homework01

zhewei xie

2024-07-18

Problem 1: playlists revisited

This problem use observed proportions/frequencies to approximate probabilities.

Part A

In the 2x2 table (Table 1) displayed below:
it shows P(plays Daft Punk | plays David Bowie) **as the bottom right entry,**
shows P(plays Daft Punk | not plays David Bowie) as the bottom left entry,
shows P(not plays Daft Punk | plays David Bowie) as the upper right entry,
shows P(not plays Daft Punk | not plays David Bowie) as the upper left entry.

Table 1: Conditional probabilities of "plays Daft Punk" and "plays David Bowie".

	not plays David Bowie	plays David Bowie
not plays Daft Punk	0.925	0.912
plays Daft Punk	0.075	0.088

Part B

Just like the frequency of Johnny Cash (Table 2) and the 2x2 table of conditional probabilities (Table 3) displayed below, it shows that:
P(plays Johnny Cash | plays Pink Floyd) ≈ 0.105,
P(plays Johnny Cash | not plays Pink Floyd) ≈ 0.055,
P(plays Johnny Cash) ≈ 0.060.
Therefore:
P(plays Johnny Cash | plays Pink Floyd) $\neq P$(plays Johnny Cash | not plays Pink Floyd) $\neq P$(plays Johnny Cash),
this implies that **the events "plays Johnny Cash" and "plays Pink Floyd" are not independent.**

Table 2: Frequency of "plays Johnny Cash".

	Freq
not plays Johnny Cash	0.940
plays Johnny Cash	0.060

Table 3: Conditional probabilities of "plays Johnny Cash" and "plays Pink Floyd".

	not plays Pink Floyd	plays Pink Floyd
not plays Johnny Cash	0.945	0.895
plays Johnny Cash	0.055	0.105

Problem 2: Super Bowl ads

This problem use observed proportions/frequencies to approximate probabilities.

Part A

Estimate $P($danger = TRUE$)$

Because $P(\text{danger = TRUE}) \approx \{\text{Frequency of danger = TRUE}\}$.
Therefore, given the frequency of danger = TRUE (Table 4) displayed below, $P(\text{danger = TRUE}) \approx 0.30$.

Table 4: Frequency of danger.

	Freq
danger = FALSE	0.70
danger = TRUE	0.30

Estimate $P($danger = TRUE $\mid$ funny = TRUE$)$

Because $P(\text{danger = TRUE} \mid \text{funny = TRUE})$

$= \frac{P(\text{danger = TRUE, funny = TRUE})}{P(\text{funny = TRUE})}$

$\approx \frac{\text{Frequency of danger = TRUE and funny = TRUE both happening}}{\text{Frequency of funny = TRUE happening}}$.

Therefore, given the 2x2 table of conditional probabilities (Table 5) displayed below:
$P(\text{danger = TRUE} \mid \text{funny = TRUE}) \approx 0.39$.

Estimate $P($danger = TRUE $\mid$ funny = FALSE$)$

Because $P(\text{danger = TRUE} \mid \text{funny = FALSE})$

$= \frac{P(\text{danger = TRUE, funny = FALSE})}{P(\text{funny = FALSE})}$

$\approx \frac{\text{Frequency of danger = TRUE and funny = FALSE both happening}}{\text{Frequency of funny = FALSE happening}}$.

Therefore, given the 2x2 table of conditional probabilities (Table 5) displayed below:
$P(\text{danger = TRUE} \mid \text{funny = FALSE}) \approx 0.12$.

Table 5: Conditional probabilities of danger and funny.

	funny = FALSE	funny = TRUE
danger = FALSE	0.88	0.61
danger = TRUE	0.12	0.39

Explanation

Given the results of estimations of $P(\text{danger} = \text{TRUE})$, $P(\text{danger} = \text{TRUE} \mid \text{funny} = \text{TRUE})$ and $P(\text{danger} = \text{TRUE} \mid \text{funny} = \text{FALSE})$, it is clear that
$P(\text{danger} = \text{TRUE}) \neq P(\text{danger} = \text{TRUE} \mid \text{funny} = \text{TRUE}) \neq P(\text{danger} = \text{TRUE} \mid \text{funny} = \text{FALSE})$. **In light of these numbers, it seem that ads using humor are more likely to feature danger than ads not using humor, and under the given condition, humor and danger are not independent.**

Part B

Estimate P(animals = TRUE)

Because $P(\text{animals} = \text{TRUE}) \approx \{\text{Frequency of animals} = \text{TRUE}\}$.
Therefore, given the frequency of animals = TRUE (Table 6) displayed below, $P(\text{animals} = \text{TRUE}) \approx 0.37$.

