Exploratory Data Analysis

- StudentID: 22000282

- Name: Yeeun Park

- 1st Major: Life Science

- 2nd Major: AI Convergence

Proposed Project Summary: The project, titled "Analyzing Happiness Indices and Associated Factors," aims to extract valuable insights from the General Social Survey (GSS) dataset. It focuses on understanding the factors that influence happiness by conducting in-depth data analysis, segmentation, and predictive modeling. The project's expected outcomes include a comprehensive understanding of happiness drivers, data-backed recommendations for enhancing happiness, and guidance for future well-being initiatives.

1. Data overview

The GSS data is substantial, with a total size of 59,599 rows and 11,232 columns. It can be considered a large dataset. However, it's important to note that there are eight columns ('hompop_labels', 'babies_labels', 'size_labels', 'HUSHREL3', 'cohort_labels', 'wtss_labels', 'wtssnr_labels', 'wtssall_labels') where all rows have missing values (NaN).

Additionally, within the dataset, there are values like "Inapplicable (IAP)," "Don't know (DK)," and "No answer (NA)," which are all considered as missing data.

I selected a total of 15 variables from the dataset, focusing on those with a substantial amount of data. These variables are: 'year,' 'wrkstat_labels,' 'age,' 'degree_labels,' 'sex_labels,' 'happy,' 'realrinc,' 'postlife_labels,' 'natsoc,' 'marital_labels,' 'satjob,' 'sibs,' 'childs,' 'race_labels,' and 'relig_labels.'

In the original dataset, each variable had a corresponding '_labels' variable to represent the labels. Therefore, for these 15 variables I selected, I used the '_labels' variables only for nominal variables, excluding 'year.' The data, after extracting these variables, consists of 10,911 rows and 15 columns.

The table below provides the overview of data descriptions, data types, and data ranges for the 15 variables I selected:

Table 1. The Summary of Selected Values from Original Data

Variable	Description	Data Type	Data Range
year	GSS YEAR FOR THIS RESPONDENT	nominal	1984 - 2014
wrkstat_labels	LABOR FORCE STATUS	nominal	'WORKING FULLTIME', 'RETIRED', 'WORKING PARTTIME', 'KEEPING HOUSE', 'SCHOOL', 'UNEMPL, LAID OFF', 'TEMP NOT WORKING', 'OTHER'
age	AGE OF RESPONDENT	numerical	19 - 89
degree_labels	RS HIGHEST DEGREE	nominal	'JUNIOR COLLEGE', 'BACHELOR', 'HIGH SCHOOL', 'GRADUATE'
sex_labels	RESPONDENTS SEX	nominal	'FEMALE', 'MALE'
happy	GENERAL HAPPINESS	ordinal	1(Very happy), 2(Pretty happy), 3(Not too happy)
realrinc	RS INCOME IN CONSTANT $	numerical	236.50 - 480144.47
postlife_labels	BELIEF IN LIFE AFTER DEATH	nominal	'Yes', 'No'
natsoc	SOCIAL SECURITY	ordinal	1(Too little), 2(About right), 3(Too much)
marital_labels	MARITAL STATUS	nominal	'NEVER MARRIED', 'MARRIED', 'DIVORCED', 'WIDOWED', 'SEPARATED'

Variable	Description	Data Type	Data Range
satjob	WORK SATISFACTION	ordinal	1(Very satisfied), 2(Mod. satisfied), 3(A Little dissat), 4(Very Dissatisfied)
sibs	NUMBER OF BROTHERS AND SISTERS	numerical	1 - 37
childs	NUMBER OF CHILDREN	numerical	1 - 8(Eight or more)
race_labels	RACE OF RESPONDENT	nominal	'WHITE', 'BLACK', 'OTHER'
relig_labels	R'S RELIGIOUS PREFERENCE	nominal	'CATHOLIC', 'JEWISH', 'PROTESTANT', 'NONE', 'OTHER', 'CHRISTIAN', 'INTER-NONDENOMINATIONAL', 'ORTHODOX-CHRISTIAN', 'HINDUISM', 'BUDDHISM', 'MOSLEM/ISLAM', 'OTHER EASTERN', 'NATIVE AMERICAN'

