Project Title:

Analysing the Role of Education, Gender, and Drinking Habit on Income using R.

Objective:

To explore how **education level**, **gender**, and **drinking habits** impact **income** using a synthetic dataset and statistical tools in R.

Tools & Libraries Used:

- **Language**: R
- **Libraries**: readxl, base R functions
- **Methods**: Descriptive statistics, Skewness, Kurtosis, Shapiro-Wilk, KS Test, t-tests, Boxplots, Confidence Intervals

1. Install and load the necessary package

```
# Install and load the package
install.packages("readxl")
library(readxl)
```

2. Load and view the data

```
# Load the data
data <- read_excel("C:/Users/Sanika/Downloads/Synthetic_Drinking_Dataset_500.xlsx")
# View data structure
head(data)
```

```
> head(data)
# A tibble: 6 x 4
  Gender Education   DrinkingHabit  Income
  <chr>  <chr>       <chr>           <dbl>
1 Male   High School Non-Drinker    29282.
2 Female High School Heavy Drinker  38521.
3 Male   PhD         Social Drinker 71589.
4 Male   High School Social Drinker 47408.
5 Male   High School Non-Drinker    29679.
6 Female Graduate    Social Drinker 43554.
```

summary(data)

```
> summary(data)
    Gender           Education         DrinkingHabit          Income     
 Length:500         Length:500         Length:500         Min.   : 4212  
 Class :character   Class :character   Class :character   1st Qu.:35133  
 Mode  :character   Mode  :character   Mode  :character   Median :45865  
                                                          Mean   :47672  
                                                          3rd Qu.:58934  
                                                          Max.   :96944  
```

str(data)

```
> str(data)
tibble [500 × 4] (S3: tbl_df/tbl/data.frame)
 $ Gender       : chr [1:500] "Male" "Female" "Male" "Male" ...
 $ Education    : chr [1:500] "High School" "High School" "PhD" "High School" ...
 $ DrinkingHabit: chr [1:500] "Non-Drinker" "Heavy Drinker" "Social Drinker" "Social Drinker" ...
 $ Income       : num [1:500] 29282 38521 71589 47408 29679 ...
```

3. Fix Incorrect Measurement Scale

Fixing Education as an ordered factor

data$Education <- factor(data$Education, levels = c("High School", "Graduate", "PhD"), ordered = TRUE)

data$Education

```
Levels: High School < Graduate < PhD
```

4. Compare Mean, Median, and 50th Percentile of Income

mean_income <- mean(data$Income)

median_income <- median(data$Income)

percentile_50 <- quantile(data$Income, 0.50)

cat("Mean Income: ", mean_income, "\n")

cat("Median Income: ", median_income, "\n")

