S

93201Q

932012

NZQA
NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA

QUALIFY FOR THE FUTURE WORLD
KIA NOHO TAKATŪ KI TŌ ĀMUA AO!

# Scholarship 2020
# Statistics

2.00 p.m. Friday 20 November 2020
Time allowed: Three hours
Total score: 40

## QUESTION BOOKLET

There are FIVE questions in this booklet. Answer ALL questions.

Pull out Formulae and Tables Booklet S–STATF from the centre of this booklet.

Write your answers in Answer Booklet 93201A.

Show ALL working. Start your answer to each question on a new page. Carefully number each question.

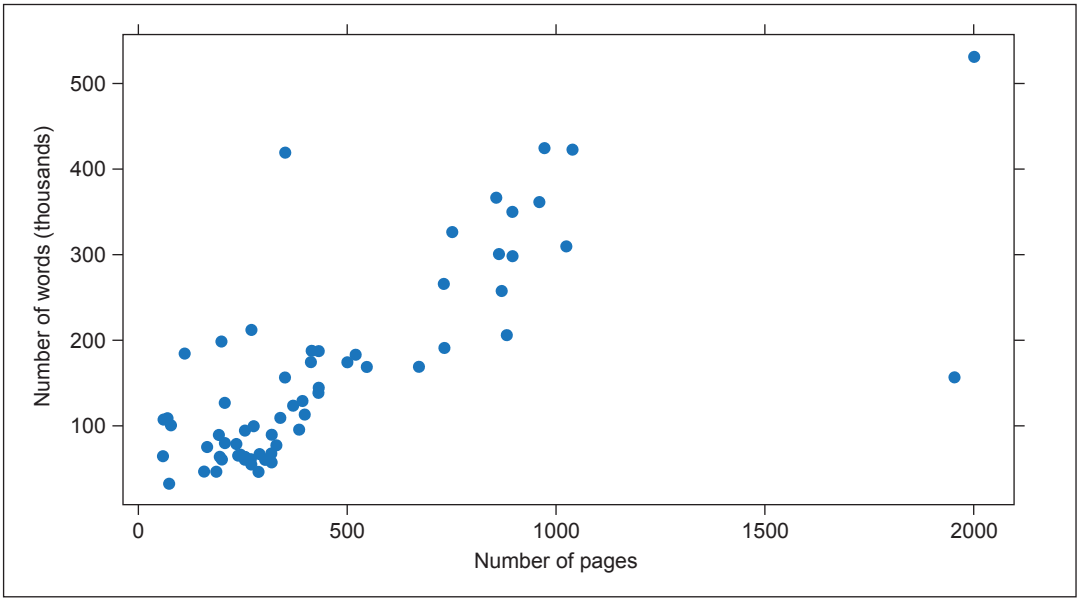
Check that this booklet has pages 2–11 in the correct order and that none of these pages is blank.

**YOU MAY KEEP THIS BOOKLET AT THE END OF THE EXAMINATION.**

## QUESTION ONE

(a) A website for authors states that for fiction books there are, on average, 250 words per page. Data was obtained on the number of words (in thousands) and the number of pages for 65 of the most popular fiction books written in English. Figure 1 shows the scatterplot produced from these books and variables.

**Figure 1: Scatterplot of popular fiction books written in English**




(i) Describe the relationship between the number of words and the number of pages for these books, identifying any notable features of the data.

(ii) Give TWO potential reasons why the number of pages of a book might not precisely predict the number of words in the book.

(iii) A linear model was fitted to the data shown in Figure 1.

The equation of this model is given below:

$$\text{Number of words (thousands)} = 57.84 + 0.2202 \times \text{Number of pages}$$

With reference to the data, the features of the scatterplot, and this linear model, discuss the suitability of the statement "for fiction books there are, on average, 250 words per page".

(b) Read the following report.