Table 6: Frequency of animals.

	Freq
animals = FALSE	0.63
animals = TRUE	0.37

Estimate P(animals = TRUE | use_sex = TRUE)

Because $P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{TRUE})$

$= \frac{P(\text{animals} = \text{TRUE, use\_sex} = \text{TRUE})}{P(\text{use\_sex} = \text{TRUE})}$

$\approx \frac{\text{Frequency of animals = TRUE and use\_sex = TRUE both happening}}{\text{Frequency of use\_sex = TRUE happening}}$.

Therefore, given the 2x2 table of conditional probabilities (Table 7) displayed below:
$P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{TRUE}) \approx 0.38$.

Estimate P(animals = TRUE | use_sex = FALSE)

Because $P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{FALSE})$

$= \frac{P(\text{animals} = \text{TRUE, use\_sex} = \text{FALSE})}{P(\text{use\_sex} = \text{FALSE})}$

$\approx \frac{\text{Frequency of animals = TRUE and use\_sex = FALSE both happening}}{\text{Frequency of use\_sex = FALSE happening}}$.

Therefore, given the 2x2 table of conditional probabilities (Table 7) displayed below:
$P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{FALSE}) \approx 0.37$.

Table 7: Conditional probabilities of animals and sexuality.

	use_sex = FALSE	use_sex = TRUE
animals = FALSE	0.63	0.62
animals = TRUE	0.37	0.38

Explanation

Given the results of estimations of $P(\text{animals} = \text{TRUE})$, $P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{TRUE})$ and $P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{FALSE})$, it is clear that
$P(\text{animals} = \text{TRUE}) \approx P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{TRUE}) \approx P(\text{animals} = \text{TRUE} \mid \text{use\_sex} = \text{FALSE})$, **so it may seem that ads using sexuality are as likely to feature animals as ads not using sexuality. This implies that the events using sexuality and featuring animals are independent.**

Part C

Estimate P(celebrity = TRUE)

Because $P(\text{celebrity} = \text{TRUE}) \approx \{\text{Frequency of celebrity} = \text{TRUE}\}$.
Therefore, given the frequency of celebrity = TRUE (Table 8) displayed below, $P(\text{celebrity} = \text{TRUE}) \approx 0.29$.

Table 8: Frequency of celebrity.

	Freq
celebrity = FALSE	0.71
celebrity = TRUE	0.29

Estimate P(celebrity = TRUE | patriotic = TRUE)

Because $P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{TRUE})$

$$= \frac{P(\text{celebrity} = \text{TRUE, patriotic} = \text{TRUE})}{P(\text{patriotic} = \text{TRUE})}$$

$$\approx \frac{\text{Frequency of celebrity} = \text{TRUE and patriotic} = \text{TRUE both happening}}{\text{Frequency of patriotic} = \text{TRUE happening}}.$$

Therefore, given the 2x2 table of conditional probabilities (Table 9) displayed below:
$P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{TRUE}) \approx 0.29$.

Estimate P(celebrity = TRUE | patriotic = FALSE)

Because $P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{FALSE})$

$$= \frac{P(\text{celebrity} = \text{TRUE, patriotic} = \text{FALSE})}{P(\text{patriotic} = \text{FALSE})}$$

$$\approx \frac{\text{Frequency of celebrity} = \text{TRUE and patriotic} = \text{FALSE both happening}}{\text{Frequency of patriotic} = \text{FALSE happening}}.$$

Therefore, given the 2x2 table of conditional probabilities (Table 9) displayed below:
$P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{FALSE}) \approx 0.29$.