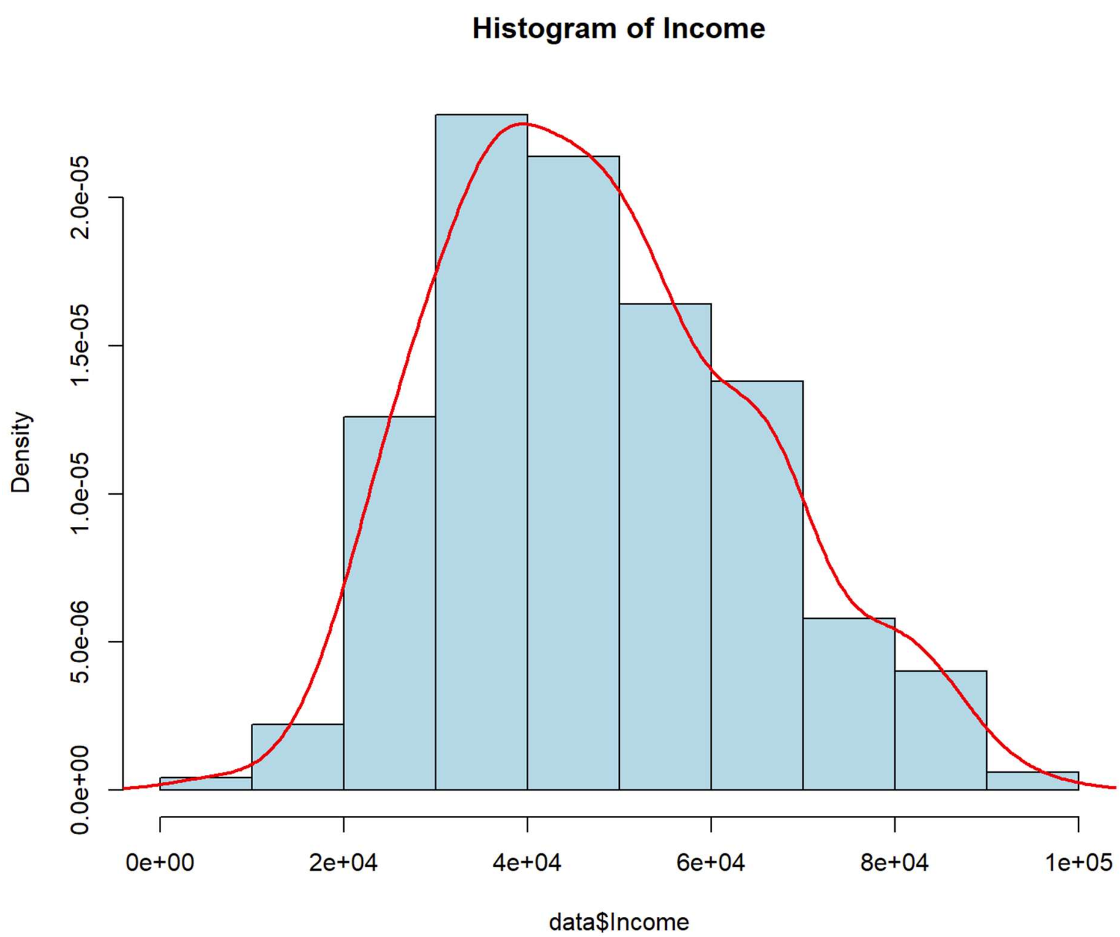
cat("50th Percentile: ", percentile_50, "\n")

```
> cat("Mean Income: ", mean_income, "\n")
Mean Income:  47672.12
> cat("Median Income: ", median_income, "\n")
Median Income:  45864.76
> cat("50th Percentile: ", percentile_50, "\n")
50th Percentile:  45864.76
```

5. Check for Normality of Income (Overall and by Education Level)

```r
# Histogram & Density
hist(data$Income, probability = TRUE, col = "lightblue", main = "Histogram of Income")
lines(density(data$Income), col = "red", lwd = 2)

# Skewness and Kurtosis
skewness_fn <- function(x) {
  m3 <- mean((x - mean(x))^3)
  s3 <- sd(x)^3
  m3 / s3
}

kurtosis_fn <- function(x) {
  m4 <- mean((x - mean(x))^4)
  s4 <- sd(x)^4
  m4 / s4
}

cat("Skewness:", skewness_fn(data$Income), "\n")
cat("Kurtosis:", kurtosis_fn(data$Income), "\n")

# Normality Tests
shapiro.test(data$Income)

# By education
phd_income <- data$Income[data$Education == "PhD"]
hs_income <- data$Income[data$Education == "High School"]

shapiro.test(phd_income)
shapiro.test(hs_income)
```

```
ks.test(phd_income, "pnorm", mean(phd_income), sd(phd_income))

ks.test(hs_income, "pnorm", mean(hs_income), sd(hs_income))
```



```
> cat("Skewness:", skewness_fn(data$Income), "\n")
Skewness: 0.3851676
> cat("Kurtosis:", kurtosis_fn(data$Income), "\n")
Kurtosis: 2.628446
> # Normality Tests
> shapiro.test(data$Income)

        Shapiro-Wilk normality test

data:  data$Income
W = 0.98307, p-value = 1.45e-05

> # By education
> phd_income <- data$Income[data$Education == "PhD"]
> hs_income <- data$Income[data$Education == "High School"]
> shapiro.test(phd_income)

        Shapiro-Wilk normality test

data:  phd_income
W = 0.98855, p-value = 0.5279

> shapiro.test(hs_income)

        Shapiro-Wilk normality test

data:  hs_income
W = 0.99119, p-value = 0.2488
```