> This report summarises the results of the second survey of book reading in New Zealand.
>
> Between 2 and 25 May 2018, 2261 adult New Zealanders responded to the online survey conducted by Horizon Research Limited for the New Zealand Book Council. Participants were recruited to represent the New Zealand population. The sample was weighted to match national demographics for age, gender, personal income, education level, employment status, and ethnicity.
>
> This research into the reading habits of New Zealanders confirms that we are a nation that loves to read. 86% of New Zealand adults had read or started to read at least one book in the past year, with on average 35 books per reader. While this is a lower percentage than in the March 2017 survey (88%, $n = 2082$), the difference is not statistically significant.
>
> It is wonderful that New Zealanders love to read, and to see that books remain an important touchstone in our society. But it's worrying to see how many of us didn't pick up a book in the past year. 14% of Kiwis didn't read a book in the past year. Males made up most (69%) of those adults who did not read a book in the past year.
>
> Adapted from: Book Reading in New Zealand, August 2018, New Zealand Book Council, https://www.read-nz.org/Downloads/Assets/Download/56854/1/2018%20Book%20Reading%20in%20NZ%20August%2027%20high-res%20final.pdf

(i) Identify ONE claim made in the report that is based on at least one survey percentage.

Evaluate the claim using a point estimate and a "rule of thumb"-based margin of error.

(ii) The report states that "86% of New Zealand adults had read or started to read at least one book in the past year, with on average 35 books per reader".

Discuss TWO potential non-sampling errors associated with people responding to being asked how many books they read in the past year.

## QUESTION TWO

(a) Project Gutenberg is an online library with over 60 000 free eBooks. The eBooks have been digitised by volunteers, with a focus on older works that are in the public domain.

A random sample of 24 books was taken from all the eBooks available from Project Gutenberg. For each book in the sample, the language it was written in and year that it was first published was determined by examining the digitised book.

(i) The age of each book was calculated using the difference, in years, between 2020 and the year the book was first published. The sample was then used to construct a bootstrap confidence interval for the median age, and this interval had limits (108, 143).

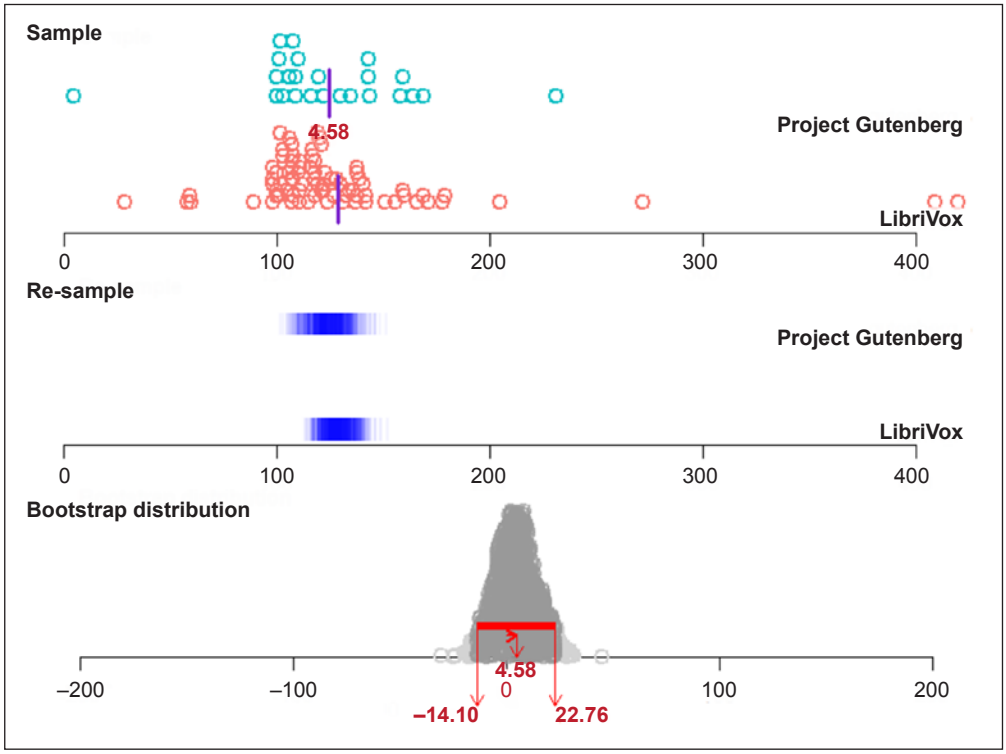
Project Gutenberg states on their website that most of their eBooks were published before 1924.

Discuss whether this claim can be supported by interpreting the confidence interval given above.

(ii) LibriVox is a website that provides over 15 000 free public domain audiobooks read by volunteers from around the world. A random sample of 80 books was taken from all the audiobooks available from the website, and for each book in the sample, the year that it was first published as a printed book was recorded.

