数据分析作业

张奥

2024-10-19

## Question 1: Analysis of Electricity Consumption and GDP

### 1. Initial Plot Comment

The initial scatter plot shows the relationship between electricity consumption (in billions of kilowatt-hours) and GDP (in billions of dollars). A positive correlation can be observed, suggesting that as GDP increases, electricity consumption tends to increase as well. However, there are outliers that may skew the linear relationship.

### 2. Identify Outliers and Replot

The two countries with the highest GDP are likely the United States and China. These countries may disproportionately influence the linear regression model due to their large GDP figures.

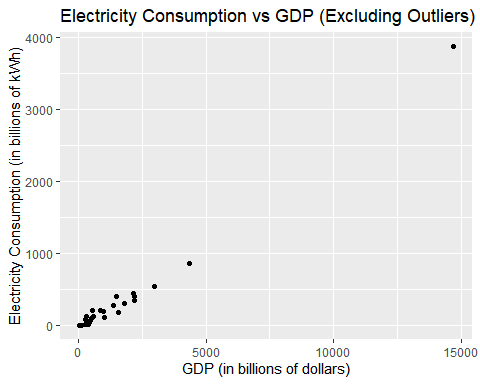
#### Replot Data

After removing the United States and China from the dataset, a new scatter plot is created to visualize the remaining countries’ electricity consumption against GDP.

# Load necessary library  
library(ggplot2)

## Warning: 程序包'ggplot2'是用R版本4.4.1 来建造的

# Load data  
data <- read.csv("electricity.csv")  
  
# Remove outliers  
data\_filtered <- data[!(data$Country %in% c("United States", "China")), ]  
  
# Create new scatter plot  
ggplot(data\_filtered, aes(x = GDP, y = Electricity)) +  
 geom\_point() +  
 labs(title = "Electricity Consumption vs GDP (Excluding Outliers)",  
 x = "GDP (in billions of dollars)",  
 y = "Electricity Consumption (in billions of kWh)")



summary("electricity.csv")

## Length Class Mode   
## 1 character character