Analysis and Prediction of Mental Health Indicators Using Readability, Sentiment, and Behavioral Features

Introduction

Overview of the Project and Dataset

This project creatively processes mental health information on behavioral features, investigating patterns, making predictions, and providing actionable insights. The dataset contains characteristics like:

Demographics: Gender, Country, and Occupation.

Behavioral Features: Days spent indoors, stress level, mood swings, difficulties in coping, interest in work, and social behavior.

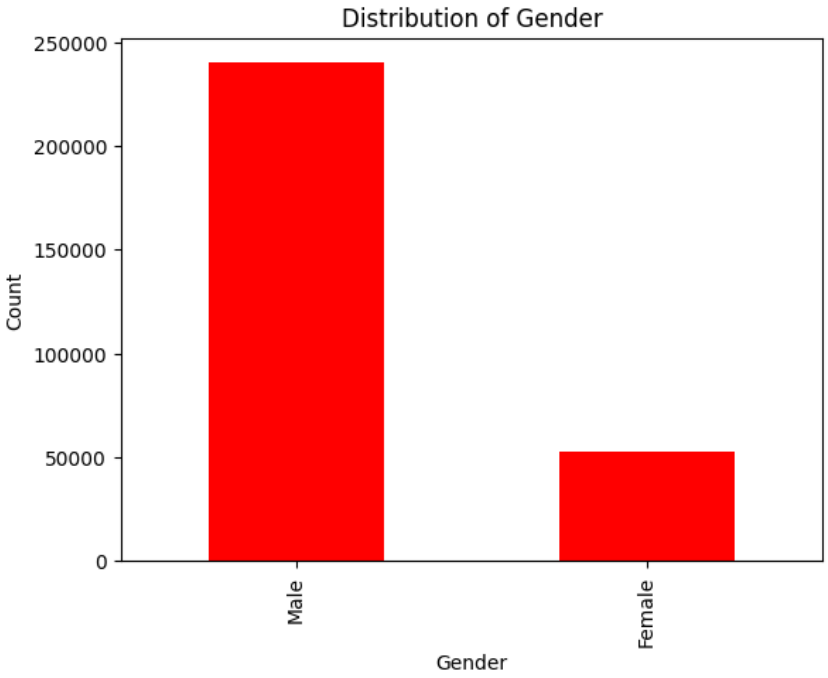
Mental Health Attributes: Family history, treatment history, and care options.

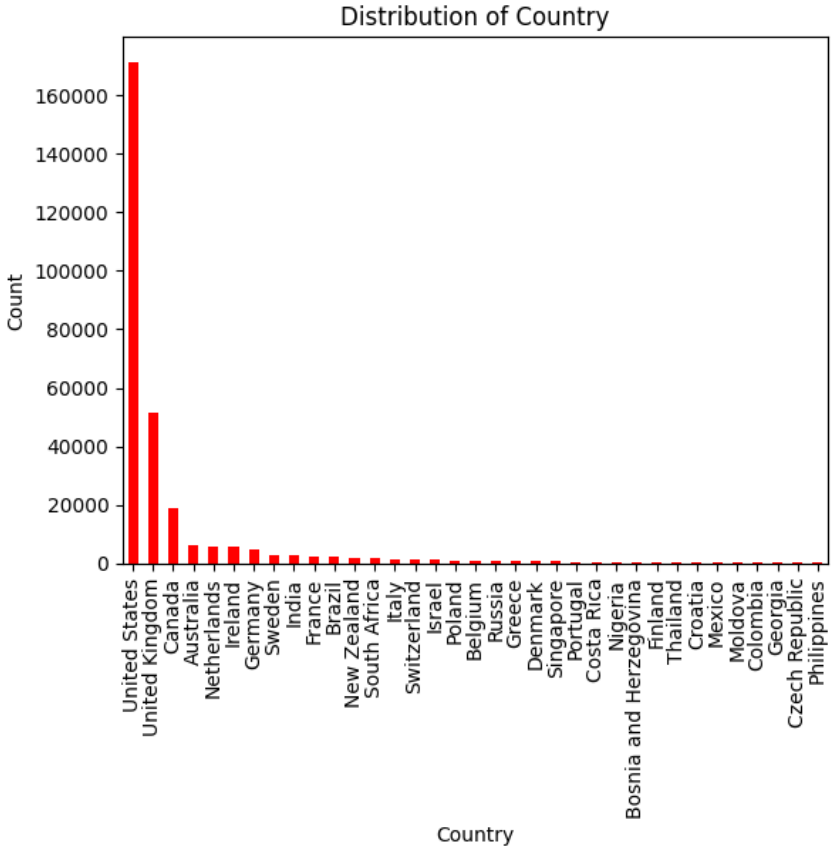
In this undertaking, a relationship will be established between these features and mental health outcomes, while mental health-related attributes will be modeled for predicting purposes.