Table 9: Conditional probabilities of celebrity and patriotic.

	patriotic = FALSE	patriotic = TRUE
celebrity = FALSE	0.71	0.71
celebrity = TRUE	0.29	0.29

Explanation

Given the results of estimations of $P(\text{celebrity} = \text{TRUE})$, $P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{TRUE})$ and $P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{FALSE})$, it is clear that
$P(\text{celebrity} = \text{TRUE}) \approx P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{TRUE}) \approx P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{FALSE})$, **so it seem that quite credibly ads using patriotic symbolism are as likely to feature celebrity endorsement as ads not using patriotic symbolism. This implies that the events using patriotic symbolism and featuring celebrity endorsement are independent.**

Problem 3: Beauty, or not, in the classroom

Part A

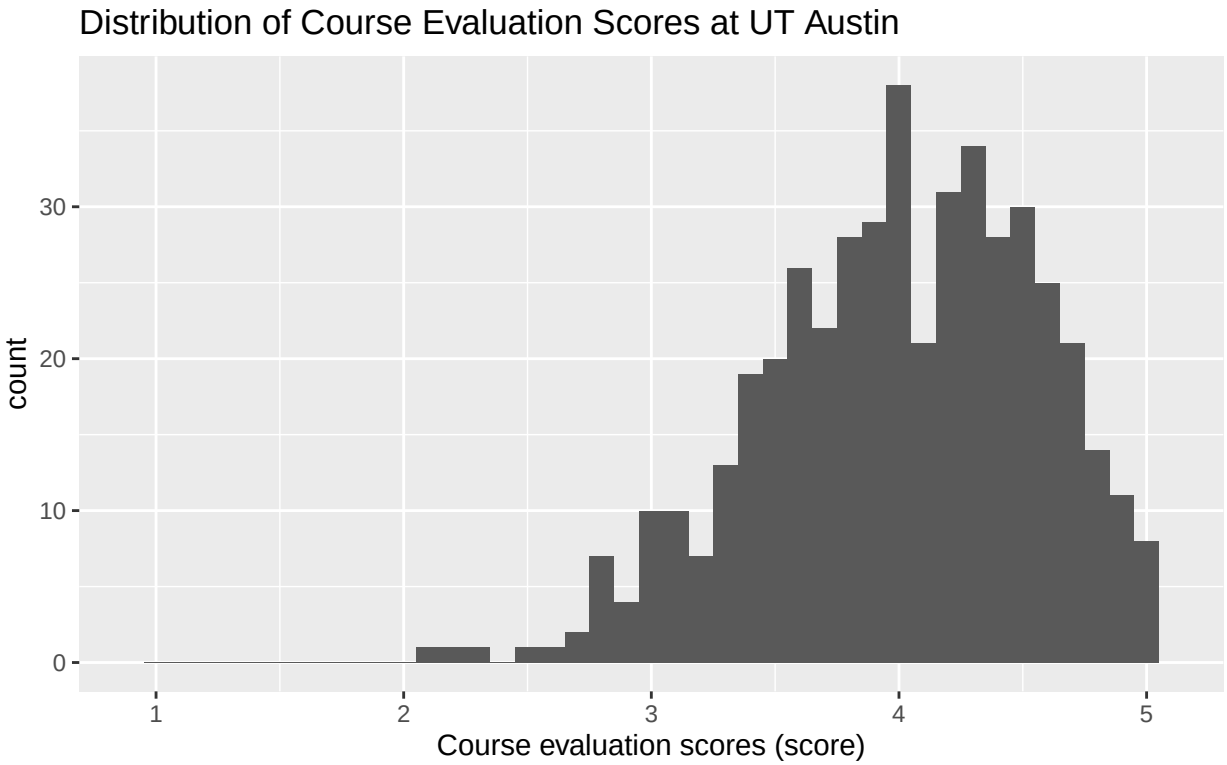

Figure 1. The distribution of course evaluation scores is left−skewed and concentrated around 4.0, indicating that the overall rating for the professor is generally positive. And the median of the distribution is 4.0, while the iqr is 0.8.

