Exploratory Data Analysis

- StudentID: 21700086

- Name:Dong Hoon Kim

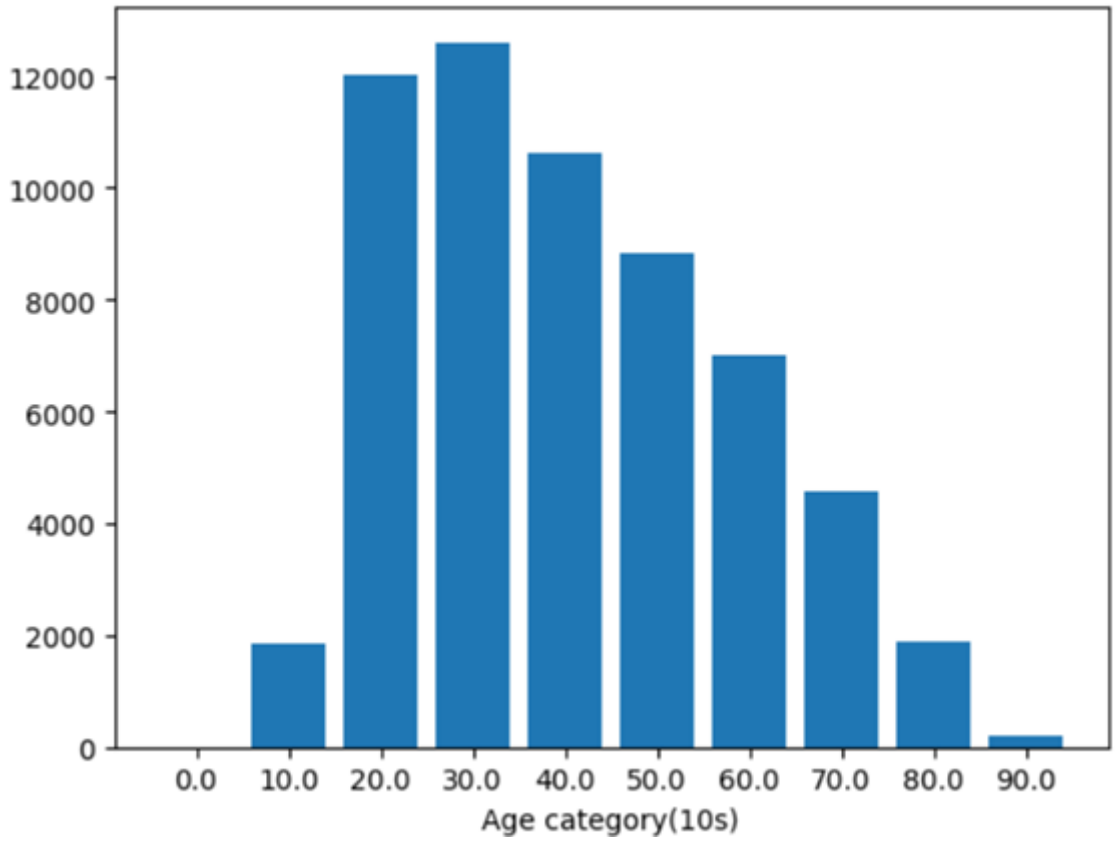
- 1st Major:Psychology

- 2nd Major:Data Science

Analyze the general factors related with 'work quality'