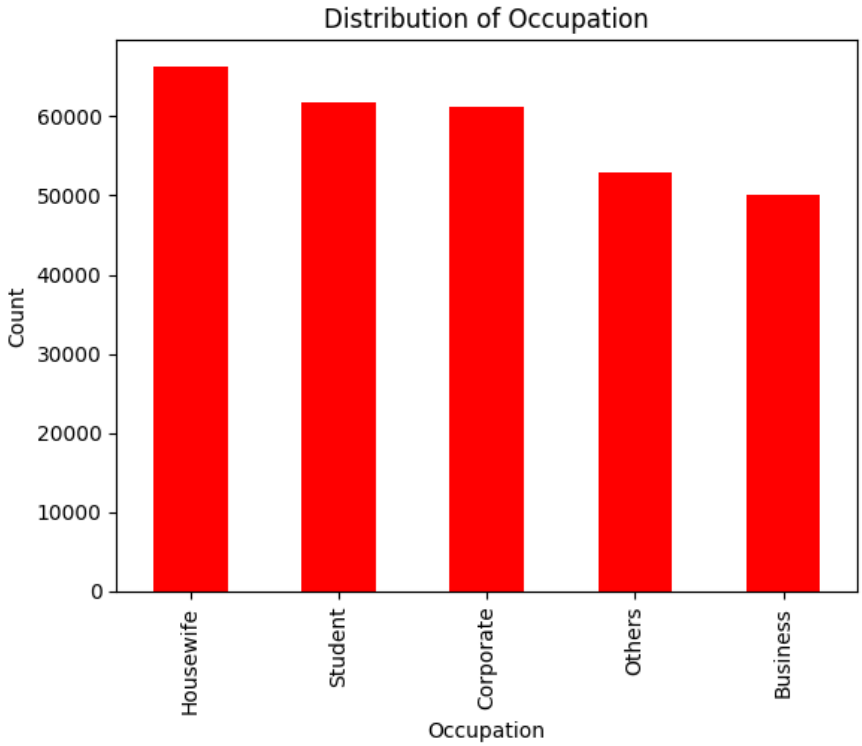
Exploratory Data Analysis (EDA) Findings

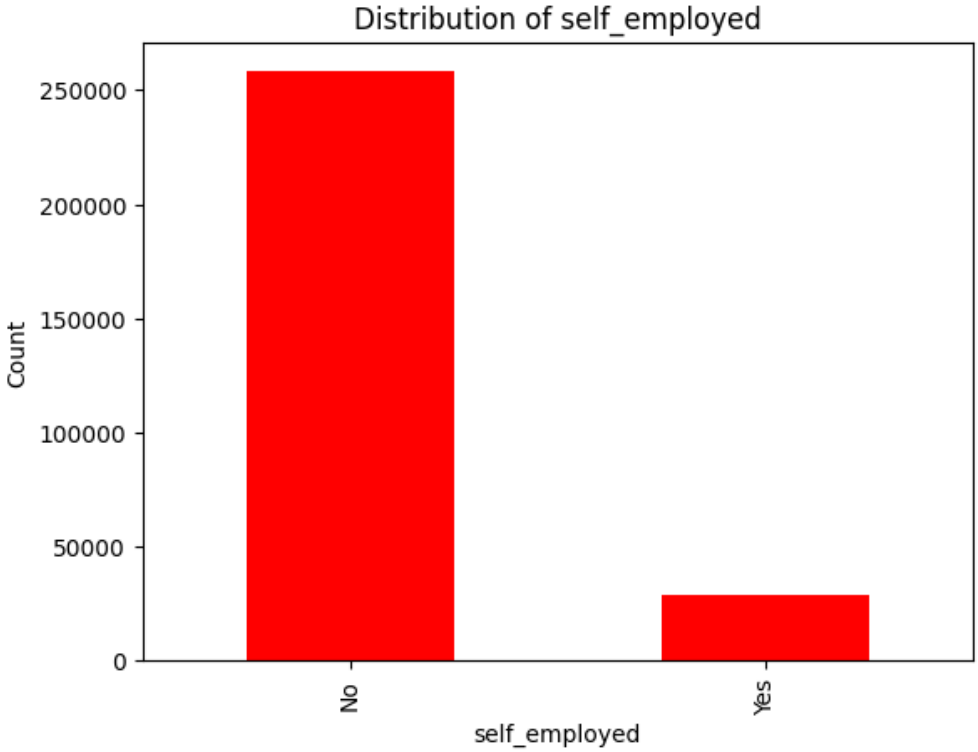
Key Patterns and Distributions

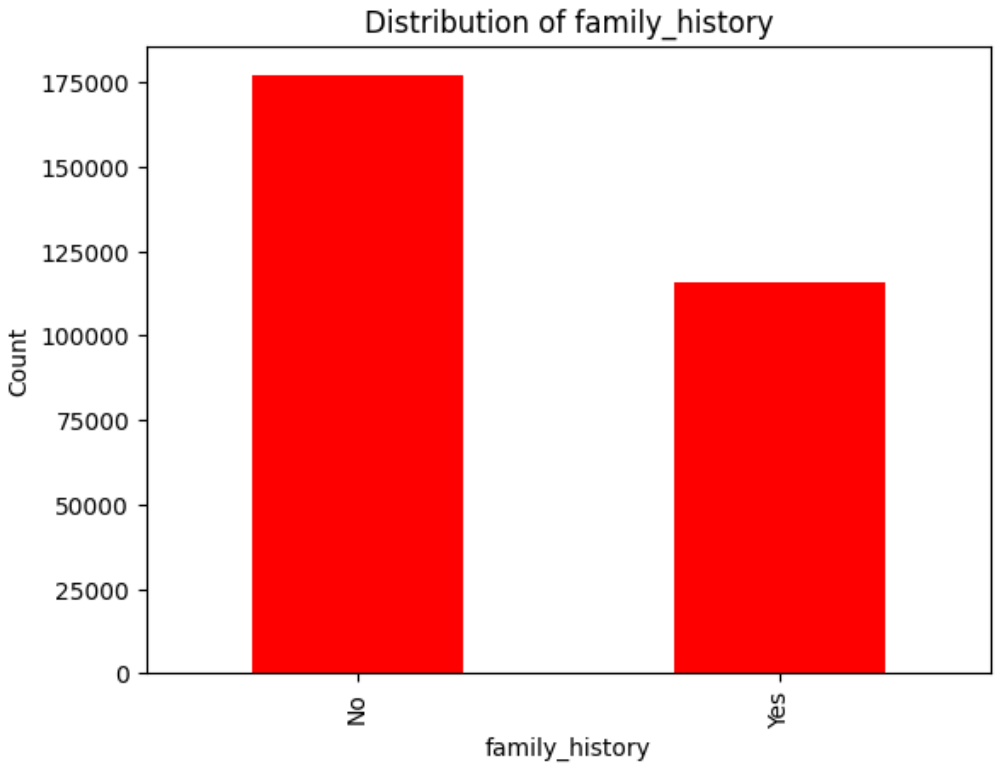
The categorical variables: The majority of responders were female and residing in the United States. Occupations are either corporate, students, or business-related. An appreciable number of respondents report family histories of mental disorders.