Part B

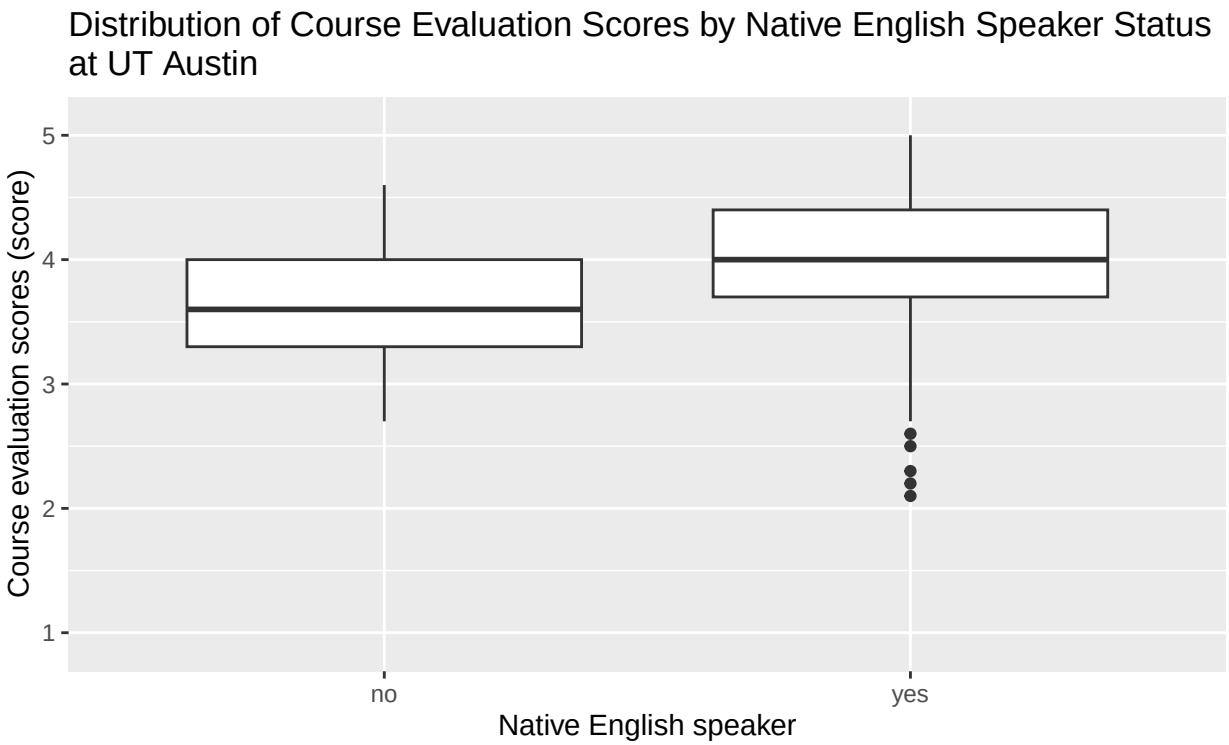

Figure 2. As English native speakers, instructors generally receive higher course evaluation scores (median = 4), but the differences in extreme cases are more pronounced (max = 5.0, min = 2.1). Non−native speakers, on the other hand, tend to have lower overall course evaluation scores (median = 3.6), but these scores are relatively clustered (max = 4.6, min = 2.7).

