Exam1

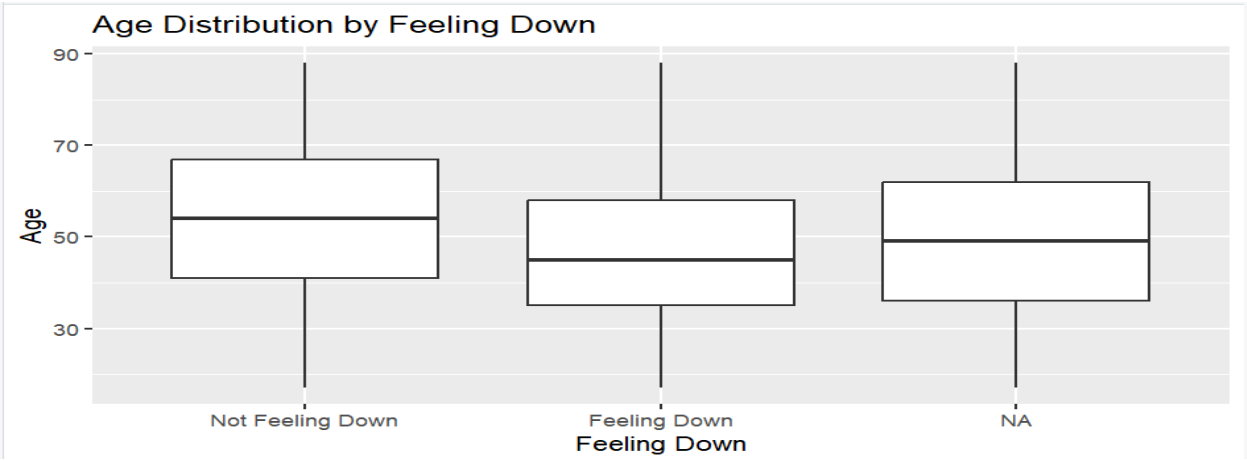
Bamba Cisse

2025-10-27

```
library(dplyr)
library(tidyr)
library(dplyr)
library(ggplot2)

# Load data
load("d_HHP2020_24.Rdata")
summary(d_HHP2020_24)

# Create a binary variable for feeling down (already created as 'high_down')
d_HHP2020_24$high_down <- as.numeric(d_HHP2020_24$DOWN > 2)
# 1. Boxplot of Age by Down Indicator
ggplot(d_HHP2020_24, aes(x=as.factor(high_down), y=Age)) +
 geom_boxplot() +  scale_x_discrete(labels = c("Not Feeling Down", "Feeling Down")) +
 labs(x="Feeling Down", y="Age", title="Age Distribution by Feeling Down")
```