Distribution of Gender
250000
200000
150000
100000
50000
0
Count
Male
Female
Gender


Distribution of Country
160000
140000
120000
100000
80000
60000
40000
20000
0
Count
United States
United Kingdom
Canada
Australia
Netherlands
Ireland
Germany
Sweden
India
France
Brazil
New Zealand
South Africa
Italy
Switzerland
Israel
Poland
Belgium
Russia
Greece
Denmark
Singapore
Portugal
Costa Rica
Nigeria
Bosnia and Herzegovina
Finland
Thailand
Croatia
Mexico
Moldova
Colombia
Georgia
Czech Republic
Philippines
Country

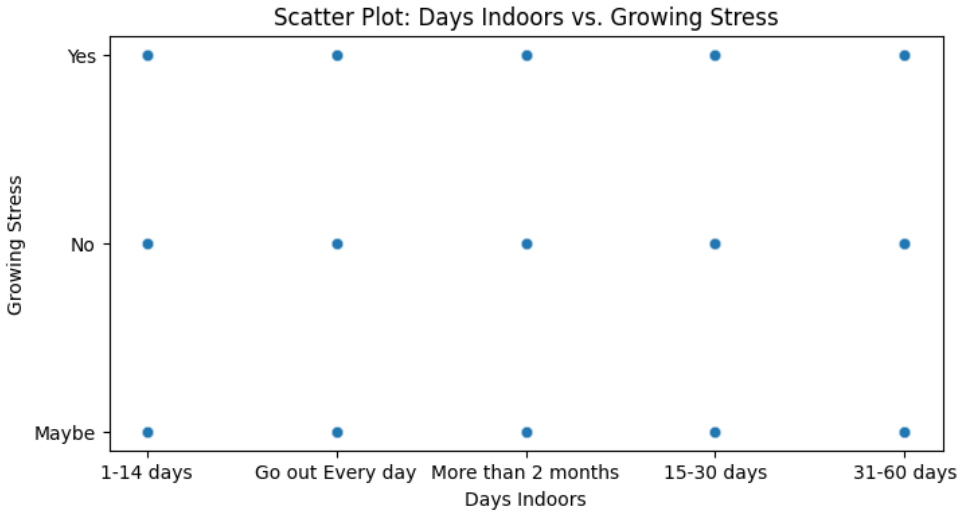

Distribution of Occupation
Count
60000
50000
40000
30000
20000
10000
0
Housewife
Student
Corporate
Others
Business
Occupation

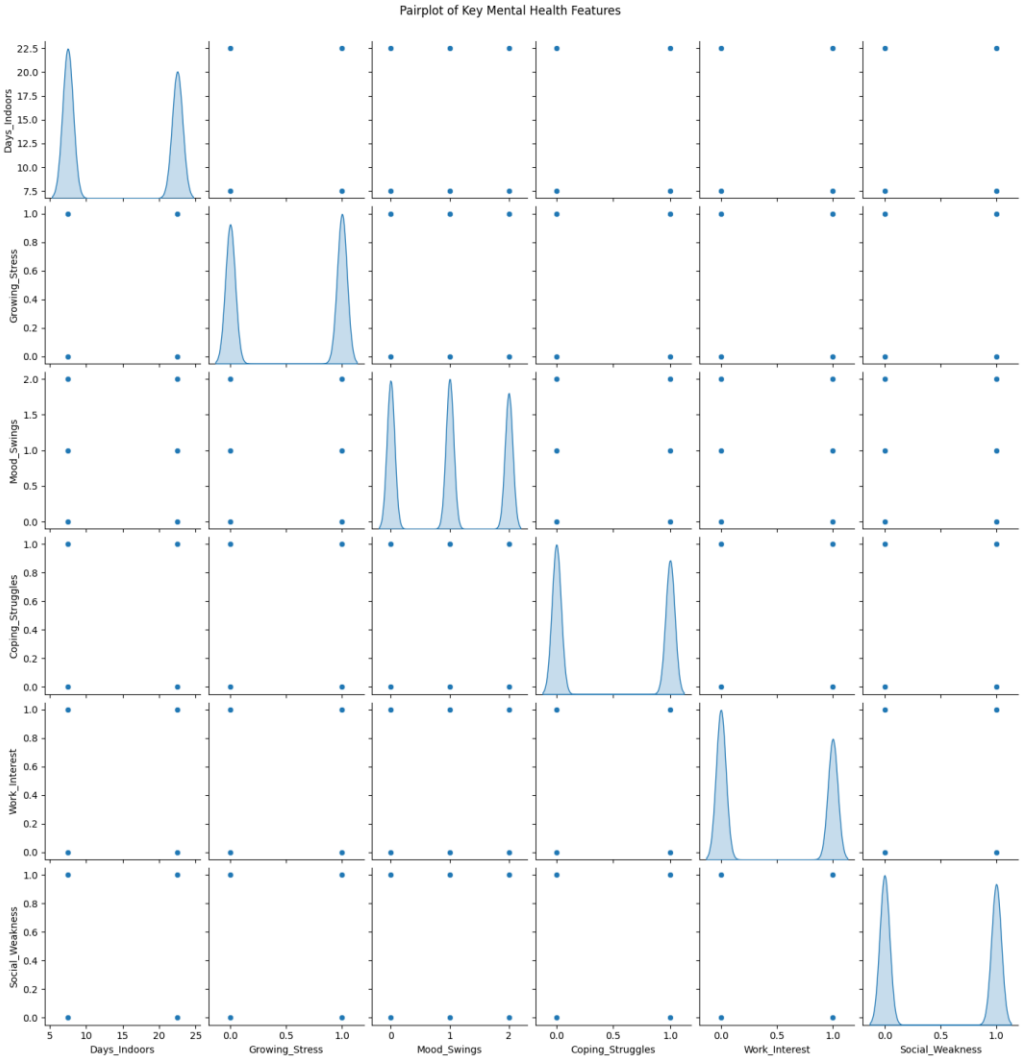

Distribution of self_employed
Count
250000
200000
150000
100000
50000
0
No
Yes
self_employed

