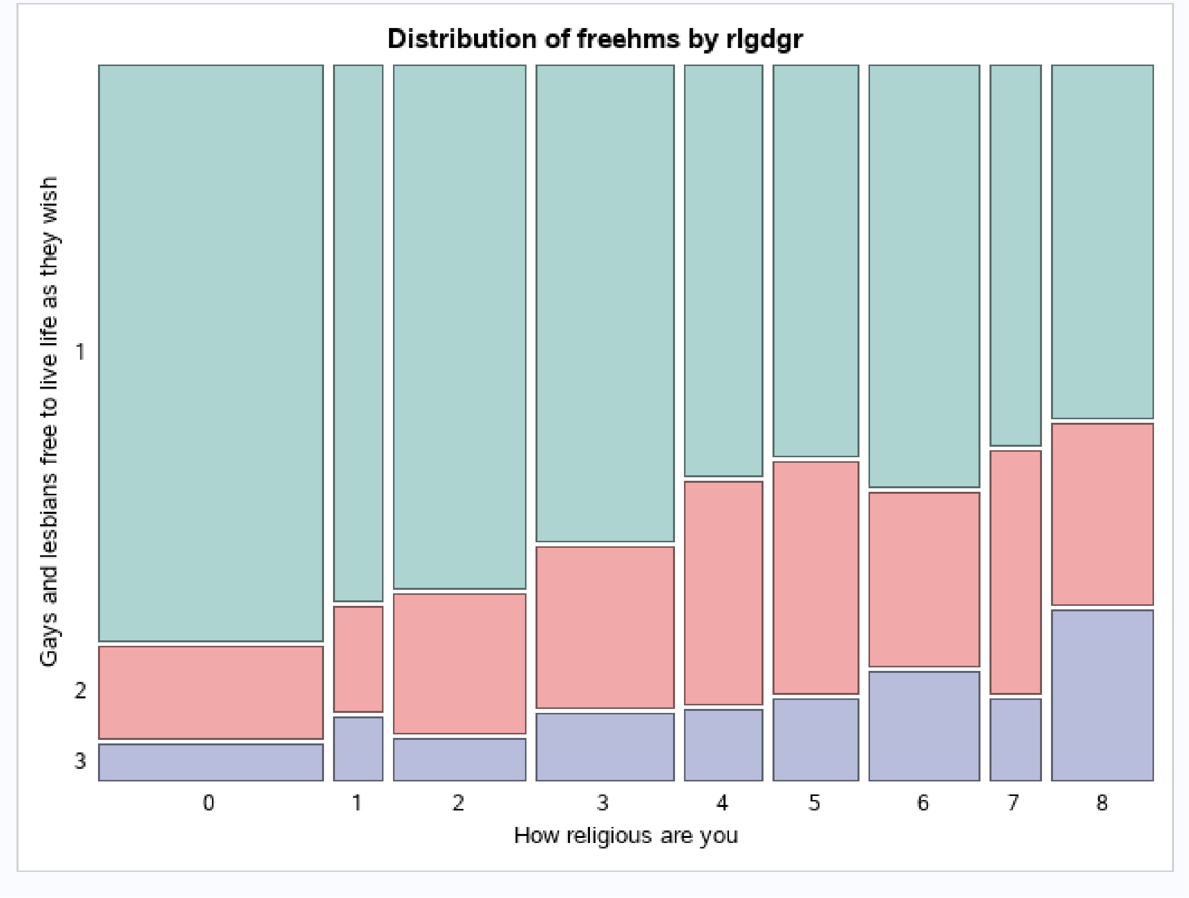


As we can see, after reducing the number of dimensions, FREEHMS variable preserved its right-skewed distribution.

RLGFGR still has its dominant in value 0 and distribution of the rest of them is nor purely skewed. It is worth to noice, that combining categories 2,3 and 4 resulted in creating the 3rd most populated category. The rest of the categories however, is not underrepresented.

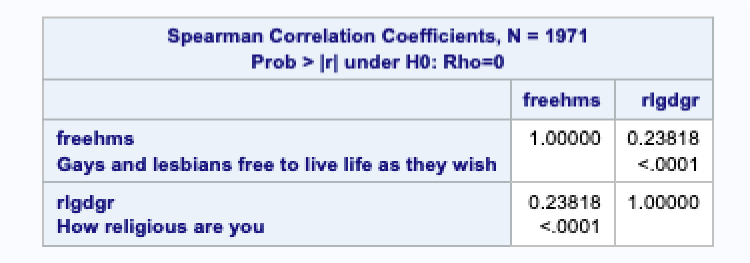
5. Relationship between FREEHMS and RLGDGR


400
300
200
100
0
Frequency
0
1
2
3
4
5
6
7
8
How religious are you
Gays and lesbians free to live life as they wish
1
2
3



When we look at the barchart with absolute frequencies, we can notice a quite high values of each of 3 categories of the target variable for rlgdgr 0 category, and a gradual decrease of strongly supportive answers along the way. Category 2 of the target seems to be stable across religiosity, and category 3 is non linearily distributed across religiosity. (there is a slight peak in rlgdgr 3rd category). However, we need to remember that most of the observations of these variables are the 1st category of freehms and 0 rlgdgr. Therefore, it may be beneficial to have a look at them in a relative way. Therefore, let's analyse the mosaic plot.

We can see a clear decrease in number of observations of category 1 of freehms variable along higher religiosity levels, along with gradual increase of category 2. Category 3 seems to be rising from religiosity 3 to 8 with exception of religiosity 7. We can notice certain interesting patterns here, however we still need to test their significance and dependency on other variables.



Spearman Correlation Coefficients, N = 1971
Prob > |r| under H0: Rho=0

	freehms	rlgdgr
freehms Gays and lesbians free to live life as they wish	1.00000	0.23818 <.0001
rlgdgr How religious are you	0.23818 <.0001	1.00000

As we can see there is a statistically significant, however minor positive correlation between our variables, which confirms our previous observation - higher values of FREEHMS tend to appear along with higher values of religiosity.

However, before build a logistic regression model we would like to select more explanatory variables to widen our understanding of the hypothesis, we're checking.

2. Selecting more explanatory variables

Intro - here we should say smth that we want to explore this problem deeper, so we chose more variables to see the effect in the broader perspective. And we can also add why we chose these variables

We decided to add following variables to our model:

a. GNDR - Gender

b. AGEA - Age of respondent, calculated

c. NETUSOFT - Internet use, how often

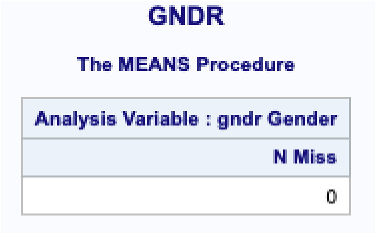
d. PPLFAIR - Most people try to take advantage of you, or try to be fair

e. EDUYRS - Years of full-time education completed

Descriptive and discriminatory performance analysis of explanatory variables

Data distribution and missing values

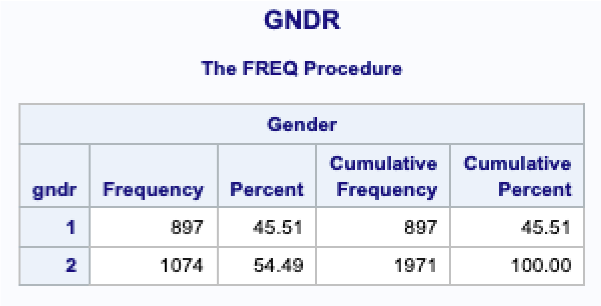
1. Gender



GNDR

The MEANS Procedure

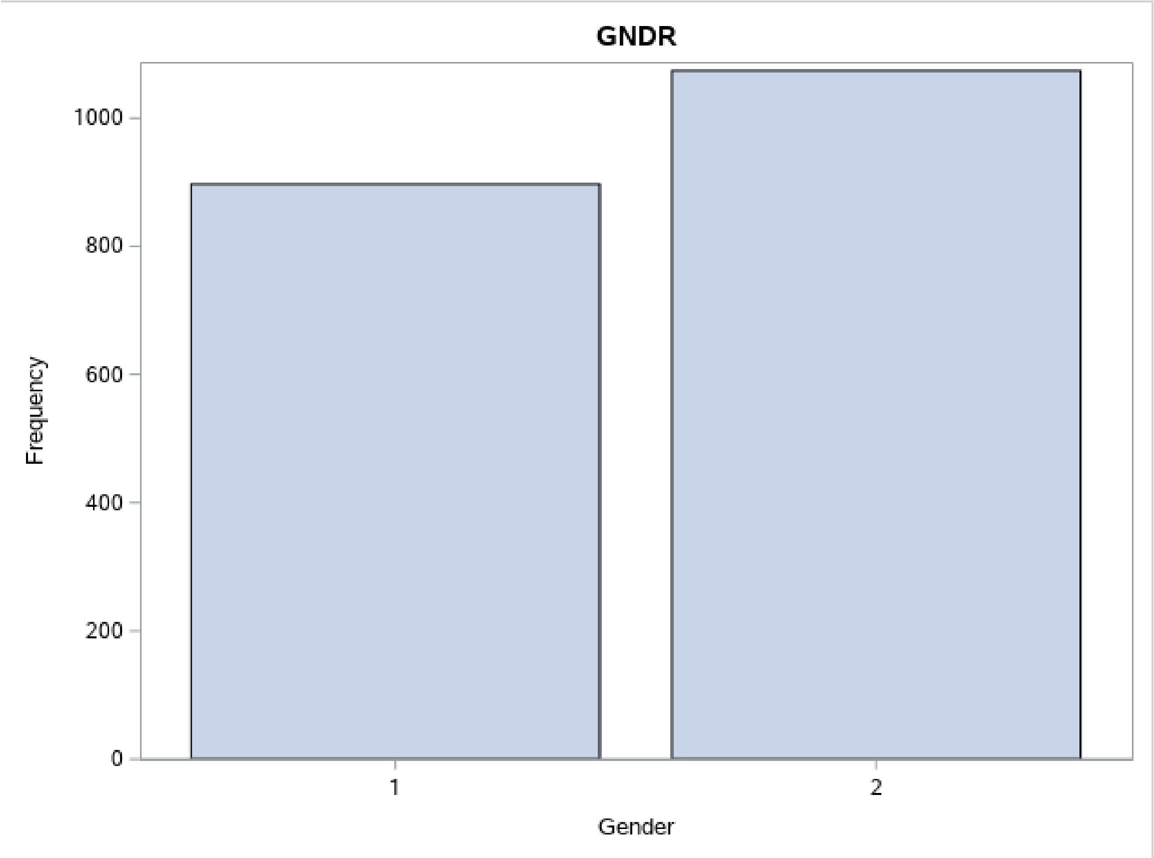
Analysis Variable : gndr Gender
N Miss
0



GNDR

The FREQ Procedure

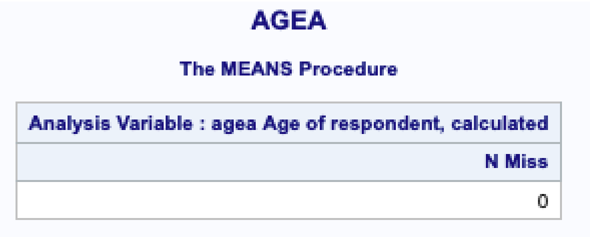
Gender				
gndr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	897	45.51	897	45.51
2	1074	54.49	1971	100.00



The distribution of gender variable is is quite even. There is 177 more women (2) than men. However, this difference is not problematic, since we still have big representation of men population. Fortunately, this variable does not contain any missing data.

b. AGEA

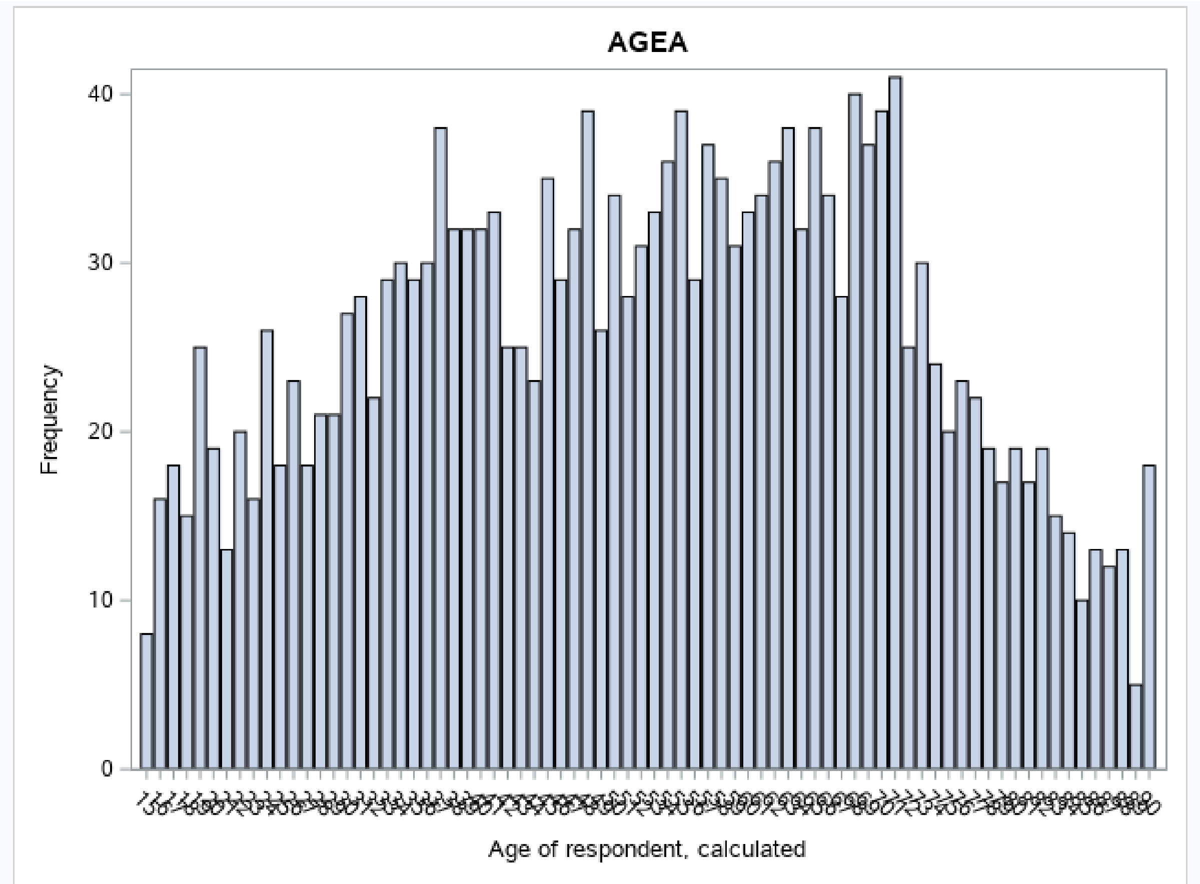
https://s3-us-west-2.amazonaws.com/secure.notion-static.com/b2a97582-680b-486c-a9eb-6e053fb787df/F01.csv

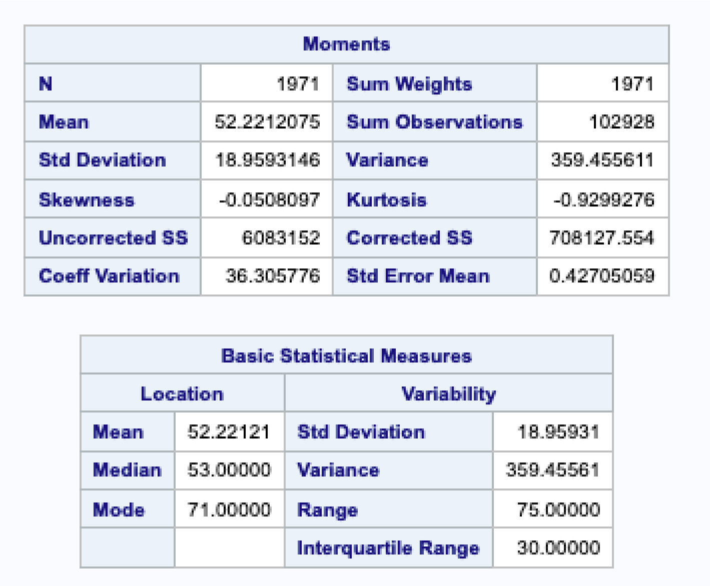


AGEA

The MEANS Procedure

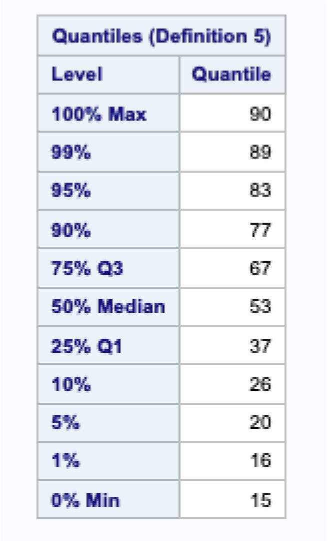
Analysis Variable : agea Age of respondent, calculated
N Miss
0