The random sample of books from LibriVox was compared to the random sample of 24 books from Project Gutenberg. The samples were used to construct a bootstrap confidence interval for the difference between the mean age of books from LibriVox and the mean age of books from Project Gutenberg. The output from this analysis is shown in Figure 2.

**Figure 2: Bootstrap confidence interval output**




Discuss what can be concluded from both the features of the sample data distributions and the confidence interval constructed using the sample data.

(iii) Five of the 24 books in the sample from Project Gutenberg were not written in English.

Use a probability distribution model to evaluate whether there is sufficient evidence to conclude that more than half of the books available from Project Gutenberg are written in English.

In your answer justify the selection of the probability distribution model that you used.

(b) A local library wants to find out how long people spend at the library when they visit in person. They suspect that people who arrive in the morning stay for longer than people who arrive in the afternoon, and want to know how much longer, on average, they stay.

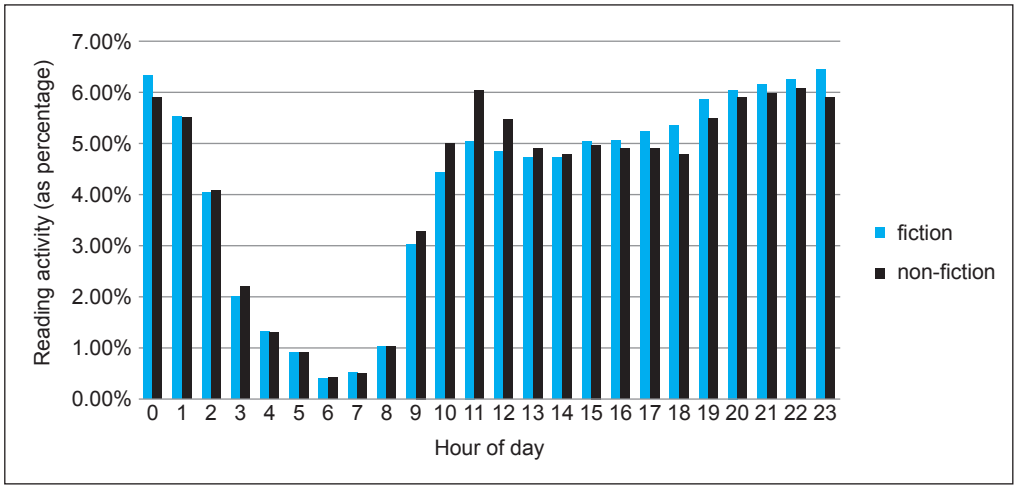
Apply the steps of the statistical enquiry cycle to this situation, giving a short description of what each step would involve.

## QUESTION THREE

A study was carried out using 10 months of reading data from an eBook subscription company. 8000 people were studied over three million reading sessions.

(a) Figure 3 shows the distribution of reading activity for fiction and non-fiction genres throughout a day.

**Figure 3: Distribution of reading activity throughout a day**




Write TWO comments comparing the similarities and differences for the reading activity of fiction and non-fiction books throughout a day. Include at least one numeric comparison of likelihood.

(b) The distribution of reading speeds (number of words read per minute) for people in the study was described in a report as "bell-shaped, with a mean around 150 words per minute".

3310 of the people in the study had a reading speed between 120 and 150 words per minute.

Using an appropriate probability distribution model, estimate the lower and upper limits for the middle 95% of reading speeds for people in the study.

(c) The study explored what percentage of a book, on average, people read before they stopped reading the book. Several thousands of books that had each been read by at least 40 people during the study were used to calculate a mean percentage completion value for each book. The mean percentage completion values for these books ranged from 0% to 100%, with 68% being the most likely value.

About half of the books had a mean completion value of 64% or less. Only about 5% of the books had a mean completion value higher than 90%.

(i) Investigate whether a triangular distribution would be a good model for the mean percentage completion values for books available from the eBook subscription company.

Support your answer with statistical reasoning and calculations.

(ii) Discuss TWO factors that might explain the variation in the mean percentage completion values for books available from the eBook subscription company.