2. Univariate analysis

2.1 'year'

Discription: GSS year for this respondent

Data Type: Nominal Data

Data Range: 1984 - 2014

Statistical Description:

- number of data: 10911
- unique values: 19
- top frequency value: 2006
- freq: 905
- dtype: object

2.2 'wrkstat_labels'

Discription: Labor force status

Data Type: Nominal Data

Data Range: 'WORKING FULLTIME', 'RETIRED', 'WORKING PARTTIME', 'KEEPING HOUSE', 'SCHOOL', 'UNEMPL, LAID OFF', 'TEMP NOT WORKING', 'OTHER'

Statistical Description:

- number of data: 10911
- unique values: 5
- top frequency value: WORKING FULLTIME
- freq: 8275
- dtype: object

2.3 'age'

Discription: Age of respondent

Data Type: Numerical Data

Data Range: 19 - 89

Statistical Description:

- number of data: 10911
- mean: 43.85
- std: 11.96
- min: 19.00
- 25%: 35.00

- 50%: 43.00
- 75%: 52.00
- max: 89.00
- dtype: float64

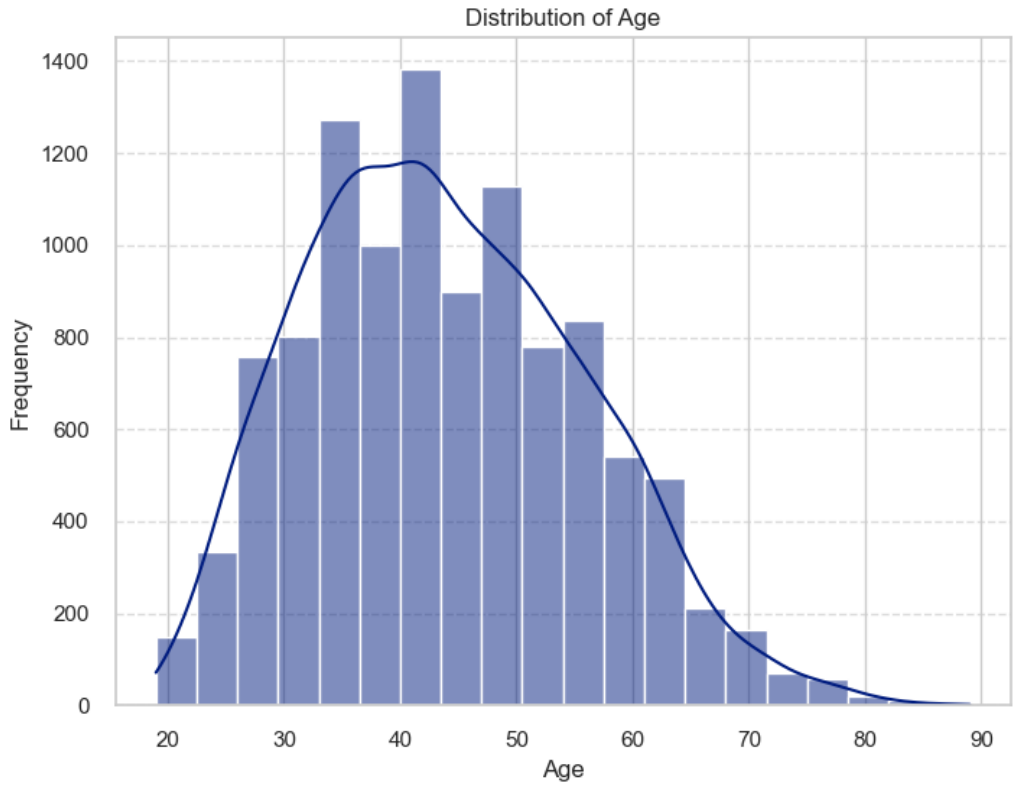

Figure 1. Distribution of Age

2.4 'degree_labels'