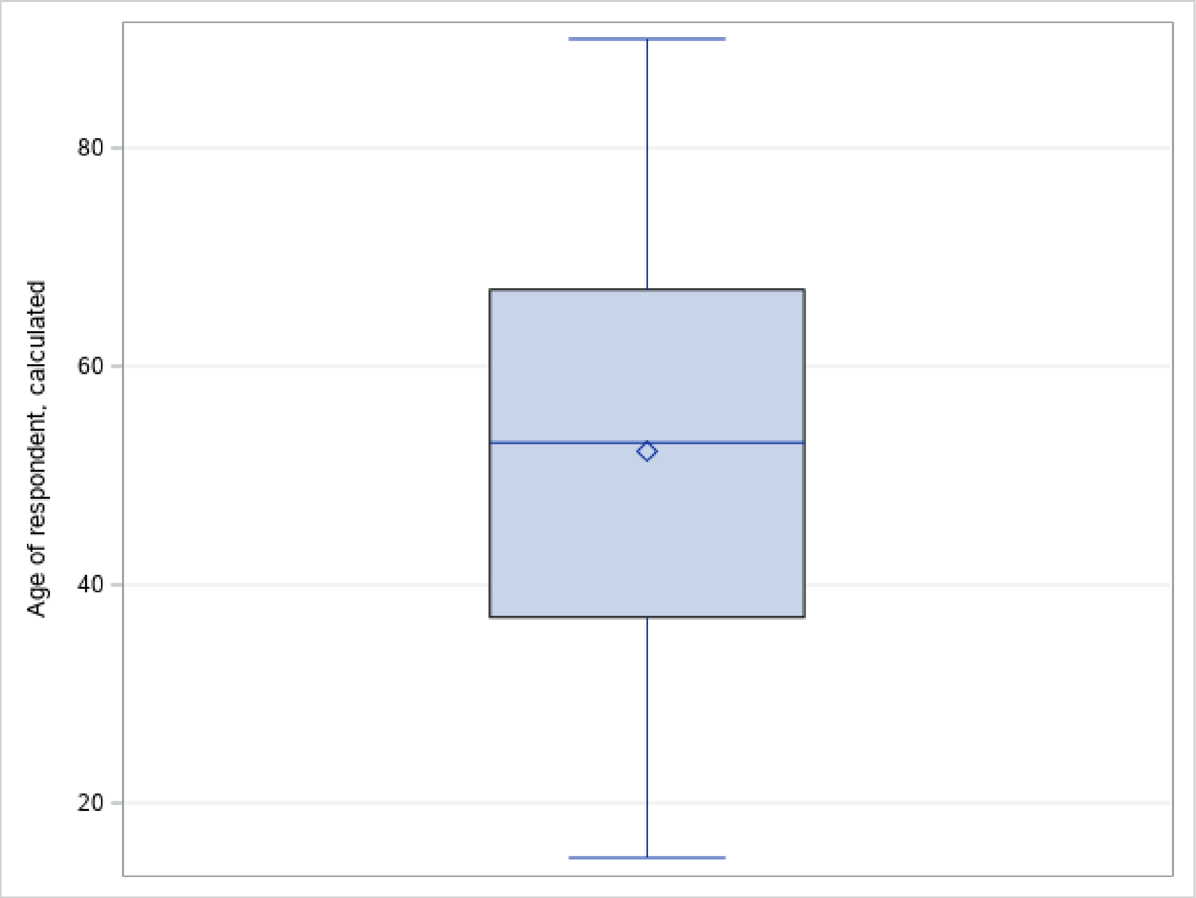


Moments			
N	1971	Sum Weights	1971
Mean	52.2212075	Sum Observations	102928
Std Deviation	18.9593146	Variance	359.455611
Skewness	-0.0508097	Kurtosis	-0.9299276
Uncorrected SS	6083152	Corrected SS	708127.554
Coeff Variation	36.305776	Std Error Mean	0.42705059

Basic Statistical Measures			
Location		Variability	
Mean	52.22121	Std Deviation	18.95931
Median	53.00000	Variance	359.45561
Mode	71.00000	Range	75.00000
		Interquartile Range	30.00000

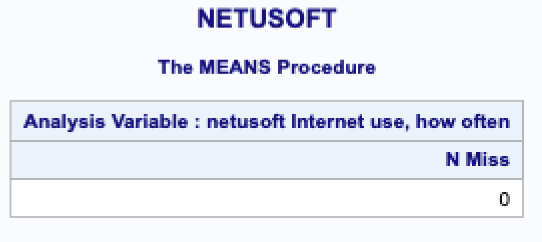


Quantiles (Definition 5)	
Level	Quantile
100% Max	90
99%	89
95%	83
90%	77
75% Q3	67
50% Median	53
25% Q1	37
10%	26
5%	20
1%	16
0% Min	15



AGEA also does not contain any missing values (1971 observations, maximum value is 90, not 999). looking at the difference between mean and median, and at the skewness and kurtosis statistics, we can observe that data distribution is close to normal distribution. Age is distributed evenly, with majority of adult respondents (especially between 50 and 70 years old), however we cannot notice vast majority of any age group. Also, there are no values below lower interquartile range or above the the upper interquartile range. We decided to leave this variable as it is.

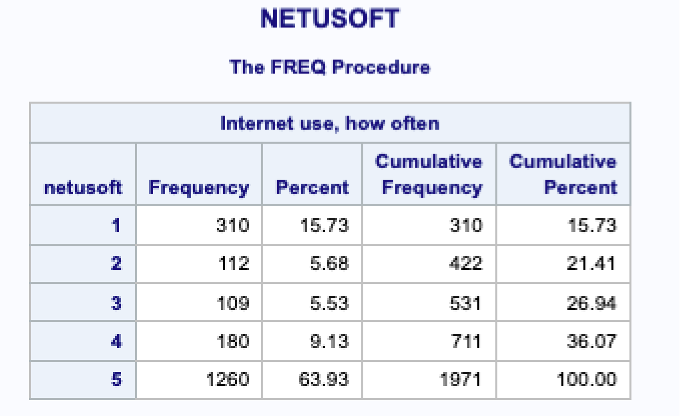
c. NETUSOFT



NETUSOFT

The MEANS Procedure

Analysis Variable : netusoft Internet use, how often
N Miss
0

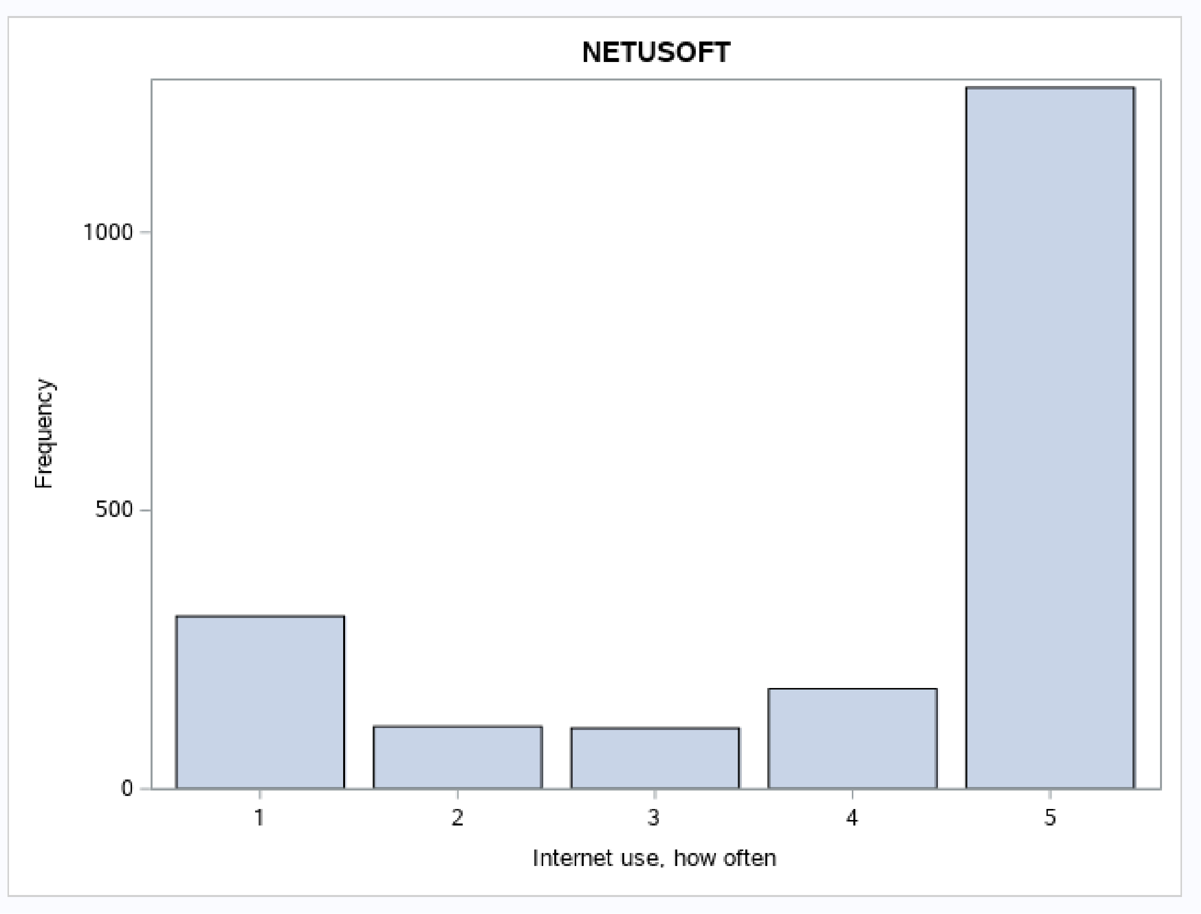


NETUSOFT

The FREQ Procedure

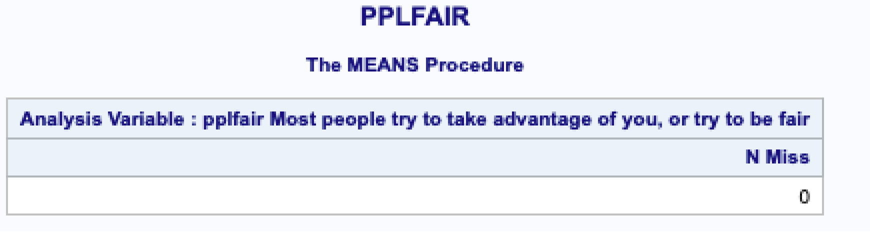
Internet use, how often

netusoft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	310	15.73	310	15.73
2	112	5.68	422	21.41
3	109	5.53	531	26.94
4	180	9.13	711	36.07
5	1260	63.93	1971	100.00



Vast majority of people respondents use internet everyday (5). It is not surprising, it sounds intuitive, that majority of people in France use internet every day.. Rest of the categories, however are frequent enough to leave the initial number of categories. We can also notice no missing values (no actual missing values, no categories 7, 8, 9). Therefore, we do not find a need to modify this variable.

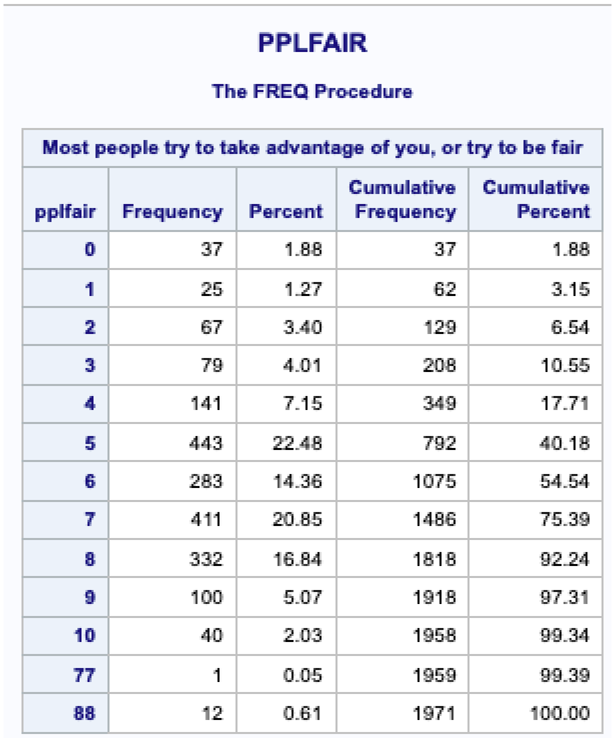
d. PPLFAIR



PPLFAIR

The MEANS Procedure

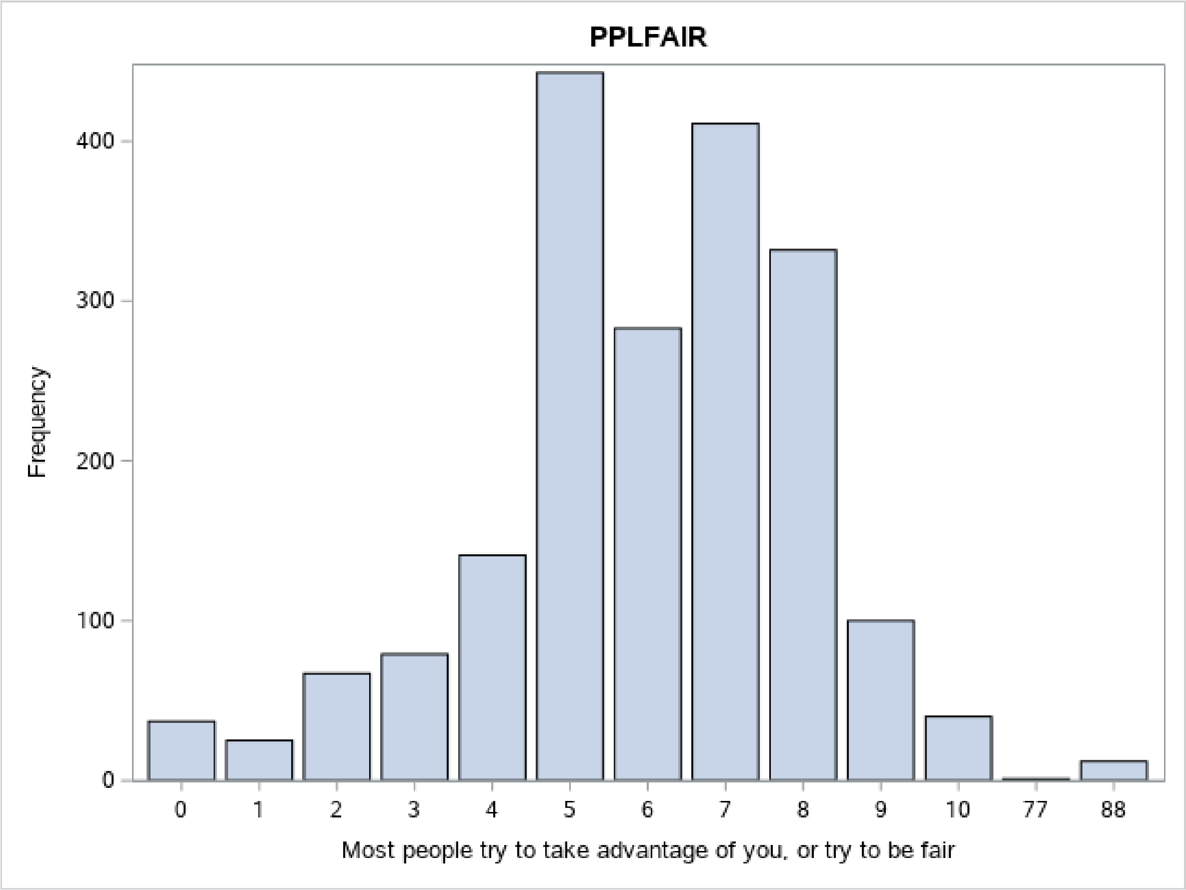
Analysis Variable : pplfair Most people try to take advantage of you, or try to be fair
N Miss
0



PPLFAIR

The FREQ Procedure

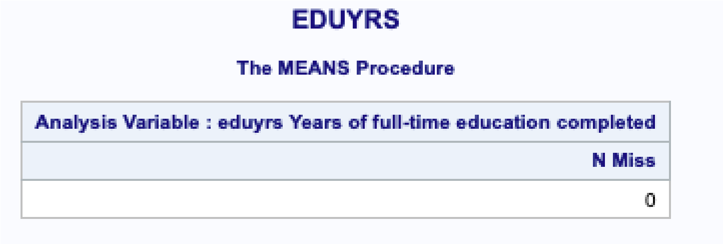
Most people try to take advantage of you, or try to be fair				
pplfair	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	37	1.88	37	1.88
1	25	1.27	62	3.15
2	67	3.40	129	6.54
3	79	4.01	208	10.55
4	141	7.15	349	17.71
5	443	22.48	792	40.18
6	283	14.36	1075	54.54
7	411	20.85	1486	75.39
8	332	16.84	1818	92.24
9	100	5.07	1918	97.31
10	40	2.03	1958	99.34
77	1	0.05	1959	99.39
88	12	0.61	1971	100.00



Most of the answers are focused over values 5 - 8, which is not surprising, as they are not that radical as the other ones. Unfortunately, we can notice 13 observations of values 77 and 88, which is missing information. We should remove them.

Also, observation in categories 0 - 3 and 10 are not sufficient. Therefore, we will reduce number of them, by merging categories 0 -3 and 9-10.

