homework02

zhewei xie

2024-07-23

Problem 1: Capital Metro UT Ridership

Question 1

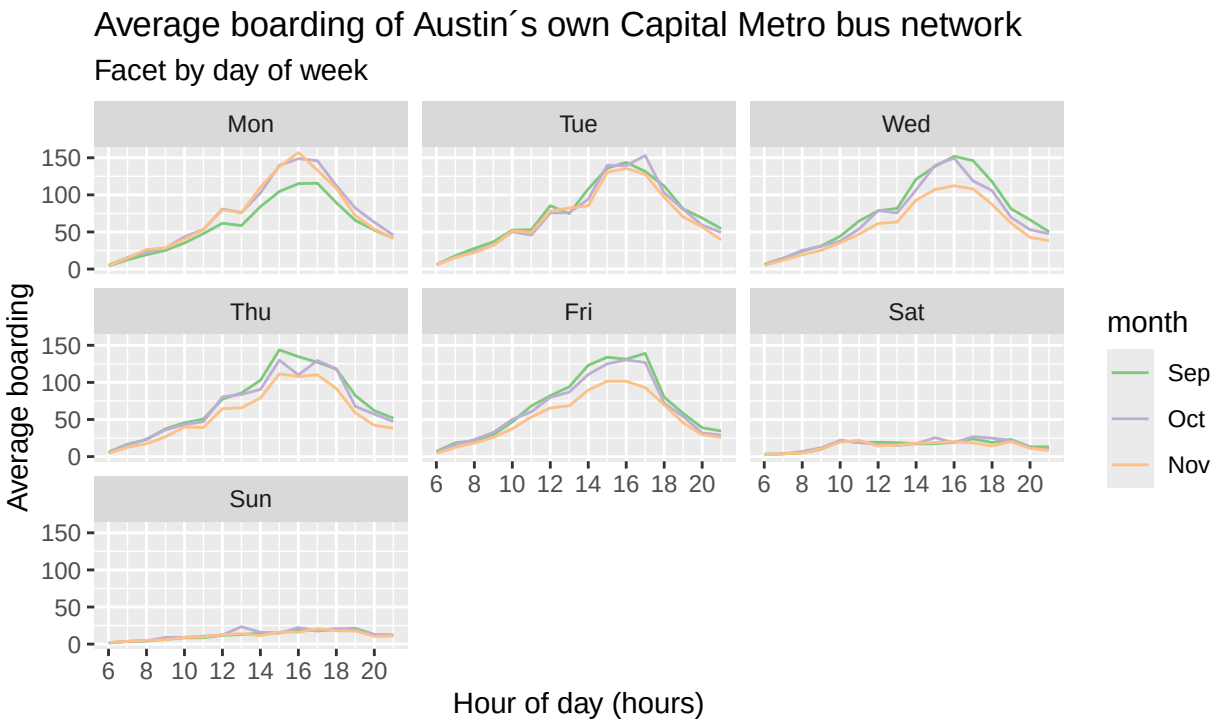

Figure 1. The average boarding of Austin´s own Capital Metro bus network, faceted by day of the week, is based on data from September to November 2018. The hour of peak boardings saw a notable decrease on weekends. Labor Day may have resulted in lower average boardings on Mondays in September. Meanwhile, Thanksgiving Day may have lowered the average boardings on Wednesdays, Thursdays, and Fridays in November.