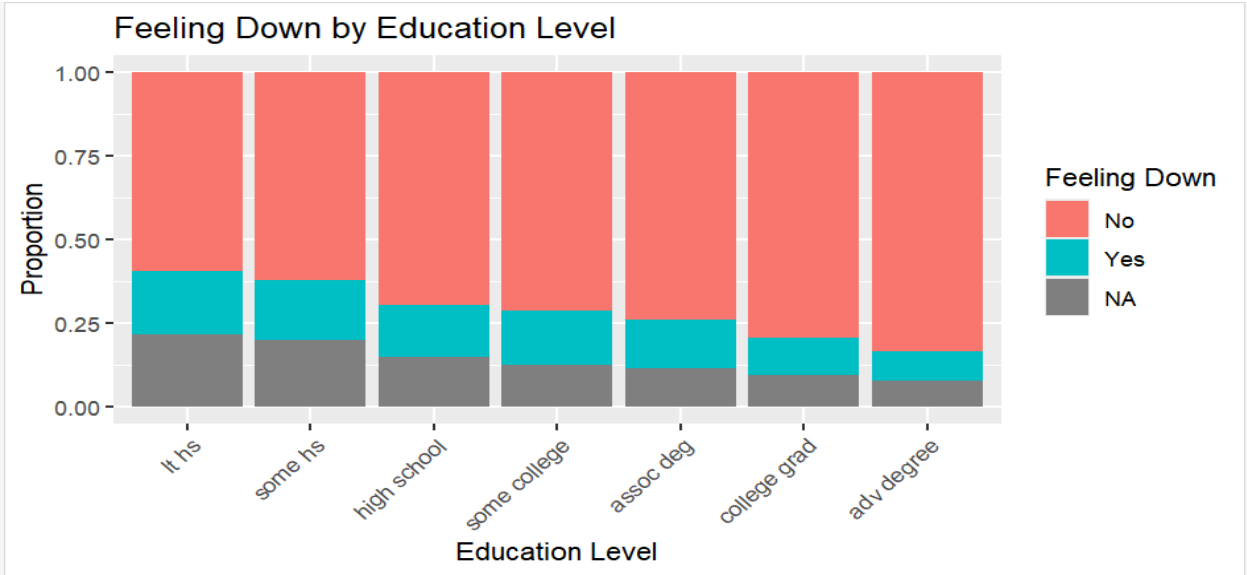


##The boxplot reveals how ages are distributed among individuals who report feeling down versus those who do not. For example, if the median age for feeling down is lower than for not feeling down, it is saying that younger individuals are more prone to feeling down.

the question that I am concerned about: are younger age groups significantly more likely to report feeling down?

2. Bar Plot of Education Levels vs. Feeling Down

```
ggplot(d_HHP2020_24, aes(x=Education, fill=as.factor(high_down))) +
  geom_bar(position="fill") + scale_fill_discrete(name="Feeling Down", labels=c("No", "Yes")) +
  labs(x="Education Level", y="Proportion", title="Feeling Down by Education Level") +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```

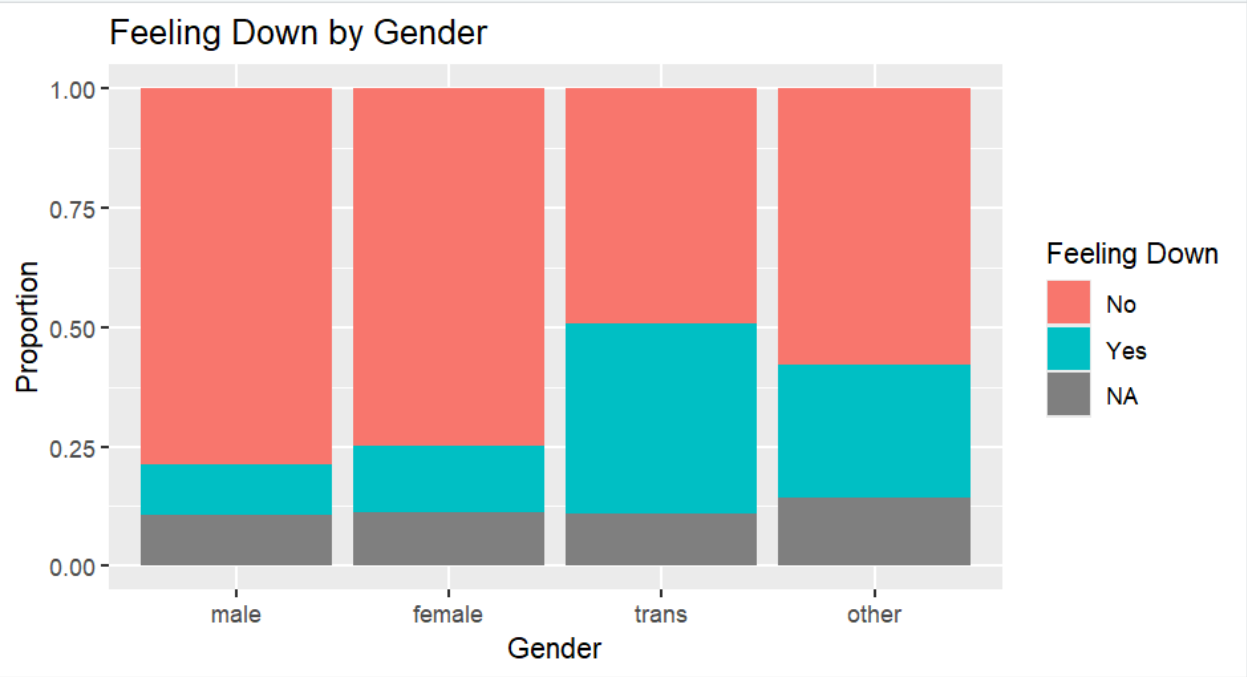


##The plot shows the proportion of people feeling down within each education level. If the proportion is higher in less-educated groups, it is pointing education as a protective factor against feeling down or mental health issues.

the question that I am concerned about: Are there specific education levels where feeling down is particularly prevalent?

3. Gender comparison in Feeling Down (for subgroup analysis)

