Final Project

Aya Ibrahim

ERM 411: Information Analytics

Data Analysis, Visualization, & Storytelling

Understanding the Complex Relationship between Unemployment, Mental Health, and Socioeconomic Status

Table of Contents

Executive Summary

This paper delves into the connection between unemployment and mental health issues. It emphasizes the impact of mental health on professional lives and explores the causative factors behind the high unemployment rates among individuals with mental health challenges. The research is motivated by the significant role that financial stability plays in mental well-being and the attention this issue received post-pandemic, as unemployment rates increased significantly.

Utilizing a dataset from Kaggle, the paper conducts a comprehensive analysis through exploratory data analysis (EDA) and logistic regression models. Key findings highlight a strong correlation between mental health conditions like anxiety, depression, and panic attacks with unemployment rates, indicating that mental health challenges significantly impact employability. The study also finds a clear link between lower socioeconomic status and higher incidences of mental health issues, suggesting a nature between poverty and mental health. Additionally, higher education levels correlate with lower unemployment and fewer mental health issues, highlighting the importance of education in securing stable employment and maintaining good mental health. Regional disparities in unemployment rates suggest that location-specific economic and social factors can influence employment status. The paper concludes that a combination of factors, including resume gaps, anxiety, and depression, significantly predicts unemployment, highlighting the complex interaction between these elements in the context of employment challenges.

I. Introduction and Background

The intersection of unemployment and mental illness is a topic of huge significance and complexity, touching upon diverse fields such as public health, sociology, and economics. The focus of this project is to delve into the details of this relationship, exploring how mental health issues and unemployment rates are connected. The motivation for selecting this topic stems from the profound impact mental health has on individuals' professional lives and vice versa. Studies, including those by the National Alliance on Mental Illness (NAMI), have confirmed high unemployment rates among individuals with mental health challenges. However, this project aims to go beyond mere statistical relationships and delve into the causation – why individuals with mental health challenges face unemployment. This is such an important topic because financial instability leads to dissatisfaction. Unemployment, surprisingly, can have a substantial impact on an individual's psychological state. Maintaining good mental health is crucial for gaining and retaining employment. Joblessness induces stress, leading to persistent physical health issues and adverse mental health effects like depression, anxiety, and low self-esteem. Post-pandemic, this issue gained prominence: Unemployment increased during the pandemic from 3.8% (6.2 million people) in February 2020 to 13% (20.5 million) in May 2020, peaking at 14.7% (23.1 million) in April. Although it decreased to 6% by March 2021, it remained above pre-pandemic levels. “At least one study predicts that following COVID-19, mental illness will be the next pandemic” (“Unemployment and Mental Health: Resources to Managing Stress and Anxiety - Sunshine Behavioral Health,” 2021).

This project is driven by a desire to understand the multifaceted aspects of this topic. Questions at the heart of this project and that I hopefully aim to answer include: How does one's education level play a role in this narrative of unemployment among those with mental illness? What stories do the gaps in employment tell about one's income and life journey? And how do the subtle, often unseen symptoms of mental illness shape one's journey in the workforce? The exploration of these questions can provide insights vital for policymakers, mental health professionals, and the broader community.

The relevance of this study lies in its potential to inform strategies to support individuals with mental health challenges in the workforce. By analyzing the complex interplay between mental health and unemployment, this project aims to contribute to the development of more effective social welfare programs and employment policies. Additionally, it seeks to broaden the understanding of the societal and economic implications of mental health issues.

Ultimately, the goal is to leverage this analysis to benefit others by providing actionable insights that can be used to craft targeted interventions and support systems. This project aspires to understand how we can form a more inclusive and empathetic society where mental health challenges are not a barrier to professional success and personal fulfillment but are to be embraced in our life journey, yet addressed effectively.

II. Dataset Description

- **Kaggle:** *Unemployment & Mental Illness Dataset*

Data Collection & Methodology:

The dataset used in this project was sourced from Kaggle, a platform for data science competitions, and comprises data collected through Survey Monkey, ensuring a broad and unbiased representation of the population. Michael Corley, a Data Scientist at Sigma Data Science in Connecticut, US, was the contributor to this survey responsible for paying respondents on Survey Monkey to participate. The sample includes individuals both with and without mental illness, but the study's focus is on those with mental illnesses. The dataset is comprehensively collected, containing 334 observations with 32 features, including demographics, employment status, mental health status, and lifestyle information.

Data Intake Report & Feature Description:

Total number of observations/instances (rows)	334
Total number of features/attributes (columns)	31
Total number of files	1
Base Format of the file	Microsoft Excel Worksheet (.xlsx)
Size of the data	45.8 KB (46,955 bytes)

26 of the features are numerical and 6 are categorical.

```
Rows: 334
Columns: 31
$ `I_am_currently_employed_at_least_part-time`                    <dbl> 0, 1, 1…
$ I_identify_as_having_a_mental_illness                           <dbl> 0, 1, 0…
$ Education                                                       <chr> "High S…
$ I_have_my_own_computer_separate_from_a_smart_phone              <dbl> 0, 1, 1…
$ I_have_been_hospitalized_before_for_my_mental_illness           <dbl> 0, 0, 0…
$ How_many_days_were_you_hospitalized_for_your_mental_illness     <dbl> 0, 0, 0…
$ I_am_legally_disabled                                           <dbl> 0, 0, 0…
$ I_have_my_regular_access_to_the_internet                        <dbl> 1, 1, 1…
$ I_live_with_my_parents                                          <dbl> 0, 0, 0…
$ I_have_a_gap_in_my_resume                                       <dbl> 1, 0, 0…
$ Total_length_of_any_gaps_in_my_resume_in_months                 <dbl> 24, 1, …
$ `Annual_income_(including_any_social_welfare_programs)_in_USD`  <dbl> 35, 22,…
$ I_am_unemployed                                                 <dbl> 1, 0, 0…
$ I_read_outside_of_work_and_school                               <dbl> 1, 1, 1…
$ Annual_income_from_social_welfare_programs                      <dbl> 0, 0, 0…
$ I_receive_food_stamps                                           <dbl> 0, 0, 0…
$ I_am_on_section_8_housing                                       <dbl> 0, 0, 0…
$ How_many_times_were_you_hospitalized_for_your_mental_illness    <dbl> 0, 0, 0…
$ Lack_of_concentration                                           <dbl> 1, 1, 0…
$ Anxiety                                                         <dbl> 1, 1, 0…
$ Depression                                                      <dbl> 1, 1, 0…
$ Obsessive_thinking                                              <dbl> 1, 0, 0…
$ Mood_swings                                                     <dbl> 0, 0, 0…
$ Panic_attacks                                                   <dbl> 1, 1, 0…
$ Compulsive_behavior                                             <dbl> 0, 0, 0…
$ Tiredness                                                       <dbl> 0, 1, 0…
$ Age                                                             <chr> "30-44"…
$ Gender                                                          <chr> "Male",…
$ Household_Income                                                <chr> "$25,00…
$ Region                                                          <chr> "Mounta…
$ Device_Type                                                     <chr> "Androi…
```

Out of the 334 cases, 80 respondents identified as having a mental illness, mirroring the general population's 20-25% estimate of mental illness prevalence.

The dataset owner claims that the dataset is clean with no missing or duplicate values. It contains a comprehensive set of features that provide insights into the respondents' socio-economic and health status. This dataset aims to provide a multifaceted view of the respondents' life circumstances and can be useful in deriving meaningful conclusions for my final analysis.

Below is a list of the features and their description:

Age: A numerical variable representing the respondent's age.
Gender: A categorical variable representing the respondent's gender. The unique values are: ["Male", "Female"].
Education: A categorical variable representing the respondent's current education level. The unique values are: ["High School or GED", "Some highschool", "Some Phd", "Completed Phd", "Completed Undergraduate", "Some Undergraduate", "Completed Masters", "Some Masters"].
Region: A categorical variable representing what region the respondent is from. The unique values are: ["Mountain", "East South Central", "Pacific", "New England", "East North Central", "South Atlantic", "Middle Atlantic", "West South Central", "West North Central"].
Household_Income: A categorical variable representing the respondent's range of household income.
Device_Type: A categorical variable representing the respondent's device type.

The unique values are: ["Android Phone / Tablet", "MacOS Desktop / Laptop", "Ios Phone / Tablet", "Other"].
I_am_currently_employed_at_least_part-time: A binary variable indicating if the respondent is employed at least part-time. The unique values are: [0, 1] for No and Yes respectively.
I_am_unemployed: A binary variable indicating if the respondent is unemployed or not. The unique values are: [0, 1] for No and Yes respectively.
I_identify_as_having_a_mental_illness: A binary variable indicating if the respondent has a mental illness or not. The unique values are: [0, 1] for No and Yes respectively.
I_have_been_hospitalized_before_for_my_mental_illness: A binary variable indicating if the respondent was hospitalized before their mental illness. The unique values are: [0, 1] for No and Yes respectively.
How_many_days_were_you_hospitalized_for_your_mental_illness: A numerical variable indicating the number of days the respondent was hospitalized for their mental illness.
How_many_times_were_you_hospitalized_for_your_mental_illness: A numerical variable indicating the number of times the respondent was hospitalized for their mental illness.
I_have_my_own_computer_separate_from_a_smart_phone: A binary variable indicating if the respondent owns both a computer and a smartphone. The unique values are: [0, 1] for No and Yes respectively.
I_am_legally_disabled: A binary variable indicating if the respondent is disabled.

The unique values are: [0, 1] for No and Yes respectively.
I_have_my_regular_access_to_the_internet: A binary variable indicating if the respondent has regular internet access. The unique values are: [0, 1] for No and Yes respectively.
I_have_a_gap_in_my_resume: A binary variable indicating if the respondent has a gap in their resume. The unique values are: [0, 1] for No and Yes respectively.
I_live_with_my_parents**:** A binary variable indicating if the respondent lives with their parents. The unique values are: [0, 1] for No and Yes respectively.
Total_length_of_any_gaps_in_my_resume_in_months: A numerical variable indicating the total numb er of gap months the respondent has in their resume.
Annual_income_(including_any_social_welfare_programs)_in_USD: A numerical variable indicating the respondent's annual income in US dollars including any welfare income earned.
Annual_income_from_social_welfare_programs: A numerical variable indicating the respondent's ann ual income only from social welfare programs in US dollars.
I_read_outside_of_work_and_school: A binary variable indicating if the respondent reads outside their career and academic life. The unique values are: [0, 1] for No and Yes respectively.
I_receive_food_stamps: *A* binary variable indicating if the respondent lives with their parents. The unique values are: [0, 1] for No and Yes respectively.

I_am_on_section_8_housing: A binary variable indicating if the respondent is on Section 8 Housing, meaning the federal program gives qualifying participants a voucher and a public housing agency pays a significant portion of their rent. The unique values are [0, 1] for No and Yes respectively.
Lack_of_concentration: A binary variable indicating if the respondent experiences a lack of concentrati on. The unique values are [0, 1] for No and Yes respectively.
Anxiety: A binary variable indicating if the respondent has anxiety. The unique values are: [0, 1] for No and Yes respectively.
Depression: A binary variable indicating if the respondent has depression. The unique values are: [0, 1] for No and Yes respectively.
Obsessive_thinking: A binary variable indicating if the respondent experiences obsessive thinking. The unique values are: [0, 1] for No and Yes respectively.
Mood_swings: A binary variable indicating if the respondent experiences mood swings. The unique values are: [0, 1] for No and Yes respectively.
Panic_attacks: A binary variable indicating if the respondent experiences panic attacks. The unique values are: [0, 1] for No and Yes respectively.
Compulsive_behavior: A binary variable indicating if the respondent has compulsive behavior. The unique values are: [0, 1] for No and Yes respectively.
Tiredness: A binary variable indicating if the respondent experiences tiredness.

The unique values are: [0, 1] for No and Yes respectively.

Data Limitations:

- Sample Representation, Bias, and Size: Since the data was collected through an online platform like Survey Monkey, there's a possibility that it may not be representative of the broader population. This limitation could be due to the digital divide, where certain demographics may not have equal access to or familiarity with online survey platforms.
- Self-Reported Data Accuracy: The data relies on self-reporting, which can introduce biases. Respondents might provide socially desirable answers, underreport sensitive information, or misunderstand questions, leading to inaccuracies.
- Limited Demographic Diversity: The survey might have limited reach in terms of demographic diversity, such as age, socioeconomic status, or geographical location, potentially skewing results towards specific groups.
- Single Time Point: If the data is capturing a single point in time, it limits the ability to understand changes over time between unemployment and mental health.
- Survey Design and Questionnaire Limitations: The way questions are framed can significantly influence the responses. Leading or ambiguous questions might skew the results. Also, the absence of certain questions might limit the depth of the analysis.

III. Exploratory Data Analysis

Data Cleaning and Pre-processing:

In the course of the analysis performed on RStudio, it was observed that the dataset contained a total of 45 missing values. A significant proportion of these, 37 instances, were identified within the "Days_Hospitalized" variable. The remaining missing values were dispersed across several other variables, including "Lack_of_concentration," "Mood_swings," "Panic_attacks," "Compulsive_behavior," "Tiredness," and "Region."

For the purpose of enhancing interpretability in upcoming analyses and visualizations, I substituted the two absent values in the "Region" variable with "Unknown." Conversely, in handling the missing data pertaining to mental illness conditions, a decision was made to exclude all instances of NA values.

I chose to retain the missing values in the "Days_Hospitalized" variable. This decision was driven by a consideration to avoid substantial data loss. Furthermore, the dataset includes other related variables, such as "Times_Hospitalized," along with a range of mental illness conditions. These variables offer a good foundation for a more detailed analysis, allowing for the construction of an analogous narrative that effectively compensates for the missing data in "Days_Hospitalized."

