Exploratory Data Analysis

- StudentID: 21900296

- Name: Yoobin Park

- 1st Major: Life Science (45)

- 2nd Major: AI convergence (33)

In this project, I tried to find out how respondents' environment and characteristics impact on their happiness. There was no strong correlation among the variables, but there was a tendency between happiness and job satisfaction and finance-related variables. Based on these findings, I suggest a project aiming to enhance the workplace satisfaction

1. Data overview

First, I picked 18 relevant variables from the original dataset since the original data was too large to handle with my computer source. To investigate the factors for the happiness, basic information, educational and economic environment of respondents and how respondents feel about his/her life and job were selected along with general happiness.

The followings are selected variables

- **Basic information of respondents**
  `YEAR` GSS YEAR FOR THIS RESPONDENT
  `ID` RESPONDENT ID NUMBER
  `SEX` RESPONDENTS SEX
  `AGE` AGE OF RESPONDENT
  `RACE` RACE OF RESPONDENT
  `WRKSTAT` LABOR FORCE STATUS
  `WRKSLF` R SELF-EMP OR WORKS FOR SOMEBODY
  `RINCOME` RESPONDENTS INCOME
  `REGION` REGION OF INTERVIEW (Assume this as a region where Respondents live)

- **Educational & Economical environment of respondents**
  `DEGREE` RS HIGHEST DEGREE
  `REALINC` FAMILY INCOME IN CONSTANT $
  `REALRINC` RS INCOME IN CONSTANT $
  `INCOME` TOTAL FAMILY INCOME

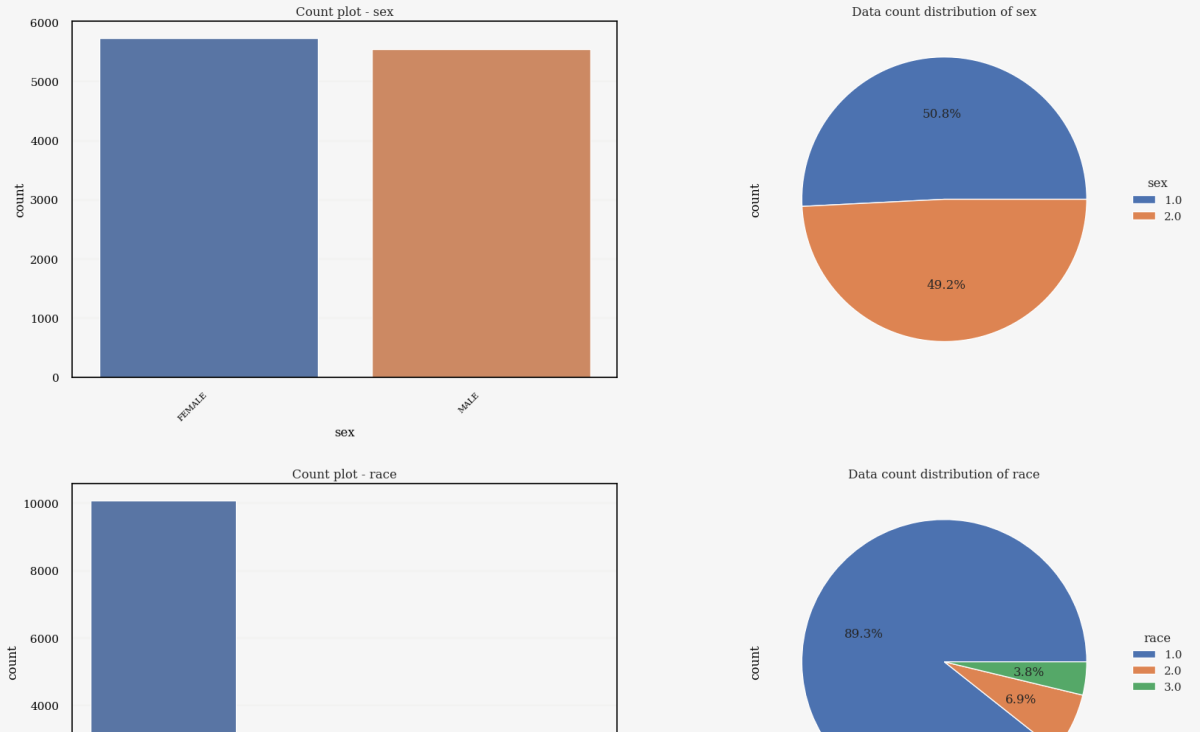
- **How respondents feel about his/her life and job**
  HAPPY GENERAL HAPPINESS
  SATJOB WORK SATISFACTION
  SATFIN SATISFACTION WITH FINANCIAL SITUATION

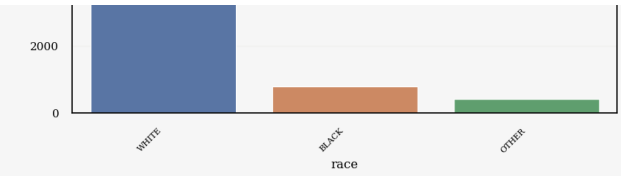
After selecting variables, the rows with `nan` values were dropped out. The rows with `IAP`(Invalid answer), `DK`(Don't know) were also deleted. To clarify the status of `HAPPY`, I encoded `VERY HAPPY`, `PRETTY HAPPY` as 1 (HAPPY) and `NOT TOO HAPPY` was encoded as 0 (NOT HAPPY). Finally, 11,268 rows and 18 columns were left after removing rows with any missing values.

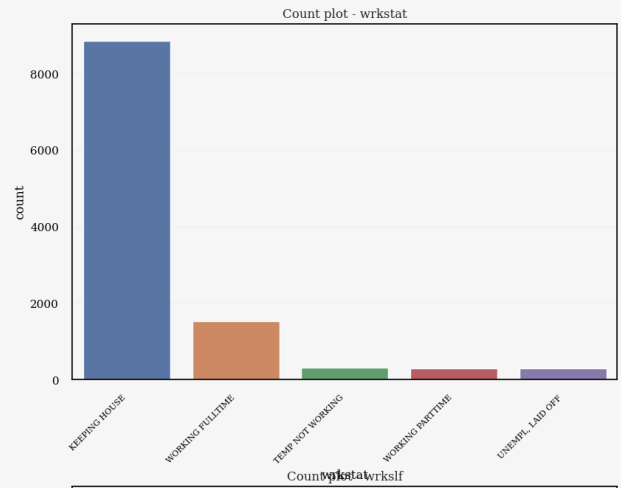
	dtypes	count	#unique	#missing	missing%	mean	std	min	25%	50%	75%	max
id	float64	11268	2950	0	0.000000	1118.529198	798.823224	1.000000	480.750000	997.000000	1536.000000	4506.000000
year	float64	11268	28	0	0.000000	1995.720714	11.047892	1974.000000	1987.000000	1996.000000	2006.000000	2014.000000
sex	float64	11268	2	0	0.000000	1.492013	0.499958	1.000000	1.000000	1.000000	2.000000	2.000000
age	float64	11268	67	0	0.000000	37.664448	11.548887	18.000000	28.000000	36.000000	45.000000	89.000000
race	float64	11268	3	0	0.000000	1.144302	0.445819	1.000000	1.000000	1.000000	1.000000	3.000000
wrkstat	float64	11268	5	0	0.000000	1.423944	1.116527	1.000000	1.000000	1.000000	1.000000	7.000000
wrkslf	float64	11268	2	0	0.000000	1.874689	0.331085	1.000000	2.000000	2.000000	2.000000	2.000000
rincome	float64	11268	12	0	0.000000	9.844693	3.078630	1.000000	9.000000	11.000000	12.000000	12.000000
region	float64	11268	9	0	0.000000	5.000532	2.562574	1.000000	3.000000	5.000000	7.000000	9.000000
degree	float64	11268	4	0	0.000000	2.088747	1.155568	1.000000	1.000000	2.000000	3.000000	4.000000
padeg	float64	11268	4	0	0.000000	1.801739	1.143051	1.000000	1.000000	1.000000	3.000000	4.000000
madeg	float64	11268	4	0	0.000000	1.547746	0.956649	1.000000	1.000000	1.000000	2.000000	4.000000
realinc	float64	11268	521	0	0.000000	42174.049283	33016.734866	236.500000	20355.000000	33016.000000	50027.000000	162607.000000
realrinc	float64	11268	550	0	0.000000	26788.413424	35663.597742	236.500000	10808.000000	19938.000000	31336.000000	480144.472857
income	float64	11268	12	0	0.000000	11.164714	1.783933	1.000000	11.000000	12.000000	12.000000	12.000000
happy	int64	11268	2	0	0.000000	0.923323	0.266091	0.000000	1.000000	1.000000	1.000000	1.000000
satfin	float64	11268	3	0	0.000000	1.943734	0.723414	1.000000	1.000000	2.000000	2.000000	3.000000
satjob	float64	11268	4	0	0.000000	1.668087	0.770601	1.000000	1.000000	2.000000	2.000000	4.000000

