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 $P(\text{plays Daft Punk} \mid \text{plays David Bowie})$ **as the bottom right entry,**
shows $P(\text{plays Daft Punk} \mid \text{not plays David Bowie})$ as the bottom left entry,
shows $P(\text{not plays Daft Punk} \mid \text{plays David Bowie})$ as the upper right entry,
shows $P(\text{not plays Daft Punk} \mid \text{not plays David Bowie})$ 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:
$P(\text{plays Johnny Cash} \mid \text{plays Pink Floyd}) \approx 0.105$,
$P(\text{plays Johnny Cash} \parallel \text{not plays Pink Floyd}) \approx 0.055$,
$P(\text{plays Johnny Cash}) \approx 0.060$.
Therefore:
$P(\text{plays Johnny Cash} \mid \text{plays Pink Floyd}) \neq P(\text{plays Johnny Cash} \mid \text{not plays Pink Floyd}) \neq P(\text{plays Johnny Cash})$,
this implies that **the events "plays Johnny Cash" and "plays Pink Floyd" 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 $P(\text{danger} = \text{TRUE}) \approx \{\text{Frequency of danger} = \text{TRUE}\}$.
Therefore, given the frequency of danger = TRUE (Table 4) displayed below, $P(\text{danger} = \text{TRUE}) \approx 0.30$.

Table 4: Frequency of danger

	Freq
danger = FALSE	0.70
danger = TRUE	0.30

Estimate P(danger = TRUE | funny = TRUE)

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

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

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

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

Estimate P(danger = TRUE | funny = FALSE)

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

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

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

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

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 $P(\text{danger} = \text{TRUE})$, $P(\text{danger} = \text{TRUE}|\text{funny} = \text{TRUE})$ and $P(\text{danger} = \text{TRUE} \mid \text{funny} = \text{FALSE})$, it is clear that
$P(\text{danger} = \text{TRUE}) \neq P(\text{danger} = \text{TRUE} \mid \text{funny} = \text{TRUE}) \neq P(\text{danger} = \text{TRUE} \mid \text{funny} = \text{FALSE})$, **so humor and danger look nearly independent of each other.** In the other words, it seem that the use of humor in ads does not have a clear relationship with wheather the ads likely to feature danger.

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

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

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

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

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

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

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 seem that may ads using sexuality are more likely to feature animals than ads not using sexuality.**

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, $P(\text{celebrity} = \text{TRUE}) \approx 0.29$.

	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:
$P(\text{celebrity} = \text{TRUE} \mid \text{patriotic} = \text{FALSE}) \approx 0.29$.

	patriotic = FALSE	patriotic = TRUE
celebrity = FALSE	0.71	0.71

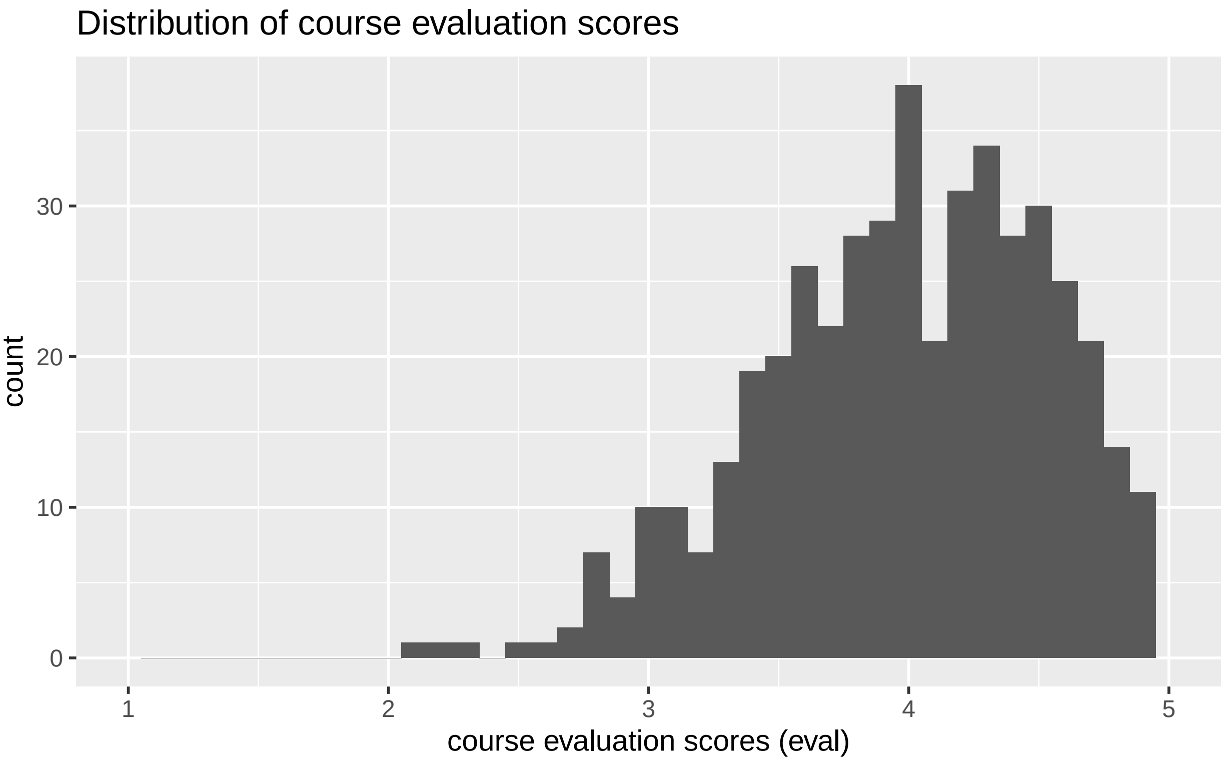
	patriotic = FALSE	patriotic = TRUE
celebrity = TRUE	0.29	0.29

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 more likely to feature celebrity endorsement than ads not using patriotic symbolism.**

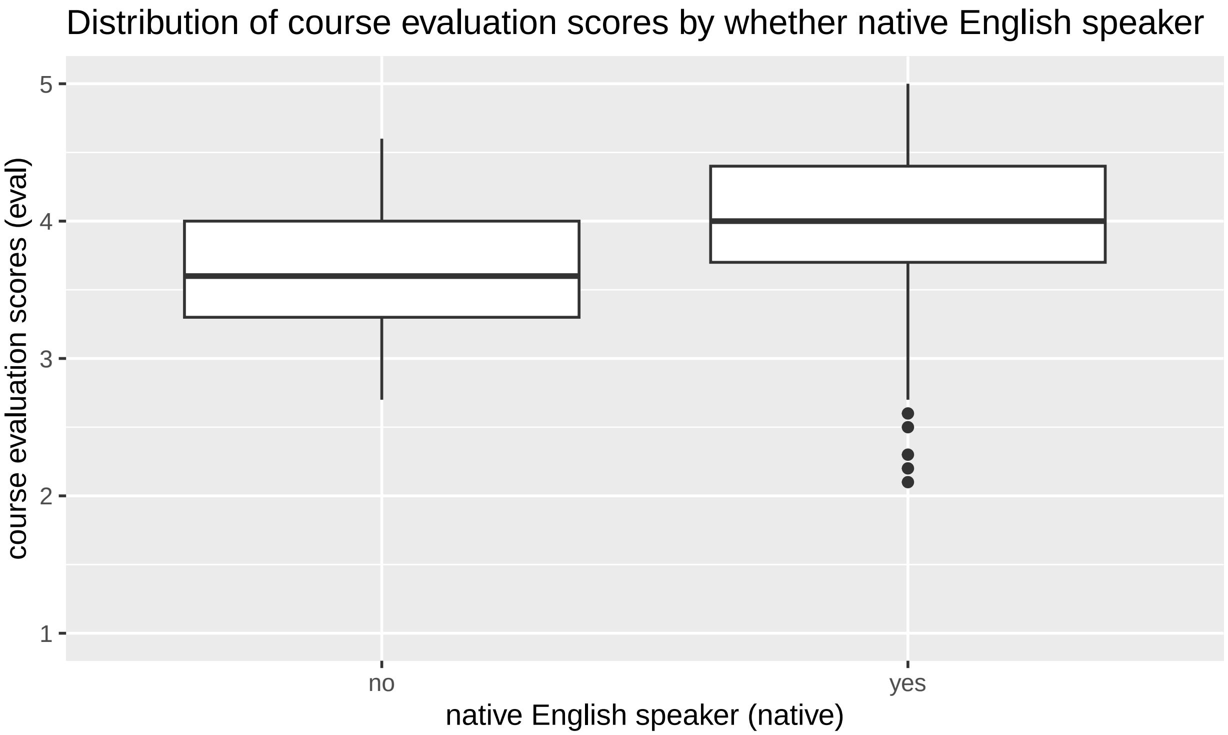
Problem 3: Beauty, or not, in the classroom

Part A



The distribution of course evaluation scores is left−skewed and nearly unimodal, with the mode being 4.0. And the median of eval distribution is 4, while the iqr is 0.8.

Part B

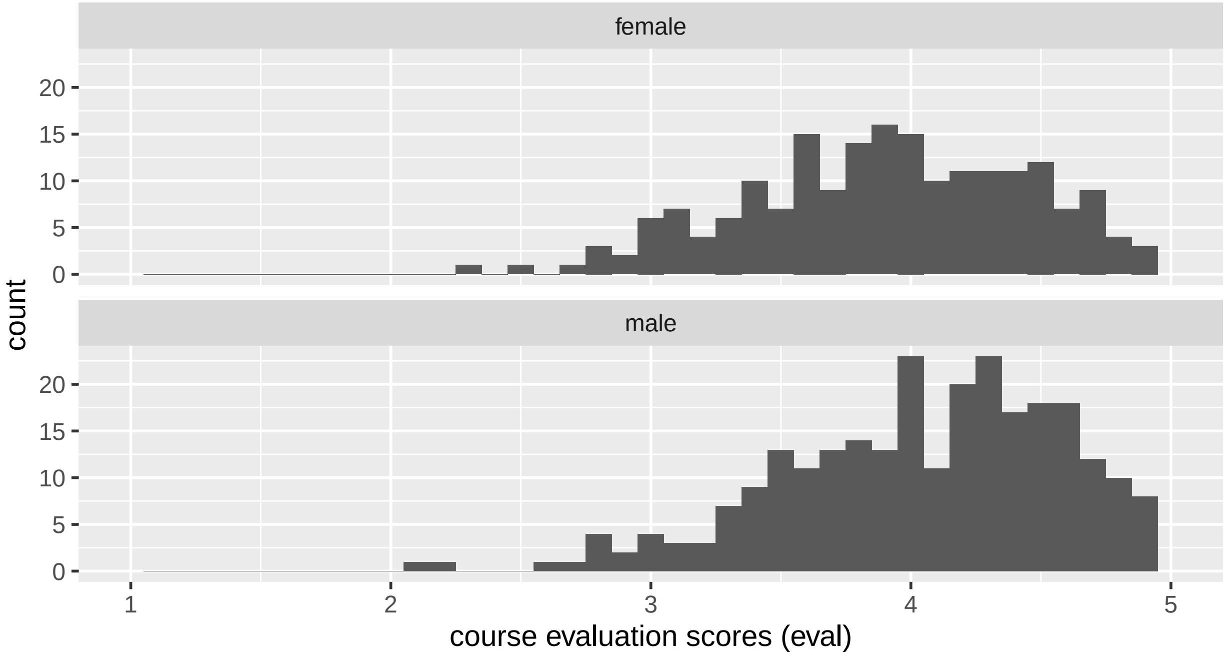


Distribution of course evaluation scores by whether native English speaker

As English native speakers, instructors generally receive higher course evaluation scores, but the differences in extreme cases are more pronounced. Non−native speakers, on the other hand, tend to have lower overall course evaluation scores, but these scores are relatively clustered.

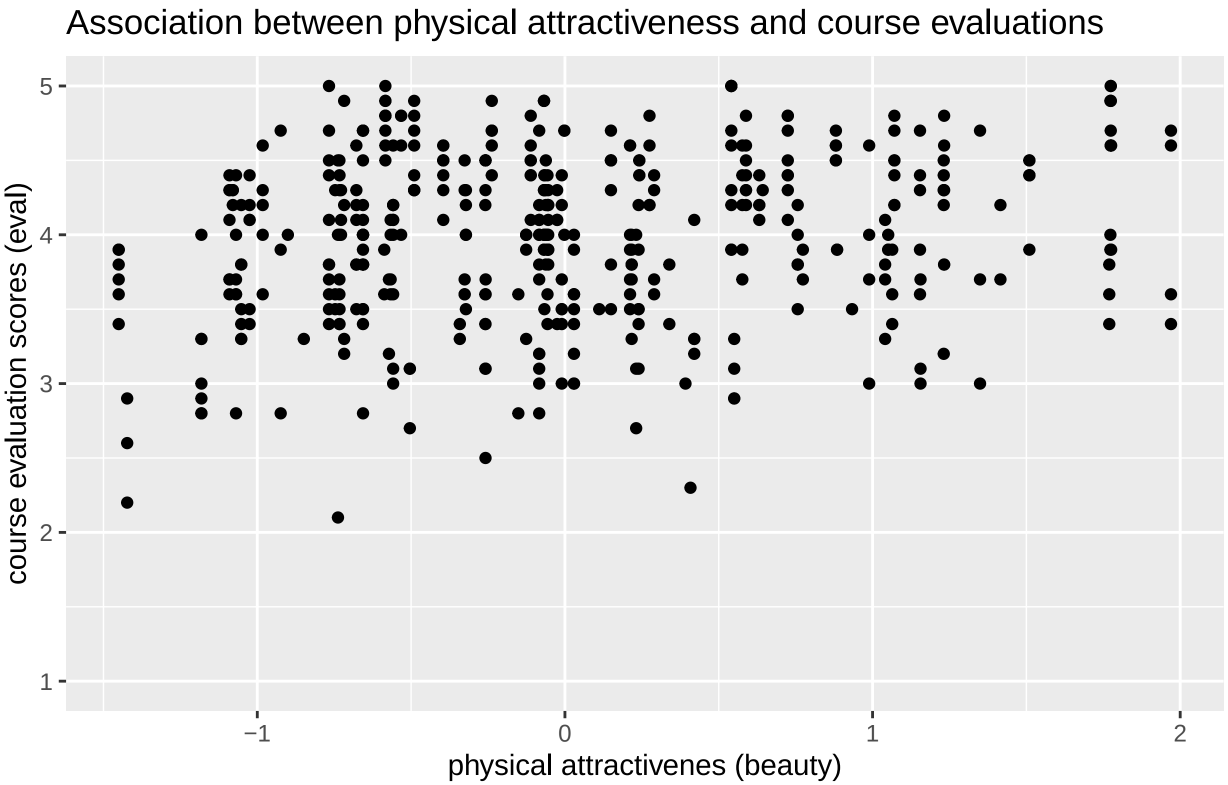
Part C

Distribution of course evaluation scores faceted by gender



Both of distributions of course evalution scores faceted by gender are left−skewed. The median course evaluation scores for male instructors (4.15) are higher than those for female instructors (3.9). However, the IQE value for males (0.8) is also higher compared to females (0.7), indicating greater distribution variability.

Part D



There is no significant linear correlation between a professor´s physical attractiveness and their course evaluations.