In the dataset utilized for this analysis, the original feature names were notably lengthy, necessitating a modification for enhanced readability, particularly in the context of visualization analysis. To address this, I abbreviated the majority of these feature names. I also factored a few of the features such as Region, Age, Education, Income-related, etc. For the region feature, I factored them from the westernmost to the easternmost after the following research:

- Mountain:

This region includes states in the western United States that are mainly mountainous, including Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming.

- East South Central:

States in this region are located in the southeastern United States and typically include Alabama, Kentucky, Mississippi, and Tennessee.

- Pacific:

This region is on the west coast of the United States and includes Alaska, California, Hawaii, Oregon, and Washington.

- New England:

This northeastern region comprises six states: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont.

- East North Central:

This region is in the northern part of the midwest and includes Illinois, Indiana, Michigan, Ohio, and Wisconsin.

- South Atlantic:

This region covers the southeastern Atlantic coast and includes Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, and the District of Columbia.

- Middle Atlantic:

This region includes states in the northeastern United States, namely New Jersey, New York, and Pennsylvania.

- West South Central:

These states are in the south-central United States and include Arkansas, Louisiana, Oklahoma, and Texas.

- West North Central:

States in the northern part of the midwestern United States, including Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota.

The following outlines all modifications made:

```
'data.frame':   333 obs. of  31 variables:
 $ Employed                   : num  0 1 1 0 1 1 1 1 1 1 ...
 $ Mental_Illness             : num  0 1 0 0 1 0 0 1 0 1 ...
 $ Education                  : Factor w/ 8 levels "Completed Phd",..: 7 2 5 6 5 7 6 6 5 NA ...
 $ Owns_Computer              : num  0 1 1 1 1 1 1 1 1 1 ...
 $ Hospitalized_Before_Illness: num  0 0 0 0 1 0 0 0 0 0 ...
 $ Days_Hospitalized          : num  0 0 0 NA 35 0 0 0 0 0 ...
 $ Legally_Disabled           : num  0 0 0 0 1 0 0 0 0 0 ...
 $ Regular_Internet_Access    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Lives_With_Parents         : num  0 0 0 1 0 1 0 1 0 0 ...
 $ Resume_Gap                 : num  1 0 0 1 1 0 0 0 0 0 ...
 $ Gap_Duration_Months        : num  24 1 0 11 33 0 0 0 0 0 ...
 $ Annual_Income_USD          : num  35 22 100 0 32 0 1 11 73 12 ...
 $ Unemployed                 : num  1 0 0 1 0 0 0 0 0 0 ...
 $ Reads_Outside_Work_School  : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Welfare_AnnualIncome_USD   : num  0 0 0 0 30 0 0 0 0 0 ...
 $ Receives_FoodStamps        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ On_Section8_Housing        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Times_Hospitalized         : num  0 0 0 0 4 0 0 0 0 0 ...
 $ Lack_of_concentration      : num  1 1 0 0 1 0 0 1 1 0 ...
 $ Anxiety                    : num  1 1 0 0 1 0 0 1 1 1 ...
 $ Depression                 : num  1 1 0 0 1 0 0 1 1 1 ...
 $ Obsessive_thinking         : num  1 0 0 0 1 0 0 0 0 0 ...
 $ Mood_swings                : num  0 0 0 0 1 0 0 0 0 0 ...
 $ Panic_attacks              : num  1 1 0 0 1 0 0 1 0 0 ...
 $ Compulsive_behavior        : num  0 0 0 0 1 0 0 1 0 0 ...
 $ Tiredness                  : num  0 1 0 0 1 0 0 1 1 1 ...
 $ Age                        : Factor w/ 4 levels "> 60","45-60",..: 3 4 3 3 3 3 4 4 3 3 ...
 $ Gender                     : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
 $ Household_Income           : Factor w/ 10 levels "$200,000+","$175,000-$199,999",..: 8 7 3 8 8 10 5 5 7 4 ...
 $ Region                     : Factor w/ 9 levels "Pacific","Mountain",..: 2 6 1 9 5 7 8 8 7 7 ...
 $ Device_Type                : Factor w/ 5 levels "Android Phone / Tablet",..: 1 4 4 3 2 1 3 3 2 1 ...
```

Statistical Summary:

Summary Statistics for Numerical Features:

```
   Employed      Mental_Illness   Owns_Computer    Hospitalized_Before_Illness
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000
Median :1.0000   Median :0.0000   Median :1.0000   Median :0.00000
Mean   :0.6787   Mean   :0.2372   Mean   :0.8739   Mean   :0.07808
3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000

Days_Hospitalized Legally_Disabled Regular_Internet_Access Lives_With_Parents
Min.   :  0.000   Min.   :0.0000   Min.   :0.000           Min.   :0.0000
1st Qu.:  0.000   1st Qu.:0.0000   1st Qu.:1.000           1st Qu.:0.0000
Median :  0.000   Median :0.0000   Median :1.000           Median :0.0000
Mean   :  3.287   Mean   :0.0961   Mean   :0.964           Mean   :0.1111
3rd Qu.:  0.000   3rd Qu.:0.0000   3rd Qu.:1.000           3rd Qu.:0.0000
Max.   :100.000   Max.   :1.0000   Max.   :1.000           Max.   :1.0000
NA's   :37
  Resume_Gap     Gap_Duration_Months Annual_Income_USD   Unemployed
Min.   :0.0000   Min.   :  0.000     Min.   :  0.00    Min.   :0.0000
1st Qu.:0.0000   1st Qu.:  0.000     1st Qu.: 12.00    1st Qu.:0.0000
Median :0.0000   Median :  0.000     Median : 30.00    Median :0.0000
Mean   :0.2462   Mean   :  8.523     Mean   : 37.46    Mean   :0.2583
3rd Qu.:0.0000   3rd Qu.:  5.000     3rd Qu.: 55.00    3rd Qu.:1.0000
Max.   :1.0000   Max.   :100.000     Max.   :100.00    Max.   :1.0000

Reads_Outside_Work_School Welfare_AnnualIncome_USD Receives_FoodStamps
Min.   :0.0000            Min.   :  0.000          Min.   :0.00000
1st Qu.:1.0000            1st Qu.:  0.000          1st Qu.:0.00000
Median :1.0000            Median :  0.000          Median :0.00000
Mean   :0.8889            Mean   :  3.336          Mean   :0.06607
3rd Qu.:1.0000            3rd Qu.:  0.000          3rd Qu.:0.00000
Max.   :1.0000            Max.   :100.000          Max.   :1.00000

On_Section8_Housing Times_Hospitalized Lack_of_concentration    Anxiety
Min.   :0.00000     Min.   :  0.000    Min.   :0.0000        Min.   :0.0000
1st Qu.:0.00000     1st Qu.:  0.000    1st Qu.:0.0000        1st Qu.:0.0000
Median :0.00000     Median :  0.000    Median :0.0000        Median :0.0000
Mean   :0.02102     Mean   :  1.198    Mean   :0.1532        Mean   :0.2973
3rd Qu.:0.00000     3rd Qu.:  0.000    3rd Qu.:0.0000        3rd Qu.:1.0000
Max.   :1.00000     Max.   :100.000    Max.   :1.0000        Max.   :1.0000

  Depression     Obsessive_thinking  Mood_swings     Panic_attacks
Min.   :0.0000   Min.   :0.0000     Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000     Median :0.0000   Median :0.0000
Mean   :0.2553   Mean   :0.1261     Mean   :0.1141   Mean   :0.1471
3rd Qu.:1.0000   3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000     Max.   :1.0000   Max.   :1.0000

Compulsive_behavior   Tiredness
Min.   :0.00000     Min.   :0.0000
1st Qu.:0.00000     1st Qu.:0.0000
Median :0.00000     Median :0.0000
Mean   :0.08709     Mean   :0.3003
3rd Qu.:0.00000     3rd Qu.:1.0000
Max.   :1.00000     Max.   :1.0000
```

From the summary above, the following insights have been observed:

- Mental Illness:

About 24% reported having mental illness.

- Mental Health Symptoms (Disability, Ever Hospitalized, Days/Times Hospitalized, Lack_of_concentration, Anxiety, Depression, Obsessive_thinking, Mood_Swings, Panic_attacks, Compulsive Behavior, Tiredness):

Disability is rare with about only 10% reporting having it.

Anxiety and Depression have higher reports of about 30% and 26 % respectively. About 15% report having Panic attacks and lack of concentration. About 12-13% have mood swings and obsessive thinking. The least is about 3% in tiredness.

A very low percentage have been hospitalized before (7-8%), and thus a relatively low mean in days and times hospitalized reported.

- Access and Living Situation (Regular_Internet_Access, Lives_With_Parents, Owns_Computer):

The majority (96%) have regular internet access. The majority (87%) own a computer besides their smartphone. The majority (89%) read outside work and school. Only 11% live with their parents.

- Work and Income:

Gap_Duration_Months has a low mean, suggesting most respondents do not have long employment gaps.

Resume_Gap has a higher mean, possibly indicating gaps in employment are somewhat common. About 25% have a resume gap.

Annual_Income_USD shows some variation, with a mean significantly higher than the median, which may indicate some high earners skewing the average. 37,000 USD is the average annual income.

However, the average income received from social welfare programs is about 3,000 USD.

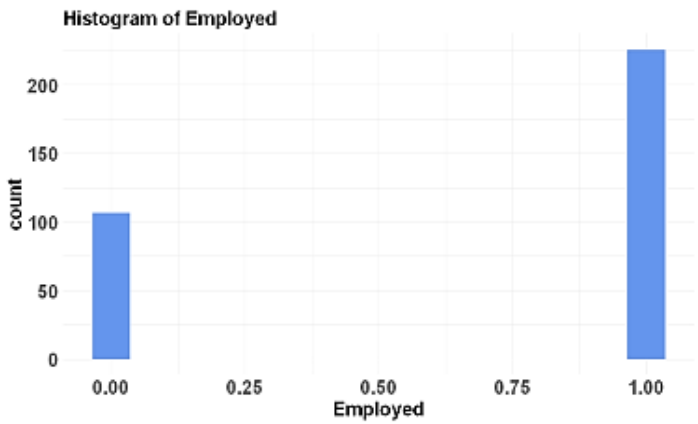
About 68% are employed. It is a binary variable with a nearly even split, as the median is 0.50.

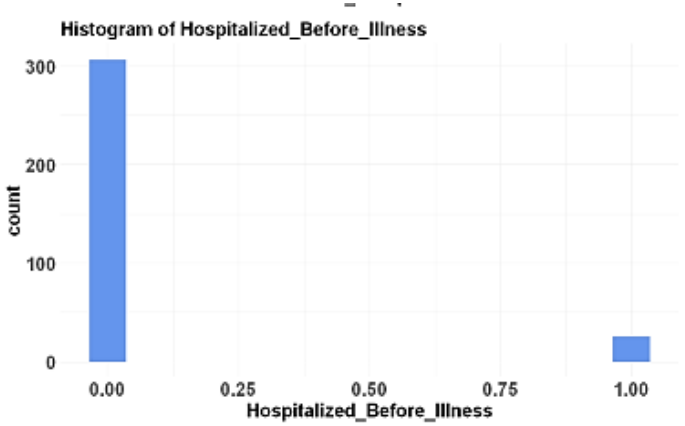
About 26% are unemployed. The mean suggests a lower rate of unemployment among respondents

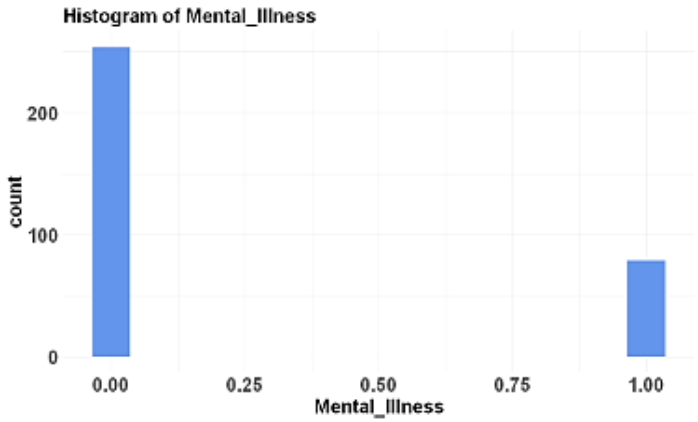
- Low Income Indicators (On_Section8_Housing, Receives_FoodStamps):

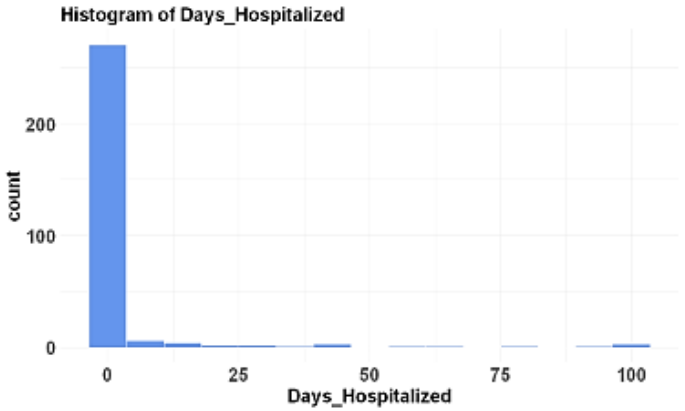
These variables show a low mean, indicating not many respondents receive food stamps or social welfare benefits. About 6% receive food stamps and 2% are on Section 8 Housing.