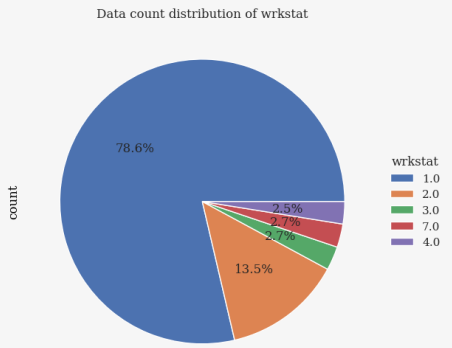
Table 1. Statistics summary of preprocessed data

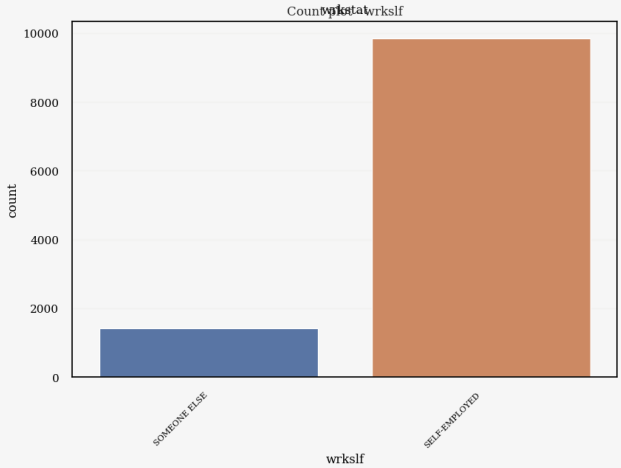
The selected data was collected from 1974 to 2014 and the average age of selected respondents is 37. All the data has 'float64' dtypes. The categorical data and continuous data distribution are visualized in Figure 1 and Figure 2, respectively.

