Bike Sharing

2024-09-16

Introduction

For this assignment, I chose the Bike Sharing dataset, which provides a comprehensive log of bike-sharing rentals collected from the Capital Bikeshare system in Washington D.C. over two years, 2011 and 2012. The rental data is combined with environmental and seasonal factors that influence bike usage. Variables like temperature, humidity, wind speed, weather conditions, and whether the day is a holiday or weekend are provided.

```
# Download the dataset
download.file('https://archive.ics.uci.edu/static/public/275/bike+sharing+dataset.zip',
              'bike-sharing.zip')
unzip('bike-sharing.zip', files = 'hour.csv')

# Load the dataset
df <- read.csv('hour.csv', header = TRUE)

# Remove zip and csv files
file.remove(list.files(pattern = "\\.zip$|\\.csv$"))
```

Problem 1. Summary Statistics Table

The dataset contains 17 columns and 17379 obs. Below are a brief overview of the columns

Variable	Role	Type	Description
instant	Feature	Integer	Record index
dteday	Feature	Date	Date of the record
season	Feature	Categorical	Season (1 = Spring, 2 = Summer, 3 = Fall, 4 = Winter)
yr	Feature	Categorical	Year (0 = 2011, 1 = 2012)
mnth	Feature	Categorical	Month (1 to 12)
hr	Feature	Categorical	Hour of the day (0 to 23)
holiday	Feature	Categorical	Whether the day is a holiday (1 = Yes, 0 = No)
weekday	Feature	Categorical	Day of the week (0 = Sunday, 1 = Monday, . . . , 6 = Saturday)
workingday	Feature	Categorical	Whether the day is a working day (1 = Yes, 0 = No)
weathersit	Feature	Categorical	Weather situation (1 = Clear, 2 = Mist, 3 = Light Snow/Rain, 4 = Heavy Rain/Snow)
temp	Feature	Continuous	Normalized temperature in Celsius (values divided by 41)
atemp	Feature	Continuous	Normalized feeling temperature in Celsius (values divided by 50)
hum	Feature	Continuous	Normalized humidity (values divided by 100)
windspeed	Feature	Continuous	Normalized wind speed (values divided by 67)
casual	Other	Integer	Count of casual (unregistered) users
registered	Other	Integer	Count of registered users

Variable	Role	Type	Description
cnt	Target	Integer	Total count of bike rentals, including both casual and registered users

This dataset is suitable for both classification and regression tasks. Key questions it can help address include:

- How can bike-sharing systems be optimized by identifying peak usage periods?
- How can planners prepare for special events or holidays?
- How does the weather impact bike rentals?

Problem 2. Bad Data Visualization

2.1. Categorical Variable (season)

If we make a Pie Chart of Total Bike Rentals by Season (category variable), it becomes difficult to compare exact differences between the seasons by comparing angles, which is harder for the human eye than comparing lengths (as in bar charts). In contrast, a bar chart would clearly show the differences by the heights of the bars, refer 3.1

```
ggplot(df, aes(x = factor(1), fill = factor(season))) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Pie Chart of Total Bike Rentals by Season", fill = "Season")
```

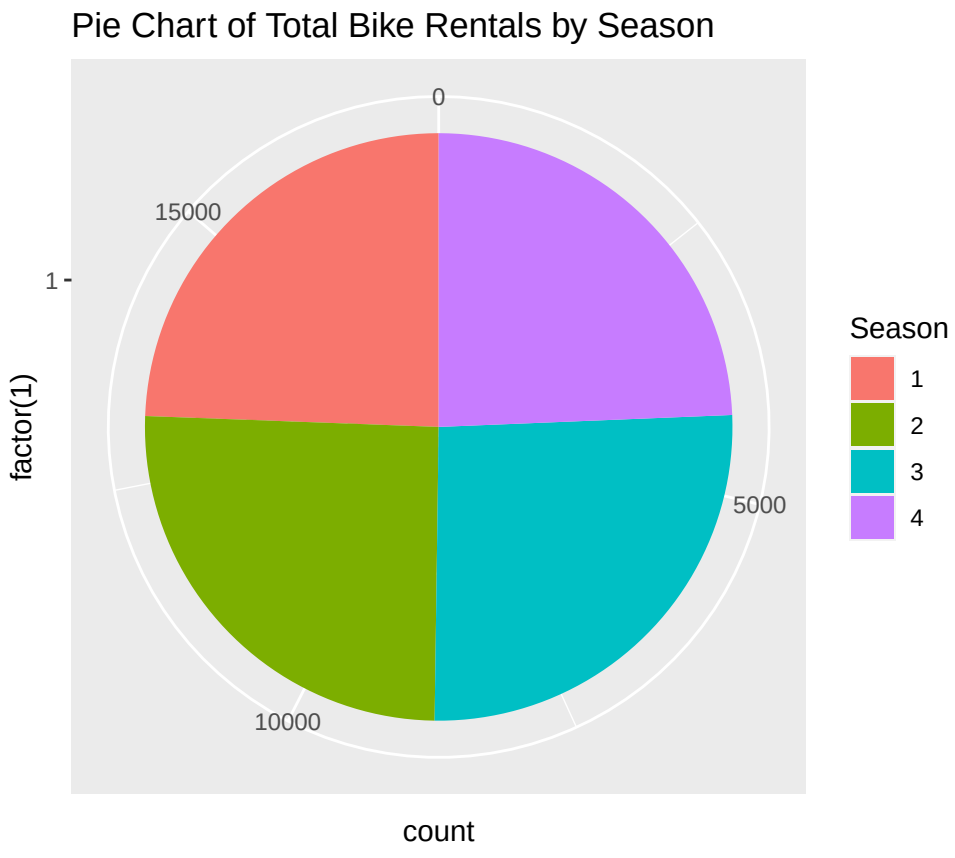


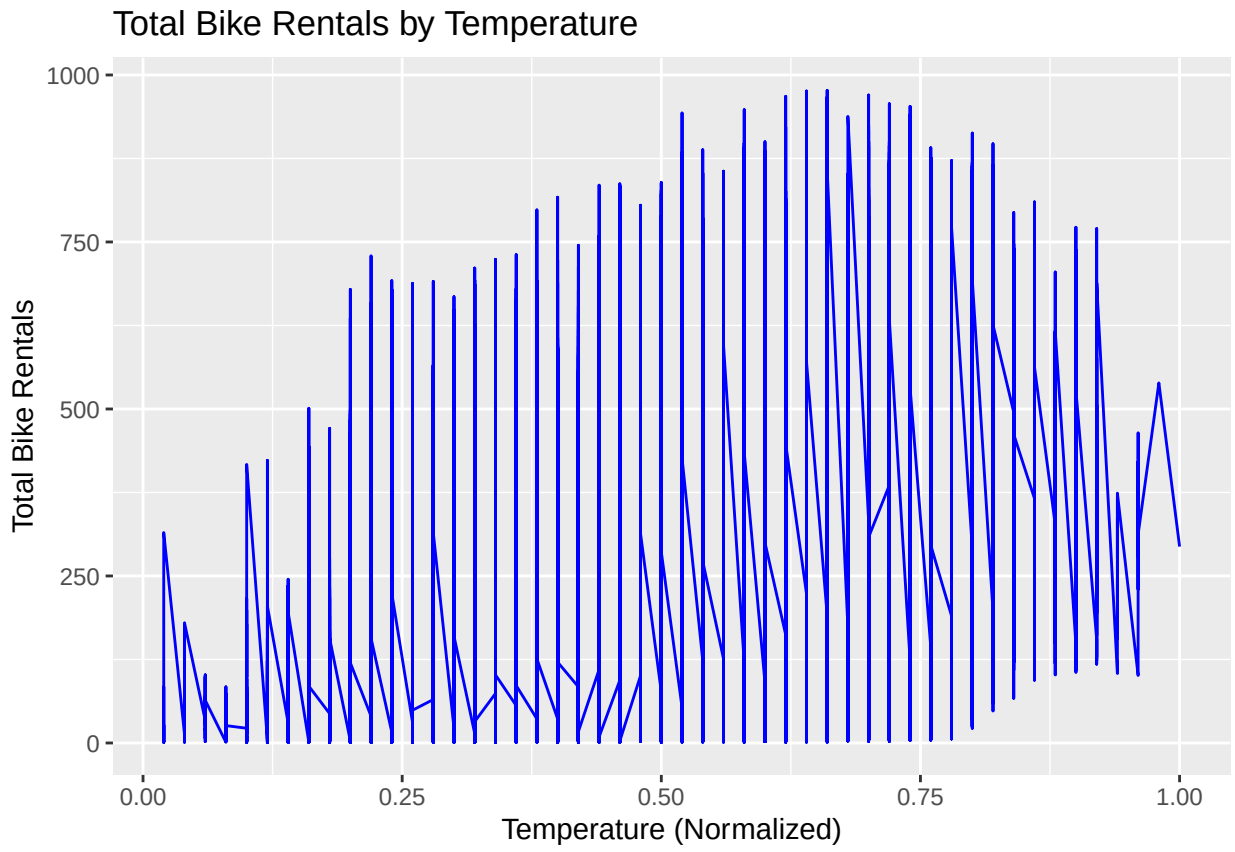
Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
instant	17379	8690	5017	1	4346	13034	17379
season	17379	2.5	1.1	1	2	3	4
yr	17379	0.5	0.5	0	0	1	1
mnth	17379	6.5	3.4	1	4	10	12
hr	17379	12	6.9	0	6	18	23
holiday	17379	0.029	0.17	0	0	0	1
weekday	17379	3	2	0	1	5	6
workingday	17379	0.68	0.47	0	0	1	1
weathersit	17379	1.4	0.64	1	1	2	4
temp	17379	0.5	0.19	0.02	0.34	0.66	1
atemp	17379	0.48	0.17	0	0.33	0.62	1
hum	17379	0.63	0.19	0	0.48	0.78	1
windspeed	17379	0.19	0.12	0	0.1	0.25	0.85
casual	17379	36	49	0	4	48	367
registered	17379	154	151	0	34	220	886
cnt	17379	189	181	1	40	281	977

2.2. Continous Variable (temp)

In below example, Line chart of Total Bike Rentals by Temperature gives the false impression of a sequential or time-based relationship and it can not show a clear trend, so this visualization offers little meaningful insight into how bike rentals respond to temperature changes.

```
ggplot(df, aes(x = temp, y = cnt)) +
  geom_line(color = "blue") +
  labs(title = "Total Bike Rentals by Temperature",
       x = "Temperature (Normalized)",
       y = "Total Bike Rentals")
```

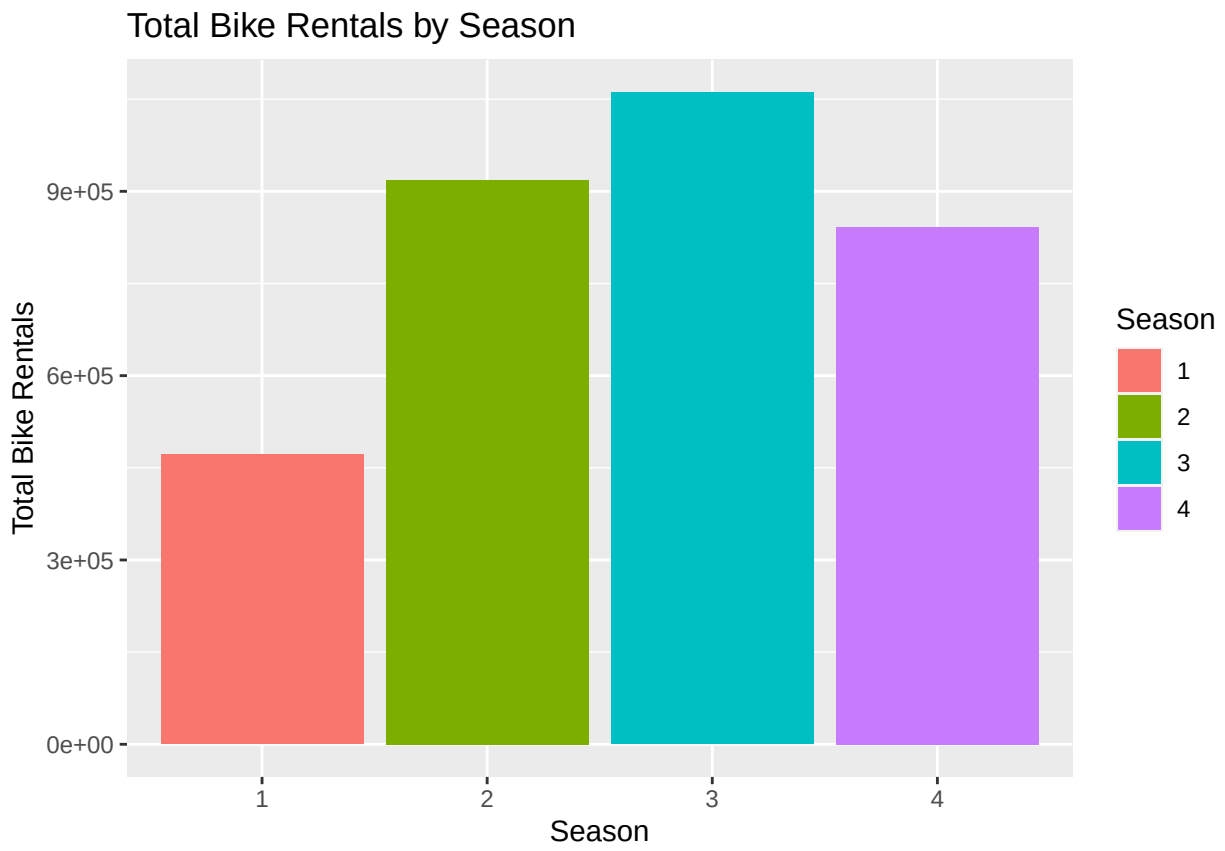


Problem 3. Good Data Visualization

3.1. Categorical Variable

For the chart shown in section 2.1, bar charts are much better for comparing categorical variables like seasons, as they show the total bike rentals for each season clearly. And we can compare the height of the bars to see which season has the highest bike rentals. In addition, the y-axis starts at zero, providing an accurate sense of scale and preventing any distortion.

