Dataset

```
data <- read_excel("USStates.xlsx")
str(data)
```

```
tibble [50 × 13] (S3: tbl_df/tbl/data.frame)
 $ State           : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ..
.
 $ Region          : chr [1:50] "S" "W" "W" "S" ...
 $ Population      : num [1:50] 4.849 0.737 6.731 2.966 38.803 ...
 $ HouseholdIncome : num [1:50] 43.3 70.8 49.8 40.8 61.1 ...
 $ HighSchool      : num [1:50] 84.9 92.8 85.6 87.1 84.1 89.5 91 86.9 87.1
85.3 ...
 $ College         : num [1:50] 24.9 24.7 25.5 22.4 31.4 37 39.8 31.7 26.5
29 ...
 $ Smokers         : num [1:50] 21.5 22.6 16.3 25.9 12.5 17.7 15.5 19.6 16
.8 18.8 ...
 $ PhysicalActivity: num [1:50] 45.4 55.3 51.9 41.2 56.3 60.4 50.9 49.7 50
.2 50.8 ...
 $ Obese           : num [1:50] 32.4 28.4 26.8 34.6 24.1 21.3 25 31.1 26.4
30.3 ...
 $ NonWhite        : num [1:50] 30.7 33.1 20.8 21.7 37.7 15.8 22.1 30 23.7
39.4 ...
 $ HeavyDrinkers   : num [1:50] 4.3 8.2 6.3 5 6.4 6.7 6.3 6.6 7.2 4.7 ...
 $ TwoParents      : num [1:50] 58.7 69.6 62.7 62 65.3 69.9 67 60.4 60.2 6
0.3 ...
 $ Insured         : num [1:50] 78.8 79.8 74.7 71.7 79.7 80 87.7 85.7 70.9
72.7 ...
```

This dataset comprises 50 records and 13 variables, capturing various state-level metrics from the U.S.

Variables in the Dataset:

1. State (Character): This represents the names of the 50 states in the U.S., such as "Alabama", "Alaska", "Arizona", and so on.
2. Region (Character): Categorical variable indicating the region of each state, e.g., 'S' for Southern or 'W' for Western.
3. Population (Numeric): Represents the population of each state in millions.
4. HouseholdIncome (Numeric): Denotes the average household income in thousands of dollars for each state.
5. HighSchool (Numeric): Indicates the percentage of individuals in each state who have completed high school.
6. College (Numeric): Represents the percentage of individuals in each state with a college degree.
7. Smokers (Numeric): Provides the percentage of the population in each state that identifies as smokers.
8. PhysicalActivity (Numeric): Indicates the percentage of individuals in each state that engage in physical activity.
9. Obese (Numeric): Represents the obesity rate, denoting the percentage of individuals classified as obese in each state.

10. NonWhite (Numeric): Captures the percentage of the population in each state that is nonwhite.
11. HeavyDrinkers (Numeric): Denotes the percentage of individuals in each state who are identified as heavy drinkers.
12. TwoParents (Numeric): Represents the percentage of households in each state with two parents.
13. Insured (Numeric): Indicates the percentage of the population in each state that has insurance coverage.

This dataset offers a comprehensive view of various socioeconomic and health related metrics across U.S. states. Such data can be instrumental for policymakers, researchers, and analysts aiming to discern patterns, disparities, or correlations between different factors at the state level.

```
# Loop through each column in the dataset to compute the summary
statistics of each column

for (col_name in names(data)) {
  cat("Summary for", col_name, ":\n")
  print(summary(data[[col_name]]))
  cat("\n")  # Add an extra newline for separation
}

Summary for State :
   Length     Class      Mode
       50 character character

Summary for Region :
   Length     Class      Mode
       50 character character

Summary for Population :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.584   1.858   4.532   6.364   6.983  38.803

Summary for HouseholdIncome :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.03   46.81   51.76   53.28   58.72   73.54

Summary for HighSchool :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  83.80   87.10   89.70   89.32   91.62   95.40

Summary for College :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.10   25.90   30.15   30.83   35.25   48.30

Summary for Smokers :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.30   16.65   19.05   19.32   21.48   27.30

Summary for PhysicalActivity :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  37.40   47.65   50.65   50.73   54.12   64.10

Summary for Obese :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.30   26.40   29.40   28.77   31.07   35.10
```

```
Summary for NonWhite :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.80   13.35   20.75   22.16   30.23   75.00

Summary for HeavyDrinkers :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.300   5.200   6.150   6.046   6.775   8.600

Summary for TwoParents :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  52.30   62.70   65.45   65.52   69.50   80.60

Summary for Insured :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  67.30   76.15   79.90   80.15   84.47   92.80
```

Task 1

Given the variables in this dataset, which variables can be considered explanatory (X), and which considered response (Y)? Can any variables take on both roles? Make a table that summarizes your conclusions.