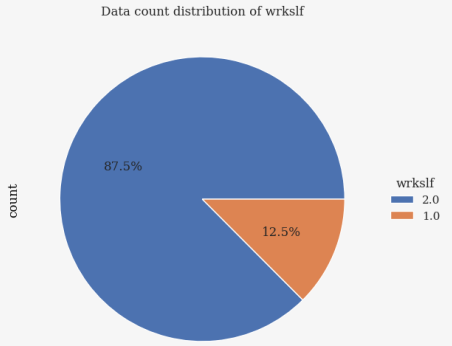


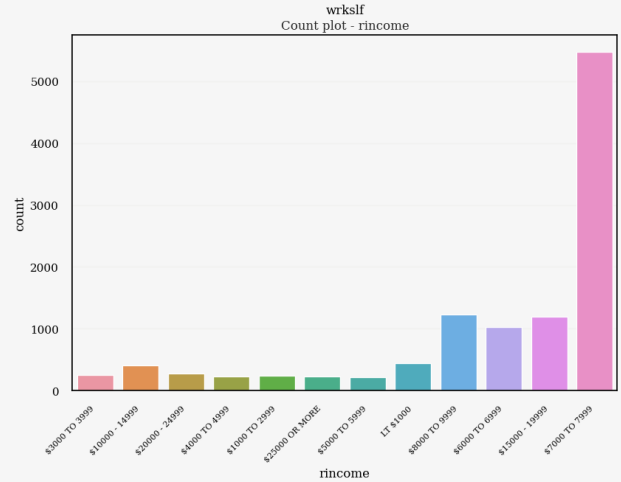


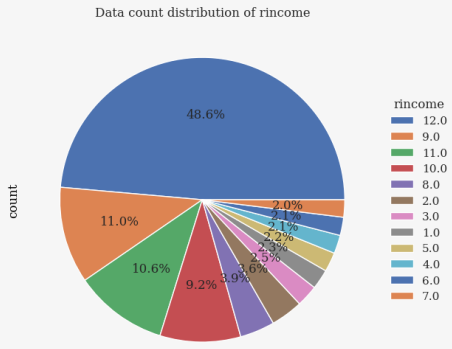


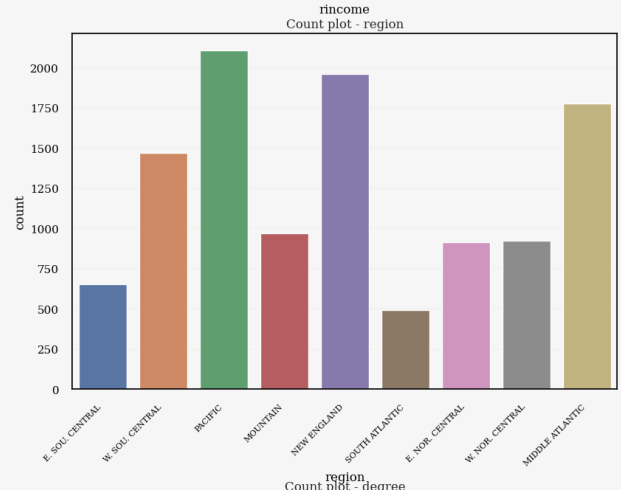


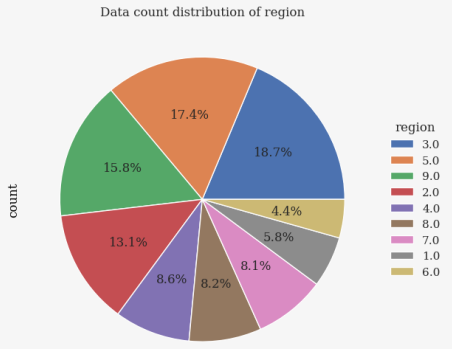




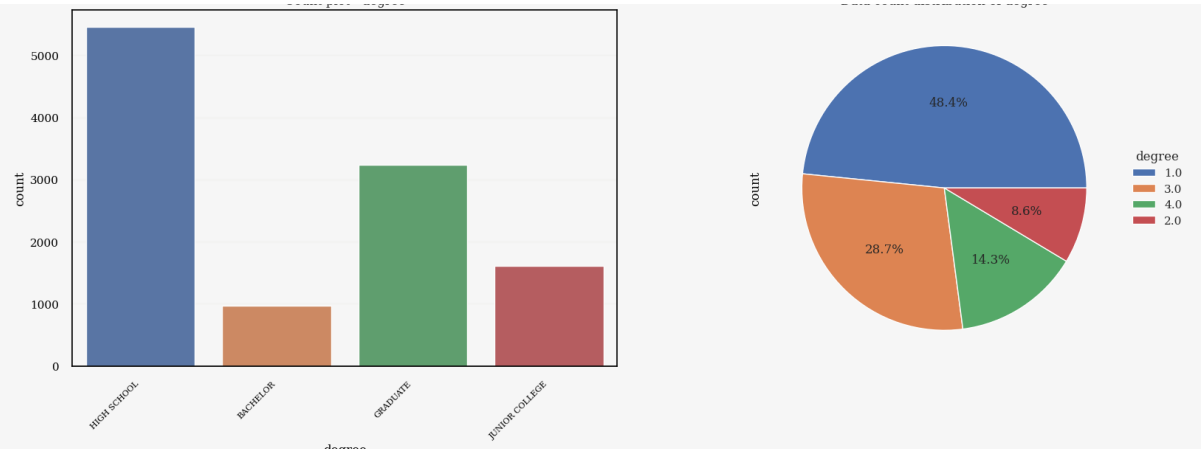






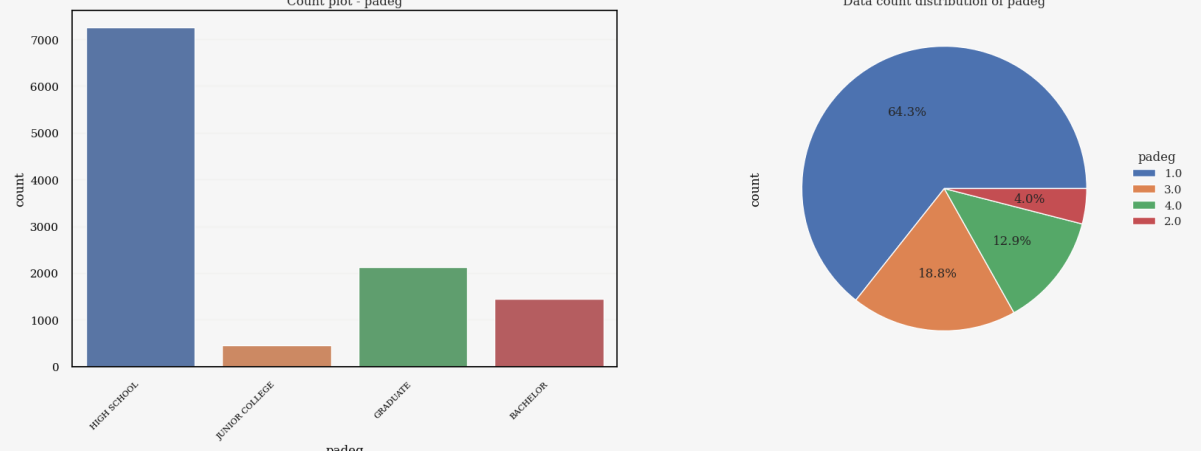


Count plot - degree Data count distribution of degree



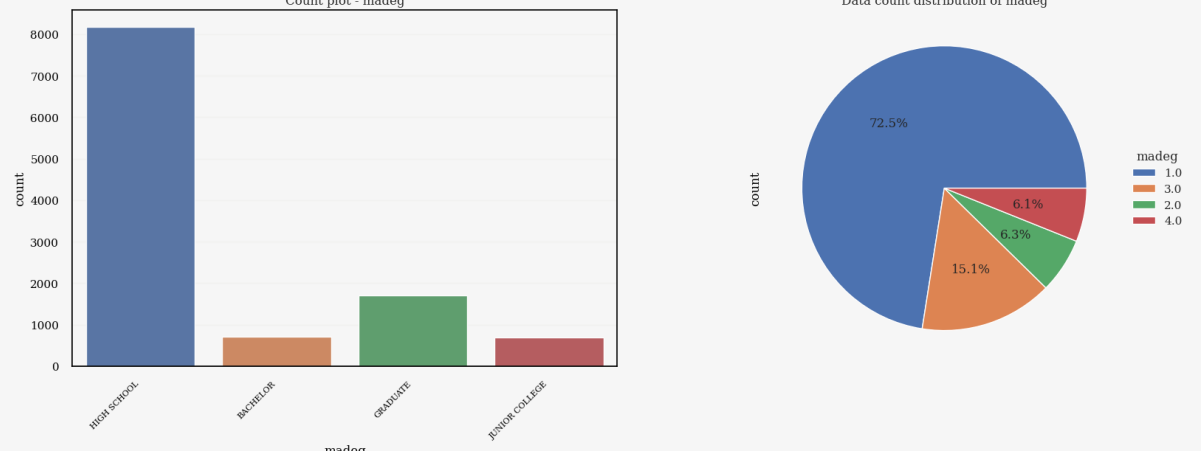
Count plot - padeg

Data count distribution of padeg



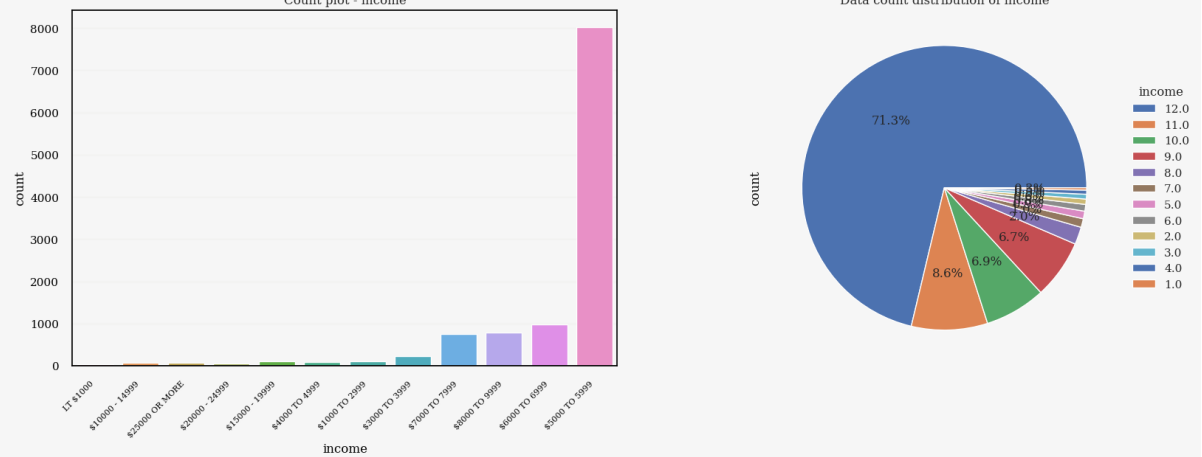
Count plot - madeg

Data count distribution of madeg



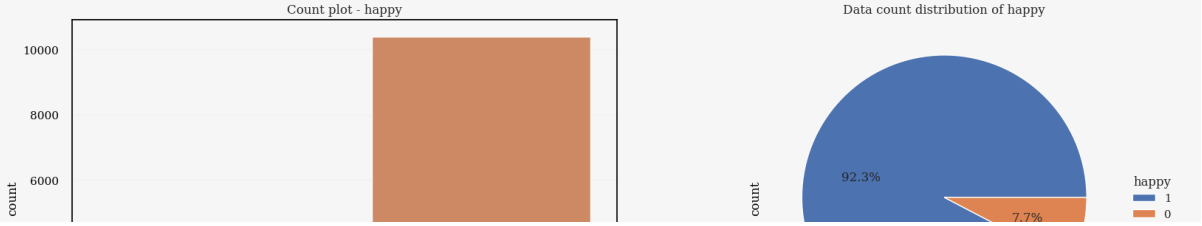
Count plot - income

Data count distribution of income



Count plot - happy

Data count distribution of happy



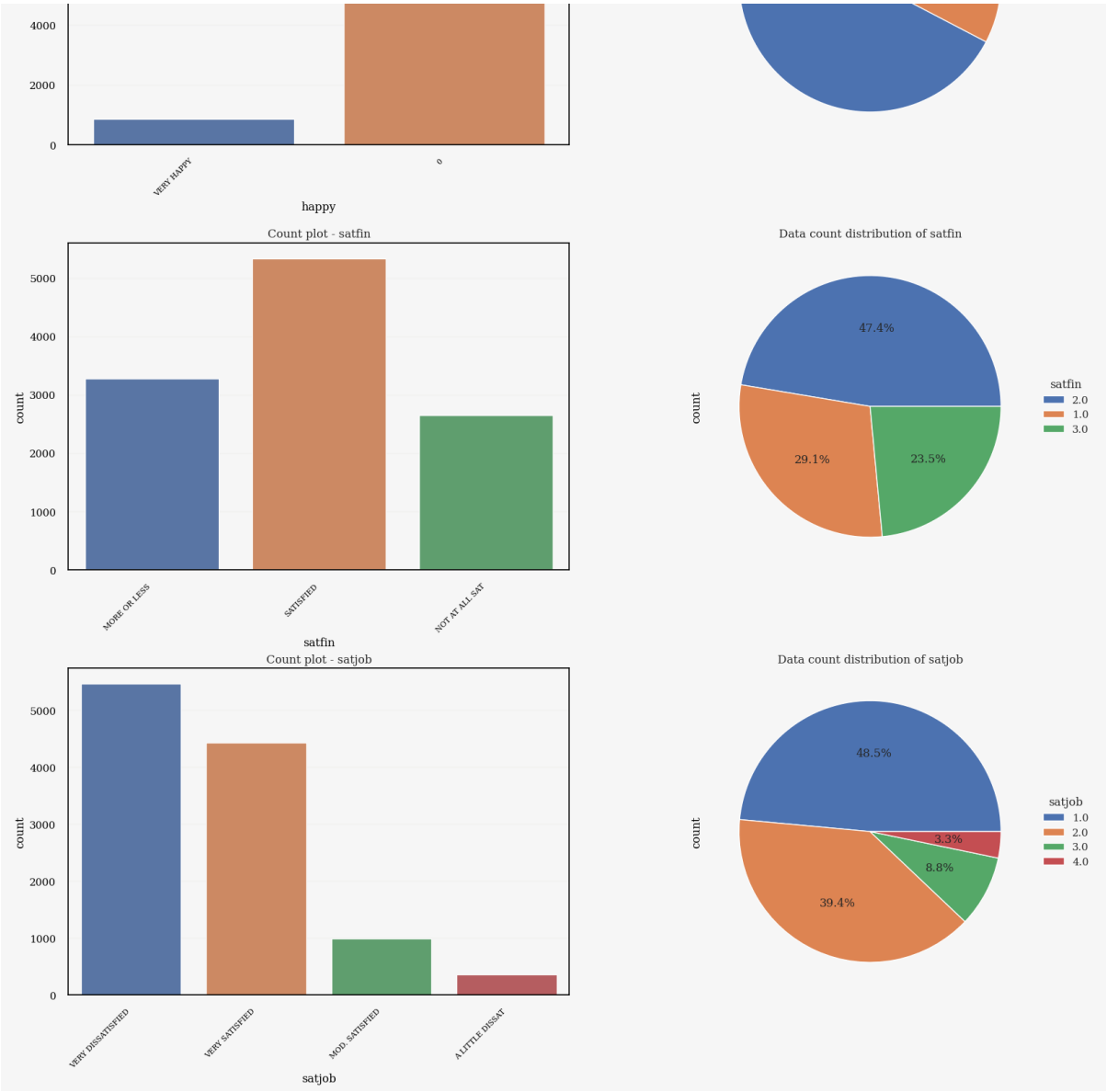