## QUESTION FOUR

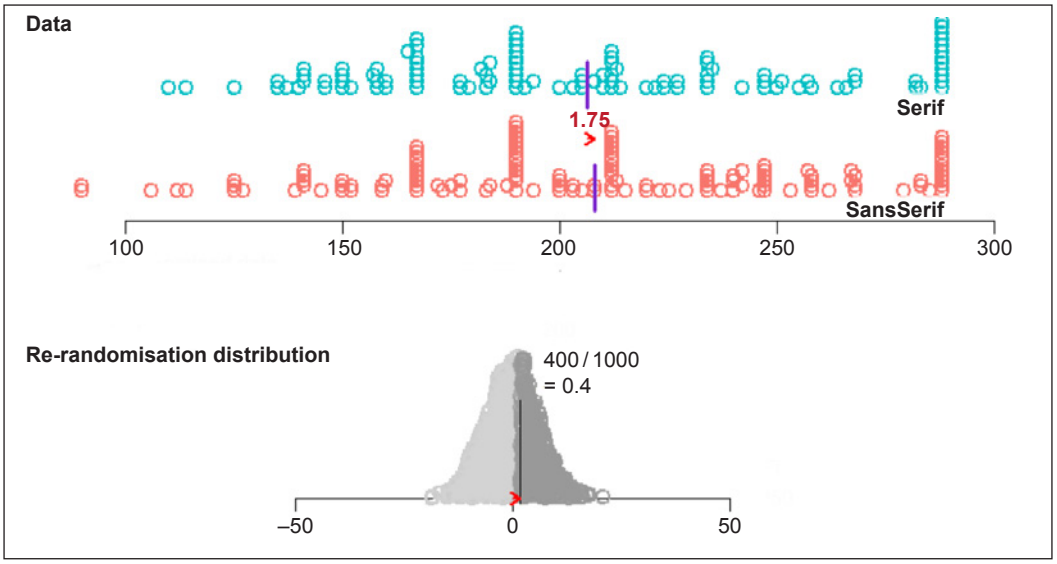
This sentence is written in a serif font. This sentence is written in a sans-serif font. It has been suggested that sans-serif fonts are easier to read than serif fonts.

An experiment was carried out to investigate whether the type of font used affected the difficulty of reading the text. 238 university medical students volunteered to participate in the study. The students were randomly allocated into two groups and given the same text to read, which contained 288 words. One group was given the text in a serif font, and the other group was given the text in a sans-serif font. The number of words read from the text in one minute was recorded for each student.

(a) Write a short paragraph that identifies the key elements of the experimental design using **appropriate statistical terminology**.

(b) A randomisation test was carried out using the difference between the mean reading speed for the two groups. Figure 4 gives some output from this test.