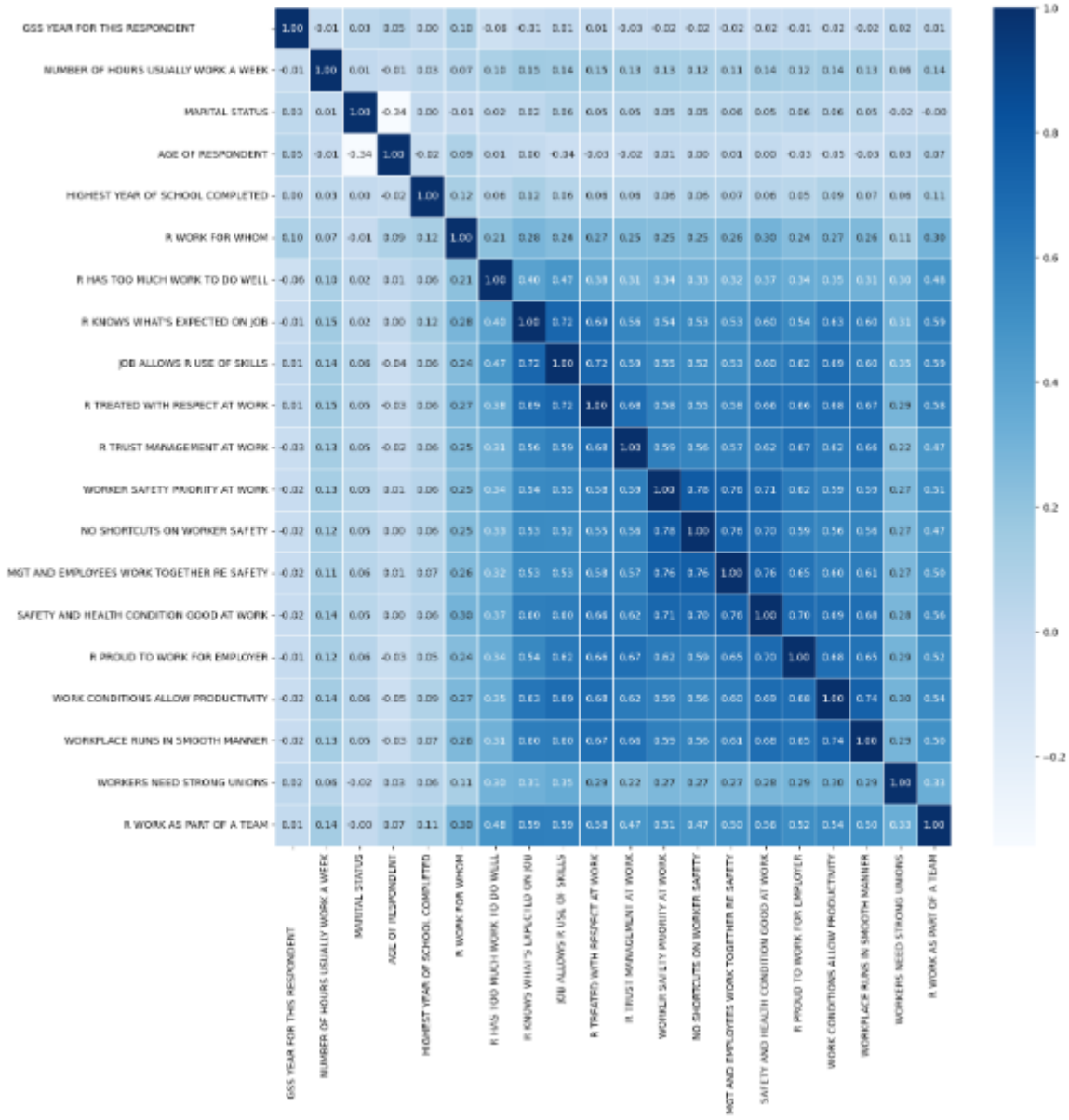
1. Data overview

Raw data consist of 11232 columns, 59599 rows
Data types was float64, and object



Age of people consisting the data skewed to right.
Most are 30s and 20s, 40s follows

2.1 'The Quality of Worklife Topical Module'



Considering 'GENERAL SOCIAL SURVEYS, 1972-2018 CUMULATIVE CODEBOOK' written in 'University of Chicago',

From the module list, I am able to pick The 'The Quality of Worklife Topical Module'

which consist of questions asking about overall condition of workplace that resphondent works at.

I conducted correlation compairison to several questions from module The codebook indicated the questions are conducted only at 2002,2006,2010,2014 year.

From ‘R WORK FOR WHOM’ to ‘R WORK AS PART OF A TEAM’ I picted these questions because the answer format is equal to these questions.

And I add some general variabls like ‘GSS YEAR FOR THIS RESPONDENT ’, ‘NUMBER OF HOURS USUALLY WORK A WEEK’,’ HIGHEST YEAR OF SCHOOL COMPLETED’, ‘MARITAL STATUS’, ‘AGE OF RESPONDENT’

The codebook indicated the questions are conducted only at 2002,2006,2010,2014 year The filtered dataframe