Question 2

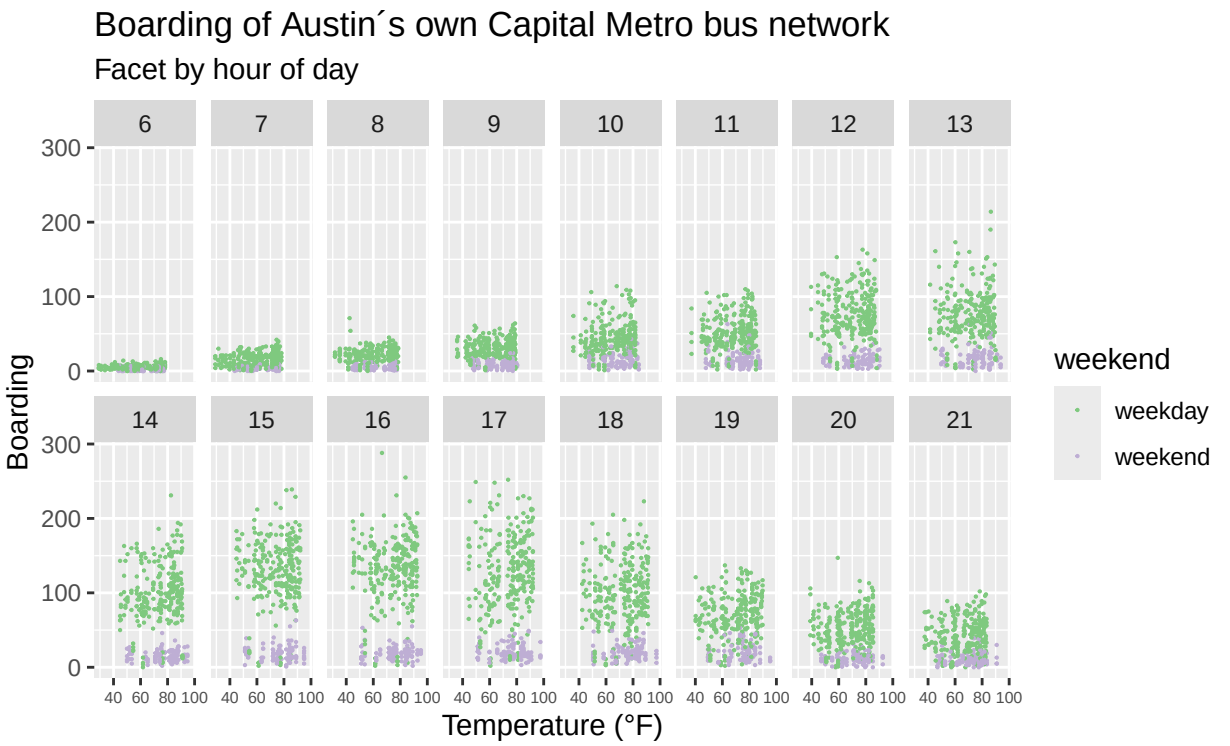

Figure 2. A scatter plot showing the average boardings of Austin´s Capital Metro bus network at different temperatures, grouped by weekend status, and faceted by the hour of the day. Temperature does not have a noticeable effect on the number of UT students riding the bus, considering the hour of the day and weekend status as constants.

Problem 2: Wrangling the Billboard Top 100

Part A

Table 1: The top 10 most popular songs since 1958, as measured by the total number of weeks that a song spent on the Billboard Top 100.

performer	song	count
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
The Weeknd	Blinding Lights	76
Jason Mraz	I'm Yours	76
LeAnn Rimes	How Do I Live	69
OneRepublic	Counting Stars	68
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
Adele	Rolling In The Deep	65
Jewel	Foolish Games/You Were Meant For Me	65
Carrie Underwood	Before He Cheats	64

Part B

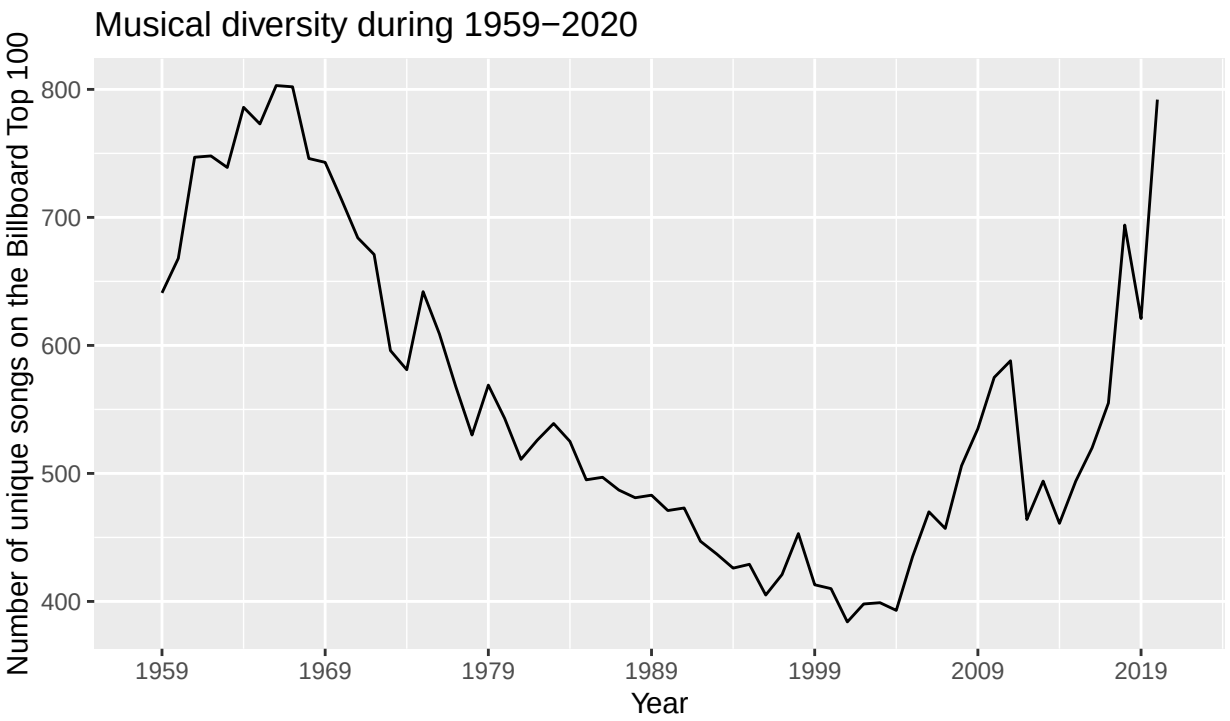

Figure 3. The trend of musical diversity from 1959 to 2020. The trend of musical diversity saw an increase from 1959 to 1967, peaking in 1966 with 803 unique songs and in 1967 with 802 unique songs. This was followed by a fluctuating decrease until 2004, reaching its lowest point at 384. It then increased until 2011, followed by a dramatic decrease the next year. The trend increased again from 2014, reaching its second peak in 2020 with 792 unique songs, which is still below the peak of the late 1960s.