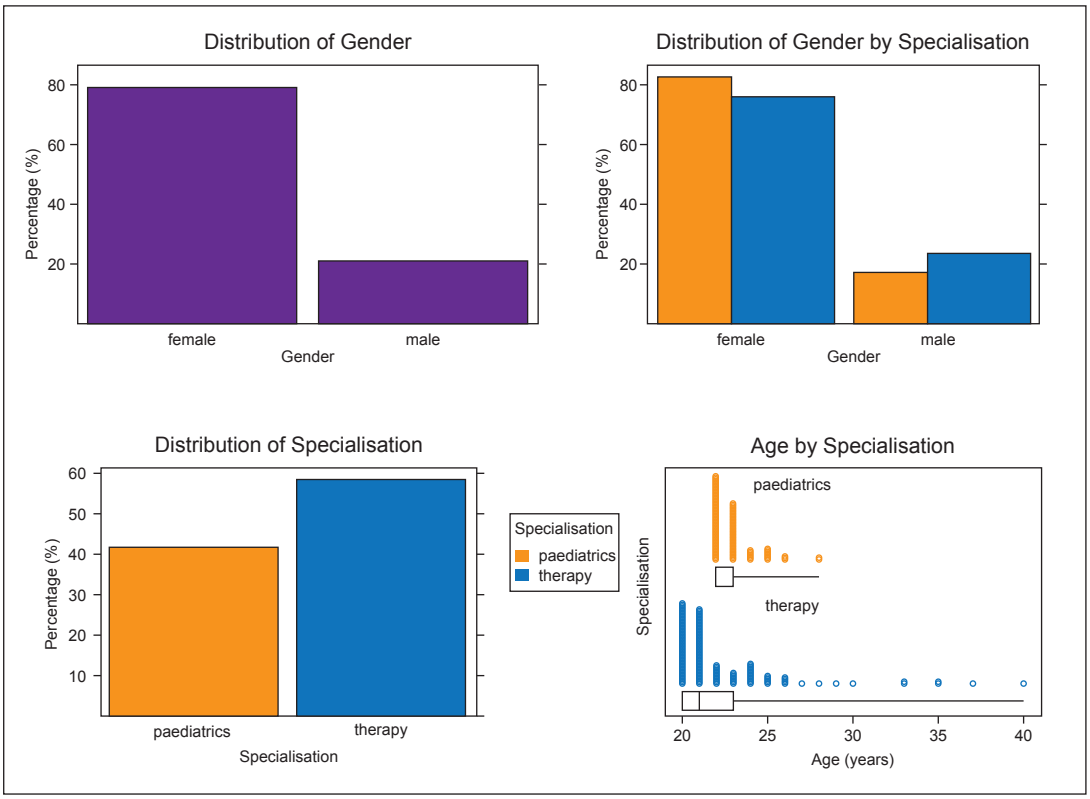
**Figure 4: Randomisation test output**




Interpret the randomisation test output, and explain why the result could have been expected in this context.

(c) The researchers also collected data from each student about their gender, age, and degree specialisation. Figure 5 shows four plots produced from this data.

**Figure 5: Plots produced using student data**




(i) Write a short paragraph summarising the information collected about the students in the study.

(ii) With respect to this information about the students in the study, discuss how the design of the experiment could be modified.

(d) This experiment was conducted with Russian medical students using text written in the Russian (Cyrillic) alphabet. The text used for the study was about the history of medicine in Russia. All participants were fluent speakers of the language used for the text and had normal or corrected vision (e.g. wore glasses).

Discuss how this new information about the study affects the generalisability of the results from the experiment.

## QUESTION FIVE

(a) Wikipedia is a free online encyclopaedia that publishes articles created and maintained by volunteers. From the start of 2001, when Wikipedia was started, until the end of 2018, Wikipedia provided data on the total number of new articles published in the English language per month.

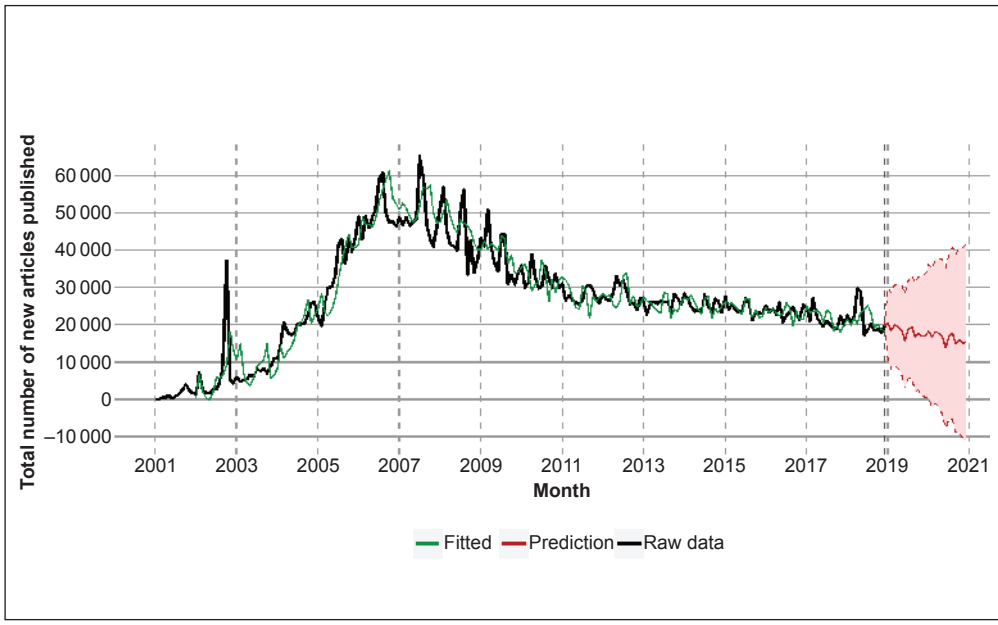
(i) Figure 6 displays the raw data for the total number of new articles published in the English language per month, with a smoothed trend curve shown in blue. Figure 6 also displays the seasonal differences, with their average (mean) shown in red.

**Figure 6: New articles published in the English language on Wikipedia, 2001–2018**


Write a short paragraph identifying key features of the data for the total number of new articles published in the English language per month, over the period 2001 to 2018.

(ii) A student used an additive Holt-Winters model to obtain a forecast for the total number of new articles published in the English language during November 2020. Figure 7 shows the raw and fitted data for the months from 2001 to 2018, and forecasts produced from the model.

**Figure 7: Holt-Winters model fitted to Wikipedia data**



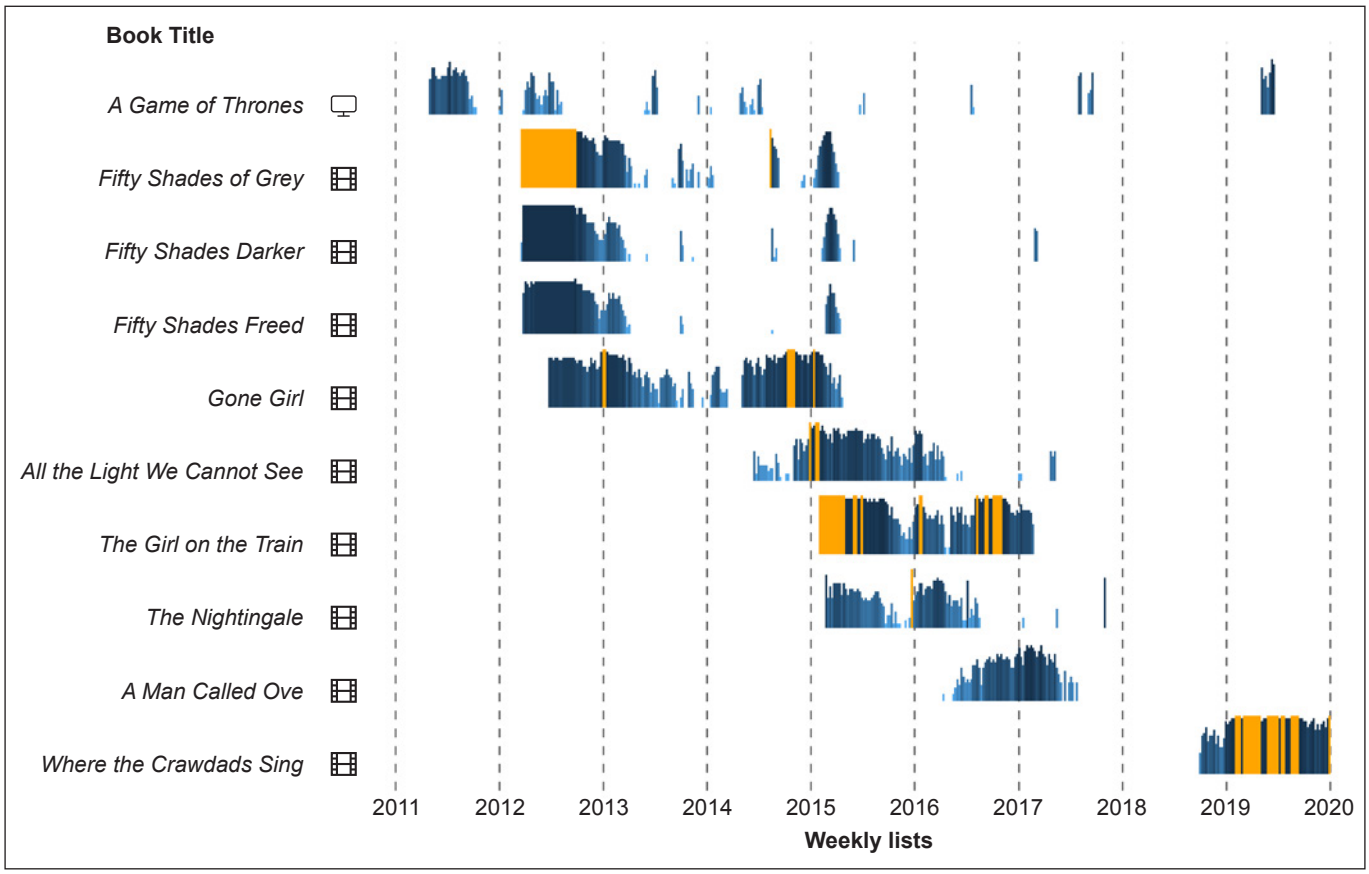

| | fit | upr | lwr |
|---|---|---|---|
| Jan 2019 | 20349.78 | 29928.19 | 10771.3609 |
| Feb 2019 | 18376.41 | 28517.80 | 8235.0080 |
| Mar 2019 | 20027.18 | 30749.31 | 9305.0603 |
| Apr 2019 | 19522.86 | 30842.68 | 8203.0446 |
| May 2019 | 18700.84 | 30634.63 | 6767.0581 |
| Jun 2019 | 15631.41 | 28194.81 | 3068.0060 |
| Jul 2019 | 18902.25 | 32110.36 | 5694.1327 |
| Aug 2019 | 19487.53 | 33354.94 | 5620.1171 |
| Sep 2019 | 16716.75 | 31257.58 | 2175.9251 |
| Oct 2019 | 17776.24 | 33004.19 | 2548.2887 |
| Nov 2019 | 17010.32 | 32938.72 | 1081.9216 |
| Dec 2019 | 17150.50 | 33792.31 | 508.6811 |
| Jan 2020 | 18535.44 | 36326.74 | 744.1341 |
| Feb 2020 | 16562.07 | 35074.91 | -1950.7692 |
| Mar 2020 | 18212.85 | 37460.33 | -1034.6365 |
| Apr 2020 | 17708.52 | 37703.42 | -2286.3772 |
| May 2020 | 16886.50 | 37641.28 | -3868.2748 |
| Jun 2020 | 13817.07 | 35343.90 | -7709.7560 |
| Jul 2020 | 17087.91 | 39398.69 | -5222.8672 |
| Aug 2020 | 17673.19 | 40779.56 | -5433.1841 |
| Sep 2020 | 14902.42 | 38815.80 | -9010.9651 |
| Oct 2020 | 15961.90 | 40693.48 | -8769.6756 |
| Nov 2020 | 15195.98 | 40756.74 | -10364.7770 |
| Dec 2020 | 15336.16 | 41736.88 | -11064.5672 |

Evaluate the use of this model to make a forecast for the total number of new articles published in November 2020.

(b) The *New York Times* publishes a list of the top 20 best-selling fiction books each week. Data was obtained on all books that appeared on this list during 2011 to 2019.

Figure 8 shows a visualisation that was created using data about the 10 fiction books that were on the best-sellers list for the highest total number of weeks from 2011 to 2019. Figure 8 is deliberately missing a legend that would help readers understand the visualisation.

**Figure 8: Visualisation of *New York Times* best-selling fiction books data**




Data about the book *The Nightingale* during 2015 is given in the table below.

| Date of weekly list | Ranking in list |
|---|---|
| 22/02/2015 | 3 |
| 1/03/2015 | 11 |
| 8/03/2015 | 7 |
| 15/03/2015 | 11 |
| 22/03/2015 | 6 |
| 29/03/2015 | 6 |
| 5/04/2015 | 6 |
| 12/04/2015 | 6 |
| 19/04/2015 | 10 |
| 26/04/2015 | 6 |
| 3/05/2015 | 8 |
| 10/05/2015 | 10 |
| 17/05/2015 | 8 |
| 24/05/2015 | 8 |

| Date of weekly list | Ranking in list |
|---|---|
| 31/05/2015 | 9 |
| 7/06/2015 | 9 |
| 14/06/2015 | 8 |
| 21/06/2015 | 9 |
| 28/06/2015 | 11 |
| 5/07/2015 | 12 |
| 12/07/2015 | 11 |
| 19/07/2015 | 12 |
| 26/07/2015 | 12 |
| 2/08/2015 | 11 |
| 9/08/2015 | 10 |
| 16/08/2015 | 9 |
| 23/08/2015 | 9 |
| 30/08/2015 | 11 |

| Date of weekly list | Ranking in list |
|---|---|
| 6/09/2015 | 13 |
| 13/09/2015 | 14 |
| 20/09/2015 | 20 |
| 27/09/2015 | 17 |
| 4/10/2015 | 17 |
| 11/10/2015 | 13 |
| 18/10/2015 | 20 |
| 25/10/2015 | 16 |
| 1/11/2015 | 20 |
| 8/11/2015 | 20 |
| 29/11/2015 | 19 |
| 13/12/2015 | 18 |
| 20/12/2015 | 1 |
| 27/12/2015 | 12 |

(i) Describe THREE graphical techniques that you think have been used in the visualisation to convey information about these best-selling fiction books.

Explain how you believe each graphical technique should be interpreted.

(ii) Write a short paragraph describing what the visualisation reveals about these 10 best-selling fiction books, including one or two particular examples.

93201Q