In data analysis, understanding the role of each variable is crucial. Whether a variable acts as explanatory (X) or as a response (Y) is not always set in stone; it often shifts based on the research objectives and the narrative or insights we intend to uncover from the data. Taking that into consideration, here's a suggested breakdown of the given variables:

1. State and Region: These are identifiers and categorical explanatory variables, respectively. They won't typically be used as response variables, but they help segment and categorize other data points.
2. Population: This would often be used as an explanatory variable, especially if looking at influences on other factors on a per capita basis. However, it's rarely used as a response variable.
3. HouseholdIncome: This could serve both as an outcome of certain state characteristics (response) and as a predictor for others (explanatory).
4. HighSchool and College: Education levels are typically explanatory, as they can predict many socio-economic or health outcomes.
5. Smokers, PhysicalActivity, Obese: These are health metrics and could be either response or explanatory. For instance, one might want to see if HouseholdIncome predicts obesity rates.
6. NonWhite: Typically, explanatory in socio-economic and health studies.
7. HeavyDrinkers: Could be either, depending on context.
8. TwoParents: Often explanatory, especially in studies about child outcomes or household stability.
9. Insured: This could serve both as an outcome (e.g., influenced by HouseholdIncome) and as a predictor for health metrics.

S. No.	Variable	Explanatory (X)	Response (Y)	Both
1	State	X		
2	Region	X		
3	Population	X		
4	HouseholdIncome	X	X	X
5	HighSchool	X		
6	College	X		
7	Smokers		X	X
8	PhysicalActivity		X	X
9	Obese		X	X
10	NonWhite	X		
11	HeavyDrinkers		X	X
12	TwoParents	X		
13	Insured	X	X	X

Task 2

What is the population of interest for this problem (yes – this is a trick question!)? Be sure your answer is clear and complete.

When discussing the "population" in statistical terms, we're referring to a complete set of items or entities of interest from which we can draw observations. It's the entire group we wish to describe or understand, and it forms the basis for any generalizations or conclusions we make from our data.

In the context of the provided dataset, which includes various socio-economic and health-related metrics for all 50 U.S. states, the statistical population of interest is the entirety of the United States at the state level. Specifically, this means:

- We're interested in understanding and drawing conclusions about the characteristics and behaviours of all states in the U.S., not just a subset or sample.
- The data encompasses a broad range of metrics from household income, education levels, health behaviours, to insurance coverage. Therefore, any analysis or interpretation would aim to provide insights that are applicable to the whole nation at the state level.
- It's significant to note that having data for all 50 states allows for a comprehensive analysis. This breadth provides a unique advantage: rather than estimating or extrapolating data based on a sample, we can make direct and confident assertions about trends, patterns, and relationships across the entire U.S. landscape.
- Furthermore, because our dataset encompasses all states, it enables us to investigate regional disparities, correlations between different metrics, and even

potential causal relationships, all while being confident that our findings pertain to the broader U.S. context.

In conclusion, the statistical population for this dataset and the related problems or questions we might explore is the full collection of U.S. states. The comprehensive nature of the dataset provides a holistic view, allowing for robust analyses and generalizable findings about the United States at the state level.

Task 3

For the duration of this assignment, let's have HOUSEHOLDINCOME be the response variable (Y). Also, consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain all possible scatterplots relating the non-demographic explanatory variables (Xs) to the response variable (Y).

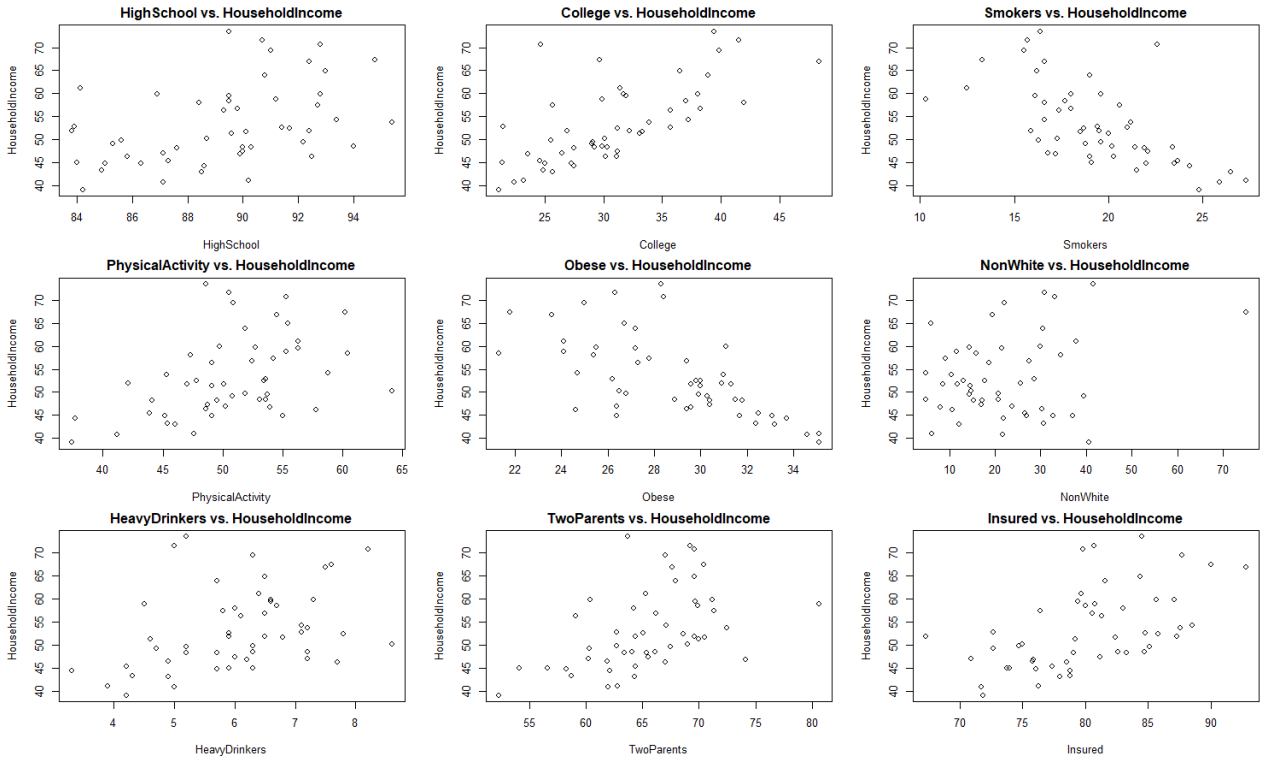
```
# Excluding character variables
numeric_data <- data %>% select_if(is.numeric)
# Calculate summary statistics
summary_stats <- numeric_data %>%
  summarise_all(list(
    n = ~sum(!is.na(.)),
    mean = ~mean(., na.rm = TRUE),
    sd = ~sd(., na.rm = TRUE),
    min = ~min(., na.rm = TRUE),
    max = ~max(., na.rm = TRUE)
  )) %>%
  gather(variable, value, everything()) %>%
  separate(variable, into = c("Variable", "Statistic")) %>%
  pivot_wider(names_from = Statistic, values_from = value)
# Print the summary using kable
kable(summary_stats)
```