Part C

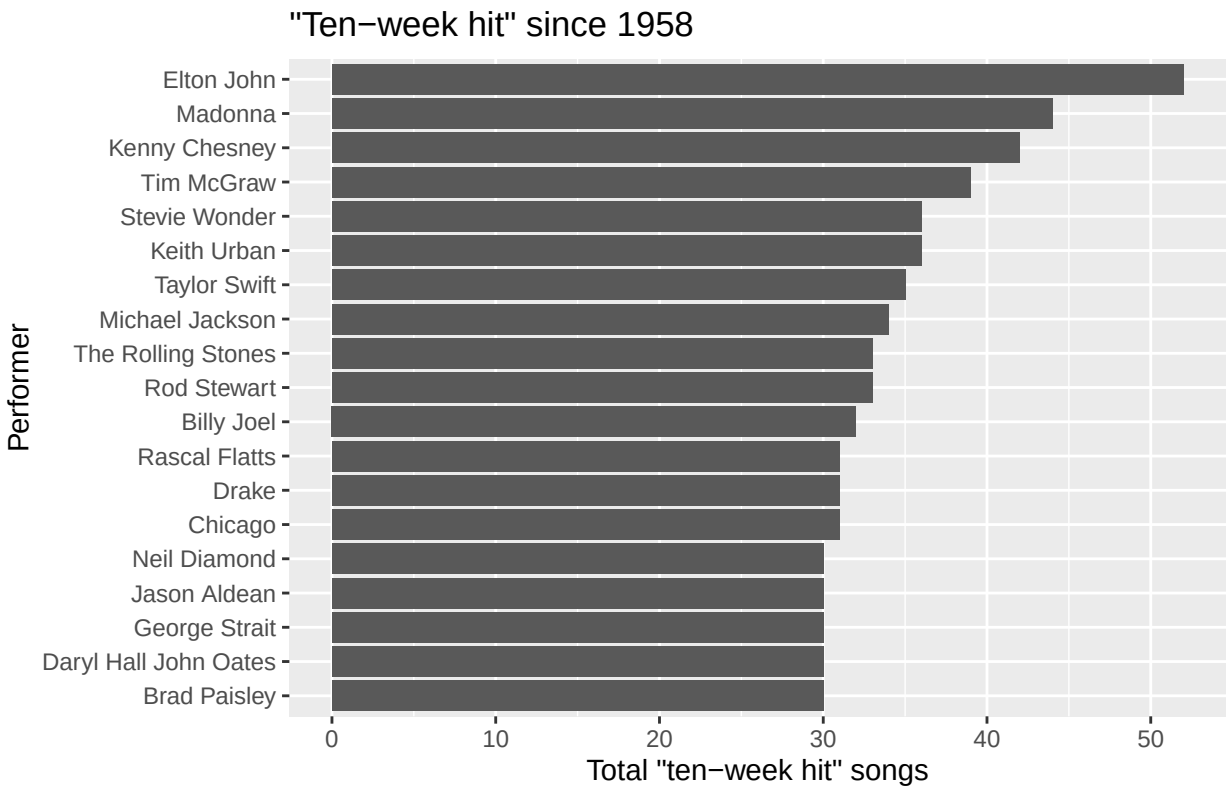

Figure 4. A barplot showing the performers who have had at least 30 songs that were "ten−week hits" since 1958 to 2021.

Problem 3: regression practice

Question A

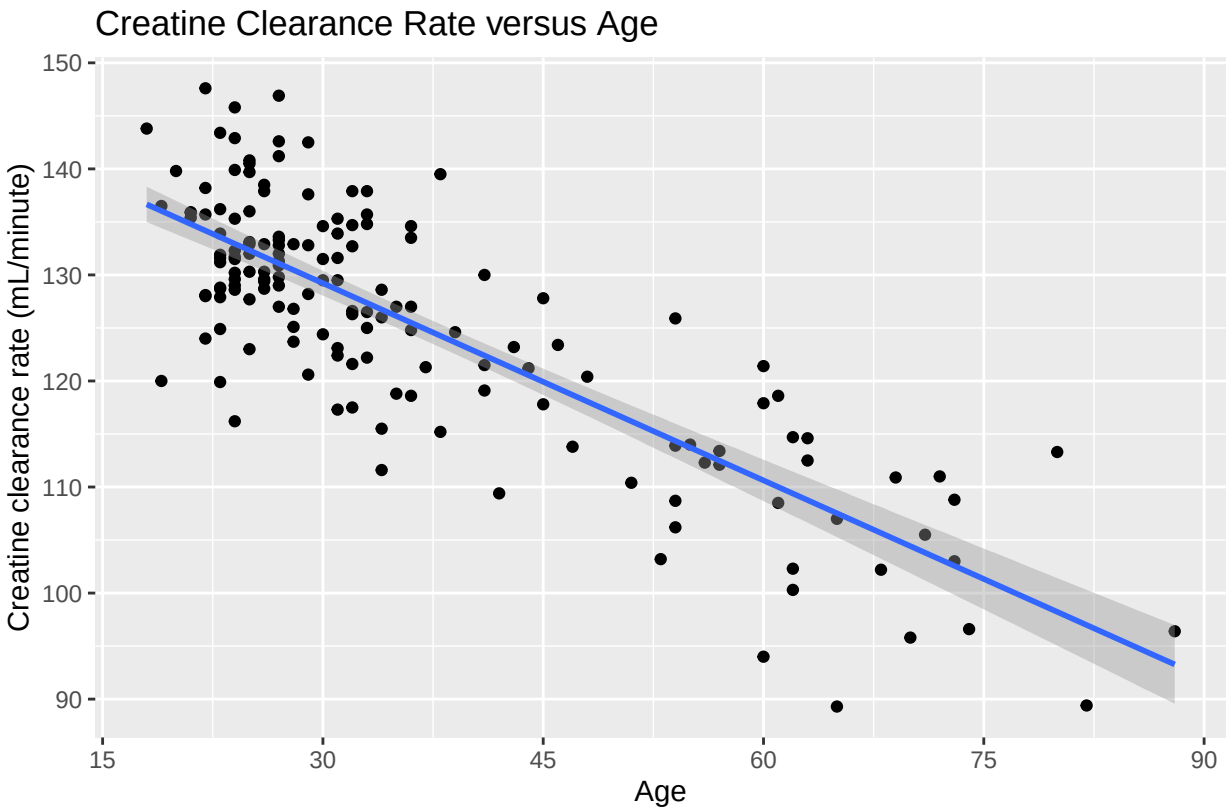

Figure 5. Creatine clearance rate versus age.