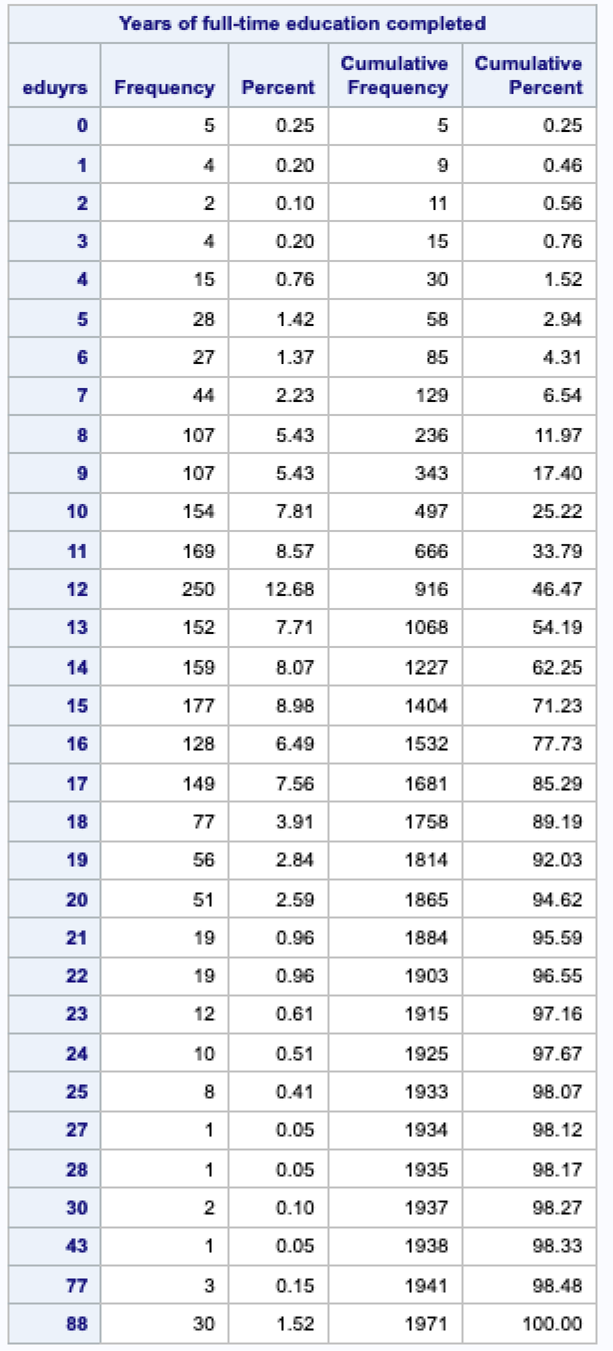
e. EDUYRS



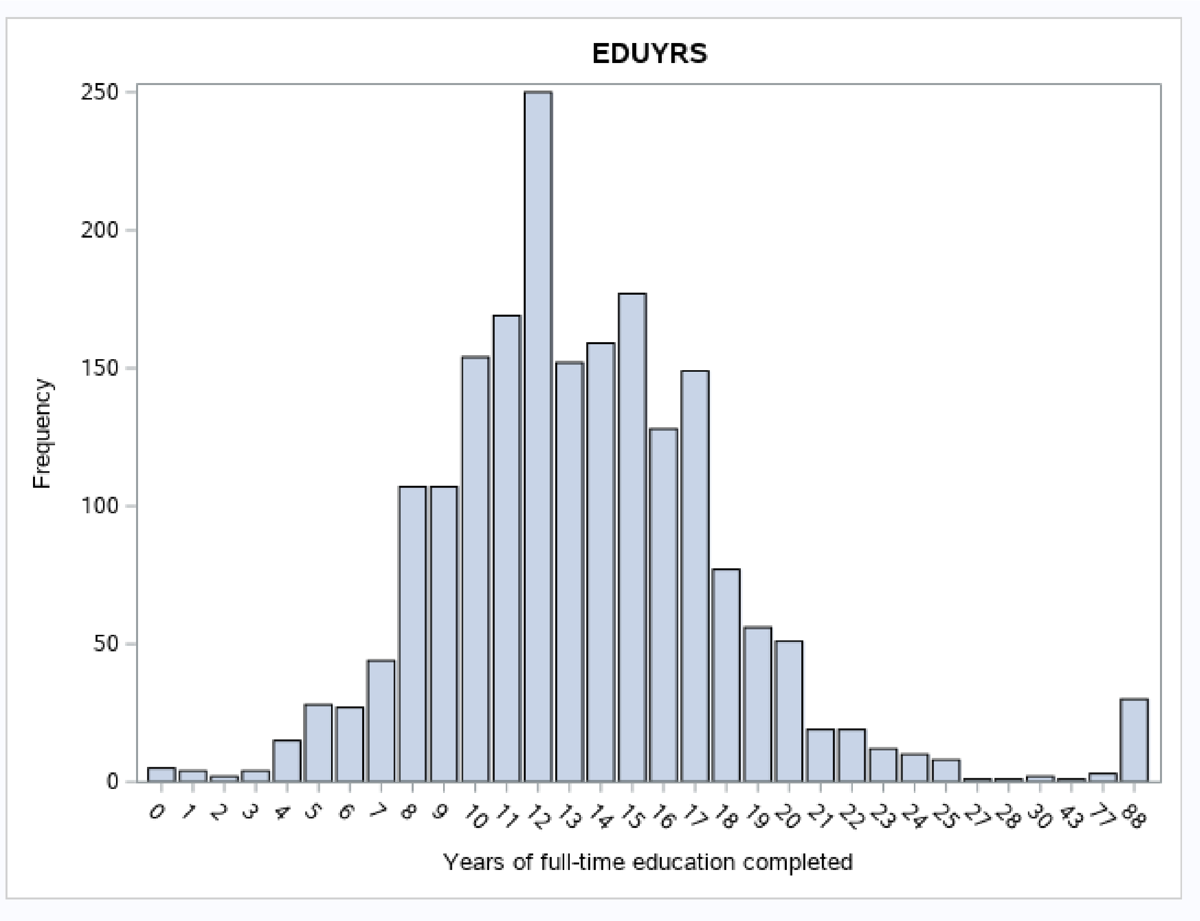
EDUYRS

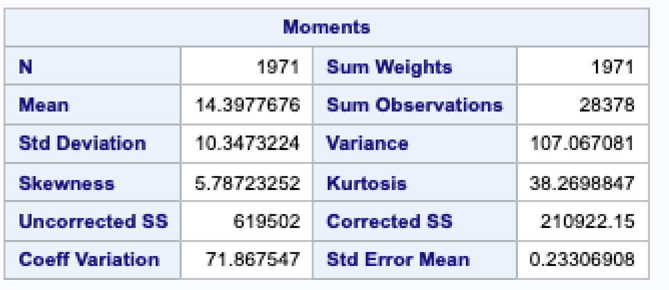
The MEANS Procedure

Analysis Variable : eduyrs Years of full-time education completed
N Miss
0

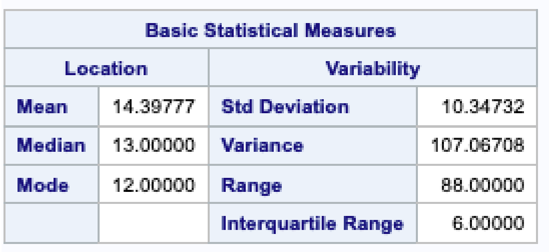


Years of full-time education completed				
eduyrs	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5	0.25	5	0.25
1	4	0.20	9	0.46
2	2	0.10	11	0.56
3	4	0.20	15	0.76
4	15	0.76	30	1.52
5	28	1.42	58	2.94
6	27	1.37	85	4.31
7	44	2.23	129	6.54
8	107	5.43	236	11.97
9	107	5.43	343	17.40
10	154	7.81	497	25.22
11	169	8.57	666	33.79
12	250	12.68	916	46.47
13	152	7.71	1068	54.19
14	159	8.07	1227	62.25
15	177	8.98	1404	71.23
16	128	6.49	1532	77.73
17	149	7.56	1681	85.29
18	77	3.91	1758	89.19
19	56	2.84	1814	92.03
20	51	2.59	1865	94.62
21	19	0.96	1884	95.59
22	19	0.96	1903	96.55
23	12	0.61	1915	97.16
24	10	0.51	1925	97.67
25	8	0.41	1933	98.07
27	1	0.05	1934	98.12
28	1	0.05	1935	98.17
30	2	0.10	1937	98.27
43	1	0.05	1938	98.33
77	3	0.15	1941	98.48
88	30	1.52	1971	100.00

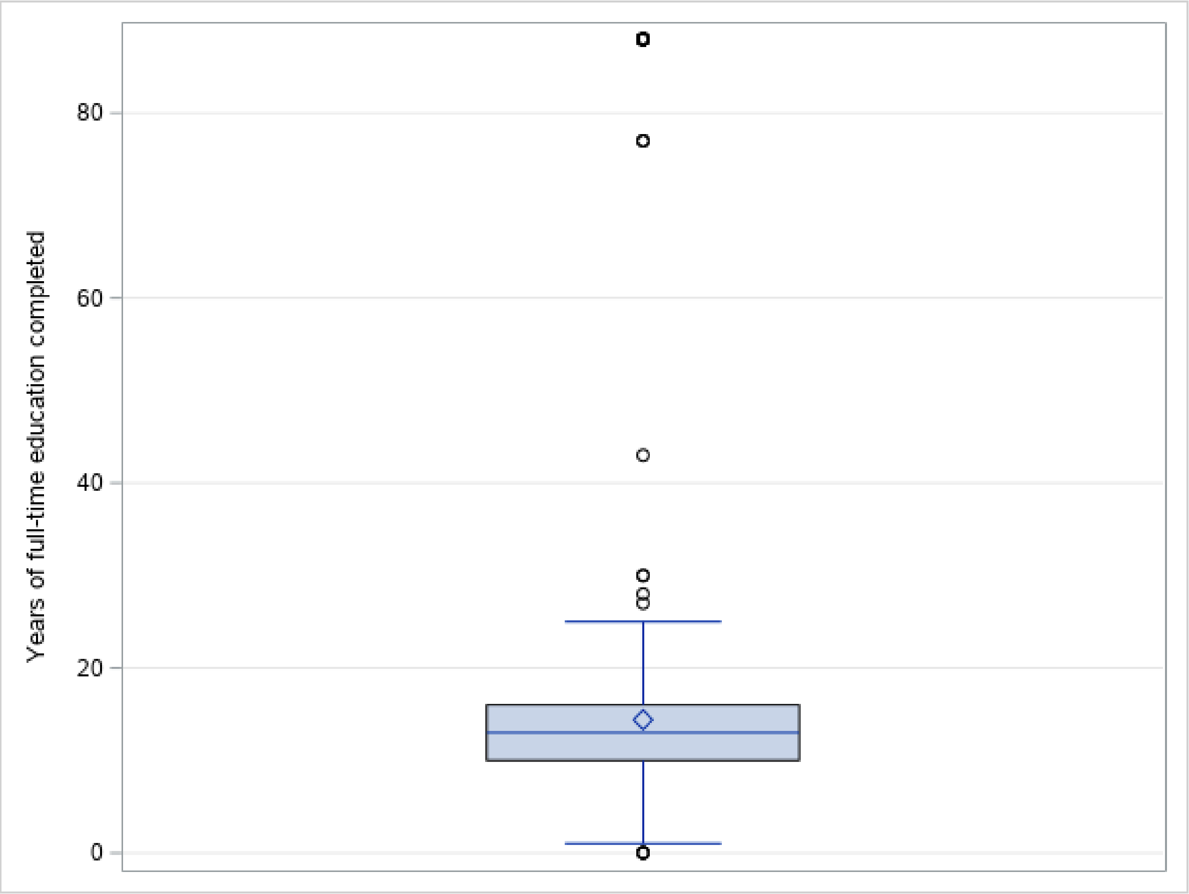


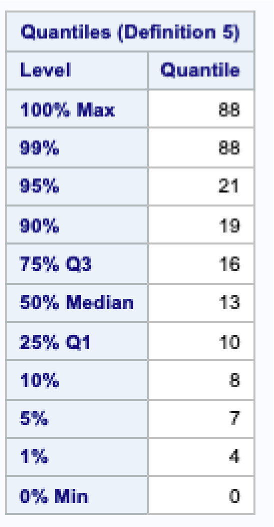


Moments			
N	1971	Sum Weights	1971
Mean	14.3977676	Sum Observations	28378
Std Deviation	10.3473224	Variance	107.067081
Skewness	5.78723252	Kurtosis	38.2698847
Uncorrected SS	619502	Corrected SS	210922.15
Coeff Variation	71.867547	Std Error Mean	0.23306908



Basic Statistical Measures			
Location		Variability	
Mean	14.39777	Std Deviation	10.34732
Median	13.00000	Variance	107.06708
Mode	12.00000	Range	88.00000
		Interquartile Range	6.00000

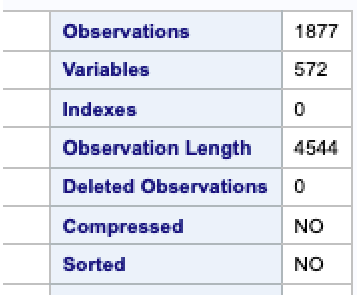




Quantiles (Definition 5)	
Level	Quantile
100% Max	88
99%	88
95%	21
90%	19
75% Q3	16
50% Median	13
25% Q1	10
10%	8
5%	7
1%	4
0% Min	0

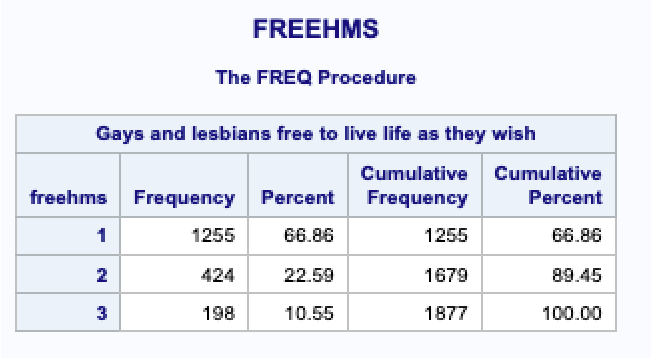
Looking at skewness and kurtosis we can see that distribution is far from normal one. We need to remember however, that this partially caused by including values 77 and 88 in the data. Also, there are a lot of values that seems to be rather uncommon - years of education below 8 and over 25. Also, boxplot support this observation. Taking into account values of the first and last quartile and interquartile range we could cut all observations below 4 and higher than 22 - these are 53 observations. This, with missing values added (also from pplfair) gives us around 100 observations we should get rid of (in worst case - maybe some missing values of pplfair and eduyrs are in the same observations). This however, raises a concern if we will still have enough number of observations in 3rd category of our target variable.

f. Look at the variables after further data reduction



Observations	1877
Variables	572
Indexes	0
Observation Length	4544
Deleted Observations	0
Compressed	NO
Sorted	NO

As we can see, we lost 94 observations in total. Let’s if it influenced our variables.

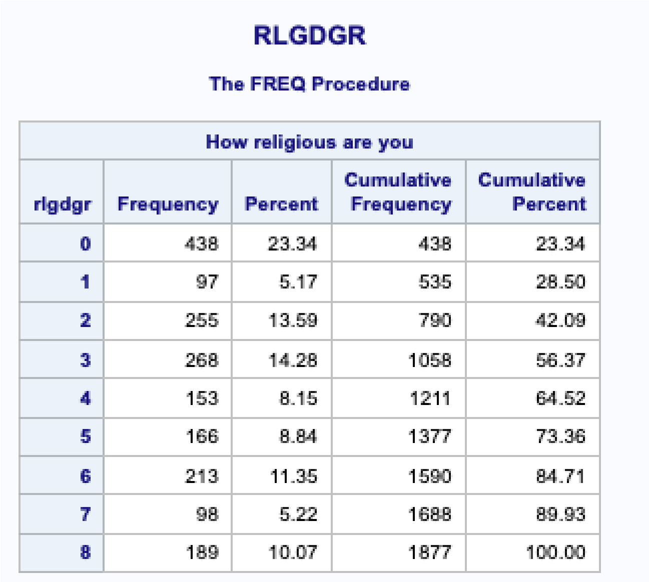


FREEHMS

The FREQ Procedure

Gays and lesbians free to live life as they wish				
freehms	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1255	66.86	1255	66.86
2	424	22.59	1679	89.45
3	198	10.55	1877	100.00

There is no significant difference between target before and after data deletion (https://www.notion.so/Project-notes-be06953a0dc0480f876b638a801e29c6#beec3bc99cee444898307b3792baa3e8) (25, 18 and 11 observations lost in each of the categories 1 -3 respectively)

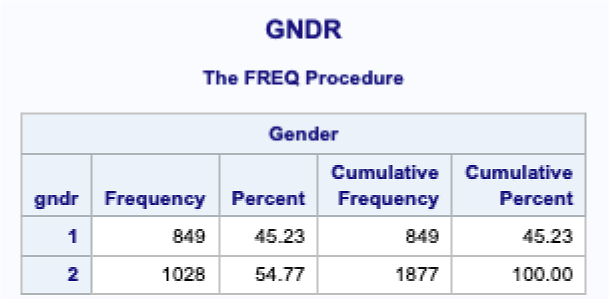


RLGDGR

The FREQ Procedure

How religious are you				
rlgdgr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	438	23.34	438	23.34
1	97	5.17	535	28.50
2	255	13.59	790	42.09
3	268	14.28	1058	56.37
4	153	8.15	1211	64.52
5	166	8.84	1377	73.36
6	213	11.35	1590	84.71
7	98	5.22	1688	89.93
8	189	10.07	1877	100.00

As we can see, there’s noticeable loss in in the 1 and 7 category of RLGDGR(https://www.notion.so/Project-notes-be06953a0dc0480f876b638a801e29c6#beec3bc99cee444898307b3792baa3e8). However, we find 97 and 98 as close enough to 100 to keep this variable as it is, especially that these values are close to the previous values, before deleting the observations with missing data.