Discription: Respondent's highest degree

Data Type: Nominal Data

Data Range: 'JUNIOR COLLEGE', 'BACHELOR', 'HIGH SCHOOL', 'GRADUATE'

Statistical Description:

- number of data: 10911
- unique values: 5
- top frequency value: HIGH SCHOOL
- freq: 6002
- dtype: object

2.5 'sex_labels'

Discription: Respondent's sex

Data Type: Nominal Data

Data Range: 'FEMALE', 'MALE'

Statistical Description:

- number of data: 10911
- unique values: 2
- top frequency value: FEMALE
- freq: 5797
- dtype: object

2.6 'happy'

Discription: General happiness

Data Type: Ordinal Data

Data Range: 1(Very happy), 2(Pretty happy), 3(Not too happy)

Statistical Description:

- number of data: 10911
- mean: 1.78
- std: 0.62
- min: 1.00
- 25%: 1.00
- 50%: 2.00
- 75%: 2.00
- max: 3.00
- dtype: float64

2.7 'realrinc'

Discription: Respondent's income in constant $

Data Type: Numerical Data

Data Range: 236.50 - 480144.47

Statistical Description:

- number of data: 10911
- mean: 24196.23
- std: 35266.32
- min: 236.50
- 25%: 9000.00
- 50%: 17394.00
- 75%: 28156.50
- max: 480144.47
- dtype: float64

2.8 'postlife_labels'

Discription: Belief in life after death

Data Type: Nominal Data

Data Range: 'Yes', 'No'

Statistical Description:

- number of data: 10911
- unique values: 2
- top frequency value: YES
- freq: 8879
- dtype: object

2.9 'natsoc'

Discription: Social security

Data Type: Ordinal Data

Data Range: 1(Too little), 2(About right), 3(Too much)

Statistical Description:

- number of data: 10911
- mean: 1.46
- std: 0.62
- min: 1.00
- 25%: 1.00
- 50%: 1.00
- 75%: 2.00
- max: 3.00
- dtype: float64

2.10 'marital_labels'

Discription: Marital status

Data Type: Nominal Data

Data Range: 'NEVER MARRIED', 'MARRIED', 'DIVORCED', 'WIDOWED', 'SEPARATED'

Statistical Description:

- number of data: 10911
- unique values: 5
- top frequency value: MARRIED
- freq: 6910
- dtype: object

2.11 'satjob'

Discription: Work satisfaction

Data Type: Ordinal Data

Data Range: 1(Very satisfied), 2(Mod. satisfied), 3(A Little dissat), 4(Very Dissatisfied)

Statistical Description:

- number of data: 10911
- mean: 1.65
- std: 0.77
- min: 1.00
- 25%: 1.00
- 50%: 1.00
- 75%: 2.00
- max: 4.00
- dtype: float64

2.12 'sibs'

Discription: Number of brothers and sisters

Data Type: Numerical Data

Data Range: 1 - 37

Statistical Description:

- number of data: 10911
- mean: 4.01
- std: 3.06
- min: 1.00
- 25%: 2.00
- 50%: 3.00
- 75%: 5.00
- max: 37.00
- dtype: float64

2.13 'childs'

Discription: Number of children

Data Type: Numerical Data

Data Range: 1 - 8(Eight or more)

Statistical Description:

- number of data: 10911
- mean: 2.41
- std: 1.32
- min: 1.00
- 25%: 2.00
- 50%: 2.00
- 75%: 3.00
- max: 8.00

- dtype: float64

2.14 'race_labels'

Discription: Race of respondent

Data Type: Nominal Data

Data Range: 'WHITE', 'BLACK', 'OTHER'

Statistical Description:

- number of data: 10911
- unique values: 3
- top frequency value: WHITE
- freq: 8491
- dtype: object

2.15 'relig_labels'