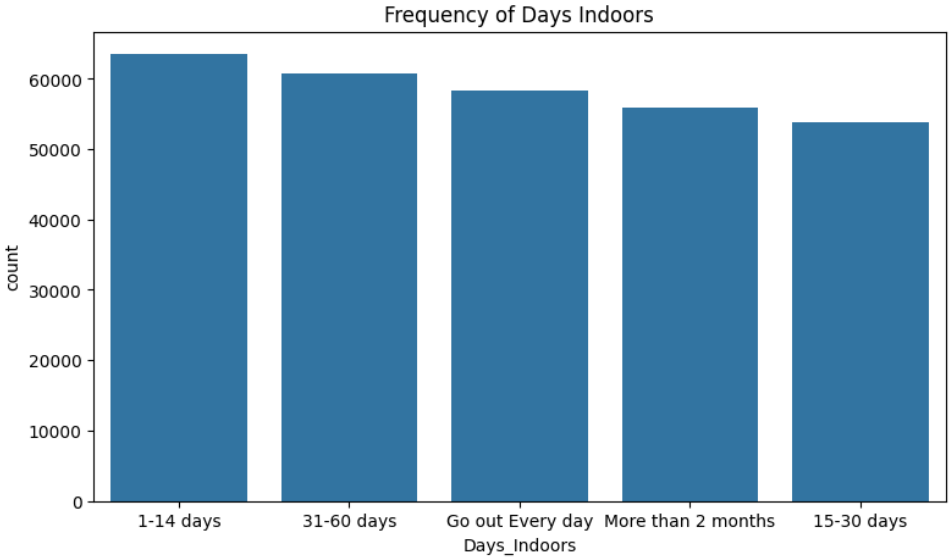


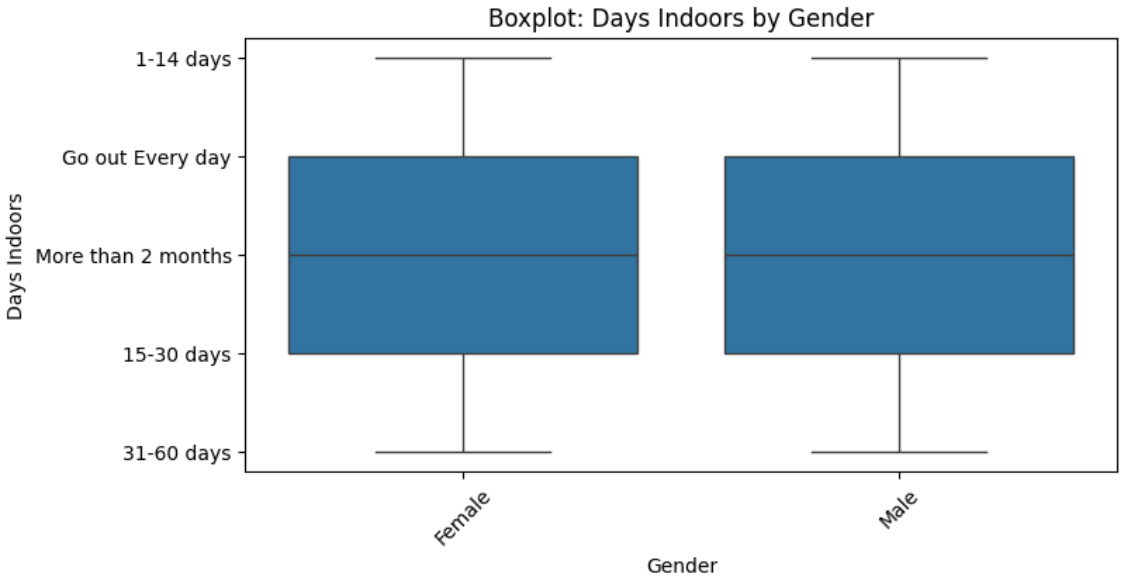
The numerical variables: Most days are in the criterion of 1-14 days being with days spent indoors. High stress is reported, many respondents reporting high stress now than before. Medium or high mood swings seemed to be the case most times.

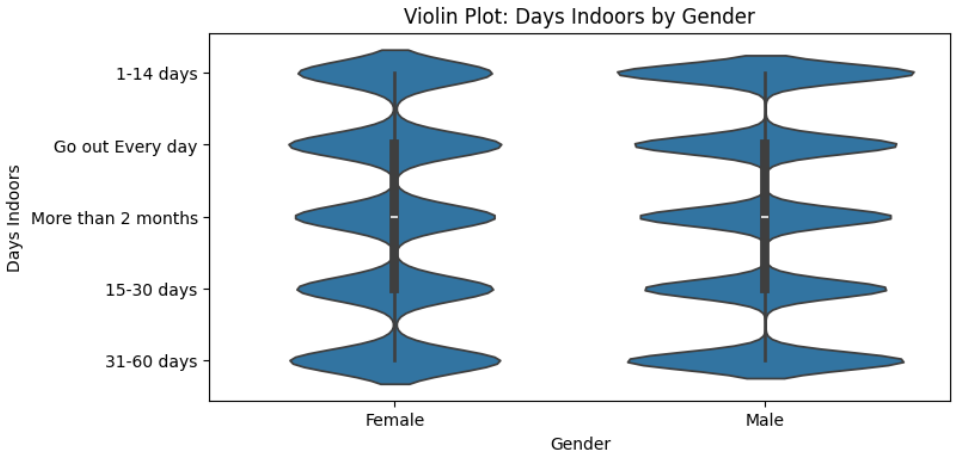
More Findings for Categorical data (Relations)