Part C

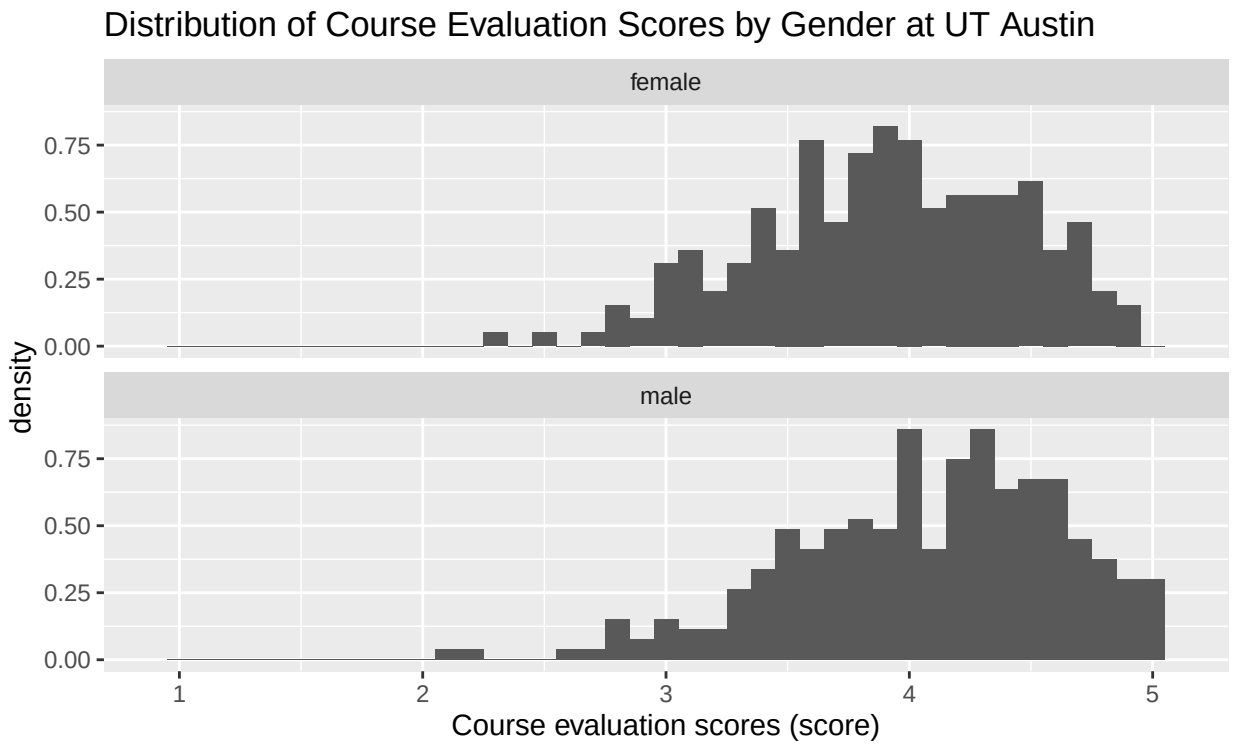

Figure 3. Both of distributions of course evalution scores faceted by gender are left−skewed, while the median course evaluation scores for male instructors (4.15) are higher than those for female instructors (3.9 However, the IQR value for males (0.8) is also higher compared to females (0.7), indicating that the scores for male instructors are more spread out or variable.