Discription: Respondent's religious preference

Data Type: Nominal Data

Data Range: 'CATHOLIC', 'JEWISH', 'PROTESTANT', 'NONE', 'OTHER', 'CHRISTIAN', 'INTER-NONDENOMINATIONAL', 'ORTHODOX-CHRISTIAN', 'HINDUISM', 'BUDDHISM', 'MOSLEM/ISLAM', 'OTHER EASTERN', 'NATIVE AMERICAN'

Statistical Description:

- number of data: 10911
- unique values: 13
- top frequency value: PROTESTANT
- freq: 6356
- dtype: object

3. Multivariate analysis

3.1 Correlation

I first examined the correlations between variables. I transformed nominal data into a binary format by "One-Hot Encoding."

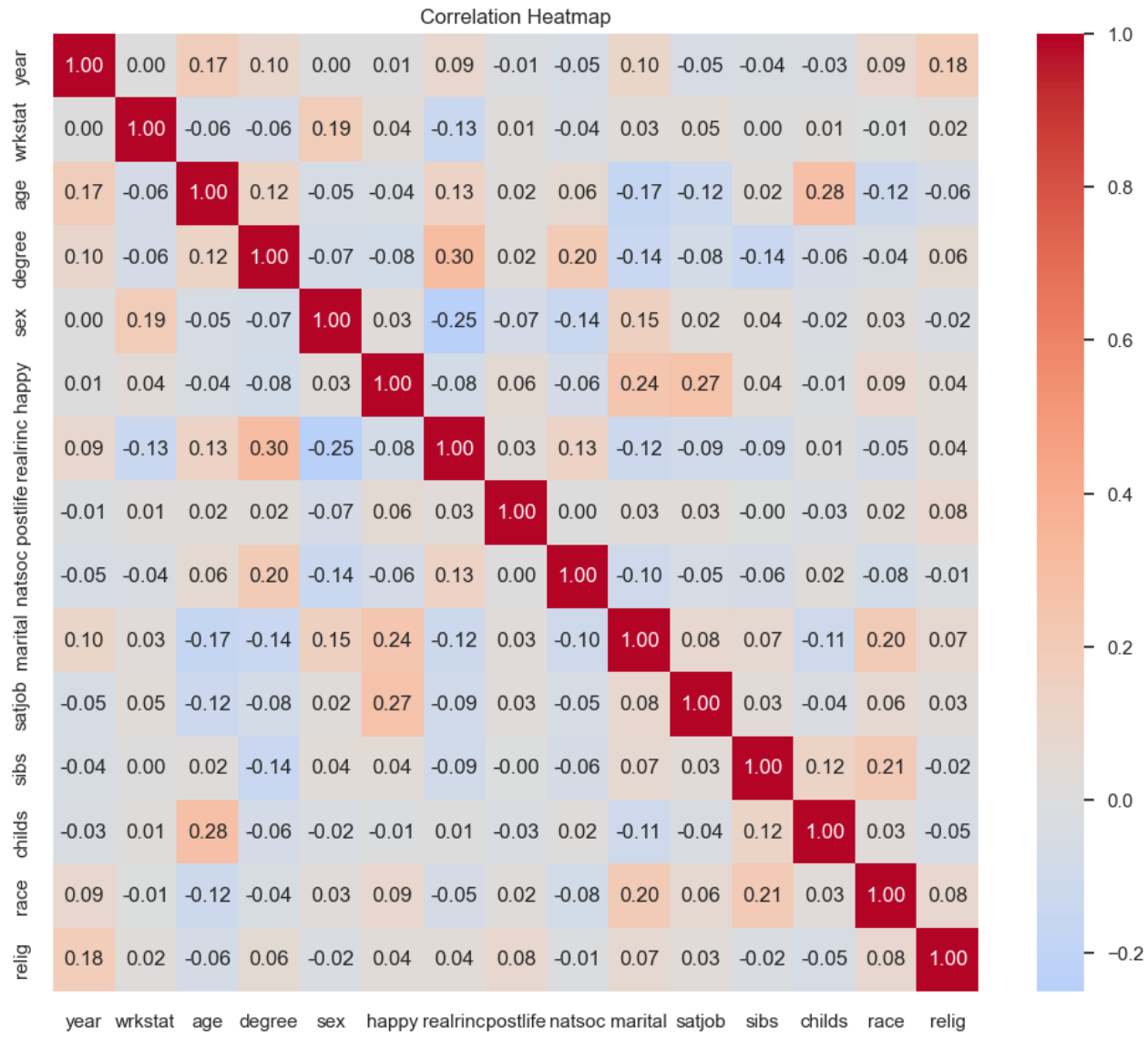

Figure 2. Correlation between Variables

However, I didn't find any strong correlations above 0.5. Among the relatively higher correlations, 'degree' and 'realrinc(income)' (0.30), 'age' and 'childs' (0.28) and 'happy' and 'satjob' (0.27) showed some noticeable associations than other variables.