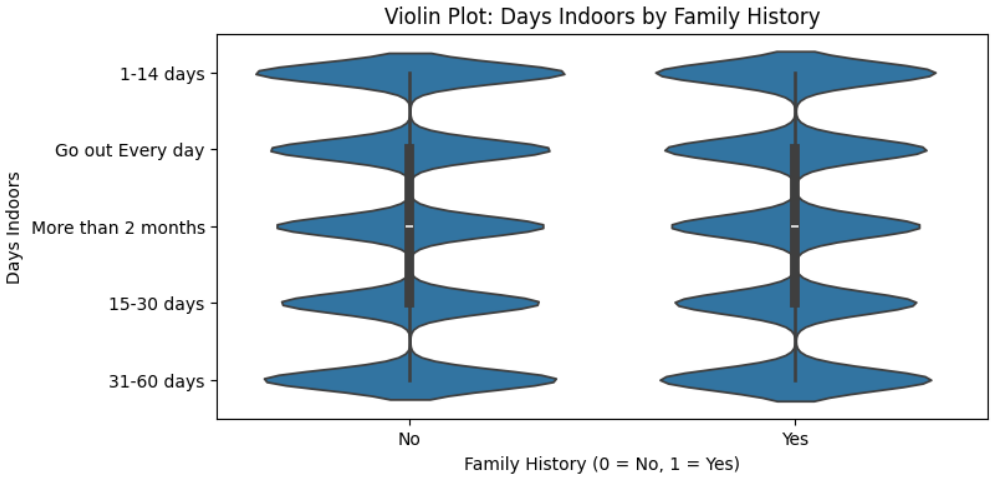



Pairplot of Key Mental Health Features
Days_Indoors
Growing_Stress
Mood_Swings
Coping_Struggles
Work_Interest
Social_Weakness

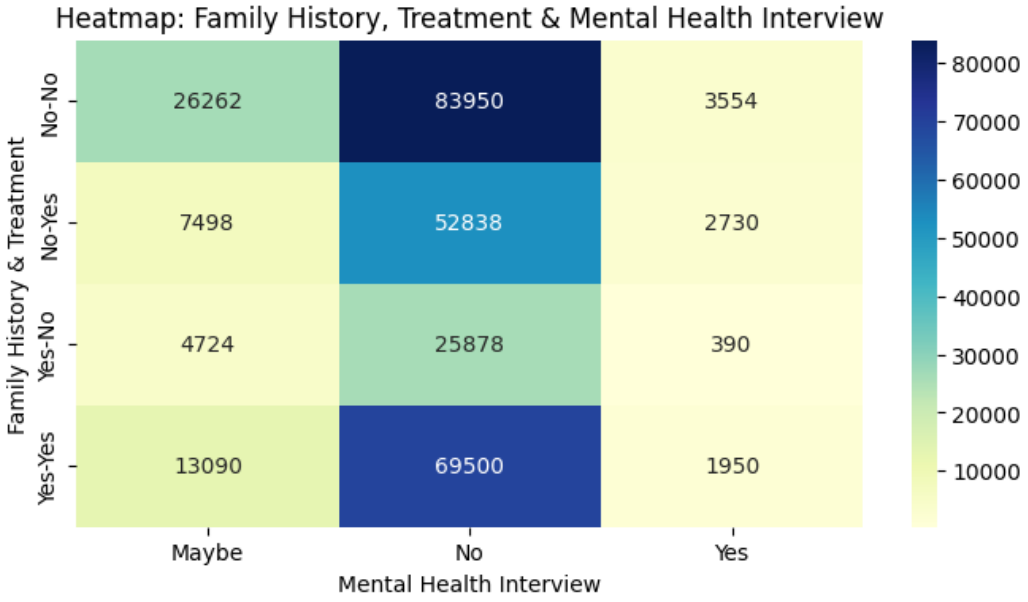

Frequency of Days Indoors
count
Days_Indoors
1-14 days
31-60 days
Go out Every day
More than 2 months
15-30 days