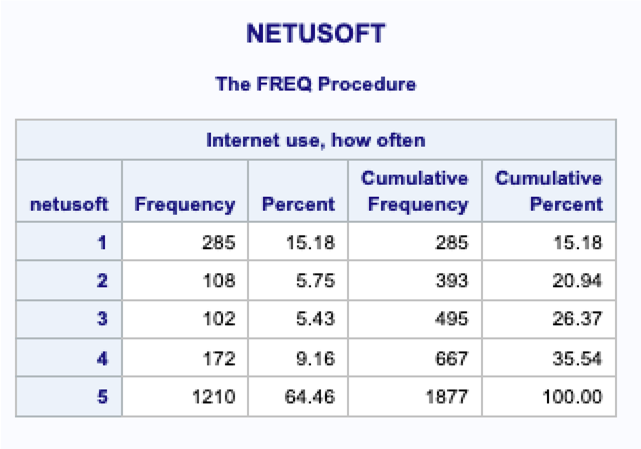


GNDR

The FREQ Procedure

Gender				
gndr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	849	45.23	849	45.23
2	1028	54.77	1877	100.00

Proportion of genders remains at similar level (https://www.notion.so/Project-notes-be06953a0dc0480f876b638a801e29c6#ff1a795de8114257858377 53e5c00f0a) Proportion around 83% remains.

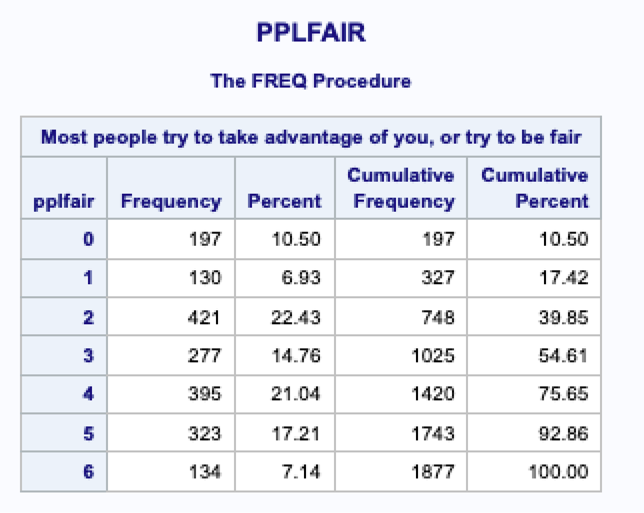


NETUSOFT

The FREQ Procedure

Internet use, how often				
netusoft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	285	15.18	285	15.18
2	108	5.75	393	20.94
3	102	5.43	495	26.37
4	172	9.16	667	35.54
5	1210	64.46	1877	100.00

NETUSOFT also remained relatively unchanged. All categories are properly poulated, and their relative frequency remains at similar level (https://www.notion.so/Project-notes-be06953a0dc0480f876b638a801e29c6#d35693b3e46a43e2980068ca17ac2dab)

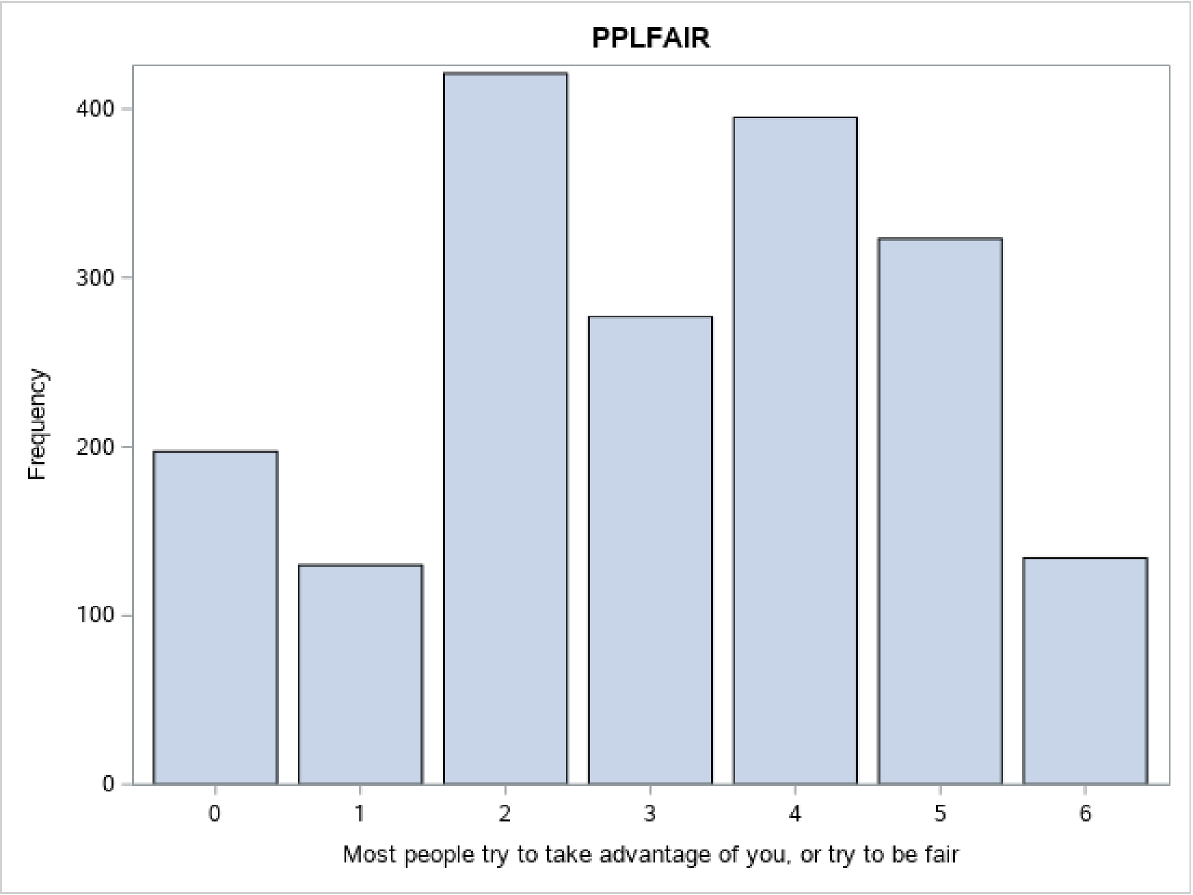


PPLFAIR

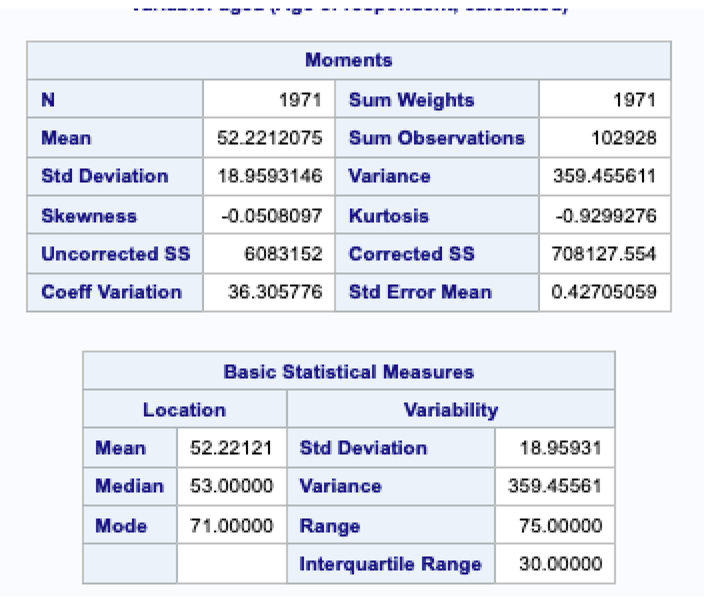
The FREQ Procedure

Most people try to take advantage of you, or try to be fair				
pplfair	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	197	10.50	197	10.50
1	130	6.93	327	17.42
2	421	22.43	748	39.85
3	277	14.76	1025	54.61
4	395	21.04	1420	75.65
5	323	17.21	1743	92.86
6	134	7.14	1877	100.00

https://www.notion.so/Project-notes-be06953a0dc0480f876b638a801e29c6#564c50349fdb4d08a0608b85258f083a



After data reduction, pplfair still has it’s property of most of the observations focused over center and right side. Also, now categories are populated with enough amount of data.

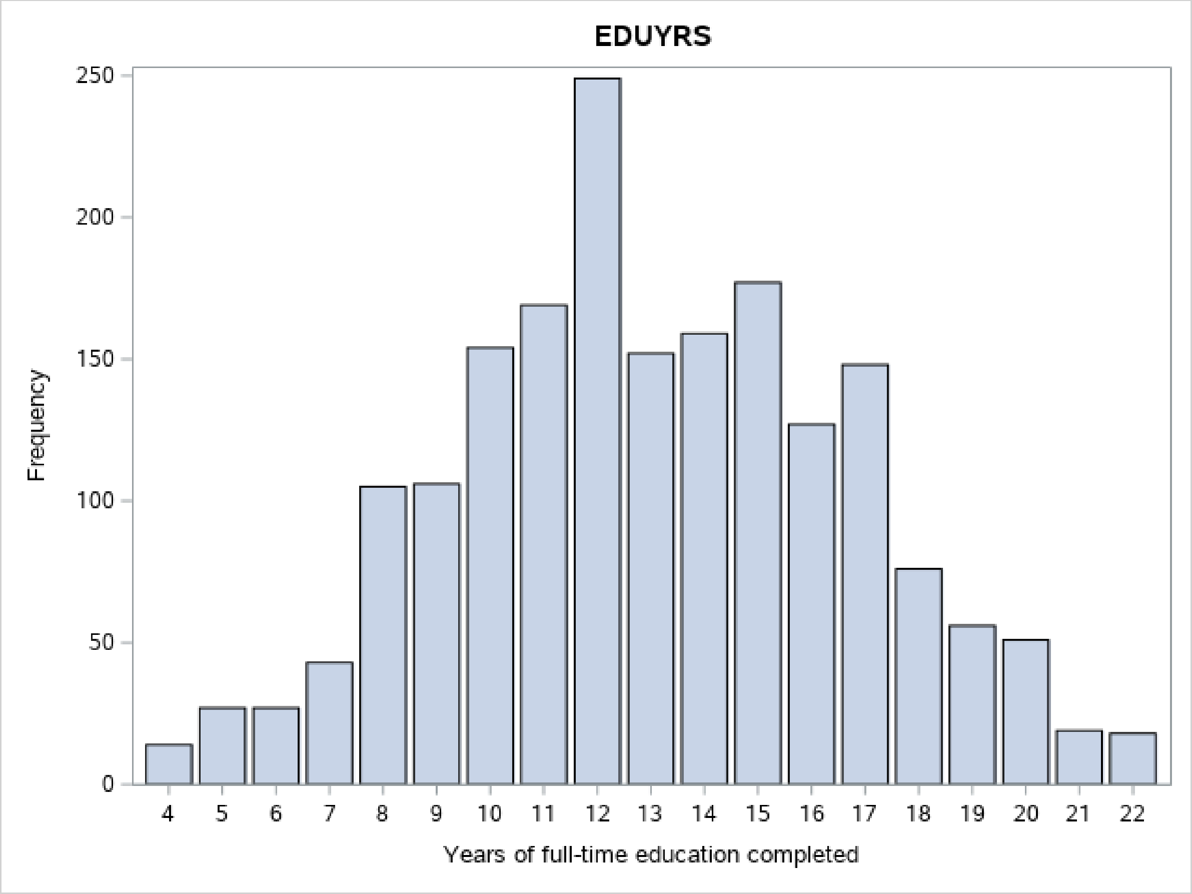


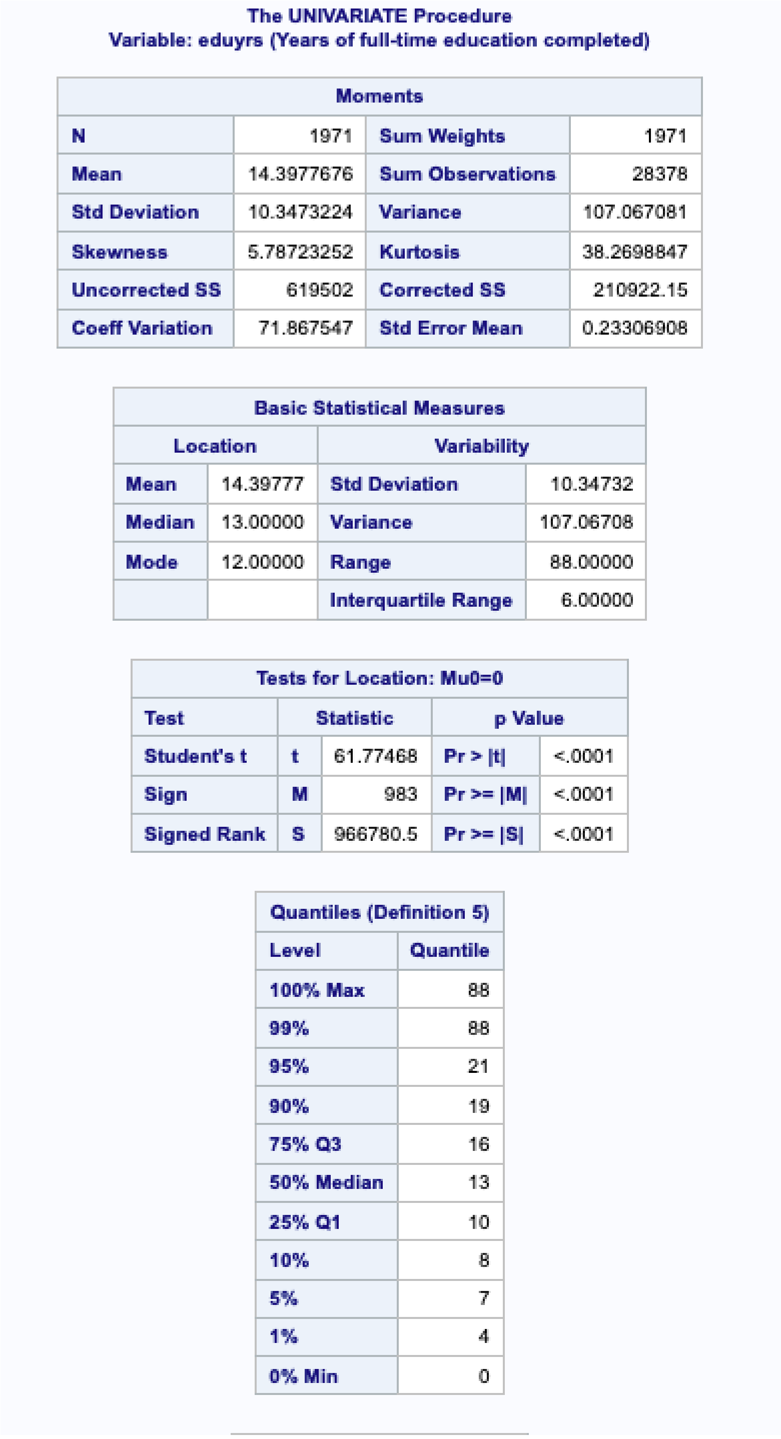
Variable: agea (Age of respondent, calculated)

Moments			
N	1971	Sum Weights	1971
Mean	52.2212075	Sum Observations	102928
Std Deviation	18.9593146	Variance	359.455611
Skewness	-0.0508097	Kurtosis	-0.9299276
Uncorrected SS	6083152	Corrected SS	708127.554
Coeff Variation	36.305776	Std Error Mean	0.42705059

Basic Statistical Measures			
Location		Variability	
Mean	52.22121	Std Deviation	18.95931
Median	53.00000	Variance	359.45561
Mode	71.00000	Range	75.00000
		Interquartile Range	30.00000

Age preserved its close to normality, regular distribution.