Variable	n	mean	sd	min	max
Population	50	6.36394	7.150960	0.584	38.803
HouseholdIncome	50	53.28428	8.690234	39.031	73.538
HighSchool	50	89.32000	3.107135	83.800	95.400
College	50	30.83000	6.078643	21.100	48.300
Smokers	50	19.31600	3.523122	10.300	27.300
PhysicalActivity	50	50.73400	5.509643	37.400	64.100
Obese	50	28.76600	3.369286	21.300	35.100
NonWhite	50	22.15600	12.685572	4.800	75.000
HeavyDrinkers	50	6.04600	1.175291	3.300	8.600
TwoParents	50	65.52400	5.170740	52.300	80.600
Insured	50	80.14800	5.494087	67.300	92.800

```
# Scatterplots of non-demographic explanatory variables vs.
HouseholdIncome
non_demographic_vars <- c("HighSchool", "College", "Smokers",
"PhysicalActivity", "Obese", "NonWhite", "HeavyDrinkers", "TwoParents",
"Insured")

# Setting up the plotting layout
par(mfrow=c(3,3))  # 3x3 grid of plots

# Plotting
for(var in non_demographic_vars) {
  plot(data[[var]], data$HouseholdIncome, main=paste(var, "vs.
HouseholdIncome"), xlab=var, ylab="HouseholdIncome")}
```



Interpreting scatter plots:

1. HighSchool vs. HouseholdIncome: This plot shows a mild positive correlation between the percentage of individuals who have completed high school and the average household income. As the percentage of individuals with a high school education increase, there seems to be a trend of increasing household income.
2. College vs. HouseholdIncome: There's a more pronounced positive correlation here. As the percentage of individuals who have attended college increases, the household income seems to rise. This indicates that states with higher college-going rates tend to have higher average incomes.
3. Smokers vs. HouseholdIncome: This plot depicts a negative correlation between the percentage of smokers and household income. It suggests that states with higher percentages of smokers tend to have a lower average household income.

4. PhysicalActivity vs. HouseholdIncome: A positive correlation is seen here. States where a higher percentage of individuals engage in physical activity seem to have higher household incomes. This might be indicative of higher health consciousness or access to recreational facilities in more affluent states.
5. Obese vs. HouseholdIncome: There's a negative correlation. States with higher obesity rates tend to have lower household incomes. This might point to lower access to healthy foods or healthcare in states with lower incomes.
6. NonWhite vs. HouseholdIncome: The relationship here seems to be scattered without a very clear trend. There's a mild positive correlation, suggesting that states with a higher percentage of non-white populations might have slightly higher average household incomes, but the correlation isn't very strong.
7. HeavyDrinkers vs. HouseholdIncome: The relationship here is mildly positive. States with a higher percentage of heavy drinkers might have slightly higher household incomes, but the correlation isn't very pronounced.
8. TwoParents vs. HouseholdIncome: There's a positive correlation. States where more children live with both parents tend to have higher household incomes. This might indicate more stable financial situations in two-parent households.
9. Insured vs. HouseholdIncome: This plot shows a pronounced positive correlation. States with a higher percentage of insured individuals tend to have higher household incomes. This suggests that financial well-being might be associated with access to health insurance.

Overall, these graphs provide insights into various socio-economic indicators and their relationships with household income across states. Some relationships, like the ones between college education, physical activity, and being insured with household income, are more pronounced, while others are milder. The patterns seen here can provide a basis for further investigation into the underlying socio-economic factors that influence household income across different states.

