Project 2: Exploratory Data Analysis with TidyTuesday Datasets

BYU STAT 250

Introduction

In this project you will conduct an exploratory data analysis (EDA) using a dataset from the TidyTuesday project. The goal is to explore individual variables and relationships between them. You are free to choose any TidyTuesday dataset that meets the following requirements:

- **At least 5 variables** (columns) that are either numeric or factors.
- **At least 2 numeric variables**.
- **At least 2 factor (categorical) variables**.
- Avoid variables such as names, addresses, or IDs that aren't suitable for visualization.

A data set may have more than just 5 good plotting variables. You will pick 5 out of all the variables included for the remainder of the tasks.

Note: Each TidyTuesday dataset file includes instructions on how to access and understand the data. Please review those instructions carefully.

Example TidyTuesday Datasets

Below are five TidyTuesday datasets that meet the above criteria. You do not have to use one of these, but they serve as examples:

Example TidyTuesday Datasets

Below are five TidyTuesday datasets that meet the above criteria. You do not have to use one of these, but they serve as examples:

1. **Netflix Movies and TV Shows (2020-04-21)**
   Dataset Link

   - *Numeric Variables*: `release_year`, `duration`

   - *Factor Variables*: `type` (Movie/TV Show), `rating`

2. **Coffee Ratings (2020-09-08)**
   Dataset Link

   - *Numeric Variables*: `rating`, `aroma`, `flavor` scores

   - *Factor Variables*: `country`, `variety`

3. **Bird Collisions (2019-09-24)**
   Dataset Link

   - *Numeric Variables*: e.g., `count`

   - *Factor Variables*: `species`, `location_type`

4. **Bike Share Data (2018-06-19)**
   Dataset Link

   - *Numeric Variables*: e.g., `trip_duration`, `start_hour`

   - *Factor Variables*: e.g., `user_type`, `station_category`

5. **Plastic Pollution (2020-10-20)**
   Dataset Link

   - *Numeric Variables*: e.g., `plastic_measurement`

   - *Factor Variables*: e.g., `region`, `country`

Data Analysis Tasks

Your analysis should include the following steps:

1. **Individual Variable Exploration**
   - Provide summary statistics and appropriate visualizations for each variable.
   - For numeric variables: use histograms, boxplots, etc.
   - For factor variables: use bar charts or pie charts.

2. **Pairwise Relationships**

- Examine relationships between variable pairs:
  - **Numeric-Numeric**: e.g., scatterplots with a trend line.
  - **Factor-Factor**: e.g., contingency tables or mosaic plots.
  - **Numeric vs. Factor**: e.g., boxplots or violin plots.

3. **Multi-Variable Visualizations**

   - Create at least two plots that incorporate three or more variables. For example:
     - A scatterplot with point size or color representing a third variable.
     - A faceted plot using `facet_wrap()` where panels are defined by a categorical variable.

4. **Summary of Findings**

   - Write a summary describing what your plots reveal about the relationships between the variables.
   - Ensure all plots include clear labels, legends, and well-formatted axes.

Getting Started with the Data

Below is an example of how you might load and inspect a TidyTuesday dataset using R. (Be sure to install the required packages if needed.)

```r
# Load required libraries
library(tidytuesdayR)  # For loading TidyTuesday data
library(dplyr)         # For data manipulation
library(ggplot2)       # For plotting

# Replace 'YYYY-MM-DD' with the date of the dataset you want to use.
# Example: To load the Netflix Movies and TV Shows dataset from 2020-04-21:
tues_data <- tidytuesdayR::tt_load('2020-04-21')

# View the names of the datasets included in this release
names(tues_data)
```

```
[1] "gdpr_text"       "gdpr_violations"
```

```r
# Follow the instructions in the dataset documentation to select the appropriate data frame.
# For example, if the dataset of interest is named 'netflix', you can access it via:
# netflix_data <- tues_data$netflix
```

Your Analysis Workflow

Perform your EDA by following these steps:

1. **Data Import and Cleaning**
    - Import the dataset and remove any unnecessary columns (e.g., names, IDs, addresses).
    - Check that your dataset meets the requirements (5 variables; at least 2 numeric and 2 factors).
    - If more variables are in the data set, choose 5 to use for the remainder of the assignment
2. **Exploratory Analysis**
    - **Univariate Analysis:** Generate summary statistics and visualizations for each variable.
    - **Bivariate Analysis:** Create visualizations to examine at least one relationship for each of the following pairs of variables (at least 3 plots needed for this section):
        - Two numeric variables
        - Two categorical variables
        - A numeric and a categorical variable
    - **Multi-Variable Visualization:** Create at least two plots that incorporate three or more variables. For instance:
        - Use `facet_wrap()` to create subplots based on a categorical variable.
        - Map a third variable to point size or color in a scatterplot.
3. **Summarizing Findings**
    - Write a summary of what your plots reveal about the relationships between variables.
    - Include insights drawn from both univariate and bivariate explorations.

Submission Guidelines

- Submit a PDF report generated from this QMD file.
- Ensure your code is well-commented and organized.
- Your final report should include:
    - All code used for analysis.
    - The generated visualizations.
    - A written summary of your findings.
- Collaboration is allowed. You may work individually or in groups.