The UNIVARIATE Procedure
Variable: eduyrs (Years of full-time education completed)

Moments			
N	1971	Sum Weights	1971
Mean	14.3977676	Sum Observations	28378
Std Deviation	10.3473224	Variance	107.067081
Skewness	5.78723252	Kurtosis	38.2698847
Uncorrected SS	619502	Corrected SS	210922.15
Coeff Variation	71.867547	Std Error Mean	0.23306908

Basic Statistical Measures			
Location		Variability	
Mean	14.39777	Std Deviation	10.34732
Median	13.00000	Variance	107.06708
Mode	12.00000	Range	88.00000
		Interquartile Range	6.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	61.77468	Pr > |t|	<.0001
Sign	M	983	Pr >= |M|	<.0001
Signed Rank	S	966780.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	88
99%	88
95%	21
90%	19
75% Q3	16
50% Median	13
25% Q1	10
10%	8
5%	7
1%	4
0% Min	0

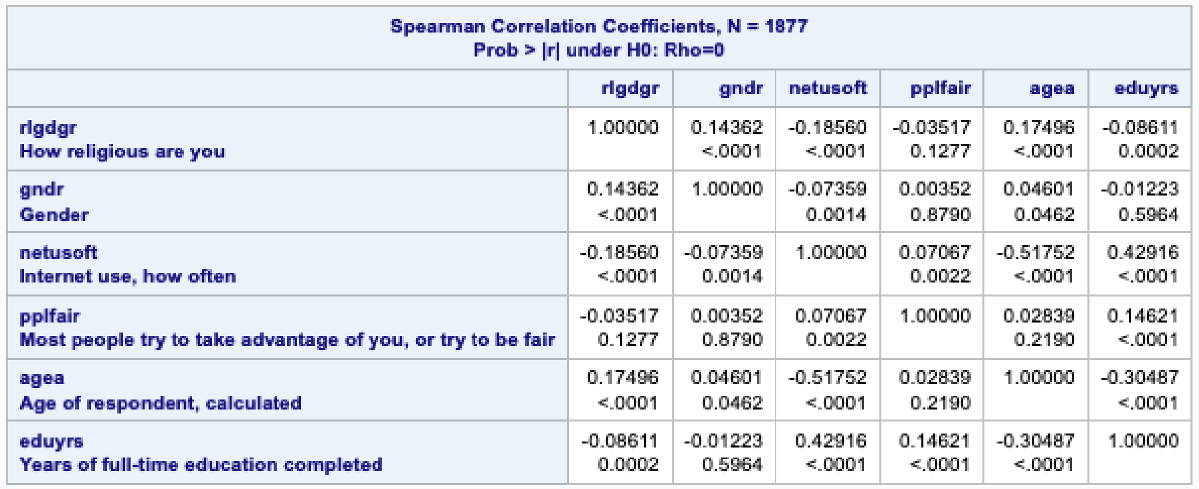
https://www.notion.so/Project-notes-be06953a0dc0480f876b638a801e29c6#b2c4500aec8d4fecb07309094c10f36b

Histogram has now much more regular shape, however values of basic statistics did not change much. This is probably due to small number of outliers and missing values. However, due to not so big change in other variables, we can keep this dataset, and compare a model built on it, with model built using previous one.

Collinearity assessment

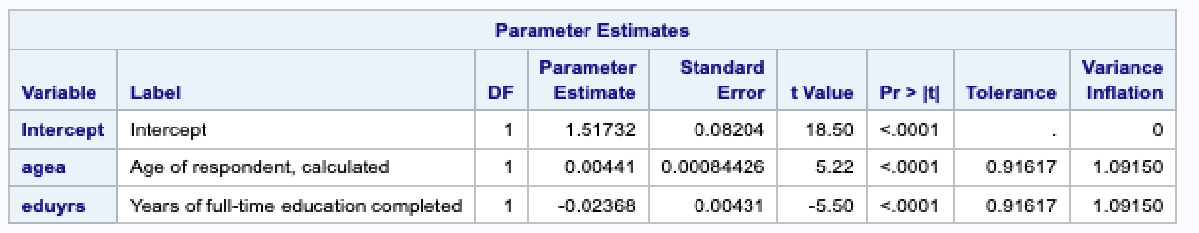
in this part we’re going to check if our explanatory variables are correlated with each other. If they are then we will have to exclude some of them, as collinearity among explanatory variables can result in unstable model.

Since we are having ordinal and ratio variables, we are going to use Pearson corerrlation for ratio variables, and Spearman correlation for the rest of them.



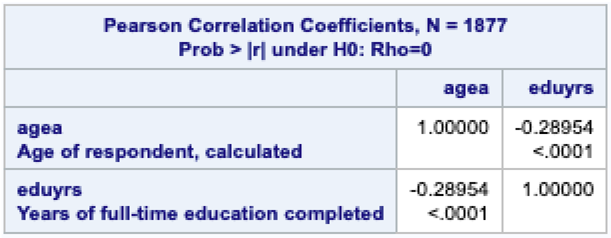
Spearman Correlation Coefficients, N = 1877
Prob > |r| under H0: Rho=0

	rlgdgr	gndr	netusoft	pplfair	agea	eduyrs
rlgdgr **How religious are you**	1.00000	0.14362 <.0001	-0.18560 <.0001	-0.03517 0.1277	0.17496 <.0001	-0.08611 0.0002
gndr **Gender**	0.14362 <.0001	1.00000	-0.07359 0.0014	0.00352 0.8790	0.04601 0.0462	-0.01223 0.5964
netusoft **Internet use, how often**	-0.18560 <.0001	-0.07359 0.0014	1.00000	0.07067 0.0022	-0.51752 <.0001	0.42916 <.0001
pplfair **Most people try to take advantage of you, or try to be fair**	-0.03517 0.1277	0.00352 0.8790	0.07067 0.0022	1.00000	0.02839 0.2190	0.14621 <.0001
agea **Age of respondent, calculated**	0.17496 <.0001	0.04601 0.0462	-0.51752 <.0001	0.02839 0.2190	1.00000	-0.30487 <.0001
eduyrs **Years of full-time education completed**	-0.08611 0.0002	-0.01223 0.5964	0.42916 <.0001	0.14621 <.0001	-0.30487 <.0001	1.00000



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	1.51732	0.08204	18.50	<.0001	.	0
agea	Age of respondent, calculated	1	0.00441	0.00084426	5.22	<.0001	0.91617	1.09150
eduyrs	Years of full-time education completed	1	-0.02368	0.00431	-5.50	<.0001	0.91617	1.09150



Pearson Correlation Coefficients, N = 1877
Prob > |r| under H0: Rho=0

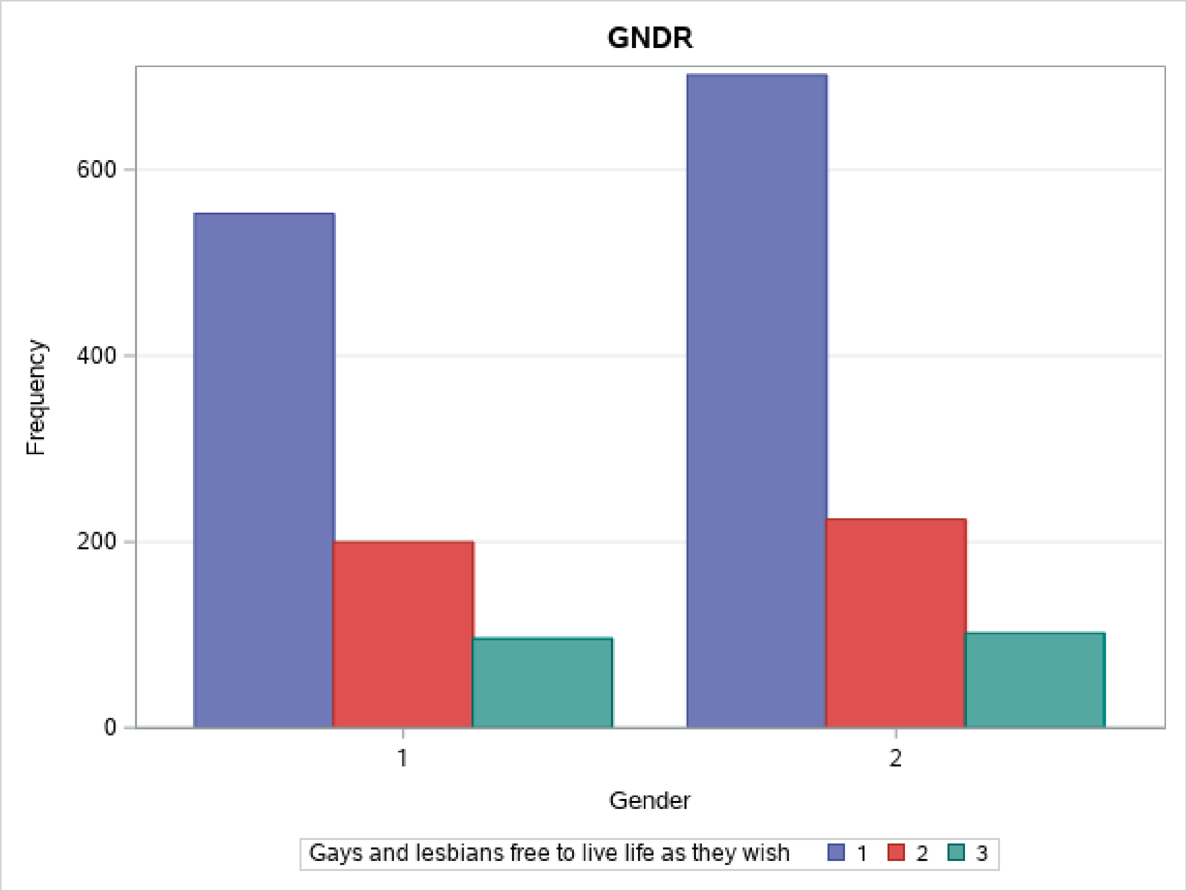
	agea	eduyrs
agea **Age of respondent, calculated**	1.00000	-0.28954 <.0001
eduyrs **Years of full-time education completed**	-0.28954 <.0001	1.00000

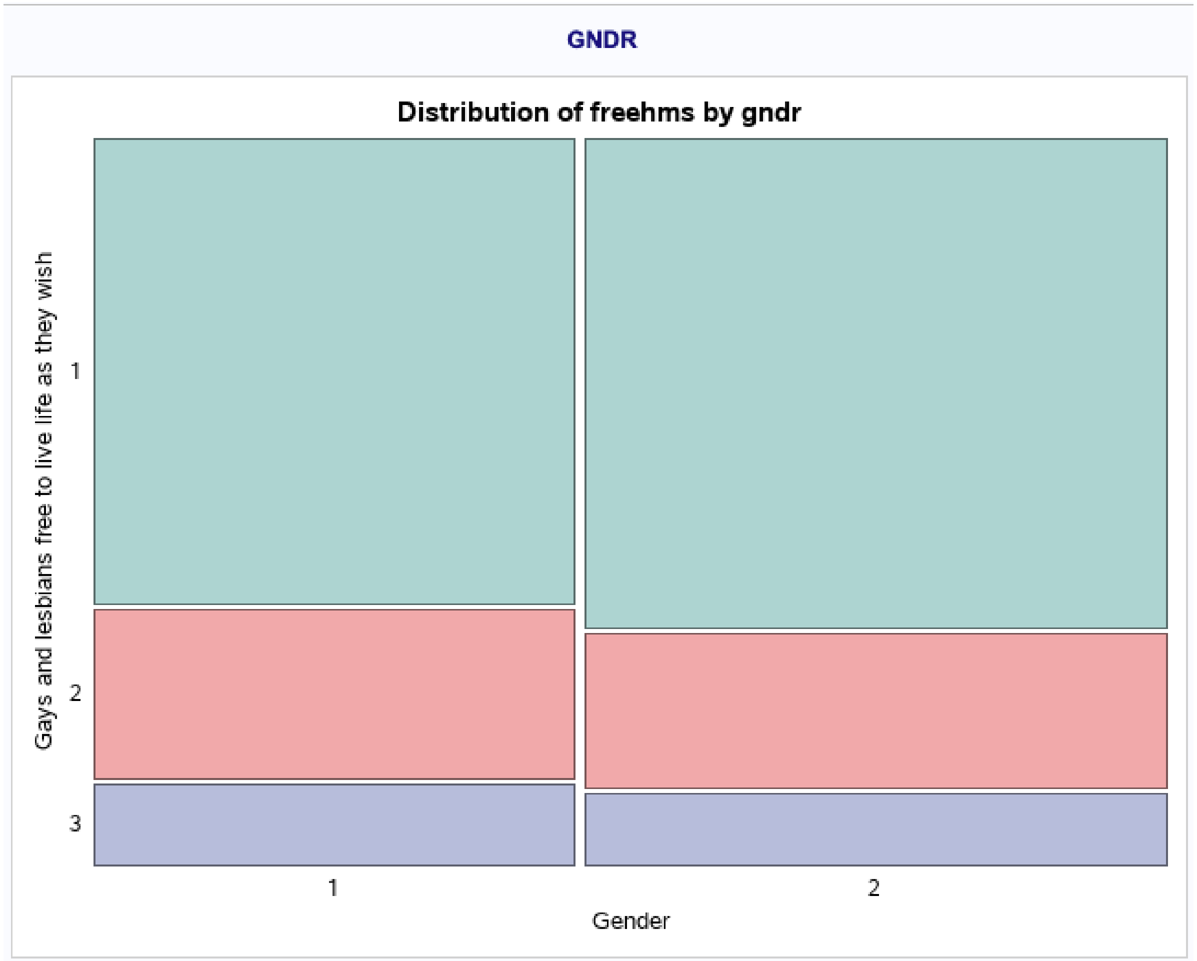
In most of the cases we have statistically significants results. However, we are not worried about this, as for many observations in the dataset it is common to obtain statistically significant results of correlation. Additionally, correlation coefficients are mostly low. There is only one coefficient that we find worrying, and this is Spearman correlation between NETUSOFT and AGEA variables. After consideration we decided to exclude NETUSOFT variable from our model.

Variables AGEA and EDUYRS have statistically important Pearson correlation. However, correlation coefficient is low, and variance inflection factor, does not indicate multicollinearity (we assume that it would indicate multicollinearity if it was equal to 10 or bigger)