Table 2: The fitted linear model parameters of creatinine (based on data from individuals aged 18 to 88).

	value
intercept	147.81
slope	-0.62

Based on the fitted linear regression model parameter, it is clear that the linear regression is:

Clearance $= 147.81 - 0.62 \cdot$ Age.

Note: Based on data from individuals aged 18 to 88.

The linear regression gives the conditional expected value of creatinine clearance rate, given someone's age.

So let's plug in Age $= 55$ into the fitted equation:

$E(\text{Clearance}|Age = 55) = 147.81 - 0.62 \cdot 55 = 113.71$

This is the expected of creatinine clearance rate for a 55-year-old, with the value is 113.71 mL/minute.

Question B

Table 3: Predictions of creatinine clearance rate based on age (based on data from individuals aged 18 to 88).

age	predicted	age	predicted	age	predicted	age	predicted
18	136.66	36	125.50	54	114.34	72	103.19
19	136.04	37	124.88	55	113.72	73	102.57
20	135.42	38	124.26	56	113.10	74	101.95
21	134.80	39	123.64	57	112.48	75	101.33
22	134.18	40	123.02	58	111.86	76	100.71
23	133.56	41	122.40	59	111.24	77	100.09
24	132.94	42	121.78	60	110.62	78	99.47
25	132.32	43	121.16	61	110.00	79	98.85
26	131.70	44	120.54	62	109.38	80	98.23
27	131.08	45	119.92	63	108.76	81	97.61
28	130.46	46	119.30	64	108.14	82	96.99
29	129.84	47	118.68	65	107.52	83	96.37
30	129.22	48	118.06	66	106.91	84	95.75
31	128.60	49	117.44	67	106.29	85	95.13
32	127.98	50	116.82	68	105.67	86	94.51
33	127.36	51	116.20	69	105.05	87	93.89
34	126.74	52	115.58	70	104.43	88	93.27
35	126.12	53	114.96	71	103.81	89	92.65

Based on the linear regression, it says that a one-year change in age is associated with a 0.62 mL/minute change in creatinine clearance rate, on average.

Question C

Based on the fitted linear regression:

Clearance $= 147.81 - 0.62 \cdot$ Age.

It is clear that $E(\text{Clearance}|Age = 40) = 147.81 - 0.62 \cdot 40 \approx 123.0$

and $E(\text{Clearance}|Age = 60) = 147.81 - 0.62 \cdot 60 \approx 110.6$.

So for the 40-year-old, $\hat{\varepsilon} = y - \hat{y} = 135 - 123.0 \approx 12.0$,

while for the 60-year-old, $\hat{\varepsilon} = y - \hat{y} = 112 - 110.6 \approx 1.4$.

Therefore, based on the differences, the 40-year-old is healthier for the age.

Problem 4: probability practice

Part A

Event A is defined as 'you will get at least one lemon among the 3 cars you purchase.' It is easier to consider the converse: getting no lemons among the 3 cars you purchase. To calculate this, you can imagine choosing 3 cars from 20 normal cars, while the probability space consists of choosing 3 cars from all 30 cars.

$P(A) = 1 - P(\overline{A}) = 1 - \frac{\binom{20}{3}}{\binom{30}{3}} \approx 0.719$

Conclusion: The probability that you will get at least one lemon when you buy 3 cars is approximately 71.9%.

Part B

Question 1

It is clear that:

odd + even = odd
odd + odd = even
even + even = even

so the numbers of a 1-6 dice could be split to 2 sets.

$Set_{odd} = \{1, 3, 5\}$

$Set_{even} = \{2, 4, 6\}$

Event A is 'the sum of the two numbers is odd.' This event can be split into two steps. Step 1 is to choose one odd number from the set of odd numbers, and Step 2 is to choose one even number from the set of even numbers. Interestingly, we could choose from the odd set first or the even set first, and it does not affect the sum. Meanwhile, the probability space is clearly the random selection of 2 numbers from 1 to 6.

$P(A) = \frac{\binom{3}{1}\binom{3}{1}2!}{\binom{6}{1}\cdot\binom{6}{1}} = \frac{1}{2}$

Conclusion: The probability that the sum of the two numbers is odd is 50%.

Question 2

The event 'the sum of the two numbers is less than 7' can be divided into subevents, as follows:

$P(x + y < 7) = P(y < 6, x = 1) + P(y < 5, x = 2) + P(y < 4, x = 3) + P(y < 3, x = 4) + P(y < 2, x = 5) + P(y < 1, x = 6)$