Figure 1. Countplot and pieplot of categorical variables

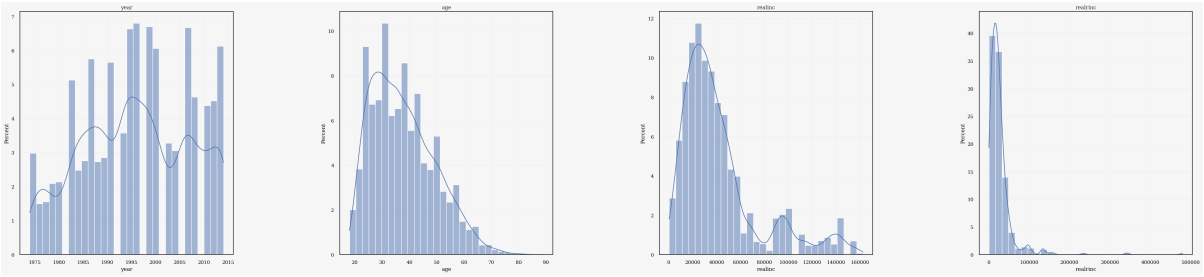


Figure 2. Histogram of continuous variables

2. Univariate analysis

To investigate how the happy status varies across different variables, I conducted a comprehensive exploratory data analysis (EDA). For continuous variables, I created histograms, box plots, bar plots, and violin plots to gain insights into how their distributions are different according to HAPPY status. For categorical variables, I used count plots and bar plots to compare the proportions of happy=1 within each category. Since the

continuous data didn't show any differences in distribution between `HAPPY` and `NOT HAPPY` except `realinc` (Fig. 3), I focused on categorical data shown in Figure 4.