3.2 Clustering

I tried to cluster the variables into three groups based on their characteristics and maps these clusters to different levels of happiness. To do so, I standarized the data, and performed K-Means clustering to group individuals into three clusters based on these features. I adjusted 'happy' values according to the cluster assignments and employed Principal Component Analysis (PCA) to visualize the clustered data in two dimensions.

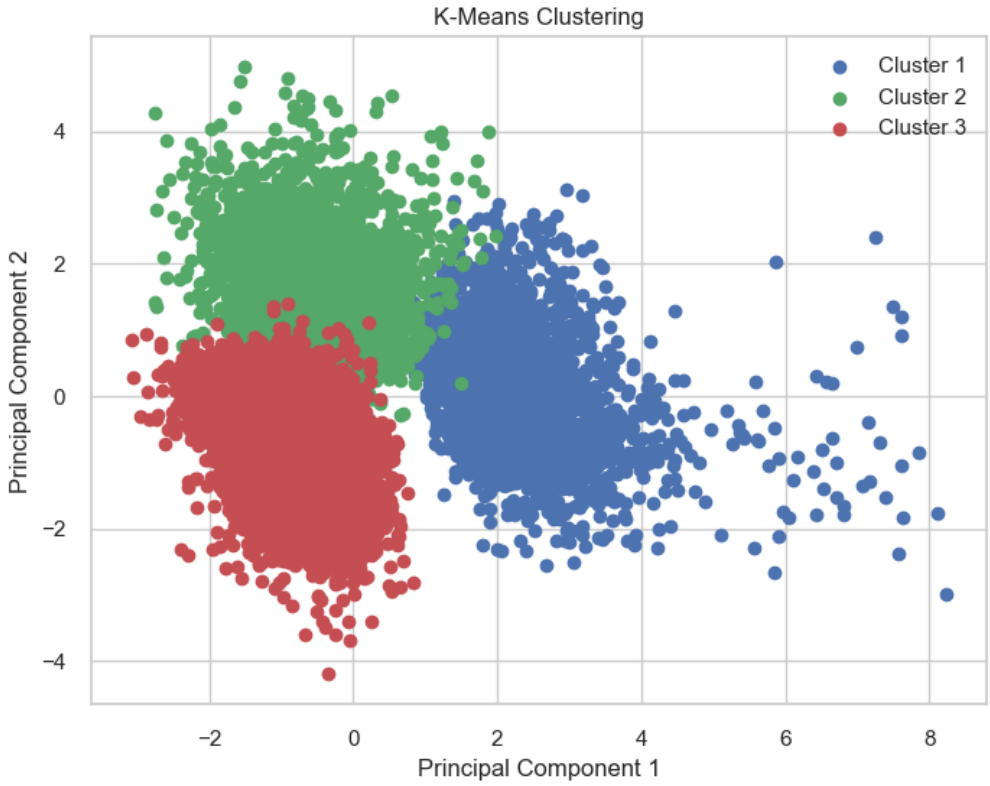

Figure 3. K-Means Clustering of Variables

Table 2. The Result of the Clustering

Cluster	year	wrkstat	age	degree	sex	happy	realrinc	postlife	natsoc	marital	satjob	sibs	childs	race	relig
0	2001	1.26	47.17	3.28	1.40	1	49419	1.19	1.79	1.42	1.48	2.84	2.22	1.13	2.12
1	2001	1.61	37.90	1.39	1.66	2	15675	1.22	1.28	3.31	1.93	4.92	2.11	1.81	2.60
2	1997	1.53	44.74	1.17	1.55	3	18894	1.14	1.41	1.41	1.57	3.68	2.49	1.07	1.59