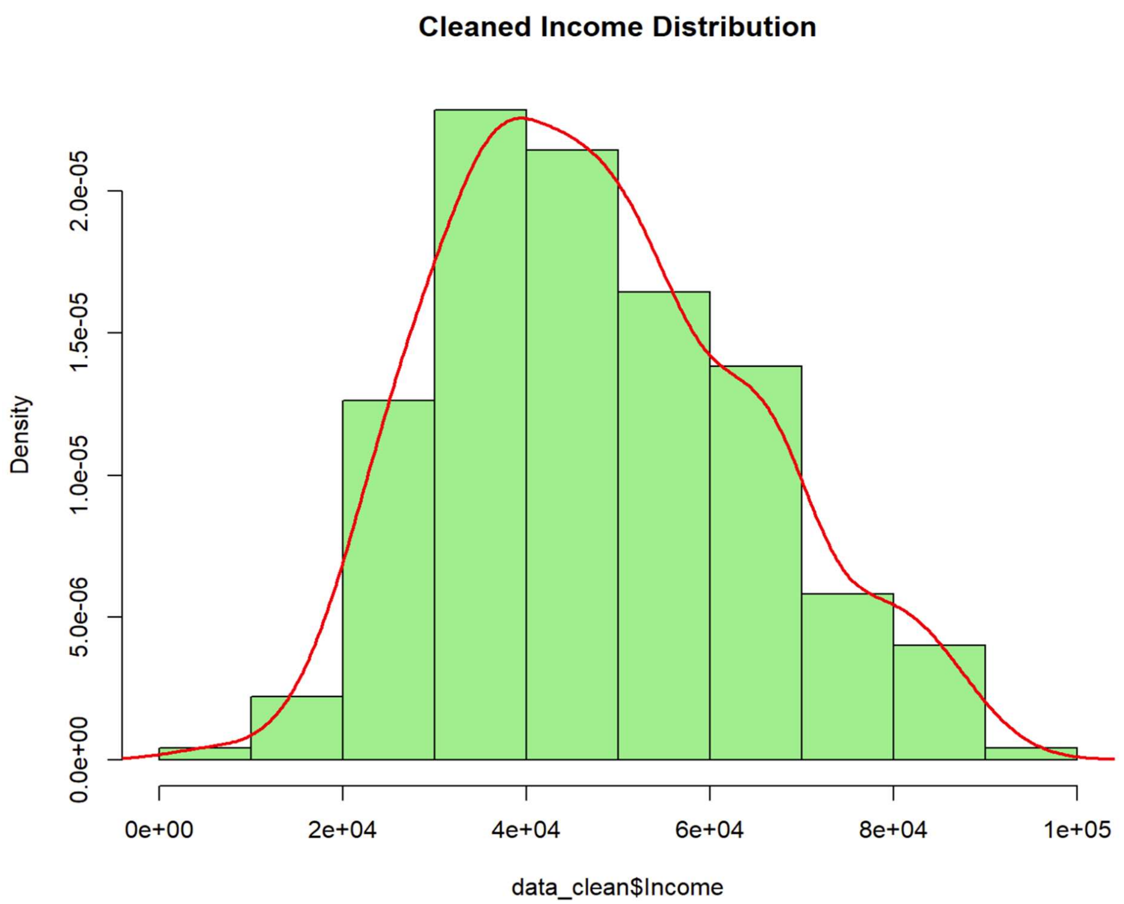
```
> hs_income <- data$Income[data$Education == "High School"]
> shapiro.test(phd_income)

        Shapiro-Wilk normality test

data:  phd_income
W = 0.98855, p-value = 0.5279

> shapiro.test(hs_income)

        Shapiro-Wilk normality test

data:  hs_income
W = 0.99119, p-value = 0.2488

> ks.test(phd_income, "pnorm", mean(phd_income), sd(phd_income))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  phd_income
D = 0.092587, p-value = 0.3404
alternative hypothesis: two-sided

> ks.test(hs_income, "pnorm", mean(hs_income), sd(hs_income))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  hs_income
D = 0.039952, p-value = 0.899
alternative hypothesis: two-sided
```