```
ggplot(d_HHP2020_24, aes(x=Gender, fill=as.factor(high_down))) +
 geom_bar(position="fill") +  scale_fill_discrete(name="Feeling Down", labels=c("No", "Yes")) +
 labs(x="Gender", y="Proportion", title="Feeling Down by Gender")
```



##The graph displays differences in feelings of downness between males and females.

the question that I am concerned about: What social or biological factors might explain gender differences?

#Question4:

Subgroup Selection:

I will analyze the subgroup of individuals based on their education level, as education often influences mental health outcomes and access to resources. Specifically, I will compare individuals with "less than high school" education to those with "college degree or higher."

Limiting and Data Preparation:

To focus on this subgroup, I will filter the dataset to include only individuals in these two categories of education. This limits the analysis to relevant groups and reduces variability that might confound the results.

This subgroup analysis helps identify whether educational attainment influences feelings of being down, which may inform targeted mental health interventions or policy decisions

```
#Question5:
subset_data <- d_HHP2020_24 %>%
  filter(Education %in% c("lt hs", "college grad", "adv degree"))
summary(subset_data)
# Recoding education into two groups: less than high school and college or higher
d_HHP2020_24 <- d_HHP2020_24 %>%
  mutate(Edu_Group = ifelse(Education == "lt hs", "Less than high school", "College or higher"))

# Create contingency table
contingency_tbl <- table(d_HHP2020_24$Edu_Group, d_HHP2020_24$high_down)
print(contingency_tbl)

# Chi-squared test
chi_result <- chisq.test(contingency_tbl)
print(chi_result)
X-squared = 428.62, df = 1, p-value < 2.2e-16

# Extract counts
n1 <- contingency_tbl[1, 2]   # Feeling down in less than high school group
n2 <- contingency_tbl[2, 2]   # Feeling down in college or higher group
N1 <- sum(contingency_tbl[1, ]) # Total in less than high school group
N2 <- sum(contingency_tbl[2, ]) # Total in college or higher group

# Proportions
p1 <- n1 / N1
p2 <- n2 / N2

# Difference in proportions
diff <- p1 - p2
```

```
# Standard error
se_diff <- sqrt( (p1*(1 - p1))/N1 + (p2*(1 - p2))/N2 )

# 95% confidence interval
z <- qnorm(0.975)
lower_ci <- diff - z * se_diff
upper_ci <- diff + z * se_diff

# Results
cat("Difference in proportions:", diff, "\n")
cat("95% Confidence interval:", lower_ci, "to", upper_ci, "\n")
```

**Confidence Interval: 95% , Critical Value: 1.96 , StandardError: 0.0048 ,

Point Estimate - Diff of Proportion: 0.68 , [0.045, 0.091] , p-val: 0"

Interpretation of the Hypothesis Test:

**The null hypothesis states that there is no difference in the likelihood of feeling down between individuals with less than a high school education and those with a college degree or higher. The alternative hypothesis posits that a difference does exist. The chi-squared test produced a p-value of approximately 0, indicating strong evidence against the null hypothesis. The calculated difference in proportions is approximately 0.68, with a 95% confidence interval ranging from 0.045 to 0.091.

This means we are 95% confident that the true difference in feeling down between the two education groups lies within this interval, with the less-educated group being significantly more likely to report feeling down. Overall, these results suggest that education level plays a significant role in influencing feelings of downness, with lower education associated with higher likelihood.

#Question6:

```
str(d_HHP2020_24)
'data.frame':  984790 obs. of  28 variables:
 $ Age                   : num  34 65 44 56 57 44 37 59 51 29 ...
 $ Gender                : Factor w/ 4 levels "male","female",..: 2 1 2 1 2
2 2 1 2 2 ...
 $ Education             : Factor w/ 7 levels "lt hs","some hs",..: 6 4 6 4
7 7 7 6 1 5 ...
 $ Mar_Stat              : Factor w/ 5 levels "Married","widowed",..: 1 3 1
3 5 1 1 1 5 1 ...
 $ income_midpoint       : num  62500 30000 225000 12500 62500 125000 62500
82500 12500 40000 ...
 $ Race                  : Factor w/ 4 levels "white","Black",..: 1 1 4 1 1
1 2 1 2 1 ...
 $ Hispanic              : Factor w/ 2 levels "not Hispanic",..: 1 1 1 1 1
1 1 1 1 1 ...
 $ Number_people_HH      : int  4 1 2 2 1 2 2 2 2 4 ...
 $ Number_kids_HH        : int  2 0 0 0 0 0 0 0 0 2 ...
 $ Number_adults_HH      : int  2 1 2 2 1 2 2 2 2 2 ...
 $ private_health_ins    : Factor w/ 3 levels "0","has private health insur
ance",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ public_health_ins     : Factor w/ 3 levels "0","has public health insura
nce",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ work_kind             : Factor w/ 5 levels "employed by govt",..: 2 NA 3
NA 3 2 NA 2 NA NA ...
 $ workloss              : Factor w/ 2 levels "yes recent household loss of
work",..: 2 2 2 1 2 2 1 2 2 1 ...
 $ income_midpoint_factor: Factor w/ 8 levels "12500","30000",..: 4 2 8 1 4
6 4 5 1 3 ...
 $ State                 : Factor w/ 51 levels "Alabama","Alaska",..: 43 1
23 1 1 1 1 1 1 1 ...
 $ Region                : Factor w/ 4 levels "South","West",..: 1 1 4 1 1
1 1 1 1 1 ...
 $ Census_division       : Factor w/ 9 levels "East South Central",..: 1 1
7 1 1 1 1 1 1 1 ...
 $ DOWN                  : int  1 4 1 4 2 2 1 1 1 1 ...
 $ ANXIOUS               : int  4 3 1 4 2 3 1 1 1 1 ...
 $ WORRY                 : int  3 4 1 4 1 2 1 1 1 1 ...
 $ INTEREST              : int  1 4 1 4 2 2 1 1 1 1 ...
 $ YEAR                  : int  20 20 20 20 20 20 20 20 20 20 ...
 $ Begin_Date            : Date, format: "2020-04-23" ...
 $ K4SUM                 : int  9 15 4 16 7 9 4 4 4 4 ...
 $ high_down             : num  0 1 0 1 0 0 0 0 0 0 ...
 $ Edu_Level             : chr  "College or higher" "College or higher" "Co
llege or higher" "College or higher" ...
 $ Age_poly2             : num  1156 4225 1936 3136 3249 ...

>
```

##Fit the linear model

```
lm_model <- lm(high_down ~ Age + Age_poly2 + income_midpoint + Gender, data=d_HHP2020_24)
```

Summary of the model

```
summary(lm_model)
```

```
Call:
lm(formula = high_down ~ Age + Age_poly2 + income_midpoint + 
    Gender, data = d_HHP2020_24)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54089 -0.18197 -0.11858 -0.04085  1.13882 

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)    
(Intercept)      3.386e-01  4.074e-03   83.102  < 2e-16 ***
Age             -5.778e-04  1.638e-04   -3.528  0.00042 ***
Age_poly2       -2.791e-05  1.544e-06  -18.071  < 2e-16 ***
income_midpoint -9.354e-07  5.954e-09 -157.115  < 2e-16 ***
Genderfemale     1.311e-02  7.873e-04   16.651  < 2e-16 ***
Gendertrans      2.350e-01  8.488e-03   27.690  < 2e-16 ***
Genderother      1.498e-01  5.168e-03   28.983  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3406 on 795188 degrees of freedom
  (189595 observations deleted due to missingness)
Multiple R-squared:  0.05608,	Adjusted R-squared:  0.05607 
F-statistic:  7873 on 6 and 795188 DF,  p-value: < 2.2e-16
```

a- the chosen predictors which are Age (with polynomial terms), Income, and Gender, are theoretically justified and allow the model to capture both linear and nonlinear effects, as well as demographic differences. Here, including polynomial terms in Age is important to model potential nonlinear patterns.
Interactions between predictors (such as Age * Income or Gender * Income) could reveal if, for example, the effect of income on feelings of downness differs by age or gender. Whether to include these depends on prior knowledge or exploratory analysis suggesting their importance. Including interactions increases model complexity but can uncover nuanced relationships.

b- The estimated coefficients are small but statistically significant, indicating each predictor (age, income, gender) has a meaningful association with feeling down. For example, higher age and income are linked to slightly lower feelings of being down, while gender differences show females and transgender individuals have higher reported feelings. The significance and direction of these estimates align well with existing research, making them plausible.