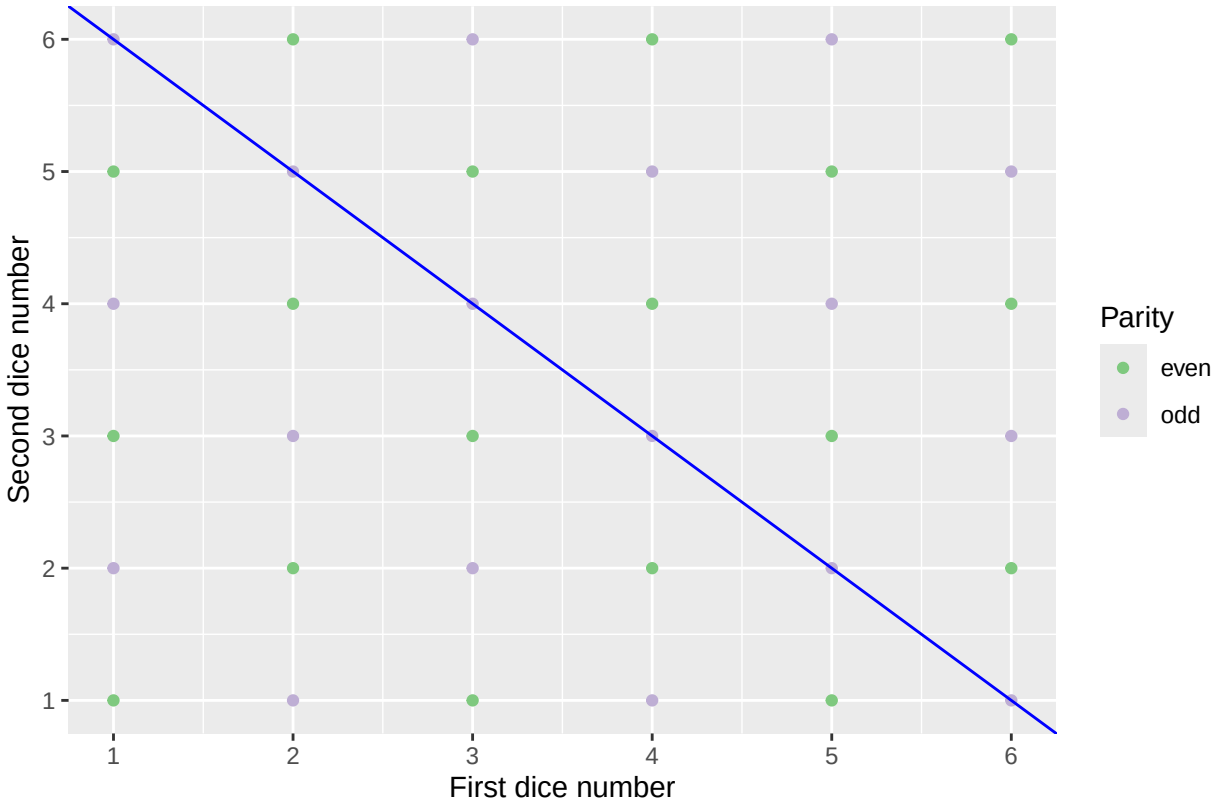

Figure 6. The event space of throwing two dice (each with the usual 6 sides, numbered 1–6).

Conclusion: Through counting the case, the probability that the sum of the two numbers is less than 7 is $\frac{15}{36} = \frac{5}{12}$.

Question 3

Event A is "the sum of the two numbers is less than 7".

Event B is "the sum of the two numbers is odd".

So $P(A|B) = \frac{P(AB)}{P(B)}$.

Given that there are 6 cases where the sum of the two numbers is less than 7 and odd, and there are 18 cases where the sum of the two numbers is odd, the probability that the sum of the two numbers is less than 7, given that it is odd, is $\frac{6}{18} = \frac{1}{3}$.

Because $P(A|\overline{B}) = \frac{P(A\overline{B})}{P(\overline{B})} = \frac{9}{18} = \frac{1}{2}$.
Clearly, $P(A|B) \neq P(A|\overline{B}) \neq P(A)$.

Conclusion: The events 'the sum of the two numbers is less than 7' and 'the sum of the two numbers is odd' are **not independent.**

Part C

Let's donate the event "Random clicker" as RC, the event "Truth clicker" as TC, the event "answer yes" as Y, the event "answer no" as N.

$P(Y) = P(Y|RC) \cdot P(RC) + P(Y|TC) \cdot P(TC)$

Given the expected fraction of random clickers is 0.3, it means that $P(RC) = 0.3$, so $P(TC) = 1 - P(\overline{TC}) = 1 - P(RC) = 1 - 0.3 = 0.7$.

Besides, because random clickers would click either one with equal probability, which means $P(Y|RC) = P(N|RC) = 0.5$ and the following survey results: 65% said Yes and 35% said No.

Therefore, $P(Y) = P(Y|RC) \cdot P(RC) + P(Y|TC) \cdot P(TC) = 0.5 \cdot 0.3 + P(Y|TC) \cdot 0.7 = 0.65$, it is clear that $P(Y|TC) = \frac{P(Y) - P(Y|RC) \cdot P(RC)}{P(TC)} = \frac{0.65 - 0.5 \cdot 0.3}{0.7} \approx 0.714$.

Conclusion: The fraction of people who are truthful clickers and answered "Yes" is approximately 71.4%.

Part D

Let's donate the event "someone has the disease" as D, the event "test positive" as TP.

Because someone has the disease, there is a probability of 0.993 that they will test positive, it is clear that $P(TP|D) = 0.993$.

Additionally, if someone does not have the disease, there is a 0.9999 probability that they will test negative, which means $P(\overline{TP}|\overline{D}) = 0.9999$ and $P(TP|\overline{D}) = 1 - P(\overline{TP}|\overline{D}) = 0.0001$.

In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it, which means $P(D) = 0.000025$ and $P(\overline{D}) = 1 - P(\overline{D}) = 0.999975$.

$P(TP) = P(TP|D) \cdot P(D) + P(TP|\overline{D}) \cdot P(\overline{D}) = 0.993 \cdot 0.000025 + 0.0001 \cdot 0.999975 = 0.0001248225$

According to Bayes' theorem, $P(D|TP) = \frac{P(TP|D) \cdot P(D)}{P(TP)} = \frac{0.993 \cdot 0.000025}{0.0001248225} \approx 0.1989$.

Conclusion: The probability that someone has the disease given that they tested positive is approximately 19.89%.

Part E

Let's donate the event "an aircraft is present in a certain area" as A, the event "a radar correctly registers its presence" as R.