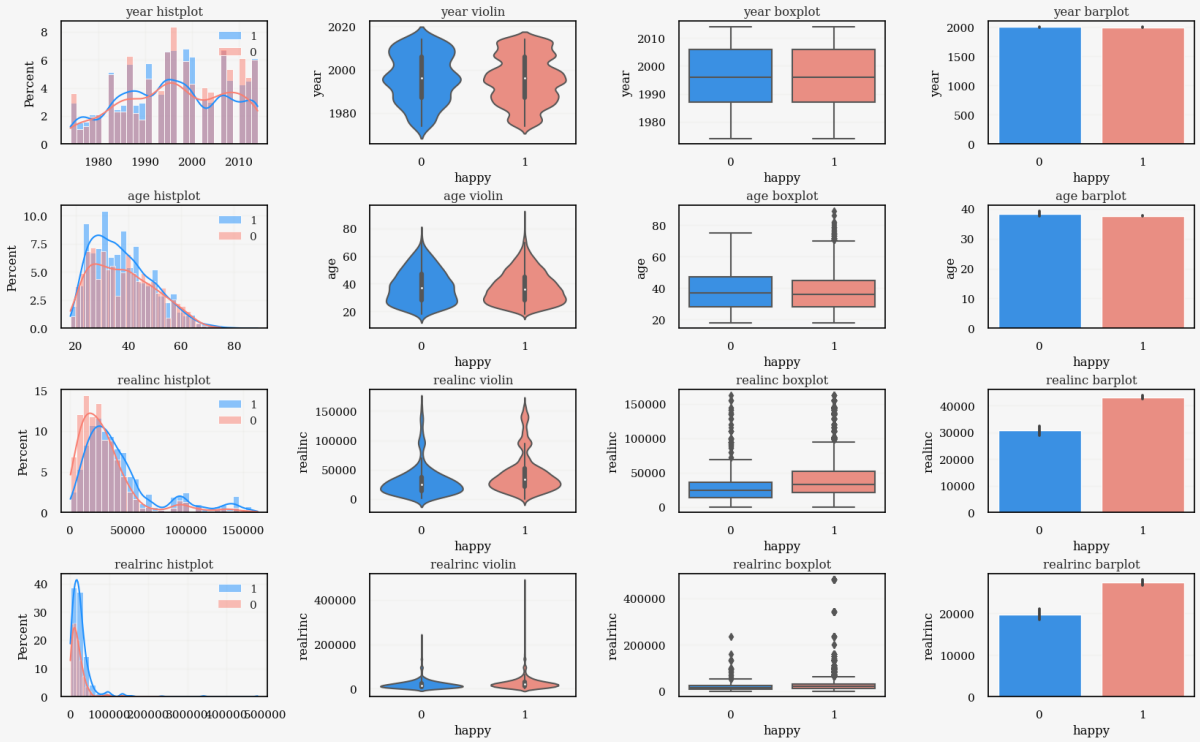
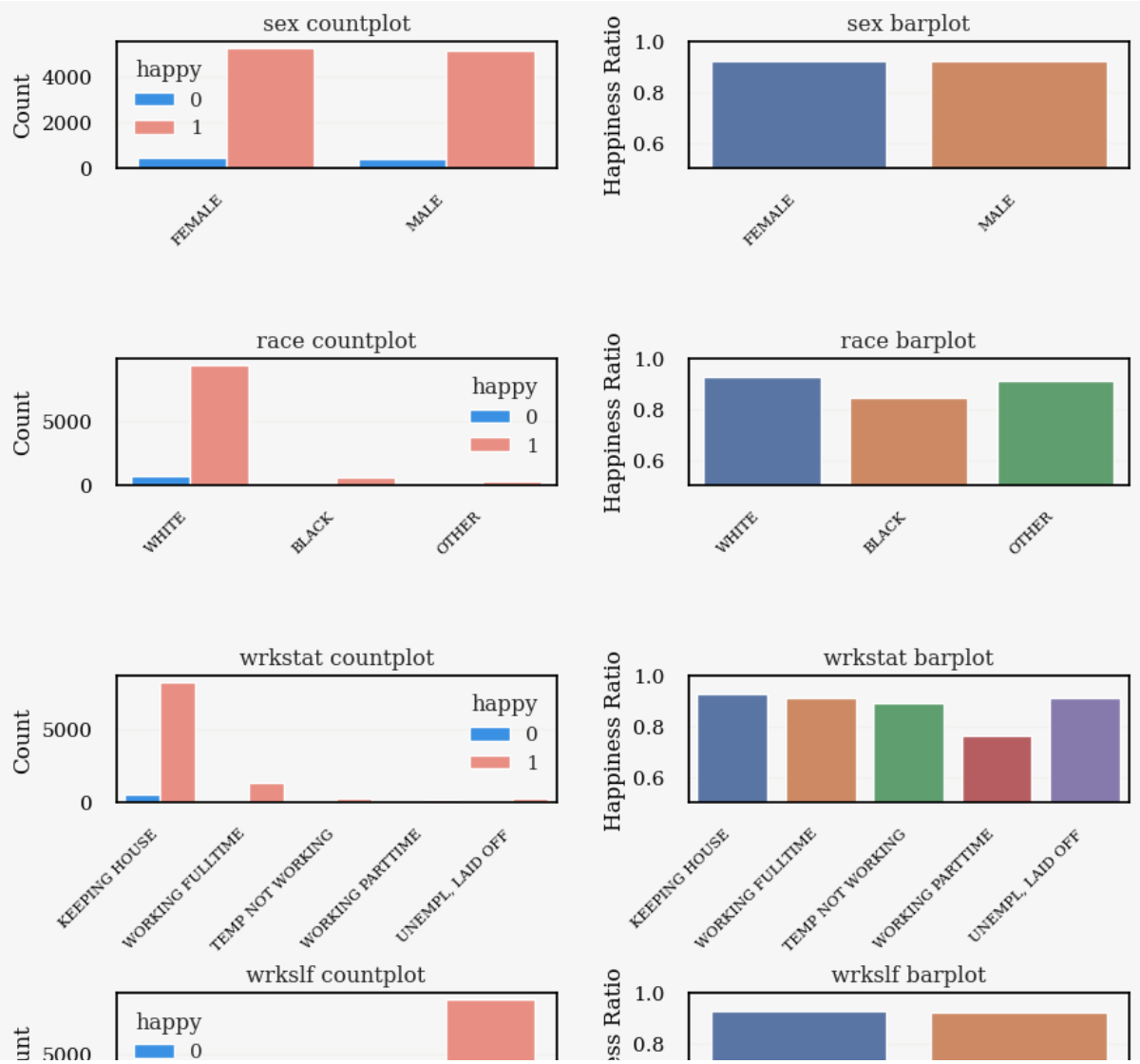
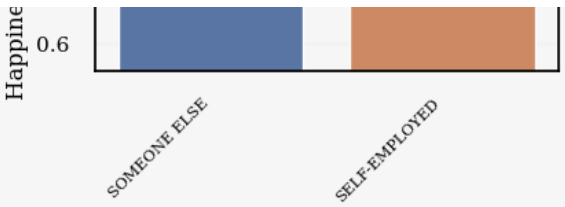
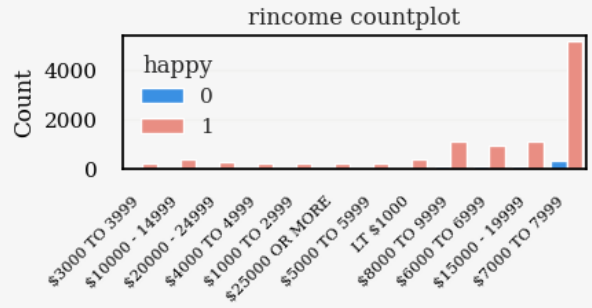


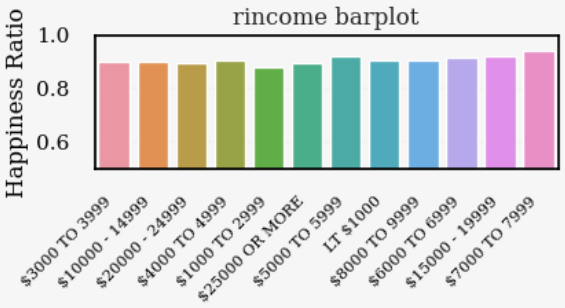
Figure 3. Histogram, Violin plot, boxplot and barplot of continuous variables