features is like below

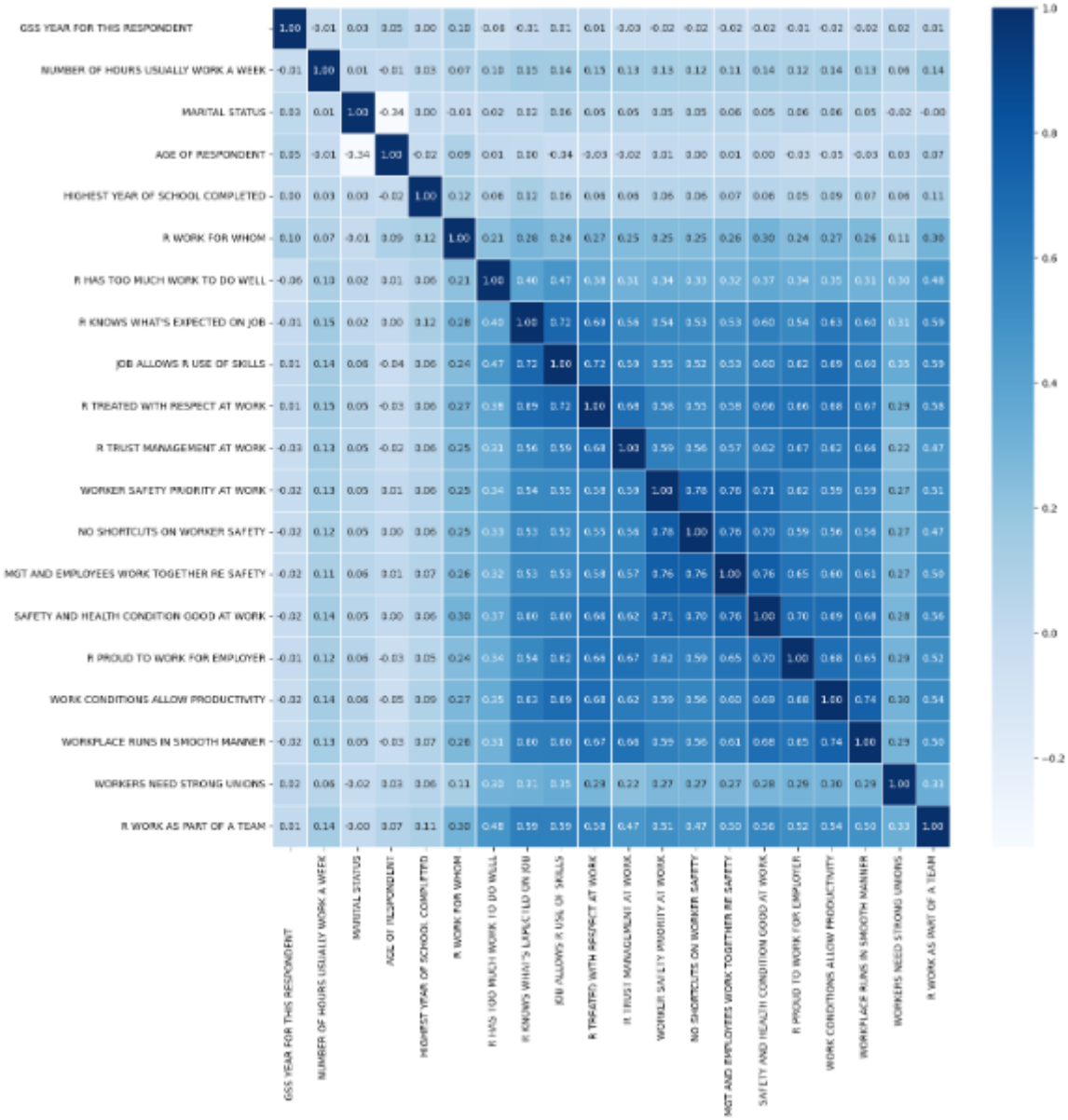
	GSS YEAR FOR THIS RESPONDENT	NUMBER OF HOURS USUALLY WORK A WEEK	MARITAL STATUS	AGE OF RESPONDENT	HIGHEST YEAR OF SCHOOL COMPLETED	R WORK FOR WHOM	Quality of work
40934	2002.0	-1.0	3.0	25.0	14.0	1.0	3.416667
40935	2002.0	-1.0	1.0	43.0	16.0	1.0	3.333333
40936	2002.0	-1.0	4.0	30.0	13.0	1.0	3.083333
40937	2002.0	-1.0	3.0	55.0	2.0	3.0	2.583333
40938	2002.0	-1.0	3.0	37.0	7.0	1.0	2.583333
...	...	...	...	...	...	...	...
59595	2014.0	-1.0	2.0	89.0	14.0	NaN	0.000000
59596	2014.0	-1.0	3.0	56.0	12.0	1.0	2.818182
59597	2014.0	-1.0	5.0	24.0	14.0	1.0	2.545455
59598	2014.0	-1.0	5.0	27.0	13.0	NaN	3.000000
59599	2014.0	-1.0	2.0	71.0	12.0	NaN	2.818182

16692 rows × 7 columns

2.2 Quality of work

Form the correlation comparison, and heat map, I can find the area consist of vadiabls correlation coefficient is over 0.5.
I made the new variable ‘Quality of work’ with mean of reversed score of variables in the area.
All question is asking about positivi asfect of work atmosphere.
And the ‘strongly agree’ is assigned to 1, ‘strongly disagree’ is assogned 4.