$P(R|A) = 0.99$
$P(R|\overline{A}) = 0.10$
$P(A) = 0.05$ and $P(\overline{A}) = 1 - P(A) = 0.95$

Because $P(R) = P(R|A) \cdot P(A) + P(R|\overline{A}) \cdot P(\overline{A}) = 0.99 \cdot 0.05 + 0.10 \cdot 0.95 = 0.1445$.

According to Bayes' theorem, $P(A|R) = \frac{P(R|A) \cdot P(A)}{P(R)} = \frac{0.99 \cdot 0.05}{0.1445} \approx 0.3426$

Conclusion: The conditional probability that an aircraft is present given that the radar registers an aircraft presence is approximately 34.26%.

Problem 5: modeling soccer games with the Poisson distribution

Question

What are the estimated probabilities of win, lose, and draw results for a match between Liverpool (home) and Tottenham (away), and a match between Manchester City (home) and Arsenal (away)?

Approach

Step 1

The article "One match to go!", by Spiegelhalter and Ng provides an algorithm to evaluate the 'attack strength' and 'defence weakness' of a team in a season.

$$\text{Attack Strength} = \frac{\text{'goals for' of the team}}{\text{the average number of goals scored by a team}}$$

$$\text{Defence Weakness} = \frac{\text{'goals against' of the team}}{\text{the average number of goals scored by a team}}$$

To use this algorithm, let's define the 'average number of goals scored by a team' as μ_{goals}, the " 'goals against' of team i" as GF_i, and the " 'goals against' of team i" as GA_i. Besides, each goal scored as 'goals against' for a home team is the same as 'goals for' for the away team.

So, using the algorithm and data, we can calculate:

$$\mu_{\text{goals}} = \frac{\sum_{i=1}^{20}(\text{GF}_i + \text{GA}_i)}{20} = 53.6$$

Note: 20 is the number of the teams of the League.

By examining the data from 'epl_2018-19_away.csv' and 'epl_2018-19_home.csv,' it is clear that the goals are categorized into two classes: home team goals and away team goals. Therefore, the following can be established:

$$\text{GF}_i = \text{GF}_{ihome} + \text{GF}_{iaway}$$

$$\text{GA}_i = \text{GA}_{ihome} + \text{GA}_{iaway}$$

Note: We should merge the two tables carefully because the order of teams in the tables is different. We should merge them using the key 'Team.'

After merging the two files by the key 'Team' and summing the GF and GA from the merged file, it is easy to use R to calculate each team's 'attack strength' and 'defence weakness' with the appropriate function:

$$\text{AttackStrength}_i = \frac{\text{GF}_i}{\mu_{\text{goals}}}$$

$$\text{DefenceWeakness}_i = \frac{\text{GA}_i}{\mu_{\text{goals}}}$$

Step 2

According to the article, it is also important to calculate the average goals per game as a baseline, categorized by goals scored by home teams and by away teams.

Since the attribute named 'GP' stands for 'games played,' it is clear that each team played 19 games as the home team and 19 games as the away team. Therefore, the total number of games is $19 \cdot 20 = 380$. 380 games have been played with 1072 goals: 596 scored by home teams (average approximately 1.57 per game); 476 by away teams (average approximately 1.25 per game).

GF_{ihome} represent the 'goals for' by team i as the home team, while GF_{iaway} represents the 'goals for' team i as the away team:

$$\text{baseline}_{home} = \frac{\sum_{i=1}^{20} \text{GF}_{i_{home}}}{19 \cdot 20} \approx 1.568421$$

$$\text{baseline}_{away} = \frac{\sum_{i=1}^{20} \text{GF}_{i_{away}}}{19 \cdot 20} \approx 1.252632$$

Step 3

According to the method outlined in the article, the expected goals for a home team are:

$$\text{expected goals}_{home} = \text{baseline}_{home} \times \text{AttackStrength}_{home} \times \text{DefenceWeakness}_{away}$$

And the expected goals of an away team are:

$$\text{expected goals}_{away} = \text{baseline}_{away} \times \text{AttackStrength}_{away} \times \text{DefenceWeakness}_{home}$$

Step 4

The goals scored by teams are modeled using a Poisson distribution, based on the independence of each game, and 'attack strength' and 'defence weakness' of each team. And the scores of a team do not give us additional information about the performance of another team.

Firstly, the number of goals scored by a team in a game named as 'expected goals' is modeled as λ. Then, the probability $P(X = x)$ represents the likelihood of the target team scoring exactly x goals.

According to the Poisson distribution:

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Given that the expected goals scored by each team are independent, the joint probability for a match is:

$$P(X = x, Y = y) = \frac{\lambda_1^x}{x!} e^{-\lambda_1} \times \frac{\lambda_2^y}{y!} e^{-\lambda_2}$$

Result 1

Following the approach outlined above, we can calculate the result of each match. For example, for the match between Liverpool (home) and Tottenham (away), to determine the probability of a 2-1 result, we multiply 27% by 34% to get approximately 9%, as demonstrated by the following calculation steps:

$\because$

$\lambda_{\text{Liverpool}_{win}} = \text{baseline}_{\text{home}} \times \text{AttackStrength}_{\text{Liverpool}} \times \text{DefenceWeakness}_{\text{Tottenham}}$

$\lambda_{\text{Liverpool}_{lose}} = \text{baseline}_{\text{away}} \times \text{AttackStrength}_{\text{Tottenham}} \times \text{DefenceWeakness}_{\text{Liverpool}}$

$\therefore$

$\lambda_{\text{Liverpool}_{win}} = 1.568421 \times 1.660448 \times 0.7276119 \approx 1.89$

