Assignment 2 HI 743

Manthan Mehta

2025-03-05

Section 1 Statistical Analysis Report

Boston Data Analysis

Introduction & Objective

The objective of this analysis is to explore the factors influencing median home prices (medv) in the Boston housing dataset. This study specifically examines the impact of the lower status population (lstat) and house age (age), along with their interaction effect, on housing prices. Additionally, we implement predictive modeling techniques to evaluate their performance in predicting house prices

Dataset Understanding & Preparation

Dataset Description

The dataset consists of 506 observations and 12 predictor variables, along with the target variable (`medv`). Below is a description of each variable:

-crim: Per capita crime rate by town.

-zn: Proportion of residential land zoned for lots over 25,000 sq. ft.

-indus: Proportion of non-retail business acres per town.

-chas: Charles River dummy variable (1 if tract bounds river; 0 otherwise).

-nox: Nitrogen oxides concentration (parts per 10 million).

-rm: Average number of rooms per dwelling.

-age: Proportion of owner-occupied units built prior to 1940.

-dis: Weighted mean of distances to five Boston employment centers.

-rad: Index of accessibility to radial highways.

-tax: Full-value property tax rate per $10,000.

-pratio: Pupil-teacher ratio by town.

-lstat: Percentage of lower-status population.

-medv: Median value of owner-occupied homes in $1000s (Target variable).

Summary Statistics of the dataset

```
summary(Boston)
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          lstat      
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00
```

Variables used in the analysis

For this analysis, we focus on:

-lstat(Percentage of lower-status population)

-age (Proportion of owner-occupied units built before 1940)

-lstate*age (Interaction term)

These variables were selected based on their potential impact on housing prices, which we explore through linear regression and interaction effects

Data Cleaning and Preprocessing

1. Checking for any missing values-

```
missing_values = Boston %>% summarise(across(everything(), ~sum(is.na(.))))
print(missing_values)
```

```
##   crim zn indus chas nox rm age dis rad tax ptratio lstat medv
## 1    0  0     0    0   0  0   0   0   0   0       0     0    0
```

No missing were found in the dataset

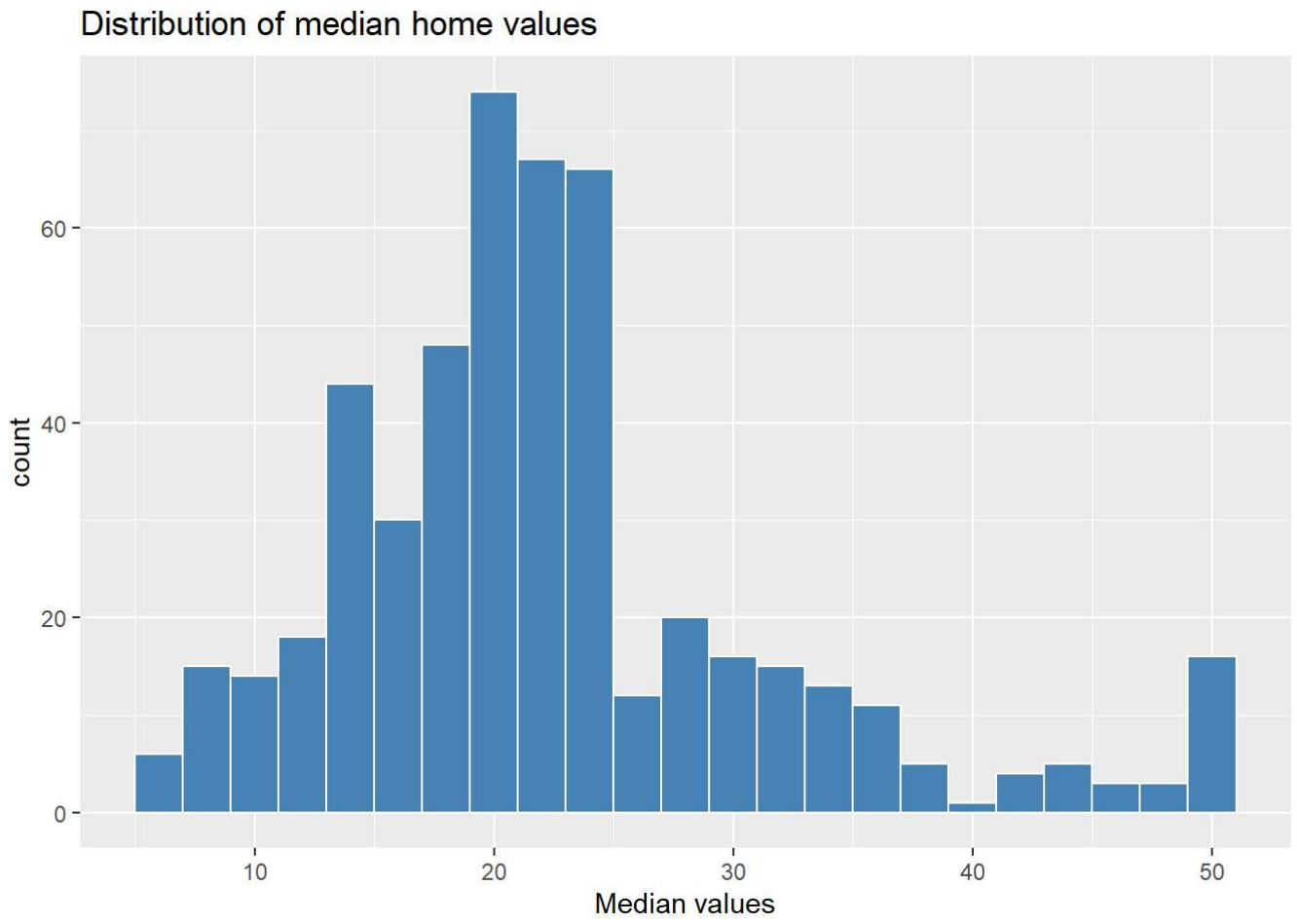
2)Splitting Data into Training and Testing Sets:

```
set.seed(123)
Boston_split = Boston %>% mutate(id=row_number()) %>% sample_frac(0.75)
Boston = Boston %>% mutate(id=row_number())
train_data = Boston_split
test_data = anti_join(Boston, Boston_split, by="id")
```

75% of the data is used for training, and 25% for testing.

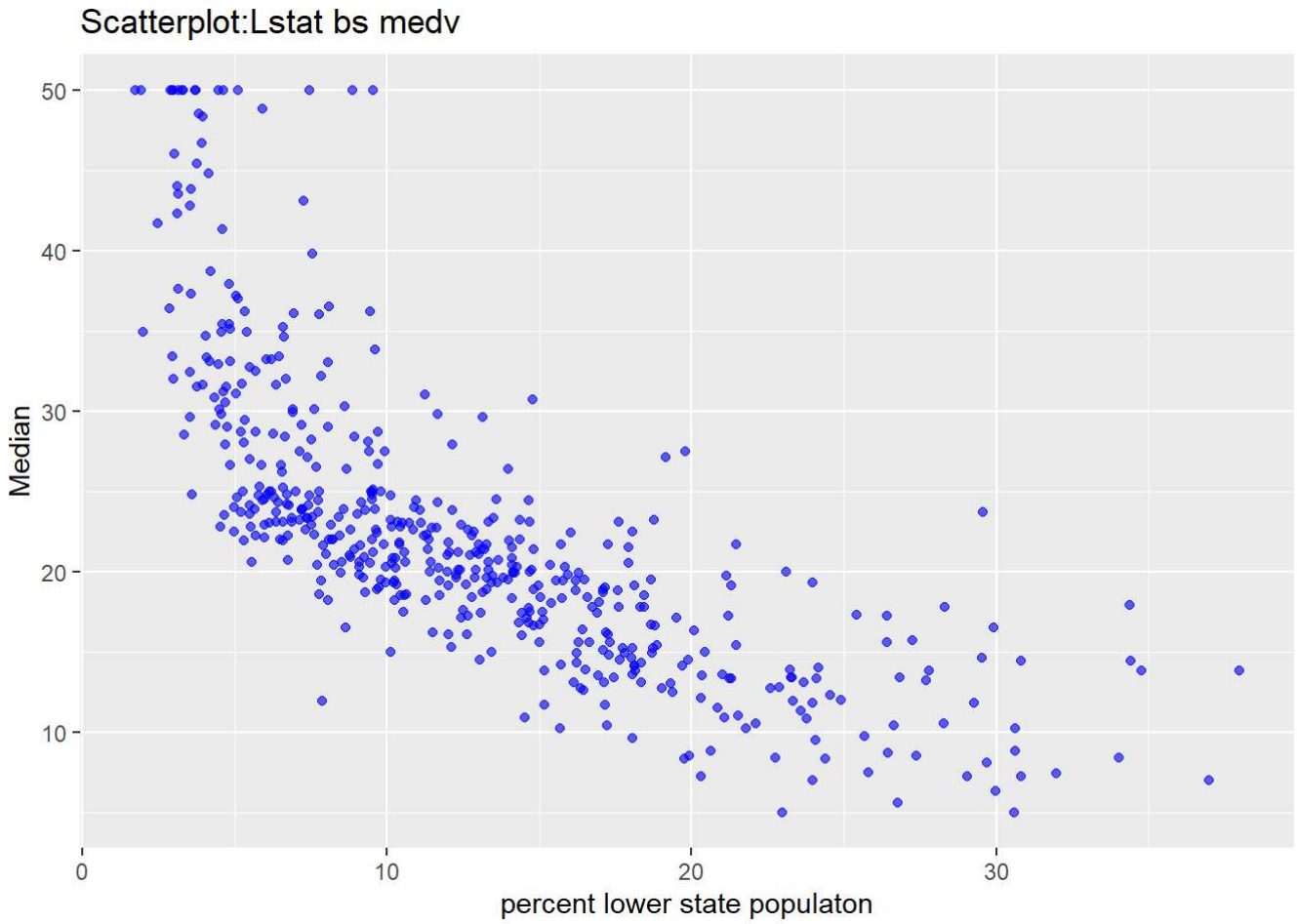
Exploratory Data Analysis:

```
ggplot(Boston,aes(x=medv))+
  geom_histogram(fill="steelblue",binwidth = 2, color="white")+
  labs(title = "Distribution of median home values",
       x="Median values",
       y="count")
```



-The histogram indicates that the distribution of median home values is right-skewed, with most values ranging between $15,000 and $25,000. There is a noticeable peak at $50,000, suggesting a possible upper capping of home values in the dataset.(maybe some prices were higher than 50000, but for this study they were capped at 50000$)

```
ggplot(Boston,aes(x=lstat,y=medv))+
  geom_point(alpha=0.6,color="blue")+
  labs(title="Scatterplot:Lstat bs medv",
       x="percent lower state populaton",
       y="Median")
```



-The scatterplot reveals a strong negative correlation between lstat and medv. As the percentage of lower-status population increases, median home values decrease.

-The relationship appears non-linear, suggesting that more complex models might better capture this trend.

-Similar to the histogram, many points are capped at $50,000, reinforcing the likeilihood of a capping issue in the dataset.

Model Implementation and Explanation

Simple linear Regression Model

```
lm.fit = lm(medv ~ lstat,data = train_data)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.068  -3.891  -1.275   1.706  24.613
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.51650    0.62851   54.92   <2e-16 ***
## lstat       -0.95795    0.04357  -21.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.131 on 378 degrees of freedom
## Multiple R-squared:  0.5612, Adjusted R-squared:  0.5601
## F-statistic: 483.5 on 1 and 378 DF,  p-value: < 2.2e-16
```

```
# Calculating Mean Squared Error (MSE)
train_mse <- mean((train_data$medv - predict(lm.fit, train_data))^2)
test_mse <- mean((test_data$medv - predict(lm.fit, test_data))^2)


print(train_mse)
```

```
## [1] 37.39408
```

```
print(test_mse)
```

```
## [1] 41.85541
```

Model summary:

-Intercept = 34.52, meaning that if lstat were zero, the predicted median home value would be $34,520.

-Slope for lstat = -0.958, showing that for each 1% increase in the lower-status population , the median home value decreases by approximately $958.

-Residual Standard Error (RSE) = 6.131, indicating the typical prediction error in thousands of dollars.

-Multiple R-squared = 0.561, meaning that 56.1% of the variability in home values can be explained by lstate variable.

-F-statistic = 483.5, with a p-value <2.2e-16, indicating that the model is highly significant.

This suggests that lstat has a strong and statistically significant negative impact on home values.

Mean Squared Error (MSE):

-Training MSE: 37.39

-Testing MSE: 41.86

Multiple Linear Regression model-

```
lm.multiple.fit = lm(medv~lstat + age+lstat*age,data = train_data)
summary(lm.multiple.fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat + age + lstat * age, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.807  -3.756  -1.185   1.859  25.509
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.897129   1.592822  21.909  < 2e-16 ***
## lstat       -1.272874   0.184461  -6.900 2.21e-11 ***
## age          0.014279   0.021794   0.655    0.513
## lstat:age    0.002649   0.002041   1.298    0.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.078 on 376 degrees of freedom
## Multiple R-squared:  0.5711, Adjusted R-squared:  0.5677
## F-statistic: 166.9 on 3 and 376 DF,  p-value: < 2.2e-16
```

```
train_mse=mean((train_data$medv - predict(lm.multiple.fit,train_data))^2)
test_mse= mean((test_data$medv - predict(lm.multiple.fit,test_data))^2)
print(train_mse)
```

```
## [1] 36.55216
```

```
print(test_mse)
```

```
## [1] 40.67891
```

Model Summary:

-Intercept = 34.90, meaning that if lstat and age were zero, the predicted median home value would be $34,900.

-Slope for lstat= -1.273, showing that an increase in lstat reduces medv significantly.

-Slope for age= 0.014, which is not statistically significant (p = 0.513).

-Interaction term = 0.00265, meaning that as houses get older, the impact of lstat on medv slightly weakens. However, this term is also not statistically significant (p = 0.195).

-Residual Standard Error (RSE) = 6.078, indicating the typical prediction error in thousands of dollars.

-Multiple R-squared = 0.571, meaning that 57.1% of the variability in home values can be explained by lstat, age, and their interaction.

-F-statistic = 166.9, with a p-value <2.2e-16, confirming that the model is highly significant.

Mean Squared Error (MSE):

-Training MSE: 36.55

-Testing MSE: 40.68

Conclusion

-The multiple regression model provides a slight improvement over the simple model, but the improvement is minimal.

-The interaction effect is not significant, so age does not meaningfully modify the effect of lstate.

-The slightly lower MSE and higher R^2 in the multiple model suggest a marginally better fit, but lstat alone remains a strong predictor of home values.

Limitations-

-The dataset may suffer from truncation at $50,000, potentially affecting predicions.

-The model assumes a linear relationship, but the scatterplot suggests possible non-linerity.

-Other important predictors like rm (number of rooms) or crim (crime rate) were not included, which could improve model accuracy.

Section 2: Follow Up Assignment

```
# Loading necessary libraries
library(tidyverse)
library(NHANES)
```

```
## Warning: package 'NHANES' was built under R version 4.3.3
```

Problem Definition & Justification

The objective of this analysis is to predict Body Mass Index (BMI) using Age, Smoking Status (SmokeNow), and Physical Activity (Physactive) for individuals aged 18 to 70. BMI is a crucial health indicator associated with chronic diseases such as obesity, diabetes, and cardiovascular conditions. Understanding the relationship between age, lifestyle factors (smoking and physical activity), and BMI can help in developing targeted public health interventions

```
# Loading the NHANES dataset
data(NHANES)

# Data Preparation: Selecting and filtering relevant data
SMOKERS <- NHANES %>%
  select(BMI, Age, SmokeNow, PhysActive) %>%
  filter(Age >= 18 & Age <= 70)
```

Data Import, Cleaning, & Exploration

Data Loading

The dataset is extracted from the NHANES (National Health and Nutrition Examination Survey) database. The variables selected for analysis include:

-BMI: Body Mass Index (kg/m^2)

-Age: Age of the individual (years)

-SmokeNow: Current smoking status (Yes or No)

-PhysActive: Physical activity status (Yes or No)

Summary Statistics

```
summary(SMOKERS)
```

```
##       BMI             Age        SmokeNow    PhysActive
##  Min.   :15.02   Min.   :18.00   No  :1389   No :2973
##  1st Qu.:23.94   1st Qu.:30.00   Yes :1406   Yes:3690
##  Median :27.63   Median :42.00   NA's:3868
##  Mean   :28.76   Mean   :42.48
##  3rd Qu.:32.34   3rd Qu.:54.00
##  Max.   :81.25   Max.   :70.00
##  NA's   :47
```

```
# Check for missing values
missing_values <- SMOKERS %>%
  summarise(across(everything(), ~sum(is.na(.))))
print(missing_values)
```

```
## # A tibble: 1 × 4
##     BMI   Age SmokeNow PhysActive
##   <int> <int>    <int>      <int>
## 1    47     0     3868          0
```

```
# Removing all BMI values greater than 40 and missing BMI values
SMOKERS <- SMOKERS %>%
  filter(BMI <= 40 & !is.na(BMI))

# Removing all rows where SmokeNow is missing
SMOKERS <- SMOKERS %>%
  filter(!is.na(SmokeNow))

# Final checking for missing values
missing_values <- SMOKERS %>%
  summarise(across(everything(), ~sum(is.na(.))))
print(missing_values)
```

```
## # A tibble: 1 × 4
##     BMI   Age SmokeNow PhysActive
##   <int> <int>    <int>      <int>
## 1     0     0        0          0
```

```
# Splitting the dataset into training (75%) and testing (25%) sets
set.seed(123)
SMOKERS <- SMOKERS %>% mutate(id = row_number())
SMOKERS_split <- SMOKERS %>% sample_frac(0.75)

train_data <- SMOKERS_split
test_data <- anti_join(SMOKERS, SMOKERS_split, by = "id")
```

Handling Missing and Extreme Values

BMI Cleaning:

-Removed all BMI values greater than 40, as they were considered extreme outliers.

-Removed all rows where BMI was missing to maintain data integrity.

SmokeNow Cleaning:

-Removed all rows where SmokeNow was missing to ensure consistency in categorical analysis.

Final dataset was verified to ensure no missing values remain.

Splitting the Dataset into Training and Testing Sets

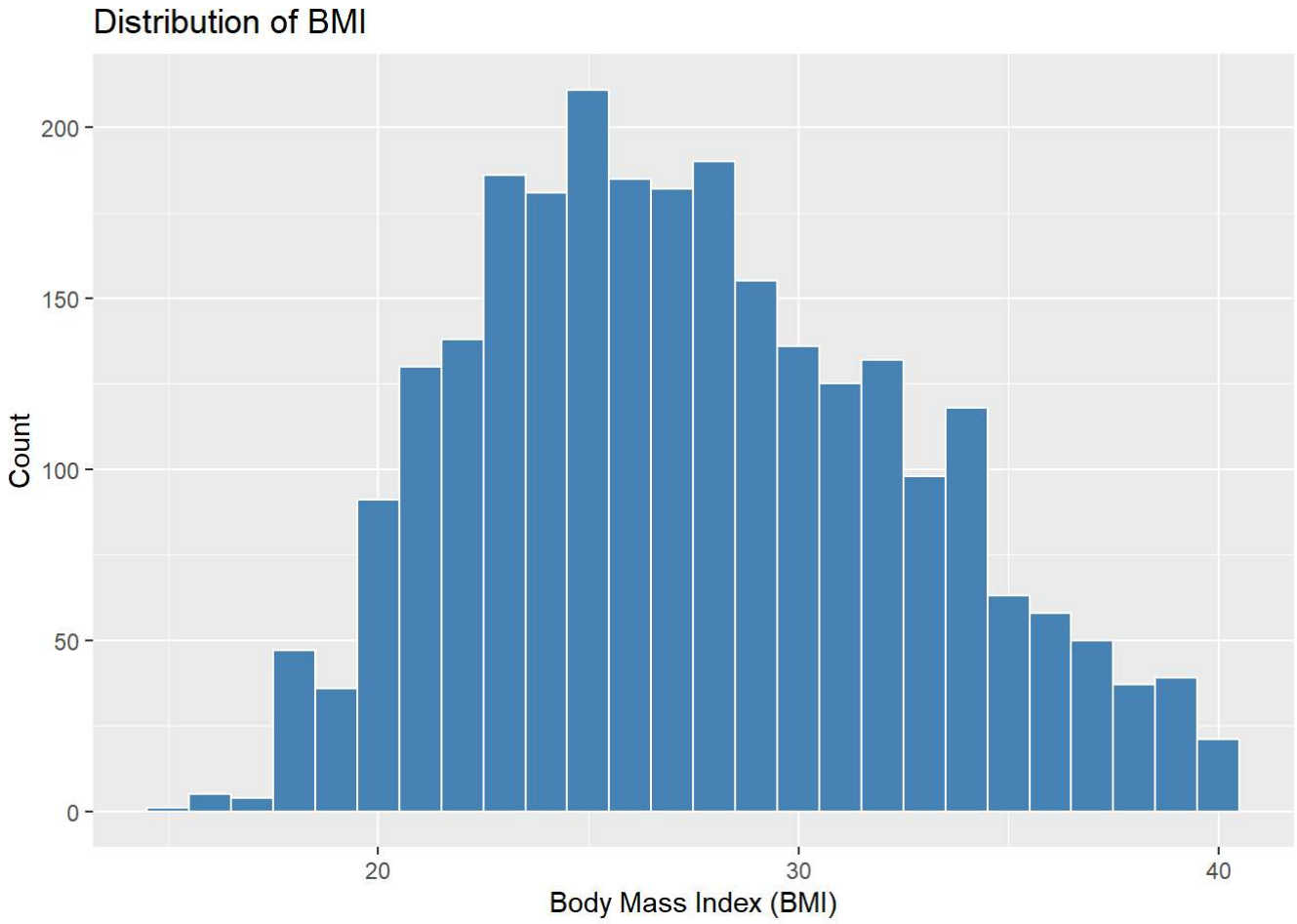
To evaluate model performance, the cleaned dataset was split into:

-75% Training Data: Used to train the predictive models.

-25% Testing Data: Used to assess the model’s performance on unseen data.

Exploratory Data Analysis

```
ggplot(SMOKERS, aes(x = BMI)) +
  geom_histogram(fill = "steelblue", binwidth = 1, color = "white") +
  labs(title = "Distribution of BMI",
       x = "Body Mass Index (BMI)",
       y = "Count")
```

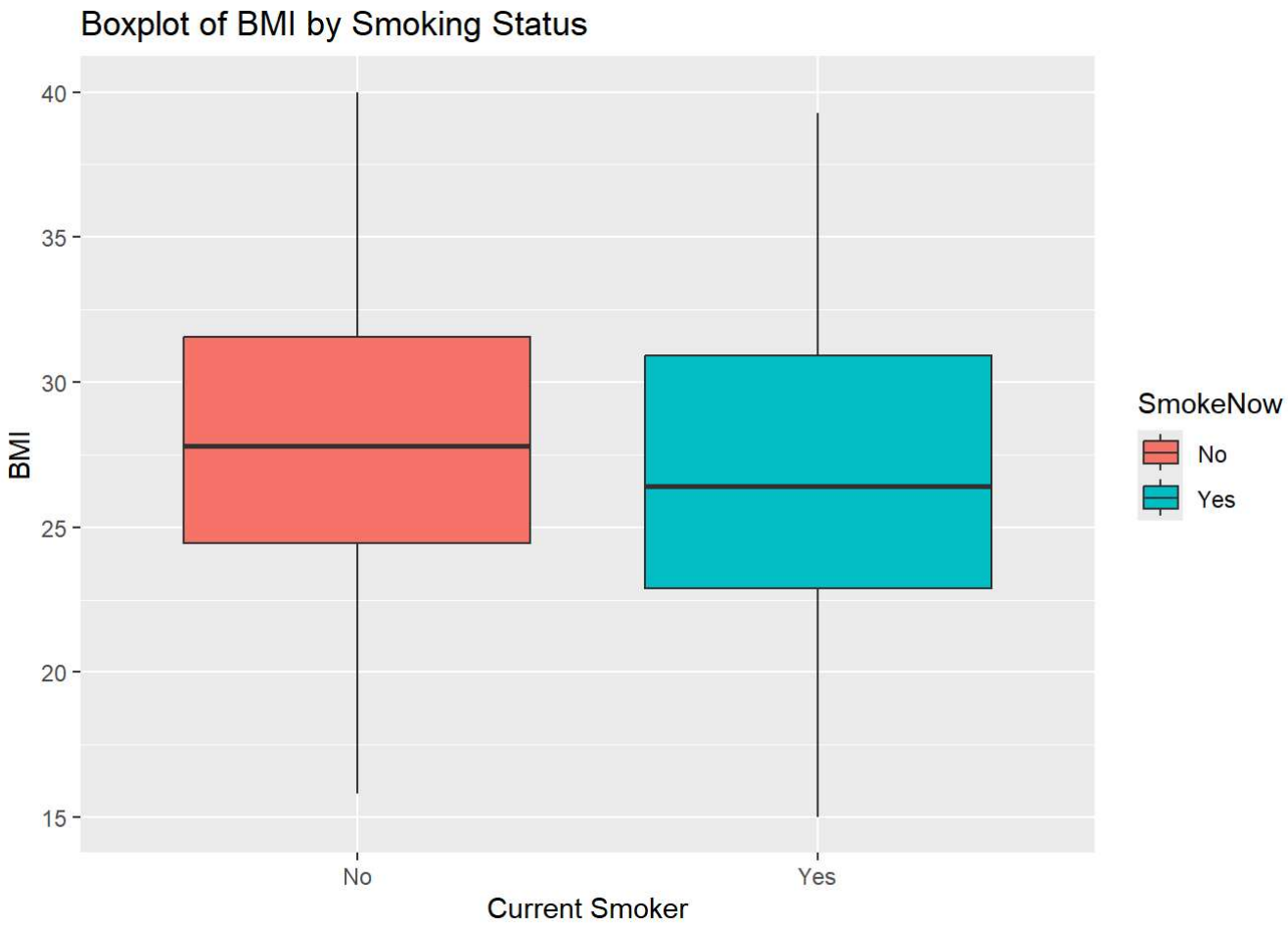


The histogram of BMI shows a roughly normal distribution, with most values ranging between 20 and 35.

The distribution is slightly right-skewed, meaning there are some individuals with higher BMI values.

The highest concentration of individuals has a BMI between 25 and 30, which falls within the overweight category.

```
ggplot(SMOKERS, aes(x = SmokeNow, y = BMI, fill = SmokeNow)) +
  geom_boxplot() +
  labs(title = "Boxplot of BMI by Smoking Status",
       x = "Current Smoker",
       y = "BMI")
```



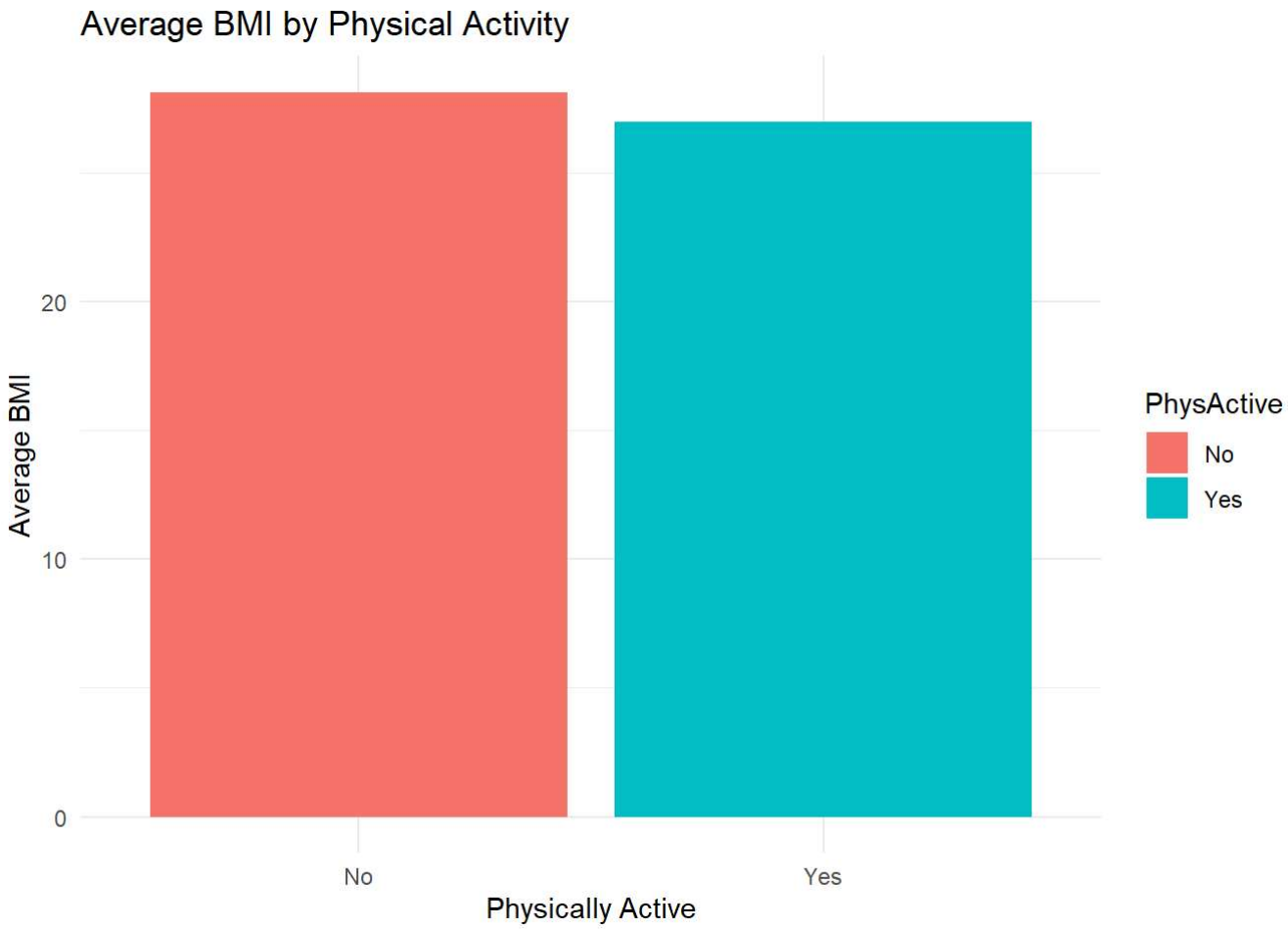
This visualization compares BMI distributions between smokers and non-smokers.

The median BMI for smokers appears slightly lower compared to non-smokers.

The spread of BMI is similar across both groups, with overlapping interquartile ranges (IQRs).

This suggests that smoking status alone may not have a strong impact on BMI, though further statistical testing is needed

```
SMOKERS %>%
  group_by(PhysActive) %>%
  summarise(Avg_BMI = mean(BMI, na.rm = TRUE)) %>%
  ggplot(aes(x = PhysActive, y = Avg_BMI, fill = PhysActive)) +
  geom_bar(stat = "identity") +
  labs(title = "Average BMI by Physical Activity",
       x = "Physically Active",
       y = "Average BMI") +
  theme_minimal()
```

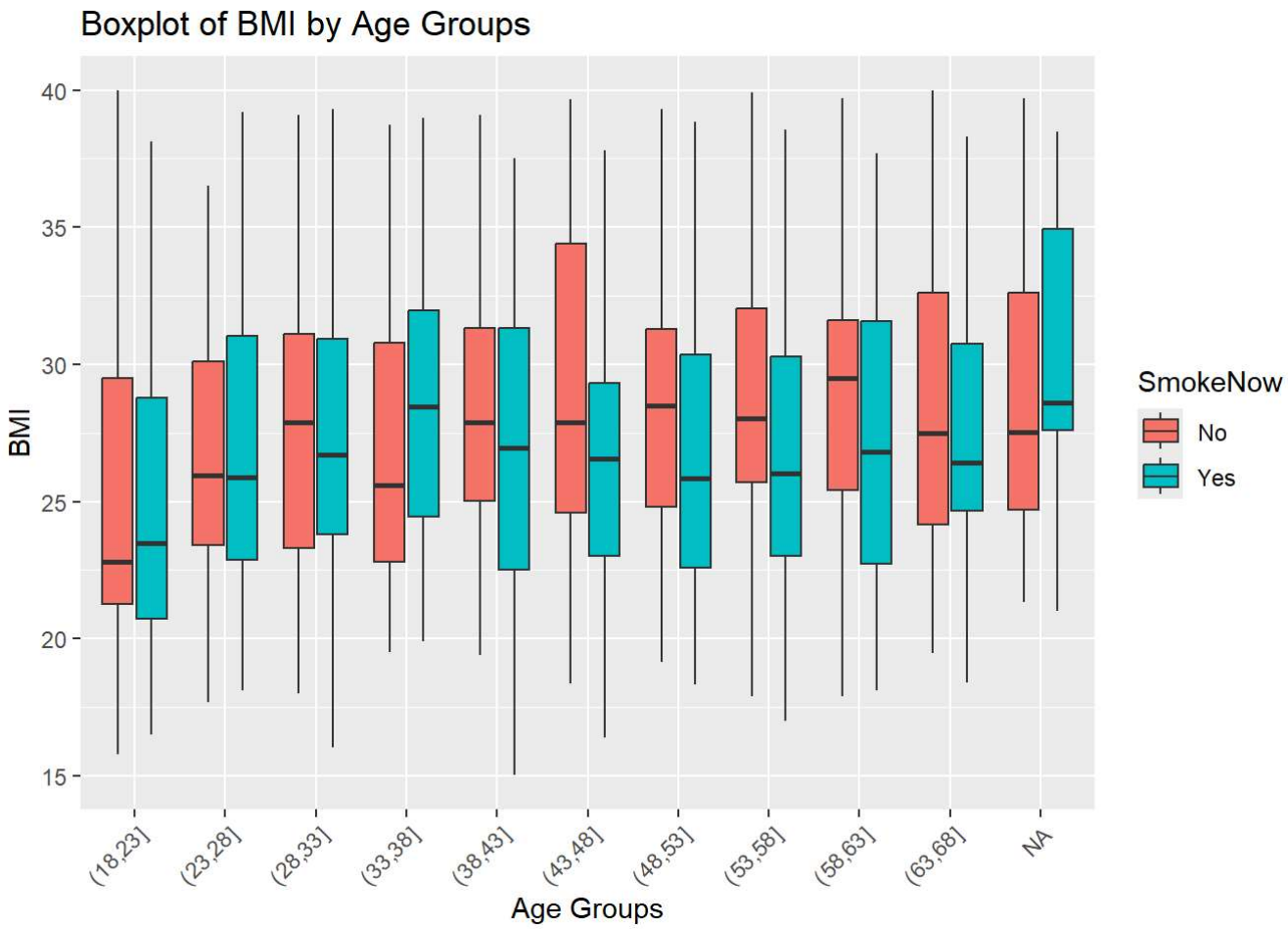


This bar chart shows the mean BMI for physically active vs. inactive individuals.

Non-active individuals tend to have slightly higher BMI on average than active individuals.

This supports the hypothesis that physical activity contributes to lower BMI, though further statistical validation is required.

```
ggplot(SMOKERS, aes(x = cut(Age, breaks = seq(18, 70, by = 5)), y = BMI, fill = SmokeNow)) +
  geom_boxplot() +
  labs(title = "Boxplot of BMI by Age Groups",
       x = "Age Groups",
       y = "BMI") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Smokers tend to have slightly lower BMI than non-smokers across age groups.

Variability in BMI increases with age, particularly among non-smokers

Model Selection & Justification

```
lm_simple <- lm(BMI ~ PhysActive, data = train_data)

summary(lm_simple)
```

```
##
## Call:
## lm(formula = BMI ~ PhysActive, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3001  -3.8301  -0.4775   3.5451  12.9451
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28.3301     0.1645 172.241  < 2e-16 ***
## PhysActiveYes  -1.3751     0.2292  -5.998 2.37e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.078 on 1962 degrees of freedom
## Multiple R-squared:  0.01801,    Adjusted R-squared:  0.01751
## F-statistic: 35.98 on 1 and 1962 DF,  p-value: 2.365e-09
```

```
train_mse_simple <- mean((train_data$BMI - predict(lm_simple, train_data))^2)
test_mse_simple <- mean((test_data$BMI - predict(lm_simple, test_data))^2)

# Print MSE values
print(train_mse_simple)
```

```
## [1] 25.75579
```

```
print(test_mse_simple)
```

```
## [1] 25.9907
```

The simple linear regression model evaluates whether physical activity significantly impacts BMI. The results indicate:

-Intercept (28.33): Represents the predicted BMI for inactive individuals.

-PhysActive Coefficient (-1.3751) Indicates that physically active individuals, on average, have a 1.38 unit lower BMI than inactive individuals, with high statistical significance ($p<0.001$).

-Residual Standard Error (5.078): Suggests that the model's predictions vary by approximately 5 BMI units.

-R^2 (0.018): The model explains 1.8% of BMI variance, indicating that physical activity alone is significant but not sufficient to explain BMI variation.

- Mean Squared Error (MSE):

  Training MSE = 25.76

  Testing MSE = 25.99

  The similarity between training and testing MSE values suggests good generalization, but BMI is influenced by additional factors beyond physical activity.

```
lm_multiple <- lm(BMI ~ Age + SmokeNow + PhysActive, data = train_data)

summary(lm_multiple)
```

```
##
## Call:
## lm(formula = BMI ~ Age + SmokeNow + PhysActive, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4217  -3.8741  -0.6066   3.5337  13.5086
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   27.856927   0.489174  56.947  < 2e-16 ***
## Age            0.027773   0.008578   3.237  0.00123 **
## SmokeNowYes   -1.388268   0.239316  -5.801 7.67e-09 ***
## PhysActiveYes -1.527168   0.233520  -6.540 7.84e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.005 on 1960 degrees of freedom
## Multiple R-squared:  0.04693,    Adjusted R-squared:  0.04547
## F-statistic: 32.17 on 3 and 1960 DF,  p-value: < 2.2e-16
```

```
# Calculating Mean Squared Error (MSE) for Training and Testing Data
train_mse_multiple <- mean((train_data$BMI - predict(lm_multiple, train_data))^2)
test_mse_multiple <- mean((test_data$BMI - predict(lm_multiple, test_data))^2)


print(train_mse_multiple)
```

```
## [1] 24.99737
```

```
print(test_mse_multiple)
```

```
## [1] 25.3672
```

The multiple regression model incorporates additional predictors. The results indicate:

- Intercept (27.86): The estimated BMI for individuals who do not smoke and are not physically active.
- Age Coefficient (0.0278): BMI increases by 0.028 units per year, suggesting a slight upward trend with age.
- SmokeNow Coefficient (-1.38826): Smokers, on average, have a 1.39 unit lower BMI than non-smokers ($p<0.001$).
- PhysActive Coefficient (-1.5272): Being physically active is associated with a 1.53 unit decrease in BMI ($p<0.001$).

- Residual Standard Error (5.005): The predictions are, on average, 5 BMI units off from actual values.
- R^2 (0.047): The model explains 4.7% of BMI variance, improving over the simple model.
- Mean Squared Error (MSE):

  Training MSE = 24.997

  Testing MSE = 25.367

Interaction model(using age and smoking status)

```
lm_interaction <- lm(BMI ~ Age * SmokeNow, data = train_data)

summary(lm_interaction)
```

```
##
## Call:
## lm(formula = BMI ~ Age * SmokeNow, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2226  -3.9186  -0.5582   3.4866  12.4728
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     25.43946    0.60283  42.200  < 2e-16 ***
## Age              0.05862    0.01199   4.888  1.1e-06 ***
## SmokeNowYes      0.83398    0.80033   1.042   0.2975
## Age:SmokeNowYes -0.04181    0.01703  -2.455   0.0142 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.051 on 1960 degrees of freedom
## Multiple R-squared:  0.02911,    Adjusted R-squared:  0.02763
## F-statistic: 19.59 on 3 and 1960 DF,  p-value: 1.631e-12
```

```
# Calculating Mean Squared Error (MSE) for Training and Testing Data
train_mse_interaction <- mean((train_data$BMI - predict(lm_interaction, train_data))^2)
test_mse_interaction <- mean((test_data$BMI - predict(lm_interaction, test_data))^2)


print(train_mse_interaction)
```

```
## [1] 25.46453
```

```
print(test_mse_interaction)
```

```
## [1] 25.22756
```

Age positively influences BMI (0.0586 per year, $p<0.001$).

Smokers have a slightly higher intercept BMI, but the effect diminishes with age ($p=0.0142$).

R^2 (0.029): The model explains 2.9% of BMI variance, indicating a weak interaction effect.

MSE:

- Training = 25.465
- Testing = 25.228

Results & Performance Evaluation

- Simple Linear Regression (BMI~PhysActive):

  This model has the highest MSE (25.99) and lowest R^2(0.018), indicating that physical activity alone is a weak predictor of BMI.

- Multiple Regression (BMI~Age+SmokeNow+PhysActive):

  This model has the lowest MSE (25.367) and highest R^2 (0.047), making it the best-performing model in explaining BMI variance.

- Interaction Model (BMI~Age*SmokeNow):

  The interaction model slightly improves on the simple model but performs worse than the multiple regression model, suggesting that the interaction effect of smoking and age is weak.

Conclusion & Discussion

- Key Insights:

  Physical activity and smoking significantly impact BMI, with active individuals and smokers having lower BMI.

  Age has a positive association with BMI, meaning BMI slightly increases as people age.

  The multiple regression model provides the best prediction accuracy, capturing more variance in BMI.

  The interaction model suggests that smoking modifies the effect of age on BMI, but the effect size is small.

- Limitations & Future Work:

  The dataset does not account for dietary habits, socioeconomic status, or genetics, which are key BMI determinants.

  More advanced models such as logistic regression or machine learning could improve predictive performance.