For convinient understanding, I decide to reverse the score. and excluded the null values.



3. Multivariate analysis

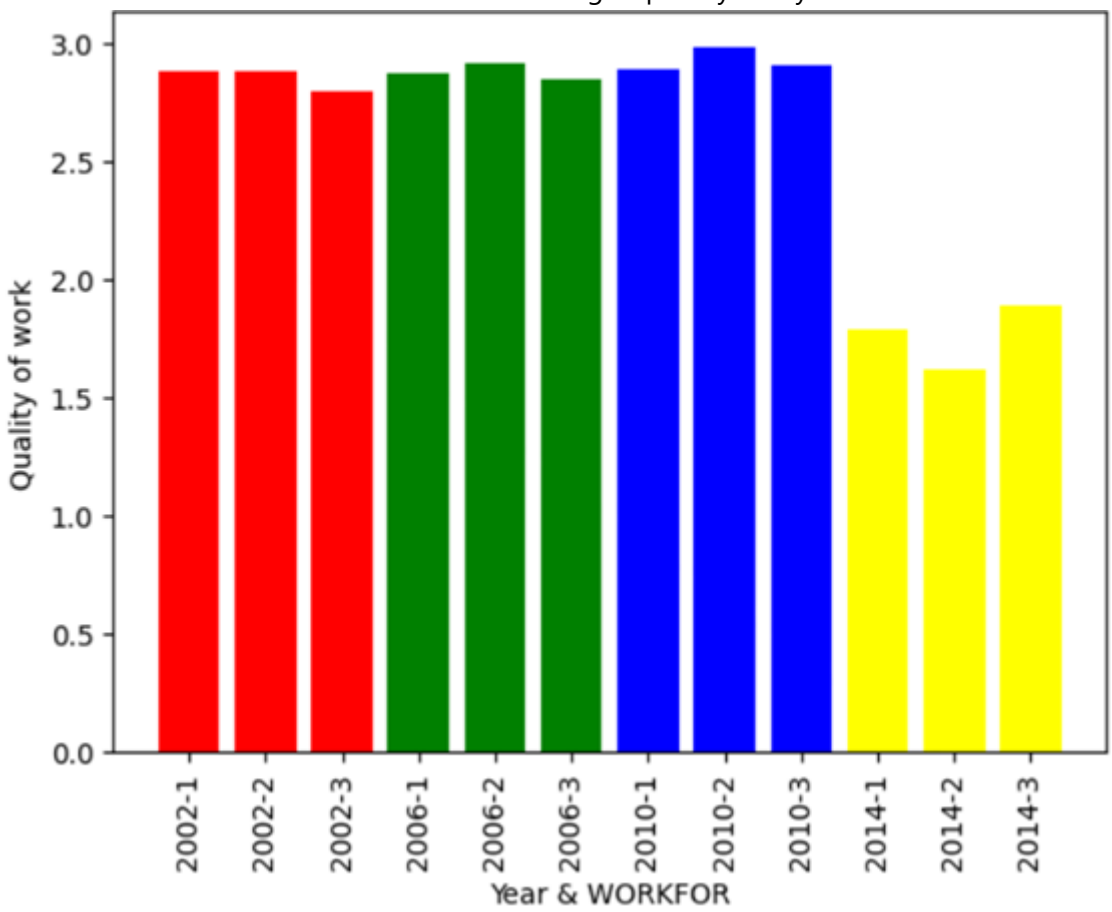
Analyzed the general values and the‘quality of work’

3.1 Work category and Quality of work

		Quality of work
GSS YEAR FOR THIS RESPONDENT	R WORK FOR WHOM	
2002.0	1.0	2.874486
	2.0	2.881792
	3.0	2.791326
	8.0	3.125000
	9.0	0.841270
2006.0	1.0	2.871529
	2.0	2.913030
	3.0	2.845866
	8.0	2.353175
	9.0	2.750000
2010.0	1.0	2.887701
	2.0	2.980556
	3.0	2.906541
	8.0	2.852834
	9.0	0.548276
2014.0	1.0	1.785427
	2.0	1.613992
	3.0	1.888889
	8.0	2.216783
	9.0	1.595455