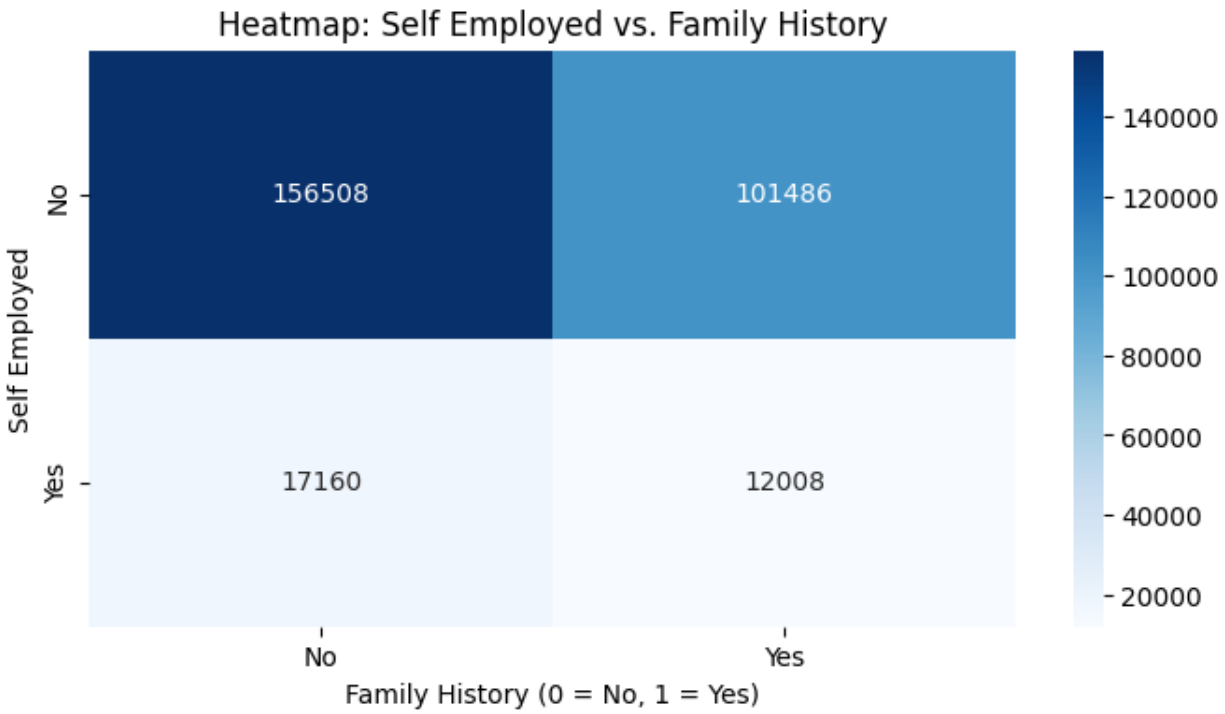

Boxplot: Days Indoors by Gender
1-14 days
Go out Every day
More than 2 months
15-30 days
31-60 days
Days Indoors
Female
Male
Gender

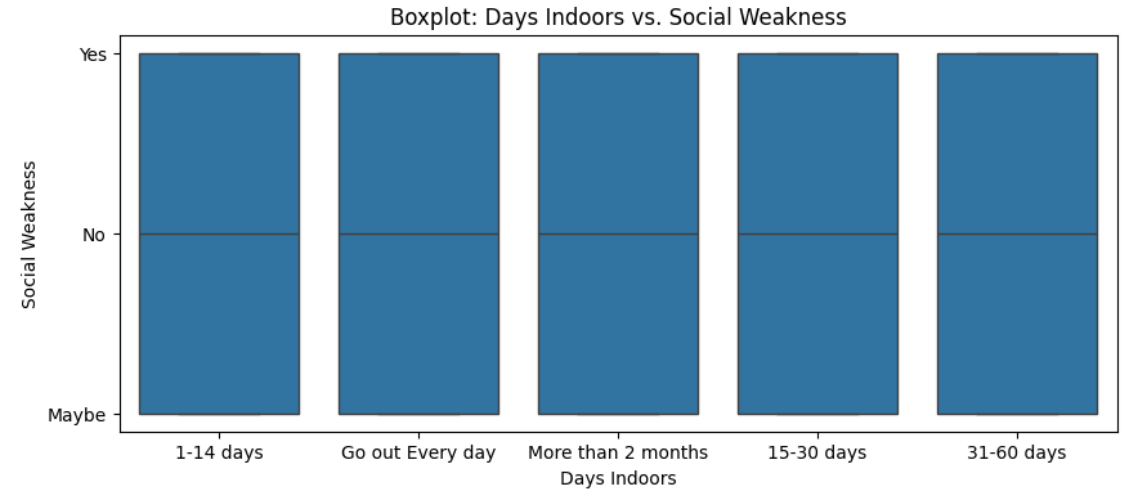

Violin Plot: Days Indoors by Gender
1-14 days
Go out Every day
More than 2 months
15-30 days
31-60 days
Days Indoors
Female
Male
Gender

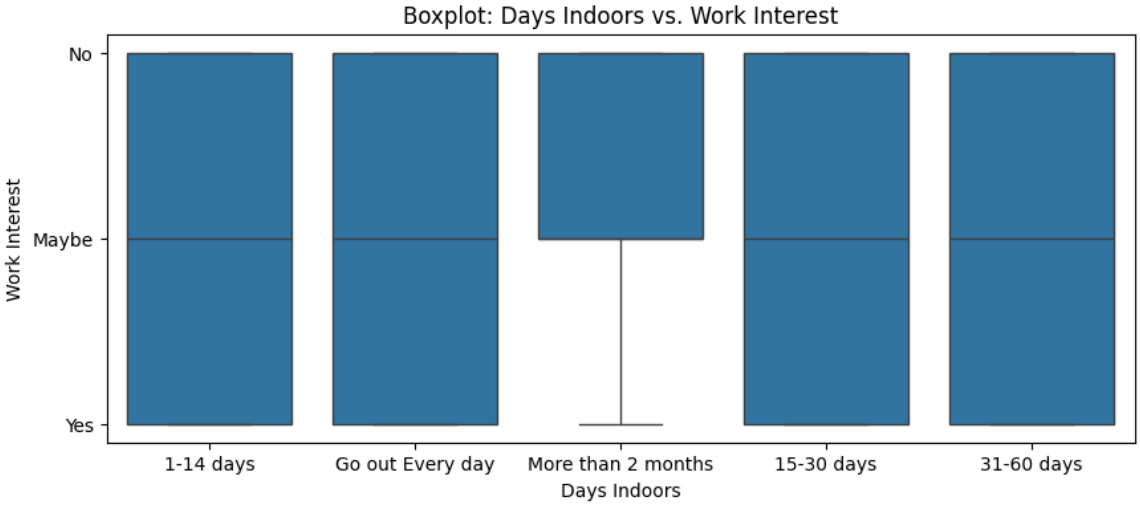

Violin Plot: Days Indoors by Family History
1-14 days
Go out Every day
More than 2 months
15-30 days
31-60 days
Days Indoors
No
Yes
Family History (0 = No, 1 = Yes)

