Lecture 1: Introduction to the Assignment and Loading the Dataset

Dr. Logan Kelly

2024-09-10

Introduction

- **Objective**:
  - The goal of this assignment is to explore the relationship between **Energy Efficiency (MPG)** and **Horsepower** using simple linear regression.
  - The analysis will involve loading and cleaning a dataset, performing exploratory data analysis (EDA), building a regression model, and evaluating its assumptions.
- **Why This Is Important**:
  - Simple linear regression is one of the most commonly used statistical methods for understanding relationships between two variables.
  - In this case, we are exploring how the horsepower of a vehicle affects its fuel efficiency. This analysis has practical applications in fields such as automotive design and consumer choice, as it can inform decisions about performance and energy efficiency.
- **Key Learning Outcomes**:
  - By the end of this assignment, students will be able to:
    - Understand the overall workflow for conducting simple linear regression analysis.
    - Load data from an external source into R.
    - Preview and understand the structure of the dataset.
    - Build and interpret a simple linear regression model.

Setting Up R and Quarto

- **Introduction to R Quarto**:

  - We will be using Quarto for this assignment. Quarto allows us to combine code and narrative text in one document.
  - The structure of the document should follow clear sections, with `#` for main sections and `##` for subsections.

- **How to Structure Your R Quarto Document**:

  - Each step of the assignment will correspond to a separate section.
  - Use headings like `# Data Loading` and `## Exploratory Data Analysis` to organize your analysis.

Loading the Dataset

Checking if `openxlsx` is Installed and Loading Necessary Libraries

- **Explanation**: Before we can load the dataset, we need to check if the **`openxlsx`** package is installed. If it is not installed, we will install it. After that, we can load the package.

```r
# Check if 'openxlsx' is installed and install if necessary
if (!require(openxlsx)) {
  install.packages("openxlsx")
}
```

```
Loading required package: openxlsx
```

```r
# Load the necessary package to read Excel files
library(openxlsx)
```

Breaking Down the Code

- **`if (!require(openxlsx)) { install.packages("openxlsx") }`**: This code checks whether the **`openxlsx`** package is already installed using the `require()` function. If it's not installed, it will automatically install the package using `install.packages("openxlsx")`.
- **`library(openxlsx)`**: After checking the installation, we load the **`openxlsx`** package, which allows us to read Excel files in R.

Loading the Dataset from the Web

- **Explanation**: We will load the dataset directly from an external URL into R using the **`read.xlsx()`** function from the **`openxlsx`** package. This dataset contains car data with various attributes, including **Energy Efficiency (MPG)** and **Horsepower**.

```r
# Load the dataset from an Excel file hosted online
car_data <- read.xlsx(
    "https://ljkelly3141.github.io/real-world-statistics-with-r/data/car_price.xlsx"
    )
```

Breaking Down the Code

- **`read.xlsx()`**: This function reads the Excel file located at the provided URL and stores it in a data frame called `car_data`.

Previewing the Dataset

- **Explanation**: Once the data is loaded, it's important to inspect the structure of the dataset to understand what variables it contains.

```r
# Preview the first few rows of the dataset to ensure it's loaded correctly
head(car_data)
```

```
   Brand Model     Trim Trim.Level  Style      Size MSRP.(USD)
1 Toyota Camry       LE       Base  Sedan   Midsize      29000
2 Toyota Camry      XSE     Medium  Sedan   Midsize      34000
3 Toyota Camry   Hybrid    Premium  Sedan   Midsize      37000
4   Ford F-150      XLT       Base Pickup Full-size      52000
5   Ford F-150   Lariat     Medium Pickup Full-size      61000
6   Ford F-150 Platinum    Premium Pickup Full-size      72000
  Energy.Efficiency.(MPG) Horsepower Engine.Size.(L) Customer.Rating
1                      32        203             2.5             4.5
2                      31        301             3.5             4.7
3                      50        208             2.5             4.8
4                      20        290             3.3             4.4
5                      18        400             5.0             4.6
6                      18        400             5.0             4.8
  Safety.Rating     Hybrid     Electric Four_Wheel_Drive Sunroof Bluetooth
1             5 Non-Hybrid Non-Electric              2WD Sunroof Bluetooth
2             5 Non-Hybrid Non-Electric              2WD Sunroof Bluetooth
```

```
3             5     Hybrid         <NA>              2WD Sunroof Bluetooth
4             5 Non-Hybrid         <NA>              4WD    <NA> Bluetooth
5             5 Non-Hybrid         <NA>              4WD Sunroof Bluetooth
6             5 Non-Hybrid         <NA>              4WD Sunroof Bluetooth
  Backup_Camera   Main.Market Average.Annual.Cost.of.Ownership.(USD)
1 Backup Camera North America                                   6200
2 Backup Camera North America                                   6400
3 Backup Camera North America                                   5800
4 Backup Camera North America                                   9100
5 Backup Camera North America                                   9500
6 Backup Camera North America                                   9800
```

Breaking Down the Code

- **`head(car_data)`**: This function displays the first six rows of the dataset, providing a quick look at its structure and helping confirm that the data has been successfully loaded.

Understanding the Structure of the Data

- **Explanation**: Understanding the structure and types of variables in the dataset is crucial for the analysis. The `str()` function helps us see the data types of each column.

```
# Check the structure of the dataset to see data types and variable names
str(car_data)
```

```
'data.frame':   44 obs. of  20 variables:
 $ Brand                                 : chr  "Toyota" "Toyota" "Toyota" "Ford" ...
 $ Model                                 : chr  "Camry" "Camry" "Camry" "F-150" ...
 $ Trim                                  : chr  "LE" "XSE" "Hybrid" "XLT" ...
 $ Trim.Level                            : chr  "Base" "Medium" "Premium" "Base" ...
 $ Style                                 : chr  "Sedan" "Sedan" "Sedan" "Pickup" ...
 $ Size                                  : chr  "Midsize" "Midsize" "Midsize" "Full-size" ..
 $ MSRP.(USD)                            : num  29000 34000 37000 52000 61000 72000 53000 70
 $ Energy.Efficiency.(MPG)               : num  32 31 50 20 18 18 22 20 36 35 ...
 $ Horsepower                            : num  203 301 208 290 400 400 355 420 158 180 ...
 $ Engine.Size.(L)                       : num  2.5 3.5 2.5 3.3 5 5 5.3 6.2 2 1.5 ...
 $ Customer.Rating                       : num  4.5 4.7 4.8 4.4 4.6 4.8 4.4 4.7 4.5 4.6 ...
 $ Safety.Rating                         : num  5 5 5 5 5 5 4 4 5 5 ...
 $ Hybrid                                : chr  "Non-Hybrid" "Non-Hybrid" "Hybrid" "Non-Hybr
 $ Electric                              : chr  "Non-Electric" "Non-Electric" NA NA ...
 $ Four_Wheel_Drive                        : chr  "2WD" "2WD" "2WD" "4WD" ...
 $ Sunroof                                 : chr  "Sunroof" "Sunroof" "Sunroof" NA ...
 $ Bluetooth                               : chr  "Bluetooth" "Bluetooth" "Bluetooth" "Bluetoo
 $ Backup_Camera                           : chr  "Backup Camera" "Backup Camera" "Backup Came
 $ Main.Market                             : chr  "North America" "North America" "North Ameri
 $ Average.Annual.Cost.of.Ownership.(USD): num  6200 6400 5800 9100 9500 9800 8800 9200 5600
```

Breaking Down the Code

- `str(car_data)`: This function provides information on the data types and structure of the dataset, such as whether a variable is numeric or categorical.

Key Variables to Focus On

- **Energy Efficiency (MPG)**: This is the miles per gallon (MPG) value, representing how efficient the vehicle is with its fuel.
- **Horsepower**: The power output of the vehicle's engine, measured in horsepower (HP).

Next Steps

Now that the dataset is loaded and understood, the next lecture will focus on performing Exploratory Data Analysis (EDA) to visualize the relationship between **Energy Efficiency (MPG)** and **Horsepower**.

Assignment for Students:

- Ensure that you have successfully loaded the dataset and reviewed its structure.
- Write a short summary of the key variables and explain why they are important for this analysis.