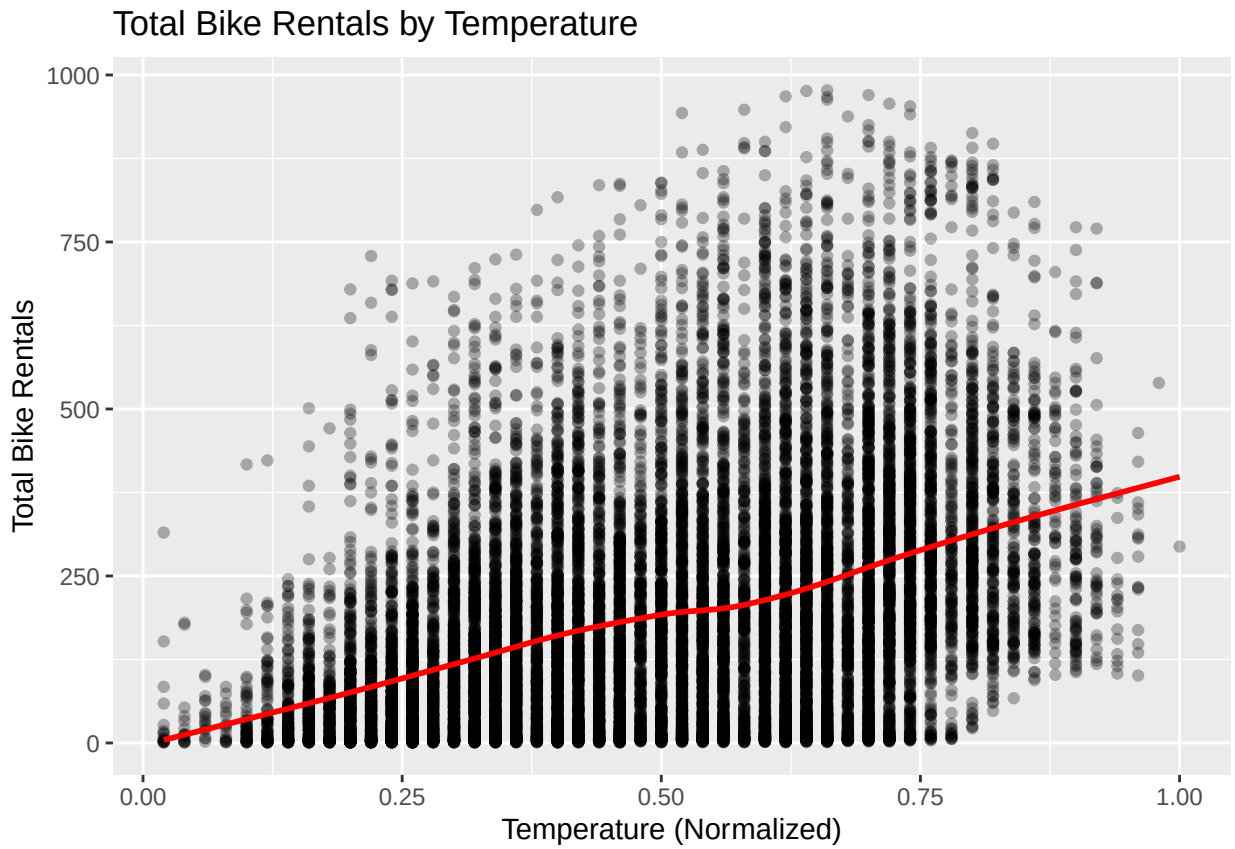
```
ggplot(df, aes(x = factor(season), y = cnt, fill = factor(season))) +
  geom_bar(stat = "summary", fun = "sum") +
  labs(title = "Total Bike Rentals by Season",
       x = "Season",
       y = "Total Bike Rentals",
       fill = "Season")
```



3.2. Continous Variable

The scatter plot is better suited for continuous data like temperature and rentals. It shows the spread and density of the data points without implying a sequential connection. On the other hand, adding a smoothing line (LOESS) or trend line gives a clear indication of the relationship between temperature and total bike rentals, helping to highlight any trends in the data.

```
ggplot(data = df, aes(x = temp, y = cnt)) +
  geom_point(alpha = 0.3) +
  geom_smooth(formula = y ~ x, method = "loess", color = "red", se = FALSE) +
  labs(title = "Total Bike Rentals by Temperature",
       x = "Temperature (Normalized)",
       y = "Total Bike Rentals")
```



Problem 4. Simple analysis

TBD