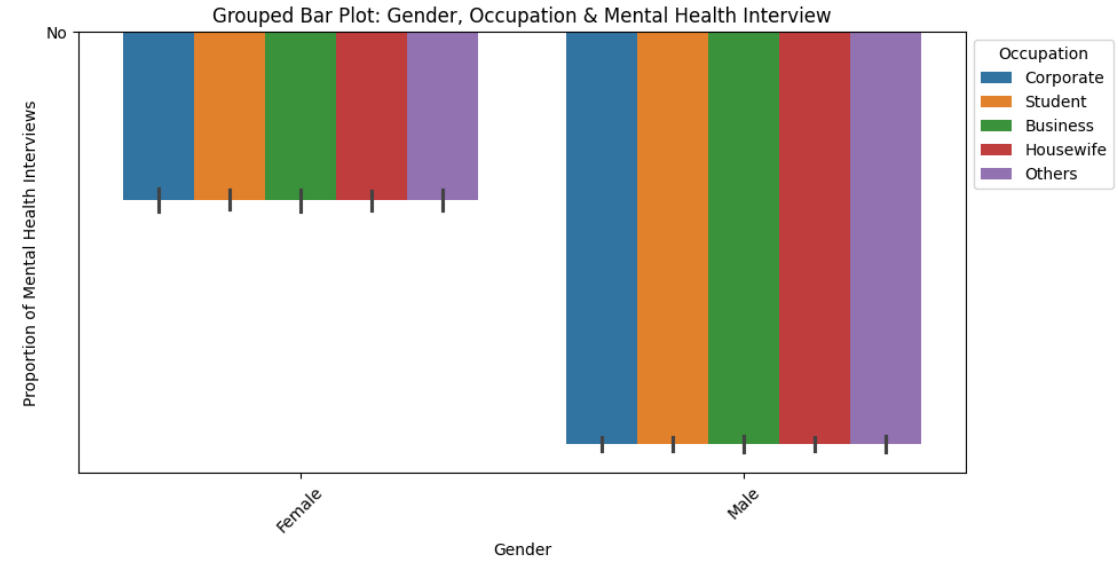
Family history and treatment-seeking behavior: People with family histories are more likely to seek treatment.