Problem 4: SAT scores for UT students

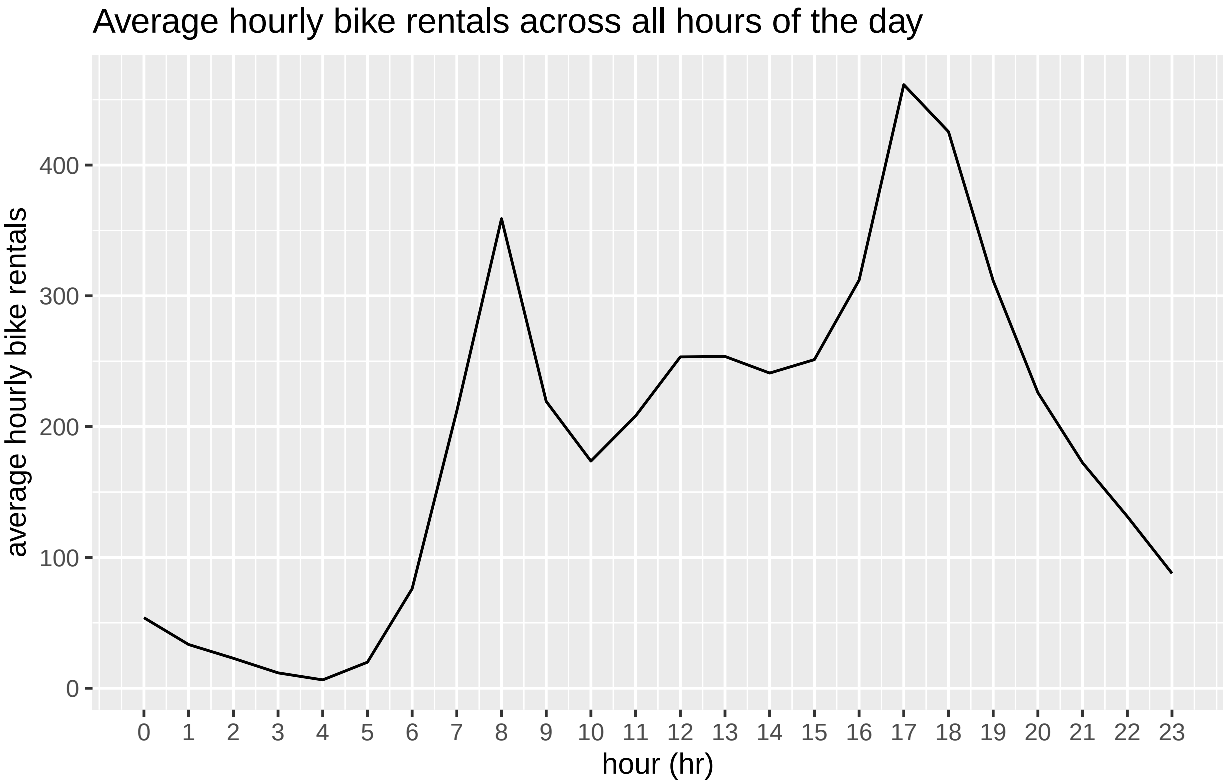
Part A

Table 10: The table shows the mean, standard deviation, interquartile rang, 5th percentile, 25th, percentile, median (50 percentile), 75th percentile, and 95th percentile of the SAT Verbal, SAT Quantitative, and GPA.

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



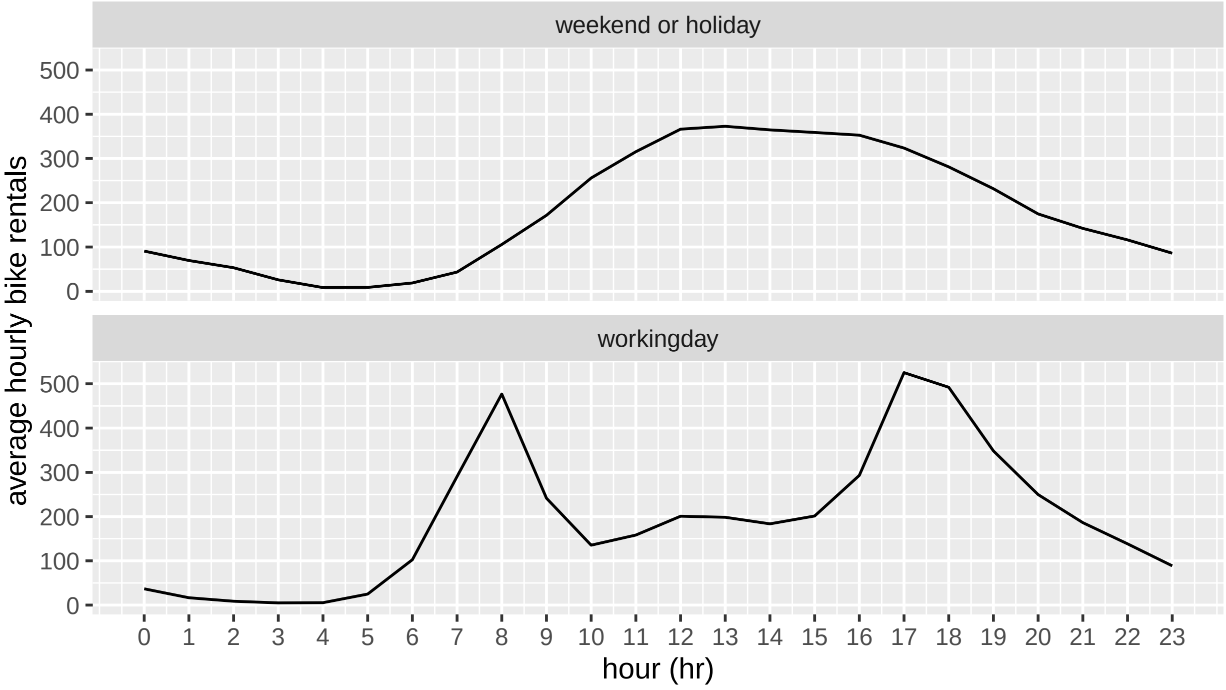
Average hourly bike rentals across all hours of the day

x−axis represent hour of the day
y−axis represent average bike rentals

Plot B

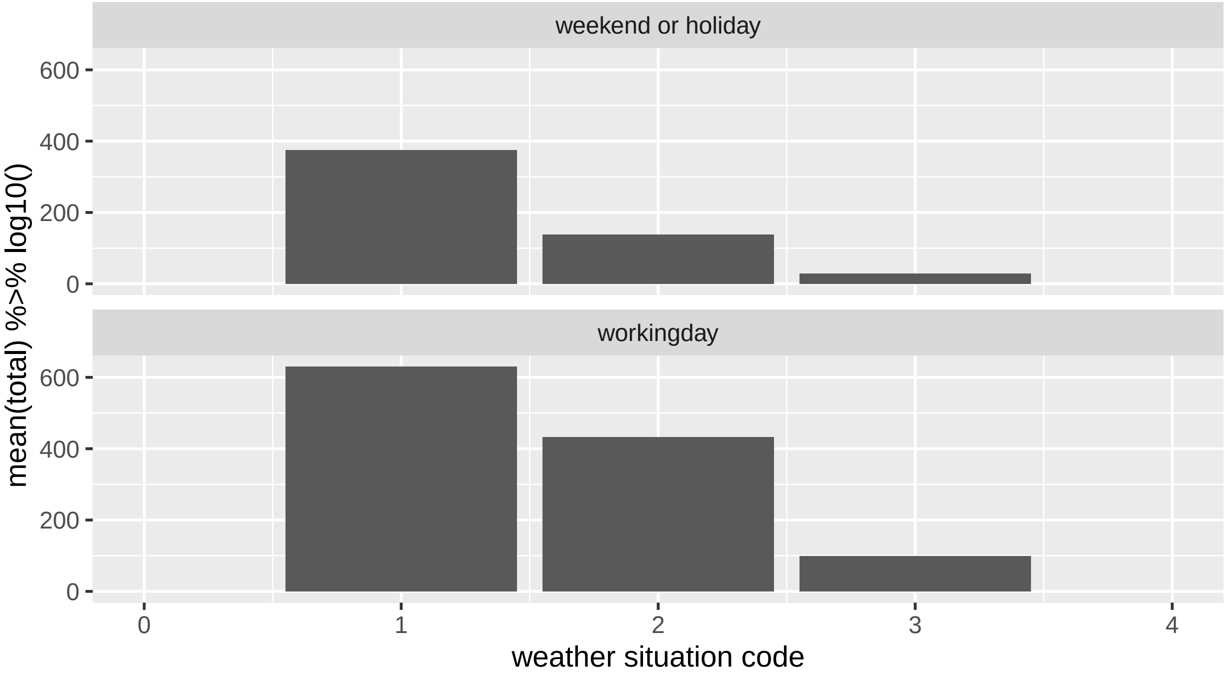
Average bike rentals by hour of the day.

Faceted according to whether it is a working day.



x−axis represent hour of the day
y−axis represent average bike rentals

Plot C



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
y−axis represent average ridership during the 9 AM hour after log10