6. Remove Outliers and Retest for Normality

```
Q1 <- quantile(data$Income, 0.25)

Q3 <- quantile(data$Income, 0.75)

IQR_val <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR_val

upper_bound <- Q3 + 1.5 * IQR_val


# Remove outliers

data_clean <- subset(data, Income >= lower_bound & Income <= upper_bound)


# Retest normality

shapiro.test(data_clean$Income)
```

hist(data_clean$Income, probability = TRUE, col = "lightgreen", main = "Cleaned Income Distribution")

lines(density(data_clean$Income), col = "red", lwd = 2)




```
> shapiro.test(data_clean$Income)

        Shapiro-Wilk normality test

data:  data_clean$Income
W = 0.98301, p-value = 1.421e-05
```

7. Count Graduate Social Drinkers by Gender

```
a <- as.data.frame(data)

a <- a[a$Education == "Graduate" & a$DrinkingHabit == "Social Drinker", ]

table(data$Gender)

nrow(data)
```

```
Female   Male
   256    244
```

8. Calculate 95% Confidence Intervals for Social Drinkers vs Non-Drinkers

```
# For Social Drinkers

t.test(data$Income[data$DrinkingHabit == "Social Drinker"])


# For Non-Drinkers

t.test(data$Income[data$DrinkingHabit == "Non-Drinker"])


# Comparison between Social Drinkers and Non-Drinkers

t.test(Income ~ DrinkingHabit, data = subset(data, DrinkingHabit %in% c("Social Drinker", "Non-Drinker")))
```

```
> t.test(data$Income[data$DrinkingHabit == "Social Drinker"])

        One Sample t-test

data:  data$Income[data$DrinkingHabit == "Social Drinker"]
t = 43.238, df = 235, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 47975.11 52555.73
sample estimates:
mean of x
 50265.42

> # For Non-Drinkers
> t.test(data$Income[data$DrinkingHabit == "Non-Drinker"])

        One Sample t-test

data:  data$Income[data$DrinkingHabit == "Non-Drinker"]
t = 35.709, df = 156, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 44472.96 49681.16
sample estimates:
mean of x
 47077.06
```

```
> # Comparison between Social Drinkers and Non-Drinkers
> t.test(Income ~ DrinkingHabit, data = subset(data, DrinkingHabit %in% c("Social Drinker", "Non-Dri
nker")))

        Welch Two Sample t-test

data:  Income by DrinkingHabit
t = -1.8139, df = 351.75, p-value = 0.07054
alternative hypothesis: true difference in means between group Non-Drinker and group Social Drinker
is not equal to 0
95 percent confidence interval:
 -6645.277   268.553
sample estimates:
   mean in group Non-Drinker mean in group Social Drinker
                    47077.06                     50265.42
```

9. 95% Trimmed Mean for Males and Females who are Heavy Drinkers

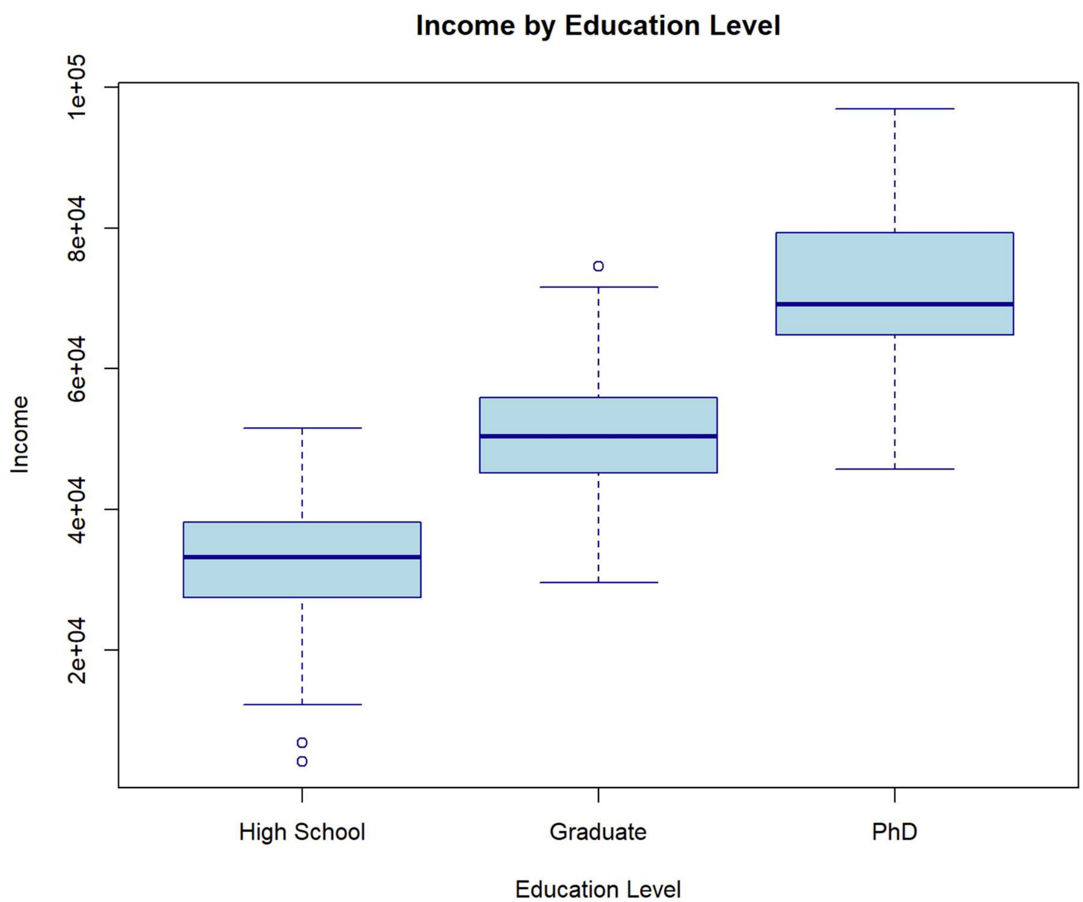
mean(data$Income[data$Gender == "Male" & data$DrinkingHabit == "Heavy Drinker"], trim = 0.05)

mean(data$Income[data$Gender == "Female" & data$DrinkingHabit == "Heavy Drinker"], trim = 0.05)

```
> mean(data$Income[data$Gender == "Male" & data$DrinkingHabit == "Heavy Drinker"], trim = 0.05)
[1] 42403.39
> mean(data$Income[data$Gender == "Female" & data$DrinkingHabit == "Heavy Drinker"], trim = 0.05)
[1] 43188.52
>
```

10.Boxplot to compare income by education level

```
boxplot(Income ~ Education,
    data = data,
    main = "Income by Education Level",
    xlab = "Education Level",
    ylab = "Income",
    col = "lightblue",
    border = "darkblue")
```



Key Insights:

1. **Mean income is higher than the median**, indicating a **right-skewed distribution** — a few high earners are pulling the average up.
2. **Income is not normally distributed** overall, as confirmed by **Shapiro-Wilk test** and high skewness/kurtosis values.
3. **Removing outliers improves normality** slightly, but income distribution still shows deviation from perfect normality.
4. **PhD holders generally earn more** than graduates and high school passouts, as shown in the boxplot comparison.
5. **Income distribution by education level** reveals that higher education tends to correlate with higher income.
6. **High School educated individuals** show the **most variability** in income, likely due to fewer structured job roles.
7. **Graduate-level social drinkers** make up a noticeable subgroup, suggesting potential sociocultural patterns in lifestyle and income.
8. **Male heavy drinkers earn more on average** than female heavy drinkers, as shown by trimmed means, though without statistical significance testing.
9. **Social drinkers tend to have higher income** than non-drinkers on average, as shown by the t-test — this could be due to networking effects, but correlation ≠ causation.
10. **The 95% confidence intervals** for social drinkers and non-drinkers do not completely overlap, indicating **a statistically significant difference** in income.
11. **Boxplot comparison of income by education** shows a clear upward trend — median income increases with education level.
12. **Kurtosis value > 3** indicates a leptokurtic distribution — income has **heavy tails**, meaning more extreme values than a normal distribution would predict.
13. **Skewness is positive**, confirming that the income distribution has a **long right tail** — some individuals earn substantially more than most.
14. **Gender and drinking habits interact with income**, but further statistical tests (e.g., ANOVA or regression) would be required for deeper causal insights.
15. **Education plays the strongest individual role** among the three variables (education, gender, drinking) in explaining **income variation**.