Discriminatory performance analysis

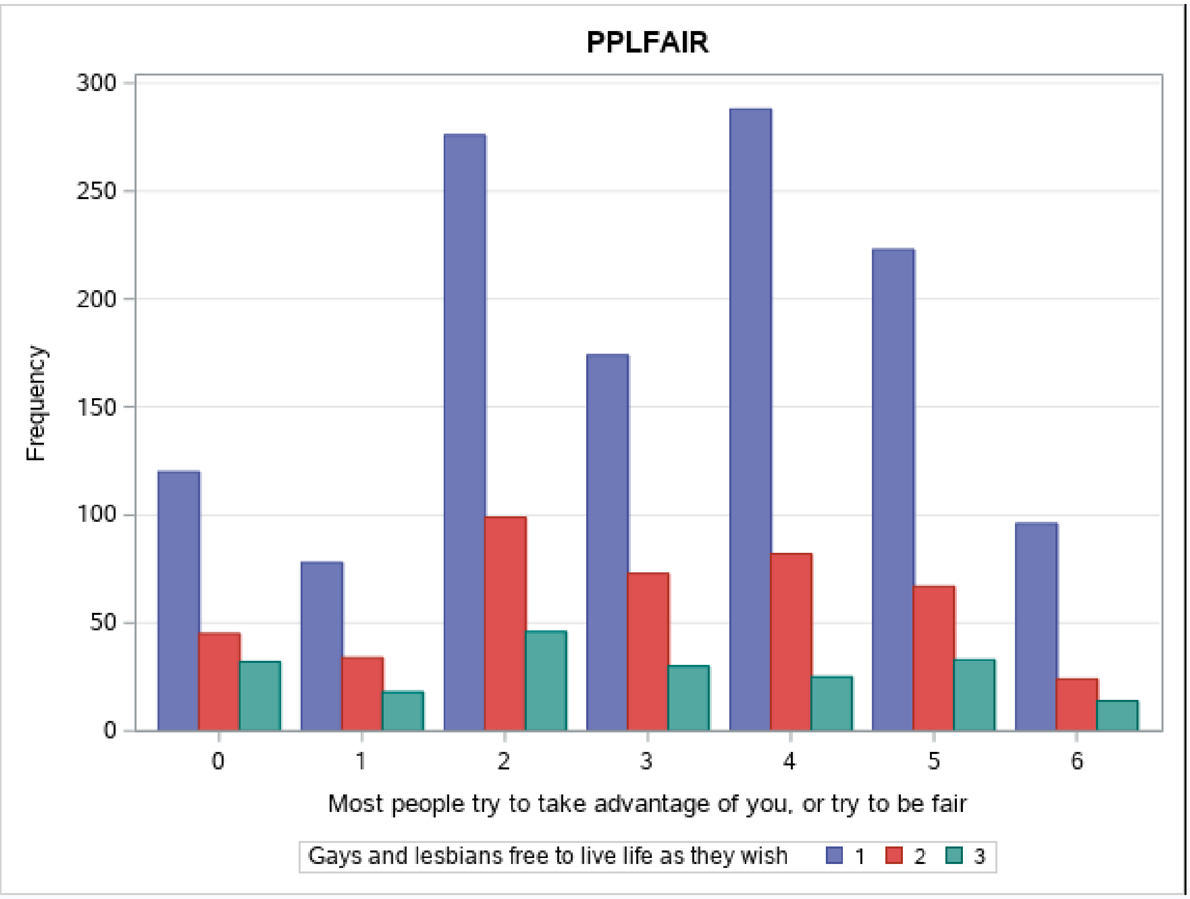
a. GNDR

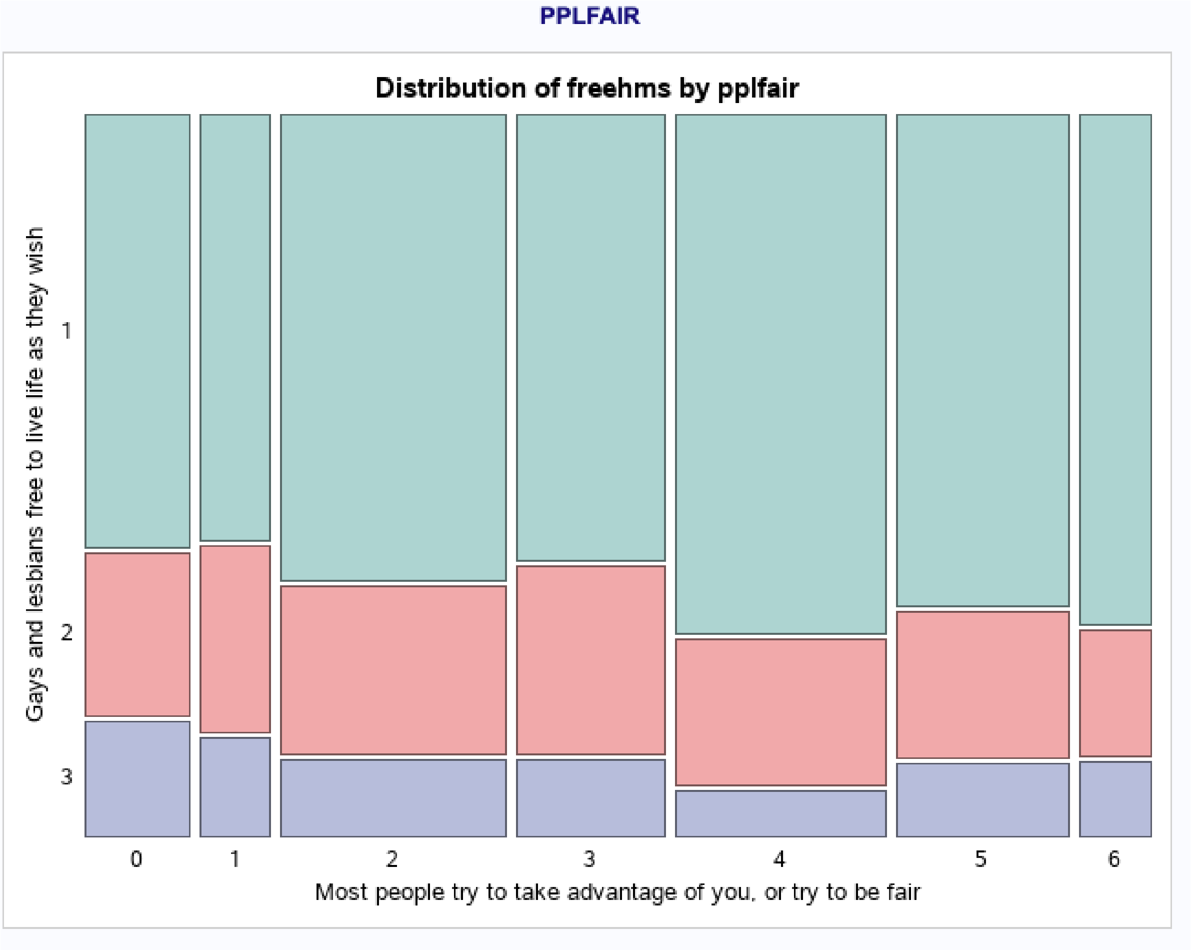




Categories of target variable are distributed rather evenly across GNDR variable. Females answer agreed or strongly agreed a bit more often than men (and men answered negatively or neutral slightly more often than women). It doesn’t look like significant difference though.

b. PPLFAIR


PPLFAIR
Frequency
300
250
200
150
100
50
0
0
1
2
3
4
5
6
Most people try to take advantage of you, or try to be fair
Gays and lesbians free to live life as they wish
1
2
3



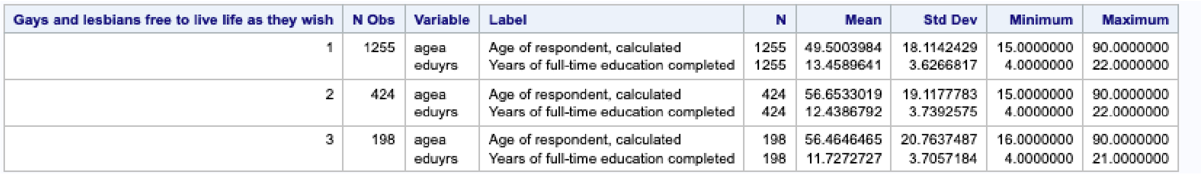
Looking at the bar chart we can see that data distribution for each of each of the FREEHMS categories resembles the distribution of PPLFAIR variable itself.

On mosaic plot, we can see that number observations of category 1 of FREEHMS raises with rise of PPLFAIR category.

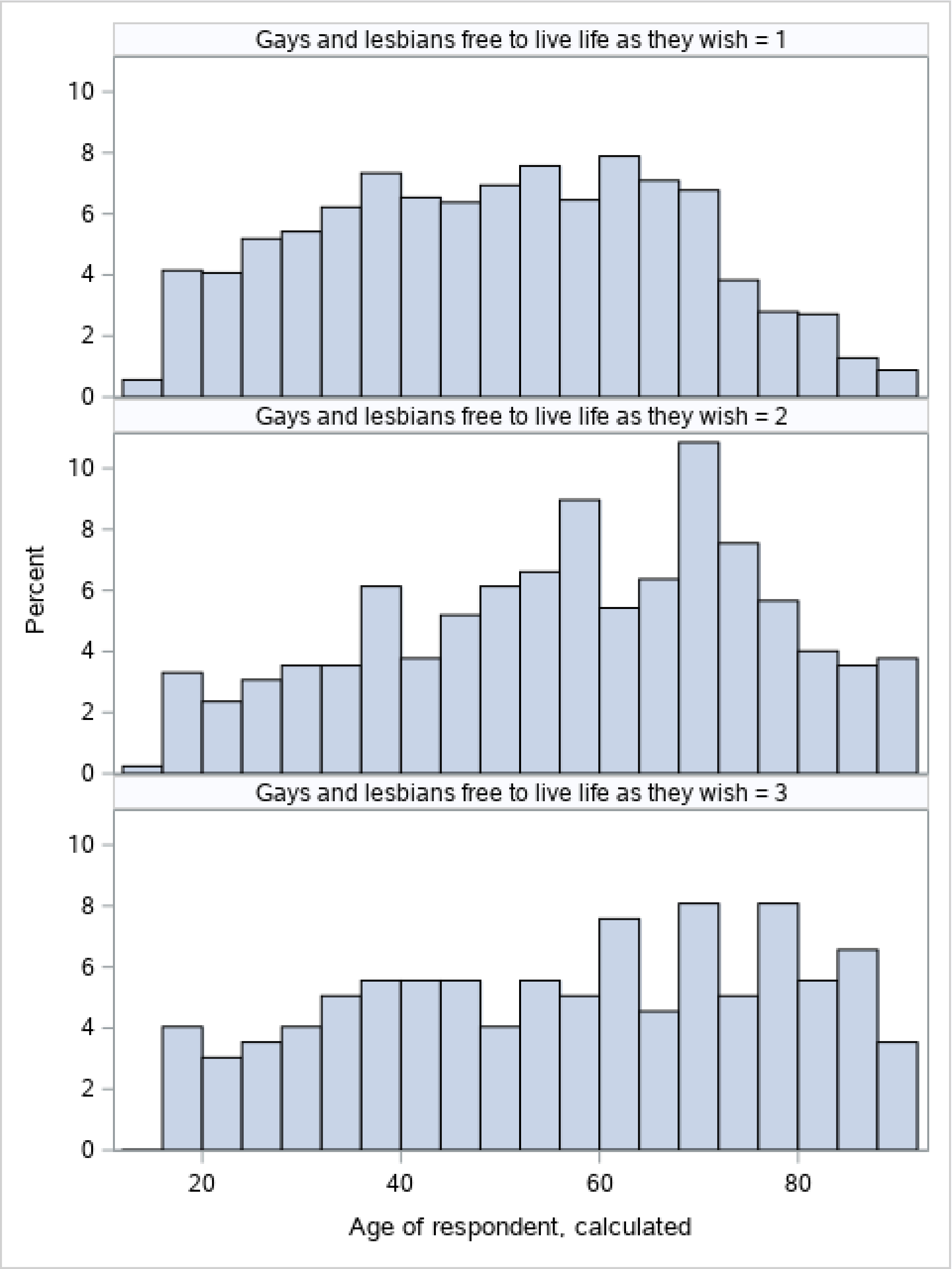
The opposite phenomena seems to be happening for 3rd category of FREEHMS variable.

It is hard to spot any tendency for the 2nd category of FREEHMS.

c. AGEA

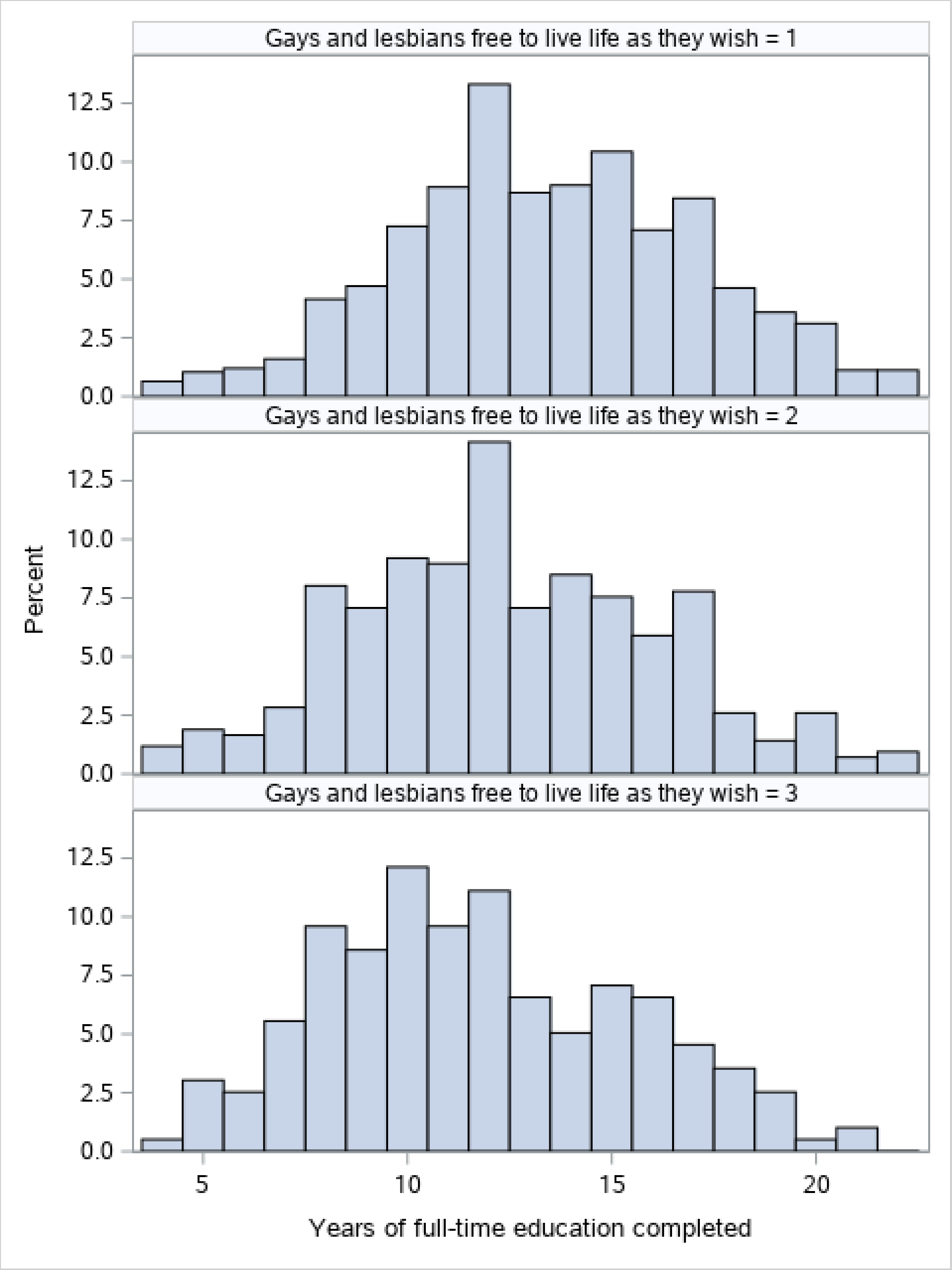


Gays and lesbians free to live life as they wish	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
1	1255	agea eduyrs	Age of respondent, calculated Years of full-time education completed	1255 1255	49.5003984 13.4589641	18.1142429 3.6266817	15.0000000 4.0000000	90.0000000 22.0000000
2	424	agea eduyrs	Age of respondent, calculated Years of full-time education completed	424 424	56.6533019 12.4386792	19.1177783 3.7392575	15.0000000 4.0000000	90.0000000 22.0000000
3	198	agea eduyrs	Age of respondent, calculated Years of full-time education completed	198 198	56.4646465 11.7272727	20.7637487 3.7057184	16.0000000 4.0000000	90.0000000 21.0000000



We can observe a raise of a mean age when we move from target category 1 to 2. There is no big change in that value between categories 2 and 3, however when we look at the histogram, we can observe a small shift towards older age.

d. EDUYRS



This variable seems to behave in opposite way. Mean values of years of education completed decrease along with growing categories of FREEHMS variable. Here the effect is more clear than in case of AGEA, however we can’t tell that either of these effects is strong.

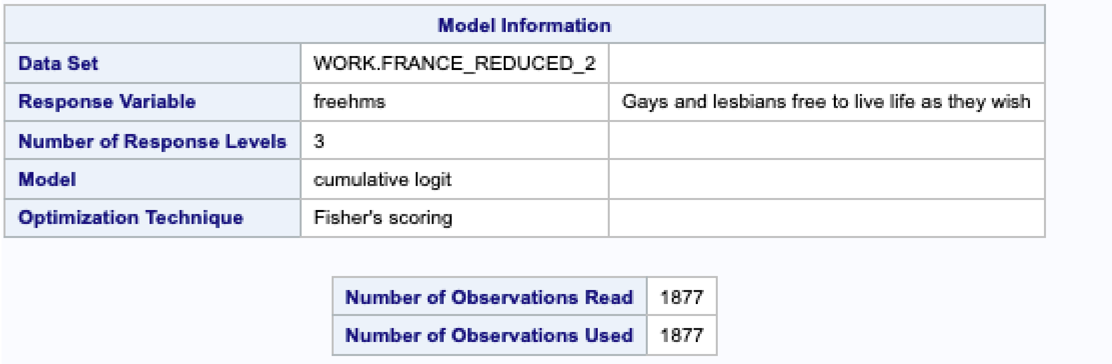
To conclude: In this chapter we conducted descriptive and discriminatory performance data analysis. We saw a distribution of each variable, reduced number of categories, deleted missing data and outliers if needed. Then we conducted collinearity assessment, what resulted in excluding NETUSOFT variable. At the end, we conducted discriminatory performance analysis, what showed us some minor, but interesting associations between our target and explanatory variables. These are:

1. RLGDGR: decrease in number of positive responses and increase of number of neutral or negative responses, with growing religiosity.
2. GNDR: Females answer agreed or strongly agreed a bit more often than men (and men answered negatively or neutral slightly more often than women)
3. PPLFAIR: The bigger trust in people, the more strongly agreeing answers and less neutral and negative ones.
4. AGEA: Mean age of the responded increased for less positive answers
5. EDUYRS: Mean years of full-time education completed decreased with less positive answers

Now, after data data cleaning and description, we are ready to build a logistic regression model, and analyse its results.

Substantiative analysis

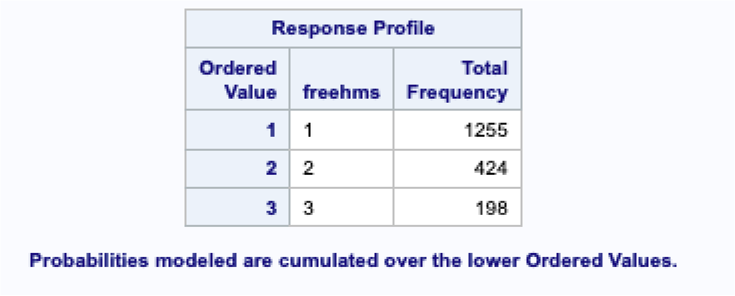
Overview



Model Information		
Data Set	WORK.FRANCE_REDUCED_2	
Response Variable	freehms	Gays and lesbians free to live life as they wish
Number of Response Levels	3	
Model	cumulative logit	
Optimization Technique	Fisher's scoring	

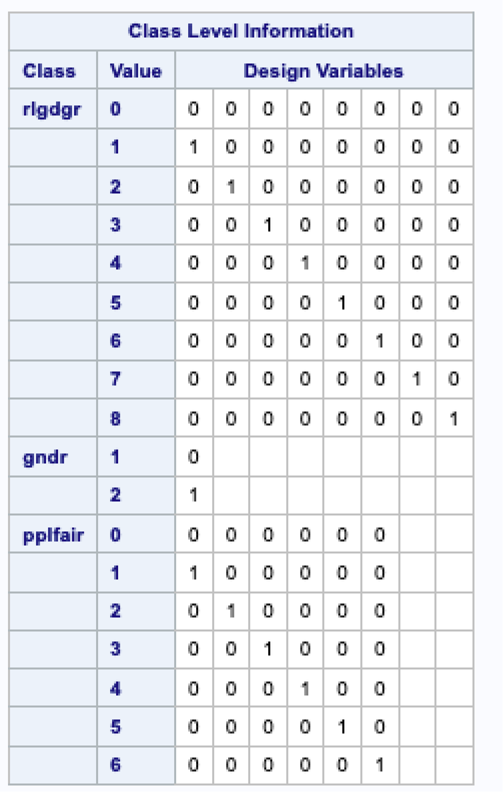
Number of Observations Read	1877
Number of Observations Used	1877

All observations were used, that means we did not omit any missing data during descriptive analysis.