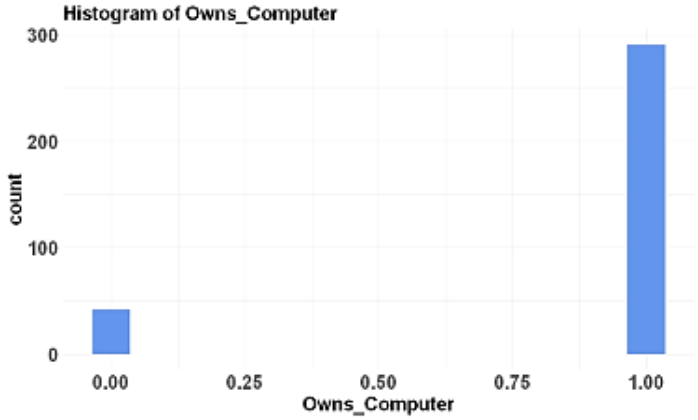
The following three pages include all histogram distributions confirming what has been observed in the summary.

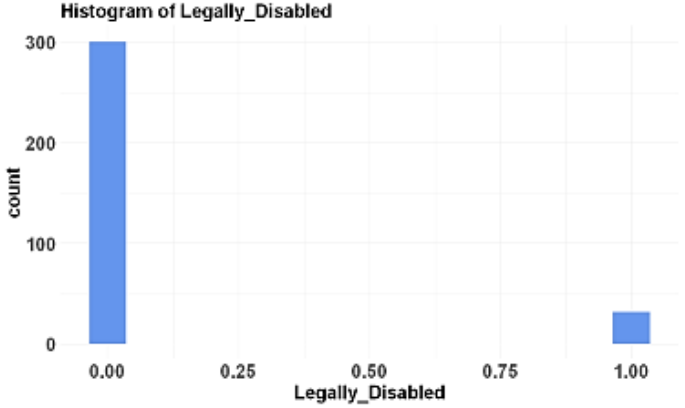

Histogram of Employed
count
Employed
0
50
100
150
200
0.00
0.25
0.50
0.75
1.00

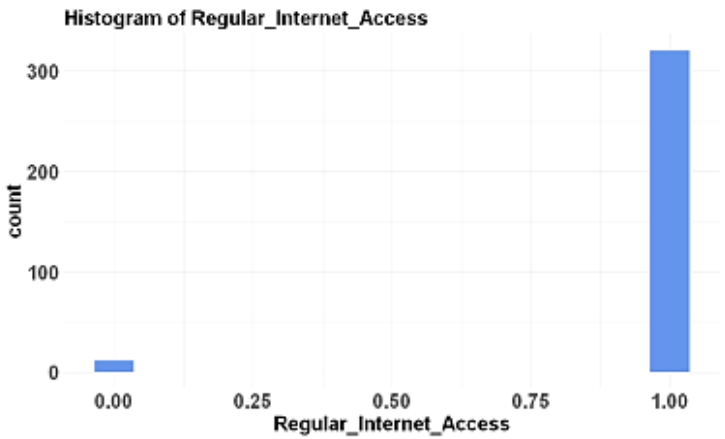

Histogram of Hospitalized_Before_Illness
count
Hospitalized_Before_Illness
0
100
200
300
0.00
0.25
0.50
0.75
1.00

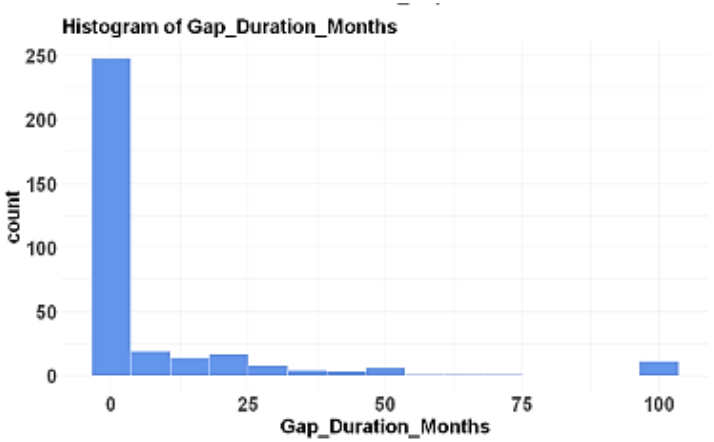

Histogram of Mental_Illness
count
Mental_Illness
0
100
200
0.00
0.25
0.50
0.75
1.00

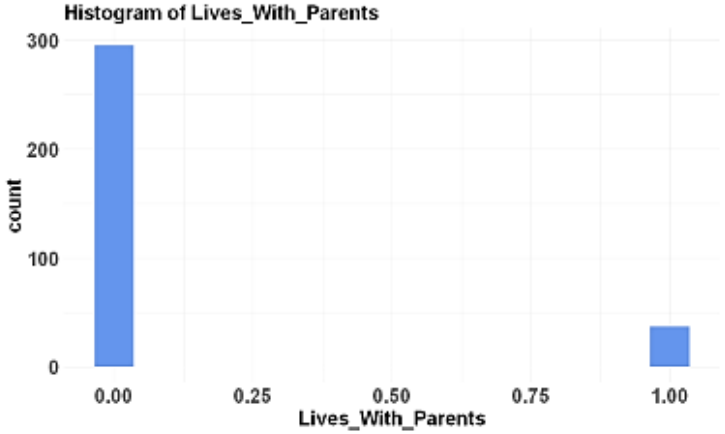

Histogram of Days_Hospitalized
count
Days_Hospitalized
0
100
200
0
25
50
75
100

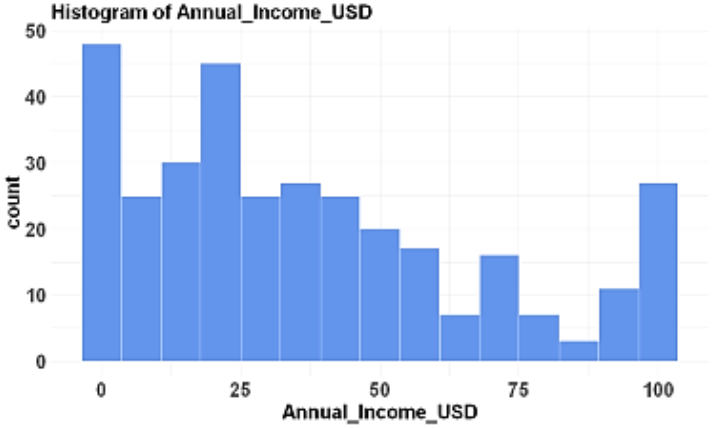

Histogram of Owns_Computer
count
Owns_Computer
0
100
200
300
0.00
0.25
0.50
0.75
1.00

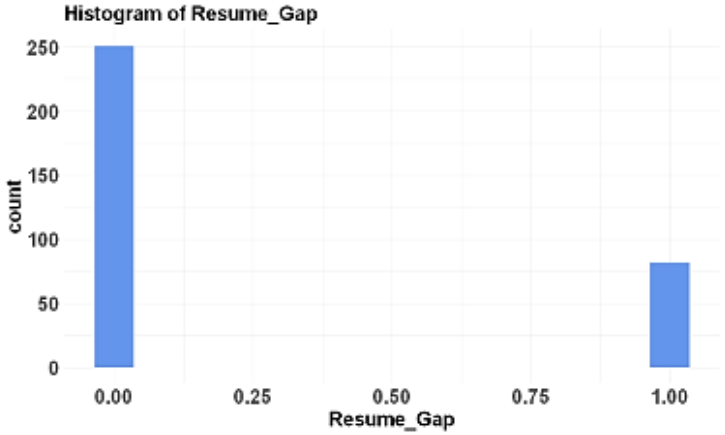

Histogram of Legally_Disabled
count
Legally_Disabled
0
100
200
300
0.00
0.25
0.50
0.75
1.00

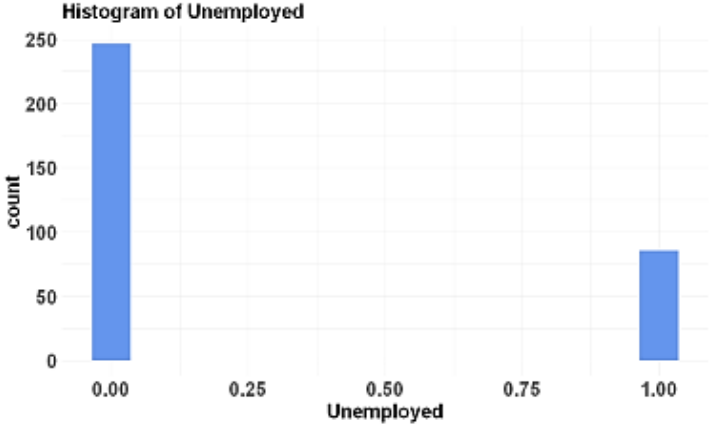

Histogram of Regular_Internet_Access
count
Regular_Internet_Access
0
100
200
300
0.00
0.25
0.50
0.75
1.00

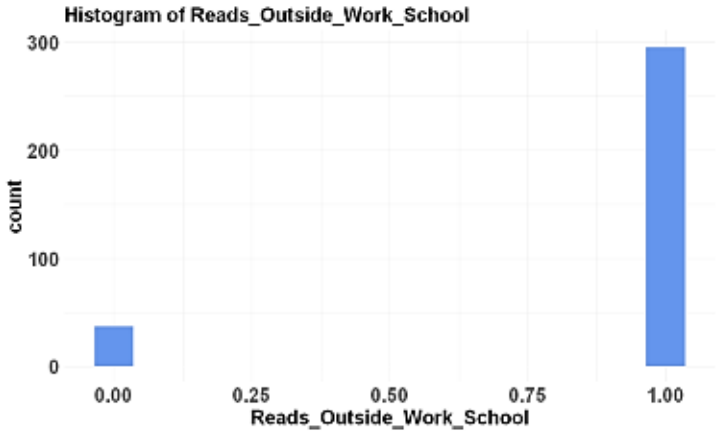

Histogram of Gap_Duration_Months
count
Gap_Duration_Months
0
50
100
150
200
250
0
25
50
75
100


Histogram of Lives_With_Parents
count
Lives_With_Parents
0
100
200
300
0.00
0.25
0.50
0.75
1.00

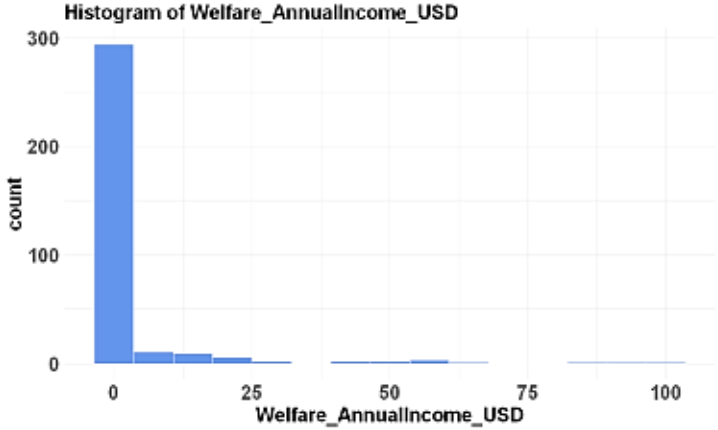

Histogram of Annual_Income_USD
count
Annual_Income_USD
0
10
20
30
40
50
0
25
50
75
100

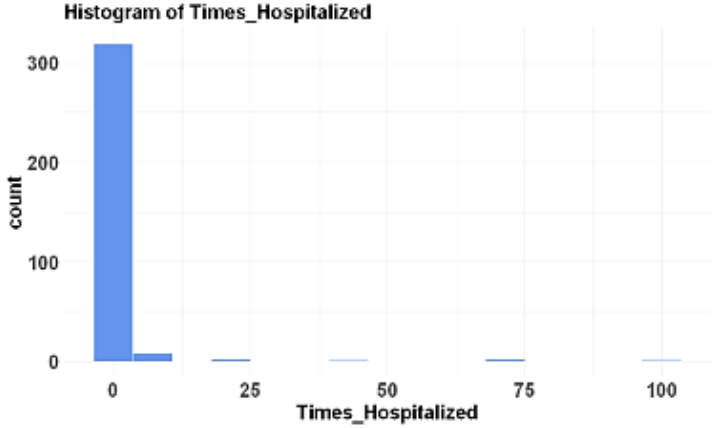

Histogram of Resume_Gap
count
Resume_Gap
0
50
100
150
200
250
0.00
0.25
0.50
0.75
1.00

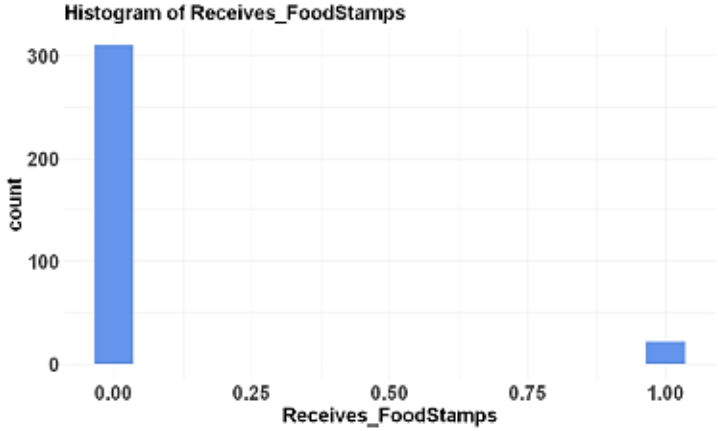

Histogram of Unemployed
count
Unemployed
0
50
100
150
200
250
0.00
0.25
0.50
0.75
1.00

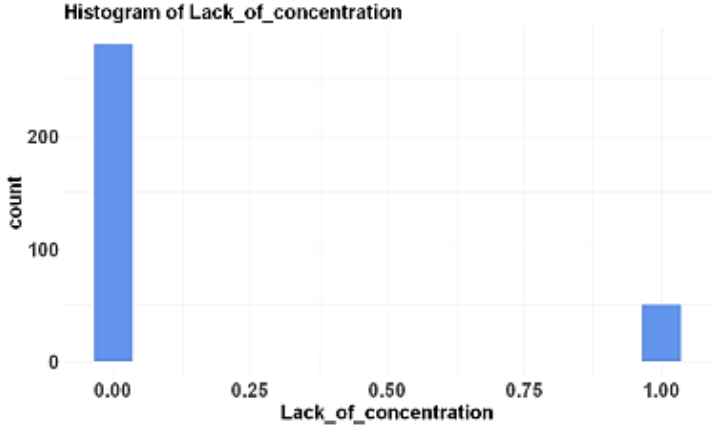

Histogram of Reads_Outside_Work_School
count
300
200
100
0
0.00
0.25
0.50
0.75
1.00
Reads_Outside_Work_School


Histogram of On_Section8_Housing
count
300
200
100
0
0.00
0.25
0.50
0.75
1.00
On_Section8_Housing


Histogram of Welfare_AnnualIncome_USD
count
300
200
100
0
0
25
50
75
100
Welfare_AnnualIncome_USD


Histogram of Times_Hospitalized
count
300
200
100
0
0
25
50
75
100
Times_Hospitalized


Histogram of Receives_FoodStamps
count
300
200
100
0
0.00
0.25
0.50
0.75
1.00
Receives_FoodStamps


Histogram of Lack_of_concentration
count
200
100
0
0.00
0.25
0.50
0.75
1.00
Lack_of_concentration

The most important insights for the purpose of my analysis are the following and these are the areas I will focus on for my future analysis:

- In the entire dataset, there are 334 cases.
- About 24% of the people surveyed report having a mental illness which is an interesting statistical piece given that this also reflects the general US population's percentage of mental illness according to research conducted by Johns Hopkins Medical Department in that 26% of Americans suffer from at least one mental illness disorder (Johns Hopkins Medicine, 2023).