- **Cluster 0** represents individuals with an average year of 2001, highest satisfaction ('happy' level 1), relatively older age, higher income ('realrinc'), and etc,.
- **Cluster 1** includes individuals with an average year of 2001, relatively moderate satisfaction ('happy' level 2), younger age, lower income, and predominantly married and job-satisfied individuals.
- **Cluster 2** is characterized by individuals with an average year of 1997, the lowest satisfaction ('happy' level 3), moderately older age, and lower income. This cluster also includes individuals who are not predominantly married and have varying job satisfaction.

3.3 Relations between 'happy' and other variables

I decided to select the 'happy' variable as the independent variable to explore its relationship with other variables in more detail.

First, when I plotted the average 'happy' values over the years in a line graph, here's what I found:

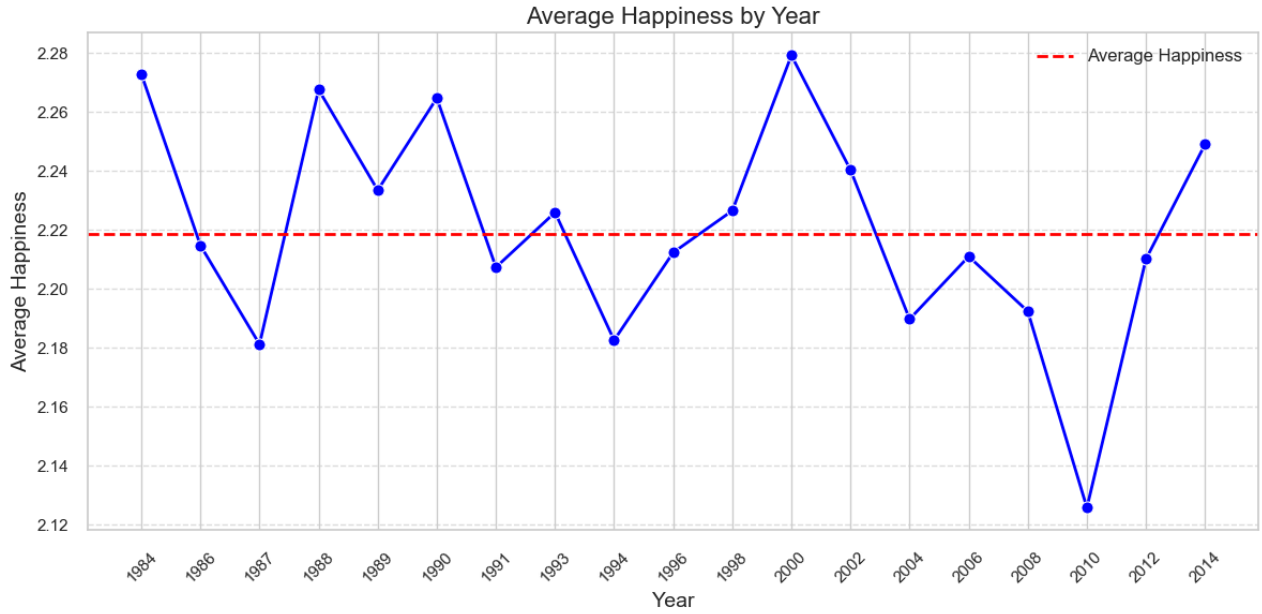

Figure 4. Average Happiness by Year

The overall average 'happy' value for the entire dataset was 2.22, which falls in the middle range among the values 1 (Very happy), 2 (Pretty happy), and 3 (Not too happy). The year with the highest 'happy' value was 2000, and the lowest was in 2010. Before 1994, the 'happy' values were generally higher, and after that year, they tend to be lower.