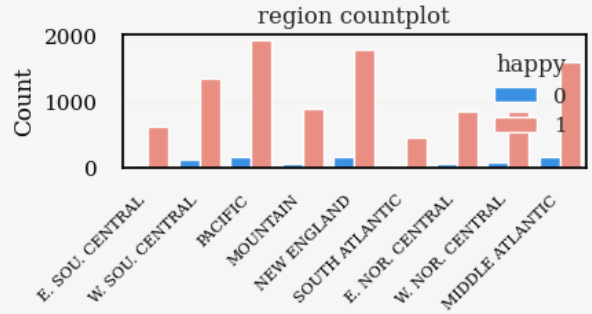


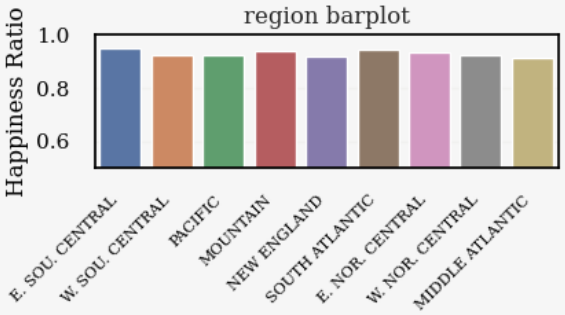

1
0
SOMEONE ELSE
SELF-EMPLOYED

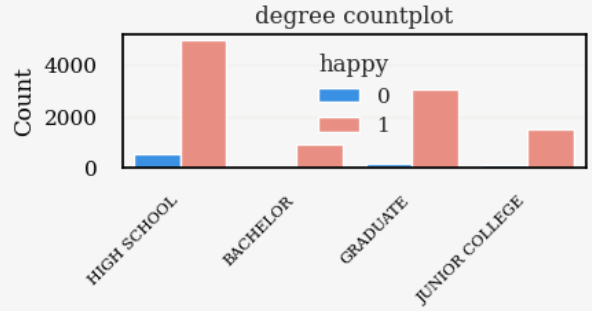

0.6
SOMEONE ELSE
SELF-EMPLOYED

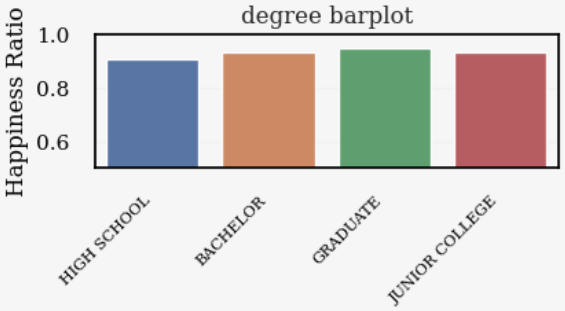

rincome countplot
Count
4000
2000
0
happy
0
1
$3000 TO 3999
$10000 - 14999
$20000 - 24999
$4000 TO 4999
$1000 TO 2999
$25000 OR MORE
$5000 TO 5999
LT $1000
$8000 TO 9999
$6000 TO 6999
$15000 - 19999
$7000 TO 7999

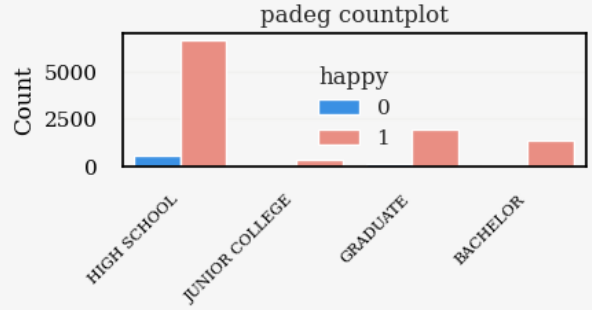

rincome barplot
Happiness Ratio
1.0
0.8
0.6
$3000 TO 3999
$10000 - 14999
$20000 - 24999
$4000 TO 4999
$1000 TO 2999
$25000 OR MORE
$5000 TO 5999
LT $1000
$8000 TO 9999
$6000 TO 6999
$15000 - 19999
$7000 TO 7999

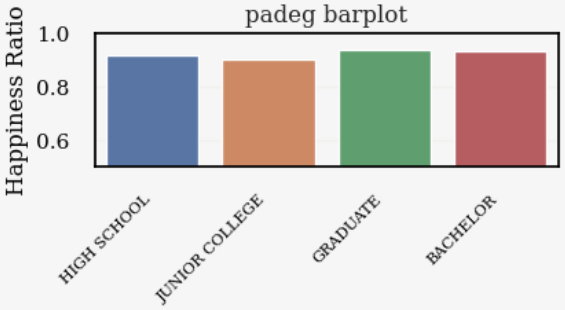

region countplot
Count
2000
1000
0
happy
0
1
E. SOU. CENTRAL
W. SOU. CENTRAL
PACIFIC
MOUNTAIN
NEW ENGLAND
SOUTH ATLANTIC
E. NOR. CENTRAL
W. NOR. CENTRAL
MIDDLE ATLANTIC

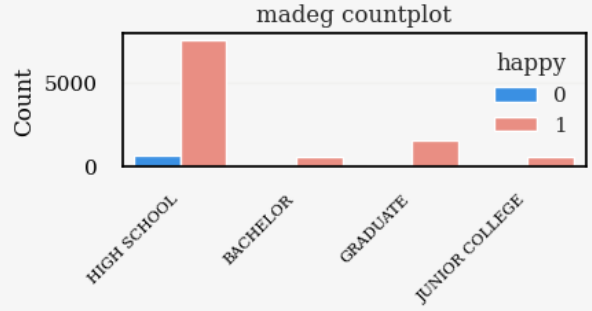

region barplot
Happiness Ratio
1.0
0.8
0.6
E. SOU. CENTRAL
W. SOU. CENTRAL
PACIFIC
MOUNTAIN
NEW ENGLAND
SOUTH ATLANTIC
E. NOR. CENTRAL
W. NOR. CENTRAL
MIDDLE ATLANTIC

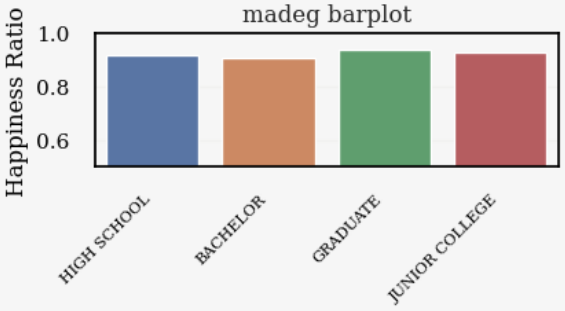

degree countplot
Count
4000
2000
0
happy
0
1
HIGH SCHOOL
BACHELOR
GRADUATE
JUNIOR COLLEGE

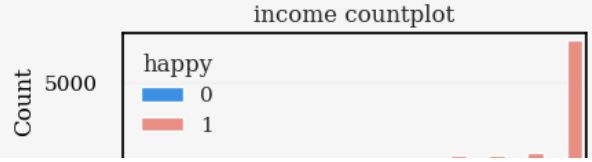

degree barplot
Happiness Ratio
1.0
0.8
0.6
HIGH SCHOOL
BACHELOR
GRADUATE
JUNIOR COLLEGE