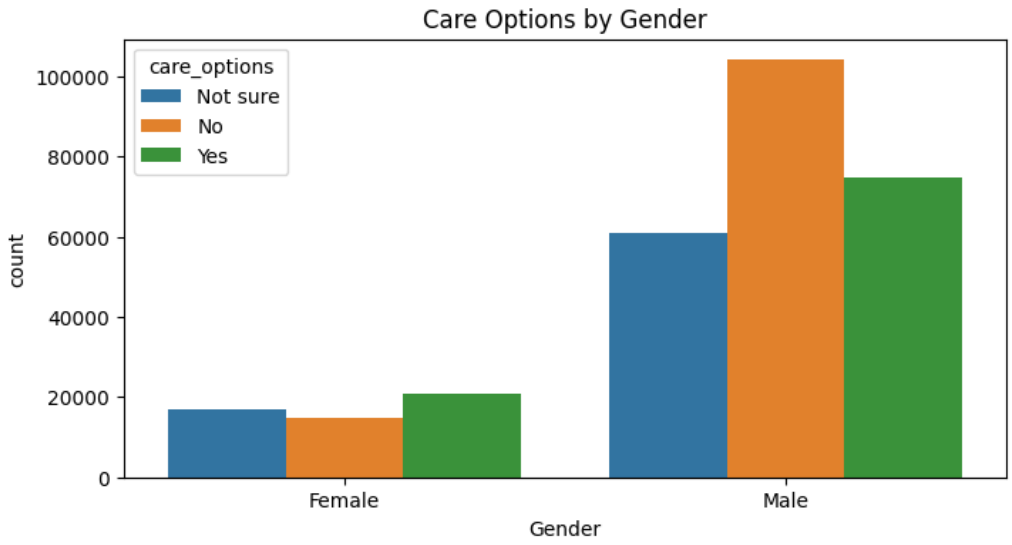


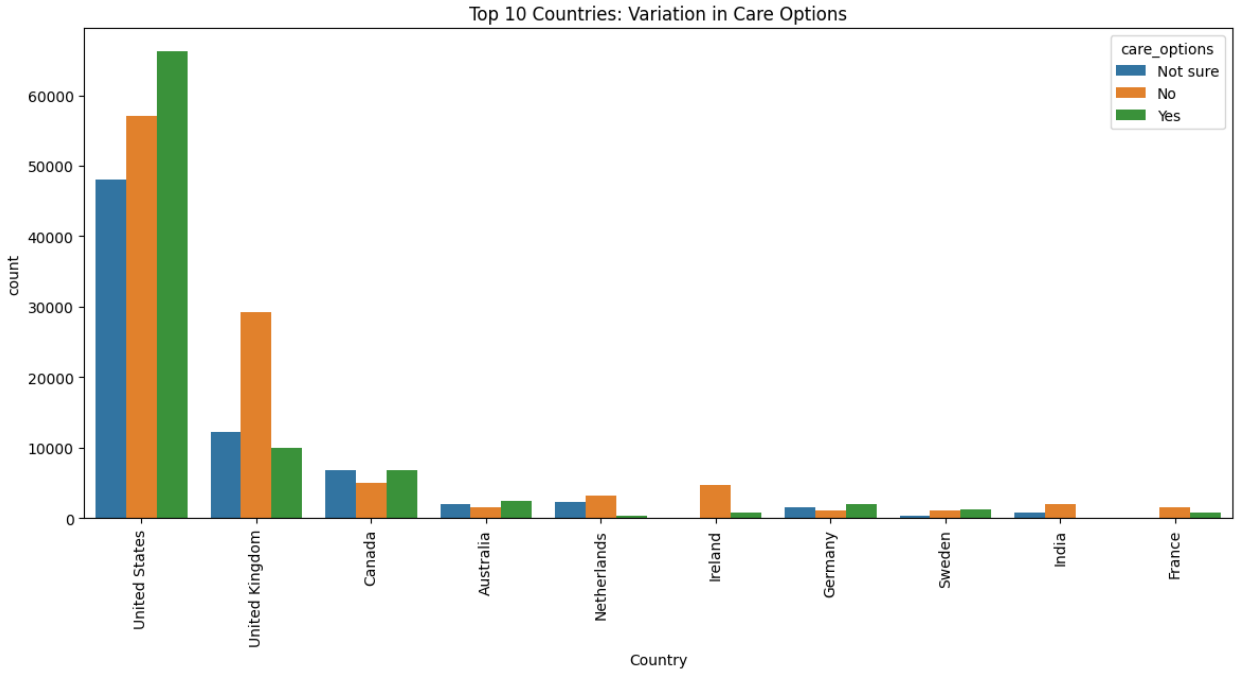

Boxplot: Days Indoors vs. Social Weakness
Social Weakness
Yes
No
Maybe
1-14 days
Go out Every day
More than 2 months
15-30 days
31-60 days
Days Indoors


Boxplot: Days Indoors vs. Work Interest
Work Interest
No
Maybe
Yes
1-14 days
Go out Every day
More than 2 months
15-30 days
31-60 days
Days Indoors


Grouped Bar Plot: Gender, Occupation & Mental Health Interview
Proportion of Mental Health Interviews
No
Female
Male
Gender
Occupation
Corporate
Student
Business
Housewife
Others

Very useful findings the relation between the Gender frequency and the care options features/attributes.





Modeling Results

Description of Models and Target Variables

1. Predicting Mental Health Treatment

• **Target Variable:** treatment

• **Objective:** Predict whether an individual is likely to seek or require mental health treatment based on factors like family history, stress levels, mood swings, and social weakness.

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.678775	0.68822	0.68002	0.675616
1	K-Nearest Neighbors	0.678775	0.68822	0.68002	0.675616
2	Logistic Regression	0.678775	0.68822	0.68002	0.675616

Linear regression is nor being used for this as this was mostly categorically distributed and all the dataset has been distributed categorically so that's why in whole project, regression (linear) is not being used.

2. Predicting Days Spent Indoors

• **Target Variable:** Days_Indoors

• **Objective:** Estimate how many days a person spends indoors based on stress levels, coping struggles, and other behavioral traits

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.968943	0.969482	0.969187	0.969326
1	K-Nearest Neighbors	0.289860	0.286472	0.283493	0.275391
2	Logistic Regression	0.286115	0.287989	0.278034	0.254778

3. Predicting Stress Levels

• **Target Variable:** Growing_Stress

• **Objective:** Use features such as Days_Indoors, Mood_Swings, and Coping_Struggles to predict an individual's stress level.

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.459015	0.459280	0.459231	0.458940
1	K-Nearest Neighbors	0.441845	0.445280	0.439689	0.435279
2	Logistic Regression	0.363894	0.354353	0.359438	0.346621

4. Predicting Changes in Habits

• **Target Variable:** Changes_Habits

• **Objective:** Predict the likelihood or extent of changes in habits based on mental health history, mood swings, and stress factors.

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.439707	0.435005	0.428874	0.423975
1	K-Nearest Neighbors	0.419903	0.418770	0.411985	0.409545
2	Logistic Regression	0.400287	0.379389	0.378269	0.346441

5. Predicting Work Interest

• **Target Variable:** Work_Interest

• **Objective:** Determine if stress, social weakness, or mental health factors impact a person's interest in work.

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.455013	0.458111	0.445852	0.444666
1	K-Nearest Neighbors	0.447198	0.452409	0.439736	0.439139
2	Logistic Regression	0.382416	0.371708	0.360353	0.314394
3	Random Forest	0.455013	0.458111	0.445852	0.444666

+ Code + Markdown

6. Predicting Likelihood of Mental Health Issues

• **Target Variable:** Mental_Health_History

• **Objective:** Predict whether a person has a history of mental health issues based on current stressors, family history, and coping mechanisms.

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.681015	0.681138	0.679052	0.679560
1	K-Nearest Neighbors	0.678860	0.678291	0.677835	0.678017
2	Logistic Regression	0.396422	0.386703	0.388681	0.378562
3	Random Forest	0.681066	0.681390	0.679290	0.679775

Discussion

Implications of Findings for Mental Health Analysis

Behavioral Patterns: High stress and mood swings signal a greater need for stress management and coping strategies.

Family history: influences a person's treatment-seeking behavior, therefore an area for focused intervention in at-risk populations.

Treatment Awareness: There is a large lack of awareness of potential treatments among survey respondents, showing a need for mental health education and access.

Potential Applications of Predictive Models

Personalized Interventions: Models have the capacity to flag an individual at risk and prescribe personalized options for patient development.

Policy Making: Findings could address important policy questions, such as raising awareness and improving access to care.

Workplace Mental Health: These models can help employers flag potential at-risk employees and intervene with support programs.

Conclusion and Recommendations

Summary of Insights and Model Performance

The dataset shows that relationships between behavioral features and mental health outcomes exist.

Decision Trees were the best-fitted models for classification problems, whereas Linear Regression fitted moderately well for stress severity prediction.

Important results were the effect of family history on treatment-seeking behavior and the prevalence of high-stress and mood swings.

The strange thing I find that logistic regression accuracy was less as compared to KNN algorithm and the decision tree and random forest analysis was also same this means that we can further improve the dataset for better analysis.

Recommendations:

I believe that data could be improved by inceasing the quantity of attributes and classes. If continous values will be introduced in the dataset then I can say that Linear Regression analysis will be work fine. For large dataset, logistic regression is also very cool along with decision tree and random forest.