In the next analysis, I used a bar graph to depict the average 'happy' values according to marital status. It was clear that individuals who were married had the highest average 'happy' values.

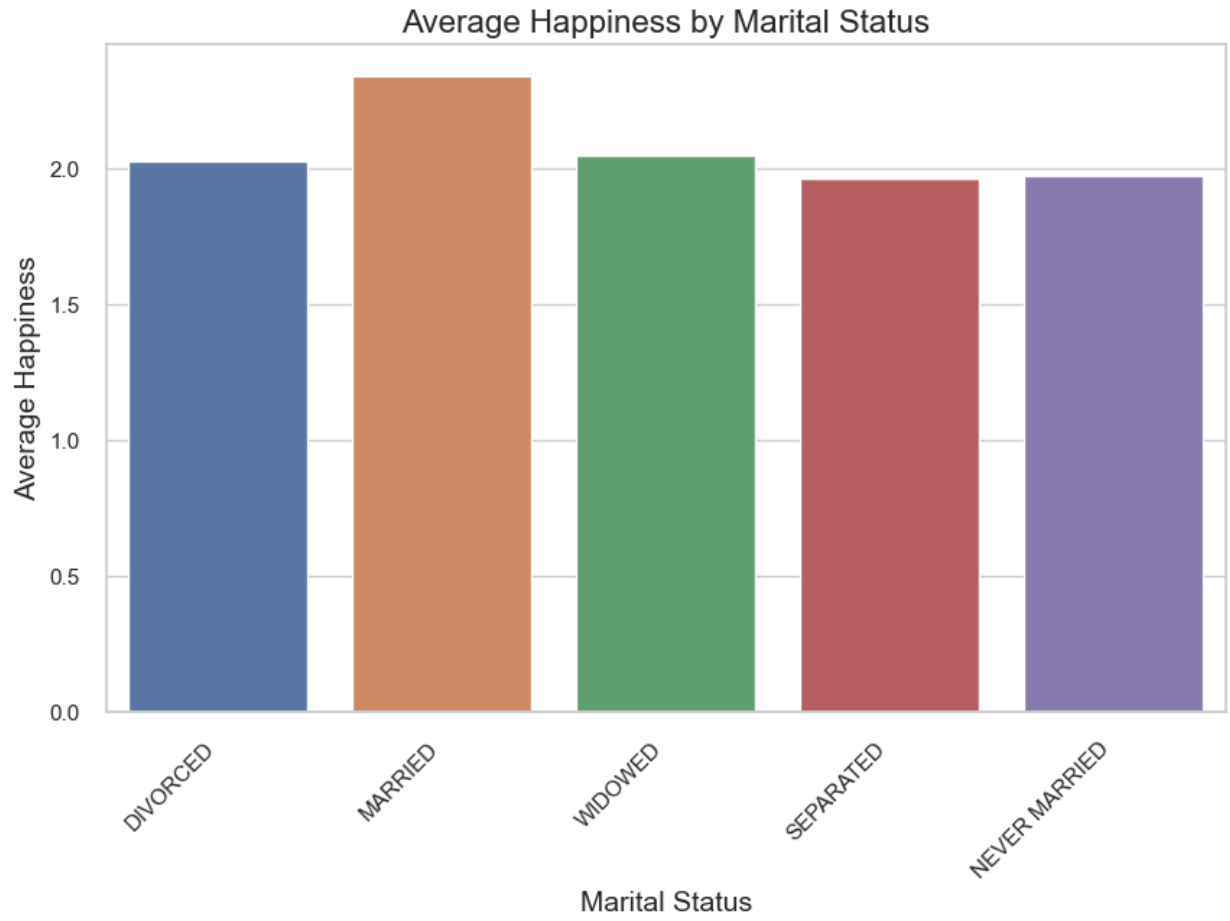

Figure 5. Average Happiness by Marital Status

Finally, I examined the relationship between belief in the post-life and 'happy' by creating a bar graph.

