Week 9 TidyVerse/GitHub CREATE assignment

Shri Tripathi

2024-24-10‘

Contents

“‘{r setup, include=FALSE}

knitr::opts_chunk$set(echo = TRUE)

Introduction

This project focuses on demonstrating the utility of the TidyVerse suite by using its packages to analyze and visualize real-world data. Specifically, we will explore the relationship between happiness levels and alcohol consumption across various countries.

Dataset Information The data used in this analysis was sourced from Kaggle:

https://www.kaggle.com/marcospessotto/happiness-and-alcohol-consumption?select=HappinessAlcoholConsumption.csv

This dataset includes 122 countries and examines factors such as happiness scores and the average consumption of beer, wine, and spirits. We’ll investigate if any patterns emerge between a country’s happiness index and its alcohol intake.

The guiding question for this analysis: *Is there a measurable relationship between alcohol consumption and happiness?*

Load Required Libraries

We begin by loading the necessary packages from the TidyVerse collection, which includes `ggplot2` for visualization, `dplyr` for data manipulation, and `tidyr` for data tidying.

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Import and Preview Data

Next, we load the dataset directly from a GitHub repository. After reading the data, we'll explore the structure by displaying the first few rows and reviewing the column names to familiarize ourselves with the data.

```
# Load the dataset from GitHub
happiness_data <- read_csv("https://raw.githubusercontent.com/Shriyanshh/Week-9-TidyVerse-GitHub-CREATE
```

```
## Rows: 122 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Country, Region, Hemisphere
## dbl (6): HappinessScore, HDI, GDP_PerCapita, Beer_PerCapita, Spirit_PerCapit...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# View column names and first six rows of data
colnames(happiness_data)
```

```
## [1] "Country"          "Region"           "Hemisphere"       "HappinessScore"
## [5] "HDI"              "GDP_PerCapita"    "Beer_PerCapita"   "Spirit_PerCapita"
## [9] "Wine_PerCapita"
```

```
head(happiness_data)
```

```
## # A tibble: 6 x 9
##   Country    Region Hemisphere HappinessScore   HDI GDP_PerCapita Beer_PerCapita
##   <chr>      <chr>  <chr>               <dbl> <dbl>         <dbl>          <dbl>
## 1 Denmark    Weste~ north                7.53   928          53.6            224
## 2 Switzerla~ Weste~ north                7.51   943          79.9            185
## 3 Iceland    Weste~ north                7.50   933          60.5            233
## 4 Norway     Weste~ north                7.50   951          70.9            169
## 5 Finland    Weste~ north                7.41   918          43.4            263
## 6 Canada     North~ north                7.40   922          42.3            240
## # i 2 more variables: Spirit_PerCapita <dbl>, Wine_PerCapita <dbl>
```

Data Transformation with `dplyr` and `tidyr`

Before visualizing the data, we perform some transformations to make it more usable. We will:

- **Create a new column** using `dplyr` to represent the total alcohol consumption per capita by summing beer, wine, and spirits consumption.

- **Handle missing values** with `tidyr` by replacing any NA values in the newly created column with 0.

```
# Use dplyr to create a new column for total alcohol consumption
happiness_data_clean <- happiness_data %>%
  mutate(TotalAlcohol = rowSums(select(., Beer_PerCapita, Spirit_PerCapita, Wine_PerCapita), na.rm = TRU

# Use tidyr to replace any missing values in the TotalAlcohol column with 0
happiness_data_clean <- happiness_data_clean %>%
  replace_na(list(TotalAlcohol = 0))

# Display the first few rows of the transformed dataset
head(happiness_data_clean)
```

```
## # A tibble: 6 x 10
##   Country    Region Hemisphere HappinessScore   HDI GDP_PerCapita Beer_PerCapita
##   <chr>      <chr>  <chr>               <dbl> <dbl>         <dbl>          <dbl>
## 1 Denmark    Weste~ north                7.53   928          53.6            224
## 2 Switzerla~ Weste~ north                7.51   943          79.9            185
## 3 Iceland    Weste~ north                7.50   933          60.5            233
## 4 Norway     Weste~ north                7.50   951          70.9            169
## 5 Finland    Weste~ north                7.41   918          43.4            263
## 6 Canada     North~ north                7.40   922          42.3            240
## # i 3 more variables: Spirit_PerCapita <dbl>, Wine_PerCapita <dbl>,
## #   TotalAlcohol <dbl>
```

Explanation of the Code:

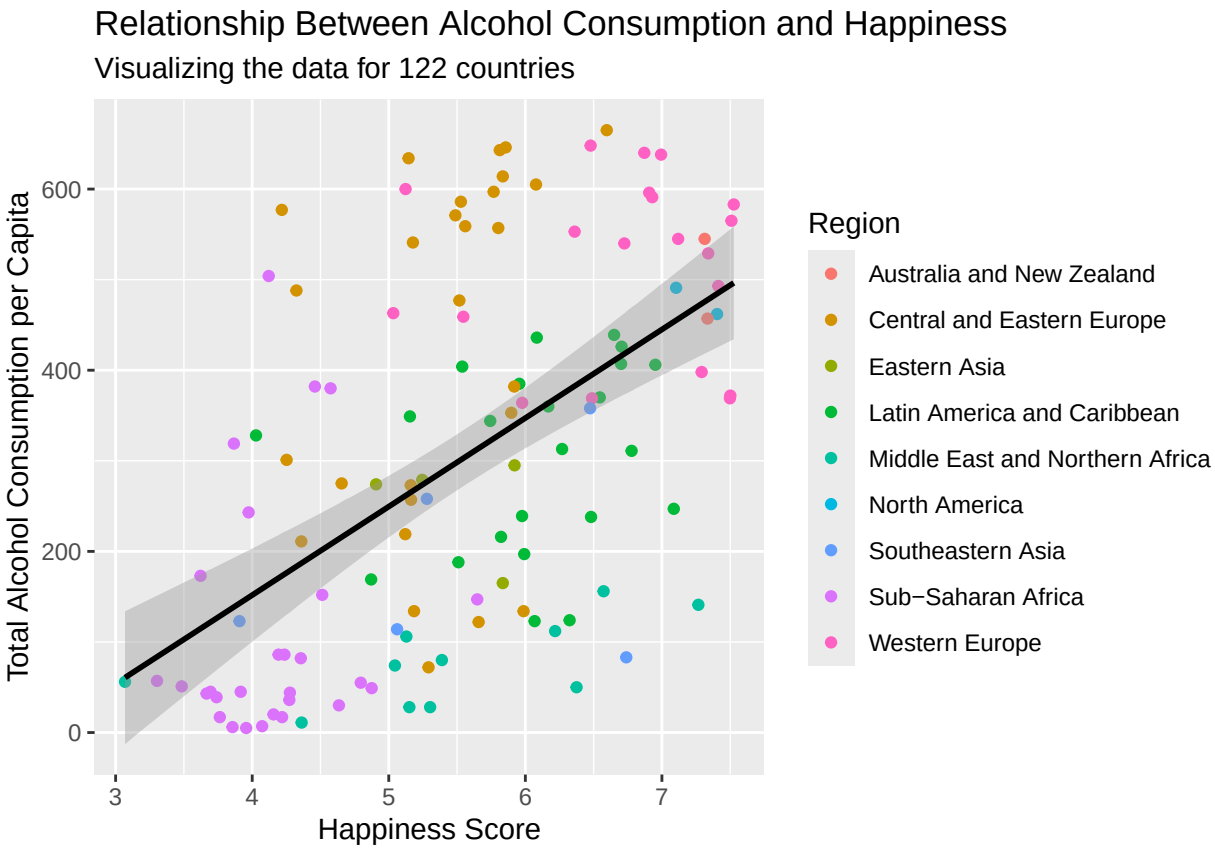
- `mutate()`: This is part of `dplyr` and allows you to create or transform columns in the dataset. Here, we use `mutate()` to create the `TotalAlcohol` column by summing the `Beer_PerCapita`, `Spirit_PerCapita`, and `Wine_PerCapita` columns using the `rowSums()` function.
- `select()`: This function, also from `dplyr`, is used inside `rowSums()` to select the relevant columns (beer, spirit, and wine consumption).
- `replace_na()`: This function from `tidyr` is used to replace missing values in the `TotalAlcohol` column. We specify `na.rm = TRUE` in the `rowSums()` to ignore any NA values during the summation process, but in case there are any remaining NAs in the dataset, we replace them with 0 using `replace_na()`.

Visualizing the Relationship with ggplot2

Next, we create a scatter plot to visualize the relationship between total alcohol consumption and happiness scores. Each country is colored according to its region, and a linear regression line is added to show the overall trend.

```
# Create a scatter plot using ggplot2
ggplot(happiness_data_clean, aes(x = HappinessScore, y = TotalAlcohol, color = Region)) +
  geom_point() +
  labs(
    title = "Relationship Between Alcohol Consumption and Happiness",
    subtitle = "Visualizing the data for 122 countries",
    x = "Happiness Score",
    y = "Total Alcohol Consumption per Capita"
  ) +
  geom_smooth(method = "lm", color = "black")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interpretation of Results

The scatter plot above suggests a positive association between happiness and alcohol consumption. However, there is substantial variability in the data, implying that while a trend exists, it is not strongly deterministic.

Further analysis would be required to explore factors such as income levels, social structures, or cultural practices that may also influence these observations.

Conclusion

This example has demonstrated the power of TidyVerse for cleaning, manipulating, and visualizing data. In particular, we used:

- `dplyr` for creating new variables and manipulating the dataset
- `tidyr` for handling missing values
- `ggplot2` for visualizing the relationship between happiness and alcohol consumption

To conclude, while we can see a general trend, the data reveals substantial variation between different regions, necessitating further exploration to draw robust conclusions.