- Anxiety and Depression have higher reports of about 30% and 26 % respectively. About 15% report having panic attacks and a lack of concentration. About 12-13% have mood swings and obsessive thinking. The least is about 3% observed in tiredness.

- 26% are unemployed.
- About 25% of respondents have resume gaps. On average, those gap durations are about 8.5 months.
- 6% receive food stamps and 2% are on Section 8 Housing.

- The majority (96%) have regular internet access and the majority (87%) own a computer besides their smartphone suggesting that access to technology isn't the problem causing unemployment.

Frequency Distribution for Categorical Features:

```
$Education

          Completed Phd                 Some Phd       Completed Masters
                     10                        8                      49
           Some Masters Completed Undergraduate      Some Undergraduate
                      0                      100                      81
     High School or GED         Some highschool
                     63                       10

$Gender

Female   Male
   175    158

$Age

 > 60 45-60 30-44 18-29
   80    99   103    51

$Household_Income

        $200,000+ $175,000-$199,999 $150,000-$174,999 $125,000-$149,999
               20                 2                14                16
$100,000-$124,999   $75,000-$99,999   $50,000-$74,999   $25,000-$49,999
               24                33                57                68
  $10,000-$24,999         $0-$9,999
               34                27

$Region

           Pacific           Mountain West North Central West South Central
                45                 32                13                32
East North Central East South Central     South Atlantic    Middle Atlantic
                50                 19                63                56
       New England
                21

$Device_Type

  Android Phone / Tablet       iOS Phone / Tablet Windows Desktop / Laptop
                      92                       93                      122
  MacOS Desktop / Laptop                    Other
                      24                        2
```

- Education:

The most common level of education among respondents is 'Completed Undergraduate', followed by 'Some Undergraduate'.
Few respondents have completed a PhD.

- Gender:

The dataset is fairly balanced in terms of gender, with slightly more female respondents than males.

- Age:

The age groups '30-44' and '45-60' are the most represented.
There are fewer respondents over 60 and between '18-29'.

- Household Income:

The most common income range is '25K-50K', followed by '50K-75K'.
Very few respondents have a household income of over 200K or under 10K.

- Region:

Respondents are distributed across various regions, with the 'Middle Atlantic' and 'West North Central' being the most represented.

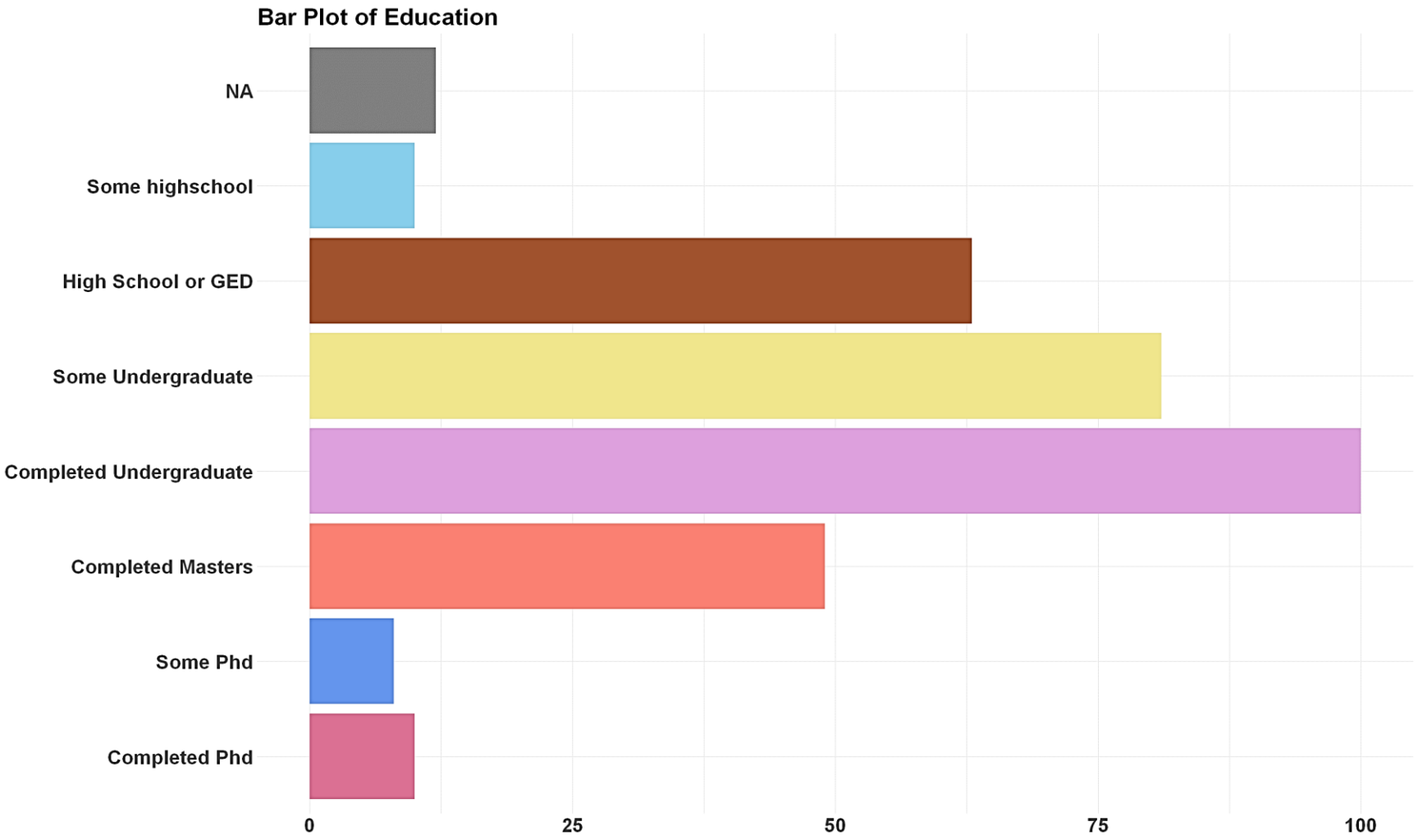
'East South Central' has the fewest respondents.

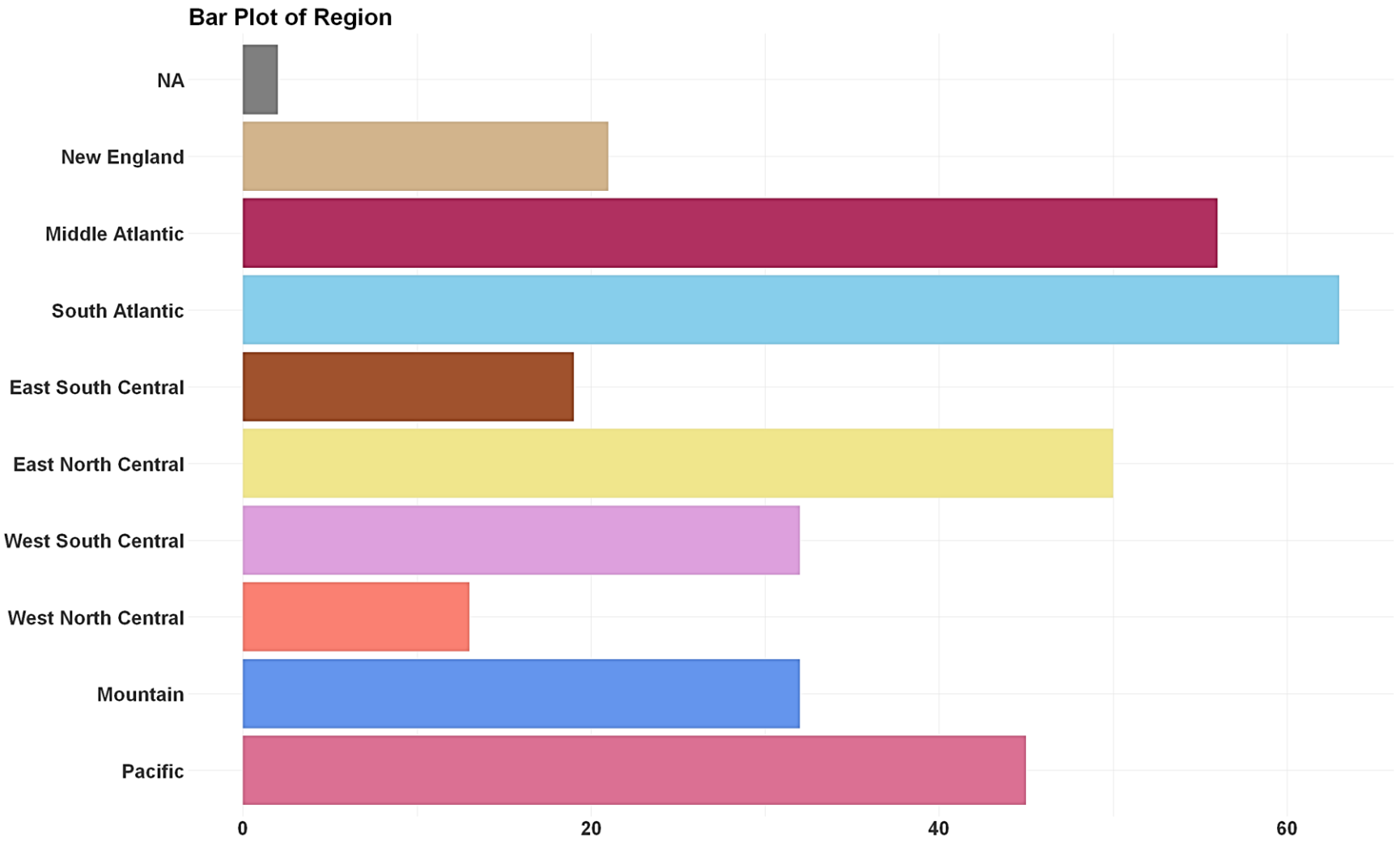
- Device Type:

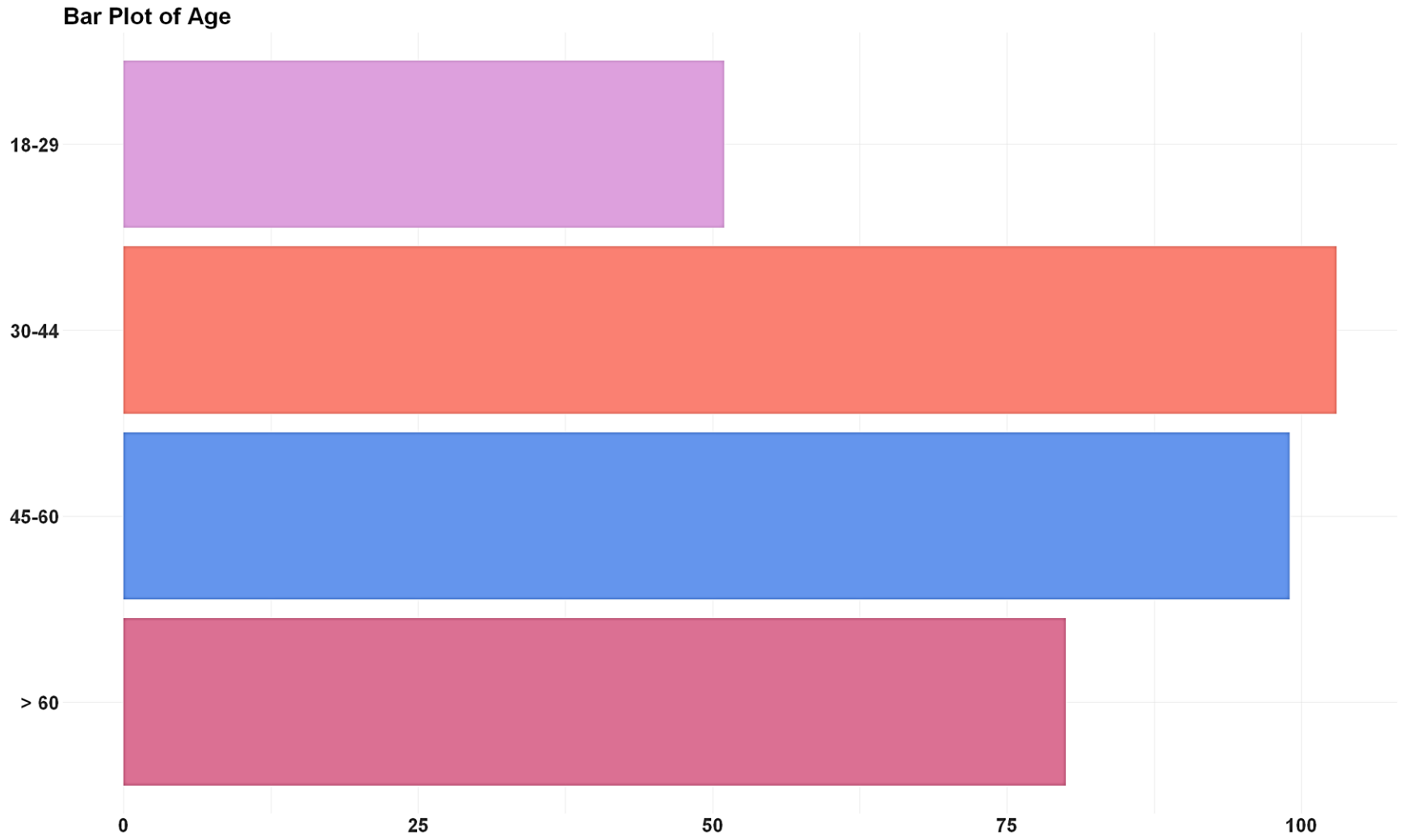
A large number of respondents use 'Android Phone/Tablet' and 'iOS Phone/Tablet', indicating mobile device prevalence.