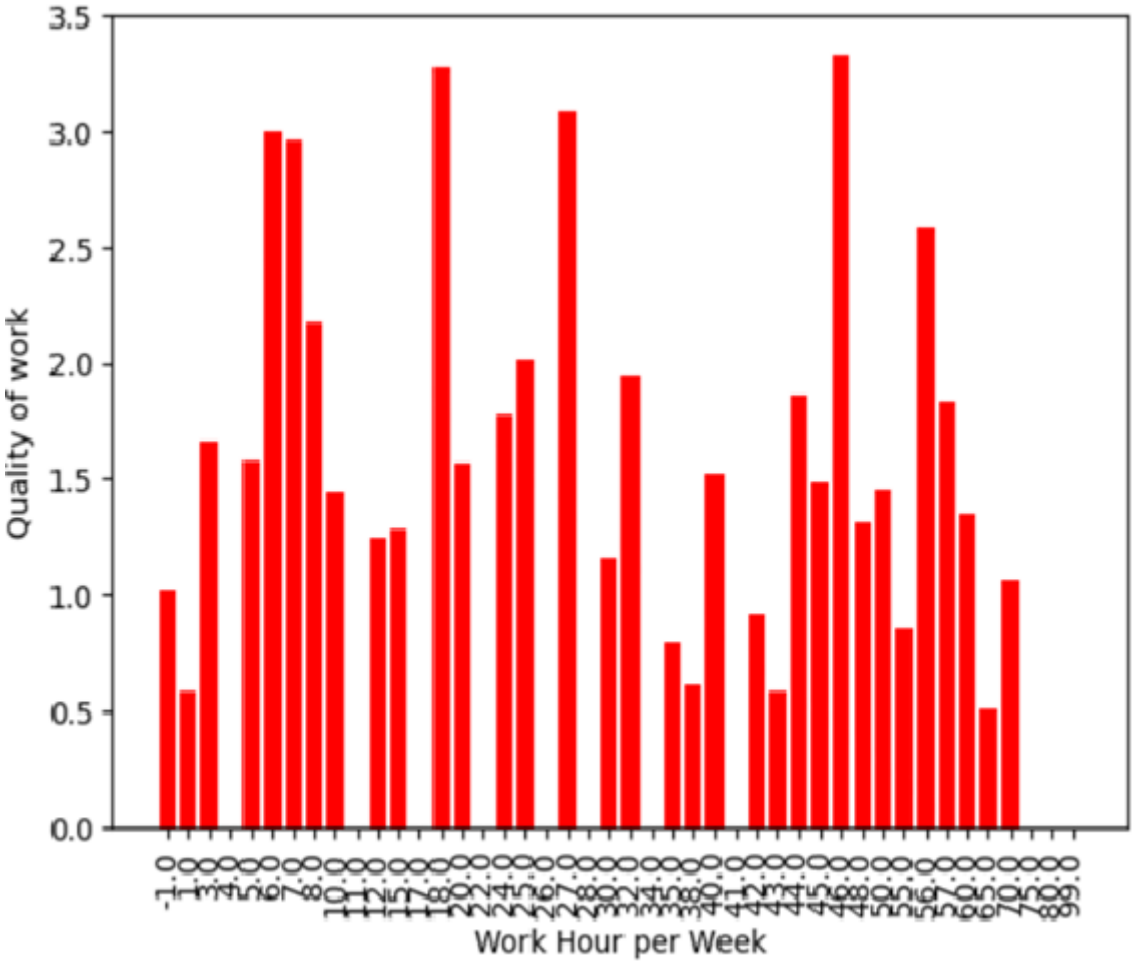
variable R WORK FOR WHOM is used.
(Index of R WORK FOR WHOM: 1-Private company 2-Non-profit organization 3-Government or government agency)

It seem some diffrence exist between the value grouped by GSS year.



However by checking visualized graph, there is no significant difference between work category and quality of work.
work quality of Non-profit organization worker reported slightly highest at 2002, 2006, 2010, but lowest at 2014

3.2 Work houre and Quality of work



NUMBER OF HOURS USUALLY WORK A WEEK is used. partecrpant reported their work hours per week.
By visualizing the mean score of 'Quality of work' by and hours of work ber week, 18h-3.3, 27h,-3.1 46h-3.4 shows the highest score.

4. Suggestion

The optimal work hour per week for workers seems to be 46h.

Considering the decreased quality of work an non-profit workers, proper reward system should be implemented to non-profit organization.