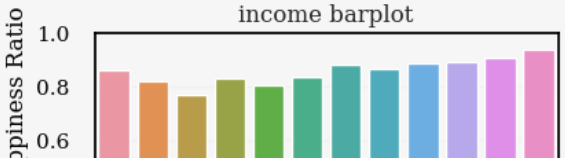

padeg countplot
Count
5000
2500
0
happy
0
1
HIGH SCHOOL
JUNIOR COLLEGE
GRADUATE
BACHELOR


padeg barplot
Happiness Ratio
1.0
0.8
0.6
HIGH SCHOOL
JUNIOR COLLEGE
GRADUATE
BACHELOR


madeg countplot
Count
5000
0
happy
0
1
HIGH SCHOOL
BACHELOR
GRADUATE
JUNIOR COLLEGE


madeg barplot
Happiness Ratio
1.0
0.8
0.6
HIGH SCHOOL
BACHELOR
GRADUATE
JUNIOR COLLEGE


income countplot
Count
5000
happy
0
1


income barplot
1.0
0.8
0.6

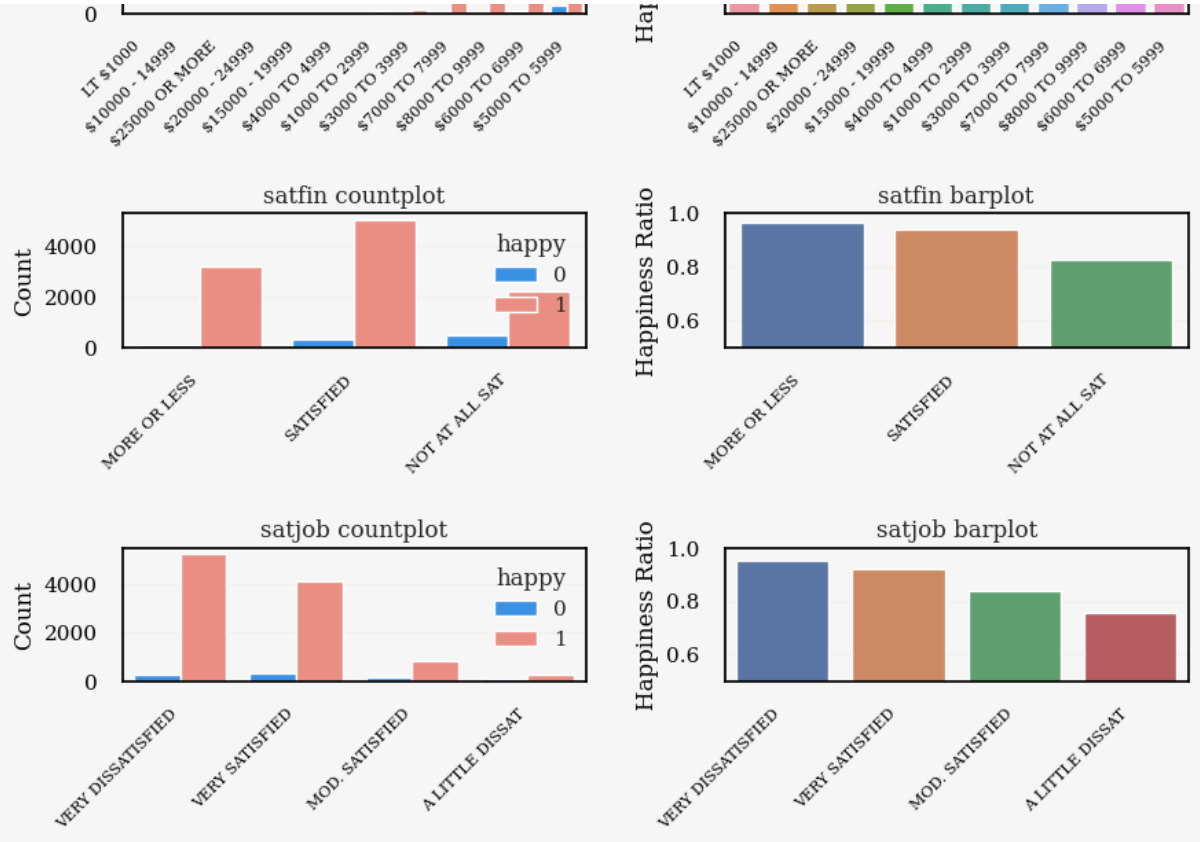


Figure 4. Countplot and Barplot of categorical variables

2.1 Relationship between 'wrkstat' and Happiness

Individuals working part-time ('wrkstat' = 'WORKING PARTTIME') tend to report lower happiness levels compared to those with other work statuses.

2.2 Relationship between 'income' and Happiness

Although there isn't a clear trend in the `rincome` variable, you can see the proportion of `HAPPY` decreases as income from total family `income` decreases.

2.3. Relationship between Financial Satisfaction ('satfin') and Happiness

People who are less satisfied with their financial situation tend to report lower happiness levels.

2.4. Relationship between Job Satisfaction ('satjob') and Happiness:

Lower job satisfaction ('satjob') seems to be associated with reduced happiness.

3. Multivariate analysis

In order to gain a deeper understanding of the relationships between multiple variables and their impact on happiness, a multivariate analysis was conducted. One of the key methods used in this analysis was the creation of a correlation heatmap. The heatmap visually represents the correlation coefficients between different variables, allowing us to identify patterns, dependencies, and insights.

3.1 Correlation between variables

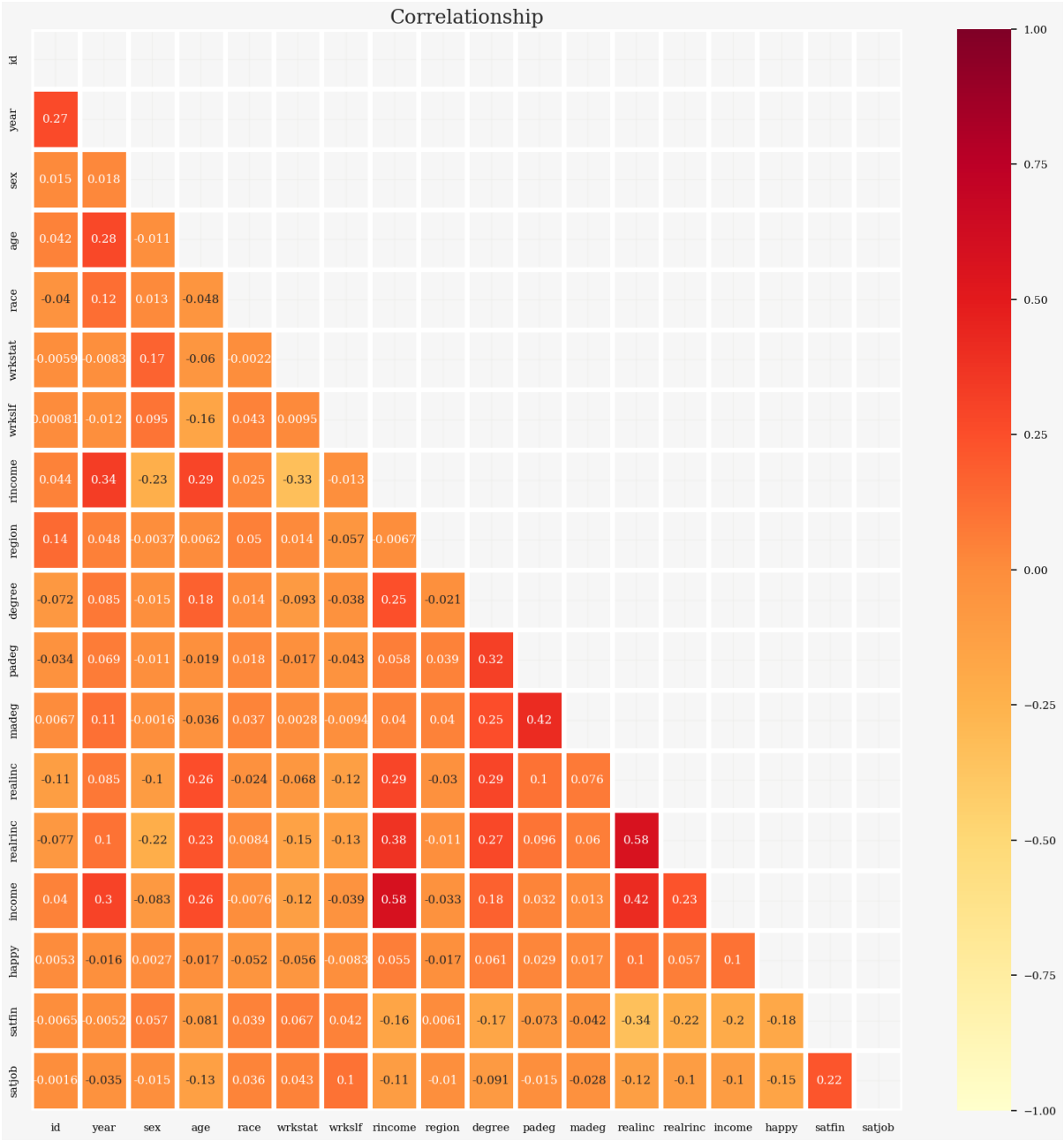


Figure 5. Correlation between variables

Unfortunately, I couldn't find any strong correlation(corr > 0.7) among any variables. Although there was no significant correlation betweeen any variables, I could see `wrkstat`, `satjob`, `realinc`, and `satfin` were relatively have high correlations to `Happy` compared to other variables as shown in Figure 5.

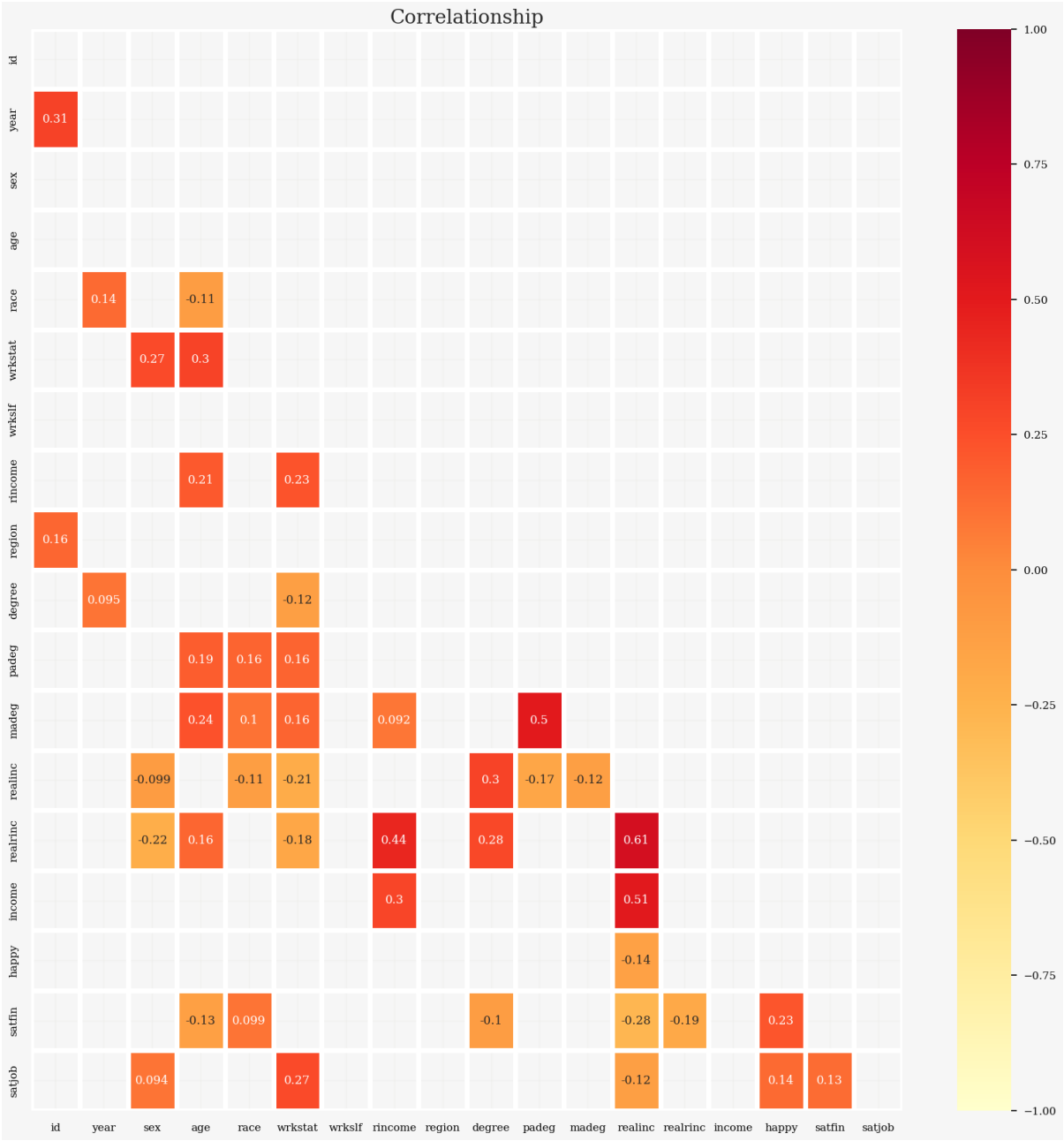


Figure 6. Correlation between variables that has bigger correlation than average correlations

4. Suggestion

Although I couldn't find any variables that have strong correlation with happiness, I could find there is a tendency between respondents' satisfaction to job and their happiness. Therefore, **enhancing workplace satisfaction for improved well-being.** This analysis has revealed that variables related to job satisfaction (`satjob`) and financial well-being (`realinc``, satfin``) partly play a role in influencing an individual's happiness (happy). However, there is no strong correlation among the variables, indicating a more complex relationship.

Therefore, we can provide following programs to enhance job satisfaction to increase happiness of individuals.

The main objective of this project will be developing a comprehensive program aimed at improving workplace satisfaction by addressing factors such as workload, work-life balance, recognition, and career growth.

1. **Financial Well-Being Workshops**:
   Organize workshops and educational programs focused on financial literacy and planning (realinc, satfin). Empower individuals with the knowledge and skills to manage their finances effectively, leading to reduced financial stress and increased financial satisfaction.

2. **Data-Driven Decision-Making**:
   Implement data collection and analytics to monitor the impact of the program on individuals' happiness. Continuously analyze the data to make data-driven decisions and improvements.