Part D

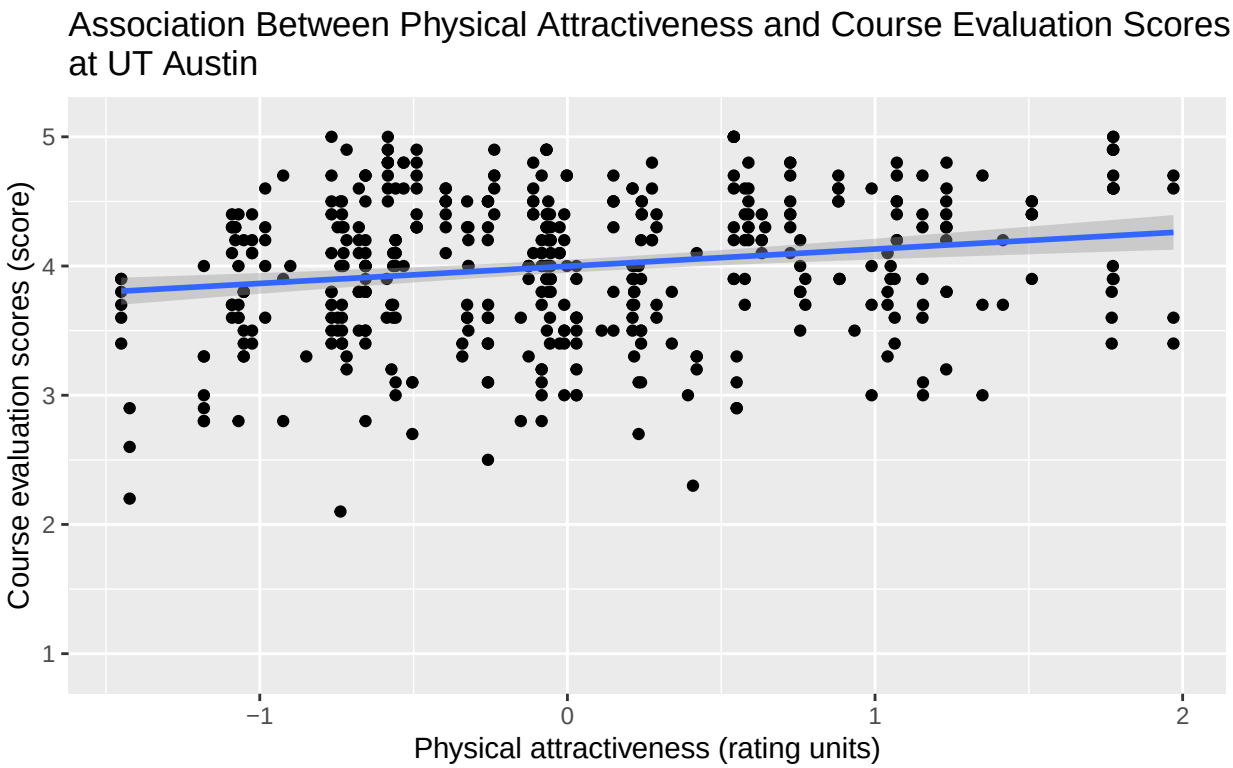

Figure 4. Regardless of the rating of physical attractiveness scores, course evaluation scores shows great spread. And there is a linear correlation (correlation = 0.189) between a professor´s physical attractive− ness and their course evaluations.

Problem 4: SAT scores for UT students

Table 10: The table shows the mean, standard deviation, inter-quartile rang, 5th percentile, 25th, percentile, median (50 percentile), 75th percentile, and 95th percentile of the SAT Verbal, SAT Quantitative, and GPA (all summaries rounded to 2 decimal places). The data contains the SAT scores and graduating college GPAs for every UT student who entered UT in a specific, recent year, and went on to graduate from UT within 6 years.

variables	avg	sd	iqr	q05	q25	q50	q75	q95
SAT.V	595.00	84.00	110.00	460.00	540.00	590.00	650.00	730.00
SAT.Q	620.00	83.00	120.00	480.00	560.00	620.00	680.00	760.00
GPA	3.21	0.48	0.72	2.36	2.87	3.25	3.59	3.92