$\lambda_{\text{Liverpool}_{lose}} = 1.252632 \times 1.25 \times 0.4104478 \approx 0.64$

$P(X = 2, Y = 1) = \frac{1.89^2}{2!}e^{-1.89} \times \frac{0.64^1}{1!}e^{-0.64} \approx 0.2698 \times 0.3375 \approx 0.091$

Table 4: Expected number of goals, and percentage chance of getting a particular score for the two teams (Liverpool (home) versus Tottenham (away)), assuming a Poisson distribution

Team	Expected goals	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)
liverpool	1.89	15	28	27	17	8	3	1
tottenham	0.64	53	34	11	2	0	0	0

To calculate probabilities using the Poisson distribution, we can use the dpois function in R. To estimate the probabilities of each possible match result, we multiply the Poisson probabilities of the two competing teams together, assuming independence. Creating a heat map to visualize these probabilities is also a useful approach.

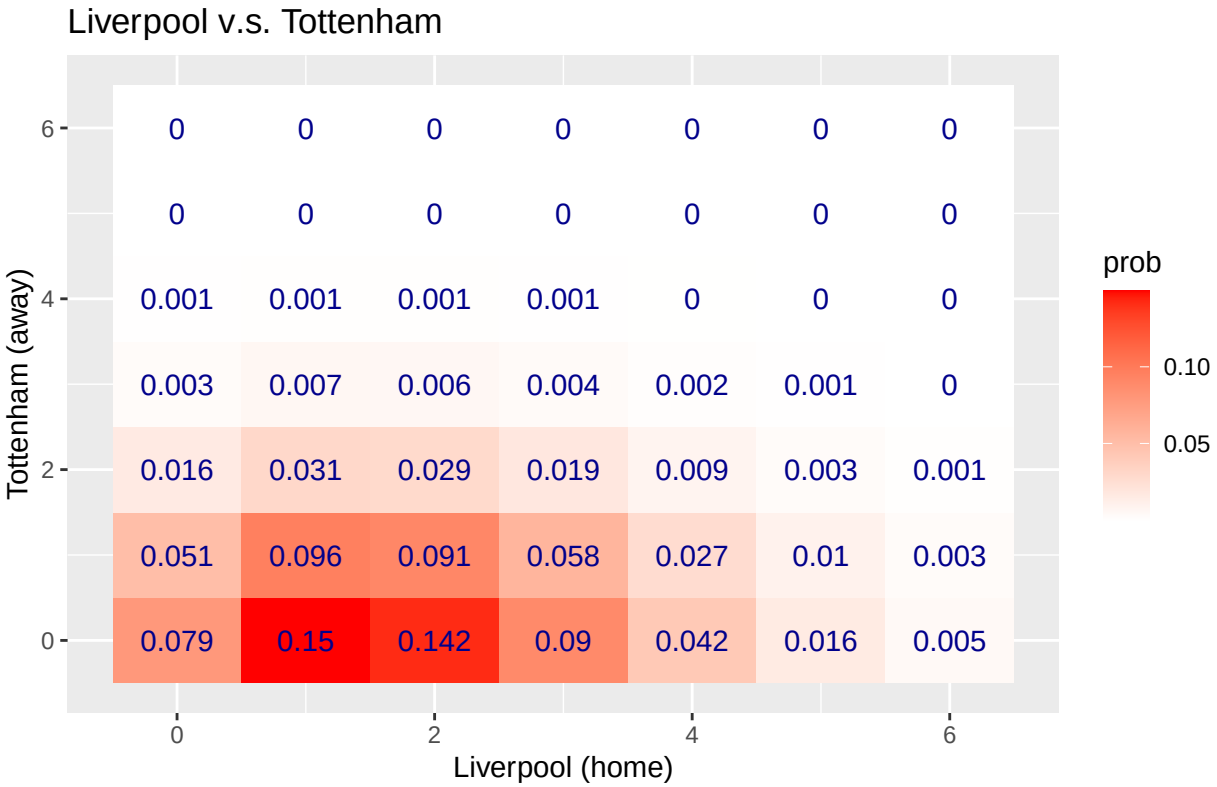

Figure 7. Probabilities of the expected goals in a match between Liverpool (home) and Tottenham (away).

Based on the probabilities, we can easily sum the probabilities of all outcomes that meet our condition to determine the win/lose/draw result for the match between Liverpool (home) and Tottenham (away).

	Proabilities
Liverpool win	0.6694094
Liverpool lose	0.1182597
Draw	0.2089292

As an alternative approach, running a Monte Carlo simulation can also provide similar probabilities for the match between Liverpool (home) and Tottenham (away).

	Proabilities
Liverpool win	0.67136
Liverpool lose	0.12108
Draw	0.20756

Result 2

Following the approach outlined above, we can calculate the result of each match. For example, for the match between Manchester City (home) and Arsenal (away), to determine the probability of a 2-1 result, we multiply 27% by 34% to get approximately 9%, as demonstrated by the following calculation steps:

$\because$

$$\lambda_{\text{Manchester City}_{win}} = \text{baseline}_{\text{home}} \times \text{AttackStrength}_{\text{Manchester City}} \times \text{DefenceWeakness}_{\text{Arsenal}}$$

$$\lambda_{\text{Manchester City}_{lose}} = \text{baseline}_{\text{away}} \times \text{AttackStrength}_{\text{Arsenal}} \times \text{DefenceWeakness}_{\text{Manchester City}}$$

$\therefore$

$$\lambda_{\text{Manchester City}_{win}} = 1.568421 \times 1.772388 \times 0.9514925 \approx 2.65$$

$$\lambda_{\text{Manchester City}_{lose}} = 1.252632 \times 1.36194 \times 0.4291045 \approx 0.73$$

$$P(X = 2, Y = 1) = \frac{2.65^2}{2!}e^{-2.65} \times \frac{0.73^1}{1!}e^{-0.73} \approx 0.2481 \times 0.3518 \approx 0.087$$

Table 7: Expected number of goals, and percentage chance of getting a particular score for the two teams (Manchester City (home) versus Arsenal (away)), assuming a Poisson distribution

Team	Expected goals	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)
man_city	2.65	7	19	25	22	14	8	3
arsenal	0.73	48	35	13	3	1	0	0

Using the same approach as in the previous question, we calculate the joint probabilities for the match between Manchester City (home) and Arsenal (away) and represent the results as a heat map.