Response Profile		
Ordered Value	freehms	Total Frequency
1	1	1255
2	2	424
3	3	198

Probabilities modeled are cumulated over the lower Ordered Values.



Class Level Information									
Class	Value	Design Variables							
rlgdgr	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0
	4	0	0	0	1	0	0	0	0
	5	0	0	0	0	1	0	0	0
	6	0	0	0	0	0	1	0	0
	7	0	0	0	0	0	0	1	0
	8	0	0	0	0	0	0	0	1
gndr	1	0							
	2	1							
pplfair	0	0	0	0	0	0	0		
	1	1	0	0	0	0	0		
	2	0	1	0	0	0	0		
	3	0	0	1	0	0	0		
	4	0	0	0	1	0	0		
	5	0	0	0	0	1	0		
	6	0	0	0	0	0	1		

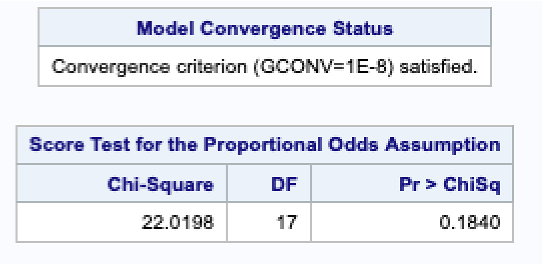
Our reference categories are:

a. RLGDGR - 0 - not at all religious

b. GNDR - 1 - Male

c. PPLFAIR - 0 - Most people try to take advantage of me

Proportional odds assumption test

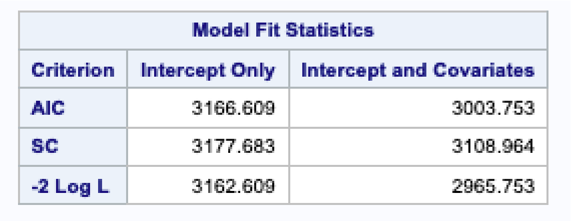


Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

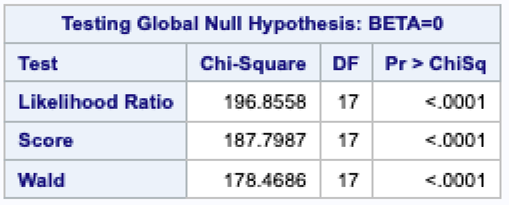
Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
22.0198	17	0.1840

Convergence criterion is satisfied. Also, we can see that p value of proportional odds assumption is higher than 0.05. That means we cannot reject the null hypothesis, that odds are proportional. This means, we can safely analyse estimates and odds ratios as we don’t have to take into account specific intercept of a certain cutoff point of the target variable (we’ve got 2 cut-off points One between category 1 and 2 with 3, and second point is between categories 1 with 2 and 3. Thanks to satisfying proportional odds assumption we don’t need to take into account one of these 2 cutoff points, as odds are proportional.)

Global beta test and model fit statistics



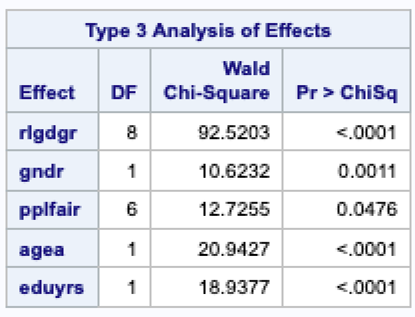
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3166.609	3003.753
SC	3177.683	3108.964
-2 Log L	3162.609	2965.753



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	196.8558	17	<.0001
Score	187.7987	17	<.0001
Wald	178.4686	17	<.0001

All results of likelihood ratio, Score and Wald test are statistically significant. That means, there’s at least one variable in our model that explains the target in statistically significant way.

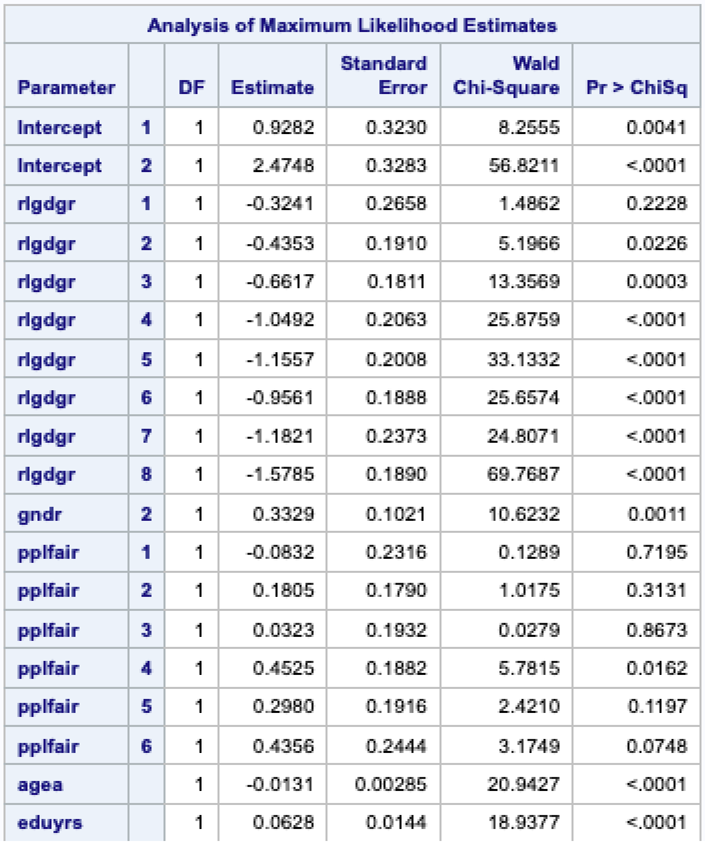
Type 3 analysis of effects



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
rlgdgr	8	92.5203	<.0001
gndr	1	10.6232	0.0011
pplfair	6	12.7255	0.0476
agea	1	20.9427	<.0001
eduyrs	1	18.9377	<.0001

All variables explain our target in a statistically significant way. However, it is worth to note, that PPLFAIR variable is vary close to alpha value (0.05).

Analysis of maximum likelihood estimates



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	0.9282	0.3230	8.2555	0.0041
Intercept	2	1	2.4748	0.3283	56.8211	<.0001
rlgdgr	1	1	-0.3241	0.2658	1.4862	0.2228
rlgdgr	2	1	-0.4353	0.1910	5.1966	0.0226
rlgdgr	3	1	-0.6617	0.1811	13.3569	0.0003
rlgdgr	4	1	-1.0492	0.2063	25.8759	<.0001
rlgdgr	5	1	-1.1557	0.2008	33.1332	<.0001
rlgdgr	6	1	-0.9561	0.1888	25.6574	<.0001
rlgdgr	7	1	-1.1821	0.2373	24.8071	<.0001
rlgdgr	8	1	-1.5785	0.1890	69.7687	<.0001
gndr	2	1	0.3329	0.1021	10.6232	0.0011
pplfair	1	1	-0.0832	0.2316	0.1289	0.7195
pplfair	2	1	0.1805	0.1790	1.0175	0.3131
pplfair	3	1	0.0323	0.1932	0.0279	0.8673
pplfair	4	1	0.4525	0.1882	5.7815	0.0162
pplfair	5	1	0.2980	0.1916	2.4210	0.1197
pplfair	6	1	0.4356	0.2444	3.1749	0.0748
agea		1	-0.0131	0.00285	20.9427	<.0001
eduyrs		1	0.0628	0.0144	18.9377	<.0001

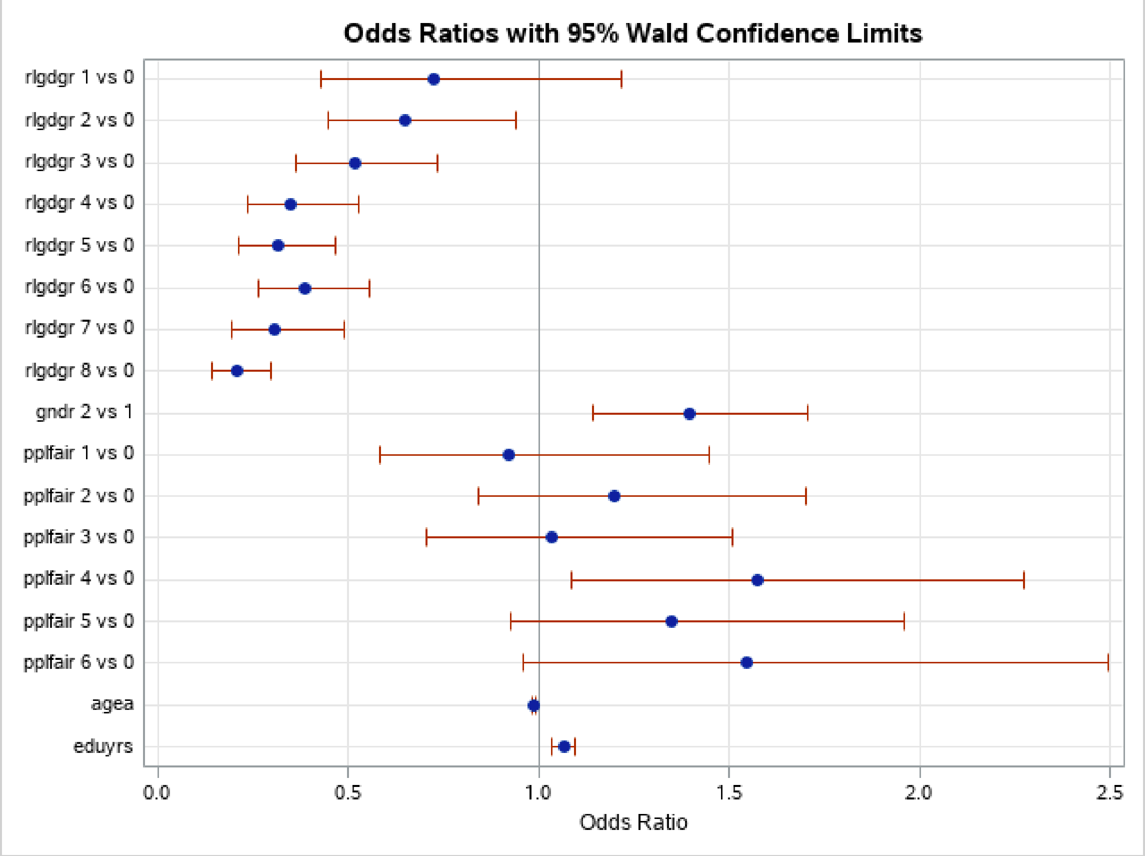
There are 6 statistically unimportant estimates. These are

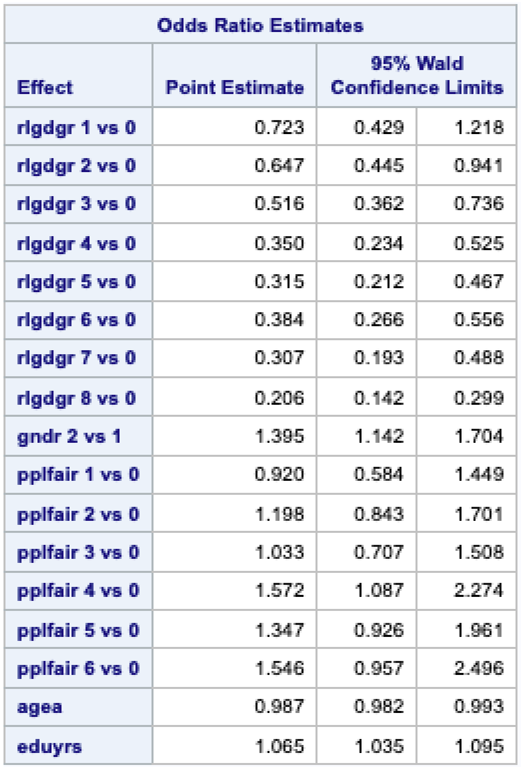
1. RLGDGR 1 vs 0, which means there is no difference if a person defines herself as 0 or 1 in terms of religiosity. The rest of religiosity categories are significant and they are inhibiting the probability of being in a lower category of FREEHMS (inhibiting the probability of more positive opinion about freedom of homosexual people)

2. PPLFAIR from 1 to 3 and 5-6 vs 0. That means that only people who answered 4 to that question, are statistically significant in explaining value of the target. This may result from dimensionality reduction. We can interpret this in such a way that people with rather balanced but a bit shifted towards "fairness of people" opinion are stimulating the probability of having more positive opinion about homosexuals living their life as they want. (please bear in mind, that category 4 was a category 7 before dimensionality reduction)

The rest of effects is statistically important. Lets analyse them deeper, looking at their odds.

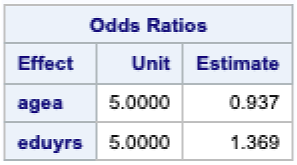
Odds Ratio Estimates analysis





Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
rlgdgr 1 vs 0	0.723	0.429	1.218
rlgdgr 2 vs 0	0.647	0.445	0.941
rlgdgr 3 vs 0	0.516	0.362	0.736
rlgdgr 4 vs 0	0.350	0.234	0.525
rlgdgr 5 vs 0	0.315	0.212	0.467
rlgdgr 6 vs 0	0.384	0.266	0.556
rlgdgr 7 vs 0	0.307	0.193	0.488
rlgdgr 8 vs 0	0.206	0.142	0.299
gndr 2 vs 1	1.395	1.142	1.704
pplfair 1 vs 0	0.920	0.584	1.449
pplfair 2 vs 0	1.198	0.843	1.701
pplfair 3 vs 0	1.033	0.707	1.508
pplfair 4 vs 0	1.572	1.087	2.274
pplfair 5 vs 0	1.347	0.926	1.961
pplfair 6 vs 0	1.546	0.957	2.496
agea	0.987	0.982	0.993
eduyrs	1.065	1.035	1.095



Odds Ratios

Effect	Unit	Estimate
agea	5.0000	0.937
eduyrs	5.0000	1.369

As mentioned earlier not every estimate is statistically significant. Now we will focus only on these significant (they can be easily identified on a plot above - significant estimates' 95% confidence interval does not cross with 1.0 odds ratio line).

Each RLGDGR estimate inhibits probability of obtaining lower FREEHMS category. Moreover, this **inhibition raises along higher RLGDGR category** (except for RLGDGR 6 vs 0). That means, that probability of being in lower FREEHMS category is less by around 35% for observations that are in 2nd RLGDGR category (or by around 6% up to around 55% with 95% confidence) when compared to observations in 0 RLGDGR category (lower FREEHMS category is higher acceptance that homosexual people should live as they wish). This probability decreases with higher RLGDGR categories down to 80% less when being in 8th RLGDGR category, when compared to RLGDGR 0.

It is worth to notice big 95% confidence interval limits for lower RLGDGR values, and their decrease in wideness with higher RLGDGR levels.

Wideness of 95% confidence intervals for RLGDGR estimates

Estimate	Beginning	End	Difference
2 vs 0	0,445	0,941	-0,50
3 vs 0	0,362	0,736	-0,37
4 vs 0	0,234	0,525	-0,29
5 vs 0	0,212	0,467	-0,26
6 vs 0	0,266	0,556	-0,29
7 vs 0	0,193	0,488	-0,30
8 vs 0	0,142	0,299	-0,16

We can notice a decrease in wideness from RLGDGR 2 vs 0 up to 8 vs 0, except for small increase for 6 vs 0, and 8vs 0. It means, that diversity in observed FREEHMS categories across RLGDGR categories was decreasing. We can interpret this phenomena in such a way that people with rising religiosity were more decided to less agree with FREEHMS question.

These estimates are in line with our observations from descriptive analysis.

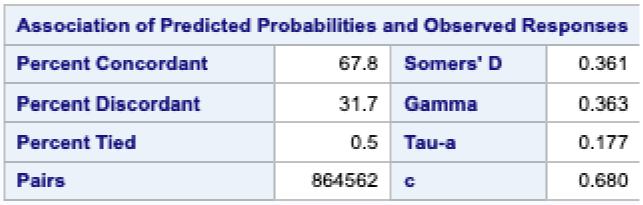
Females have 39,5% bigger chance to be in lower FREEHMS category than men (from around 14% up to around 70% with 95% confidence). These results are also in line with our observations.

It is hard to interpret effect on our target. As we mentioned above, a possible interpretation can be that people being slightly positive about fairness of other people are around 57% more likely to be in lower FREEHMS category when compared to people who agrees with a statement, that other people mostly want to take advantage of them. However considering very wide 95% confidence interval (from around 9% bigger probability up to around 270%, which is more than twice as big probability) and insignificance of the rest of remaining categories’ estimates we consider this variable as insignificant in predicting values FREEHMS variable.

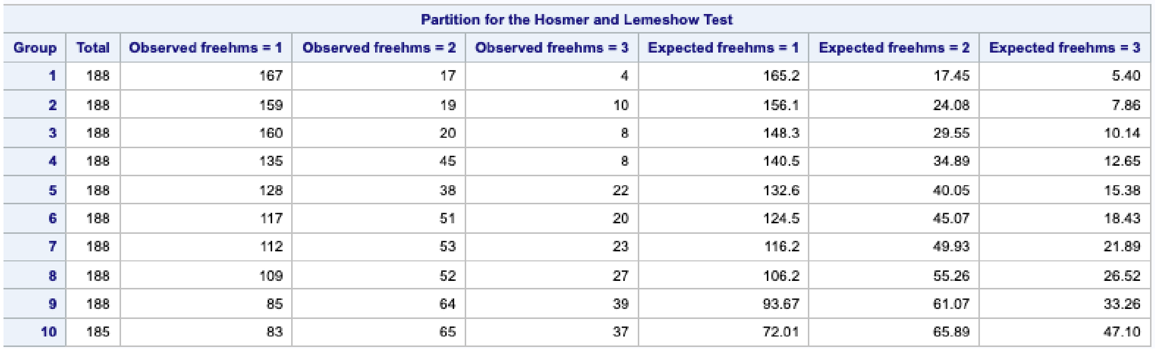
According to our previous expectations age is an inhibitor of lower FREEHMS values. As wee can see person who is 5 years older than another one is around 6% less likely to be in lower FREEHMS category. (or, from 5,5% to 6.6% with 95% confidence).

Also, number of completed years of full-time education explains our target in expected way. Person who completed 5 more years of education has approximately 37% greater probability of being in lower FREEHMS category (from around 33% to 40% with 95% of confidence).

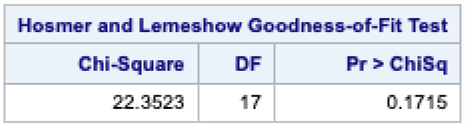
Analysis of predictive power



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	67.8	Somers' D	0.361
Percent Discordant	31.7	Gamma	0.363
Percent Tied	0.5	Tau-a	0.177
Pairs	864562	c	0.680



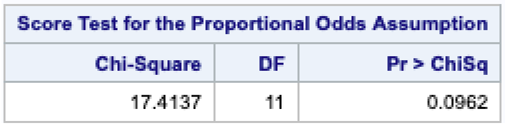
Partition for the Hosmer and Lemeshow Test							
Group	Total	Observed freehms = 1	Observed freehms = 2	Observed freehms = 3	Expected freehms = 1	Expected freehms = 2	Expected freehms = 3
1	188	167	17	4	165.2	17.45	5.40
2	188	159	19	10	156.1	24.08	7.86
3	188	160	20	8	148.3	29.55	10.14
4	188	135	45	8	140.5	34.89	12.65
5	188	128	38	22	132.6	40.05	15.38
6	188	117	51	20	124.5	45.07	18.43
7	188	112	53	23	116.2	49.93	21.89
8	188	109	52	27	106.2	55.26	26.52
9	188	85	64	39	93.67	61.07	33.26
10	185	83	65	37	72.01	65.89	47.10



Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
22.3523	17	0.1715

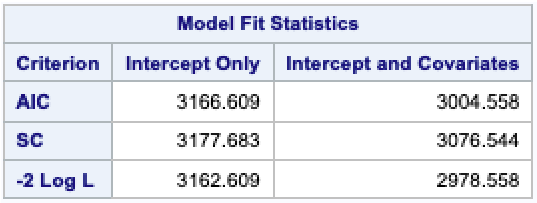
Our model has almost 68% of concordant pairs (pairs that were assigned with lower FREEHMS value, when probability of obtaining such a value was bigger for a given profile in observed data). We consider that as a near to satisfactory number. More importantly, area under the curve (c test) is 0.68. That means that our model made good predictions 18% percent points more often than random model (which has 0.5 AUC). Plus, p value for Hosmer Lemeshow test is bigger than 0.05, which means we cannot reject null hypothesis of equal frequency of expected and observed target variable categories. That means that our model correctly reproduced these frequencies.

Comparison with model without PPLFAIR variable



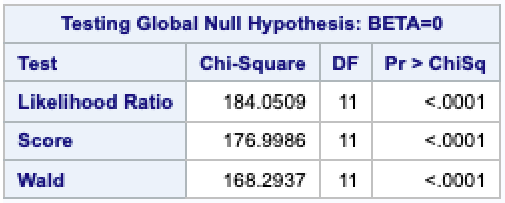
Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
17.4137	11	0.0962



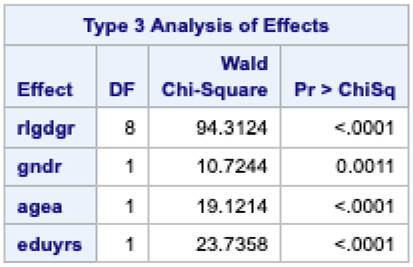
Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	3166.609	3004.558
SC	3177.683	3076.544
-2 Log L	3162.609	2978.558

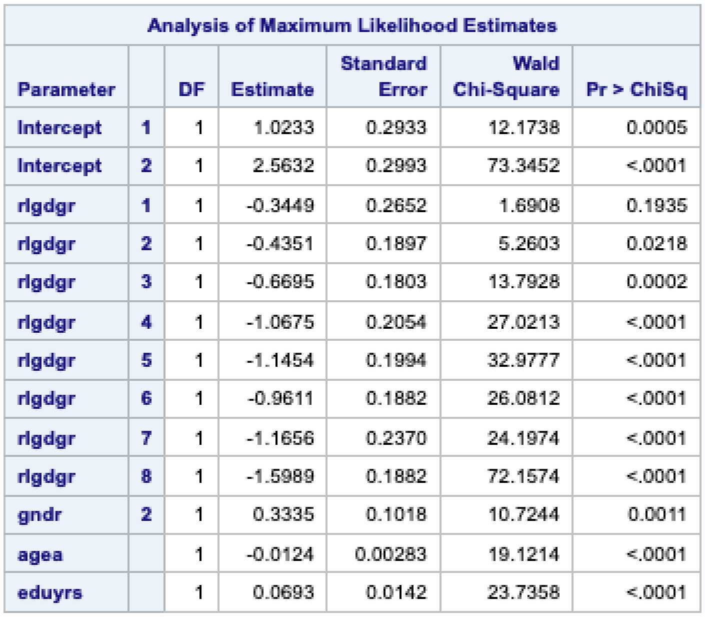


Testing Global Null Hypothesis: BETA=0

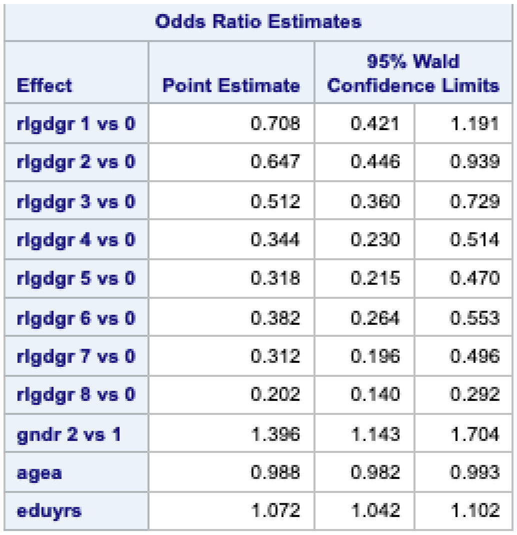
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	184.0509	11	<.0001
Score	176.9986	11	<.0001
Wald	168.2937	11	<.0001



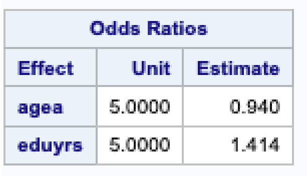
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
rlgdgr	8	94.3124	<.0001
gndr	1	10.7244	0.0011
agea	1	19.1214	<.0001
eduyrs	1	23.7358	<.0001



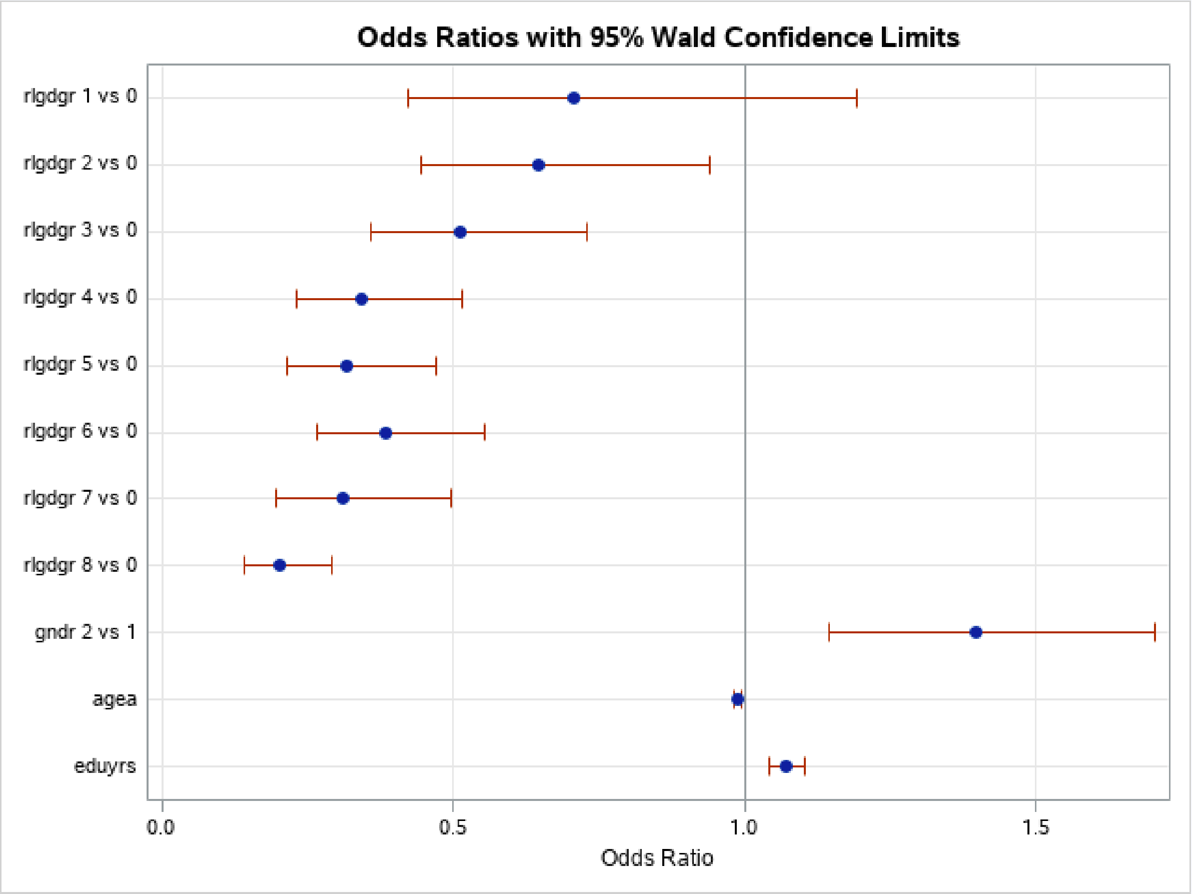
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	1.0233	0.2933	12.1738	0.0005
Intercept	2	1	2.5632	0.2993	73.3452	<.0001
rlgdgr	1	1	-0.3449	0.2652	1.6908	0.1935
rlgdgr	2	1	-0.4351	0.1897	5.2603	0.0218
rlgdgr	3	1	-0.6695	0.1803	13.7928	0.0002
rlgdgr	4	1	-1.0675	0.2054	27.0213	<.0001
rlgdgr	5	1	-1.1454	0.1994	32.9777	<.0001
rlgdgr	6	1	-0.9611	0.1882	26.0812	<.0001
rlgdgr	7	1	-1.1656	0.2370	24.1974	<.0001
rlgdgr	8	1	-1.5989	0.1882	72.1574	<.0001
gndr	2	1	0.3335	0.1018	10.7244	0.0011
agea		1	-0.0124	0.00283	19.1214	<.0001
eduyrs		1	0.0693	0.0142	23.7358	<.0001

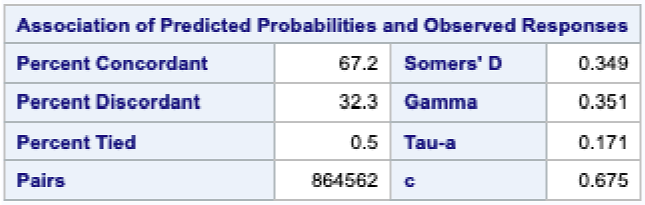


Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
rlgdgr 1 vs 0	0.708	0.421	1.191
rlgdgr 2 vs 0	0.647	0.446	0.939
rlgdgr 3 vs 0	0.512	0.360	0.729
rlgdgr 4 vs 0	0.344	0.230	0.514
rlgdgr 5 vs 0	0.318	0.215	0.470
rlgdgr 6 vs 0	0.382	0.264	0.553
rlgdgr 7 vs 0	0.312	0.196	0.496
rlgdgr 8 vs 0	0.202	0.140	0.292
gndr 2 vs 1	1.396	1.143	1.704
agea	0.988	0.982	0.993
eduyrs	1.072	1.042	1.102

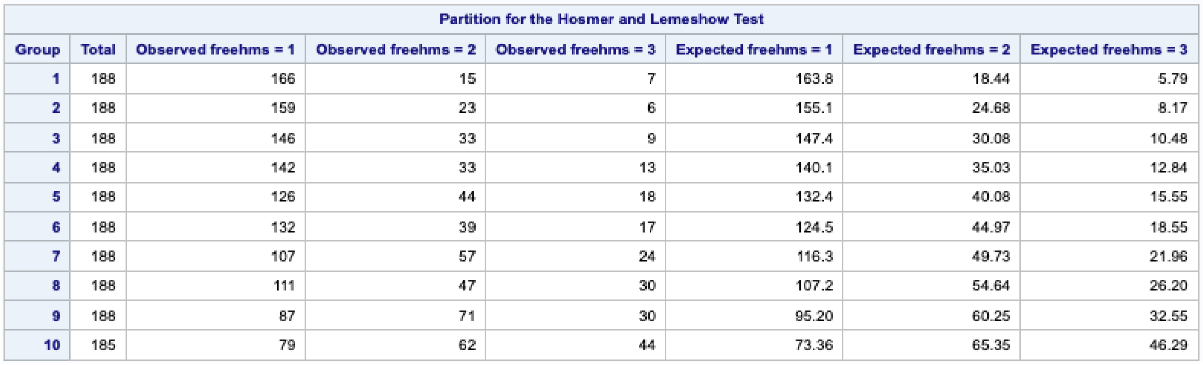


Odds Ratios		
Effect	Unit	Estimate
agea	5.0000	0.940
eduyrs	5.0000	1.414



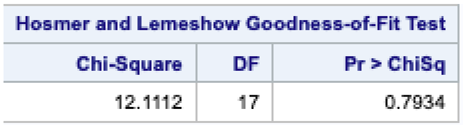


Association of Predicted Probabilities and Observed Responses			
Percent Concordant	67.2	Somers' D	0.349
Percent Discordant	32.3	Gamma	0.351
Percent Tied	0.5	Tau-a	0.171
Pairs	864562	c	0.675



Partition for the Hosmer and Lemeshow Test

Group	Total	Observed freehms = 1	Observed freehms = 2	Observed freehms = 3	Expected freehms = 1	Expected freehms = 2	Expected freehms = 3
1	188	166	15	7	163.8	18.44	5.79
2	188	159	23	6	155.1	24.68	8.17
3	188	146	33	9	147.4	30.08	10.48
4	188	142	33	13	140.1	35.03	12.84
5	188	126	44	18	132.4	40.08	15.55
6	188	132	39	17	124.5	44.97	18.55
7	188	107	57	24	116.3	49.73	21.96
8	188	111	47	30	107.2	54.64	26.20
9	188	87	71	30	95.20	60.25	32.55
10	185	79	62	44	73.36	65.35	46.29



Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
12.1112	17	0.7934

As we can see model without PPLFAR still satisfies proportional odds assumption, and explains target statistically significantly according to global beta test All variables explain target in signficantly. The same estimates are statistically significant, and odds ratios are very similar. Also percent of concordant pairs is similar as well as value of c test. C test value for model without PPLFAIR is only 0.5 smaller than for the model with PPLFAIR. Therefore, PPLFAIR value does not change much, and we consider it insignificant.

Conclusions

We can clearly see that the more religious the person is the lower is the probability for her to be in 1 or 2 FREEHMS category. Of course, this variable is also associated with years of education and age of the respondent. In general, we can say that less educated, older and more religious person is more likely to agree less or disagree with sentence “Gays and lesbians free to live life as they wish”.

When it comes to PPLFAIR vairable - as we saw it didn’t change much in our model. However, it is still worth to explore it more, as it is possible that if we had proper frequencies of all initial categories of these variable, it would start to be more meaningful. The most important category of this variable, the 4th one, was category 7 from initial distribution (didn’t merged with any other category). We cannot deny though, that combined categories 0, 1, 2, 3 and 9 with 10 contained some interesting patterns, that were lost after dimensionality reduction.

But going straight to the point - can we say, that our model supports our hypothesis that The religiosity of a person stimulates his/her negative attitude towards gays and lesbians living life as they wish? We need to bear in mind a crucial fact: due to underrepresentation of categories 4 and 5 of our target (disagree and strongly disagree) **we combined them with category 3 - neither agree nor disagree.**

Therefore, we cannot deny that it is still possible that more religious people **tend to be more neutral towards freedom of homosexual people instead of being more negative.** Therefore, we conclude that these study does not confirm our hypothesis. It is not denying it, though. We showed that the religiosity of a person inhibits his/her positive attitude towards gays and lesbians living life as they wish. We cannot say this is the same as stimulation of a negative opinion, but as we saw, there is definitely an association that is worth further study.