Problem 5: bike sharing

Plot A

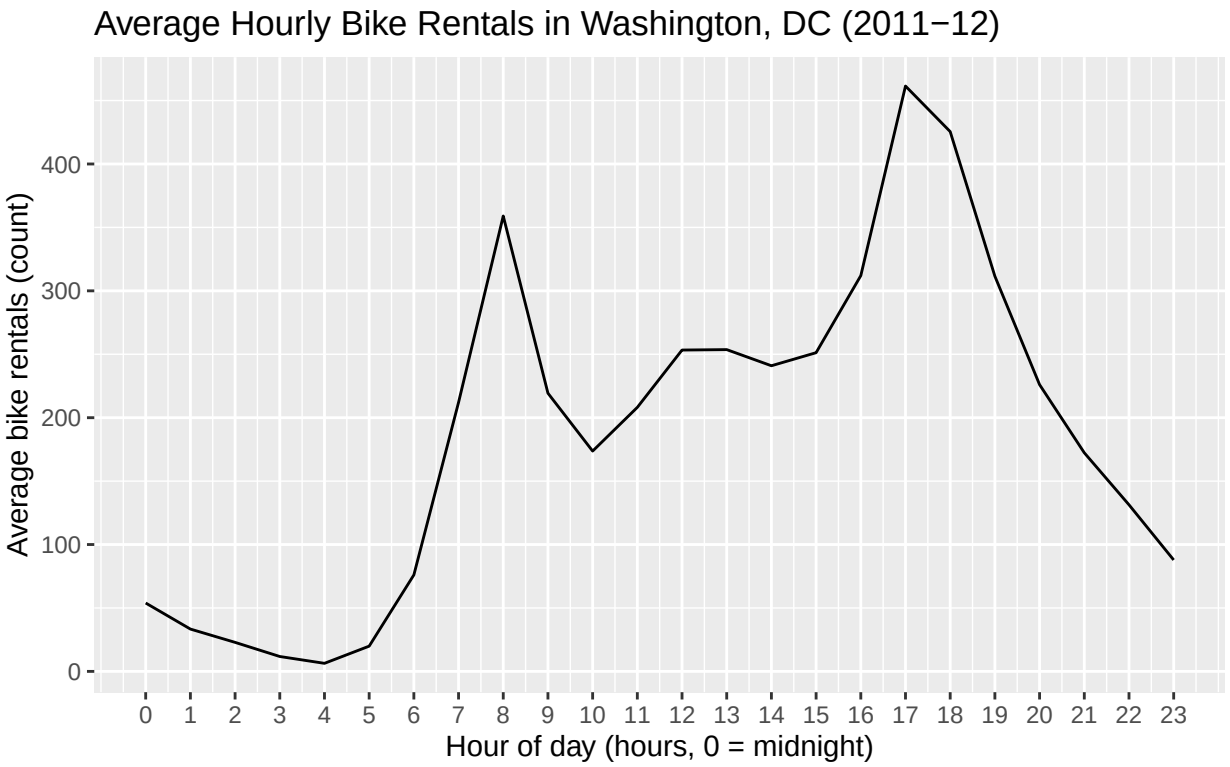

Figure 5. Average hourly bike rentals in Washington, DC (2011−12). The distribution is bimodal, with peak during the morning and evening rush hours. Meanwhile, the number of rentals during the evening rush hour is significantly higher than during the morning rush hour.

Note: The significantly higher number of bicycle rentals during the evening peak compared to the morning peak implies that there may be other influencing factors. The subsequent Plot B, which provides statistics on whether the rentals occur on working days, offers a possible explanation.