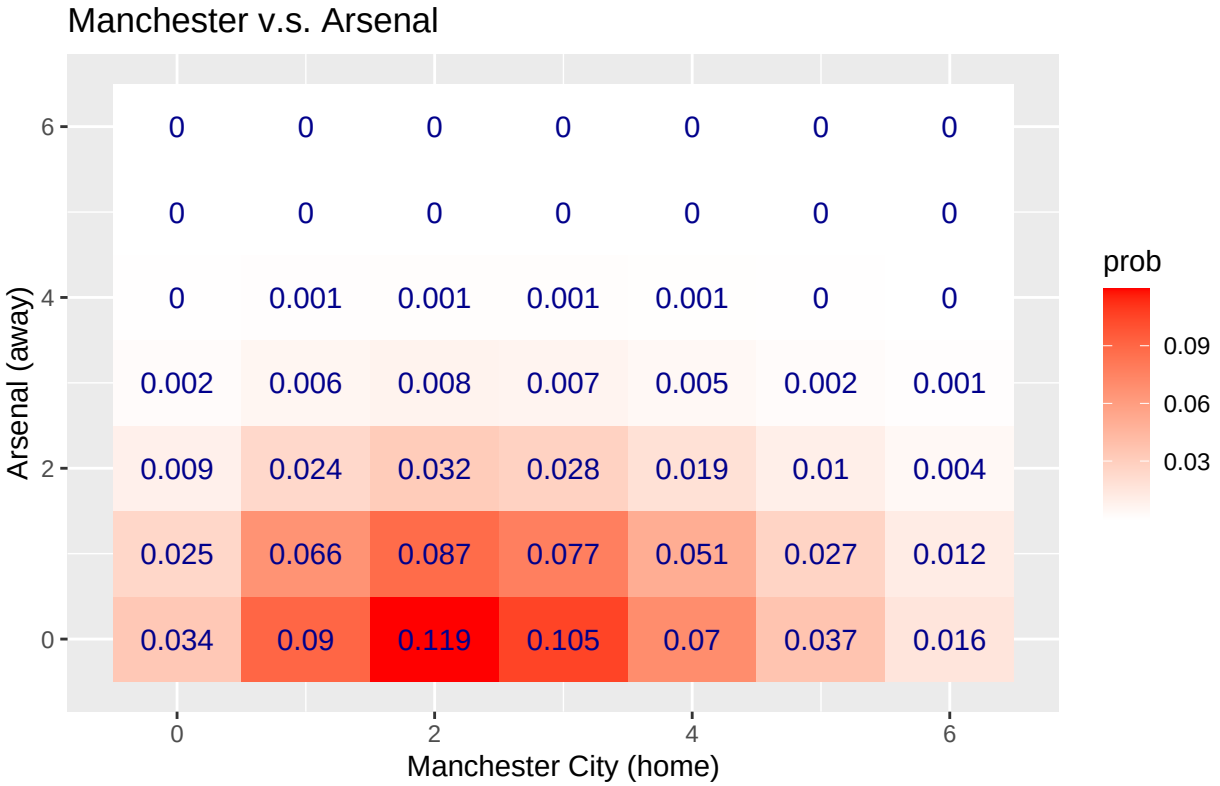

Figure 8. Probabilities of the expected goals in a match between Manchester City (home) and Arsenal (away).

Based on the probabilities, we can easily sum the probabilities of all outcomes that meet our condition to determine the win/lose/draw result for the match between Manchester City (home) and Arsenal (away).

	Proabilities
Manchester City win	0.7619712
Manchester City lose	0.0793087
Draw	0.1400616

As an alternative approach, running a Monte Carlo simulation can also provide similar probabilities for the match between Manchester City (home) and Arsenal (away).

	Proabilities
Manchester City win	0.78106
Manchester City lose	0.07885
Draw	0.14009

Conclusion

In conclusion, based on the statistical model, we can expect the following probabilities for the matches:

- The probability of Liverpool winning as the home team against Tottenham as the away team is approximately 67.3%,
- The probability of Liverpool losing the game is approximately 11.7%,
- The probability of a draw is approximately 21.0%.

For the match between Manchester City (home) and Arsenal (away):

- The probability of Manchester City winning is approximately 77.9%,
- The probability of losing the game is approximately 8.0%,
- The probability of a draw is 14.1%."

The estimations above are obtained by running Monte Carlo simulations to provide the probabilities, ensuring that the sum of the three possible game results equals 1. All of these estimates are based on data from the 2018-19 English Premier League soccer season. This model is very simple and does not account for recent factors. Additionally, the model naively assumes that the performance of each team is independent, which may not be realistic. The extent to which the data from previous seasons can accurately predict outcomes is likely overly idealistic.

Note: This model is used only as a simple estimation of match results for the English Premier League soccer season and should not be used to make betting decisions. Any decisions made based on this model are at the user's own risk.