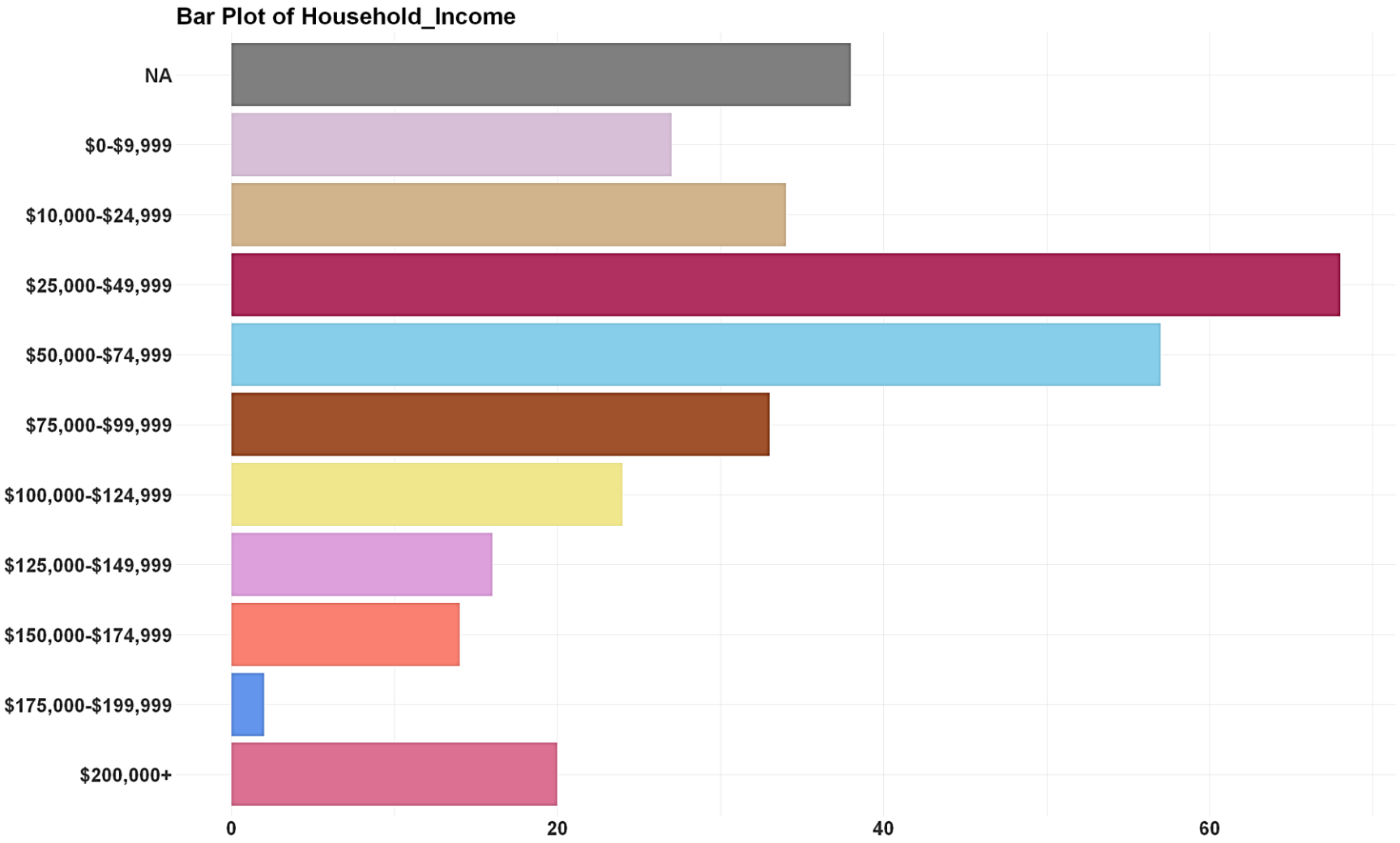
'Windows Desktop/Laptop' is also common, while 'MacOS Desktop/Laptop' and 'Other' devices are less used.

The following pages will include a few of the categorical distribution bar plots that will prove to be the most important for the purpose of my analysis.


Bar Plot of Education
NA
Some highschool
High School or GED
Some Undergraduate
Completed Undergraduate
Completed Masters
Some Phd
Completed Phd
0
25
50
75
100

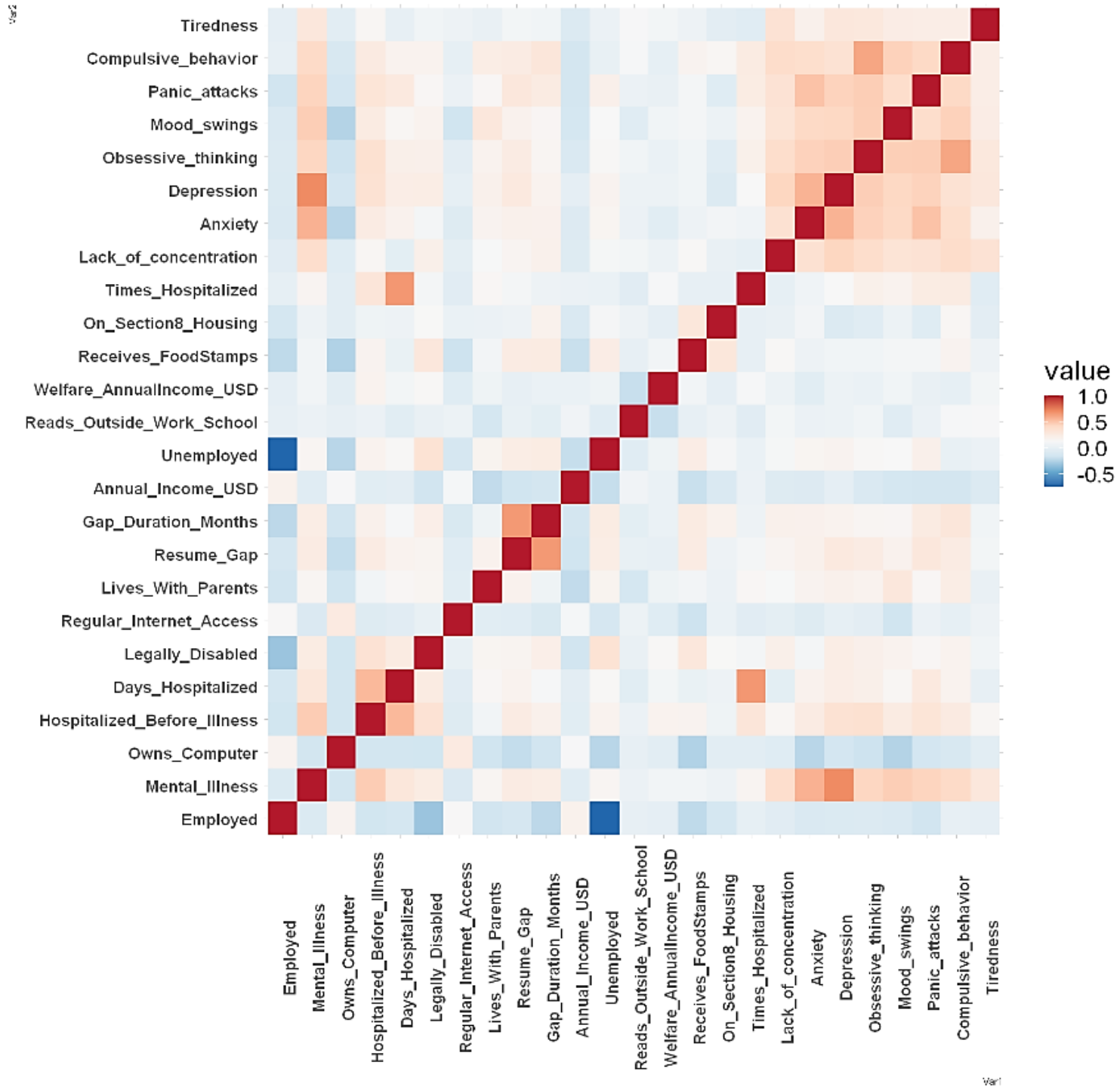

Bar Plot of Region
NA
New England
Middle Atlantic
South Atlantic
East South Central
East North Central
West South Central
West North Central
Mountain
Pacific
0
20
40
60


Bar Plot of Age
18-29
30-44
45-60
> 60
0
25
50
75
100


Bar Plot of Household_Income
NA
$0-$9,999
$10,000-$24,999
$25,000-$49,999
$50,000-$74,999
$75,000-$99,999
$100,000-$124,999
$125,000-$149,999
$150,000-$174,999
$175,000-$199,999
$200,000+
0
20
40
60

I proceeded to generate a correlation heatmap. I found this to be such an important step in my analysis as it helped me focus on certain variables that I can clearly see a correlation between from the beginning before I dive deeper into the analysis process and thus, give them more attention throughout the rest of my analysis.

- From the correlation heatmap on the following page, variables related to mental health conditions such as tiredness, compulsive behavior, panic attacks, mood swings, obsessive thinking, Depression, Anxiety, and even legal disability are highly related to one another and with mental illness. This suggests that these conditions might co-occur or have common underlying factors.
- Economic variables like being on Section 8 housing and receiving food stamps and welfare income are all related to unemployment and are positively correlated, indicating a link between employment status and the need for financial assistance programs.
- Mental illness is also correlated to having a resume gap (periods of unemployment reported on their resume) and the gap duration variable.
- There is a noticeable negative correlation between mental illness and employment, suggesting that those with a mental illness diagnosis are less likely to be employed.


Tiredness
Compulsive_behavior
Panic_attacks
Mood_swings
Obsessive_thinking
Depression
Anxiety
Lack_of_concentration
Times_Hospitalized
On_Section8_Housing
Receives_FoodStamps
Welfare_AnnualIncome_USD
Reads_Outside_Work_School
Unemployed
Annual_Income_USD
Gap_Duration_Months
Resume_Gap
Lives_With_Parents
Regular_Internet_Access
Legally_Disabled
Days_Hospitalized
Hospitalized_Before_Illness
Owns_Computer
Mental_Illness
Employed
Employed
Mental_Illness
Owns_Computer
Hospitalized_Before_Illness
Days_Hospitalized
Legally_Disabled
Regular_Internet_Access
Lives_With_Parents
Resume_Gap
Gap_Duration_Months
Annual_Income_USD
Unemployed
Reads_Outside_Work_School
Welfare_AnnualIncome_USD
Receives_FoodStamps
On_Section8_Housing
Times_Hospitalized
Lack_of_concentration
Anxiety
Depression
Obsessive_thinking
Mood_swings
Panic_attacks
Compulsive_behavior
Tiredness
value
1.0
0.5
0.0
-0.5

Group-Specific Summary:

The summary statistics below are grouped by employment status. My research question for this type of analysis is: How does employment status influence mental health and financial stability?

```
# A tibble: 2 × 31
  Unemployed Annual_Income_Mean Annual_Income_SD Annual_Income_Min
       <dbl>              <dbl>            <dbl>             <dbl>
1          0               41.5             30.4                 0
2          1               25.9             28.5                 0
  Annual_Income_Max Welfare_Income_Mean Welfare_Income_SD Welfare_Income_Min
              <dbl>               <dbl>             <dbl>              <dbl>
1               100                2.92              13.2                  0
2               100                4.52              10.4                  0
  Welfare_Income_Max ResumeGap_Duration_Months_Mean ResumeGap_Duration_Months_SD
               <dbl>                          <dbl>                         <dbl>
1                100                           5.18                          14.7
2                 50                          18.1                           30.5
  ResumeGap_Duration_Months_Min ResumeGap_Duration_Months_Max
                          <dbl>                         <dbl>
1                             0                           100
2                             0                           100
  Internet_Access_Mean Owns_Comp_Mean Times_Hospitalized_Mean
                 <dbl>          <dbl>                   <dbl>
1                0.976          0.915                   0.899
2                0.930          0.756                   2.06
  Times_Hospitalized_SD Times_Hospitalized_Min Times_Hospitalized_Max
                  <dbl>                  <dbl>                  <dbl>
1                  7.81                      0                    100
2                  8.98                      0                     69
  Depression_Mean Anxiety_Mean Tiredness_Mean Compulsivebehavior_Mean
            <dbl>        <dbl>          <dbl>                   <dbl>
1           0.211        0.251          0.251                  0.0810
2           0.384        0.430          0.430                  0.105
  PanicAttacks_Mean MoodSwings_Mean ObsThinking_Mean ConcentrationLack_Mean
              <dbl>           <dbl>            <dbl>                  <dbl>
1             0.101          0.0891           0.0972                  0.134
2             0.279          0.186            0.209                   0.209
  Disabled_Mean Mental_Illness_Mean Receives_FoodStamps_Mean
          <dbl>               <dbl>                    <dbl>
1        0.0364               0.202                   0.0283
2        0.267                0.337                   0.174
  Section8_Housing_Mean
                  <dbl>
1                0.0121
2                0.0465
```

The summary suggests that:

- Those employed have an annual income of about $41,000 (SD=30.4). Their welfare mean income is about $3,000.
- Those unemployed have an annual income of about $26,000 (SD = 28.5). Their welfare mean income is about $4,500.

- Standard Deviation (SD) is a measure of the amount of variation or dispersion in a set of values.
- The slightly higher SD in the employed group suggests a greater diversity in income levels compared to the unemployed group. This is expected as employed individuals can have a wide range of salaries depending on their jobs.

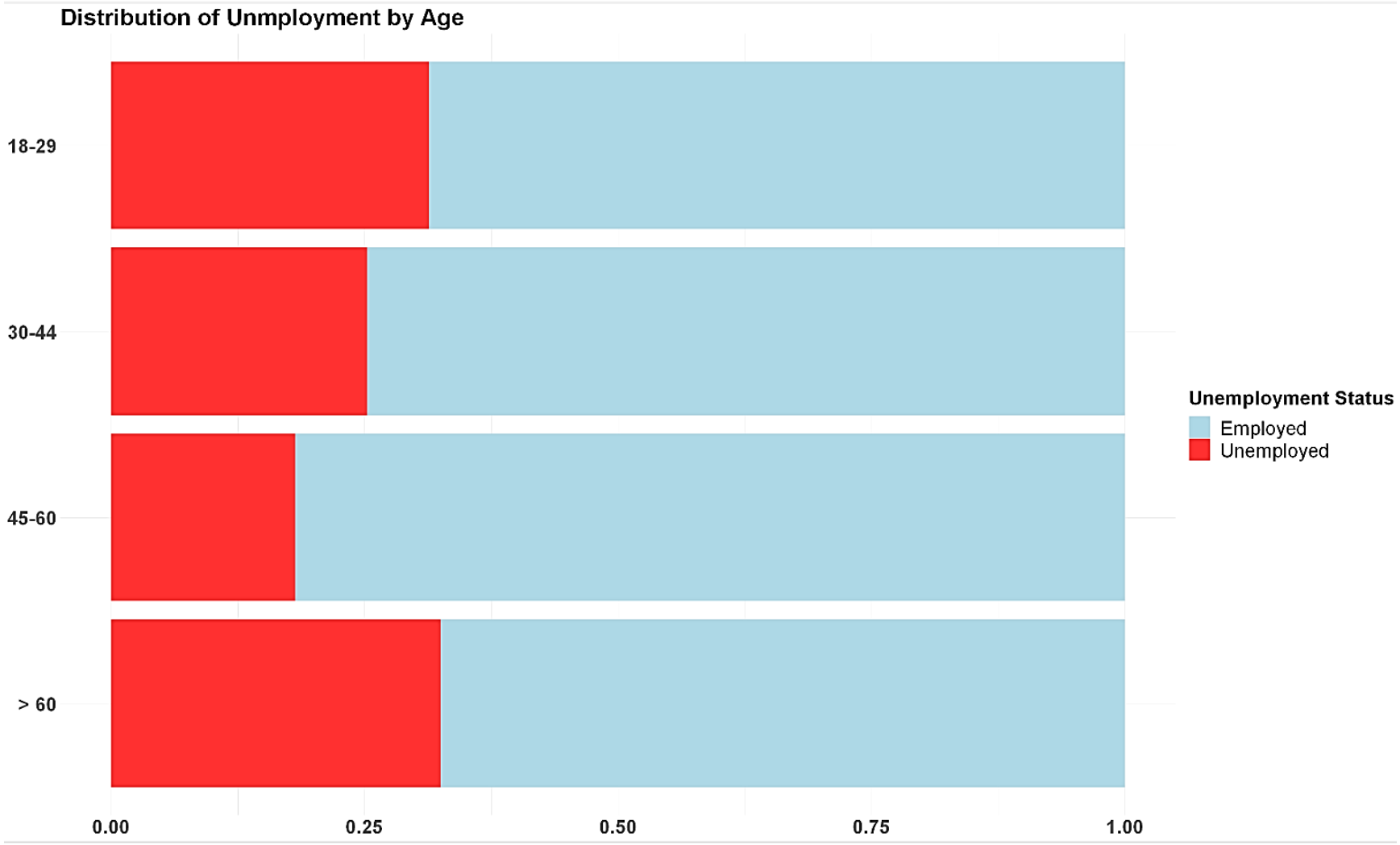
- Of those employed, the average resume gap duration in months is: 5.18, while for those unemployed it is: 18.1

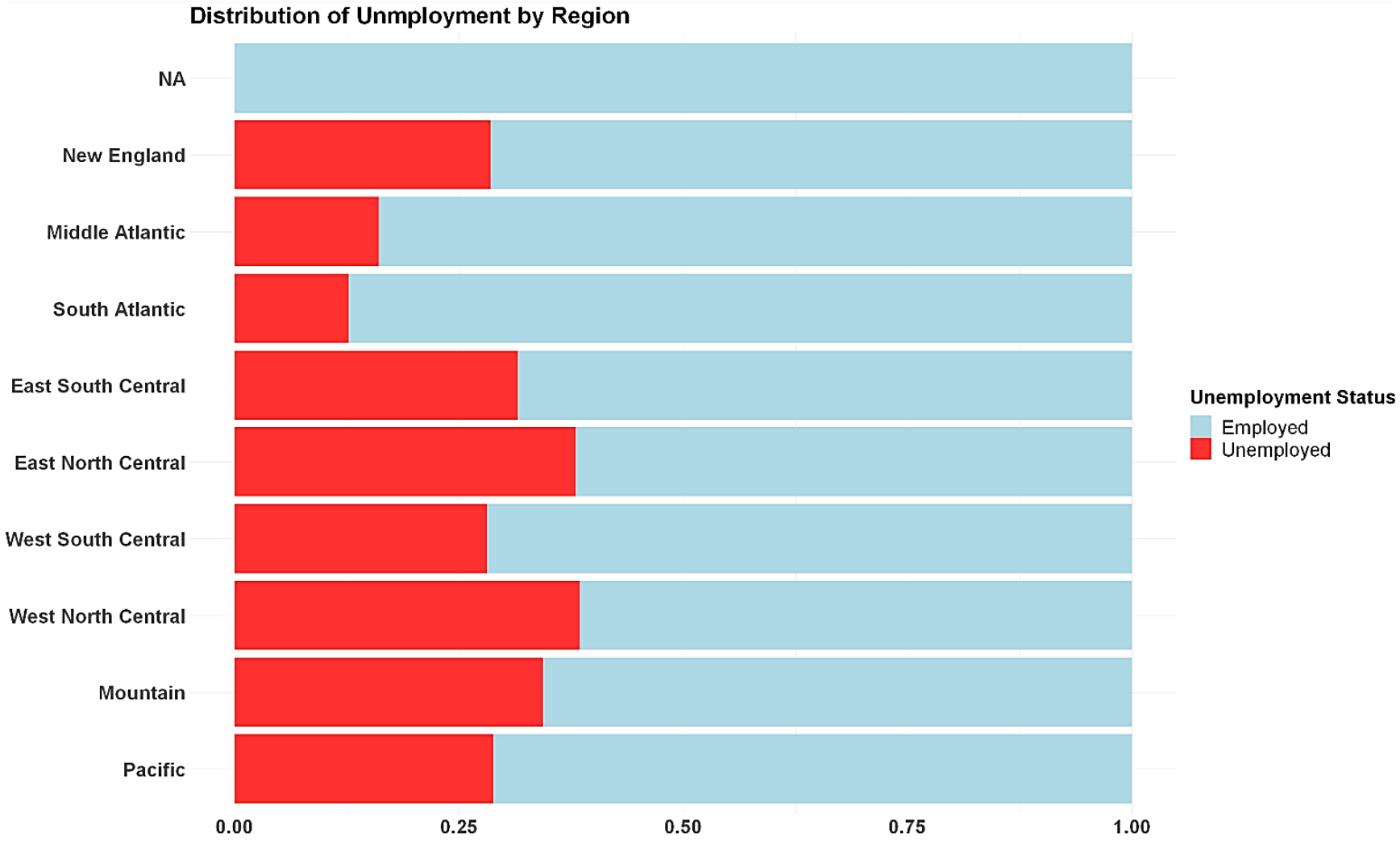
- In the entire dataset, 24% reported having mental illness.
- Taking the unemployed subgroup, 33% reported having a mental illness which is less than half and the top mental disorders were:
- Anxiety and Tiredness (43%) following that is Depression (38%) and Panic Attacks (27%)

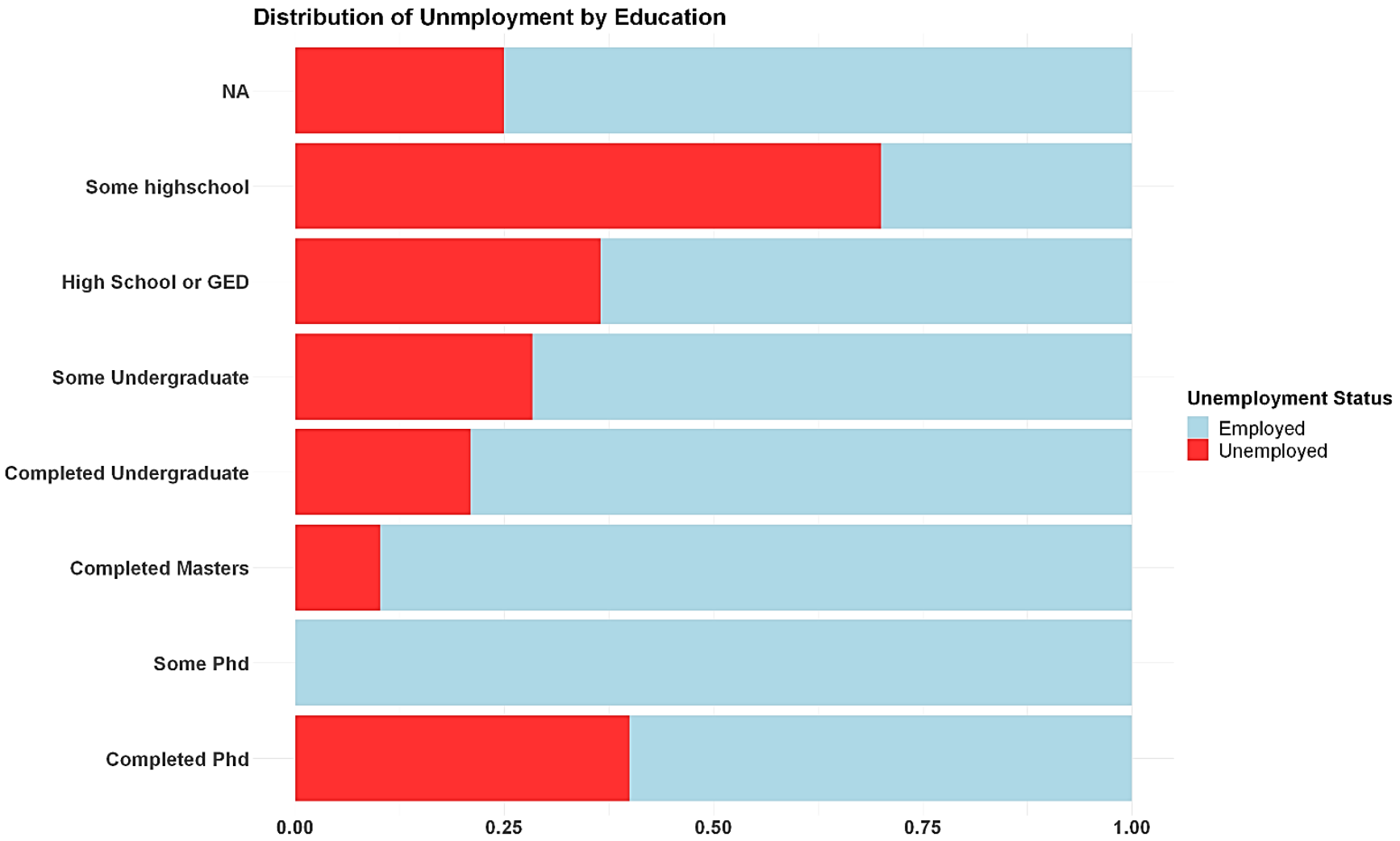
- 17% of the unemployed group receive food stamps compared to 2% among employed.
- 4% of unemployed individuals are in Section 8 housing and 1% of employed are in Section 8 housing.

The following stacked bar graphs are also grouped by employment status to further show some categorical variables that didn't appear in the table above. The bar graphs suggest the following insights:

- Unemployment rates increase as an individual tends to be younger and increase again once they're older than 60 possibly suggesting retirement.
- East North Central and West North Central regions show the highest unemployment rates and the Atlantic regions show the lowest.
- Unemployment rates decrease with a higher education level attained which is expected.


Distribution of Unmployment by Age
18-29
30-44
45-60
> 60
0.00
0.25
0.50
0.75
1.00
Unemployment Status
Employed
Unemployed


Distribution of Unmployment by Region
NA
New England
Middle Atlantic
South Atlantic
East South Central
East North Central
West South Central
West North Central
Mountain
Pacific
0.00
0.25
0.50
0.75
1.00
Unemployment Status
Employed
Unemployed


Distribution of Unmployment by Education
NA
Some highschool
High School or GED
Some Undergraduate
Completed Undergraduate
Completed Masters
Some Phd
Completed Phd
0.00
0.25
0.50
0.75
1.00
Unemployment Status
Employed
Unemployed

Based on the insights from the previous bar graphs, I formulated my first statistical model. The following is a logistic regression model, a statistical test that will provide me with valuable insights on the probability of a certain feature (dependent variable) being in a certain category or occurring based on other certain features (independent variables).

This mode predicts the likelihood of an individual being unemployed based on a couple of independent variables: the individual's region of residence, age, and education level.

In [61]:
```
logistic_model1 <- glm(Unemployed ~ Region + Age + Education,
                       family = "binomial", data = df)

summary(logistic_model1)
```

```
Call:
glm(formula = Unemployed ~ Region + Age + Education, family = "binomial",
    data = df)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   1.06588    0.83382   1.278  0.20114
RegionMountain                0.24616    0.55106   0.447  0.65509
RegionWest North Central     -0.07203    0.70656  -0.102  0.91880
RegionWest South Central     -0.23946    0.57145  -0.419  0.67519
RegionEast North Central      0.22609    0.49556   0.456  0.64823
RegionEast South Central     -0.17404    0.66441  -0.262  0.79337
RegionSouth Atlantic         -1.25568    0.56730  -2.213  0.02687 *
RegionMiddle Atlantic        -0.78115    0.54489  -1.434  0.15169
RegionNew England            -0.03291    0.63799  -0.052  0.95886
Age45-60                     -1.16710    0.41349  -2.823  0.00476 **
Age30-44                     -0.93562    0.40656  -2.301  0.02137 *
Age18-29                     -0.63909    0.45112  -1.417  0.15658
EducationSome Phd           -16.60138  811.95790  -0.020  0.98369
EducationCompleted Masters   -2.44433    0.85802  -2.849  0.00439 **
EducationCompleted Undergraduate -1.49688    0.74541  -2.008  0.04463 *
EducationSome Undergraduate  -0.98842    0.74796  -1.321  0.18634
EducationHigh School or GED  -0.52908    0.74484  -0.710  0.47750
EducationSome highschool      0.88121    0.99368   0.887  0.37518
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 365.74  on 318  degrees of freedom
Residual deviance: 315.58  on 301  degrees of freedom
  (14 observations deleted due to missingness)
AIC: 351.58

Number of Fisher Scoring iterations: 15
```

The model confirmed what has been observed in the previous bar graphs that there is less likelihood for individuals to be unemployed if:

- They live in the South Atlantic regions (Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, and the District of Columbia).
- They are middle-aged.
- They have at least an undergraduate or master's degree.

On the following page is a boxplot showing how income varies with region. One can conclude that there is some relationship between region of residence and income, as the median income and spread vary across regions. This suggests that certain regions are associated with higher or lower incomes.

There appears to be a variation in median income across regions. Some regions have higher median incomes, and others lower.

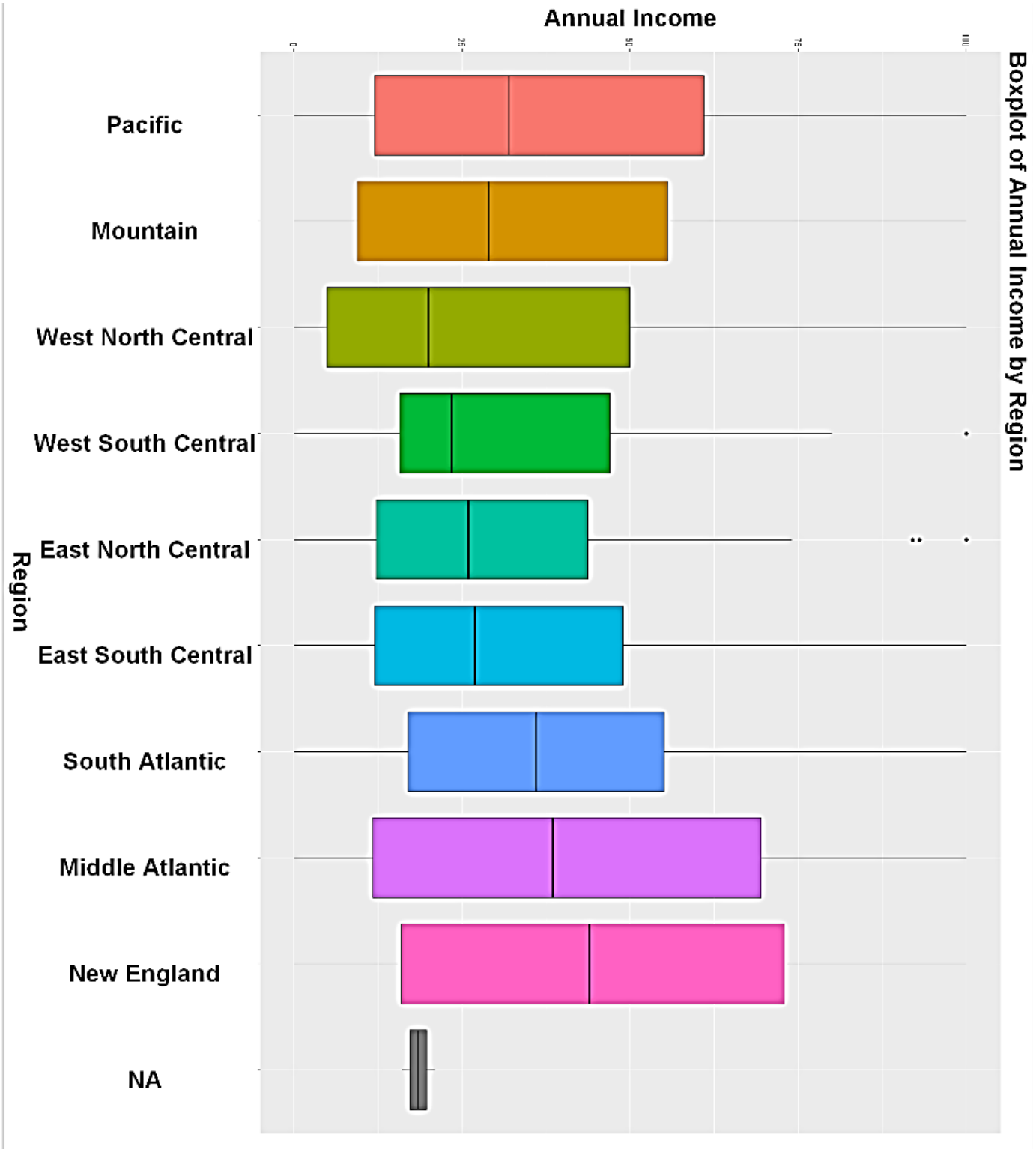
- Wealthier Regions:

Middle Atlantic and New England: These regions show higher median incomes compared to the others. Their interquartile range (IQR) is also higher, indicating a larger spread of income around the median.

South Atlantic: This region has a wide range of incomes as evidenced by the long box and whiskers, suggesting a high variance in income.

- Less Wealthy Regions:

West North Central and East South Central: These regions appear to have lower median incomes. The boxes are shorter, suggesting less variation in income.


Boxplot of Annual Income by Region
Annual Income
Region
Pacific
Mountain
West North Central
West South Central
East North Central
East South Central
South Atlantic
Middle Atlantic
New England
NA

Because Education & Income affect mental illness and employment rates as observed, I wanted to further analyze if the region of residence plays a significant role through performing a statistical test.

I chose to proceed with a chi-square test, which is a statistical analysis that tests if there's a significant association between two categorical (nominal/ordinal) variables.

This is useful for a number of reasons:

- Validity:

  Checking for statistical significance helps validate the observations from the data. Instead of relying on just visual or intuitive differences between groups, statistical tests give a measure of how likely these differences are due to actual associations versus occurring by chance.

- Decision Making:

  P-values and statistical tests assist in decision-making.

- Refining Focus:

  By understanding which variables are statistically significant, researchers can refine their focus, making subsequent analyses or experiments more targeted and efficient.

It has been observed that there are certain regions with higher Employment Status and Lower/Higher Mental Illness.

Let's see if that is caused by chance or because of certain factors such as: Education or Household Income.

- Research Question:

What is the relationship between the Region of Residence and Individual's Education Levels and Household Income?

- Directional Alternative Hypothesis (H1):

There is a relationship between both variables: Individuals in certain regions are more likely to have higher education levels compared to individuals in other regions. Individuals in certain regions are more likely to have higher household incomes compared to individuals in other regions.

- Null Hypothesis (H0):

The variables are independent; There is no relationship.

In [26]:
```
# Chi-Square Test for Region and Education
chisq_test1 <- chisq.test(df$Region, df$Education)
print(chisq_test1)

# Chi-Square Test for Region and Household Income
chisq_test2 <- chisq.test(df$Region, df$Household_Income)
print(chisq_test2)
```

```
        Pearson's Chi-squared test

data:  df$Region and df$Education
X-squared = 36.326, df = 48, p-value = 0.8916


        Pearson's Chi-squared test

data:  df$Region and df$Household_Income
X-squared = 70.186, df = 72, p-value = 0.5385
```

- Region & Education:

The p-value is 0.9039, which is much higher than the common alpha level of 0.05. This high p-value suggests that there is no statistically significant association between Region and Education levels in the dataset.

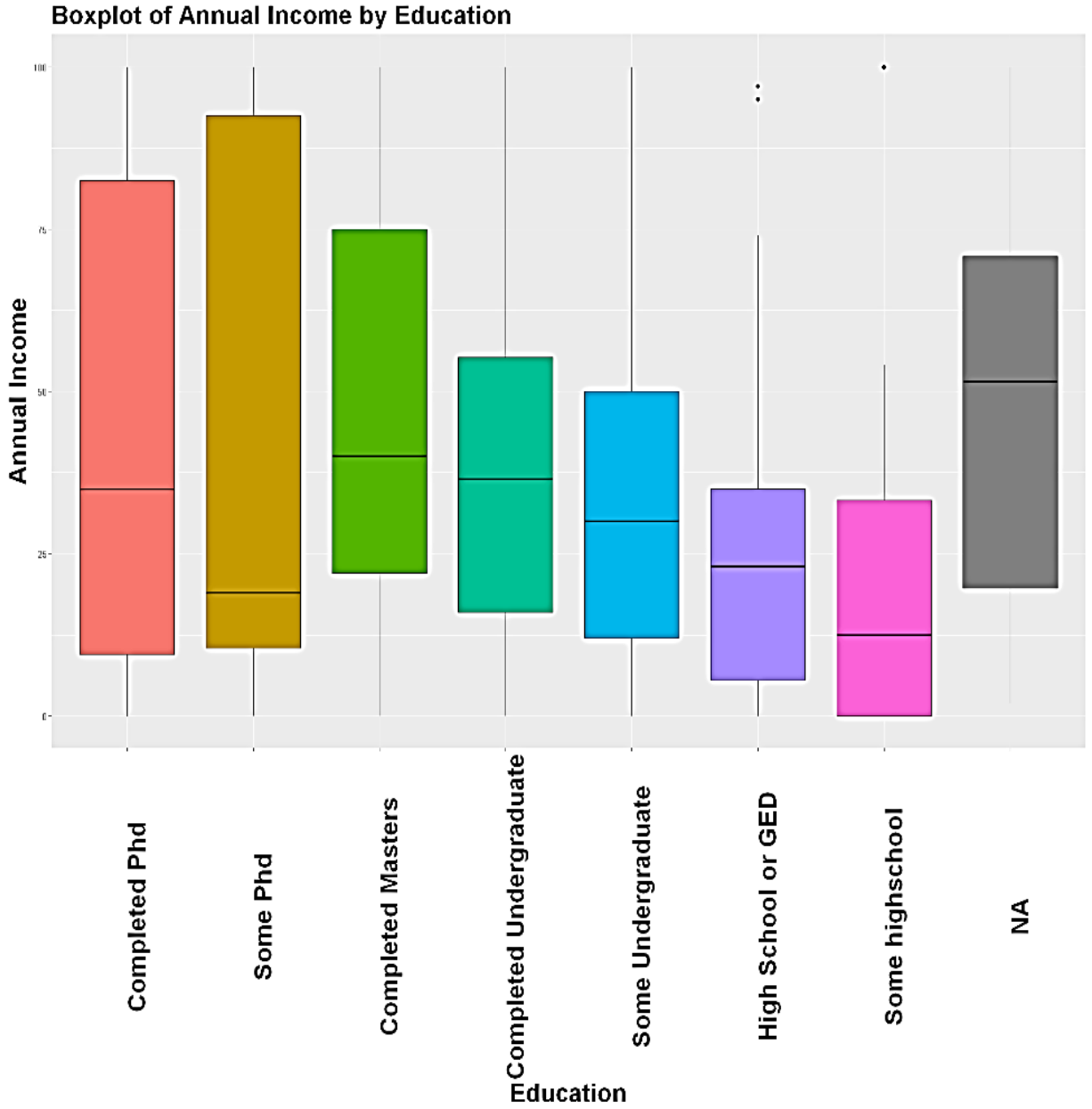
We fail to reject the null hypothesis, which states that the two variables are independent. This implies that, based on the data, the region where individuals live does not significantly affect their education level, or vice versa.