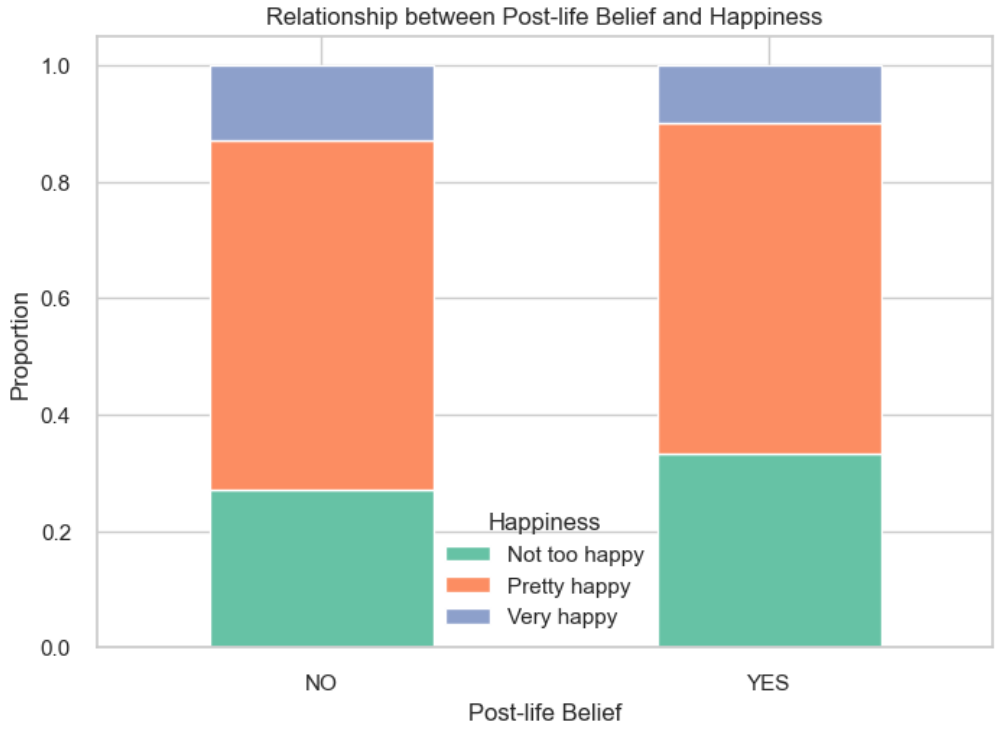

Figure 6. Average Happiness by Marital Status

The results showed that individuals who didn't believe in the post-life had a slightly higher proportion of "very happy" compared to those who believed.

4. Suggestion

Project Title: Analyzing Happiness Indices and Associated Factors

Project Description: This project is dedicated to uncovering valuable insights from the General Social Survey (GSS) dataset, with the primary objective of understanding the factors that influence happiness. By employing advanced data analysis techniques, the project aims to gain a deeper understanding of the relationships between various socio-economic and personal variables and an individual's happiness levels.

- **In-Depth Data Analysis**
  - Conduct thorough data analysis to identify correlations, trends, and patterns related to happiness within the GSS dataset.
  - Leverage advanced statistical methods to uncover hidden insights.
- **Segmentation and Profiling**
  - Segment the dataset into distinct groups based on happiness levels, such as "Very Happy," "Pretty Happy," and "Not Too Happy."
  - Create detailed profiles of each group, highlighting the demographic and behavioral characteristics that distinguish them.
- **Predictive Modeling**
  - Develop predictive models to forecast an individual's happiness based on specific variables.
  - Evaluate the accuracy of these models and provide recommendations for optimizing happiness.

Expected Outcomes:

- A comprehensive understanding of the key drivers of happiness, including the most influential variables.
- Data-backed recommendations for individuals, policymakers, and organizations to enhance happiness.
- Insights that can guide future research and initiatives aimed at improving well-being.