Task 4

Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with the response variable Y and report the correlations in a table. Given the scatterplots from step 3) and these correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or why not?

```
# List of non-demographic variables
non_demographic_vars <- c("HighSchool", "College", "Smokers",
"PhysicalActivity", "Obese", "NonWhite", "HeavyDrinkers", "TwoParents",
"Insured")

# Calculate correlations with HouseholdIncome
correlations <- sapply(non_demographic_vars, function(var) {

  cor(data[[var]], data$HouseholdIncome, method="pearson",
use="complete.obs")
```

```
})

# Present correlations in a table
correlation_table <- data.frame(
  Variable = non_demographic_vars,
  Correlation_with_HouseholdIncome = correlations
)


print(correlation_table)
```

```
 Variable          Correlation_with_HouseholdIncome
 HighSchool                               0.4308448
 College                                  0.6855909
 Smokers                                 -0.6375225
 PhysicalActivity                         0.4404166
 Obese                                   -0.6491116
 NonWhite                                 0.2529418
 HeavyDrinkers                            0.3730143
 TwoParents                               0.4776443
 Insured                                  0.5496786
```

Interpretation of Correlation Values:

1. HighSchool (0.4308): There's a moderate positive correlation with HouseholdIncome. This suggests that states with a higher percentage of high school graduates tend to have higher household incomes. Simple linear regression might be moderately appropriate here.
2. College (0.6856): A strong positive correlation with HouseholdIncome. This indicates that states with higher college graduation rates are likely to have higher average household incomes. Simple linear regression seems appropriate for this variable.
3. Smokers (0.6375): A strong negative correlation with HouseholdIncome. States with a higher percentage of smokers tend to have lower average household incomes. Simple linear regression seems appropriate for this variable.
4. PhysicalActivity (0.4404): Moderate positive correlation with HouseholdIncome, indicating that states with higher physical activity rates might have higher average incomes. Simple linear regression might be moderately appropriate.
5. Obese (0.6491): A strong negative correlation with HouseholdIncome. This suggests states with higher obesity rates tend to have lower household incomes. Simple linear regression seems appropriate for this variable.
6. NonWhite (0.2529): A weak positive correlation with HouseholdIncome. Given the relatively weak correlation, simple linear regression may not be very effective for this variable.
7. HeavyDrinkers (0.3730): There's a moderate positive correlation with HouseholdIncome. This suggests that states with higher percentages of heavy

drinkers might have slightly higher household incomes. Simple linear regression might be moderately appropriate.

8. TwoParents (0.4776): A moderate positive correlation with HouseholdIncome, indicating that states with more two parent households might have higher average household incomes. Simple linear regression might be moderately appropriate.
9. Insured (0.5497): There's a relatively strong positive correlation with HouseholdIncome, suggesting states with higher insurance coverage rates tend to have higher average household incomes. Simple linear regression seems appropriate for this variable.

Conclusion:

Simple linear regression seems appropriate for many of the variables, especially those with strong correlations (absolute value > 0.6) with `HouseholdIncome`, such as `College`, `Smokers`, `Obese`, and `Insured`. For variables with moderate correlations (like `HighSchool`, `PhysicalActivity`, `HeavyDrinkers`, and `TwoParents`), linear regression might still provide insights but might not capture all the variance in `HouseholdIncome`. For weak correlations (like `NonWhite`), the linear regression might not be as effective.

It is essential to remember that correlation does not imply causation. Even if there's a strong correlation, it doesn't mean one variable causes the other. Also, the appropriateness of regression isn't solely based on correlation values. Other assumptions of linear regression (like linearity, independence, homoscedasticity, and normality of residuals) should be checked before finalizing its use.

Task 5

Fit a simple linear regression model to predict Y using the COLLEGE explanatory variable. Use the base STAT lm(Y~X) function. Why would you want to start with this explanatory variable? Call this Model 1. Report the prediction equation for Model 1 and interpret each coefficient of the model in the context of this problem. In addition, report and interpret the R-squared statistic for Model 1.

```
# Fitting the simple linear regression model
model1 <- lm(HouseholdIncome ~ College, data=data)

# Display the summary of the model
summary(model1)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-7.319 -4.245 -2.203  2.652 23.484

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.0664     4.7187   4.888 1.18e-05 ***
College       0.9801     0.1502   6.525 3.94e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.392 on 48 degrees of freedom
Multiple R-squared:   0.47,    Adjusted R-squared:  0.459
F-statistic: 42.57 on 1 and 48 DF,  p-value: 3.941e-08
```

Why start with the `College` explanatory variable?

1. Strong Correlation: As mentioned earlier, the `College` variable has a strong positive correlation with `HouseholdIncome`. This suggests that it might be a significant predictor and using it as a starting point can offer meaningful insights into its relationship with household income.
2. Statistical Significance: The coefficient for `College` is statistically significant (p-value is much smaller than 0.05), confirming that there's a meaningful relationship between the percentage of college graduates in a state and the average household income of that state.

Equation:

$$\text{HouseholdIncome} = 23.0664 + 0.9801 * \text{College}$$

Interpretation of the Coefficients:

1. Intercept (23.0664): When the percentage of college graduates (`College`) in a state is 0, the predicted average `HouseholdIncome` is approximately $23,066.40. While this provides a baseline, it's not practically meaningful in this context, as a state with 0% college graduates is an unrealistic scenario.
2. College Coefficient (0.9801): For every 1% increase in the population of a state that has a college degree, the average household income is predicted to increase by approximately $980.10. This is a significant rise, indicating the value of higher education in influencing average household incomes.

R-squared Statistic:

R-squared (0.47): This statistic indicates that 47% of the variance in `HouseholdIncome` can be explained by the `College` explanatory variable in this model. In other words, the percentage of college graduates in a state explains nearly half of the variability in the average household income of that state. While this is a substantial portion, it also suggests that there are other factors (constituting the remaining 53% of the variance) not included in this model that influence household income.

In conclusion, the `College` explanatory variable is a good starting point for understanding the predictors of `HouseholdIncome` because of its strong correlation and significant coefficient. The model shows a clear relationship between higher education and average household incomes, though there are other factors at play, given that the R-squared value isn't closer to 1.

Task 6

(From your Model 1 results for task 5) – Specify the null and alternative hypothesis separately for each of the two parameters in the model. Report and interpret the results of the T-tests for these hypotheses. In addition, state the null and alternative hypotheses for

the omnibus (i.e. overall) model. Report the ANOVA table and interpret the results of the F-test.

Null and Alternative Hypotheses for the Model Parameters:

1. Intercept (β0):
   - H_0: $\beta 0 = 0$ (Null Hypothesis: The intercept is 0.)
   - H_a: $\beta 0 \neq 0$ (Alternative Hypothesis: The intercept is not 0.)
2. College Coefficient (β1):
   - H_0: $\beta 1 = 0$ (Null Hypothesis: The coefficient of the `College` variable is 0, meaning there is no linear relationship between `College` percentage and `HouseholdIncome`.)
   - H_a: $\beta 1 \neq 0$ (Alternative Hypothesis: The coefficient of the `College` variable is not 0, indicating there is a linear relationship.)

T-test Results:

1. Intercept (β0): The t-value is 4.888 with a p-value of 1.18e-05, which is statistically significant at conventional levels (e.g., 0.05). This indicates that we reject the null hypothesis for the intercept, implying that the intercept is not 0 and has statistical significance in predicting `HouseholdIncome`.
2. College Coefficient (β1): The t-value is 6.525 with a p-value of 3.94e-08, which is also statistically significant. This means we reject the null hypothesis for the `College` coefficient, suggesting that there's a statistically significant linear relationship between the percentage of college graduates in a state and its average household income.

Null and Alternative Hypotheses for the Overall Model:

- H_0: All coefficients (excluding the intercept) in the model are 0. (The model doesn't fit the data better than a model with no predictors.)
- H_a: At least one coefficient in the model is not 0. (The model fits the data better than a model with no predictors.)

The F-statistic tests the overall significance of the regression model.

Given the F-statistic value of 42.57 with a p-value of 3.941e-08:

The p-value is very close to 0 and much less than conventional significance levels (e.g., 0.05). Thus, we reject the null hypothesis of the omnibus test, indicating that the regression model with `College` as a predictor fits the data better than a model with no predictors. In other words, `College` contributes significantly to explaining the variability in `HouseholdIncome`.

To sum it up, both the t-tests for individual parameters and the F-test for the overall model indicate that the `College` variable is a statistically significant predictor of `HouseholdIncome`.

```
anova(model1)
Analysis of Variance Table

Response: HouseholdIncome
           Df Sum Sq Mean Sq F value    Pr(>F)
College     1 1739.4 1739.36  42.572 3.941e-08 ***
Residuals 48 1961.1   40.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of the ANOVA table for the regression model:

1. Response: The dependent variable (or response variable) being predicted is `HouseholdIncome`.
2. Df (Degrees of Freedom):
   - College: 1 - This tells us that there's one predictor (College) in the model.
   - Residuals: 48 - This is the remaining degrees of freedom after accounting for the predictor. It's essentially the number of observations minus the number of predictors minus one.
3. Sum Sq (Sum of Squares):
   - College: 1739.4 - This is the variation in `HouseholdIncome` that's explained by the `College` predictor.
   - Residuals: 1961.1 - This is the variation in `HouseholdIncome` that's not explained by the model. It's essentially the discrepancy between the observed values and the values predicted by the model.
4. Mean Sq (Mean Squares):
   - College: 1739.36 - This is the explained variance by `College` per degree of freedom. Calculated as the Sum of Squares for College divided by its degrees of freedom (1739.4/1).
   - Residuals: 40.86 - This is the unexplained variance per degree of freedom. Calculated as the Sum of Squares for Residuals divided by its degrees of freedom (1961.1/48).
5. F value (F-statistic): 42.572 - This measures the overall significance of the predictor in the model. It's the ratio of the Mean Square for `College` to the Mean Square for Residuals. A high F value indicates that the predictor significantly improves the fit of the model over a model with no predictors.
6. Pr(>F) (p-value for F-statistic): 3.941e-08 - This p-value tests the null hypothesis that the `College` coefficient is zero (meaning it has no effect). Given its very low value, it's much less than conventional significance levels (like 0.05), and therefore, we reject the null hypothesis. This indicates that the percentage of college graduates in a state is a statistically significant predictor of the average household income of that state.

Overall Interpretation:

The ANOVA table suggests that the `College` variable significantly explains the variation in `HouseholdIncome`. The high F-statistic and the extremely low associated p-value confirm the importance of the `College` predictor in the regression model. Given these results,

there's a statistically significant relationship between the percentage of college graduates in a state and its average household income.

Task 7

Use the predicted values and the original response variable Y to create a variable of residuals (i.e., residual = Y – Y_hat = observed minus predicted) for Model 1. Using the original Y variable, the predicted, and/or residual variables, write R-code to:

- Square each of the residuals and then add them up. This is called sum of squared residuals or sums of squared errors.
- Deviate the mean of the Y's from the value of Y for each record (i.e., Y – Y_bar). Square each of the deviations and then add them up. This is called sum of squares total.
- Deviate the mean of the Y's from the value of predicted (Y_hat) for each record (i.e., Y_hat – Y_bar). Square each of these deviations and then add them up. This is called the sum of squares due to regression.
- Calculate a statistic that is: (Sum of Squares due to Regression) / (Sum of squares Total)

Verify and note the accuracy of the ANOVA table and R-squared values from the regression printout from part 4), relative to your computations here. Report your R-code for these computations.

```
# Predicted values using Model 1
Y_hat <- predict(model1, data)

# Residuals
residuals <- data$HouseholdIncome - Y_hat

# Sum of Squared Errors
sse <- sum(residuals^2)

# Sum of Squared Totals
Y_bar <- mean(data$HouseholdIncome)   # Mean of Y
sst <- sum((data$HouseholdIncome - Y_bar)^2)

# Sum of Squares due to Regression
ssr <- sum((Y_hat - Y_bar)^2)

# Compute Desired Statistic
statistic <- ssr / sst

# Predicted values and Residuals
Y_hat <- predict(model1, data)
residuals <- data$HouseholdIncome - Y_hat

# Sum of Squared Residuals (SSE)
sse <- sum(residuals^2)

# Sum of Squares Total (SST)
Y_bar <- mean(data$HouseholdIncome)
sst <- sum((data$HouseholdIncome - Y_bar)^2)

# Sum of Squares due to Regression (SSR)
ssr <- sum((Y_hat - Y_bar)^2)

# Calculate the statistic
statistic <- ssr / sst

# Print out the results
```

```
cat("SSE:", sse, "\n")
cat("SST:", sst, "\n")
cat("SSR:", ssr, "\n")
cat("Statistic:", statistic, "\n")
```

```
> # Print out the results
> cat("SSE:", sse, "\n")
SSE: 1961.13
> cat("SST:", sst, "\n")
SST: 3700.488
> cat("SSR:", ssr, "\n")
SSR: 1739.359
> cat("Statistic:", statistic, "\n")
Statistic: 0.4700349
```

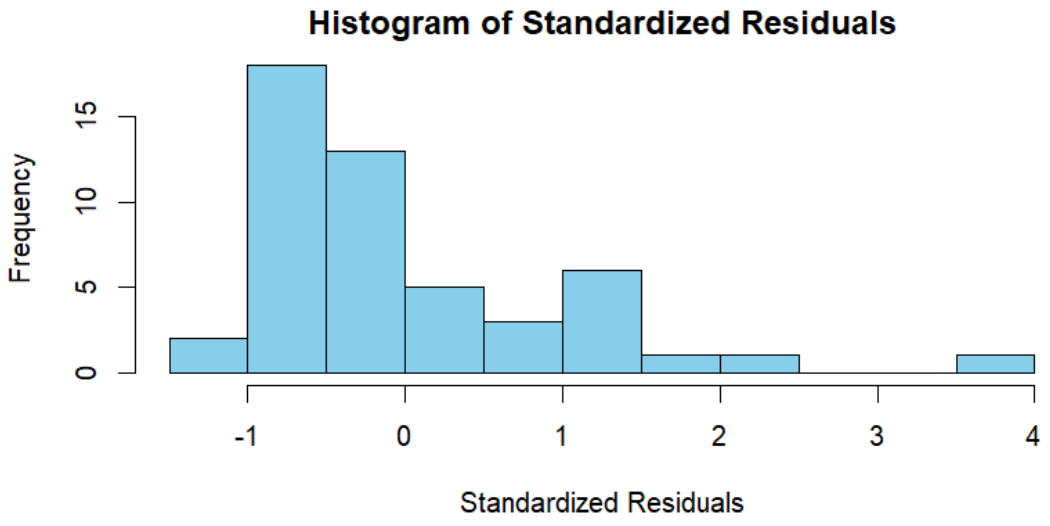
Task 8

From task 7 you created a variable of residuals for Model 1. Write R-code to standardize the residuals. Do not use residuals from the lm(). Plot the standardized residuals using a histogram. Also, plot the standardized residuals in a scatterplot with the predicted values. Discuss what you see in these two graphs.

```
# Standardize the Residuals
std_residuals <- residuals / sd(residuals)

# Plot the Standardized Residuals using a Histogram:
hist(std_residuals, main="Histogram of Standardized Residuals",
xlab="Standardized Residuals", col="skyblue", border="black")

# Scatter plot of Standardized Residuals vs. Predicted Values:
plot(Y_hat, std_residuals, main="Scatterplot of Standardized Residuals vs.
Predicted Values", xlab="Predicted Values", ylab="Standardized Residuals",
pch=19, col="blue")
abline(h=0, col="red") # horizontal line at y=0
```

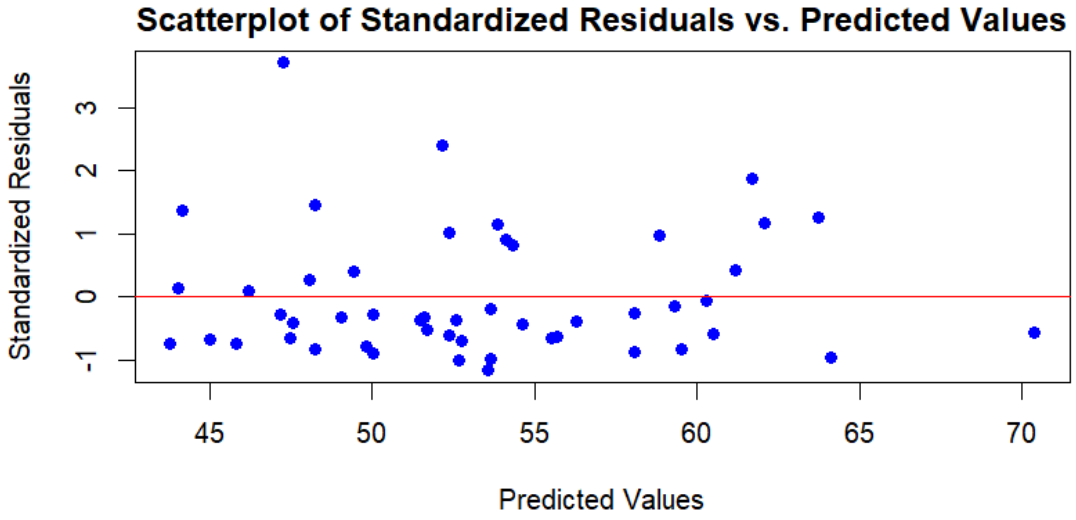


Histogram of Standardized Residuals:

- Distribution: The residuals seem somewhat symmetrically distributed around zero, but it doesn't exactly follow a perfect normal distribution. There's a peak slightly to the left of zero and then a gradual decline as we move further away, with a few residuals on the extreme right tail.

- Skewness: The distribution appears to have a slight right skew since there are a few more extended bars on the positive side than on the negative side.
- Tail Behavior: There are residuals in the positive tail that reach up to a value of 4, suggesting the presence of some large positive residuals (potential outliers) that might be influencing the fit of the regression model.

In summary, while there's a general bell shape indicative of normality, the presence of tail values and the slightly skewed nature of the distribution suggests a deviation from the assumption of normally distributed errors.



Scatterplot of Standardized Residuals vs. Predicted Values:

- Distribution Around Zero: Many residuals cluster around the horizontal line at y=0, which suggests that the model does a reasonably good job predicting the response for those points.
- Pattern of Spread: There isn't a clear funneling or any systematic pattern, indicating the residuals display relatively consistent variance across different predicted values (homoscedasticity).
- Outliers: Some points, especially in the upper region, are further away from the zero line. These might represent outliers or influential observations that might warrant further investigation.
- Lack of Systematic Curvature: There isn’t an obvious Ushape or inverted Ushape, which means that nonlinearity isn’t a major concern with this model, at least based on this visual check.

In summary, the scatterplot suggests a reasonable fit of the regression model to the data, but potential outliers might need closer inspection.

Conclusion:
Both graphs collectively provide insights into the model's assumptions and its fit. While the scatterplot suggests a good linear relationship without obvious heteroscedasticity, the histogram of standardized residuals indicates a potential deviation from normality and the presence of influential observations. Before drawing definitive conclusions or making

predictions, it would be advisable to further investigate the potential outliers and consider transformations or other strategies to address the nonnormality in the residuals.

Task 9

Select a different explanatory variable and use that variable in a Simple Linear Regression model to predict Y, HOUSEHOLDINCOME. Call this Model 2. Report and interpret the results of Model 2. Which is the better model, Model 1 or Model 2? Give evidence to justify your answer.

```
# Linear Regression Model with "Smokers" is the variable for Model 2

model2 <- lm(HouseholdIncome ~ Smokers, data=data)
summary(model2)


Call:
lm(formula = HouseholdIncome ~ Smokers, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.285  -4.074  -1.100   2.434  22.640

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.6593     5.3840  15.539  < 2e-16 ***
Smokers      -1.5725     0.2743  -5.733  6.4e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.765 on 48 degrees of freedom
Multiple R-squared:  0.4064,   Adjusted R-squared:  0.3941
F-statistic: 32.87 on 1 and 48 DF,  p-value: 6.396e-07
```

Interpretation of the output of the linear regression model where "HouseholdIncome" is the response variable and "Smokers" is the predictor.

Equation:

HouseholdIncome = 83.6593 - 1.5725 *Smokers

1. Coefficients:
   - Intercept (83.6593): This is the predicted value of 'HouseholdIncome' when the 'Smokers' variable is 0. In the context of this data, it represents the expected average household income in states where no one smokes (which is a theoretical and not a practical scenario).
   - Smokers (-1.5725): This coefficient represents the change in 'HouseholdIncome' for each additional 1% increase in the smoking population. Specifically, for every 1% increase in the population that smokes, the average household income decreases by approximately $1,572.50. This suggests a negative relationship between the percentage of smokers in a state and its average household income.

2. Significance:

- Both the intercept and the coefficient for 'Smokers' are highly significant with p-values less than 0.05 (actually, both are much smaller than 0.05). The '*' next to the coefficients indicates a high level of statistical significance.

3. Model Fit and Variability:
   - Residual Standard Error (6.765): This measures the average amount that the observed values of 'HouseholdIncome' deviate from the values predicted by the model. The average deviation is approximately $6,765.
   - Multiple R-squared (0.4064): This tells us that approximately 40.64% of the variability in 'HouseholdIncome' is explained by the 'Smokers' variable.
   - Adjusted R-squared (0.3941): It adjusts the R-squared for the number of predictors in the model. It's almost the same as the R-squared since we only have one predictor.
   - F-statistic (32.87) and its associated p-value (6.396e-07): This tests the overall significance of the model. The small p-value suggests that the model with the 'Smokers' predictor is a better fit than a model with no predictors.

Overall Interpretation: The linear regression model suggests a significant negative relationship between the percentage of smokers in a state and its average household income. For every 1% increase in smokers, there's an associated decrease in the average household income by approximately $1,572.50. Given the R-squared value, the percentage of smokers in a state can explain about 40.64% of the variability in its average household income.

Comparing Model 1 and Model 2:

Model 1 (Predictor: College):
- Adjusted Rsquared: 0.459
- Multiple Rsquared: 0.47
- Residual Standard Error: 6.392
- Significant predictor: Yes (pvalue very small)

Model 2 (Predictor: Smokers):
- Adjusted Rsquared: 0.3941
- Multiple Rsquared: 0.4064
- Residual Standard Error: 6.765
- Significant predictor: Yes (pvalue very small)

1. Adjusted R-squared: Model 1 has a higher adjusted R-squared (0.459) compared to Model 2 (0.3941). This means Model 1 explains more of the variance in 'HouseholdIncome' than Model 2.
2. Multiple R-squared: Similarly, the multiple R-squared value for Model 1 (0.47) is higher than for Model 2 (0.4064).
3. Residual Standard Error: The residual standard error for Model 1 (6.392) is lower than for Model 2 (6.765). A lower residual standard error indicates that Model 1 predictions are, on average, closer to the actual observations than those of Model 2.
4. Significance of Predictor: Both models have predictors that are statistically significant. However, the strength or magnitude of the relationship (as denoted by the coefficient)

and the practical implications should also be considered. For instance, the relationship between 'College' attendance and 'HouseholdIncome' might have different implications than the relationship between 'Smokers' and 'HouseholdIncome'.

Conclusion: Based on the metrics provided, Model 1 (with 'College' as the predictor) appears to be the better model. It explains a higher proportion of the variance in 'HouseholdIncome', and its predictions are, on average, closer to the actual values than those of Model 2.

Task 10

For this last task, you are welcome to fit any Simple Linear Regression model that you wish on the US States data. You'll need to decide on the response variable as well as the explanatory variable. Call this Model 3. Report and interpret the results of Model 3.

```
# Model 3: Predicting "HouseholdIncome" using the "Insured" variable.

model3 <- lm(HouseholdIncome ~ Insured, data=data)
summary(model3)


Call:
lm(formula = HouseholdIncome ~ Insured, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-8.896 -5.963 -1.976  5.200 17.865

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.4004    15.3210  -1.070     0.29
Insured       0.8695     0.1907   4.559 3.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.335 on 48 degrees of freedom
Multiple R-squared:  0.3021,   Adjusted R-squared:  0.2876
F-statistic: 20.78 on 1 and 48 DF,  p-value: 3.557e-05
```

Equation:

HouseholdIncome = -16.4004 + 0.8695 *Insured

1. Intercept (-16.4004): This coefficient represents the predicted value of the `HouseholdIncome` when the `Insured` percentage is 0. Theoretically, if no one in a state is insured, the predicted average household income would be -$16,400. This doesn't have a real-world interpretation, as it's unlikely to have a state where no one is insured, and negative income isn't practical. Moreover, the p-value for the intercept is 0.29, indicating it's not statistically significant at conventional levels.
2. Insured (0.8695): This coefficient indicates that for every 1% increase in the population that is insured, the average household income is expected to increase by approximately $869.50. The p-value associated with the `Insured` coefficient is very small (3.56e-05), suggesting that the relationship between the percentage of insured individuals and the average household income is statistically significant.

3. Model Fit:
    - Adjusted R-squared (0.2876): This metric tells us that approximately 28.76% of the variability in `HouseholdIncome` can be explained by the percentage of insured individuals in the state. The closer this value is to 1, the better the model fits the data.
    - Multiple R-squared (0.3021): This metric indicates that the `Insured` variable accounts for about 30.21% of the variance in `HouseholdIncome`.

4. Significance Tests:

    - t-value for Insured (4.559): This statistic measures how many standard errors the coefficient is away from zero. A larger t-value suggests that it's less likely that the coefficient is zero (indicating no relationship).
    - F-statistic (20.78) and its associated p-value (3.557e-05): These metrics evaluate the overall significance of the model. In this case, the p-value is extremely low, indicating that the model with `Insured` as a predictor is significantly better at explaining the variability in `HouseholdIncome` than a model with no predictors.

In conclusion, there is a statistically significant positive relationship between the percentage of insured individuals in a state and its average household income. Every 1% increase in insured individuals is associated with an average rise in household income by $869.50. Given the R-squared values, the percentage of insured individuals can explain between 28.76% and 30.21% of the variability in a state's average household income.