- Region & Household Income:

The p-value here is 0.544, also well above the typical alpha level of 0.05. This indicates a lack of statistically significant association between the Region and Household Income categories in the dataset.

We also fail to reject the null hypothesis, suggesting that the region of residence is not significantly associated with the household income levels of individuals in the sample.

Next, I proceeded to generate the same type of staitsical test after visualizing how income levels vary with both age and education as follows. Although it is highly expected for income to vary with such factors, proving it significantly helps strenghten my project objective.




Boxplot of Annual Income by Age
Annual Income
100
75
50
25
0
> 60
45-60
30-44
18-29
Age

- Research Question:

Is there an association between household income with education level and age in the population?

- Null Hypothesis (H0):

There is no association.

- Directional Alternative Hypothesis (H1):

There is an association between household income and education level. The distribution of education levels and ages vary across different household income brackets.

In [28]:
```
# Chi-Square Test for Education and Household Income
chisq_test3 <- chisq.test(df$Household_Income, df$Education)
print(chisq_test3)

# Chi-Square Test for Age and Household Income
chisq_test4 <- chisq.test(df$Age, df$Household_Income)
print(chisq_test4)
```

```
        Pearson's Chi-squared test

data:  df$Household_Income and df$Education
X-squared = 126.27, df = 54, p-value = 1.011e-07


        Pearson's Chi-squared test

data:  df$Age and df$Household_Income
X-squared = 47.348, df = 27, p-value = 0.009071
```

- Household Income & Education:

The p-value is very low approximately and is significantly lower than the conventional alpha level of 0.05. This suggests that there is a statistically significant association between Household Income and Education levels in the dataset.

We reject the null hypothesis of independence, implying that in the data set, an individual's education level is likely to be related to their household income. This could reflect the impact of educational attainment on earning potential.

- Household Income & Age:

The p-value is less than the standard threshold of 0.05. This indicates that the association between Age and Household Income is statistically significant.

We reject the null hypothesis in this case, it suggests that there is a statistically significant relationship between a person’s age and their household income. This could be indicative of various life cycle income patterns, such as increasing income with age due to career advancement or decreasing income post-retirement.

Next, to further understand what factors play the most significant role in causing and/or predicting unemployment, I proceeded to perform logistic regression statistical tests.

I will first explore unemployment with various mental illness factors paying most attention to those who appeared to have a positive correlation in the correlation heatmap discussed previously.

- Null hypothesis (H0):

Mental Illness does not have any association with the likelihood of being unemployed.

- Directional Alternative Hypothesis (H1):

Mental Illness or at least one type of mental illness disorder can influence unemployment likelihood.

In [55]:
```
logistic_model2 <- glm(Unemployed ~ Mental_Illness + Panic_attacks + Tiredness + Depression + Anxiety,
                       data = df, family = "binomial")
summary(logistic_model2)
```

```
Call:
glm(formula = Unemployed ~ Mental_Illness + Panic_attacks + Tiredness +
    Depression + Anxiety, family = "binomial", data = df)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.37425    0.17540  -7.835 4.69e-15 ***
Mental_Illness1 -0.09203    0.42645  -0.216    0.829
Panic_attacks    0.90022    0.39883   2.257    0.024 *
Tiredness       -0.07715    0.29911  -0.258    0.796
Depression       0.42342    0.42681   0.992    0.321
Anxiety          0.25294    0.36627   0.691    0.490
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 380.44  on 332  degrees of freedom
Residual deviance: 363.38  on 327  degrees of freedom
AIC: 375.38

Number of Fisher Scoring iterations: 4
```

From the model, the coefficients represent log odds in the context of a logistic regression model, which means they show the log odds of being unemployed for a one-unit increase in the predictor variable, holding all other variables constant.

- Intercept: The model's intercept is statistically significant, indicating that when all other variables are at their reference levels (which is typically the absence or baseline level of the conditions), the log odds of being unemployed is -1.38952.
- Panic Attacks: The coefficient for panic attacks is positive and significant ($p = 0.0217$), indicating that individuals who experience panic attacks have a higher likelihood of being unemployed. The effect size is moderate, with an odds ratio of exp(0.98045), indicating that the odds of unemployment are about 2.67 times higher for those with panic attacks compared to those without.

Therefore, we reject the null hypothesis because at least on factor of mental illness significantly predicted unemployment.

In the next model, I will test unemployment against factors like resume gap presence and resume gap durations to see if they significantly affect someone’s chance at employment.

```
In [45]: logistic_model3 <- glm(Unemployed ~ Resume_Gap + Gap_Duration_Months,
                                data = df, family = "binomial")
         summary(logistic_model3)

         Call:
         glm(formula = Unemployed ~ Resume_Gap + Gap_Duration_Months,
             family = "binomial", data = df)

         Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
         (Intercept)         -1.432599   0.159523  -8.980   <2e-16 ***
         Resume_Gap1          0.689148   0.360080   1.914   0.0556 .
         Gap_Duration_Months  0.017155   0.007577   2.264   0.0236 *
         ---
         Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         (Dispersion parameter for binomial family taken to be 1)

             Null deviance: 380.44  on 332  degrees of freedom
         Residual deviance: 355.16  on 330  degrees of freedom
         AIC: 361.16

         Number of Fisher Scoring iterations: 4
```

From the model above, the following insights were found meaningful:

- Resume Gap: The presence of a resume gap (Resume_Gap1) has a positive coefficient, indicating that individuals with a gap in their resume are more likely to be unemployed. However, the p-value is just above the conventional threshold for significance ($p = 0.0556$), suggesting a marginal effect.

- Gap Duration: Each additional month in the duration of the resume gap (Gap_Duration_Months) increases the likelihood of being unemployed. The coefficient is statistically significant ($p = 0.0236$), which means that as the length of the employment gap increases, so does the probability of being unemployed.
- Model Significance: The intercept is significant, meaning that for individuals with no resume gap and a gap duration of zero, the log odds of being unemployed is -1.432599. Given that the intercept is significant, the relationship between the variables and unemployment is not purely due to chance.
- Effect Size and Interpretation: For Resume_Gap1, the odds ratio is exp(0.689148), which suggests that having a resume gap almost doubles the odds of unemployment (1.99 times more likely). For Gap_Duration_Months, the odds ratio is exp(0.017155), indicating that for each additional month of gap duration, the odds of being unemployed increase by about 1.7%.
- The model suggests that both having a resume gap and the length of that gap are relevant factors in unemployment. This could have implications for job seekers and those supporting them, highlighting the importance of minimizing employment gaps and returning to work swiftly.

Now that both mental illness and resume gaps each have an individual significant effect on unemployment. I proceeded to perform models that tested their combination against employment. This will predict how likely an individual is to be unemployed if they have both a mental illness and a resume gap.

In [57]:
```
logistic_model5 <- glm(Unemployed ~ Resume_Gap * Depression,
                       family = "binomial", data = df)

summary(logistic_model5)
```

```
Call:
glm(formula = Unemployed ~ Resume_Gap * Depression, family = "binomial",
    data = df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.40487    0.17645  -7.962  1.7e-15 ***
Resume_Gap1            0.50408    0.37324   1.351   0.1768
Depression            -0.06147    0.40974  -0.150   0.8808
Resume_Gap1:Depression  1.57536    0.62821   2.508   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 380.44  on 332  degrees of freedom
Residual deviance: 349.89  on 329  degrees of freedom
AIC: 357.89

Number of Fisher Scoring iterations: 4
```

In [54]:
```
logistic_model4 <- glm(Unemployed ~ Resume_Gap * Anxiety,
                       family = "binomial", data = df)

summary(logistic_model4)
```

```
Call:
glm(formula = Unemployed ~ Resume_Gap * Anxiety, family = "binomial",
    data = df)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.4534     0.1851  -7.851 4.12e-15 ***
Resume_Gap1           0.5844     0.3788   1.543   0.1228
Anxiety               0.1472     0.3634   0.405   0.6854
Resume_Gap1:Anxiety   1.2609     0.5952   2.118   0.0342 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 380.44  on 332  degrees of freedom
Residual deviance: 351.11  on 329  degrees of freedom
AIC: 359.11

Number of Fisher Scoring iterations: 4
```

Interpretation: The key takeaway from this model is that the interaction between having a resume gap and experiencing anxiety and depression is a significant predictor of unemployment. While neither having a resume gap nor experiencing anxiety or depression alone significantly predicts unemployment, their combined effect does. This suggests that these two factors may interact in a way that increases the likelihood of being unemployed.

Legal disability also showed to have a high correlation on the heatmap. I proceeded to check if that relation is significant or due to chance.

In [45]:
```
logistic_model6 <- glm(Unemployed ~ Resume_Gap * Legally_Disabled,
                       family = "binomial", data = df)
summary(logistic_model6)
```

```
Call:
glm(formula = Unemployed ~ Resume_Gap * Legally_Disabled, family = "binomial",
    data = df)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.6672     0.1792  -9.301  < 2e-16 ***
Resume_GapYes                  1.1876     0.3072   3.865 0.000111 ***
Legally_Disabled               2.3603     0.5312   4.444 8.84e-06 ***
Resume_GapYes:Legally_Disabled -0.5815     0.8767  -0.663 0.507169
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 380.44  on 332  degrees of freedom
Residual deviance: 331.89  on 329  degrees of freedom
AIC: 339.89

Number of Fisher Scoring iterations: 4
```

The model shows that having a resume gap predicts unemployment and is highly significant. The same is seen with having a disability.

It is concerning to see that those suffering from any form of mental illness or disability have significantly fewer chances of being employed.

Therefore, the next part of the analysis would be how socioeconomic support systems correlate with employment status and mental health conditions.

The overall model below is significant but not any of the individual factors are in relation to mental illness.

```
In [40]: logistic_model5 <- glm(Mental_Illness ~ Receives_FoodStamps + Welfare_AnnualIncome_USD +
                                On_Section8_Housing, data = df, family = "binomial")
         summary(logistic_model5)

         Call:
         glm(formula = Mental_Illness ~ Receives_FoodStamps + Welfare_AnnualIncome_USD +
             On_Section8_Housing, family = "binomial", data = df)

         Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
         (Intercept)           -1.254386   0.140245  -8.944   <2e-16 ***
         Receives_FoodStampsYes  0.635964   0.489762   1.299    0.194
         Welfare_AnnualIncome_USD  0.010613   0.009182   1.156    0.248
         On_Section8_HousingYes  -0.084818   0.901689  -0.094    0.925
         ---
         Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         (Dispersion parameter for binomial family taken to be 1)

             Null deviance: 364.88  on 332  degrees of freedom
         Residual deviance: 361.71  on 329  degrees of freedom
         AIC: 369.71

         Number of Fisher Scoring iterations: 4
```

The second shown below also proved to be overall significant specifically in relation to receiving food stamps. This means that most of those who are unemployed receive food stamps suggesting that the government helps those with low income maintain good nutrition but not significantly in other aspects such as hoising and welfare social income.

In [44]:
```
logistic_model6 <- glm(Unemployed ~ Receives_FoodStamps + Welfare_AnnualIncome_USD +
                       On_Section8_Housing, data = df, family = "binomial")
summary(logistic_model6)
```

```
Call:
glm(formula = Unemployed ~ Receives_FoodStamps + Welfare_AnnualIncome_USD +
    On_Section8_Housing, family = "binomial", data = df)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.245962   0.140036  -8.897  < 2e-16 ***
Receives_FoodStampsYes    1.881257   0.496328   3.790  0.00015 ***
Welfare_AnnualIncome_USD  0.006762   0.009524   0.710  0.47771
On_Section8_HousingYes    0.486053   0.891471   0.545  0.58560
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 380.44  on 332  degrees of freedom
Residual deviance: 360.91  on 329  degrees of freedom
AIC: 368.91

Number of Fisher Scoring iterations: 4
```

IV. Final Takeaways & Interpretations

- Correlation Between Mental Health and Unemployment: The analysis reveals a strong correlation between mental health conditions like anxiety, depression, and panic attacks, and unemployment rates. This suggests that mental health challenges significantly impact an individual's ability to obtain and maintain employment.
- Socioeconomic Status and Mental Health: There is a clear link between lower socioeconomic status, as indicated by reliance on financial assistance programs like food stamps and Section 8 housing, and higher incidences of mental health issues. This highlights the cyclical nature of poverty and mental health challenges.
- Impact of Education on Employment and Mental Health: Higher levels of education correlate with lower unemployment rates and fewer mental health issues. This underscores the importance of educational attainment in securing stable employment and maintaining good mental health.
- Regional Variations: Certain regions, such as the East North Central and West North Central, showed higher unemployment rates. This regional disparity suggests that location-specific factors, possibly including economic opportunities and social support systems, play a significant role in employment status.

- Income, Education, and Age Connection: The analysis confirms a significant association between household income and education levels, as well as household income and age. This implies that as people age and attain higher education, their income potential increases, reducing their risk of unemployment and associated mental health issues.
- Resume Gaps and Unemployment: Having a resume gap and the duration of the gap are strongly associated with higher unemployment rates. This highlights the challenges faced by individuals with gaps in their employment history, potentially due to mental health issues.
- Interaction of Multiple Factors: The combined effect of having a resume gap and experiencing anxiety and depression is a significant predictor of unemployment. This indicates that these factors may interact in ways that intensify the challenges of securing employment.

V. References

Johns Hopkins Medicine. (2023). Mental Health Disorder Statistics. Retrieved from https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics

Kaggle. (n.d.). Unemployment and mental illness survey. Retrieved November 29, 2023, from https://www.kaggle.com/datasets/michaelacorley/unemployment-and-mental-illness-survey/data

Sunshine Behavioral Health. (2021, April 19). Unemployment and Mental Health: Resources to Managing Stress and Anxiety. Retrieved November 29, 2023, from https://sunshinebehavioralhealth.com/resources/unemployment-and-mental-health-resources/