Plot B

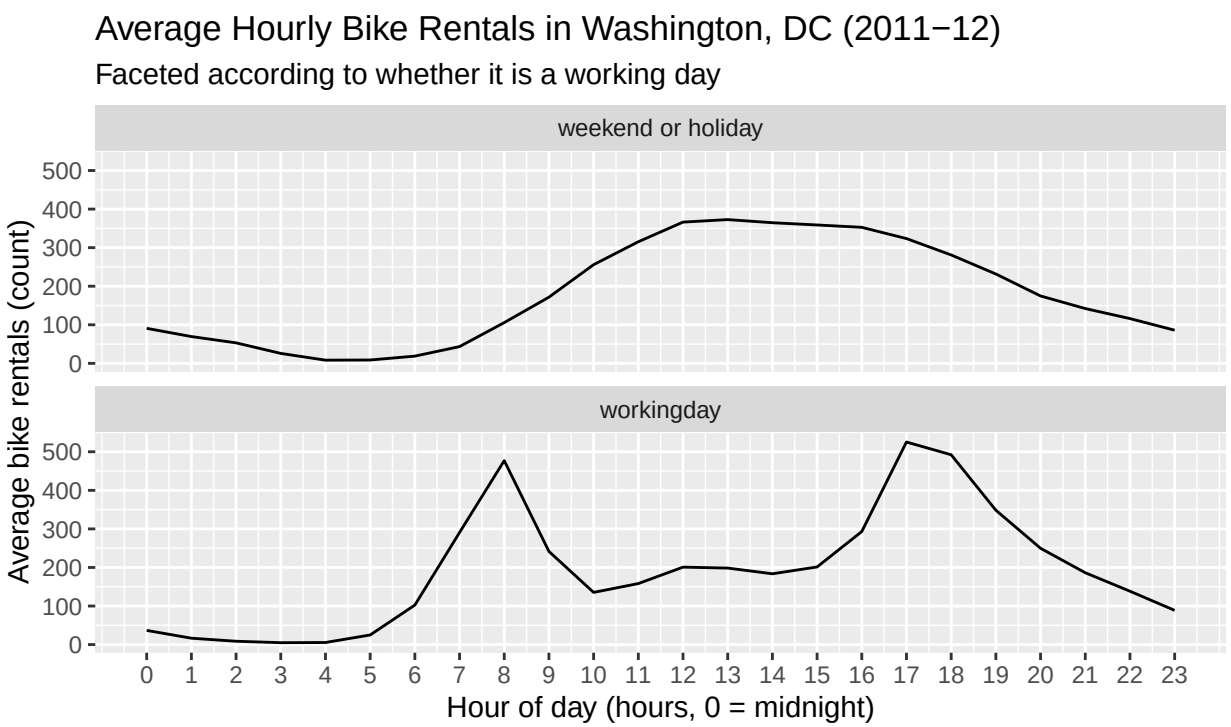

Figure 6. Average hourly bike rentals in Washington, DC (2011−2012) grouped by whether it is a working day or not. During working days, bike rentals peak during the morning and evening rush hours, and the number of rentals is quite similar during these times. In contrast, on holidays or weekends, the distribution of bike rentals is a gentle hill−like curve, and the number of bicycle rentals is not very high in the morning, increases from noon to afternoon, and gradually decreases in the evening.

Plot C

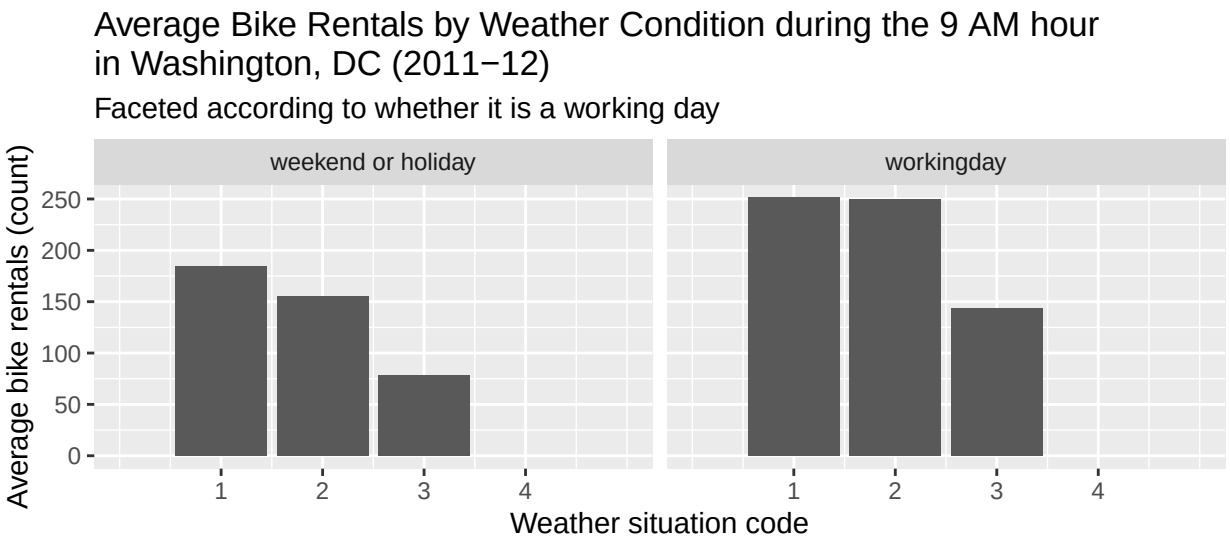

Figure 7. Average bike rentals grouped by weather condition and working day status during the 9 AM in Washington, DC (2011−12). The number of bike rentals on working days is generally higher than on non−working days. As the weather conditions worsen, people gradually reduce their bike rentals, but they tend to tolerate bad weather more on working days than on non−working days. This is evident as bike rentals on working days are almost the same under weather conditions 1 and 2, whereas they decrease on non−working days. In the worst weather condition (weather situation code 4), people do not choose to travel by renting bikes at all.

x−axis represent weather situation code, which has the following values:
− 1: Clear, Few clouds, Partly cloudy, Partly cloudy
− 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
− 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
− 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog