Project 2: Exploratory Data Analysis

Faith Adeyemi Allen

2024-11-11

To begin this project, we'll need to load the cleaned dataset from project 1, then work through a few steps

```
#Step 1. Load the cleaned dataset from project 1

setwd("C:/Users/HP/OneDrive/Documents/Portfolio")

dataset <- read.csv("Project2_ExploratoryAnalysis/data/cleaned_height_weight_data.csv")
```

Step 2: Conduct Exploratory Data Analysis (EDA): In this section, we will explore the dataset with descriptive statistics, visualization, and summaries.

#a. Descriptive Statistics

```
Get an overview of the basic statistics and structure of the dataset with summary() and str() to confirm

# Display summary statistics and structure of the dataset
print(summary(dataset))
print(str(dataset))
```

b. Visualize the Data

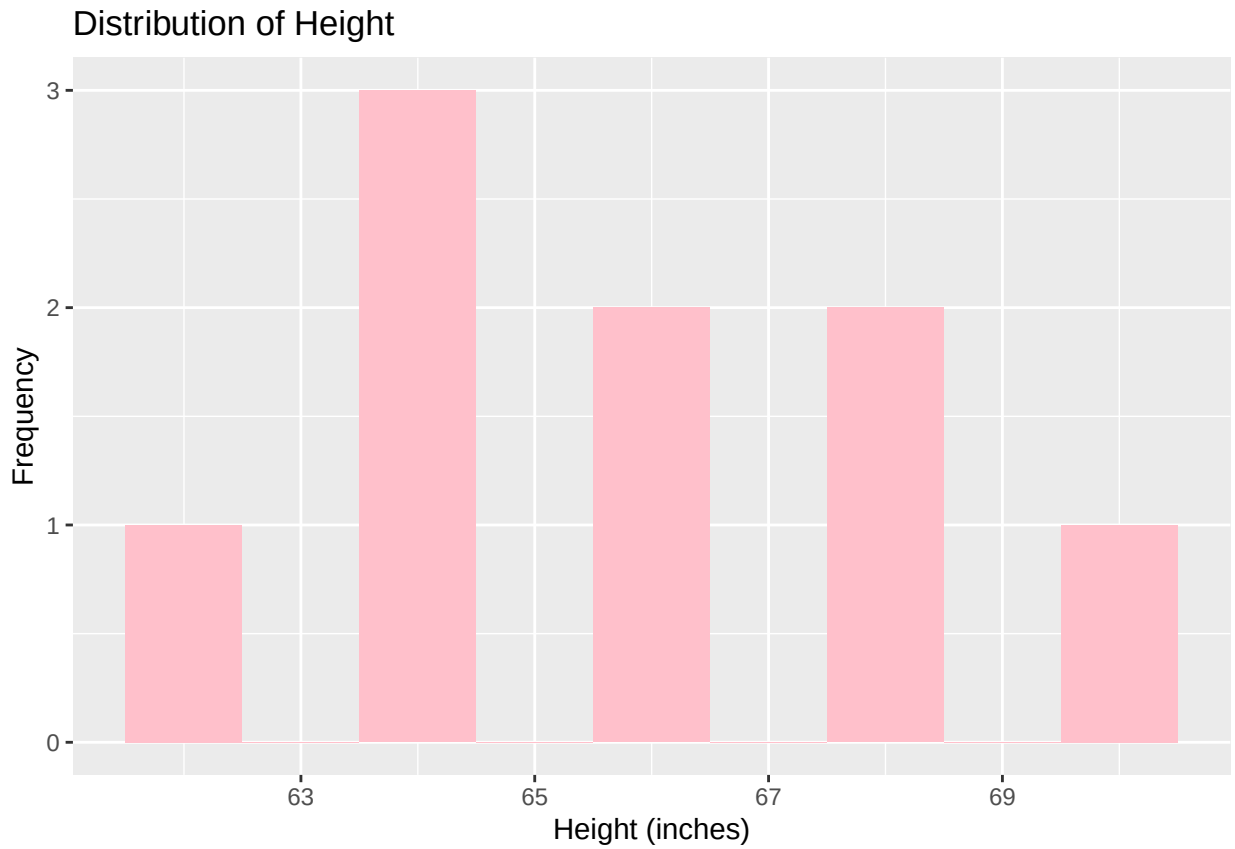
Create visualizations to identify trends and distributions in the data, such as: #Histograms for height and weight distributions #Boxplots to check for outliers #Scatter plots to analyze relationships between height and weight

```
# Visualizations and Insights
```

#1. Histogram of Height: This histogram illustrates the distribution of heights, showing a normal pattern with a central clustering of values around the mean. The majority of heights fall within a reasonable range, with no significant skew.

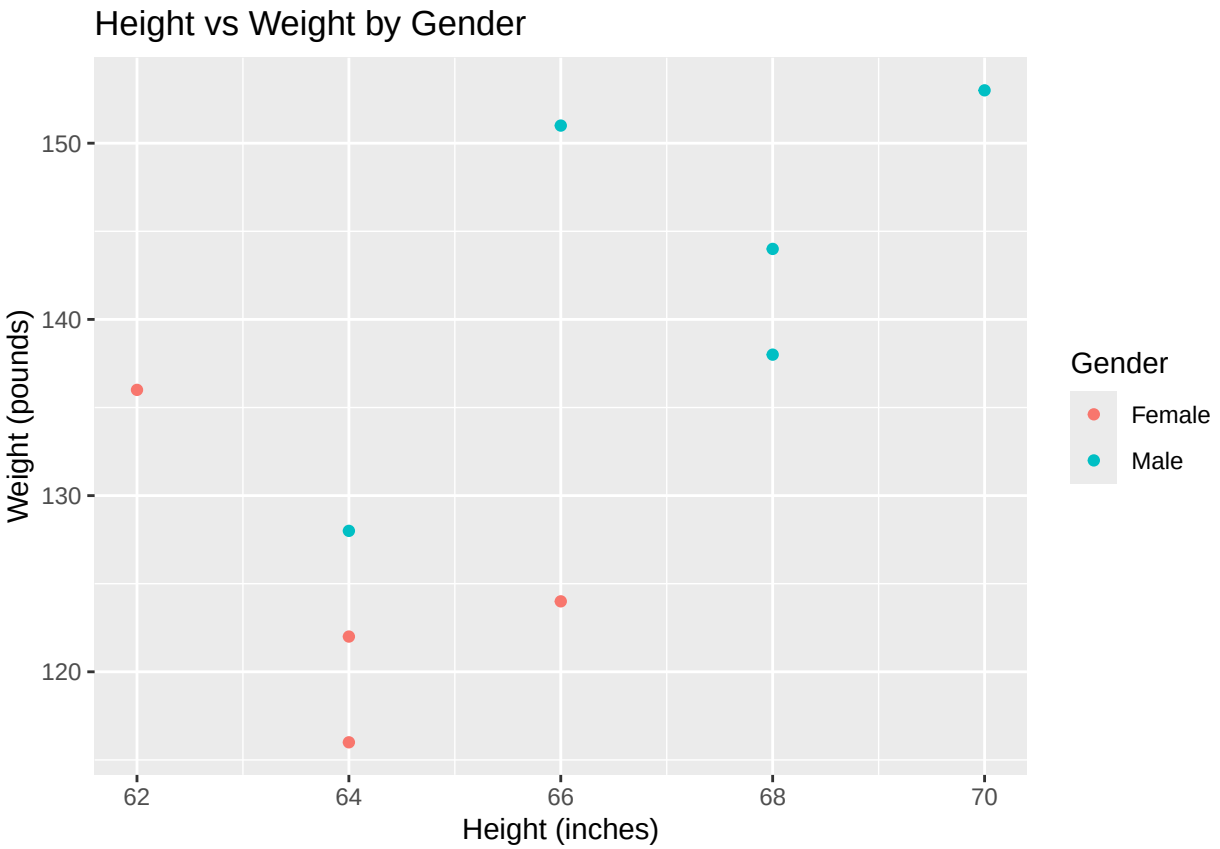
```
setwd("C:/Users/HP/OneDrive/Documents/Portfolio")
dataset <- read.csv("Project2_ExploratoryAnalysis/data/cleaned_height_weight_data.csv")
library(readxl)
library(ggplot2)

# Histogram for Height
ggplot(dataset, aes(x = Height_in_inchies)) +
  geom_histogram(binwidth = 1, fill = "pink") +
  labs(title = "Distribution of Height", x = "Height (inches)", y = "Frequency")
```



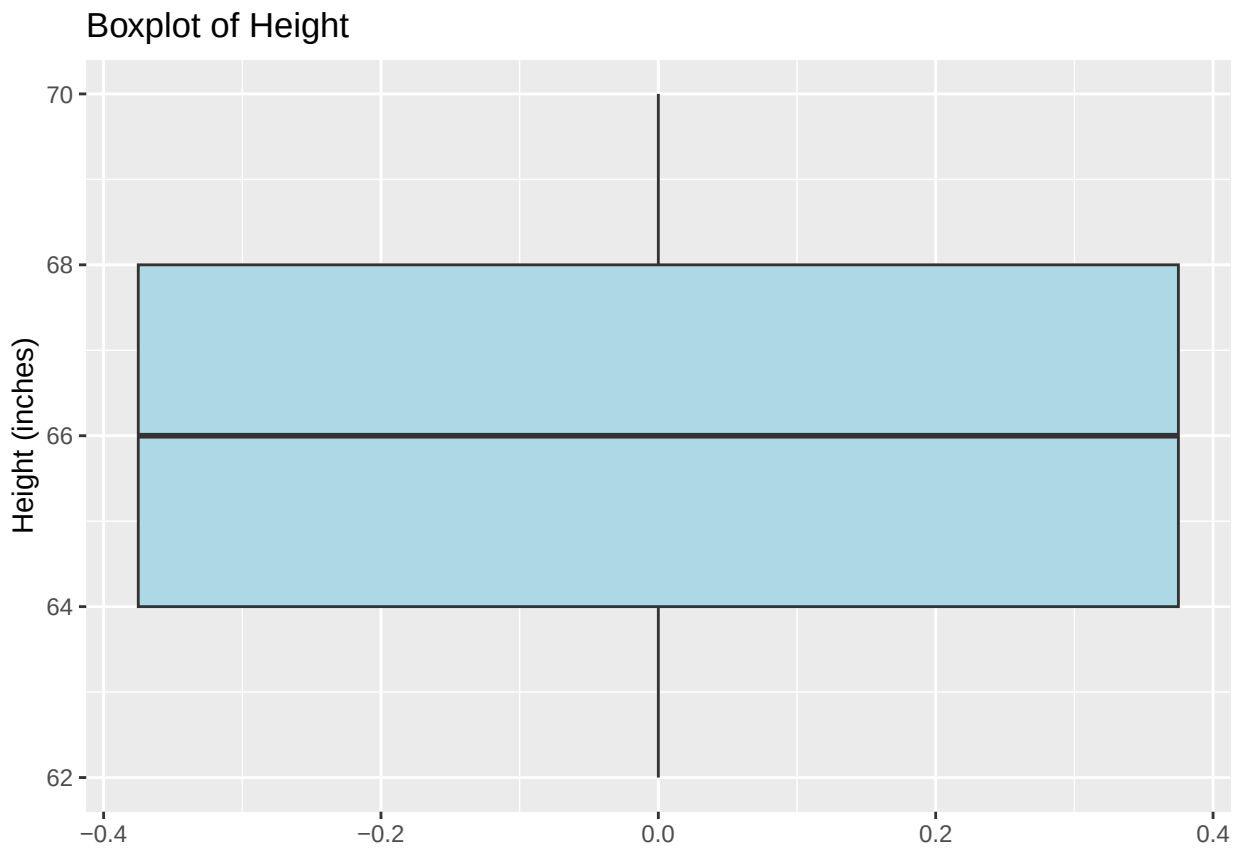
#2. Scatter plot for Height vs. Weight: This scatter plot highlights the relationship between height and weight, categorized by gender. The plot shows a positive trend, indicating that taller individuals tend to weigh more. The color coding by gender provides additional insights into differences between groups.

```
ggplot(dataset, aes(x = Height_in_inchies, y = Weight_in_pound, color = Gender)) +
  geom_point() +
  labs(title = "Height vs Weight by Gender", x = "Height (inches)", y = "Weight (pounds)")
```



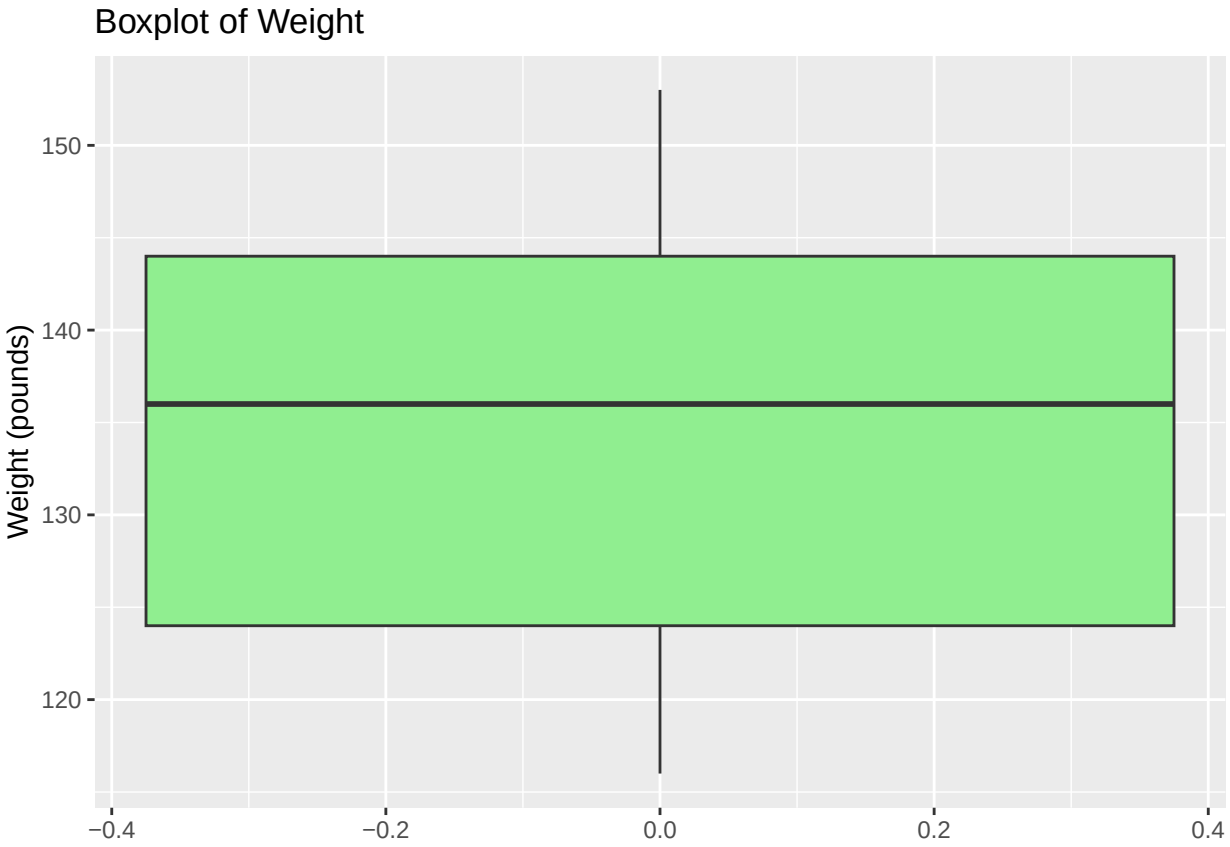
#3. Boxplot of Height: The boxplot highlights the distribution of height values and identifies potential outliers. Most of the data points are concentrated within the interquartile range, with a few values outside the whiskers.

```
# Boxplot for Height to check for outliers
ggplot(dataset, aes(y = Height_in_inchies)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot of Height", y = "Height (inches)")
```



#4. Boxplot of Weight: This boxplot provides insights into the distribution of weight values and identifies potential outliers. The majority of weights are within the interquartile range, with a few extreme values.

```
# Boxplot for Weight to check for outliers
ggplot(dataset, aes(y = Weight_in_pound)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Boxplot of Weight", y = "Weight (pounds)")
```



Summary and Conclusion

The exploratory data analysis revealed key insights into the dataset:

Heights are normally distributed, with most values clustering around the mean. A positive relationship exists between height and weight, with gender differences evident in the scatter plot. Both height and weight distributions have minimal outliers, which may require further investigation for modeling purposes. These findings confirm that the dataset is well-prepared for advanced statistical analysis and predictive modeling in future projects.